Journalist's Resource
Research on today's news topics

HARVARD Kennedy School
SHORENSTEIN CENTER
on Media, Politics and Public Policy

Research chat: Sarah Cohen of the New York Times on the state of data journalism and what reporters need to know

By John Wihbey

August 29, 2014

Sarah Cohen is a leading practitioner and educator in the field of data journalism, and she now serves as both editor of computer-assisted reporting at the *New York Times* and board president of Investigative Reporters & Editors (IRE). She was the Knight Chair in computational journalism at Duke University prior to joining the *Times*. As part of a reporting team at the *Washington Post*, she earned a Pulitzer Prize in 2002 for an investigation into the child welfare system in the District of Columbia. She continues to teach part-time at the Columbia University Graduate School of Journalism.

Journalist's Resource recently caught up with Cohen to ask about the knowledge she's developed both in the classroom and in the field. The following is an edited transcript:

———

Journalist's Resource: What are your reflections now that you've been out of full-time academia for a few years, having gone back into daily journalism?

Sarah Cohen: I run the reporting team at the *Times* that does data journalism, so I'm not in the graphics or the interactive groups, and not in "The Upshot." We have four different data-journalism teams. It's very specialized here.

One of the things we were trying to do at Duke was to simplify working with difficult datasets and simplify, in particular, unstructured data — documents and audio and things like that. We learned a lot, but no matter what, the tools just aren't working. Even things that seem like they might work don't. It's amazing to me that, 10 years in, we still don't have a decent way to deal with PDFs. So much of our job remains trying to get public records from one format to another. It's just so frustrating, because you want to get on to the side of doing the actual analysis and 90% of your time is spent getting things from one format to another.

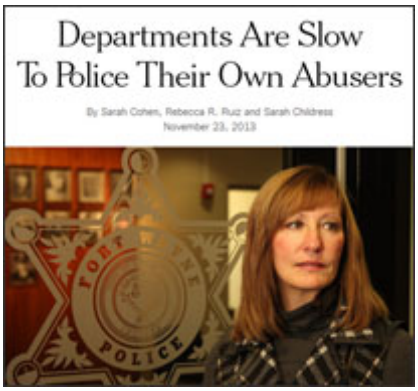
In my work at Columbia, where I'm an adjunct, I've really come around to the view that, yes, students need to know how to do analysis and presentation. But their biggest obstacle is going to be getting

data in the right form.

JR: There was a piece in the *Times* recently on just that issue: "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights."

Sarah Cohen: I was glad to see that we in journalism are not alone, right? I've been playing around with machine learning on some difficult document sets, while a member of my team is doing it the old way. I'm trying to do it the new way and see if it works. One of the ways we're different from the data science people is that data journalists tend to only use a dataset once. We're not dealing with a huge dataset over and over again. We tend to deal with data of modest size — maybe 20,000 or 100,000 records, not millions and millions. And we're only looking for one thing about them. What we're finding is that these techniques are really only time-efficient if you're going to be repeating the same analysis over and over again, not a one-time thing.

JR: So, for example, somebody who is doing data for a health analytics firm might spend the time to master techniques that are very complex, but there's a payoff because they will be using them again and again. But you're saying a journalist must be more versatile, and it might not make sense to invest heavily in certain complex data techniques.


Departments Are Slow To Police Their Own Abusers

By Sarah Cohen, Rebecca R. Ruiz and Sarah Childress
November 23, 2013

Sarah Cohen: That's very much what I mean. I do think it's worth it to know what sorts of techniques are out there. One example is a story that we were involved in last year on Chinese art, and part of it was about forgeries. We had a whole bunch of images scraped from a website of sales of Chinese art and we were sure that there was some way to know if two images were the same. Google knows what the Eiffel Tower is, for example, based on just the image. There must be an algorithm, but we didn't know how to do it. One of the things that's super important for people going into the newsroom now is to have in their field of vision some knowledge of what may be possible. A lot of the people we spoke to said, "You can try to do this, but the tools don't work." Well, what they meant was that it was only 99% accurate. That's OK for us — we'll look at all the results. One of the things we're trying to do is keep our eyes on the horizon for what is possible.

JR: For a cub reporter — not necessarily a data journalist, but someone who can at least speak your language — what's the basic set of tools and competencies she should have when she shows up at Sarah's desk?

Sarah Cohen: If they are concentrating on street reporting and writing, I would like them to have a pretty good familiarity with public records and how to find and request them for data purposes. Most reporters, even if they have some experience with public records, have really never tried to negotiate for databases and they don't realize how different it is. So we end up, too many times, having to go in behind somebody and re-request records because they weren't efficiently requested in the first place. The second thing I would like them to have is some imagination about what is possible, and at the same time an idea about what the limits are. One of my team members has a little statue on his desk that he calls the "data unicorn." There is this idea that there is this "data unicorn" out there that you

and that we shouldn't be teaching closed-source tools. But it is still on everybody's desktop; it is still used across government and will be the way they will be giving you data. And the alternatives are still not great.

