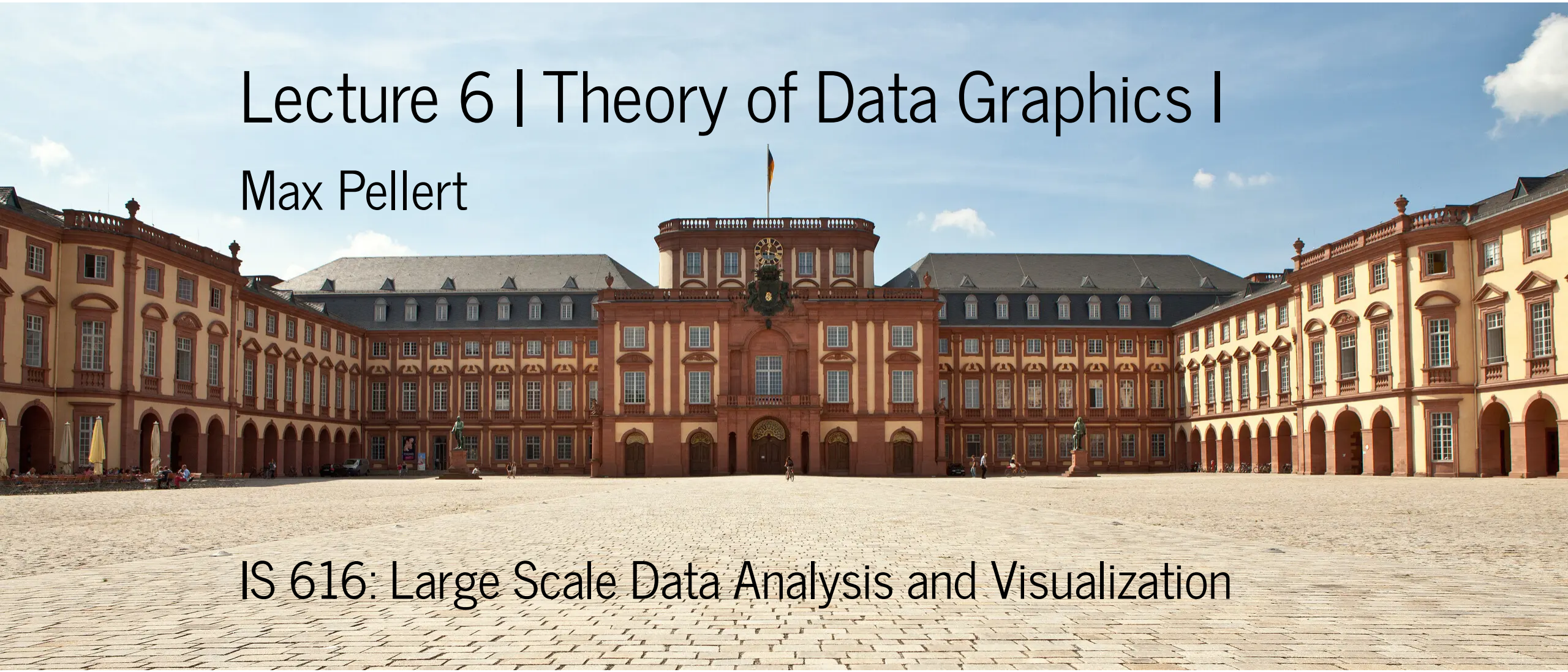
Lecture 6 | Theory of Data Graphics I

Max Pellert



IS 616: Large Scale Data Analysis and Visualization

Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading, and color.

The use of abstract, non-representational pictures to show numbers is a surprisingly recent invention, perhaps because of the diversity of skills required—the visual-artistic, empirical-statistical, and mathematical.

It was not until 1750–1800 that statistical graphics—length and area to show quantity, time-series, scatterplots, and multivariate displays—were invented, long after such triumphs of mathematical ingenuity as logarithms, Cartesian coordinates, the calculus, and the basics of probability theory.

Theory of Data Graphics

The emphasis is on maximizing principles, empirical measures of graphical performance, and the sequential improvement of graphics through revision and editing.

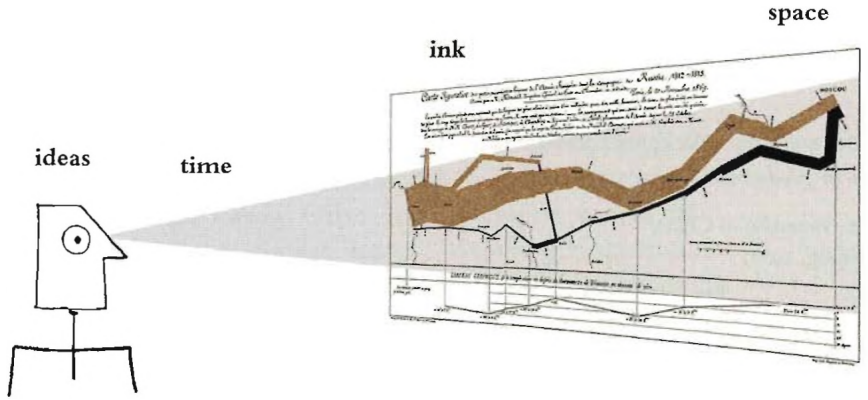
Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency. Graphical displays should

- Show the data
- Induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- Avoid distorting what the data have to say
- Present many numbers in a small space
- Make large data sets coherent

Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency. Graphical displays should

- Encourage the eye to compare different pieces of data
- Reveal the data at several levels of detail, from a broad overview to the fine structure
- Serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- Be closely integrated with the statistical and verbal descriptions of a data set.

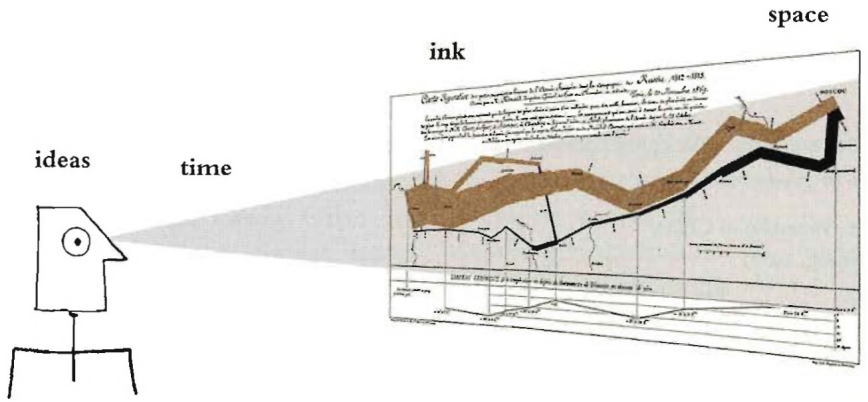
Principles of graphical excellence



Graphical excellence is the well-designed presentation of interesting data—a matter of *substance*, of *statistics*, and of *design*.

Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency.

Principles of graphical excellence



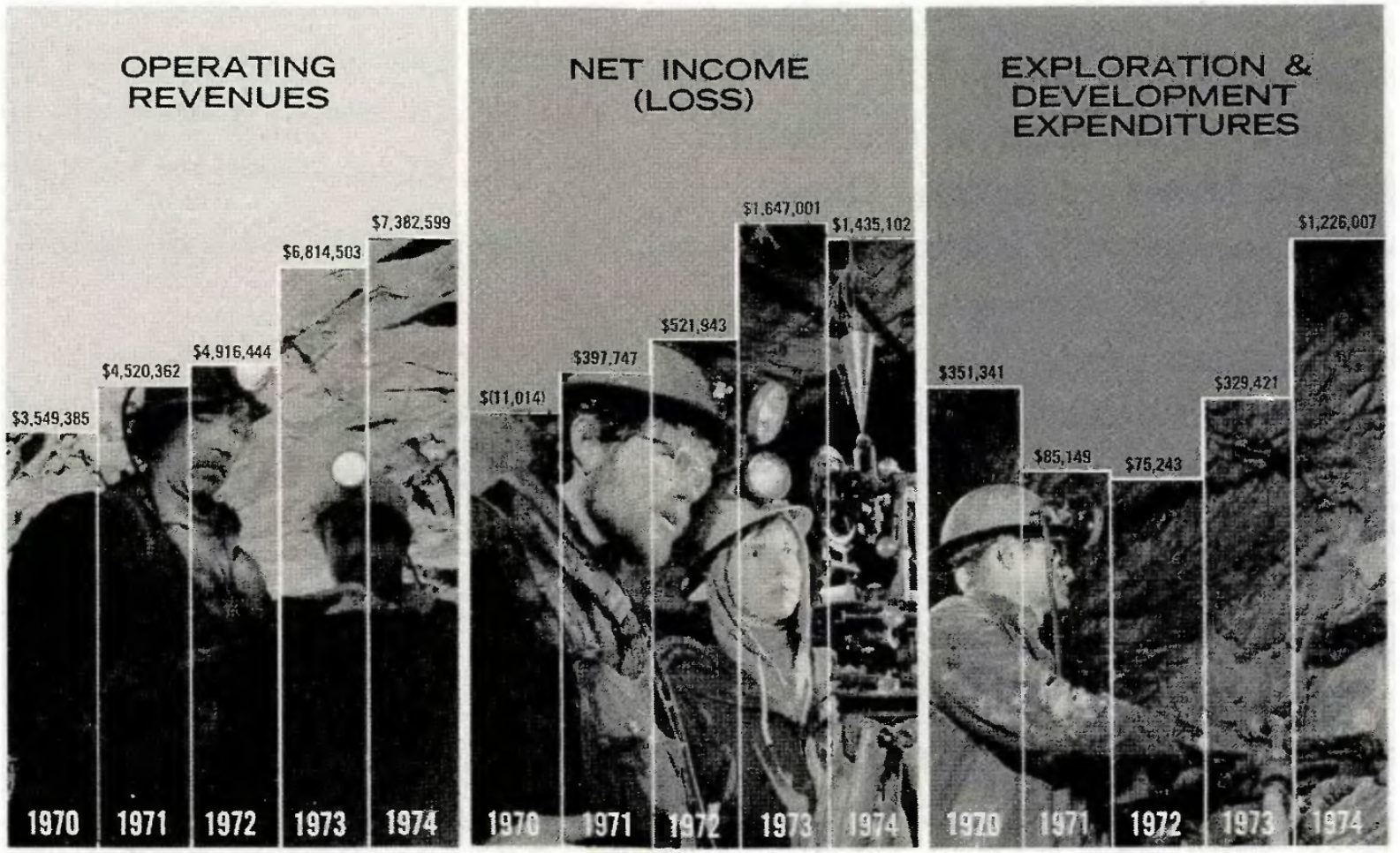
Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.

