Calculate *and* communicate

Robert Grant says statisticians have plenty to learn from designers when it comes to data visualisation, but the opposite also applies. Both sets of skills are needed

I am a statistician. I went to university and studied mathematics. I expected the compulsory statistics courses to be boring, but I was soon bitten by the bug of discovery and practical applications. I studied some more and got some experience working on hospital quality and safety data. I learnt lots of techniques for analysing data, and the importance of communicating it to the people who were going to act on the findings.

None of the courses I took taught me about graphics, though. Sometimes I made some fairly basic images with my data: scatter plots, bar charts and such. They always seemed to have the biggest impact on the audience, so I thought I had better learn what made for good statistical graphics.

I soon found that visualising data, statistics and predictions was – to my pleasant surprise – an established field of study, and people like William Cleveland had already made plenty of recommendations about good practice. It seemed odd that this was not widely taught as part of statistics courses, given the importance of effective communication.

But then a funny thing happened. Graphs suddenly became cool, and lots of new writing about them sprang up. New formats were appearing too, with new techniques for getting the message across, like animations and interactive web content. The field even got a new name: data visualisation, or just *dataviz*.

There have been several recent books on dataviz that document these developments, often written by people from a design or journalism background. They have injected some much-needed design thinking and verve into the subject, but they have had little to say about analytical aspects.

This feels like an oversight to me. Anyone going into a career visualising data will have to combine analysis and design thinking. Increasingly, they will find themselves working in diverse data science teams that might combine statisticians with designers, web developers and computer scientists. They may be challenged by different norms and language among these colleagues, but they also stand to learn a lot from one another.

Learning from design

Statisticians can learn from designers about creative thinking, sketching and user testing. If you want to know whether your visualisation is as good as it could be, you will need to explore many potential formats, generate lots of sketches, throw away the no-hopers, and show the remainder to people who are as similar as possible to your intended audience. Do they draw the right conclusions from your graph? Do they find it interesting, or dull, or confusing? Do they remember the message a short while later? In dataviz, there is generally no right or wrong approach, which is one of the reasons why it is such fun.

To make enough preliminary sketches – and bearing in mind that nothing cramps creative thinking quite like a ticking clock – you have to work quickly and in rough outline. I encourage people to step away from the computer, as software tends to lead us down one design

Robert Grant is a statistician specialising in data visualisation and Bayesian modelling, and is the founder of BayesCamp, which provides training and coaching in Bayesian methods and software. His new book, *Data Visualization: Charts, Maps, and Interactive Graphics*, is out now, published by Chapman and Hall/CRC.

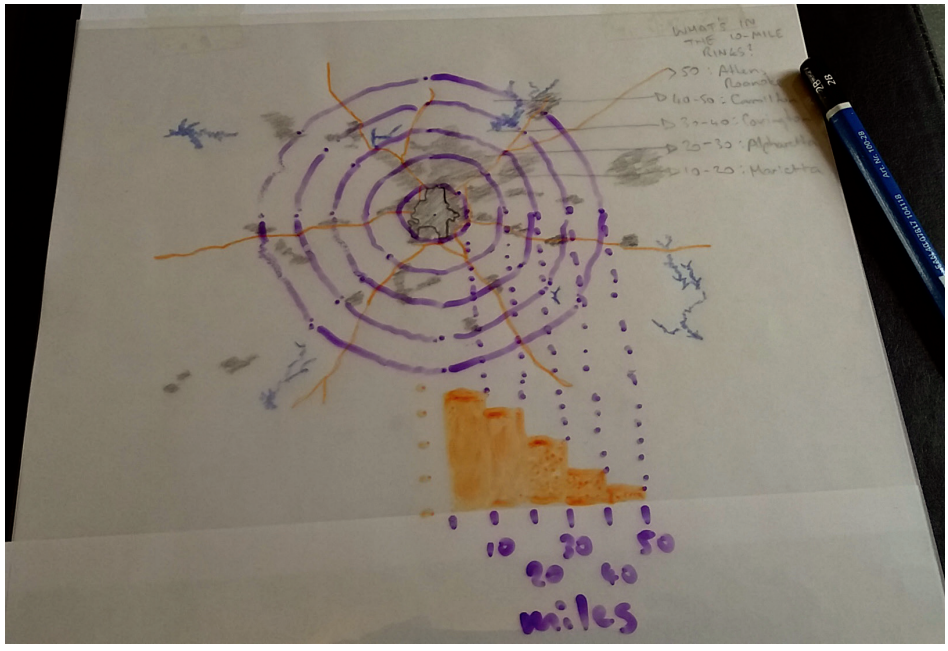

FIGURE 1 Author's sketch of a visualisation of commuting distances for the city of Atlanta, Georgia, using a data set sourced from *Statistics: Unlocking the Power of Data*.[2]

