Statistical Graphics and InfoVis — Twins separated at Birth?

Nicholas Lewin-Koh and Martin Theus, Eds.

This volume features two articles both looking at the aspects of "graphical displays of quantitative data". In the first paper "Visualization: It's More than Pictures!" by Robert Kosara, Robert sheds a light from the point of view of an InfoVis person, i.e. someone who primarily learned how to design tools and techniques for data visualization. With the second article "Visualization, Graphics, and Statistics" by Andrew Gelman and Antony Unwin, we get a similar view, but now from someone whose primary training is in math and/or statistics.

Given this set-up, we might think that we have a good idea how both sides would argue, and what would be the assets the one and the other side would claim: computer scientists are good at designing tools for data visualization and statisticians are good at doing the analysis; and consequently, they both don't know much about the expertise of the other discipline.

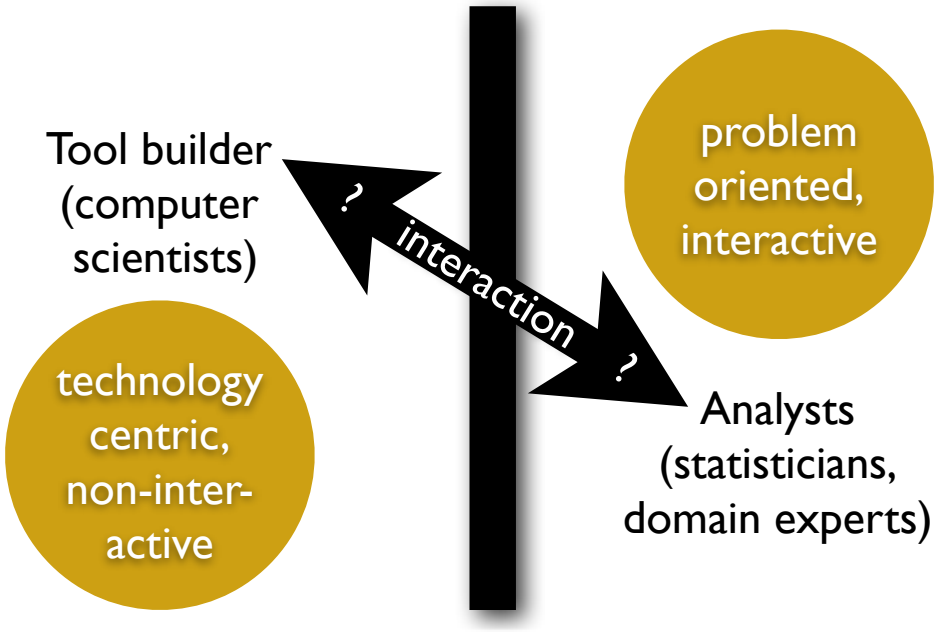


Figure 1: Is there a "wall" between the two promotors of graphical displays? (Taken from `http://www.theusrus.de/blog/the-wall-what-wall/`.)

Reading the two papers you will find out that, while there is certainly some truth behind this simple classification, the overlap and agreement is larger than one would probably think. The common and most important understanding is that there is a story to be told with the data. Graphics are the most powerful tool to do this, no matter what your training and background is.

Nonetheless, there is still a lot to be learned from each other and the one or the other difference or misunderstanding might spur the discussion between the two sides. As a platform for this **discussion** you can use the post at `http://www.theusrus.de/blog/InfoVis-and-StatGraphics/` — we are looking forward to a lively exchange, which might even end up in a collaboration!

Visualization: It's More than Pictures!

Robert Kosara

Introduction

Information visualization is a field that has had trouble defining its boundaries, and that consequently is often misunderstood. It doesn't help that InfoVis, as it is also known, produces pretty pictures that people like to look at and link to or send around. But InfoVis is more than pretty pictures, and it is more than statistical graphics.

The key to understanding InfoVis is to ignore the images for a moment and focus on the part that is often lost: interaction. When we use visualization tools, we don't just create one image or one kind of visualization. In fact, most people would argue that there is not just one perfect visualization configuration that will answer a question [4]. The process of examining data requires trying out different visualization techniques, settings, filters, etc., and using interaction to probe the data: filtering, brushing, etc.

The term *visual analytics* captures this process quite well, and it also gives a better idea of what most visualization is used for: analysis. Analysis is not a static thing, and can rarely be done by looking at a static image. Visualization and visual analytics use images, but the images are only one part of visualization.

