Are Infovis and Statistical Graphics Really All That Different?

A response by Stephen Few of Perceptual Edge to the article
"Infovis and Statistical Graphics: Different Goals, Different Looks"
by Andrew Gelman and Antony Unwin
Journal of Statistical Computing and Graphics
Jan 20, 2011

I work in the field of information visualization. I am not a statistician by training, but statistics informs much that I do. Through books, articles, blog posts, presentations, and courses, I teach people of all kinds, in organizations of all types, simple and practical ways to use information visualization to explore, make sense of, and communicate quantitative data, primarily to support better decisions. When reading this article by Gelman and Unwin, I found much that is useful, articulate, and true. I also found ways in which we disagree; in some cases quite fundamentally.

As statisticians and ardent practitioners of information visualization, Gelman and Unwin know and appreciate both of these worlds. Despite this fact, however, they characterize information visualization in a narrow way that misrepresents its scope and reduces its worth. But I don't think this was their intention. They seem to genuinely respect information visualization, but do so in a way that most practitioners would find misleading and belittling.

Before addressing points of disagreement, I want to acknowledge the ways in which we agree and show my appreciation for the authors' good intentions. Here are five points of agreement in particular:

1. Statisticians should be more familiar with information visualization and should apply it to their work more routinely.

2. There is insufficient interaction between the worlds of statistics and information visualization. A greater degree of interaction would benefit both groups.
3. When trying to communicate statistical information to an audience that is not already interested, it helps to design a graphical display that is visually engaging in a way that draws them into the data. Statistical graphics can be beautiful and engaging as well as clear and accurate.
4. Despite the previous point, many infographics that are popular today display data ineffectively. “The very beauty of many professionally-produced images may, paradoxically, stand in the way of better understanding of data in many situations.”
5. Yau’s best visualizations of 2008, McCandless’ map of airplane crashes, Nightingale’s polar area graph of deaths in the Crimean War, Uberti’s health spending and life expectancy plot, and the network diagram that described how to win the war in Afghanistan are not good examples of effective data presentation. Wattenberg’s Baby Name Voyager visualization, on the contrary, is a good example of form and function skillfully integrated to produce a rich analytical experience. “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.”

It is because we agree on so much that I find our points of disagreement surprising and confusing. I agree with Gelman and Unwin that visualization is underappreciated and insufficiently utilized by most statisticians. Statisticians often attend my courses and presentations. Even though I am not a statistician, they almost always leave these events with a greater appreciation for the potential of information visualization to expand their abilities. As statisticians themselves, Gelman and Unwin have a better opportunity than I do to show fellow statisticians ways in which information visualization is useful, sometimes critically so, and to warn them against poor visualization practices. In this paper, I believe that they failed to do this—not entirely, but in some important ways—in part by suggesting a seldom appropriate and impractical solution when a simple and accessible solution is at hand.

I'll identify and respond to five claims made by the authors that I consider erroneous.

Error #1: Information visualization equals infographics

The authors define "information visualization" as "infographics," especially those that are cool looking. "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." They often refer to information visualization practitioners as "graphic designers." The authors were careful to point out that quantitative graphics are used for many purposes and that judgments about their effectiveness must be based on the creator's intentions. In light of this, it's strikingly inconsistent to define information visualization narrowly as infographics, which represent a small subset of information visualization that just happens to be getting a lot of attention right now, mostly due to poorly designed but eye-catching examples that appear on the Web. Relatively few practitioners of information visualization are graphic designers. Apart from information visualization researchers (mostly at universities) and infographic designers (many who are journalists), the vast majority are people who use graphics and visual analysis techniques in the workplace to explore, make sense of, and then communicate quantitative data. They exist in all parts of the organization and go by many titles, some of which include the word statistician or analyst, but just as often include words like manager or even administrative assistant. They use information visualization to do their work, mostly to support better decisions.

Using Yau's list of the five best information visualizations of 2008 to characterize information visualization is not appropriate. Yau's interest in flashy infographics is only typical of a relatively small group of people involved in information visualization. As the authors point out, none of the visualizations on his list qualify as the best information visualizations of 2008 or any other year. I assume that Yau found these visualizations appealing because they were novel, not because they were effective. Few practitioners would agree that Yau's list features exemplars that should be emulated, except perhaps as exploratory experiments by researchers searching for new approaches.

Error #2: Statisticians should collaborate with graphic designers

The authors suggest that statisticians should collaborate with graphic designers to improve their data communication skills. While some principles and practices of graphic design can certainly be applied to statistical graphics to improve them (good color choices, typography, balance, use of space, etc.), most graphic designers know little or nothing about communicating quantitative information, because it isn't part of their training.

Most of the graphical design practices that can be used by statisticians to communicate their findings more effectively are quite simple and easy to learn. This is the important message for statisticians to hear today. Not, "you're incapable of presenting data effectively, so partner up with a graphic designer to get this done," but "what you need to present data effectively is easy to learn given the right resources and a little practice."

