Visualizing Data: An Interview with Anselm Spoerri

VISUALIZING DATA STARTS WITH PUTTING THE DATA INTO A FORM YOU CAN VISUALIZE, THEN UNDERSTANDING WHO YOUR AUDIENCE IS AND WHAT THEY WANT TO KNOW.

BY JOCELYN MCNAMARA, MLIS

If a picture is worth a thousand words, how many data is an infographic worth?

Typically not many, says Anselm Spoerri. An infographic—often (wrongly) used as a catch-all term for everything from simple bar charts to complex interactive visualizations—is best used to convey a few data points that represent key patterns or trends. Visualizations, on the other hand, provide context for the data and often allow the audience to manipulate the data to reveal more complex relationships among the factors that affect the data.

Spoerri, a faculty member at the School of Communication and Information at Rutgers University, has conducted research in the field of information visualization for the last 20 years and teaches students how best to visualize data. Earlier this year, he was honored with a Professional and Scholarly Excellence (PROSE) Award for DataVis Material Properties, a tool he helped design for McGraw-Hill Education.

Information Outlook interviewed Spoerri about the techniques and tools for visualizing data and information and the biggest challenges facing information professionals who work with data.

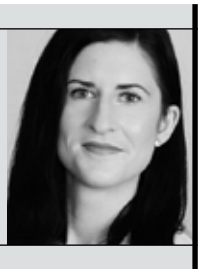
The terms *graphics*, *infographics*, and *visualization* are used frequently and often interchangeably in regard to data presentation. Can you define these words and explain their differences and similarities?

The way I like to think about it is, infographics usually are static. They give you a high-level view of the key salient patterns in the data that have been identified by an analyst, who then—either by using some tools or hiring a graphic designer—comes up with a visual representation of those key results. So, from the perspective of data visualization as a whole, infographics are what you use when you are communicating with a very general audience or you just want to communicate highlights in a quick way.

But the moment you want to understand more about the data, then you come into the realm of what's considered interactive data visualization, meaning you can actually drill into the data and filter it and look at the subsets and understand the data yourself. The questions to ask about interactive data visualization are, one, who is doing it, and two, for whom is it being done?

An infographic is designed for end users who don't need to have a lot of content understanding or domain understanding, so they can very quickly grasp key things that are going on with the data. But you can't communicate

JOCELYN MCNAMARA is deputy director at LAC Federal in Rockville, Maryland, and a member of the Information Outlook Advisory Council. She can be reached at mcnamarajocelyn@gmail.com.



more complex or subtle relationships with data using infographics.

If you're an analyst and you have a data set, you'd like to have interactive data visualization tools so you can better understand the data you're trying to analyze. What's starting to happen—and you're seeing this in newspapers like *The New York Times* and *The Washington Post*—is that readers are being given access to interactive data visualization tools, so the readers themselves can explore the data and not just have to take the word of journalists that these are the key insights to be gained from the data.

It's critical that the right kinds of displays are used to make the underlying patterns of the data visible. Often, you need to first figure out what those patterns are, so you need to go through an interactive or iterative loop to figure out what's going on in the data and what's the best way to see it and understand it and then present it. The distinction, in a way, is between the *presentation* of data and *understanding* and *analyzing* data. Sometimes the same tools are being used for both purposes.

If you want to go one level deeper, you get into what's called data analytics. This is where the data sets are so huge that, yes, you can visualize them, but to be able to really understand them you need machine learning or artificial intelligence methods to find the interesting subsets. Then you can use visualization to further examine these data sets or fine-tune the algorithms that are crunching through them.

So the idea is that you don't need to be a subject matter expert to be able to drill down and understand what's going on, correct?

It depends. For example, you can view an infographic as an access ramp into data to get the key ideas. The reader gets an understanding, then moves on. He doesn't linger with the data, maybe doesn't even have a desire to, because often data is messy.

Think of it almost like being in a big city and wanting to show someone the major sights. You position your camera and take a snapshot, and you say, here is an interesting sight. Infographics is like that—you take a snapshot of an interesting pattern, and then you move on and explore the data and find another interesting connection or relationship, and you take another snapshot. And those snapshots are what you put into your presentation.

Anselm Spoerri

Using infographics is really a question of who your audience is and how much they care to know. Infographics are great for giving you the highlights, and depending on the sophistication of the designer, sometimes it can be done in a very memorable way so that even a lay person can understand quickly that the data are about cars or animals or the climate. There are pictorial representations that immediately tell people the context the data are in.

The challenge with what I would call "cold" or "sober" data visualization is that you have a series of displays that you look at and you don't even know what they're about. You have to go look at the axes and read the labels, and it's all very abstract.

