Elementary Plotting

Claudio Silva Luis Gustavo Nonato

Steve Callahan created a large number of these slides for the Sci Vis course at the University of Utah

Elementary Plotting Techniques I

Motivation

- Everyone uses plotting
- It is easy to lie or to deceive people with bad plots
- Default plotting tools are terrible
- Most people ignore or are unaware of simple principles

5

Elementary Plotting Techniques

Plots, charts, and graphs are often used interchangeably.

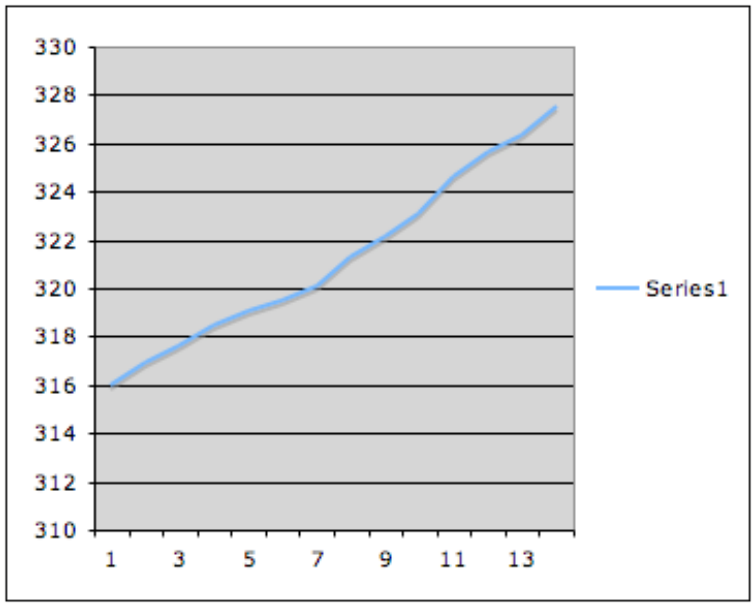
Plotting data is one of the oldest forms of visualization. In fact, many of the standard plotting techniques were introduced in the late 18th century by William Playfair [Playfair 86, Playfair 01], a pioneer in information visualization. Even today, plotting is by far the most prevalant method for analyzing, correlating, condensing, and presenting scientific data. This is because, with a properly created plot, our visual system is easily able to distinguish patterns that may lead to insight about the underlying data. Conversely, with a bad plot, it is easy to confuse or even deceive the observer about the underlying data. Learning good plotting techniques should not be underestimated because of its importance in the scientific community for publishing and presenting results of hypotheses and experiments. Yet, the subject is often entirely left out of the curriculum for most college students in scientific disciplines!

It is important to note that the goals for plotting in a scientific setting are not the same as they are for those used in general media settings, such as newspapers and magazines. A more advanced knowledge base can be assumed about the scientific reader—less emphasis can be placed on extraneous or superfluous information and more emphasis can be placed on the data itself. The techniques described in this chapter are directed at the scientific community, though many of the principles apply in a more general setting.

There are two basic purposes for plots: data analysis and data communication. As readers and observers of publications and presentation, we are generally more familiar with the latter. However, the former may be of greater importance during the research phase where hypotheses are formed and tested. In either case, the process of creating a useful plot is more iterative than direct. The tack of performing experiments and gathering data can be time consuming, do not expect the analysis to be any different.

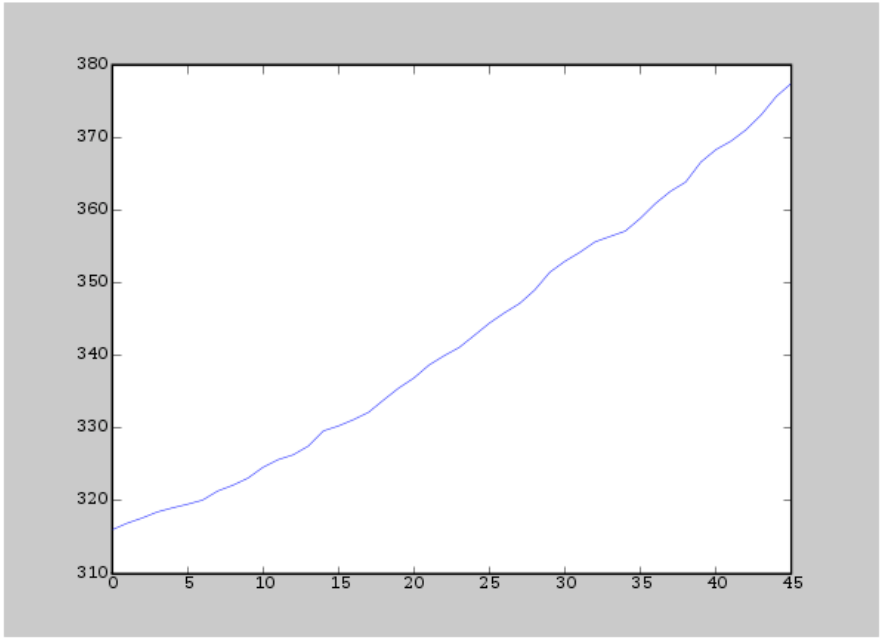
In a simplistic view, plotting is just reducing a large amount of information to a smaller form that is more easily understood. There is often a misconception that plotting is a way of presenting the data itself, taking the place of a table or list of the actual values. To the contrary, plotting should be used for displaying relationships within the data. Understanding the information that is being displayed

Motivation



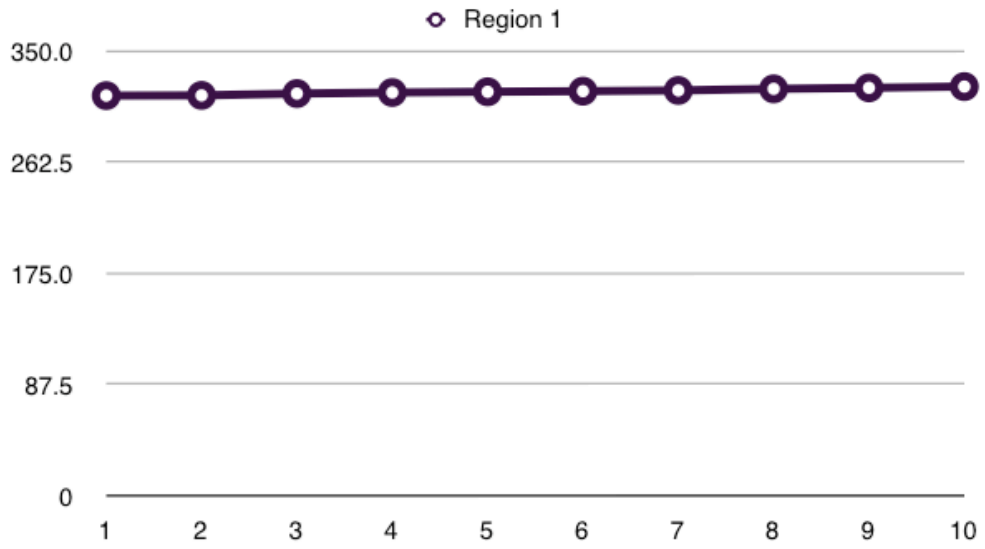
Default Excel Plot

Motivation



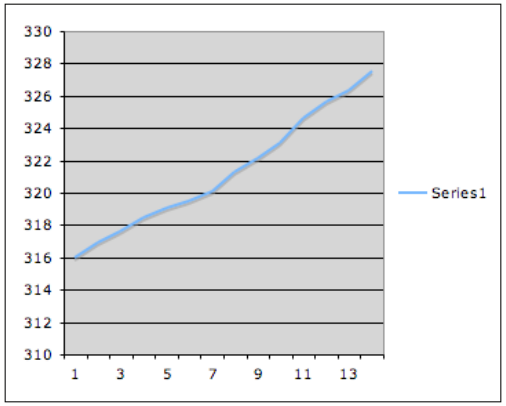
Default Matplotlib/Matlab Plot

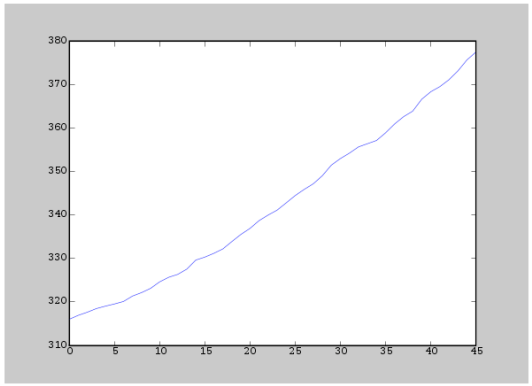
Motivation

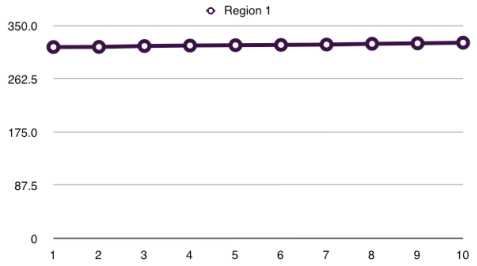


Default Pages Plot

Motivation



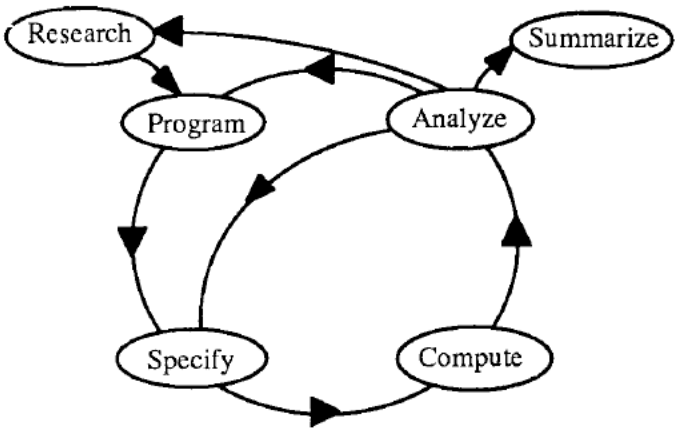




- Why are they all different?
- What is good/bad about each?

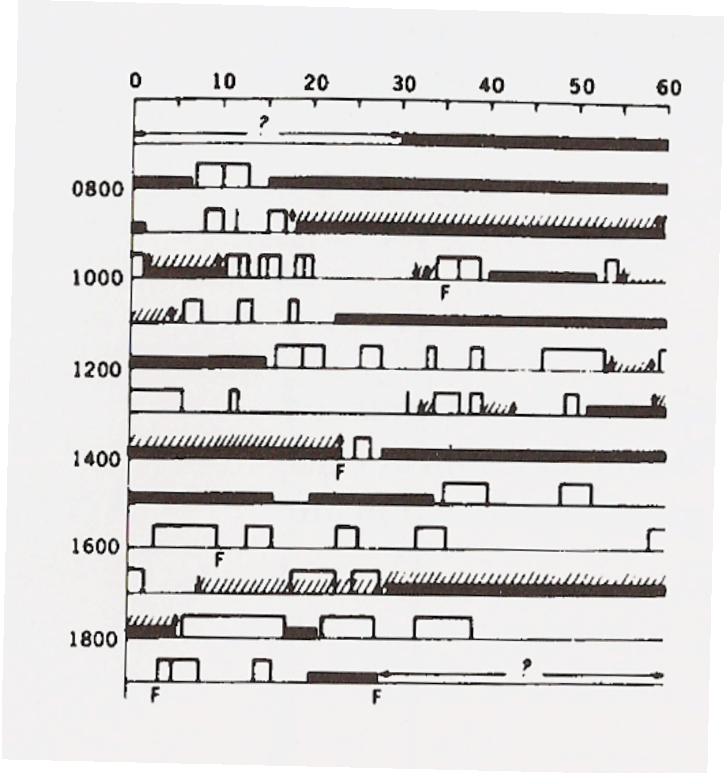
Fundamentals of plotting

- Analysis vs. Communication
- Presenting data vs. Presenting correlation
- Vision vs. Understanding



Clear Vision

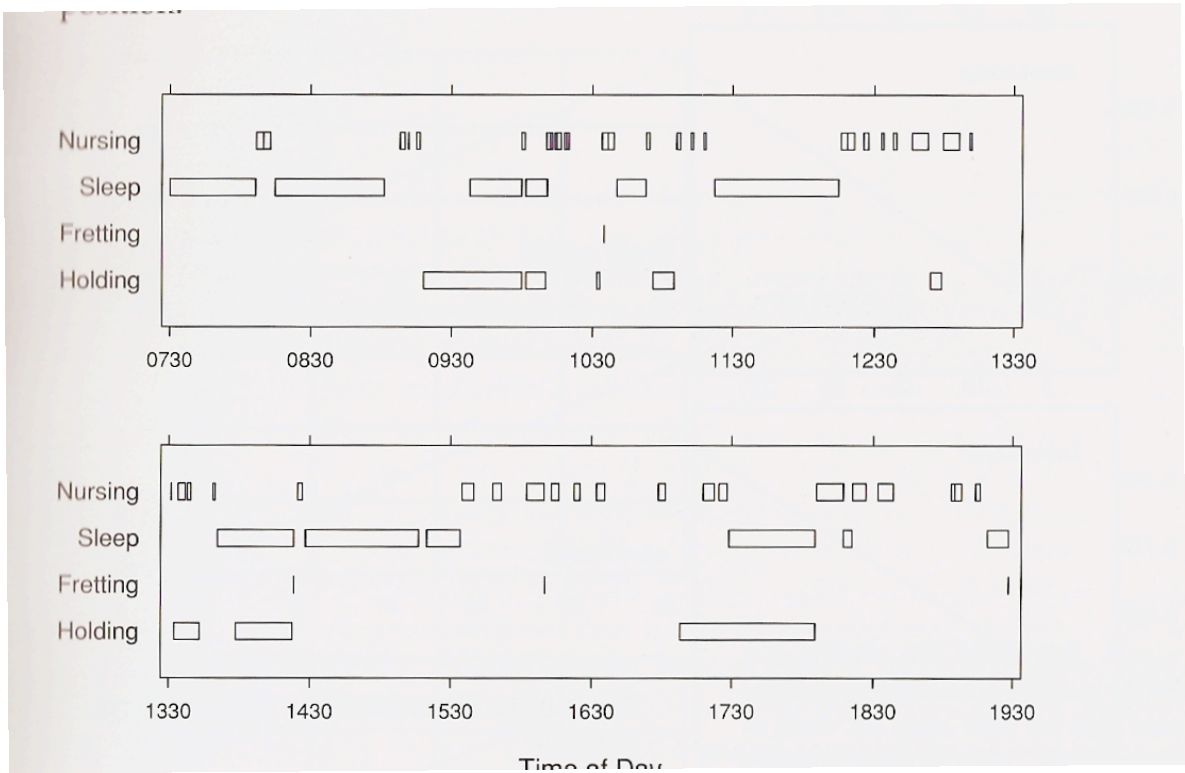
- Principle 1: Make data stand out
  - Avoid superfluidity, clutter, or chartjunk.



Activities of a !Kung woman and her baby
Open Bar and Vertical Lines: Nursing times
Closed Bars: Sleeping
F: Fretting
Slashed Lines: Held by mother
Arrows: Picking up and setting down

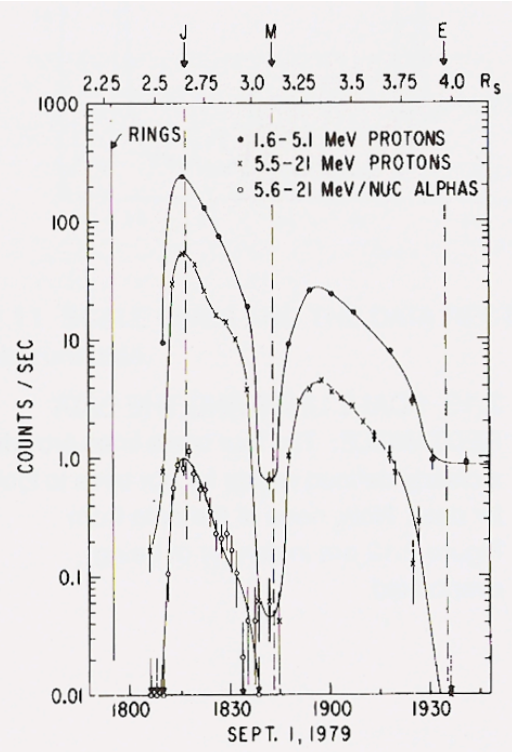
Clear Vision

- Principle 1: Make data stand out
  - Avoid superfluidity, clutter, or chartjunk.



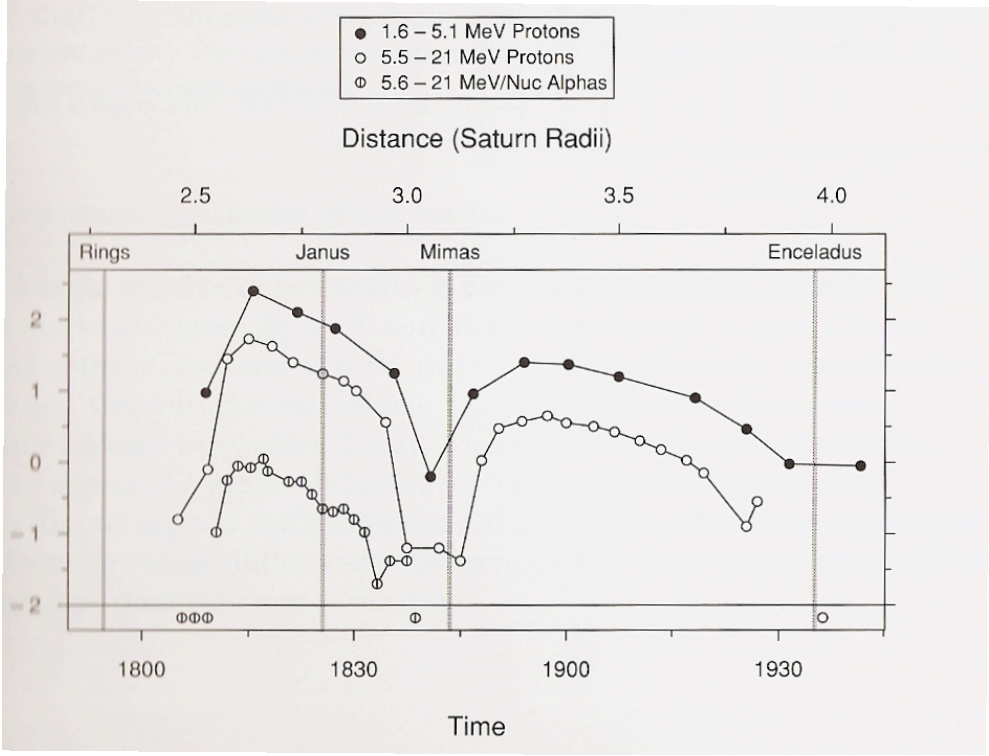
Clear Vision

- Principle 1: Make data stand out
  - Avoid superfluidity, clutter, or chartjunk.



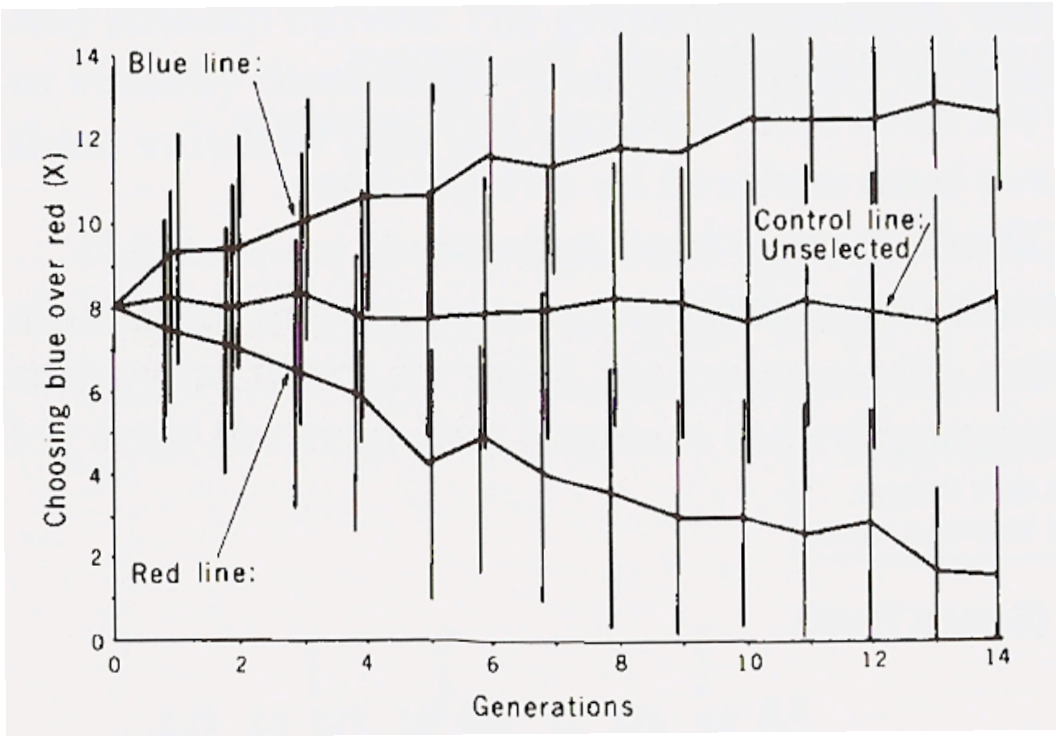
Clear Vision

- Principle 1: Make data stand out
  - Avoid superfluidity, clutter, or chartjunk.



