Good Graphs: Graphical Perception and Data Visualization

Nina Zumel*

August 28th, 2009

1 Introduction

What makes a good graph? When faced with a slew of numeric data, graphical visualization can be a more efficient way of getting a feel for the data than going through the rows of a spreadsheet. But do we know if we are getting an accurate or useful picture? How do we pick an effective visualization that neither obscures important details, or drowns us in confusing clutter? In 1968, William Cleveland published a text called *The Elements of Graphing Data* (`http://www.stat.purdue.edu/~wsc/elements.html`), inspired by Strunk and White's classic writing handbook *The Elements of Style* (`http://www.amazon.com/Elements-Style-50th-Anniversary/dp/0205632645`). *The Elements of Graphing Data* puts forward Cleveland's philosophy about how to produce good, clear graphs not only for presenting one's experimental results to peers, but also for the purposes of data analysis and exploration. Cleveland's approach is based on a theory of graphical perception: how well the human perceptual system accomplishes certain tasks involved in reading a graph. For a given data analysis task, the goal is to align the information being presented with the perceptual tasks the viewer accomplishes the best.

> When a graph is made, quantitative and categorical information is encoded by a display method. Then the information is visually decoded. This visual perception is a vital link. No matter how clever the choice of the information, and no matter how technologically impressive the encoding, a visualization fails if the decoding fails. Some display methods lead to efficient, accurate decoding, and others lead to inefficient, inaccurate decoding. It is only through scientific study of visual perception that informed judgments can be made about display methods. The display methods of *Elements* rest on a foundation of scientific enquiry.
>
> from the preface of *The Elements of Graphing Data*

A revised edition of *The Elements of Graphing Data* was published in 1994, along with a companion volume, *Visualizing Data* (*http://www.stat.purdue.edu/~wsc/visualizing.html*),

*`http://www.win-vector.com/`

which is oriented towards the implementation and technical details of different graphing techniques. I highly recommend *The Elements of Graphing Data* as a guidebook for creating graphs, as well as for its excellent survey of several useful techniques. Cleveland, along with other colleagues at Bell Labs, developed the Trellis display system (`http://stat.bell-labs.com/project/trellis/s.html`), a framework for the visualization of multivariable databases, using the ideas developed in his texts. Trellis, in turn, influenced Deepayan Sarkar's Lattice graphics system for R. Lattice implements many of Cleveland's ideas, and I also recommend Sarkar's Lattice manual (`http://lmdvr.r-forge.r-project.org/figures/figures.html`) if you do data visualization in R.

It's important to note here that Cleveland writes for researchers and decision-makers who use graphs to analyze data, or to convey scientific results to colleagues in an (ideally) objective manner. This distinguishes him from Darrell Huff, whose 1954 *How to Lie with Statistics* (`http://www.amazon.com/How-Lie-Statistics-Darrell-Huff/dp/0393310728`) considered the use of graphs (and statistics in general) as rhetorical devices for convincing others of one's point of view. Hence, some of Cleveland's recommendations and guidelines actually contradict Huff's.[1]

Edward Tufte also explored the idea that the choice of graphical display should be influenced by the viewer's cognitive processes, in his 1990 book *Envisioning Information* (`http://www.edwardtufte.com/tufte/books_ei`). Tufte tends to be more broadly concerned with the gestalt of a graph, beyond its use as an analysis tool; he is also more concerned than Cleveland is with aesthetic considerations.

Cleveland's philosophy might be summarized as: *minimize the mental gymnastics that the viewer must go through to understand the graph.* This leads to some obvious advice: avoid clutter and occlusion, make graphing symbols or color-coding unambiguous, use scale-lines on all four sides of the graph, and so on. It also leads to advice that perhaps should be as obvious, but isn't: *make the aspect of the data that you want to analyze as clear as possible.* But what does this mean in practice?

2 Make important differences large enough to perceive

Weber's Law is a well known observation from the psychophysics literature, which states that the "just noticeable" change in a stimulus is a constant ratio of the original stimulus. Put another way, people are only capable of detecting a change in a stimulus that is greater than a certain percentage k of the original stimulus. Here, "stimulus" can refer to any perceivable physical quantity: weight, intensity, length, orientation. The percentage k will vary with stimulus, and with observer.

[1] *How to Lie with Statistics* is an entertaining (if a little dated) discussion of how to read statistical and quantitative claims critically, and is definitely worth a read.

