Just Because You Can Doesn't Mean You Should
–
The Elements of Graphing Data Well

Steve Figard, Ph.D., Abbott Laboratories, Abbott Park, IL

ABSTRACT

Most software packages that revolve around and/or include the creation of graphs and the representation of data in graphical format provide today's users with an unprecedented (and sometimes confusing) plethora of options. Data analysis and presentation are highly facilitated by the proper use of graphics. Indeed, the Graph Builder platform of JMP® 8 is the greatest thing since sliced bread for the rapid visual presentation of quantitative information. With the great power of these and other graphics packages come, not just great responsibility (thank you, Spidey), but also an increased probability of lack of clarity, error, and actual abuse due to simple ignorance of principles of good graph construction. The software can only provide so much protection, for as Richard Cook (science fiction author, *The Wizardry Compiled*) has so well observed,

> "Programming today is a race between software engineers striving to build bigger and better idiot-proof programs, and the Universe trying to produce bigger and better idiots. So far, the Universe is winning."

The objective of this presentation is to provide the audience with the principles gleaned from such giants as Cleveland and Tufte within the context of JMP in an effort to combat graphic entropy.

INTRODUCTION

The graphing of data generally serves one or both of two primary purposes: *to present* the data visually and/or *to analyze* the data visually. The human brain can often recognize patterns faster with the eye than it can by numbers or by sophisticated programming techniques. Thus, the "visual display of quantitative data" should have two primary principles of operation:

1. clarity in revealing "the story" the data has embedded within, and
2. ease of visually analyzing the plotted data, which often (but not always[1]) may be translated into the speed with which the viewer can see "the story" the data contains.

The ability to generate graphics of quantitative data has exploded with the personal computer and software designed for just this purpose. Options abound, and all too often the inexperienced user can't resist simultaneously utilizing all the bells and whistles provided, resulting in what Dr. Tufte has called "chartjunk" which does little to inform no matter how impressive it may look. As Dr. Cleveland says in the preface of his *Elements of Graphing Data* (*emphasis* added):

> 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.

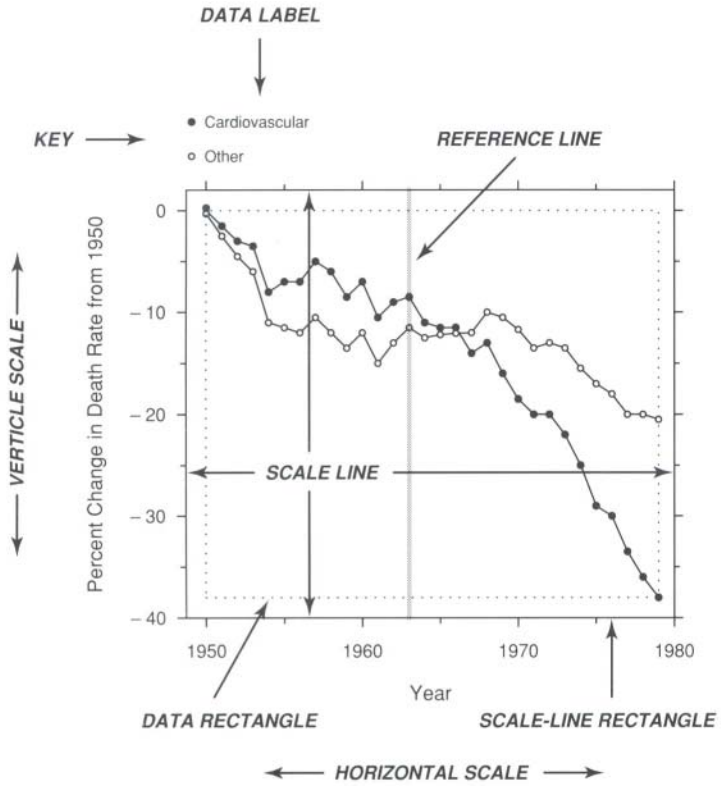
Thus, the Grand Unification Philosophy (GUP) underlying good graphics might be summarized as: *minimize the mental gymnastics that the viewer must go through to understand the graph.* It is the purpose of the remainder of this paper to flesh out this philosophy with some practical principles and "rules of thumb" for creating readily understood scientific graphics that allow for quick visual analysis of the information contained therein. Although the context for this dissertation is the statistical graphics of JMP, most of my examples and figures are from elsewhere as the principles are "generic."

TERMINOLOGY

[1] "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 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."
- Cleveland, pages 115, 117.

The place to begin our odyssey will be to define some of the terminology, and here, visually labeling the elements of graphics is the easiest way to do so. Figure 1 defines terminology using a graph of two sets of data – death rates due to cardiovascular disease and death rates due to all other diseases – superimposed. Note particularly the difference between the data rectangle and the scale-line rectangle.

Figure 1:



The scale lines are usually also designated the x axis for the horizontal, and the y axis for the vertical. Figure 2 on the next page defines some additional terms with the same data plotted in one figure that contains two different panels.[2]

[2] Figures 1 and 2 from Cleveland, pages 22, 24.

Figure 2:

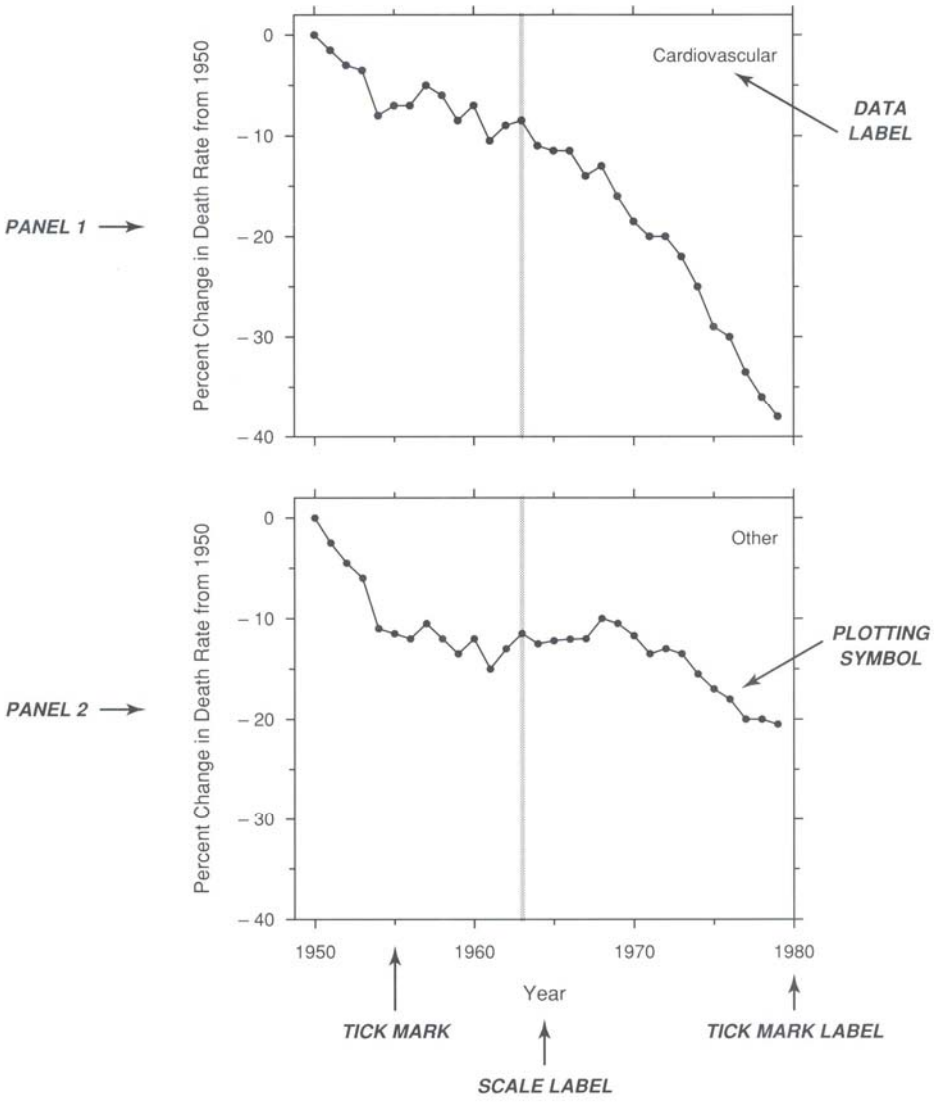


Figure 19. AGE-ADJUSTED DEATH RATE. The data are the percent changes from 1950 in death rate in the United States due to cardivascular disease and other diseases.

THE TEN COMMANDMENTS OF GOOD GRAPHICS

The following compilation of graphic principles have their foundation in a combination of practical experience and the science of perception, and have been gleaned from Cleveland and Tufte as primary sources. While I have labeled them "The Ten Commandments," they have not been issued from Mount Sinai and are not written on stone tablets, so if you have *good* reason for ignoring any of these, you will have no need to seek absolution. They are not presented in any particular order of priority as they tend to operate synergistically towards the final comprehension of the graphic presentation.

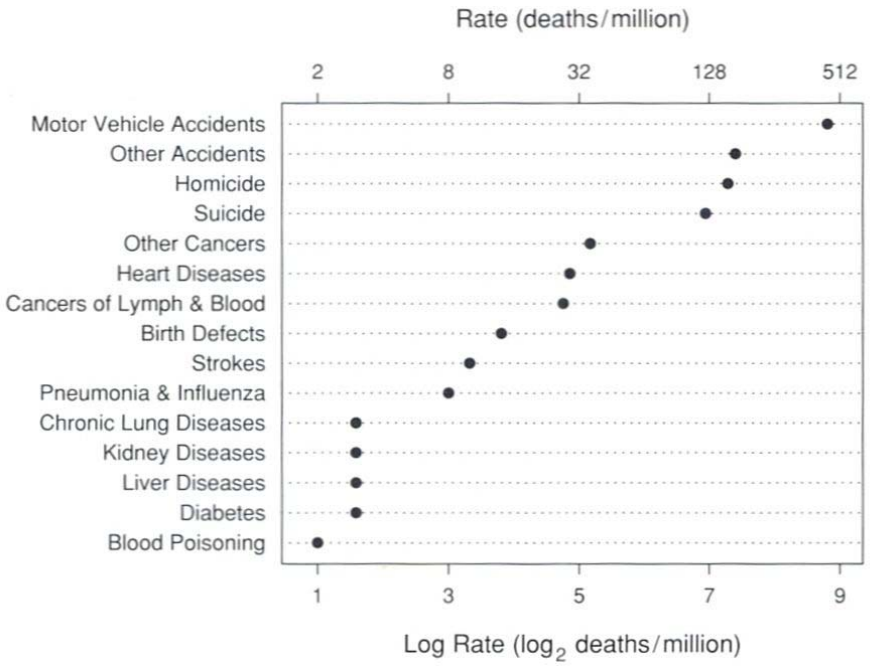
#1: *Thou shalt pay very close attention to thy axes, for therein lieth great opportunity to succeed or to fail.*

There are at least five aspects of axes to which that close attention should be paid, both in graphic construction and in reading the graphs of others:

1. the units of measure employed,
2. the range of those units of measure,
3. the number of tick marks shown,
4. the presence or absence of breaks in the axis, and
5. the size or length of the axis on the page.

Units of measure: Presumably your intended audience will be familiar with whatever standard unit of measure (UOM) is being employed, but often there is more than one way to display those units. For example, when the range of the data requires a log scale to see all the data clearly, one encounters the problem that most people don't relate to log scales very well. Figure 3 shows one such graph with a simple solution, a second set of scale labels on the top denoting the actual values rather than the logs:[3]

Figure 3:



Related to UOM is the choice of just what to graph. When the point of a graph is the analysis of the difference in two measurements over time, it is often more informative to calculate and plot that difference directly rather than plotting the two series of measurements and asking your viewer to mentally do the subtraction as they look at your graph. Cleveland provides us with an interesting example, shown in Figure 4 on the next page.[4] The top graph is the actual data, but the real story is in the difference between the two lines. The lower graph plots that difference and tells the story in the data with greater effectiveness.

