4 Graphical Perception

When a graph is constructed, information is *encoded*. The *visual decoding* of this encoded information is *graphical perception*. The decoding is the vital link, the raison d'être, of the graph. No matter how intelligent the choice of information, no matter how ingenious the encoding of the information, and no matter how technologically impressive the production, a graph is a failure if the visual decoding fails. To have a scientific basis for graphing data, graphical perception must be understood. Informed decisions about how to encode data must be based on knowledge of the visual decoding process.

The only route to an understanding of graphical perception is rigorous study. There is much to draw on to carry this out. First, we can exploit the massive body of knowledge that resides in several scientific disciplines that study vision generally — psychophysics, cognitive psychology, and computational vision [25,75]. Second, we can carry out controlled experiments, systematically varying aspects of display methods, and study the effects on the decoding process. We can observe effects in two ways. One is to simply observe the systematically varied visual displays [8]. Such *simple demonstration* can be sufficient. But in many cases we must run more formal experiments with subjects, viewing protocols, and response measures [25,36,56,75,113]. The experimental technique of vision research is important for such *measured-response experiments*.

Rigorous study has produced a nascent *model* for graphical perception [25]; much of this chapter uses material directly from this source. The model is a paradigm in the sense of Thomas S. Kuhn [79] — a framework that organizes knowledge about a subject. The model sets out definitions, organizes tasks of graphical perception, and hypothesizes that the study of certain visual operations can yield important information about how to design display methods to enhance visual decoding.

The model is presented in Section 4.1 (pp. 223–227). In subsequent sections, the model serves as a framework for studies of display methods. The studies accomplish two goals. They provide the justification for a number of display methods introduced in earlier chapters, and they provide guidance for carrying out other studies.

Section 4.2 (pp. 227–230) studies the display of two or more curves in the same scale-line rectangle. Superposed curves are simple to draw and ubiquitous, so it might be thought that there are no issues, but superposed curves create a major decoding problem.

Section 4.3 (pp. 230–233) is about color encoding: selecting the hues, saturations, and lightnesses of colors that encode data. There are two cases: color encoding a categorical variable, and color encoding a quantitative variable.

Section 4.4 (pp. 234–239) presents the rigorous study that led to texture symbols for encoding different groups of points on scatterplots.

Section 4.5 (pp. 240–243) studies visual reference grids for enhancing the comparison of data on juxtaposed panels. One of the basic laws of psychophysics, Weber's Law, is at the core of this issue.

Section 4.6 (pp. 244–250) studies methods for ordering categories on dot plots and multiway dot plots. These ordering methods can substantially enhance our visual decoding.

Section 4.7 (pp. 251–256) reviews the theory, the mathematics, and the experimentation that led to the discovery of banking to 45°.

Section 4.8 (pp. 256–258) presents an interesting aspect of scatterplots that affects our perception of the correlation between two variables.

Section 4.9 (pp. 259–261) studies encoding by position along a common scale, a display method that leads to efficient visual decoding.

Section 4.10 (pp. 262–269) is about pop charts: visual displays used largely in the mass media and in certain business presentations where the transmission of information is less of an issue, and glitzy display that does not tax the intellect is paramount. The section shows why three pop charts — pie charts, divided bar charts, and area charts — do not result in efficient visual decoding compared with other display methods.

4.1 The Model

Quantitative and Categorical Information

The information inside the data rectangle of a graph can be partitioned into *quantitative values* and *categorical values.* This is illustrated in the multiway dot plot in Figure 4.1. The data are from an agricultural experiment [63]. The response is barley yield and the factors are the growing site (6 categories), the barley variety (10 categories), and the year (2 categories). The yield variable is quantitative; it takes on numerical values. The site, variety, and year variables are categorical; their values are categories. For example, the site variable takes on one of six categorical values — "Waseca", "Crookston", "University Farm", and so forth. A year variable in many applications is quantitative, but in this application it is categorical in that it simply indicates two different growing seasons.

Scale Information and Physical Information

Categorical information and quantitative information each can be described in two ways: *scale* and *physical.* For a quantitative value, scale information is the value in the units of the data. For a categorical value, it is the name of the category. For the observation with the largest yield in Figure 4.1, the scale information is the following: (1) yield = 65 bushels/acre; (2) site = "Waseca"; (3) variety = "No. 462"; (4) year = "1931". The physical information on a graph is a description of the quantitative and categorical information after the removal of the tick labels and the names of categories. We retain the geometry of graphical elements and a breakup of the data into different subsets according to the categories. For the observation with the largest yield, the physical information is the following: (1) yield = 3.9 cm (along the horizontal scale); (2) site = "top panel"; (3) variety = "panel-level two from the bottom"; (4) year = "+".

Pattern Perception and Table Look-Up

The visual decoding of scale information is *table look-up.* For example, we use operations of table look-up to decode the scale

information about the observation with the largest yield — 65 bushels/acre, "No. 462", "Waseca", and "1931". The visual decoding of physical information is pattern perception. For example, in Figure 4.1, we can perceive the following: on each panel except that for Morris, the "+" symbols as a group are shifted to the right with respect to the "○"; on the Morris panel, the "+" symbols are shifted to the left; and the magnitudes of the shifts vary with the site.

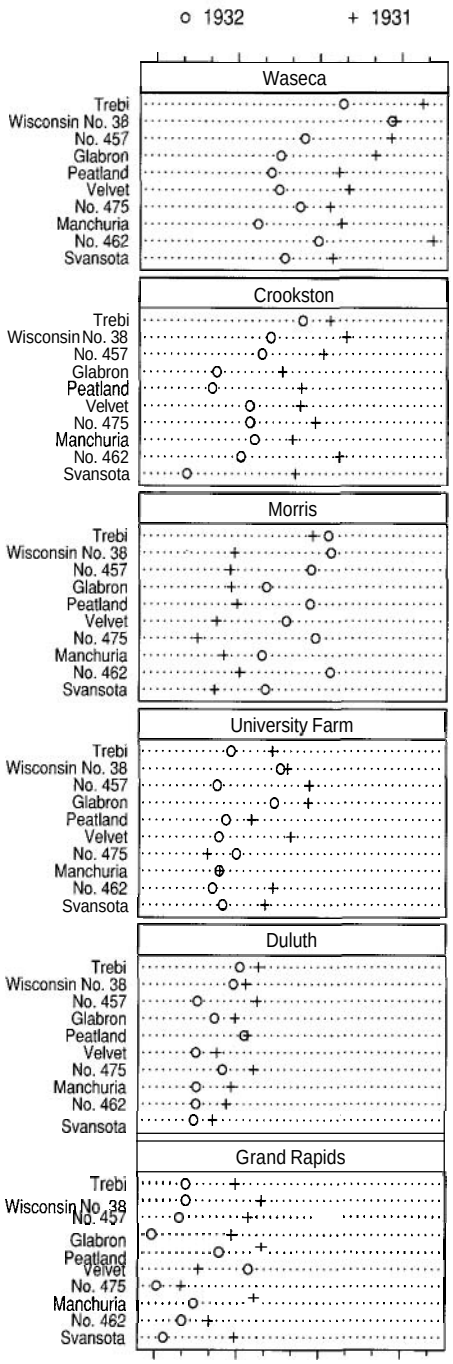
It is crucial to distinguish between pattern perception and table look-up. The output is different, and the visual processing that is necessary to produce the output is different. Table look-up tends to be focused on individual values of quantitative and categorical variables, one after another, with no visual connection of successive look-ups. In pattern perception, we detect and assemble geometric objects to see patterns. The visual processes of pattern perception can vary from slower sequential operations where we compare physical magnitudes by focused attention, to exceedingly fast processes that appear to operate in parallel to produce objects, or gestalts [120].

Three Operations of Pattern Perception

The three visual operations of pattern perception in the model are detection, assembly, and estimation. The efficiency of a visual operation is the speed and accuracy with which it is carried out. Fast operations lead to a perceptually salient formation of patterns. Accurate operations result in valid conclusions about the structure of the data. Often, the two are closely associated; fast operations tend to be accurate.

Detection is the visual recognition of a geometric aspect that encodes a physical value. In Figure 4.1, each plotting symbol encodes physical values of yield, variety, site, and year. Recognition of a symbol is a detection operation. On each level of the figure, the horizontal line segment between the "+" and the "○" encodes the magnitude of the difference in yield between 1931 and 1932 for one combination of variety and site. Recognition of the segment is a detection operation.

Assembly is the visual grouping of detected graphical elements. On each panel of Figure 4.1, we can view the "○" symbols as a whole, mentally filtering out the other elements. This grouping of the "○" symbols is an assembly operation.



4.1 A MODEL FOR GRAPHICAL PERCEPTION. The model for graphical perception in this chapter provides a framework for studies of display methods. The model divides visual operations of graphical perception into pattern perception and table look-up.

Estimation is the visual assessment of the relative magnitudes of two or more quantitative physical values, a and b. There are three progressive levels of this operation: *discrimination, ranking,* and *ratioing.* Discrimination is a judgment of whether $a = b$ or $a \neq b$. Ranking is a judgment of whether $a > b$, $a < b$, or $a = b$. Ratioing is a judgment of the value of a/b. As we move from discrimination to ranking to ratioing there is an increase in the amount of information derived from the judgment. In Figure 4.1, we estimate the lengths of the horizontal line segments connecting the "○" symbols to the "+" symbols to visually decode the magnitudes of the 1931 and 1932 yield differences. This is an estimation operation.

Three Operations of Table Look-Up

The three operations of table look-up in the model are *scanning, interpolation,* and *matching.* As with pattern perception, the efficiency is determined by speed and accuracy.

Consider the largest yield in Figure 4.1. To look up the scale value of yield we scan and interpolate: (1) scan perpendicularly up or down to fix a point along a horizontal scale line; (2) interpolate by estimating the distance from the point to the tick mark to the left or to the right as a fraction of the distance between tick marks. Then another process takes over; we read the tick mark labels and use the physical interpolation to convert to an interpolation in data units. For most people, the decoded scale value would be about 65 bushels/acre. To look up the scale value for variety we scan to the left to get "No. 462". To look up the scale value for site, we scan to the top of the panel to get "Waseca". To look up the scale value for year, we scan to the key and match the "+" encoding the value with the "+" in the key.

Using the Model to Study Display Methods

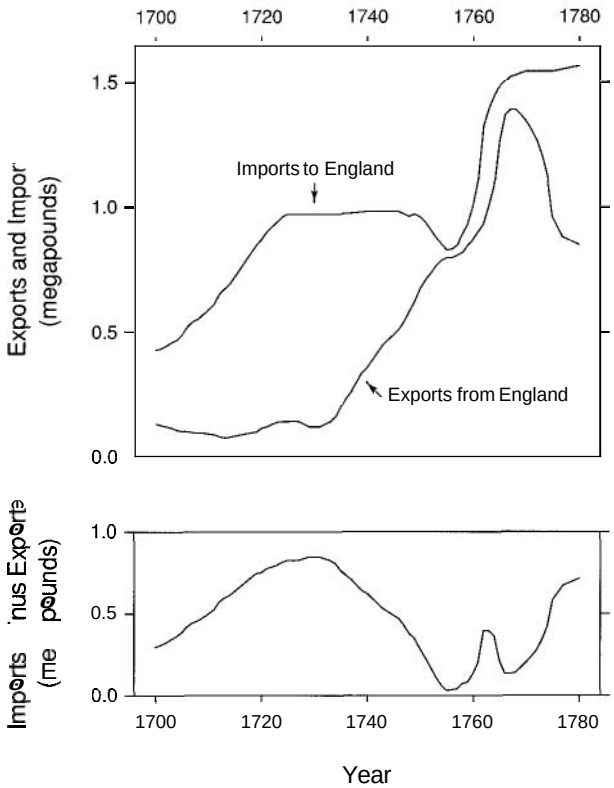
The model provides a collection of visual operations that are to be considered in the study of a display method. It does not prescribe which of its visual operations are important in any particular case, nor does it provide information about the efficiency of the important visual operations once they have been determined. Determinations of

importance and efficiency must come from basic reasoning, the theory of visual perception, simple demonstration, and measured-response experiments. But the model does play a vital role in specifying a collection of operations that are central to graphical perception. Studying these operations can yield important information for judging the performance of display methods. The examples in the next sections demonstrate this.

4.2 Superposed Curves

Figure 4.2, also shown in Section 1.3 (pp. 16–21), graphs data published by William Playfair in 1786 [108]. The top panel displays the values of imports and exports between England and the East Indies. To visually decode the import data we judge the import curve. Each point along the curve encodes imports at a specific point in time. The detection and assembly of the points is exceedingly efficient, and the result is a gestalt, a visual whole that appears as a single object on the graph. The same statements hold for the export curve.

Also encoded by this graph are the amounts by which imports exceed exports. These quantities are encoded by vertical line segments that connect the two curves. Thus to visually decode differences between the curves we must detect the segments, assemble them, and then estimate their lengths. In this section we study the visual operations of pattern perception and table look-up for our decoding of the differences of curves.



