Pages from
Data Visualization: A Successful Design Process
Andy Kirk, Packt Publishing, December 2012

An example of finding and telling stories

Before we move on, to help embed the understanding of data familiarization, visual analysis and the difference between finding stories and telling stories, let's work through a basic example.

Take the following sample table of data. The subject matter is the Olympic games and specifically the total medals won by the top eight participating nations over five recent events. The selection of the top eight is based on them being the top ranked countries at the Beijing Olympics in 2008.

Suppose you were briefed to unearth some key stories around Olympics medal winning trends in recent years, how would you go about it?

Total medals won in the Summer Olympics

Country	2008	2004	2000	1996	1992
United States of America	110	103	92	101	108
People's Republic of China	100	63	59	50	54
Russian Federation	72	92	88	63	112*
Great Britain	47	30	28	15	20
Australia	46	49	58	41	27
Germany	41	49	56	65	82
France	40	33	38	37	29
Republic of Korea	31	30	28	27	29
ALL	951	929	925	842	815

** When part of former Soviet Union. Data from http://www.databaseolympics.com/index.htm*

Let's start by just scanning the data with our eyes to find anything that stands out.

The main data issue appears to be that the Russian Federation medals total for 1992 was actually when it was known as the Soviet Union. It is noticeably higher than for all the other Olympic events, due to the contributions of additional member states that then made up the Soviet Union but who are now independent countries competing in their own right. As it will be hard to unpick this value to isolate just those athletes who would now be considered part of the Russian Federation, it will be sensible to just ignore this value from our analysis. Otherwise, it will skew our interpretations.

We can see that the event order goes from left to right in reverse chronological order and the vertical sorting is organized by the most successful nations as at 2008. In addition to the medal winning totals for the selected countries, we also have the aggregate of all medals across all countries.

We now continue our examination by noting some of the dataset's descriptive and statistical properties to develop an increased level of familiarity:

- Two variables: Country and event year
- Country is a categorical nominal variable with nine values (each country and the aggregate)
- Event year is a quantitative (interval-scale) variable with five values
- The maximum country medal count value is 110 medals, the minimum is 15
- The maximum aggregate value is **951** and the minimum is 815 (but that includes the Russian Federation contribution)
- Each event year is spaced 4 years apart
- The longest country name is People's Republic of China, the shortest is France

This gives us a sense of the physicality of the data and the potential influencing attributes that might shape our visualization architecture.

What other data preparation tasks might we undertake?

We have no real transformation activities to undertake in terms of addressing data quality aside from already deciding to ignore the Russian Federation total.

For transforming the data for its use in analysis we may decide to create some calculations to show the percentage of medals won out of each event total. You may also decide to abbreviate some of the county values to potentially help accommodate the space required for labeling.

We also need to consider data consolidation. For the purpose of this demonstration, we are going to stick to our original dataset on its own but there could be many different options to enhance and contextualize this subject matter, including the following:

- The details behind the medal totals of how many golds, silvers, and bronzes each country has won
- The full dataset of medal statistics for all the other countries who have competed, not just the recent top eight
- The full dataset of medal statistics for every Olympic games
- The number of competitors who were taking part in the games for each country, in order to understand the percentage of success of each team
- The split of performances between the different sporting events

- Population figures to contextualize the achievements, maybe even sporting participation figures if they were recorded
- Historical milestones of socio-political and geo-political issues to help us appreciate the status of the different countries at these key points in time
- You might look to bolster the ingredients of your visualization design resources with national flags' image files or URL links to national Olympic associations

Whether we could obtain these additional data items is another matter and they may not even help with our stories. But it is always good to let your imagination roam and explore ideas for content that could really enhance your work.

