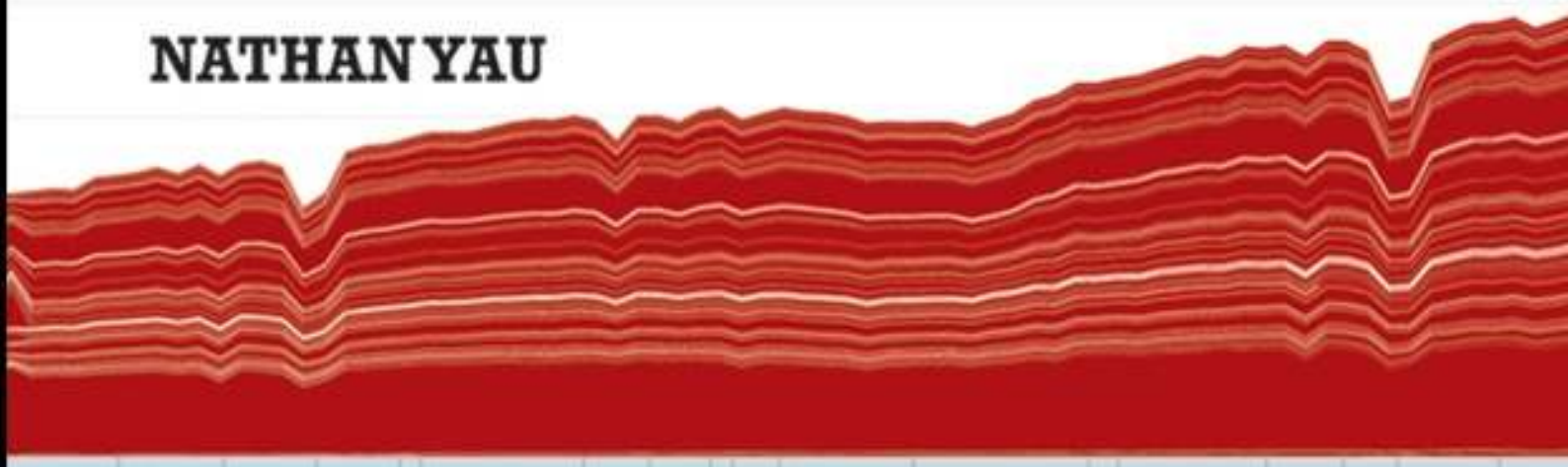
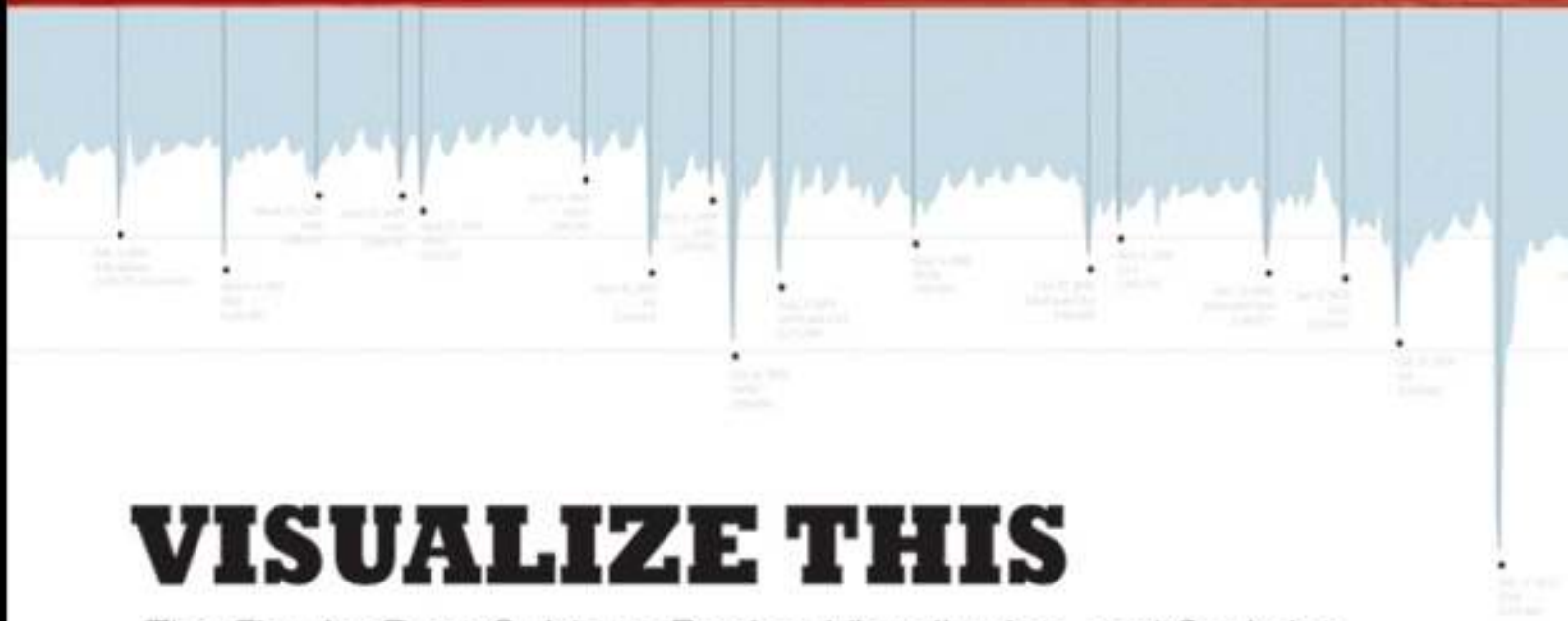
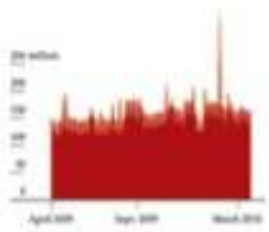
NATHAN YAU





VISUALIZE THIS

The FlowingData Guide to Design, Visualization, and Statistics



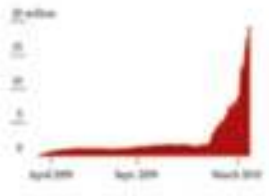


Table of Contents

Introduction

Learning Data

Chapter 1

Telling Stories with Data

Think of all the popular data visualization works out there—the ones that you always hear in lectures or read about in blogs, and the ones that popped into your head as you were reading this sentence. What do they all have in common? They all tell an interesting story. Maybe the story was to convince you of something. Maybe it was to compel you to action, enlighten you with new information, or force you to question your own preconceived notions of reality. Whatever it is, the best data visualization, big or small, for art or a slide presentation, helps you see what the data have to say.

More Than Numbers

Face it. Data can be boring if you don't know what you're looking for or don't know that there's something to look for in the first place. It's just a mix of numbers and words that mean nothing other than their raw values. The great thing about statistics and visualization is that they help you look beyond that. Remember, data is a representation of real life. It's not just a bucket of numbers. There are stories in that bucket. There's meaning, truth, and beauty. And just like real life, sometimes the stories are simple and straightforward; and other times they're complex and roundabout. Some stories belong in a textbook. Others come in novel form. It's up to you, the statistician, programmer, designer, or data scientist to decide how to tell the story.

This was one of the first things I learned as a statistics graduate student. I have to admit that before entering the program, I thought of statistics as pure analysis, and I thought of data as the output of a mechanical process. This is actually the case a lot of the time. I mean, I did major in electrical engineering, so it's not all that surprising I saw data in that light.

Don't get me wrong. That's not necessarily a bad thing, but what I've learned over the years is that data, while objective, often has a human dimension to it.

For example, look at unemployment again. It's easy to spout state averages, but as you've seen, it can vary a lot within the state. It can vary a lot by neighborhood. Probably someone you know lost a job over the past few years, and as the saying goes, they're not just another statistic, right? The numbers represent individuals, so you should approach the data in that way. You don't have to tell every individual's story. However, there's a subtle yet important difference between the unemployment rate increasing by 5 percentage points and several hundred thousand people left jobless. The former reads as a number without much context, whereas the latter is more relatable.

