CS-E4840
Information Visualization
Lecture 3: Theory of data graphics

Tassu Takala <tapio.takala@aalto.fi>
4 March 2019

Theory of data graphics

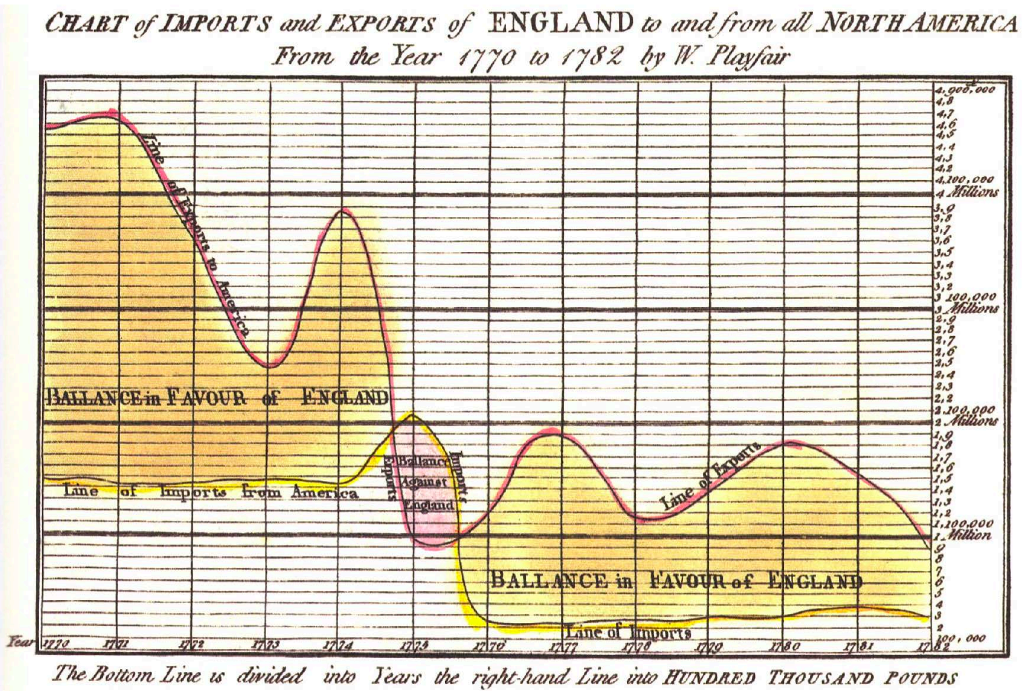
- Influential concepts
- Not an exact theory (no quantitative truths, a common sense is still needed)
- Theory of human perception will be discussed in later lectures
- How is it useful for me?
    - Knowing these things helps to see the difference between the good and bad solutions in information visualization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

Theory of data graphics

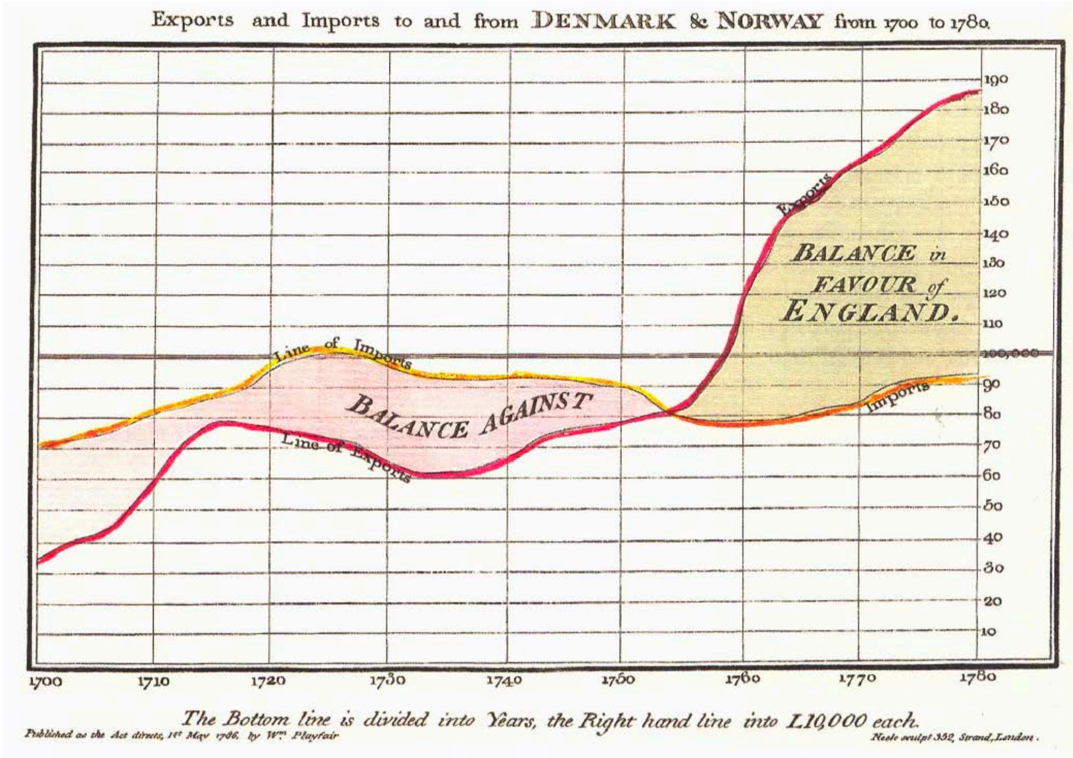
- The idea:
  - Give the viewer the greatest number of ideas in the shortest time
  - Use the least amount of ink
  - Don’t waste space
  - Eliminate non-essentials and redundancies
- Or:
  - Make the graphics as easy to read and as simple as possible, while displaying the data fully.

Which is better?

(a)



(b)



Theory of data graphics

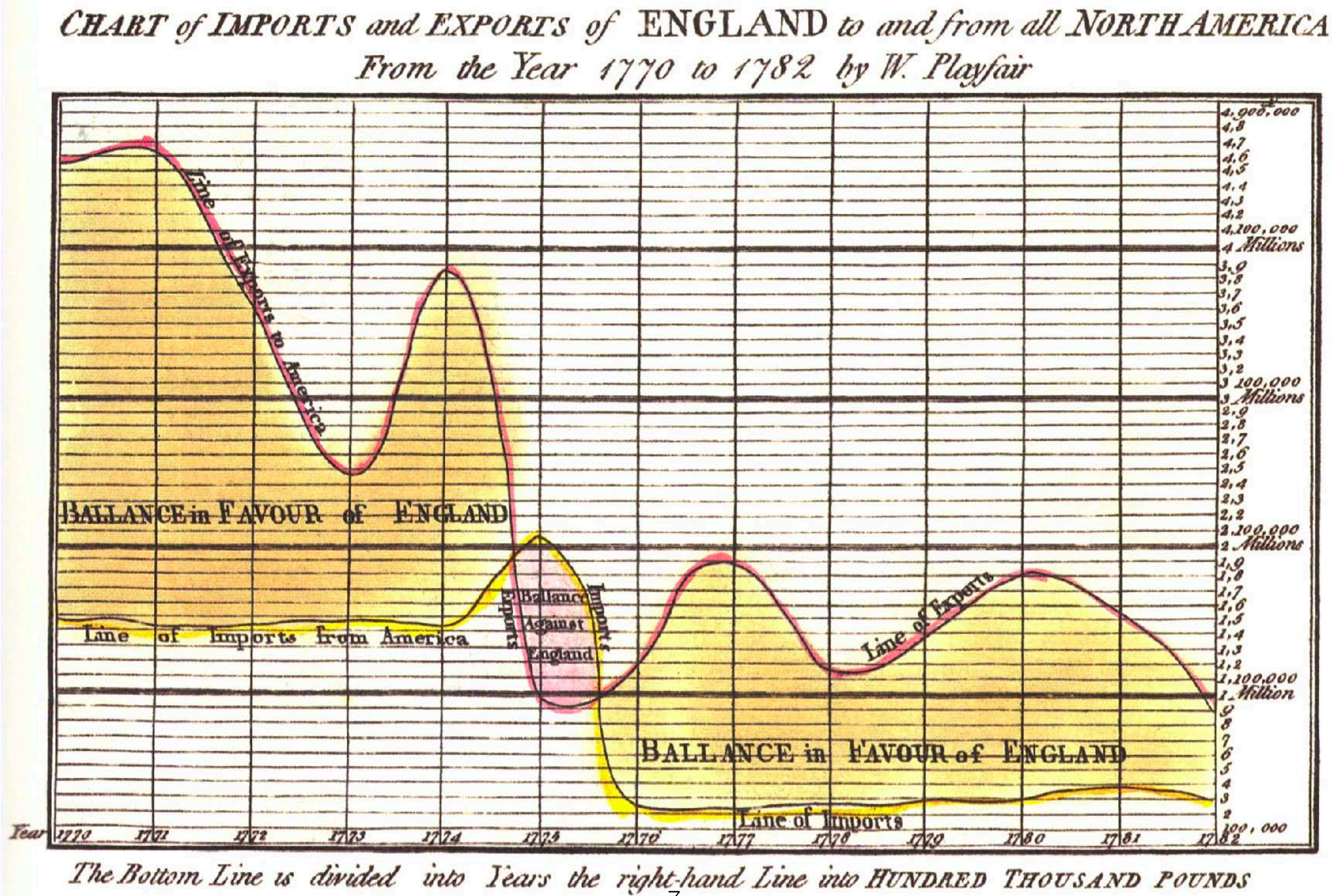
- Data-ink
- Chartjunk
- Multifunctioning graphical elements
- Data density and small multiples
- Aesthetics and techniques

Data-ink

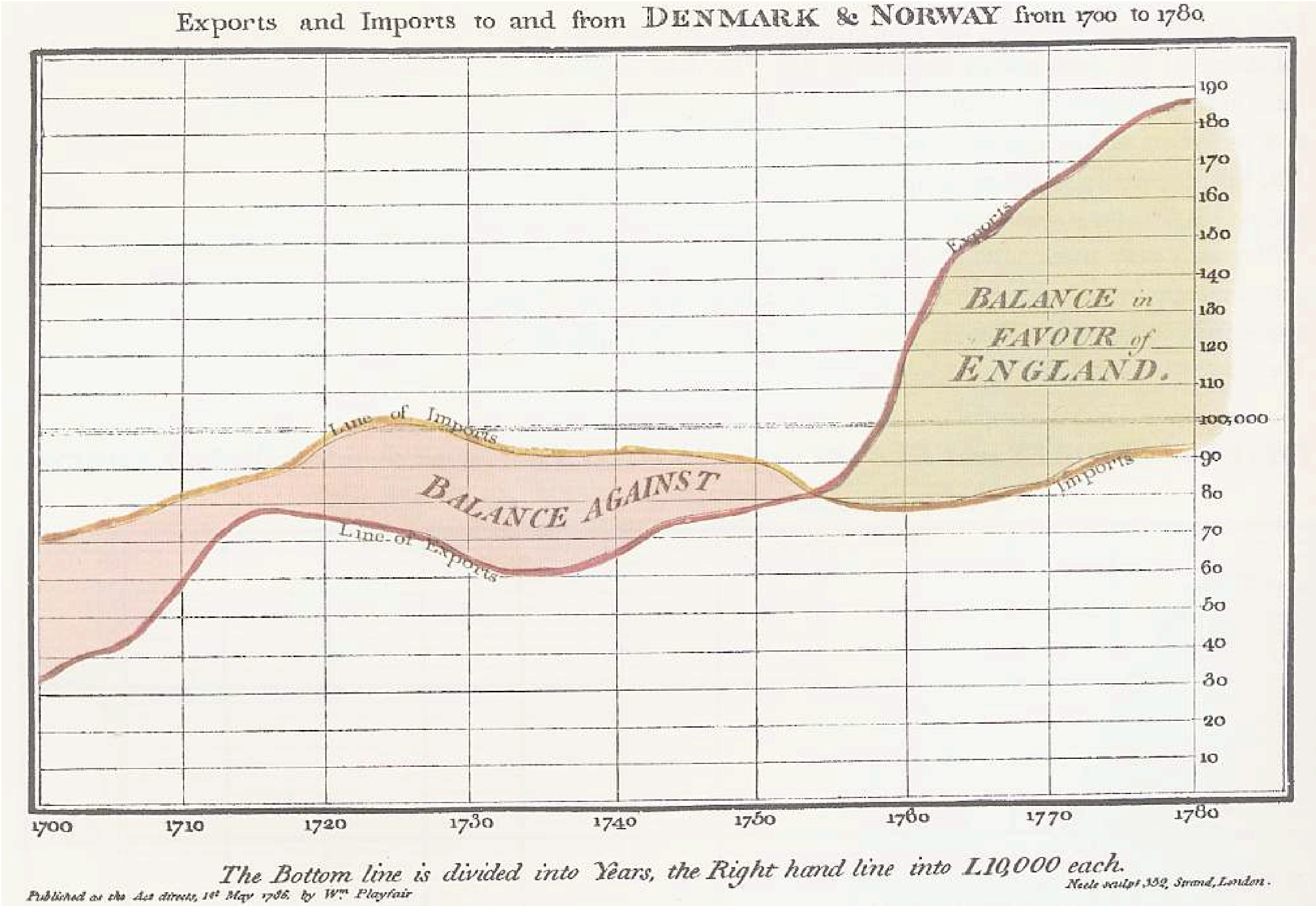
- Data consists of *empty space* (white paper) and *ink*
- *Data-ink* is the non-erasable and non-redundant core of graphics. Erasing data-ink would reduce the amount of information transmitted by the graphics

$$\text{data-ink ratio} = \frac{\text{data-ink}}{\text{total ink used to print the graphics}}$$

Example: low data-ink ratio



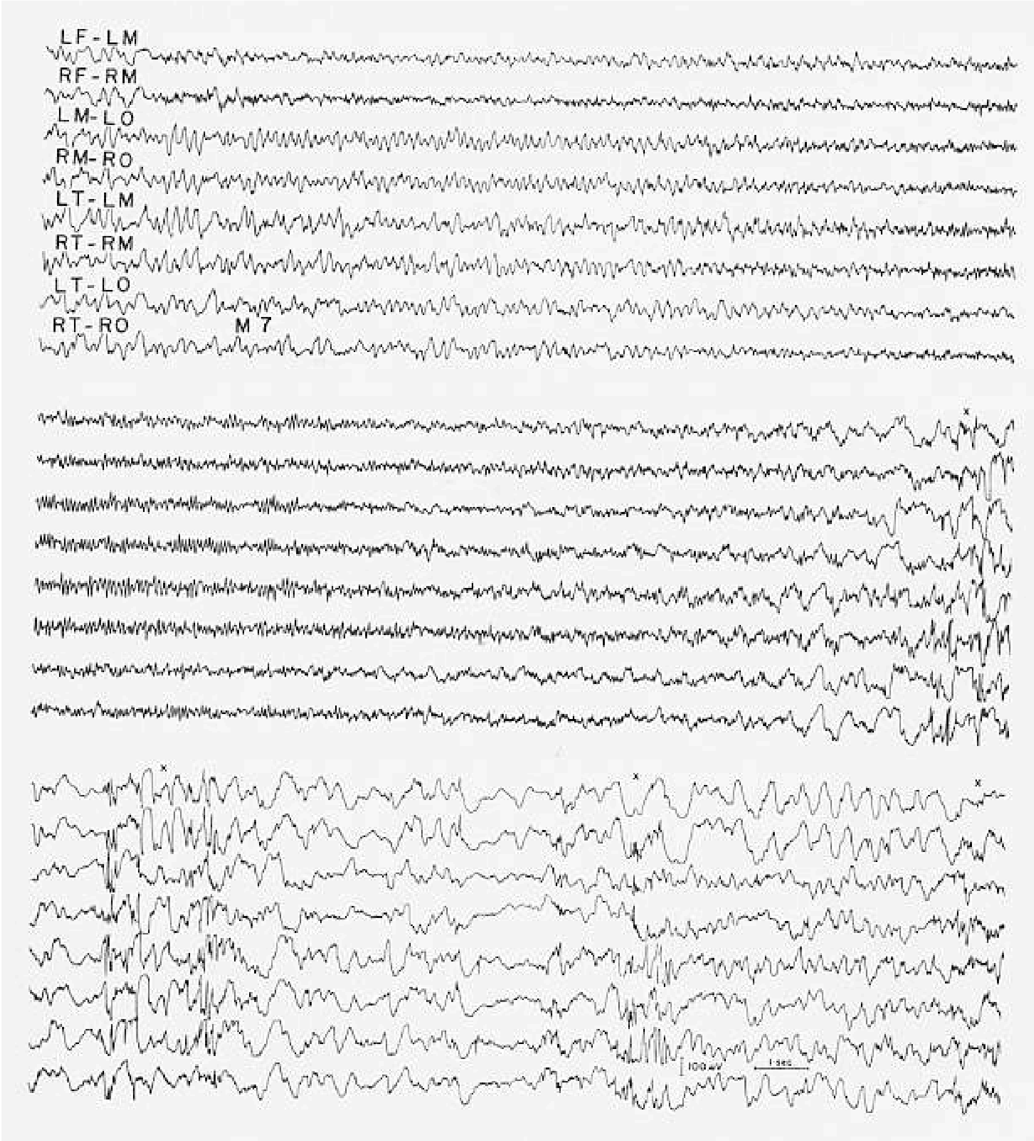
Example: high-data ink ratio



8

William Playfair, 1786 [T 92].

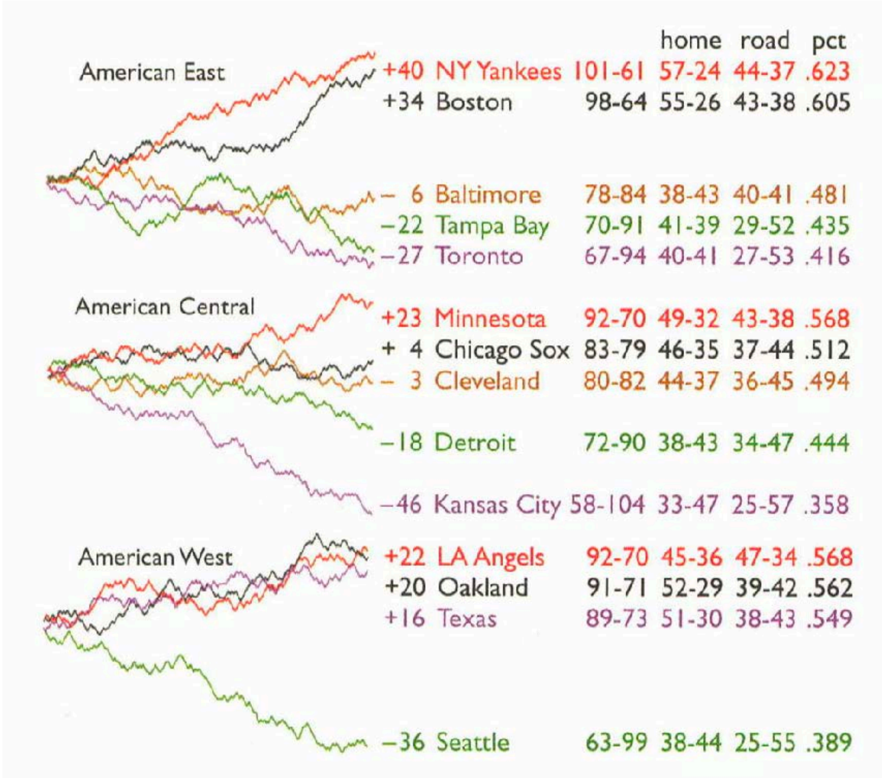
100% data-ink ratio

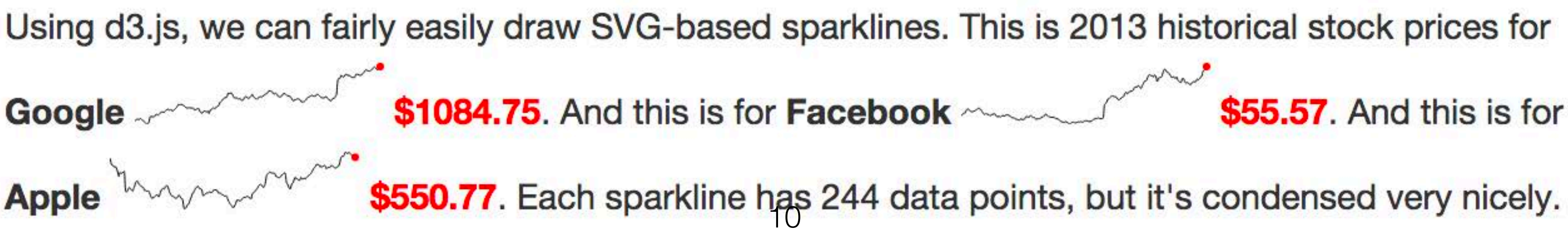


Kenneth A. Kooi, 1971 [T 93].

Sparklines

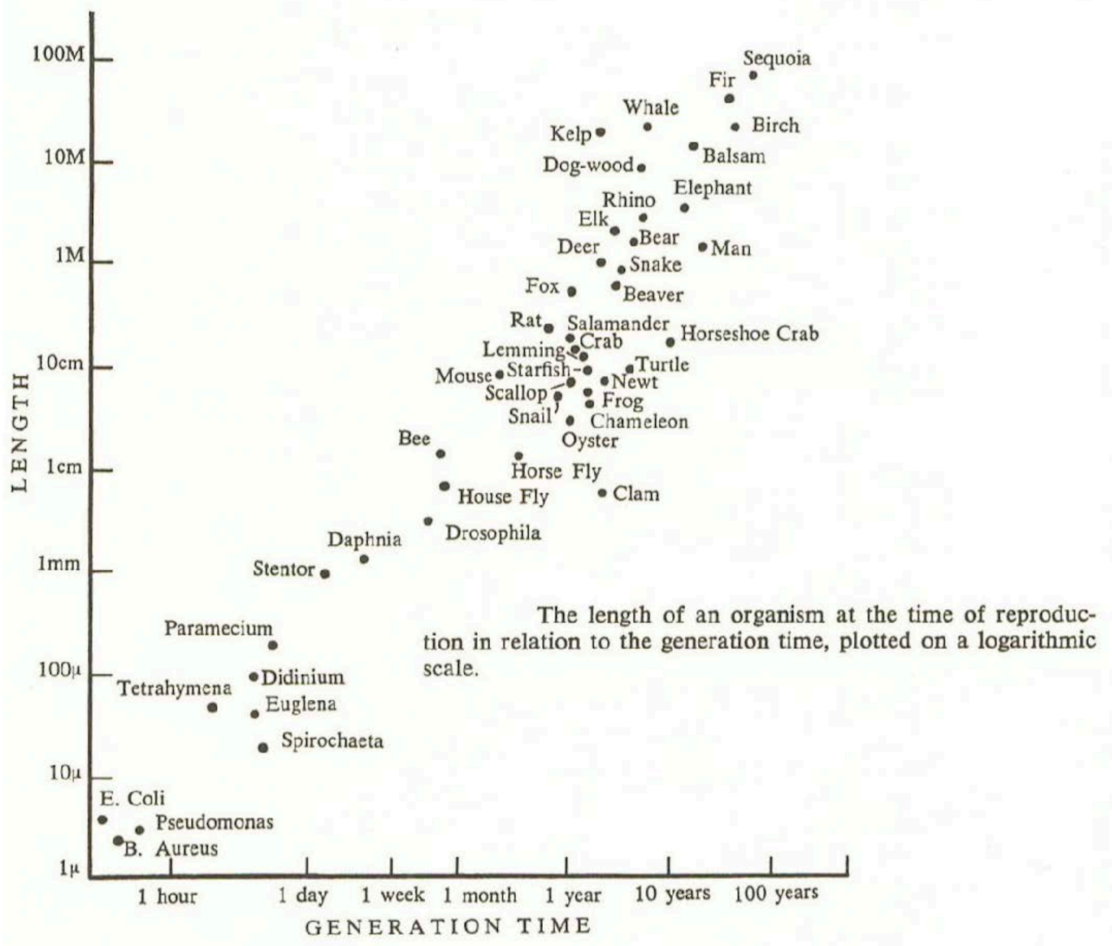
- compact time series
- data-ink ratio = 1
- labels clear from context
- can be used inline with main text




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

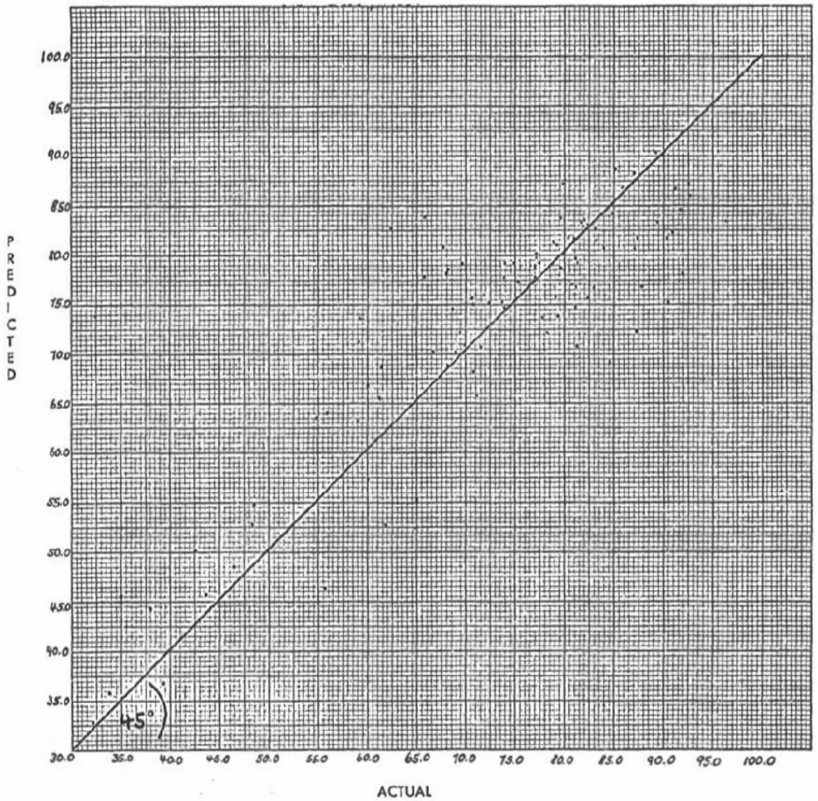
- Most of the ink here is data-ink
    - the dots and labels on the diagonal
- with, 10-20 percent non data-ink
    - the grid ticks and the frame

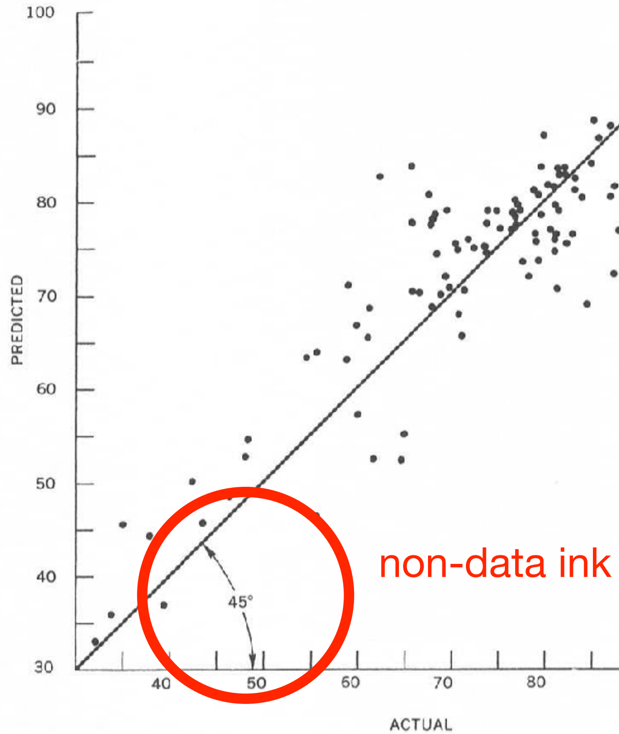


The length of an organism at the time of reproduction in relation to the generation time, plotted on a logarithmic scale.

