2

PRINCIPLES OF GRAPH CONSTRUCTION

This chapter is about the basic elements of graph construction — scales, legends, plotting symbols, reference lines, keys, labels, panels, markers, 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 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.

In this chapter there are many examples of graphs from science and technology that have problems. Such problems are pervasive because graphing data is a complex task. (See Sections 3 and 4 of Chapter 1.) The principles are applied to the examples to show how the problems can be solved.

Section 2.1 defines terms. Section 2.2 gives principles that make the elements of a graph visually clear, and Section 2.3 gives principles that contribute to a clear understanding of what is graphed. Section 2.4 is about scales, and Section 2.5 discusses principles that are general strategies for graphing data. Finally, Section 2.6 lists the principles of graph construction given in the previous sections.

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 same data in two different ways; the words in boldface convey the terminology. For the most part, the terms are self-explanatory, but a few comments are in order.

In Figures 2.1 and 2.2 the data are the percent changes from 1950 in death rates in the United States due to cardiovascular disease and due to all other diseases [87]. In Figure 2.1 the two data sets are *superposed* and in Figure 2.2 they are *juxtaposed*. The *marker* along the horizontal scale on each graph 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 in the *data regions*.

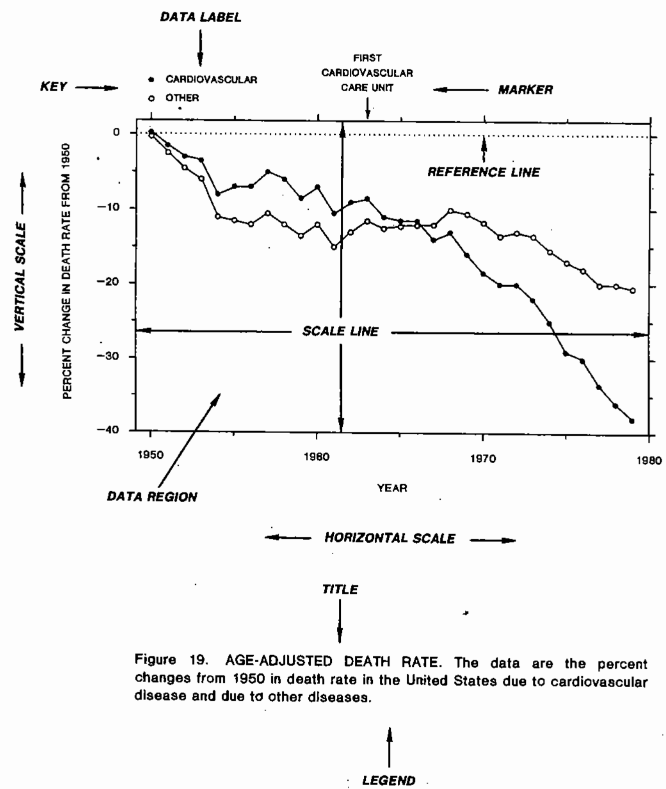

Figure 2.1 TERMINOLOGY. This figure and the next define terminology. The two sets of data — death rates due to cardiovascular disease and death rates due to all other diseases — are superposed. The data labels are in the key on this graph.

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

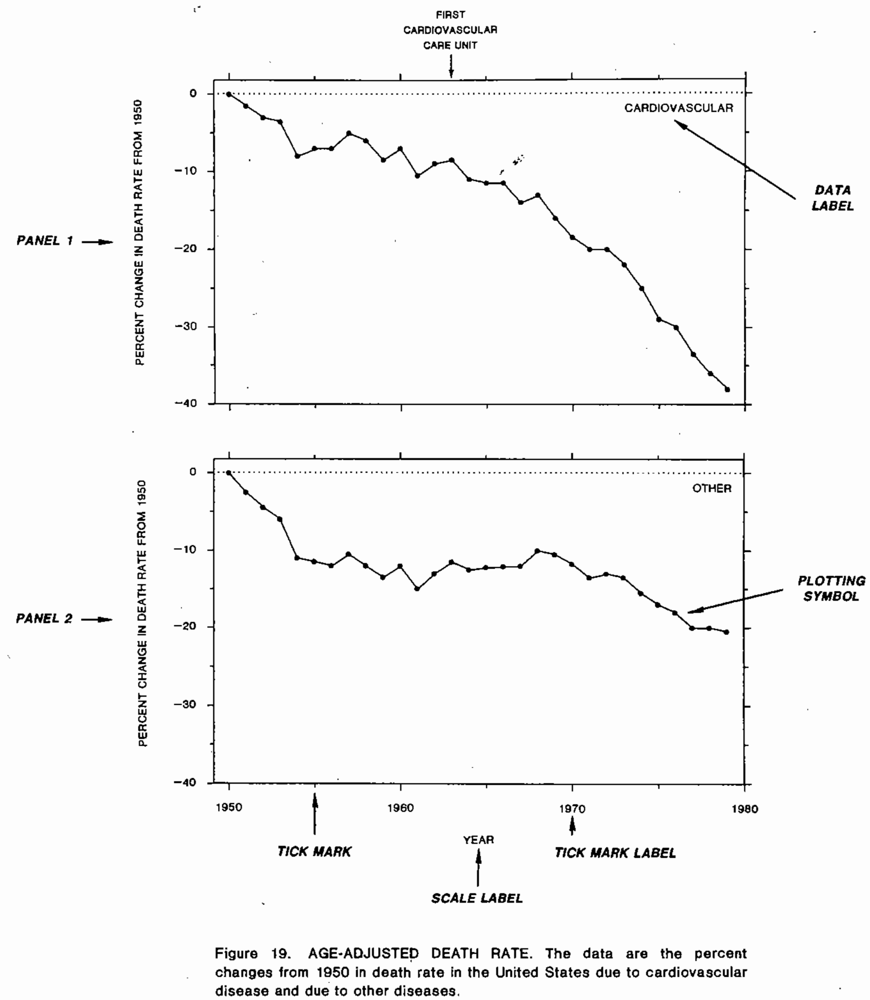

Figure 2.2 TERMINOLOGY. This figure also defines the meaning of terms. The two sets of data are juxtaposed by using two panels. Each panel on this graph has a data label.

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 graphs. 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 — the quantitative and qualitative information in the data region — are the reason for the existence of the graph. The data should stand out. It is too easy to forget this. One of the major problems uncovered in the studies of graphs in scientific publications described in Section 4 of Chapter 1 was the data not standing out. 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* [123].

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 [80]. 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.

In 1980 two Harvard anthropologists, Melvin Konner and Carol Worthman, put forward a likely solution to the puzzle [80]. 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 tall 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 legend is necessary. A minor problem with Figure 2.3 is that the arrows for picking up and setting down are superfluous.

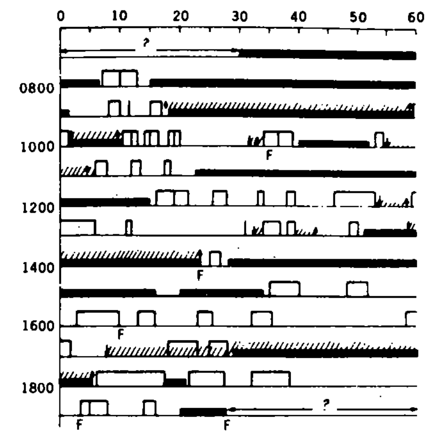

Figure 2.3 SUPERFLUITY AND STANDING OUT. The graph shows the activities of a !Kung woman and her baby. The open bars and tall vertical lines are nursing times; the closed bars show times when the baby is sleeping; F means fretting; and slashed lines are intervals when the baby is held by the mother,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

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 legend 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.

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.

Use visually prominent graphical elements to show the data.

On the graph in Figure 2.5 [25] the data do not stand out. 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.

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; scales issues such as these will be discussed in Section 2.4.

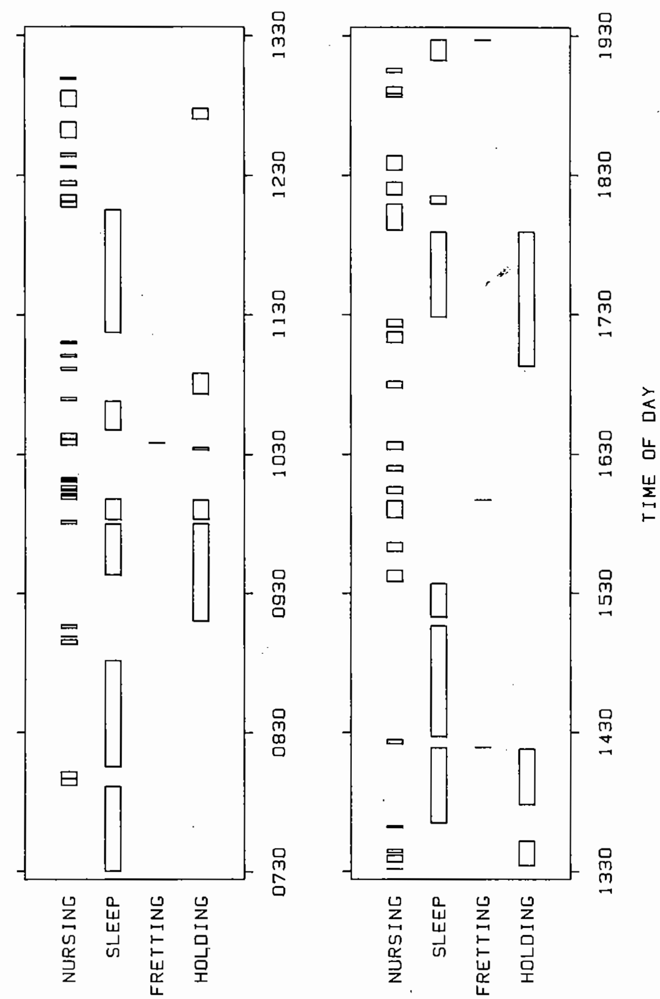

Figure 2.4 **SUPERFLUITY AND STANDING OUT.** ***Make the data stand out. Avoid superfluity.*** **These are two broad principles that guide the specific principles to follow in this section. The data from Figure 2.3 are regraphed. It is now easier to see the activity times and their interactions, constant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

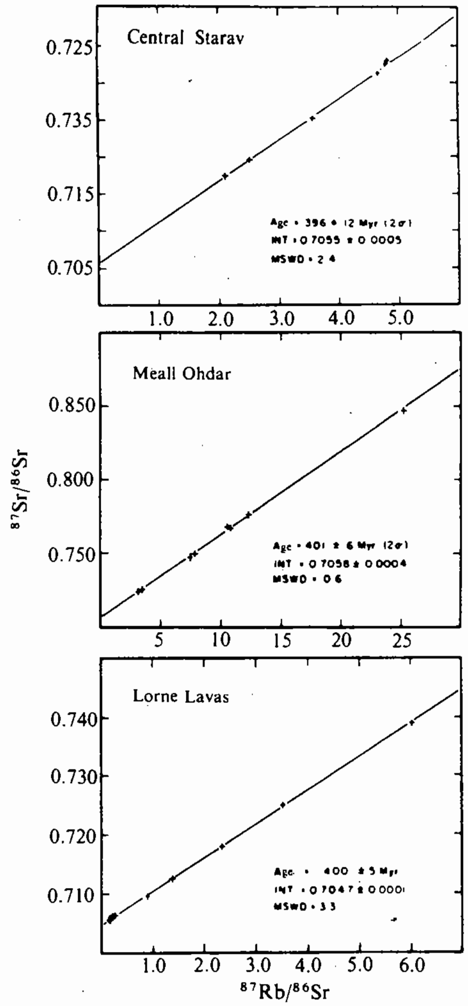

Figure 2.5 VISUAL PROMINENCE. The data do not stand out.
Figure republished by permission from Nature [25]. Copyright © 1983 Macmillan Journals Limited.

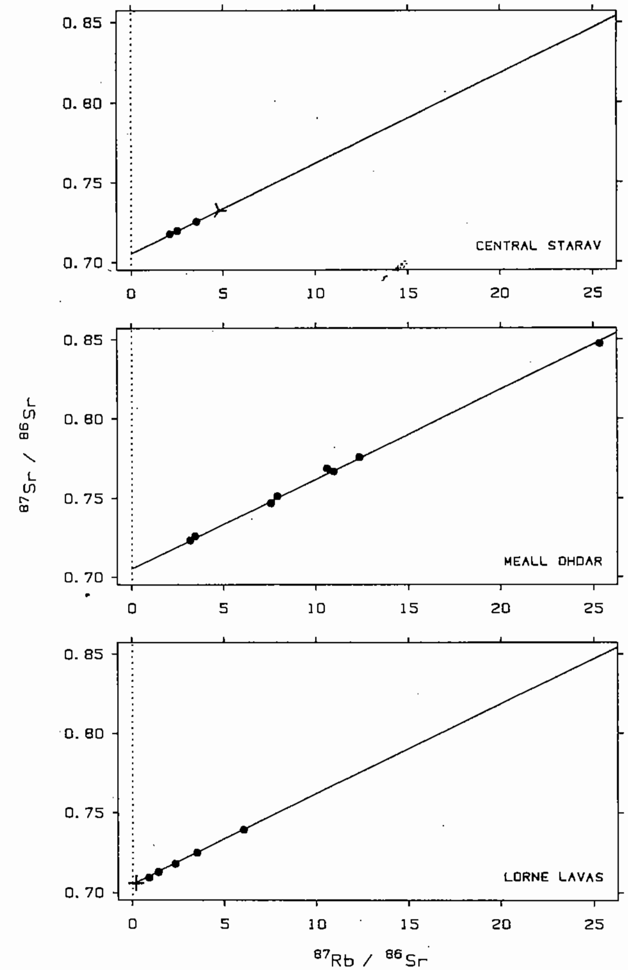

Figure 2.6 VISUAL PROMINENCE. *Use visually prominent graphical elements to show the data.* Now the data from Figure 2.5 can be seen. The symbols that look like the spokes of a wheel represent multiple points; each spoke is one observation.

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 [17]. 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 [56]. The graph shows good news: After the ban on the use of DDT, the average number of young per site began increasing.

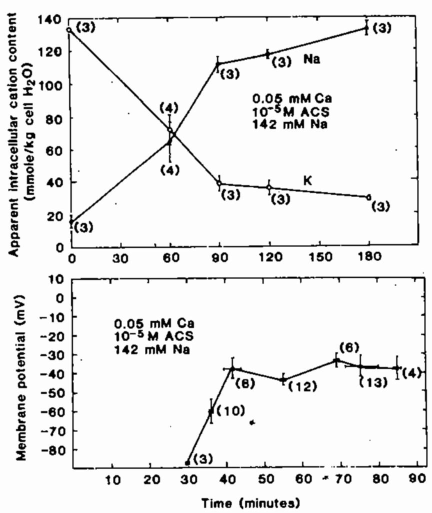

Figure 2.7 VISUAL PROMINENCE. The data on this graph do not stand out because the graphical elements showing the observations and their standard errors are not prominent enough to prevent being obscured by the connecting lines.
Figure republished from [17]. Copyright 1983 by the AAAS.

Use a pair of scale lines for each variable. Make the data region the interior of the rectangle formed by the scale lines. Put tick marks outside of the data region.

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

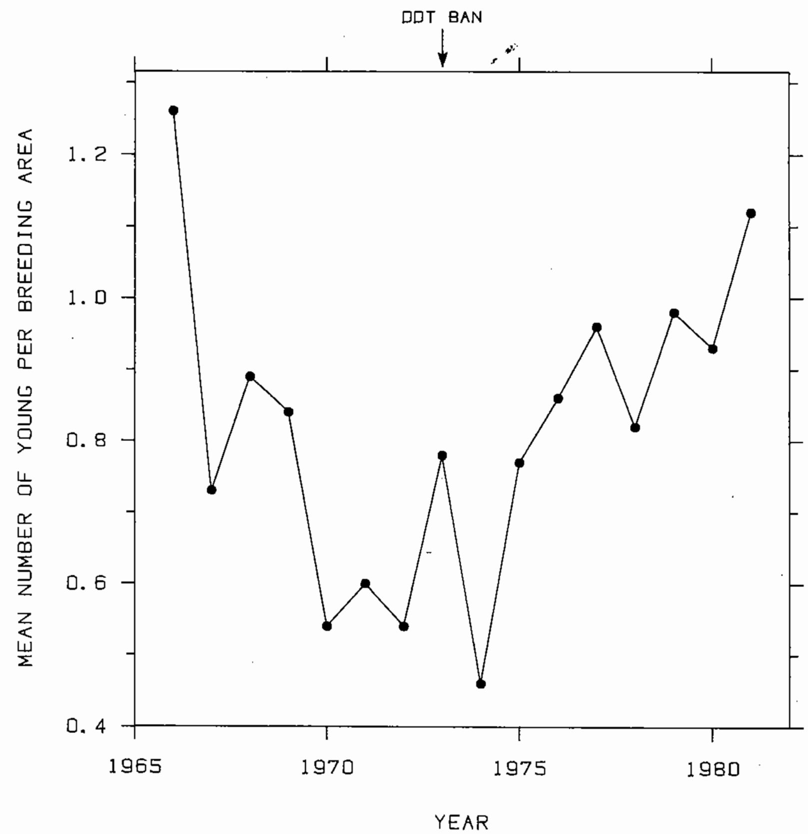

Figure 2.8 VISUAL PROMINENCE. The plotting symbols on this graph are prominent enough to prevent being obscured by the connecting lines.

interesting experiment run by four Harvard anatomists — Charles Lyman, Regina O'Brien, G. Cliett Greene, and Elaine Papafrangos [89]. 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."

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

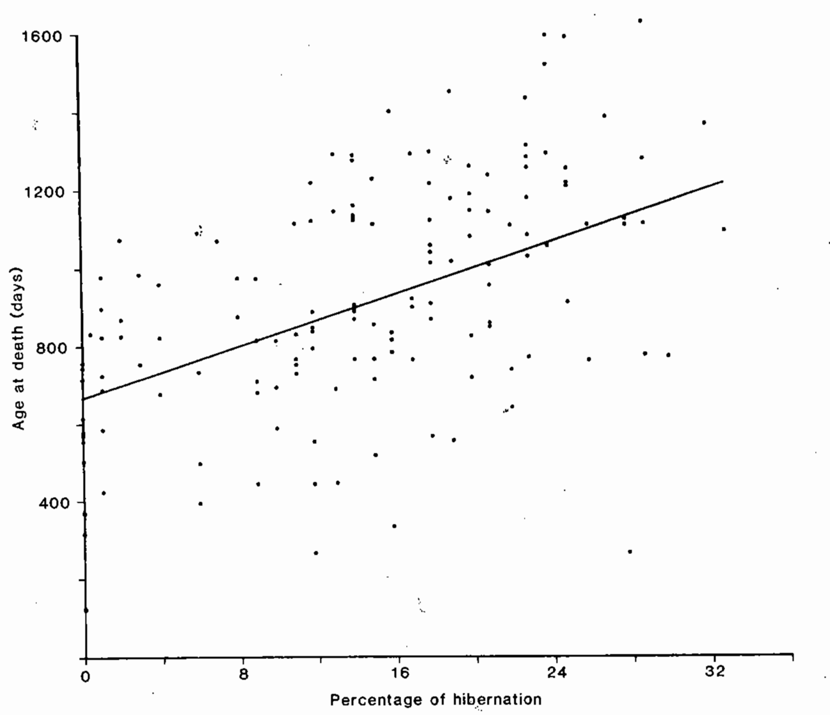

Figure 2.9 SCALE LINES AND THE DATA REGION. The data for zero hibernation are obscured by the left vertical scale line.
Figure republished from [89]. Copyright 1981 by the AAAS.