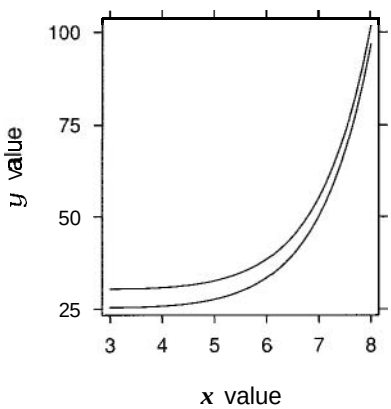
4.2 DIFFERENCES OF CURVES. This display does not convey an accurate visual decoding of the differences of the two curves in the top panel because our detection of the vertical line segments between the two curves is inefficient. If two curves are superposed and judging differences is important, then it is necessary to also graph the differences as in the bottom panel of this display.

Pattern Perception

Surprisingly, pattern perception for differences of curves can be exceedingly inaccurate. It is surprising because the superposing of two or more curves is such a ubiquitous graphical method and because the formation of the individual curve gestalts is so efficient. But simple demonstration shows clearly the inaccuracy.

Figure 4.2 provides a simple demonstration. During the period just after 1760 when both curves are rapidly increasing, the visual impression is that imports minus exports is not large and does not change by much. This is not the case. In the bottom panel of Figure 4.2, imports minus exports are graphed directly. Now, the differences form a curve, and the result is a far more accurate decoding of the information about imports minus exports. The behavior just after 1760 is quite different from how it appears in the top panel; there is a rapid rise to a peak and then a decrease.

Figure 4.3 provides another demonstration with made-up data. The visual impression is that the differences decrease from left to right, but the differences are constant.



4.3 DIFFERENCES OF CURVES. The differences of the curves appear to decrease from left to right but they are constant.

The problem in judging the difference of two curves is not inaccurate estimation, but rather inaccurate detection. When the slopes of two superposed curves are steep, it is exceedingly difficult for our visual system to focus on the vertical distances between the curves. Our visual system detects minimum distances. For example in Figure 4.3 the minimum distances lie along perpendiculars to the tangents of the curves. As the slope increases, the distance along the perpendicular decreases, so the curves look closer as the slope increases. Our visual system does an accurate job of detecting, assembling and estimating the minimum distance segments; the problem is that they are the wrong segments and we cannot force our visual system to process the right segments without using slow sequential search.

Table Look-Up

Figure 4.2 also shows that table look-up for differences of superposed curves is much less efficient than for a direct graphing of the differences. Consider the decoding of the scale value of the difference for 1740. In the bottom panel a single horizontal scan to a vertical scale line together with an interpolation result in a value of about 0.7 megapounds. In the top panel of Figure 4.2 we must either perform two such scans and interpolations or match the vertical distance between the curves at 1740 with a vertical distance along a vertical scale line.

Remedies

Playfair graphed the import-export data using the method of the top panel of Figure 4.2, yet much of his discussion focused on the balance of payments, the differences of the curves, so his display method was not adequate to the task.

One display method for curves is to superpose them on one panel and graph differences on another. This is an attractive method if there are just two curves; Figure 4.2 is one example. The remedy provides both improved pattern perception and improved table look-up.

Another remedy is to juxtapose curves on separate panels. While pattern perception of differences is not as efficient as a direct graphing of differences, and table look-up is not enhanced, the juxtaposition does eliminate the distortion of pattern perception caused by superposition. This remedy is attractive when there are many curves as, for example, on the conditioning plots of functions in Section 3.11 (pp. 203–205). And when this juxtaposition is used, visual reference grids can enhance our pattern perception; the reason for this is given in Section 4.5 (pp. 240–243).

4.3 Color Encoding

There are two uses of color that genuinely enhance the visual decoding of information from graphs. One is encoding different categories of graphical elements in different colors. Figure I at the beginning of the book is one example. Another use is the color level plot: a display of a function of two variables, $z = g(x, \mathrm{y})$, where x and y are encoded by position in the plane and z is encoded by color at the position. Figure II at the beginning of the book is an example. These two uses of color were discussed in Section 3.13 (pp. 209–212). Here, we discuss in detail the color encodings.

The two uses of color are distinct and require different methods. Coloring different groups of graphical elements is an encoding of a categorical variable. For example, in Figure I the categorical variable is the animal group. The goal is an encoding that makes the visual operation of assembly as efficient as possible. For a color level plot a

quantitative variable is encoded. For example, in Figure II the encoded variable is soil resistivity. The goal is an encoding that makes the visual operation of estimation as efficient as possible.

Color Specification by HSL

The processing of light by our visual system is an amazing feat. Light with a single color is a mixture of energies at different wavelengths in the visible spectrum ranging from about 380 nanometers to 770 nanometers. The variation in the amounts of radiation at the different wavelengths accounts for our different color perceptions.

The physical characterization of the light requires the specification of the amounts of radiation at all wavelengths in the visible spectrum. But our perception of color can be described accurately by just three numbers that are derivable from the radiation amounts [50,119]. If two different light sources with different mixtures of wavelengths have the same three numbers, we judge their colors to be the same.

One three-value system is the HSL system: hue, saturation and lightness [50,119]. Hue is what we typically mean in everyday language when we refer to color; hue is described by terms such as green, blue, yellow, magenta, and so forth. Hue is measured in degrees from 0° to 360° since there is a circularity to our perception of hue. (0° and 360" describe the same hue.) The following are the hues of six colors that are equally spaced around the color circle: red = 0°, yellow = 60°, green = 120°, cyan = 180°, blue = 240°, magenta = 300°. In Figure I there are four hues: cyan, magenta, orange (30°), and green. In Figure II there are two hues: cyan and magenta. Lightness refers to how light or dark a color appears. Saturation refers to how pale or deep a color appears. For example, a deep red is highly saturated but a pink is desaturated. If the hue and lightness of a color are kept constant, and the saturation is decreased, the color becomes paler until, when the saturation is zero, the color becomes a gray. For the magentas in Figure II the saturation increases and the lightness decreases as resistivity decreases from the middle of the scale to the smallest values. For the cyans, the saturation increases and the lightness decreases as resistivity increases from the middle of the scale to the largest values.

Assembly and Ranking

There are three properties of color perception that provide principles for the two color encoding tasks: encoding a categorical variable and encoding a quantitative variable.

First, changing hue can provide efficient discrimination of colors. This is strongly supported by both theory and experiment [83,120]. Thus we can use changing hue to provide efficient visual assembly when we encode a categorical variable. In Figure I, assembly of the different groups of points is efficient because the different hues are discriminable. Shortly, we will specify five colors that provide good assembly; four of them are used in Figure II.

Second, if hue is held fixed we can perceive an ordering as either lightness changes or saturation changes. Thus we can use changing lightness and saturation to provide efficient ranking when we encode a quantitative variable. Consider the magentas in Figure II. As resistivity increases from the smallest values to the middle values, the increase in lightness and the decrease in saturation act in concert to provide a strong sense of order. A similar statement holds for the cyans. Shortly, we will give a precise specification of the colors used in Figure II.

Third, we cannot effortlessly perceive an ordering to changing hue. Thus an encoding of a quantitative variable that incorporates changes through many hues does not result in efficient ranking. As Travis [119] puts it, "It is erroneous to assume that we have some hard-wired intuitions for a spectral sequence (i.e. red, orange, yellow, green, blue, indigo, violet)." And as Tufte puts it [122], "the mind's eye does not readily give an order to ROYGBIV." In Figure II, there are two hues, which might seem to ignore this principle of hue perception. But two hues are used to achieve another perceptual goal: clearly perceived boundaries between adjacent levels of the color encoding. Two hues provide somewhat more color variation than just one. True, we do now need to remember that values encoded by cyan are greater than those encoded by magenta, but this does not unduly burden the ranking operation.

CMYK Specification of Colors for Color Encoding

Another method for specifying color is the CMYK system: cyan, magenta, yellow, and black [50,99]. This is the standard system used by printers to produce color on paper. First, each of the four basis colors can be mixed with white; a 75%–cyan is 75% cyan and 25% white. Colors are described by combinations of mixed basis colors. For example, in Figure I the orange mixes a 50%–magenta and a 100%–yellow in equal amounts. We will use the CMYK system to specify a set of colors for encoding a categorical variable and another set for encoding a quantitative variable; we use CMYK because it leads to a simple description. But there are transformations that can translate CMYK to HSL values or to RGB, the red, green, and blue system used for computer screens [50,99].

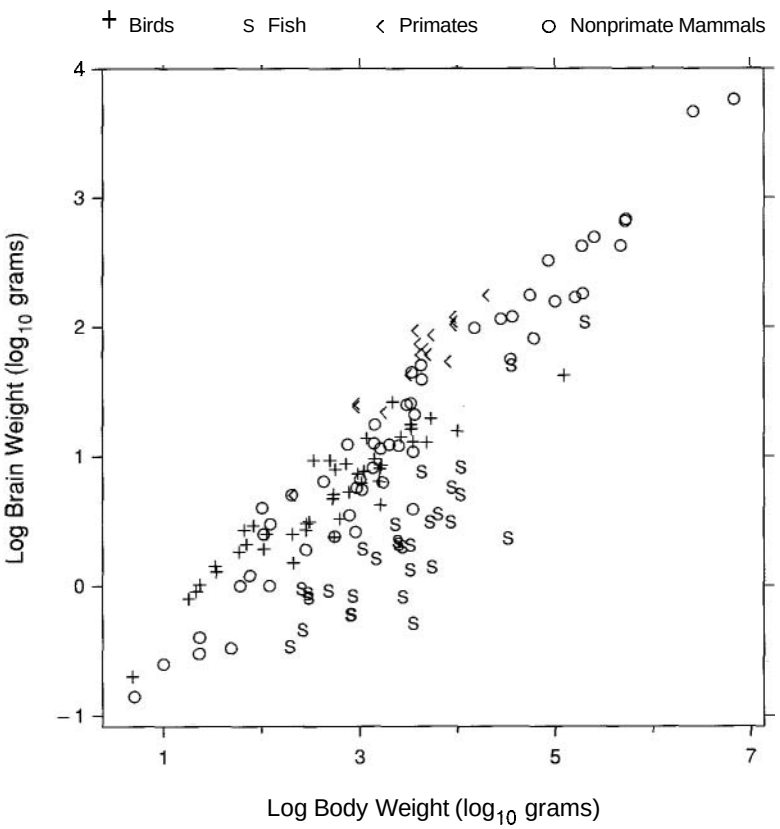
Five colors that provide good visual assembly of the categories of a categorical variable are 100%–cyan, 100%–magenta, 100%–yellow mixed equally with 100%–cyan (green), 50%–magenta mixed equally with 100%–yellow (orange), and 100%–cyan mixed equally with 50%–magenta (light blue). The first four of these are used in Figure I. Actually, a better statement is that the first four are attempted in Figure I since the achieved colors of hardcopy output inevitably vary somewhat from what is specified.

The color encoding in Figure II provides efficient ranking of a quantitative variable. From the middle to the extremes, the cyan ranges from 20%–cyan to 100%–cyan in steps of 20%–cyan, and the magenta ranges from 20%–magenta to 100%–magenta in steps of 20%–magenta. This method provides efficient ranking because it allows accurate ordering and it allows a sufficient number of distinct colors. But there is a bound on the number of colors that can be used if the distinction is to be maintained. Because of the delicacy of color reproduction, only 10 have been used in Figure II. At a computer screen it is possible to drive the number up to 15 or so, but using a significantly greater number typically results in a perceptual merging of some of the adjacent colors. Actually, more than 10 colors are used in Figure II, but the additional ones occur just at boundaries of level regions and are only barely perceptible. This *anti-aliasing* uses standard methods to give the boundaries a smooth look [26].

4.4 Texture Symbols

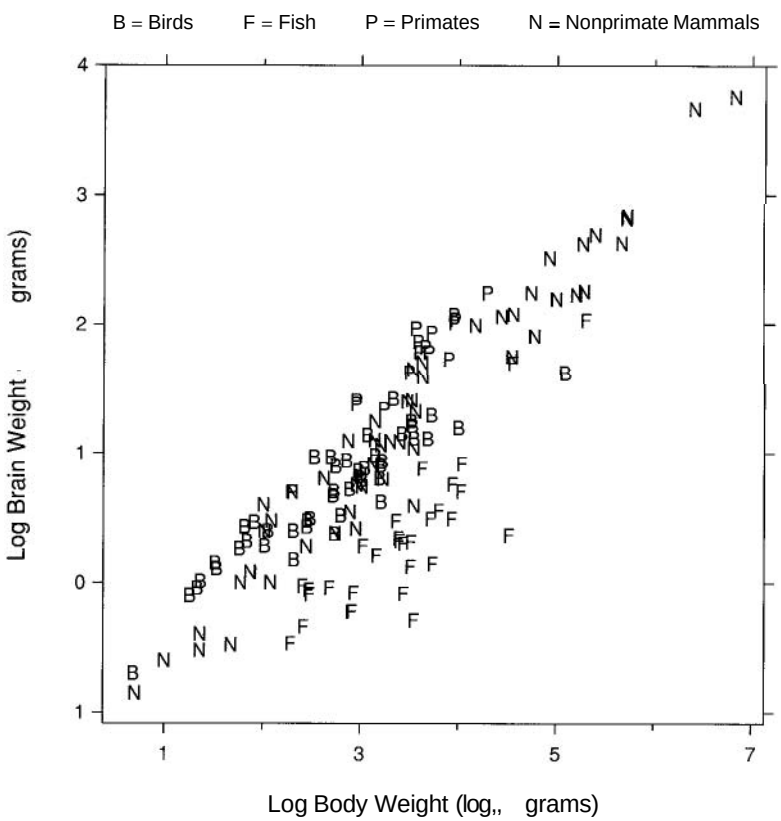
Figure 4.4 graphs brain weights against body weights for four groups of animals. The animal group, a categorical variable, is encoded by different symbol types. The selection of the plotting symbols for such an encoding on a scatterplot substantially affects our ability to detect and assemble each group of points.