The reason behind this difficulty in perception is that people are good at perceiving perpendicular distances between two curves, but not the difference in height, which is what actually measures the difference on an x-y plot. Figure 5 on the next page gives a simpler example that can almost be labeled an "optical illusion."[5] The difference between the two lines is, indeed, a constant one unit, but the lines appear to converge as the two curves ascend in y values. Our eyes are drawn to the perpendicular distance between the two lines rather than the vertical distance.

[3] Cleveland, page 62.
[4] Cleveland, page 21.
[5] Zumel, page 12.

Figure 4:

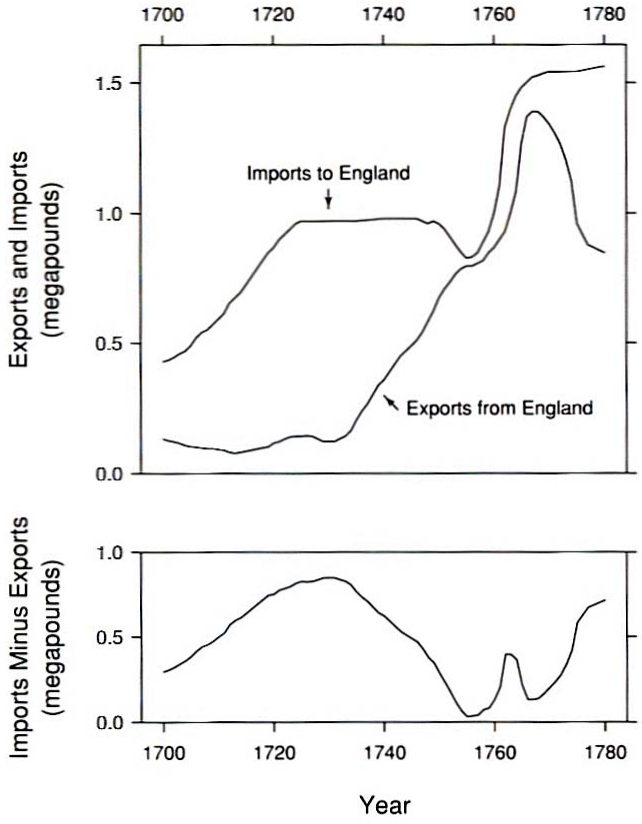
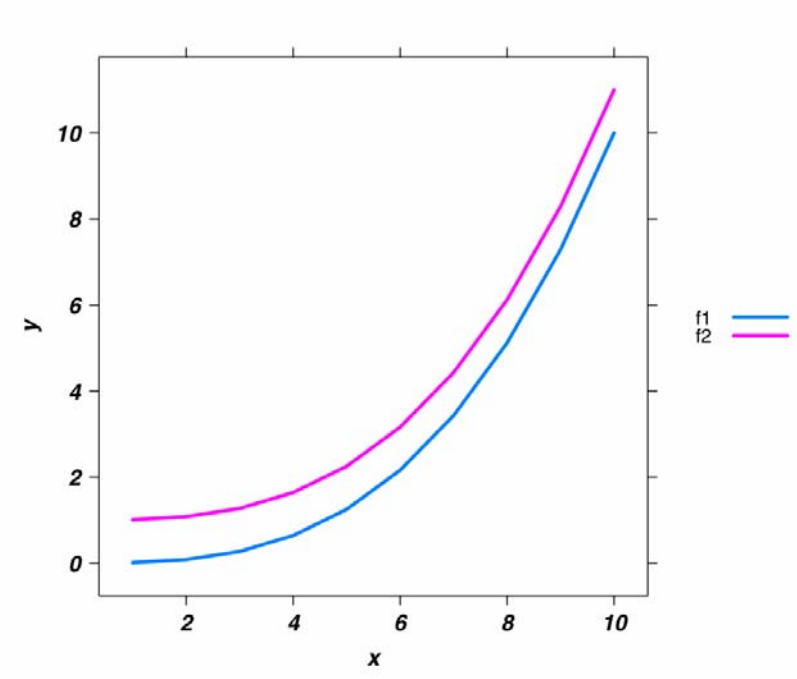
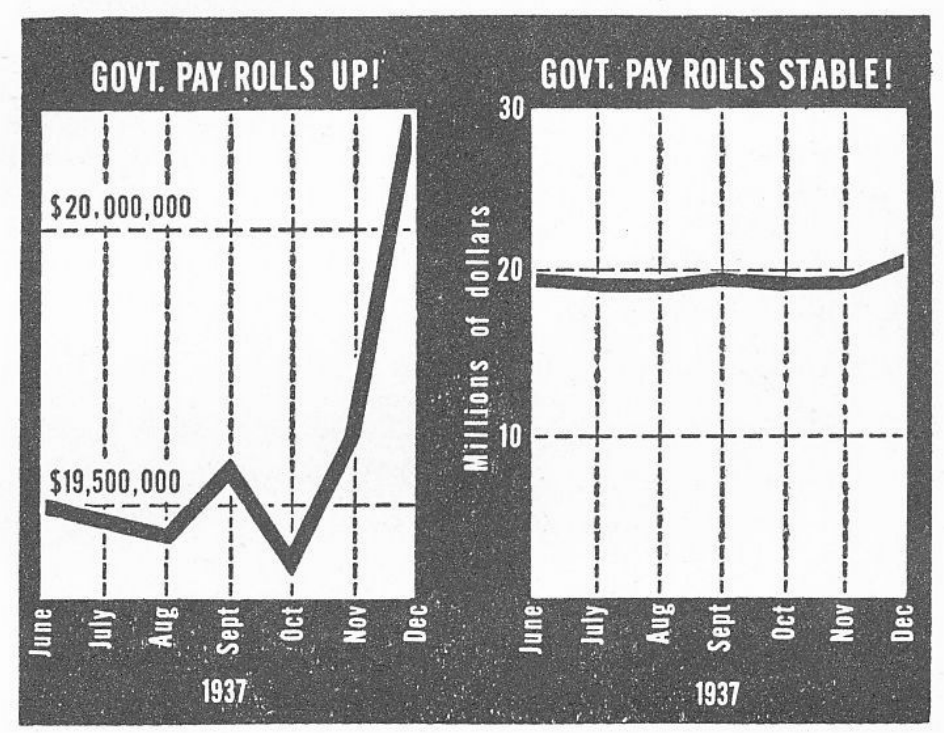


Figure 5:



Range of units: Several principles apply here. First, it is usually best to choose your range so that the data rectangle fills up as much of the scale-line rectangle as possible. As mentioned above, sometimes it will be helpful to use a pair of scale lines for a variable to show two different UOM. Do not insist that the zero always be included on a scale showing magnitude, but

note, herein lays great opportunity to lie with statistics. A classic example showing the same data with two different y axis scales is the one Darrell Huff serves up[6] in *How to Lie with Statistics*, shown in Figure 6:



On the left, the graph obeys our corollary to adjust the y axis so that the data rectangle fills up as much of the scale-line rectangle as possible. The graph on the right does the opposite, but perhaps gives a better picture of the actual trend in government payrolls. In the final analysis, a table of values may have served the purpose of clear presentation of the data rather than either of these graphs.

An important consideration in the selection of the range of units is the purpose for which the graph is intended. Alas and forsooth, agendas political and otherwise often distort the selection process of this axis parameter, violating the rules of objectivity in data analysis, and confirming the observation of that master statistician who said,

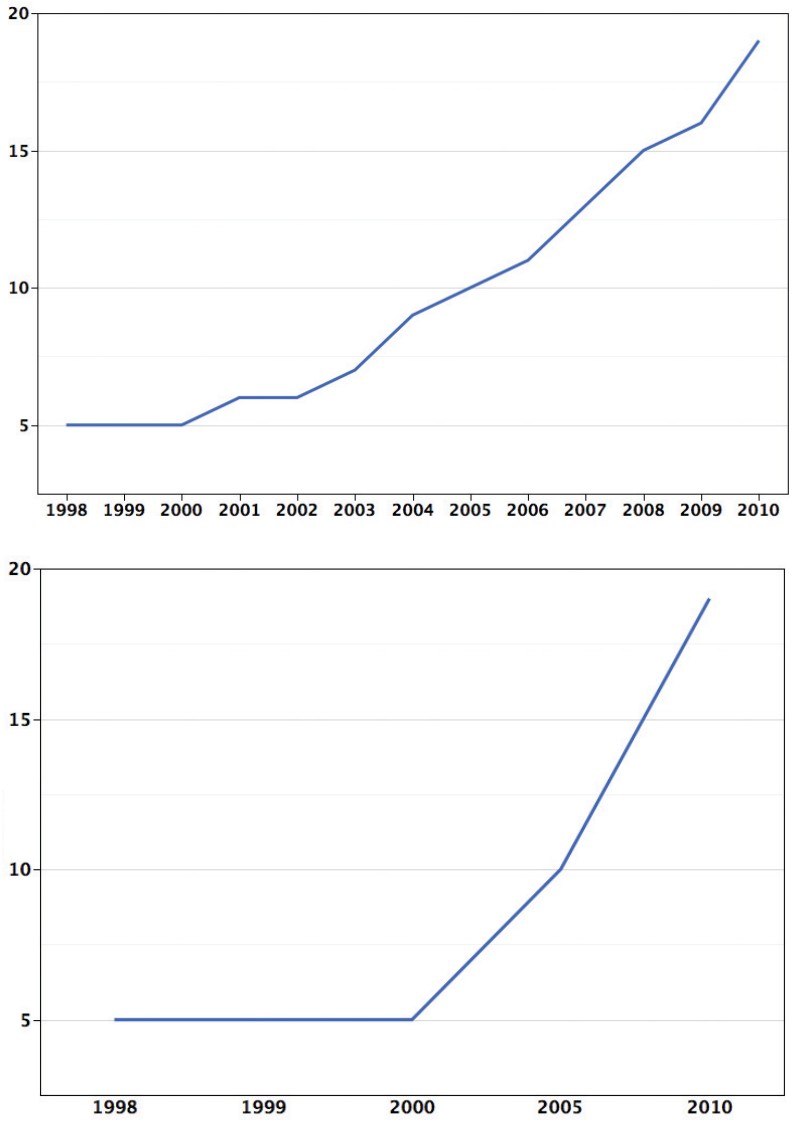
> "Get your facts first, and then you can distort them as much as you please. (Facts are stubborn, but statistics are more pliable)."[7]

Number of tick marks: Too many tick marks add clutter without added value, and decrease the visual prominence of the data. Too few make the reader "guesstimate" too much in mentally dividing the space between tick marks. From 3 to 10 tick marks are generally sufficient. Tick marks are used to give a broad sense of the measurement scale and to enable sufficiently accurate table look-up.

One way to abuse tick marks is to skew those of a time axis to change the interval represented by the tick mark. Chuck Pirrello supplies us with an example in his white paper, *Effective Visualization Techniques for Data Discovery and Analysis*. As shown in Figure 7 on the next page, "The top graph shows a more gradual increase over time. The second graph misleads the viewer into thinking the rise over time is much more dramatic by changing midstream from yearly increments to five-year increments."[8]

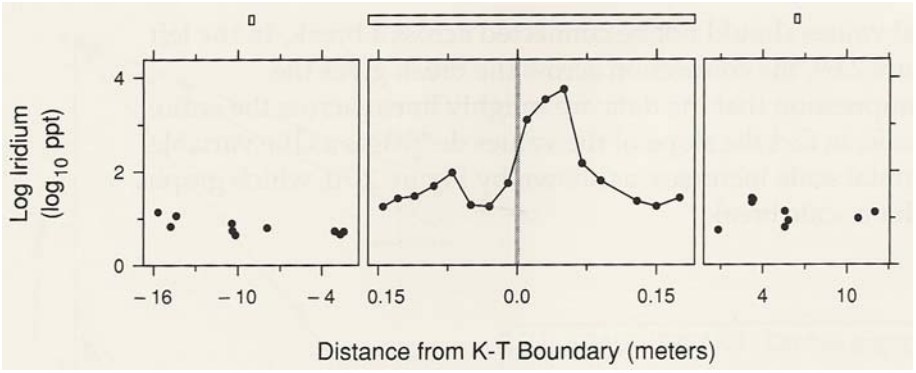
[6] Huff, page 65.
[7] Mark Twain (1835-1910)
[8] Pirrello, page 6.

