Visualization

Bernhard Schmitzer

Uni Göttingen, summer term 2022, April 17, 2023

1 Introduction

1.1 Basic information

Time, BBB, videos.

- Time: Mondays, 10:15-11:45 via the same BBB link. Please share responsibly.
- Lecture will be recorded and made available afterwards in StudIP.
- If you are uncomfortable with your questions being part of the video, please let me know and I will edit them out. But please do not let it stop you from asking and participating.

Exercises and final exam

- One 'problem sheet' every other week, starting next week.
  - Requirement for admission to final exam: 50% of problems solved (or reasonably worked on). General rule: we will be very 'tolerant' here. Not meant as a measure to bar students from taking the exam, more as an 'encouragement' to stay engaged. Submissions will be checked by my doctoral student Olga Minevich.
  - Problem sheets will usually be 'extended examples', not theoretical questionnaires. So it should really be well accessible.
  - Submission online via StudIP. After every sheet there will be a virtual discussion session where the solutions will be discussed. Participation in these is **not mandatory**.
- Final test will be a short practical project, in **groups of two to three**.
  - Should take about 2-3 weeks to prepare.
  - Finally an oral exam in groups, approximately 20mins per student
  - will suggest list of potential topics later during the lecture; suitable suggestions by students are always welcome
- some examples for previous topics:
  - demographics of German parliament (distribution over parties, age, gender, development over time); someone did this for US congress instead
  - trajectories of space probes through the solar system

  - analysis of manga tv shows, how are different genres related, how do viewers decide what to watch and rate shows?
  - extending an open source fitness app for smartphones to provide a better visual presentation of the past workout data
  - analysis of transfers and cash flow between European soccer clubs
  - analysis CO2 emissions or precipitation data, relation to climate change

1.2 What is this lecture about?

What is visualization?

- this lecture is about creating good figures.
  - but not in the sense of: 'how can we render fancy 3d graphics'
  - more in the following sense: good figures can convey large amounts of data or complex structures to the human brain;
  - our brains can process them almost effortlessly, build mental model of data which cannot be achieved by other means (such as reading a description text of the data)
- we want to make the most of this powerful communication tool, need to understand the following things:
  - how do the human eyes and brains process visual information?
  - based on this: what are good design rules, how do we make figures 'compatible' to the human visual system
  - how can we represent data visually (colors, positions, lengths, graphs,...)
  - how can we prepare and transform data for better representation
  - how to keep track of the whole pipeline from data to figure

The role of computers.

- some books on visualization like to emphasize that visualization is not about how to create figures with a computer
  - that a good course on visualization is not a tutorial for a particular piece of software
  - makes sense: software changes, good design principles are independent of software
- but: computers are incredible useful for visualization
  - we can apply much more complicated processing to much more complicated data and transform it into ways that can be visualized
  - we can set up algorithmic pipeline that automatically generates figures for many datasets or after parameter changes
  - we can generate dynamic and interactive visualizations to handle even more complexity
- modern data analysis and visualization impossible without computers, which is why so many universities now offer courses on data science

  - similarly: modern data science and visualization require mathematics and statistics, to transform, analyze and simplify data

- in this spirit: this lecture will not be a tutorial session for a particular software environment
  - but: practical examples will be indispensable
  - for this will mainly use python/matplotlib in jupyter notebooks
  - some other programs and libraries will be mentioned here and there

Python: a prototypical computational data analysis environment.

- open source, available on all platforms, long term support, immense availability of / compatibility with libraries and software
- simple installation, management of components via
  - Anaconda/Miniconda:
    https://docs.conda.io/en/latest/miniconda.html
  - or the Python Package Index and pip
- alternatively, try Google colab (https://colab.research.google.com) or GWDG Jupyter cloud (https://jupyter-cloud.gwdg.de/)
- core packages:
  - numpy/scipy: scientific computing
  - matplotlib: plotting (including high-quality export e.g. for LaTeX manuscripts)
  - imageio: saving/loading raster image formats, animations
  - pandas: managing tables, simple import/export
  - scikit-learn: basic statistical analysis and machine learning tools
  - jupyter (or similar): fast interactive scripting, sharing and presenting results (use jupytext plugin for better compatibility with version control)
- support for data formats (built-in or popular packages):
  - raw binaries, mat, csv, json, xml, ods, xls...

