

Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

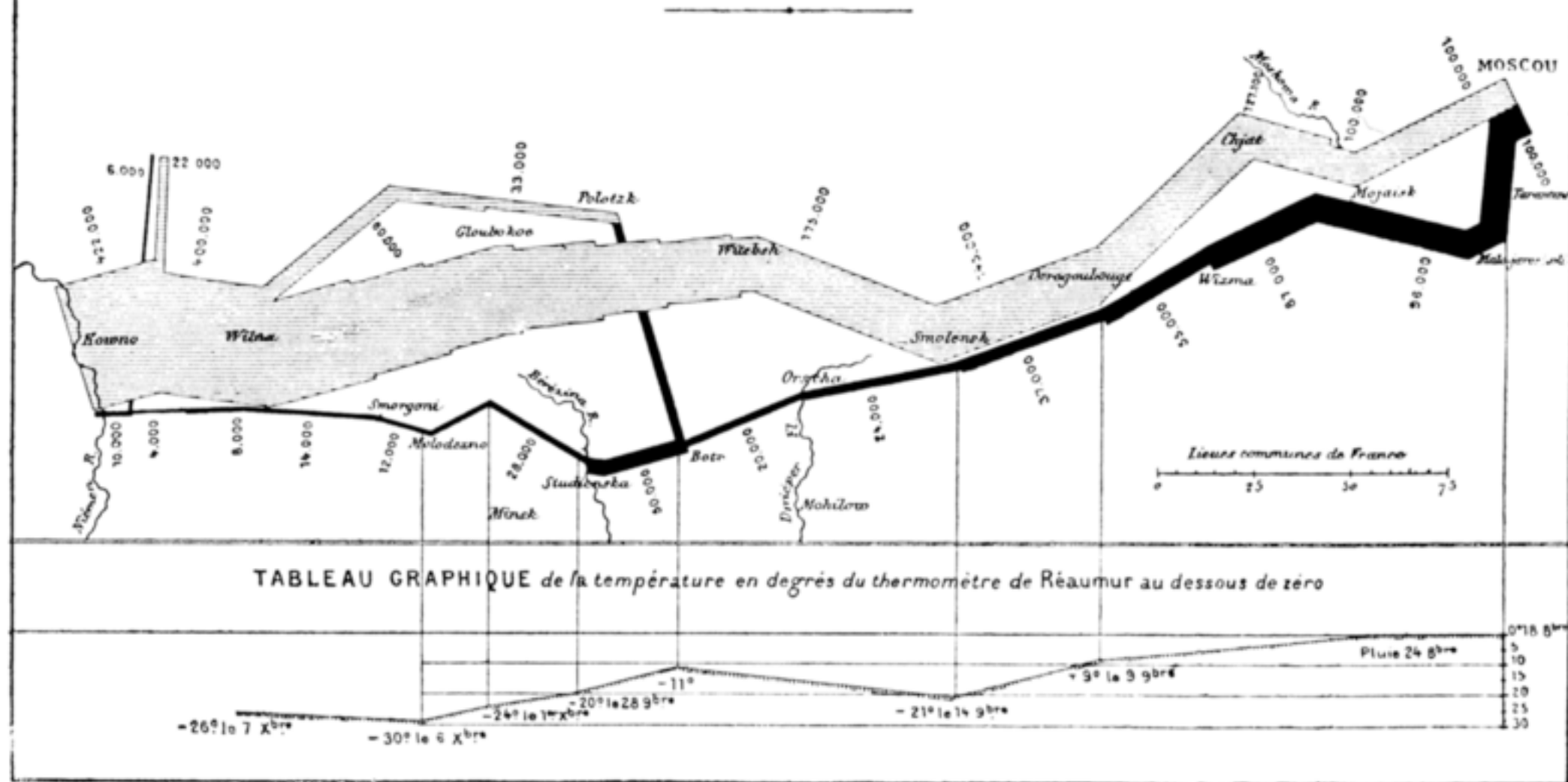
- Homework
- Visualization: Tufte
- Coming up...Preparing data, more viz, asking questions
- HW for Monday

the art of graphing: presenting data clearly

The best statistical graphic ever?

CARTE FIGURATIVE des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

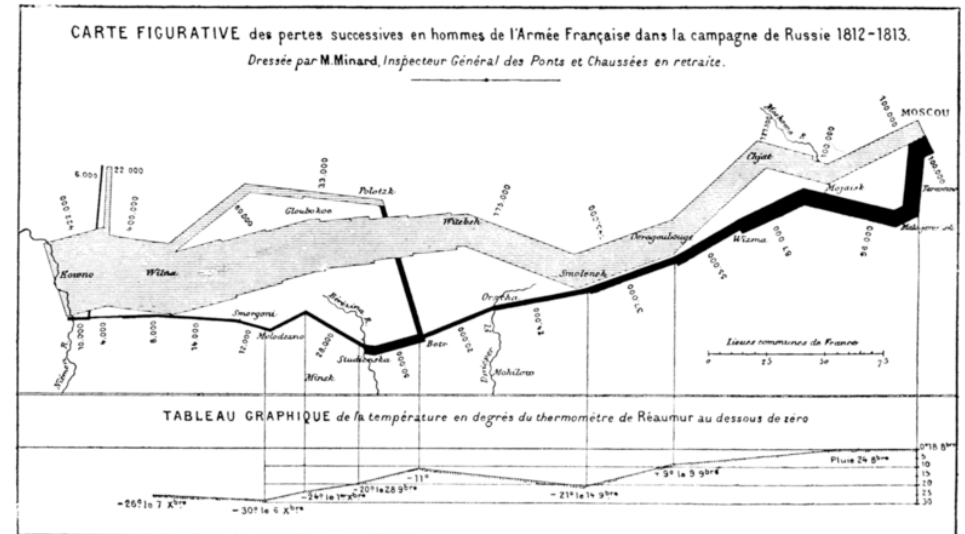
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.



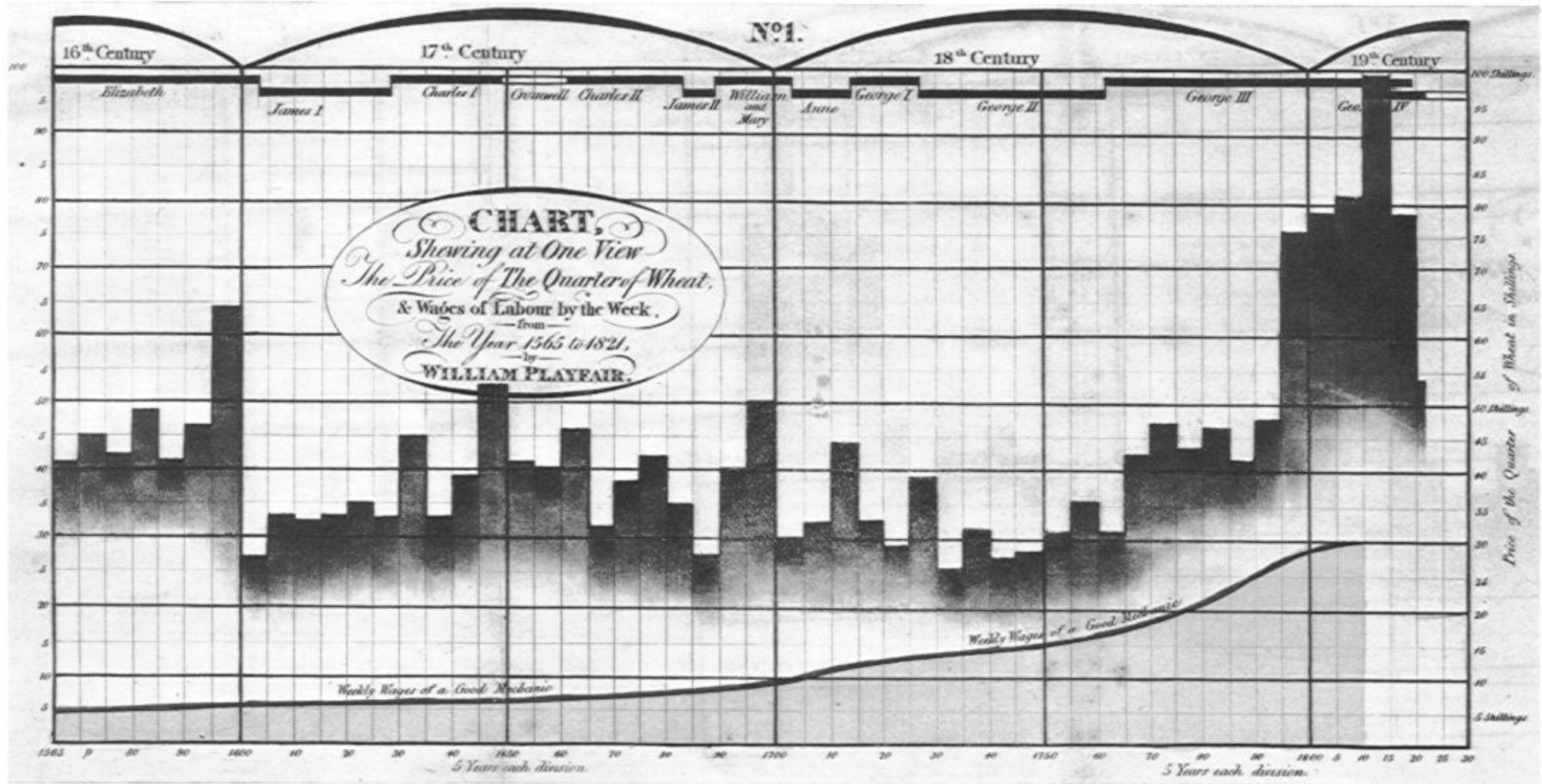
the art of graphing: presenting data clearly

The best statistical graphic ever?

- 6 dimensions of information:
 - Time, size of army, temperature, space (2D), direction of march
- All presented clearly individually, and in a way that it is easy to grasp how they relate.

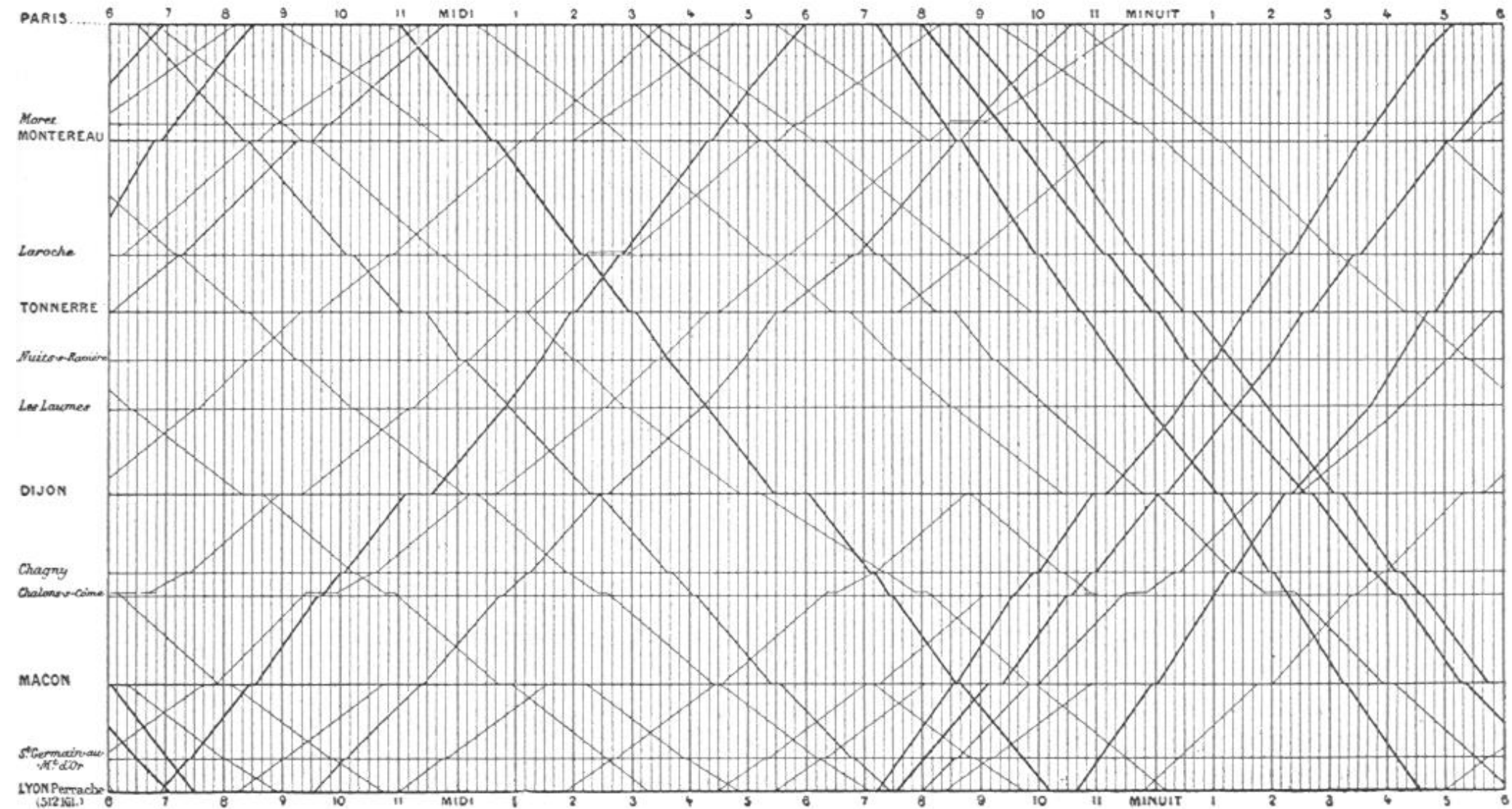


the art of graphing: presenting data clearly



The price of wheat compared to labour wages, William Playfair (1759-1823)

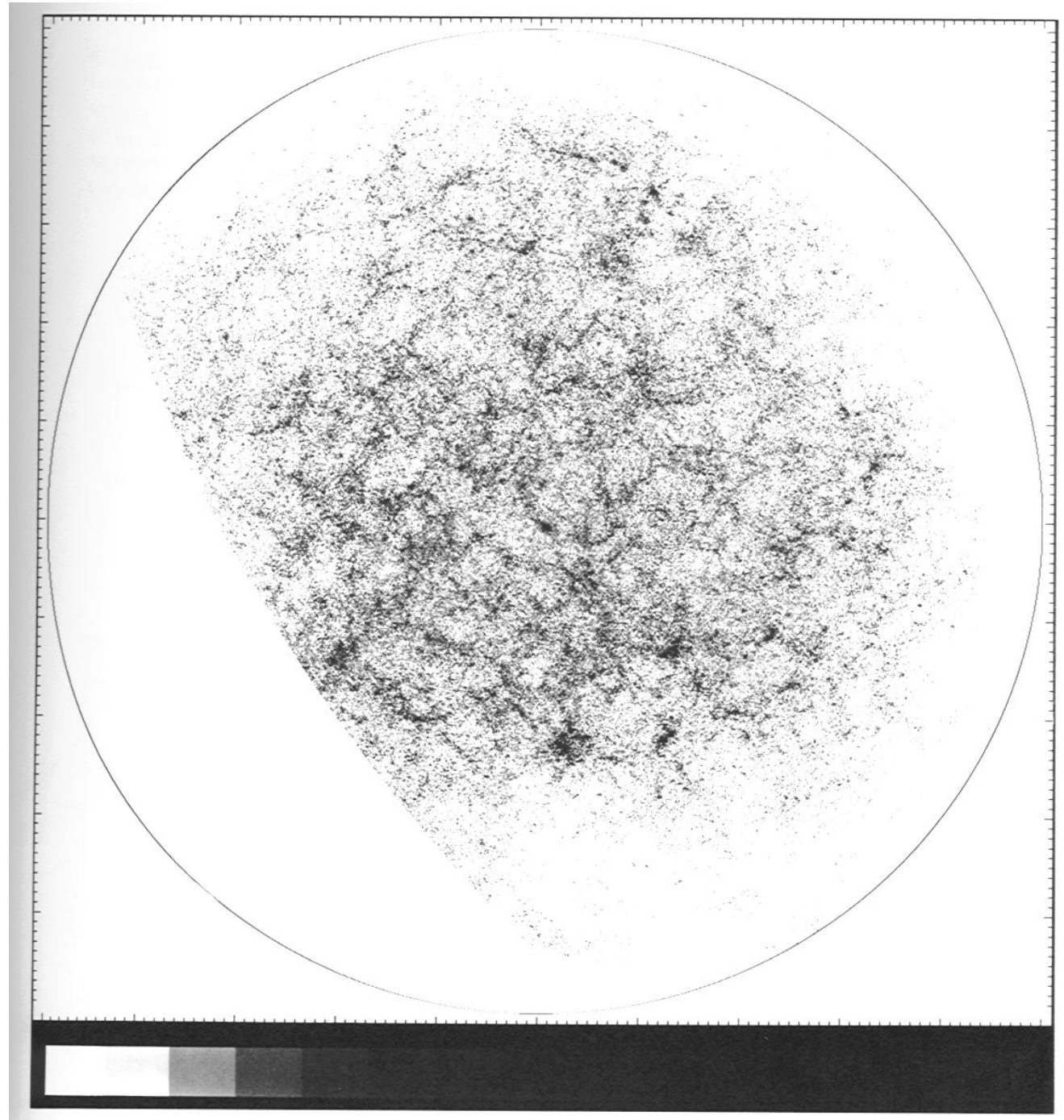
the art of graphing: presenting data clearly



