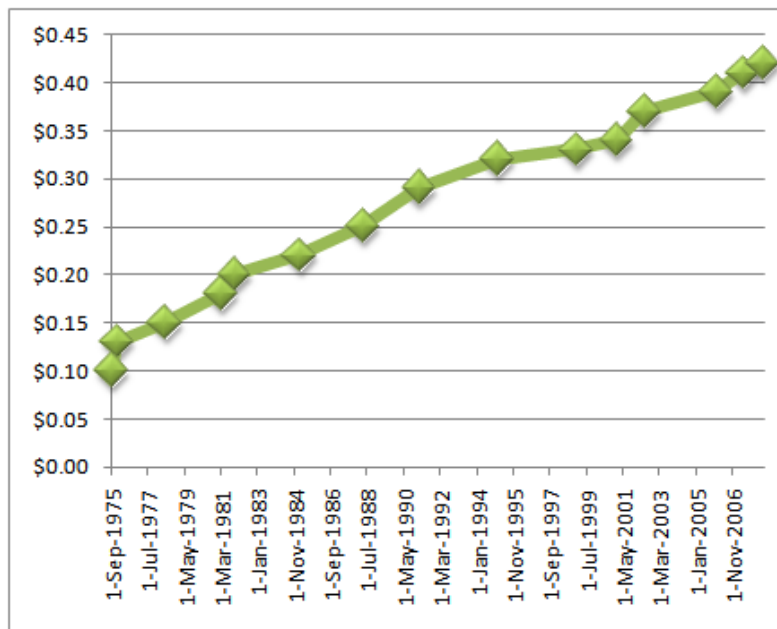


## Line Graphs and Irregular Intervals: An Incompatible Partnership

Stephen Few, Perceptual Edge  
*Visual Business Intelligence Newsletter*  
 November/December 2008

I recently received an email from Colin Banfield, a fellow who has read my books and follows my work. In it, Colin invited me to join a discussion about line graphs and unequal intervals of time. The occasion for his email was a disagreement between Jon Peltier and him that you can read on Jon's blog at [www.peltiertech.com](http://www.peltiertech.com). In his email, Colin wrote: "I have a long-standing battle with Jon Peltier, who believes that it's OK to chart unequal date intervals on a line chart." Jon is an Excel expert extraordinaire—one of a few Excel aficionados that I correspond with from time to time. In contrast to our usual agreement, Jon and I appear to differ on this particular matter.

The debate between Jon and Colin originally began in response to the following graph of U.S. postage stamp rates, borrowed from [www.thefinancejourney.com](http://www.thefinancejourney.com), which Jon used in his blog as an example of data better suited to a step chart:

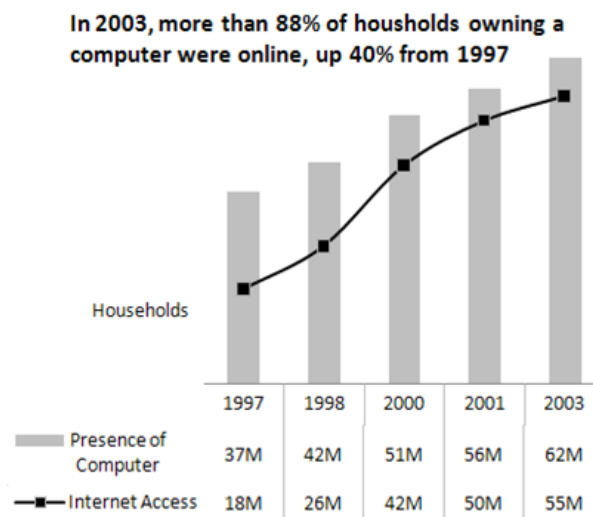


When Colin objected to the unequal intervals of time that appear along the X-axis, Jon responded:

*I don't understand the obsession with an equal date interval. A line chart need not show the trend of only evenly-spaced data. Suppose I am observing temperatures, and I decide for simplicity that where the temperature hasn't changed, or where it has been changing steadily, I do not need to record every value. Overnight after the temperature has dropped, I can characterize my temperature profile with one point per hour. As the sun rises, I may need more frequent recordings to capture the morning warm up. Then the clouds blow over, it starts to rain, then it clears up again; I may need minute-by-minute data points to track this. When I make my plot, is it any less relevant because the spacing of the data ranges from minutes to hours?*

On this matter, I side with Colin. Using a line to connect values along unequal intervals of time or to connect intervals that are not adjacent in time is misleading.

Their debate came to life again this month when Jon featured a [guest blog by Mike Alexander](#), the author of *Excel 2007 Dashboard and Reports for Dummies* (an excellent book), in which Mike presented “Ten Chart Design Principles.” After reading Mike’s blog, Colin breathed fresh life into the debate because a graph that appears several times in various forms displays unequal intervals of time. Here’s an example of the graph:



Notice that the years 1999 and 2002 are missing, yet nothing in the graph alerts us to this fact. Colin agreed with Mike’s charting tips, but couldn’t let these line charts with unequal intervals of time slip by without comment.

