CHAPTER TWO

Once Upon a Stacked Time Series

The Importance of Storytelling in Information Visualization

Matthias Shapiro

THE ART OF INFORMATION VISUALIZATION is something of a strange beast. Very few disciplines require such a range of skills from their practitioners. The best visualizations not only require several talents, but may require their creators to move between these different talents quickly. Furthermore, during the process of creating the final visual, the creators may realize that certain information that was discarded early on is vital to a full understanding, or that a calculation made early in the process did not produce an accurate result.

In his exceptional book *Visualizing Data* (O'Reilly), Ben Fry identifies seven stages of creating an information visualization: acquire, parse, filter, mine, represent, refine, and interact. Each stage requires a certain level of technical or artistic talent, and information visualization necessitates the close integration of these talents. When acquiring and parsing the data, the information visualization artist may be imagining how to interact with it. As he refines the representation, he may recall a step in the filtering process that excluded data that turns out to be relevant. The best visualizations tend to be dreamed up and executed by either single individuals with abilities across a wide range of disciplines, or small teams working very closely together. In these small, agile environments, the full range of talents can intersect and produce a stunning image or interactive product that can communicate a concept in a way that is more natural to human comprehension than a string of digits.

While many of the talents required for creating good information visualizations are widely recognized, there is one that is commonly overlooked in more formal settings—probably because nearly every visualization author engages in it subconsciously and because it is such a natural part of the process that is hardly seems worth mentioning. This talent is the art of storytelling.

Stories have a marvelous way of focusing our attention and helping us to discern why the data presented is important or relevant to some part of our lives. It is only inside of a context that data is meaningful, and using the data as part of a story is an excellent way of allowing the data to make a lasting impact. The most effective information visualizations will make themselves a pivotal point in a story or narrative within the viewers' (or users') minds.

Not every information visualization requires a story. Some are simply beautiful to look at and can exist merely as fine works of art. However, most visualizations have a goal or purpose and present their data in a meaningful way, in the context of some kind of story.

Question + Visual Data + Context = Story

Most visualization stories begin with some kind of question that orients the viewer to the topic and context within which the data is most meaningful. This can be done explicitly or implicitly, but the context must be clear. The question contains the premise and introduction to the story, and leads us up to the point at which the data can take over the storyline.

Many of the key parts of a story are related as part of the process of placing the visualization in a context. We frequently find the visualization context as part of an introductory text to an infographic or visualization. The context provides information that answers questions such as:

- What data are we looking at?
- In what time frame does this data exist?
- What notable events or variables influenced the data?

Consider the visualization in Figure 2-1. Assuming the user is coming to this from a place of relative ignorance, we can be confident only that he will understand that the data is mapped along a timeline and that the timeline is in some way relevant to an election. Outside of that, there is almost no valuable context to guide the user in making sense of this visualization.

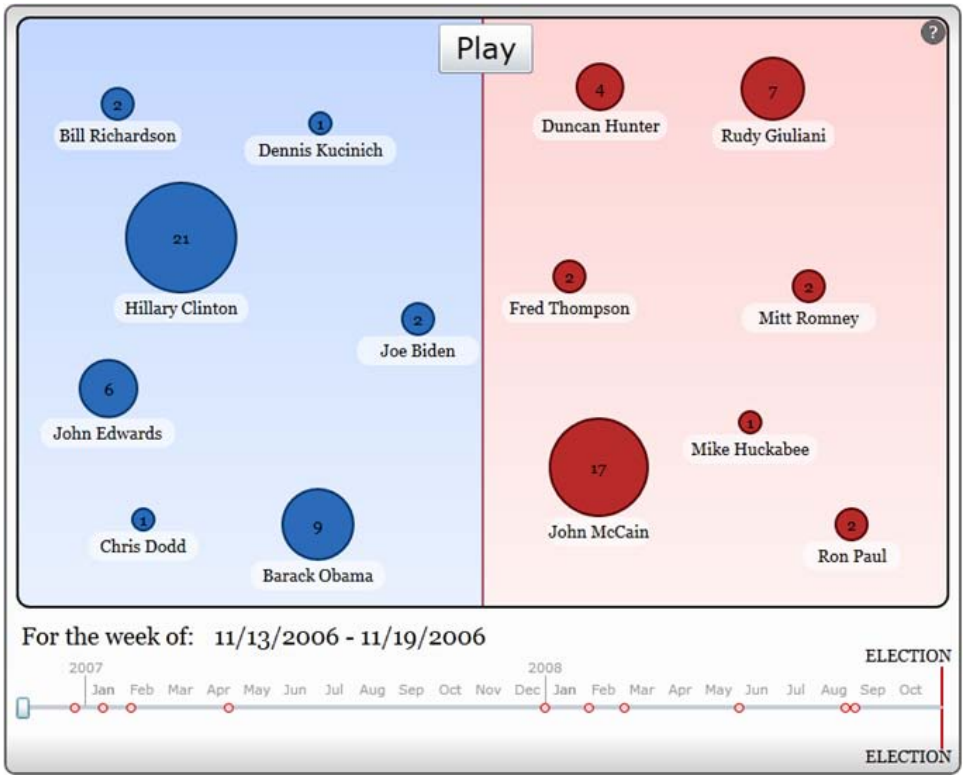

Figure 2-1. *Visualization from Designer Silverlight*[*]

If we take a step forward and assume that our user is familiar with some of the more famous names on the visualization, we can assume he will know that this visualization measures some metric related to presidential candidates in the two years preceding the 2008 U.S. presidential election.

The full context is only revealed if the user clicks on the question mark in the upper-right corner, at which point he is informed that the visualization maps the number of times each presidential candidate was mentioned in the *New York Times* in a given week. Once this information is revealed, the user can see that this is a rough map of newsworthiness as determined by the *New York Times* writers.

