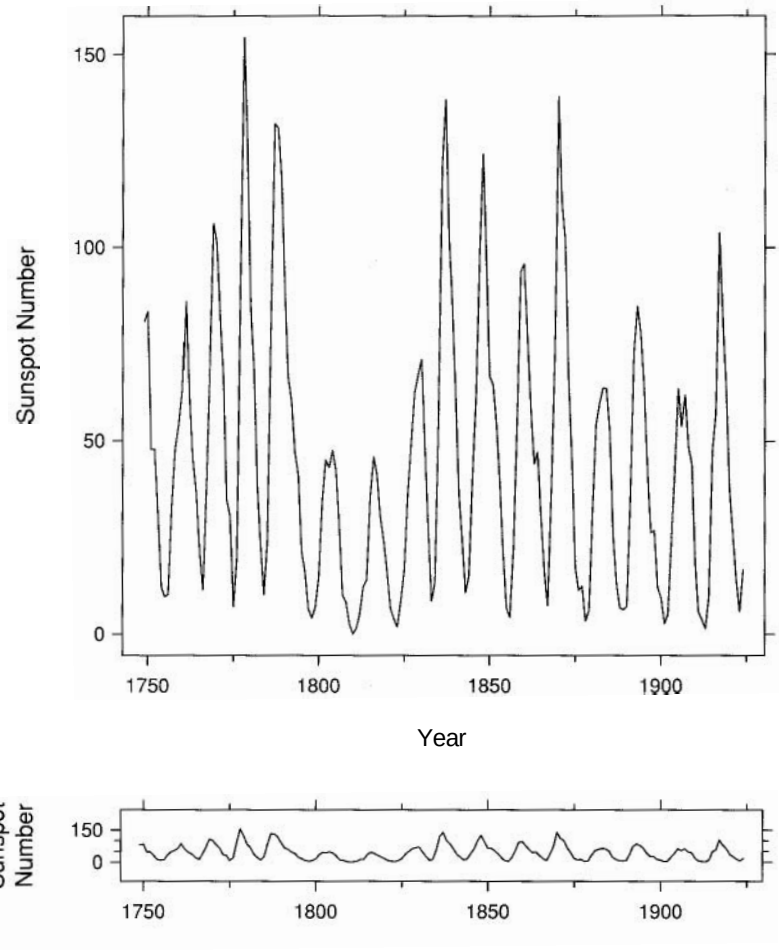


1.1 GRAPHICAL METHODS AND PRINCIPLES. The visualization of data requires basic principles and methods. Both panels of this graph show the yearly sunspot numbers from 1749 to 1924. A display method, banking to 45°, has been used to choose the shape, or aspect ratio, of the bottom panel. The method allows us to perceive an important property of the sunspots that is not revealed in the top panel — the sunspots rise more rapidly than they fall.

I Introduction

Data display is critical to data analysis. Graphs allow us to explore data to see overall patterns and to see detailed behavior; no other approach can compete in revealing the structure of data so thoroughly. Graphs allow us to view complex mathematical models fitted to data, and they allow us to assess the validity of such models.

But realizing the potential of data visualization requires methods and basic principles. Figure 1.1 illustrates this. The top panel graphs the yearly sunspot numbers from 1749 to 1924. The dominant frequency component of variation in the data is the cycles with periods of about 11 years. The existence of the cycles is clearly revealed, but an important property of them is not. And this property is critical to understanding the variation in the cycles, which in turn is critical to developing theories of solar physics that explain the origin of the sunspots. The problem is the shape, or aspect ratio, of the graph, a square. The data are graphed again in the bottom panel; a method called banking *to 45°,* which will be introduced in Chapter 2, is used to determine the aspect ratio, and the result is a narrow rectangle. Now the graph reveals the important property. The cycles typically rise more rapidly than they fall; this behavior is most pronounced for the cycles with high peaks, is less pronounced for those with medium peaks, and disappears for those cycles with the very lowest peaks.

This book is about methods and basic principles that help the data analyst to realize the potential of visualization. The next three chapters of the book divide the material into principles of graph construction, graphical methods, and graphical perception. In this chapter, Section 1.1 (pp. 6–9) demonstrates the power of visualization, Section 1.2 (pp. 9–15), demonstrates how easy it is for the graphing of data go wrong, and Section 1.3 (pp. 16–21) briefly describes the content of the next three chapters.

