

CS-E4840

Information visualization D

Lecture 3: Theory of data graphics
Mar 6, 2023

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

Basic principle:

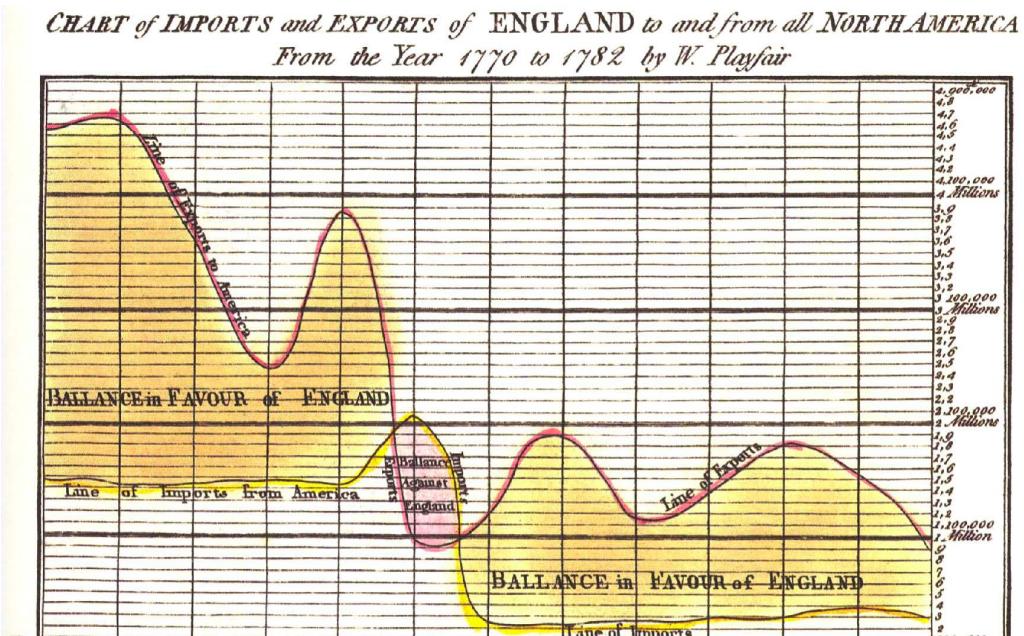
- Give the viewer the greatest number of ideas (or, most complete understanding)
- ... in the shortest time
- ... using the least amount of ink

By:

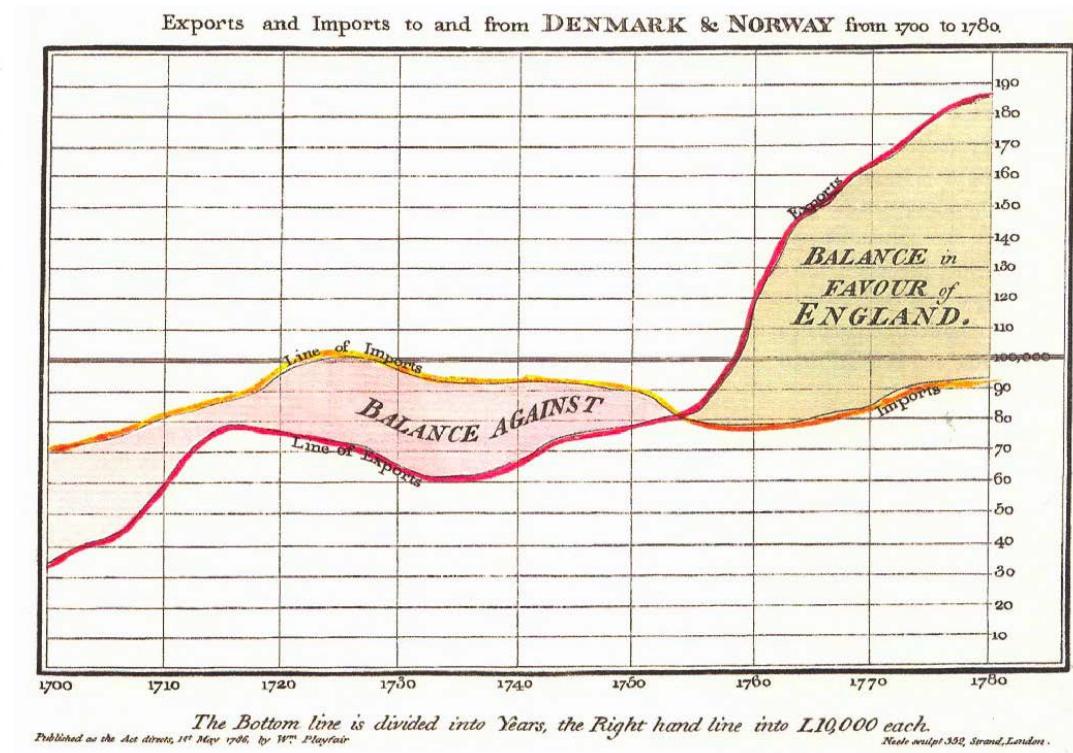
- Not wasting space
- Eliminating non-essentials and redundancies
- Making the graphics as easy to read and as simple as possible, while displaying the data fully.

Which is better?

(a)



(b)



Theory of data graphics

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

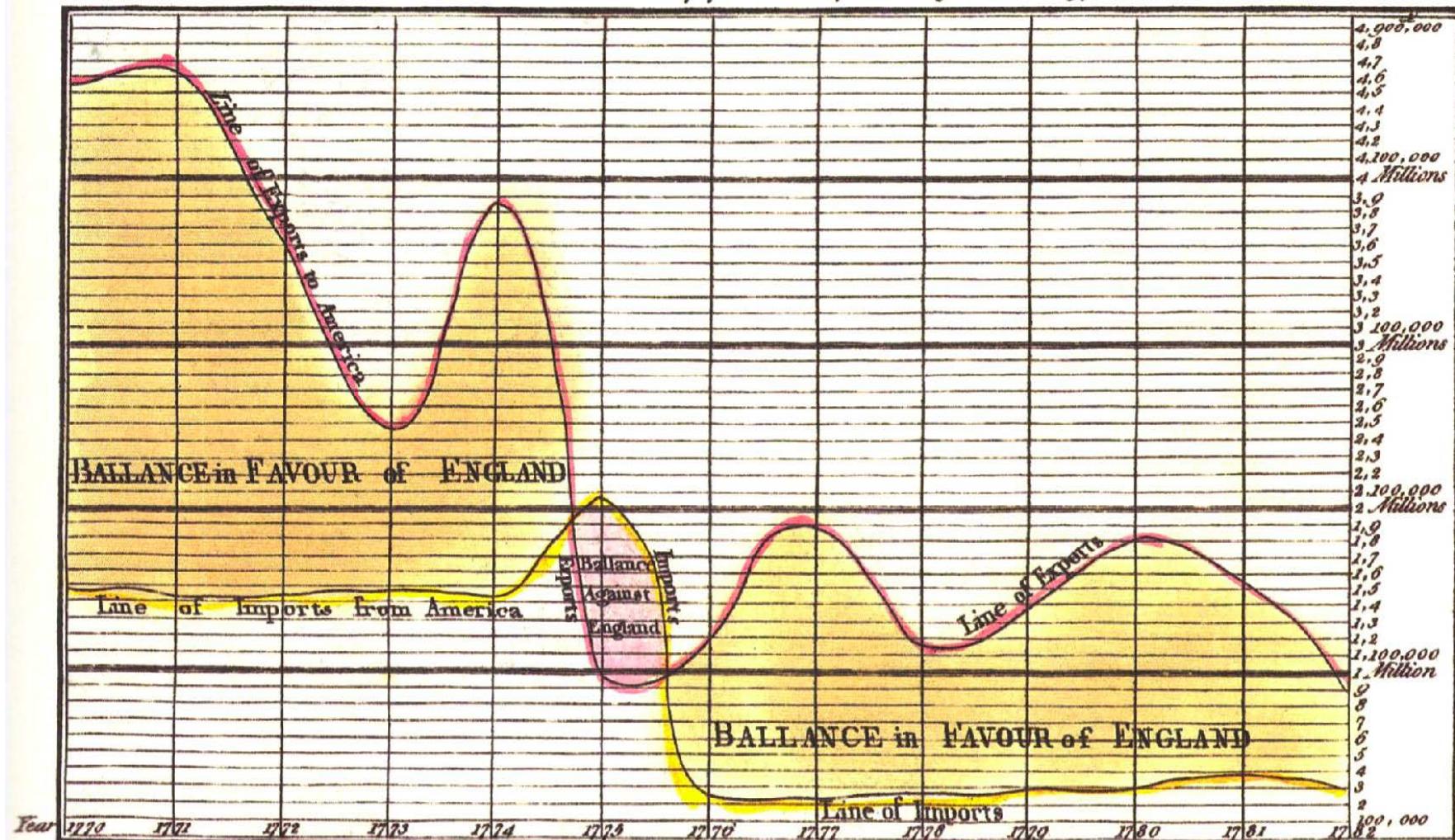
Data-ink

- 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

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

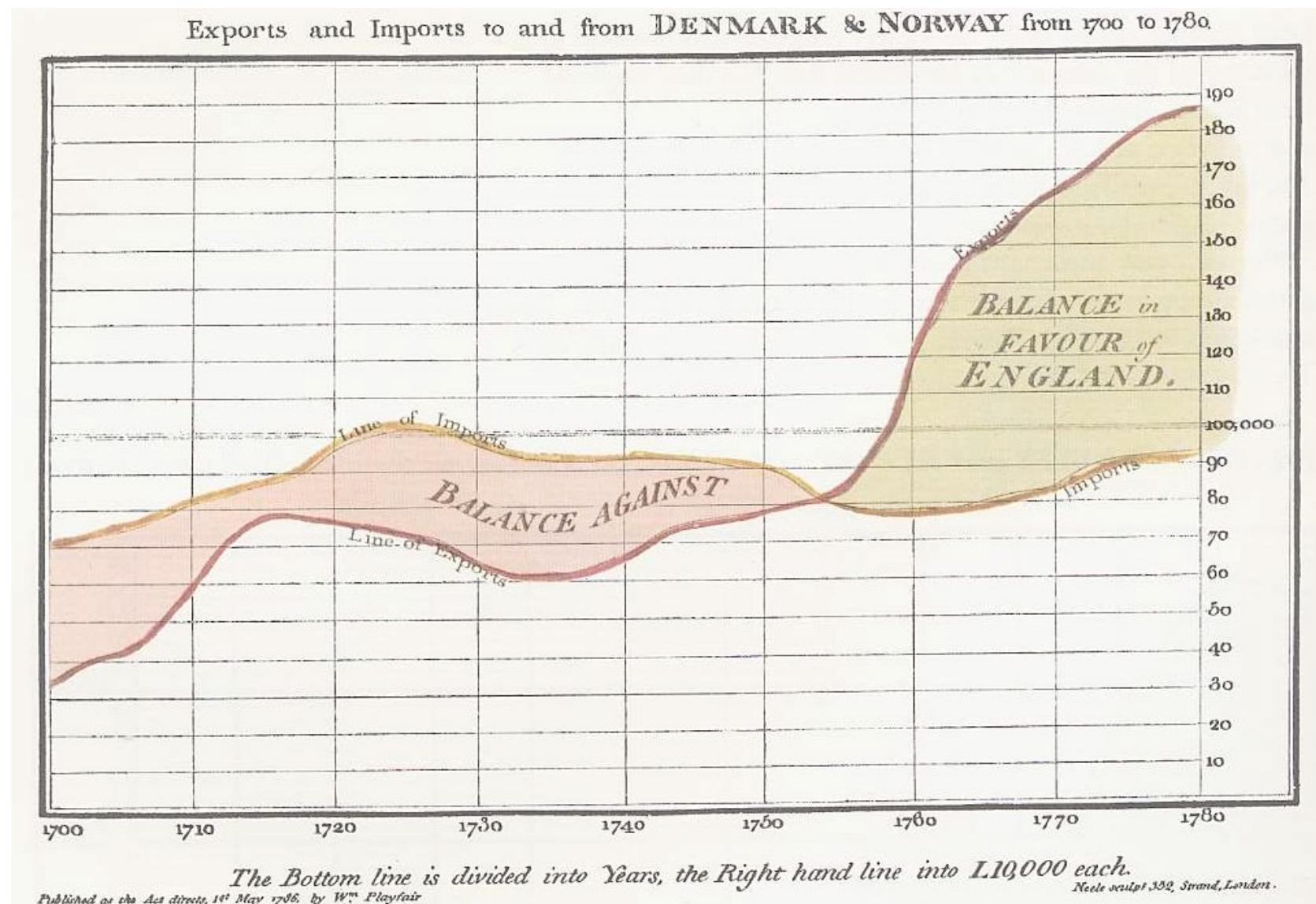
Example: low data-ink ratio

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

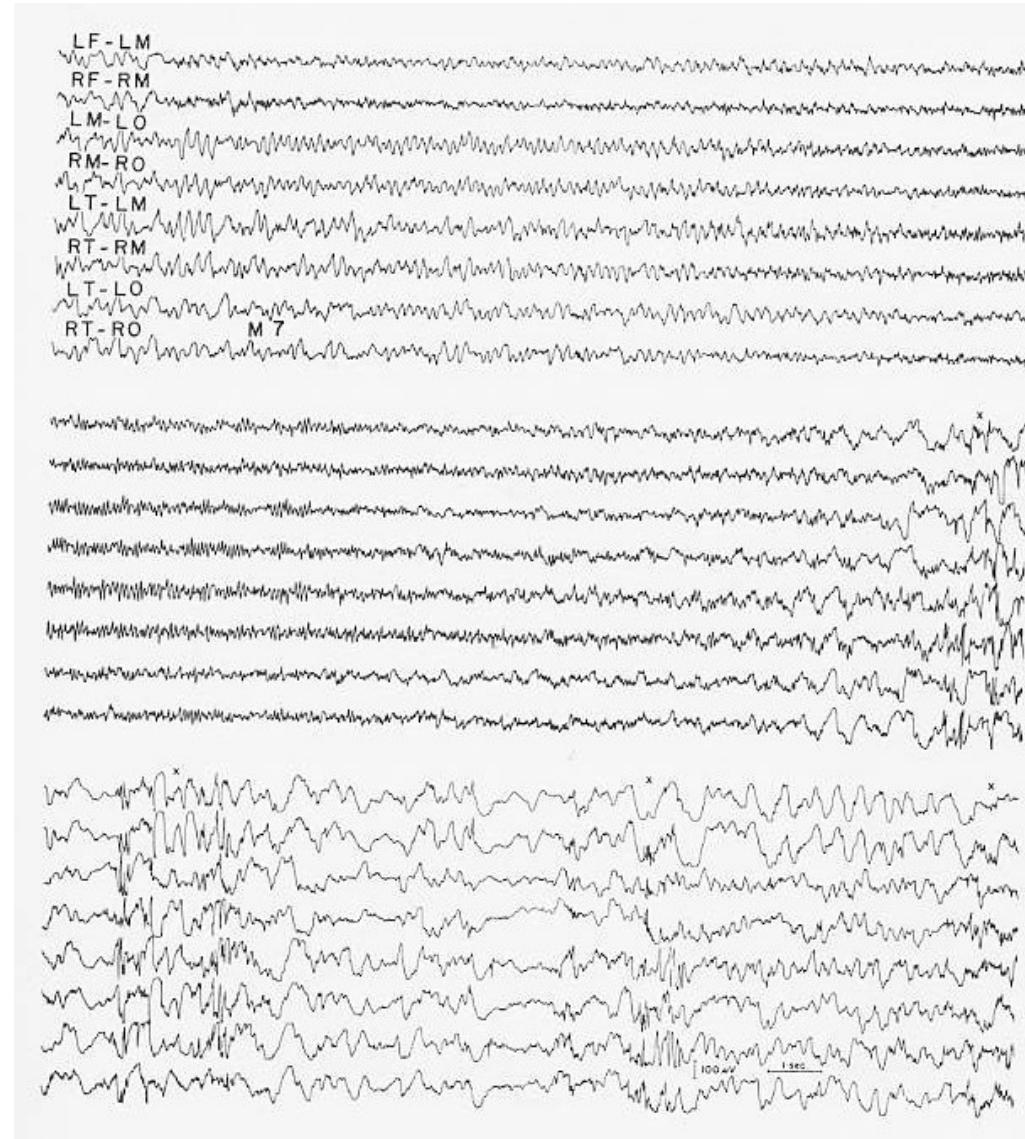


The Bottom Line is divided into Years the right-hand Line into HUNDRED THOUSAND POUNDS

Example: high-data ink ratio



100% data-ink ratio



Kenneth A. Ko¹⁰, 1971 [T 93].

Sparklines

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



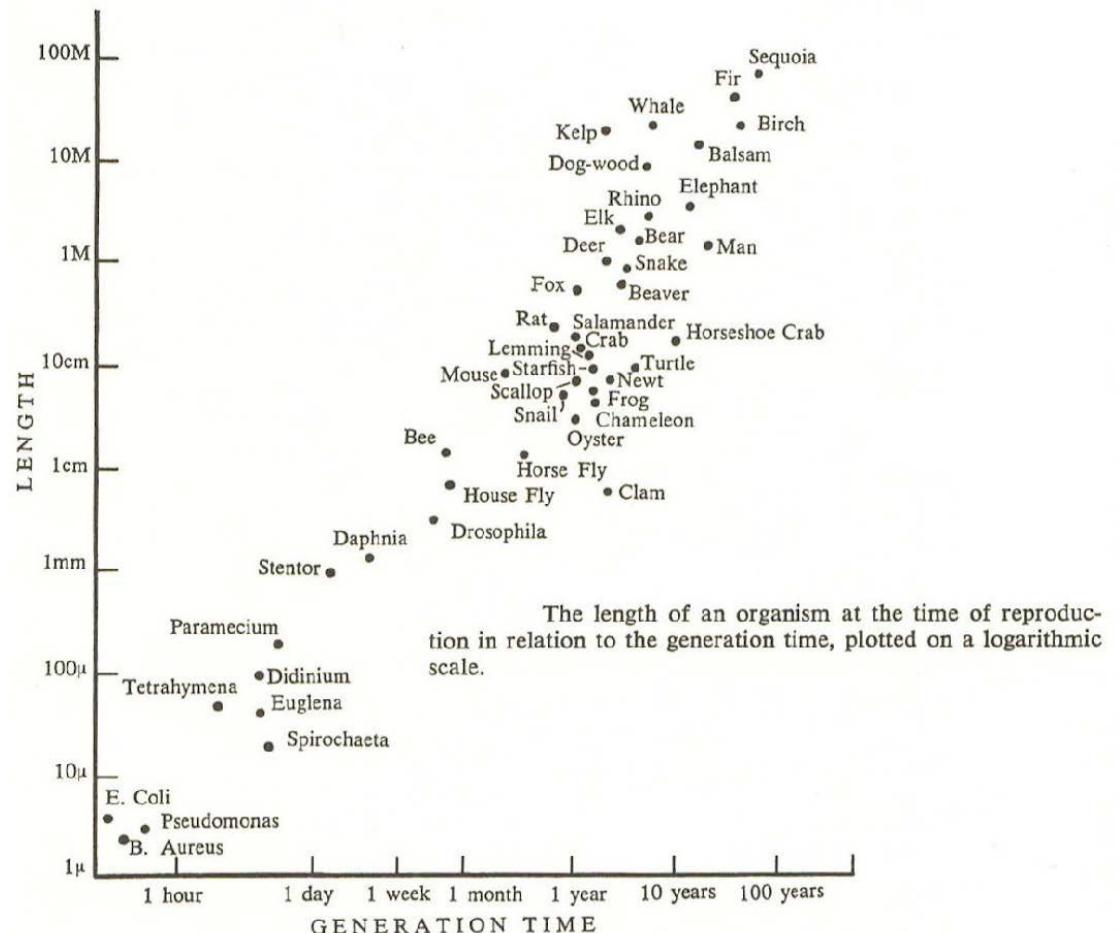
Using d3.js, we can fairly easily draw SVG-based sparklines. This is 2013 historical stock prices for

Google **\$1084.75**. And this is for **Facebook** **\$55.57**. And this is for

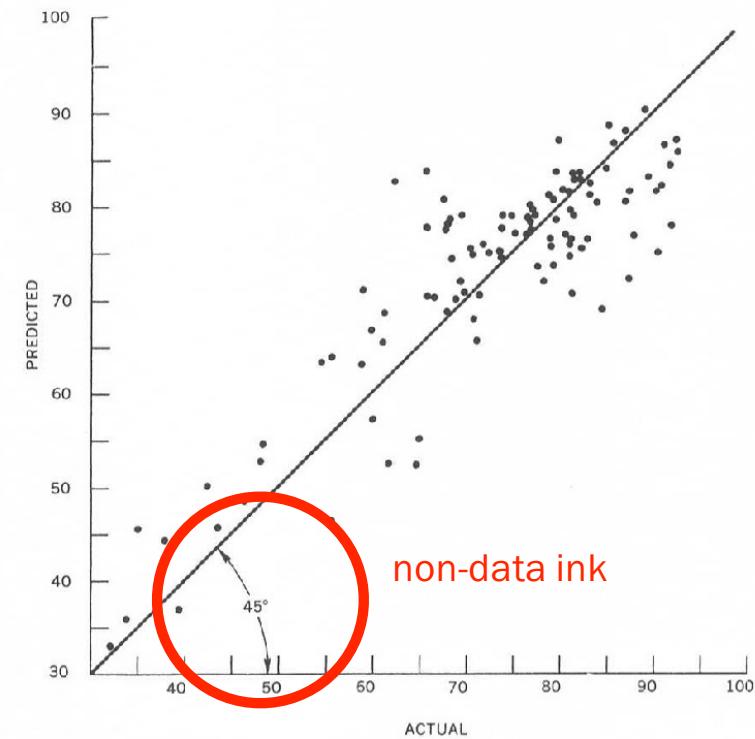
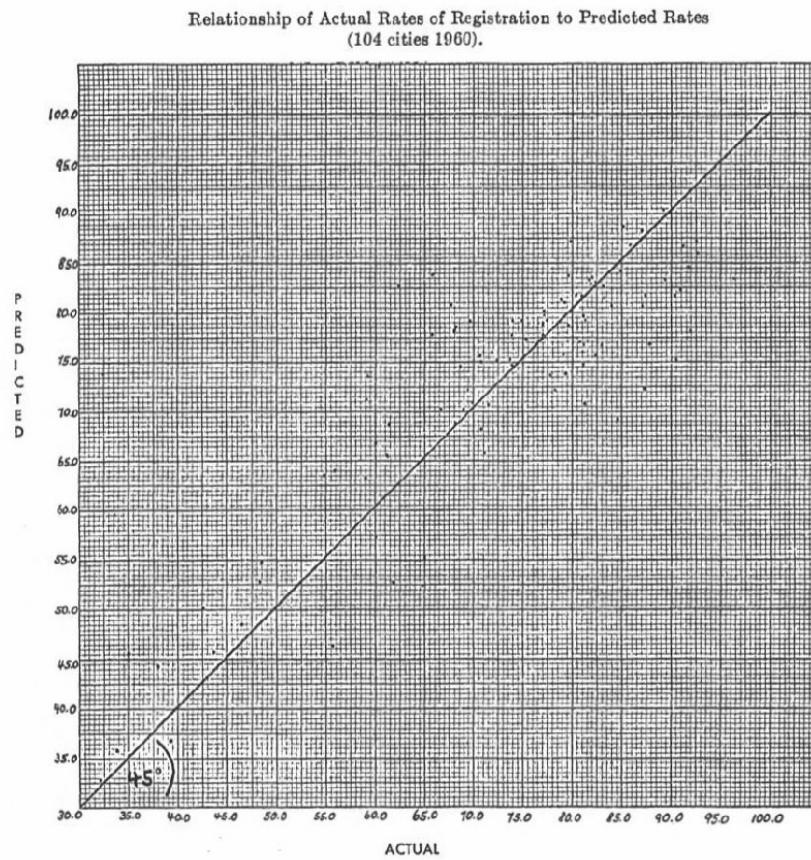
Apple **\$550.77**. Each sparkline has 244 data points, but it's condensed very nicely.

