4 Hello, Data Journalism

We are at an exciting moment, when every field has taken a computational turn. We now have computational social science, computational biology, computational chemistry, or digital humanities; visual artists use languages like Processing to create multimedia art; 3-D printing allows sculptors to push further into physical possibilities with art. It's thrilling to consider the progress that has been made. However, as life has become more computational, people haven't changed. Just because we have open government data doesn't mean we don't have corruption. The tech-facilitated gig economy has exactly the same problems as labor markets have had since the beginning of the industrial age. Traditionally, journalists have investigated these types of social problems to create positive social change. In the computational world, the practice of investigative journalism has had to go high tech.

Many of the people who push the boundaries of what is technologically possible in journalism call themselves *data journalists*. Data journalism is a bit of a catchall term. Some people practice data journalism by making data visualizations. Amanda Cox, the editor of the *New York Times* section "The Upshot," is a master of this kind of visual journalism. Consider a 2012 story, "All of Inflation's Little Parts," which led to Cox winning the American Statistical Association's award for excellence in statistical reporting. The underlying data in the story is taken from the consumer price index, which is compiled by the Bureau of Labor Statistics monthly and used to measure inflation. In the graphic, a large half-circle is broken into colored mosaic tiles. The different sizes correspond to the percentages of Americans' spending.

A large shape, gasoline, represents 5.2 percent of spending. It is part of a category, transportation, that takes 18 percent of the average person's income. A smaller shape, representing eggs, is highlighted as part of the 15 percent of income spent on food and beverages. “High oil prices and drought in Australia are among the factors that have made food prices rise faster than they have since 1990,” reads Cox’s text. “Strong European demand for eggs has also affected prices in that category.”[1] This text, and the intriguing shapes, open a window into the amazing ways that global citizens are connected via a complex web of trade. Eggs are global? Of course they are! Countries don’t produce all of their own food anymore. Food is a global trade market. Western Australia contains a massive wheat belt. Australia overall exported $27.1 billion in food from 2010 to 2011 according to the Australian government’s Department of Agriculture, Fisheries, and Forestry. The drought in the wheat belt led to less wheat production. Poultry feed in the United States is based primarily on cereal grains. Corn is preferred, but producers will use wheat if wheat is cheaper than corn. Less wheat available globally means that wheat prices are more expensive, which means that poultry feed producers either pay more for wheat or turn to also-expensive corn. As poultry farmers pay higher prices for chicken feed, they pass along the costs and charge higher prices for eggs. That increase is passed along to the consumer in the supermarket. The data provides a way to think through how a drought in Australia leads to higher egg prices at a North American supermarket, which is also a story about globalization, interconnectedness, and the environmental consequences of climate change. Cox uses her storytelling skills, her knowledge of how complex systems work in the world, her technological skills, and her keen design sense to create a visually exciting computational artifact that both informs and delights.

Other data journalists collect their own data and analyze it. In 2015, the *Atlanta Journal-Constitution* (AJC) gathered data on doctors who sexually abuse their patients. An AJC investigative reporter discovered that in Georgia, two-thirds of doctors who had been disciplined for sexual misconduct with their patients were permitted to practice again. This would have been enough for a story, but the reporter wondered if Georgia was typical or unusual. The story became a team investigation. The team gathered data from across the United States and analyzed more than one hundred thousand medical board orders from 1999 to 2015 related to disciplinary action against doctors. Their findings were shocking. All across the country, doctors were forgiven and allowed to resume practicing medicine after being found guilty of abusing patients. The very worst cases were horrifying. A pediatrician, Earl Bradley, was believed to have drugged over one thousand children with lollipops and molested them on video. He was indicted in 2010 on 471 charges of rape and molestation, and sentenced to fourteen life terms without parole. The AJC story led to awareness and positive reform, thankfully.[2]

In Florida, *Sun Sentinel* data journalists sat on the side of the highway and noted when police cars came by; they later requested the data from the police transponders at toll booths, and found that the police were systematically traveling at high speeds that endangered citizens. After the investigation, police speeding dropped 84 percent. This dramatic, positive public impact helped the story win the 2013 Pulitzer Prize for Public Service.[3] A lot of good data journalism comes out of Florida. For one thing, the narrative possibilities are endless. “Florida has long eclipsed California as the place where the bizarre, unusual and outlandish have become commonplace,” Jeff Kunerth wrote in the *Orlando Sentinel* in 2013.[4] Everything that the US government does is public by default, but Florida has “sunshine laws” that guarantee that the public has access, and tapes, photographs, films, and sound recordings are also considered public records. Great open records laws mean that it’s easy to get official government data, which means that a lot of data journalism is done in and about Florida.