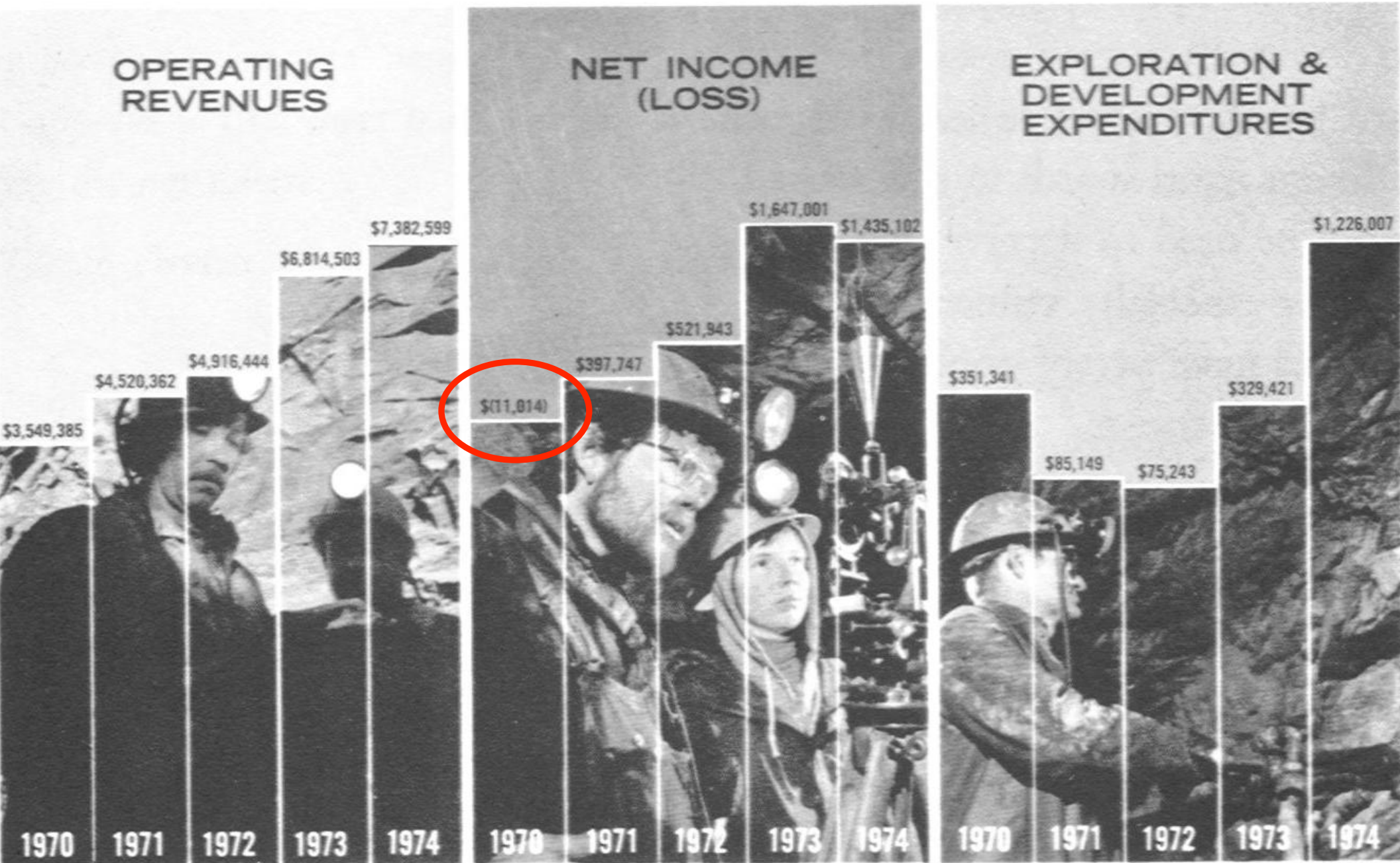
French train schedule, as drawn by E.J. Marey (1830-1904)

the art of graphing: presenting data clearly

**Map of the northern
galactic hemisphere
(1.3 million galaxies
shown)**



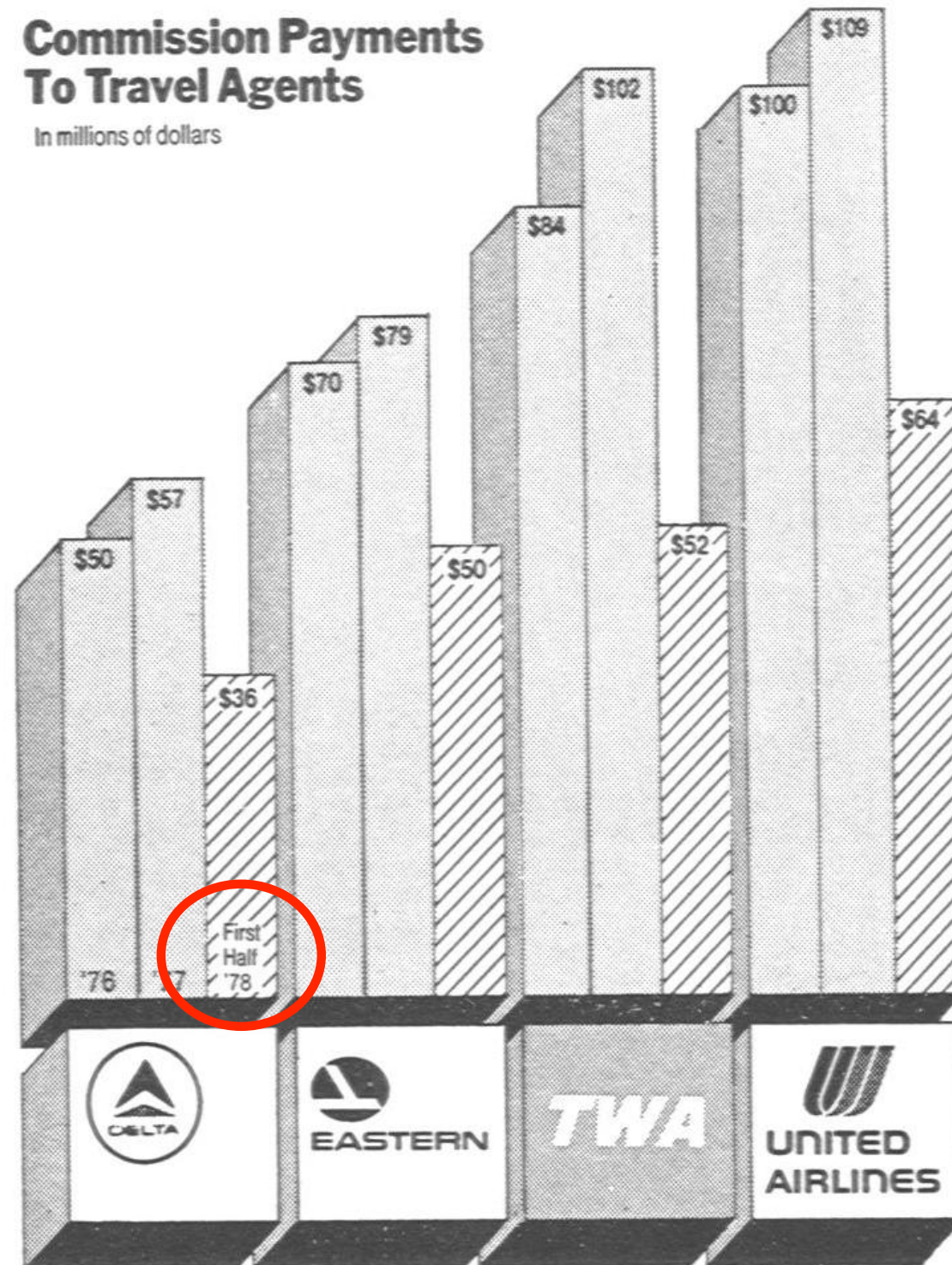
bad graphics!



bad graphics!

Commission Payments To Travel Agents

In millions of dollars



bad graphics!

Comparative Annual Cost per Capita for care of Insane in
Pittsburgh City Homes and Pennsylvania State Hospitals.

\$147



South Mountain

\$172



Pittsburgh

\$198



Harrisburg

\$213



Norristown

\$214

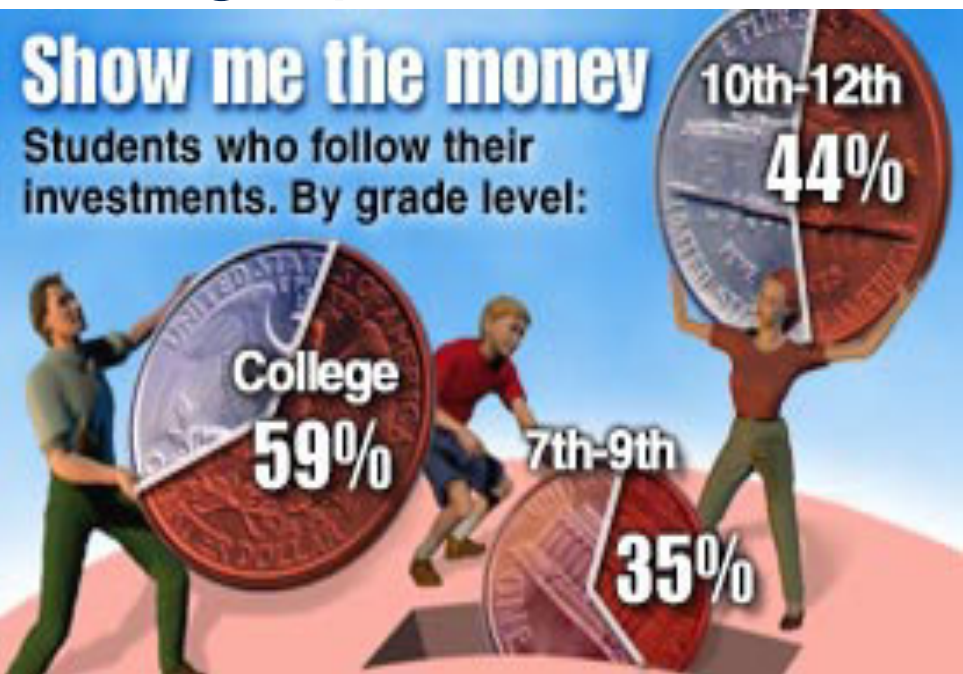


Warren

bad graphics!



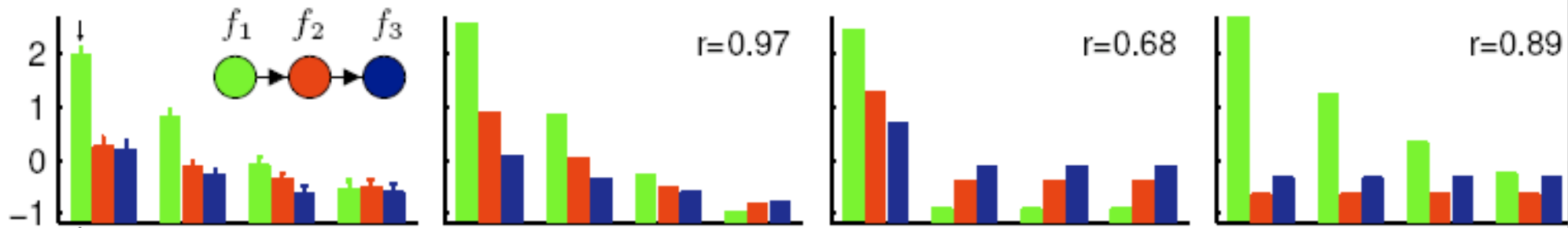
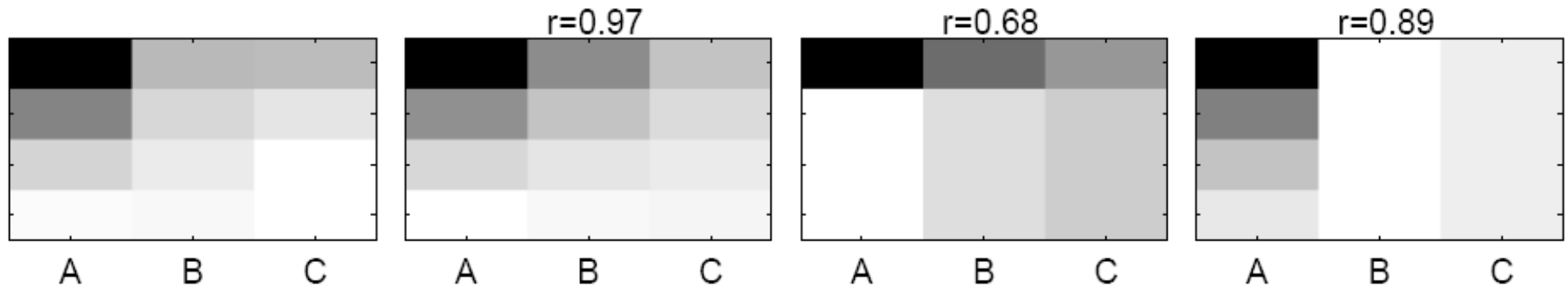
bad graphics!



bad graphics!



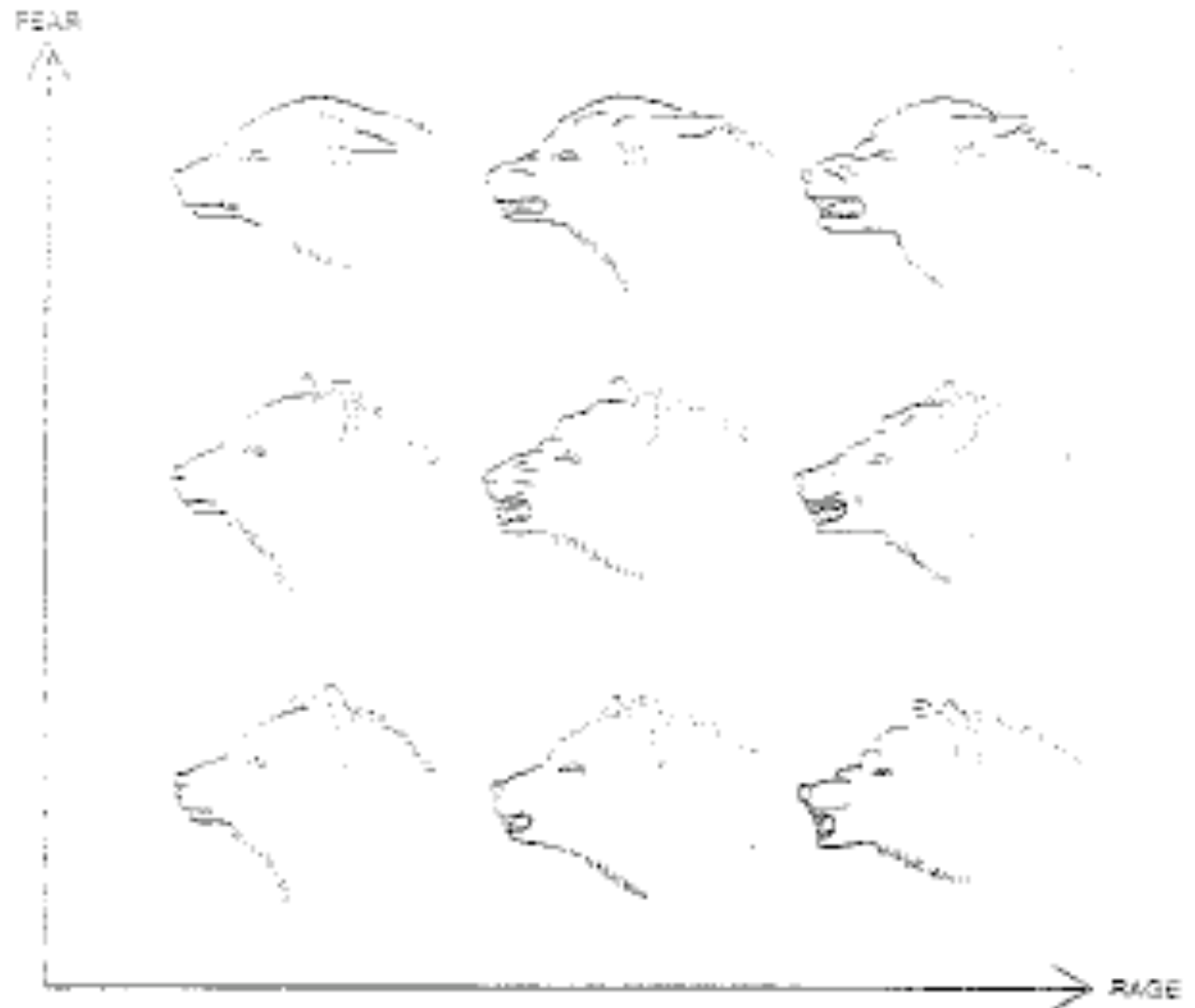
the art of graphing: presenting data clearly



the art of graphing: presenting data clearly

A scatterplot with data points that themselves contain data!

