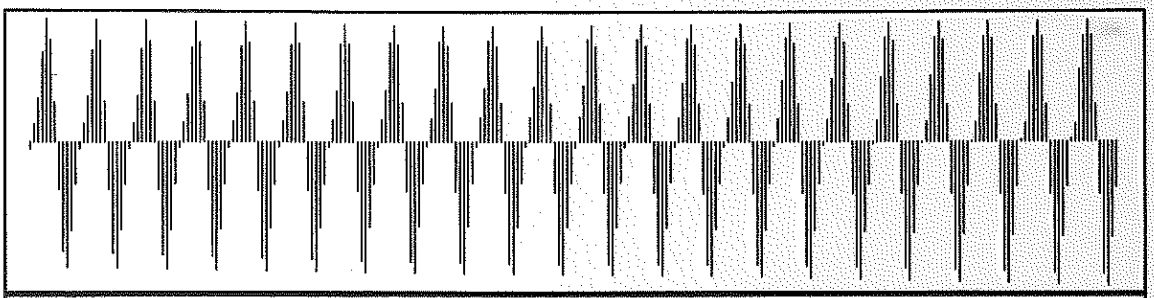
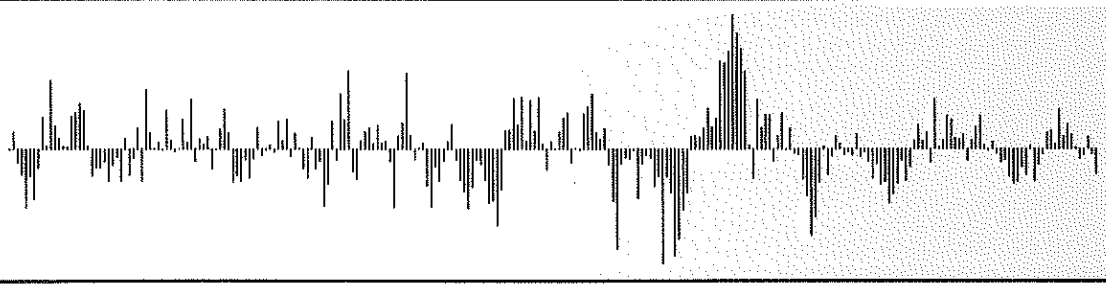


The Elements of Graphing Data

William S. Cleveland



2 *Principles of Graph Construction*

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

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

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

2.1 *Terminology*

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

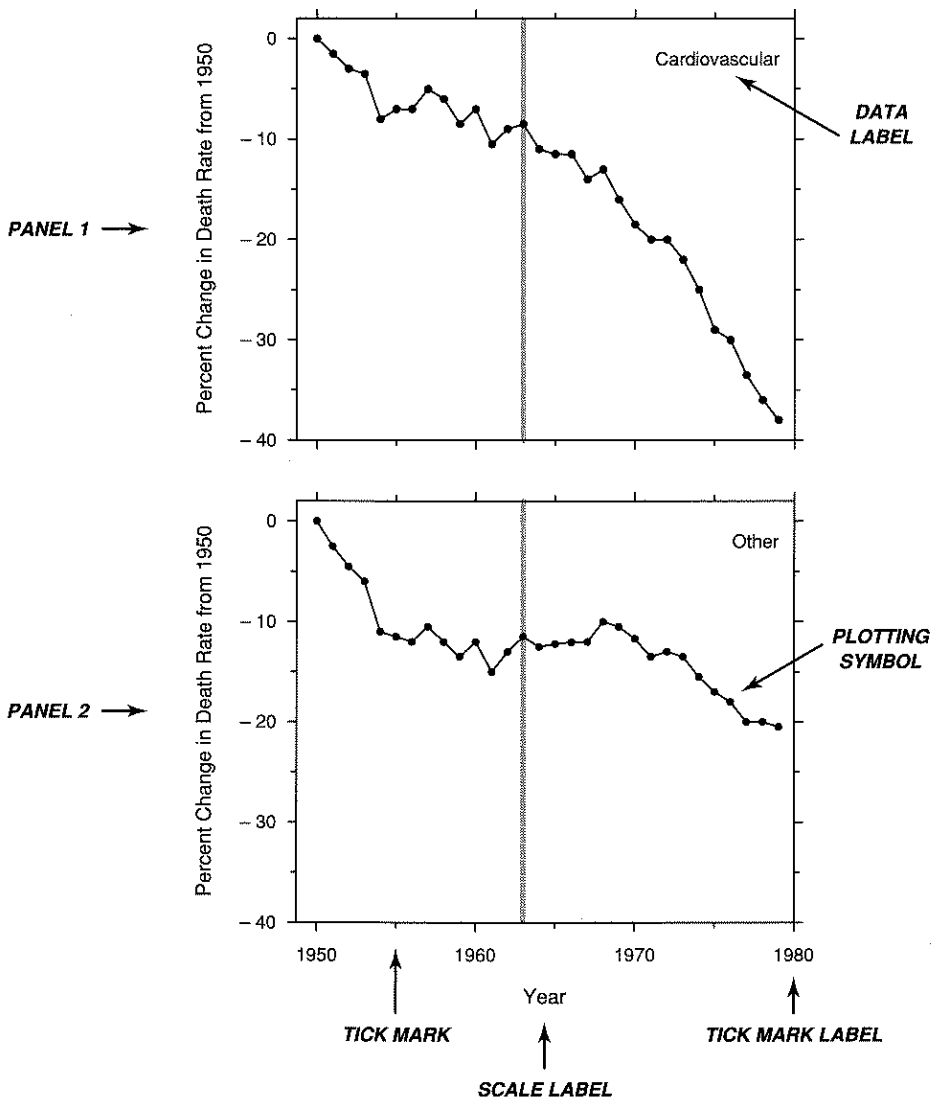


Figure 19. AGE-ADJUSTED DEATH RATE. The data are the percent changes from 1950 in death rate in the United States due to cardivascular disease and other diseases.

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

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

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

2.2 Clear Vision

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

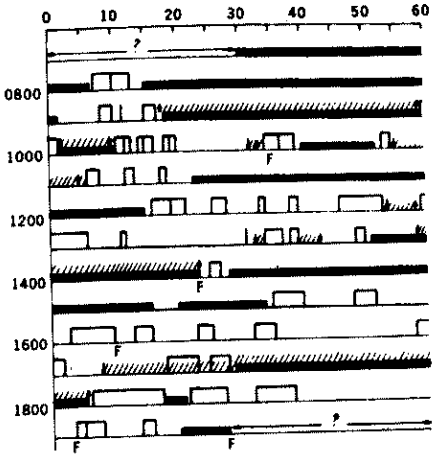
Make the data stand out. Avoid superfluity.