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 region — the place where the symbols representing the data are allowed to be — is in the interior of the rectangle formed by the scale lines. Now the values with zero hibernation can be seen clearly.

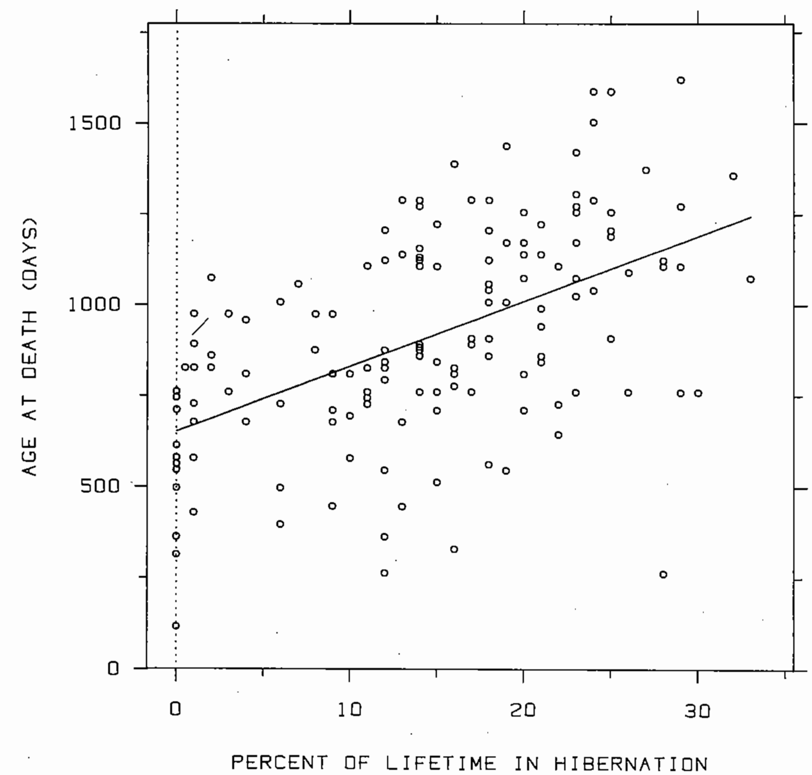

Figure 2.10 SCALE LINES AND THE DATA REGION. *Use a pair of scale lines for each variable. Make the data region the interior of the rectangle formed by the scale lines. Put tick marks outside of the data region.* This format prevents data from being obscured. Using two scale lines for each of the two variables on this graph, instead of the more usual one, allows easier judgment of the values of data on the top or on the right of the graph.

Ticks are put outside the data region in Figure 2.10 because ticks can obscure data, as is illustrated in the upper panels of Figure 2.11 [64].

Four scale lines are used in Figure 2.10 rather than the two of Figure 2.9. Judging the value of a point by judging its position along a scale line is easier 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 assess 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 assess than those at the bottom. By using four scale lines, the graph treats the data in a more nearly equitable fashion.

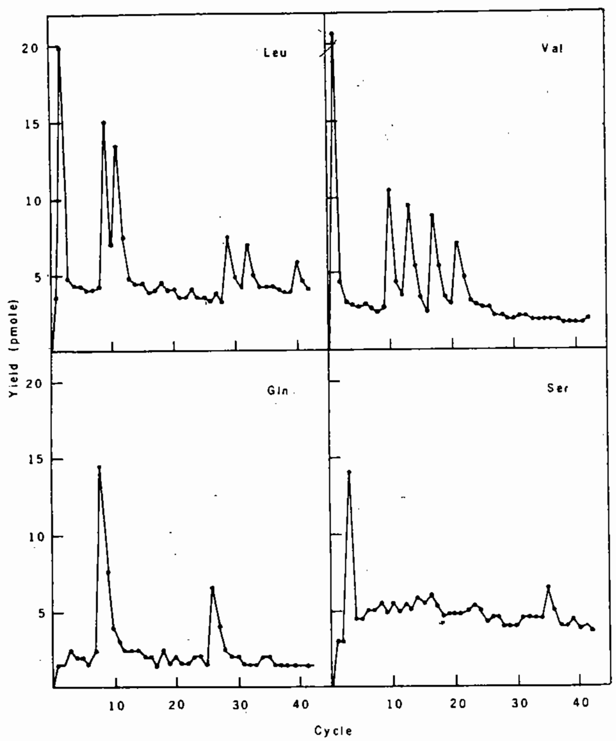

Figure 2.11 SCALE LINES AND THE DATA REGION. Tick marks can obscure data.
Figure republished from [64]. Copyright 1980 by the AAAS.

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 [139]; 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.

Making the data region the entire interior of the rectangle formed by the scale lines means the plotting symbol for a data point could just touch a scale line. But just touching has the potential to camouflage points. So it is probably best to interpret "interior" as a rectangle slightly inside the scale line rectangle.

Do not clutter the data region.

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

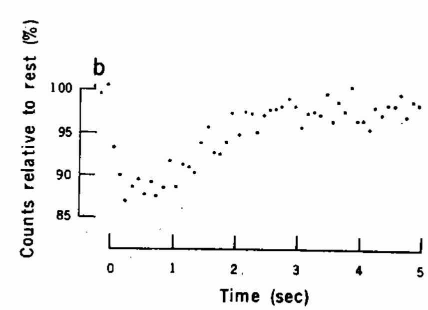

Figure 2.12 SCALE LINES AND THE DATA REGION. The three points in the upper left are camouflaged.
Figure republished from [139]. Copyright 1980 by the AAAS.

is illustrated in Figure 2.14 [122]. The data are particle counts from an exciting scientific exploration: the passage of the Pioneer II spacecraft by Saturn. Inside the data region 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.

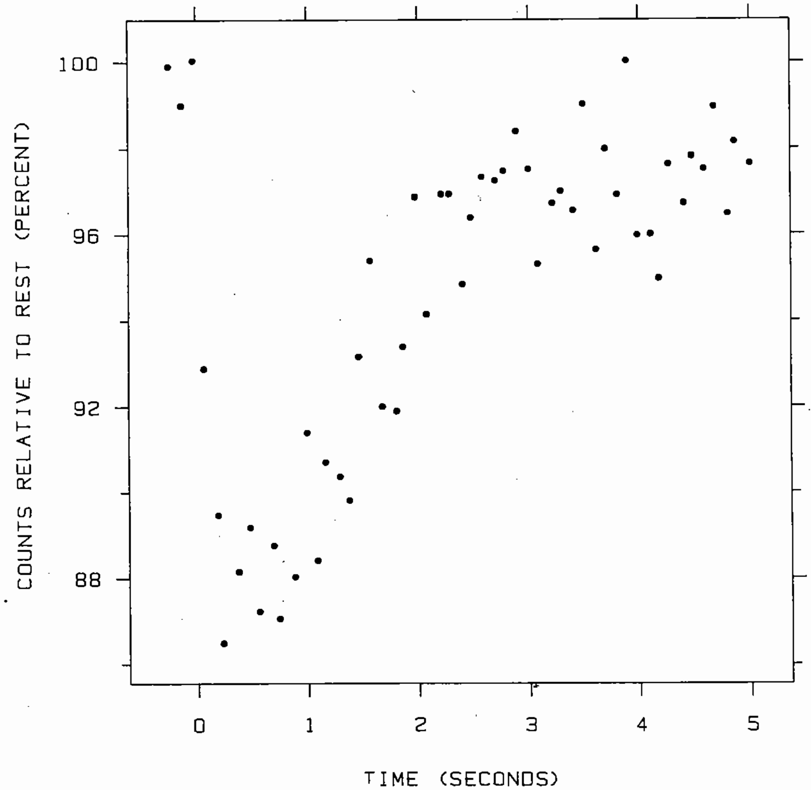

Figure 2.13 SCALE LINES AND THE DATA REGION. The four scale lines provide a clearly defined region for our eyes to look for data. Now, none of the data from Figure 2.12 are in danger of being overlooked.

The data are shown again in Figure 2.15. The clutter in the data region 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 the counts increase. (It is possible that accuracy is nearly constant on a scale of $(\text{counts/sec})^{1/2}$ 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.

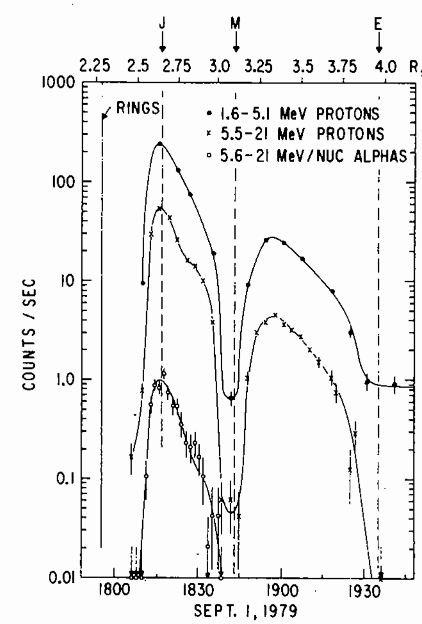

Figure 2.14 CLUTTER. This graph is cluttered. The result is that different graphical elements in the data region obscure one another.
Figure republished from [122]. Copyright 1980 by the AAAS.

The clutter in the data region also has been reduced by some alterations. The key is outside the data region, the label for rings is outside the data region, the arrows showing values below 0.01 counts/sec in Figure 2.14 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 and thus have reduced the clutter.

Figure 2.16 [81] is also cluttered; the error bars interfere with one another so much that it is hard to see the values they portray. One

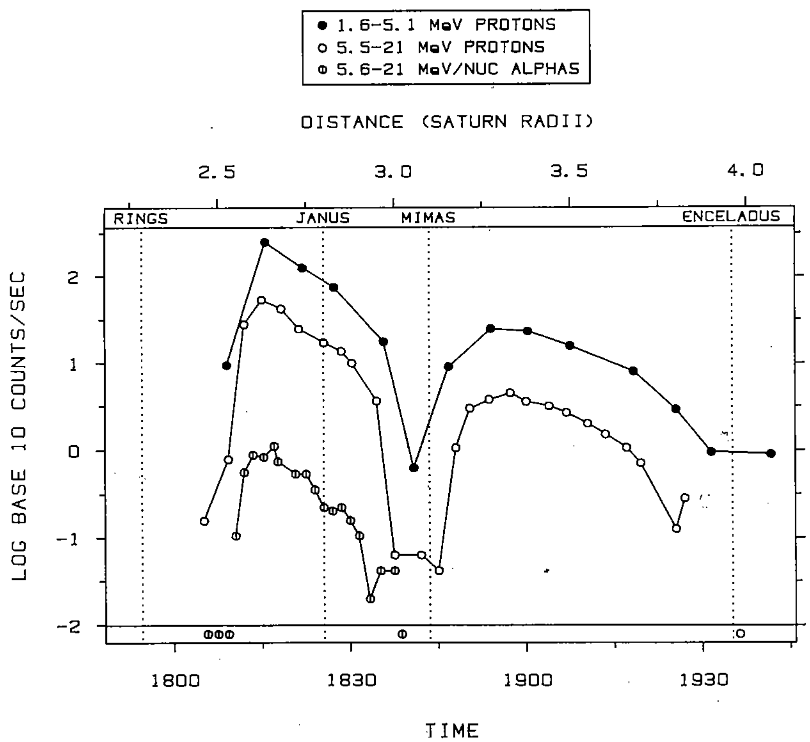

Figure 2.15 CLUTTER. *Do not clutter the data region.* The clutter of Figure 2.14 has been removed by alteration and excision. For example, the number of tick marks has been reduced.

solution is shown in Figure 2.17. In the top three panels the three data sets are juxtaposed and in the bottom 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.

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. Copious tick marks date back to a time when numerical values were communicated on graphs more than they are today. In our high-tech age we have photocopies of tables, computer tapes, disk packs, and telecommunications networks to transfer data. 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 [113] 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

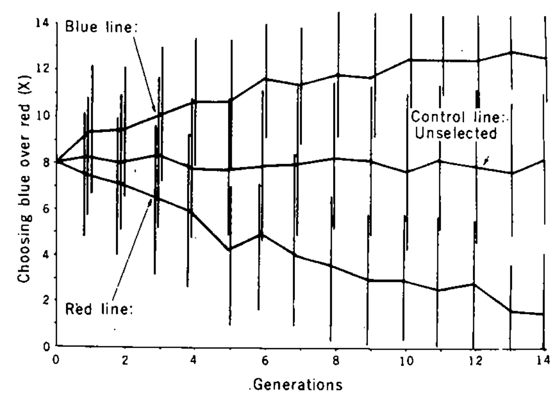

Figure 2.16 CLUTTER. This graph is also cluttered.
Figure republished from [81]. Copyright 1980 by the AAAS.

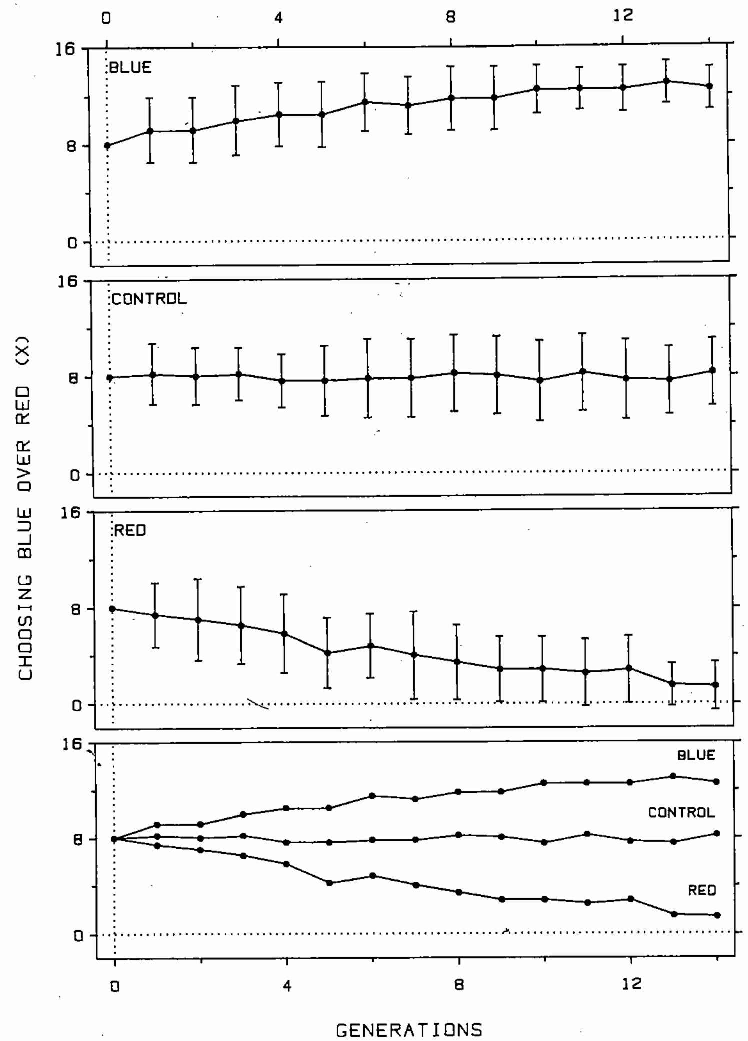

Figure 2.17 CLUTTER. The clutter of Figure 2.16 has been eliminated by graphing the data on juxtaposed panels. The bottom panel is included so that the values of the three data sets can be more effectively compared.

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.

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.

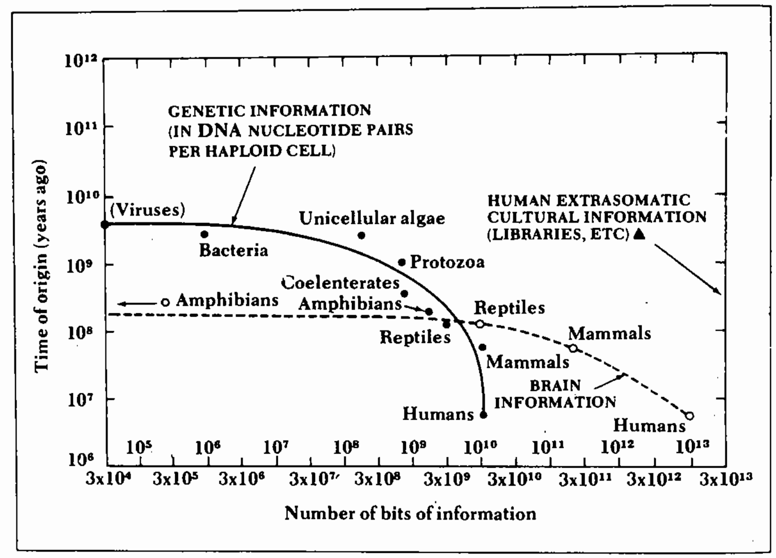

Figure 2.18 TICK MARKS. There are too many tick marks and tick mark labels on this graph. Using tick mark labels of the form 3×10^k as an approximation of $10^{k+0.5}$ is confusing.

Figure republished from *The Dragons of Eden: Speculations on the Evolution of Human Intelligence*, by Carl Sagan, p. 26. Copyright © 1977 by Carl Sagan. Reprinted by permission of Random House, Inc.

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

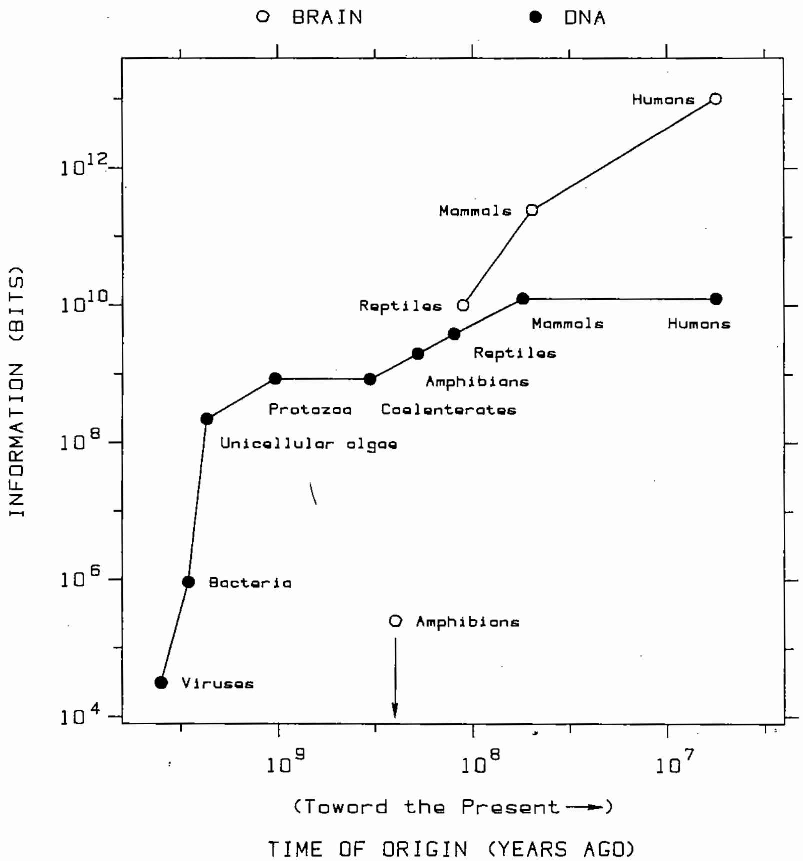

Figure 2.19 TICK MARKS. *Do not overdo the number of tick marks.* The vertical axis of this graph, previously the horizontal axis of Figure 2.18, has a sensible number of tick marks and labels.