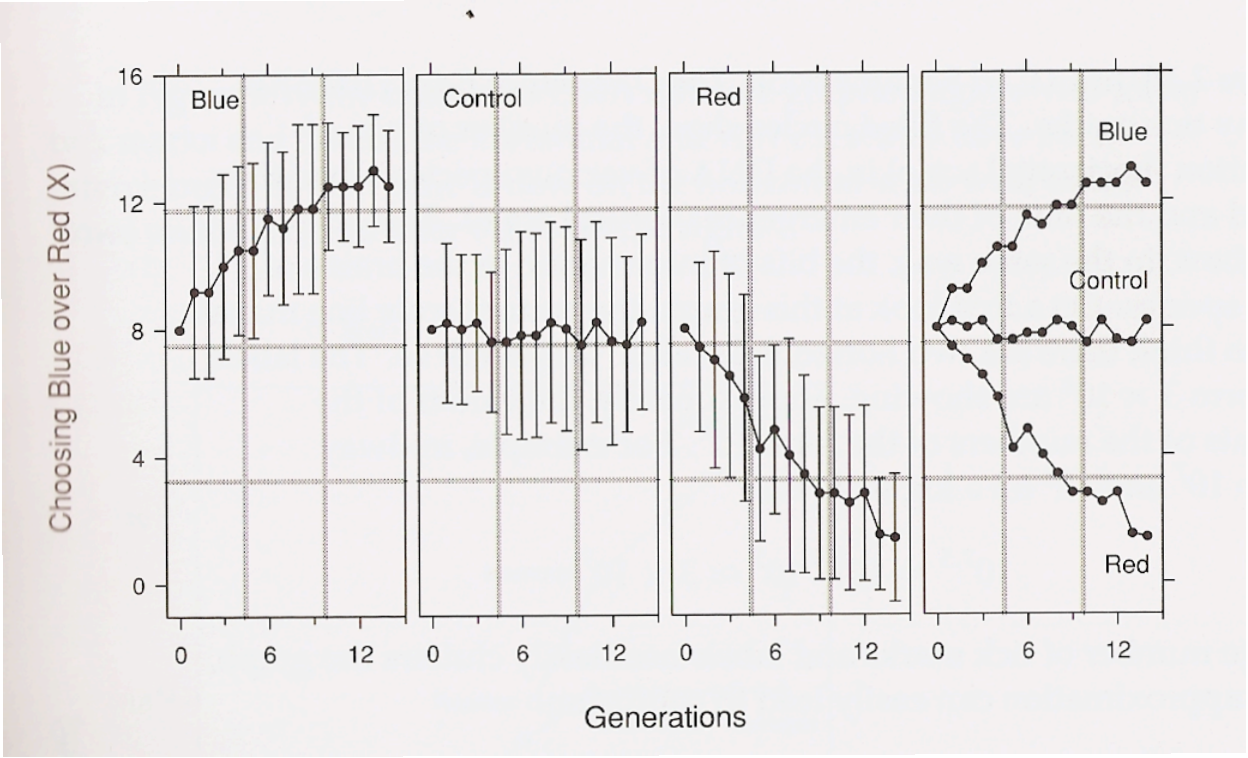
Clear Vision

- Principle 1: Make data stand out
  - Avoid superfluidity, clutter, or chartjunk.



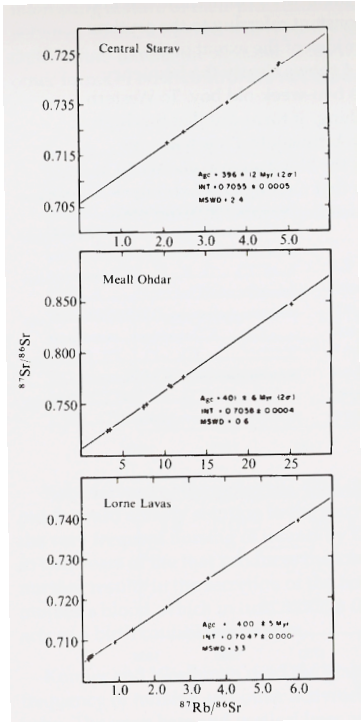
Clear Vision

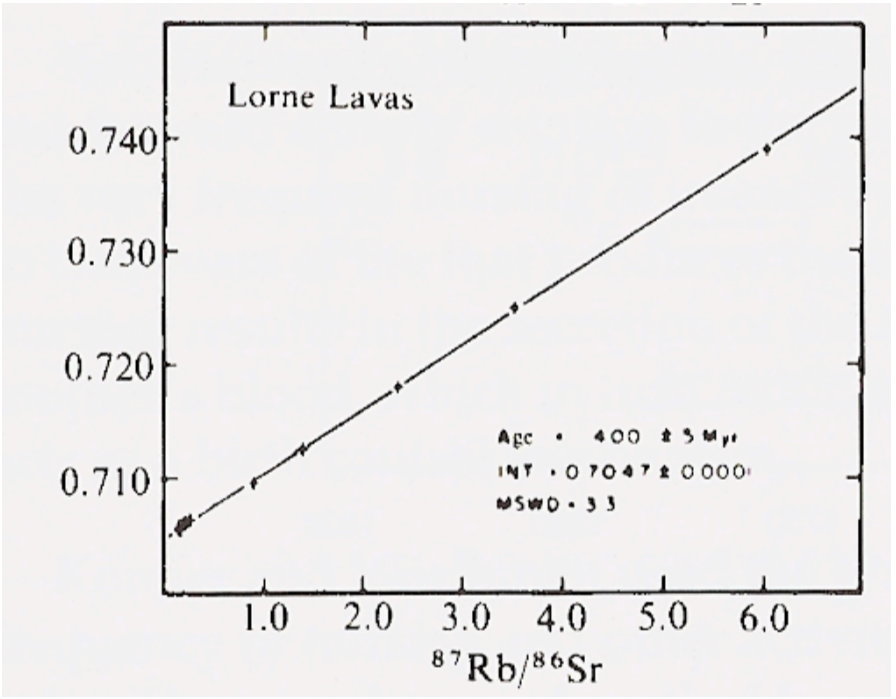
- Principle 1: Make data stand out
  - Avoid superfluidity, clutter, or chartjunk.



Clear Vision

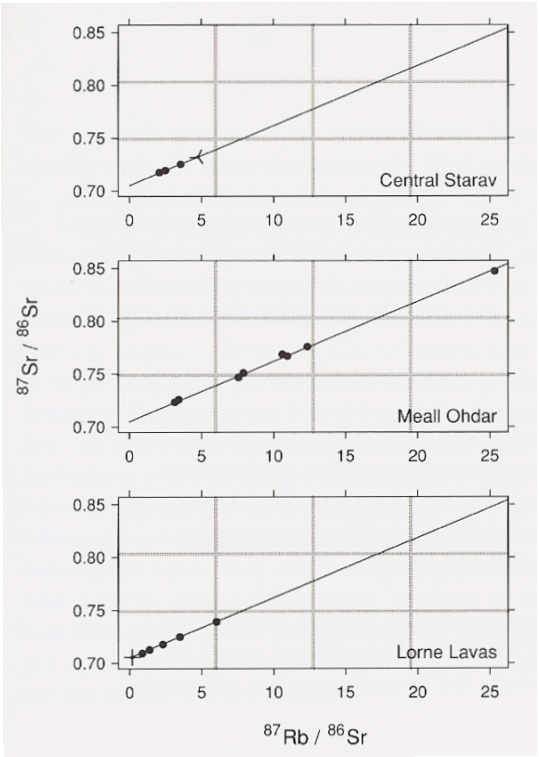
- Principle 2: Visual prominence
  - Use visually prominent graphical elements to show the data.

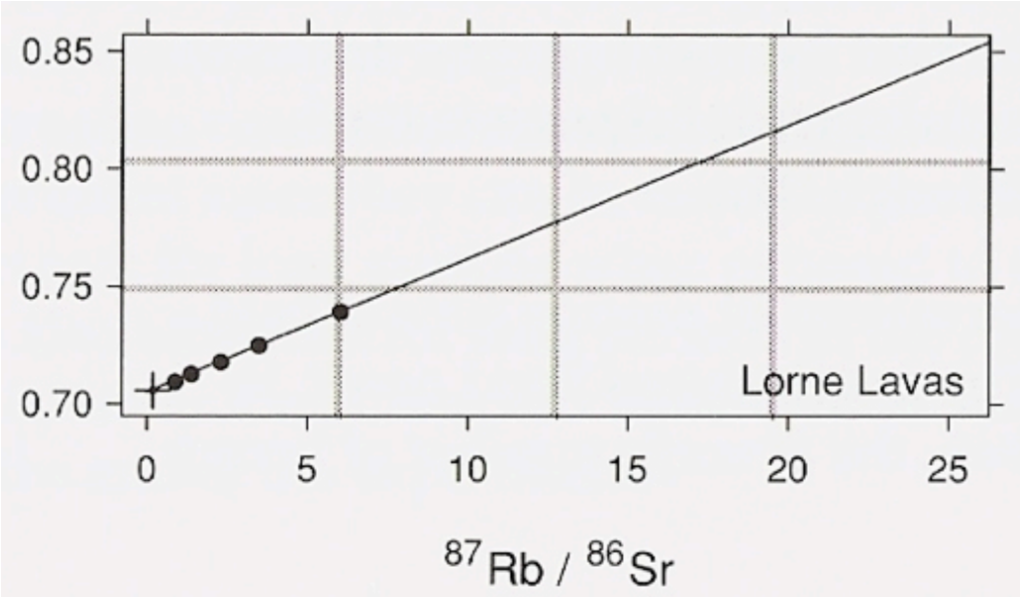




Clear Vision

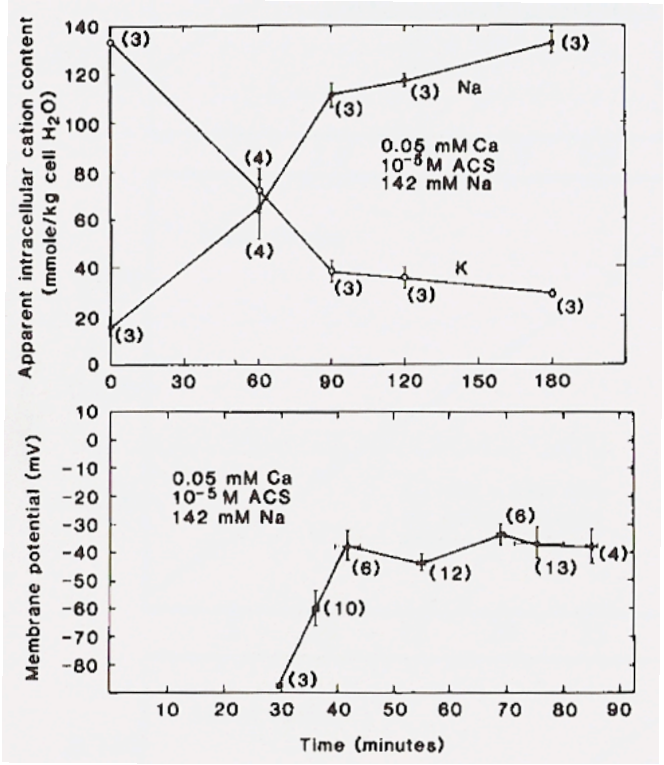
- Principle 2: Visual prominence
  - Use visually prominent graphical elements to show the data.

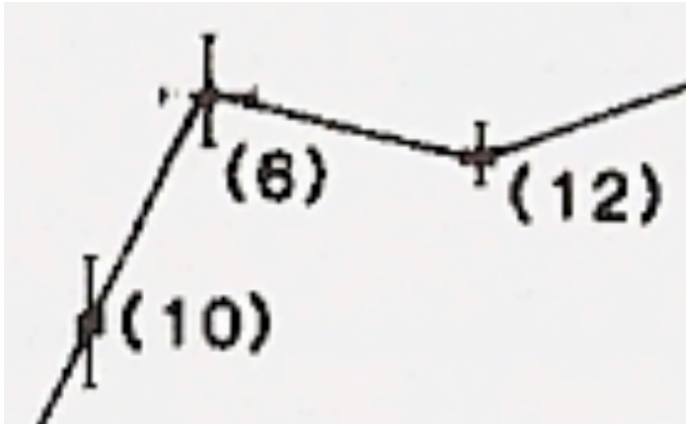




Clear Vision

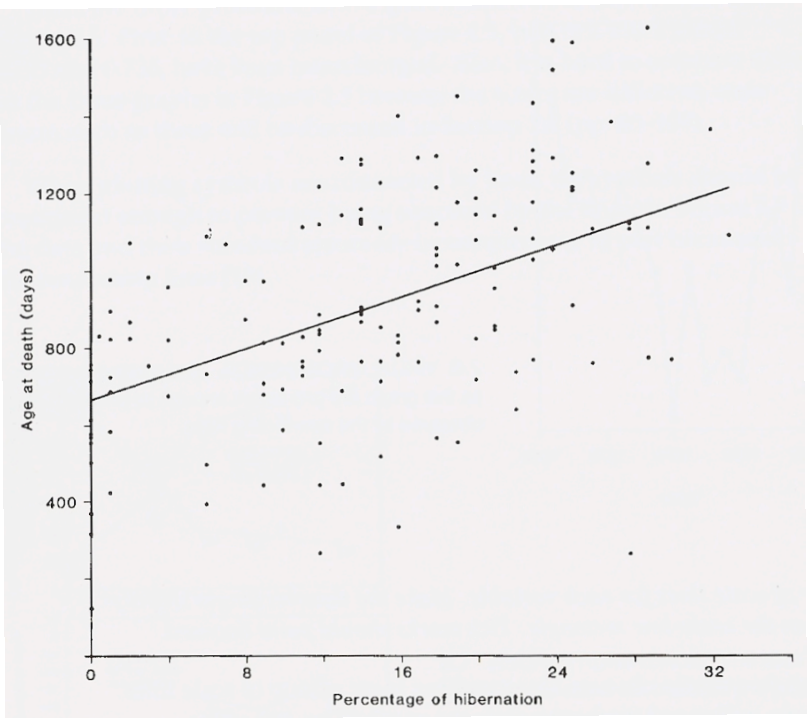
- Principle 2: Visual prominence
  - Use visually prominent graphical elements to show the data.





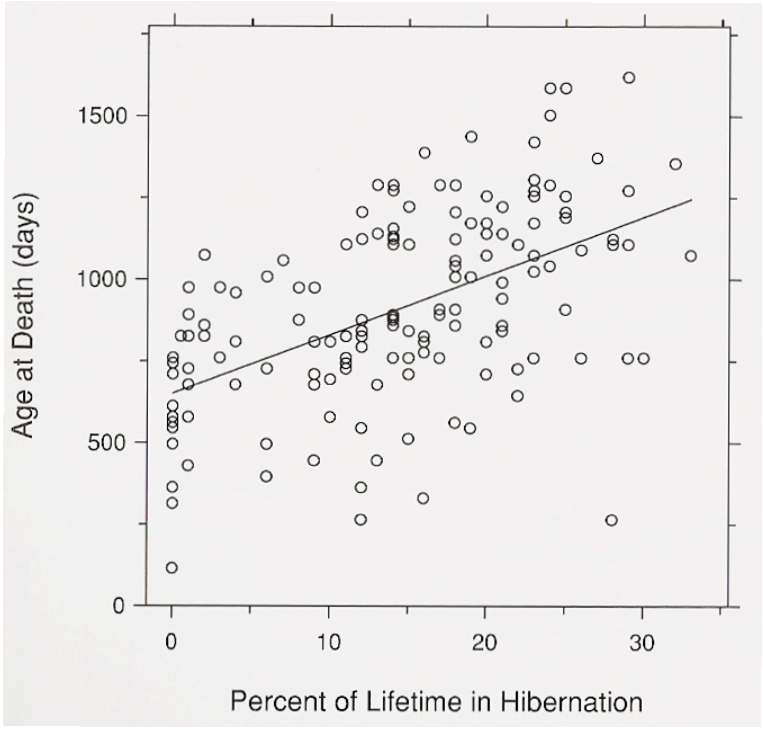
Clear Vision

- Principle 3: Scale lines and the data rectangle
  - Use two scale lines (box), add margins for data, tick-marks out, 3-10 tick marks.



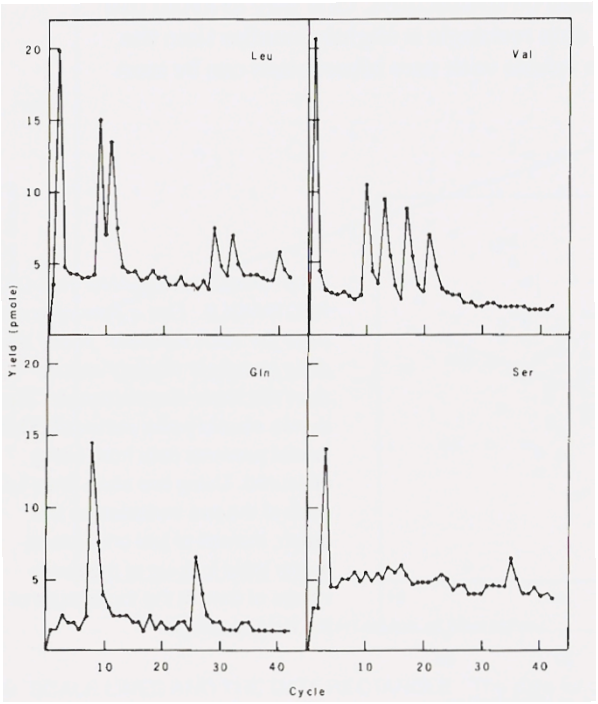
Clear Vision

- Principle 3: Scale lines and the data rectangle
  - Use two scale lines (box), add margins for data, tick-marks out, 3-10 tick marks.



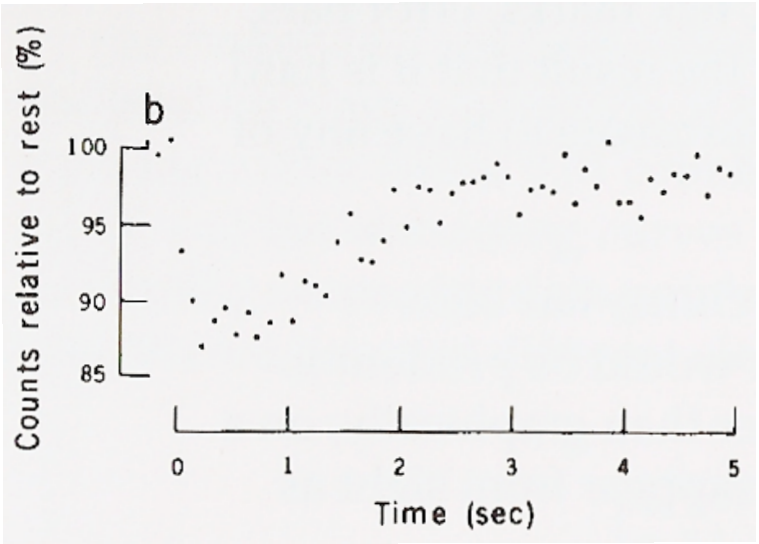
Clear Vision

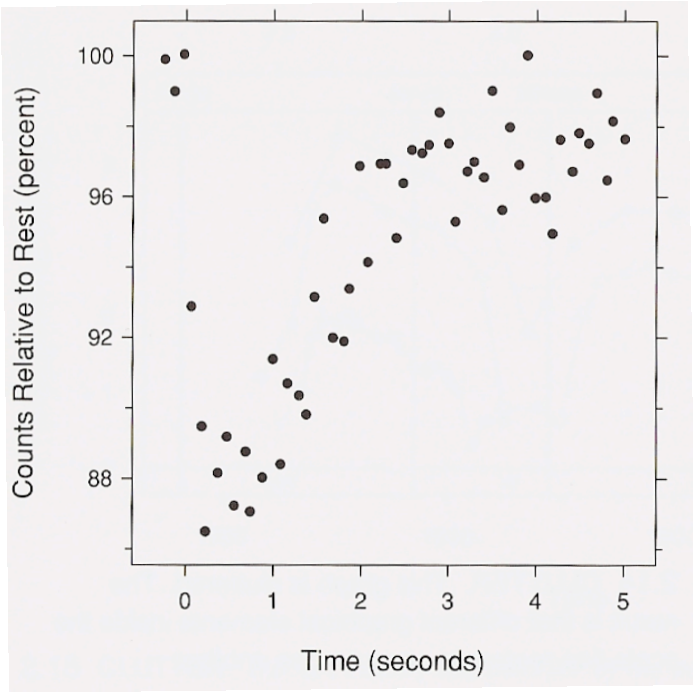
- Principle 3: Scale lines and the data rectangle
  - Use two scale lines (box), add margins for data, tick-marks out, 3-10 tick marks.



