The ladder of abstraction in statistical graphics[*]

Andrew Gelman

31 May 2023

1. Overview

Graphical forms such as scatterplots, line plots, and histograms are so familiar that it can be easy to forget how abstract they are. As a result, we often produce graphs that are difficult to follow. We propose a strategy for graphical communication by climbing a ladder of abstraction (a term from linguistics that we borrow from Hayakawa, 1939), starting with simple plots of special cases and then at each step embedding a graph into a more general framework. We demonstrate with two examples, first graphing a set of equations related to a modeled trajectory and then graphing data from an analysis of income and voting.

2. Visualizing a mathematical model: basketball shooting

Consider the following simple problem in applied mathematics: What is the optimal angle for shooting a basketball, if you would like to be able to throw the ball as softly as possible and still reach the hoop? The solution will be a function of your distance from the hoop and the altitude at which you release the ball.

We would like to reach a general solution but we start with the specific case in Figure 1: a player standing $d = 10$ meters from the basketball hoop, releasing the ball at an altitude of $a = 1.7$ meters. The hoop itself is $h = 3.05$ meters off the ground.

The left plot in Figure 2 shows the path of a ball shot at a $\theta = 30^\circ$ angle at an initial velocity of $v = 15$m/sec. From Newton's laws of motion and assuming no air resistance, the trajectory is $x(t) = v\cos(\theta)\,t$, $y(t) = a + v\sin(\theta)\,t - 0.5gt^2$, where $g = 9.8\text{m/sec}^2$ is the gravitational constant. The red line on this graph directly shows the ball's trajectory. The only way to make it more vivid and less abstract would be an animated presentation that would dynamically show the moving ball.

We then ask what would happen at this same launch angle but with different initial speeds. The right plot in Figure 2 shows trajectories corresponding to $v = 5$, 10, 15, and 20 m/sec. This graph is slightly more abstract than the one on the left in that it displays several paths at once. In addition, because all the curves start at the same point, we moved the label of the initial velocity at the end of each path, which is not quite intuitive. This is an example of the conceptual distancing that can be required when increasing the information content of a plot. We find it helpful to present both graphs of Figure 2 in order to guide the reader through the process of abstraction.

Our next step is to figure out the initial velocity needed to reach the basket, given the shooting angle. This is a straightforward algebra problem. First we figure out the time T required for the ball to reach the horizontal position of the hoop: $d = v\cos(\theta)\,T$, thus $T = \frac{d}{v\cos(\theta)}$. Then we solve for the initial velocity v so that the ball's vertical postion at time T is that of the hoop: $h = a + v\sin(\theta)\frac{d}{v\cos(\theta)} - 0.5g\frac{d^2}{v^2(\cos(\theta))^2} = a + d\tan(\theta) - 0.5\frac{gd^2}{v^2(\cos(\theta))^2}$, so

$$v = \sqrt{\frac{0.5gd^2}{(\cos(\theta))^2(d\tan(\theta) + a - h)}}\,. \tag{1}$$

[*] We thank Howard Wainer, Ron Yurko, Dianne Cook, and Alessandra Casella for helpful comments.

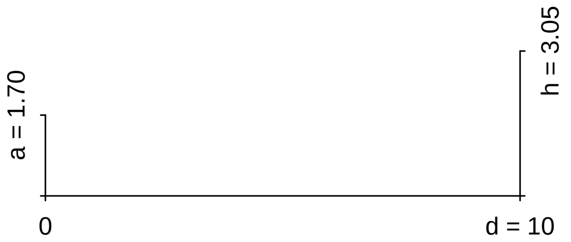

Figure 1: *Diagram of a person standing 10 meters away from a basketball hoop, with the ball released at an altitude of 1.7 meters and the hoop being 3.05 meters high. This is the first step of a series of graphs that will culminate in displaying a optimal shooting angle.*