*Unfortunately, ... the line chart examples are misleading. You cannot simply compare equal intervals for some data points and then switch to a different interval and expect the result to be a meaningful trend of the plotted data...For all we know, there might have been declines in the years 1999 and 2002—the line chart constructed would mask this perfectly.*

## Corresponding Quantity and Matching Visual Encoding

When we present quantitative information graphically, we should take care to match the visual features of the graph to the perceptual inclinations of the reader’s mind. When lines are used in a graph to connect unequal or non-adjacent intervals of time, they misrepresent the information. Edward Tufte expressed the problem as follows:

*The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.*

*(The Visual Display of Quantitative Information, Edward R. Tufte, Graphics Press, Cheshire, CT, 1983, p. 77.)*

*Each part of a graphic generates visual expectations about its other parts and, in the economy of graphical perceptions, these expectations often determine what the eye sees. Deception results from the incorrect extrapolation of visual expectations generated at one place on the graphic to other places.*

*A scale moving in regular intervals, for example, is expected to continue its march to the very end in a consistent fashion, without the muddling or trickery of non-uniform changes.*

*(The Visual Display of Quantitative Information, Edward R. Tufte, Graphics Press, Cheshire, CT, 1983, p. 60)*

Nicely put. When encoding time graphically as distance along an axis, equal intervals of time should be displayed as equal intervals of distance.

Stephen Kosslyn, a Harvard cognitive psychologist who has applied his knowledge of perception and cognition to graphical design, also expressed the problem insightfully.

*Our visual system and memory system tend to make a direct connection between the properties of a pattern and the properties of the entities symbolized by that pattern. A continuous rise and fall of a line will naturally be taken to reflect a continuous variation in the entity being measured. If the changes in that entity are in fact not continuous but discrete, the continuity implied by a line graph is misleading; a bar graph would better represent the actual situation being depicted. The specific principle here is compatibility. The properties of the visual pattern itself should reflect the properties of what is symbolized.*

(*Elements of Graph Design*, Stephen M. Kosslyn, W. H. Freeman and Company, New York, p. 8)

Connecting values with a line along an interval scale such as time accurately represents reality only if (1) the intervals are equal in size, (2) the intervals are in proper order, and (3) values have been recorded for all intervals. Missing values constitute a discontinuity in the data that is not meaningfully represented by a continuous line. Connection suggests a constant slope of change between two points in time, when in fact the intervening states that are missing might have been quite different.

## Line Graph Best Practices

Based on these principles, we can derive a set of guidelines for line graphs.

### 1. Lines should only be used to connect values along an interval scale (with a couple exceptions).

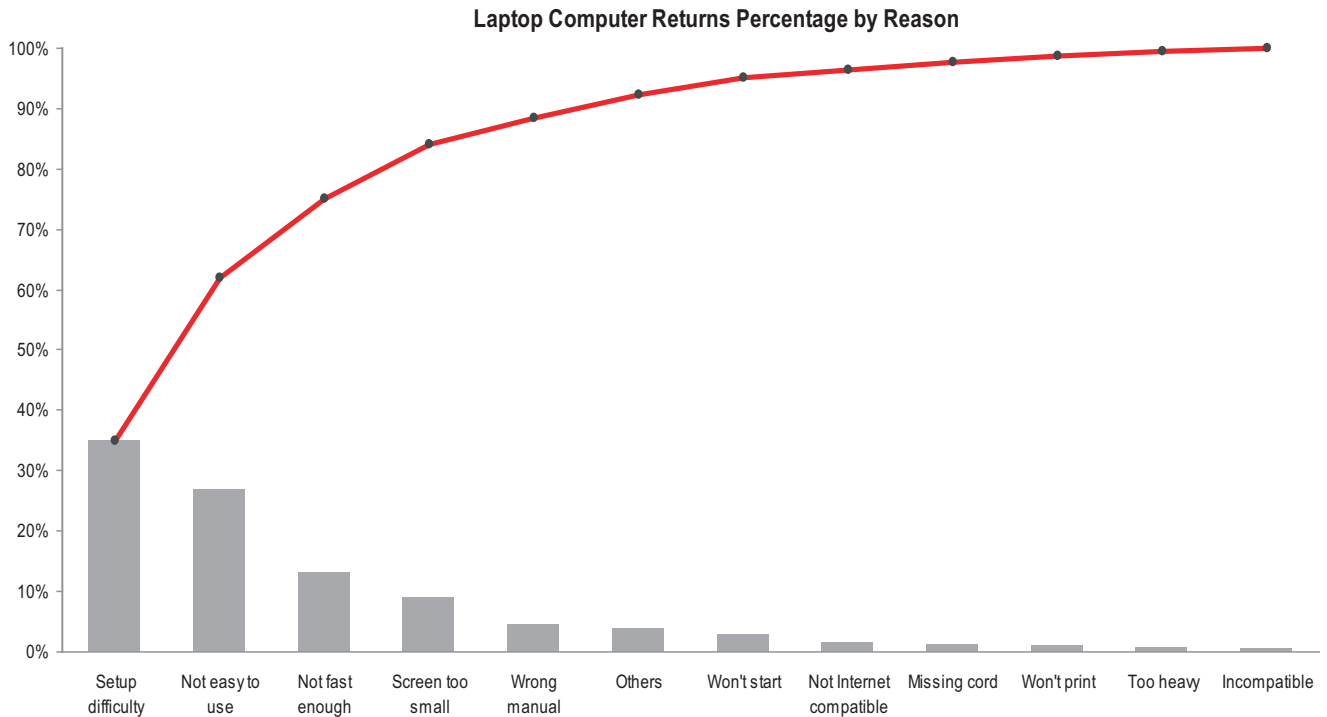
An interval scale is one that divides a continuous range of quantitative values into equal intervals. For example, if you wanted to display the incidence of obesity among elementary school children at a particular school by age, you could divide the age range of 5 to 11 years into equal intervals of one year each and then count the number of children who are obese per age group. This range of ages is an example of an interval scale.