If color is available, the color encoding method for categorical variables given in Section 4.3 (pp. 230–233) provides highly efficient detection and assembly that typically performs better than different symbol types in black and white. But the assumption in this section is that color is not available.



4.4 TEXTURE SYMBOLS. Animal group, which is a categorical variable, is encoded by texture symbols, which provide efficient detection and assembly of the four groups of points.

The symbols in Figure 4.4 are from the texture symbol set introduced in Section 3.5 (pp. 154–165). Compared with many other symbol sets, the texture set provides highly efficient detection and assembly. In Figure 4.5 another encoding set is used — the first letters of the group names — and the assembly is poor. In this section we will present the scientific investigations that led to the choice of the texture symbols.



4.5 LETTERS. The plotting symbols are the first letters of the group names. Detection and assembly are less efficient than for the texture symbols.

Detection

Suppose that symbols overlap as in Figure 4.4. To enhance detection we will choose from symbols that are made up of curves and line segments. Such symbols enhance the visual operation of detection when the symbols overlap. Filled symbols such as filled circles tend to form uninterpretable blobs in the presence of overlap, so detection is degraded.

Assembly

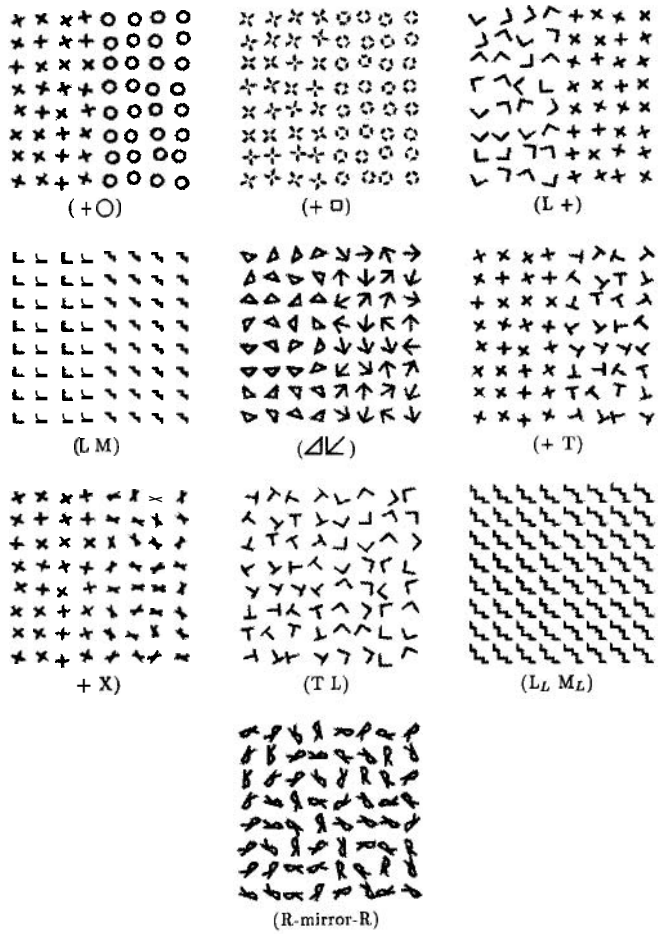
Imagine a room with a plaid rug and a polka-dot wall. The wall and the rug have distinct *textures:* micropatterns with much local variation that nevertheless take on a uniform appearance to our visual system [69]. Where the wall and rug meet, the visual system perceives a clearly defined boundary between the texture patterns. But different micropatterns with very different local features can appear much like one another or even the same.

Each panel of Figure 4.6 has two micropatterns, left and right, with different local features. There is substantial variation in the perceptual salience of the boundaries between the patterns. There is a strong boundary for "+" and "○", a weaker one for "T" and "L", and none for " R and "mirror-R.

What properties of the local features of micropatterns result in strong boundaries and what properties are ignored? This question has been intensively studied by vision researchers because texture perception is one of the important processes of our visual algorithms [69,78,88,90,120]. Fortunately for our search for plotting symbols on graphs with overlap, many of these studies investigated micropatterns formed by symbols consisting of lines and curves, exactly the domain of symbols that are candidates for graphs. Two symbols that provide strong boundaries on micropatterns such as those in Figure 4.6, also provide efficient assembly on scatterplots, so we can use these studies in texture perception to select a symbol set.

Krose [77,78] ran experiments in which subjects determined the presence or absence of a single occurrence of a symbol of a certain type drawn among a texture pattern of symbols of another type. They saw each picture for 80 milliseconds. The percentage of correct answers, corrected for guessing, measures the degree of texture discrimination. The second column of Table 4.1 shows percentages from the Krose experiments for seven of the symbol pairs in Figure 4.6.

A second method of study of texture perception is the image processing of texture patterns using algorithms of computational vision that attempt to reproduce the algorithms of the human visual system [90]. Malik and Perona used methods of computational vision to study the 10 pairs of micropatterns in Figure 4.6 [88]. They applied a texture algorithm to each pair to produce a texture image. To determine the degree of texture discriminability, they computed a measure from each image of how distinct the vertical boundary is between the two regions of symbols. This measure is shown in the third column of Table 4.1 for the 10 symbol pairs. The order from smallest to largest of the algorithmic measure is nearly the same as that for the experimental.



4.6 TEXTURES. Each panel shows two micropatterns formed by symbols. The perceptual salience of the boundaries between the pairs varies substantially. Studies of texture perception provide information for selecting plotting symbols for graphs.

Table 4.1 TEXTURE DISCRIMINATION.

Texture Pair	Experiment	Algorithm
$+ \circ$	100	407
$+ \square$	88.1	225
$L +$	68.6	203
$L\ M$		165
$\Delta \downarrow$	52.3	159
$+ T$	37.6	120
$+ X$	30.3	104
$T\ L$	30.6	90
$L_L\ M_L$		85
R mirror-R		50

The information from Table 4.1 together with a number of simple demonstrations were the basis for the choice of the five texture symbols given in Section 3.5 (pp. 154–165). The symbols, which have an order, are "○", "+", "<", "s", and "w". If there are two groups, the first two symbols are used. If there are three groups, the first three are used, and so forth. Clearly, Table 4.1 gives strong support for using "○" and "+" for two or more groups; this pair has the highest level of discrimination of the 10 pairs. The symbol "<" is third since it does well against "+" in Table 4.1, and simple demonstration suggests that it does well against "○". The fourth symbol, "s", was picked solely on the basis of simple demonstration. The fifth symbol, "w", is an approximation of the symbol labeled " M in Figure 4.6.

4.5 Visual Reference Grids

Visual reference grids were discussed in Section 3.6 (pp. 166–167) and have been used throughout the book to enhance the visual decoding of displays with juxtaposed panels.

Table Look-Up vs. Pattern Perception

Grids on graphs are, of course, an old idea. Originally, they were drawn to enhance table look-up. This was important because graphs were in part archival: they recorded data for detailed recovery of their values later. Also, grids appeared because they were used as an aid to graph production, which was done by hand. Because of their two purposes, table look-up grids were drawn at simple numbers, the simple numbers used for tick marks today. Because the purpose of visual reference grids is to enhance pattern perception and not table look-up, it is not important that the grid lines be drawn at simple numbers; it is only vital that the same grid be drawn on all panels.

Table look-up grids have largely been abandoned in the display of scientific data because of the ubiquitous use of computers to produce graphs and communicate data. But visual reference grids remain as an important display method.

Weber's Law

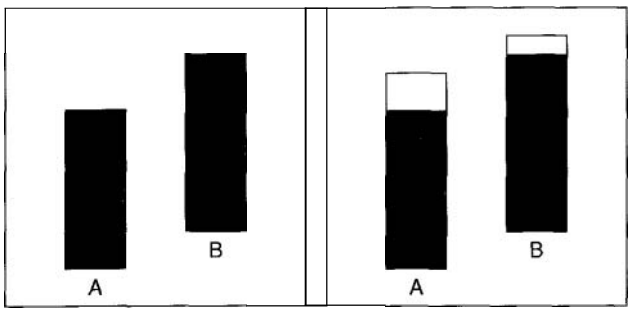
Our study of visual reference grids will make use of *Weber's Law*, formulated by the 19th century psychophysicist E. H. Weber [4]. It is one of the fundamental laws of human perception. Suppose x is the magnitude of a physical attribute; to be specific let it be the length of a line segment. Let $w_p(x)$ be a positive number such that a line of length $x + w_p(x)$ is discriminated with probability p to be longer than the line of length x. Weber's Law states that for fixed p,

$$w_p(x) = k_p x \, ,$$

where k_p does not depend on x. The law appears to describe reality extremely well for many perceptual judgments including length, area, and volume.

One implication of Weber's Law is that we need a fixed percentage increase in line length to achieve detection. For example, it is easy to detect a difference between two lines of length 2 cm and 2.5 cm, because the percentage increase of the second over the first is 25%; however, it is much harder to detect a difference between two lines that are 50 cm and 50.5 cm, even though the difference is also 0.5 cm, because the percentage increase is only 1%.

In the left panel of Figure 4.7 there are two solid rectangles with unequal vertical lengths. It is difficult to determine which length is greater. In the right panel, the two solid rectangles are embedded in two frames that have equal vertical lengths. Now we can accurately discriminate the vertical lengths of the solid bars; the right length is greater.



4.7 WEBER'S LAW. The solid rectangles in the left panel have unequal vertical lengths that are difficult to discriminate. In the right panel the solid rectangles are embedded in frames that have equal vertical lengths and now we can accurately discriminate the lengths. Weber's Law explains why.

Weber's Law explains our differing accuracies of length estimation in Figure 4.7. In the left panel our estimation is not accurate enough to discriminate the length differences because the percentage difference is small. In the right panel the two rectangles formed by the empty spaces at the tops of the frames have vertical lengths whose absolute differences — that is, differences in cm — are the same as those of the solid bars, but whose percentage differences are much greater. Because of Weber's law, we can readily perceive that the vertical length of the left empty rectangle is greater. The vertical lengths of the two frames, which have been drawn to be equal, do indeed appear equal to us. The visual system combines this conclusion of equality of the frames with the conclusion that the vertical length of the left empty rectangle is greater to infer that the vertical length of the right solid rectangle is greater.

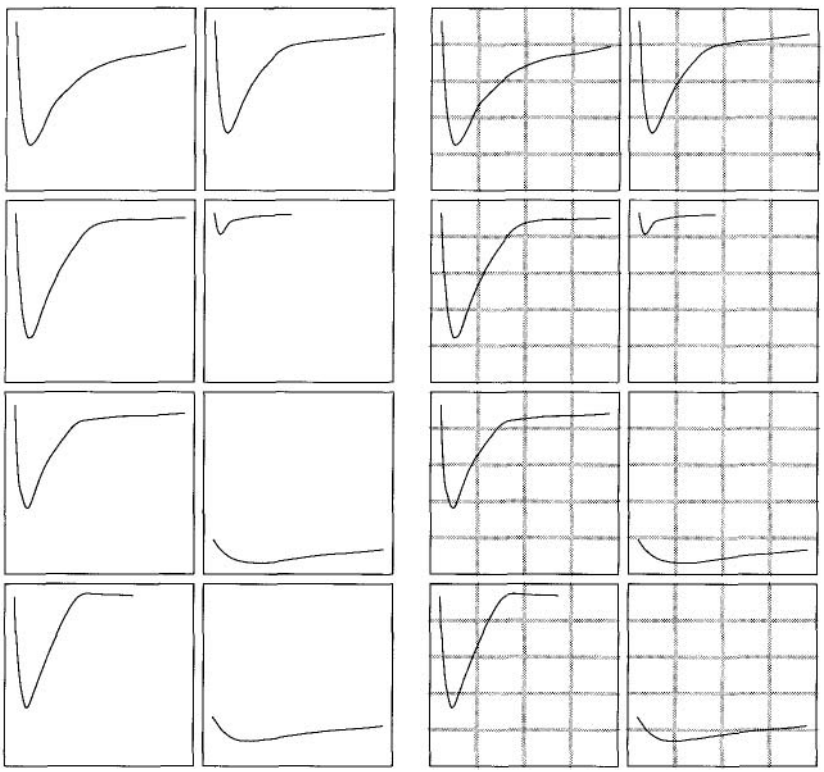
Pattern Perception

In Figure 4.8 eight curves are graphed twice in juxtaposed panels, once in the two columns of panels to the left and once in the two columns to the right.

The operations of detection, assembly, and estimation for decoding information about each curve in Figure 4.8 are exceedingly efficient. The result is a gestalt, a curve object that we see as a single perceptual unit. But the comparison of curves on different panels is an entirely different visual process. Consider the decoding of the minima of the eight curves in the left two columns. There are different ways this can be done. One is to detect and assemble the vertical line segments that extend from the bottom lines of the boxes to the points on the curves where the minima occur, and then carry out estimation of the lengths. The detection and assembly must be carried out by a highly attentive sequential search process with shifts of our eyes from one panel to the next. No gestalt forms as it does for the visual decoding of a single curve because the visual operations are much slower. We cannot readily hold all segments in short term memory, and even our cognitive processing, our conscious conclusion about relative values, must be built up from a sequence of comparisons of subsets of values. Nevertheless, the comparison of minima and other commensurate aspects of curves on separate panels is an informative visual process.