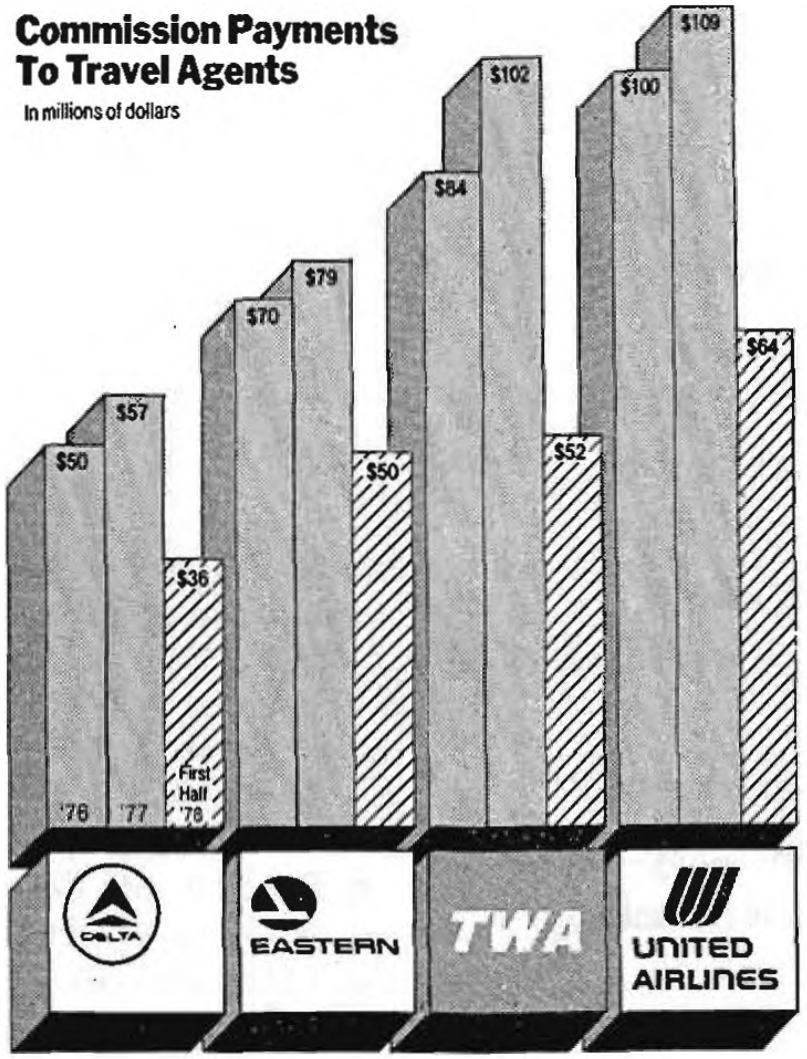
Graphical excellence is nearly always multivariate.

And graphical excellence requires telling the truth about the data.

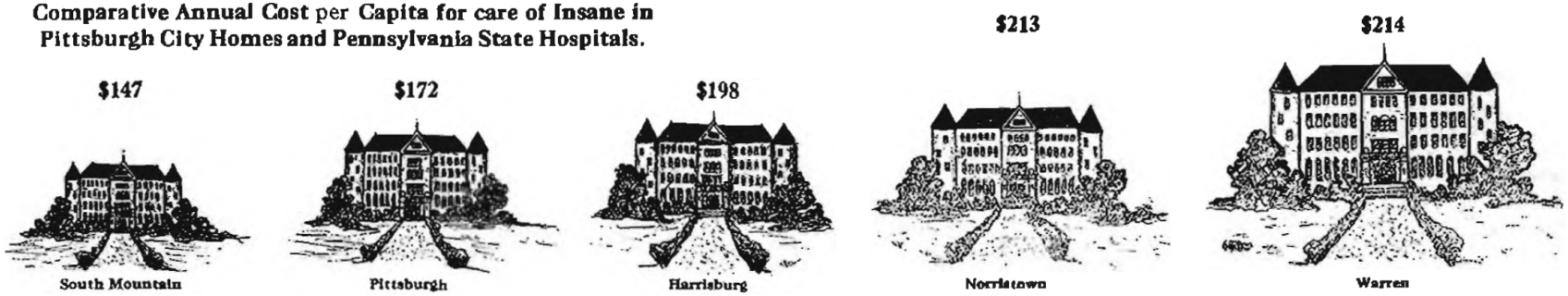
The principles

Avoid distorting what the data have to say: **The Lie Factor**

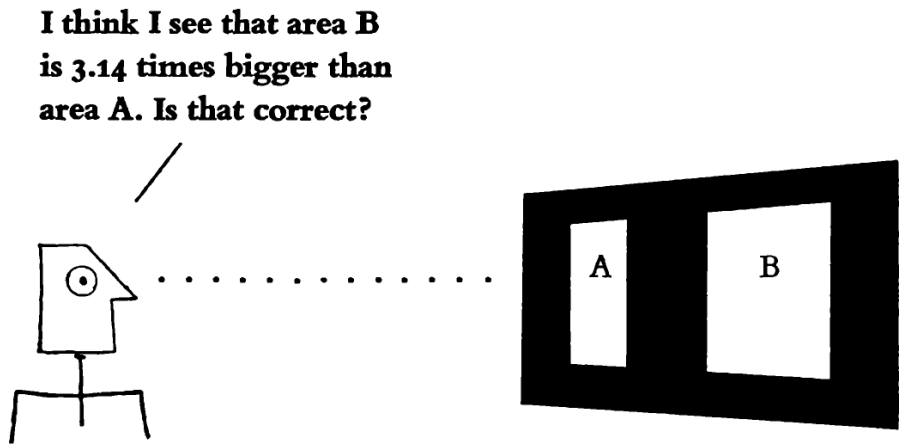

OPERATING REVENUES
$3,549,385
$4,520,362
$4,916,444
$6,814,503
$7,382,599
1970
1971
1972
1973
1974
NET INCOME (LOSS)
$(11,014)
$397,747
$521,943
$1,647,001
$1,435,102
1970
1971
1972
1973
1974
EXPLORATION & DEVELOPMENT EXPENDITURES
$351,341
$85,149
$75,243
$329,421
$1,226,007
1970
1971
1972
1973
1974


Commission Payments
To Travel Agents
In millions of dollars
$50
$57
$36
$70
$79
$50
$84
$102
$52
$100
$109
$64
'76
'77
First Half '78
DELTA
EASTERN
TWA
UNITED AIRLINES

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



Distortions of data



"A graphic does not distort if the visual representation of the data is consistent with the numerical representation."

"At any rate, given the perceptual difficulties, the best we can hope for is some uniformity in graphics (if not in the perceivers) and some assurance that perceivers have a fair chance of getting the numbers right."

Distortions of data

Two principles lead toward these goals and, in consequence, enhance graphical integrity:

The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.