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

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

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

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



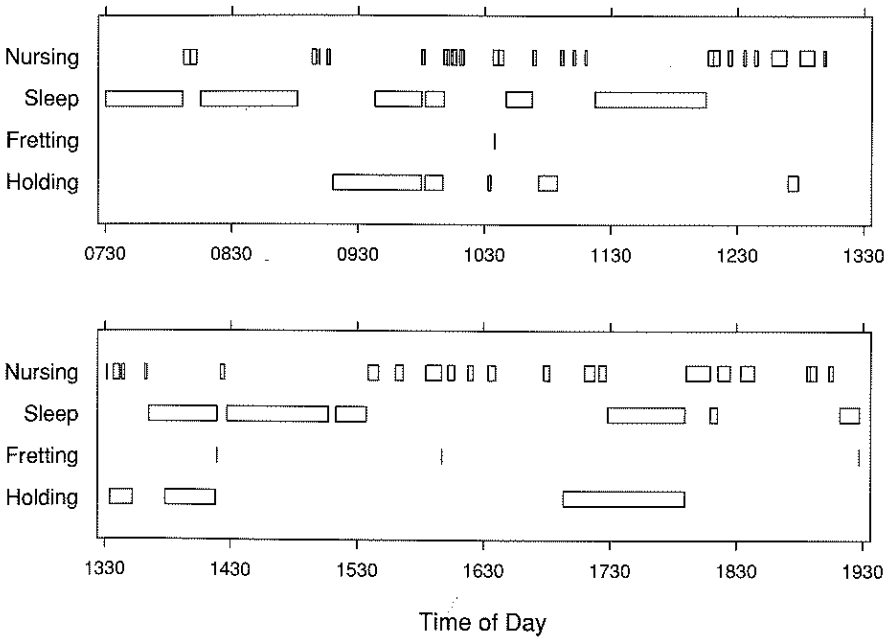
2.3 SUPERFLUITY AND STANDING OUT. The graph shows the activities of a !Kung woman and her baby. The open bars and vertical lines are nursing times; the closed bars show times when the baby is sleeping; F means fretting; and slashed lines are intervals when the baby is held by the mother, with arrows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

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

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

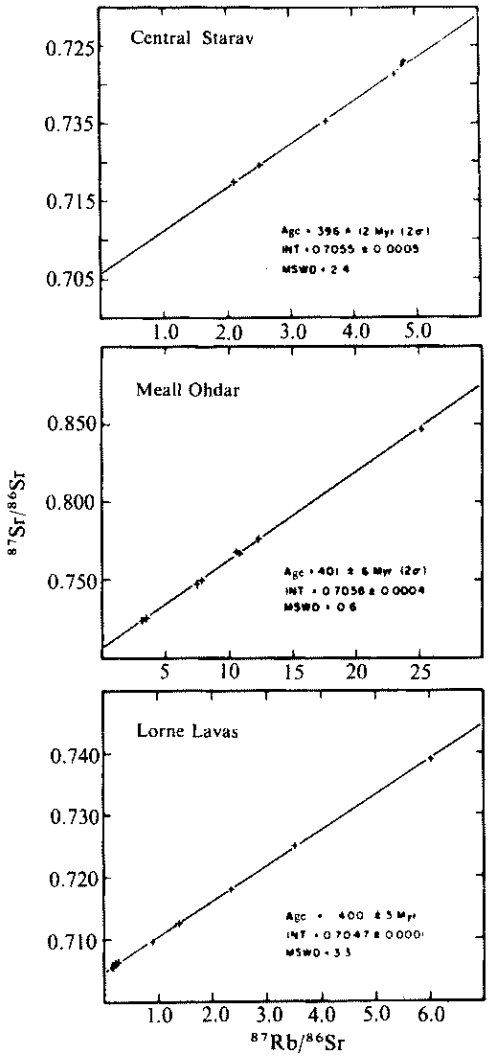
variability of each activity and it is difficult to remember which symbol goes with which activity, so that constant referring to the caption is necessary. A minor problem with Figure 2.3 is that the arrows for picking up and setting down are superfluous.

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



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

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

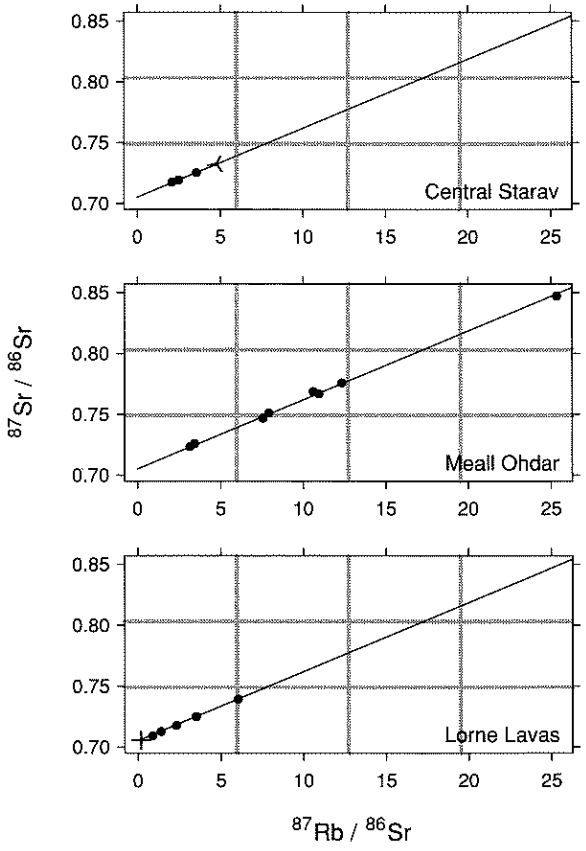


2.5 VISUAL PROMINENCE. The data do not stand out.

Use visually prominent graphical elements to show the data.