The Application of Weber's Law

Weber's Law explains why visual reference grids enhance pattern perception. The grids allow us to convert estimation of lengths with small percentage differences to estimation with much larger percentage differences which, by Weber's Law, means an increase in accuracy. Figure 4.8 illustrates this. Consider the values of the minima of the five curves with deep troughs. As we just saw, in the left two columns we can compare the minima by judgment of the lengths of the vertical line segments that extend from the bottom lines of the boxes to the points on the curves where the minima occur. But in the right two columns we can compare the minima by estimation of the five line segments extending from the minima to the horizontal grid lines just above them, or by



4.8 VISUAL REFERENCE GRIDS. Eight curves are graphed twice, once in the two columns of panels to the left and once to the right. The right columns have visual reference grids that enhance pattern perception; they enable more efficient comparisons of patterns on different panels. The increased efficiency is explained by Weber's Law.

estimation of the five line segments extending down to the grid lines just below. Each of these two collections of five segments has lengths whose pairwise percentage differences are greater than those of the segments to the bottom lines of the boxes.

One response to this discussion might be that we should superpose as many data sets as possible to reduce the comparison of patterns on different panels. For certain types of graphs — for example, for scatterplots — this can make sense. But as we saw in Section 4.2 (pp. 227–230), superposed curves introduce serious biases in our judgments of differences of curves because of faulty detection.

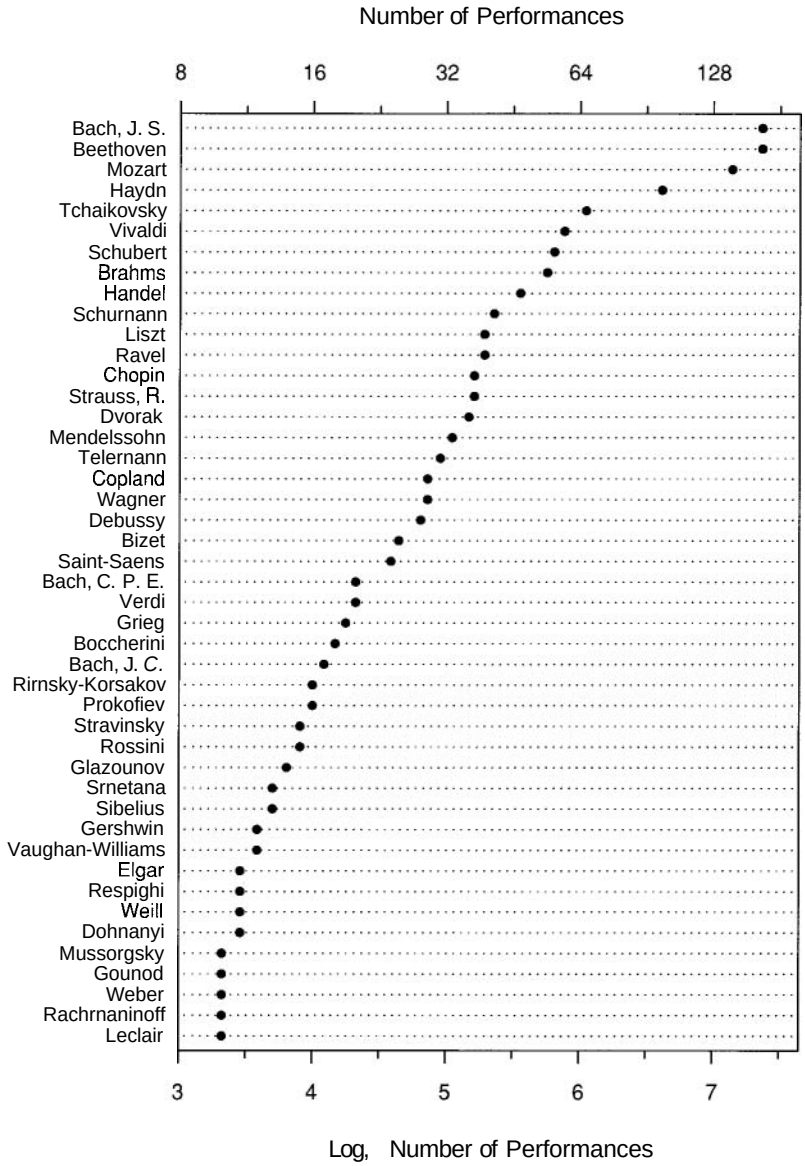
4.6 Order on Dot Plots

Section 3.4 (pp. 150–154) presented dot plots, a graphical method for displaying measurements with labels that are formed from one or more categorical variables. The order of the categories for each categorical variable is an important aspect of the dot plot display method that substantially affects our visual decoding.

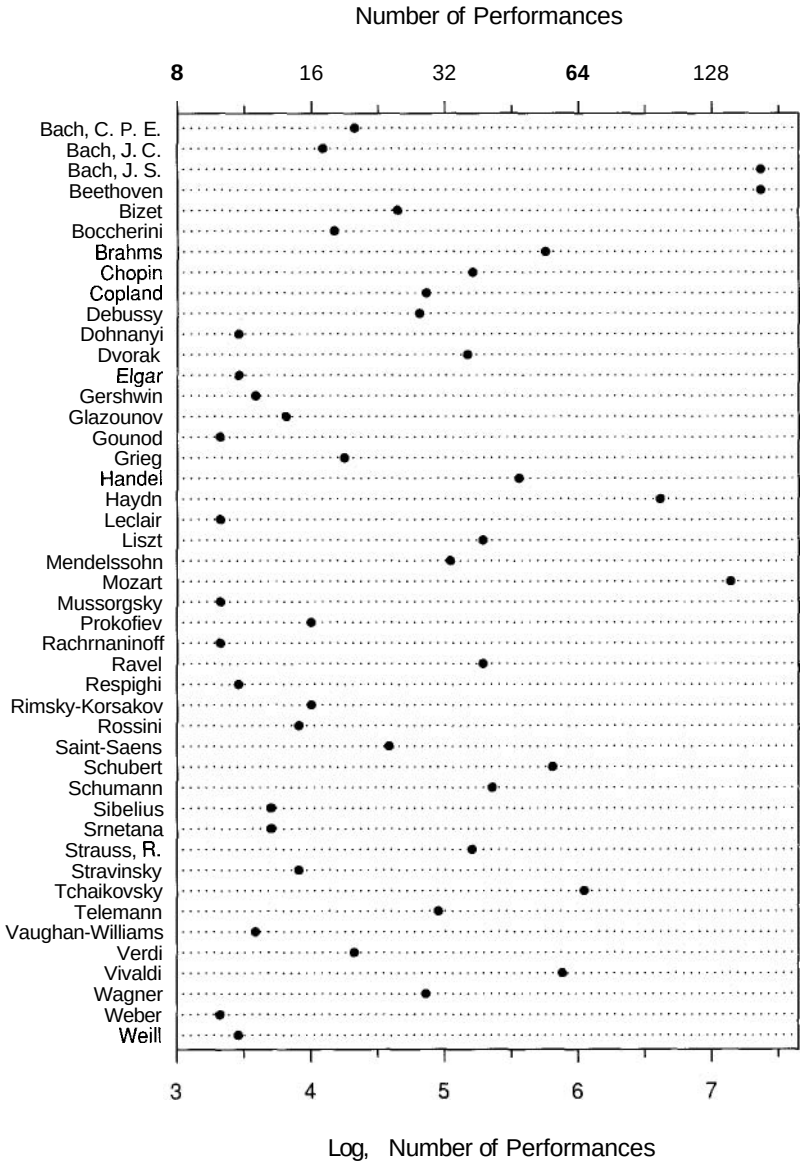
Figure 4.9 is a dot plot of measurements of a classical music station in New York City. The data were collected to resolve a debate about which composer is played the most on the station; one advocate was sure it was Beethoven and another was sure it was Johann Sebastian Bach. To resolve the debate a survey was conducted using the monthly program guide for two months. The number of performances of pieces for each composer aired during this time period was counted. (Actually, the count is only for longer pieces since compositions less than about ten minutes are not included in the section of the guide that was used.) Repeat performances of a composition were counted; for example, Beethoven's Ninth Symphony was performed three times, so Beethoven got a count of three for that. The number of performances for all composers with more than 10 are shown in Figure 4.9. Almost as if to make both advocates happy, Bach and Beethoven tied for first place with 165 performances each.

Figure 4.9 shows some interesting information in addition to who was first. The very most popular composers are way ahead of the rest of the field. Fifth place Tchaikovsky, whom most would think of as quite a popular composer, is beaten by Bach and Beethoven by a factor of about 2.25, and tenth place Schumann is beaten by the winning pair by a factor of about four. Mozart has about 50% more performances than Haydn, even though their compositions, to many ears, are similar. It is surprising to see Aaron Copland with so many performances. As the list shows; few 20th century composers get much exposure in classical music circles, which makes the word "classical" appropriate.

In Figure 4.9 the data are ordered from bottom to top. Figure 4.10 graphs the data again, but arranged this time with composers in alphabetical order. For most purposes, Figure 4.9 is more informative than Figure 4.10. When we study a distribution of values such as the composer performances, we want to know what is large, what is medium, and what is small. The organization in Figure 4.9 allows us to easily assemble and estimate the large values, or the medium values, or the small values. We cannot do this nearly as effectively in Figure 4.10 because each of these sets of values is scattered throughout the graph. For example, in the above discussion of the data, we focused on the values of the top five composers and carefully compared them visually from Figure 4.9. It is easier to assemble the large values because they are spatially grouped by the ordering. And estimation is more accurate because the symbols encoding the values are closer to one another [36].



4.9 ORDER FOR DOT PLOTS. The data on this dot plot are ordered from smallest to largest. This enhances our visual decoding of the distribution of the values along the measurement scale.

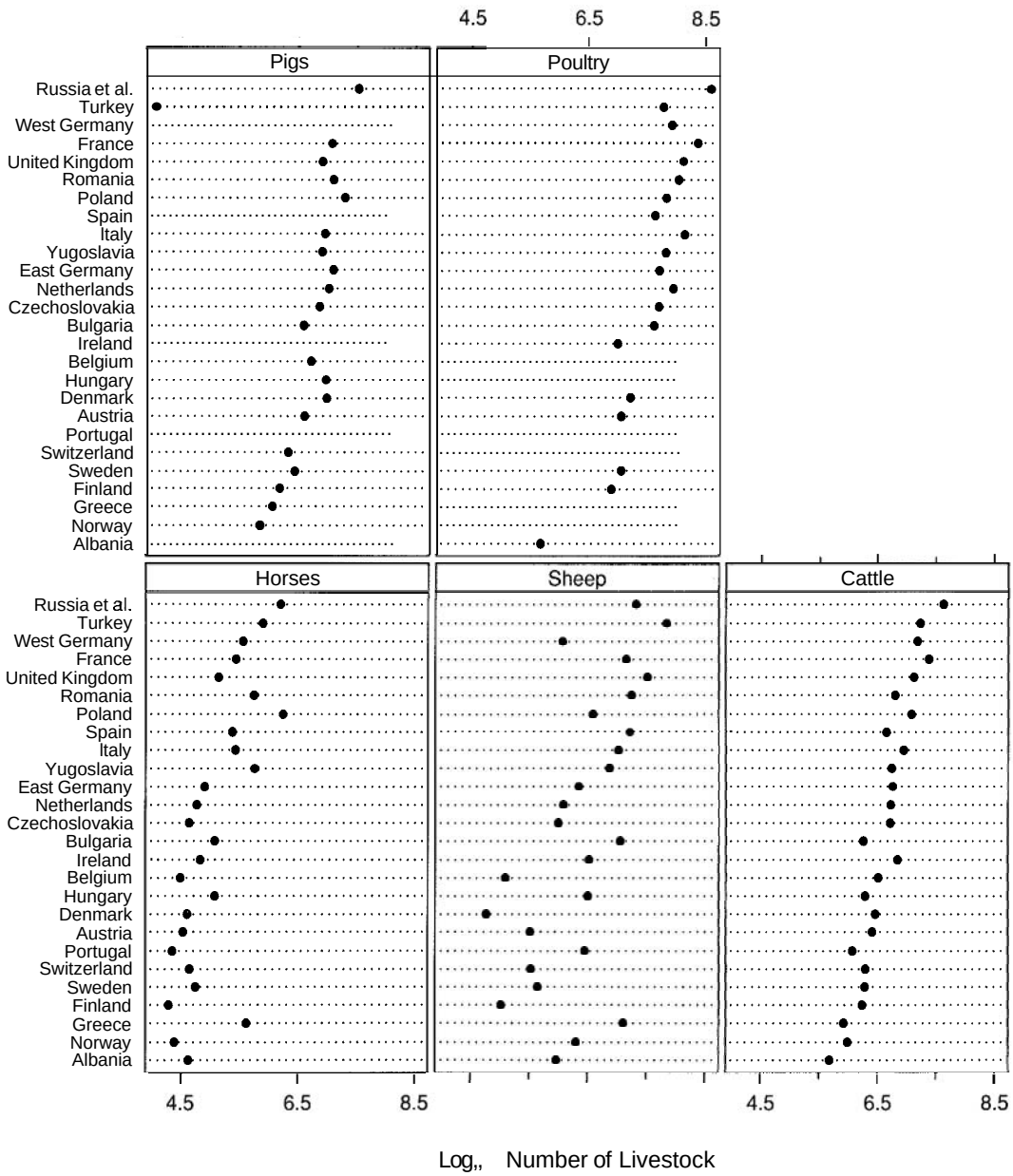


4.10 ORDER FOR DOT PLOTS. The data on this dot plot are ordered alphabetically. This degrades our visual decoding of the distribution of the values along the measurement scale.

