Elementary Plots

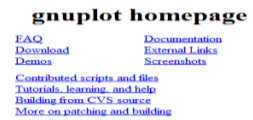
Why Should We Care?

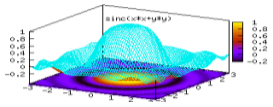
- Everyone uses plotting
- But…most people ignore or are unaware of simple principles
- Default plotting tools are not always the best

- More importantly, it is easy to lie or deceive people with bad plots

Some examples of available plotting tools !

http://plasma-gate.weizmann.ac.il/Grace/





http://www.gnuplot.info/

http://software.proindependent.com/pricing.html

http://office.microsoft.com/en-us/excel/default.aspx

http://www.mathworks.com/

http://www.aptplot.com/

http://www.sigmaplot.com/products/sigmaplot/sigmaplot-details.php

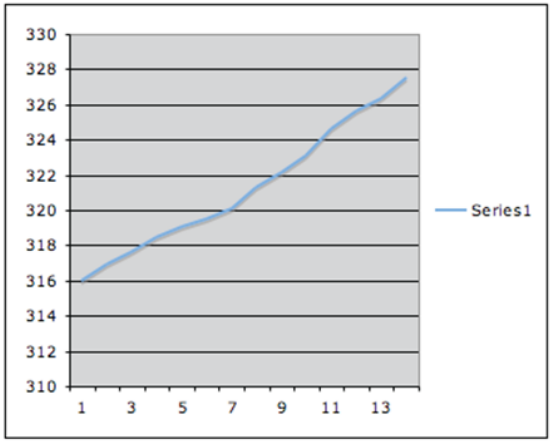
http://matplotlib.sourceforge.net/

http://www.wolfram.com/

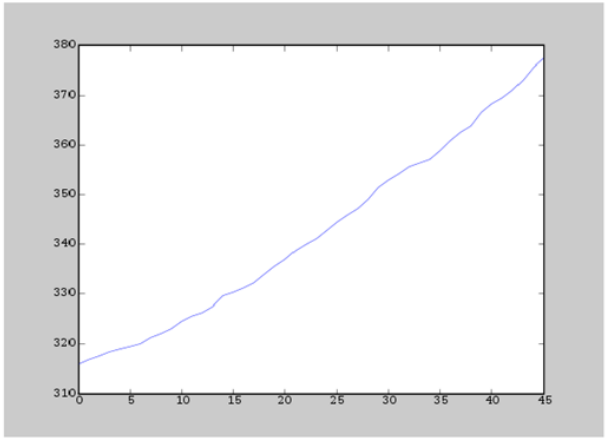
What Can Plots Do?

- **Data analysis and communication**
  - In a simplistic view, plotting reduces a large amount of information to a smaller form that is more easily understood via certain graphical representation.
  - Reduction of the data to its simplest and cleanest form, such that the relationships inherent in the data (points) are easily perceived.

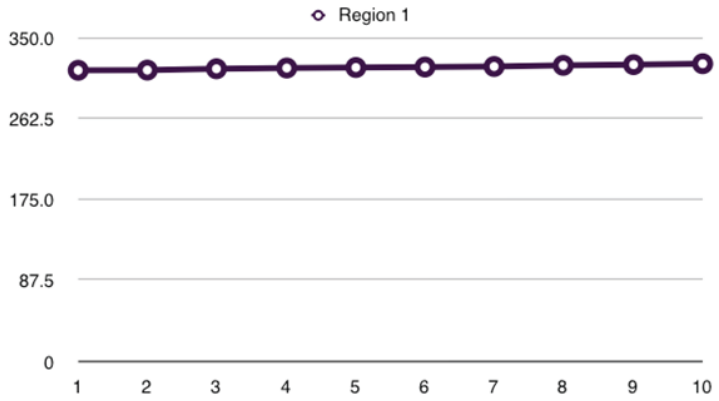
Examples of plots generated by a number of tools using their default setting



Default Excel Plot

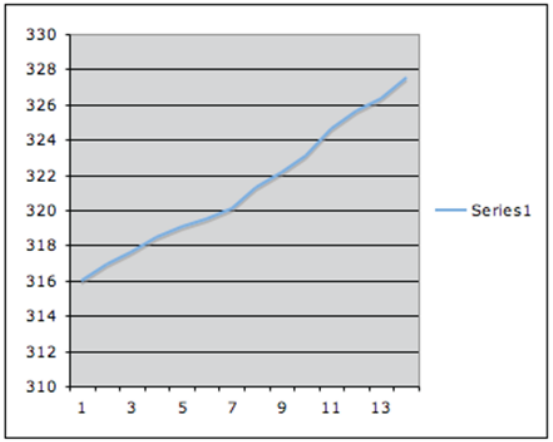


Default Matplotlib/Matlab Plot

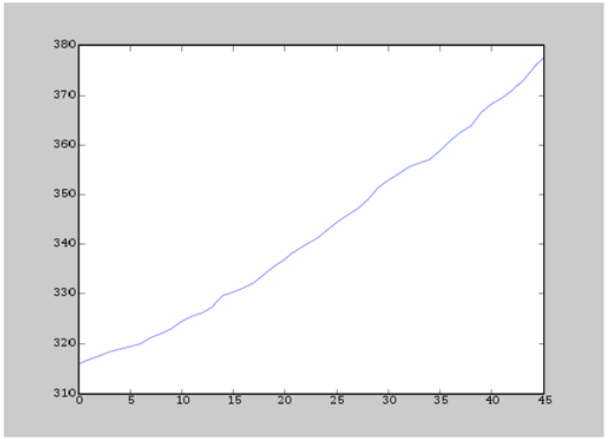


Default Pages Plot

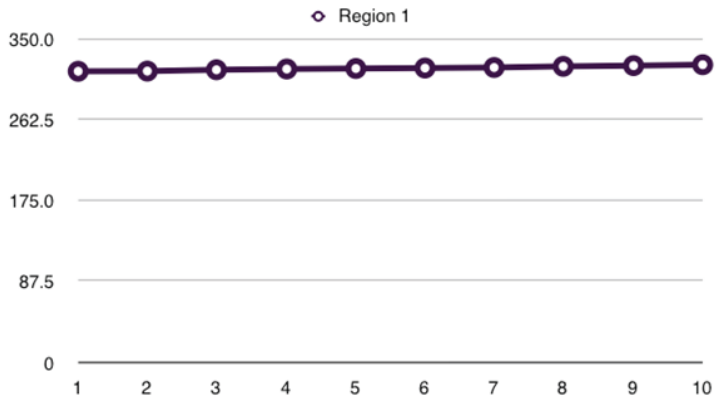
Examples of plots generated by a number of tools using their default setting



Default Excel Plot



Default Matplotlib/Matlab Plot



Default Pages Plot

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

These plots demonstrate two important points:

First, there is no obvious standard for what a plot should look like. This is easy to see by the differences in the axes and scale lines, the data rectangle inside the plot, and the actual representation of the data values.

Second, creating a plot is an iterative process that can not be generally applied to all types of data.

There are no magic formulas for creating a useful plot. However, some general principles have been advocated that can be applied to plots to improve their likelihood of being useful.

The information provided here should be considered as guidelines

PRINCIPLES OF PLOTTING

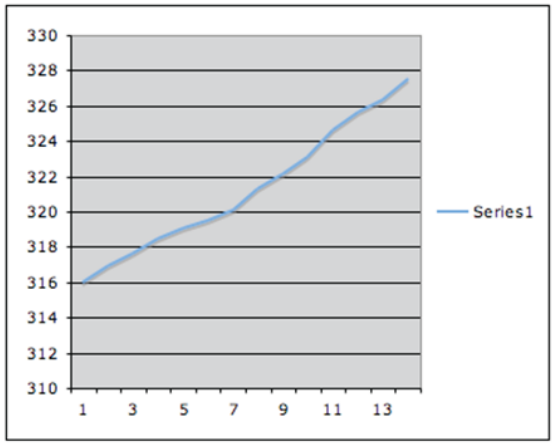
Visualizing Data [Cleveland 93] and *Elements of Graphing Data* [Cleveland 94] by William S. Cleveland

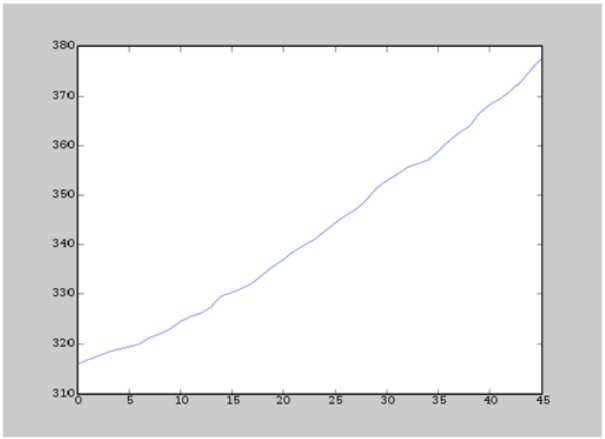
Principles of Plotting

- Improving the vision
  - Improve the readability of the plot
- Improving the understanding
  - Ensure that the analysis of the plot is effectively communicated.

Improving the Vision

- Principle 1: Reduced clutter, Make data stand out
  - The main focus of a plot should be on the data itself, any superfluous elements of the plot that might obscure or distract the observer from the data needs to be removed.

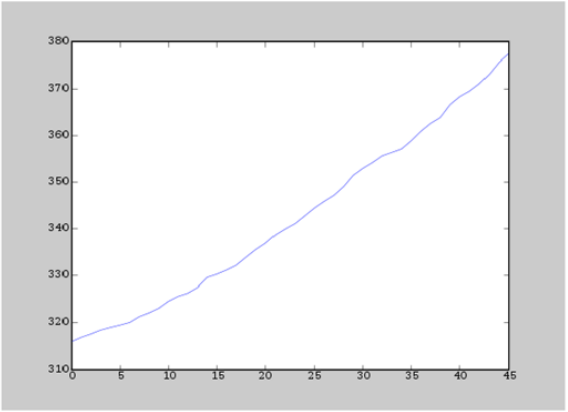


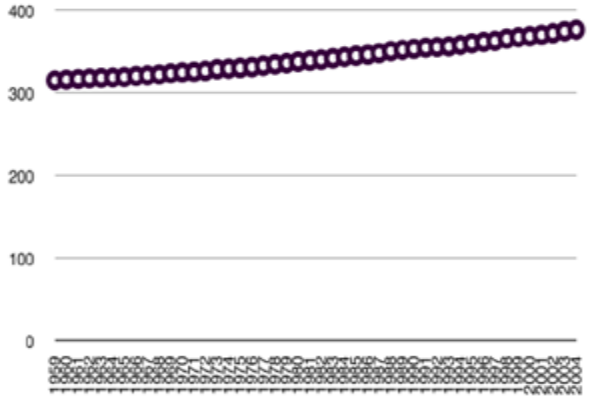


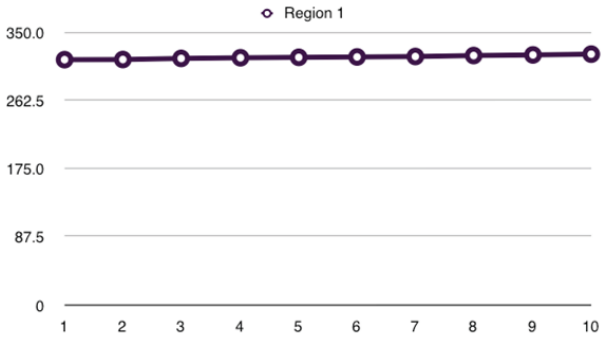
Which one is better?

