CHAPTER 2

The Rise of Investigative Data Journalism

Andrew W. Lehren

My journey into data journalism did not begin with some massive database. It was a long way off before I would get involved in combing through secret diplomatic cables or the Snowden trove of spy documents. It was before I really appreciated how statistics could further my journalism, to the point where it would be using statistics that would lead to the recall of a quarter of a million Nissan Altimas because its airbags were blinding passengers. The arrests of those involved in defrauding a major railway's disability pension system. Or spark the Pentagon to overhaul the way it delivers medical care to its soldiers and their families. It was before I learned about computerized mapping, and how it could help uncover policing that disproportionally targets African Americans, or how real estate developers reshape flood maps that later haunt unsuspecting property owners when hurricanes strike.

I was at a small weekly newspaper in Philadelphia, covering City Hall and trying to find enterprise stories that the bigger media organizations were missing. I sometimes got a good tip. But those often went to the veteran reporters at other publications. A friend told me about Investigative Reporters & Editors, a non-profit in the United States that helps journalists learn how to dig. I went to a conference and learned

A. W. Lehren (✉)
NBC News, New York, NY, USA

O. Hahn and F. Stalph (eds.), *Digital Investigative Journalism*,
https://doi.org/10.1007/978-3-319-97283-1_2

how to do a better job combing through public records. I started checking out court filings more often, became sharper going through business documents and began examining campaign reports. That was a start.

Poring through an endless pile of dusty documents can be a slog. Piles of paper. What would keep me going would be the results. The questionable economic deals that would be spiked because of what I found. The massive government contract that would be rescinded. The politician who claimed his prowess with finances only to be shown to be not everything he advertised.

Then other journalists started saying that if I learned how to use a spreadsheet, I could begin to do a better job going through public records. If I learned how to use a database manager, I could go bigger. I paid my way to go to some of the first seminars for journalists in the United States to see if this would live up to its promise. I became hooked.

If you wonder why I mention going through documents, it is because that is the image I have kept in my mind in the more than two decades I have been doing data journalism. It contains fundamental points that I sometimes worry about with the facile way information is talked about sometimes these days.

The documents I pored through would sometimes be incomplete. Pages could be inexplicably missing. Or someone filled out government forms incorrectly. Or to really understand an inspection report, you needed to learn the codes mysteriously noted in a corner. You needed to understand exactly why the documents were collected, to begin with, and what might be hidden from public inspection.

You also needed to know these were not an end, but a start. They sparked more questions. They altered the way you would do interviews. And who you asked questions. More importantly, you had to keep in mind that a pile of documents was fundamentally boring to most readers. Your job was to find the story—the important journalism—and find an engaging way to incorporate it in the telling.

At one point at *The New York Times*, I was feeling a bit burnt out. I had finished contributing to one project that won a Pulitzer. Run ragged helping cover the massive BP oil spill in the Gulf of Mexico. And worked on several enterprise stories on US casualties in its wars in Iraq and Afghanistan.

I asked my editors if I could spend a couple weeks just writing some stories for sports. Just for a change of scenery. I have been a long-time runner. I had some half-baked ideas about marathons. I thought I could use a brief busman's holiday before diving back re-energized into weightier subjects.

But the last thing I wanted to do was a story that someone else had already done. I would leave the heart-warming feature about a runner who overcame obstacles for another day. I started casting around. What public records even existed about running? I thought of the phone apps that friends and family use to track loved ones in races. I called the organizers of the New York City Marathon to see if I could get the underlying data from the most recent race. It turned out to be the same data that newspapers and websites use to list results. Pretty boring stuff.

I put it into a database manager. About 50,000 records. And started asking questions to see if anything would pop up. For those unfamiliar with marathons, this is how it works in New York. You start off in Staten Island. Cross the arching Verrazano Bridge into Brooklyn. Cut straight into Queens. Take the 59th Street Bridge into Manhattan, where runners are greeted by throngs of cheering crowds. Turn up to the Bronx. Cut back down through Harlem and into Central Park. You finish having covered all five boroughs 26.2 miles. In one of the world's biggest running races.

What I could not figure out was why I could follow most runners throughout the race. But not all of them. The race organizers have computer chips for all the participants, no matter how slow. They could track them every few miles. I had times for almost everyone throughout the race. But not everyone. It could not be a faulty computer chip. The chips worked at the beginning and end of the races. But for five or more split points in a row along the course, there were no times.

I began wondering about other documents that are public. One feature of every marathon is the souvenir photograph. Get a picture to commemorate your accomplishment. And the photo sites are public. Of course. What better way to make it easy for you or your family to find that image. I started using it to check the runners with five or more missing split times. No images of them along those stretches of the course either.

Now I had some interesting questions for the race organizers. They quickly conceded: Welcome to the dirty underbelly of large marathons. Dozens and dozens of people cheat every time. Cheaters hop on subways in Brooklyn. Or bleed into the crowd greeting them when they enter

Manhattan. And then magically re-emerge in Central Park to collect their medals.