Returning to the questions listed previously, we now know what data we're looking at and what the time frame is. This visualization is interactive: if the user presses the "Play" button at the top, dots along the timeline pop out to reveal important events that may have influenced the data one way or another (Figure 2-2).

* See *http://tr.im/I2Gb*.

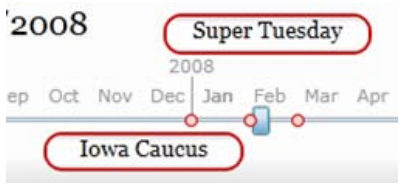

Figure 2-2. *The visual draws attention to important events that may have influenced the perceived newsworthiness of the candidates*

In addition to these cues, the user can draw on his knowledge of the presidential race to supply additional context to the data. He may recall that in the Democratic party there was a knock-down, drag-out primary contest between Hillary Clinton and Barack Obama, which is reflected in the fact that they maintained a high level of newsworthiness into April and May of 2008, while John McCain (who secured the Republican nomination in early March) lagged behind them both.

From the question "How often did the *New York Times* mention each candidate during the course of the 2008 presidential campaign?," a story emerges. This visualization provides an engaging visual component to that story and helps the user relive the drama of the two-year presidential campaign in the space of a minute.

Steps for Creating an Effective Visualization

When creating an information visualization, I typically walk through the following key steps:

1. Formulate the question.
2. Gather the data.
3. Apply a visual representation.

Formulate the Question

Asking the question that drives the story you're trying to tell is not necessarily a task that must be done at the beginning of the visualization journey. Don't feel bad if you start digging into the data before you have a finalized question in your head. Often, it is not until we have a good understanding of the data that we know how to ask a good question about it. However, asking a question (or at least keeping a question or set of questions in mind) can be useful when gathering and filtering the necessary data.

You may want to start with a topic to focus your data search and refine your question as you gather more data. For example, let's say we want to communicate that carrying out the U.S. Census is an enormous task. This is a good topic to start us out in our data search because it is broad enough that there are many pieces of data that can help give context to this idea. We could find the relevant data and create a visualization based on:

- The number of surveys filled out
- The number of pencils used
- The number of miles census workers walked

My favorite U.S. Census–related data to watch is the number of federal employees over time. Statistics show a spike of 200,000–300,000 federal employees between March and July of a census year. These employment figures then drop off as the census process completes.

The specific question that we ultimately ask will have a heavy impact on the final representation of the visual. For example, we might ask "How much paper does it take to record all the information necessary for a census?" and show sheets of paper covering a small city as a representation of the surveys, or we might ask "How many people does it take to count everyone in the country?" and use icons of people to represent the spike in federal employment figures at census time. These questions both relate to the original topic of the scope of the U.S. Census, but they draw from different sets of data and result in drastically different visuals.

When asking a question for the purposes of creating an information visualization, we should focus on questions that are as data-centric as possible. Questions that begin with "where," "when," "how much," or "how often" are generally good starting points: they allow us to focus our search for data within a specific set of parameters, so we're more likely to find data that lends itself to being mapped visually.

Be especially careful if you find your question starts with "why." This is a good sign that you are moving from a more formal portrayal of data into data analysis.

Gather the Data

Finding exactly the data you want can be a difficult task. Often, instead of trying to gather your own data, you're better off taking data that is already available and trying to find a way to portray it. That is, it may be better to start (as mentioned earlier) with a dataset and construct a question as you find patterns in the data. If you're creating a data visualization for a purpose other than as a hobby or out of pure curiosity, it is likely that you already have a dataset to work from. However, there are still several datasets available that may inspire or inform some aspect of your work.

There are many good places to start looking at data. One of the largest and most diverse repositories can be found at Data.gov (*http://www.data.gov*). This site houses an enormous collection of data, from migratory patterns of birds to patent bibliographies to Treasury rate statistics and federal budget data.

Other excellent sources of data include:

- The Census Bureau (*http://www.census.gov*) for a wide variety of demographic and geographic data

- The Bureau of Labor Statistics (*http://www.bls.gov*) for extensive data on employment in the United States (click on the "Databases and Tables" tab and scroll down to the Historical News Release Tables for the easiest access to the data)
- The New York Times APIs (*http://developer.nytimes.com*) for easy API access to huge sets of data including congressional votes, bestseller lists, article searches, movie reviews, real estate listings and sales in New York City, and more

Once you have the raw data, you may want parse it, organize it, group it, or otherwise alter it so that you can identify patterns or extract the specific information you wish to portray. This process is known as "data munging" and is usually an ad hoc attempt to "play around" with the data until interesting patterns emerge. If this process sounds a little opaque or nonspecific, don't worry; we'll walk through an example of data munging in the hands-on tutorial in the next section.

Apply a Visual Representation

Now that we have the data, we come to the task of deciding how to portray it. This means making decisions about what kind of visual representation of the data will aid viewers in understanding it.

A visual representation is some kind of visual dimension that can change in correspondence to the data. For example: an *XY* graph is a simple visual presentation that maps an *x*, *y* data point in a two-dimensional plane. Map enough points, and an obvious visual pattern may emerge even if there is no immediately identifiable pattern in the raw data itself.

Let's take a look at the most commonly used visual representations.

Size

Size is probably the most commonly used visual representation, and for good reason. When differentiating between two objects, we can judge very quickly between sizes. Moreover, using size helps cut through the fog of comparing two unfamiliar numbers. It is one thing to hear or read that methadone is the most lethal recreational drug in the UK and quite another to see that information in the context of deaths caused by other dangerous drugs, as shown in Figure 2-3.

