acmqueue A Conversation with Jeff Heer, Martin Wattenberg, and Fernanda Viégas

Sharing visualization with the world

Visualization can be a pretty mundane activity: collect some data, fire up a tool, and then present it in a graph, ideally with some pretty colors. But all that is changing. The explosion of publicly available data sets on the Web, coupled with a new generation of collaborative visualization tools, is making it easier than ever to create compelling visualizations and share them with the world.

To explore some of the key developments in this fruitful new age of visualization, we've assembled an all-star cast of researchers who are responsible for some of today's most popular and engaging new visualization tools: Jeff Heer, Martin Wattenberg, and Fernanda Viégas. Heer, who graciously agreed to be our interviewer, is an assistant professor at Stanford University who focuses on human-computer interaction, visualization, and social computing. He led the design of the popular open source visualization toolkits Prefuse (http://prefuse.org/), Flare (http://flare.prefuse.org/), and Protovis (http://vis.stanford.edu/protovis/). He received his B.S., M.S., and Ph.D. degrees in computer science from the University of California, Berkeley, and in 2009 he was named to *MIT Technology Review's* TR35, a list recognizing 35 innovators under the age of 35.

On the other side of the table are two research scientists from IBM's VCL (Visual Communication Lab), Martin Wattenberg and Fernanda Viégas. Wattenberg, VCL's founder, previously worked at SmartMoney.com as director of research and development, where he led development on the innovative and extremely popular Map of the Market visualization (http://www.smartmoney.com/map-of-the-market/). He holds a Ph.D. in mathematics from UC Berkeley.

Viégas is an alumna of MIT's Media Lab, where her interest in visualization and visualization-driven storytelling took root. Her work on visualizing people's chat histories and e-mail archives showed that visualizations can express an emotional, human element beyond just presenting raw data.

Wattenberg and Viégas have collaborated on many groundbreaking visualization projects together, the latest of which is Many Eyes (www.many-eyes.com), a Web site that encourages users to upload their own data sets, create visualizations from them, and share and discuss these visualizations with others. Both are interested in the artistic component of visualization and have had work shown in museums including the Museum of Modern Art, the Boston Institute of Contemporary Art, and the Whitney Museum of American Art.

JEFF HEER What sparked your interest in visualization?
MARTIN WATTENBERG I first got interested in visualization when I was working at *Smart Money* magazine. This was in the mid-'90s, and putting news on the Web was new. Part of the mission of the magazine was to make complicated financial data understandable. We tried a whole bunch of ways of doing this, but the one that really caught fire was visualization. We created a visualization called the Map of the Market, which presented a maplike view of the U.S. stock market using what was then a fairly new technology called a tree map. It brought market data into focus in a way that I had never seen before.

I began relying on it all the time, and I felt like it was almost giving me an advantage. I could talk to reporters who were talking to traders, and we seemed to understand a lot of what was going on. A bunch of reporters got addicted to it at the magazine, and then we put it on the Web site and it became a hit with readers.

This really made me enthusiastic about visualization, and I began to believe it could increase your awareness in some amazing ways.

FERNANDA VIÉGAS I come to visualization from a very different perspective. I had never even heard the term *visualization* before getting to MIT because I had been doing traditional graphic design. When I got to the Media Lab I became interested in interfaces for communication. That's when I started wondering if there were graphical alternatives for things like chat rooms, which back then tended to be just avatar based. I was thinking you could use graphics not just to reinvent reality as it looks, but also to convey patterns of communication.

Tons of data get produced when people communicate online, so I began asking, "Is there a way to make these archives talk?" I realized there are visual ways to make them talk.

The tricky thing in being a graphic designer is that your control over what's going to happen on the screen is given off to the data, which is interesting because you don't really know what things are going to end up looking like.

JH Can you walk me through the design and development of a favorite visualization? I'm curious how you get to the point of deciding which diagram to do. But first, how do you choose which data sets to attack, and how do you then engage in prototyping and design to converge on your visualizations?

MW It's a pretty long process, and I think you're right to focus on the data as the first part. When Fernanda and I did our history-flow project on Wikipedia (http://www.research.ibm.com/visual/projects/history_flow/), it was preceded by a couple of months of lunch and coffee meetings where we tried to figure out what we really wanted to look at online and what data was available. What was dramatic about Wikipedia at that time is that it had a huge amount of data available, and it had not been studied to death.

Once we knew our general area of inquiry, we started just poking through by hand. We would go to Wikipedia and look at articles and click on the history links, and through that we got a general sense of the size of the data. This definitely confirmed that we would need some tools to sift through all of the data.

One important decision we made early on was to look at the content of articles and not just the metadata. In subsequent projects we have looked at the metadata around which people did editing and things like that, but I think one of the things that led to the success of this project was focusing on the text. Once we decided to look at the text and the history, we needed to be able to connect one version to the next. That meant we had to find a good algorithm for doing a diff between two versions of text. That was actually subtler than I would have assumed, but we figured it out.