Journalism

A graphics internship at *The New York Times* drove the point home for me. It was only for 3 months during the summer after my second year of graduate school, but it's had a lasting impact on how I approach data. I didn't just learn how to create graphics for the news. I learned how to

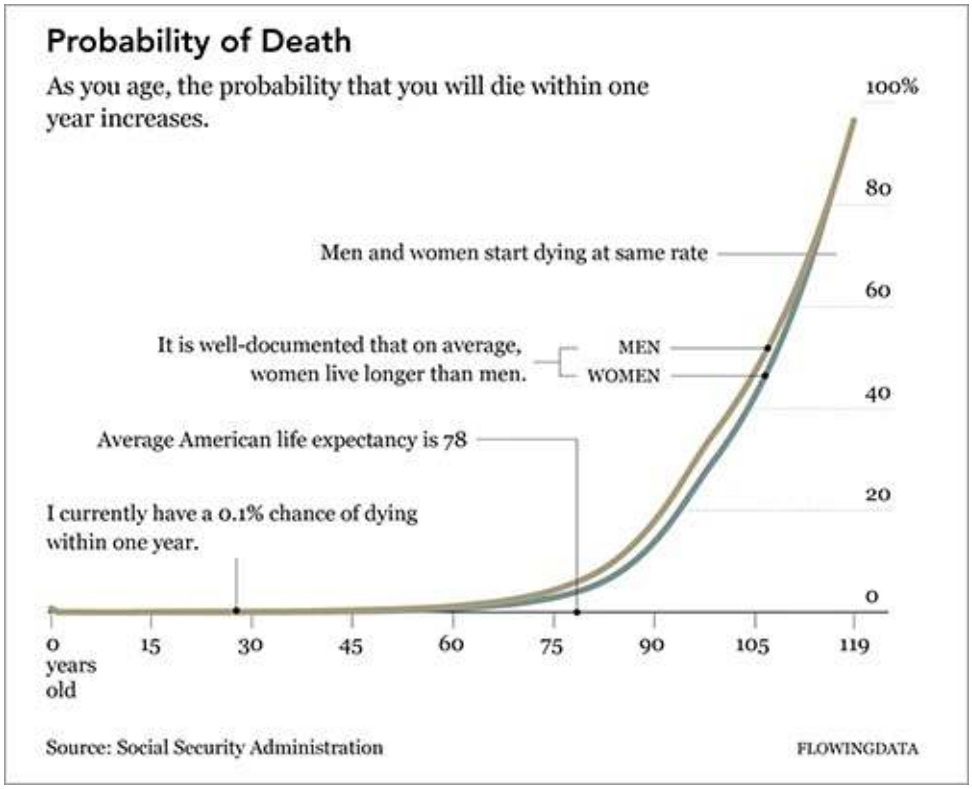
report data as the news, and with that came a lot of design, organization, fact checking, sleuthing, and research.

There was one day when my only goal was to verify three numbers in a dataset, because when *The New York Times* graphics desk creates a graphic, it makes sure what it reports is accurate. Only after we knew the data was reliable did we move on to the presentation. It's this attention to detail that makes its graphics so good.

Take a look at any *New York Times* graphic. It presents the data clearly, concisely, and ever so nicely. What does that mean though? When you look at a graphic, you get the chance to understand the data. Important points or areas are annotated; symbols and colors are carefully explained in a legend or with points; and the *Times* makes it easy for readers to see the story in the data. It's not just a graph. It's a graphic.

The graphic in Figure 1-1 is similar to what you will find in *The New York Times*. It shows the increasing probability that you will die within one year given your age.

Figure 1-1: Probability of death given your age



> Check out some of the best *New York Times* graphics at http://datafl.ws/nytimes.

The base of the graphic is simply a line chart. However, design elements help tell the story better. Labeling and pointers provide context and help you see why the data is interesting; and line width and color direct your eyes to what's important.

Chart and graph design isn't just about making statistical visualization but also explaining what the visualization shows.

> **Note**
>
> See Geoff McGhee's video documentary "Journalism in the Age of Data" for more on how journalists use data to report current events. This includes great interviews with some of the best in the business.

The New York Times is objective. It presents the data and gives you the facts. It does a great job at that. On the opposite side of the spectrum, visualization is less about analytics and more about tapping into your emotions. Jonathan Harris and Sep Kamvar did this quite literally in *We Feel Fine* (Figure 1-2).

Figure 1-2: We Feel Fine by Jonathan Harris and Sep Kamvar

The interactive piece scrapes sentences and phrases from personal public blogs and then visualizes them as a box of floating bubbles. Each bubble represents an emotion and is color-coded accordingly. As a whole, it is like individuals floating through space, but watch a little longer and you see bubbles start to cluster. Apply sorts and categorization through the interface to see how these seemingly random vignettes connect. Click an individual bubble to see a single story. It's poetic and revealing at the same time.

> Interact and explore people's emotions in Jonathan Harris and Sep Kamvar's live and online piece at http://wefeelfine.org.

There are lots of other examples such as Golan Levin's *The Dumpster*, which explores blog entries that mention breaking up with a significant other; Kim Asendorf's *Sumedicina*, which tells a fictional story of a man running from a corrupt organization, with not words, but graphs and charts; or Andreas Nicolas Fischer's physical sculptures that show economic downturn in the United States.

> See FlowingData for many more examples of art and data at http://datafl.ws/art.

The main point is that data and visualization don't always have to be just about the cold, hard facts. Sometimes you're not looking for analytical insight. Rather, sometimes you can tell the story from an emotional point of view that encourages viewers to reflect on the data. Think of it like this. Not all movies have to be documentaries, and not all visualization has to be traditional

charts and graphs.

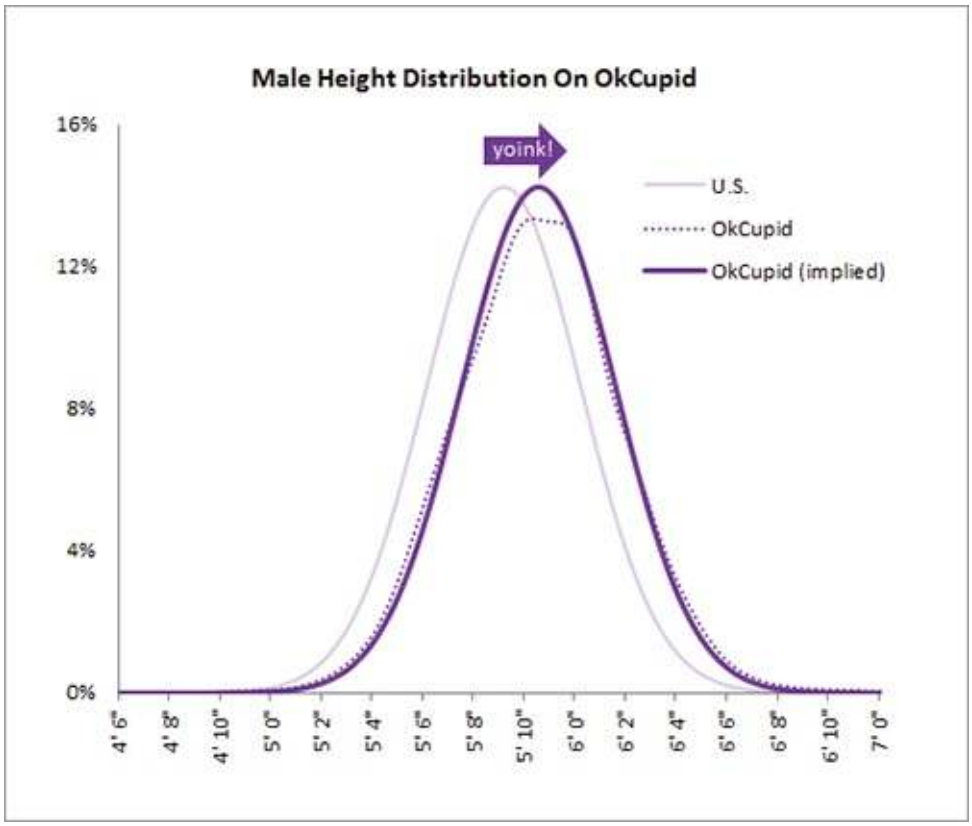
Entertainment

Somewhere in between journalism and art, visualization has also found its way into entertainment. If you think of data in the more abstract sense, outside of spreadsheets and comma-delimited text files, where photos and status updates also qualify, this is easy to see.

Facebook used status updates to gauge the happiest day of the year, and online dating site OkCupid used online information to estimate the lies people tell to make their digital selves look better, as shown in Figure 1-3. These analyses had little to do with improving a business, increasing revenues, or finding glitches in a system. They circulated the web like wildfire because of their entertainment value. The data revealed a little bit about ourselves and society.

Facebook found the happiest day to be Thanksgiving, and OkCupid found that people tend to exaggerate their height by about 2 inches.

Figure 1-3: Male Height Distribution on OkCupid



> Check out the OkTrends blog for more revelations from online dating such as what white people really like and how not to be ugly by accident: http://blog.okcupid.com.