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).

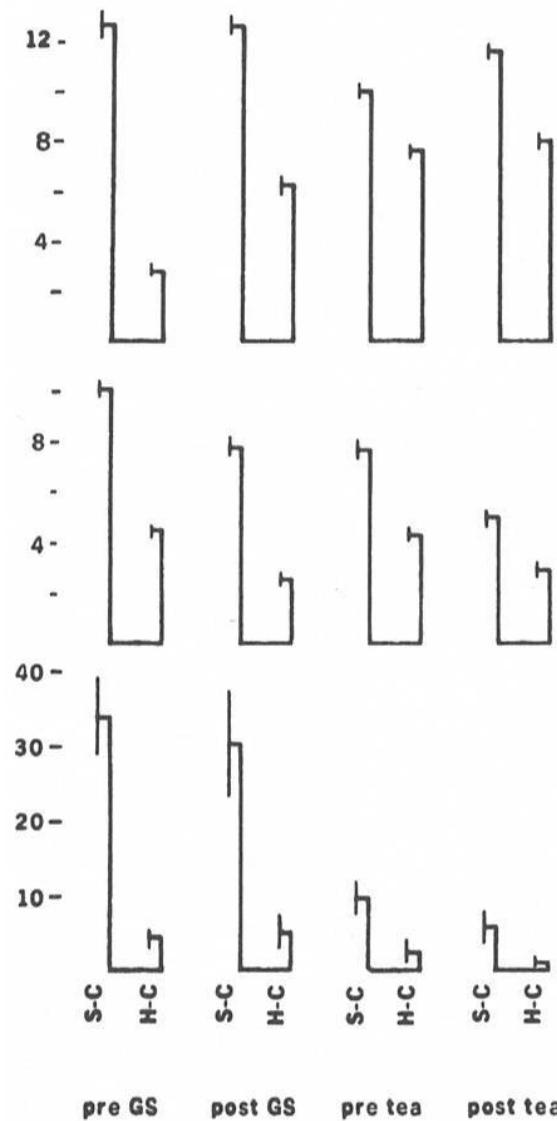
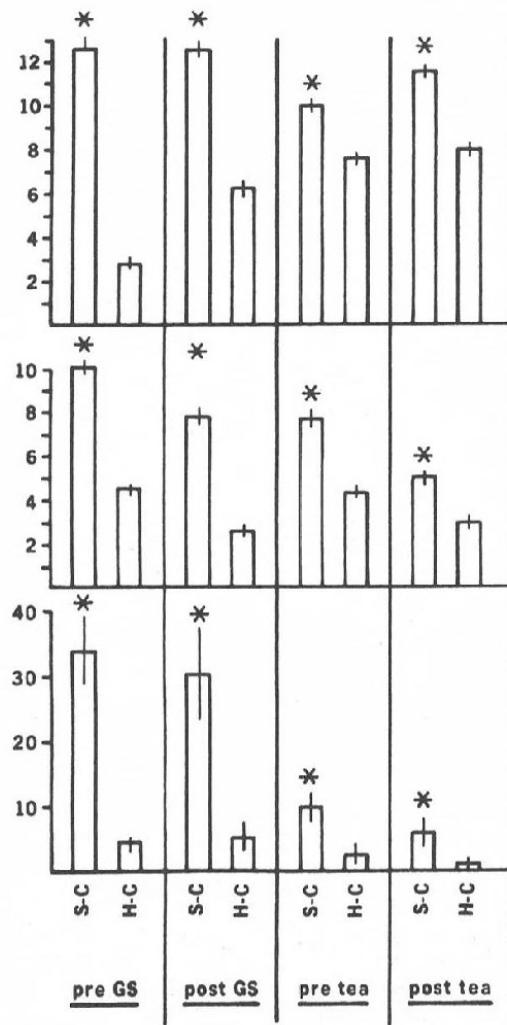
Maximize data-ink

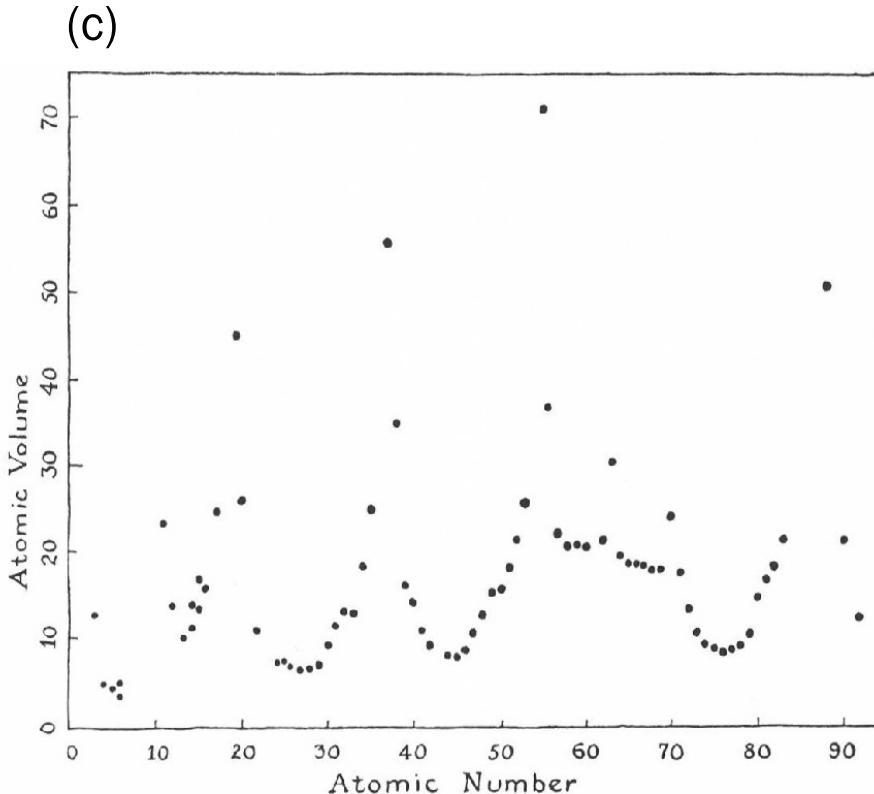
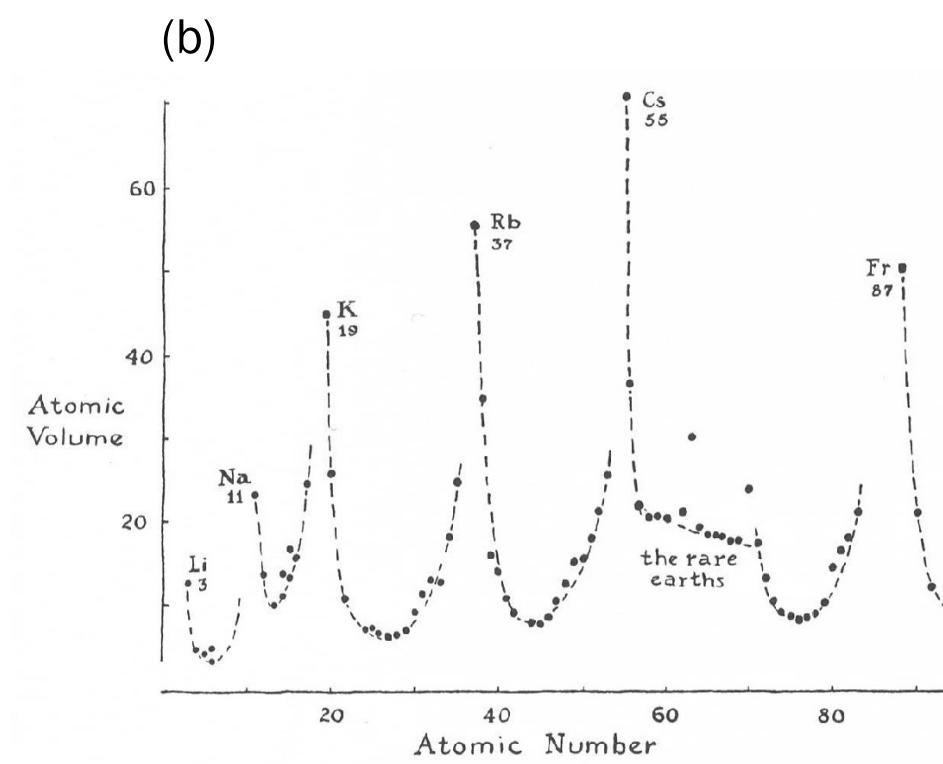
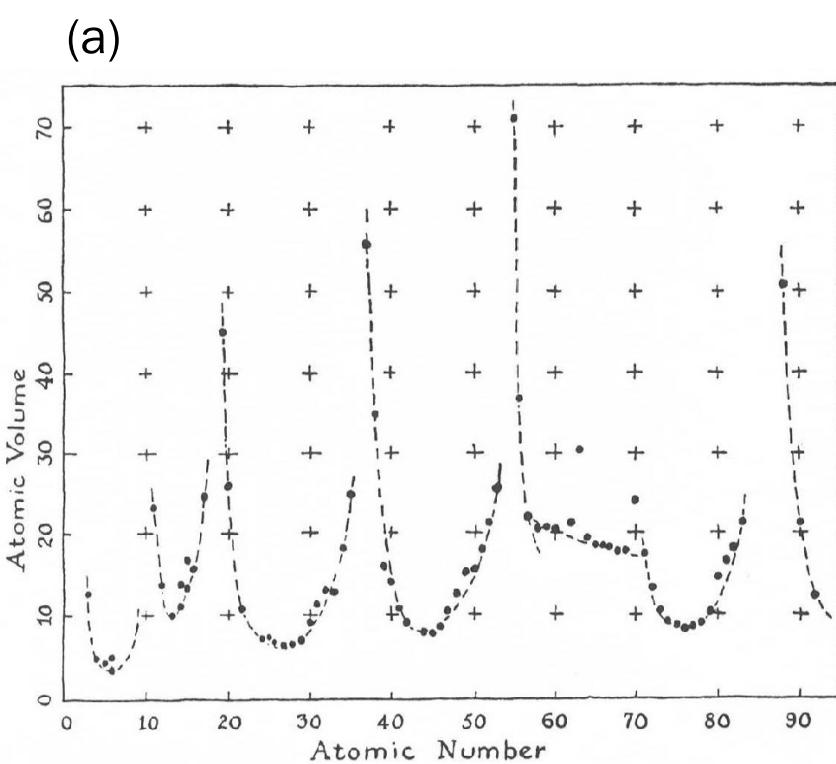
- It is always a good idea to maximize the data-ink ratio, within reason
- The larger the share of data-ink the better, other matters being equal
 - every bit of ink on a graphic needs a reason
 - nearly always that reason being that the ink presents new information
- Ink that fails to depict statistical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

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 does not add to the total information content

Edit and redesign



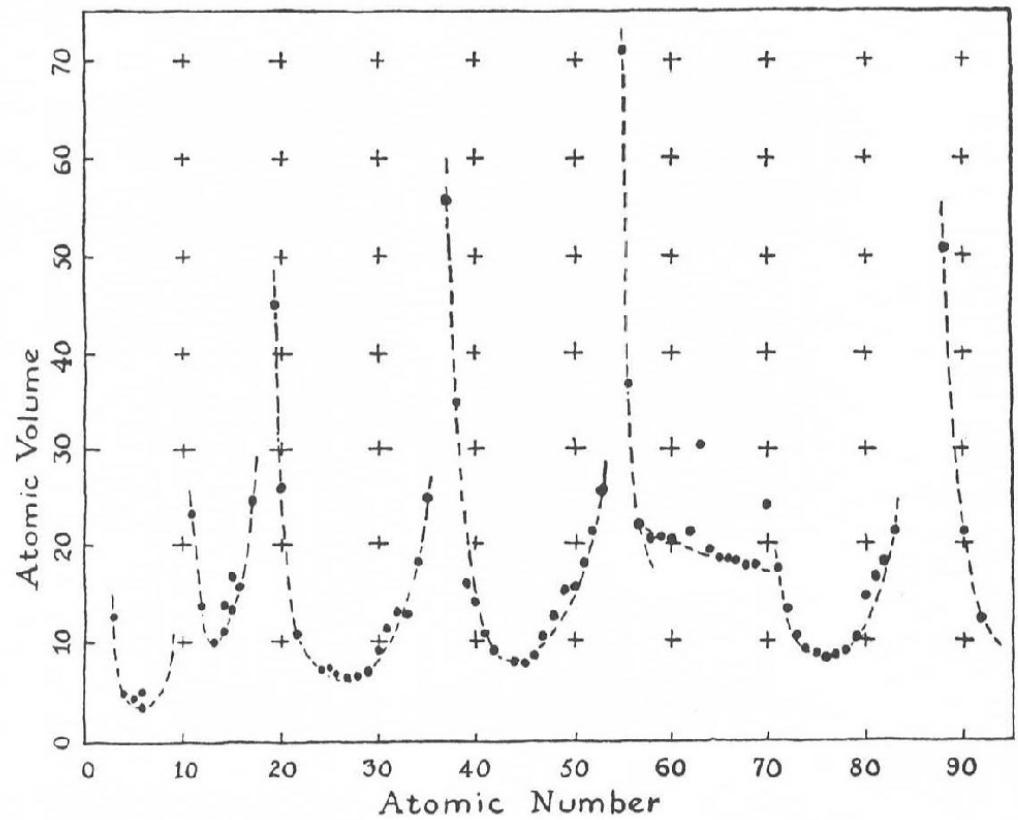


Atomic volume vs. atomic number

Which do you think is the best,
and why?

Edit and redesign

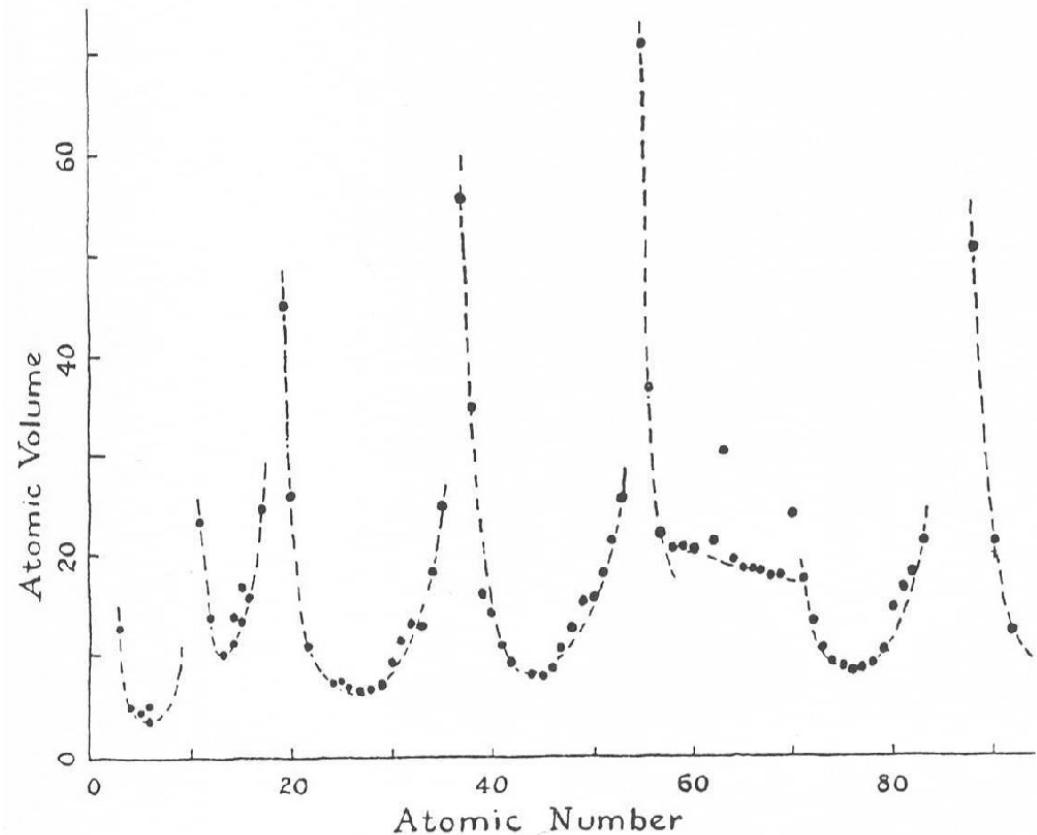
- The data-ink ratio is about 0.6
 - 76 data points and the reference curve are obscured by 63 grid marks
- The grid and part of the frame can be erased to improve the data-ink ratio



Linus Pauling, General Chemistry, p. 64, 1947

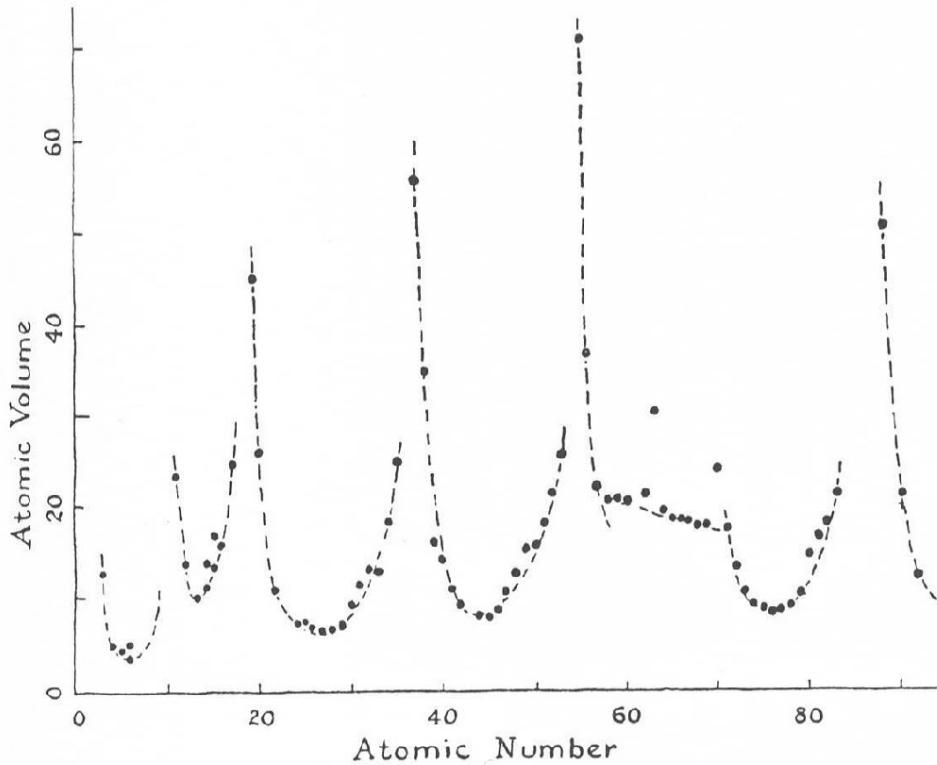
Edit and redesign

- 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

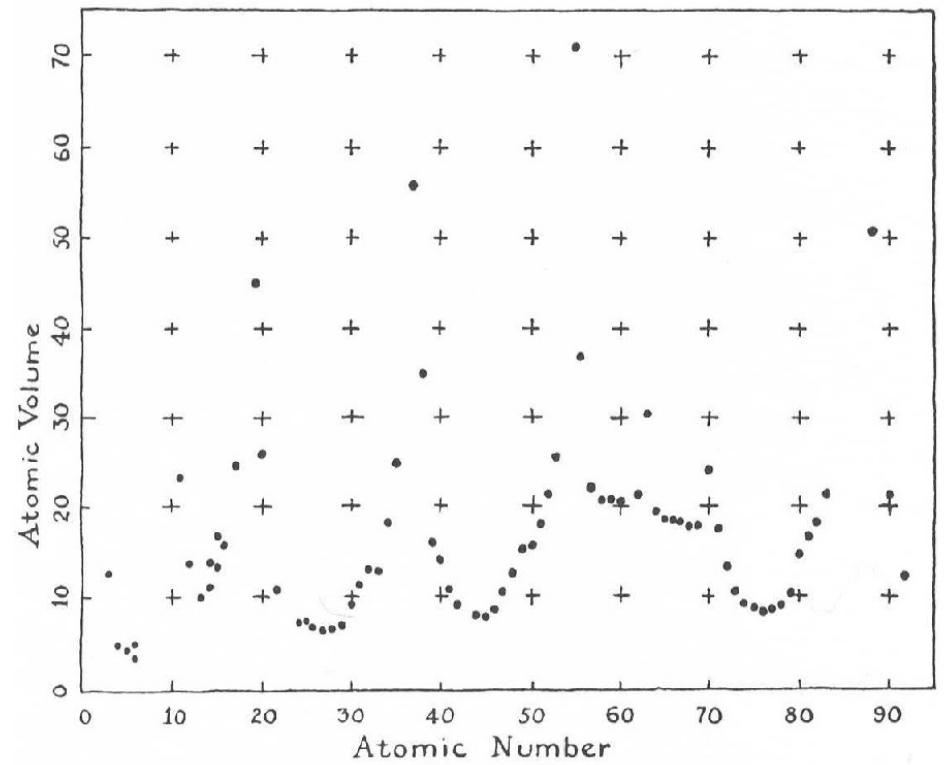


The reference curve is essential in organizing the data, and shows the periodicity (the message) by creating a structure, and by giving ordering and hierarchy

Edit and redesign



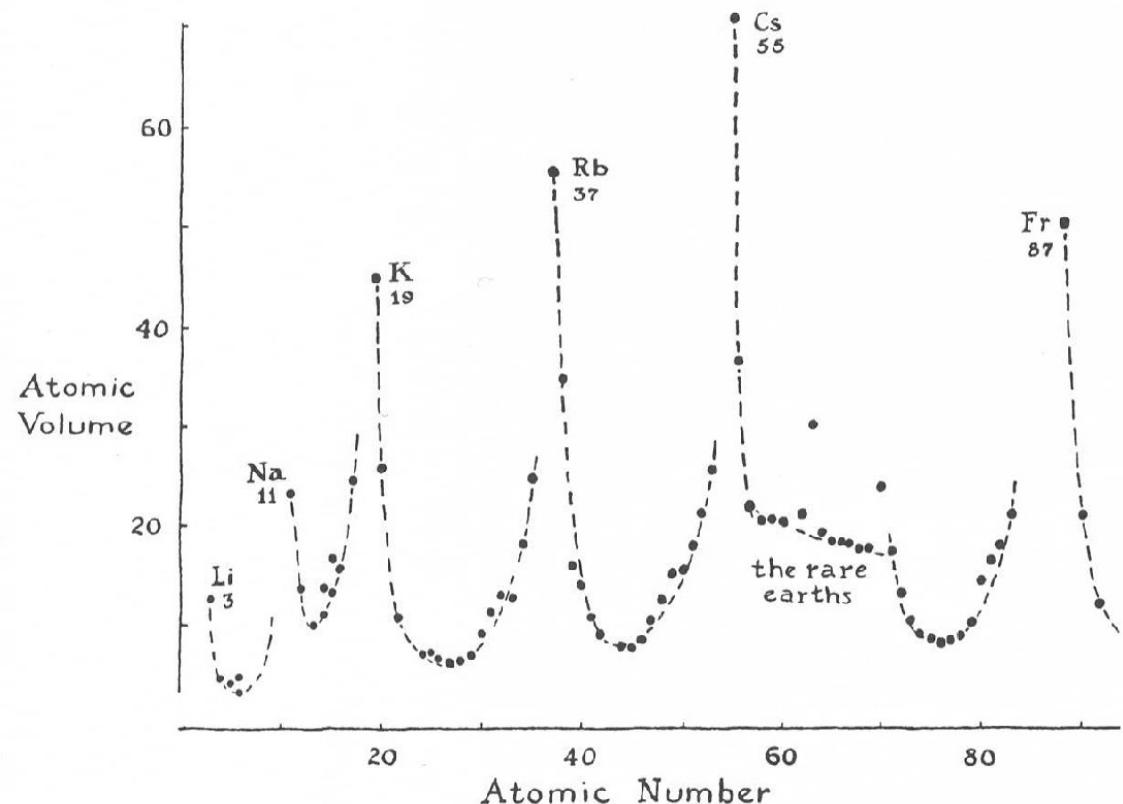
Without the curve we hardly detect the periodicity. The curve becomes necessary because the eye needs guidance