Clear Vision

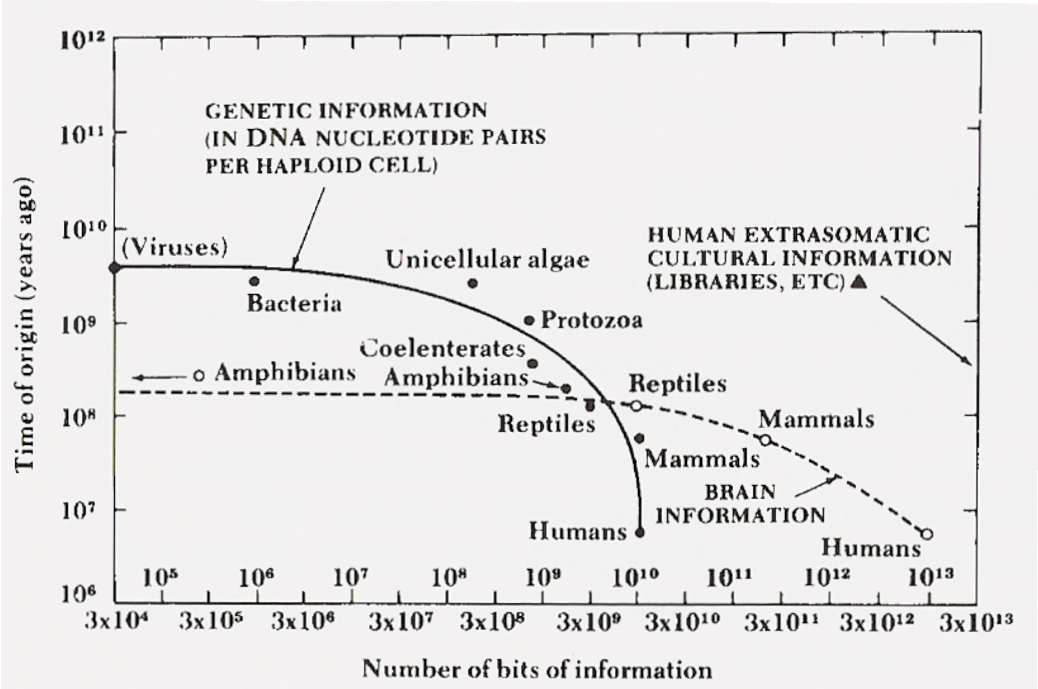
- Principle 3: Scale lines and the data rectangle
  - Use two scale lines (box), add margins for data, tick-marks out, 3-10 tick marks.





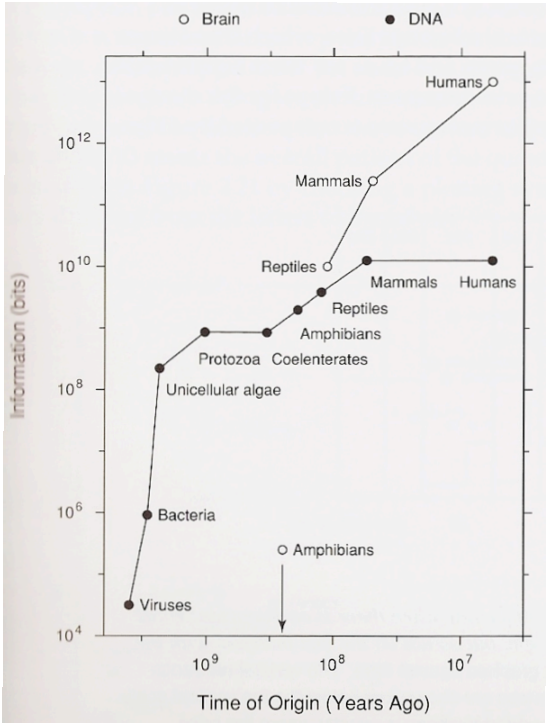
Clear Vision

- Principle 3: Scale lines and the data rectangle
  - Use two scale lines (box), add margins for data, tick-marks out, 3-10 tick marks.



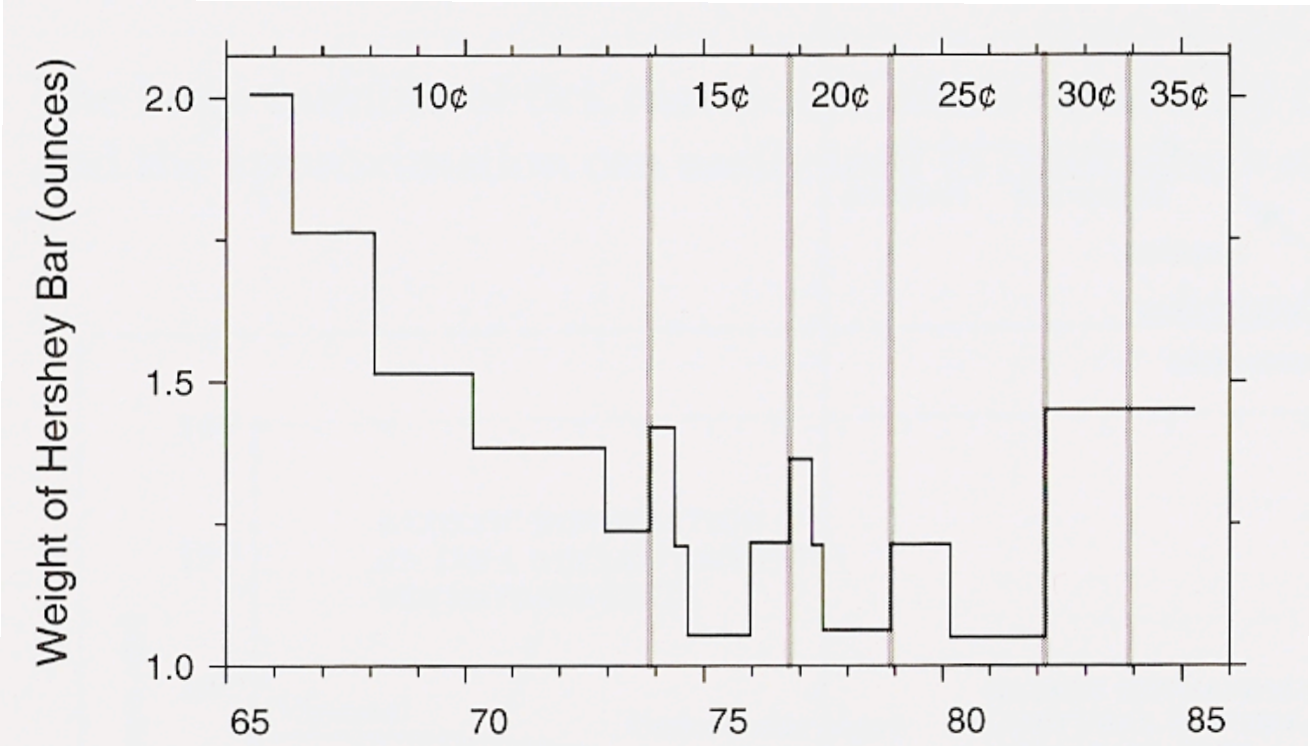
Clear Vision

- Principle 3: Scale lines and the data rectangle
  - Use two scale lines (box), add margins for data, tick-marks out, 3-10 tick marks.



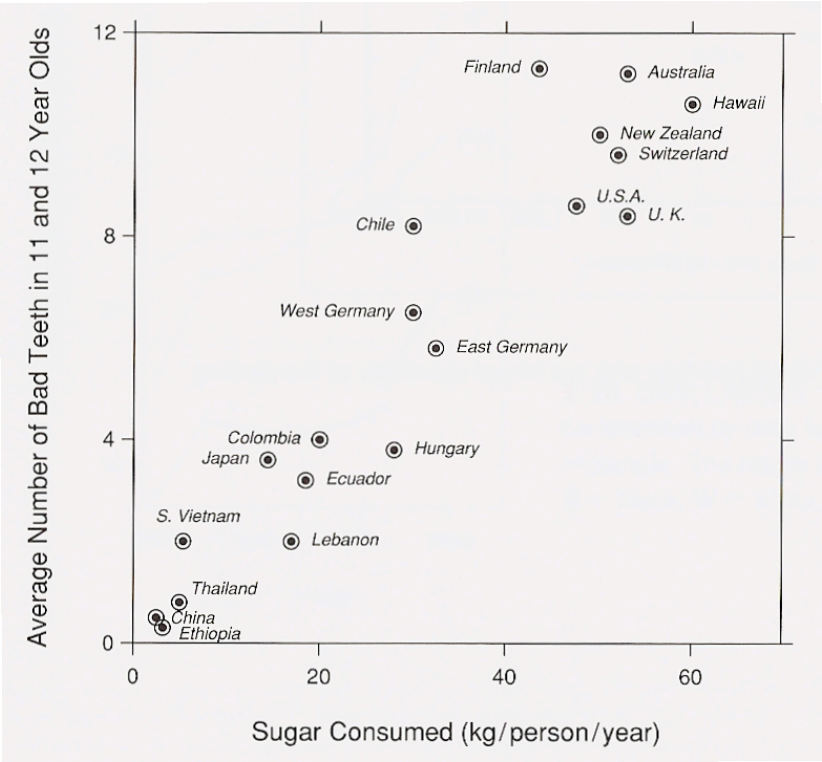
Clear Vision

- Principle 4: Reference lines, labels, notes, and keys
  - Only use when necessary and don’t let them obscure data.



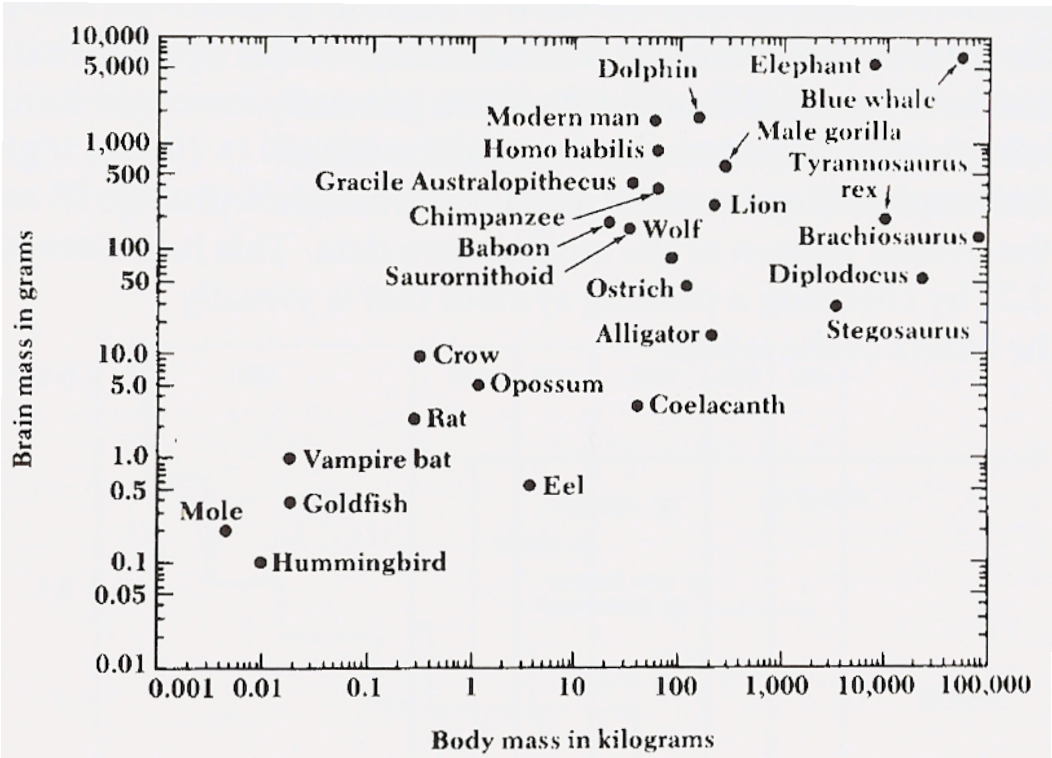
Clear Vision

- Principle 4: Reference lines, labels, notes, and keys
  - Only use when necessary and don't let them obscure data.



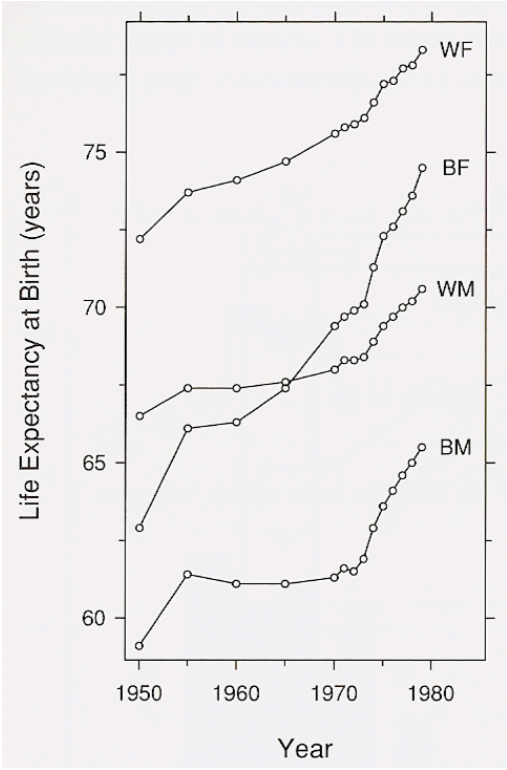
Clear Vision

- Principle 4: Reference lines, labels, notes, and keys
  - Only use when necessary and don't let them obscure data.



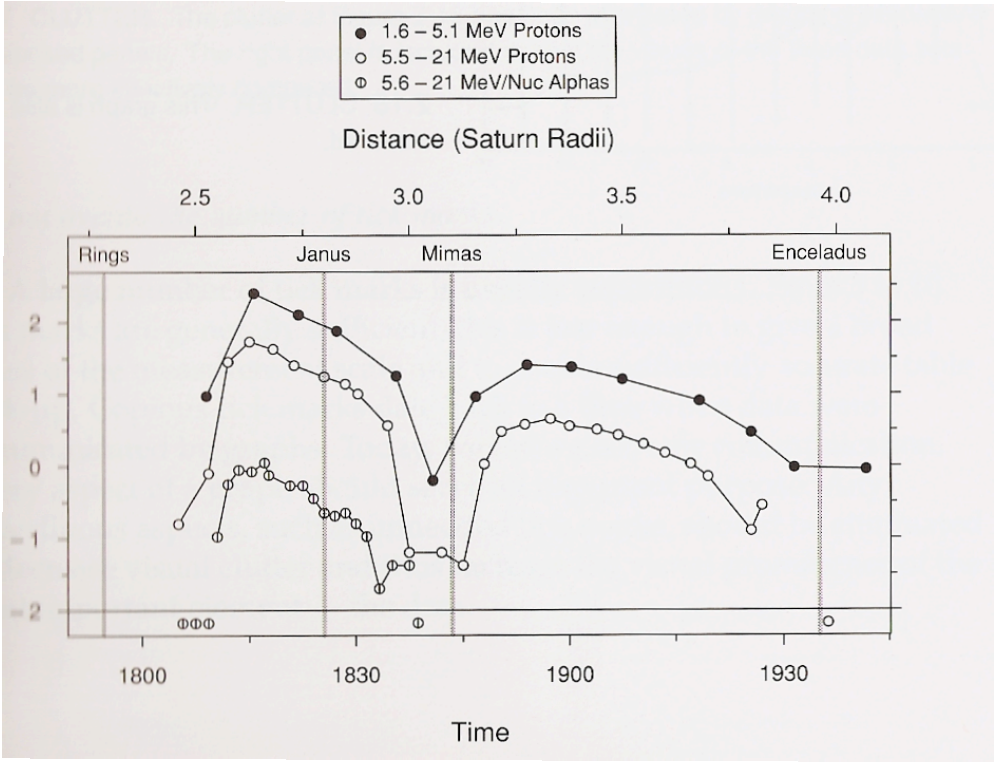
Clear Vision

- Principle 4: Reference lines, labels, notes, and keys
  - Only use when necessary and don’t let them obscure data.



Clear Vision

- Principle 4: Reference lines, labels, notes, and keys
  - Only use when necessary and don't let them obscure data.



Clear Vision

- Principle 4: Reference lines, labels, notes, and keys
  - Only use when necessary and don't let them obscure data.

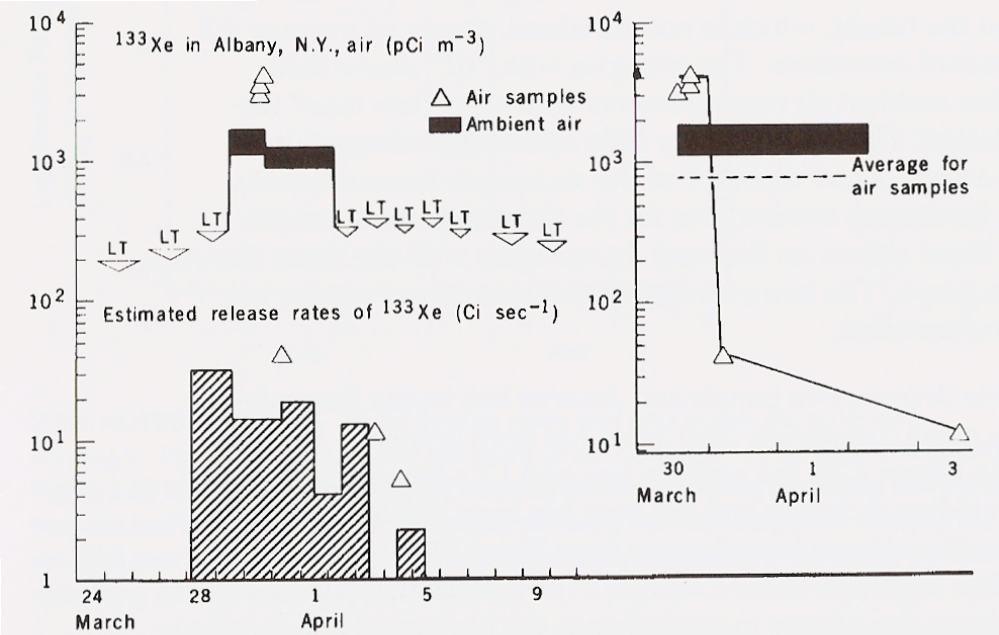
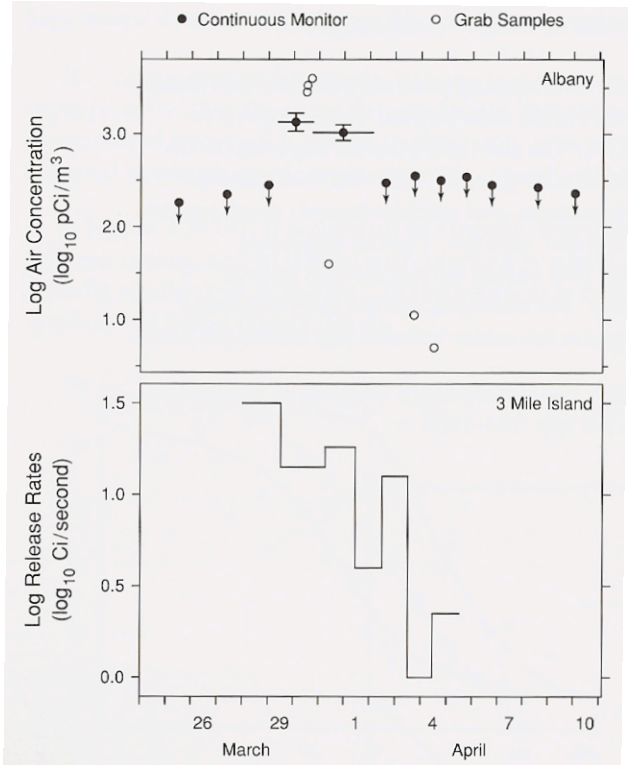


Fig. 1. Xenon-133 activity (picocuries per cubic meter of air) in Albany, New York, for the end of March and early April 1979. The lower trace shows the time-averaged estimates of releases (curies per second) from the Three Mile Island reactor (*2*). The inset shows detailed values for air samples (gas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

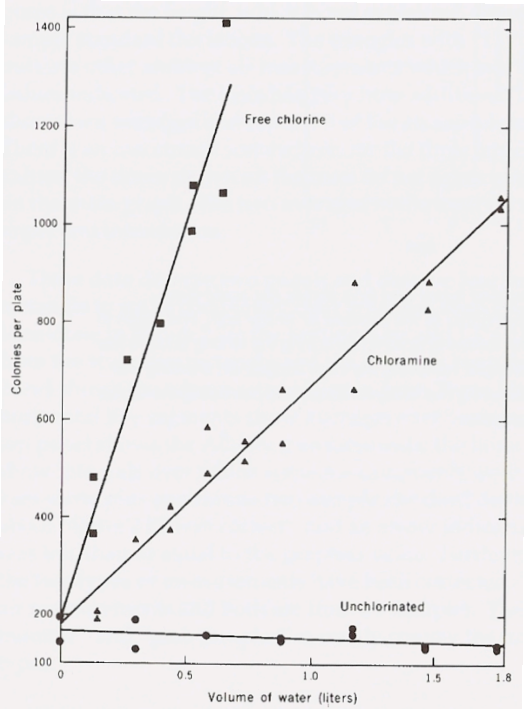
Clear Vision

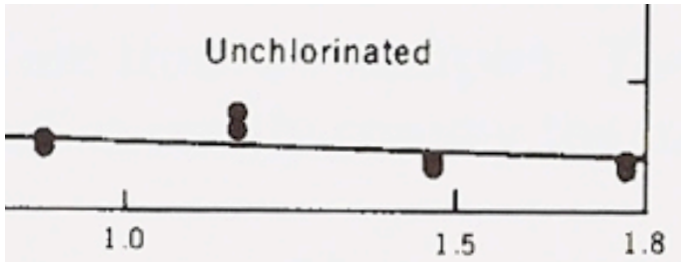
- Principle 4: Reference lines, labels, notes, and keys
  - Only use when necessary and don’t let them obscure data.