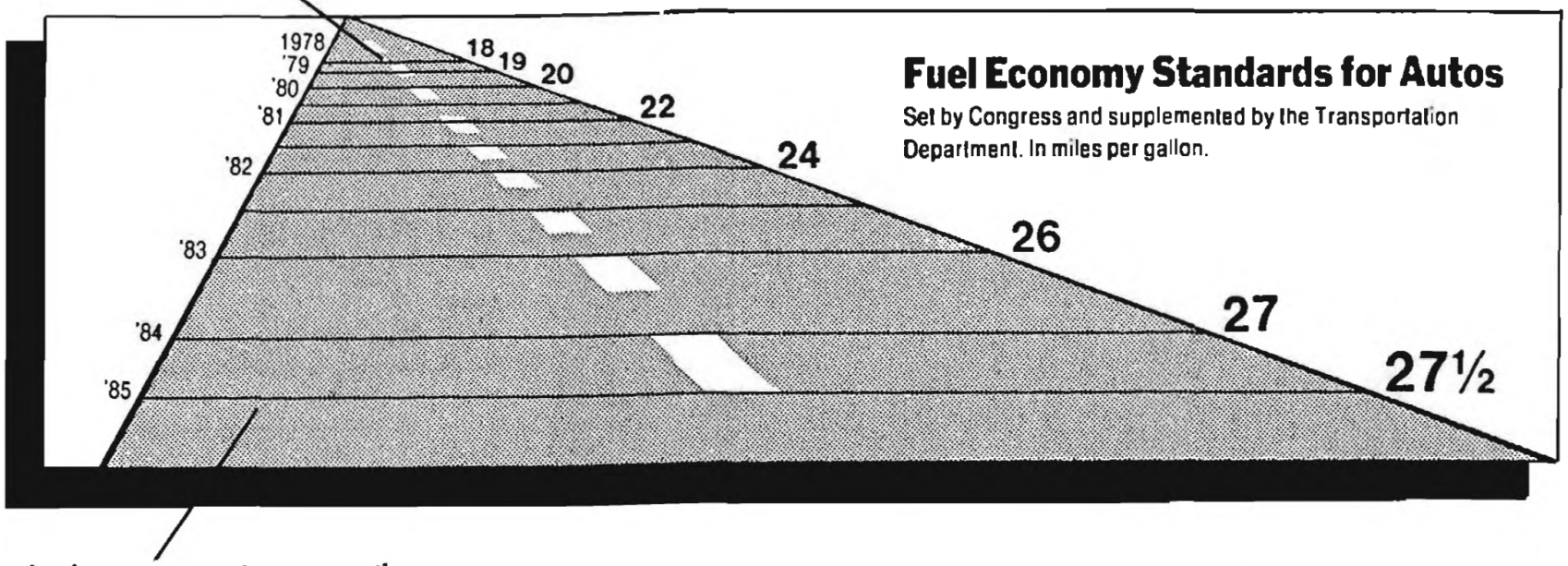
Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.

Violations of the first principle constitute one form of graphic misrepresentation, measured by the

$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

If the Lie Factor is equal to one, then the graphic might be doing a reasonable job of accurately representing the underlying numbers. Lie Factors greater than 1.05 or less than .95 indicate substantial distortion, far beyond minor inaccuracies in plotting.

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Here is an extreme example. A newspaper reported that the U.S. Congress and the Department of Transportation had set a series of fuel economy standards to be met by automobile manufacturers, beginning with 18 miles per gallon in 1978 and moving in steps up to 27.5 by 1985, an increase of 53 percent:

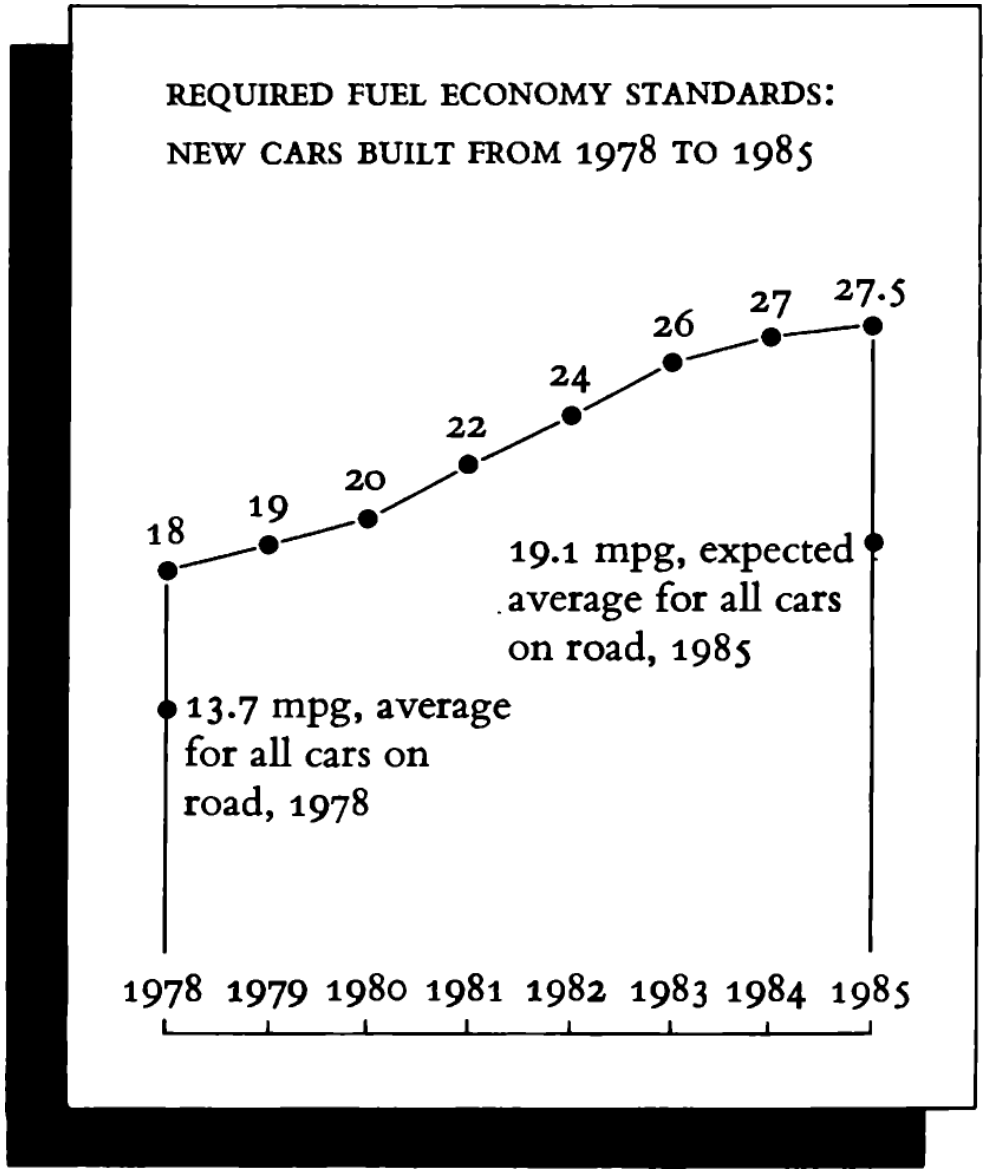
$$\frac{27.5 - 18.0}{18.0} \times 100 = 53\%$$

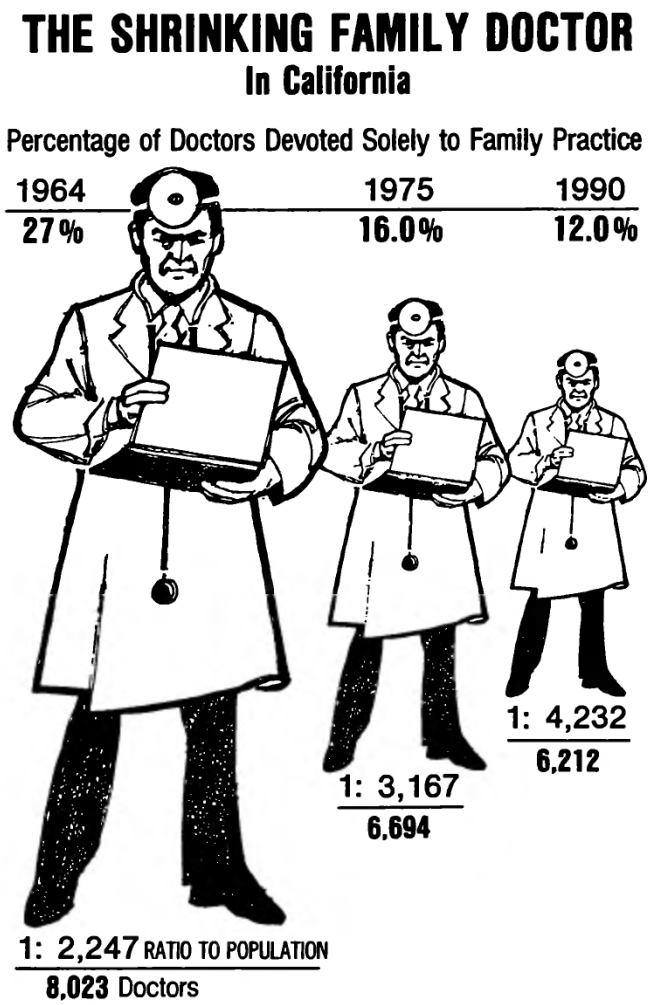
The magnitude of the change from 1978 to 1985 is shown in the graph by the relative lengths of the two lines:

$$\frac{5.3 - 0.6}{0.6} \times 100 = 783\%$$

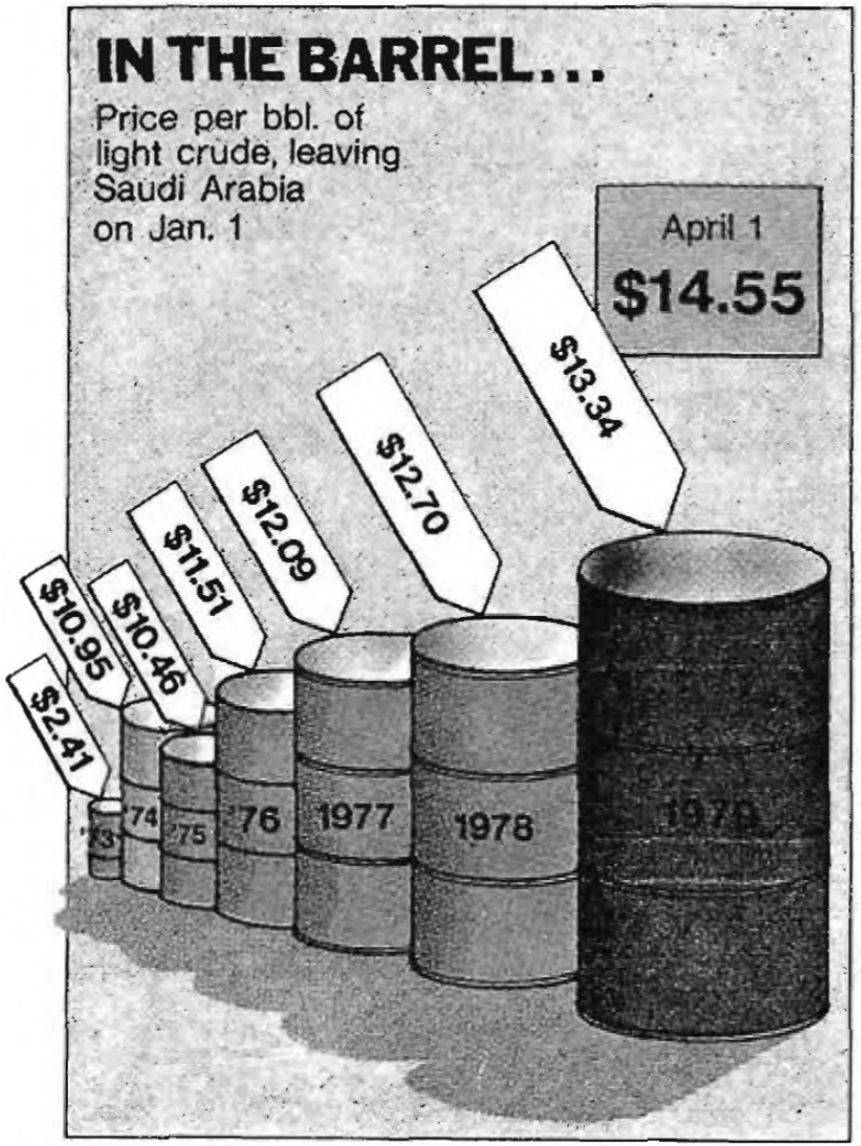
Thus the numerical change of 53 percent is presented by some lines that changed 783 percent, yielding

$$\text{Lie Factor} = \frac{783}{53} = 14.8$$


REQUIRED FUEL ECONOMY STANDARDS:
NEW CARS BUILT FROM 1978 TO 1985
18
19
20
22
24
26
27
27.5
19.1 mpg, expected average for all cars on road, 1985
13.7 mpg, average for all cars on road, 1978
1978 1979 1980 1981 1982 1983 1984 1985



Using areas for one-dimensional data with Lie Factor of 2.8


IN THE BARREL...
Price per bbl. of light crude, leaving Saudi Arabia on Jan. 1
April 1
$14.55
$13.34
$12.70
$12.09
$11.51
$10.46
$10.95
$2.41
'73
74
75
76
1977
1978
1979

Here an increase of 454% is depicted as an increase of 4,280%, for a Lie Factor of 9.4

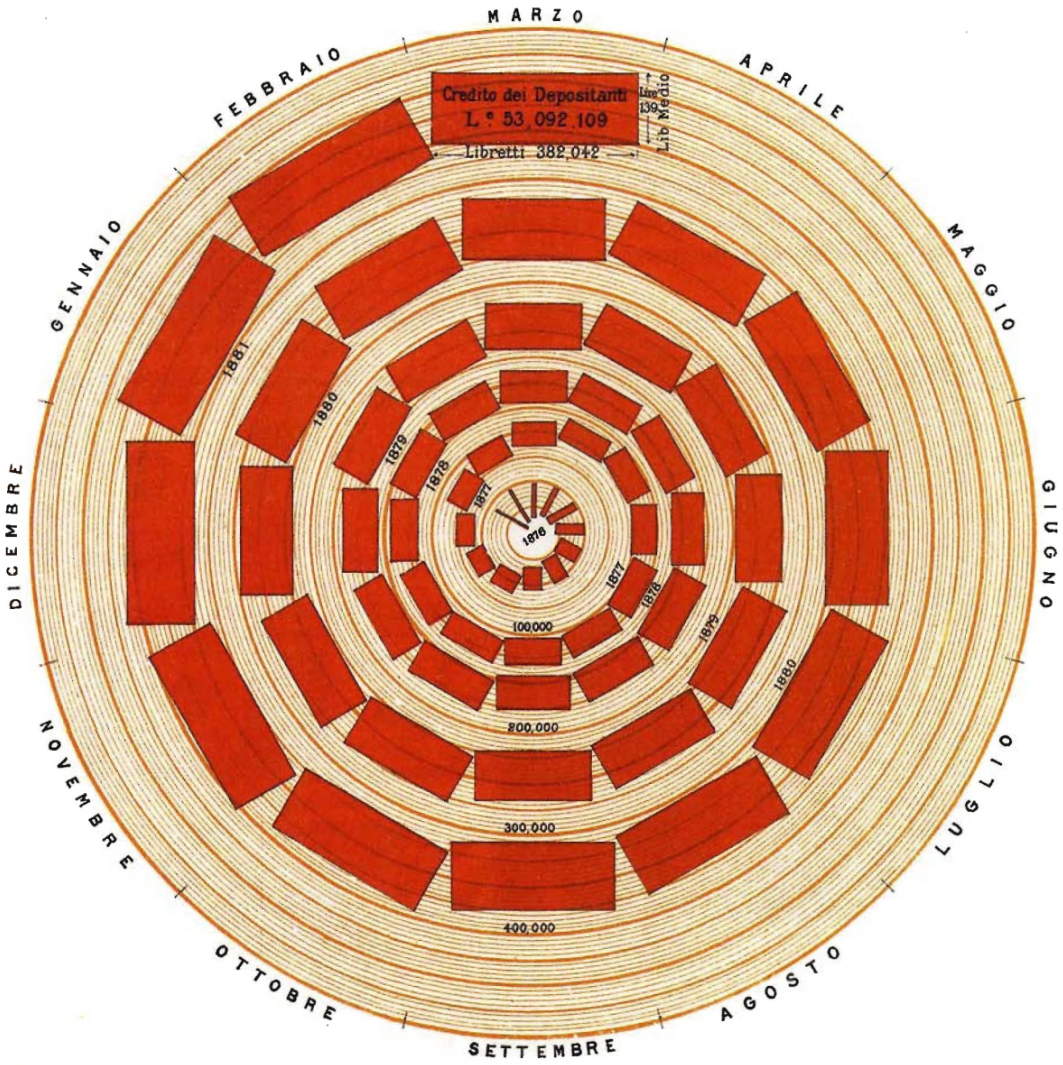
There are considerable ambiguities in how people perceive a two-dimensional surface and then convert that perception into a one-dimensional number. Changes in physical area on the surface of a graphic do not reliably produce appropriately proportional changes in perceived areas. The problem is all the worse when the areas are tricked up into three dimensions:

By surface area, the Lie Factor for this graphic is 9.4. But, if one takes the barrel metaphor seriously and assumes that the *volume* of the barrels represents the price change, then the volume from 1973 to 1979 increases 27,000 percent compared to a data increase of 454 percent, for a Lie Factor of 59.4, which is a record.