Compelling

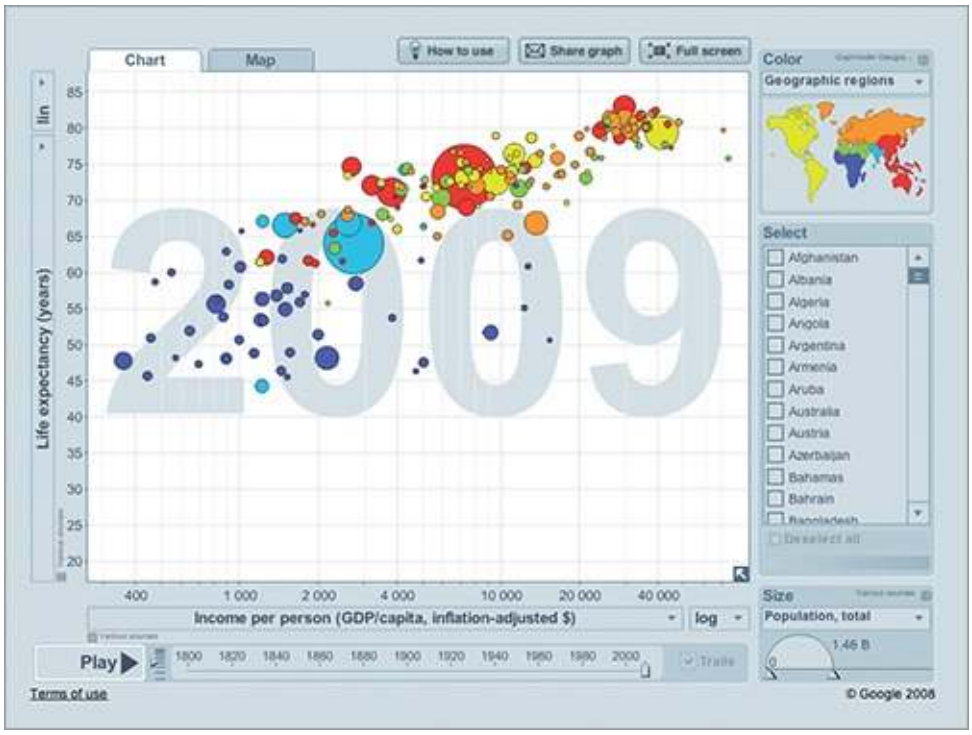
Of course, stories aren't always to keep people informed or entertained. Sometimes they're meant to provide urgency or compel people to action. Who can forget that point in *An Inconvenient Truth* when Al Gore stands on that scissor lift to show rising levels of carbon dioxide?

For my money though, no one has done this better than Hans Rosling, professor of International Health and director of the Gapminder Foundation. Using a tool called Trendalyzer, as shown in Figure 1-4, Rosling runs an animation that shows changes in poverty by country. He does this

during a talk that first draws you in deep to the data and by the end, everyone is on their feet applauding. It's an amazing talk, so if you haven't seen it yet, I highly recommend it.

The visualization itself is fairly basic. It's a motion chart. Bubbles represent countries and move based on the corresponding country's poverty during a given year. Why is the talk so popular then? Because Rosling speaks with conviction and excitement. He tells a story. How often have you seen a presentation with charts and graphs that put everyone to sleep? Instead Rosling gets the meaning of the data and uses that to his advantage. Plus, the sword-swallowing at the end of his talk drives the point home. After I saw Rosling's talk, I wanted to get my hands on that data and take a look myself. It was a story I wanted to explore, too.

Figure 1-4: Trendalyzer by the Gapminder Foundation



> Watch Hans Rosling wow the audience with data and an amazing demonstration at http://datafl.ws/hans.

I later saw a Gapminder talk on the same topic with the same visualizations but with a different speaker. It wasn't nearly as exciting. To be honest, it was kind of a snoozer. There wasn't any emotion. I didn't feel any conviction or excitement about the data. So it's not just about the data that makes for interesting chatter. It's how you present it and design it that can help people remember.

When it's all said and done, here's what you need to know. Approach visualization as if you were telling a story. What kind of story are you trying to tell? Is it a report, or is it a novel? Do you want to convince people that action is necessary?

Think character development. Every data point has a story behind it in the same way that every character in a book has a past, present, and future. There are interactions and relationships between those data points. It's up to you to find them. Of course, before expert storytellers write novels, they must first learn to construct sentences.

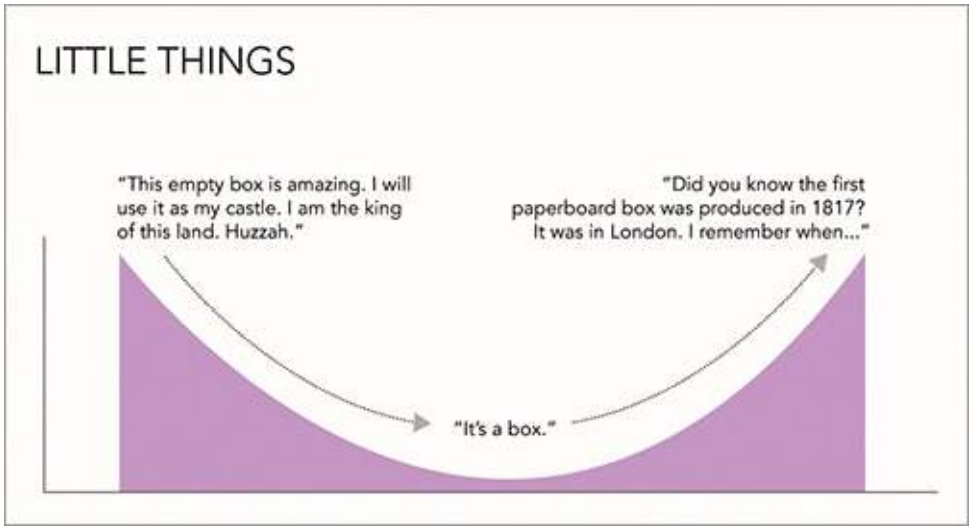
What to Look For

Okay, stories. Check. Now what kind of stories do you tell with data? Well, the specifics vary by dataset, but generally speaking, you should always be on the lookout for these two things whatever your graphic is for: patterns and relationships.

Patterns

Stuff changes as time goes by. You get older, your hair grays, and your sight starts to get kind of fuzzy (Figure 1-5). Prices change. Logos change. Businesses are born. Businesses die. Sometimes these changes are sudden and without warning. Other times the change happens so slowly you don't even notice.

Figure 1-5: A comic look at aging



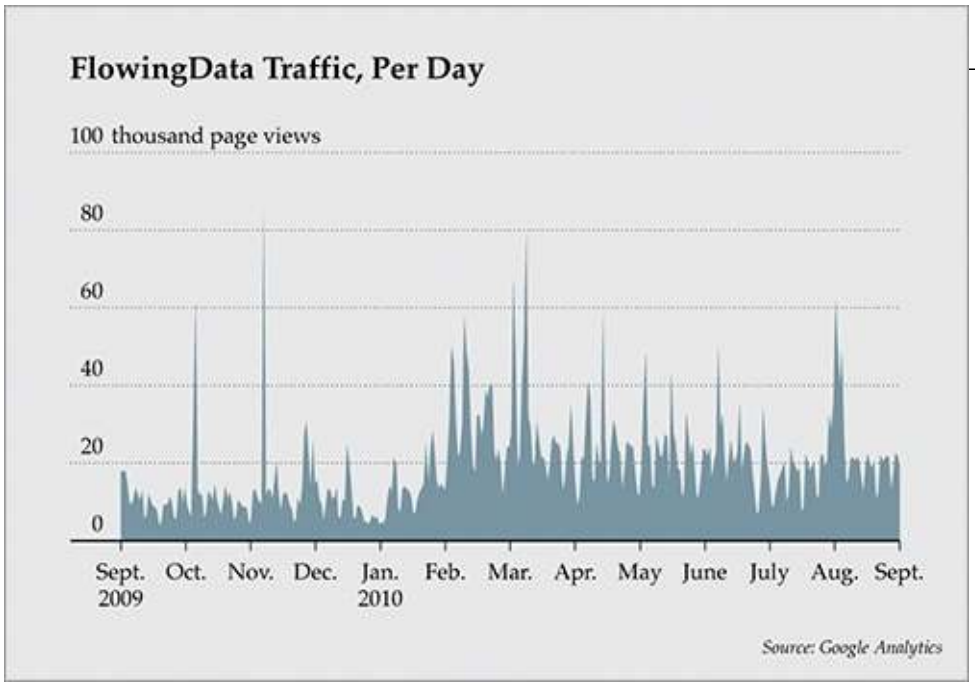
Whatever it is you're looking at, the change itself can be interesting as can the changing process. It is here you can explore patterns over time. For example, say you looked at stock prices over time. They of course increase and decrease, but by how much do they change per day? Per week? Per month? Are there periods when the stock went up more than usual? If so, why did it go up? Were there any specific events that triggered the change?

As you can see, when you start with a single question as a starting point, it can lead you to additional questions. This isn't just for time series data, but with all types of data. Try to approach your data in a more exploratory fashion, and you'll most likely end up with more interesting answers.