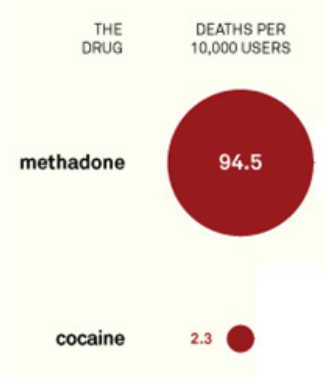

Figure 2-3. *From David McCandless's information visualization "World's Deadliest Drugs"*

While size is an extremely useful and intuitive representation, it is also often overused. Many poorly constructed graphs misinform and confuse simply because their creators wanted to visualize some data, but knew of only one way to visually present it.

Color

Color is a fantastic representation method for enormous sets of data. We can identify many gradations and shades of color and can see differences in a high resolution. This makes color a natural choice for representing big-picture trends, like what we might see in weather maps. For this reason, it is commonly used for identifying patterns and anomalies in large datasets.

Figure 2-4 is a zoomed-out view of a set of data about stocks over the course of just over three months.

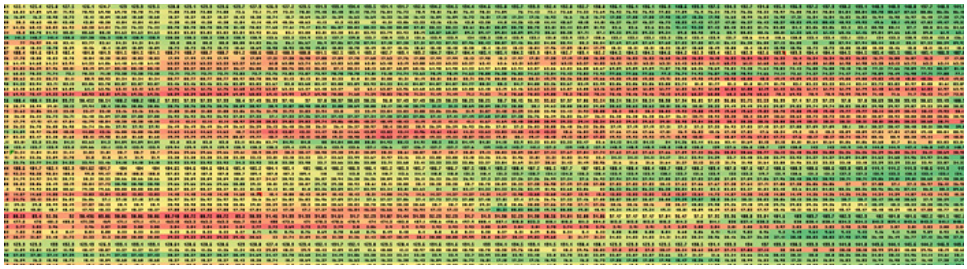

Figure 2-4. *The 30 most watched Motley Fool CAPS stocks tracked over several months and visualized using a red-to-green color scale*

Even though the type is far too small to read, we can easily recognize rows that show positive or negative growth. We can also make an overall assessment of the trends in the data with very little intellectual effort expended.

Color is less useful for smaller datasets or data that is differentiated by small ranges. If there are not stark ranges in the data, it can be difficult for even a trained eye to spot important differences.

As an example, let's assume a dataset with a range between 1 and 100 and a color scheme that ranges from red (representing 1) to yellow (50) to green (100). In such a scheme, consider the 10-point difference in Figure 2-5. As you can see, the difference is subtle and may not be easily distinguishable to many viewers.

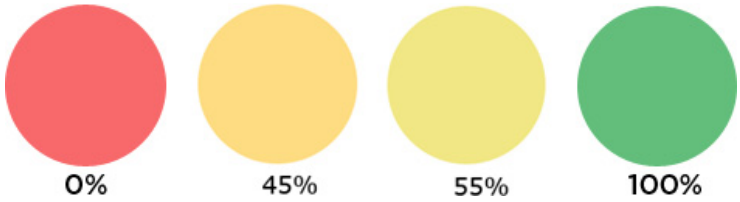

Figure 2-5. *Color image representing the difference between 45% and 55% in a color visualization*

If you're creating a visualization in which it is important for viewers to be able to distinguish between data points at 45% and 55%, you may need to alter the points at which the colors change or steer away from using color as your primary representation method.

A word should also be put in for those who suffer from colorblindness, which affects nearly 1 out of 10 individuals. If you need your visualization to reach the largest possible audience, you may want to consider using ranges like black-to-white instead of green-to-red. For more information about design and colorblindness, consider visiting We Are Colorblind (*http://wearecolorblind.com*), a website devoted to designing in a way that is accessible to the colorblind.

Location

A location representation method attaches data to some kind of map or visual element corresponding to a real or virtual place. An everyday example of a locative visualization is when we are presented with a simple outline of an airplane or a theater in order to choose a seat.

In Figure 2-6, we see the county-by-county crime rates for 1996 and 2008 rendered onto a map of Florida.

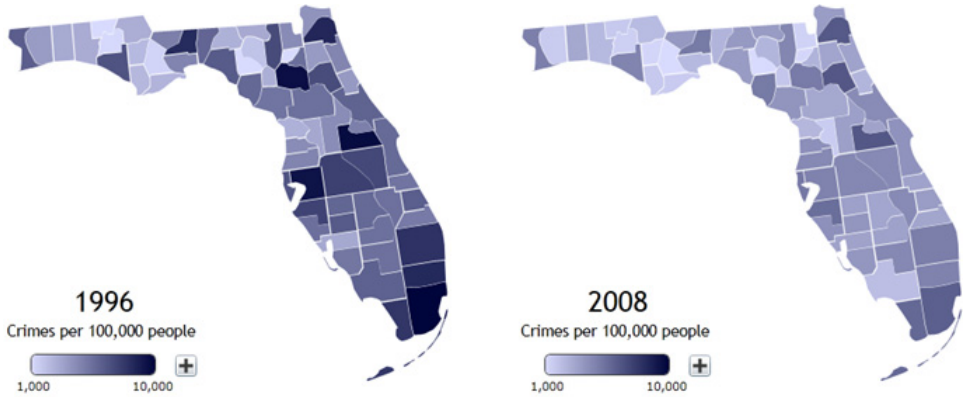

Figure 2-6. *Florida county map shaded to indicate crime rate by county*

Location presentation methods are especially valuable when the audience has some familiarity with the location being portrayed. Such familiarity allows the audience members to project their personal contexts onto the visualization and draw conclusions based on their personal experience with the area.

Networks

A network presentation shows binary connections between data points and can be helpful in viewing the relationships between those data points. A number of online network visualizations have sprung up that allow people to see maps of their friends on Facebook or their followers on Twitter.

Figure 2-7 shows a network visualization of my Facebook friends and how many of them have "friended" one another.