rabbit hole or another, encouraging lots of tiny tweaks to code to make the perfect image without stopping to ask whether the format itself is most appropriate. Instead, grab some coloured pencils, tracing paper and acetate transparencies, and experiment.

I wanted to make these ideas prominent in my book,[1] so I took a very simple data set of commuting distances for the city of Atlanta, Georgia (from *Statistics: Unlocking the Power of Data*[2]) and made a sketch.

Figure 1 shows a histogram of the commuting distances, as well as a map of the area around Atlanta. I used tracing paper to get minimal information from the map, choosing the city boundaries, shading for built-up areas, blue for reservoirs and lakes, and orange for major roads. I felt that this would be enough information for local people to recognise, and I could use the orange again in the histogram. I attached the tracing paper to a blank background and added an acetate transparency sheet on top. That allowed me to try lots of different designs over the map.

I drew rings on the transparency at 10-mile intervals around a central point, which I thought would help to relate the histogram bins directly to the real world. When I saw the effect of all the orange, I decided it would be better to switch colours for the rings and have some faint lines linking them to the histogram visually. Then I added annotations to the sheet underneath. I decided that I did not want to clutter the map with names of towns, but it would be helpful to local readers to see the main place names in each ring/histogram bin.

This is only one sketch, for illustrative purposes. In reality, it helps to make lots of sketches. There are many different considerations that need to be brought together before you advance to the next stage and tidy up your design for user-testing. In my case, bars are understood unambiguously, while a pie chart would have been harder for readers to convert mentally from image to numbers.

In my first digital draft (Figure 2), I used mapbox.com to generate a bespoke minimal map with my preferred colour scheme. I put that image into a Scalable Vector Graphics (SVG) file and added the rest of the image by manually editing the SVG code (my experience has been that good dataviz always requires some manual editing – there is no free lunch). The histogram got inverted because, if the dotted lines extended above the bars, they gave a potentially confusing impression of the bars being somehow elongated without explanation. I reluctantly added the actual percentages because they were the bottom line most people wanted to end with. The place names were added in the same area, to avoid clutter. The title was added not at the top but at bottom left, to draw the reader's eye down through the image.

Adding the map means that the 10-mile bins of the histogram relate directly to something concrete, which the audience (local commuters or transport planners) already know well. This means that they will absorb the quantitative information in the visualisation quickly and with minimal effort, and the familiar locations will get them interested in a way that simple numbers do not.

Annotations help guide the reader through the image, as long as they are used sparingly and do not clutter it. Colour is also used sparingly: I decided on one shade of red, which links the concentric rings to the bars, and not to make the visual task harder for the reader by making this the same colour as the roads.