You can split your time series data in different ways. In some cases it makes sense to show hourly or daily values. Other times, it could be better to see that data on a monthly or annual basis. When you go with the former, your time series plot could show more noise, whereas the latter is more of an aggregate view.

Those with websites and some analytics software in place can identify with this quickly. When you look at traffic to your site on a daily basis, as shown in Figure 1-6, the graph is bumpier. There are a lot more fluctuations.

Figure 1-6: Daily unique visitors to FlowingData

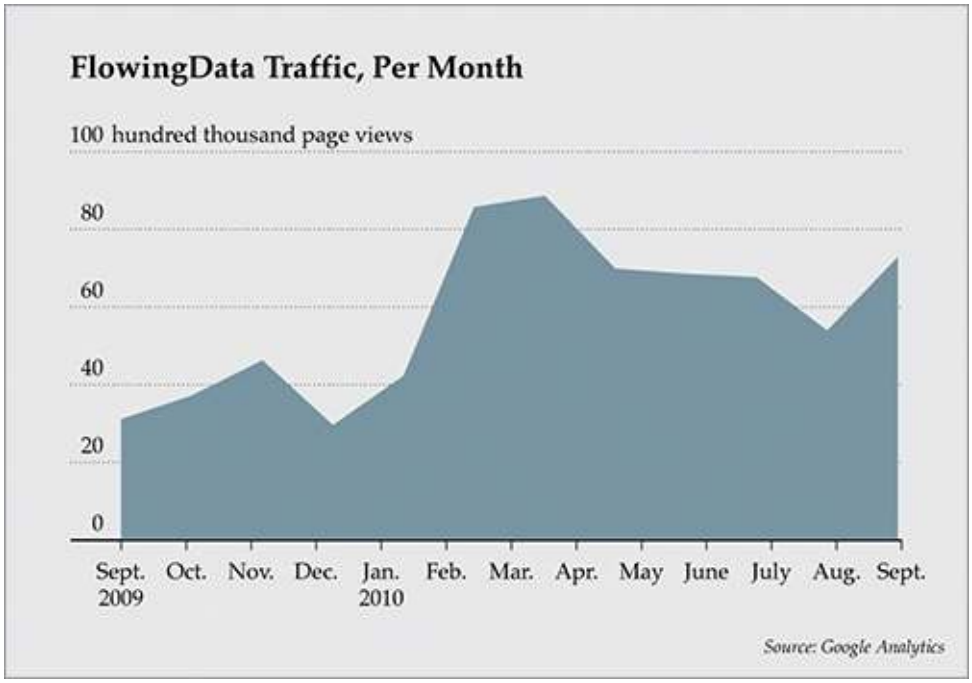


When you look at it on a monthly basis, as shown in Figure 1-7, fewer data points are on the same graph, covering the same time span, so it looks much smoother.

I'm not saying one graph is better than the other. In fact, they can complement each other. How you split your data depends on how much detail you need (or don't need).

Of course, patterns over time are not the only ones to look for. You can also find patterns in aggregates that can help you compare groups, people, and things. What do you tend to eat or drink each week? What does the President usually talk about during the State of the Union address? What states usually vote Republican? Looking at patterns over geographic regions would be useful in this case. While the questions and data types are different, your approach is similar, as you'll see in the following chapters.

Figure 1-7: Monthly unique visitors to FlowingData



Relationships

Have you ever seen a graphic with a whole bunch of charts on it that seemed like they've been

randomly placed? I'm talking about the graphics that seem to be missing that special something, as if the designer gave only a little bit of thought to the data itself and then belted out a graphic to meet a deadline. Often, that special something is relationships.

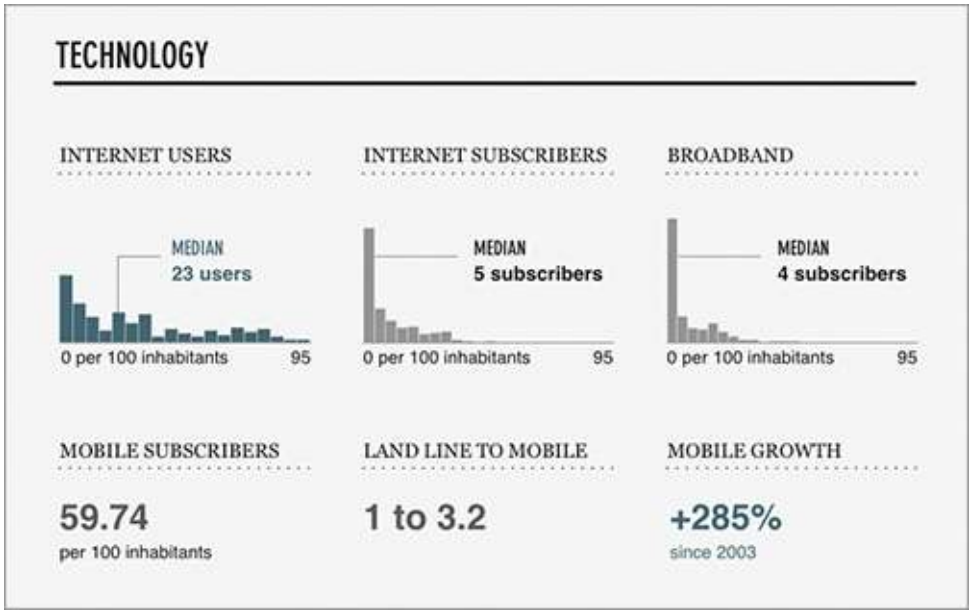
In statistics, this usually means correlation and causation. Multiple variables might be related in some way. Chapter 6, "Visualizing Relationships," covers these concepts and how to visualize them.

At a more abstract level though, where you're not thinking about equations and hypothesis tests, you can design your graphics to compare and contrast values and distributions visually. For a simple example, look at this excerpt on technology from the *World Progress Report* in Figure 1-8.

> The World Progress Report was a graphical report that compared progress around the world using data from UNdata. See the full version at http://datafl.ws/12i.

These are histograms that show the number of users of the Internet, Internet subscriptions, and broadband per 100 inhabitants. Notice that the range for Internet users (0 to 95 per 100 inhabitants) is much wider than that of the other two datasets.

Figure 1-8: Technology adoption worldwide



The quick-and-easy thing to do would have been to let your software decide what range to use for each histogram. However, each histogram was made on the same range even though there were no countries who had 95 Internet subscribers or broadband users per 100 inhabitants. This enables you to easily compare the distributions between the groups.

So when you end up with a lot of different datasets, try to think of them as several groups instead of separate compartments that do not interact with each other. It can make for more interesting results.

Questionable Data

While you're looking for the stories in your data, you should always question what you see. Remember, just because it's numbers doesn't mean it's true.

I have to admit. Data checking is definitely my least favorite part of graph-making. I mean, when someone, a group, or a service provides you with a bunch of data, it should be up to them to

make sure all their data is legit. But this is what good graph designers do. After all, reliable builders don't use shoddy cement for a house's foundation, so don't use shoddy data to build your data graphic.

Data-checking and verification is one of the most important—if not *the* most important—part of graph design.

Basically, what you're looking for is stuff that makes no sense. Maybe there was an error at data entry and someone added an extra zero or missed one. Maybe there were connectivity issues during a data scrape, and some bits got mucked up in random spots. Whatever it is, you need to verify with the source if anything looks funky.

The person who supplied the data usually has a sense of what to expect. If you were the one who collected the data, just ask yourself if it makes sense: That state is 90 percent of whatever and all other states are only in the 10 to 20 percent range. What's going on there?

Often, an anomaly is simply a typo, and other times it's actually an interesting point in your dataset that could form the whole drive for your story. Just make sure you know which one it is.

Design

When you have all your data in order, you're ready to visualize. Whatever you're making, whether it is for a report, an infographic online, or a piece of data art, you should follow a few basic rules. There's wiggle room with all of them, and you should think of what follows as more of a framework than a hard set of rules, but this is a good place to start if you are just getting into data graphics.

Explain Encodings

The design of every graph follows a familiar flow. You get the data; you encode the data with circles, bars, and colors; and then you let others read it. The readers have to decode your encodings at this point. What do these circles, bars, and colors represent?

William Cleveland and Robert McGill have written about encodings in detail. Some encodings work better than others. But it won't matter what you choose if readers don't know what the encodings represent in the first place. If they can't decode, the time you spend designing your graphic is a waste.

Note

See Cleveland and McGill's paper on *Graphical Perception and Graphical Methods for Analyzing Data* for more on how people encode shapes and colors.

