Why is Data Visualization Important? What is Important in Data Visualization?

Antony Unwin¹

¹Institute of Mathematics, University of Augsburg, Augsburg, Germany

Published on: Jan 31, 2020

DOI: <https://doi.org/10.1162/99608f92.8ae4d525>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Column Editor's Note: *Data visualization, facilitated by the power of the computer, represents one of the fundamental tools of modern data science. Professor Antony Unwin from the University of Augsburg describes different ways in which data visualization is used, explores the opportunities for future research in the area, and looks at how data visualization is taught.*

The What and Why of Data Visualization

Data visualization means drawing graphic displays to show data. Sometimes every data point is drawn, as in a scatterplot, sometimes statistical summaries may be shown, as in a histogram. The displays are mainly descriptive, concentrating on 'raw' data and simple summaries. They can include displays of transformed data, sometimes based on complicated transformations. One person's statistics may be another person's raw data. As with other aspects of working with graphics, it would be useful to have an agreed base of concepts and terminology to build on. The main goal is to visualize data and statistics, interpreting the displays to gain information.

Data visualization is useful for data cleaning, exploring data structure, detecting outliers and unusual groups, identifying trends and clusters, spotting local patterns, evaluating modeling output, and presenting results. It is essential for exploratory data analysis and data mining to check data quality and to help analysts become familiar with the structure and features of the data before them. This is a part of data analysis that is underplayed in textbooks, yet ever-present in actual investigations. Look, for instance, at the one-sided peaks in the distributions of marathon finishing times (marastats, 2019).

Graphics reveal data features that statistics and models may miss: unusual distributions of data, local patterns, clusterings, gaps, missing values, evidence of rounding or heaping, implicit boundaries, outliers, and so on. Graphics raise questions that stimulate research and suggest ideas. It sounds easy. In fact, interpreting graphics needs experience to identify potentially interesting features and statistical nous to guard against the dangers of overinterpretation. Just as graphics are useful for checking model results, models are useful for checking ideas derived from graphics (for more on models, see Hand, 2019).

This overview concentrates on static graphics. Dynamic graphics and, more especially, interactive graphics are in an exciting stage of development and have much to add. They require an article of their own. Superb examples include Human [Terrain](#), a dynamic graphic showing the world's population in 3-D, and the interactive [NameVoyager](#).

'A Picture Is Worth a Thousand Words'

Famous sayings have a way of developing a life of their own. A picture is not a substitute for a thousand words; it needs a thousand words (or more). For data visualization you need to know the context, the source of

the data, how and why they were collected, whether more could be collected, the reasons for drawing the displays, and how people with the necessary background knowledge advise they might be interpreted. There is a story that M. G. Kendall reviewed a book of R.A. Fisher's with the words: "No one should read this book who has not read it already." It is like that with graphics. If you have read all the supporting text, the display is often memorable and readily understandable. If you have not, it is not. Graphics on their own are insufficient, they are part of a whole. They complement text and are complemented by text. Student's reanalysis of the Lanarkshire Milk Experiment (Student, 1931) is an excellent example (and is also interesting as an early analysis of a large data set).

The potential synergy of text and graphics can be appreciated by talking through your own graphics, explaining them to others. Why have you drawn those graphics? How have you drawn them? What can be seen? Are there interesting patterns? What could be changed and improved? Which other graphics might be drawn? How can conclusions be checked? There should be more talking about graphics and less relying on the graphics to speak for themselves.

When it comes to graphics you have not drawn yourself, the same kinds of questions are still relevant, although they may be more difficult to answer. Edward Tufte described Charles Minard's display of Napoleon's Russian campaign as the best statistical graphic ever drawn (Tufte, 2001). It is a magnificent graphic, fully deserving of the praise heaped on it, yet as Lee Wilkinson has pointed out in his book *The Grammar of Graphics* (Wilkinson, 2005), there are inaccuracies and imprecisions in the display. Why did no one point them out before? We are too used to accepting graphics uncritically, not asking enough questions of them.