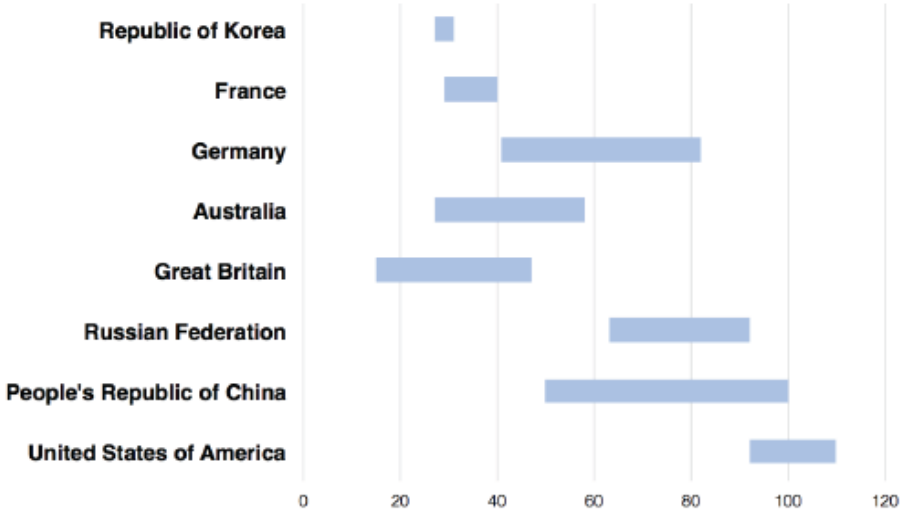
Our data is now in good shape. Next up, we look to develop our editorial focus, specifically considering the following:

- What initial sparks of curiosity crossed our minds when we were given the brief and initially saw the data?
- What dimensions of analysis do we think might be of interest or relevance about this subject matter?
- What data questions will we seek to answer in our visualization design?

To refine our focus we need to commence our visual analysis work to explore our dataset and see what comparisons, trends, patterns, and relationships we can identify. Out of this we will hope to unearth some interesting stories to tell.

Given we have a small dataset with only two variables we shouldn't need to embark on too much varied visual analysis.

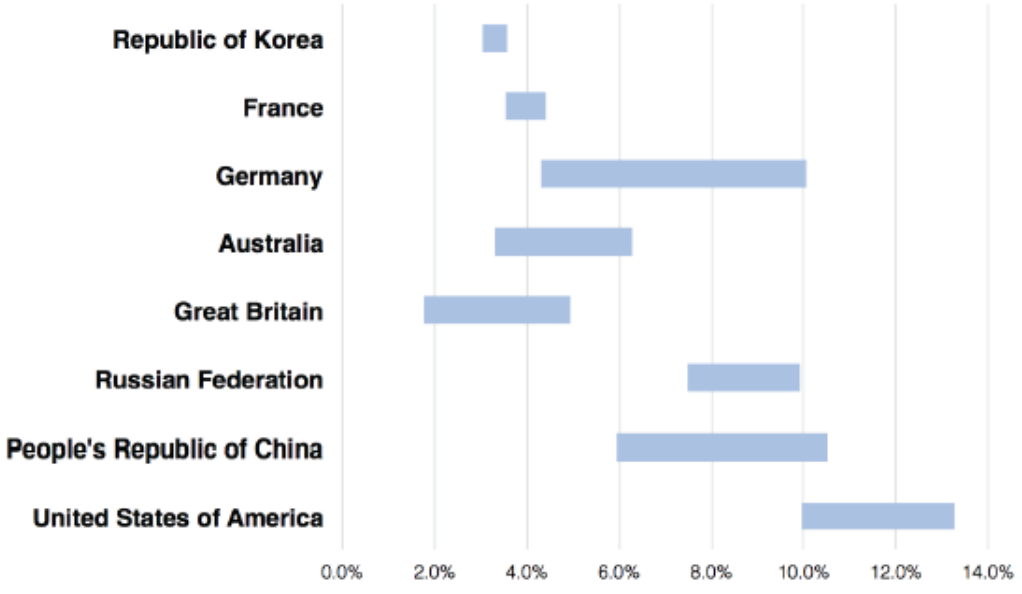
The first graphic takes a look at the variation of medal winning across the years, showing the range of totals for each country using a floating bar chart:



Through interpreting this chart in conjunction with the descriptive statistics we just collected, we are able to form some interesting data questions about the subject and start to get a feel about the main stories, such as:

Question	Answer
Which countries have experienced a significant change in their medal-winning performance levels?	We're looking for the widest bars to show the variability, this could be improvement, decline or inconsistency. We would identify the spread of Germany and China as being particularly interesting.
Which countries have maintained consistency in their performance levels?	Now we're looking for the narrowest bars, the tightest of value ranges. This leads to noticing the USA, France, and especially Republic of Korea.
What have been the most interesting country stories in terms of the transition of their performance and rankings?	Possibly too hard to see with this chart, but there is potentially something going on with the bars that intersect and exceed the lengths of others. At this stage, the story of China seems to stand out as being something to look out for.