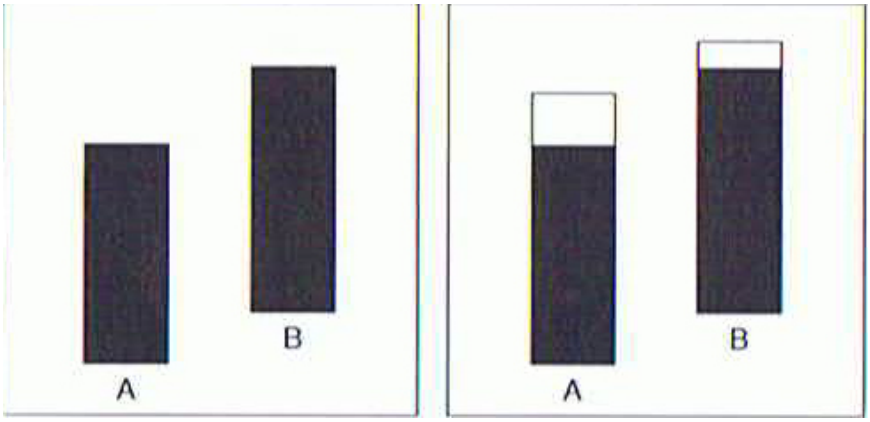

Figure 1: From Cleveland, *The Elements of Graphing Data*

Figure reffig:1 shows the application of Weber's law to lengths. The bars A and B are of different lengths, but the difference is such a small fraction of the "base" length (say, A's length, to be specific) that is difficult to tell whether or not they are different, or which is longer. On the right, the bars have been embedded in frames of identical length, and now it is easy to see that B is longer. Why? Because the difference in lengths of the *white* intervals is a much larger percentage of the white "base" length (say the white A interval). It is easy to see that the white B interval is shorter than the white A interval, and therefore, the black B interval is longer than the black A interval.

The moral is that you always want the viewer to be estimating changes or differences with respect to a short base length. You can do this with reference grids, as demonstrated below.

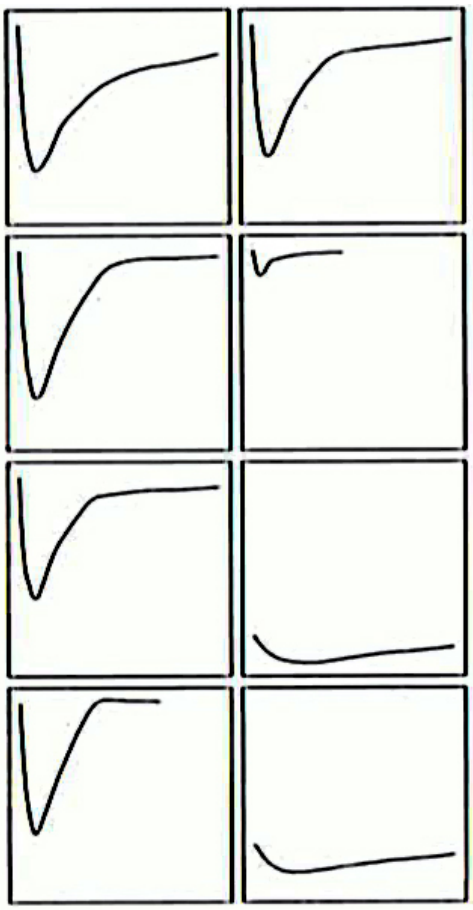

Figure 2: From Cleveland, *The Elements of Graphing Data*

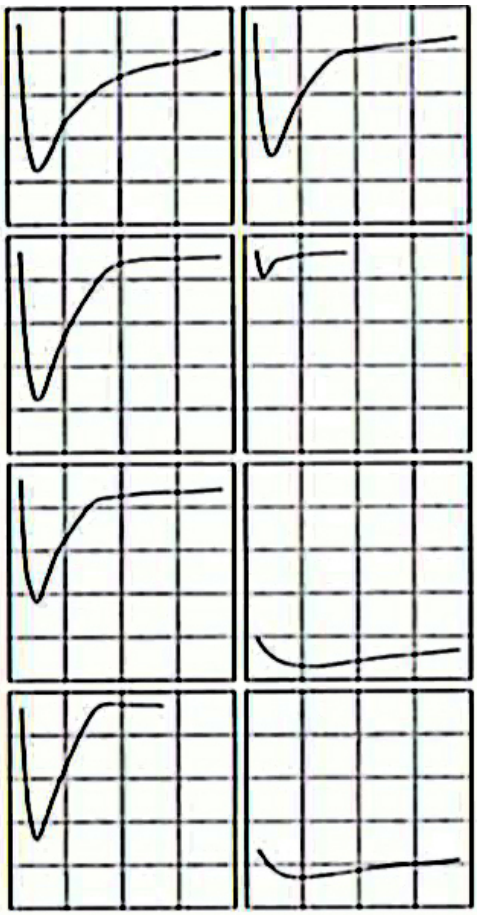

Figure 3: From Cleveland, *The Elements of Graphing Data*

Figure 2 shows eight curves. Which one dips to the lowest minimum? Are the high curves approaching the same value, and which one is rising the fastest? Are the low curves dipping to the same minimum? Are they going to the same steady state? Figure 3 shows the same curves, graphed with identical reference grids. The grids shorten the base lengths that are being compared, and it is now much easier to compare highs, lows, and steady state behavior.

But wouldn't it be better to compare the graphs by superposing them? For two or three curves, perhaps. But in this case, eight curves can clutter the graph, and use up the symbol or color space, making it difficult to distinguish the different datasets – increasing the mental gymnastics.

Reference grids are useful even for a single curve, especially one with slowly varying segments, such as these graphs have. The reference grid makes it easier to answer questions like: does the process return to the initial state, or to a different steady state? Has the process reached steady state, or is it still growing?