Improving data-ink ratio

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





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

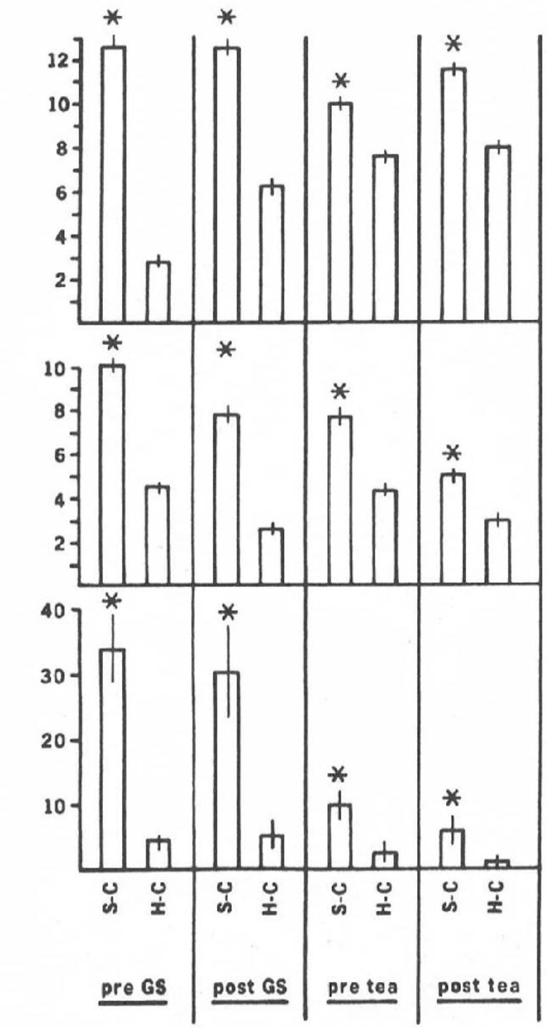
Maximize data-ink

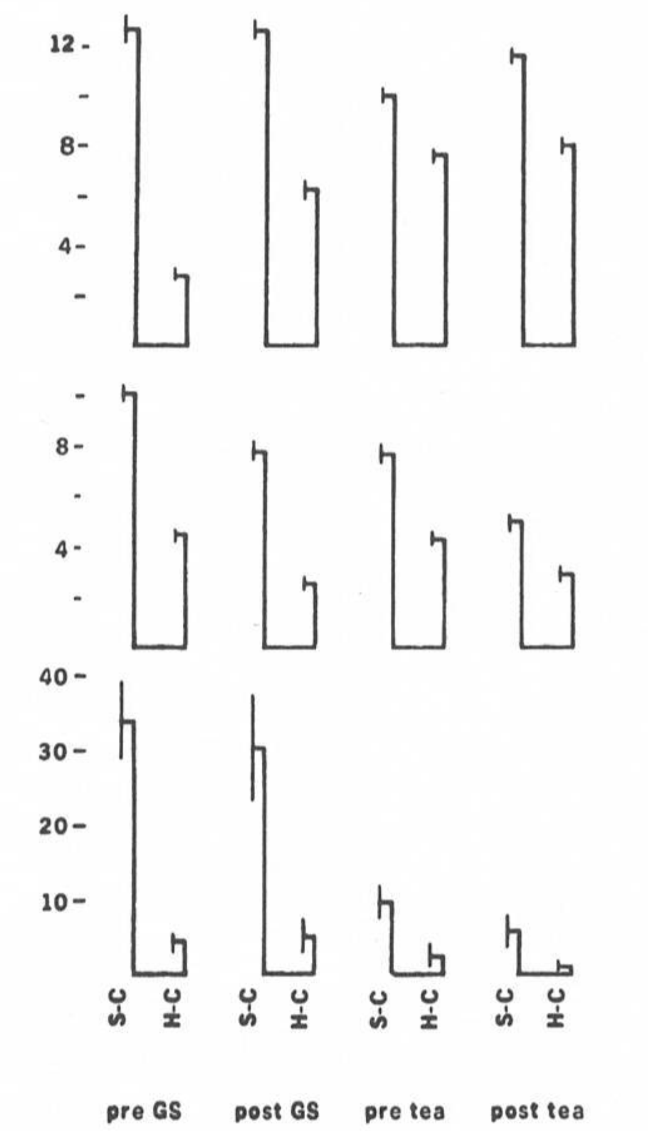
- It is always a good idea to maximize the data-ink ratio, within reason
- The larger the share of data-ink the better, other matters being equal
  - every bit of ink on a graphic needs a reason
  - nearly always that reason being that the ink presents new information
- Ink that fails to depict statistical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

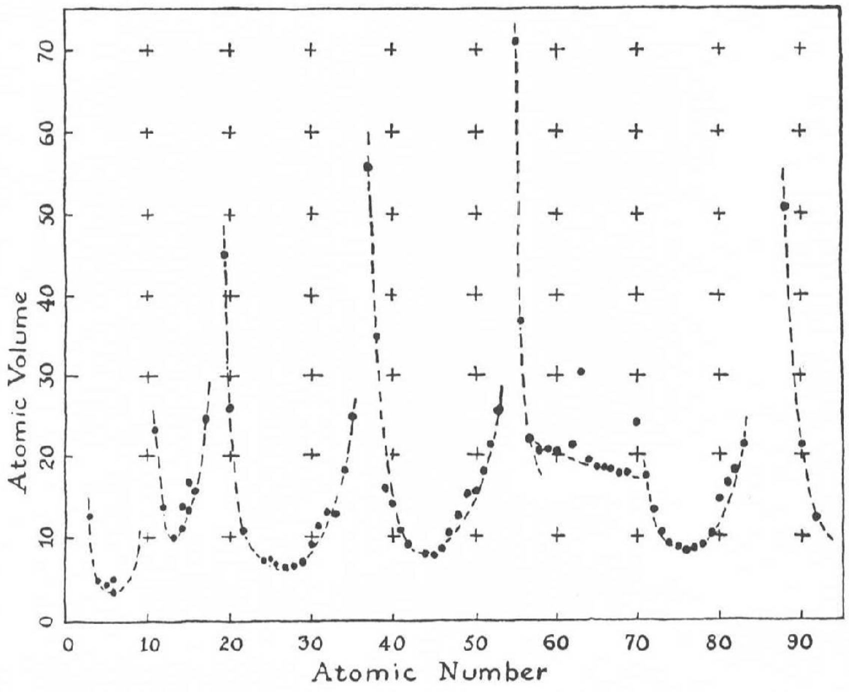
Edit and redesign



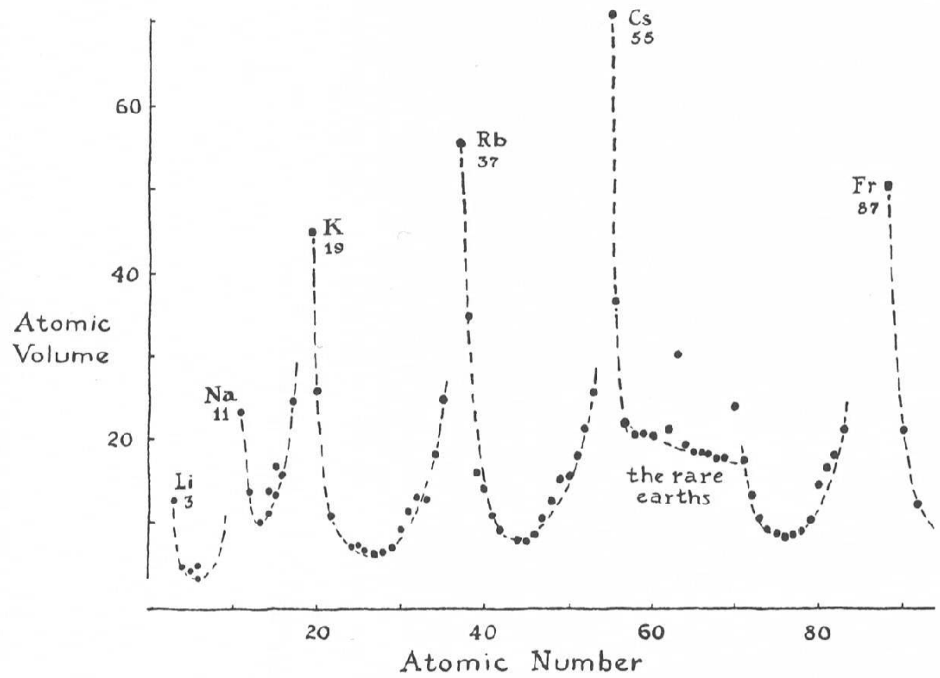


Kuznicki et al. Journal of Experimental Psychology: General, 108 (1979), 76.

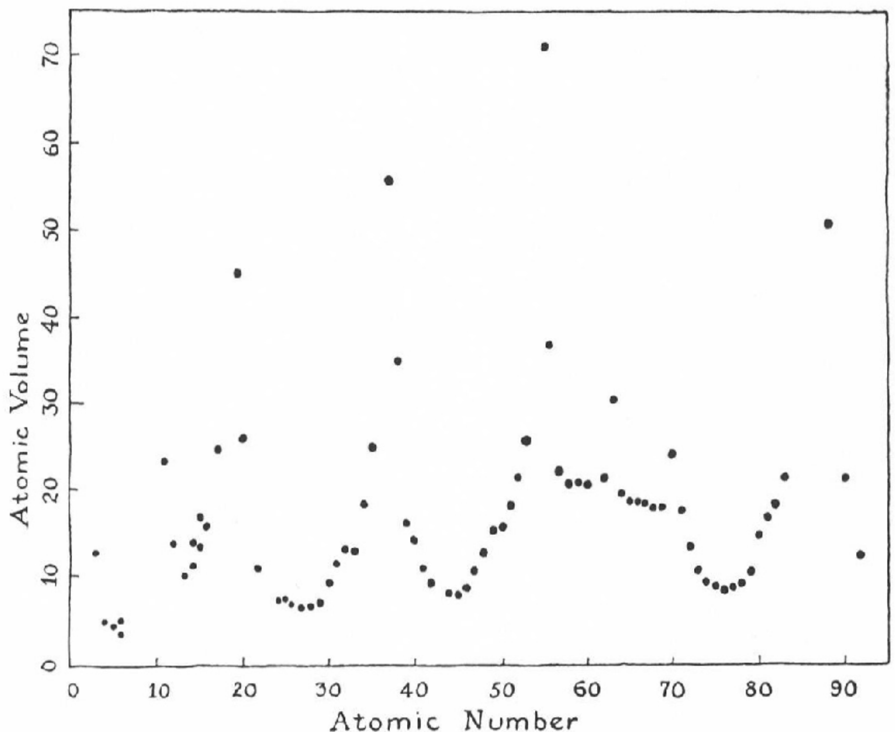
(a)



(b)



(c)

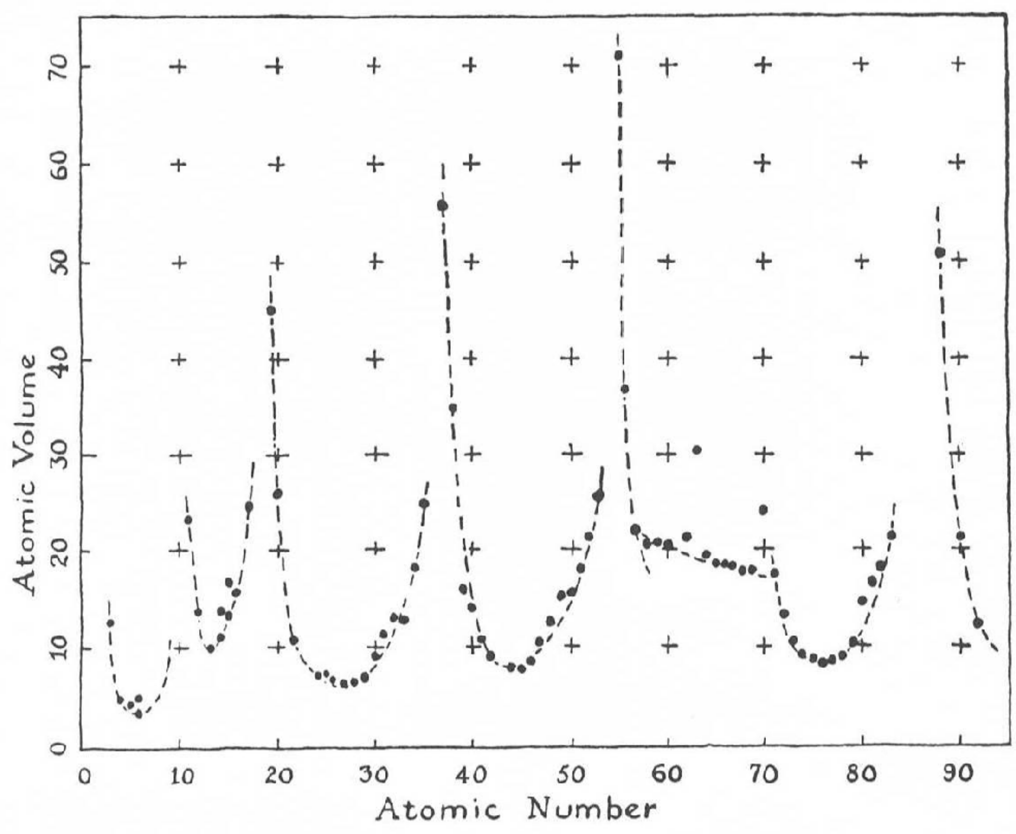


Atomic volume vs. atomic number

Which do you think is the best, and why?

Edit and redesign

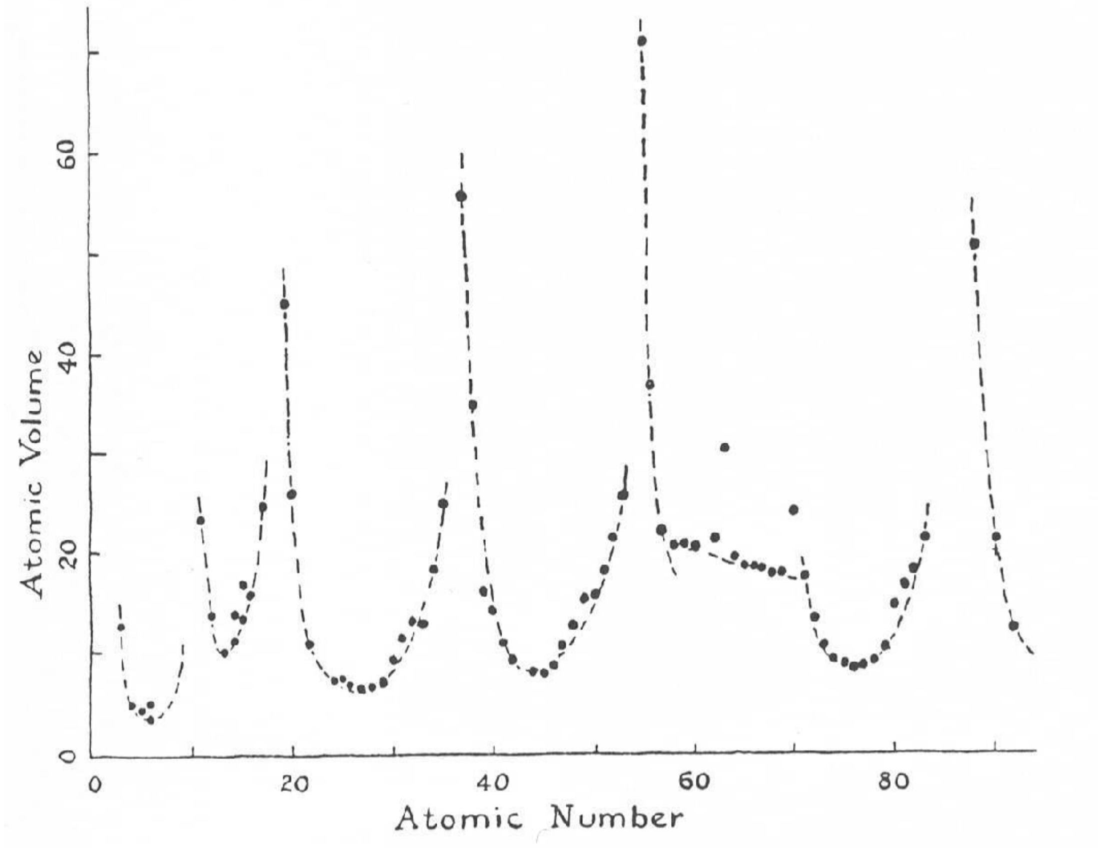
- The data-ink ratio is about 0.6
    - 76 data points and the reference curve are obscured by 63 grid marks
- The grid and part of the frame can be erased to improve the data-ink ratio



Linus Pauling, General Chemistry, p. 64, 1947

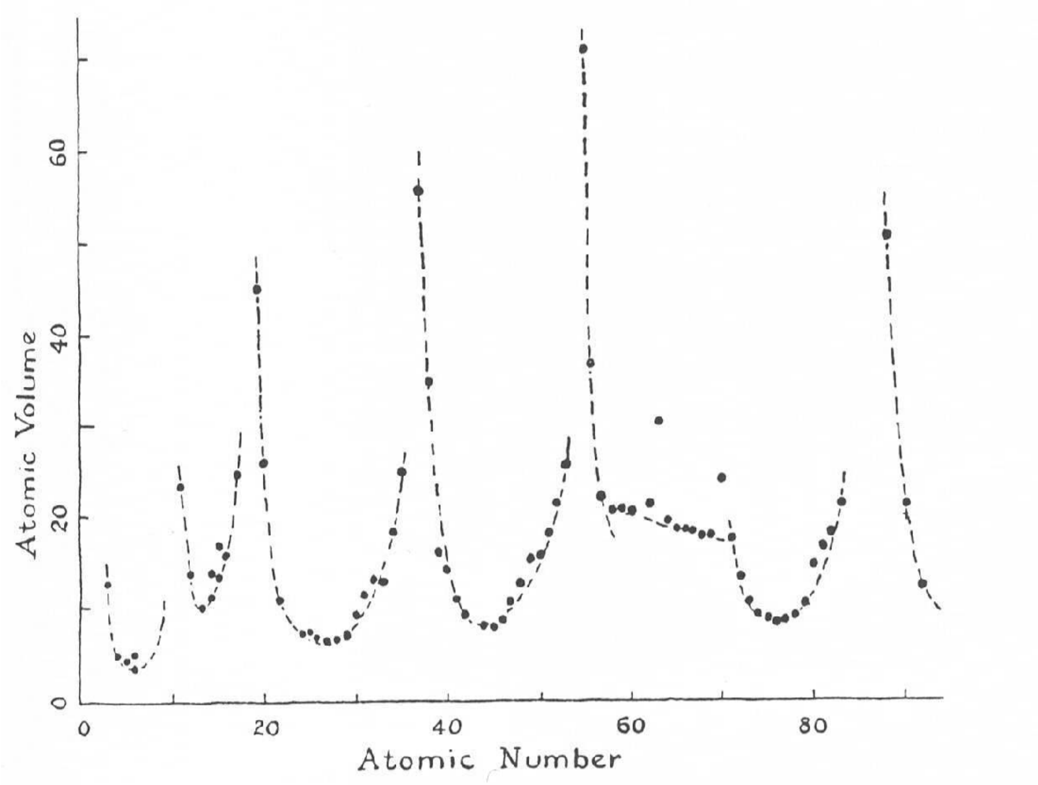
Edit and redesign

- 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

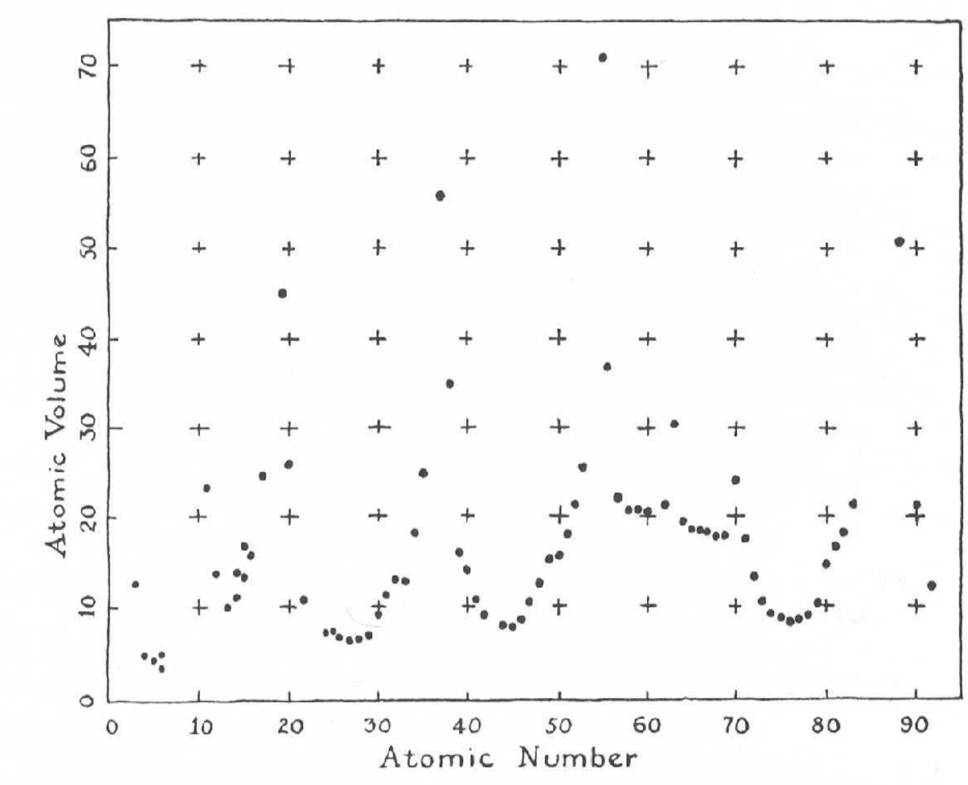


The reference curve is essential in organizing the data, and shows the periodicity (the message) by creating a structure, and by giving ordering and hierarchy

Edit and redesign



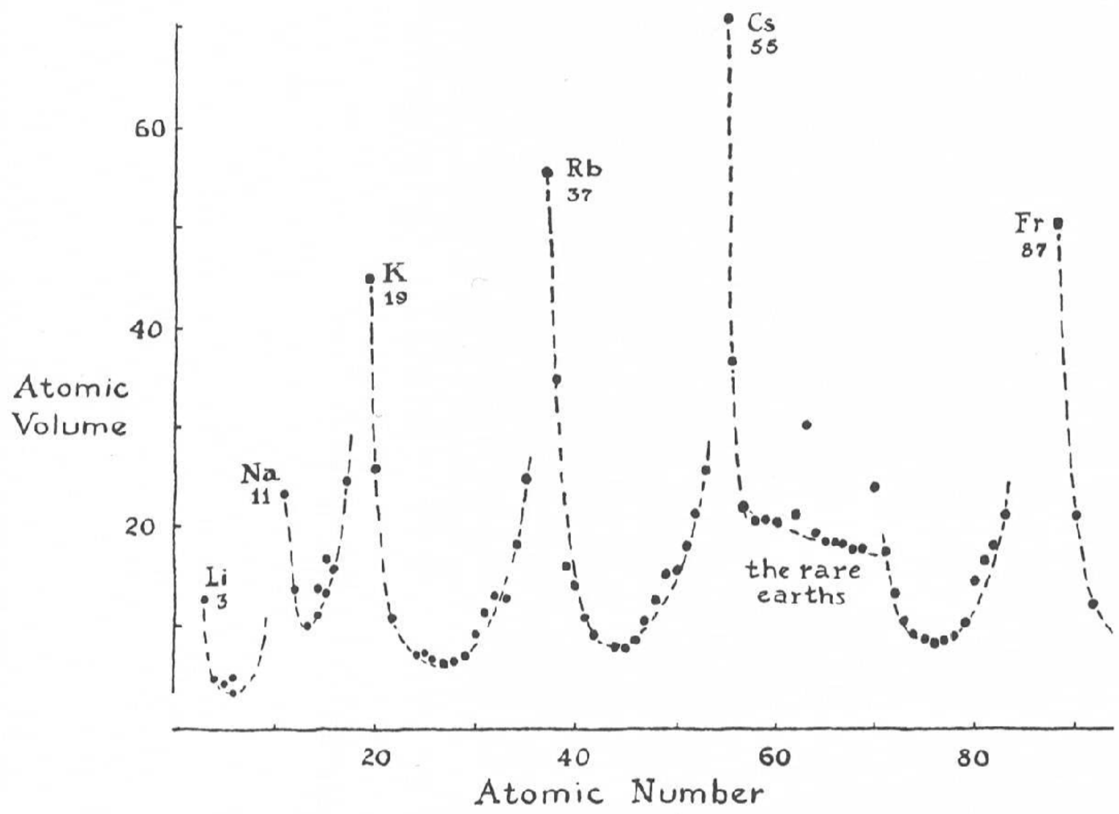
Without the curve we hardly detect the periodicity. The curve becomes necessary because the eye needs guidance



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!



