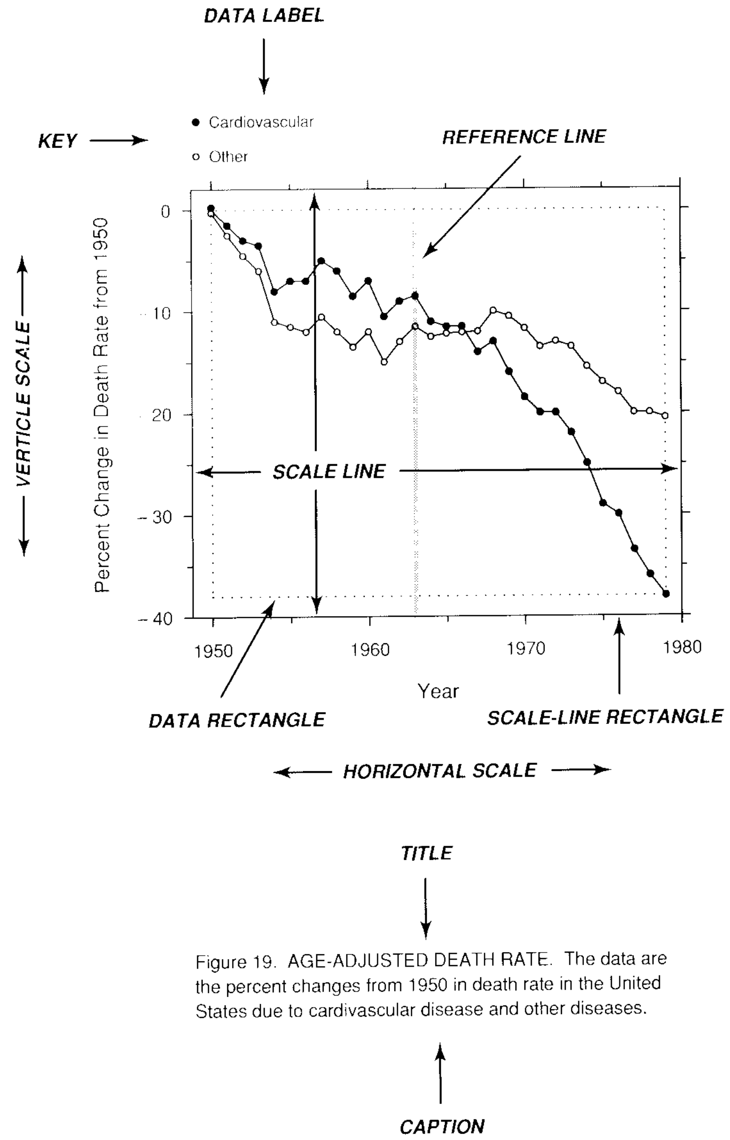


2.1 TERMINOLOGY. This figure and the next define terminology. The two sets of data — death rates due to cardiovascular disease and death rates due to all other diseases — are superposed. The data labels are in the key on this graph.

2 *Principles of Graph Construction*

This chapter is about the basic elements of graph construction — scales, captions, plotting symbols, reference lines, keys, labels, panels, and tick marks. Principles of graph construction are given that can enhance the ability of a graph to show the structure of the data. The principles are based on the study of graphical perception, the topic of Chapter 4. They are relevant both for data *analysis*, when the analyst wants to study the data, and for data *communication*, when the analyst wants to present quantitative information to others.

Graphing data is difficult, and without principles of construction problems can occur. The chapter contains many examples of graphs from science and technology that have problems. The principles are applied to the examples to solve the problems.

Section 2.1 (pp. 23–25) defines terms. Section 2.2 (pp. 25–54) gives principles that make the elements of a graph visually clear, and Section 2.3 (pp. 54–66) gives principles that contribute to a clear understanding of what is graphed. Section 2.4 (pp. 66-79) is about the aspect ratio of a graph — its height divided by its width. Section 2.5 (pp. 80–109) is about scales, and Section 2.6 (pp. 110–118) discusses general strategies for graphing data.

2.1 *Terminology*

Terminology for graphical displays is unfortunately not fully developed and usage is not consistent. Thus, in some cases we will have to invent a few terms and in some other cases we will pick one of several possible terms now in use. Terminology is defined in Figures 2.1 and 2.2, which display the percent changes from 1950 in death rates in the United States due to cardiovascular disease and due to all other diseases [82]. The words in boldface convey the terminology. For the most part, the terms are self-explanatory, but a few comments are in order.

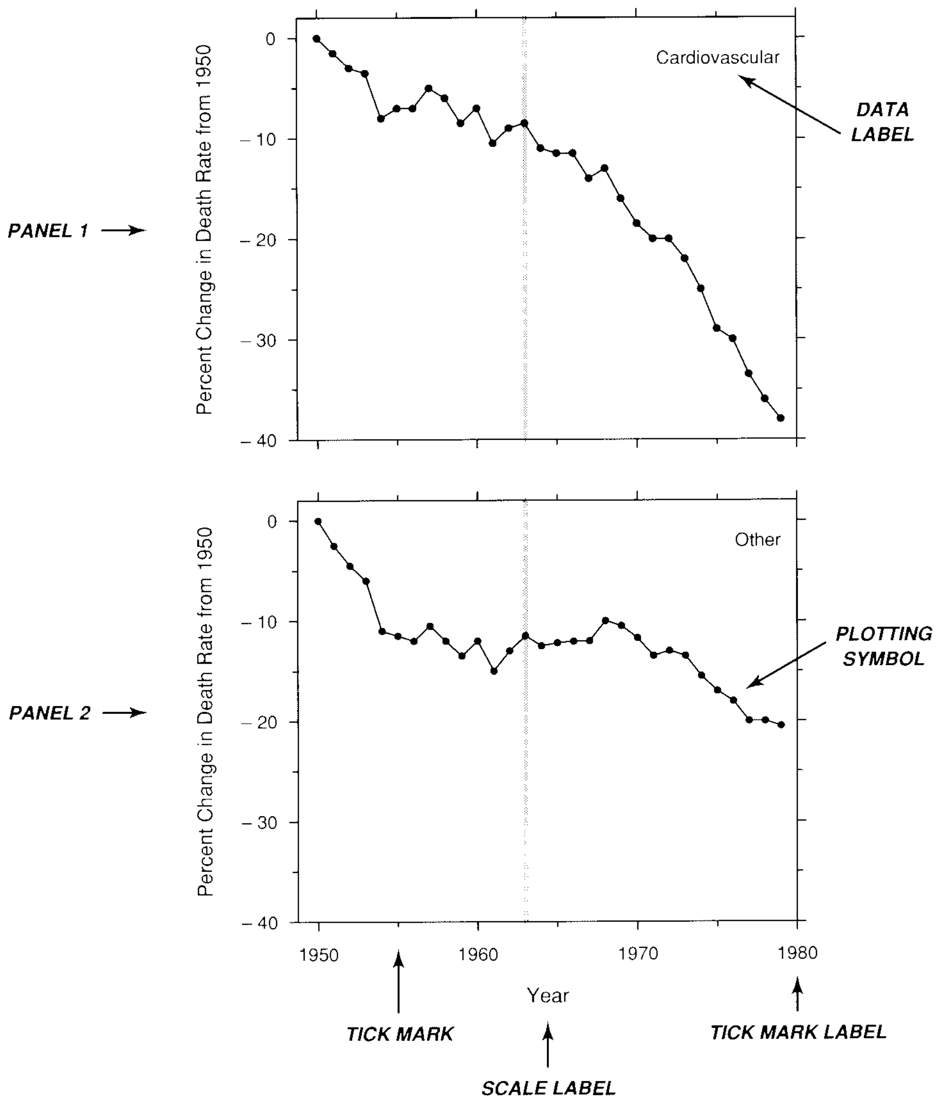


Figure 19. AGE-ADJUSTED DEATH RATE. The data are the percent changes from 1950 in death rate in the United States due to cardivascular disease and other diseases.

2.2 TERMINOLOGY. This figure also defines the meaning of terms. The two sets of data are juxtaposed by using two panels. Each panel on this graph has a data label.

The *scale-line rectangle* is the rectangle formed by the scale lines. The *data rectangle* is the rectangle that just encloses the data. In Figure 2.1 the two data sets are *superposed* and in Figure 2.2 they are *juxtaposed.* The *reference line* shows the time of the first specialized cardiovascular care unit in a hospital in the United States. In Figure 2.1 the *data labels* are part of the *key*, but in Figure 2.2 they are inside the scale-line rectangles.

Scale has two meanings in graphical data display. One is the ruler along which we graph the data; this is the meaning indicated in Figure 2.1. But scale is also used by some to mean the number of data units per cm. This meaning will not be used in this book. Instead, the phrase, *number of units per cm*, will be used. Not every concept needs a single-word definition.

2.2 Clear Vision

Clear vision is a vital aspect of graphing data. The viewer must be able to visually disentangle the many different items that appear on a graph. In this section elementary principles of graph construction are given to help achieve clear vision.

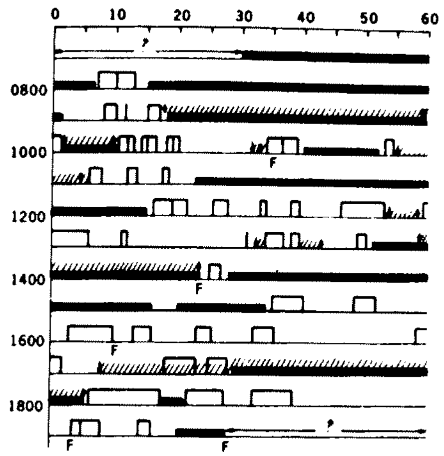
Make the data stand out. Avoid superfluity.

Make the data stand out and *avoid superfluity* are two broad strategies that serve as an overall guide to the specific principles that follow in this section.

The data on a graph are the reason for the existence of the graph. The data should stand out. It is too easy to forget this. There are many ways to obscure the data, such as allowing other elements of the graph to interfere with the data or not making the graphical elements encoding the data visually prominent. Sometimes different values of the data can obscure each other.

We should eliminate superfluity in graphs. Unnecessary parts of a graph add to the clutter and increase the difficulty of making the necessary elements — the data — stand out. Edward R. Tufte puts it aptly; he calls superfluous elements on a graph *chartjunk* [121].

Let us look at one example of implementing these two general principles where the result is increased understanding of the data. Figure 2.3 shows data on a !Kung woman and her baby [74]. The !Kung are an African tribe of hunter-gatherers from Botswana and Namibia whose present culture provides a glimpse into the history of man. One interesting feature of their procreation is that there is a long interval between births; a mother will typically go three years after the birth of a child before having the next one. This was puzzling since abortion or other forms of birth control are not used.



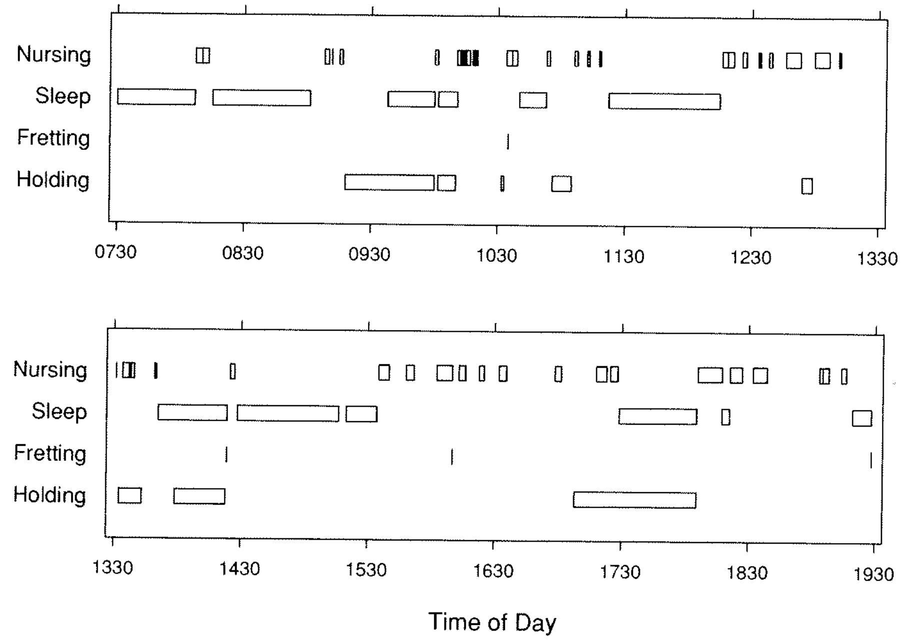
2.3 SUPERFLUITY AND STANDING OUT. The graph shows the activities of a !Kung woman and her baby. The open bars and vertical lines are nursing times; the closed bars show times when the baby is sleeping; F means fretting; and slashed lines are intervals when the baby is held by the mother, with arrows for picking up and setting down. The data do not stand out on this graph.

Two Harvard anthropologists, Melvin Konner and Carol Worthman, put forward a likely solution to the puzzle [74]. They argued that it was the very frequent nursing of infants by their mothers during the first one to two years of life that produces the long inter-birth interval. The nursing results in the secretion of the hormone prolactin into the mother's blood, which in turn reduces the functions of the gonads. This acts as a birth control mechanism.

Konner and Worthman used the graph in Figure 2.3 to show the frequency of nursing and other activities of one !Kung woman and her baby. The open bars and vertical lines are nursing times; the closed bars show times when the baby is sleeping; F means fretting; and slashed lines represent the time held by the mother with arrows for picking up and setting down. A major problem with Figure 2.3 is that the data do not stand out. It is hard to get a visual summary of the extent and

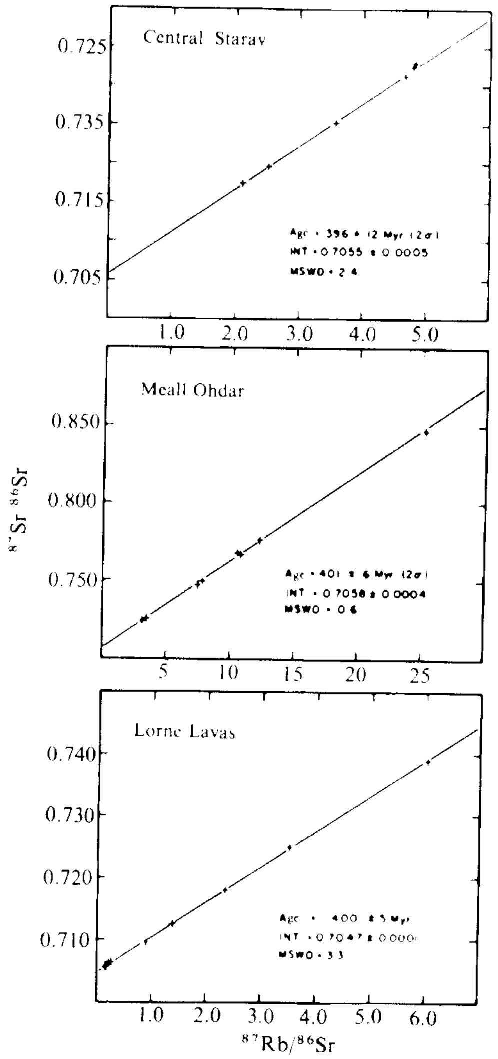
variability of each activity and it is difficult to remember which symbol goes with which activity, so that constant referring to the caption is necessary. A minor problem with Figure 2.3 is that the arrows for picking up and setting down are superfluous.

Figure 2.4 is an improved graph of Figure 2.3. The data stand out and there are no superfluous elements. The constant referring to the caption is not necessary and we get a much better idea of the extent of the activities and their interactions. Figure 2.4 shows clearly the frequency and duration of the nursing bouts for this two-week-old boy. To Western eyes the frequency of the bouts is astonishing. It turns out that this high frequency is needed to make the prolactin birth control mechanism work, since the hormone has a half-life in the blood stream of only 10 to 30 minutes. The figure also shows clearly that nursing and holding infrequently occur together; presumably feeding is done in some prone position.



2.4 SUPERFLUITY AND STANDING OUT. *Make the data stand out. Avoid superfluity.* These are two broad principles that guide the specific principles to follow in this section. The data from Figure 2.3 are regraphed. It is now easier to see the activity times and their interactions, constant referring to the caption is not necessary, and there are no superfluous graphical elements.

The specific principles that follow in this section will allow us to achieve the two general goals of making the data stand out and avoiding superfluity.

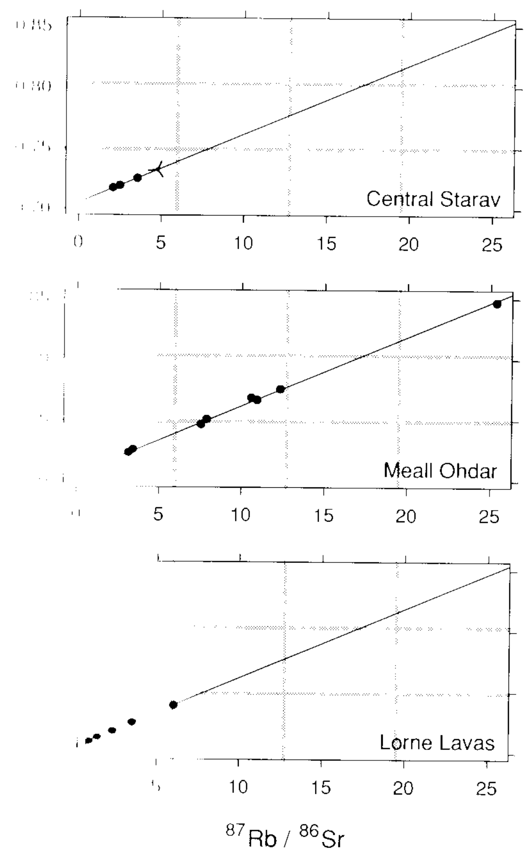


2.5 VISUAL PROMINENCE. The data do not stand out.

Use visually prominent graphical elements to show the data.

On the graph in Figure 2.5 the data do not stand out [20]. The plotting symbols are not visually prominent, and in the bottom panel we cannot tell how many data values make up the black blob in the lower left corner.

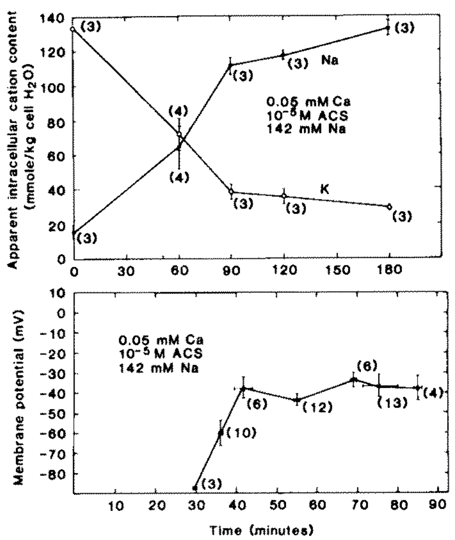
A good way to help the data to stand out is to show them with a graphical element that is visually prominent. This is illustrated in Figure 2.6; the data from Figure 2.5 are regraphed. The symbols showing the data stand out, and now the data can be seen. The symbols that look like the spokes of a wheel represent multiple points; each spoke is one point. For example, the spoked symbol in the Lorne Lavas panel represents four data values.



2.6 VISUAL PROMINENCE. *Use visually prominent graphical elements to show the data.* Now the data from Figure 2.5 can be seen. The symbols that look like the spokes of a wheel represent multiple points; each spoke is one observation.

There are other problems with Figure 2.5 that have been corrected in Figure 2.6. First, in the top panel of Figure 2.5, two tick mark labels, 0.725 and 0.735, have been interchanged. Also, it is hard to compare data on the three graphs in Figure 2.5 because the scales are different; scale issues such as these will be discussed in Section 2.5 (pp. 80–109).

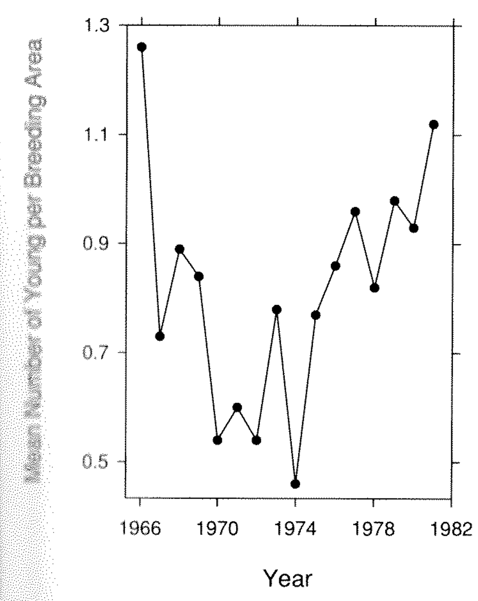
When plotting symbols are connected by lines, the symbols should be prominent enough to prevent being obscured by the lines. In Figure 2.7 the data and their standard errors are inconspicuous, in part because of the connecting lines [10].



2.7 VISUAL PROMINENCE. The data on this graph do not stand out because the graphical elements showing the observations and their standard errors are not prominent enough to prevent being obscured by the connecting lines.

In Figure 2.8 visually prominent filled circles show the data. These large, bold plotting symbols make the data amply visible and ensure that the connecting of one datum to the next by a straight line does not obscure the data. The connection is useful since it helps us to track visually the movement of the values through time.

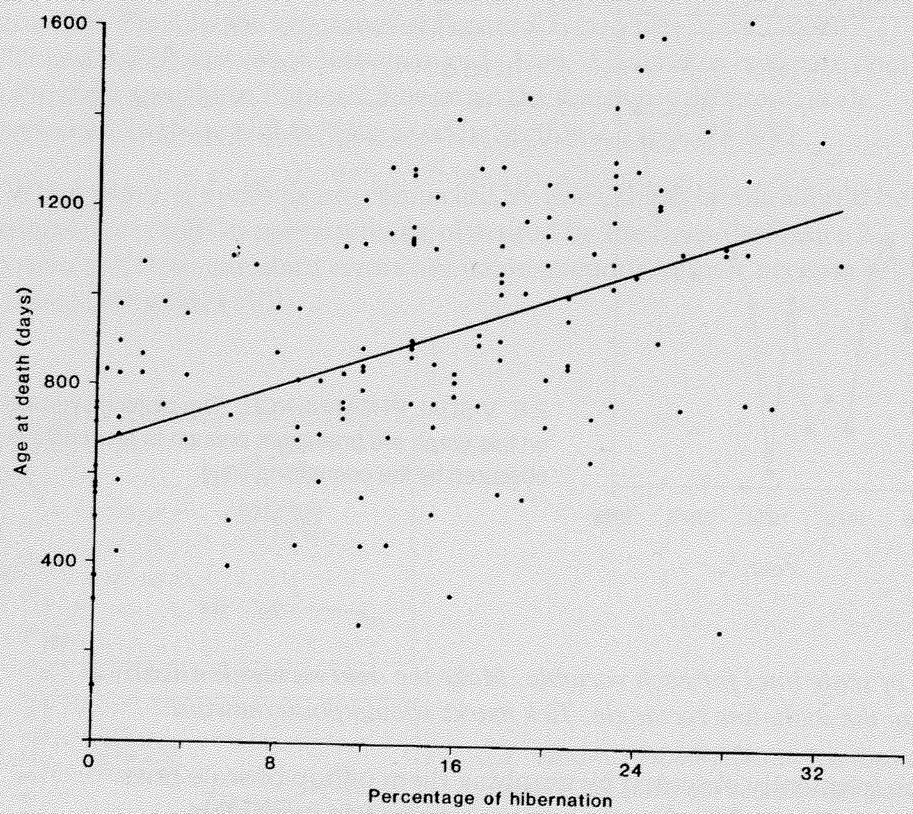
The data in Figure 2.8 are from observations of nesting sites of bald eagles in northwestern Ontario [55]. The graph shows good news: in 1973 DDT was banned, and after the ban, the average number of young per site began increasing.



2.8 VISUAL PROMINENCE. The plotting symbols on this graph are prominent enough to prevent being obscured by the connecting lines.

Use a pair of scale lines for each variable. Make the data rectangle slightly smaller than the scale-line rectangle. Tick marks should point outward.

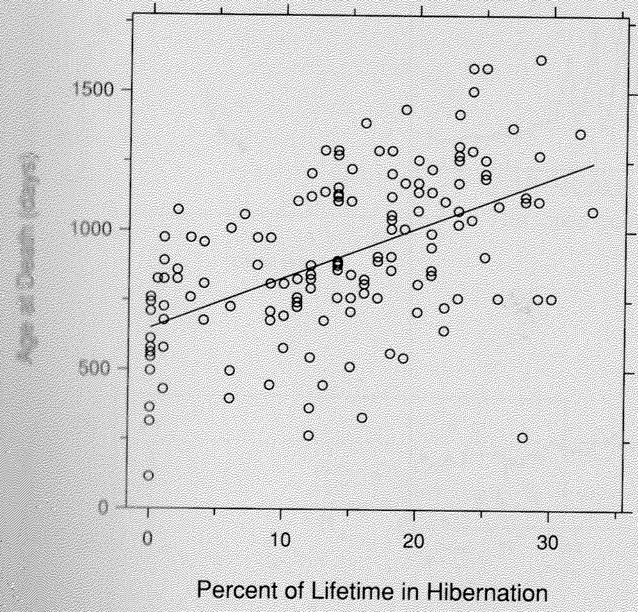
Data are frequently obscured by graphing them on top of scale lines. One example is Figure 2.9 where points are graphed on top of the vertical scale line. The graph and data of Figure 2.9 are from an interesting experiment run by four Harvard anatomists — Charles Lyman, Regina O'Brien, G. Cliett Greene, and Elaine Papafrangos [85]. In the experiment, the researchers observed the lifetimes of 144 Turkish hamsters (*Mesocricetus brandti*) and the percentages of their lifetimes that the hamsters spent hibernating. The goal of the experiment was to determine whether there is an association between the amount of hibernation and the length of life; the hypothesis is that increased hibernation *causes* increased life. Hamsters were chosen for the experiment since they can be raised in the laboratory and since they hibernate for long periods when exposed to the cold. Certain species of bats also hibernate for long periods in the cold but, as the experimenters put it, "their long life-span challenges the middle-aged investigator to see the end of the experiment."



2.9 SCALE LINES AND THE DATA RECTANGLE. The data for zero hibernation are obscured by the left vertical scale line.

The graph in Figure 2.9 suggests that hibernation and lifetime are associated; while this does not *prove* causality it does support the hypothesis. The graph also shows one deviant hamster that spent a large fraction of its life hibernating but nevertheless died at a young age. Hibernation cannot save a hamster from all of the perils of life.