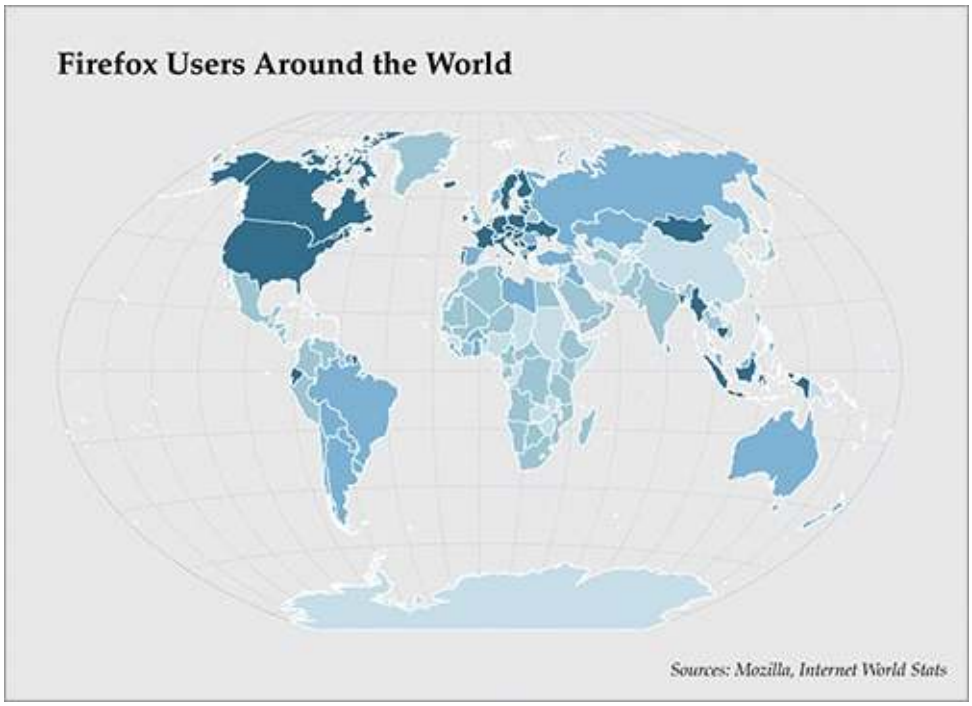
You sometimes see this lack of context with graphics that are somewhere in between data art and infographic. You definitely see it a lot with data art. A label or legend can completely mess up the vibe of a piece of work, but at the least, you can include some information in a short description paragraph. It helps others appreciate your efforts.

Other times you see this in actual data graphics, which can be frustrating for readers, which is the last thing you want. Sometimes you might forget because you're actually working with the data, so you know what everything means. Readers come to a graphic blind though without the

context that you gain from analyses.

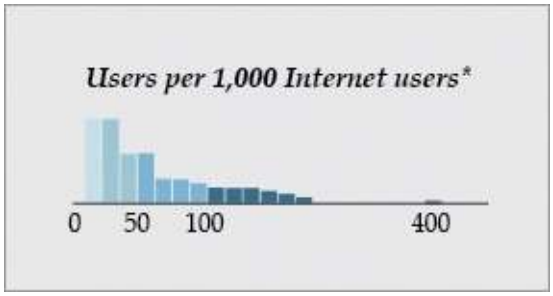
So how can you make sure readers can decode your encodings? Explain what they mean with labels, legends, and keys. Which one you choose can vary depending on the situation. For example, take a look at the world map in Figure 1-9 that shows usage of Firefox by country.

Figure 1-9: Firefox usage worldwide by country



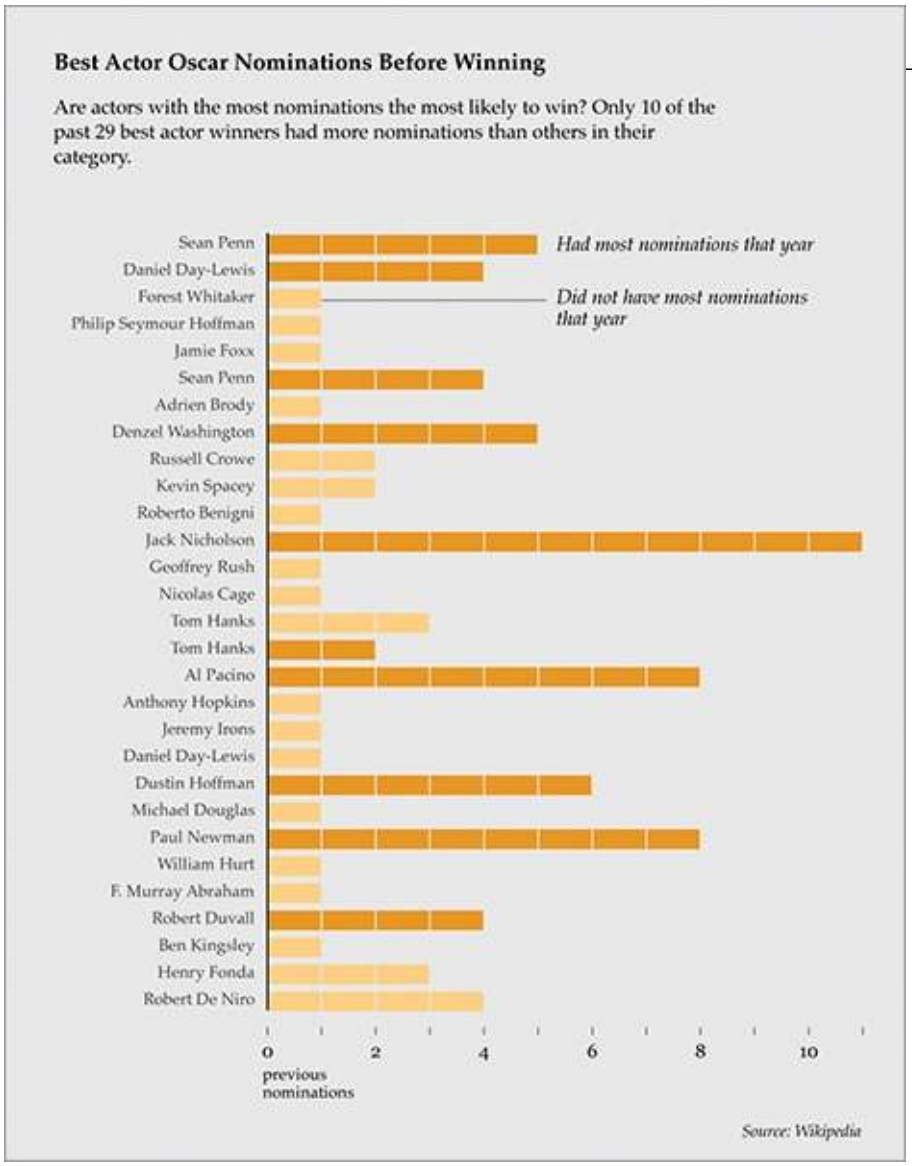
You can see different shades of blue for different countries, but what do they mean? Does dark blue mean more or less usage? If dark blue means high usage, what qualifies as high usage? As-is, this map is pretty useless to us. But if you provide the legend in Figure 1-10, it clears things up. The color legend also serves double time as a histogram showing the distribution of usage by number of users.

Figure 1-10: Legend for Firefox usage map



You can also directly label shapes and objects in your graphic if you have enough space and not too many categories, as shown in Figure 1-11. This is a graph that shows the number of nominations an actor had before winning an Oscar for best actor.

Figure 1-11: Directly labeled objects



A theory floated around the web that actors who had the most nominations among their cohorts in a given year generally won the statue. As labeled, dark orange shows actors who did have the most nominations, whereas light orange shows actors who did not.

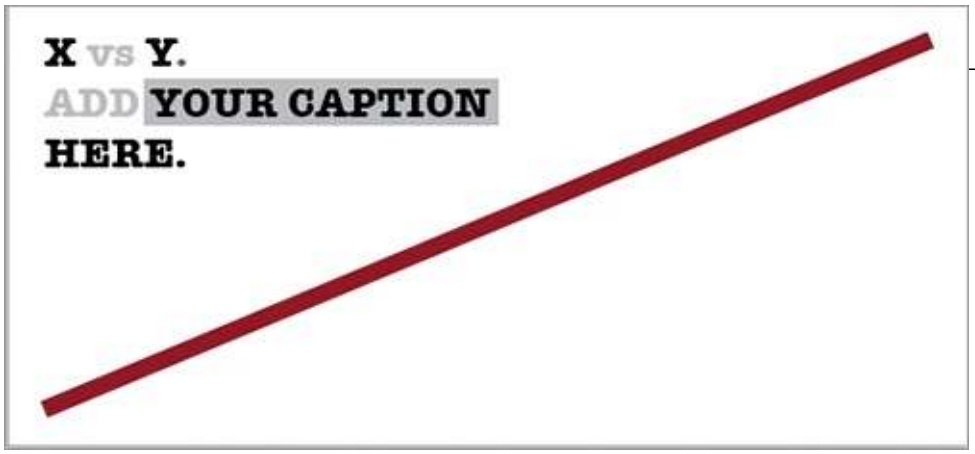
As you can see, plenty of options are available to you. They're easy to use, but these small details can make a huge difference on how your graphic reads.