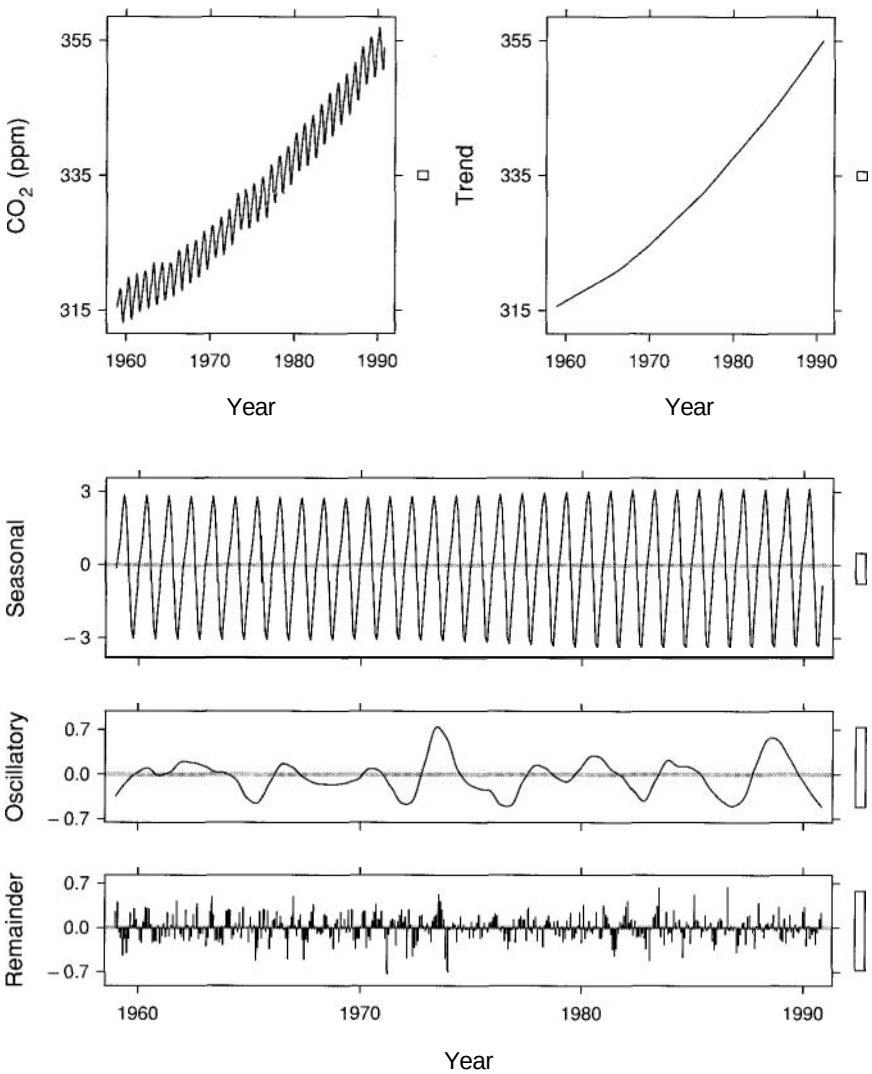
1.1 The Power of Graphical Data Display

Figure 1.2 illustrates the power of visualization to reveal complex patterns in data. The top left panel is a graph of monthly average atmospheric carbon dioxide concentrations measured at the Mauna Loa Observatory in Hawaii [9,71]. These data woke up the world. Charles Keeling pioneered their collection and fostered them amidst the adversity of nature at the top of a volcano and the controversy of man closer to sea level. The controversy raged first in science and then later in politics [108]. Earlier data had hinted that atmospheric CO_2 was rising due to man-made emissions, but Keeling's data proved the case, signaling the danger of global climate change.

The remaining panels of Figure 1.2 show a numerical decomposition of the data into four frequency components of variation whose sum is equal to the CO_2 concentrations. The decomposition was carried out by a statistical procedure, STL [21]. On the five vertical scales of the figure, the number of units per cm varies. The heights of the bars on the right sides of the panels provide a visualization of the relative scaling; the heights represent equal changes in parts per million on the five vertical scales.

The component graphed in the upper right panel is a trend component that describes the persistent long-term increase in the level of the concentrations. This rise, if continued unabated, will eventually cause atmospheric temperatures to rise, the polar ice caps to melt, the coastal areas of the continents to flood, and the climates of different regions of the earth to change radically [57,80,108]. And the graph shows that the rate of increase of CO_2 is itself increasing through time.

The component graphed in the third panel from the bottom is a seasonal component: a yearly cycle in the concentrations due to the waxing and waning of foliage in the Northern Hemisphere. When foliage grows in the spring, plant tissue absorbs CO_2 from the atmosphere, depositing some of the carbon in the soil, and atmospheric concentrations decline. When the foliage decreases at the end of the summer, CO_2 returns to the atmosphere, and the atmospheric concentrations increase. The graph shows that the amplitudes of these seasonal oscillations have increased slightly through time.



1.2 THE POWER OF GRAPHICAL DATA DISPLAY. Visualization provides insight that cannot be appreciated by any other approach to learning from data. On this graph, the top left panel displays monthly average CO_2 concentrations from Mauna Loa, Hawaii. The remaining panels show frequency components of variation in the data. The heights of the five bars on the right sides of the panels portray the same changes in ppm on the five vertical scales.

An oscillatory component, graphed in the second panel from the bottom, is made up mostly of variation with periods in a band centered near three years. This variation is associated with changes in the Southern Oscillation index, a measure of the difference in atmospheric pressure between Easter Island in the South Pacific and Darwin, Australia. Changes in the index are also associated with changes in climate. For example, when the index drops sharply, the trade winds are reduced and the temperature of the equatorial Pacific increases. This warming, which has important consequences for South America, often occurs around Christmas time and is called El Niño — the child [73].

