

Infovis and Statistical Graphics: Different Goals, Different Looks¹

Andrew Gelman² and Antony Unwin³

20 Jan 2012

Abstract. The importance of graphical displays in statistical practice has been recognized sporadically in the statistical literature over the past century, with wider awareness following Tukey’s *Exploratory Data Analysis* (1977) and Tufte’s books in the succeeding decades. But statistical graphics still occupies an awkward in-between position: Within statistics, exploratory and graphical methods represent a minor subfield and are not well-integrated with larger themes of modeling and inference. Outside of statistics, infographics (also called information visualization or Infovis) is huge, but their purveyors and enthusiasts appear largely to be uninterested in statistical principles.

We present here a set of goals for graphical displays discussed primarily from the statistical point of view and discuss some inherent contradictions in these goals that may be impeding communication between the fields of statistics and Infovis. One of our constructive suggestions, to Infovis practitioners and statisticians alike, is to try not to cram into a single graph what can be better displayed in two or more.

We recognize that we offer only one perspective and intend this article to be a starting point for a wide-ranging discussion among graphics designers, statisticians, and users of statistical methods. The purpose of this article is not to criticize but to explore the different goals that lead researchers in different fields to value different aspects of data visualization.

Recent decades have seen huge progress in statistical modeling and computing, with statisticians in friendly competition with researchers in applied fields such as psychometrics, econometrics, and more recently machine learning and “data science.”

But the field of statistical graphics has suffered relative neglect. Within the field of statistics, exploratory methods represent a subfield with relatively small influence. For example, as Howard Wainer has noted, the papers in the *Journal of Computational and Graphical Statistics* are about 80% computation and 20% graphics, and in applied work, graphics is

¹ We thank Nathan Yau for posting the infographics and commentary that motivated this work, Jessica Hullman, Hadley Wickham, Lee Wilkinson, Chris Volinsky, Kaiser Fung, Alfred Inselberg, Martin Wattenberg, and conference participants at the University of Kentucky, Iowa State University, and the University of California for helpful comments, the Institute of Education Sciences for grants R305D090006-09A and ED-GRANTS-032309-005, and the National Science Foundation for grants SES-1023189 and SES-1023176.

² Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu, <http://www.stat.columbia.edu/~gelman/>

³ Department of Mathematics, University of Augsburg, unwin@math.uni-augsburg.de

typically thought of as a way to help with simple tasks such as data cleaning and exploration, before getting to the serious task of inference. Meanwhile, outside of statistics, data graphics have become hugely popular, with innovative visualizations appearing regularly on the web and in the *New York Times*.

We see an unfortunate lack of interaction between the worlds of statistical graphics and information visualization. As statisticians, we recognize that our graphs are not generally catching fire in the modern media environment: we are unhappy at being left behind and wonder what we are doing wrong. The sorts of graphs we do best are being squeezed from both ends, from one direction by the good-enough bar plots that just about anyone can make in Excel, and from the other by the beautiful professionally-designed images that are the Armani suits to our clunky plaid shirts with pocket protectors.

But this is more than a concern about losing market share. We are also disturbed that many talented information-visualization experts do not seem interested in the messages of statistics, most notably the admonitions from William Cleveland and others to consider the effectiveness of graphical displays in highlighting comparisons of interest. We worry that designers of non-statistical data graphics are not so focused on conveying information and that the very beauty of many professionally-produced images may, paradoxically, stand in the way of better understanding of data in many situations.

The purpose of the present article is to start a conversation between practitioners in statistical graphics and information visualization. We hope that by identifying the different goals that motivate work in these two areas, we can get the best researchers in these fields to learn from each other.

An important part of statistical practice is graphical communication of data and models, and this is an important part of statistical theory as well, as has been recognized by Tukey (1972, 1977) and others. Much has been written about statistical graphics in recent years but very little, we believe, on the different goals involved in visual data displays.

We believe it is vital for the field of statistics to engage with those outside the field who use related methods to attack related problems. Thus, in the spirit of outreach efforts by statistical researchers to understand econometrics, expert systems, machine learning, fuzzy logic, neural nets, and other approaches developed outside the field for learning from data, we seek here to stimulate a discussion on information visualization—those grabby graphs that are in many ways the public face of modern-day statistics.

We anticipate that this article will spark disagreement, but we think it is best to have such debates in the open rather than for statisticians, computer scientists, and graphic designers operating independently. There is no right answer in graphics, and we hope that a forthright discussion of different goals will be helpful to all.

Looking at Infovis through statisticians' eyes

We begin our story in December, 2008, when statistician and graphic designer Nathan Yau published on his influential Flowing Data blog⁴ a list of what he viewed as the five best data visualizations of the year (Yau, 2008b). We were struck by the discrepancy between the visual appeal of these displays and their divergence from the usual principles of statistical graphics.

Upon further reflection, and after a blog exchange (Yau, 2009, Gelman, 2009a), we decided that our difficulties with some popular visualizations arose from an insufficient understanding—by ourselves and others—of the multiplicity of goals involved in data display. These goals reflect the differing interests and approaches of two different groups: On the statistical side, data analysts and statisticians are interested in finding effective and precise ways of representing data, whether raw data, statistics or model analyses. Providing the right comparisons is important, numbers on their own make little sense, and graphics should enable readers to make up their own minds on any conclusions drawn, and possibly see more. On the Infovis side, computer

⁴ It may be unusual for a journal article to be reacting to a blog—but the blog in question has approximately 15,000 subscribers, about three times more than the most prominent academic statistics blogs and more readers per day than many scientific journals get per year. And the issue is not just circulation. Thanks to Yau and his commenters, Flowing Data is a thoughtful forum on the interface between statistics and graphic design.

scientists and designers are interested in grabbing the readers' attention and telling them a story. When they use data in a visualization (and data-based graphics are only a subset of the field of Information Visualization), they provide more contextual information and make more effort to awaken the readers' interest. We might argue that the statistical approach concentrates on what can be got out of the available data and the Infovis approach uses the data to draw attention to wider issues. Both approaches have their value and it would probably be best if both could be combined.

The present article comes, unavoidably, from a statistical perspective, and we discuss ways in which several popular data visualizations do not serve statistical goals. We are not writing this as a critique of the Infovis approach; our intent is to consider the goals being served by graphical displays that we might not choose ourselves, with the ultimate goal of improving communication among graphics designers, statisticians, and users of statistical methods. This is not a division between disciplines so much as a debate going on within all these fields.

One issue that arises is the familiar distinction between exploratory and presentation graphics. With presentation graphics you prepare some small number of graphs, which may be viewed by thousands, and with exploratory graphics you prepare thousands of graphs, which are viewed by one person, yourself. Exploratory graphics is all about speed and flexibility and alternative views. Presentation graphics is all about care and specifics and a single view. Presentation graphics can really benefit from a graphic designer's contribution; for exploratory graphics it's not so relevant. That said, the first consumer of any graph is the person who makes it, and it can often be useful to use "presentation" skills to communicate to ourselves as well as to others. In either context, much can be gained by thinking carefully about goals.

In this paper we will be writing primarily about static presentation graphics, a well-established and mature field. Modern exploratory data analysis involves using interactive graphics, and such tools are also frequently used for Infovis graphics in web displays, along with sound and video. However, these approaches are very much in a development phase and we would prefer to encourage further experimentation rather than comment of what are still early efforts.

Data graphics are increasingly being produced in all sorts of contexts. We are happy to see the increasing recognition of the importance of visualizing data, but we have some concerns that practitioners are not fully aware of the multiplicity of goals that arise in graphical presentation. In the present article, we lay out some of these conflicting goals and discuss how awareness of some underlying principles of statistical communication could improve the work of statisticians and graphics designers alike.

Understanding and dialogue rather than pure criticism

Our point is not to criticize or to pass judgment (whether positive or negative) on graphs made by non-statisticians, but rather to explore and understand, through examples, the different goals and different approaches of the two groups. To the extent that our perspectives differ from the creators and audiences of the information visualizations, we highlight these differences—not to say or imply that we are right and they are wrong, but to clarify the many different goals involved in data display.

Why is it important to have this dialogue?

Statistical graphics are fine but they don't always serve to communicate outside the academic/professional bubble. Infovis is much more popular than statistical graphics, and it behooves us to understand why.

From the other direction, in an Infovis setting, the public often gets a pretty graph or raw data with nothing in between. If designers can understand the goals of statistical graphics (and how they differ from the typical goals of Infovis), perhaps they can see the value of line plots, scatterplots, and more sophisticated statistical graphs as a useful tool for understanding data, for those readers who have been drawn in by the infograph and now are ready for more insight.

As we shall see, the very features that make an effective information visualization can be detrimental to statistical presentation of data—and vice-versa. Hence both statisticians and

designers can benefit from understanding each other's perspective, with the aim not being a single display that makes everyone happy but a set of different data views that serve different purposes.

Sources for the two points of view

There are several fine books on presentation graphics for statisticians, including the theoretical works of Bertin and Wilkinson, the style advice books of Cleveland and others, and the attractively polemical books of Tufte. In fact some of Tufte's publications are a bridge to the Infovis world, which has a newer and more scattered literature. Articles by Heer, Kosara, Munzner, and Shneiderman are a good starting point and Kosara's blog, eagereyes.org, is a useful place to look for enlightened discussion of the issues. Amongst other contributions, Shneiderman (1996) has proposed and promoted his mantra: Overview first, zoom and filter, then details-on-demand. In effect this is a drill down for details and there is no mention of any comparisons. For statisticians there always have to be comparisons; numbers on their own are not enough.

There is a series of Infovis workshops, BELIV (BEyond time and errors: novel evaLuation methods for Information Visualization), concerned with the evaluation of visualizations. Substantial progress has not been made, but the aim of trying to determine what insights may be obtained from a graphic and how well they are presented is well worth pursuing and such research should encourage statisticians too to think more formally of what they are trying to achieve with their graphics.

A clear presentation of a computer science perspective on visualization as "augmented cognition" appears in the introduction to Card et al. (1999). Within engineering and computer science there have been debates about the role of aesthetics and design in the user experience, beyond direct functionality. In an article with the subtitle, "Attractive things work better," Norman (2002) expresses the importance of design and writes: "Affect makes us smart . . . Affect therefore regulates how we solve problems and perform tasks. Negative affect can make it harder to do even easy tasks: positive affect can make it easier to do

difficult tasks.” Translating this to statistical graphics: the aspects of a display that register as “beautiful” or “cool” can make a reader more comfortable with the statistical information and substantive context of a graph. At the same time, we must remember that a graph can look good—even look appropriately “functional”—while failing as a data display. Just as some notorious buildings designed by modernist architects in the mid-twentieth century looked functional but actually had serious practical problems with leaks in the ceilings, poor air circulation, inability to adapt to new uses, and so on.

Some goals involving the visual display of quantitative information

Why is visualization important for data analysis? Tufte (1983) writes, “At their best, graphics are instruments for reasoning about quantitative information,” a surprisingly weak statement really! Cleveland (1985) writes, “Graphs are exceptionally powerful tools for data analysis.” Chambers et al. (1983) begins, “There is no statistical tool that is as powerful as a well-chosen graph.” The chapter on diagrams in *Statistical Methods for Research Workers* (Fisher, 1925) begins, “Diagrams prove nothing, but bring outstanding features readily to the eye.” These last three statements by statisticians are all very positive but remain general. Tukey (1993) was more specific about what he called the true purpose of graphic display, which he set down in four parts:

1. Graphics are for the qualitative/descriptive—conceivably the semiquantitative—never for the carefully quantitative (tables do that better).
2. Graphics are for comparison—comparison of one kind or another—not for access to individual amounts.
3. Graphics are for impact—interocular impact if possible, swinging-finger impact if that is the best one can do, or impact for the unexpected as a minimum—but almost never for something that has to be worked at hard to be perceived.
4. Finally, graphics should report the results of careful data analysis—rather than be an attempt to replace it. (Exploration—to guide data analysis—can make essential interim use of graphics, but unless we are describing the exploration process rather than its results, the final graphic should build on the data analysis rather than the reverse.)