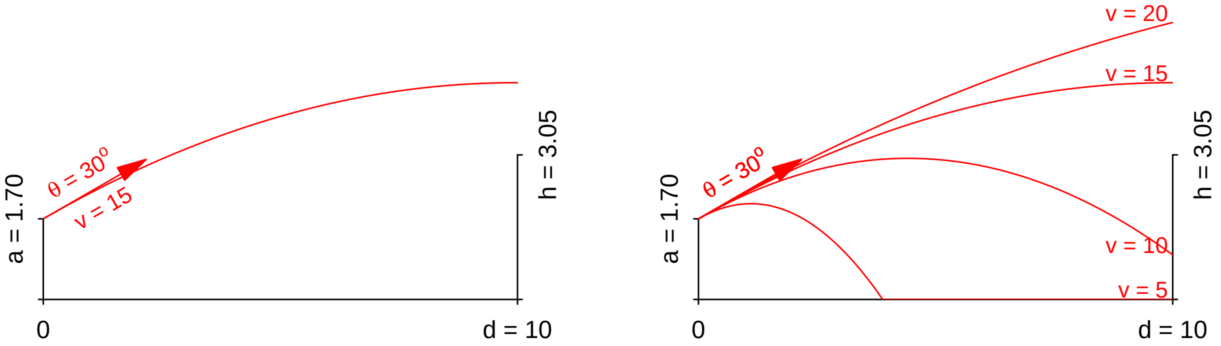

Figure 2: *(Left:) Trajectory of a shot taken at a* 30° *angle at a velocity of 15 m/sec, as computed using Newtonian mechanics ignoring air resistance. The shot goes too high and misses the hoop. (Right:) Trajectories of a series of shots with different initial velocities. The velocity required to reach the hoop must be somewhere between 10 and 15 m/sec. The left graph is at a low level of abstraction in that the red line directly shows the path of the ball. The graph on the right is more abstract in that it shows several trajectories at once.*

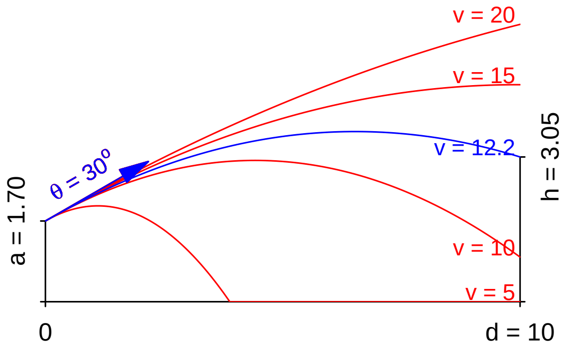

Figure 3: *A shot at angle* 30° *from altitude 1.7 m, from a distance of 10 m, will reach the hoop if its initial velocity is 12.2 m/sec.*

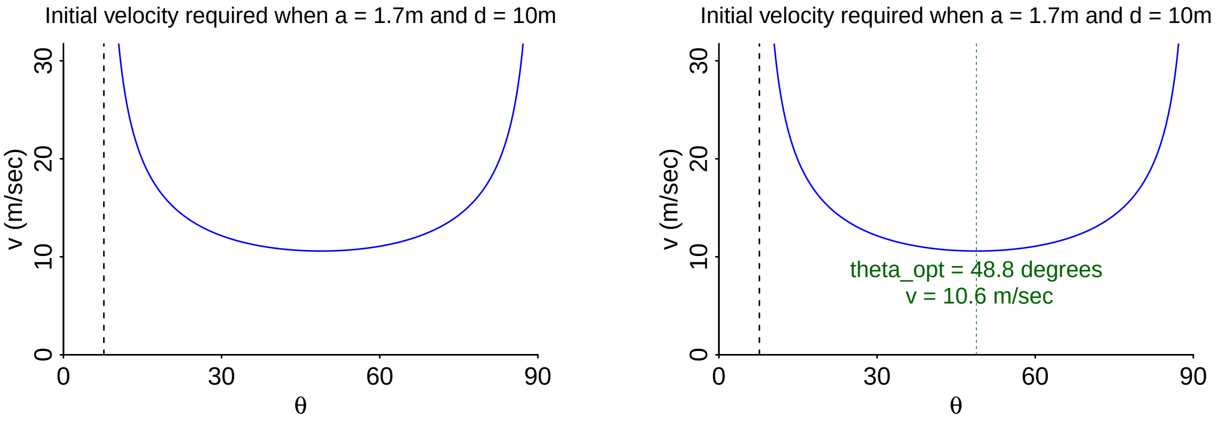

Figure 4: *(Left:) Velocity required to reach the basketball hoop for a shot released at an altitude of 1.7 meters, 10 meters away from the basket. (Right:) On this graph, the required velocity v is minimized at an angle $\theta = 48.8°$, at which point $v = 10.6$ m/sec. The blue color of the lines matches the blue used to display the hoop-reaching solution in Figure 3.*

Plugging in $a = 1.70, \theta = 30°, d = 10, h = 3.05, g = 9.8$ into this equation yields $v = 12.2$. The resulting trajectory is the blue line in Figure 3. It is helpful to see it in the context of the red curves that show other velocities.

Now that we have an equation for the required initial velocity for the ball to reach the hoop, we can plot it as a function of the shooting angle, θ. The left plot of Figure 4 shows the result. For the lowest angles ($\theta \leq \tan^{-1}(\frac{h-a}{d})$, indicated by the dashed vertical line on the plot), the ball will go below the hoop no matter how hard it is thrown. The lowest initial velocities are needed for shots taken from the 30°–60°, with harder shots required when the angle goes outside that range.

