Chapter 2

The Gospel According to Tufte

"Above all else show the data"
—— Edward Tufte (1983)

2.1 Data-Ink

One of Tufte's major themes is the good graphics present their message as simply as possible. A good principle, surely, but as vague as the Christian theme of "Love Thy Neighbor". Who is thy neighbor? And what is graphical simplicity?

To turn simplicity into more practical ideas, Tufte defined the following:

Definition 1 (DATA-INK) *The non-erasable core of a graphic.*

Definition 2 (DATA-INK RATIO)

$$\begin{aligned} \textit{data-ink ratio} &= \frac{\textit{data-ink}}{\textit{total ink used to print the graphic}} \\ &= \textit{the proportion of a graphic's ink devoted to the non-redundant} \\ &\qquad\qquad \textit{display of data-information} \\ &= 1.0 - \textit{proportion of a graphic that can be erased without} \\ &\qquad\qquad \textit{loss of data-information} \end{aligned}$$

The concept of "data-ink" doesn't completely solve the problem because the question of what actually is "non-erasable" depends on both the problem and the readership. This concept is much more concrete than the vaguer idea of "simplicity", however. Tufte further elaborates his concepts in the form of five maxims.

Tufte's Five Laws of Data-Ink:

- Above all else show the data.
- Maximize the data-ink ratio
- Erase non-data-ink.
- Erase redundant data-ink.
- Revise and edit.

2.1.1 Show the Data

This is the most important part of the five maxims because the “data-ink” is undefined until one has first developed a purpose for the graphic. Composition teachers press their students to begin the creation of an essay by first writing a “topic sentence” that summarizes: What is the essay about? What is the big idea I want to get across to my readers?

Writing a topic sentence for each graph before you begin to compose it is not a bad idea for visualization either. Exploratory graphics is different; in the early stages of a study, one may simply mess around with different views of a hydrodynamic flow, looking for something interesting. When the time comes to design a graph for a journal article or a thesis, one had darn well better be able to supply a topic sentence for the graph, or the work simply isn’t ready for a final draft.

“This graph will show the variation of the growth rates of the Charney instability as a function of the zonal wavenumber.” This is an example of a topic sentence. With this in hand, we can then begin to ask a slew of other questions before we begin to draw.

First, what do we need to compare with the Charney growth rates? In other words, can this curve stand alone, or is it useful only if plotted with other curves on the same axes? Second, is it useful to compare the graph with other figures, such as the growth rates as a function of latitudinal wavenumber? We can’t plot both curves on the same axis since the zonal wavenumber is different from the latitudinal wavenumber. We might nevertheless wish to compare them to show that growth rate is insensitive to the latitudinal wavenumber, but varies strongly with the zonal wavenumber. In this situation, we make it easier for the reader if both graphs have a common format, linestyle, consistent labels and so on and also if both graphs have clear labels that call the reader’s attention to the fact the ordinate is the same, but the abscissa variable is different.

Third, what wider context needs to be explained to the reader so that the concept of “Charney growth rates” is meaningful? The “Charney problem” is a model for the development of midlatitude storm systems. To connect the Charney problem with the real world, it might be necessary to include some graphs of atmospheric observations, perhaps copied with permission from other authors, in an introductory section. And what are “growth rates”? Context is a mixture of good graphs and explanations that orient the reader like the compass rose on a map.

2.1.2 Emphasize the Data

“Maximize the data-ink ratio” is a very general precept that motivates Tufte’s remaining three maxims. Unfortunately, almost all graphics require some non-data elements, such as axis lines, tic marks and labels. One can, however, deemphasize these elements.

One way is to draw the data curves using thicker lines than the axis lines and frames. Fig. 2.1 compares two graphs which are identical except that the right panel has a higher data-ink ratio because more ink has been used to draw the thick line for the data.

Emphasizing the data is purely an artistic touch because the cognitive content of the graph is not altered. However, design touches do matter because scientists and engineers always have too many papers to read and too little time. A paper with clear, easy-to-decode graphs will make a much more lasting impression than one with confusing illustrations that require a lot of concentrated attention.

2.1.3 Erase Non-Data-Ink or Down with the Grid!

Tufte and Howard Wainer both regard grid lines as the Spawn of Satan. This is perhaps too extreme, but one certainly should be selective in the use of grid lines.

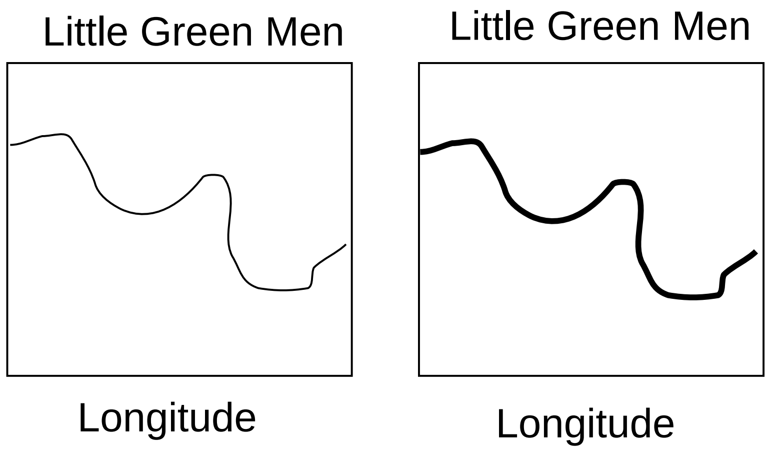


Figure 2.1: Two graphs identical except that the data-curve has been thickened on the right. Which graph do you prefer?

Historically, grid lines became part of the mental framework of technologists because in the pre-computer days, figures were usually drawn by hand on "graph paper". This was special paper which was printed with a fine checkboard pattern of both horizontal and vertical grid lines. This network of grid lines enormously simplified the task of plotting individual points by hand.

Because scientists were accustomed to seeing grid lines when MAKING graphs, they came to expect grid lines when READING graphs, too. At the very least, scientists of the pre-computer-graphics era were very good at ignoring grid lines and focusing on the data.

Even in those ancient days, however, there was a certain ambiguity about grid lines. First, printed "graph paper" always had the lines printed FAINTLY so as to avoid clashing with the data curves — sound practice for twenty-first century grids, too. Second, the graph lines were printed in a special "non-reproducing" green ink. When a hand-drawn graph was xeroxed, the green ink was NOT VISIBLE on the copy.

In the age of computer graphics, we should be even more sceptical of grid lines. Our readership is different from that of 1970: scientists and engineers under thirty-five have rarely used graph paper. They do not expect grid lines; their visual systems are not so good at unconsciously filtering them from the graph when the mind is trying to concentrate on the data-curves.

Fig. 2.2 compares two graphs. The data-curve is much easier to read in the right panel because it does not have to compete with the grid.

Are there exceptions when one should include a grid? One exception is when the figure is a NOMOGRAM, that is, a graph which will be used as a sort of graphical calculator. Grid lines are very helpful if the reader has to use the figure to generate numbers.

Nomograms used to be very common, but are now reserved for special applications. One reason for the decline of the nomograph is that many key numbers, such as the maxima and minima of a curve and associate numbers with curve features. Longer strings of numbers can be supplied as a TABLE. Third, in those rather desperate situations where one needs access to a lot of numbers with high accuracy, one option is to archive a data file on a Web site or a CD-ROM. Another option is to print a table which is a listing for a short computer code to reproduce the graph or to list the coefficients of a curve-fitting polynomial.

A second exception to the general disparagement of grid lines is when the author expects the reader to carefully study a curve and try to pick out the heights of local maxima and minima. Again, the numbers could all be given in the caption, but when trying to absorb a

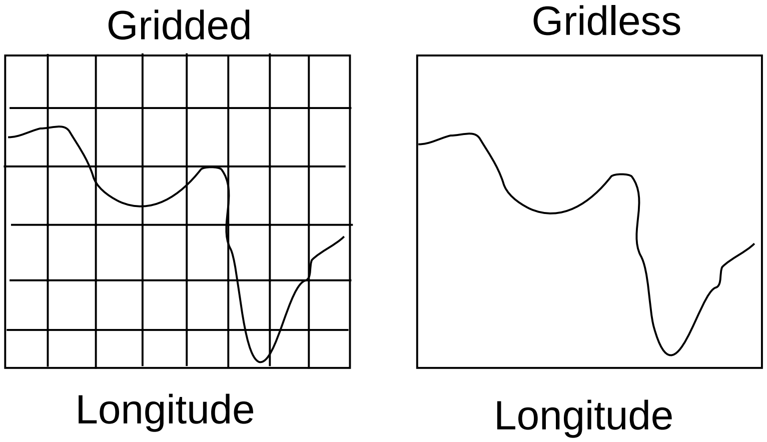


Figure 2.2: Two graphs identical except that the grid has been omitted on the right.

data-curve, the reader doesn't want to keep his eyes moving back and forth between graph and caption. Grid lines can be helpful when the reader wants to do some "qualitative nomography", so to speak.

A third exception is for certain species of three-dimensional diagrams where a grid box may be helpful in supplying visual clues to correctly interpret ink patterns on a two-dimensional plane as representing a shape in three-dimensional space (Fig. 2.3).

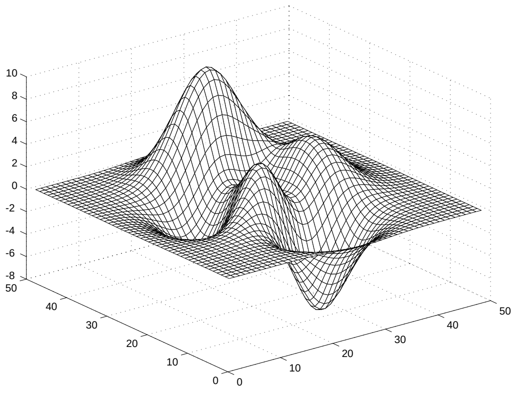


Figure 2.3: A surface mesh diagram with a faint grid box to help the reader visualize the three-dimensional shape.

A fourth exception is when the grid itself carries the message. Fig. 2.4 is a graph twenty years ahead of its time. Pierre Welander showed through this illustration that a smooth vortex flow could distort a blob of fluid into a very wierd object. Although the word "fractal" did not exist in his time, he understood that as time increases, the shape of the marked portion of the fluid would tend to an object with an infinite boundary but a finite area. Carl-Gustav Rossby, perhaps the most celebrated meteorologist of the century, reproduced this figure in a 1959 review article. It was written for a book to honor Rossby's sixtieth birthday, but he died of a heart attack shortly after finishing his review, and it appeared as part of the *Rossby Memorial Volume*. It was not until the appearance of Benoit Mandelbrot's book *Fractals* in 1973 that a word for such wierdly-shaped objects, and an appreciation for their ubiquitous appearance in nature, gained currency.

The most amazing thing about this figure, and the article in which it appeared, is that Welander had a good understanding of ideas that became popular only many years later. However, it is also a graphical masterpiece. The curve that bounds the marked region of fluid, initially a square, carries the most important message, which is that the perimeter of the blob is being stretched and stretched and stretched without limit. However, the checkerboard of black squares and white squares conveys another message: that distortion is also occurring *within* the marked blob of fluid.

Note that this grid is not used for the usual purpose: To facilitate the conversion of the data curve into numerical values. Welander's grid has no axis labels to allow such conversion. The checkerboard grid is not to serve as an adjunct to the data curves, but to be itself a graphical representation of the data.

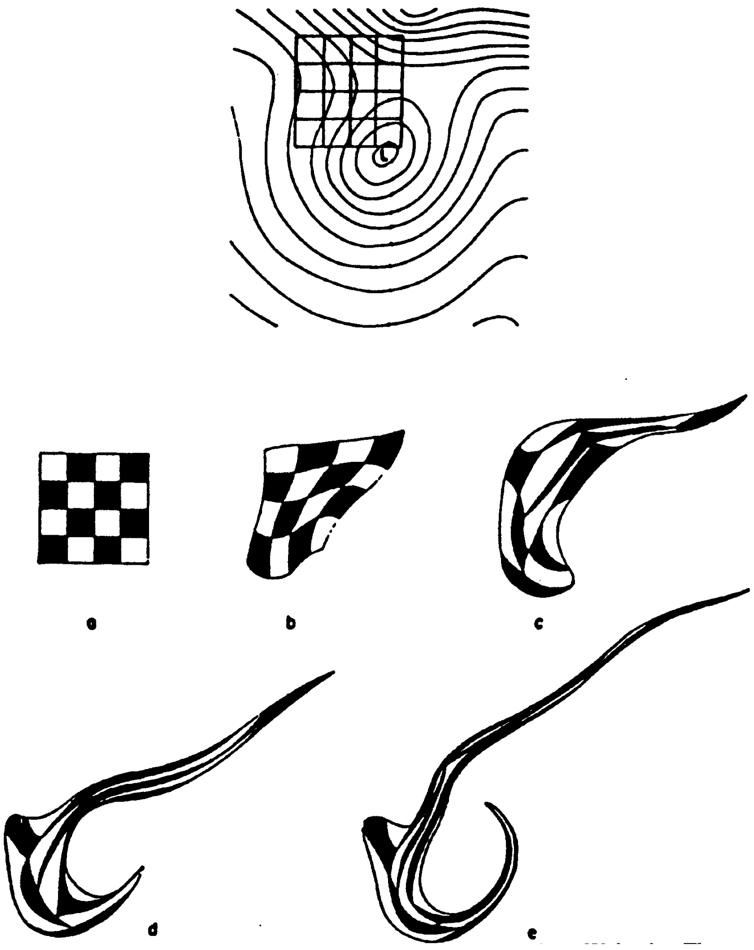


Figure 2.4: This diagram by the late Pierre Welander was published in *Tellus*(1955). The quasi-circular contours in the top diagram are the streamlines of an atmospheric or ocean vortex; the velocity vector is everywhere parallel to the streamlines. The lower figures show how the 25-cell checkerboard superimposed on the vortex is distorted over time by the vortex flow. The crucial point is that fluid mass is teased out into long, stringy filaments. To use modern language, although the vortex flow is laminar, the advection distorts the initial square of fluid into a curve whose perimeter is increasing without bound even though the area is constant (because of mass conservation). As $t \rightarrow \infty$, the perimeter of the marked fluid approaches a *fractal*, that is, an object which has a fractional dimension in the sense that it is an object of infinite length with a finite area.

The key guidelines for a grid are:

- Don't use a grid unless you really have to.
- Make the grid lines faint compared to the data-curves by drawing the grid as thin lines or dotted lines or by using a thick line for the data.

Fig. 2.5 illustrates these guidelines.

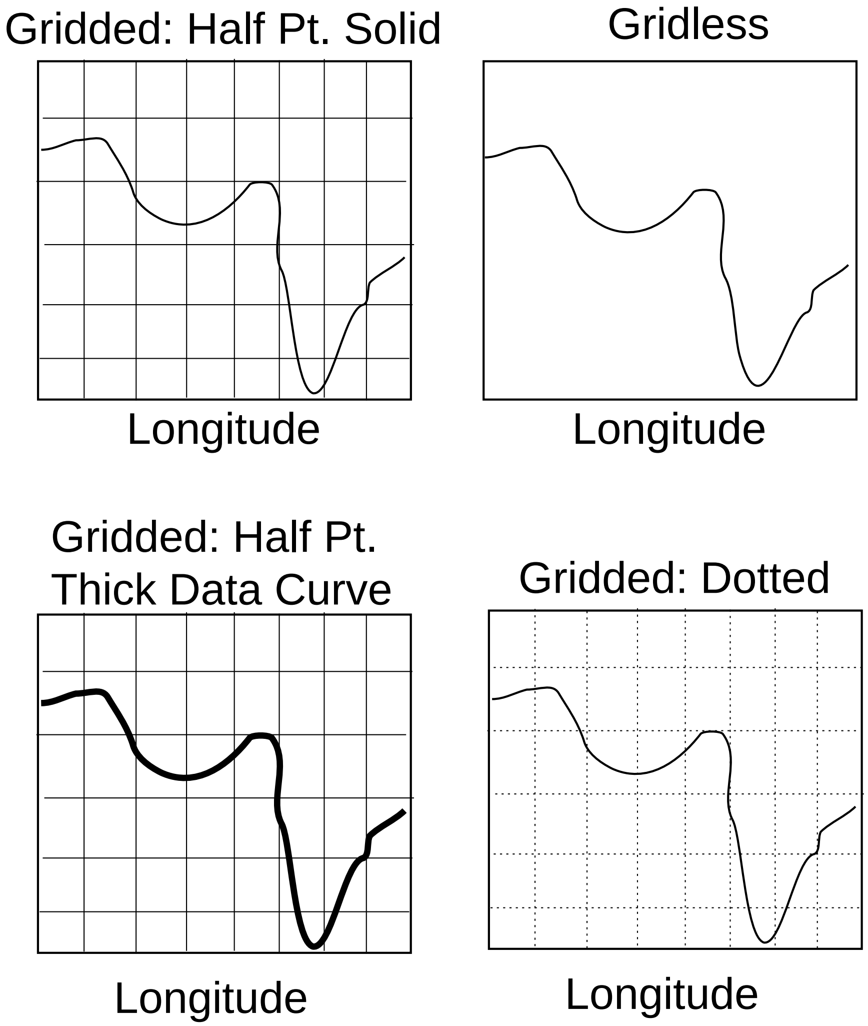


Figure 2.5: Strategies for a grid. The data curve is most visible when there is no grid at all (upper right). If a grid is used, it should be muted relative to the data curve. This can be done by using very thin grid lines (upper left), by thickening the data-curve so it is much darker than the grid lines (lower left), or by making the grid lines dotted (lower right).

2.1.4 Erase Non-Data-Ink: Hurrah for Half-Framing!

Tufte and some other technical artists such as Mary Helen Briscoe (1996) are advocates of another simplification: half-framing, which is to say, drawing only the the usual horizontal and vertical axes and omitting framing lines on top and the right. In contrast, the default in Matlab is to draw a full frame: whenever axes are drawn, a second line parallel to the horizontal axis is drawn at the top of the graph while a second vertical parallel to the vertical axis is drawn on the right so that the data curves are enclosed in a rectangle. Fig. 2.6 contrasts these two styles of framing.

I have no strong opinions. A full frame is common, and I personally am most comfortable with this since it is what I am used to. However, no data is lost or hidden by using only a half-frame. The very notion of a half-frame emphasizes the "data-ink" theme: strive for simplicity and include nothing irrelevant.

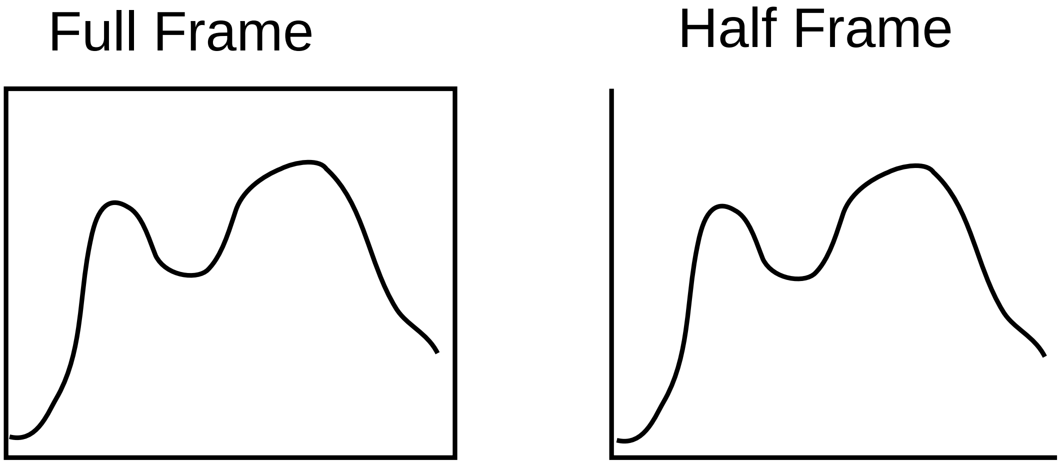


Figure 2.6: Left: a graph with a full rectangular frame. Right: a "half-framed" graph in which only the two axes are visible boundaries to the data-curves.

2.1.5 Erase Non-Data-Ink: Example of the Simplified Bar Graph

Tufte (1983) redesigned a bar chart to illustrate his theme that much can be erased in a graph without compromising the data. The bars have been replaced by thick vertical line segments; short horizontal segments connect each pair of bars, linking them into a single group as in the original bar graph shown on the left in Fig. 2.7.

How much information has been lost in the redesign? Nothing. The bars themselves are redundant in the sense that they can be replaced by a simpler element (here, a line segment) that also has a length. Placing pairs of bars side-by-side and shading one of each pair is unnecessary, too. The bars can be grouped by a short, connecting segment. The shading is unnecessary because the left and right bars in each group are always distinguishable. Finally, one can replace a full frame by a half frame.

The only disadvantage of Tufte's format is unfamiliarity. Because a bar chart is familiar and intrinsically rather simple, a bar chart is quickly and easily decoded even though it is not, strictly speaking, maximally simple. Tufte's new format is unfamiliar, and the reader is likely to stare hard at the text and caption to ask: Why did he choose this novel format? What information is supposed to be conveyed here? Do I understand this strange species of graph, or have I missed something important?

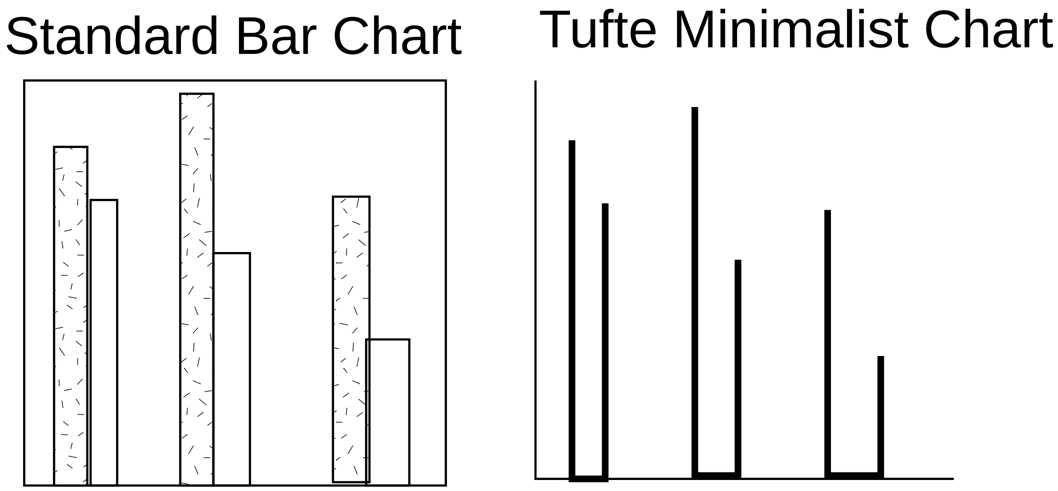


Figure 2.7: Left: a standard bar graph with multiple comparisons. Right: the same in Tufte's minimalist bar chart format.

2.1.6 Erase Redundant Data-Ink: Symmetry and Wrap-Around and All That

If a function has the property

$$f(x) = f(-x), \qquad \text{for all } x \qquad \Rightarrow \text{Symmetric} \tag{2.1}$$

then the function is said to be "symmetric with respect to $x = 0$". Similarly, if for all x,

$$f(x) = -f(-x), \qquad \text{for all } x \qquad \Rightarrow \text{Antiymmetric} \tag{2.2}$$

then f is "antisymmetric with respect to $x = 0$". A function which is either symmetric or antisymmetric is said to be of "definite parity" where symmetry is "even parity" and antisymmetry is "odd symmetry".