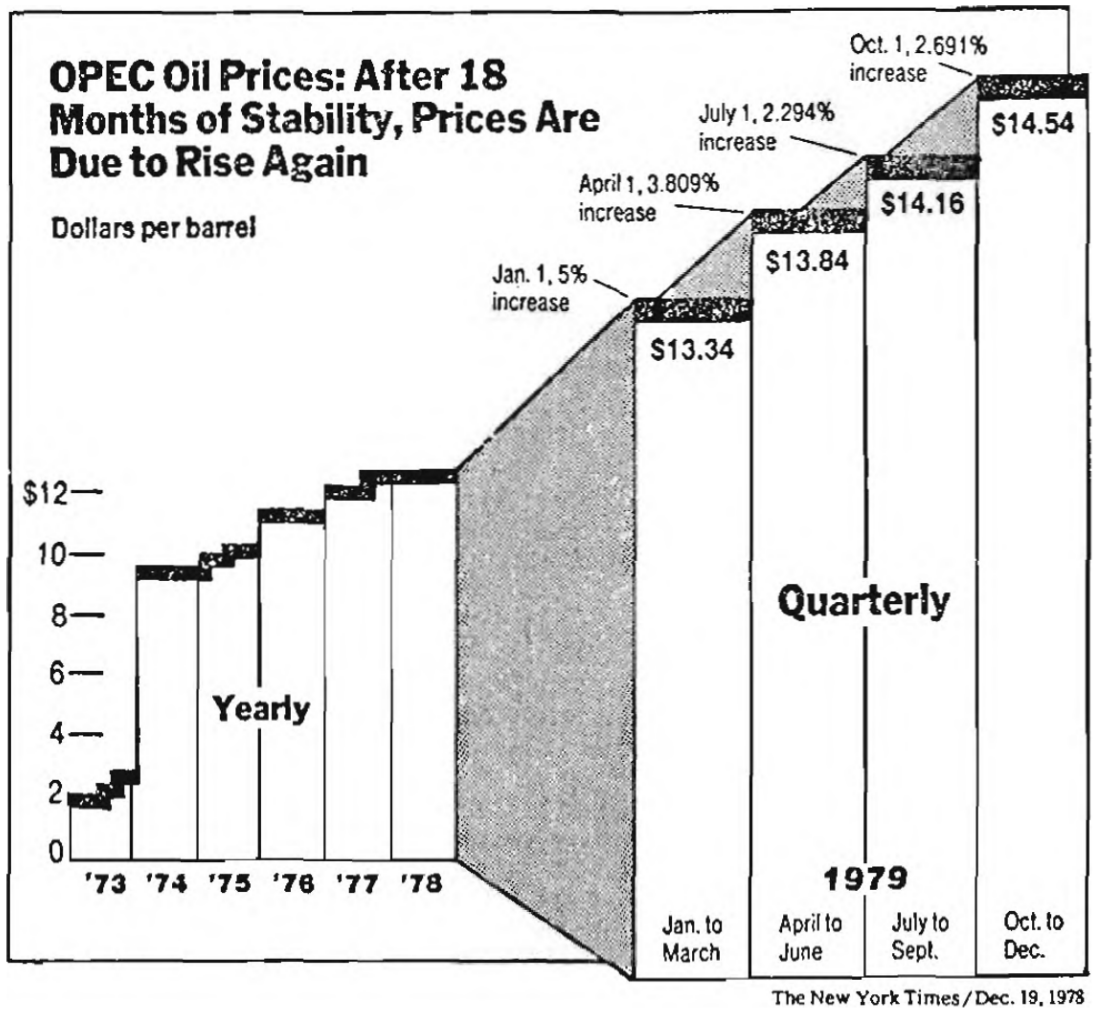
The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.

CASSE POSTALI DI RISPARMIO ITALIANE

Numero dei Libretti, Libretto medio e Deposito totale
al fine di ogni mese



Design variation corrupts this display:



Show data variation, not design variation.

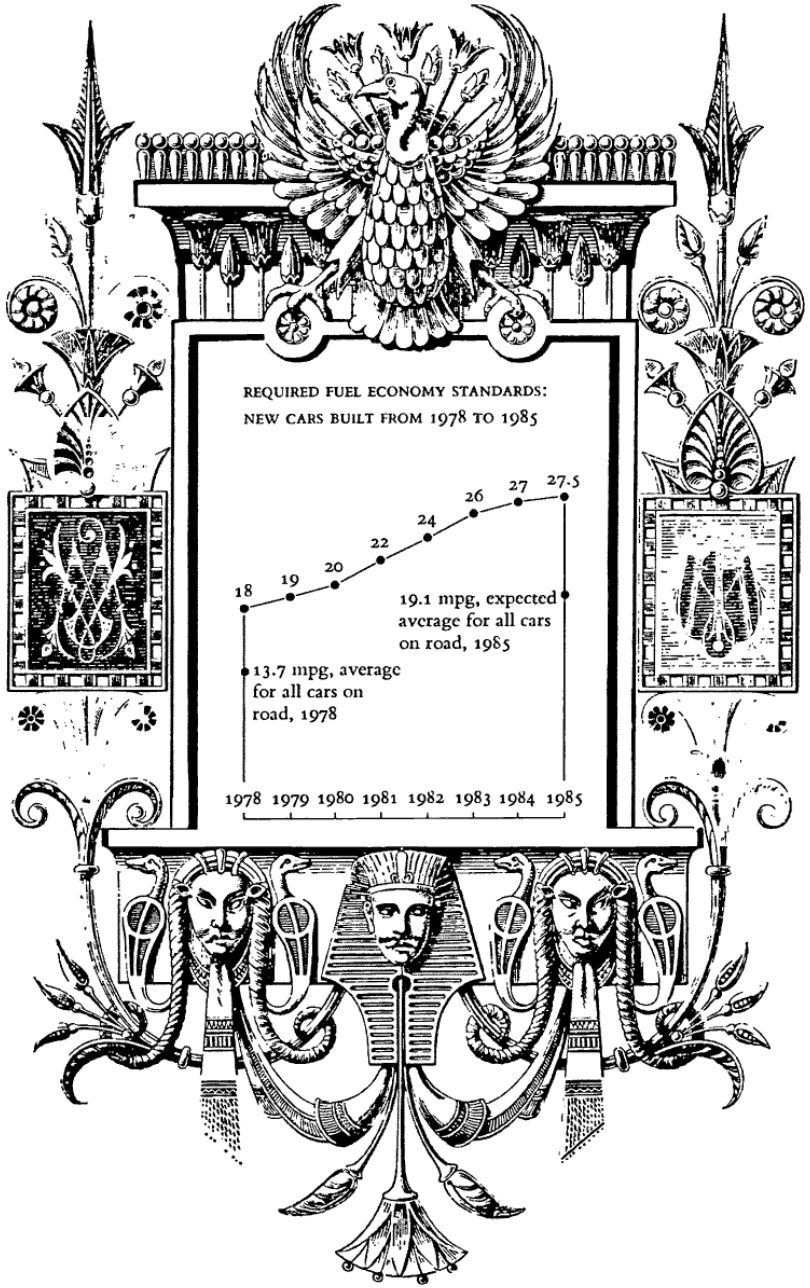
Show the data & induce the viewer to think about the substance: avoid **Chartjunk**

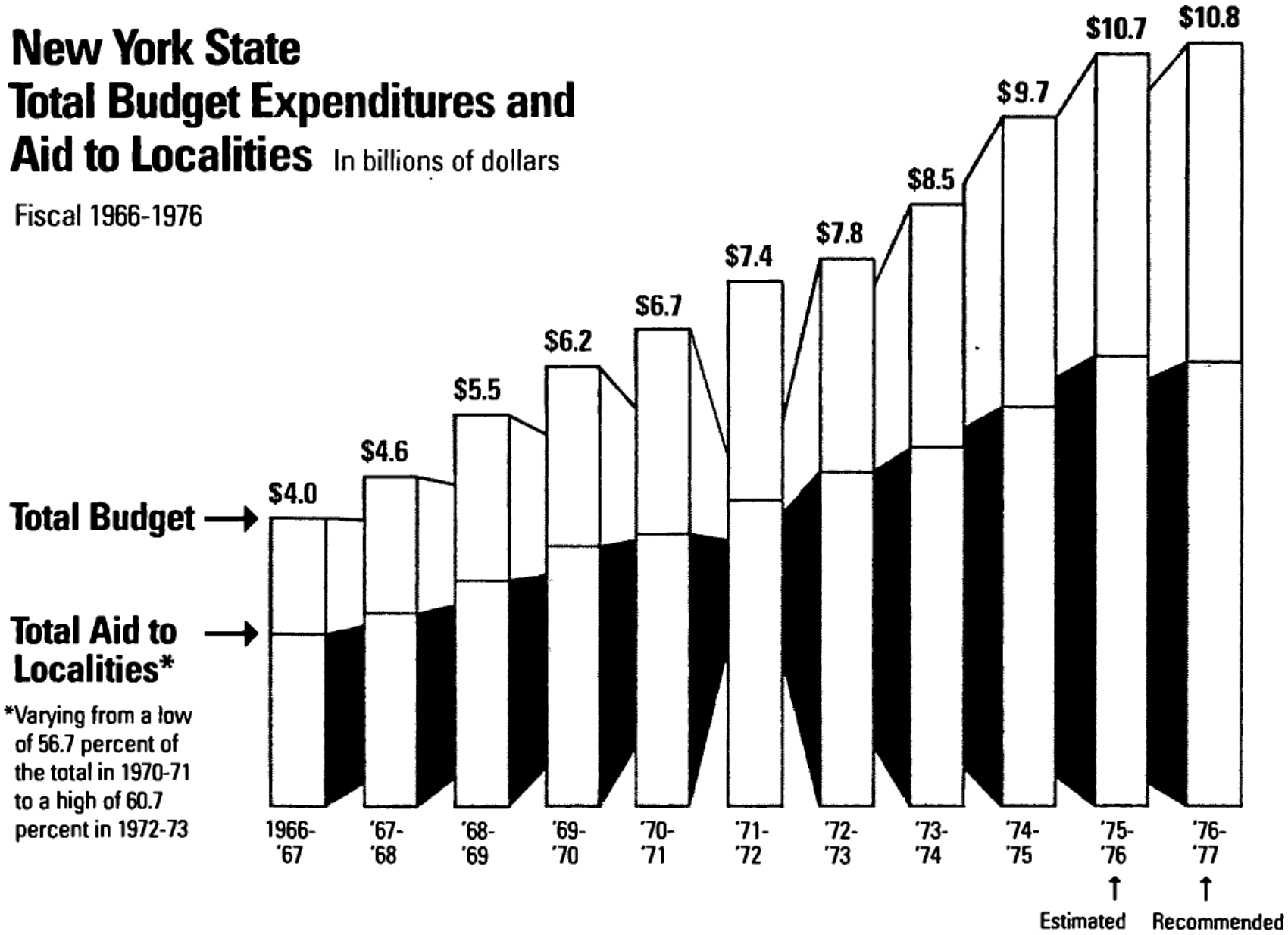
Much of the “winter” in data graphics from the early 20th until roughly 1970 is due to the strong assumptions then…

