Visualize This

The FlowingData Guide to Design, Visualization, and Statistics

Nathan Yau

WILEY
Wiley Publishing, Inc.

Designing with a Purpose

9

When you explore your own data, you don't need to do much in terms of storytelling. You are, after all, the storyteller. However, the moment you use your graphic to present information—whether it's to one person, several thousand, or millions—a standalone chart is no longer good enough.

Sure, you want others to interpret results and perhaps form their own stories, but it's hard for readers to know what questions to ask when they don't know anything about the data in front of them. It's your job and responsibility to set the stage. How you design your graphics affects how readers interpret the underlying data.

Prepare Yourself

You need to know your source material to tell good stories with data. This is an often overlooked part of designing data graphics. When you start, it's easy to get excited about your end result. You want something amazing, beautiful, and interesting to look at, and this is great; but you can't do any of that if you have no idea what you're visualizing. You'll just end up with something like Figure 9-1. How can you explain interesting points in a dataset when you don't know the data?

Learn about the numbers and metrics. Figure out where they came from and how they were estimated, and see if they even make sense. This early data gathering process is what makes graphics in *The New York Times* so good. You see the end results in the paper and on the web, but you miss all the work that goes into the graphics before a single shape is drawn. A lot of the time, it takes longer to get all the data in order than it does to design a graphic.

TIP

Visualization is about communicating data, so take the time to learn about what makes the base of your graphic, or you'll just end up spouting numbers.

So the next time you have a dataset in front of you, try not to jump right into design. That's the lazy person's way out, and it always shows in the end. Take the time to get to know your data and learn the context of the numbers.

Punch some numbers into R, read any accompanying documentation so that you know what each metric represents, and see if there's anything that looks weird. If there is something that looks weird, and you can't figure out why, you can always contact the source. People are usually happy to hear that someone is making use of the data they published and are eager to fix mistakes if there are any.

After you learn all you can about your data, you are ready to design your graphics. Think of it like this. Remember that part in *The Karate Kid* when Daniel is just starting to learn martial arts? Mister Miyagi tells him to wax a bunch of cars, sand a wooden floor, and refinish a fence, and then Daniel is frustrated because he feels like these are useless tasks. Then of course, it turns out that blocking and punching all of a sudden come natural to him because he's been working on all the right motions. It's the same thing with data. Learn all you can about the data, and the visual storytelling will come natural. If you haven't seen the movie, just nod your head in agreement. And then go add *The Karate Kid* to your Netflix queue.