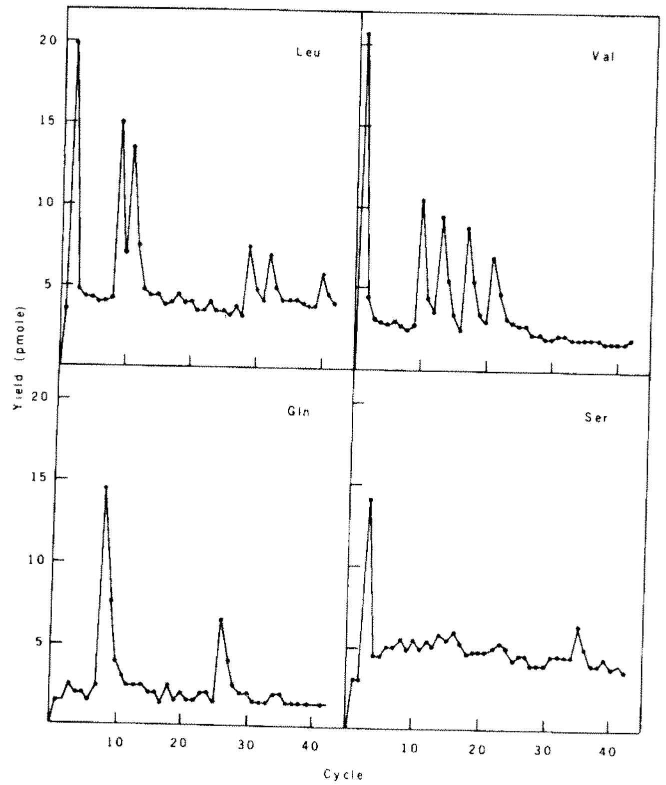
One unfortunate aspect of Figure 2.9 is that the data for hamsters with zero hibernation are graphed on top of the vertical scale line. This obscures the data to the point where it is hard to perceive just how many points there are. No data should be so obscured. One way to avoid this is shown in Figure 2.10. The data rectangle is slightly smaller than the scale-line rectangle. Now the values with zero hibernation can be seen clearly.



2.10 SCALE LINES AND THE DATA RECTANGLE. *Use a pair of scale lines for each variable. Make the data rectangle slightly smaller than the scale-line rectangle. Tick marks should point outward.* This format prevents data from being obscured. Using two scale lines for each of the two variables on this graph, instead of just one, allows easier table look-up of the scale values of data at the top or right of the data rectangle.

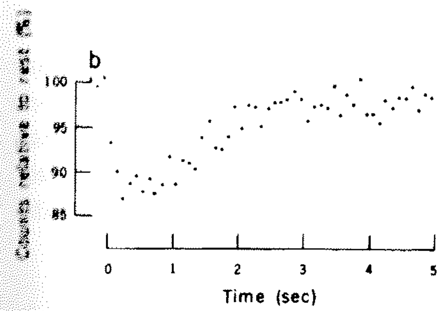
Four scale lines are used in Figure 2.10 rather than the two of Figure 2.9. Table look-up — judging the scale value of a point by judging its position along a scale line — is easier and more accurate as the distance of the point from the scale line decreases. The consequence of one vertical scale line on the left is that the vertical scale values of data to the right are harder to look up than those of data to the left because the rightmost values are further from the line; similarly, when there is just one horizontal scale line, the horizontal scale values of data at the top are harder to look up than those at the bottom. By using four scale lines, the graph treats the data in a more nearly equitable fashion.

Ticks point outward in Figure 2.10 because ticks that point inward can obscure data, as is illustrated in the upper panels of Figure 2.11 [62].

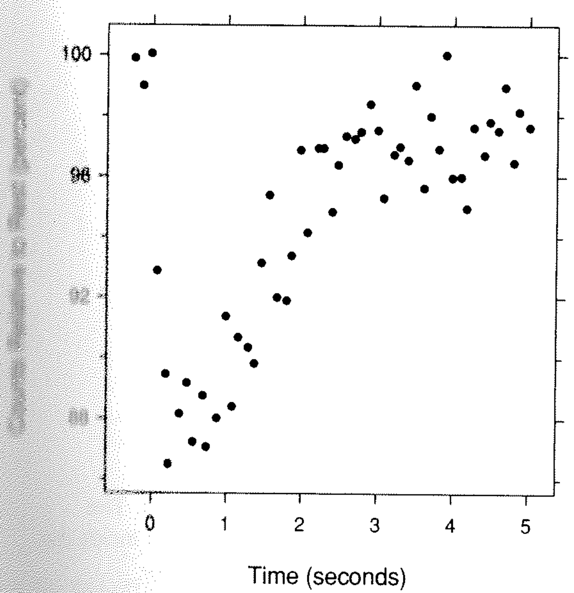


2.11 SCALE LINES AND THE DATA RECTANGLE. Tick marks that point inward can obscure data.

The four scale lines also provide a clearly defined region where our eyes can search for data. With just two, data can be camouflaged by virtue of where they lie. This is true for the data in Figure 2.12 [133]; it is easy to overlook the three points hidden in the upper left corner. In Figure 2.13 the graph has four scale lines and the three points are more prominent.



2.12 SCALE LINES AND THE DATA RECTANGLE. The three points in the upper left are camouflaged.

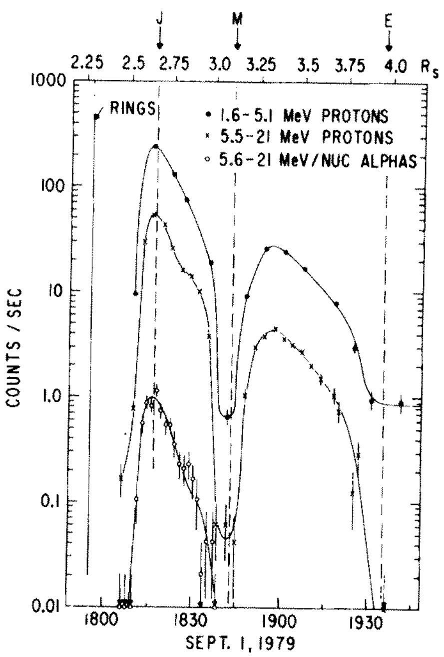


2.13 SCALE LINES AND THE DATA RECTANGLE. The four scale lines provide a clearly defined region for our eyes to look for data. Now, none of the data from Figure 2.12 are in danger of being overlooked.

Do not clutter the interior of the scale-line rectangle.

Another way to obscure data is to graph too much. It is always tempting to show everything that comes to mind on a single graph, but graphing too much can result in less being seen and understood. This is illustrated in Figure 2.14 [118]. The data are particle counts from an exciting scientific exploration: the passage of the Pioneer II spacecraft by Saturn. In the interior of the scale-line rectangle we have reference lines, a label, arrows, a key, symbols showing the data, tick marks, error bars, and smooth curves. The graph is cluttered, with the result that it is hard to visually disentangle what is graphed. It is unfortunate to have any of these valuable data obscured.

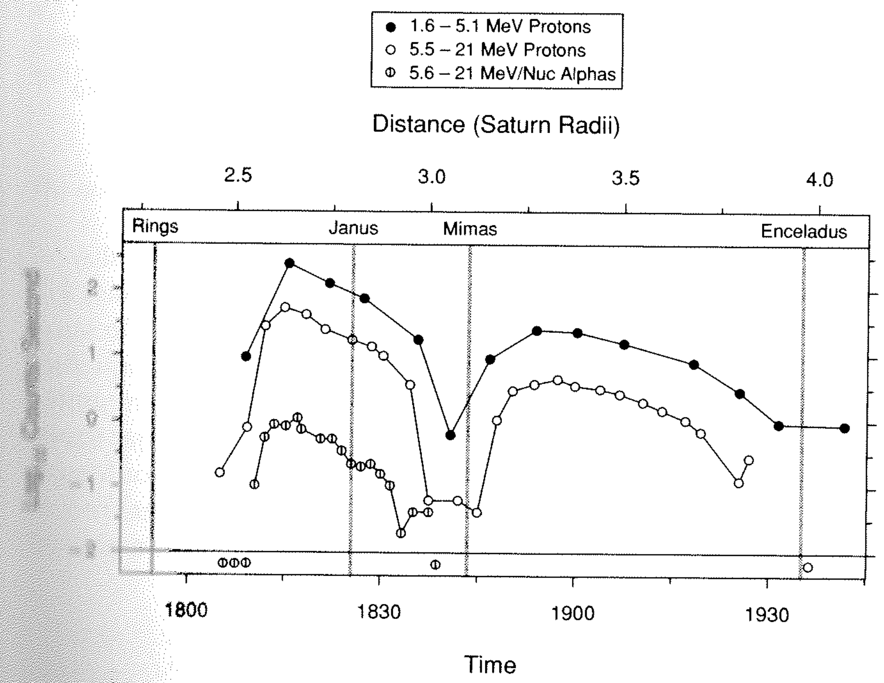
The data are shown again in Figure 2.15. The clutter has been alleviated, in part, by removing the error bars. It would be prudent to convey accuracy for these data numerically rather than graphically; on a log scale the error bars decrease radically and disappear from sight as



2.14 CLUTTER. This graph is cluttered. The result is that different graphical elements inside the scale-line rectangle obscure one another.

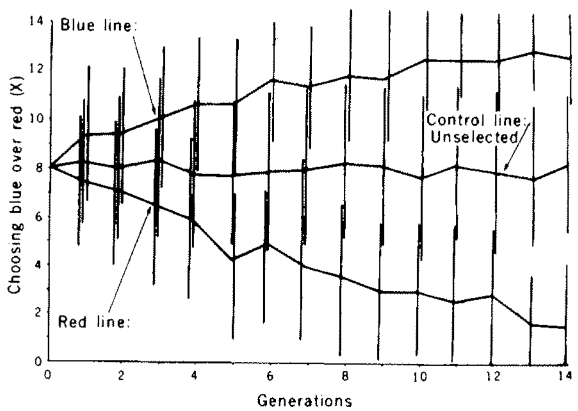
the counts increase. (It is possible that accuracy is nearly constant on a scale of square root counts/sec since count data of this sort tend to have a Poisson distribution. Thus accuracy might be conveyed more readily on the square root scale rather than on the log scale.) Other removals have taken place. The plethora of tick marks on the vertical scale has been reduced, as well as the number of tick mark labels on the top horizontal scale line. Also, the top horizontal scale line is labeled in Figure 2.15, but not in Figure 2.14.

The clutter also has been reduced by some alterations. The key and the label for rings are outside of the scale-line rectangle, the arrows showing values below 0.01 counts/sec have been replaced by a separate panel, and the wandering curves have been replaced by straight lines connecting successive data points. These changes have reduced interference between different elements of the graph.

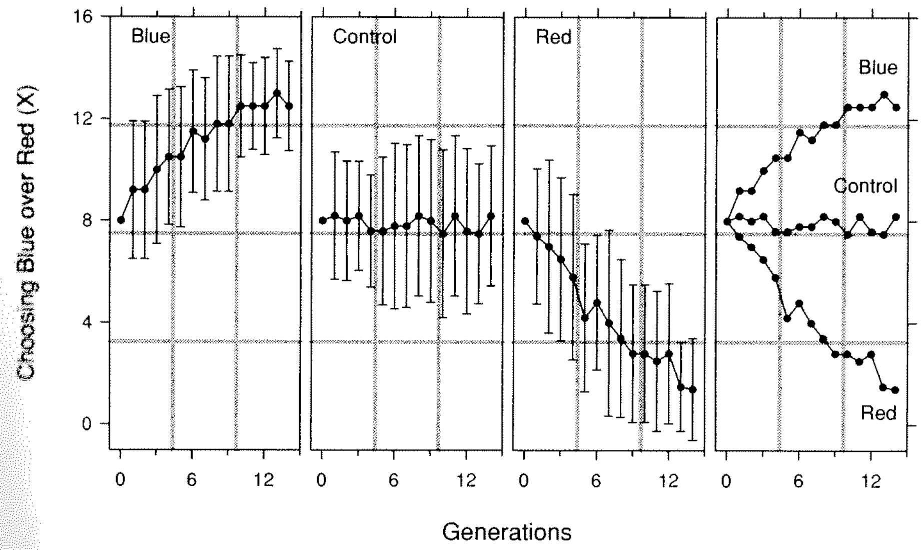


2.15 CLUTTER. *Do not clutter the interior of the scale-line rectangle.* The clutter of Figure 2.14 has been removed by alteration and excision. For example, the number of tick marks has been reduced.

Figure 2.16 [76] is also cluttered; the error bars interfere with one another so much that it is hard to see the values they portray. One solution is shown in Figure 2.17. In the left three panels the three data sets are juxtaposed and in the right panel they are superposed, but without the error bars. The juxtaposition allows us to see clearly each set of data and its error bars; the superposition allows us to compare the three sets of data more effectively.



2.16 CLUTTER. This graph is also cluttered.



2.17 CLUTTER. The clutter of Figure 2.16 has been eliminated by graphing the data on juxtaposed panels. The right panel is included so that the values of the three data sets can be more effectively compared.

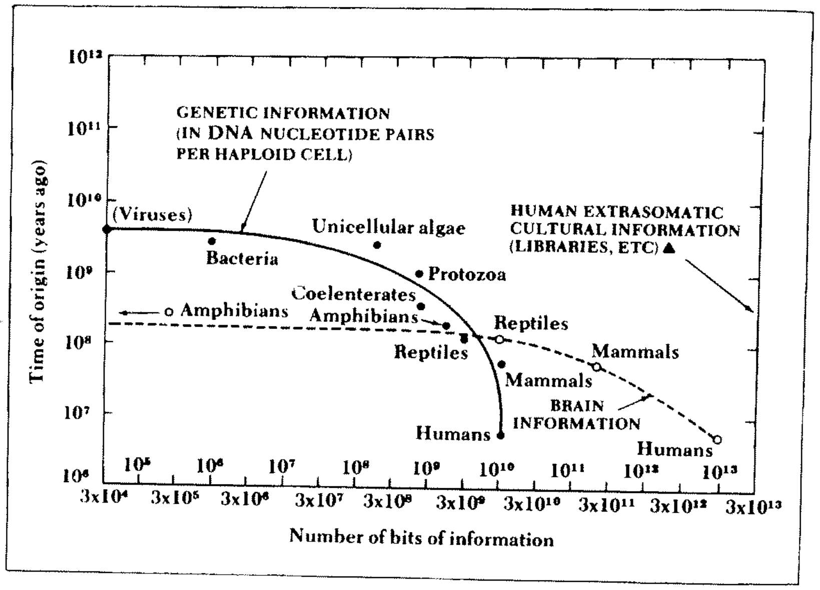
Do not overdo the number of tick marks.

A large number of tick marks is usually superfluous. From 3 to 10 tick marks are generally sufficient; this is just enough to give a broad sense of the measurement scale and to enable sufficiently accurate table look-up. Copious tick marks date back to a time when data were communicated by graphs. Today, we have electronic communication. Every aspect of a graph should serve an important purpose. Any superfluous aspects, such as unneeded tick marks, should be eliminated to decrease visual clutter and thus increase the visual prominence of the most important element — the data.

Figure 2.18, from Carl Sagan's book, *The Dragons of Eden* [107], has too many tick marks. The filled circles show the number of bits of information (horizontal scale) in the DNA of various species when they emerged and the time of their emergence (vertical scale). The open circles show, in the same way, the bits of information in the brains of various species. On a first look at this graph, the bottom scale line makes it easy to think there are two horizontal scales. This is not so. The labels of the form 3×10^k are showing, approximately, the values of the midpoints of the numbers of the form 10^k. For example, midway between 10^7 and 10^8 on a log scale is

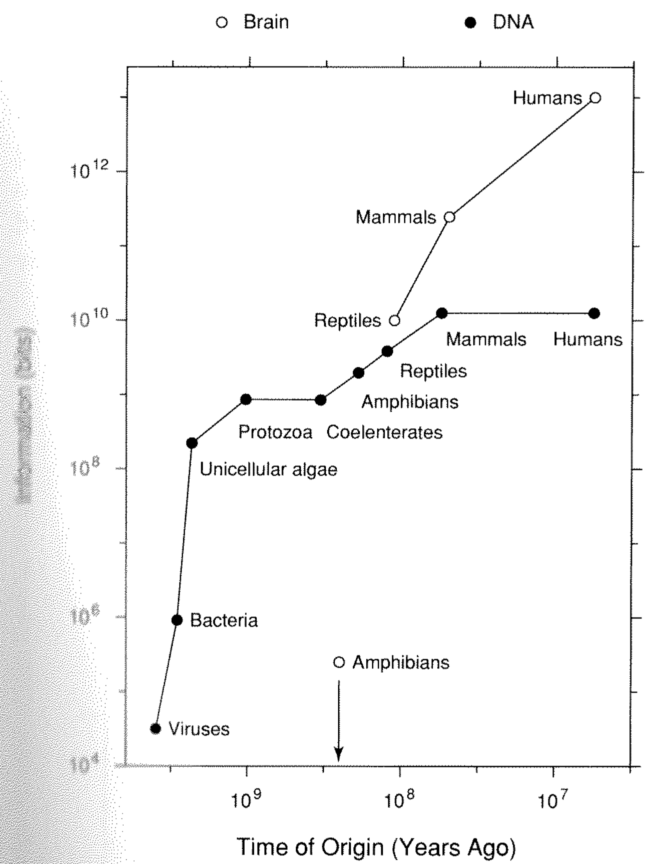
$$10^{7.5} = 10^{0.5}\, 10^7 \approx 3 \times 10^7 .$$

The large number of tick marks and labels needlessly clutters the graph, and the approximation can easily lead to confusion.



2.18 TICK MARKS. There are too many tick marks and tick mark labels on this graph. The tick mark labels on the horizontal scale are confusing.

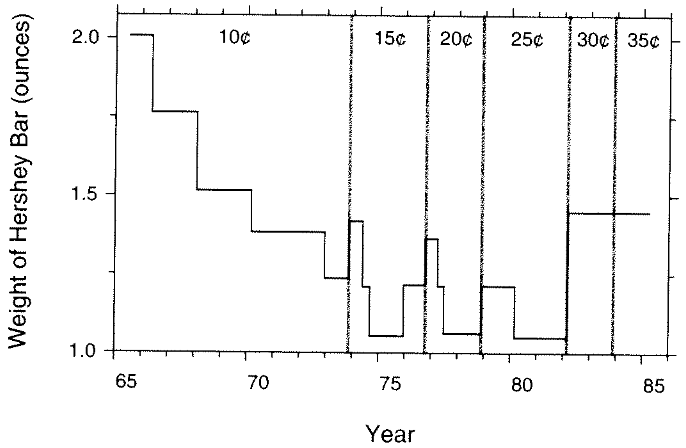
In Figure 2.19 the brain and DNA data are graphed again with fewer tick marks and labels; the horizontal and vertical scales have been interchanged so that time is now on the horizontal scale with earlier times on the left and later times on the right.



2.19 TICK MARKS. *Do not overdo the number of tick marks.* The vertical scale of this graph, previously the horizontal scale of Figure 2.18, has a sensible number of tick marks and labels.

Use a reference line when there is an important value that must be seen across the entire graph, but do not let the line interfere with the data.

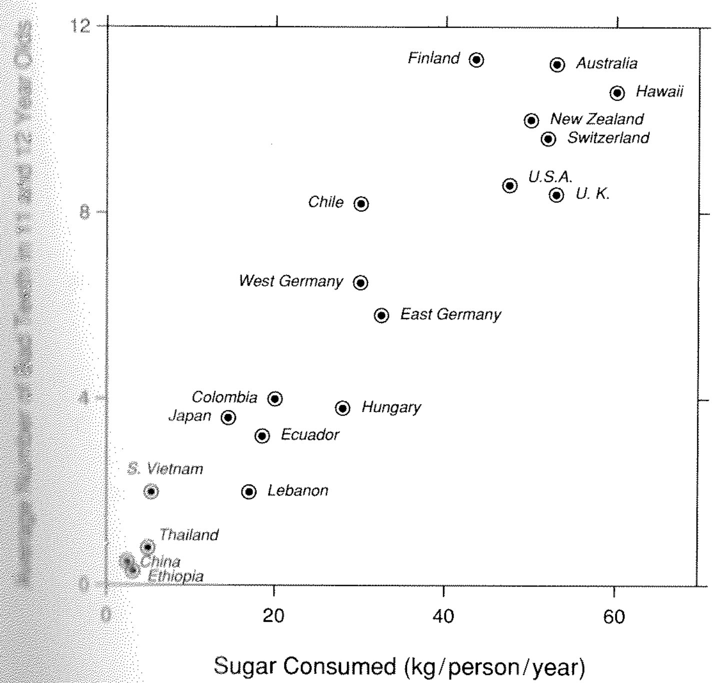
Reference lines are used in Figure 2.20. The data are the weights of the Hershey Bar, the famous American candy bar. These data, and Stephen Jay Gould's analysis of them [53], are discussed in detail in Section 3.8 (pp. 180–192). The vertical reference lines, which show times of price increases, cross the entire graph and let us see what happened to weight exactly at the times of the price increases. Except for the change from 30 cents to 35 cents, all price increases were accompanied by a size increase.



2.20 REFERENCE LINES. *Use a reference line when there is an important value that must be seen across the entire graph, but do not let the line interfere with the data.* The weight of the Hershey Bar is graphed against time. The vertical reference lines divide time up into price epochs; prices are shown just below the top vertical scale. The precision of the reference lines is needed to show us exactly where the price increases occur.

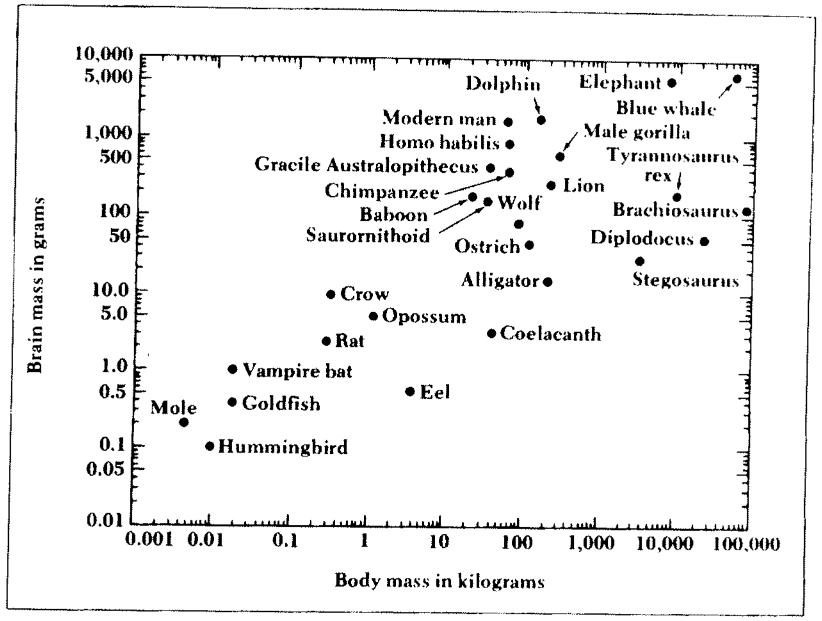
Do not allow data labels in the interior of the scale-line rectangle to interfere with the quantitative data or to clutter the graph.

Figure 2.21 shows the relationship between the average number of bad teeth in 11 and 12 year old children and the per capita sugar consumption per year for 18 countries and the state of Hawaii [97]. When it is important to convey the names for the individual values of a data set, data labels inside of the scale-line rectangle are generally unavoidable. In so doing we should attempt to reduce the visual prominence of the labels so that they interfere as little as possible with our ability to assess the overall pattern of the quantitative data. This has been done in Figure 2.21 by choosing a plotting symbol that is visually very different from the letters of the labels.



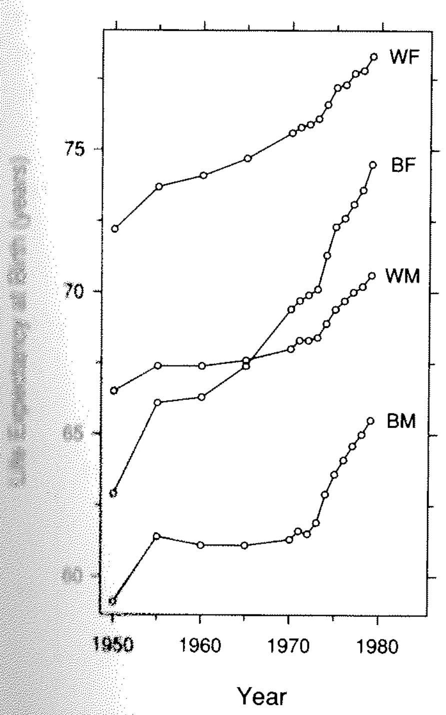
2.21 DATA LABELS. *Do not allow data labels in the interior of the scale-line rectangle to interfere with the quantitative data or to clutter the graph.* The data labels on this graph are needed to convey the names. The visual impact of the labels has been lessened so that they interfere as little as possible with our visual assembly of the plotting symbols.

In Figure 2.22 [107], discussed in Section 1.3 (pp. 16–21), the plotting symbols are not sufficiently visually distinguishable from the labels. The result is that the point cloud is camouflaged by the labels.



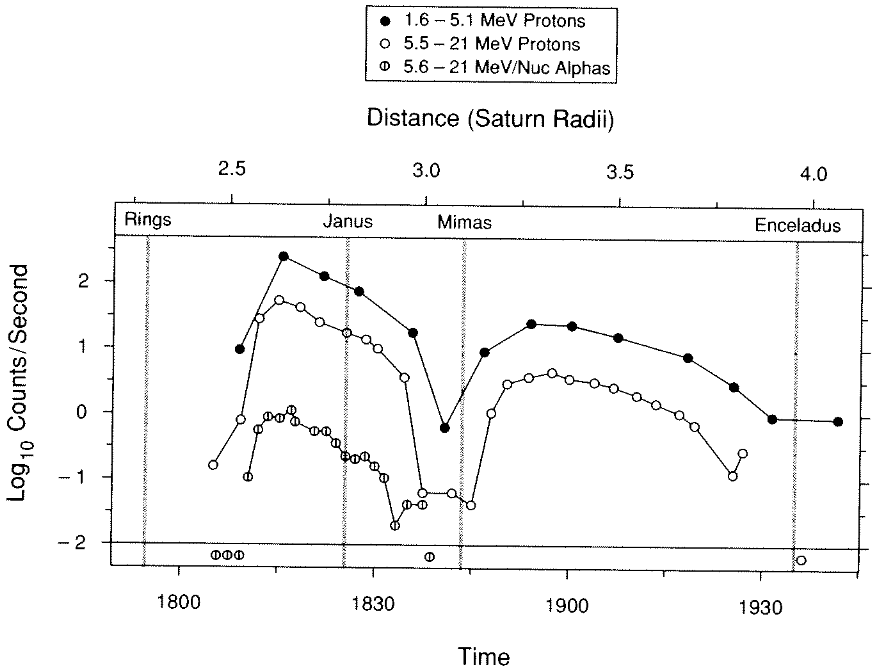
2.22 DATA LABELS. The data labels interfere with our visual assembly of the plotting symbols.