We can then use a general-purpose numerical optimizer to find the value of θ that minimizes (1), given $a = 1.70, d = 10, h = 3.05, g = 9.8$. The result is $\theta = 48.8°$, yielding $v = 10.6$. This result is displayed as a green dotted line in the right plot of Figure 4. Again, we find it helpful to display both graphs, as the show the development of the analysis: first the curve, then the minimum, which we label as the "optimum" in the sense of requiring the softest throw to reach the hoop.

Compared to Figure 3, Figure 4 represents a leap in abstraction: the angle θ and the initial velocity v are represented by linear coordinates, and the curve is not a trajectory; it si the solution to an equation. Again, in statistics we are so familiar with this sort of graph that we don't always remember how abstract it is.

The punch line of Figure 4 is the recommended shooting angle, θ_{opt}, and corresponding initial velocity v, for a shot taken at an altitude of 1.7 meters, 30 meters from the basket. The next level of abstraction is to ask how these parameters change as a function of distance. Obviously v will increase with distance, but at what rate? And what will happen to θ_{opt}?

We answer this question by fixing $a = 1.7$ and considering a grid of values of d, ranging from 1 to 15 meters from the hoop. For each value of d, we numerically optimize and find the value of θ that minimizes (1), along with the corresponding value of v.

The left panel of Figure 5 shows the results. When you are standing close to the basket you need to shoot at a steep angle to reach the hoop; as the distance increases, the optimal angla approaches 45°. The required release velocity increases gradually, at a rate slightly less than linearly with distance. We have lined up the two plots vertically so they can share an x-axis and the reader can see how both θ_{opt} and v vary together.

Finally, we gain one more level of generality by considering different release points: 1.2 m, 1.7 m,

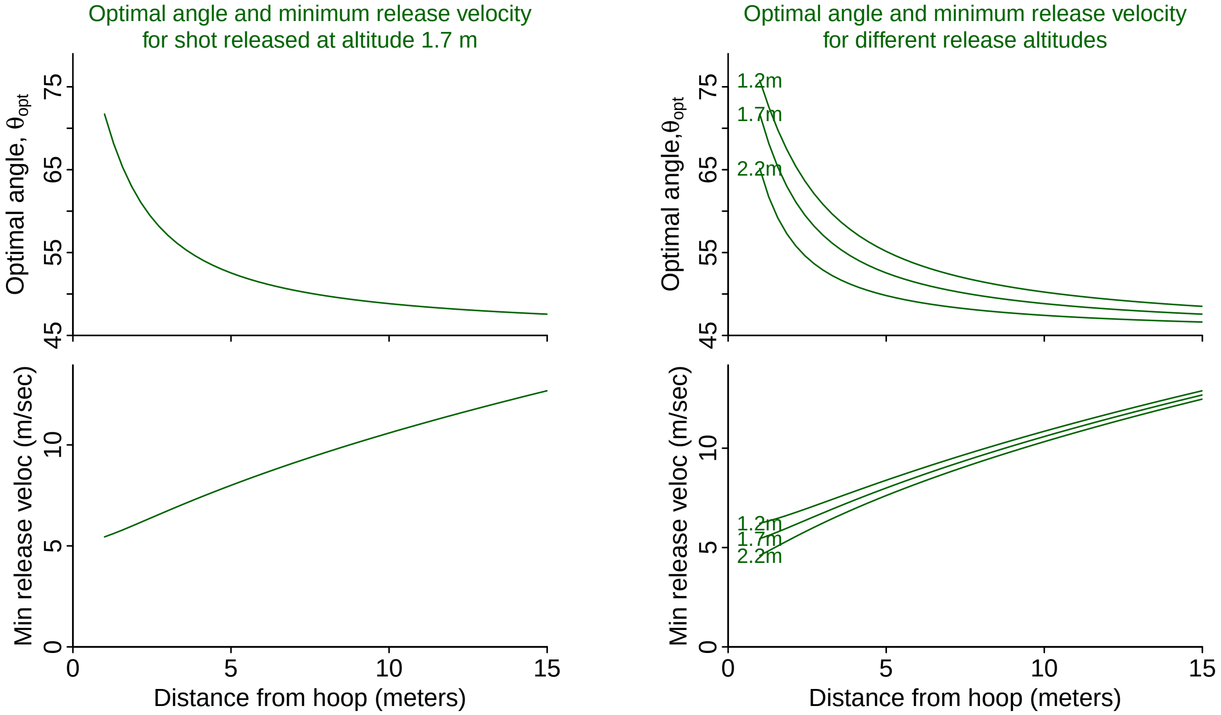

Figure 5: *(Left:) Shooting angle requiring the softest shot and the corresponding release velocity, as a function of the distance from the hoop, for a shooter releasing the ball at an altitude of 1.7 meters. (Right:) Same graphs, also showing the solutions for release altitudes of 1.2 and 2.2 meters. The green color of the lines matches the green used to display the minimum-initial-velocity solution in Figure 4.*

and 2.2 m off the ground, which might correspond to shooting from chest level, head level, or above the head. The right panel of Figure 5 shows curves for each of these three heights. Unsurprisingly, releasing from a greater altitude requires a lower shooting angle and a lower initial velocity.

