CS-E4840
Information visualization D

Lecture 1: Introduction, practicalities, history

Feb 27, 2023

Petter Holme, Aalto University

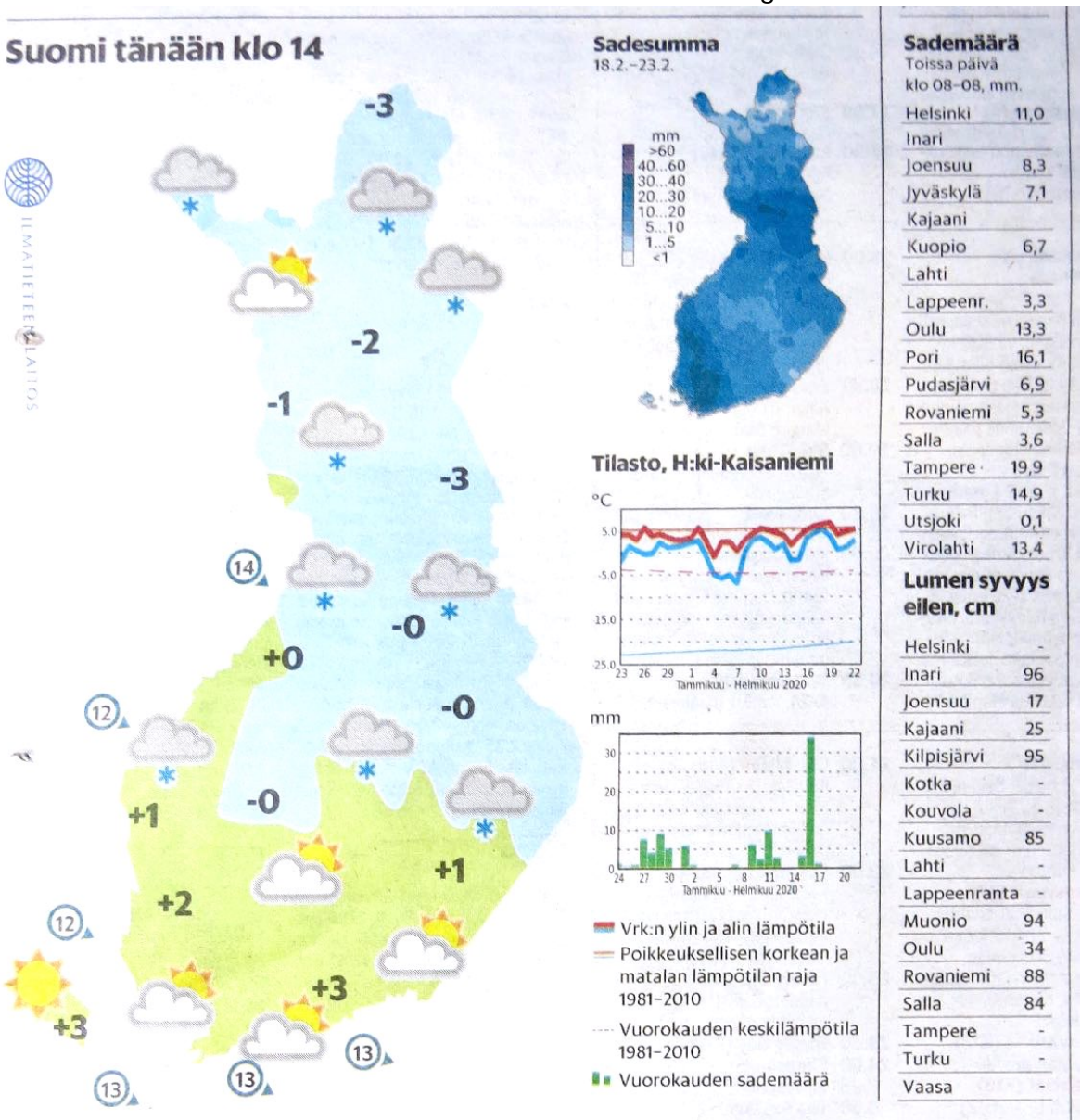
Information visualization is a language

- This course may be your only formal training.
- We go over it from the phonics, to vocabulary, to school compositions, and try to give a feeling for the poetry.
- To know what to say it is crucial to think about how others will understand what you say.

Data visualization gives…

- Quick overview
- Summaries and trends
- Details in context (e.g. on the map)
- Relations between items
- etc.

Helsingin Sanomat 24 Feb 2020



Why visualization matters

- In data analysis we try to understand and find interesting patterns from complex data
- Visualizing data properly is important for
    - understanding the data
    - properly communicating discovered results
- The course goal is to teach how to
    - design good plots
    - recognize bad and manipulative plots (and they are everywhere!)
- Necessary for research and data science
- But also, a general methods course.

Anscombe's quartet

1		2		3		4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All of the four Anscombe's datasets 1-4 have the same linear statistics.

$$y \approx \hat{y} = x/2 + 3$$

$$\mathrm{cor}(x, y) = 0.816$$

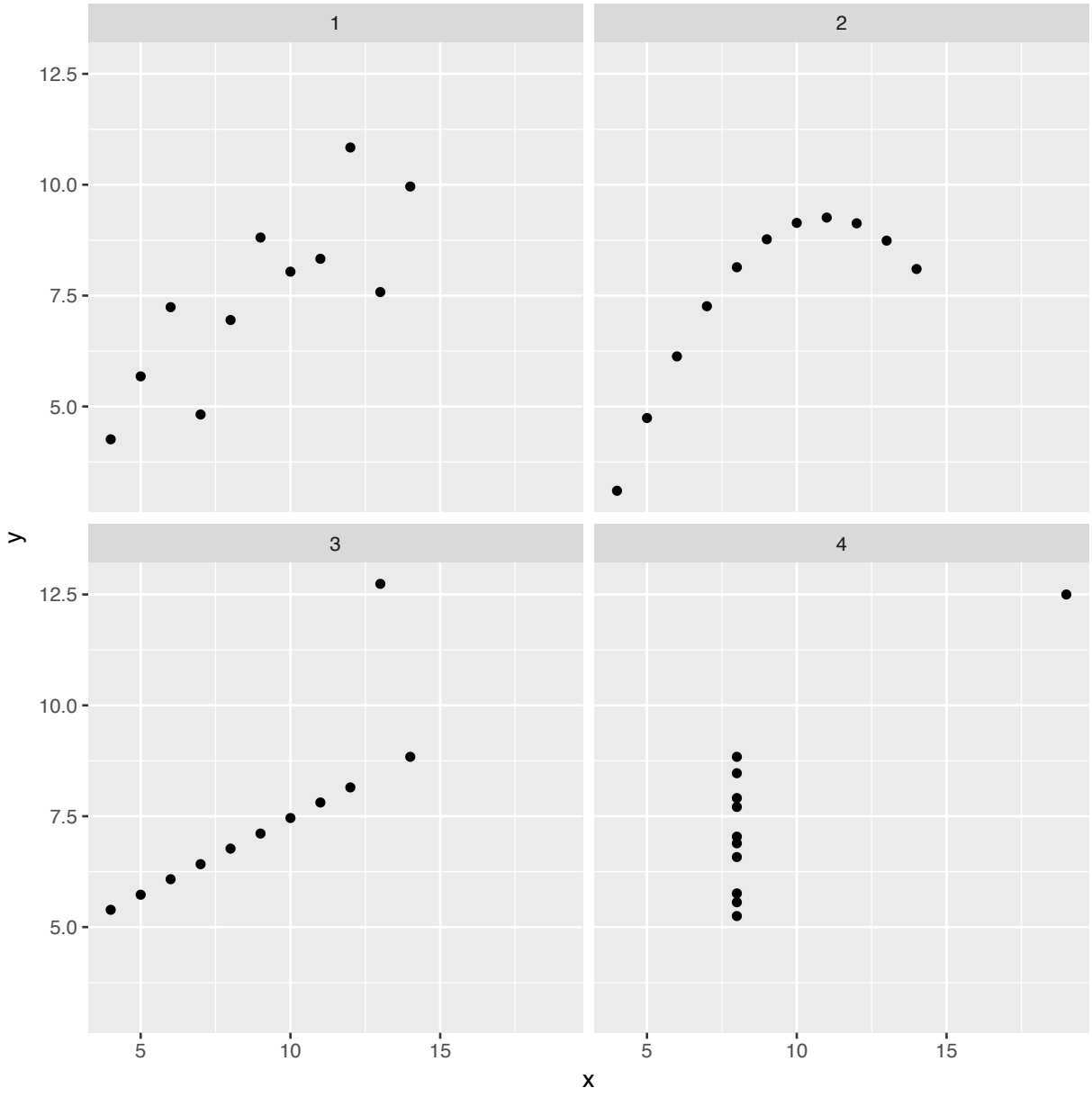
$$\overline{x} = 9 \qquad \overline{y} = 7.5$$

$$\sigma_x^2 = 11 \qquad \sigma_y^2 = 4.127$$

What is the difference?

Anscombe's quartet

1		2		3		4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Course topics

1. General principles, with basics of data graphics and definitions of good information visualisations

2. Basics of human visual system, with respect to the problem of designing and developing good visualizations

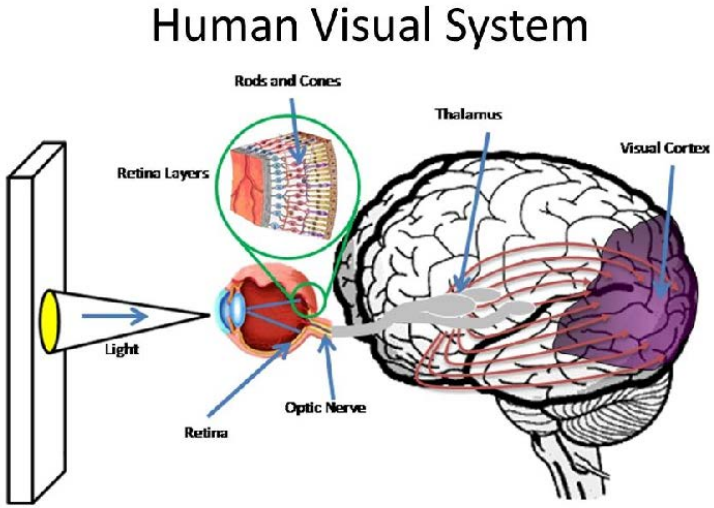
3. Visualisation methods for complex data and interaction.

Course topics

	A	B	C	D	E	F	G
1	0,76144	0,99926	0,59353	0,26766	0,58828	0,08342	0,71821
2	0,70548	0,87626	0,93543	0,6459	0,85224	0,91625	0,17509
3	0,843	0,6885	0,63091	0,18007	0,58733	0,80476	0,16237
4	0,08748	0,34491	0,94111	0,84336	0,79541	0,80996	0,07987
5	0,85647	0,88321	0,30905	0,79756	0,79529	0,35804	0,02648
6	0,22338	0,64558	0,13572	0,53257	0,95441	0,64331	0,5895
7	0,11769	0,83481	0,43029	0,35643	0,31803	0,67361	0,3808
8	0,58369	0,85472	0,21644	0,13686	0,99648	0,35249	0,85745
9	0,01584	0,3643	0,87598	0,69975	0,73019	0,68512	0,68624
10	0,3066	0,13121	0,30138	0,28631	0,81899	0,28214	0,7823
11	0,5898	0,61903	0,68734	0,69408	0,23265	0,42369	0,44631
12	0,87785	0,71118	0,26225	0,75308	0,45452	0,66544	0,71188
13	0,80458	0,60053	0,63635	0,97262	0,05896	0,76963	0,6336
14	0,78174	0,49842	0,28218	0,97796	0,16879	0,4536	0,6072
15	0,97982	0,39325	0,43348	0,10431	0,29396	0,82928	0,86148
16	0,42392	0,17357	0,30216	0,16662	0,72002	0,07476	0,33337
17	0,1985	0,43727	0,78689	0,04252	0,3221	0,40792	0,94561
18	0,19861	0,44761	0,3822	0,09014	0,9653	0,49958	0,24562
19	0,01229	0,05561	0,40269	0,08393	0,27243	0,28443	0,67197

Part 3. how to select the right data?