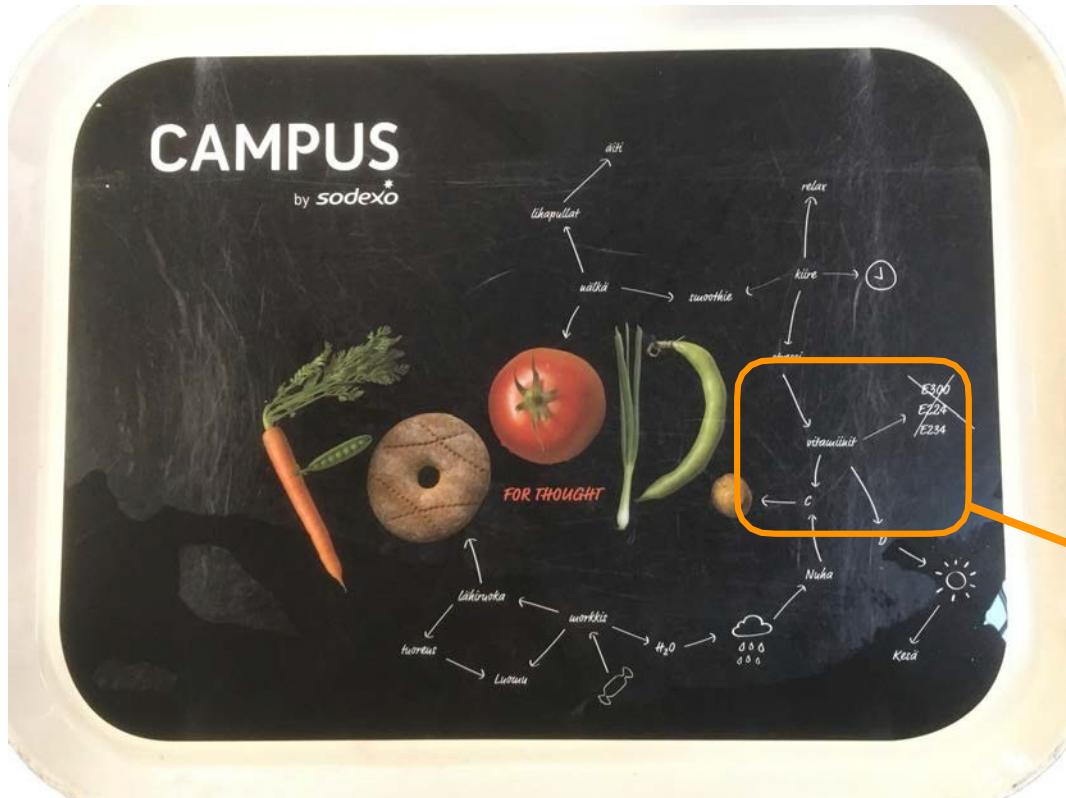
Restoring the grid totally fails to organize the data. The grid marks are too powerful and induce visual vibration.

Edit and redesign

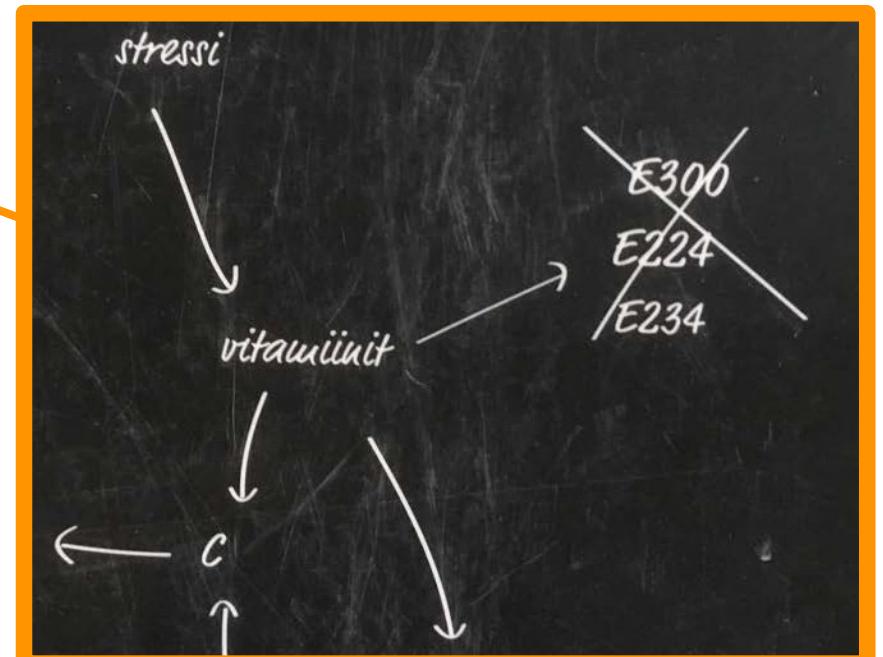
- We can use the erased space
 - labels for the initial elements of each period
 - unusual rare-earths
- also, turned label and numbers on the vertical axis
- Message: Always be critical about the first version of your graphics!



Check your information!



for discussion:
what's wrong?

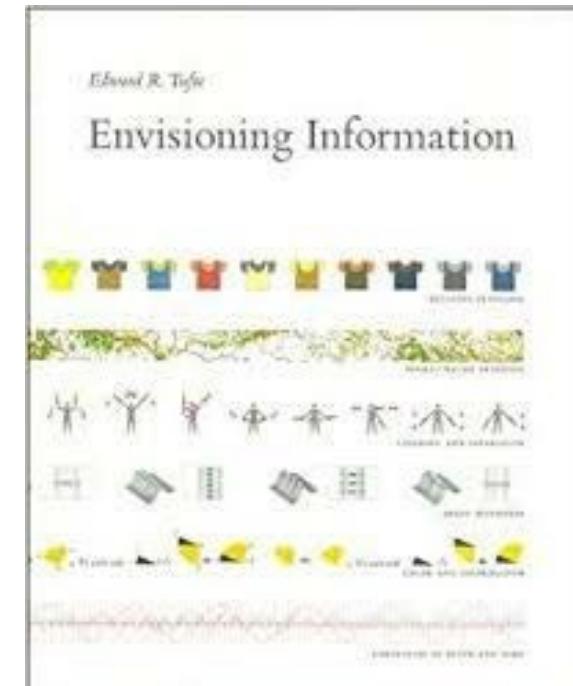


what is E300 ?

https://en.wikipedia.org/wiki/E_number

Edit and redesign

- Example of how to improve standard R scatterplot
- <http://www.iki.fi/kaip/p/iris.nb.html>
- examples from
Tufte: Envisioning Information

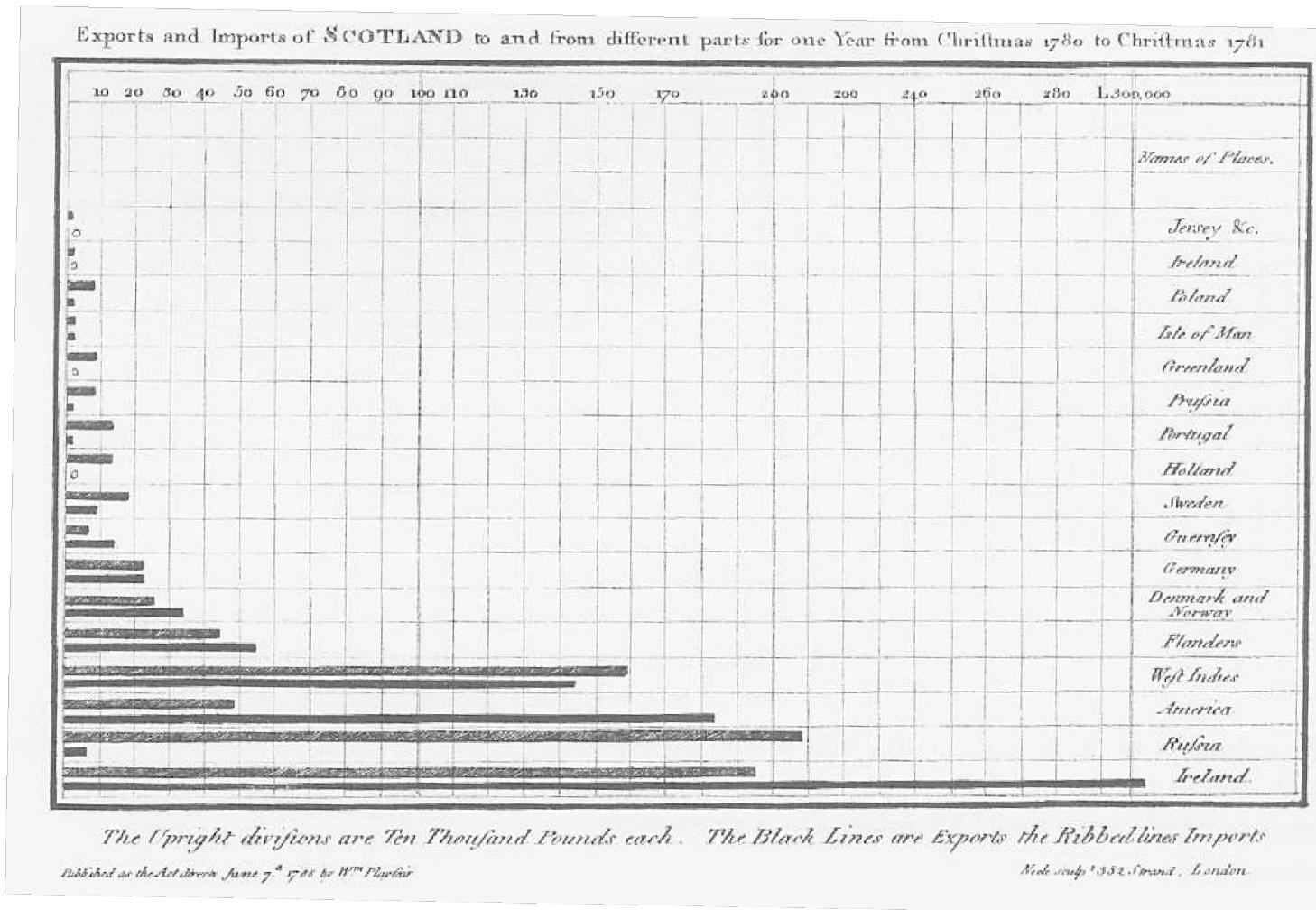


The five data-ink principles

- 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**.

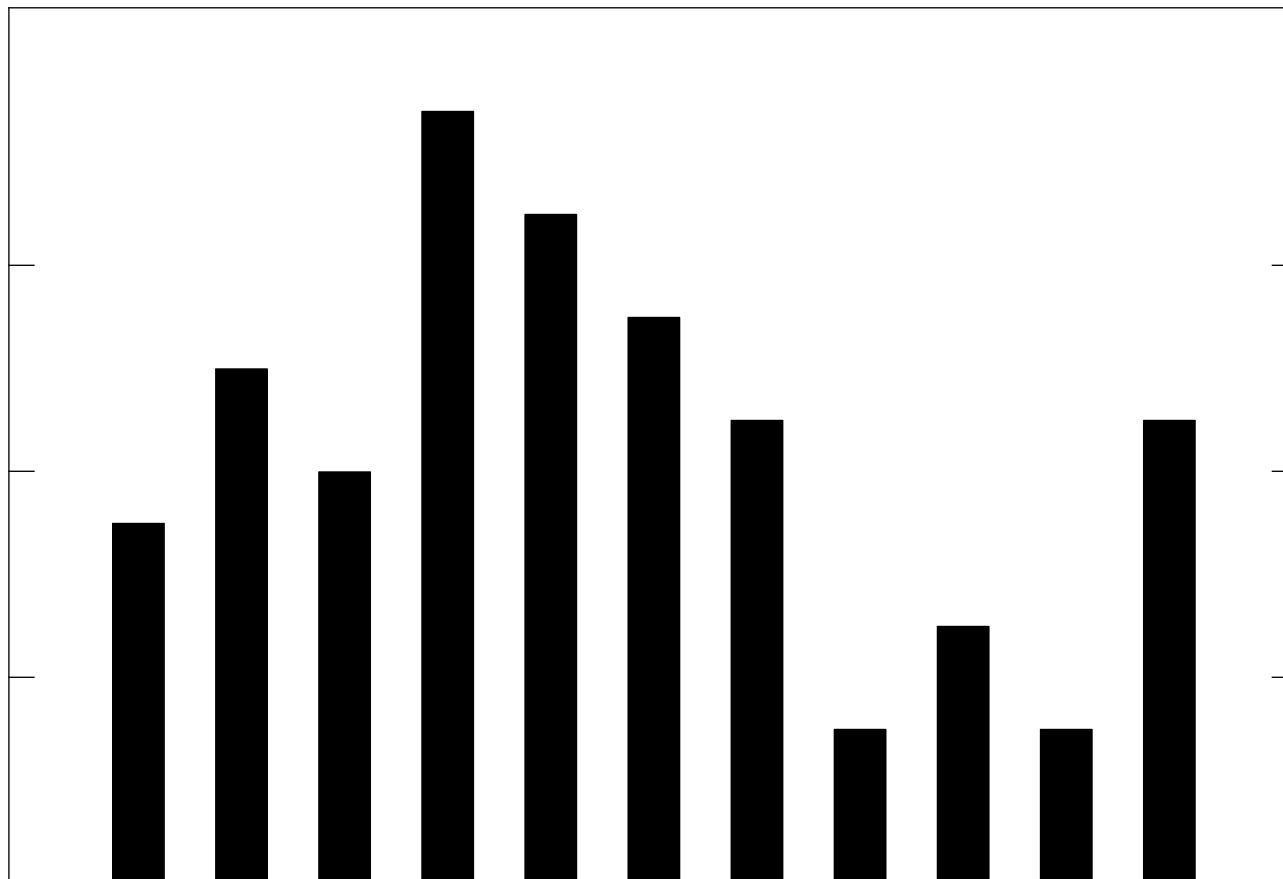
Bar chart

- One of the basic designs.



William Playfair, 1786.

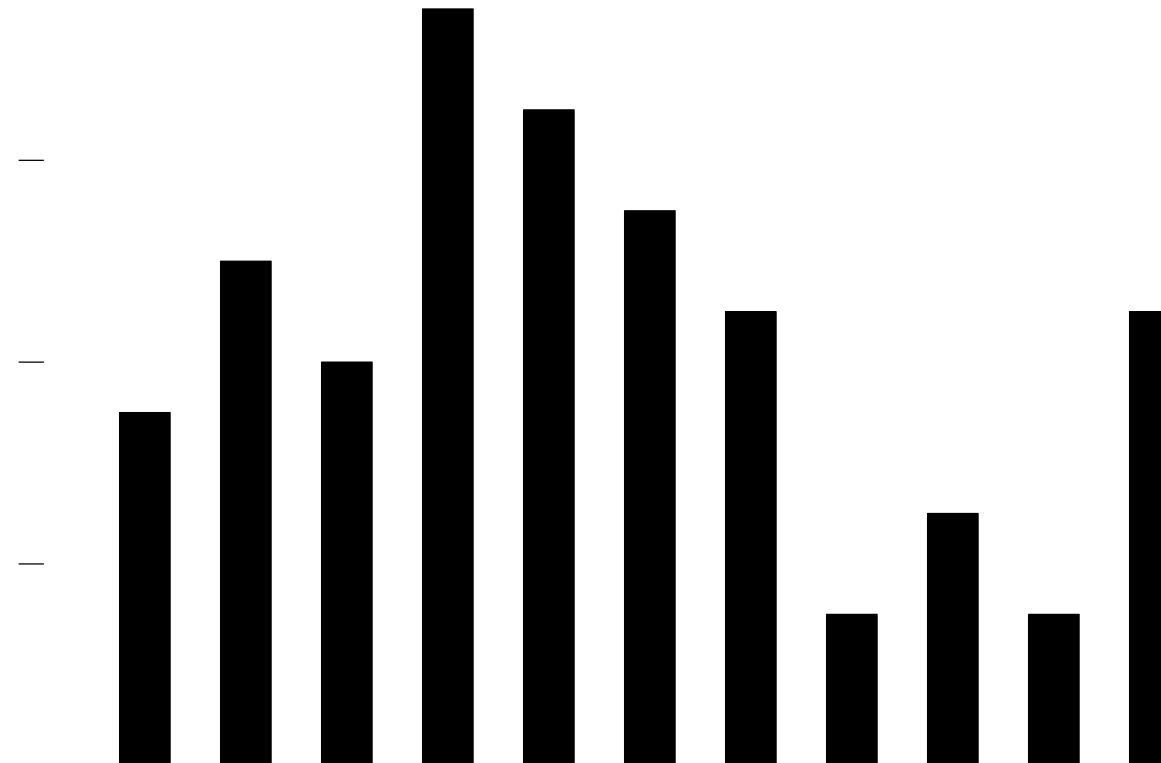
Bar chart



The standard bar chart

Bar chart

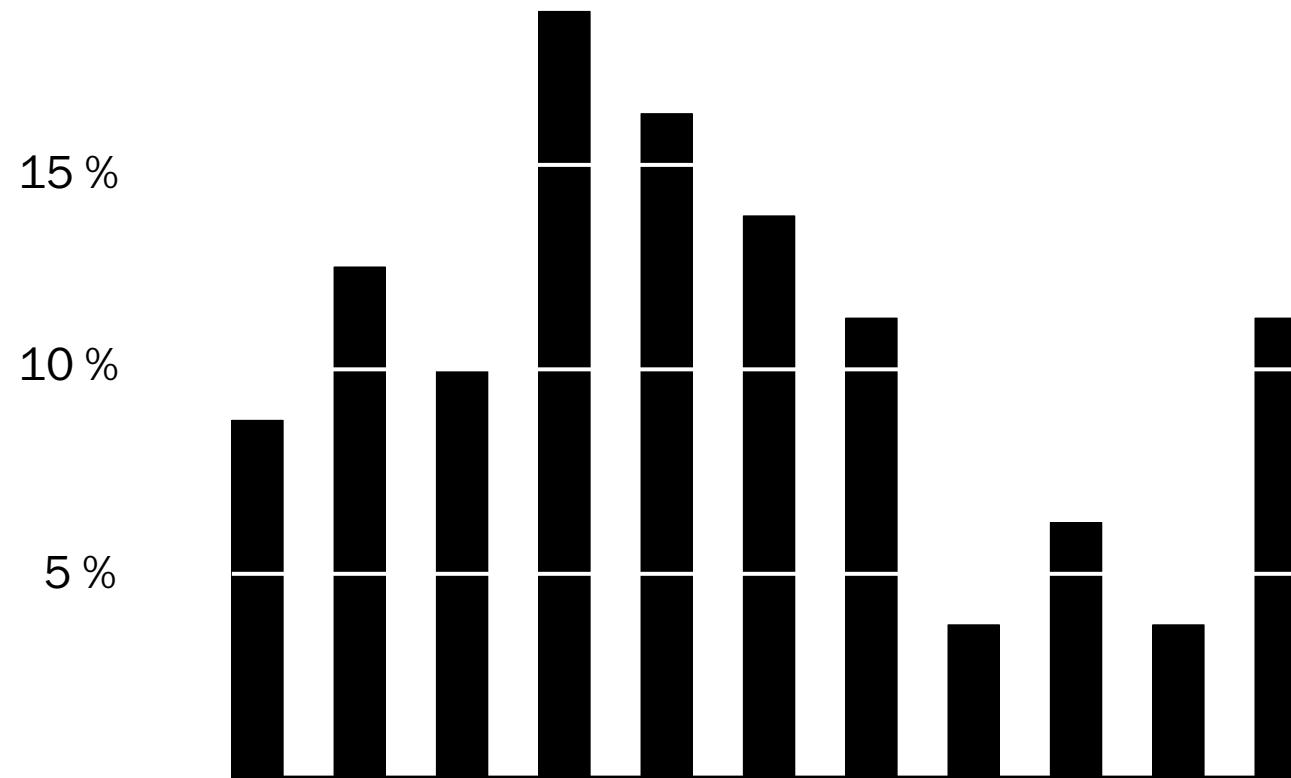
The box and vertical axis can be erased:



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

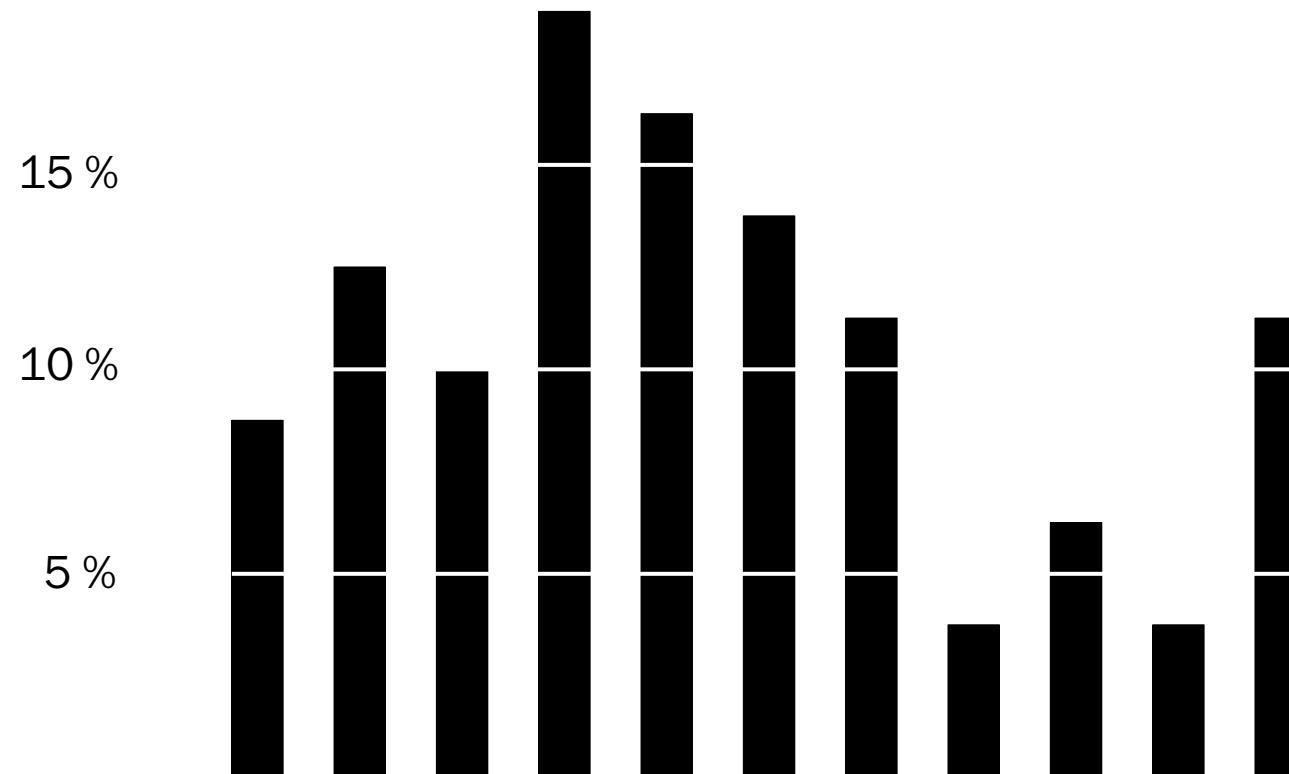
Bar chart

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



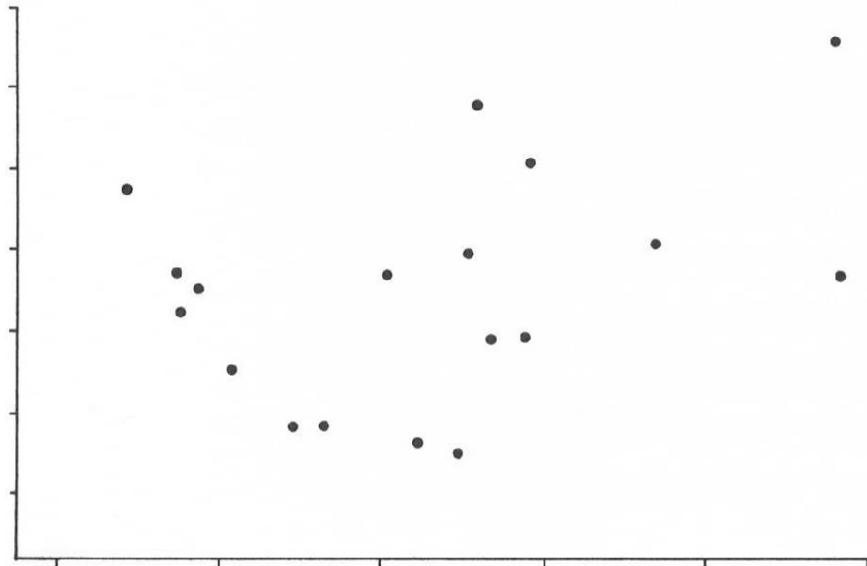
Bar chart

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

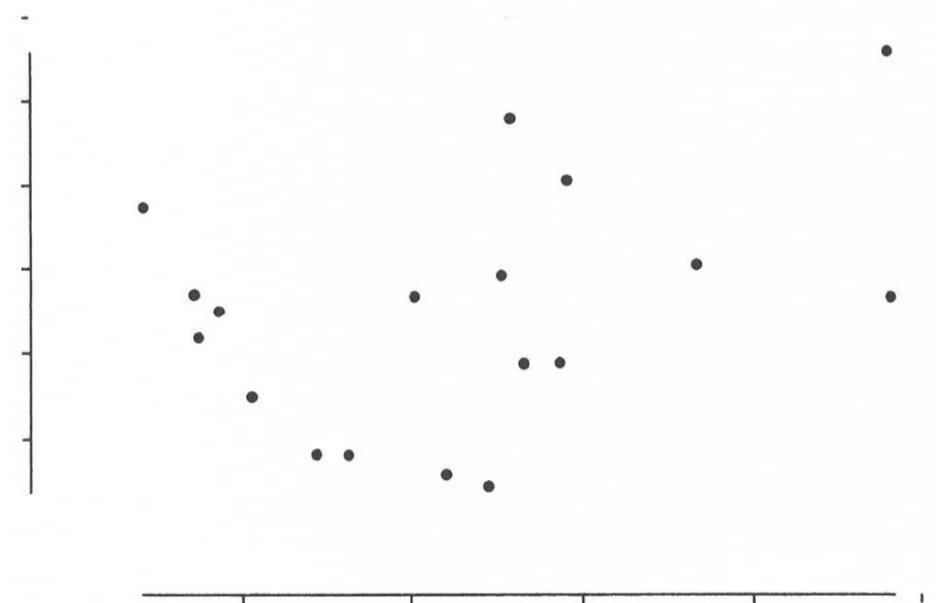


This might however be overdoing it (thin baseline looks good).

Scatterplot



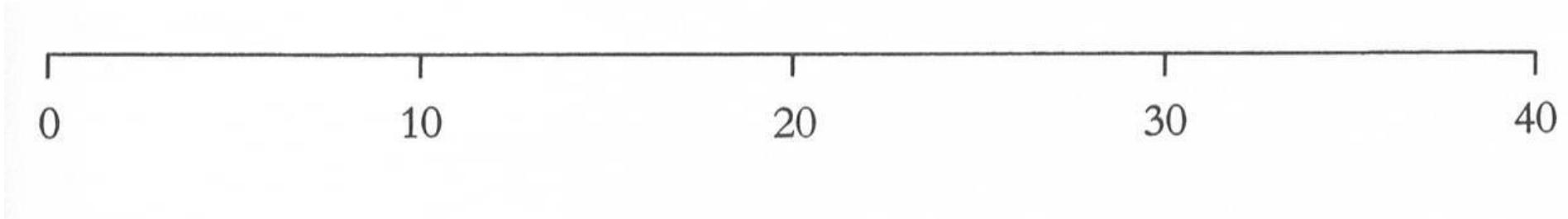
standard bivariate scatterplot



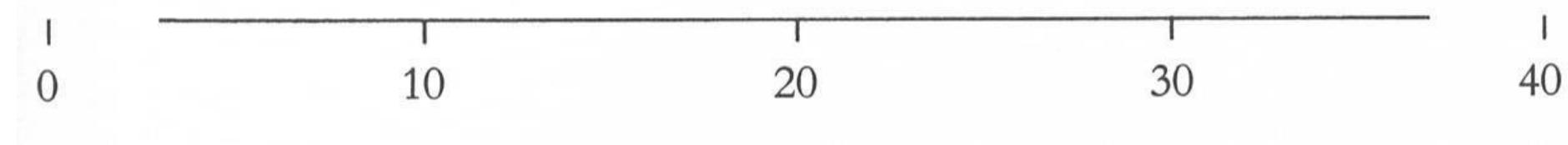
range frames indicate minimum and maximum values

Range frames

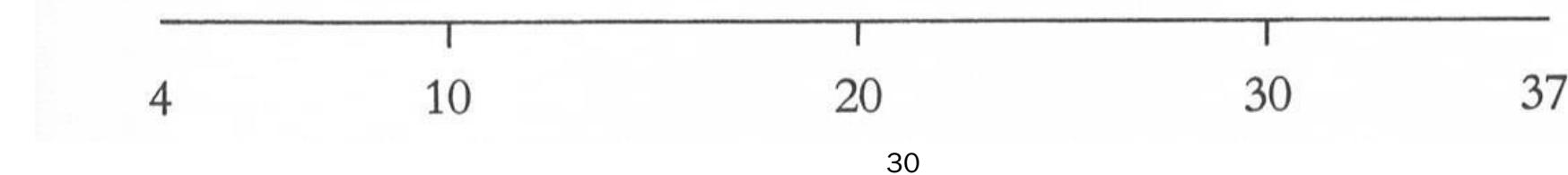
Standard frame:



Range frame:



Range frame with explicit limits (may work better):



Data values as frame

20.3

15.2
14.6

11.3
10.1
8.4

5.1

81

123

182

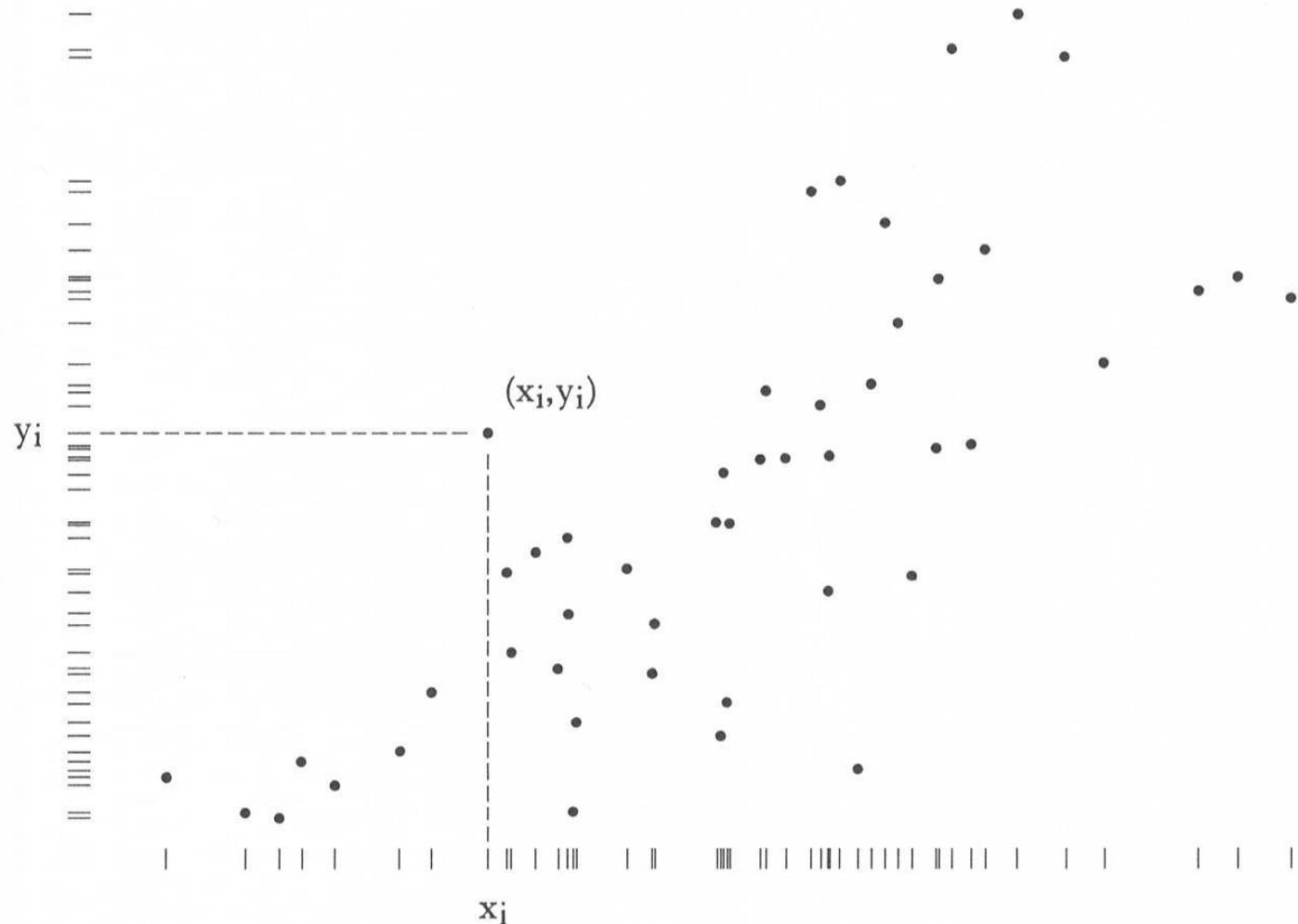
227

255

291

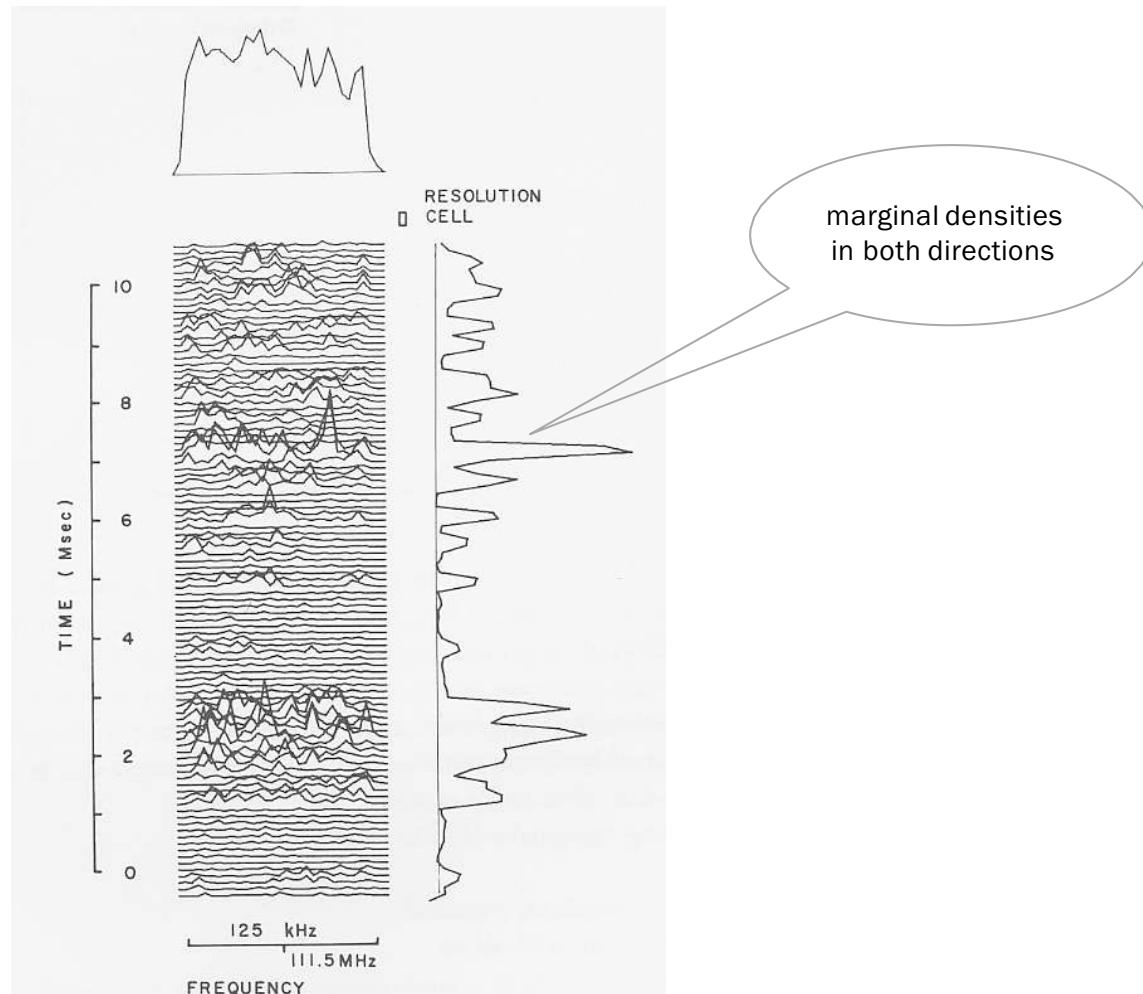
357

Rug plot as frame

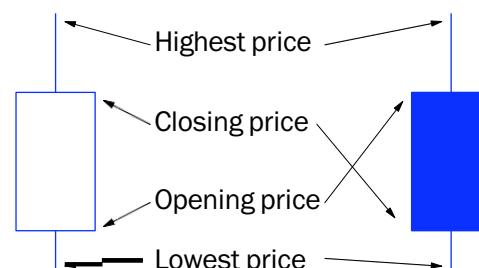
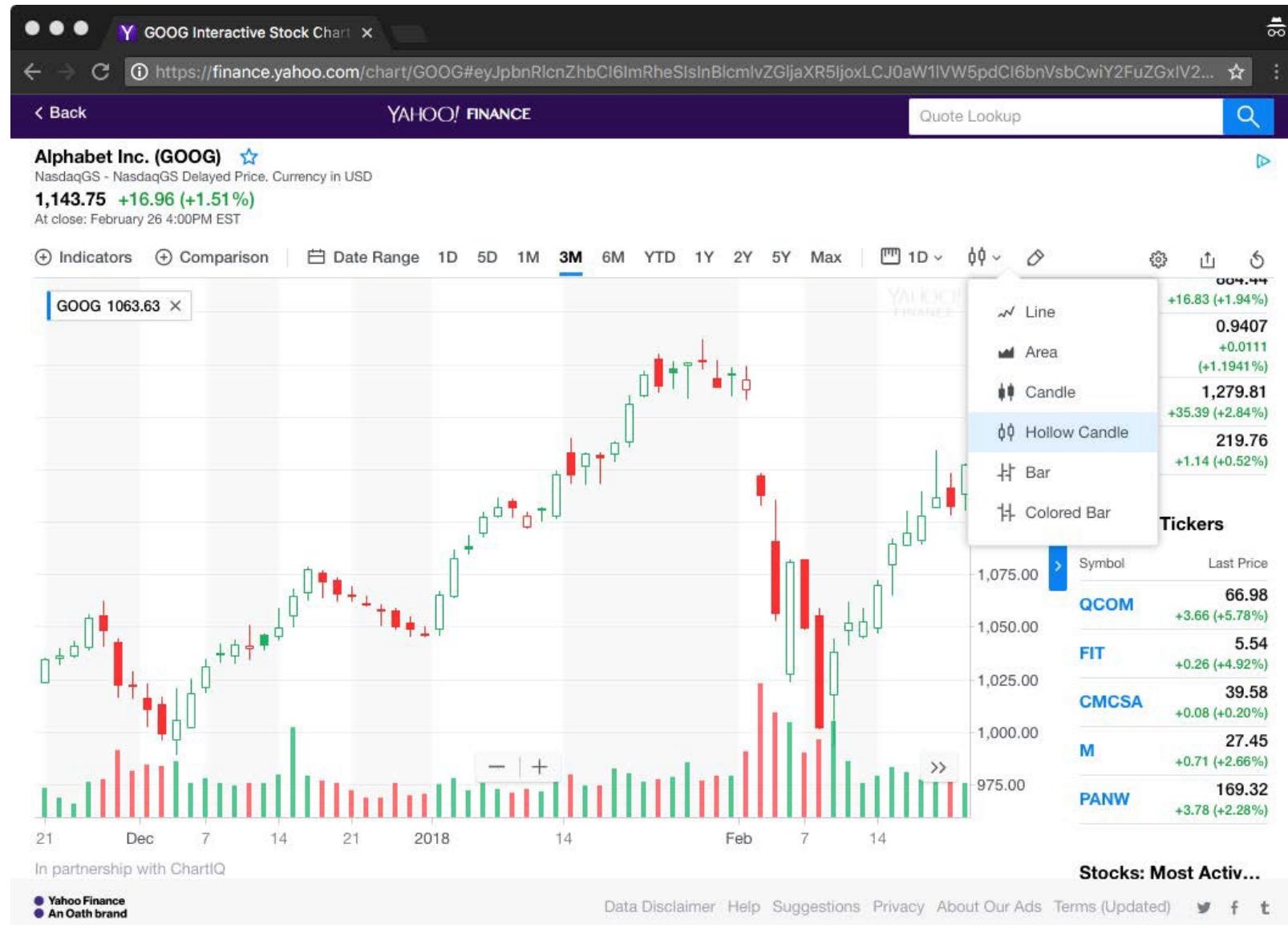


Dot-dash plot

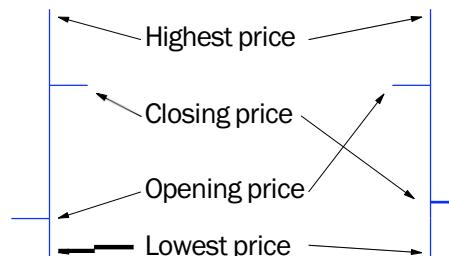
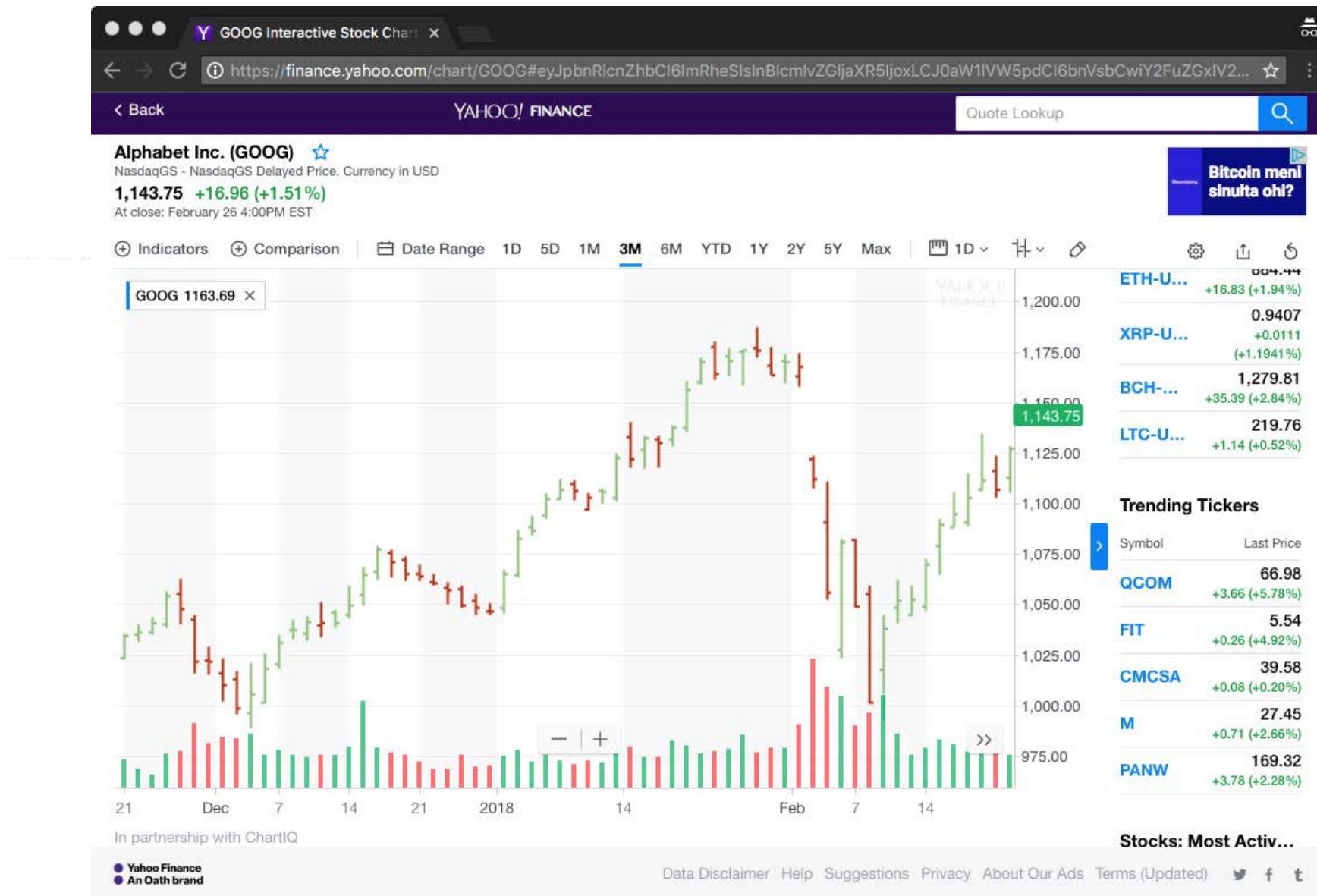
development
from rug plot



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



Range bars show opening and closing prices and price variation



The five data-ink 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.

