Graphs

Aaron Rendahl
slides by Sanford Weisberg & G. Oehlert

School of Statistics
University of Minnesota

February 9, 2009

Good Graphics

Fundamental Principal of Statistical Graphics

Above all else show the data.

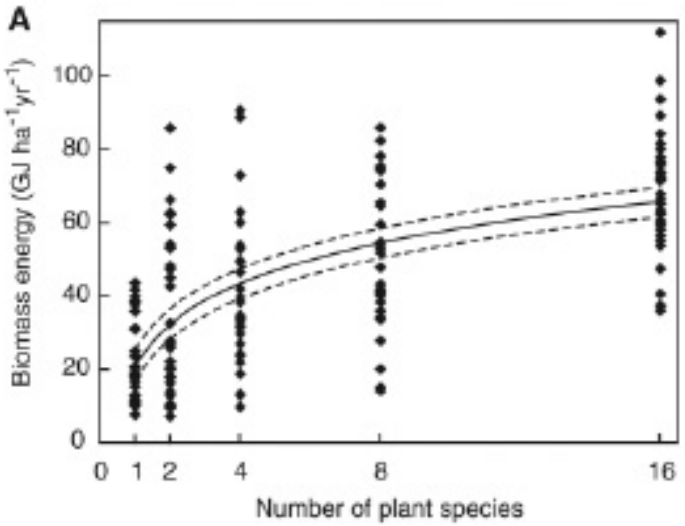
Ed Tufte

Graphics can be

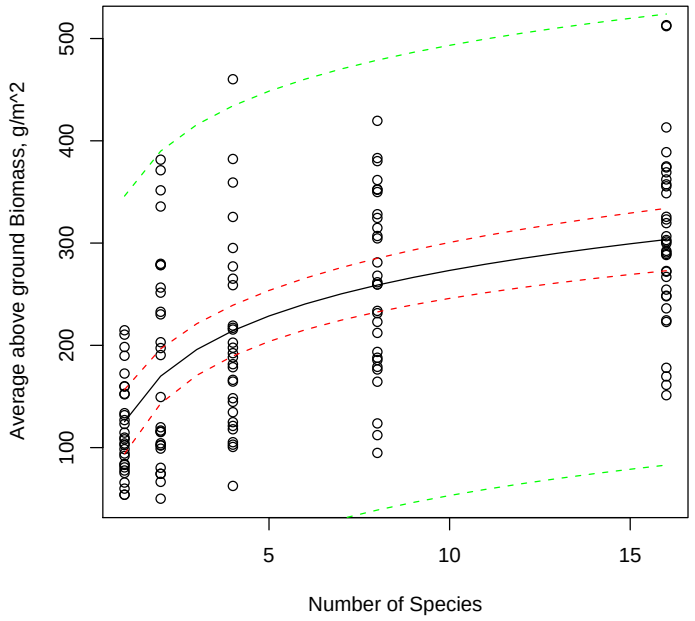
- . . . all that is read in an article
- . . . efficiently summarize a problem
- . . . very aesthetic
- . . . misleading or otherwise awful

We must use them well, or else who will?

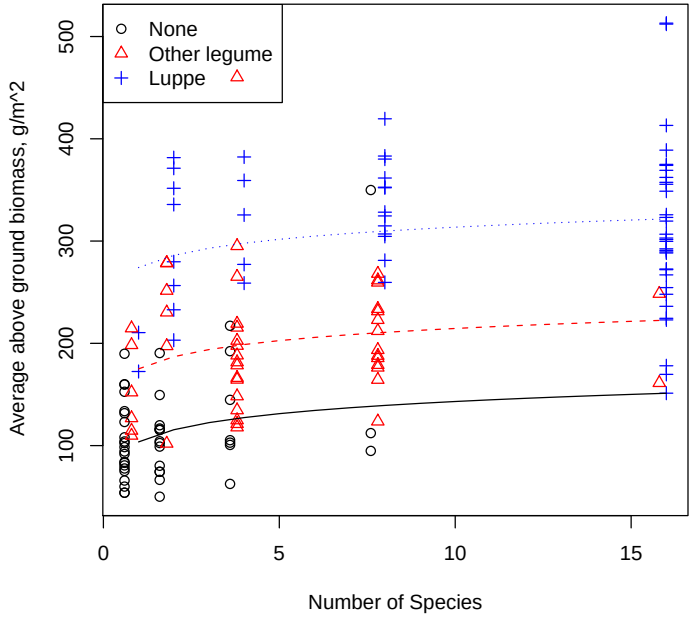
From Tilman, Hill and Lehman (2006) *Science*, p. 1598