Statistical graphics rarely need to be beautiful to communicate, but they do need to feature the essential data simply and without distraction, in clear, accurate, and aesthetically pleasing ways. The aesthetic qualities that Donald Norman talks about in the book *Emotional Design* that make things pleasurable to interact with, including statistical graphics, and thus put readers in the right frame of mind to assimilate and understand information, need not rise to the level of beautiful. By “beautiful,” I’m referring to an extraordinary quality that makes something stunning in appearance. There are occasions when something must be stunning to catch an audience’s attention, such as when people are browsing the Web for something interesting, but only in such circumstances are the effort and graphic design skills that are required to accomplish this necessary. The overall aesthetic qualities that statistical graphics should routinely exhibit may be summarized as “pleasing to the eye.” The display doesn’t look cluttered. The colors work well together and are not overly bright, except in cases when brightness causes something important to stand out. The text is legible. Non-data components of the display (e.g., axes on a graph) do their part subtly, without undue salience. The most important information stands out above the rest. These are a few aspects of graphical design that go a long ways toward making a reader’s interaction with information inviting and pleasurable, drawing the reader into the information rather than distracting her with meaningless embellishments.

With a little training, these aesthetic qualities are easy to achieve.

Error #3: Statistical graphics and information visualizations are quite distinct

The authors draw a stark contrast between statistical graphics and information visualization, which establishes a false dichotomy. Statisticians and information visualization practitioners share a common goal: to find, understand, and then present important truths that reside in data. The best known and most commonly quoted definition of “information visualization” appears in the book *Readings in Information Visualization: Using Vision to Think*, by Stuart Card, Jock Mackinlay, and Ben Shneiderman: “Information visualization is the use of computer supported, graphical representations of abstract data to augment cognition.” When information visualizations display quantitative data, which is usually the case, they are in fact “statistical graphics.” In other words, such a display can be properly called either a statistical graphic or an information visualization. Regardless of what we call them, the principles that we follow to make them effective are precisely the same.

The authors state: “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.” Actually, people in both fields pursue unique and distinctive displays to solve unfamiliar problems, but mostly rely on proven generic methods. The fact that “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),” is just as true of information visualization. It is true that, “In the world of Infovis, design goals **can** be pursued at the expense of statistical goals.” However, it is not true that they **should** be pursued in this way any more than in the world of statistics.

Error #4: Difficulty in deciphering graphics can be useful for communication

The authors believe that eye-catching infographics, which look “cool” but are difficult to decipher, can effectively serve as a door opener to “bland” but decipherable and accurate statistical graphics, along with tables of numbers for the raw details. They state: “Inefficient graphical displays can be effective in implicitly extracting a commitment from the reader to think hard about the data.” The authors provide no evidence to support this claim nor have I seen any credible research that supports it elsewhere.

Error #5: Infographics need not always inform

The authors say that infographics (keep in mind that to them infographics and information visualization are one and the same) don't always need to inform. "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 information." I agree that there is a place for comic illustrations and the like to catch the reader's eye, but we shouldn't call these information visualizations. Illustrations of this type should serve their attention-getting role in a way that is obvious to readers that this is their sole intention so they don't waste their time trying to decipher them. I've seen far too many cool-looking infographics that entice readers to play with flashy, spinning charts, thinking that information is being gleaned when they are in fact learning little or nothing.

I'll use one of the author's examples to show that statistical graphics can exhibit attractive design by following simple principles that are easily learned. I'm not talking about the flashy stuff that will win a beauty contest, which is rarely needed, but statistical graphics that tell their story in a clear and aesthetically pleasing way. In their article, the authors used Florence Nightingale's polar area graph of deaths in the Crimean War as an example of an infographic that was engaging and successful, but not easy to decipher. As an alternative, they created their own graphics to tell the story more clearly, but in a way that was less attractive. In a previous article titled "A Response to Visualization, Graphics, and Statistics," which covered the same topic, the authors described what they had in mind as follows:

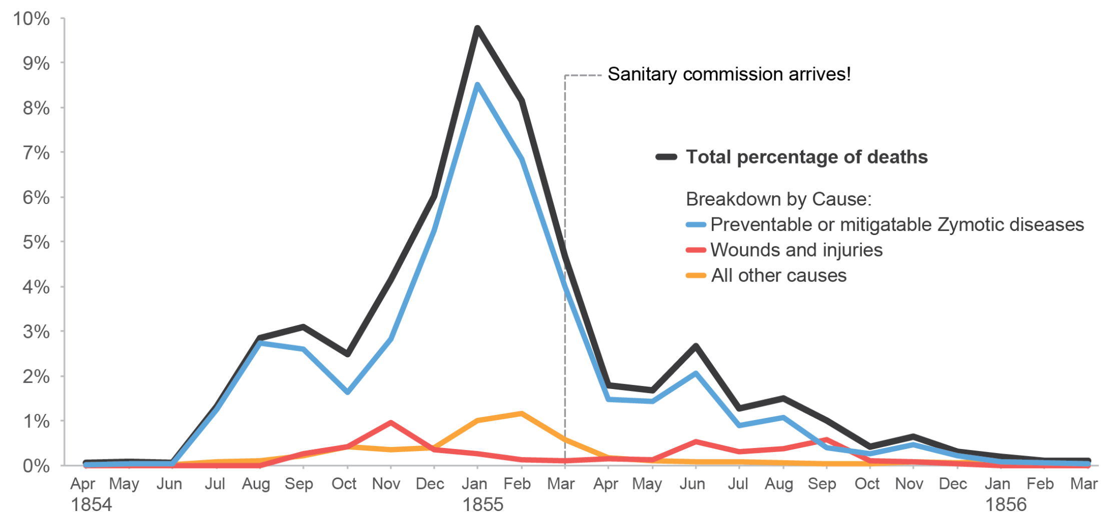
> *In a modern computing environment, a display such as Nightingale's could link to a more direct graphical presentation such as ours, which in turn could link to a spreadsheet with the data…We have an interesting dataset and several interesting conclusions to present and we would like to do it in an 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.