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

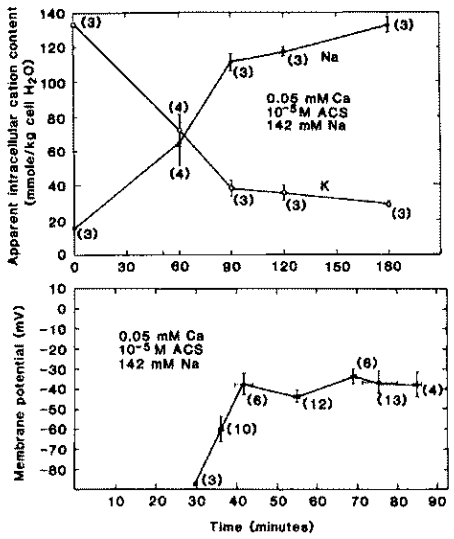
A good way to help the data to stand out is to show them with a graphical element that is visually prominent. This is illustrated in Figure 2.6; the data from Figure 2.5 are regraphed. The symbols showing the data stand out, and now the data can be seen. The symbols that look like the spokes of a wheel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



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

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

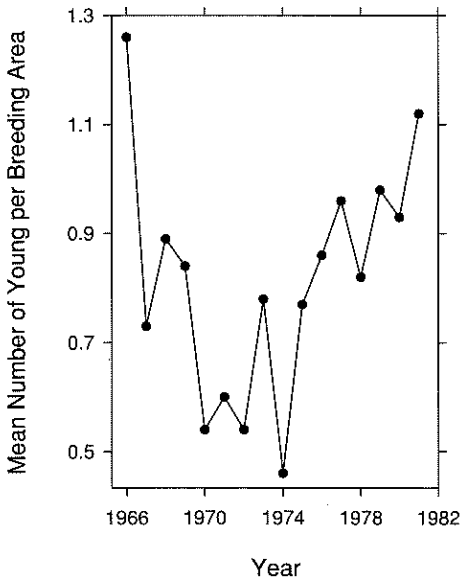
When plotting symbols are connected by lines, the symbols should be prominent enough to prevent being obscured by the lines. In Figure 2.7 the data and their standard errors are inconspicuous, in part because of the connecting lines [10].



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

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

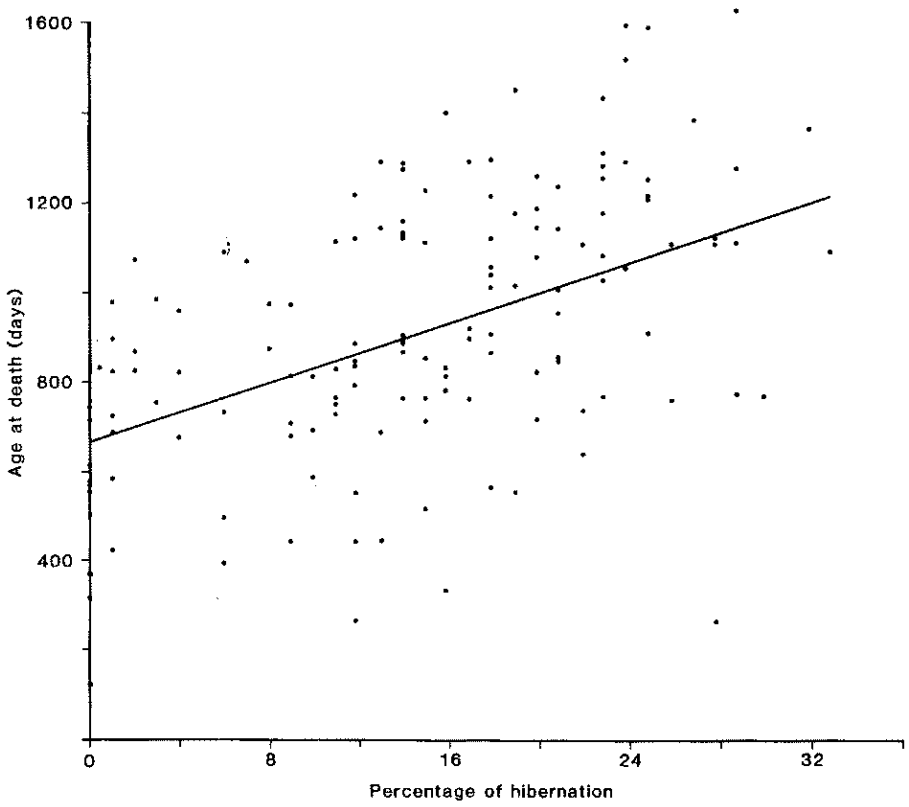
The data in Figure 2.8 are from observations of nesting sites of bald eagles in northwestern Ontario [55]. The graph shows good news: in 1973 DDT was banned, and after the ban, the average number of young per site began increasing.



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

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

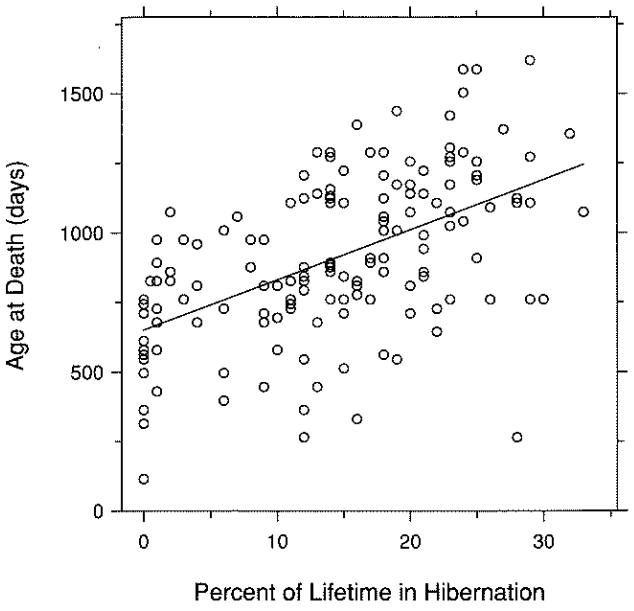
Data are frequently obscured by graphing them on top of scale lines. One example is Figure 2.9 where points are graphed on top of the vertical scale line. The graph and data of Figure 2.9 are from an interesting experiment run by four Harvard anatomists — Charles Lyman, Regina O'Brien, G. Cliett Greene,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



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

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

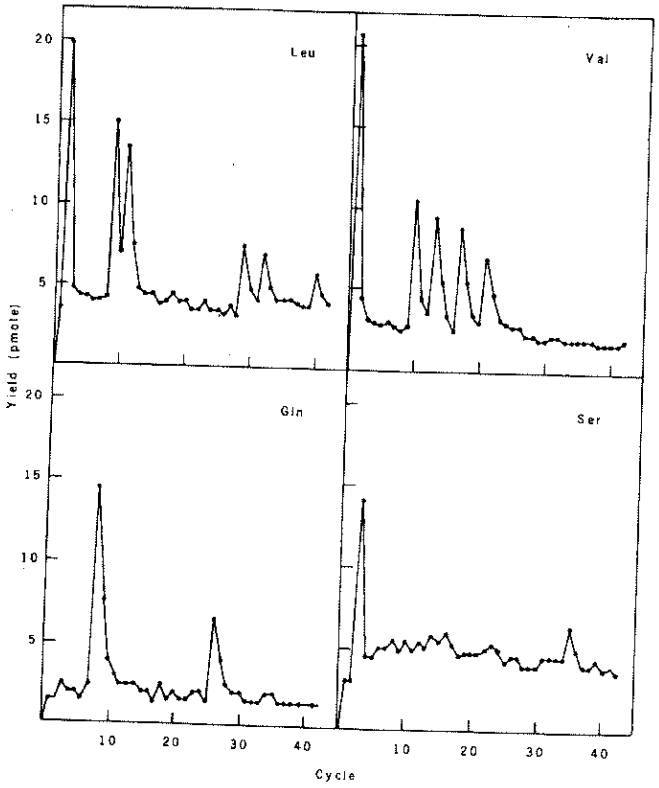
One unfortunate aspect of Figure 2.9 is that the data for hamsters with zero hibernation are graphed on top of the vertical scale line. This obscures the data to the point where it is hard to perceive just how many points there are. No data should be so obscured. One way to avoid this is shown in Figure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