Part 1. how to design a presentation?



Part 2. how we see it?

Part 0: Practicalities

what to do and how to pass the course

Basic course info

- Information Visualization D, 5 cr., CS-E4840
- Home page https://mycourses.aalto.fi/course/view.php?id=36686
- MyCourses “Announcement” is the main channel for information from the teachers.
- Use MyCourses “General/Forum” for questions and discussions (rather than email as much as possible). It’s monitored by the teachers. Feel free to answer fellow students there.
- All course-related email should be sent to cs-e4840@aalto.fi

Basic course info

- Spring term (period IV) 2023.
- 12 lectures (from February 27 to April 13).
- 3 assignments at your own time (no exercise sessions)
- Exam April 20: 13:30–16:30. Retake exam in September (and possibly in December if there is enough interest).
- Check Sisu for updates about the exam. Don’t forget to register.

Staff

- Lecturer: Petter Holme, http://petterhol.me
- Teaching assistants:
  - Beverley Goh
  - Chenyu Li
  - Duy Vu
  - Yajing Wang

Language

The course language is English. All communication will be in English with the only exception that you are allowed (but discouraged) to answer the exam in Finnish or Swedish.

Participation

To participate in the course, you need:

- To be registered as a student at Aalto University.
- You need a student number and a valid registration in Sisu.
- In exceptional cases, we may send you email to the address registered in Sisu.

Course structure and goals

- Information visualization is a diverse topic
- The goal of this course is to discuss/understand/see information visualization from different points of view
- Thematically, course is divided in 3 parts:
  - Basic elements and core guidelines of visualization (four lectures)
  - Human perception and its relation to visualization (three lectures)
  - Algorithms for visualizing complex (such as high-dimensional) data, and interactive visualization (four lectures)

Prerequisites

- BSc degree or equivalent knowledge (math etc.)
  - e.g., optimization techniques such as gradient descent, as well as linear algebra (Part 3)
- Basic programming skills for plotting
  - Get familiar with R, Matlab, Python, or some similar software (Excel is not flexible enough)
- Acquiring these prerequisites is up to you

Grading

- The course grade is determined by the exam and 3 assignments:

- Exam: 6-7 questions of various kinds (to be solved with pen and paper), 30 points in total.

- Assignments: 3 of them with 14-15 points each, 44 points in total.

Grading

The total points are computed with the following formula,

total points = examination points + assignment points.

The maximum number of total points is 30 + 44 = 74.

To pass the course, you need to both:

- get at least 15 points from the examination and
- get at least 74 / 2 = 37 total points.

The grade is defined on a scale from 1 to 5 by the total points with the following grade limits (lowest total points that would give a specific grade): 1: 37, 2: 44, 3: 51, 4: 59, and 5: 66.

The results of the examination and assignments will be valid until the end of 2023.

Exam

You need to pass one exam

The date for the first exam is April 20

The second (retake) exam will take place after the summer

You only need to pass one of these exams.

You may take more than one exam, in which case your highest score will give your grade.

Assignments

- Three of them with deadlines 23:59 (March 15, 29, and April 12).
- Submit your answers as a PDF file via MyCourses.
- Late submissions is penalized starting with –0.5 points and increasing –0.5 points every 4 hours passed the deadline.
- Scores and feedback posted at MyCourses.
- You should do the assignments by yourself. Plagiarism and the use of AI writing aids (like Chat-GPT) is not allowed.

Assignments

- Typically, most of the points come from assignments.

- So, take the assignments seriously, and start early (some of them take time). (Last year Assignment 3 took longer than Assignment 2 which puzzled students.)

- The results of the assignment are only valid 2023 as next year the grading may change.

Course material

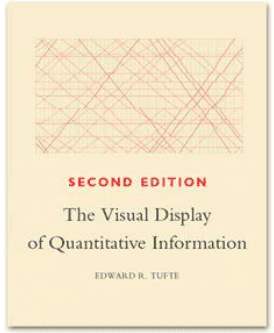
- Lecture slides
  - Available online.
  - Based on previous years' slides by Kai Puolamäki, Antti Ukkonen, Pekka Marttinen, Francesco Corona, Amaury Lendasse, Luana Micallef, Tassu Takala, and Nikolaj Tatti.
  - The **purpose of lecture slides is to support presentation**, not to act as self-contained reading material (use books instead or attend the lectures!)
- Lecture recordings
  - I will share as much as I can.
  - Some lectures will unfortunately only be online (zoom) or prerecorded.
  - A complete sent of recordings from the 2021 edition is online for your reference.
- Books
- Software

Lectures

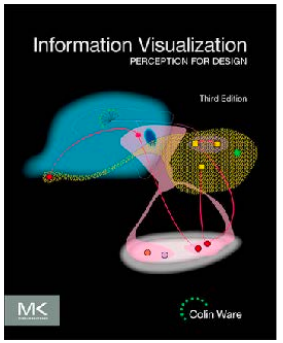
1. Introduction, practicalities, history
2. Introduction and graphical excellence
3. Graphical practice (prerecorded)
4. Theory of data graphics (prerecorded)
5. Visualization techniques (prerecorded)
6. Human perception (zoom)
7. Human vision
8. Visual attention
9. Gestalt laws and more about perception
10. Dimensionality reduction and statistics aiding explorative visualization
11. Network visualization & Dr. Holme’s visualization clinic
12. Recap and preparation of the exam

Main course books

- Edward R. Tufte, **The visual display of quantitative information**, 2nd Ed., Graphics Press (2001). A classic work on good graphical practice



- Colin Ware, **Information Visualization: Perception for Design**, 3rd Ed., Morgan Kaufmann (2012). A good summary on human perception in relation to visualization [ebook available]



- Stephen Few, **Now You See It: Simple Visualization Techniques for Quantitative Analysis**, Analytics Press (2009)

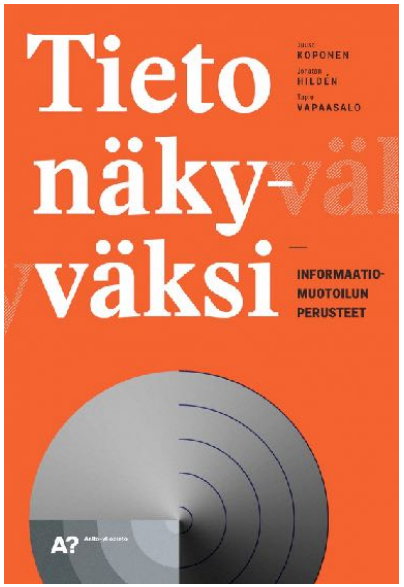
Other useful books

- Alan Izenman. Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer (2008).
- Tamara Munzner, Visualization Analysis and Design, CRC Press (2014).
- Jacques Bertin, Semiology of graphics. The University of Wisconsin Press (1983).
- William S. Cleveland, The Elements of Graphing Data, Revised Edition, Hobart Press (1994).
- John A. Lee and Michel Verleysen, Nonlinear dimensionality reduction, Springer (2007).
- Manuel Lima, Visual complexity: Mapping patterns of information, Princeton Architectural Press (2011).
- Other books by Tufte: Envisioning information, Visual explanations, and Beautiful evidence .

Also in Finnish

- Juuso Koponen, Jonatan Hildén, Tapio Vapaasalo, **Tieto näkyväksi: informaatiomuotoilun perusteet**, Aalto Arts Books (2016).

- Good coverage of graphical design, practical techniques and appications.



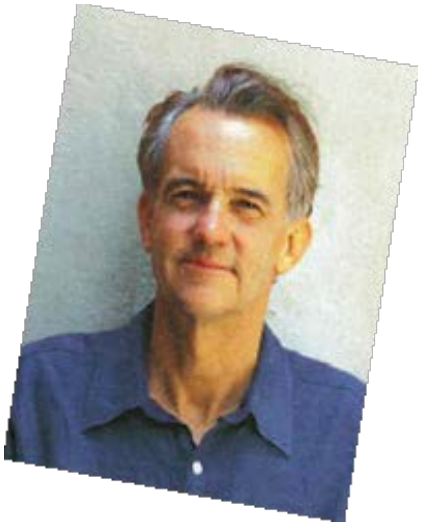
Software

- Useful software for your exercises/assignment (not all is needed and there is no "official" course software, but you will need some programming skills)
  - **R** (www.r-project.org) - recommended
  - **R Shiny** (shiny.rstudio.com) - recommended for interactive data visualisation
  - **RStudio** (www.rstudio.com) - R and R Shiny in a nice package, with a website with links to tutorials etc.
  - Matlab (installed in Aalto computers) or Octave
  - Python and matplotlib
  - Prefuse and Flare: prefuse.org and flare.prefuse.org
  - Processing: processing.org
  - D3.js: d3js.org
  - Open office or MS Excel
  - ...or any other software you are comfortable with

Part 1:
Graphical excellence

what is a well-designed presentation

Edward Tufte



- Edward Tufte (1942-), American statistician and Yale University emeritus professor of political science, computer science, and statistics. www.edwardtufte.com

- **The visual display of quantitive information** (1983 and 2009) is a classic on data graphics, charts and tables

  - *A landmark book, a wonderful book*. Frederick Mosteller, Harvard

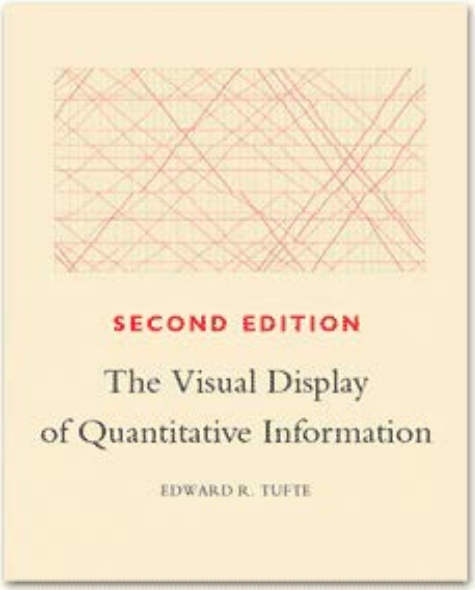
  - *A tour de force*. John Tukey, Bell Labs & Princeton

  - *The century's best book on statistical graphics.* Computing Reviews

  - *One of the best books you will ever see*.Datamation

  - *Best 100 non-fiction books of the 20th century.* Amazon.com

  - *Reading it is a must to understand how you are being lied to by politicians.* djinni111@thepiratebay



Graphical excellence

- Graphical excellence is all about the **well-designed** presentation of **interesting** data
    - you need to have good data
    - your (statistical) analysis needs to be solid
    - the plot needs to be well-designed
    - complex ideas communicated with clarity, precision, and efficiency
- Graphical excellence gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space
    - nearly always multivariate, complex data
    - tells the truth about the data

Goals of visualization

- Presentation (4,000 years)
    - starting point: facts to be presented
    - goal: visualization which makes the facts apparent
    - “You do not really understand something unless you can explain it to your grandmother.” (Albert Einstein ?)
- Confirmative analysis (200 years)
    - starting point: hypothesis about the data
    - goal: confirmation or rejection of the hypothesis
- Explorative analysis (≈ 20 years)
    - starting point: no hypothesis about the data
    - goal: hypothesis about the data