Clear Vision

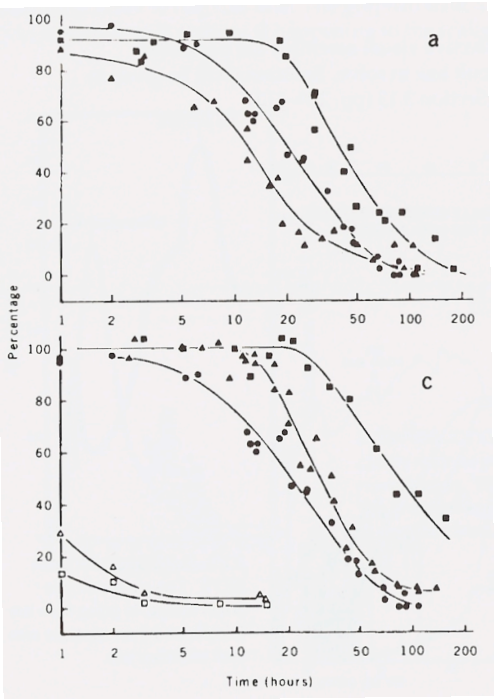
- Principle 5: Superposed data sets
  - Symbols should be separable and data sets should be easily visually assembled.





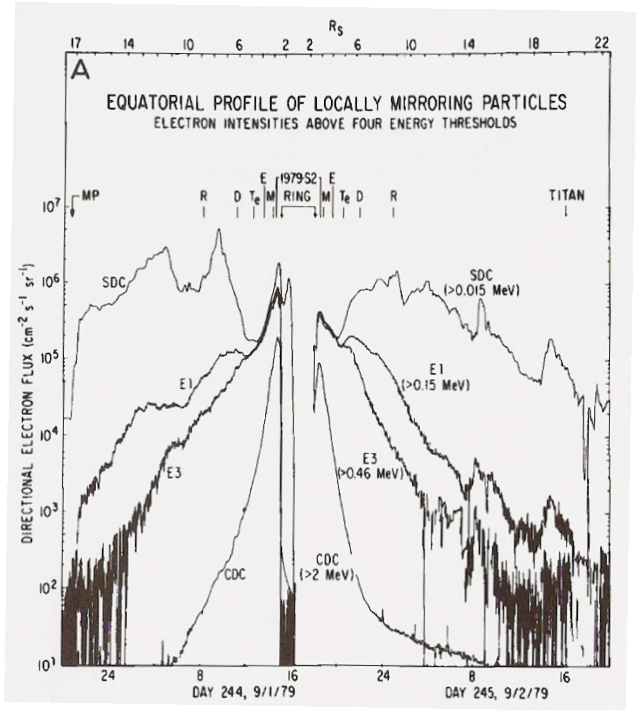
Clear Vision

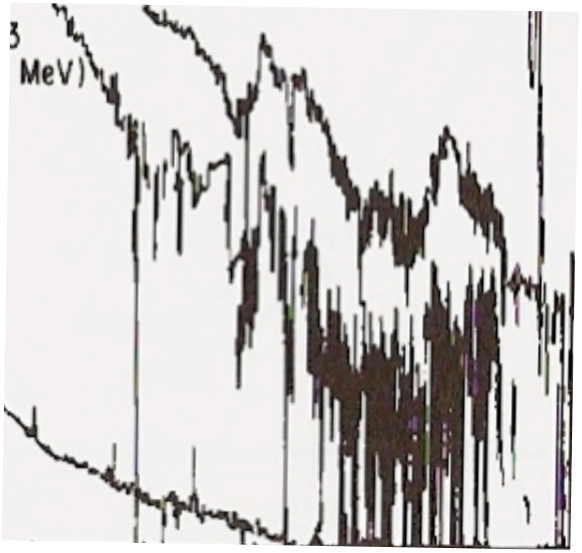
- Principle 5: Superposed data sets
  - Symbols should be separable and data sets should be easily visually assembled.



Clear Vision

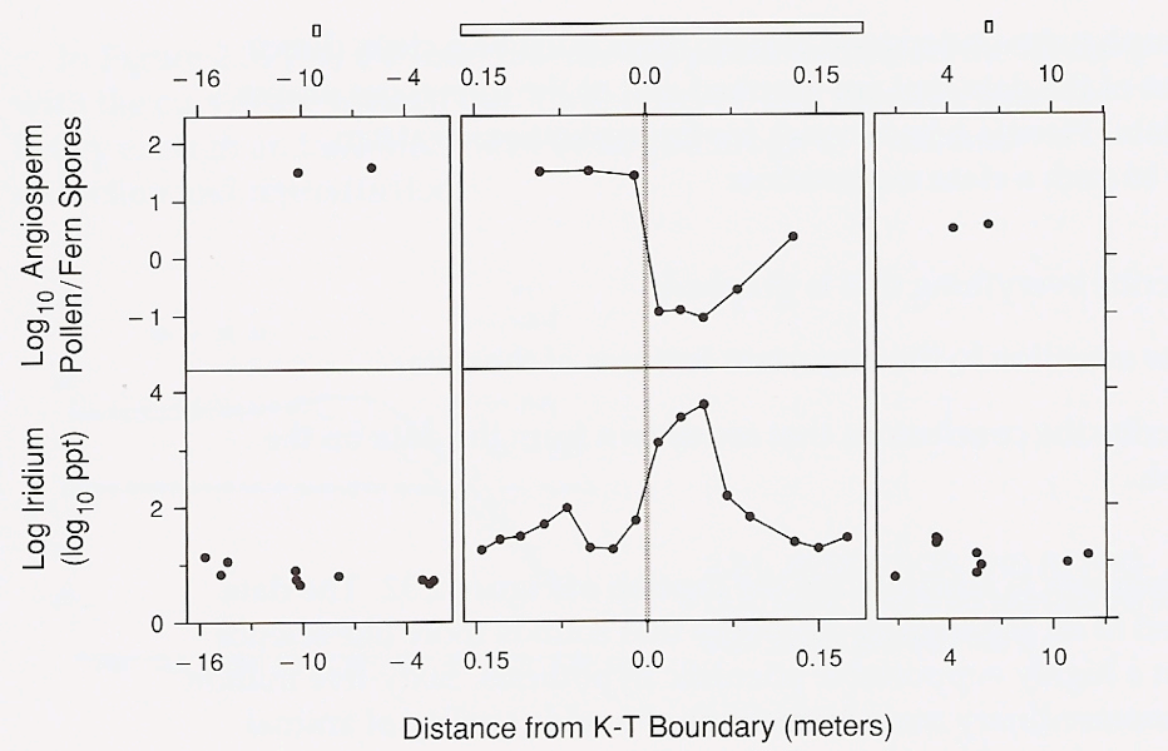
- Principle 5: Superposed data sets
  - Symbols should be separable and data sets should be easily visually assembled.





Clear Understanding

- Principle 1: Explanation and conclusions
  - Describe everything, draw attention to major features, describe conclusions



Clear Understanding

- Principle 1: Explanation and conclusions
  - Describe everything, draw attention to major features, describe conclusions

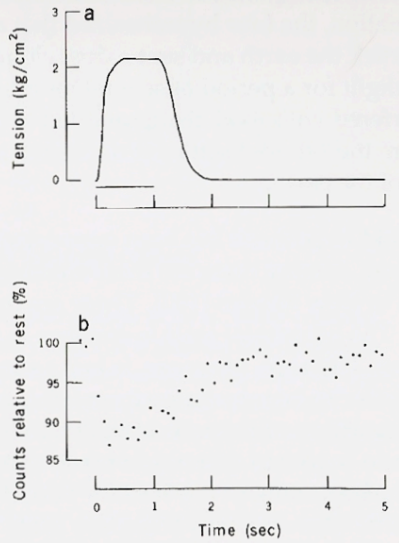


Fig. 2. Tension and the intensity of the 42.9-nm layer line during 1-second tetanus at the sarcomere length of 2.2 μm. (a) Tension record averaged over the 40 tetanic contractions required for obtaining the time course of the layer-line intensity. A sartorius muscle was dissected from *Rana catesbeiana* and tetanized for 1 second at 2-minute intervals. The horizontal line represents the period of stimulation. Tension was recorded with an isometric tension transducer (Shinkoh, type UL). (b) Intensity of the first-order myosin layer line at 42.9 nm. The x-ray source was a rotating-anode generator (Rigaku FR) with a fine focus (1.0 by 0.1 mm) on a copper target. This was operated at 50 kV with a tube current of 70 mA; such a high power was possible with an anode of a large diameter (30 cm) rotating at a high speed (9000 rev/min). A bent-crystal monochromator was used at a source-to-crystal distance of 25 cm with a viewing angle of 6°. The intensity of the myosin layer line was measured with a scintillation counter combined with a mask; the mask had apertures at the positions of the off-meridional parts of the first-order layer line. The meridional reflection at 14.3 nm is known to be slightly displaced during contraction, suggesting a minute change in the myosin periodicity (*1*, *3*). It is, therefore, possible that the 42.9-nm layer line is also slightly displaced. However, the possible displacement (14 μm at the position of the mask) would be insignificant compared with the width of each aperture (0.8 mm). The intensity measured at the resting state was 1400 count/sec. The intensities during and after tetanus were expressed as percentages of the resting intensity and plotted against time after the first stimulus of each set of stimuli. Each point represents the intensity averaged over a 100-msec period. The first three points represent the measurements made before stimulation.

Clear Understanding

- Principle 1: Explanation and conclusions
  - Describe everything, draw attention to major features, describe conclusions

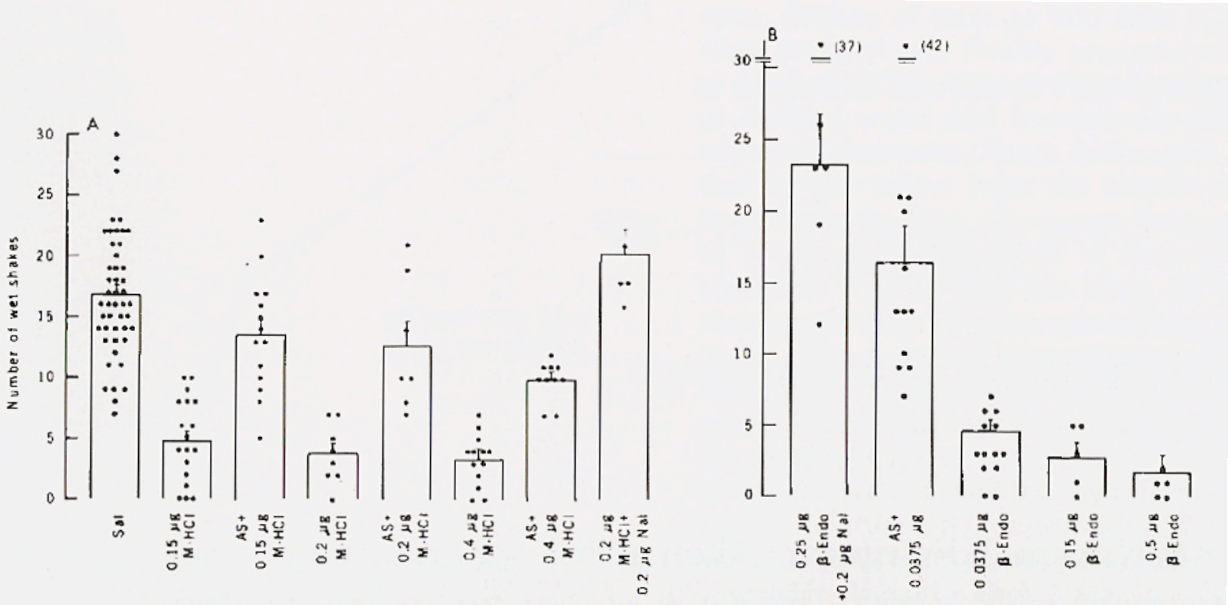
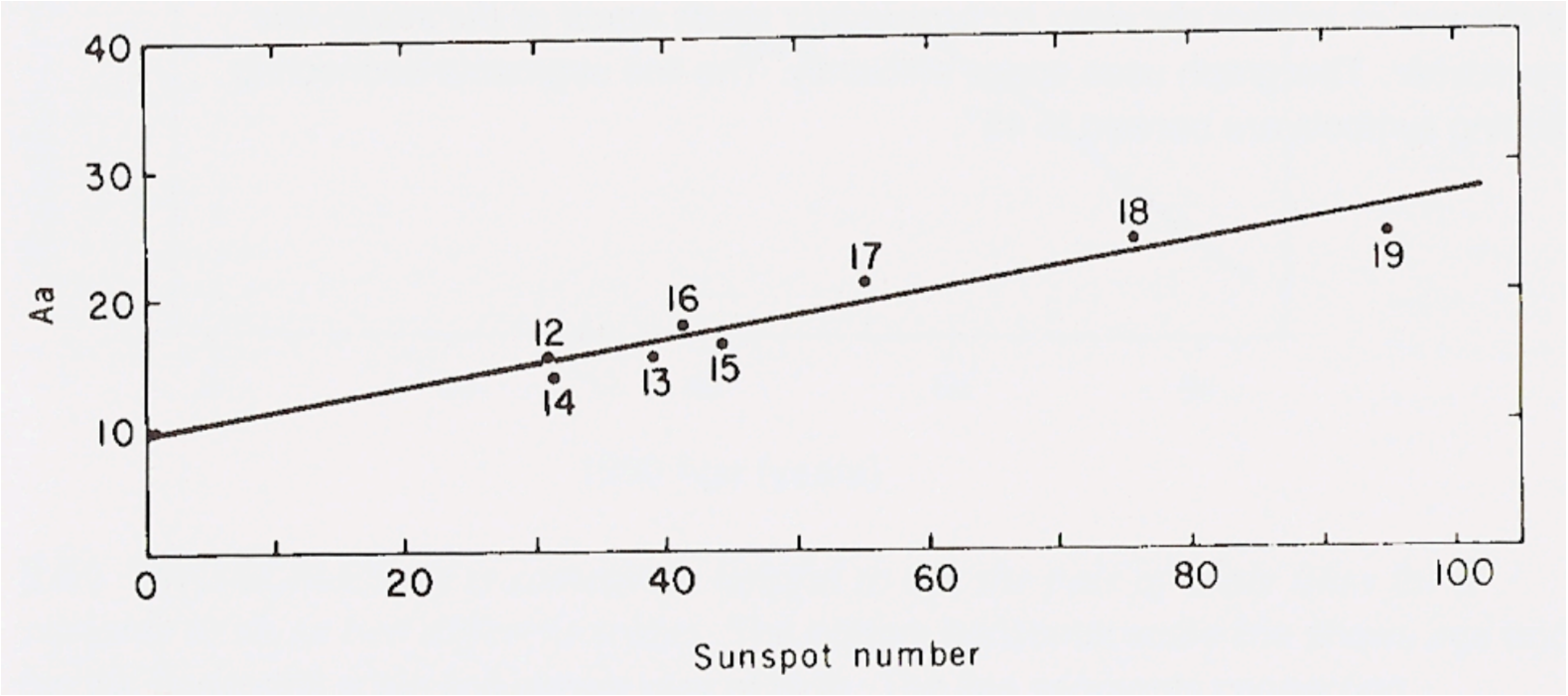


Fig. 1. Inhibitory effect of morphine hydrochloride (A) and β-endorphin (B) on wet-shake behavior in rats. Antagonism was by antibody to cerebroside sulfate (*AS*). A volume of 2 μl was delivered into the PAG in 1-μl increments with a 1-minute interval in between. The control consisted of saline (*Sal*). Morphine HCl (*M-HCl*) and β-endorphin were preceded by saline, AS, or naloxone (*Nal*) as indicated. The dose of morphine HCl and naloxone refers to the chloride salt. Each point is the datum for one animal. The AS + morphine HCl groups are all significantly different (*t*-test, $P < .005$) from the corresponding morphine HCl groups alone (A). The group receiving AS + 0.0375 μg of β-endorphin is also significantly different ($P < .005$) from the group receiving 0.0375 μg of β-endorphin (B).

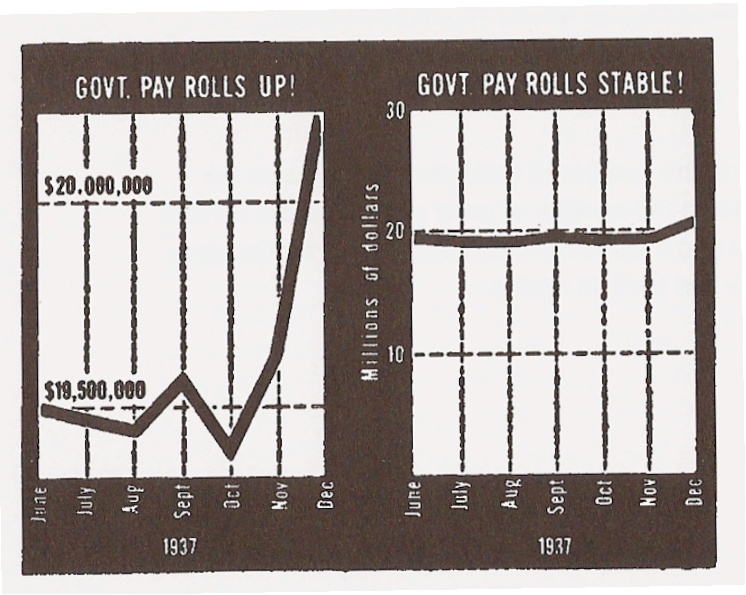
Clear Understanding

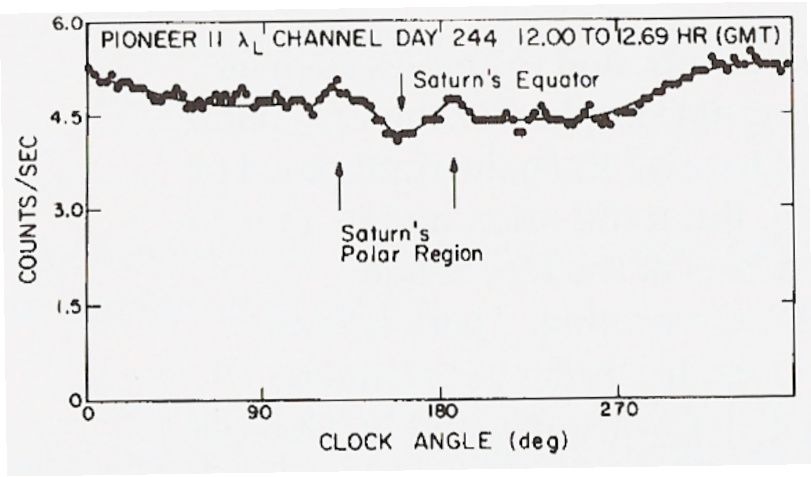
- Principle 2: Use all of the available space
  - Fill the data rectangle, only use zero if you need it



Clear Understanding

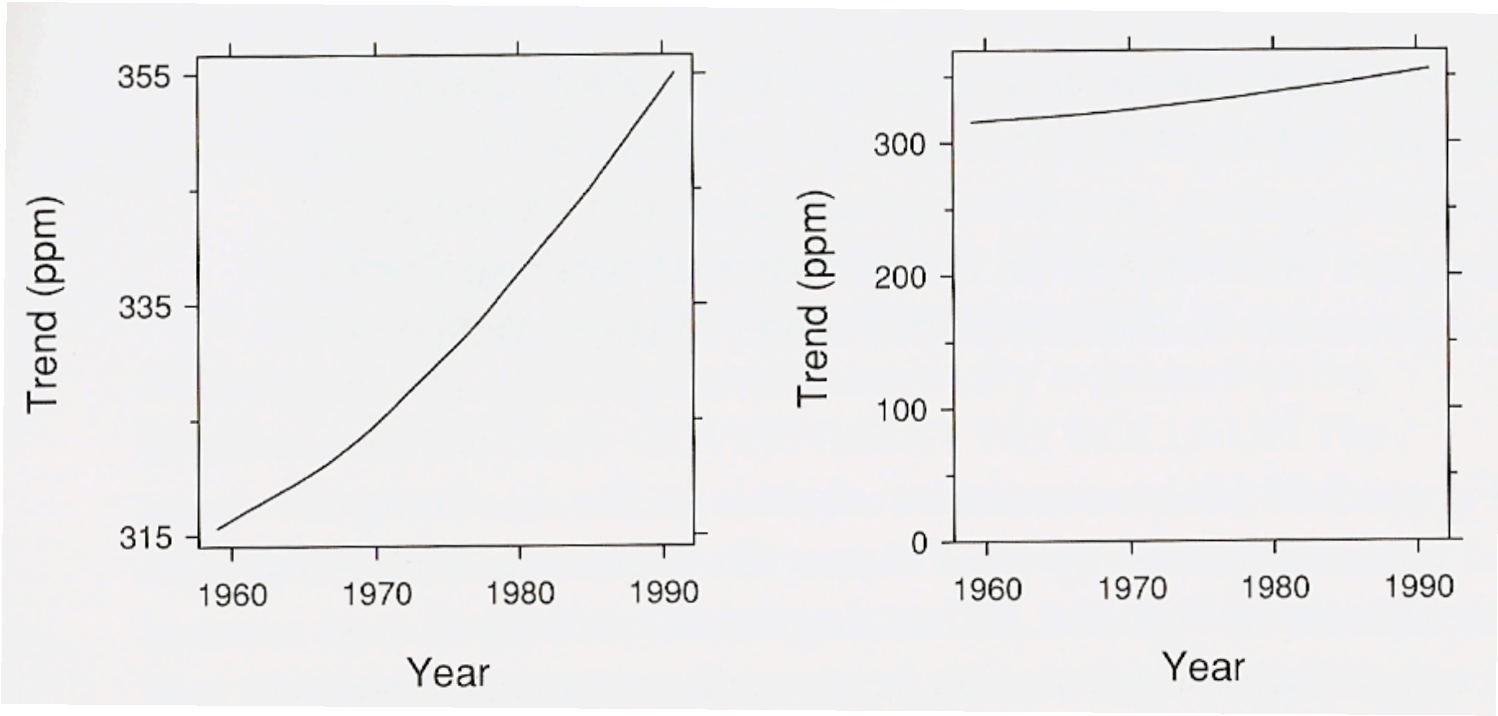
- Principle 2: Use all of the available space
  - Fill the data rectangle, only use zero if you need it