The component shown in the bottom panel has no apparent, strong, time pattern and behaves, for the most part, like random noise.

Figure 1.2 conveys a large amount of information about the CO_2 concentrations. We have been able to summarize overall behavior and to see detailed information. As the eminent statistician W. Edwards Deming would have put it [45], "the graph retains the information in the data."

Many techniques of data analysis have data reduction as their first step. For example, classical statistical procedures, widely used in science and technology, fall in this category. The first step is to take all of the data and reduce them to a few statistics such as means, standard deviations, correlation coefficients, variance components, and t-tests. Then, inferences are based on this very limited collection of values. Using only numerical reduction methods in data analyses is far too limiting. We cannot expect a small number of numerical values to consistently convey the wealth of information that exists in data. Numerical reduction methods do not retain the information in the data.

Contained within the data of any investigation is information that can yield conclusions to questions not even originally asked. That is, there can be surprises in the data. The progress of science depends heavily on formulating hypotheses and probing them by data collection. Darwin, in a letter to Henry Fawcett in 1861, writes [54]: "How odd it is that anyone should not see that all observation must be for or against some view if it is to be of any service." But analyses of data should not narrowly focus on just those hypotheses that led to collection. This inhibits finding surprises in the data. To regularly miss surprises by failing to probe thoroughly with visualization tools is terribly inefficient

because the cost of intensive data analysis is typically very small compared with the cost of data collection.

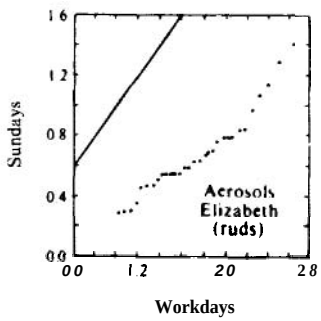
A graph of CO_2 concentrations similar to that of Figure 1.2 produced a surprise discovery. For a long time it was thought that the amplitude of the seasonal component was stable and not changing through time, but eventually three groups — one at CSIRO in Australia [102], a second at Scripps Institution of Oceanography in the United States [3], and a third at AT&T Bell Laboratories in the United States [30] — independently discovered the small, but persistent change in the Mauna Loa seasonal cycles. For the Bell Labs group, the discovery was serendipitous. The goal of the analysis had been to study the relationship between CO_2 and the Southern Oscillation index. The first step in the analysis was to decompose the CO_2 concentrations as in Figure 1.2 to get the oscillatory component so it could be correlated with the index. Fortunately the group graphed all of the components, and the graph showed clearly the persistent change in the amplitude of the seasonal component. This surprise was so exciting that the group switched its mission to the seasonal behavior of CO_2 and abandoned the original mission. No one yet has a good understanding of what is causing the change. It might be a harbinger of changes in the earth's climate or it might be simply part of the natural variation in CO_2.

1.2 The Challenge of Graphical Data Display

Visualization is surprisingly difficult. Even the most simple matters can easily go wrong. This will be illustrated by three examples where seemingly straightforward graphical tasks ran into trouble.

Aerosol Concentrations

Figure 1.3 is a graphical method called a q-q *plot* which will be discussed in detail in Chapter 3; the figure shows the graph as it originally appeared in a *Science* report [31]. As with almost all of the reproduced graphs in this book, the size of the graph is the same as that of the source. The display compares Sunday and workday concentrations of aerosols, or particles in the air. First, the graph has a construction error: the 0.0 label on the horizontal scale should be 0.6. Unfortunately, the error makes it appear that the left corner is the origin; many readers probably wondered why the line $y = x$, which is drawn on the graph, does not go through the origin. A second problem is that the scales on the graph are poorly chosen; comparison of the Sunday and workday values would have been enhanced by making the horizontal and vertical scales the same. Scale issues such as these are discussed in Chapter 2. Finally, the display of the data misses an opportunity to see the behavior of the data more thoroughly. On this single panel it is not easy to compare the vertical distances of the points from the line $y = x$; the solution is a graphical method called the *Tukey mean-difference plot*, which will be introduced in Chapter 3.