- The pictures here contain more information than could be described otherwise
- Pictures are arranged in an order that shows how they relate



the art of graphing: presenting data clearly

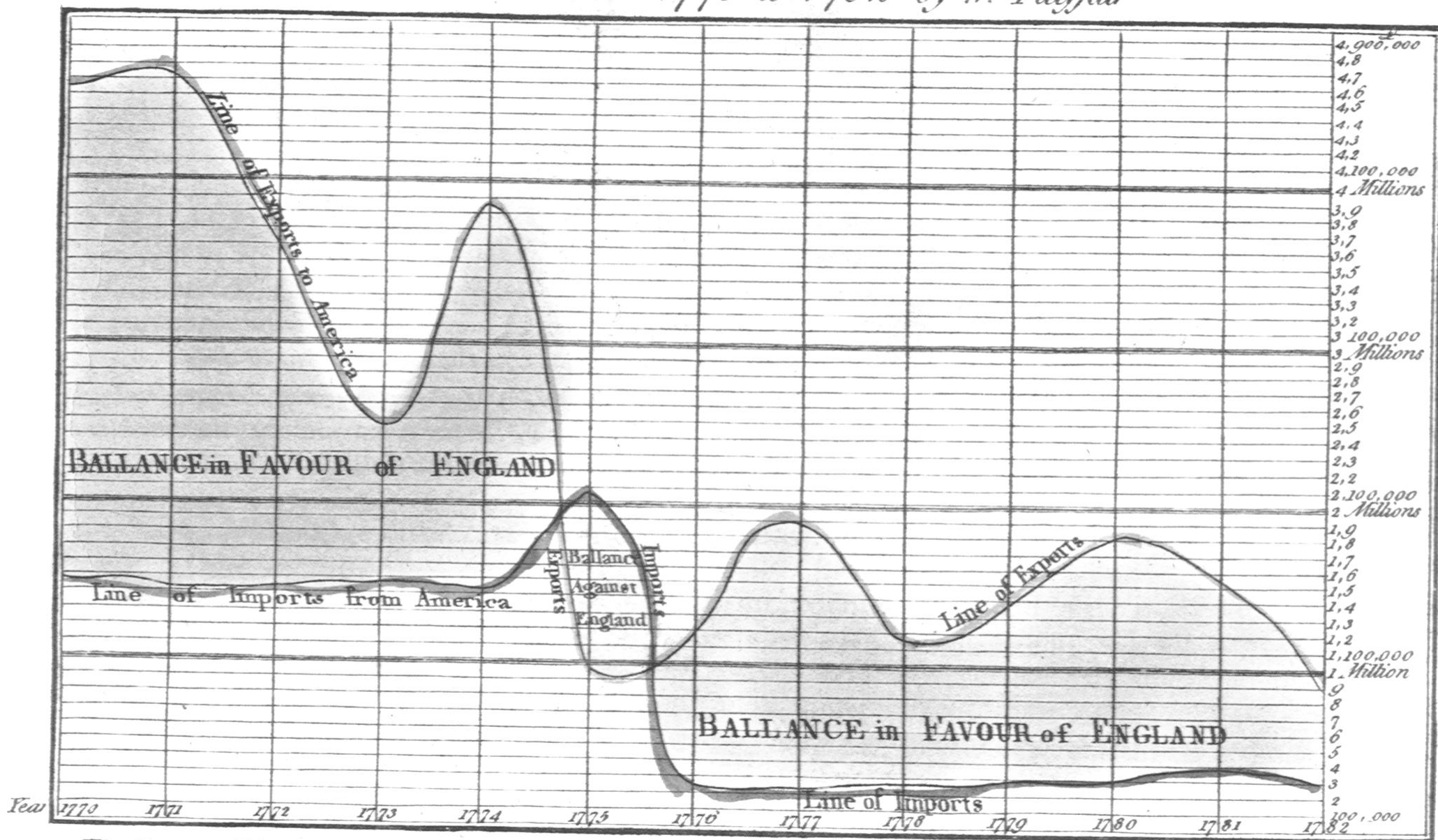
$$\begin{aligned}\text{Data-ink ratio} &= \frac{\text{data-ink}}{\text{Total ink used to print graphic}} \\ &= \text{Proportion of a graphic's ink devoted to the non-redundant display of data-information.} \\ &= 1.0 - \text{proportion of graphic that can be } \textit{erased} \text{ without the loss of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

the art of graphing: presenting data clearly

Tufte presents some principles of data graphics

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

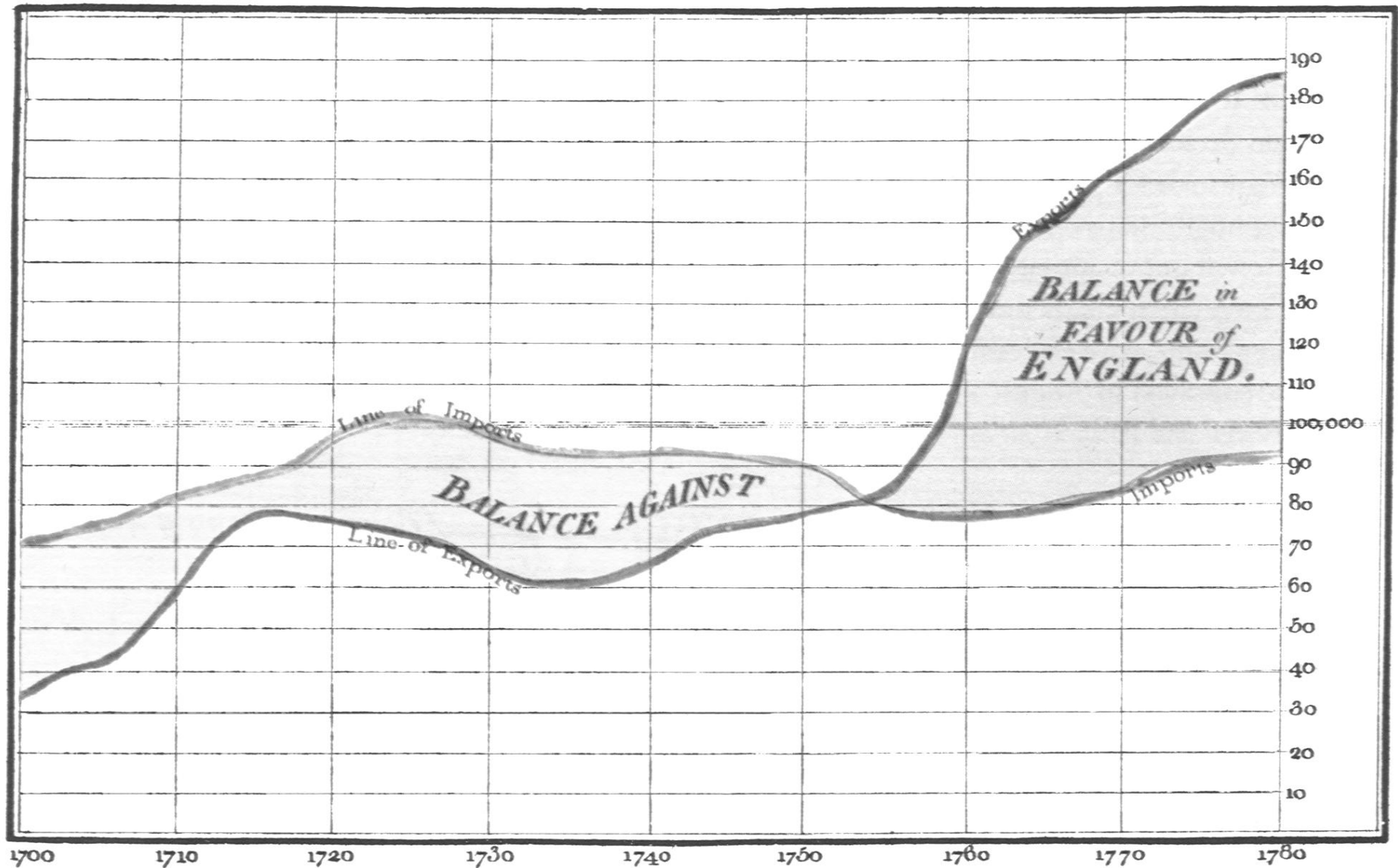
*CHART of IMPORTS and EXPORTS of ENGLAND to and from all NORTH AMERICA
From the Year 1770 to 1782 by W. Playfair*



The Bottom Line is divided into Years the right-hand Line into HUNDRED THOUSAND POUNDS

the art of graphing: presenting data clearly

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



The Bottom line is divided into Years, the Right hand line into £10,000 each.

Published as the Act directs, 14th May 1786. by W^m Playfair

Needle sculpt 352, Strand, London.

the art of graphing: presenting data clearly

Bad graphing practices:

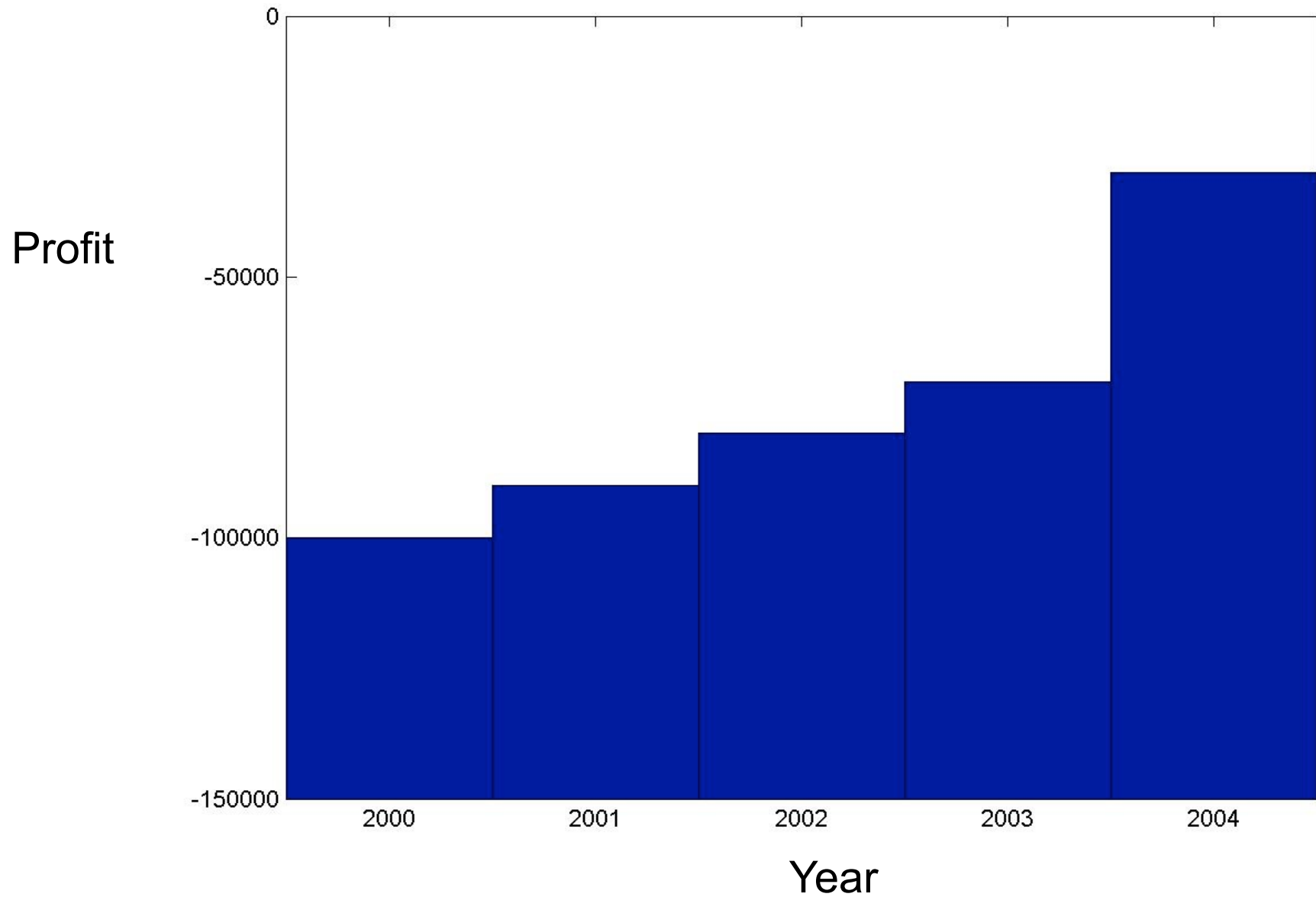
- Setting false baselines
- Comparing apples and oranges
- 2 dimensional graphics representing 1 dimensional data
- Distorting effects
- Ineffective comparisons
- Extra non-information carrying graphics
- Excessive prettiness

from: Tufte, E.R. (1983). The visual display of quantitative information.

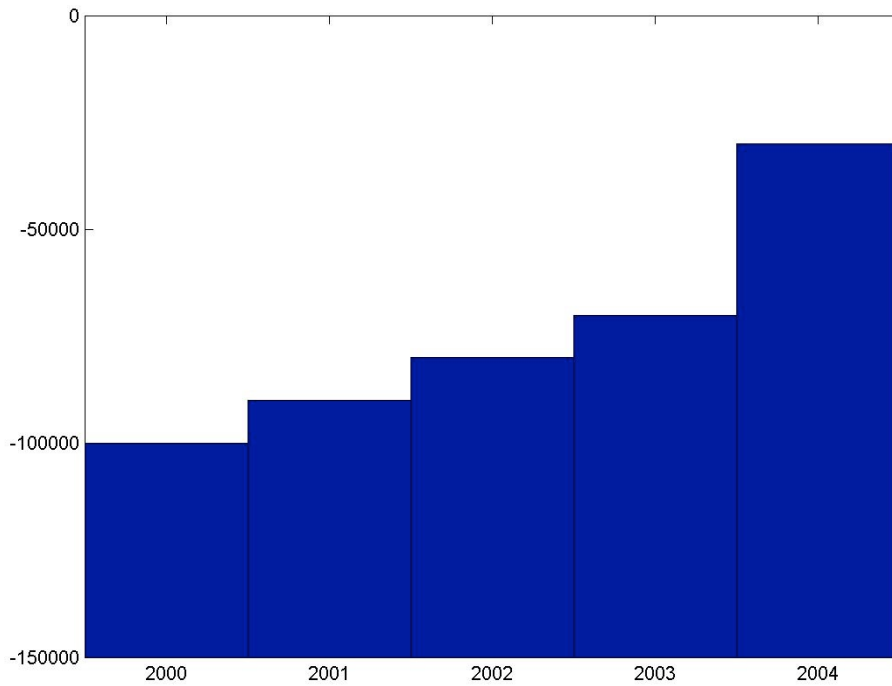
General principles:

- Maximize the amount of data presented for the amount of ink used.

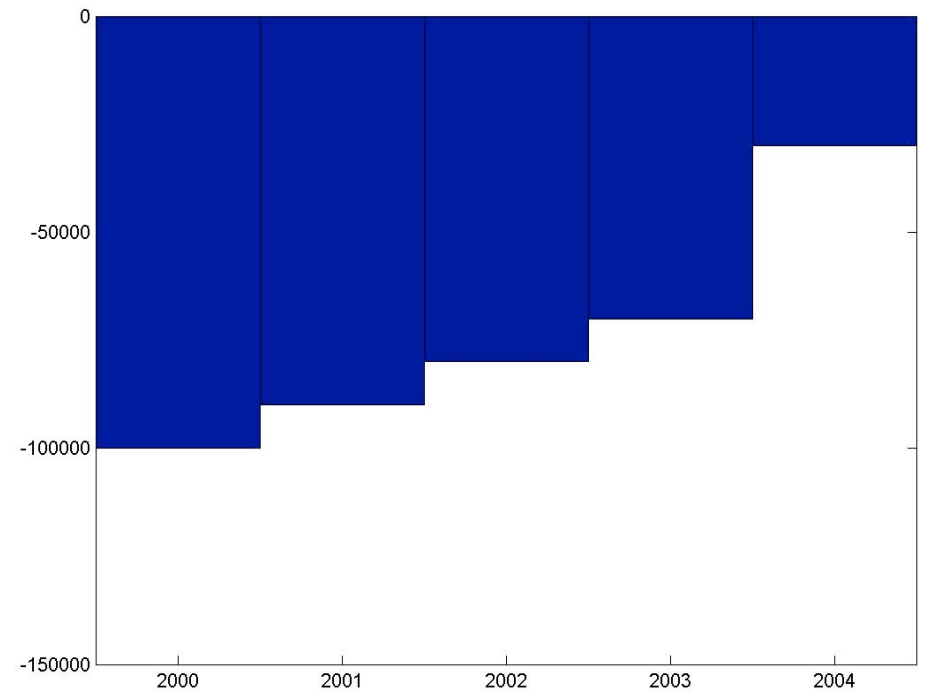
bad practice: false baselines



bad practice: false baselines



Are we making more money?



No! We are just losing less!