Theory of data graphics

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

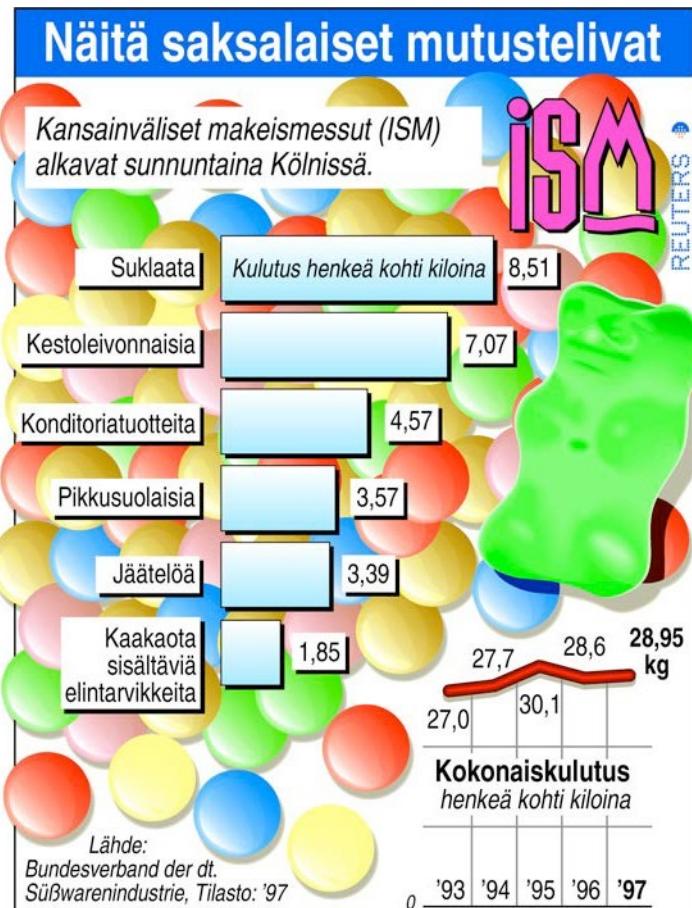
Chartjunk

- *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

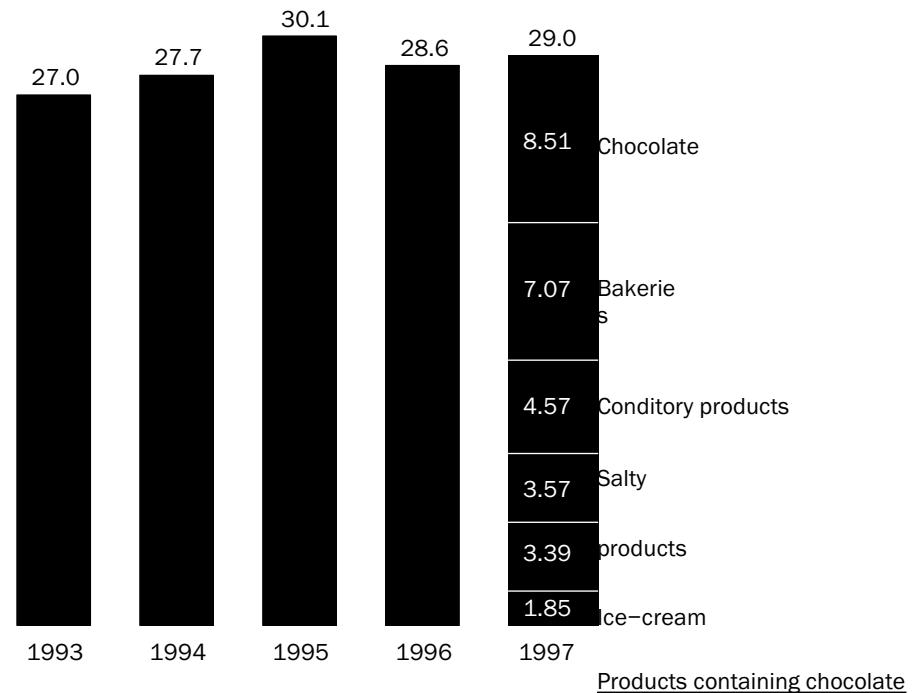
- Ducks (eye candy and self-promoting graphics)
- Vibrations
- Grids

Eye candy



Consumption of sweets in Germany

(kilograms per capita in a year)



H. Spissler, Reuters, 1999.

How much ice cream did a German eat in 1997?

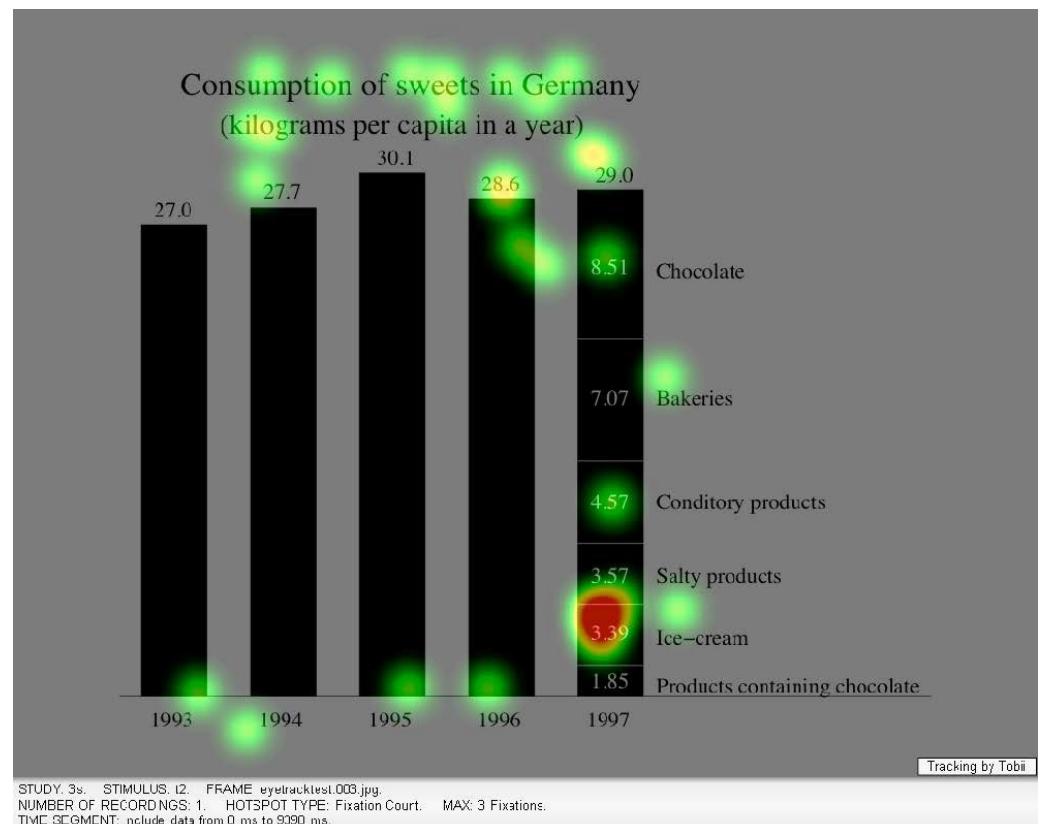
Eye tracking: what people really look at...



How much ice cream did a German eat in

Eye candy

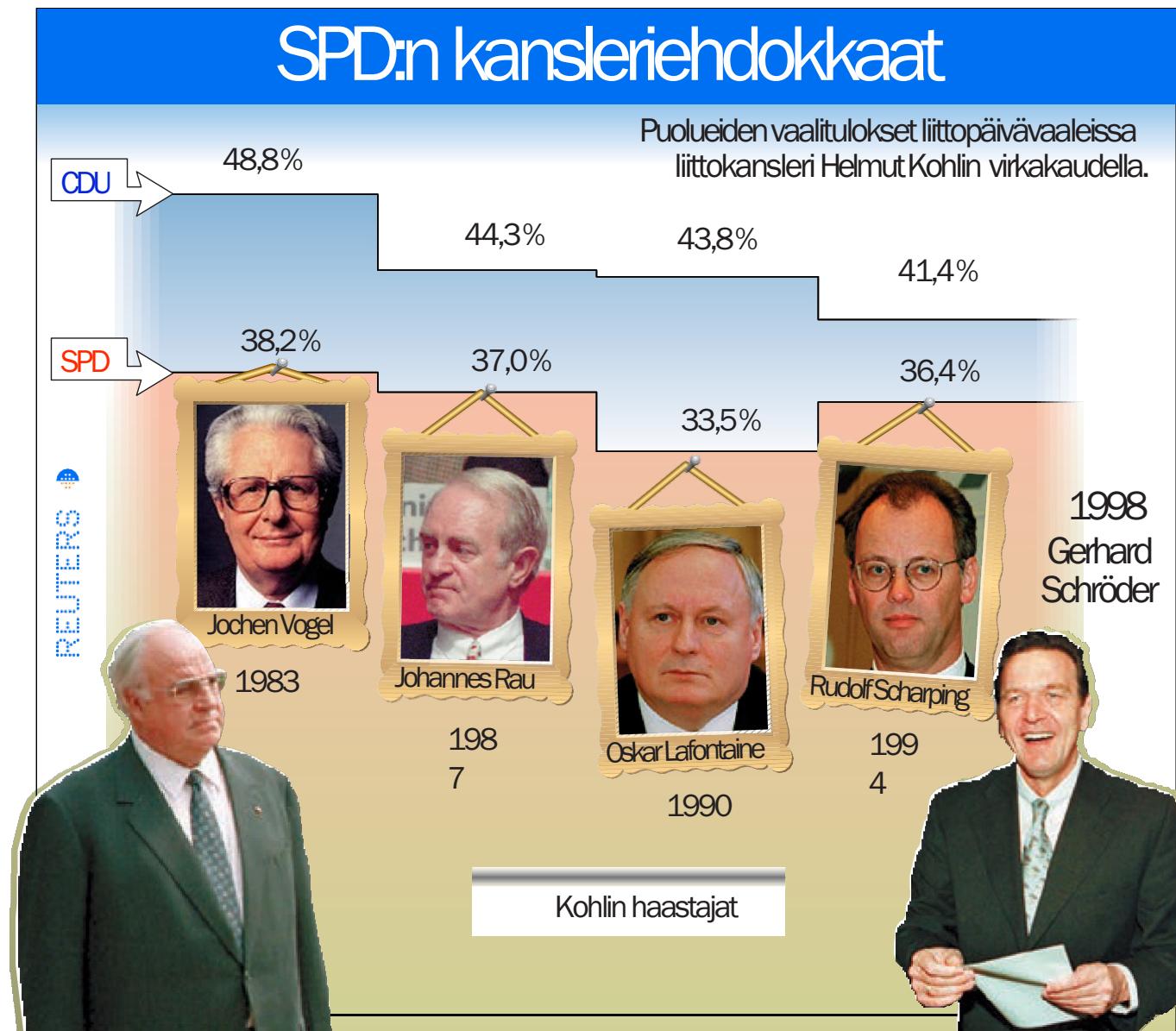
Where did they really look:



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

Facejunk

Reuters/H.
Spissler

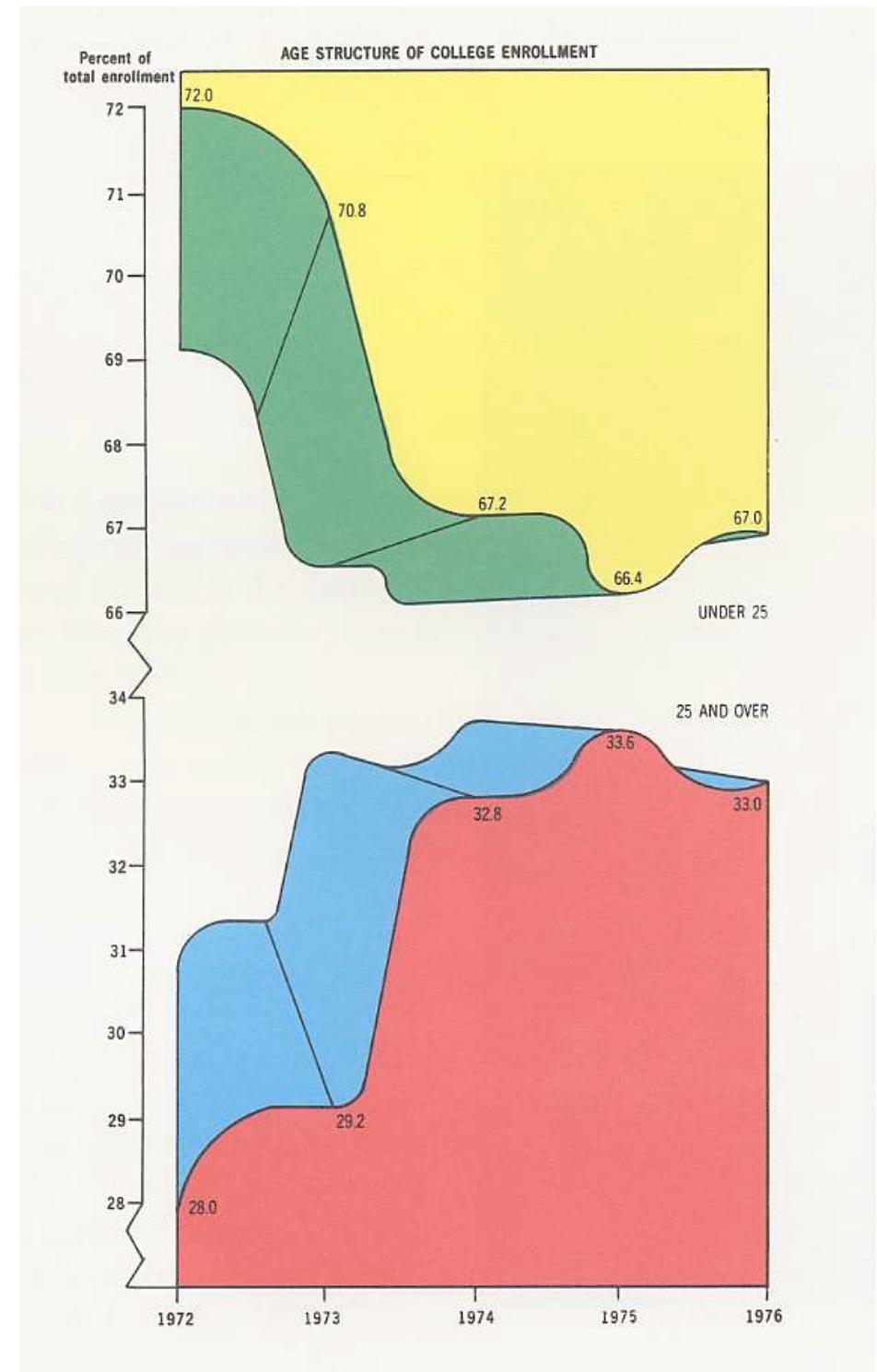


Self-promoting graphics

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

American education [T 118].

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



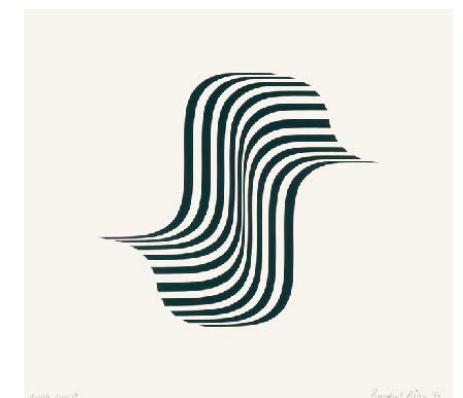
Visual stress (vibrations)

- Striped patterns cause visual stress in most people.
- The following combination is most potent:

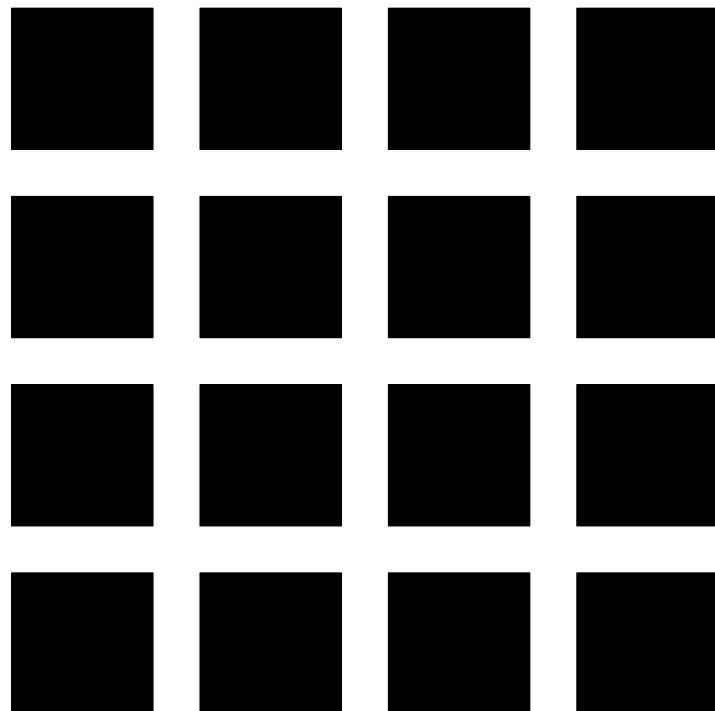
- about 3 cycles per degree
- flicker rate of about 20 Hz
- large patterns



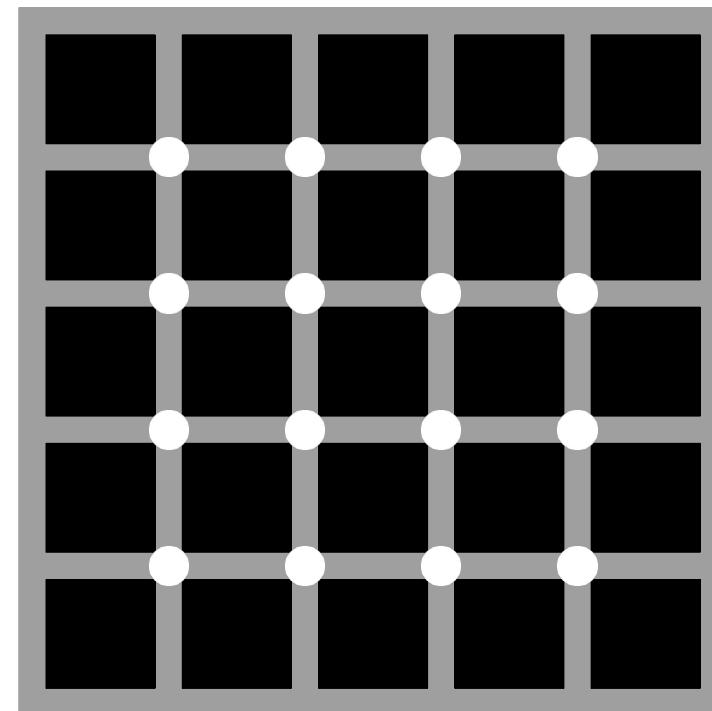
op-art by
Bridget Riley



Optical effects



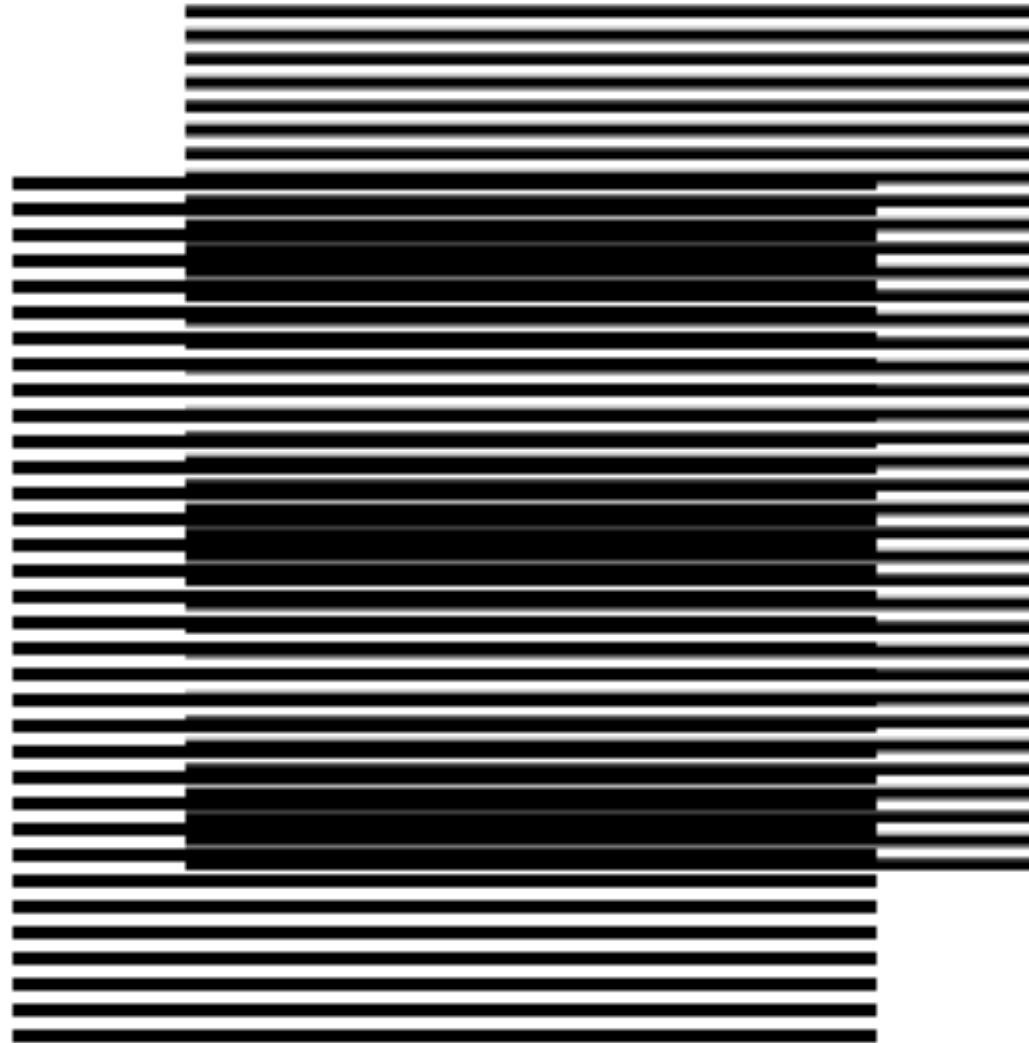
Hermann Grid illusion



Modified Hermann Grid illusion

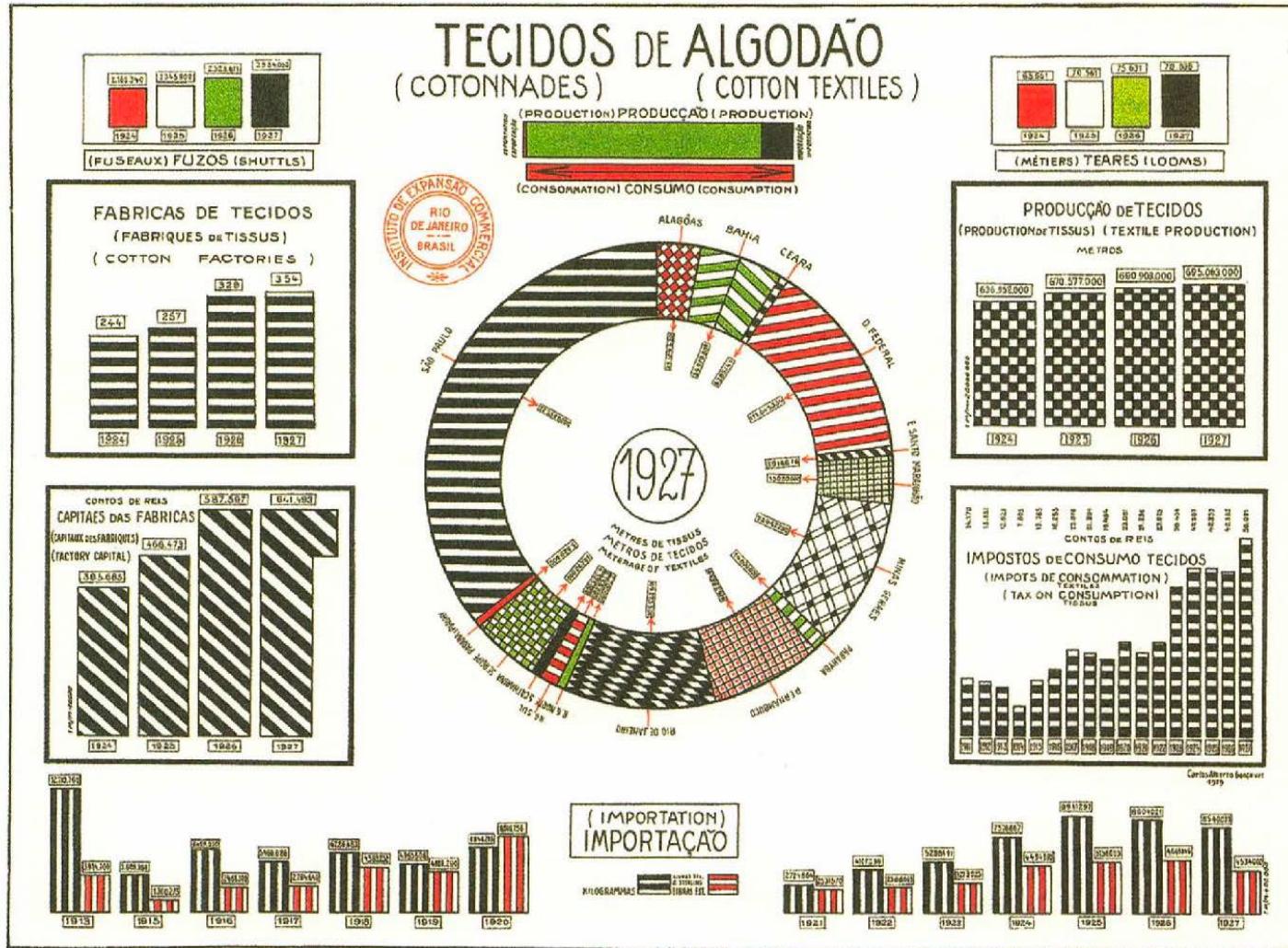
There appear to be dark spots at the intersection of the bright lines. Similar effects can appear in data graphics.