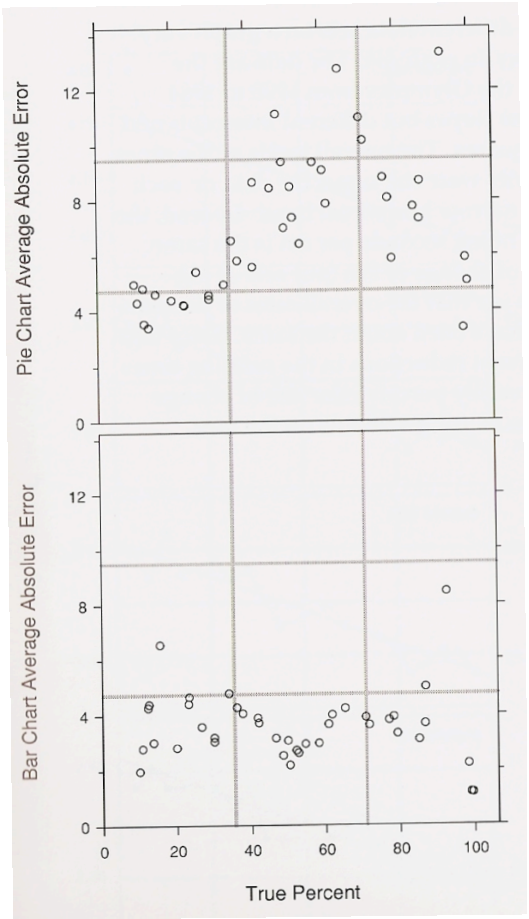
Clear Understanding

- Principle 2: Use all of the available space
    - Fill the data rectangle, only use zero if you need it



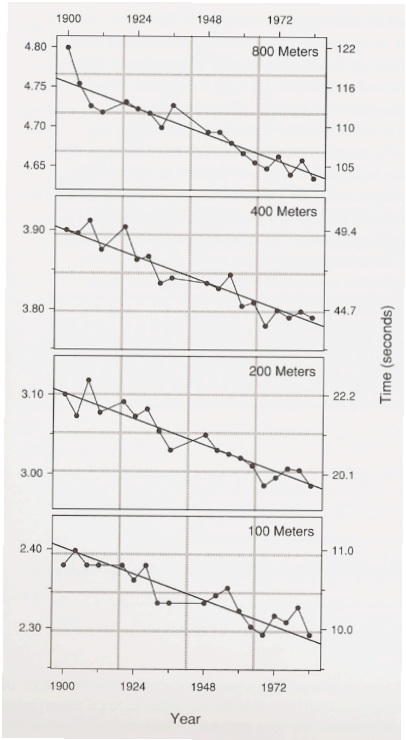
Clear Understanding

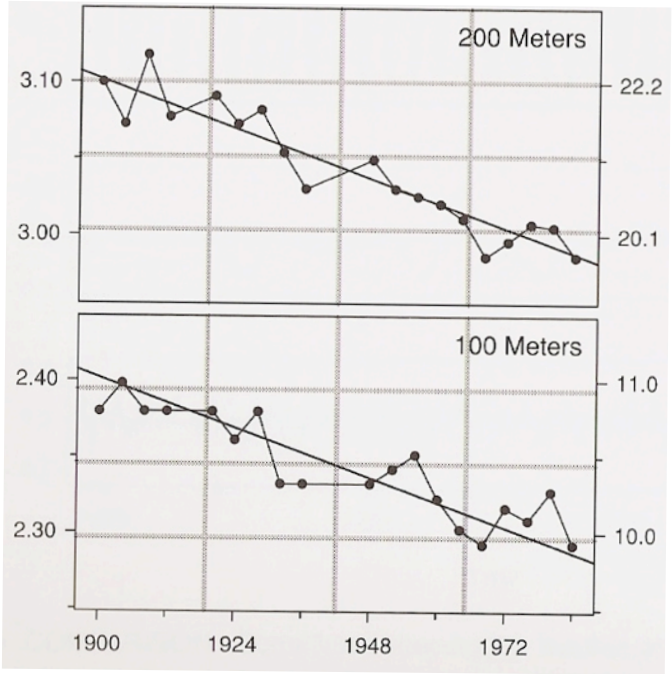
- Principle 3: Juxtaposed data sets
  - Make sure scales match and graphs are aligned



Clear Understanding

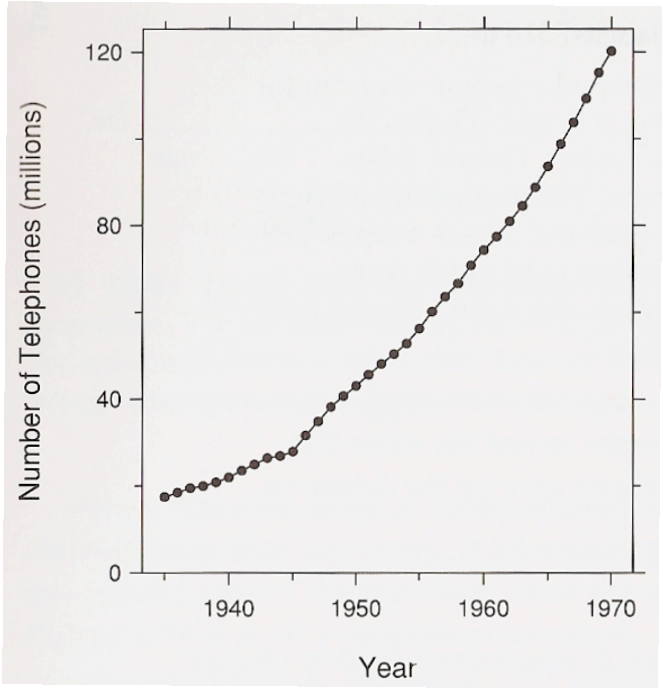
- Principle 3: Juxtaposed data sets
  - Make sure scales match and graphs are aligned

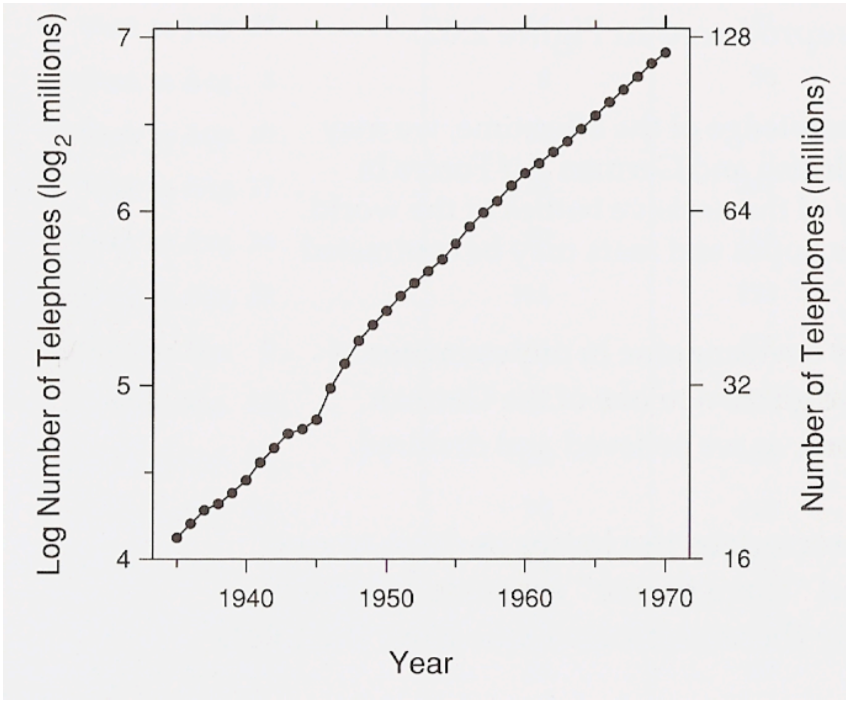




Clear Understanding

- Principle 4: Log scales
  - Used to show percentage change, multiplicative factors and skewness





Clear Understanding

- Principle 4: Banking to 45°
  - Aspect ratio is important for judging rate of change

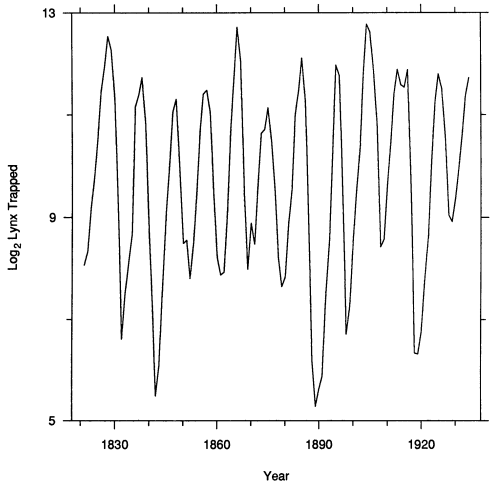


Figure 2. The Effect of Shape. The data are the Canadian lynx trappings from 1821 to 1934. The shape parameter is 1. The orientations of the line segments connecting successive data points are too close to 90° and −90° to allow us to see an important property of the data.

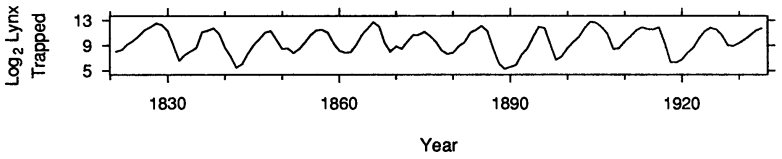


Figure 3. The Effect of Shape. The shape parameter of the graph is .074. The orientations of the line segments are in a range that allows better visual decoding of the slopes. We can now see what we could not see in Figure 2—the numbers tend to rise more slowly than they fall.

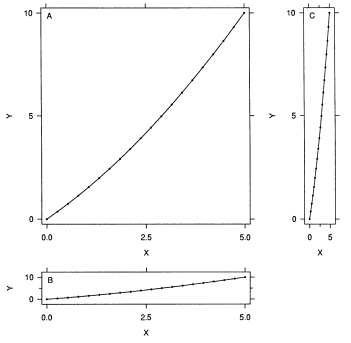


Figure 5. Orientation Resolution. The same data are graphed in all three panels. The orientation resolution is 18.4° in Panel A, 4.0° in Panel B, and 4.0° in Panel C. In Panels B and C it is difficult to perceive the curvature that is apparent in Panel A, because the resolutions are so much smaller. Orientation resolution is maximized when the midangle is 45°.

https://www.dropbox.com/s/vs1ry31zpc0kwhj/2288843.pdf?dl=0

- Principle 4: Banking to 45°
  - Aspect ratio is important for judging rate of change

IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 12, NO. 5, SEPTEMBER/OCTOBER 2006

Multi-Scale Banking to 45º

Jeffrey Heer and Maneesh Agrawala

Abstract—In his text *Visualizing Data*, William Cleveland demonstrates how the aspect ratio of a line chart can affect an analyst's perception of trends in the data. Cleveland proposes an optimization technique for computing the aspect ratio such that the average absolute orientation of line segments in the chart is equal to 45 degrees. This technique, called *banking to 45°*, is designed to maximize the discriminability of the orientations of the line segments in the chart. In this paper, we revisit this classic result and describe two new extensions. First, we propose alternate optimization criteria designed to further improve the visual perception of line segment orientations. Second, we develop *multi-scale banking*, a technique that combines spectral analysis with banking to 45°. Our technique automatically identifies trends at various frequency scales and then generates a banked chart for each of these scales. We demonstrate the utility of our techniques in a range of visualization tools and analysis examples.

Index Terms—Information visualization, banking to 45 degrees, line charts, time-series, sparklines, graphical perception

1 Introduction

Few visualizations are more common than the line chart, used to communicate changes in value over a contiguous domain, with the slopes of line segments encoding rates of change. Mathematical functions, stock prices, and all varieties of time-series data are plotted in this form. Accordingly, techniques which improve the graphical perception of such displays could be of great practical benefit. Given the human visual system's sensitivity to line orientation [6], it is not surprising that the relative orientation of line segments in a chart can strongly impact an analyst's perception of trends in the data. One way of intentionally manipulating these orientations is through the choice of the *aspect ratio* (width/height) of the chart.

Both in his book *Visualizing Data* [2] and elsewhere [1,3,4], William Cleveland demonstrates how the choice of a line chart's aspect ratio can impact graphical perception. Figure 1 shows plots of average monthly carbon dioxide measurements made at the Mauna Loa observatory, an example used in [2]. The first plot (1a) clearly shows an upward trend in the data. There is an inflection in the curve, indicating that the increase in carbon dioxide is accelerating. The second chart (1b) shows the same data plotted at a wider aspect ratio. The slow onset and quick decay of the yearly oscillations is clearly visible in this plot. However the inflection that was visible in the first chart becomes difficult to see.

Noting that an average orientation of 45° maximizes the discriminability of adjacent line segments, Cleveland introduces a technique called *banking to 45°*, which determines the aspect ratio such that the average orientation of all line segments in a chart is 45 degrees [2,3,4]. While Cleveland's optimization procedure makes it easier to see higher frequency oscillations in the data, it can obscure lower-frequency trends of interest. The aspect ratio used in Figure 1b is the result of banking to 45°. As we have noted, the low-frequency inflection point is difficult to discern in this chart. Cleveland addresses this issue, describing a manual process of fitting smooth regression curves to the data and banking the resulting low frequency curve. Finding interesting trends in the data becomes a tediously iterative trial-and-error process. Users must manually consider each smoothness level (*i.e.*, the frequency scale), bank it to 45º, and then visually check for an interesting trend.

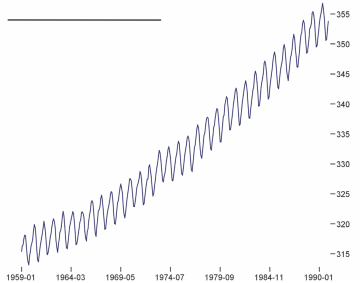


Figure 1a. CO_2 measurements, aspect ratio = 1.17. The x-axis shows time, in monthly increments, while the y-axis shows carbon dioxide measurements taken at the Mauna Loa observatory [2]. Note the bend in the trend of increasing values, suggesting an accelerating increase.

Figure 1b. CO_2 measurements, aspect ratio = 7.87. The wider aspect ratio enables the viewer to see that the ascent of each yearly cycle is more gradual than its decay. However, the bend in the lower-frequency trend is now difficult to see. The choice of aspect ratios for Figures 1a and 1b were automatically determined using multi-scale banking.

In this paper we extend Cleveland's work in two ways. First, we explore alternate optimization criteria for Cleveland's banking procedure. These criteria are designed to find an aspect ratio that further improves the visual perception of line segment orientations. Second, we develop *multi-scale banking*, a technique that combines spectral analysis with banking to 45°. Our technique automatically identifies frequency scales that may be of interest and then generates a banked chart for each of these scales. The aspect ratios for the charts in Figures 1a and 1b were automatically chosen using our multi-scale banking approach. While our technique finds the same aspect ratio as Cleveland's original approach in Figure 1b, it also returns the aspect ratio which reveals the low-frequency inflection point in Figure 1a. We have incorporated multi-scale banking into tools for both static and dynamic visualizations, and present the results of applying the technique on a collection of real-world data sets.

- *Jeffrey Heer is with the Computer Science Division of the University of California, Berkeley. E-Mail: jheer@cs.berkeley.edu.*
- *Maneesh Agrawala is with the Computer Science Division of the University of California, Berkeley. E-Mail: maneesh@cs.berkeley.edu.*

Manuscript received 31 March 2006; accepted 1 August 2006; posted online 6 November 2006. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

https://www.dropbox.com/s/a0ajc6l29nntsi3/2006-Banking-InfoVis.pdf?dl=0

https://www.dropbox.com/s/zroyl5vfkpi47ze/2012-SlopeComparison-InfoVis.pdf?dl=0

An Empirical Model of Slope Ratio Comparisons

Justin Talbot, John Gerth, and Pat Hanrahan

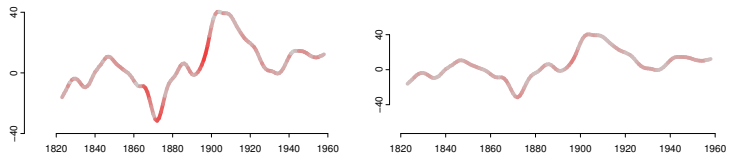


Fig. 1. Both plots show the same data set—the change in the length of a day (in microseconds) over 140 years—with different aspect ratios. The segment redness corresponds to the error that viewers will make when comparing the slope of that segment to all other slopes in the plot—as predicted by our empirical model. The aspect ratio on the right minimizes the total absolute predicted error.

Abstract—Comparing slopes is a fundamental graph reading task and the aspect ratio chosen for a plot influences how easy these comparisons are to make. According to *Banking to 45°*, a classic design guideline first proposed and studied by Cleveland et al., aspect ratios that center slopes around 45° minimize errors in visual judgments of slope ratios. This paper revisits this earlier work. Through exploratory pilot studies that expand Cleveland et al.'s experimental design, we develop an empirical model of slope ratio estimation that fits more extreme slope ratio judgments and two common slope ratio estimation strategies. We then run two experiments to validate our model. In the first, we show that our model fits more generally than the one proposed by Cleveland et al. and we find that, in general, slope ratio errors are *not* minimized around 45°. In the second experiment, we explore a novel hypothesis raised by our model: that visible baselines can substantially mitigate errors made in slope judgments. We conclude with an application of our model to aspect ratio selection.

Index Terms—Banking to 45 degrees, slope perception, orientation resolution, aspect ratio selection.

1 Introduction

Banking to 45° is a classic design guideline in information visualization and statistical graphics due to Cleveland, McGill, and McGill [5, 4, 3]. It recommends that the aspect ratio of a plot be chosen such that the slopes of the plot's line segments are centered around 45°. The guideline has been widely and successfully used by visualization designers to manually select aspect ratios. And it has inspired a variety of algorithms [5, 3, 10, 15] that automate this task.

Despite the practical success of this guideline, its perceptual underpinnings remain unclear. Cleveland et al. justified the guideline with an experiment that showed that placing the mid-angle of two lines (the angle halfway between them) at 45° minimizes errors made in judging the ratio of their slopes. However, examination of their experimental design suggests that this conclusion might not be generally applicable. First, their experiment only tested moderate slope ratios and moderate mid-angles. It's unclear if their observed trends will hold for more extreme slope comparisons. Second, they restricted how their subjects made the slope ratio comparisons, instructing them to compare only the heights (*y*-extents) of the line segments. Thus, their results may not apply if other comparison methods are also used in practice.

This paper seeks to improve our understanding of slope ratio estimation in line plots through empirical modeling and experimentation. The paper is organized as follows. Section 2 provides details on the original study and other background material. In Section 3 we describe our methodological approach. Section 4 describes our pilot studies and Section 5 presents a new empirical model for slope ratio judgments. Sections 6 and 7 describe two experimental studies of the new model. The first study demonstrates that our new model fits observed data well. Additionally this model provides strong evidence that the minimum slope ratio estimation error does not occur at 45 degrees. The second study explores a novel hypothesis raised by the new model—that slope ratio estimates are dramatically affected by the presence or absence of a baseline. In Section 8 we briefly discuss the implication of the new model for aspect ratio selection. We conclude with a discussion of the limitations of our study and recommendations for future work.

2 Previous Work

This section described related work in three categories—perception, aspect ratio selection algorithms, and line chart visualization design.