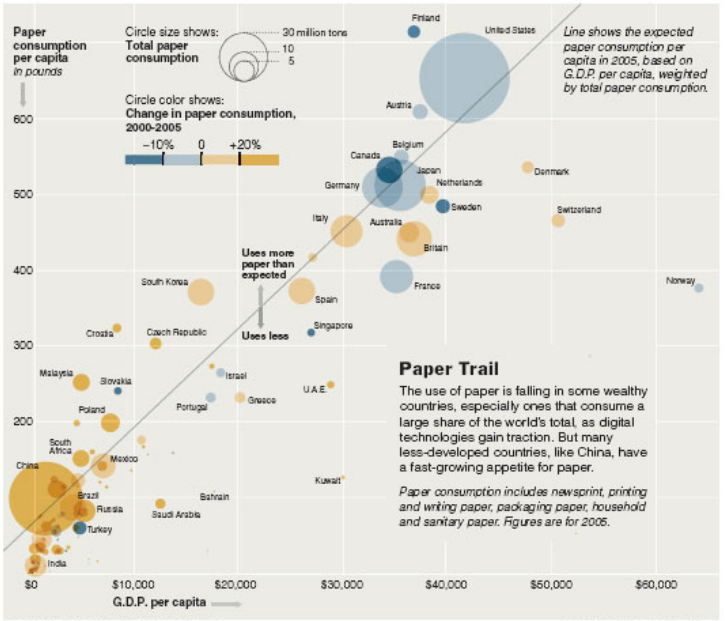
. . . adding prediction intervals



. . . adding species indicator



Paper usage, New York *Times*, Feb. 10, 2008



Bush Tax Cuts, New York *Times*, Feb. 10, 2008

The Cost of Allowing Tax Cuts to Expire ...

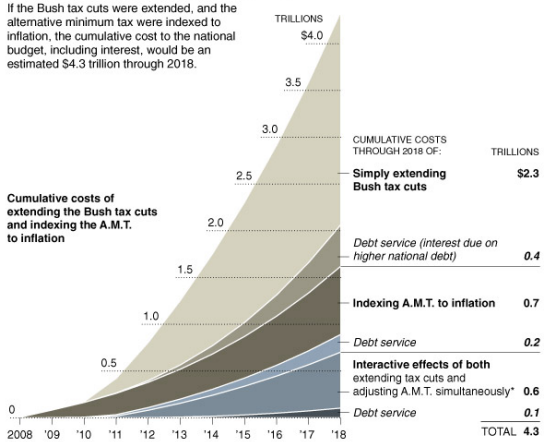
If the Bush tax cuts were allowed to expire as scheduled at the end of 2010, here are some of the potential consequences for taxpayers:

	CURRENTLY	IF ALLOWED TO EXPIRE
INCOME TAX RATES 2008 TAXABLE INCOME	CURRENT TAX RATE	NEW TAX RATE
$0 to $8,025	10%	15%
$8,025 to $32,550	15%	15%
$32,550 to $78,850	25%	28%
$78,850 to $164,550	28%	31%
$164,550 to $357,700	33%	36%
$357,700 and over	35%	39.6%
CAPITAL GAINS AND DIVIDENDS		CAPITAL GAINS / DIVIDENDS
15% bracket or below	0%	10% / same as income tax rate
Brackets above 15%	15%	20% / same as income tax rate
ESTATE TAX	Estates of $2 million or less are exempt, with a top rate of 45%. Tax declines and then disappears in 2010, but only for one year.	Estates of $1 million or less would be exempt, with a top rate of 55%.
CHILD TAX CREDIT	$1,000 per child	$500 per child
SO-CALLED MARRIAGE PENALTY	Deduction for a married couple is 200% of that for a single person.	Deduction for a married couple would be 167% of that for a single person.

Source: Tax Policy Center, Congressional Budget Office

... And the Cost of Extending Them

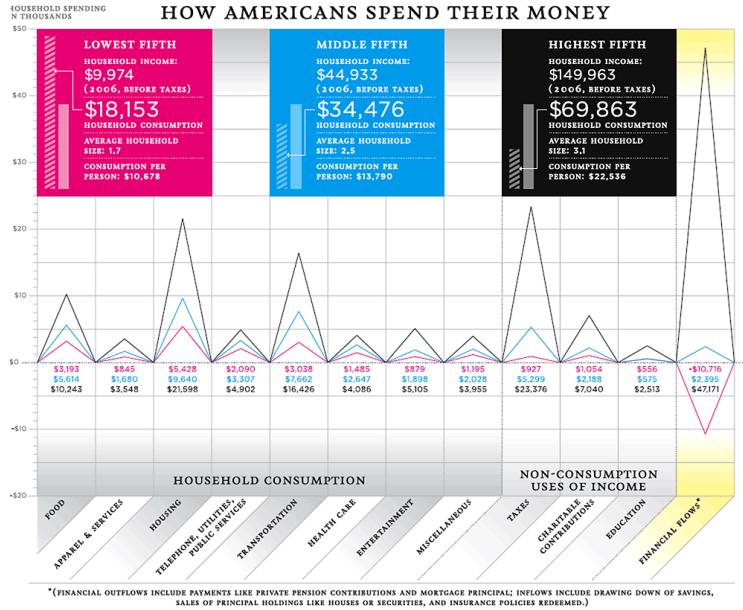
If the Bush tax cuts were extended, and the alternative minimum tax were indexed to inflation, the cumulative cost to the national budget, including interest, would be an estimated $4.3 trillion through 2018.