Let's now repeat the same chart type but apply it to a transformed version of the data that has been standardized to show the medals won as a percentage of the overall total:



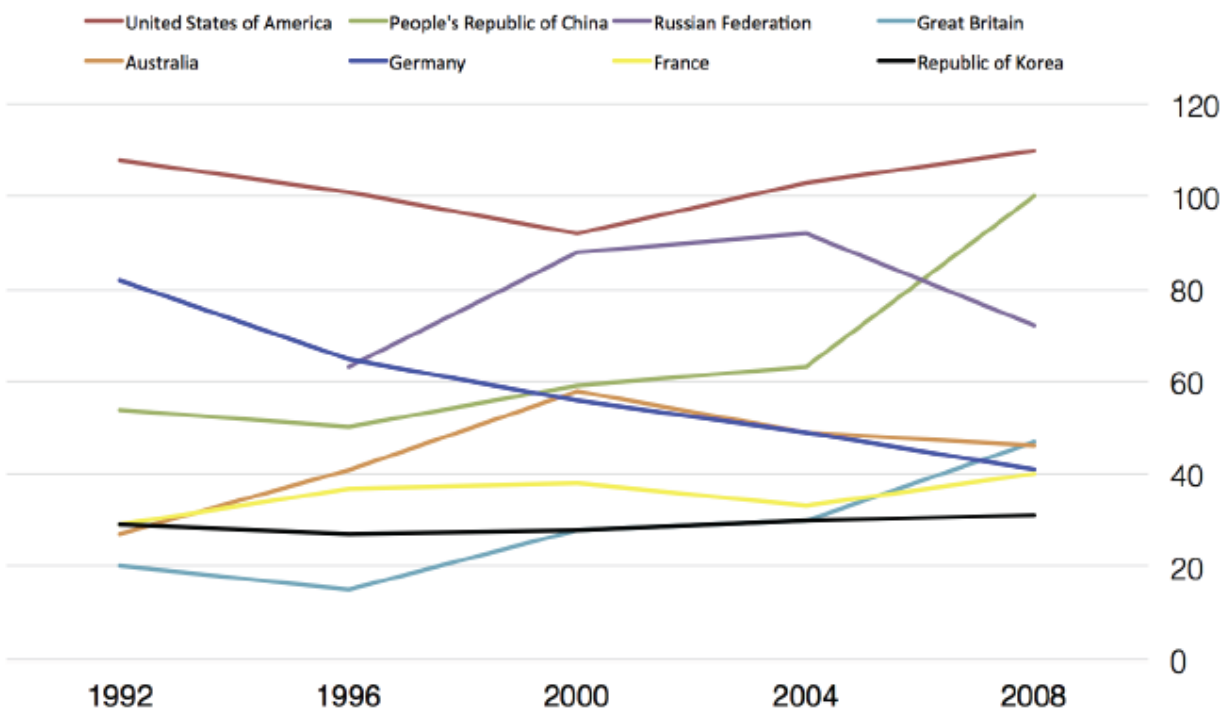
Does this alter the focus of our questioning or change our impressions of the main insights?

If anything it reinforces them, especially our interest in the varied performance levels for Germany and China. It also emphasizes the remarkable consistency of Republic of Korea and France.

At this point, we have definitely established a scent for the story. We have started to articulate the data questions that best interrogate this data and most likely reflect what the readers of a visualization about this subject will wish to learn.