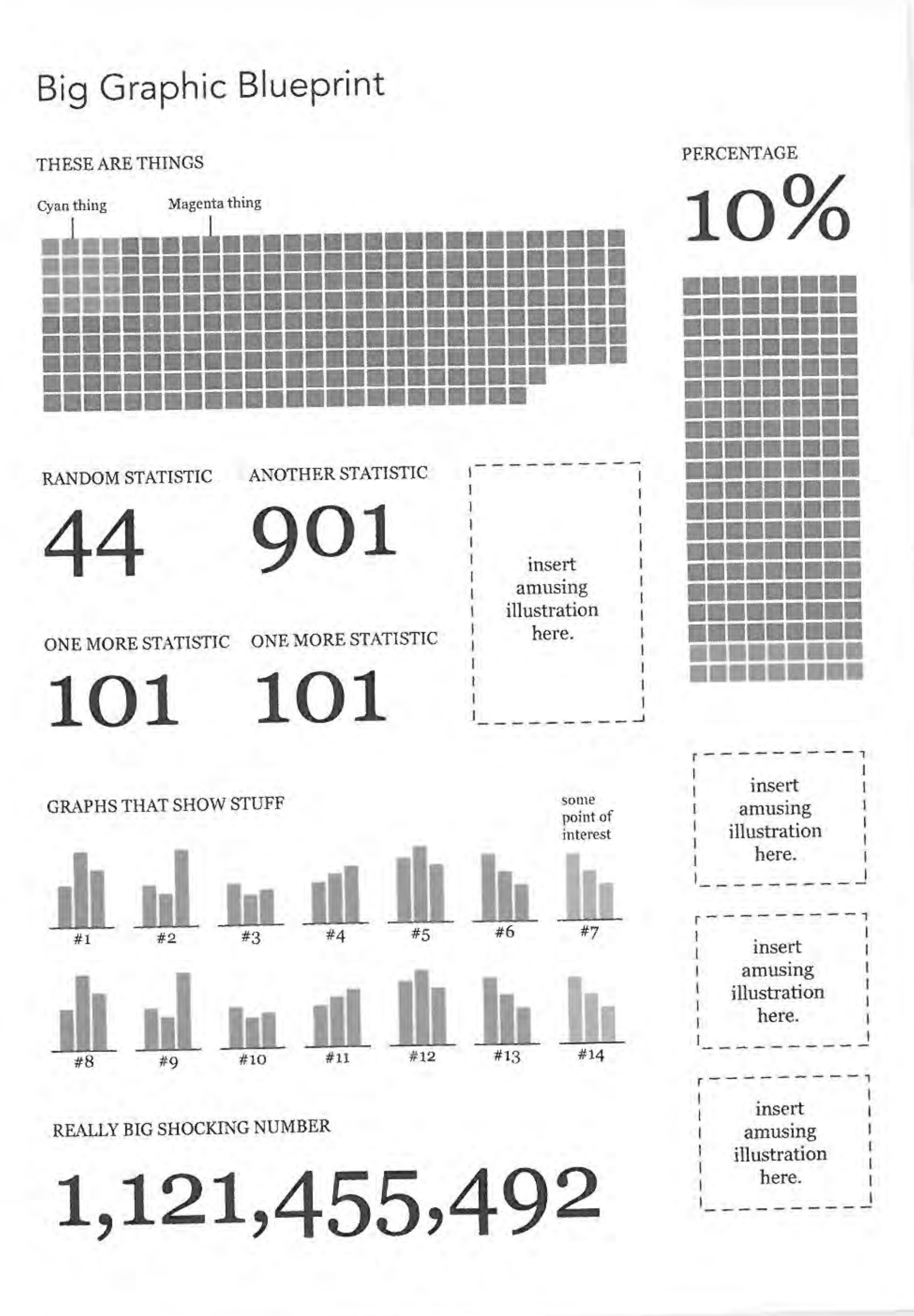


FIGURE 9-1 Big graphic blueprint. Go big or go home.

Prepare Your Readers

Your job as a data designer is to communicate what you know to your audience. They most likely didn't look at the data, so they might not see the same thing that you see if there's no explanation or setup. My rule of thumb is to assume that people are showing up to my graphics blindly, and with sharing via Facebook and Twitter and links from other blogs, that's not all that far off.

For example, Figure 9-2 shows a screenshot of an animated map I made. If you haven't seen this graphic before, you probably have no clue what you're looking at. Given the examples in Chapter 8, "Visualizing Spatial Relationships," your best guess might be openings for some store.

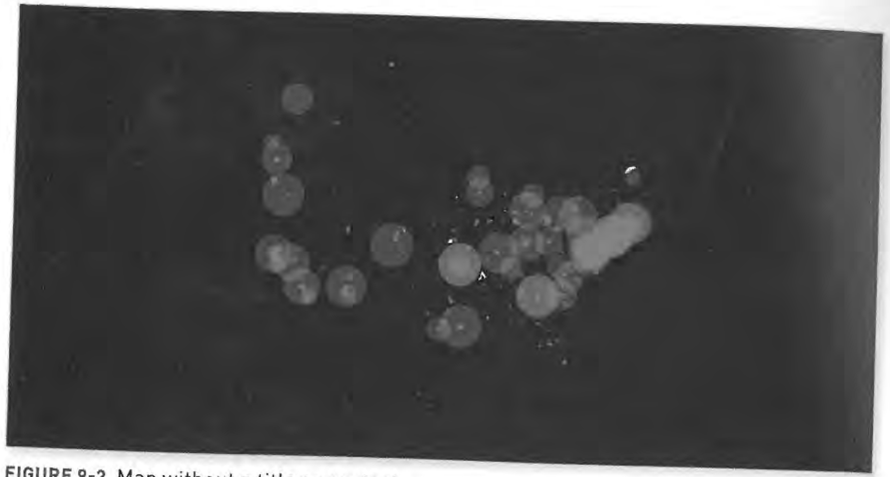


FIGURE 9-2 Map without a title or context

▶ Watch the full map animation at `http://datafl.ws/19n`.

The map actually shows geotagged tweets that were posted around the world during the inauguration of President Barack Obama on Tuesday, January 20, 2009, at noon Eastern Standard Time. The animation starts early Monday morning, and as the day moves on, more people wake and tweet at a steady rate. The number of tweets per hour increases as the event nears, and Europe gets in on some of the action as the United States sleeps. Then Tuesday morning starts, and then boom—there's huge excitement as the event actually happens. You can easily see this progression in Figure 9-3. Had I provided this context for Figure 9-2, it probably would've made a lot more sense.