*The Bush tax cuts greatly expand the number of taxpayers affected by the A.M.T. If the cuts are extended, while at the same time the A.M.T. is restricted, the cost goes up.

THE NEW YORK TIMES

New York *Times*, Feb. 10, 2008



The Aesthetics of Graphics

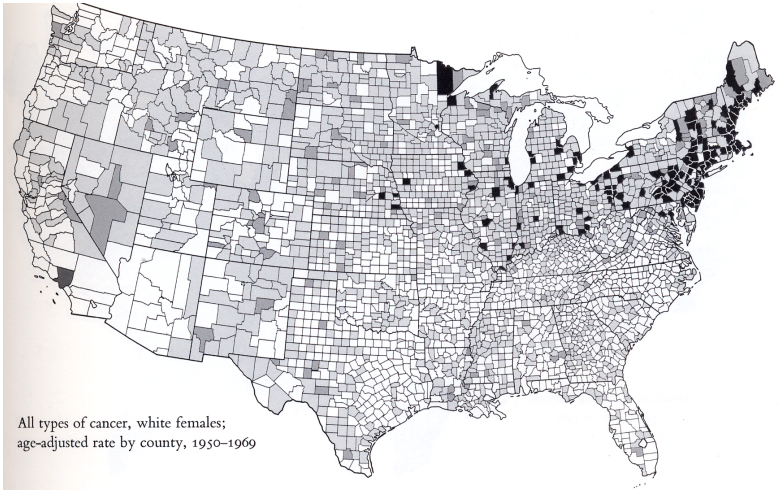
Ed Tufte is at the top of the pantheon of statistical graphics gods.

Tufte has three **extremely** influential books on graphics.

Not everyone agrees with Tufte, but no one can ignore him.
Other important sources:

- Lee Wilkenson (**The Grammar of Graphics**)
- Bill Cleveland (**The Elements of Graphing Data**)
- Howard Wainer (lots of articles)