Edit and redesign

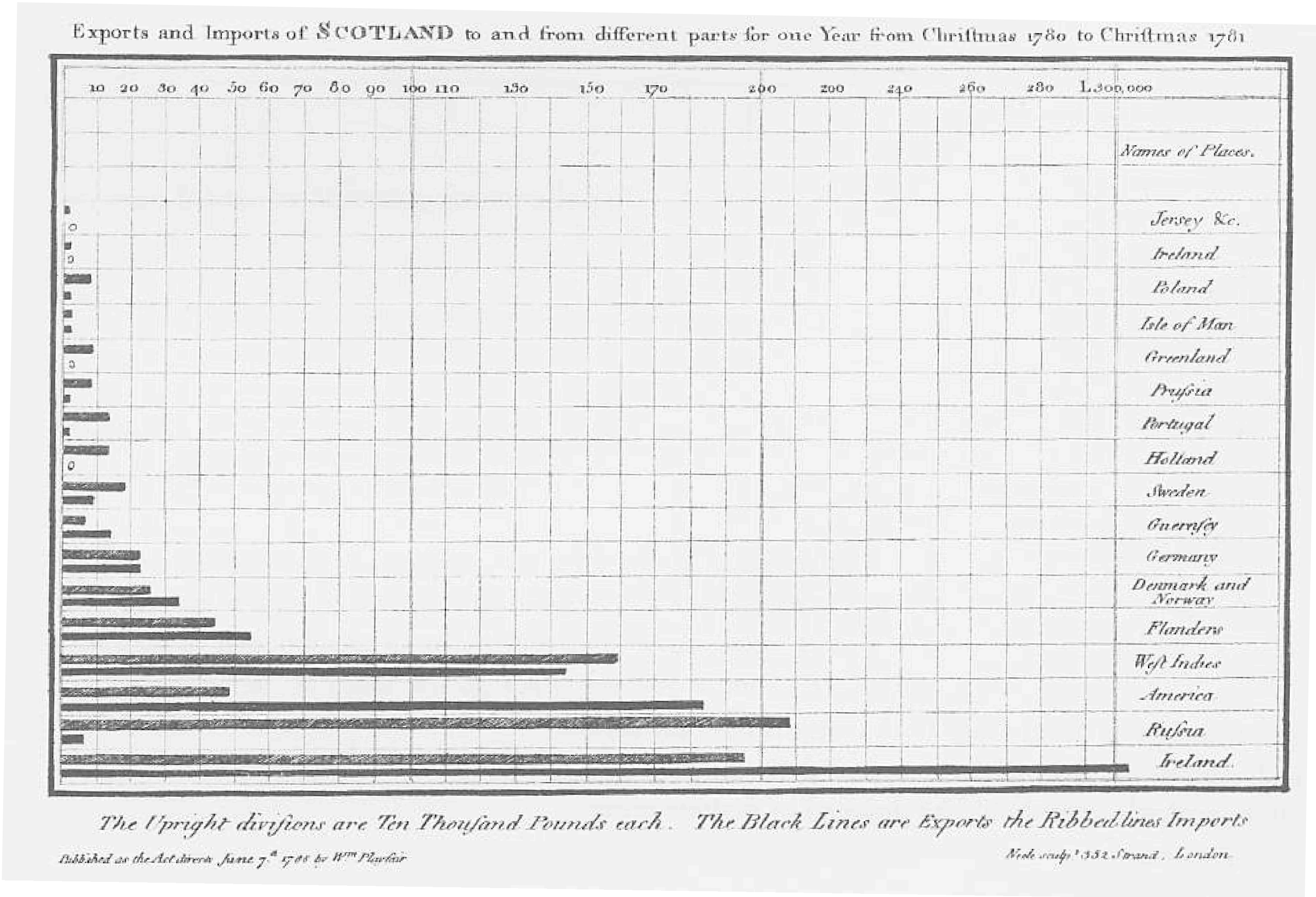
- Example of how to improve standard R scatterplot
- http://www.iki.fi/kaip/p/iris.nb.html

The five principles

- **Above all show the data**
- The larger the share of data-ink, the better (all other things being equal): **Maximize the data-ink ratio, within reason.**
- Maximizing the data-ink ratio implies minimizing the amount of non-data ink:
  - **Erase non-data-ink, within reason.**
  - **Erase redundant data-ink.**
- **Revise and edit.**

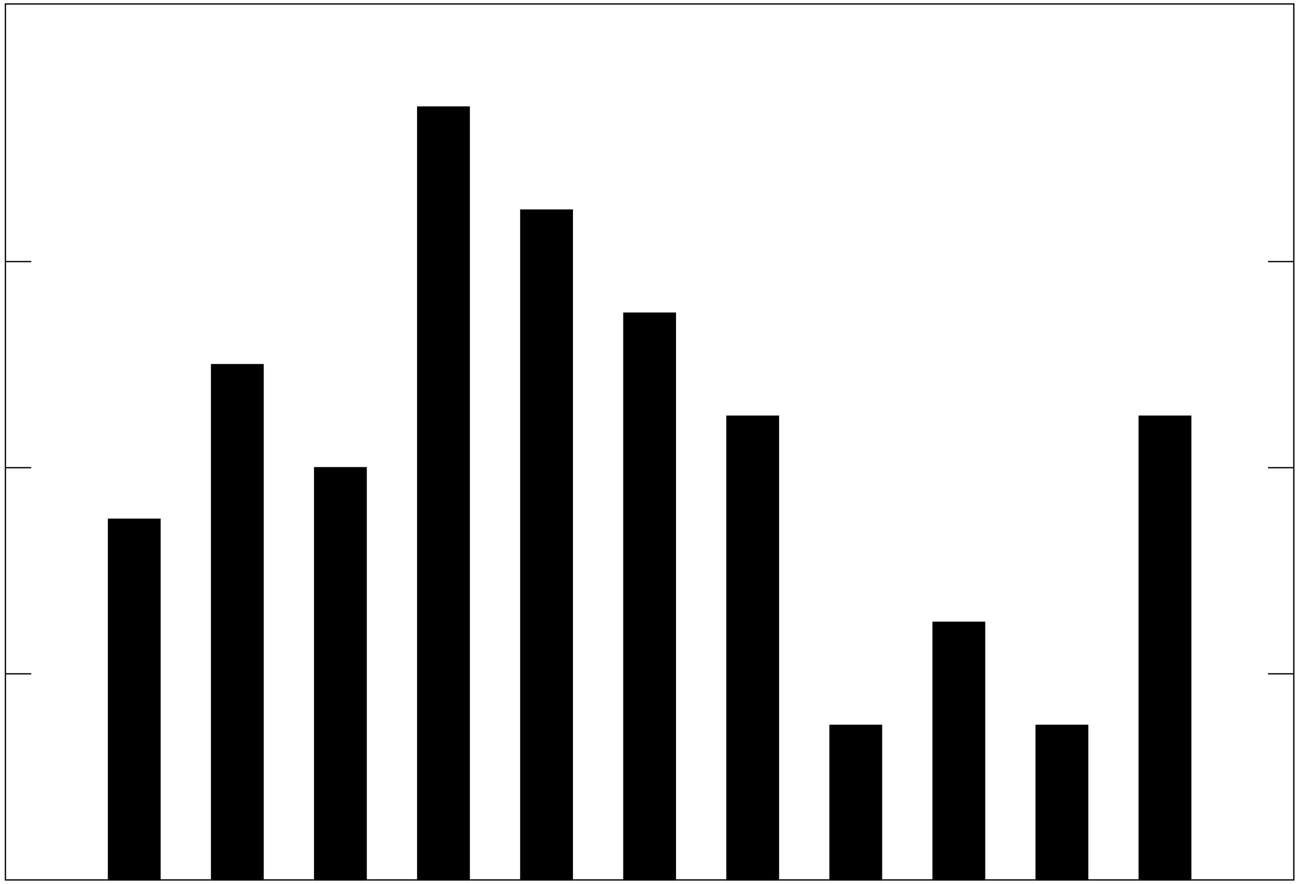
Bar chart

- One of the basic designs.



William Playfair, 1786.

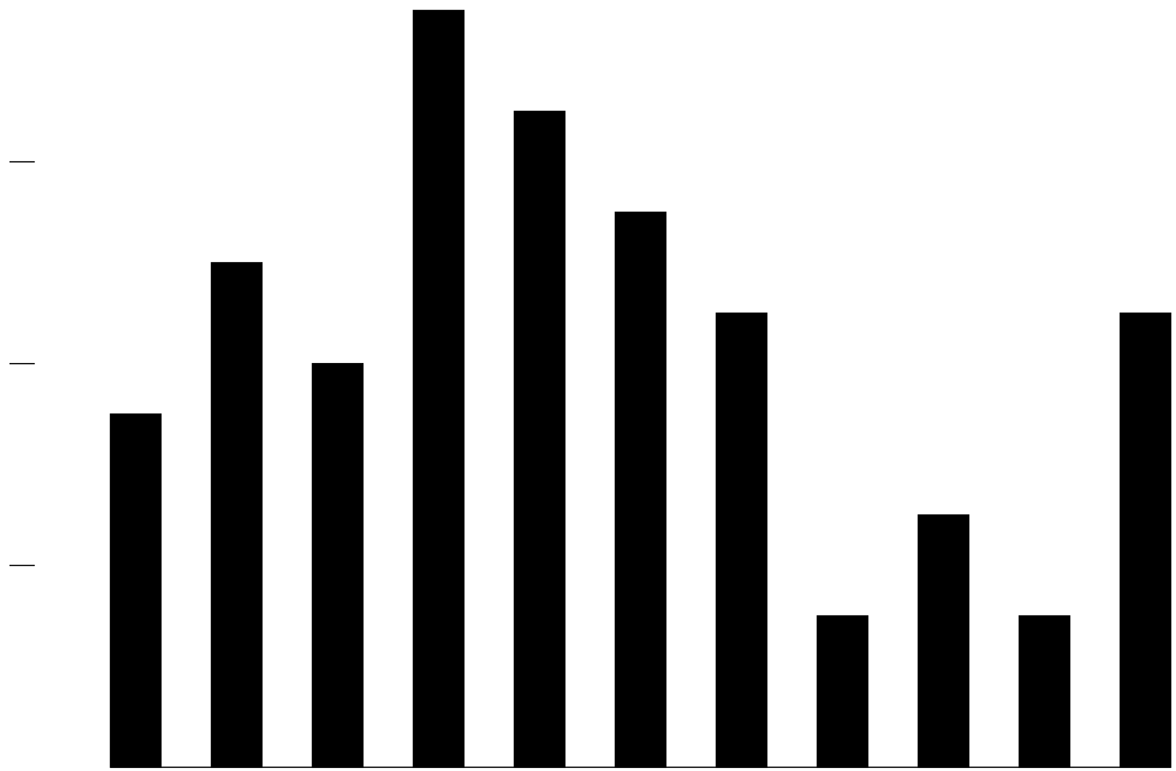
Bar chart



The standard bar chart

Bar chart

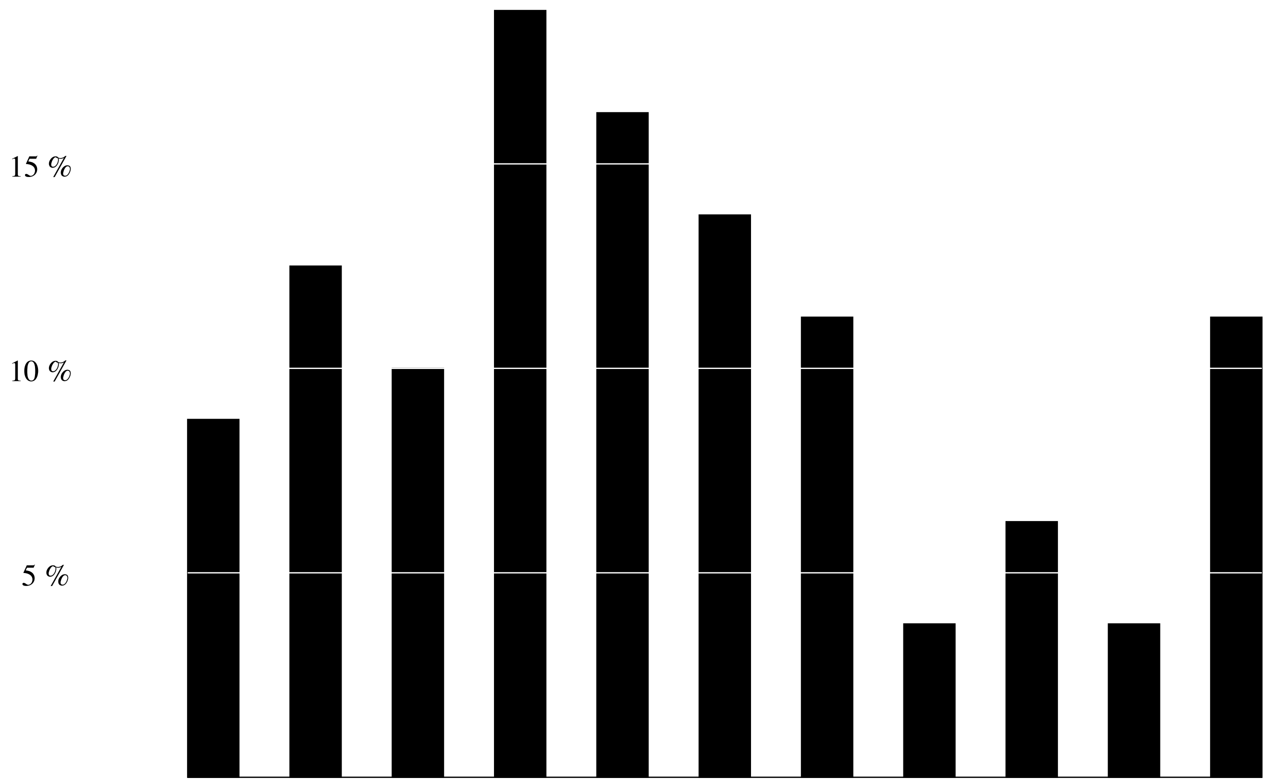
The box and vertical axis can be *erased*:



The ticks are needed to show the coordinate lines. Or are they?

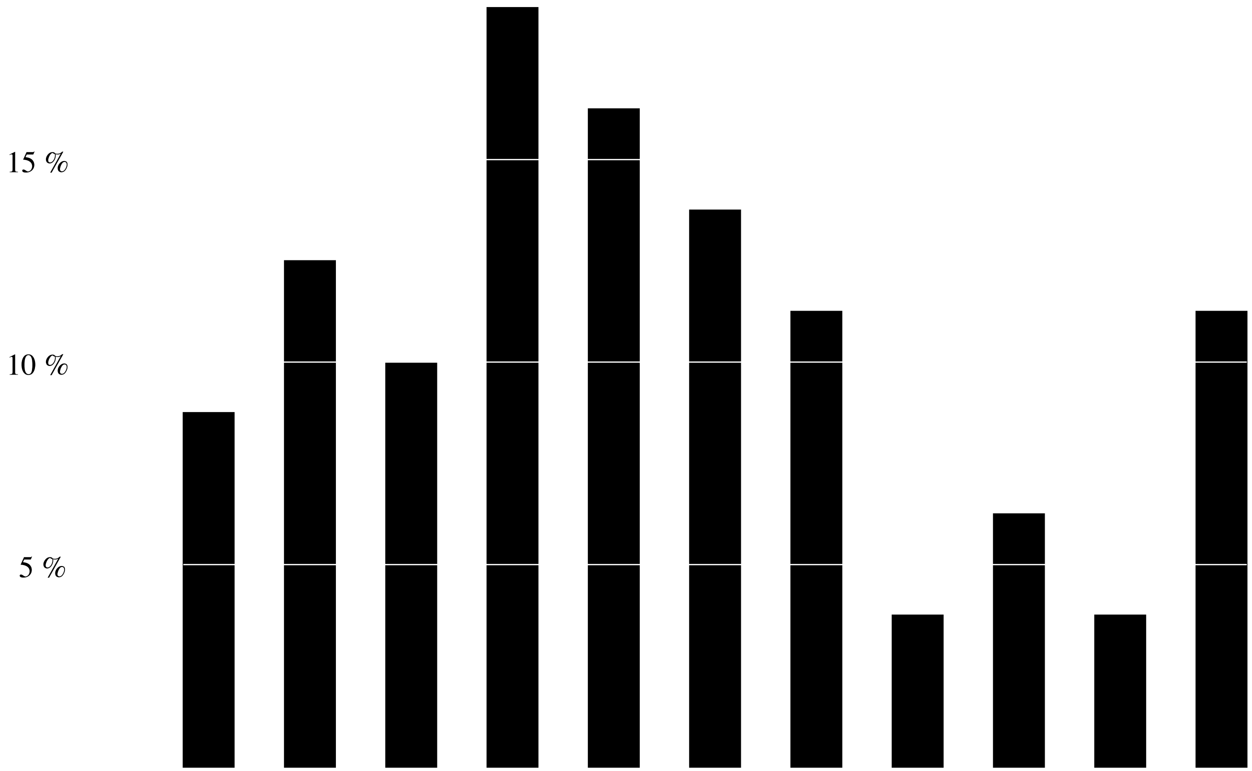
Bar chart

The white lines show the coordinate lines more precisely than ticks, which are no longer needed:



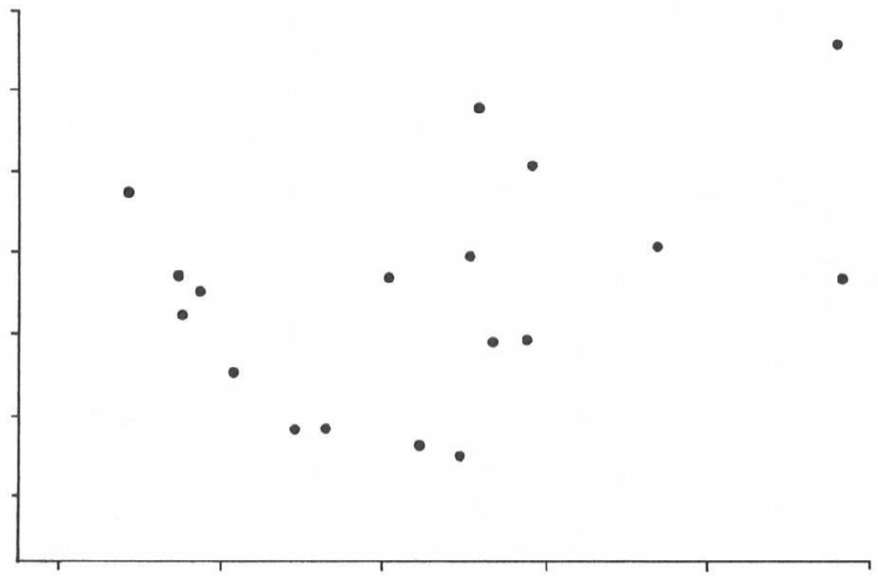
Bar chart

We could still erase the base line since the bars define the end-point at the bottom:

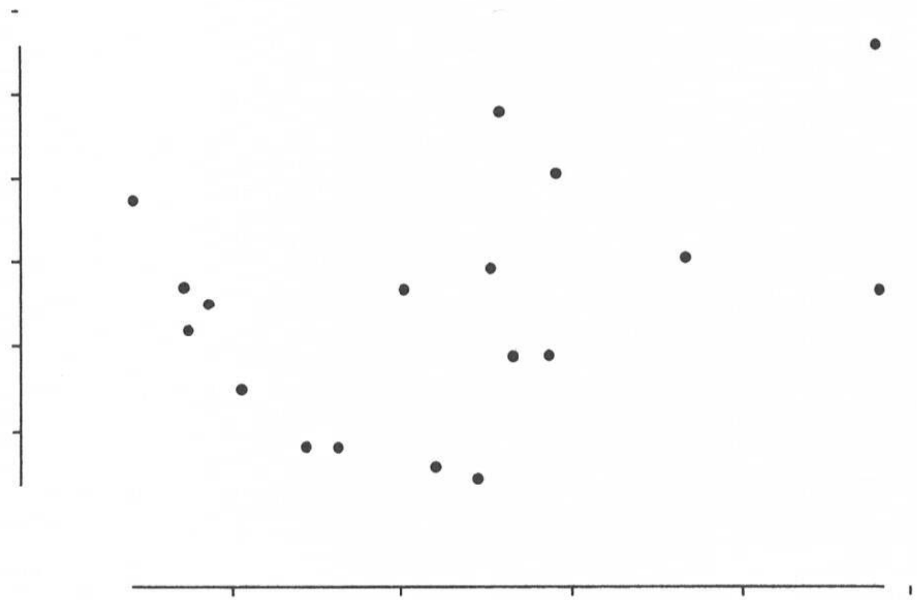


This might however be overdoing it (thin baseline looks good).