One hardly needs to learn these jargon terms to know that symmetry and antisymmetry are very common in both the natural and man-made worlds. Human beings and most animals are approximately symmetric about a line draw from the middle of the head to between the feet. The right and left sides of a car or truck are externally symmetric but unsymmetric on the interior.

Because parity is so common, our brains are very good at visualizing the complete object from a picture of only the left or right sides. Indeed, psychology experiments have shown that when a person studies a complete picture of a symmetric object, she looks carefully at one half and then barely glances at the other half, just enough to confirm that the object does indeed have definite parity. It follows that if the reader isn't going to bother to look at the left half of a symmetric or antisymmetric object, one can save a lot of space by erasing the unlooked-at-half. (The figure caption must clearly state that the other half has been omitted because the object has parity, and also clearly state whether it is symmetric or antisymmetric.)

Nevertheless, I sometimes graph both halves of a symmetric object. Why? The reason is that the most powerful way to convey the parity of an object is to allow the reader to see it for himself. As a verbal description, the key word "symmetry" might be missed. A graph that shows only positive values of the abscissa does not necessarily imply symmetry. For example, when polar coordinates are employed, the radial coordinate r is defined only for $r \in [0, \infty]$. It follows that if coordinate labels values are all positive, it is not possible to conclude that the object has definite parity. One needs additional information from the caption or text to unambiguously identify the symmetry of the figure. It is sometimes best, especially if symmetry is an important property that deserves emphasis, to simply SHOW the symmetry.

Fig. 2.9 shows two versions of flows in a global ocean. To depict the surface of a sphere, the ocean flow must be sliced along some longitude and then projected onto a flat map. The perceptual problem is that the left and right edges of the map are the same longitude. The reader must mentally connect the left and right edges to truly visualize the flow.

Tufte's proposed revision is to expand the graph so that it depicts two-thirds of the globe twice. Because of this redundancy — a necessary and helpful redundancy in Tufte's opinion — the reader can see the flow in all parts of the globe without the need for mental reconstruction.

The broad message is that sometimes redundancy IS necessary for complicated visualization. This is certainly true. However, the narrow issue of whether redundancy is good for a particular illustration depends on the context and the readership.

Being a political scientist, Tufte is not particularly adept at mentally stitching opposite sides of the globe together. However, oceanographers ALWAYS have to perform this mental task because the earth is always round, and every physical oceanography article is forced

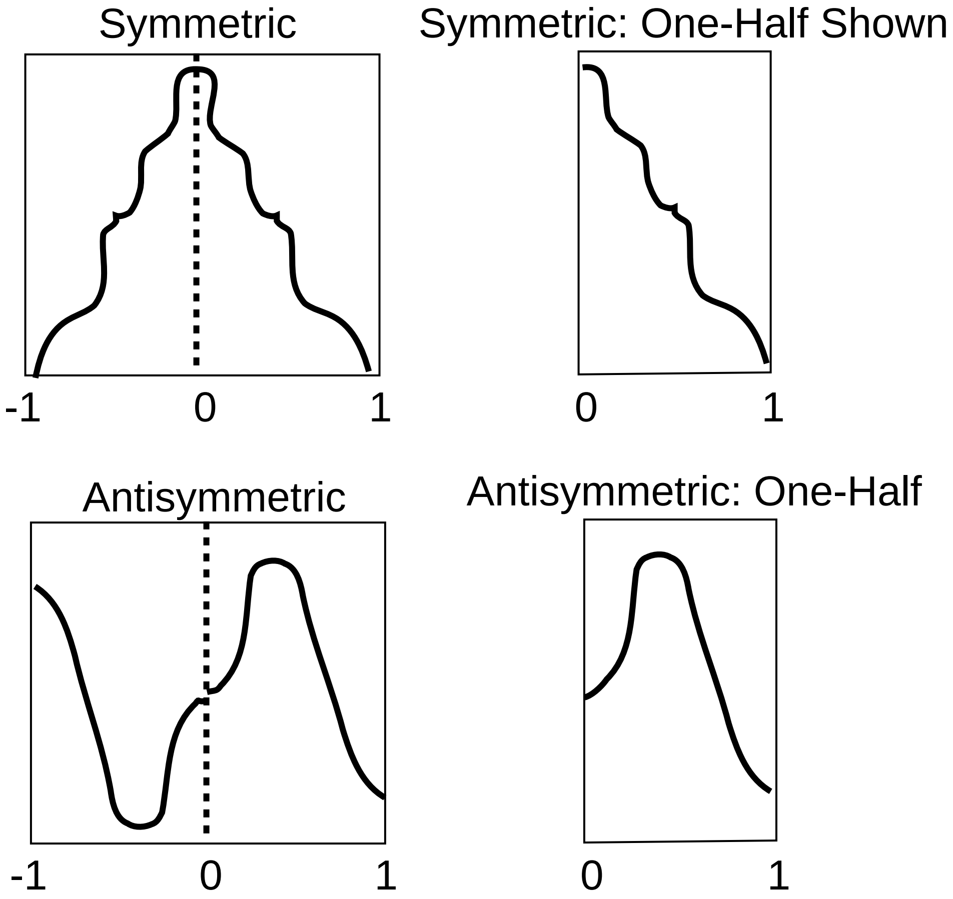


Figure 2.8: Schematic of a symmetric function ($f(x) = f(-x)$ for all x, top pair of graphs) and an antisymmetric function ($f(x) = -f(-x)$, bottom pair.) The left graphs illustrate the full function; the dotted line at $x = 0$ denotes the symmetry or antisymmetry plane. The right pair of graphs shows how the graphs can be simplified by plotting the right half of $f(x)$.

to describe spherical geometry through a flat map. So, oceanographers are not nearly as bothered by the Bryan-Cox diagram as Tufte himself. His suggested redundancy has rarely, if ever, been applied in oceanography.

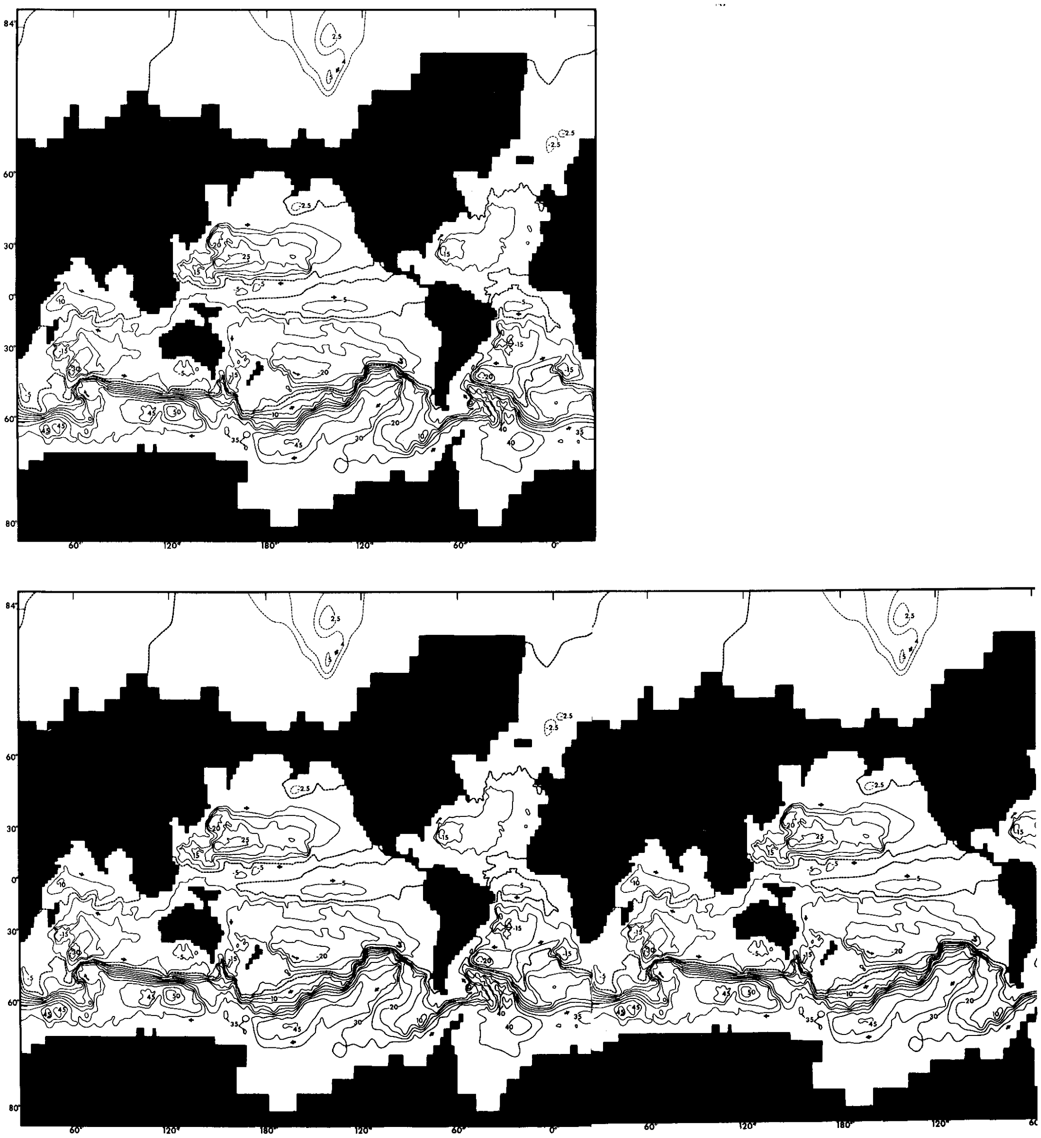


Figure 2.9: Top: Streamlines of ocean flow in a computational model as originally published by Kirk Bryan and the late Mike Cox in 1972. Bottom: Tufte's suggestion revision, which shows 5/3 of the globe so that one may see all ocean basins without the need to mentally connect the left and right sides of the top figure, which are the same longitude although on opposite sides of the map.

2.1.7 Erasing: Eliminating the Graph Entirely

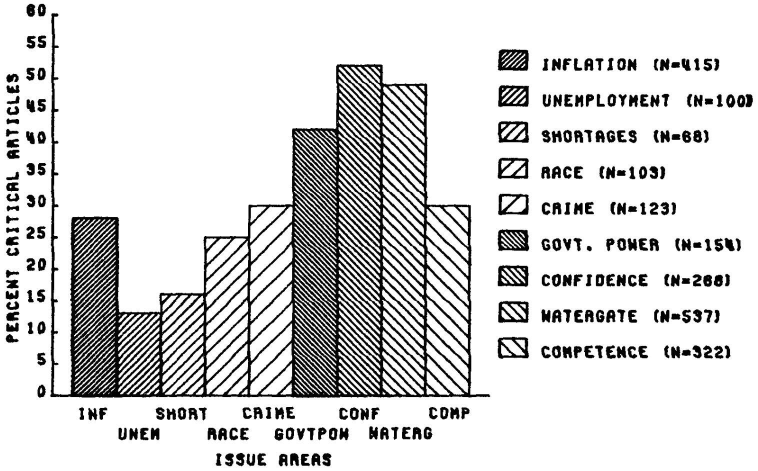


Figure 2.10: A graph that should be a table. So many bars are needed that the labels must be nested at three different levels. Furthermore, it is necessary to shade the bars with different patterns to make them distinctive; the rather small number of data numbers is almost obliterated by all the shimmering Moiré patterns.

Sometimes the best way to cope with a flawed graph is to eliminate the illustration entirely, and use a table instead. Fig. 2.10 is a graph that Tufte asserts is better replaced by a table.

The figure's first flaw is that one should eliminate the legend, and directly label each bar. (This is sound general practice.) Unfortunately, it is not possible here because there simply isn't room to tack nine lengthy labels under the horizontal axis. So, the authors stuck abbreviated, confusing labels — on three levels, no less — under the graph and added a legend box.

The second flaw is that the shading patterns add a lot of extra visual elements, distracting attention from the HEIGHT of the bars which actually carries the information. However, if the shading were omitted, it would be hard for the eye to distinguish one bar from another.

One could do better by turning this bar graph on its side so that the bars project horizontally. This allows long labels for each bar. Furthermore, since the page is longer than it is wide, one can add enough white space between the bars so that they are distinguishable without Moiré patterns in each bar.

However, Tufte's more radical solution is to replace the graph with Table 2.1.

Table 2.1: Table replacing Graph from pg. 121 of Tufte(1983)

Content and tone of front-page articles in 94 U. S. newspapers, October and November, 1974	Number of articles	Percent of articles with negative criticism of specific person or policy
Watergate: defendants and prosecutors, Ford's pardon of Nixon	537	49%
Inflation, high cost of living	415	28%
Government competence: costs, quality, salaries of public employees	322	30%
Confidence in government: power of special interests, trust in political leaders, dishonesty in politics	266	52%
Government power: regulation of business, secrecy, control of CIA and FBI	154	42%
Crime	123	30%
Race	103	25%
Unemployment	100	13%
Shortages: energy, food	68	16%

2.1.8 Revise and Edit

Writing teachers emphasize the important of multiple drafts. Indeed, “first draft” has become a perjorative term. The initial version of a document can almost always be clarified by revision and editing. Similarly, fine arts teachers and graphic designers stress the usefulness of preliminary sketches and of techniques for making changes on the painting or drawing itself.

Revision and editing are equally important for scientific visualization. The intellectual content is not changed by editing, only the clarity. However, in an era of “Short Attention Span Theater”, a graph is pointless unless it is clear.

Tufte illustrates the revise-and-edit cycle through a case study. The figure was originally published in a 1947 textbook by the Nobel Laureate Linus Pauling. It was drawn by Roger Hayward, whom Tufte describes as a “distinguished science illustrator” even though he strongly criticizes the published figure, which is shown as the upper left panel of Fig. 2.11.

On the upper right are all the graphical elements that Tufte considers redundant. First, the full frame has been reduced to a half-frame by eliminating the boundary lines on the top and right. Second, the original figure labeled every tick mark, but labelling every other tick mark is satisfactory. Lastly, the grid, which here is a lattice of plus signs, is redundant.

The middle left figure shows the result when all these redundant elements are eliminated. Version 2 is a vast improvement over the original. However, a pretty good graph is not the same as the best graph, so Tufte experimented with other variations.

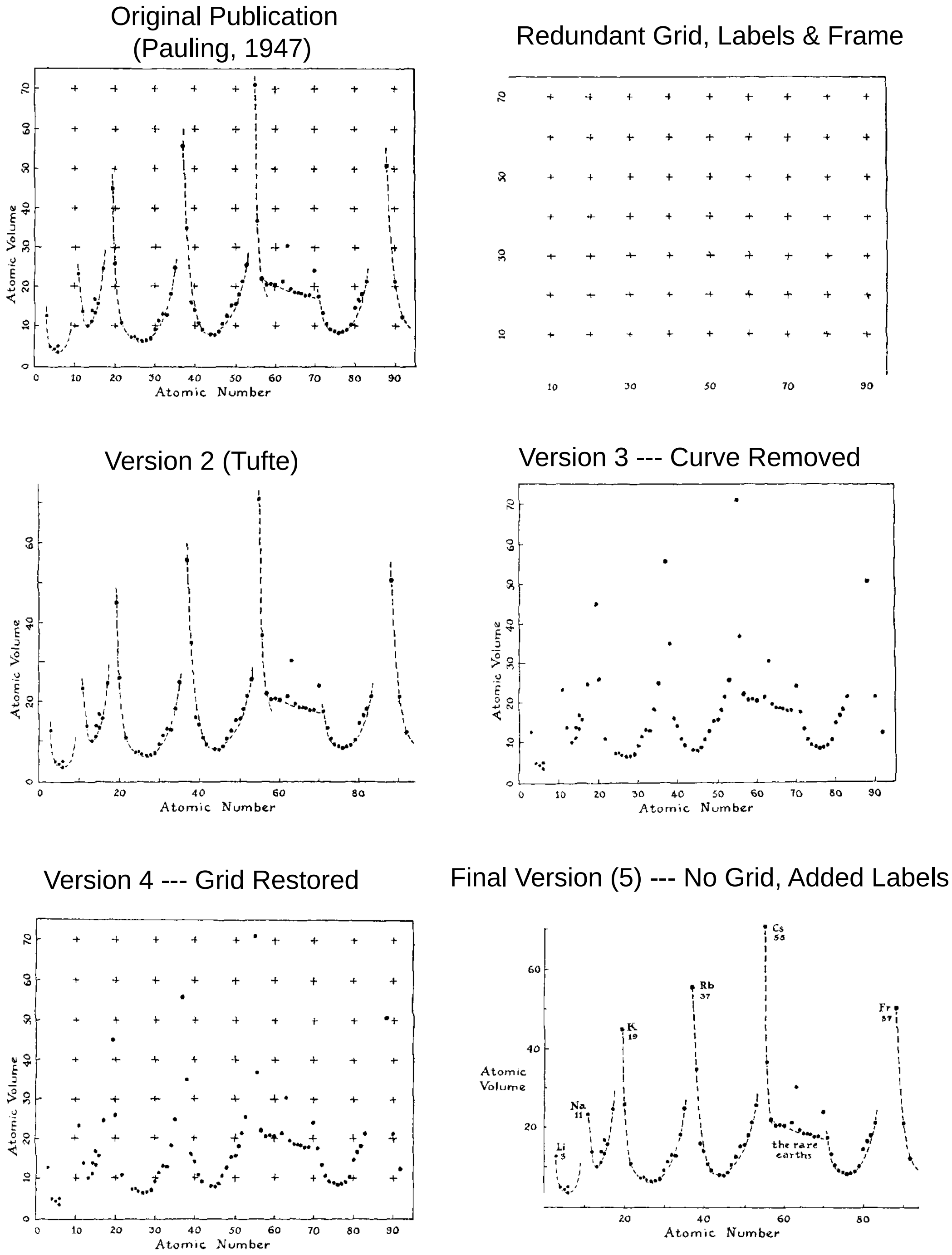
In Version 3 (middle right), the dashed curve connecting most of the points has been eliminated. Although it is unnecessary for cognitive content — the data points are still there — the curve is helpful in guiding the eye to make sense of a pattern of points. Without the dashed curve, the figure is simply harder to read.

The next version (lower left) restores the grid. The result is both confusing and ugly. This was not a serious experiment, but rather another opportunity for Tufte to attack grids.

His final version (lower right) is similar to Version 2 except that it has more labels. Each of the peaks is now labelled with the name of the element and its atomic number. The plateau between atomic numbers 60 and 70, which breaks the pattern for rapid rise or fall of the rest of the graph, is now labelled “the rare earths”. Because of the labels and the reference curve, the significance of the peaks and the plateau is almost instantly perceptible.

This case study makes several points. First, experiment: try several versions of the same figure and save the one that you like the best. Experimentation is easy with Matlab, which allows one to add or subtract labels, delete or add axes and so on, with a single command.

Second, “redundancy” is a matter of rather subtle judgment. One could, for example, thin out the axis labels still more, delete the dashed curve, and all the labels of the peaks and the plateau, and the data points would all still be there. The point is not that redundancy or non-data-ink is necessarily bad, but rather that non-data-ink should be there for a reason: to make the graph clearer.



Original Publication (Pauling, 1947)

Redundant Grid, Labels & Frame

Version 2 (Tufte)

Version 3 --- Curve Removed

Version 4 --- Grid Restored

Final Version (5) --- No Grid, Added Labels

Figure 2.11: Upper left: a figure drawn by Roger Hayward for *General Chemistry* by L. Pauling. Upper right: the graphical elements that can be erased from the graph without damaging its intellectual content. Remaining panels: four different versions of Pauling's graph. The lower right diagram is Tufte's favorite.

2.2 High Data Density

Another theme of Tufte's is that good graphics have high "data density". To calculate this quantity, note that the numbers to be plotted in a graphic can be written as an array of one or more dimensions.

Definition 3 (Date Density)

$$\textit{data density} = \frac{\textit{number of entries in data array}}{\textit{area of data graphic}}$$

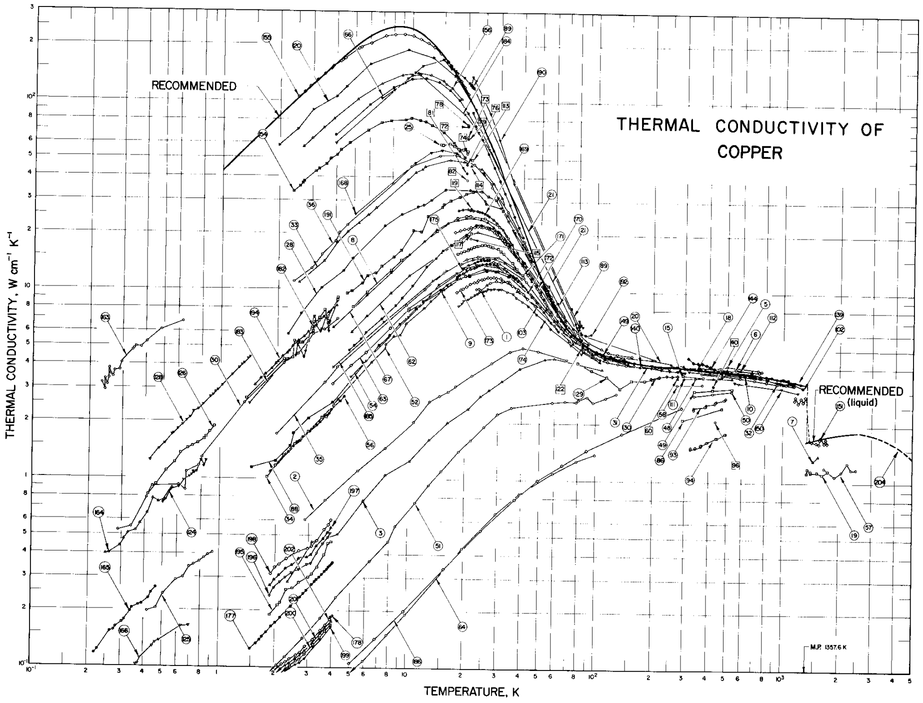


Figure 2.12: High data density graphic: illustration of many experimental measurements of the thermal conductivity of copper. Each string of dots, connected by a thin curve, is from a single publication which is identified by a number enclosed in a circle. Each of these 200-plus articles is listed with its identification number in the bibliography of the source of this figure, C. Y. Ho, R. W. Powell, P. E. Liley, "Thermal conductivity of the elements: A comprehensive review", supplement no. 1, *J. Phys. Chem. Reference Data, 3*, pg. 1-244 (1974). The thick solid line is the recommended curve.

Not withstanding this formal definition, the concept of a "high data density" graphic is perhaps best illustrated through some examples. Fig. 2.12 compares a large number of simultaneous measurements of thermal conductivity as a function of temperature. Because more than 200 different curves, each labelled, are combined into a single graph, Tufte has canonized this as a Good Example. Is it?

Well, yes, but only up to a point. First, there is no really good way to present 200 data sets. A single figure is certainly much easier to grasp than a whole of lot of figures, or a verbal description. The authors did as good a job as possible this vast amount of information.