Outline of lecture.

- examples (historical, and 'live' or 'personal')
- brief discussion of 'theory'
  - design principles by Tufte
  - grammar of graphics by Wilkinson
- the human visual system
  - how do the eyes and the brain generate 'the image in our head'?
  - what does this imply for the creation of good figures?

- data types and processing pipeline
- the role of colors
- fundamental visual data representations
  - points, lines, bars, pies, histograms...
  - boxes, violins, errorbars, ...
  - images, heatmaps,
  - contours, vector fields, transformations
  - graphs, meshes, networks
- high-dimensional or complex data
  - aggregations, filters, slices
  - dimensionality reduction and embeddings
- how to lie with charts?
- animations and interactive visualization
- visualization on the web

Literature.

- Alberto Cairo: The Functional Art, New Riders, Berkeley, 2013
- Alberto Cairo: How charts lie, W. W. Norton & Company, 2019
- Andy Kirk: Data Visualisation, SAGE Publications Ltd, 2019
- Robert Spence: Information Visualization, Springer, 2014
- Edward Tufte: The Visual Display of Quantitative Information, Graphics Press Cheshire, 1983
  (and similar, more recent books by same author)
- Leland Wilkinson: The grammar of graphics, Springer, 2005

1.3 A few examples

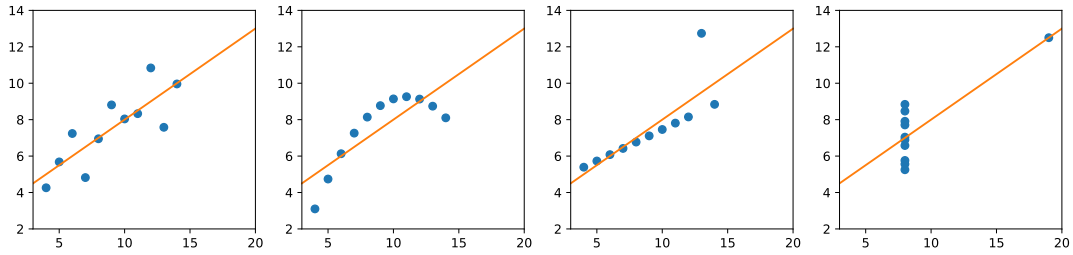
1.3.1 Tables versus plots

- taken from F. J. Anscombe: Graphs in Statistical Analysis, The American Statistician, 1973, 27, 17–21

Data series represented as table:

x_1	10.00	8.00	13.00	9.00	11.00	14.00	6.00	4.00	12.00	7.00	5.00
y_1	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
y_2	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
y_3	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
x_2	8.00	8.00	8.00	8.00	8.00	8.00	8.00	19.00	8.00	8.00	8.00
y_4	6.58	5.76	7.71	8.84	8.47	7.04	5.25	12.50	5.56	7.91	6.89

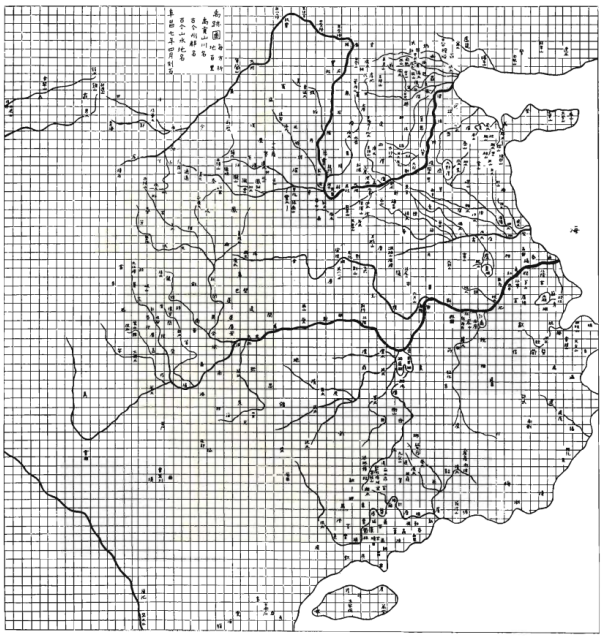
- hard to interpret as numbers in table, try basic statistical analysis
- all x_i and y_i sequences have same mean and variance
- sequences of pairs (x_1, y_1), (x_1, y_2), (x_1, y_3) and (x_2, y_4) all yield essentially the same linear regression:
  - same slope, intercept, correlation coefficient, standard error for slope estimation
- graphic representation immediately tells us four different stories



