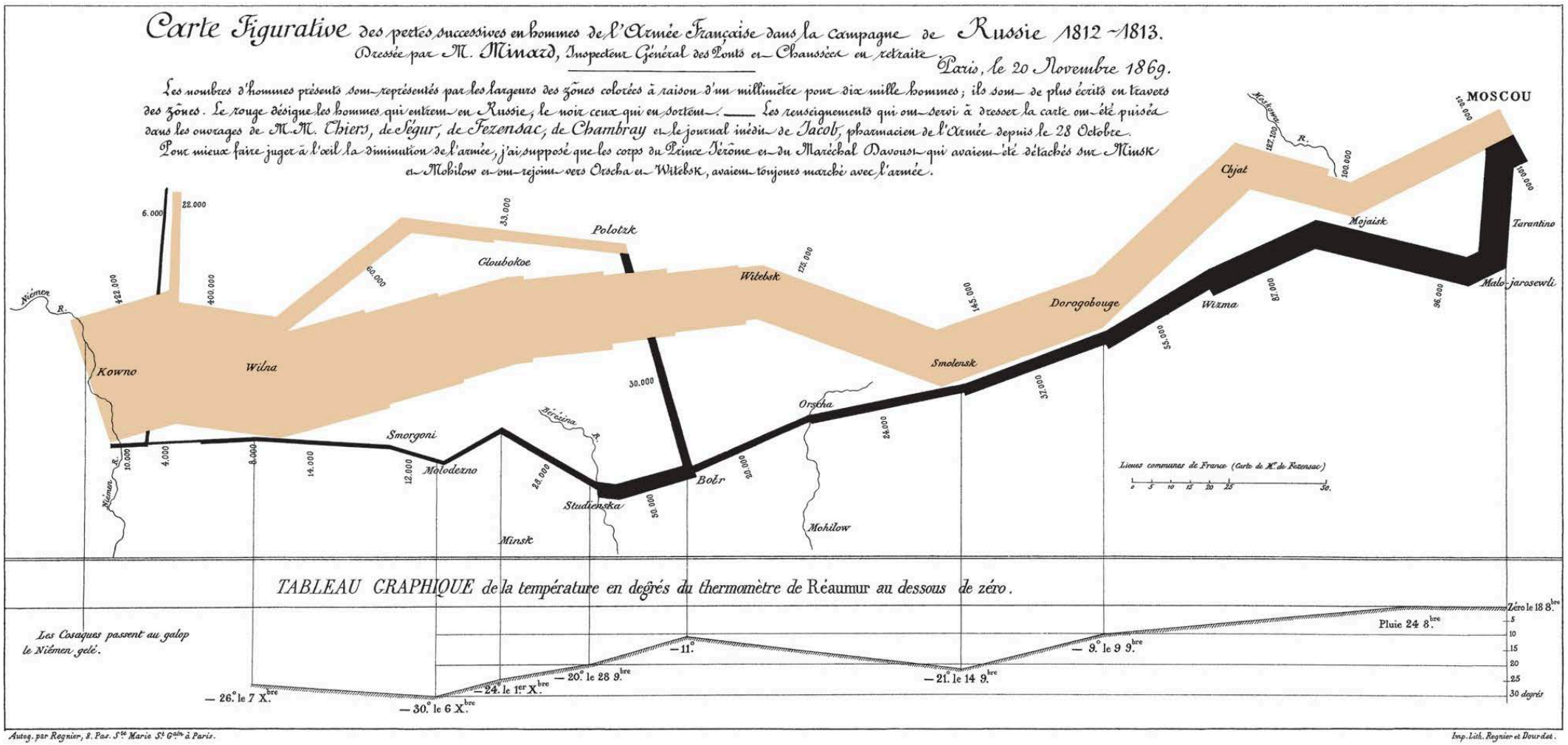
CS-E4840
Information Visualization
Lecture 2: Graphical practice

Tassu Takala <tapio.takala@aalto.fi>
4 March 2021

Recap

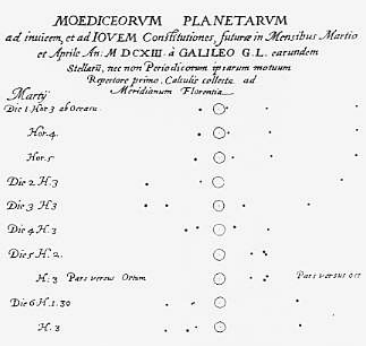
examples of good data graphics

Carte Figurative by Minard



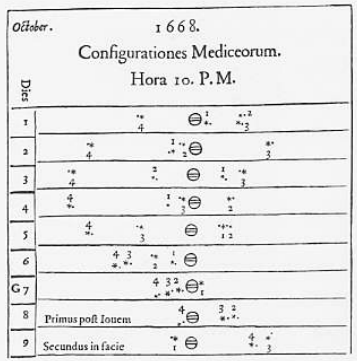
Jovian moons

- On 10 January 1610 Galileo Galilei was able to separate the motion of the Jovian satellites from that of the planet.
- It took 300 years to move from dots to continuous curves, with muted horizontal lines, that report every position of the moons.



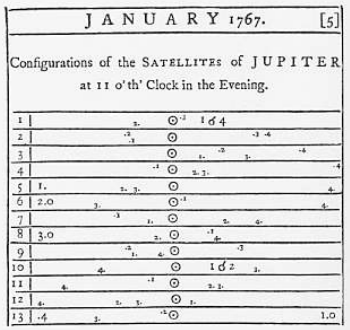
MOEDICEORVM PLANETARVM

Galileo Galilei, *Istoria e dimostrazioni intorno alle macchie solari* . . . [Welser sunspot letters], (Rome, 1613), illustration of satellites (called by Galileo "Medicean stars" in honor of his patron) following p. 150.



October. 1668.
Configurationes Mediceorum.
Hora 10. P.M.

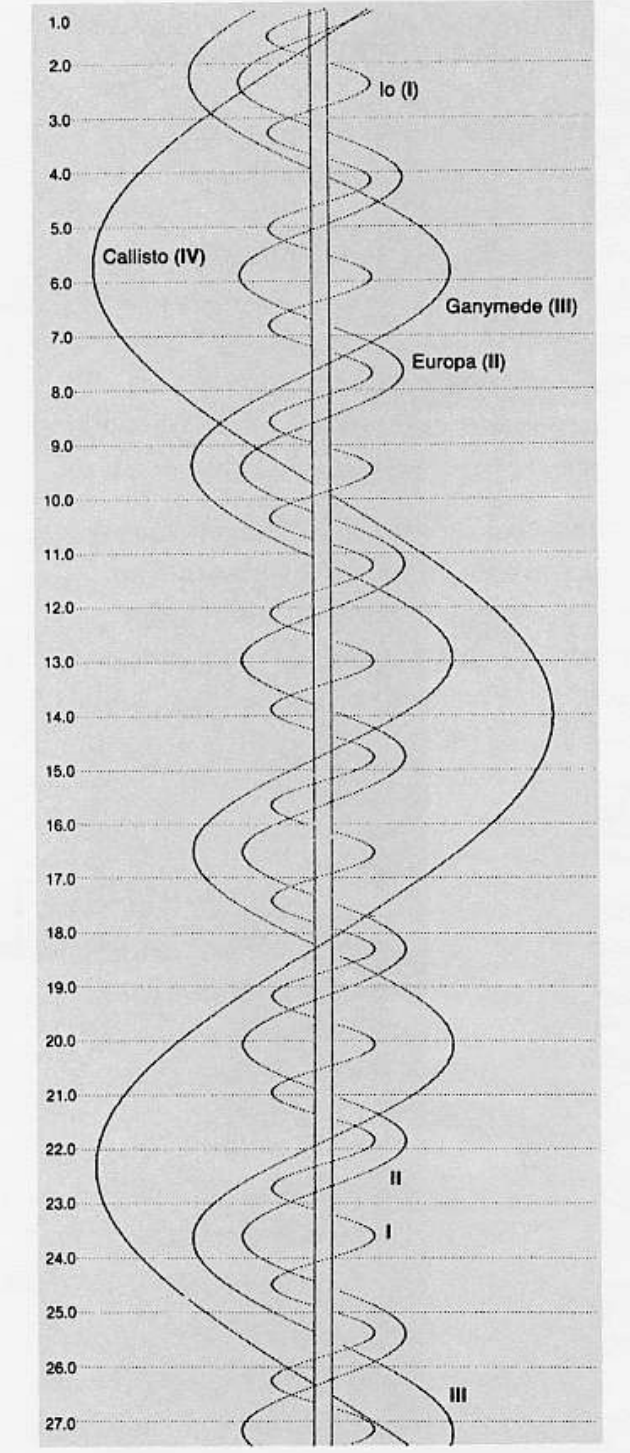
Jean Domenique Cassini, *Ephemerides Bononienses Mediceorum syderum ex hypothesibus, et tabulis Io,* (Bologne, 1668), p. 34.



JANUARY 1767. [5]
Configurations of the SATELLITES of JUPITER at 11 o'th' Clock in the Evening.

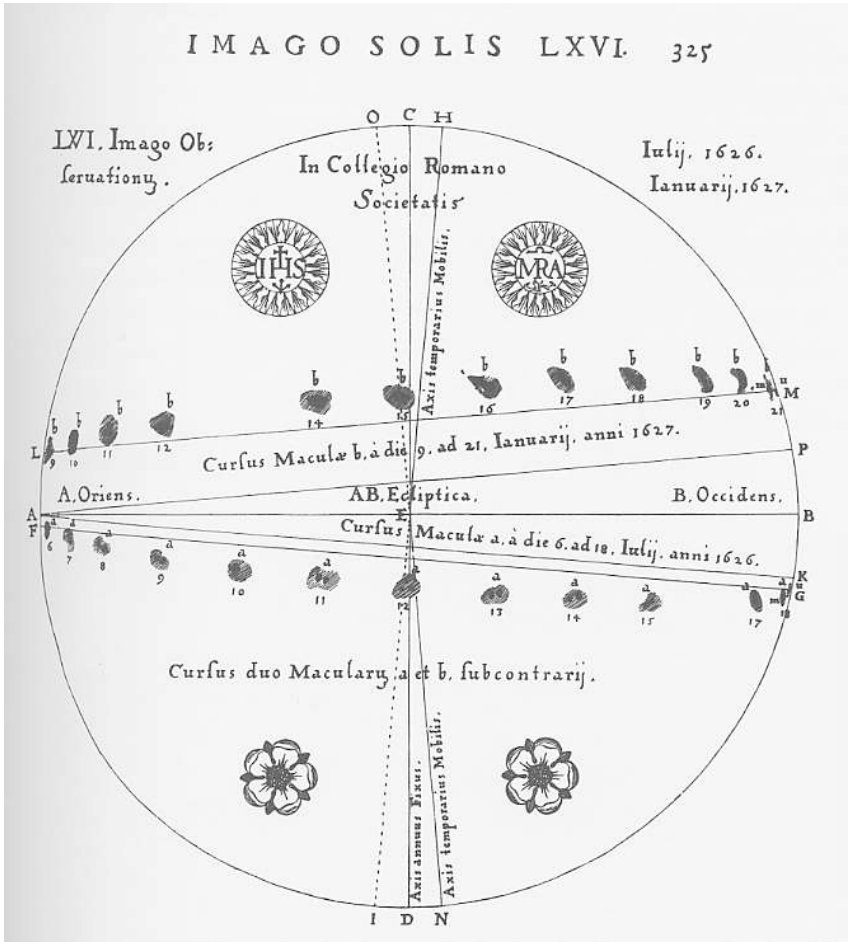
Bureau des Longitudes, *Connaissance des Temps* (Paris, 1766), p. 5.

[2] Translation of *The Starry Messenger* by Stillman Drake, in his *Telescopes, Tides, and Tactics* (Chicago, 1983), pp. 59–63.

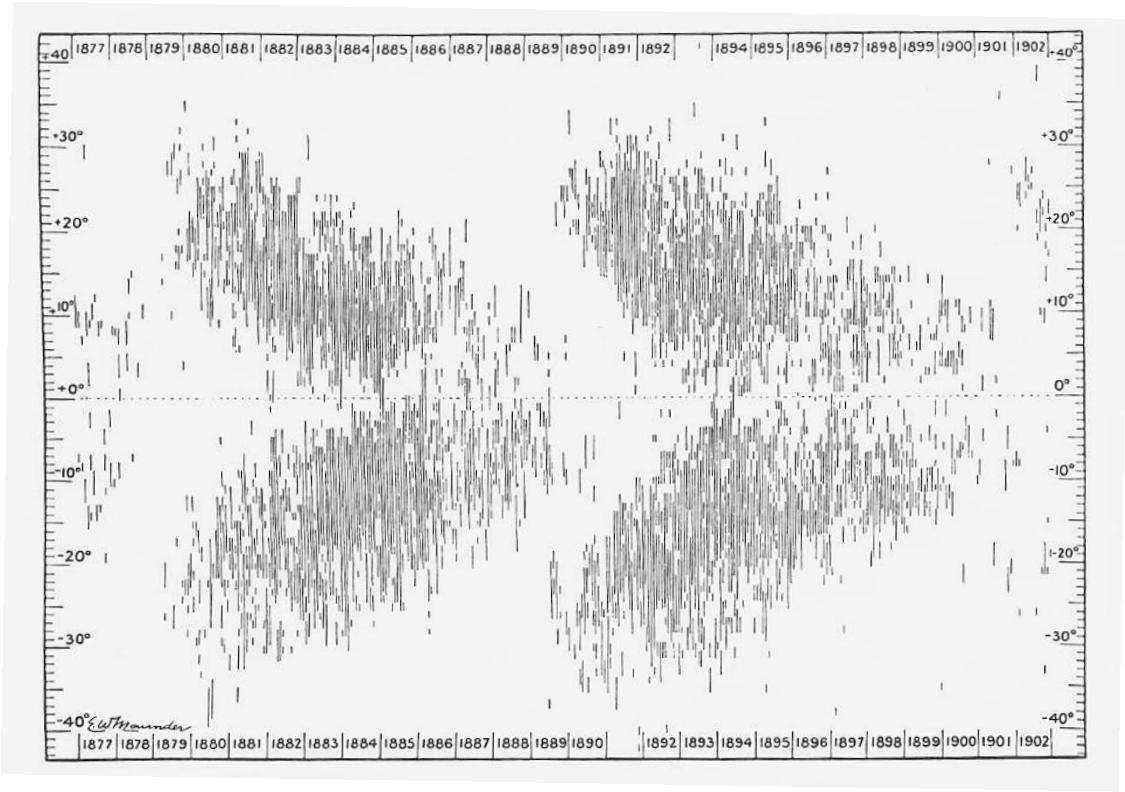


Sky & Telescope, 1988 [El 100].

Sunspots

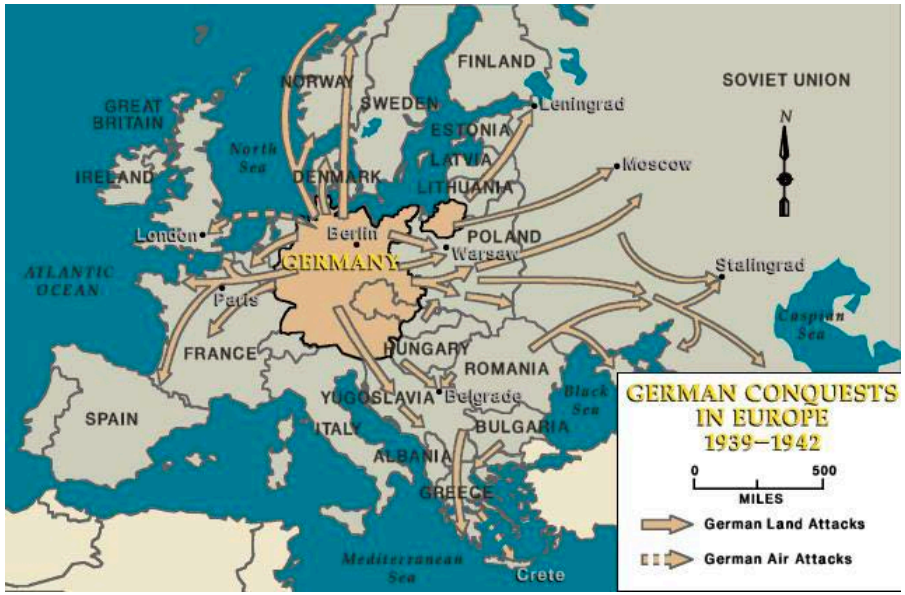


Christopher Schneier, 1630 [EI 21].



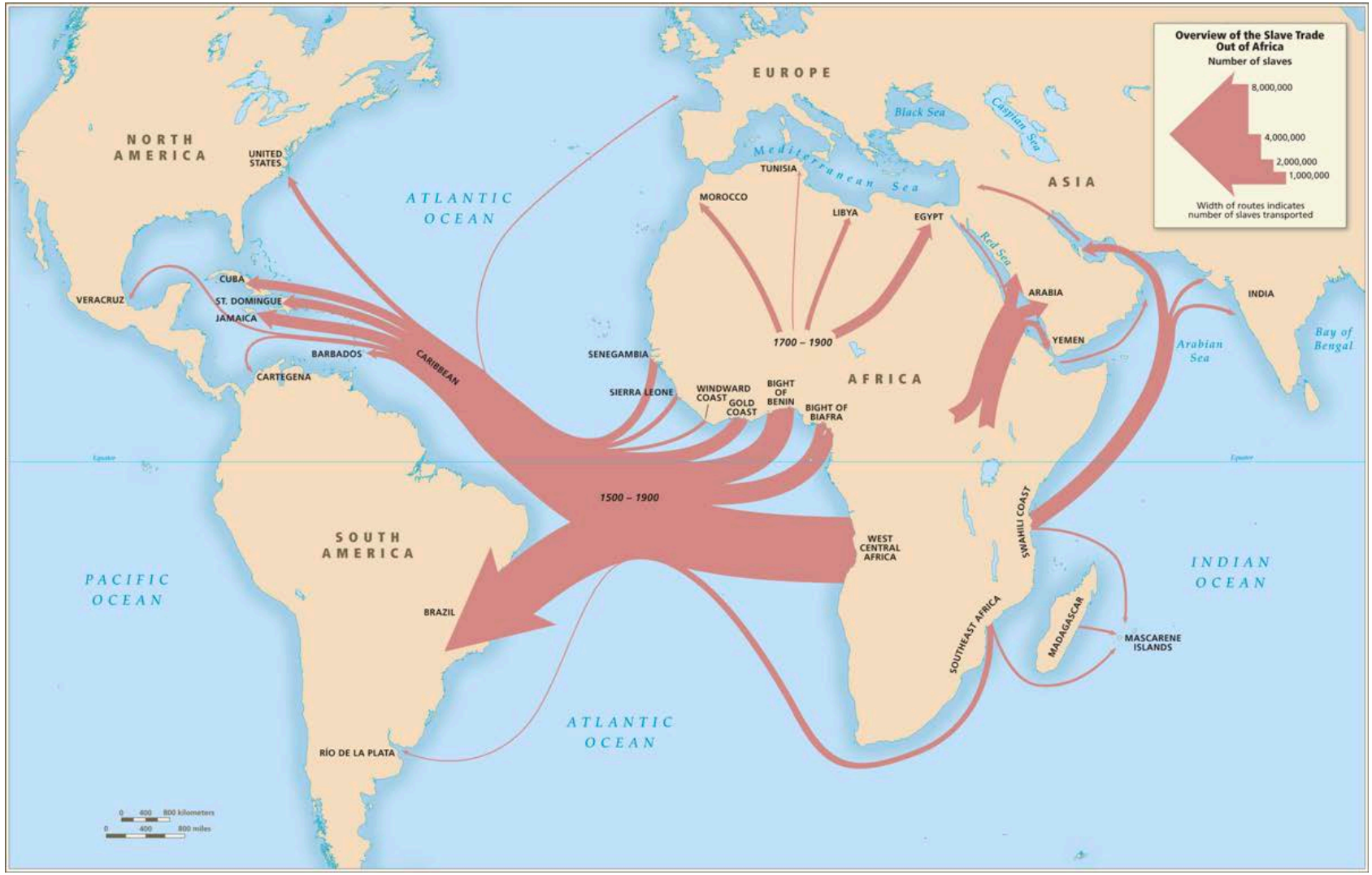
E. W. Maunder, 1904 [EI 22].