The right panel of Figure 5 represents the greatest level of abstraction in this series of graphs, and, again, we find it helpful to build up to it rather than simply presenting these results and then expecting readers to follow. That said, we have tried to use graphical cues as much as possible to make plots more intuitive:

- The shared x-axis signals that the two plots represent a single analysis,
- The x-axis has a hard bound at 0 because it is not possible to shoot from a negative distance,
- The y-axis on the top graph has a hard bound at 45° because, with the release point lower than the basket, 45° is a lower bound for the optimal angle and also a useful reference point,
- The y-axis on the bottom graph has a hard bound at 0 because it is not possible for the velocity to be be negative, and 0 is a useful reference point for comparing velocities.

These choices depend on the context of the problem, which implies that some direct involvement on our part was needed; it would not be enough to simply use default plotting options, even if such defaults could give us good starting points.

In addition, throughout this example we have attempted to guide the reader using a succession of colors:

- *Black* for the basketball court in Figures 1—3 and the axes in Figures 4—5,
- *Red* lines for the trajectories in Figures 1—3.
- *Blue* line for the trajectory in Figure 3 that reaches the hoop, and also for the curves in Figure 4 that represent the initial velocity as function of angle for these shots,
- *Green* for the optimal angle that minimizes the required initial velocity in the right plot of Figure 4, and also for the curves in Figure 5 that show this optimal angle and corresponding initial velocity as a function of distance and release altitude.

The colors climb the ladder of abstraction, from direct mapping of trajectories to the solutions of a series of equations. When presenting the series of graphs we do not draw attention to the colors—that might feel like overkill and distract readers from the content being graphed—but we hope the colors help, in the same way that good design can make a tool easier to use even for users who are unaware of these choices.

3. Visualizing data: income and voting within and between states

When writing a book on quantitative trends in politics (Gelman et al., 2009), we were particularly attuned to the challenges of communicating to a general audience. It was important to us to use graphs and not just text because we wanted to actively engage readers in the process of understanding and discovery. Rather than simply *saying*, for example, that richer people were more likely than poor people to vote for Republicans, but Democrats did better in rich states, we wanted to *show* it.

Figure 6 juxtaposes the two patterns for George W. Bush in 2004, showing he did better among richer voters but worse within poorer states. Both these graphs are abstract, but in different ways:

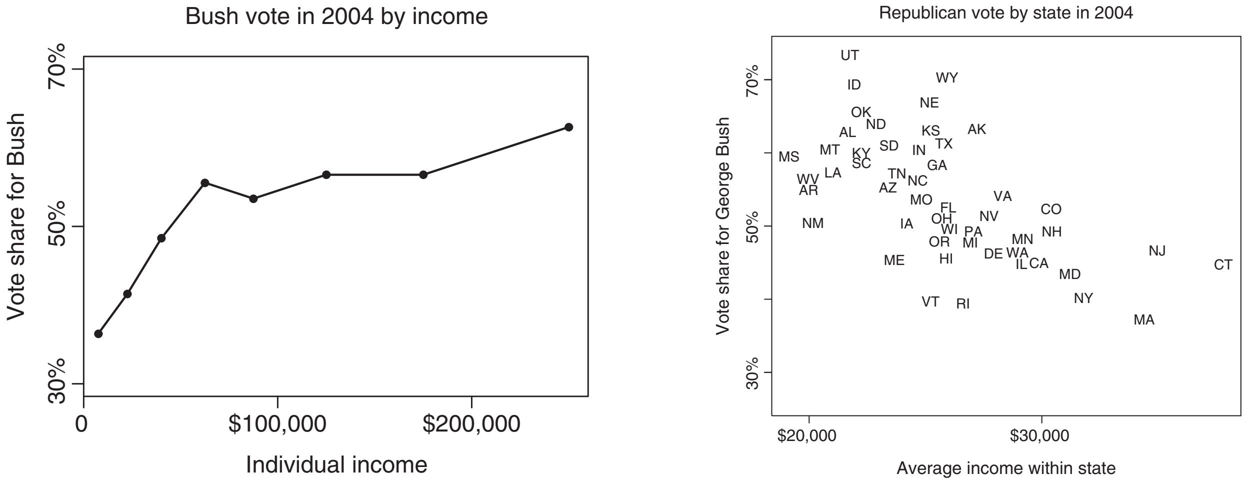

Figure 6: *(Left:) From exit polls, national voting patterns showing that George W. Bush,the Republican candidate for president in 2004, did better among richer voters. (Right:) At the same time, Bush was more successful in poorer states.*

in the left plot, each dot represents a set of survey respondents with a common response to the income question; in the right plot, the horizontal and vertical positions of each point show average income and vote share in a state.