Stephen Jay Gould's analysis of them [54], are discussed in detail in Section 4 of Chapter 3.) 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¢ to 35¢, all price increases were accompanied by a size increase.

In Figure 2.21 only a marker is used to show the time of the first cardiovascular care unit since the high precision of a reference line is not needed. We can see the position clearly enough to perceive that somewhere after that point, the death rate for cardiovascular disease decreased more rapidly; a reference line is avoided since we want, as always, to reduce the visual burden in the data region.

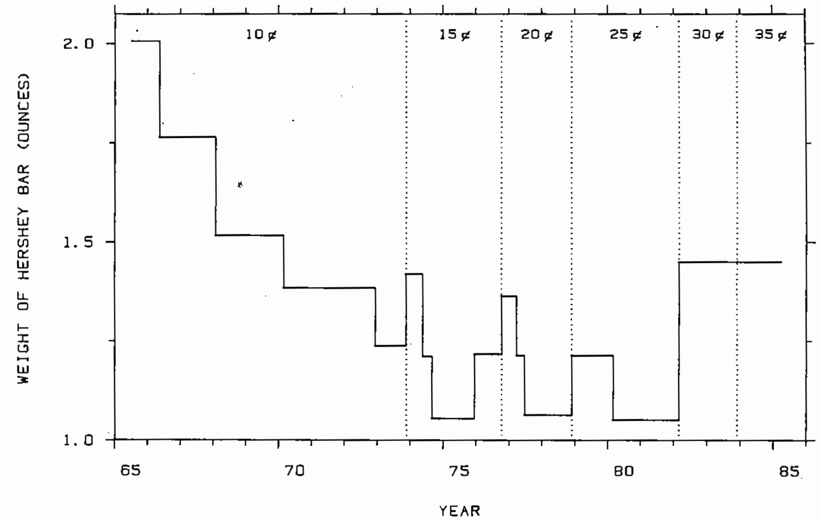

Figure 2.20 REFERENCE LINES. *Use a reference line when there is an important value that must be seen across the entire graph, but do not let the line interfere with the data.* The weight of the Hershey Bar is graphed against time. The vertical reference lines divide time up into price epochs; prices 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 data region to interfere with the quantitative data or to clutter the graph.

Figure 2.22 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 [101]. When it is important to convey the names for the individual values of a data set, data labels in the data region 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 2.22 through the use of several methods: the plotting symbol is visually very different from the letters of the labels, the letters of the labels are small, and when possible a label has been placed outside of the region formed by the point cloud rather than inside.

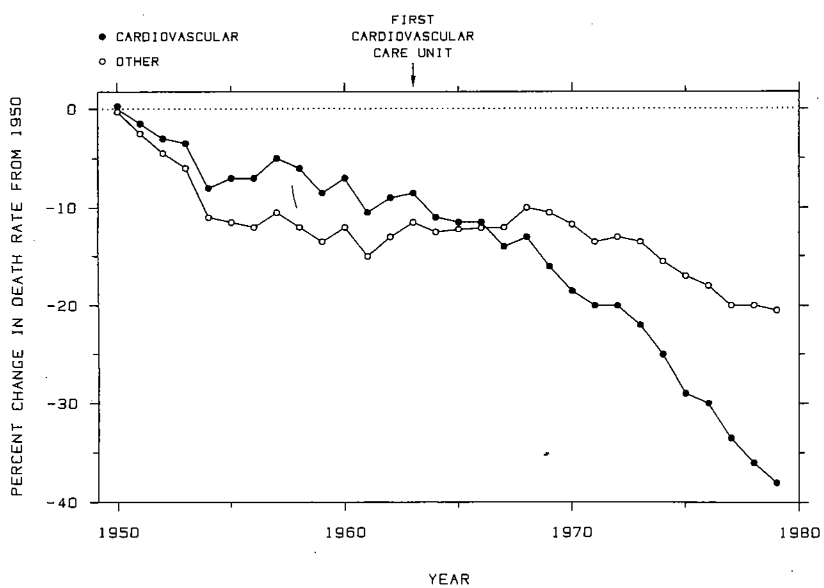

Figure 2.21 REFERENCE LINES. Only a marker is used to show the time of the first cardiovascular care unit since the high precision of a reference line is not needed.

In Figure 2.23 [113] the plotting symbols are not sufficiently visually distinguishable from the labels. The result is that the point cloud is camouflaged by the labels.

Figures 2.22 and 2.23 show one type of data label; each value in the data set has its own name. Sometimes the quantitative information on a

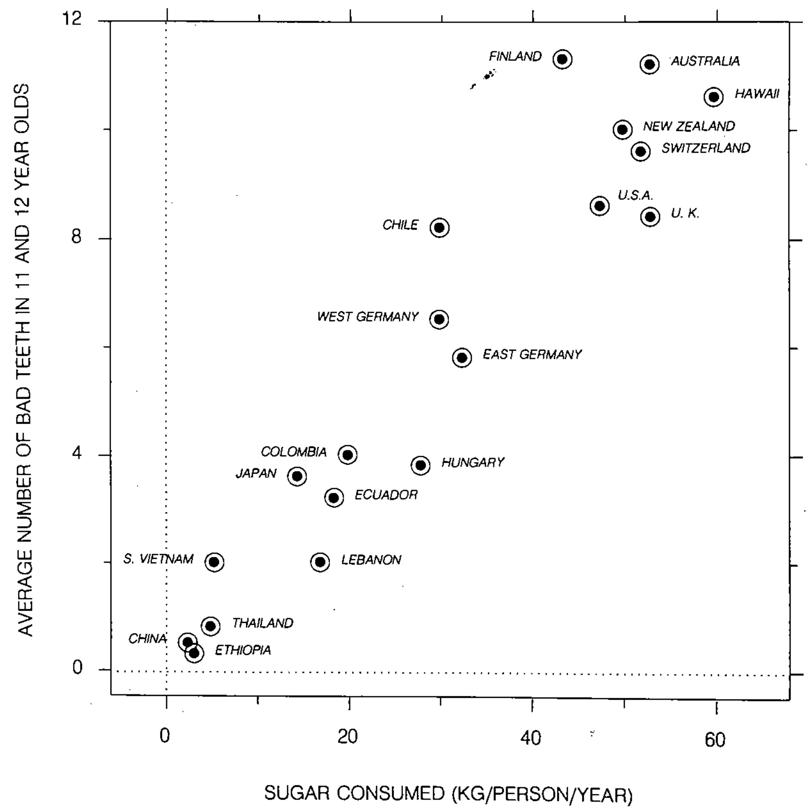

Figure 2.22 DATA LABELS. *Do not allow data labels in the data region to interfere with the quantitative data or to clutter the graph.* The data labels on this graph are needed to convey the names. The visual impact of the labels has been lessened so that they interfere as little as possible with our assessment of the overall pattern of the quantitative data.

graph consists of different data sets where each data set has a name that we want to convey. This is illustrated in Figure 2.24, which shows life expectancies for four groups of people: black females, black males, white females and white males [129, p. 71]. Four data labels in the data region convey the data set names without obscuring the data or cluttering the data region.

Sometimes a key with the data labels is needed to identify data sets, either because data labels in the data region 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 2.25 for both reasons. On this graph the data labels are long and the data region 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.

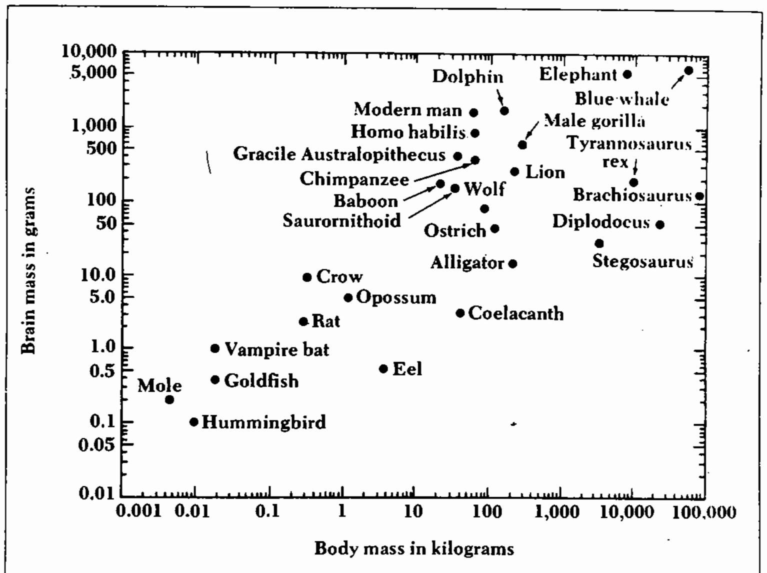

Figure 2.23 DATA LABELS. The data labels interfere with our assessment of the overall pattern of the quantitative data.
Figure republished from *The Dragons of Eden: Speculations on the Evolution of Human Intelligence*, by Carl Sagan, p. 39. Copyright © 1977 by Carl Sagan. Reprinted by permission of Random House, Inc.

Avoid putting notes, keys, and markers in the data region. Put keys and markers just outside the data region and put notes in the legend or in the text.

We should approach the data region 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 markers, keys, and notes need to appear in the data region.

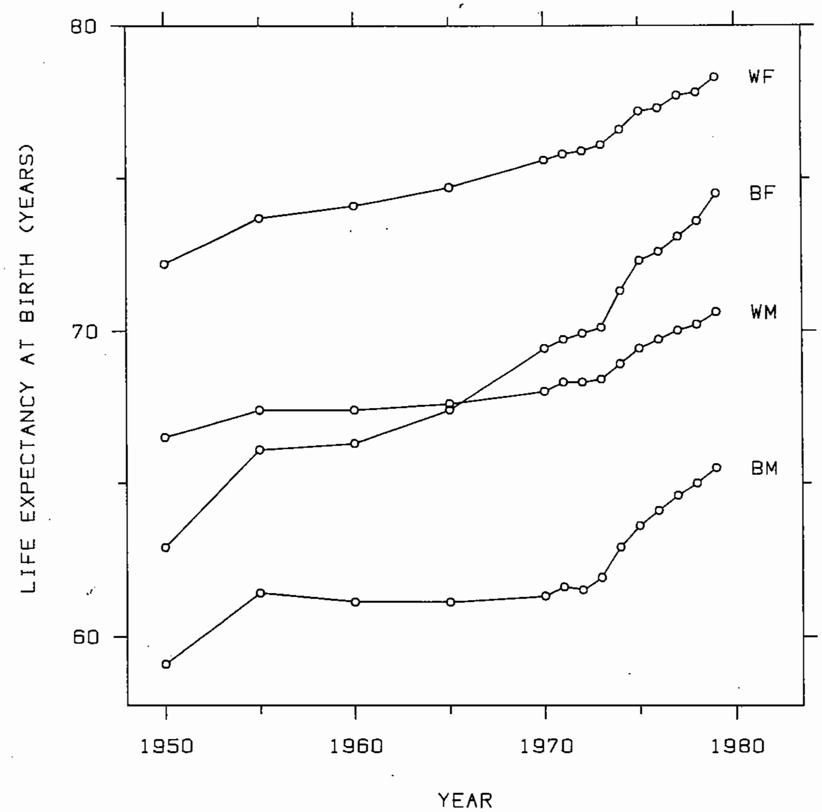

Figure 2.24 DATA LABELS. Groups of data values often can be identified by data labels in the data region. The labels are abbreviations in which B = black, W = white, M = male, and F = female.

Keys and markers can go outside the data region and notes can go in the text or the legend. This has not been done in Figure 2.26 [133] and the result is needless clutter and a confusing graph. The main graph (not including the inset) shows release rates of xenon-133 from the Three Mile Island nuclear reactor accident in 1979 and concentrations of xenon in the air of Albany, N.Y. during 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.

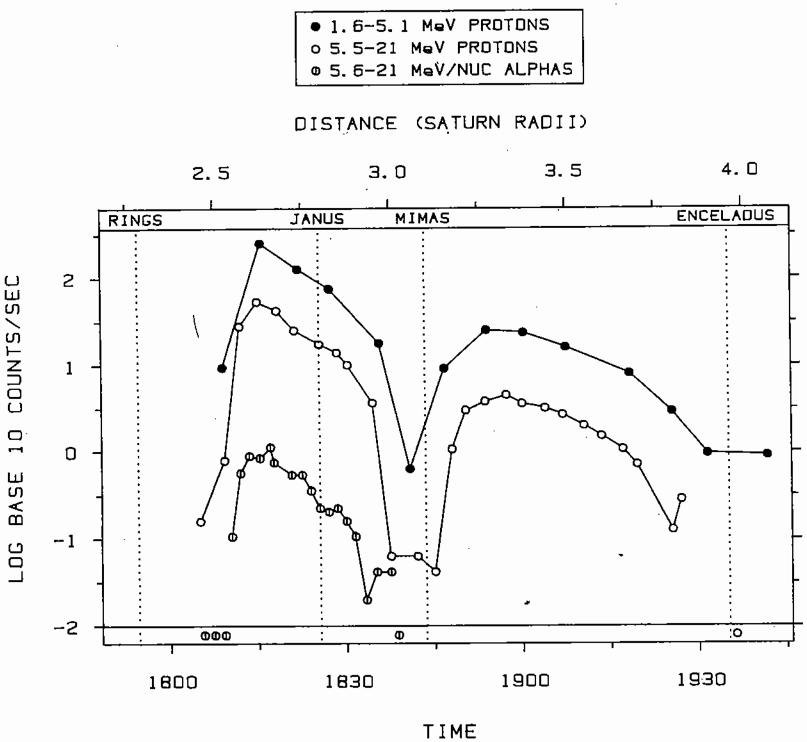

Figure 2.25 DATA LABELS. Groups of data values also can be identified by a key. One disadvantage, compared with data labels in the data region, is that identification is slightly harder because we must look back and forth between the key and the quantitative data. One advantage over data labels in the data region, an important one in this example, is that clutter in the data region is reduced. Furthermore, a key is needed in this example 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.

Figure 2.26 has a number of problems arising from some unusual and unexplained conventions and from putting too much in the data region. The writing in the data region 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, which impinges on the data region, has very little additional information; it shows two averages and repeats 5 of

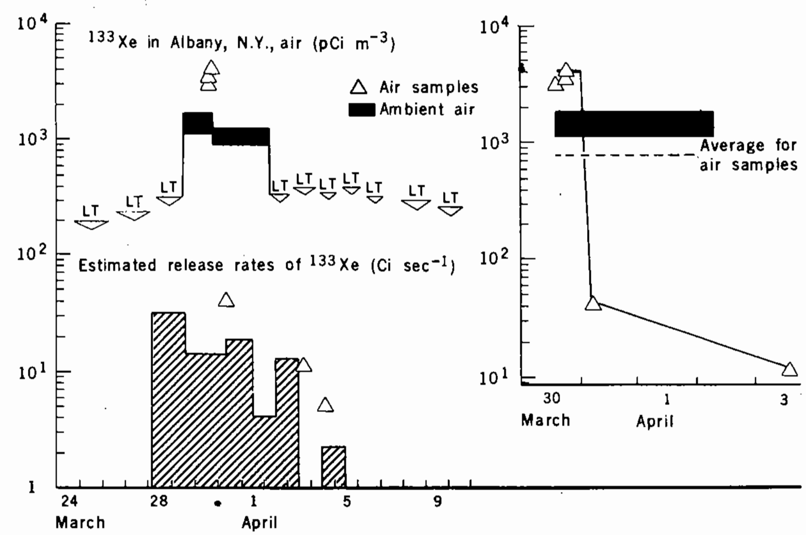

Fig. 1. Xenon-133 activity (picocuries per cubic meter of air) in Albany, New York, for the end of March and early April 1979. The lower trace shows the time-averaged estimates of releases (curies per second) from the Three Mile Island reactor (2). The inset shows detailed values for air samples (gas counting) and concurrent average values for ambient air (Ge diode). Abbreviation: *LT*, less than.

Figure 2.26 NOTES, KEYS, AND MARKERS. Everything — including the scale labels, a key, "LT" (meaning less than), and an inset — has been thrown into the data region of this graph. The result is confusing.

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 in the data region to make completely clear what has been graphed. This has been done in Figure 2.27; the writing, key, and LT's have been removed from the data region 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 2.26 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.

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 2.28 [23]. 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 4 of Chapter 3.

Superposed data sets must be readily visually discriminated.

It is very common for graphs to have two or more data sets superposed within the same data region. We already have encountered many such graphs in Chapters 1 and 2. The studies reported in Section 4 of Chapter 1 revealed that one of the most serious

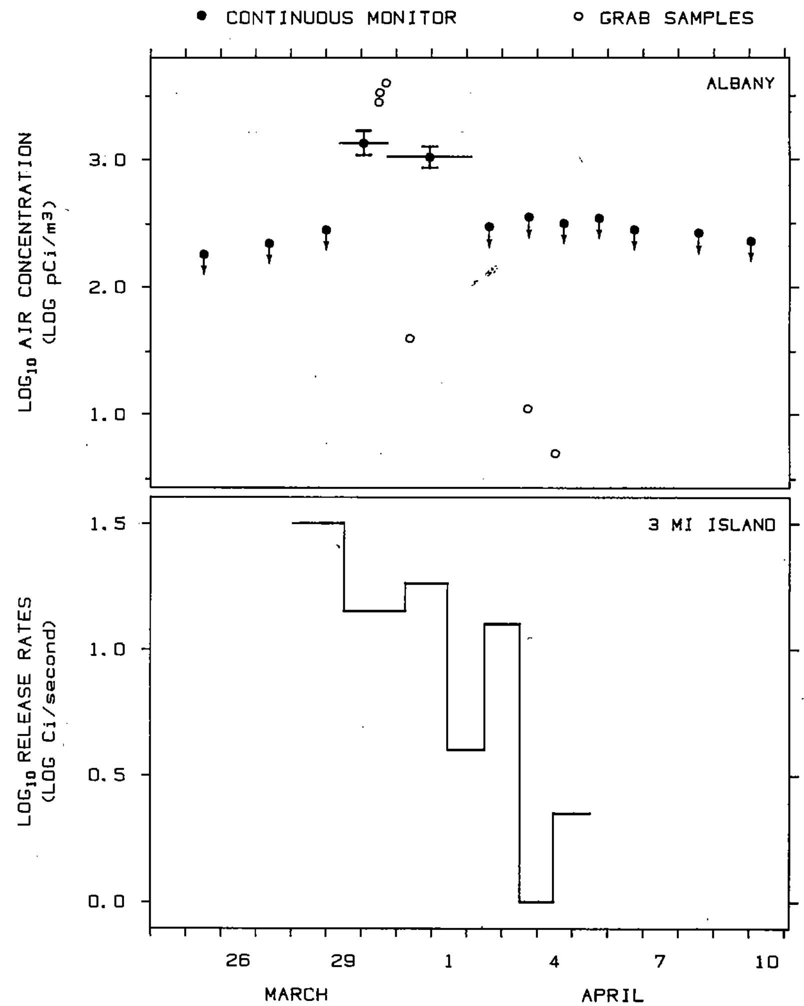

Figure 2.27 NOTES, KEYS, AND MARKERS. *Avoid putting notes, keys, and markers in the data region. Put keys and markers just outside the data region and put notes in the legend or in the text.* The graph in Figure 2.26 has been improved by the following actions: removing the writing, the key, and the inset from the data region; showing the two data sets on separate panels; removing the idiosyncrasies; correcting the labels describing the two types of measurement.

shortcomings in graphs in science and technology was poor visual discrimination of the different data sets on graphs employing superposition.