Space-time narratives



(from the United States Holocaust Memorial Museum)

Space-time narratives



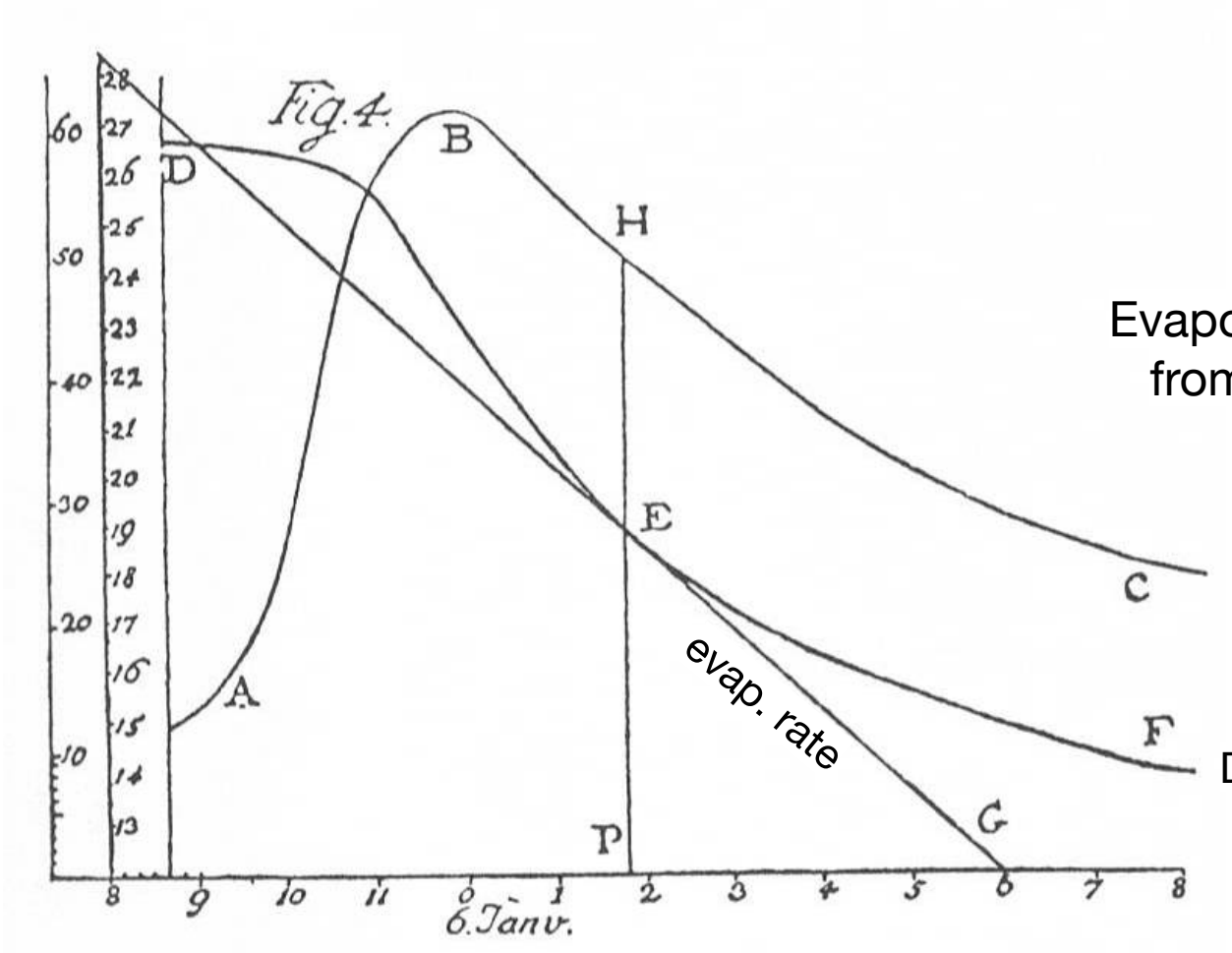
un.org

David Eltis and David Richardson, Atlas of the Transatlantic Slave Trade (New Haven, 2010).

Relational graphics

- The invention of data graphics required replacing the coordinates of the map with more abstracts measures, not based on geographical analogy
    - moving from maps and time series to fully abstract statistical graphics was a **big** step
    - thousands of years passed before this step was taken
    - Lambert, Playfair, and others in the 18th century

From overlapping time series...



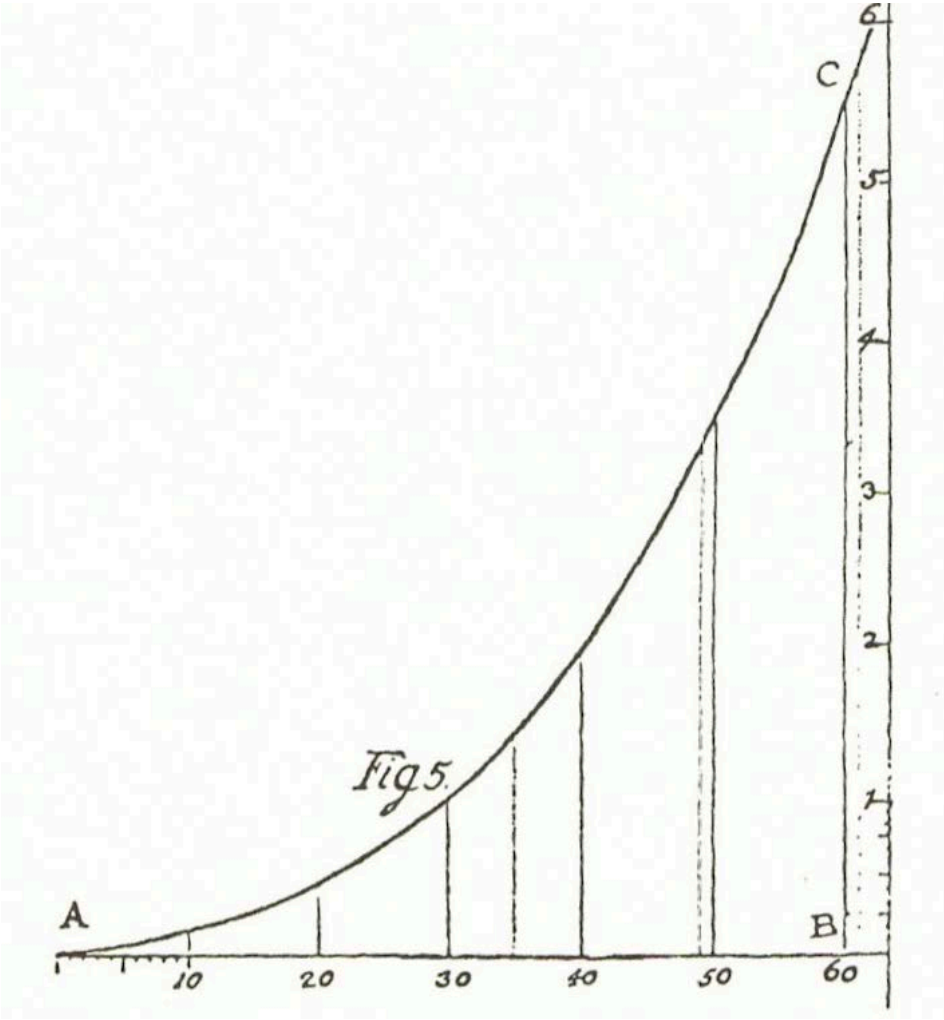
Evaporation of water from a container:

ABC = temperature

DEF = water level

J. H. Lambert, Essai d'hygrométrie ou sur la mesure de l'humidité, Mémoires de l'Académie Royale des Science et Belles-Lettres, 1769.

...to relational graphics



X = temperature

Y = measured rate

Graphical excellence

- In summary, graphical excellence is the well-designed presentation of interesting data
    - it is a matter of substance, of statistics, and of design
- Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency or, it should give to the viewer
    - the greatest number of ideas
    - in the shortest time
    - with the least ink
    - in the smallest space

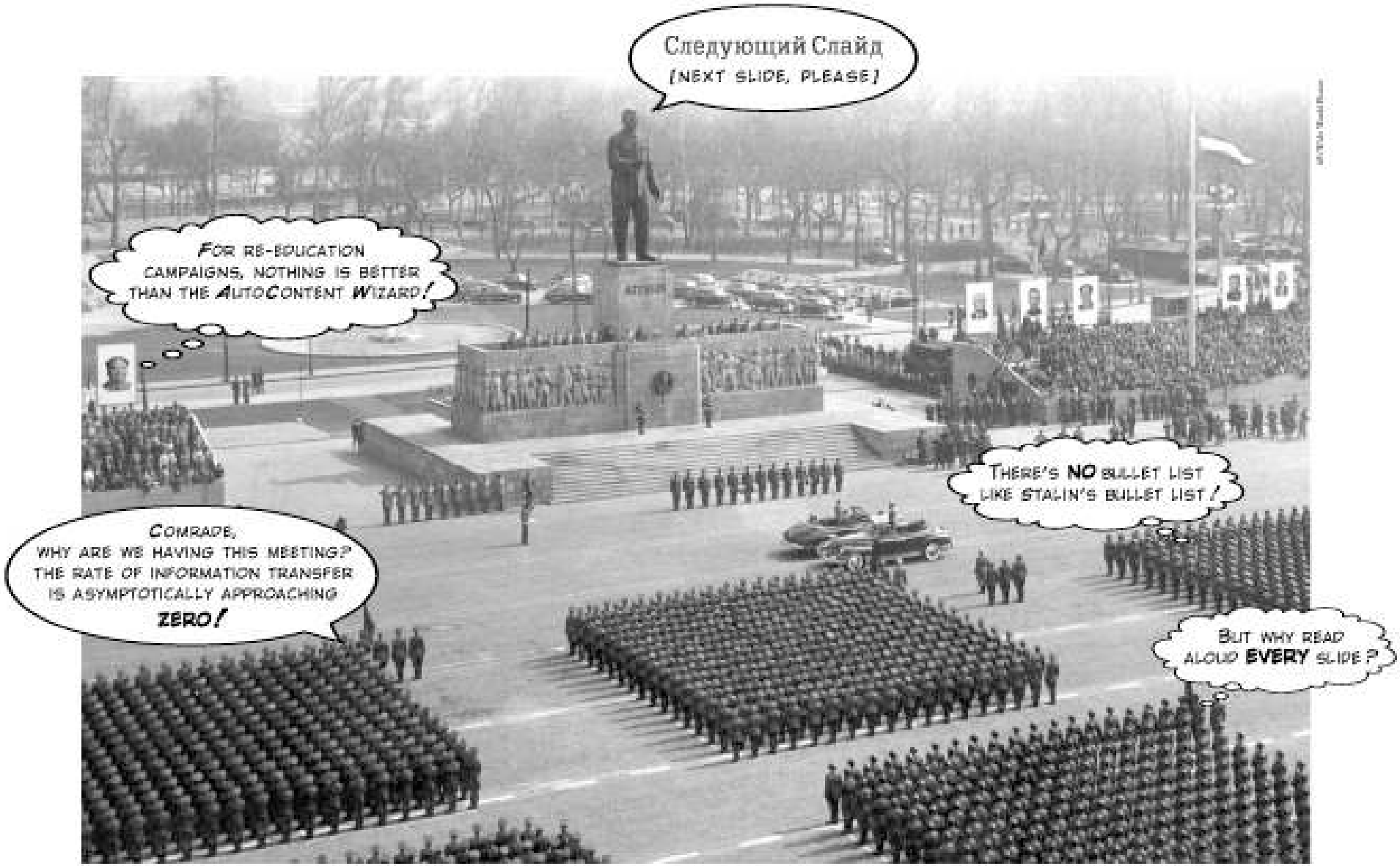
Today

- examples of bad graphics
- a theory of good graphics

Visualisations in decision-making

- John Snow's discovery of cholera was an example of a successful use of visualisations as part of decision making
- What if things don't go well?

…popular PowerPoint templates (ready-made designs) usually weaken verbal and spatial reasoning, and almost always corrupt statistical analysis.



Edward Tufte, *The Cognitive Style of PowerPoint*

https://www.edwardtufte.com/tufte/powerpoint

Tufte:

"As a consumer of presentations, you should not trust speakers who rely on the PP cognitive style. It is likely that these speakers are simply serving up PowerPointPhluff to mask their lousy content, just as this massive tendentious pedestal in Budapest once served up Stalin-cult propaganda to orderly followers feigning attention."

Space shuttle Columbia disaster in 2003

- A piece of insulating foam from the external fuel tank damaged the thermal protection system of the left wing during the launch.
- During re-entry the breach to the thermal protection system allowed superheated air to penetrate the left wing, eventually destroying the orbiter.

How the problem was reported:

Orbiter Assessment of STS-107 ET Bipod Insulation Ramp Impact

P. Parker

D. Chao

I. Norman

M. Dunham

January 23, 2003

BOEING

https://www.nasa.gov/pdf/2203main_COL_debris_boeing_030123.pdf

Meaning: "The review of test data shows that the models cannot be used to reliably analyse tile damage"

Review of Test Data Indicates Conservatism for Tile Penetration

Unexplained abbreviation

- **The existing SOFI on tile test data used to create Crater was reviewed along with STS-87 Southwest Research data**
  - **Crater overpredicted penetration of tile coating significantly**
    - **Initial penetration to described by normal velocity**
      - Varies with volume/mass of projectile (e.g., 200ft/sec for 3cu. In)
    - **Significant energy is required for the softer SOFI particle to penetrate the relatively hard tile coating**
      - Test results do show that it is possible at sufficient mass and velocity
    - **Conversely, once tile is penetrated SOFI can cause significant damage**
      - Minor variations in total energy (above penetration level) can cause significant tile damage
  - **Flight condition is significantly outside of test database**
    - **Volume of ramp is 1920cu in vs 3 cu in for test**

significantly = modelling error

significant = large amount of energy

It is finally revealed at the bottom that the models are used outside their area of validity

significant = everyone dies

6 levels of hierarchy!

significantly = we are outside the validity area of the model

Better?

Foam strike may have caused critical damage to Columbia

- We reviewed the data used to create the *Crater damage model*, used to model the spray-on foam insulation (SOFI) collisions with tiles
  - *Crater* overpredicted SOFI penetration of tile coating
  - It is possible that the SOFI piece did not penetrate Columbia's tiles
- However, STS-107 flight condition was far outside of test database!
  - The volume of the SOFI pieces that struck Columbia was 640 times larger than what has been tested (1920 in^3 vs. 3 in^3)
- The pieces that hit Columbia may have had enough energy for the soft SOFI to penetrate the relatively hard tile coating
  - Test results show tile penetration is possible when SOFI impacts with sufficient mass and velocity
- SOFI can cause critical damage if it penetrates the tile!

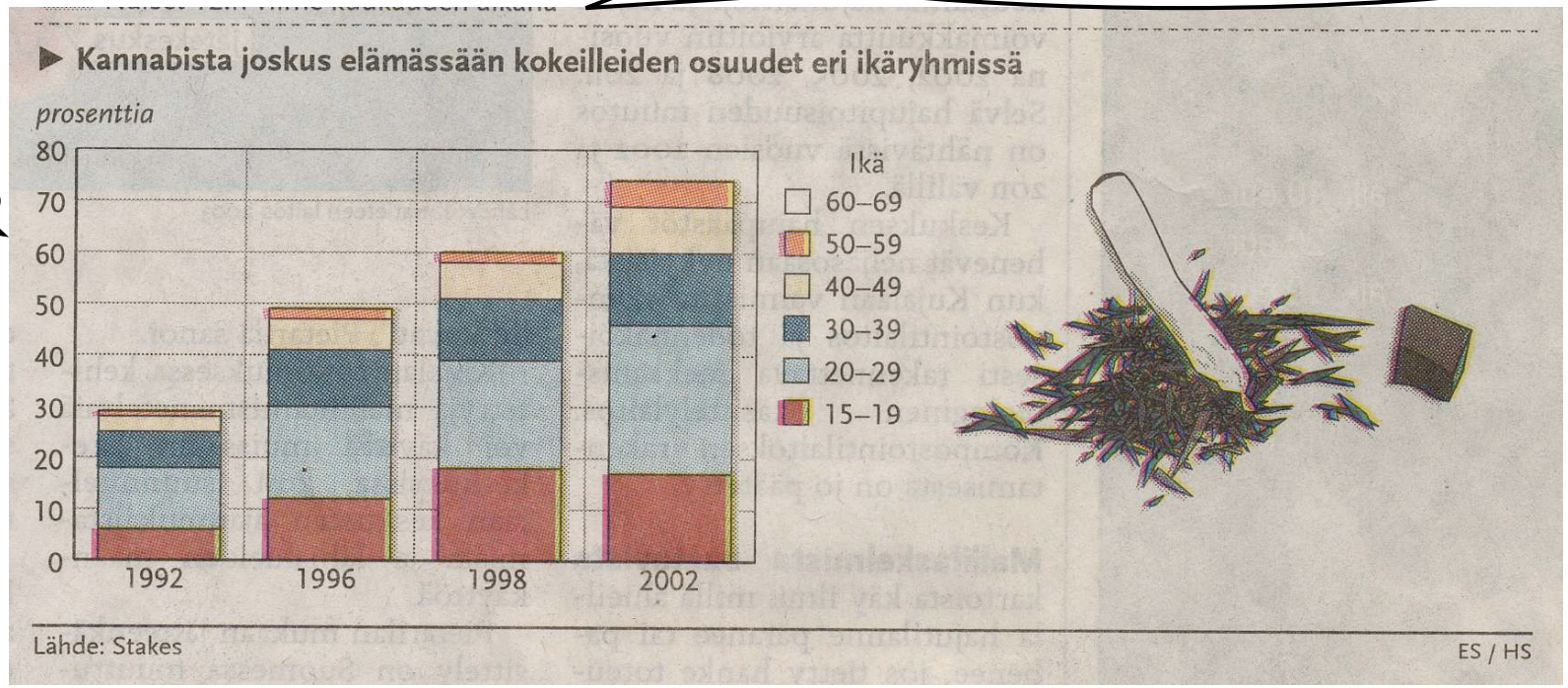
Bottom line: the foam strike is a safety-of-flight issue!

Telling the truth about the data

- It is easy to make misleading visualizations by purpose or by mistake
- A visualization can be misleading, even if it is "technically correct"

The fractions of people having tried cannabis at some point of their lives in various age groups

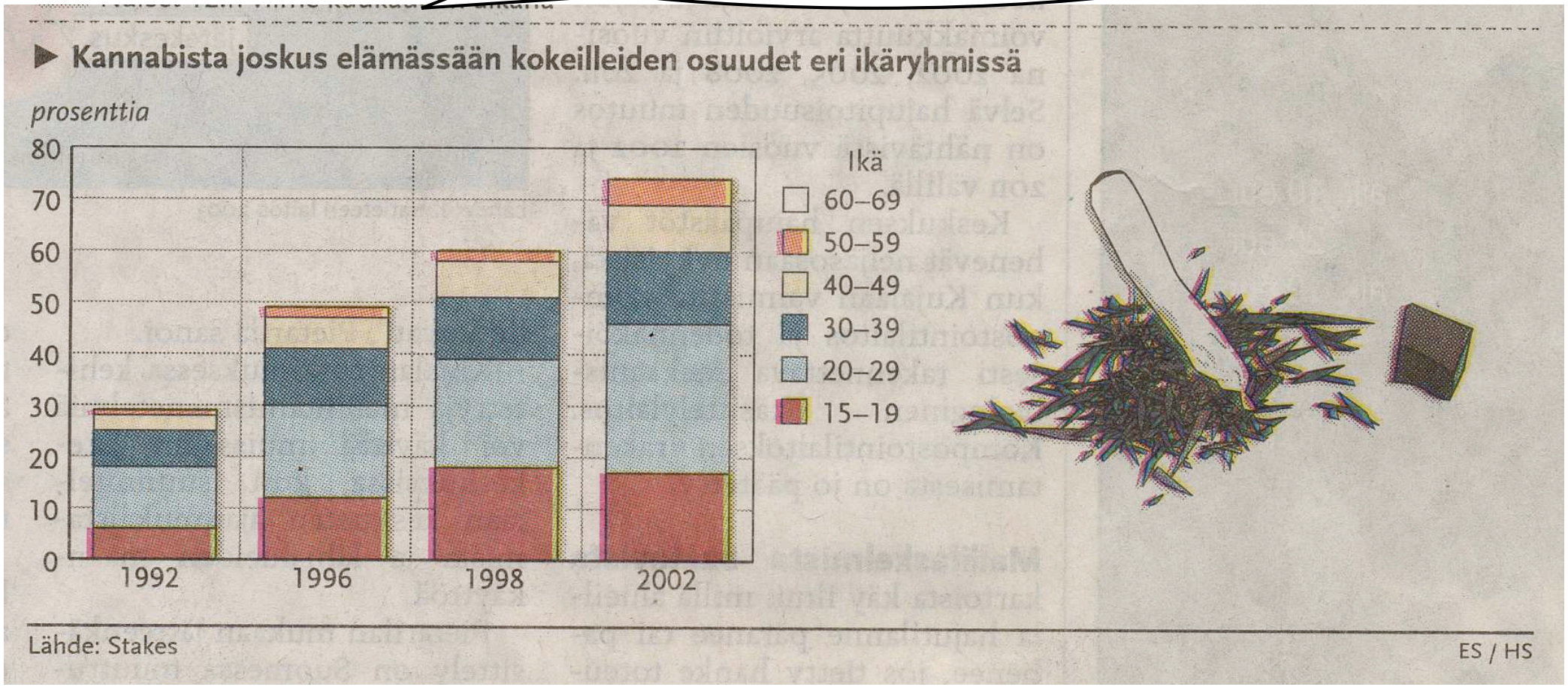
Percent



Have 75% of the Finns tried cannabis?

Helsingin Sanomat, 9 May 2003.