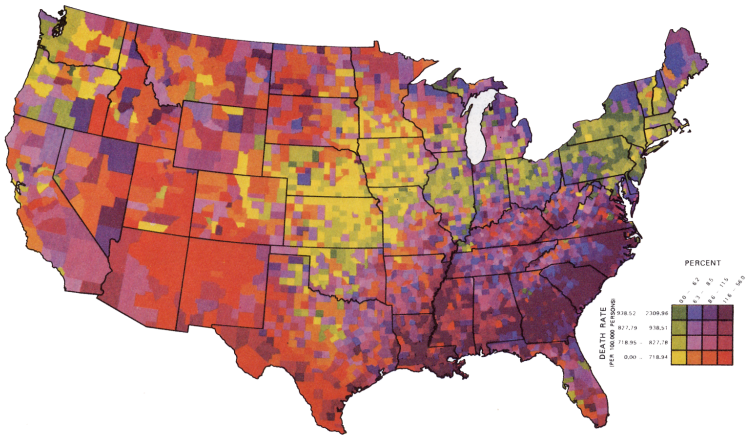
Map of Cancer Rates



All types of cancer, white females;
age-adjusted rate by county, 1950–1969

Avoid puzzles

Try to figure this one out

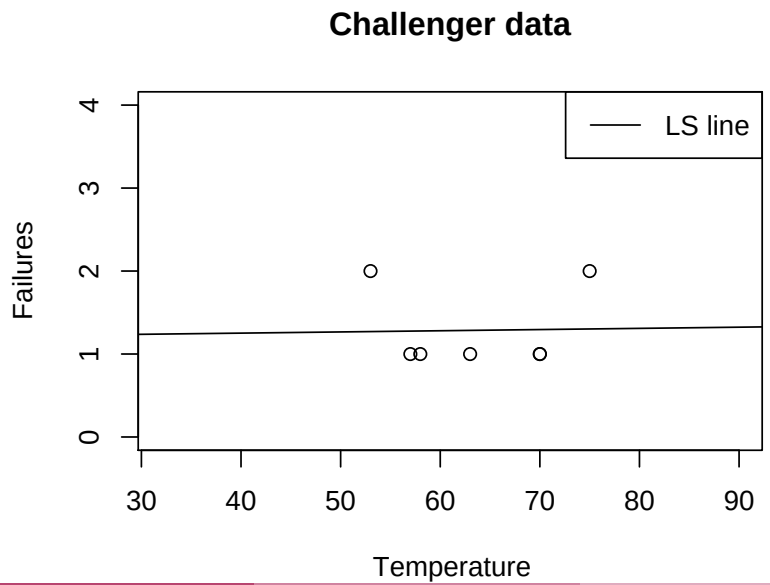


John Snow, Cholera & the Broad St. Pump

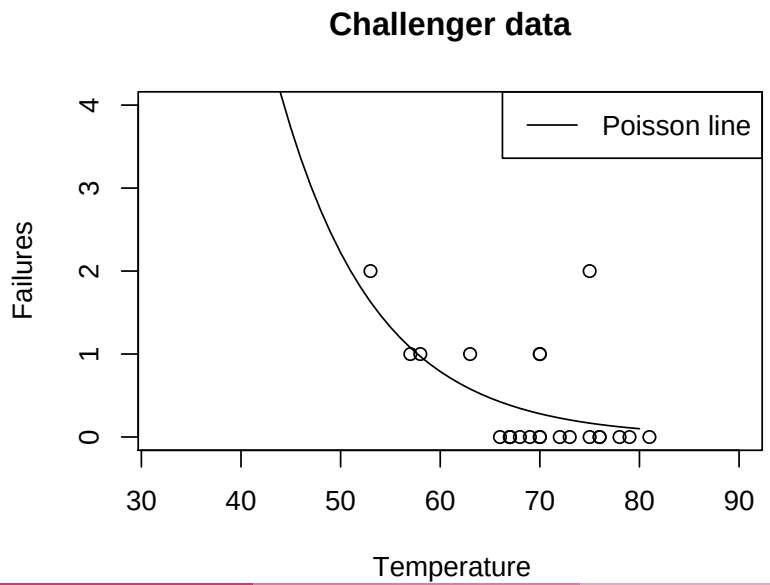
The best graphs ever

www.economist.com/printedition/
displayStory.cfm?Story_ID=10278643

The Worst Graph Ever



What they should have done



Technique

- Graphs may be the only part of an article that is read.
- Good format and design
  Aesthetics, elegance, and style difficult to prescribe.
  Construct, revise, edit, try again
- Words/numbers/graphics together
- Data graphics are paragraphs about numbers (Tufte, p 181).
- Graphics and tables must always reinforce message and text.

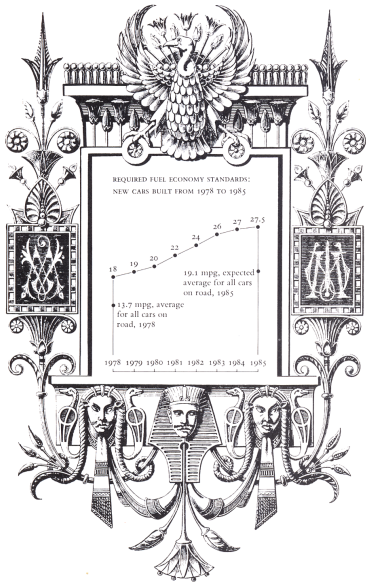
Don't...

1. ...Mislead
2. ...Use mysterious abbreviations
3. ...Include too much clutter (forest for the trees)
4. ...Misuse placement of origin
5. ...Include graphs without explanation
6. ...Use gratuitous color/line variation
7. ...SHOUT (use all capital letters)
8. ...use chart junk
9. ...use pie charts

Do...

1. ...use accessible friendly graphic
2. ...include axis labels, titles and legends
3. ...use sensible tick marks
4. ...facilitate comparisons between graphs by using common scales.
5. ...avoid unclear abbreviations.