In Figure 2.29 [95] it is difficult to visually disentangle the solid squares, circles, and triangles; such plotting symbols are in general visually similar, but in Figure 2.29 the problem is exacerbated by the symbols not being crisply drawn. In Figure 2.30 [50] the different

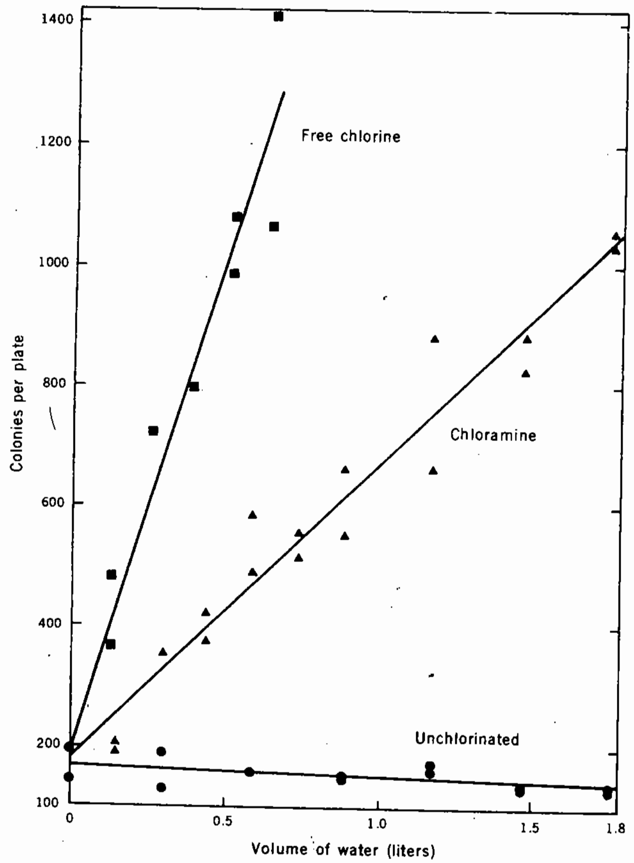

Figure 2.28 OVERLAPPING PLOTTING SYMBOLS. *Overlapping plotting symbols must be visually distinguishable.* On this graph, because of exact and near overlap, some of the data cannot be seen. Methods for combatting overlap are given in Chapter 3.

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

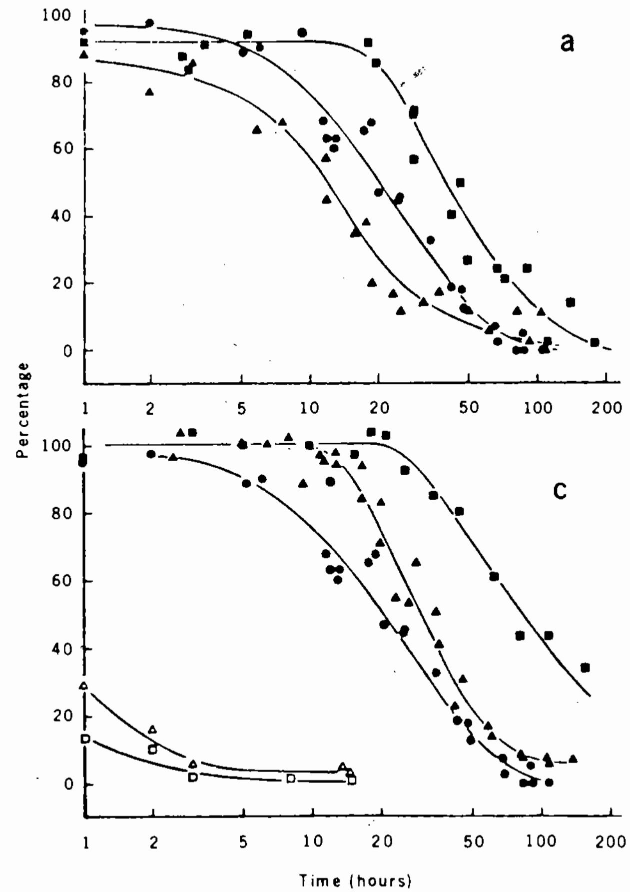

Figure 2.29 SUPERPOSED DATA SETS. *Superposed data sets must be readily visually discriminated.* One of the pervasive problems of graphs in science and technology is the lack of visual discrimination of different data sets superposed in the same data region. On this graph we cannot easily visually discriminate the circles, squares, and triangles.

cognitive mental effort to follow each curve visually, rather than the rapid, easy discrimination 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 see a single curve as a whole, mentally filtering out the other curves.

Graphs that fail to allow effective visual discrimination are pervasive because the problem is a difficult one to solve. Solutions will be given in Section 5 of Chapter 3.

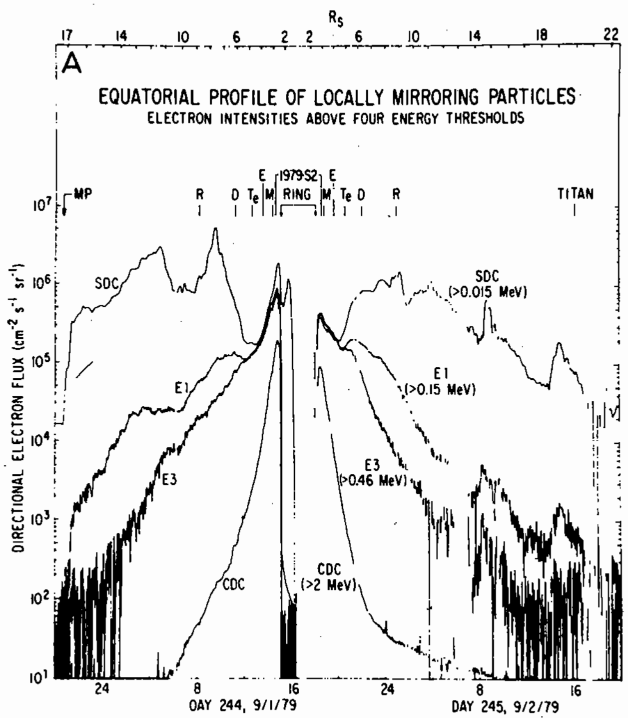

Figure 2.30 SUPERPOSED DATA SETS. The curves on this graph merge, in going from left to right, and then separate with their identities lost. Methods for graphing different data sets and maintaining visual discrimination are given in Chapter 3.

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 2.31 [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 2.31 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 region as if nothing is happening to the scales.

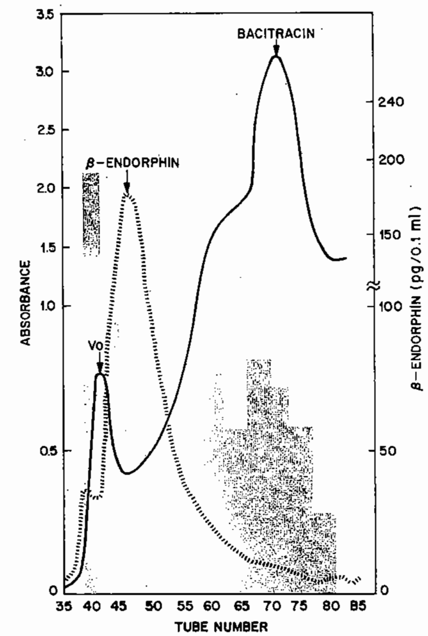

Figure 2.31 REDUCTION AND REPRODUCTION. *Visual clarity must be preserved under reduction and reproduction.* This did not happen on this graph. The ghostly image in the background was supposed to represent immunoreactivity.

In Figure 2.32 [98] 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.

One good test to check the ability of a graph original to stand up to both reduction and reproduction is to put it through a reducing photocopier. If a reduction to 2/3 is available, copy the original and then copy the copy. If the second copy, which is a reduction to $(2/3)^2 = 4/9$, is still visually clear, then it is likely that the original will withstand most reduction and reproduction processes. Clearly other strategies, depending on the photocopy equipment available and on the way the graph original will be utilized, can be tailored to each situation.

2.3 CLEAR UNDERSTANDING

Graphs are powerful tools for communicating quantitative information in, for example, technical reports and journal articles. 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 legends comprehensive and informative.

Communication of the results of scientific and technological studies, when the results involve quantitative issues, can be greatly enhanced by graphs that speak to the essence of the results. Graphs and their

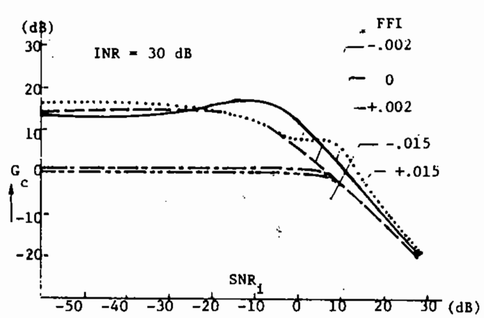

Figure 2.32 REDUCTION AND REPRODUCTION. The lines from the curves to their labels are washed out.
Figure republished from [98]. © 1983 IEEE.

legends can incisively communicate important data and important conclusions drawn from the data. One good approach is to make the sequence of graphs and their legends 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 legends 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 *legends* 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 legend of Figure 2.33. 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.

What could have caused such a calamity? In 1980 Luis Alvarez, Walter Alvarez, Frank Asaro, and Helen Michel at Berkeley discovered 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 [104]. These data are shown in Figure 2.33.

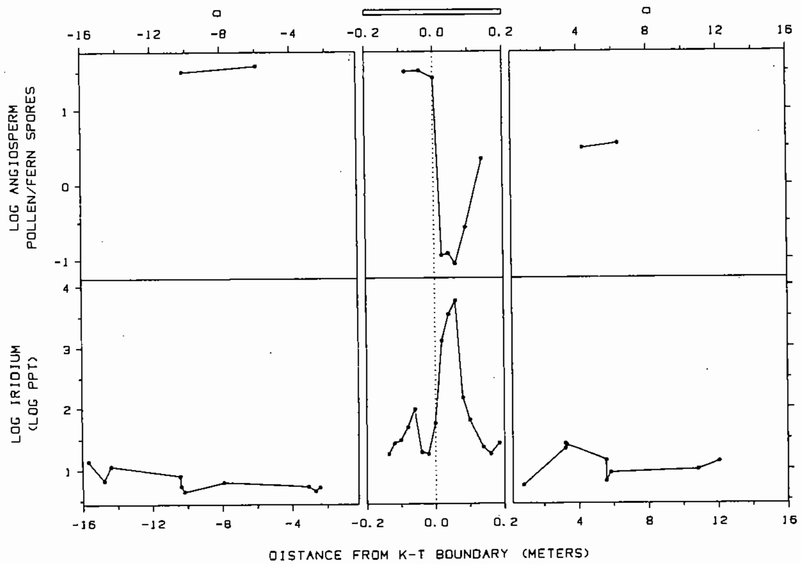

Figure 2.33 EXPLANATION. *Put major conclusions into graphical form. Make legends comprehensive and informative.* Describe everything that is graphed and convey the conclusion drawn from the data. The following is a legend, including the title, that might accompany this graph in its original subject 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 upper panel shows the ratio of angiosperm pollen to fern spores on a log base 10 scale; 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, also on a log base 10 scale; 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 darkened the earth for months or years, leading to the large number of extinctions, including the dinosaurs, that occurred at the beginning of the Tertiary period.

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 point at which the ratio of pollen to fern spores begins to decrease is taken to be the K-T boundary because at the beginning of the Tertiary period angiosperms declined relative to ferns. At this boundary there is a corresponding peak in the iridium concentrations, shown in the bottom panel.

The legend of Figure 2.33 follows the three-step guidelines presented earlier. The graph and its legend 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, legend, 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 legend that makes a graph as self-contained as possible. If there are several graphs, the legends collectively can be an independent piece; for example, a detailed description of a data set described in one graph legend does not need to be repeated in a subsequent graph legend.

It is possible, though, to overdo a comprehensive legend. Putting a description of the experimental procedure in the legend — conventional in medical and biological writings and mandatory in some circles — seems to go too far. It burdens the graph and makes what should be a concise summary into a tome. Figure 2.34 [139] is an example. The ratio of legend 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 legend, which is a summary.

Too little detail, however, occurs more frequently in graphs in science and technology than too much detail. The studies of graphs in scientific publications described in Section 4 of Chapter 1 revealed an alarming percentage of graphs containing elements not explained either in the text or in the legend. Figure 2.35 [39] 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.

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.

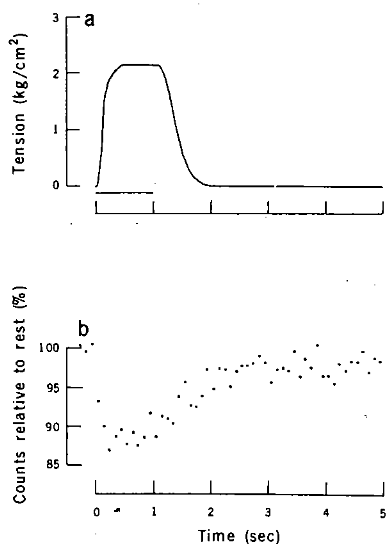

Fig. 2. Tension and the intensity of the 42.9-nm layer line during 1-second tetanus at the sarcomere length of 2.2 μm. (a) Tension record averaged over the 40 tetanic contractions required for obtaining the time course of the layer-line intensity. A sartorius muscle was dissected from *Rana catesbeiana* 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.

Figure 2.34 EXPLANATION. It is possible to overdo the explanation in a legend. The complete description of the experimental procedure in this legend is too much detail. The ratio of the legend area to the graph area is 2.8.
Figure republished from [139]. Copyright 1980 by the AAAS.

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,

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

The sample standard deviation of the data is

$$s = \left[\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2} .$$

An estimate of the standard error of the mean is

$$s/n^{1/2} \, .$$

If the data are from a normal distribution then a 95% confidence interval for the population mean is $(\bar{x} - k\ s/n^{1/2}, \bar{x} + k\ s/n^{1/2})$, 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 2.36 [66]. In the last sentence of the figure legend we are told that the graphed values "represent means of

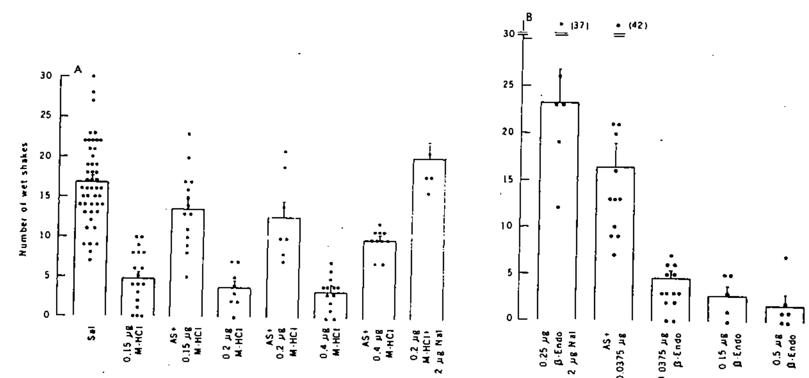

Fig. 1. Inhibitory effect of morphine hydrochloride (A) and β-endorphin (B) on wet-shake behavior in rats. Antagonism was by antibody to cerebroside sulfate (*AS*). A volume of 2 μl was delivered into the PAG in 1-μl increments with a 1-minute interval in between. The control consisted of saline (*Sal*). 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 also significantly different ($P < .005$) from the group receiving 0.0375 μg of β-endorphin (B).

Figure 2.35 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 legend.
Figure republished from [39]. Copyright 1980 by the AAAS.

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:

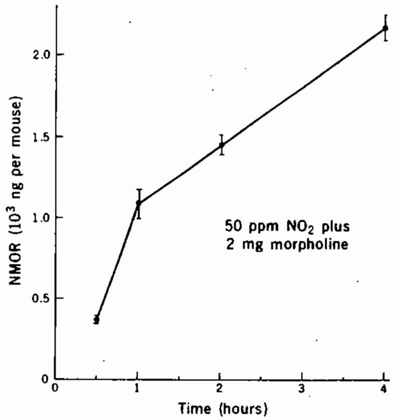

Fig. 1. Time course of NMOR biosynthesis in mice. Groups of three to four male ICR mice were gavaged with freshly prepared solutions of 2 mg of MOR (Aldrich Chemical) in 0.2 ml of distilled water and immediately placed in exposure chambers (Nalge desiccators, modified for gas inflow from the bottom and 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 NMOR levels and for the DPN standard recoveries and represent means of three to four mice ± the standard deviation.

Figure 2.36 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 statistics the error bars portray.

(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 7 of Chapter 3.

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

The dot chart in Figure 2.37 shows death rates for the leading causes of death of people in the age group 15 to 24 years in the United States [99]. 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. The bottom horizontal scale line uses another method for labeling; the tick mark labels and the scale labels correspond, but both are describing the variable on the log scale.

Proofread graphs.

Graphs should be proofread and carefully checked for errors. In the study of the graphs in the journal *Science* described in Section 4 of Chapter 1, construction errors were uncovered in 6.4% of the graphs. Any of them could have been detected by careful proofreading. Figure 2.38 [115] shows such an error for a graph of measurements of Saturn's magnetic field made by the Pioneer II spacecraft; the exponents for the tick mark labels on the vertical scale line are missing. 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.

Strive for clarity.

Strive for clarity is really a summation of the principles presented so far; in Section 2.2 the principles contribute to making a graph visually clear and in Section 2.3 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.