Figures 2.21 and 2.22 show one type of data label; each value in the data set has its own name. Sometimes the quantitative information on a graph consists of different data sets where each data set has a name that we want to convey. This is illustrated in Figure 2.23, which shows life expectancies for four groups of people: black females, black males, white females and white males [127]. Four data labels inside the scale-line rectangle convey the data set names without obscuring the data or causing clutter.



2.23 DATA LABELS. Groups of data values often can be identified by data labels inside the scale-line rectangle. The labels are abbreviations in which B = black, W = white, M = male, and F = female.

Sometimes a key is needed to identify data sets, either because data labels inside the scale-line rectangle would add too much clutter or because the values for each data set cannot be identified without using different plotting symbols for the different data sets. A key is used in Figure 2.24 for both reasons. On this graph the data labels are long and the data rectangle is already host to many things. Furthermore, a key is needed because there is no other convenient way to allow identification of the values below $-2 \log_{10}$ (counts/sec), which are shown at the bottom of the graph.

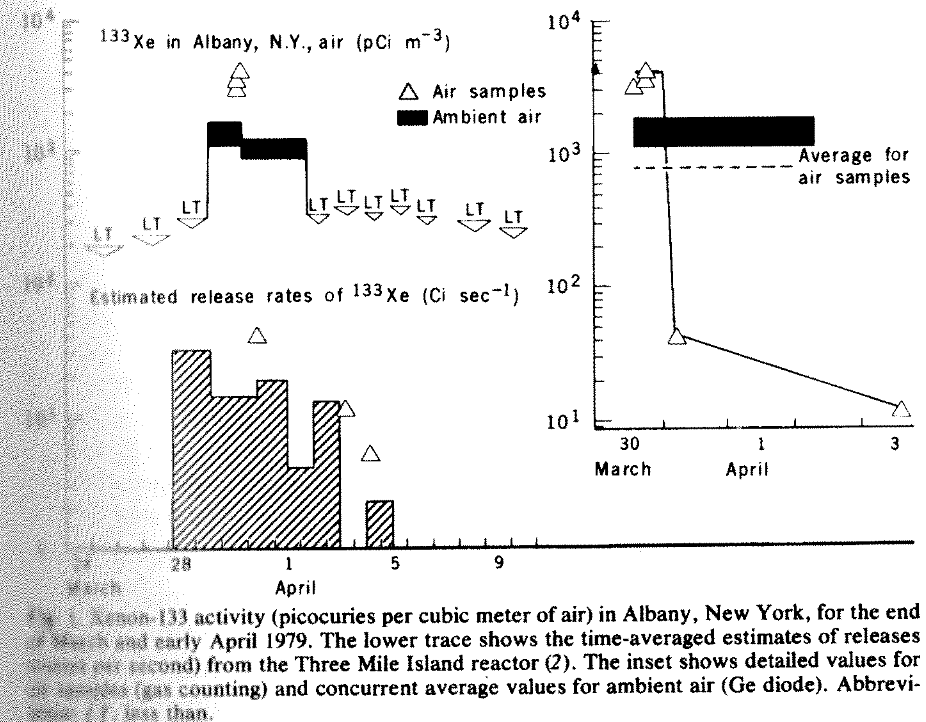


2.24 DATA LABELS. Groups of data values also can be identified by a key. One disadvantage, compared with data labels inside the scale-line rectangle, is that identification is slightly harder because we must look back and forth between the key and the data. However, one advantage over data labels inside, an important one in this example, is that clutter is reduced.

Avoid putting notes and keys inside the scale-line rectangle. Put a key outside, and put notes in the caption or in the text.

We should approach the interior of the scale-line rectangle with a strong spirit of minimalism and try to keep as much out as possible. Not doing so can jeopardize our relentless pursuit of making the data stand out. There is no reason why keys and notes need to appear in the interior.

Keys can go outside of the scale-line rectangle and notes can go in the text or the caption. This has not been done in Figure 2.25 [130] and the result is needless clutter and a confusing graph. The main graph shows release rates of xenon-133 from the Three Mile Island nuclear reactor accident and concentrations of xenon in the air of Albany, N.Y. during



2.25 NOTES AND KEYS. Everything — including the scale labels, a key, and "LT" (meaning less than) — has been thrown into the interior of the scale-line rectangle of this graph. The result is confusing.

the same time period. The purpose of the graph is to show that in Albany, about 500 km from Three Mile Island and downwind during the period of the accident, xenon concentrations rose after the accident.

Figure 2.25 has a number of problems arising from some unusual and unexplained conventions and from putting too much inside the scale-line rectangle. The writing inside is really two scale labels, complete with units. The top label describes two types of Albany air concentration measurements. The bottom label describes the Three Mile Island release rates. Part of the difficulty in comprehending this graph is that three Albany air samples are below the label for release rates, which gives an initial incorrect impression that they are air samples measuring the release rates. The ambient air measurements are shown in a somewhat unconventional way. The two solid rectangles are averages over two intervals; the width shows the averaging interval and a good guess is that the height, which is not explained, shows an average ± 2 sample standard deviations. The triangles with "LT" above them indicate other ambient air measurements which are "less than" the values indicated. The inset has very little additional information; it shows two averages and repeats 5 of the air sample measurements. There is an inaccuracy somewhere; for the three largest air sample values, the times shown on the inset do not agree with the times shown on the main graph. The two averages in the inset do not convey any important information.

These data deserve two panels and deserve less inside the scale-line rectangle to make completely clear what has been graphed. This has been done in Figure 2.26; the writing, key, and LT's have been removed from the scale-line rectangle and the inset has been deleted. The bottom panel shows the release rates of xenon from Three Mile Island; the horizontal line segments show averages over various time intervals. The top panel shows the Albany measurements; the horizontal line segments show intervals over which some measurements were averaged, the error bars show plus and minus two sample standard deviations (if the guess about Figure 2.25 was correct), and an arrow indicates the actual value was less than or equal to the graphed value. Furthermore, the labels for the two types of measurements have been corrected. Both are ambient air measurements and both are from air samples. The terms "continuous monitor" and "grab samples" correctly convey the nature of the two types.

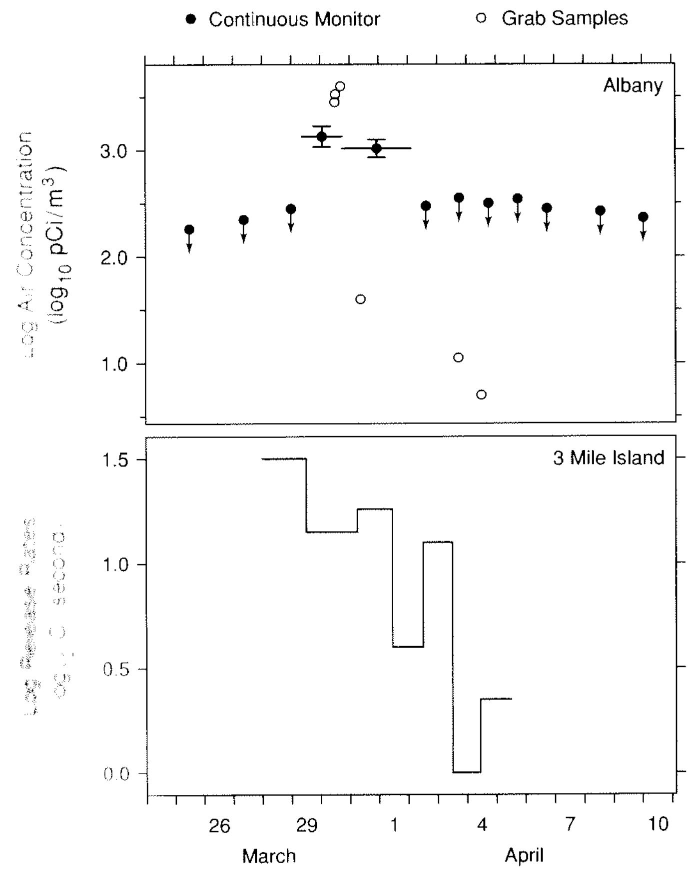
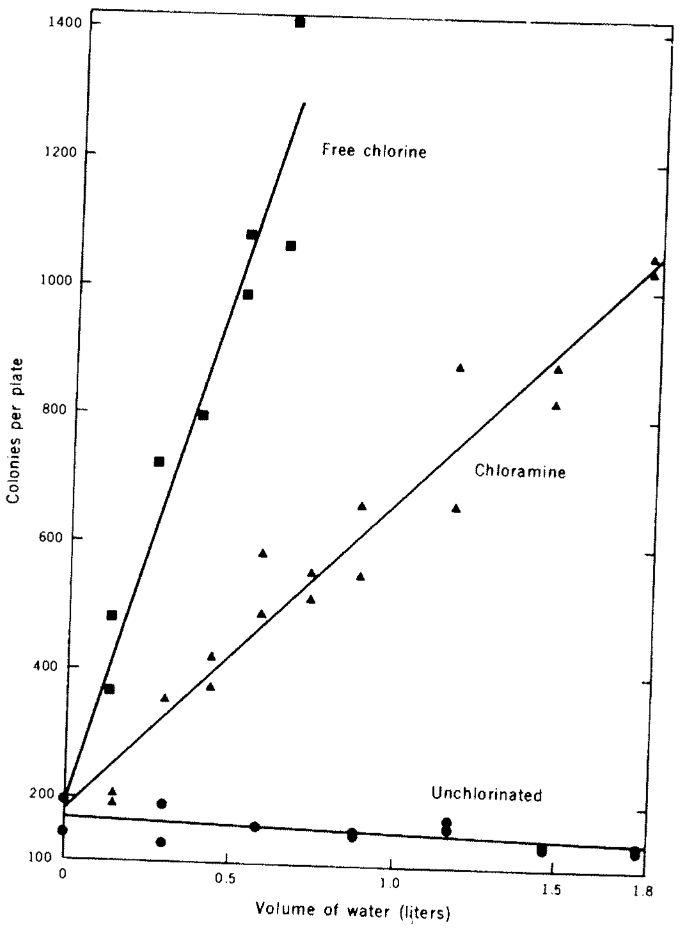


Figure 2.26 NOTES AND KEYS. *Avoid putting notes and keys inside the scale-line rectangle. Put a key outside, and put notes in the caption or in the text.* The graph in Figure 2.25 has been improved by the following actions: removing the writing and the key from the interior of the scale-line rectangle; removing the inset altogether; showing the two data sets on separate panels; removing the idiosyncrasies; and correcting the labels describing the two types of measurements.

Overlapping plotting symbols must be visually distinguishable.

Unless special care is taken, overlapping plotting symbols can make it impossible to distinguish individual data points. This happens in several places in Figure 2.27 [18]. The data are from an experiment on the production of mutagens in drinking water. For each category of observation (free chlorine, chloramine, and unchlorinated) there are two observations for each value of water volume. That is, duplicate measurements were made. But two values do not always appear because of exact or near overlap. For example, for the unchlorinated data only one observation appears for water volume just above 0.5 liters.

This problem of visual clarity is a surprisingly tough one. Several solutions are given in Section 3.5 (pp. 154–165).

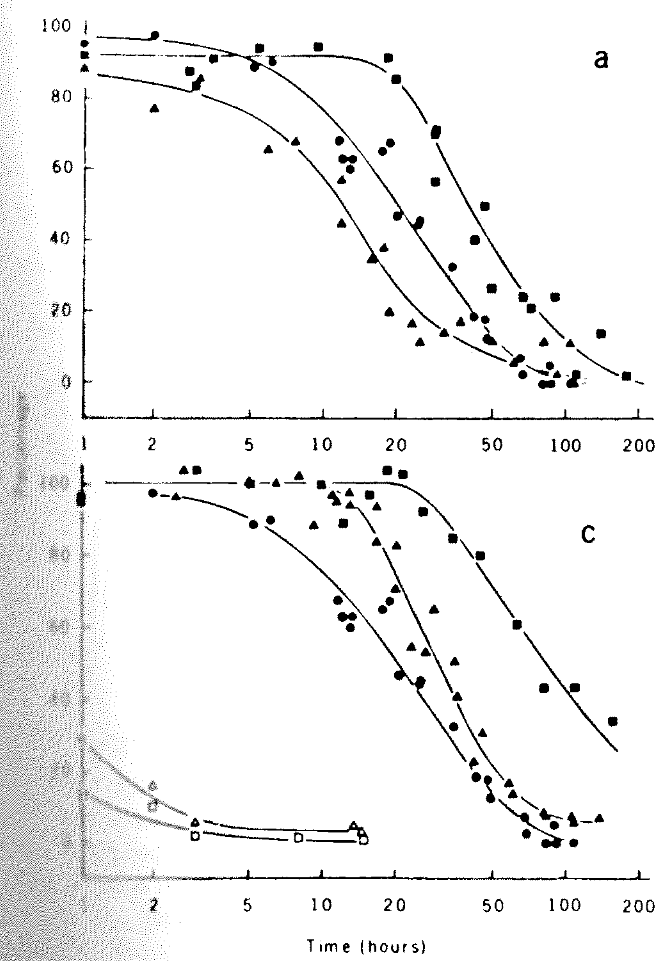


2.27 OVERLAPPING PLOTTING SYMBOLS. *Overlapping plotting symbols must be visually distinguishable.* On this graph, because of exact and near overlap, some of the data cannot be seen.

Superposed data sets must be readily visually assembled.

It is very common for graphs to have two or more data sets superposed within the same data rectangle. We already have encountered many such graphs in this book. Special methods are often required to ensure good visual assembly of each of the different data sets.

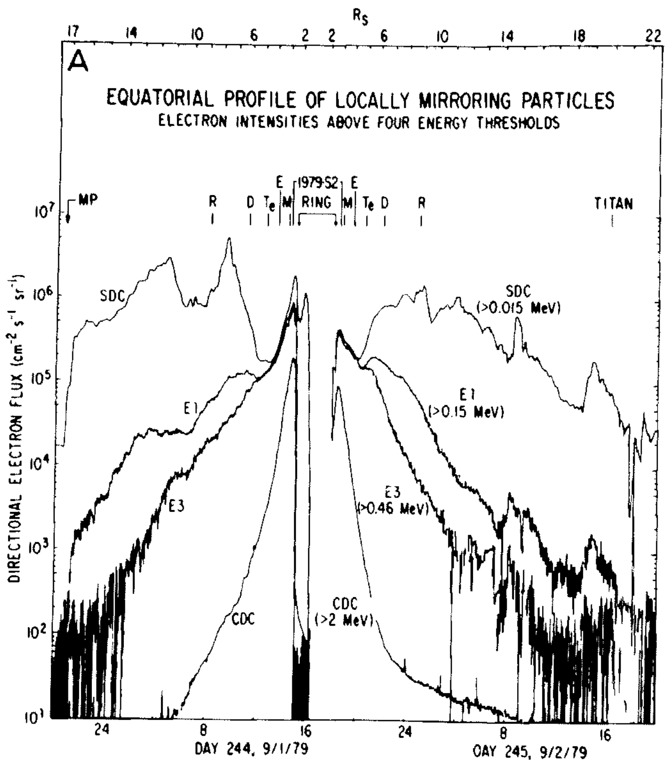
In Figure 2.28 [91] it is difficult to visually disentangle the solid squares, circles, and triangles; such plotting symbols are in general visually similar, but in Figure 2.28 the problem is exacerbated by the symbols not being crisply drawn.



2.28 SUPERPOSED DATA SETS. *Superposed data sets must be readily visually assembled.* On this graph we cannot easily visually assemble the circles as a group, or the squares, or the triangles.

In Figure 2.29 [49] the different curves are hard to disentangle in many places and impossible in others. For example, on the left of the graph between 8 and 16 hours, curves E1 and E3 merge and then join CDC in a triple junction; a little later one curve splits off, but it is impossible to tell which it is. More copious labeling might help but it still would require a concentrated and highly cognitive mental effort to follow each curve visually, rather than the rapid, easy assembly that we should strive for when data sets are superposed. We do not want to have to visually follow a curve on a graph the way we have to visually follow a twisting secondary road on a detailed map; rather, we want to be able to visually assemble a single curve as a whole, mentally filtering out the other curves.

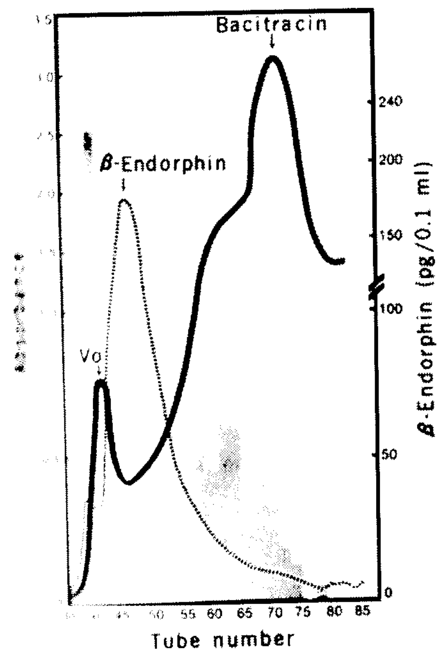
Graphs that fail to allow effective visual assembly are pervasive because the problem is a difficult one to solve. Solutions will be given in Section 3.5 (pp. 154–165) and Section 3.13 (pp. 209–212).



2.29 SUPERPOSED DATA SETS. The curves on this graph merge, in going from left to right, and then separate with their identities lost.

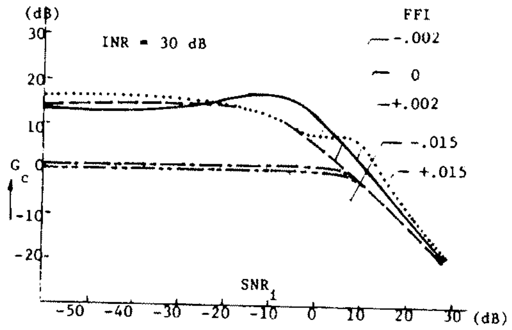
Visual clarity must be preserved under reduction and reproduction.

Graphs that communicate data to others often must undergo reduction and reproduction; these processes, if not done with care, can interfere with visual clarity. In Figure 2.30 [60] the ghostly image in the background should be a shaded area representing immunoreactivity, but the shading is barely visible due to poor reproduction. Figure 2.30 has other problems. The scales are poorly constructed. The right vertical scale shows a break; in fact it is not a break in the usual sense of a gap in the scale, but rather the number of units per cm suddenly changes. The same type of change occurs on the left vertical scale, but the authors have chosen not to flag this one. The graphed data move through the data rectangle as if nothing is happening to the scales.



2.30 REDUCTION AND REPRODUCTION. *Visual clarity must be preserved under reduction and reproduction.* This did not happen on this graph. The ghostly image in the background was supposed to represent immunoreactivity.

In Figure 2.31 [94] the lines that are supposed to connect the labels with the curves are washed out. Lines, curves, and lettering must be heavy enough and symbols must be large enough to withstand reduction and reproduction.



2.31 REDUCTION AND REPRODUCTION. The lines from the curves to their labels are washed out.

2.3 Clear Understanding

Graphs are powerful tools for communicating quantitative information in written documents. The principles of this section, which are oriented toward the task of communication, contribute to a clear understanding of what is graphed.

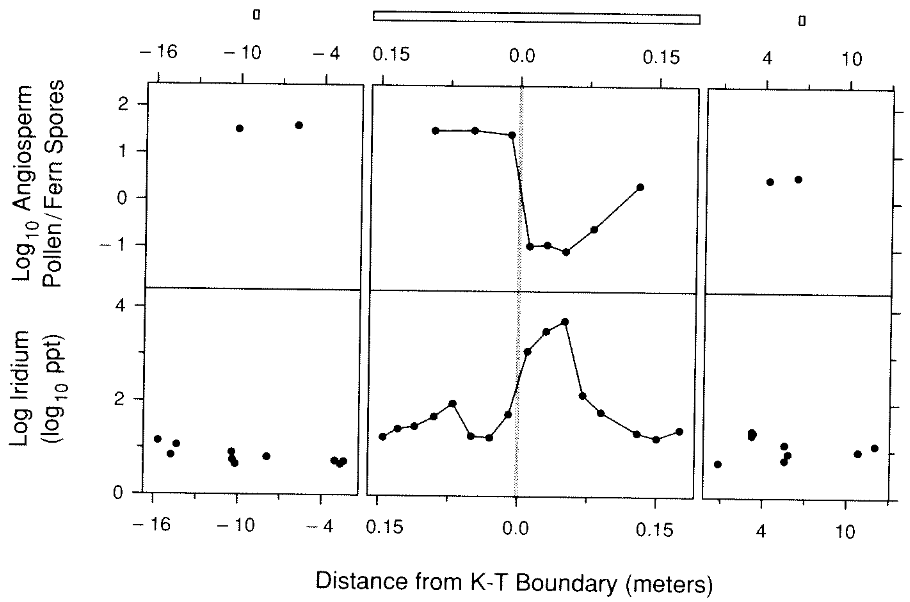
Put major conclusions into graphical form. Make captions comprehensive and informative.

Communication of the results of technical studies, when the results involve quantitative issues, can be greatly enhanced by visual displays that speak to the essence of the results. Graphs and their captions can incisively communicate important data and important conclusions drawn from the data. One good approach is to make the sequence of graphs and their captions as nearly independent as possible and to have them summarize evidence and conclusions. This book has been constructed in this way; the graphs and their captions summarize the ideas, and the text has been written around the sequence of graphs. This is to be expected of a book on graphs, but it is also an effective device for other writings in science and technology.

For a graph to be understood clearly, there must be a clear, direct explanation of the data that are graphed and of the inferences drawn from the data. Here is a framework for figure *captions* that can contribute to such a clear explanation:

1. Describe everything that is graphed.
2. Draw attention to the important features of the data.
3. Describe the conclusions that are drawn from the data on the graph.

The framework is illustrated in the caption of Figure 2.32. The data are involved in an astounding discovery that sounds more like science fiction than a highly supportable scientific hypothesis. Sixty-five million years ago extraordinary mass extinctions of a wide variety of animal species occurred, marking the end of the Cretaceous period and the beginning of the Tertiary. The dinosaurs died out along with the marine reptiles and the flying reptiles such as the ichthysaur. Many marine invertebrates also became extinct; ocean plankton almost disappeared completely.



2.32 EXPLANATION. *Put major conclusions into graphical form. Make captions comprehensive and informative.* Describe everything that is graphed and convey the conclusion drawn from the data. The following is a caption, including the title, that might accompany this graph in its original subject matter context: ANGIOSPERM-FERN RATIO AND IRIDIUM NEAR THE *K–T* BOUNDARY. The graph shows measurements of a core from northeastern New Mexico. The horizontal scale is in meters from the boundary between the Cretaceous and the Tertiary periods; negative values are below the *K–T* boundary so time goes from earlier to later in going from left to right. The widths of the three rectangles at the top of the graph show the same number of meters on the horizontal scales of the three panels. The top panel shows the ratio of angiosperm pollen to fern spores; the *K–T* boundary is taken to be the time point at which these values begin to decrease. The bottom panel shows concentrations of iridium; the concentrations begin a dramatic rise and fall at the boundary. Since the principal source of iridium is extraterrestrial, its rise and fall supports the hypothesis that an asteroid struck the earth causing a cloud of dust in the upper atmosphere; this is argued to have caused the large number of extinctions, including the dinosaurs, that occurred at the beginning of the Tertiary period.

What could have caused such a calamity? Luis Alvarez, Walter Alvarez, Frank Asaro, and Helen Michel at Berkeley made a fortuitous discovery that suggested a cause. They found unusually high levels of iridium right at the K-T (Cretaceous-Tertiary) boundary in sediments from Italy, Denmark, and New Zealand [2]. It is likely that the high iridium levels have an extraterrestrial cause; asteroids and meteors are rich in iridium while the earth's crust is not because this heavy element sank to the core during the earth's molten years. From these data and other information, the four hypothesized that an asteroid, 10 ± 4 km in diameter, struck the earth and sent a dust cloud into the atmosphere that blocked sunlight for a period of several months or even years. The loss of light interfered with food chains and led to the mass extinctions. As the dust from the asteroid settled it deposited an iridium-rich layer on the surface of the earth.

The asteroid hypothesis has been supported by subsequent measurements. Among them are measurements of pollen, fern spores, and iridium in New Mexico [100]. These are the data shown in Figure 2.32. The horizontal scale is distance in the sediment from the K-T boundary. Distance, of course, is just a surrogate for time, which goes from earlier to later as we go from left to right. The top panel graphs the logarithm of the ratio of pollen to fern spores. The point at which the ratio begins to decrease is taken to be the K-T boundary because at the beginning of the Tertiary period angiosperms declined relative to ferns. Just after this boundary there is a peak in the iridium concentrations, shown in the bottom panel.

The caption of Figure 2.32 follows the three-step guidelines presented earlier. The graph and its caption can nearly stand alone as a document that conveys the basic idea of the asteroid-impact hypothesis and the quantitative information that gives it credence.

The interplay between graph, caption, and text is a delicate one that requires substantial judgment. No complete prescription can be designed to allow us to proceed mechanically and to relieve us of thinking hard. However, a viewer is usually well served by a caption that makes a graph as self-contained as possible. If there are several graphs, the captions collectively can be an independent piece; for example, a detailed description of a data set described in one graph caption does not need to be repeated in a subsequent graph caption.

It is possible, though, to overdo a comprehensive caption. Putting a description of the experimental procedure in the caption seems to go too far. It burdens the graph and makes what should be a concise summary into a tome. Figure 2.33 [133] is an example. The ratio of caption area to graph area is 2.8; this is too much detail. The details of an experimental procedure must be communicated, but surely there is a better place than a figure caption, which is a summary.