However, it is extremely unlikely that anyone would actually correlate all two hundred-odd curves with the bibliographic references given many pages later. In that sense, most of the labels are useless for a given reader.

Furthermore, for high temperatures (greater than the 250^{o}K., say) the measurements are all superimposed on one another. The information presented in this range is really (i) a single curve plus (ii) a verbal description, graphically expressed by the article numbers, that all the measurements agree closely.

For low temperatures (below 80^{o}K.), many curves can be visually distinguished. However, these curves are not useful in and of themselves because most of them are gravely in error. The information conveyed in this range is: measurements of thermal conductivity at cryogenic temperatures is difficult and published estimates vary wildly. This sentence conveys the information almost as well as the graph.

The authors also provide a solid curve which represents the "recommended" curve. This immediately raises an important question: How can one "recommend" anything when the measurements disagree by as much a factor of one hundred? One purpose of a good graphic is to raise questions, and this does. However, the answer must come from the text.

This graph IS a good figure. However, in designing high density graphs, one must recognize that the reader is never going to be able to remember or even visually identify all the curves, all the data. The crucial design principle for high density graphs is not: Let's find a good way to cram lots of curves/points/arrows/labels in a small area. Rather, one should identify key THEMES or GOALS of the figure, and then design and edit so the graphs successfully presents these themes.

Tufte is practically beside him in praising Fig. 2.13. The sky was divided into more than two million rectangles, and then the 1.3 million galaxies of the Lick Catalog were binned into one of the rectangles. The density within each rectangle was represented by a grayscale. The map was very influential because it suggested that there might be very large-scale structures, chains and filaments, in addition to the clusters and superclusters that were previously known. Later work has extended the graphics into three dimensions and confirmed that on a very large scale, the universe is filled with a kind of "cosmic foam": bubbles of void surrounded by sheets with a relatively high density of galaxies.

Is this a good graphic? Yes, but high data density is only part of the reason. First, the figure is good because its scientific message is important. This is the first qualification for graphic greatness!

Second, the figure is good because it allows one to see new things in old data. The Lick Catalog had been expanding for many years before this 1974 illustration was extracted from it. Every one of the galaxies had been individually measured from a photographic plate. Nevertheless, filaments and clumps in the map were TERRA INCOGNITA[1] before this illustration was published.

However, high data density is only indirectly responsible for the usefulness of the graph. First, note that if the Lick Catalog had contained only half a million galaxies or as many ten million galaxies, the usefulness of the figure would not have been changed. The discovery of filaments is a qualitative property that cannot be altered by adding more galaxies. (However, the authors might have gained additional confidence in their discovery with a larger sample). Similarly, reducing the number of galaxies by two-thirds would have made the filaments fuzzier and less distinct, but not enough to have prevented the authors from noticing them.

Second, one can analyze a high density graphic from the perspective of the branch of engineering known as "Information Theory". This was created by Claude Shannon in the

[1] "Terra incognita" is Latin for "unknown lands".

1930's to analyze telecommunications system. The amount of bandwidth needed to transmit a message obviously depends on the information in the message, but real messages have noise and redundancy. Shannon was able to quantify these complications.

His definition of "information" is the number of computer bits needed to encode a message, sans redundancy. For graphics, it would perhaps be more meaningful to quantify the "information" as the number of words needed for a verbal description of the key themes of the graph or the number of brush strokes necessary for an artist to draw a copy of the graph that contains the key ideas.

By either a verbal or artistic measure, the information theory-content of the galaxy map is much smaller than two million grey tones. The verbal information is one sentence: "The galaxies in the cosmos are clumped into filaments". The artistic description is fuzzier. However, an artist could trace over the map and outline the clumps and filaments with considerably less than two million brushstrokes. Indeed, less than a thousand would suffice. The freehand sketch of the clumps and filaments (Fig. 2.14) took only a few minutes, and yet depicts the essential content of the density map. Indeed, it shows the filaments much more clearly than the grey tones of the figure it imitates.

A reduced-density graph could have been made more systematically by applying a filter; points above a certain threshold density of galaxies are black while points of lower density are white. Again, this would show the filaments more clearly.

The conclusion is that high data density is a virtue but should not be elevated above the more fundamental principles: clear labels, clear theme, and so on. It is very easy for a high density graphic to be fuzzy or hard to read. Indeed, perhaps the best way to present the important implications of the Lick Catalog would be to publish two graphs: a high density gray scale map to show the raw data with minimal editing, and a reduced density black-and-white graph to show the filaments and clumps more clearly.

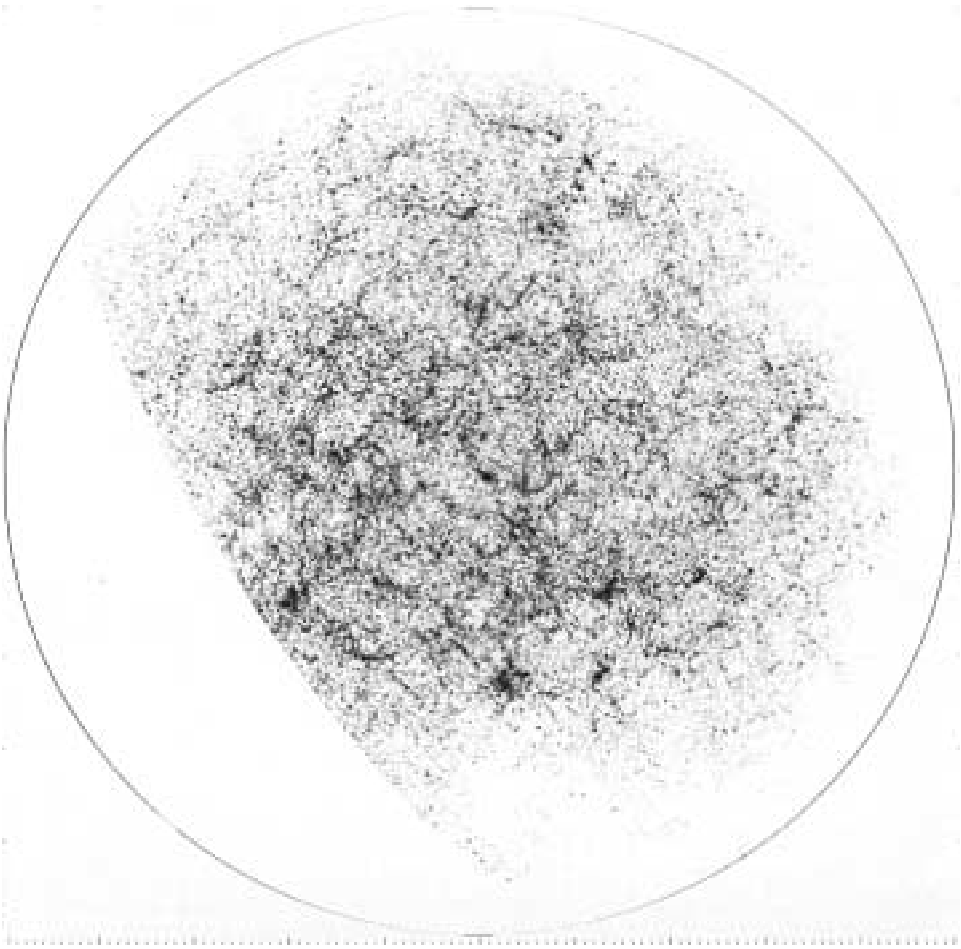


Figure 2.13: Density of galaxies. The sky was divided into more than two million rectangles. The number of galaxies in each rectangle is encoded by ten different gray tones. Although similar filaments appear even in randomly created data, the filamentary structure apparent in the map is real. Map created by Michael Seldner, B. H. Siebers, Edward J. Groth and P. James Peebles, "New reduction of the Lick Catalog of galaxies", *Astronomical Journal, 82*, pg. 249-314 (1974).

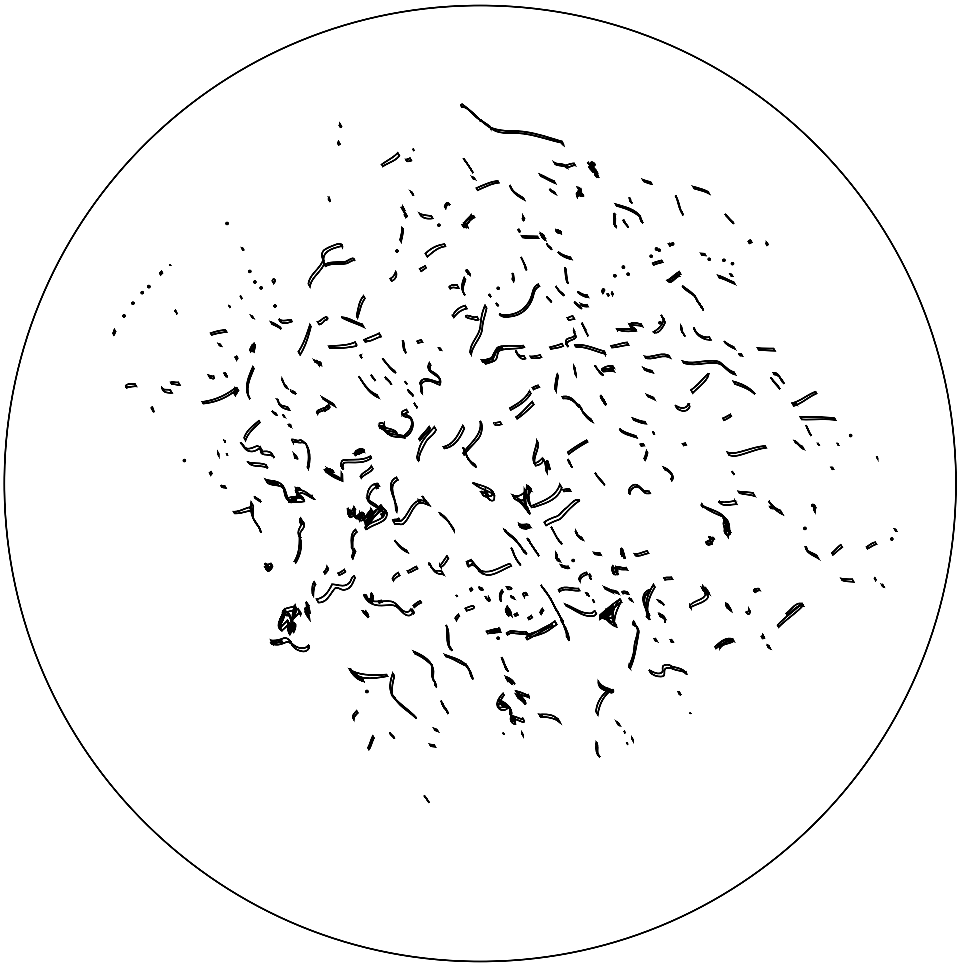


Figure 2.14: Sketch of the clumps and filaments in the previous figure of a galaxy density map, drawn in less than fifteen minutes in Adobe Illustrator

2.2.1 The Shrink Principle

Consistent with his theme that High-Data-Density-is-Good, Tufte advocates complicated multi-panel graphs that pack a lot of information into a single composite figure. This in turn is possible because of the following:

Definition 4 (Shrink Principle)
Graphics can be shrunk way down.
— Tufte(1983), pg. 169

Fig. 2.15 illustrates this principle. Six one-dimensional graphs and six two-dimensional graphs are combined into a single composite. The top and middle pairs of two-dimensional graphs are contour plots; the bottom pair are satellite photos. The one-dimensional graphs on the left are a mixture of line graphs (time series of continuously recorded data) and histograms (precipitation, which is normally described by accumulations rather than rates).

Several factors contribute to the success of this figure for its intended audience. First, this forecast was prepared as a special handout for attendees of an international meteorological conference, IAMAP '89. Because the readers are all professional meteorologists, extensive labels and elaborate explanatory captions were unnecessary, which helps enormously in fitting a dozen graphs on a single page. Second, the purpose of the figure was mostly public relations; it was prepared by the European Centre for Medium-Range Forecasting, which generates weather prognostications for an eighteen-nation consortium, to illustrate the work of the Centre for conference guests who were given a tour of the Centre's facilities. If the figure is "over-shrunk" or the labels too few, this is much less serious for public relations than for a research journal.

Second, the figure filled the entire 8.5" x 11" (including the margins). If reproduced in a digest-sized journal, the diagrams would have to be shrunk considerably with a corresponding loss of legibility. High density, heavily shrunk graphics must be designed to fit the limits of the publication medium.

Satellite photos and pressure contour plots are widely available on the Internet. However, no site attempts to pack twelve diagrams into a single screen. Instead, the photos or contour plots are presented one at a time. Composite figures are often used to organize THUMBNAIL sketches to present a kind of visual INDEX of the data. However, at the usual screen resolution of only 72 dots-per-inch, it is impossible to identify fronts and vortices from a thumbnail sketch an inch square. High density is good, but know your target!

Fig. 2.16, originally published in Bertin's classic *Semiology of Graphics*, is Tufte's epitome of a good high density graphic. Seventeen different plots of several different graphics species are combined into a single composite figure. The data curves, however, are clearly visible so that this graph is successful.

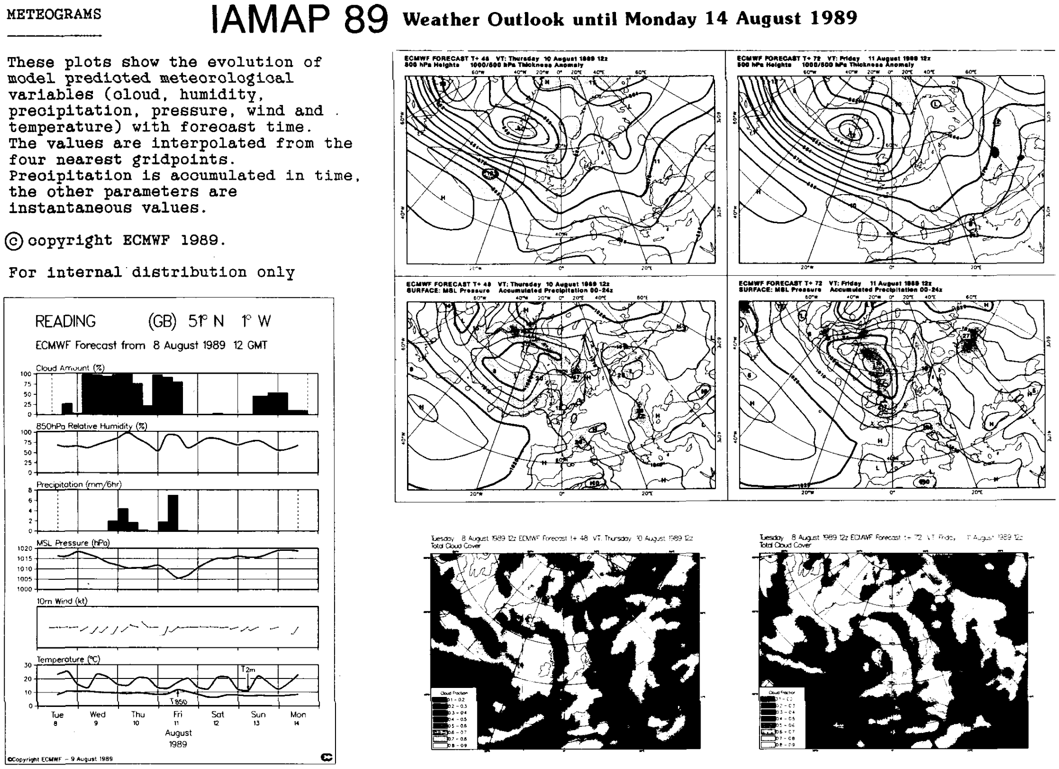
However, success must be defined carefully. At this density, almost all labels must be omitted from the graphs because there just isn't room. Full frames, and not merely half-frames, are helpful in visually distinguishing one plot from another.

Further, the graph is effective because none of the data curves are oscillating too wildly. Curves of complicated shape must be presented at a larger size than curves with but a single maximum and minimum and smooth, monotonic variations in between. The degree of allowable shrinking is data-dependent as well as resolution-dependent.

With these caveats, Tufte's principles of high data density, shrinking plots, and combining many graphs into a composite figure are all sensible. It is much easier for a reader to grasp a set of related figures if they are all one one page in a single multi-panel figure than if they are scattered over many pages.

In the days before computer graphics, cartoon animators would draw a series of sketches and then check their work by rapidly flipping the sketches so that they would flash past the eye quickly enough to produce the illusion of continuous motion, as in the final filmed cartoon. Far too many scientific papers and doctoral theses require their readers to be "Disney animators", treating the paper as a "flipbook".

What cannot be animated or converted into hypertext (as on a Web site) should be COMBINED wherever possible.



METEOGRAMS

IAMAP 89 Weather Outlook until Monday 14 August 1989

These plots show the evolution of model predicted meteorological variables (cloud, humidity, precipitation, pressure, wind and temperature) with forecast time. The values are interpolated from the four nearest gridpoints. Precipitation is accumulated in time, the other parameters are instantaneous values.

© copyright ECMWF 1989.

For internal distribution only

READING (GB) 51° N 1° W
ECMWF Forecast from 8 August 1989 12 GMT
Cloud Amount (%)
100
75
50
25
0
850hPa Relative Humidity (%)
100
75
50
25
0
Precipitation (mm/6hr)
8
6
4
2
0
MSL Pressure (hPa)
1020
1015
1010
1005
1000
10m Wind (kt)
Temperature (°C)
30
20
10
0
T2m
T850
Tue 8
Wed 9
Thu 10
Fri 11
Sat 12
Sun 13
Mon 14
August
1989
©Copyright ECMWF – 9 August 1989
ECMWF FORECAST T+ 48 VT: Thursday 10 August 1989 12z
500 hPa Heights 1000/500 hPa Thickness Anomaly
ECMWF FORECAST T+ 72 VT: Friday 11 August 1989 12z
500 hPa Heights 1000/500 hPa Thickness Anomaly
ECMWF FORECAST T+ 48 VT: Thursday 10 August 1989 12z
SURFACE: MSL Pressure Accumulated Precipitation 00-24z
ECMWF FORECAST T+ 72 VT: Friday 11 August 1989 12z
SURFACE: MSL Pressure Accumulated Precipitation 00-24z
60°W
40°W
20°W
0°
20°E
40°E
60°E
H
L
Tuesday 8 August 1989 12z ECMWF Forecast t+ 48 VT: Thursday 10 August 1989 12z
Total Cloud Cover
Tuesday 8 August 1989 12z ECMWF Forecast t+ 72 VT: Friday 11 August 1989 12z
Total Cloud Cover
Cloud Fraction

Figure 2.15: A "meteogram", created by the European Centre for Medium-Range Forecasting in Bracknell, England, for distributions to attendees of an international conference (IAMAP) in the adjoining town of Reading, England. Each of the four contour plots is a snapshot of two different weather variables, represented by isolines and by shading, respectively. The upper two show 500 mb geopotential height and 500-1000 mb thickness anomalies; the lower pair show surface pressure and accumulated precipation for the preceding 24 hours. The two photos at lower right show cloud thickness as observed by satellite. The line graphs and histograms on the left depict a variety of variables at the conference site of Reading.

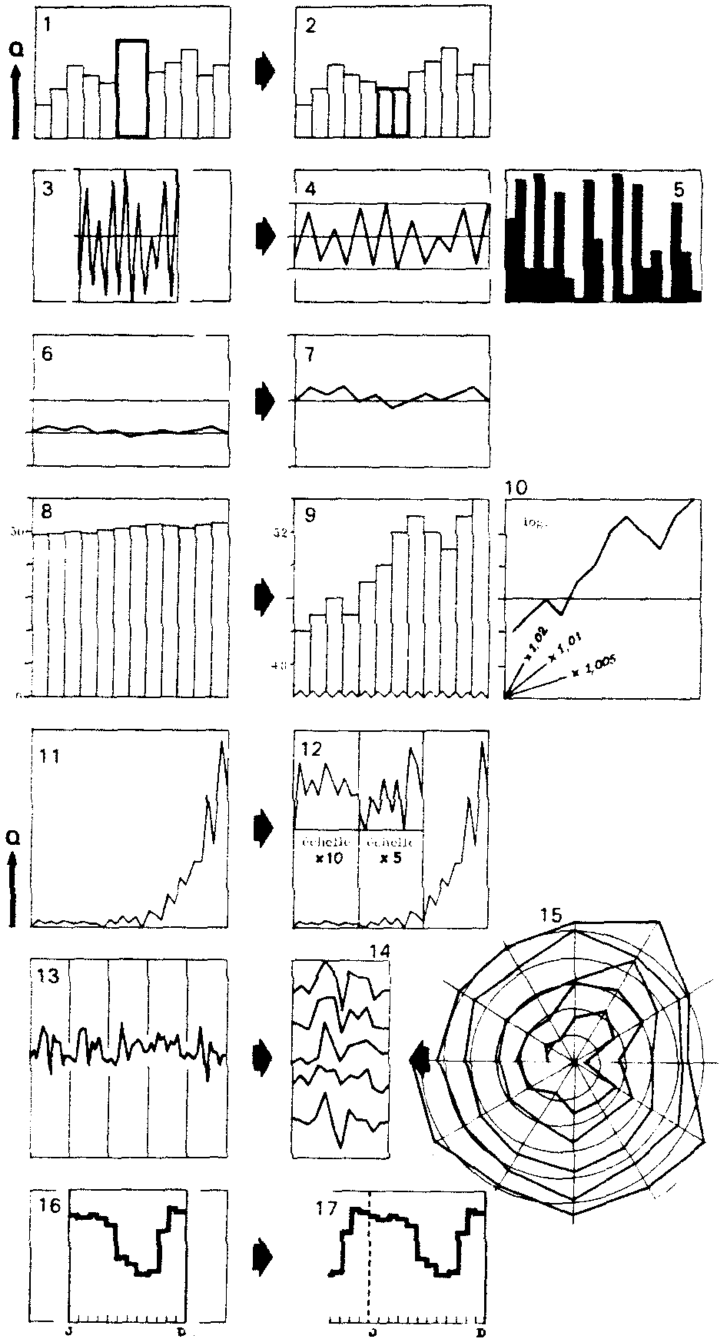


Figure 2.16: A seventeen-panel figure, originally from J. Bertin's *Semiology of Graphics.*

2.3 Multifunctioning Graphical Elements

"Mobilize every graphical element, perhaps several times over, to show the data."
Tufte(1983), pg. 139

This good advice is not easily implemented, but on those rare occasions when it can, the results can be very useful to the reader. The best way to illustrate this theme is through concrete examples.

The first example is Table 2.2, which is a table that also functions as a graphic. The reason that this table is effective for its intended audience of military specialists is that in the American army, the basic unit is the *division*, and each division is identified by a unique number. The 82d and 101st divisions, for example, became elite paratroop units that parachuted into Normandy on D-day. Later, these units converted to helicopters and became "air cavalry", playing an important role in Vietnam. To a military man or woman, the division number conveys valuable information.

Because this table has been designed so cleverly, one can almost instantly deduce three pieces of information:

- The date each division came to France
- The number of columns in France each month (by summing the nonzero entries in a column)
- The duration of each division's stay in France (by summing the nonzero entries in each row)

Table 2.2: The American Build-up in World War I

Table-graphic by Col. Leonard P. Ayres for his book *The War with Germany*, (1919), pg. 102. Each horizontal column indicates the months that a particular army division, denoted by its unique identification number, was present in France.