Once we had those diffs, then the question was how to represent that visually. We spent some time drawing on whiteboards in probably the loosest, sketchiest way you can imagine. When we finally came up with the idea of connecting pieces from one version to matching pieces from another version, we didn't really waste a lot of time refining that on paper—we immediately began coding.

That's when the project became exciting, because we had no idea what we would see. When

2

we first looked at the patterns that emerged, it was fun because they were all sort of weird. It was like landing on an alien planet and seeing something moving and wondering, "Is that something blowing in the wind or an animal that's about to eat me?"

Once we had that initial look, we got a sense of how to refine it. We added various coloring schemes and refinements for scaling. We also added a lot of interactivity. But, it's funny, all of that design happened after we had written some code. That's one of the ways we tend to work: at first we don't have real data, but once we have the data, we very quickly jump to the software stage.

JH One theme I hear is the importance of working with real data—and having that infuse your process.

FV That is definitely the way we tend to do things; otherwise, you find yourself doing things with imaginary data that don't pan out. Let me give you a concrete example. When we were doing the Wikipedia history flow one of the hypotheses we had in mind was this: What if paragraphs change places all the time and you end up with all these crossings on the diagram because things are moving all over the place? Then we would have to make our visualizations perform better to show those kinds of things. As soon as we visualized the first few pages, however, it was striking how rarely that happened. It was one of the hypotheses we had been working with that turned out to be completely false.

MW You can almost make the analogy that having real data as part of your project is as important as having real users look at your project. I have seen people try to design visualizations where they create mockups with fake data, show them to real users, get all sorts of feedback, and in the end it's not worth anything. They eliminate some of the user risk but they don't eliminate the data uncertainty. It's almost as if the data is one of the stakeholders in the project and you need its input from the beginning.

FV You also need to be a good judge of how close to the data you want to fly, because one of the things you could do is be so worried about the real data that you lose track of your bigger question with the visualization.

For example, one of the things we wanted to visualize was what a person looks like on Wikipedia. What does an individual's editing history look like? What are the different kinds of roles that people play?

That was pretty hard. We had the editors' data, but we couldn't come up with a useful way of showing it. Martin and I, along with our extremely helpful collaborator, Kate Hollenbach, spent the whole summer trying to figure out a good way to visualize editors, but we kept getting these not-very-useful results. At one point we tried just to get a sense of the shape of the data using bar charts, line graphs, and stack graphs, but that wouldn't tell us anything, either.

Eventually we decided to try out a very weird technique, which was mapping streams of text to colors. This makes you lose a lot of information because text is really rich and you can use only so many colors, but we wanted to see what we would get. We decided to look at the three first letters of a string and map that to a given color.

All of a sudden we saw patterns. Someone was going around all of Wikipedia correcting typos; another person was working on images; another was working on stub sorting, which is the way you do categorization on Wikipedia. There was a plethora of different kinds of projects that people were engaging in that snapped into focus because of this visualization.

Looking back, we feel that the very first experiments we did with the data were on too high of a

level. They were abstracting too much away from the data and not giving you this sort of messiness that Wikipedia has, which is everybody's there, everyday, making minute changes. It's these minute changes that add up to patterns.

This notion of how close to the data you want to be and what is your question—*what is the story you want to tell?*—seems to be really important.

JH As you engage in this process, how do you gauge the quality of your visualizations?

MW There are a few milestones. The first sign that a visualization is good is that it shows you a problem in your data. Every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.

The next sign, which is sort of a weird, subjective thing, is that you will be programming a visualization and making progress and going really fast, and then, all of a sudden, your progress will slow down because you've started playing with the visualization instead of debugging it. That's also a really good sign. The moment something just becomes fun to play with, almost like a video game, you hit some level of engagement and it becomes really interesting.

The third sign is that you'll just start talking to people. Fernanda and I have worked on many visualizations together, and each of the successful ones has led us to these long conversations where we sit in front of the computer screen saying, "OK, let's look at another one, let's look at another one." That's incredibly enjoyable. This need just to start talking about it and showing people is important.

The fourth sign of success is more traditional. This occurs after you have hit these other three milestones and you know that you're onto something, but you really have to show it to other people for feedback. You immediately discover, for example, that there's all sorts of labeling stuff that you haven't done. Then it becomes more of a traditional software process, where you watch your users deal with your visualization and see whether they get insight or not.

JH One of your more recent projects is Many Eyes (http://many-eyes.com), which is a Web site where users can publicly share data sets and visualizations. What series of events led you to build such a site?

FV Martin and I both had personal experiences with visualization building that made us rethink what visualization is about and what kinds of insights it empowers.