Improving the Vision

- Principle 2: Use visually prominent graphical elements to show the data.
  - Connecting lines should never obscure points and points should not obscure each other.
  - If multiple data sets are represented in the same plot (superposed data), they must be visually separable.
  - If this is not possible due to the data itself, the data can be separated into adjacent plots that share an axis

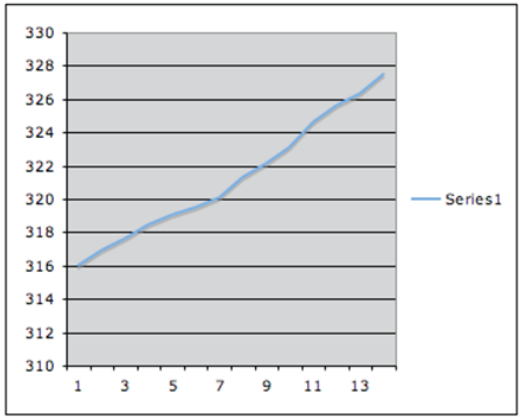


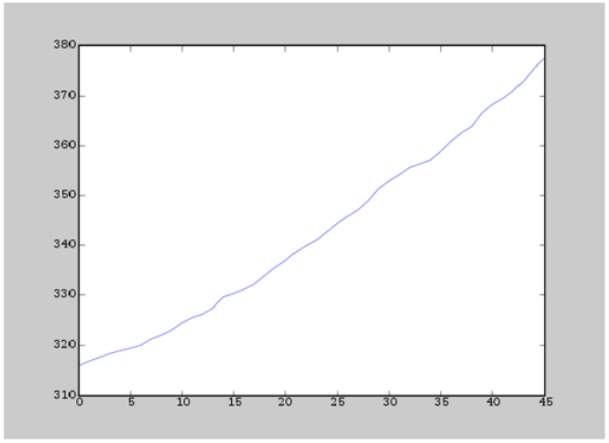


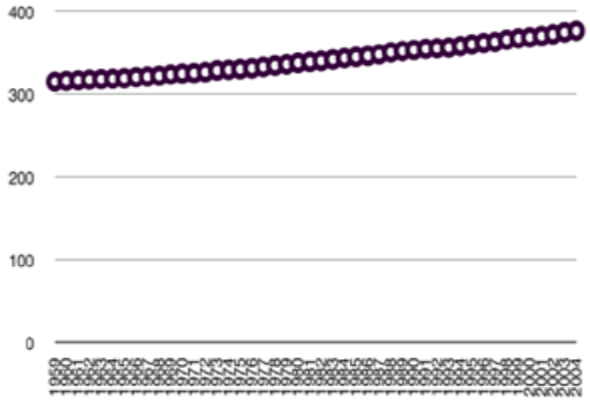


Improving the Vision

- Principle 3: Use proper scale lines and a data rectangle.
  - Two scale lines should be used on each axis (left and right, top and bottom) to frame to data rectangle completely.
  - Add margins for data to make the plot prominent.
  - Tick-marks outs and 3-10 for each axis.

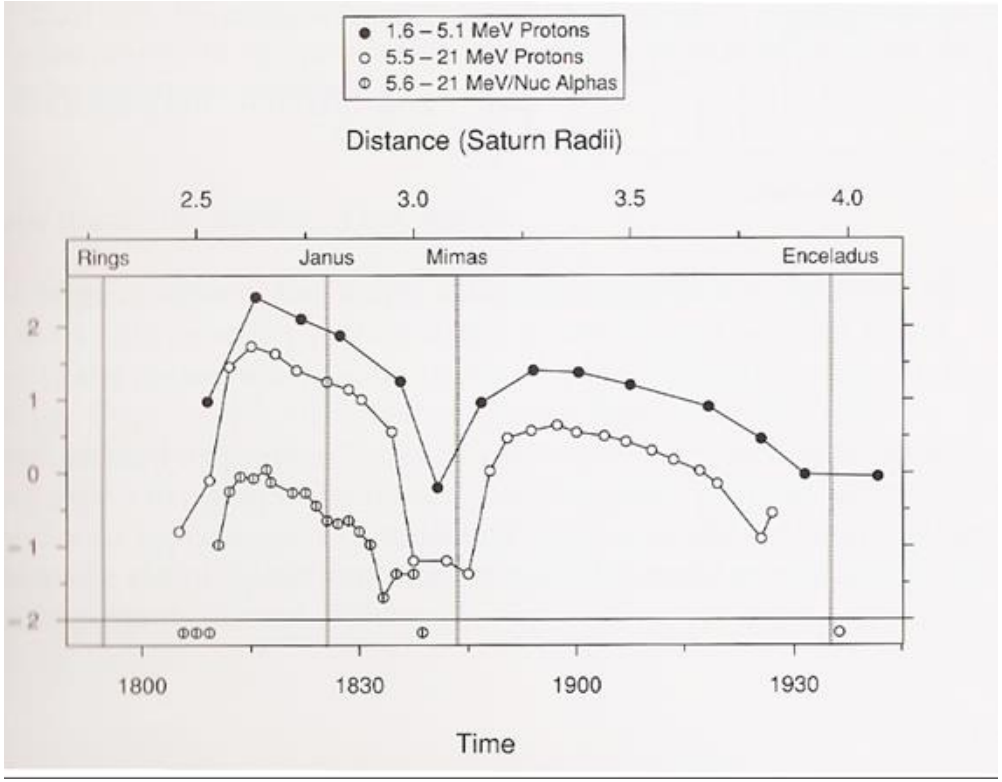






Improving the Vision

- Principle 4: Reference lines, labels, notes, and keys.
  - Reference lines are only used to show the thresholds within data.
  - Only use them sparsely when necessary and don't let them obscure data.



Improving the Vision

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

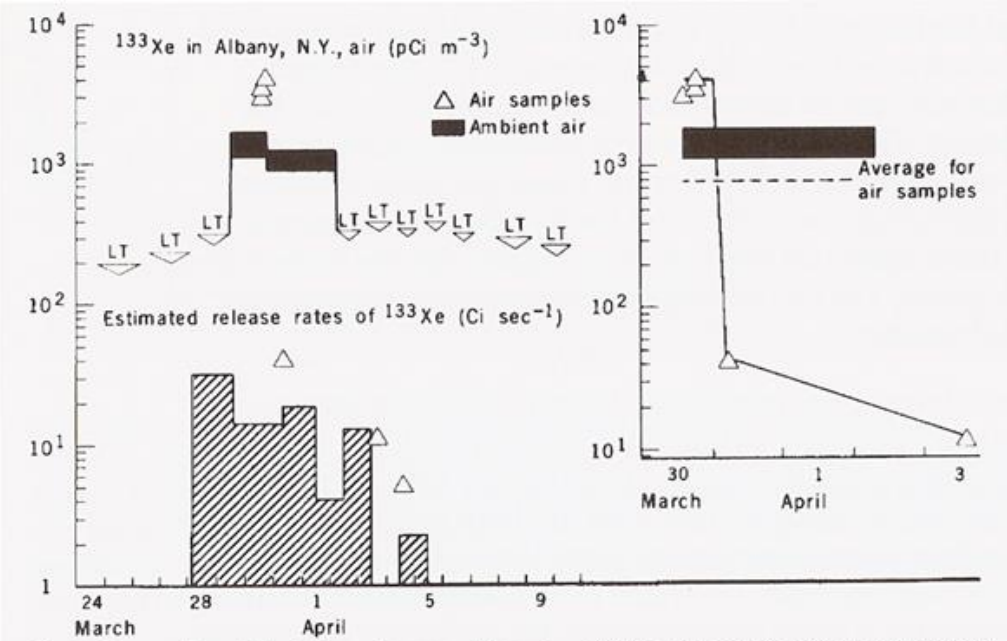
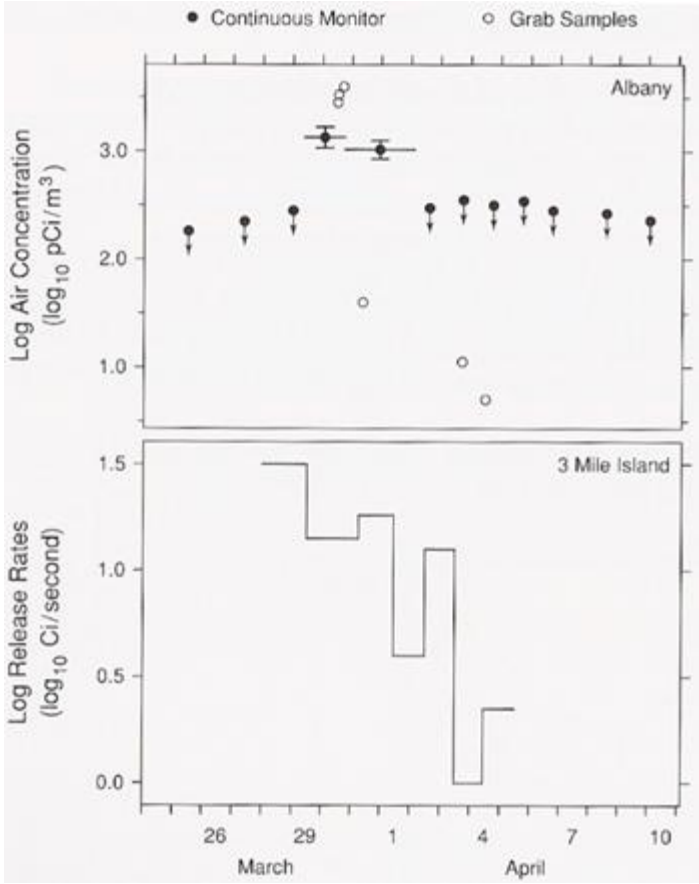
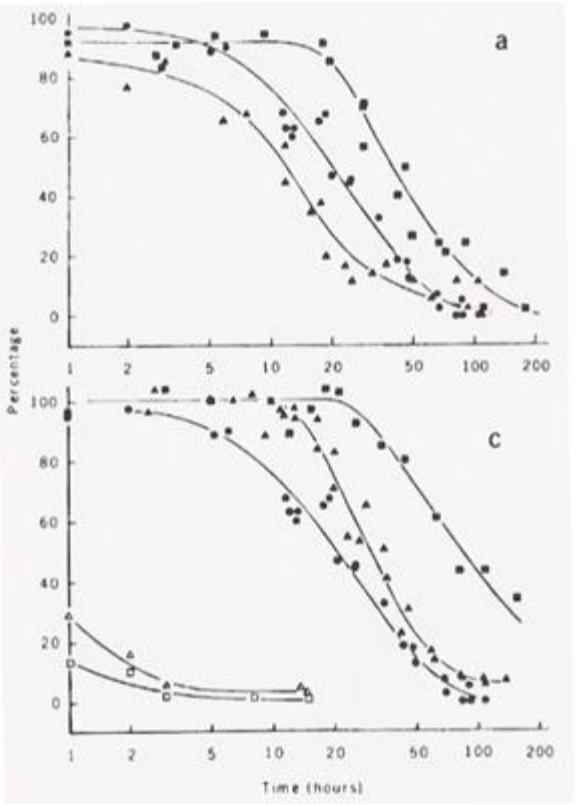


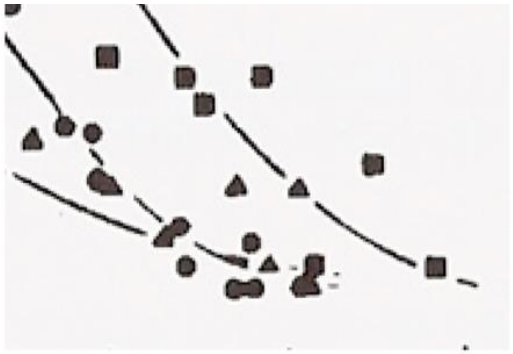
Fig. 1. Xenon-133 activity (picocuries per cubic meter of air) in Albany, New York, for the end of March and early April 1979. The lower trace shows the time-averaged estimates of releases (curies per second) from the Three Mile Island reactor (2). The inset shows detailed values for air samples (gas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



Improving the Vision

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





Improving the Understanding

- Principle 1: Provide explanations and draw conclusions
  - A graphical representation is often the means in which a hypothesis is confirmed or results are communicated.
  - Describe everything, draw attention to major features, describe conclusions

Explain everything in the plot.
Do not let them guess.