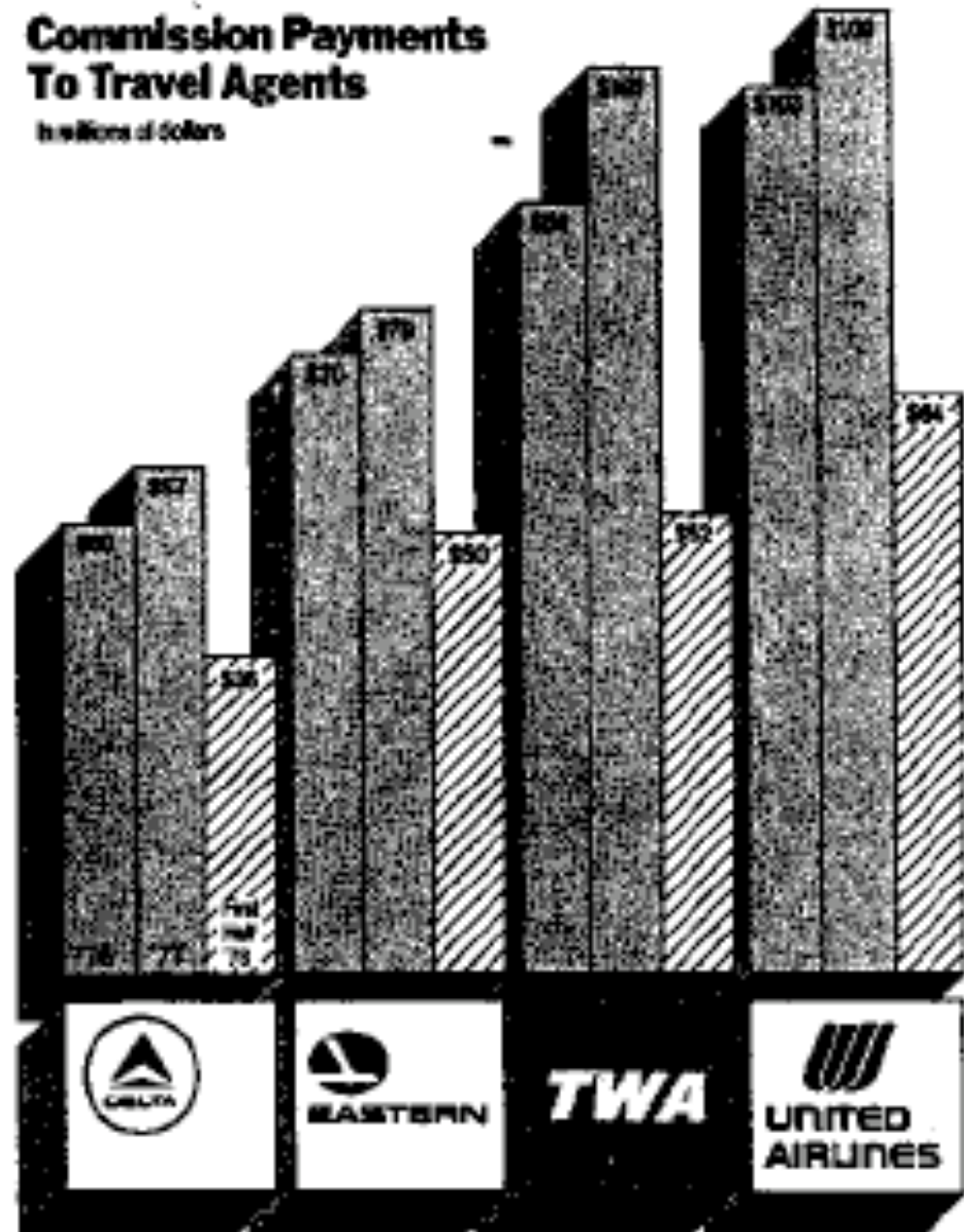
bad practice: comparing apples and oranges

Notice how revenues appear to be sharply declining!

Oh, no! Are the airlines going to fail?

No...the hashed bars are only from the first half of 1978!

Note how the hashing against the dark bars exaggerates the effect.

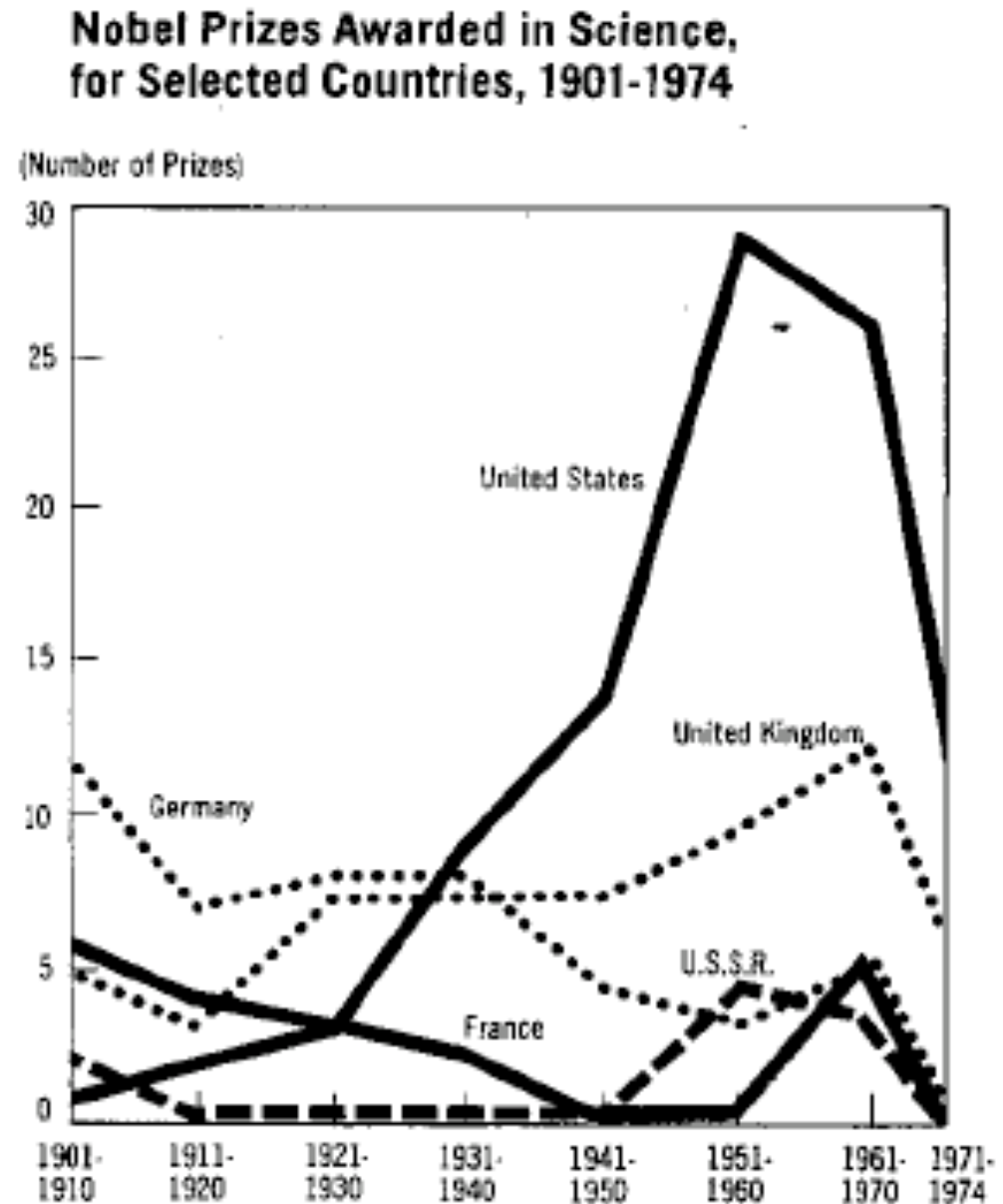


bad practice: comparing apples and oranges

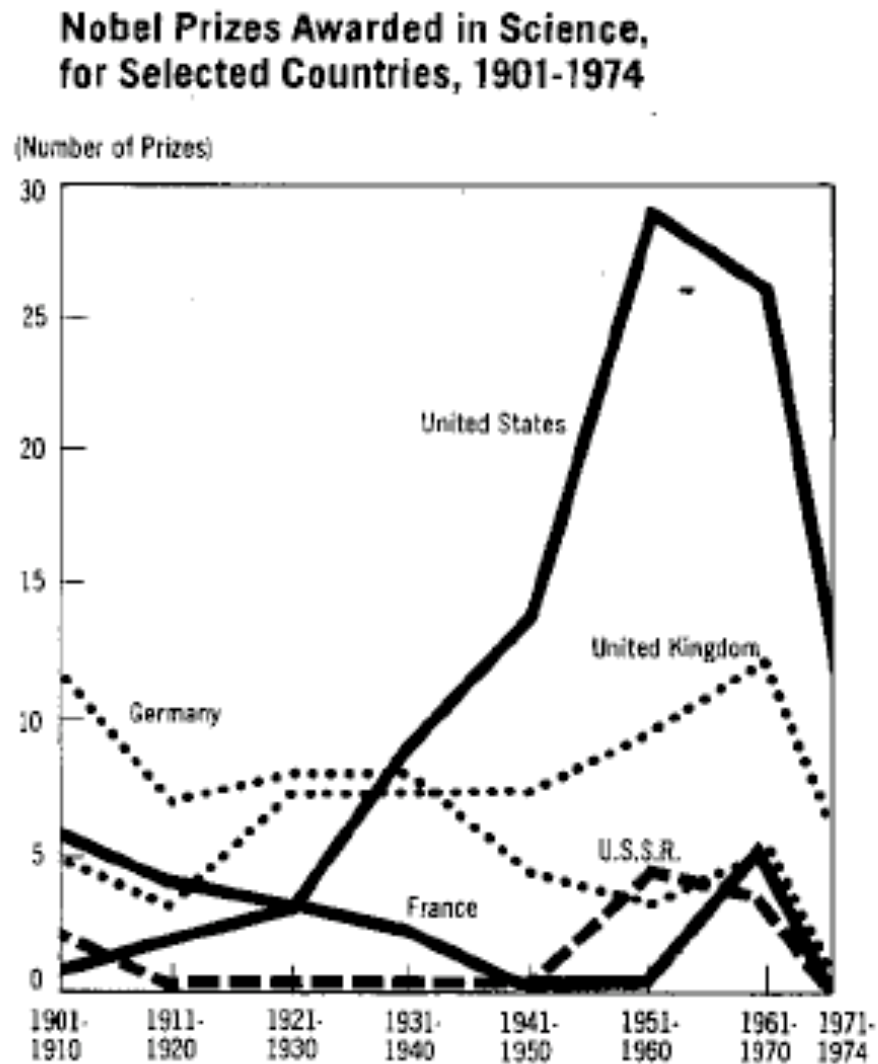
Oh no! Nobel prizes for the US appear to be sharply declining!

Now look at the X axis:
the label for the last tick
– that data is only for 3
years (not 10).

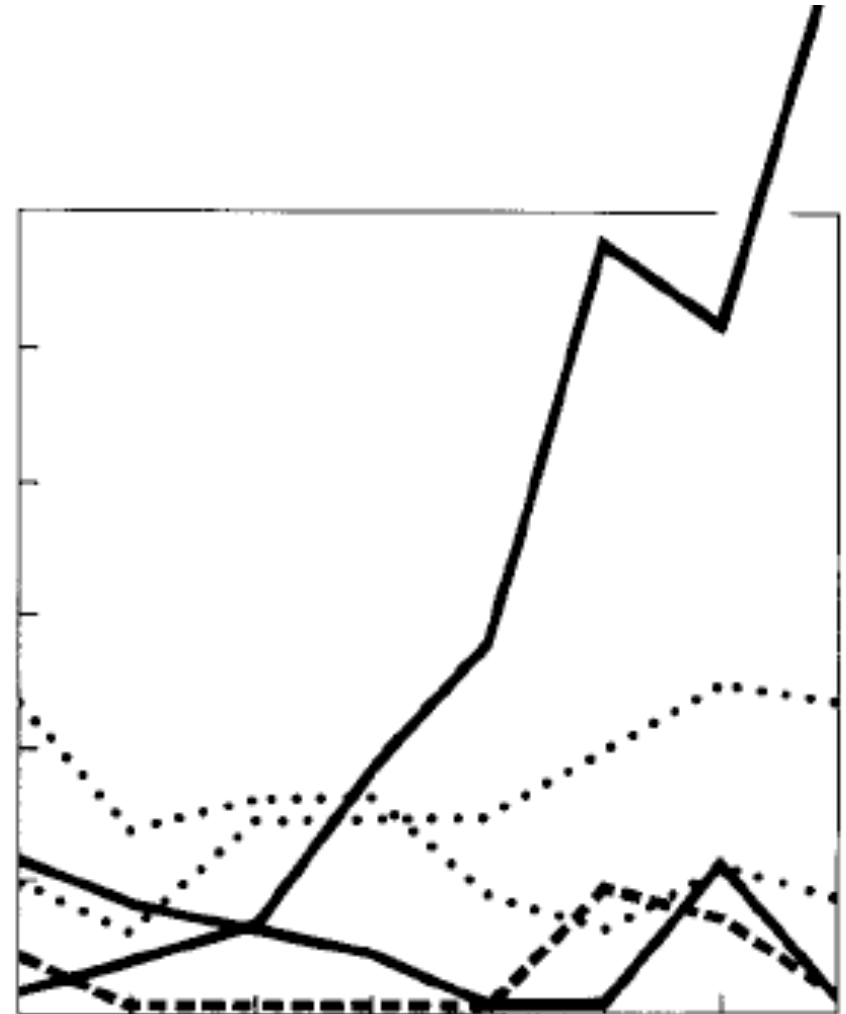
If we extrapolated to a
full 10 years, based on
this data, the graph
would look like this:



bad practice: comparing apples and oranges



Bad graph



Better graph

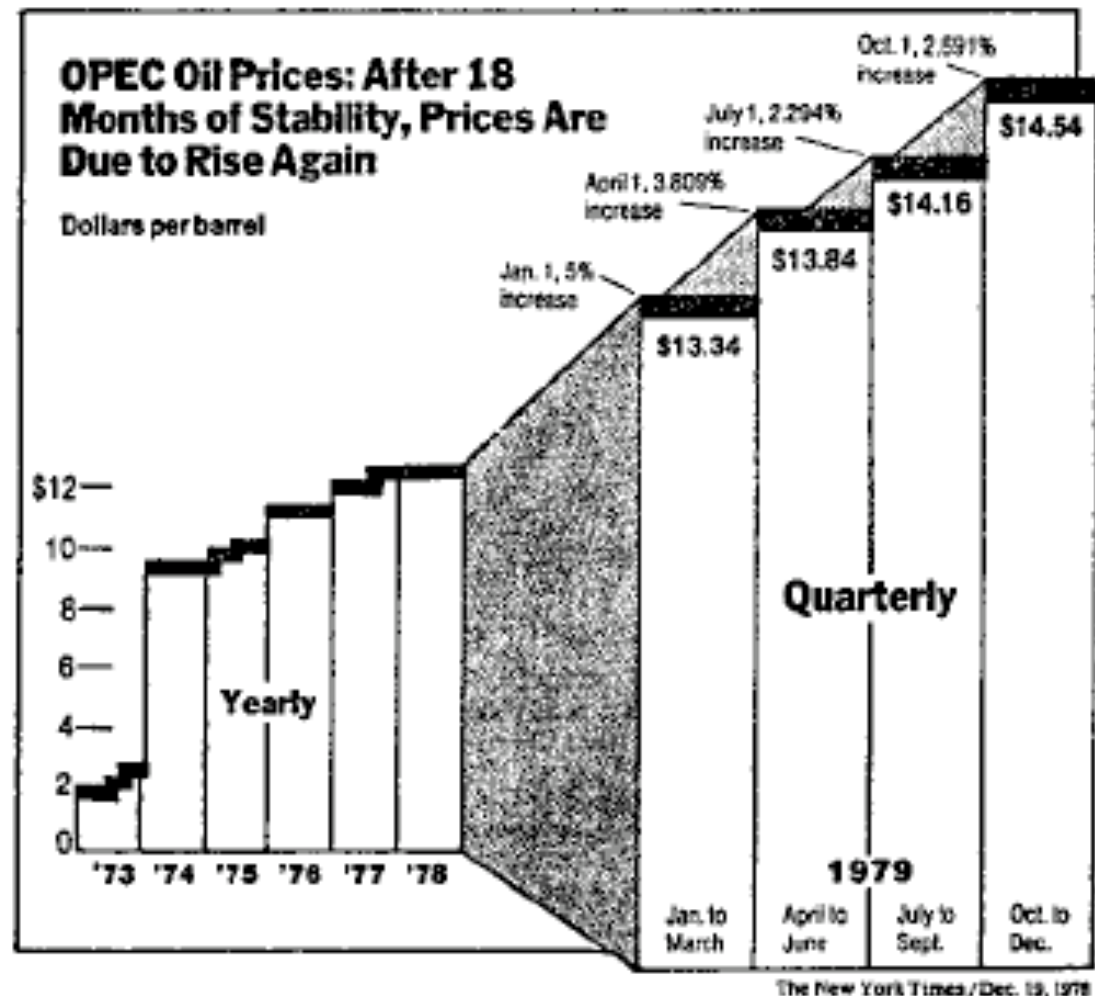
bad practice: comparing apples and oranges