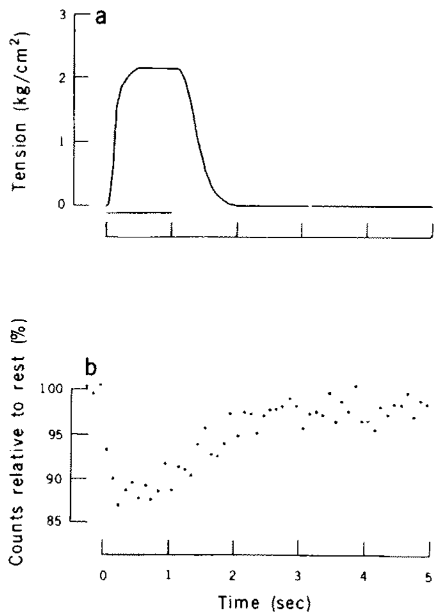


Fig. 2. Tension and the intensity of the 42.9-nm layer line during 1-second tetanus at the sarcomere length of 2.2 μm. (a) Tension record averaged over the 40 tetanic contractions required for obtaining the time course of the layer-line intensity. A sartorius muscle was dissected from *Rana catesbeiana* and tetanized for 1 second at 2-minute intervals. The horizontal line represents the period of stimulation. Tension was recorded with an isometric tension transducer (Shinkoh, type UL). (b) Intensity of the first-order myosin layer line at 42.9 nm. The x-ray source was a rotating-anode generator (Rigaku FR) with a fine focus (1.0 by 0.1 mm) on a copper target. This was operated at 50 kV with a tube current of 70 mA; such a high power was possible with an anode of a large diameter (30 cm) rotating at a high speed (9000 rev/min). A bent-crystal monochromator was used at a source-to-crystal distance of 25 cm with a viewing angle of 6°. The intensity of the myosin layer line was measured with a scintillation counter combined with a mask; the mask had apertures at the positions of the off-meridional parts of the first-order layer line. The meridional reflection at 14.3 nm is known to be slightly displaced during contraction, suggesting a minute change in the myosin periodicity (*1*, *3*). It is, therefore, possible that the 42.9-nm layer line is also slightly displaced. However, the possible displacement (14 μm at the position of the mask) would be insignificant compared with the width of each aperture (0.8 mm). The intensity measured at the resting state was 1400 count/sec. The intensities during and after tetanus were expressed as percentages of the resting intensity and plotted against time after the first stimulus of each set of stimuli. Each point represents the intensity averaged over a 100-msec period. The first three points represent the measurements made before stimulation.

2.33 EXPLANATION. It is possible to overdo the explanation in a caption. The complete description of the experimental procedure in this caption is too much detail. The ratio of the caption area to the graph area is 2.8.

Too little detail, however, occurs more frequently in graphs in science and technology than too much detail. Figure 2.34 [40] is an example. The bars and error bars are not explained anywhere. One good guess is that they are sample means and estimates of the standard errors of the means; guessing should not be necessary.

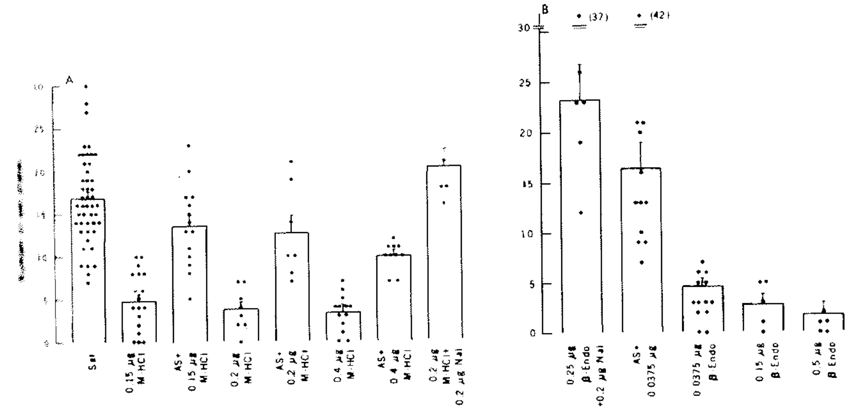


Fig. [illegible] Inhibitory effect of morphine hydrochloride (A) and β-endorphin (B) on wet-shake behavior in rats. Antagonism was by antibody to [illegible] sulfate (*AS*). A volume of 2 μl was delivered into the PAG in 1-μl increments with a 1-minute interval in between. The control [illegible] of saline (*Sal*). Morphine HCl (*M-HCl*) and β-endorphin were preceded by saline, AS, or naloxone (*Nal*) as indicated. The dose of morphine HCl and naloxone refers to the chloride salt. Each point is the datum for one animal. The AS + morphine HCl groups are all significantly different (*t*-test, $P < .005$) from the corresponding morphine HCl groups alone (A). The group receiving AS + 0.0375 μg of β-endorphin is [illegible] significantly different ($P < .005$) from the group receiving 0.0375 μg of β-endorphin (B).

2.34 EXPLANATION. The more common problem of scientific data display is too little explanation, rather than too much. The bars and error bars on this graph are not explained in the text or in the caption.

Error bars should be clearly explained.

Error bars are a convenient way to convey variability in data. Unfortunately, terminology is so inconsistent in science and technology that it is easy for an author to say one thing and a viewer to understand something else. Error bars can convey one of several possibilities:

(1) The *sample standard deviation* of the data.

(2) An *estimate of the standard deviation* (also called the *standard error*) of a statistical quantity.

(3) A *confidence interval* for a statistical quantity.

As an example, let us consider a particular case, also the most frequent one. Suppose the data are $x_1, \ldots, x_n$ and the statistical quantity being graphed is the sample mean,

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \, .$$

The sample standard deviation of the data is

$$s = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2} \, .$$

An estimate of the standard error of the mean is

$$s/\sqrt{n} \, .$$

If the data are from a normal distribution then a 95% confidence interval for the population mean is $(\overline{x} - k\, s/\sqrt{n},\ \overline{x} + k\, s/\sqrt{n})$, where k is a value that depends on n; if n is larger than about 60, k is approximately 1.96.

Error bars are used in Figure 2.35 [64]. In the last sentence of the figure caption we are told that the graphed values "represent means of three to four mice ± the standard deviation." What are we being shown? Is it (1) or (2) above? It is probably (1), but we should not have to deal with probability in understanding what is graphed.

Error bars should be unambiguously described. For the three cases cited above, the following is some terminology that can prevent ambiguity:

(1) The error bars show plus and minus one sample standard deviation of the data.

(2) The error bars show plus and minus an estimate of the standard deviation (or one standard error) of the statistic that is graphed.

(3) The error bars show a confidence interval for the statistic that is graphed.

Unambiguous description is only one issue with which we need to concern ourselves in showing error bars on graphs. A second important issue is whether they convey anything meaningful. This statistical issue is discussed in Section 3.14 (pp. 212–220).

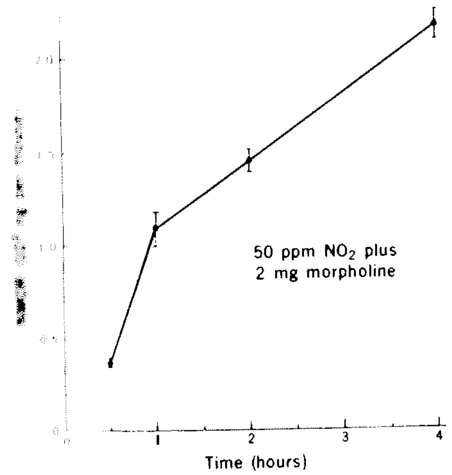
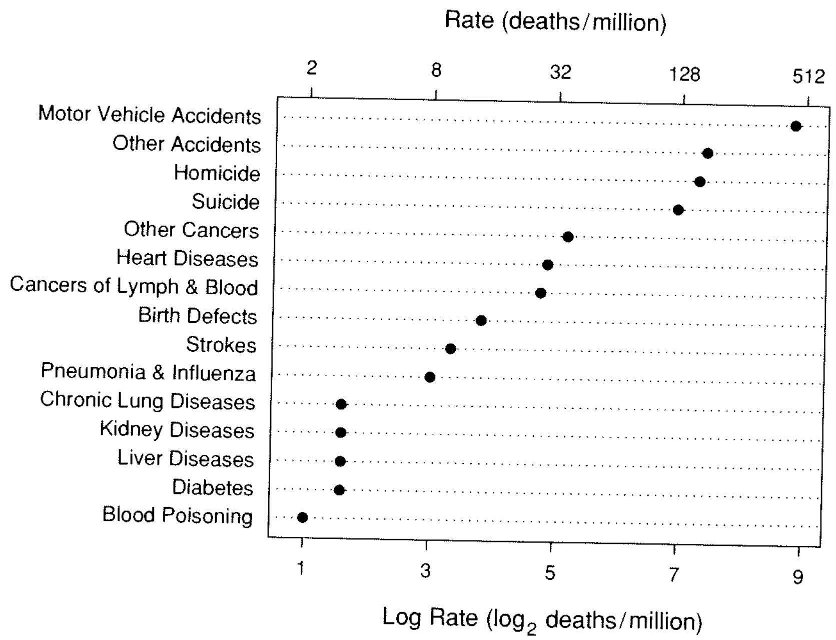


Fig. 1. Time course of NMOR biosynthesis in mice. Groups of three to four male ICR mice were gavaged with freshly prepared solutions of 2 mg of MOR (Aldrich Chemical) in 0.2 ml of distilled water and immediately placed in exposure chambers (Nalge desiccators, modified for gas inflow from the bottom and exhaust from the top). Mice were then exposed to 50 ppm of NO_2 (three to four mice per chamber, 5 cubic feet per hour, 20 volume changes per hour) at intervals of from 0.5 to 4 hours. The required concentrations of NO_2 were produced by mixing stock NO_2 (custom grade, Union Carbide) with air at an appropriate flow rate, prior to introduction into the chambers; we checked the accuracy of the exposure mixtures by periodically monitoring and analyzing the NO_2 in the exhaust from the chambers, using the Griess-Saltzman reaction (*19*). Concurrent controls consisted of two mice exposed in separate chambers to NO_2 alone for 4 hours, additional controls were gavaged with 2 mg of MOR or 0.2 ml of distilled water and exposed to air for identical periods in separate chambers. After exposure to NO_2, the mice were killed by freezing in liquid nitrogen and blended to a fine powder (*20*). Two or three aliquots (approximately 8 g each) were taken from each mouse powder and blended with 75 ml of ice-cold 35 percent aqueous methanol in a Waring Blendor (5 minutes, medium speed); a known amount of a nitrosamine standard [152 ng of di-*n*-propylnitrosamine (DPN), Aldrich] was then added, and blending continued for 1 to 2 minutes. Homogenates were divided in half and centrifuged (5000*g*, 25 minutes, 5°C; swinging bucket), supernatant was removed, and the pellets were extracted again with cold 35 percent methanol. The pooled supernates were extracted (twice) with an equal volume (total, 150 ml) of dichloromethane [(DCM), Burdick and Jackson] (*21*), and the organic layer was dried by passage through a cotton gauze (Ex-tube, Analytichem International) and concentrated to 2 ml in a Kuderna-Danish concentrator (Kontes, 250 ml) kept in a 65°C bath. Aliquots (20 μl) of the concentrates from each of two or three powder samples were injected into the thermal energy analyzer gas chromatograph (Thermo Electron modified model TEA-502) (*22*) for NMOR analysis. Peaks were identified and quantitated by comparison with the retention time and response of reference nitrosamines (*23*). The plotted values are corrected for any background control levels and for the DPN standard recoveries and represent means of three to four mice ± the standard deviation.

2.35 ERROR BARS. *Error bars should be clearly explained.* It is important to distinguish between the sample standard deviation and an estimate of the standard deviation of the sample mean (the standard error of the mean). It is not clear from the explanation of this graph which of these two deviations the error bars portray.

When logarithms of a variable are graphed, the scale label should correspond to the tick mark labels.

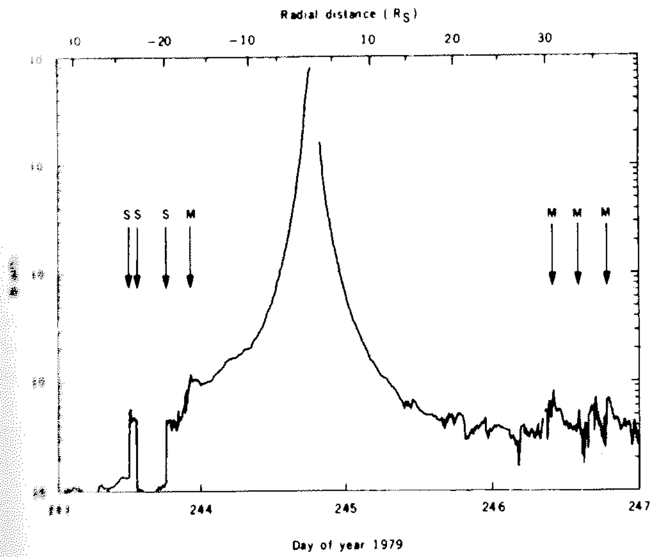
The dot plot in Figure 2.36 shows death rates for the leading causes of death of people in the age group 15 to 24 years in the United States [95]. The logarithms of the data are graphed; that is, equal increments on the horizontal scale indicate equal increments of the logarithm of death rate. On the top horizontal scale line the tick mark labels show the values of the data on the original scale; the scale label describes the variable and its units on the original scale, to correspond to the tick mark labels. On the bottom horizontal scale line, the tick mark labels are in log units; the scale label describes the variable and its units on the log scale.



2.36 LABELS FOR LOGS. *When logarithms of a variable are graphed, the scale label should correspond to the tick mark labels.* The logarithms of the data are graphed on this dot plot. On the top horizontal scale line the tick mark labels are in the units of the data on the original scale, so the scale labels describe the data on the original scale. On the lower scale line the tick mark labels are expressed in log units of the data, so the scale label describes the logarithms of the data.

Proofread graphs.

Graphs should be proofread and carefully checked for errors. Figure 2.37, a graph of measurements of Saturn's magnetic field made by the Pioneer II spacecraft, has an error; the exponents for the tick mark labels on the vertical scale line are missing [110]. This is quite unfortunate since the magnitude of the magnetic field is of much interest. The authors write about the graph: "This is shown in Figure 1.1, which presents an overview of the encounter as evident in the magnitude of the ambient magnetic field." It is unfortunate to have a graph error degrade the communication of such exciting, high-quality scientific work.



2.37 PROOFREAD. *Proofread graphs.* Graphs should be proofread, just as we do text. On this graph, lack of careful proofreading resulted in missing exponents on the tick mark labels of the vertical scale.

Strive for clarity.

Strive for clarity is really a summation of the principles presented so far; in Section 2.2 (pp. 25–54) the principles contribute to making a graph visually clear, and in this section the principles contribute to a clear understanding of what is graphed. Striving for clarity should be done consciously. We should ask of every graph, "Are the data portrayed clearly?" and "Are the elements of the graph clearly explained?" Let us consider one example.

The data in Figure 2.38 [129] are percentages of degrees awarded to women in several disciplines of science and technology during three time periods. The elements of the graph are not fully explained; little is said in the text, so we must rely on the labeling and the caption to understand what is graphed. At first glance the labels suggest the graph is a standard divided bar chart with the length of the bottom division of each bar showing the percentage for doctorates, the length of the middle division showing the percentage for master's, and the top division showing the bachelor's. This is not so. It would imply that in most cases the percentage of bachelor's degrees given to women is generally lower than the percentage of doctorates.

A little detective work reveals that the three values of the data for each discipline during each time period are determined by the three adjacent, vertical, dashed lines. The top end of the left line indicates the position of the value for doctorates, the middle line indicates the master's degrees, and the right line indicates the bachelor's degrees; the thick horizontal line segments are drawn at the ends of the vertical lines. The horizontal segments are what one sees most prominently, but without intensive visual scrutiny of the graph, it is not possible to determine the degree associated with a particular segment.

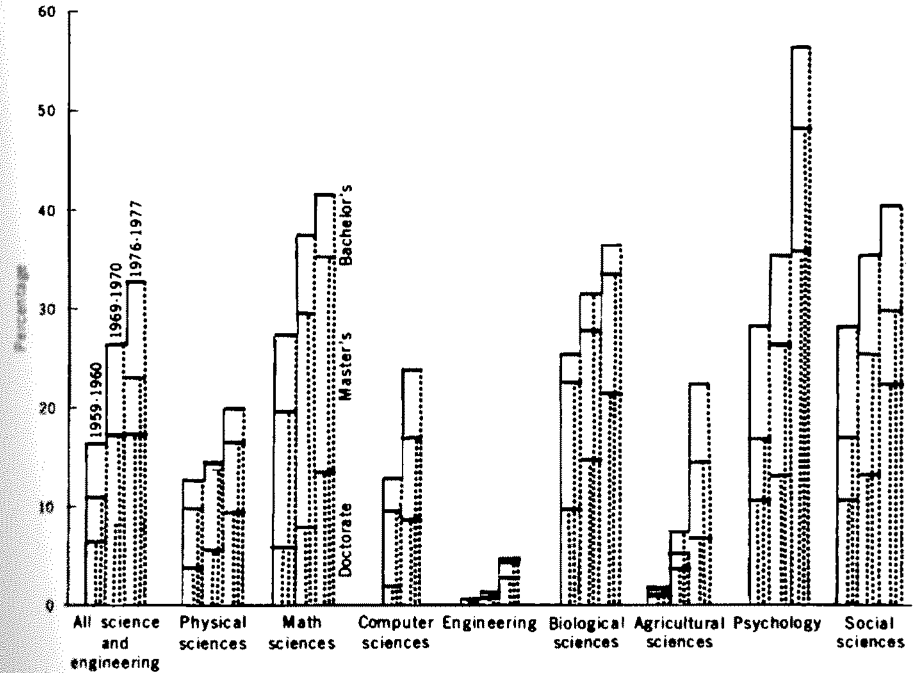
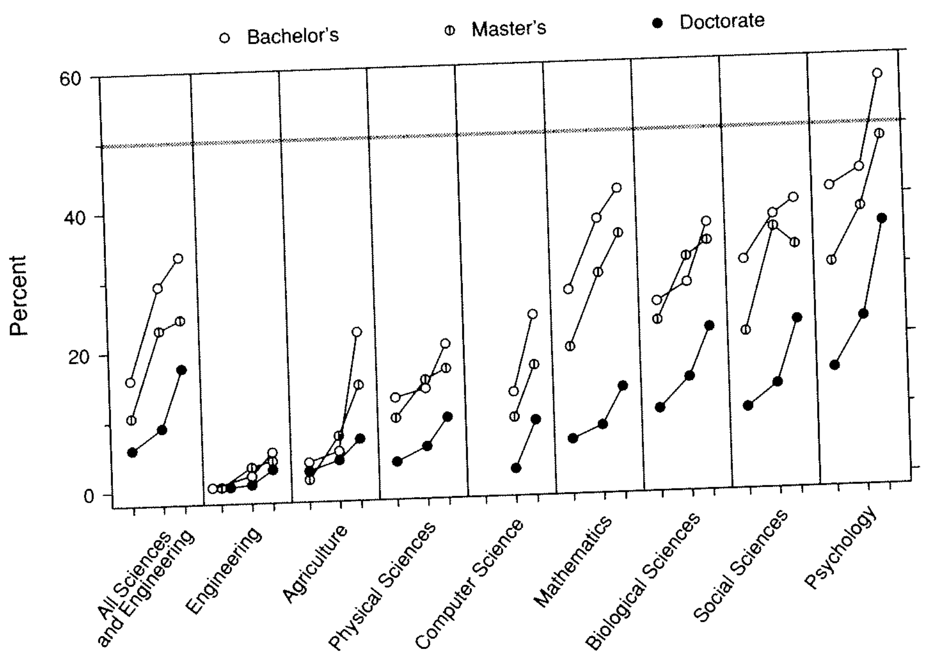


Fig. 1. Proportion of degrees in science and engineering earned by women in 1959 to 1960, 1969 to 1970, and 1976 to 1977 (6). Included in the social science degrees are anthropology, sociology, economics, and political science.

2.38 CLARITY. This graph fails both in clarity of vision and clarity of explanation.

There are other problems with this graph. Only two bars are shown for computer science, with no explanation. One can only assume that the 1959-1960 time period is missing. There is a construction error; the horizontal line for doctorates in all science and engineering in 1969-1970 is missing. Another difficulty with the graph is visual; the bar chart format makes it hard to visually connect the three values of a particular degree for a particular discipline to see change through time.

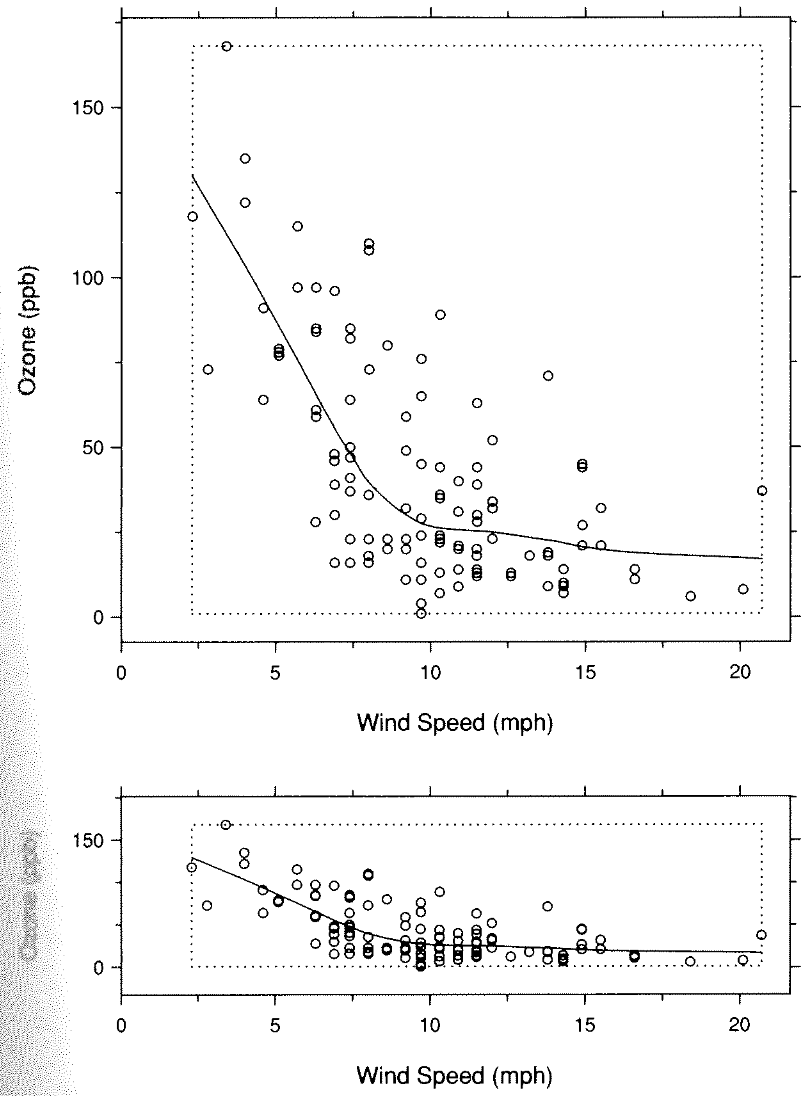
In Figure 2.39 the data from Figure 2.38 are regraphed. There has been a striving for clarity. It is clear how the data are represented, and the design allows us to see easily the values of a particular degree for a particular discipline through time. Finally, the figure caption explains the graph in a comprehensive and clear way.



2.39 CLARITY. *Strive for clarity.* This graphing of the data from Figure 2.38 strives for clarity. It shows the percentage of degrees earned by women for three degrees, three time periods, and nine disciplines. For each discipline, the three tick marks indicate the years 1959–1960, 1969–1970, 1976–1977.

2.4 Banking to 45°

The *data rectangle*, defined at the beginning of this chapter, just encloses the data. In Figure 2.40 the same data are graphed twice. On each panel the dotted rectangle is the data rectangle; the shapes of the two rectangles are different because the vertical scale line is shorter in the bottom panel.



2.40 ASPECT RATIO. The dotted rectangle that just encloses the data on each panel is the data rectangle. The aspect ratio is the height of the data rectangle divided by the width. In the top panel, the aspect ratio is 1 vcm/hcm, and in the bottom panel, it is 0.25 vcm/hcm.

The *aspect ratio* of a graph is the height of the data rectangle divided by the width. In the top panel of Figure 2.40 the aspect ratio is 1 vcm/hcm, where vcm means vertical centimeters and hcm means horizontal centimeters, and in the bottom panel it is 0.25 vcm/hcm.

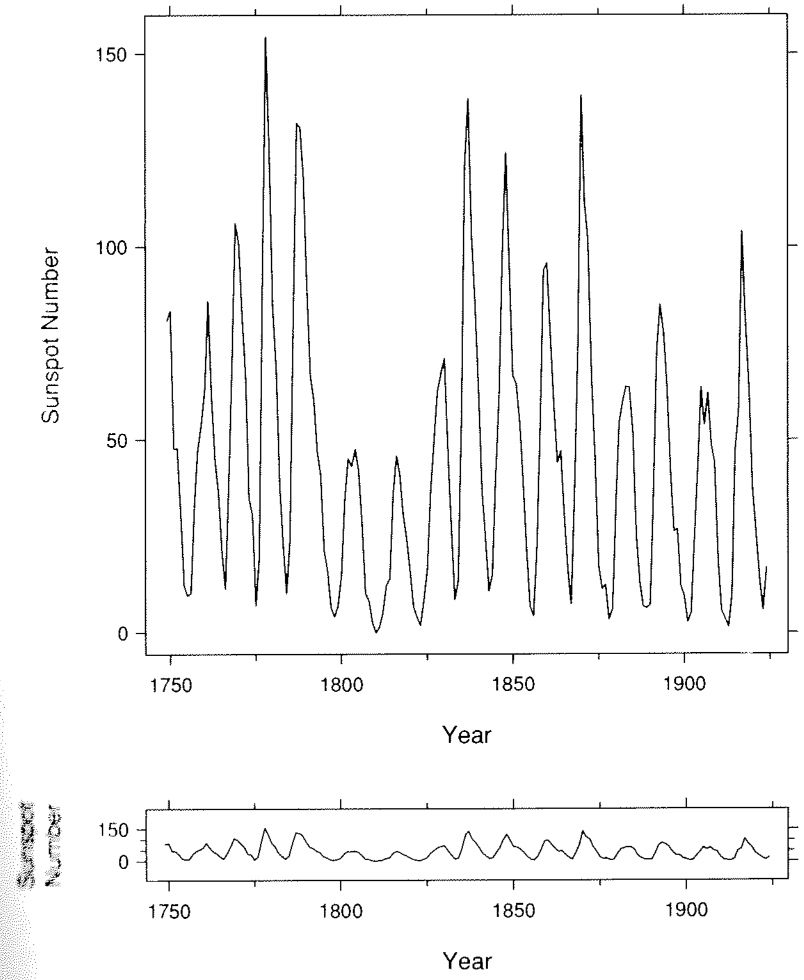
The two panels of Figure 2.40 graph the air pollutant ozone against wind speed for 111 days in New York City from May 1 to September 30 of one year [13]. Superposed on each panel is a smooth curve that describes the dependence of ozone on wind speed. For example, we can see that the general pattern is for ozone to decrease as wind speed increases because of the increased ventilation of air pollution that higher wind speeds bring. At each value of wind speed the curve has a slope that encodes the rate of change of ozone as a function of wind speed. The units of this rate of change are ppb/mph. For low values of wind speed, the slope is very negative; in other words, a small increment in wind speed brings a large reduction in ozone. For high values of wind speed, the slope is close to zero; in other words, a small increment in wind speed brings a small reduction in ozone.