3 Make important shape changes large enough to perceive: Banking to 45 degrees.

The aspect ratio of a graph is important when trying to understand shape. Rate of change information is encoded in the slope of the curve, which the viewer estimates by changes in the orientation of the local tangents at each point of the graph. Weber's Law tells us that very small changes in this orientation will be difficult to detect. For a given (physical) curve, the local orientation changes will be dependent on the aspect ratio of its graphical presentation, as shown (to an exaggerated degree) in Figure 4. Here, the same curve (two line segments) is plotted at three different aspect ratios, one that centers the graph at 45 degrees, one that forces the curve to be nearly vertical, and another that forces it to be nearly horizontal. In the last two cases, the change in orientation of the two line segments is so small as to be nearly undetectable.

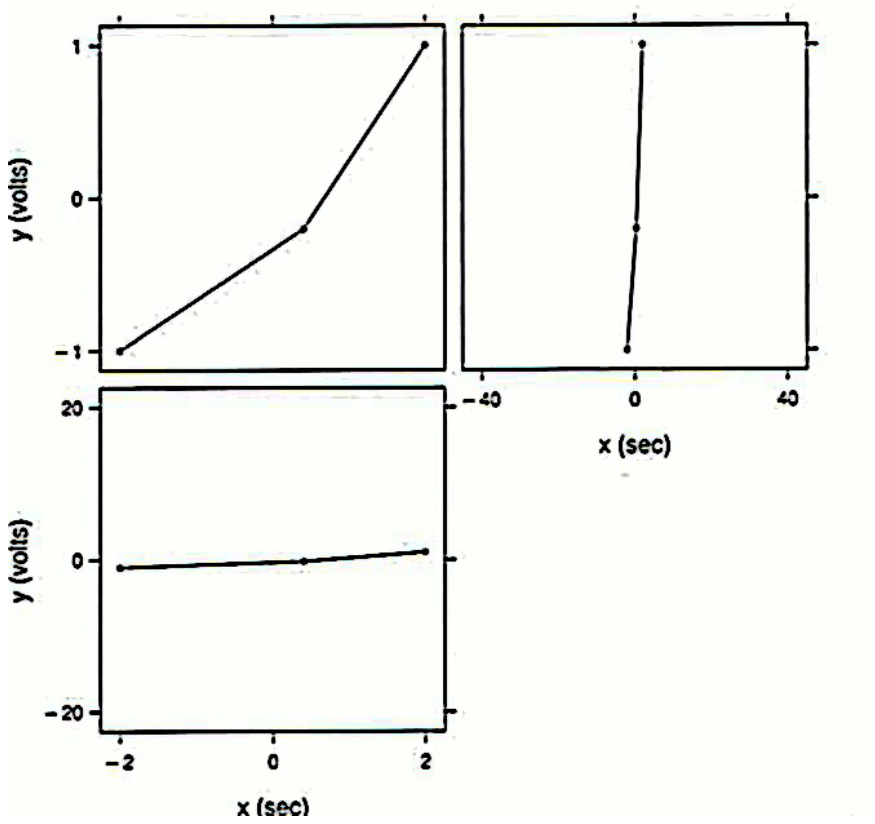

Figure 4: From Cleveland

For two line segments with positive, unequal slopes, a simple geometric argument shows that their absolute difference in orientation is maximized by the aspect ratio that sets their average orientation to 45 degrees (the first graph in Figure 4). Empirical studies by Cleveland and others have indeed verified that a viewer's ability to judge the relative slopes of line segments on a graph is maximized when the absolute values of the orientations of the segments are centered on 45 degrees.

This result leads to a technique called *Banking to 45*, whereby the aspect ratio of the graph is chosen so that the average slope of the entire graph is 45 degrees. The details are discussed in Cleveland, and many of the plots in R's Lattice package also have an option to bank the graph to 45 degrees.

This deliberate exaggeration of slope is something that Darrell Huff deplores. In *How to Lie with Statistics*, Huff refers to these graphs as "gee-whiz" graphs and in the context of his discussion of statistics as rhetoric, they are:

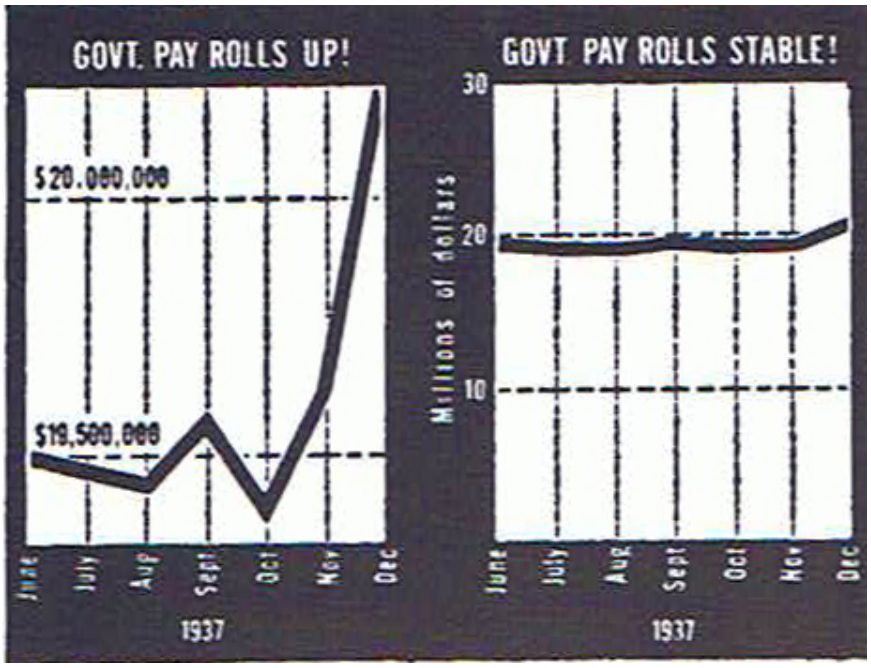

Figure 5: From Huff, *How to Lie With Statistics*

To insist that a graph should always include a zero line and that units be in proportion may be good advice from a rhetorical perspective; but it is poor advice if the purpose of the graph is data analysis. As Figure 6 below demonstrates, we can lose resolution if we always insist on including the zero. Does the trend line in the left graph increase linearly, superlinearly, or sublinearly? The convexity of the curve is more apparent when it is banked to 45, as on the right. Assuming that the scientist reads the axis and is cognizant of the actual magnitude changes involved, the graph on the right conveys more information.

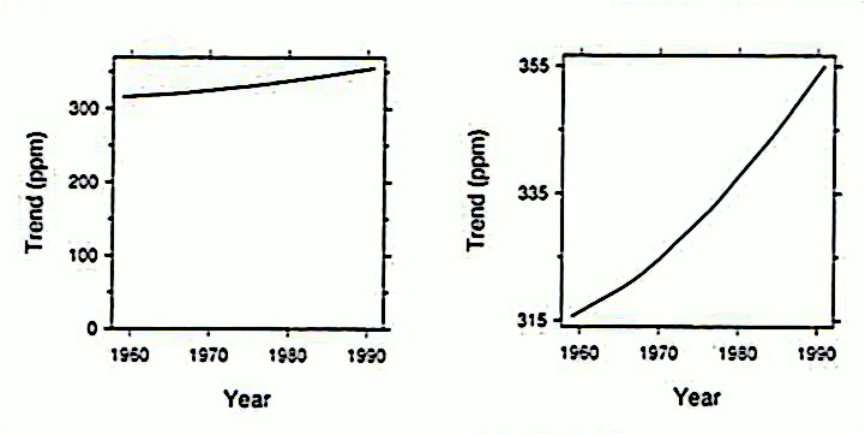

Figure 6: From Cleveland

4 Make sure all the data is equally well resolved.

It is quite common for positive data word frequencies, populations, price distributions, just to name a few examples to be skewed: most of the data is bunched towards low values, the rest of it is spread out on a very long tail. This long tail squashes the majority of the data into a tiny interval of a very narrow dynamic range, as in Figure 7, making it difficult to evaluate the data.

