By MaryJo Webster
@MaryJoWebster
Mjwebster71@gmail.com

What is data journalism?

There are several terms for what we're going to learn this semester. The one that has been in use the longest is "computer-assisted reporting." But nearly everyone — even it's most ardent practitioners — don't like the term.

When CAR was first coined in the late 1980s, its definition largely fit the computer technology of the day. It encompassed everything from looking up a single piece of information using online databases (a predecessor to the Internet), to doing some math calculations in Excel, to analyzing large government datasets in database software.

During this era, most of the CAR work was done by investigative reporters who realized they needed to learn how to navigate spreadsheets and database software because increasingly the government records they relied on for their investigations were being converted to computer systems.

One story that helps explain this huge transition is that before computers, candidates for public office filed their campaign finance reports (listing each donation) on paper. Journalists who wanted to analyze the donations for patterns — such as whether a candidate was being unduly influenced by a particular sector of industry — would copy each donation onto an index card, noting the industry that the person worked in.

After creating hundreds, potentially thousands, of cards, the journalist would lay them out on a large conference table and start making piles. If they wanted to look at how much money came from each industry, they would put the cards in piles based on the industry. Then use a calculator to tally up the dollar amounts for each pile. Then put all the dollar amounts on a piece of paper, listing them in order from highest to lowest. This process could take days, if not weeks!

With computers — using the skills that we will learn in this class — journalists can now get those campaign finance reports electronically and can do that same analysis in a matter of hours.

You'll also hear the term "precision journalism," which was coined by Phil Meyer, who is considered the grandfather of analyzing data for journalism purposes. He did sophisticated analyses using mainframe computers back in the 1960s and 1970s, then wrote a seminal book, "Precision Journalism" about using social science methods (i.e. regression analysis, conducting surveys, etc) for journalism. (Meyer recently published his autobiography, which includes some great insights into how he did those stories. The book is called "Paper Route.")

The use of social science methods is still being used among the CAR or data journalism community. A couple recent examples:

USA TODAY collected soil samples and hired a lab to test them for a project they did in early 2012 about old lead smelting sites that were being neglected by state and federal environmental agencies. http://www.usatoday.com/news/nation/smelting-lead-contamination

The Atlanta Journal Constitution did an impressive project that relied on regression analysis of school test scores to look for potential cheating in schools throughout the U.S. http://www.ajc.com/news/cheating-our-children-suspicious-1397022.html

More recently the term "data journalism" has come into use. This one aims to encompass a wider variety of ways that journalists with "data" skills are contributing to news products. This includes analyzing data to look for patterns and trends (just like CAR), but it tilts heavily into the digital world where data skills are in demand for building web applications (searchable databases, interactive maps and graphics, etc). The "Data Journalism Handbook", a recommended reading for this class, approaches the topic from this angle.

What kinds of stories can I do?

Education:

--Analyze data comparing student demographics (race, ethnicity) with teacher demographics to look for schools where the demographics generally match and where they are quite opposite.

--Analyze data on teachers to see if the least experienced/least educated teachers are clustered in the poor schools.

Local government:

--Analyze changes in the annual budget, comparing back to previous years

--Study campaign finance donations to candidates for mayor, city council, etc. You can look for patterns of influence (by the industry/occupation of the donor) or look to see how much money is coming from outside the city or outside the councilor's district.

--Analyze contracts that local government has with private companies. You can look for unusually high-priced contracts, or study whether or not the city is living up to its promise to ensure minority- and women-owned businesses are getting contracts; or you can look to see if friends or family members of city officials are getting an unfair advantage. How you approach this would largely depend on the issues at play in the community you're covering, at that time.

--Compare government employee salaries for several equally-sized communities to see if the pay is the same or different for similar jobs

--Get data on rental property inspections (some cities don't have such a program) to find "problem" properties and to study whether or not the city is being harsh enough on problem landlords.

Public safety:

--Get police incident data to see where the police are spending the most time and what types of incidents they are most often responding to. Compare back to earlier years to see if there are significant changes.

--Analyze jail bookings data to see how much time/resources the sheriff's department spends on repeat offenders (particularly those who seem to show up nearly every week)

Business/Economy:

--Get records on foreclosures in your county to see where they are occurring. Look at data over time to see rise and fall in total number of foreclosures.

--Build a database of salaries of CEOs of large companies, comparing their full compensation packages (bonuses, profit sharing, stocks, etc) and seeing how that has changed over time

Sports:

-- Build a database of salaries for high school football (or other sports) coaches in your area. You could compare the salaries against each coach's win-loss record to see which district is getting the most bang for their buck.

--Get sports participation data (this is generally reported to the state high school athletics organizations that runs the state tournaments) to look at participation rates over time or comparing among sports.

Community:

--Use census data to look at the changing demographics of one or more communities. Is it becoming older? More diverse? What are the effects on a community that is aging — what changes will the school district, city, county, etc., need to make to adjust? Same with diversity? Is there a big exodus of young people, particularly well-educated young people (what's known as a "brain drain")?