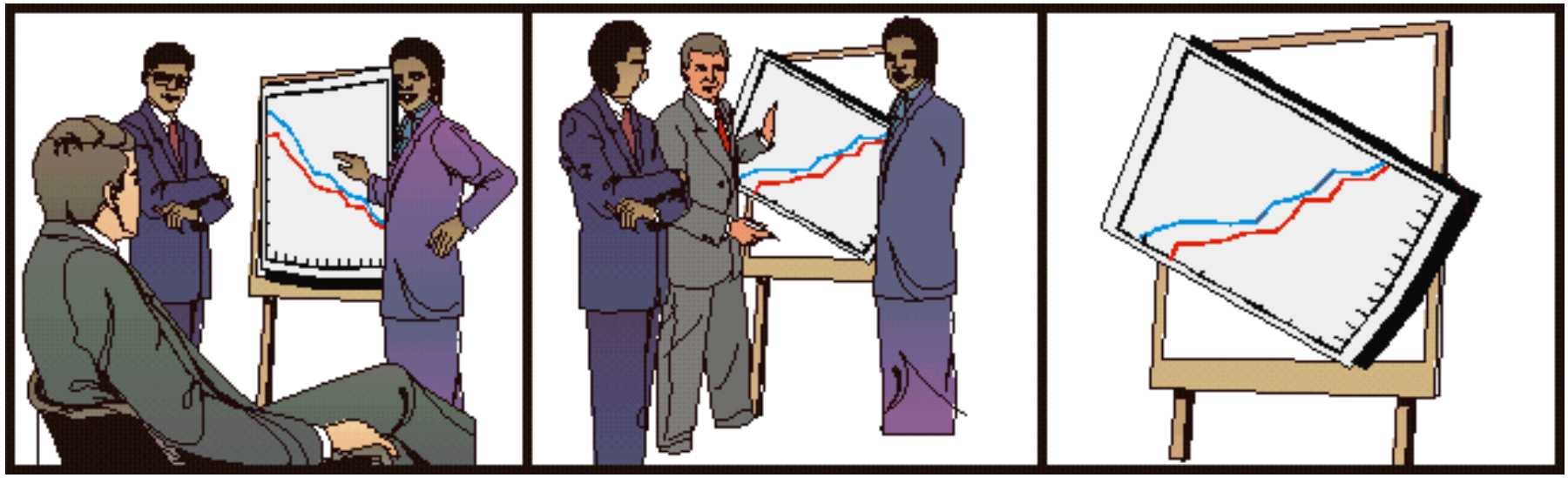
The fractions of people having tried cannabis at some point of their lives in various age groups



Helsingin Sanomat, 9 May 2003.

Lying with graphics

- It is easy to lie using visualisation
- It is important to know how visual quantities are perceived (even "technically correct" visualisation can be misleading)



This and some of the following illustrations by H. Ingo

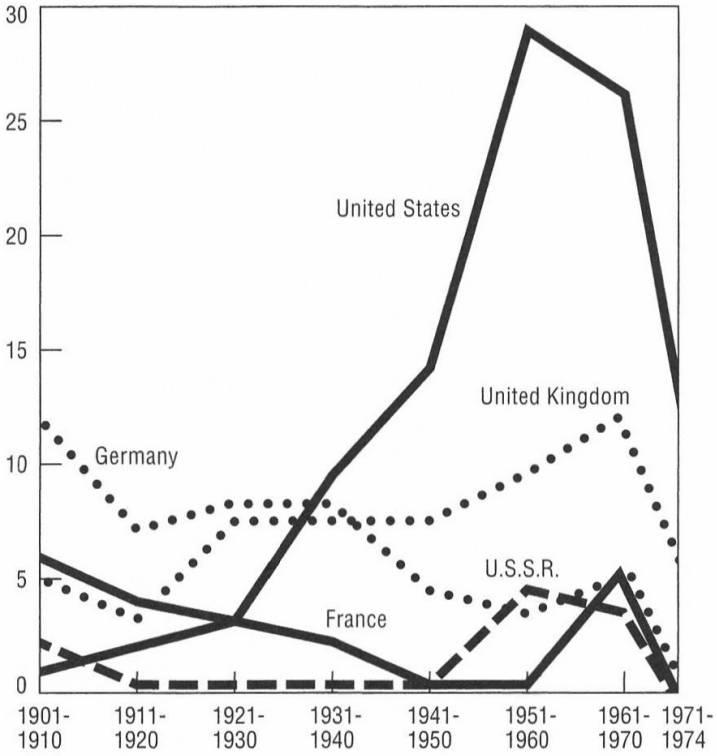
Graphical integrity

- Much of the (first half of) 20th century focused on the question of how charts might fool a viewer
    - the use of graphics for serious data analysis was largely ignored
- Data graphics were meant only for showing the **obvious** to the **ignorant**, which led to two fruitless paths
    - the graphics had to be alive, communicatively dynamic, overdecorated, and exaggerated (otherwise, the dullards would fall asleep)
    - the main task of graphical analysis was to detect and denounce deception (because the dullards could not protect themselves)

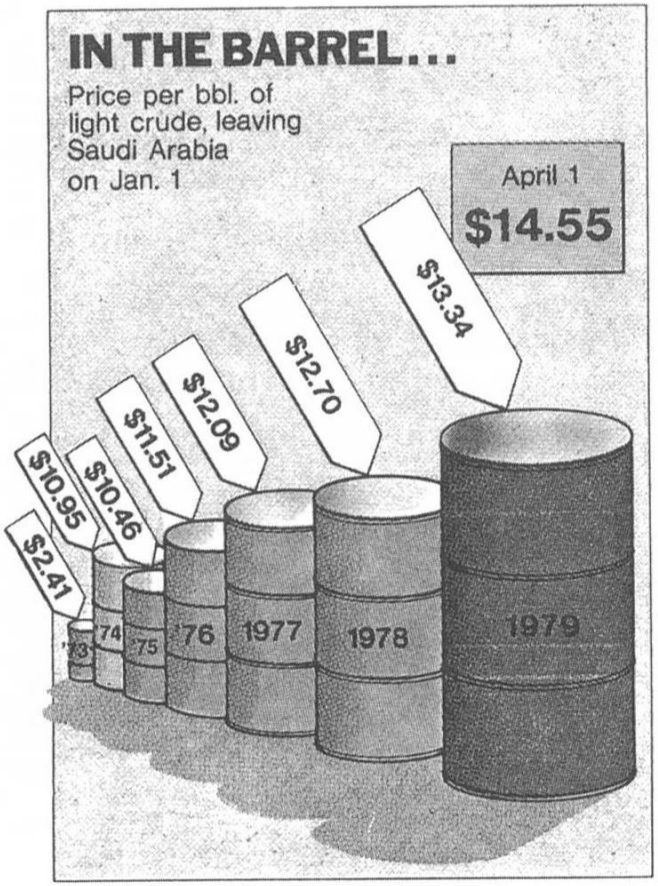
Graphical integrity

- John **Tukey** (1915-2000) started making statistical charts respectable in 1960s, putting an end to the view that graphics were only for decorating numbers
    - American mathematician and a world-class data analyst
    - new designs and their effective use in the exploration of complex data
    - forerunner of interactive graphics
    - invented box plot
    - invented the word "bit"
- focus how can we use graphics as instruments for reasoning
- less focus on deceptive graphics

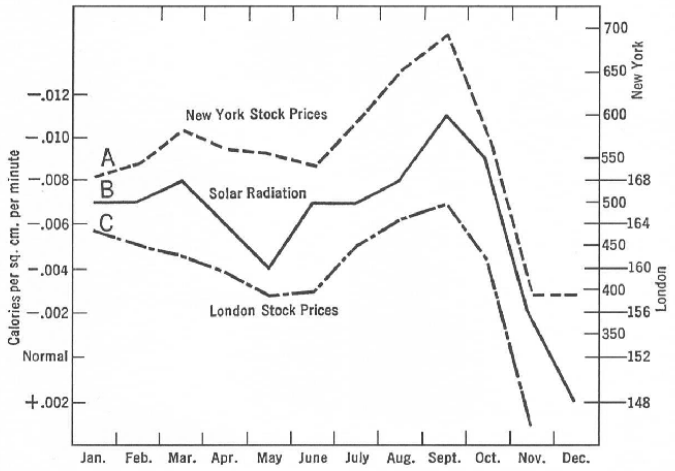
(a)



(b)



(c)



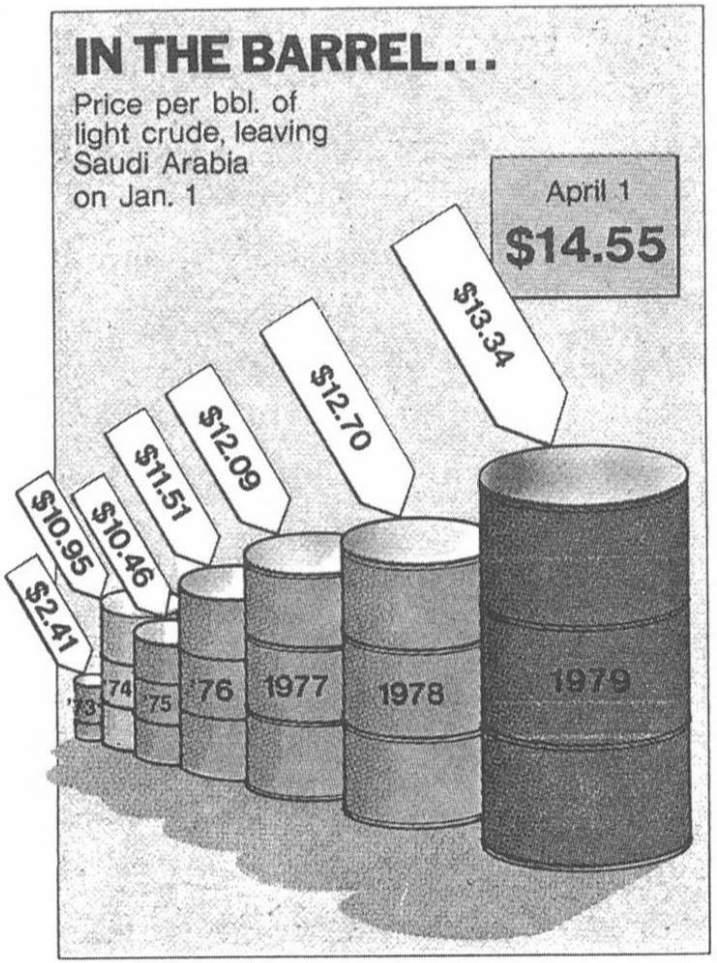
SOLAR RADIATION AND STOCK PRICES

A. New York stock prices (Barron's average). B. Solar Radiation, inverted, and C. London stock prices, all by months, 1929 (after Garcia-Mata and Shaffner).

Discuss and try to figure out what is wrong in each graph.

Distortion in data graphics

- A graphic does not distort if the visual and numerical representations are consistent
- What then is the visual representation of the data?
  - as physically **measured** on the surface of the graphics?
  - or, as the **perceived** visual effect?
- How do we know that the visual image really represents the underlying numbers?



Distortion in data graphics

- Perception of graphics varies: we can only hope for
  - presenting graphics in a consistent manner, and
  - assurance that the perceiver has a fair chance to get the numbers right
- Two principles lead towards those goals:
  1. if numbers are represented by visual elements then they should be directly proportional to perceived magnitude of the visual elements
  2. use clear, detailed and thorough labelling to defeat graphical distortion and ambiguity
     - write out explanations, and label important events in the data

Lie factor

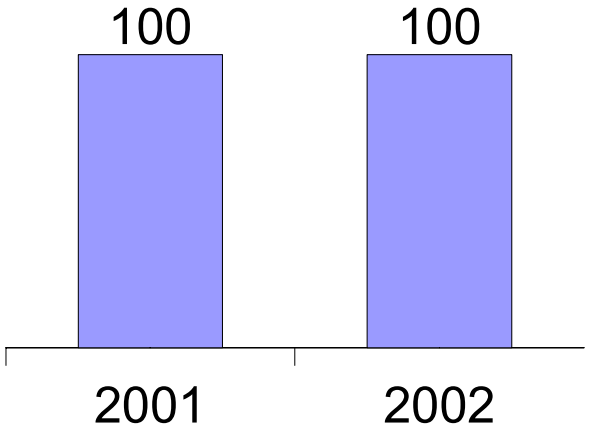
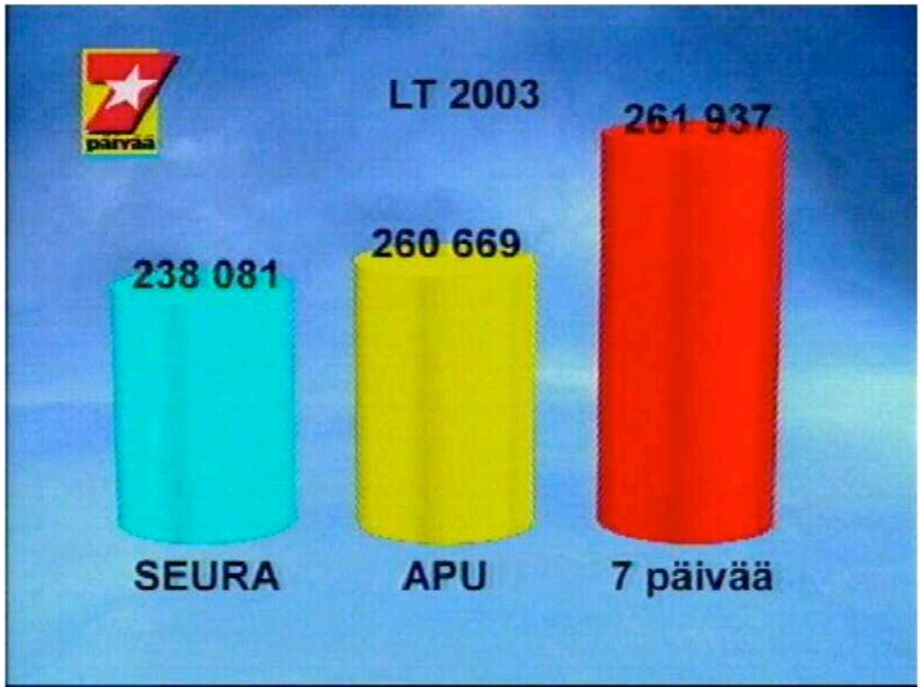


Chart 1: Truthfull representation of data

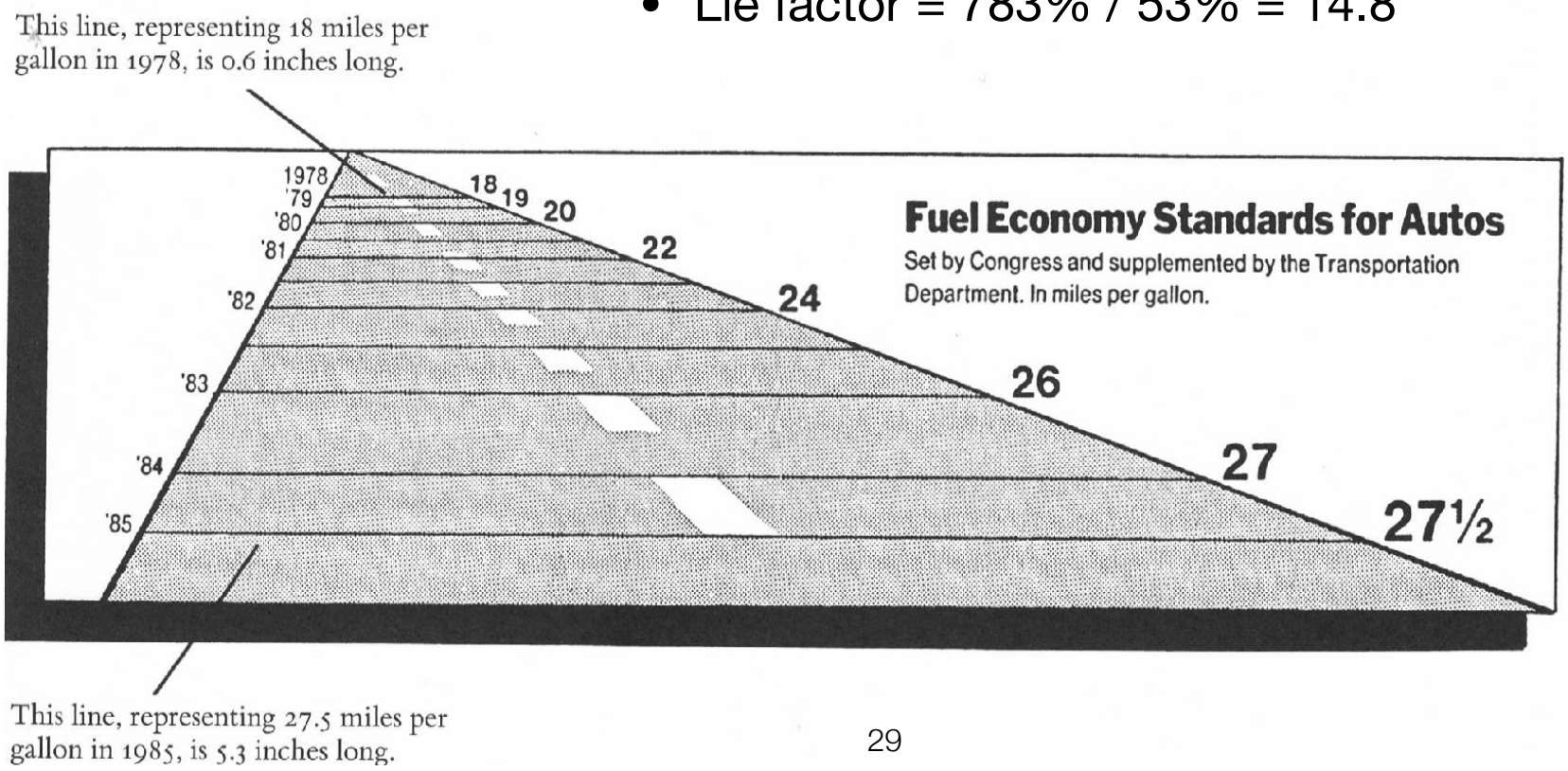


Circulation of periodicals 2003

$$\text{lie factor} = \frac{\text{size of effect shown in graphic}}{\text{actual effect in data}}$$

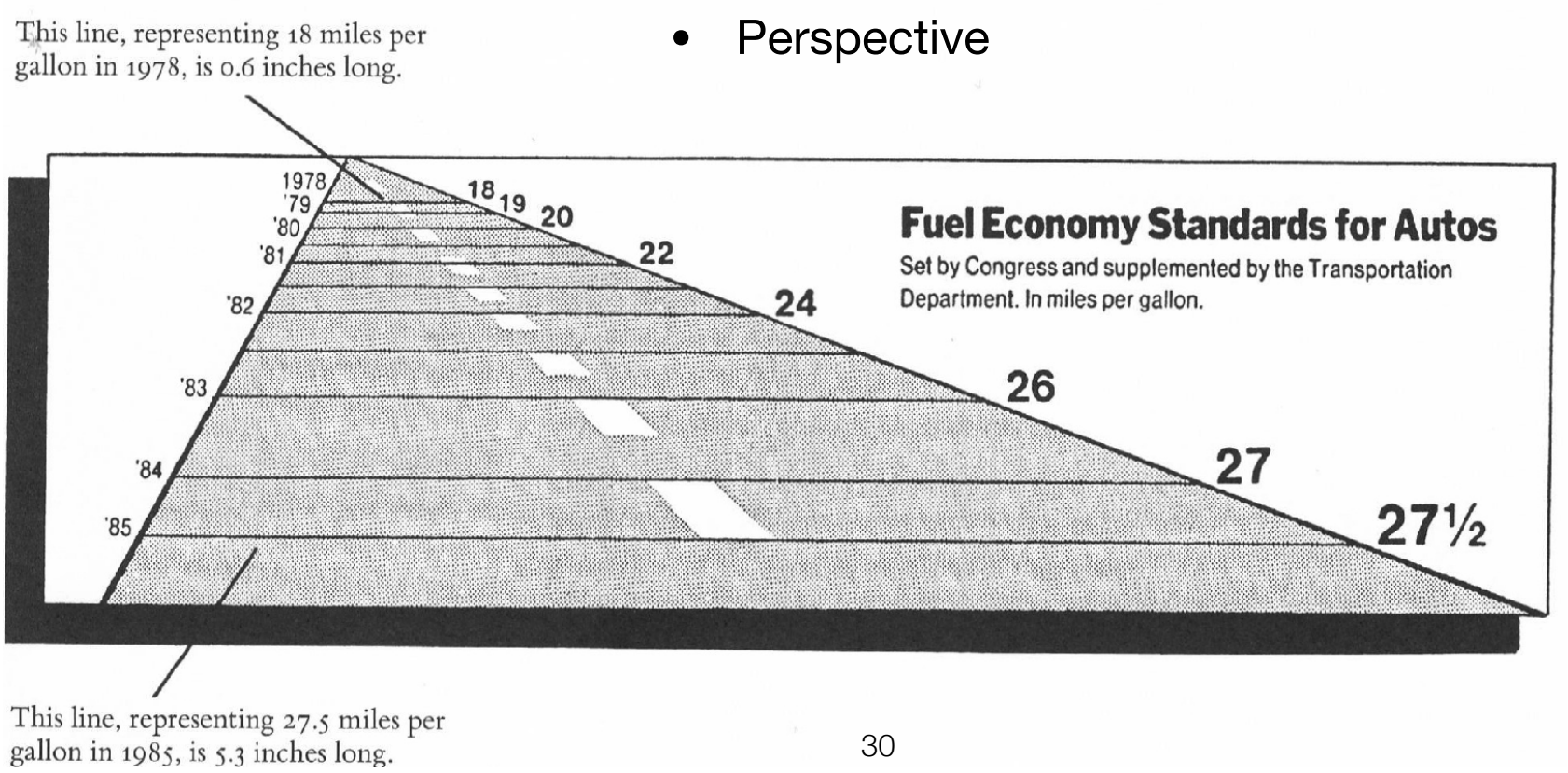
Example: fuel economy standards (1978)