Through this network mapping, we can perceive at a glance the different social networks to which I belong (or belonged). Furthermore, the density of the groups corresponds fairly well to the social intimacy of those groups.

One thing to keep in mind with network visualizations is that if they are not carefully constructed, the thousands of data points may just turn into a visually messy glob of connections that isn't helpful in increasing our understanding of how those connections are meaningful.

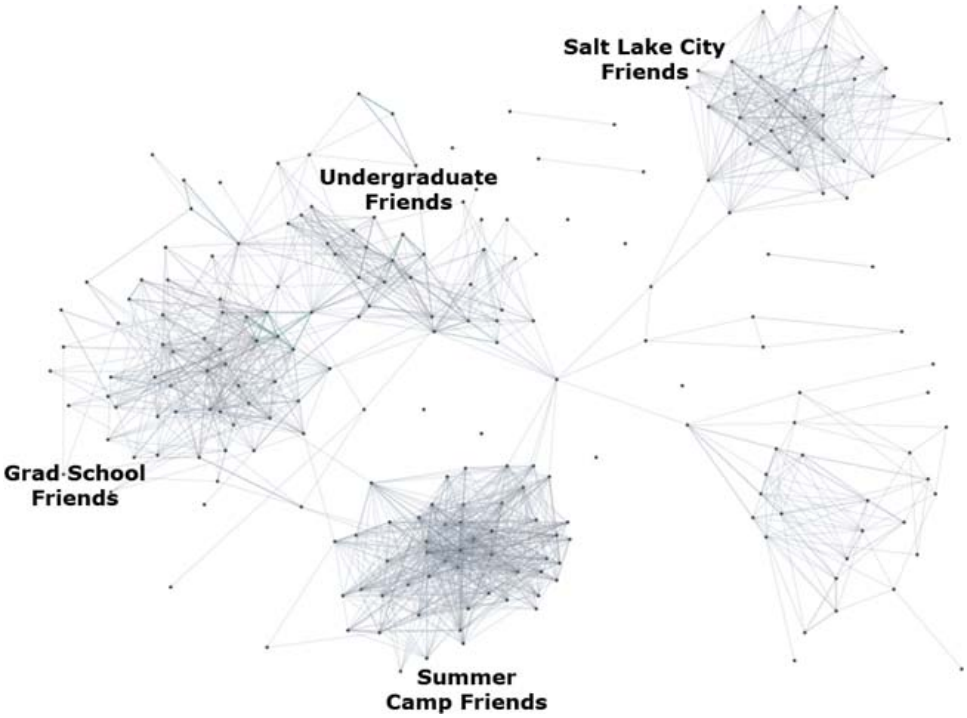

Figure 2-7. *Nexus rendering of a network visualization of my Facebook friends*

Time

Data that changes over time (stock quotes, poll results, etc.) is traditionally portrayed along a timeline. In recent years, though, software with animation capabilities has allowed us to portray such data in a different manner. Animations like the *New York Times*'s "Twitter Chatter During the Super Bowl"* (shown in Figure 2-8) compress a longer period of time so that we can watch the data change in an accelerated environment.

Pressing the "Play" button in the top-left corner starts the animation, and the most popular words used in Super Bowl–related tweets across the country grow and shrink according to their frequency of use through the course of the game.

This visualization gives users a series of helpful contextual clues along the timeline indicating when major events happened in the game. By doing this, the authors provide valuable context and relieve the users from the task of remembering how the game played out. Instead, they can focus on the words being used in tweets across the country and let the application alert them when a key event is driving the data.

* See *http://www.nytimes.com/interactive/2009/02/02/sports/20090202_superbowl_twitter.html*.

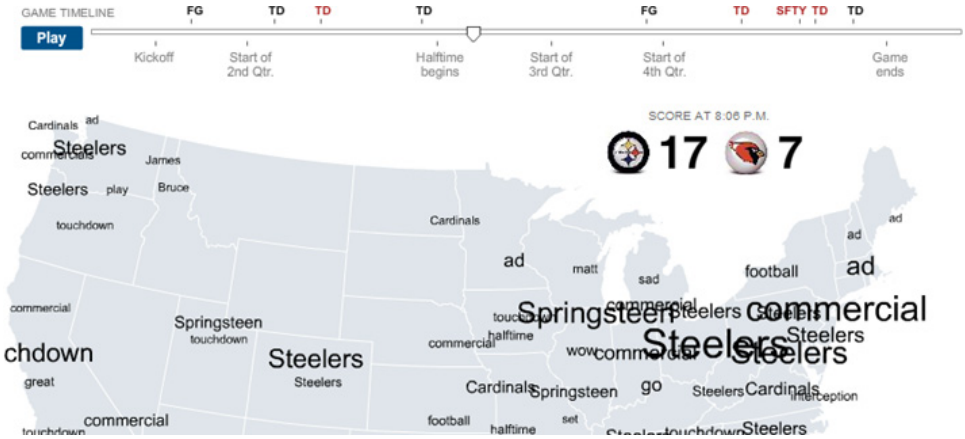

Figure 2-8. *New York Times visualization of commonly used words in 2009 Super Bowl–related tweets*

Using multiple visual presentation methods

Many excellent information visualizations use more than one of these visual presentation methods to give a full picture of the data. In the online application *NameVoyager* (*http://www.babynamewizard.com/voyager*), users can type in the first few letters of a name and see a history of how many people have given their child a name beginning with those letters (Figure 2-9).

