

Assignment 3: Summary of *Effectively Communicating Numbers: Selecting the Best Means and Manner of Display*, by Stephen Few (2005).

David Boose

September 21, 2015

Stephen Few, a principal at [Perceptual Edge](#), a business consulting company, prepared a [white paper](#) for the software company [ProClarity](#) on the effective display of data in graphical format. ProClarity software is designed to make graphing data easy, but as with any such software, it will do whatever the user asks it to do. Thus, it is just as easy to make a useless or misleading graph as it is to make a beautiful and effective one. Producing the latter instead of the former requires some knowledge, thought, and careful decision-making on the part of the graph designer, and it is those things that Few's white paper is designed to support. The paper lays out six steps in the process of creating an effective visual communication of data, and each step represents a set of decisions that need to be made. Following the steps and thinking carefully about the answers to the questions can make it much more likely that your graph will say what you want it to say, and your audience will understand its message clearly.

1 What do you want to say, and which data say it?

Visual processing is extremely fast and effective, so a viewer can create meaning from a graph in a very short time. But if the graph is not well designed, the meaning that is created by the viewer may not be the one intended by the designer. So it is important that before you begin to create the graph (or table), you know what message you want the graph to convey. Are there specific patterns you want to highlight? Specific ranges of values that are more important than others, or specific comparisons you want the viewer to make? Once those decisions are made, you need to find the specific variables in your data that will make those comparisons possible.

2 What is the most effective format, a table or a graph?

Both tables and graphs can depict relationships among variables. But they have different strengths and weaknesses. A graph is very effective at showing *patterns* of relationships among variables. In some formats, large numbers of data points can be plotted to give an overall picture of the relationship, while the specific values of those data points are de-emphasized. If the precise values of individual data points are the focus of the presentation, then a table will be more effective. But a large table with many values can cause the reader's eyes to glaze over, and individual comparisons will be harder. So for a relatively small number of specific values, a table should be used. For patterns emerging from larger numbers of data points, a graph is the better choice.

3 How should the data be displayed?

This is really the key step in graph design. Once the message has been determined and the format chosen, the designer must choose the way in which the data will be displayed. Few refers to this as *encoding* the data, which means, what symbols on the graph will be used to represent the real-world values of the data (which graphing experts call the *referents*.) The most common symbols used in graphs are points, lines, bars, and

boxes, and each has advantages and disadvantages. The type of coding to be used is determined by the types of variables to be shown, and the type of relationship they display.

3.1 Types of variables

There are two basic categories of variables that are likely to be displayed graphically: Quantitative and qualitative. **Quantitative variables** (also known as continuous variables) are represented by numerals and can (at least theoretically) take an infinite number of values. **Qualitative variables** (also known as categorical variables) take values that can either be numerical or textual, but always depict categories with finite boundaries. There is no meaning to values between categories. Within categorical variables, Few identifies three different subtypes:

- Nominal variables. These variables identify categories that do not have a particular relationship. That is, there is no order or relative value to the categories. Variable values are simply the names of the different categories, for example, countries.
- Ordinal variables. These are categorical variables that have an intrinsic order, but do not have a quantitative relationship. Differences among adjacent categories are not necessarily equal, or even defined quantitatively. Examples include rankings such as “good” “better,” and “best.”
- Interval variables. Interval variables are actually quantitative variables that have been divided into discrete groups. Sometimes, this is a result of the nature of the variable itself. For example, counts of individual units (e.g., cases, customers, or abdominal bristles on a fruit fly) will always be interval variables. Interval variables also often show up in histograms, where the number of data points falling within a particular range of values is plotted. The ranges of values are “categories” created by placing arbitrary boundaries between groups of values.

The types of variables to be displayed will play a large role in determining the format of a graph. Comparisons between two continuous variables are almost always displayed with a scatter plot. Relationships between a continuous variable and one or more categorical variables can be displayed in a variety of formats, depending on the particular relationship being shown.

3.2 Types of representations

Among the features that the human visual system processes most quickly are 2-d location in a plane and the length and slope of lines. Points, lines, bars, and boxes take advantage of this, and can be used to depict patterns and relationships that the eye can discern quickly and accurately.