Information visualization

- Data graphics display measured quantities using and combining
    - points and lines, a coordinate system, numbers, words, shading, and colour
- Showing numbers with abstract pictures, without direct connection to the physical world, is a surprisingly recent invention,
    - perhaps due to the diversity of skills required: visual-artistic, empirical-statistical, and mathematical
- Statistical graphics were invented around 1750–1800, long after Cartesian coordinates, logarithms, the calculus, and the basics of probability theory

History of data graphics

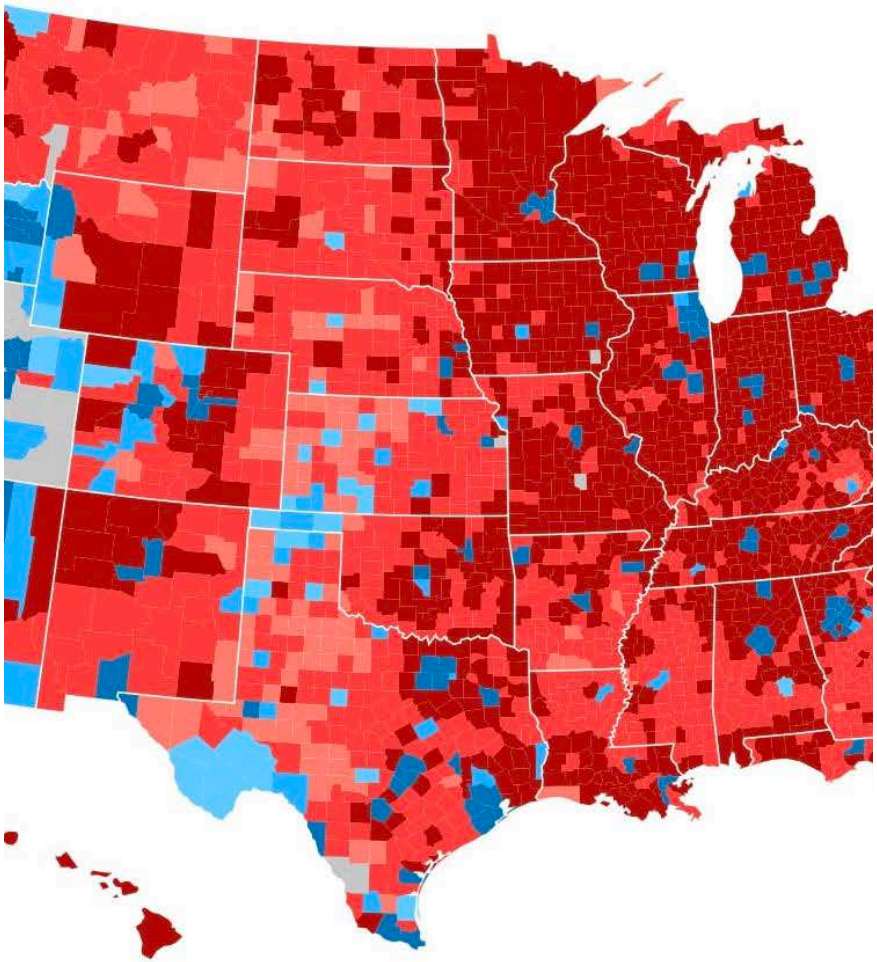
- History of graphics
  - maps, time series, narratives of space and time, abstract graphics
- These illustrations serve multiple purposes
  - providing a set of high-quality graphics
  - helping to demonstrate the terminology
  - telling about the history of graphical development
  - seeing how good statistical graphics can be
  - understanding that visual designs that we take for granted sometimes took even thousands of years (!) to be perfected

Cave paintings

http://colophon.com/gallery/minsky/caves.htm

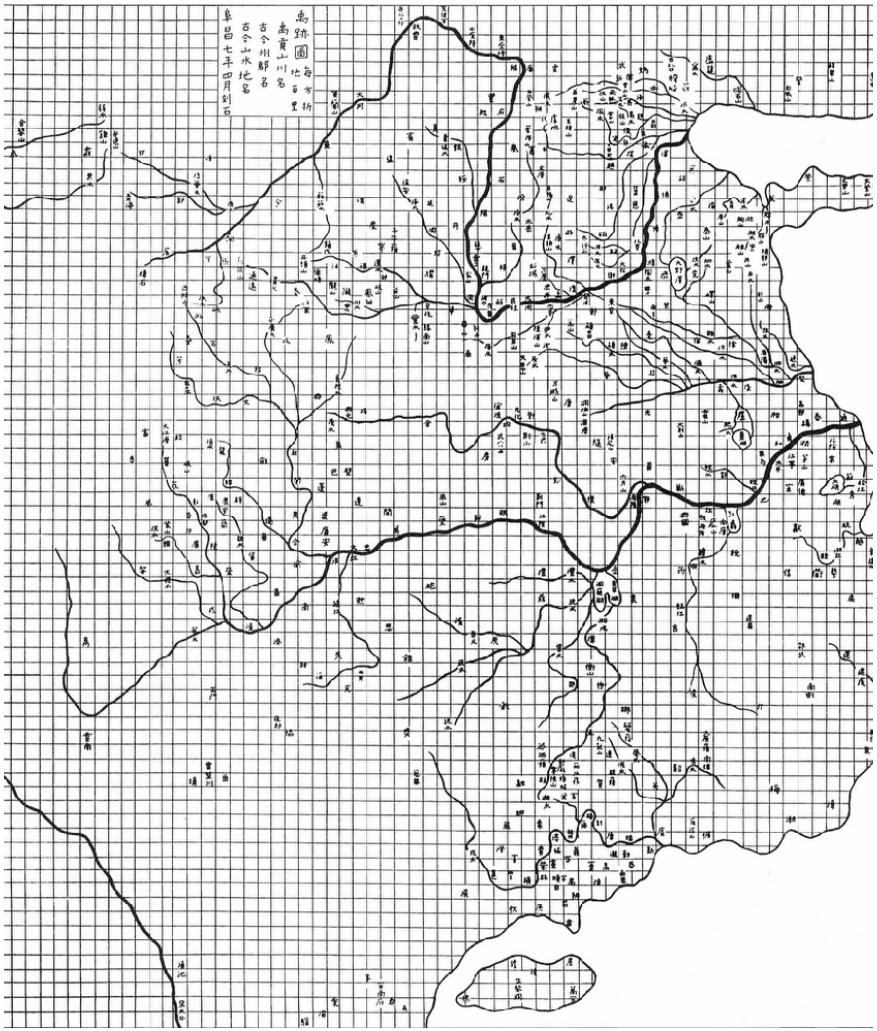
Data maps

- Data map is a graphic showing data on a map
- In the 17th century the combination of cartography and statistical skills required to construct a data map came together
- 5000 years after the first geographic maps were drawn on clay tablets
- Many highly sophisticated geographic maps were produced centuries before the first map containing any statistical material was drawn



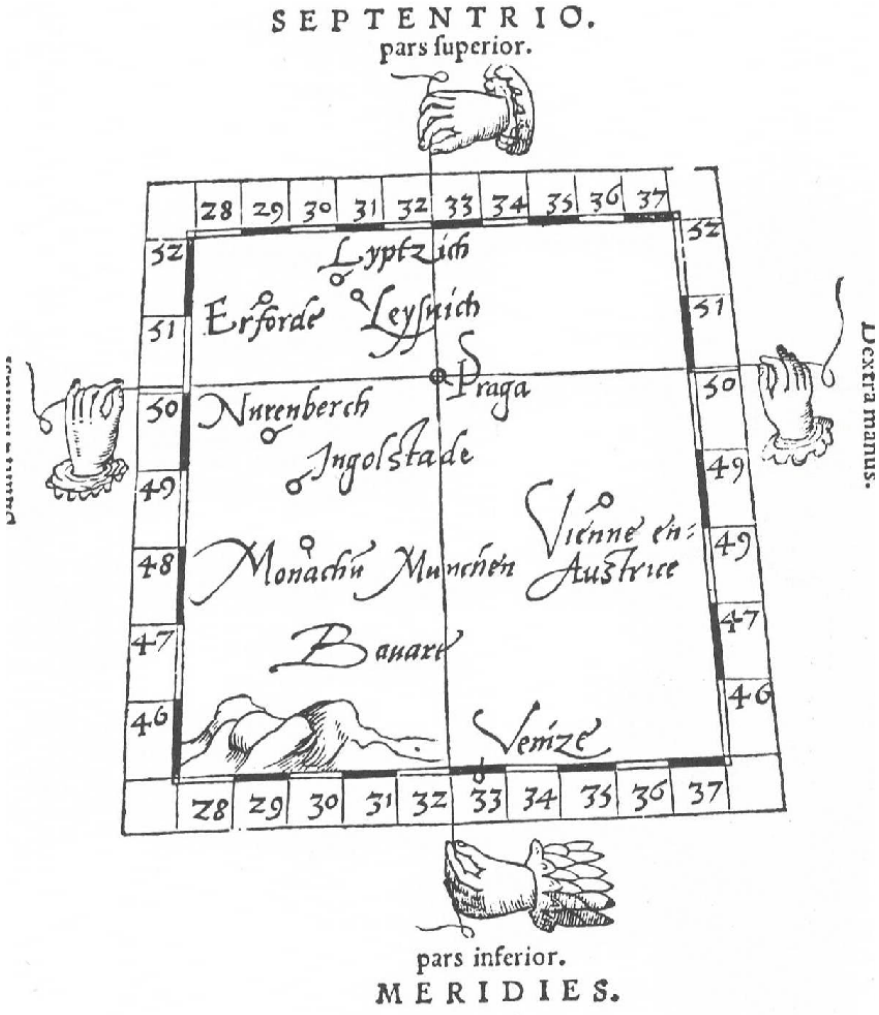
Cartography in China

- The map of the tracks of Yü the Great , detailed map engraved during the 11th century.
- Joseph Needham in Science and civilization of China (1959) described this as "...the most remarkable cartographic work of its age, in any culture".
  - Full grid (100 li scale), a relatively firm coastline, an extraordinary precision of the network of rivers
  - There is nothing like it in Europe till about 1550



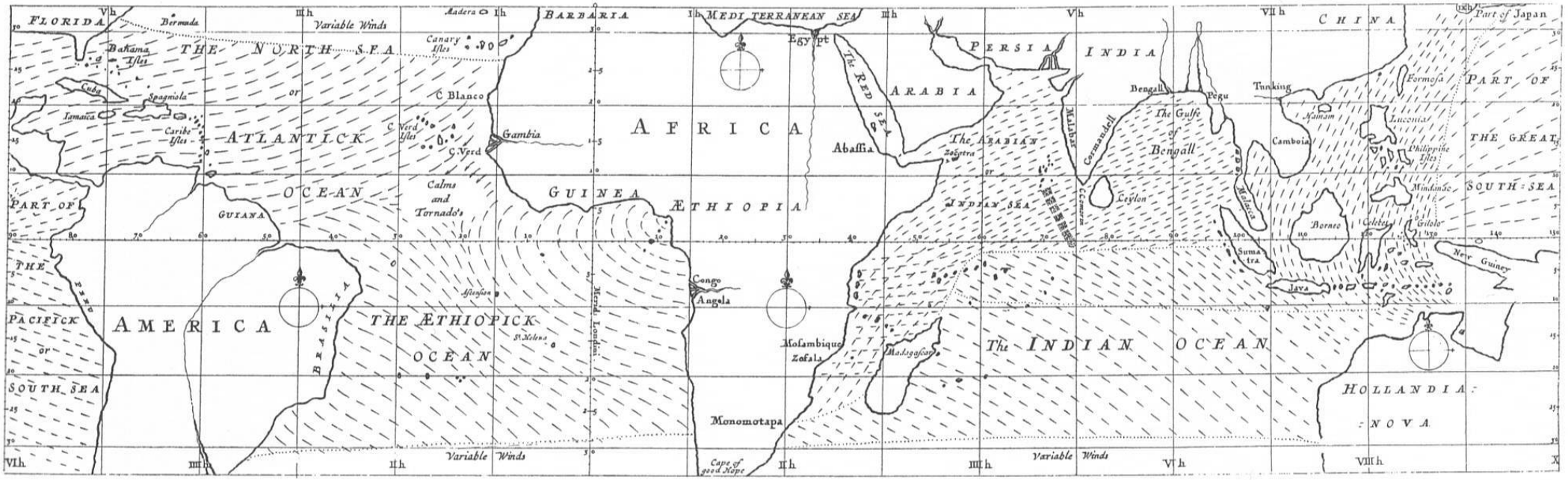
Approaching data maps

- By that time, European cartography had come close to achieving statistical graphicacy, even approaching scatterplots but
  - no one had made the quantitative abstraction of placing a measured quantity on the map's surface
  - let alone the more difficult abstraction of replacing latitude and longitude with some other dimensions