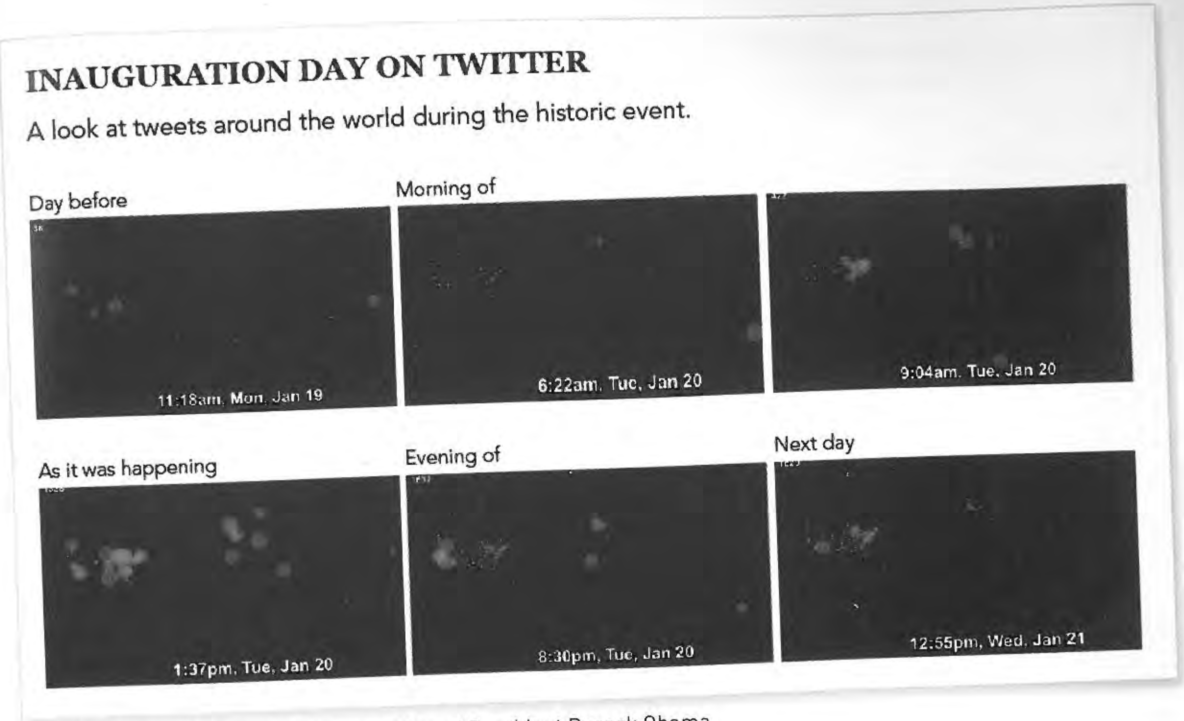
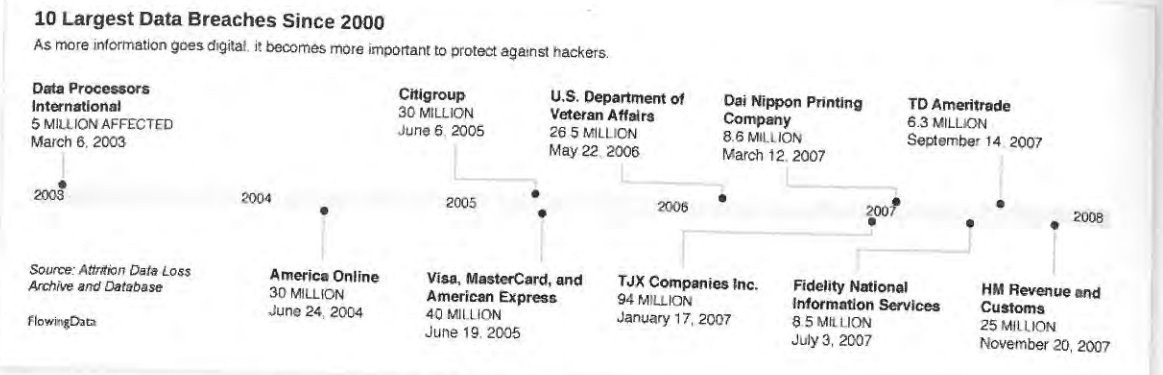


FIGURE 9-3 Tweets during the inauguration of President Barack Obama

You don't have to write an essay to accompany every graphic, but a title and a little bit of explanation via a lead-in are always helpful. It's often good to include a link somewhere on your graphic so that people can still find your words even if the graphic is shared on another site. Otherwise, it can quickly become like a game of Telephone, and before you know it, the graphic you carefully designed is explained with the opposite meaning you intended. The web is weird like that.

As another example, the graphic in Figure 9-4 is a simple timeline that shows the top ten data breaches at the time.

It's basic with only ten data points, but when I posted it on FlowingData, I brought up how the breaches grow higher in frequency as you move from 2000 to 2008. The graphic ended up getting shared quite a bit, with a variant even ending up in *Forbes* magazine. Almost everyone brought up that last bit. I don't think people would've given the graphic much thought had I not provided that simple observation.



10 Largest Data Breaches Since 2000

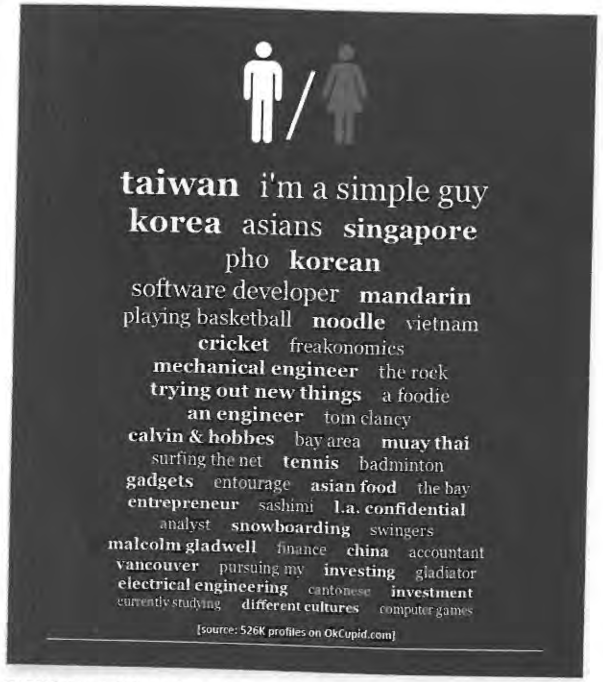
As more information goes digital, it becomes more important to protect against hackers.