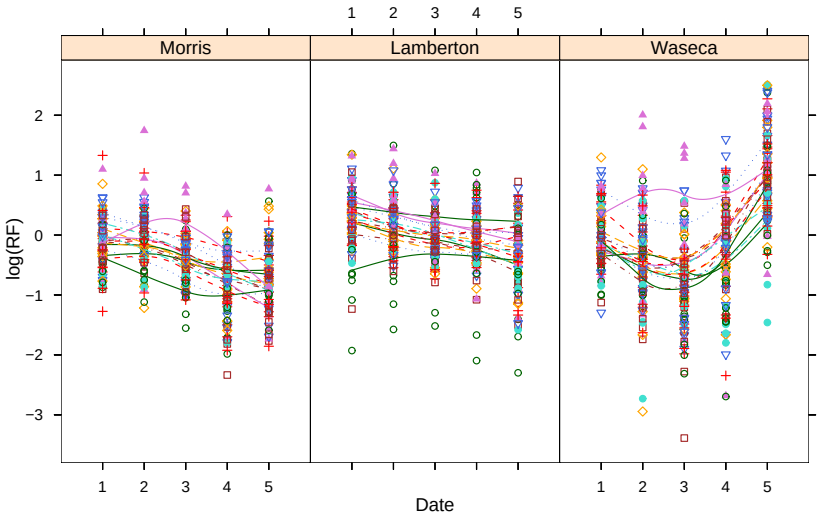
Content-free decoration



Graphs in R

1. Basic graphs use plot, pairs, boxplot
   - Uses sensible defaults, but not always
   - Reasonably, but not completely, flexible
2. Lattice graphics
   - Very aesthetic and moderately flexible
   - Very hard to use well
3. ggplot2
   - I've not used it
   - Should be very flexible and maybe easier to use

A lattice graph




```
xyplot(log(RF)~Date|Location,data=scn1,groups=Treatment,
   auto.key=FALSE,layout=c(3,1),
   panel=function(x,y,subscripts,...){
      panel.superpose(x,y,subscripts,...) },
   panel.groups=function(x,y,...){
      panel.loess(x,y,...)
      panel.xyplot(x,y,...)}
   )
```

I couldn't figure out how to get a reasonable legend added to the plot to name the colors/symbols, or how to label dates.

Tufte's Data Ink

Definition (Data ink)

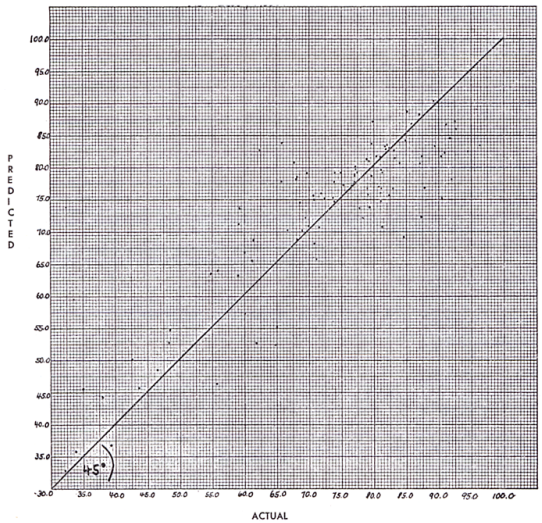
Data ink is the "ink" that displays non-redundant data information.

Definition (Data ink ratio)

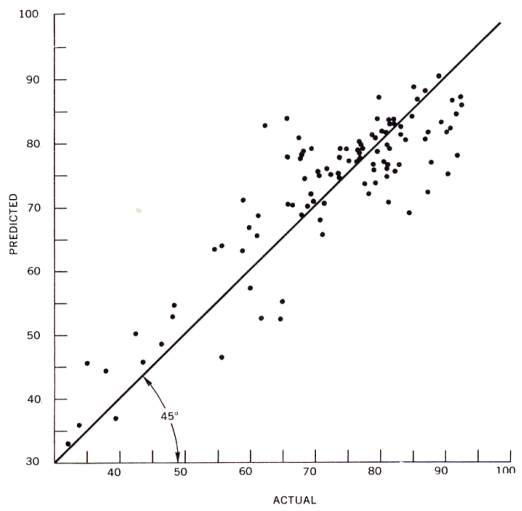
Proportion of a graphic's ink devoted to the non-redundant display of data information.

1. Maximize data ink ratio, within reason
2. Erase non data ink, within reason
3. Erase redundant data ink, within reason