You could use bars to represent the number of children who are obese per age group. When bars are used to display a frequency distribution (in this case the frequency of obesity by age), the graph is called a histogram. Because an interval scale represents a continuous range of quantitative values, an intimate connections exists from one interval to the next. As such, rather than using bars, it would be fine if you wished to use a line to display this frequency distribution, connecting one age group to the next, because the line would meaningfully represent a connection that exists in the data.

Another more common example of an interval scale is one that divides a continuous range of time into equal intervals, such as years, quarters, months, or days. (Yes, I know that not all months include an equal number of days, but usually when we display change through time by month, we consider the months equal in size.) Rarely does any other form of graph display the shape of change through time better than a simple line graph. Not only does a line meaningfully suggest fluid connection from one point in time to the next, but it displays changes in value clearly as up and down slopes of varying magnitudes along the line.

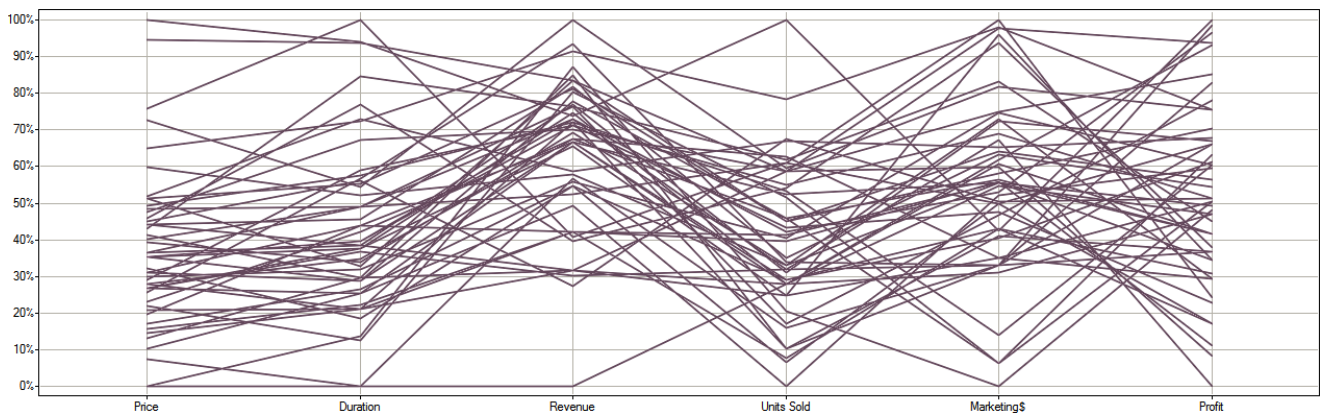
There are a couple of exceptions to the general rule that lines should only be used in graphs to connect values along an interval scale: Pareto charts and Parallel Coordinates plots. Both of these exceptions are acceptable because the connection between values that's suggested by a line is meaningful, although different from what it means when a line connects values along an interval scale.

Beginning with a Pareto chart, here's an example:



In this example, the reasons that customers have returned laptops that they purchased appear in order of rank from most to least frequent, and each bar's height represents the percentage of total returns associated with a particular reason. The red line displays the cumulative percentage of returns. Each point along the line is the sum of the percentage of laptops returned for that particular reason and all preceding reasons, which increases with each item until it reaches 100% of all returns at the right-hand edge of the graph. The scale along the X-axis is not quantitative in nature, and thus not an interval scale. The items that make up this scale are independent from one another; not intimately connected like items along an interval scale. Yet a line has been used to connect the cumulative value along the scale. In this case, the line is meaningful, because each value is the sum of that item's independent value and the values of all previous items, which connects the values intimately. In this respect, the line makes sense and therefore works.

I hesitate to show an example of a parallel coordinates plot, because they look odd, complicated, and overwhelming until you learn how they work, which I won't be able to adequately explain in this article. Brace yourself—here's an example:

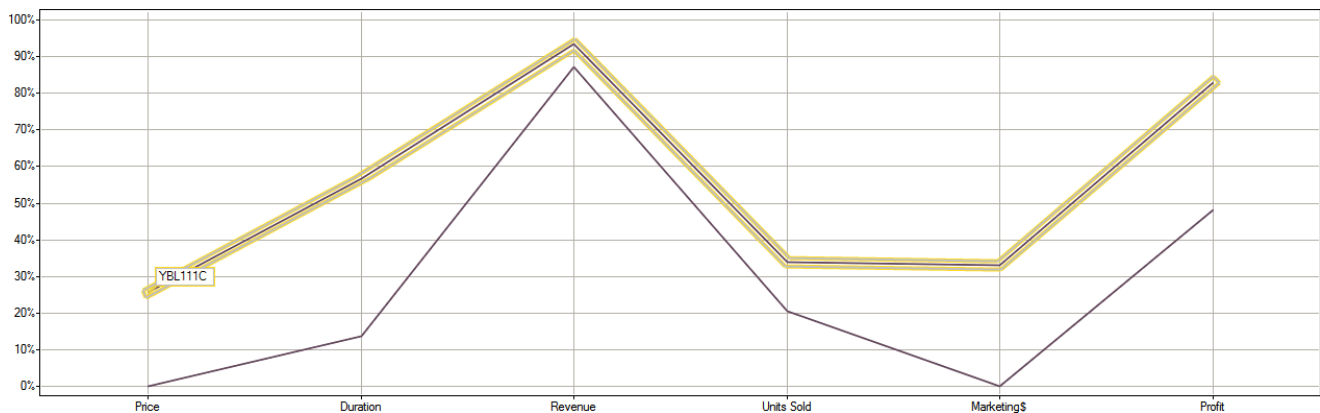


Parallel coordinate plots were designed to support a particular type of multivariate analysis, which seeks to find similarities and differences between things across several variables that make up composite multivariate profiles of those things. This particular example displays 49 different products, one per dark gray line that

extends from the left vertical axis titled “Price” all the way across several axes to the final one titled “Profit” at the right. The position where one of these lines intersects a vertical axis indicates that product’s value for that variable, for each of the following six:

- Price (the product’s list price)
- Duration (how long the product has been on the market)
- Revenue (the amount of revenue the product has generated during the current year)
- Units Sold (the quantity of the product that has sold during the current year)
- Marketing\$ (the amount of money that was spent on marketing the product during the current year)
- Profit (the profit that was earned during the current year resulting from sales of the product)

This article isn’t about parallel coordinates or multivariate analysis, so I’ll refer you to an article that I wrote in September of 2006 titled “[Multivariate Analysis Using Parallel Coordinates](#)” for more information about them and focus now only on why the lines are meaningful. Not only don’t the lines connect values along an interval scale in parallel coordinates plots, they don’t even connect values that are associated with the same variable. In these graphs, the lines serve a different, but nevertheless meaningful purpose: they form a pattern that represents the multivariate profile of a single entity, such as each of the products represented in the example above, and do so in a way that makes it possible for us to compare the multivariate profiles of many entities to one another. I haven’t seen any other graphical object that does this as well as a line that moves up and down to intersect each of the axes at the appropriate point along its scale, resulting in an overall pattern that can be easily compared to the patterns formed by other lines. To illustrate what I’m saying, in the following example I’ve selected a single product (the one that’s highlighted and labeled YBL111C) and the other product out of the total of 49 that is most like it across the six variables as a whole. Although they vary, when the plot is uncluttered by all 49 lines, it is easy to see the similar multivariate profiles of these products as represented by the pattern formed by these two lines.



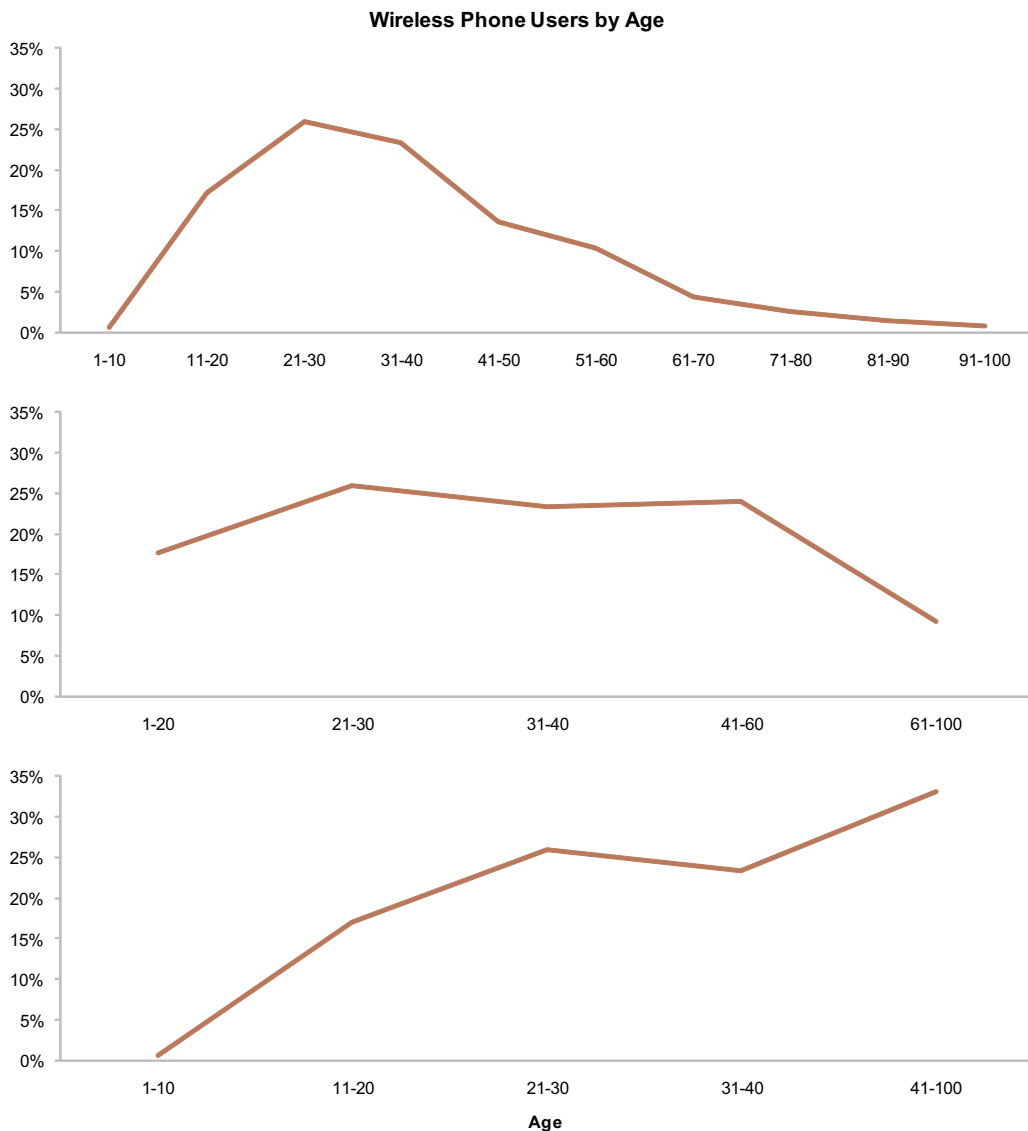
## 2. Intervals should be equal in size.