I don't need people who know how to code for publication on the Web. I don't need them to know how to make an app that will scale to our audience. That's a whole other job here at the *Times*, and it is in most places. You still usually don't have most people doing the reporting and the production together in the same job. If they want to do reporting, I would say they should come in with some ability to program, enough to scrape a reasonably difficult website. I don't need them to be able to make websites — though it's helpful if they know how to make small-scale applications for internal data sharing — but I do need them to be able to get data from websites. That's a key skill right now, and it's not easy to do because websites are getting more and more complicated.

The last thing that I'd like to see from them is interest in something more data science-y. What I mean is, maybe, exploratory data analysis using R with graphics — trying to visualize data. Or geographic or network analysis. It might mean exploring machine learning or natural language processing. I don't need them to have mastered any of this, but I'd like to see a clear indication that they have an interest in moving toward the more sophisticated work, including high-level statistical methods beyond simple regressions. Our hope is that, for the simple stuff, we can build tools for reporters who don't want to learn to do it for themselves — or that we can teach them to do it for themselves. I just did an Excel class for our investigative group in metro. The reason we want people to be more self-sufficient is so we can do the things that nobody else is doing, the things that are more sophisticated and have more promise. So I want people who are interested in doing that.

JR: Let's get technical for a moment. In terms of scraping websites, which language or technique do you want them to use? And what do you use?

Sarah Cohen: I don't care what they use. I've seen people do scraping with Visual Basic, I've seen people use Python. We have someone on the team who uses Ruby, and one who still uses Perl. I use Python. I've tried to get more into R, but I haven't yet gotten to the point where I can use it with dirty data. The reason I go with Python, as opposed to Ruby is that it has more statistics and data science built into it. It does matrix algebra pretty easily. There are tools for Python that make it a pretty good fit for what we do.

JR: Do you use Google data products, such as Fusion Tables?

Sarah Cohen: No, mainly because our datasets are too big for Google spreadsheets, no less Fusion Tables. They have a pretty strict limit on the number of cells you can have — I think it's 100,000 cells, which is not that much data. We might use other tools from Google for visualizations. We still are using ArcView for mapping, though we might switch to some of the other open source tools over the next year or so. In general, though, we're pretty tool agnostic. As part of IRE, everybody can get a free copy of Tableau, so I'm trying to learn more about that.

JR: Speaking of IRE, why should everyone sign up?

Sarah Cohen: As its president, I think everyone who is interested in learning more about in-depth reporting can get a lot out of IRE. There is a growing body of online training, practice datasets and tutorials to learn Excel if you feel like you're behind on that. All of the tip sheets from conferences are available on the website, like how to find data on your beat. They just put out a tip sheet on all of the datasets relating to the militarization of local police departments. They'll put out things on breaking news topics. When you are exploring a story, you can search the database of contest entries and find detailed descriptions of how other reporters overcame obstacles, and as an IRE member you can usually call those reporters if you need more help. It's a great deal for the $25 annual membership fee for students.

JR: Trying to look over the horizon, how can educators best prepare students for the next decade or so of newsroom changes and demands? How will data journalism evolve and the newsroom shift accordingly?

Sarah Cohen: For about a decade, people who were doing the reporting were moving away from the people who were doing production. Graphics, interactive and reporting were pretty separate jobs, but I'm seeing those coming back together a little bit. Not all the way, because the skills really are different. Some people totally disagree with me, but I think you need to know what you want your job to be and school is a great time to explore the different options, even within data journalism. For instance, at "The Upshot," or FiveThirtyEight or Vox — if that's what draws you to data journalism — it's really closer to science writing than it is to doing traditional computer-assisted reporting. The tools for visualization have just gotten so much better for telling stories. If you get really into that, that can be a whole career, and it won't go away anytime soon. And frankly, our data-journalism reporting job won't go away. Maybe in the newsroom of the future, we'll be called the "data science team" or something like that.

I sometimes work in our R&D lab, and more often with our interactive and graphics teams. I have even worked with people on the business side who are doing data science. They can teach us a lot. There was someone working on the cooking app who just got really interested in a problem I had and she was incredibly helpful — she showed me how to tweak several algorithms and got something from 30% accurate to 90% accurate in one day. And she could show me how to do it. There are people all over the building who are really skilled, and they love working with the newsroom. So we need to be able to look around and find people who can help us.

This segues finally into possibly the most important characteristic students need: to effectively work in teams. There are very few jobs that you can do by yourself anymore. If you can learn to work with a team, to communicate with a team, share with a team, that's going to be the key thing across all newsrooms. It's already pretty important.