																1
																38
																31
															34	34
															86	86
															84	84
															87	87
														40	40	40
														39	39	39
														88	88	88
														81	81	18
														7	7	7
														65	65	65
													36	36	36	36
													91	91	91	91
													79	79	79	79
													76	76	76	76
												29	29	29	29	29
												37	37	37	37	37
												90	90	90	90	90
												92	92	92	92	92
												89	89	89	89	89
												83	83	83	83	83
												78	78	78	78	78
											80	80	80	80	80	80
											30	30	30	30	30	30
											33	33	33	33	33	33
											6	6	6	6	6	6
											27	27	27	27	27	27
											4	4	4	4	4	4
											28	28	28	28	28	28
											35	35	35	35	35	35
											82	82	82	82	82	82
										77	77	77	77	77	77	77
									3	3	3	3	3	3	3	3
									5	5	5	5	5	5	5	5
								32	32	32	32	32	32	32	32	32
						41	41	41	41	41	41	41	41	41	41	41
					42	42	42	42	42	42	42	42	42	42	42	42
			26	26	26	26	26	26	26	26	26	26	26	26	26	26
		2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
			1917								1918					

Fig. 2.17 is sort of the converse of Ayres: instead of a table that functions also as a graph, this drawing is a graph that functions also as a table. The thick vertical lines are the bars of a standard bar chart. The artist wanted to label each bar with its numerical value so that the graph would provide the same precise numbers as a table. Unfortunately, there are too many bars to allow direct labelling of each bar. Instead, the artist added thin horizontal lines to attach the top of each bar to the corresponding label. The labels can thus be stacked vertically, and contain as many digits or letters as necessary without crowding; the label stack functions as a table. The horizontal lines themselves have heights equal to the heights of the bars and thus carry data-information themselves (although this is redundant to the heights of the bars, which provide the same visual sense of magnitude).

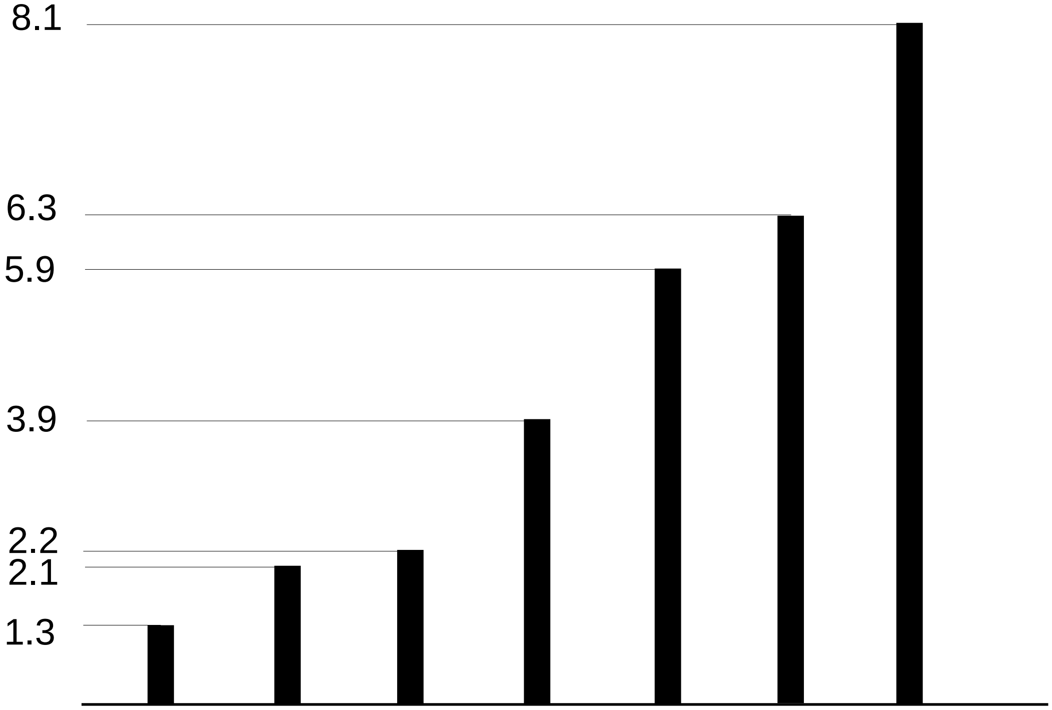


Figure 2.17: A graph that is also a table in the sense that the numerical heights of each bar are tabulated in a column on the left. An imitation of one by Carol Moore for a corporate annual report and reprinted in Tufte(1983), pg. 152 (bottom), and also in Walter Herdeg, *Graphic/Diagrams*, Zurich (1976), pg. 23.

Ayres' and Moore's table-as-graph, graph-as-table inspired Tufte to try his hand at redesigning the flight-of-steps graph which is common in statistics. Fig. 2.18 shows a standard step chart on the left. On, the right, the same graph is presented in the style of Tufte's curve-free graph on pg. 151 of his 1983 book. The integers are now "multifunctioning" in the sense that these not only label different steps, but their height on the graph conveys the same numerical information as the height of the steps does in the standard flight-of-stairs chart. The precise numerical values of each step are vertically stacked along the vertical axis. Thus, this is a graph that is a table, too.

Multifunctioning graphs are always clever, but this isn't the same as being effective. The floating pattern of integers in Tufte's redesign does not give as good a visual sense of the steps as the standard flight-of-stairs graph. The figure would look less strange if the stairsteps were added, perhaps as a thin dotted curve in back of the digits.

Similarly, Moore's style of bar chart contains horizontal lines that are redundant. Furthermore, the eye must follow the thin horizontals rightward, then the thick bars downward to find the bar labels on the horizontal axis, and so connect them with the corresponding number. (The horizontal labels are omitted from our schematic, but would be included as essential information on any real graph.) In a table, the numbers and the alphanumerical labels would be side-by-side, which makes it much easier to determine that 2.3 was the number of barrels in millions sold by the Schlitz company this past year (or whatever).

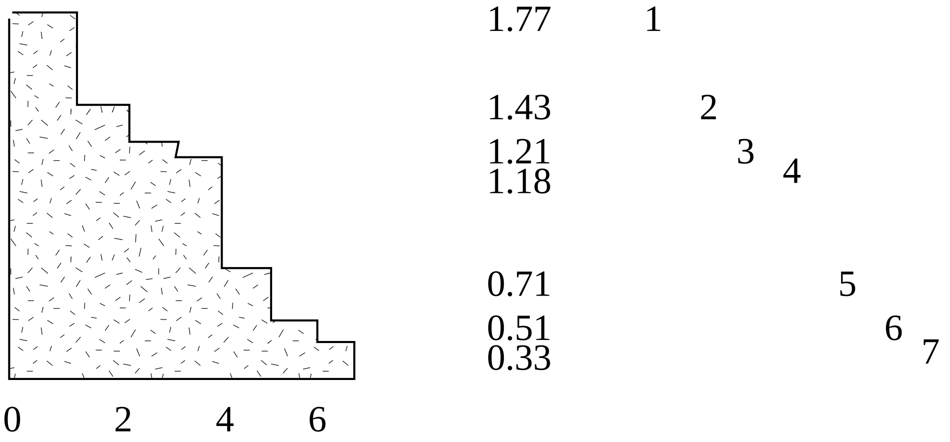


Figure 2.18: Left: a standard step chart, similar to the exemplar on the style sheet of the Journal of the American Statistical Association. Right: The same plot in Tufte's graph-into-numbers style.

Another Tufte example is a comparison of taxes as a percentage of gross domestic product for two different years for a variety of countries, printed on pg. 158 of Tufte(1983). Similar graphs are common in quantum chemistry and atomic physics where they are known as "level-crossing" diagrams. Fig. 2.19 is a schematic example. The reason for the name "level-crossing" is that the perturbation, whatever it may be, can change the ordering of the atomic states so that the third lowest energy state, "2p", becomes the second lowest when the perturbation is at full strength. The numerical labels make this graph function as a table. However, the visual elements — the connecting lines between the two stacks of labels and numbers — are important, too, because they allow the reader to instantly grasp which states have switched relative positions when energy from lowest energies to highest.

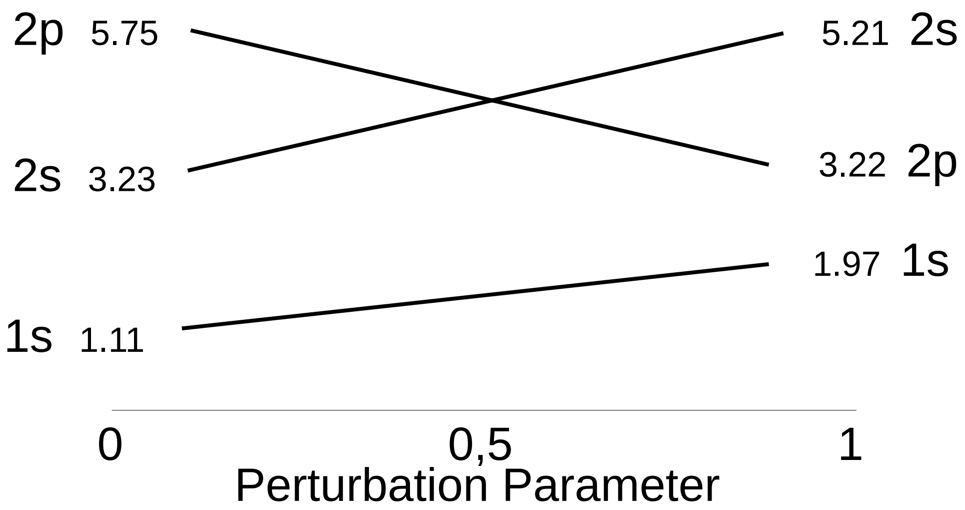


Figure 2.19: A schematic "level-crossing" diagram from physics. Left: atomic states (by name) and energy levels for the unperturbed atom. Right: the same when the perturbation parameter is equal to one. The lines show how the eigenvalues [energy levels] of the Schroedinger equation vary with the strength of the perturbation parameter.

2.4 Small Multiples or Animations-on-a-Page

Even though he devotes at least one chapter to "multiples" in each of his three books, Tufte never precisely defines the concept of "small multiples". Indeed, he eventually abandoned the adjective "small" in his latest (1997) book where a chapter is entitled "Multiples in Space and Time". The closest he comes to a definition is the following from *Envisioning Information*, pg. 67:

> "Illustrations of postage-stamp size are indexed by category or a label, sequenced over time like the frames of a movie, or ordered by a quantitative variable not used in the single image. Information slices are positioned within the eyespan, so that viewers make comparisons at a glance — uninterrupted visual reasoning. Constancy of design puts the emphasis on changes in data, not changes in data frame."

In short, a "small multiple" is really an animation-on-a-page. The trick which underlies animation is that when a sequence of discrete images is shown rapidly, the mind can be fooled into perceiving continuous motion. However, a necessary condition for this illusion is that the individual frames must differ by a SMALL amount.

A "small multiple" figure is a collection of miniature illustrations, arrayed as a single figure, which are designed to be perceived as one. Just like a true animation, the frames must differ by only a small amount from one another, or the illusion is spoiled and each frame is perceived as an individual rather than as part of the whole. This requires that each graph in such a multipanel figure must be the same size, same graph species, and same all other aspects of the design. Only the data itself, and perhaps the number which indicates the time or time-like parameter that orders the sequence, can differ from one miniature graph to the next.

The necessity for small differences from one frame to the next so that the mind can combine a dozen graphs into a single mental filmstrip is the reason for Tufte's modifier "small". "Multiple" means multiple images of the same quantity for different times or parameter values, apparently.

Still, "multiples of time and space" and "small multiples" seems remarkably sloppy terminology for a proponent of graphical precision. "Animation-on-a-page", "flipbook-on-a-page", and "quantitative storyboard" are more descriptive. And yet these labels have problems, too.

The difficulty is that terms like "animation", "flipbook" or "storyboard" are all borrowed from the jargon of movies and television. A multi-panel figure has its own rules for good design, and these are different from those for *The Simpsons* or *Star Wars*.

Fig. 2.20 shows a composite graph which is rather close to an animation: the frames are merely snapshots from the time evolution of a plasma. Even here, however, there are differences from the cinema. First, a movie film animates at a rate of twenty-four frames per second. To fit the graphs into a single figure, this "animation-on-a-page" shows only six different times.

Second, to make it easier to understand the flow, each time is illustrated in two different ways: once as a mesh plot and the other as a contour graph. Steven Spielberg has used a split screen occasionally, no doubt, but this breaks the illusion of continuous motion; one has to consciously study both representations.

Fig. 2.21 is a step further from cinema. First, there are only four frames. Second, the ordering parameter is not time, but rather the width of the hyperbolic secant function. There is still a sense of continuous variation from a narrow graph to a wide graph, and these four panels are meant to be perceived as a single, coherent message.

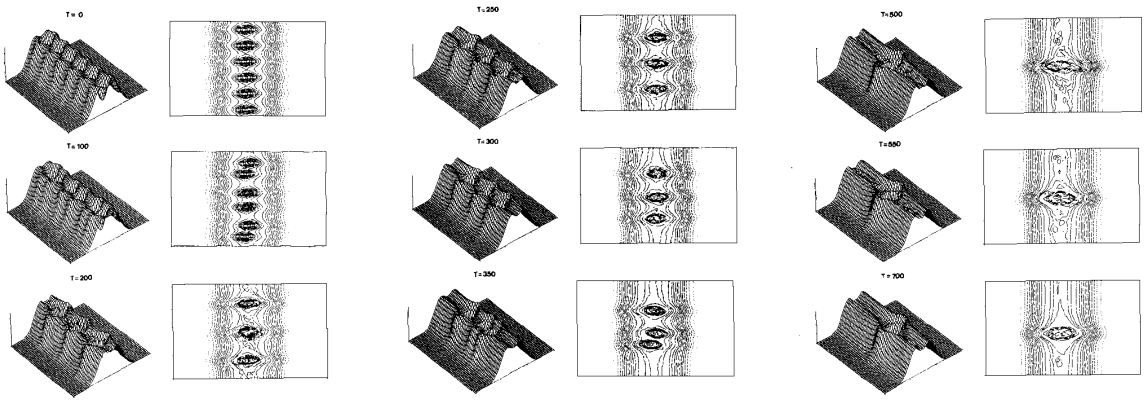


Figure 2.20: Plasma fields over time, beginning in the upper left. Each field is shown as both a three-dimensional mesh diagram and also, immediately to its right, as contour plot. From Ghizzo *et al., Phys. Fluids, 31*, 72-82 (1988).

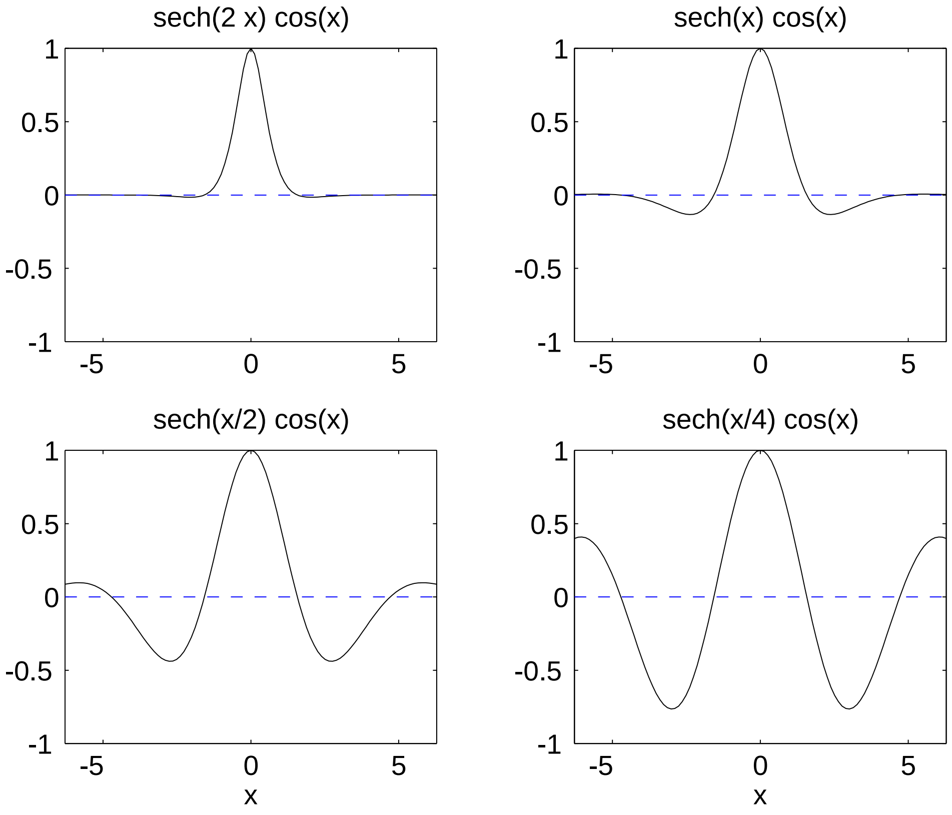


Figure 2.21: An illustration of the principle of "small multiples". Each of the four panels is IDENTICAL except for changes in a single parameter: the width of the hyperbolic secant ("sech") function.

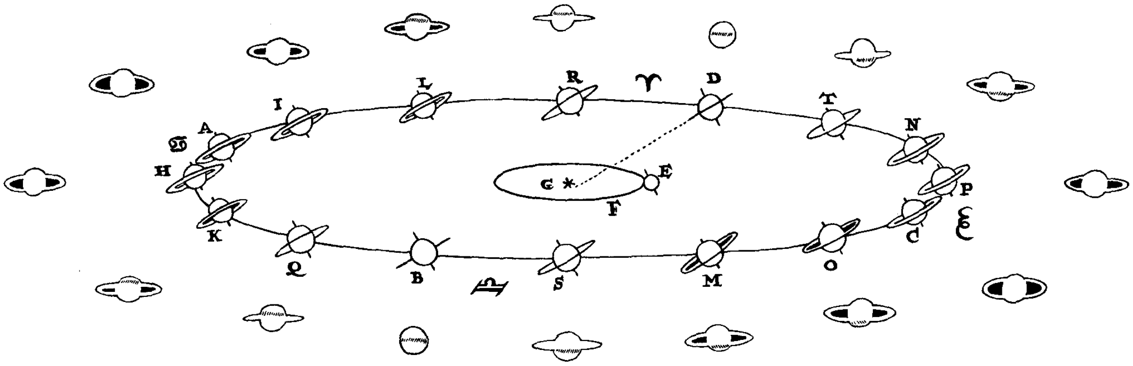


Figure 2.22: Saturn and its rings as viewed through telescope from various positions in its orbits. Christian Huygens, *Systema Saturnium*, (The Hague, 1659),pg. 55.

Fig. 2.22 was drawn by Christian Huyghens centuries before there were movies. Nevertheless, one could easily imagine morphing the images into a Quicktime movie to follow the visual appearance of Saturn's rings through one complete Saturnian year. However, Huyghens had a deeper purpose. Through a telescope, with the image jiggled continuously by small-scale turbulence in the atmosphere, Saturn was drawn differently by many different observers, depending both on individual interpretation and also when in the planet's 29-year orbit it was seen by a particular astronomer. Huyghens' small multiple figure shows that these different images can be explained by a single model: a planet with rings tilted at an angle to the earth. He wants the reader to linger over the individual frames, recognizing Galileo's model and Brahe's image, and then to gather the whole sequence into a single theory. The animation of the Saturnian system is only half the story, and a Quicktime movie of Huyghen's frames would miss a major portion of his message.

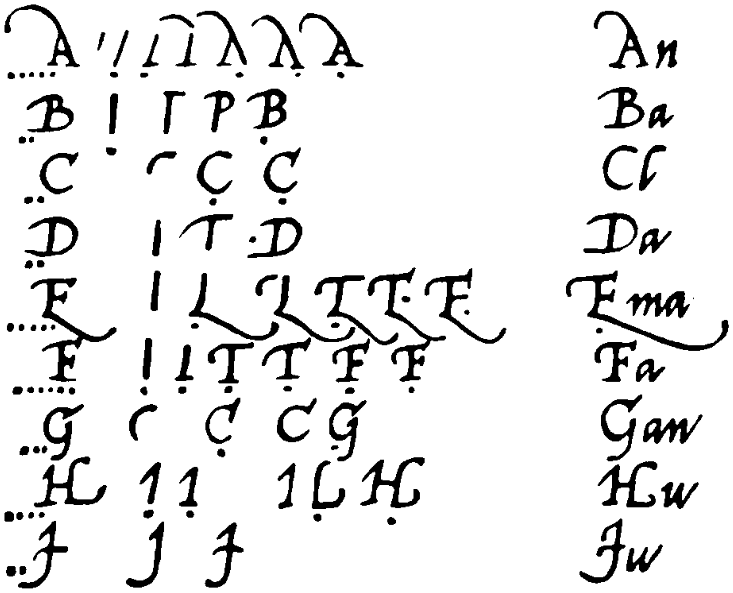


Figure 2.23: Sequences illustrating the strokes in drawing capital letters. Gerardus Mercator, *Literarum Latinarum, quas Italicas cursoriasque vocant, scribendarum ration [The Method of Writing the Latin Letters, Which are Called Italic and Cursive]* (Louvain, 1540), chapter 6.

Our final example would make a good storyboard for a training video on penmanship, although it was drawn by the famous mapmaker Mercator more than four centuries before Philo Farnsworth, Fig. 2.23. One critical point is that the figure illustrates the individual strokes for drawing a letter. Because each letter requires several strokes, the instruction for a given letter cannot be animated by flowing smoothly from one image to another. Instead, a real training video would likely stop-and-go: a burst of smooth video to illustrate the first stroke, then a few seconds in which the completed stroke is frozen on the screen while the narrator explains how the pen must be tilted for the next stroke, then an animation of the second stroke followed by a few seconds when the partially completed letter is again motionless, and so on.

In addition, because the sequence for a given letter occupies only a small amount of space, the nine rows of the figure each illustrate a different letter. In this sense, Mercator's figure is not an animation but rather nine separate brief animations.

Any sequence of frames that are "multiple" with "small" variations from frame to frame can be animated. If there are too few frames to make a smooth movie by themselves, one can *interpolate* between the frames — in computer graphics, this interpolation is called "morphing" — until the animation is pleasing to the eye. In principle, all "animation-on-a-page" or "small multiple" graphs can be converted into actual Quicktime or MPEG movies.

However, the symbolic equation

$$\text{“small multiple” graph} \leftrightarrow \text{one animation with morping between frames} \qquad (2.3)$$

is quite inadequate to describe the diversity of multipanel figures with slow frame-to-frame variations. Because of this diversity, there isn't really a good term for this class of graphs.

Nevertheless, the ability to condense many small graphs into a single mental picture is as important in science as in cinema.