Each curve in Figure 2.40 consists of a collection of short connected line segments, the *local line segments*. We visually decode the information about the relative local rate of change of ozone with wind speed by judging the *orientations* of these local segments. The orientation of a segment is its angle with the horizontal. Segments with positive slopes have positive orientations and segments with negative slopes have negative orientations. A segment with slope 1 has an orientation of 45°, a segment of slope -1 has an orientation of $-45°$, and a segment with zero slope has an orientation of 0°. The orientations of line segments on a graph change if the aspect ratio of the graph changes. If the aspect ratio increases, there is an increase in the steepness of the orientation of any segment that has a slope other than zero. For example, in the top panel of Figure 2.40, each nonzero orientation is steeper than its counterpart in the bottom panel because the aspect ratio is greater in the top panel.

The aspect ratio of a graph is an important factor for judging rate of change.

The aspect ratio of a display such as Figure 2.40 greatly affects the accuracy of our visual decoding of the rate of change of y with x; as we just saw, the aspect ratio controls the overall steepness of the orientations of line segments, and it is such orientations that we judge to decode information about rate of change. This effect of the aspect ratio is demonstrated in Figure 2.41. The display, discussed previously in Chapter 1, graphs the yearly sunspot numbers from 1749 to 1924. In the top panel, the shape parameter is 1 vcm/hcm, and in the bottom panel,

it is 0.055 vcm/hcm. The dominant frequency component of variation of the data is the cycles whose average period is about 11 years. These cycles are evident in both panels of the figure, but the top panel fails to reveal an important property — the cycles rise more rapidly than they fall. The faster rise than fall is most pronounced for the cycles with high peaks, is less pronounced for those with medium peaks, and disappears for those cycles with the very lowest peaks.



41 BANKING. *The aspect ratio of a graph is an important factor for judging rate of change.* The sunspot numbers are graphed in both panels. The aspect ratio is 1 vcm/hcm in the top panel. In the bottom panel the aspect ratio is 0.055 vcm/hcm; the line segments are banked to 45°, which allows us to more accurately visually decode information about rate of change.

The importance of the aspect ratio was recognized as early as 1914 [12], and for many decades there was much discussion. But no one attempted to base discussion on a rigorous investigation of the effect of the aspect ratio on our visual decoding of information. There was no science. The result was numerous, contradictory opinions [33]. This is not surprising since with no facts arising from either convincing theoretical arguments or careful experiments, there was no scientific body of material around which opinions could coalesce.

In the late 1980s the problem was finally attacked using mathematics, the theory of visual perception, and controlled experiments, and the solution was found [25, 33]. Here, we will simply present the results, but in Section 4.7 (pp. 251–256) the scientific enquiry that led to the solution is described.

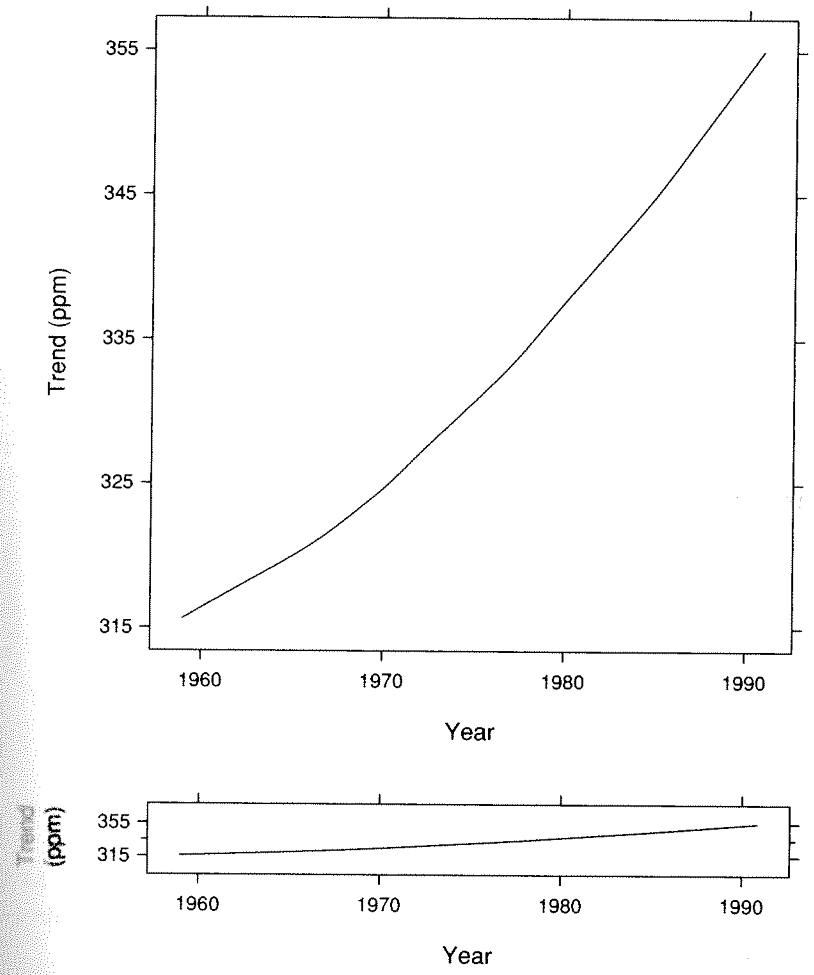
When the orientations of line segments are judged to decode information about rate of change, bank the segments to 45°.

The judgments of the orientations of line segments are optimized when the aspect ratio is chosen so that the absolute values of the orientations of the segments are centered on 45°. This tends to center the segments with positive slopes on 45° and the segments with negative slopes on −45°. This centering is called *banking to 45°*, a display method whose name suggests the banking of a road to affect its slope. In Section 4.7 (pp. 251–256) a formula is given for the aspect ratio that achieves such banking.

In the bottom panel of Figure 2.41 the local line segments that make up the curve are banked to 45°; the aspect ratio is 0.055 hcm/vcm. It is this banking that allows us to see the faster rise than fall of the sunspots. In the top panel the aspect ratio is 1.00 vcm/hcm; the absolute orientations are centered on an angle much greater than 45°, which degrades the accuracy of our visual decoding of rate of change.

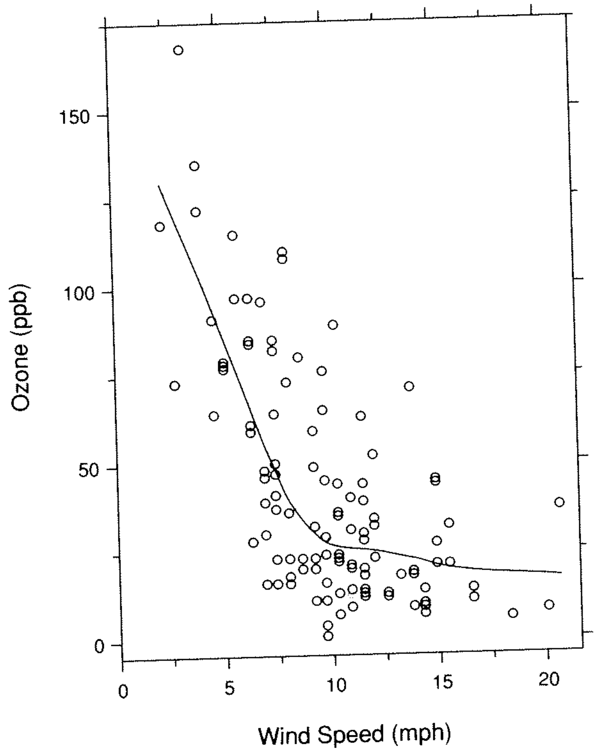
Banking to 45° is used in the top panel of Figure 2.42; the aspect ratio is 1.00 vcm/hcm. The data are a trend curve for atmospheric CO_2 concentrations at Mauna Loa Observatory in Hawaii; the curve was graphed earlier in one panel of Figure 1.2. Through our judgment of the orientations of the banked, local line segments, we see that the curve is convex; this means that the rate of increase of CO_2 is increasing through

time. In the bottom panel of Figure 2.42 the aspect ratio is 0.055 vcm/hcm, the optimum value for the sunspot data. But for the CO_2 data, this aspect ratio centers the absolute orientations on an angle much less than 45°; the result is a less accurate perception of rate of change.



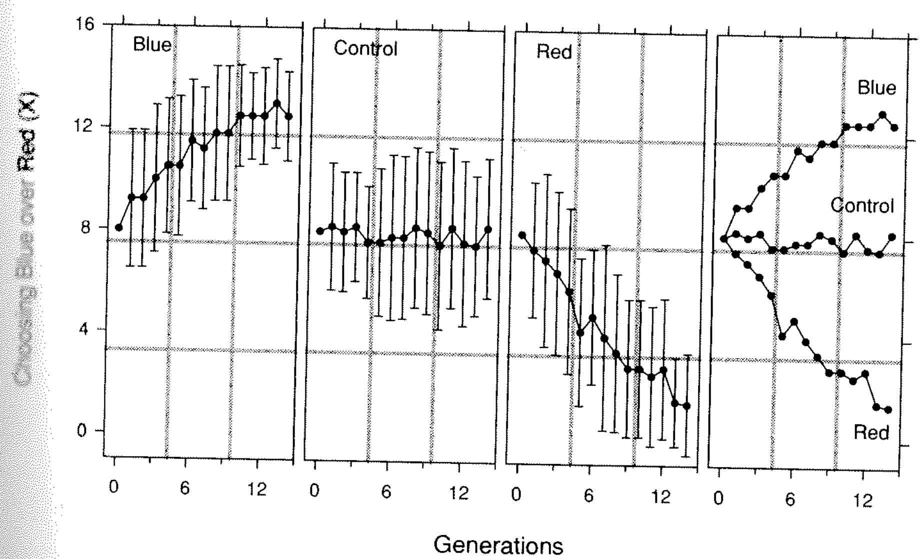
2.42 BANKING. *When the orientations of line segments are judged to decode information about rate of change, bank the segments to 45°.* The CO_2 trend curve is graphed in both panels. In the top panel, the segments are banked to 45° and the aspect ratio is 1.00 vcm/hcm; we can readily judge the convexity of the curve. In the bottom panel, the aspect ratio is 0.055 vcm/hcm; the absolute orientations are centered on an angle much less than 45°, which interferes with our judgment of rate of change.

The principle of banking to 45° is a universal one and applies to the judgment of orientation of any collection of line segments, not just a single curve by itself on a graph as in Figures 2.41 and 2.42. For example, it applies to a curve superposed on a set of points. This is illustrated in Figure 2.43, which graphs the ozone and wind speed data from Figure 2.40. The local segments of the curve are banked to 45°.



2.43 BANKING. The segments that make up the smooth curve are banked to 45°.

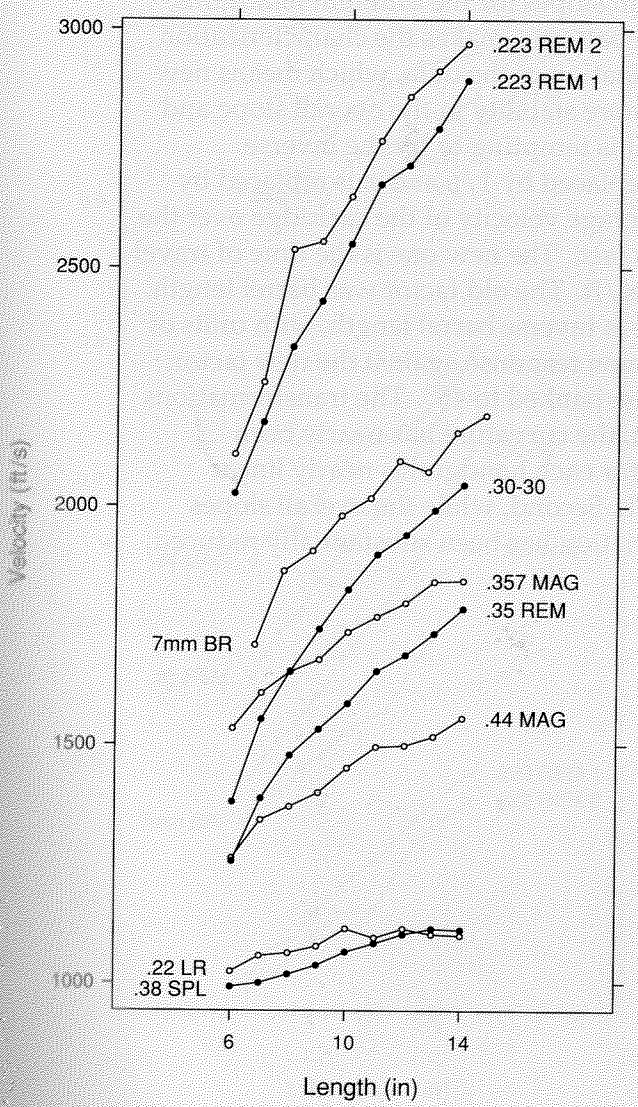
The principle applies to the judgment of orientations of segments on juxtaposed panels with the same scales. This is illustrated in Figure 2.44, shown earlier in Section 2.2 (pp. 25–54). The collection of 84 segments that connect the measurements on the four panels is banked to 45°. There is a simple way to think of this banking. Suppose the data were graphed on a single panel. There would be a single data rectangle, and we would choose the aspect ratio to bank the collection of segments to 45°. For the multipanel display, the scales are exactly the same as they would be on the single-panel display. In other words, for the multipanel display we make four copies of the scale-line rectangle of the single-panel display, and then graph part of the data on each panel.



2.44 BANKING. The entire collection of line segments connecting plotting symbols is banked to 45°.

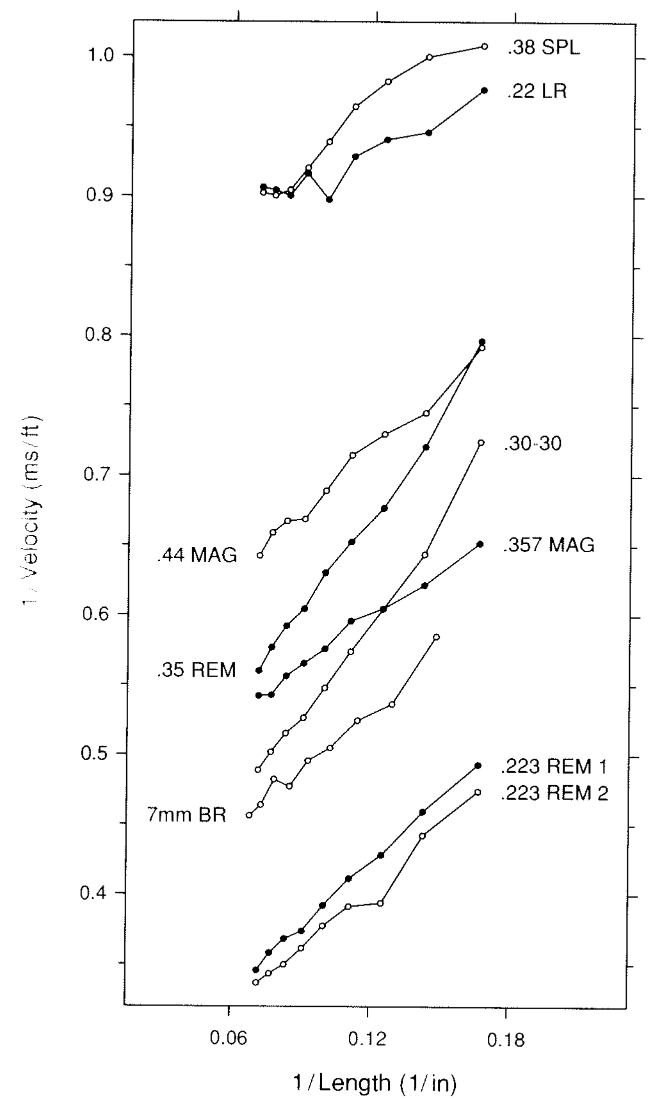
Banking is critical to seeing important effects in Figure 2.45. The data, collected by Bob Milek [93], are from a handgun experiment to study how cartridge velocity, the response, depends on barrel length, the factor. For each handgun, the average cartridge velocity for the first 12 feet from the muzzle was measured for each of several different barrel lengths. On the graph, velocity is graphed against length for nine handguns. One handgun, .223 REM, appears twice with two different types of ammunition. The segments connecting successive values of all of the handgun curves are collectively banked to 45°.

The ability to effectively judge rate of change in Figure 2.45 allows us to see important patterns in the data. First, the overall slopes of the handgun curves tend to increase as we go from bottom to top on the graph. In other words, the increase in velocity as a function of barrel length tends to be greater for handguns with fast cartridges than for those with slow ones. For example, the .223 REM handguns at the top of Figure 2.45 have a much greater rate of increase than the .38 SPL and the .22 LR at the bottom. Second, the underlying pattern in the data for each handgun tends to be concave. In other words, the rate of change of velocity as a function of barrel length tends to decrease as the length increases.



2.45 BANKING. The line segments that connect successive observations are banked to 45°. This allows us to readily perceive that the slopes of the underlying patterns increase as the overall velocity levels increase.

The large variation in the overall slopes for the different handguns, together with the curvature of the patterns, makes the characterization of the data complicated. Transformation of the data, which means new measurement units, might yield more stability in the overall slope and remove or reduce the curvature. It is tempting to try the inverse transformation. This means y is replaced by $1/y$ and x is replaced by $1/x$. The old response was the average velocity of the cartridge over the distance of 12 feet, with units of ft/sec. The new one is the time of travel of the cartridge, with units of msec/ft. The old factor was barrel length, with units of inches. The new one is inverse barrel length, with units of 1/inches. Figure 2.46 graphs the new response against the new factor. Again, the connecting segments are banked to 45°. The transformations have simplified the patterns. First, the curvature has nearly been removed; the underlying pattern for each handgun is nearly linear except possibly that for the .38 SPL. Second, while the overall slopes appear to vary somewhat, the variation has been substantially reduced.



2.46 BANKING. The segments that connect successive observations are banked to 45°. The graph shows that the slopes of the patterns do not depend on the levels.

Suppose v denotes velocity and l denotes barrel length. Figure 2.46 shows that for each handgun,

$$v^{-1} = \alpha + \beta l^{-1},$$

where α and β depend on the characteristics of each handgun, and thus vary from one handgun to the next. Values of α and β were estimated for each handgun by fitting lines to the transformed data using least-squares. The fitted lines are graphed in Figure 2.47 and are banked to 45°.

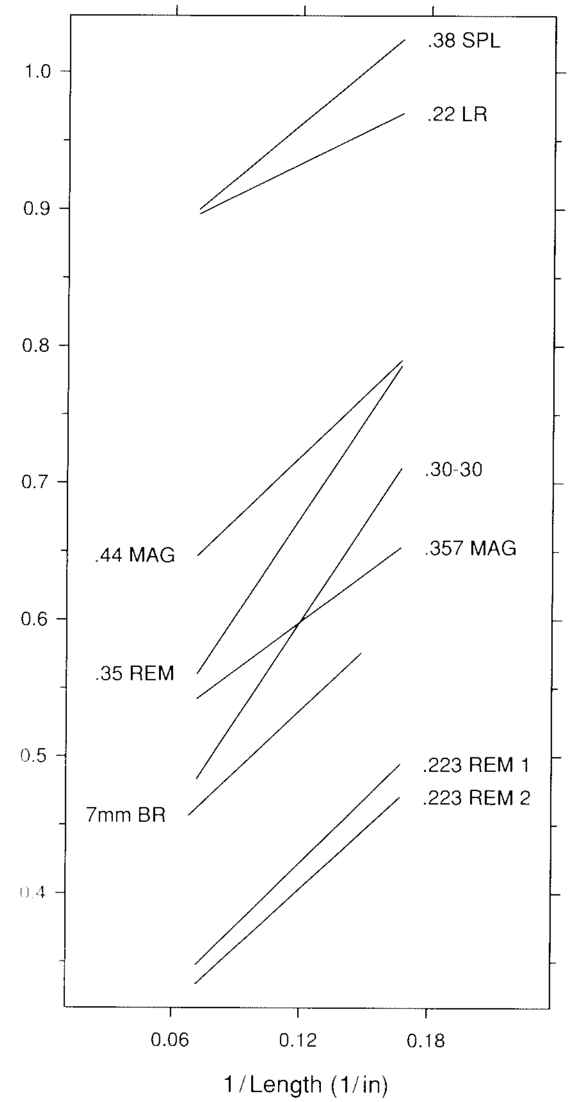
The display of the velocity data using banking has led to an empirical law for the relationship of v and l. The law stipulates that v depends nonlinearly on l through the mathematical equation

$$v = \frac{l}{\alpha l + \beta}.$$

The equation now allows the derivation of properties of the relationship. For example, Milek, the experimenter, was keen to determine the sensitivity of velocity to an increase in barrel length. The derivative of v with respect to l, which measures the rate of change of v with l, is

$$\frac{dv}{dl} = \frac{\beta}{(\alpha l + \beta)^2}.$$

Thus the derivative decreases monotonically, which means the sensitivity of velocity to length decreases monotonically as the length increases.



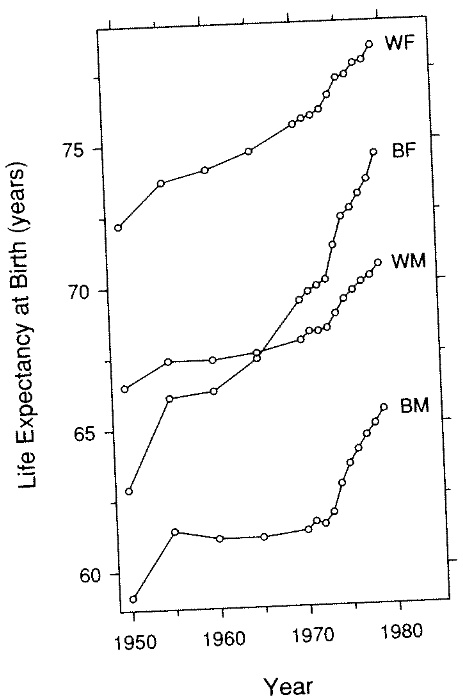
2.47 BANKING. The line segments, which show the least-squares lines fitted to the data, are banked to 45°.

2.5 Scales

Scales are fundamental. A graph is a graph, in part, because it has one or more scales. Graphing data would be far simpler if these basic, defining elements of graphs were straightforward, but they are not; scale issues are subtle and difficult.

Choose the range of the tick marks to include or nearly include the range of the data.

The interval from the minimum to the maximum of a set of values is the range of the values. It is a good idea to have the range of the data on a graph be included or nearly included in the range of the tick marks to allow an effective assessment of all of the data. In Figure 2.48 the range of the data on the horizontal scale is included in the range of the tick marks, and the data on the vertical scale are nearly included in the range.

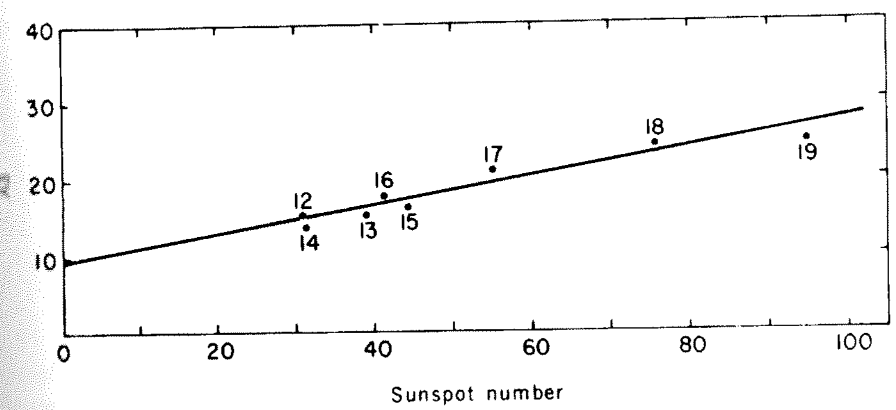


2.48 RANGES. *Choose the range of the tick marks to include or nearly include the range of the data.* The range of the data on the vertical scale is nearly contained within the range of the tick marks. On the horizontal scale the range of the data is completely contained within the range of the tick marks. The line segments connecting successive plotting symbols are banked to 45°.

Subject to the constraints that scales have, choose the scales so that the data rectangle fills up as much of the scale-line rectangle as possible.

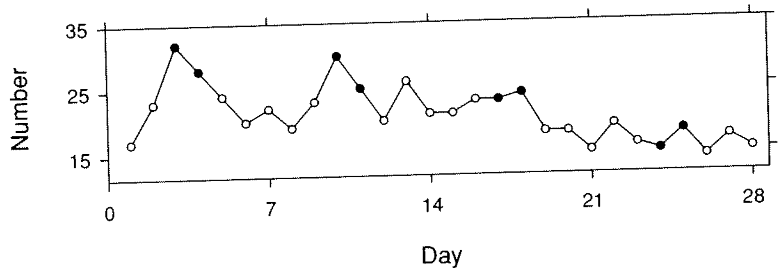
There are a number of constraints that affect the choice of scales on graphs. One, just discussed, is that the range of the tick marks should encompass or nearly encompass the range of the data. Another is that we do not want data to be graphed on scale lines. Also, in some cases we want a particular value to be included in the scale; the most common example is showing a zero value. (More will be said later about including zero.) Finally, when different panels of a graph are compared, we will often want the scales to be the same on all panels.

But subject to these constraints, we should attempt to use as much of the interior of the scale-line rectangle as possible. In other words, the data rectangle should lie slightly inside the scale-line rectangle. This is not done in Figure 2.49 [116]. Only 26% of the vertical scale is taken up by the data. Space is wasted on this graph.



2.49 FILLING THE SCALE-LINE RECTANGLE. Only 26% of the vertical scale is used by the data.

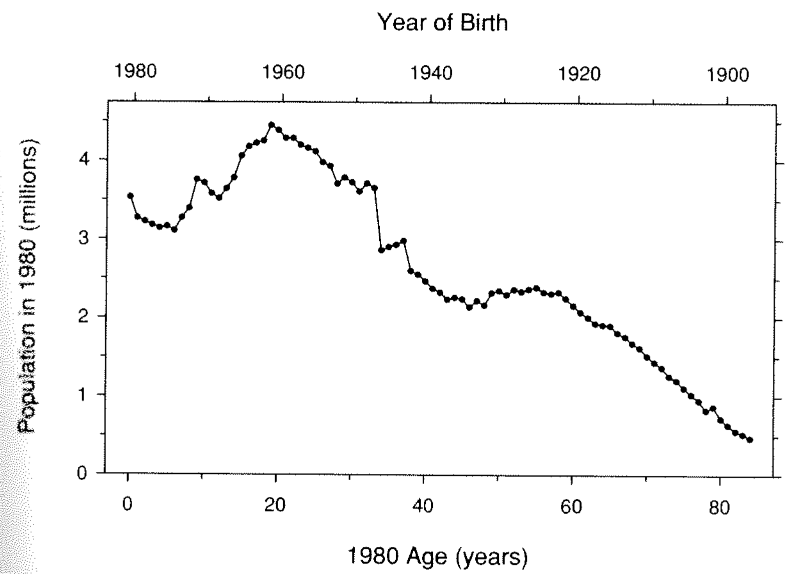
In contrast, Figure 2.50 utilizes space more efficiently. The data span most of the range of the scales without getting too close to the scale lines. The data are the number of cigarettes consumed daily by a smoker in a 28-day program to quit smoking; after the 28 days the smoker quit altogether. A "day" is defined as starting at 6:00 a.m. and ending 24 hours later. The open circles are the days Monday to Friday and the closed circles are Saturdays and Sundays.