We now need a different visual representation. Using the floating bar we have seen the categorical view of the countries and their performances. Now, we need to switch our perspective to the other main variable, that of event year, to pursue our curiosities about the transition of medal-winning performances and the transition in ranking of the individual countries across the five Olympic Games.

For this next visual sketch we turn to a line chart. On this single chart we plot the eight countries, differentiated by color, showing the absolute medal wins from left to right across the five Olympic events:



It looks a bit messy doesn't it? Don't worry. Remember, this is an exploratory visualization for ourselves. We are the audience and we just want to see if we can discover some interesting physical properties about the data in this display.

You wouldn't and shouldn't publish an isolated, cluttered, and poorly-annotated chart like this to convey a story to others, but when it is a visualization serving yourself, it is a different matter. You created it and you know what you're looking out for. Quick and dirty is absolutely fine.

The decision to place all countries onto one graphic is to enable visibility of the interesting transitions, the crossovers, the seemingly cluttered parts, and the empty parts. You could separate each country out into its own line chart and assess a matrix of eight small-multiples. However, this would only show you the individual country stories. Our keen interest here is in the relationship between the countries.

The chart shows how Germany's (blue) wide range of results, actually reflects their general decline in medal winning levels and, by extension, their relative rank.

By contrast, China's wide distribution shows a country on the rise over the past four games at least. The extended fascination of this trend would be whether they will catch up and possibly overtake the US once we have the results and data for the 2012 Games (not available at the time of writing this book!).

Elsewhere, Russia can be seen to have moved up and down over the years and has now been overtaken by China. There is an interesting chunk of white space for the 2008 results either side of the Russian value, leaving them quite comfortably in third position. Interestingly, the UK has seemingly demonstrated a very similar pattern of improvement relative to the Chinese over the past five events.

Sometimes no change is as interesting as some change and, in this respect, the consistency of Republic of Korea is quite stark given the different generation of competitors who will have contributed to those totals.

Otherwise there is nothing else really of significant interest. The charts have served their purpose in discovering and confirming some relevant and interesting stories concerning the contrasting experiences of China, Germany and, potentially, the Republic of Korea.

Of course, sometimes you simply may not find a story. There just might not be anything of substance to convey to others visually, in which case a table of data may prove to be the most appropriate solution.

However, we *have* found our stories, so how do we tell them? As a bridge to the next chapter, where we will be focusing on design matters around presenting our stories, let's attempt a quick solution.

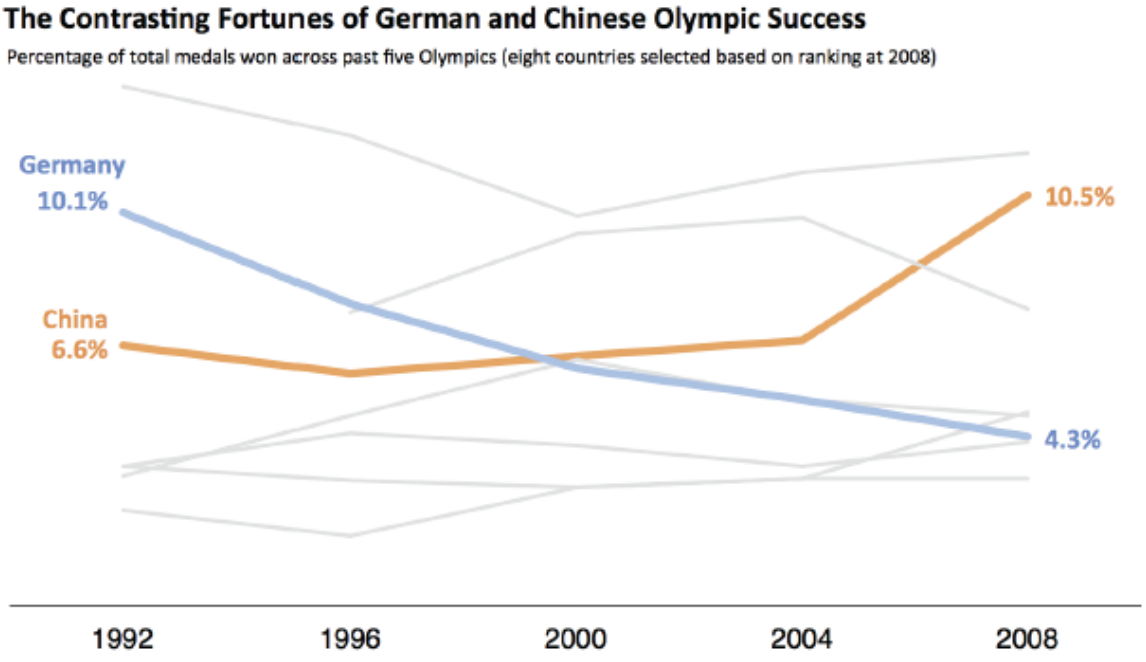
Remember the quote we saw earlier from Amanda Cox: "different forms do better jobs and answering different questions"? Let's reduce the story to a simple contrast between China and Germany. Our main data question will be something like "how have the medal-winning performances of China and Germany compared over the past five events?"