Source: Attrition Data Loss Archive and Database

FlowingData

FIGURE 9-4 Major data breaches since 2000

The lesson: Don't assume your readers know everything or that they can spot features in your graphic. This is especially true with the web because people are used to clicking to the next thing.

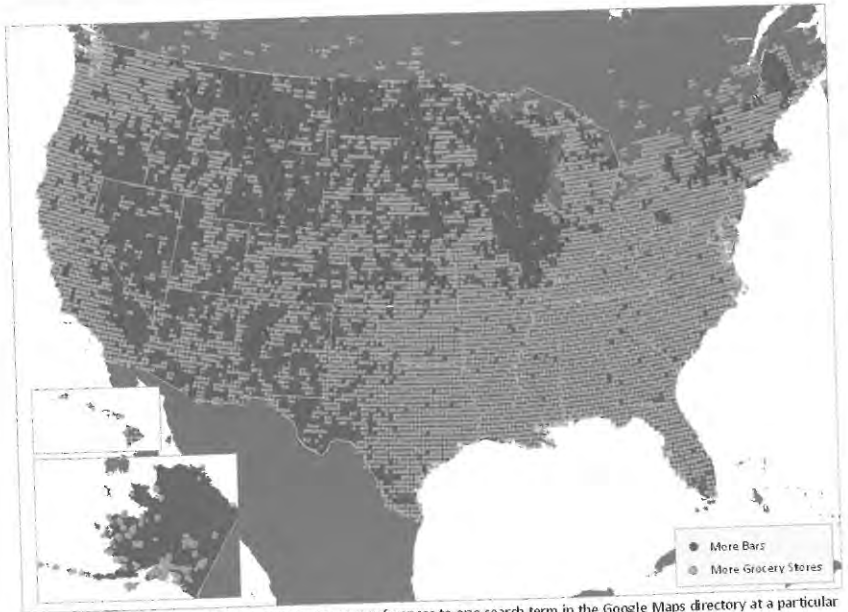


[source: 526K profiles on OkCupid.com]

FIGURE 9-5 What Asian guys like based on OkCupid online dating profiles

That's not to say that people won't spend time looking at data. As you might have seen, the OkCupid blog has been writing relatively long posts presenting results from thorough analyses of its online dating dataset. Titles include "The Best Questions for a First Date" and "The Mathematics of Beauty."

Posts on the blog have been viewed millions of times, and people love what the OkCupid folks have to say. In addition to the tons of context in the actual post, people also come to the blog with a bit of context of their own. Because it is data and findings about dating and the opposite sex, people can easily relate with their own experiences. Figure 9-5, for example, is a graphic that shows what Asian guys typically like, which is from an OkCupid post on what people like, categorized by race and gender. Hey, I'm Asian *and* a guy. Instant connection.



BARS AND GROCERY STORES: Shading represents more references to one search term in the Google Maps directory at a particular location, e.g., a red circle indicates more references to bars than grocery stores. Data collected in August 2008.

Copyleft Floatingsheep.org, 2010

FIGURE 9-6 Where bars outnumber grocery stores in the United States

On the other hand, when your graphic's topic is pollution levels or global debt, it can be a tough sell to a general audience if you don't do a good job of explaining.

Sometimes, no matter how much you explain, people simply don't like to read online, and they'll just skim. For example, I posted a map by FloatingSheep that compares number of bars to number of grocery stores in the United States, as shown in Figure 9-6. Red indicates areas where there are more bars than grocery stores, and orange indicates vice versa. The FloatingSheep guys called it the "beer belly of America."

Toward the end of the post, I wondered about the accuracy of the map and then finished up with, "Anyone who lives in the area care to confirm? I expect your comment to be filled with typos and make very little sense. And maybe smell like garbage." The lesson? Dry humor and sarcasm doesn't translate very well online, especially when people aren't used to reading your writing. I didn't actually expect comments to smell like garbage. Most people got the joke, but there were also a good number of insulted Wisconsinites. Like I said, the web is an interesting place (in a good way).

Visual Cues

In Chapter 1, "Telling Stories with Data," you saw how encodings work. Basically, you have data, and that data is encoded by geometry, color, or animation. Readers then decode those shapes, shades, and movement, mapping them back to numbers. This is the foundation of visualization. Encoding is a visual translation. Decoding helps you see data from a different angle and find patterns that you otherwise would not have seen if you looked only at the data in a table or a spreadsheet.

These encodings are usually straightforward because they are based on mathematical rules. Longer bars represent higher values, and smaller circles represent smaller values. Although your computer makes a lot of decisions during this process, it's still up to you to pick encodings appropriate for the dataset at hand.

Through all the examples in previous chapters, you've seen how good design not only lends to aesthetics, but also makes graphics easier to read and can change how readers actually feel about the data or the story you tell. Graphics with default settings from R or Excel feel raw and mechanical. This isn't necessarily a bad thing. Maybe that's all you want to show for an academic report. Or if your graphic is just a supplement to a more important body of writing, it could be better to not detract from what you want people to focus on. Figure 9-7 shows a generic bar plot that is about as plain as plain can be.

If, however, you do want to display your graphic prominently, a quick color change can make all the difference. Figure 9-8 is just Figure 9-7 with different background and foreground colors.