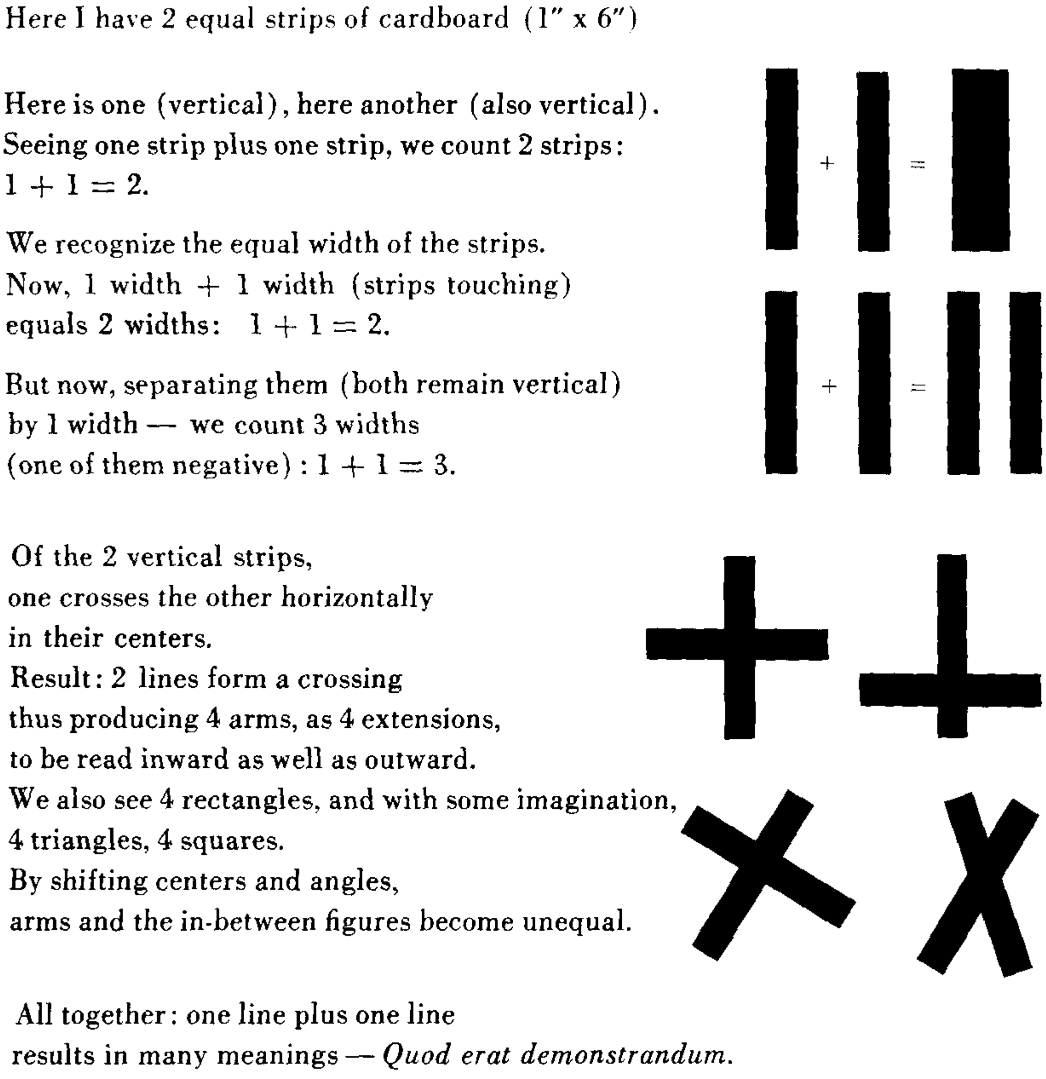
2.5 One Plus One Is Three

Josef Albers wrote a famous essay, much admired and quoted by Tufte, with the provocative title, *"One Plus One Equals Three or More: Factual Facts and Actual Facts"*. He illustrates his message that in graphics

$$1 + 1 = 3 \tag{2.4}$$

What he means by this zen-like *koan* is that elements of a graph interact with one another. Because of this, a good graph is more than a simple addition of its various pieces.

In designing a graph, the first question is: What to include? There is, however, a second phase, which is to ask: How does these parts interact? What should be emphasized with a thick line? What should be deemphasized with a thin line? Can the data curves be easily picked out from the frame, labels and other elements?



Here I have 2 equal strips of cardboard (1″ x 6″)

Here is one (vertical), here another (also vertical).
Seeing one strip plus one strip, we count 2 strips:
1 + 1 = 2.

We recognize the equal width of the strips.
Now, 1 width + 1 width (strips touching)
equals 2 widths: 1 + 1 = 2.

But now, separating them (both remain vertical)
by 1 width — we count 3 widths
(one of them negative) : 1 + 1 = 3.

Of the 2 vertical strips,
one crosses the other horizontally
in their centers.
Result: 2 lines form a crossing
thus producing 4 arms, as 4 extensions,
to be read inward as well as outward.
We also see 4 rectangles, and with some imagination,
4 triangles, 4 squares.
By shifting centers and angles,
arms and the in-between figures become unequal.

All together: one line plus one line
results in many meanings — *Quod erat demonstrandum.*

Figure 2.24: Taken from Albers' book, *Search Versus Re-Search* (Hartford, 1969), pp. 17–18.

2.6 Layering, Separation and Rubrication

Cartoon animation is all about layers. Before 1920, a few pioneering cartoons were produced by laboriously drawing each figure from scratch. Then a new idea reduced reduced costs enormously. The background was painted in as before, but the characters were painted onto thin sheets of transparent celluloid. Each sheet or "cell" was laid in place over the background painting and photographed. To further reduce costs, later low budget animators used one cell for a character's body and then overlaid additional cells with mouth, eyes and eyebrows and so on. Only the mouth cell is changed while the character speaks. A collection of stock cells showing a character's mouth in various positions enables a clip of a few seconds to be filmed with only two drawings: one for the background and one for the body.

Drawing programs such as Adobe Illustrator allow a figure to be separated into *layers*. The reason is that like cartoons, complex illustrations subdivide into groups which are closely related, but logically distinct from the other groups of the figure. It is just as convenient for the illustrator to be able to manipulate background, body, and face separately as it is for the animator.

Similarly, most scientific and engineering figures are overlays of several layers that are logically distinct, even though they may appear as a single image in the final printed product. The frame and tick marks are one logical layer; the axis labels and title are a second, text layer; the data curves are a third layer. It is very advantageous to recognize that an illustration can be decomposed into layers because these layers are of varying degrees of importance. Much of Tufte's philosophy can be summed up as: emphasize the data curve layer, and mute the other layers.

(Parenthetically, note that this layering occurs in the computer representation of a graph in an object-oriented software system such as Matlab through the concept of curves, labels, and so on as "children" of a set of axes, which in turn are children of the figure.)

Layering implies "separation" in the sense that different layers should be separated by space, linestyle, color, type font and other properties — insofar as practical — so that groups of elements which are *logically* distinct are also *graphically* distinct.

This principle is easier to state than to do. For example, Fig. 2.25 shows two versions of a railway table with two different strategies for the "separation" of different rows. The top figure employs the not-very-good solution of using heavy grid lines, often called "rules" in graphic design, for "separation".

The difficulties with this were stated more than sixty years ago by a well-known graphics designer:

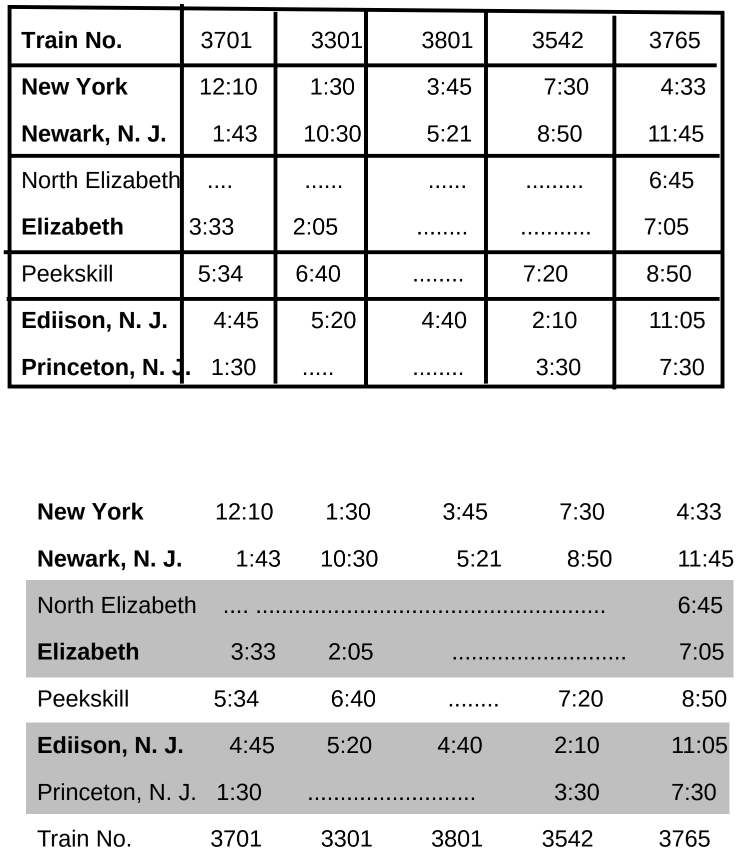
"The setting of tables, often approached with gloom, may with careful thought be turned into work of great pleasure. First, try to do without rules altogether. They should be used only when they are absolutely necessary. Vertical rules are needed only when the space between columns is so narrow that mistakes will occur in reading without rules. Tables without vertical rules look better; thin rules are better than thick ones."

Jan Tschichold, *Asymmetric Topography*, Basel, 1935), pg. 62.

The lower version of the table has dumped all the horizontal and vertical rules. Instead, groups of related trains are visually linked through grey-shading or the absence of same.

The original table was badly ordered because the train number, used only by railroad employees, was assigned prominence-of-place by being the first row. It was moved to the bottom in the revised table.

In addition, the little rows of dots in each blank cell have been replaced by long continuous lines of dots, which guide the eyes to the next number.



Train No.	3701	3301	3801	3542	3765
New York	12:10	1:30	3:45	7:30	4:33
Newark, N. J.	1:43	10:30	5:21	8:50	11:45
North Elizabeth	....	......	......	....	6:45
Elizabeth	3:33	2:05	....	....	7:05
Peekskill	5:34	6:40	....	7:20	8:50
Ediison, N. J.	4:45	5:20	4:40	2:10	11:05
Princeton, N. J.	1:30	.....	....	3:30	7:30

New York	12:10	1:30	3:45	7:30	4:33
Newark, N. J.	1:43	10:30	5:21	8:50	11:45
North Elizabeth	....				6:45
Elizabeth	3:33	2:05	....		7:05
Peekskill	5:34	6:40	....	7:20	8:50
Ediison, N. J.	4:45	5:20	4:40	2:10	11:05
Princeton, N. J.	1:30	....		3:30	7:30
Train No.	3701	3301	3801	3542	3765

Figure 2.25: Two versions of a railway timetable with the "bad" version on top. Inspired by an example from pgs. 54-55 of Tufte(1990).

Another illustration from Tufte's chapter on "Layer and Illustration" is imitated as Fig. 2.26. It is desirable to visually isolate the warning from the rest of the text, which is advertising puffery, copyright warnings and so on. A good graphic designer always emphasizes that which is likely to kill his readers! However, the black box in the top figure almost overwhelms the text. Tufte likes the much more muted frame in the lower figure. His philosophy is:

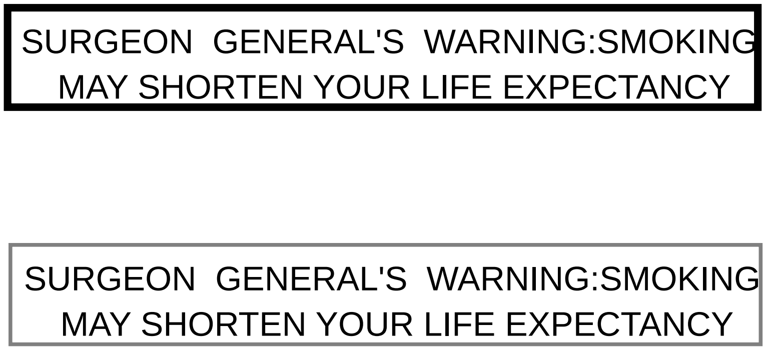
"For background elements, grey is better."

(Better than black, that is.) But is it? Most smokers don't read the text on the cigarette package. One could make a case that the black box is better because it is so striking that a veteran smoker is likely to notice the box and be reminded that the smoking is hazardous without reading the familiar text inside the box.

There are other familiar examples where a strong background element is okay. At an intersection, the red octagon moves the foot from accelerator to brake even before the eyes have read the word "STOP" in the middle. For most scientific graphs, though, the data curve is unfamiliar, and a muted background is better.

Color and greyscale are both powerful tools for separation. To a printer, "separation" actually refers to the generation of four separate images of a figure, one for each of the primary colors, which will then be used to make a separate photolithography plate. The page is printed four times, each time with a different plate and different color of ink, to receive the full-color illustration. This layering-by-color which happens in printing can also be reversed; one can conceptually layer the figure first and then assign separate colors to each layer. (The conceptual layers will then have a physical manifestation as the separate lithography plates.)

The only disadvantage of separation-by-color is that it is EXPENSIVE — four times as many photolithography plates, four times as many impressions and a very costly, multi-pass printing press. If one can afford it, though, separation-by-color is very valuable.



SURGEON GENERAL'S WARNING:SMOKING
MAY SHORTEN YOUR LIFE EXPECTANCY

SURGEON GENERAL'S WARNING:SMOKING
MAY SHORTEN YOUR LIFE EXPECTANCY

Figure 2.26: Two versions of the mandatory warning label on cigarettes. The grey box allows the message to be read more easily while still setting the warning off from the rest of the text on the cigarette package (not shown). Imitation of Tufte (1990), pg. 62.

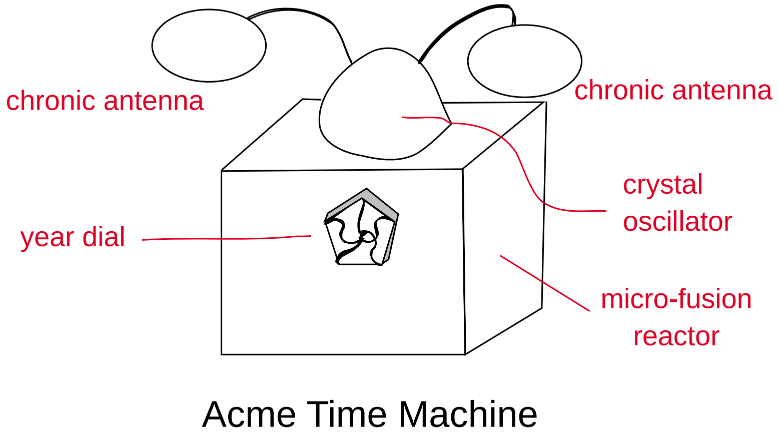


Figure 2.27: A whimsical illustration of an imaginary technology; red is used to separate the labels from the parts themselves.

Indeed, the principle of emphasis-and-logical-separation-by-color is so basic that it was in widespread use long before the invention of printing. During the Middle Ages, church missals, which contain the prayers recited during various services of the Christian Church, had the main text in black ink. The stage directions — "bishop turns and blesses the congregation", "acolyte rings the bells three times " — were written in red ink. This made it impossible, even in a smoky, candle-lit church, for a priest to confuse a prayer with a stage-direction. Red ink was also routinely used for emphasis in other illuminated manuscripts of all kinds.

Indeed, the use of red was so ubiquitous that it added words to the language. The verb "rubricate" means "to add stage-directions or marginal comments". (The word is derived from the Latin for "red".) Churchmen became so familiar with seeing liturgical directions in red that these became known as "rubrics", and are still so known in the modern Catholic Church.

Multi-colored printing followed Gutenberg's work rather early. The first two-colored book in English was printed by a man who is known to us only by his job description as "The Schoolmaster of St. Alban's". In this small, provincial town, this unknown teacher supervised the printing of a small volume for his students that had part of the text in red, part in black, to make it easier for his students to learn. Half a millenium later, separation-by-color is still a valuable tool for learning and teaching both in texts and graphs.

Tufte (1990) reprints a splendid modern example on his pg. 54: an exploded view of an IBM copier drawn by Gary E. Graham for the parts manual. The components are in black, the numerical labels and the lines from label to part are in muted red. The diagram is too detailed to scan and reprint here; the vast number of parts was why the "rubrication" of part numbers was so helpful to the reader, who was able to see all the parts, and their relationship to one another in the copier, in a single figure that simultaneously allowed him to look up the parts in the numerical key. Fig. 2.27 is a less cluttered example of a "rubricated" diagram.

When color is not an option, greyscale shading may be satisfactory. However, the shading needs to be *light* so that it will not drown the text and data curves even if the grey prints darker than it appeared on the computer screen.

"Separation" may also be physical proximity. Fig. 2.28 shows good and bad ways to label a curve. The most fundamental mistake is to put the label so far from the curve that the reader becomes confused about what element the label is supposed to identify. The middle two boxes are mildly deprecated. It is more difficult and slower to read curve-following text than horizontal letters. We are all used to the convention that labels go *above* the thing which is labelled — note that the title of a graph always goes on top, for example — and we should follow this convention wherever possible. The middle two panels have been identified as only "semi-bad" because when a graph is crowded with labels, one may be forced to make the text follow the curve so that the text will be unamiguously associated with the curve; similarly, one sometimes must put the label below the curve because there is simply no room to fit it above. The rightmost option is best. The goal of a label is be as LITTLE SEPARATE as possible from the curve it identifies because label-and-curve are really a single, logical element.

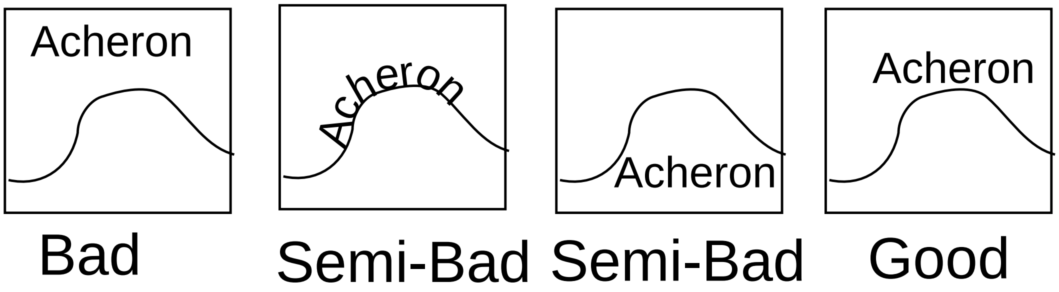


Figure 2.28: Good and bad line labels. The really bad example at far left is too far from the line it labels. The example that is second from the left is clear, but the curve-following type is harder to reader than a horizontal word. The label in the next box is *below* the line it labels; this is harder to read, and easier to misread, than a label which is *above* the curve. The rightmost example is good because it is close, horizontal, and above the curve it labels. Imitation of a figure by Eduard Imhof, reprinted in Tufte(1990), pg. 62.

2.7 Word-Labels Are Better Than Letter-Labels

Another general theme emphasized by Tufte and others: make labels as clear and explicit on the graph itself. For very complicated figures, it may be necessary to use a legend box or to provide a verbal key to the lines (solid: gold, dashed: silver, dotted: brass, etc.) in the caption. However, as much as possible, one write out labels as whole words or numbers.

Fig. 2.29 shows two versions of the same diagram. The left illustration is incomprehensible as it stands because the parts of the ear have been labelled with letters which are meaningless in the absence of a key or letter box. The figure on the right conveys much more meaning in the graph itself because the labels are the actual names of the parts. The left version of the graph forces "tennis spectator syndrome" on the reader, forcing the eyes to go back-and-forth, back-and-forth between caption and drawing.

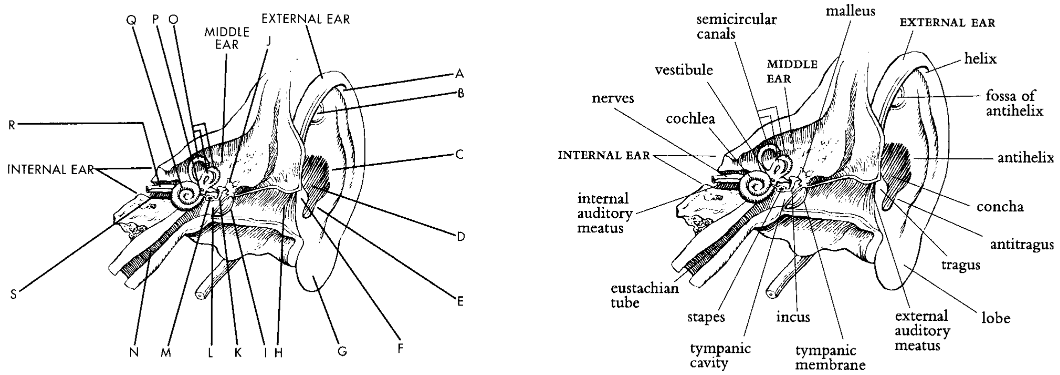


Figure 2.29: Letter-labels (left), which can be only interpreted by turning from the graph to a key in the caption or a legend box, are inferior to word-labels (right), which can be understood without taking one's eyes off the figure.

2.8 Collapsing a Dimension or Escaping Flatland

In elementary particle physics, string theory asserts that the three-dimensionality of the universe is a mirage. Rather, there are at least ten dimensions, but the extra dimensions are “compactified”, that is, curled up on a microscopic scale so as to have no visible role. Sometimes a similar compactification or collapsing of dimensions is equally fundamental to good graphics on two-dimensional paper.

Fig. 2.30 employs a convention popular in biological illustration: one coordinate, time, is overlaid on a spatial coordinate so that left-to-right in space also conveys a sense of time. The vertical lines, each labeled with a month, provide a minimal framing that makes this diagram a multipanel animation-on-a-page. However, the framing is so muted that the diagram is also perceived by the eye as a single figure rather than a composite. Taken this way, as a whole, this diagram of the life-cyle of the Japanese beetle is physically unrealizable as a portrait of an adult human talking to himself as a small child. And yet, because the passing months are clearly labeled at the top, this life-cycle diagram is not confusing. The conversion of time to space, graphically speaking, has made it possible to mute the frame to vertical dividing lines only.