Figure 7:



Breaks in the axis: The basic rule of thumb here is to use them only when necessary. This is one of the more exotic features of graphical software; it is not offered in many packages. This option is usually considered when the data range spans several orders of magnitude. In such cases, first try a log scale, a common remedy for skewness and the transformation of non-normal data into a normal dataset. If it is determined that a scale break is necessary, do a full scale break if possible (Figure 8).[9] Above all, do not connect numerical values across a break, obscuring the fact that there is a break in the data representation (Figure 9).[10]

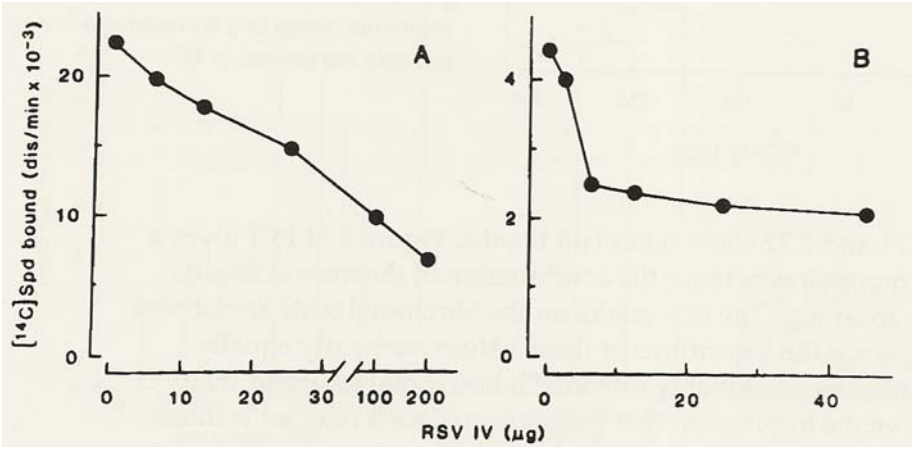
Figure 8:



[9] Cleveland, page 105.
[10] Ibid.

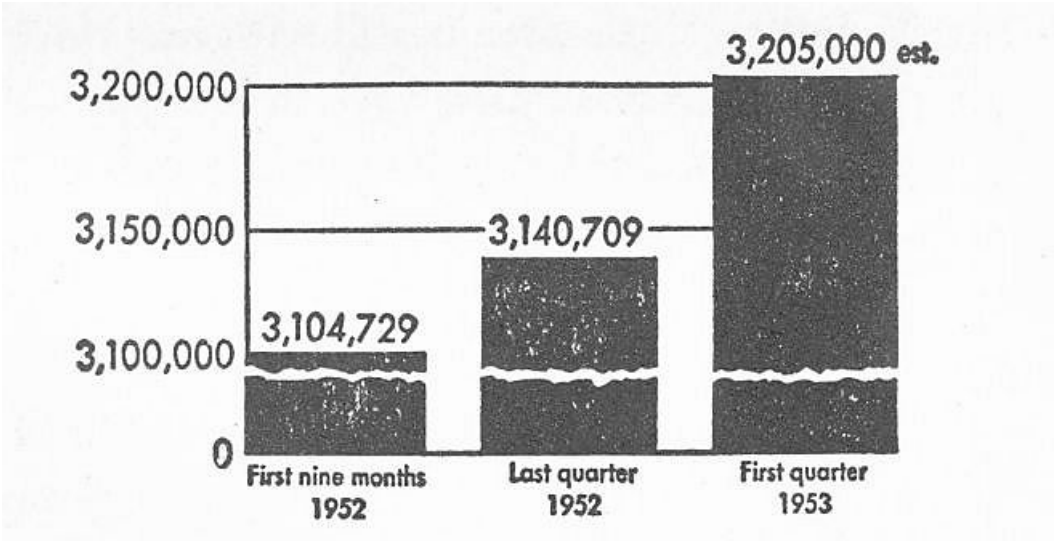
The full breaks in Figure 8 signal changes in number of units per cm on the horizontal scale and do so forcefully. The graph also includes the rectangles at the top to portray the same number of horizontal scale units on each panel to clarify the perspective differences in the scales.

Figure 9:



The partial scale break on the x axis of the left panel of Figure 9 is masked by the more prominent connection of data points, conveying the misleading impression of roughly linear data.

Yet another example of distorting the data with a data break[11] is given in Figure 10:



Here the break is fairly obvious, but its use results in artificially enlarging the differences between the three values graphed. Again, a data table would have served the function of presenting the data without distortion better than this bar chart.

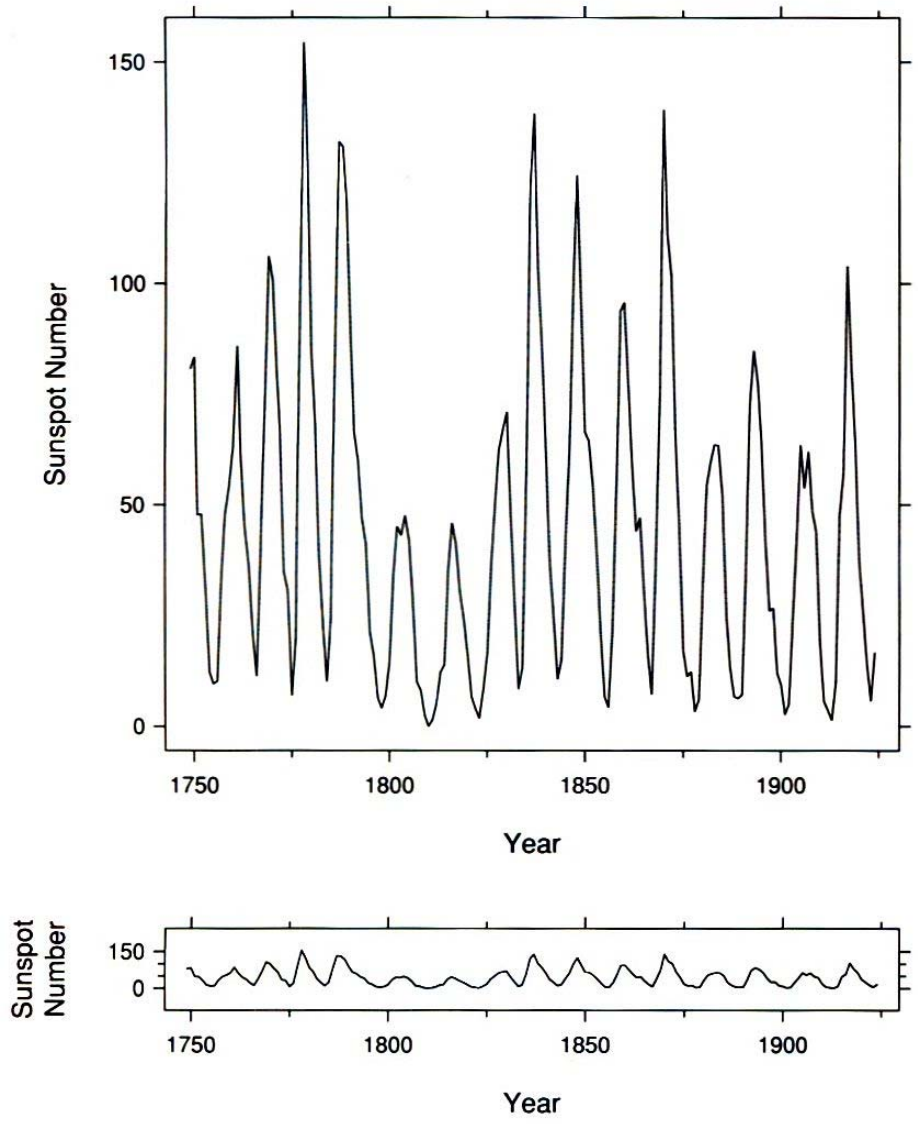
Length of axis on the page: Most graphics packages default to a square or large rectangle for the scale-line rectangle, but there are times when this masks important features of the data. Figure 11 below shows one such example.[12] The upper graph is how one would typically create the plot, which at first glance appears to be a series of oscillating spikes in the sunspot number over time. However, when one maintains the same relationship between the data rectangle and the scale-line rectangle while shrinking the length of the y-axis, a feature of the data becomes more obvious: the peaks are not normal distributions, but rise more sharply than they fall. In other words, they are skewed, with a trailing off of sunspot activity after the rapid rise, and this phenomenon appears to by relatively cyclic.

JMP is somewhat unique in its dynamic ability to drag the axes to various and sundry lengths, making this kind of evaluation effortless. Doing so also falls under the ninth commandment and will be reviewed in more detail there. The default values for the axes for JMP analyses tend to conform well with the above admonitions, while allowing the freedom to change the axes as needed. One option missing in JMP is the ability to put in axis breaks (at least I haven't found how to do it yet; Excel

[11] Huff, page 65.
[12] Cleveland, page 4.

doesn't allow for it either. In fact, SigmaPlot is the only package of which I am aware that does.). This is consistent with the rules of thumb above.

Figure 11:



#2: *Thou shalt use color to categorize, not accessorize.*

Color can be a powerful method for encoding data, genuinely enhancing the visual decoding of the information in the data. But it can also merely "accessorize," that is, be used to no purpose related to data decoding. As Cleveland notes:

> "Our tendency is to be mislead into thinking we are absorbing relevant information when we see a lot. But the success of a graphical method should be based solely on the amount we learn about the phenomenon under study, not on the amount of glitz on the display."[13]

In actual fact, there are only two uses of color that transmit useful information to the viewer. One is to encode categorical data to provide efficient assessment of each category as a whole, visually filtering out the other categories based on their different colors. The second use is to encode the values of a quantitative numerical function of two factor variables by a color level plot, more frequently known as a contour plot.

Encoding a categorical variable: Figure 12 shows the same data plotted with texture symbols (12a)[14] and with color (12b).[15] The color coding allows for a more efficient visual assembly of the categories.

[13] Cleveland, page 209.
[14] Cleveland, page 211.
[15] Cleveland, Figure I at beginning of book.

Figure 12a:

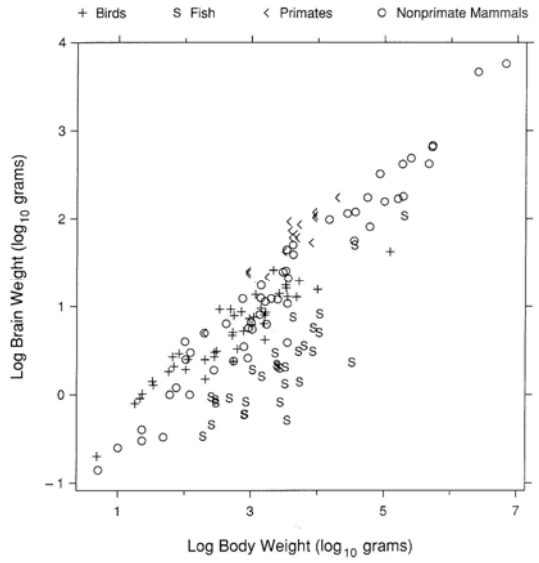
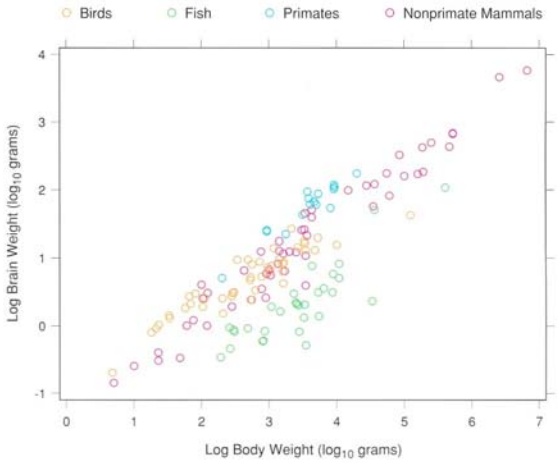


Figure 12b:



Encoding a quantitative variable: Contour plots are two dimensional plots of a function defined by two variables plotted on the x and y axes, the color denoting the value of the function at each combination of the two variables. Figure 13 below[16] is a good example.