I had not seen that story before. This took me about a day or two to puzzle out. I could probably pitch this for the front page. I checked the archives of various publications. There had been occasional stories about the odd cheater here and there. But nothing about the scale. The sports editor was intrigued.

I suppose I could have filed something in a few days. Track down some of the cheaters. I had their names. Amusing. But I paused and asked, what's the harm? What's the foul? So some people feel compelled to cut corners and collect a finisher's medal they did not earn. The overall winners were not affected. I paused and started re-evaluating results. Then I realized that in subcategories, like the fastest man in the aged 60–65 category, the winners were being affected. I tracked down those involved.

When I reached Alan Miller, who should have known on race day that he won his age grouping for older men, he asked how I even figured out his story. I told him. He said, Listen, I've run the marathon 25 times. I get ready for it with my buddies from the New York Police Department. Afterwards, we all go out for drinks at Micky Mantles. One of the landmark waterholes then in New York. If I had known I had won that day, it would have been the biggest day of my life. The drinks would have been on me. I would have celebrated with my friends like there was no tomorrow. And now I have cancer. I may never run another marathon.

This is why I will comb through data, no matter how boring. To find the stories of the Alan Millers of the world. To find bigger patterns that other journalists may miss.

It is this mindset that I had used about a decade earlier working for NBC News.

A young unarmed black male, Timothy Thomas, had just been shot and killed by Cincinnati police. Riots broke out in the streets as African Americans protested what they argued was a criminal justice system that consistently worked against them unfairly. Front page stories followed the demonstrators. The protests dominated cable and evening news broadcasts.

The president of the news division asked me and another producer to spend a few weeks, and see if we could find a story others had missed. I watched, again and again, the dash camera footage of the officer in the minutes leading up to the shooting. All the officer really knew about

Thomas was a description and that he had 14 outstanding warrants. For a 19-year-old, that was a lot. Sounded like a hardened criminal. That was the impression in much of the coverage.

I thought I would check the court records while others covered the protests. It turned out every one of the warrants stemmed from a traffic stop. More than that, they were not typical citations. Not for speeding or running stop signs. They were for small things an officer would have a hard time knowing before he pulled over Thomas, like not wearing a seat belt and not having proper paperwork.

Was this unusual? I used Ohio's open records laws to get data from the Cincinnati police. When it came to moving violations like speeding, blacks and whites were pulled over at about the same rate. But when it came to these non-moving violations, blacks got these tickets three times as often as whites.

This took about two weeks to sort out. We told our boss. He said that sounded promising. Now see if this pattern held true in cities across America. A fantastic assistant, Ben Vient, helped me file open records requests for more than 15 cities we knew were required to keep similar data. Almost everywhere, the patterns held up.

Criminal justice experts told us the key ratios to examine. When police said disparate ticketing happened only in high-crime neighbourhoods, we mapped the data. That was not true. Blacks and whites were treated differently regardless of neighbourhood crime rates.

The result, produced with Jason Samuels and hosted by John Larson, was an hour-long documentary that was the most-watched programme that evening in America. It went on to win a slew of awards. More important, it raised compelling questions about policing.

In the wake of the Ferguson, Missouri riots in 2014, I revisited this work with Sharon LaFraniere at *The New York Times*. Sadly, the patterns still held true. We also delved deeper into statistics on searches, and found police in many places conducted speculative searches far more often on African American drivers. Particularly since the searches were supposed to uncover guns or drugs. The statistics showed that in most places, police were far better on average when they searched whites. When they searched blacks, which they did far more often, they disproportionally failed to discover contraband. Our story was centred on North Carolina, and the resulting front-page story led to a storm of public hearings about policing there.

Beyond database work, the statistics and mapping were important. Each discipline has its own key variables. Like preparing for any interview with a public official, knowing the right measures to check can lead to most revealing answers.

Several years earlier, when Sharon and I were combing through lawsuits that the United States lost in federal court, we found disturbing anecdotal cases involving military hospitals. There are 55 such hospitals around the world, set up to help more than 9 million service members and their families. We wanted to know the big picture. We asked the Pentagon for data and statistics. Every major civilian hospital chain maintains such metrics. The military countered that it kept no such information.

We fought a Freedom of Information Act request for more than a year. Eventually, we started getting data the military has said it did not maintain. When the Army surgeon general, who oversees all its hospitals, told us his facilities had no problems, we pointed out several with troubling surgical complication figures. Shortly the head of one of the hospitals lost that position. We learned about the importance of shoulder dystocia in birthing; how it is a marker for troubled deliveries. Checking those figures pointed out a disturbing pattern particularly among the military's smaller hospitals. The resulting series of stories led then US Defence Secretary Chuck Hagel to revamp oversight of military hospitals.