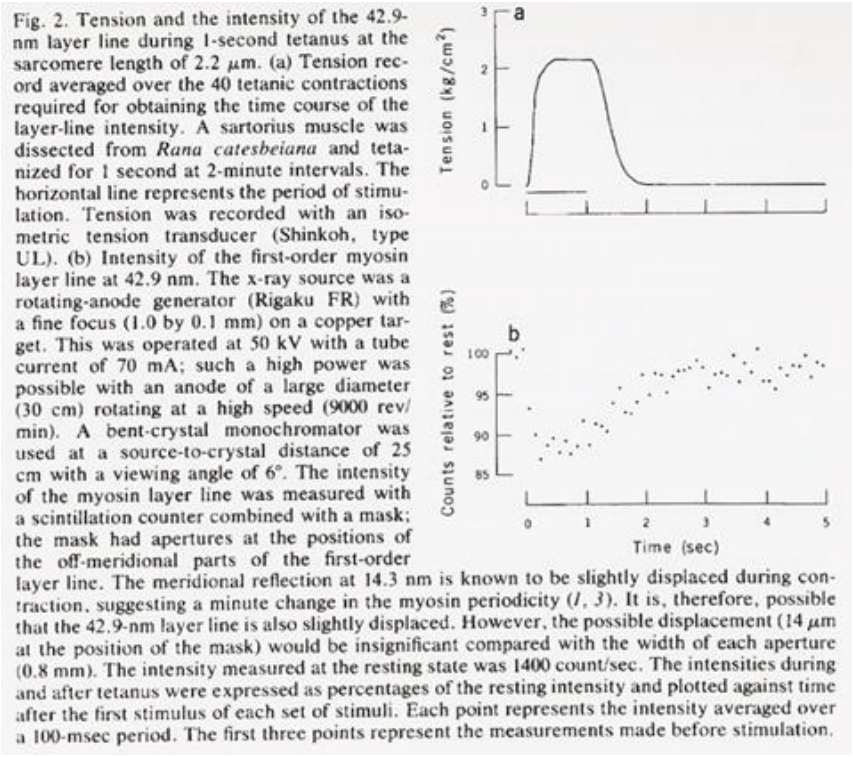
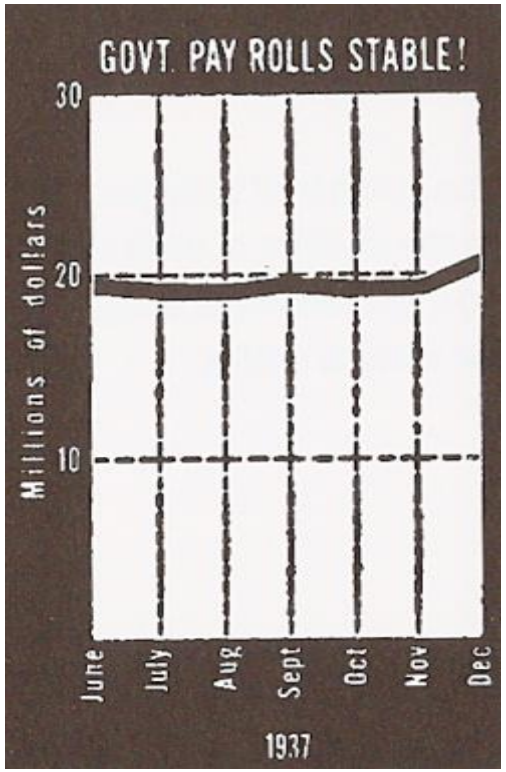
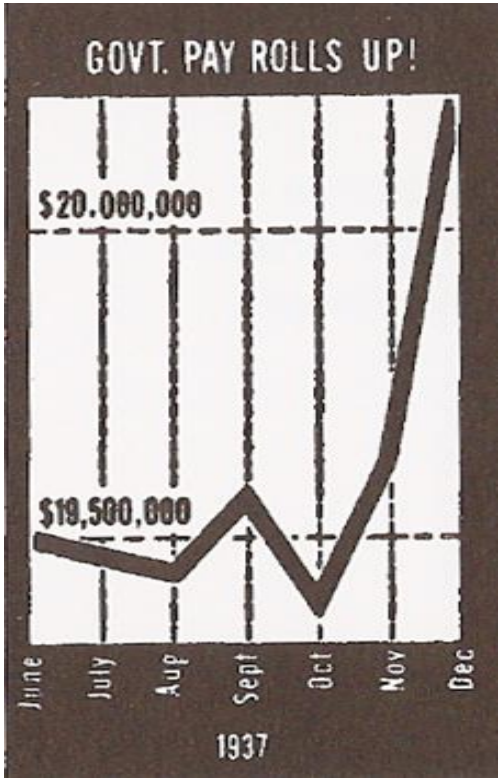


Fig. 2. Tension and the intensity of the 42.9-nm layer line during 1-second tetanus at the sarcomere length of 2.2 μm. (a) Tension record averaged over the 40 tetanic contractions required for obtaining the time course of the layer-line intensity. A sartorius muscle was dissected from *Rana catesbeiana* 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.

Improving the Understanding

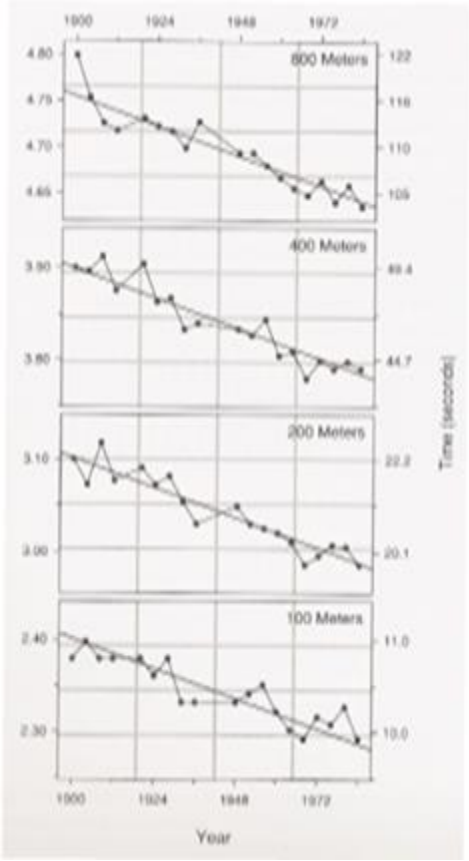
- Principle 2: Use all available space.
  - Fill the data rectangle as much as you can, only use zero if you need it (*for scientific data*)

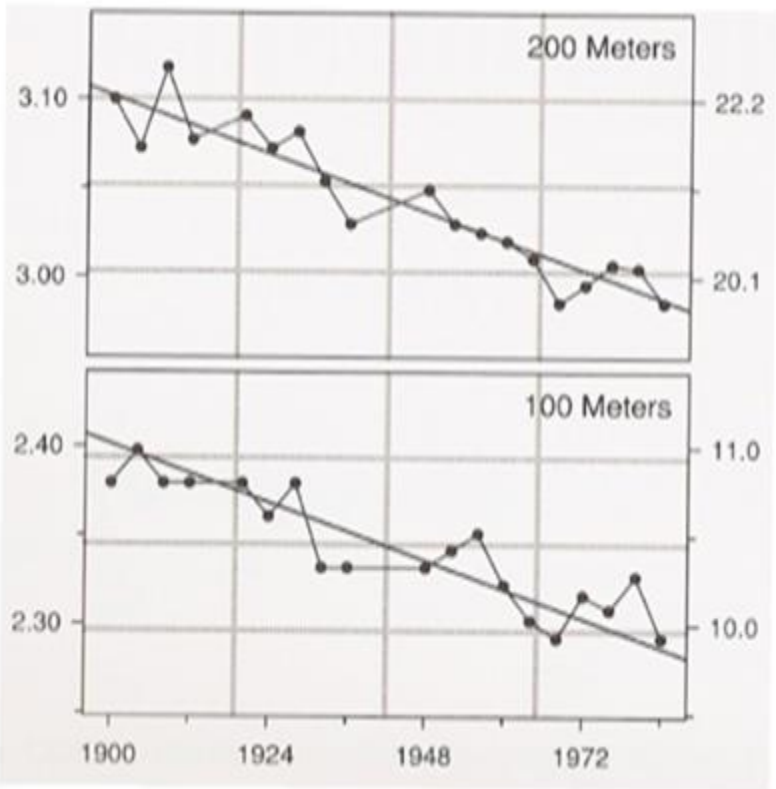




Improving the Understanding

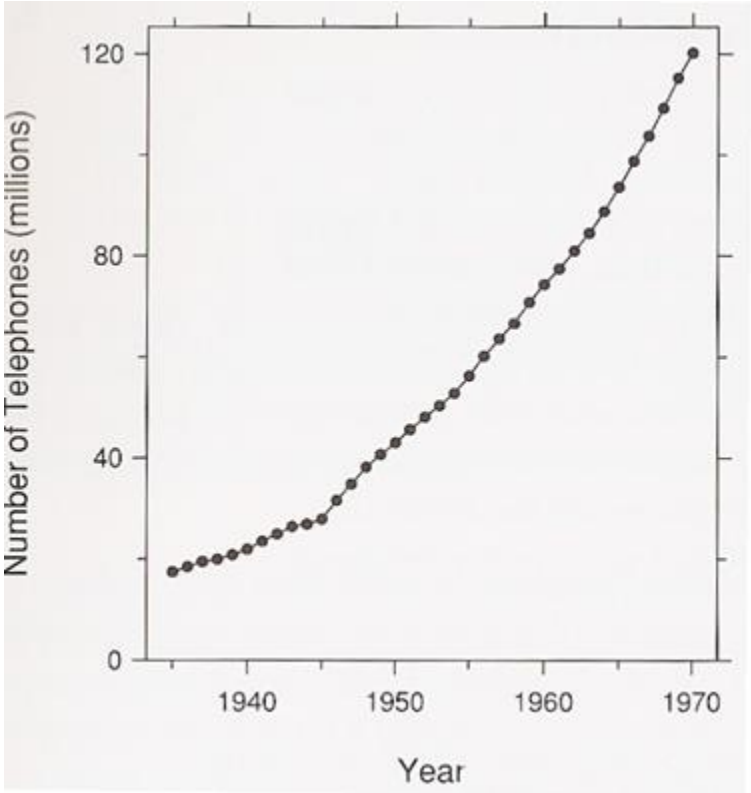
- Principle 3: Align juxtaposed plots
  - Make sure scales match and graphs are aligned