Figure 13:

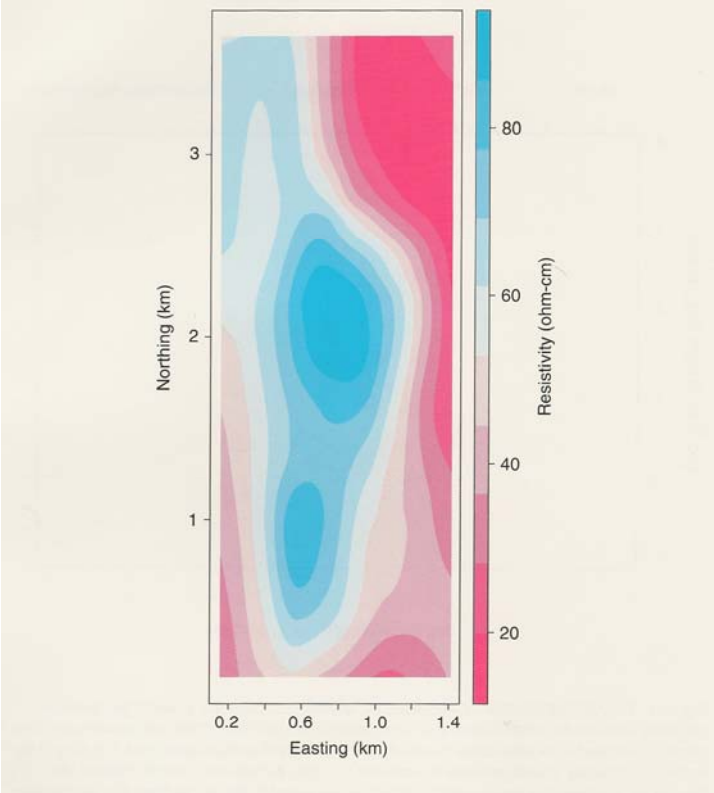


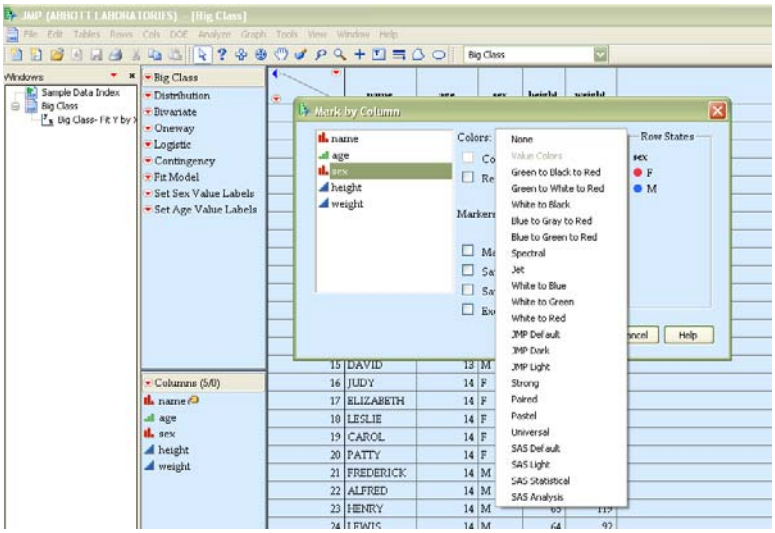
Figure II. COLOR ENCODING A QUANTITATIVE VARIABLE. The second use of color is to display a function of two variables on a level plot. In this figure, color encodes soil resistivity to show how resistivity varies geographically. The encoding method achieves two goals: effortless perception of the order of the encoded quantities, and clearly perceived boundaries between adjacent levels. A number of encoding methods — for example, the much used rainbow encoding — do not achieve both of these goals.

[16] Cleveland, Figure II at beginning of book.

As noted in the caption of Figure 13, the choice of color for a contour plot must achieve two goals: effortless perception of the order of the values (i.e., we do not want to be constantly referring to a key), and clearly perceived boundaries between adjacent levels, two goals that are harder to achieve simultaneously than they sound.

JMP provides the user with both standard color schemes and several other color sequences (see Figure 14) that give range to customization that should be chosen with care, as noted by this commandment.

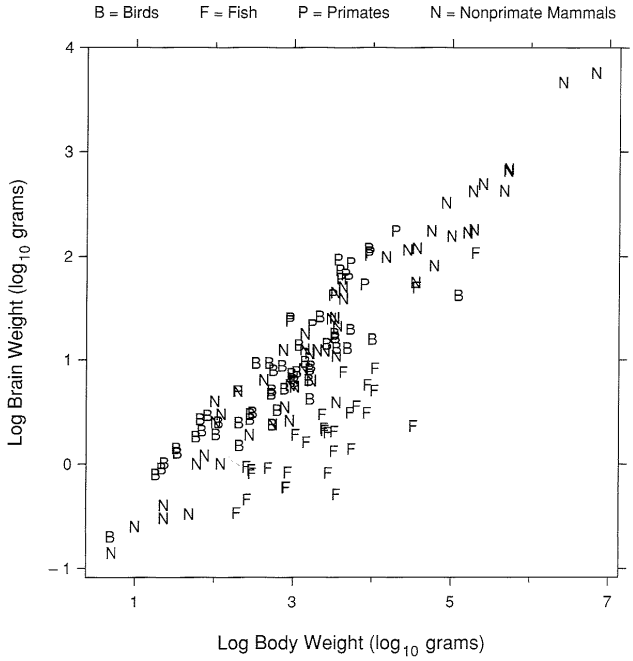
Figure 14:



#3: *Thou shalt choose symbols that can be easily distinguished from one another.*

This "commandment" is primarily concerned with plots in which the data overlaps so that discerning the different datasets being plotted becomes critical. The data from Figure 12 above show one such instance in which the categorical distinctions are readily indentified and grouped visually by either the texture symbols in Figure 12a or better, by the color encoding in Figure 12b. Figure 15 below shows a more dismal effort in which the first letter of each category is used, and the confusion which results in the area where overlap is predominant.

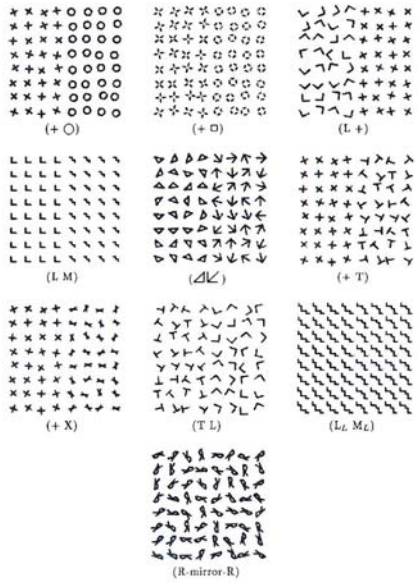
Figure 15:



Each symbol will have its own unique "texture," a micropattern with local variations in its formation that nevertheless take on a uniform appearance to our visual system. Some micropatterns will be more readily distinguished from others, forming

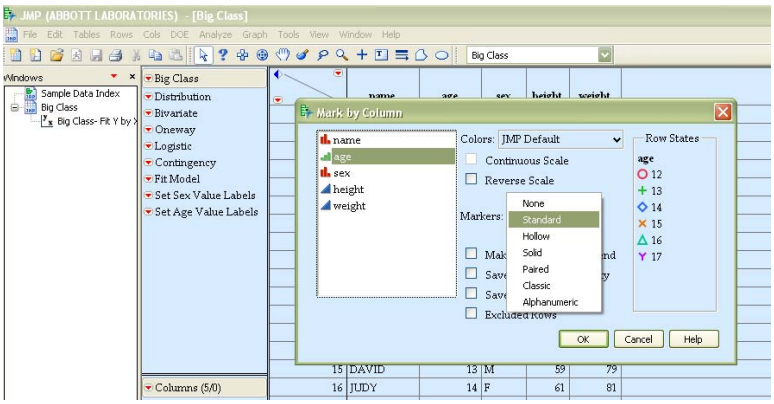
easily observed boundaries which allow for easier decoding when these symbols are close or overlapping. Figure 16[17] shows ten contrasting textures which range from a strong boundary for "+" and "o", to none for "R" and "mirror-R".

Figure 16:



JMP provides several sets of markers that should be evaluated with this commandment in mind (see Figure 17).

Figure 17:



#4: *Thou shalt not employ "chartjunk."*

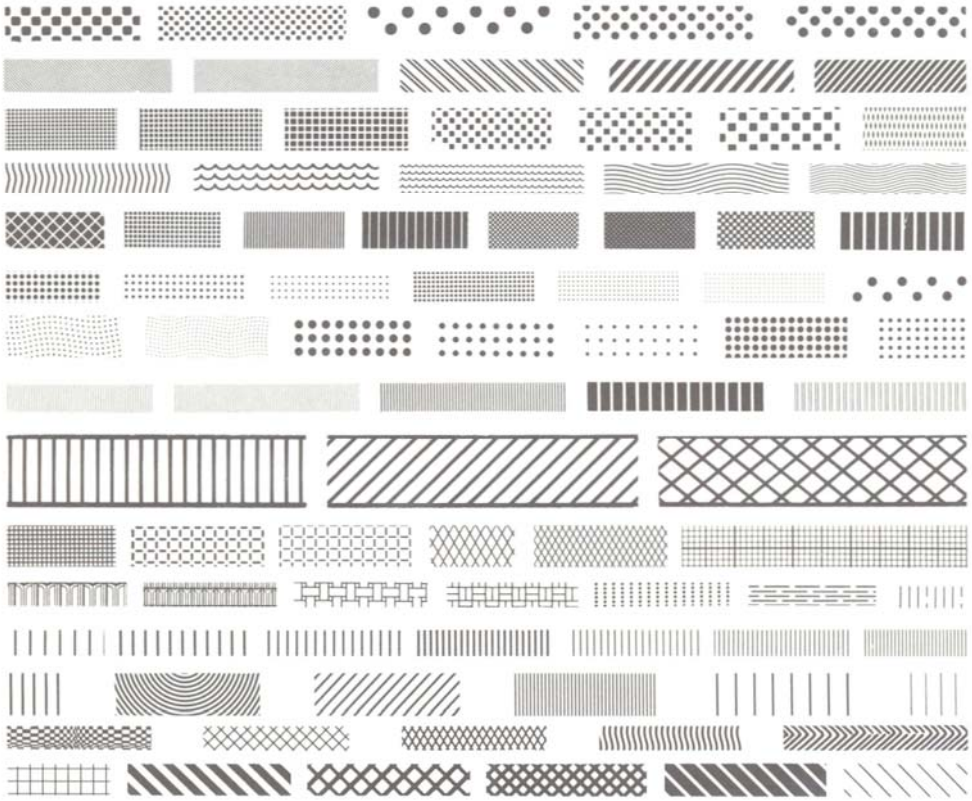
This term is Tufte's creation, but it aptly describes graphical decoration that he catalogues into three types:

1. unintentional optical art,
2. the dreaded grid, and
3. the self-promoting graphical duck!

Unintentional optical art: The moiré effect describes the phenomenon when the graphic design interacts with the physiological tremor of the eye to generate the distracting appearance of vibration and movement. This adds unnecessary noise to the decoding process and is particularly prominent in the "fill" for bar charts. Figure 18 shows a selection of such fills that fall into this category.[18]

[17] Cleveland, page 238.
[18] Tufte, page 111.