2.1 Perception

In the Cleveland et al. study [5], 16 subjects were each asked to estimate the ratio (as a percentage) between the slopes of 44 pairs of lines with equal x-extents by comparing the y-extents of the two lines. The line pairs were chosen so that the true ratio percentage between their slopes (p_{ij}) varied between 50% and 100% and the slopes themselves varied from about 0.1 to slightly more than 1. Subjects were shown each pair of lines for only 2.5s encouraging quick estimates. The resulting estimates, $\hat{p}_{ij}$, were then used to fit an empirical model of the absolute estimation error [4]:

$$|\hat{p}_{ij} - p_{ij}| = 4.39 - 0.47(p_{ij} - 100) - 1.14r_{ij} + \varepsilon_{ij} \quad (1)$$

where p_{ij} is the true percentage and r_{ij} is the angle between the lines in degrees. The model indicated that the absolute error was minimized when the lines were centered around 45°.

- *Justin Talbot is with Stanford, e-mail: jtalbot@stanford.edu.*
- *John Gerth is with Stanford, e-mail: gerth@graphics.stanford.edu.*
- *Pat Hanrahan is with Stanford, e-mail: hanrahan@cs.stanford.edu.*

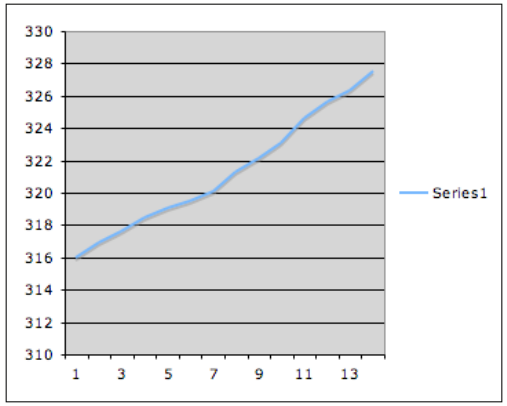
Manuscript received 31 March 2012; accepted 1 August 2012; posted online 14 October 2012; mailed on 5 October 2012.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

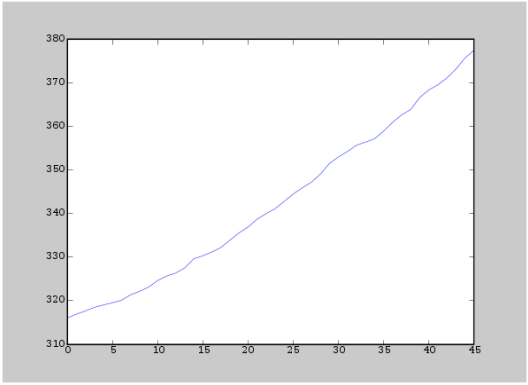
Summary of Principles

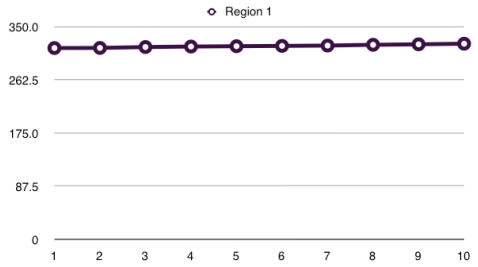
- Clear Vision
  1. Make data stand out
  2. Visual prominence
  3. Scale lines and data rectangle
  4. Superposed data sets

- Clear Understanding
  1. Explanations and conclusions
  2. Use all available space
  3. Juxtaposed data sets
  4. Log scaling
  5. Banking to 45°

Summary of Principles



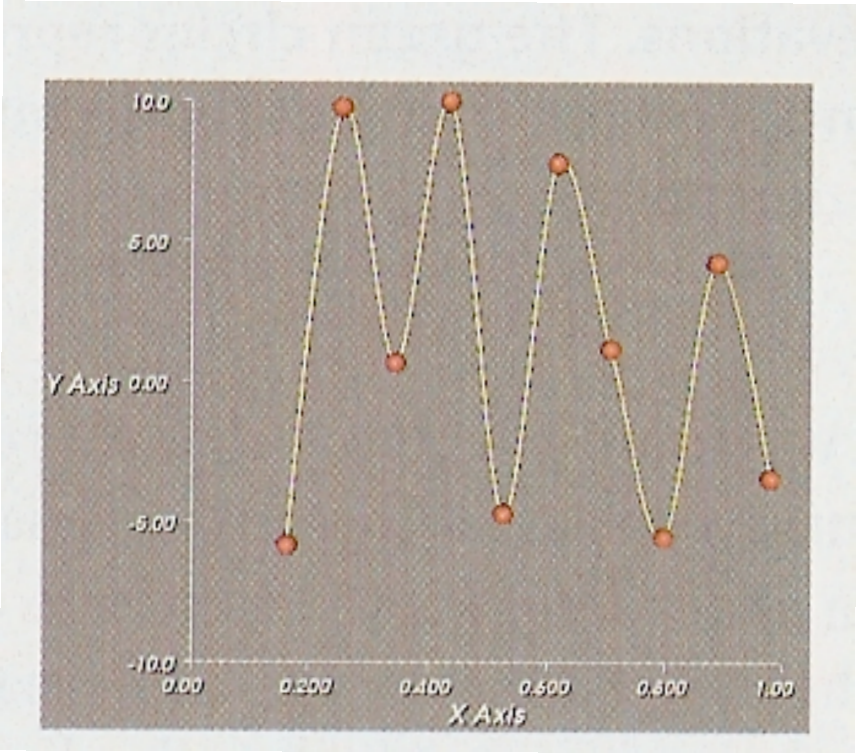




- Why are they all different?
- What is good/bad about each?

Quiz on Principles

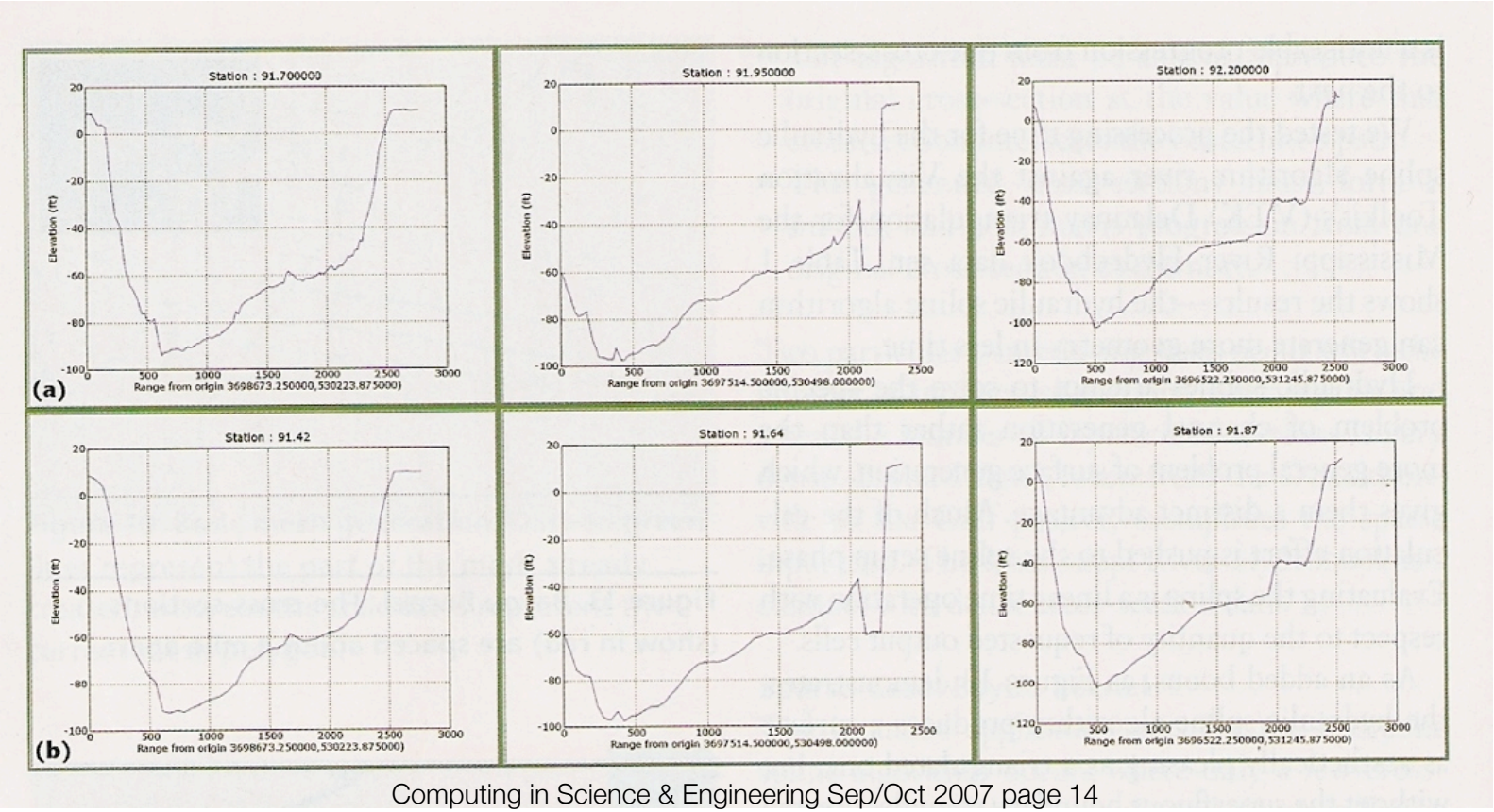
- What is wrong with this plot?



Computing in Science & Engineering
Sep/Oct 2007
page 8

Quiz on Principles

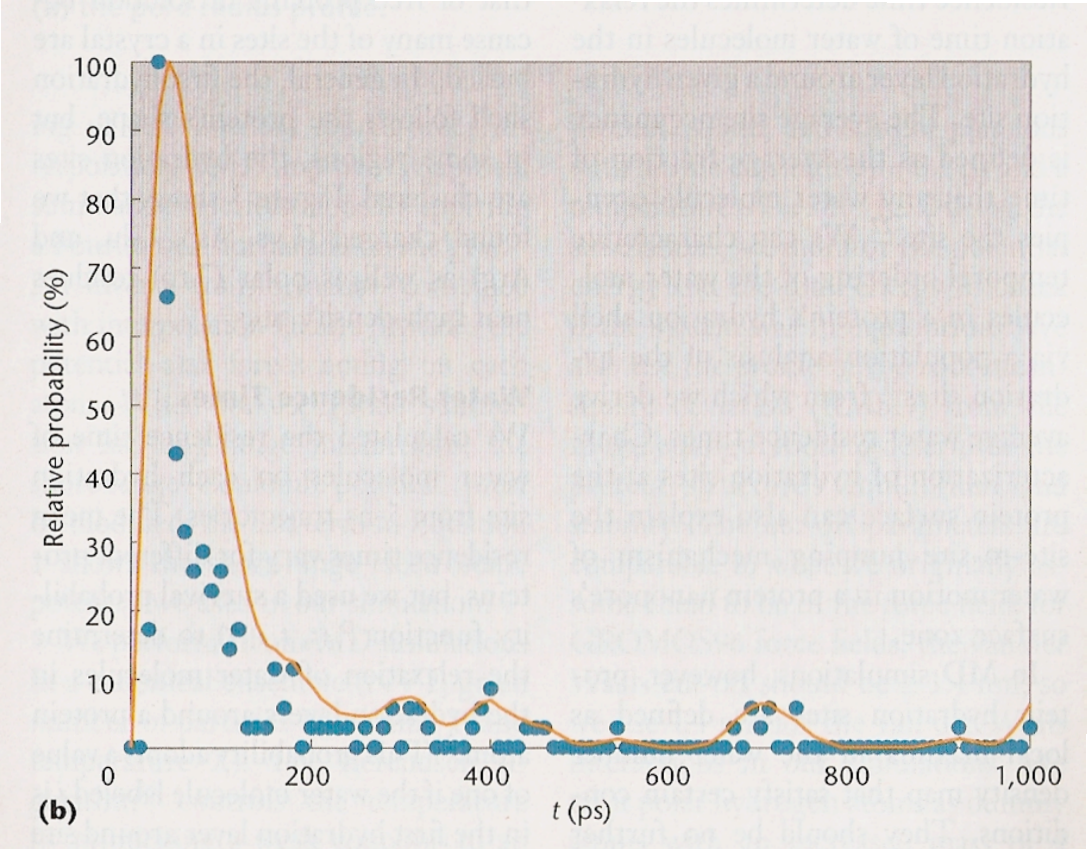
- What is wrong with this plot?



Computing in Science & Engineering Sep/Oct 2007 page 14

Quiz on Principles

- What is wrong with this plot?



Computing in Science & Engineering
Sep/Oct 2007
page 94

Elementary Plotting Techniques II

Motivation

- Given a certain type of data, what plotting technique should I use?
- What plotting techniques should be avoided?
- How do I encode additional information in my plot?

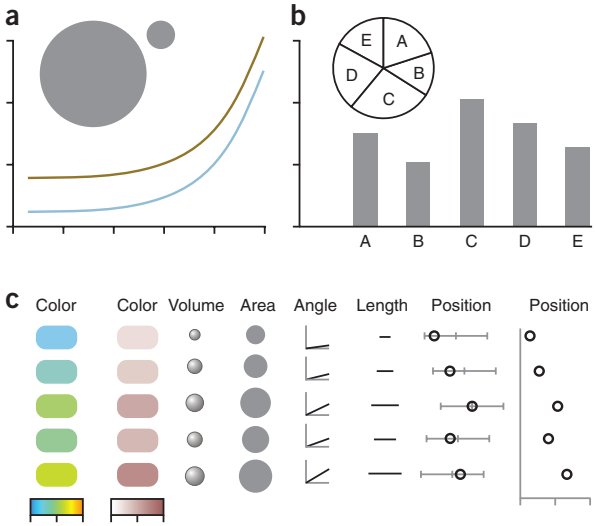


Figure 1 | Some visual estimations are more easily carried out than others. (**a**) Examples illustrating the difficulty in interpreting graphs and charts accurately. (**b**) Same data presented in a bar chart and in a pie chart. (**c**) Different visual variables encoding the same five values.

Table 1 | Elementary perceptual tasks

Rank	Aspect to compare
1	Positions on a common scale
2	Positions on the same but nonaligned scales
3	Lengths
4	Angles, slopes
5	Area
6	Volume, color saturation
7	Color hue

Tasks are ordered from most to least accurate. Information adapted from ref. 2.

Graphical Perception and Graphical Methods for Analyzing Scientific Data

William S. Cleveland and Robert McGill

Graphs provide powerful tools both for analyzing scientific data and for communicating quantitative information. The computer graphics revolution, which began in the 1960's and has intensified during the past several years, stimulated the invention of graphical methods: types of graphs and types of quantitative information to be shown on graphs (*1–4*). One purpose of this article is to describe and illustrate several of these new methods.

Summary. Graphical perception is the visual decoding of the quantitative and qualitative information encoded on graphs. Recent investigations have uncovered basic principles of human graphical perception that have important implications for the display of data. The computer graphics revolution has stimulated the invention of many graphical methods for analyzing and presenting scientific data, such as box plots, two-tiered error bars, scatterplot smoothing, dot charts, and graphing on a log base 2 scale.

What has been missing, until recently, in this period of rapid graphical invention and deployment is the study of graphs and the human visual system. When a graph is constructed, quantitative and categorical information is encoded, chiefly through position, shape, size, symbols, and color. When a person looks at a graph, the information is visually decoded by the person's visual system. A graphical method is successful only if the decoding is effective. No matter how clever and how technologically impressive the encoding, it fails if the decoding process fails. Informed decisions about how to encode data can be achieved only through an understanding of this visual decoding process, which we call graphical perception (*5*).

Our second purpose is to convey some recent theoretical and experimental investigations of graphical perception. We identify certain elementary graphical-perception tasks that are performed in the visual decoding of quantitative information from graphs; theory and experimental data are then used to order the tasks on the basis of accuracy. The ordering has an important application: data should be encoded so that the visual decoding involves tasks as high in the ordering as possible, that is, tasks performed with greater accuracy. This is illustrated by several examples in which some much-used graphical forms are presented, set aside, and replaced by new methods.

The authors are statistical scientists at AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974.

Elementary Tasks for the Graphical Perception of Quantitative Information

The first step is to identify elementary graphical-perception tasks that are used to visually extract quantitative information from a graph. (By "quantitative information" we mean numerical values of a variable, such as frequency of radiation and gross national product, that are not highly discrete; this excludes categorical information, such as type of metal and nationality, which is also shown on many graphs.) Ten tasks with which we have worked, in our theoretical investigations and in our experiments, are the following: angle, area, color hue, color saturation, density (amount of black), length (distance), position along a common scale, positions on identical but nonaligned scales, slope, and volume (Fig. 1).

Visual decoding as we define it for elementary graphical-perception tasks is what Julesz calls preattentive vision (*6*): the instantaneous perception of the visual field that comes without apparent mental effort. We also perform cognitive tasks such as reading scale information, but much of the power of graphs—and what distinguishes them from tables—comes from the ability of our preattentive visual system to detect geometric patterns and assess magnitudes. We have examined preattentive processes rather than cognition.

We have studied the elementary graphical-perception tasks theoretically, borrowing ideas from the more general field of visual perception (*7, 8*), and experimentally by having subjects judge graphical elements (*1, 5*). The next two sections illustrate the methodology with a few examples.

Study of Graphical Perception: Theory

Figure 2 provides an illustration of theoretical reasoning that borrows some ideas from the field of computational vision (*8*). Suppose that the goal is to judge the ratio, r, of the slope of line segment BC to the slope of line segment AB in each of the three panels. Our visual system tells us that r is greater than 1 in each panel, which is correct. Our visual system also tells us that r is closer to 1 in the two rectangular panels than in the square panel; that is, the slope of BC appears closer to the slope of AB in the two rectangular panels than in the square panel. This, however, is incorrect; r is the same in all three panels.

The reason for the distortion in judging Fig. 2 is that our visual system is geared to judging angle rather than slope. In their work on computational theories of vision in artificial intelligence, Marr (*8*) and Stevens (*9*) have investigated how people judge the slant and tilt (*10*) of the surfaces of three-dimensional objects. They argue that we judge slant and tilt as angles and not, for example, as their tangents, which are the slopes. An angle contamination of slope judgments explains the distortion in judgments of Fig. 2. Let the angle of a line segment be the angle between it and a horizontal ray extending to the right (θ in Fig. 3). The angles of the line segments in the square panel of Fig. 2 are not as similar in magnitude as the angles in either of the rectangular panels; this makes the slopes in the rectangular panels seem closer in value.