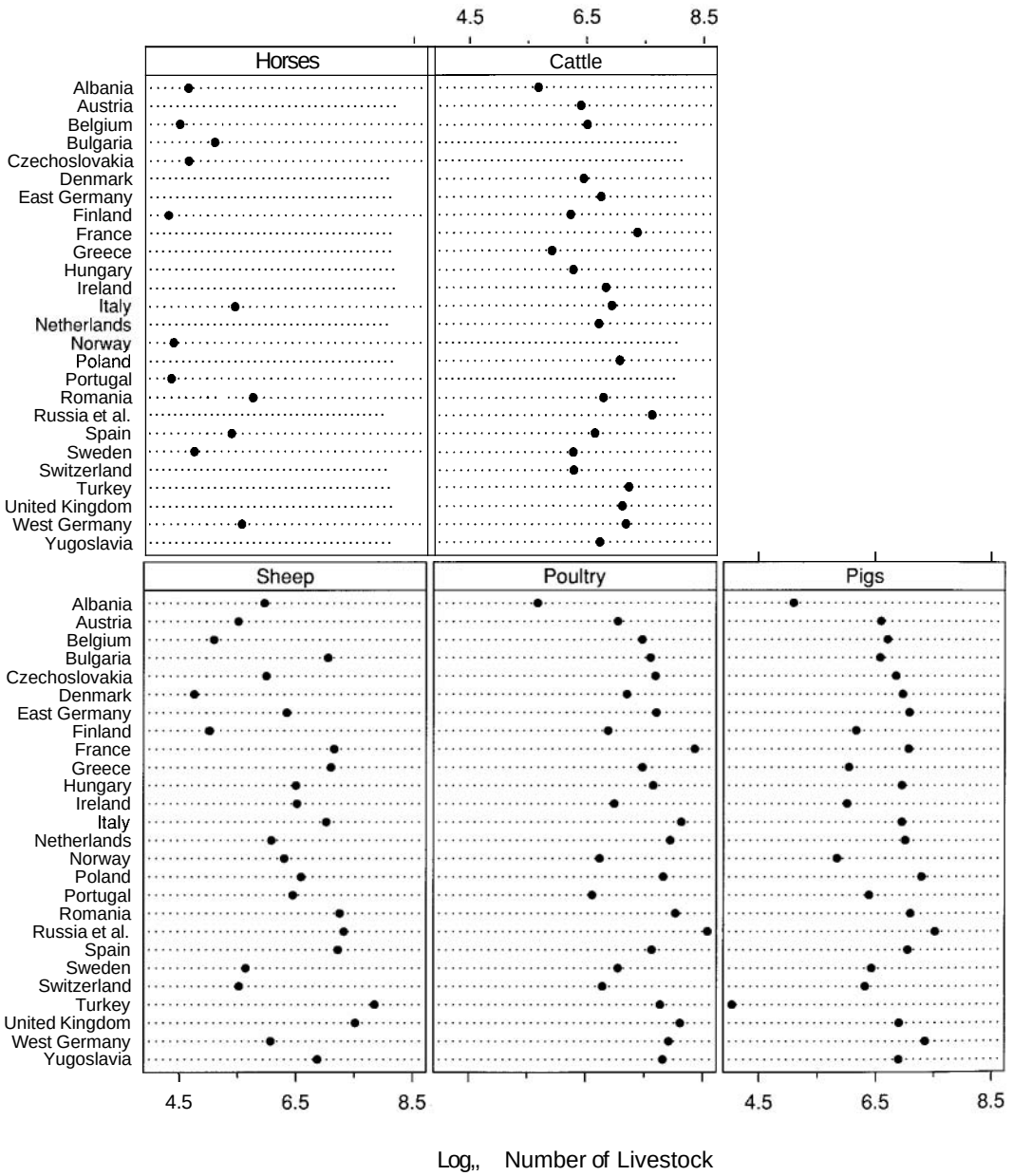
Order is important for multiway dot plots as well. Figure 4.11 graphs the logarithms of livestock counts from a census of farm animals in 26 countries [14]. The livestock variable is encoded by the panels and the country variable is encoded by the levels of each panel. The countries are assigned to the levels so that the country medians increase from bottom to top. The median of the five observations for Albania is the smallest country median. Norway has the next smallest country median, and Russia *et al.* has the largest. The panels are ordered so that the livestock medians increase from left to right and from bottom to top. The median for horses is the smallest, and the median for poultry is the largest.

The category orderings in Figure 4.11 are crucial to the perception of effects. The ordering of the countries by the country medians establishes gestalts on each panel that are easier to compare from one panel to the next. For example, this allows us to see that the sheep data behave differently from the data of the other livestock types. The amount of variation in the log counts for sheep is greater than that of the other livestock types; in other words, the ordering of the sheep data agrees the least well with the ordering of the country medians. Also, the median ordering of the countries provides a benchmark for each log count — the values of the nearby log counts in the same panel. For example, Figure 4.11 shows that the small log cattle count in Albania is not unusually small given the overall rank of Albania, but the log pig count in Ireland is unusually small given the overall rank of Ireland.

In Figure 4.12, the log counts are displayed again with the levels ordered alphabetically and the panels ordered arbitrarily. Many of the effects readily seen in Figure 4.11 are not revealed. The sheep data no longer stand out as particularly unusual, and we cannot see that the log pig count in Ireland is small or that the log cattle count in Albania fits the pattern of the data.



4.11 ORDER FOR MULTIWAY DOT PLOTS. On this multiway dot plot the countries are ordered so that the country medians increase from bottom to top, and the panels are ordered so that the livestock medians increase from left to right and from bottom to top.



4.12 ORDER FOR MULTIWAY DOT PLOTS. On this multiway dot plot the countries are ordered alphabetically. Properties of the data apparent in Figure 4.11 cannot be seen.

4.7 Banking to 45°

Banking to 45°, used extensively in the book, was introduced and discussed in detail in Section 2.4 (pp. 66-79). So was the aspect ratio, which controls banking. Recall that the aspect ratio is the height of the data rectangle divided by its width. This section contains both the studies in graphical perception that led to banking [33,37] and the details of the banking method [26].

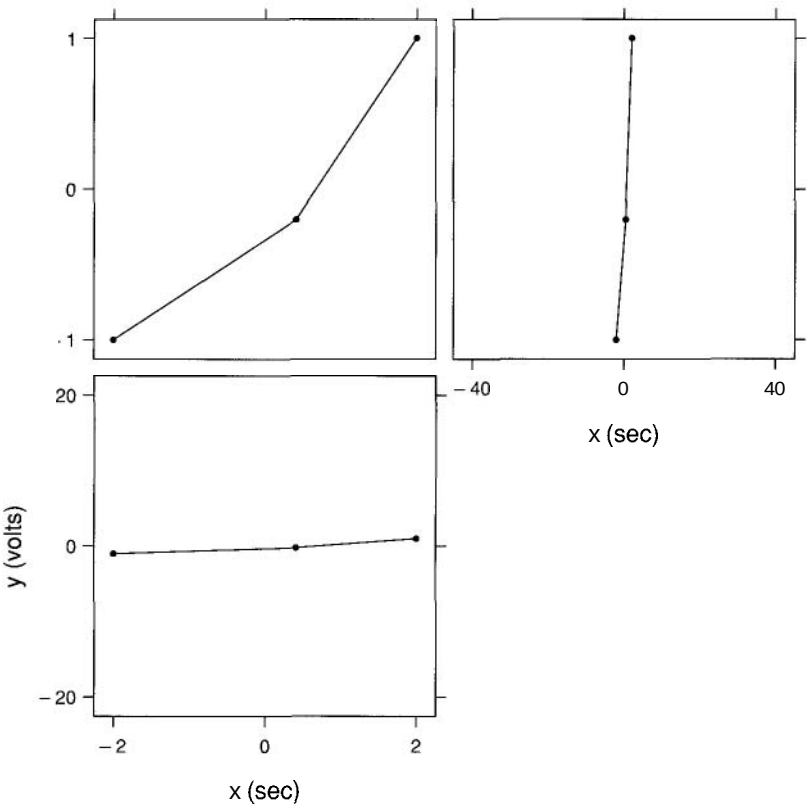
Insight, Hypothesis, and Experiment

The issues involved in banking to 45° are intricate and required extensive investigation to confirm hypotheses. But as with many complicated issues, the essential insight is relatively simple.

Consider Figure 4.13. The top left panel shows two line segments with positive unequal slopes. The scale slopes — that is, the slopes of the segments measured in scale units — are 0.75 volts/sec for the upper right segment and 0.33 volts/sec for the lower left segment. The ratio of the larger slope to the smaller is 2.25. The physical slopes — that is, the slopes measured in physical units such as cm — are 1.5 vcm/hcm and 0.67 vcm/hcm where "vcm" means vertical cm and "hcm" means horizontal cm. The ratio of the physical slopes is the same as the ratio of the scale slopes, 2.25.

The *orientations* of the segments are their angles with the horizontal. If θ is the orientation of a segment and s is the physical slope then $s = \tan(\theta)$. We measure orientation in degrees; segments with positive slopes have positive orientations and segments with negative slopes have negative orientations. A segment with a physical slope of 1 vcm/hcm has an orientation of 45" and a segment with a slope of −1 has an orientation of $-45°$. For the two segments in the top left panel of Figure 4.13, the orientations are 56.3" and 33.7".

The slopes of line segments on a graph encode information about rate of change. For example, the slopes of the local line segments that make up a curve, $y = f(x)$, encode information about the rate of change of y as a function of x. We estimate the orientations of the segments to provide information about the relative rate of change. Thus it is not slope itself we judge but rather a transformation of the slope.



4.13 THE 45° PRINCIPLE. Centering two line segments on 45° maximizes their angular separation, which enhances our judgment of rate of change. In the upper left panel the two segments are centered on 45° in the sense that the average of the two orientations is 45°. In the other panels the aspect ratios are very different from the one that centers the segments on 45°, and the angular separations in both cases are far smaller.

In the right panel of Figure 4.13 the segments are graphed again with a different horizontal scale. The aspect ratio is now very large, the orientations are much steeper, and the difference of the orientations is much smaller. The ratio of the two slopes is the same, so if it were slope that we estimated directly there would be no problem. But we estimate the orientations, and the relationship of the orientations has changed quite substantially. In particular, because the difference of the two orientations is small, it is difficult to discriminate a difference between them; in other words, the slopes appear to be the same or very nearly so. This is a general phenomenon. If the aspect ratio of a display gets too

big, we can no longer discriminate two positive slopes or two negative slopes because the orientations get too close. A similar statement holds when the aspect ratio is too small. This is illustrated in the bottom panel of Figure 4.13.

The important insight in banking to 45° was to ask the following question: since the aspect ratio controls the angular separation of two line segments with positive, unequal slopes, what aspect ratio maximizes the absolute difference of their orientations? The answer is the aspect ratio that makes the arithmetic average of the two orientations equal to 45°. Similarly, the absolute difference of the orientations for two segments with negative, unequal slopes is maximized when their average orientation is −45°.

A simple argument proves this 45° principle. We will treat the case of two positive, unequal slopes. Let $\widehat{a}$ be the value of the aspect ratio that makes the average of the two orientations equal to 45°. When the average is 45° each orientation is 90° minus the other, so if $\widehat{s}$ is one of the physical slopes the other is $1/\widehat{s}$ because of the identity

$$\tan(\theta) = 1/\tan(90^\circ - \theta)$$

Suppose it is $\widehat{s}$ that is the bigger of the two slopes. If we multiply the aspect ratio $\widehat{a}$ by a factor f, the new aspect ratio is $a = f6$, the physical slopes are $f\widehat{s}$ and $\widehat{f}/s$, and the difference of the orientations is

$$d(f) = \arctan(f\widehat{s}) - \arctan(f/\widehat{s}) .$$

The derivative of d at f is

$$\begin{aligned} d'(f) &= \frac{\widehat{s}}{1 + f^2\widehat{s}^2} - \frac{1/\widehat{s}}{1 + f^2/\widehat{s}^2} \\ &= \frac{(1 - f^2)(\widehat{s} - 1/\widehat{s})}{(1 + f^2\widehat{s}^2)(1 + f^2/\widehat{s}^2)} . \end{aligned}$$

$d'(f)$ is positive for $f < 1$, zero for $f = 1$, and negative for $f > 1$. Thus $d(f)$ is a maximum when $f = 1$.

This fact about the absolute difference of orientations led to the 45° hypothesis: the orientations of two line segments with positive slopes are most accurately estimated when the average of the orientations

is 45°, and the orientations of two line segments with negative slopes are most accurately estimated when the average of the orientations is −45°.

An experiment was run to confirm the 45° hypothesis. Subjects where shown pairs of line segments and were asked to estimate what percent the slope of one segment was of the other. The experiment unequivocally confirmed the hypothesis.

The Details of Banking to 45°.