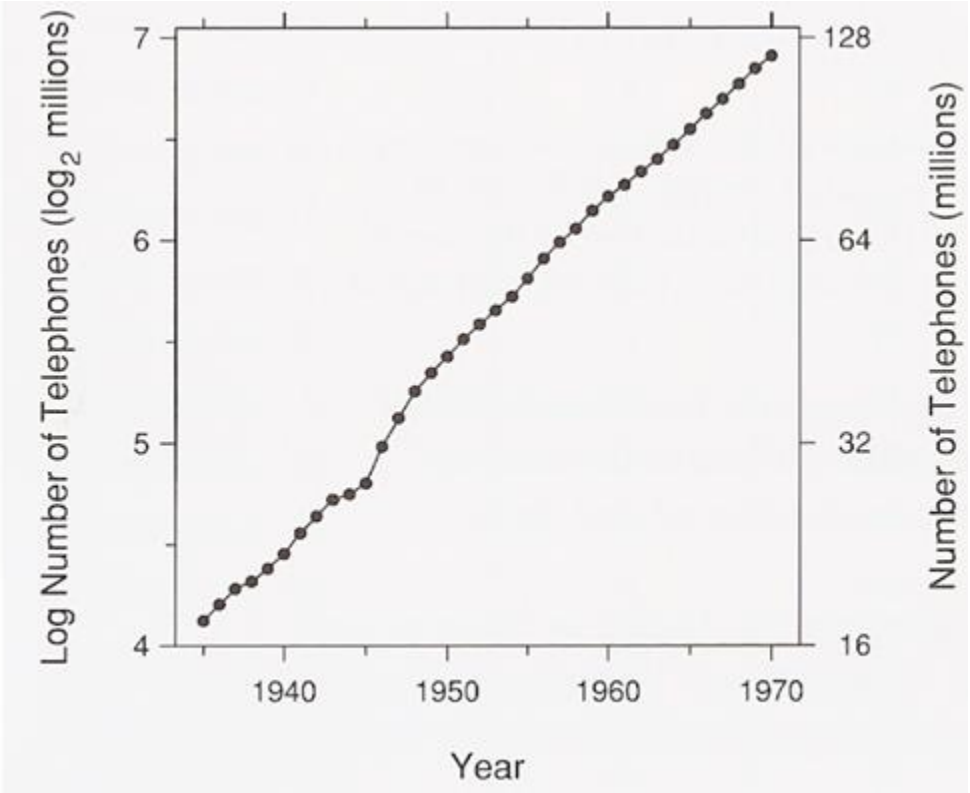




Improving the Understanding

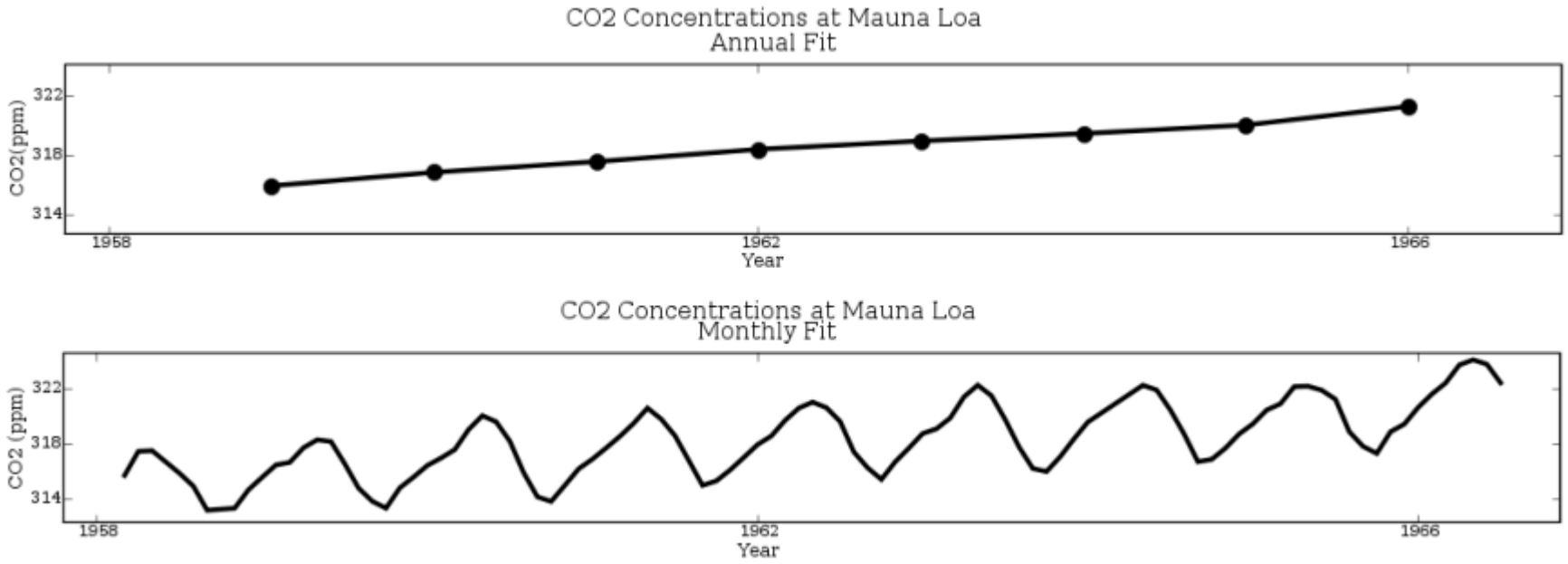
- Principle 4: Use log scales when appropriate
  - Used to show percentage change, multiplicative factors and skewness





Improving the Understanding

- Principle 5: Bank to 45^o(optional!!!)
  - Optimize the aspect ratio of the plot



Summary of Principles

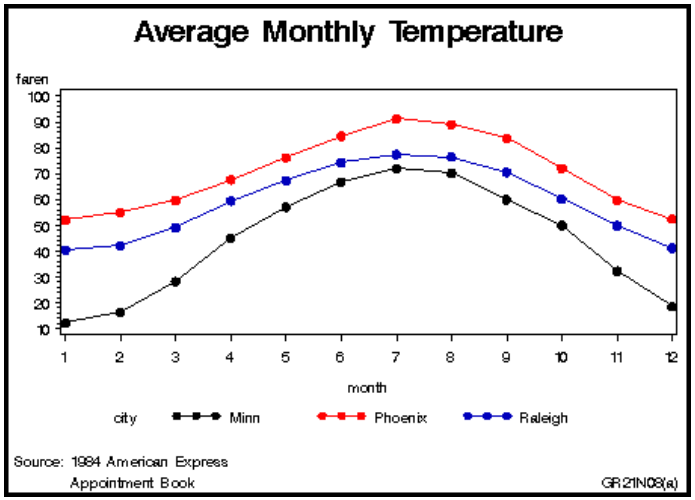
- Improve vision
  1. Reduced clutter, Make data stand out
  2. Use visually prominent graphical elements
  3. Use proper scale lines and a data rectangle
  4. Reference lines, labels, notes, and keys
  5. Superposed data set

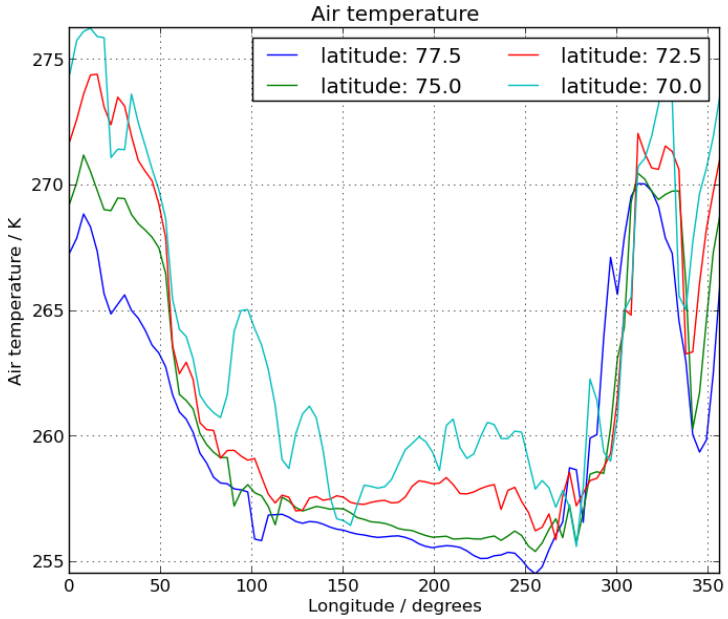
- Improve understanding
  1. Provide explanations and draw conclusions
  2. Use all available space
  3. Align juxtaposed plots
  4. Use log scales when appropriate
  5. Bank to 45^o