Again, let θ be the angle of a line segment. Suppose a second line segment has an angle $\theta + \Delta\theta$ where $\Delta\theta$ is small but just large enough that a difference in the orientations of the line segments can

https://www.dropbox.com/s/ey4tfxwpcfs9rxh/nmeth0910-665.pdf?dl=0

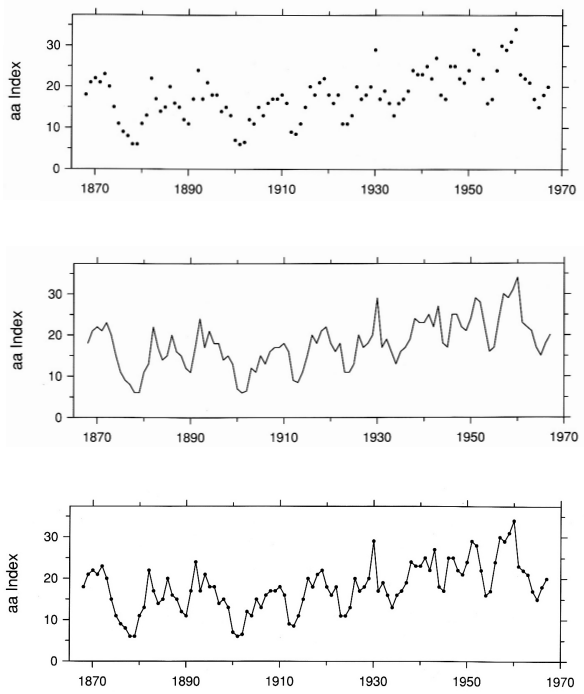
https://www.dropbox.com/s/5tog0adts3u6dcj/843c2c0e60915709268ac4224894469d82d5.pdf?dl=0

Summary

- Basic Plotting
  - Connected Symbol Plots
  - Dot Plots
  - Scatter Plots
  - Histograms
  - Others
- Advanced Plotting
  - Multimodal Data
  - Higher Dimensional Data
  - Correlation
  - Uncertainty and Variation

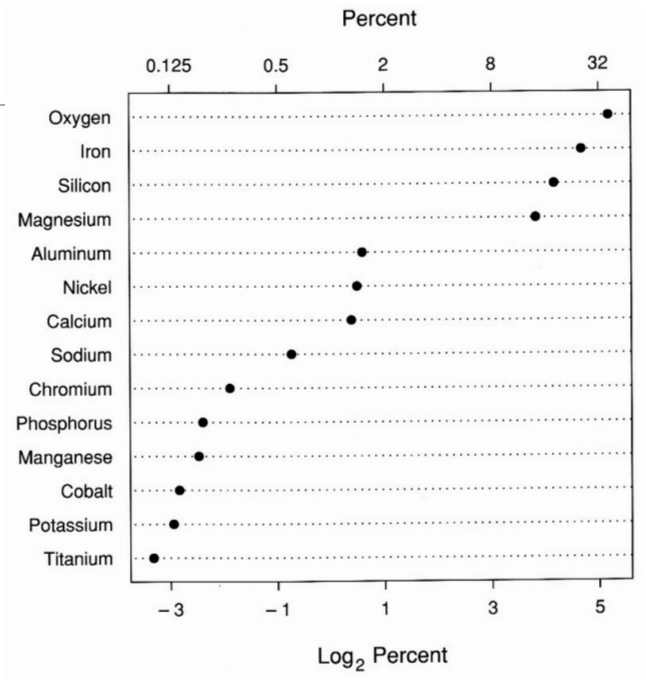
Basic Plotting

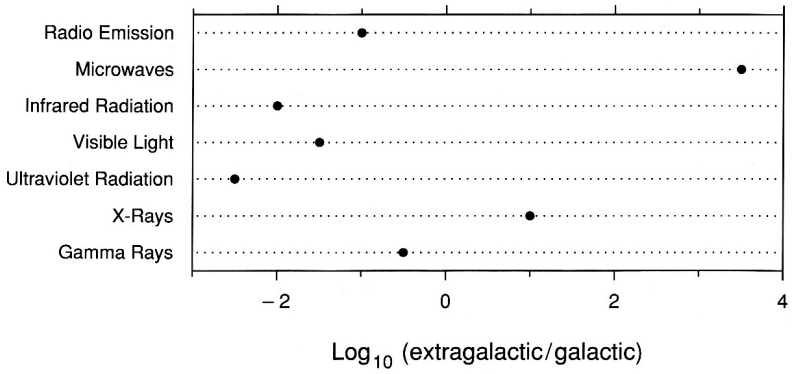
- Connected Symbol Plots
  - Used for graphing a time series or other 1D data
  - Symbols, connections, or connected symbols can be used
    - Symbols: High frequency data (spikey) where only the low frequency trend is important
    - Connections: Low frequency data (smooth) where points do not add additional information
    - Connected Symbols: In between data where the points can show concentrations of data and the connections can show the trend of the data



Basic Plotting

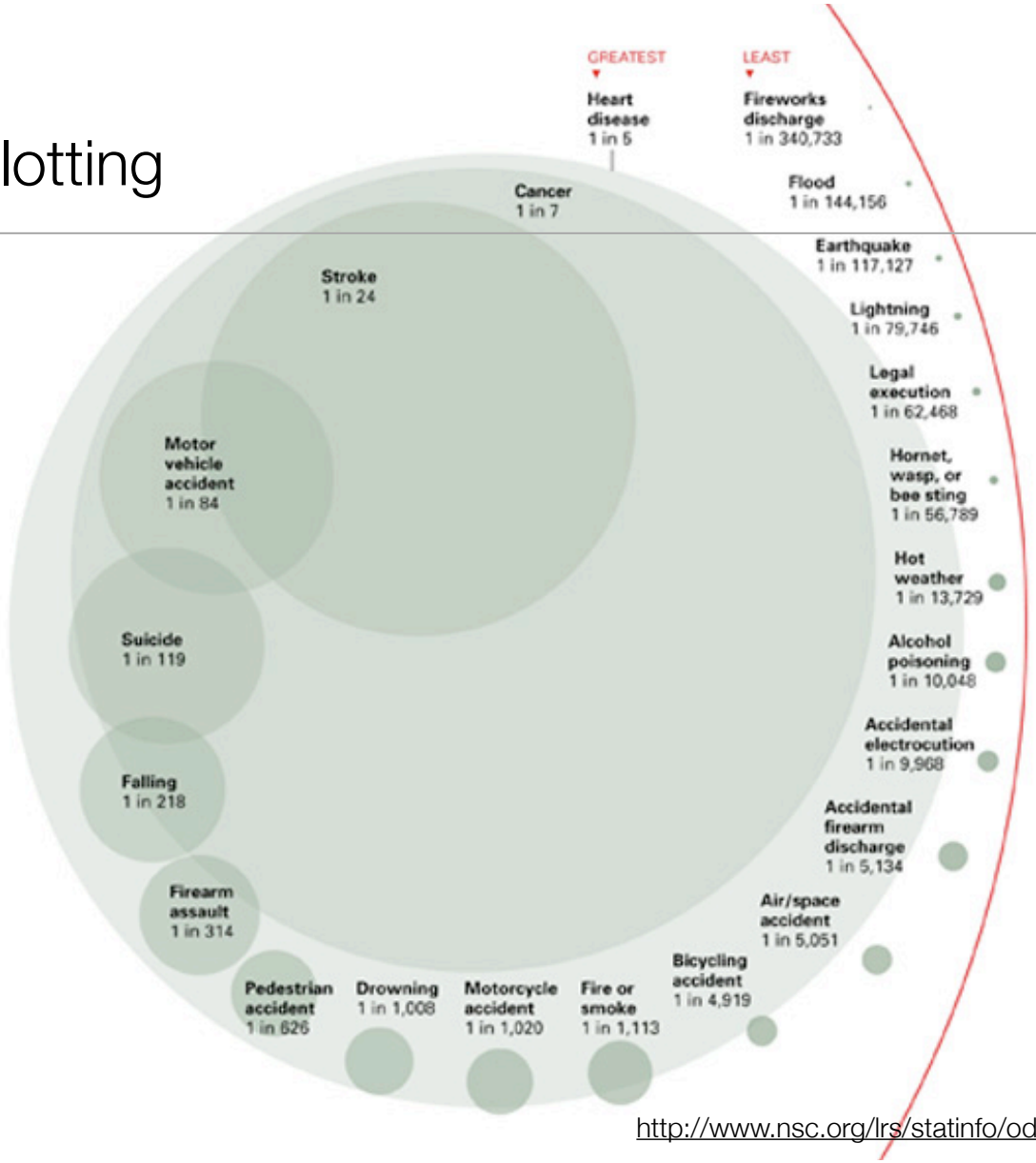
- Dot Plots
  - Used for quantitative, labeled data
  - Similar to the more familiar bar charts and pie charts
  - Order the plot in one of two ways:
    - Data: Sort from highest to lowest going from top to bottom
    - Label: Sort by label if it has an inherent order





Basic Plotting

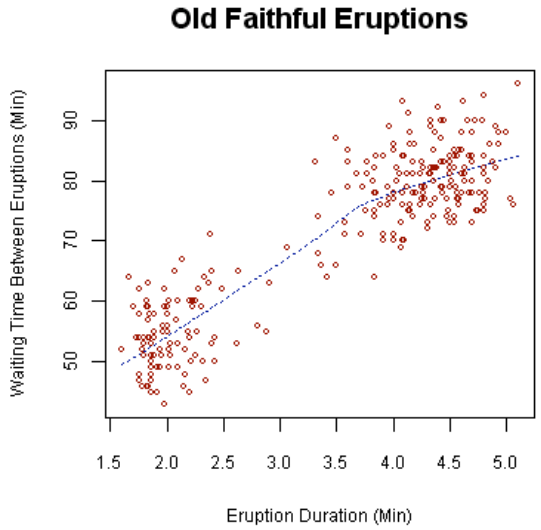
- Dot Plots

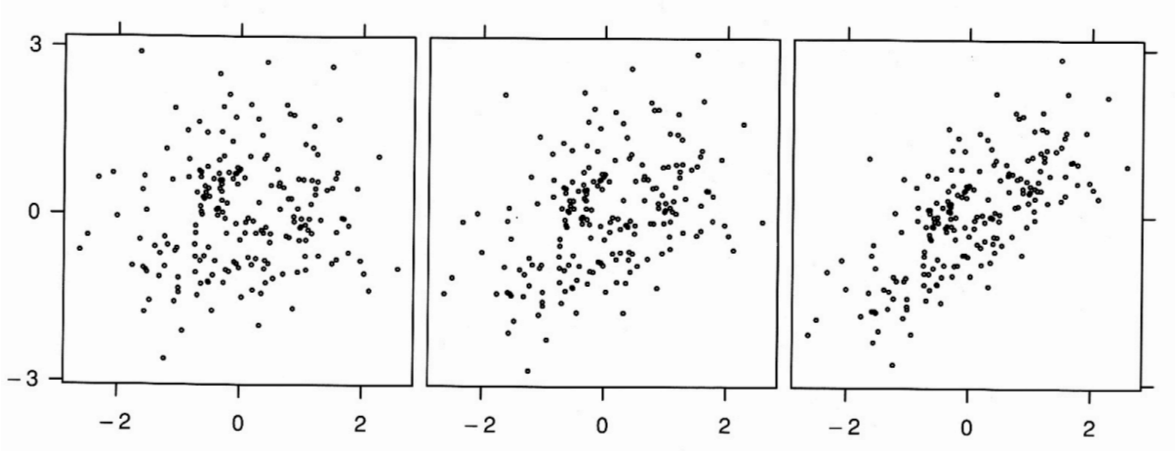


http://www.nsc.org/lrs/statinfo/odds.htm

Basic Plotting

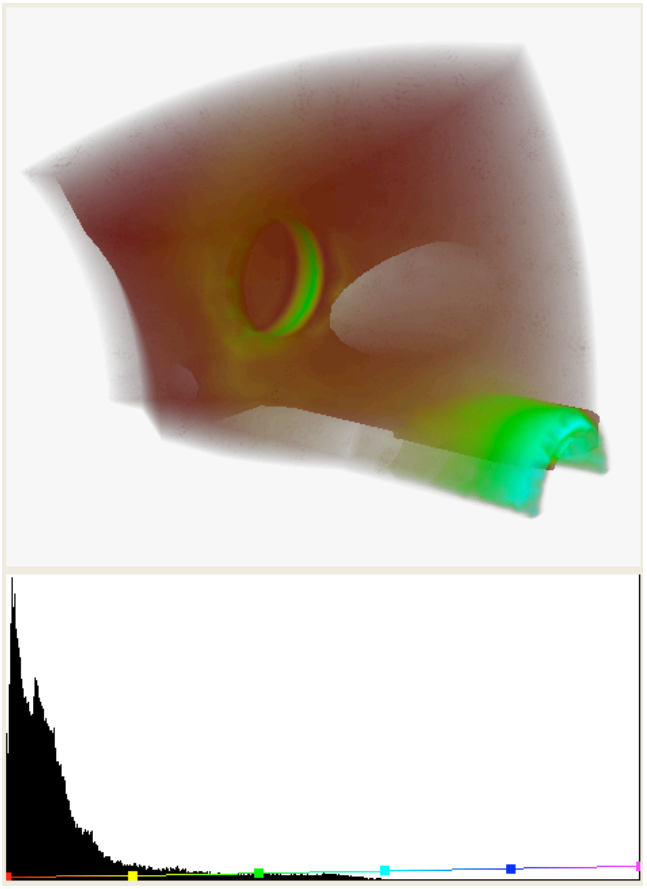
- Scatter Plots
  - Used to show how one variable is affected by another (correlation) in 2D data
    - Works well for lots of data samples
    - High vs Low vs. No Correlation





Basic Plotting

- Histograms
  - Used for analyzing distributions in large amounts of quantitative data
    - Horizontal axis is a set of bins (intervals) of the data
    - Vertical axis is the number of entries in the bin
    - Reduces the amount of data, interval selection is important



Histogram Bin Size

Another important choice to make for most cost functions is the histogram bin size, or equivalently, number of bins. All the cost functions investigated in section 3.2 require some form of intensity binning. For the Correlation Ratio and Woods function the intensity bins are only used for the reference volume in order to determine the iso-sets, whereas for the entropy-based functions, intensity binning occurs for both images.

The number of intensity bins determines two things: (1) how good the statistics will be in reflecting the ideal, continuous distribution and (2) the effective fidelity of the intensities (that is, using 256 bins is equivalent to that image having an 8 bit intensity range). In most cases it is the first point that is of most interest, as the Signal to Noise Ratio in a typical image limits the maximum practical fidelity.

Determining how many histogram bins should be used for estimating distributions is a problem in non-parametric statistics, although histogram-based methods are not the only form of distribution estimation (for an overview see [Izenman, 1991]). However, histogram-based methods are the most practical as other methods usually involve too much computational overhead to be useful for this problem.

It has been shown [Scott, 1979] that the optimal histogram bin size, which provides the most efficient, unbiased estimation of the probability density function, is achieved when:

$$W = 3.49\sigma N^{-1/3} \tag{38}$$

where W is the width of the histogram bin, σ is the standard deviation of the distribution and N is the number of available samples. In practice, the estimated standard deviation, s, must be used. A similar, but more robust, result was also obtained by Freedman and Diaconis (summarised in [Izenman, 1991]), which gives the bin width as:

$$W = 2\,(IQR)\,N^{-1/3} \tag{39}$$

where IQR is the interquartile range (the 75th percentile minus the 25th percentile).

Biometrika (1979), **66**, 3, *pp.* 605–10
Printed in Great Britain

On optimal and data-based histograms

BY DAVID W. SCOTT
Department of Mathematical Sciences, Rice University, Houston, Texas

SUMMARY

In this paper the formula for the optimal histogram bin width is derived which asymptotically minimizes the integrated mean squared error. Monte Carlo methods are used to verify the usefulness of this formula for small samples. A data-based procedure for choosing the bin width parameter is proposed, which assumes a Gaussian reference standard and requires only the sample size and an estimate of the standard deviation. The sensitivity of the procedure is investigated using several probability models which violate the Gaussian assumption.

Some key words: Frequency distribution; Histogram; Nonparametric density estimation; Optimal bin width.

1. INTRODUCTION

The histogram is the classical nonparametric density estimator, probably dating from the mortality studies of John Graunt in 1662 (Westergaard, 1968, p. 22). Today the histogram remains an important statistical tool for displaying and summarizing data. In addition it provides a consistent estimate of the true underlying probability density function. Present guidelines for constructing histograms do not directly address the issues of estimation bias and variance. Rather, they draw heavily on the investigator's intuition and past experience. In this paper we propose new guidelines that reduce the subjectivity involved in histogram construction by considering a mean squared error criterion.

2. BACKGROUND

We consider only histograms defined on an equally spaced mesh $\{t_{ni}; -\infty < i < \infty\}$ with bin width $h_n = t_{n(i+1)} - t_{ni}$, where n denotes the sample size and emphasizes the dependence of the mesh and bin width on the sample size. For a fixed point x, the mean squared error of a histogram estimate, $\hat{f}(x)$, of the true density value, $f(x)$, is defined by

$$\text{MSE}\,(x) = E\{\hat{f}(x) - f(x)\}^2.$$

For a random sample of size n from f, Čencov (1962) proved that MSE (x) asymptotically converges to zero at a rate proportional to $n^{-2/3}$, that is, MSE $(x) = O(n^{-2/3})$. This rate is fairly close to the Cramér–Rao lower bound of $O(n^{-1})$. The integrated mean squared error represents a global error measure of a histogram estimate and is defined by

$$\text{IMSE} = \int E\{\hat{f}(x) - f(x)\}^2\,dx.$$

Since it is the shape of the density that is of most interest, the IMSE is more relevant than the mean squared error of the density height. The IMSE of a histogram also converges to zero as $O(n^{-2/3})$.

To achieve these rates of convergence requires proper choice of the two parameters of the histogram, the bin width h_n and the relative position of the mesh. The latter is determined by

5. DATA-BASED HISTOGRAMS

The optimal choice for h_n requires knowledge of the true underlying density f. This knowledge is rare. In another context Tukey (1977, p. 623) has suggested using the Gaussian density as a reference standard, to be used cautiously but frequently. Therefore, we propose the data-based choice for the bin width

$$h_n = 3{\cdot}49sn^{-1/3}, \tag{6}$$

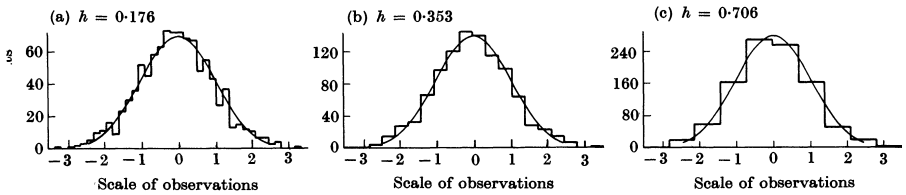


Fig. 2. Histograms of 1000 pseudorandom Gaussian numbers for three bin widths: the data-based choice and that choice perturbed by a factor of 2.

https://arxiv.org/abs/0807.4820

Data analysis recipes: Choosing the binning for a histogram[1]

David W. Hogg
Center for Cosmology and Particle Physics, Department of Physics
New York University
david.hogg@nyu.edu

Data points are placed in bins when a histogram is created, but there is always a decision to be made about the number or width of the bins. This decision is often made arbitrarily or subjectively, but it need not be. A jackknife or leave-one-out cross-validation likelihood is defined and employed as a scalar objective function for optimization of the locations and widths of the bins. The objective is justified as being related to the histogram's usefulness for predicting future data. The method works for data or histograms of any dimensionality.

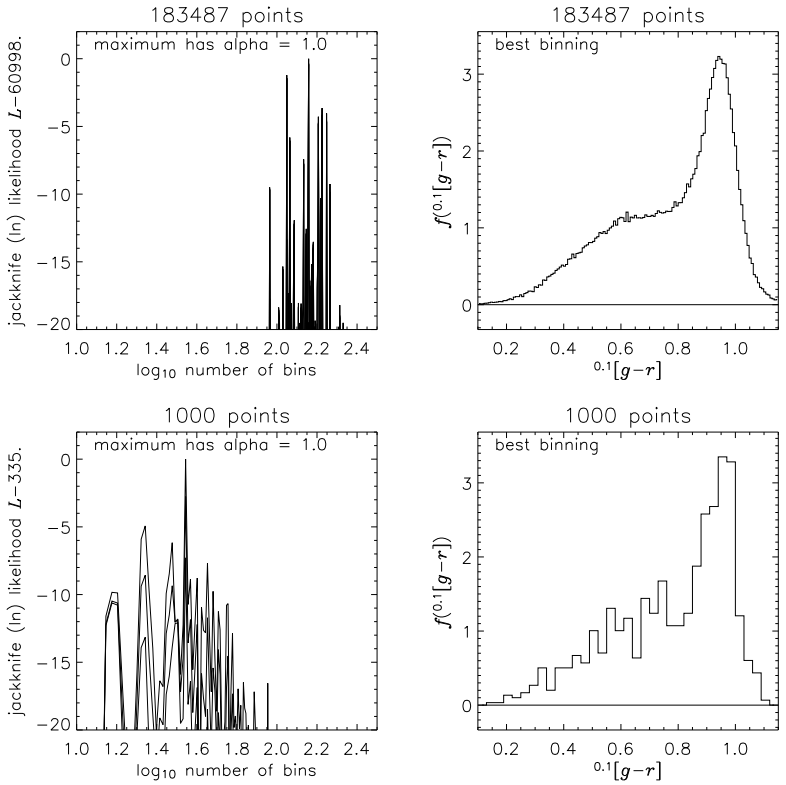


Figure 1: Constant-width binning of a set of measured galaxy colors. The top-left panel shows grid searches in binsize for the eight possible combinations of smoothing $\alpha = (10, 1, 0.1, 0.01)$ and binning phase $\delta = (0, 0.5)$ (see text for definitions). The top-right panel shows the data binned with the maximum-likelihood binning parameters. The bottom panels show the same, but for a randomly chosen subsample.

https://arxiv.org/abs/0807.4820

Data analysis recipes: Choosing the binning for a histogram[1]

David W. Hogg
Center for Cosmology and Particle Physics, Department of Physics
New York University
david.hogg@nyu.edu

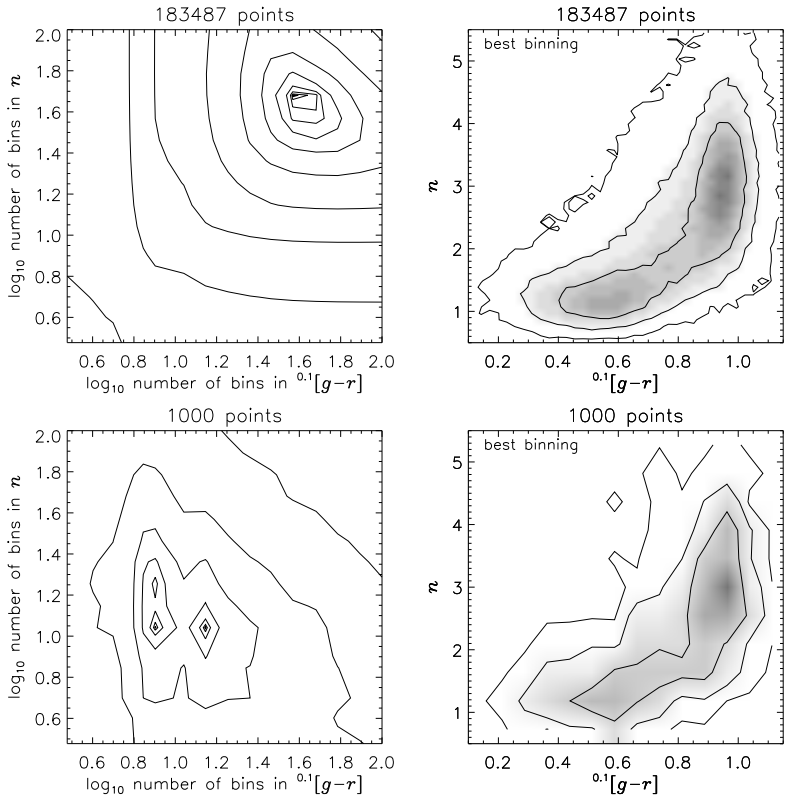
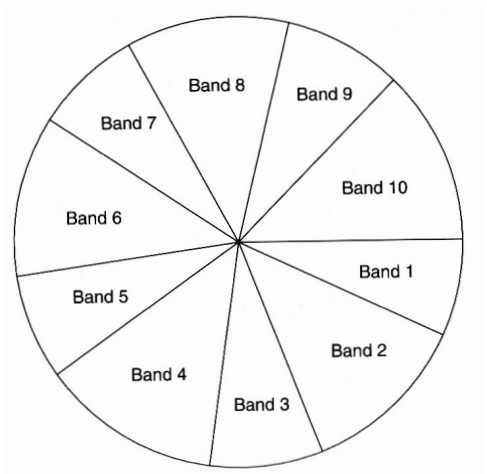
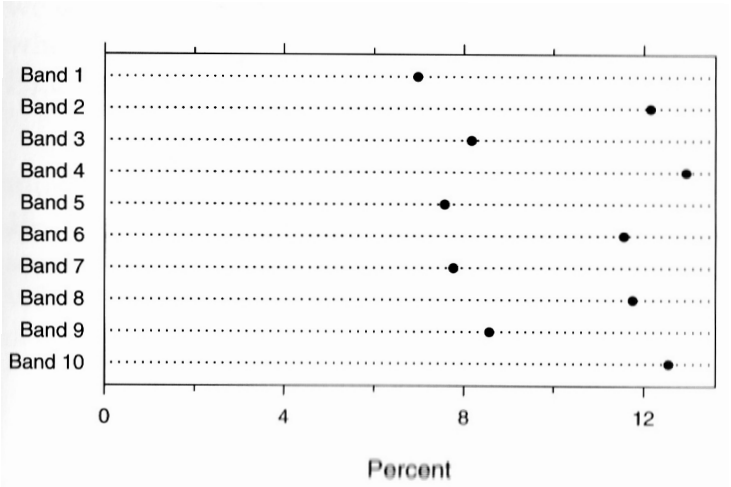


Figure 2: Two-dimensional constant-width binning of a set of measured galaxy colors and radial profile shapes (as parameterized by the Sérsic index n). The top-left panel shows a grid search in the two binsizes, with smoothing fixed at $\alpha = 1.0$ and both phases fixed at $\delta = 0$. The top-right panel shows the data binned with the maximum-likelihood binning parameters, plus contours at 2, 10, 25, 50, and 75 percent of the maximum value. The bottom panels show the same, but for a randomly chosen subsample.