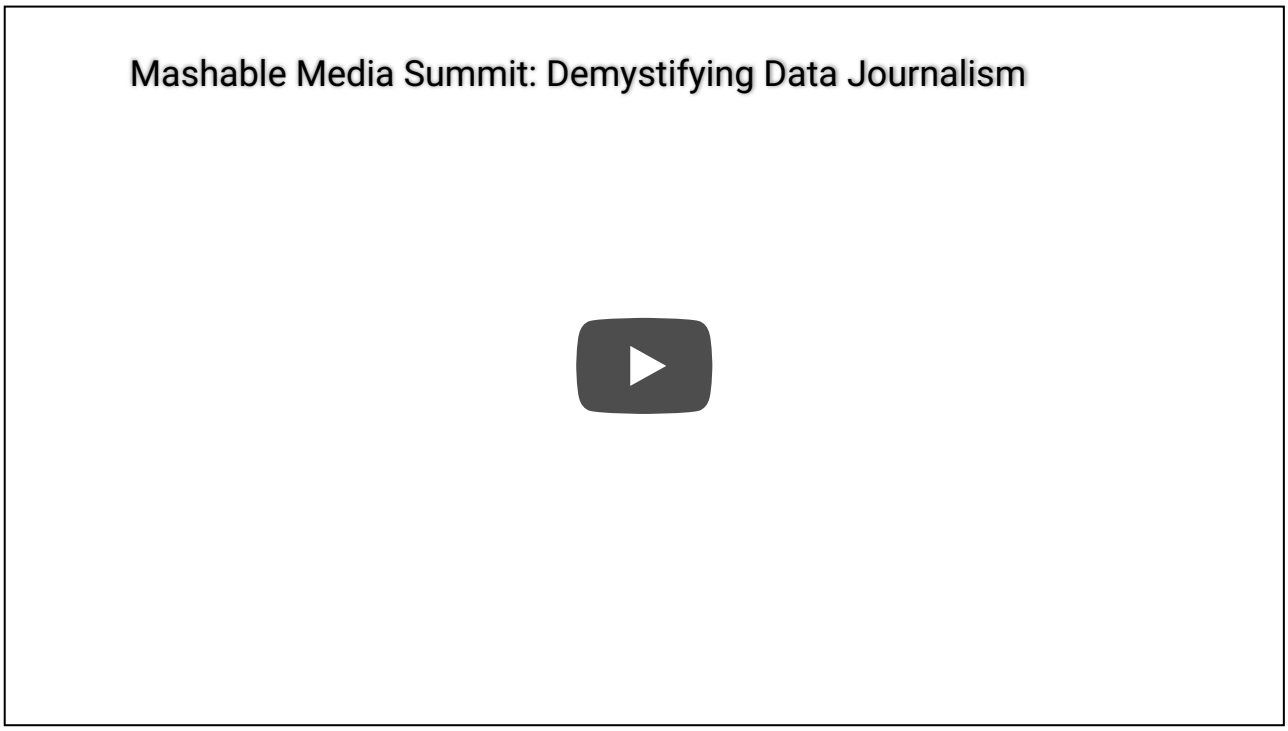
Related: Read Cohen's co-authored paper "Computational Journalism," for Communications of the ACM, in which she discusses the "ability of computer scientists to strengthen the hands of the

can just shake and it will come out with your answers. It's also critical that reporters be at least conversant with a spreadsheet. You'll find people losing patience with you if you need an "expert" to sort a list or to do the most basic calculations.

JR: Is there a recent *Times* project that is an example of the kinds of public records reporting you are pointing to?

Sarah Cohen: I did a story with a reporter here, Ginger Thompson, on immigration that involved a fairly straightforward data analysis on deportations and returns going back about 20 years — "More Deportations Follow Minor Crimes, Records Show." We didn't get the ideal dataset — maybe 2 or 3 million records in crummy old spreadsheets — but they weren't that hard to put together and analyze. That story is an example of how we can get much more out of individual records, where there's a line for each deportation, versus working from government statistics.

See Sarah Cohen discuss data journalism:



JR: What skill set would you like to see a young data journalist show up with, someone maybe even just out of J-school?

Sarah Cohen: The first priority is reporting rather than data. Ideally, I'd like to see a young data journalist spend two or three years in a beat reporting and writing job, or at least having to produce three or four reported, edited pieces a week. Sometimes that background isn't necessary, but it changes how you view any data-journalism job. Then, no matter what others say, a strong facility with Excel is pretty much a baseline. It's the tool of choice for so many *other* people that if you are not proficient with it, you really can't even get started. That doesn't mean that you need to be able to write macros or program in VBA or anything like that, but you should be proficient in data cleaning, sharing data, doing things like PivotTables. That's a given. I know there are people who think Excel is terrible

remaining professional reporters and engage new players in the watchdog process."

Keywords: data journalism, reporting, computer-assisted reporting, research chat

We welcome feedback. Please contact us here.

Foundations, News Media, Reporting, Research data journalism, research chat