Visualization as the workhorse of data journalism

Sarah Cohen
Knight Professor of the Practice, Duke University
March 2012

Before you launch into trying to chart or map your data, take a minute to think about the many roles that static and interactive graphic elements play in your journalism.

In the reporting phase, visualizations can:

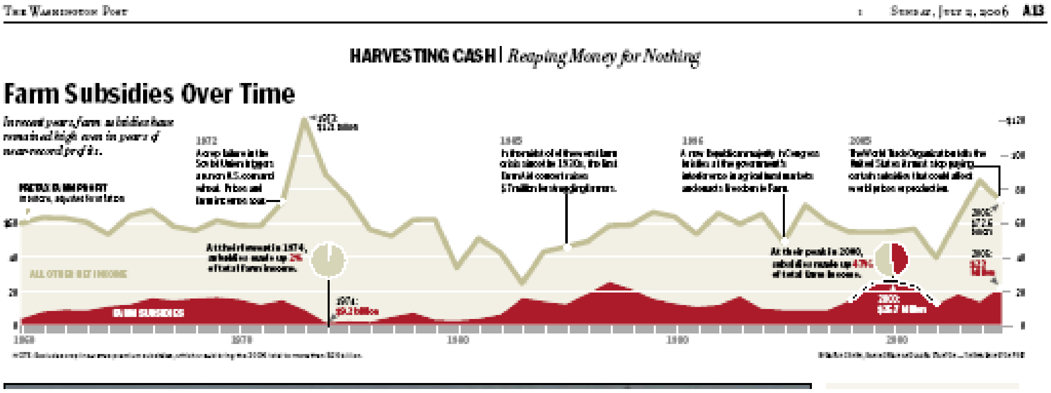
- Help you identify themes and questions for the rest of your reporting
- Identify outliers – good stories, or perhaps errors, in your data
- Help you find typical examples
- Show you holes in your reporting

Visualizations also play multiple roles in publishing:

- Illustrate a point made in a story in a more compelling way
- Remove unnecessarily technical information from prose
- Particularly when they are interactive and allow exploration, provide transparency about your reporting process to your readers

These roles suggest you should start early and often with visualizations in your reporting, whether or not you start electronic data or records. Don't consider it a separate step – something to be considered after the story is largely written. Let this work help guide your reporting.

Getting started sometimes means just putting in a visual form the notes you've already taken. Consider this graphic, which ran in the Washington Post in 2006:

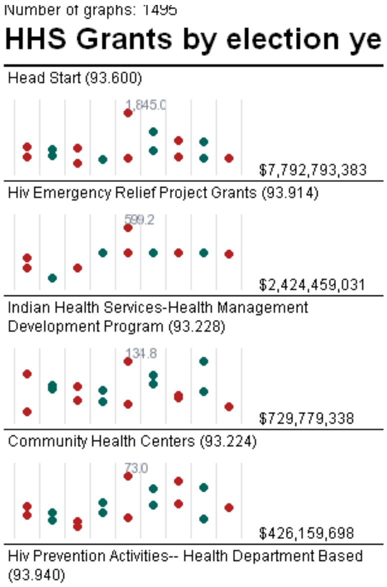


It shows the portion of farm income associated with subsidies and key events over the past 45 years, and was built over a series of months. Finding data that could be used over time with similar definitions and similar meanings was a challenge. Investigating all of the peaks and troughs helped us keep context in mind as we did the rest of our reporting. It also meant that one chore was pretty much finished before the stories were written.

Tips for early exploration

Use small multiples to quickly orient yourself in a large dataset and find examples for further reporting.

I used this technique at the Washington Post when we were looking into a tip that the George W. Bush administration was awarding grants on political, not substantive, grounds. Most of these aid programs are done by formula, and others have been funded for years, so we were curious whether we might see the pattern by looking at nearly 1,500 different discretionary streams.



I created a graph for each program, with the red dots indicating a presidential election year and the green dots indicating a congressional year. The problem: Yes, there was a spike in the six months before the presidential election in several of these programs – the red dots with the peak numbers next to them – but it's the wrong election year. Instead of George W. Bush's re-election bid, the peak as consistently for the 2000 presidential election, when Bill Clinton was in the White House and his vice president, Al Gore, was running for the office.

This was really easy to see in a series of graphs rather than a table of numbers, and an interactive form let us check various types of grants, regions and agencies. Maps in small multiples can be a way to show time and place on a static image that's easy to compare – sometimes even easier than an interactive.