Five different vertical scales show the price:

<u>During this time</u>	<u>one vertical inch equals</u>
1973-1978	\$8.00
January-March 1979	\$4.73
April-June 1979	\$4.37
July-September 1979	\$4.16
October-December 1979	\$3.92

And two different horizontal scales show the passage of time:

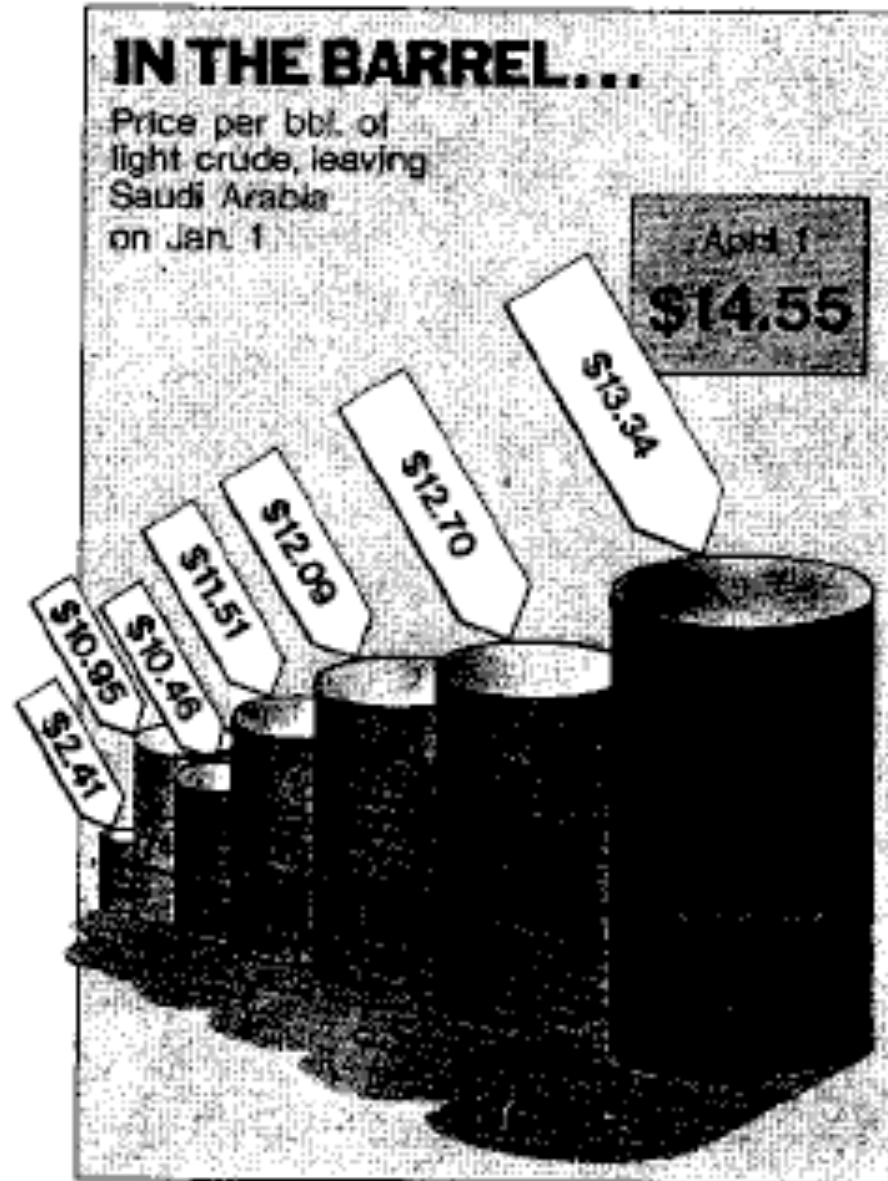
<u>During this time</u>	<u>one horizontal inch equals</u>
1973-1978	3.8 years
1979	0.57 years



bad practice: 1D data on 2D plot

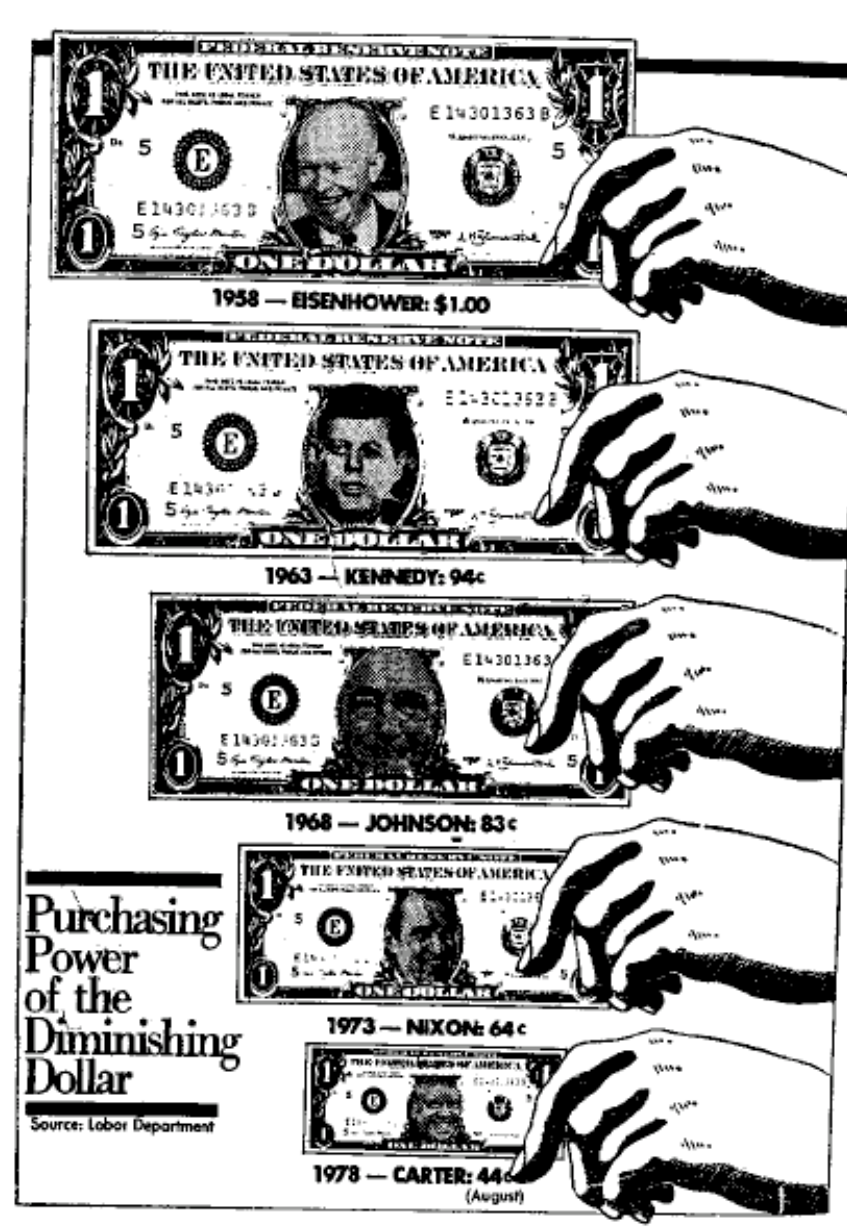
Price changed
554%.

Area of the barrels
changes roughly
1300%!



bad practice: 1D data on 2D plot

If the visual area correctly represented the change, the 1978 dollar should be two times larger!

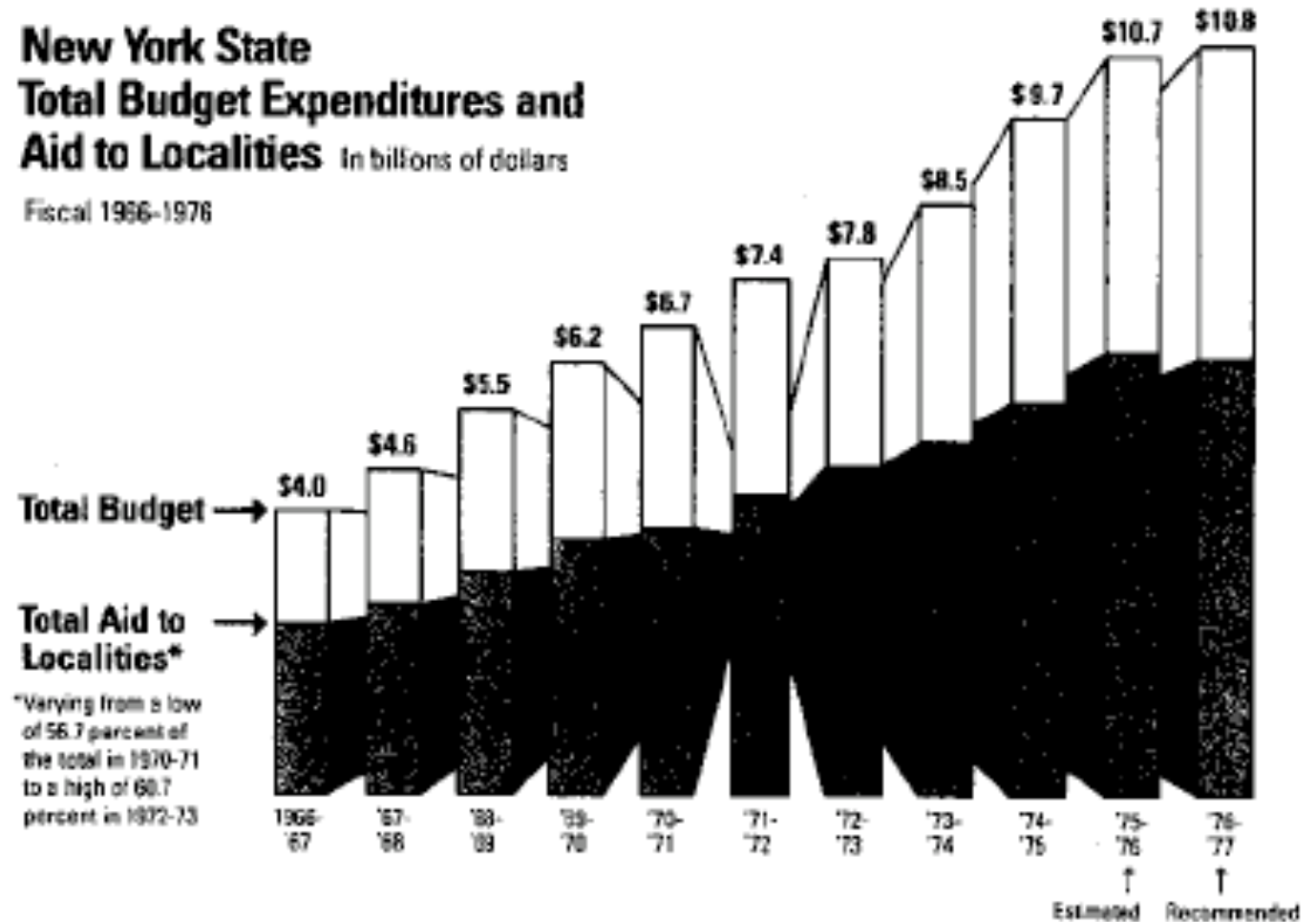


bad practice: distorting effects

Note:

- The weird 3D thing going on
- Apparent variation in depth
- Use of arrows
- Labels are squeezed next to smallest bar

Compare this to when junk is deleted:

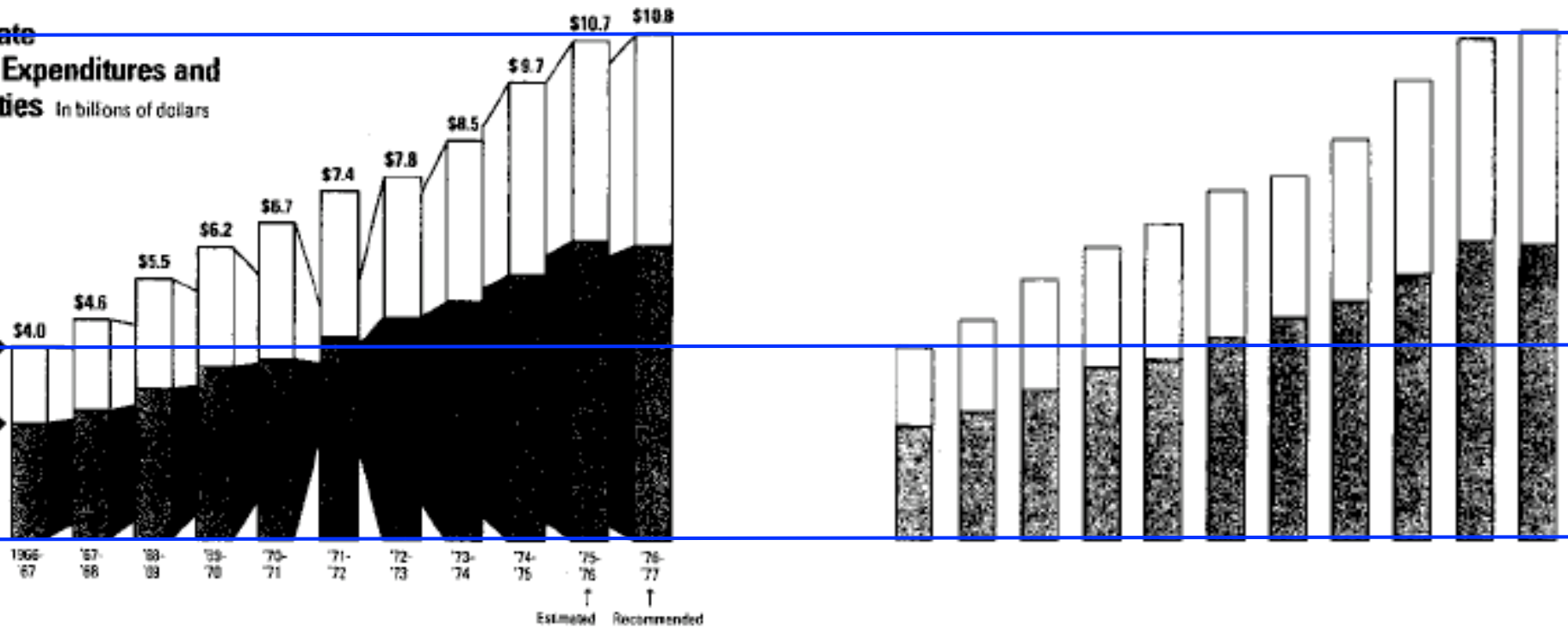


bad practice: distorting effects

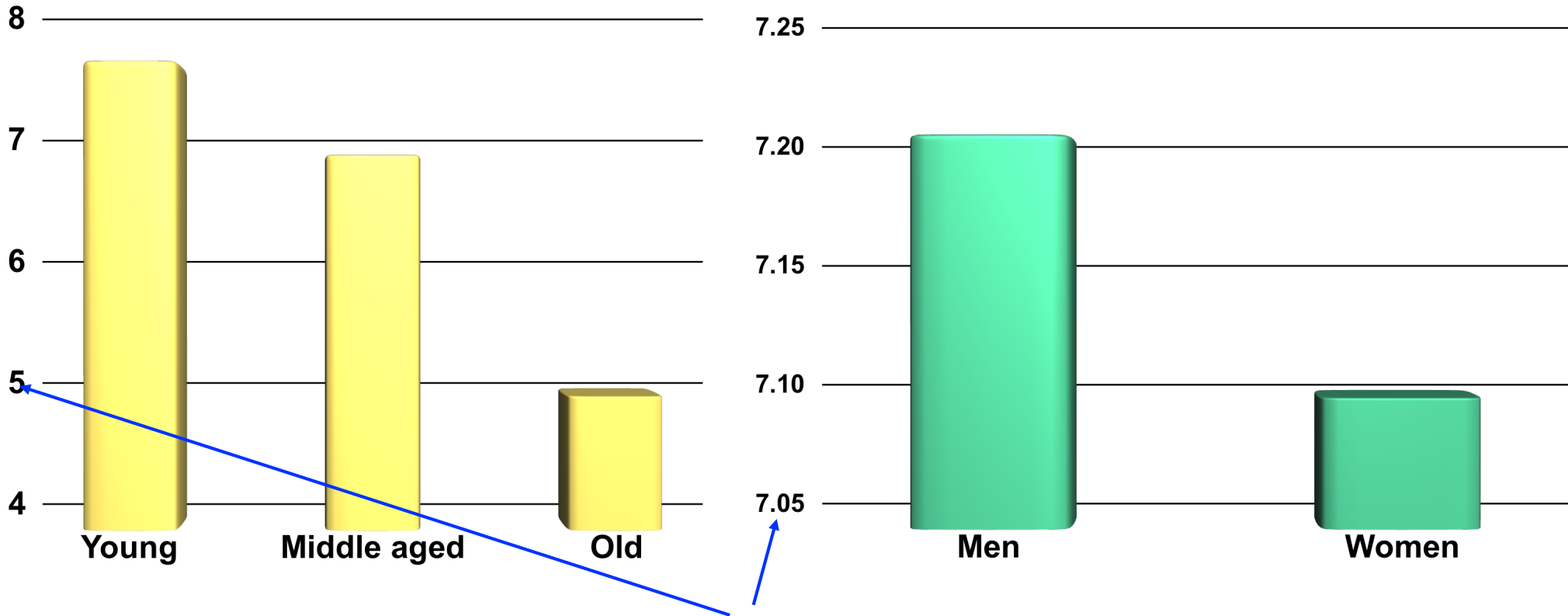
**New York State
Total Budget Expenditures and
Aid to Localities** In billions of dollars
Fiscal 1966-1976