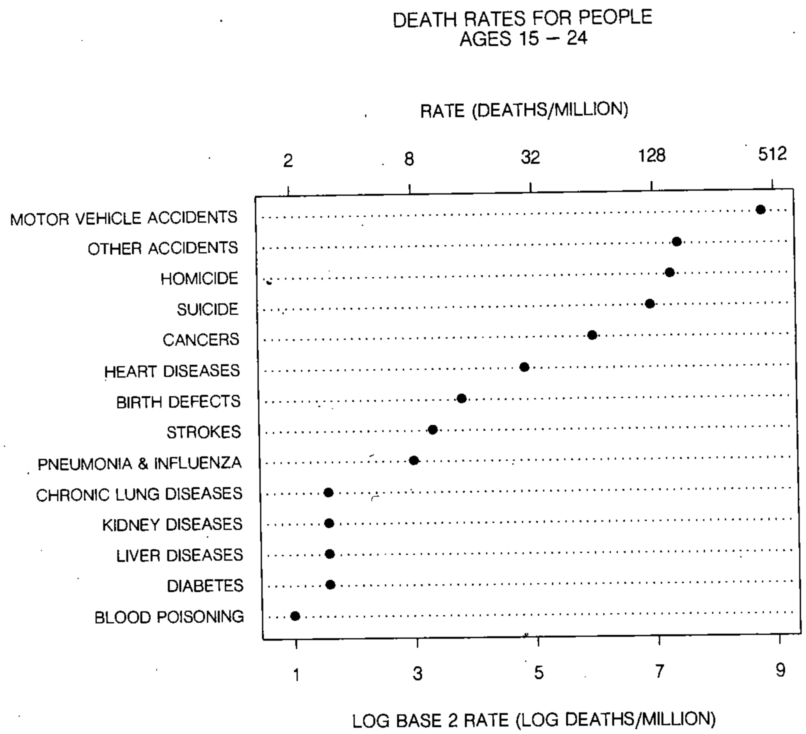

Figure 2.37 LABELS FOR LOGS. *When logarithms of a variable are graphed, the scale label should correspond to the tick mark labels.* The logarithms of the data are graphed on this dot chart. On the top horizontal scale line the tick mark labels are in the units of the data on the original scale, so the 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.

The data in Figure 2.39 [132] are percentages of degrees awarded to women in several areas 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 legend 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 makes it clear that the total distance from the zero baseline to the top encodes the percentage for bachelor's, the total distance from the baseline to the top of the middle division encodes the percentage for master's, and the length of the bottom division encodes doctorate's. This type of graph works only because the percentages decrease in going from bachelor's degrees to master's degrees to doctorates for every category.

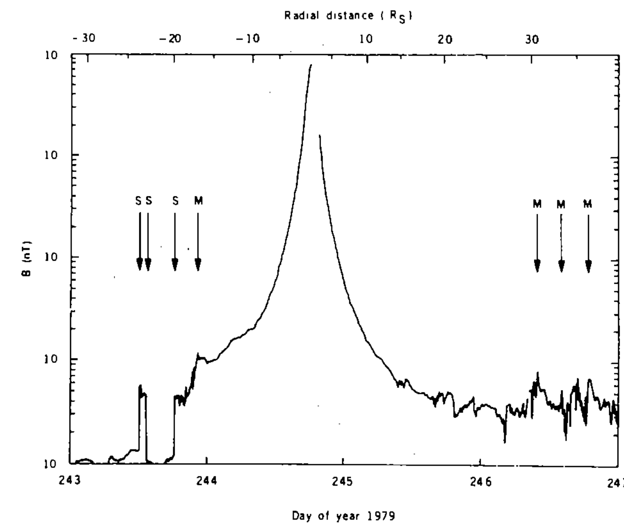

Figure 2.38 PROOFREAD. *Proofread graphs.* Graphs should be proofread, just as we do text. On this graph, lack of careful proofreading resulted in missing exponents on the tick mark labels of the vertical scale. Figure republished from [115]. Copyright 1980 by the AAAS.

There are other problems with this graph. Only two bars are shown for computer science, with no explanation. One can only assume, since majoring in computer science is a new phenomenon, 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 subject.

In Figure 2.40 the data from Figure 2.39 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 subject through time. Finally, the figure legend explains the graph in a comprehensive and clear way.

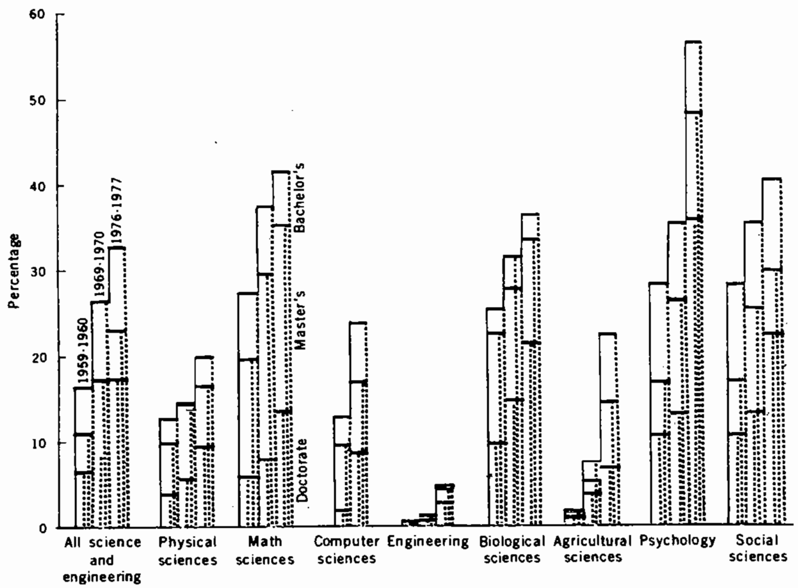

Fig. 1. Proportion of degrees in science and engineering earned by women in 1959 to 1960, 1969 to 1970, and 1976 to 1977 (6). Included in the social science degrees are anthropology, sociology, economics, and political science.

Figure 2.39 CLARITY. This graph fails both in clarity of vision and clarity of explanation.

Figure republished from [132]. Copyright 1980 by the AAAS.

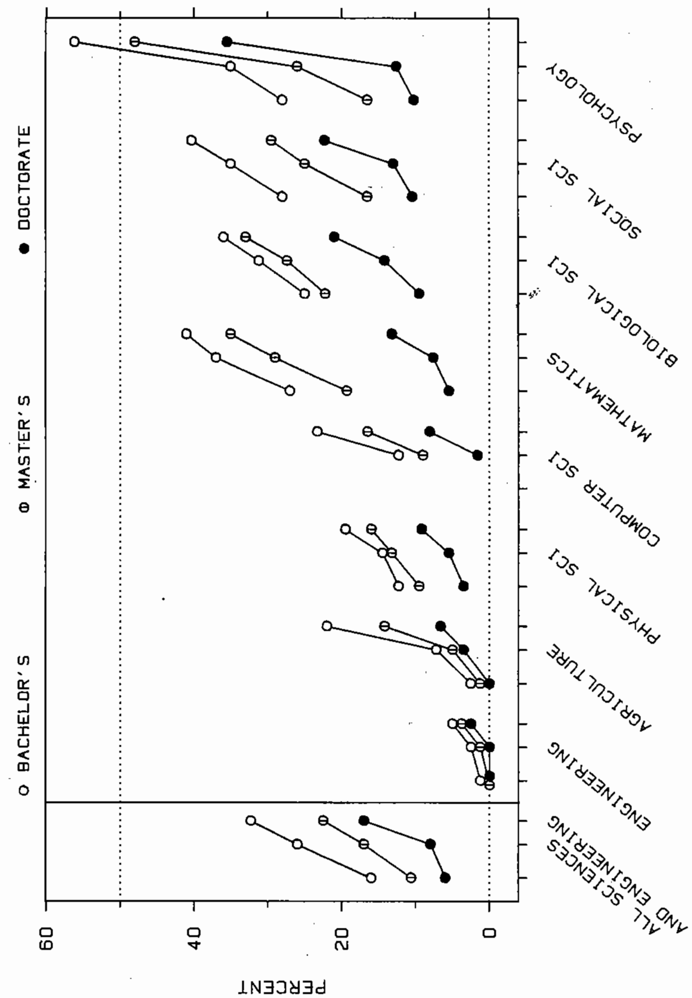

Figure 2.40 CLARITY. *Strive for clarity.* This is a summation of the principles in Sections 2.2 and 2.3. This graphing of the data from Figure 2.39 strives for clarity. It shows the percentage of degrees earned by women for three degree categories, three time periods, and nine categories of subjects. For each subject category the three tick marks indicate the years 1959-1960, 1969-1970 and 1976-1977.

2.4 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. This section is about constructing scale lines, comparing scales, including zero, taking logarithms, and breaking a scale.

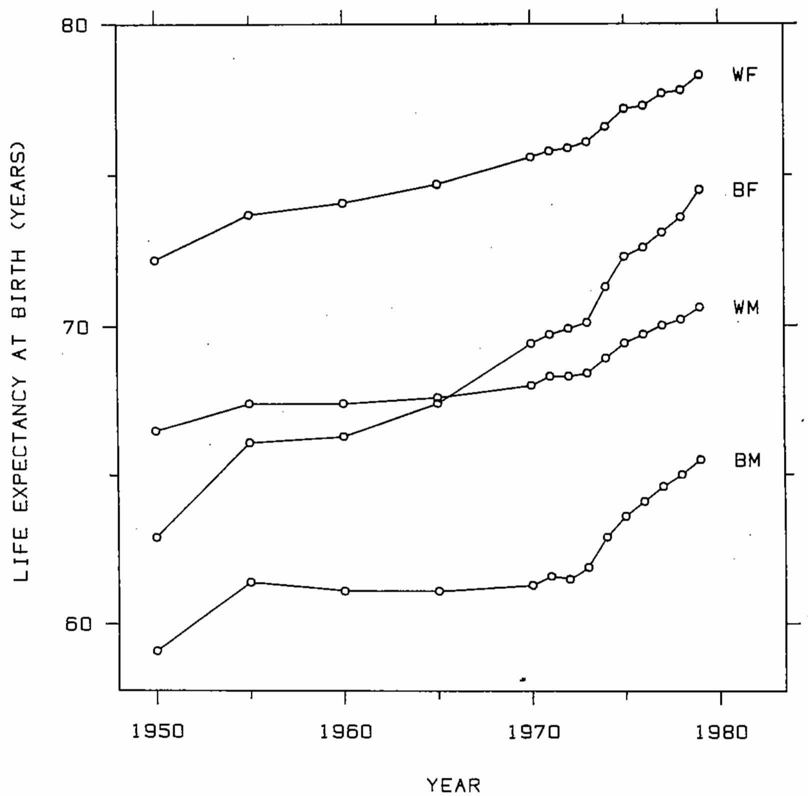

Figure 2.41 RANGES. *Choose the range of the tick marks to include or nearly include the range of the data.* The range of the values on the vertical scale are nearly contained within the range of the tick marks. On the horizontal scale the values are completely contained within the range of the tick marks.

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 2.41 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.

Subject to the constraints that scales have, choose the scales so that the data fill up as much of the data region 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 data region as possible. This is not done in Figure 2.42 [121].

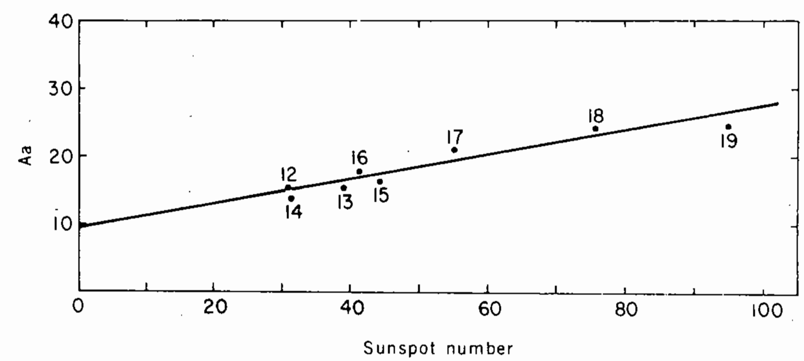

Figure 2.42 FILLING THE DATA REGION. Only 26% of the vertical scale is used by the data.
Figure republished from [121]. Copyright 1980 by the AAAS.

Only 26% of the vertical scale is taken up by the data. Space is wasted on this graph. In contrast, Figure 2.43 utilizes the data region more efficiently. The data span most of the range of the scales without getting too close to the frame. 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.

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

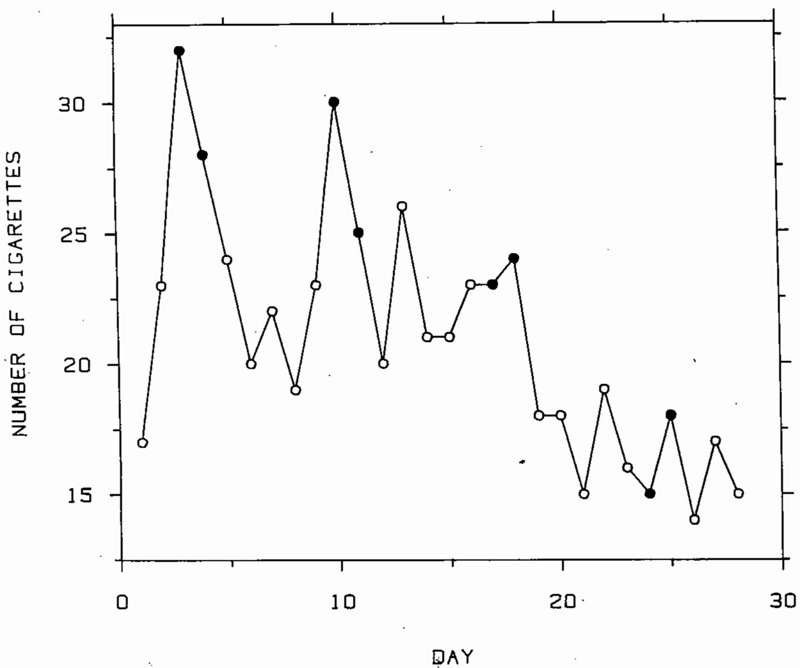

Figure 2.43 FILLING THE DATA REGION. *Subject to the constraints that scales have, choose the scales so that the data fill up as much of the data region as possible.*

information of a second scale often can be helpful. One example is Figure 2.44. The data, which are from the 1980 census [130], are the number of people in the United States in 1980 for each age from 0 to 84. The bottom horizontal scale line shows the age and the top horizontal scale line shows the year of birth.

When the logarithms of data are graphed there is an opportunity to use two scales. In Figure 2.45 the death rates for people 15 to 24 years

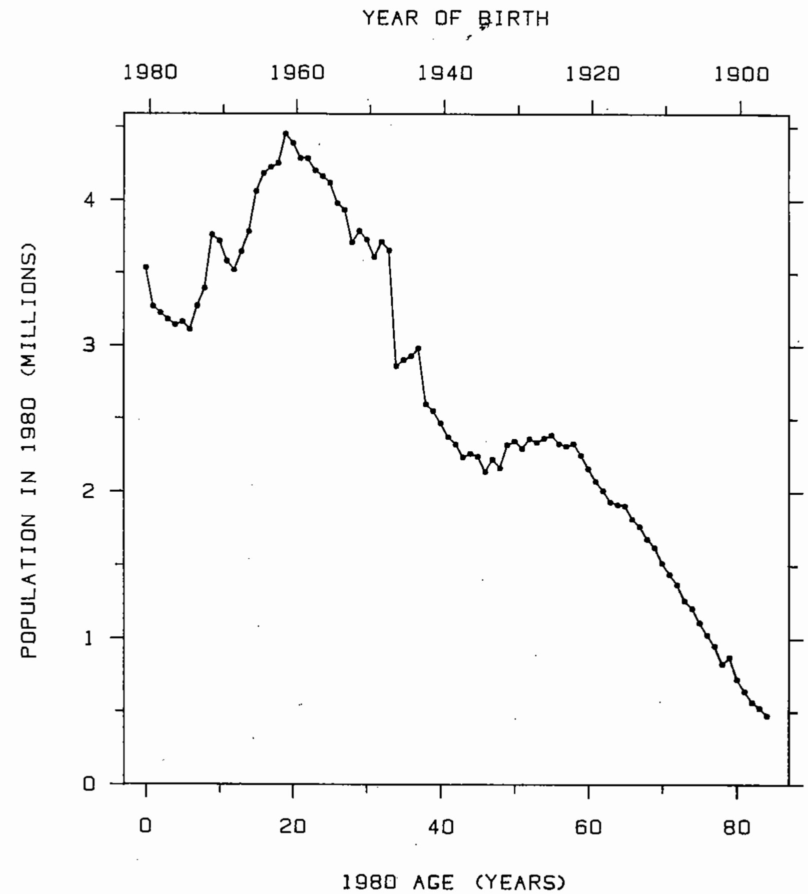

Figure 2.44 TWO SCALES. *It is sometimes helpful to use the pair of scale lines for a variable to show two different scales.* The bottom horizontal scale line shows age and the top horizontal scale line shows year of birth.

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.

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 2.46 is a graph of averages of

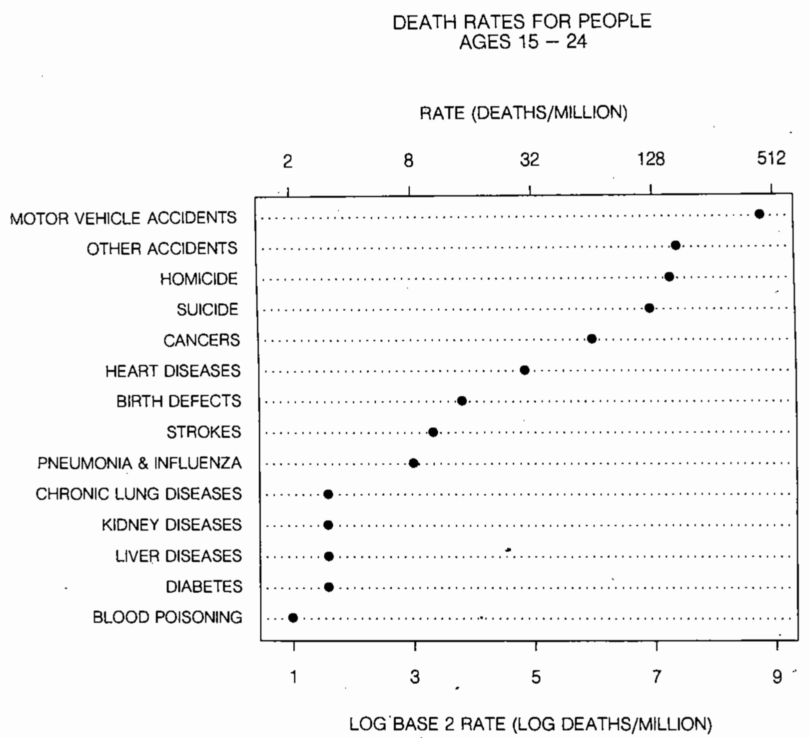

Figure 2.45 TWO SCALES. The bottom horizontal scale line shows log death rate in log deaths/million and the top horizontal scale line shows death rate in deaths/million.

the mathematics Scholastic Aptitude Test scores for selected years from 1967 to 1982 [131, p. 158]. 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%.