- from 18 mpg (1978) to 27.5 mpg (1985)
- (27.5-18)/18 = 53% increase in **data**
- (5.3-0.6)/0.6 = 783% increase in (perceived?) **length**
- Lie factor = 783% / 53% = 14.8



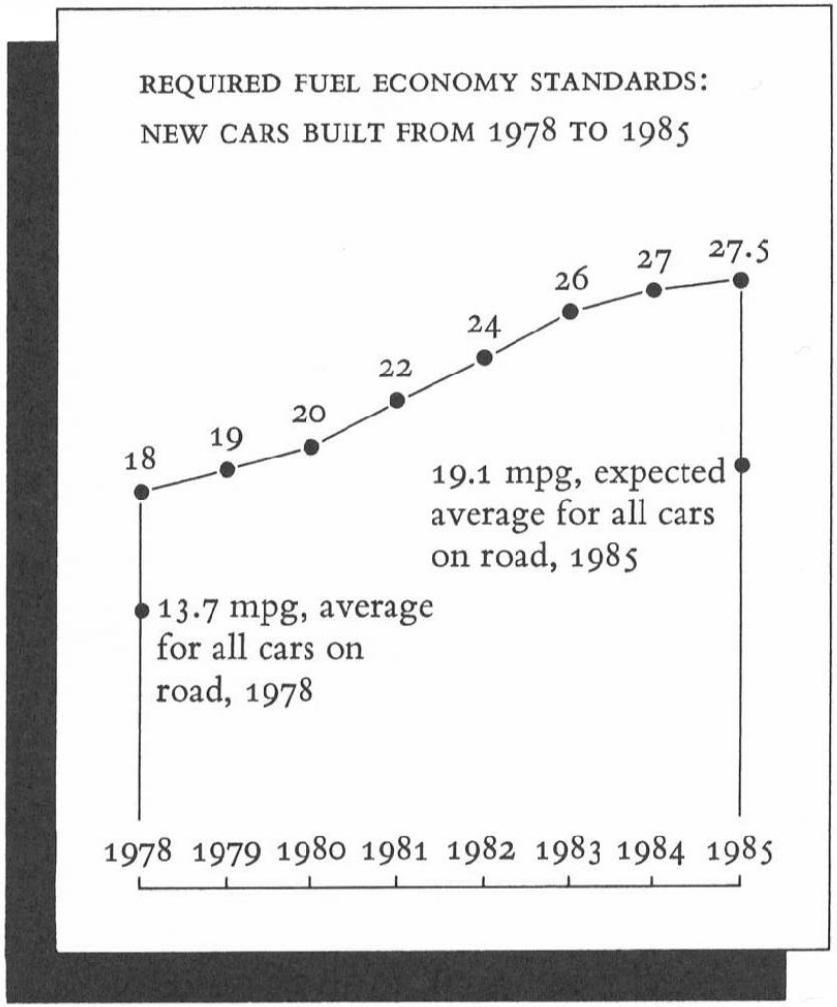
Example: fuel economy

- Peculiarities:
  - Future is in the front (not in the horizon)
  - The dates remain constant size
- The numbers and the road shrinking because of
  - Change in values
  - Perspective

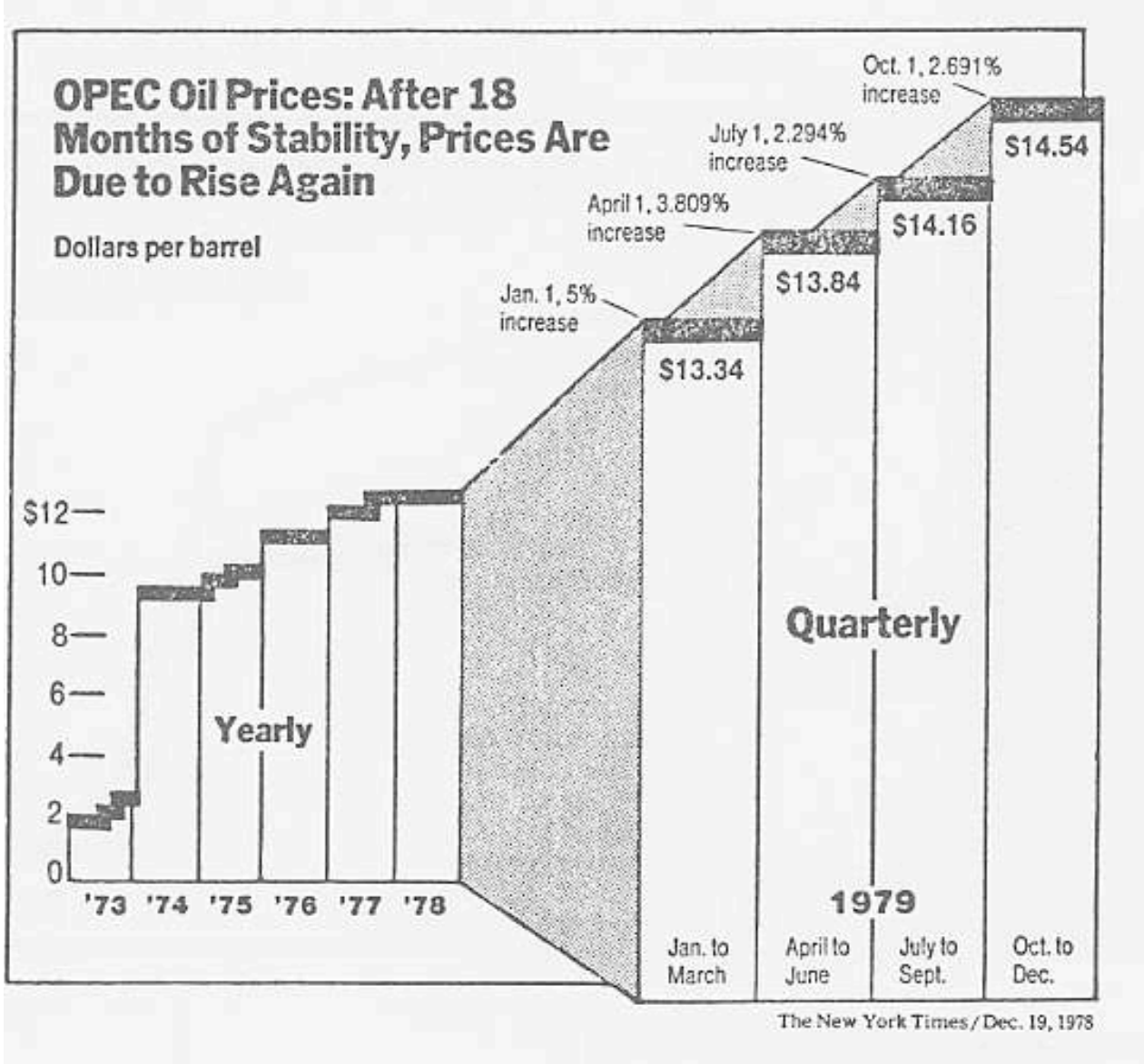


Example: fuel economy

- The non-lying version, with proper context
  - new cars standards compared with average cars on the road
- plot reveals dynamics of fuel economy
  - slow startup
  - fast growth
  - final stabilization



Another example



Yet another trick: manipulating the axis

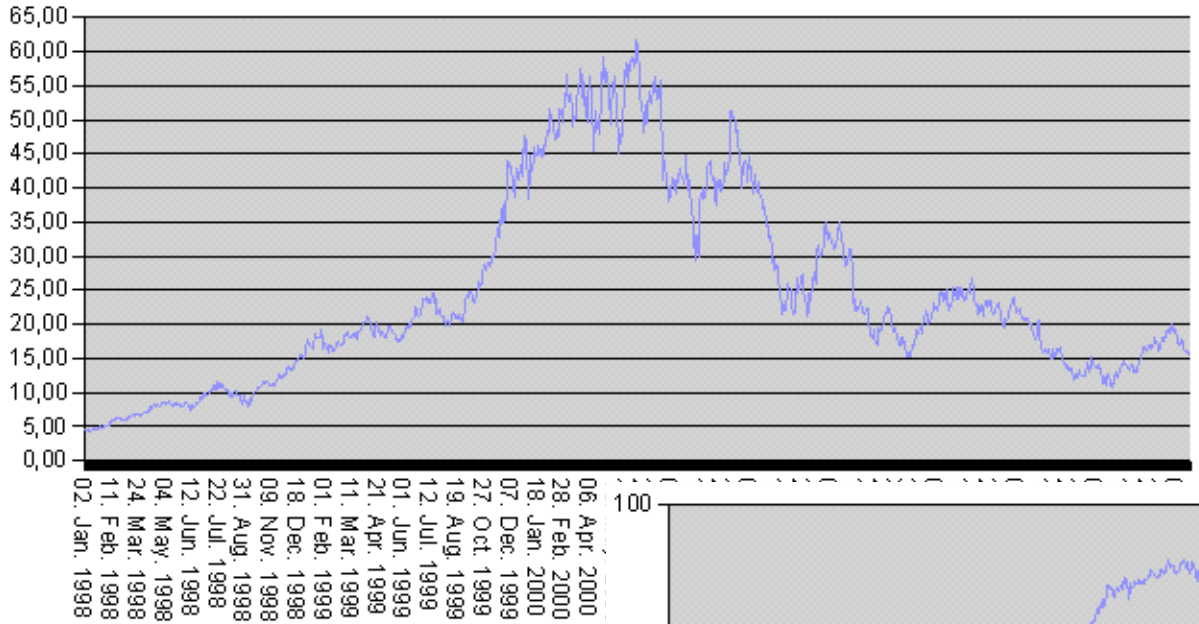


Chart 9: 5 year history of

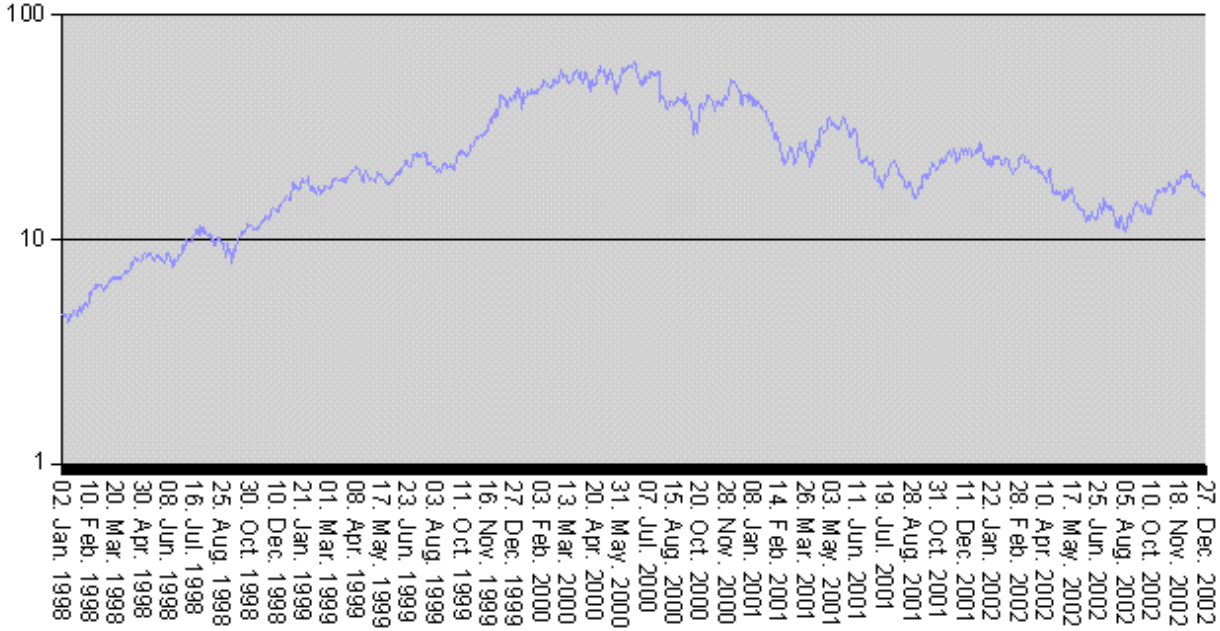
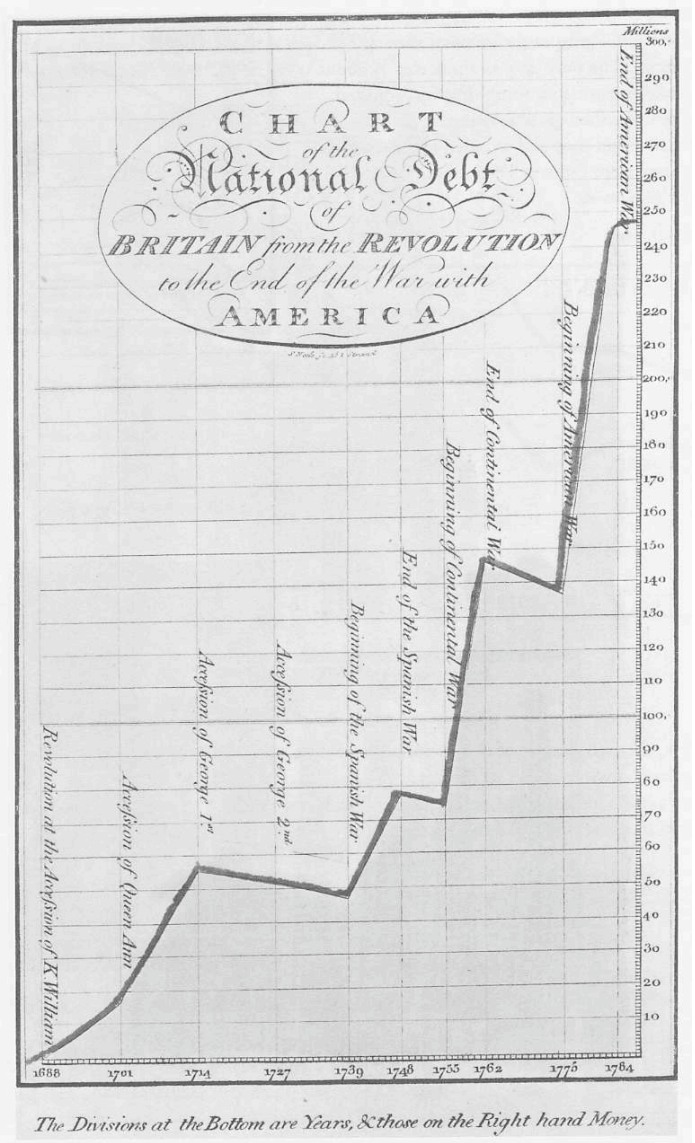


 Chart 10: 5 year history of Nokias stock, log10 curve

Axis trick 2: inflation

Playfair was first to make a plot on government spending without taking the inflation into account. The impression of skyrocketing government spendings was increased by the fact that the graph is taller than it is wide.



William Playfair, 1786 [T 64]

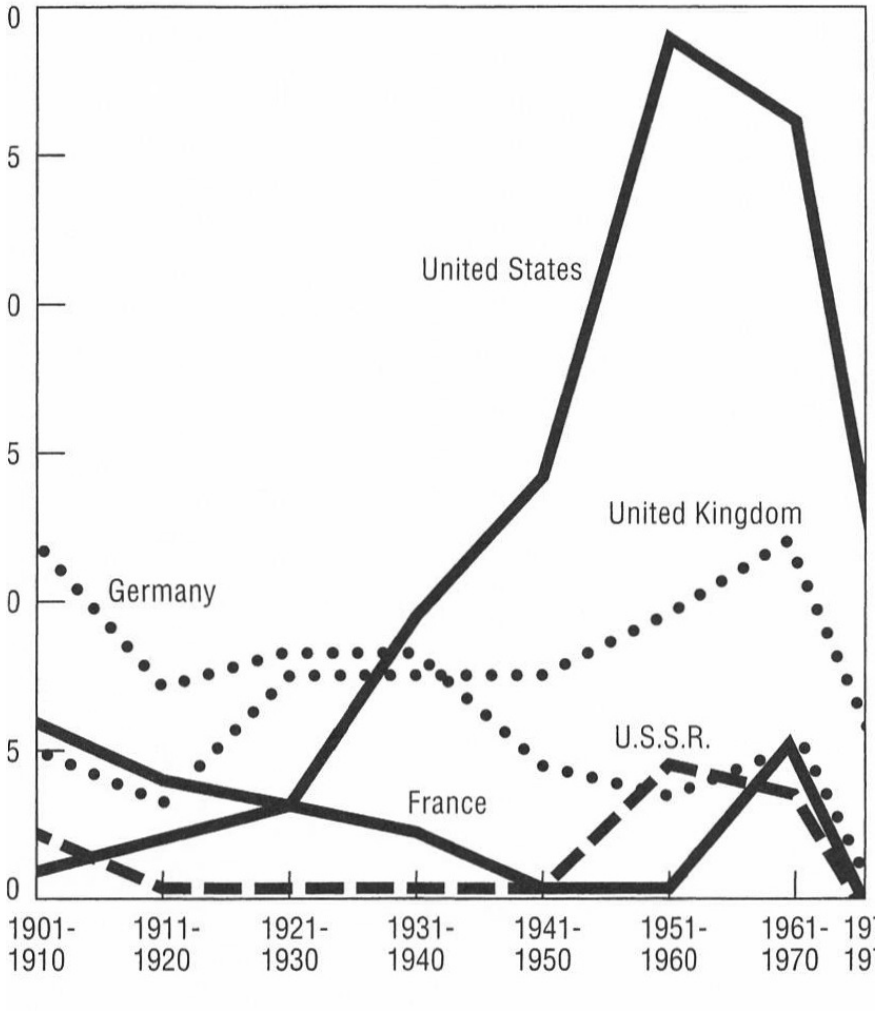
Design and data variation

- Each part of a graphic generates **visual expectations** about its other parts
  - these expectations often determine what the eyes actually see
  - incorrect extrapolation of visual expectations, generated at one place on the graphics, deceives in other places
- A typical example is a scale moving at regular intervals, ... we expect it to move to the very end in a coherent fashion
- Let’s see what happens with the muddling or trickery of non-uniform changes...

Example: Nobel prizes

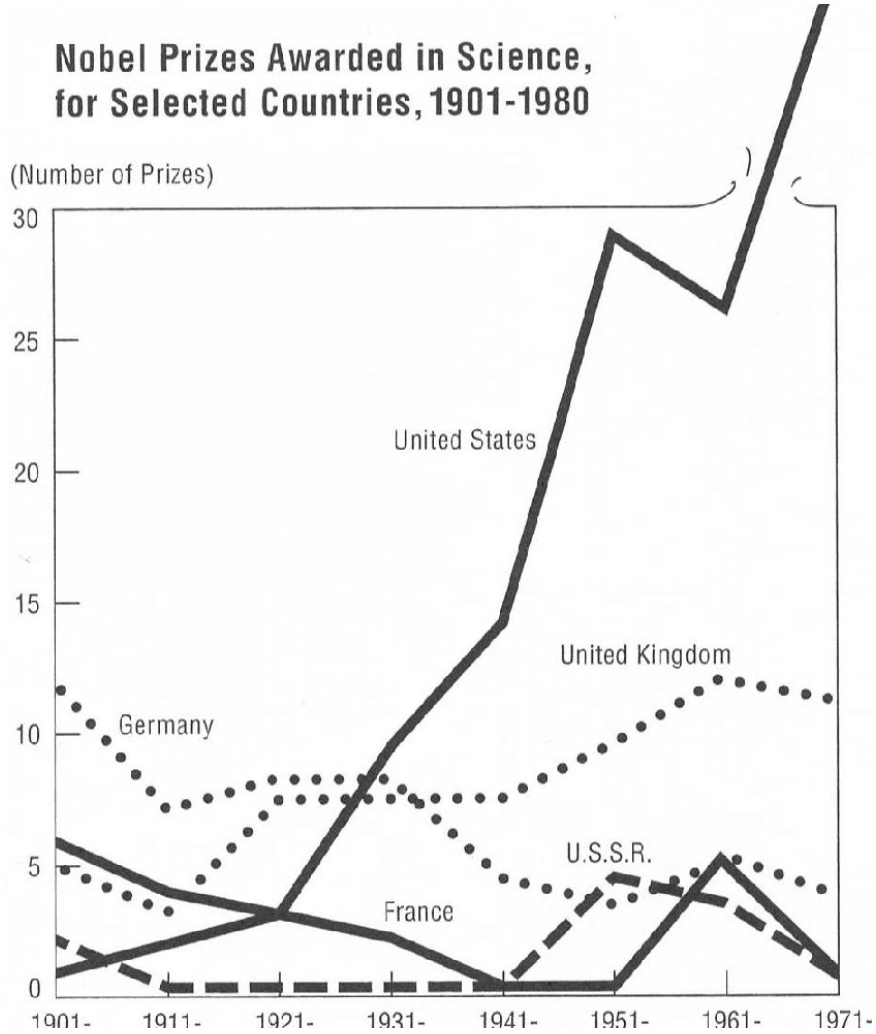
- Irregular scale concocts a pseudo-decline
  - the first seven intervals are 10-yr long
  - the rightmost is only 4-yr long
- As a result, a conspicuous feature of the chart is an apparent fall of all the curves
- The sole effect is design variation!