With his first two points Tukey emphasized that for looking up numbers you should use tables not graphics, a point that is true but basically obsolete with modern computing (see Friendly and Kwan, 2003). His second point, on comparisons, is key for statisticians. His third point is about showing what you want to show clearly, and the phrase “swinging-finger impact” suggests he was not against using graphics for persuasive purposes. Tukey’s final point is expressed rather negatively, implying he saw too many graphics without proper supporting analysis. In practice we would argue that graphics and analysis are complementary. Like Tukey we do not like graphics without supporting analysis. We also do not like analysis without supporting graphics (and we are sure Tukey would have agreed with that too). To put it another way, a picture may be worth a thousand words, but a picture plus 1000 words is more valuable than two pictures or 2000 words.

Tukey was primarily writing of small datasets. Our proposed goals for graphics take more account of the vastness of data nowadays and more account of communication through graphics.

Discovery goals:

- Giving an overview—a qualitative sense of what is in a dataset, checking assumptions, confirming known results, looking for distinct patterns.
- Conveying the sense of the scale and complexity of a dataset. For example, graphs of networks notoriously reveal very little about underlying structure but, if constructed well, can give an impression of interconnectedness and of central and peripheral nodes. And maybe that is the point. The picture tells the story as well as, and in less space than, the equivalent thousand words.
- Exploration: flexible displays to discover unexpected aspects of the data; small multiples or, even better, interactive graphics to support making comparisons.

Communication goals:

- Communication to self and others: displaying information from the dataset in a readily understandable way. Information density is great, but only if this information can be visually extracted from the graph! (cf. Tukey’s third point)

- Telling a story. This is really another form of communication. If we communicate well, we call it storytelling. Consider, for example, Minard's Napoleon-in-Russia graph popularized by Tufte (1983).
- Attracting attention and stimulating interest. Graphs are grabby. Not so much in submitted journal manuscripts (where, by convention, they may be placed in a pile at the end of the article) but in newspaper articles, blogs, and so forth. The flip side of this is that graphs are often viewed as intimidating; for example, Barabasi (2010) writes, echoing Hawking (1988), "There is a theorem in publishing that each graph halves a book's audience." This may be one reason that no graphs appear in data-rich books such as *Freakonomics* (Levitt and Dubner, 2005) that one might expect to be full of visual data displays.

Each of these goals can be important but it would be meaningless to try to achieve all of them at once. The most general goals we can think of in data display are *discovery* (the first three of our goals above) and *communication* (the second three goals). These can go together—we want to communicate our discoveries!—but also to some extent lead in different directions. To put it simply, we communicate when we display a convincing pattern, and we discover when we observe deviations from our expectations. These may be explicit in terms of a mathematical model or implicit in terms of a conceptual model. How a reader interprets a graphic will depend on their expectations. If they have a lot of background knowledge, they will view the graphic differently than if they rely only on the graphic and its surrounding text. Consider the analogy of fine art and advertising. A painting often has much detail and demands a great deal of your attention. You expect to work at it, and you know that extra information, not included in the painting, may be helpful to your understanding. With advertising, you either get an immediate hit or nothing (though repetition—you may see the same advertisement many times—is an additional complicating factor). Statisticians are analysts; they want to get down to the essentials and display specifics in a precise way. They assume they already have the reader's interest and involvement. Designers are artists and want to display an environment, a complex whole. They want to attract attention and encourage involvement. By giving the sense of that complex whole, you may distract from conveying the main information. The very best data visualizations (for example, the Baby Name Wizard, discussed near the end of this article) manage to do both.

Making graphics attractive can help motivate readers to understand them. And, once the reader is willing to put in the work, innovative and even inefficient graphical displays can be effective in implicitly extracting a commitment from the reader to think hard about the data.⁵ For graphics designers, novelty is an end in itself, with the goal of eliciting the reaction, “Hey, I’ve never seen this before,” followed by, “Of course—that’s a really good way to display these data.”

Statisticians tend to use standard graphic forms (for example, scatterplots and time series), which enable the experienced reader to quickly absorb lots of information, but may leave other readers cold. We personally prefer repeated use of simple graphical forms, which we hope draw attention to the data rather than to the form of the display. In addition, with familiar structures, visual conventions carry some of the work of exposition. Consider, for example, the use of the horizontal axis for a predictor variable and the vertical axis for the outcome, or, in a bar chart, the idea that the areas of the bars convey relative numbers. You need experience with a graphic form to use it well. For judging graphics in the media, statisticians should perhaps take the likely audience more into consideration.

We regard graphics as part of a story rather than as isolated objects. A graphic does not live on its own. There can be (working from inside to out) annotations, a legend, a title, a caption, accompanying text, an overall story, and a headline. The elements should all be consistent and in tune with one another (which is often not the case in the news media, where different people are independently involved at different stages of preparation). And the graph itself is often part of a grid of graphs, or even one of several associated displays within an online posting, an article, or a book. Using several graphics allows you to present a selection of views, each one of which makes an additional contribution to the overall picture.

⁵ When exposed to novel graphics, readers have to make an effort of understanding. Having made the effort, they have an emotional commitment to the graphic (nobody wants to admit they wasted their time). Whether they have actually learnt anything useful is difficult to tell and would make for an interesting research study. This effect is related to the positive emotional buzz we can get from working out how to do something in R. Up to a certain point, the longer it takes us to work out how to do what we want to do, the more satisfaction we get from finding the solution—even if in retrospect it was something we should have known. And all that time has been “wasted,” it could have been spent on the real statistical problem and not on the computing problem.

Beyond this, statisticians should remember that, contrary to the impressions they may have received from a hasty reading of Tukey, graphics are not just for visualizing data. For example, the parallel coordinate plot (Inselberg, 2009) is a modern standby, an excellent tool for the clear display of multivariate data, but it was originally developed as a way of visualizing high-dimensional structures in pure math, with no data in sight. Visualization has its own principles which are relevant to statistics without being part of it.

Newspaper and magazine articles are often illustrated by photographs and cartoons which are pretty or shocking or attention-grabbing or interesting and in some way complement the substance of the article. We don't generally criticize newspaper illustrations as being noninformative; they're not really expected to convey substantive information in the first place. From that perspective, an infographic can be a good choice even if it does not clearly display patterns in the data.

Bateman et al. (2010) have studied people's recall of information, comparing embellished graphics with plain vanilla ones and found that recall after a two or three week gap was much better for the (dramatically) embellished graphics. This should not surprise us. We would probably remember a young man in green and white striped jeans better than one in a smart business suit. Does this mean we should dress exotically, if we want to be noticed? It obviously depends on what we want to be noticed for. The same goes for visualizations and we should think of graphics as being part of a whole, not view them in isolation.

Balancing the conflicting goals of designers and statisticians would be easier if the two groups worked together more. Designers would doubtless find statistical graphics dull and austere, while Chatfield (1995) captured many statisticians' view when he wrote, "the features of a graph which make it visually attractive (e.g., colour, design complexity) may actually detract from comprehension. Some people deliberately choose to present fancy graphics which look pretty but are not interpretable. In other words they use graphics to hide information while appearing to do the opposite. The misuse of graphics is particularly hard to combat because most people think they understand graphs."

Background

The recent popularity of data visualization can perhaps be traced to developments in computer technology and the influence of Tufte's 1983 cult classic *The Visual Display of Quantitative Information* and his subsequent books in the area. In the popular press, the state of the art of graphical display has moved from the goofy front-page charts in *USA Today* (so memorably parodied by *The Onion*; see Figure 1) and the sober time series and scatterplots of *The Economist* (Figure 2), through high-tech, information-rich graphs of the stock market and the weather (Figure 3), to a variety of creative visualizations that have appeared in the *New York Times*.

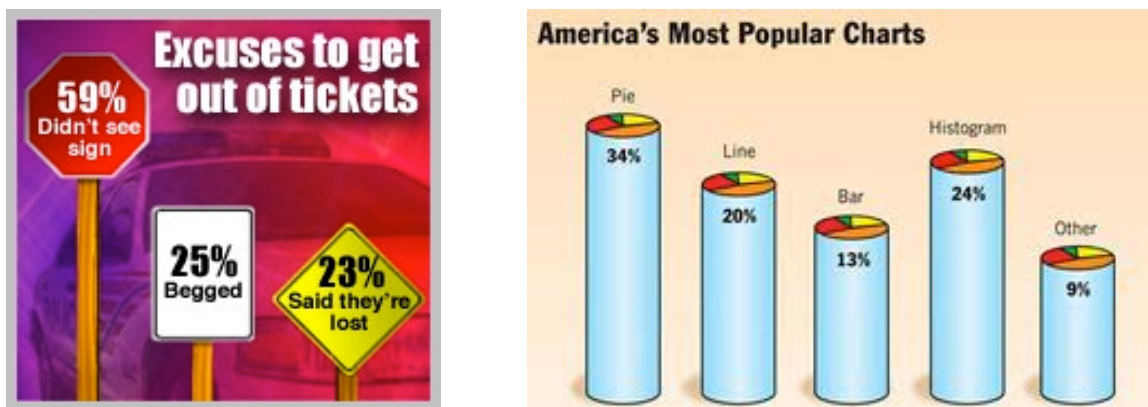


Figure 1. Goofy graphics. The real thing (from *USA Today*) and the parody (from *the Onion*).

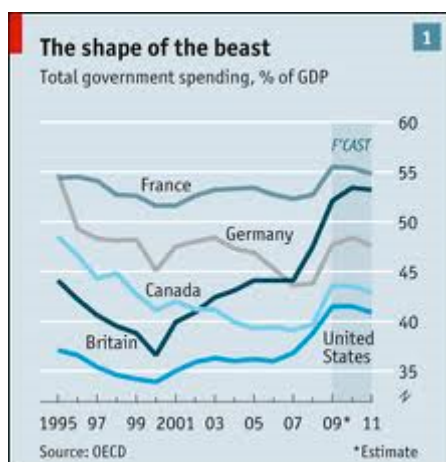


Figure 2. *The Economist* continues its long tradition of understated, informative data displays.

2009 Weather in NYC

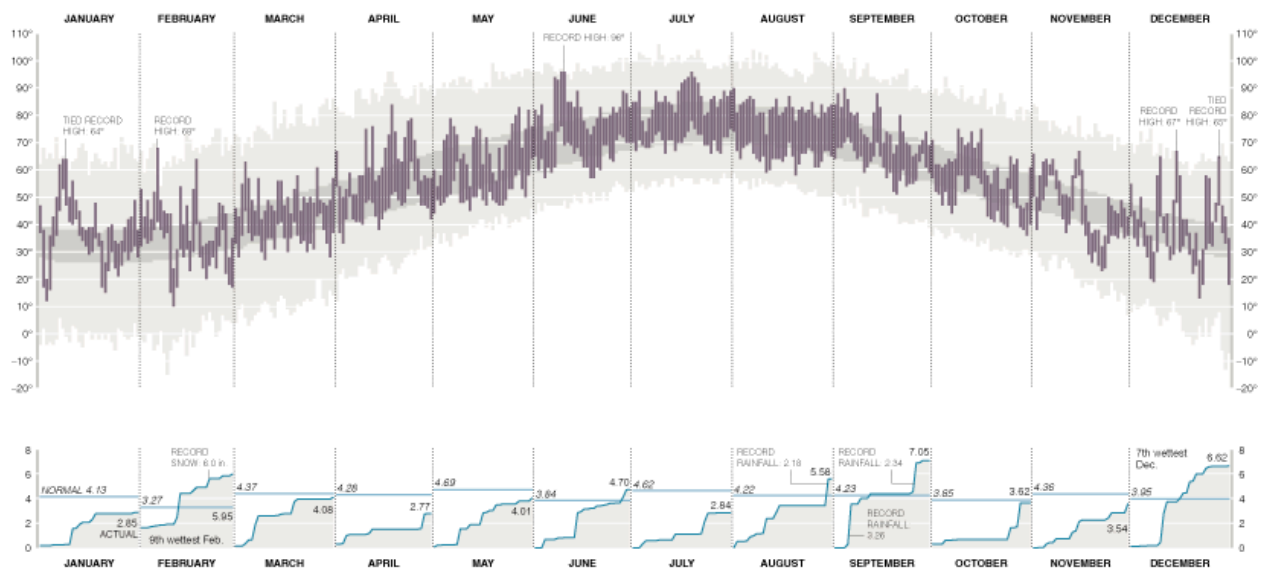


Figure 3. This weather chart is a technically impressive data visualization, but for just about any practical purpose it offers an excessive amount of detail. If you accept the goal of displaying the exact high and low temperatures for every day in the year, the graph does an excellent job—but it is not clear why these numbers should be of interest to a newspaper reader, so many months later. One concern both with statistical graphics and with information visualization is a focus on technique without always a sense of larger purpose. It should be possible to view this as a wonderful display of a time series, while at the same time questioning its utility.