The 45" principle applies to the estimation of the slopes of two line segments. But we seldom have just two segments to judge on a display, and the aspect ratio that centers one pair of segments with positive slopes on 45" will not in general center some other pair of segments with positive slopes on 45°. Banking to 45° is a compromise method that centers the absolute values of the orientations of the entire collection of line segments on 45° to enhance overall the estimation of the rate of change.

Let *v* be the length in physical units of a vertical side of the data rectangle of a graph. Let *9* be the length of the vertical side in scale units. Similarly, let *h* be the length in physical units of a horizontal side of the data rectangle, and let h be the length in scale units.

Consider the length of any interval on the vertical scale of a graph. The value $v/\ddot{v}$ is the conversion factor that takes the length in scale units and changes it to a length in physical units on the graph. Similarly, h/h is the conversion factor for the horizontal scale.

The aspect ratio of the data on the graph is

$$a(h, v) = v/h\ .$$

Suppose the units of the data are fixed so that *9* and h are fixed values. The values of *v* and *h* are under our control in graphing the data, and the aspect ratio is determined by our choices of them.

Consider a collection of n line segments inside the data region. Let $\ddot{v}_i$ be the absolute value of the difference in scale units of the vertical scale values of the two end points of the ith line segment. Let $v_i(v)$ be the

absolute difference in physical units when the vertical length of the data rectangle is v. Define $\ddot{h}_i$ and $h_i(h)$ similarly. Let

$$\bar{v}_i = \ddot{v}_i/\ddot{v}$$

and

$$\bar{h}_i = \ddot{h}_i/\ddot{h}\,.$$

The absolute value of the orientation of the ith segment is

$$\theta_i(h,v) = \arctan\left(\frac{v_i(v)}{h_i(v)}\right) = \arctan\left(a(h,v)\bar{v}_i/\bar{h}_i\right)$$

The physical length of the ith segment is

$$\ell_i(h,v) = \sqrt{h_i^2(h) + v_i^2(v)} = h\sqrt{\bar{h}_i^2 + a^2(h,v)\bar{v}_i^2}\,.$$

One method for banking the n segments to 45° is to choose $a(h,v)$ so that the mean of the absolute orientations weighted by the line segment lengths is 45°. (Strictly speaking this does not adhere to the 45° principle because the centering solution for two segments derived earlier sets the unweighted average to 45°, but experimentation with many data sets suggested that the weighted criterion performed somewhat better for the case of many segments, the common one.) Thus

$$\frac{\sum_{i=1}^n \theta_i(h,v)\ell_i(h,v)}{\sum_{i=1}^n \ell_i(h,v)} = \frac{\sum_{i=1}^n \arctan\left(a(h,v)\bar{v}_i/\bar{h}_i\right)\sqrt{\bar{h}_i^2 + a^2(h,v)\bar{v}_i^2}}{\sum_{i=1}^n \sqrt{\bar{h}_i^2 + a^2(h,v)\bar{v}_i^2}}$$

is equal to 45°. Notice that the right side of this formula depends on v and h only through $a(h,v)$. As is intuitively clear, if we multiply v and h by the same factor, the orientations of the segments do not change. Only their ratio matters. Thus it is the aspect ratio that controls banking.

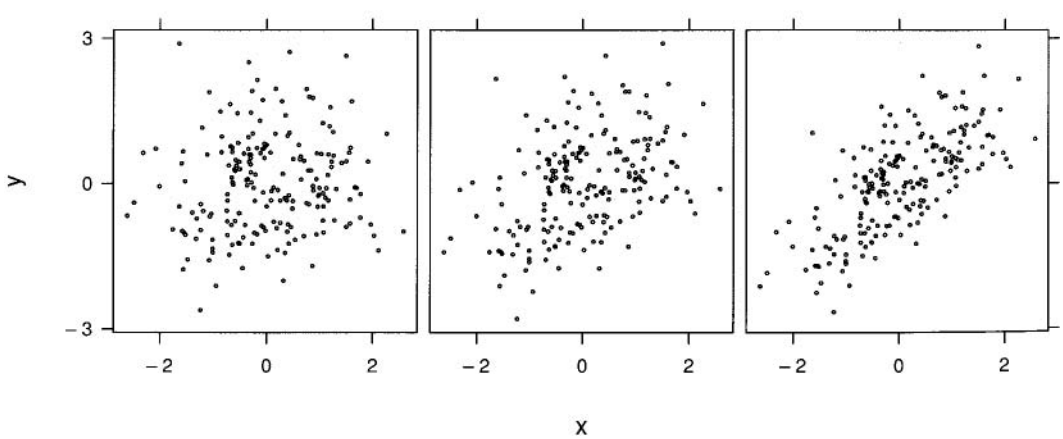
There is no closed-form solution for the aspect ratio that makes the above weighted mean absolute orientation equal to 45°. The value of a needs to be found by iterative approximation. But the approximation can be fast because the weighted mean is a monotone function of a.

Table Look-Up

Table look-up of information about rate of change is exceedingly inefficient. Consider what we must do to decode the scale slope of a single segment in the upper left panel of Figure 4.13. We must scan horizontally from the two endpoints onto the vertical scale line and interpolate to determine two scale values in volts, do mental arithmetic to get a volt difference, scan vertically from the two endpoints onto the horizontal scale and interpolate to determine two scale values in seconds, do mental arithmetic to get a seconds difference, and then do a mental division to get a slope in volts/second. Clearly this is too complicated to be useful. If a visualization requires greater efficiency for table look-up of slope, then we must derive and graph slope directly.

4.8 Correlation

Scatterplots are widely used because they provide incisive study of the properties of bivariate data. One property that can be assessed is the amount of correlation — that is, how closely paired sets of measurements are associated. For example, the three scatterplots in Figure 4.14 allow estimation of the relative amounts of correlation in the three sets of measurements; the correlation appears to increase from left to right.



4.14 ESTIMATION OF CORRELATION. Scatterplots allow us to estimate the relative amounts of correlation in different data sets. Here, the amount appears to increase from left to right.

There are also numerical measures of correlation. The most widely used is the correlation coefficient. Let the two sets of measurements be x_i and y_i. Then the correlation coefficient is

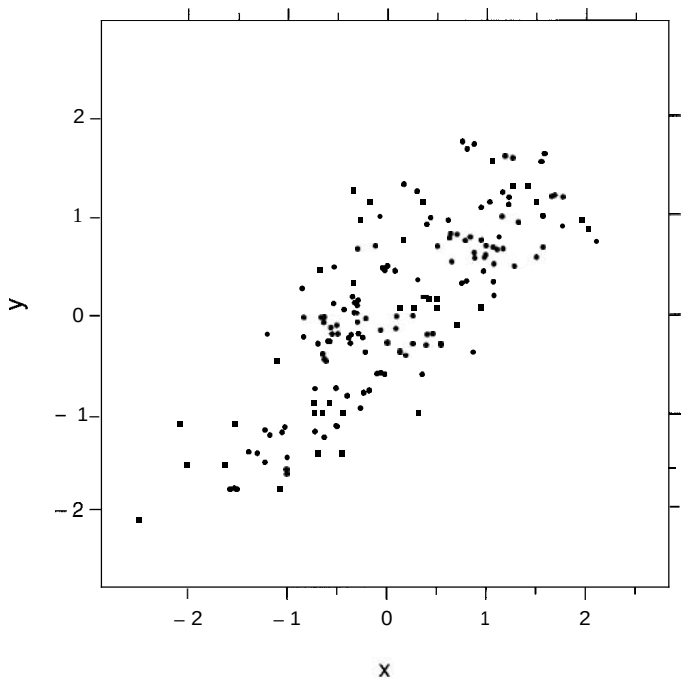
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} .$$

But any numerical measure can be defeated. For example, r measures merely the amount of linear correlation, and r can be severely distorted by outliers in the data, even a very small number of outliers.

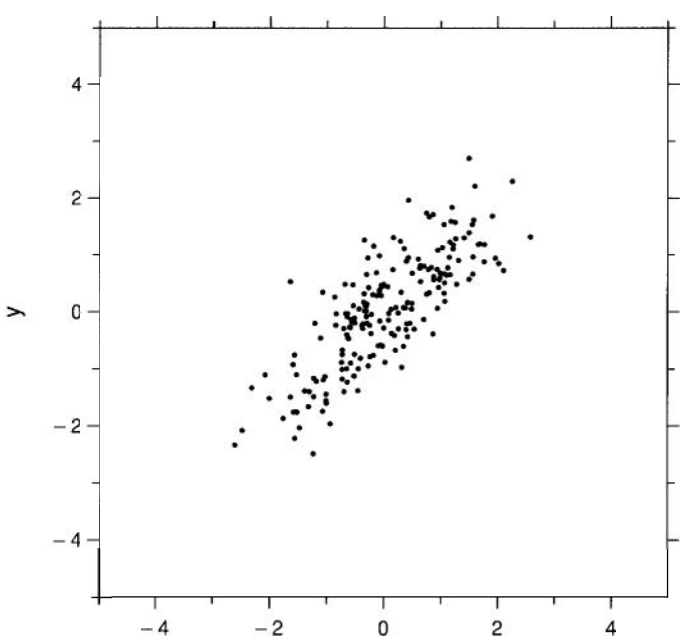
Experiments in graphical perception showed interesting properties of our estimation of correlation from scatterplots, one of them quite surprising [29].

First, our perceptual scale is closer to $\text{sign}(r)\, r^2$ than it is to r, where $\text{sign}(r)$ is 1 if $r > 0$, 0 if $r = 0$, and -1 if $r < 0$. This is illustrated in Figure 4.14. The amounts of correlation in the left and center scatterplots appear to be closer in value than the amounts of correlation in the center and right scatterplots. The values of r are equally spaced — 0.1, 0.4, 0.7. The values of $\text{sign}(r)\, r^2$ are closer to our perceptions — 0.01, 0.09, and 0.36.

The surprise is that our estimation of correlation is strongly affected by the area of the data rectangle relative to the area of the scale-line rectangle. As the ratio of the first to the second decreases, our perception of correlation increases. In Figure 4.15, the ratio is close to 1. In Figure 4.16, which graphs the same data, the ratio is much smaller and the amount of correlation appears to be greater. The lesson for data display is that we should keep the ratio constant when we judge the relative amounts of correlation on different scatterplots and that it makes sense to accompany such judgments with appropriate numerical measures.



4.15 ESTIMATION OF CORRELATION. Our estimation of correlation is affected by the area of the data rectangle divided by the area of the scale-line rectangle. In this example, the ratio is close to 1.



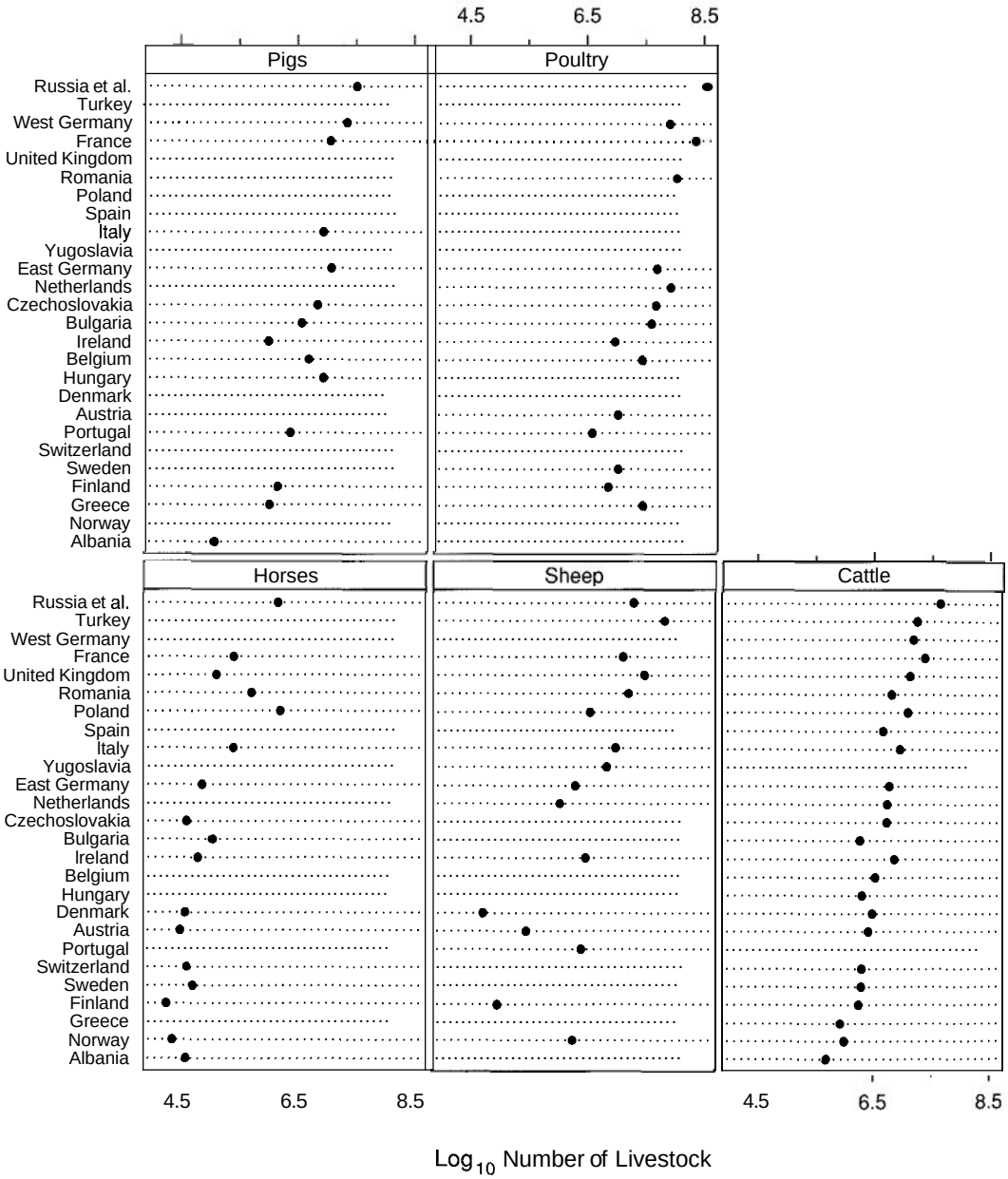
4.16 ESTIMATION OF CORRELATION. The data of Figure 4.15 are graphed again and the area of the data rectangle divided by the area of the scale-line rectangle is much smaller. The amount of correlation now appears greater.