Some data journalists get data from official sources and analyze it to find insights. These insights can lead to uncomfortable truths. For example, one example of a successful academic-industrial partnership came about when data journalist Cheryl Phillips of the Stanford Computational Journalism Lab organized a class project in which her students requested data on police stops from all fifty states. They analyzed the data for nationwide trends and released it online for reuse by other journalists. The Stanford journalists and all of the other journalists found that people of color were stopped far more often than white people in every state.[5]

Data journalism also includes algorithmic accountability reporting, which is the small corner of the field I occupy. Algorithms, or computational processes, are being used increasingly to make decisions on our behalf. Algorithms determine the price you see for a stapler when you're shopping online; they also determine how much you pay for health insurance. When you submit a job application or resume via an online job site, an algorithm generally determines whether you meet the criteria to be evaluated by a human or whether you're rejected outright. The role of the free press in a democracy has always been to hold decision-makers accountable. Algorithmic accountability reporting takes this responsibility and applies it to the computational world.

A prominent example of algorithmic accountability reporting is ProPublica's story "Machine Bias," published in 2016.[6] ProPublica reporters found that an algorithm used in judicial sentencing was biased against African Americans. Police administered a questionnaire to people who were arrested, and the answers were fed into a computer. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm then spit out a score that "predicted" how likely the person was to commit a crime in the future. The score was given to judges in the hopes that the score would allow judges to make more "objective," data-driven decisions about sentencing. However, this resulted in African Americans receiving longer jail sentences than whites.

It's easy to see how technochauvinism blinded the COMPAS designers from seeing how their algorithm might be harming people. When you believe that a decision generated by a computer is better or fairer than a decision generated by a human, you stop questioning the validity of the inputs to the system. It's easy to forget the principle of garbage in, garbage out—especially if you really *want* the computer to be correct. It's important to question whether these algorithms, and the people who make them, are making the world better or worse.

The practice of using data in journalism is older than most people think. The first data-driven investigative story appeared in 1967, when Philip Meyer used social science methods and a mainframe computer to analyze data on race riots in Detroit for the *Detroit Free Press*. "One theory, popular with editorial writers, was that the rioters were the most frustrated and helpless cases at the bottom of the economic ladder, who rioted because they had no other means of advancement or expression," Meyer wrote. "The theory was not supported by the data."[7] Meyer conducted a large survey and performed a statistical analysis of the results using a mainframe. He discovered that the participants in the riots came from a variety of social classes. He won a Pulitzer for his reporting. Meyer called the application of social science to journalism *precision reporting*.

Later, when desktop computers came to every newsroom, reporters started using spreadsheets and databases to track data and find stories. Precision reporting evolved to what became called c*omputer-assisted reporting*. Computer-assisted reporting is the type of investigative journalism used in the movie *Spotlight*, which dramatizes the *Boston Globe*'s Pulitzer-winning investigation into Catholic priests who were sexually abusing children—and the forces that were covering up the problem. To keep track of the hundreds of cases and hundreds of priests and parishes, the reporters used spreadsheets and data analysis. In 2002, this was state-of-the-art investigative practice.

As the Internet grew and new digital tools emerged, computer-assisted reporting evolved into what we now call *data journalism*, which encompasses visual journalism, computational journalism, mapping, data analysis, bot-making, and algorithmic accountability reporting (among other things). Data journalists are journalists first. We use data as a source, and we use a variety of digital tools and platforms to tell stories. Sometimes those stories are about breaking news; sometimes the stories are entertaining; sometimes the stories are investigative. They are always informative.

ProPublica, established in 2008, and the *Guardian* have been leaders in the field.[8] ProPublica, started by *Wall Street Journal* veteran Paul Steiger with philanthropic backing, quickly made a name for itself as an investigative powerhouse. Steiger has a deep investigative background: he served as the managing editor of the *Wall Street Journal* from 1991 to 2007, during which time members of the publication's newsroom staff were awarded sixteen Pulitzer Prizes. ProPublica reporters received the first of their many Pulitzer Prizes in May 2010. The organization's 2011 Pulitzer Prize for national reporting was the first such prize ever for stories not published in print.

Many Pulitzer projects have had data journalists, or people who identify as data journalists, on the team. Journalist and programmer Adrian Holovaty, who created the Django programming framework that's used by many newsrooms, published an online rant titled "A

Fundamental Way Newspaper Sites Need to Change” in September 2006.[9] Holovaty advocated for newsrooms going beyond the traditional story model by integrating structured data into reporters’ ordinary methods. His rant led Bill Adair, Matt Waite, and their team to create the PolitiFact fact-checking site, which won a Pulitzer in 2009. Waite wrote of the launch: “The site is a simple, old newspaper concept that’s been fundamentally redesigned for the web. We’ve taken the political ‘truth squad’ story, where a reporter takes a campaign commercial or a stump speech, fact checks it and writes a story. We’ve taken that concept, blown it apart into its fundamental pieces, and reassembled it into a data-driven website covering the 2008 presidential election.”[10]

Holovaty went on to make EveryBlock, a pioneering news app that integrated crime data and geolocation. It was the first to use the Google Maps API, leading Google to make the feature available to everyone.[11]

The *Guardian* broke ground in data journalism in 2009 when a team of reporters and programmers sought to review, via crowdsourcing, 450,000 records of expenses generated by members of Parliament. This effort was a follow-up to a scandal in which it was discovered that MPs were using government funds to pay for household and office expenses. The *Guardian* team also gained expertise in using computational methods to analyze large troves of leaked documents, as in their analysis of the Afghanistan and Iraq war logs.[12]

One important project in the field is an investigation by the *Wall Street Journal* into price discrimination.[13] Major chains like Staples and Home Depot were charging different prices on their websites depending on the zip code in which visitors seemed to be. The journalists used computational analysis tools to discover that customers in wealthier zip codes were being charged less than customers in poorer zip codes.