Scatterplot



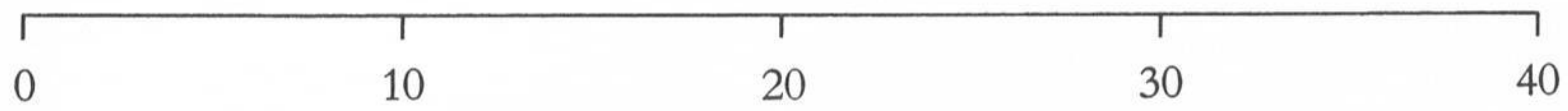
standard bivariate scatterplot



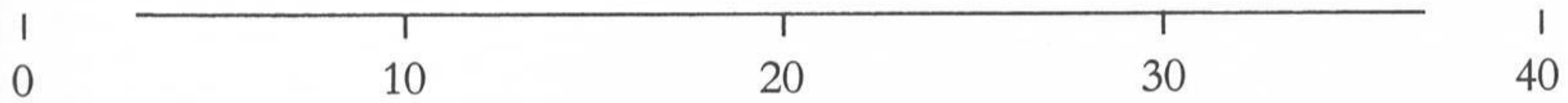
range frames indicate minimum and maximum values

Range frames

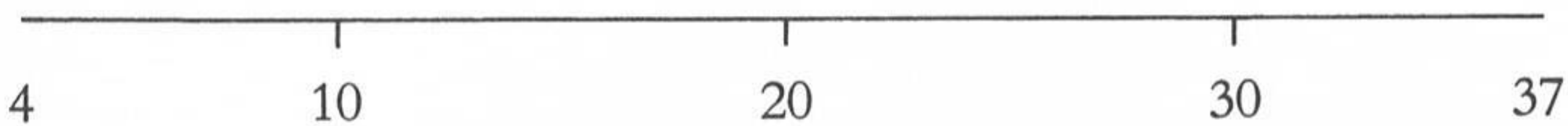
Standard frame:



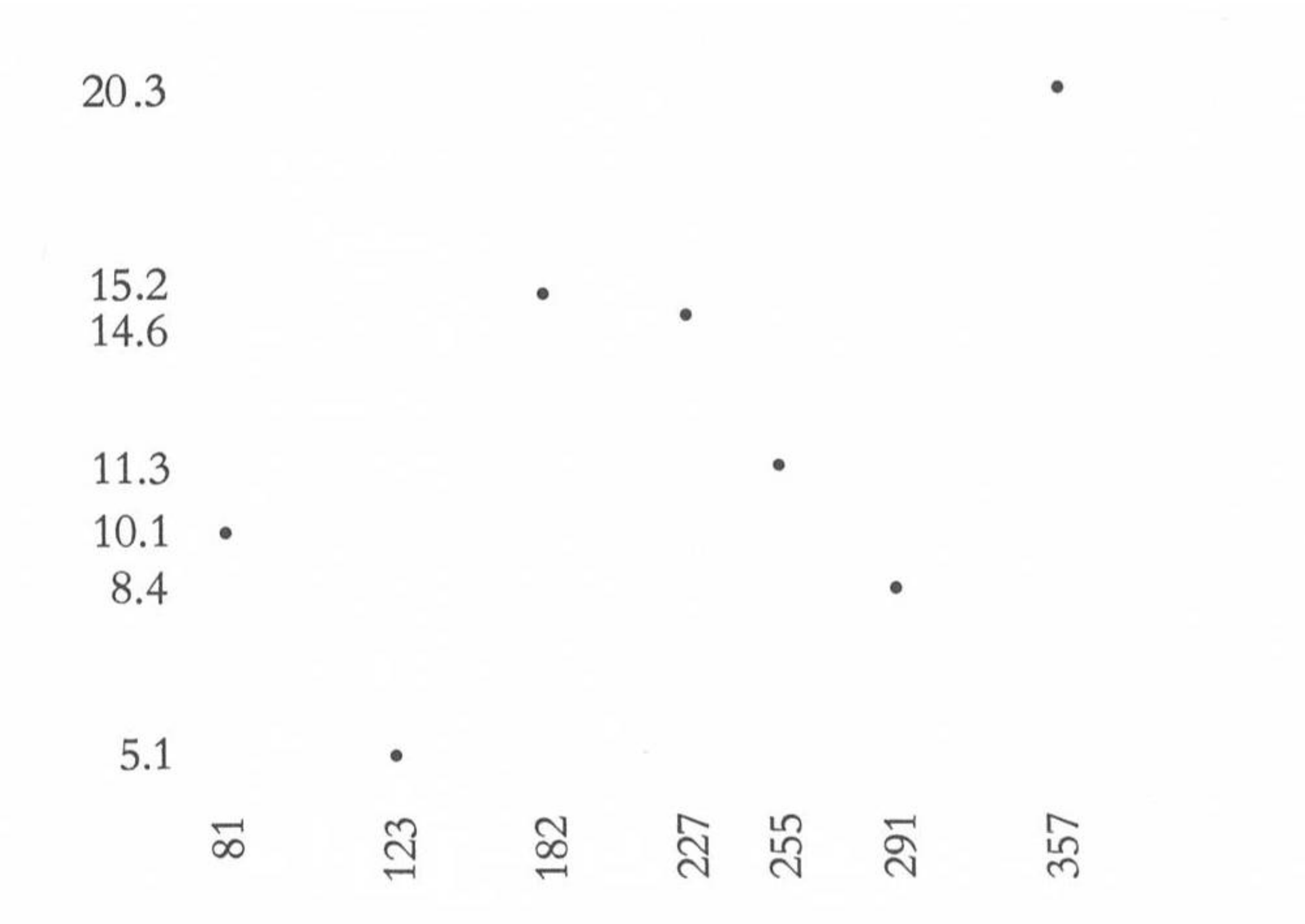
Range frame:



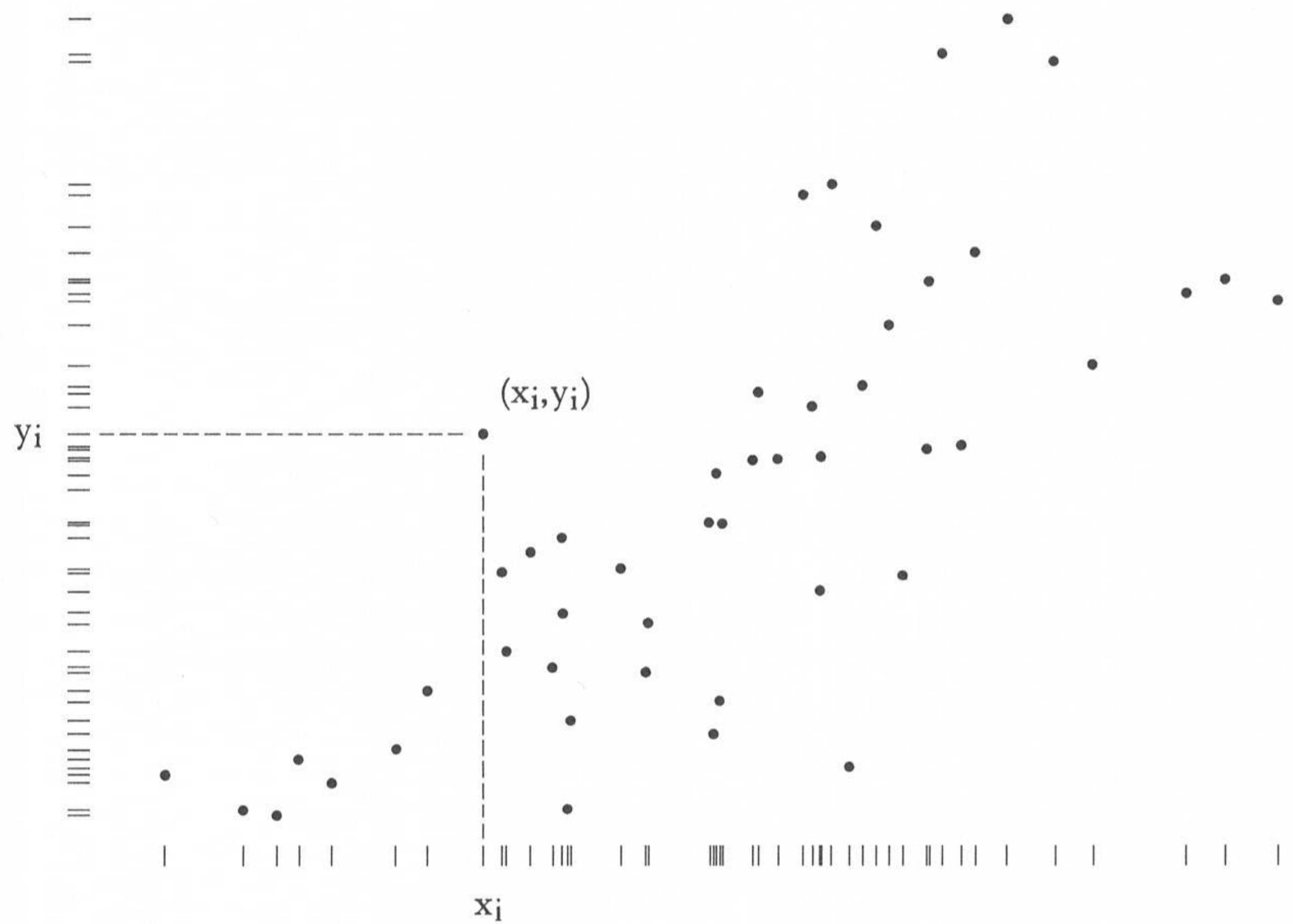
Range frame with explicit limits (may work better):



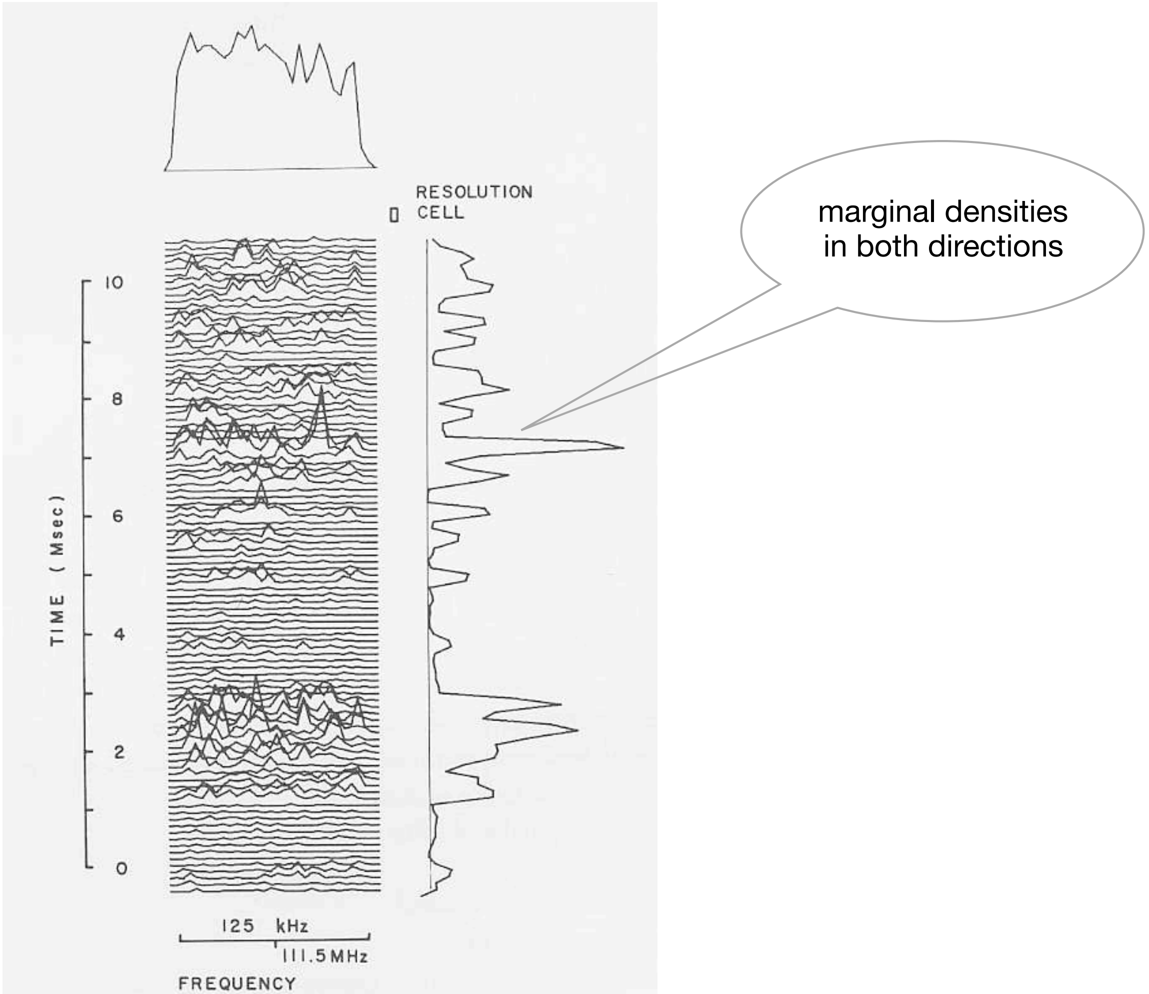
Data values as frame



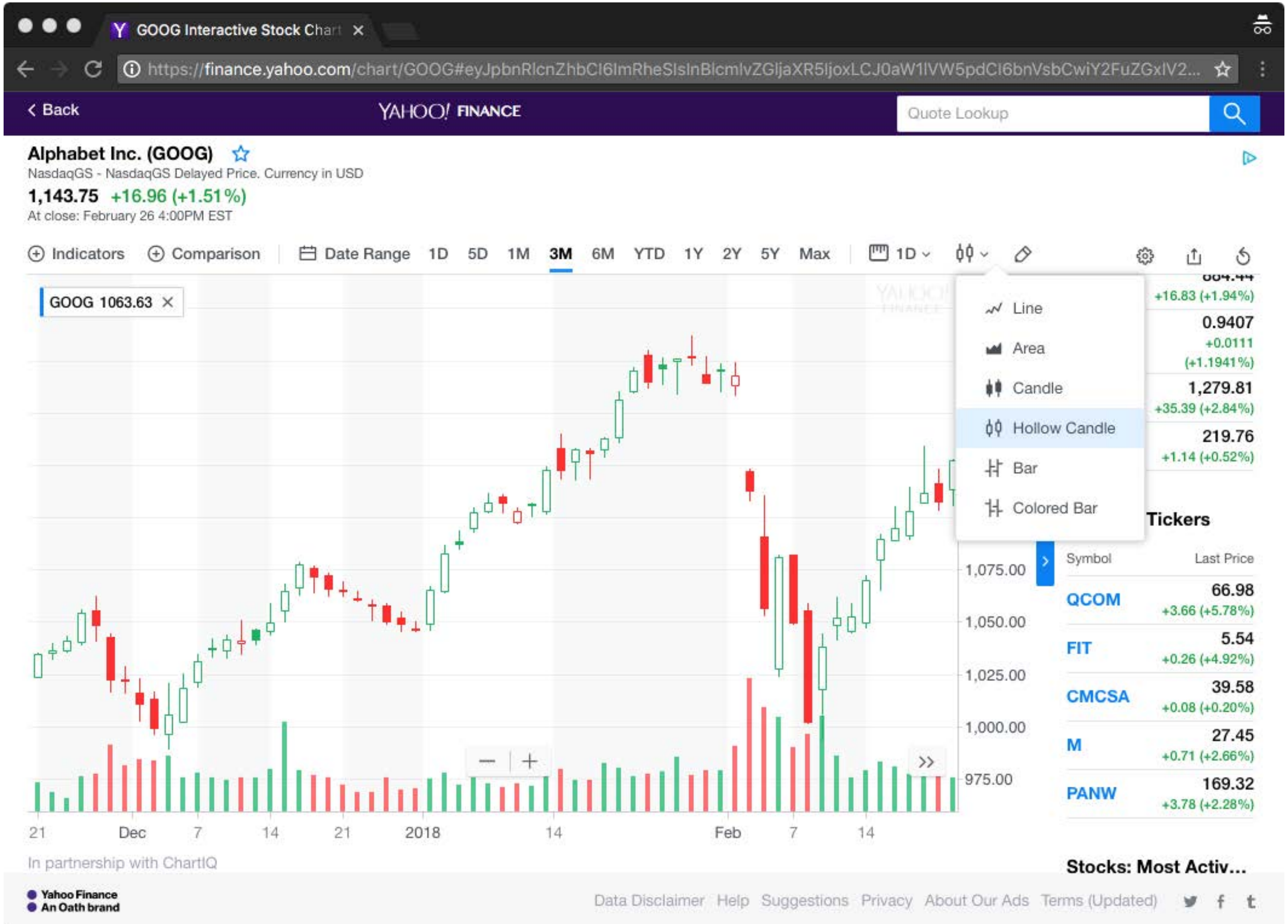
Rug plot as frame

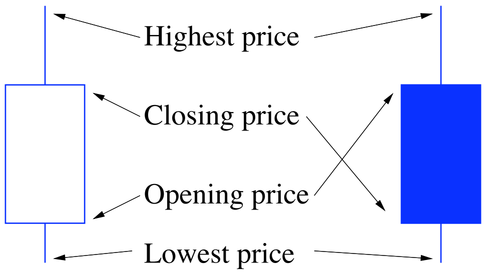


Dot-dash plot

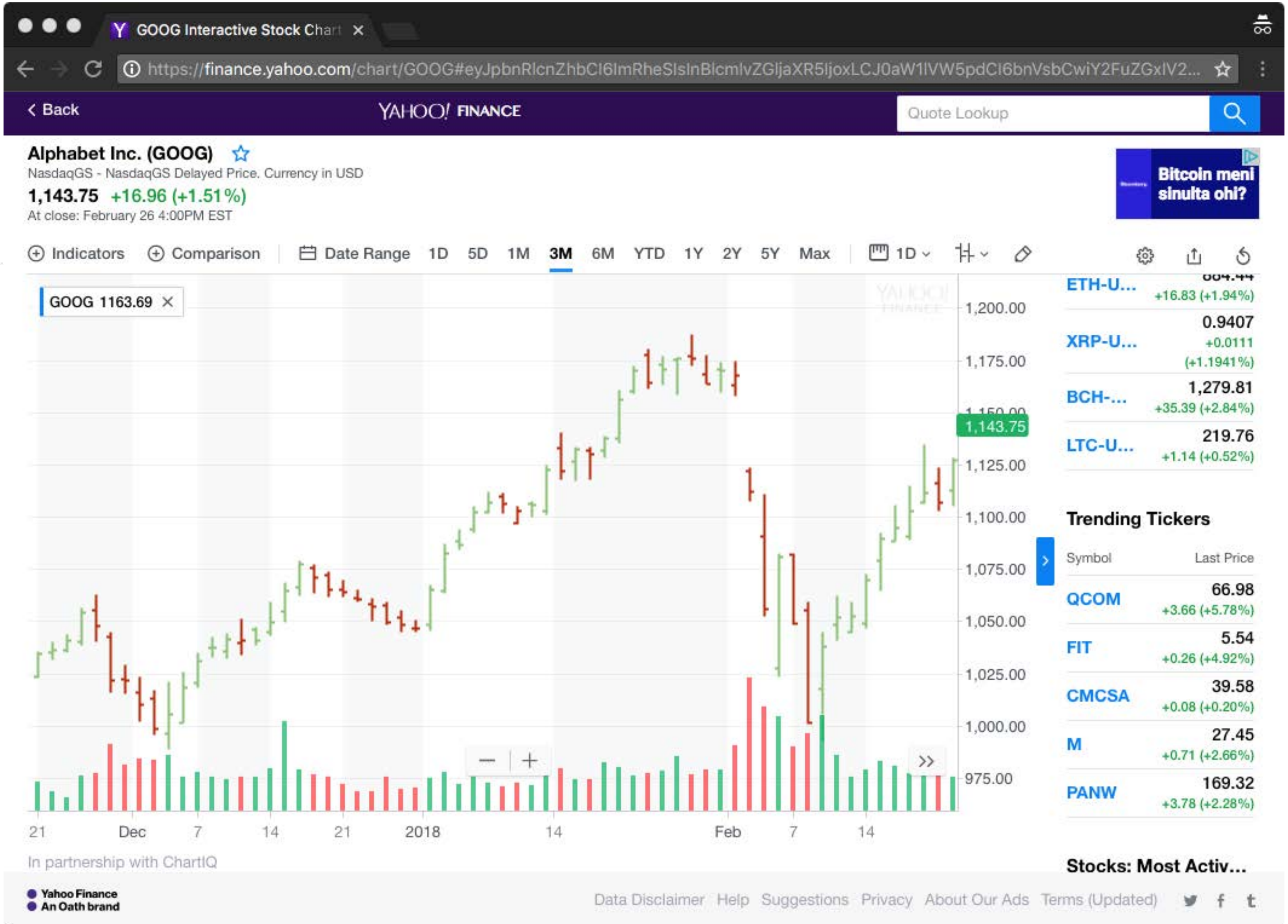


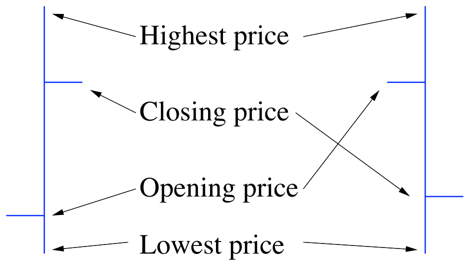
Timothy H. Hankins, Barney J. Rickett, 1975 [T 134].





Range bars show opening and closing prices and price variation


GOOG Interactive Stock Chart
https://finance.yahoo.com/chart/GOOG#eyJpbnRlcnZhbCI6ImRheSIsInBlcmlvZGljaXR5IjoxLCJ0aW1lVW5pdCI6bnVsbCwiY2FuZGxlV2...
Back
YAHOO! FINANCE
Quote Lookup
Alphabet Inc. (GOOG)
NasdaqGS - NasdaqGS Delayed Price. Currency in USD
1,143.75 +16.96 (+1.51%)
At close: February 26 4:00PM EST
Bitcoin meni sinulta ohi?
Indicators
Comparison
Date Range
1D
5D
1M
3M
6M
YTD
1Y
2Y
5Y
Max
1D
GOOG 1163.69
1,200.00
1,175.00
1,143.75
1,125.00
1,100.00
1,075.00
1,050.00
1,025.00
1,000.00
975.00
21
Dec
7
14
21
2018
14
Feb
7
14
In partnership with ChartIQ
ETH-U...
+16.83 (+1.94%)
XRP-U...
0.9407
+0.0111
(+1.1941%)
BCH-...
1,279.81
+35.39 (+2.84%)
LTC-U...
219.76
+1.14 (+0.52%)
Trending Tickers
Symbol
Last Price
QCOM
66.98
+3.66 (+5.78%)
FIT
5.54
+0.26 (+4.92%)
CMCSA
39.58
+0.08 (+0.20%)
M
27.45
+0.71 (+2.66%)
PANW
169.32
+3.78 (+2.28%)
Stocks: Most Activ...
Yahoo Finance
An Oath brand
Data Disclaimer
Help
Suggestions
Privacy
About Our Ads
Terms (Updated)


Highest price
Closing price
Opening price
Lowest price

The five principles
(on data-ink)

- **Above all show the data**
- The larger the share of data-ink, the better (all other things being equal): **Maximize the data-ink ratio, within reason.**
- Maximizing the data-ink ratio implies minimizing the amount of non-data ink:
  - **Erase non-data-ink, within reason.**
  - **Erase redundant data-ink.**
- **Revise and edit.**

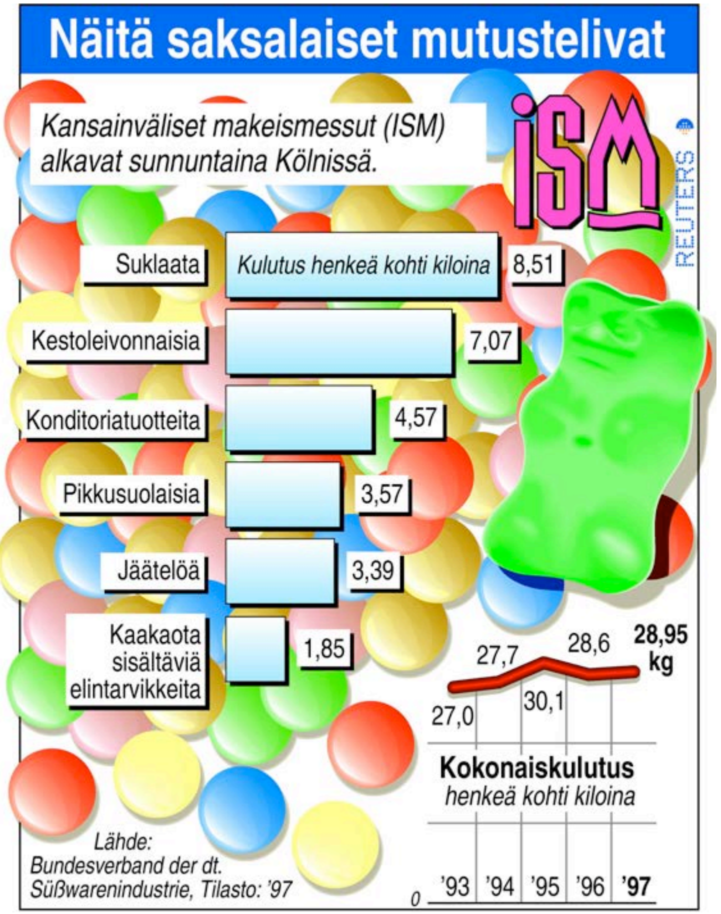
Chartjunk

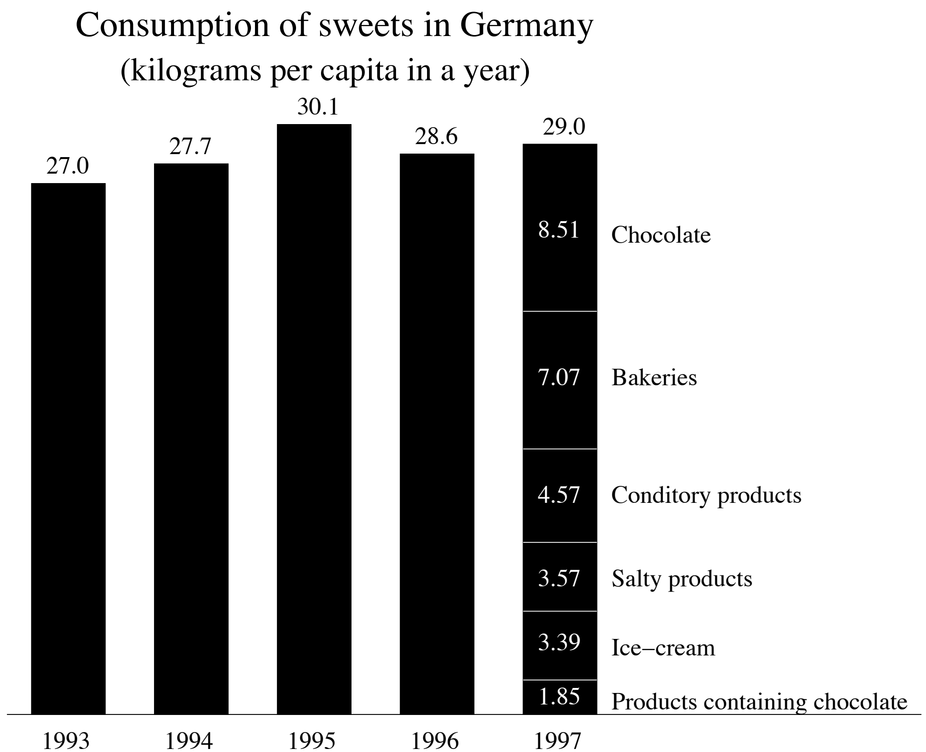
- *Chartjunk* is the interior decoration of graphics that does not tell the viewer anything new
- The purpose of the chartjunk may be to
  - make the graphics appear more scientific and precise (grid lines, excess ticks, redundant representations of simple data etc.)
  - decorate the graphics
  - make the data appear more lively