A darker color scheme might be used for a somber topic, whereas a brighter color scheme can feel more happy-go-lucky (Figure 9-9).

Of course, you don't always need a theme. You can use a neutral color palette if you like, as shown in Figure 9-10.

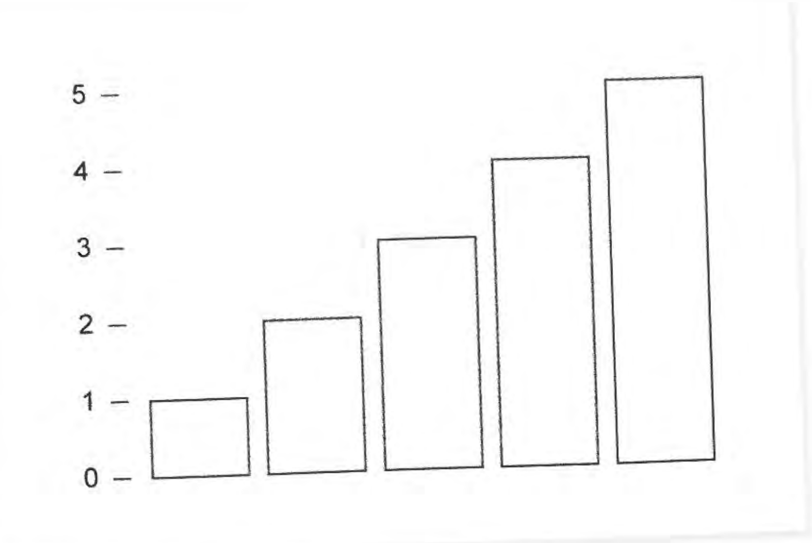


FIGURE 9-7 Plain bar plot

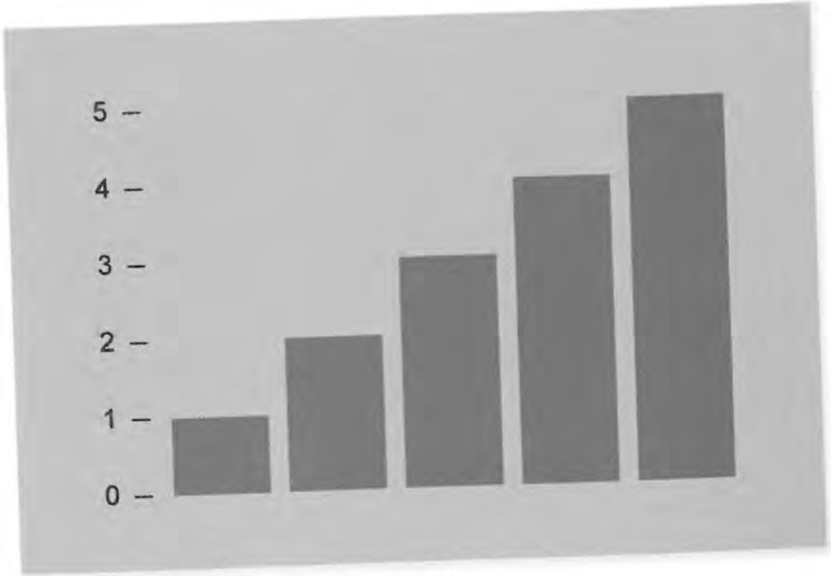


FIGURE 9-8 Default graph with dark color scheme

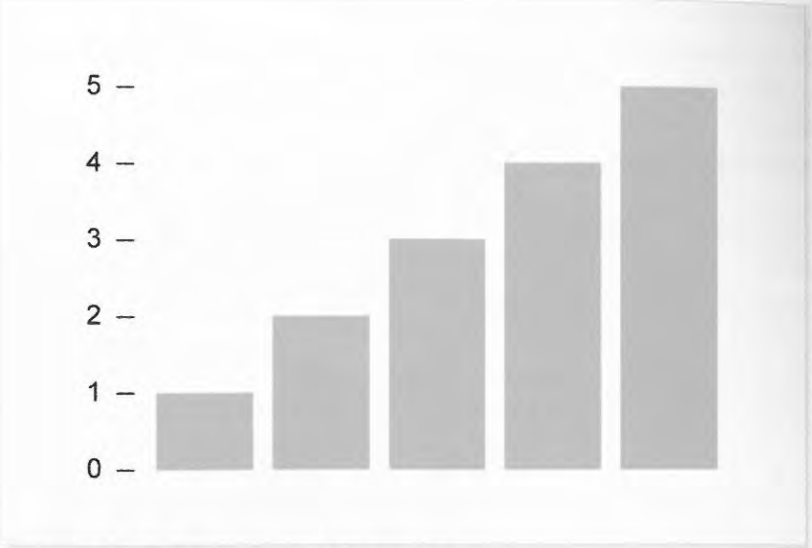


FIGURE 9-9 Default graph with light color scheme

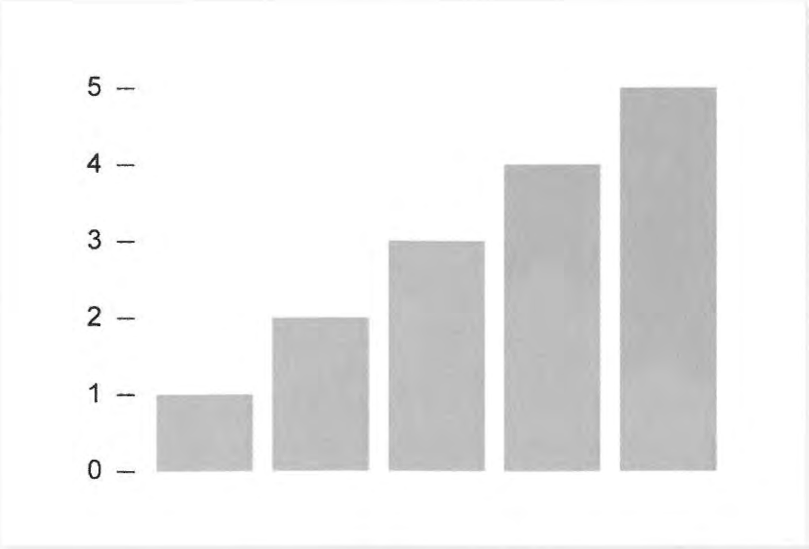


FIGURE 9-10 Default graph with neutral color scheme

The main point is that color choice can play a major role in data graphics. It can evoke emotions (or not) and help provide context. It's your responsibility to choose colors that represent an accurate message. Your colors should match the story you are trying to tell. As shown in Figure 9-11, a simple color change can change the meaning of your data completely. The graphic by designer David McCandless and design duo Always With Honor, explores the meaning of colors in different cultures. For example, black and white are often used to represent death; however, blue and green are more commonly used in Muslim and South American cultures, respectively.

Similarly, you can change geometry for a different look, feel, and meaning. For example, Figure 9-12 shows a randomly generated stacked bar chart with visualization researcher Mike Bostock's Data-Driven Documents. It has straight edges and distinct points, along with peaks and valleys.