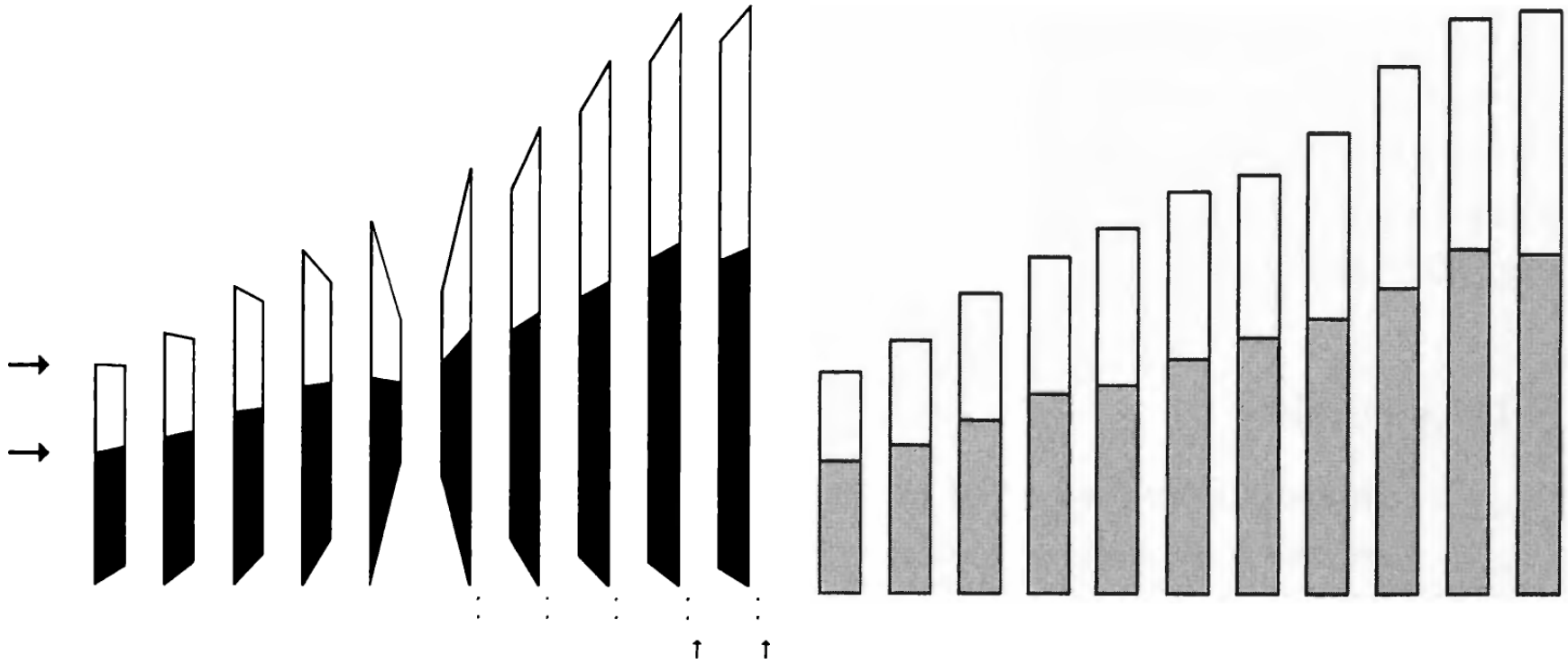
> that graphics had to be “alive,” “com-
> municatively dynamic,” overdecorated and exaggerated (other-
> wise all the dullards in the audience would fall asleep in the face
> of those boring statistics).

This lead to some bizarre ornaments and other things that designer added to scientific visualizations

Luckily, this is not the most pressing issue anymore as the by now predominant computerized way of creating graphics often got the seperate profession of such “chart designers” out of the way, but many other unnecessary elements can still be chartjunk nowadays


REQUIRED FUEL ECONOMY STANDARDS:
NEW CARS BUILT FROM 1978 TO 1985
18
19
20
22
24
26
27
27.5
19.1 mpg, expected average for all cars on road, 1985
13.7 mpg, average for all cars on road, 1978
1978 1979 1980 1981 1982 1983 1984 1985


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 60.7 percent in 1972-73
$4.0
$4.6
$5.5
$6.2
$6.7
$7.4
$7.8
$8.5
$9.7
$10.7
$10.8
1966-'67
'67-'68
'68-'69
'69-'70
'70-'71
'71-'72
'72-'73
'73-'74
'74-'75
'75-'76
'76-'77
Estimated
Recommended



Chartjunk

Some minimalistic pre-set theme (for example `theme_bw` in ggplot2) can often be a quick fix to already get rid of some chartjunk

But usually it takes some tinkering, programmatically and also using external tools (such as Inkscape) to remove everything that is unimportant

But: Chartjunk has also attracted some interests in academia, particularly about the effect on memorability or engagement.

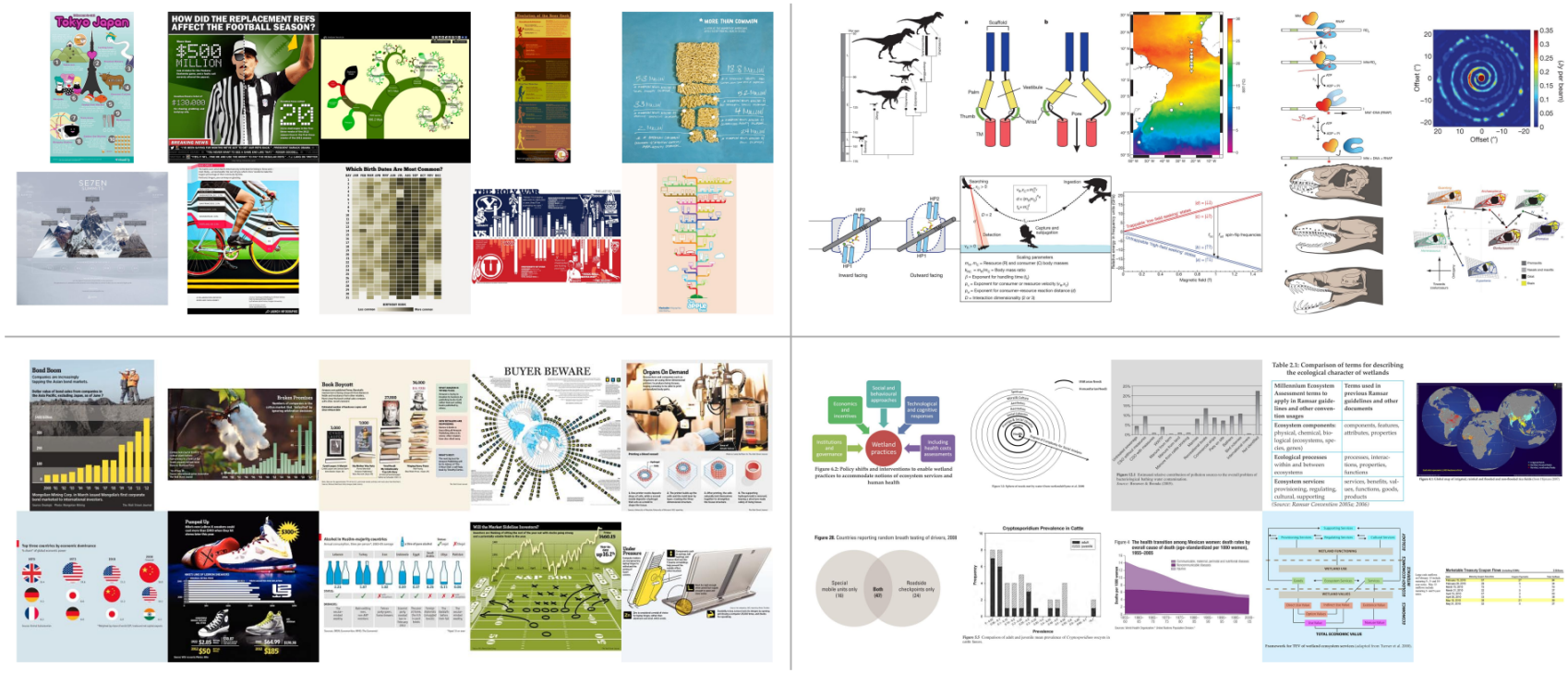
ABSTRACT

Guidelines for designing information charts often state that the presentation should reduce 'chart junk' – visual embellishments that are not essential to understanding the data.

In contrast, some popular chart designers wrap the presented data in detailed and elaborate imagery, raising the questions of whether this imagery is really as detrimental to understanding as has been proposed, and whether the visual embellishment may have other benefits. To investigate these issues, we conducted an experiment that compared embellished charts with plain ones, and measured both interpretation accuracy and long-term recall.