Presentation and Exploratory Graphics

Presentation and exploratory graphics are quite different animals. In presenting your results, you may have space for only one graphic and no idea how many people may see it. If it appears in a newspaper or on television or the Web, your audience could be millions of people. The graphic should be well-designed and well-drawn with an effective accompanying explanatory text. On the other hand, if you are exploring data, then you need many, many graphics and they are for an audience of one: yourself. The individual graphics need not be perfect, but they should provide alternative views and additional information. Presentation graphics are used to convey known information and are often designed to attract attention. Exploratory graphics are used to find new information and should direct attention to information.

Published graphics tend to be graphics for presentation, partly because they are for publication and partly because no one wants to see hundreds of quick graphics that may or may not have been helpful. It is rather like mathematical proofs: articles contain the elegant and concise final versions, not the scribbled notes and random ideas that came before. How many graphics may have been drawn before the striking display was chosen to show the resignations of U.K. cabinet ministers in recent years (Institute for Government, 2019)?

Exploratory graphics take advantage of how easy it is now to draw and redraw graphics. What used to be a slow and wearisome process, even including having to print out displays, has become fast and flexible. At the same time, new, additional skills are required. Identifying interesting features and knowing how to check them in more detail among a myriad of possible graphics is not just a matter of drawing many graphics, you need interpretative skills and an appreciation of which graphics will provide what kinds of information. There is so much that can be varied: the variables displayed, the types of graphics, the sizes of graphics and their aspect ratios, the colors and symbols used, the scales and limits, the ordering of categorical variables, the ordering of variables in multivariate displays. Selecting from the wide range of graphics wisely, and understanding how to gain insights, are not trivial tasks. The lack of a theory of data visualization to guide and build on is a key issue.

Data Visualization Has Become More Important

Better hardware has meant more precise reproduction, better color (including alpha-blending), and faster drawing. Better software has meant easier and more flexible drawing, consistent themes, and higher standards. Computer scientists have become much more involved, both on the technical side and in introducing new approaches. There has been progress in developing a theory of graphics, especially thanks to Wilkinson's *Grammar of Graphics* (2005) and Hadley Wickham's implementation of it in the R package *ggplot2* (Wickham, 2016). There is continuing work and better understanding of the problems of color and perception. Graphics that were rarely used and difficult to draw, such as parallel coordinate plots (e.g., Theus, 2015) and mosaicplots (e.g., Unwin, 2015), have been refined and developed. Much larger data sets can be analyzed and visualized and graphics can play a valuable role in diagnosing the strengths and weaknesses of complex models. Data visualizations can be found everywhere, in scientific publications, in newspapers and TV, and on the Web. There are many Web pages where graphics are discussed and debated. This is a huge improvement over the situation of even 20 years ago.

Research in Data Visualization

There are great opportunities for future research in data visualization. Principles are needed on how to decide which of many possible graphics to draw. It is not a matter of drawing a single, 'optimal' graphic, if such a thing even existed; it is a matter of choosing a group of graphics that will provide more information. It is like taking photographs of a complicated object, a single one would not be enough, and taking pictures from every possible angle and distance would be far too many. Sets of graphics are useful for providing context, as the scatterplots in Klimek, Yegorov, Hanel, and Thurner (2012) demonstrate.

More understanding of combining and linking graphics is needed, whether in static ensembles or in interactive displays, just as better software is needed for these. The value of alignment and common scaling for making effective comparisons, for instance, with small multiples and faceting (displaying many graphics of the same form conditioning on other variables) is one part of this. It is a historical curiosity that the current exciting work on interactive graphics on the Web still lags behind standalone systems that were already available more

than 30 years ago in linking multiple windows. Data Desk and JMP were commercial examples at the time (see Velleman, 2019, and Sall, 2019, for current versions).