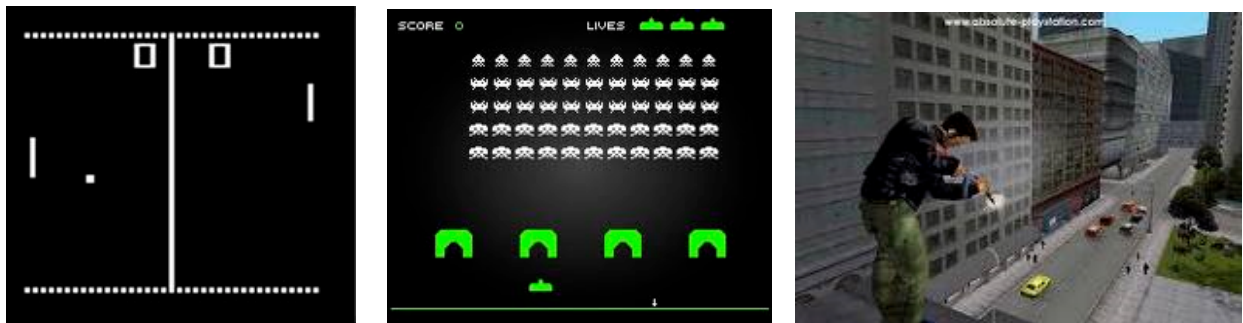


Figure 4. Progress in computer graphics, from Pong to Space Invaders to Grand Theft Auto.

At the same time, computer graphics have made incredible leaps, from Pong and Space Invaders to Grand Theft Auto (see Figure 4) and the three-dimensional imaging that has become routine in children's television cartoons. It is no surprise that data visualization tools have moved beyond those of the Pac-Man era.

How does this all relate to statistical theory and practice? The statistical literature on visualization tends to focus on the display of raw data (for example, the book *Graphics of Large Datasets: Visualizing a Million*, by Unwin et al., 2006) but graphical visualization can also be important in understanding and checking the fit of complex models (Gelman, 2003, 2004, Buja et al., 2009, Wickham et al., 2010) and for exploration across models (Unwin, Volinsky, and Winkler, 2003, Urbanek, 2006, Wickham, 2006).

The present article explores two related but distinct practices which we define in ideal forms:

(1) *Statistical data visualization*, which is focused not on visual appeal but on facilitating an understanding of patterns in an applied problem (recall the Discovery goals listed above), both in directing readers to specific information and allowing the readers to see for themselves.

(2) *Infographics*, which ideally should be attractive, grab one's attention, tell a story and encourage the viewer to think about a particular dataset, both as individual measurements and as a representation of larger patterns (as in our Communication goals).

Statistical visualization could itself be separated into visualization of data and of models, but we prefer to lump these together, partly to sharpen the contrast with non-statistical infographics and partly because data and model visualization go together: the most effective data graphs can often be viewed as implicit or explicit comparisons to models (Gelman, 2004); and, conversely, when graphing models, we prefer to display data alongside to give a sense of model fit.

Infographics and statistical visualization are both important, and we should respect the different goals that they address.

The “5 best data visualization projects of the year”

We illustrate our graphical ideas on the examples that motivated these thoughts, the list by statistician and Flowing Data blogger Nathan Yau of the best data visualizations of 2008. In

using these examples to explore different perspectives on data graphics, we are not in any way trying to criticize the graphics design of these projects. We recognize that these projects look better and in many ways function better than most if not all of the graphics that we have made. One of our key goals in writing this article is ultimately to strengthen connections between statisticians and graphics designers. When we engage in criticism here, the purpose is to highlight our differences in perspectives.

Wordle. This popular program created by Jonathan Feinberg for displaying word frequencies received Yau's honorable mention. Figure 5 shows an example of Wordle in action.

Yau writes, perceptively:

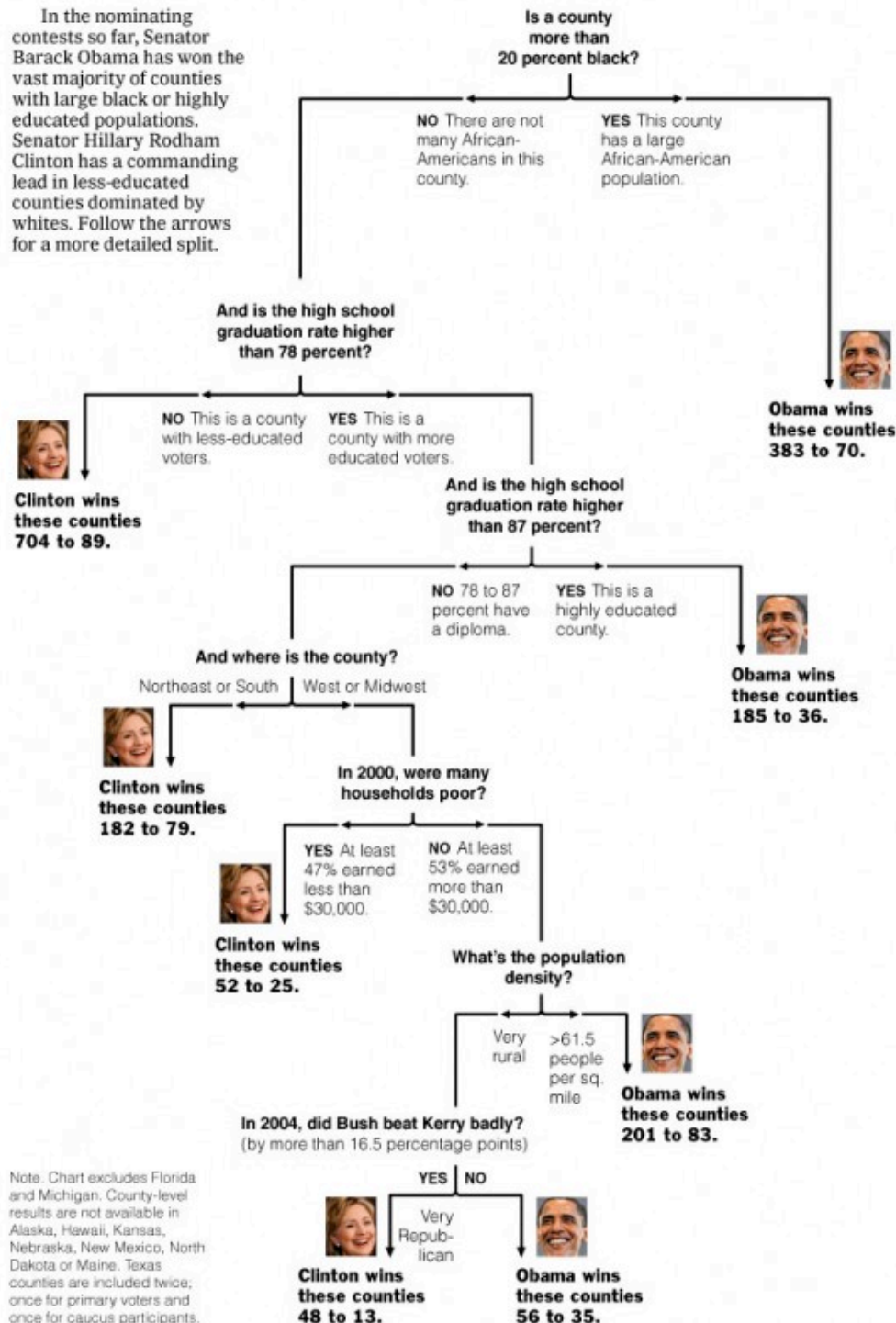
It's hard to say what exactly made Wordle so popular, but I [Yau] think it was a mix of randomness, aesthetics, and customization options.

From our perspective as creators and users of statistical graphics, we see Wordle as conveying a small but important amount of information (the most common words in a document and their approximate relative frequencies) in what we see as an eye-catching but confusing way. The advantages lie in the lack of ordering (you may see something you don't expect) and the fitting in of the most frequent substantive words. The disadvantages lie in the lack of ordering (every Wordle displays looks very different and you may miss important words due to color, orientation or position) and fitting in of too many words. The alternative of listing the words in order of frequency and also using size and color would be consistent, but would not encourage the kind of random word linking which Wordle encourages.

Wordle certainly grabs attention and stimulates thought (goal 6) and it provides an overview (goal 1). To some extent it encourages exploration (goal 3), since it is easy to try other variants, but the exploration is random rather than under the reader's control. But our biggest problem with Wordle is that study of the image sends the viewer not toward a deeper understanding of the original data but rather toward engagement with Wordle itself. The how-did-they-do-it and how-does-it-work aspects of Wordle overwhelm the data it was intended to display.

Decision Tree: The Obama-Clinton Divide

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.



Note: Chart excludes Florida and Michigan. County-level results are not available in Alaska, Hawaii, Kansas, Nebraska, New Mexico, North Dakota or Maine. Texas counties are included twice; once for primary voters and once for caucus participants.

Sources: Election results via The Associated Press; Census Bureau; Dave Leip's Atlas of U.S. Presidential Elections

AMANDA COX/
THE NEW YORK TIMES

Figure 6. A visualization of voting in the 2008 primary election campaign that we dislike because it displays a model that we do not think usefully summarizes political attitudes or behavior.

votes.) This visualization gives a kind of overview (goal 1), though we have no idea how good it is. It doesn't encourage exploration and it implies the divisions into groups are simpler than they really are, so we can strike goals 2 and 3. It does seem to be trying to tell a story (goal 5) and the photos probably help attract the reader's attention (goal 6).

Radiohead music video. This cool four-minute video displays a movie reconstructed from data from a three-dimensional scanner. Figure 7 shows a still shot which captures the general look of the images (but the video is much more impressive).

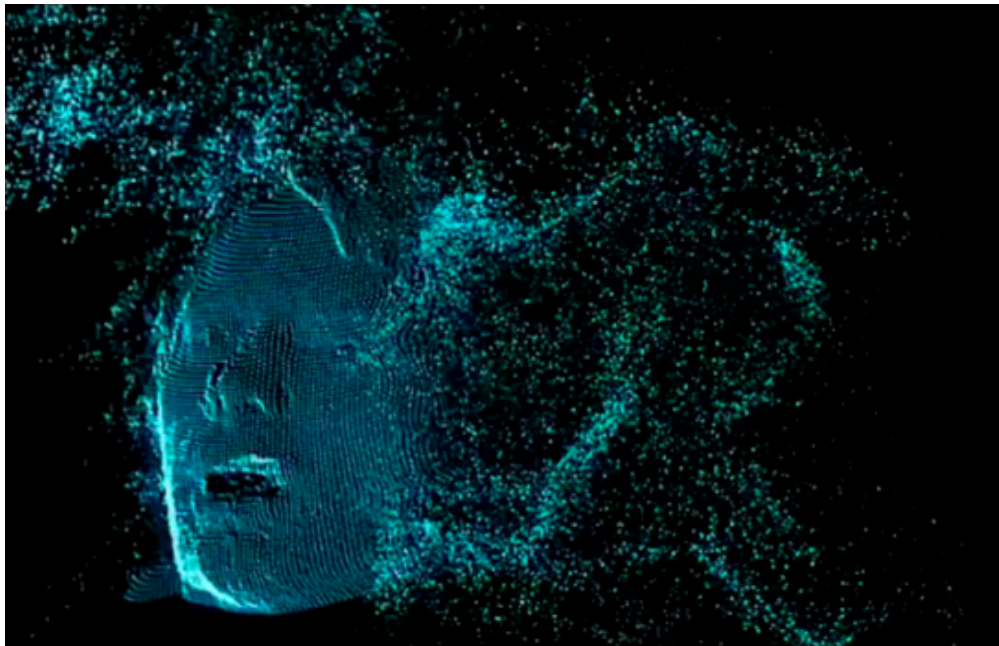


Figure 7. This snapshot from a Radiohead music video is an attractive demonstration of graphics technology but is not itself a data visualization.