Academic research is an important complement to data journalism. Data journalists tend to rely on established scholarly research methods. Part of being a good journalist is knowing when to turn to a subject matter expert; another part is telling the difference between an expert and a shill. Data journalists synthesize expertise from a wide variety of fields. Georgia Tech professor Irfan Essa organized the first Computation + Journalism Symposium in 2008. At this annual event, journalists come together with researchers from communication, computer science, data science, statistics, human-computer interaction, visual design, and more to share their research and promote understanding. Northwestern professor Nicholas Diakopoulos, one of the conference co-founders, has written important works about reverse-engineering algorithms as part of holding decision-makers accountable. His paper "Algorithmic Accountability: Journalistic Investigation of Computational Power Structures"[14] describes some of his work and the work of other journalists in investigating algorithmic black boxes.

Though it is closely linked to computer science, data journalism is generally considered a social science. The most in-depth explorations of the field can be found in social science literature. C. W. Anderson published "Towards a Sociology of Computational and Algorithmic Journalism" in 2012,[15] in which he united Schudson's four approaches to studying news with ethnographic insights gained from fieldwork at a Philadelphia newspaper between 2007 and 2011. Nikki Usher contributed additional ethnographic context with her book *Interactive Journalism: Hackers, Data, and Code*,[16] which is based on both fieldwork and interviews with data journalists at the *New York Times*, the *Guardian*, ProPublica, WNYC (New York Public Radio), AP, National Public Radio (NPR), and Al Jazeera English. Cindy Royal's work on journalists producing code[17] was important for understanding how journalists used code inside the newsroom, and it also prompted understanding of how journalism schools could integrate computational skills into their curricula. James T. Hamilton's 2016 book *Democracy's Detectives* outlined how crucial data-driven investigative journalism is for the public good—and how much this public service can cost. High-impact investigative data journalism stories cost hundreds of thousands of dollars to produce. "Stories can cost thousands of dollars to produce but deliver millions in benefits spread across a community," Hamilton writes.[18]

In 2010, Tim Berners-Lee gave the new field the computational stamp of approval when he said: "Journalists need to be data-savvy. It used to be that you would get stories by chatting to people in bars, and it still might be that you'll do it that way some times. But now it's also going to be about poring over data and equipping yourself with the tools to analyse it and picking out what's interesting. And keeping it in perspective, helping people out by really seeing where it all fits together, and what's going on in the country."[19] By the time Nate Silver launched FiveThirtyEight.com and published his book *The Signal and the Noise* in 2012, the term *data journalism* was in widespread use among investigative journalists.[20]

As computers have evolved, human nature has not. People need to be kept honest. I hope that this book will help you think like a data journalist so that you can challenge false claims about technology and uncover injustice and inequality embedded in today's computational systems. Using a journalist's skepticism about what can go wrong can help us move from blind technological optimism to a more reasonable, balanced perspective on how to live lives that are enhanced but not threatened or compromised by technology.

Notes

[1] Cox, Bloch, and Carter, "All of Inflation's Little Parts."

[2] Hart, Robbins, and Teegardin, "How the Doctors & Sex Abuse Project Came About."

[3] Kestin and Maines, "Cops Hitting the Brakes—New Data Show Excessive Speeding Dropped 84% since Investigation."

[4] Kunerth, "Any Way You Look at It, Florida Is the State of Weird."

[5] Pierson et al., "A Large-Scale Analysis of Racial Disparities in Police Stops across the United States."

[6] Angwin et al., "Machine Bias."

[7] Meyer, *Precision Journalism*, 14.

[8] Lewis, "Journalism in an Era of Big Data"; Diakopoulos, "Accountability in Algorithmic Decision Making"; Houston, *Computer-Assisted Reporting*; Houston and Investigative

Reporters and Editors, Inc., *The Investigative Reporter's Handbook*.

9. Holovaty, "A Fundamental Way Newspaper Sites Need to Change."

10. Waite, "Announcing Politifact."

11. Holovaty, "In Memory of Chicagocrime.org."

12. Daniel and Flew, "The Guardian Reportage of the UK MP Expenses Scandal"; Flew et al., "The Promise of Computational Journalism."

13. Valentino-DeVries, Singer-Vine, and Soltani, "Websites Vary Prices, Deals Based on Users' Information."

14. Diakopoulos, "Algorithmic Accountability."

15. Anderson, "Towards a Sociology of Computational and Algorithmic Journalism"; Schudson, "Four Approaches to the Sociology of News."

16. Usher, *Interactive Journalism*.

17. Royal, "The Journalist as Programmer."

18. Hamilton, *Democracy's Detectives*.

19. Arthur, "Analysing Data Is the Future for Journalists, Says Tim Berners-Lee."

20. Silver, *The Signal and the Noise*.