The contrasting patterns for individuals and states motivated us to look at income and voting within states. Figure 7 we shows: Mississippi (the poorest state), Ohio (a middle-income state), and Connecticut (the state with highest median income). You might notice that these are smooth lines rather than the connected points in the left panel of Figure 6. In Figure 7 the lines display summaries of a fitted hierarchical logistic regression model—each of the lines is a part of a logistic curve, with each being on a narrow enough range that the lines are nearly straight.

It might seem like cheating in Figure 7 to display fitted curves rather than raw data, but you can think of these as smoothed data or as an alternative to presenting logistic regression coefficients. Going up one rung on the ladder of abstraction allows us to visually compare the states' patterns.

Once we have abstracted state and national voting patterns as in Figure 7, we can visualize changes over time by juxtaposing a series of such plots, which we do in Figure 8. These new graphs have an additional feature that connects the patterns within and between states. Within each state, open circles indicate the relative proportion of voters of each of the five income categories, and the solid circles plot the Republican candidate's vote share in the state against the average income on this five-point scale. The pattern of the three solid circles in each graph shows how poorer states are more Republican-leaning, even while there is a positive correlation between income and Republican voting within each state.

Incidentally, the lower-right graph in Figure 8 corresponds to 2004, the same year as displayed in Figure 7, but the lines are slightly different because the estimates come from a different data source. Perhaps it would make sense to include uncertainty estimates in these graphs to convey estimation errors? We considered options such as error bars or shaded uncertainty zones but decided not to, as we were worried about cluttering that would make it more difficult to make visual comparisons. In addition, the variation of the estimates from year to year gives some sense of estimation uncertainty. In any case, there is no right answer here; there is an inevitable tradeoff between clarity and the presentation of more information.

Indeed, one could take another step of abstraction and summarize each fitted curve by an

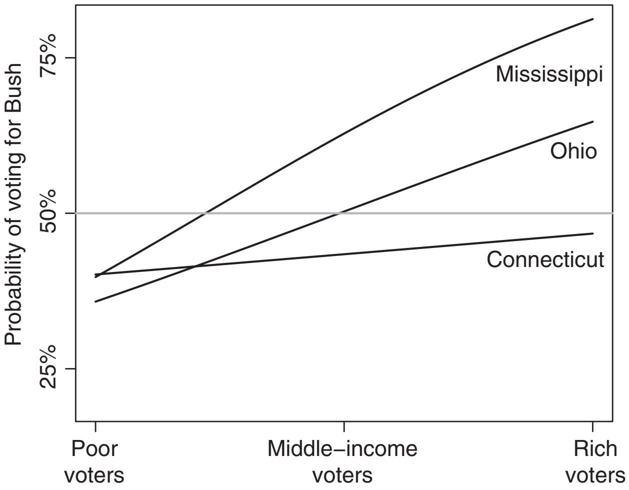

Figure 7: *From exit polls, George W. Bush's vote share in 2004, as a function of income, in a poor state, a middle-income state, and a rich state.*