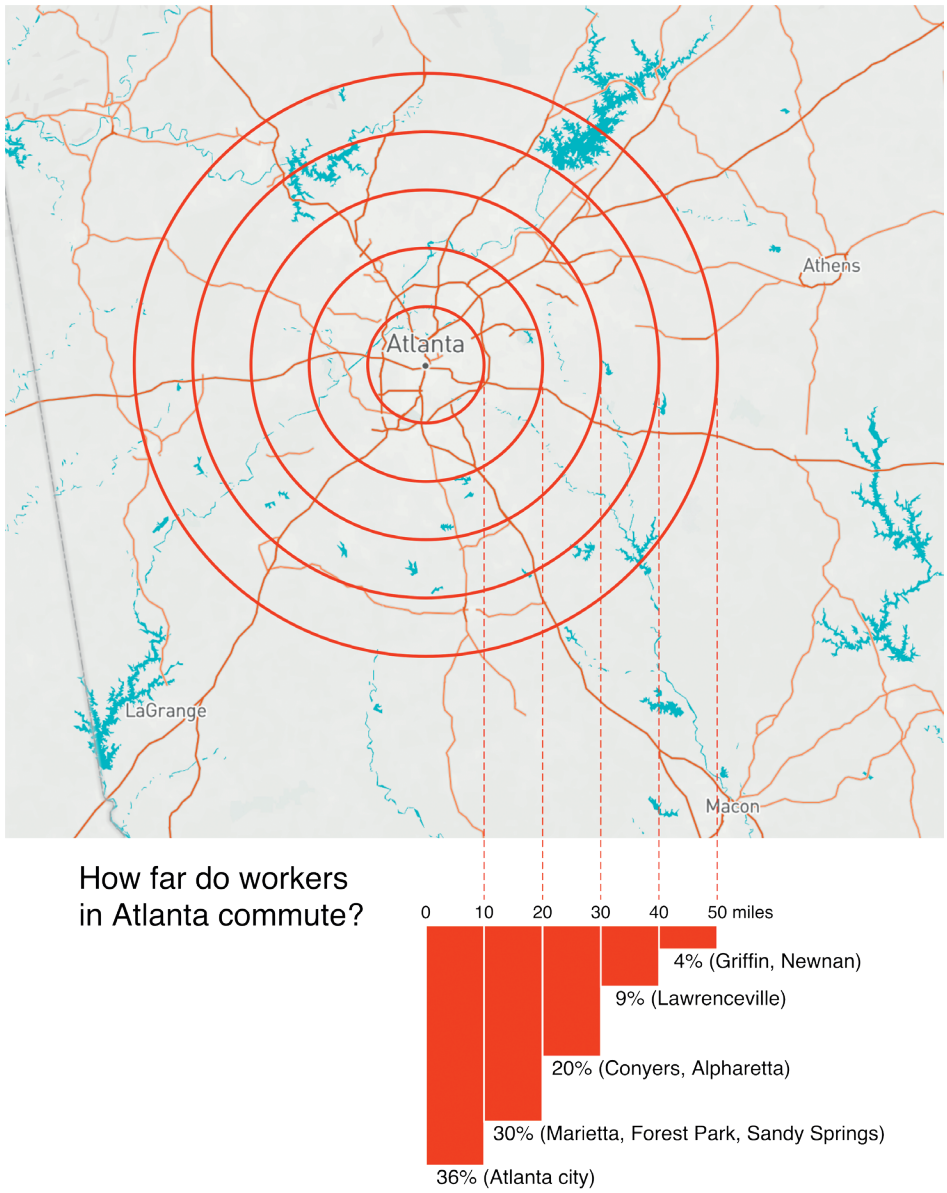

FIGURE 2 A refined, digital version of Figure 1, a visualisation of commuting distances for the city of Atlanta, Georgia, using a data set sourced from *Statistics: Unlocking the Power of Data*.[2]

Learning from statistics

While statisticians hone their design skills, designers can improve their statistical thinking. Some (but not all) of the dataviz writing I have seen recently suggests that a standard workflow begins with finding an interesting data set, then looking through it for patterns, correlations, outliers and other anomalies, then deciding which of these makes a good story, then visualising that.

Statistically trained readers will probably have mental alarm bells ringing now. Scouring quantitative data for a story will often lead to false conclusions, especially if you found the data second-hand and do not know the limitations of how and why it was collected. As the saying goes: if you torture the data long enough, it will confess to anything. By understanding the limitations of the provenance of the data, defining our questions up front, and quantifying uncertainty, we can guard against this (but even then, not entirely).

You can imagine the value to an investigative journalist of searching, say, the Panama Papers for some interesting story, but the same does not apply to inferences made from quantitative data sets. You can look at a league table to find the worst hospital in a country, or the worst school, or the most crime-ridden neighbourhood, but league position alone cannot tell you whether neglect, mismanagement, or bad luck is to blame. The story you think you have might just be noise.

Some visualisation formats make anomalies pop out of the page more visibly than others. A time series with a few transient outliers, shown as a line chart, will produce spikes, where the outliers put a lot of "ink" on the "page" compared to the rest of the series. If we use a scatter plot, the outliers are just more dots, albeit in unusual places. So the analytical and design considerations are interwoven and support each other.

If we consider the questions *a priori* and decide that these outliers really are the story of interest, then a line chart will work well. Otherwise, it will be a poor choice. If we conclude that the outliers are not to be trusted, and that the story is about average trends, then it may be best to smooth the time series visually and relegate the outliers to an accompanying data table.