Label Axes

Along the same lines as explaining your encodings, you should always label your axes. Without labels or an explanation, your axes are just there for decoration. Label your axes so that readers know what scale points are plotted on. Is it logarithmic, incremental, exponential, or per 100 flushing toilets? Personally, I always assume it's that last one when I don't see labels.

To demonstrate my point, rewind to a contest I held on FlowingData a couple of years ago. I posted the image in Figure 1-12 and asked readers to label the axes for maximum amusement.

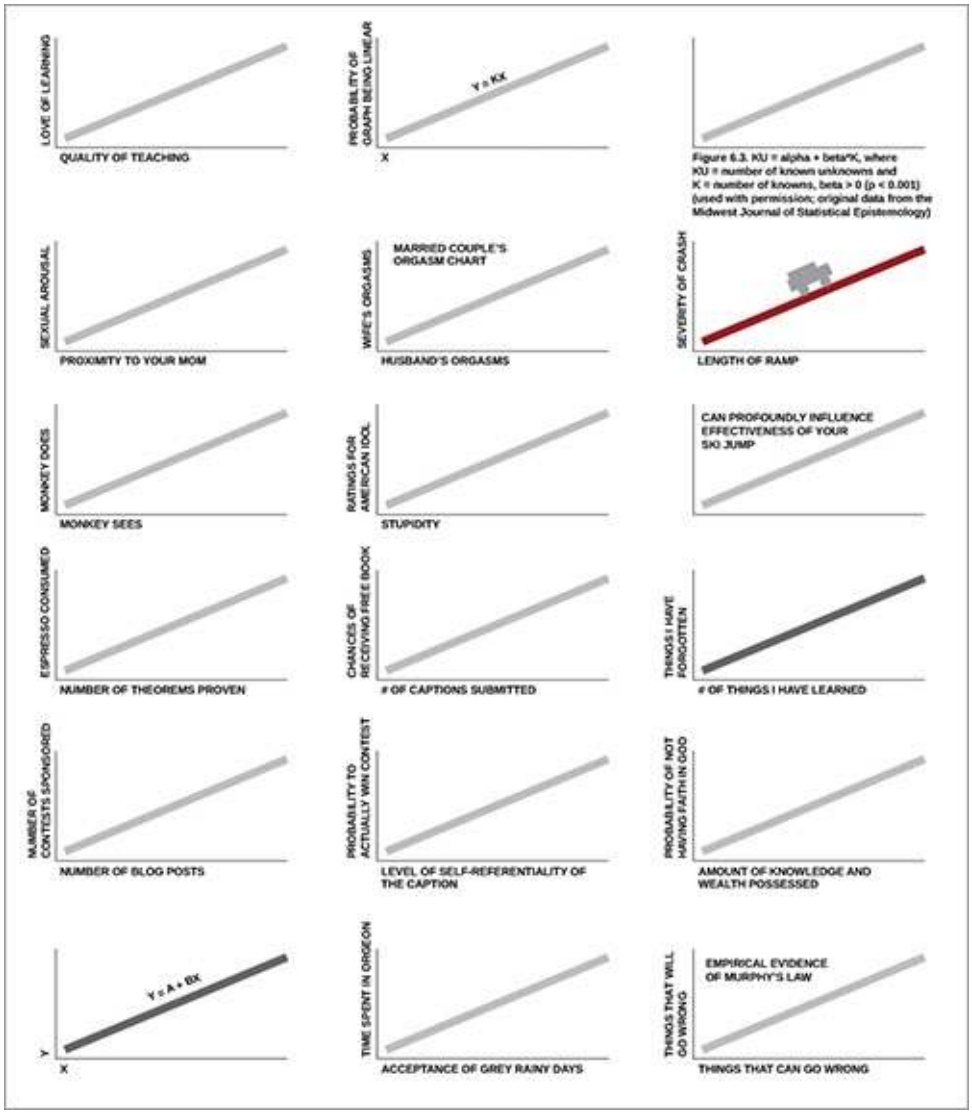
Figure 1-12: Add your caption here.



There were about 60 different captions for the same graph; Figure 1-13 shows a few.

As you can see, even though everyone looked at the same graph, a simple change in axis labels told a completely different story. Of course, this was just for play. Now just imagine if your graph were meant to be taken seriously. Without labels, your graph is meaningless.

Figure 1-13: Some of the results from a caption contest on FlowingData

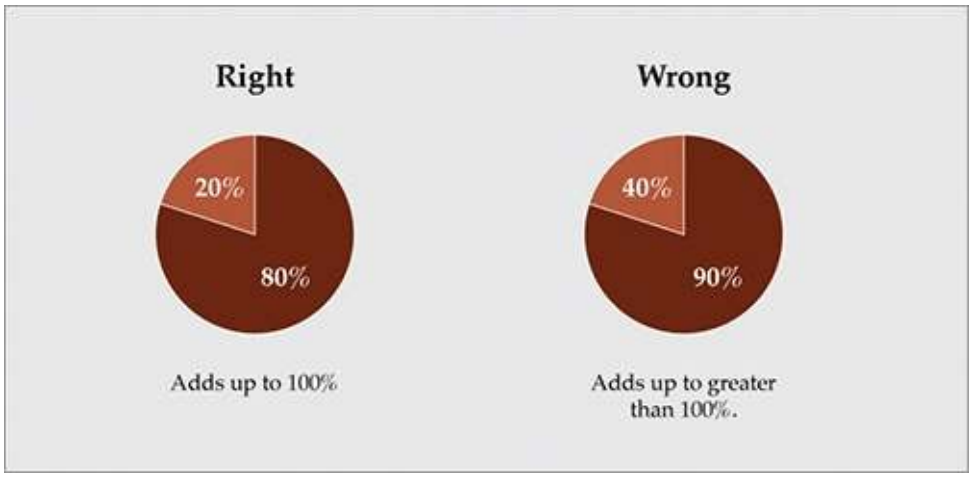


Keep Your Geometry in Check

When you design a graph, you use geometric shapes. A bar graph uses rectangles, and you use the length of the rectangles to represent values. In a dot plot, the position indicates value—same thing with a standard time series chart. Pie charts use angles to indicate value, and the sum of the values

always equal 100 percent (see Figure 1-14). This is easy stuff, so be careful because it's also easy to mess up. You're going to make a mistake if you don't pay attention, and when you do mess up, people, especially on the web, won't be afraid to call you out on it.

Figure 1-14: The right and wrong way to make a pie chart



Another common mistake is when designers start to use two-dimensional shapes to represent values, but size them as if they were using only a single dimension. The rectangles in a bar chart are two-dimensional, but you only use one length as an indicator. The width doesn't mean anything. However, when you create a bubble chart, you use an area to represent values. Beginners often use radius or diameter instead, and the scale is totally off.

Figure 1-15 shows a pair of circles that have been sized by area. This is the right way to do it.

Figure 1-15: The right way to size bubbles

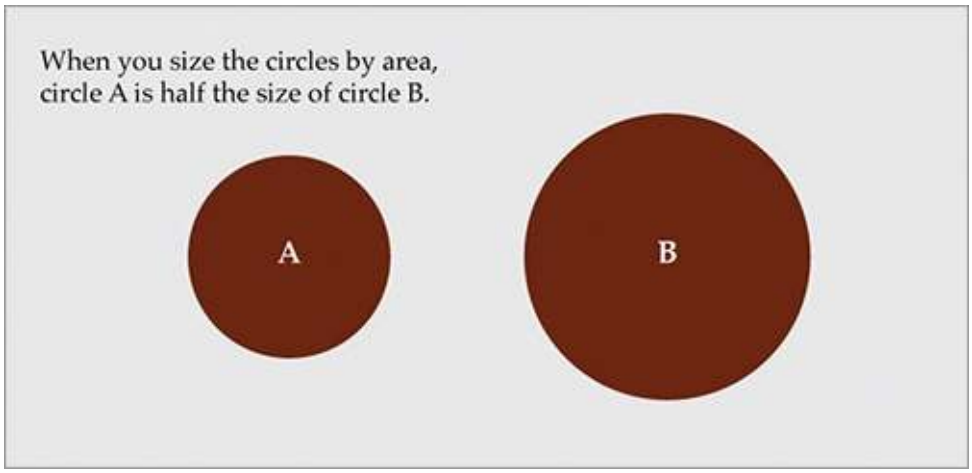
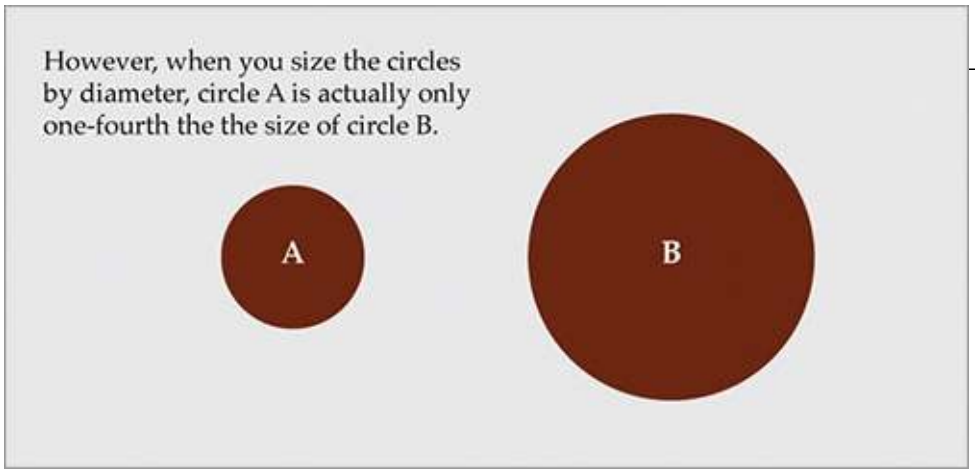