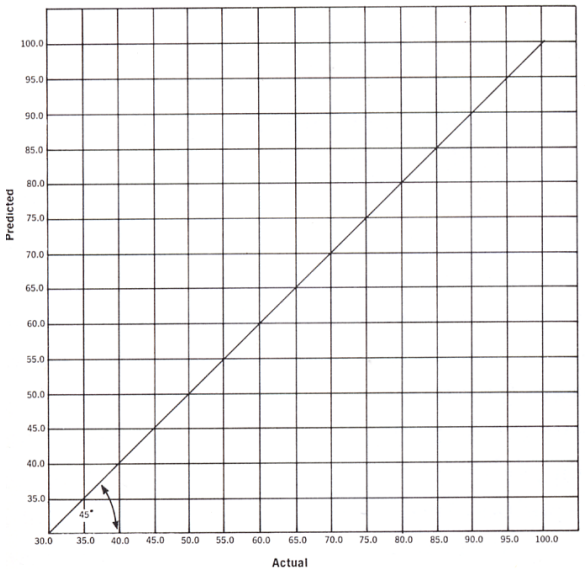
Bad data-ink ratio



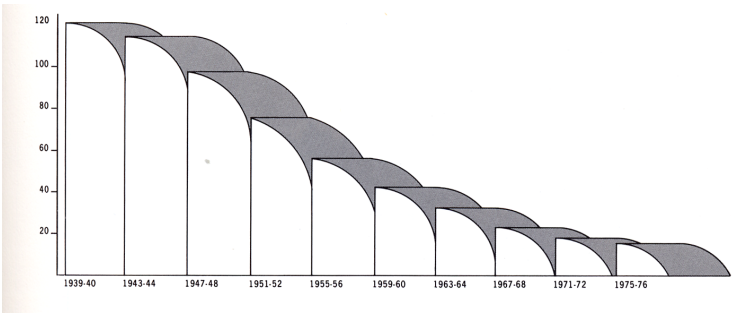
Good data-ink ratio



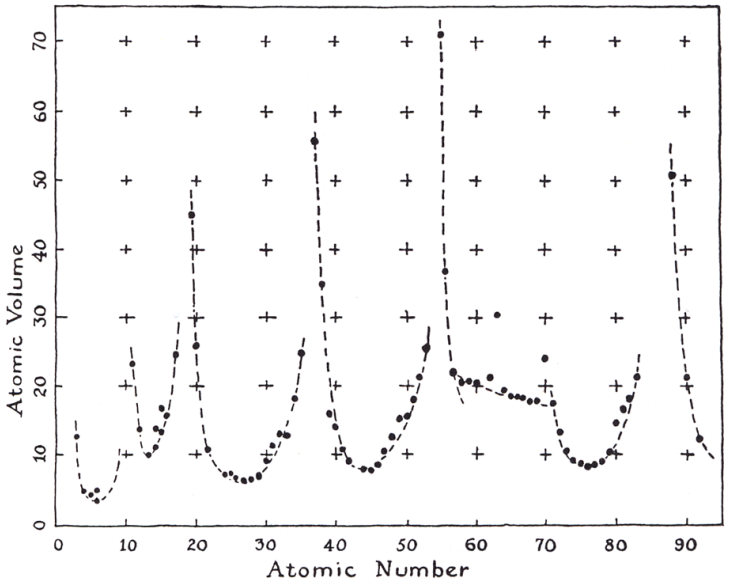
Zero data-ink ratio



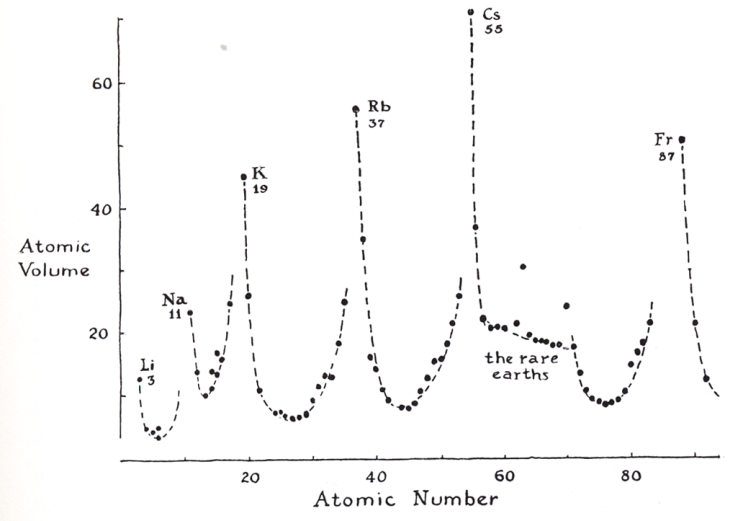
Erasable non-data ink



Erasable non-data ink



Improved non-data ink



Mighty Ducks

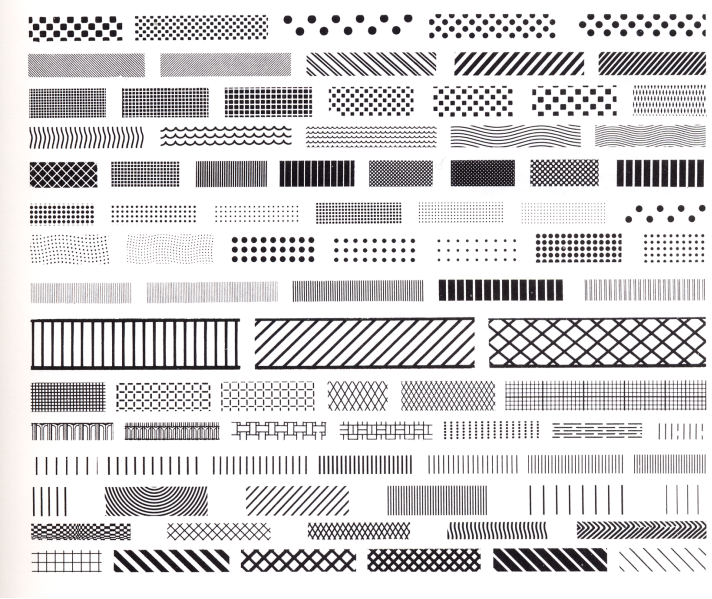
Non-data ink can be chartjunk.