Types of chartjunk

- Ducks (eye candy and self-promoting graphics)
- Vibrations
- Grids

Eye candy





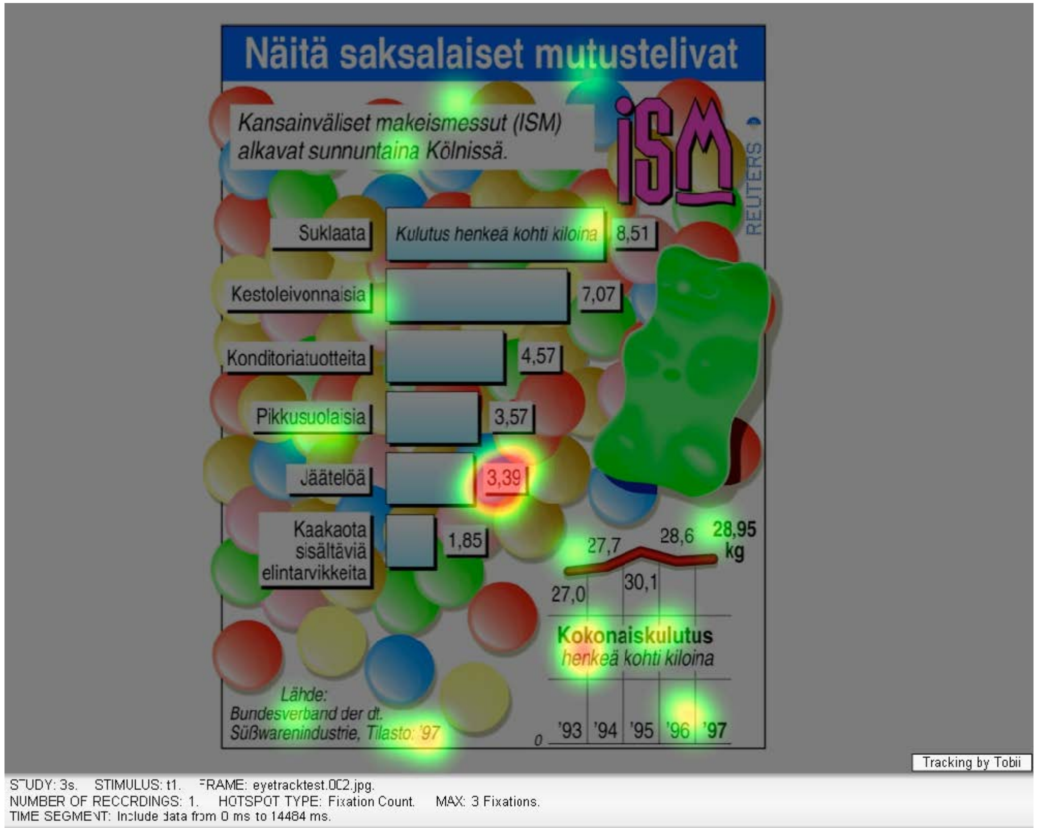
H. Spissler, Reuters, 1999.

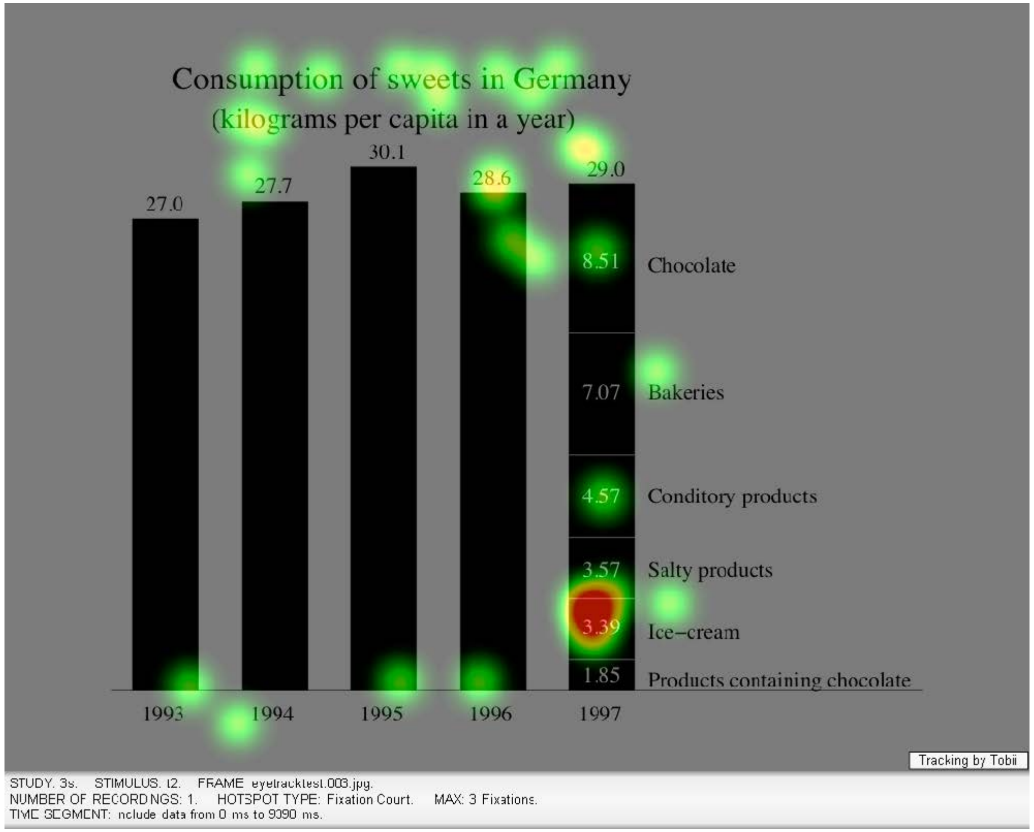
How much ice cream did a German eat in 1997?

How much ice cream did a German eat in 1997?

Eye candy

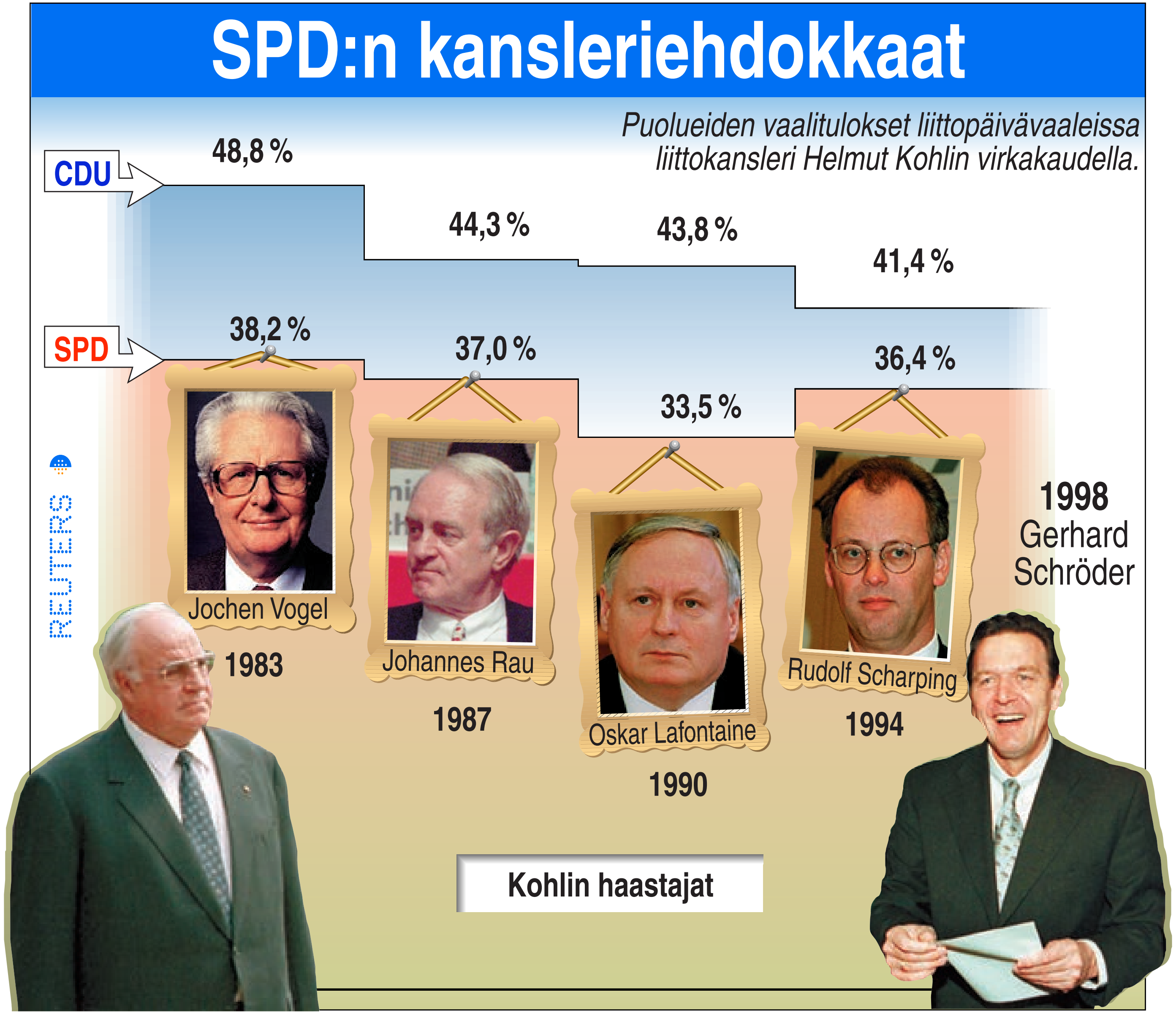
Where did they really look:





Differences: more unnecessary fixations (wasted time, missed information) with chartjunk? (I do not know if this is true in general.)

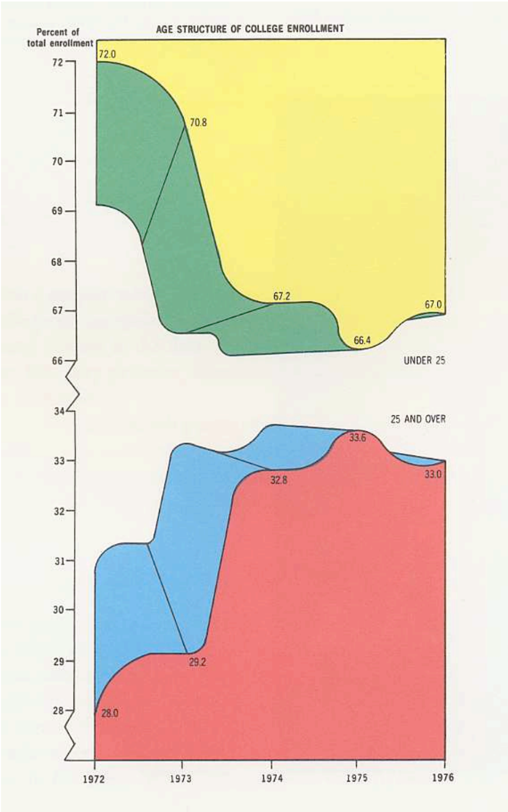
Facejunk



Reuters/H. Spissler

Self-promoting graphics

The graphics becomes *self-promoting* when the graphical style takes precedence over data structures.



American education [T 118].

The above chart could have been represented by a table of five numbers.

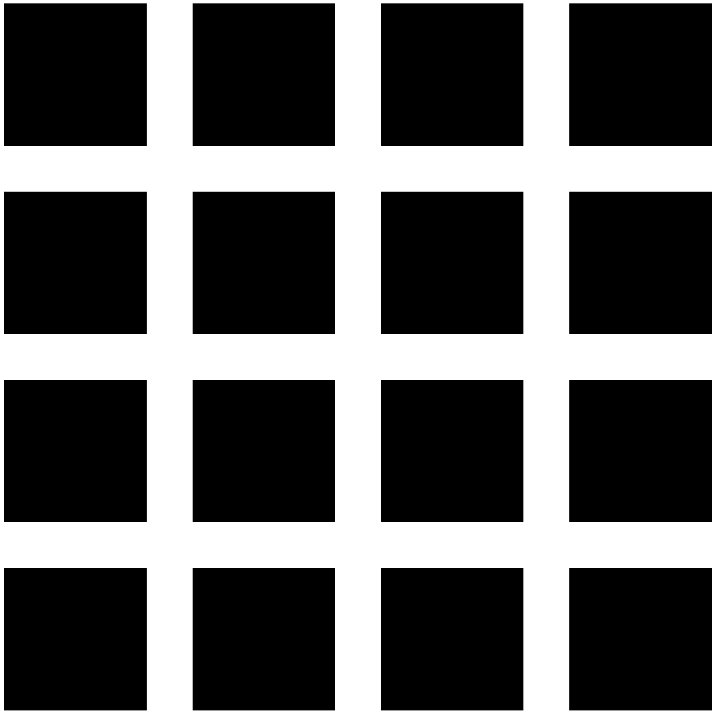
Visual stress

(vibrations)

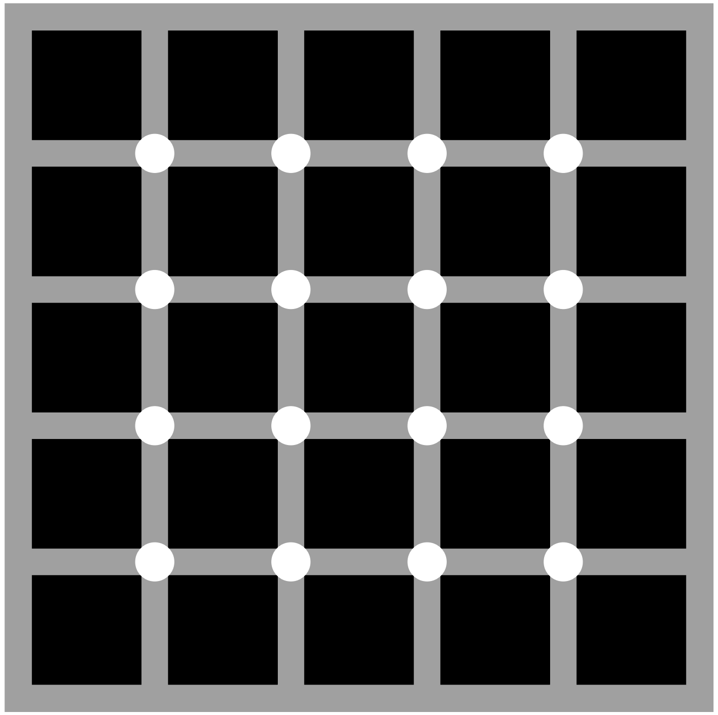
- Striped patterns cause visual stress in most people.
- The following combination is most potent:
  - about 3 cycles per degree
  - flicker rate of about 20 Hz
  - large patterns

op-art by Bridget Riley

Optical effects



Hermann Grid illusion

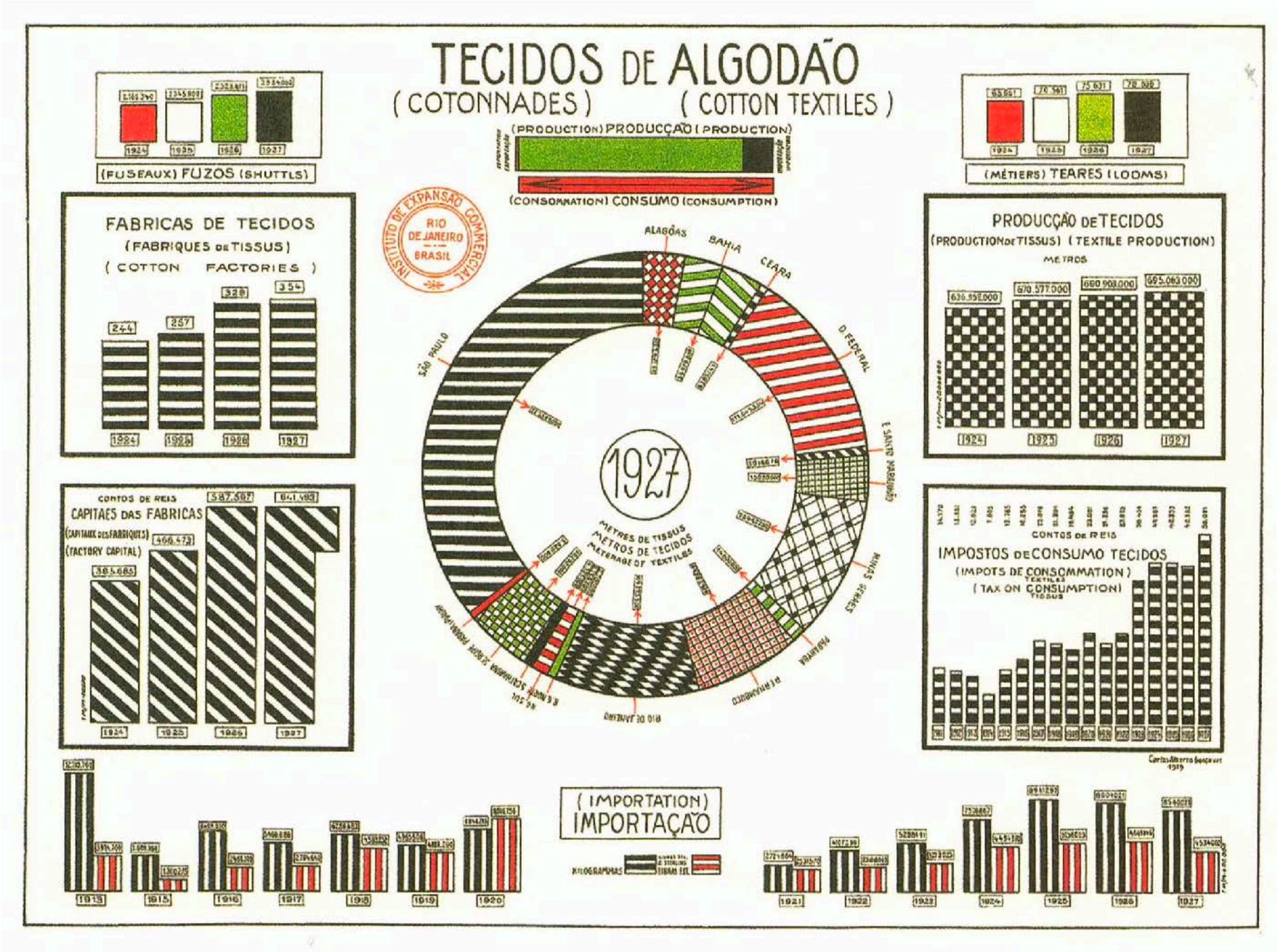


Modified Hermann Grid illusion

There appear to be dark spots at the intersection of the bright lines. Similar effects can appear in data graphics.

Moirè effects

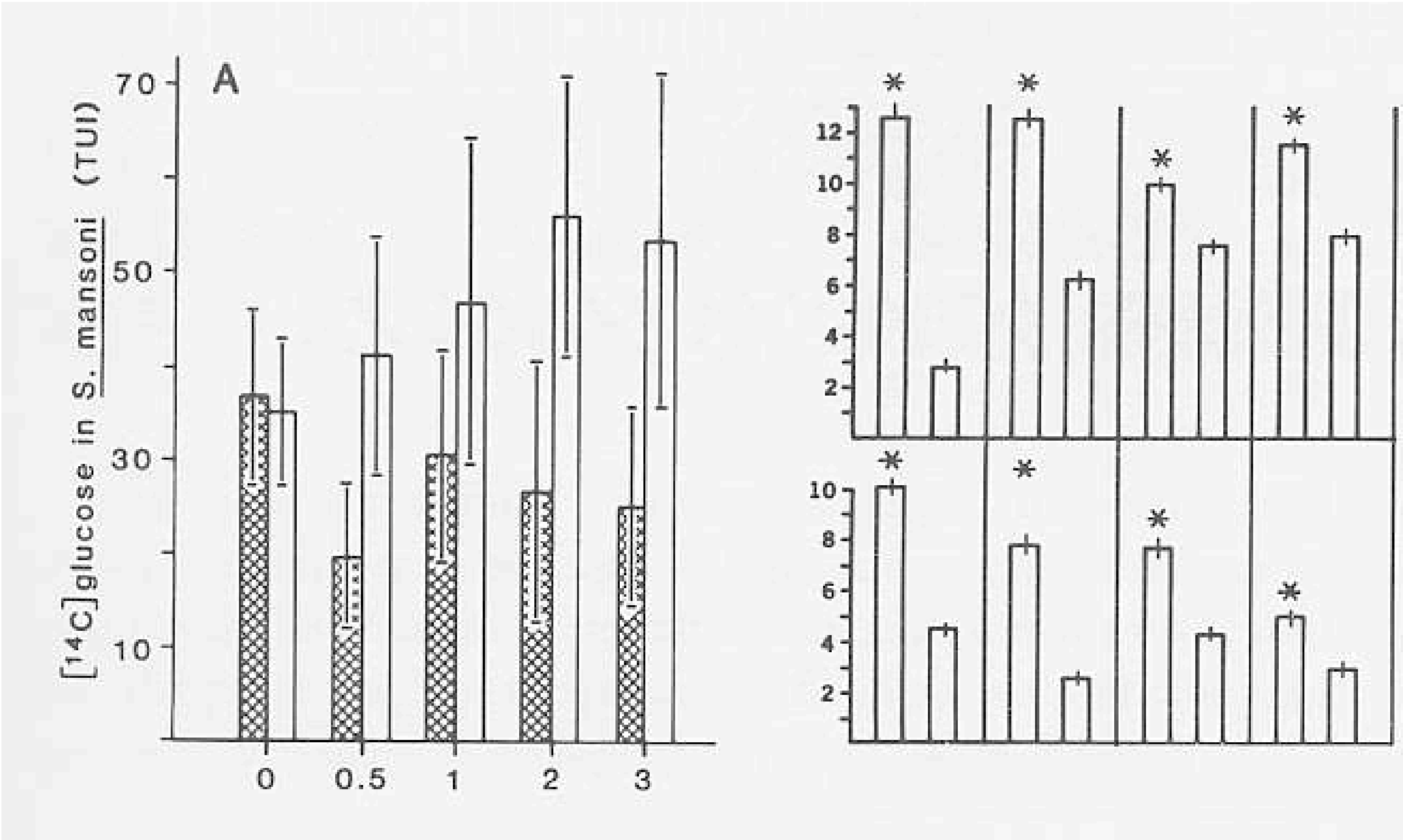
The vibration caused by repeating lines and optical effects are called *Moirè effects*.



Instituto de Expansao Commercial, Brazil, 1929 [T 108].

Moirè effects

Moirè vibration appears at a maximum for equally spaced bars:



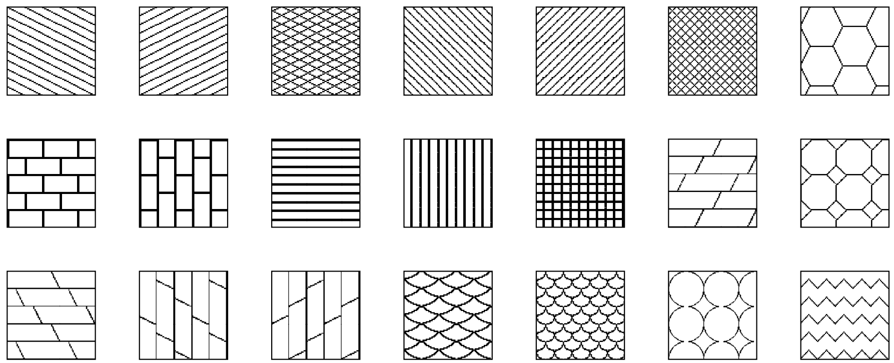
James T. Kuznicki, N. Bruce McCutcheon, 1979 [T 109].