This is pretty; we just wouldn't call this sort of thing "data visualization." Rather, it is a use of data visualization *tools* to make art, and it's also a demonstration of statistical methods of image reconstruction. Both of these are great but seem to us to fall into a different category than statistical graphics. Maybe the relevant point here is that graphics can be identified by their technical tools as much as by their statistical goals. In that sense, the Radiohead video, 3-D

cartoons, and all sorts of computer graphics (many of which are statistically-based) serve as baselines for us in thinking about the possibilities of modern imaging, of which statistical graphics is a small subset.

Box office streamgraphs. Yau writes this of Lee Byron’s fascinating-looking stacked graphs of movie ticket sales:

Discussion burst out across the Web—about the technique and what people were seeing in the data—that I am convinced would not have come about if instead of a Streamgraph, they used say, a stacked bar chart.

Figure 8 shows a portion of the streamgraph. The online version is bigger and better, with interactive features that allow you to scroll over time, search for movies, and click on parts of the image to get details. And the graph does convey information; as Yau (2008a) writes, “You can see Oscar contenders attract a smaller audience than the holiday and summer blockbusters and kind of slowly build an audience.” Well, does it really? How can you tell which films were Oscar contenders and whether they had that kind of pattern?

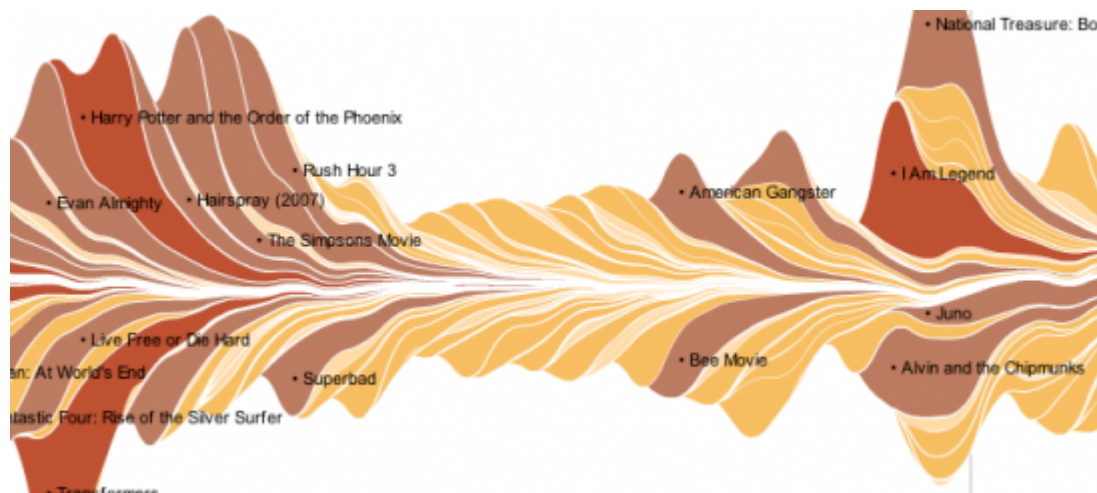


Figure 8. This information-rich display of movie box-office receipts is attractive and intriguing but, to us, is flawed as a data visualization because it is difficult to compare most of the quantities being displayed.

The visualization does indeed look cool, but the strategy of stacking the curves on top of each other makes the visuals for individual films almost impossible to interpret. We would prefer two graphs, one showing total movie sales over time (and thus capturing the overall shape of the curve) and another showing the trajectories for the individual movies, possibly identifying Oscar contenders by color.

In this case we believe the designers made a common error of statistical graphics: trying to cram into a single graph what can be better displayed in two. On the other hand they achieved the goal of grabbing attention superbly, it's just that they did not achieve any other goal.

I Want You to Want Me. In featuring this image and an accompanying video by Jonathan Harris and Sep Kamvar (see Figure 9), Yau writes that “this blend of art, computer science, and mathematics is beautiful.”



Figure 9. This data-based video is visually appealing but does not give an overview of the quantitative information that is essential for a statistical graphic.

We agree that the *I Want You to Want Me* video is visually appealing, but, again, we don't really see it as an effective way to convey the data. It's more of a way to get attention, but then we would want a pointer toward a better data visualization to learn more. Our point here is not to criticize this work as graphic design or as art but rather to focus on the different goals that we have in data display.

That said, we only will learn by trying new things, and that, for graphics, it's good to have new tools. Who knows if the eye-catching graphics you display in *I Want You to Want Me* might be altered to display data in some informative way? So we don't want to discourage experimentation. The perils come when a snazzy display is used to obscure information. As statisticians, we should have a way of pointing this out—of connecting the visuals to the goals of subject-matter understanding—without alienating people and obscuring our own message.

Britain from Above. Figure 10 shows a still photo from a series of videos that Nathan Yau selected as the top visualization project from 2008.



Figure 10. A snapshot from a video of air traffic over Britain that could conceivably be part of an effective data visualization if it were placed in the context of relevant summary displays.

The videos show air traffic over Britain, and they represent an impressive computing and statistical achievement, putting together information from satellite images to visualize the transportation network and, in Yau's words, "bring data to life." Although this work is not a statistical graphic in the traditional sense, we could imagine it being combined with some data reductions (for example, average flows of different kinds) to achieve statistical goals of communication and discovery. From a statistical point of view it is hard to argue in favor of the distortion that occurs when presenting Britain from an angle rather than directly from above, though the idea is presumably to suggest you are actually in a plane looking at the flight traces. At the end of the video, attention is drawn to areas which no flight paths cross, possibly locations

of secret military installations or high security prisons. This visualization certainly satisfies goals 5 (telling a story) and 6 (grabbing attention), to some extent it communicates some information (e.g., the gaps), but statistical goals are not met.

Overview. The “best data visualizations of the year” are eye-catching graphics that in several cases use state-of-the-art methods in statistics and computer science, while at the same time not attempting to achieve traditional goals of statistical graphics. We would characterize all these graphs as visually attractive and data-related, so at the very least they can serve as inspirations to statisticians and other designers who are thinking about future data display challenges.

Statistical problems with other highly praised infographics

Even infographics that are clever and beautiful can have problems when viewed as statistical data visualizations. We illustrate with some examples that have appeared recently in the press.

Plane crashes. Figure 11, created by David McCandless (2009), was part of a series of graphs that won the *Guardian* newspaper’s Visualization Contest.

Final Destination
Density of fatal accidents 1942-2009

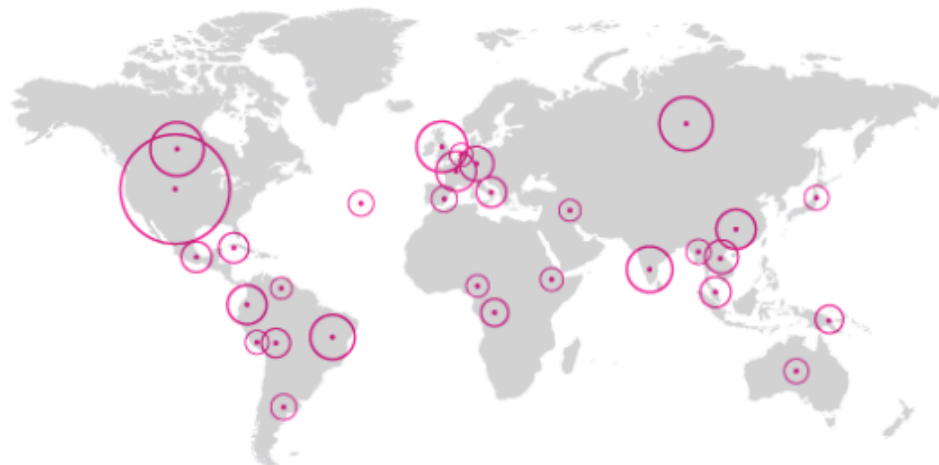


Figure 11. This award-winning infographic of air accidents is nearly useless in that it presents events unnormalized by base rates. The graph is not improved by redundantly displaying each country with both a dot and a circle.

Even though we agree fully with Lakeland's comments, we are wary of criticizing this sort of visualization. For the goal of conveying information, it is horrible, but for sparking interest in their topics and motivating readers to look carefully at the numbers, maybe it is still useful. To give it a prize in a visualization contest, though . . . well, we wouldn't do that. That would be like deciding who won the Indy 500 by picking the car with the snazziest paint job. Might it help to shade the circles according to their accident rate? Even then we would still be left with the problem that the graphic shows the density by country, not by location, so the circle for the USA doesn't even include any of the East coast, where probably many of the accidents occurred. In the case of Russia, it may even be the case that none of the Russian accidents occurred at locations covered by its circle! This display grabs attention but does little else.

Florence Nightingale's coxcomb. Consider the famous image drawn by Florence Nightingale (1858), which is often considered as an exemplar of data display (see Figure 12). In a recent discussion of the coxcomb plot, Rehmeier (2008) writes:

The conventional way of presenting this information would have been a bar graph, which William Playfair had created a few decades earlier. Nightingale may have preferred the coxcomb graphic to the bar graph because it places the same month in different years in the same position on the circle, allowing for easy comparison across seasons. It also makes for an arresting image. She said her coxcomb graph was designed “to affect thro’ the Eyes what we fail to convey to the public through their word-proof ears.”

Given the context, the graph is impressive and important (see Small, 1998, for further background on this and other work of Nightingale). But given what we know today, we would prefer it as a line plot (not a bar graph, which, as Rehmeier notes, unfortunately is indeed the default choice for many if not most producers of graphs).