American National Science Foundation (1976)



Example: Nobel prizes

- corrected figure with the actual data
  - mixing design variation with actual data variation generates ambiguity and deception
  - the eye can mix up changes in the design with changes in the data
- Show data variation, not design variation!



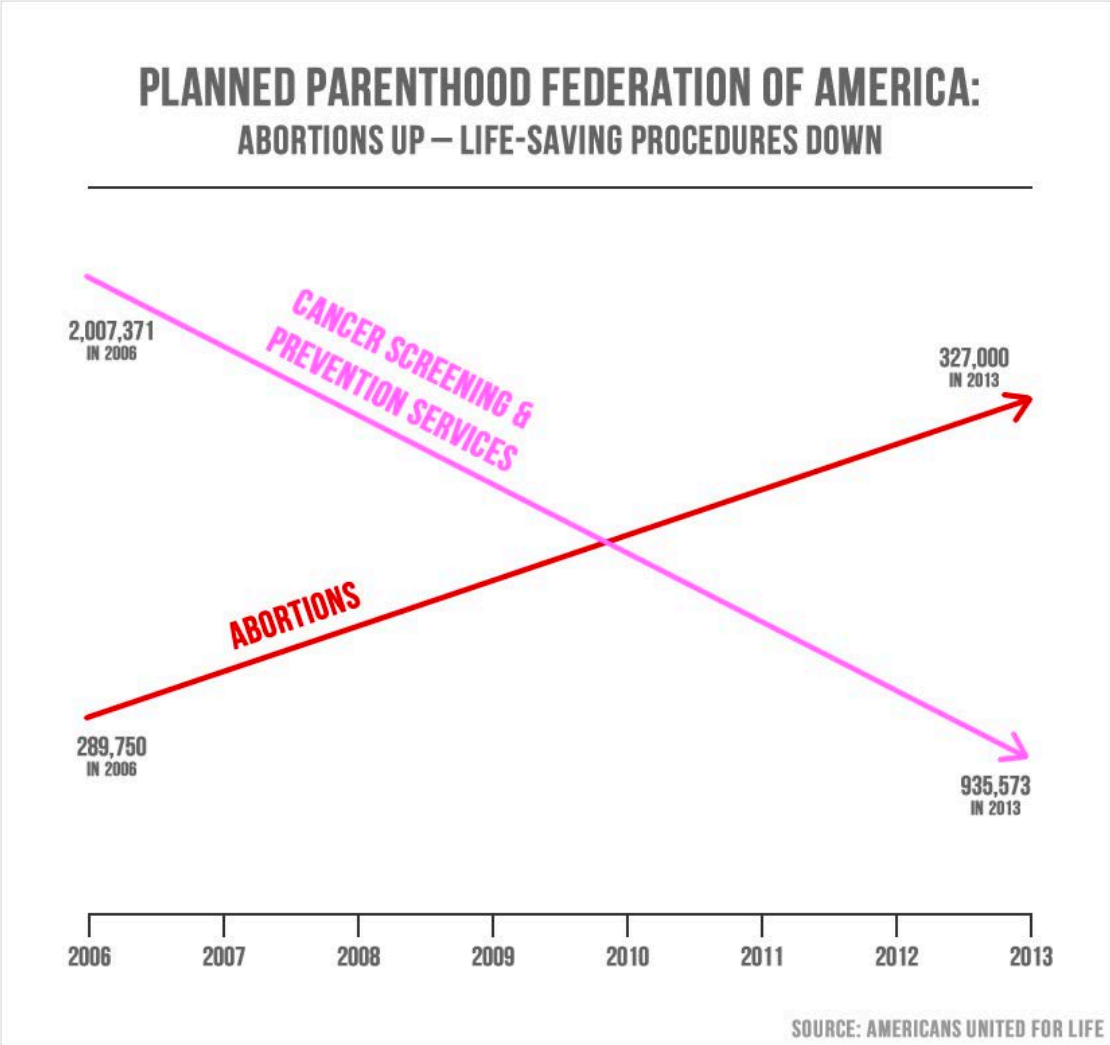
Example: car speedometer

OK

What is wrong with the scale?

Example: Planned Parenthood

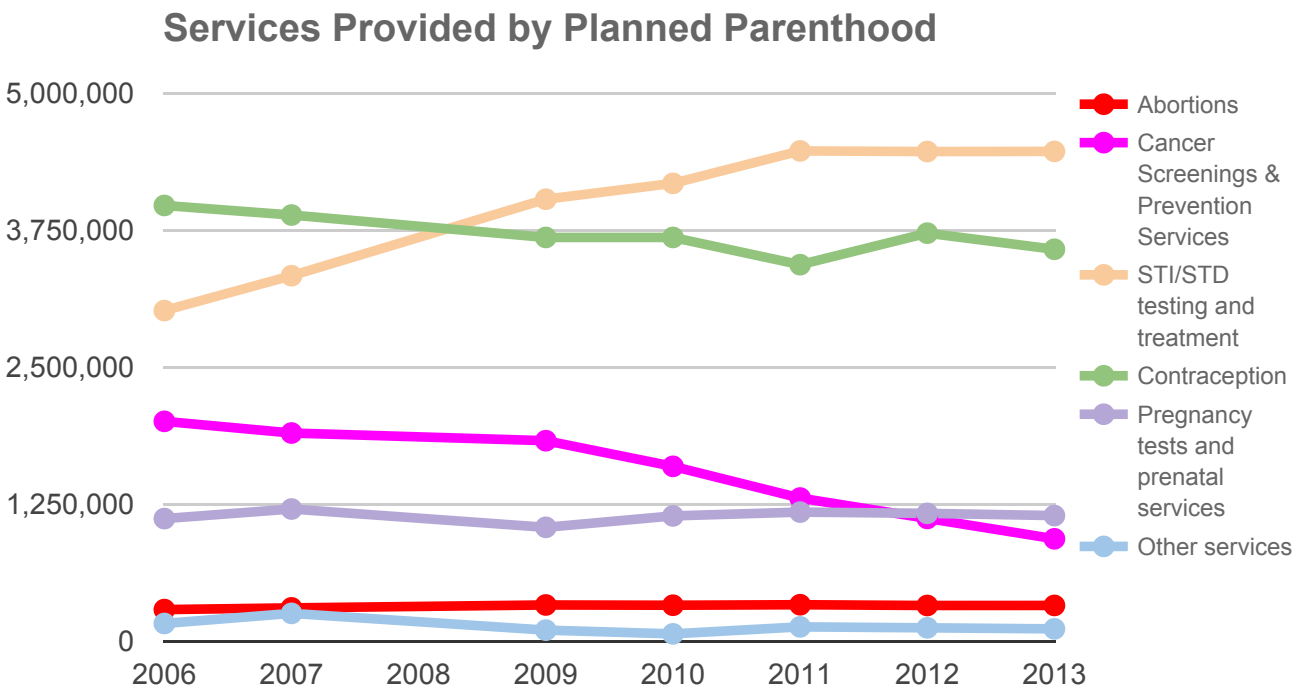
- An actual chart cited by an actual US congressman!
- "*That graphic is a damn lie. Regardless of whatever people think of this issue, this distortion is ethically wrong*" Alberto Cairo



http://www.politifact.com/truth-o-meter/statements/2015/oct/01/jason-chaffetz/chart-shown-planned-parenthood-hearing-misleading-/

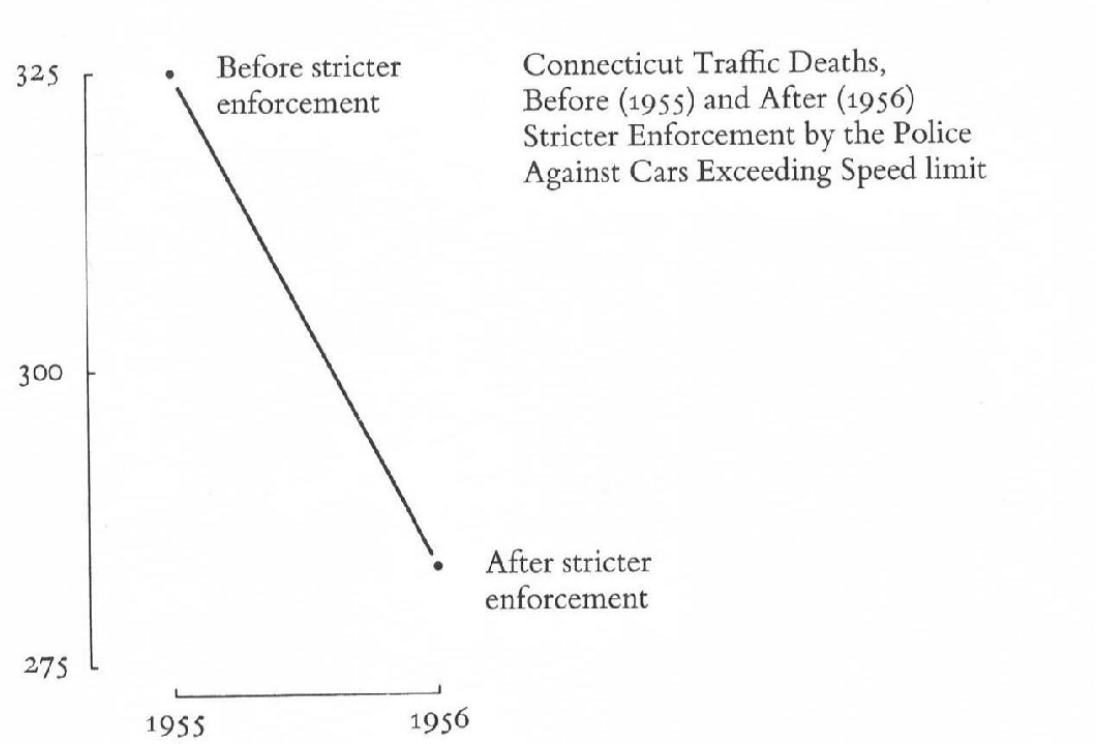
Example: Planned Parenthood

- same y-axis for all lines
- missing intermediate points added
- additional context from other spending

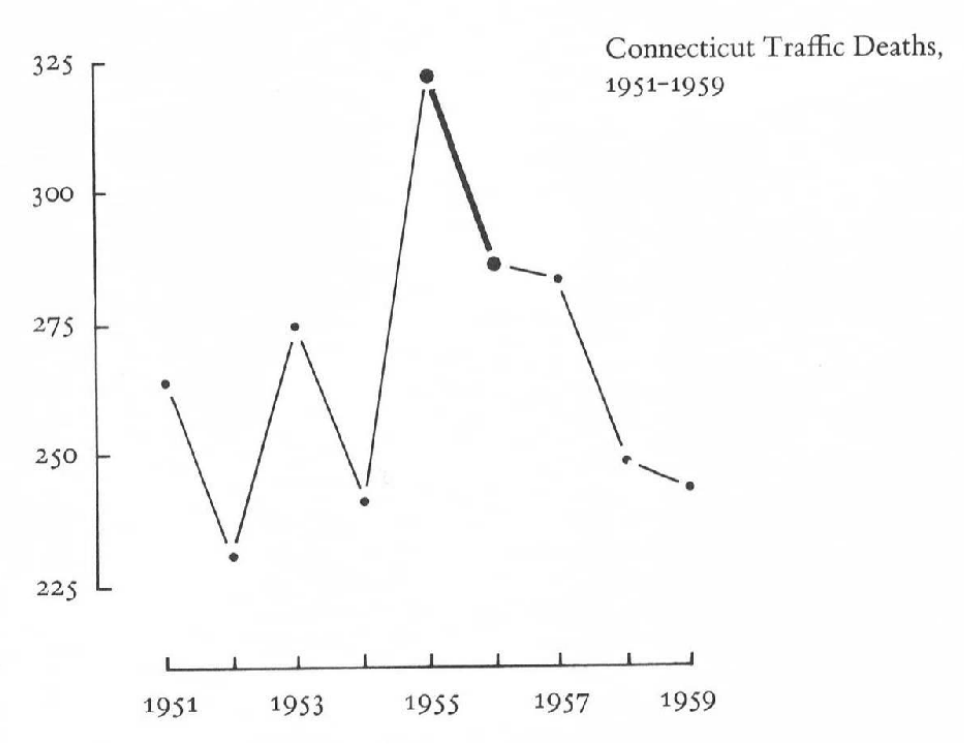


http://www.politifact.com/truth-o-meter/statements/2015/oct/01/jason-chaffetz/chart-shown-planned-parenthood-hearing-misleading-/

Context matters



without context

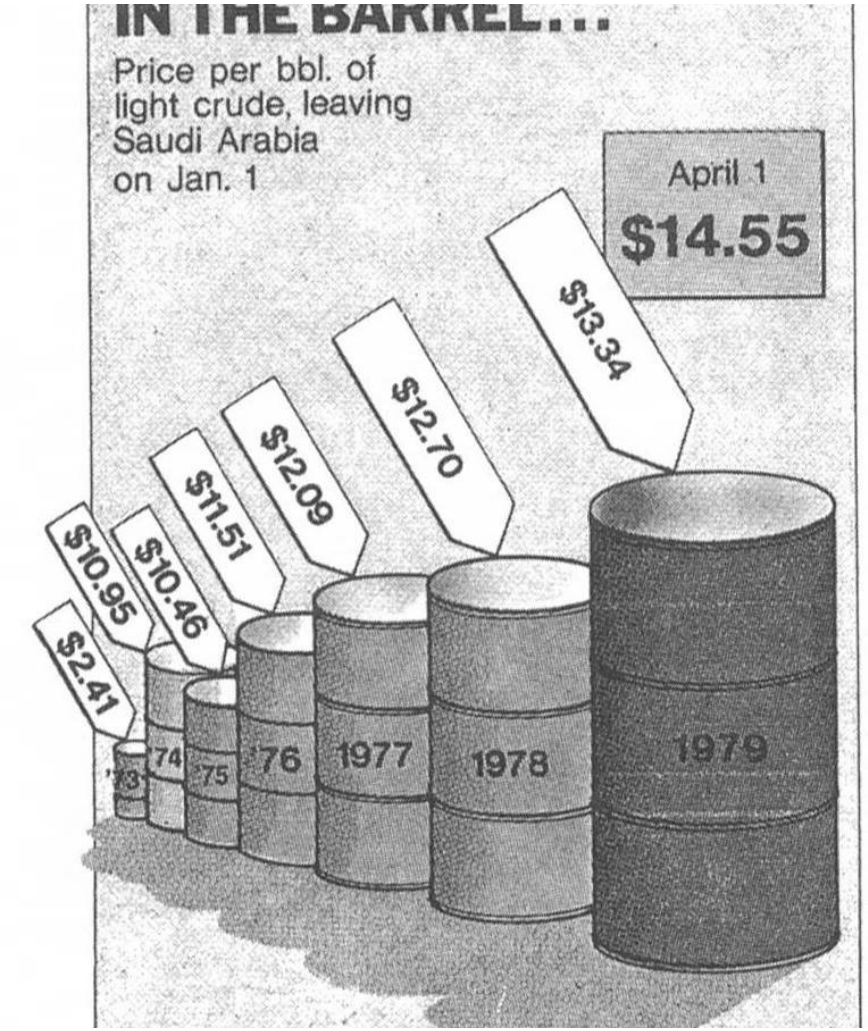


with context

Visual area and numerical measure

- an increase of 454% is depicted as an increase of 4280%
- Lie Factor?
- The viewer gets mixed up by the fact that a barrel (3D) represented by area (2D) is used to show 1D data,

The Time magazine (1979)



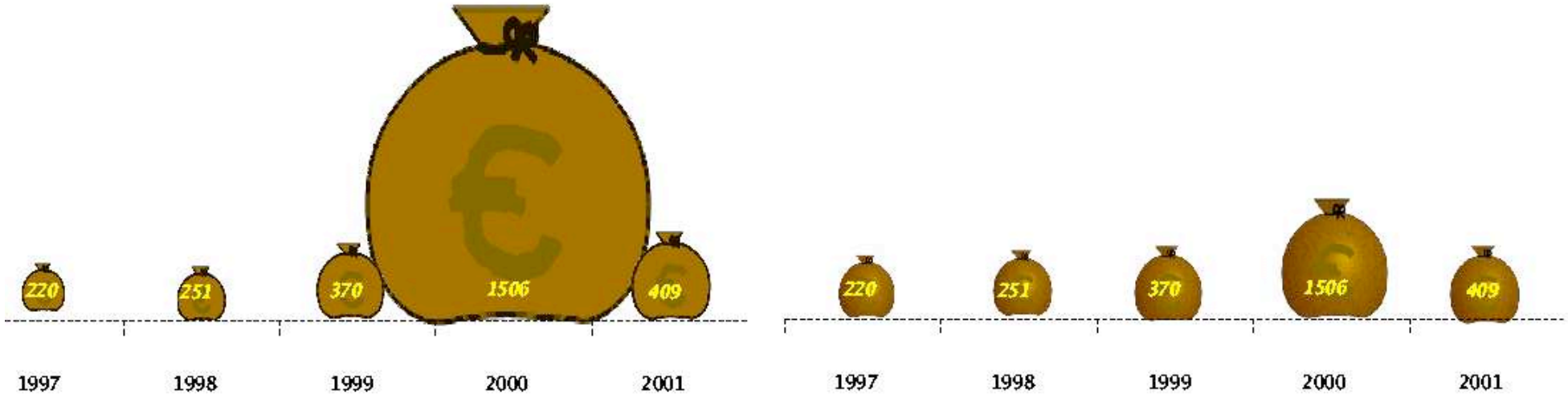
Perception of surface

- Many experiments on the visual perception of graphics have been conducted
  - people look at lines of varying length, circles of different areas and then recording their assessments of the numerical quantities
- E.g., the perceived area of a circle grows more slowly than the actual measured area,

$$\text{perceived area} = (\text{actual area})^{\alpha} \ , \ \alpha = 0.8 \pm 0.3$$

- However, different persons see the same areas somewhat differently
  - perceptions change with experience
  - perceptions are context-dependent

Height or volume?



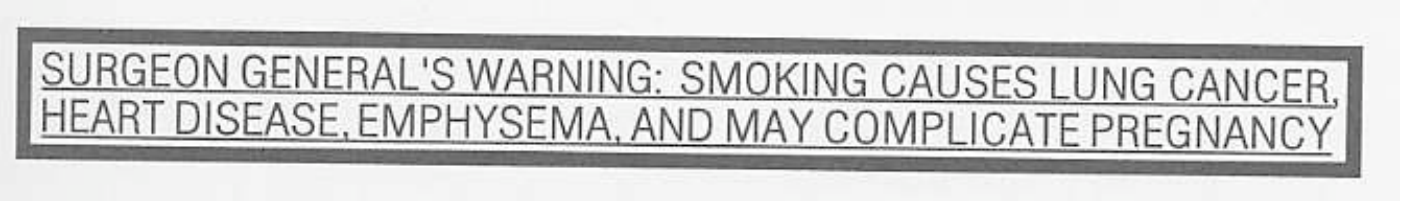
H. Ingo

Soneras profits 1997–2001

$$\text{lie factor} = \frac{\left(\frac{1506}{220}\right)^3}{\left(\frac{1506}{220}\right)} = 47$$

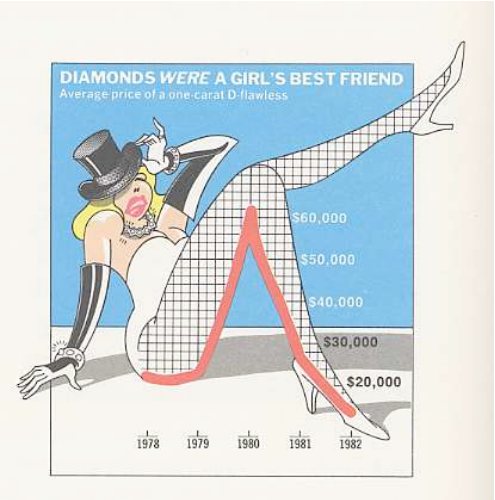
Distractions

Decorating graphics does not necessarly mean lying, but decorations can be used to distract the reader.