Choose appropriate scales when graphs are compared.

Figure 2.47 shows data from an experiment on graphical perception [33] that will be discussed in Section 3 of Chapter 4. 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

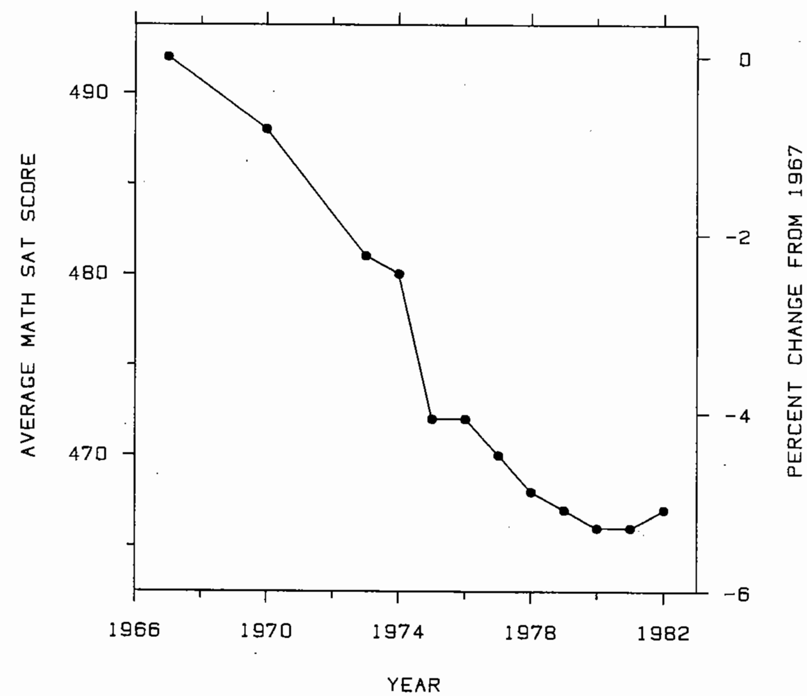

Figure 2.46 TWO SCALES. The left vertical scale line shows SAT score and the right vertical scale line shows percent change from 1967.

two values and visually judging what percent the smaller was of the larger. The left panel of Figure 2.47 shows the 40 average judgment errors (averaged across subjects) graphed against the true percents for the 40 pie chart judgments. The right panel shows the same variables for the bar chart judgments. The two smooth curves were computed using the lowess procedure that will be described in Section 4 of Chapter 3. 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 would fill the data region.

Unfortunately, scales cannot always be made the same; we must forego equal scales when the result is poor resolution. The next best thing is to have the same number of units per cm; this is illustrated in Figure 2.48. The data are the winning times of four track races at the Olympics from 1900 to 1984 [22, 138, p. 833]. The four lines have the same slopes but different intercepts and were fit to the data using least squares. If the vertical scales had been the same, the wide variation in

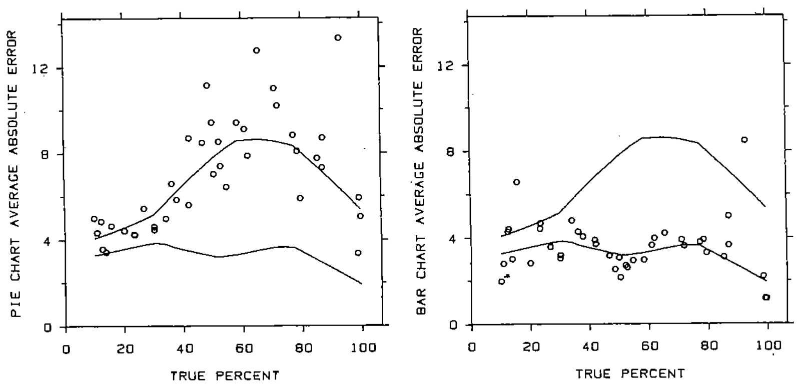

Figure 2.47 COMPARISON. *Choose appropriate scales when graphs are compared.* Scales on different panels should be made as commensurate as possible when the data on the different panels are compared. On this graph the scales on the left panel are the same as those on the right.

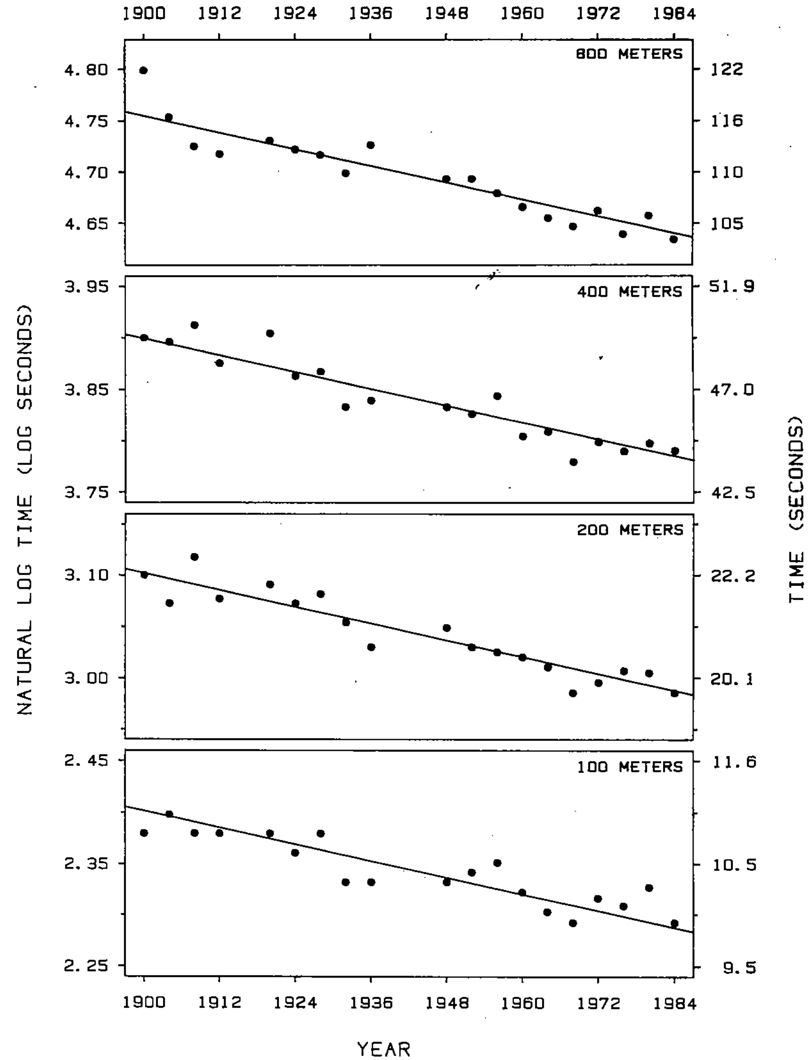

Figure 2.48 COMPARISON. Sometimes making scales identical ruins the resolution of one or more panels. The next best thing is illustrated on this graph — the number of units per cm is the same on the four vertical scales. The four lines on the panels have the same slope.

the times for the different races would have ruined the resolution. Instead, the number of log seconds per cm is the same. This allows us to compare changes in the data on different panels. For example, we can see that the overall rate of decrease through time for the four sets of data is about the same. Since logs are graphed, this means that the percent reduction in the running times for the four distances has been the same. Our ability to see this easily comes from having the same number of log units per cm on the vertical scales of the four panels.

Sometimes even the number of units per cm cannot be the same without ruining the resolution. In Figure 2.49 the data in the top panel are the monthly measurements of atmospheric CO_2 concentrations that were discussed in Section 2 of Chapter 1. The remaining panels are statistical descriptions of the trend in the data, the seasonal variation in the data, and all other variation. The range of the data and of the trend are very similar, so the vertical scales of the top two panels are the same. The variation of the data in the bottom two panels is much less than that in the top panels; if the number of units per cm were the same on the vertical scales of all panels, the resolution in the bottom two panels would be degraded. One way to appreciate how the scales change on the four panels is to study the tick mark labels and the distances between them, but this is a difficult mental-visual task. To make appreciation of the scale change easier in Figure 2.49, 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 four panels.

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 resolution of the data on the graph.

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* [62, pp. 64-65] goes so far as to say that a graph of magnitudes without a zero line is dishonest. Referring to Figure 2.50 he writes:

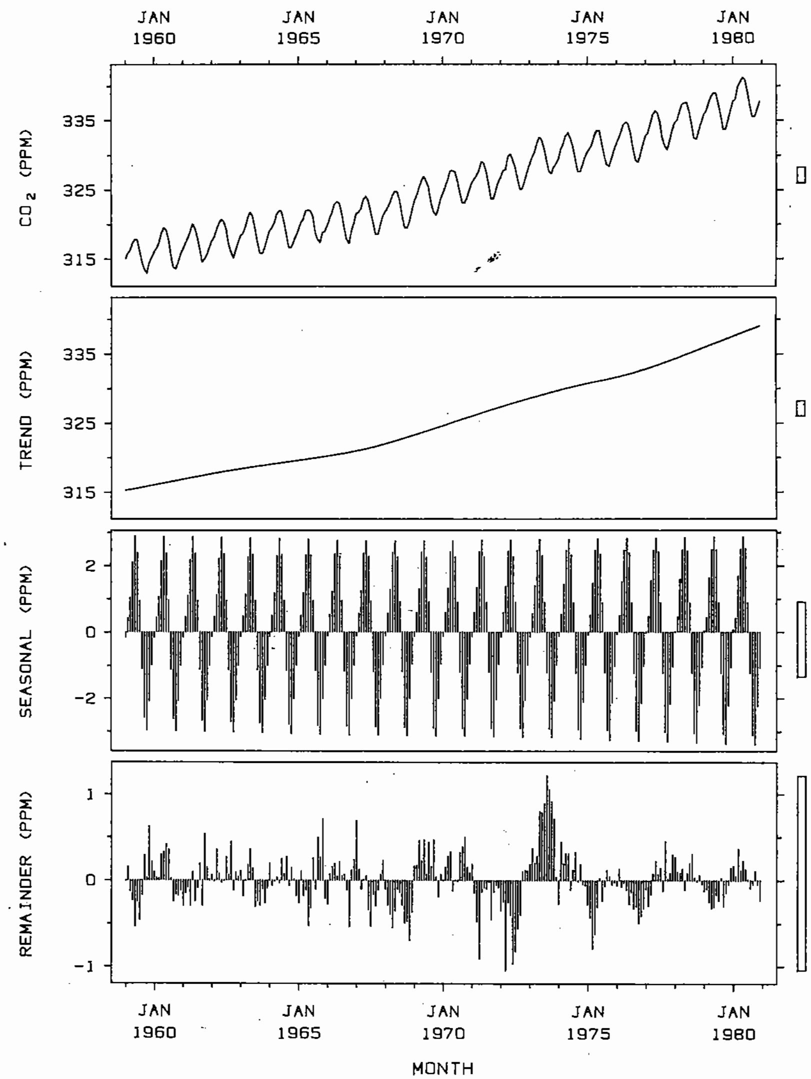

Figure 2.49 COMPARISON. Sometimes keeping the number of units per cm the same also ruins the resolution. On this graph the number of units per cm on the vertical scales varies. The rectangles on the right help to show the relative scaling; the vertical lengths portray changes of the same magnitudes on the four panels.

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 $19,500,00 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.

Huff's presumption is that viewers will not look at scale labels and apply the most trivial of quantitative reasoning. The result, the graph

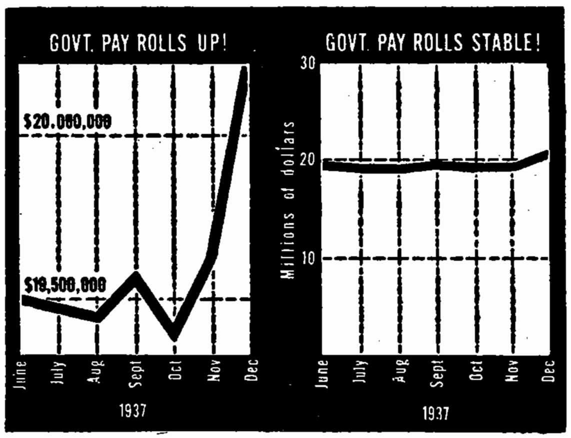

Figure 2.50 ZERO. The compulsion to include zero on a scale has ruined many graphs. Darrell Huff in *How to Lie With Statistics* argues the left graph is misleading, but the right graph is a waste of space that shows very little quantitative information beyond what could be conveyed in one sentence.
Figure republished by permission of W. W. Norton & Company, Inc. [62, p. 65]. Copyright 1954 and renewed 1982 by Darrell Huff and Irving Geis.

on the right in Figure 2.50, is a waste of space because the resolution is so poor; the simple statement, "government payrolls were 10.5 million dollars in June and rose by about 4% from June to December" is much more incisive and efficient. The graph on the left in Figure 2.50 conveys more quantitative information; for example, we can determine from the left graph that the rise is roughly 4%, but not from the right.

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 2.51 [69] 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 2.51 has degraded the visual resolution of the data. It is quite unlikely that a graph of these data with the vertical scale going from 4.0 to 5.5, which includes the range of the data, would lead space physicists to think the percent variation in the emission signals is larger than it really is.

Resolution has been ruined in Figure 2.52; including zero is ludicrous. The graph shows the CO_2 data and trend curve that were graphed in Figure 2.49. Figure 2.53 shows the data in the sensible way; now the changes in CO_2 through time can be seen far more clearly.

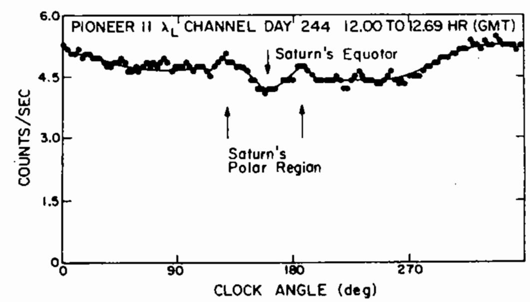

Figure 2.51 ZERO. The resolution of this graph is degraded by including zero.
Figure republished from [69]. Copyright 1980 by the AAAS.

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 [62, pp. 61-62] 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

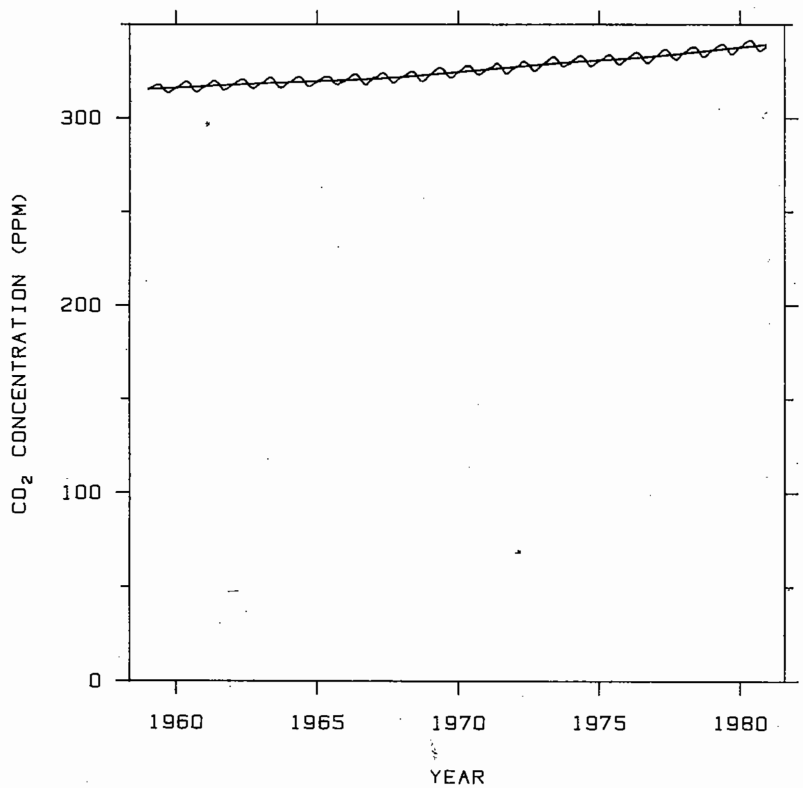

Figure 2.52 ZERO. Including zero here is ludicrous. It is reasonable to include zero if, unlike this graph, it does not ruin the resolution.

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

$$\frac{a}{b} = \frac{c}{d}$$

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 2.54. The data are the number of telephones in the U.S. from 1935 to 1970

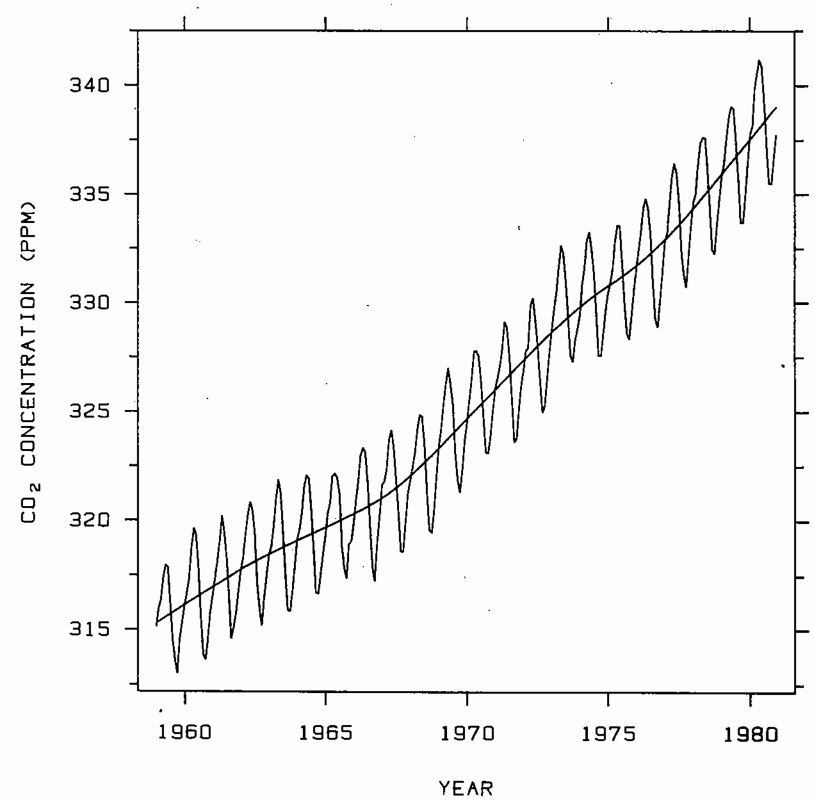

Figure 2.53 ZERO. *Do not insist that zero always be included on a scale showing magnitude.* This graph conveys much more quantitative information than Figure 2.52.

[128, p. 783]. 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 2.54 without reading off values from the vertical scale and doing arithmetic.