In theory, the circular plot could have advantages for displaying monthly regularities, but, first, such regularities are not so strong in this dataset (or, if they are, the circular plot is not a good

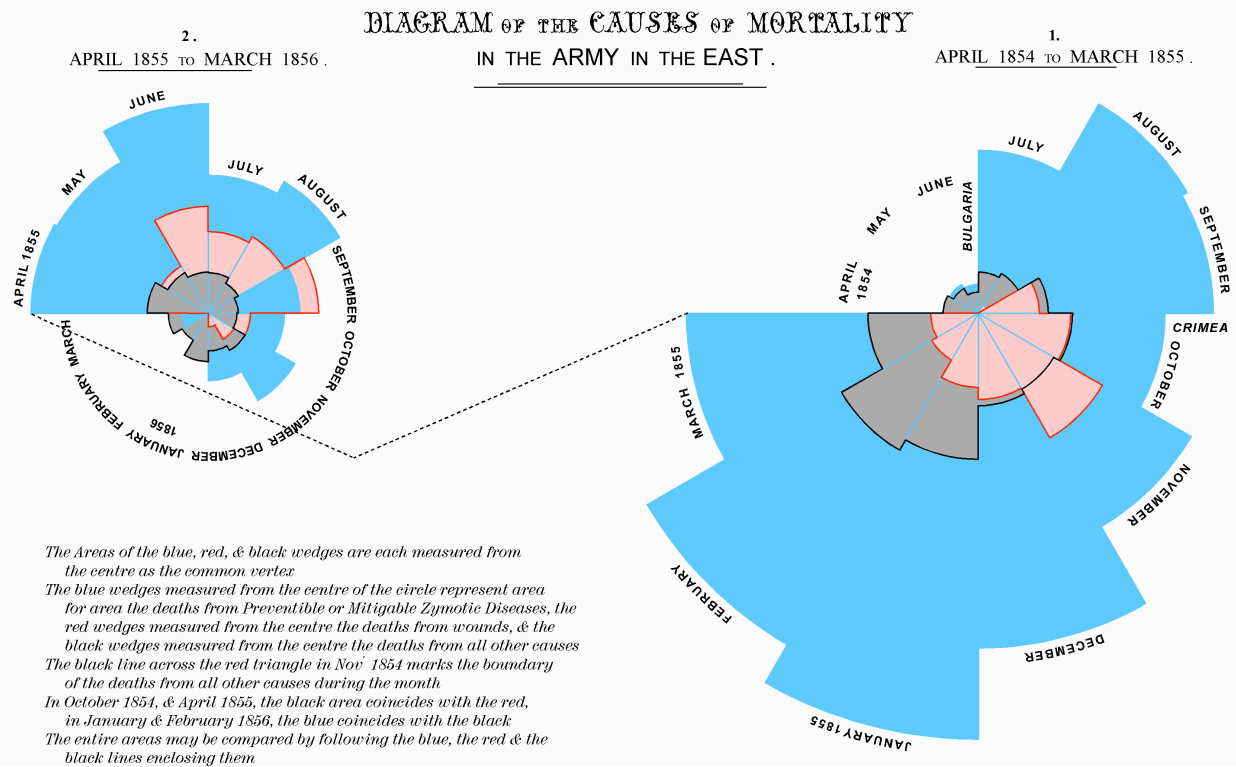


Figure 12. Florence Nightingale's celebrated circular plot of Crimean War mortality is a landmark in infographics, but a modern-day statistician would prefer to display such data using simple time-series plots such as shown in Figure 13. The ability of the circular plot to line up months in different years obscures the patterns in the data and offers little if any benefit in this example. At the same time, the attractive and unusual appearance of Nightingale's graph—even its puzzle-like nature—might well have helped draw attention to the public health problems she was working on. This illustrates the differing goals of statistical graphics (intended for understanding patterns in data and departures from these patterns) and information visualization (intended to attract attention and stimulate thought in people who might otherwise not have been interested in the topic).

way to reveal them!), and, second, one could always do the comparisons by month better by overlaying several years with line plots. We find it surprising that Rehmeyer claims the display allows “easy comparison across seasons.” The circular structure of the above image is indeed beautiful but we don't think it conveys the information very well. In addition, each color is measured from the center, so that the total numbers of deaths per month from the three causes is impossible to determine. Overlaps complicate the interpretation further. It is also unusual that the data for the first twelve months are drawn in the right plot and the data for the second twelve months in the left plot. Conventions help and should only be disregarded for very good reason.

Figure 13 shows an alternative we prepared. The upper graphs show the dramatic rise and fall in the death rates from “zymotic” (basically, infectious) diseases, while emphasizing the comparatively low rates of deaths from other causes. The lower graph shows that the army size varied, but not nearly as much. The time series for the army size has purposely been drawn as a bar graph as it represents counts and also because this makes it clear that we are dealing with a different kind of series.⁶

Our point here is not to claim that our plots are optimal but rather to contrast the virtues of Nightingale’s display—its unique appearance and the visual appeal of the areas and the twelve-

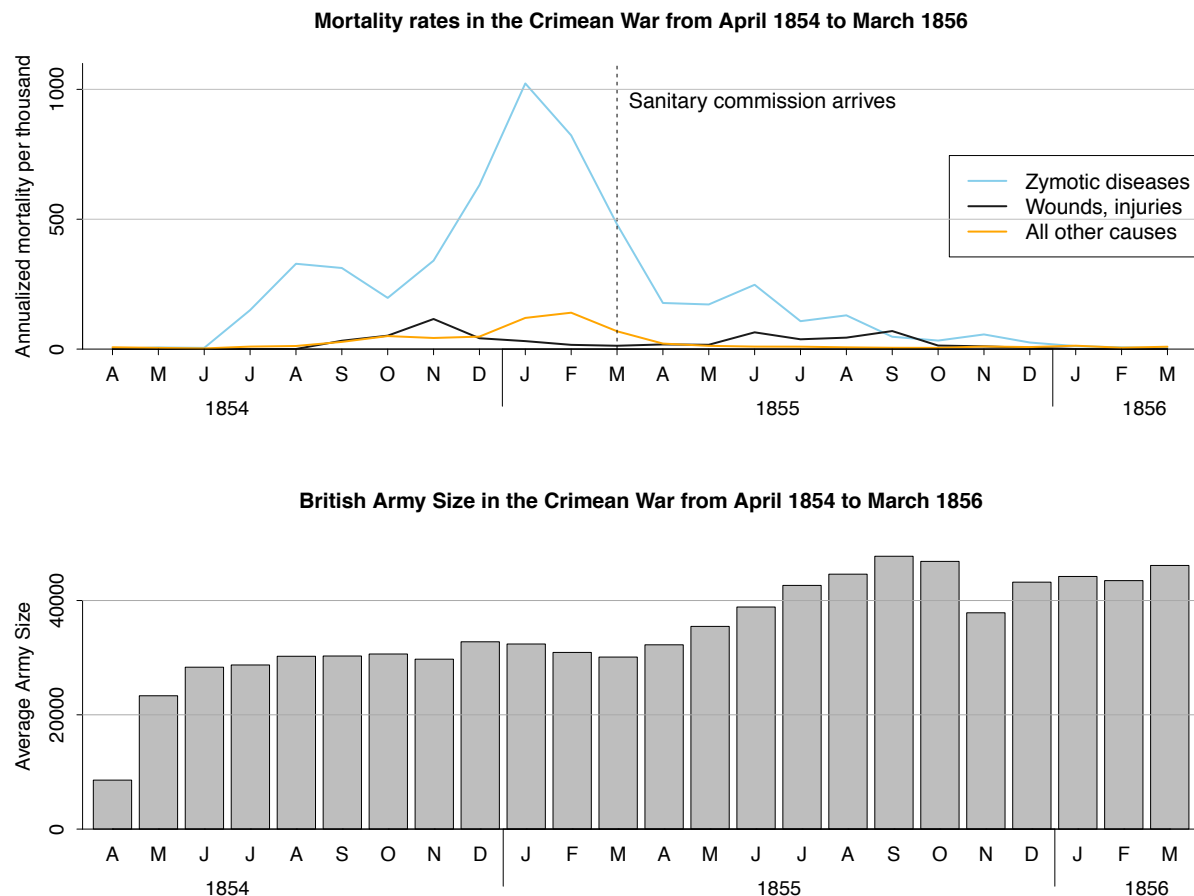


Figure 13. The Crimean War mortality data as time series graphs. The use of different forms (line plot for death rates and bars for army size) is a visual cue that two sorts of data are being presented: monthly rates (multiplied by 12 to be annualized, following Nightingale’s calculations) and absolute numbers.

⁶ The data used in Nightingale’s graph are available at <http://understandinguncertainty.org/node/214>

month circles—with our strategy, characteristic of statistical graphics, to choose a bland, conventional background to allow the reader to more clearly see the changes within and between the time series.

In the language of the present paper, the Nightingale graph is an excellent example of “infographics”—it is attractive, grabs one's attention, and gets you thinking—but it is not so great as “statistical graphics” in that it does not directly facilitate a deeper understanding of the data. In Nightingale’s political context, the goal of attracting attention was arguably much more important than the goal of understanding and communicating subtle patterns in the data.

If we were presenting these alternatives on the web, it could be appealing to offer both the Infovis and statistical graphics. The display would start with the Nightingale plot to draw attention. Then, clicking on the coxcomb would switch the display to a more statistical graphic such as our Figure 13. Finally, another click could bring up a spreadsheet with the data. Another option could be to include decoration to draw attention, and have it fade away on clicking to enable the reader to concentrate on the data information.

We do not claim that our graphs are better than Nightingale’s classic; rather, the two displays serve different purposes, and in the modern high-bandwidth era, there is room for both. The first step is to understand the different goals involved in a graphical display.

Health spending and life expectancy. Figure 14, produced by Oliver Uberti in 2009 for *National Geographic*, dramatizes that Americans spend much more on health care, compared to residents of a range of other countries, without seeing any apparent benefit in terms of life expectancy.

Figure 15 (from Gelman, 2009c) contains the same information while following the standard principles of statistical graphics. The standard way to display two variables is a scatterplot, in this case health-care spending vs. life expectancy. (The original display also contains information on frequency of doctor visits, but this third variable appears to be a minor part of the story.) The scatterplot reveals the arbitrariness of the scaling of the parallel coordinate plot in this example. In particular, the original graph gives a sense of convergence, that spending is all over the map

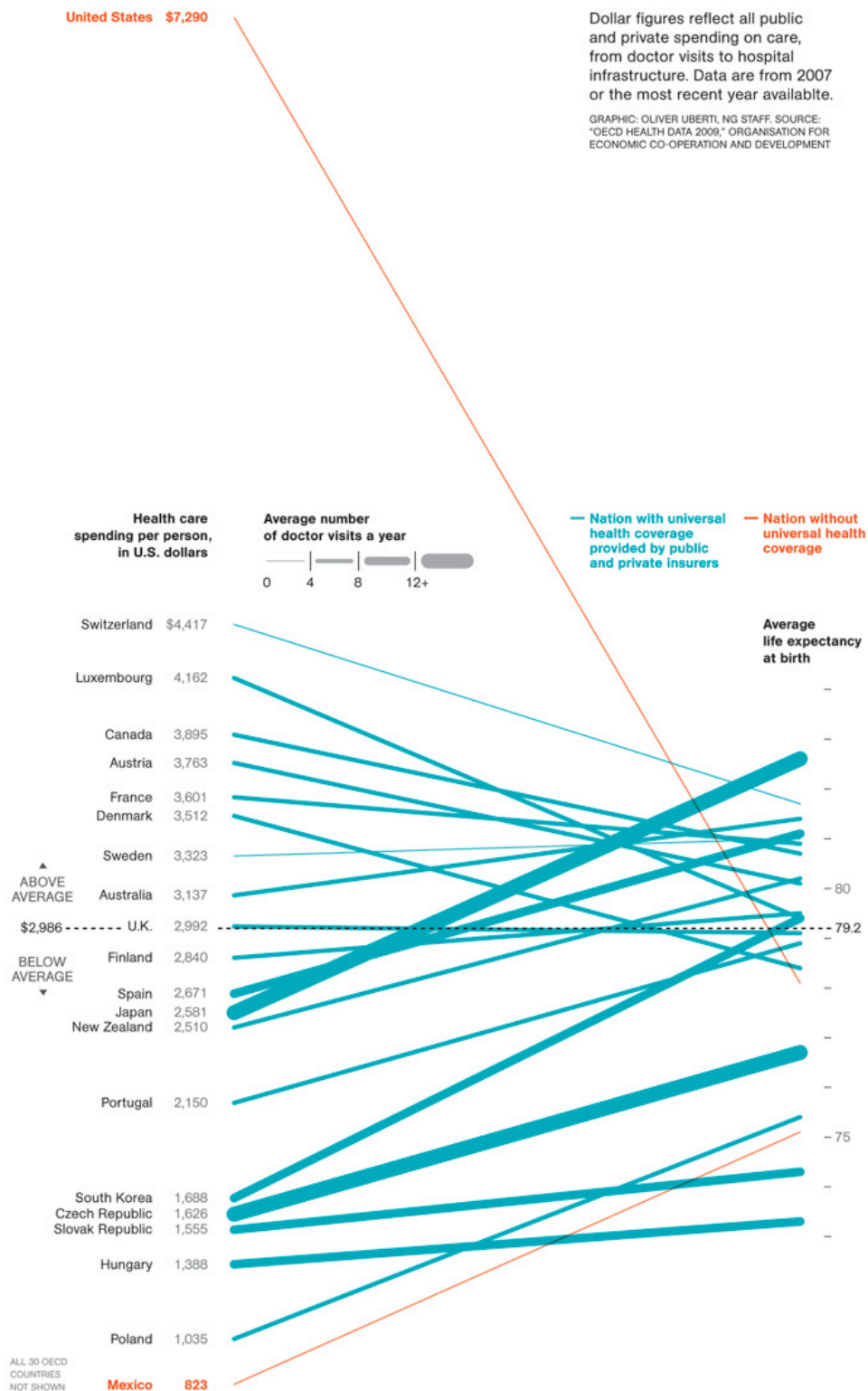


Figure 14. A dramatic display showing that the United States spends much more on health care per person than other developed countries but remains in the middle of the pack in life expectancy. From the standpoint of statistical data visualization, we would prefer a simple scatterplot such as shown in Figure 15, but Figure 14 does a great job of dramatizing to non-statisticians the outlier status of the U.S.