Moirè effects



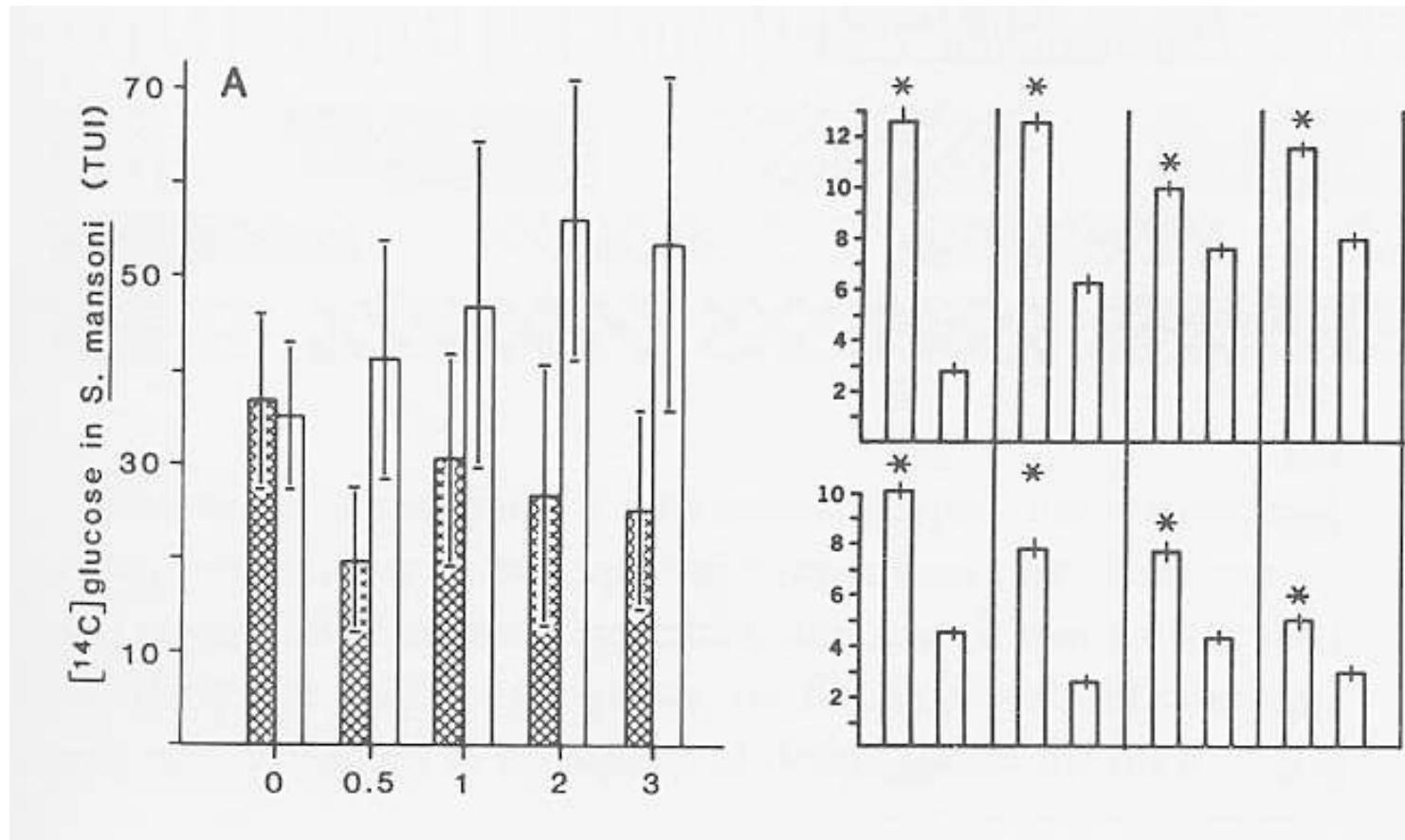
Moirè effects

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



Moirè effects

Moirè vibration appears at a maximum for equally spaced bars:

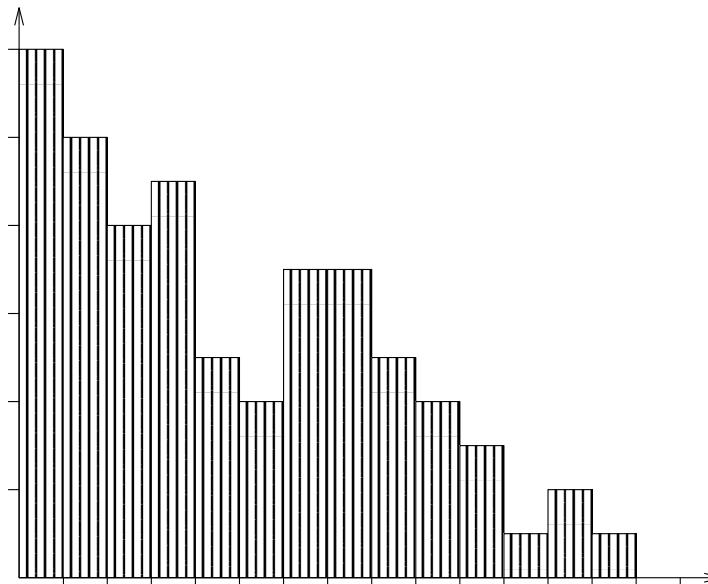
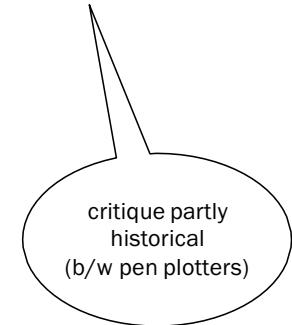
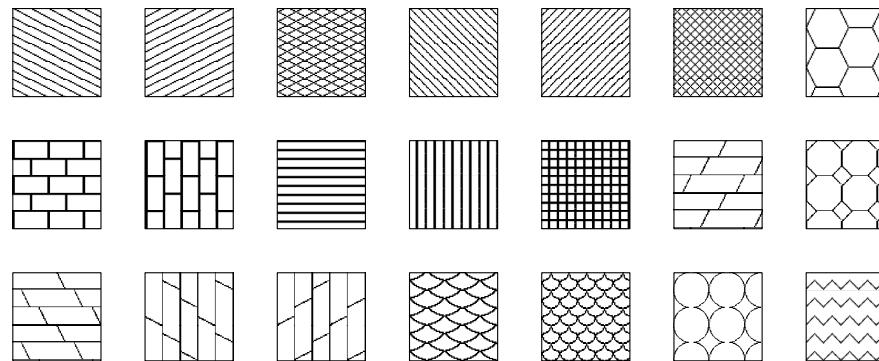


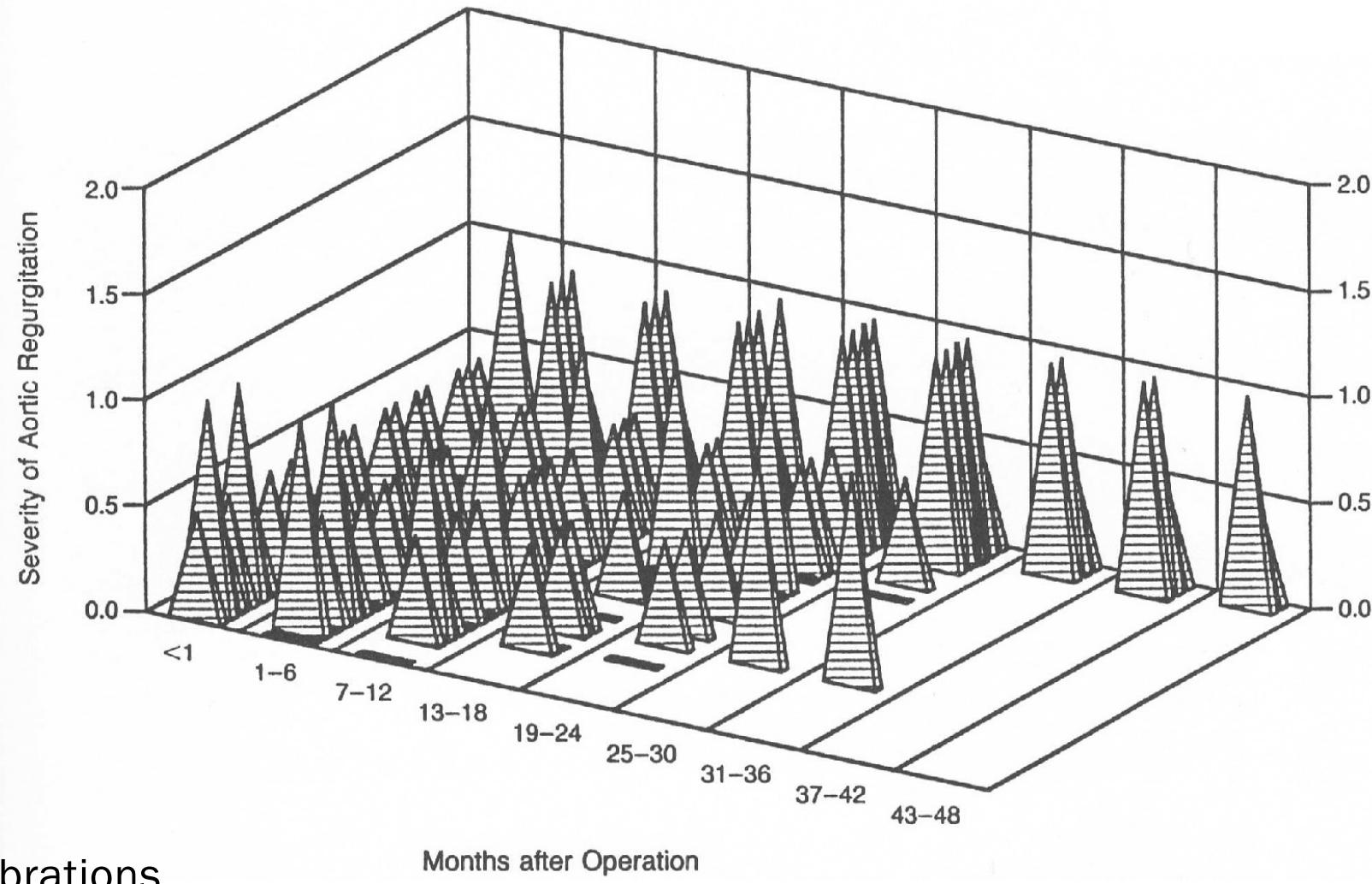
James T. Kuznicki, N. Bruce McCutcheon, 1979 [T 109].

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

Moirè effects

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





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

Grid lines

455865876864565749286555584765298742309847249473247

324879427149572389742982479280742938742564875647654

902842968476745464274784674573847648562484789847985

455865876864565749286555584765298742309847249473247

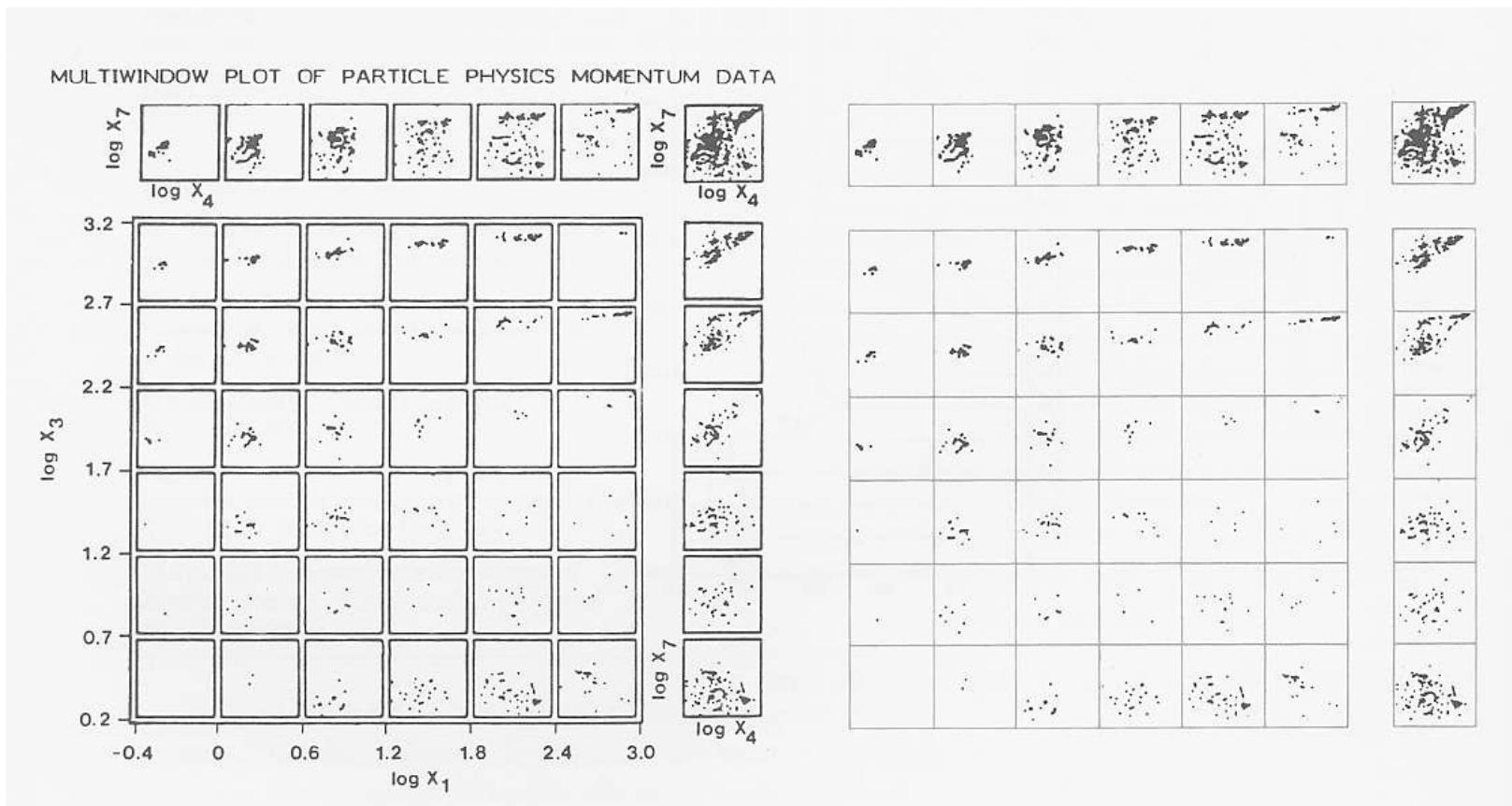
324879427149572389742982479280742938742564875647654

902842968476745464274784674573847648562484789847985

- 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

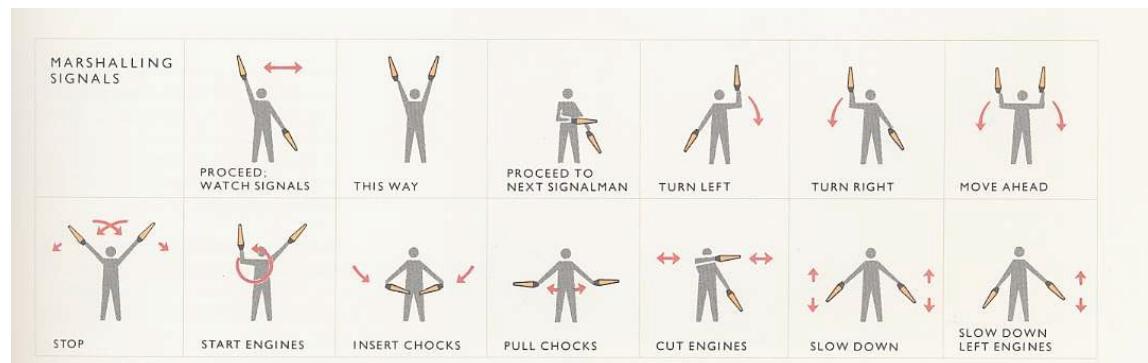
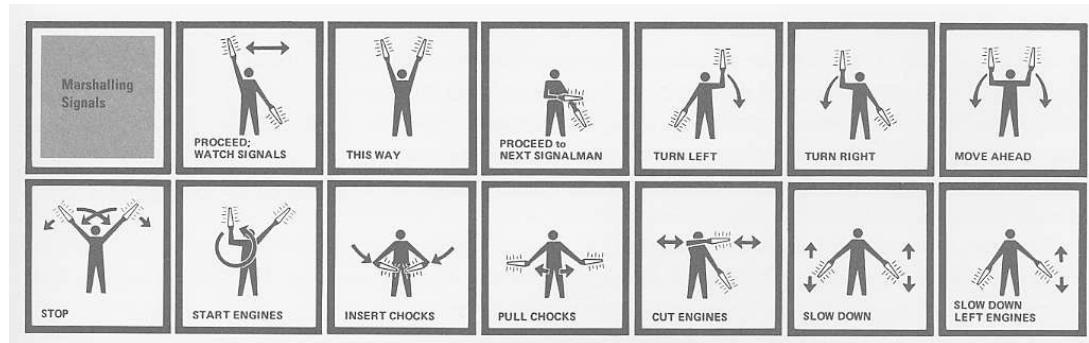
The doubled grid consumes 18 % of the area of this plot. Optical white dots appear at the intersection of the grid lines. Redrawing eliminates the vibration:



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

Grid lines

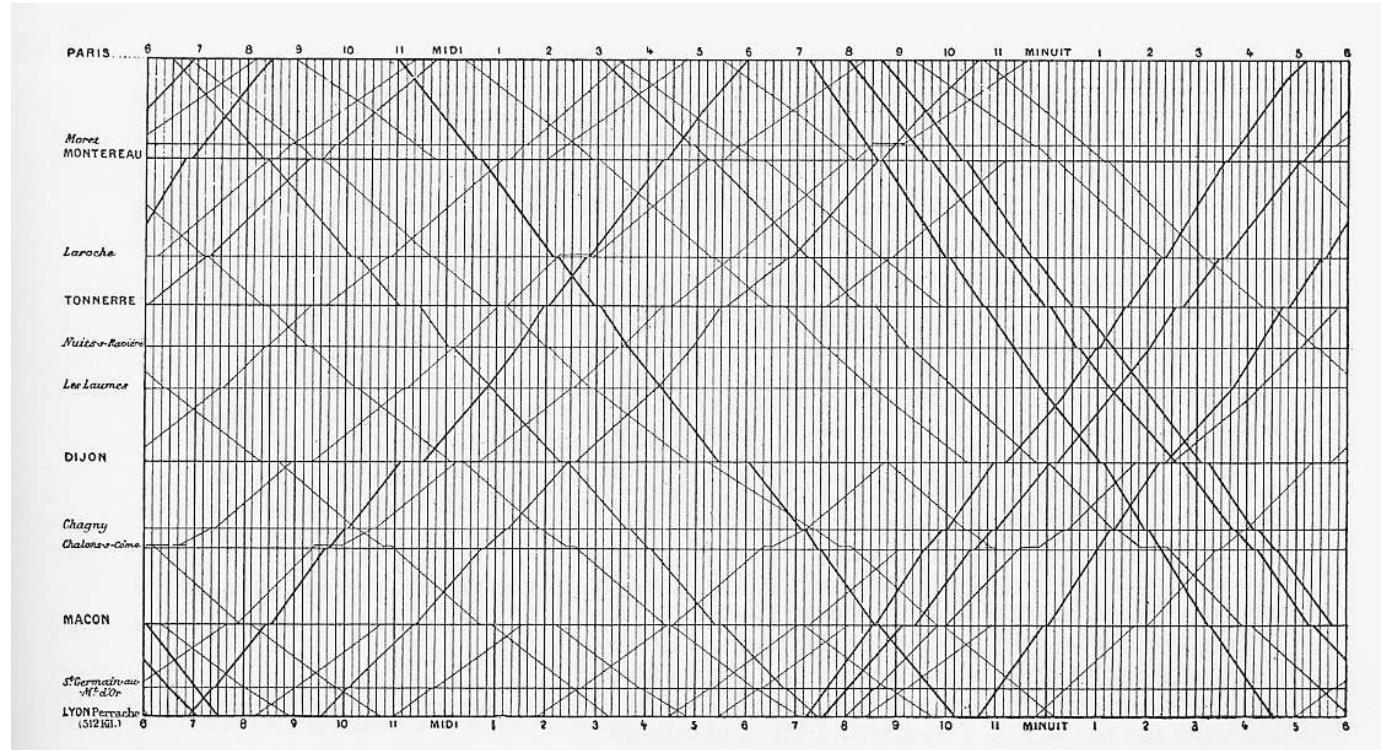
The grid dominates the graphics. The font is disproportionately weak as compared to the grid. Optical dark spots appear at the intersection of the white grid lines. Redrawing fixes this. The information content is further emphasized by conservative use of color.