Published graphics are sometimes attractive and beautifully produced. The content does not always match. That may be because authors and publishers do not expect the graphics to be examined in any detail. They may be added as illustrations to balance the layout and make it look more agreeable. If you do not have a suitable photograph, cartoon, or map, you could use a colorful statistical graphic. I have many times heard people say that they do not understand numbers and were bad at mathematics in school. No one has ever said to me they do not understand graphics, perhaps because they regard them as illustrations and not as central parts of an argument. There is work to be done in educating researchers and readers in the value of graphics.

Research into new and innovative graphics is exciting and productive. Simultaneously, it is essential to make the best use of known and well-understood graphics. There is a risk of emphasis on novelty at the expense of familiarity. New, innovative graphics need instruction and experience to interpret them. Their designers have spent much time developing them and reasonably enough believe that what is obvious to them should be obvious to everyone. Just think of the humble scatterplot. It is only in recent years that scatterplots have appeared in the media, although they are one of the most important statistical graphics. If you have never seen one before, they can be intimidating, even more so when you are told ‘It is clear that...’ or ‘You can easily see that...’ We should build on the familiar to carry our readers along with us.

Examples and Sources

The visualizations I like may not be the visualizations you like. I urge you to search extensively and judge for yourselves. Much interesting and thought-provoking material can be found in Tufte's classic books (e.g., Tufte, 2001), and in the displays by the *New York Times* over the years (e.g., *New York Times*, 2018). Other newspapers and media have also produced excellent work. These are, of course, presentation graphics, but they offer much to engage with. It is difficult to make a choice among the many individual Web pages providing examples and discussion, but [Visualising Data](#) is one site that recommends highlights across the web. The current interest and activity in graphics are very welcome.

What Happens Now?

Educating people in choosing, drawing, and interpreting graphics is more difficult than you might think. Data visualization is not taught badly, it is just not taught very much at all. Ideally, there should be better theory, and consequently better graphics. That will take time. In the meantime, we should:

- discuss more graphics more;
- interpret more graphics more;
- teach more graphics more.

Disclosure Statement

Antony Unwin has no financial or non-financial disclosures to share for this article.

References

- Daniels, M. (2018). *Human terrain*. https://pudding.cool/2018/10/city_3d/
- Hand, D. (2019). What is the purpose of statistical modelling? *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.4a85af74>
- Institute for Government. (2019). Ministerial resignations outside reshuffles, by prime minister. Retrieved August 14, 2019, from <https://www.instituteforgovernment.org.uk/charts/ministerial-resignations-outside-reshuffles-prime-minister>
- Klimek, P., Yegorov, Y., Hanel, R., & Thurner, S. (2012). Statistical detection of systematic election irregularities. *PNAS*, 109(41), 16469–16473. <https://doi.org/10.1073/pnas.1210722109>
- marastats. (2019). *General marathon stats*. Retrieved August 14, 2019, from <https://marastats.com/marathon/>
- New York Times. (2018, December 31). *2018: The year in visual stories and graphics*. <https://www.nytimes.com/interactive/2018/us/2018-year-in-graphics.html>
- Sall, J. (2019). *JMP*. Retrieved August 8, 2019, from <http://www.jmp.com>
- Student. (1931). The Lanarkshire milk experiment. *Biometrika*, 23, 398–406.
- Theus, M. (2015). *Tour de France 2015*. Retrieved August 14, 2019, from <http://www.theusrus.de/blog/tour-de-france-2015/>
- Tufte, E. (2001). *The visual display of quantitative information* (2nd ed.) Cheshire, CT: Graphics Press.
- Unwin, A. (2015). *Studying multivariate categorical data*. Retrieved August 14, 2019, from <http://www.gradaanwr.net/content/ch07/>
- Velleman, P. (2019). *Data desk*. Retrieved August 8, 2019, from <http://www.datadesk.com>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). New York, NY: Springer-Verlag. <https://doi.org/10.1007/978-3-319-24277-4>
- Wilkinson, L. (2005). *The grammar of graphics* (2nd ed.). New York, NY: Springer. <https://doi.org/10.1007/0-387-28695-0>

©2020 Antony Unwin. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.