My personal story has to do with visualizing e-mail archives, for which I needed real data and real users. I had to be very explicit with my potential users about the privacy implications and the fact that I would never publish anything with their data without their consent—and even then, I would make the data anonymous. As soon as I got that out of the way, people felt more comfortable giving me years' worth of e-mail archives.

When I put them in front of the computer to play with the visualizations, the first thing they wanted to do when they saw something interesting was to share those images with others. They would take screen shots and e-mail them to friends and family. I saw them bringing people to sit with them in front of the computer and showing them where numerous love affairs had started and ended. And here I am thinking, "But we just talked about privacy!"

All of a sudden it dawned on me that one of the most powerful aspects of the visualizations I was creating was communication. The images functioned much as photographs, where they serve as social artifacts around which people have conversations, reminisce about the past, and exchange stories.

4

If we stop thinking about visualization as solely an exploratory, insight-driven technology, then what else can we do? Can we build these tools to be more social?

MW My own experience in parallel came when my wife wrote a book about baby names, and she and I worked together to create a visualization of name popularity over time called NameVoyager (http://www.babynamewizard.com/voyager).

We put this on the Web right as her book was published, and the reaction was really interesting. We had expected that parents who were expecting a child would use this and might talk about potential names. But we discovered, by looking at blogs and discussion forums online, that all sorts of people who were not really interested in babies, but who were interested in names or who liked using the visualization, were having these long discussions about what it meant.

That was really interesting to me, because watching people analyze what was a pretty complicated data set in this very gamelike, fun way completely changed my expectations of visualization. Before then, I thought about it as largely a serious tool, the kind you might use to do scientific research or to decide about your investments; but after watching the level of engagement people were having—analyzing thousands of time series, not even thinking about the fact they were doing statistics—made me realize there is something about the social use of visualizations that is really critical.

When we started working together at IBM, we had a series of conversations where we asked, "What is next for visualization? What should we work on?" We both settled on the social use of visualization as the idea that we wanted to test and use to guide our research. That led to the Many Eyes system, and it's up there today where anyone can use it.

JH How did you decide which elements to include in the site? Were there any stumbling blocks along the way?

FV We knew a couple of things right off the bat. We knew that anything we put out had to include the kinds of graphics that Excel does—bar charts, pie charts, and so on. We also knew that because we wanted everything to be interactive, providing the ability to point at things was going to be a challenge we had to address. That's how we came up with the whole bookmarking notion. As you're interacting with a visualization on Many Eyes, bookmarking provides a way of easily capturing that state and attaching a comment to it. If I see some peak that I'm interested in, I might say, "OK, what is this about?" Then other people seeing the same visualization have an easy way of clicking on that bookmark and looking at exactly the peak I'm referring to.

We knew these things right away. Something that was not as clear to us was what kinds of less traditional visualization techniques we should launch with. If we launched with a tree map, would people really use it? Would they be able to make sense of it? Even back then there was interest in social networking, so we thought maybe we should have a visualization technique for that.

MW This is a very small element, but it illustrates the kind of thinking we had to do. One of the chart types on Many Eyes is called the bubble chart. You have a list of numbers—for example, countries and their populations—and the chart is just a collection of circles whose areas correspond to those numbers. It's a very simple chart, and you could probably make various complaints about it from a statistical point of view.

It turned out to be really popular, partly for the reasons that we added it: (a) it's a little bit off the beaten path—it's not something that you find in Excel; (b) you can apply it to pretty much any list of numbers that people are likely to have. Part of what we wanted to do was have types

of visualizations that people could apply to data very simply, that didn't have a whole lot of complicated requirements to sort through.

JH One of the goals of Many Eyes is to enable what you call social data analysis: groups of people coming together to make sense of interesting patterns in data. Along those lines, what do you consider to be the most telling success stories of Many Eyes users to date?

FV One success story happened early on and was quite unexpected because we didn't even know the community existed. A Many Eyes user created a visualization of the New Testament, based on name co-occurrence. Whenever a verse in the Bible had two or more names, that became a data point, and then this person visualized this set of co-occurrences.

The user then blogged about this visualization, and it was rapidly picked up by other Christian blogs. There were tons of comments back on this person's blog, and then it spread over the next couple of days to places outside of the Christian blogosphere. Sites such as Boing Boing were picking it up; and someone even created a YouTube video of himself playing with this visualization and showing what kinds of new insights he was getting. This led to other people uploading Bible data to Many Eyes, creating their own visualizations, and blogging about them.

This was a success story for a couple of reasons. It exposed an entire community to these visualization tools. This community already had a bunch of data sets about Bible statistics, but all of a sudden they were empowered to look at those from a very different perspective. They were actually experimenting with a bunch of different visualization techniques. They used a tree map, a social network, and were one of our first sets of users to figure out how to embed visualizations live on their sites.