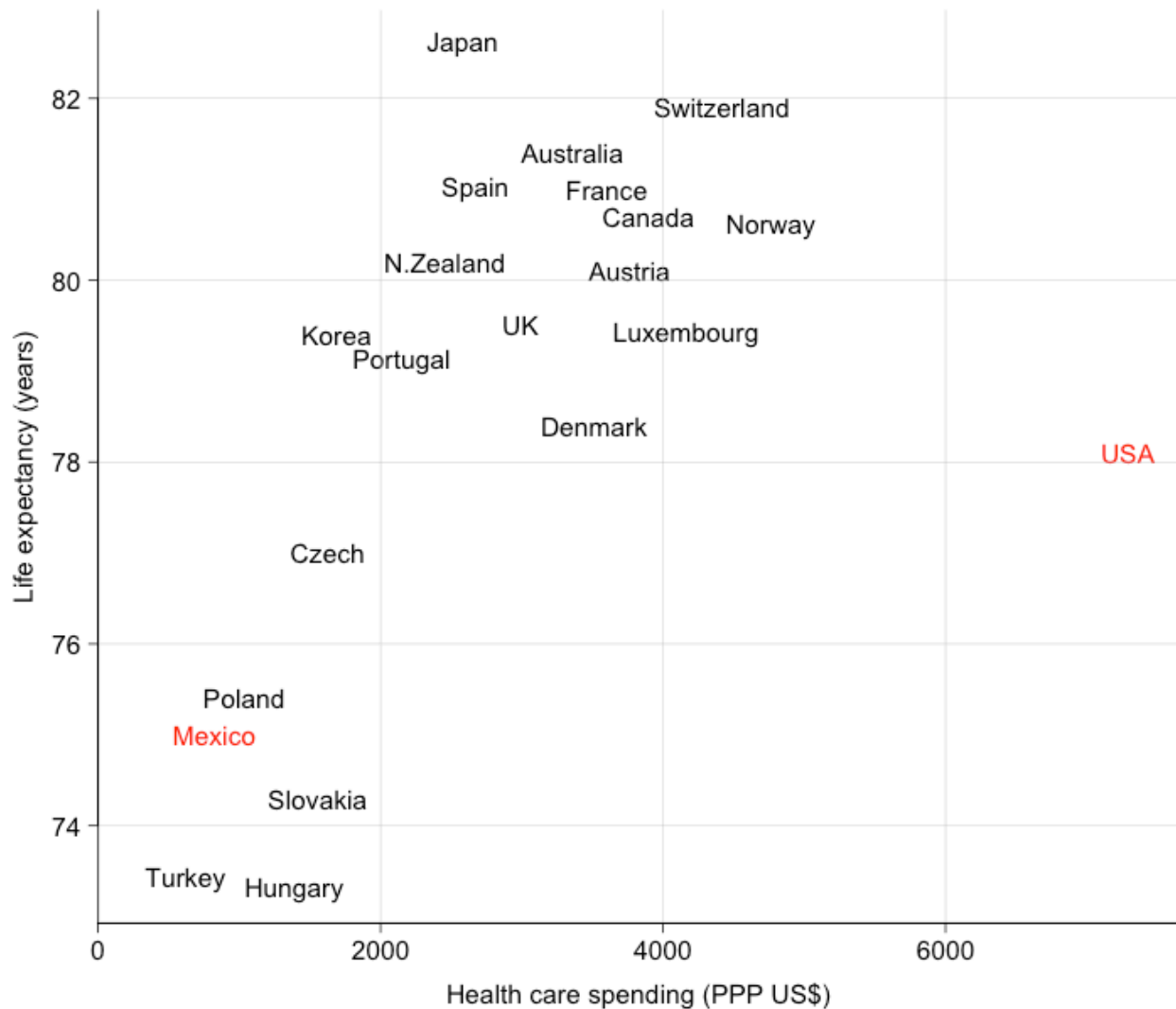


Figure 15. A scatterplot that displays the health-spending/life-expectancy data in Figure 14, but more transparently, more informatively, and in less space. Figure 14 is a dramatic display, while Figure 15 shows the pattern of the two variables more clearly. The two displays serve different goals, and an online display might start with Figure 14 and then reveal Figure 15 with a click.

but all countries have pretty much the same life expectancy—look at the way the lines converge to a narrow zone as you follow the lines from the left to the right of the plot.

But once you remove the U.S., there is a strong correlation between spending and life expectancy, and this jumps out of the scatterplot, much more than in that eye-catching parallel coordinate plot. (For simplicity, we have removed the data on doctor visits from our plot; a display of that additional information reveals no particular relation with the two major variables.)

But data graphs are not just judged on informativeness. Another consideration is novelty. The scatterplot in Figure 15 looks like lots of other graphs we've all seen. This is a plus—familiar graphical forms are easier to read—but also a minus, in that it probably looks boring to many readers. The parallel-coordinate plot in Figure 14 is not really the right choice for the goal of conveying information in this case, but it's exciting and new to many people, and that's maybe why one of the commentators at the *National Geographic* website hailed it as “a masterpiece of succinct communication.” The goal is not just to display information but also to grab the eye.

Ultimately, we think the solution is to do both — in this case, to make a scatterplot in some pretty, eye-catching way. Not being experts on graphic design, we just did the first part and will leave it to others to figure out good ways of making the display more eye-catching.

How to win in Afghanistan. The graph reproduced in Figure 16 was prepared by a military contractor for the Office of the Joint Chiefs of Staff. Neither of us has ever been involved in any planning more complicated than setting up a M.A. program, and that had a budget approximately one-zillionth that of the Afghan war. Without any experience in large projects, we will limit our comments to the graph itself.

To start, we think the graph would be improved by making the arrows lighter—gray rather than black—and maybe reducing the number of arrows overall. We understand the goals of showing the connections between the nodes, but as it is, the graph is dominated by the tangle of lines.

A larger problem is that the picture gives no sense of priorities. All the items are the same size and it is not clear where the focus should be. The full presentation (PA Consulting Group, 2008) puts all the nodes in context and makes the story clearer. But we can't really see what is gained from the image. We can understand the value of a complicated graph showing suppliers and contractors and purchasers and so forth, but we don't see what you get out of this sort of map where most of the nodes are vaguely-defined concepts.⁷

⁷ We looked carefully at the graph and could only find one node that is an orphan (that is, with no arrows pointing toward it). That node: “Media Sensationalism Bias.” Perhaps there could be another node leading to it, labeled “Pay \$\$ to friendly journalists.”

Afghanistan Stability / COIN Dynamics – DEVELOPMENT

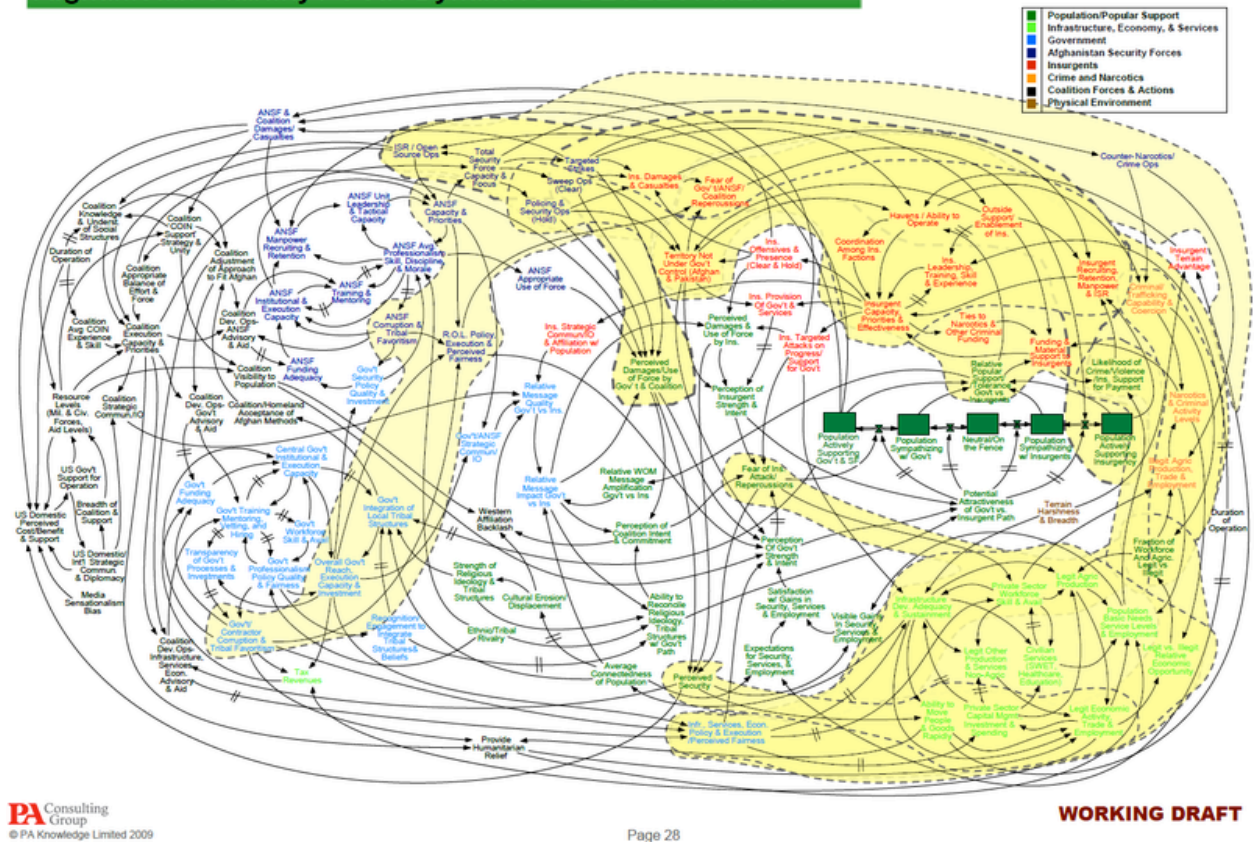


Figure 16. This flowchart—part of a powerpoint presentation from a military contractor—effectively displays the complexity of planning during the Afghan war but would not be good if the goal is data or model visualization.

As noted above, we are complete strangers to the world of military planning, and we are reacting based on our understanding of graphical display. We are suspicious of the combination of a complex display and lack of precision in the details. Similarly, we suspect the graph displayed above does not do much to directly help the planning for Afghanistan, but it certainly does a good job of conveying the complexity of the situation! Maybe that was the point.

As statisticians, one might simply label Figure 16 as “junk” and leave it at that. Our point, though, is not merely to offer judgment (although we are happy to use our professional expertise in that way as necessary) but to use this example, quite different from the usual histograms and scatterplots of statistics texts, to consider the goals of graphical displays.

The Afghanistan flow chart is neither a data visualization nor a statistical graph, but it raises a point that is relevant to our discussion: this display is not useful for conveying information, but it is useful for giving context. If someone mentions a concept, then you find it on the display and see what other concepts are related to it. In that sense, it is more like a (non-statistical) map than a graph. At least, that's the theory. In practice, we are skeptical that the above display is useful even as a conceptual map, but we will leave that for the subject-matter experts to judge. Our point here is to connect the visual format of the image to the goals that motivated its creation. It is through considering these goals that we as statisticians can better offer constructive criticism.