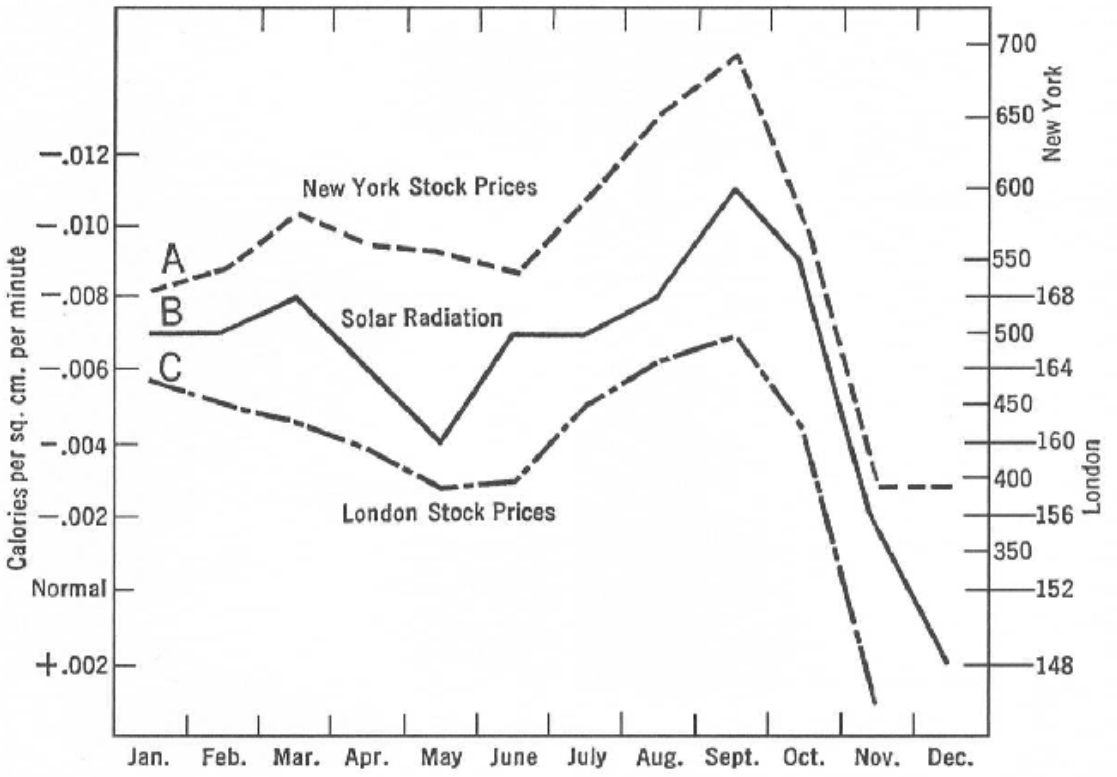
[VE 65].

Strong frames and capital letters make the text hard to read. There is a negative white space between the words and the border.



?, [EI 34].

False Conclusions from True Premises?

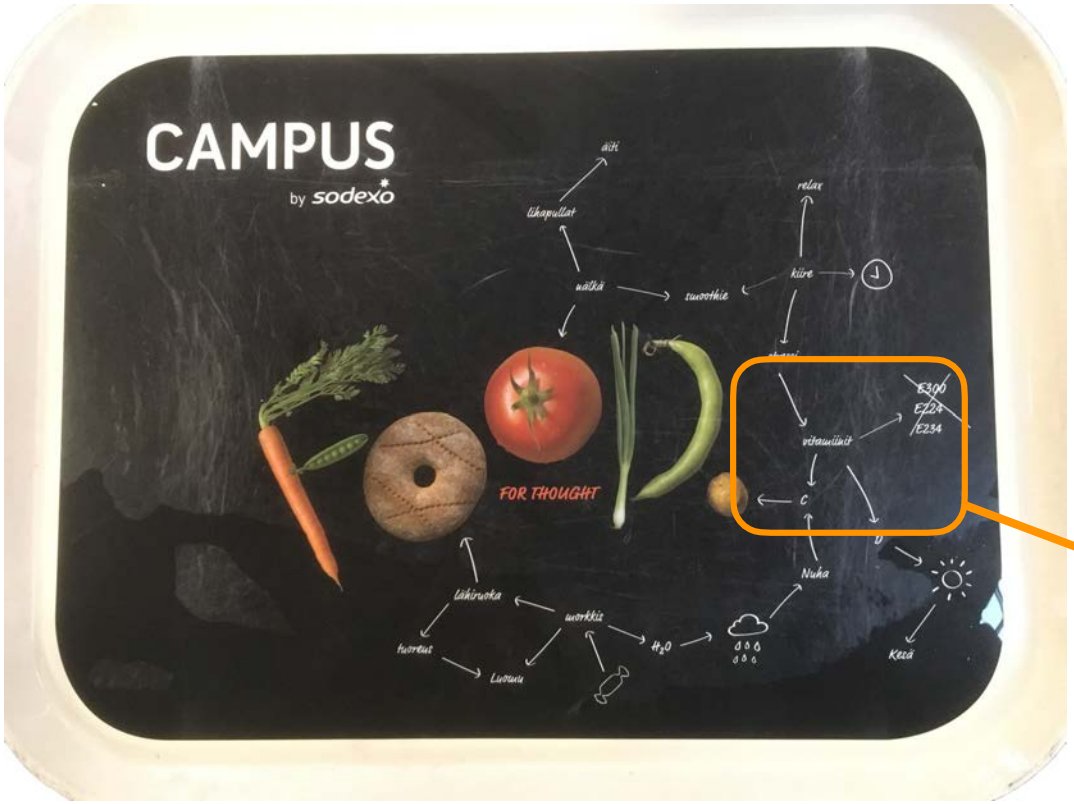


SOLAR RADIATION AND STOCK PRICES

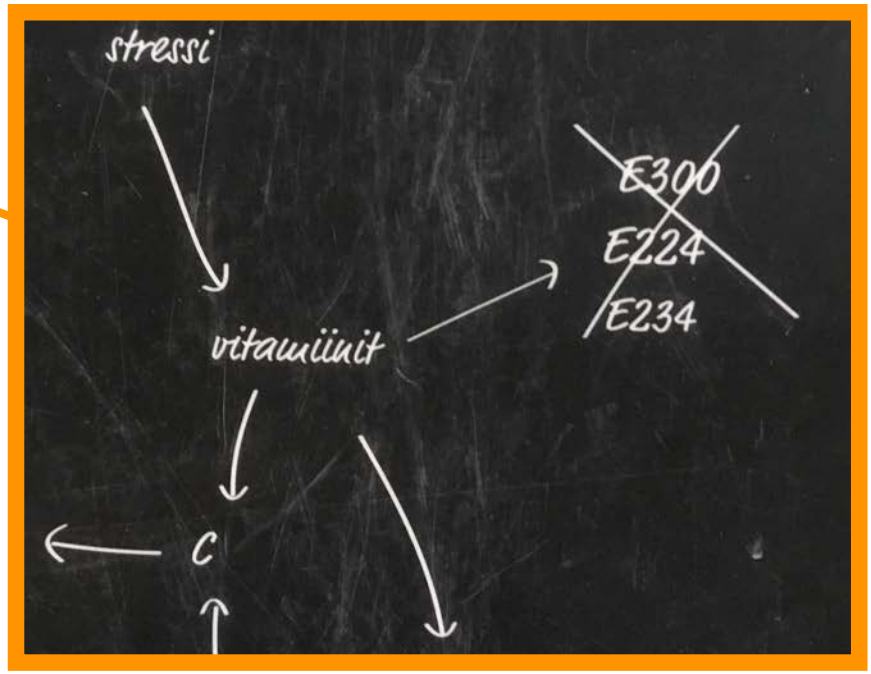
A. New York stock prices (Barron's average). B. Solar Radiation, inverted, and C. London stock prices, all by months, 1929 (after Garcia-Mata and Shaffner).

http://www.tylervigen.com/spurious-correlations

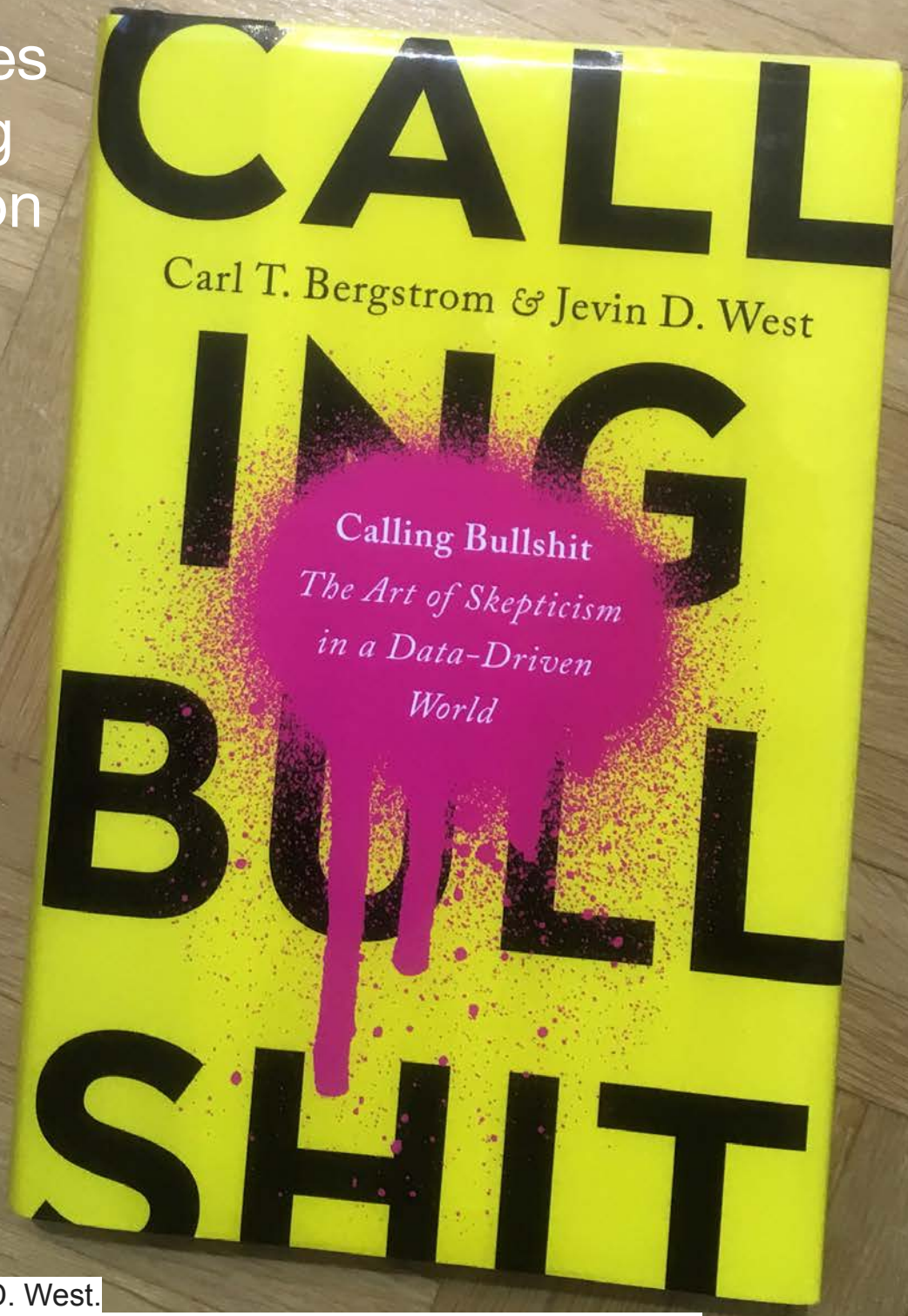
Check your information!



for discussion
on next lecture:
what's wrong?



Great examples of misleading communication and how to avoid it.



Bergstrom, Carl T., and Jevin D. West.
Calling bullshit: the art of skepticism in a data-driven world. Random House, 2020.

Theory of data graphics
(by Tufte)

- Some patterns of good graphics start to emerge...

Theory of data graphics

- Influential concepts
- Not an exact theory (no quantitative truths, a common sense is still needed)
- Theory of human perception will be discussed in later lectures
- How is it useful for me?
  - Knowing these things helps to see the difference between the good and bad solutions in information visualization
  - Terminology is good to know

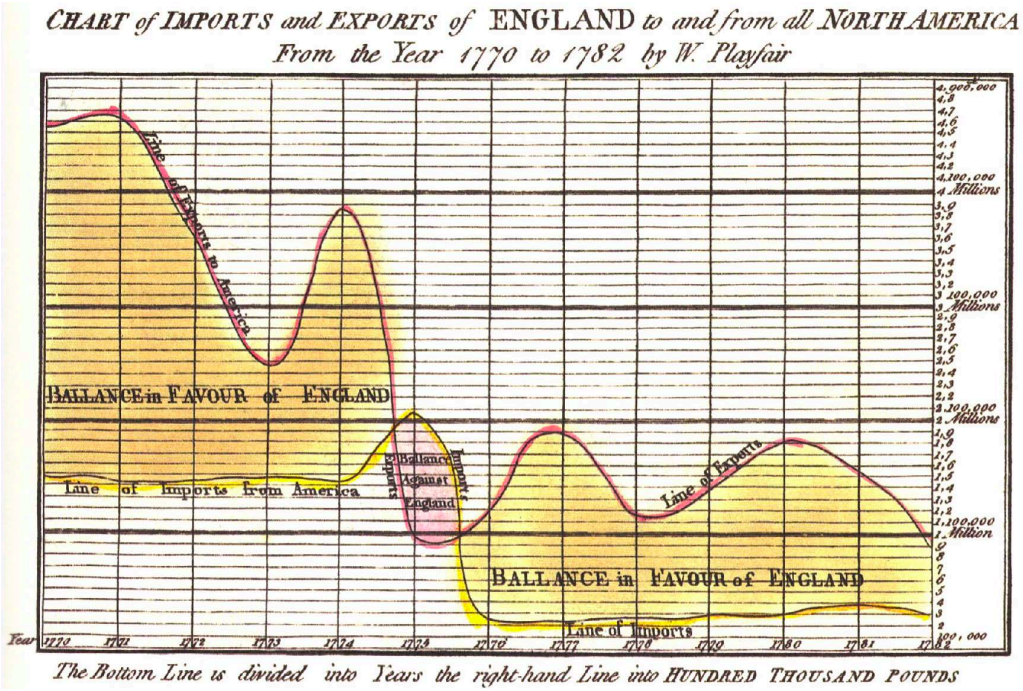
Theory of data graphics

- The idea:
  - Give the viewer the greatest number of ideas in the shortest time
  - Use the least amount of ink
  - Don't waste space
  - Eliminate non-essentials and redundancies

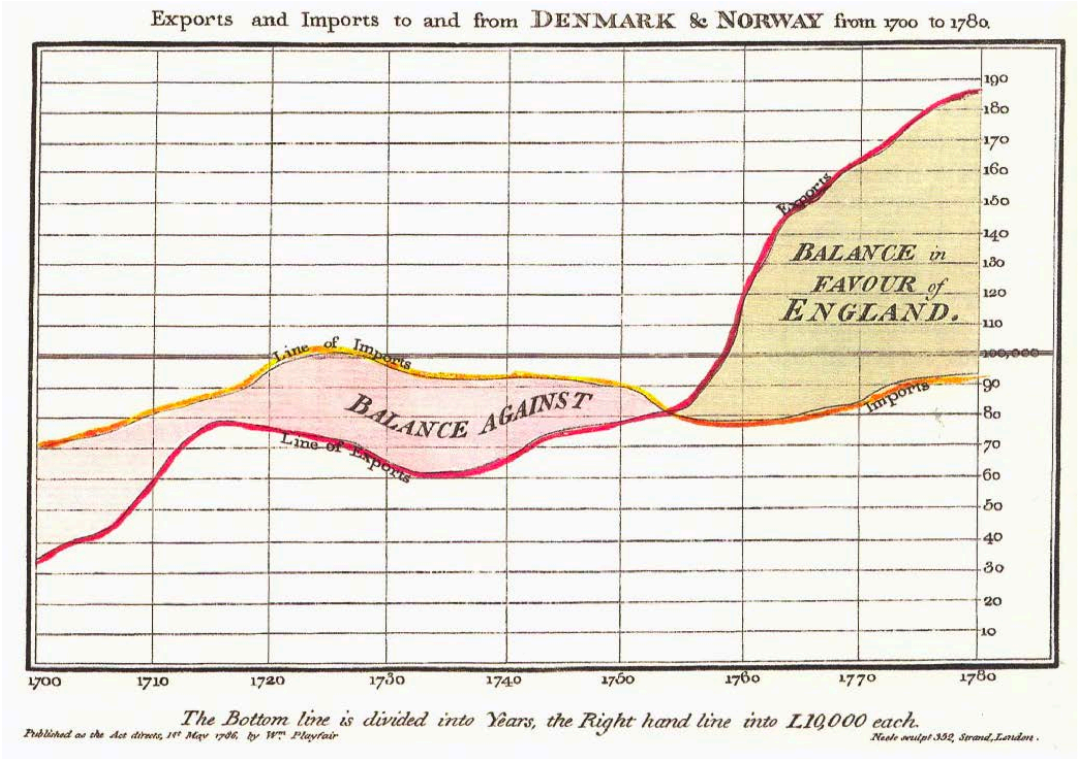
- Or:
  - Make the graphics as easy to read and as simple as possible, while displaying the data fully.

Which is better?

(a)



(b)



Theory of data graphics

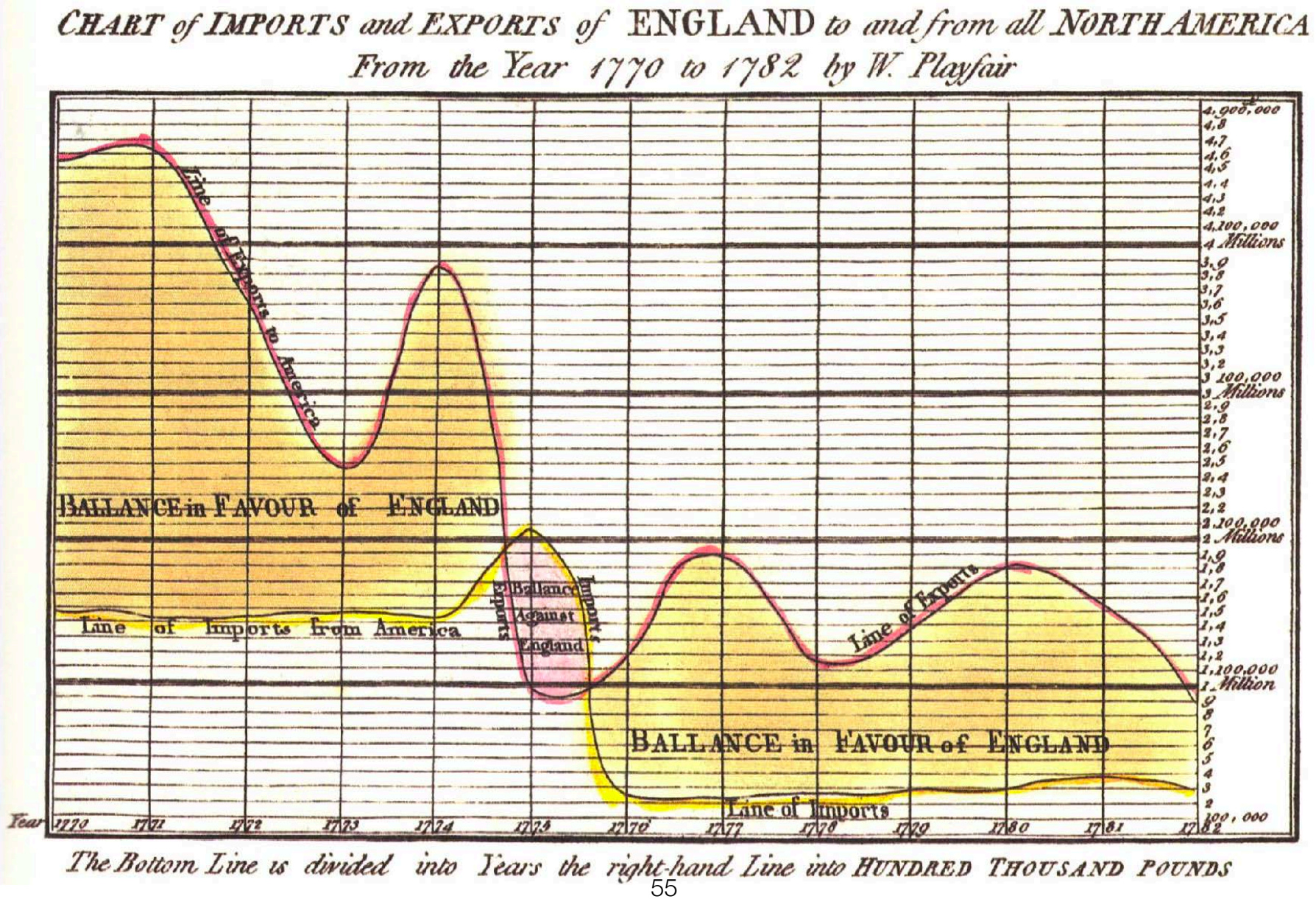
- Data-ink
- Chartjunk
- Multifunctioning graphical elements
- Data density and small multiples
- Aesthetics and techniques