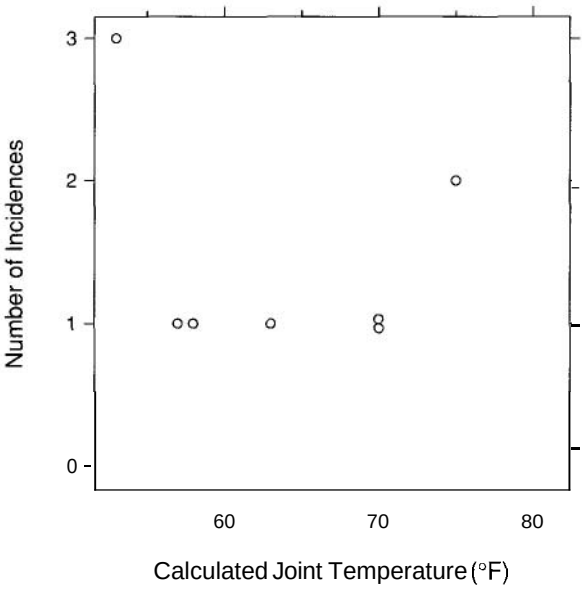


1.3 THE CHALLENGE OF GRAPHICAL DATA DISPLAY. This graph compares Sunday and workday concentrations of aerosols. The line shown is $y = x$. The graph has problems. There is a construction error: the 0.0 label on the horizontal scale is wrong and should be 0.6. The horizontal and vertical scales should be the same but are not. Furthermore, it is hard to judge the deviations of the points from the line $y = x$.

0-Ring Data

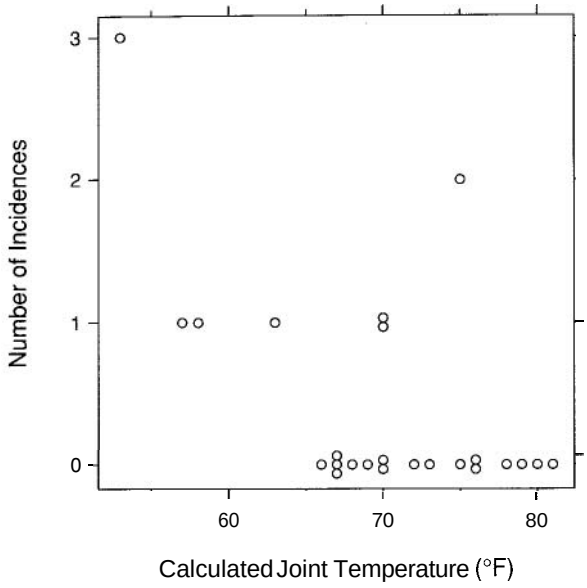
On January 27,1986, the day before the last flight of the space shuttle Challenger, a group of engineers met to study an alarm that had been raised. The forecast of temperature at launch time the following day was 31°. There was a suggestion that the low temperature might affect the performance of the O-rings that sealed the joints of the rocket motors.

To assess the issue, the engineers studied a graph of the data shown in Figure 1.4. Each data point was from a shuttle flight in which the O-rings had experienced thermal distress. The horizontal scale is O-ring temperature, and the vertical scale is the number of O-rings experiencing distress. The graph revealed no effect of temperature on the number of stress problems, and Morton Thiokel, the rocket manufacturer, communicated to NASA the conclusion that the "temperature data [are] not conclusive on predicting primary O-ring blowby" [43]. The next day Challenger took off, the O-rings failed, and the shuttle exploded, killing the seven people on board.



1.4 STATISTICAL REASONING. These data were graphed by space shuttle engineers the evening before the Challenger accident to determine the dependence of O-ring failure on temperature. Data for no failures was not graphed in the mistaken belief that it was irrelevant to the issue of dependence. The engineers concluded from the graph that there is no dependence.

The conclusion of the January 27 analysis was incorrect, in part, because the analysis of the data by the graph in Figure 1.4 was faulty. It omitted data for flights in which no O-rings experienced thermal distress. Figure 1.5 shows a graph with all data included. Now a pattern emerges. The Rogers Commission, a group that intensively studied the Challenger mission afterward, concluded that the engineers had omitted the no-stress data in the mistaken belief that they would contribute no information to the thermal-stress question [43].

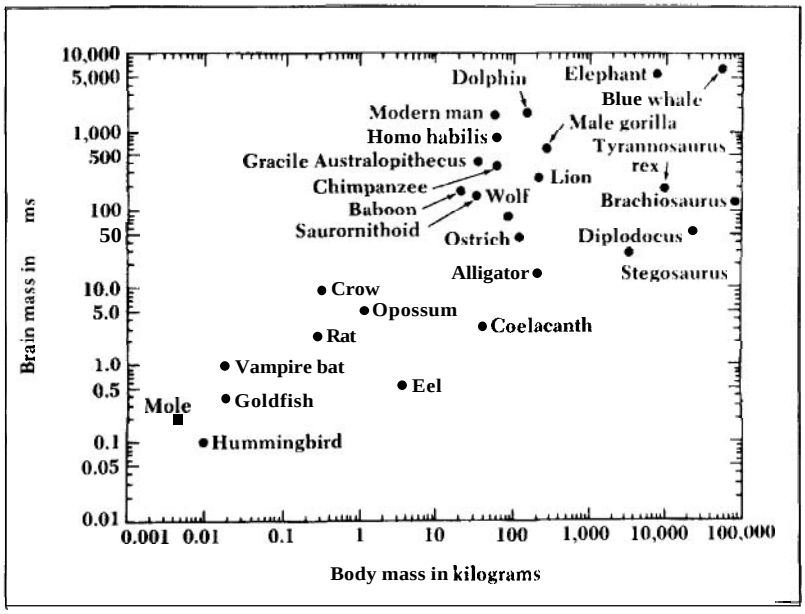


1.5 STATISTICAL REASONING. The complete set of O-ring data is now graphed, including the observations with no failures. A dependence of failure on temperature is revealed.