SIMPLE PLOTTING TECHNIQUES

Connected Symbol Plots

- The most common plotting technique
- Used to plot time series or other 1D data



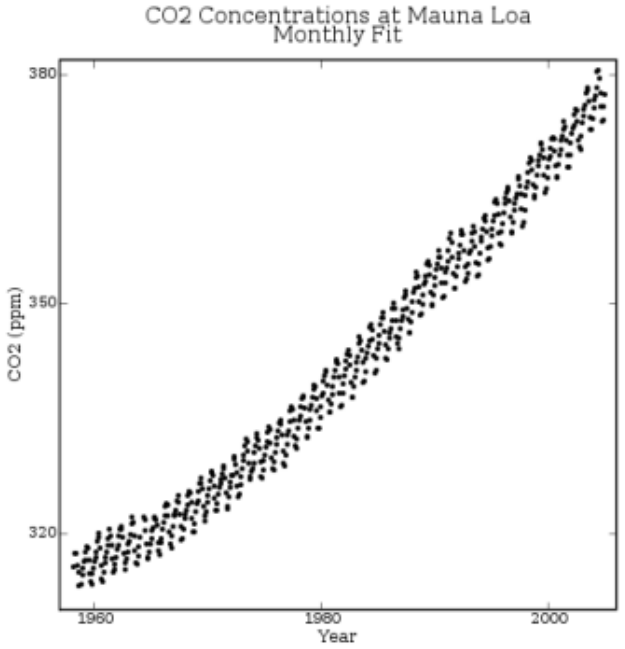


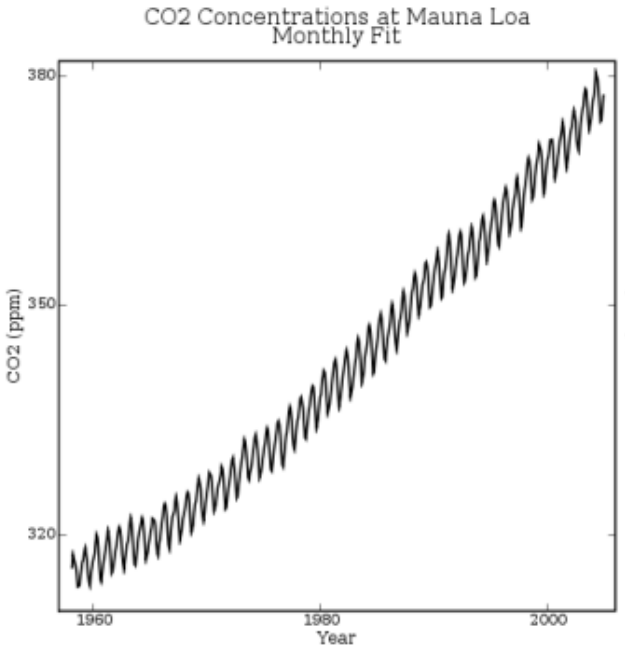
Connected Symbol Plots

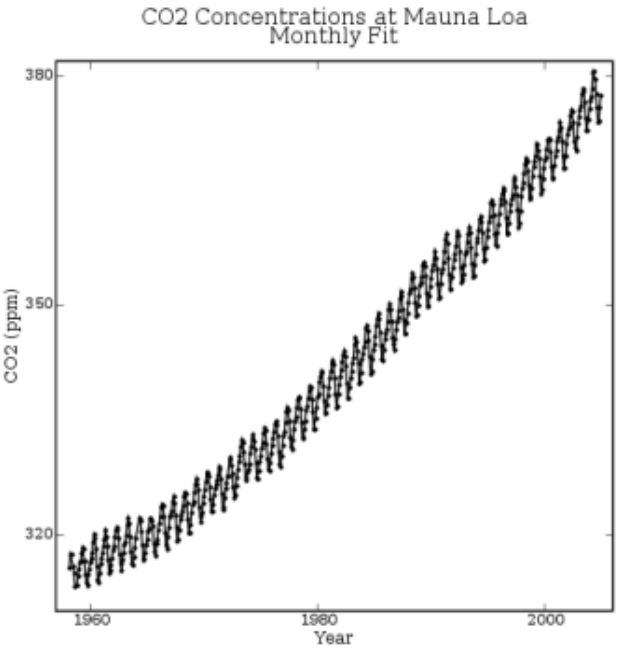
Symbols. For noisy data that shows high frequency characteristics

Connections. For smooth data that shows low frequency characteristics

Connected Symbols. The symbols demonstrate the actual concentrations of the data, while the path that the data takes can be better followed using connections.





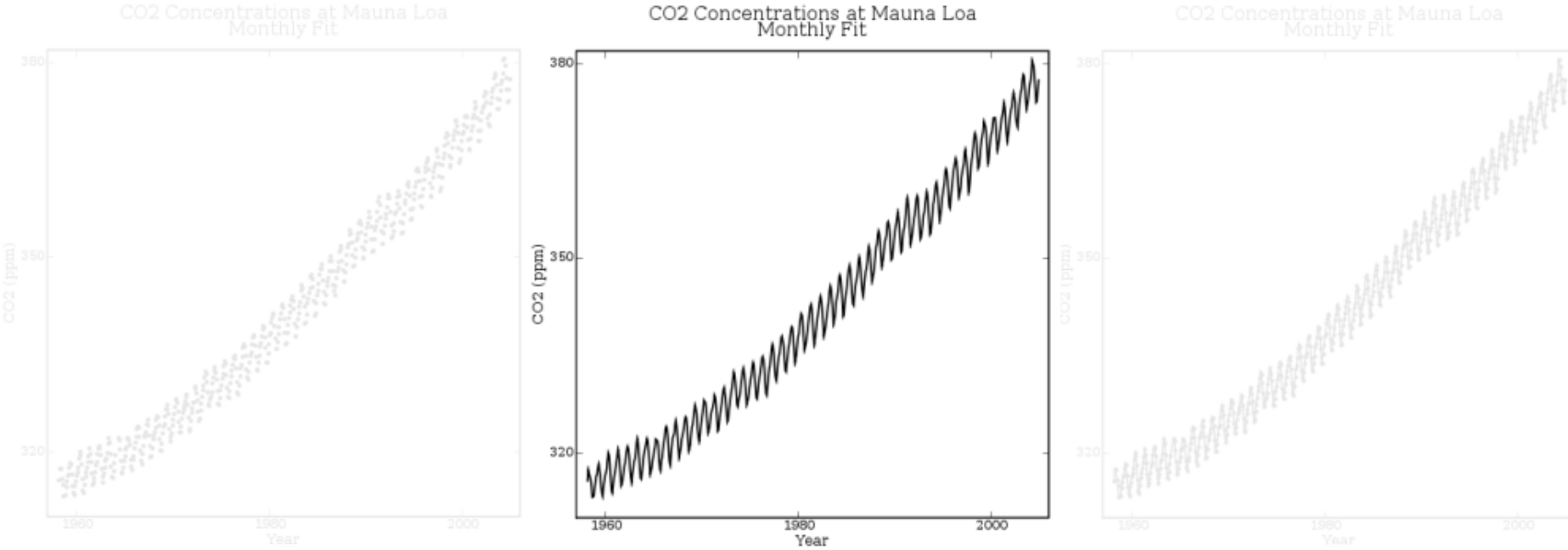


Connected Symbol Plots

Symbols. For noisy data that shows high frequency characteristics

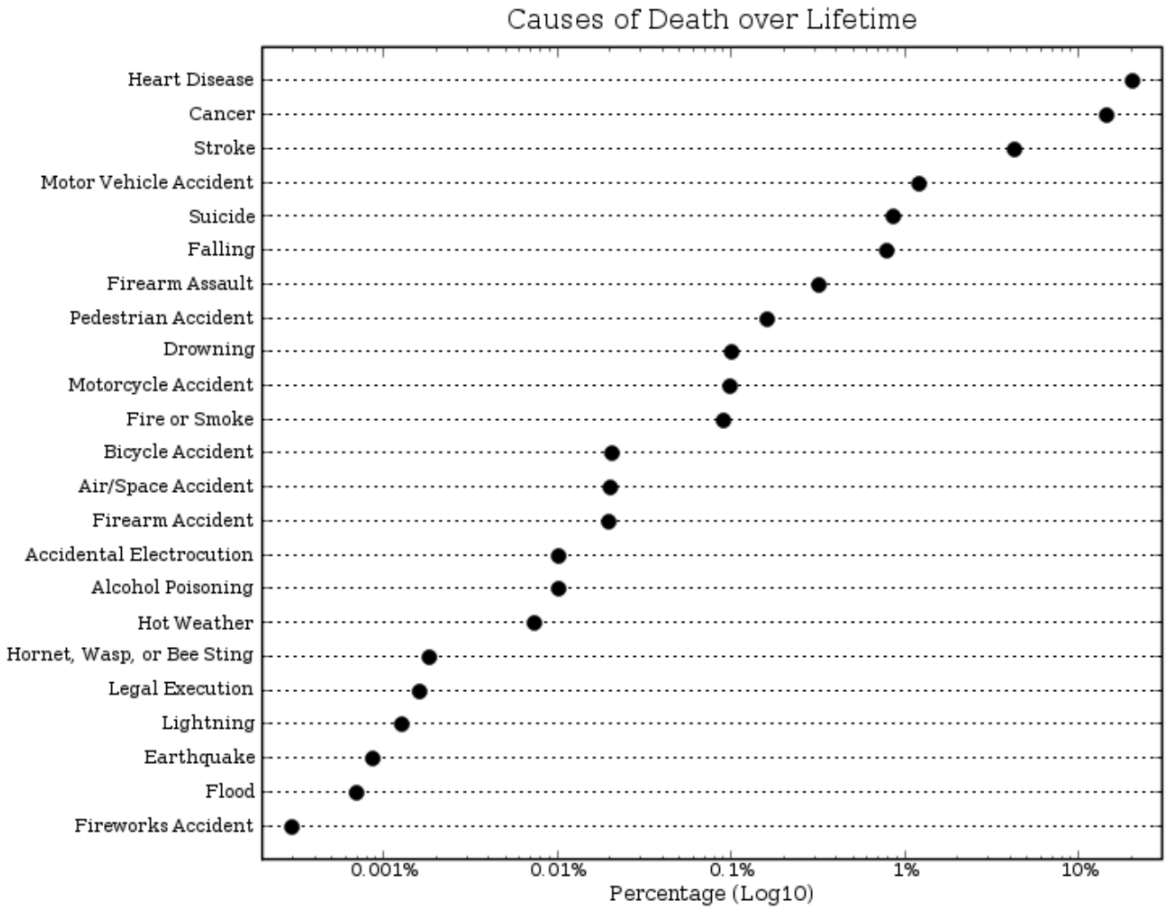
Connections. For smooth data that shows low frequency characteristics

Connected Symbols. The symbols demonstrate the actual concentrations of the data, while the path that the data takes can be better followed using connections.



Dot Plots

- Similar in nature to bar charts or pie charts
- Should be used for quantitative labeled data



The data points do not have sequential relation!!

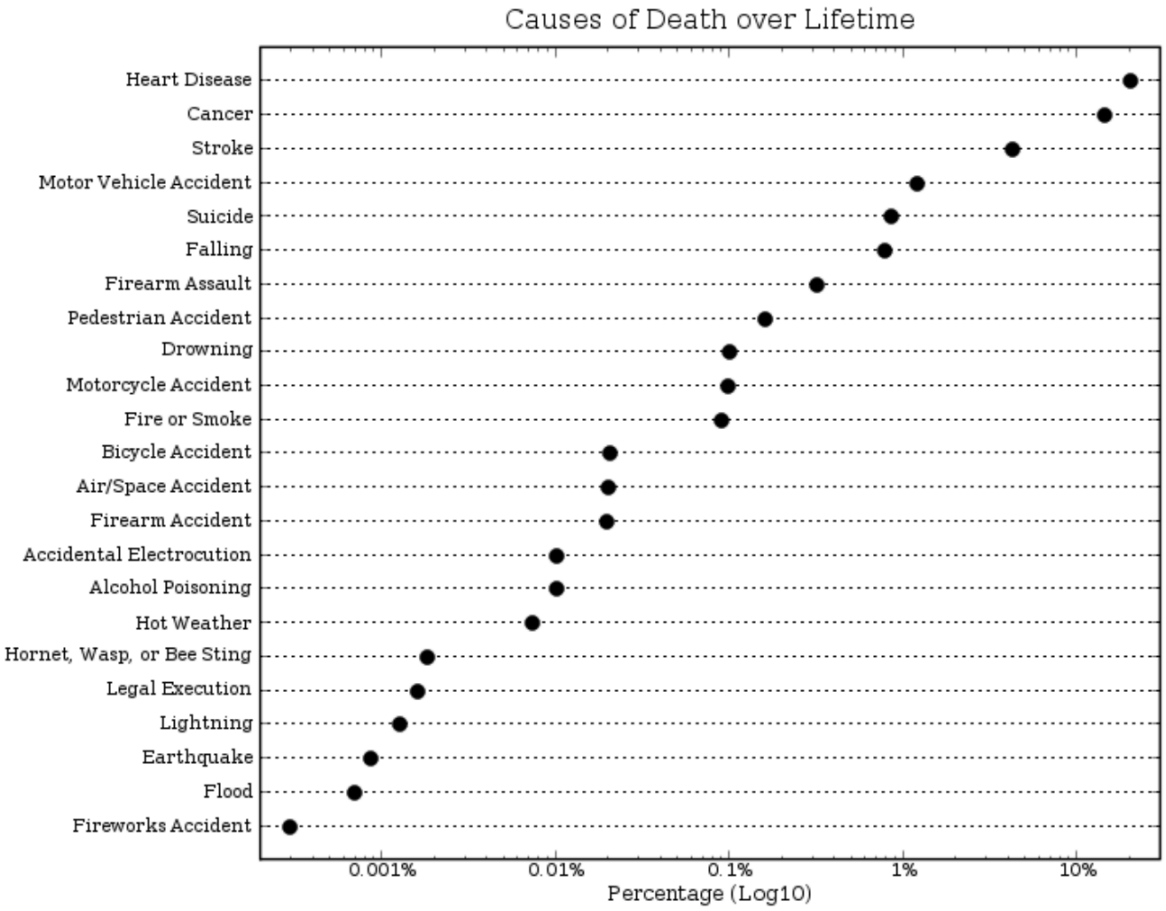
A dot plot showing the odds of dying.