The most suitable method for giving form to and answering this question will still be a line chart. Similar to the one we used for the visual analysis, we are trying to show the relationship between these two countries' respective performance over time.

However, the design execution will be different. This time we're conveying the story to others, so we need to refine the visuals in order to make it an explanatory piece:

- We need to elevate the important features of the main story and relegate any background context and secondary content.
- We need to ensure that there are annotations for labels, values, and captions so the reader is entirely clear about what is being communicated.

Here is a proposed solution for telling this story:



The first thing to point out is that we have used the calculated data for medals won as a percentage of the total. This is more appropriate for this story as it helps standardize and contextualize the performance across all events in a more comparable way.

The aim here is to provide a clear visual hierarchy emphasizing the two main countries in our story and diminishing the contextualizing six nations into the background. We could have removed the other six countries but, through the use of a subtle shade of grey, we can still see them well enough to get a sense of the overall rankings. That is all we need from them—context.

The title neatly frames the story, the subheading describes the chart and the data, and the labels help the reader compare the two countries' relative trajectory.

The use of color attempts to help imply the positive improvement (orange = hot = good) of China and the negative decline (blue = cold = bad) of Germany. Only the bare minimum chart apparatus (the axis line) is included, once again, to allow the main story to come to the fore.

Contrast this design approach for telling a story (explanatory) with the design of the same chart method we used to find the stories (exploratory); here we provide nothing more and nothing less than the reader requires to easily interpret the story. This use of contrasting visual approaches for the same chart types but for different intentions is important to recognize in your design work.

Annotation

Here is a quote from Amanda Cox (`http://eyeofestival.com/speaker/amanda-cox/`):

> *"The annotation layer is the most important thing we do... otherwise it's a case of here it is, you go figure it out."*

Our next layer is one that can often be neglected. However, as this quote suggests, annotating visualization is such an important features of our design. It is about taking care of your audience, recognizing who they are, what they might know already, and what they don't know.

Done well, annotation can help explain and facilitate the viewing and interpretive experience. It is the challenge of creating a layer of user assistance and user insight: how can you maximize the clarity and value of engaging with this visualization design?

As discussed in the first chapter, a key objective for effective data visualization design is the facilitation of accessibility into a subject through intuitive design. The degree of accessibility is enhanced through the effective inclusion of useful explanation across all features of your visualization solution. We shouldn't assume that readers or users are instantly and easily going to be able to navigate their way around our designs and so we need to carefully consider the best ways to assist them; explained as follows:

- **Titles**: A compelling title can help to attract an audience and articulate the focus of the visualization subject matter. Sometimes, especially in explanatory visualizations, you can look to exploit this prominent space to tell readers about a key insight or headline. However, make sure it is an accurate reflection of the content of the visualization otherwise it will be misleading.
- **Introductions**: These are really important instructive elements to explain t he project's background and context, describing the background motivation and what your intentions are in terms of how it should be used.
- **User guides**: While intuitive accessibility is stated as an overall goal, many projects often warrant further explanation, particularly with interactive pieces and those that have inherently complex subjects or frameworks.