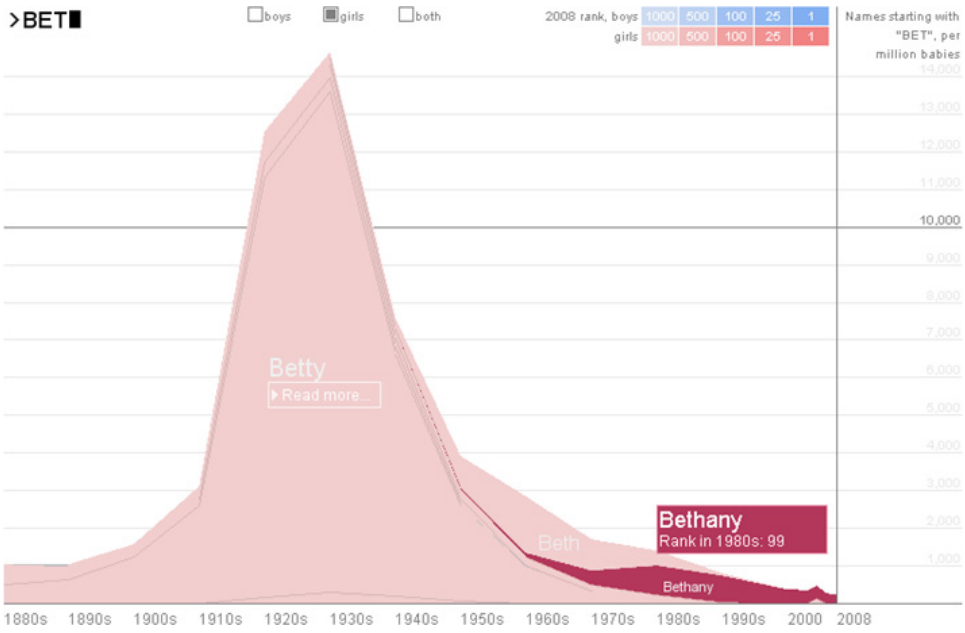

Figure 2-9. *The NameVoyager baby name explorer charts name frequency by year*

Here, two visual dimensions are presented. The first is time: we see the frequency with which names beginning with the entered letters were used represented along a timeline. The second is size: shaded areas on the graph indicate how many children were given certain names in certain years.

This particular type of graph is called a *stacked time series* and is a fairly standard way of visualizing several pieces of information in a combined but separate manner.

Hands-on Visualization Creation

Now that we've covered the basics of information visualization in a general manner, let's walk through the process of building a visualization. We'll create a static visualization, commonly referred to as an *infographic*.

To do this walkthrough, we will need the following tools:

- Microsoft Excel (or Google Documents in a pinch)
- Adobe Photoshop (GIMP, a free image-manipulation program, will also work)

In order to replicate the process as closely as possible, I'll walk through the discovery process in the order in which it actually happened rather than following the "Question-Data-Presentation" method described earlier.

Data Tasks

When constructing this tutorial visualization, I started out messing around with the data and formulated the question as the shape of the information became clear. Because the process of sifting through data is often very ad hoc, I'll simply describe my discovery in general terms. We'll walk through the details later in this section.

Gathering the data

I decided to use easily accessible, publicly available data for this tutorial, so I started looking at a number of pieces of data collected by the U.S. government and placed online in the interest of transparency. I settled on data about vehicles traded in and purchased via the Car Allowance Rebate System (CARS), commonly referred to as the "Cash for Clunkers" program. The data I used is available in two separate Excel files at *http://www.cars.gov/carsreport*. It is also available in CSV or MDB format.

Sorting the data: The discovery version

When we're done with this visualization, we want to feel like it provides some kind of insight into the individual transactions that make up this dataset. We can imagine someone driving in a beat-up clunker thinking to herself that she will soon be able to rid herself of her old, inefficient vehicle and replace it with a beautiful new one.

What kind of vehicle is she driving? Is she looking to replace it with something similar but newer and more efficient (an "old sedan to new sedan" trade)? Or does she want to swap her vehicle for something totally different (a trade more along the lines of "SUV for two-door coupe")?

The data we're looking at is the result of over 650,000 individual stories that each required motivation, drive, time, and effort to report. We won't be able to tease out those individual stories from the data, but our visualization will help tell a larger story about those people's choices. Our goal is to find a way to tell a story that is interesting and new to our users/viewers.

Here are the steps I took in sorting and filtering through the data as I was trying to discover that story.

After downloading the dataset, I started looking at the trade-in data and tried to group it in many different ways. Grouping it by car model seemed interesting at first, but this was somewhat tedious because the vehicles are grouped by engine and transmission type, so the same model might have several different entries.

However, in the process of looking at the vehicles by model, I was struck by the fact that several makes had a fairly high number of trade-ins. I became curious to see if people were more eager to trade one make of vehicle over another, so I began sorting the vehicles by make.

Warning: Asking questions like "are people eager to trade in one make over another?" is a dangerous thing to do when creating a visualization. The data can tell us a large number of things, but it is rare that data will give us good information on things that are as complex as human motivation. It is one thing to portray the data as it is and another thing to interpret what the data means. It would be a mistake to state as a part of your visualization that, because more Ford vehicles were traded in than any other make, people were eager to get rid of Fords. Such a statement would dismiss dozens of important variables, including things like market share, type of vehicles sold, Ford's position in large vehicle sales, age of the vehicles, etc. It is a good rule of thumb to restrict a visualization to stating things that can be seen from the data alone and allow the users or viewers to draw their own conclusions.

With all of that said, asking these kinds of questions internally can be an effective driver for discovery, so don't shy away from asking them at this early stage—just shy away from answering them in the final visual.

I began sorting vehicles by make and tallying up the sums for the trade-in vehicles, and I thought it would be interesting to see a comparison of the makes of the trade-ins (Honda, Toyota, GM, Ford, Chrysler) versus the makes of the new vehicles purchased. As I started collecting that data, it became clear that there were so many vehicle makes,

it would be difficult to clearly portray that many separate data points. As a result, I started trying to group by "parent make"—i.e., grouping together vehicle makes by the companies that owned those makes. For example, Lexus is a division of Toyota, so I grouped Lexus and Toyota trade-ins together under the parent company, Toyota.