Dot Plots

The values should normally be sorted such that the largest value is at the top.

- Exception: the data has an inherent order that must be preserved

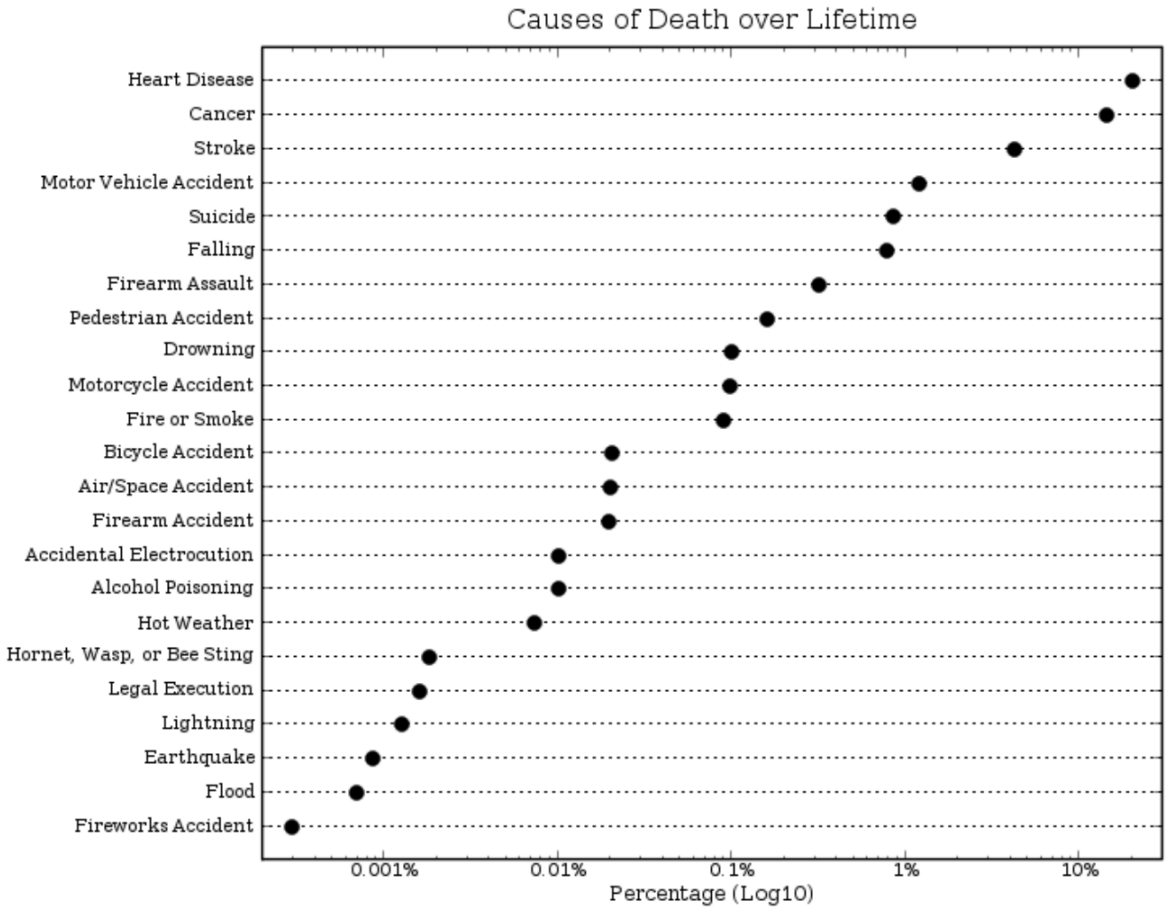
A log scale should be used to reduce skewness in the data



A dot plot showing the odds of dying.

Dot Plots

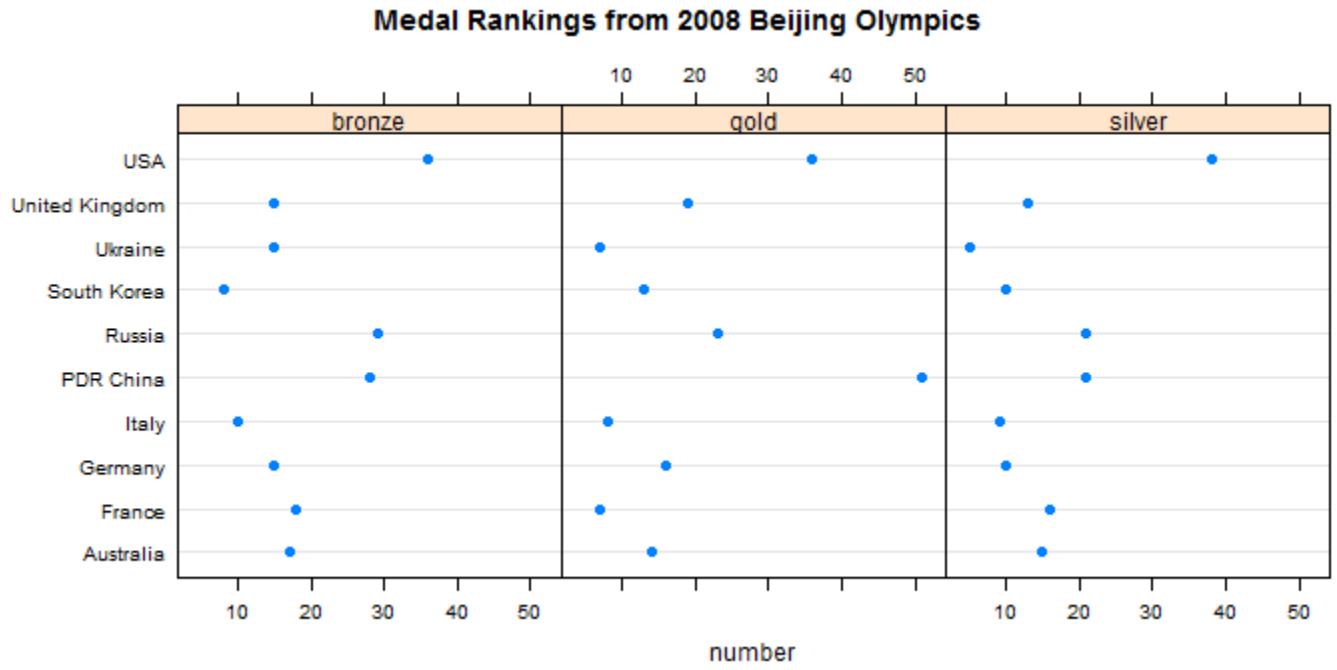
- Real world data is not always univariate.
- To represent multi-dimensional data, a multiway dot plot can be used



A dot plot showing the odds of dying.

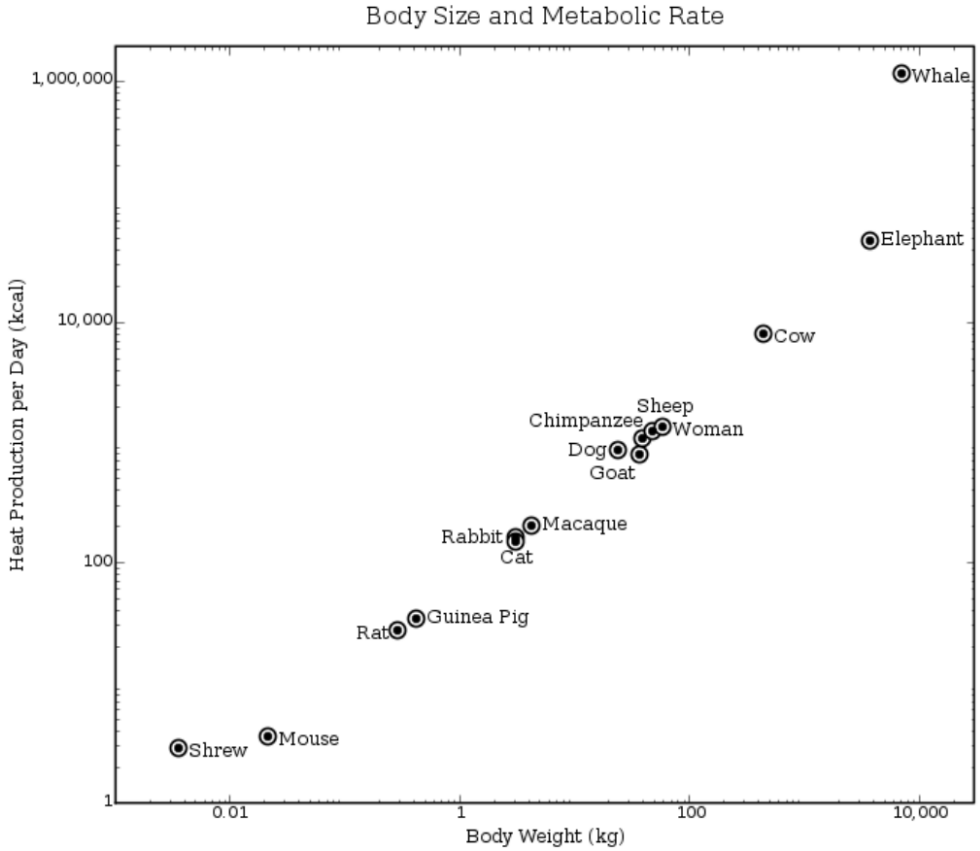
Dot Plots

- A multiway dot plot is just several dot plots that share common labels and are juxtaposed such that they share an axis.



Scatter Plots

- Scatter plots are used to show how one variable is affected by another, or correlated, in 2D data.