How could we trust graphical representations of time series or frequency distributions if their shapes could have been altered by inconsistently manipulating the sizes of intervals along the scale, either arbitrarily or intentionally to deceive? We can derive meaning from patterns and trends that these graphs display only if the intervals are consistent. The three graphs below display the same time-series data, but they look quite different.



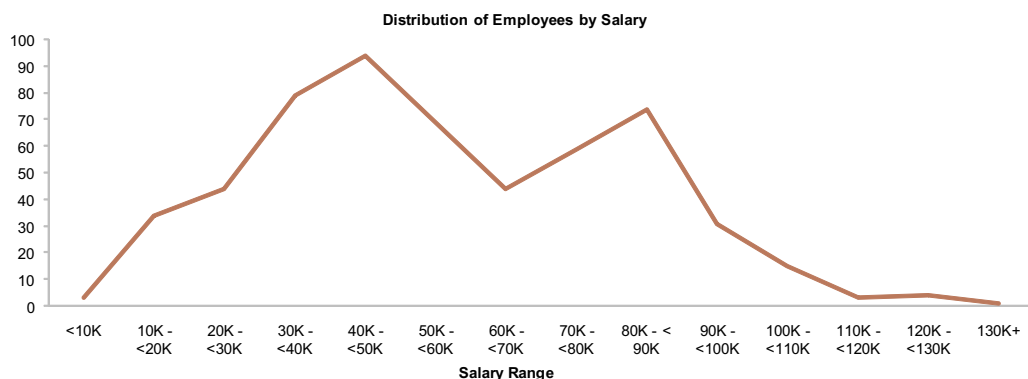
Only the first graph effectively displays the pattern of change through time.

All three graphs below are based on the same frequency distribution data, but they tell completely different stories.



Again, only the first display tells the story effectively.

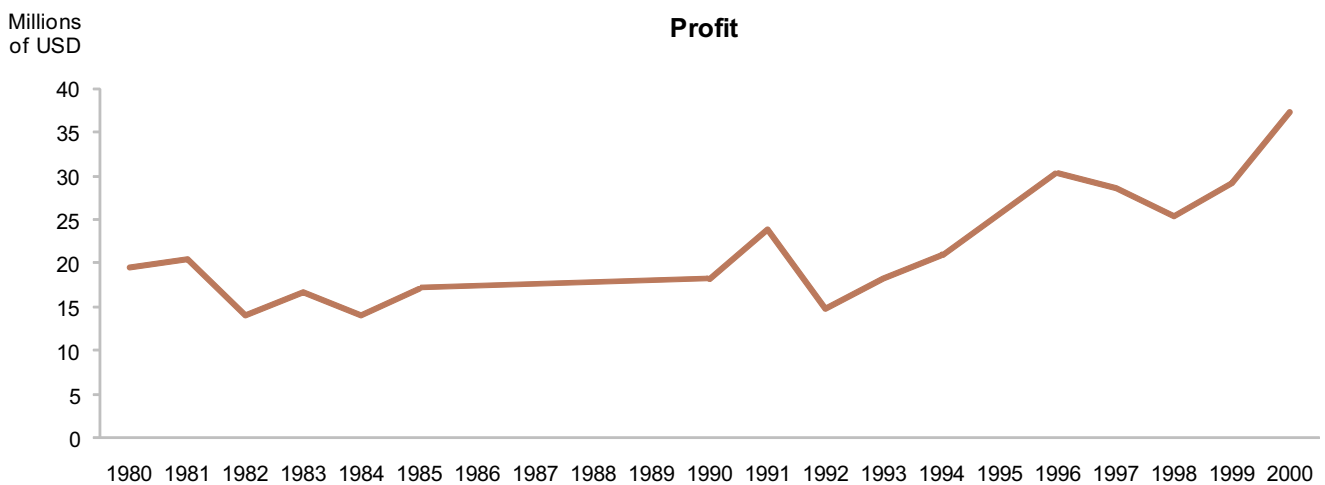
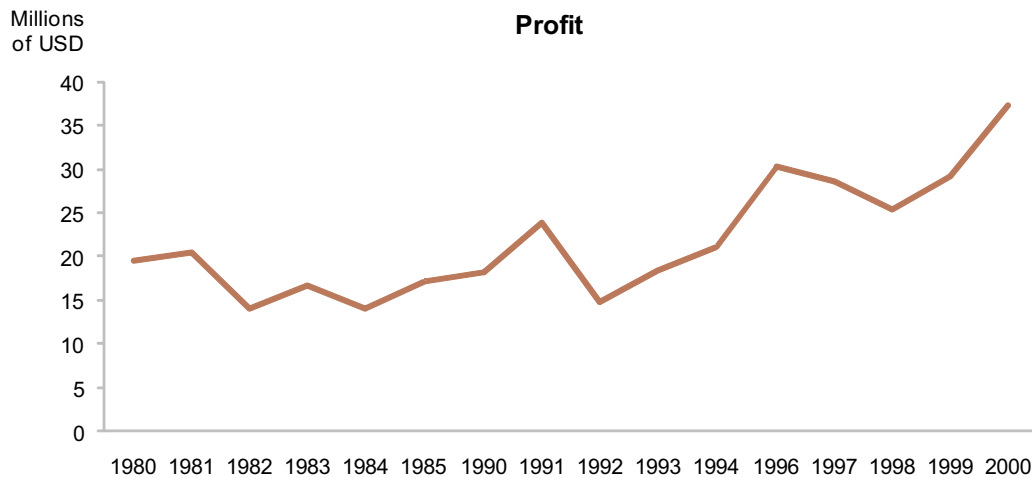
There is one situation when breaking the rule of equal intervals is justified, which I've illustrated in the following example.