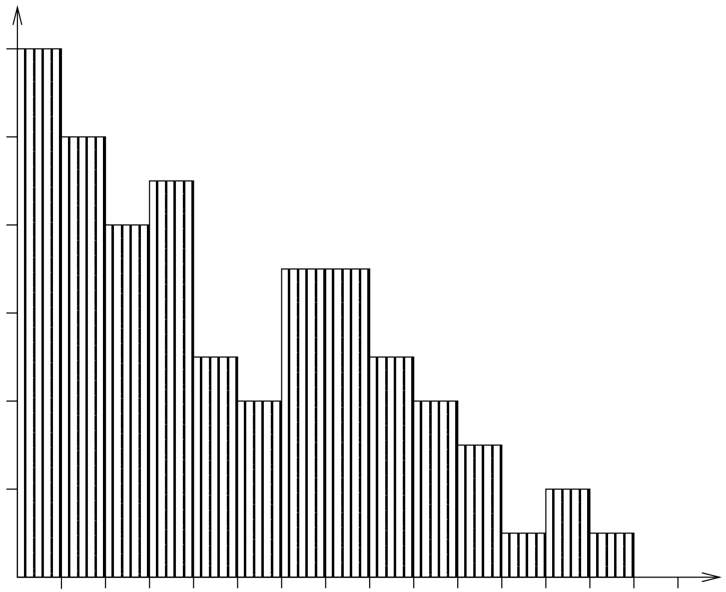
Eain M. Cornford, Marie E. Huot, Science, 1981 [T 109].

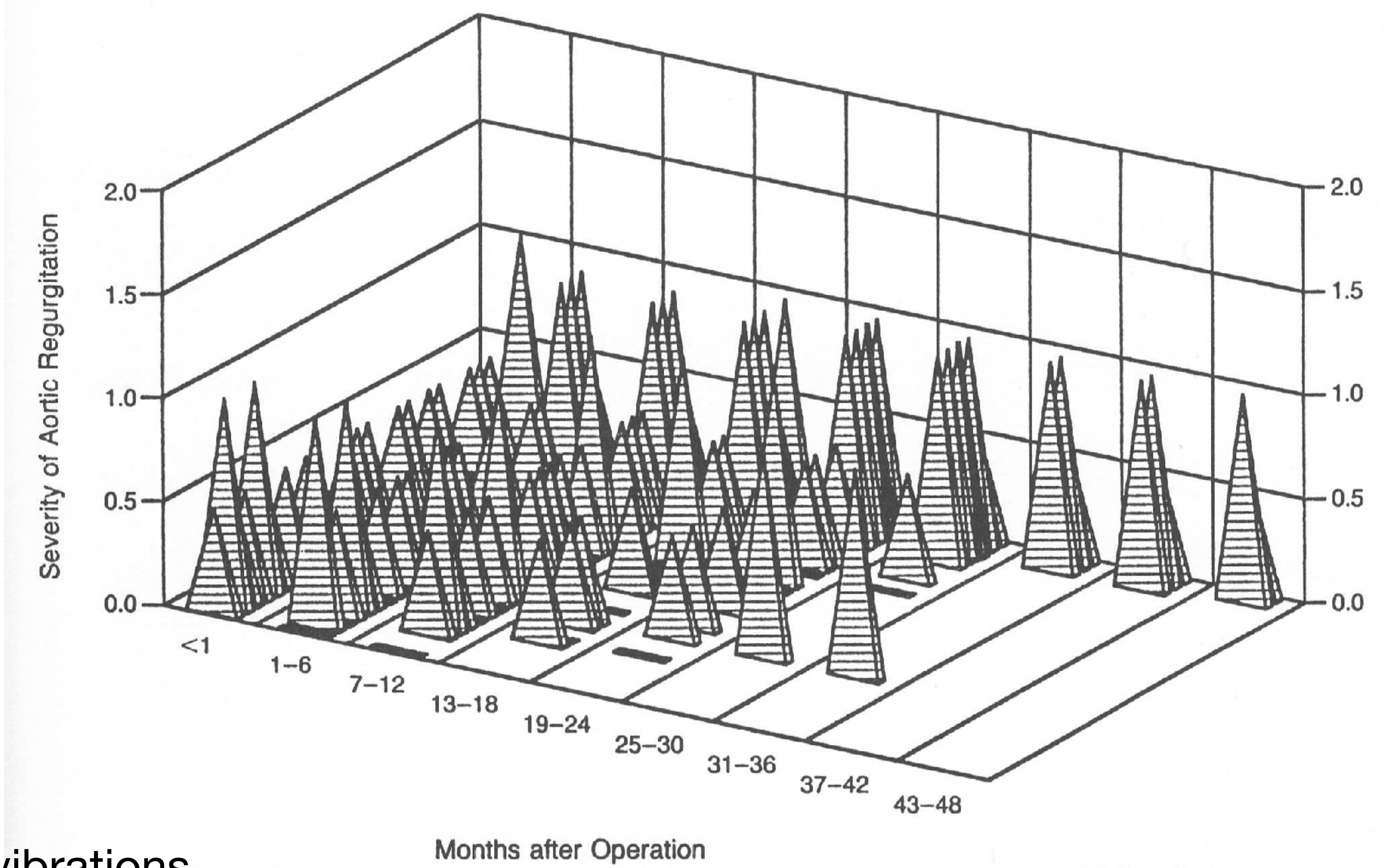
Moirè effects

Moirè vibration is extremely easy to produce with computer graphics tools:

critique partly
historical
(b/w pen plotters)







- vibrations
- necker illusion
- pyramids conceal each other
- also, the stacked depth of the pyramids has no label or scale

Grid lines

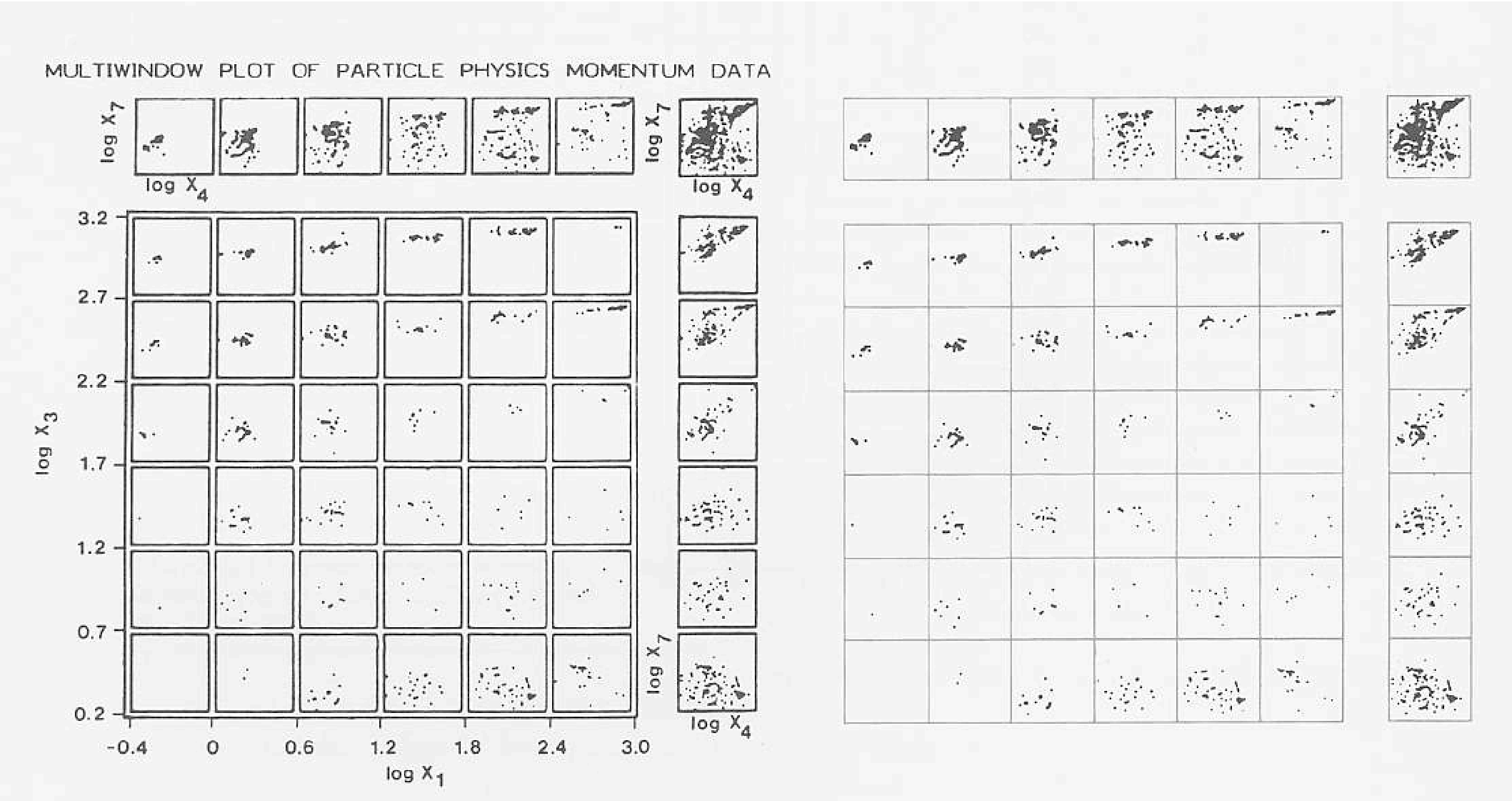
455865876864565749286555584765298742309847249473247
324879427149572389742982479280742938742564875647654
902842968476745464274784674573847648562484789847985

4558658768645657492865555847652987423**3**0984724947**3**247
32487942714957 2**3**89742982479280742 9**3**8742564875647654
9028429684767454642747846745 7**3**847648562484789847985

- As with the numbers above, *the grids* should be muted or suppressed so that the data (3's) can be separated pre-attentively
- Dark grids are *chartjunk*. They carry no information and clutter the graphics.

Grid lines

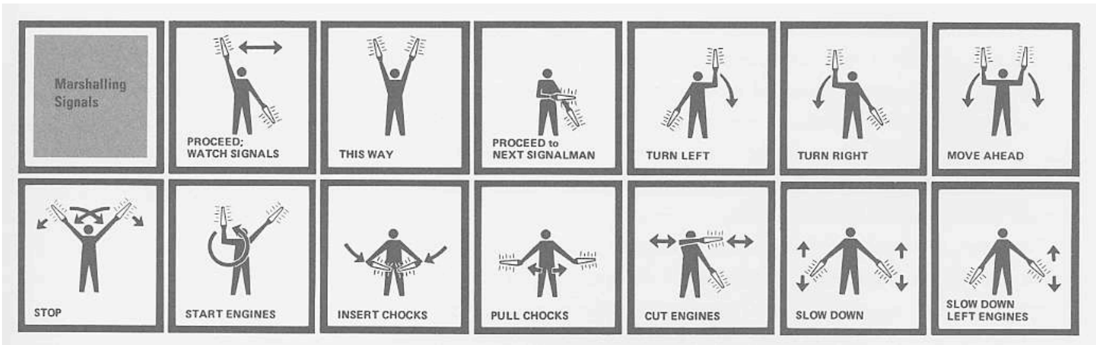
The doubled grid consumes 18 % of the area of this plot. Optical white dots appear at the intersection of the grid lines. Redrawing eliminates the vibration:

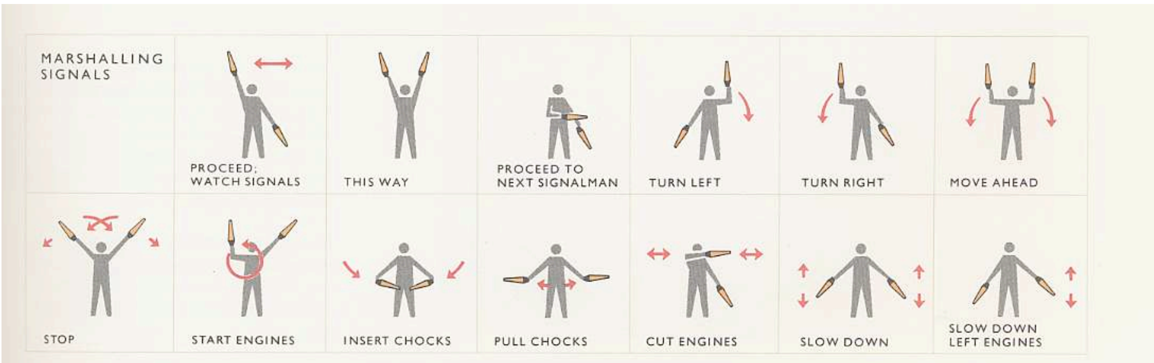


Paul A. Tukey, John W. Tukey, 1981 [T 114].

Grid lines

The grid dominates the graphics. The font is disproportionately weak as compared to the grid. Optical dark spots appear at the intersection of the white grid lines. Redrawing fixes this. The information content is further emphasized by conservative use of color.

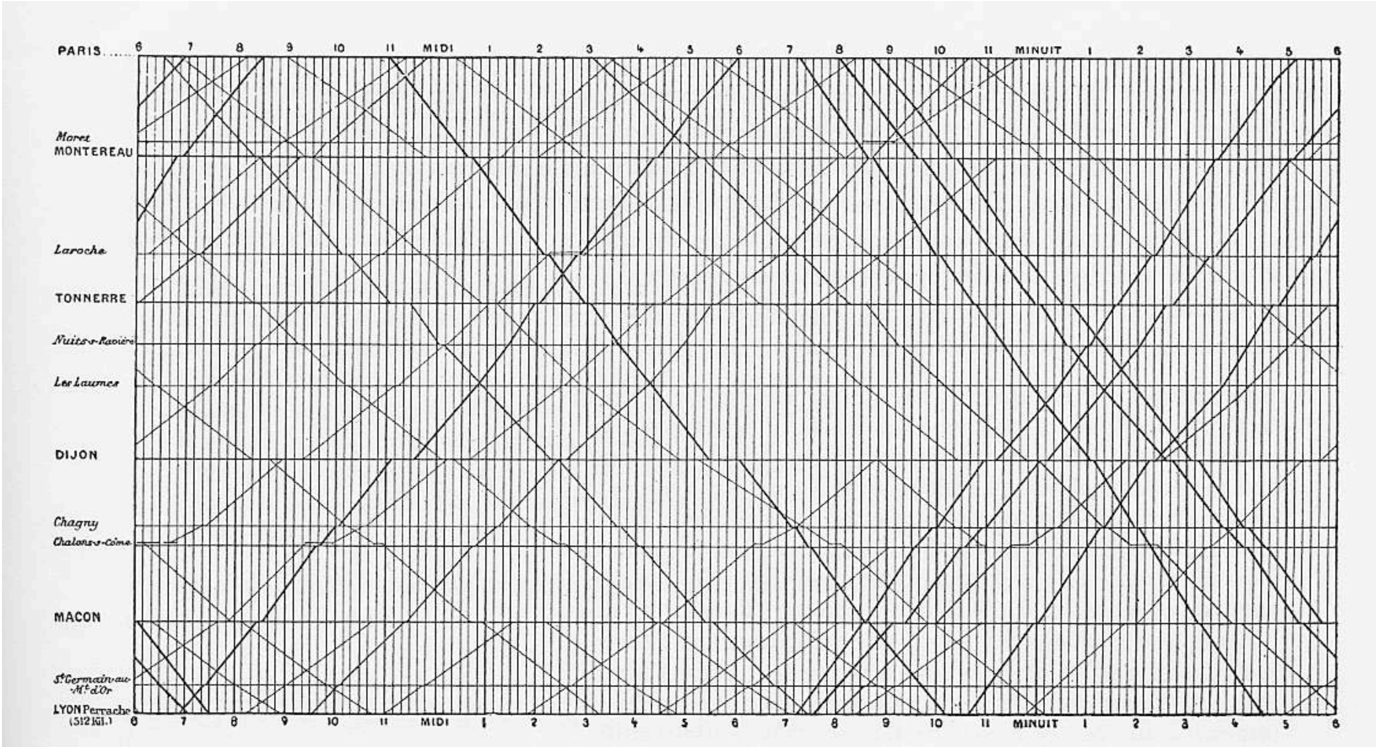




[EI 63].

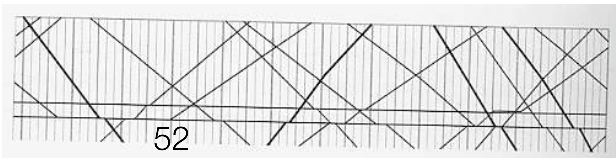
Grid lines

The train schelude by Marey has some Moirè vibration:



E. J. Marey, 1885 [T 31].

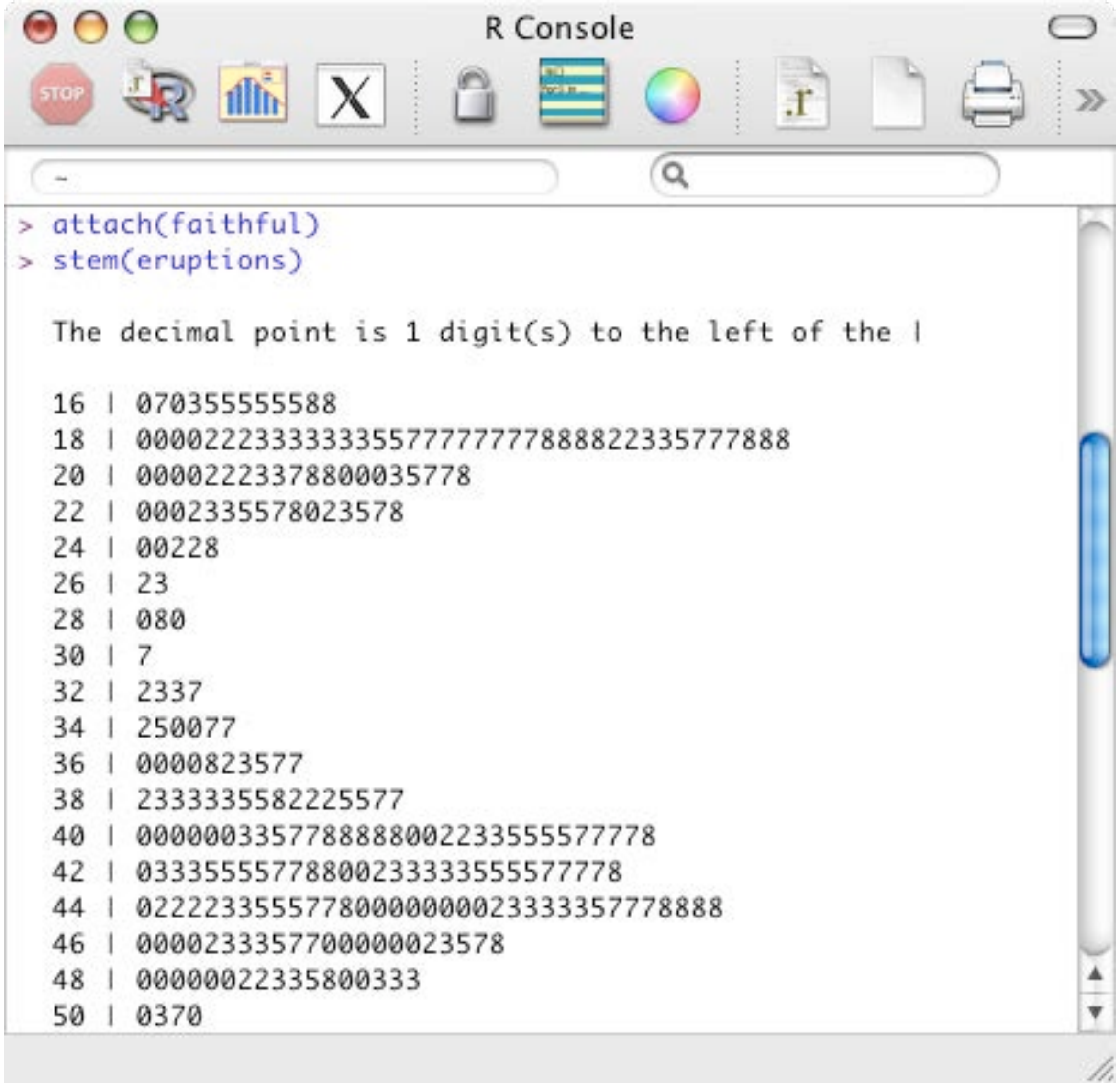
Thinning (or removing!) the grid lines helps:



Multifunctioning graphical elements

- A single *multifunctioning graphical element* can effectively display complex, multivariate data
- Example: a blob on the map specifies not only the geographic coordinates, but also shape of the feature and other properties are specified by color and shading
- Multifunctioning graphical elements will create puzzles, if applied wrongly

Eruption times of Old Faithful geyser



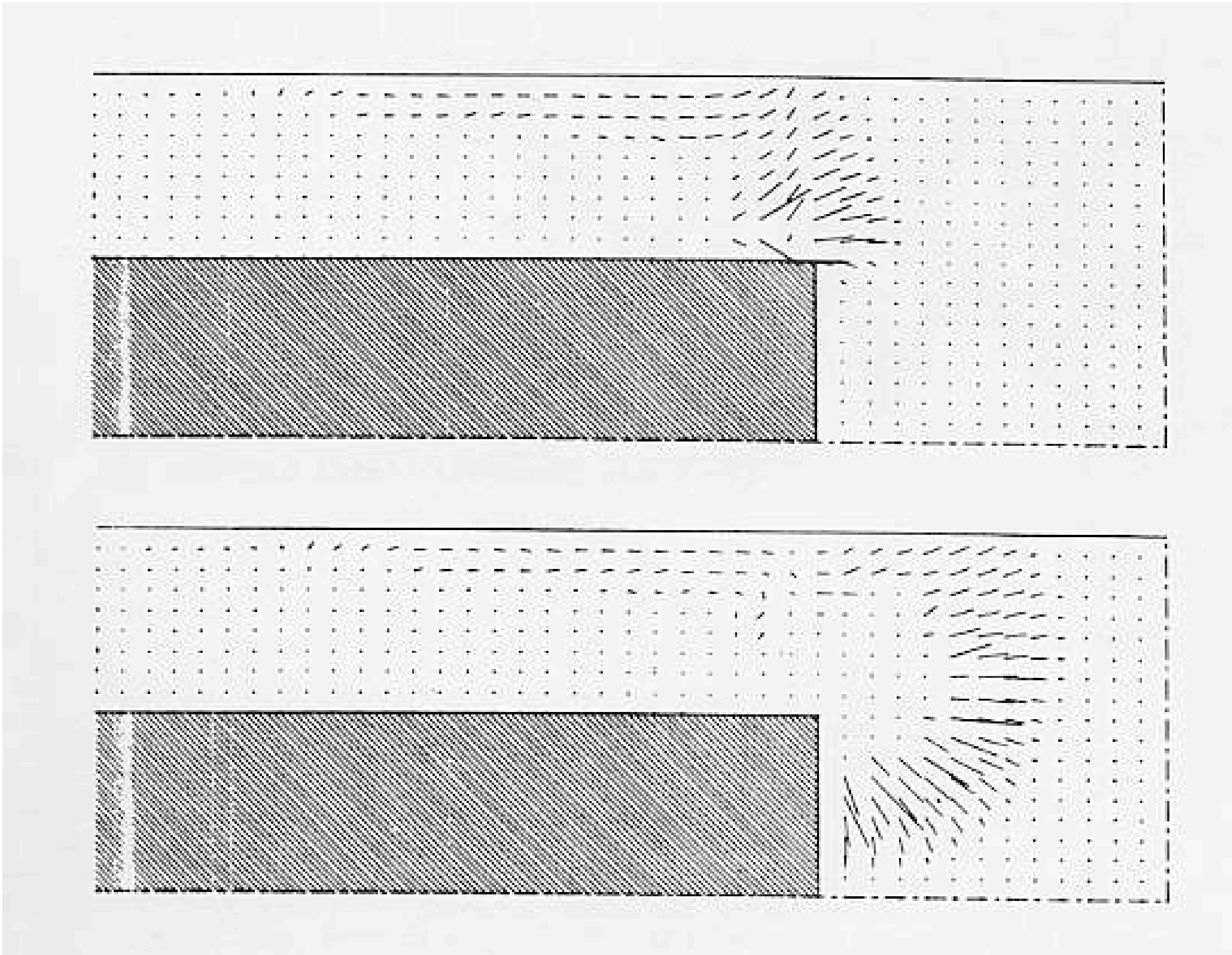
```
R Console

> attach(faithful)
> stem(eruptions)

  The decimal point is 1 digit(s) to the left of the |

  16 | 070355555588
  18 | 000022233333335577777777888822335777888
  20 | 00002223378800035778
  22 | 0002335578023578
  24 | 00228
  26 | 23
  28 | 080
  30 | 7
  32 | 2337
  34 | 250077
  36 | 0000823577
  38 | 2333335582225577
  40 | 0000003357788888002233555577778
  42 | 03335555778800233333555577778
  44 | 02222335557780000000023333357778888
  46 | 0000233357700000023578
  48 | 00000022335800333
  50 | 0370
```

Stem-and-leaf display

The numbers specify *exact* eruption times (minutes up to one decimal) *and* form a bar chart.