Cheap Thrills

It is no wonder that many people think that visualization is primarily about pretty and colorful pictures, even smart people like Andrew Gelman. What readers see on popular websites like FlowingData [8] and infosthetics [3], and what makes them so popular, are the pictures. In many cases, they provide only minimal context, and readers are mostly left to look at the images as images, rather than figure out what they are actually trying to tell them.

Another issue is the blurred boundary between actual visualization and data art, which is often ignored on purpose to have more interesting images to choose from. The result is that the expectation many people have of visualization images is similar to that of a piece of art: that you can look at it and like or don't like it, but don't get any actual information out of it. In fact, I have argued that what Gelman calls "that puzzling feeling" is actually what sets pragmatic visualization apart from data art [2].

Data art clearly has its place, and the more pragmatic visualization community can learn from it. But when we're talking about visualization in the context of statistics and the analysis of data, we need to draw a clear distinction. Visualization is not art any more than statistics is.

Goals

So what does visualization do, then? The main idea is to provide insight into data. This is how scientific visualization got started in the 1980s: the huge amounts of data produced by the then-recent supercomputers required new ways of analysis. Scientific visualization made it possible to see the effects of design changes on the pressure distribution of an airplane wing, for example. The same thing could be done with number crunching in theory, but it was a lot more immediate and obvious where things went wrong when the model was actually shown as an image.

Another, more recent, goal is making data accessible. A lot of data is already available in principle, but not in a form that normal people would want to play with. There is still a difference between data being technically available and actually being accessible to a broad audience. Creating a visualization makes it possible for people to start poking around in the data and perhaps discover interesting facts that nobody has seen before.

Finally, to borrow Tableau's tagline [6], the goal of visualization is *to make analytics fast*. Sure, a lot of questions can be asked of a data warehouse by writing 150-line SQL queries, but changing parameters or exploring variations is going to be difficult this way. An interactive visualization system makes it possible to do that and ask many more questions in much less time. This is not only a worthwhile goal in a business context, but also in the sciences and many other fields: the easier and quicker it is to ask questions, the more questions can be asked.

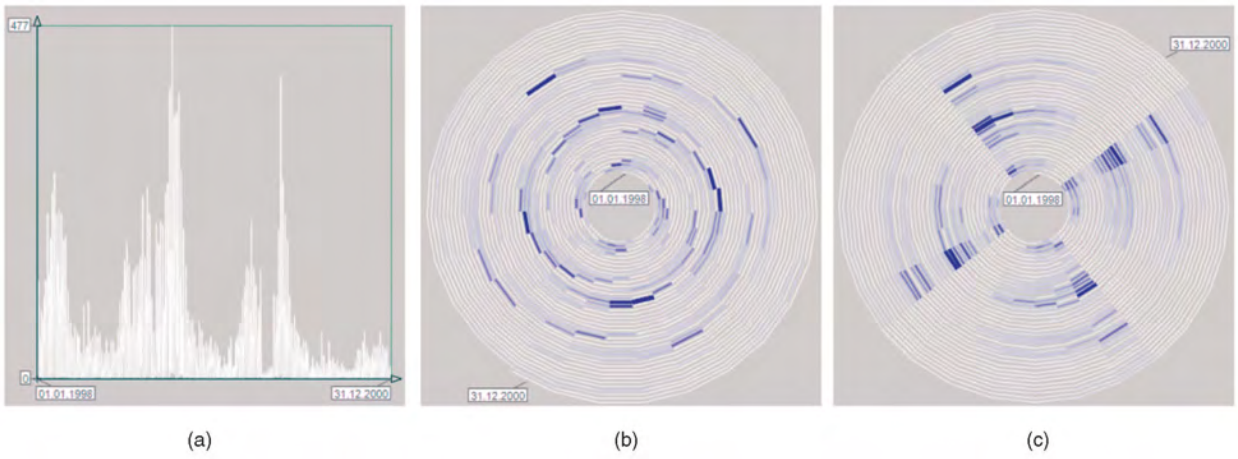


Figure 1: Spirals are useful for finding periodicity in data (from [1]). (a) The bar chart shows no obvious periodic pattern; (b) the spiral set to 25 days hints at a periodic pattern, but this is clearly not the correct time frame; (c) at 28 days, the pattern is very clearly visible.