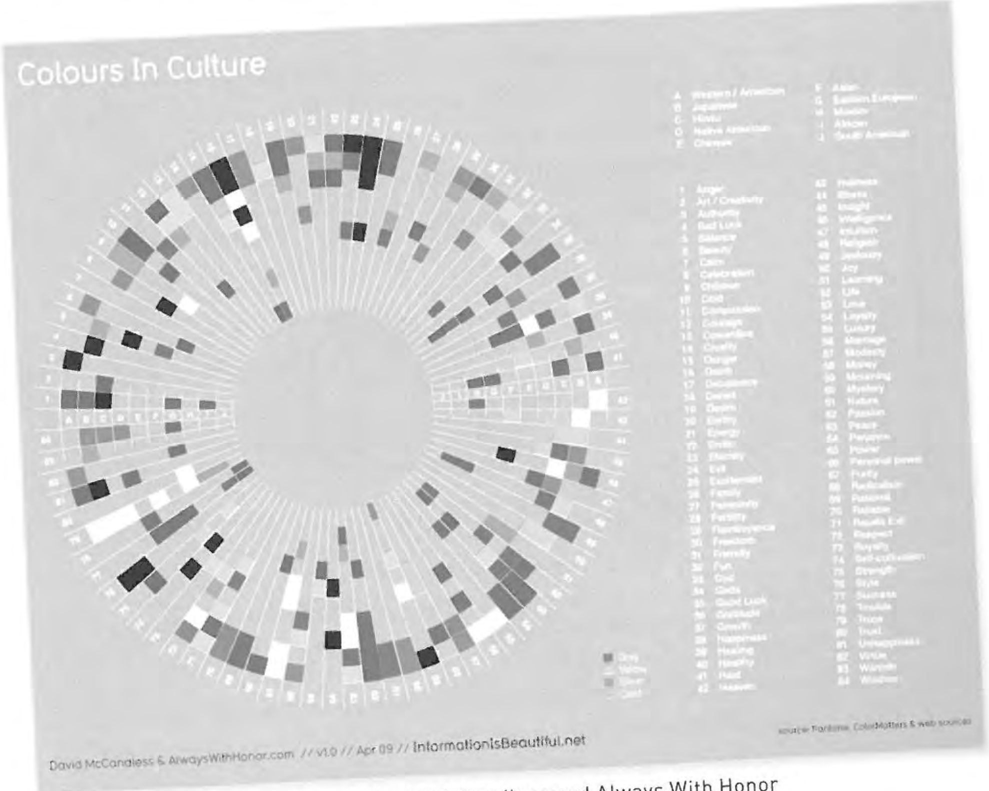


FIGURE 9-11 Colours In Culture by David McCandless and Always With Honor

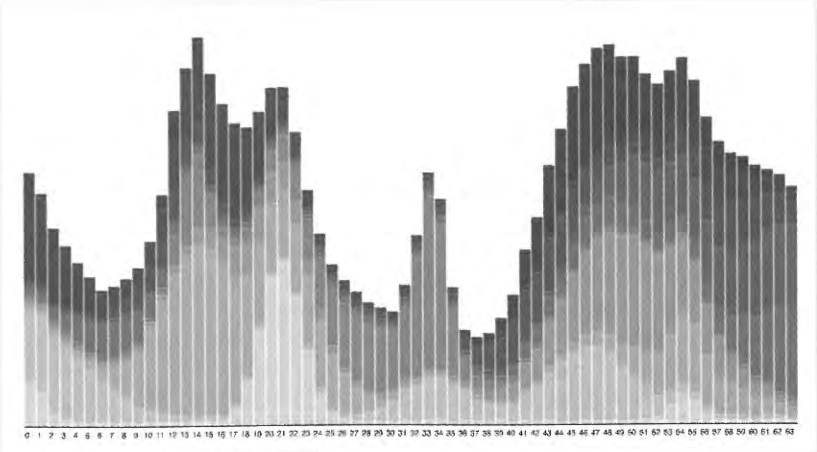


FIGURE 9-12 Randomly generated stacked bar chart

▶ Check out Lee Byron and Martin Wattenberg's paper, "Stacked Graphs—Geometry and Aesthetics" for more information on streamgraphs. Several packages are also available, such as Protovis and D3, that enable you to design your own.