The graphical analysis of the O-ring data failed, not because of the display method used, as with the aerosol data, but rather because of a poor choice of the statistical information selected for the graph. This arose because of a flaw in the statistical reasoning that underlay the graph. The flaw violated a basic statistical principle: in the analysis of failure data, the values of a causal variable when no failures occur are as relevant to the analysis as the values when failures occur. Statistical thinking is vital to data display. A number of statistical principles are discussed in Chapters 2 and **3.**

Brain Masses and Body Masses of Animal Species

Figure 1.6 is a graph from Carl Sagan's intriguing book, *The Dragons of Eden* [107]. The graph shows the brain masses and body masses, both on a log scale, of a collection of animal species. We can see that log brain mass and log body mass are correlated, but this was not the main reason for making the graph.



1.6 THE CHALLENGE OF GRAPHICAL DATA DISPLAY. This graph shows brain and body masses of animal species. The intent was for viewers to judge an intelligence measure, but the judgments require a visual operation that is too difficult.

What Sagan wanted to describe was an intelligence scale that has been investigated extensively by Harry J. Jerison [65]. Sagan writes that this measure of intelligence is "the *ratio* of the mass of the brain to the total mass of the organism." Later he adds, referring the reader to the graph, "of all the organisms shown, the beast with the largest brain mass for its body weight is a creature called *Homo sapiens.* Next in such a ranking are dolphins."

The first problem is that Sagan has made a mistake in describing the intelligence measure; it is not the ratio of brain to body mass but rather is $(\text{brain mass})/(\text{body mass})^{2/3}$. If we study a group of related species, such as all mammals, brain mass tends to increase as a function of body mass. The general pattern of the data is reasonably well described by the equation

$$\text{brain mass} = c\,(\text{body mass})^{2/3}\,.$$

Since the densities of different species do not vary radically, we may think of the masses as being surrogate measures for volume, and volume to the 2/3 power behaves like a surface area. Thus the empirical relationship says that brain mass depends on the surface area of the body; Stephen Jay Gould conjectures that this is so because body surfaces serve as end points for so many nerve channels [52]. Now suppose a given species has a greater brain mass than other species with the same body mass; what this means is that

$$(\text{brain mass})/(\text{body mass})^{2/3}$$

is greater. We might expect that the big-brained species would be more intelligent since it has an excess of brain capacity given its body surface. This idea leads to measuring intelligence by this ratio.

Let us now return to Figure 1.6 and consider the graphical problem, which is a serious one. How do we judge the intelligence measure from the graph? Suppose two species have the same intelligence measure; then both have the same value of

$$\frac{(\text{brain mass})}{(\text{body mass})^{2/3}} = r\ .$$

Thus

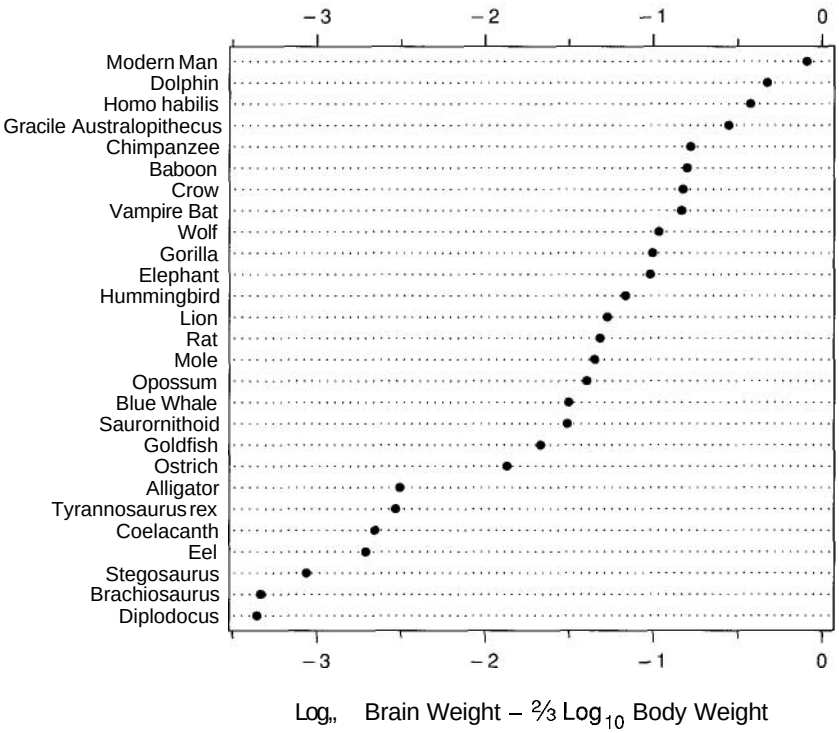
$$\log(\text{brain mass}) = 2/3 \log(\text{body mass}) + \log(r)$$

for both species. This means that in Figure 1.6, the two equally intelligent species lie on a line with slope 2/3. Suppose one species has a greater value of r than another; then the smarter one lies on a line with slope 2/3 that is to the northwest of the line on which the less intelligent one lies. In other words, to judge the intelligence measure from Figure 1.6 we must mentally superpose a set of parallel lines with slope 2/3. (If we attempt to judge Sagan's mistaken ratios, we must superpose lines with slope 1.) This visual operation is simply too hard.