Example: Perceive Patterns

A common question in time series data is whether the data is periodic, and if yes, what the period is. A common way of finding out is drawing the data on a spiral [1]. By changing the number of data points that is shown per full round the spiral makes (that number is constant, of course), patterns become visible. Figure 1 shows an example of sick leave data that has an interesting periodic pattern: in 28 days, there are four periods, which means that there is a weekly pattern: more people call in sick on Mondays than later in the week.

The way this pattern was discovered is deceptively simple. All it took was to play with a slider that allowed the user to change the number of days on shown on the spiral. Slide it back and forth, and soon you will see a pattern (if there is one). With a bit of practice, you can even tell when you're getting close, as there are telltale signs around the optimal value.

The key here is not just the way of displaying the data, but also the interaction. Without it, it would take much longer to find the correct interval, or require some very educated guessing. The power of visualization is that it allows the user to find things he or she may not have expected, and thus would not have been looking for.

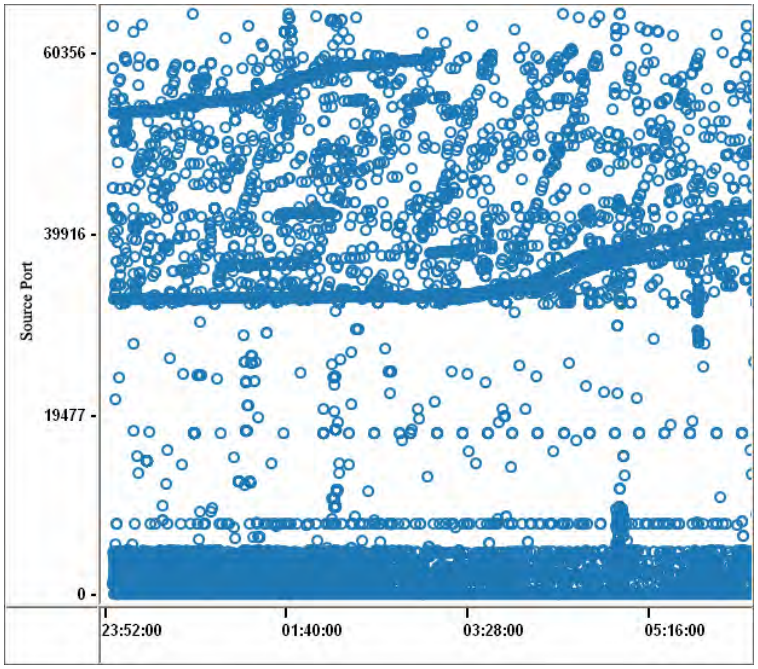
Example: Filter the Flood

A beautiful example of the integration of analysis and visualization is a system for visualizing network traffic data [7]. To be able to deal with the enormous amount of data, the system includes a declarative logic system that can apply rules to find certain patterns in the data. The idea is to identify patterns of known good data, and filter that data out, so that what remains is the data that needs to be examined more closely (Figure 2).

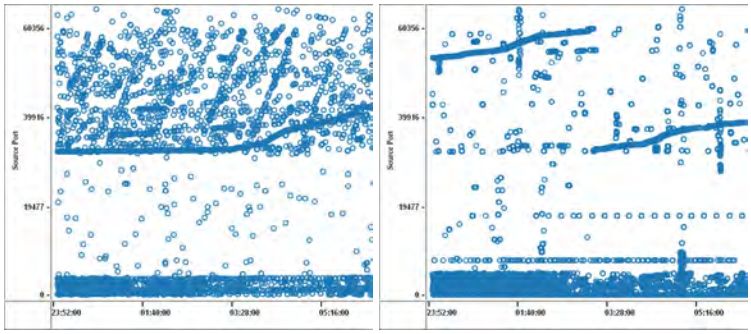
Instead of having to write the declarations by hand, however, the system allows the user to select data points and creates a rule from the selection. The user can then apply that rule to other traffic to see if it matches the right data, and even examine and edit the actual definition directly. Creating and refining definitions of different traffic patterns is relatively straight-forward this way, especially for a network security expert.

One of the most clever design decisions in this system is to focus on the known good traffic, rather than trying to define what is suspicious. New types of scans and attacks are developed all the time, so keeping up with them is practically impossible. Also, defining the bad traffic would defeat a big advantage of the visual part of this system: being able to see new patterns as they emerge.