2.50 FILLING THE SCALE-LINE RECTANGLE. *Subject to the constraints that scales have, choose the scales so that the data rectangle fills up as much of the scale-line rectangle as possible.* This graph uses space efficiently. The line segments connecting successive plotting symbols are banked to 45°.

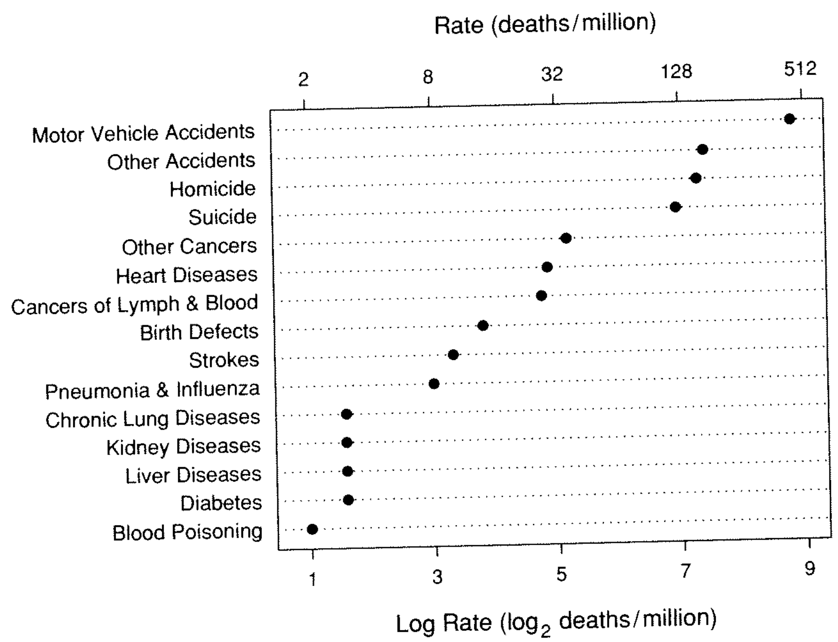
It is sometimes helpful to use the pair of scale lines for a variable to show two different scales.

The two scale lines for a variable on a graph provide an opportunity to show two different scales for the variable; the additional information of a second scale often can be helpful. One example is Figure 2.51. The data are the number of people in the United States in 1980 for each age from 0 to 84 [128]. The bottom horizontal scale line shows the age and the top horizontal scale line shows the year of birth.



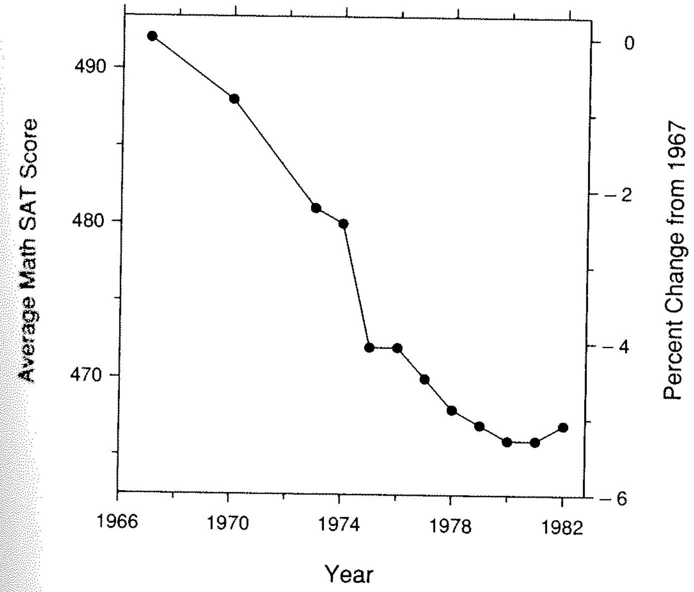
2.51 TWO SCALES. *It is sometimes helpful to use the pair of scale lines for a variable to show two different scales.* The bottom horizontal scale line shows age and the top horizontal scale line shows year of birth. The line segments connecting successive plotting symbols are banked to 45°.

When the logarithms of data are graphed there is an opportunity to use two scales. In Figure 2.52 the death rates for people 15 to 24 years old are graphed on a log scale. The bottom horizontal scale line shows log death rate in log deaths/million. The tick mark labels on this scale line allow us to see quickly by how much two values of the data differ in multiples of two. For example, the death rate due to automobile accidents is four times larger than that for suicide. The top scale line shows death rate on the original scale in deaths/million. This scale is added to allow an assessment of the magnitudes of the death rates without having to take powers of two in our heads.



2.52 TWO SCALES. The bottom horizontal scale line shows log death rate in log deaths/million and the top horizontal scale line shows death rate in deaths/million.

When magnitudes are shown on a graph we can use two scales to show the data in their units of measurement and to show percent change from some baseline value. Figure 2.53 is a graph of averages of the mathematics Scholastic Aptitude Test scores for selected years from 1967 to 1982 [127]. The left vertical scale line shows the scores and the right vertical scale line shows percent change from 1967. Without the right scale it takes some mental arithmetic to determine the percent changes, for example, to see that the change from 1967 to 1982 was about 5%.

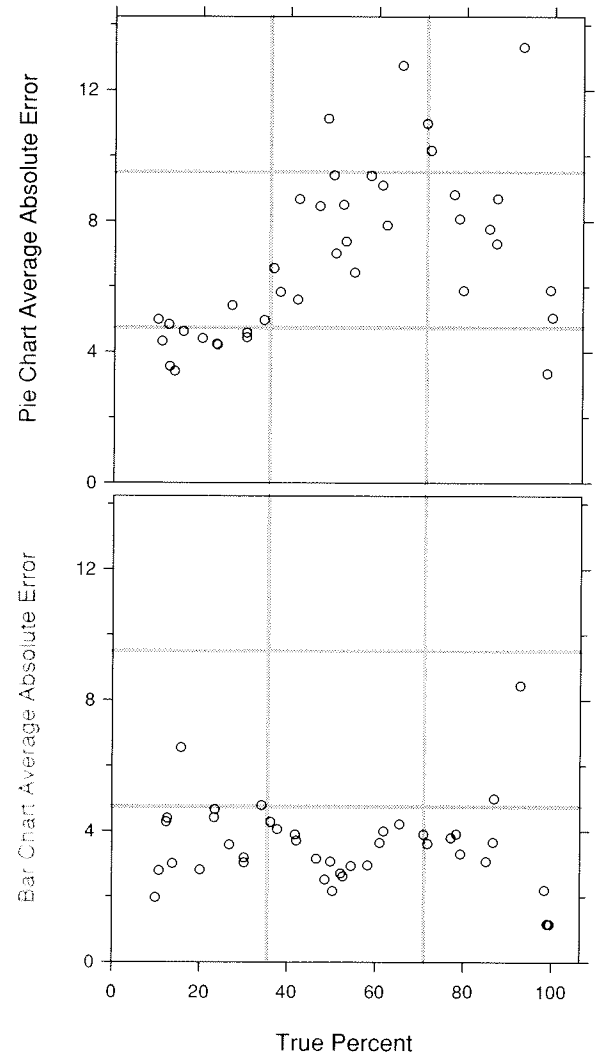


2.53 TWO SCALES. The left vertical scale line shows SAT score and the right vertical scale line shows percent change from 1967. The line segments connecting successive plotting symbols are banked to 45°.

Choose appropriate scales when data on different panels are compared.

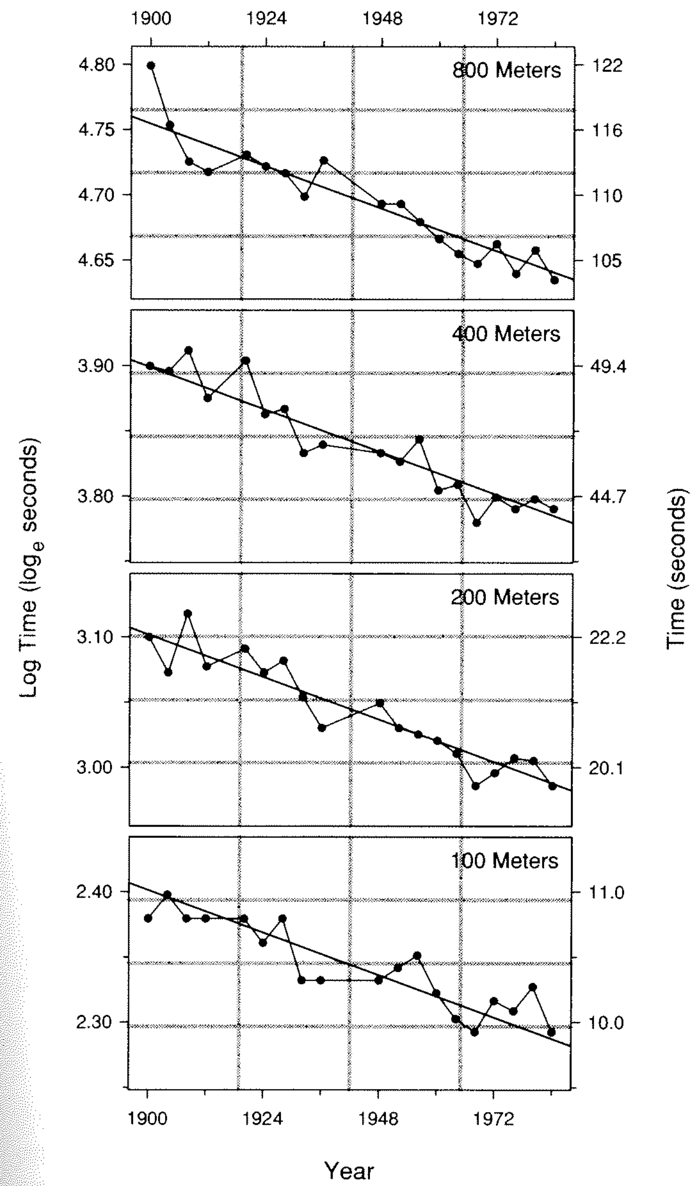
Figure 2.54 shows data from an experiment on graphical perception [34]. A group of 51 subjects judged 40 pairs of values on bar charts and the same 40 pairs on pie charts; each judgment consisted of studying the two values and visually judging what percent the smaller was of the larger. The top panel of Figure 2.54 shows the 40 average judgment errors (averaged across subjects) graphed against the true percents for the 40 pie chart judgments. The bottom panel shows the same variables for the bar chart judgments. To enhance the comparison of the bar chart and pie chart values, the scales on the two panels are the same; this allows us to see very clearly that the pie chart judgments are less accurate than the bar chart judgments. One result of the common scale is that the data do not fill either panel; we should always be prepared to forego the fill principle to achieve an effective comparison. But note that if all of the data were put on one of the panels, the data rectangle would nearly fill the scale-line rectangle.

Using the same scales in the two panels of Figure 2.54 allows a number of quantitative comparisons to be made of the two sets of data. We can compare the average level of the absolute error of each chart type for each true percent. For example, we can see that the average level is about the same for true percents less than about 35%, but generally for percents greater than 35%, the average level is greater for pie charts. Furthermore, we can compare the variation in the errors for the two chart types. There is greater change in the pie chart errors than in the bar chart errors as the true percent changes; the pie chart errors increase but the bar chart errors have a flat pattern. The same scales can be used in this example, in part, because the overall levels of the two sets of data are not radically different.



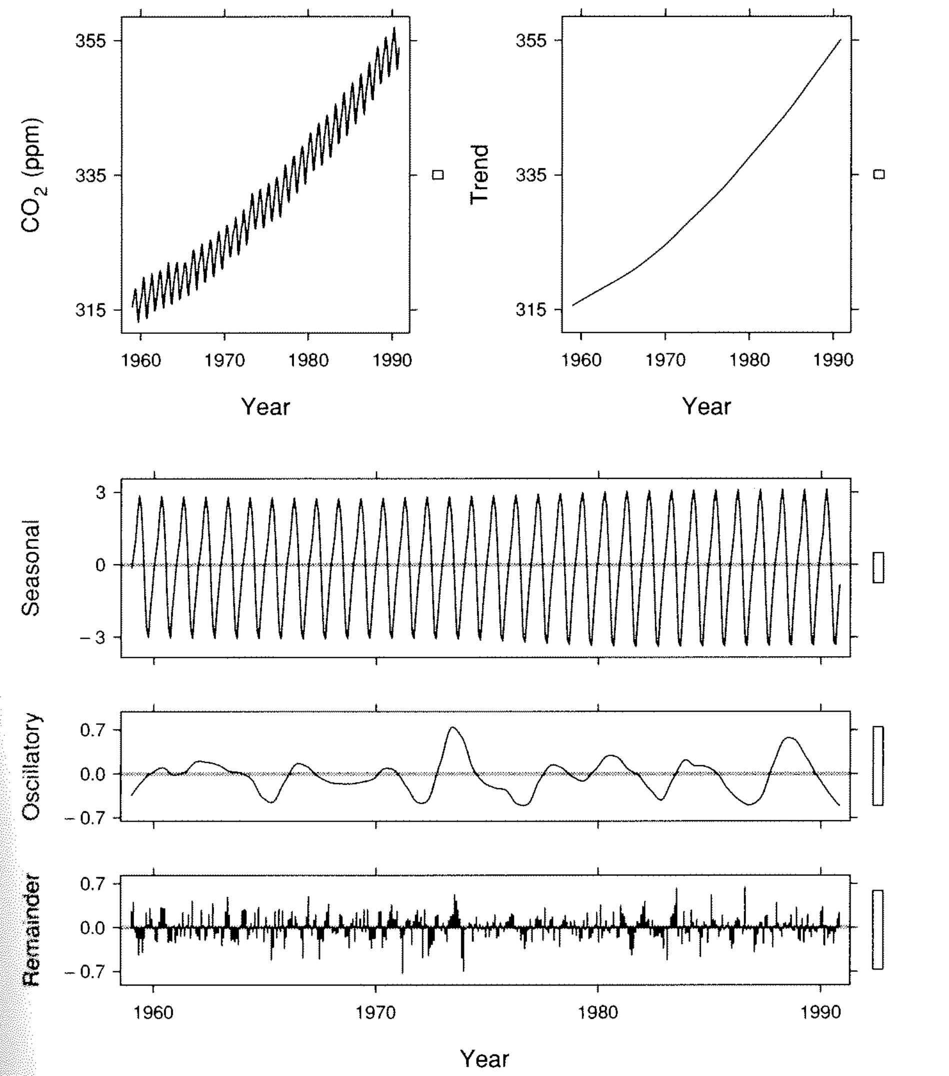
2.54 COMPARISON. *Choose appropriate scales when data on different panels are compared.* Scales on different panels should be made as commensurate as possible when the data on the different panels are compared. On this graph the scales on the top panel are the same as those on the bottom.

But suppose the overall levels of different data sets on a graph vary by a large amount. Figure 2.55 shows an example. The data are the winning times of four track races at the Olympics from 1900 to 1984 [17, 98]. The four lines have the same slopes but different intercepts and were fitted to the data using least-squares. The overall levels of the times are quite different; if the vertical scales were the same, the data on each panel would be contained in a very narrow horizontal band. Instead, the vertical scales vary, but the number of log seconds per cm is the same. This allows us to compare the rates of change of the four sets of log running times. For example, we can see that the overall rates of decrease through time for the four distances have been about the same. Since logs are graphed, this means that the percent reductions in the running times have been about the same. We can readily perceive the rate of change because the segments connecting the plotting symbols on the four panels are collectively banked to 45°.



2.55 COMPARISON. Sometimes making scales identical prevents effects from being seen. The next best thing is illustrated on this graph — the number of units per cm is the same on the four vertical scales. The four lines on the panels have the same slope. The line segments connecting successive plotting symbols on the four panels are banked to 45°.

Sometimes even the number of units per cm cannot be the same without ruining our judgment of rate of change. In Figure 2.56, the data in the top left panel are the monthly measurements of atmospheric CO_2 concentrations that were discussed in Section 1.1 (pp. 6–9). The other panels graph frequency components of the data. The variation of the data in the bottom two panels is considerably less than that in the top two panels; were we to make the number of units per cm the same on all panels, the vertical scales of the bottom two panels would be too small. One way to appreciate the changes in the number of units per cm on the five panels is to study the tick mark labels and the distances between them, but this is a difficult visual operation. To make appreciation of the scale change easier, rectangles have been put to the right of the panels. The vertical lengths of the rectangles represent equal changes in parts per million on the five vertical scales.



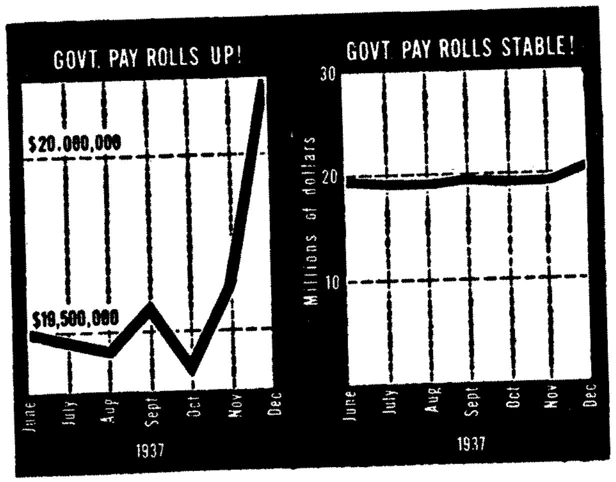
2.56 COMPARISON. Sometimes keeping the number of units per cm the same on different panels is not possible. On this graph the number of units per cm on the vertical scales varies. The rectangles on the right show the relative scaling; the vertical lengths represent the same change in ppm on the five vertical scales.

Do not insist that zero always be included on a scale showing magnitude.

When the data are magnitudes, it is helpful to have zero included in the scale so we can see its value relative to the value of the data. But the need for zero is not so compelling that we should allow its inclusion to ruin the judgment of the variation in the data.

There has been much polemical writing about including zero when graphs are used to communicate quantitative information to others. Too frequently zero has been endowed with an importance it does not have. Darrell Huff in his book *How to Lie with Statistics* [61] goes so far as to say that a graph of magnitudes without a zero line is dishonest. Referring to Figure 2.57 he writes:

> An editorial writer in *Dun's Review* in 1938 reproduced a chart from an advertisement advocating advertising in Washington, D.C., the argument being nicely expressed in the headline over the chart: GOVERNMENT PAY ROLLS UP! The line in the graph went along with the exclamation point even though the figures behind it did not. What they showed was an increase from about $19,500,000 to $20,000,000. But the red line shot from near the bottom of the graph clear to the top, making an increase of under four percent look like more than 400. The magazine gave its own graphic version of the same figures alongside — an honest red line that rose just four percent, under this caption: GOVERNMENT PAY ROLLS STABLE.

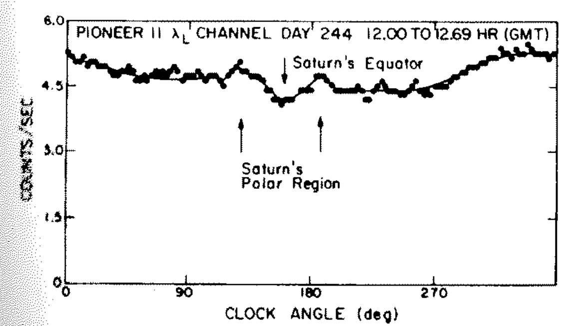


2.57 ZERO. The compulsion to include zero on a scale has ruined many graphs. Darrell Huff in *How to Lie With Statistics* argues the left graph is misleading, but the right graph is a waste of space that shows very little quantitative information beyond what could be conveyed in one sentence.

Huff's presumption is that viewers will not look at tick mark labels and apply the most trivial of quantitative reasoning. The result, the graph on the right in Figure 2.57, is a waste of space because the absurdly small aspect ratio makes it impossible to judge how spending depends on time, or even to look-up spending values along the vertical scale. The simple statement, "government payrolls were 19.5 million dollars in June and rose by about 4% from June to December" is much more incisive and efficient than this graph. The graph on the left in Figure 2.57 conveys more quantitative information; for example, we can use table look-up to determine from the left graph that the rise is roughly 4%.

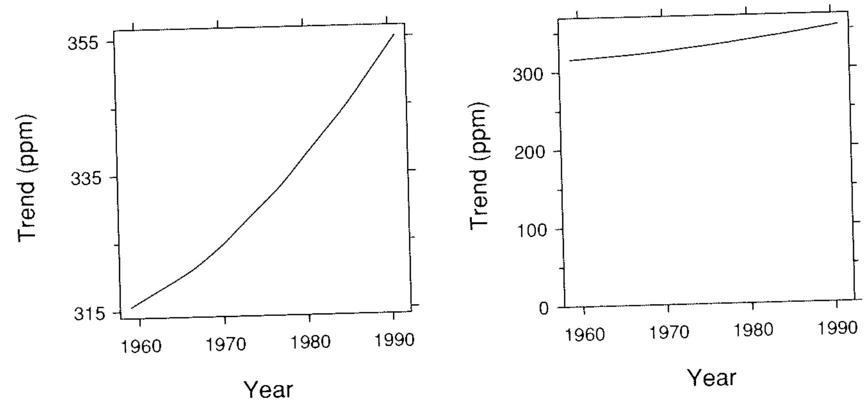
For graphical communication in science and technology *assume the viewer will look at the tick mark labels and understand them.* Were we not able to make this assumption, graphical communication would be far less useful. If zero can be included on a scale without wasting undue space, then it is reasonable to include it, but never at the expense of resolution.

The data in Figure 2.58 [66] are emission signals in the λ_L channel from Saturn and were measured by the Pioneer II spacecraft. Including zero on the vertical scale in Figure 2.58 has degraded our visual decoding of the quantitative information because the aspect ratio is too small. It is quite unlikely that a graph of these data with the vertical scale going from 4.0 counts/sec to 5.5 counts/sec would lead space physicists to think the percent variation in the emission signals is larger than it really is.



2.58 ZERO. The zero line on this graph interferes with our judgment of the data.

The left panel of Figure 2.59 graphs the CO_2 trend curve from Figure 2.56. The sensible thing has been done; there is no zero and the segments are banked to 45°. The right panel includes zero and the result is ridiculous, even worse, misleading because the increase in the rate of change of CO_2 with time is not readily perceived because the orientations of the segments that make up the curve are so close to 0°. Were we to attempt both banking to 45° and including 0, keeping the width of the data rectangle the same, the height of the scale-line rectangle would be 32 cm, clearly impractical.



2.59 ZERO. *Do not insist that zero always be included on a scale showing magnitude.* The left panel displays the trend curve sensibly; the curve is banked to 45°. The right panel, which includes zero, does not allow effective judgment of the change through time because the aspect ratio is too small.

When this zero issue is contemplated calmly, and examples such as the previous ones are given, it all seems quite simple. For the CO_2 data, one would expect the zero-line issue not to arise since the right panel of Figure 2.59 is such a preposterous graph. But it did arise, and in a forum of great importance. In March 1981 members of the U.S. Senate convened scientists for testimony on global warming. There was an exchange between U.S. Senator Albert Gore, Jr., trying to galvanize scientists and politicians into action on global warming, and N. Douglas Pewitt, a witness for the U.S. Department of Energy who was resisting action. The exchange is described by Stephen Schneider in his book *Global Warming* [108]. It needs no comment:

Gore responded that the Mauna Loa observatory data had clearly demonstrated increases in carbon dioxide and that it continued to show buildups. He pushed Pewitt further, "This is a rather impressive body of data that continues to accumulate. Doesn't that lead you to look at it in a different light?"

. . . Pewitt came back to the Mauna Loa CO_2 trend chart, calling it very deceptive: "It is a clever piece of chartology, in that it is intellectually accurate but can be subject to being read the wrong way." Gore was incredulous, and countered, pointing out that "if you look at a longer-range chart going back to 1880, you see virtually the same thing . . ." Gore continued, "Dr. Pewitt, I don't want to put words in your mouth, but I got the impression that you were saying the significance of the chart should be substantially discounted."

Pewitt responded, "No. What I said is that that is chartology. It is intellectually just exactly correct. It displays 315 going to 336, but it appears to be going from 0 to very large amounts."

Gore was stunned. "It is clearly labeled 316 to . . ."

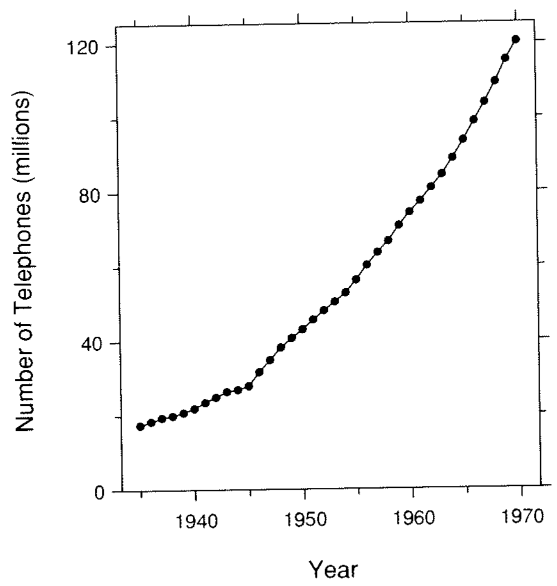
Pewitt interrupted, "That is correct, and a person very careful in their review of this chart will see that it is true. That is the reason why I said it was intellectually correct. I mean, it is objectively a statement of fact that is true. The impression it leaves on some people is quite a different impression. I have always had trouble playing with chartology. It appears to be going from nothing to a vast amount and that is a fact. We have a lawyer here who is nodding in agreement — it is true."

By this time most of us in the hearing room were giggling at this double talk. Gore simply read into the record the entire CO_2 emission chart from 1860 through 1980 and gave up the debate on that point.

Use a logarithmic scale when it is important to understand percent change or multiplicative factors.

There are some who feel that including a zero line on a graph helps us to better understand percent change and multiplicative factors. Darrell Huff [61] states that a graph with a zero baseline is beneficial "because the whole graph is in proportion and there is a zero line at the bottom for comparison. Your ten percent *looks* like ten percent."