Figure 1-16 shows a pair of circles sized by diameter. The first circle has twice the diameter as that of the second but is four times the area.

It's the same deal with rectangles, like in a treemap. You use the area of the rectangles to indicate values instead of the length or width.

Figure 1-16: The wrong way to size bubbles



Include Your Sources

This should go without saying, but so many people miss this one. Where did the data come from? If you look at the graphics printed in the newspaper, you always see the source somewhere, usually in small print along the bottom. You should do the same. Otherwise readers have no idea how accurate your graphic is.

There's no way for them to know that the data wasn't just made up. Of course, you would never do that, but not everyone will know that. Other than making your graphics more reputable, including your source also lets others fact check or analyze the data.

Inclusion of your data source also provides more context to the numbers. Obviously a poll taken at a state fair is going to have a different interpretation than one conducted door-to-door by the U.S. Census.

Consider Your Audience

Finally, always consider your audience and the purpose of your graphics. For example, a chart designed for a slide presentation should be simple. You can include a bunch of details, but only the people sitting up front will see them. On the other hand, if you design a poster that's meant to be studied and examined, you can include a lot more details.

Are you working on a business report? Then don't try to create the most beautiful piece of data art the world has ever seen. Instead, create a clear and straight-to-the-point graphic. Are you using graphics in analyses? Then the graphic is just for you, and you probably don't need to spend a lot of time on aesthetics and annotation. Is your graphic meant for publication to a mass audience? Don't get too complicated, and explain any challenging concepts.

Wrapping Up

In short, start with a question, investigate your data with a critical eye, and figure out the purpose of your graphics and who they're for. This will help you design a clear graphic that's worth people's time—no matter what kind of graphic it is.

You learn how to do this in the following chapters. You learn how to handle and visualize data. You learn how to design graphics from start to finish. You then apply what you learn to your own data. Figure out what story you want to tell and design accordingly.

Chapter 2

Handling Data

Before you start working on the visual part of any visualization, you actually need data. The data is what makes a visualization interesting. If you don't have interesting data, you just end up with a forgettable graph or a pretty but useless picture. Where can you find good data? How can you access it?

When you have your data, it needs to be formatted so that you can load it into your software. Maybe you got the data as a comma-delimited text file or an Excel spreadsheet, and you need to convert it to something such as XML, or vice versa. Maybe the data you want is accessible point-by-point from a web application, but you want an entire spreadsheet.

Learn to access and process data, and your visualization skills will follow.

Gather Data

Data is the core of any visualization. Fortunately, there are a lot of places to find it. You can get it from experts in the area you're interested in, a variety of online applications, or you can gather it yourself.

Provided by Others

This route is common, especially if you're a freelance designer or work in a graphics department of a larger organization. This is a good thing a lot of the time because someone else did all the data gathering work for you, but you still need to be careful. A lot of mistakes can happen along the way before that nicely formatted spreadsheet gets into your hands.

When you share data with spreadsheets, the most common mistake to look for is typos. Are there any missing zeros? Did your client or data supplier mean six instead of five? At some point, data was read from one source and then input into Excel or a different spreadsheet program (unless a delimited text file was imported), so it's easy for an innocent typo to make its way through the vetting stage and into your hands.

You also need to check for context. You don't need to become an expert in the data's subject matter, but you should know where the original data came from, how it was collected, and what it's about. This can help you build a better graphic and tell a more complete story when you design your graphic. For example, say you're looking at poll results. When did the poll take place? Who conducted the poll? Who answered? Obviously, poll results from 1970 are going to take on a different meaning from poll results from the present day.

Finding Sources

If the data isn't directly sent to you, it's your job to go out and find it. The bad news is that, well, that's more work on your shoulders, but the good news is that's it's getting easier and easier to

find data that's relevant and machine-readable (as in, you can easily load it into software). Here's where you can start your search.

Search Engines

How do you find anything online nowadays? You Google it. This is a no-brainer, but you'd be surprised how many times people email me asking if I know where to find a particular dataset and a quick search provided relevant results. Personally, I turn to Google and occasionally look to Wolfram|Alpha, the computational search engine.

> See Wolfram|Alpha at http://wolframalpha.com. The search engine can be especially useful if you're looking for some basic statistics on a topic.

Direct from the Source

If a direct query for "data" doesn't provide anything of use, try searching for academics who specialize in the area you're interested in finding data for. Sometimes they post data on their personal sites. If not, scan their papers and studies for possible leads. You can also try emailing them, but make sure they've actually done related studies. Otherwise, you'll just be wasting everyone's time.

You can also spot sources in graphics published by news outlets such as *The New York Times*. Usually data sources are included in small print somewhere on the graphic. If it's not in the graphic, it should be mentioned in the related article. This is particularly useful when you see a graphic in the paper or online that uses data you're interested in exploring. Search for a site for the source, and the data might be available.

This won't always work because finding contacts seems to be a little easier when you email saying that you're a reporter for the so-and-so paper, but it's worth a shot.

Universities

As a graduate student, I frequently make use of the academic resources available to me, namely the library. Many libraries have amped up their technology resources and actually have some expansive data archives. A number of statistics departments also keep a list of data files, many of which are publicly accessible. Albeit, many of the datasets made available by these departments are intended for use with course labs and homework. I suggest visiting the following resources:

- Data and Story Library (DASL) (http://lib.stat.cmu.edu/DASL/)—An online library of data files and stories that illustrate the use of basic statistics methods, from Carnegie Mellon
- Berkeley Data Lab (http://sunsite3.berkeley.edu/wikis/datalab/)—Part of the University of California, Berkeley library system
- UCLA Statistics Data Sets (www.stat.ucla.edu/data/)—Some of the data that the UCLA Department of Statistics uses in their labs and assignments

General Data Applications

A growing number of general data-supplying applications are available. Some applications provide large data files that you can download for free or for a fee. Others are built with developers in mind with data accessible via Application Programming Interface (API). This lets

you use data from a service, such as Twitter, and integrate the data with your own application. Following are a few suggested resources:

- Freebase (www.freebase.com)—A community effort that mostly provides data on people, places, and things. It's like Wikipedia for data but more structured. Download data dumps or use it as a backend for your application.
- Infochimps (http://infochimps.org)—A data marketplace with free and for-sale datasets. You can also access some datasets via their API.
- Numbrary (http://numbrary.com)—Serves as a catalog for (mostly government) data on the web.
- AggData (http://aggdata.com)—Another repository of for-sale datasets, mostly focused on comprehensive lists of retail locations.
- Amazon Public Data Sets (http://aws.amazon.com/publicdatasets)—There's not a lot of growth here, but it does host some large scientific datasets.
- Wikipedia (http://wikipedia.org)—A lot of smaller datasets in the form of HTML tables on this community-run encyclopedia.

Topical Data

Outside more general data suppliers, there's no shortage of subject-specific sites offering loads of free data.

Following is a small taste of what's available for the topic of your choice.

Geography

Do you have mapping software, but no geographic data? You're in luck. Plenty of shapefiles and other geographic file types are at your disposal.

- TIGER (www.census.gov/geo/www/tiger/)—From the Census Bureau, probably the most extensive detailed data about roads, railroads, rivers, and ZIP codes you can find
- OpenStreetMap (www.openstreetmap.org/)—One of the best examples of data and community effort
- Geocommons (www.geocommons.com/)—Both data and a mapmaker
- Flickr Shapefiles (www.flickr.com/services/api/)—Geographic boundaries as defined by Flickr users

Sports

People love sports statistics, and you can find decades' worth of sports data. You can find it on *Sports Illustrated* or team organizations' sites, but you can also find more on sites dedicated to the data specifically.