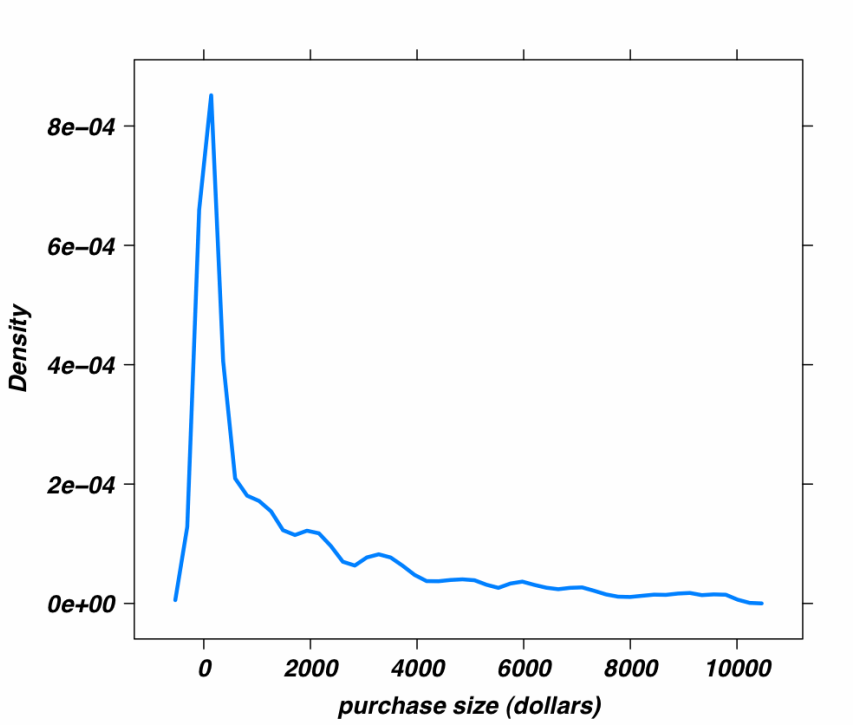

Figure 7: Long-tailed distribution of purchase sizes

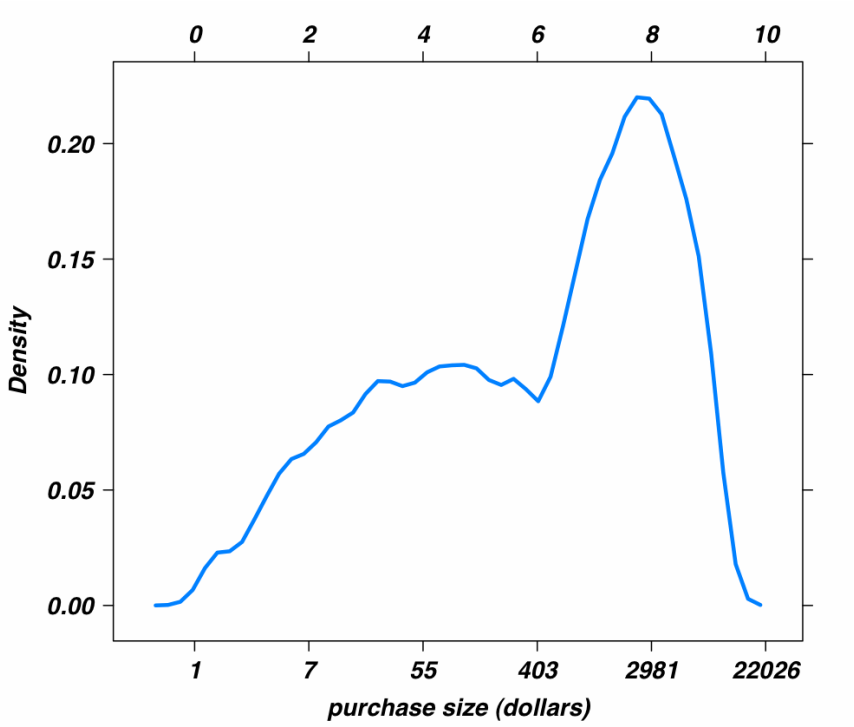

Figure 8: Distribution of log(purchase size)

Imagine that Figure 7 represents the distribution of average purchase size across an online merchant's customers: average purchase size is plotted on the x-axis, and the y-axis represents the fraction of the total customer population whose average purchase size is a given value (the area under the graph integrates to one). According to this graph, most customers make fairly small purchases on average, but there is a long tail of big spenders trailing out into the range of several thousand dollars. Obviously, one would like a little more resolution on the big spike of customers near zero. One could simply "zoom in" on this range, by chopping off some long chunk of the tail, but you may potentially lose sight of some global patterns in the data by doing so.

Graphing the distribution of log(purchase size) enables you to increase the resolution near zero, while preserving the global view. Figure 8 shows the distribution of log(purchase size), revealing two spending populations: a population of high spenders who tend to make purchases in the $3000 range (in log space), and another population whose purchases are centered (in log space) around $60. The existence of these two distinct populations is not apparent in the original graph.

Notice that Figure 8 has two x-axis scales: the top axis is marked in log units, while the bottom axis is marked in absolute dollars, spaced on a log scale. This accords with the principle of minimizing mental gymnastics, since the viewer of the graph will typically be concerned about prices in dollars, not log dollars. In fact, it would have been better yet to have plotted the distribution of $\log_2$ or $\log_{10}$ of the data; the former would allow us to see at a glance the doubling of price ranges, the latter to see price changes in factors of ten.