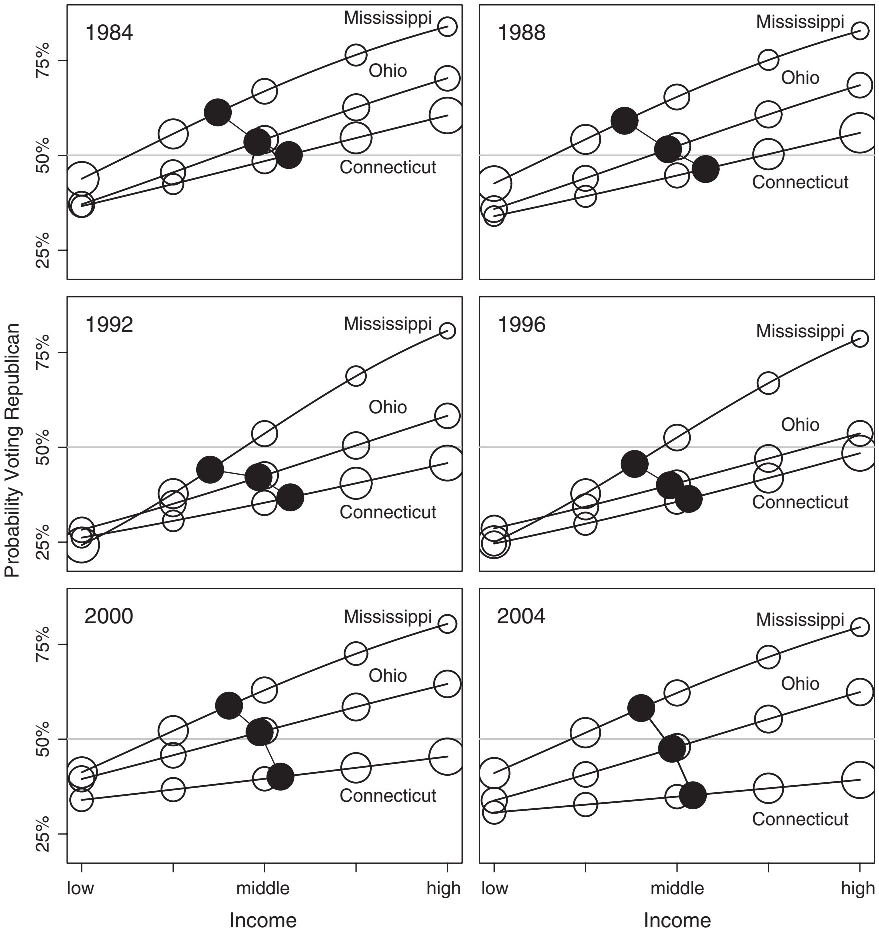

Figure 8: *Probability of Republican vote as a function of income in poor, middle-income, and rich states in six straight presidential elections. The open circles on the plot show the relative size of each income group in each state (thus, Mississippi has more poor people than average, and Connecticut has more rich people). The solid circles show the average Republican vote and average income in each state.*

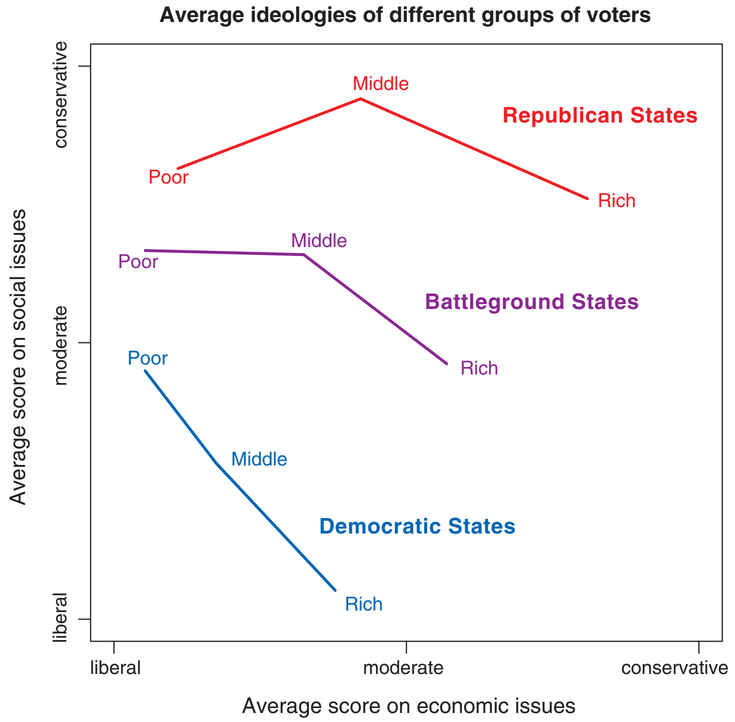

Figure 9: *From a large national survey in 2000, average positions of respondents on social and economic issues among voters in different income levels, within Republican-leaning, battleground, and Democratic-leaning states. Richer voters are consistently more conservative on economic issues, but the patterns on social issues are more complicated.*

intercept and slope on the logistic scale (with the income predictor coded as $-2, -1, 0, 1, 2$ so the intercept would be directly interpretable and then make a scatterplot of slope vs. intercept with two-letter state abbreviations. Such a graph could be difficult to explain on its own but could be a natural followup to Figure 7, showing fifty states instead of just three.

To better understand what was happening with income and voting, we went to polling data and created left-right scores on economic and social issues by combining several survey items in each area. The estimates were based on a large national poll from 2000 with enough data to get stable estimates for subgroups characterized by individual incomes and state political leanings. Figure 9 shows the result, which uses lines, colors, and labels to enable comparisons across voters and states in both dimensions (social and economic).

Once Figure 9 has been understood, we can go one more level of abstraction to Figure 10, which displays the same comparisons but separately for people who attend church regularly, occasionally, or not at all. As in the earlier graphs, we did not want to clutter the display with standard errors for all these averages, but uncertainty can be deduced from apparently random variation in the plots.