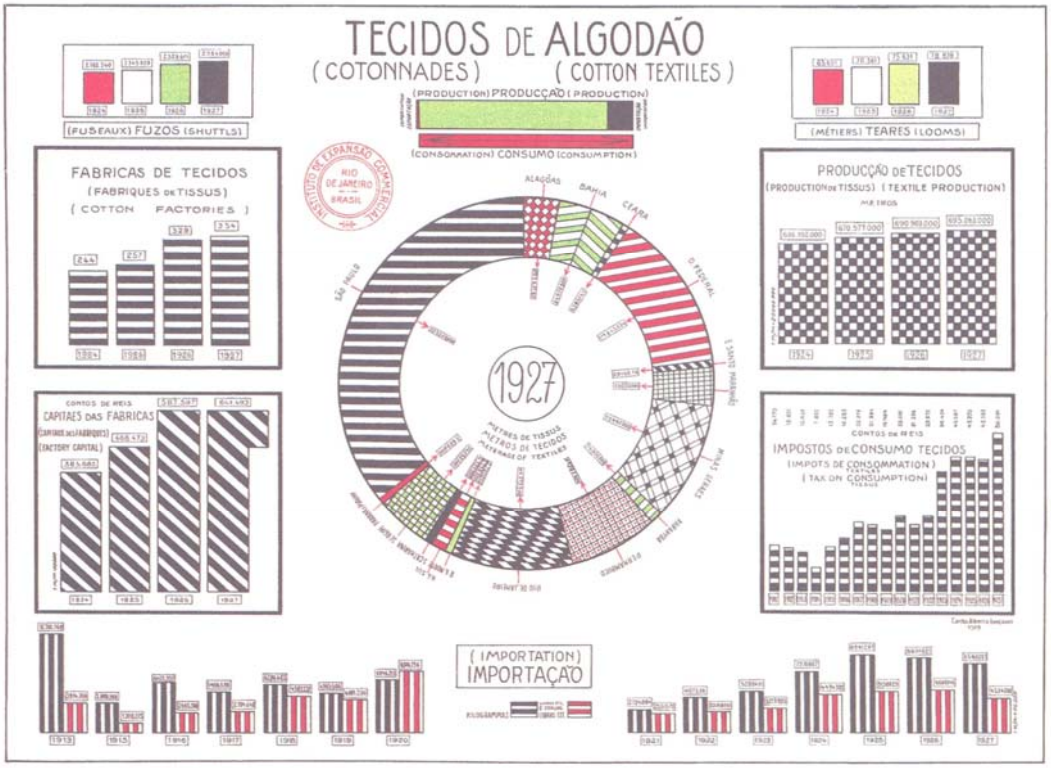
Figure 18:



Those of us who "grew up" on Microsoft Excel charts will recognize this selection of fill patterns immediately.

Figure 19 gives us an example of what not to do with these patterns. The word "garish" comes to mind!

Figure 19:



The dreaded grid: The purpose of grid lines, when used at all, should be to enhance the reading of the data. Dark grid lines enter the realm of chartjunk by competing with the data for the attention of the viewer. Thus, they should be muted or completely suppressed. Figure 20 shows us an instance in which the data is barely visible in the mass of grid lines.[19]

Figure 20:

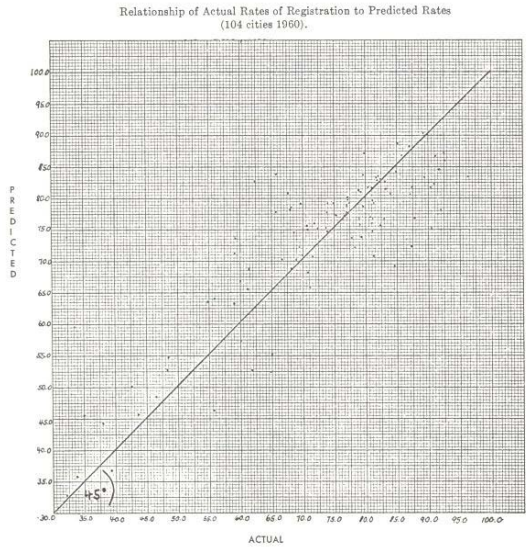


Figure 21:

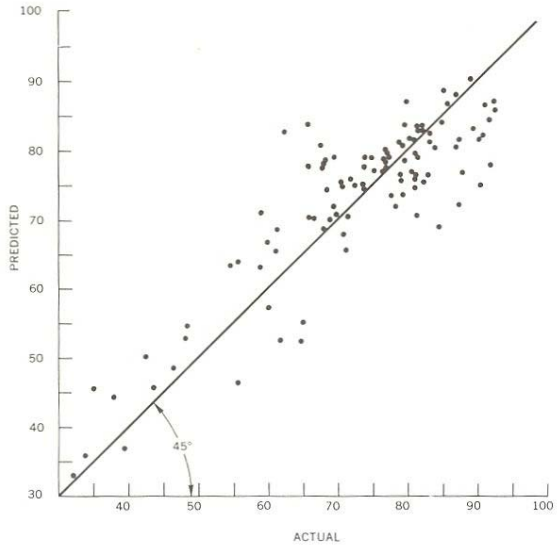
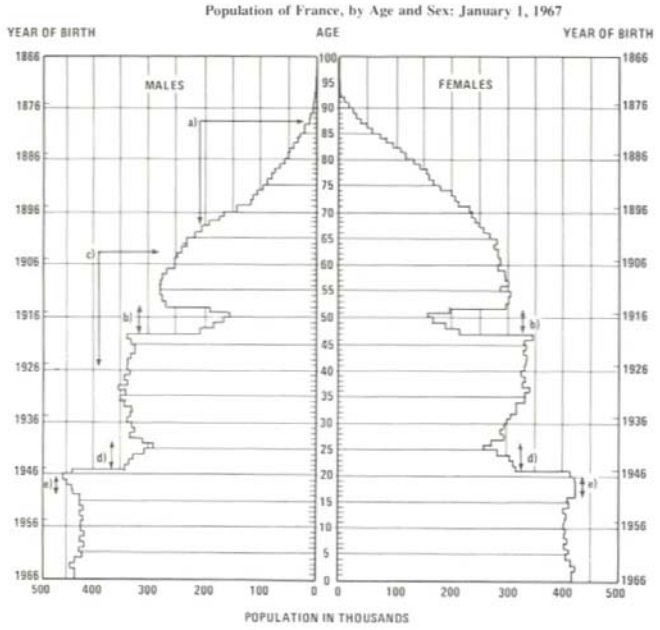


Figure 21 shows the same data without the grid lines. The difference in clarity is stunningly obvious.[20]

Of course, not all cases are so obvious. Figure 22 shows an example in which the grid makes the profile of the data more difficult, but not impossible, to read.[21]

Figure 22:



The self-promoting graphical duck: This last category had best be described in Tufte's own words (see Figure 23):[22]

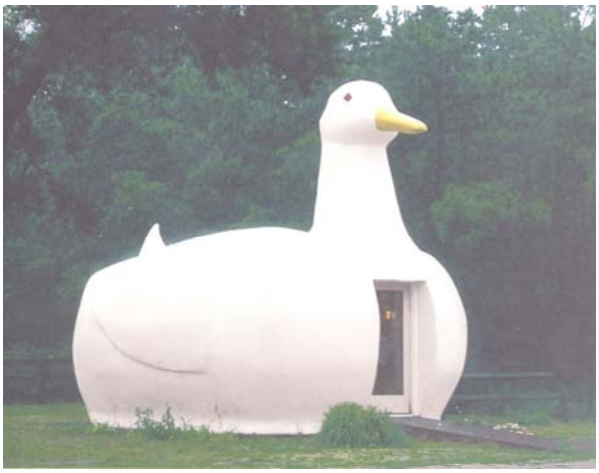
> When a graphic is taken over by decorative forms or computer debris, when the data measures and structures become Design Elements, when the overall design purveys Graphical Style rather than quantitative information, then that graphic may be called a *duck* in honor of the duck-form store, "Big Duck." For this building the whole structure is itself decoration, just as in the duck data graphic. In *Learning from Las Vegas*, Robert Venturi, Denise

[19] Tufte, page 94.
[20] Tufte, page 95.
[21] Tufte, page 113.
[22] Tufte, pages 116-117.

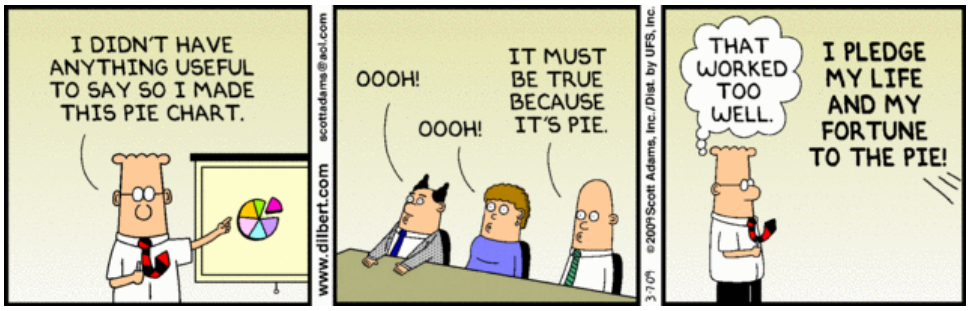
Scott Brown, and Steven Izenour write about the ducks of modern architecture – and their thoughts are relevant to the design of data graphics as well:

> "When Modern architects righteously abandoned ornament on buildings, they unconsciously designed buildings that *were* ornament. In promoting Space and Articulation over symbolism and ornament, they distorted the whole building into a duck. They substituted for the innocent and inexpensive practice of applied decoration on a conventional shed the rather cynical and expensive distortion of program and structure to promote a duck.... It is now time to reevaluate the once-horrifying statement of John Ruskin that architecture is the decoration of construction, but we should append the warning of Pugin: It is all right to decorate construction but never construct decoration."

Figure 23: *Big Duck*



One can also point out that, in most cases, the charts widely used in mass media and business publications, to wit, the pie chart, divided bar charts, and area charts, will violate this commandment when their use is attempted in science and technology.

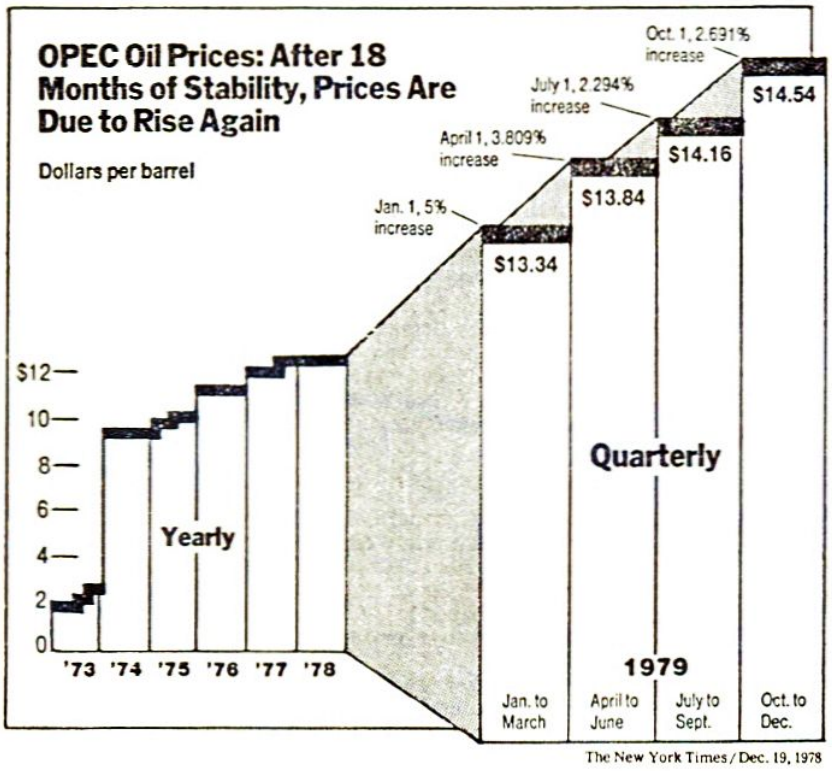


#5: *Thou shalt show variation in thy data, not in thy design.*

A common source for ambiguity and deception when agendas other than clear data presentation are at work is the confounding of *design variation* with *data variation*. Data in the real world (particularly the biological data with which the current author works) invariably contain variation. That variation the graphic designer wants to show with clarity. Indeed, that is often the story the data are telling. But when variation in design is used, the eye may confuse the changes in design with changes in the data, and even worse, may not recognize this confusion, resulting in incorrect conclusions and subjective impressions of the data that are misleading. Figure 24 shows an example of a chart in which design variation corrupts the data display.[23]

[23] Tufte, page 61.