Eventually, I decided that the most compelling portrayal of the information would be to group the makes together under the parent *country*. This approach has the benefit of reducing the number of data points to about a dozen, as well as grouping the information in a way that isn't immediately apparent in the data. By doing this, we're able to get a new and fresh look at the data.

Sorting the data: The technical version

Now that we've walked through the thought process, let's walk through replicating that process in the files.

If you download the Excel files, you can open them up and see that the data is arranged first by vehicle category (with trucks first and cars second) and then alphabetically by vehicle make (Acura, Audi, BMW, and so on). In order to sort the data for our purposes, the easiest thing to do will be to categorize the data by vehicle make. Later, we will determine which makes correspond to the various countries in which the parent companies are based.

To sort the data in Excel, simply select the New_Vehicle_Make column in the *new-vehicles* file or the Trade_in_make column in the *trade-in-vehicles* file and select "Sort & Filter->Sort A to Z." If Excel asks you if you want to expand the selection, accept that option.

You can add together all the cars purchased or traded for a particular make by entering `=SUM(` and using the mouse to select all the cells in the Count column for a particular make. As a method of checking your first attempt, add up all the Acura purchases. The result should be 991. Gather sums for all the makes and, if it helps you to look at the data, move the results to another page.

This is the perfect time to play around with the data if you're so inclined. Try to figure out which cars sold the best, or which year's models were traded in the most frequently. Even in a dataset as small as this one, there are dozens of interesting questions to ask. One of them might pop out at you and inspire a new and compelling visualization. At the very least, this is an excellent opportunity to practice looking at data.

There are many ways to sort this kind of data. It might be more efficient (and would certainly be impressive) to write a script or small program that walks through the CSV file and pulls the data into a summary file that is easy to look at. The reason for using Excel in this example was to try to help people who are not familiar with programming engage with the data and participate in creating visualizations.

Formulating the Question

At this point in the process, we should have a firm enough grasp of what we want to do that we can formulate a solid question for this visualization. Our question is, "In the 'Cash for Clunkers' program, what proportion of vehicles were purchased from manufacturers based in which countries?"

Within the context of this question, we can choose to establish a number of relevant pieces of information as an appropriate setup for the visualization, keeping in mind that our target audience may not be intimately familiar with the topic. Here are a few items that will help contextualize the data:

- The program cost $2,850,162,500 and provided money for 677,081 vehicle purchases.
- For each vehicle that was purchased, a vehicle was traded in and scrapped.
- The program ran from July 1, 2009 until August 24, 2009.
- Vehicles eligible for trade-in had to get less than 18 miles per gallon (MPG).
- Vehicles eligible for purchase had to get more than 22 MPG.

For the purposes of this visualization, we're most interested in the fact that there was a correspondence between vehicles purchased and vehicles scrapped. This creates an interesting balance (and hence a certain kind of drama) between the kinds of vehicles people wanted to get rid of and the vehicles they wanted to purchase. As we put together the data and visualization, we'll keep this balance in mind and orient the visuals accordingly.

With the question in hand, we have a solid basis for manipulating the data further by grouping and sorting it as guided by our question.

Grouping the data

This step takes a little bit of research. In order to group the makes by country, we need to find out which vehicle makes correspond to which companies. There are over 50 makes represented in these two files, so the research could take some time. In this task, Wikipedia is your friend since it will provide quick answers regarding the ownership of various vehicle makes (for example, Chrysler owns or owned six makes that are represented in this dataset) as well as the countries in which they are headquartered.

I've provided a helpful table containing this data, to save you time (Table 2-1).

Table 2-1. *Vehicles grouped by make, company, and country*

Make	Owned by	Country	Make	Owned by	Country
Jaguar	Tata	England	Hyundai	Hyundai	South Korea
Land Rover	Tata	England	Kia	Hyundai	South Korea
BMW	BMW	Germany	Volvo	Volvo	Sweden
MINI	BMW	Germany	Saab		Sweden
Mercedes-Benz	Daimler	Germany	American Motor	Chrysler	U.S.
smart	Daimler	Germany	Chrysler	Chrysler	U.S.
Audi	Volkswagen	Germany	Dodge	Chrysler	U.S.
Porsche	Volkswagen	Germany	Eagle	Chrysler	U.S.
Volkswagen	Volkswagen	Germany	Jeep	Chrysler	U.S.
Acura	Honda	Japan	Plymouth	Chrysler	U.S.
Honda	Honda	Japan	Ford	Ford	U.S.
Isuzu	Isuzu	Japan	Lincoln	Ford	U.S.
Mazda	Mazda	Japan	Mercury	Ford	U.S.
Mitsubishi	Mitsubishi	Japan	Merkur	Ford	U.S.
Infiniti	Nissan	Japan	Buick	GM	U.S.
Nissan	Nissan	Japan	Cadillac	GM	U.S.
Subaru	Subaru	Japan	Chevrolet	GM	U.S.
Suzuki	Suzuki	Japan	GMC	GM	U.S.
Lexus	Toyota	Japan	Hummer	GM	U.S.
Scion	Toyota	Japan	Oldsmobile	GM	U.S.
Toyota	Toyota	Japan	Pontiac	GM	U.S.
			Saturn	GM	U.S.

Keep in mind, however, that grouping the makes this way raises some questions about the data that we'll need to answer before we continue. For example, Jaguar is a quintessentially British company with its headquarters in England. Nevertheless, it is owned by the Indian company Tata Motors. Should we categorize Jaguar as an English car or an Indian one?

The "correct" method of dealing with these kinds of questions is largely a matter of personal preference. The important thing to remember is to maintain consistency in the representation of this decision and to indicate to the viewer that you have made the decision one way or another. Usually, a footnote at the corner of the visualization is sufficient.