See other visualizations at https://en.wikipedia.org/wiki/Trade_winds

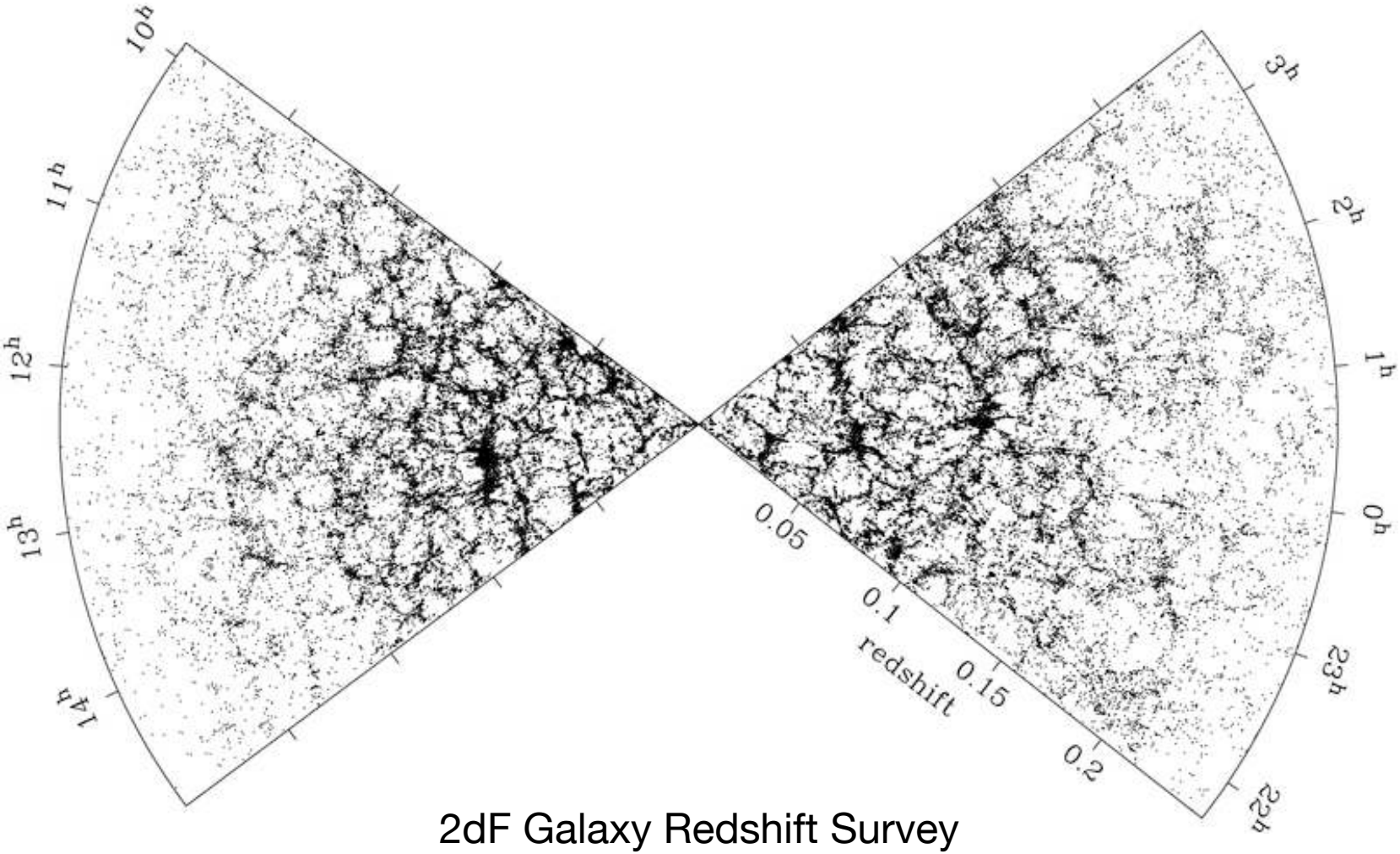
Data maps (wind maps)



- One of the first data maps is a world chart showing trade winds and monsoons (1689)

- Edmond Halley (1656-1742), English astronomer, geophysicist, mathematician, meteorologist, and physicist (best known for computing the orbit of a comet)

Modern data maps



2dF Galaxy Redshift Survey

Modern data maps

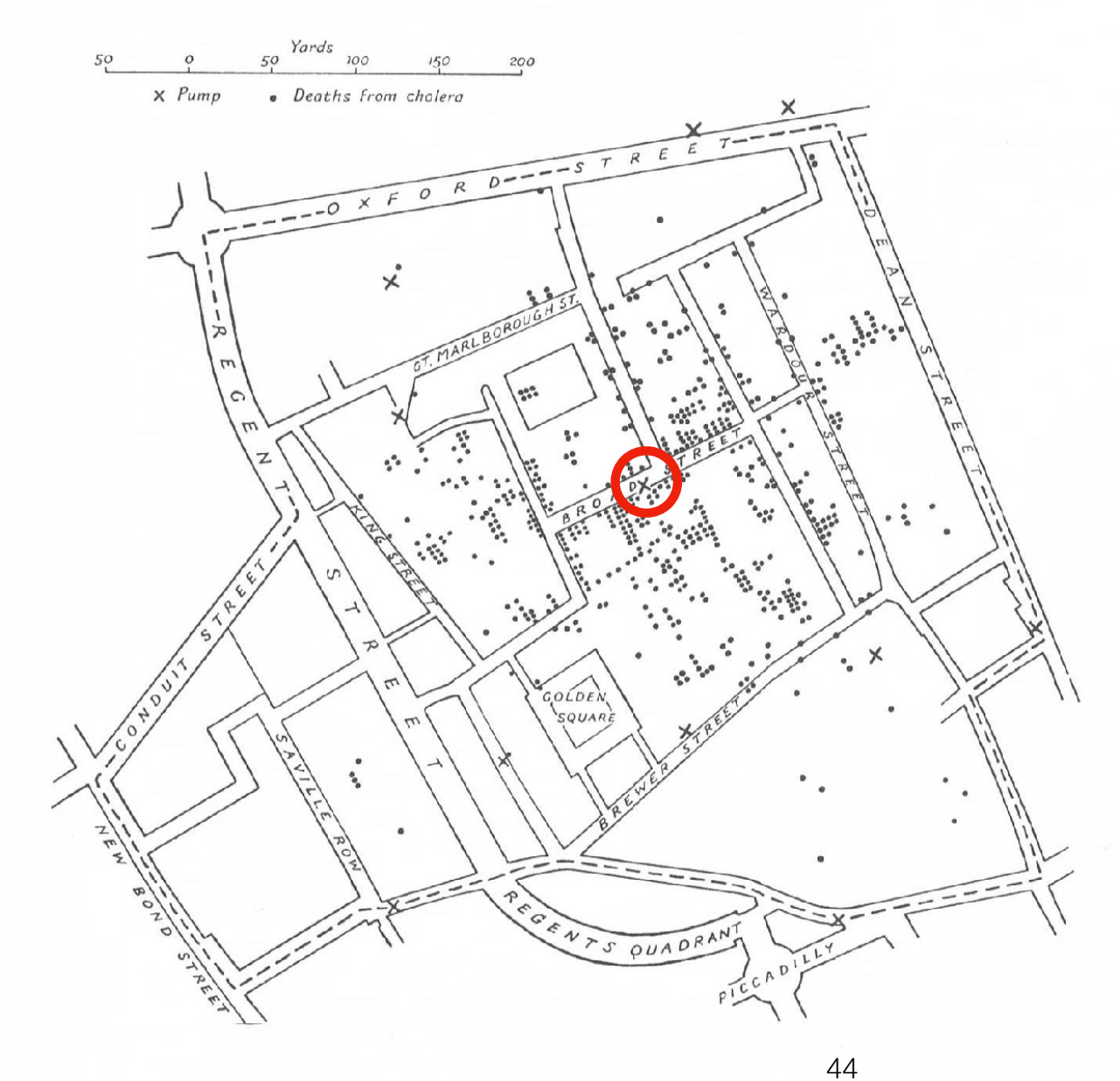
http://paulbutler.org/archives/visualizing-facebook-friends/

Data maps (cholera deaths)

- An early and worthy use of a map to chart non-geographical patterns
- Cholera broke out in Broad Street area in London on 31 August 1854, with over 500 deaths
- John Snow, M.D., obtained a list of deaths and by persistent case-by-case detective work he discovered the probable cause for the epidemics: a water pump at the Broad Street
- The pump handle was removed on 7 September and the epidemics ended.
- Previously it was thought that cholera spread via impure air etc.

Kai Puolamäki at Broad Street

Data maps (cholera deaths)



Cholera deaths are marked with dots

11 water pumps are marked with crosses

Why was Dr. Snow taken seriously?

- Data was placed in an appropriate context to make the **relation between cause and effect** apparent. Time series, for instance, would not have been useful for finding the cause in this case.

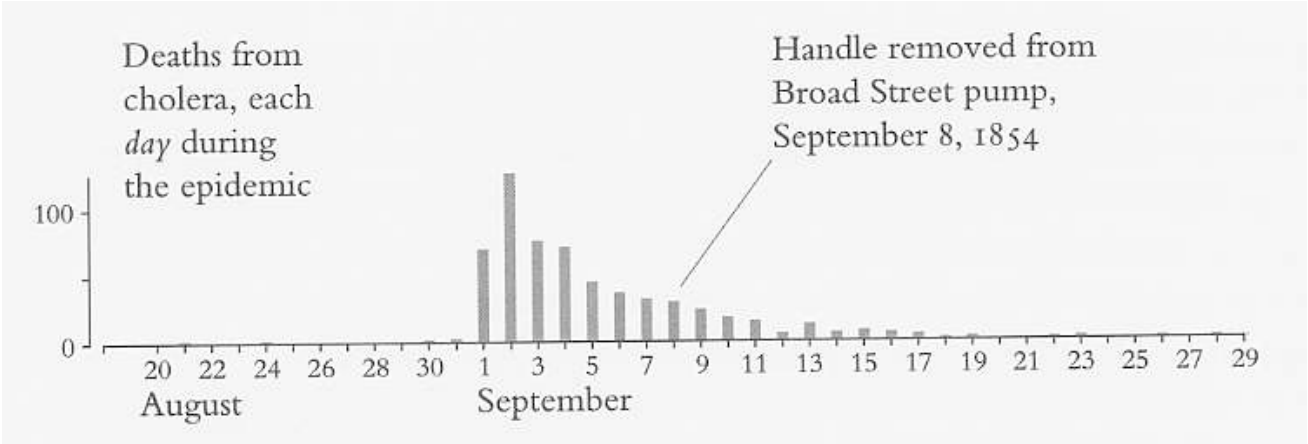
- **Quantitative comparisons** were made. For example, Snow found that the employees of the adjacent brewery were saved because they didn't drink the water from the polluted well. They were saved by the beer(!).