Figure 1.6 can be greatly improved, at least for the purpose of showing the intelligence measure, by graphing the measure directly on a log scale, as is done in the dot plot of Figure 1.7. Now we can see strikingly many things not so apparent from Figure 1.6. Happily, modern man is at the top. Dolphins are next; interestingly, they are ahead of our ancestor Homo *habilis*.

The problems with Figure 1.6 do not stop here. Five of the labels are wrong. The following are the corrections: "saurornithoid" should be "wolf," "wolf" should be "saurornithoid,""hummingbird" should be "goldfish," "goldfish should be "mole," and "mole" should be "hummingbird." The correct labels yield the satisfying result that a hummingbird is smaller than a mole.

It should be emphasized that for some purposes, a corrected version of Figure 1.6 is a useful graph. For example, it shows the values of the brain and body masses and gives us information about their relationship. The point is that it does a poor job of showing the intelligence measure.



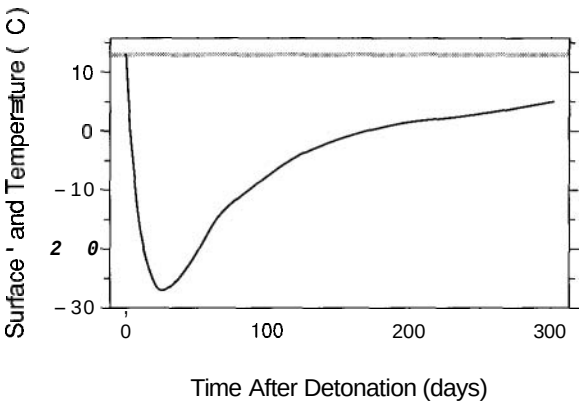
1.7 DOT PLOT. The intelligence measure is shown directly by a dot plot. (Both masses are expressed in grams for this computation.) The values of the measure can be judged far more readily than in Figure 1.6. For example, we can see modern man is at the top, even ahead of our very clever fellow mammals, the dolphins. Incorrect labels on Figure 1.6 have been corrected here.

1.3 The Contents of the Book

Chapter 2: Principles of Graph Construction

Figure 1.8 graphs an estimate of average temperature in the Northern Hemisphere following a nuclear war involving 10,000 megatons of nuclear weapons. The data are from a *Science* article, "Nuclear Winter: Global Consequences of Multiple Nuclear Explosions," by Turco, Toon, Ackerman, Pollack, and Sagan [125]. The temperatures are computed from a series of physical models that describe a script for the nuclear war, for the creation of particles, for radiation production, and for convection. Figure 1.8 shows that the predicted temperature drops to about −25°C and then slowly increases toward the current average ambient temperature in the Northern Hemisphere, which is shown by the horizontal line on the graph.

In Figure 1.8 there are four scale lines that form a rectangle, the tick marks are outside of the rectangle, the size of the rectangle is set so that no values of the data are graphed on top of it, and there are tick marks on all four sides of the graph. Principles of graph construction such as these are the topic of Chapter 2. The focus is on the basic elements: tick marks, scales, captions, plotting symbols, reference lines, keys, and labels. These details of graph construction are critical controlling factors whose proper use can greatly increase the accuracy of the information that we visually decode from displays of data.