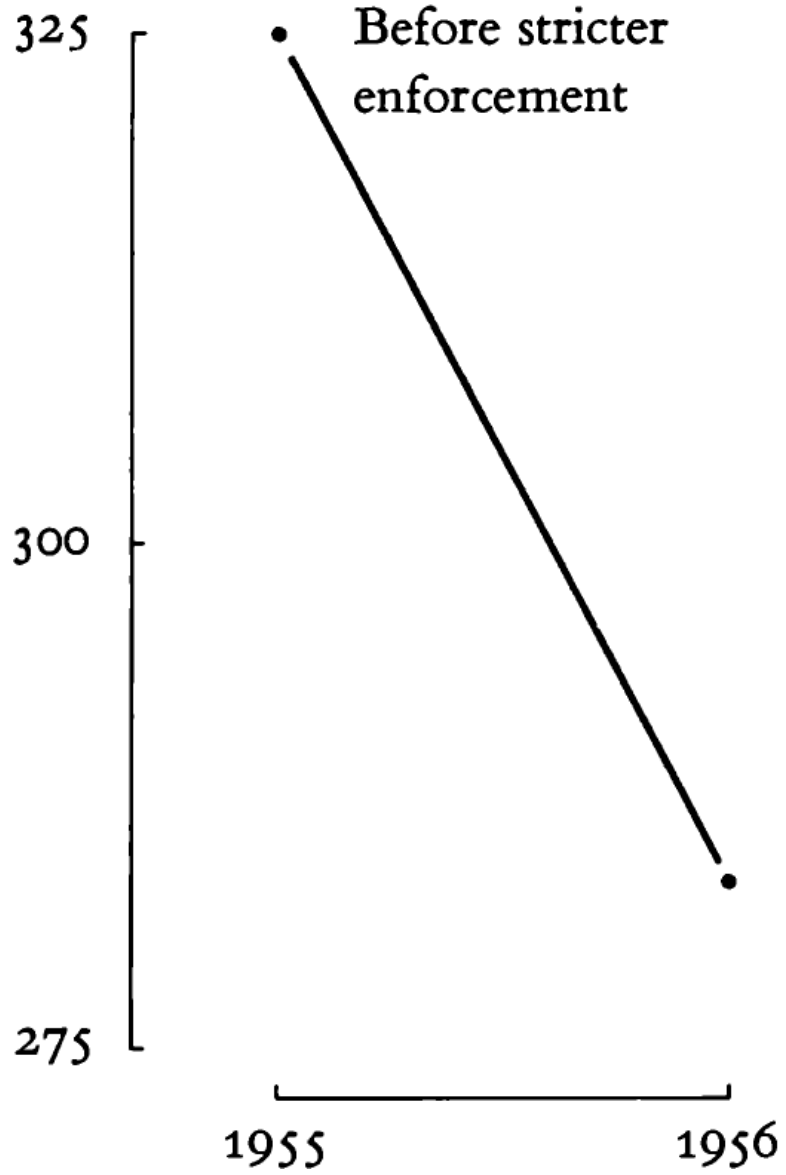
We found that people's accuracy in describing the embellished charts was no worse than for plain charts, and that their recall after a two-to-three-week gap was significantly better. Although we are cautious about recommending that all charts be produced in this style, our results question some of the premises of the minimalist approach to chart design.

Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., McDine, D., & Brooks, C. (2010). Useful junk?: The effects of visual embellishment on comprehension and memorability of charts. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2573–2582. https://doi.org/10.1145/1753326.1753716



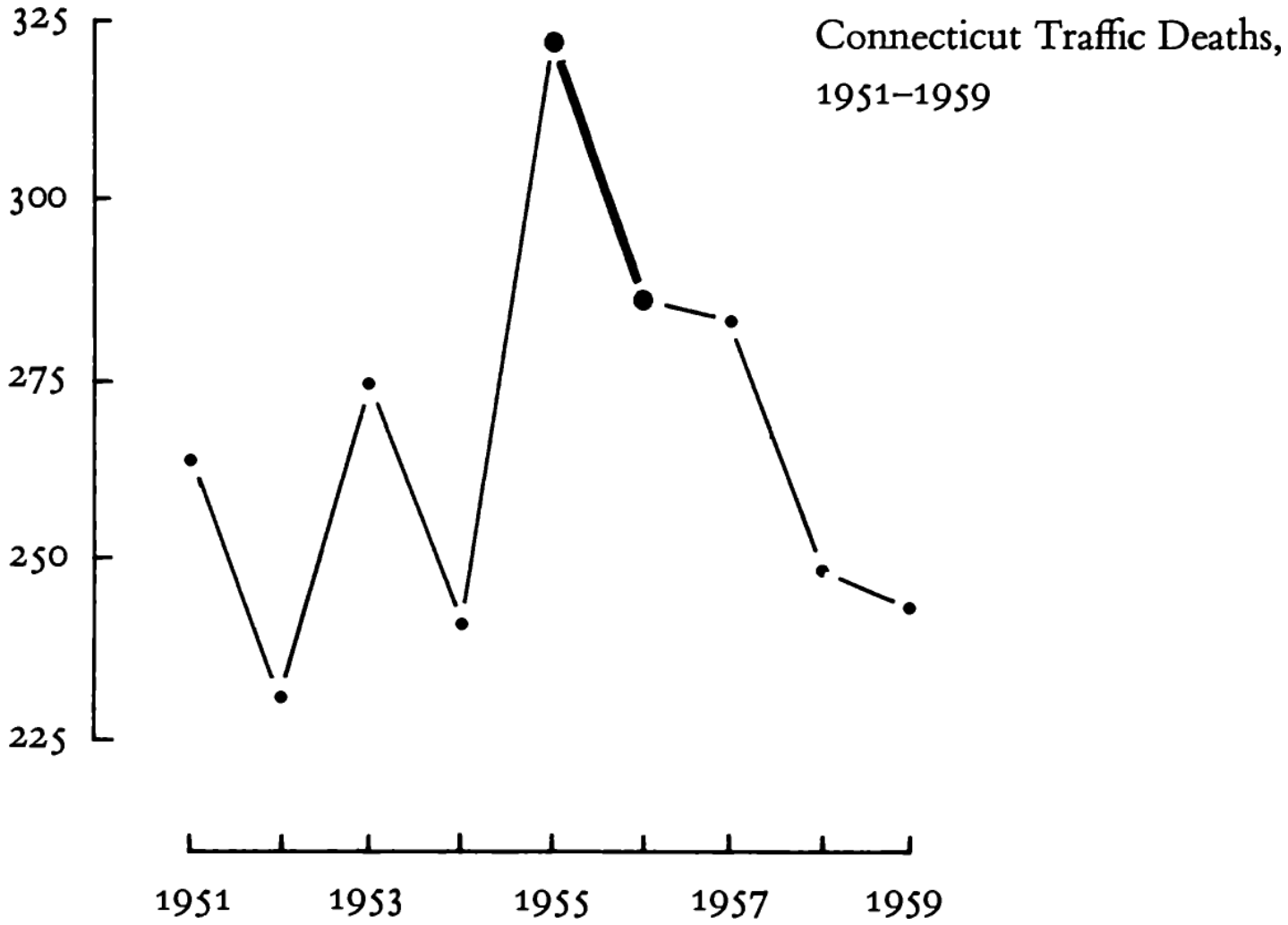
Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What Makes a Visualization Memorable? IEEE Transactions on Visualization and Computer Graphics, 19(12), 2306–2315. https://doi.org/10.1109/TVCG.2013.234

Context is essential for graphical integrity



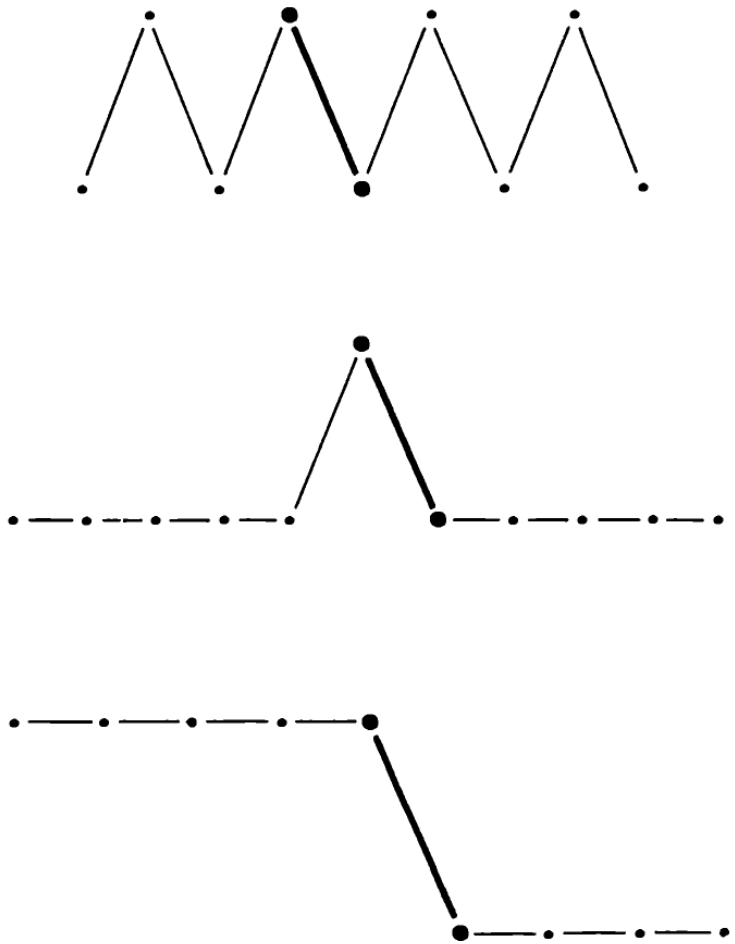
Connecticut Traffic Deaths,
Before (1955) and After (1956)
Stricter Enforcement by the Police
Against Cars Exceeding Speed limit