4.9 Graphing Along a Common Scale

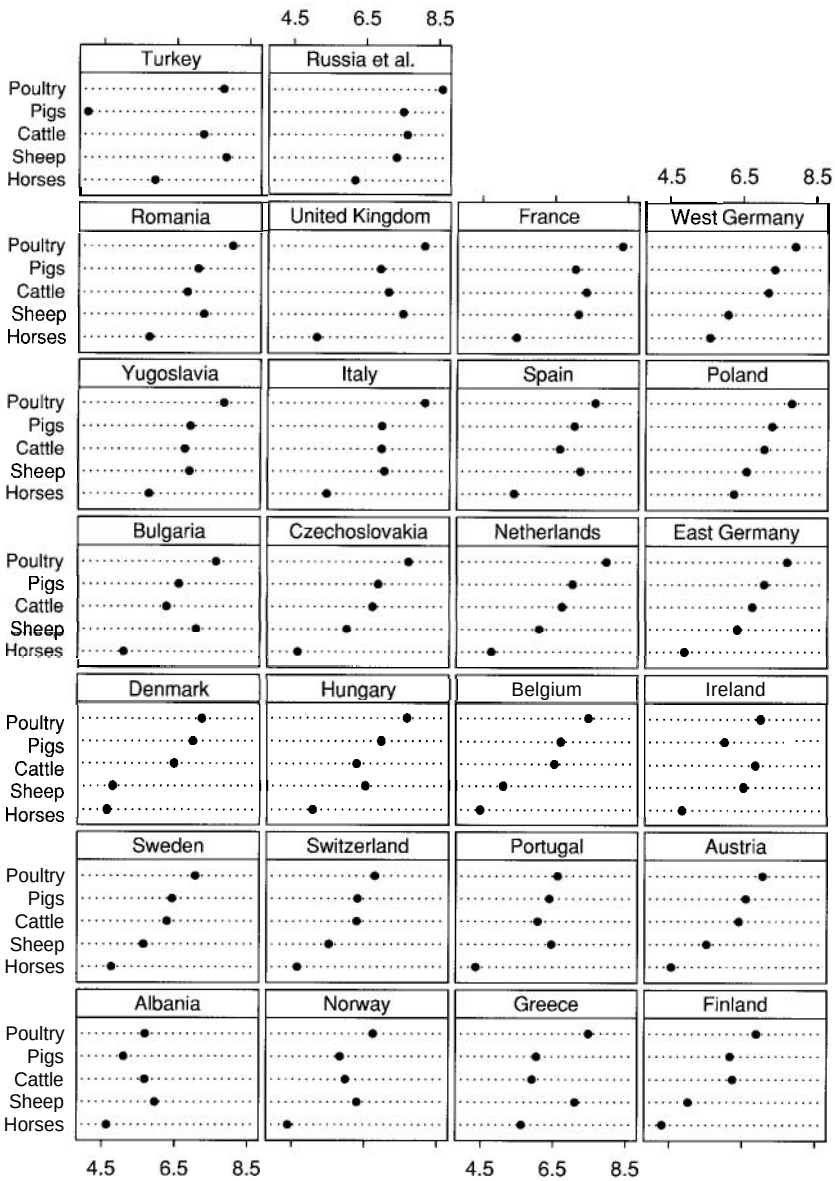
Figure 4.17, shown earlier in Section 4.6 (pp. 244–250), graphs the logarithms of livestock counts from a census of farm animals in 26 countries. The livestock variable is encoded by the panels and the country variable is encoded by the levels of each panel. The result is that the 26 values for each livestock type are on one panel; the 26 values are encoded by *position along a common scale*. But the five values for each country are graphed on the five panels, one value per panel; that is, they are graphed on *identical but nonaligned scales*.

The consequence of graphing the 26 values for each livestock type along a common scale, but not the five values for each country, is that we can more efficiently decode the values for each livestock type than the values for each country. There are two reasons for this. First, the assembly of values on a single panel is far more efficient. In Figure 4.17 we can effortlessly assemble the values for a particular livestock type, but the assembly of values for a particular country requires a slow sequential search. The second reason is a detection issue. Graphing values along a common scale allows us to detect geometric aspects that contribute substantially to our pattern perception. These aspects are not detectable when the values are graphed on identical, nonaligned scales. Consider the lower left panel in Figure 4.17. Going from the bottom to the top of the panel, the line segments connecting successive dots provide information about the magnitudes of the differences of the successive values. Also, through a visual projective process, we can also compare differences of any two values on the display. For example, to compare the horizontal separation between horses in Poland and horses in Greece, we can visually project the Poland dot down to the dotted line for Greece and then estimate the distance between the Greece dot and the projection.

This asymmetry for a multiway dot plot — better visual decoding for one categorical variable than another — is typically undesirable because we are typically interested equally in the effects of all categorical variables. For example, for the data in Figure 4.17, we are as interested in decoding values for each country as we are in decoding values for each livestock type. The solution is to make as many multiway dot plots as there are categorical variables, with each variable assigned once to the panels. In Figure 4.18 the country variable is assigned to the panels and now we can more effectively judge values for each country.



4.17 POSITION ALONG A COMMON SCALE. The data for each livestock type are graphed by position along a common scale, which allows us to effectively decode the distribution of the values for each type.



4.18 POSITION ALONG A COMMON SCALE. The data for each country are graphed by position along a common scale, which allows us to effectively decode the distribution of the values for each country.

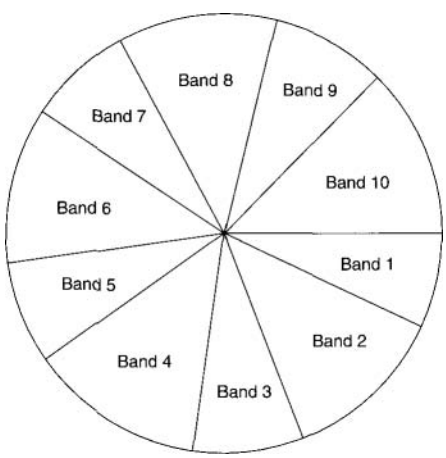
4.10 ***Pop*** *Charts*

Three graphical methods — pie charts, divided bar charts, and area charts — are widely used in mass media and business publications but are used far less in science and technology. Because of their use, we will call these graphical methods pop *charts.*

Any data that can be encoded by one of these pop charts can also be encoded by either a dot plot or a multiway dot plot that typically provides far more efficient pattern perception and table look-up than the pop-chart encoding. Interestingly, the better pattern perception results from a detection operation, a phenomenon that has been missed in previous studies of pop charts.

Pie Charts

Figure 4.19 is a pie chart that graphs 10 percentages. The labels, ten band numbers, are an ordered categorical variable; that is, Band 1 is first, Band 2 is second, and so forth. Kosslyn has argued that the sizes of sectors of a pie-chart encoding should increase circularly [75], just as we have argued in Section 4.6 (pp. 244–250) that categories of dot plots should be ordered so that the numerical values increase from bottom to top. But for our pie-chart example here, since the categories are ordered, we have not ordered the numerical values circularly, just as we do not order dot plots by the numerical values when the categories are ordered.

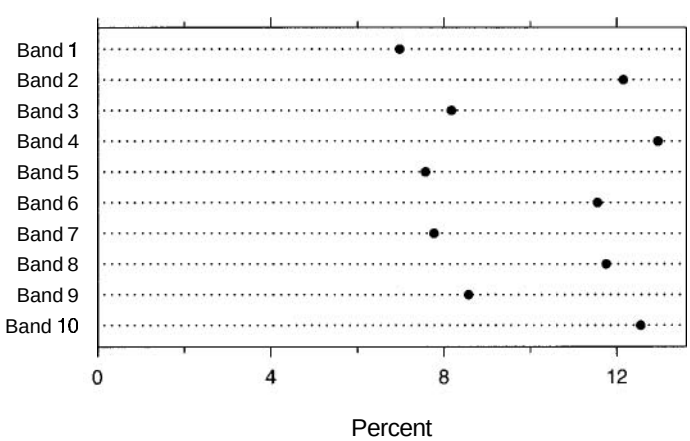


4.19 PIE CHART. The pie chart falls in the category of a pop chart — a graphical method used frequently in the mass media and certain business presentations but far less in science and technology. Both table look-up and pattern perception are less efficient for pie charts than for dot plots.

Figure 4.20 is a dot plot of the ten percentages from Figure 4.19. Pattern perception is far more efficient for this display than for the pie chart. We can effortlessly see a number of properties of the data that are either not apparent at all on the pie chart or are just barely noticeable. First, the percentages have a bimodal distribution; odd numbered bands cluster about 8% and even numbered bands cluster about 12%. Furthermore, the shape of the pattern for the odd values as the band number increases is the same as the shape for the even values; each even value is shifted with respect to the preceding odd value by about 4%.

The poor performance of pattern perception for pie charts is not restricted to cases with ordered labels. Many other simple demonstrations show that the poor performance is pervasive [8,121].

A dot plot graphs data by position along a common scale. As discussed in Section 4.9 (pp. 259–261), one strength of such position encoding is the detection of line segments that provide information about the differences of the graphed values. For example, in Figure 4.20 our ability to detect and cluster the line segments between the odd values and the line segments between the even values results in two gestalts that appear to be horizontal translations of one another, which allows us to readily perceive the 4% shift. There is no corresponding detection operation for the pie chart that allows effortless decoding of differences. The result is degraded pattern perception.



4.20 DOT PLOT. The data from Figure 4.19 are graphed by a dot plot. Patterns emerge that cannot be decoded from Figure 4.19.

This insight about detection has been missed in previous comparisons of pie charts and graphing along a common scale. Many studies treated estimation of the quantities directly graphed as the fundamental issue [36,75,114]. For example, in Figure 4.20 the horizontal line segments from the data dots to the left side of the scale-line rectangle encode the percentages. The studies compared estimation of ratios of such line segments with estimation of ratios of sector sizes of pie charts. Typically, the line segment estimation was found to be more accurate than the sector size estimation, and this was assumed to be the fundamental issue for the poorer pattern perception from pie-chart encodings. But the fundamental issue is the efficient detection of differences of values for position along a common scale; this is the cause of the better estimation observed in the studies.

In other studies it has been argued that the pie chart sometimes provides better estimation of the individual values of the percents [42,109]. The experimental information appears to support this assertion, but the result is largely an artifact of the experimental protocol. The increased accuracies for the pie charts in the experiments occur for percents in the vicinity of 25% and 50% and result from our ability to very accurately judge the sizes of 90° and 180° angles; these values serve as *anchors* [109]. The increase could be eliminated in such experiments simply by drawing reference lines at 25% and 50% on the displays that encode the data by position along a common scale. But more fundamentally, such experiments confuse table look-up with pattern perception. The decoding of individual values in isolation is not a salient visual operation for pattern perception.

Table look-up is also more efficient on dot plots than on pie charts. Most pie charts are drawn with no scales so we must judge the sizes of angles to infer percentages. This is both slower and less accurate than the scanning and interpolation discussed in Section 4.1 (pp. 223–227), which we use to carry out table look-up for data graphed along a common scale. It would be possible to draw a circular scale around a pie chart, but to carry out table look-up would still be slower and less accurate than for a dot plot because it would typically require two scans and interpolations to decode a single value.