A better and certainly more practical approach in most cases, however, would involve a single statistical display that is clear, accurate, and aesthetically pleasing, such as the following:

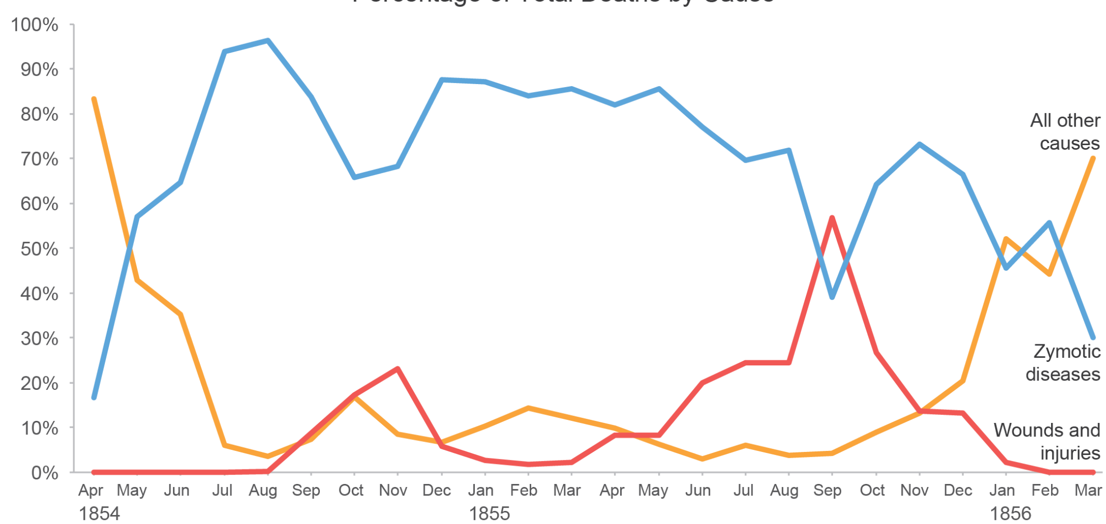
Most Deaths Among Soliders in the East are Preventable

Percentage of Active Soldiers Who Died per Year and Cause



Notice in the graph above that most deaths in the war effort (black line represents total deaths) were caused by preventable and mitigatable Zymotic diseases (blue line, which closely hugs the black line).

Percentage of Total Deaths by Cause



Notice that wounds and injuries (red line) at one point in time only–September 1855–accounted for a greater percentage of deaths than Zymotic diseases and otherwise accounted for relatively few.

These two simple line graphs combined with a few words (titles, labels, annotations, and two descriptive sentences) tell the story in a way that is accessible, informative, and pleasing to the eye. I chose semi-saturated and easy to distinguish colors for the lines. I reduced the salience of less important elements that play only a supporting role, such as axis lines and scales. I chose a sans-serif font that can be easily read even on a screen that is not high resolution. I used white space, alignment, and the positioning of items to create a clean, uncluttered appearance.

Besides aesthetic choices, I also used aspects of visual design for semantic distinctions. I made the black line, which represents the total number of deaths, darker and slightly thicker to make it stand out as different and greater than the other lines. I made the main title the same color as the blue lines in the graphs because those lines illustrate the message of the title.

I expressed the values in particular ways to make them easier to decipher and understand. Rather than requiring readers to compare the number of soldiers who died in one graph to the total number of soldiers in another graph to get a sense of proportion as Gelman and Unwin did in their version, I simplified this aspect of the story in the upper graph by displaying these proportions directly (percentage of soldiers who died out of the total who were active). Rather than expressing deaths per cause as raw counts or rates per one thousand, in the lower graph I expressed them as percentages of total deaths summing to 100% to spread the values across more physical space in the plot area, making cause-specific patterns easier to see and compare.

I created this easily using Excel. I used Excel, not because it’s the best graphing tool available—it certainly isn’t—but because everyone has Excel and I wanted to demonstrate that you can communicate data effectively even if you have nothing else.

Every statistician should have these skills, because every statistician should be able to communicate their findings to the people who need them. Because these skills are so easy to learn, there is no excuse for displaying data poorly.