Data-ink

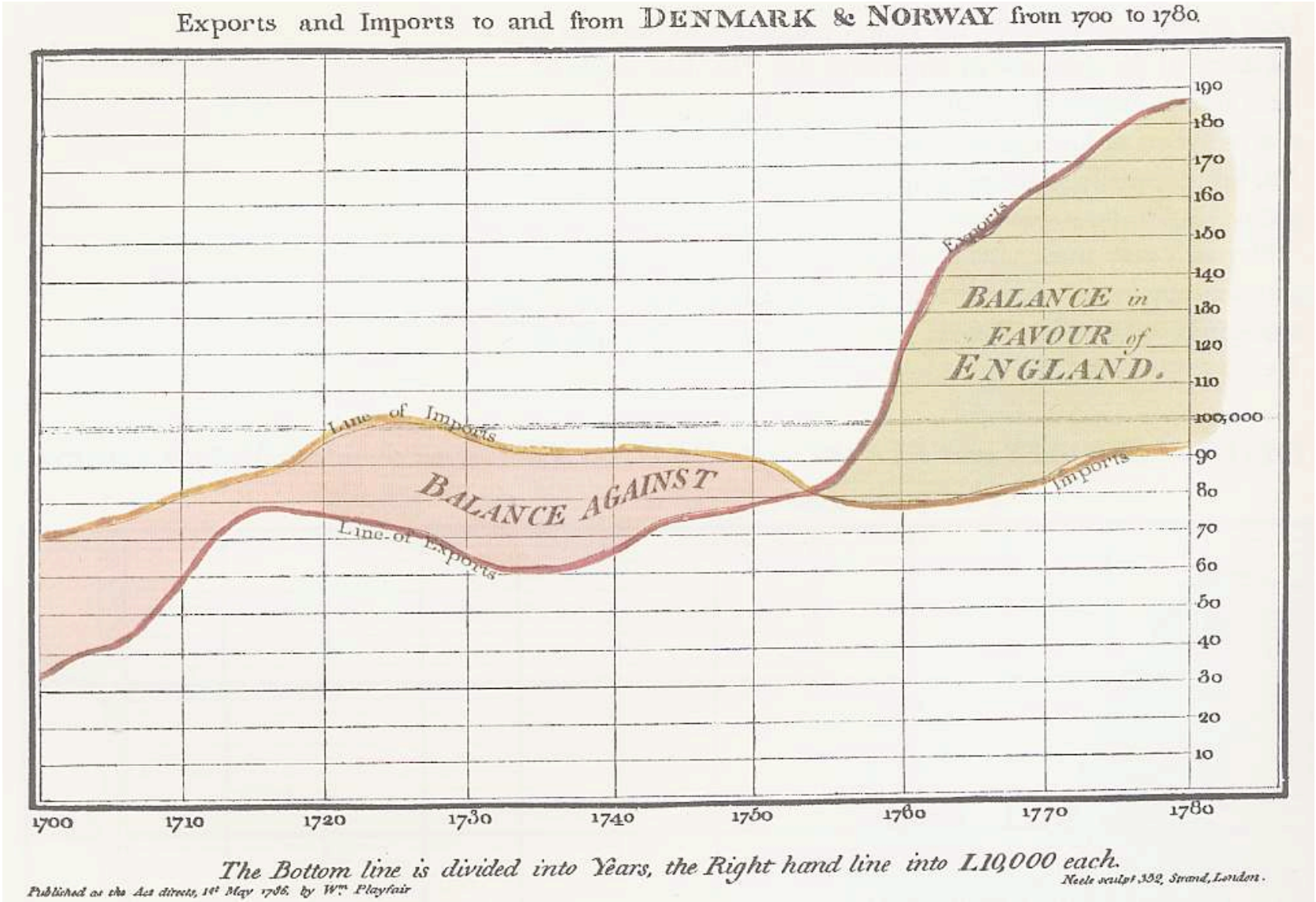
- Data consists of *empty space* (white paper) and *ink*
- *Data-ink* is the non-erasable and non-redundant core of graphics. Erasing data-ink would reduce the amount of information transmitted by the graphics

$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used to print the graphics}}$$

Example: low data-ink ratio

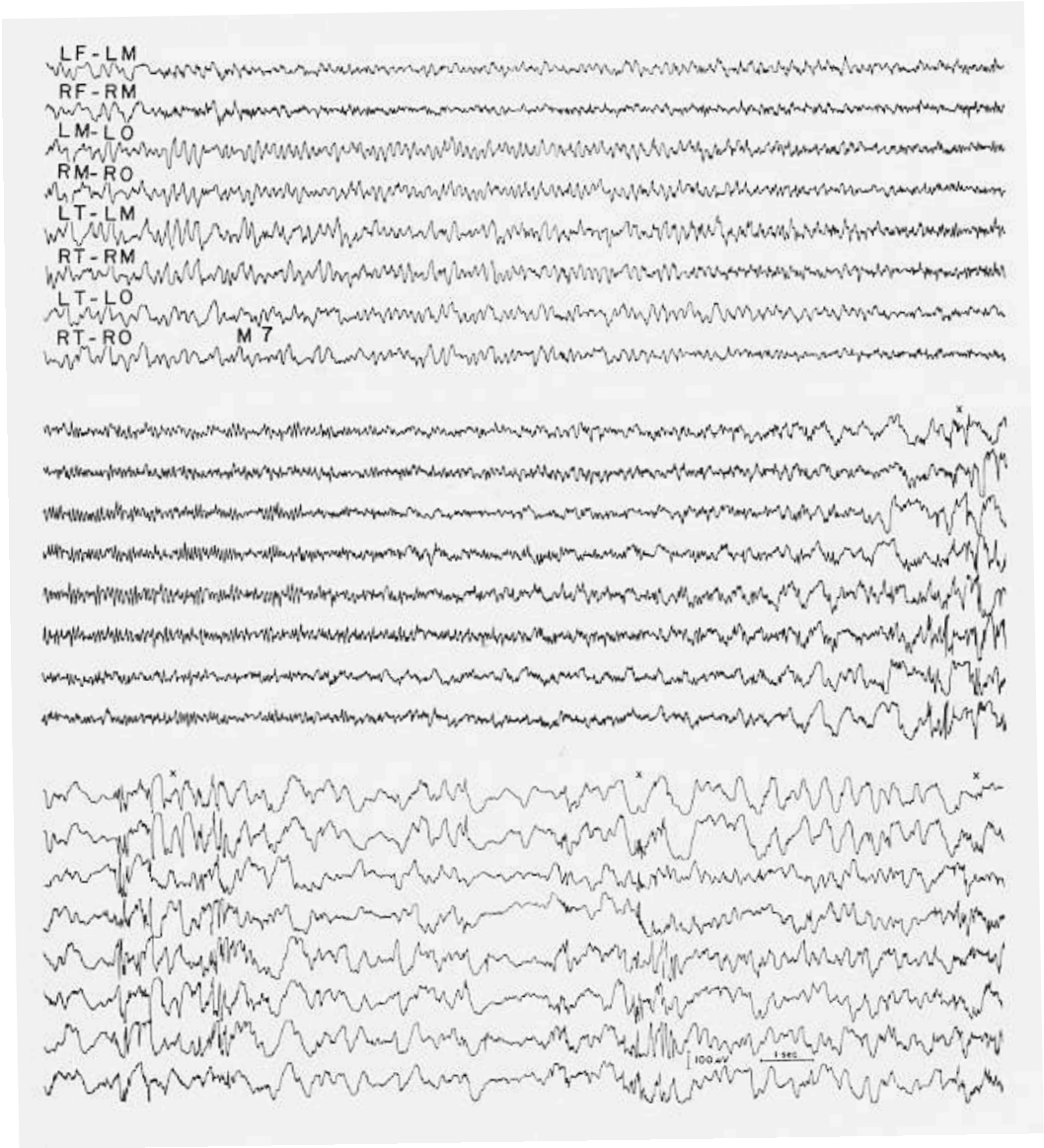


Example: high-data ink ratio



William Playfair, 1786 [T 92].

100% data-ink ratio

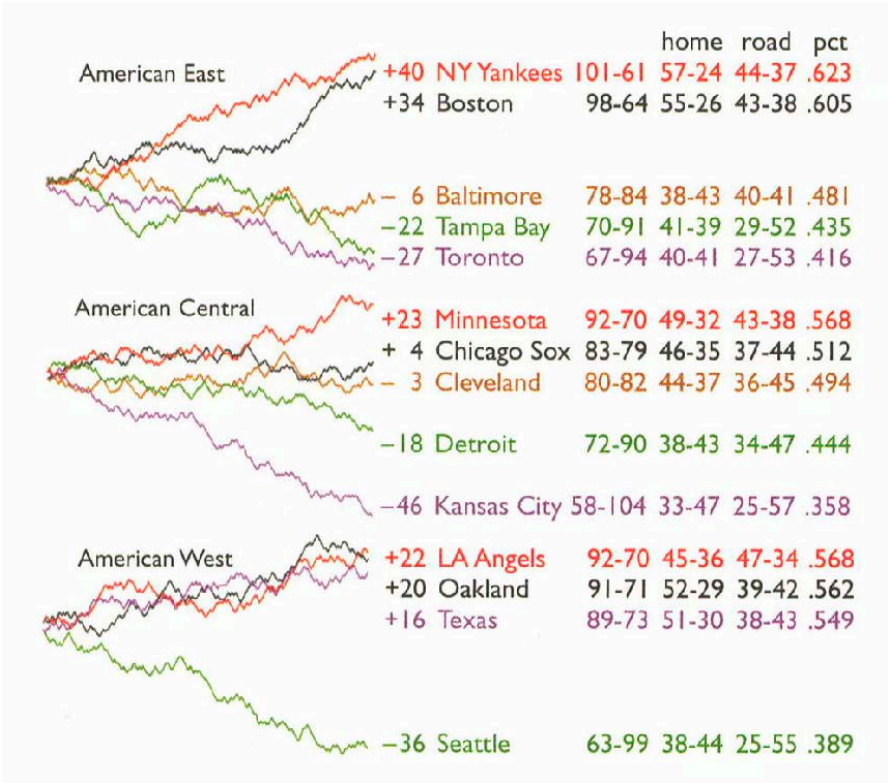


… but you need to know how to read it.

Kenneth A. Kooi, 1971 [T 93].

Sparklines

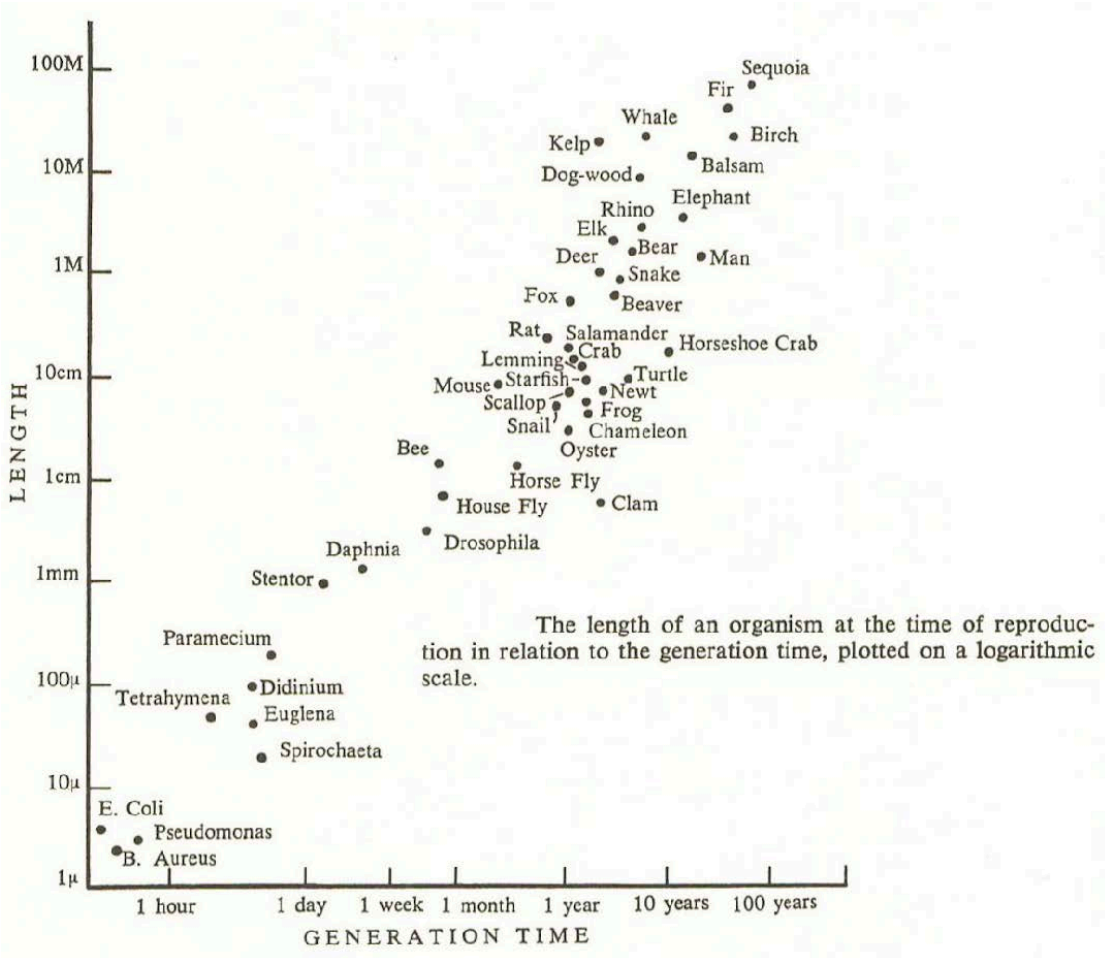
- compact time series
- data-ink ratio = 1
- labels clear from context
- can be used inline with main text



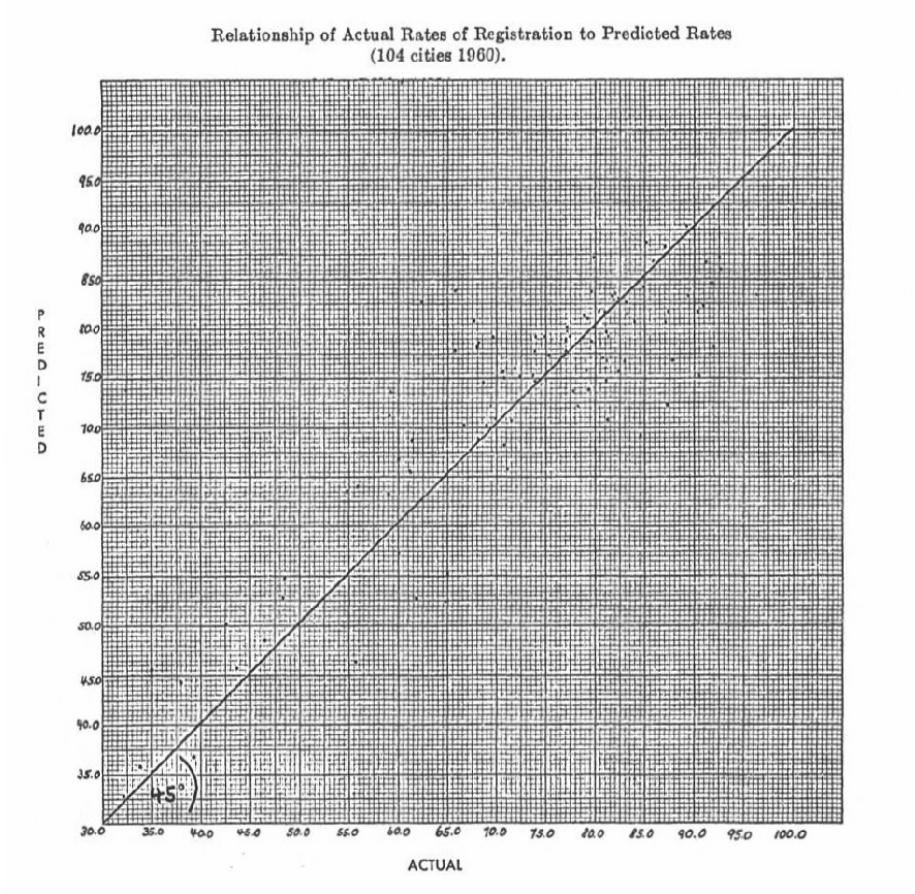
Using d3.js, we can fairly easily draw SVG-based sparklines. This is 2013 historical stock prices for **Google** **$1084.75**. And this is for **Facebook** **$55.57**. And this is for **Apple** **$550.77**. Each sparkline has 244 data points, but it's condensed very nicely.

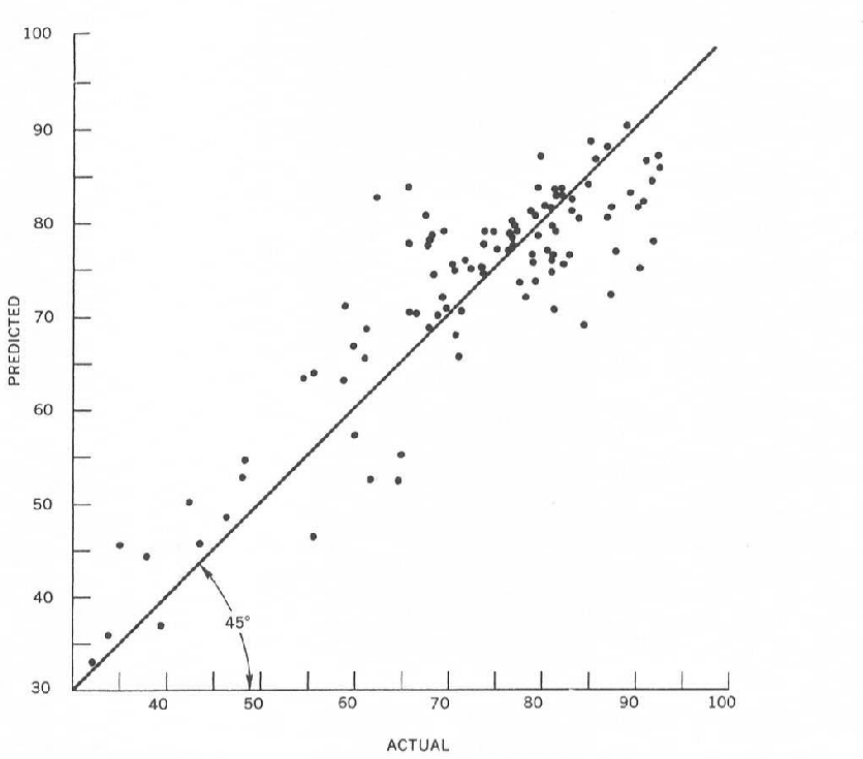
Another example

- Most of the ink here is data-ink
  - the dots and labels on the diagonal
- with, 10-20 percent non data-ink
  - the grid ticks and the frame



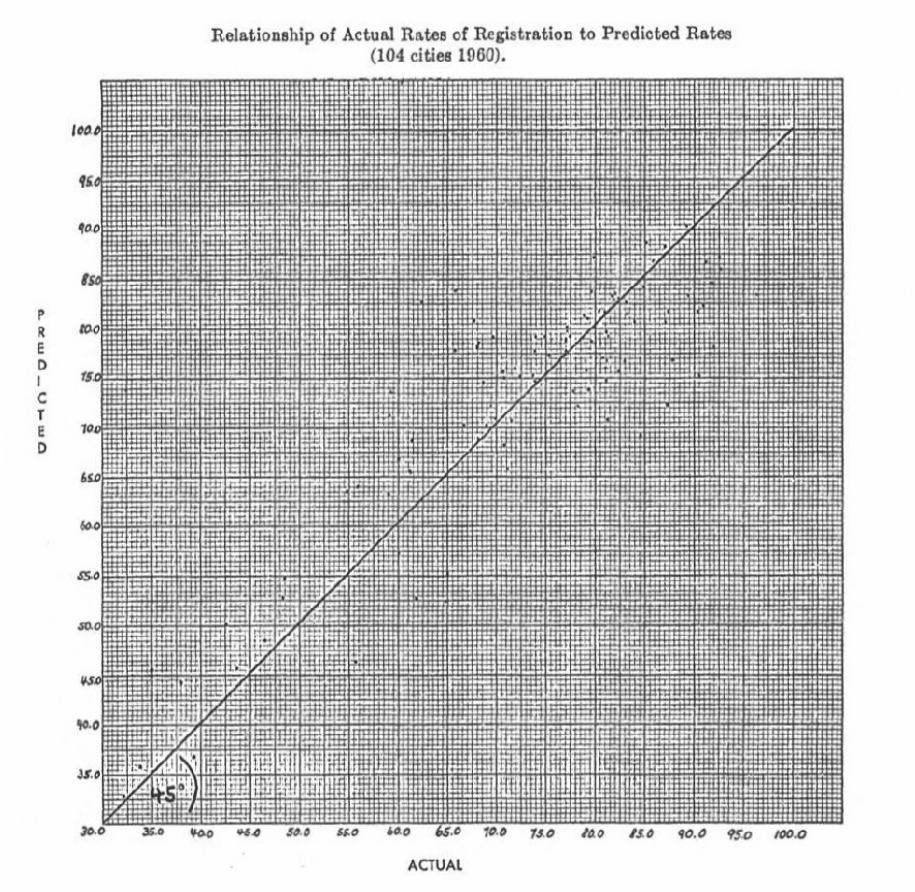
Improving data-ink ratio

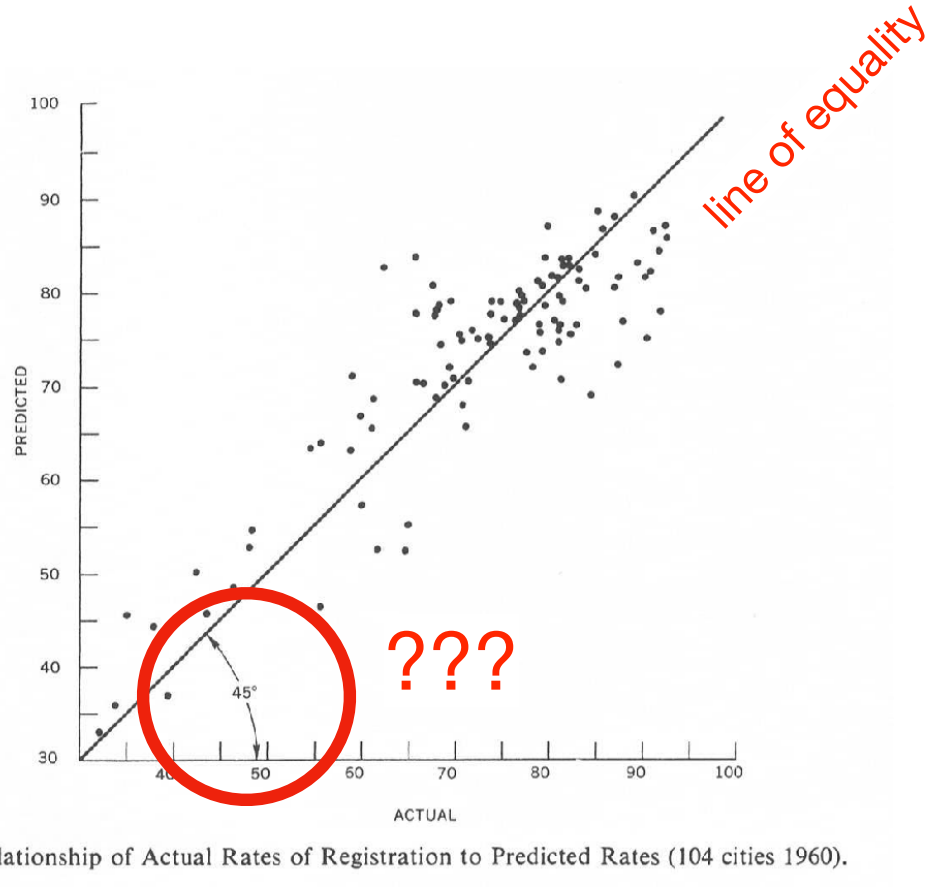




Relationship of Actual Rates of Registration to Predicted Rates (104 cities 1960).