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

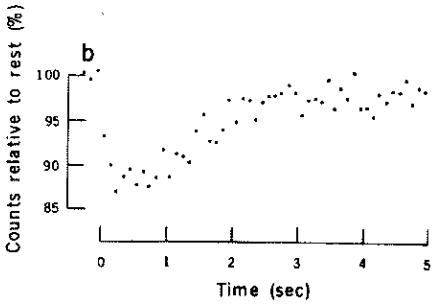
Four scale lines are used in Figure 2.10 rather than the two of Figure 2.9. Table look-up — judging the scale value of a point by judging its position along a scale line — is easier and more accurate as the distance of the point from the scale line decreases. The consequence of one vertical scale line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

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

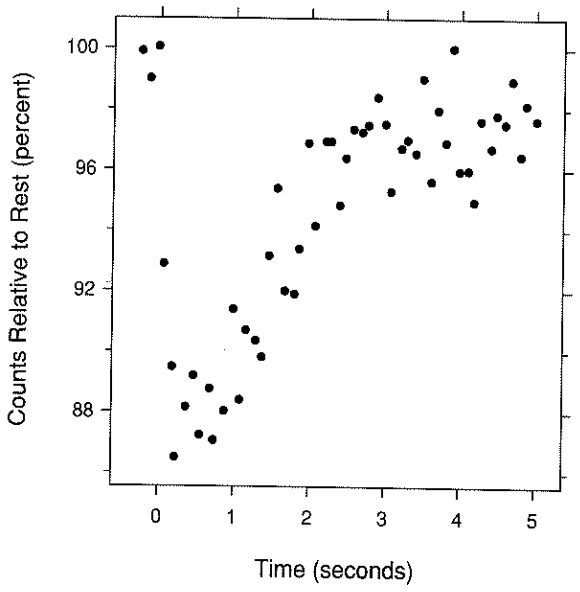


2.11 SCALE LINES AND THE DATA RECTANGLE. Tick marks that point inward can obscure data.

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



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

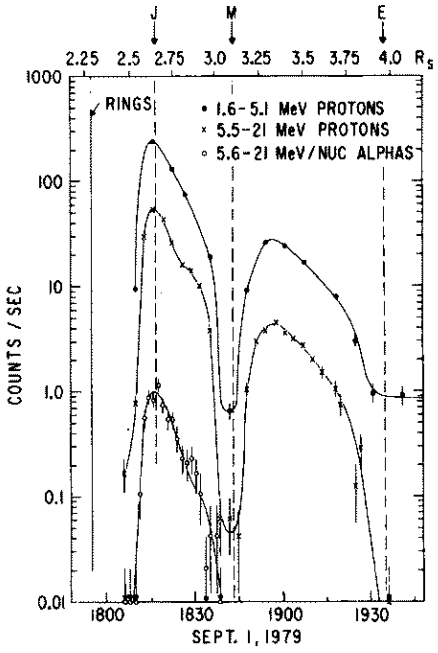


2.13 SCALE LINES AND THE DATA RECTANGLE. The four scale lines provide a clearly defined region for our eyes to look for data. Now, none of the data from Figure 2.12 are in danger of being overlooked.

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

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

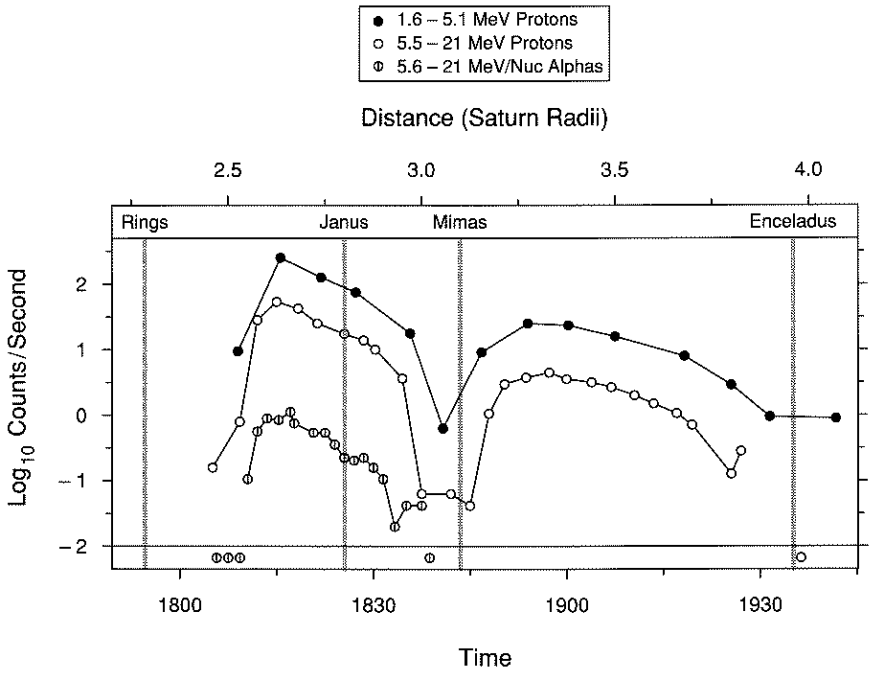
The data are shown again in Figure 2.15. The clutter has been alleviated, in part, by removing the error bars. It would be prudent to convey accuracy for these data numerically rather than graphically; on a log scale the error bars decrease radically and disappear from sight as



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

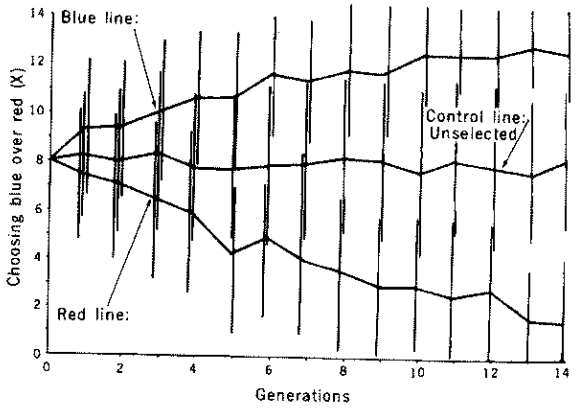
the counts increase. (It is possible that accuracy is nearly constant on a scale of square root counts/sec since count data of this sort tend to have a Poisson distribution. Thus accuracy might be conveyed more readily on the square root scale rather than on the log scale.) Other removals have taken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