- Points. Points make use of our ability to preattentively notice differences in location. Because they can be used to indicate distances from two (or even three) mutually perpendicular axes simultaneously, they are very effective in depicting relationships among multiple quantitative variables simultaneously. They also contain no extraneous non-data ink, and can differ in shape or color to denote additional qualitative variables.
- Lines. Lines make trends in data obvious. Because a line can have both position and slope, we can quickly make comparisons in both magnitude and rate of change of the values depicted by a line. Because connecting sequential data points with a line implies that the space between those points has meaning, lines are only appropriate when the predictor variable is continuous or interval in nature. The space between categorical values does not have meaning, so suggesting a trend continuing through that space would be misleading. Interval-scale axes can use either lines or non-continuous depictions such as bars or points. Non-continuous depictions will emphasize the specific values being shown, while lines will emphasize the rate and direction of change between those values.

- **Bars.** Preattentive visual processing can rapidly detect differences in the lengths of lines, so bars can be very effective in conveying this information quickly. They are excellent for showing comparisons of a quantitative response variable across a number of categorical values, and are most often used for this purpose. As described above, bars will call attention to specific values of the data, but are less effective at showing trends and overall patterns. One exception to this is the histogram, in which bars are typically used to show the number of data points falling within an interval range of quantitative values. In this case, the exact height of the bar is not as important as the overall shape of the distribution. In fact, in a histogram, the edges of the bars are typically contiguous, which reduces the ability to make visual comparisons between individual pairs of bars, and instead emphasizes the 2-d positions of the tops of the bars.
- **Boxes.** Boxes are a very efficient use of data-ink, because they combine both length and position to show a range of values. Boxes are often combined with perpendicular lines across them to show specific values within that range, such as the median or quartiles. In this case, these crossing lines are really acting like points, because they show a specific value, and Tufte has argued that using points would be more efficient from a design standpoint.

3.3 Types of relationships

In his white paper, Few identifies seven different types of relationship that he says are the most common ones found in business data. Each is displayed most effectively by a particular type of graph.

- **Time-series.** In a time-series relationship, a quantitative response variable is plotted against time as the predictor variable. Time is a continuous variable that is often divided into intervals, such as days, months, quarters, or years. If the number of intervals is relatively large, and the goal is to show trends in the response variable, a line graph is the best choice. In contrast, if the number of intervals is relatively small (such as four quarters in a year, for example), then representing the quantitative variable with bars would facilitate more specific comparisons.
- **Comparisons among nominal categories.** This is probably the simplest type of graph. The predictor variable is categorical, and the response variable is a quantitative value associated with each category of the predictor. Vertical or horizontal bar graphs are most appropriate here. Typically, the categorical variable does not have an inherent order to it (i.e., it is not an ordinal variable), in which case the effectiveness of the graph may be increased by ranking the categories in order of the values of the response variable. This then creates the next type of relationship:
- **Ranking relationships.** Ordering the categorical predictor variable according to the values of the quantitative response variable shows a ranking relationship. This type of graph can be very effective, because the human visual system can detect differences in the length of lines very quickly. When bars are ranked in ascending or descending order, it creates a pattern that is immediately discernable.
- **Part-to-Whole relationships.** In these relationships, the categorical variable represents parts that make up a whole unit—regions of a country for example, or divisions of a company. The quantitative variable represents the fractional contribution of each category to the whole, so the quantitative values sum to unity. The default representation in this situation is often a pie chart, because the circular “pie” easily represents the whole, and the wedge-shaped “pieces” can be seen as fractional components of that whole. The problem with pie charts is that the brain has a much harder time making comparisons among different areas than it does among varying lengths or positions. And if the number of categories is relatively large (greater than five, say), then distinctions become nearly impossible. A better solution with only a single categorical variable would be a ranked bar graph, which would convey the important information much more quickly. If two categorical variables are to be shown (such as percentage of total sales by region by quarter), a stacked bar graph is effective. One variable (e.g., quarter) is represented by different locations along the axis, and sales values in each region for that quarter would be represented by different length bars stacked on top of one another, creating a bar of total length 1.