1.3.2 Historical examples

Chinese cartography

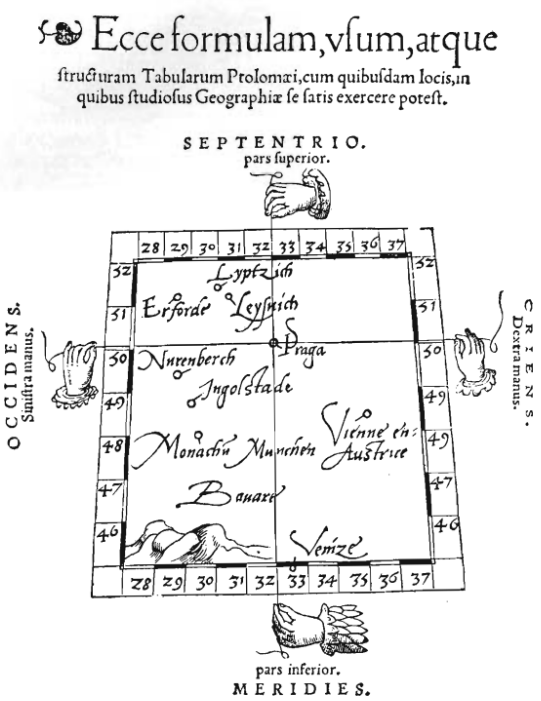
- taken from [Tufte: The visual display of quantitative information]
- approx 1100 AD



European cartography

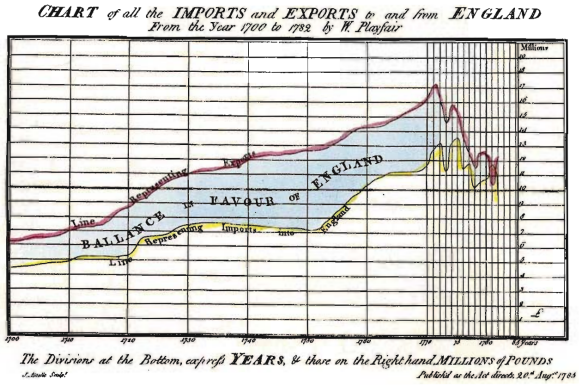
- taken from [Tufte: The visual display of quantitative information]

- 1546 by Petrus Apianus, generalization to two-dimensional plots still took some time