It may well be that a zero line contributes a little to such judgments, but our ability to judge percents and factors is at best extremely poor. If we want to make such judgments it is far better to take logarithms. Suppose a, b, c, and d are all positive numbers with $a/b = c/d$ and b a few times bigger than d. Then on a graph of the four numbers it is quite hard to judge that the ratios are equal because on the graph, b is further from a than c is from d. This is illustrated in Figure 2.60. The data are the number of telephones in the U.S. from 1935 to 1970 [126]. The zero line is there, but it is very difficult to judge percents. Consider the following basic question: how is the percent increase in phones changing through time? For example, how does the percent change from 1935 to 1953, the middle of the time period, compare with the percent change from 1953 to 1970? It is very difficult to judge from Figure 2.60 without reading off values from the vertical scale and doing arithmetic.

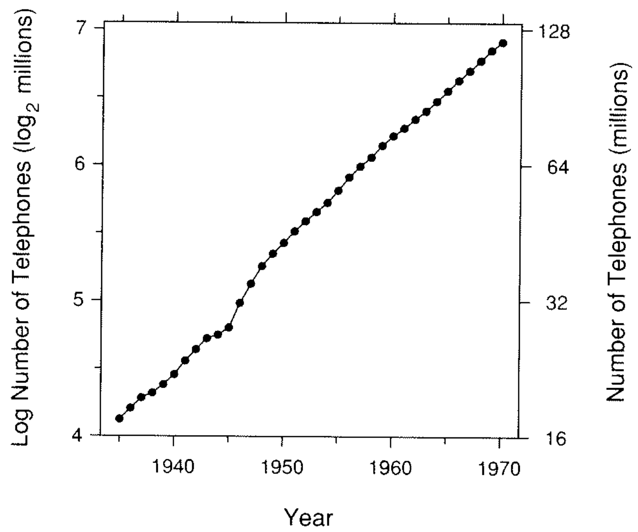


2.60 LOGS FOR FACTORS. The data are the number of telephones in the United States each year from 1935 to 1970. It is nearly impossible to judge whether the percentage increase is constant, decreasing, or increasing. The line segments connecting successive plotting symbols are banked to 45°.

When magnitudes are graphed on a logarithmic scale, percents and factors are easier to judge since equal multiplicative factors and percents result in equal distances throughout the entire scale. For our four numbers above,

$$\log(b) - \log(a) = \log(c) - \log(d) \ .$$

So $\log(b)$ is the same distance along the log scale from $\log(a)$ as $\log(c)$ is from $\log(d)$. This is illustrated in Figure 2.61. A log base 2 scale is used on the vertical scale for the telephone data. Now we can see that the percent increase in telephones through time has been roughly stable, since the trend in the data is roughly linear. Now we can see easily that telephones increased from 1935 to 1953 by about the same factor ($2^{1.5} \approx 2.8$) as they did from 1953 to 1970.



2.61 LOGS FOR FACTORS. *Use a logarithmic scale when it is important to understand percent change or multiplicative factors.* The data in Figure 2.60, are graphed by taking logarithms base 2. Now it is clear that the percentage increase in telephones was roughly stable from 1935 to 1970. The line segments connecting successive plotting symbols are banked to 45°.

Many phenomena obey multiplicative laws, even laws of aircraft engagement in battle, as Winston Churchill discovered. Churchill, known to all as a person of prodigious talent in matters of state, was remarkably facile in quantitative thinking. As an historian, he liberally used tables, graphs, and maps. For example in his epic account, *The Second World War*, there are 96 diagrams of military operations, 62 tables, 17 graphs, and 37 maps [19].

Churchill's discussion of the Battle of Britain in *Their Finest Hour*, the third volume of *The Second World War*, reveals the multiplicative law. This battle, fought between the Luftwaffe and the R.A.F. from July to October of 1940, until then a dark period for the British who suffered loss after loss, established British air superiority over the South of England and the English Channel. This was a turning point, a first demonstration that the German war machine was not invincible. A British loss in this air battle would have meant an imminent German invasion of the British Isles and a near-certain loss of the war. Churchill writes:

> Our fate now depended upon victory in the air. The German leaders had recognized that all their plans for the invasion of Britain depended on winning air supremacy above the Channel and the chosen landing places on our south coast.

Later in his account, Churchill analyzes the German and British losses, referring to a table of data that is reproduced in Figure 2.62:

> In cold blood, with the knowledge of the after-time, we may study the actual losses of the British and German Air Forces in what may well be deemed one of the decisive battles of the world. From the table on page 339 our hopes and fears may be contrasted with what happened.
>
> No doubt we were always oversanguine in our estimates of enemy scalps. In the upshot we got two to one of the German assailants, instead of three to one, as we believed and declared.

In the last sentence, Churchill characterizes the losses as "two to one" German losses to British losses and "three to one" estimated German losses to British losses. Presumably these numbers came from the totals at the bottom of his table:

$$\frac{\text{1733 actual German losses}}{\text{915 actual British losses}} = 1.89 \approx 2$$

$$\frac{\text{2698 estimated German losses}}{\text{915 actual British losses}} = 2.95 \approx 3\,.$$

Thus Churchill chose to quote factors of 2 and 3 rather than absolute differences of losses. Because

$$1733 - 915 = 818 \approx 800$$

and

$$2698 - 915 = 1783 \approx 1800$$

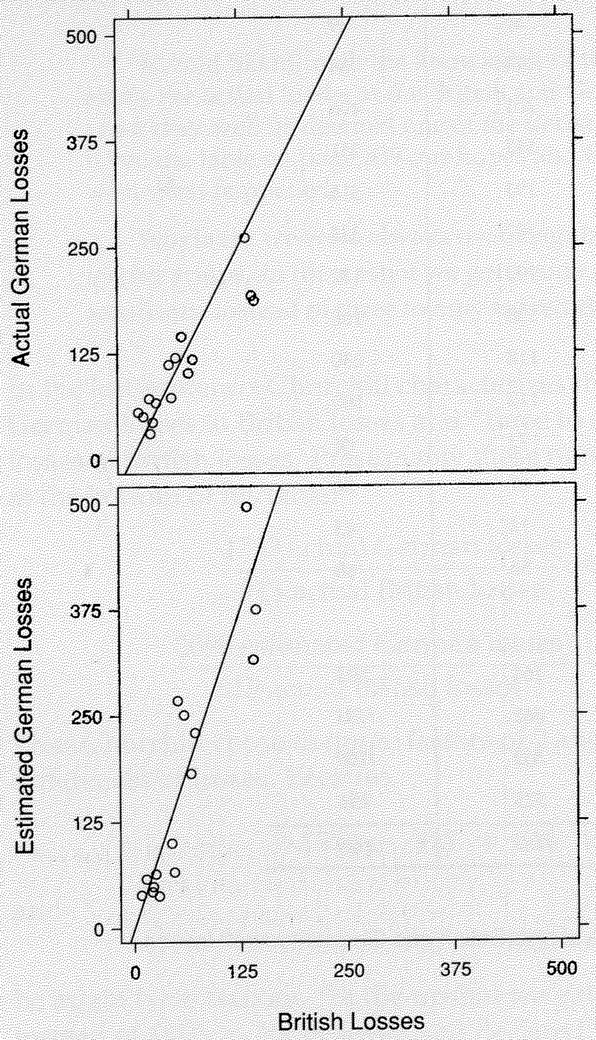
he might have written, "In the upshot we got 800 more than they, instead of 1800, as we believed and declared." But Churchill saw multiplicative factors as the right way to think.

AIRCRAFT LOSSES

	British Fighters Lost by R.A.F. (complete write-off or missing)	Enemy Aircraft Actually DESTROYED (according to German records)	Enemy Aircraft CLAIMED by us (Fighter Command, A.A., Balloons, etc.)
WEEKLY TOTALS:			
July 10–13	15	45	63
Week to July 20	22	31	49
Week to July 27	14	51	58
Week to Aug. 3	8	56	39
Week to Aug. 10	25	44	64
Week to Aug. 17	134	261	496
Week to Aug. 24	59	145	251
Week to Aug. 31	141	193	316
Week to Sep. 7	144	187	375
Week to Sep. 14	67	102	182
Week to Sep. 21	52	120	268
Week to Sep. 28	72	118	230
Week to Oct. 5	44	112	100
Week to Oct. 12	47	73	66
Week to Oct. 19	29	67	38
Week to Oct. 26	21	72	43
Oct. 27–31	21	56	60
MONTHLY TOTALS:			
July (from July 10)	58	164	203
August	*360*	*662*	*1133*
September	*361*	*582*	*1108*
October	136	325	254
Totals	*915*	*1733*	*2698*

2.62 LOGS FOR FACTORS. Churchill's table shows British and German aircraft losses in World War II.

Figure 2.63 demonstrates that Churchill's use of factors rather than differences was right on target. German losses, estimated and actual, are graphed against actual British losses; the line $y = 2x$ is superposed in the top panel and the line $y = 3x$ is superposed in the bottom. Clearly the relationships are multiplicative, an empirical law of air battle that prescribes ratios of losses are more stable than differences of losses, presumably because the latter inevitably depend on the size of the engagement. Thus it makes sense to take logs, which is done in

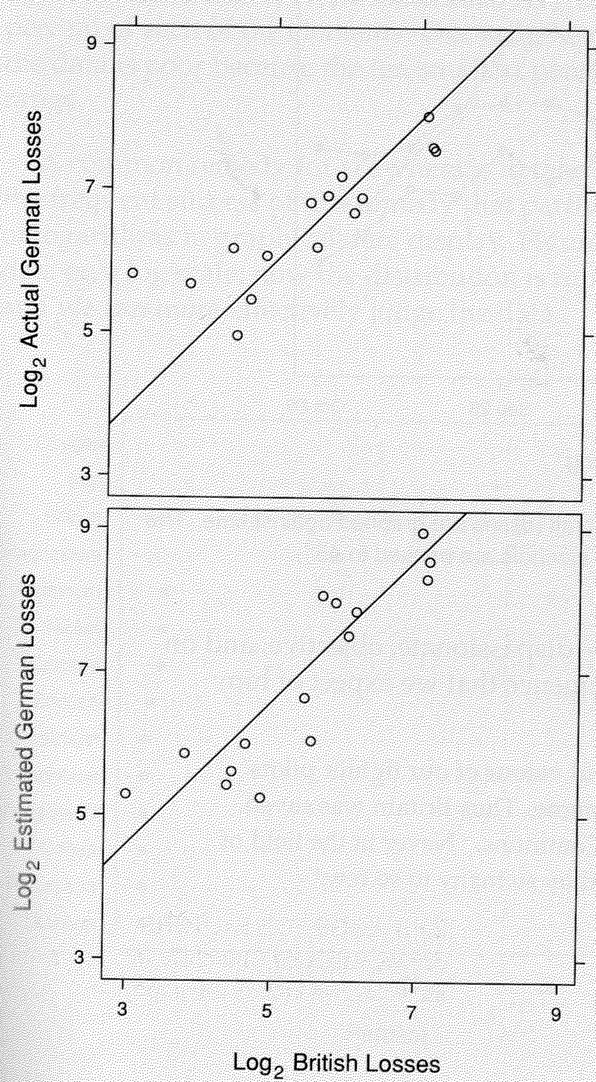


2.63 LOGS FOR FACTORS. The display validates Churchill's intuition that multiplicative factors are important.

Figure 2.64. The relationship $y = kx$ becomes

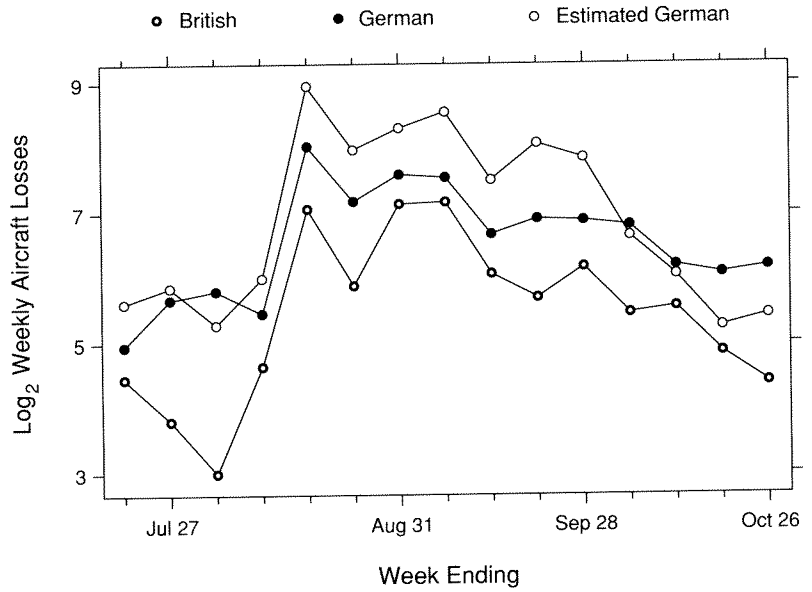
$$\log_2(y) = \log_2(k) + \log_2(x)$$

on the new graph. Thus a line with slope 1 and intercept $\log_2(2) = 1$ is superposed in the top panel, and a line with slope 1 and intercept $\log_2(3) = 1.58$ is superposed in the bottom panel.



2.64 LOGS FOR FACTORS. Logs are taken of Churchill's data since multiplicative factors are important.

It makes sense for us to think in terms of factors, as Churchill did, and take logs in looking further at these data. This is done in Figure 2.65 where weekly losses are graphed against time to see the course of the Battle of Britain as it unfolded during those crucial months.



2.65 LOGS FOR FACTORS. The logs of aircraft losses are graphed against time. The line segments connecting successive plotting symbols are banked to 45°.

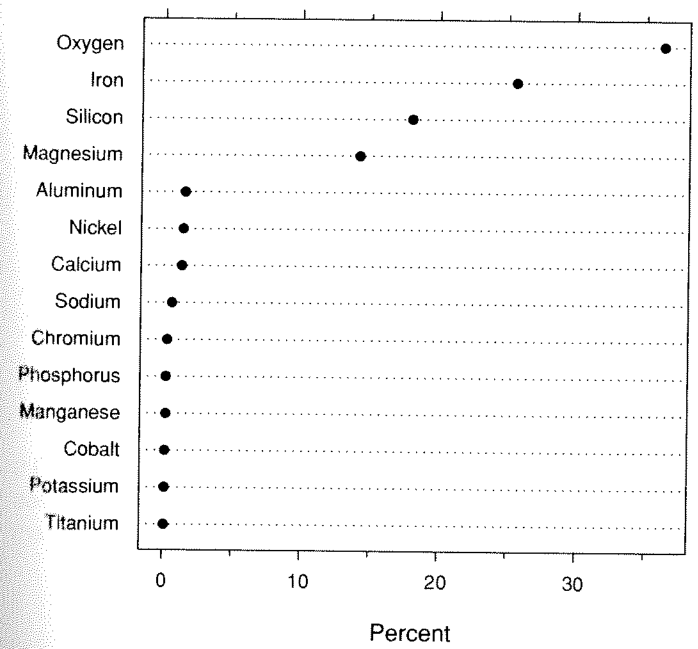
Churchill's quantitative analysis — dispassionate, objective, and on target — then concludes with the eloquence that we expect of him:

> At the summit the stamina and valour of our fighter pilots remained unconquerable and supreme. Thus Britain was saved. Well might I say in the House of Commons, "Never in the field of human conflict was so much owed by so many to so few."

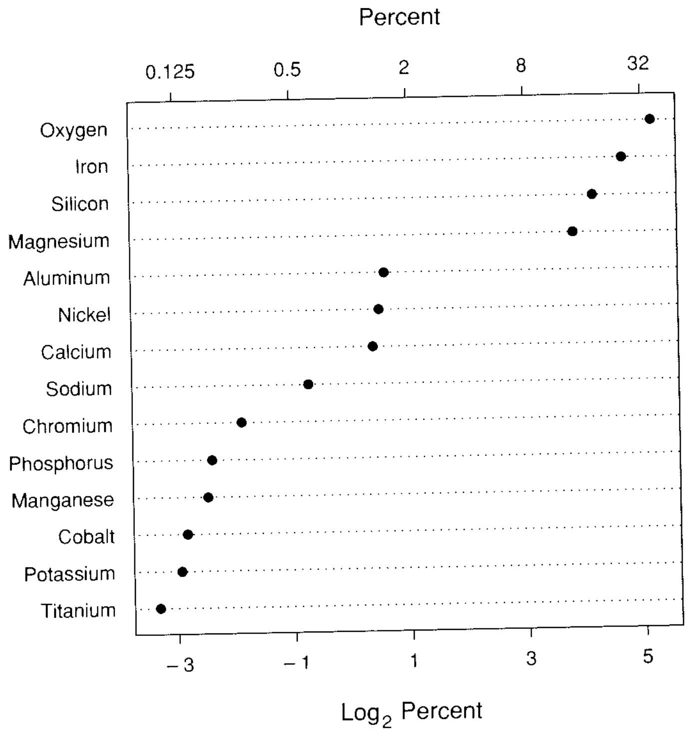
Showing data on a logarithmic scale can cure skewness toward large values.

It is common for positive data to be *skewed toward large values*: some values bunch together at the low end of the scale and others trail off to the high end with increasing gaps between the values as they get higher. Severe skewness causes most of the data to be squashed into a narrow interval, which degrades our judgment of the data. An example of skewed data is given in Figure 2.66. The graph shows the 14 most abundant elements in stone meteorites [47]; the data are the average percent of each of the elements. Our judgment of the relative values of the data is poor because the ten smallest percents vary over a small range.

A common remedy for skewness is to take logarithms. Indeed, it is the frequent success of this remedy that partly accounts for the large use of logarithms in graphical data display. Figure 2.67 shows the meteorite data on a log scale; now the distribution is much more nearly uniform and we can more effectively judge the data.



2.66 LOGS FOR SKEWNESS. Because the data on this graph are severely skewed, we cannot accurately judge the relative values of the data.

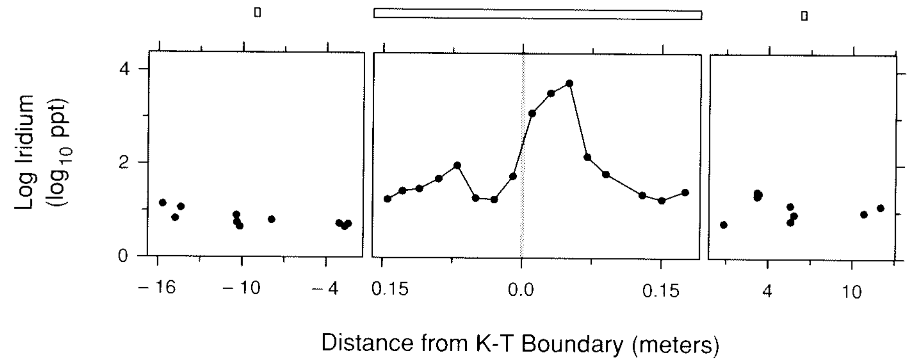


2.67 LOGS FOR SKEWNESS. *Showing data on a logarithmic scale can cure skewness toward large values.* The logs of the data in Figure 2.66 are graphed, which has eliminated the skewness.

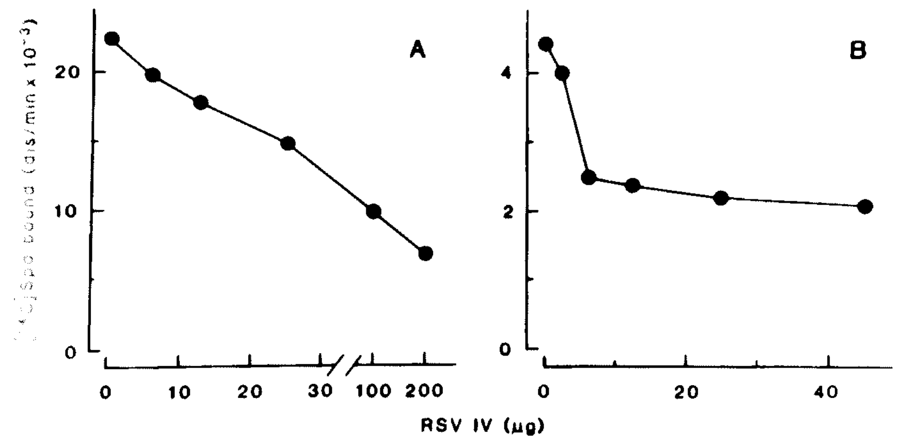
Use a scale break only when necessary. If a break cannot be avoided, use a full scale break. Do not connect numerical values on two sides of a break. Taking logs can cure the need for a break.

Figure 2.68 shows the iridium data discussed in Section 2.3 (pp. 54–66). Two *full scale breaks* are used to signal changes on the horizontal scale; the middle panel has a smaller number of data units per cm. The widths of the rectangles at the top of the graph portray the same number of data units on the panels.

A change or gap in a scale is shown forcefully by a full break. Some indicate a change or gap in the scale of a graph by a *partial scale break*: two short wavy parallel curves or two short parallel line segments breaking a scale line. This is illustrated on the horizontal scale line of the left panel in Figure 2.69 [101]. But the partial scale break is a weak indicator that the reader can fail to appreciate fully; visually, the graph is still a single panel that invites the viewer to see patterns between the two scales.

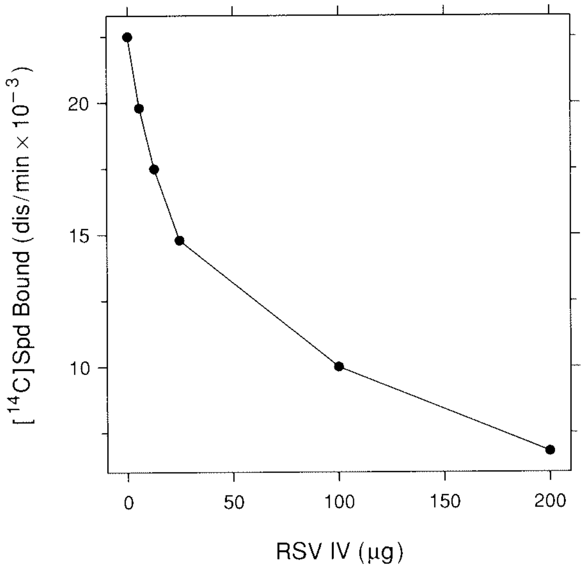


2.68 SCALE BREAKS. *Use a scale break only when necessary. If a break cannot be avoided, use a full scale break. Do not connect numerical values on two sides of a break. Taking logs can cure the need for a break.* This graph uses full scale breaks on the horizontal scale to signal changes in the number of units per cm. The full breaks show the scale breaks forcefully. Without the breaks, the data in the center panel would lie very nearly on a vertical line and there would be no time resolution. The rectangles at the top of the graph portray the same number of horizontal scale units on each panel.



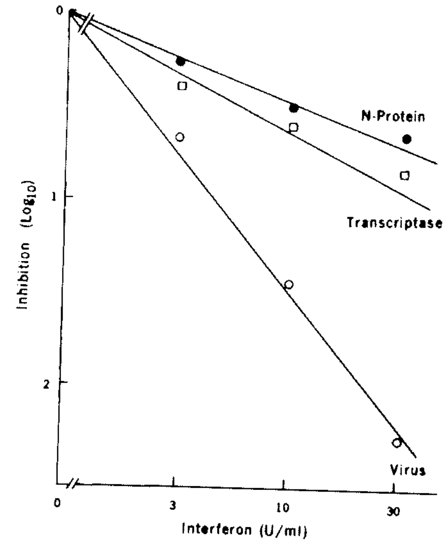
2.69 SCALE BREAKS. The partial scale break on the horizontal scale of the left panel does not give a forceful indication of a break. The connection of numerical values across the break gives the misleading impression that the data are roughly linear.

Numerical values should not be connected across a break. In the left panel of Figure 2.69, the connection across the break gives the misleading impression that the data are roughly linear across the entire horizontal scale; in fact the slope of the values decreases as the variable on the horizontal scale increases, as shown by Figure 2.70, which graphs the data with no scale break.

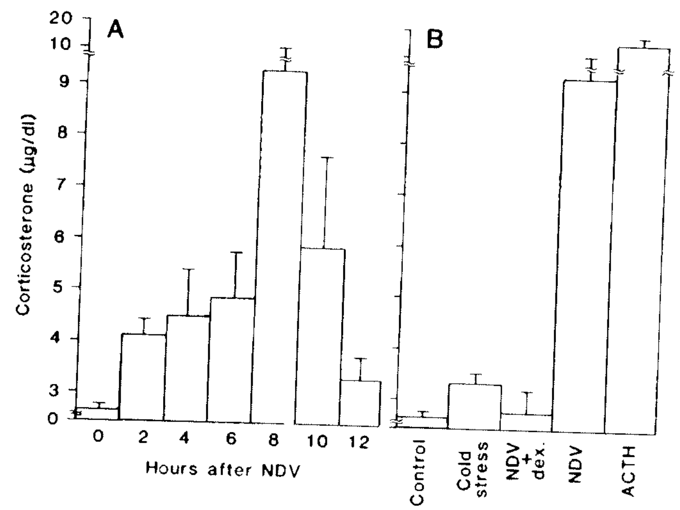


2.70 SCALE BREAKS. The data from the left panel of Figure 2.69 are graphed without a scale break. Now it is clear that the data are not roughly linear and that the slope decreases as the variable on the horizontal scale increases. The line segments connecting successive plotting symbols are banked to 45°.

Figures 2.71 and 2.72 show other bad breaks. Figure 2.71 [87] gives a misleading impression because the continuation of the lines across the break has no meaning. The tick marks on the horizontal scale are labeled 3, 10, and 30; since the logarithms of these values are nearly equally spaced, the authors presumably intended a horizontal log scale. The three lines give the impression that the pattern of each data set is linear through the origin. But a value of zero U/ml of interferon is off at minus infinity on the horizontal log scale, so the three lines could not possibly go through the origin. In Figure 2.72 [111] bars and error bars are allowed to barge right through two scale breaks. The bar lengths and areas, important and prominent visual aspects of the graph, are meaningless.