The effectiveness of graphs such as Figure 10 will depend on the problem being studied and on the data being plotted. Here it helped that the lines for the three groups of states were clearly separated and that the paths from poor to middle-income to rich voters and generally similar trajectories. As with any applied problem, this is not the end of the story. There are many other ways of looking at data on income and voting, and the patterns we showed above have changed in the years since those graphs were published (see, for example, Feller et al., 2012, and Gelman and

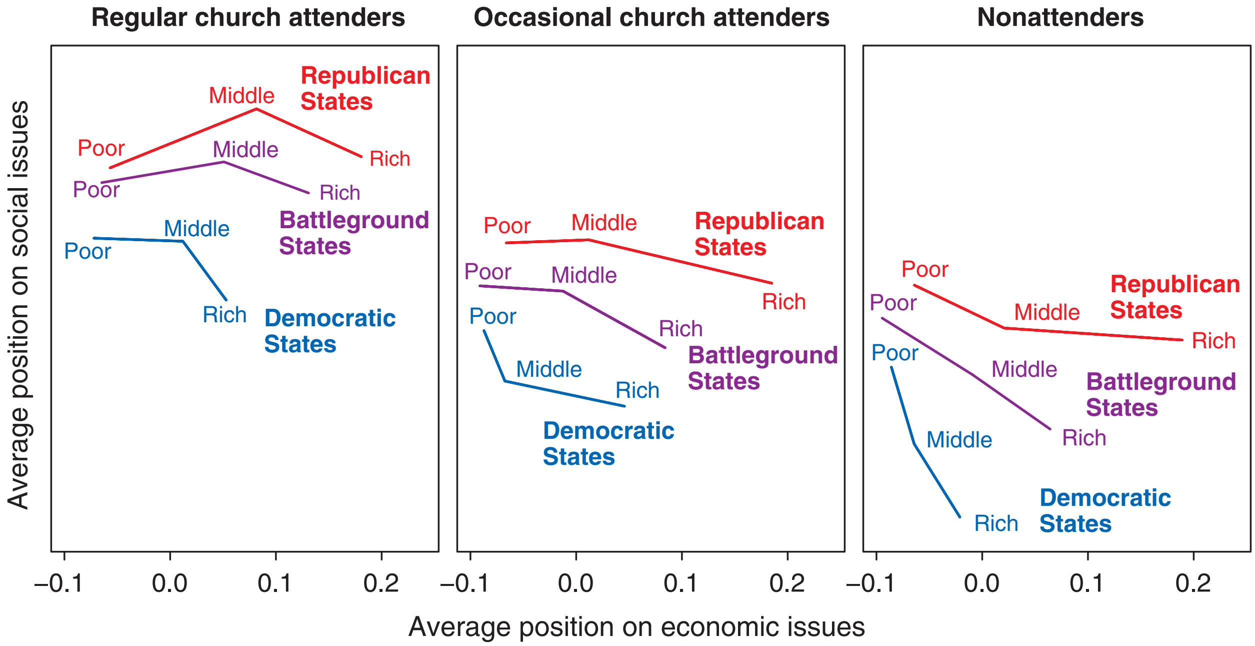

Figure 10: *Continuation of Figure 9, further characterizing people by religious attendance. This plot is complicated, but it is possible to read it by building up from earlier, less-abstract graphs.*

Azari, 2017).

4. Discussion

There is a long history of abstraction in visual representation, going far beyond statistical graphics. Consider, for example, the phrase "the map is not the territory" (Korzybski, 1933) and the satirical vignette of Borges (1946) pointing out the uselessness of a perfectly faithful 1-1 scale map. Figure 11 gives an amusing example discussed by Wainer and Friendly of an early infographic that cleverly conveys information directly through a non-abstract image. But when it comes to displaying mathematical or statistical relationships, we typically take for granted the abstraction of a scatterplot or a graph of y versus x.

Modern treatments of data visualization or the grammar of graphics consider graphs not as static drawn objects but as overlays of information, so that a useful plot is built by adding elements to a general structure (Wilkinson, 2005, Wickham, 2016). The ladder of abstraction discussed in the present paper represents a different sort of forward progression: rather than adding information to a graph, we are using increasing levels of generality. Our ideas follow the spirit of the grammer of graphics, however, in considering statistical graphics as a form of communication of quantitative information to the reader, with the specific graph being just an instantiation of this goal. As emphasized by Tufte (1983) and Cleveland (1985), choices in data display can be understood in with respect to the specific goals of communication. In present paper, we emphasize the challenge that useful visualizations can be highly abstract representations of the information being presented, hence our proposal of stepping readers through a series of graphs. We have demonstrated this idea in two different applications of statistical graphics: continuous curves and discrete data.