Of course, we must not hide parts of the data that do not suit our message. Seeing the story we want to see is only one of many cognitive and quantitative biases that statisticians are taught to resist. Without understanding these potential pitfalls, dataviz people are constantly at risk of producing something good-looking that may later ruin their reputation.

Visualising uncertainty is one of the most important tools to counteract an over-enthusiastic reading of data. Three broad approaches are shown in Figure 3: (a) shows many estimates piled up, preferably in a semi-transparent colour so that the densest region is easily seen; (b) gives a best estimate and an interval of pre-specified probability around it; and (c) shows a collection of contours as the likelihood/probability goes up, peaking at the best estimate. Statistical techniques like bootstrapping or Bayesian statistics can help with this, but are unknown to the great majority of dataviz design people.

Learning to compromise

All of this weighing of opposing influences and considerations to arrive at a final visualisation means that dataviz has to be thoughtful and must aim to be useful to its audience. Not everything can be shown; compromises have to be made, information curated, and decisions along the way have to be justified. But that's what makes dataviz a stimulating area to work in. You get to analyse data, learn new computational skills, talk to people about what it means to them, and – hopefully – see your work make a difference.

As statistics and data science develop, new problems arise for the dataviz designer. For example, how can we show the billions of observations that might be contained within a "big data" set? We can count them in bins like a histogram or hexagonal grid, but even getting a computer to do that is no mean feat. And if interest switches from the distributional shape of the data or the modes to the outliers, what then?

There are considerable challenges for contemporary dataviz, but also considerable opportunities: for shared learning and collaboration, experimentation, and making your mark in this rapidly developing field. If you analyse data, then surely you do so to inform or influence someone. To achieve your goals, you will need to calculate *and* communicate – and dataviz is spookily effective at getting the message across. ■

References

1. Grant, R. (2018) *Data Visualization: Charts, Maps, and Interactive Graphics*. Boca Raton, FL: Chapman and Hall/CRC.

2. Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F. and Lock, D. F. (2012) *Statistics: Unlocking the Power of Data*. Hoboken, NJ: Wiley.

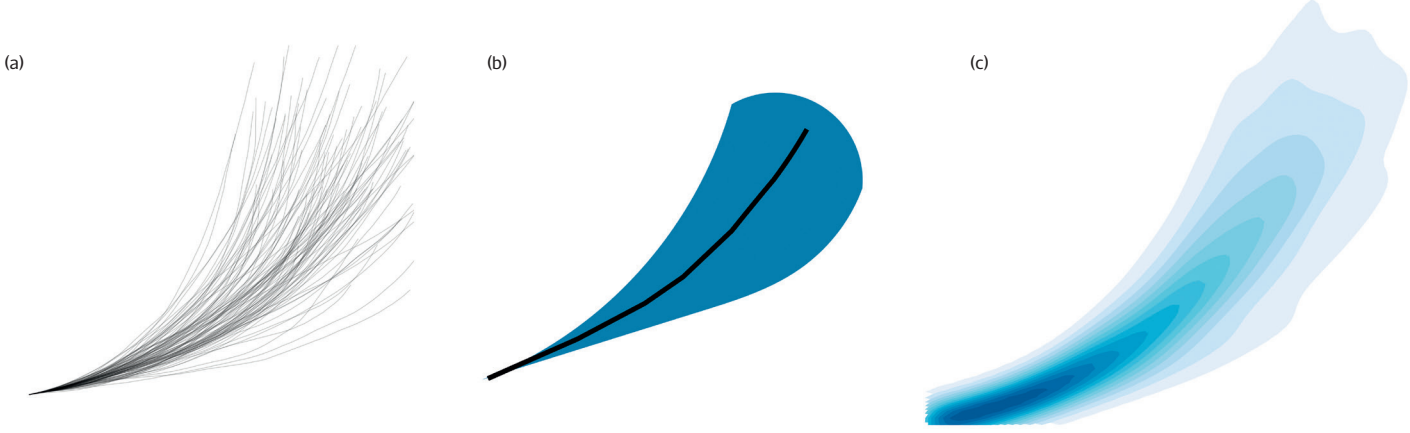

FIGURE 3 Three approaches to visualising uncertainty: (a) shows a collection of estimates, (b) a best estimate within a cone of uncertainty, and (c) a collection of contours of likelihood/probability.