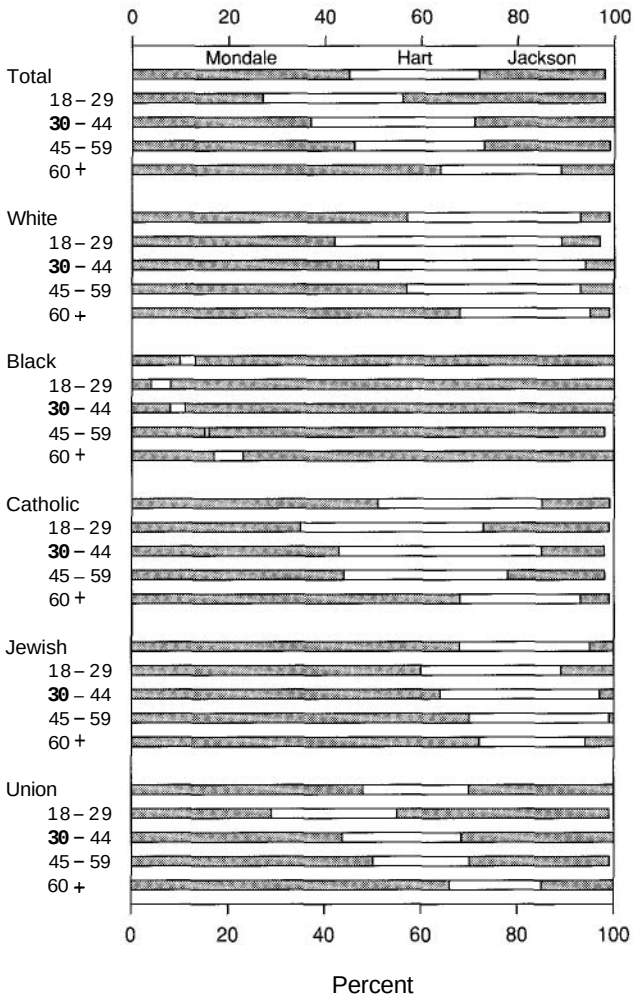
Divided Bar Charts

Figure 4.21 is a divided bar chart that shows the percentages of the vote for each of three candidates — Mondale, Hart, and Jackson — in a sample of 2016 voters leaving polling places in the 1984 New York State Democratic primary in the U. S. [117]. There are four age groups and six categories of voters. The percentages for the three candidates do not add to 100 in all cases because of rounding of the reported data, voting for others, or omitted answers.

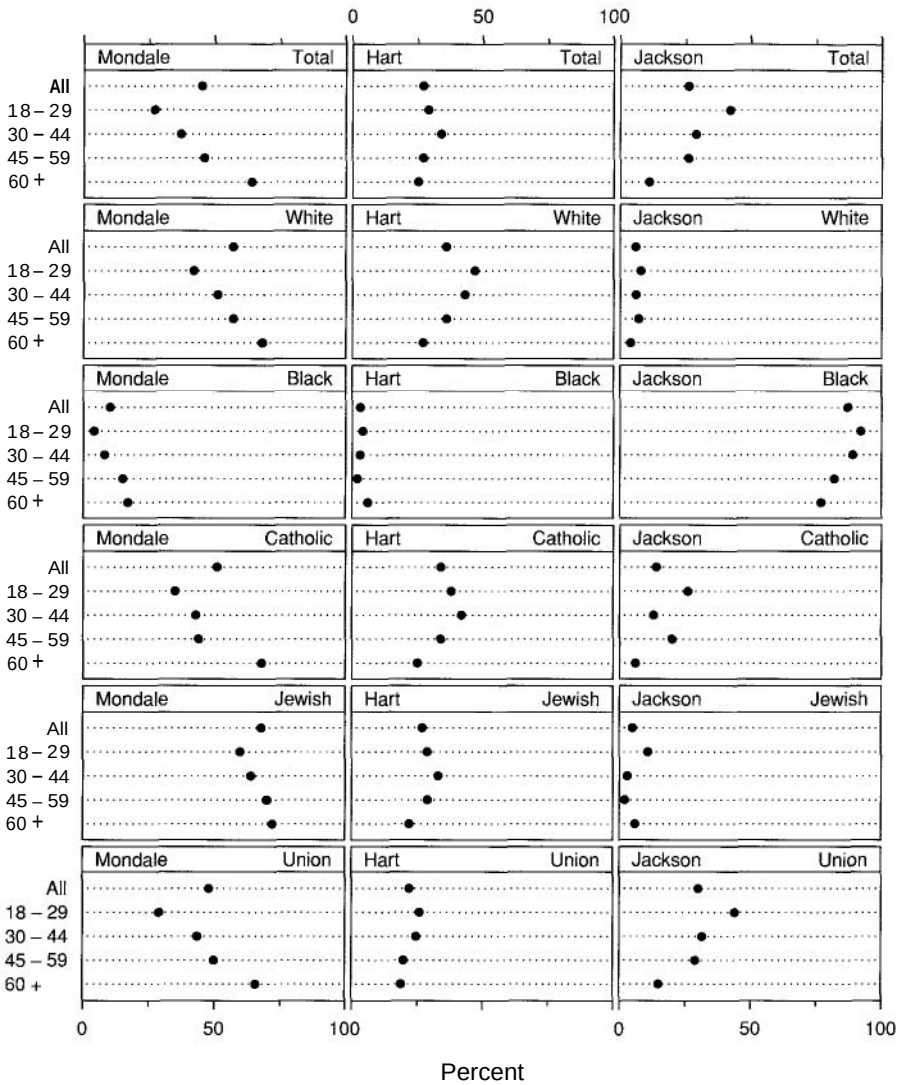
The Mondale bars in Figure 4.21 all have a common baseline at the left of the graph, and the positions of the right ends of the bars encode the Mondale vote. In other words, the Mondale values are encoded by position along a common scale and we have the benefit of increased detection to judge differences of Mondale values. We cannot do this for the Hart or the Jackson values; neither has a fixed baseline and so the lengths of the bars convey the values, not their positions. Consequently, there is no efficient visual mechanism for detecting differences of values. This reduces the efficiency of pattern perception for the Hart and Jackson values.

Figure 4.22 is a two-way dot plot of the voting data. Now the data for each combination of candidate and voter group are encoded by position along a common scale, not just Mondale. This allows us to perceive patterns in the data that are not apparent in Figure 4.21. For example, we see a Hart age effect. The 30-44 age group is often Hart's strongest; when it was not, the 18-29 age group is usually the strongest. In Figure 4.21 the Hart age effect is not readily apparent.

Table look-up is also far less efficient for divided bar charts than for multiway dot plots. Consider the value for the Hart white vote in the 18-29 age group in Figure 4.22. Scanning and interpolation to the top scale line readily provides a value of about 50%. To decode the same value from Figure 4.21 requires two scans and interpolations which is slower and less accurate.



4.21 DIVIDED BAR CHART. A divided bar chart is used to show the percentage of the vote for three candidates in the 1984 New York Democratic primary election. The Mondale values are graphed by position along a common scale, but the Hart values and the Jackson values are not and our visual decoding of these latter two sets of values is less accurate than for the Mondale values.

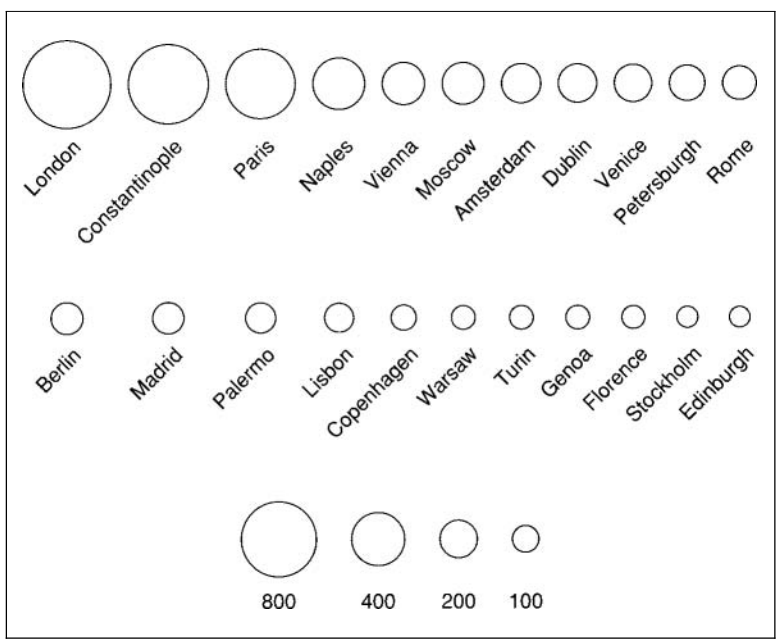


4.22 MULTIWAY DOT PLOT. The data from Figure 4.21 are graphed by a multiway dot plot. Now the Hart values and the Jackson values are encoded by position along a common scale. Now we can perceive a Hart age pattern.

Area Charts

In 1801, William Playfair published his *Statistical Breviary* 11051, which contains many displays of economic and demographic data. On one display Playfair encoded the populations of 22 cities by the areas of circles, although, as we saw in Section 3.2 (pp. 126–132) his encoding had errors as large as ±15%. Playfair may have been the first person to make such an area chart.

Figure 4.23 graphs the Playfair population data by circles whose areas are proportional to the data. As with pie charts and divided bar charts, area charts do not provide efficient detection of geometric objects that convey information about differences of values.

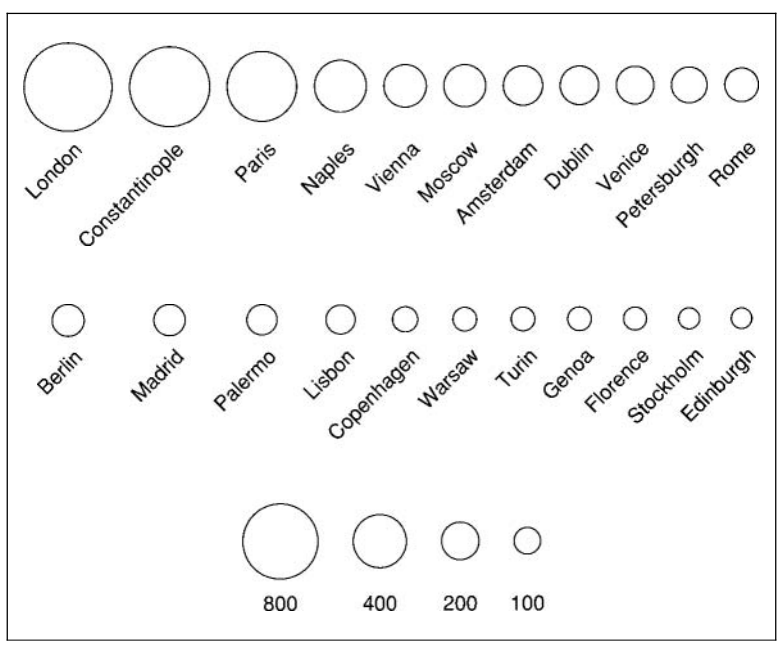


4.23 AREA CHART. Populations of European cities are graphed by an area chart. As with pie charts and divided bar charts, area charts do not have geometric aspects that can be detected to provide information about differences of graphed values.

Area Charts

In 1801, William Playfair published his *Statistical Breviary* [105], which contains many displays of economic and demographic data. On one display Playfair encoded the populations of 22 cities by the areas of circles, although, as we saw in Section 3.2 (pp. 126–132) his encoding had errors as large as ±15%. Playfair may have been the first person to make such an area chart.

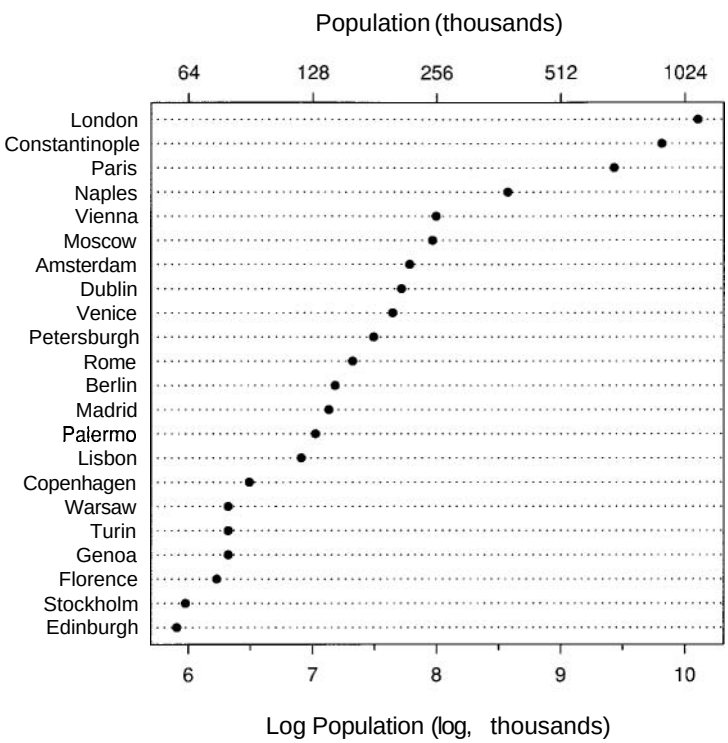
Figure 4.23 graphs the Playfair population data by circles whose areas are proportional to the data. As with pie charts and divided bar charts, area charts do not provide efficient detection of geometric objects that convey information about differences of values.



4.23 AREA CHART. Populations of European cities are graphed by an area chart. As with pie charts and divided bar charts, area charts do not have geometric aspects that can be detected to provide information about differences of graphed values.

Figure 4.24 is a dot plot of the population data using a log scale. Now the data are graphed by position along a common scale and pattern perception is far more efficient than in Figure 4.23. For example, it is hard from Figure 4.23 to detect a change in the circle areas from Petersburgh to Lisbon, but Figure 4.24 shows that the populations vary by a large factor.

Table look-up is far more accurate and rapid from Figure 4.24 than from Figure 4.23. The matching operations necessary to decode values from Figure 4.23 are both slower and less accurate than the scanning and interpolation operations that provide table look-up from Figure 4.24.



4.24 POSITION JUDGMENTS. The data from Figure 4.23 are graphed by a dot plot with a log scale. Now the data are encoded by position along a common scale and the efficiency of pattern perception and table look-up is much greater.