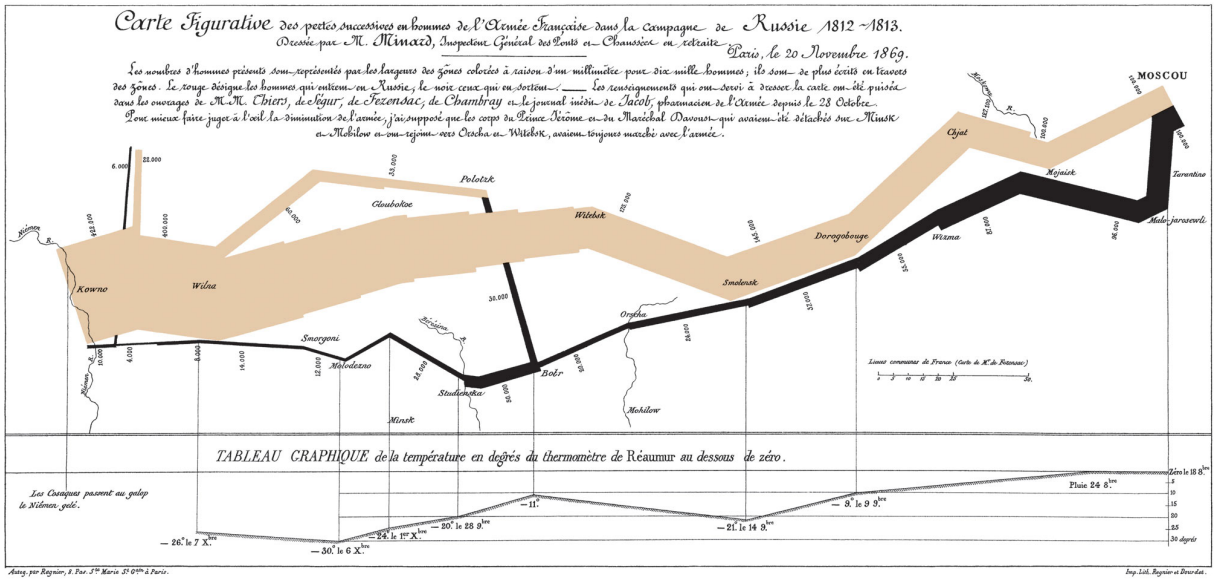
Playfair

- taken from [Tufte: The visual display of quantitative information]
- William Playfair: The commercial and political atlas, 1786



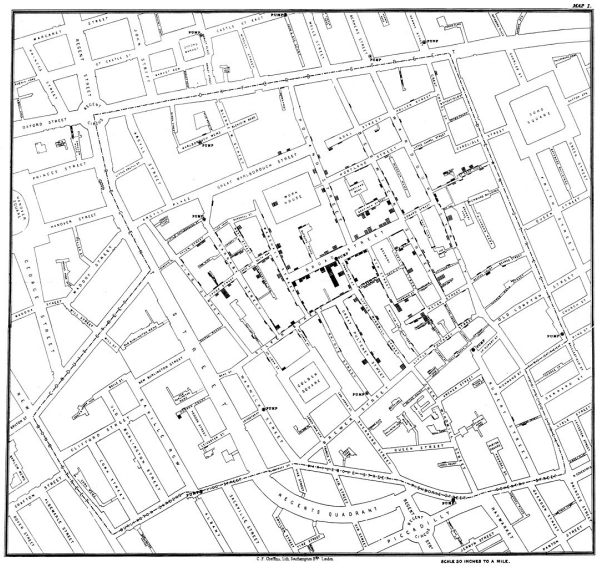
Napoleon's march to Moscow.

- taken from [Spence: Information Visualization]
- figure in public domain, available at https://en.wikipedia.org/wiki/File:Minard.png
- Charles Joseph Minard, 1869



Water pumps in London.

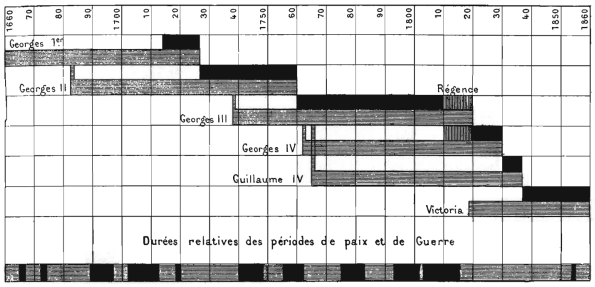
- taken from [Spence: Information Visualization]
- figure in public domain, available at https://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg
- John Snow, 1854



Regency chart

- taken from [Tufte: The visual display of quantitative information]
- E. J. Marey: La méthode graphique, 1885

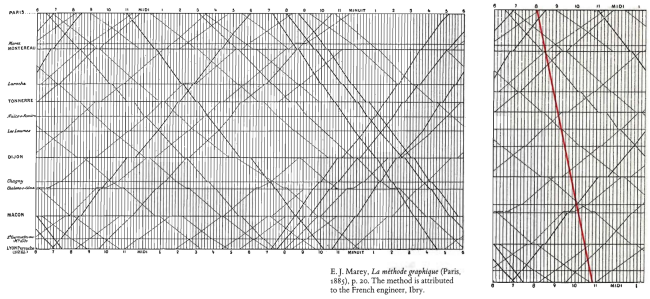
- Note: George II was the founder of Göttingen University, Wilhelmsplatz is named after William IV.



E. J. Marey, *La Méthode Graphique* (Paris, 1885), p. 6.

Train timetable.