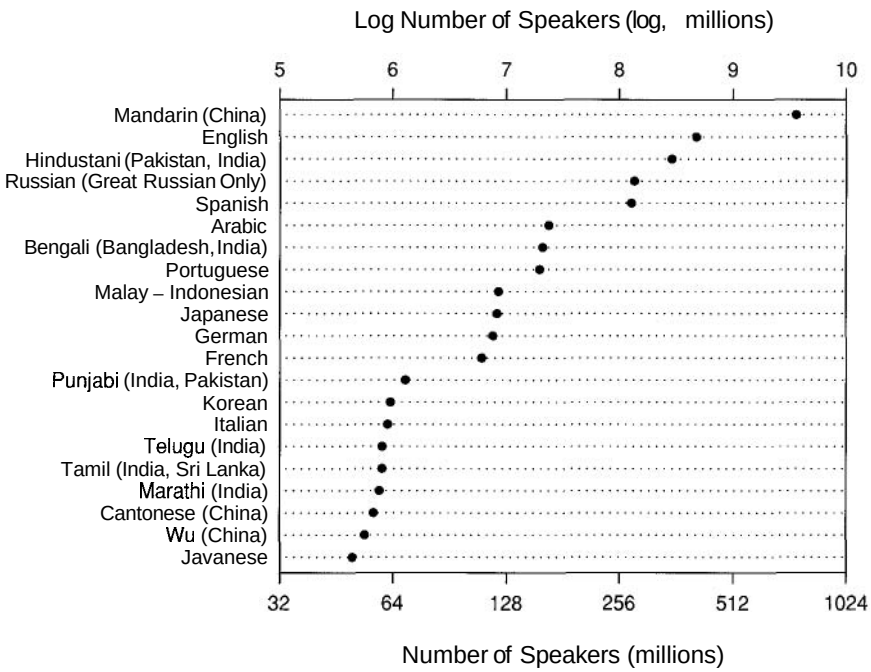


1.8 CHAPTER 2. On this graph there are four scale lines that form a rectangle, the tick marks are outside of the rectangle, the size of the rectangle is set so that no values of the data are graphed on top of it, and there are tick marks on all four sides of the graph. Chapter 2 is about principles of graph construction such as these.

Chapter 3: Graphical Methods

Figure 1.9 is a dot plot, a graphical method that was invented to display measurements with labels [23,26]. The large dots convey the values and the dotted lines enable us to visually connect each value with its label. The dot plot has several different forms depending on the nature of the data and the structure of the labels.

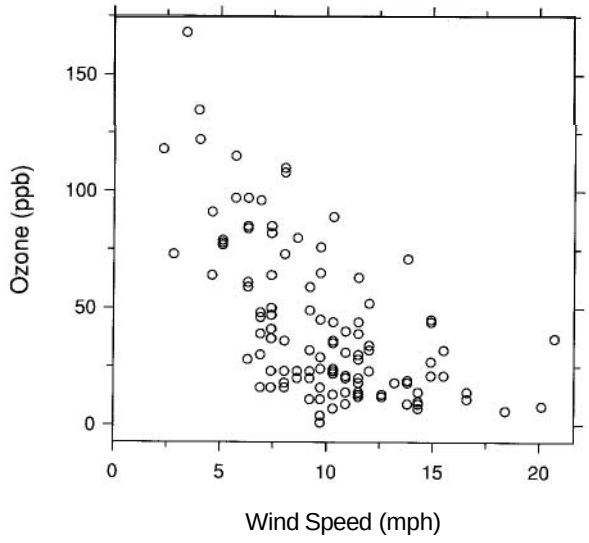
The data in Figure 1.9 are the number of speakers for 21 of the world's languages [98]. Only languages spoken by at least 50 million people are shown. The data are graphed on a log base 2 scale, so moving from left to right, values double from one tick mark to the next.



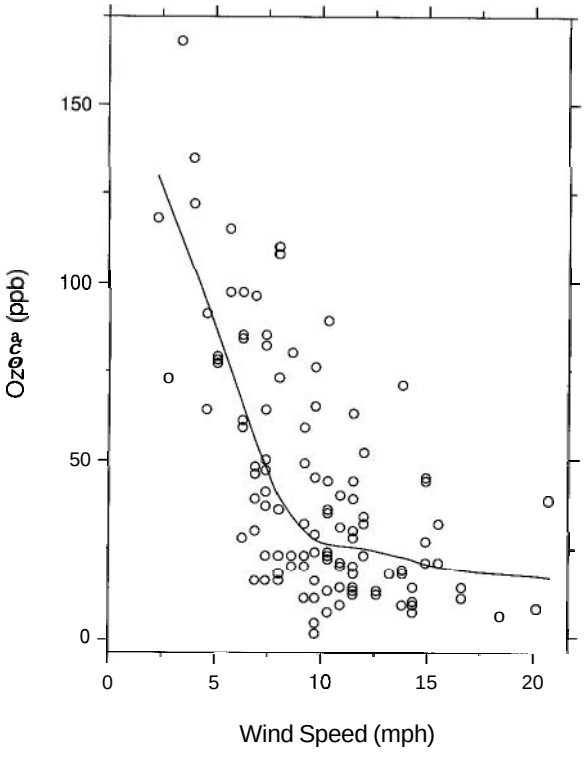
1.9 CHAPTER 3. The figure shows a graphical method called a dot plot, which can be used to show data where each value has a label. The data are the number of speakers for the world's 21 most spoken languages. The data are graphed on a log base 2 scale, so values double in moving left to right from one tick mark to the next.

Figure 1.10 is a graph of ozone against wind speed for 111 days in New York City from May 1 to September 30 of 1973 [13]. The graph shows that ozone tends to decrease as wind speed increases due to the increased ventilation of air pollution that higher wind speeds bring. However, because the pattern is embedded in a lot of noise, it is difficult to see more precise aspects of the pattern, for example, whether there is a linear or nonlinear decrease. In Figure 1.11 a smooth curve has been added to the graph of ozone and wind speed. The curve was computed by a method called *locally* weighted regression, often abbreviated to lowess, or loess [22,26,28]. Loess provides a graphical summary that helps our assessment of the dependence; now we can see that the dependence of ozone on wind speed is nonlinear. One important property of loess is that it is quite flexible and can do a good job of following a very wide variety of patterns.