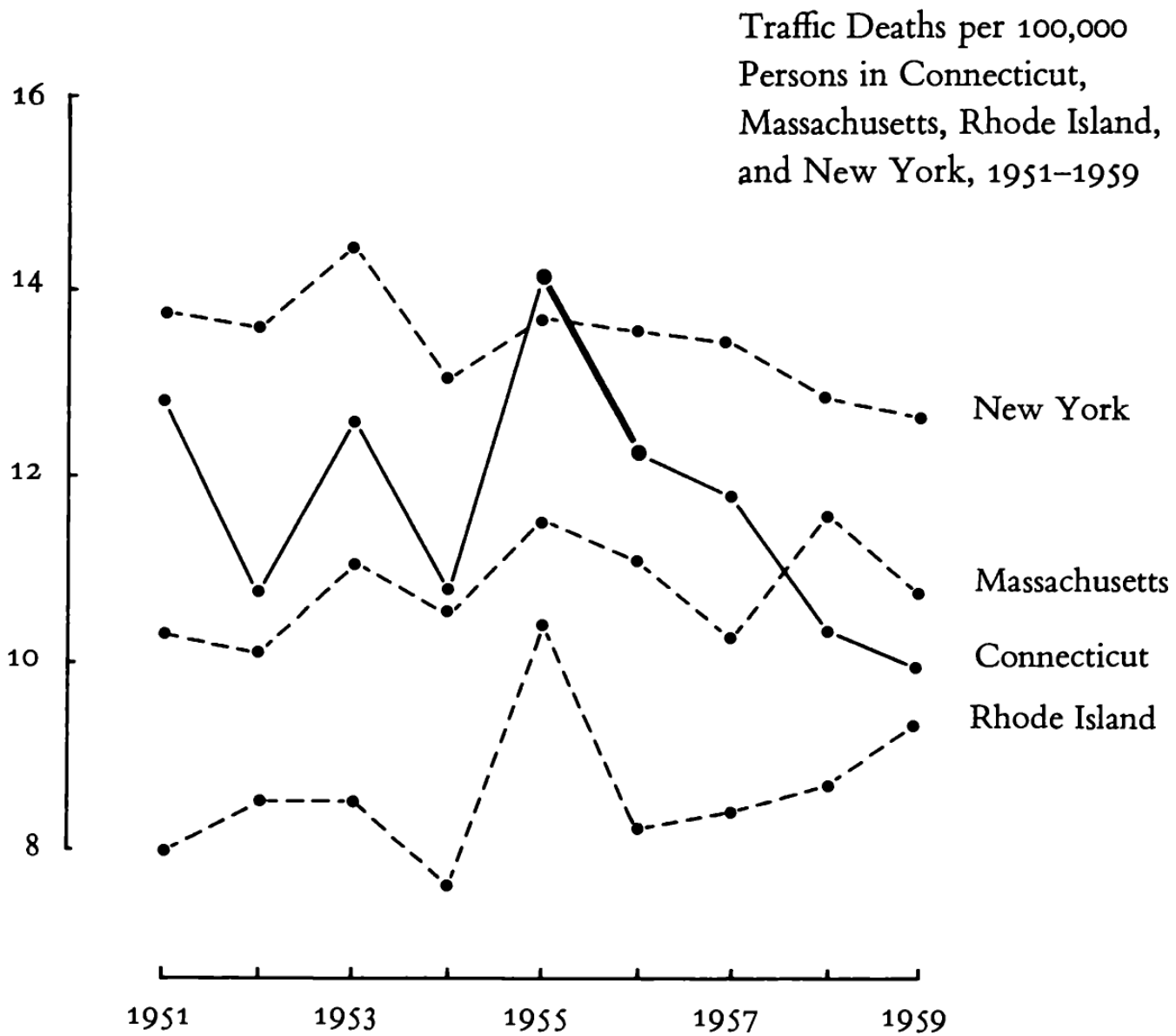
A few more data points add immensely to the account:

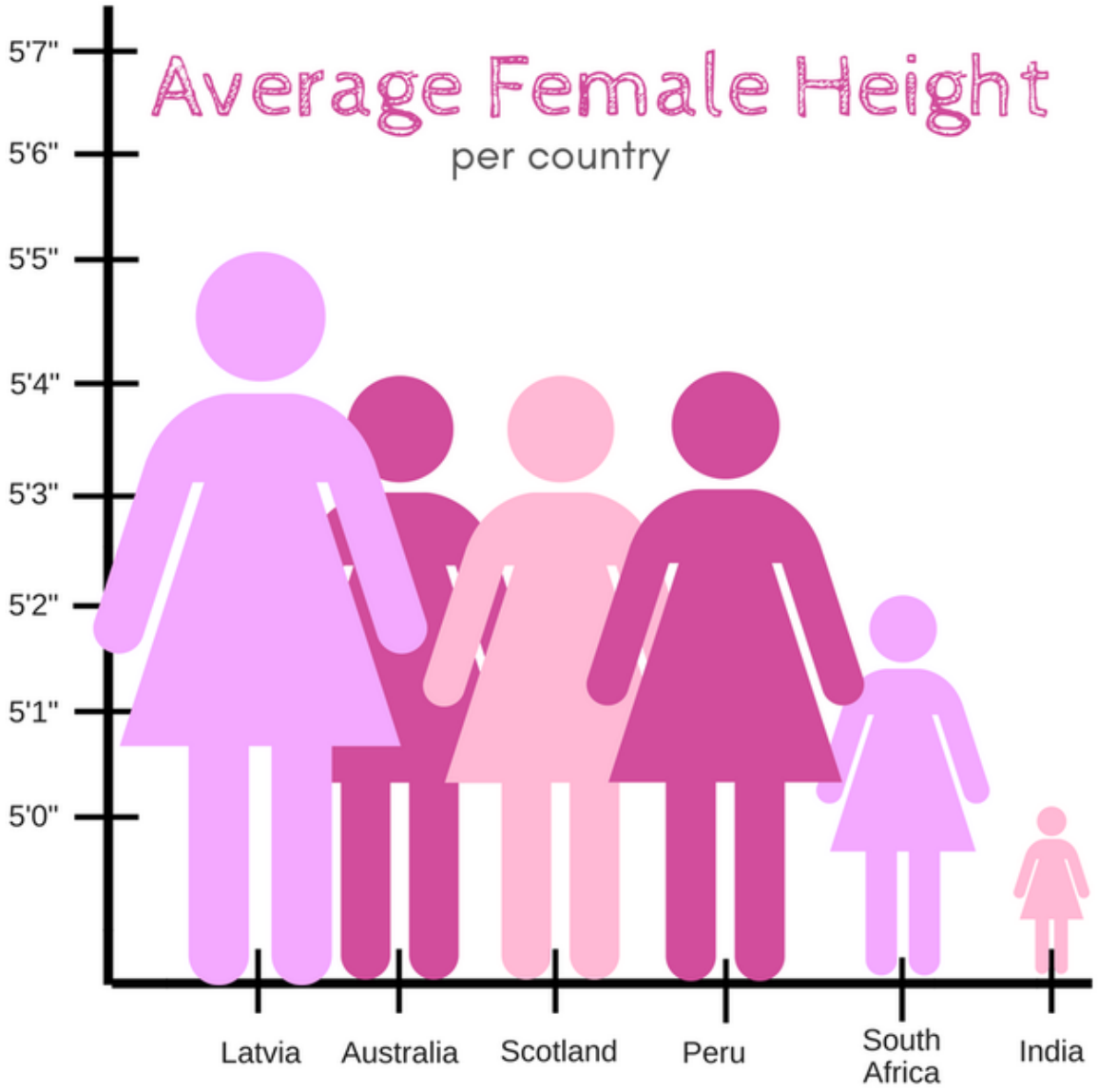


Connecticut Traffic Deaths,
1951–1959

Imagine the very different interpretations other possible time-paths surrounding the 1955–1956 change would have:




Traffic Deaths per 100,000 Persons in Connecticut, Massachusetts, Rhode Island, and New York, 1951–1959
16
14
12
10
8
1951
1953
1955
1957
1959
New York
Massachusetts
Connecticut
Rhode Island


Average Female Height
per country
5'7"
5'6"
5'5"
5'4"
5'3"
5'2"
5'1"
5'0"
Latvia
Australia
Scotland
Peru
South Africa
India

https://twitter.com/reina_sabah/status/1291509085855260672

Graphics must not quote data out of context.

> The emaciated, data-thin design should always provoke suspicion, for graphics often lie by omission, leaving out data sufficient for comparisons.

Leads to lack of integrity

Connected to cherry-picking the data

Be very careful with truncating axes!

The best graphic cannot help conceal selective reporting

The minimalist perspective of Tufte advocates plain and simple charts that maximize the proportion of data-ink (the ink in the chart used to represent data):

The Ink Data Ratio

Acknowledgements

https://yy.github.io/dviz-course/

https://yyahn.com/dviz-course/m05-design/class/