Basic Plotting

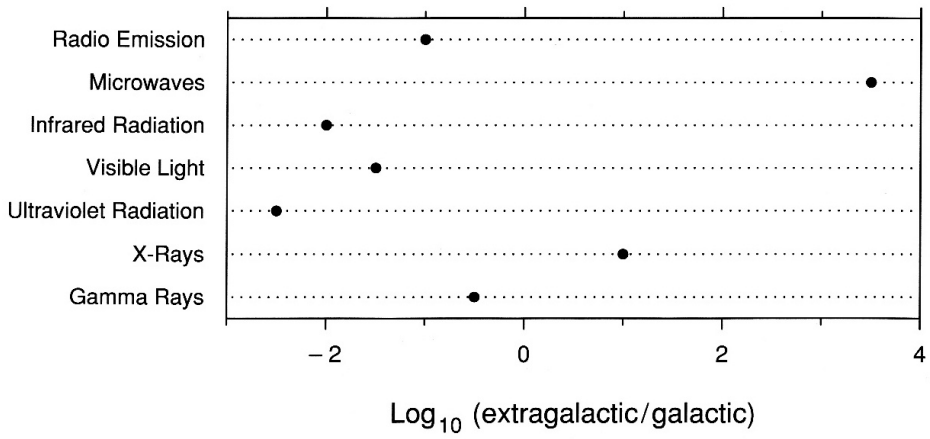
- Others
  - Pie Charts
    - Don’t use for scientific data, use a dot plot instead
    - Poor pattern perception: judging area is difficult!





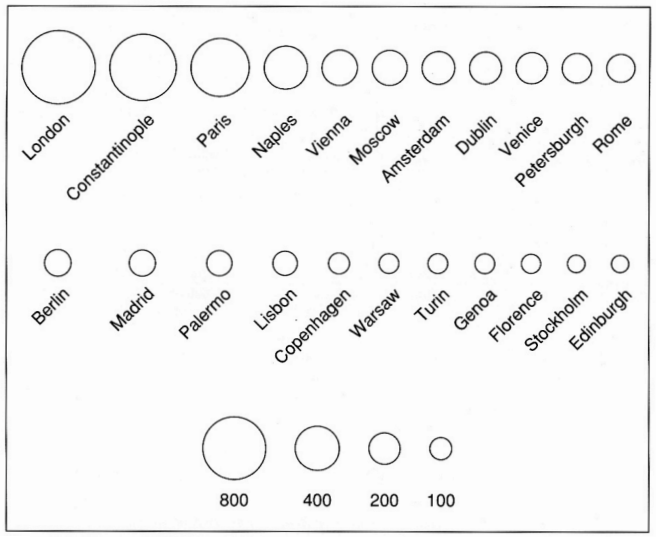
Basic Plotting

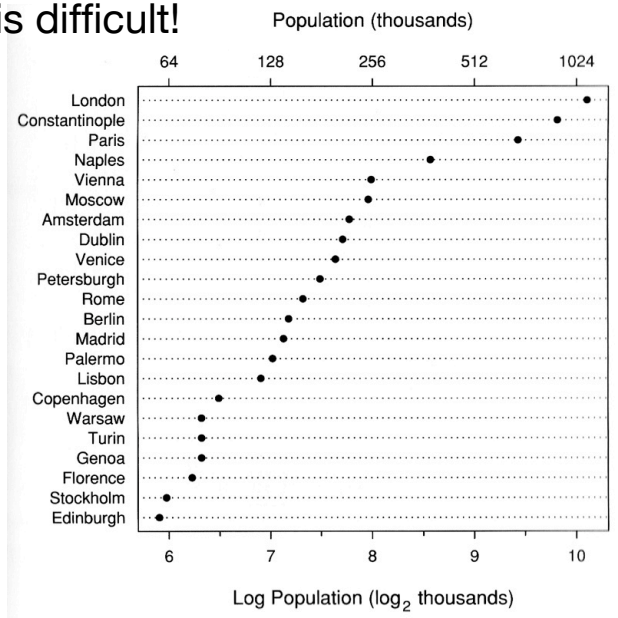
- Others
  - Bar Charts
    - Don’t use for scientific data, use a dot plot instead
    - How do you show data that does not have a zero baseline?



Basic Plotting

- Others
  - Area Charts
    - Don’t use for scientific data, use a dot plot instead
    - Poor pattern perception: judging area is difficult!

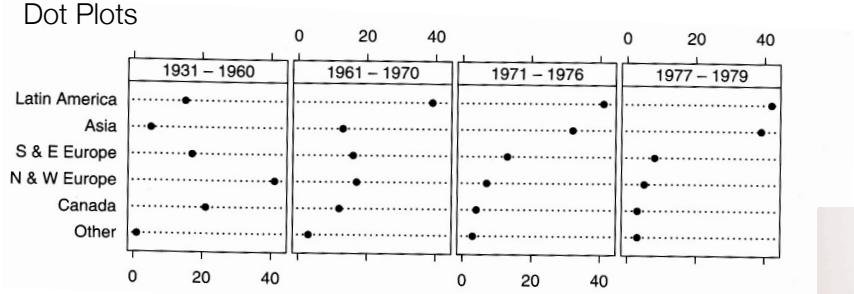


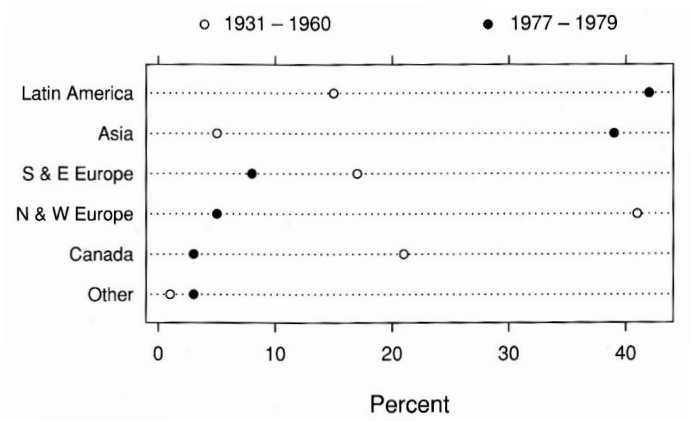


Advanced Plotting

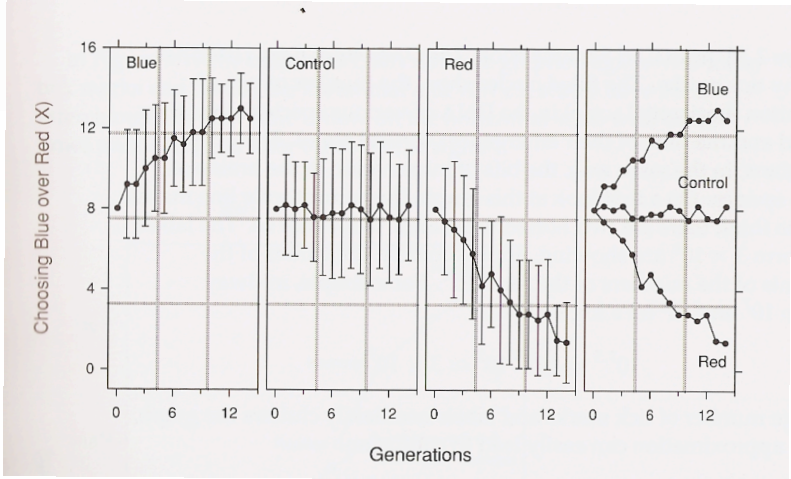
- Multimodal Data
  - Juxtaposed vs. Superposed

Dot Plots



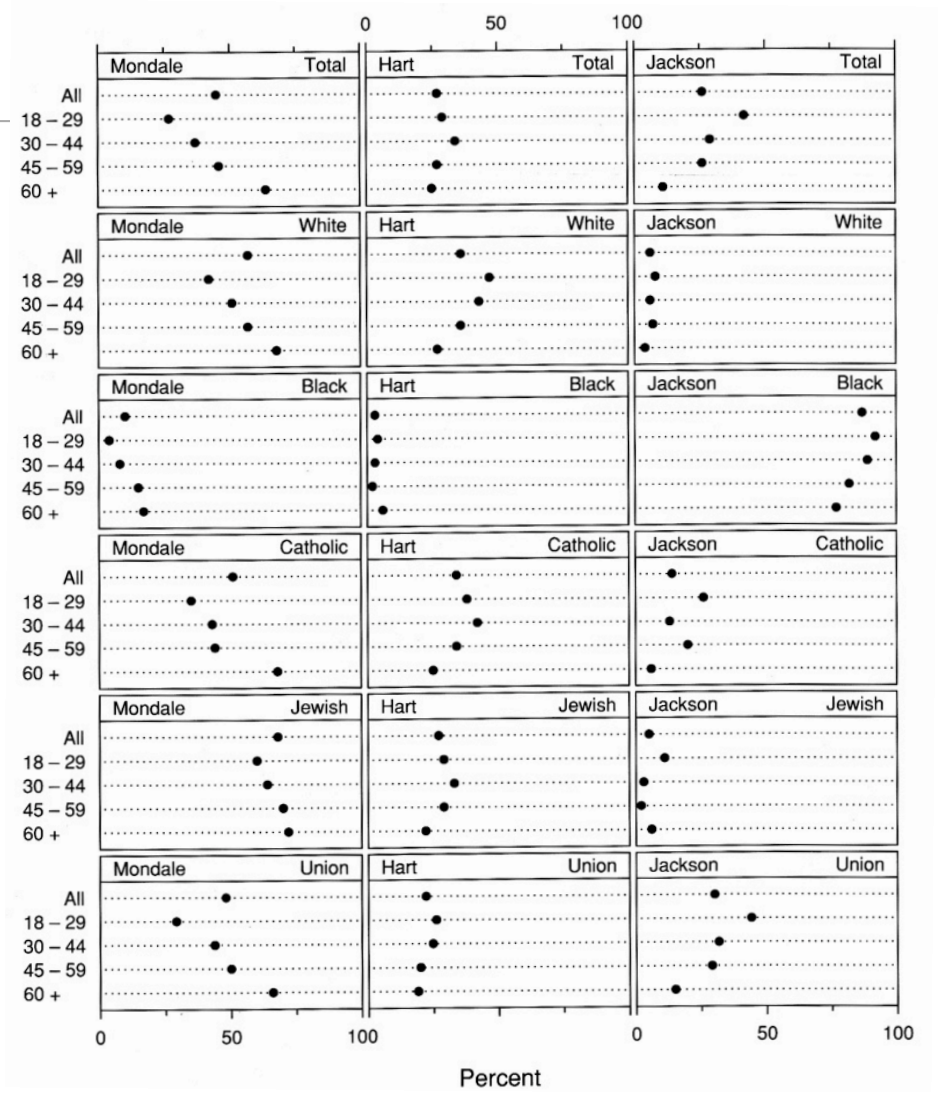


Symbol and Connection Plots



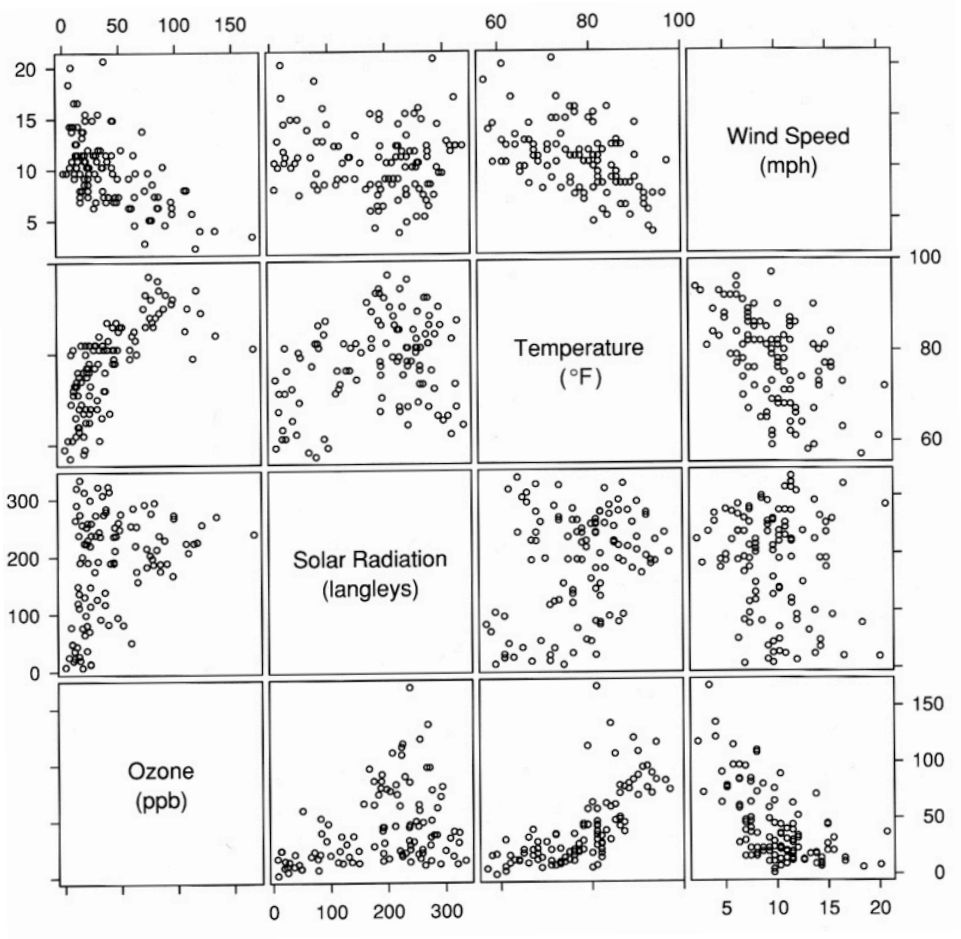
Advanced Plotting

- Higher Dimensional Data
  - Multiway Dot Plots



Advanced Plotting

- Higher Dimensional Data
  - Scatterplot Matrices



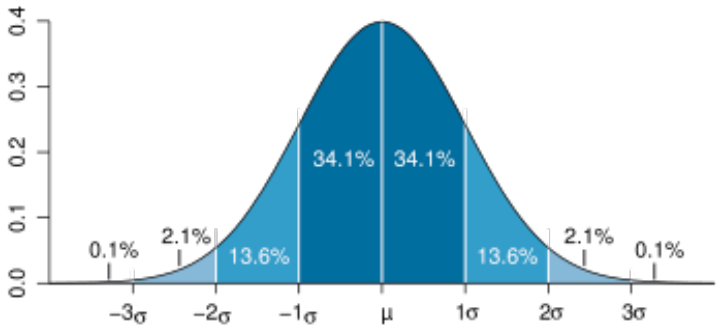
Advanced Plotting

- Correlation
  - Linear Regression using least squares
    - Find the regression line: $y = a_0 + a_1 x$
    - Where the summed squares of the vertical distances: $\Delta = \sum_0^n (y_i - f(x_i))^2$
    - And the best parameter set for the fit is achieved when the sum of the squares of the distance Δ is mimimal for the approximation: $\frac{\delta \Delta}{\delta a_i} = 0$

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

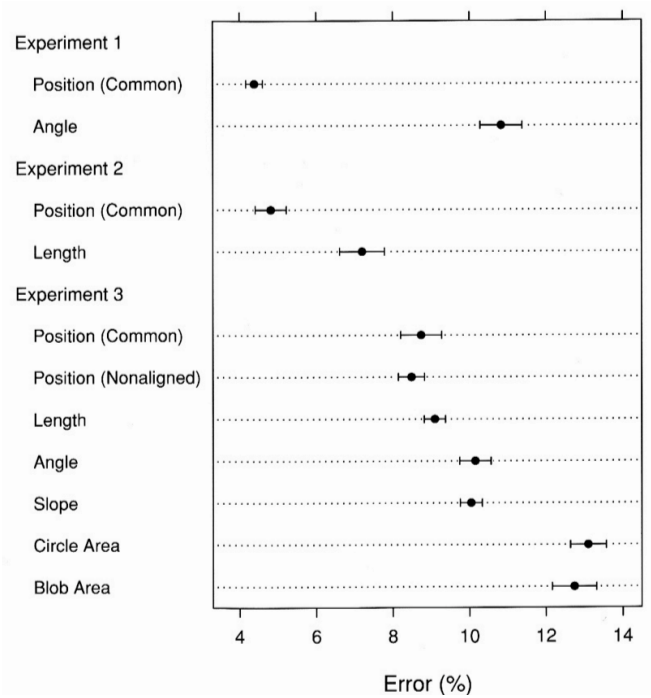
Advanced Plotting

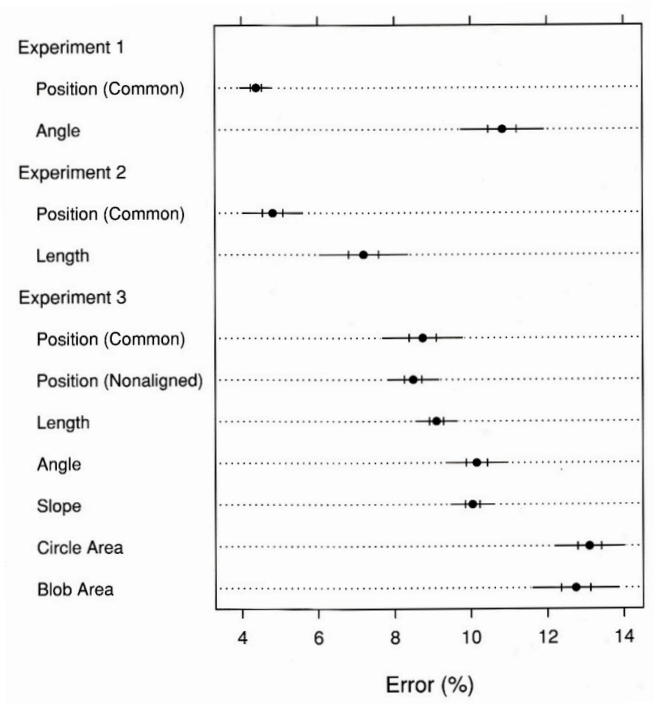
- Uncertainty and Variation
  - Two types of variation
    - Empirical Distribution: The variation captured in the experiment
    - Sample to Sample Variation: The variation that occurs in statistics from a sparse sampling and a denser sampling
  - Represent mean, standard deviation, and confidence intervals for a normal distribution



Advanced Plotting

- Uncertainty and Variation
  - Error Bars: Mean and one standard deviation or mean, 50%, and 95% confidence intervals





Advanced Plotting

- Uncertainty and Variation
  - Box Plots (Tukey Bars): First quartile, second quartile (mean), third quartile, adjacents (first-1.5r, third+1.5r), and outside.