In this example, the CEO of the company makes a \$3,000,000 salary, which is so much higher than anyone else's salary, to extend the \$10,000 intervals all the way from the second highest salary of \$127,000 to \$3,000,000 would produce an unreasonably wide graph. Consistent intervals could be maintained by changing their size from \$10,000 to \$500,000, but then all but one salary would be lumped into the first interval and nothing about the shape of the distribution could be discerned. In situations when extreme outliers at the high end, low end, or both ends of a frequency distribution would either force the graph to cover an impractical number of intervals or to make the intervals so large that useful information about the shape of the distribution would be lost, it is acceptable to lump the extremes at one or both ends of the scale into a single interval that is larger than the others. When you do this, however, it's important to point out what you've done, so others can take this inconsistency into account when examining the graph.

### 3. Lines should only directly connect values in adjacent intervals.

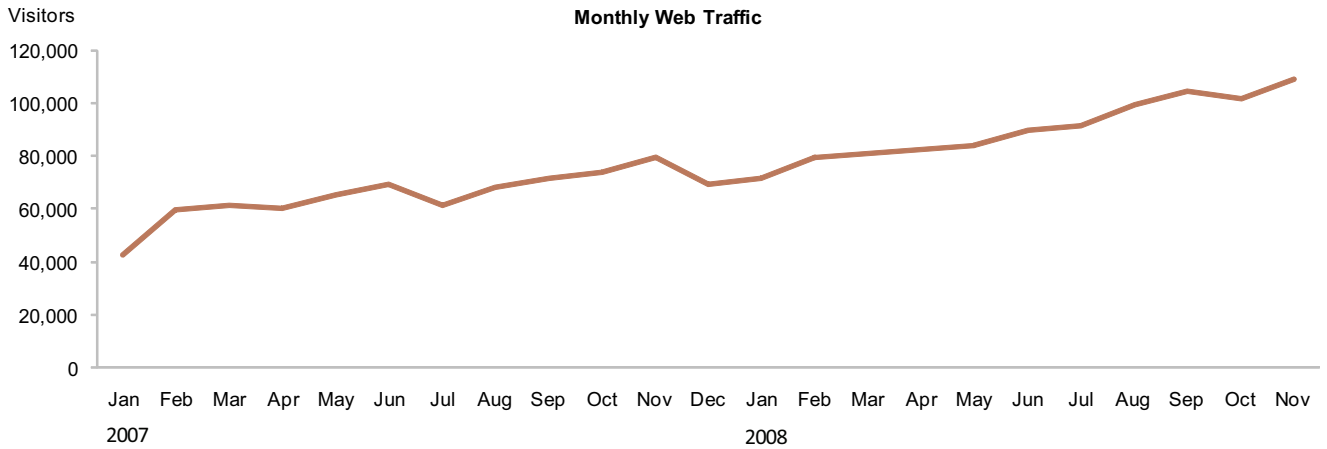
Both of the graphs below display the same series of values, but neither represents it correctly. Both use the line to directly connect intervals that aren't adjacent in time. The top graph omits several intervals (the years 1986-1989 and 1995) from the times-series scale altogether, but still directly connects the years with a line as if none were missing. The bottom graph includes the full set of years along the X-axis, but fails to display values on the line for the same years that are missing in the top graph.