[EI 63].

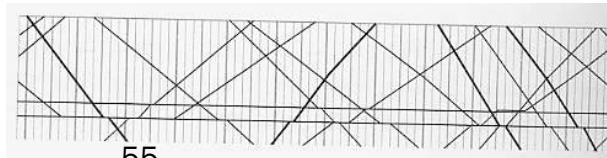
Grid lines

The train schedule by Marey has some Moirè vibration:



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

Thinning (or removing!) the grid lines helps:



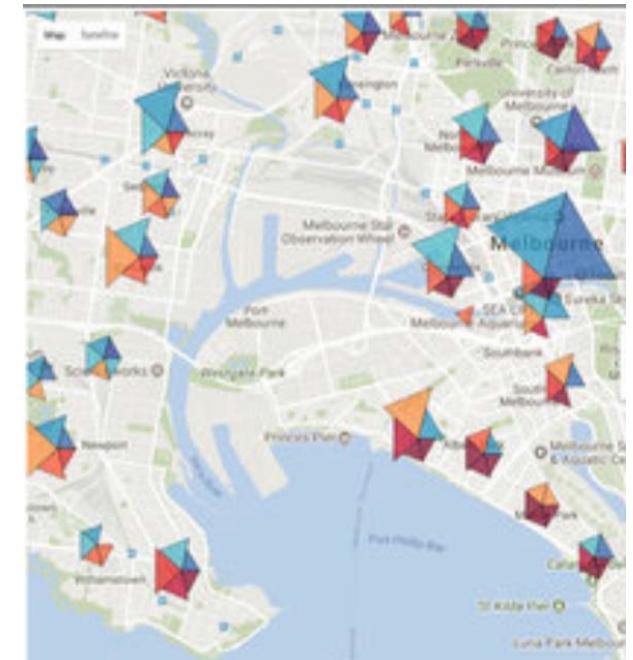
Theory of data graphics

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

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

glyphs on a map



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 | 00002223333333557777777888822335777888
 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 | 0222233555778000000023333357778888
 46 | 0000233357700000023578
 48 | 00000022335800333
 50 | 0370
```

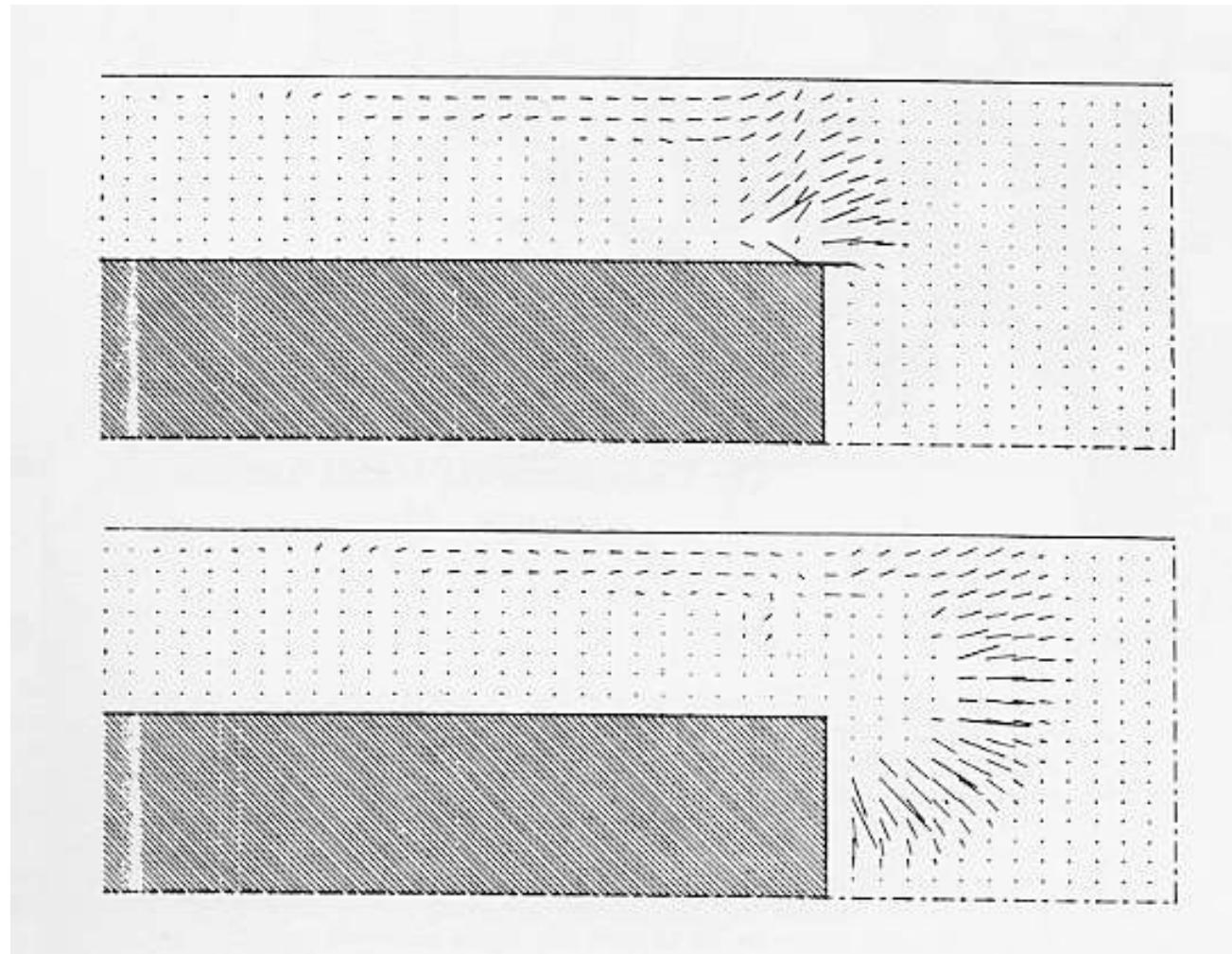
Stem-and-leaf display

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



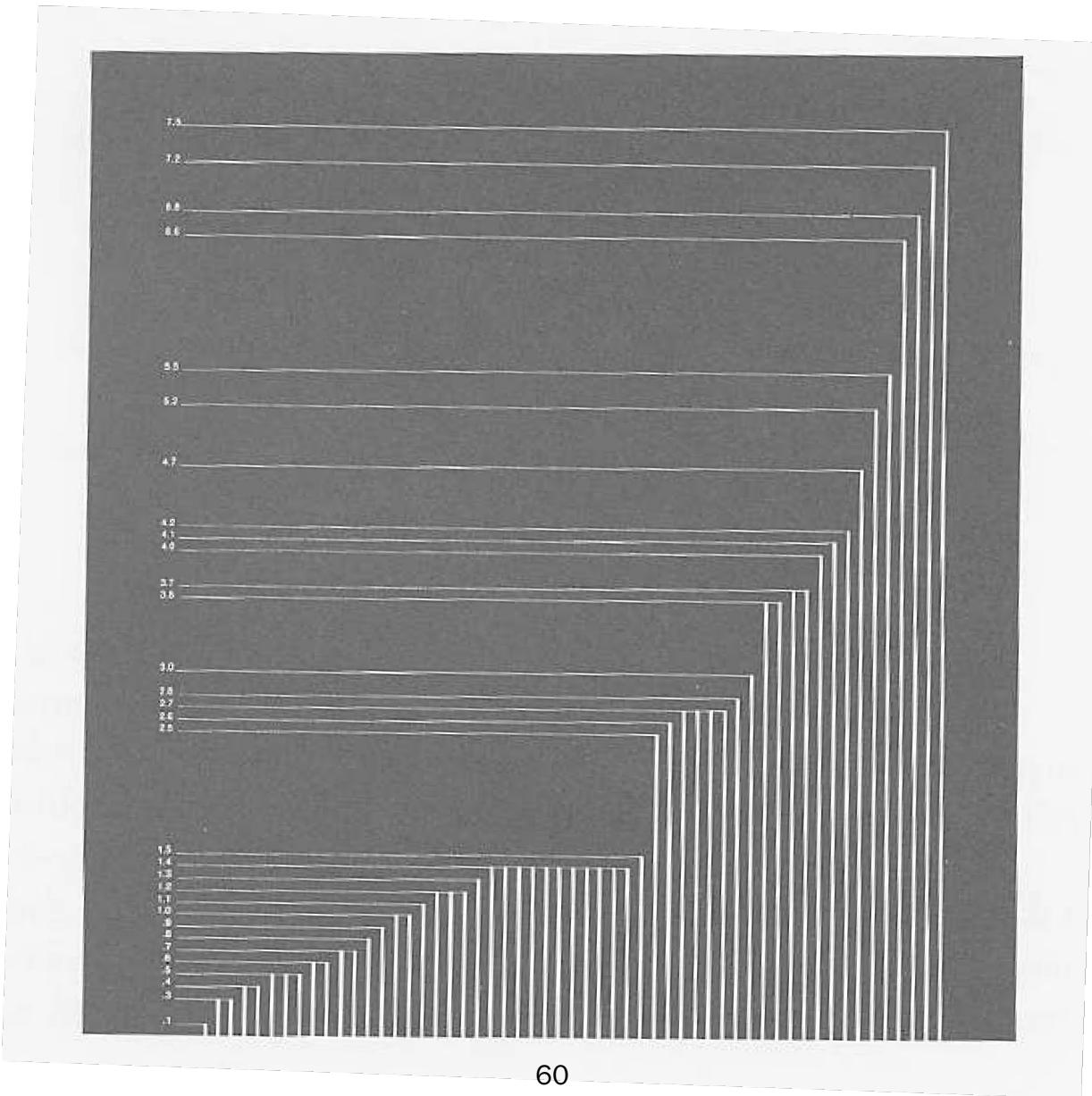
Data-based grid

Sometimes the data grid can be used to report data:

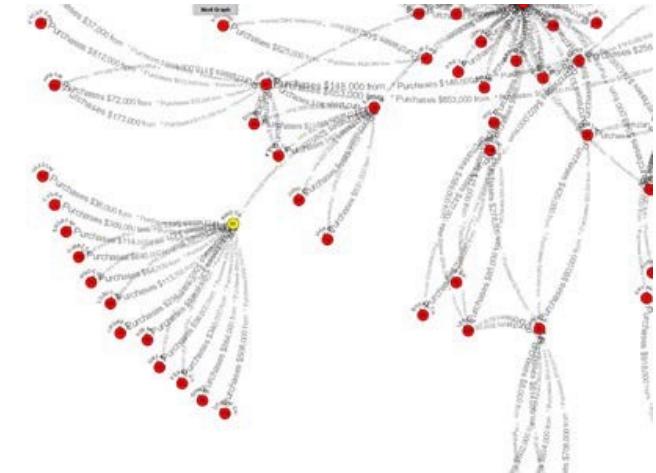
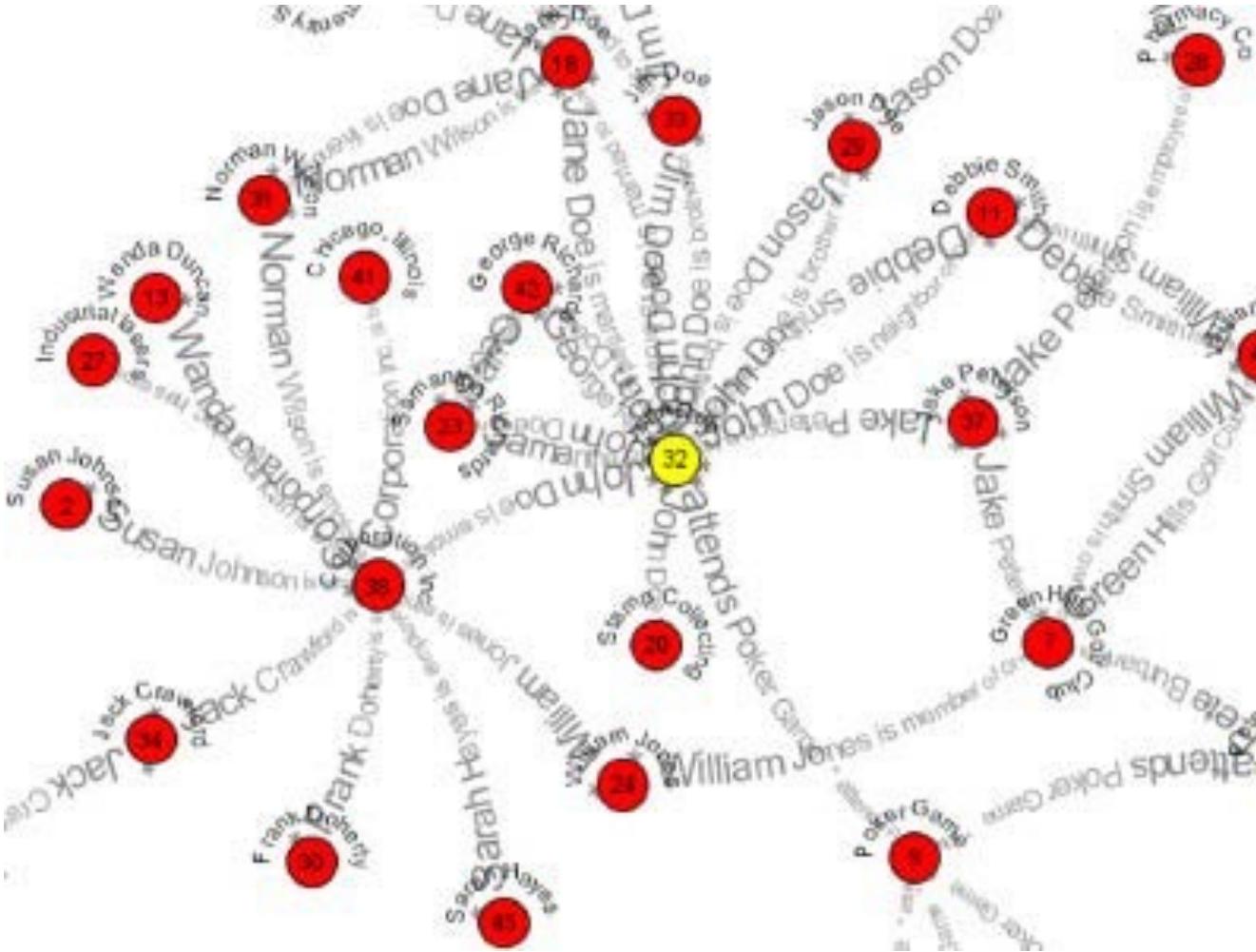


K. V. Roberts, D. E. Potter, 1970 [T 145].

Data-based labels



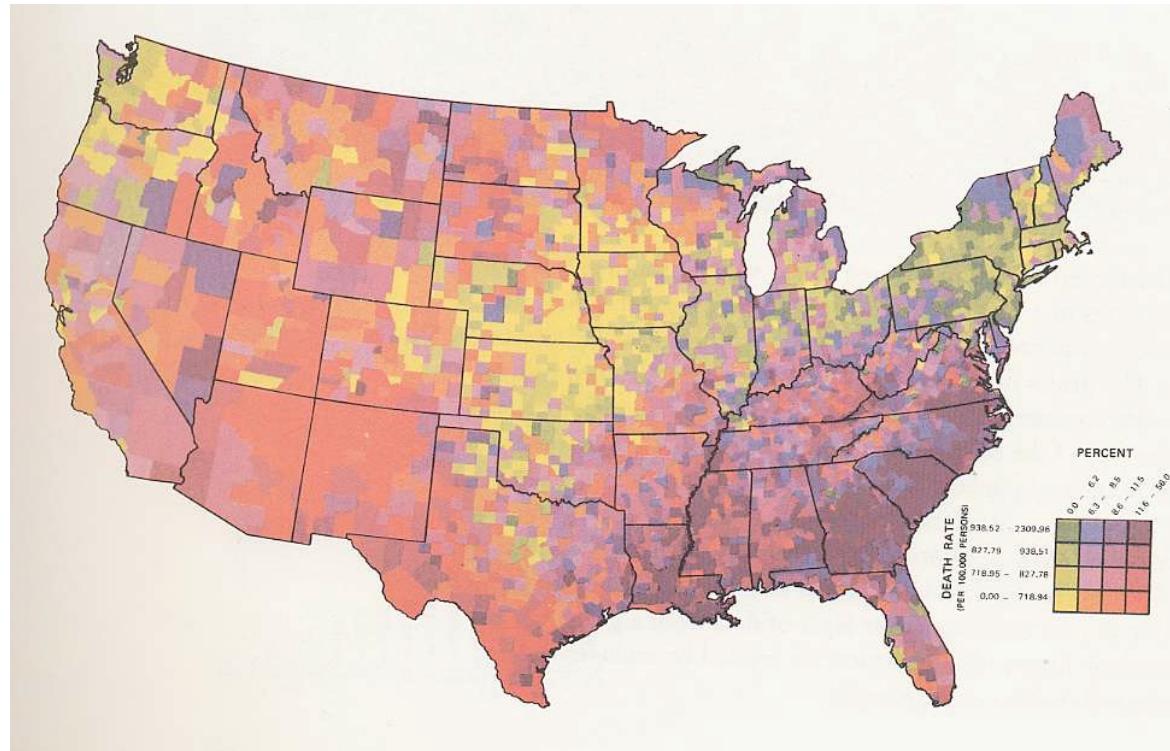
Graphs with extended labels



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

Graphical puzzles

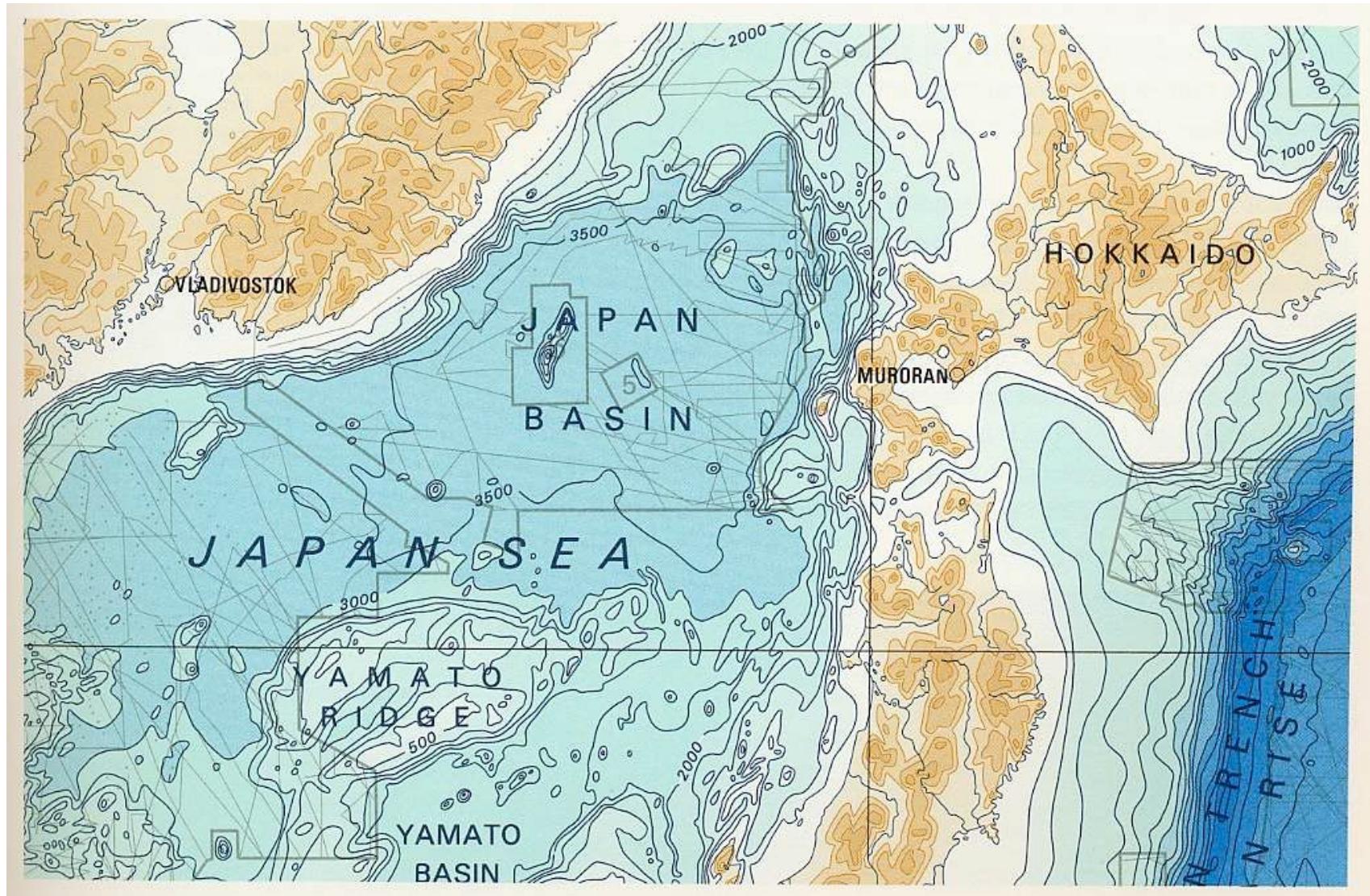
The complexity of multifunctioning elements may turn the data into visual puzzles.



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

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

Visually intuitive use of colors



Theory of data graphics

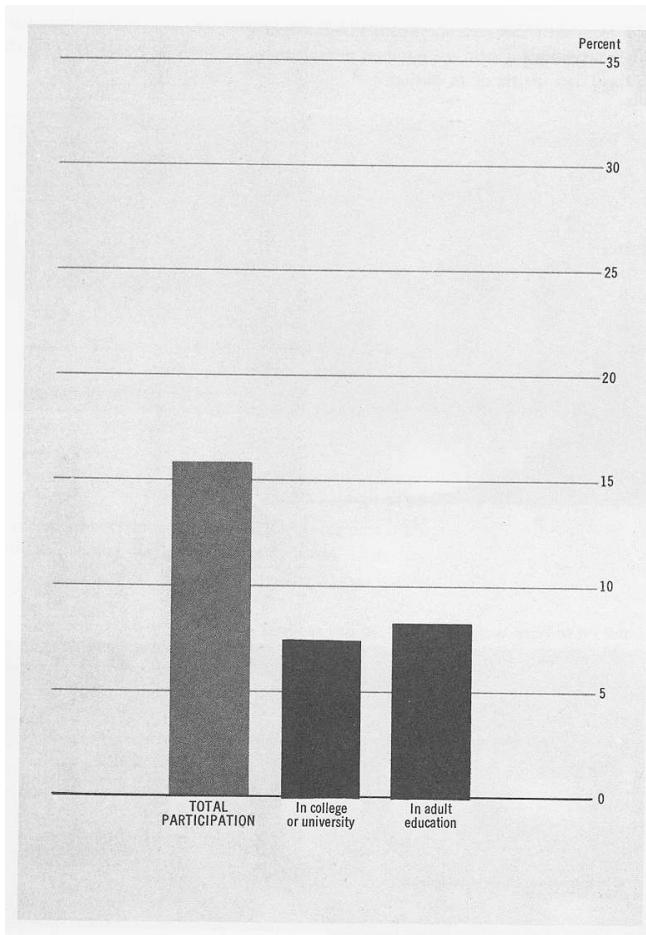
- Data-ink
- Chartjunk
- Multifunctioning graphical elements
- **Data density and small multiples**
- Aesthetics and techniques

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)

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

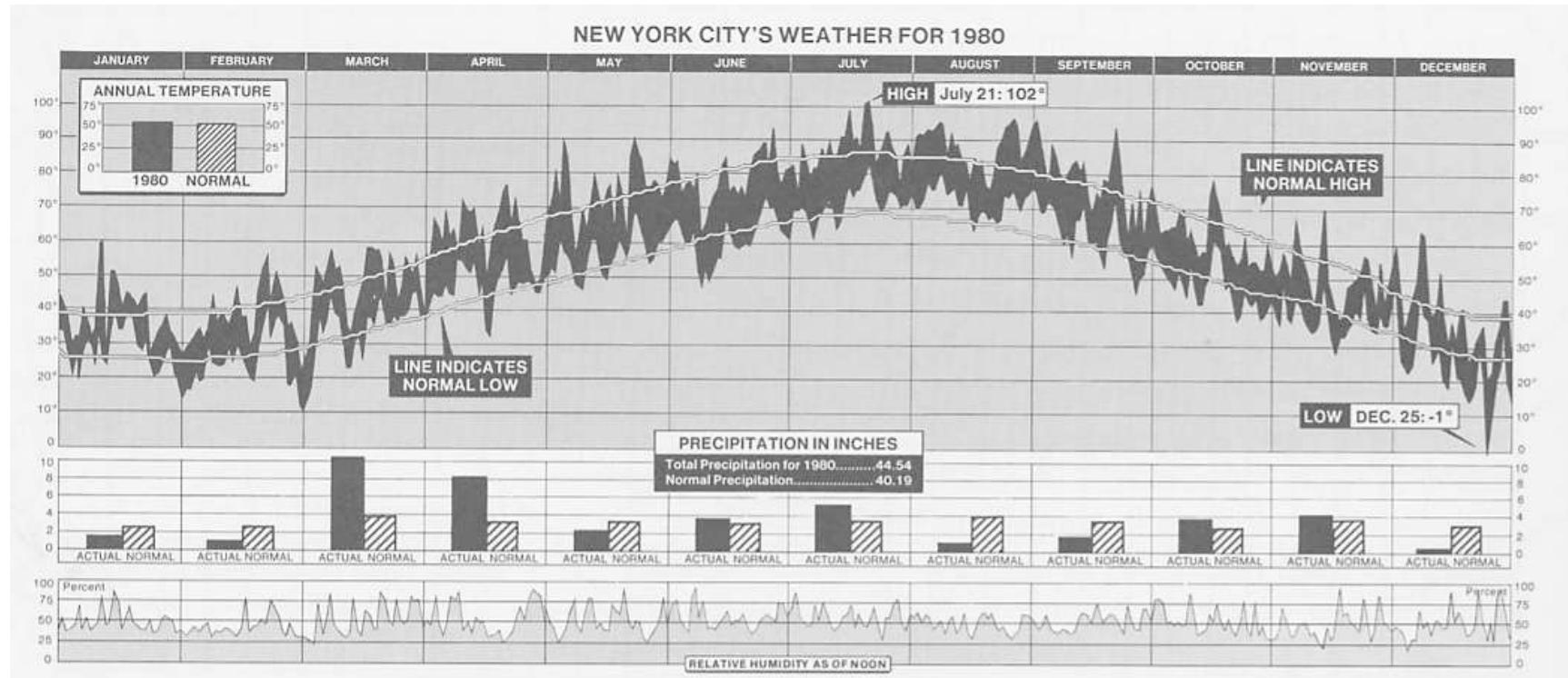
Low data density



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

Data density = 0.02 numbers per cm².

High data density

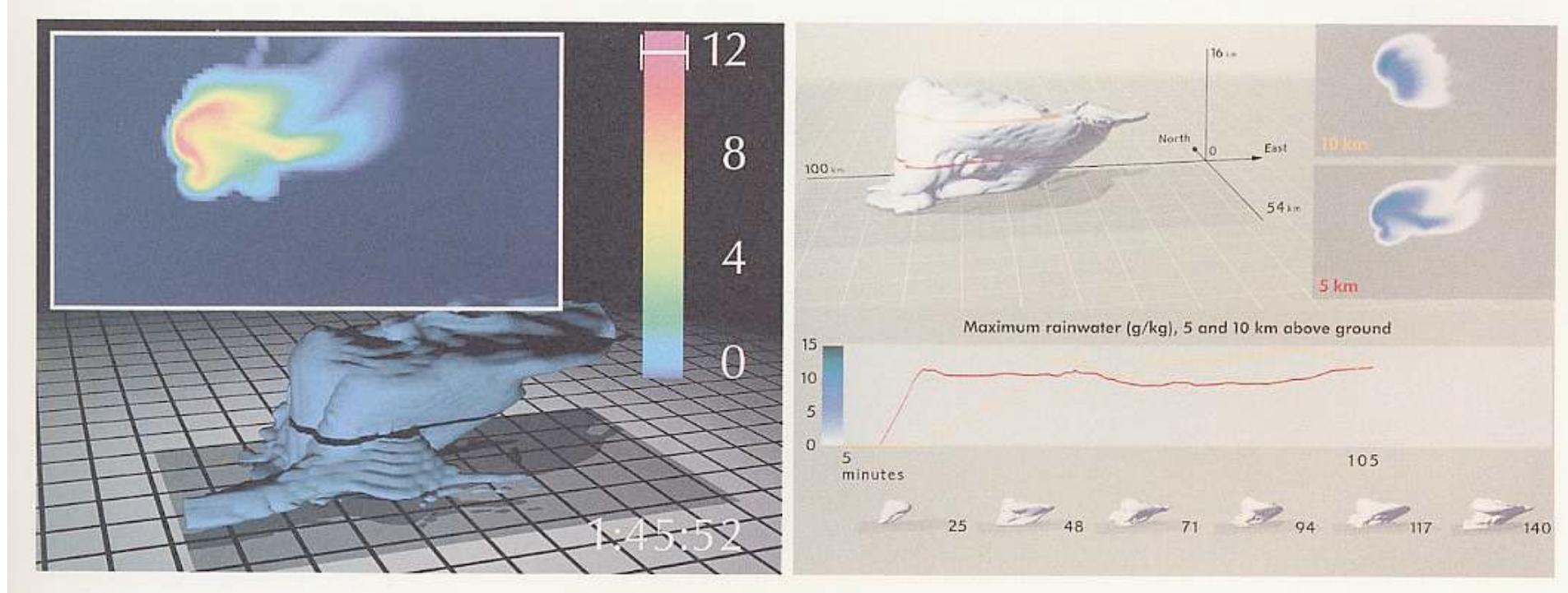


New York Times, 11 January 1981 [T 30].

Data density = 28 numbers per cm².

Use small differences

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

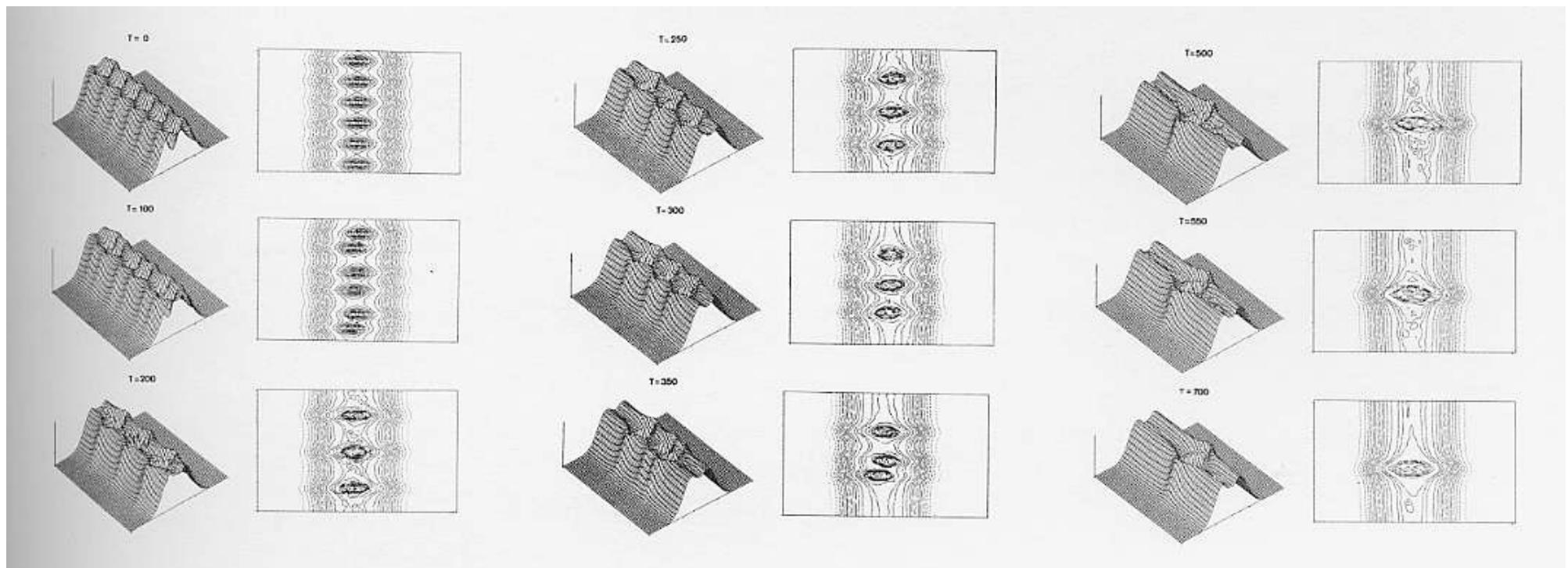


Matthew Arrot et al (original), E. R. Tufte, Polly Baker et al (revised) [VE 75].

Large distinctions generate clutter. Smaller distinctions highlight the data.

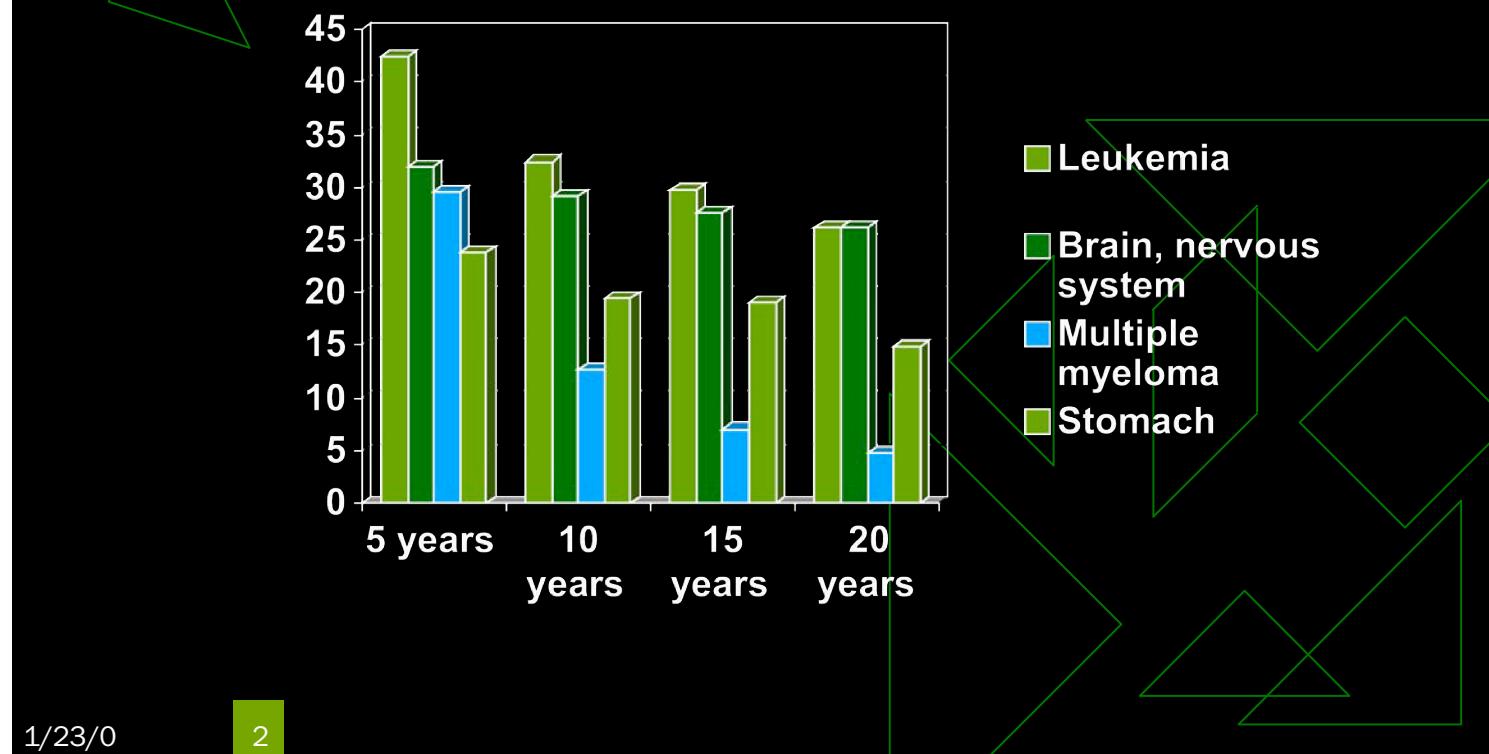
Using small multiples to make comparisons

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



Graphs and tables

I. Cancer Survival Rates



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

Graphs and tables

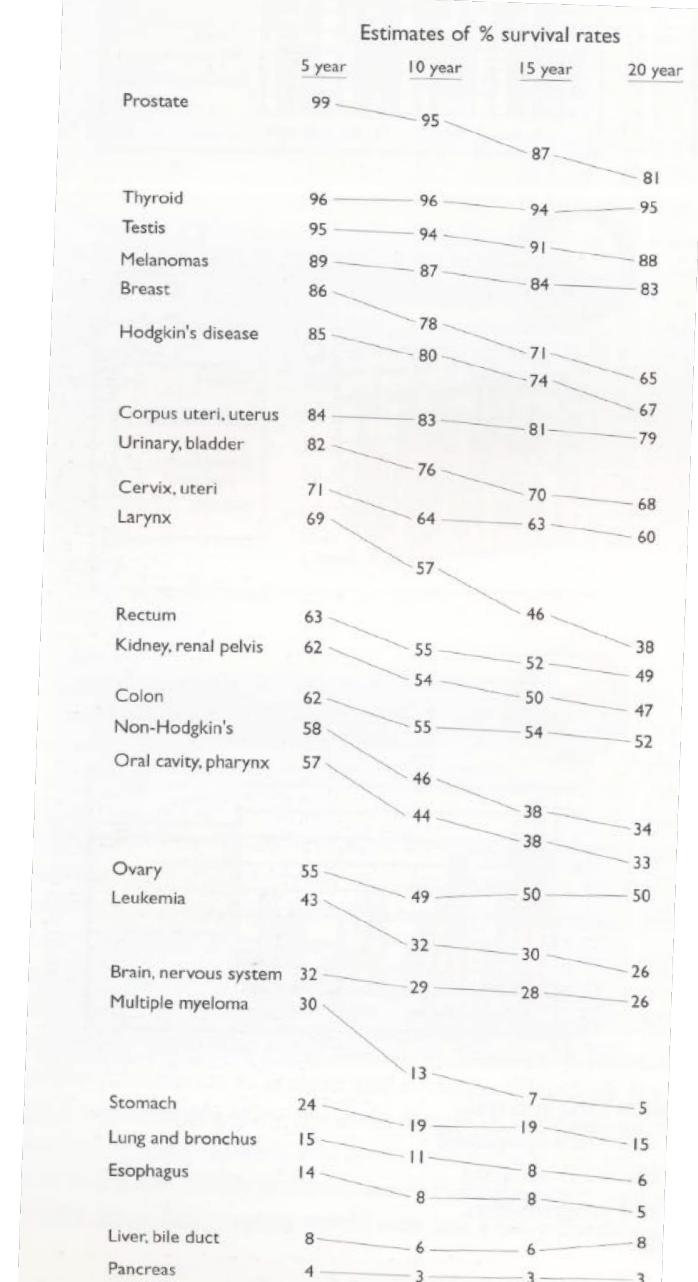
	Cancer survival rates (%)			
	5 year	10 year	15 year	20 year
Leukemia	42.5	32.4	29.7	26.2
Brain, nervous	32.0	29.2	27.6	26.1
Multiple myeloma	29.5	12.7	7.0	4.8