- **Alternative explanations were considered.** Snow also analysed deaths that occurred far away from the Broad Street.

- Snow made a **honest estimation of errors** reported in his map. Snow's map does not for example show the population density.

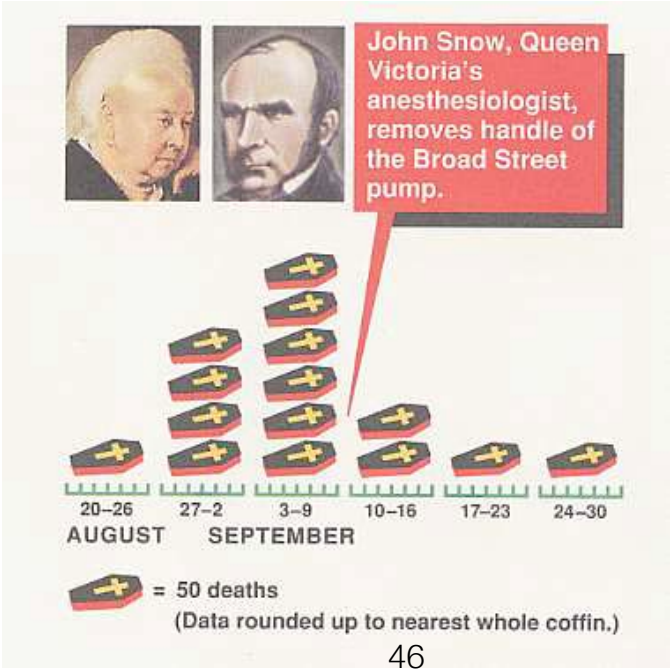
Epilogue

E. R. Tufte, 1998 [VE 36].

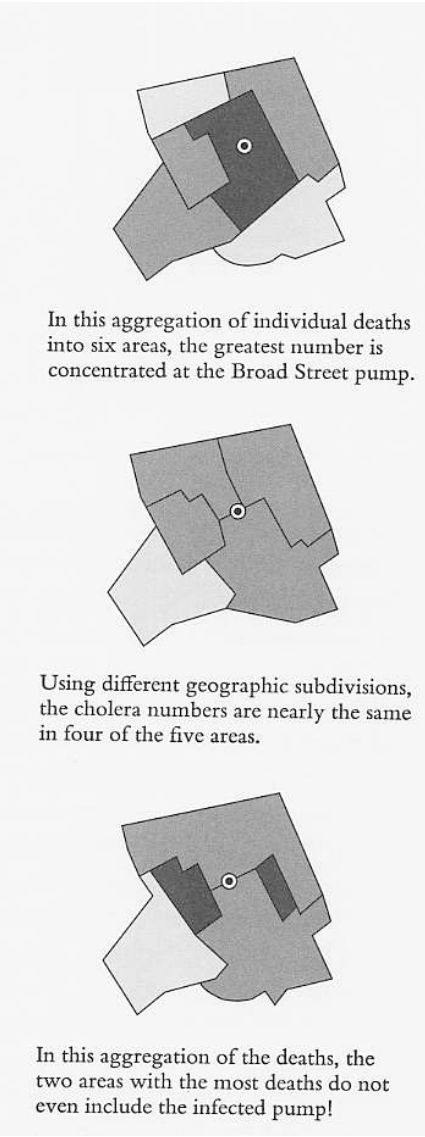


Did Dr. Snow's action end the epidemics or would it have ended by itself...?

John W. Tukey, Science, 1977 [VE 37].



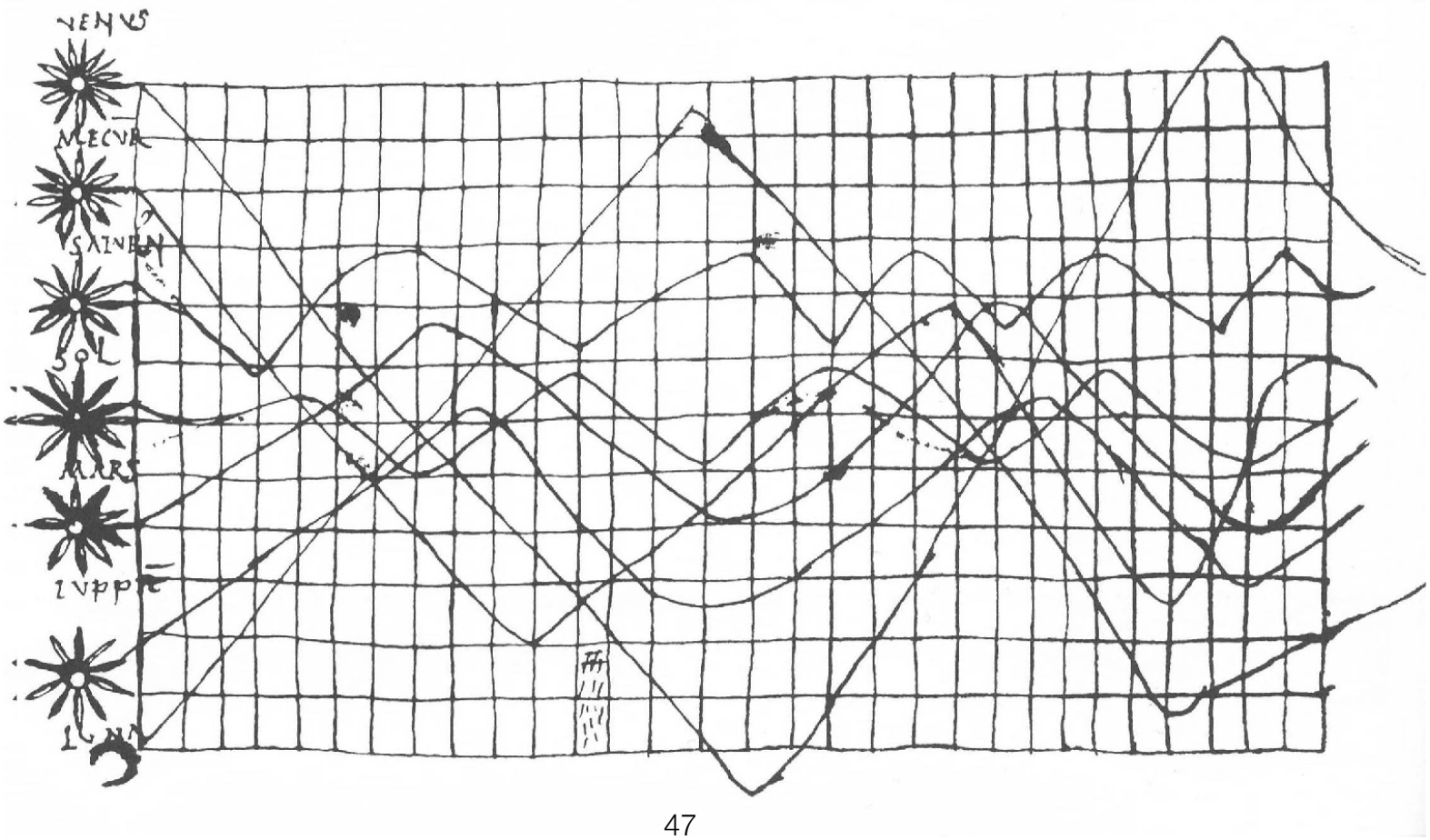
How not to do it:



Mark Monmonier, 1991 [VE 35].

Time series

- Time series plot is one of the oldest and most frequent graphic design
- Part of the text for monastery schools, 900–1100 [T 28].

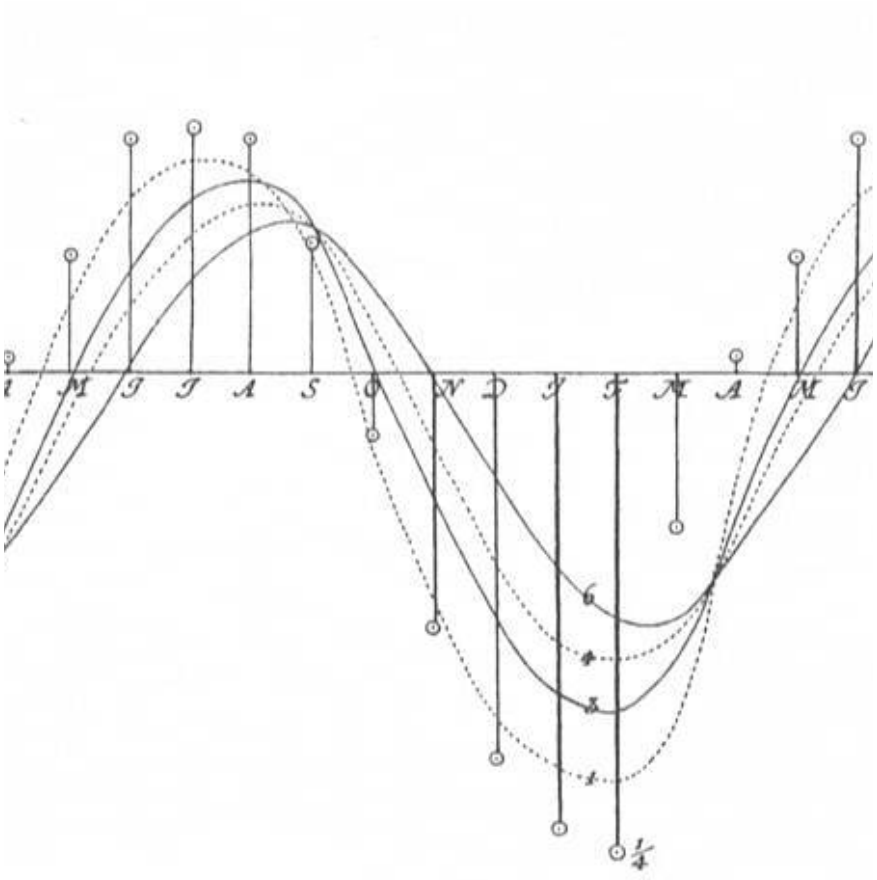


Time series

- Johann Heinrich **Lambert** (1727–1777)
  - Swiss mathematician, physicists, astronomer, ...
  - hyperbolic geometry and properties of map projections
  - the law of light absorption (Beer-Lambert Law)
- William **Playfair** (1759–1823)
  - Scottish engineer and a political economist
  - founder of graphical methods of statistics
  - pioneer of information graphics