  In this next project, titled *Political Moneyball* and created by the Wall Street Journal, we see a demonstration of exceptional care for the audience's understanding of how to optimize this visualization's use. Not only does it include thorough written annotation and labeling to help users understand all the features of this incredibly immersive tool, but there is also a video tutorial to offer that extra degree of support. The designers of this piece astutely recognize the potential depth and interpretive complexity of the subject matter and I imagine also want to do justice to their efforts to bring this deep subject to fruition.

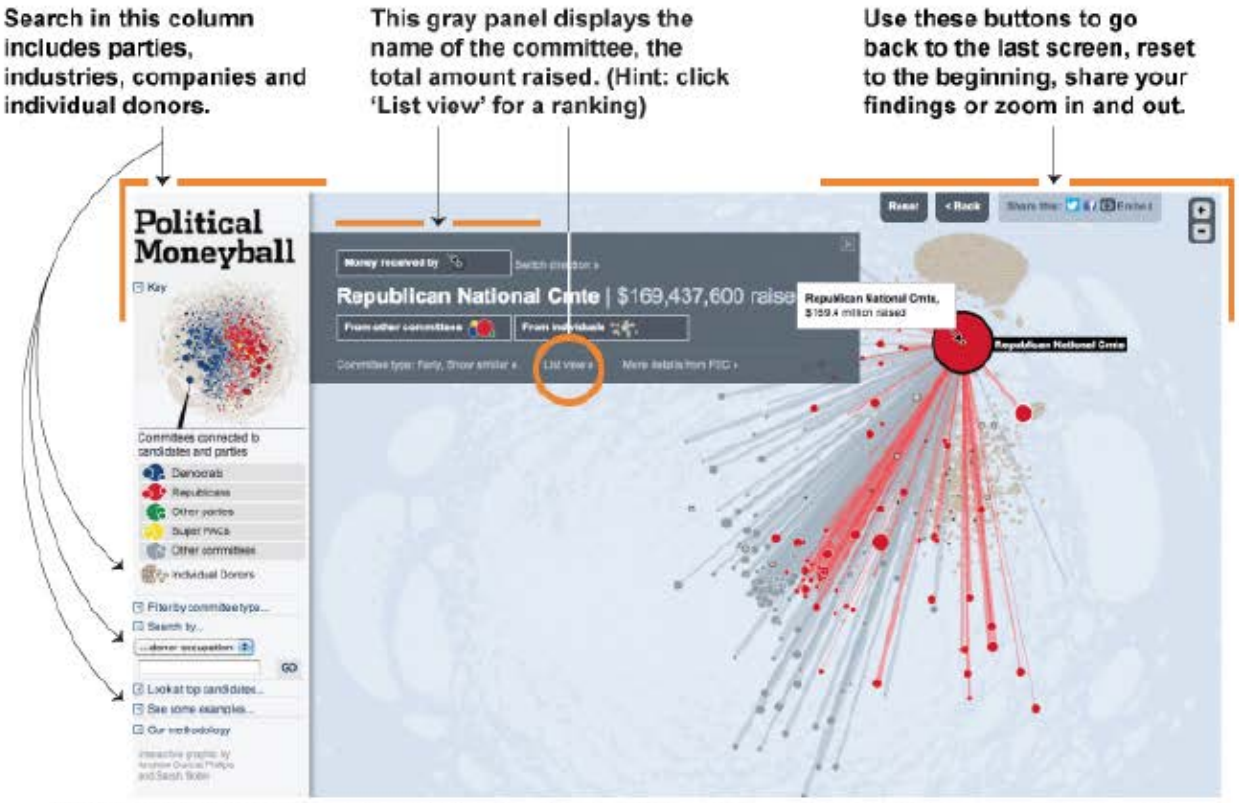


Image from "Political Moneyball" (`http://graphics.wsj.com/political-moneyball/#`), created by Andrew Garcia Phillips and Sarah Slobin of the Wall Street Journal

- **Labels**: In the interactive section, we discussed the potential of labels to reveal extra details about data values. As we see in the previous project, labeling is an incredibly simple but useful device to help explain matters. Often, these are hidden and interactively revealed through selection or by hovering.
- **Captions and narrative**: In addition to the potential use of the title to offer a key headline, sometimes you may wish to surface important insights and findings to help fast-track the reader's interpretation process. You might draw out the good and the bad or maybe the expected or unexpected. You should also consider the potential value, in certain projects, of the "what next?" question—what should the user do with this information? what actions need to be taken?