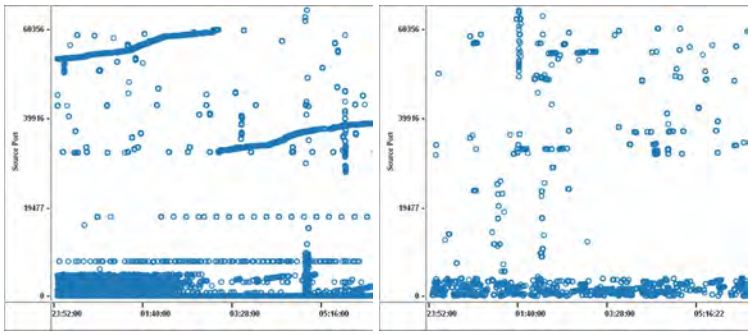
By treating the known good traffic as irrelevant, it can be removed, and the user can focus on the



(a)



(b) (c)



(d) (e)

Figure 2: Event diagrams showing flows during residual analysis (from [7]). (a) Original unidentified traffic (b) Flows with "mail" label (c) The residual after filtering out "mail" from Figure 6a. (d) Flows with the "scan" label (e) The residual after filtering out the "scan" label from Figure 6(c).

parts that may be suspicious. Each part is done by the component that is best suited for it. The machine uses the rules to sift through and filter large amounts of data, and the user tries to understand what remains and tweaks the rules (or finds a way to fend off a break-in attempt).

Conclusions

Visualization cannot exist without visual representations, and those representations need to be designed so that they can be effectively and efficiently perceived. There is no question that more effective visual representations will result in better analysis and easier comprehension of data. But the images aren't everything.

There is also a vast open field of research that makes good use of statistics to enhance visualization. A few attempts at this exist [5], but a lot more can be done. Despite the relatively new field of visual analytics, visualization research is still very strongly focused on visual representation, with too little attention being paid to interaction, analysis, and cognitive effects.

And yet, visualization is much, much more than what it appears to be at first glance. The real power of visualization goes beyond visual representation and basic perception. Real visualization means interaction, analysis, and a human in the loop who gains insight. Real visualization is a dynamic process, not a static image. Real visualization does not puzzle, it informs.

Bibliography

[1] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visual methods for analyzing time-oriented data. *Transactions on Visualization and Computer Graphics*, 14(1):47–60, 2008.

[2] R. Kosara. Visualization criticism — the missing link between information visualization and art. In *Proceedings of the 11th International Conference on Information Visualisation (IV)*, pages 631–636. IEEE CS Press, 2007.

[3] A. V. Moere. information aesthetics. `http://infosthetics.com/`.

[4] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualization. In *Proceedings Visual Languages*, pages 336–343. IEEE Computer Society Press, 1996.

[5] C. A. Steed, P. J. Fitzpatrick, T. Jankun-Kelly, and J. E. S. II. Practical north atlantic hurricane trend analysis using parallel coordinates and statistical techniques. In *Proceedings of the 2008 Workshop on Geospatial Visual Analytics*, 2008.

[6] Tableau software. `http://tableausoftware.com/`.

[7] L. Xiao, J. Gerth, and P. Hanrahan. Enhancing visual analysis of network traffic using a knowledge representation. In *Proceedings Visual Analytics Science and Technology (VAST)*, pages 107–114. IEEE CS Press, 2006.

[8] N. Yau. Flowingdata. `http://flowingdata.com/`.

Robert Kosara
UNC Charlotte
`rkosara@uncc.edu`
`http://eagereyes.org/`

Visualization, Graphics, and Statistics

Andrew Gelman and Antony Unwin [1]

Quantitative graphics, like statistics itself, is a young and immature field. Methods as fundamental as histograms and scatterplots are common now, but that was not always the case. More recent developments like parallel coordinate plots are still establishing themselves. Within academic statistics (and statistically-inclined applied fields such as economics, sociology, and epidemiology), graphical methods tend to be seen as diversions from more "serious" analytical techniques. Statistics journals rarely cover graphical methods, and Howard Wainer has reported that, even in the *Journal of Computational and Graphical Statistics*, 80% of the articles are about computation, only 20% about graphics.

Outside of statistics, though, infographics and data visualization are more important. Graphics give a sense of the size of big numbers, dramatize relations between variables, and convey the complexity of data and functional relationships. Journalists and graphic designers recognize the huge importance of data in our lives and are always looking out for new modes of display, sometimes to more efficiently portray masses of information that their audiences want to see in detail (as with sports scores, stock prices, and poll reports), sometimes to help tell a story (as with annotated maps), and sometimes just for fun: a good data graphic can be as interesting as a photograph or cartoon.