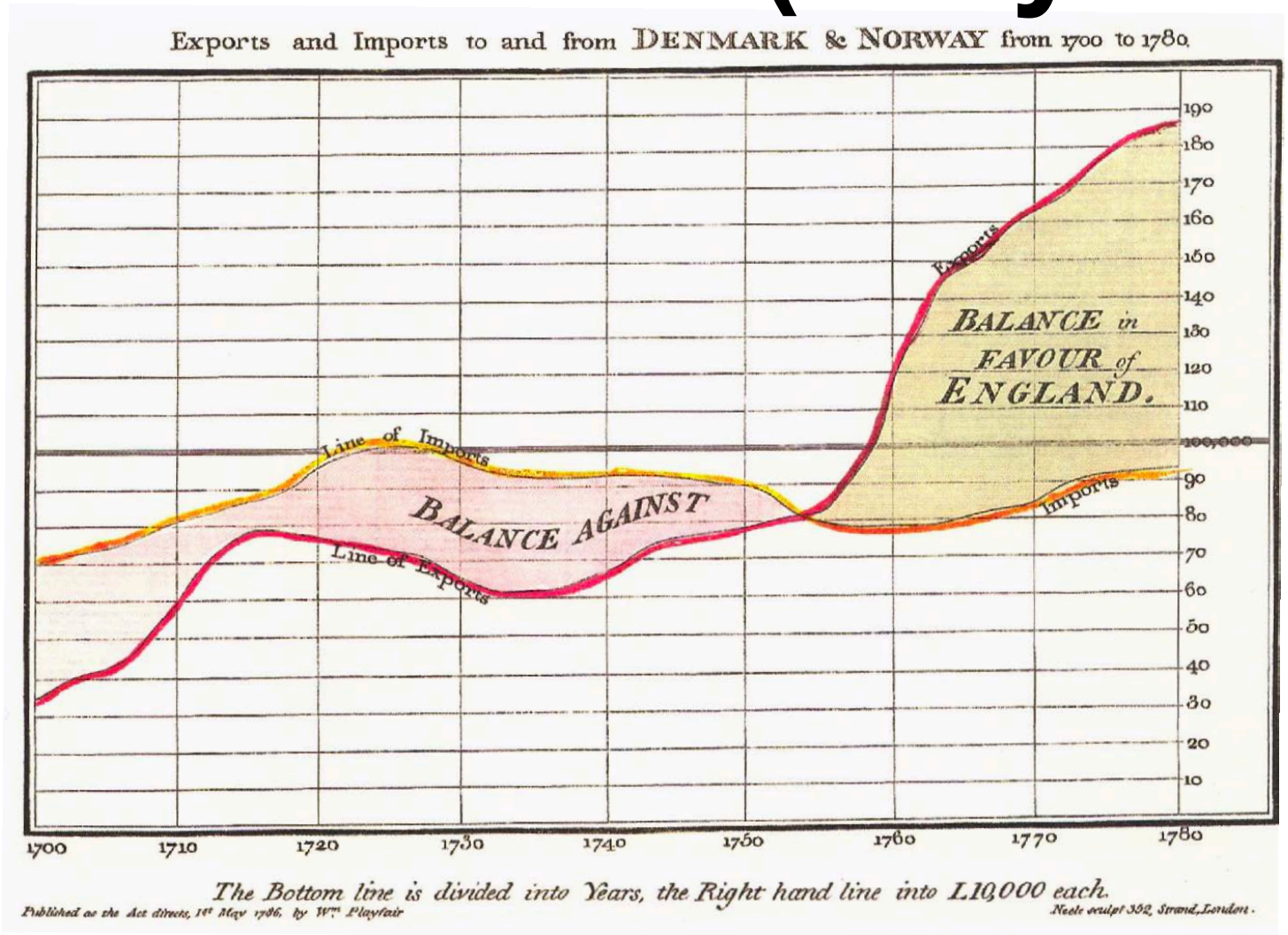
Time series (Lambert)

- This drawing of Johann Lambert (1779) shows the periodic variation in soil temperature in relation to the depth under the surface
  - the greater the depth, the greater the time lag in temperature responsiveness
  - modern time series graphics differ little from those of Lambert



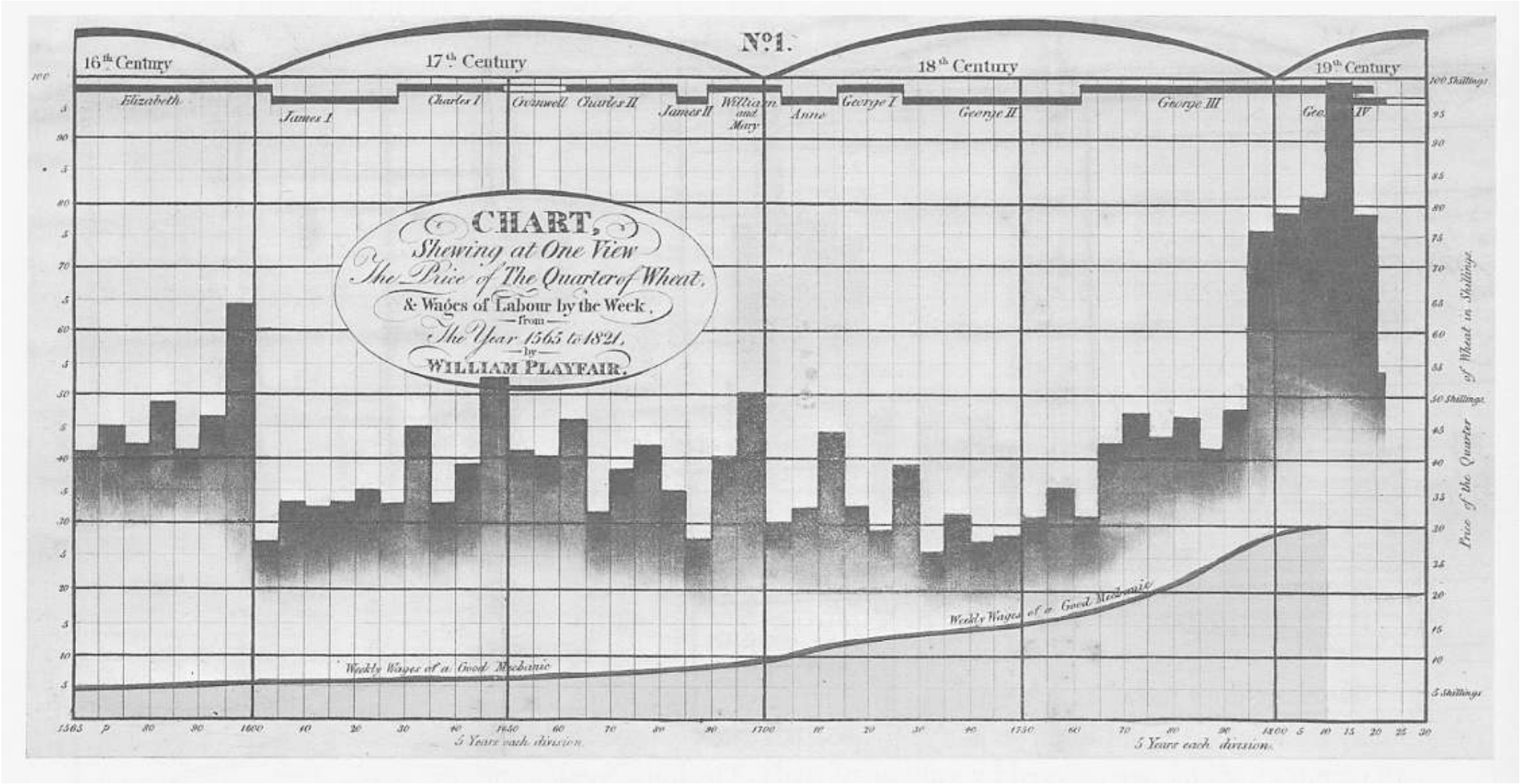
Lambert, *Pyrometrie* (Berlin, 1

Time series (Playfair)

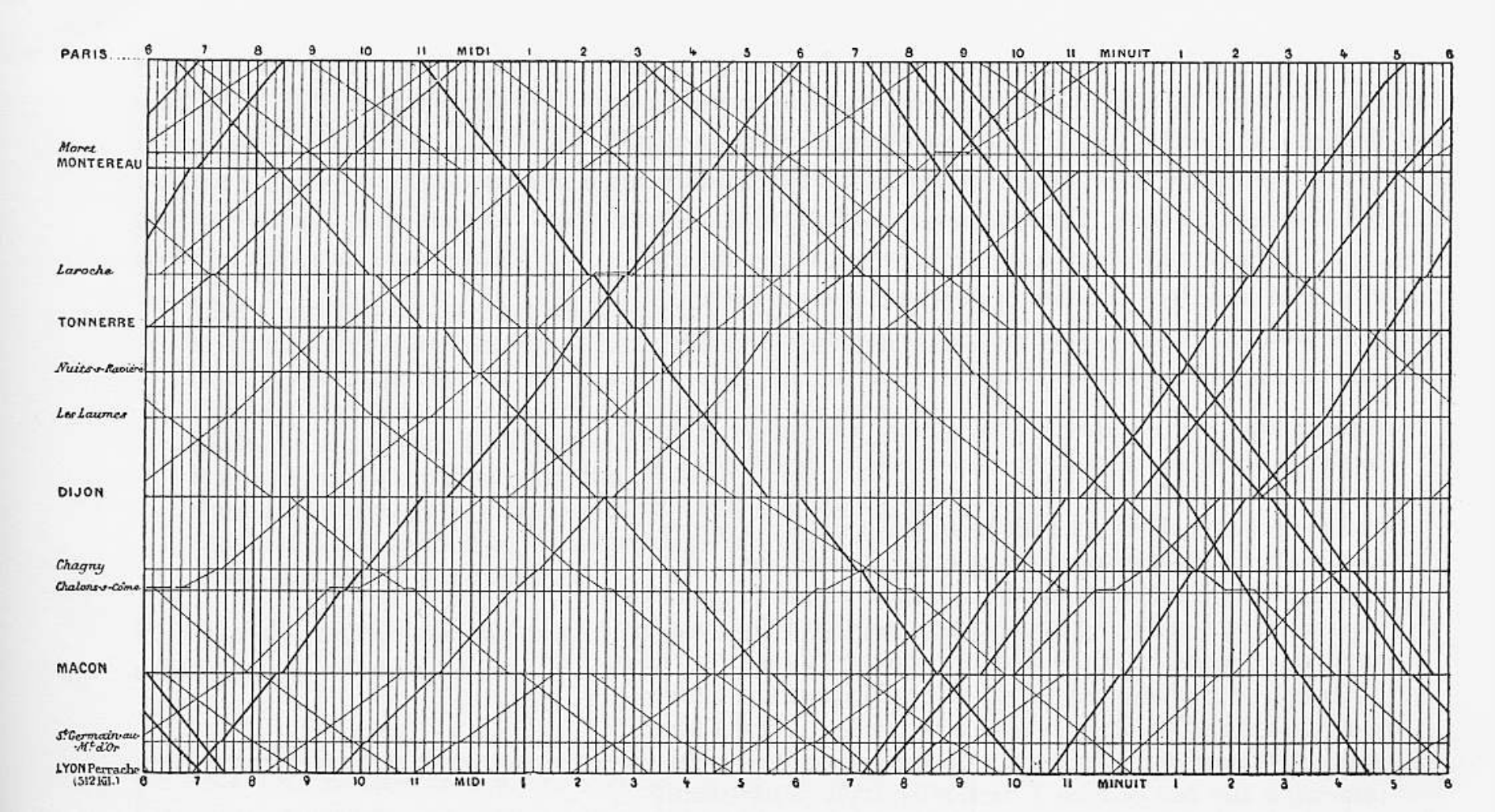


- Playfair was the first to plot economic time series in his book *Commercial and political atlas* (London, 1786).