Not only that, there was a real conversation going on, as they were trying to analyze what these visualizations were showing them. Different people would come up with different kinds of data and different visualizations to add to the conversation, so it really felt like these visualizations were empowering a debate that was already going on.

We also saw this viral effect, where the visualization started in a very specific community that was interested in statistics about the Bible and then appeared on a bunch of different kinds of blogs.

That was exactly the kind of thing we had dreamed of when we first created Many Eyes, but we had no way of knowing if it would happen or not. It was a textbook example of what can be done when you just let visualization be free.

MW Another example that sticks out in my mind involves the Sunlight Foundation, which is a watchdog agency that tracks government action on many different fronts. One of the topics it is interested in is earmarks in bills, where money gets spent on very specific entities, so it created a visualization—a bubble chart—of earmarks.

That chart went out on a couple of well-known blogs, and then I noticed that Lawrence Lessig (known for his work in law and technology) included it in an online presentation about government corruption. I assume he got it from the Sunlight Foundation rather than the Many Eyes site, but to me that showed the power of putting these things out in public so that they can spread and have an impact in a variety of ways.

It also showed something else, which is that analysis is part of the story but communication and storytelling are at least as important.

JH What challenges have you faced in trying to take these types of explorations and scale them, either in terms of the data or in terms of the number of people involved?

MW I think you want as many people as possible to contribute to a visualization and make comments on it, but I still don't know of a magic bullet to cause that kind of engagement to happen. We did a study that confirmed some of what we've talked about anecdotally so far, which is that people like to talk about these things in their own communities. A lot of the really interesting conversations occur when people can embed something in their own context—in their own blogs or wherever—so I think the portability of these tools is very important.

FV One interesting phenomenon that took off last year was Wordle (http://www.wordle.net/), which creates "word clouds" from text provided by the user. One of the things that creator Jonathan [Feinberg] did there was give folks a way to customize their visualization. That's something that visualization tools, including Many Eyes, don't do very well.

We did a study of Wordle, and one of the surprising results was that almost 90 percent of its users said they felt highly creative when making a Wordle, yet all they were doing was clicking a couple of buttons. The notion that these kinds of visual tools tap into people's creativity is very underexplored. It's something we don't usually think about when we create visualization tools, and it's something that the community should be looking into.

JH That's a really exciting point. Typically people post visualizations and then there will be textual commentary and discussion, but I can't help but think about some of the more interesting and more meaningful forms of discussion, such as the video response on YouTube. Do you think there are opportunities for dialogue among visualizations themselves as one way of trying to facilitate this increased creative output on the part of users?

MW There's a lot of opportunity for that kind of dialogue. On Many Eyes, we've seen examples where someone will put up a visualization that is left leaning, and then a right-leaning visualization will appear on the site seemingly in response.

There are many other ways you can do it. For example, your work with the sense.us project (http://vis.berkeley.edu/papers/sense.us/) includes graphical annotation capabilities, and I think they help people feel creative while allowing them to make very precise, analytical comments. I would also like to see all visualization tools have the linking capabilities you've included that let you go seamlessly between comments and views.

One of the reasons that you can have a conversation on YouTube through videos is because videos are so easy to make. So, one of the questions is just how do we make visualizations easier for people to create? Your recent work with Mike Bostock on Protovis (http://protovis.org), not to mention all of your toolkit work before (http://prefuse.org/), is a great example. But that's a huge area where there's room for a lot of different approaches, from the simplest level where people can just plug in stuff to things that they can create completely. A goal that a lot of us share is that it should be as easy to create these visualizations as it is to write a comment. It should be as fluid, as flexible, and as expressive as writing.

JH What do you envision as the future of social data-analysis tools? Five or maybe 10 years from now, what would you like to see in terms of how people are interacting around data and visualizations?

FV One of the things I think is really promising is visualizing text. That has been mostly ignored so far in terms of information visualization tools, and yet a lot of the richest information we have is in text format.

We also need to consider what it means to think about visualization as a medium. How does

that change the kinds of tools we're building? I think the question you just asked is right on point. If communicating around visualization is so important, why are all the comments outside of the visualization? How can we integrate them? We need to begin thinking about visualization in a multimedia universe, which is not currently being done.

MW I also think we'll see people contributing data, as well as analysis. We've seen this explosion of crowd sourcing in all sorts of contexts. For example, people can go out to their neighborhood stores and compare the prices of milk. WNYC did a piece on this, which led to a visualization (http://www. wnyc.org/shows/bl/gouge_map_milk_07.html).

What's nice is when you get people participating in the full life cycle. It's not just, "Here's the data, analyze it," but they're also gathering the data and are deeply invested in the whole process. **Q**

LOVE IT, HATE IT? LET US KNOW
feedback@acmqueue.com