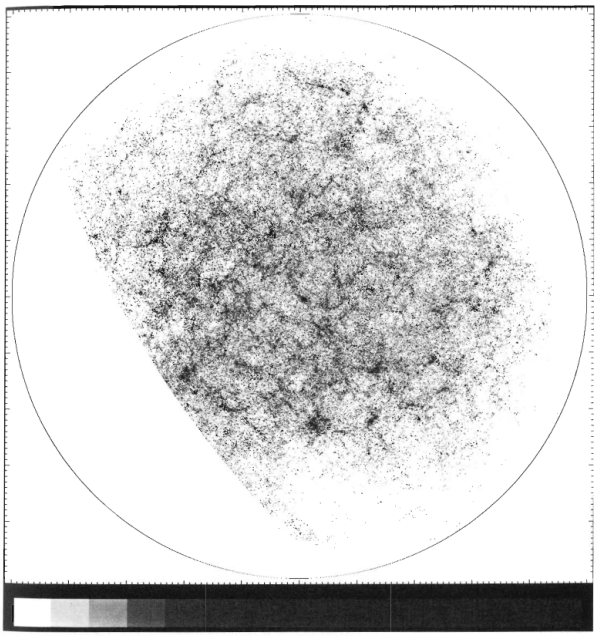
- taken from [Tufte: The visual display of quantitative information]
- E. J. Marey: La méthode graphique, 1885



E. J. Marey, *La méthode graphique* (Paris, 1885), p. 20. The method is attributed to the French engineer, Ibry.

Galaxy distribution.

- taken from [Tufte: The visual display of quantitative information]
- example for computerized cartography, 1977



1.3.3 Example by Alberto Cairo: world population

- taken from [Cairo: The Functional Art, Chapter 1]
- Cairo read book 'The Rational Optimist: How Prosperity Evolves' by Matt Ridley
- chapter on world population made the hypothesis that it will soon stabilize
  - rapid decrease in fertility in developing countries
  - slight increase (back to 'replacement rate' of 2.1 children per woman) in developed countries
- provided figure did not support the hypothesis, did not display appropriate data, different simultaneous trends cannot be distinguished in summarized data
- showing all individual trajectories not helpful either: all necessary data is shown, but hard to process visually
- final version: highlight representatives from different clusters. supports hypothesis, allows for further more detailed exploration (e.g. China, Brazil, Niger)

1.3.4 EEA: Chart dos and don'ts

https://www.eea.europa.eu/data-and-maps/daviz/learn-more/chart-dos-and-donts

- show full y-axis
- consistent x-axis intervals
- Edward Tufte in a nutshell: remove clutter
- highlight what's important
- sorting

- do not use 3d or other visual effects
- direct labeling where possible
- avoid pie charts
- avoid stacked charts
- do not use maps for everything with spatial dimension
- avoid animations, use small multiples

- show level of confidence
- tell the ‘why’ and ‘how’
- how to treat missing data
- do not confuse causation and correlation
- do not compare apples with oranges
- adjust for inflation

- do not forget color deficiency
- ask others for opinion

2 Edward Tufte

2.1 Introduction

About the author.

- Born 1942, professor emeritus for political science, statistics and computer science at Yale University. Pioneer in the field of data visualization. 'e' at the end of name is pronounced. (https://en.wikipedia.org/wiki/Edward_Tufte).
- First influential book on topic: The Visual Display of Quantitative Information, 1983.
- Promoted a philosophy of minimalist design in information graphics, apparently driven by a trend that graphics were only perceived as means to dumb down information or to make statistical data less boring, assuming the audience would be stupid or not interested.
- Adopts a polemic language in his books, seems to enjoy deconstruction of bad examples, likes to formulate lists principles.
- The following section is based on [The Visual Display of Quantitative Information] and examples are taken from there.

Graphical excellence according to Tufte. Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency. Graphical display should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- be closely integrated with the statistical and verbal descriptions of the data set.

Graphical integrity.

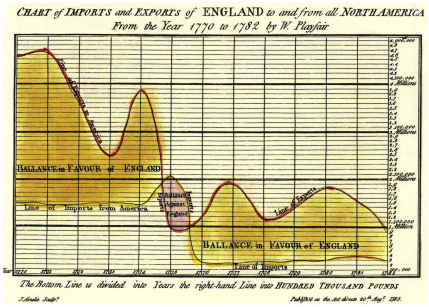
- of course it was already well known that graphics may distort the data, by ignorance or by intention
- Tufte even had the impression that graphics had a general reputation for 'lying to' or 'fooling' viewers, see [The Visual Display of Quantitative Information, Chapter 2]:
  - 'For many people the first word that comes to mind when they think about statistical charts is *lie*.'