Total Budget →
Total Aid to Localities* →

*Varying from a low of 56.7 percent of the total in 1970-71 to a high of 68.7 percent in 1972-73

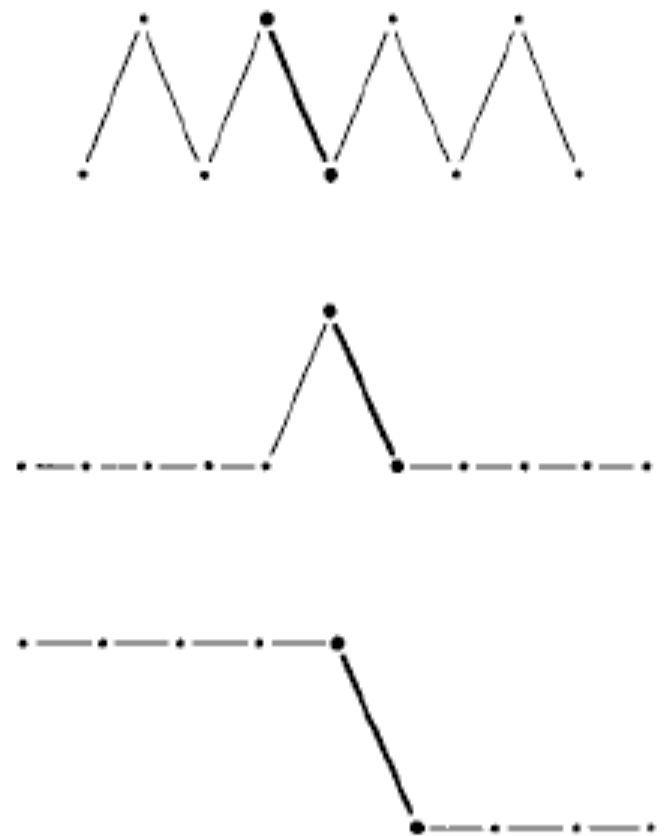
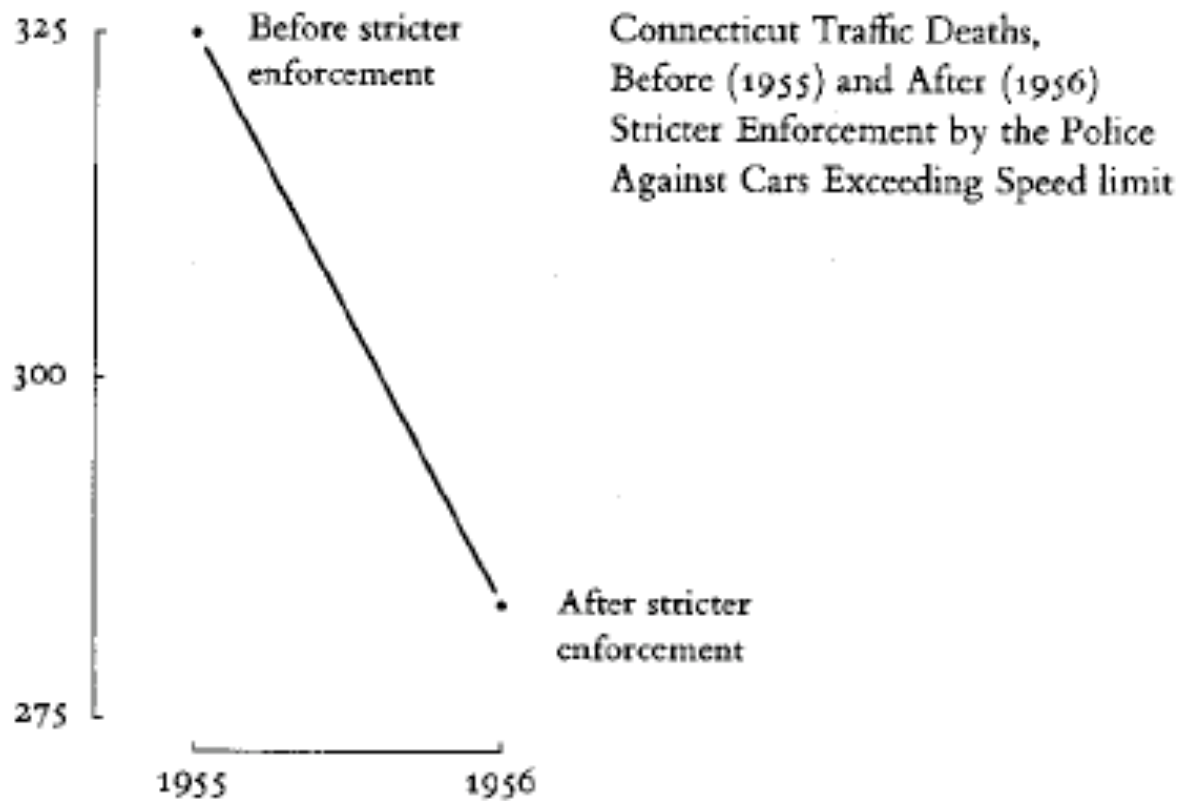


bad practice: ineffective (or no) comparison

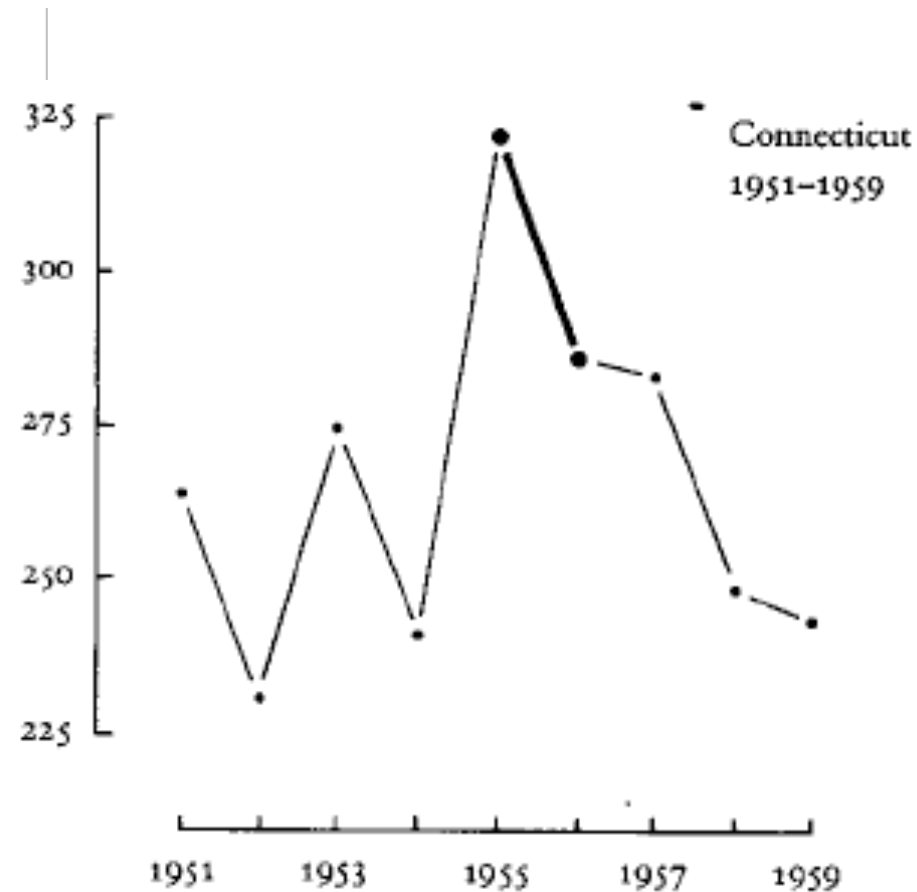
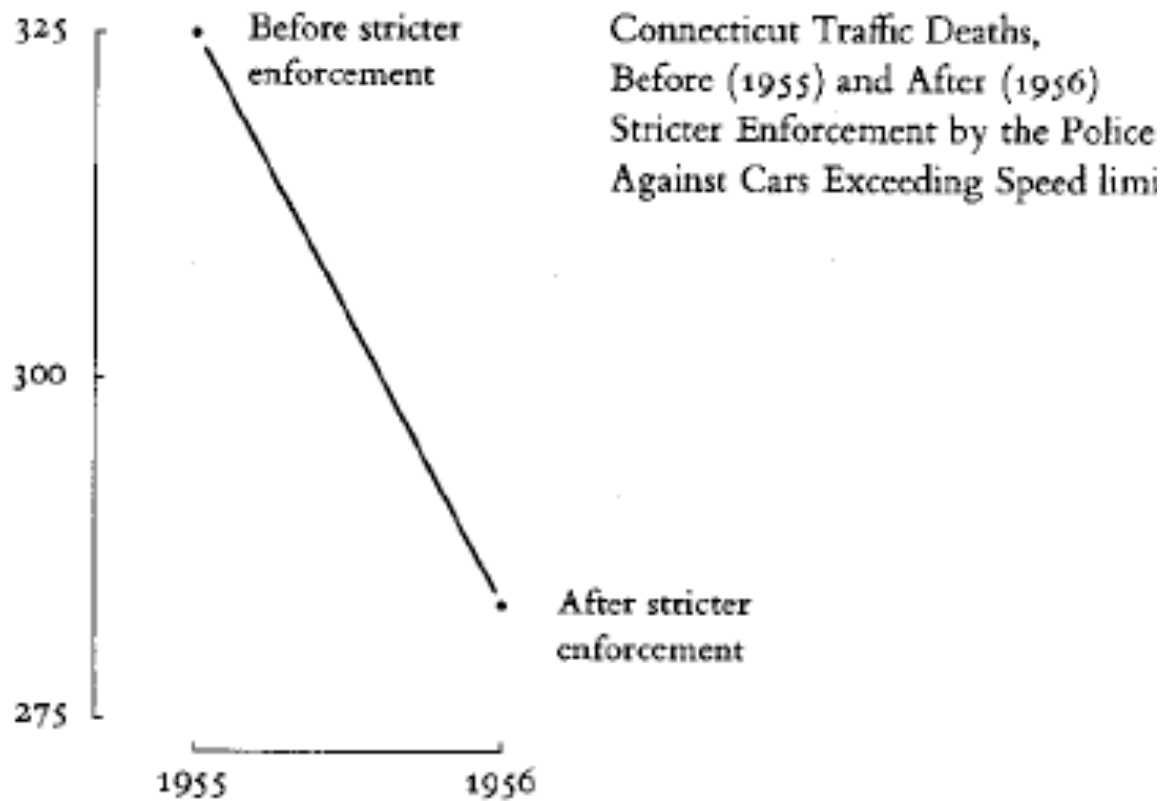


Misleading comparison! Look at the y axis labels!

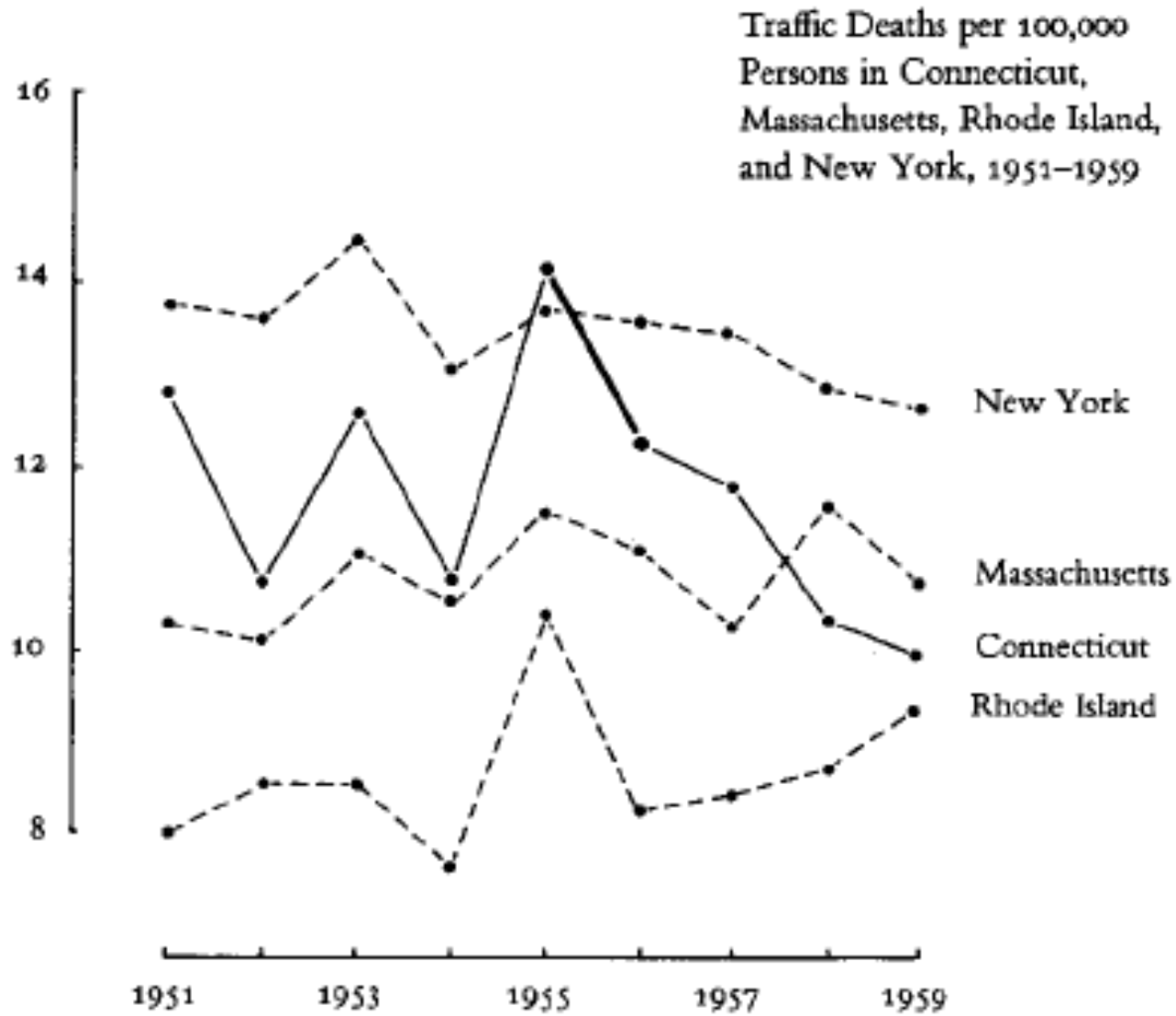
bad practice: ineffective (or no) comparison



bad practice: ineffective (or no) comparison



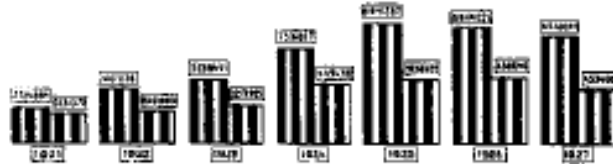
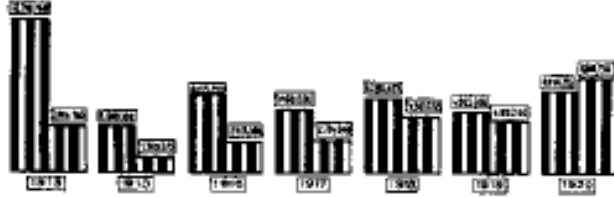
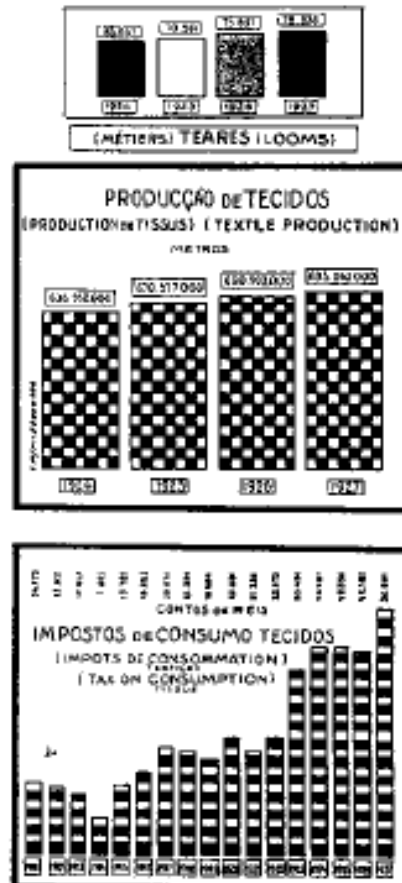
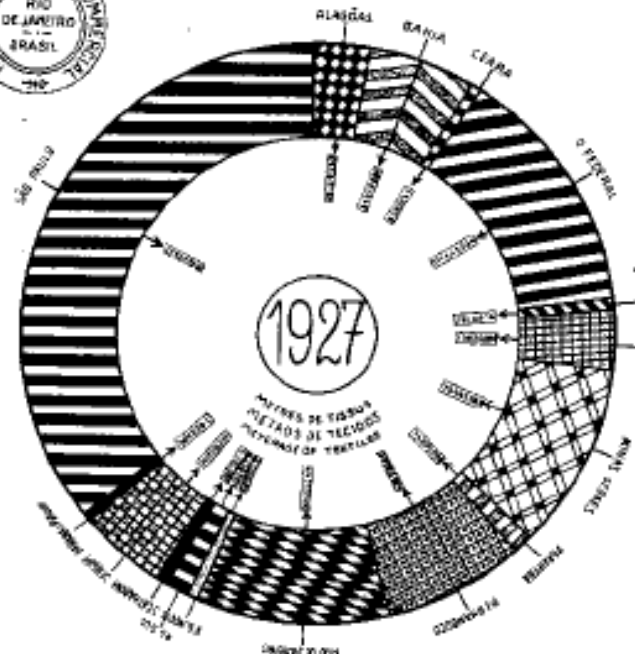
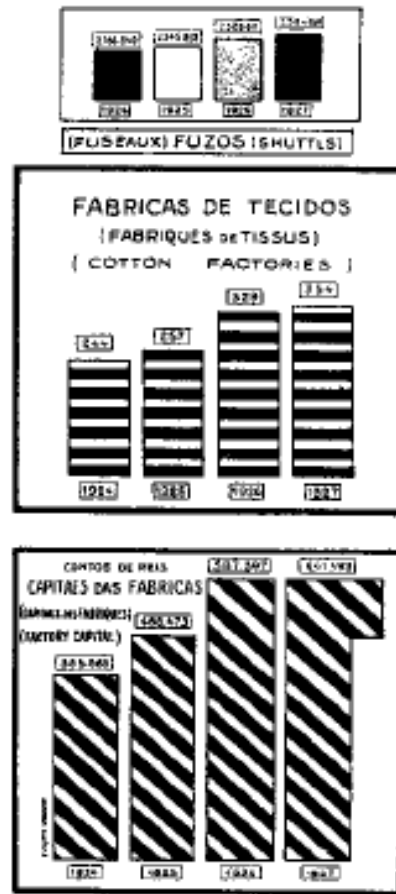
bad practice: ineffective (or no) comparison



bad practice: chart junk!