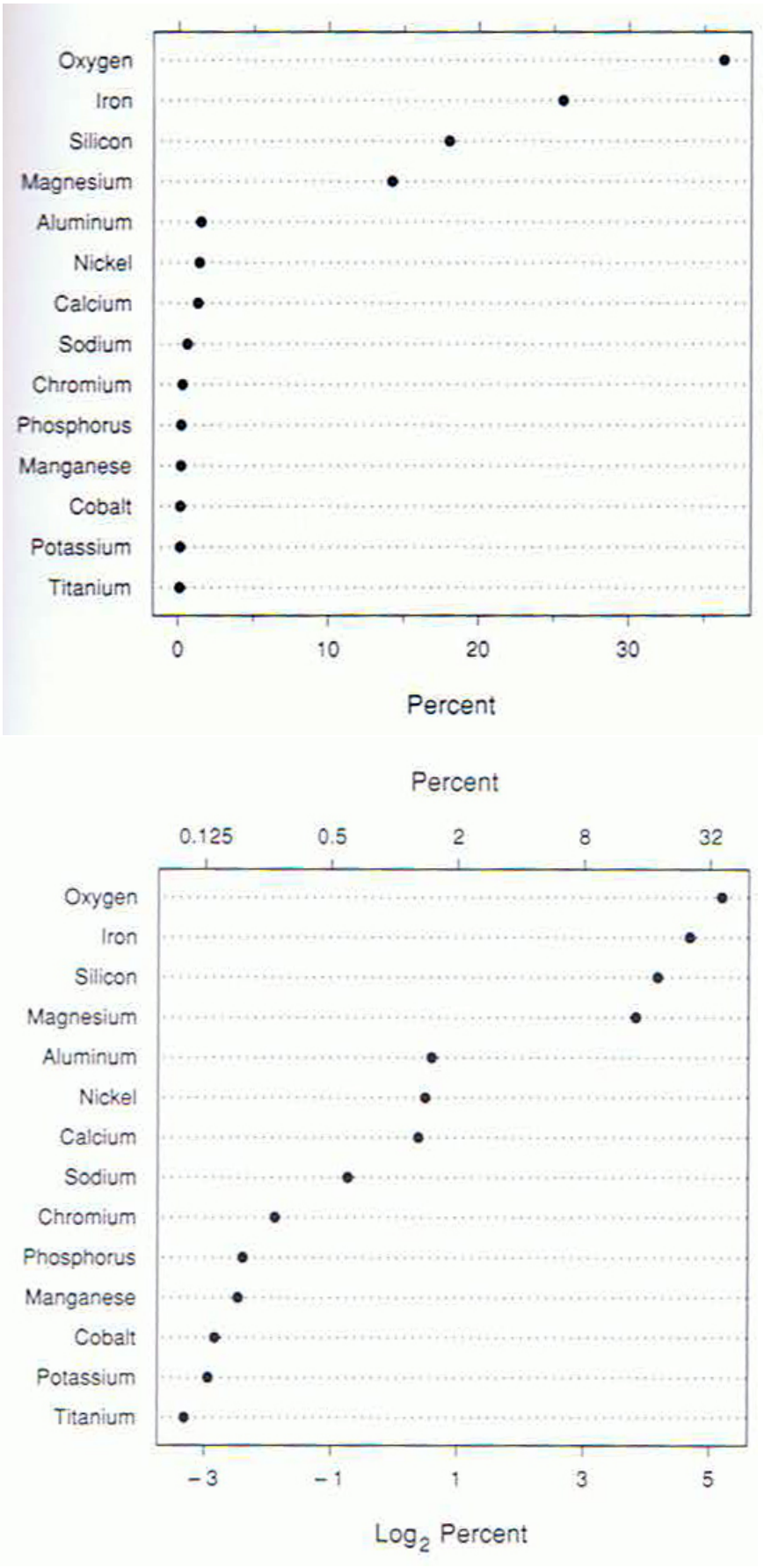

Figure 9: The 14 most abundant elements in meteorites. From Cleveland

Figure 9 shows another example: the fourteen most abundant elements in meteorites, specifically the average percent of each of the elements. If we graph the percentages directly, as on the left, we cannot easily distinguish the differences in the elements from aluminum on down. Graphing $\log_2$ of the percentages, as on the right, improves the resolution. Again, we have two x-axes on the graph of the log data.

5 If you want to analyze the difference between two processes, then graph the difference, not the processes (or graph both).

Suppose that we are comparing the two processes f1 and f2 that are shown in Figure 10. As x increases, the two processes appear to be approaching each other that is, the difference between the two seems to be decreasing. In reality, the difference between the two is constant: f2 = f1+1.