We and other graphically-minded statisticians have been thinking a lot recently about the different perspectives of statisticians and graphic designers in displaying data. But first we would like to emphasize some key places in which we agree with the infographics community, some reasons why we and they generally prefer numbers to be graphed rather than written.

- A well-designed graph can display more information than a table of the same size, and more information than numbers embedded in text. Graphical displays allow and encourage direct visual comparisons.
- It has been argued that tables are commonly read as crude graphs: what you notice in a table of numbers are (a) the minus signs, and thus which values are positive and which are negative, and (b) the length of each number, that is, its order of magnitude. In a table of statistical results you might also note the boldface type or stars that indicate statistical significance. A table is a crude form of log-scale graph. If we really must display numbers in tables with many significant figures, it would probably generally be better to display them like this: 3.14_{16}, so as not to distract the readers with those later unimportant digits.
- A graph can tell a story so easily. A line going up tells one story, a line going down tells another, and a line that goes up and then down is yet another possibility. It is the same with scatterplots and more elaborate displays. Yes, a table of numbers can tell a story too—especially in an area such as baseball where, as sabermetrician Bill James wrote, numbers such as .406 or 61 evoke images and history—but in general the possibilities of storytelling are greater and more direct with a graph. Storytelling is important in journalism and advertising (of course) but also in science, where data can either motivate and illustrate a logical argument or refute it.

In short, graphs are a good way to convey relationships and also reveal deviations from patterns, to display the expected and the unexpected.

Now we turn to differences between statistical graphics and infovis. In statistical graphics we aim for transparency, to display the data points (or derived quantities such as parameter estimates and standard errors) as directly as possible without decoration or embellishment. As indicated by our remarks above, we tend to think of a graph as an improved version of a table. The good thing about this approach is it keeps us close to the data. The bad thing is that it limits our audience. We as statisticians think we're keeping it simple and clean when we display a grid of scatterplots, but the general public—and even researchers in many scientific fields—don't have practice reading these

[1] We thank 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

graphs, and can often miss the point or simply tune out.

In contrast, practitioners of information visualization use data graphics more generally as a means of communication, in competition (and collaboration with) photographs, cartoons, interviews, and so forth. For example, a news article about health care costs might include some reportage (perhaps with some numbers gleaned from government documents), quotes from experts, an interview with a sick person who cannot get health insurance, a photograph of a high-tech MRI machine, a how-much-do-you-know quiz on the prices of medical procedures—and a data visualization showing medical costs and service use in different parts of the country. The visualization is graded partly on how cool it looks: "cool" grabs the reader's attention and draws him or her into the story.

We hope that, by recognizing our different goals and perspectives, graphic designers and statisticians can work together. For example, a website might feature a dramatic visualization that, when clicked on, reveals an informative static statistical graphic that, when clicked on, takes the interested reader to an interactive graphic and a spreadsheet with data summaries and raw numbers.

We illustrate some of our points with two examples. The first is Florence Nightingale's famous visualization of deaths in the Crimean War. Here is Nightingale's graph from 1958 (for more details, see Hugh Small's presentation at `http://www.florence-nightingale-avenging-angel.co.uk/Coxcomb.htm`):