The other question in my mind is one of motivation—how much do I care about the data, and how much value is in the data? For example, if you look at *The New York Times*, they did quite a lot of visualizations around the 2016 elections. Why? Because they believe their readers care about that kind of data—they want to know more about it, and they may even want to drill into it. So it really comes down to the value of the data set and what people are hoping to do with the information they gain from the data set.

What's also happening with infographics is what I call "animated presentations." These are like mini-videos that tell you, in an animated way, a story. They're highly scripted, like a movie. And somebody has to figure out the highlights to present.

So the interpretation has already been done for the viewer?

Yes. And the questions there are, how skillful is the analyst who interpreted the data, and how skillful is the graphic designer who created the presentation?

I can readily imagine that data could be interpreted and communicated one way,

but someone might view it and draw a different and unintended conclusion.

If it's a rich data set, there are multiple ways you can slice it and filter it. For example, go back to the 2016 election. People are still trying to understand what happened, how we got the result we got, and who voted for whom. The question is, what types of data variables do you need to take into account? Income? Religion? Education? What variables do we have, and how do we bring them into the data space so that a viewer can start to see relationships and correlations?

An Infographic simplifies, either because a person has decided that what's being presented is good enough, or there's a very established way to think about it. The key is to make the information as accessible and consumable and understandable as possible. For me, if an infographic is done well, it's like a well-designed documentary or movie. It's also there for quick consumption. That's what it's designed for.

If you look at what's happening right now, infographics are becoming very popular. The news media are using infographics to communicate with their readers; infographics are also being used in education to make certain facts available, to make numbers not appear as cold or abstract—to make them more relatable. Infographics is more about how data can be *consumed*. When we talk about data visualization, we are moving into a higher-dimensional realm of interaction, inquiry, examination, asking questions, and developing hypotheses.

Going back to your election example, people often talk about data as something that's very objective and that we can refer to in a highly scientific way. Oftentimes, we forget that the variables we come up with at the front end of the inquiry are being determined by people. I think the human bias is overlooked a lot when we decide how to construct and interpret data sets.

Sometimes you have to work with what you have—this is what we measured, or this is what was easy to measure, so now we'll work with it even though it's not really the best way to think about this phenomenon. That's one issue. Another issue is, how important is this data? Is it worth it for me to develop new ways of measuring things so I can start collecting this data?

And there are always errors in data, because data is noisy and messy. So when I teach students about data visualization, I tell them the hardest part is getting the data in a form that you can visualize. That's a lot of work. And then you have to ask, how valid is your data, and what is the provenance of your data?

Another thing that's happening—and I think this is where big data is coming in—is that people are starting to collect so much data that, at the end of the day, the noise in it doesn't matter as much. If you were to look at one data variable on its own, you might say there's too much noise in it, there's too much measurement error, or there's too much bias. But if you add another data variable, and then another and another, soon you have millions and millions of data points. And with the statistical methods and machine learning techniques available today, you can squeeze out some decent patterns that, if you had much less data, wouldn't be as good.

Yes, the data is noisy, and yes, it can be biased. But there comes a point where, at certain volumes of it, and combined with other data sets—this is another thing that's happening, people are combining multiple databases—when you start comparing them and relating them, you can sort of get the noise out, so to speak.

Looking at big data sets is something we haven't previously done before. And that raises all sorts of black swan scenarios, where we're trying to look for patterns that are emerging, but we don't even know what to look for because we haven't analyzed data at this level before.

You mentioned the black swan phenomenon, which is something that has a very low probability of happening, but when it happens, it has a catastrophic impact. True, there is a possibility that we don't know yet how to use all of this data and make sense of it. But we have this computer infrastructure now that has gotten so cheap that we can really go for it. Before, it was way too expensive to work with these large data sets, so you had to be more "clever" to extract value.

Now, let's bring it back to infographics. I'm in this sea of data and I'm using visualization to navigate the data and get some sense of it. In a way, data visualization is a communication tool. The question is, with whom am I communicating? Am I communicating with myself, meaning I'm looking at the data and I want to understand it so I can make a decision? If so, I'm using visualization as an analyst of the data.

Now I need to tell others about the data. If I have a team of specialists, we work together to build up know-how and maybe we even customize some visualization tools so we can better understand the data. Then we may need to communicate the data to others who are not experts, who are not as involved with the nitty-gritty of the data. And the question there is, what's the attention span of the audience?

If it's higher management and they just want the highlights, I might say we should create an infographic for them—here are the key facts, and here are some charts that are easy to read and don't require special knowledge to understand. Or maybe we need to have a feedback loop and an interactive tool, so that when we're talking about the data we can be in the moment and change the filter settings and the view of the data so we can start exploring the data and communicate to others how rich and complex it is.

It's all a question of who's doing it, why they're doing it, what they're hoping to get out of it, and how much time they're willing to invest in it.

So, what are the main techniques for visually displaying data, and what visualization tools do you think would be

most useful for librarians and information professionals to learn?