- Deviation relationships. Sometimes what is of interest is the difference between observations and a particular value—a long-term mean, zero, or a value predicted by theory, for example. In this case, vertical or horizontal bars extend up and down or left and right from a line indicating the comparison value. For greater visual impact, bars can be colored differently on either side of the comparison value, such as black for positive values and red for negative values of net revenues.
- Distributions. Technically, this is not a relationship among variables. Rather, it shows relationships among values of a single quantitative variable—their distribution across the possible range of values. The objective here is to visually assess the data for such things as its range, central tendency, skew, and the presence or absence of outliers. The most common type of graph used here is a bar graph, with the horizontal axis being an interval variable of arbitrary ranges of values. The height of each bar depicts the number of values in the data set that fall within that particular range.
- Correlations: When two quantitative variables are being compared, one type of relationship that may exist is a correlation, in which the two variables vary in a predictable way relative to one another. If high values of one variable are generally associated with high values of the other, and *vice versa*, this represents a positive correlation. The opposite pattern is a negative correlation. The simplest representation of this relationship is a scatter plot, where corresponding pairs of values of the two variables are plotted in an x-y plane.

4 Where should each variable be displayed?

Once the representations of the variables have been chosen, the next question is where on the graph they should be arranged. Some of this may be driven by convention in a particular field. For example, it is most common to have predictor variables (especially categorical ones) depicted along the horizontal axis, and response variables (which are almost always quantitative) along the vertical axis. But there may be good reasons to switch this arrangement. For example, if the category labels are long, it may be better to plot values as horizontal bars, so that the category labels can be easily read from left to right. If a graph is to display three variables, the two to be compared most directly should be placed in closest proximity to one another, or distinguished by differences that our eyes easily discern (e.g., contrasting shapes or colors). For example, sets of adjacent bars separated from other such sets by a larger space will make for easy comparisons within the sets, while still allowing for the detection of differences in patterns across the sets. If more than three variables need to be displayed, a set of several smaller graphs, each differing in the value of one categorical variable, will generally be more effective.

5 How should the remaining elements be displayed?

The remaining elements of the graph are what Tufte refers to as non-data ink, and according to him, they should be reduced to the minimum necessary for interpreting the graph. These include the axes, of course, as well as any tick-marks and labels along those axes. For quantitative variables, the axis should always include the value of zero, to ensure that relative differences in spatial location correspond correctly to relative differences in magnitude. Grid lines may be useful in some cases, where comparisons of specific values within the graph or across several graphs is important, but they should be made as light as possible so as not to create distraction. Other elements that might be needed (or not) are axis titles, an overall descriptive title or caption, and a legend explaining the coding of the variables.

6 Should particular values be emphasized, and if so, how?

Sometimes, the designer wants to emphasize certain values on a graph—one company among the others being compared, for example, or a particular instance of an experimental result. If that is the case, then the

way of emphasizing that value should make use of differences that are quickly and easily perceived visually. Highlighting a bar or a data point with a different color or intensity is very effective, for example, while making a point slightly larger may be more difficult to detect. Line width or type (e.g., dotted, dashed, continuous) is also easy to compare, and may be more readily detected than some color differences, depending on the shades chosen. Borders around bars and varying shapes of points are also effective means of conveying emphasis.

7 What else can I get rid of?

While Few does not make this one of his six steps, an important final consideration in graph design is removing any distractions that may make it more difficult for the viewer to perceive the message being conveyed. This is non-data ink that adds nothing to the interpretation of the graph, which [John Boyd](#) called “chartjunk.” These distractions include things like background images or patterns that can make it more difficult to see the differences in the data. Even the components used to display the data can be distracting, and should be examined. Poor use of color is a particular problem in this regard; a lack of contrast or the use of too many shades can make it difficult to discern real differences. Conversely, strong contrasts in color or shade can draw attention to certain parts of the graph that are not necessarily more important than others. And while tick marks along axes are helpful in conveying the ranges of values, too many of them can be distracting.

Because the human visual system is such a rapid and powerful creator of meaning, a well-designed graph can be extremely effective at telling a detailed story in a small space. But a poorly designed graph will hinder that process at least, and in the worst cases, will actually convey meaning that is not intended. Stephen Few’s six steps to designing a graph will go a long way toward producing an effective piece of visual communication and reducing the likelihood of causing confusion for the reader.