- Basketball Reference (www.basketball-reference.com/)—Provides data as specific as play-by-play for NBA games.
- Baseball DataBank (http://baseball-databank.org/)—Super basic site where you can download full datasets.
- databaseFootball (www.databasefootball.com/)—Browse data for NFL games by team, player, and season.

World

Several noteworthy international organizations keep data about the world, mainly health and development indicators. It does take some sifting though, because a lot of the datasets are quite sparse. It's not easy to get standardized data across countries with varied methods.

- Global Health Facts (www.globalhealthfacts.org/)—Health-related data about countries in the world.
- UNdata (http://data.un.org/)—Aggregator of world data from a variety of sources
- World Health Organization (www.who.int/research/en/)—Again, a variety of health-related datasets such as mortality and life expectancy
- OECD Statistics (http://stats.oecd.org/)—Major source for economic indicators
- World Bank (http://data.worldbank.org/)—Data for hundreds of indicators and developer-friendly

Government and Politics

There has been a fresh emphasis on data and transparency in recent years, so many government organizations supply data, and groups such as the Sunlight Foundation encourage developers and designers to make use of it. Government organizations have been doing this for awhile, but with the launch of data.gov, much of the data is available in one place. You can also find plenty of nongovernmental sites that aim to make politicians more accountable.

- Census Bureau (www.census.gov/)—Find extensive demographics here.
- Data.gov (http://data.gov/)—Catalog for data supplied by government organizations. Still relatively new, but has a lot of sources.
- Data.gov.uk (http://data.gov.uk/)—The Data.gov equivalent for the United Kingdom.
- DataSF (http://datasf.org/)—Data specific to San Francisco.
- NYC DataMine (http://nyc.gov/data/)—Just like the above, but for New York.
- Follow the Money (www.followthemoney.org/)—Big set of tools and datasets to investigate money in state politics.
- OpenSecrets (www.opensecrets.org/)—Also provides details on government spending and lobbying.

Data Scraping

Often you can find the exact data that you need, except there's one problem. It's not all in one place or in one file. Instead it's in a bunch of HTML pages or on multiple websites. What should you do?

The straightforward, but most time-consuming method would be to visit every page and manually enter your data point of interest in a spreadsheet. If you have only a few pages, sure, no problem.

What if you have a thousand pages? That would take too long—even a hundred pages would be tedious. It would be much easier if you could automate the process, which is what *data scraping* is for. You write some code to visit a bunch of pages automatically, grab some content from that page, and store it in a database or a text file.

> **Note**
>
> Although coding is the most flexible way to scrape the data you need, you can also try tools such as Needlebase and Able2Extract PDF converter. Use is straightforward, and they can save you time.

Example: Scrape a Website

The best way to learn how to scrape data is to jump right into an example. Say you wanted to download temperature data for the past year, but you can't find a source that provides all the numbers for the right time frame or the correct city. Go to almost any weather website, and at the most, you'll usually see only temperatures for an extended 10-day forecast. That's not even close to what you want. You want actual temperatures from the past, not predictions about future weather.

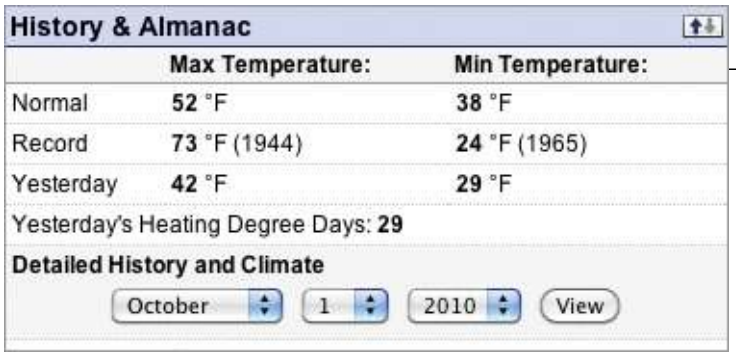
Fortunately, the Weather Underground site does provide historic temperatures; however, you can see only one day at a time.

> Visit Weather Underground at http://wunderground.com.

To make things more concrete, look up temperature in Buffalo. Go to the Weather Underground site and search for **BUF** in the search box. This should take you to the weather page for Buffalo Niagara International, which is the airport in Buffalo (see Figure 2-1).

Figure 2-1: Temperature in Buffalo, New York, according to Weather Underground

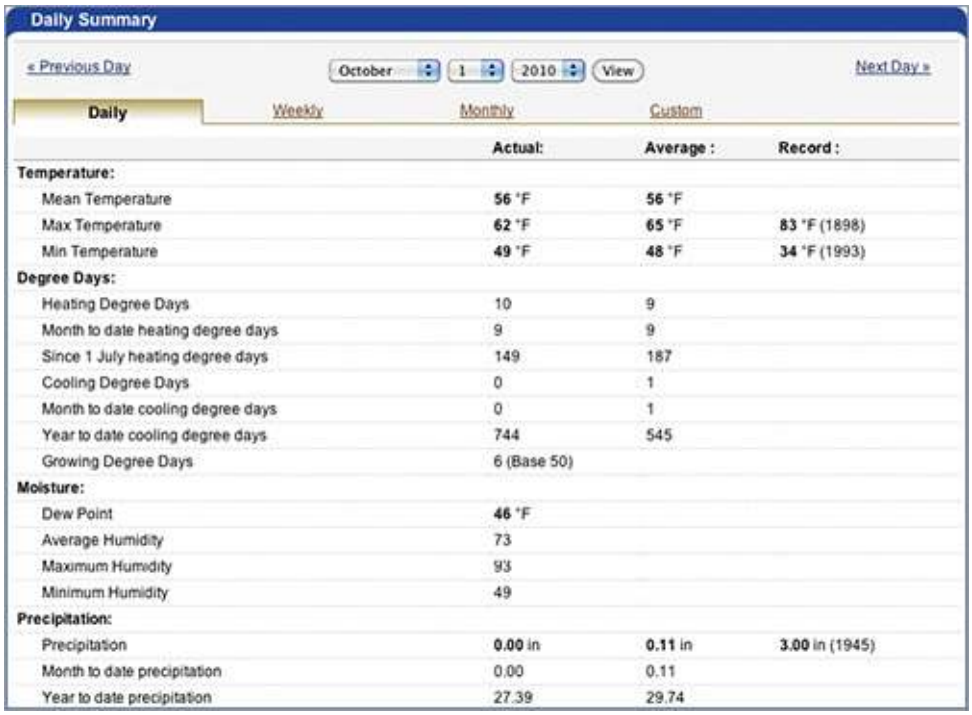
Figure 2-2: Drop-down menu to see historical data for a selected date



The top of the page provides the current temperature, a 5-day forecast, and other details about the current day. Scroll down toward the middle of the page to the History & Almanac panel, as shown in Figure 2-2. Notice the drop-down menu where you can select a specific date.

Adjust the menu to show October 1, 2010, and click the View button. This takes you to a different view that shows you details for your selected date (see Figure 2-3).

Figure 2-3: Temperature data for a single day



	Actual:	Average:	Record:
Temperature:			
Mean Temperature	56 °F	56 °F	
Max Temperature	62 °F	65 °F	83 °F (1898)
Min Temperature	49 °F	48 °F	34 °F (1993)
Degree Days:			
Heating Degree Days	10	9	
Month to date heating degree days	9	9	
Since 1 July heating degree days	149	187	
Cooling Degree Days	0	1	
Month to date cooling degree days	0	1	
Year to date cooling degree days	744	545	
Growing Degree Days	6 (Base 50)		
Moisture:			
Dew Point	46 °F		
Average Humidity	73		
Maximum Humidity	93		
Minimum Humidity	49		
Precipitation:			
Precipitation	0.00 in	0.11 in	3.00 in (1945)
Month to date precipitation	0.00	0.11	
Year to date precipitation	27.39	29.74	

There's temperature, degree days, moisture, precipitation, and plenty of other data points, but for now, all you're interested in is maximum temperature per day, which you can find in the second column, second row down. On October 1, 2010, the maximum temperature in Buffalo was 62 degrees Fahrenheit.

Getting that single value was easy enough. Now how can you get that maximum temperature value every day, during the year 2009? The easy-and-straightforward way would be to keep changing the date in the drop-down. Do that 365 times and you're done.

Wouldn't that be fun? No. You can speed up the process with a little bit of code and some know-how, and for that, turn to the Python programming language and Leonard Richardson's Python library called Beautiful Soup.

You're about to get your first taste of code in the next few paragraphs. If you have programming experience, you can go through the following relatively quickly. Don't worry if you don't have any programming experience though—I'll take you through it step-by-step. A lot of people like to keep everything within a safe click interface, but trust me. Pick up just a little bit of programming