Could be shading, hatching, grid, etc.

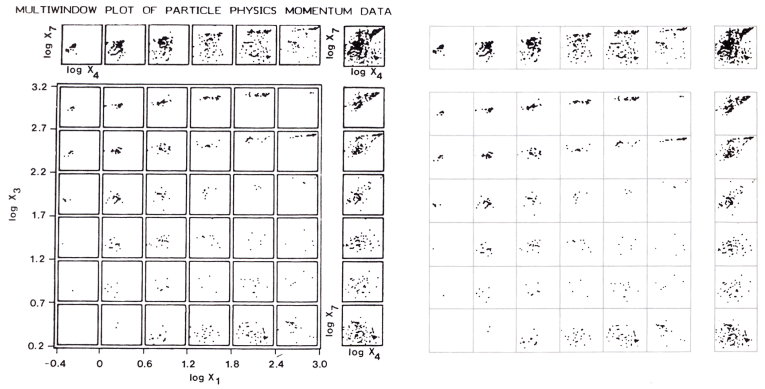
Really egregious examples are "ducks".

Get rid of it!

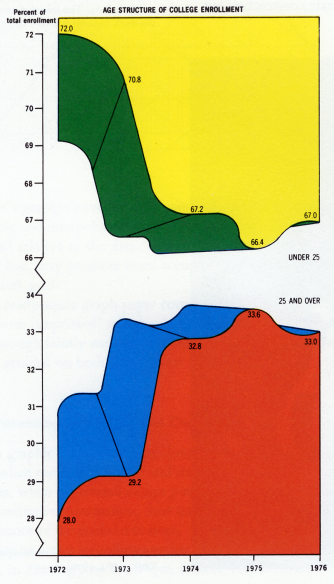
Moiré patterns



Data, not frames



Quack



Don't lie with graphics

Lies, damned lies, and statistics

could also be

Lies, damned lies, and graphics.

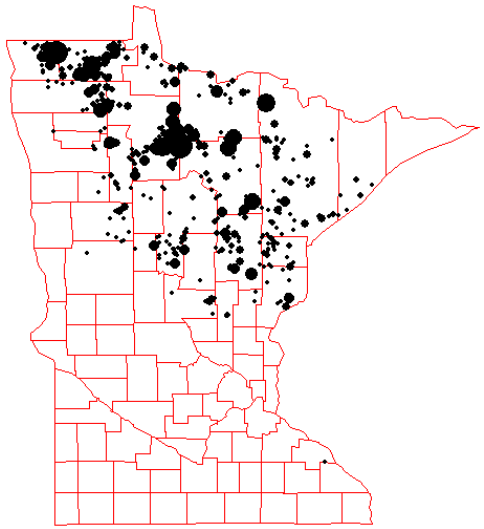
What can we do to avoid misleading?

Data, area and dimension

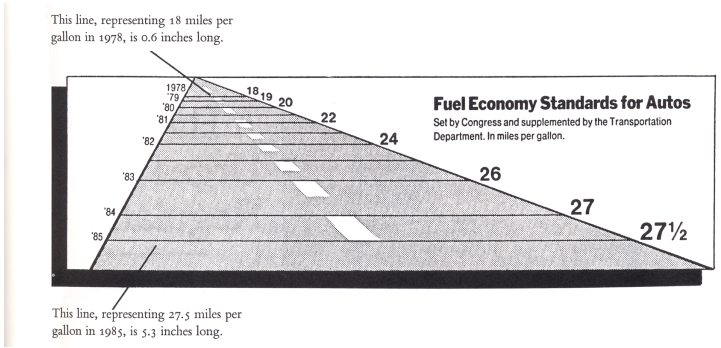
The size of the representation of a number should be proportional to the number

The number of information carrying dimensions should not exceed the dimension of the data.

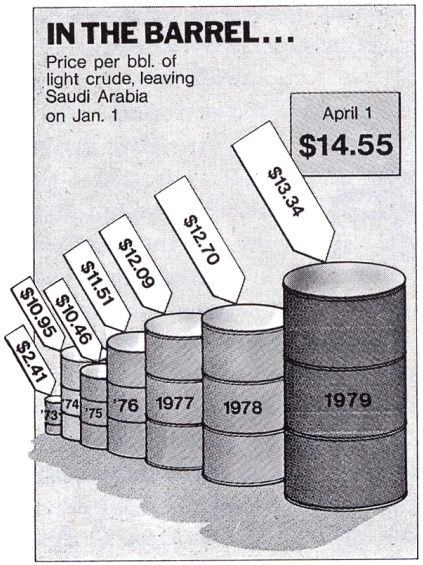
Wolf depredations



Backward in time?