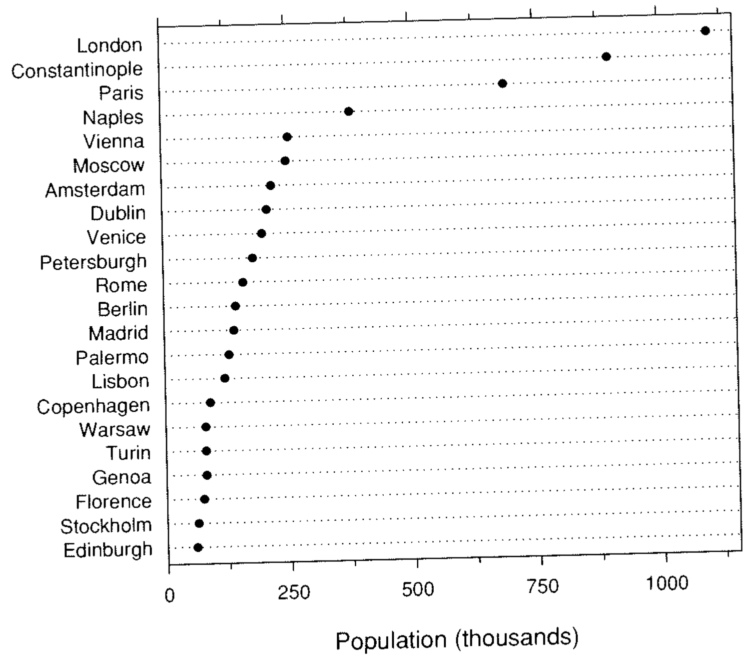


2.71 SCALE BREAKS. On this graph the lines drawn through the partial scale break have no meaning and give the misleading impression that the pattern of the data goes linearly through the origin. Since the horizontal scale is logarithmic, zero is actually at minus infinity.

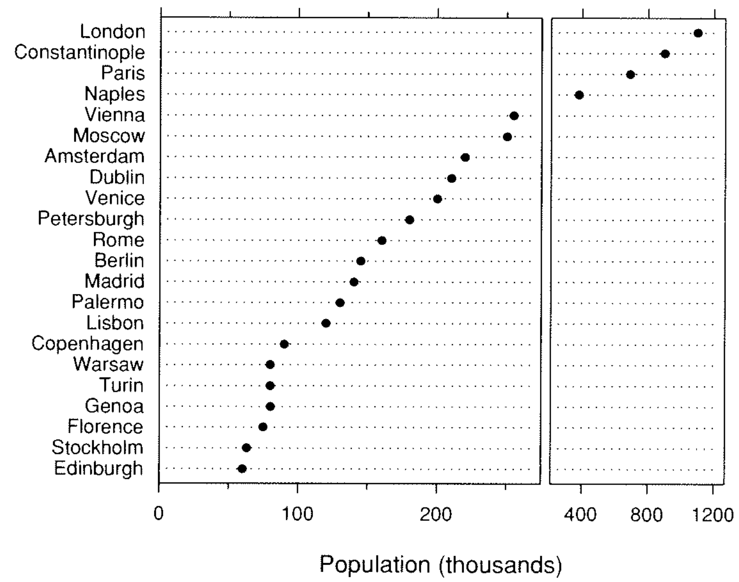


2.72 SCALE BREAKS. The lengths of the bars that barge right through the scale breaks have no meaning.

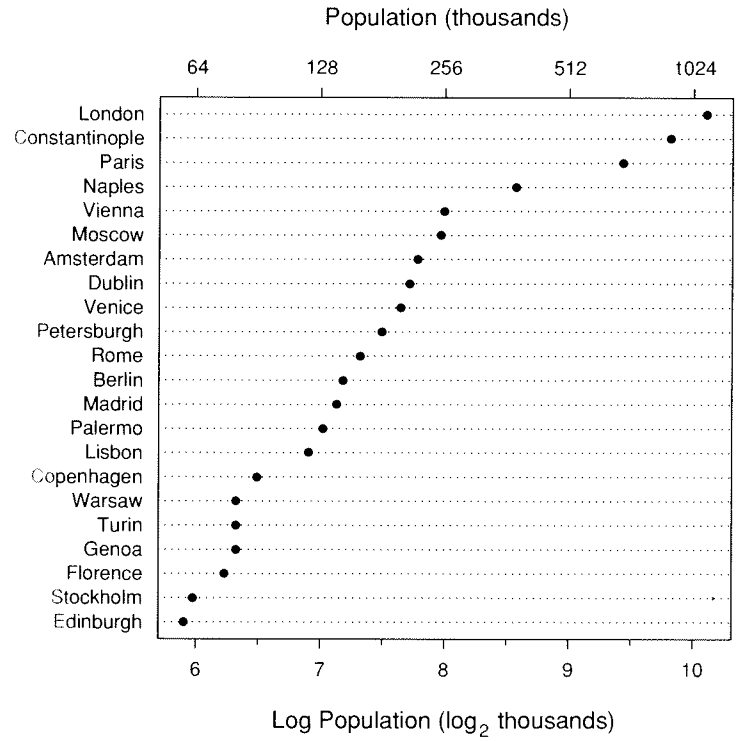
Full scale breaks should be used only when necessary. Figure 2.70 shows the break of Figure 2.69 is not needed. Taking logarithms of the data can often relieve the need for a scale break. Figure 2.73 shows data from William Playfair's *Statistical Breviary* [105], published in 1801. The data, which are the populations of 22 European cities, are forced into a small region of the scale. Figure 2.74 graphs the data with a break. True, the data are now not squashed, but we have paid a great price. The values in the right panel cannot be graphically compared with those on the left. The best we can do is a tedious table look-up by reading the scales of each panel. Figure 2.75 uses a log scale, a better solution than the broken scale because all of the data can be readily visually compared.



2.73 LOGS. The dot plot graphs populations of 22 European cities. The data are skewed to the right, which degrades the resolution of the graphed values.



2.74 LOGS. The populations are graphed with a scale break. This prevents graphical comparison of all values.



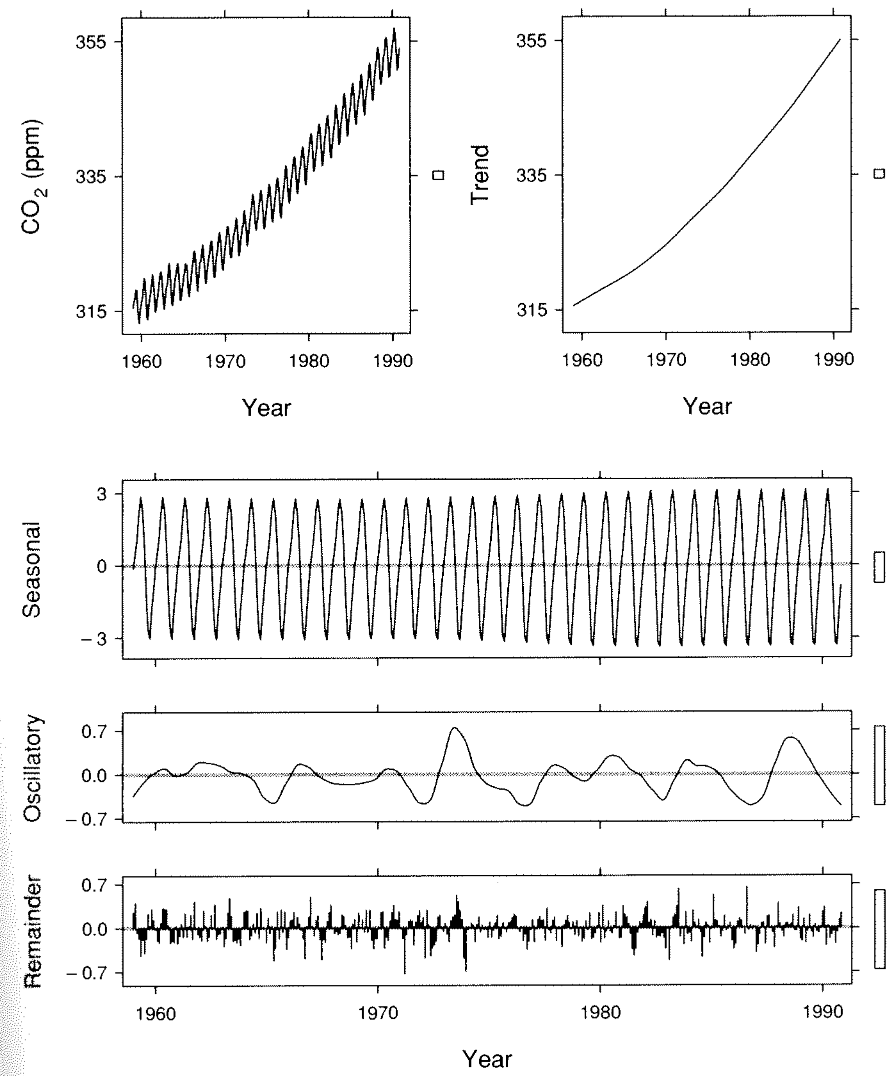
2.75 LOGS. The populations are graphed on a log scale, which relieves the need for a scale break.

2.6 General Strategy

Graphing is much like writing. Our written language has grammatical and syntactical rules that govern the details of word and sentence construction; most of the graphical principles in the previous sections — Clear Vision, Clear Understanding, Banking to 45°, and Scales — are analogous to these rules. But there are also more general guidelines — that is, overall strategies — for writing; these are more nebulous rules aimed at producing clear, interesting prose. For example, William Strunk Jr. and E. B. White [115] encourage clarity by "Use definite, specific, concrete language," and encourage brevity by "Do not overwrite." The first two principles of this chapter — make the data stand out and avoid superfluity — are general strategies for graphs. (Note the similarity between the two Strunk and White principles and these two general graphical principles. Edward R. Tufte once made the insightful remark that Strunk and White's book on the elements of writing is one of the best treatises on graphing data.) In this section several general strategies for graphing data are discussed.

A large amount of quantitative information can be packed into a small region.

Previous principles in this chapter have stipulated that graphs should not be cluttered and should not have superfluous elements, but this does *not* preclude a large amount of quantitative information being shown on a graph, even a small graph. It is possible to put a large data set on a graph in an uncluttered way. Figure 2.76, the graph of the CO_2 data and its four components that we have seen before, is an example. There are 396 monthly data points on each of the panels of this graph, which is 1920 points altogether. Each data point consists of two numbers, a value on the horizontal scale and a value on the vertical scale. Thus 3840 numbers are shown on this graph.



2.76 PACKING DATA. *A large amount of quantitative information can be packed into a small region.* The computer graphics revolution has given us the capability to graph a large amount of quantitative information in a small space. There are 1920 data points on this graph; each portrays two numerical values, so 3840 numbers are shown.

Graphing data should be an iterative, experimental process.

Iteration and experimentation are important for all of data analysis, including graphical data display. Graphing needs to be iterative because a graph can help discover unknown aspects of the data, and once the unknown is known, we frequently find ourselves formulating a new question about the data. Even when we understand the data and are graphing them for presentation, a graph will look different from what we had expected; our mind's eye frequently does not do a good job of predicting what our actual eyes will see.

Figure 2.77 is a simulation of an actual graph session and its iteration of graph making as it might have occurred in real life. The data are the number of doctorates in the physical sciences and in the mathematical sciences in the United States each year from 1960 to 1981 [96].

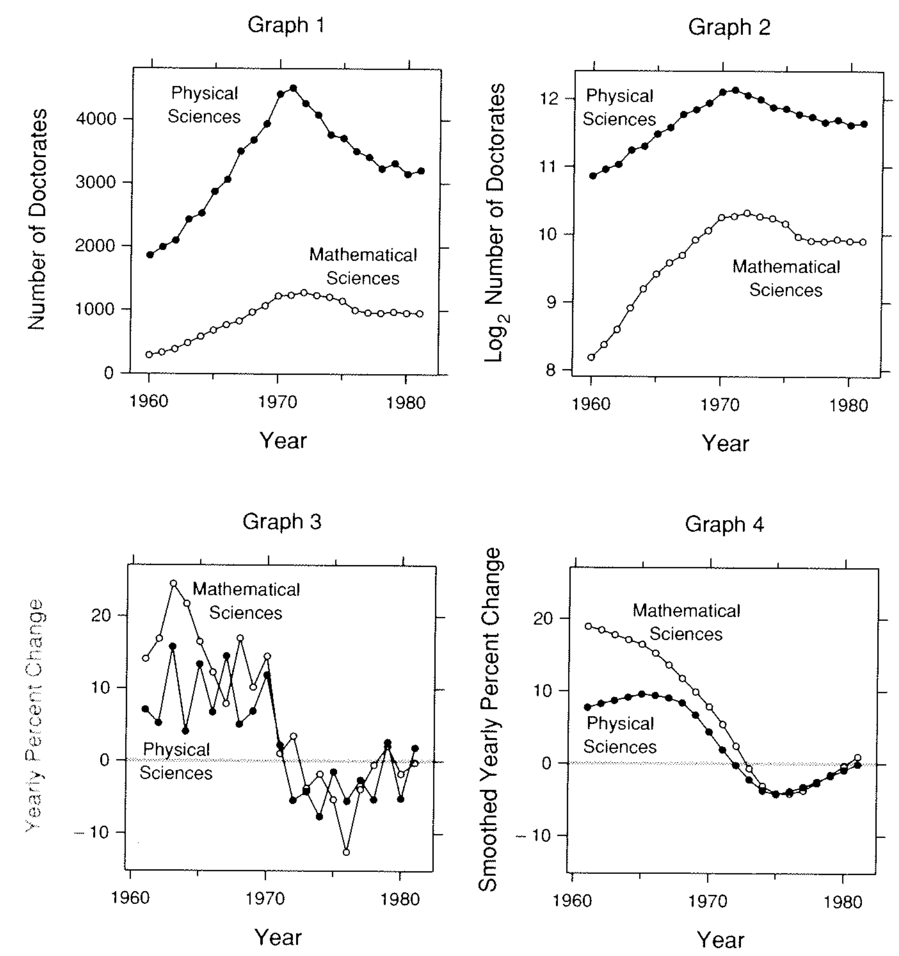
The first try, Graph 1, is a reasonable one and shows each data set graphed against time. We can see similar trends in both series; there is a rise to a peak just after 1970 and then a decline. The rise and decline for the physical sciences is greater, but the number of doctorates in the physical sciences is greater. This prompts asking how the percent changes in the two series compare; the response is Graph 2, where the logarithms of the data are shown. The graph suggests that in the early years the percent increases in the mathematical science degrees are greater, but that starting in the late 1960s the percent changes are similar.

Graph 2 allows us to study percent change between any two values. However, if we want to see just year-to-year percent change, graphing these values directly can give us a more incisive look. This has been done in Graph 3. The values confirm our impression of the overall trend in yearly percent change shown in Graph 2, but they also show more precise quantitative values — for example, we can see that the yearly increases in physical science doctorates oscillated around 10% in the early years.

One problem with Graph 3 is a large amount of year-to-year fluctuation that interferes somewhat with our ability to judge the overall trends. One solution is to smooth the data. Graph 4 shows the data after smoothing by the numerical procedure loess, which will be described in Section 3.7 (pp. 168–180). The distracting fluctuations have been removed and now we can see that in 1960 the percent increase in

mathematical science doctorates was about double that for the physical sciences, but that the trends in the two sets of rates grew closer and became virtually identical after about 1975.

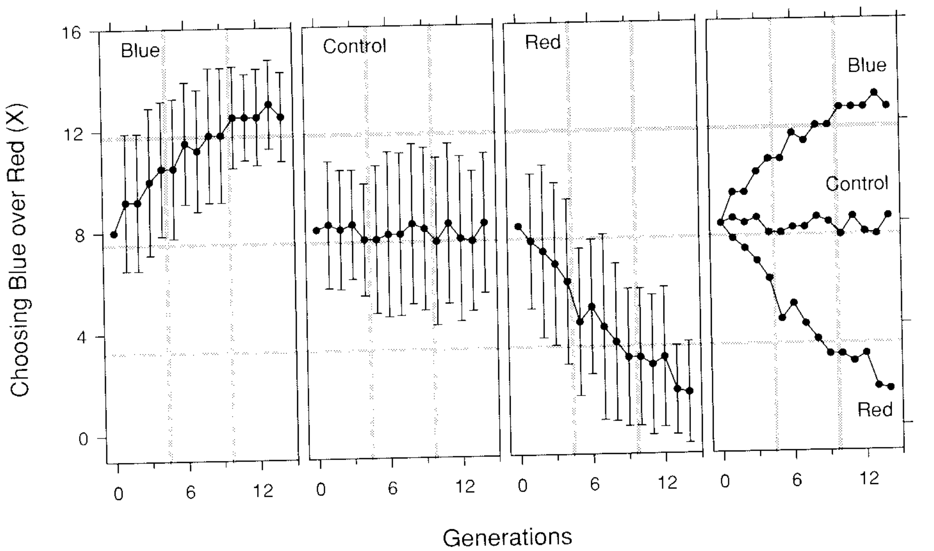
This depiction in Figure 2.77 of graph iterations is actually oversimplified. It is likely that in a real-life graphing of these data the choice of plotting symbols, the placement of the data labels, and the choice of the amount of smoothing would require several more iterations.



2.77 ITERATION. *Graphing data should be an iterative, experimental process.* The four graphs in this figure are four successive looks at the data; each of the last three is inspired by its predecessor.

Graph data two or more times when it is needed.

A corollary of the previous principle on iteration is that, whether we are in the mode of analyzing data or presenting data to others, we should not hesitate to make two or more graphs of the same data. Two different ways of graphing data sometimes bring out aspects that only one way cannot. For example, in a presentation of the doctorate degree data of Figure 2.77, it would be entirely sensible to use Graph 2 and Graph 4; both show interesting aspects of the data. Figure 2.78, shown earlier in Sections 2.2 (pp. 30–55) and 2.4 (pp. 63-73), is another example. Each of the three sets of data is shown twice. Graphing each data set separately in the left three panels allows the error bars to be perceived without interfering with one another. Graphing the three data sets together in the right panel allows them to be more effectively compared.



2.78 REGRAPHING. *Graph data two or more times when it is needed.* Each data set is graphed twice, once in one of the three left panels to allow an unobstructed view of the error bars and once in the right panel to allow an effective comparison of the data sets.

Many useful graphs require careful, detailed study.

There are some who argue that a graph is a success only if the important information in the data can be seen within a few seconds. 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 next two graphs illustrate these extremes.

Cyril Burt was a giant in psychology until his world began to crumble in 1974, three years after his death. Burt was one of the leading proponents of the theory that intelligence, as measured by IQ scores, is largely inherited. Burt's data strongly supported this view — too strongly, as it turns out. In 1974 suspicions were raised about the authenticity of some of Burt's data and his analyses [70]. For five years doubts about Burt's integrity grew, culminating in a biography by Hearnshaw who concluded, as others already had, that Burt faked much of his data, invented collaborators, and sent letters to journals from fictitious people who supported his work [58].

Table 2.1 shows data that Burt published in 1961 in the *British Journal of Statistical Psychology* [15]. The numbers are part of a larger data set that were widely quoted in subsequent scientific work until D. D. Dorfman, a psychologist at the University of Iowa, gave a convincing argument in 1978 that the numbers were made-up, either in whole or in part [46]. The values in Table 2.1 were purported to be mean IQ scores of 40,000 father-child pairs divided into six social classes.

Table 2.1 CYRIL BURT DATA.

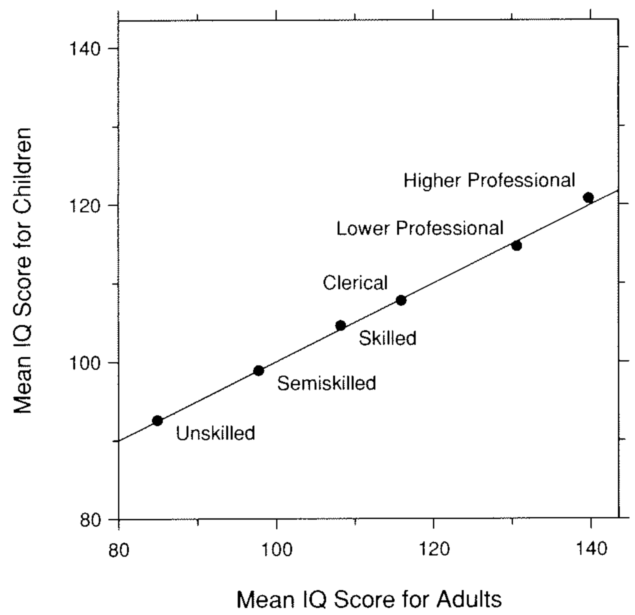
	Adult Mean IQ	Child Mean IQ
Higher Professional	139.7	120.8
Lower Professional	130.6	114.7
Clerical	115.9	107.8
Skilled	108.2	104.6
Semiskilled	97.8	98.9
Unskilled	84.9	92.6

The data in Table 2.1 look innocent enough until they are graphed. Figure 2.79 is a graph of the mean scores for the children against the corresponding values for adults. The impugnment of these data is based, in part, on the notion that the mean scores are simply too good to be true. In 1959, J. Conway [39] had put forward the equation

$$\text{child score} - 100 = 0.5\,(\text{adult score} - 100)$$

as a method for predicting the mean IQ score of children in a given class from the mean IQ score of the fathers in the class; this predictive line is shown in Figure 2.79. The line lies extraordinarily close to the data. Thus for Burt's data, Conway's predictive method, with its mathematically elegant coefficient of 0.5, makes nearly perfect predictions.

Figure 2.79 requires only a quick look to absorb the important quantitative information. The main message — that the mean scores are very close to the line — can be absorbed almost instantaneously.



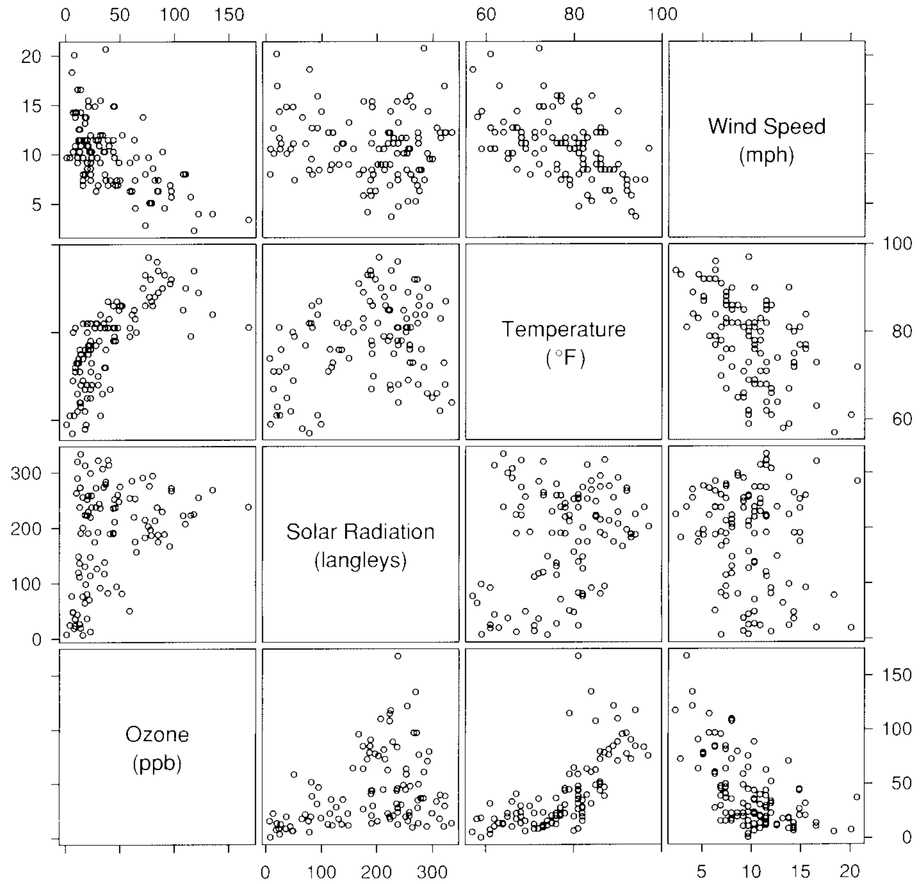
2.79 DETAILED STUDY. The important information on this graph, that Cyril Burt's fictitious data lie very close to the line, can be extracted with just a quick look.

Some graphs, however, require long and detailed scrutinizing. This is entirely reasonable. 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. If a graphical display requires hours of study to make a discovery that would have gone undetected without the graph, then the display is a success.

Figure 2.80 is a graph that requires detailed study. The graphical method used in the figure, an exceedingly useful one called a *scatterplot matrix*, will be discussed in Section 3.9 (pp. 193–197). The data in Figure 2.80 are measurements of four variables: wind speed, temperature, solar radiation at ground level, and concentrations of the air pollutant, ozone [13]. There is one measurement of each variable on each of 111 days.

Each panel of Figure 2.80 is a scatterplot of one variable against another. For the three panels in the bottom row, the vertical scale is ozone, and the three horizontal scales are solar radiation, temperature, and wind speed. So the graph in position (2,1) in the matrix — that is, the second column and first row — is a scatterplot of ozone against solar radiation; position (3,1) is a scatterplot of ozone against temperature; position (4,1) is a scatterplot of ozone against wind speed.

The scatterplot matrix reveals much about the four variables. A discussion of what is seen, since it is long and detailed, will be postponed to the full discussion of scatterplot matrices in Section 3.9 (pp. 193–197); it suffices to say here that the revelations come only after careful, detailed study of the graph. It might well be expected that a graph with 1332 points on it, each encoding two numbers for a total of 2662 numbers, would require careful study.



2.80 DETAILED STUDY. *Many useful graphs require careful, detailed study.* Compared with that needed for Figure 2.79, this scatterplot matrix of ozone and meteorological measurements requires lengthy study to extract the information. But the lengthy study reveals information that would be very difficult or impossible to get by other means.

3 Graphical Methods

This chapter is about graphical methods: types of graphs and ways of encoding quantitative information on graphs. The methods allow us to analyze both the overall structure of the data and the detail of the data.

Section 3.1 (pp. 120–126) discusses the logarithm, a basic tool that is useful in all areas of graphical data analysis.

Section 3.2 (pp. 126–132) is about another basic tool of data display — graphing residuals from the fit of a mathematical function to a set of data.

Section 3.3 (pp. 132–149) is about distributions. Suppose we have one or more sets of measurements of a single quantitative variable. The graphical methods of the section display the distribution of the data — where the measurements lie along the measurement scale.

Section 3.4 (pp. 150–154) is about dot plots, which display measurements of a quantitative variable where each measurement has a ·abel. When the labels are a cross-classification of two or more ·ategorical variables, the display becomes a multiway dot plot.

Section 3.5 (pp. 154–165) is about plotting symbols and curve types on ·aphs with two quantitative variables, a mundane issue for data display ·it an issue filled with problems that need methods. In Section 2.2 ·p. 25–54) we saw that two-variable graphs can easily fail because ·aphical elements are obscured or different data sets are not easily ·ually assembled. The methods of the section attack these problems.

·ection 3.6 (pp. 166–167) shows how visual reference grids enhance · comparison of data on the different panels of a graph with two or ·re juxtaposed panels.

·ection 3.7 (pp. 168–180) is about loess, a method for fitting curves to ·terplots. When the purpose of a scatterplot is made to study how ·variable, a response, depends on another variable, a factor, noise in