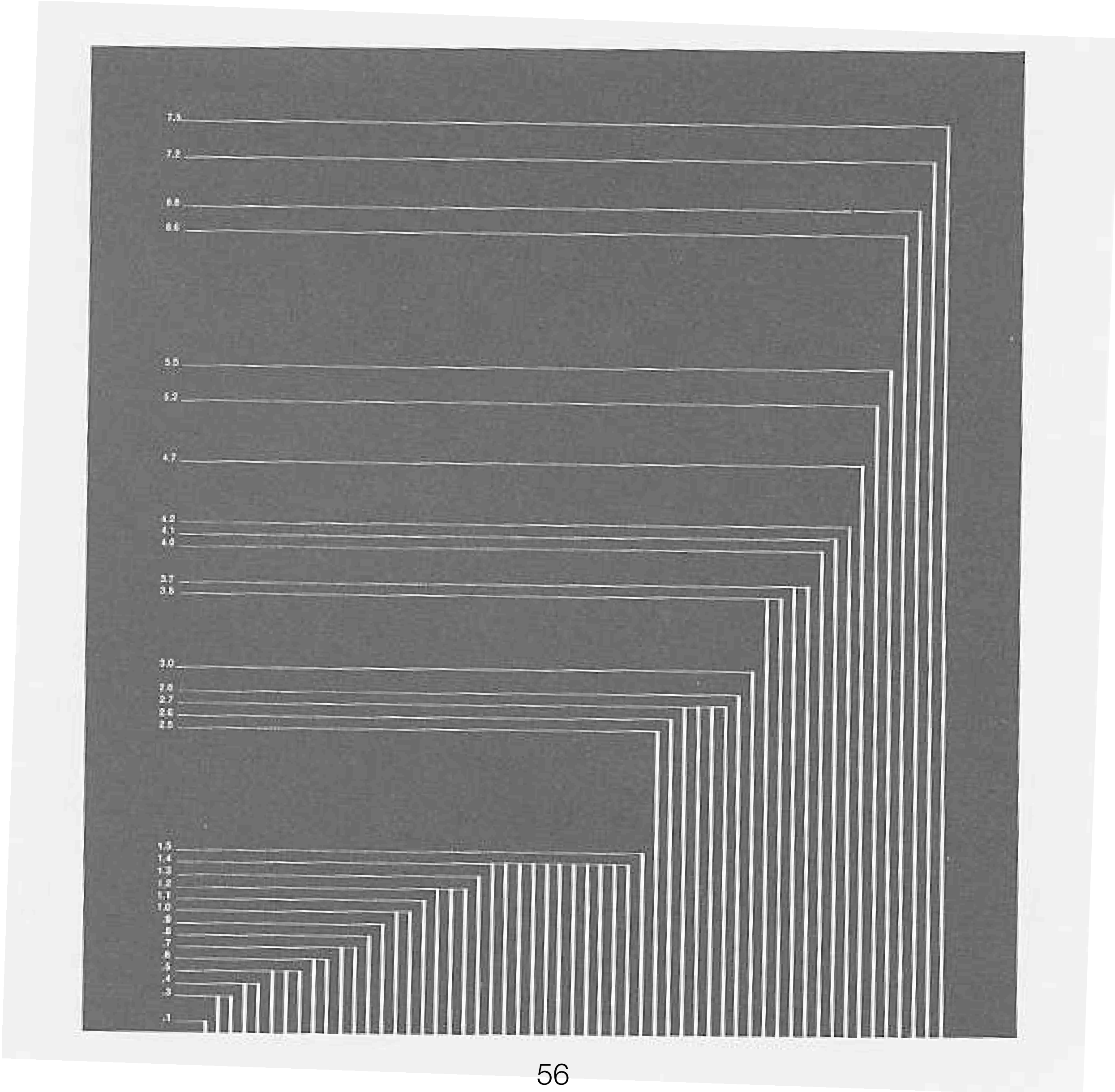
Data-based grid

Sometimes the data grid can be used to report data:

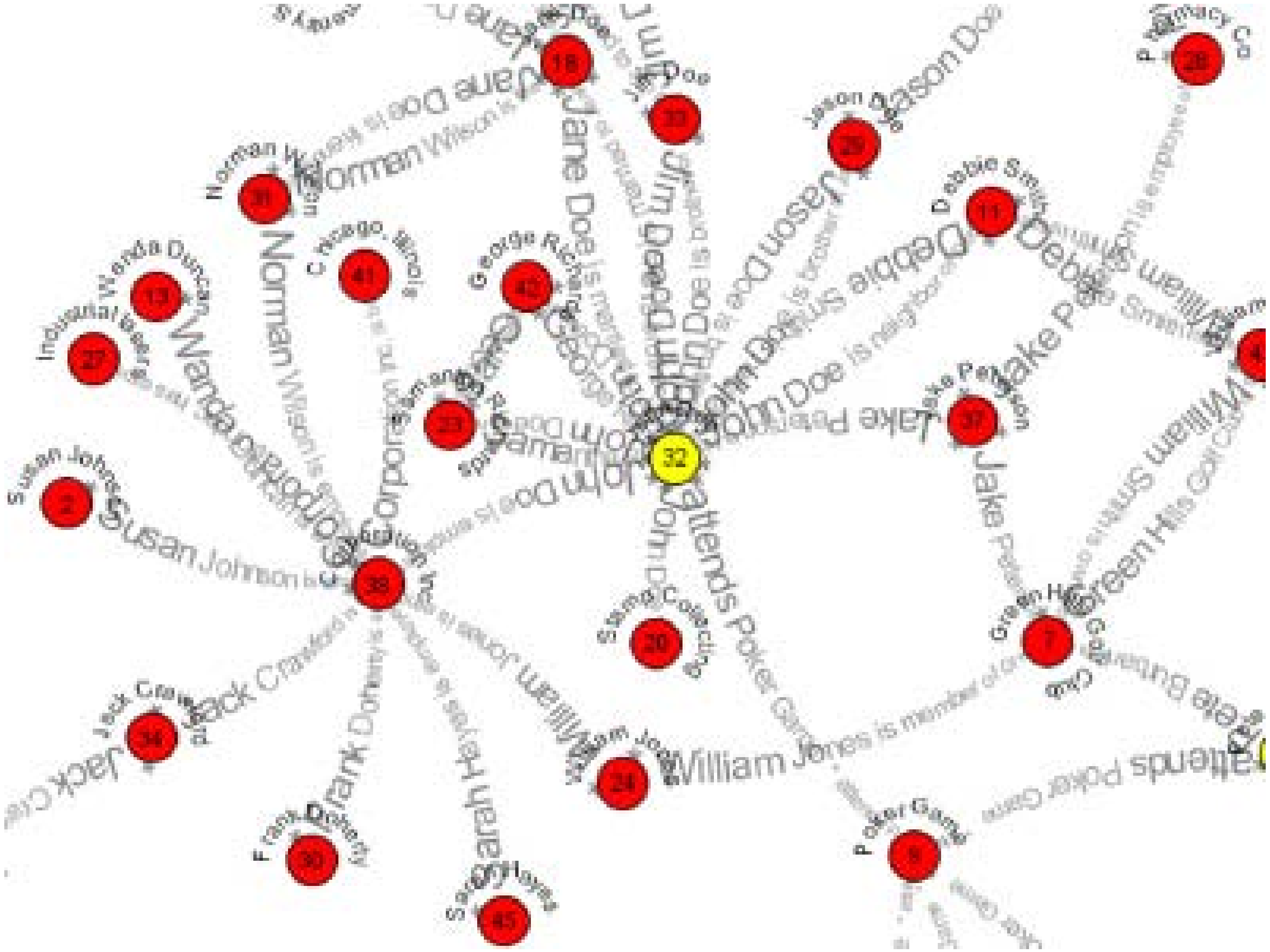


K. V. Roberts, D. E. Potter, 1970 [T 145].

Data-based labels



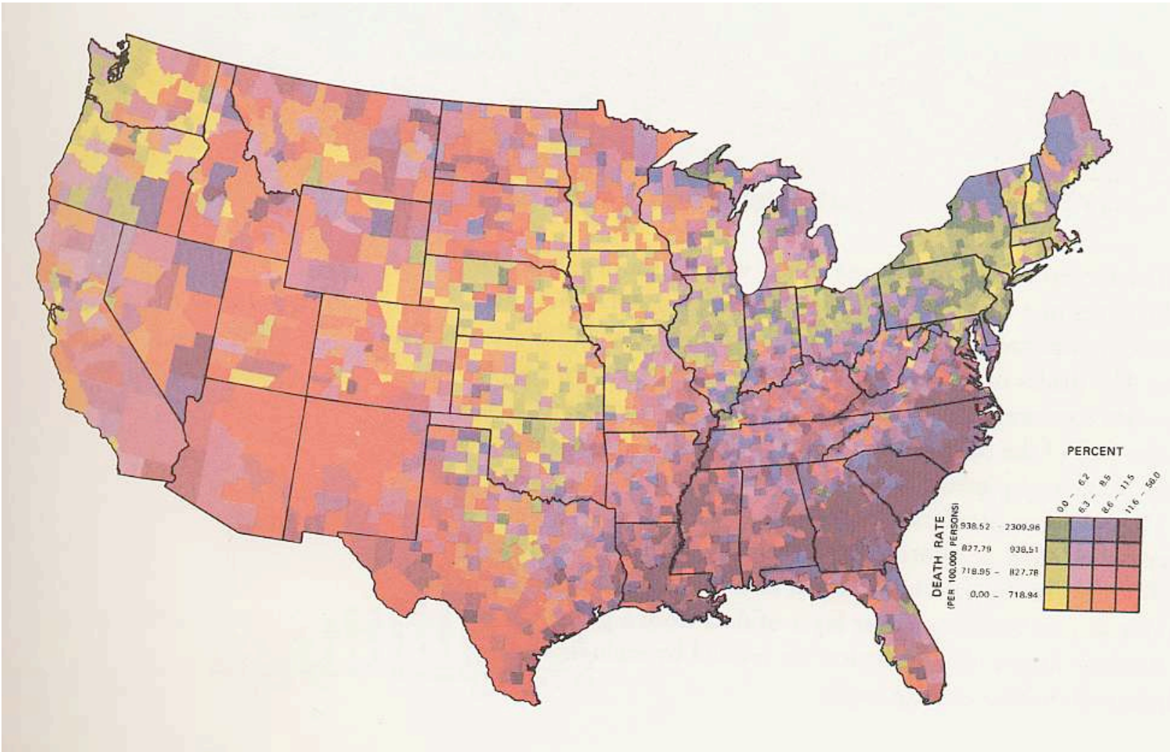
Graphs with extended labels



Wong, Mackey, Perrine, Eagan, Foote, Thomas, *Dynamic Visualization of Graphs with Extended Labels*. InfoVis '05.

Graphical puzzles

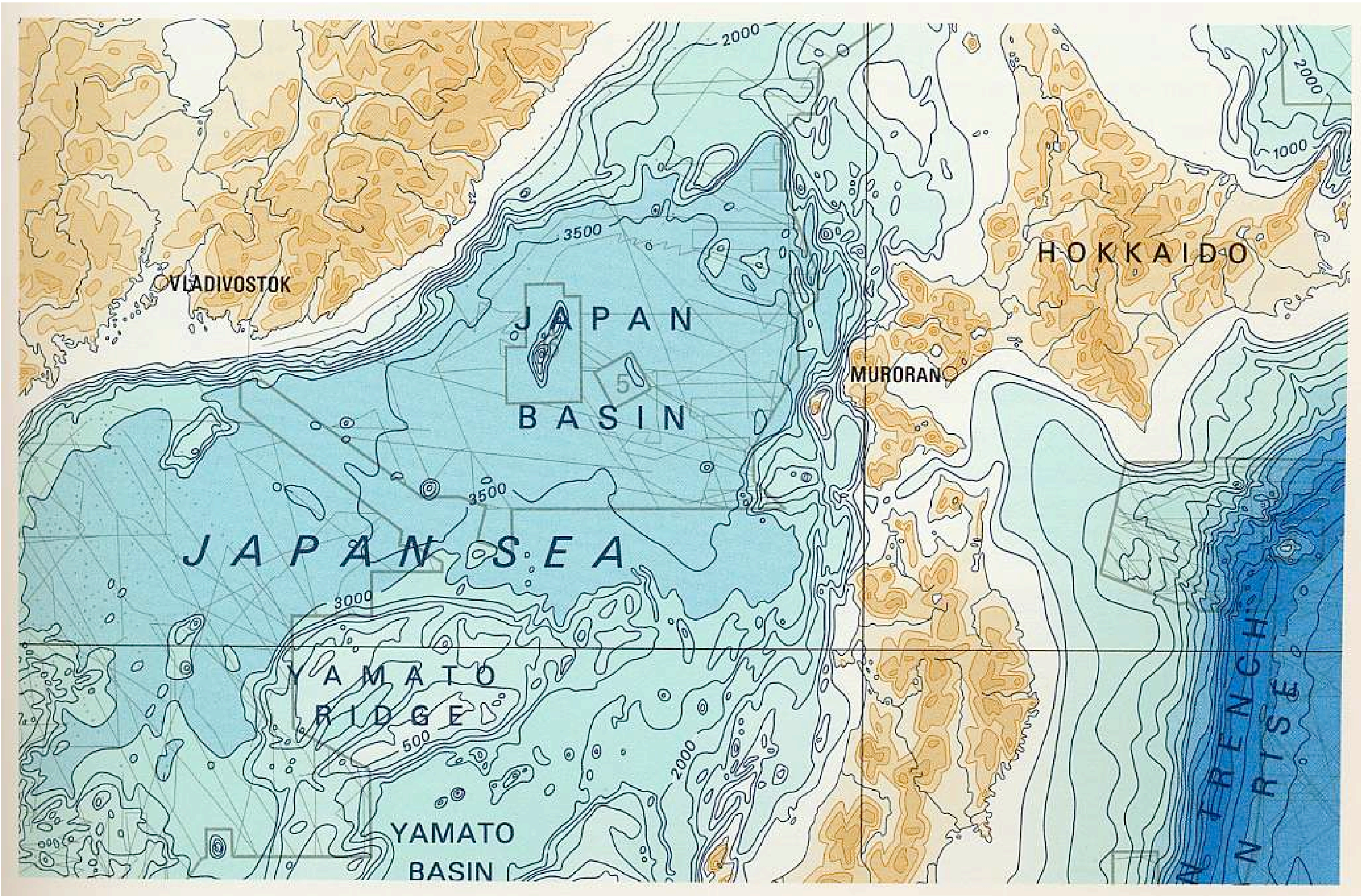
The complexity of multifunctioning elements may turn the data into visual puzzles.



P. Barabba, Alva L. Finker, 1978 [T 153].

This map must be interpreted through verbal rather than visual process.

Visually intuitive use of colors



General Bathymetric Chart of the Oceans, 1984 [EI 91].

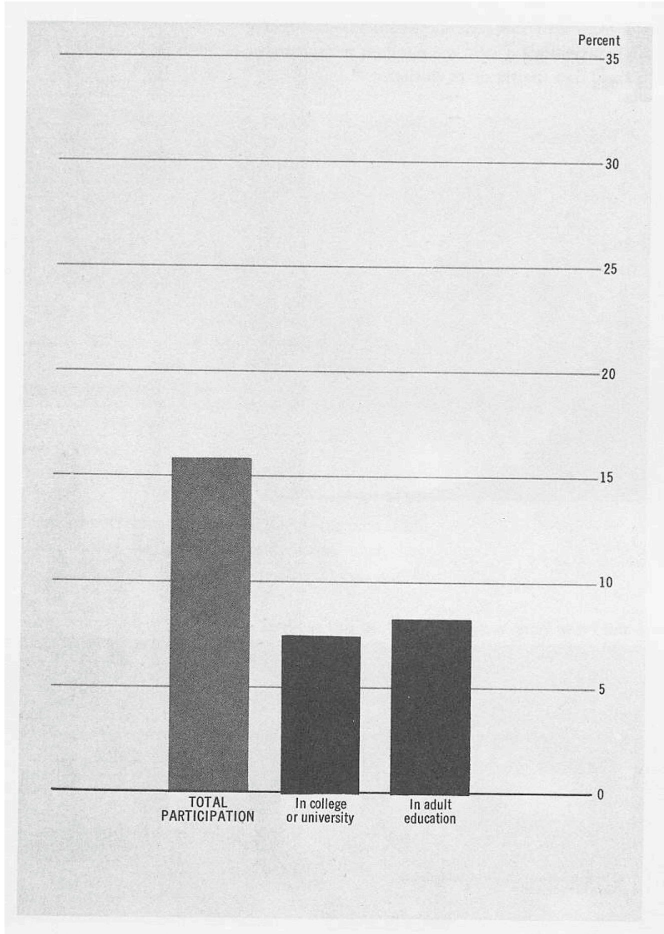
Data density and small multiples

- Eye can distinguish patterns of about 10 (or even 60) cycles per degree

- In computer graphics, the resolution may be lower due to limitations in hardware (typical monitor at typical distances has a resolution of about 40 cycles per degree, 150 cycles per degree would be optimal)

$$\text{Data density} = \frac{\text{Number of entries in data matrix}}{\text{Area of graphics}}$$

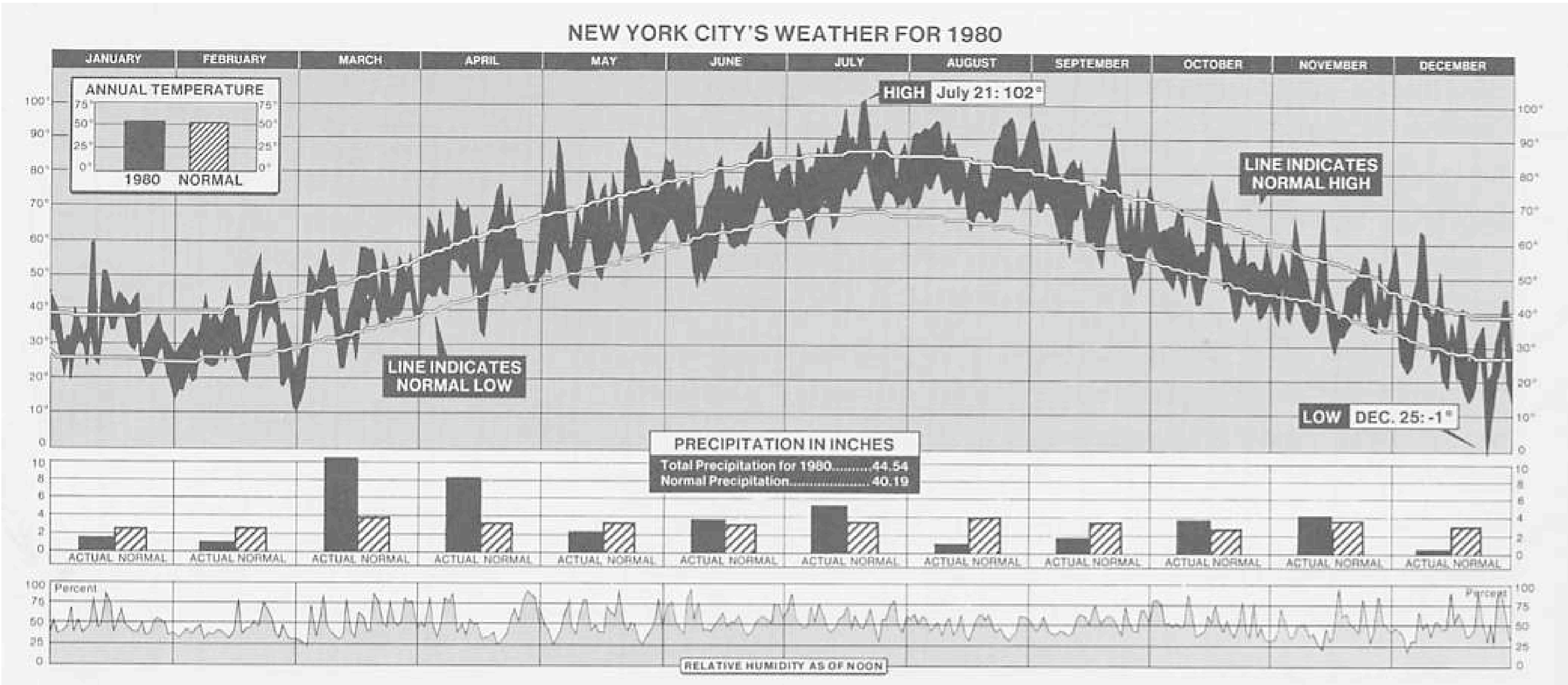
Low data density



Executive Office of the President, 1973 [T 163].

Data density = 0.02 numbers per cm^2.

High data density

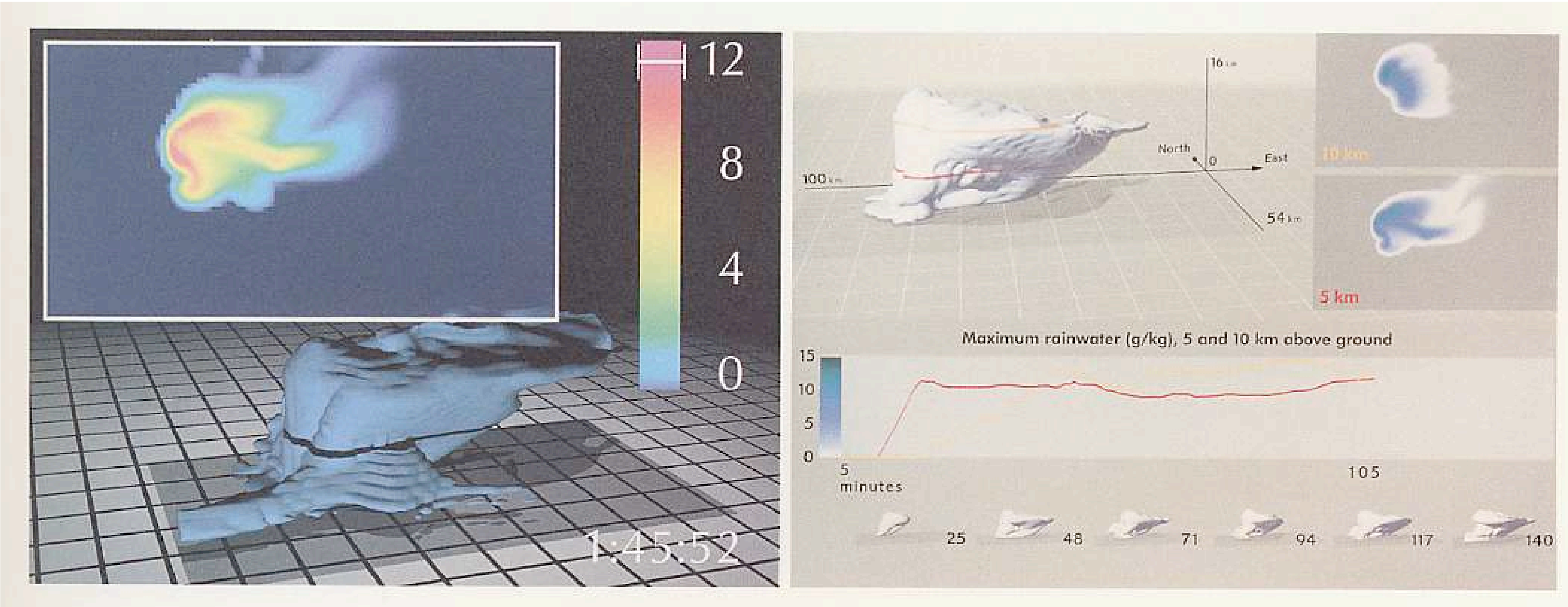


New York Times, 11 January 1981 [T 30].

Data density = 28 numbers per cm^2.

Use small differences

Make all visual distinctions as subtle as possible, but still clear and effective.