When magnitudes are graphed on a logarithmic scale, percents and factors are easier to judge since equal multiplicative factors and percents

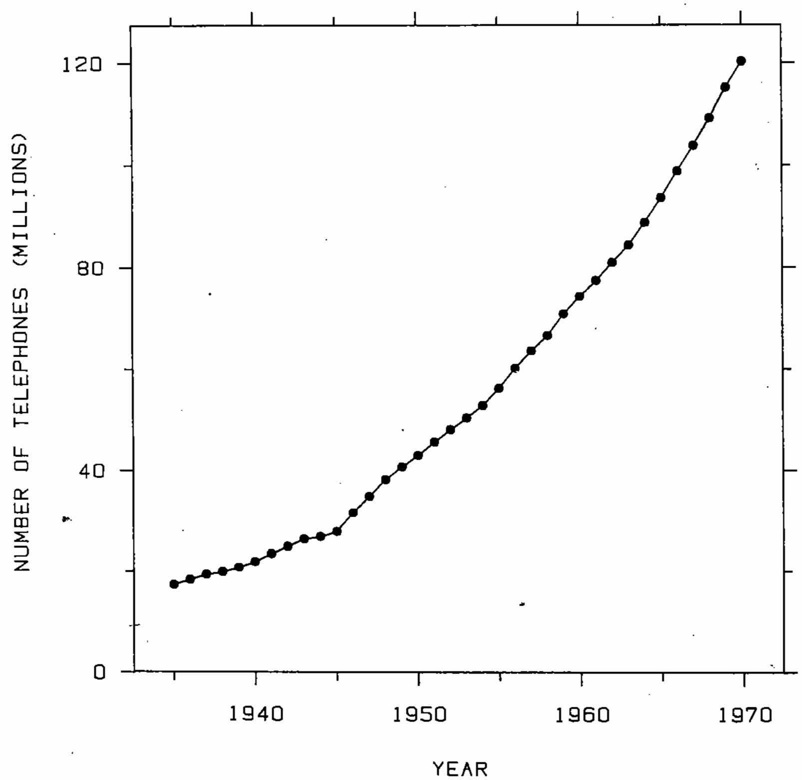

Figure 2.54 LOGS FOR FACTORS. The data are the number of telephones in the United States each year from 1935 to 1970. It is nearly impossible to judge whether the percentage increase is constant, decreasing, or increasing.

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 2.55. A log base 2 scale is used on the vertical axis 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.

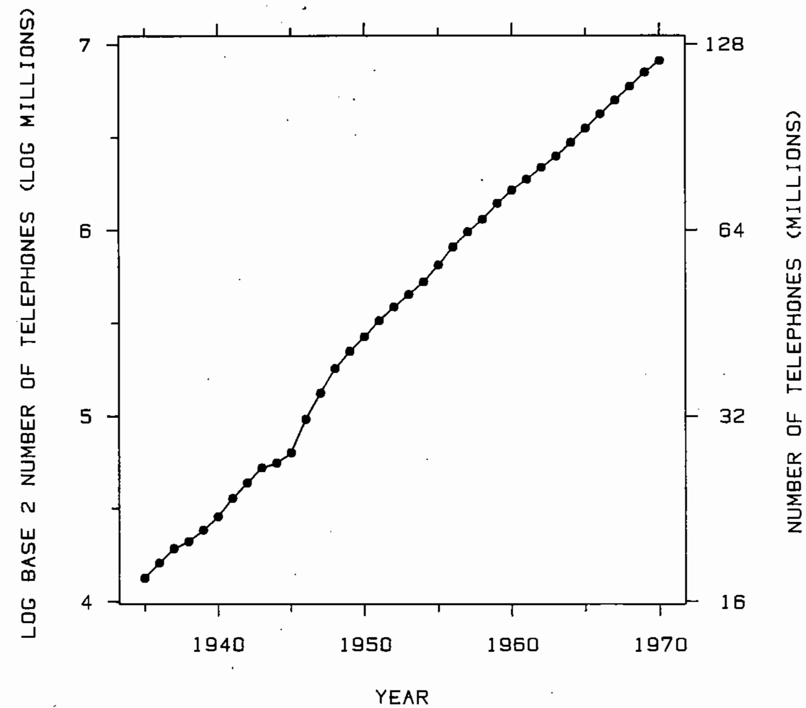

Figure 2.55 LOGS FOR FACTORS. *Use a logarithmic scale when it is important to understand percent change or multiplicative factors.* The data in Figure 2.54, are graphed by taking logarithms base 2. Now it is clear that the percentage increase in telephones was roughly stable from 1935 to 1970.

Showing data on a logarithmic scale can improve resolution.

It is common for positive data to be *skewed to the right*: 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. Such data can cause severe resolution problems on graphs, and the common remedy 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.

An example of skewed data is given in Figure 2.56. The graph shows the 14 most abundant elements in stone meteorites [48]; the data are the average percent of each of the elements. The resolution on the graph is poor because the ten smallest percents vary over a very small range. Figure 2.57 shows the data on a log scale; now the distribution is much more nearly uniform and the resolution is greatly improved.

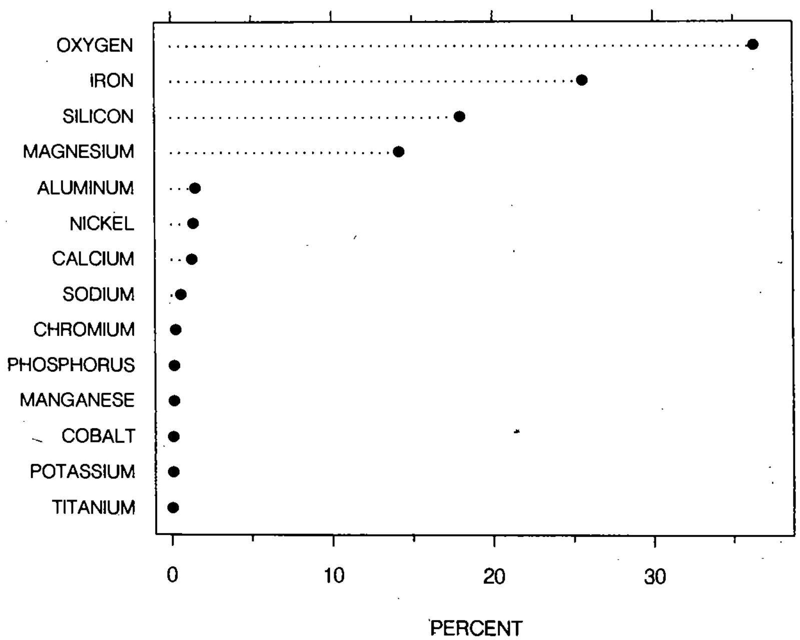

Figure 2.56 LOGS FOR RESOLUTION. Because the data on this graph are skewed to the right, the resolution of the majority of the values on the horizontal scale is poor.

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.

Figure 2.58 shows the iridium data discussed earlier in Figure 2.33. Two *full scale breaks* are used to signal changes on the horizontal scale. The middle panel has a much smaller number of data units (meters) per cm; the widths of the rectangles at the top of the graph portray the same number of horizontal scale units (meters) on the panels. A full break shows a change or gap in a scale about as forcefully as possible.

In science and technology today the convention for indicating a change or gap in the scale of a graph is a *partial scale break*: two short

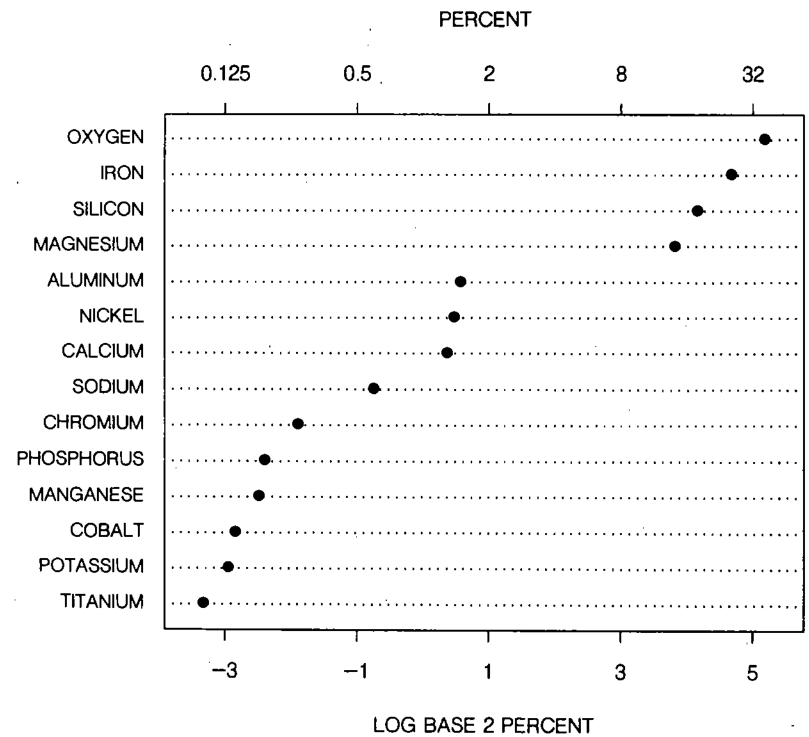

Figure 2.57 LOGS FOR RESOLUTION. *Showing data on a logarithmic scale can improve resolution.* The logs of the data in Figure 2.56 are graphed and the resolution has improved substantially.

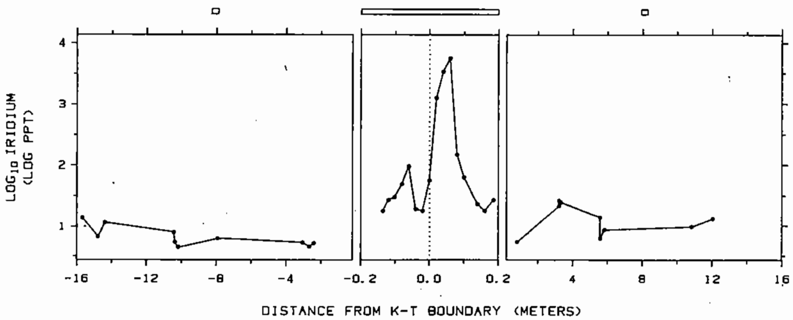

Figure 2.58 SCALE BREAKS. ***Use a scale break only when necessary. If a break cannot be avoided, use a full scale break. Do not connect numerical values on two sides of a break.*** **This graph uses full scale breaks on the horizontal scale to signal changes in the number of units per cm. The full breaks 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.**

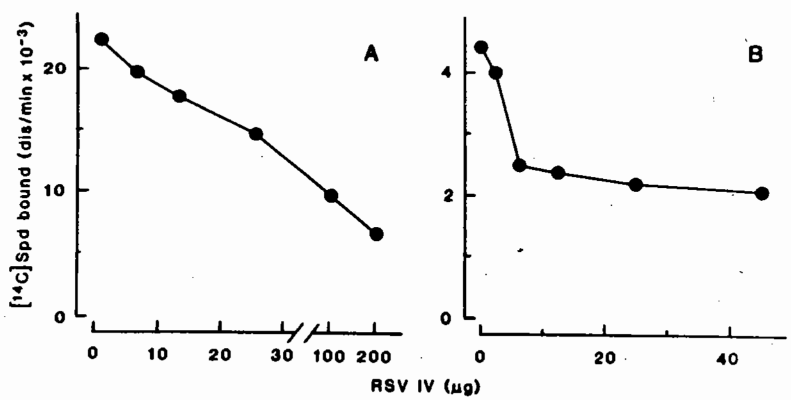

Figure 2.59 SCALE BREAKS. **The partial scale break on the horizontal scale of the left panel does not give a forceful indication of a break. The connection of numerical values across the break gives the misleading impression that the data are roughly linear.**

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 2.59 [105]. 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.

Numerical values should not be connected across a break. In the left panel of Figure 2.59, 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 2.60, which graphs the data with no scale break.

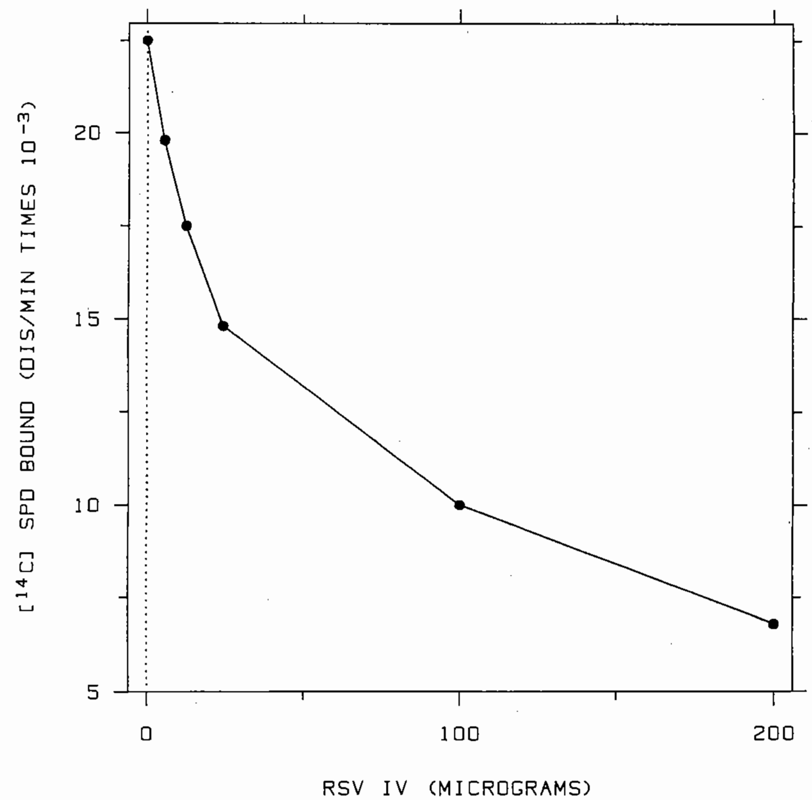

Figure 2.60 SCALE BREAKS. **The data from the left panel of Figure 2.59 are graphed without a scale break. Now it is clear that the data are not roughly linear and that the slope decreases as the variable on the horizontal axis increases.**

The problem in Figure 2.59 is not rare. The studies of graphs in scientific publications discussed in Section 4 of Chapter 1 revealed widespread problems caused by scale breaks. Figures 2.61 and 2.62 are other examples. Figure 2.61 [92] 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 2.62 [116] bars and error bars are allowed to barge right through two scale breaks. This renders meaningless the bar lengths and areas, important and prominent visual aspects of the graph.

Full scale breaks should be used only when necessary. Figure 2.60 shows the break of Figure 2.59 is not needed. Taking logarithms of the data can often relieve the need for a scale break. Figure 2.63 shows data

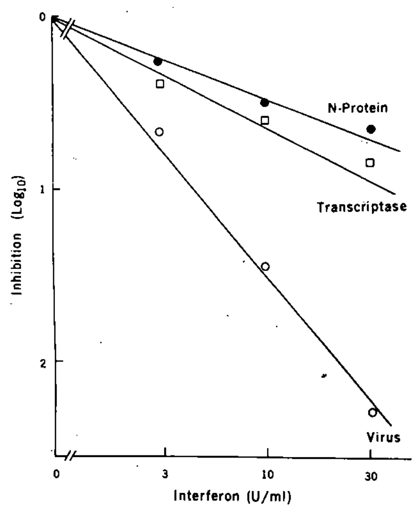

Figure 2.61 SCALE BREAKS. Scale breaks are a major cause of problems for graphs in science and technology. On this graph the lines drawn through the partial scale break have no meaning and give the misleading impression that the pattern of the data goes linearly through the origin. Since the horizontal scale is logarithmic, zero is actually at minus infinity.
Figure republished from [92]. Copyright 1980 by the AAAS.

from William Playfair's *Statistical Breviary* [109], published in 1801. The data are the populations of 22 European cities. Without a break, most of the data would be forced into a small region of the scale, which would degrade the resolution. (The dotted lines are allowed to cross the break because they carry no quantitative information that is distorted by the break.) The log scale in Figure 2.64 also improves the resolution. For most purposes a log scale is preferable to a broken one; all data can be readily compared with the log scale, whereas values on different panels of a broken scale can only be compared by the highly cognitive task of looking at the tick mark labels, reading off the values, and comparing them by doing mental arithmetic.

2.5 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, 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 [120, p. 21, p. 72] encourage clarity by "Use definite, specific, concrete language," and encourage brevity by "Do not

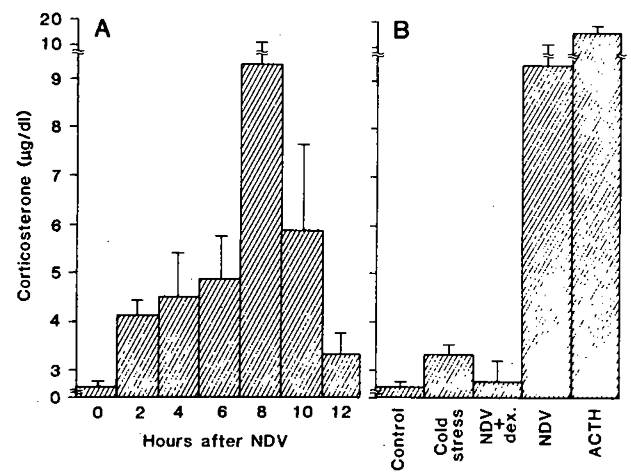

Figure 2.62 SCALE BREAKS. The lengths of the bars that barge right through the scale breaks have no meaning.
Figure republished from [116]. Copyright 1982 by the AAAS.

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.

In the past, the number of values that could be put on a graph was limited by the graph having to be made by hand. Computer graphics has removed these shackles. Now the number of values is limited only by the resolution of graphics devices and the perceptual ability of our visual system.

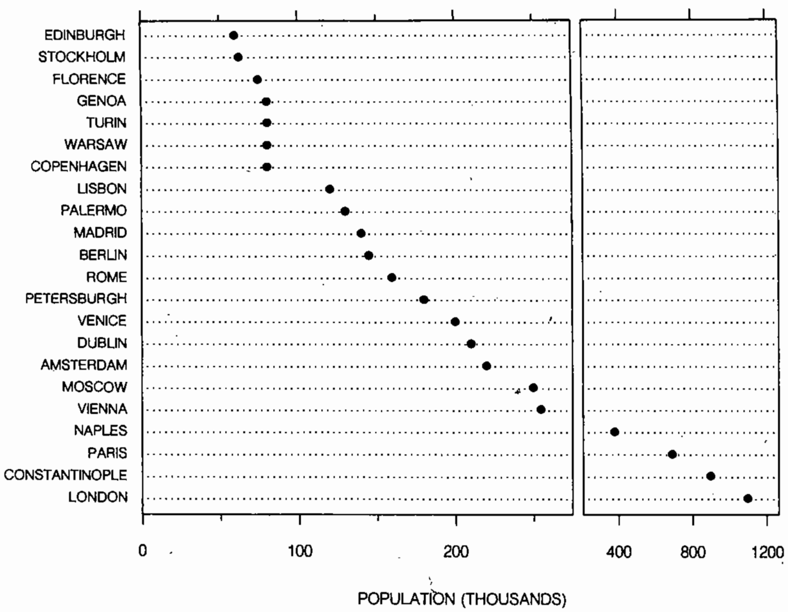

Figure 2.63 SCALE BREAKS. Without the scale break used on this graph, most of the data would be forced into a small region of the graph, which would degrade the resolution.