Static statistical graphics: timeless or simply old-fashioned?

Most of the “best” data visualizations used interaction and dynamic graphics in some way. Although interactive graphics are increasingly being used for data analysis (Buja et al., 1996), they are still not used much for display of results. What interactive presentations there are to be found on the web, with occasional exceptional examples, are still in an early stage of development, we can expect better in the near future. For this reason, we have written mainly from the point of view of static statistical graphics, the routine graphic tools of the profession.

Eye-catching data graphics tend to use designs that are unique (or nearly so) without being strongly focused on the data being displayed. In the world of Infovis, design goals can be pursued at the expense of statistical goals. In contrast, default statistical graphics are to a large extent determined by the structure of the data (line plots for time series, histograms for univariate data, scatterplots for bivariate non-time-series data, and so forth), with various conventions such as putting predictors on the horizontal axis and outcomes on the vertical axis. Most statistical graphs look like other graphs, and statisticians often think this is a good thing. To use a literary analogy, statisticians tend to prefer Orwell's (1946) dictum that “good prose is like a window pane,” while infographics experts prefer the pyrotechnic style of a Martin Amis or the make-the-reader-work style of a Chris Ware.

One of the challenges of Edward Tufte's books is that people read Tufte (1983, 1990) and then want to make cool graphs of their own. But cool like Amis, or cool like Orwell? These are two different directions, as illustrated by the health-care spending graphs above.

Two qualifications are needed here. First, it can take a lot of work to write clear prose, just as it can take a lot of work and a lot of practice to prepare clear graphs. Even our simple health-care scatterplot required a bit of work (and a foundation of applied experience) to look as clean as it does. Second, the vivid writing of a Martin Amis or T.S. Eliot can be fun in itself and also point the way forward: yesterday's experiments can be tomorrow's standards. Much of Ezra Pound is not so readable today, but he had a big influence.

To draw another analogy, consider pie charts, which take a lot of work to draw by hand but are trivial to construct on the computer. It should be possible to argue both of the following:

- (1) Pie charts are helpful: They have introduced millions of people to data, giving people a physical sense of numerical relationships where the data are shares of a whole.
- (2) Pie charts are a dead end: Elaborations on pie charts (3-d pie charts, exploding pie charts, and all the rest) make things worse, and they can stand in the way of more direct data displays.

For that matter, default graphics in Excel can be a useful research and presentation tool. The problem occurs when people assume that the Excel output is enough. Think of all the research papers in economics where the authors must have spent dozens of hours trying all sorts of different model specifications, dozens of hours writing and rewriting the prose of the article . . . and then spent 15 minutes making the graphs. They just don't realize that more can be done. And, from this perspective, Wordle and all the rest don't really help. As data analysts, we see a large and continuing role for traditional display tools such as line plots, and there is a place for thinking seriously about the connection of these methods to the data and inferential problems at hand. It is not all about making something that looks pretty and has data in it. It should be about presenting relevant data fairly and effectively (using appropriate scaling and encouraging informative comparisons) and using design to make the graphic more attractive and interesting.

We are not criticizing the general idea of snazzy graphics—we find Flowing Data and other infographics sites to be often inspirational. It was more that we had problems with the specific displays labeled as Best of the Year, which led us to recognize a divergence of goals. Let's praise the innovators who design wacky, eye-catching tools such as Wordle, but let's also think about how to use these tools to give us a fuller understanding of the world around us.⁸

The optimal situation would be a close cooperation between data analysts and designers. Each does what they are good at and that is maybe why we have dependable but dull graphics from data analysts and attractive but unreliable graphics from designers. More statisticians need education and encouragement in computing and design. They can't be expected to become experts, but they should have an appreciation of what experts in other fields could do for them

Statisticians prefer old-fashioned tools such as dot plots and line plots, with even the more modern innovations (for example, small multiples) being decades old, though implemented more effectively in a modern era of high-resolution color graphics. Infovis researchers are working much more on the technological bleeding edge.

As statisticians, we're not quite sure how to think about our own preferences. Are dot plots and line plots really the best possible choice? Or does it just take many decades before new graphical technology can become usefully implemented as statistical methodology?

The Baby Name Wizard—an example we like

We conclude our series of examples with an interactive tool created by Laura and Martin Wattenberg (2005) that, in our view, combines the eye-catching beauty of the best Infovis examples with the directness and simplicity of the best statistical data visualizations. Figure 19 shows a screenshot of the most popular names beginning with “Cr.”

⁸ Throughout, we are talking only about graphs made in good faith. Just about any graphical technique can be abused, and just about any graphical technique can become a mess, if placed in the hands of a suitably dishonest or incompetent person. These concerns are important but are outside the scope of the present article.

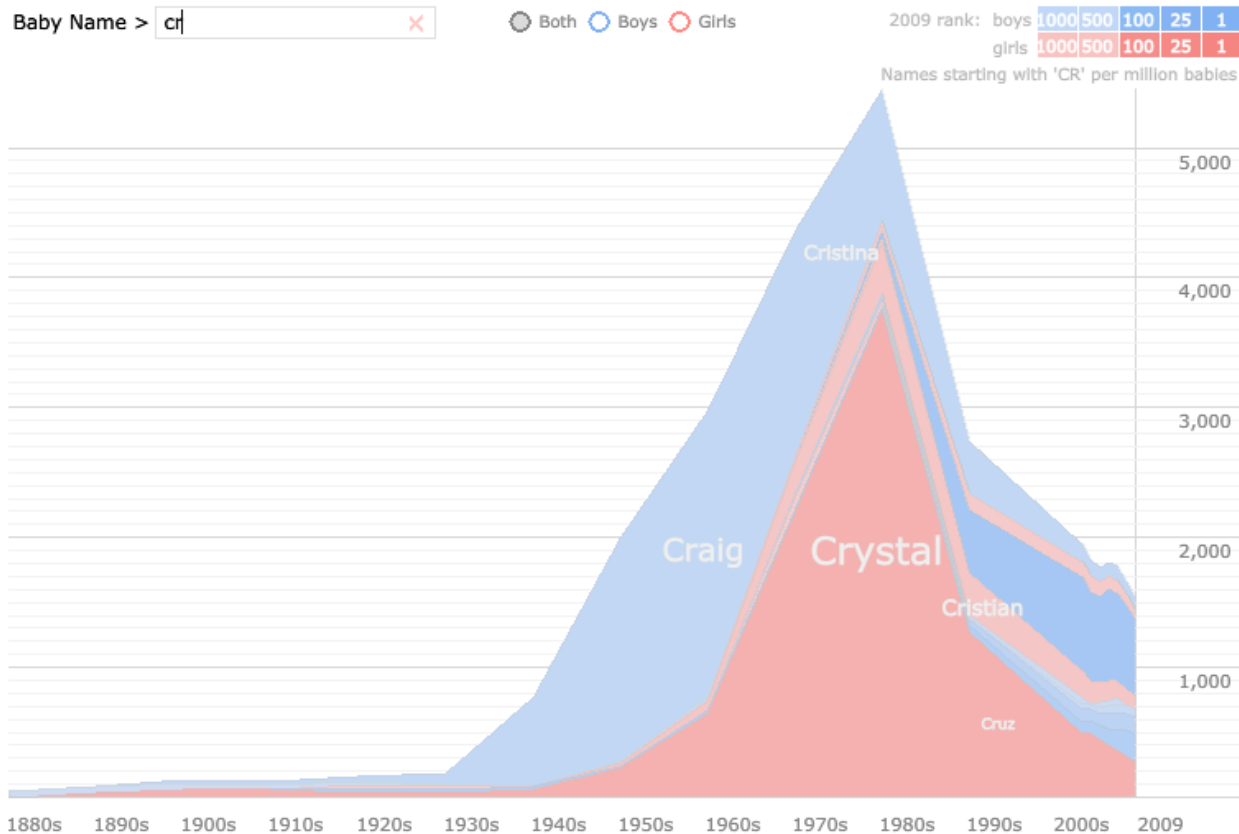


Figure 19. A screenshot from the Baby Name Wizard, an interactive tool for visualizing the popularity of first names over time. The Baby Name Wizard combines the visual appeal and excitement of the best Infovis work with the clarity of the best statistical data visualizations.

Beyond being fun and addictive (once we hit this website, we couldn't stop typing in letter combinations to see names and their trends), the display follows the principles of statistical data display as recommended by Tufte, Cleveland, and others. Colors are used sparingly (to convey information rather than as decoration), axes go down to zero and are labeled clearly but gently, the names are labeled directly on the graph, and each name can be individually located by running the cursor over the interactive version of the graph which appears online. All the six goals we listed at the beginning are met by this display.

The Wattenbergs have also used their database to make more traditional statistical graphics, one of which we have adapted for Figure 20 to show trends in popularity of different *last letters* of boys' names.

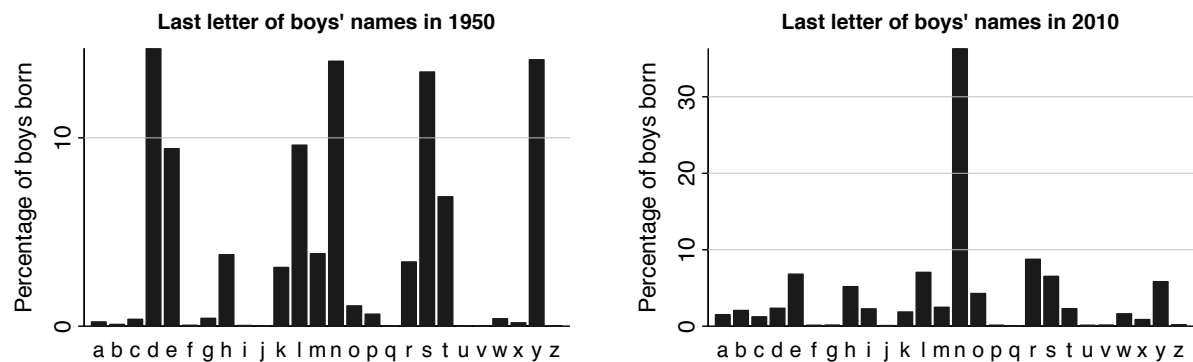


Figure 20. Histograms from the baby name database (adapted from Wattenberg, 2007), showing a dramatic change in boys' names in the past sixty years. These bare-bones graphs are effective in revealing and displaying a fascinating and unexpected pattern in the data and illustrate that the goals of exploratory and presentation graphics can often overlap.

The quick story is that at the start of the twentieth century [graph not shown here], there were about 10 last letters that dominated; sixty years ago, the number of popular last letters declined slightly, to about 6; but now, a single letter stands out: an amazing 36% of baby boys in America have names ending in N. This is, in a word, cool. As a commentator wrote, there should be some sort of award for finding the largest effect “in plain sight” that nobody has noticed before.

But, beyond pure data-coolness, what does this mean? Our story, based on the discussion of Wattenberg (2007), goes as follows. A hundred years ago, parents felt very constrained in their choice of names (especially for boys). A small set of very common names (John, William, etc.) dominated. And, beyond that, people would often choose names of male relatives. Little flexibility, a few names being extremely common, resulted in a random (in some sense) distribution of last letters.

Nowadays, parents have a lot of freedom in choosing their babies' names. As a result, there are lots and lots of names that seem acceptable, but the most common names are not as common as they were fifty or a hundred years ago. With so much choice, what do people do? Wattenberg suggests they go with popular soundalikes (for example, Aidan/Jaden/Hayden), which leads to clustering in the last letter. Even so, the pattern with N is *so* striking, there must be more to say about it.

In any case, we like the paradox: A century ago, the distribution of *names* was more concentrated but the distribution of *sounds* (as indicated by last letters) was broader. Nowadays, the distribution of names is more diffuse but the distribution of sounds is more concentrated.

