InfoVis Is So Much More: A Comment on Gelman and Unwin and an Invitation to Consider the Opportunities

Robert Kosara, Tableau Software
rkosara@tableausoftware.com

I welcome the opportunity to respond to Andrew Gelman and Antony Unwin's article, *Infovis and Statistical Graphics: Different Goals, Different Looks*. Their view of information visualization is very distorted, but unfortunately not uncommon. In the following, I will try to give readers a sense of what information visualization (InfoVis) is really about, show some recent contributions, list some challenges, and show that there is a lot of opportunity for collaboration between InfoVis and statistics.

Gelman and Unwin base their selection of work on the blog Flowing Data, run by Nathan Yau. While Yau has a large number of readers, his blog does not represent the state of the art in InfoVis research. He tends to focus on communication-oriented and artistic pieces, and rarely delves into the depth of data analysis using visualization on the blog itself. He does some of that in the paid members-only section, as well as in his book [7].

Yau's focus clearly colors Gelman and Unwins perception of visualization, and is the basis for their claim that, "[on] the infovis side, computer scientists and designers are interested in grabbing the readers attention and telling them a story." This is, quite simply, not true.

Information Visualization: Exploration, Analysis, Presentation

Visualization, of which information visualization is a part, is generally concerned with three types of tasks: exploration, analysis, and presentation. It is not surprising that the bulk of online sources focus on the latter, since presentation is the easiest to understand (by design), and tends to be the most visually appealing. There is also a large number of artistic projects that turn data into colorful images, but often without the goal to inform.

The first two tasks, exploration and analysis, is where the majority of published work in the visualization literature has been done.

Exploration of data is based on little, if any, knowledge of a particular dataset. This can be because a user really knows little or nothing about the data, or because she tries to look at it with fresh eyes. The goal of exploration is generally to find out interesting pieces of information, understand the overall relationships in the data, and perhaps make little discoveries.

Data analysis in the visualization sense involves knowledge about the data and at least some starting hypotheses. This phase is also much more involved and typically takes much

longer than exploration.

Novel techniques are only one type of work that gets published in the visualization literature. There is also a wide variety of evaluation papers, from ones looking at basic perception of sizes, shapes, colors, etc., to comparisons of different techniques against each other for particular tasks. Case studies provide insights into how visualization is used in practice, while theory papers give us tools to better understand users or explore the visualization design space.

What is Information Visualization?

The key issue that is perhaps the most fundamental, and the most misunderstood, about InfoVis is how the mapping process between the data and the visual representation works. There are many ways to turn numbers into pictures, with many different kinds of results. Some are visually appealing, others are not. Some are bare and minimal, others are exuberant and colorful.

But what sets visualization apart from other mapping processes is a simple criterion: readability. If a visual representation of data cannot be read, if it is not possible to map what is seen back to the data, it is not a visualization in the InfoVis sense. Since there is almost always a loss of precision, a truly bijective mapping between the data and its visual representation is not possible. However, the user needs to be able to relate the visual patterns back to the data for the visualization to be of any use.

Examples of one-way visual mappings include music visualizations (the shapes and colors change with the music, but they dont give you actual information about the music), as well as many artistic projects. The goal of the latter is often more to raise awareness and create interesting pieces, but not to analyze the actual data. In fact, analysis is somewhat contrary to artistic uses of visualization [1].

Representing data that has no obvious visual equivalent is a challenge and an opportunity. Finding the best way to represent data is not a trivial task, and often depends on particular properties of the data as well as the tasks envisioned with it.

InfoVis and Statistics

While there is not currently much interaction between visualization and statistics or statistical graphics, there is a lot of opportunity for really interesting work. The infovis community is very open to more statistically-informed ideas, and every year, a few papers are published at the VisWeek conference that bridge the gap. The person who has been the most successful in navigating the boundary between the two fields is Hadley Wickham, together with his collaborators Dianne Cook and Heike Hofmann [4, 5].

Information visualization builds on many ideas from statistics, and is informed by the work of statisticians like John Tukey, William Cleveland, Leland Wilkinson [6], and others. While there are differences in the methods, the goals of the two fields are very similar. The differences are a strength: by combining complimentary work, we can solve problems that are hard or impossible with just one set of techniques.