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

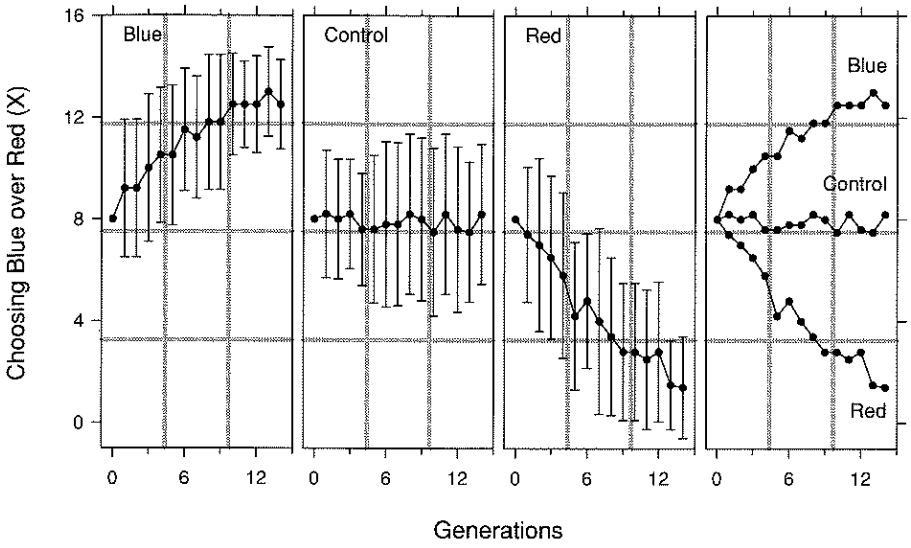


2.15 CLUTTER. *Do not clutter the interior of the scale-line rectangle.* The clutter of Figure 2.14 has been removed by alteration and excision. For example, the number of tick marks has been reduced.

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



2.16 CLUTTER. This graph is also cluttered.



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

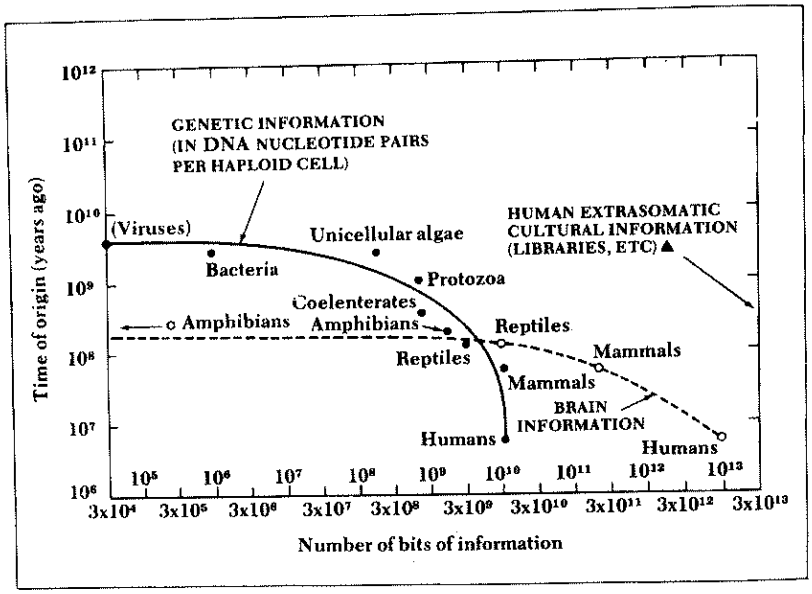
Do not overdo the number of tick marks.

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

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

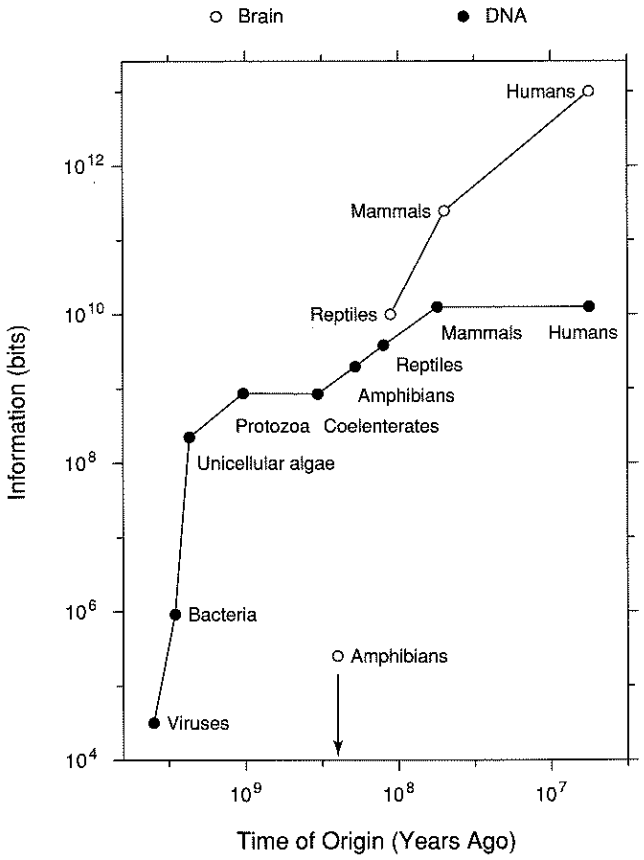
$$10^{7.5} = 10^{0.5}\, 10^7 \approx 3 \times 10^7 \,.$$

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



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

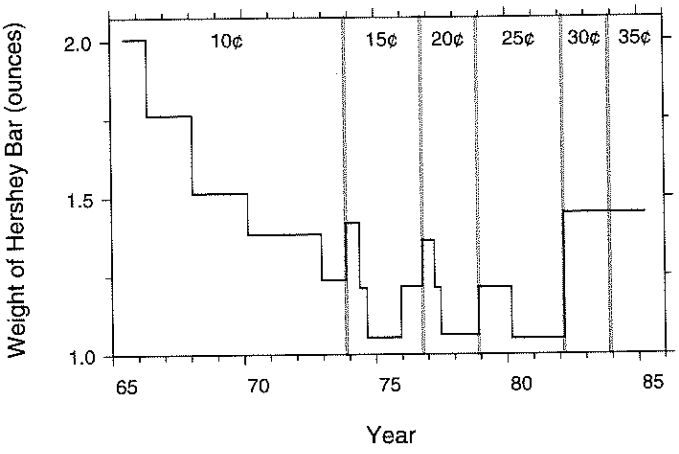
In Figure 2.19 the brain and DNA data are graphed again with fewer tick marks and labels; the horizontal and vertical scales have been interchanged so that time is now on the horizontal scale with earlier times on the left and later times on the right.