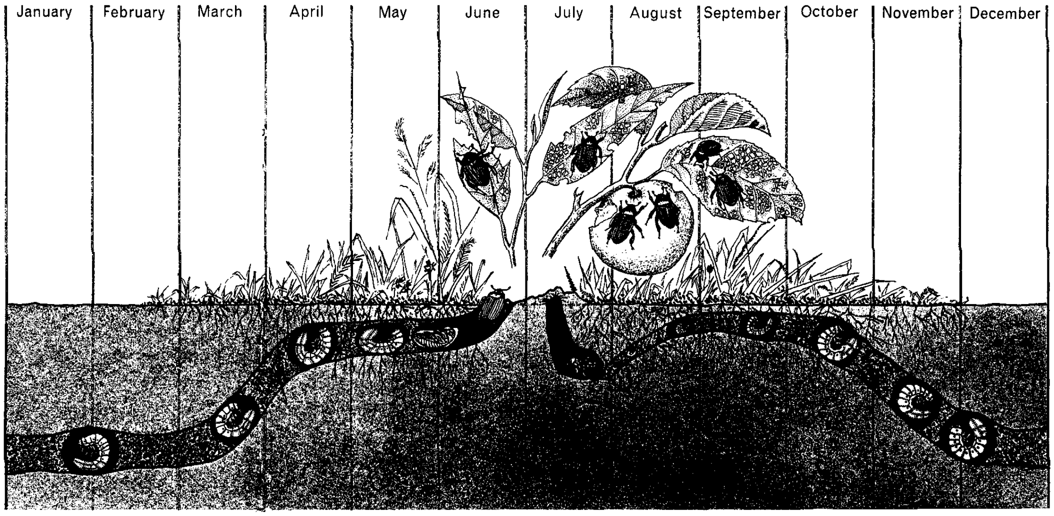
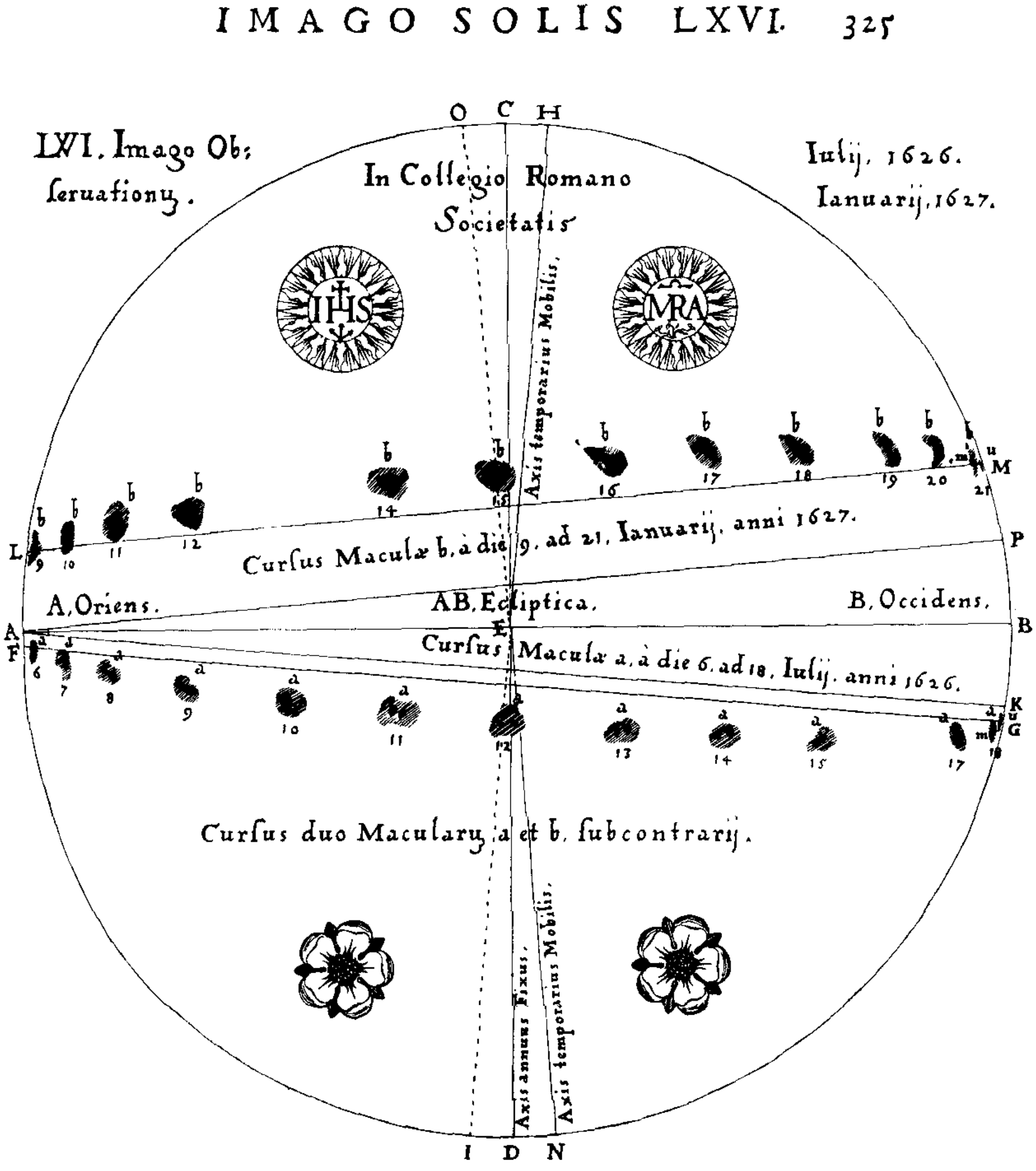


Figure 2.30: The annual life cycle of the Japanese beetle. Original from L. Hugh Newman, *Man and Insects*, (London, 1965), pg. 104-105.

Fig. 2.31 illustrates the motion of sunspots across the face of the sun. The horizontal coordinate is used in its usual role in the sense that the sunspots really do move, and the location of each icon of a sunspot depicts its actual spatial location on a given day. Because many such icons for many different days are superimposed on a single solar disk and also because each icon shows the time-varying shape of the sunspot, space is also serving as a proxy for time; following the images across the sun, one sees the life cycle of the sunspot. The vertical dividing lines of the Japanese beetle life cycle have disappeared; Scheiner's sunspot diagram is a multipanel animation-on-a-page in which the framing of individual panels has completely disappeared.

The odd-looking disks at top and bottom are symbols of Scheiner's religious order and financial patron. Three centuries before television, it was still necessary to say, metaphorically, "and now a word from our sponsor". The hunt for research funds took up much of a scientist's time even in the Renaissance.

Tufte, who reprints this figure, is rather scornful of the appearance of extraneous symbols on the graph. However, many federal agencies and industrial companies require that the agency or company logo appear on every slide or transparency of conference presentations. Fortunately, journal editors have usually been successful in banishing such unabashed advertising from their pages.



IMAGO SOLIS LXVI. 325
O C H
LXVI. Imago Obſeruationũ.
In Collegio Romano Societatis
Iulij, 1626.
Ianuarij, 1627.
Axis temporarius Mobilis.
b b b b b b b b b b b b
L 9 10 11 12 14 15 16 17 18 19 20 21 M
Curſus Maculæ b. à die 9. ad 21. Ianuarij. anni 1627.
P
A. Oriens.
AB. Ecliptica.
B. Occidens.
A F E B
Curſus Maculæ a. à die 6. ad 18. Iulij. anni 1626.
a a a a a a a a a a a a
6 7 8 9 10 11 12 13 14 15 17 18
K G
Curſus duo Macularũ a et b. ſubcontrarij.
Axis annuus Fixus.
Axis temporarius Mobilis.
I D N

Figure 2.31: From Christopher Scheiner, *Rosa Ursina sive Sol* (Bracciani, 1626-1630). The disk-shaped symbols at top and bottom are the emblems of Scheiner's religious order and patron.

Using space as a substitute for time is not the only strategy for "compactifying dimensions". Another is to ignore irrelevant dimensions.

Fig. 2.32 is a plot of sunspot activity as a function of latitude and time. Maunder simply ignored longitude to make a very powerful diagram. As the sun goes through an eleven year cycle, the band of latitude spanned by sunspots moves towards the equator. One could express this as an ordinary line graph by compactifying latitude, too, by plotting the centroid of sunspot positions. However, Maunder's graph is superior because it presents an additional theme which would be lost in an average-latitude-versus-time plot: the number of sunspots abruptsly drops to almost nothing when the sunspots have moved closest to the equator. During a significant part of the cycle (a couple of years), there is almost no sunspot activity at all.

Does this matter? Maunder's major scientific achievement was to show that the period known as the "Little Ice Age" coincided with a major drought of sunspot activity now known as the "Maunder minimum".

How does one know that ignoring one dimension is okay, but compactifying a second dimension erases the most important part of the signal? Experience helps; Maunder spent much of his life studying sunspots. In research, though, experience can be misleading: in Werner von Braun's amusing quote, "Research is what you do when you don't know what you're doing."

The reason message is: Experiment. Plot the data in several different forms, and publish an illuminating subset of the graphs.

There is an old joke about prolific authors: "He was determined to never have an unpublished thought." Some scientists, especially graduate students, approach visualization in the same spirit: Every graph of the previous five years is reproduced in the thesis.

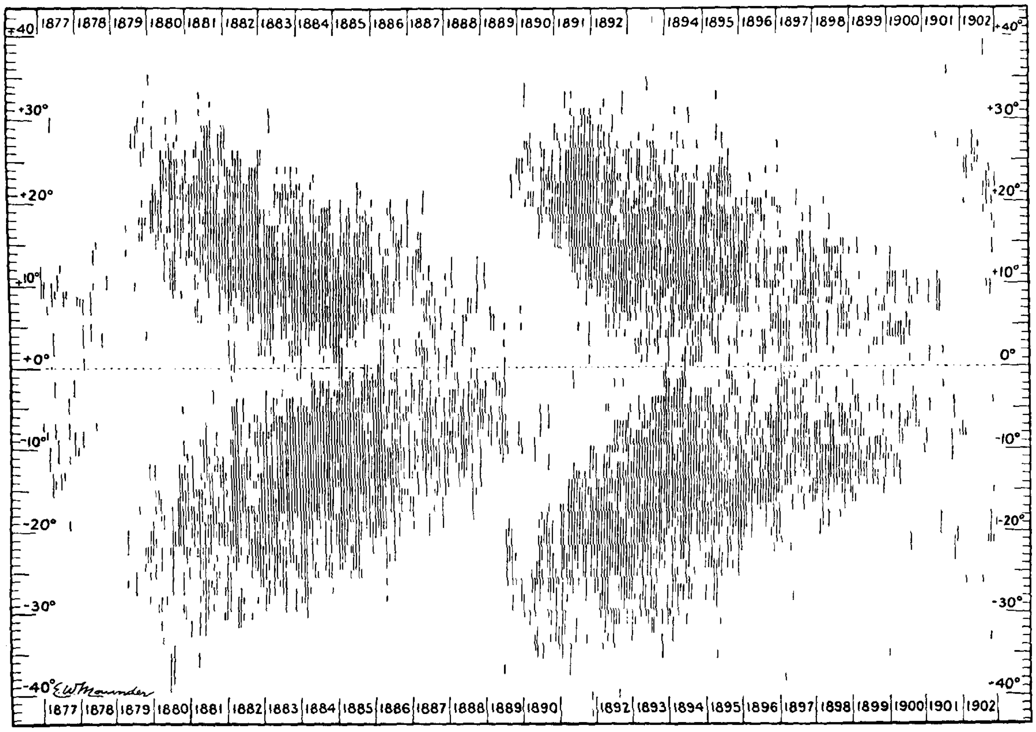


Figure 2.32: The vertical line segments denote the latitudinal extent of sunspots as observed versus time, which is the horizontal axis. Two space dimensions are collapsed into one by ignoring the longitudinal position and width of the sunspots. Note that Maunder has increased the "data-ink" ratio by restricting the latitudinal range to ±40^o; there is only negligible sunspot at higher latitudes, so expanding the vertical axis would have added nothing. From E. W. Maunder, "Notes on the Distribution of Sun-Spots in Heliographic Latitude, 1874-1902," *Royal Astronomical Society Monthly Notices, 64*, 747-761 (1904).

Fig. 2.33 is an improved modern version of Maunder's diagram. One improvement is that it uses a timeline five times as long and a higher density of data. Another improvement is that a line graph has been added in parallel underneath the "butterfly" plot to show how an overall measure of sunspot activity — the area covered by sunspots — varies with time. The lower graph is important because it shows that there are considerable differences from one sunspot cycle to the next. And yet, the steady fluttering of the pattern of "butterfly wings" in the top diagram shows that great similarities from one cycle to the next also.

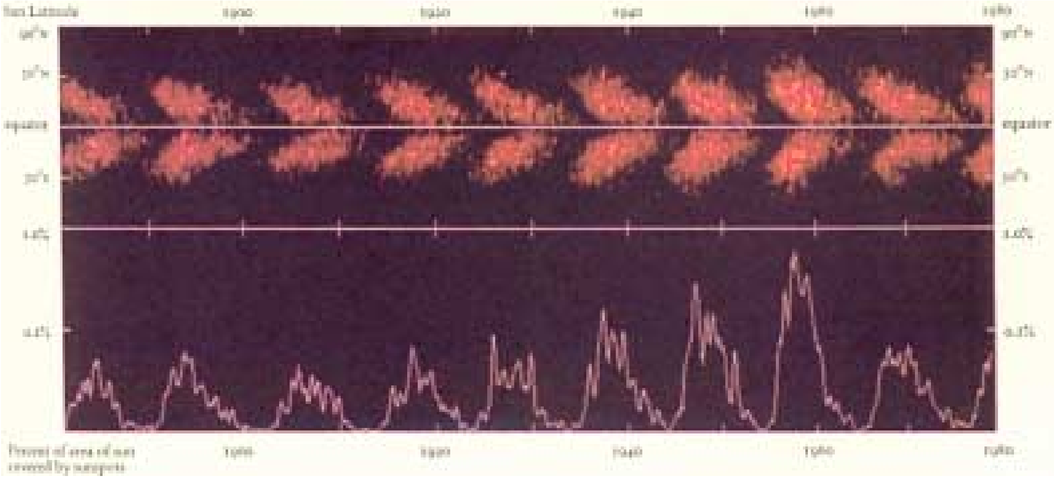


Figure 2.33: Updated version of Maunder's "butterfly" diagram of sunspot activity, so-called because the plot of the latitudinal extent of sunspots versus time traces a shape reminiscent of the wings of butterfly over the course of the 11-year sunspot cycle. The lower graph shows the percentage of the solar surface covered by sunspots as a function of time. Diagram by David H. Hathaway, Marshall Space Flight Center, NASA.

2.9 Supplementary Material

2.9.1 Small Multiples or Animations-on-a-Page

Fig. 2.34 is another example of a "small multiple". The parameter which increases from one frame to the next is not time, but rather the number of terms retained in a truncated power series. The number of oscillations jumps by one by with increase in the truncation k, so it is not possible, even in principle, to smoothly animate this sequence. Nevertheless, Gibbs' Phenonomen is mostly easily described by an animation-like sequence of graphs.

One good feature of this graph is that the frames are all shown in a single row with identical format, making comparisons easy. Another good feature is that the important features, the horizontal and vertical position of the peak which is nearest to the discontinuity, are marked with arrows and numerical labels.

It would normally be bad practice to use numerical labels that extend to six decimal places. However, this is necessary here to show that the location of the peak and its height are both converging to precise values, predicted by Gibbs' analytical theory, as $k \rightarrow \infty$.

The six graphs are given very limited framing; a bottom line plus a horizontal line on the right which is split into two discontinuous segments with a lot of blank space between. One might dub this "3/8" framing. Tufte would approve!

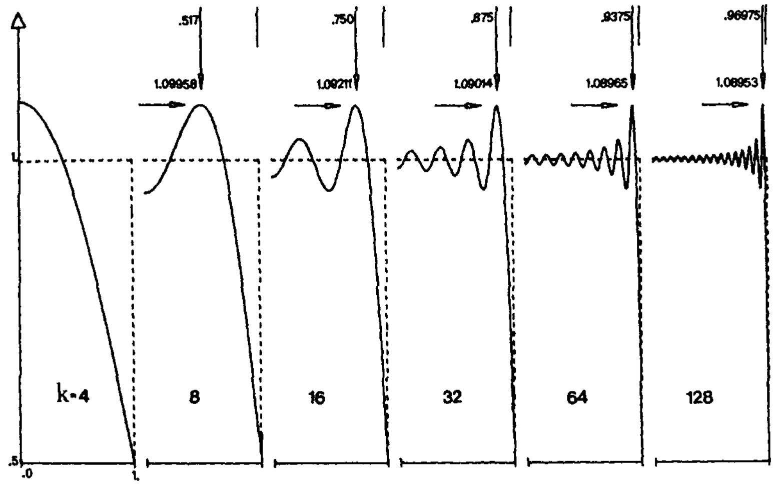


Figure 2.34: An illustration of Gibbs' Phenomenon, which is the non-uniform convergence of a Fourier series to a function with a discontinuity. The integer k, which increases by powers of 2 from one frame to the next, rightward, is the number of terms in the truncation of the Fourier series, which is shown as the solid curve. The piecewise constant function which is being approximated by the Fourier series is shown as the dashed line. (It jumps from 1 to -1 at $x = \pi$.) Because the error is large only near the endpoint $x = \pi$ where the discontinuity is located, the graph does not show the whole interval $x \in [-\pi, \pi]$, but only one-quarter of this interval, $x \in [\pi/2, \pi]$. Similarly, although the range of the function is $[-1, 1]$, only the vertical range above 0.5 is illustrated. The upper labels and arrow in each frame point to the maxima of the truncated series; the message is that it moves closer and closer to discontinuity, proportional to $1/k$. The lower label and horizontal arrow indicate that the maxima of the series converges to an overshoot of 0.0895, as asserted through analytical arguments by Gibbs at the turn of the century. The only unsound aspect of the graph is that horizontal distances (top labels) are measured in fractions of grid width instead of the actual distances, which are $\pi/2$ larger, or fractions of the expansion interval, which would give numbers four times smaller. From *Integral Transforms in Science and Engineering* by Kurt Bernardo Wolf, Plenum Press, New York, pg. 167, (1979).

2.9.2 Separation: Inset Graphs

Fig. 2.35 shows a skillful use of inset graphs. Each of the two cases is shown in two ways. The larger graph illustrates the spatial structure $u(x,t)$ at a particular time t. The inset graph shows the Fourier transform of $u(x,t)$.

These two cases could have been compared as a four-panel composite graph. However, the reader would have to look carefully to see which graph goes with what. Usually, the two graphs of a single case are shown side-by-side, but this convention is not always followed, and the reader has to examine the caption carefully to confirm that convention has been followed. Even when this has been done, the reader must figure out: are both graphs of the same function, or of different quantities for the same case?

The use of the large graph/inset graph visual format ties the two graphs of a given case more closely together. The graphic proximity reflects the intellectual closeness: both graphs are different representations, one in physical space and the other in Fourier transform space, of the SAME function. The two cases are separated in space to reflect a greater distance in the mind.

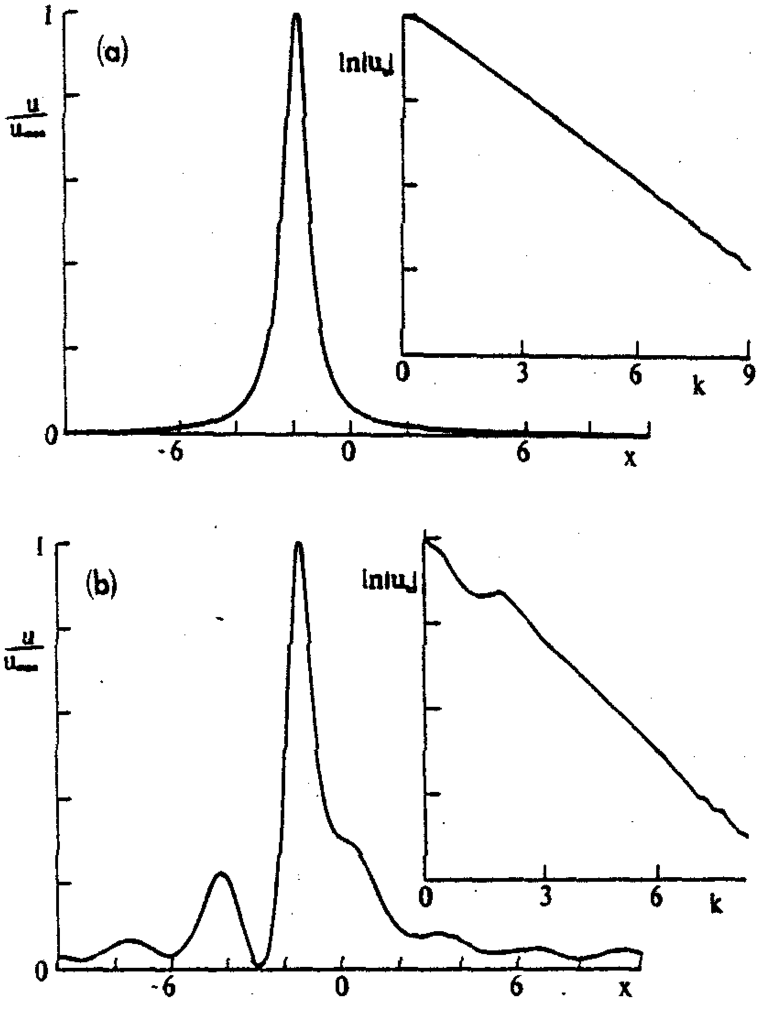


Figure 2.35: An exemplar of inset graphs. The upper pair of figures illustrate the solitary wave of the so-called "Benjamin-Davis-Ono" equation; the lower pair show the effects of radiative damping on these waves. The inset graphs show the Fourier transforms; the larger graphs show the waves in space. From Pereira and Redekopp(1980).

2.10 Wide is Wonderful: Aesthetics of Aspect Ratio

Definition 5 (External Aspect Ratio of a Graph) *The "external aspect ratio" of a graph is the ratio of its width to its height as it appears on the printed page:*

$$\mathcal{R}_E \equiv \textit{width on page/height on page} \tag{2.5}$$

It is often impractical to control the external aspect ratio of a figure, either because of software limitations or to avoid geometric distortion. (A square, two-dimensional domain is best graphed as a square on the printed page, or the image will distort the true geometry.)

When it is possible to design the aspect ratio, most designers include Tufte recommend

$$\mathcal{R}_E \approx 1.5\,\mathrm{to}\,1.6 \qquad \leftrightarrow \qquad \mathrm{width} \approx (3/2)\mathrm{height} \tag{2.6}$$

There are several reasons why this is desirable.

The first is the human visual system. Up-and-down was a lot less dangerous than left-and-right; few humans have been mugged by an earthworm. While the occasional very bold hawk may have dive-bombed *homo africanus*, lions and tigers and bears, oh, my! So, millions of years of evolution have created a human visual system that has broader vision to the sides than to the top and bottom.

Movie makers recognized this a long time ago, which is why cinema screens are wider than they are tall. Because of the limitations of early technology, TV screens have been almost square. As a result, it is difficult to show movies on TV; some movie channels show the film in "letterbox" format in which strips at the top and bottom of the TV screen are blank while in the center, the image is shown in its original aspect ratio of about 1.5.

High definition TV, which will be coming very soon, has adopted a movie-like aspect ratio. The new TVs will look very queer because of their wide screens, but they will be much more comfortable to watch, and not only because of higher pixel density.

The second reason that wider-is-better is that Latex software, at least in the absence of non-standard style files, is unable to wrap text around illustrations. It does have the flexibility to choose the print width of a figure through an optional "width=" statement in the figure command. If one chooses to print a graph of square shape at a width of 3 inches in a text column that is 5.5 inches wide, then almost half the width of the column will be wasted white space. If one blows up the figure to fill the full width of the text, then it will necessarily be 5.5 inches high. Given that the page must also accomodate header, footer, page number, top and bottom margins and the figure caption, such a figure will necessarily appear on a page devoid or almost devoid of text.

If, on the other hand, a graph is wider than tall, a figure 5.5 inches wide will be perhaps only 3 inches tall. It can then comfortably share the page with the text that describes it.

Lastly, wide figures make it easier to provide full-word (or even full-sentence) *horizontal* labels for the data-curves.

So, remember: whenever possible, try to be as smart as your (twenty-first century) TV set!