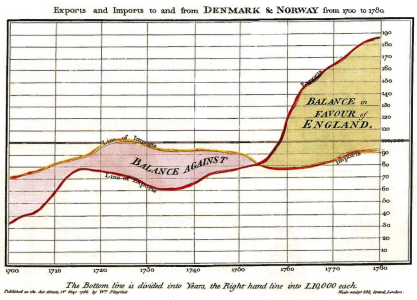
  - 'Much of twentieth-century thinking about statistical graphics has been preoccupied with the question of how some amateurish chart might fool a naive viewer.'

- postpone detailed discussion until session on 'how to lie with charts'.

2.2 Data-ink

Example: Evolution of charts by Playfair.



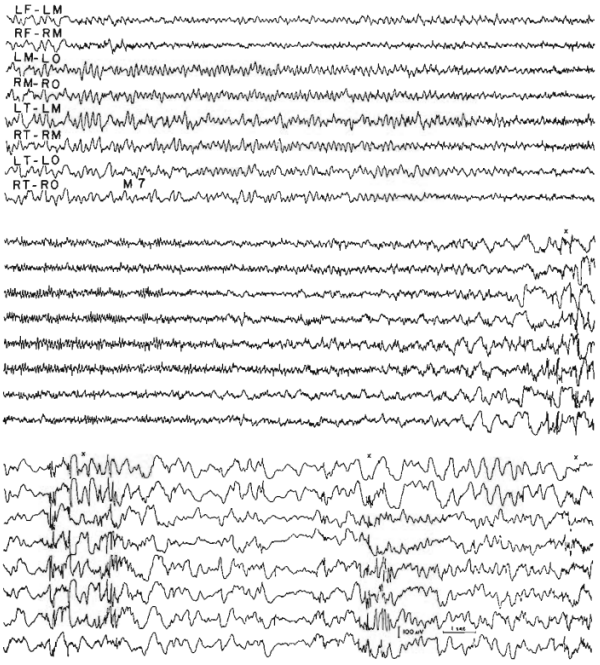


- first example: 1785, early pages of 'The Commercial and Political Atlas'
- second example: created on year later, already much more mature, removed much of the 'background'
- Tufte formulates a fundamental principle: Above all else show the data.

Definition.

- data-ink is the ink in a graphic that represents data / information
  - non-data-ink: frames, grids, (unnecessary) ticks, decoration
  - data-ink: data points, (necessary) labels, derived data (e.g. marginal distributions, indication of minimal or maximal values)
- data-ink ratio $= \dfrac{\text{data-ink}}{\text{total ink}}$

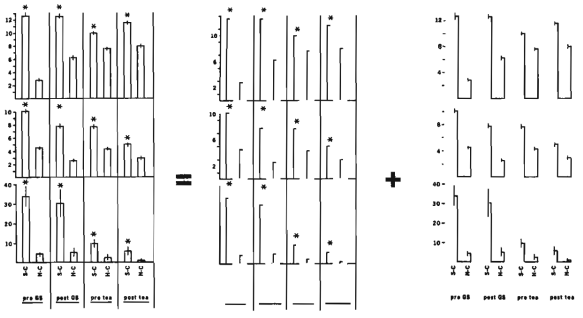
Example: electroencephalogram.



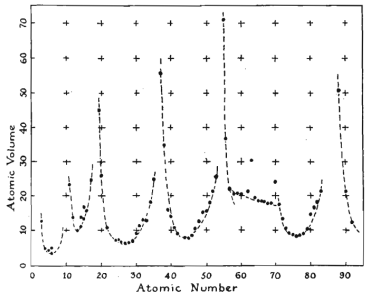
- extreme example: almost exclusively data-ink, but can only be read by specialists

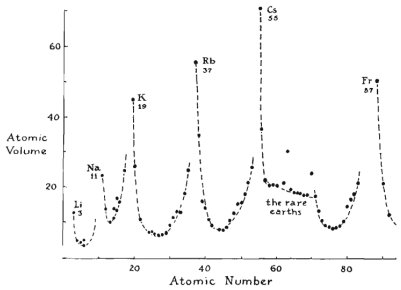
Example: sour taste.