Applying the Visual Presentation

At this point, we should have all of our data in exactly the format we want: vehicles traded or purchased, organized by country. It's time to choose our visual presentation of the data.

We'll be representing two dimensions of information in this visualization. The first is the quantity of cars organized by country, and the second is a visual differentiation between cars purchased and cars traded in. The differentiation between purchased vehicles and "clunked" vehicles is an "either-or" differentiation, so there won't be any gradations in the information, which will simplify the presentation. To differentiate between vehicles purchased and traded, we can use a simple color method: red to represent "traded" and green to represent "purchased."

Since we're dealing with a few points of data with enormous variation, it makes the most sense to use size to represent the information. This presentation choice will call attention to the scope of this variation in an intuitive and compelling way. The easiest implementation will be to use circles or bars of varying sizes to represent the numbers of trades and purchases.

A note about area and circles

If we're using circles to represent the data, we need to remember that we're going to be varying the area, not the radius or diameter, of the circle. If we take the number of U.S. vehicles purchased (575,073) and choose to represent it with a radius of 50 pixels, we will use the following equation in Excel to determine the size of each of the other circles:

```
SQRT((US_Baseline_Radius^2 * Target_Vehicles)/US_Vehicles)
```

I'm taking the time to point this out because this is probably one of the most common mistakes when creating information visualizations with circles or with area in general. Scaling a circle by linearly increasing the radius or diameter will result in exponential increases and decreases of the area of the circle, as shown in Figure 2-11; the correct relationship is shown in Figure 2-10.

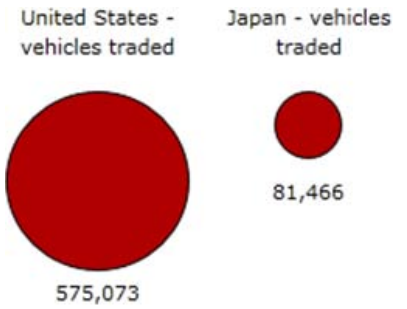

Figure 2-10. *Correct (scaling the area)*

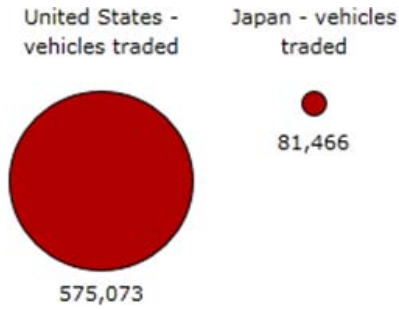

Figure 2-11. *Incorrect (scaling the radius)*

Having said all of that, we're not going to use circles. Don't worry, I have a good reason.

Presenting the data with country maps

Since our information story centers on countries, we're going to use shape maps of the various countries and size those maps appropriately. This provides a couple of valuable additions to our visualization.

First, using the shapes of the countries will give this project a *visual hook*. If their home countries are on the list, the viewers will be able to pick them out immediately and it will draw their attention. Along these same lines, we will be able to hook into any emotions our users may have concerning their home countries or any other countries with which they are familiar. A hook like this makes it more likely that the audience will remember or recommend the visualization.

Second, using country shapes instead of circles will enable the visualization to communicate at a number of different sizes. Even at thumbnail size, the shape of a country is so recognizable that the users will know that the visualization has something to do with different countries. A set of circles reduced down to thumbnail size just looks like a set of circles.

Third, if we used only circles or bars, we would be reliant on text to convey the names of the countries in the visualization. This isn't necessarily bad, but comprehension time would be increased, as the users would have to read the text before they could understand the visualization. This would increase the risk of reducing the immediate impact of the visualization.

Finally, the audience is accustomed to seeing these different countries in the context of a world map where the relative sizes are always the same. Taking these familiar shapes out of that context and placing them in a context where South Korea is larger than Germany or the United States is smaller than Japan creates interest by violating expectations. Think of it as a "twist" in the plot of the story.

Having decided that we should use countries instead of circles, we need to find visual representations of the countries on our list. Our best bet on that count is to search for a country name along with the *.svg* file extension. SVG stands for scalable vector graphics and is an open standard for vector images maintained by the World Wide Web Consortium (W3C). It is a popular vector image standard, particularly for free images and maps, and many vector manipulation applications support it.

Wikimedia Commons (*http://commons.wikimedia.org*) has a number of free, high-quality maps in vector format. These maps scale very well and are excellent for this kind of project. Some of the countries that are hard to find can also be pulled from vector maps of the world that are available on Wikimedia Commons. These files can be opened as editable vector files in Adobe Illustrator or Inkscape (*http://www.inkscape.org*) or as bitmaps in GIMP. From Illustrator, the vector objects can be copied and pasted directly into Photoshop.

In the interest of simplicity, we'll display only countries responsible for a certain minimum (1,000+ vehicles) of either the traded-in or purchased cars. This means we should have maps for the United States, Japan, South Korea, Germany, Sweden, and the United Kingdom.

Once we have images of the countries we want, we're ready to size them for the final visualization.

Building the Visual

Having moved the visuals into an image-manipulation program, we need to size them so that they appropriately represent the proportions of vehicles traded in and purchased.

My methodology for this is to take the largest piece of data (in this case, it is the number of U.S.-made vehicles that were traded in: 575,073) and scale it to a size that fits comfortably on the canvas of the infographic. This kind of anchor shape is just a practical way of making sure that none of the graphic elements becomes too large for elegant display. This piece of data becomes the anchor against which we will scale all the other elements.

Once we have the size of the anchor shape, we need to calculate how many pixels are in it. There is a trick available in Photoshop and GIMP that lets us easily count the pixels we have selected in a particular layer. Both applications have a window called "Histogram" that displays the number of pixels that are currently selected. Using this tool, we can determine the number of pixels in the anchor shape and calculate how many pixels our other shapes need to be using the following formula:

Target_Size = Target_Number * Anchor_ Size / Anchor_Number

For example, 81,466 Japanese vehicles were traded in. If we size the U.S. map so that it comprises 25,000 pixels, the equation for determining how large to make the map of Japan would be:

Japan_Size = 81,466 * 25,000 / 575,073 = 3,542 pixels

I generally use Excel to make these calculations so that they are easy to save, double-check, and replicate.