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

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

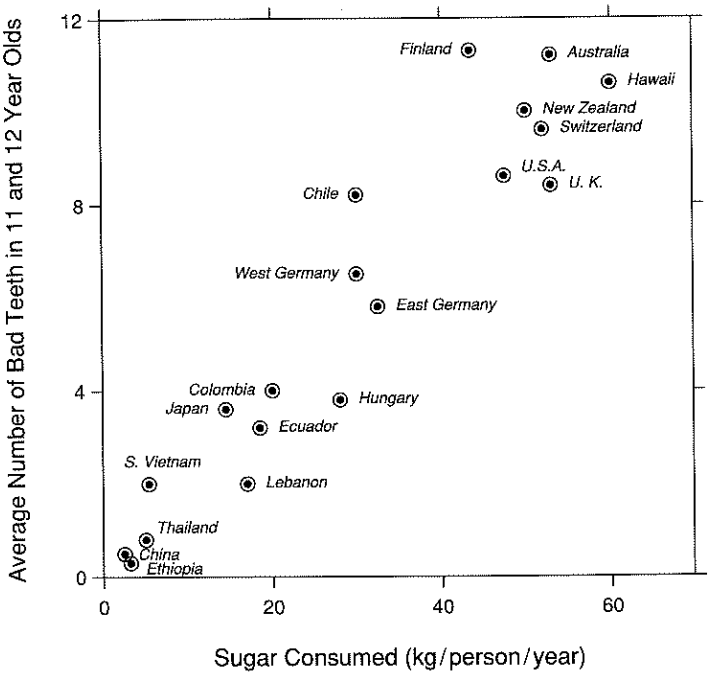
Reference lines are used in Figure 2.20. The data are the weights of the Hershey Bar, the famous American candy bar. These data, and Stephen Jay Gould's analysis of them [53], are discussed in detail in Section 3.8 (pp. 180–192). The vertical reference lines, which show times of price increases, cross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



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

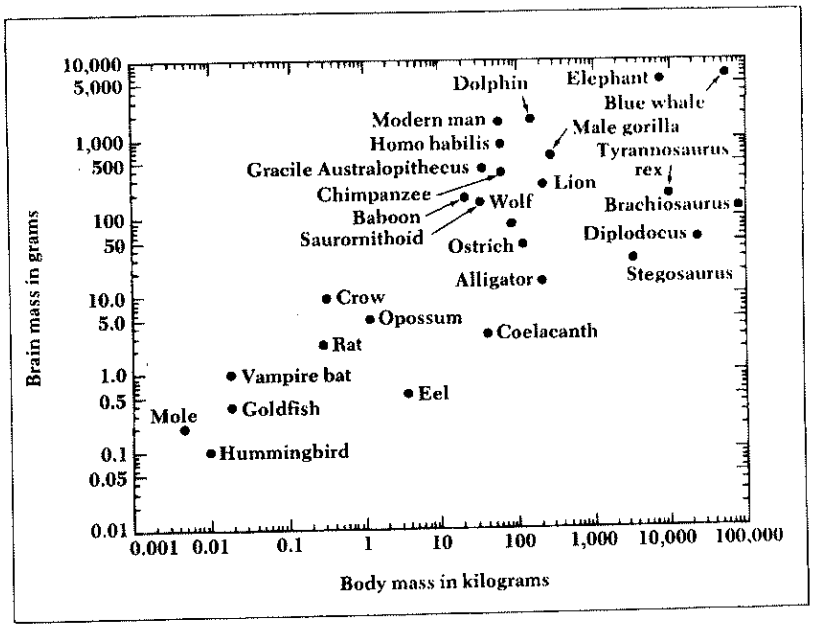
Do not allow data labels in the interior of the scale-line rectangle to interfere with the quantitative data or to clutter the graph.

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



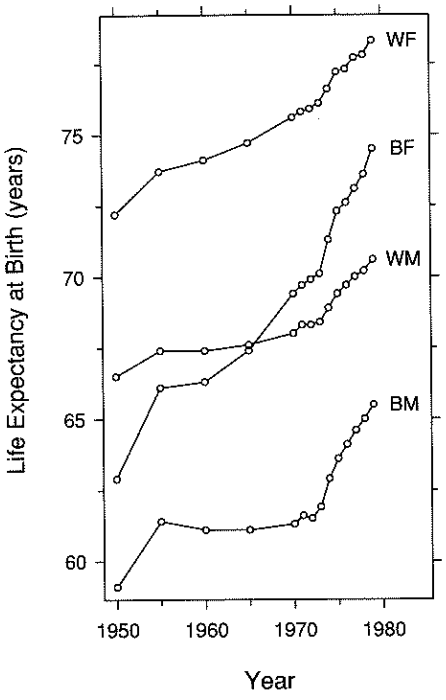
2.21 DATA LABELS. *Do not allow data labels in the interior of the scale-line rectangle to interfere with the quantitative data or to clutter the graph.* The data labels on this graph are needed to convey the names. The visual impact of the labels has been lessened so that they interfere as little as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

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



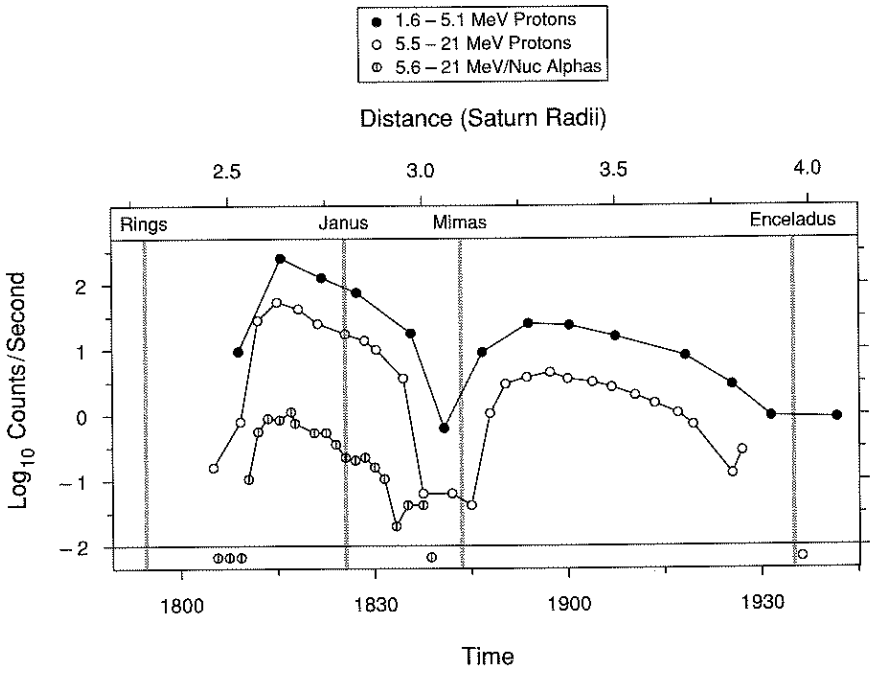
2.22 DATA LABELS. The data labels interfere with our visual assembly of the plotting symbols.