To find examples of recent stories that have published using CAR skills, check out the "Extra! Extra!" blog from Investigative Reporters and Editors. The blog links to investigative work by news organizations around the world, but much of it involves CAR on some level. Look for stories that refer to data analysis or similar terms. http://ire.org/blog/extra-extra/

Why should I learn this?

Regardless of what title you use --- CAR or data journalism or database reporting — the key fact is that there is a huge demand in journalism today for people with data skills.

First, most of the story ideas listed above would either not be possible or would be very limited in scope if a journalist could not analyze the data themselves. Sure, there are lots of government reports about those topics. But how current are those reports? Do they ask the specific question that you have on your mind? Are they available for the specific city or county or school district that you're covering?

Many times the answer is no.

In addition, those reports often hit the highlights. As a reporter, you might want more detail. Analyzing the data yourself allows you to summarize the information (just as those reports do), but also see all the details. This can be especially helpful for finding people or specific anecdotes to help bring your story to life.

Second, stories that rely on your own data analysis will automatically be unique. It will set you apart from what everyone else is doing and almost certainly guarantee you a spot on the front page or the top of the most-read list or the top of the TV news broadcast.

In this era of layoffs and limited job openings, being unique and having special skills is the only way to survive.

Third, there are many new positions being created in the data journalism realm, but not enough people to fill them. Unlike the rest of the journalism industry, this area is growing.

The tough part is that it requires learning a lot of skills. Think of it like the difference between training to be a family practice doctor versus training to become a surgeon.

Here are a couple of recent job postings:

CAR Residency at Chicago Tribune:

The Chicago Tribune is seeking a CAR journalist who can serve a 2-year residency for the newsroom. The candidate for this Computer Assisted Reporting position should have proven watchdog skills and experience with every facet of story development, from filing FOIA requests to finishing a polished draft of a newspaper story and its digital package. Clips should show proven results of this kind of work.

A key part of the job is to work with suburban reporters assigned to specific towns to tell stories that can be told only by those with command of hard data. This journalist also will be looking independently for enterprise stories.

This reporter also will be a member of our CAR team and will have the opportunity to work on other projects with seasoned database reporters.

USA TODAY Data Journalist:

USA TODAY's data team has an opportunity for an experienced journalist who's skilled at mining stories out of data and documents. Our team works on topics ranging from sports to the Census, the economy, health, campaign finance, education and entertainment. You'll join some of the news industry's most experienced data journalists to pursue daily and investigative stories solo and with teams of reporters. The job is based at our headquarters in McLean, Va.

Preferred qualifications:

-- Bachelor's degree in journalism or relevant major.

-- At least three years of experience in daily journalism.

-- Strong reporting and writing skills. You've filed FOIA requests and successfully negotiated with government agencies for public records, including data. Your writing is crisp and lively.

-- Creative thinker. We will favor candidates who have a track record of unique enterprise, of creatively using data to find stories, and who know how to monitor data routinely to find trends.

-- Strong collaborative skills. You're as comfortable working with a team as you are working solo, and you enjoy helping reporters and editors grow in data journalism skills.

-- Strong data analysis skills. Experience using tools such as SQL database managers, SPSS, R, SAS and/or Excel to interview data, find trends and quantify them. We'll favor candidates who have experience with statistical tests such as regression or correlation.

-- Strong data handling skills. You know how to use tools or a programming language to clean dirty data, scrape a website or fetch data from an API. You've left Microsoft Access in the dust in favor of raw SQL. You know a scripting language such as Python or have a strong desire to learn it.

-- Strong mapping skills. You've used ArcGIS or QGIS or another mapping platform to find spatial patterns.

Bonus points for experience with Python, Django, C#, ASP.NET, JavaScript, SQL Server.

You probably just read that last job posting and had your eyes glaze over. There are a lot of things mentioned in there that sound like a foreign language.

The skinny is that the USA TODAY job is for someone who is a veteran in data journalism and even though all those skills are listed in that job posting, the guy who did the hiring for that acknowledged that they know they aren't going to find someone with ALL of those skills. The posting represents their ideal candidate. The skills are listed in that job posting in order of importance. Note that traditional journalism skills are at the top.

That's the trick that data journalism is struggling with right now — finding the balance between traditional journalism skills and high-tech computer skills that are closer to an IT job. Ideally, news organizations want someone who is comfortable on both sides.

So where do you start?

You start by learning how to approach stories with a different frame of mind. I refer to it as a "data state of mind." Basically, this boils down to thinking about data as a source — just like your human sources and the web pages that you frequently visit — and also learning how to do stories where you analyze data to measure or quantify something (instead of relying on someone else to do it)

Traditional journalism relied heavily on reports and studies generated by others — nonprofits, government agencies, auditors. Now, journalists have the ability to do those same kinds of analyses BEFORE the others, or tackle topics that the others are not getting to.

This is especially crucial for "watchdog" journalism. We can use government data to make sure our government organizations and leaders are doing their jobs properly and to look for other problems, such as corruption, fraud, waste, etc.