- **Visual annotation**: Annotation goes beyond just written explanations and we should consider how to use chart or graphic devices to help draw out important insights visually. Simple options include features such as gridlines, axes labels, and tick marks. In *Chapter 3, Demonstrating Editorial Focus and Learning About Your Data*, we saw an example of effective visual annotation. Here, reference lines and background shading is used effectively to help the reader achieve distinction between different tiers of interpretation, as you explore the relationship between what countries spend on education and the military.

Image from "In Numbers: Education Around the World" (`http://visualdata.dw.de/specials/bildung/en/index.html`), created by Gregor Aisch for Deutsche Welle

- **Legends and keys**: Always explain the use of color schemes or the varying size of shapes in terms of their categorical or quantitative representation.
- **Units**: You should include details of the units of values being displayed to ensure you don't create ambiguities and potential misinterpretation. As with many of these annotated features, this is an obvious requirement, something we've had drilled into us since our school days, but you'd be surprised how often they can be left out.

- **Data sources**: It is vital to include detailed references about from where you have accessed your data or any other sourced element (such as imagery). Where you have chance to offer a more detailed narrative, you may wish to explain what treatment you have applied to the data in terms of its quality or analytical transformation.
- **Attribution**: Don't forget to acknowledge those who have either contributed directly, influenced the construction of the design, or those people whose work has acted as a source inspiration.

The final thing to mention about annotation is that this is likely to be the first time we have to consider our typography selections. There are, of course, plenty of established guides and sources of literature to help influence your choice of fonts for all pieces of written annotation. However, this is another aspect of design that you will be able to ultimately judge best using your own design instinct. Many designers have their favorites and like to maintain this identity but also many projects may be required to observe certain visual identity rules like we outlined in the color section.

Arrangement

You have established how you are going to represent your data, you've identified your visual identity through color, the choices around static or interactive design have been rationalized, and you have identified the range of annotation requirements.

For our final layer, we need to consider how to arrange our design in terms of the layout, placement, and organization of all visible elements. How can we piece everything together most effectively?

As we've just discussed in relation to annotation, our intention with the arrangement and architecture of our design is to deliver as intuitive an experience as possible. The level of intuitiveness and smooth access into the subject matter is strongly influenced by the logic and implied meaning behind the arrangement of our chart elements, the interactive features, and annotation devices.

The key overall aim is to reduce the amount of work the eye has to undertake to navigate around the design and to decipher the sequence and hierarchy of the display. For the brain, once again, we're looking to minimize the amount of thinking and "working out" that goes on. We therefore need to carefully consider the choices we make around the size, positioning, grouping, and sorting of all that we show. As with all visualization design layers, we need to be able to justify the decisions we make about every visible property presented.

Here is a simple, but effective, demonstration of the careful consideration of arrangement. It is just one example out of many we could refer to from the projects shown in this book.

Observe the positioning of the chapter navigation slider across the top, the size of the space afforded to the main map display, the narrative found on the right-hand side, the proximity of the legend to the data, and the location of the pan and zoom device—all these decisions are very deliberate and designed to maximize the logic and meaning behind the layout of this project's data, its interactive features and annotated elements.

Image from "The Growth of Newspapers Across the U.S.: 1690-2011" (`http://www.stanford.edu/group/ruralwest/cgi-bin/drupal/visualizations/us_newspapers`), created by Rural West Initiative, Bill Lane Center for the American West, Stanford University.

On the matter of arrangement, it is important to mention an important paper produced by Edward Segal and Jeff Heer of the Stanford Vis group and titled *Narrative Visualization: Telling Stories with Data* (`http://vis.stanford.edu/papers/narrative`).

As the title suggests, this article provides an excellent outline of the different design strategies for arranging and structuring the layout of your visualizations that will help maximize the potential telling of stories through data.