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



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

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



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

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

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

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

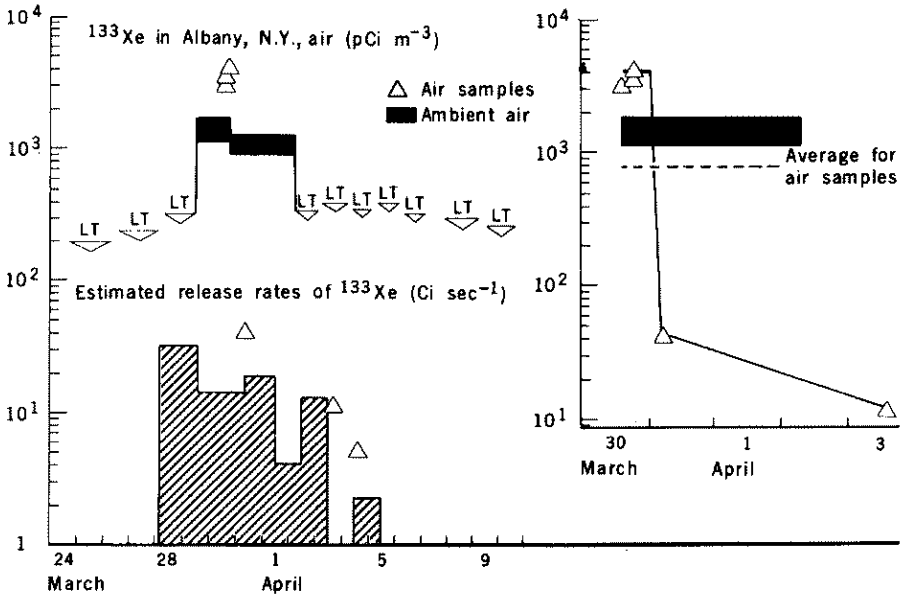


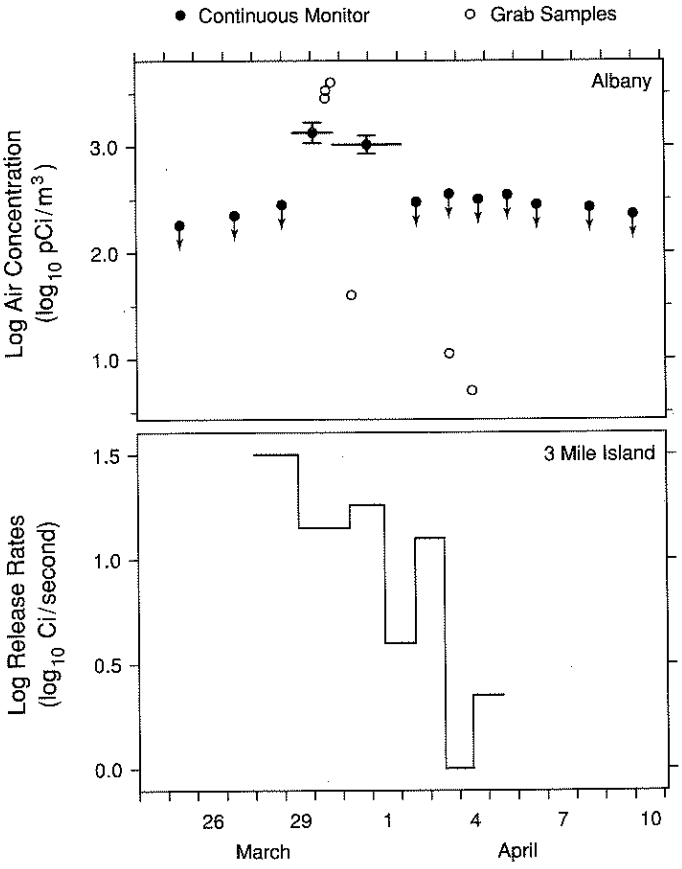
Fig. 1. Xenon-133 activity (picocuries per cubic meter of air) in Albany, New York, for the end of March and early April 1979. The lower trace shows the time-averaged estimates of releases (curies per second) from the Three Mile Island reactor (2). The inset shows detailed values for air samples (gas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

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

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

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

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

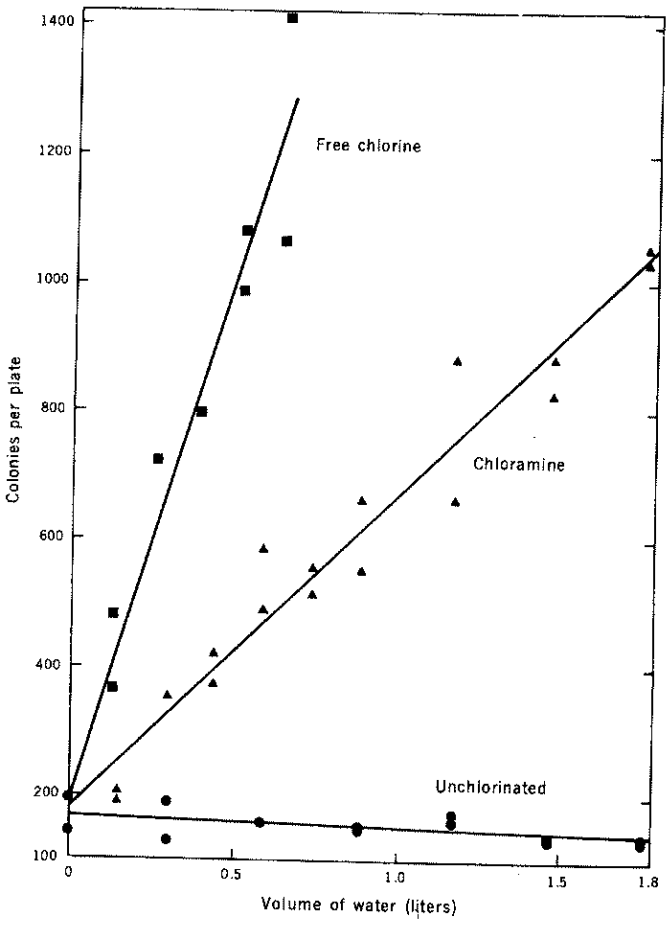


2.26 NOTES AND KEYS. *Avoid putting notes and keys inside the scale-line rectangle. Put a key outside, and put notes in the caption or in the text.* The graph in Figure 2.25 has been improved by the following actions: removing the writing and the key from the interior of the scale-line rectangle; removing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

Overlapping plotting symbols must be visually distinguishable.