If you think in terms of visualization, there are standard display types—bar charts, line charts, pie charts, bubble charts, scatter plots, and maps. These are the workhorses of data visualization, and you'll also see them in infographics. People are familiar with them; they've been in use for some time. And if you think in terms of how designers visually encode information, they do it in a way that makes it quite easy for human visual systems to read it, because they produce visual patterns that the human visual system is good at detecting.

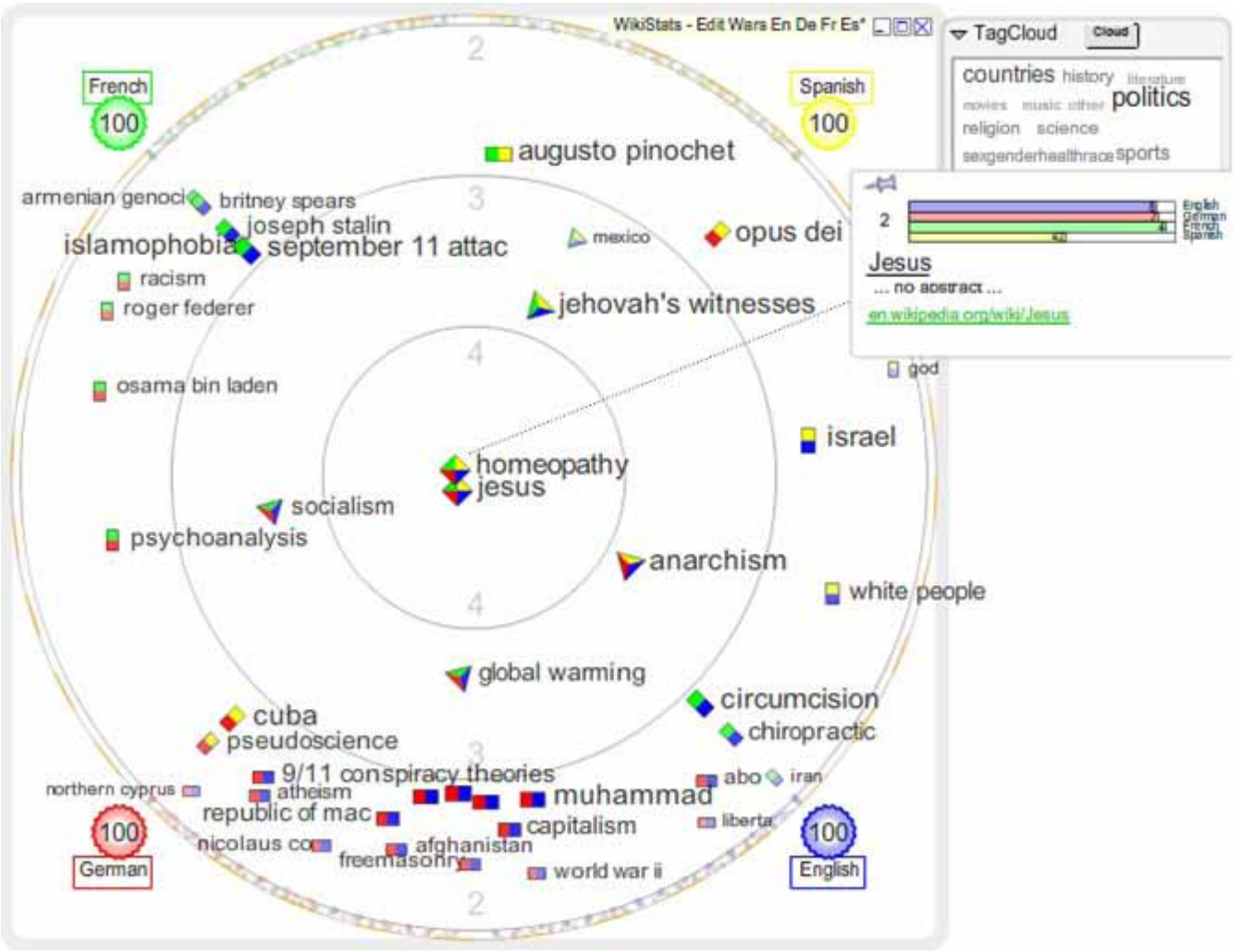
Now the question is, when do you use a bar chart, when do you use a line chart, and so on. Most often, librarians will be using these types of tools. Some of them you can generate in Excel to create static visualizations, or you can use a tool like Tableau that has a free public version. Or maybe you get a license because you have data that's coming from different databases that needs to be merged, and then you create visualizations and presentations for your management or for the public.

Google has free tools you can use; there's a whole slew of tools available out there. The question then becomes, do you also want to do some statistical analysis? There are packages that are optimized for that. So it really becomes a question of how rich your data is—the larger the data set, the more sophisticated the tool you need to understand the data.