Oil



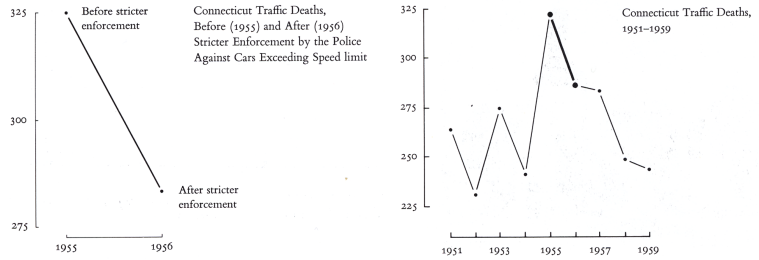
Context and labels

Keep data in context.

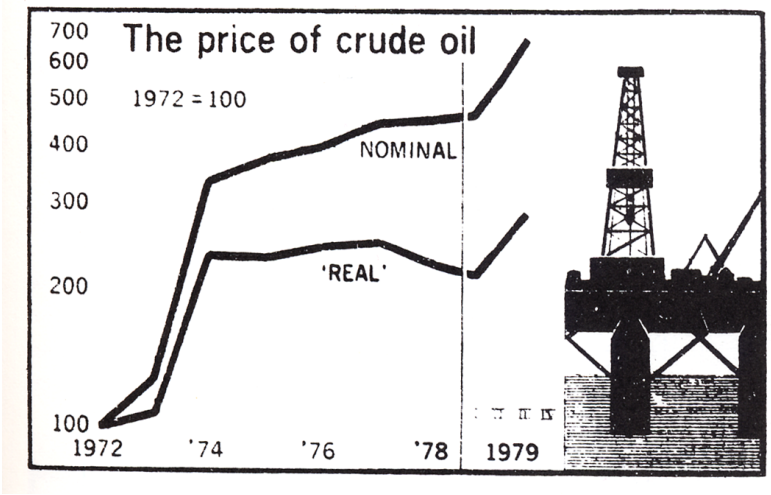
Use clear and thorough labels to avoid distortion and ambiguity.

Oil

A few more data points add immensely to the account:



Oil

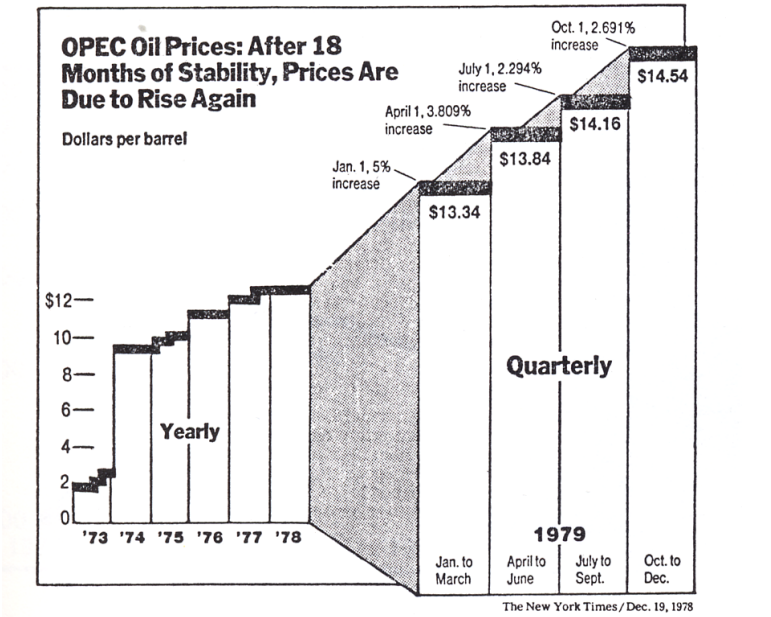


Appropriate data

Use consistent graphic design.

Deflate monetary time series.

Opec



How to Display Data Badly (Wainer)

1. Show as few data as possible.
2. Hide what data you do show.
3. Ignore the visual metaphor.
4. Only order matters.
5. Graph data out of context.
6. Change scales in mid-axis.
7. Emphasize the trivial, not the important.
8. Jiggle the baseline.
9. Austria first.
10. Label illegibly, incompletely, inaccurately, and ambiguously.

Summary

Many, many ways to do things badly.

- Show the data.
- Do not distort.
- Cause no pain.