Time series (Playfair)



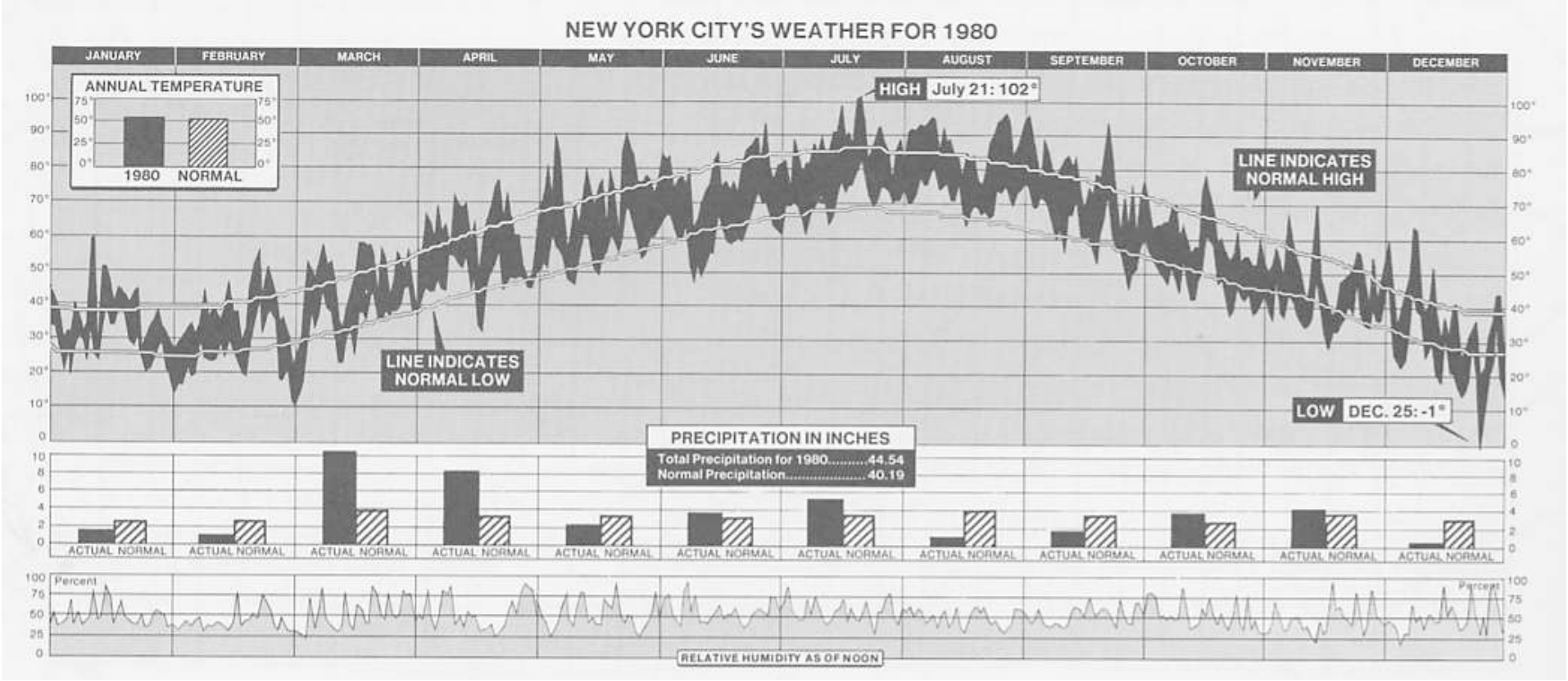
Time series (train schedule)



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

Complex time series

- Time series are best used for data sets that include rich, variable and complex statistical material. Smaller and simpler data sets are described better with a few numbers and a table.



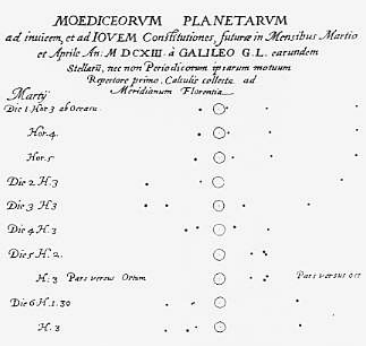
New York Times, 11[53]January 1981 [T 30].

Narratives of space and time (= map + time series)

- Adding spatial dimensions can enhance greatly the explanatory power of time series displays
  - the data are moving over space (2–3 dimensions) as well as over time
  - space-time story graphics can illustrate how multivariate complexity can be subtly integrated into graphical architecture
  - the integration can be gentle and unobtrusive that viewers (or users) are hardly aware that they are looking to a world of four or more dimensions

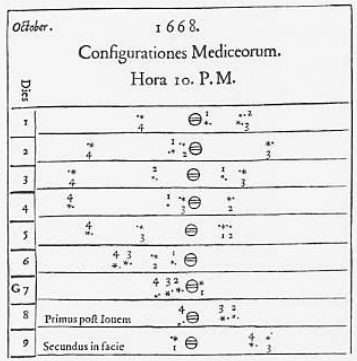
Jovian moons

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



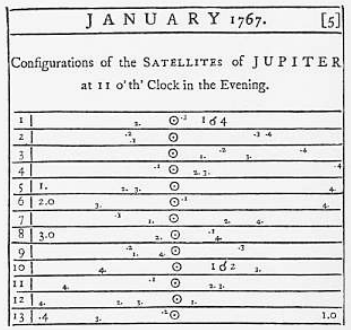
MOEDICEORVM PLANETARVM

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



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

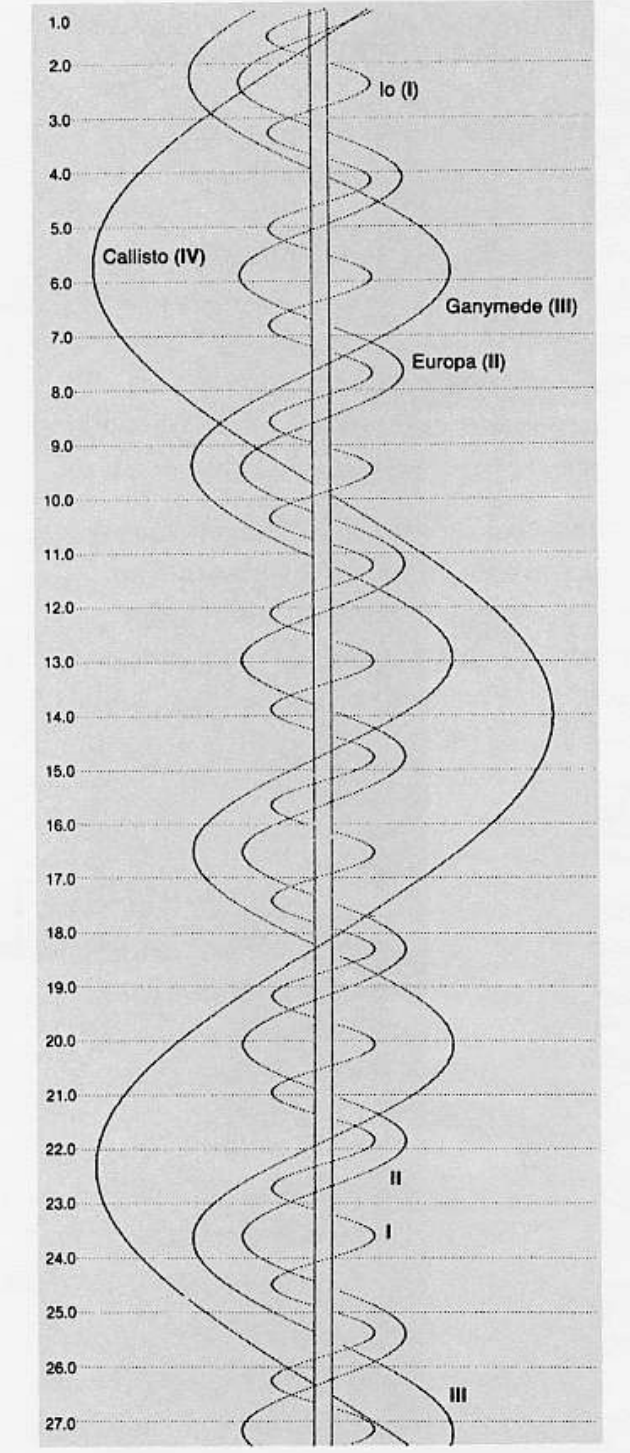
Jean Domenique Cassini, *Ephemerides Bononienses Mediceorum syderum ex hypothesibus, et tabulis Io,* (Bologne, 1668), p. 34.



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

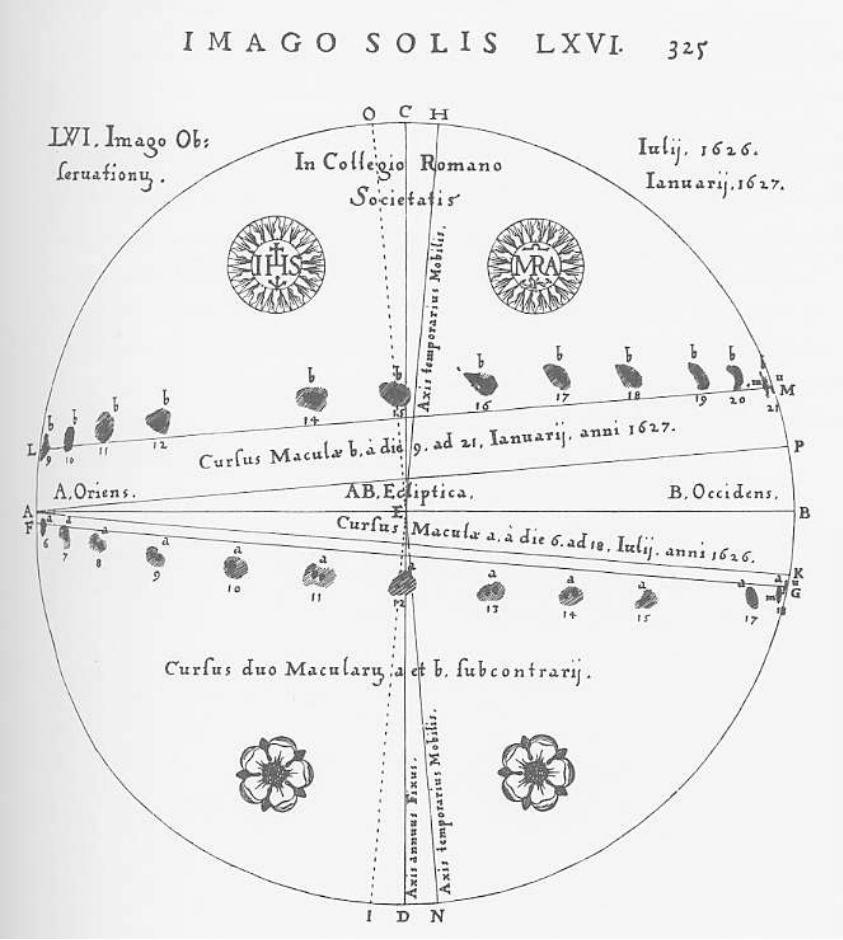
Bureau des Longitudes, *Connaissance des Temps* (Paris, 1766), p. 5.