If instead you used a streamgraph to show similar data, as shown in Figure 9-13, you clearly get a different feel. It's more free-flowing and continuous, and instead of peaks and valleys, you have tightening and swelling. At the same time though, the geometry between the two chart types is similar. The streamgraph is basically a smoothed stacked bar chart with the horizontal axis in the center instead of on the bottom.

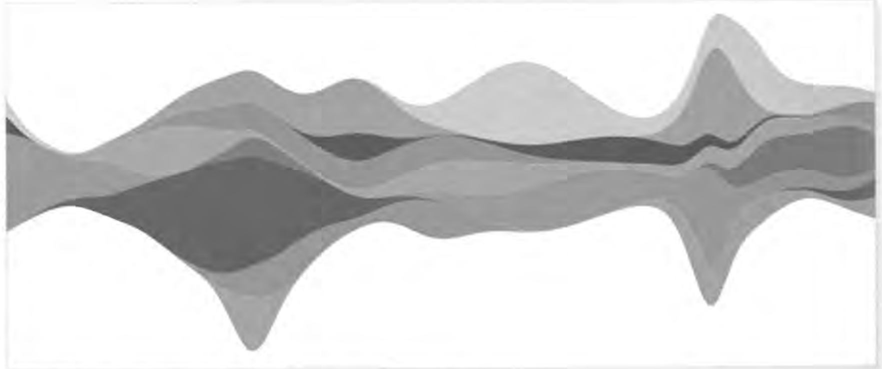
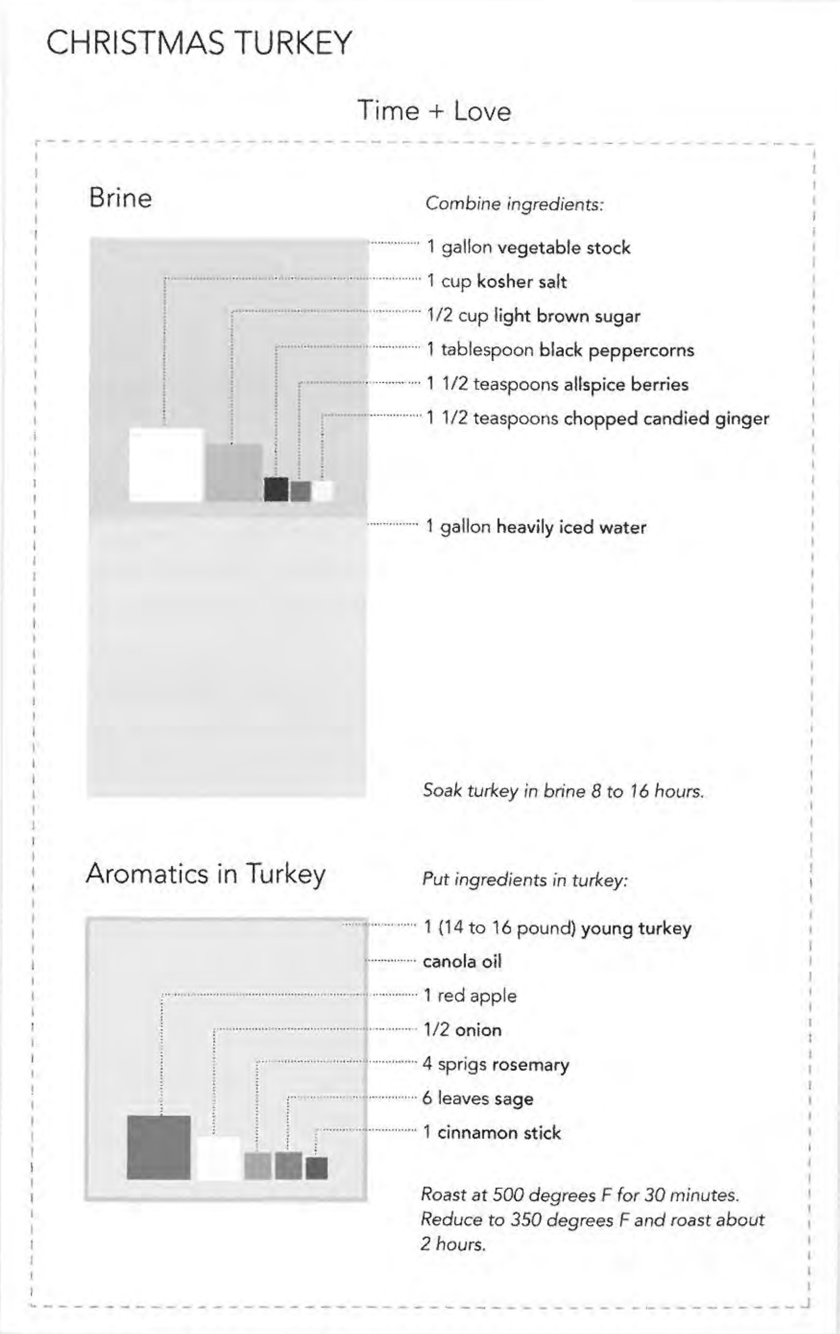