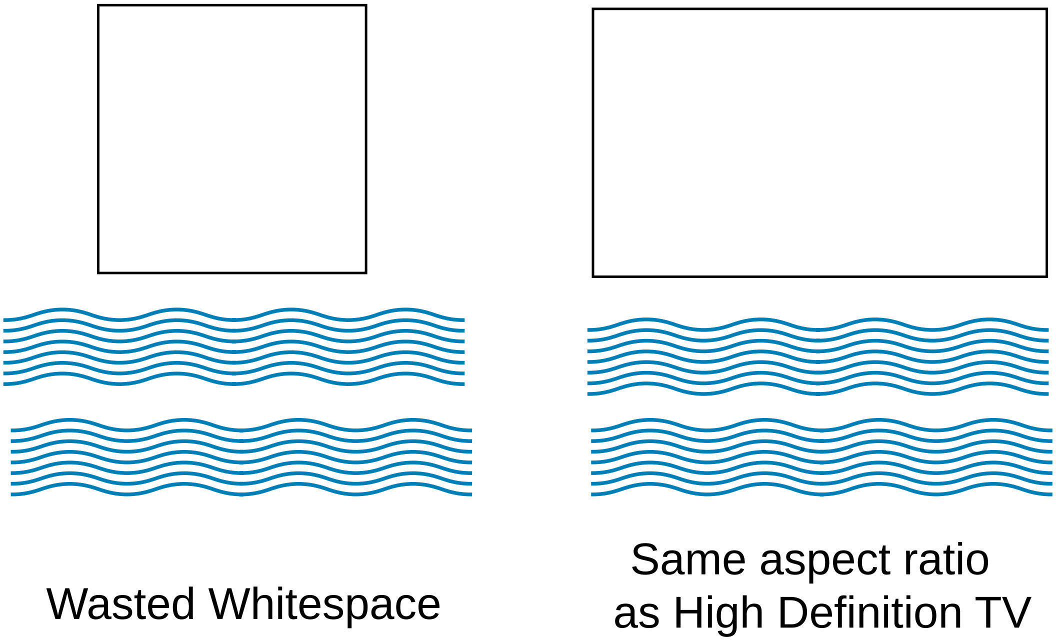


Figure 2.34: Schematic depiction of the advantages of graphs that are WIDER than TALL. The wavy lines represent the main text. A square figure leaves a lot of blank space on either side of itself; a wider-than-tall figure of the same height fills the entire column. The wider figure also matches better with the intrinsic aspect ratio of the human visual system. For this reason, wider-than-tall is standard in the movies and will become standard for TV in the twenty-first century.

2.11 Color or Why the Rainbow Isn't Golden

Color is one of the powerful tools in visualization. Unfortunately, it is also one of the easiest to misuse.

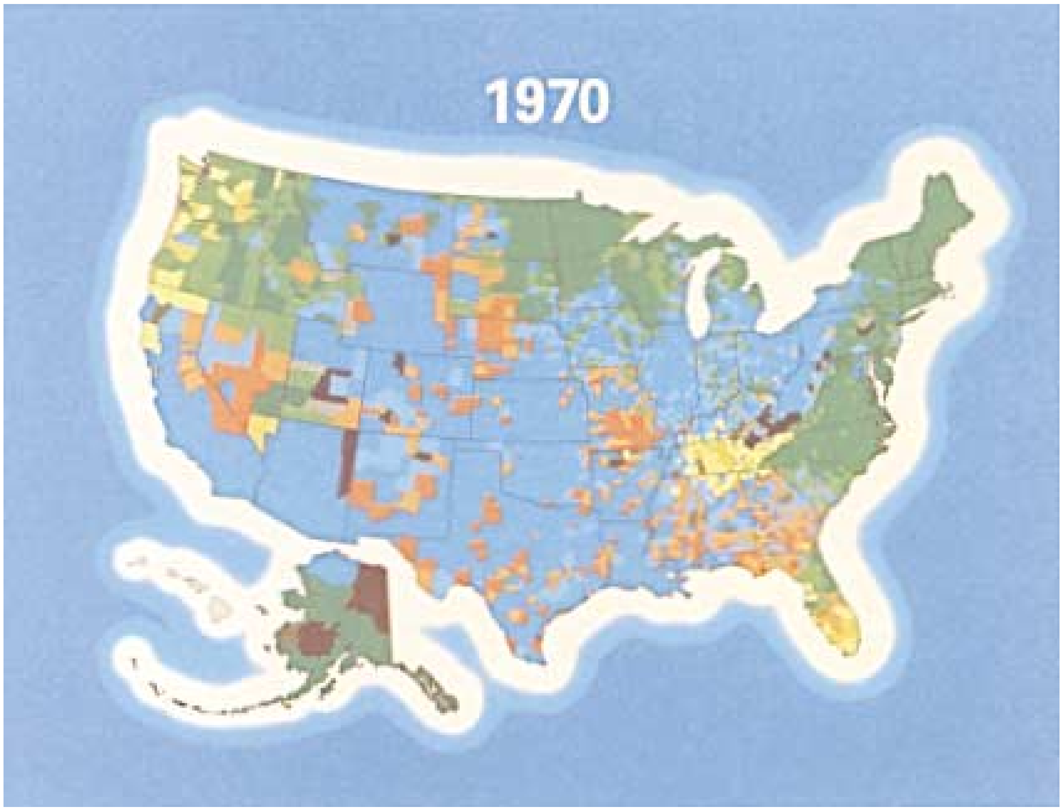


Figure 2.35: Terrible use of color; the coastal waters are a band of white so bright it seems to shimmer. The colors on land depict primary home heating fuel — the meaning of the colors is presumbably explained somewhere in the original text — in 1970. From a publication GE-70 of the U. S. Bureau of the Census. Reprinted in Tufte(1990), pg. 82

Fig. 2.35 illustrates one peril: color can easily emphasize the wrong elements through sheer inadvertence. The intellectual content of the graph is displayed as the little blobs of green, blue, tan and yellow on the land masses. However, the land is surrounded by a great ribbon of white which is so bright that it is far and away the most dramatic, eye-catching part of the graph. This would be good if the graph was supposed to illustrate U. S. territorial waters, whose width is roughly that of the ribbons. Unfortunately, the point of the graph is completely different.

Fine art and interior decorating tend to use rather muted colors; cartoons, posters and crayons use very bright, solid colors. Computers tend to be rather poster-like; subtle colors are difficult to do well because the screen display is RGB, based on additive colors, whereas printers use the CMYK subtractive system, which makes it very challenging to print the screen image exactly.

An even more fundamental reason is that the default set of colors in graphing software is limited to bright, primary colors. In Matlab, 'r' is for bright red, not a subtle lavender pink. The default palette is made of the same shades as a child's box of crayons. Indeed,

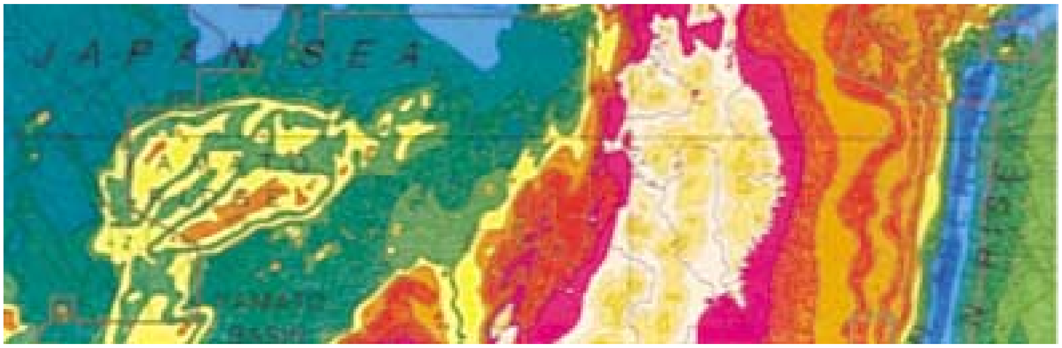


Figure 2.36: A map of the Sea of Japan that uses a "full box of crayons" as shown in the colorbar below (next figure). All the essential information is included, but the bright colors make the map hard to read. One must stare long and hard at the colorbar to recognize that bright yellow is a depth range only slightly smaller than light green.

many color printers dissolve the dyes in wax, and thus are drawing in wax just like a crayon.

With crayon colors, it is very easy for computer graphs to shriek "Look at me! Look at me!" Unfortunately, the loudest shriek is often for something completely irrelevant or even imaginary, like the "ivory moat" around the continents and islands in Fig. 2.35.

In most aspects of graphics, it is desirable to use as many weapons as possible — as many marker shapes, as many linestyles, as many linewidths, as many colors. However, using the full range of bright crayon colors can be a mistake.

In Fig. 2.36, a rainbow assortment of colors is unsatisfactory. The bright colors in the ocean make it difficult to distinguish the land from the sea. The depth of the ocean changes continuously, but the color leaps discontinuously from bright yellow to light green, then jumps again to dark green, then again to very dark green. One is forced to constantly to refer the legend because there is nothing natural to the colorcode: Is it em obvious that yellow is a depth of 1000 meters to 1250 meters?

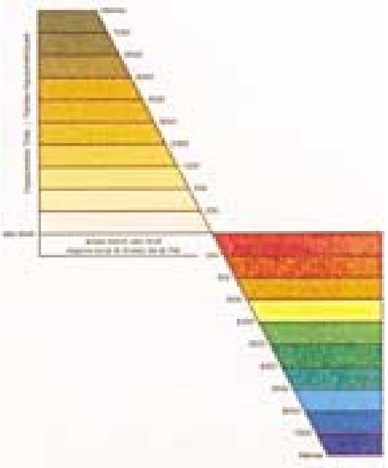


Figure 2.37: Colorbar for the preceding map.

Fig. 2.38 shows the same area, but with a palette which is restricted to shades of blue for the ocean and shades of brown for land. It is now easy to see what is land and what is ocean. There is a natural coding of ocean depth because the deepest sea is very dark blue and the near-surface waters are very pale blue. Even without checking the colorbar, one has an approximate idea of what depth is associated with a given color: middle blue is mid-depth.

In addition, the pale tints of the most of the diagram make it much easier to read the labels — Yamato Basin, etc. This is important because a good map is a pattern of labels. It is much easier refer to high turbulence activity in the Yamato Basin than in a "patch of pale blue just to the west of the big brown mass"!

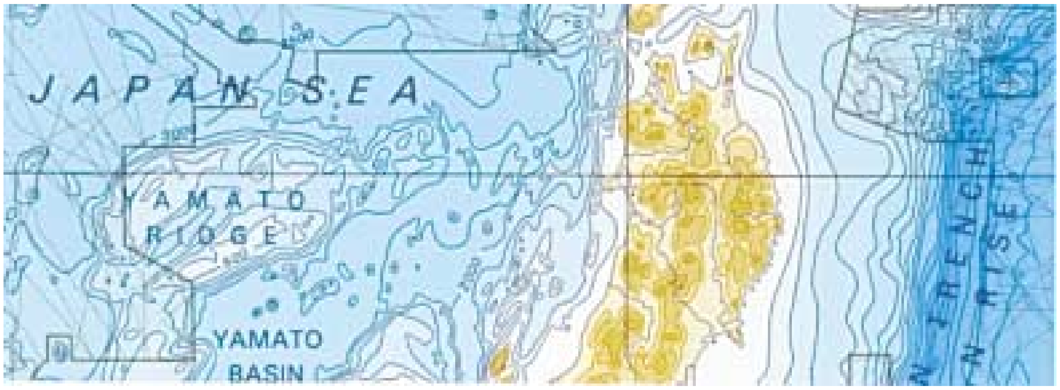


Figure 2.38: The same land-sea bathymetry map of the same region (Sea of Japan) as the previous map, but with a restricted range of colors. Leaving most crayon-colors in the box has created a much more readable map that employs only shades of blue and brown.

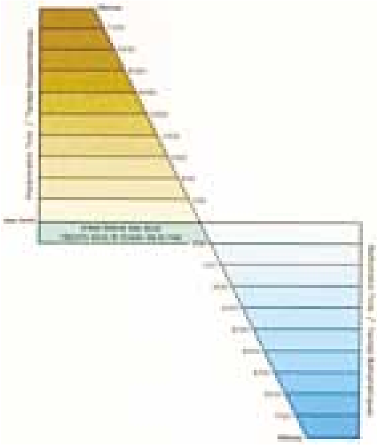


Figure 2.39: Colorbar for the previous figure.

Similar considerations apply to other species of three-dimensional graphs. Hanselman and Littlefield, *Mastering Matlab 5*, assert that surface plots with exterior lighting — **surfl** in Matlab — "look better in a single color" (pg. 345). Fig. 2.40 illustrates what they mean. When a surface plot is displayed using a colormap with rapidly varying, multiple colors, the result is both ugly and confusing (Fig. 2.41).

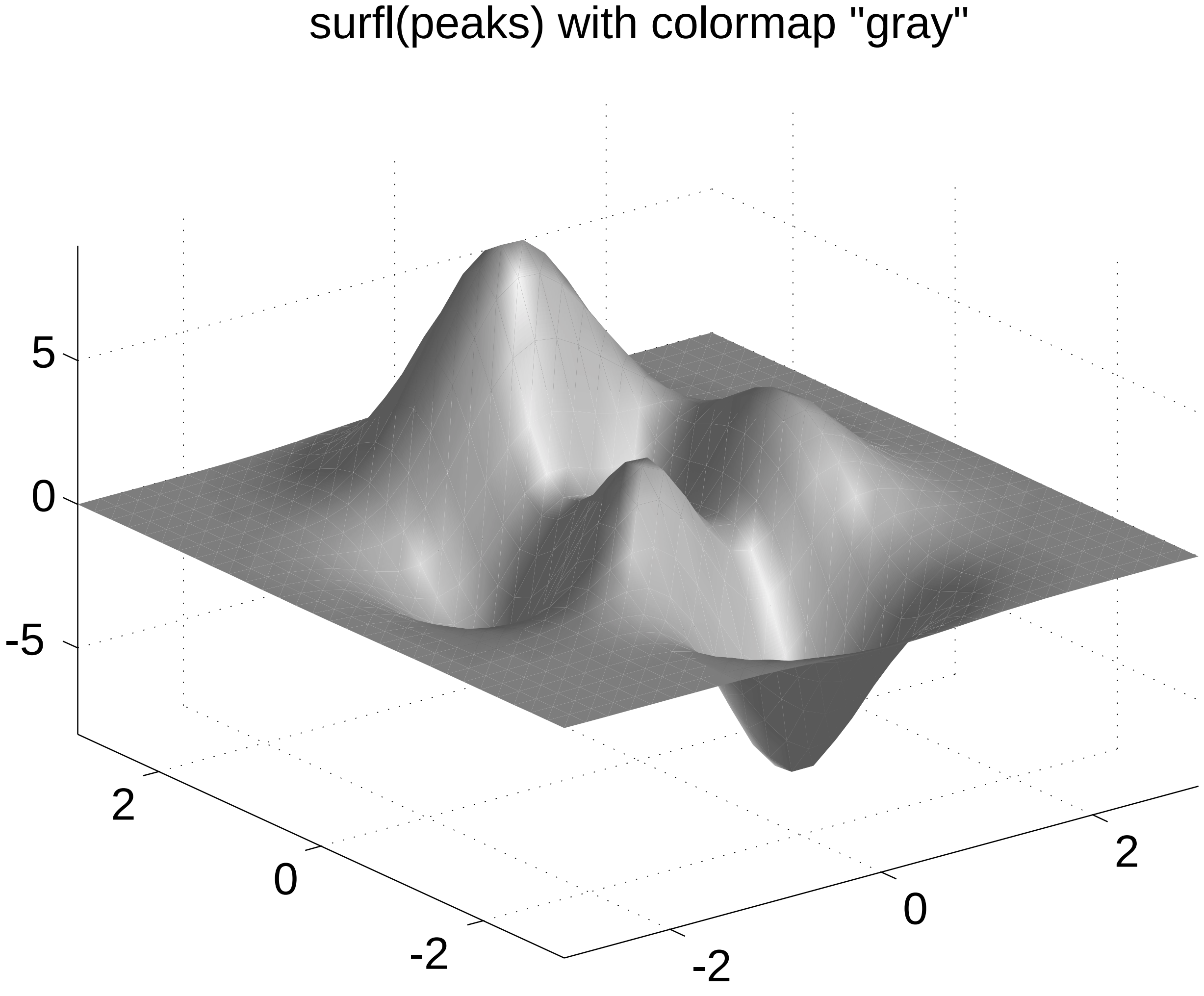


Figure 2.40: A surface plot with lighting (Matlab **surfl**) of the Matlab "peaks" function. With colormap "gray", which is only a single color, the lighting is effective.

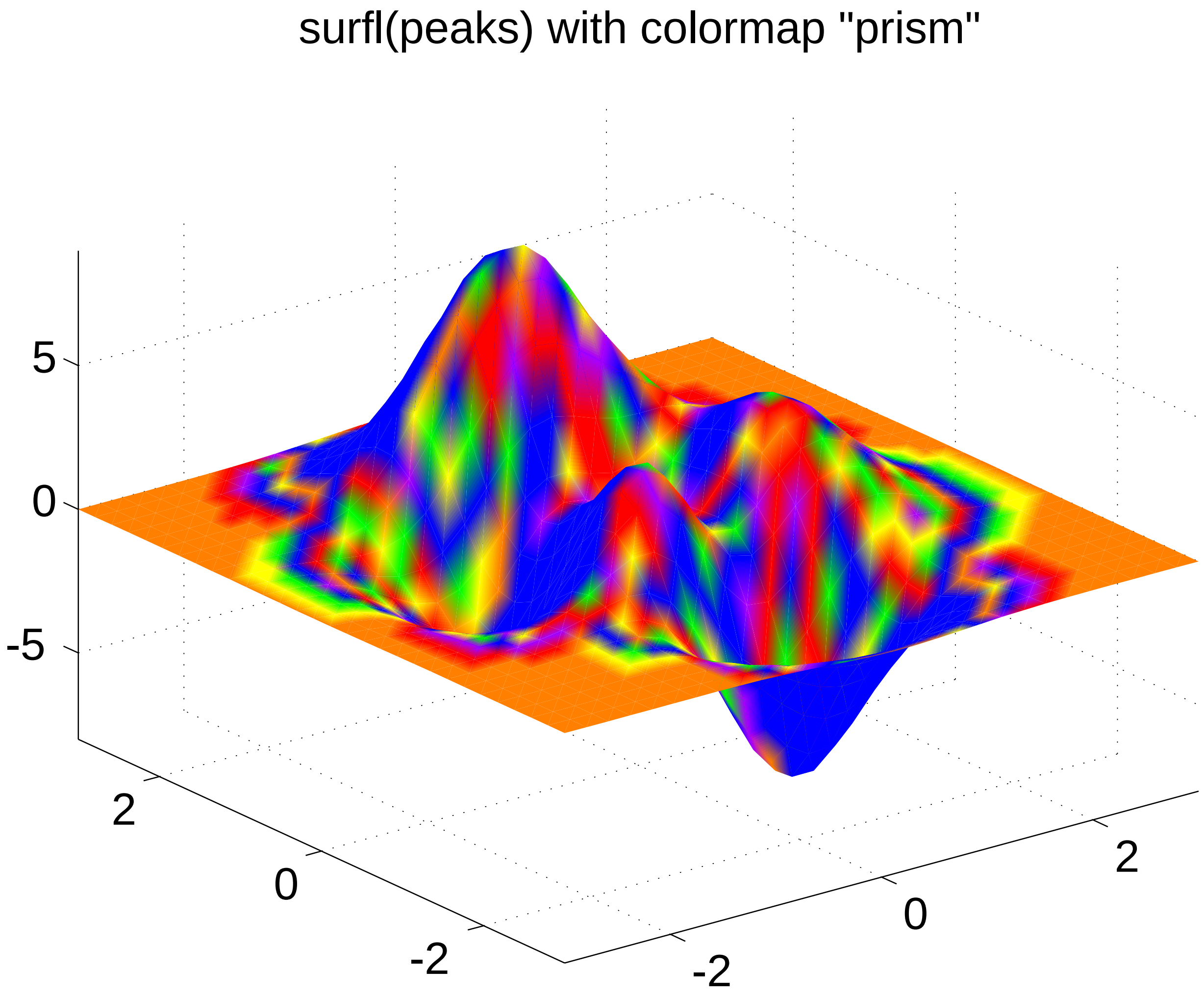


Figure 2.41: Same as previous figure but with colormap "prism". Note that surfl has the odd feature that the colormapping is turned sideways to match the exterior lighting so that the "highlights" correspond to the colors at the top of the colormap rather than the highest elevations on the surface. The reason for this oddity is that effect of exterior level would be overwhelmed by the color scheme if the usual height=color coding was employed.

However, the stricture to leave most of the crayon-colors in the box must be interpreted to varying degrees for different species of graphs. Fig. 2.42 shows the same function as in the previous surface plots, but this graphed as *pseudocolor* image, which is the Matlab **pcolor** command.

The multicolored map creates a confusing image ("prism" map). However, the two monochrome images at the top are rather bland and uninformative, too. The best image (lower right) uses the "jet" map, which uses only shades of TWO colors.

This is the same color philosophy as the good map of the Sea of Japan. One color, in various shades, is used for heights above zero — brown in the bathymetric map, red in the pseudocolor plot with the "jet" colormap. The second color is used for negative heights — blue in both images.

The two color schemes are effective in both examples because the color-coding seems natural and intuitive, even though it is artificial. In the topographic map, the association of brown with land and blue with water is obvious because much of the land *is* brown or tan and the water *is* blue. Similarly, the pseudocolor map is easier to absorb because we instinctively associate red with highs — highs of temperature as in the phrase "red-hot" — and blue with lows or with cool temperature. Even when the quantity which is plotted is not temperature, the mapping of red with high and blue with valleys seems very natural and therefore easily remembered. Furthermore, this red-high/blue-low mapping has been widely used in scientific pseudocolor maps. (Trust us on this!) When one has seen a dozen earlier plots that used the same convention, it becomes much easier to get past the colormap and into the data.

A final comment: red-green color blindness affects 5% to 10% percent of the population. (Almost all victims are male). Therefore, a two-color scheme should avoid using red and green; red and blue are easier for the color-blind to distinguish. A rainbow palette of colors is almost certain to create problems for the color-blind.