For me, the main driver is this: At the core is the data. Does the data matter? Depending on how rich the data is and how much it matters, are there meaningful patterns in the data? What are those patterns? You pick the right display to make those patterns visible. Then, if you want to design for a short attention span or for people who just want the key facts, you move to an infographic; if you want to open it up so there can be more exploration, you use an interactive data visualization tool.

With an interactive tool, the question is, do you give people access to all of the data, or do you create a curated data set that contains the key facts

FIGURE 1: **The 100 Most Controversial Pages on Wikipedia** (by language)



and patterns, but you don't put all of it in there because it's not necessary or because it's proprietary or there's something you don't want everyone to see? And in order to get to that, you need to have some data specialists who, depending on the size of the data, have to dig around and analyze it until they get it to a place where it's presentable.

There's a notion out there that if I have a data visualization tool and I connect it to the data set, I will magically understand the data. Not so. You need to have certain skills and understandings so that you know how best to make visible what's in the data. And sometimes there's nothing in the data—there's nothing strange or earth-shaking, or you already knew it or had a sense about it, and there are no major insights lurking in there.

Is there a risk that data visualization can be overused—that it will lose its impact?

This brings me back to infographics. If you ask me my opinion about infographics, I'm sort of skeptical. I understand the value of a well-designed infographic—if it's done well, it can be extremely powerful. But the key is that there has to be something in it that's of meaning, of value.

I think sometimes visualization, because we're moving into a more "visual" world, is used as an attention-getter. If you look at some infographics, they use very saturated colors and large fonts in a way to draw the eye. It's like putting a little sugar out there just to get somebody to pay attention. So it loses its power, but maybe for a certain consumption pattern, it's perfect. I get a little taste, a little flavor, a little visual stimulus, and I can quickly make sense of it. No harm is done, but no deep insights are generated.

Can visualization be overused? In certain contexts, it is overused—the visual event that's being created is not warranted. But you also have that danger with videos, when people are creating animations. Where there is data that has value in it, visualization will play its role, either to help communicate the value that's there or to help people find the value that's there.

Where do you see data visualization going, and what may be some of its long-term applications?

I said before that visualization is a communication tool. It's either a communication tool for me to understand the data as a data analyst, or a tool to communicate what I found in the data to others. The question here is the sophistication. Will this become part of a general tool set that's available for free—meaning I will buy a computer or a smartphone and it will have this capability built in—or will I have to buy some sort of specialized software?

The trend for the last few years has been that more and more has to go through the browser. Many people, even heavy-duty users, would like a way to use their browser so they can access the tool and view the data. That puts pressure on what can be done in the browser, whereas if you build specialized applications that are optimized for, say, rendering data points very quickly, that would be easier, but it would require either a download or a special license, and that will make it harder for people to become part of the conversation. So the trend is toward making it more ubiquitous in terms of what you can visualize and how you can visualize it, and there the browser plays a key role as a gateway to view the data.

I think the other trend is that visualization is used in different contexts to serve different purposes. For example, there are certain visualizations that I would call entertainment. They're designed to entertain, to move, to elicit an emotion. Yes, there's an overall theme, but it's not about being able to reliably infer certain things. It's more to get a feeling about the data.

Then there are tools that are more analytical, so that I can ask questions. If I do certain interactions, I get the same display, whereas if there's a more artful visualization, it changes each time how it chooses to display the data.

Another trend is the diversity of display types that are becoming available, and that really depends on the specific need and also on the value proposition in the data. It's the same as in other contexts—if there's something that has special value, we will build specialized tools to extract that value.

Do you have a favorite resource to recommend for those of us who want to learn more?

There's Alberto Cairo, who has written some great books that are not just for specialists. He comes from a journalism background, but he's also a data visualization researcher.

To me, what's interesting is what's happening with the newspapers. If you look at *The New York Times* and *The Washington Post* and *The Guardian*, they're using visualizations in their storytelling. And I don't mean infographics—I mean interactive data visualization for a lay audience to better get a sense of the data and understand and explore it.

I think the other thing to think about is how it's being consumed. Some of the visualizations require a large screen; otherwise, it is difficult to make sense of them. If it's being consumed on a smartphone, that immediately limits what you can do and how much richness you can communicate. So I would be guided by the attention span of your audience. That guides what you can do and what you should be doing in terms of what you present. **SLA**

RESOURCES

Scatterplot matrix (http://mbostock.github.io/d3/talk/20111116/iris-splom.html)

Motion Chart: https://bost.ocks.org/mike/nations/

Treemap of folder hierarchy of one million computer files: http://www.cs.umd.edu/hcil/millionvis/million-treemap.gif at http://www.cs.umd.edu/hcil/millionvis/