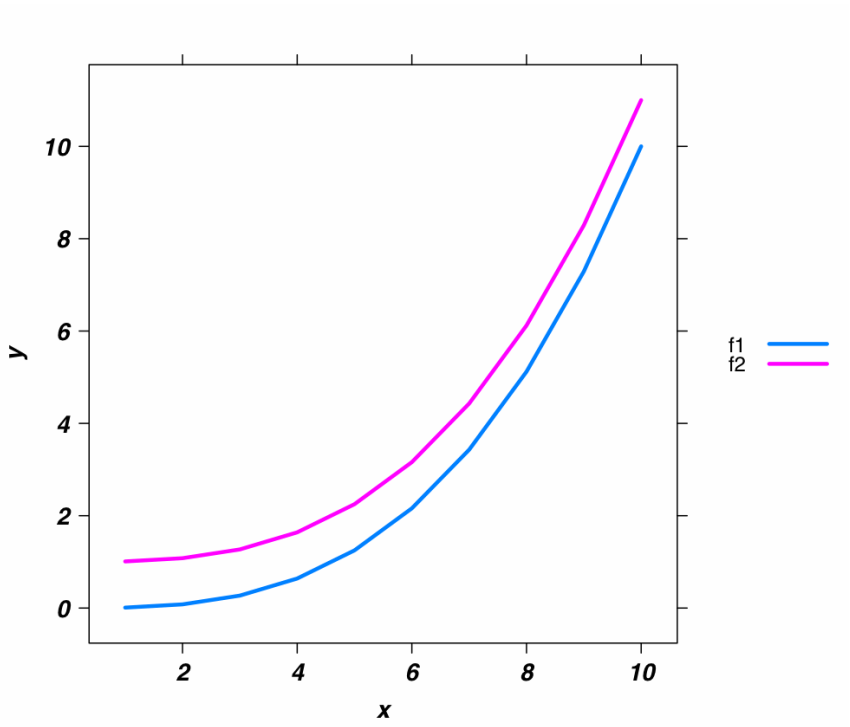

Figure 10: The illusion of convergence

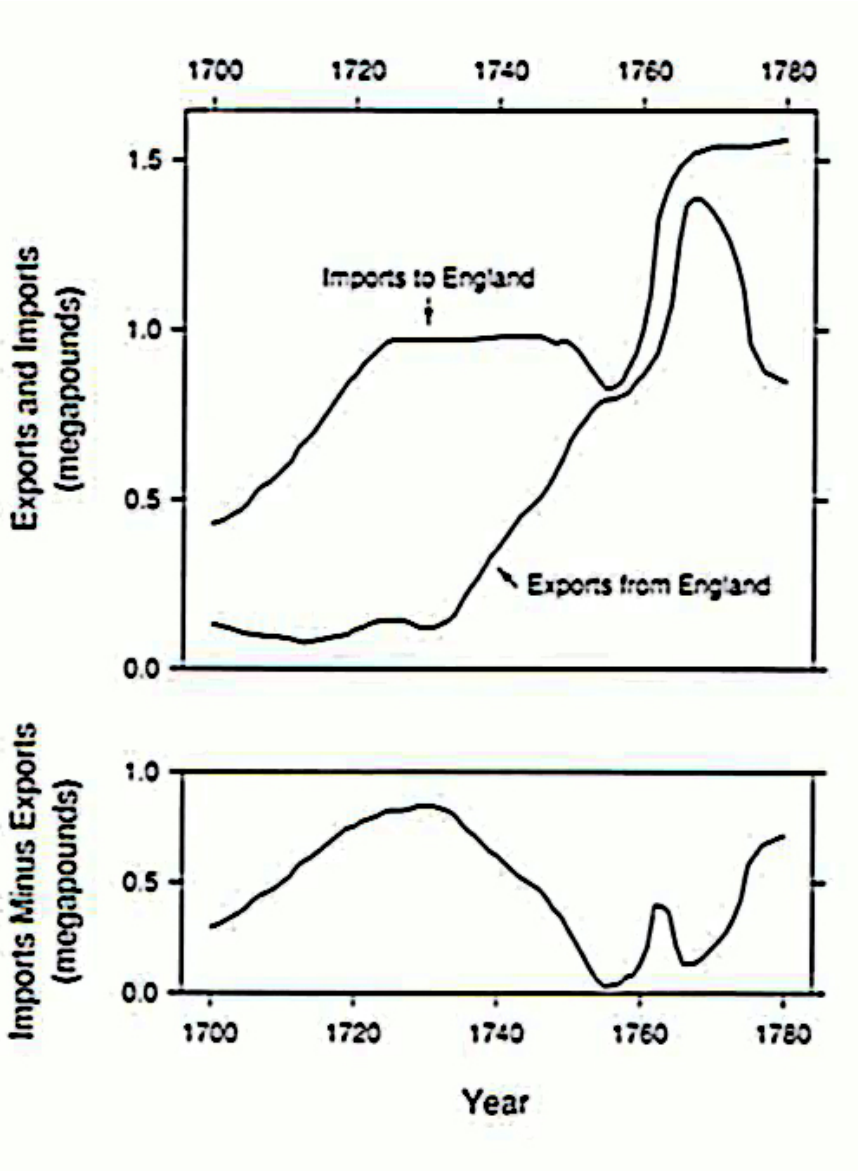

Figure 11: British Imports and Exports. From Cleveland

It turns out that people are good at perceiving the perpendicular difference between two curves, but not the differences in height, which is what we are actually interested in here. When we try to infer the differences from the process graph, we may not only miss key information, we may actually draw incorrect conclusions.

A less toy example is given in Figure 11. Here the imports to and exports from England are graphed over the first 80 years of the 18th century. In the difference graph on the bottom, we can see a local peak in (imports-exports) just after 1760; this is not obvious from simply comparing the two processes (top graph).

6 If you are interested in rate of change, then graph rate of change.

In Figure 12, we see the population figures for a given community from 1990 to 2009. Obviously, the population is steadily increasing, but how quickly? Is the rate of population growth increasing over time, or is it decreasing? If we are interested in these questions, then simply graphing the population over time is not enough. We need to look at the rate of change directly.