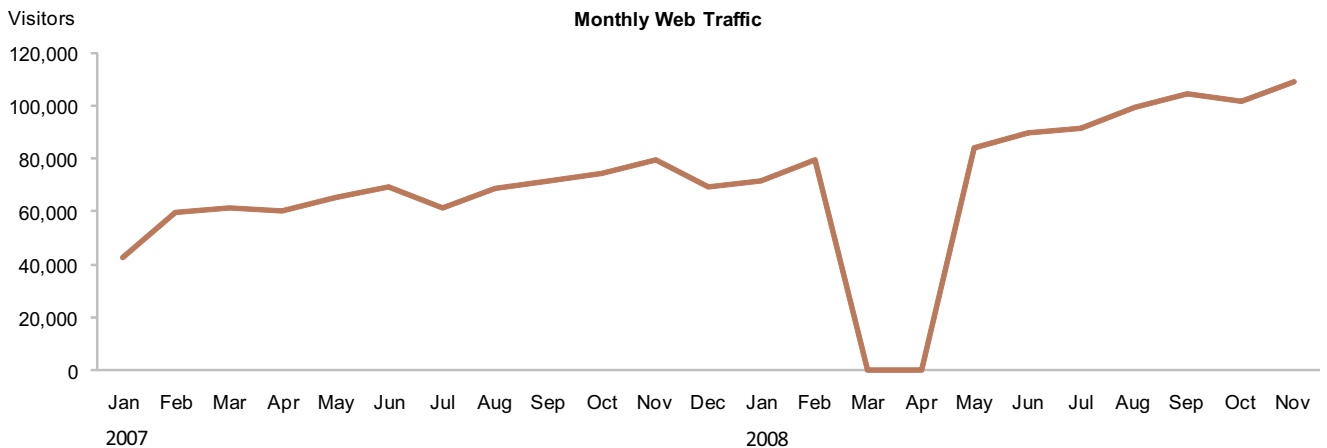
The smooth slope of the line between 1985 and 1990 in the bottom graph suggests a pattern of change that's probably quite different from what actually happened. Graphs like the bottom example are sometimes produced as a way to hide what actually happened. Whether done intentionally to deceive or innocently in error, we misrepresent the truth when we use a line to directly connect values in intervals that are not adjacent to one another.



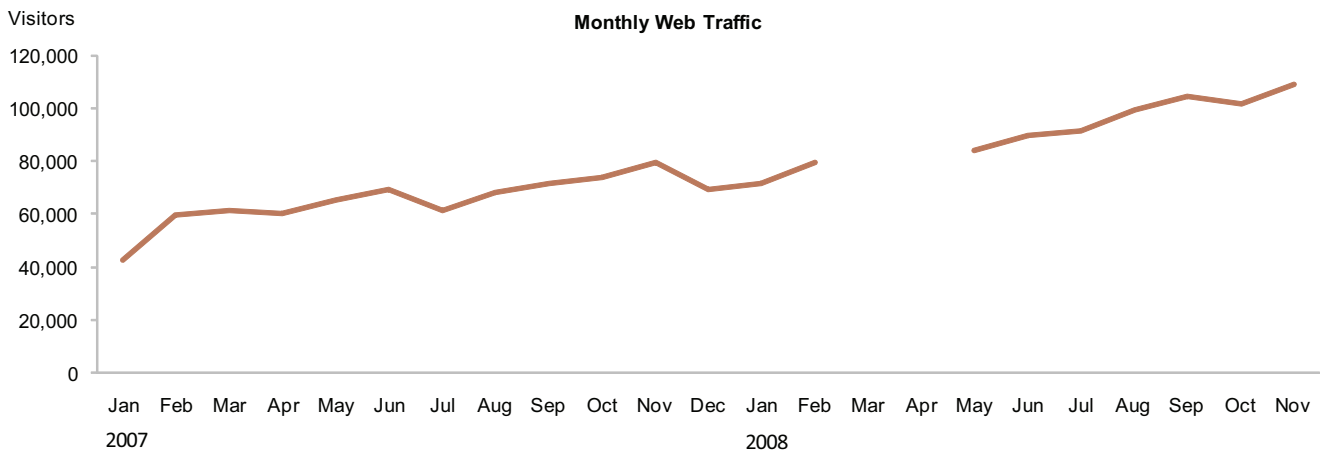
This misrepresentation sometimes occurs because values for some intervals are missing. Consider the graph below of visits to a website. If the site's web host failed to collect accurate visitor statistics for every month, which has happened to me a few times (but hopefully no longer now that I've switched web hosting companies), you might be tempted to directly connect the last accurate month prior to the problem to the first accurate month following the problem, but this not only suggests a smooth pattern between those values when the pattern is unknown, it also fails to show the important fact that some values are missing.