Cross-hatching
is bad.

Stripes are
too...



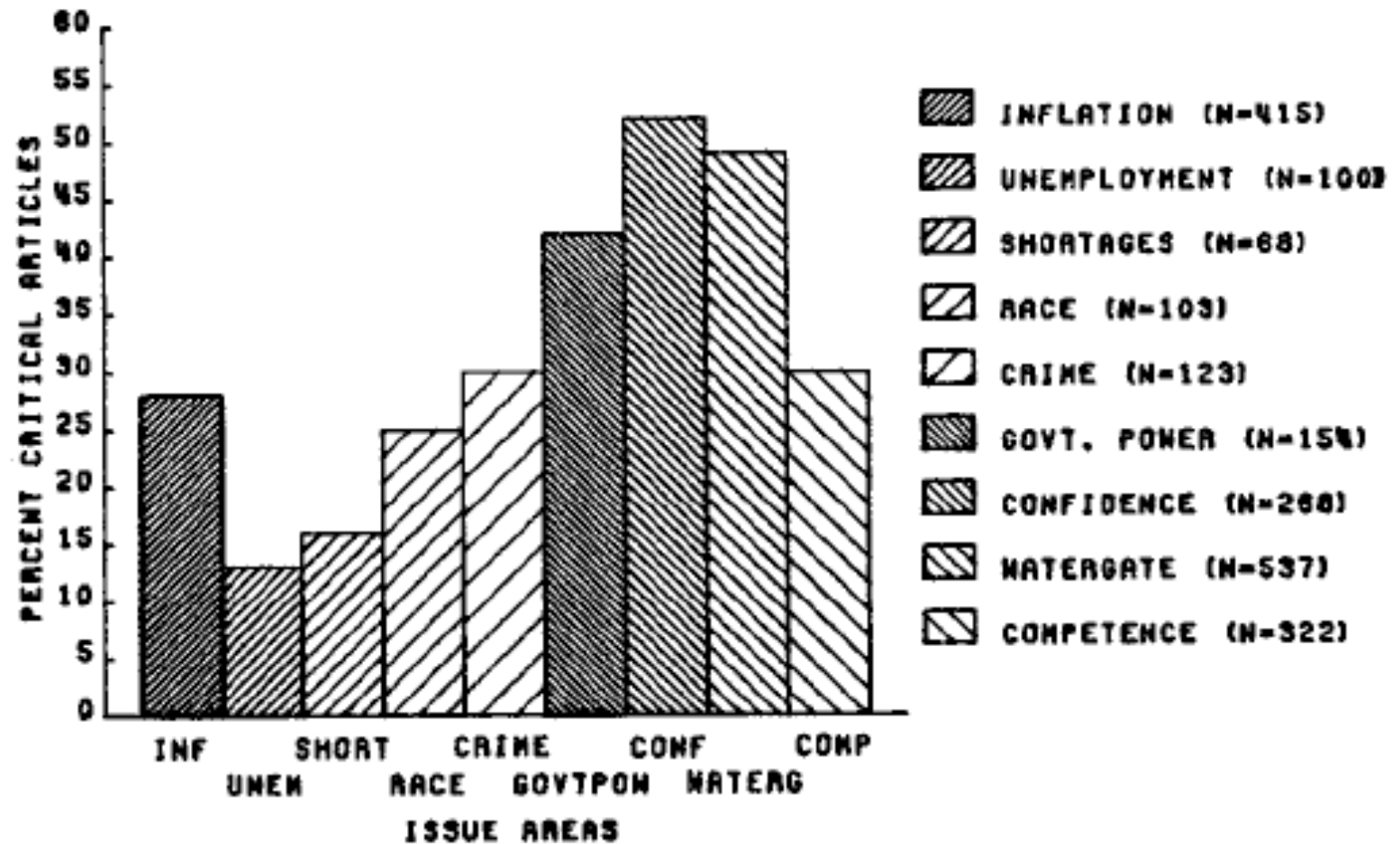
bad practice: chart junk!



Stripes create
strange visual
effects!

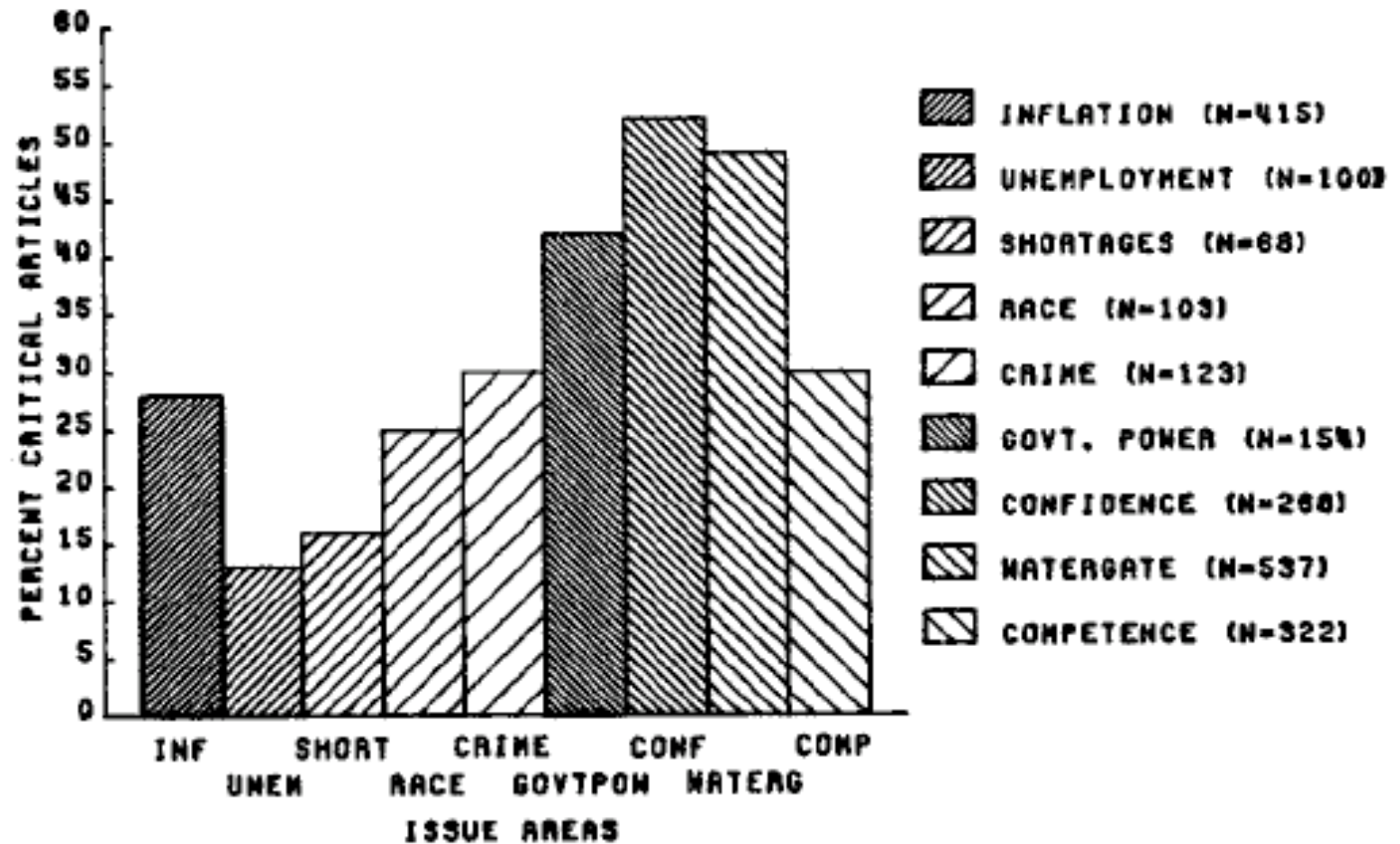
bad practice: chart junk!

Stripes create
strange visual
effects!



bad practice: chart junk!

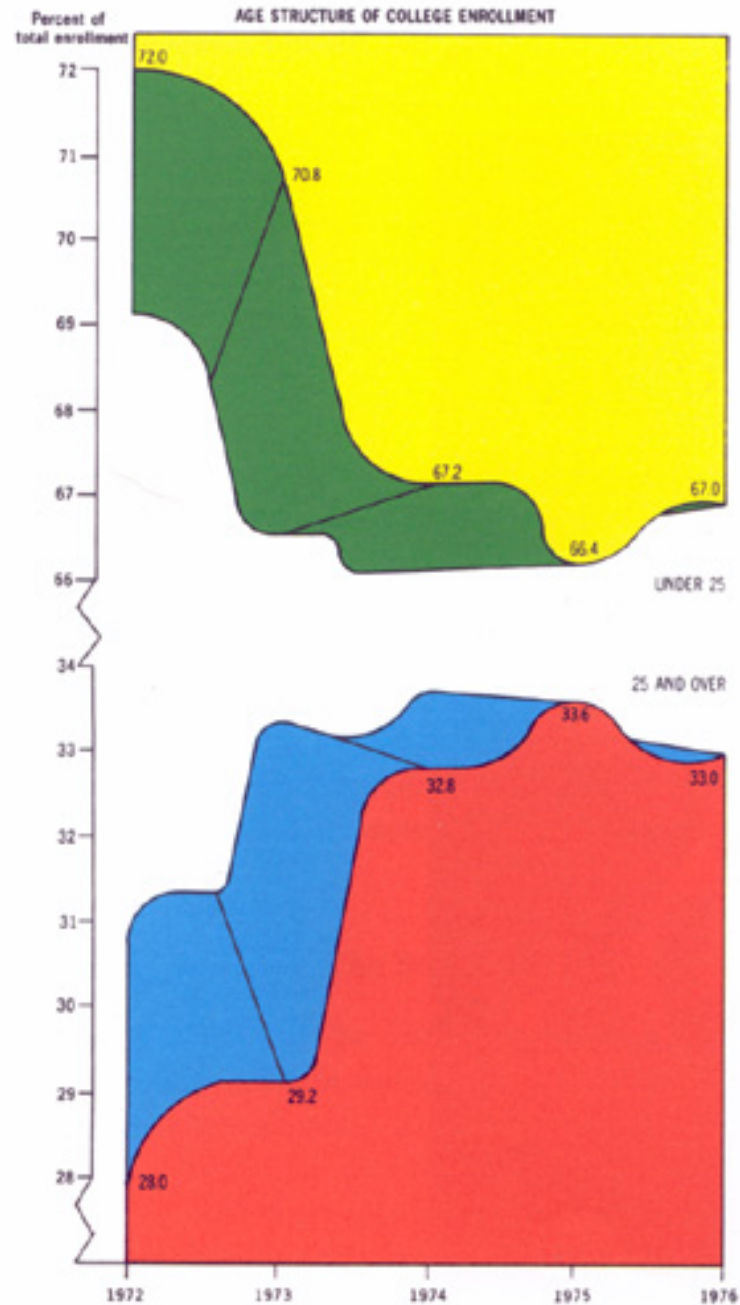
Stripes create
strange visual
effects!



bad practice: chart junk!

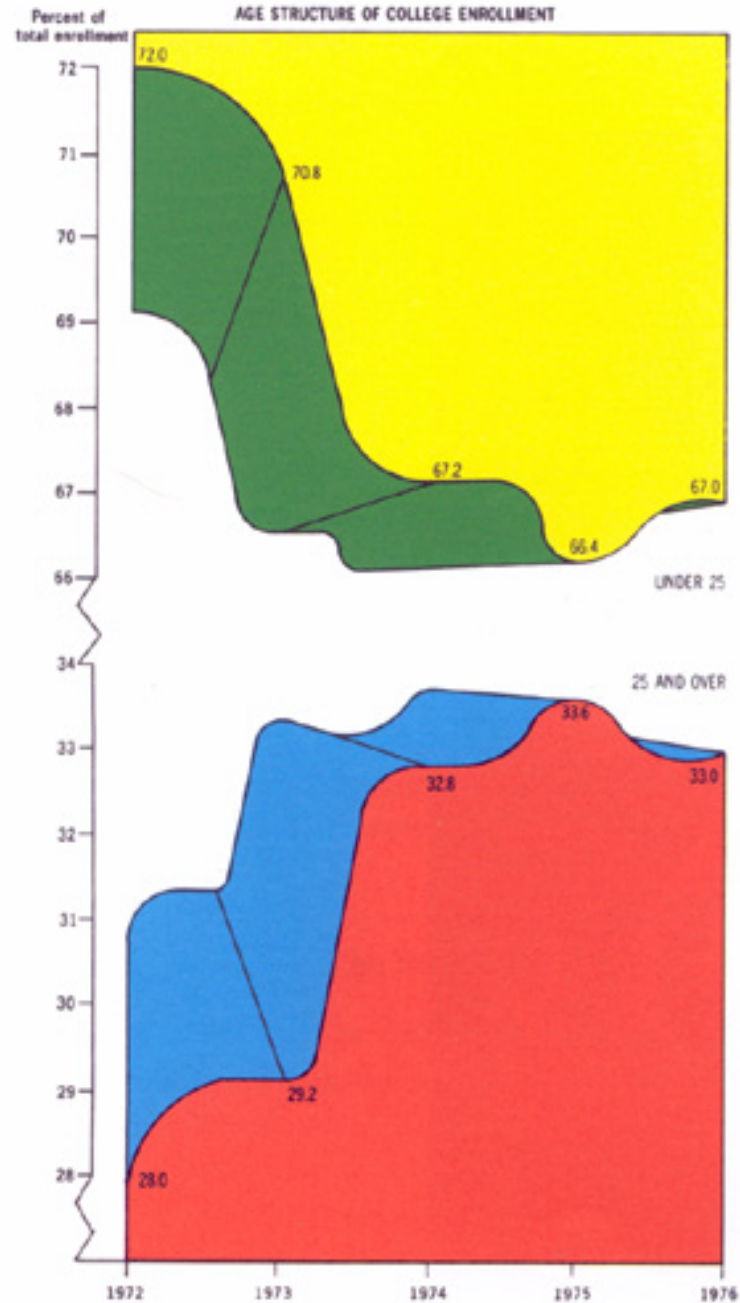
Minimize use of colors!

Never use red/green to contrast important groups!



bad practice: chart junk!

Data first, pretty
second!



the art of graphing: presenting data clearly

Bad graphing practices:

- Setting false baselines
- Comparing apples and oranges
- 2 dimensional graphics representing 1 dimensional data
- Distorting effects
- Ineffective comparisons
- Extra non-information carrying graphics
- Excessive prettiness

from: Tufte, E.R. (1983). The visual display of quantitative information.

General principles:

- Maximize the amount of data presented for the amount of ink used.