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

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

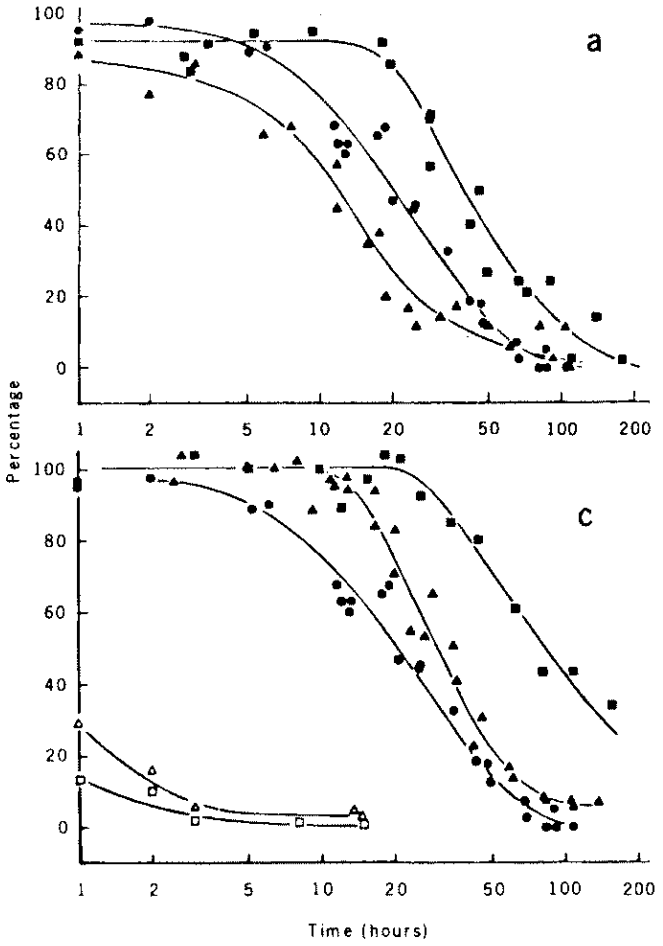


2.27 OVERLAPPING PLOTTING SYMBOLS. *Overlapping plotting symbols must be visually distinguishable.* On this graph, because of exact and near overlap, some of the data cannot be seen.

Superposed data sets must be readily visually assembled.

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

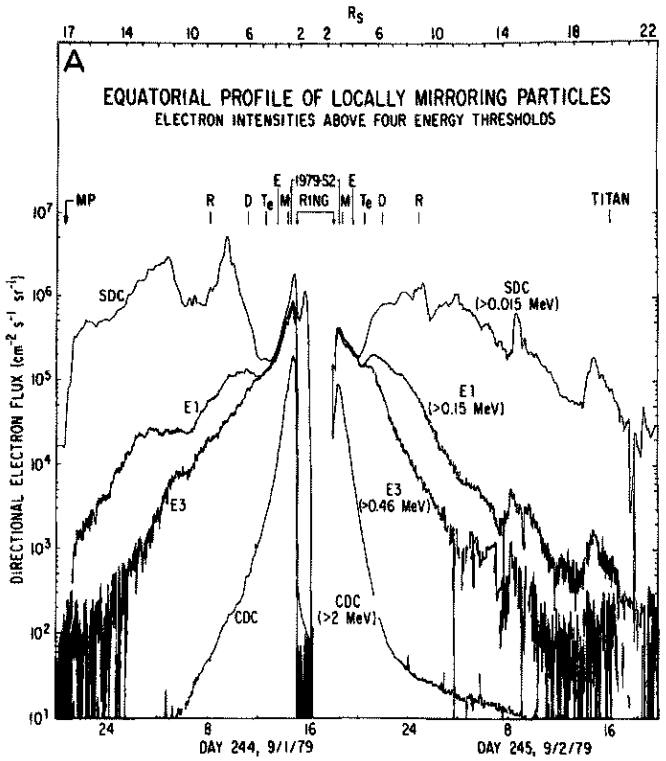
In Figure 2.28 [91] it is difficult to visually disentangle the solid squares, circles, and triangles; such plotting symbols are in general visually similar, but in Figure 2.28 the problem is exacerbated by the symbols not being crisply drawn.



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

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

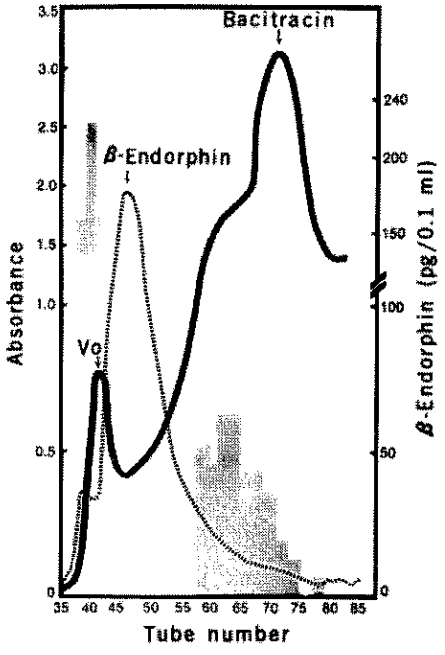
Graphs that fail to allow effective visual assembly are pervasive because the problem is a difficult one to solve. Solutions will be given in Section 3.5 (pp. 154–165) and Section 3.13 (pp. 209–212).



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

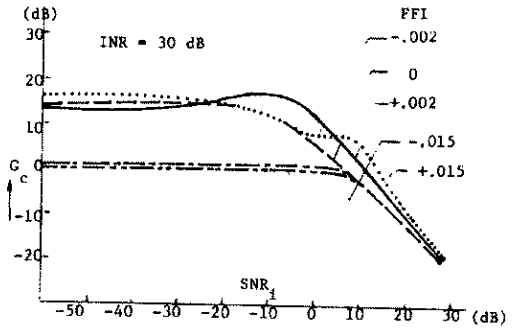
Visual clarity must be preserved under reduction and reproduction.

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



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

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



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

2.3 *Clear Understanding*

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

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

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

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

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

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