- original image taken from: Kuznicki and McCutcheon: Cross-enhancement of the sour taste on single human taste papillae. Journal of Experimental Psychology: General, 108(1), 68–89, 1979. (study effect of sucrose on the perceived intensity of sour taste)
- Tufte's introduction in book: '[The display] compares each long bar with the adjacent short bar to show the viewer that, under the various experimental conditions, the long bar is longer.'
- Tufte removes: frames, some ticks and labels, one side of each bar, stars (marking the longer bars), text decoration (underline)
- lines connecting adjacent bars are kept (data, since they show which experiments belong together)
- extreme example. My humble opinion: should be seen as illustration of principle rather than as concrete suggestion



Example: periodic system.





- original image created by science illustrator Roger Hayward for chemistry textbook by Linus Pauling, 1947 (intruduced covalent bond in chemistry, two nobel prices, chemistry and peace)
- remove grid, try removing guidelines (but rather not), add individual labels

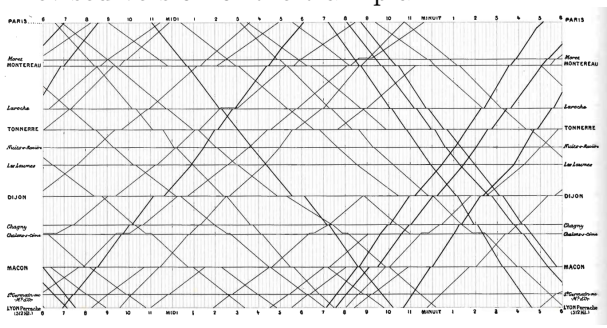
More principles.

- above all else show the data
- maximize the data-ink ratio
- erase non-data-ink
- erase redundant data-ink
- revise and edit
- everything 'within reason':
  - sometimes data-ink ratio is ill-defined or not appropriate
  - sometimes redundancy may be helpful: train schedule example

2.3 Chartjunk: Vibrations, Grids and Ducks

Avoid moiré patterns.

Keep grids subtle. A revised version of the train plan.

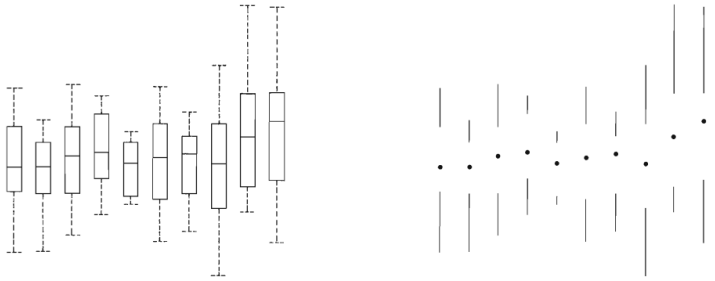


Ducks.

- a duck is a graphic which is just entirely decoration, e.g. self-promotion of graphical techniques instead of information display
- sometimes a table may be better than a pointless graphic

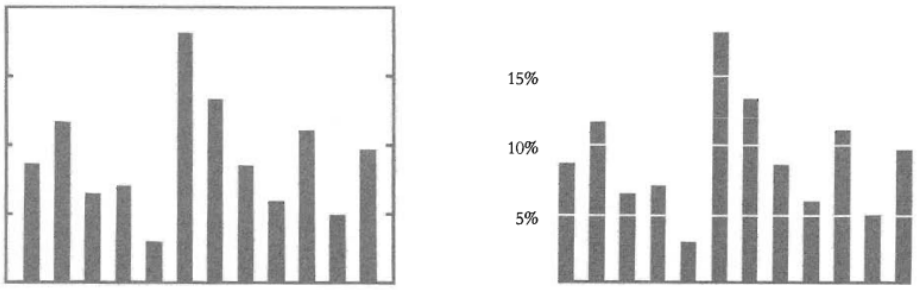
2.4 New graphical design suggestions

Simplification of box plot (quartile plot).