Improving data-ink ratio





Relationship of Actual Rates of Registration to Predicted Rates (104 cities 1960).

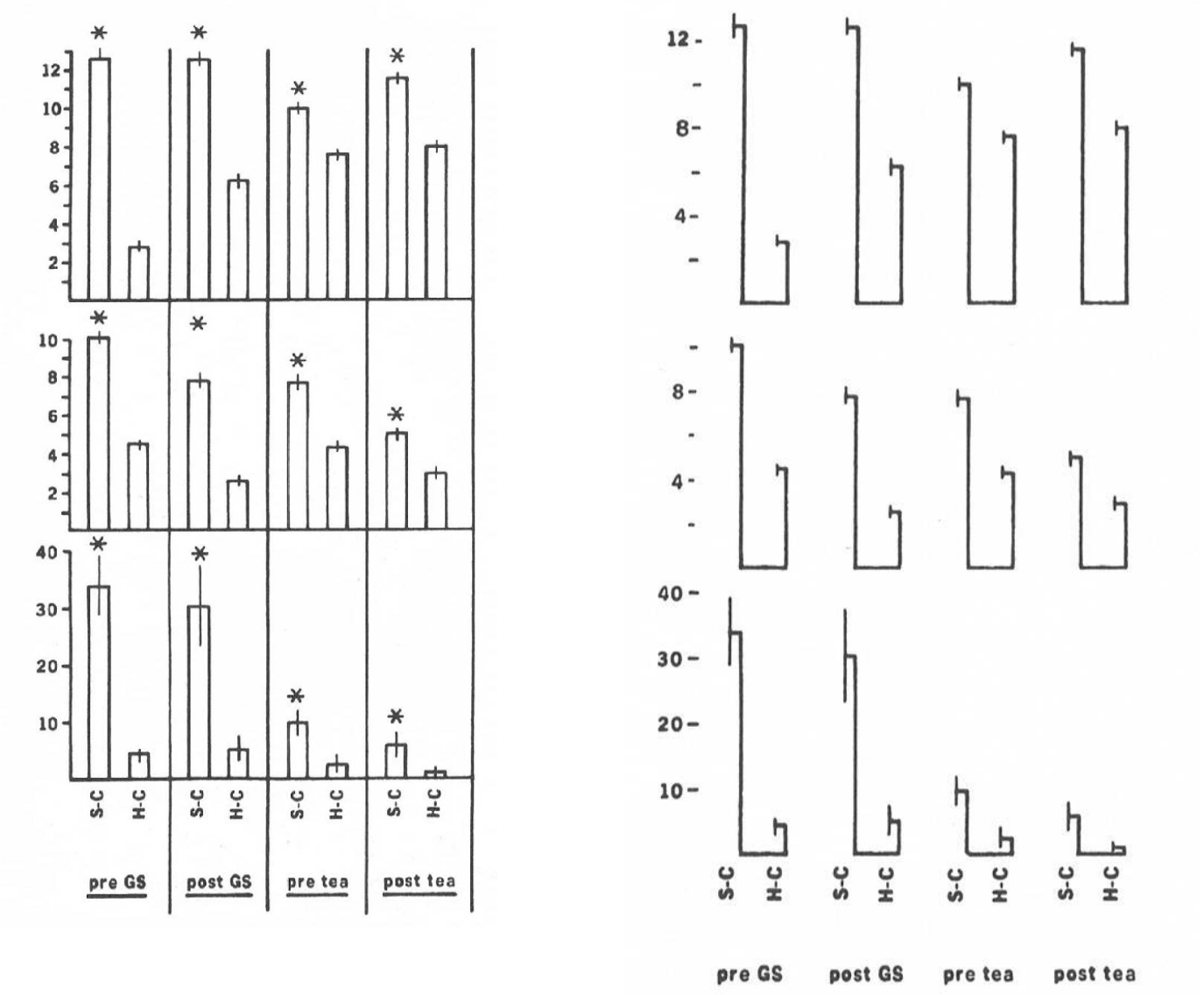
Maximize data-ink

- It is always a good idea to maximize the data-ink ratio, within reason
- The larger the share of data-ink the better, other matters being equal
  - every bit of ink on a graphic needs a reason
  - nearly always that reason being that the ink presents new information
- Ink that fails to depict statistical information is uninteresting, and often it is also dull

Maximize data-ink

- To increase the proportion of data-ink use two erasing principles
  - erase non data-ink
  - erase redundant data-ink
- Non data-ink is ink that fails to depict information, it has little interest to the viewer
  - sometimes, such non-data-ink clutters up the data
  - sometimes, such non-data-ink helps set the stage
- Redundant data-ink depicts information but it does it showing it over and over

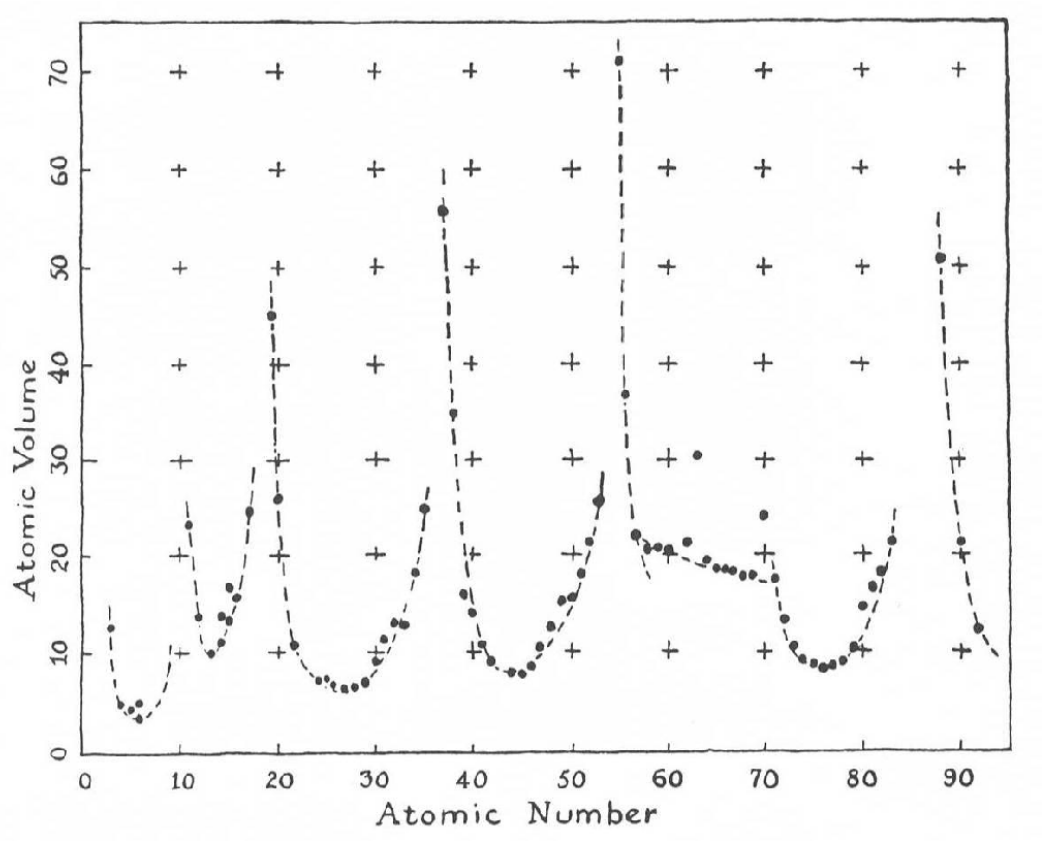
Edit and redesign



Kuznicki et al. Journal of Experimental Psychology: General, 108 (1979), 76.

Edit and redesign

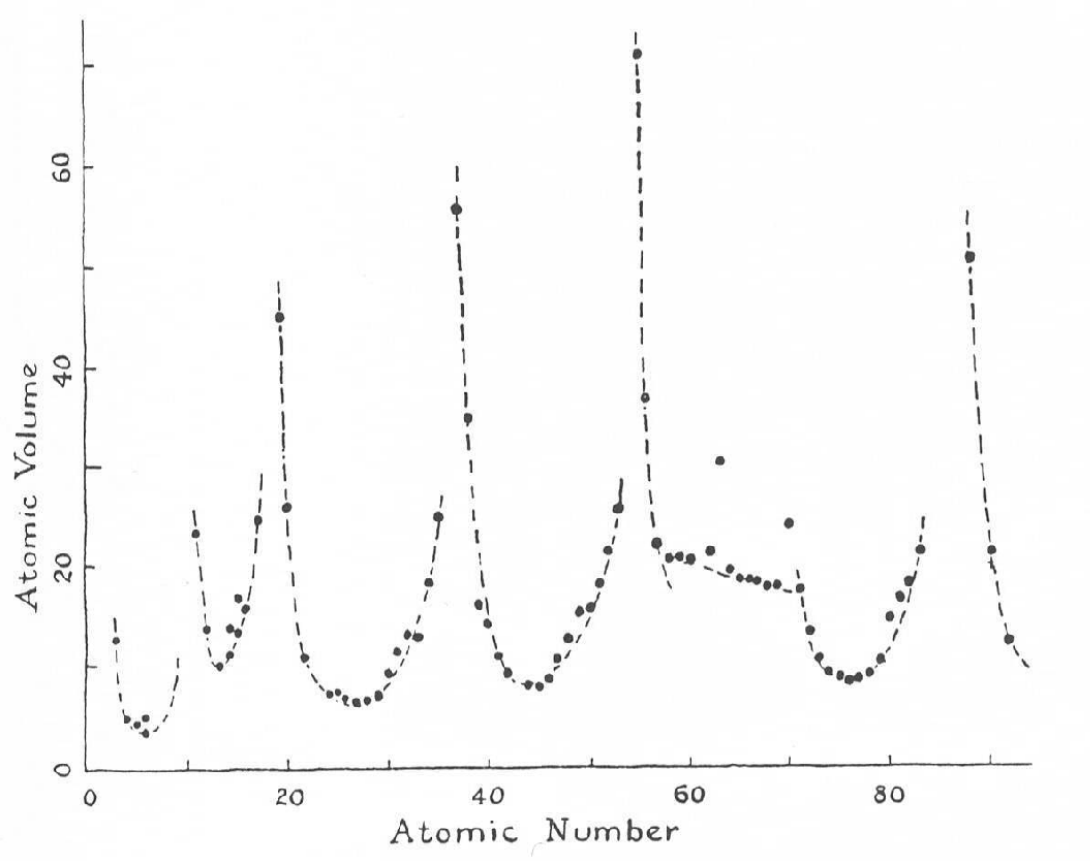
- The data-ink ratio is about 0.6
  - 76 data points and the reference curve are obscured by 63 grid marks
- The grid and part of the frame can be erased to improve the data-ink ratio



Linus Pauling, General Chemistry, p. 64, 1947

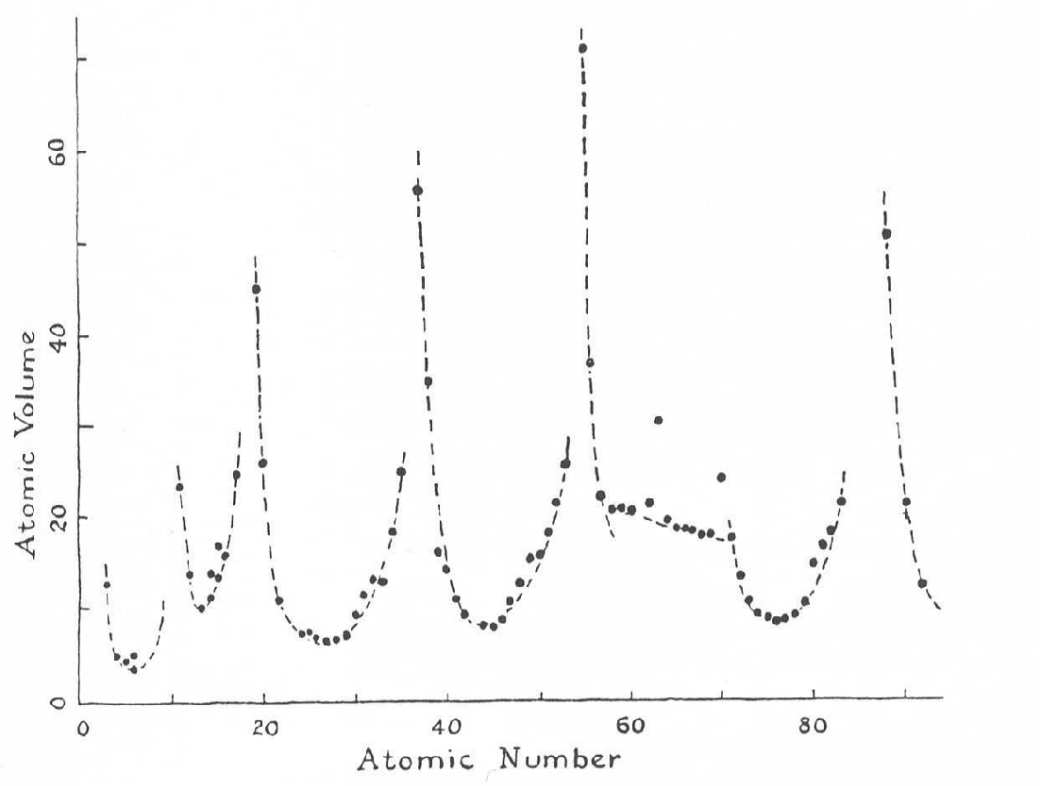
Edit and redesign

- Data-ink ratio improves to 0.9
    - only the frames line are uninformative
- erasing the grid marks highlights that several of the elements do not fit the smooth theoretical curve so well

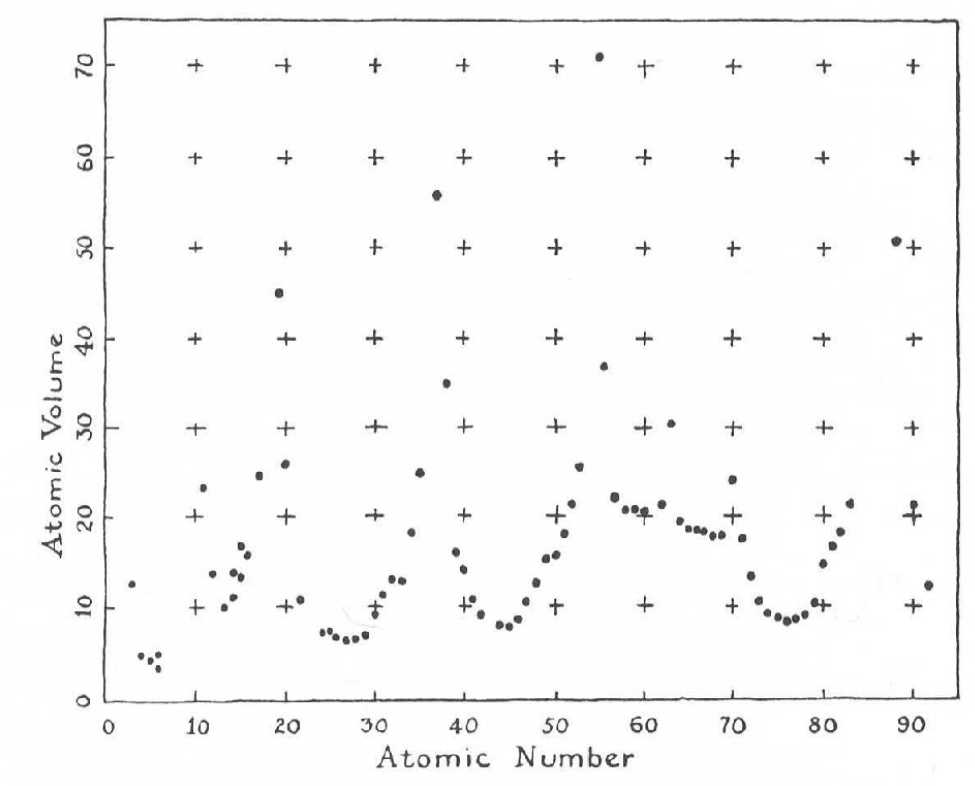


The reference curve is essential in organizing the data, and shows the periodicity (the message) by creating a structure, and by giving ordering and hierarchy

Edit and redesign



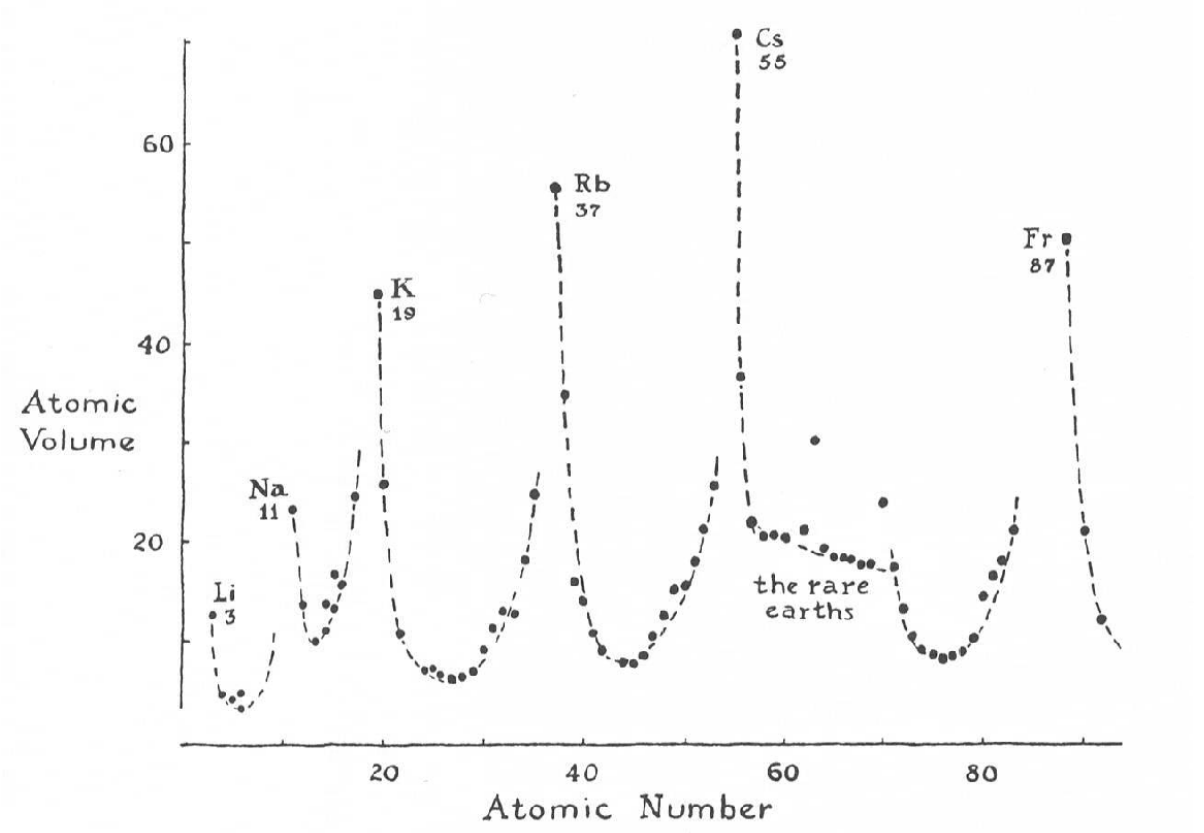
Without the curve we hardly detect the periodicity. The curve becomes necessary because the eye needs guidance



Restoring the grid totally fails to organise the data. The grid marks are too powerful and induce visual vibration.

Edit and redesign

- We can use the erased space
  - labels for the initial elements of each period
  - unusual rare-earths
- also, turned label and numbers on the vertical axis
- **Message: do not be happy with the initial version of your graphics!**



Next lecture

- Theory of data graphics (Tufte, Part II)