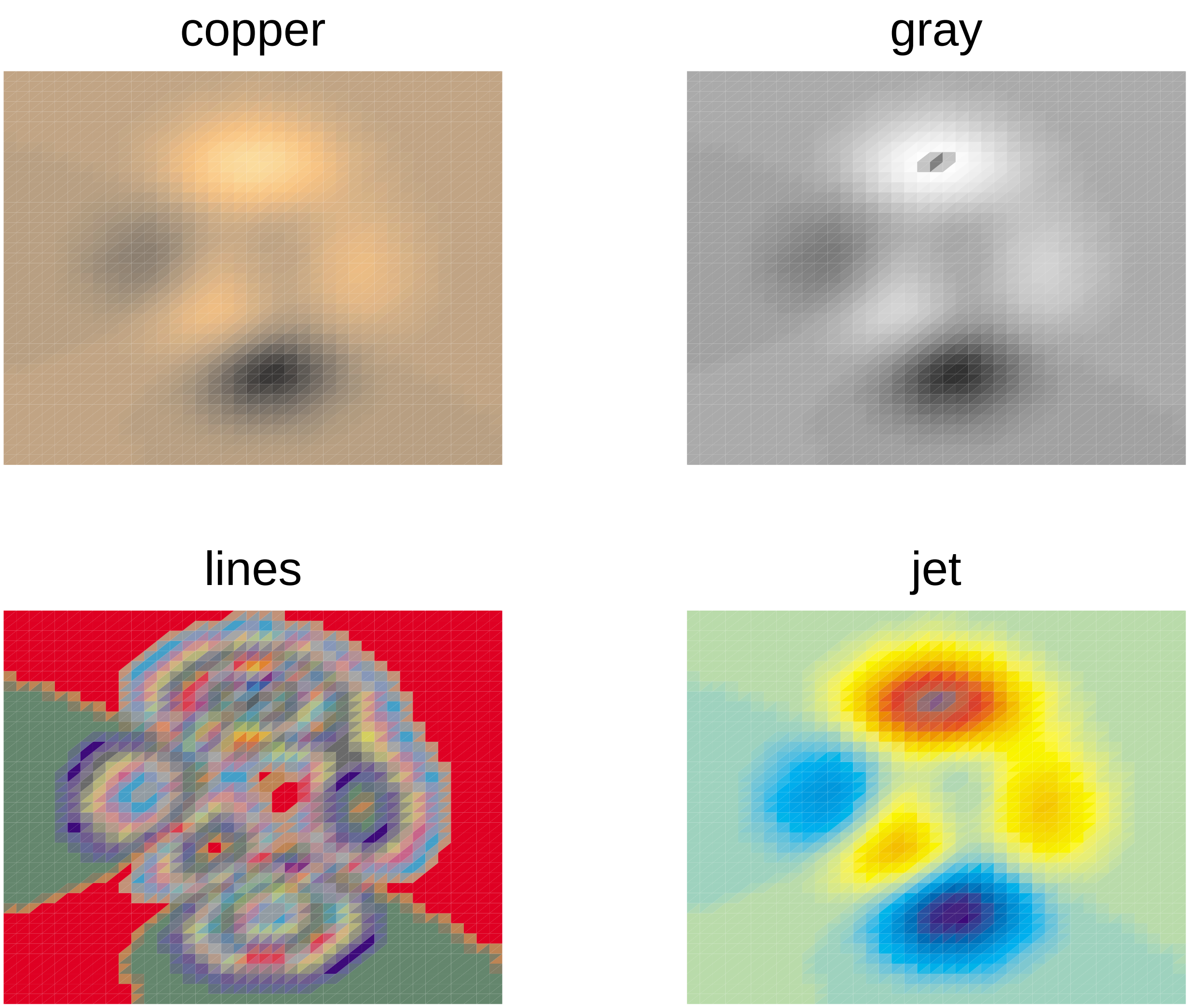


Figure 2.42: A pseudocolor map of the Matlab "peaks" function using four different Matlab colormaps. (Parenthetically, note that it is not possible to directly employ multiple colormaps in a single figure in Matlab. The workaround is to define a composite colormap which includes all four of the colormaps displayed: **mycolormap = [jet(32);lines(32);gray(32);copper(32)];**. One must then declare the color axis to have a range which is four times larger than the actual range of the "peaks" function: **caxis([-32 32]);**, and repeat this statement in each subplot. Finally, in each subplot, the "peaks" function with a shift was graphed: **pcolor(X,Y,Z+24);** and similarly with different shifts in the other subplots.)

Unfortunately, monochrome and two-color schemes have a failing: there is some loss of precision as subtle shades blend into one another. A colormap that is punctuated by white bars like "prism" or "flag" shows the messiness of turbulent jet flows or other phenomena with a lot of small-scale structure more clearly than a two-color map. There is a trade-off between garishness and precision.

One compromise is to superimpose contours on top of a pseudocolor map. When there are a lot of contours, the result is usually called a "filled contour plot", and there is a command in Matlab, **contourf**, specifically to do this (Fig. 2.43).

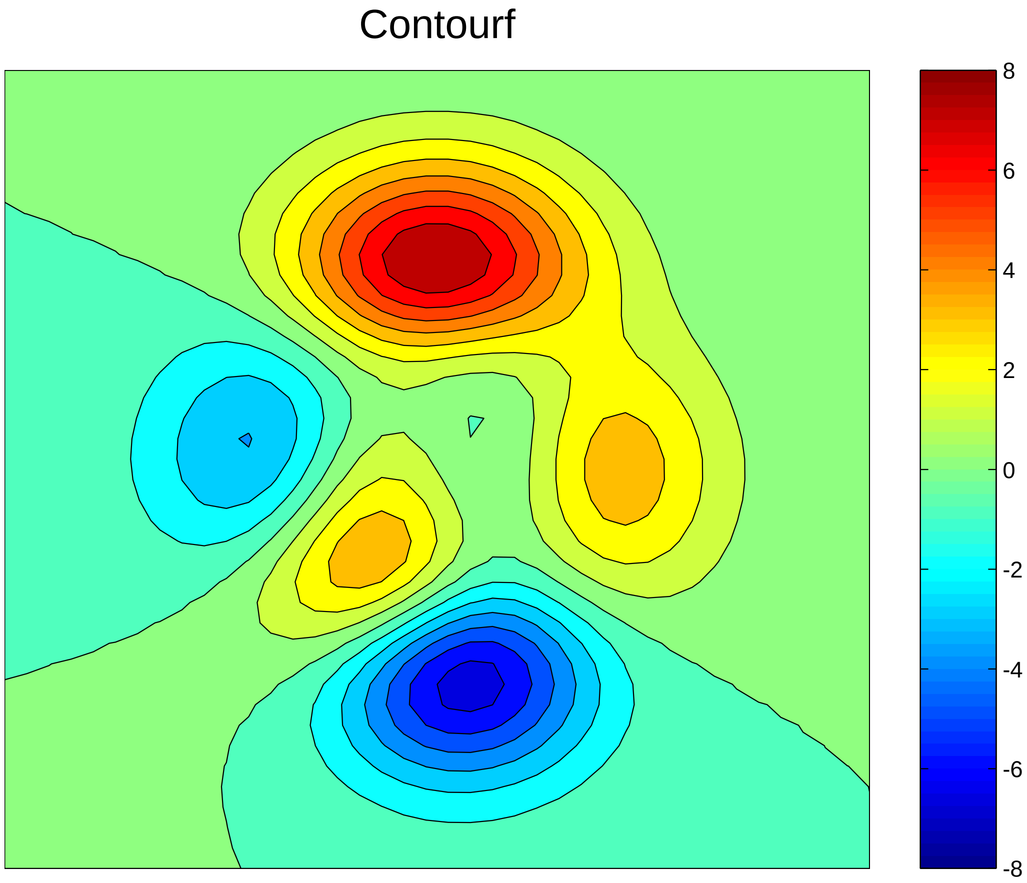


Figure 2.43: Note that in Matlab, interpolated shading (**shading interp**) is not allowed in filled contour plot; the space between each pair of contours is always filled with a single shade.

Sometimes, one only needs a single explicit contour such as the zeroline. There is a good way and a bad way to add this contour. The bad way is to modify a standard colormap so as to change a single level to black or white. As shown in Fig. 2.44, the problem with this strategy is that wide regions where the peaks have decayed to near-zero values are turned into a sea of white. The rub is that each level of a standard colormap is a color assigned to a RANGE of values equal to 1/64 of the range of values spanned by the colors. Visually, the pseudocolor map is dominated by the white seas where the function is boringly flat.

Fig. 2.45 illustrates a better way, which is to superimpose a contour plot that has been instructed to graph just a single contour.

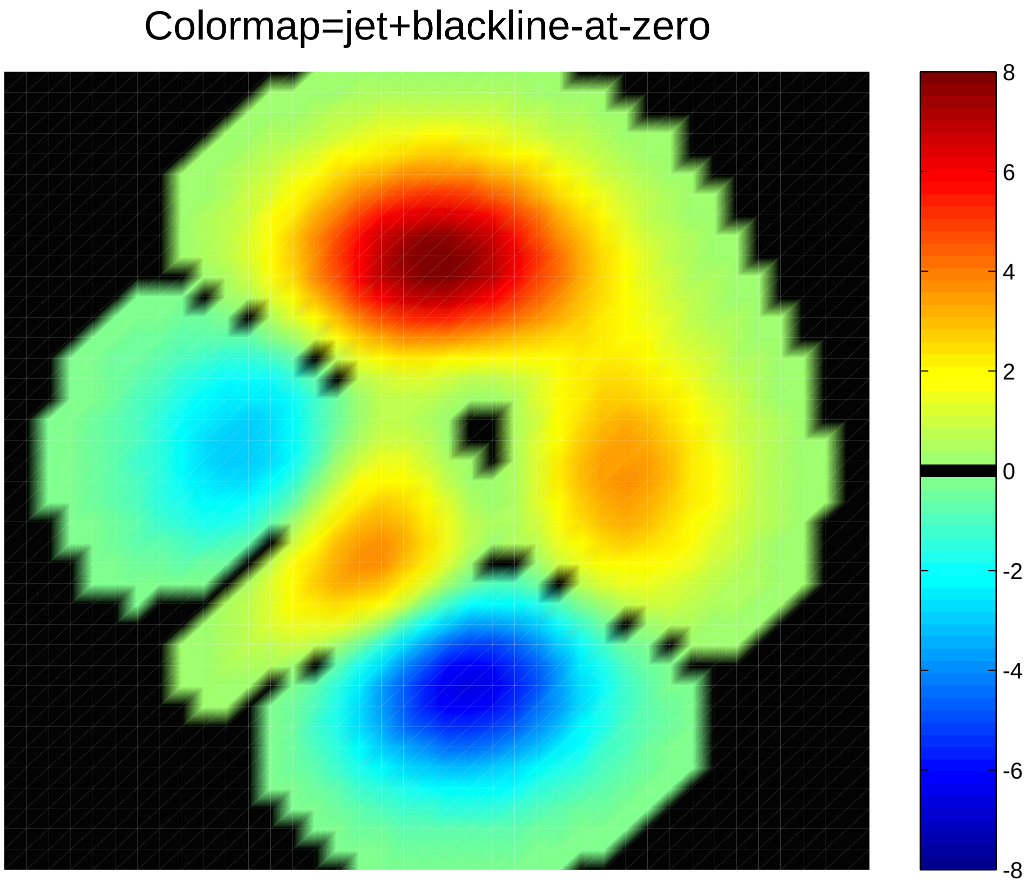


Figure 2.44: Created by the Matlab commands **blackline=jet(64); blackline(32,:)=[0.99 0.99 0.99]; colormap(blackline)**

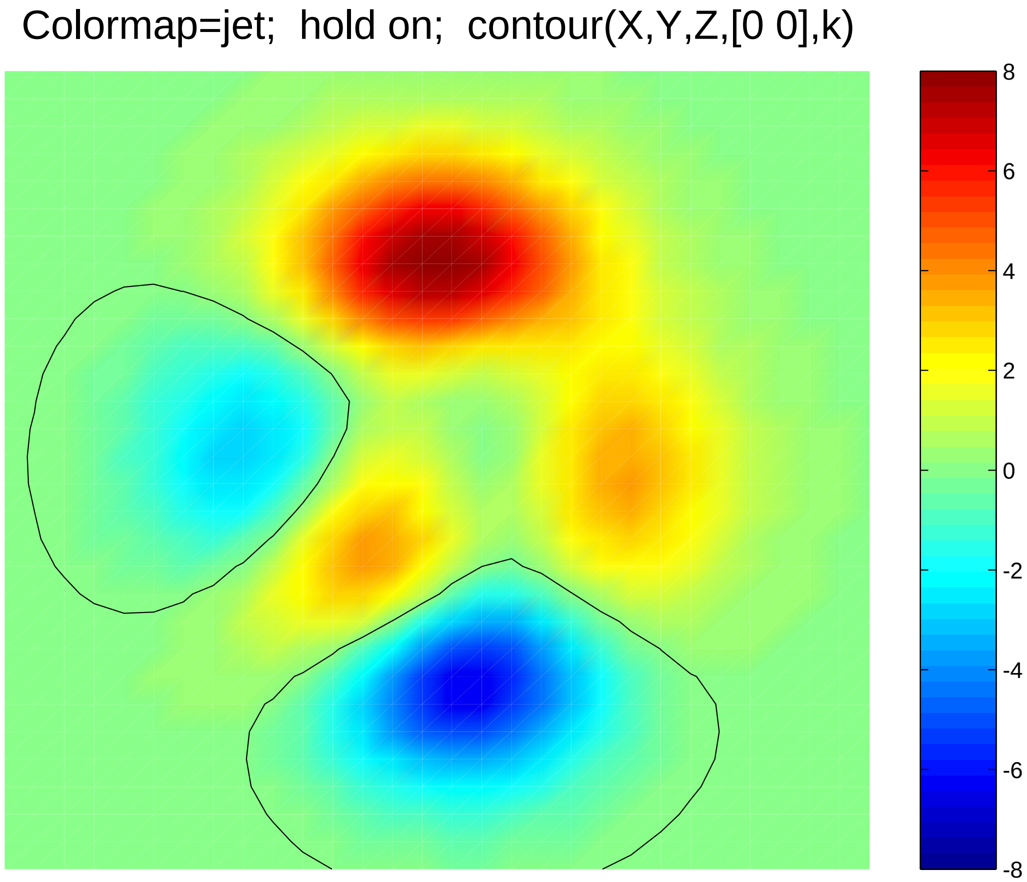


Figure 2.45: Created in Matlab by first making a standard pseudocolor map using the built-in colormap “jet”. After the **hold** command, the statement **contour(X,Y,Z,[0 0],’k’)** will superimpose a single contour line where the plotted function is equal to zero.

2.12 Parallelism

When multiple images are combined in parallel, the message is easier to grasp because the axes, format and so on are constant and only the data varies. Parallelism is closely related to "small multiples" and "animations-on-a-page". In the most favorable cases, the parallelism implicit in these concepts can be translated into explicit geometry.

In the late 60's, Stephen Dole of the Rand Corporation combined all that was then known of solar system formation into a computer model. It was thought then, and still believed now, that chance plays a fairly large role in the number and size of the planets as well as non-stochastic parameters such as the total mass of the parent star and the planetary nebula. Dole's model therefore generated ENSEMBLES of solar systems.

Fig. 2.46 illustrates a subset of his ensemble of imaginary solar systems compared with our own real system at the bottom. The size of the disks represents the mass of the planets, but the labels give the numerical values — good since it is hard for humans to accurately estimate area. Because each system is laid out on a single line, and the lines are all parallel, it is easy to visually compare different systems.

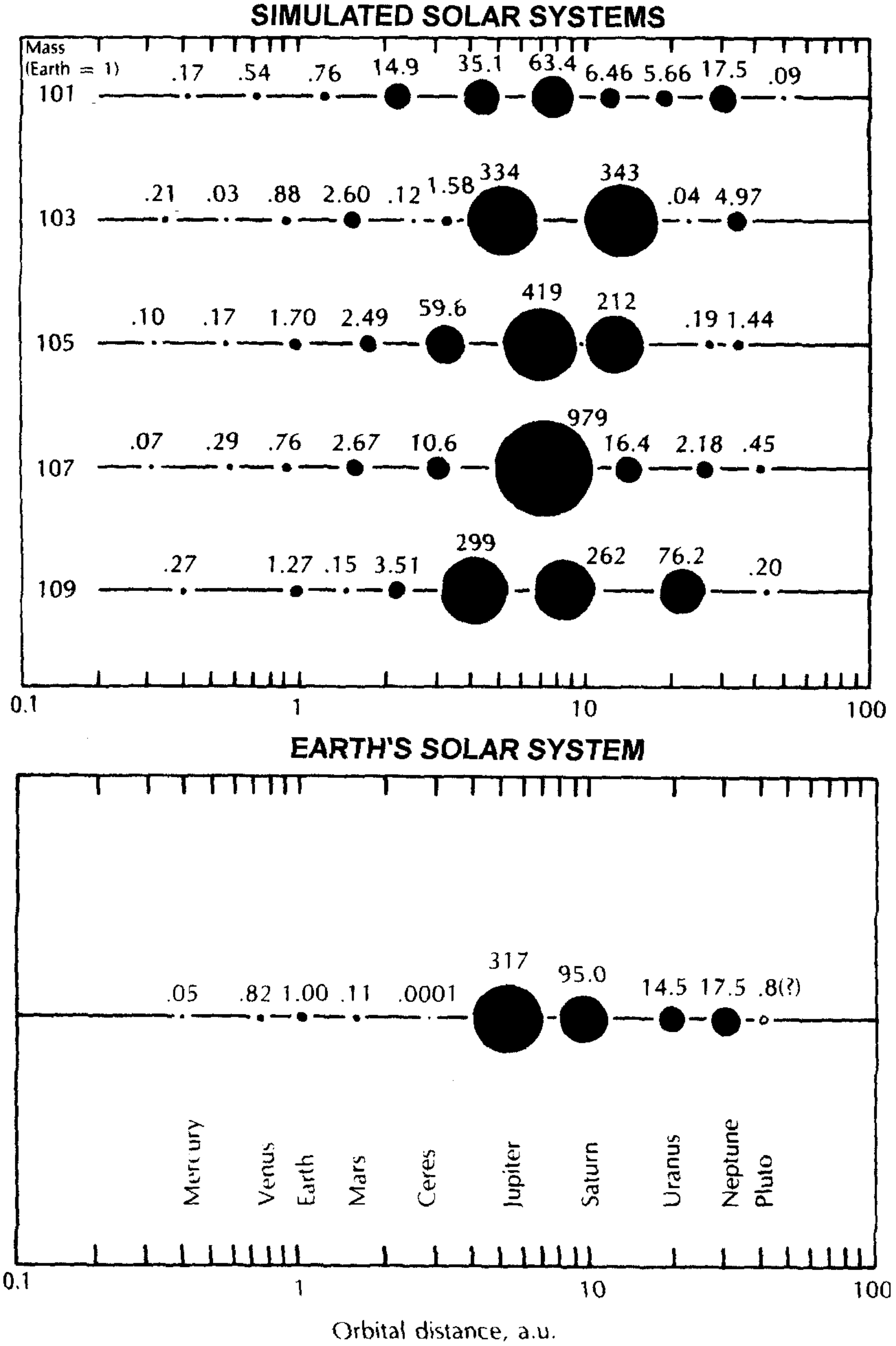


Figure 2.46: From *Habitable Planets for Man* by Stephen Dole, Elsevier, New York (1970).

Fig. 2.47 shows a similar vertical stack of plots which are different variables associated with a single model flow. Although the aspect ratio of such plots is greater than the value recommended earlier — each graph is perhaps three times as wide as tall — this does not seem bothersome when such elongated rectangles are arrayed one atop the other. Because the plots are wide, it is easy to label them with horizontally-written text.

It is possible to stack plots that are very tall and narrow side-by-side, and sometimes this is useful to facilitate comparisons between the ordinates. However, it is harder to use a side-by-side orientation successfully because there is less room for horizontal labels.

In either event, a good strategy for transmitting ideas that are conceptually parallel is through a plot which is parallel in layout, too.

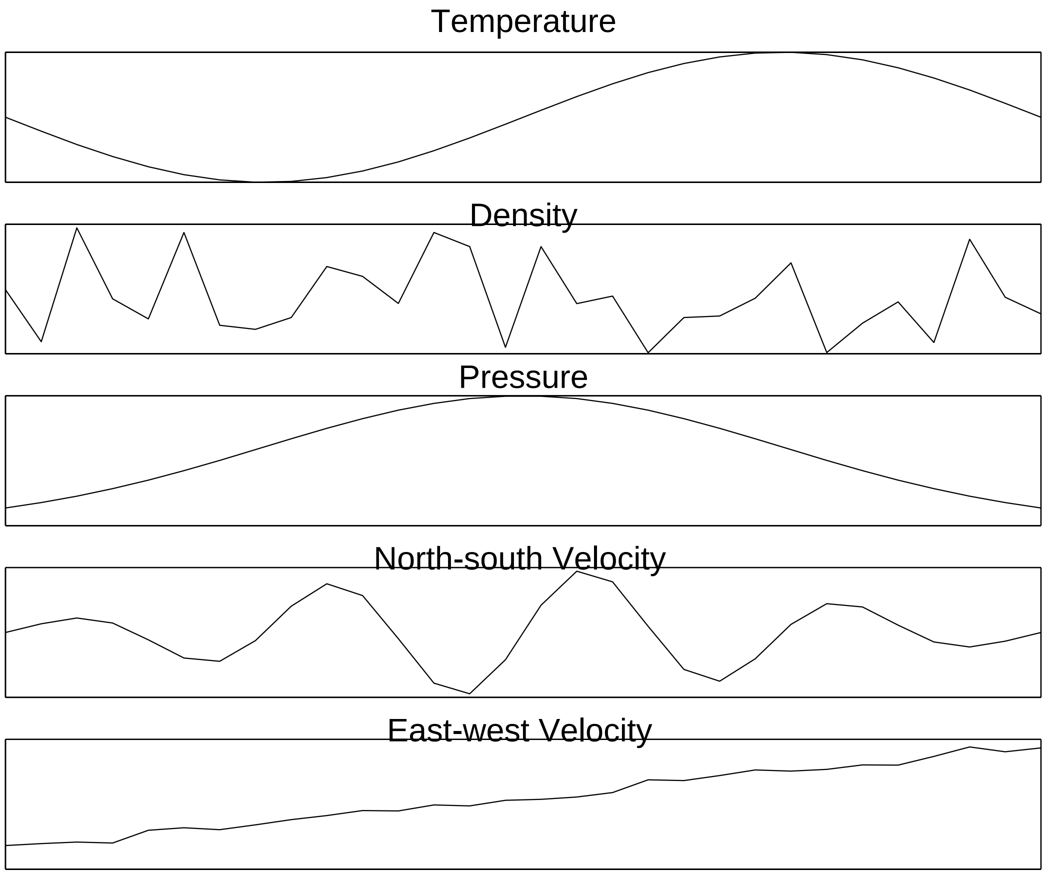


Figure 2.47: Five vertically-stacked plots for different variables of a single case study, all sharing a common horizontal axis.

2.13 The Friendly Graphic

Table 2.3: The Friendly Data Graphic, pg. 183 of Tufte(1983)

Friendly	Unfriendly
words are spelled out, mysterious and elaborate encoding avoided	abbreviations abound, requiring the viewer to sort through text to decode abbreviations
words run from left to right, the usual direction for reading occidental languagues	words run vertically, particularly along the Y-axis; words run in several different directions
little messages help explain data	graphic is cryptic, requires repeated references to scattered text
elaborately encoded shadings, cross-hatching, and colors are avoided; instead, labels are placed on the graphic itself; no legend is required	obscure codings require going back and forth between legend and graphic
graphic attracts viewer, provokes curiosity	graphic is repellent, filled with chartjunk
colors, if used, are chosen so that the color-deficient and color-blind (5 to 10 percent of viewers) can make sense of the graphic (blue can be distinguished from other colors by most color-deficient people)	design insensitive to color-deficient viewers; red and green used for essential contrasts
type is clear, precise, modest; lettering may be done by hand	type is clotted, overbearing
type is upper-and-lower case, with serifs	type is all capitals, sans serif