Once presented, the ladder of abstraction seems natural, but we have rarely seen it in published work, including our own. Our usual approach in graphical communication is to develop the graphs we want to display and then explain them in words. The examples here suggest that a more explicit

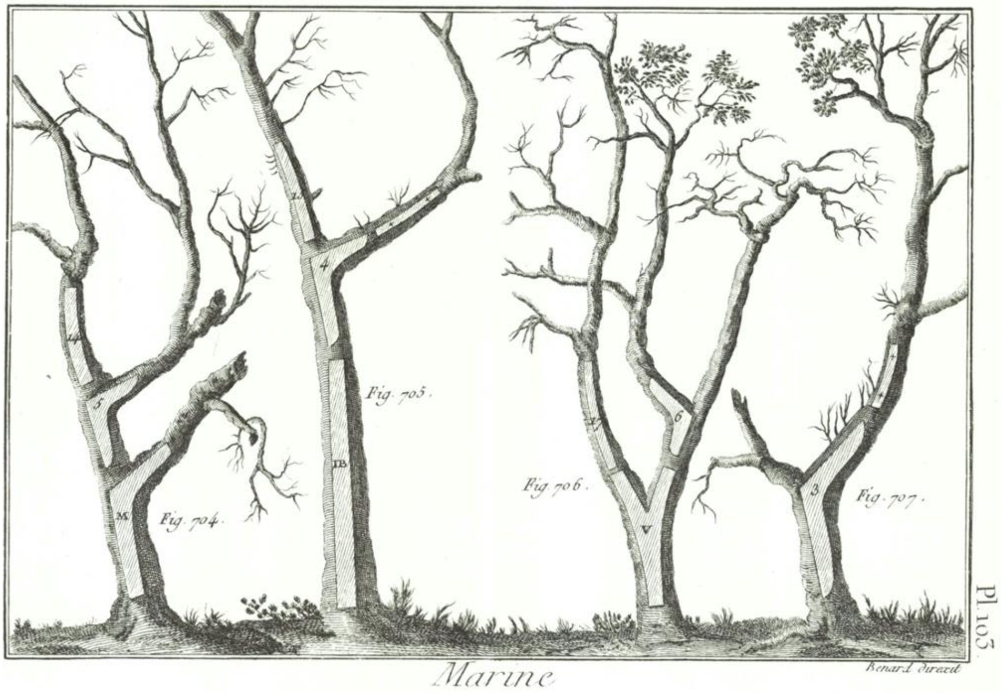

Figure 11: *Graph from Panckoucke (1783) illustrating the portions of trees that could be used to construct different parts of a ship. This is an example of a non-statistical visual representation at a low level of abstraction.*

unfolding could be useful, by analogy to the paradigm in statistical workflow of fitting increasingly complicated approximations, with each understood in comparison to what came before, an idea we have found helpful for modeling as well as computation (Gelman et al., 2020).

The present article has three messages. First, some level of abstraction exists in all but the simplest graphs, and we should be aware of this when using graphs to communicate, especially to readers who are not familiar with data graphics. Second, graphical abstraction is not just a difficulty, it also is a useful tool that allows us to understand patterns at a greater level of sophistication, and it can be valuable to see this by adding one level of abstraction at a time rather than trying to jump to a single plot that shows everything. Finally, these principles apply not just to graphs of data but also to graphs of mathematical functions such as fitted models.

References

Borges, J. L. (1946). Del rigor en la ciencia. *Los Anales de Buenos Aires* **1** (3).

Cleveland, W. S. (1985). *The Elements of Graphing Data.* Monterey, Calif.: Wadsworth.

Feller, A., Gelman, A., and Shor, B. (2012). Red state / blue state divisions in the 2012 presidential election. *The Forum* **10**, 127–141.

Gelman, A., and Azari, J. (2017). 19 things we learned from the 2016 election (with discussion). *Statistics and Public Policy* **4** (1), 1–10.

Gelman, A., Park, D., Shor, B., and Cortina, J. (2009). *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*, second edition. Princeton University Press.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Brkner, P. C., Kennedy, L., Gabry, J., and Modrk, M. (2020). Bayesian workflow. `https://arxiv.org/abs/2011.01808`

Hayakawa, S. I. (1939). *Language in Action.* New York: Harcourt, Brace.

Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer* **1**, 69–91.

Korzybski, A. (1933). *Science and Sanity*, 747–761. New York: International Non-Aristotelian Library Publishing Company.

Panckoucke, C. J. (1783). *Encyclopdie Mthodique Marine.* Paris: Chez Panckoucke.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information.* Cheshire, Conn.: Graphics Press.

Wainer, H., and Friendly, M. (2021). Displaying Information: From woolly mammoths to the Great Migration. In *International Encyclopedia of Education*, 4th edition, ed. R. Tierney, F. Rizvi, K. Ercikan, and G. Smith. Philadelphia: Elsevier.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag. `ggplot2.tidyverse.org`

Wilkinson, L. (2005). *The Grammar of Graphics*, second edition. New York: Springer-Verlag.