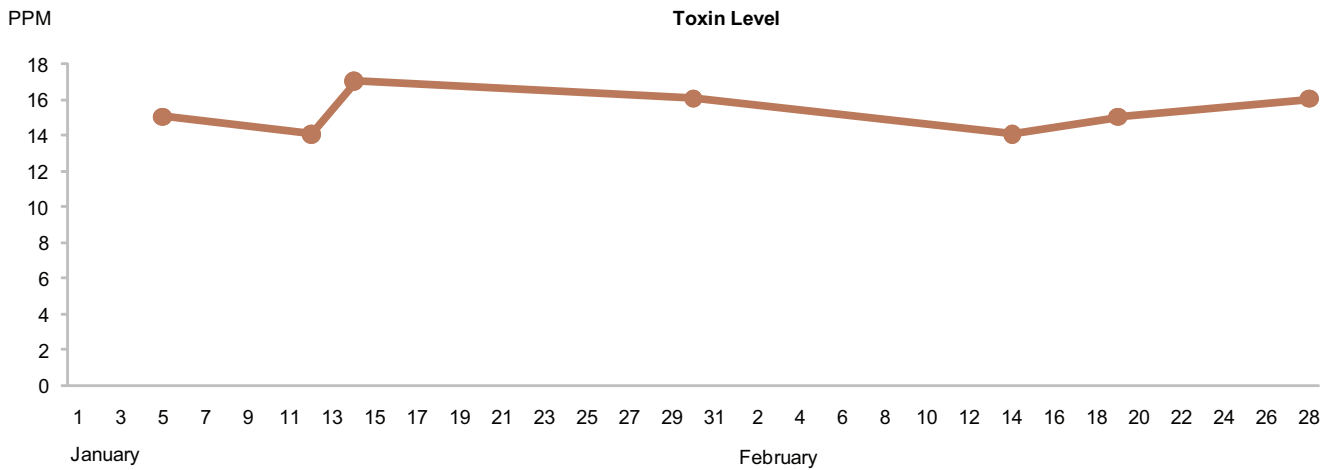
People sometimes handle situations like this by displaying the missing values as zeroes, as illustrated below, but this also misrepresents the facts. The values are not zeroes, they're missing.



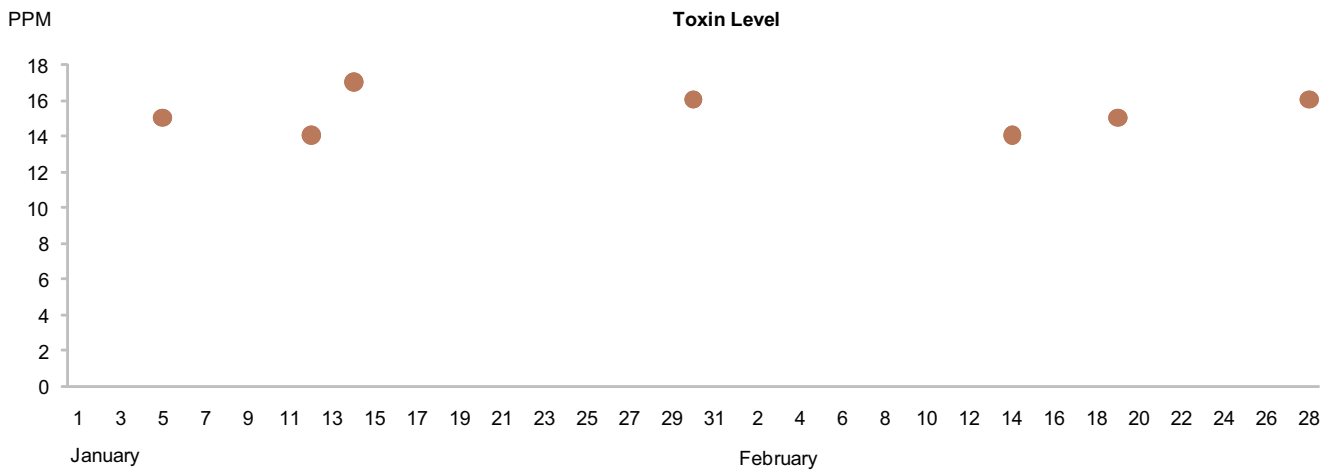
The best way to display the fact that values are missing is to omit the line altogether for those intervals, as I've done here.



Sometimes time-series data are routinely intermittent in nature. Consider the following example, which displays toxin levels in a stream, measured by a researcher at irregular intervals of time.



We have no idea what the toxin levels were on the missing days. To connect these intermittent measurements with lines suggests a pattern that might have no relation whatsoever to the changes in toxin levels that actually occurred. In such cases, I omit the line altogether, as shown here.



Even if on a few occasions along this timeline toxin levels were measured on consecutive days, I would still prefer to omit the line altogether, because its absence visually reinforces the fact that the measurement process in general was intermittent in nature.

## In Summary

These guidelines for line graphs are rooted in an understanding of human perception and practical experience. It is certainly possible that you will encounter rare occasions when one or more of these guidelines should be broken. One of the privileges and responsibilities of expertise in any field is the ability to break the rules when circumstances demand something different. Do so only when you can justify an alternative just as rationally, given the special circumstances, as I've justified the general rules for normal circumstances.

As a reminder, here are the guidelines once again:

1. Lines should only be used to connect values along an interval scale (with a couple exceptions).
2. Intervals should be equal in size.
3. Lines should only directly connect values in adjacent intervals.

---

## About the Author

Stephen Few has worked for over 25 years as an IT innovator, consultant, and teacher. Today, as Principal of the consultancy Perceptual Edge, Stephen focuses on data visualization for analyzing and communicating quantitative business information. He provides training and consulting services, writes the monthly *[Visual Business Intelligence Newsletter](#)*, speaks frequently at conferences, and teaches in the MBA program at the University of California, Berkeley. He is the author of two books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten* and *Information Dashboard Design: The Effective Visual Communication of Data*. You can learn more about Stephen's work and access an entire [library](#) of articles at [www.perceptualedge.com](http://www.perceptualedge.com). Between articles, you can read Stephen's thoughts on the industry in his [blog](#).