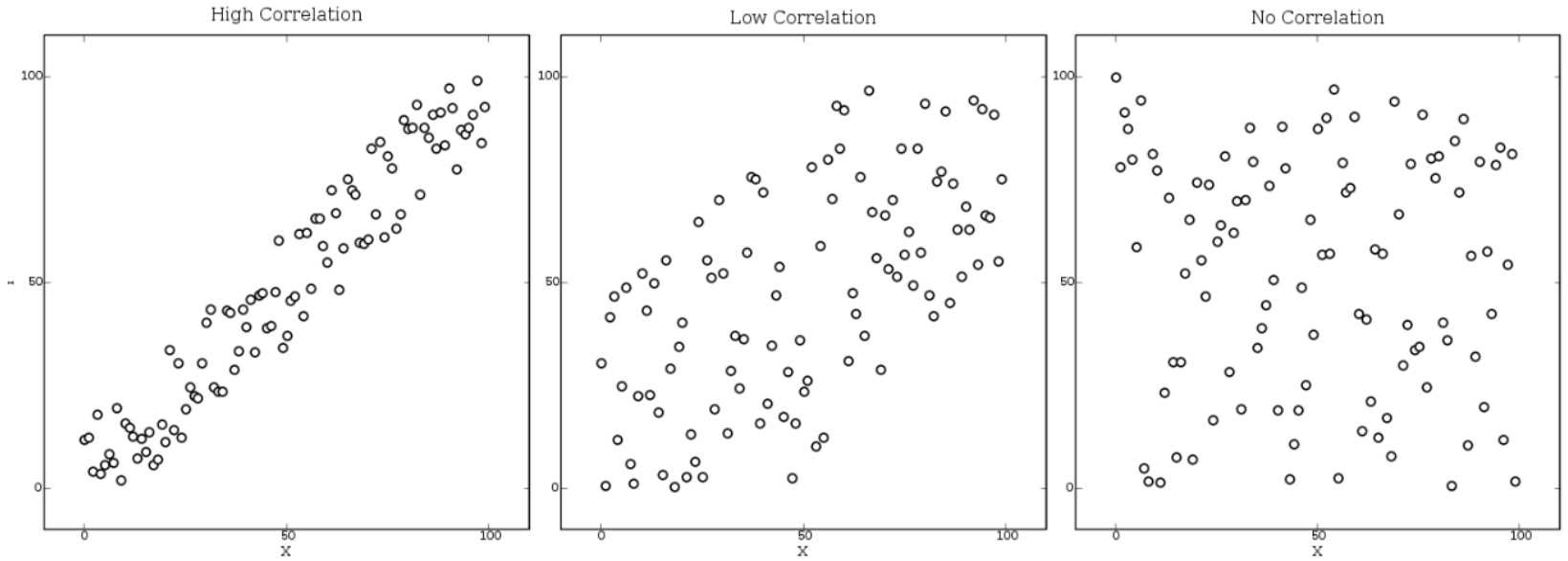


Need to make the symbols in the data stand out and keep the labels from obscuring the data and making the trend difficult to perceive

A scatter plot showing the biological principle of scaling for mammals. For each sample, the metabolic rate is plotted against the body mass to show a high correlation between the two variables. The points have also been labeled to provide additional information.

Scatter Plots

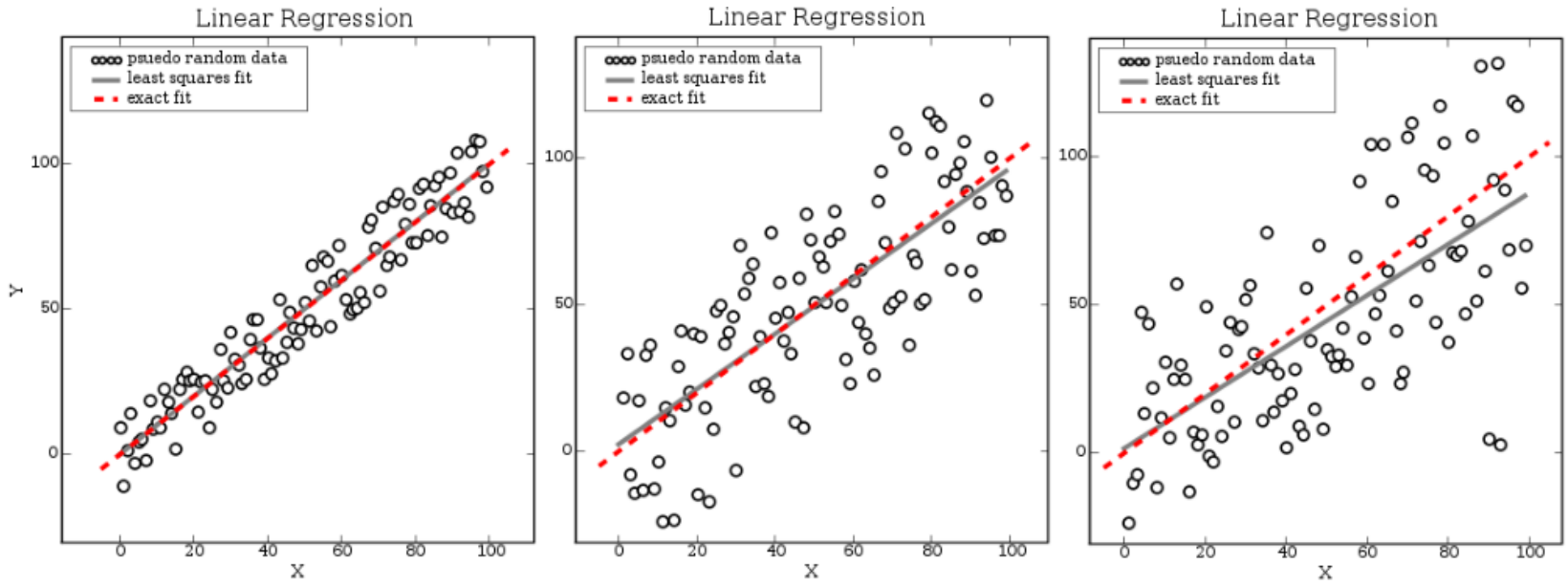
- If used properly, the correlation of the data can easily be discerned.



Scatter plots showing different levels (high, low, and no, respectively) of correlation for points generated with different magnitudes of randomness.

Scatter Plots

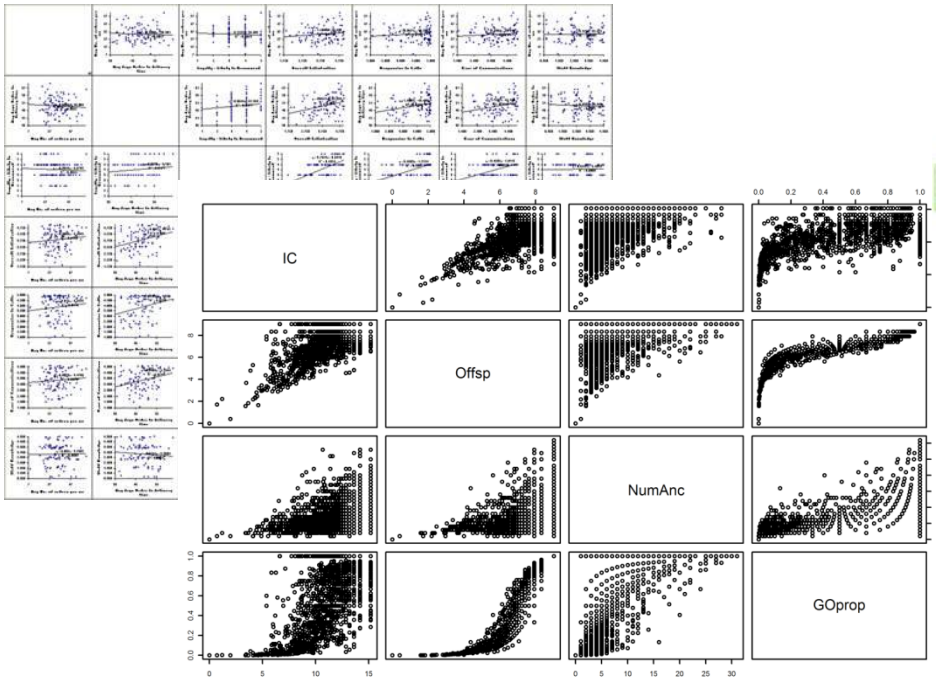
- It is often desirable to express the correlation as a line that provides the best fit for the data.

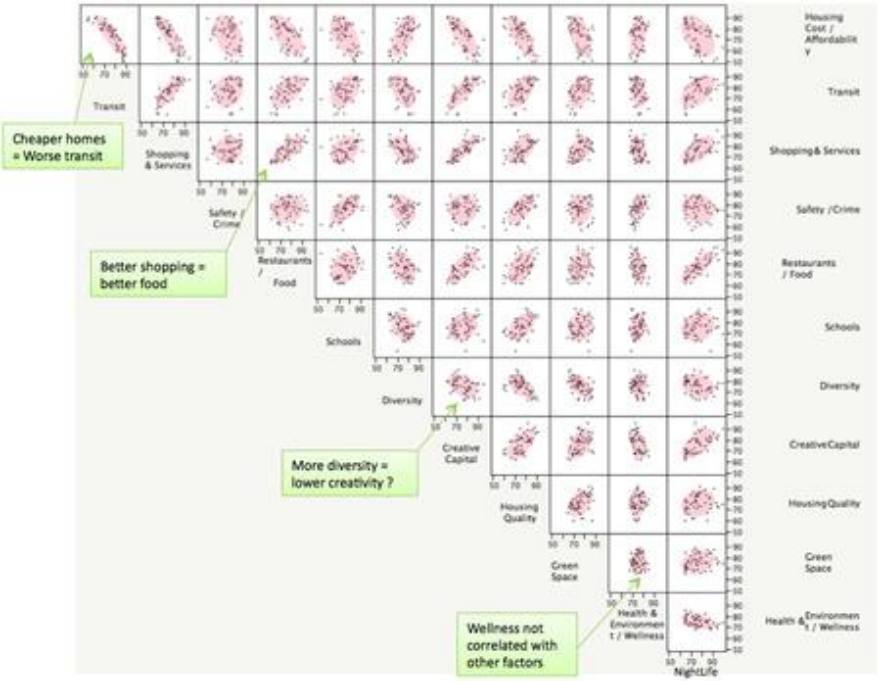


Linear regression using least squares fits a line to the data. The fit is good for high and low correlation (left and middle), but can result in problems in the case of outliers (right)

Scatter Plots

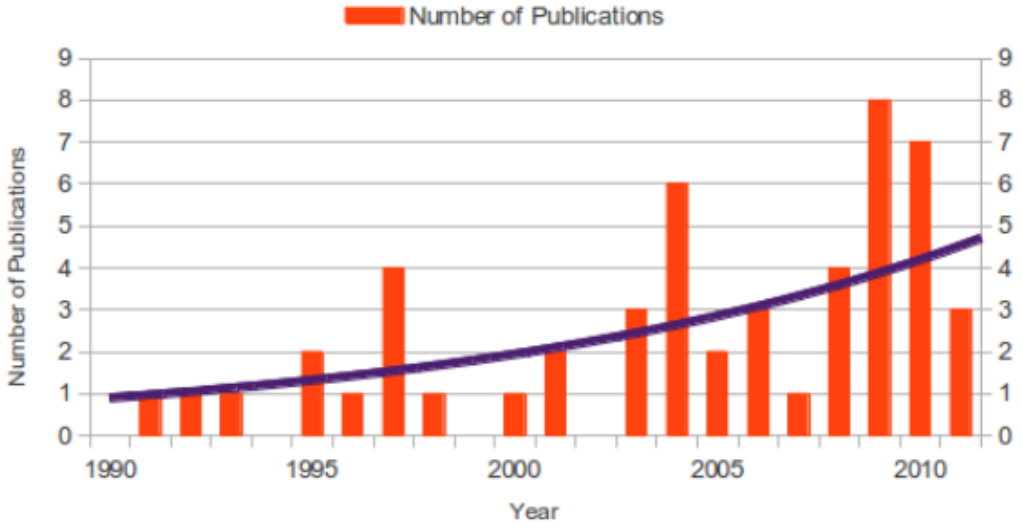
- As with dot plots, scatter plots can be used to represent data in higher dimensions. This is frequently done with a scatter plot matrix.
- This assigns each dimension of the plot to a single row and column in the matrix. The variables are then plotted against each other as a standard scatter plot for each entry in the matrix.