Stomach	23.8	19.4	19.0	14.9

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.
 - Tables are the best choice only for presenting for a small collection of numbers

Some Winners and Losers in the Forecasting Game								
			Chase Econometrics: 7.4%					
Council of Economic Advisors: +4.7%			Wharton Econometric Forecasting: 6.8%					
Data Resources: +4.5%			Conference Board: 6.7%					
Nat. Assoc. of Business Economists: +4.5%			Nat. Assoc. of Business Economists: 6.7%					
Wharton Econometric Forecasting: +4.5%			I.B.M. Economics Department: 6.6%					
Congressional Budget Office: +4.4%			Data Resources: 6.5%					
Conference Board: +4.2%			Congressional Budget Office: 6.3%					
I.B.M. Economics Department: +4.1%			Wharton Econometric Forecasting: +21%					
Real G.N.P. Growth: +3.8%			Council of Economic Advisors: 6.3%					
Industrial Production Growth: +5.6%			Change in Consumer Prices: +7.7%					
Chase Econometrics: +2.8%			Corporate Profits Growth: +13.3%					
Conference Board: +5.5%			Unemployment Rate: 6%					
Data Resources: +5.2%								
Wharton Econometric Forecasting: +4.8%								
Chase Econometrics: +1.9%								
Chase Econometrics: +5.9%								
Council of Economic Advisors: +5.0%								
Wharton Econometric Forecasting: +5.4%								
<i>Forecasters are not listed in categories for which they did not make a prediction.</i>								
<i>*After taxes</i>								

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:

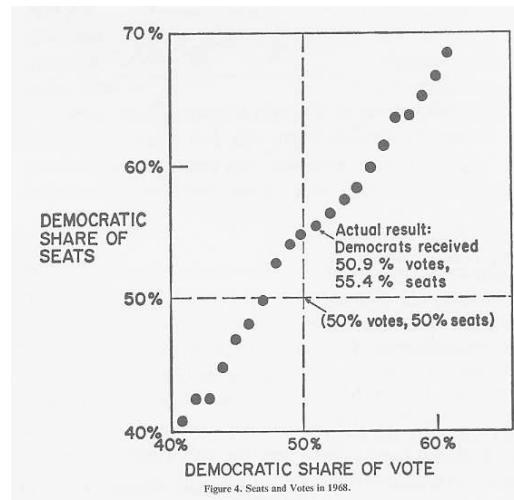
Group B	46 %
Group A	53 %
Group C	57 %

Theory of data graphics

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

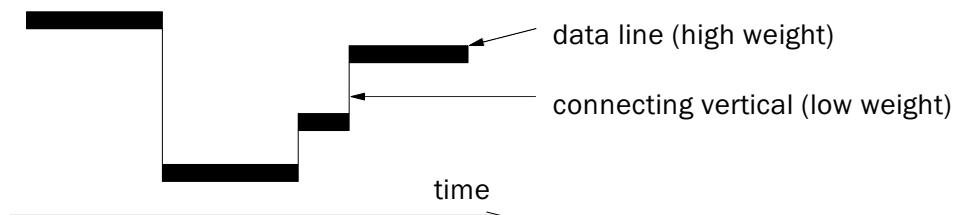
Line weight and lettering

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

A 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.

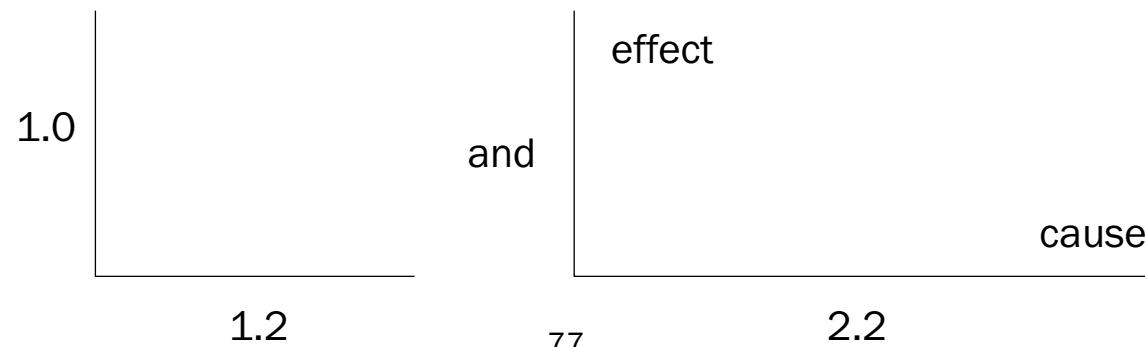
CRIMINAL ACTIVITY OF GOVERNMENT INFORMANTS							
CRIME	CARDINALE	LOFARO	MALONEY	POLISI	SENATORE	FORONJY	CURRO
MURDER	X	X					
ATTEMPTED MURDER		X	X				
HEROIN POSSESSION AND SALE	X	X		X			X
COCAINE POSSESSION AND SALE	X		X	X			X
MARIJUANA POSSESSION AND SALE							X
GAMBLING BUSINESS		X		X		X	
ARMED ROBBERIES	X		X	X	X		X
LOANSHARKING		X		X			
KIDNAPPING			X	X			
EXTORTION			X	X			
ASSAULT	X		X	X			
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			
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

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.

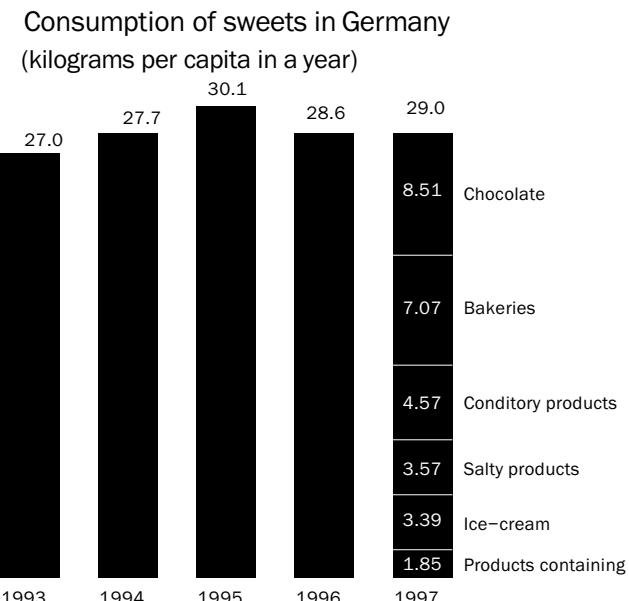


Summary

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

Conclusion

- Communicating ideas and information is difficult. Wlio'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 the damage.
- Theory of data graphics: Show only the essentials in a way that makes facts obvious. Don't waste space. Eliminate non-essentials and redundancies.



Next lecture

- More visualization techniques