This example was created with a short program written in PHP, but it's now much easier to do with Excel 2007 and 2010's sparklines. Edward Tufte, the visualization expert, invented these "intense, simple, word-like graphics" to convey information in a glance across a large dataset. You now see them everywhere, from the little graphs under stock market quotations to win-loss records in sports.

Look at your data upside down and sideways:

When you're trying to understand a story or a dataset, there's no wrong way to look at it – try it every way you can think of, and you'll get a different perspective. If you're reporting on crime, you might look at one set of charts with change in violent crimes in a year; another might be the percent change; the other might be a comparison to other cities; and another might be a change over time. Use raw numbers, percentages and indexes.

Look at them on different scales. Try following the rule that the x-axis must be zero. Then break that rule and see if you learn more. Try out logarithms and square roots for data with odd distributions.

Keep in mind the research done on visual perception. William Cleveland's experiments showed that the eye sees change in an image when the average slope is about 45 degrees. This suggests you ignore the admonitions to always start at zero and instead work toward the most insightful graphic. Other research in epidemiology has suggested you find a target level as a boundary for your chart. Each of these ways helps you see the data in different ways. When they've stopped telling you anything new, you know you're done.

Don't assume

Now that you've looked at your data a variety of ways, you've probably found records that don't seem right – you may not understand what they meant in the first place, or there are some outliers that seem like they are typos, or there are trends that seem backwards.

If you want to publish anything based on your early exploration or in a published visualization, you have to resolve these questions and you can't make assumptions. They're either interesting stories or mistakes; interesting challenges to common wisdom or misunderstanding.

It's not unusual for local governments to provide spreadsheets filled with errors, and it's also easy to misunderstand government jargon in a dataset.

First, walk back your own work. Have you read the documentation, its caveats and does the problem exist in the original version of the data? If everything on your end seems right, then it's time to pick up the phone. You 're going to have to get it resolved if you plan to use it, so you might as well get started now.

That said, not every mistake is important. In campaign finance records, it's common to have several hundred postal codes that don't exist in a database of 100,000 records. As long as they're not all in the same city or within a candidate, the occasional bad data record just doesn't matter.

The question to ask yourself is: if I were to use this, would readers have a fundamentally accurate view of what the data says?

Avoid obsessing over precision

The flip side of not asking enough questions is obsessing over precision before it matters. Your exploratory graphics should be generally correct, but don't worry if you have various levels of rounding, if they don't add up to exactly 100 percent or if you are missing one or two years' data out of 20. This is part of the exploration process. You'll still see the big trends and know what you have to collect before it's time for publication.

In fact, you might consider taking away labeling and scale markers, much like the charts above, to even better get an overall sense of the data.

Create chronologies of cases and events

At the start of any complex story, begin building chronologies of key events and cases. You can use Excel, a Word document or a special tool like TimeFlow for the task, but at some point you will find a dataset you can layer behind it. Reading through it periodically will show you what holes are in your reporting that have to be filled out.

Meet with your graphics department early and often

Brainstorm about possible graphics with the artists and designers in your newsroom. They will have good ways to look at your data, suggestions of how it might work interactively, and know how to connect data and stories. It will make your reporting much easier if you know what you have to collect early on, or if you can alert your team that a graphic isn't possible when you can't collect it.

Exploration to publication

You might have spent only a few days or few hours on your exploration, or your story might have taken months to report. But as it becomes time to move to publication, two aspects become more important.

Remember that missing year you had in your early exploration? All of a sudden, you can't go any further without it. All of that bad data you ignored in your reporting? It's going to come back to haunt you.

The reason is that you can't write around bad data. For a graphic, you either have everything you need or you don't, and there's no middle ground.

Match the effort of the data collection with the interactive graphic

There's no hiding in an interactive graphic. If you are really going to have your readers explore the data any way they want, then every data element has to be what it claims to be. Users can find any error at any time, and it could haunt you for months or years.

If you're building your own database, it means you should expect to proof read, fact check and copy edit the entire database. If you're using government records, you should decide how much spot-checking you'll do, and what you plan to do when you find the inevitable error.

Design for two types of readers

The graphic – whether it's a standalone interactive feature or a static visualization that goes with your story – should satisfy two different kinds of readers. It should be easy to understand at a glance, but complex enough to offer something interesting to people who want to go further. If you make it interactive, make sure your readers get something more than a single number or name.

Convey one idea – then simplify

Make sure there is one single thing you want people to see? Decide on the overwhelming impression you want a reader to get, and make everything else disappear. In many cases, this means removing information even when the Internet allows you to provide everything. Unless your main purpose is in transparency of reporting, most of the details you collected in your timeline and chronology just isn't very important. In a static graphic, it will be intimidating. In an interactive graphic, it will be boring.