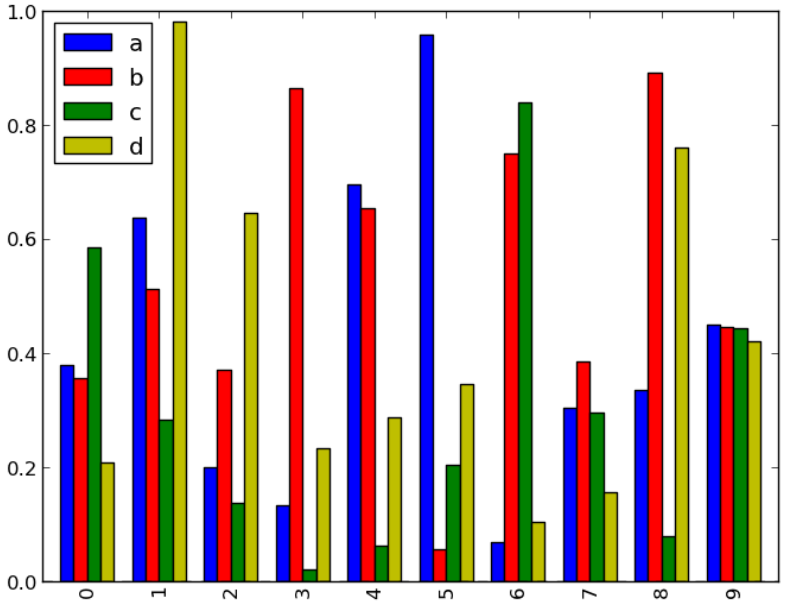




Histograms

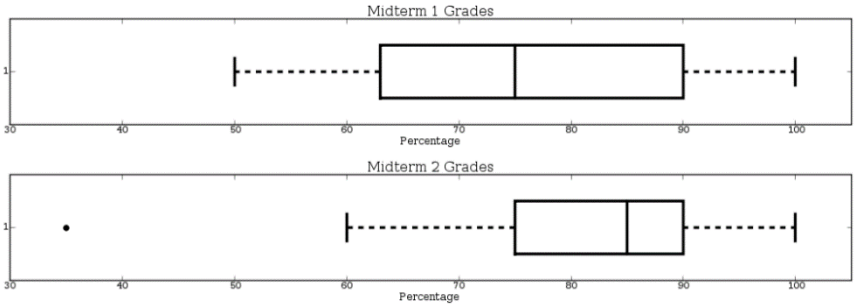
- Histograms are a special type of bar charts used for plotting distributions in data.
- The horizontal axis represents fixed intervals of the data and the vertical axis represents the number of values that lie within the intervals.

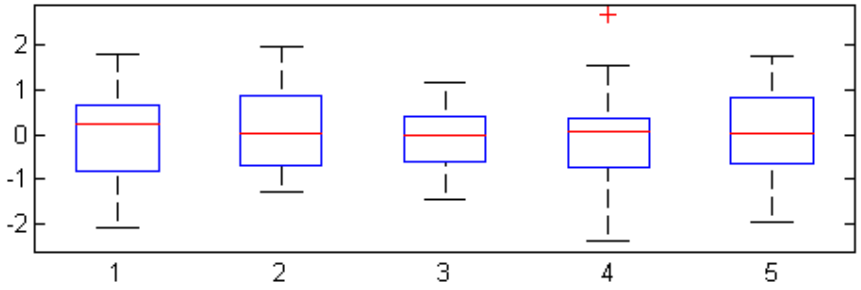


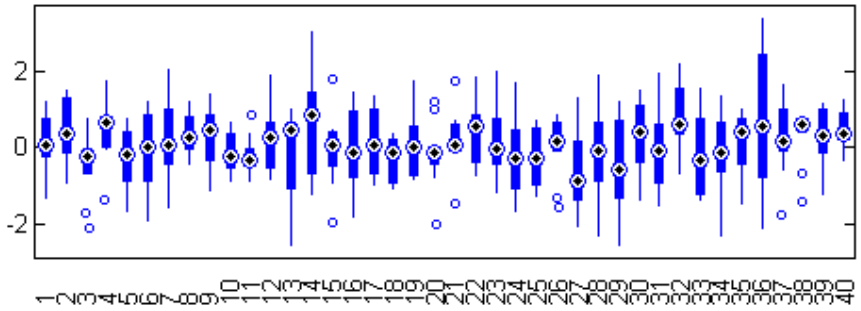


Box Plots

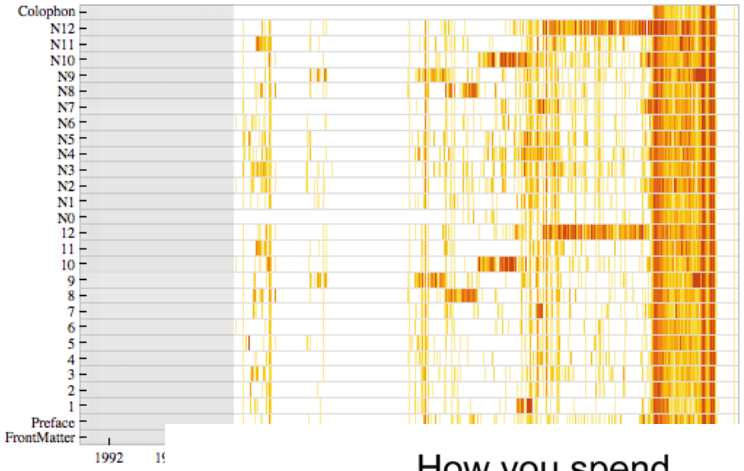
- Box plots are typically used to represent the statistical variation in the data

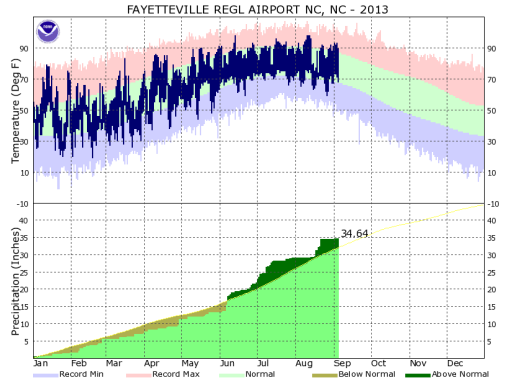






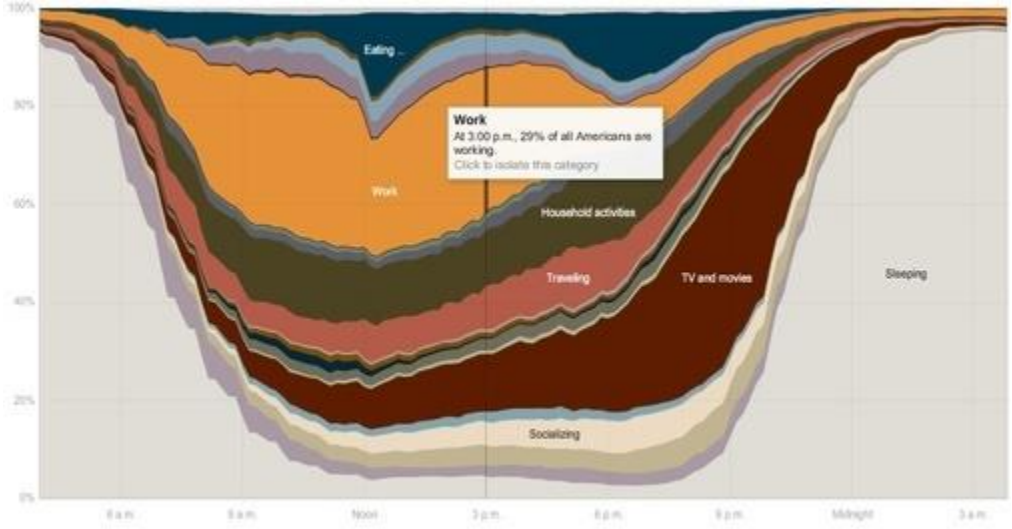
Others


Colophon
N12
N11
N10
N9
N8
N7
N6
N5
N4
N3
N2
N1
N0
12
11
10
9
8
7
6
5
4
3
2
1
Preface
FrontMatter
1992


FAYETTEVILLE REGL AIRPORT NC, NC - 2013
Temperature (Deg F)
Precipitation (Inches)
34.64
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Record Min
Record Max
Normal
Below Normal
Above Normal


Katrina's Diaspora


How you spend your time:
"Other" (2.4 hours)
a.k.a. Facebook
Napping
Sleeping in class
Sleeping (8.5 hours)
Usually Hungover
Uneasy sleep
Lie awake pondering your life choices
Surfing the Internet
Actual Sports
Leisure & Sports (3.7 hours)
Reading about Sports while surfing Internet
Work & Related Activities (2.9 hours)
"Related Activities" (shooting the breeze with your officemates)
"Work"
Surfing the Internet
Educational Activities (3.3 hours)
Eating & Drinking (1.0 hours)
"Stuff"
Free Food
Boozing
Snacking
"Grooming" (0.5 hours)
Toilet time
Texting
Traveling (1.5 hours)
SOURCE: Bureau of Labor Statistics, American Time Use Survey of Full Time University Students.
JORGE CHAM © 2013
WWW.PHDCOMICS.COM


Eating
Work
At 3:00 p.m., 29% of all Americans are working.
Click to isolate this category
Household activities
Traveling
TV and movies
Sleeping
Socializing
100%
80%
60%
40%
20%
0%
6 a.m.
9 a.m.
Noon
3 p.m.
6 p.m.
9 p.m.
Midnight
3 a.m.

Brief Overviews of Types of Graphs

2D Graphs
- Bar/Column
- Bar Dev
- Bar Left Y
- Bar Right Y
- Bar Top
- Bar X
- Box
- Detrended Probability
- Half-Normal Probability
- Hanging Bar Histograms
- Histograms
- Line
- Pie Charts
- Probability
- Probability-Probability
- Quantile-Quantile
- Range
- Scatterplots
- Sequential/Stacked
- Voronoi Scatterplot

3D XYZ Graphs
- Contour
- Deviation
- Scatterplots
- Space
- Spectral
- Trace

3D Sequential Graphs
- Bivariate Histograms
- Box
- Range
- Raw Data Contour/Discrete
- Sequential Contour
- Sequential Surface
- Raw Data Spikes
- Raw Data Surface

4D/Ternary Graphs
- Scatterplots
- 3D Ternary
- Contour/Area
- Contour/Line
- 3D Deviation
- 3D Space

2D Categorized Graphs
- Detrended Probability
- Half-Normal Probability
- Normal Probability
- Probability-Probability
- Quantile-Quantile

3D Categorized Graphs
- Contour
- Deviation
- Scatterplots
- Space
- Spectral
- Surface

Ternary Categorized Graphs
- Ternary Contour/Area
- Ternary Contour/Line
- Ternary Scatterplot

nD/Icon Graphs
- Chernoff Faces
- Columns
- Lines
- Pies
- Polygons
- Profiles
- Stars
- Sun Rays

Matrix Graphs
- Columns
- Lines
- Scatterplot

http://www.statsoft.com/Textbook/Graphical-Analytic-Techniques