Less constraint → more diffuse distribution of names → more concentrated distribution of last letters. And we gained this bit of social science insight from a powerful combination of high-tech interactive visualization, exploratory data analysis, and static statistical graphics to crisply summarize the result.

Discussion

One key difference between the two approaches is that Infovis prizes unique, distinctive displays, while statisticians are always trying to develop generic methods that have a similar look and feel across a wide range of applications. Few statisticians are trying to develop anything new; they are using the standard well-tried tools. Infovis places a high value on creativity and difference, whereas statistics is centered on objectivity and replication. Recognizability is important: every novel graph requires more work and without experience we may miss some of the secondary features. On the other hand novelty attracts attention. This is an old discussion which is not only found in data graphics, as in Wilkinson (2005) where he takes as one of his principles “Less is more,” but also, for instance, in architecture, where you can find the extreme view associated with Loos that ornamentation is a crime.

Another important difference is in the expected audience. Statisticians assume that their viewers are already interested and want to provide structured information, often a carefully prepared argument. For statisticians, graphics are part of an explanation. Even exploratory analysis typically has a clear structure. In contrast, Infovis designers want to draw attention to their graphics and thus to the subject matter. For them, graphics are more of a door opener. This is reflected in how both groups use interactivity. Infovis graphics often have video or animation, which adds to the attraction and engages viewers, who can control the animation and perhaps change colors or shapes, allowing different perspectives on the images. Statisticians, when they

use interactivity, use it to link to other graphics or to models, allowing viewers to explore the argument.

We feel that, of all the visualizations discussed in this article, the Baby Name Wizard did the best at conveying information in an attractive, interesting, and open-ended way. If, for a moment, you accept this judgment, it leads to the surprising conclusion that the five best developments from 2008 were lower in quality than something that was done in 2005—which is a bit of a disappointment given the improvements in technology during the intervening period. On the other hand, technology always lags. Most of the graphs in our book about U.S. politics (Gelman et al., 2009d) are line plots or scatterplots that could have been made 50 or 100 years ago or even earlier (had the data been available). On the other hand, back then such graphs took more effort to make—you couldn't play around and make thousands of graphs and then choose hundreds for a book. In some ways, this was good—it's not such a bad idea to have to think hard before acting—but in practice the result at the time was that tables were common, graphs were rare, and it was more difficult for readers—or researchers—to integrate large amounts of information. Even now, tables predominate over graphs in statistics journals, despite various exhortations by statisticians otherwise.

We have talked a lot about the different goals involved in creating statistical graphics, and so it is appropriate to end our discussion with a statement of our goals in writing this article. We would like to broaden the communication between graphic designers, software designers, statisticians, and users of statistical methods. This is an old point but one that bears repeating. By recognizing the diversity of goals involved in data graphics, developers and users alike might be in a better position to create eye-catching as well as informative visualizations of data and models, even if both these goals are not typically achieved in a single display.

Progress in infovis is important—should be important—to statisticians, even if the newest and prettiest developments are not yet particularly effective at revealing patterns in data. Today's infographic may become tomorrow's statistical display. Consider a hundred years ago or more, when the standard statistical graphic was the data table. With care and effort, tables can be informative, readable, and even exciting—take a look, for example, at an old volume of the

Statistical Abstract of the United States. Back in the nineteenth and early twentieth century, there were some very attractive time series graphs, scatter plots, maps and more complicated statistical graphics—but these were artisan work, the infographics of their day. They were not really used routinely enough for statistical researchers to get a sense of what worked and what did not (and they often misfired, as with Florence Nightingale’s coxcomb plot above, which, although beautiful, is ultimately less informative than a simple time series plot). But progress on these led to our current state in which graphs, not tables, are the standard in data communication. Perhaps the infographics of today will evolve into the statistical data visualization tools of future decades, and we hope our discussion of goals and examples will help move this process along.

Ultimately the interpretation of a graph is a joint product of the data, the designer, and the viewer. With the increasing prominence of innovative infographics in the news media and the web, viewers are changing their expectations of data presentation, and, more than ever before, statisticians should consider the diversity of means to achieving the very different goals of attracting attention, displaying patterns, displaying data in a way that allows for discovery, and getting viewers intellectually involved with data. As is illustrated in the historical reviews such as Wainer (1997) and Friendly (2006), there is a centuries-long tradition of data graphics that are both informative and beautiful. We should seek to continue this collective endeavor, and we hope the present article sparks a discussion among statisticians, computer scientists, graphic designers, psychologists, and others who are interested in the graphical presentation of data and inferences.

References

- Barabasi, A. L. (2010). *Bursts: The Hidden Pattern Behind Everything We Do*. New York: Dutton.
- Bateman, S., Mandryk, R.L., Gutwin, C., Genest, A.M., McDine, D., Brooks, C. 2010. Useful junk? The effects of visual embellishment on comprehension and memorability of charts. In *ACM Conference on Human Factors in Computing Systems (CHI 2010)*, Atlanta. 2573-2582.
- BBC (2008). Britain from above. Broadcast from 10 Aug.
<http://www.bbc.co.uk/britainfromabove/>
- Bloch, M., Byron, L., Carter, S., and Cox, A. (2008). The ebb and flow of movies: Box office receipts 1986-2008. *New York Times*, 23 Feb.
http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html
- Buja, A., Cook, D., and Swayne, D. (1996). Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics* 5, 78-99.
- Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E. K., Swayne, D. F., and Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A* 367, 4361-4383.
- Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. San Francisco: Morgan Kaufmann.
- Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. New York: Chapman and Hall.
- Chatfield, C. (1995). *Problem Solving: A Statistician's Guide*. London: Chapman and Hall.

Cleveland, W. S. (1985). *The Elements of Graphing Data*. Pacific Grove, Calif.: Wadsworth.

Cox, A. (2008). Decision tree: The Obama-Clinton divide. *New York Times*, 16 Apr.
<http://graphics8.nytimes.com/images/2008/04/16/us/0416-nat-subOBAMA.jpg>

Feinberg, J. (2009). Wordle. <http://www.wordle.net/>

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. London: Oliver & Boyd.

Friendly, M. (2006). A brief history of data visualization. In *Handbook of Computational Statistics: Data Visualization*, volume 3, ed. C. Chen, W. Hardle, and A. Unwin.

Friendly, M., and Kwan, E. (2003). Effect ordering for data displays. *Computational Statistics and Data Analysis* 43, 500-539.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review* 71, 369-382.

Gelman, A. (2004). Exploratory data analysis for complex models (with discussion). *Journal of Computational and Graphical Statistics* 13, 755-787.

Gelman, A. (2009a). More on data visualization, beauty, etc. Statistical Modeling, Causal Inference, and Social Science blog, 22 Apr.

http://www.stat.columbia.edu/~cook/movabletype/archives/2009/04/more_on_data_vi.html

Gelman, A. (2009b). A data visualization manifesto. Statistical Modeling, Causal Inference, and Social Science blog, 31 May.

http://www.stat.columbia.edu/~cook/movabletype/archives/2010/05/a_data_visualiz.html

Gelman, A. (2009c). Healthcare spending and life expectancy: a comparison of graphs. Statistical Modeling, Causal Inference, and Social Science blog, 30 Dec. http://www.stat.columbia.edu/~cook/movabletype/archives/2009/12/healthcare_spen.html

Gelman, A., Park, D., Shor, B., and Cortina, J. (2009d). *Red State, Blue Stat, Rich State, Poor State: Why Americans Vote the Way They Do*, second edition. Princeton University Press.

Harris, J., and Kamvar, S. (2008). I want you to want me. In the *Design and the Elastic Mind* exhibit at the Museum of Modern Art, New York. <http://iwantyouwantme.org/index.html>

Hawking, S. (1988). *A Brief History of Time*. New York: Bantam.

Inselberg, A. (1985). The plane with parallel coordinates. *Visual Computer* 1, 69-91.

Inselberg, A. (2009). *Parallel Coordinates: Visual Multidimensional Geometry and Its Applications*. New York: Springer.

Kosara, R. (2007). Visualization criticism – the missing link between information visualization and art. *Proceedings of the 11th International Conference on Information Visualisation (IV)*, 631–636.

Levitt, S., and Dubner, S. (2004). *Freakonomics*. New York: William Morrow.

McCandless, D. (2009). Reduce your odds of dying in a plane crash. Information is Beautiful blog, 1 Jul. <http://www.informationisbeautiful.net/visualizations/reduce-your-chances-of-dying-in-a-plane-crash/>

Nightingale, F. (1858). Diagram of the causes of mortality in the army in the east. Reproduced at <http://www.florence-nightingale-avenging-angel.co.uk/Coxcomb.htm>

Norman, D. A. (2002). Emotion and design: Attractive things work better. *Interactions Magazine* 9, 36-42.

Orwell, G. (1946). Why I write. London: Gangrel.

PA Consulting Group (2008). Dynamic planning for COIN in Afghanistan. Link from <http://atwar.blogs.nytimes.com/2009/12/03/finding-victory-in-a-plate-of-pasta/>

Radiohead (2008). House of Cards music video. <http://code.google.com/creative/radiohead/>

Rehmeyer, J. (2008). Florence Nightingale: The passionate statistician. *Science News*, 26 Nov. http://www.sciencenews.org/view/generic/id/38937/title/Math_Trek__Florence_Nightingale_The_passionate_statistician

Segel, E., and Heer, J. (2010). Narrative visualization: telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, 336-343. Washington: IEEE Computer Society Press.

Small, H. (1998). Florence Nightingale's statistical diagrams. <http://www.florence-nightingale-avenging-angel.co.uk/GraphicsPaper/Graphics.htm>

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire: Graphics Press.

Tufte, E. R. (1990). *Envisioning Information*. Cheshire: Graphics Press.

Tukey, J. W. (1972). Some graphic and semigraphic displays. In *Statistical Papers in Honor of George W. Snedecor*, ed. T. A. Bancroft. Ames: Iowa State University Press.

- Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley.
- Tukey, J. W. (1993). Graphic comparisons of several linked aspects: alternatives and suggested principles. *Journal of Computational and Graphical Statistics* 2, 1-33.
- Uberti, O. (2009). The cost of care. *National Geographic*, 18 Dec.
http://blogs.ngm.com/blog_central/2009/12/the-cost-of-care.html
- Unwin, A., Theus, M., and Hofmann, H. (2006). *Graphics of Large Datasets: Visualizing a Million*. New York: Springer.
- Unwin, A., Volinsky, C., and Winkler, S. (2003). Parallel coordinates for exploratory modelling analysis. *Computational Statistics and Data Analysis* 43, 553-564.
- Urbanek, S. (2006). *Exploratory Model Analysis: An Interactive Graphical Framework for Model Comparison and Selection*. Books on Demand GmbH.
- Wainer, H. (1997). *Visual Revelations*. New York: Springer.
- Wattenberg, L. (2007). Where all boys end up nowadays. Baby Name Wizard blog, 19 Jul.
<http://www.babynamewizard.com/archives/2007/7/where-all-boys-end-up-nowadays>
- Wattenberg, L., and Wattenberg, M. (2005). Name voyager. On the Baby Name Wizard website. <http://www.babynamewizard.com/voyager#>
- Wickham, H. (2006). Exploratory model analysis with R and GGobi. Technical report.
<http://had.co.nz/model-vis/2007>
- Wickham, H., Cook, D., Hofmann, H., and Buja, A. (2010). Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics* 16, 973-979.

Wilkinson, L. (2005). *The Grammar of Graphics*. New York: Springer.

Yau, N. (2008a). Ebb and flow of box office receipts over past 20 years. Flowing Data, 25 Feb.
<http://flowingdata.com/2008/02/25/ebb-and-flow-of-box-office-receipts-over-past-20-years/>

Yau, N. (2008b). 5 best data visualization projects of the year. Flowing Data blog, 19 Dec.
<http://flowingdata.com/2008/12/19/5-best-data-visualization-projects-of-the-year/>

Yau, N. (2009). Narrow-minded data visualization. Flowing Data blog, 22 April.
<http://flowingdata.com/2009/04/22/narrow-minded-data-visualization/>