FIGURE 9-13 Randomly generated streamgraph

Sometimes context can simply come from how you organize shapes and colors. Figure 9-14 shows a graphic that I made for fun to celebrate the holidays. The top part shows the ingredients that go into brining your turkey, and on the bottom is what goes into the turkey when you roast it in the oven.



CHRISTMAS TURKEY

Time + Love

Brine

Combine ingredients:

1 gallon **vegetable stock**
1 cup **kosher salt**
1/2 cup **light brown sugar**
1 tablespoon **black peppercorns**
1 1/2 teaspoons **allspice berries**
1 1/2 teaspoons **chopped candied ginger**

1 gallon **heavily iced water**

Soak turkey in brine 8 to 16 hours.

Aromatics in Turkey

Put ingredients in turkey:

1 (14 to 16 pound) **young turkey**
canola oil
1 **red apple**
1/2 **onion**
4 sprigs **rosemary**
6 leaves **sage**
1 **cinnamon stick**

Roast at 500 degrees F for 30 minutes. Reduce to 350 degrees F and roast about 2 hours.

FIGURE 9-14 Recipe for Christmas turkey

The bottom line: At its most basic level, visualization is turning data, which can be numbers, text, categories, or any variety of things, into visual elements. Some visual cues work better than others, but applicability also varies by dataset. A method that's completely wrong for one dataset could fit perfectly for another. With practice, you can quickly decide what fits your purpose best.

Good Visualization

Although people have been charting and graphing data for centuries, only in the past few decades have researchers been studying what works and what doesn't. In that respect, visualization is a relatively new field. There still isn't a consensus on what visualization actually is. Is visualization something that has been generated by a computer following a set of rules? If a person has a hand in the design process, does that make it not a visualization? Are information graphics visualization, or do they belong in their own category?

Look online, and you can find lots of threads discussing differences and similarities between information graphics and visualization or essays that try to define what visualization is. It always leads to a never-ending back and forth without resolution. These opposing opinions lead to varied criteria for what makes a data graphic good or bad.

Statisticians and analysts, for example, generally think of visualization as traditional statistical graphics that they can use in their analyses. If a graphic or interactive doesn't help in analysis, then it's not useful. It's a failure. On the other hand, if you talk to graphic designers about the same graphic, they might think the work is a success because it displays the data of interest fairly and presents the data in an engaging way.

What you need to do is smush them all together, or at least get them in the same room together more often. The analytically minded can learn a lot from designers about making data more relatable and understandable, whereas design types can learn to dig deeper into data from their analytic counterparts.

I don't try to define what visualization is because the definition doesn't affect how I work. I consider the audience, the data in front of me, and ask myself whether the final graphic makes sense. Does it tell me what I want to know? If yes, then great. If no, I go back to the drawing board and figure out what would make the graphic better so that it answers the questions I have about the data. Ultimately, it's all about your goals for the graphic, what story you want to tell, and who you tell it to. Take all of the above into account—and you're golden.

Wrapping Up

A lot of data people see design as just a way to make your graphics look pretty. That's certainly part of it, but design is also about making your graphics readable, understandable, and usable. You can help people understand your data better than if they were to look at a default graph. You can clear clutter, highlight important points in your data, or even evoke an emotional response. Data graphics can be entertaining, fun, and informative. Sometimes it'll just be the former, depending on your goal, but no matter what you try to design—visualization, information graphic, or data art—let the data guide your work.

When you have a big dataset, and you don't know where to begin, the best place to start is with a question. What do you want to know? Are you looking for seasonal patterns? Relationships between multiple variables? Outliers? Spatial relationships? Then look back to your data to see if you can answer your question. If you don't have the data you need, then look for more.

When you have your data, you can use the skills you learned from the examples in this book to tell an interesting story. Don't stop here, though. Think of the material you worked through as a foundation. At the core of all your favorite data graphics is a data type and a visualization method that you now know how to work with. You can build on these for more advanced and complex graphics. Add interactions, combine plots, or complement your graphics with photographs and words to add more context.

Remember: Data is simply a representation of real life. When you visualize data, you visualize what's going on around you and in the world. You can see what's going on at a micro-level with individuals or on a much larger scale spanning the universe. Learn data, and you can tell stories that most people don't even know about yet but are eager to hear. There's more data to play with than ever before, and people want to know what it all means. Now you can tell them. Have fun.

Index