Figure 24:



The corruption becomes apparent after one does a few measurements and finds that five different vertical scales are used to show the price and two different horizontal scales to show the passage of time without one indication of these changes (not even a scale break)!

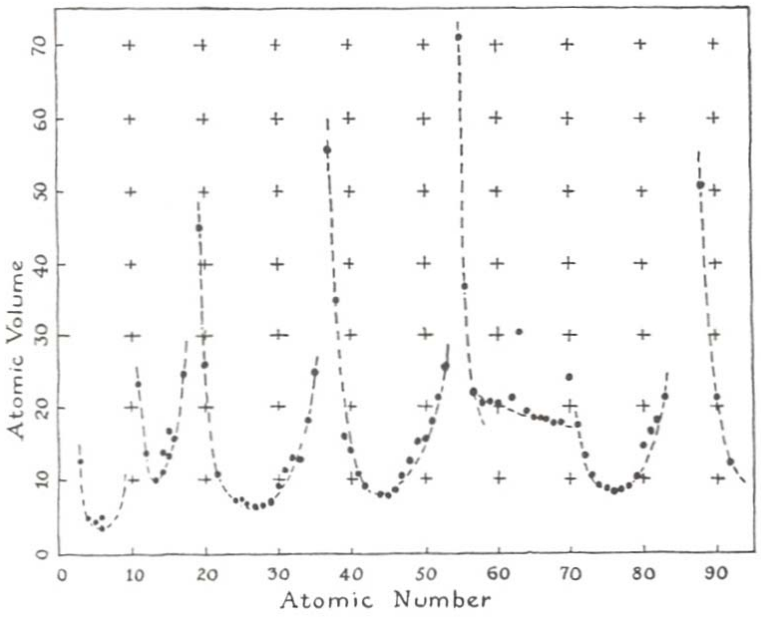
During this time	*One vertical inch equals*	*One horizontal inch equals*
1973-1978	$8.00	3.8 years
January-March 1979	$4.73	0.57 years
April-June 1979	$4.37	
July-September 1979	$4.16	
October-December 1979	$3.92	

#6: *Thou shalt maximize the data-ink ratio in thy graphs.*

The data-ink ratio is quite simply the amount of "ink" used to depict the actual data divided by the total "ink" used to print the graphic.[24] In other words, it is the proportion of a graphic display given to the non-redundant portrayal of data-information. This "commandment" is really the converse of #4 that commands avoiding chartjunk. Redundant data-ink depicts the same number repeatedly, and gratuitous decoration (for example, three dimensional effects added to bar charts) can easily be erased with no loss in the information communicated by the graphic. Consider the graphic of Figure 25, drawn by the distinguished science illustrator Roger Hayward showing the elements of the periodic table by atomic number and atomic volume.[25] As Tufte notes, "The data-ink ratio is less than 0.6, lowered because the 76 data points and the reference curves are obscured by the 63 dark grid marks arrayed over the data plane like a precision marching band of 63 mosquitoes."

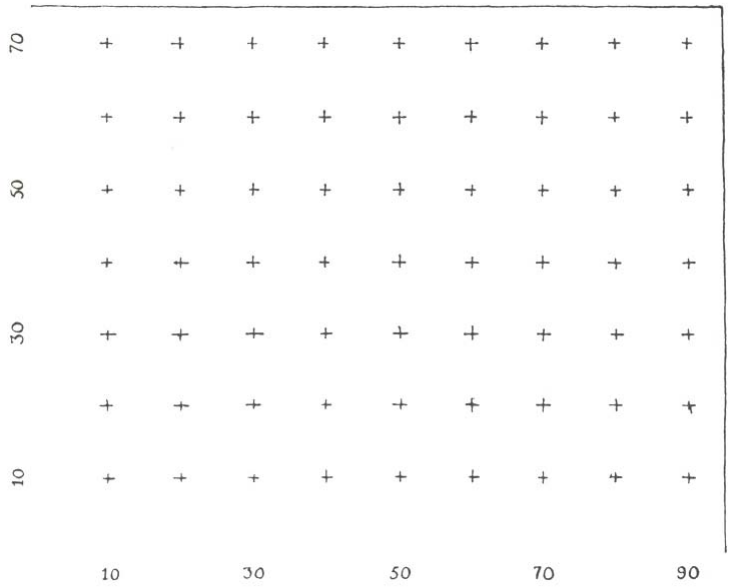
[24] I put "ink" in quotes in recognition of the fact that more often today graphics rarely use ink or see cellulose, being confined to the pixels of a computer screen. Nonetheless, the concept remains valid.
[25] Tufte, page 102.

Figure 25:



A significant amount of ink can be removed from this graphic, rendering the data with more clarity. Figure 26 shows the amount of "ink" that is cluttering up the display.[26]

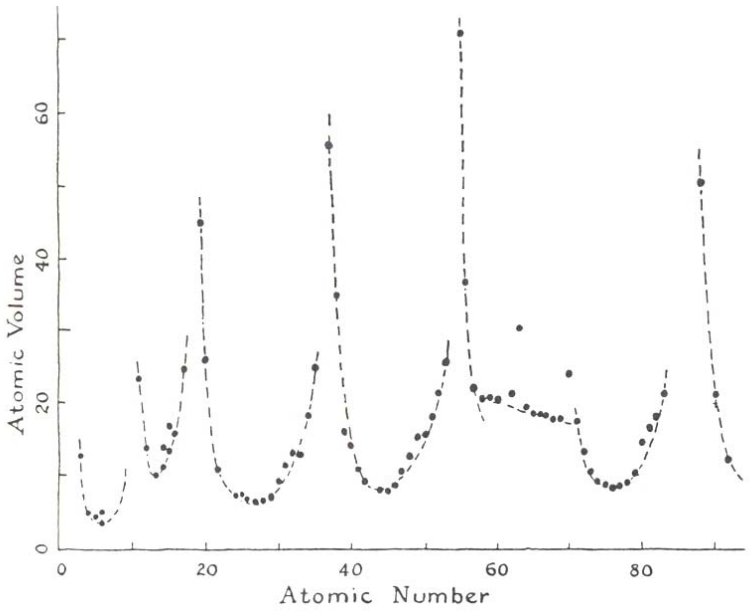
Figure 26:



Included are some of the tick labels, part of the frame, and our squadron of precision mosquitoes! Figure 27 presents us with a significantly de-cluttered display of the data in which it is revealed that several of the elements do not fit the smooth theoretical curves all that well.[27] Removal of the components in Figure 25 has increased the data-ink ration to about 0.9.

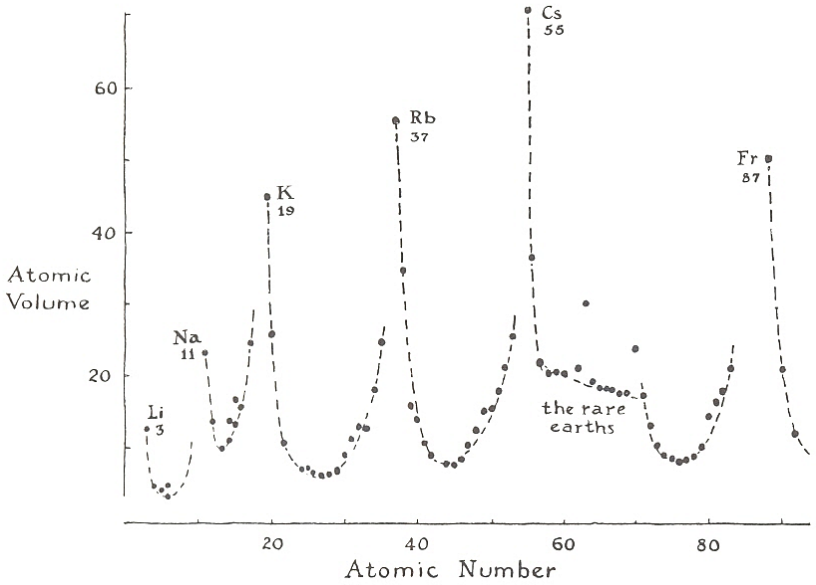
[26] Tufte, page 103.
[27] Ibid.

Figure 27:



Tufte then goes on to show that the reference curves are essential for organizing the data to show the periodicity by providing a structure and an ordering to the flow of the information. Likewise, restoring just the grid fails to do this. However, he also notes that the space opened up in the graphic can be put to effective use by the inclusion of some labeling of some of the relevant characteristics of the elements. In addition, the y axis label and numbers can be rotated to make them easier to read, more viewer friendly (see Figure 28).[28]

Figure 28:

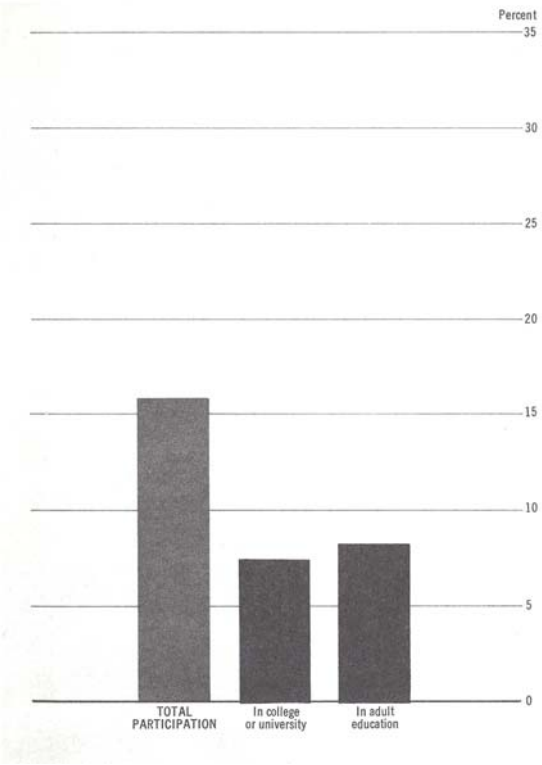


#7: *Thou shalt maximize thy data density and the size of thy data matrix (within reason).*

The human eye has the ability to detect large amounts of information in small spaces. This commandment tells us to take advantage of that ability by seeking to maximize the number of entries in the data matrix per unit area of a data graphic, i.e., the data density. Figure 29 an extreme example of low data density.[29]

[28] Tufte, page 105.
[29] Tufte, page 163.

Figure 29:

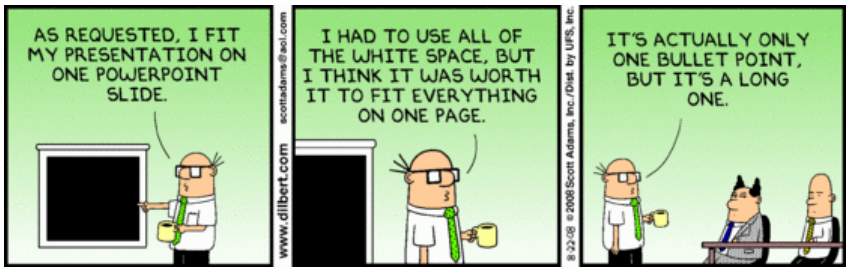


Here the data matrix contains only four entries, the names and the numbers for the two bars on the right. The bar on the left is the sum of the other two. The original graph covers 26.5 square inches, and dividing 4 by 26.5 gives us the rather thin data density of 0.15 numbers per square inch. Contrast this with Figure 30.[30] This map of France was originally 27 square inches (close to that of Figure 29). It shows the location and boundaries of 30,000 French communes. To recreate the data of the map would require somewhere in the neighborhood of 240,000 numbers: 30,000 latitudes, 30,000 longitudes, and an average of six numbers describing the shape of each commune. The data density thus works out to be nearly 9,000 numbers per square inch.