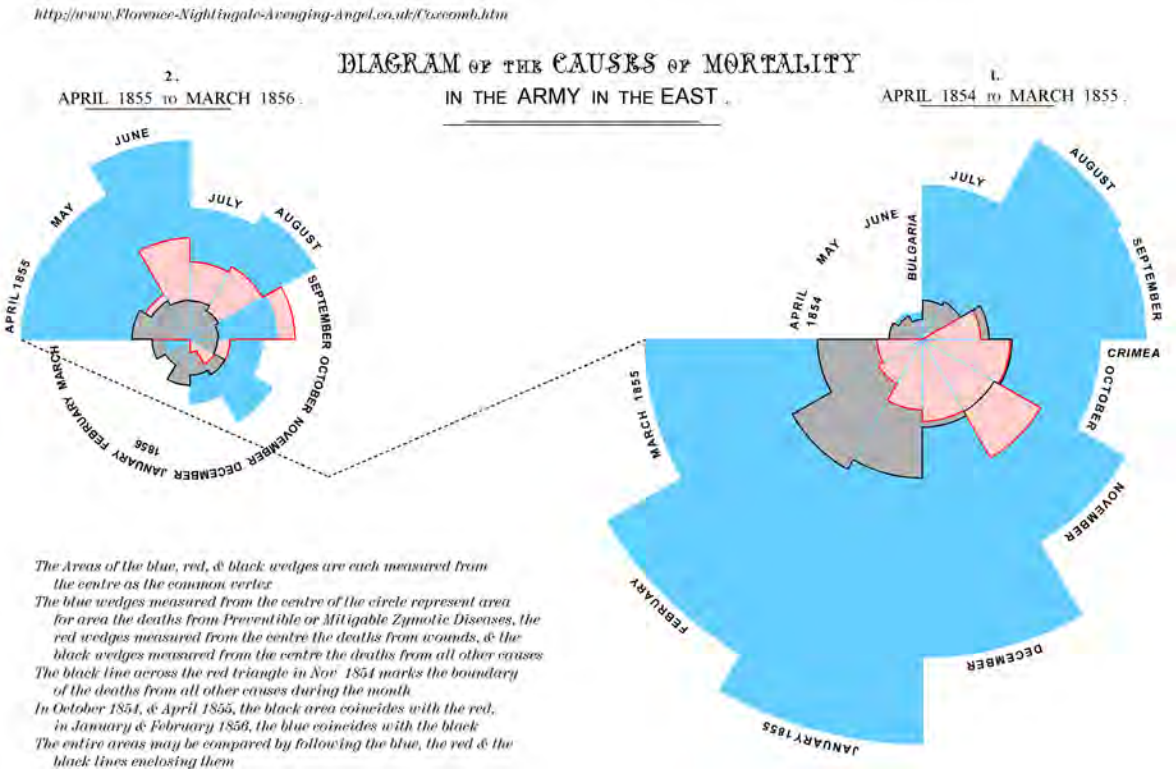


Figure 1: Florence Nightingale's famous visualization of deaths in the Crimean War is attractive and draws the viewer in closer so as to understand what is being conveyed.

And now our presentation of the same information using R:

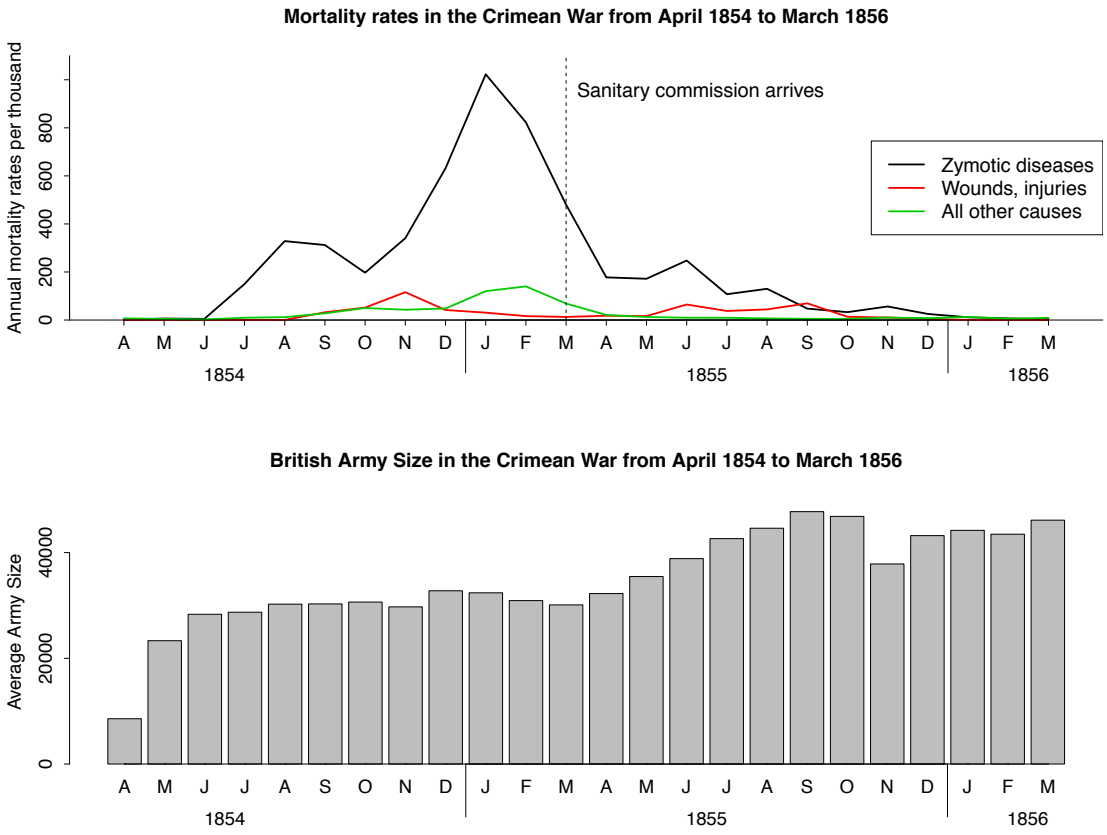


Figure 2: Our re-plotting of Nightingale's data shows the data and their patterns much more clearly, but in a visually less striking way. As is often the case, two smaller plots can show data much more directly than is possible from a single graph, no matter how clever.

Nightingale's visualization and ours both have their strengths. When it comes to displaying the data and their patterns, we much prefer the plain statistical graphs. The most salient visual feature of Nightingale's graph is that a year is divided into twelve months, a fact that we already knew ahead of time. The trends and departures from trend are much clearer when plotted directly as time series. This is no criticism of Nightingale: the standard statistical techniques of today were not so easily available in the mid-1800s, and in any case her graph did the job of attracting attention better than ours do, in any era.

Nightingale's graph is intriguing and visually appealing—much more so than our bland graph—and, as is characteristic of the best infographics, the appeal is centered on the data display itself. A reader who sees this graph is invited to stare at it, puzzle it out, and understand what it is saying. In some ways, the weaknesses of the graph from a statistical point of view—it is difficult to read, the main conclusions to be drawn from the data are not clear, indeed it is a bit of a challenge to figure out exactly what the graph is saying at all—are strengths from the infovis perspective. Given that the graph is attractive enough, and the subject important enough, to motivate the reader to go in deeper, the challenges in reading the graph induce a larger intellectual investment in the viewer and a motivation to see the raw data.

And once policymakers were alerted by Nightingale's dramatic visualization, they were able to scan the columns of numbers directly and understand what was going on: the patterns in

these time series are clear enough that we imagine a careful study of a tabular display would suffice. The role of the graph was to dramatize the problem and motivate people to go back and look at the numbers.

In a modern computing environment, a display such as Nightingale's could link to a more direct graphical presentation such as ours, which in turn could link to a spreadsheet with the data. The statistical graphic serves as an intermediate step, allowing readers to visualize the patterns in the data.

Our second example concerns the survival rates of different groups who sailed on the Titanic's maiden voyage. Here is a doubledecker plot showing the survival rates by sex (males on the left and females on the right) and within sex by class (first, second, third, crew). The widths of the bars are proportional to the numbers in each group, so that we get a rough idea of their relative sizes, though it is the survival rates that are of most interest.

It is easy to see two expected conclusions, that female survival rates were higher than males for all possible comparisons, and that female survival rates went down with class. It is also obvious, though more surprising, that the lowest male survival rate was in the second class. The fact that the male crew survival rate was higher than the male survival rates in the second and third classes must at least partly be due to the lifeboats being manned with crew members to accompany the passengers. All of these conclusions may be drawn directly from the display, but no one would claim it is an attention-grabbing graphic! We looked on the internet (a.k.a. googled) to see if these data had been presented in an infographic display and found several statistical displays, not all either clear cut or easy to read, though no infographic ones. This is a good example where cooperation between statisticians and infographics experts could really pay off: we have an interesting dataset and several interesting conclusions to present and we would like to do it in an attractive and stimulating way without losing any statistical clarity. Just wanting to do that is not enough, we need design expertise, and we look forward to someone from the infographics side taking up the challenge of helping us.

Andrew Gelman
Dep. of Statistics and Department of Political Science
Columbia University, New York
gelman@stat.columbia.edu
http://www.stat.columbia.edu/~gelman/

Antony Unwin
Department of Mathematics
University of Augsburg
unwin@math.uni-augsburg.de
http://www.rosuda.org

Figure 3: A doubledecker plot showing the survival rates on the Titanic by sex and, within sex, by class. This graph shows several interesting comparisons but could benefit from improvement in graphic design.