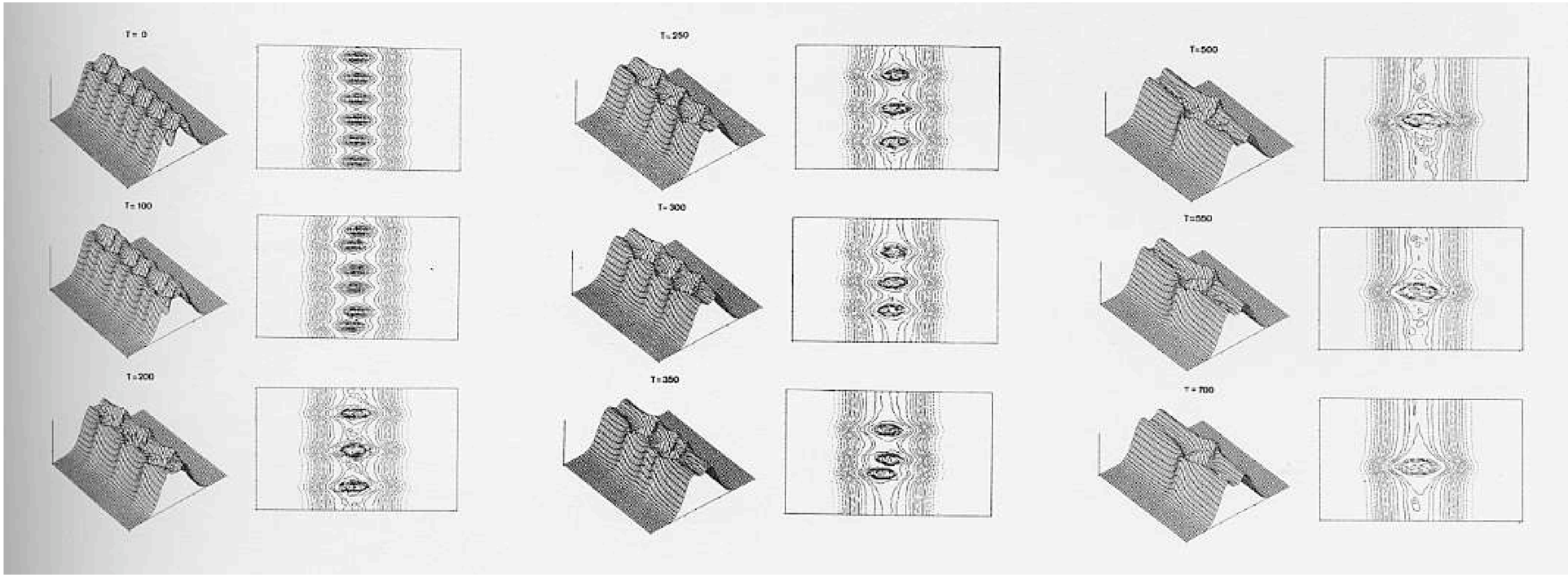


Matthew Arrot et al (original), *E. R. Tufte, Polly Baker et al* (revised) [VE 75].

Large distinctions generate clutter. Smaller distinctions highlight the data.

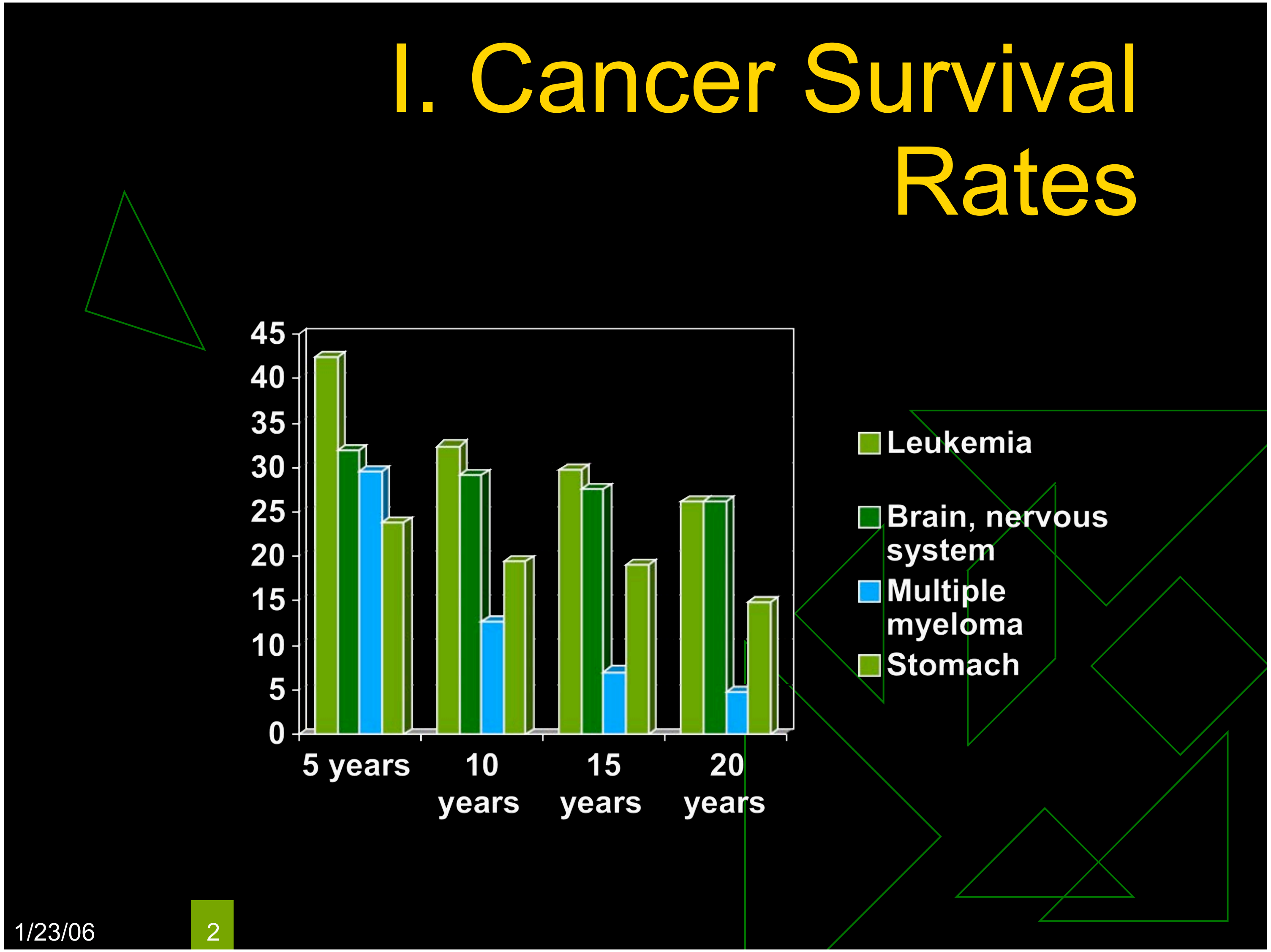
Using small multiples to make comparisons

- Comparisons must be positioned within the eyespan for the viewer to make comparisons at glance
- Show changes in data, not in design.



A. Ghizzo et al, 1988 [EI 67].

Graphs and tables



What is the 15 year survival rate for brain and nervous system cancer?

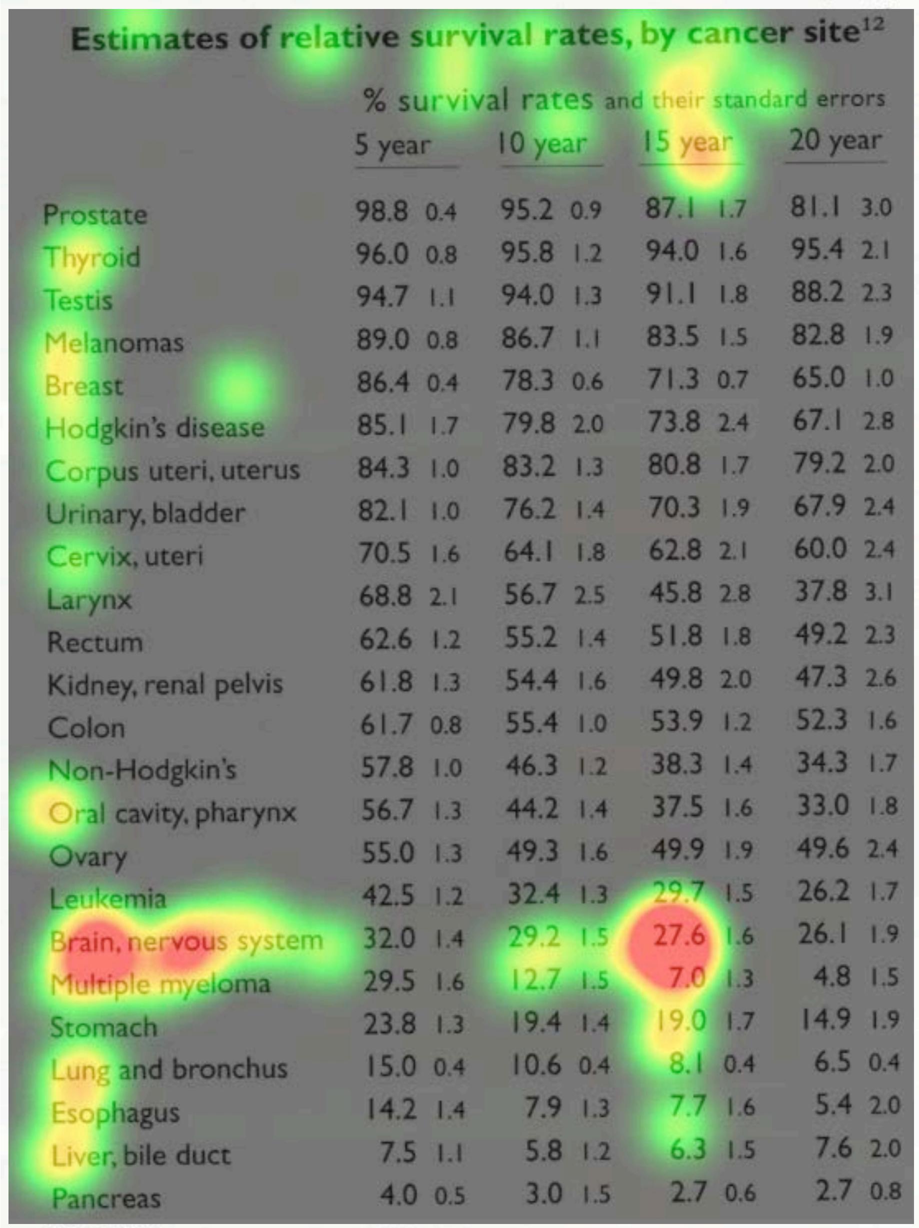
Graphs and tables

	Cancer survival rates (%)			
	5 year	10 year	15 year	20 year
Leukemia	42.5	32.4	29.7	26.2
Brain, nervous	32.0	29.2	27.6	26.1
Multiple myeloma	29.5	12.7	7.0	4.8
Stomach	23.8	19.4	19.0	14.9

What is the 15 year survival rate for brain and nervous system cancer?

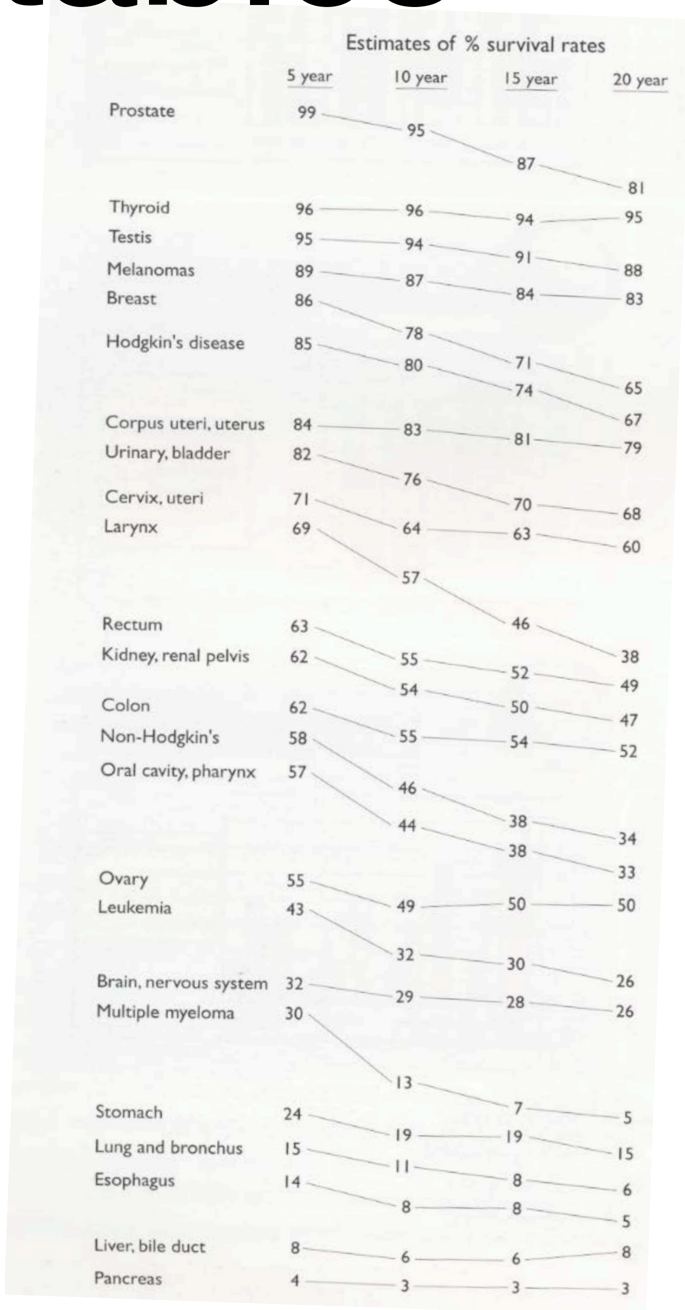
What is the trend over the years?

Graphs and tables



Estimates of relative survival rates, by cancer site[12]

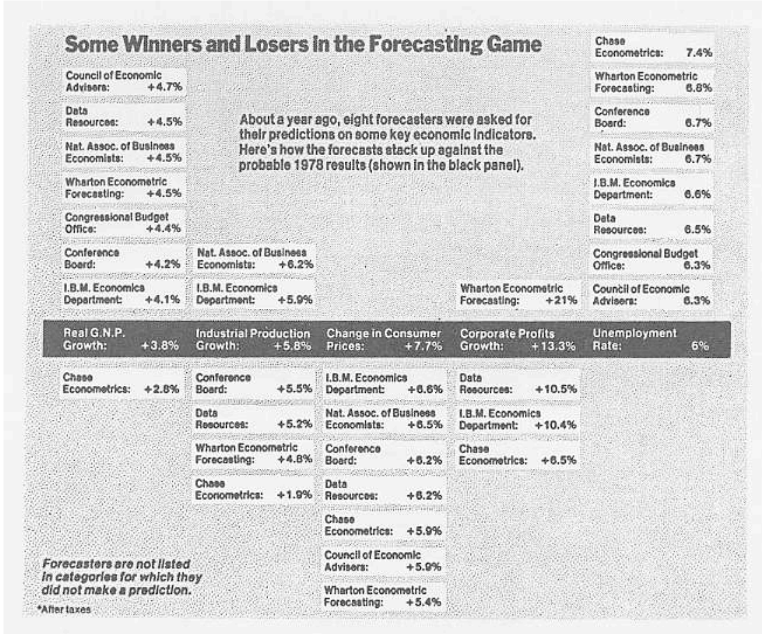
	% survival rates and their standard errors							
	5 year		10 year		15 year		20 year	
Prostate	98.8	0.4	95.2	0.9	87.1	1.7	81.1	3.0
Thyroid	96.0	0.8	95.8	1.2	94.0	1.6	95.4	2.1
Testis	94.7	1.1	94.0	1.3	91.1	1.8	88.2	2.3
Melanomas	89.0	0.8	86.7	1.1	83.5	1.5	82.8	1.9
Breast	86.4	0.4	78.3	0.6	71.3	0.7	65.0	1.0
Hodgkin's disease	85.1	1.7	79.8	2.0	73.8	2.4	67.1	2.8
Corpus uteri, uterus	84.3	1.0	83.2	1.3	80.8	1.7	79.2	2.0
Urinary, bladder	82.1	1.0	76.2	1.4	70.3	1.9	67.9	2.4
Cervix, uteri	70.5	1.6	64.1	1.8	62.8	2.1	60.0	2.4
Larynx	68.8	2.1	56.7	2.5	45.8	2.8	37.8	3.1
Rectum	62.6	1.2	55.2	1.4	51.8	1.8	49.2	2.3
Kidney, renal pelvis	61.8	1.3	54.4	1.6	49.8	2.0	47.3	2.6
Colon	61.7	0.8	55.4	1.0	53.9	1.2	52.3	1.6
Non-Hodgkin's	57.8	1.0	46.3	1.2	38.3	1.4	34.3	1.7
Oral cavity, pharynx	56.7	1.3	44.2	1.4	37.5	1.6	33.0	1.8
Ovary	55.0	1.3	49.3	1.6	49.9	1.9	49.6	2.4
Leukemia	42.5	1.2	32.4	1.3	29.7	1.5	26.2	1.7
Brain, nervous system	32.0	1.4	29.2	1.5	27.6	1.6	26.1	1.9
Multiple myeloma	29.5	1.6	12.7	1.5	7.0	1.3	4.8	1.5
Stomach	23.8	1.3	19.4	1.4	19.0	1.7	14.9	1.9
Lung and bronchus	15.0	0.4	10.6	0.4	8.1	0.4	6.5	0.4
Esophagus	14.2	1.4	7.9	1.3	7.7	1.6	5.4	2.0
Liver, bile duct	7.5	1.1	5.8	1.2	6.3	1.5	7.6	2.0
Pancreas	4.0	0.5	3.0	1.5	2.7	0.6	2.7	0.8



What is the 15 year survival rate for brain and nervous system cancer?

Graphs and tables

- The data can be shown in
  - sentences,
  - tables or
  - graphics.
- Table is usually the best choice for (small) collection of numbers



New York Times, 2 January 1979 [T 180].

Nearly 53 % of group A did something compared to 46 % of group B and 57 % of C.

Same using a table:

Group A	53 %
Group B	46 %
Group C	57 %

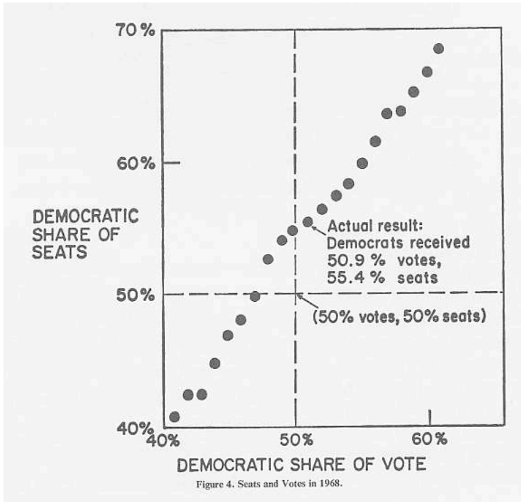
Better(?) order:

Group B	46 %
Group A	53 %
Group C	57 %

Aesthetics:

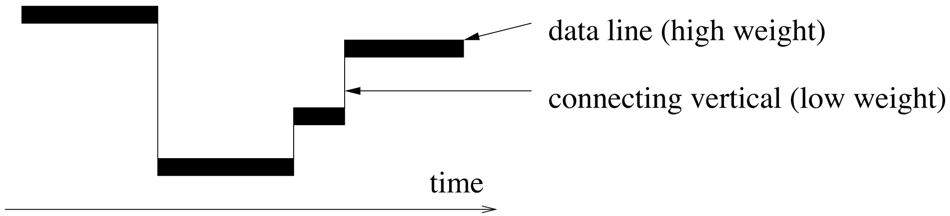
Line weight and lettering

The weights of the letters should be in proportion to the other visual elements:



E. R. Tufte, 1973 [T 184].

The heavier weight should be given to data measures:



Line weight and lettering

An excellent summary of crimes committed by state's witnesses in a Mafia trial. Notice the thick glyphs and how the most horrid crimes are listed first and last.

CRIMINAL ACTIVITY OF GOVERNMENT INFORMANTS

CRIME	CARDINALE	LOFARO	MALONEY	POLISI	SENATORE	FORONJY	CURRO
MURDER	X	X					
ATTEMPTED MURDER		X	X				
HEROIN POSSESSION AND SALE	X	X		X			X
COCAINE POSSESSION AND SALE	X		X	X			
MARIJUANA POSSESSION AND SALE							X
GAMBLING BUSINESS		X		X		X	
ARMED ROBBERIES	X		X	X	X		X
LOANSHARKING		X		X			
KIDNAPPING			X	X			
EXTORTION			X	X			
ASSAULT	X		X	X			X
POSSESSION OF DANGEROUS WEAPONS	X	X	X	X	X		X
PERJURY		X				X	
COUNTERFEITING					X	X	
BANK ROBBERY			X	X			
ARMED HIJACKING				X	X		
STOLEN FINANCIAL DOCUMENTS			X	X	X		
TAX EVASION				X		X	
BURGLARIES	X	X		X	X		
BRIBERY		X		X			
THEFT: AUTO, MONEY, OTHER			X	X	X	X	X
BAIL JUMPING AND ESCAPE			X	X			
INSURANCE FRAUDS					X	X	
FORGERIES				X	X		
PISTOL WHIPPING A PRIEST	X						
SEXUAL ASSAULT ON MINOR							X
RECKLESS ENDANGERMENT							X

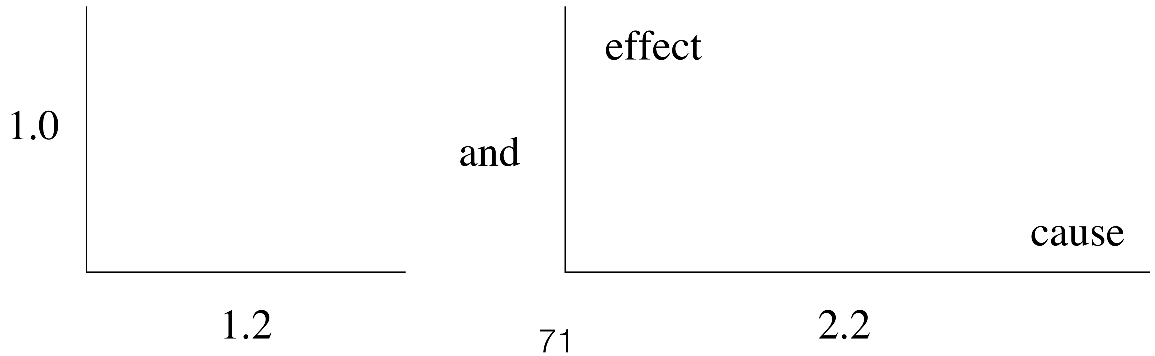
United States vs. Gotti et al, 1987 [EI 31].

Aesthetics:
Proportion of graphics

Graphics should usually have greater length than height:

- Our eye is practiced in detecting deviations from the horizon. Thus e.g. horizontal time-series are easier to read.
- It is easier to write words and labels horizontally.
- Longer horizontal helps to emphasize the causal variable

Preferred height/length ratios vary depending on the circumstances; the golden ratio 1:1.618 is a good rule of thumb.

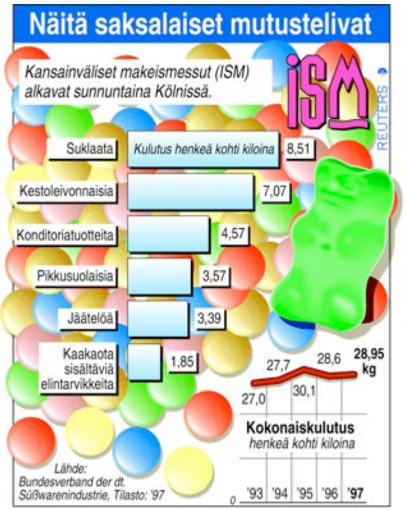


Summary

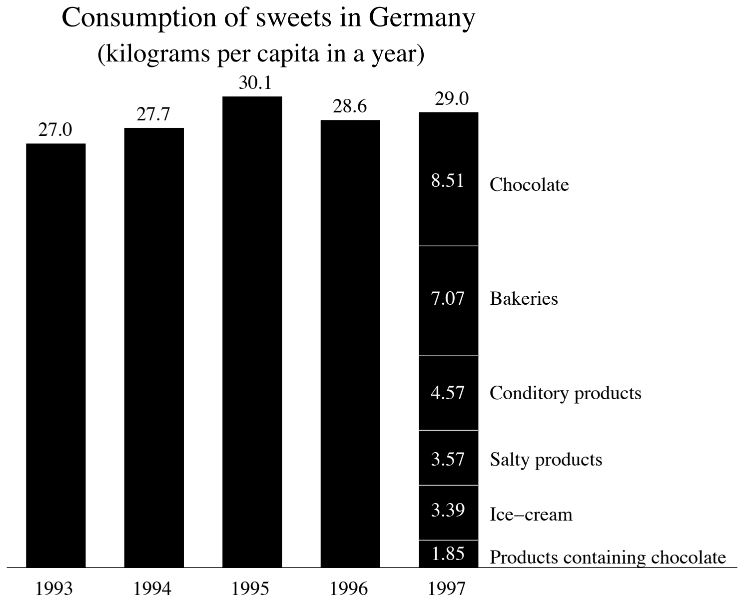
- Data-ink
- Chartjunk
- Multifunctioning graphical elements
- Data density and small multiples
- Aesthetics and techniques

Conclusion

- Communicating ideas and information is difficult. Wiio's laws:
  - Communication usually fails, except by accident.
  - If a message can be interpreted in several ways, it will be interpreted in a manner that maximizes damages.
- Theory of data graphics: show only the essential in a way that makes the facts obvious. Don't waste space and eliminate all non-essentials and redundancies



H. Spissler, Reuters, 1999.



Next lecture

- Visualization techniques
- The assignments