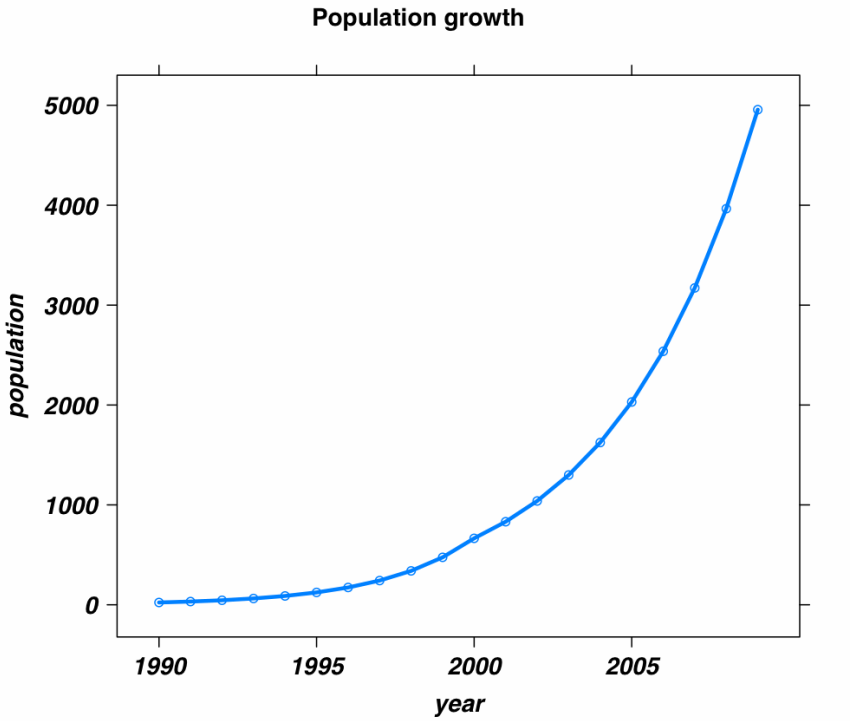

Figure 12: Population

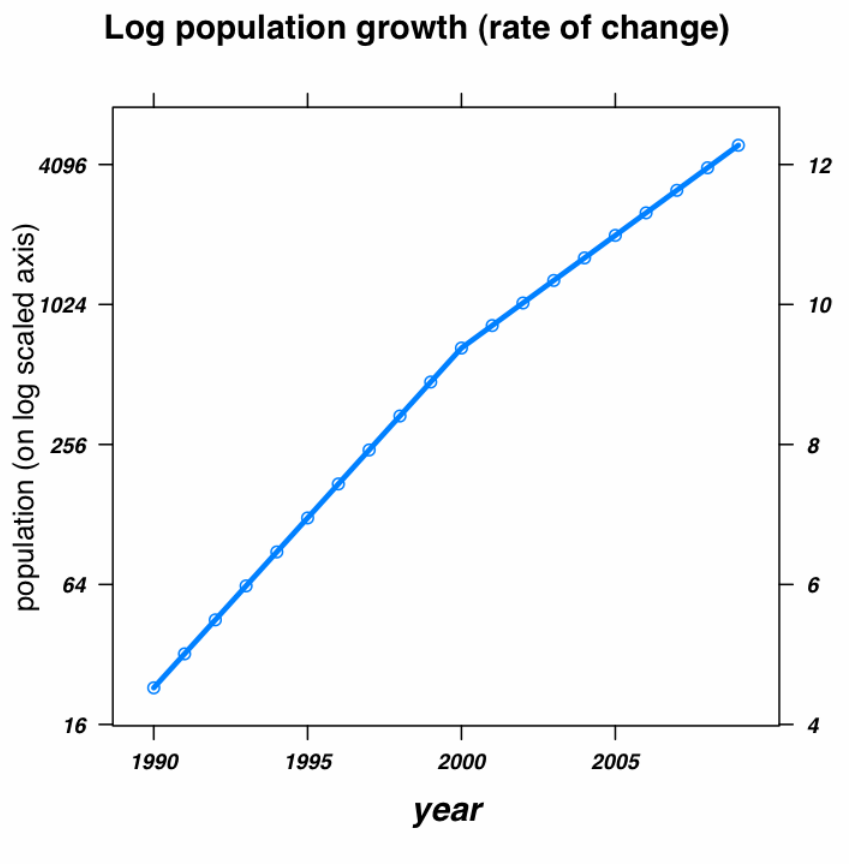

Figure 13: Log Population

The classic way to do this is by graphing the logarithm of the data. In Figure 13, we have graphed $\log_2$ of the population over time, with the log scale printed on the right hand y-axis, and the actual population numbers printed at a log scale on the left hand axis. Now we can see that the population increased at a constant rate from 1990 to 2000, quadrupling approximately every four years, and then slowed down (to a lower constant rate) after 2000.

7 Graphs as a research tool

Throughout this discussion, we have considered graphs as a tool for data exploration and initial understanding. It is an iterative process – as questions arise, the data will be reprocessed and re-plotted to highlight the new issues to be examined. A good research graph must display this information directly, with a minimum of mental gymnastics, but – as with any research tool – there can be a learning curve. For example, densityplots (such as those shown in Figures 7 and 8) are in my opinion more useful than histograms for understanding how numerical data is distributed – and I am constantly surprised at the amount of explanation that they require when I show them to people who are unfamiliar with them. A number of very useful graphs that are discussed in Cleveland's texts meet with the same reaction from people who encounter that style of graph for the first time. This is a

disadvantage, relative to using a more fashionable graph, when attempting to communicate results. But the insight into the data that these graphs provide often make it worth spending the time to educate clients or peers on how to read the graph.

Even so, a good graph still may not be a quick read. As Cleveland writes:

> While there is a place for rapidly-understood graphs, it is too limiting to make speed a requirement in science and technology, where the use of graphs ranges from detailed in-depth data analysis to quick presentation. ...
>
> The important criterion for a graph is not simply how fast we can see a result; rather it is whether through the use of the graph we can see something that would have been harder to see otherwise or that could not have been seen at all.
>
> - *The Elements of Graphing Data*, Chapter 2