Figure 30:

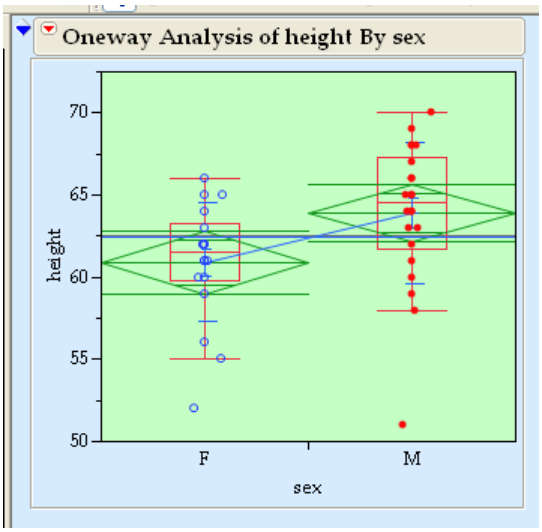
[30] Tufte, page 166

Of course, one can take this to an extreme, and the phrase, "within reason," should be aptly applied here.



As with most software, JMP does, indeed, allow for "too much" information to be shown on a graph. Creator discretion is needed to comply with this commandment (see Figure 31).

Figure 31:



#8: *Thou shalt draw the viewer's eye to the data, not to other design elements.*

The whole purpose of statistical graphics is to tell the story contained in the data, and without the data, there is no story. Moreover, simple honesty requires that the data be prominent to ensure that the story you are trying to communicate is the one the data truly contains. Thus, the fundamental principle of good statistical graphics is: above all else show the data. Many of the foregoing commandments could be considered corollaries of this one (remember, I did say these did not come from Mt. Sinai!). Some additional ways to comply with this principle include:

- Use visually prominent graphical elements to show the data
- Don't clutter the interior of the scale-line rectangle with legends, labels, and lines
- Tick marks should generally face outward
- Use reference lines only when an important value must be seen across the entire graph, and then use a color, weight and style of line that does not overpower the data symbols
- If data labels are used inside the scale-line rectangle, don't allow them to interfere with the data or to clutter the graph
- Don't put notes and keys inside the scale-line rectangle; notes should go in a caption or the accompanying text
- When datasets are superimposed, choose color, symbol, line weights and styles, and other such graphical elements so that the datasets can be readily visually distinguished

JMP default values for its graphs comply with most of the above, and allow for customization in the rare instances where modification of the defaults is needed.

#9: *Thou shalt do and redo thy graphs to determine which one telleth the story best.*

Just as a good writer works and reworks his efforts, so too the creator of statistical graphics should not rely upon the "we've always done it that way" approach to graphical design. The concept of seeking continual improvement in whatever field of endeavor we are active should become a part of our thinking here as well. Never fear to revise and edit your graphic design in a search for that which communicates most clearly to the largest audience. It would be well to remember that this process is a complex and multivariate one in which not only efficiency, but complexity, structure, density, and even beauty have a role to play in the generation of the final product. Thus, experimentation should be encouraged in as many of these areas as possible. JMP does an admirable job of providing not only a multiplicity of ways to graph data, but a myriad of options to

"play" with the resulting output to do such experimentation in presentation. The Graph Builder's ability to redo plots "on the fly" is of tremendous benefit here and is unique to my knowledge in this flexibility. For example, Figures 32 – 35 below show several possible plots from the Big Class example data, all created in moments with this JMP feature.

Figure 32: Simplest

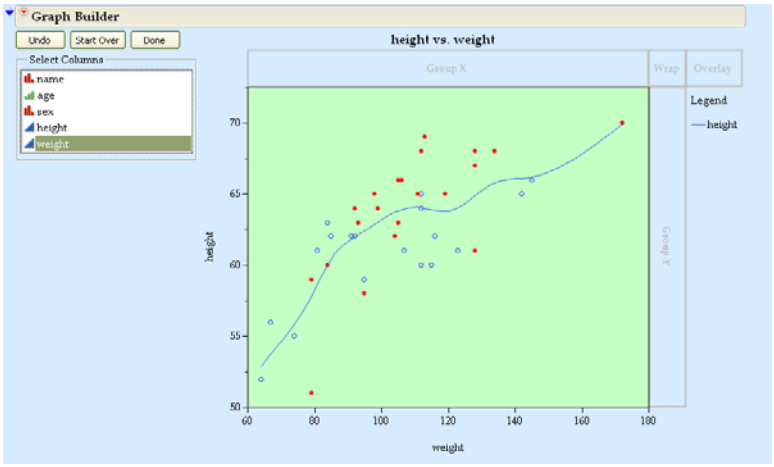


Figure 33: Wrapping by sex

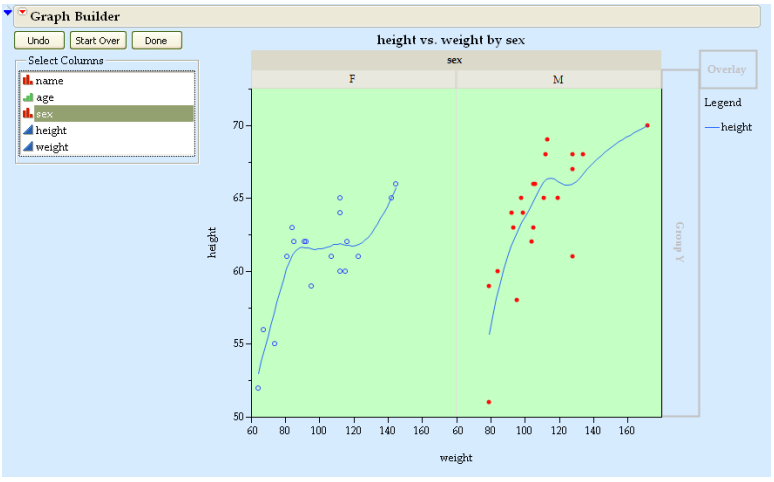
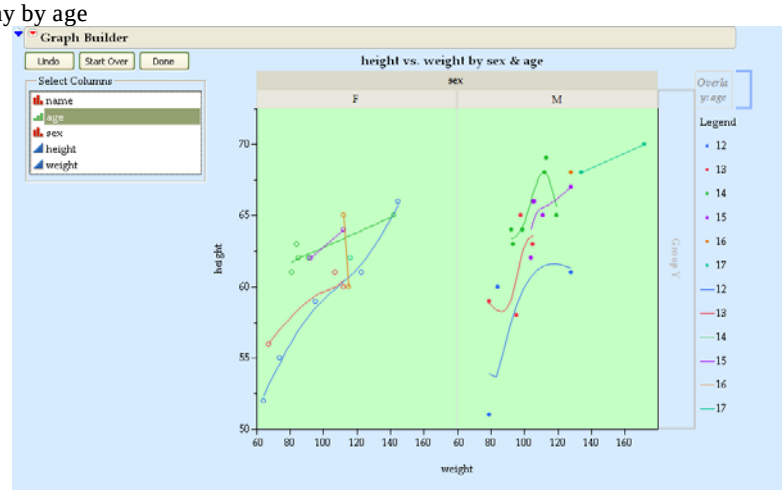
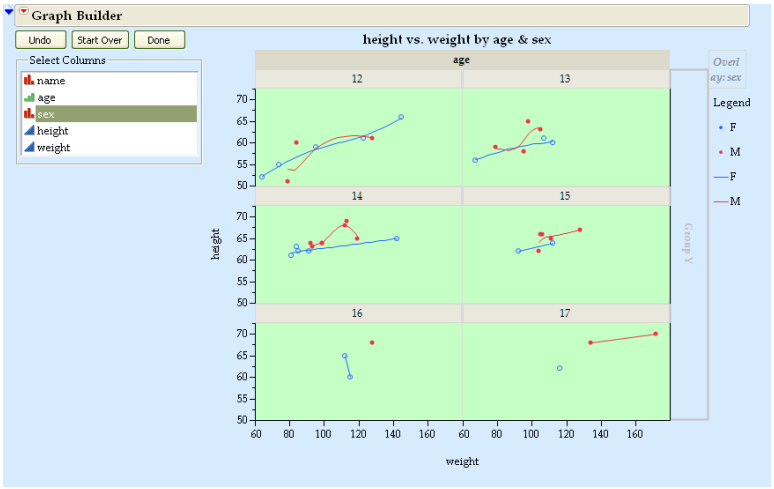


Figure 34: Adding overlay by age



Don't like that? Hit undo twice, wrap by age and overlay by sex gives Figure 35. The possibilities and potential are extraordinary (and no, SAS didn't pay me to say that!).

Figure 35: Redo



#10: *Thou shalt not create "unfriendly" but "friendly" data graphics.*

Our final commandment is a way of saying that the creator of a graphic should try to keep in mind the viewer of his creation and not get lost in his own world of graphical elements. Tufte provides us with a nice compilation of some of the specific characteristics of friendly versus unfriendly data.[31]

Friendly	*Unfriendly*
Words are spelled out, mysterious and elaborate encoding avoided	Abbreviations abound, requiring the viewer to sort through text to decode abbreviations
Words run from left to right, the usual direction for reading occidental languages	Words run vertically, particularly along the y-axis; words run in several different directions
Little messages help explain data	Graphic is cryptic, requires repeated references to scattered text
Elaborately encoded shadings, cross-hatching, and color are avoided; instead, labels are placed on the graphic itself; no legend is required	Obscure codings require going back and forth between legend and graphic
Graphic attracts viewer, provokes curiosity	Graphic is repellent, filled with chartjunk
Colors, if used, are chosen so that the color-deficient and color-blind (5 to 10 percent of viewers) can make sense of the graphic (blue can be distinguished from other colors by most color-deficient people)	Design insensitive to color-deficient viewers; red and green used for essential contrasts
Type is clear, precise, modest; lettering may be done by hand	Type is clotted, overbearing
Type is upper-and-lower case, with serifs	Type is all capitals, sans serif

With regard to typography, Tufte has a very interesting quote[32] from Josef Albers (*emphases* added):

> The concept that "the simpler the form of a letter the simpler its reading" was an obsession of beginning constructivism. It became something like a dogma, and is still followed by "modernistic" typographers.... Ophthalmology has disclosed that *the more the letters are differentiated from each other, the easier is the reading*. Without going into comparisons and details, it should be realized that *words consisting of only capital letters present the most difficult reading* – because of their equal height, equal volume, and with most, their equal width. When comparing serif letters with *sans-serif*, the latter provide *an uneasy reading. The fashionable preference for sans-serif in text shows neither historical nor practical competence*.

The astute reader at this point will note that this author has practiced what he preaches by employing a serif font for this paper!

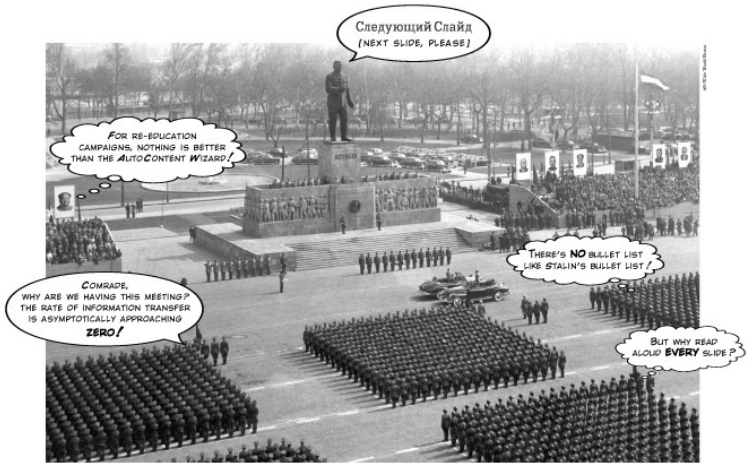
A WORD ABOUT POWERPOINT

Microsoft PowerPoint is a powerful presentation software program that has been blamed for everything from two space shuttle disasters[33] to failure of military operations in Iraq.[34] Indeed, one of the heros of this paper, Dr. Edward Tufte, has

[31] Tufte, page 183.
[32] Ibid.
[33] See http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001yB&topic_id=1
[34] See, e.g., http://blogs.zdnet.com/storage/?p=878&tag=nl.e540

written a book entitled, *The Cognitive Style of PowerPoint*, in which he charges PowerPoint with the following weaknesses that impact presenters and audience alike:

- **Cognitive style.** Presenter-focused, not content or audience focused.
- **Low resolution.** Little info per slide - so more slides are needed. Data graphics are weak: average of 12 numbers per graphic.
- **Bullets.** Bullet lists can show only 3 logical flows: sequence; priority; or membership. Multivariate models with feedback and simultaneity can't be listed. This encourages lazy thinking, generic ideas and ignores critical relationships and assumptions.