- Tufte: conventional box plot is highly redundant. Suggests minimalistic version, argues via number of placings of straightedge
- my humble opinion: oversimplified
  - (information about) data is there; but weight of ink does not align with weight of data (usually higher density within quartiles, otherwise not proper plotting device anyway)
  - Tufte frequently makes data density calculations: how much numbers are encoded in a figure and equates this with amount of numbers that are transferred into viewers brain
  - but the visual system/brain do not extract a list of numbers from a graphic (at least not at "first glance"), but coarse structures and trends
  - coarse structure more accurately visually reflected by original design

Simplification of bar chart.



- Tufte proposes changes to basic bar chart
- my humble opinion: misses the point. simple bar chart is not the right format for the discussed example in the first place.

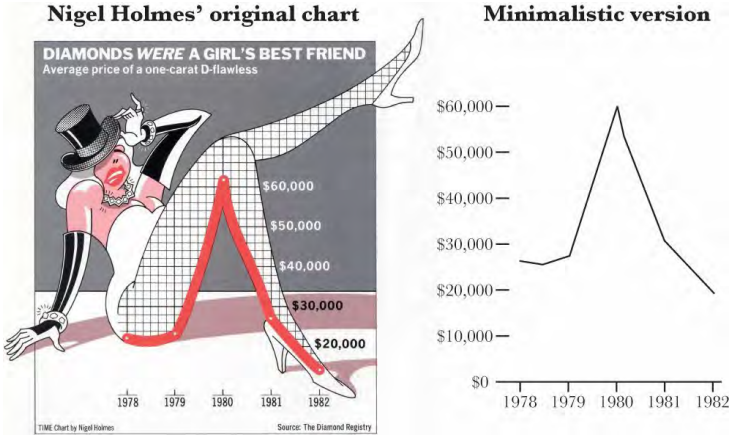
- what is the type of the x-axis?
  - a **time series** or other continuous, one-dimensional variable? then show scatter plot, possibly with lines
  - a **nominal type**? (city, country, ...) then data should be sorted, or at least grouped (e.g. by continent?)
  - does the chart represent a **histogram** with contiguous intervals (of equal width)? then remove gaps between bars
  - if gaps correspond to empty intervals: chart is perfect. it emphasizes a very peculiar property of data.

2.5 Critical discussion by Alberto Cairo

Dumbing down.

- apparently a widespread misconception: graphics are for 'dumbing down' data, flashy presentation of 'boring statistics'
- when struggling with interpretability of a graphic, reflex is to simplify instead of clarifying
- this appears to have been particularly common in 80s (and onward with rise of computers) and seemingly motivated Tufte's work
- Tufte: we should not think that our readers are stupid.
- Of course this is true. But also keep in mind: readers/listeners need time to absorb, process, pause and digest new information. Not all required background-knowledge may be present (or has become a little diffuse). We get tired. Human brains are not computers, eyes are not cameras. Motivation, patience, redundancy and good graphics are key to including the audience.

Edward Tufte vs Nigel Holmes.



- Nigel Holmes was art director for Time magazine
- example: illustration of diamond prices, 1980s

- ‘anti-Tufte’: very low data-ink ratio, data-density, full of decoration and ‘chart junk’, (and blatantly sexist)

However:

- Tufte’s principles are not rigorously based on scientific research but also on aesthetic preferences
- there is no empirical evidence that data-ink ratio is indeed a good measure for the quality of a graphic (in terms of readability)
- mixed results in studies:
  - Ben-Gurion University, 2007, 87 students: compare bar charts with minimalistic versions.
    No significant difference in interpretation performance;
    students aesthetically preferred ‘classical’ charts.
  - University of Saskatchewan (Canada), 2010, 20 students: compare four Nigel Holmes illustrations with minimalistic versions.
    Subjects interpreted both versions equally well.
    After a waiting period, subjects could answer questions about Holme’s graphics with higher accuracy (were not told that they would be questioned)
  - these are not conclusive, representative studies (e.g. very small sample groups) but tempting naive conclusion: decoration may help the brain remember a graphic (and thus also its data)

2.6 Data density and small multiples

- Tufte: most graphics can be reduced in size without losing readability
- small multiple: use same ‘graphical encoding scheme’ multiple times in a row, to display sequence/series of data
- one the reader has encoded the first graphic, he immediately can read all the others
- in many contexts preferrable to animations (in print, even in presentations: viewer scan ‘scroll’ individually, can compare simultaneously)