Applying principles

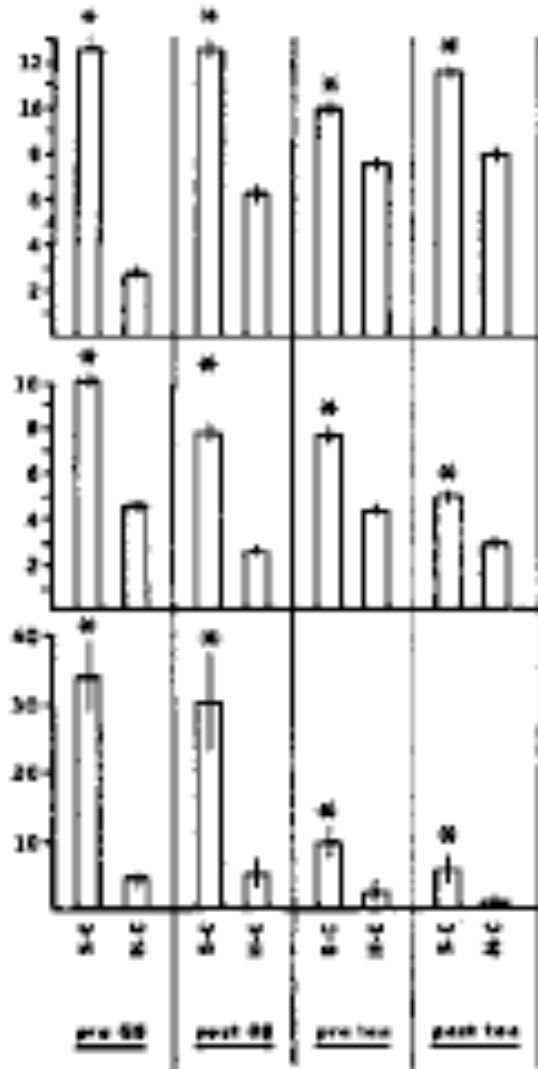
Maximize data-ink:

- Superior graphics use all of their ink to present important information about their data

Case study: two takes on the bar chart

Applying principles

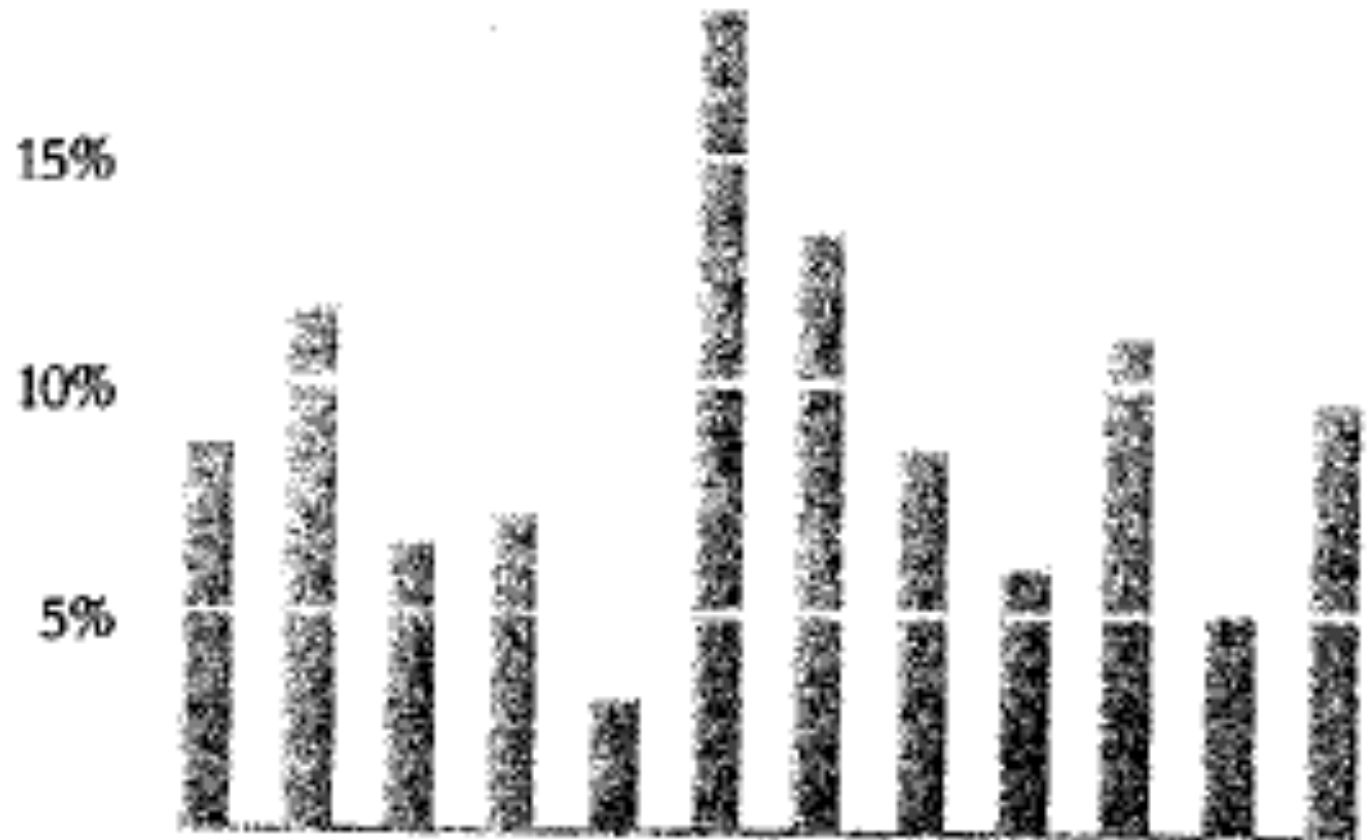
Maximizing data-ink on bar chart



Applying principles

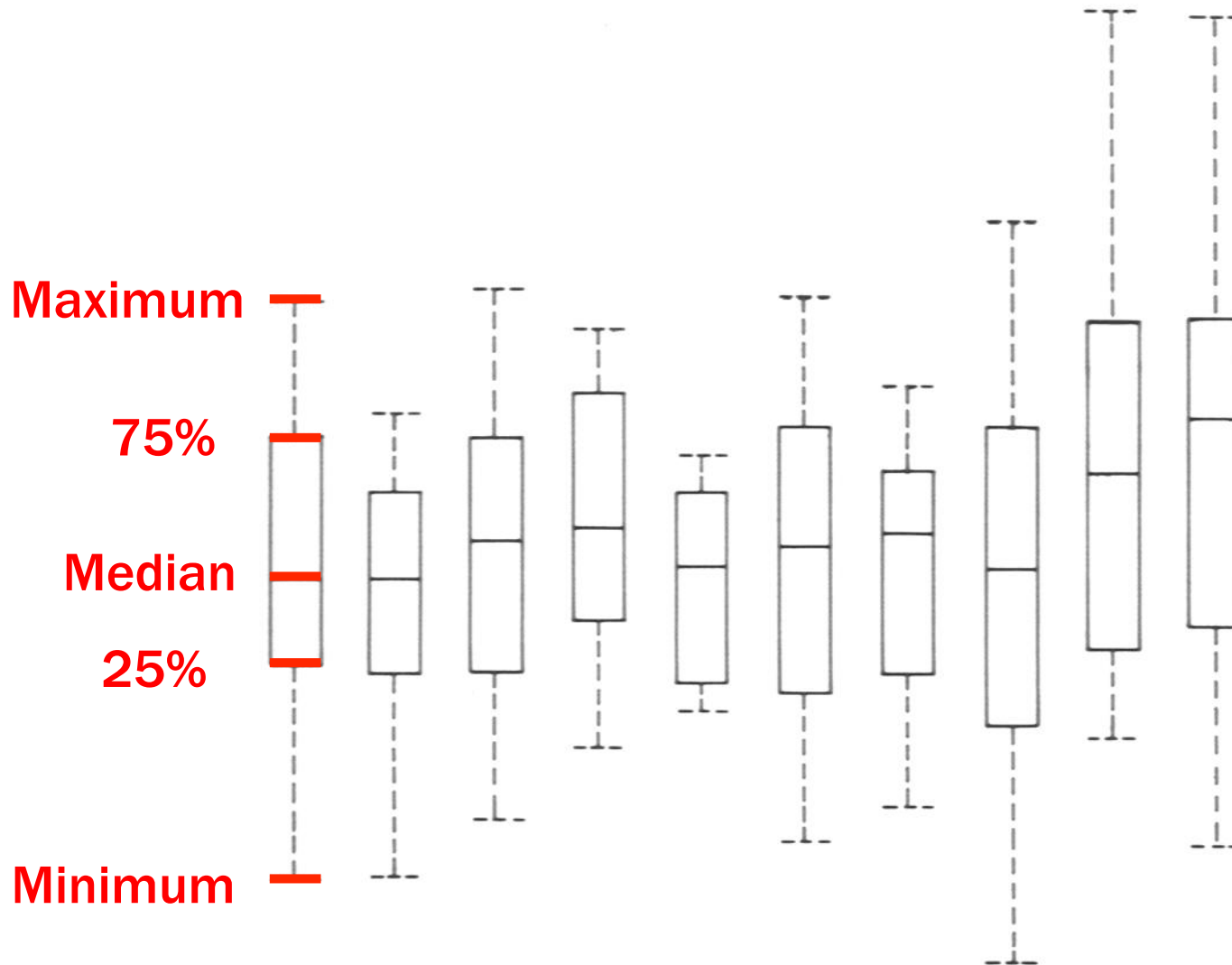
Maximizing data-ink on bar chart

- Note:
- No y axis line
 - No tick marks
 - No grid (grid is shown by breaks in bars)



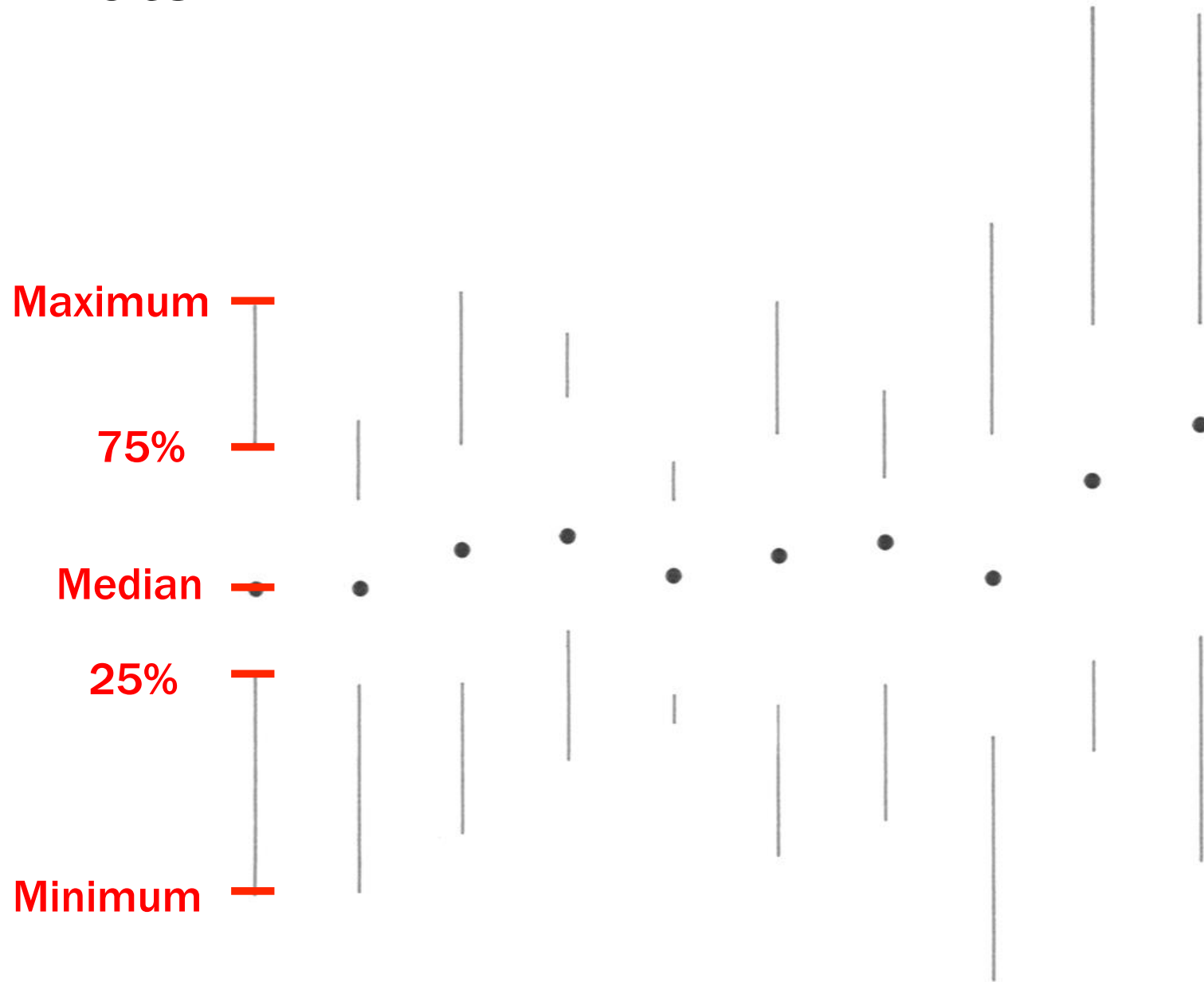
Applying principles

Box Plots



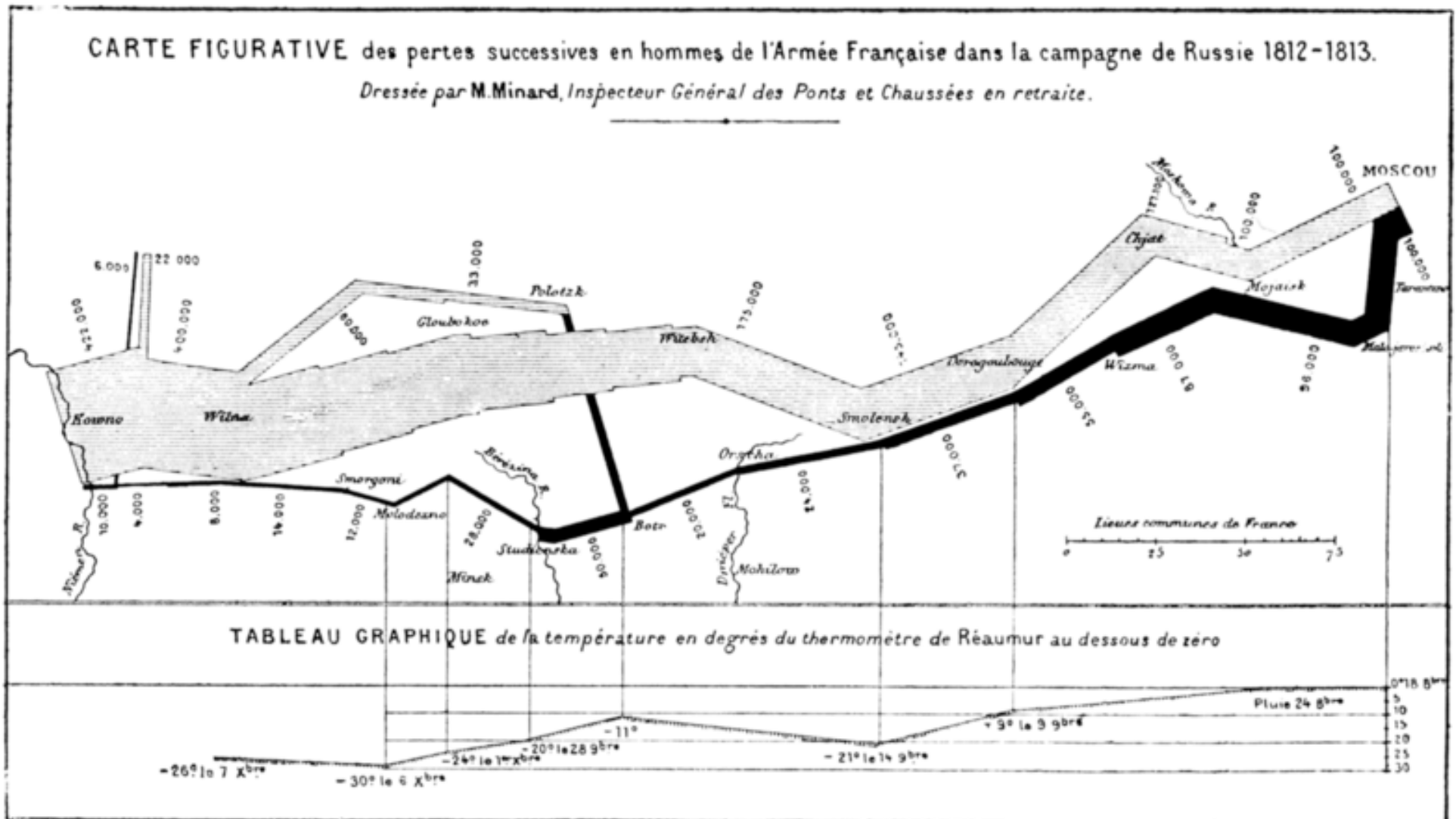
Applying principles

Box Plots



the art of graphing: presenting data clearly

The best statistical graphic ever?



Putting principles to work

- Google X, find all images that are presenting data, argue about whether they are good or bad
 - X = data science
 - X = visualization
 - X = statistics
 - X = infographic
 - X = tufte

Homework for Monday

- If you want to, resubmit the HW due today (it will be marked late, but I will give full credit)
- Create 3 plots of the baseball data in seaborn
 - (You choose what to plot)