Chapter *3* is about graphical methods such as the dot plot, loess, and graphing on a log base 2 scale. Some of the graphs are methods by virtue of the design of the visual vehicle used to convey the data; the dot plot is an example. Other methods use the standard Cartesian graph as the visual vehicle, but are methods by virtue of the quantitative information that is shown on the graph; graphing a loess curve is an example of such a method.



1.10 CHAPTER 3. An air pollutant, ozone, is graphed against wind speed. From the graph we can see that ozone tends to decrease as wind speed increases, but judging whether the pattern is linear or nonlinear is difficult.



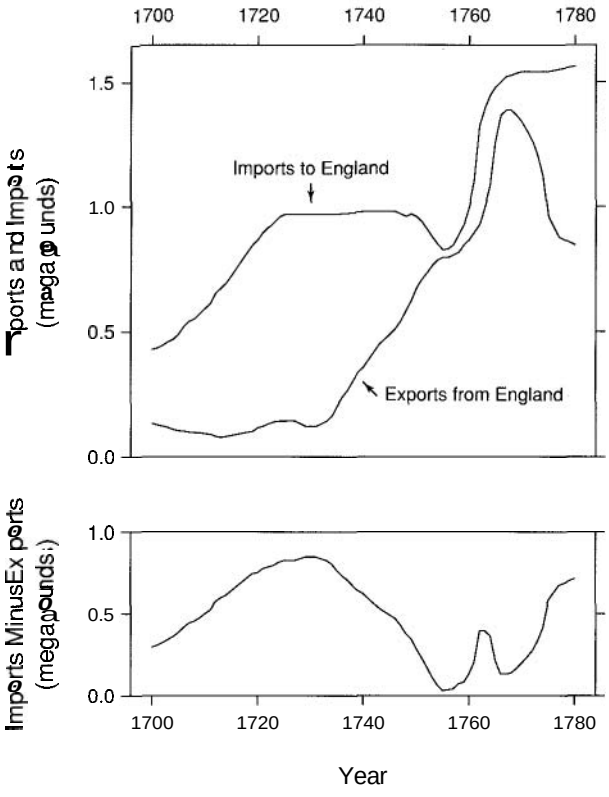
1.11 CHAPTER 3. Loess, a method for smoothing data, is used to compute a curve summarizing the dependence of ozone on wind speed. With the curve superposed, we can now see that the dependence of ozone on wind speed is nonlinear. Chapter 3 is about graphical methods such as loess, dot plots, and graphing on a log base 2 scale.

Chapter 4: Graphical Perception

When a graph is constructed, quantitative and categorical information is encoded, chiefly through position, size, symbols, and color. When we study the graph, the information is visually decoded. A graphical method is successful only if the decoding process is effective. Informed decisions about how to encode data can be achieved only through an understanding of the visual decoding process, which is called graphical perception.

A display method that leads to inefficient visual decoding can prevent important aspects of data from being detected or can lead to distortions in the perception of information. One example was discussed earlier in Section 1.1 (pp. 6–9); the faster rise than fall of the sunspot numbers could not be perceived in the top panel of Figure 1.1.

Figure 1.12 shows another example. The top panel graphs the values of imports and exports between England and the East Indies. The data were first displayed in 1786 by William Playfair [104]. To visually decode the import data we can make judgments of positions along the vertical scale; the same is true of exports. Another important set of quantitative values encoded on this graph is the amounts by which imports exceed exports. To visually decode these values we must judge the vertical distances between the two curves. But we perform this visual operation inaccurately; our visual system tends to judge minimum distances between two curves rather than vertical distances. For example, from the top panel of Figure 1.12 imports minus exports appear not to change by much during the period just after 1760 when both series are rapidly increasing. This is incorrect. Imports minus exports are graphed directly in the bottom panel of Figure 1.12 so that the values can be visually decoded by judgments of position along a common scale, and now we can see there is a rapid rise and fall just after 1760.



1.12 CHAPTER 4. The top panel is a graph of exports and imports between the East Indies and England. The data are from a graph published by William Playfair in 1786. It is difficult to visually decode imports minus exports, which are encoded by the vertical distances between the curves. Imports minus exports are graphed directly in the bottom panel, and now we can see that their behavior just after 1760 is quite different from what we visually decode in the top panel. Chapter 4 deals with issues of graphical perception such as this.

The only route to an understanding of display methods is rigorous study of graphical perception. Chapter 4 is about such rigorous study. First, a model for graphical perception is presented that provides a framework for investigations of graphical perception. Then the model is used to investigate a number of display methods introduced in earlier chapters. This provides both a justification of the methods and guidance for carrying out other investigations. The rigorous study contrasts with the approach of many past discussions of display methods, where, in medieval-science fashion, pure opinion dominates with no facts to provide guidance.