Using the Histogram trick, we can resize the irregular shapes of the target countries and scale them until they contain the number of pixels appropriate for the corresponding data point visualization.

I decided to arrange the countries along a vertical axis in order to accommodate the medium in which this visualization will be viewed (a page in this book). This approach also gives symmetry to the color elements and reinforces the green/red, bought/clunked dichotomy of the data.

We now have the core of our visualization done. Providing some context in an introductory blurb and adding a footnote about our decision regarding the country of origin for Jaguars and Land Rovers gives us the result shown in Figure 2-12.

This visualization now meets our criteria. It sets up the story with an introduction at the top, it provides a compelling layout that draws the viewers' attention, and it is instantly understandable. We've set up the "bought/clunked" dichotomy with color-coding and reinforced it with symmetrical physical placement (important if we want individuals who are colorblind to be able to understand our infographic). Our visualization tells what we hope is a compelling story in the minds of our viewers.

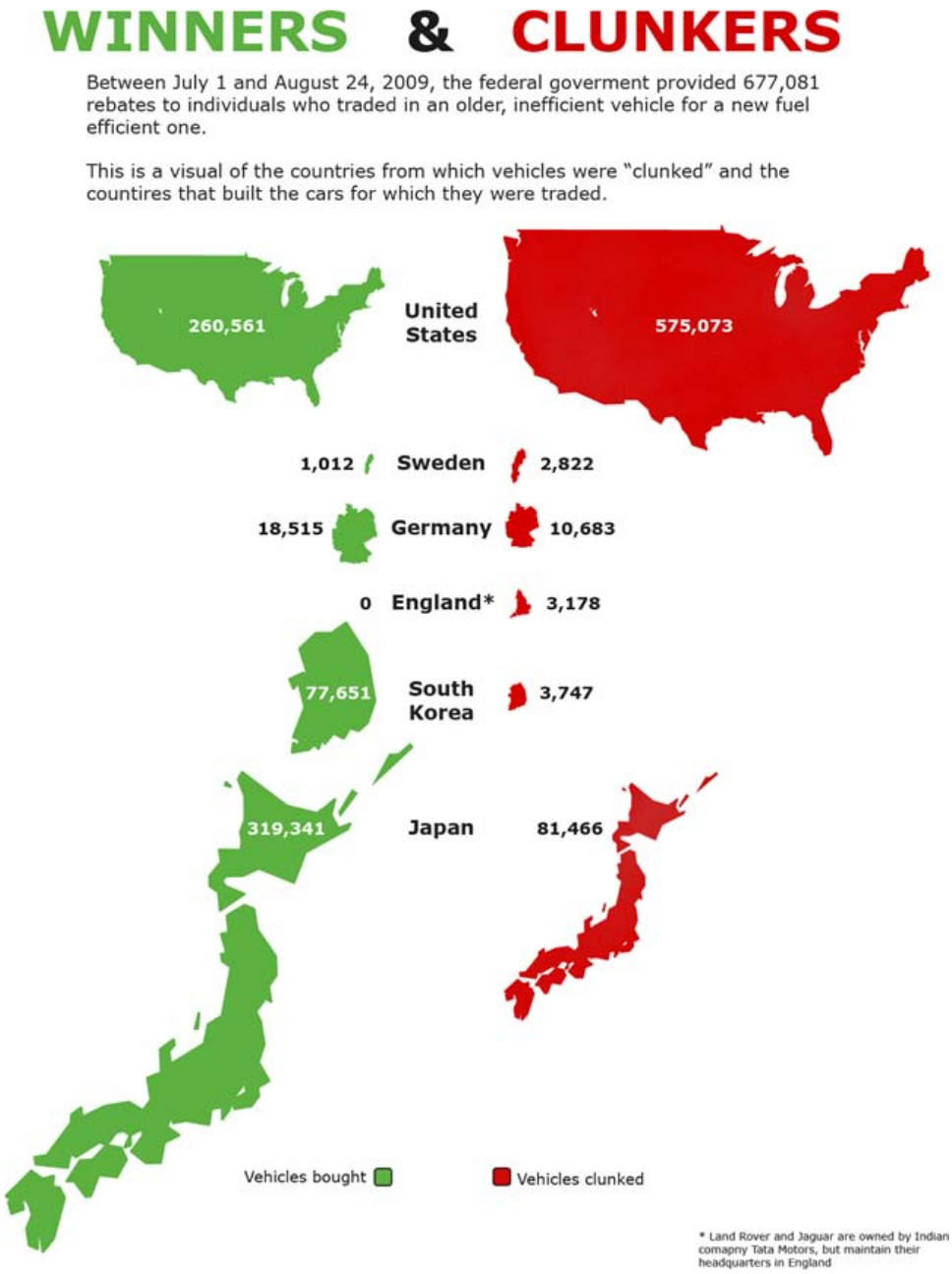

Figure 2-12. *Final visualization*

Conclusion

This tutorial has touched on only a small subset of the skills that can be used to create effective visualizations. A deeper foundation in fields like color theory, typography, computational data mining, and programming, as well as a background in the data subject, will all be valuable aids in creating compelling visualizations.

Despite the variety of fields that inform the visualization creation process, they are unified by the fact that every visualization is part of some kind of story. Even the simplest bar graph displaying a company's earnings data is drawing from information that is more memorable and more valuable within the larger context (perhaps a change in management style). It is these contexts and the stories that we associate with them that give visualizations their long-lasting impact and power.