In conjunction with developing a "data state of mind," you can start learning how to use the various tools. The tools can be broken down into different types:

--Tools for data analysis

--Tools for data cleaning

--Tools for publishing data online/creating web applications

THE TOOLBOX:

Data Analysis Tools:

Most people start by learning to do some simple analyses with spreadsheet software, such as Microsoft Excel.

I refer to Excel as a "gateway drug," because it's quite easy to learn, yet it can yield some powerful results and get you hooked.

Excel has some limitations, though, so journalists who want to tackle large amounts of data or something that requires a little more sophistication will need other tools.

In this class, we're going to learn Microsoft Access, which is a database manager. Like Excel, it stores data in rows and columns, but you can store much larger amounts of data and you can do things that Excel cannot.

Beyond this class, here are some other data analysis tools that are available — some that are quite expensive and others that are open source/free:

1) More powerful database software such as Microsoft SQL Server or MySQL (open source). These are similar to Access but are a bit harder to learn and they are much more powerful. For example, Access will have trouble with very large datasets (more than 1 million records) but SQL Server and MySQL won't. SQL Server and MySQL can also be used as a place to store data that will be published on the Internet (Access can't do that).

2) Mapping software such as ArcGIS (made by ESRI) or Quantum GIS (open source). These tools allow you to do analyses that look for geographic patterns in data. This is different than doing a Google map. These tools allow you to look for "hotspots" or calculate things such as population per square mile. And the most powerful piece is to be able to match data with boundaries. For example, if you had data listing each crime report with an address, you could plot all of those locations on a map, then have the mapping software calculate how many incidents occurred per square mile in each neighborhood.

3) Statistical software such as SPSS, SAS or R (open source). These are the tools for doing the social science methods that Phil Meyer encourages in his book, "Precision Journalism." The most common type of analysis that journalists use is a regression analysis, which allows you to determine if one or more things has a relationship with something else (and how strong that relationship is). For example, regression can tell you that poverty has a very strong relationship on school test scores — the higher the poverty, the lower the test scores.

Data Cleaning Tools:

This is where you start getting into tools that IT professionals are more likely to use, but increasingly we are getting free and easier-to-use tools.

1) Open Refine (formerly Google Refine) This is a free tool that is designed to do data standardization (such as cleaning up the name variations). http://code.google.com/p/google-refine/

2) A text editor such as UltraEdit or Programmer's File Editor (PFE). These allow you to open data files before you import them into spreadsheets or databases. UltraEdit has a pretty steep price tag, but PFE is free, http://www.lancs.ac.uk/staff/steveb/cpaap/pfe/, and there are many others out there.

3) Regular expressions. This is a sort of programming language that allows you to do pattern matching. You would use another tool — such as a text editor or a programming language — to make it work. For example, there are regular expressions that allow you to do a massive "find and replace" looking for odd things like "every instance of any word followed by any number."

4) Programming languages: Python and Perl are probably the most common in use among the CAR community right now. VB Script is an older one. JavaScript is another one that is widely used for web programming. The main reason you might want to write a program is to repeat a task. Let's say you have a dataset that you request from a government agency and you get a new set of data every month. You could write a script that could do any necessary data cleaning and/or import into your database software. You can also use Python and Perl to "scrape" data off web pages (this is useful in situations where government agencies refuse to provide you the data, yet they post it on their website)

5) Excel or Access (or other database software). There are ways to do simple data cleaning in any of these programs, but that's not their strength.

6) Mr. People: This is a cool little tool created by a data journalist at the NY Times. It will standardize names (like the campaign finance example I gave above).
http://mrpeople.ericson.net/

Tools for publishing data:

There are far too many options in this category to cover all of them here, but I'll hit the highlights (particularly the free ones).

1) Google Fusion Tables: You can build interactive maps and some simple graphics, then "embed" them in a website. The out of the box tools are quite easy to learn (and we'll learn them in class), but you can also get much more sophisticated by learning some JavaScript.
http://www.google.com/fusiontables/Home/

2) Tableau Public: Great for building interactive graphics, and some simple maps. Free for journalists. This is a powerful tool that doesn't require programming skills. There's a bit of a learning curve, but not too bad considering the results are very slick and professional looking.
http://www.tableausoftware.com/public/community

3) Various tools for building timelines, graphics, etc. There are many possibilities for this group; here are just a few: Google Charts API, Timeline JS (http://timeline.verite.co/), OpenHeatMap (http://www.openheatmap.com/)

4) freeDive: a free tool for building searchable databases on the web.
http://multimedia.journalism.berkeley.edu/tools/freedive/

5) Caspio: This is a fee-based service that allows you to build simple searchable databases and embed them on your website. It was designed for journalists and is widely used by news organizations. Easy to learn, but has some limitations in terms of size of data and sophistication of the search functions you can provide readers. http://www.caspio.com/

6) Various "frameworks" or programming languages for building web apps: These require programming skills. Includes ASP.NET, Django, and PHP. With these tools you can make more sophisticated searchable databases than what you can do with Caspio, but the learning curve is much steeper. Django, https://www.djangoproject.com/, is particularly catching on in the CAR community.

7) CartoDB – tool for creating interactive maps
http://cartodb.com/