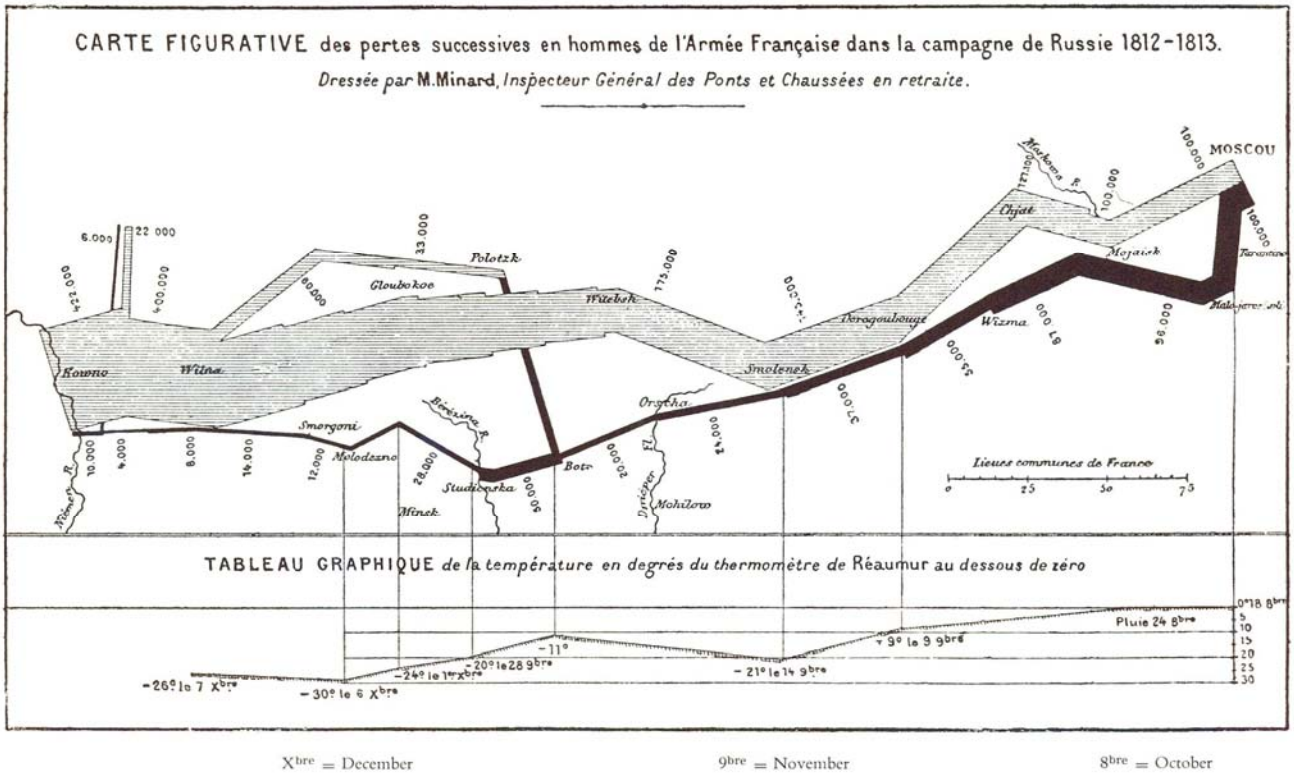
Edward Tufte, *The Cognitive Style of PowerPoint*

While this author agrees that there is some point to these charges, they all seem to ignore the fact that PowerPoint, or any other presentation software, is just a tool. To blame the tool for its misuse is to kill the messenger for his message. What is needed is not condemnation of the tool but proper instruction of the use of that tool. [Note: the following is a politically incorrect personal opinion based on observation; reader discretion is advised.] Unfortunately, that requires that we humble ourselves to admit to the problem in the first place. Our current society is a reflection of an education system that has been teaching our students for decades *what* to think rather than *how* to think. We are reaping the results of this flawed system and the abuse of PowerPoint is just one symptom of the problem.

COMPETITORS FOR BEST AND WORST

A fitting penultimate point for this paper is to view what Professor Tufte has nominated for the best and worst graphics ever produced as lessons in the practical application of, or ignoring of, the principles delineated in this work. First, the "Best."

Figure 36:



Xbre = December 9bre = November 8bre = October

This is the classic multivariate depiction of Charles Joseph Minard (1781-1870) drawn in 1861 showing the devastating losses of Napoleon's army in its march into and out of Russia in 1812. It is a combination of a data map and a time-series plotting no less than *six* variables: the size of the army (1), its location on a two-dimensional surface (2, 3), direction of the army's movement (4), and temperature (5) on various dates (6) during the retreat from Moscow. Tufte describes the history from the graphic concisely:

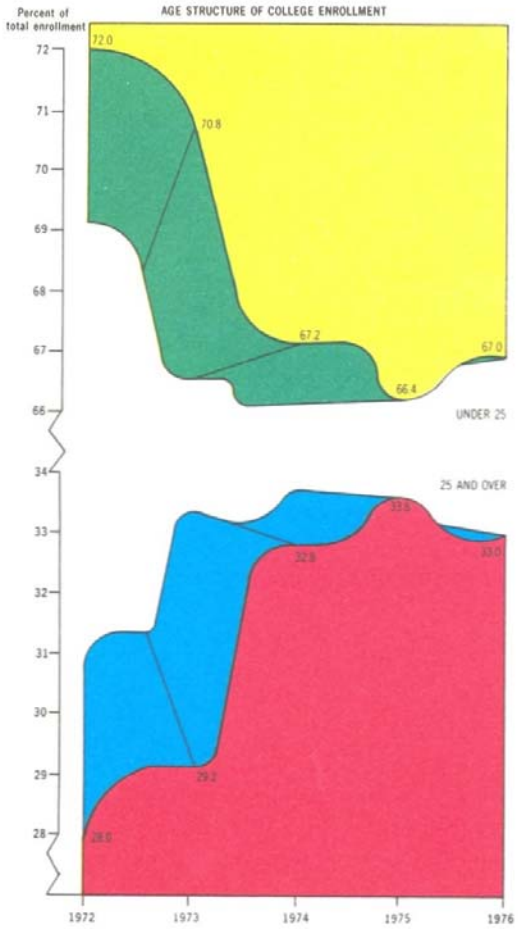
> Beginning at the left on the Polish-Russian border near the Niemen River, the thick band shows the size of the army (422,000 men) as it invaded Russia in June 1812. The width of the band indicates the size of the army at each place on the map. In September, the army reached Moscow, which was by then sacked and deserted, with 100,000 men. The path of Napoleon's retreat from Moscow is depicted by the darker, lower band, which is linked to a temperature scale and dates at the bottom of the chart. It was a bitterly cold winter, and many froze on the march out of Russia. As the graphic shows, the crossing of the Berezina River was a disaster, and the army finally struggled back into Poland with only 10,000 men remaining. Also shown are the movements of auxiliary troops, as they sought to protect the rear and the flank of the advancing army.[35]

In the words of E. J. Marey, this graphic seems "to defy the pen of the historian by its brutal eloquence."[36]

And for the "Worst" we have a three-dimensional display from the magazine *American Education* published in the 1970s. In Tufte's words, a series of such displays "delighted connoisseurs of the graphically preposterous."[37] In these works of art, five apparently randomly chosen colors report only five pieces of data. See if you can identify the other commandments broken in the creation of this masterpiece.

[35] Tufte, page 40.
[36] Ibid.
[37] Tufte, page 118.

Figure 37:



THE CONCLUSION OF THE MATTER

Tufte's epilogue to *The Visual Display of Quantitative Information* is so eloquent in putting an overall perspective on this subject that I will close with it verbatim with some *emphases* added.[38]

> Design is choice. The theory of the visual display of quantitative information consists of principles that generate design options and that guide choices among options. The principles should not be applied rigidly or in a peevish spirit; they are not logically or mathematically certain; and *it is better to violate any principle than to place graceless or inelegant marks on paper*. Most principles of design should be greeted with some skepticism, for word authority can dominate our vision, and we may come to see only through the lenses of word authority rather than with our own eyes.
>
> What is to be sought in designs for the display of information is the *clear portrayal of complexity*. Not the complication of the simple; rather *the task of the designer is to give visual access to the subtle and the difficult – that is,*
>
> *the revelation of the complex.*

[38] Tufte, page 191.

REFERENCES CITED IN TEXT

Cleveland, William S. 1994. *The Elements of Graphing Data*. Revised Edition. Summit, New Jersey: Hobart Press.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2nd Edition. Cheshire, Connecticut: Graphics Press.

Huff, Darrell. 1954. *How to Lie with Statistics*. New York, New York: W. W. Norton & Company.

Zumel, Nina. 2009. *Good Graphs: Graphical Perception and Data Visualization*. http://www.win-vector.com/, accessed 4 June 2010.

Pirrello, Chuck. 2010. *Effective Visualization Techniques for Data Discovery and Analysis*. Cary, North Carolina: SAS Institute, Inc.

ADDITIONAL USEFUL REFERENCES

Few, Stephen. 2009. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Oakland, California: Analytics Press.

Cleveland, William S. 1993. *Visualizing Data*. Summit, New Jersey: Hobart Press.

Tufte, Edward R. 1990. *Envisioning Information*. Cheshire, Connecticut: Graphics Press.

Annesley, Thomas M. 2010. *Put Your Best Figure Forward: Line Graphs and Scattergrams*. Clinical Chemistry 56 (8): 1229-1233.

Bessler, LeRoy. 2004. *Communication-Effective Use of Color for Web Pages, Graphs, Tables, Maps, Text, and Print*. SUGI 29, Montreal, Canada. http://www2.sas.com/proceedings/sugi29/176-29.pdf, accessed 2 July 2010.

AUTHOR BIO

Steve received his Bachelor's (neurobiology) from Cornell University, a Master's (chemistry) from Northern Illinois University, and was finally ejected with a Ph.D. (biochemistry) from Florida State University in 1984. After a two year post-doc at Los Alamos National Laboratory, he escaped to Abbott Laboratories in Chicagoland, where he's been hiding ever since, developing automated *in vitro* diagnostic immunoassays. Having traumatized his Ph.D. major advisor by consistently yet unsystematically changing multiple variables simultaneously between experiments, Steve became proficient over the ensuing years in a methodology that actually requires him to do so (DOE). Meanwhile, having acquired enough statistics to be dangerous, Steve persists in trying to share that knowledge with the world at large.

ACKNOWLEDGMENTS

I would like to acknowledge John Wass, statistician extraordinaire, former co-worker, and good friend, for motivation, helpful hints, and dutiful review of the content of this paper and subsequent presentation. . Also putting up with my idiosyncrasies in reviewing my prose was one of my current co-workers, Dr. Jennifer Steinhaus.

CONTACT INFORMATION

Steve Figard
Abbott Laboratories
Dept 09FE, Building AP20
100 Abbott Park Road
Abbott Park, IL 60064-6015
(847) 937-9089
steve.figard@abbott.com