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 dataset on a graph in an uncluttered way. Figure 2.65, the graph of the CO_2 data and its three components that we have seen before, is an example. There are 276 monthly data points on each of the panels of this graph, which is 1104 points altogether. Each data point consists of two numbers, a value on the horizontal scale and a value on the vertical scale. Thus 2208 numbers are shown on this graph.

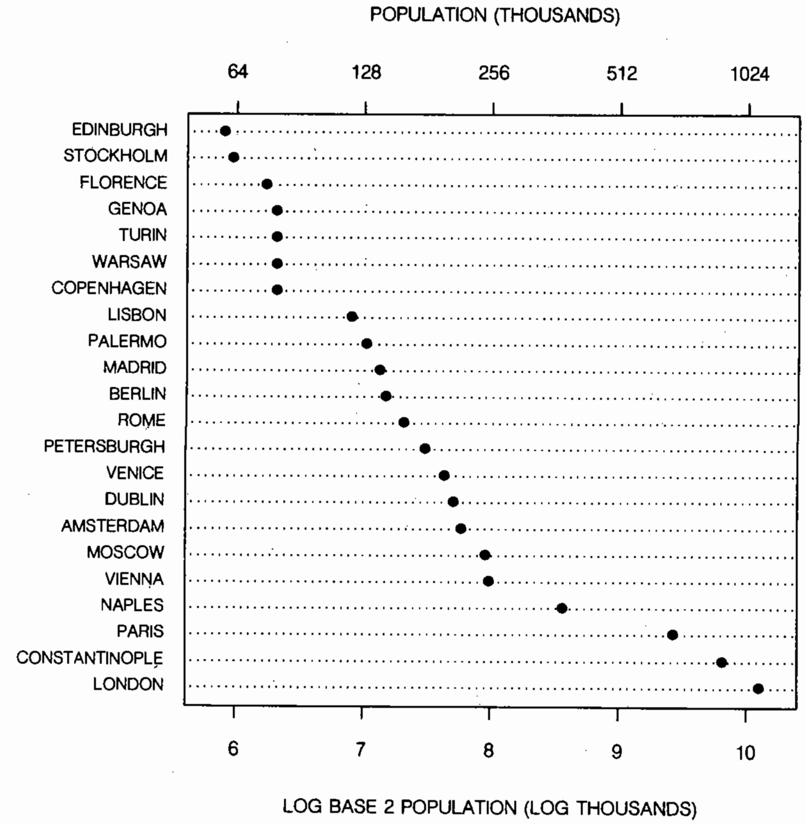

Figure 2.64 SCALE BREAK. The data from Figure 2.63 are graphed on a log scale, which relieves the need for a scale break.

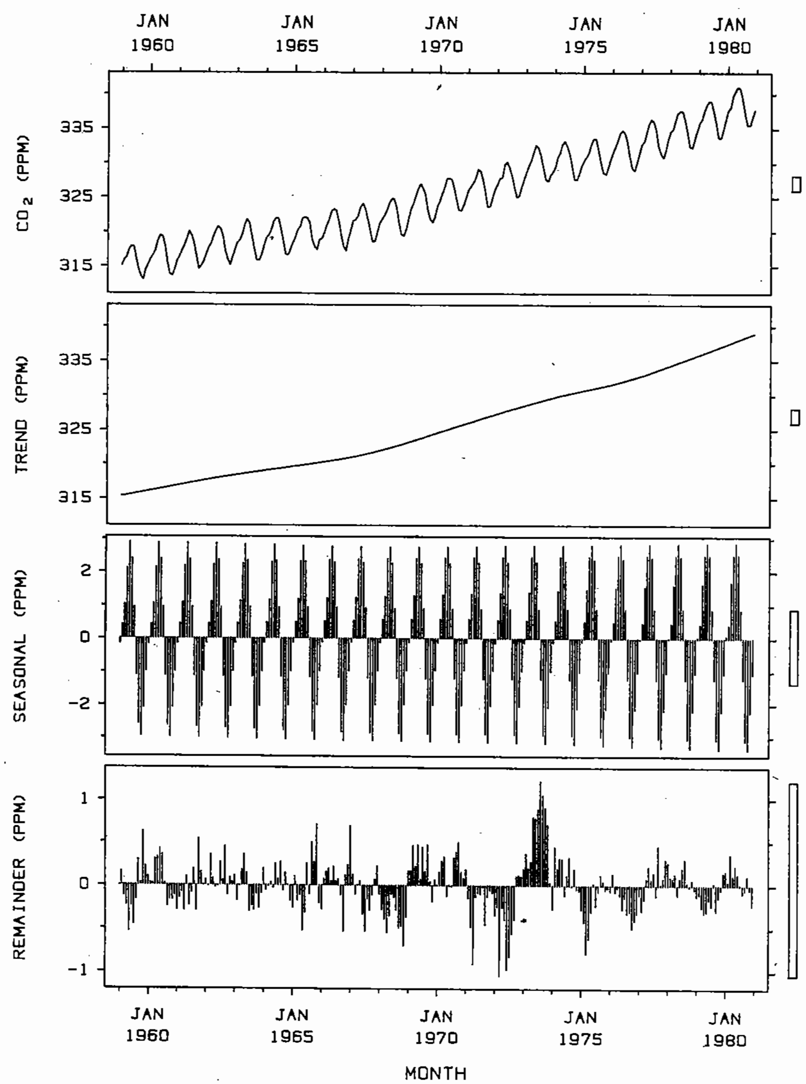

Figure 2.65 PACKING DATA. *A large amount of quantitative information can be packed into a small region.* The computer graphics revolution has given us the capability to graph a large amount of quantitative information in a small space. There are 1104 data points on this graph; each portrays two numerical values, so 2208 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. In many cases when we make a graph it is immediately clear that some aspect is inadequate and we regraph the data. In many other cases we make a graph, and all is well, but we get an idea for studying the data in a different way with a different graph; one successful graph often suggests another.

In part, graphing data needs to be iterative because we often do not know what to expect of the data; 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 2.66 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 [100].

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 a numerical procedure called lowess that will be described in Section 4 of Chapter 3. 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 2.66 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. (In fact, the real-life graphing that produced Figure 2.66 did require more iterations.)

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 2.66, it would be entirely sensible to use Graph 2 and Graph 4; both show interesting aspects of the data. Figure 2.67 is another example. Each of the three sets of data is shown twice. Graphing each data set separately in the top three panels allows the error bars to be perceived without interfering with one another. Graphing the three data sets together in the bottom panel allows them to be more effectively compared.

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 [73]. 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].

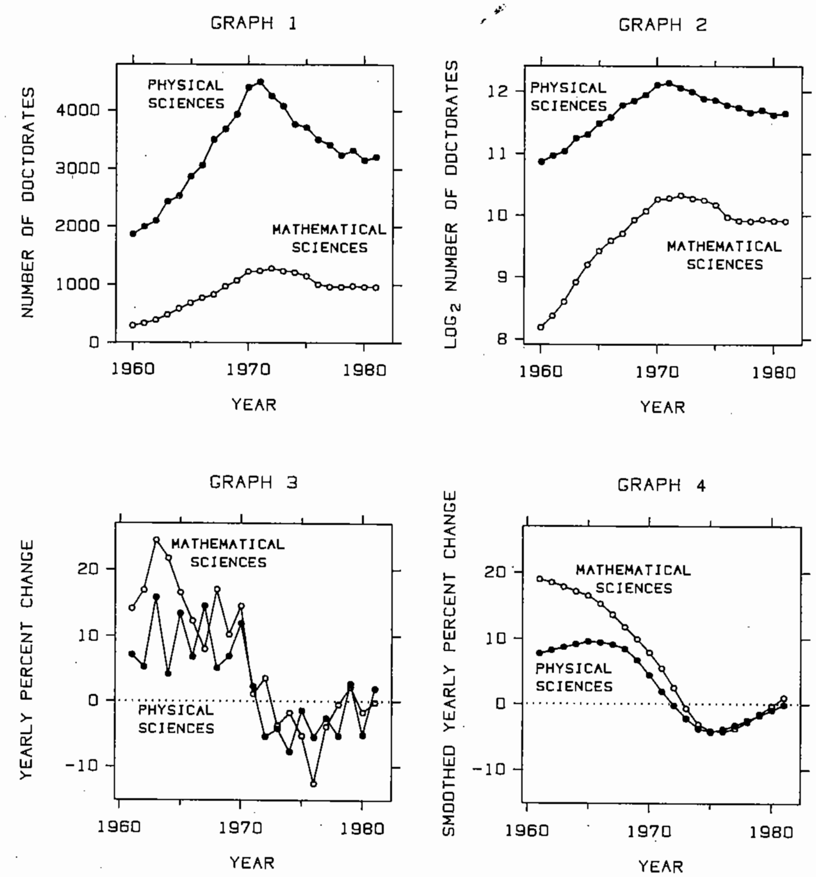

Figure 2.66 ITERATION. *Graphing data should be an iterative, experimental process.* The four graphs in this figure are four successive looks at the data; each of the last three is inspired by its predecessor.

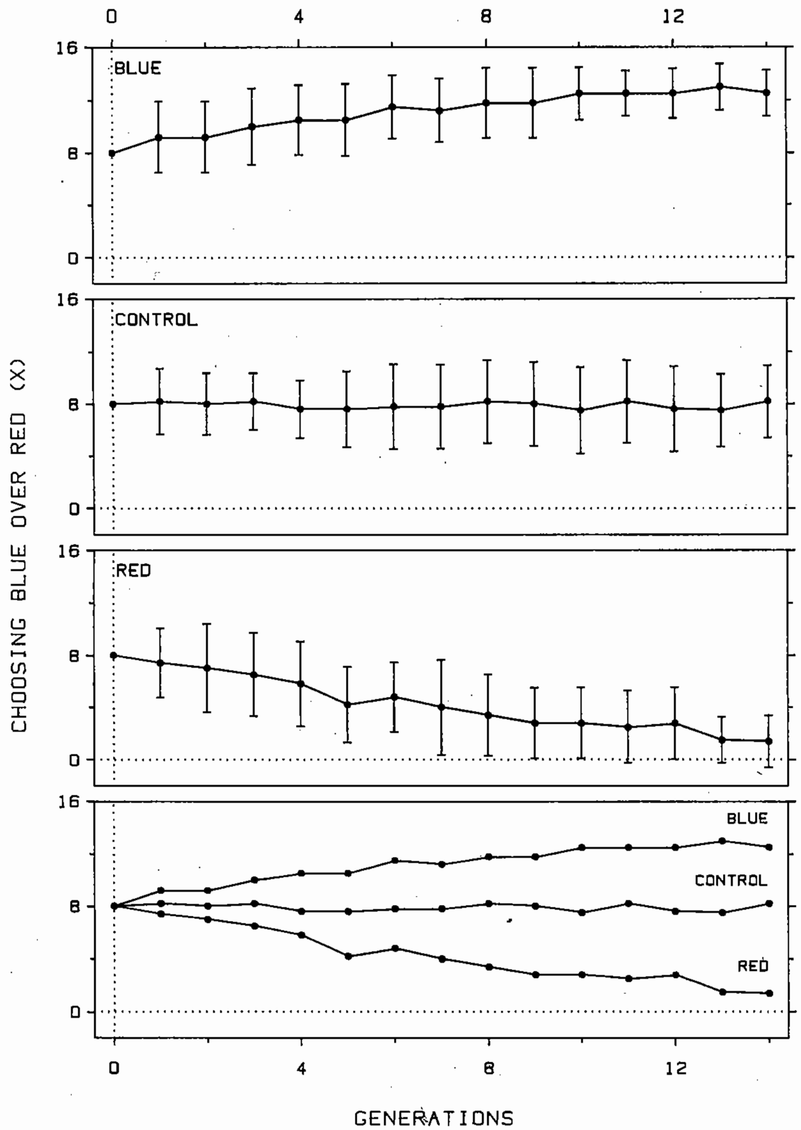

Figure 2.67 REGRAPHING. *Graph data two or more times when it is needed.* Each data set is graphed twice, once in one of the three top panels to allow an unobstructed view of the error bars and once in the bottom panel to allow an effective comparison of the data sets.

Table 2.1 shows data that Burt published in 1961 in the *British Journal of Statistical Psychology* [20]. 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. The data are means of adult IQ scores and means of child IQ scores for six social classes. The means were computed from IQ scores for 40,000 father-child pairs.

	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 2.68 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 [38] had put forward the equation

$$\text{child score} - 100 = \frac{1}{2}(\text{adult score} - 100)$$

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 2.68. 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 2.68 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.

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 2.69 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 6 of Chapter 3. The data in Figure 2.69 are measurements of four variables: wind speed, temperature, solar radiation at ground level, and concentrations of the air pollutant, ozone [18]. There is one measurement of each variable on each of 111 days.

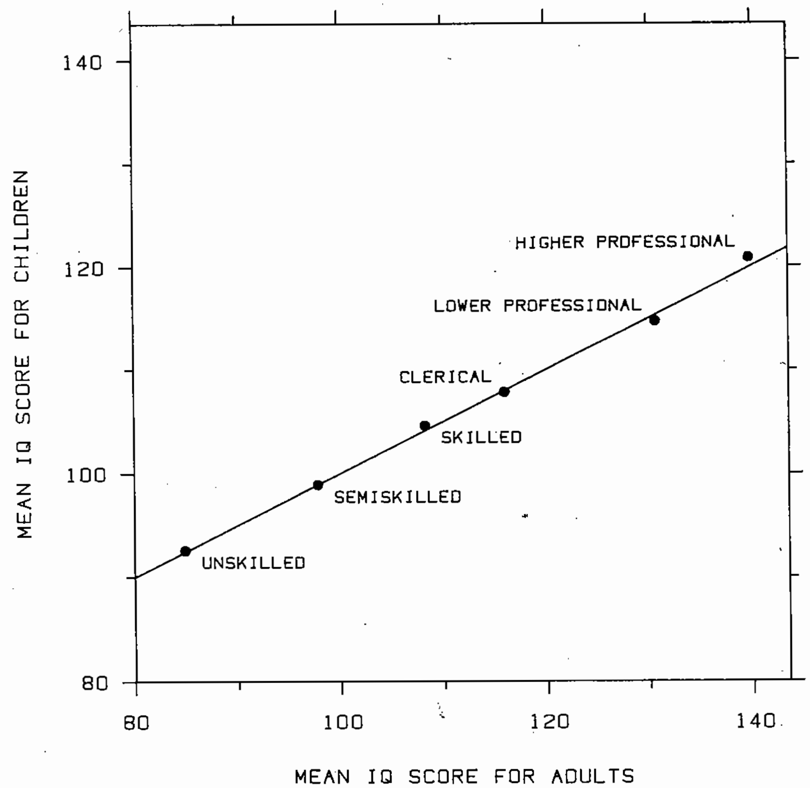

Figure 2.68 DETAILED STUDY. The important information on this graph, that Cyril Burt's fictitious data lie very close to the line, can be extracted with just a quick look.

Each panel of Figure 2.69 is a scatterplot of one variable against another. For the three panels in the second 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 row and first column — is a scatterplot of ozone against solar radiation; position (2,3) is a scatterplot of ozone against temperature; position (2,4) 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 Chapter 3; it

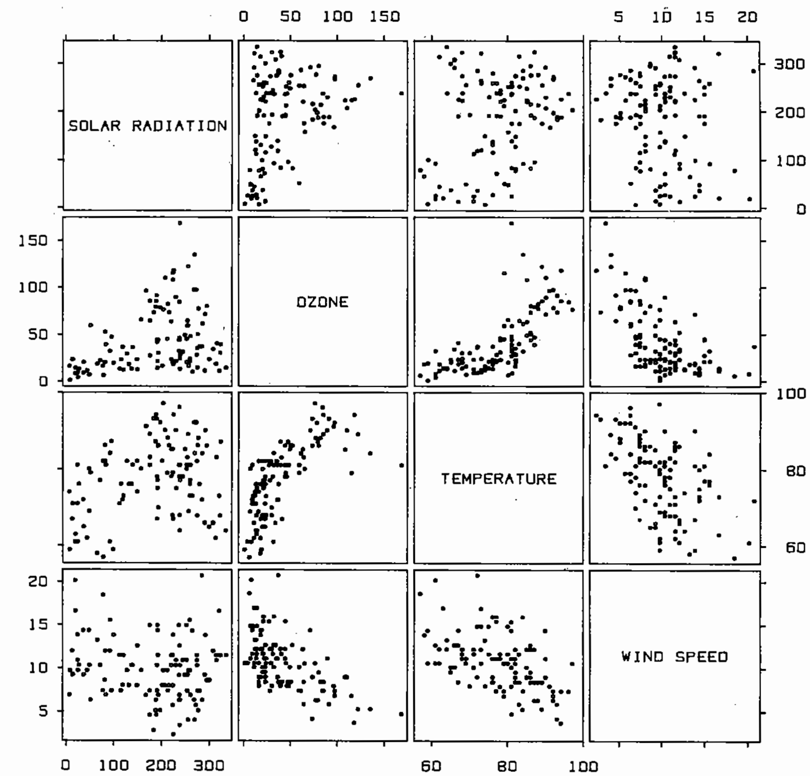

Figure 2.69 DETAILED STUDY. *Many useful graphs require careful, detailed study.* Compared with that needed for Figure 2.68, this scatterplot matrix of ozone and meteorological measurements requires lengthy study to extract the information. But the lengthy study reveals information that would be very difficult or impossible to get by other means.

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 2664 numbers, would require careful study.

2.6 A LISTING OF THE PRINCIPLES OF GRAPH CONSTRUCTION

Clear Vision

Make the data stand out. Avoid superfluity.

Use visually prominent graphical elements to show the data.

Use a pair of scale lines for each variable. Make the data region the interior of the rectangle formed by the scale lines. Put tick marks outside of the data region.

Do not clutter the data region.

Do not overdo the number of tick marks.

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.

Do not allow data labels in the data region to interfere with the quantitative data or to clutter the graph.

Avoid putting notes, keys, and markers in the data region. Put keys and markers just outside the data region and put notes in the legend or in the text.

Overlapping plotting symbols must be visually distinguishable.

Superposed data sets must be readily visually discriminated.

Visual clarity must be preserved under reduction and reproduction.

Clear Understanding

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

Error bars should be clearly explained.

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

Proofread graphs.

Strive for clarity.

Scales

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

Subject to the constraints that scales have, choose the scales so that the data fill up as much of the data region as possible.

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

Choose appropriate scales when graphs are compared.

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

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

Showing data on a logarithmic scale can improve resolution.

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.

General Strategy

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

Graphing data should be an iterative, experimental process.

Graph data two or more times when it is needed.

Many useful graphs require careful, detailed study.