The Five Best Data Visualization Projects of 2008

As I have tried to show, Gelman and Unwin's selection of examples is not representative of the

work done in infovis, and this is also true of the "five best visualization projects" Nathan Yau picked from what he published in 2008. I largely agree with their critiques, even if we differ in some minor points.

Other than the streamgraph movie visualization (and perhaps Wordle), none of these examples would be considered information visualizations by anybody in the field. The streamgraph was published in the Information Visualization conference, but there has been much debate about its usefulness. It does serve its purpose as a presentation technique, but it certainly does not lend itself to data analysis. Wordle, and word clouds in general, are problematic because of a variety of factors that make judging word sizes difficult: longer words naturally look bigger even if they don't occur more often, words with wider characters are larger, etc.

The other examples are all nice to look at but do not qualify as visualizations because they are not readable: it is not possible to actually learn anything meaningful about the data from them.

Different Goals, Different Looks

For all the issues with the article the title is certainly spot on: InfoVis and statistical graphics do, in fact, have different goals and they also tend to look quite different. InfoVis emphasizes exploration and visual discovery, often at the expense of statistical rigor. The goal is to create images that communicate the data in a way that makes it possible for the human visual system to recognize patterns, including correlation [2], clusters [3], and randomness [4]. InfoVis techniques tend to show a lot of data, at least thousands of data points, often many more. This makes InfoVis a very human-centered field, which cares first and foremost about being easy to understand and informative, and also innovative in its variety of ways to show data. In addition to the display techniques themselves, there are interaction techniques (since large and multi-dimensional data can often not be shown in a single view) and user studies to find out which techniques work, and why.

While my understanding of the statistical graphics community is limited, it seems that statistical graphics is much more centered on the statistical properties first, with the visual appearance and ability to see patterns more of a side product. Without an understanding of how the data has been processed and transformed, it is often difficult, if not impossible, to understand the graphs. Interaction also does not appear to be a priority, with many statistical graphics seemingly being created for print.

Which approach is better is clearly a question of taste as much as of the task: do I want to quickly dig into my data or do I care about precise statistical properties? However, it is easy to see that both fields can learn from each other; each has deficiencies in areas the other field does well.

Where To Find The Real InfoVis

To get a better sense of the InfoVis field, I would like to point readers to the yearly VisWeek conference, which takes place in the U.S. mid- to late October and covers a broad range of topics, including scientific visualization, information visualization, visual analytics, etc., as well as the similarly positioned EuroVis conference in early June. The main journals in the field are the IEEE *Transactions on Visualization and Computer Graphics* (which pub-

lishes the VisWeek proceedings), the *Information Visualization* journal, the Eurographics *Computer Graphics Forum* journal (which publishes the EuroVis proceedings), and IEEE *Computer Graphics and Applications*. These venues are much more focused on data exploration and analysis than presentation, are peer-reviewed, and have the acceptance rates and impact factors one expects from the leading publications in a field.

While there is much to criticize in visualization, Gelman and Unwin largely miss the point because most what they criticize is not actually visualization. A well-reasoned critique of real visualization papers from a statistics point of view would be very interesting and extremely valuable for the field.

Unfortunately, their distorted view of information visualization is quite common. I invite them and all readers to explore the journals I have listed, attend our conferences, and learn more about this fascinating field. There is a lot of opportunity for collaboration and cross-pollination between our fields.

References

[1] Robert 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.

[2] Jing Li, Jean-Bernard Martens, and Jarke J van Wijk. Judging Correlation from Scatterplots and Parallel Coordinate Plots. *Information Visualization*, 9(1):13–30, 2010.

[3] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A Taxonomy of Visual Cluster Separation Factors. *Computer Graphics Forum*, 31(3):1335–1344, June 2012.

[4] Hadley Wickham, Dianne Cook, Heike Hofmann, and Andreas Buja. Graphical Inference for InfoVis. *Transactions on Visualization and Computer Graphics*, 16(6):973–979, 2010.

[5] Hadley Wickham and Heike Hofmann. Product Plots. *Transactions on Visualization and Computer Graphics*, 17(12):2223–2230, 2011.

[6] Leland Wilkinson. *The Grammar of Graphics*. Springer, 2005.

[7] Nathan Yau. *Visualize This: The Flowing-Data Guide to Design, Visualization, and Statistics*. Wiley, 2011.