That was not my first time closely examining military numbers. Jim Dao and I were trying to turn around a quick look for *The New York Times* on military suicides when we noticed numbers not adding up. The published rates, and the testimony by military leaders before Congress, compared with overall suicide figures, did not make sense. We paused and dug deeper. It turned out the Pentagon had been systematically underestimating its suicide rate. It was not using established methods deployed by the US Centers for Disease Control, the touchstone for such calculations. This meant the Pentagon missed for years the rising suicides happening during the Afghanistan and Iraq wars, and if it was measuring these properly, it would have known it had a problem and perhaps instituted measures that could have saved some lives. Even more, it had neglected to even track suicides involving reservists and others. The Pentagon, in the wake of the stories, revamped its tracking of suicides.

I should pause here and be clear. I am just one of the hundreds of data journalists now doing this kind of work in the United States and around the world. If anything I write can encourage more to enter this field, then the writing of this chapter is worthwhile. I know there are plenty doing incredible work and do not mean to put mine ahead of any others, only to say that, if anything, it is part of a larger mosaic. Also, there are visionary journalists before me who did groundbreaking work well before me. There was Phil Meyer, who won a Pulitzer examining the forces behind the race riots in Detroit in the 1960s. There were Barlett and Steele, two Philadelphia Inquirer reporters who looked at how judges were meting out sentences in the 1970s. They used computers, but to feed in data, they had to use punch cards stacked for old mainframe machines.

There were nurturing stalwarts like Pat Stith, a North Carolina journalist who helped IRE organize its first computer-assisted reporting conference. Stith would go on to win his Pulitzer exposing how corporate hog farmers had hijacked state oversight and were damaging the environment.

Sometimes journalists can pause and wonder if using some new technology is a proper way to do reporting. The debate happened more than a century ago when the telephone was introduced. Knowing how to cull data is just one more tool.

Journalists can approach data differently than those more trained in computer sciences. Take, for instance, matching databases. Traditional IT managers compare data sets that were designed to talk with each other. Journalists may wonder if the payroll list of school teachers includes registered sex offenders.

When *The New York Times*, along with *The Guardian* and *Der Spiegel*, was given access to the Iraq War Logs funnelled by Chelsea Manning to Wikileaks, I joined the team of reporters going through the data. Unlike the later trove of diplomatic cables, these were pocked with military abbreviations and jargon. Some reporters went through record by record, looking for a big find.

Knowing database tools led us to check the completeness of the files. Unlike the earlier Afghanistan war logs, we could find almost every incident involving a US casualty. That was reassuring, particularly since we could not expect the Pentagon to vouch for the completeness of a purloined database. Jim Glanz and I wondered about violent incidents involving private security contractors. Most readers knew about a particularly controversial firefight involving Blackwater. Jim, a former Baghdad bureau chief for *The Times*, long suspected there were more incidents than officials acknowledged.

I stepped away from the secret room where we had sequestered the data and began going through public records. Audits, contracts, hearing transcripts, news accounts. I assembled a list of more than 100 private security contractors that were known to be involved in Iraq. I took that database and flung it against the text inside the war logs. I then ran more queries against the war logs to see if we missed others. Jim and I were then able to stitch together a front-page story about how indeed there were far more incidents than the public had ever been made aware of. To tell the story, we mined social media. Barbara Gray, an ace researcher at *The Times*, trolled LinkedIn resumes to find contractors that had been present at the incidents we now could document. That helped confirm what we were seeing, and helped provide a more compelling narrative that the staccato abbreviations in so much of the data.

We took similar approaches with the diplomatic cables, stitching together from seemingly unconnected passages how companies around the world had helped North Korea's missile programme and provided precursor ingredients for Syria's chemical weapons programme. Seemingly unconnected records helped us document how unregulated tainted Chinese chemicals were becoming part of the world's brand-name pharmaceuticals, sometimes with deadly consequences.

You may have noticed a pattern. Because data journalism can be time-consuming and complicated, you can get perhaps the best results when you work with colleagues to share the burden. I cannot acknowledge enough all the great reporters I have had the chance to work with on stories like these. More minds may lead to better questions for the data, and then stronger stories. As more newsrooms collaborate on data-driven stories, the depth of the reporting only increases.

When I was running IRE's database operations in the mid-1990s, I saw the benefits of many reporters examining the same data. That small staff included five journalists who would go on to be involved in Pulitzer winning journalism.

Another pattern: the tools keep evolving. Mining social media is becoming ever more important. Programmes that can help tease out connections in social networks, and can help report on people, companies and governments are playing a larger role. I recently rejoined NBC News as a senior editor on its investigative team. The organization now has several journalists whose sole job is to comb through social media and find information. The tools and methods are only expanding.

Data journalism can be a great leveller. It helped me find stories way back in Philadelphia that my more established rivals had not found. I teach a class at a New York journalism school, and I have seen my investigative reporting students find big-picture patterns in data for stories that their more seasoned competitors overlooked. I look now in awe at the international growth in these skills transforming journalism in countries across the world. The Panama Papers is evidence of that. So is the work at the Global Investigative Journalism Network and so many other investigative reporting centres spanning the world.

The types of tools out there keep increasing. If it ever seems daunting, I think back to that small newsroom in Philadelphia, and remember: It is all about finding stories that matter.