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

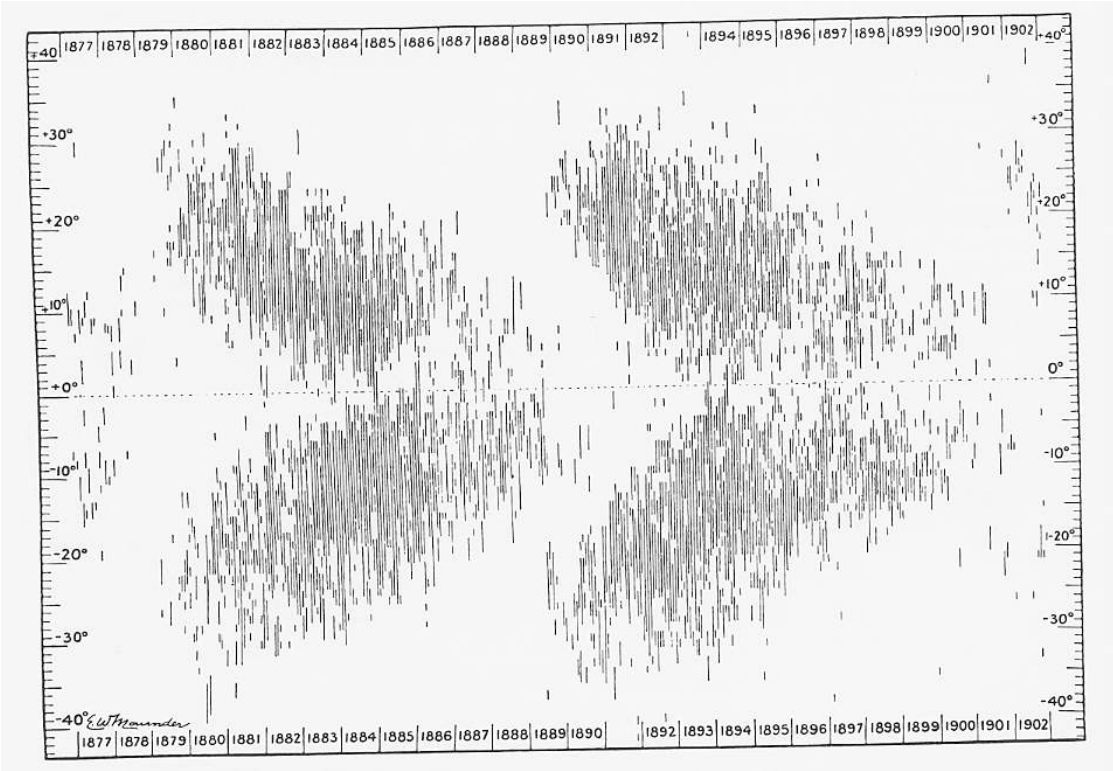


Sky & Telescope, 1988 [EI 100].

Sunspots



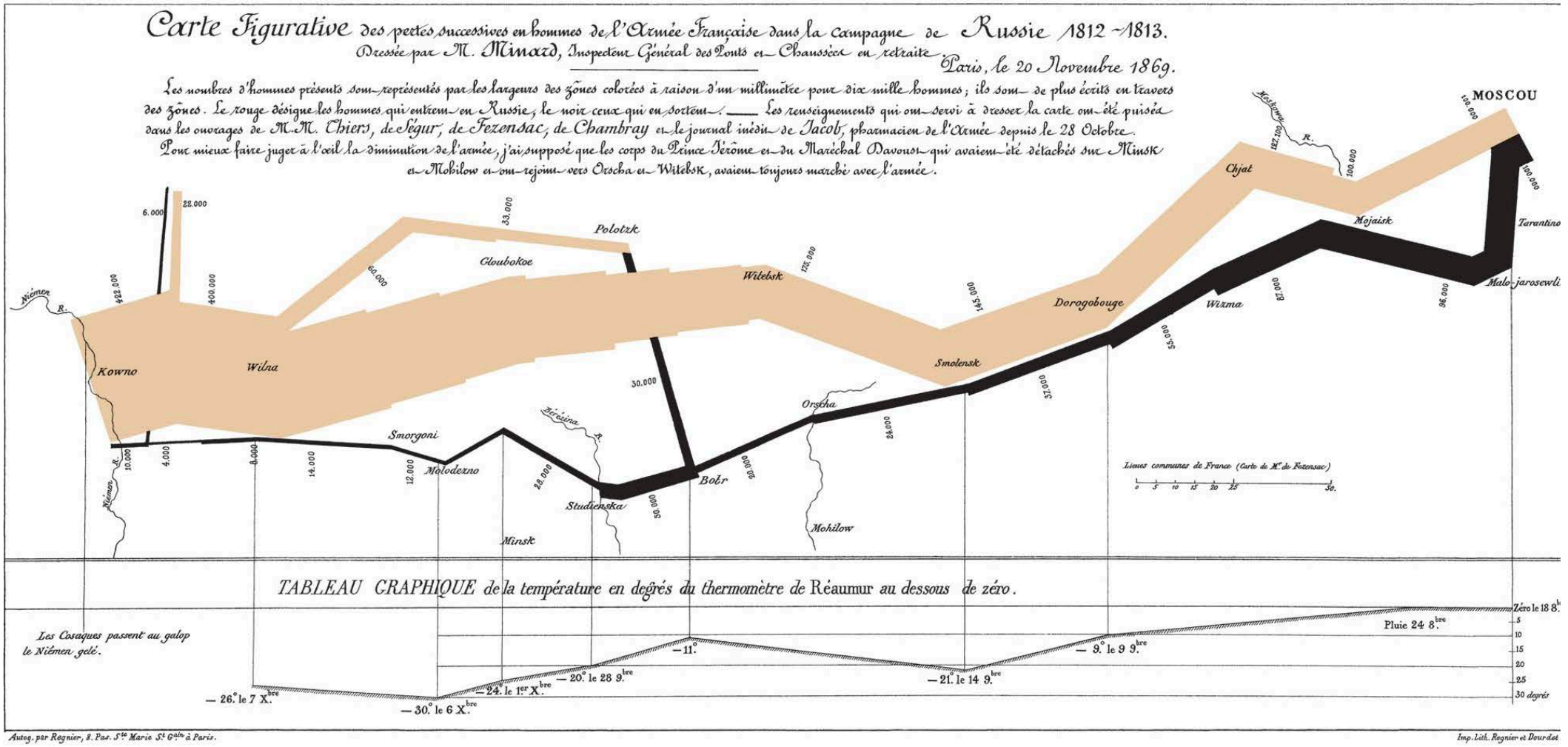
Christopher Schneier, 1630 [EI 21].



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

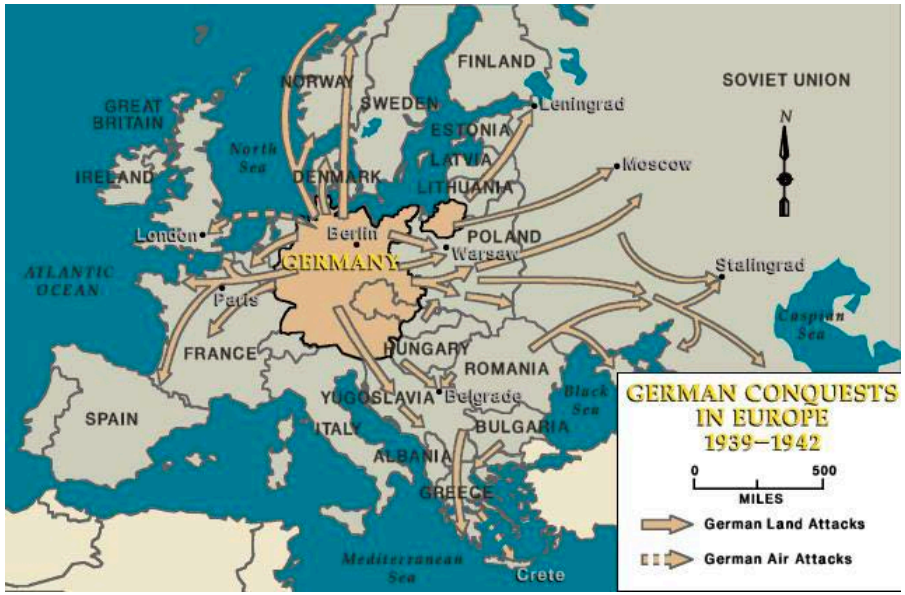
Carte Figurative by Minard

- Charles Joseph **Minard** (1781-1870)
  - French civil engineer
  - designed graphic of the terrible faith of Napoleon’s army in the Russian campaign (1869)
- Several variables are plotted
  - the size of the army
  - its location on a two-dimensional surface
  - the direction of the army’s movement
  - the temperature at various dates during the retreat
- Tufte: ”It may well be the best statistical graphic ever drawn”



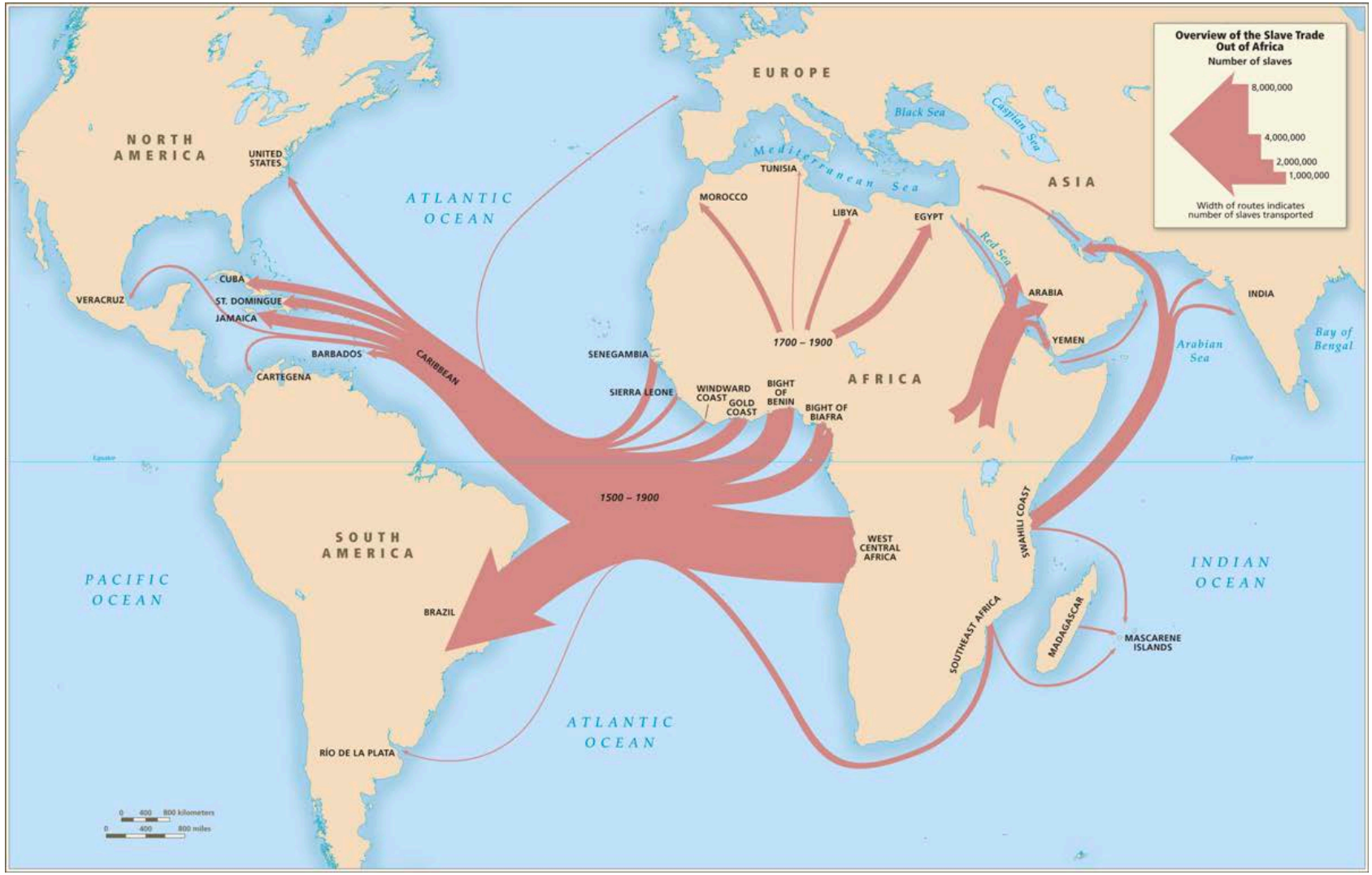
Discuss with your neighbour and find 1-3 reasons why this is good (or bad?)

Space-time narratives



(from the United States Holocaust Memorial Museum)

Space-time narratives

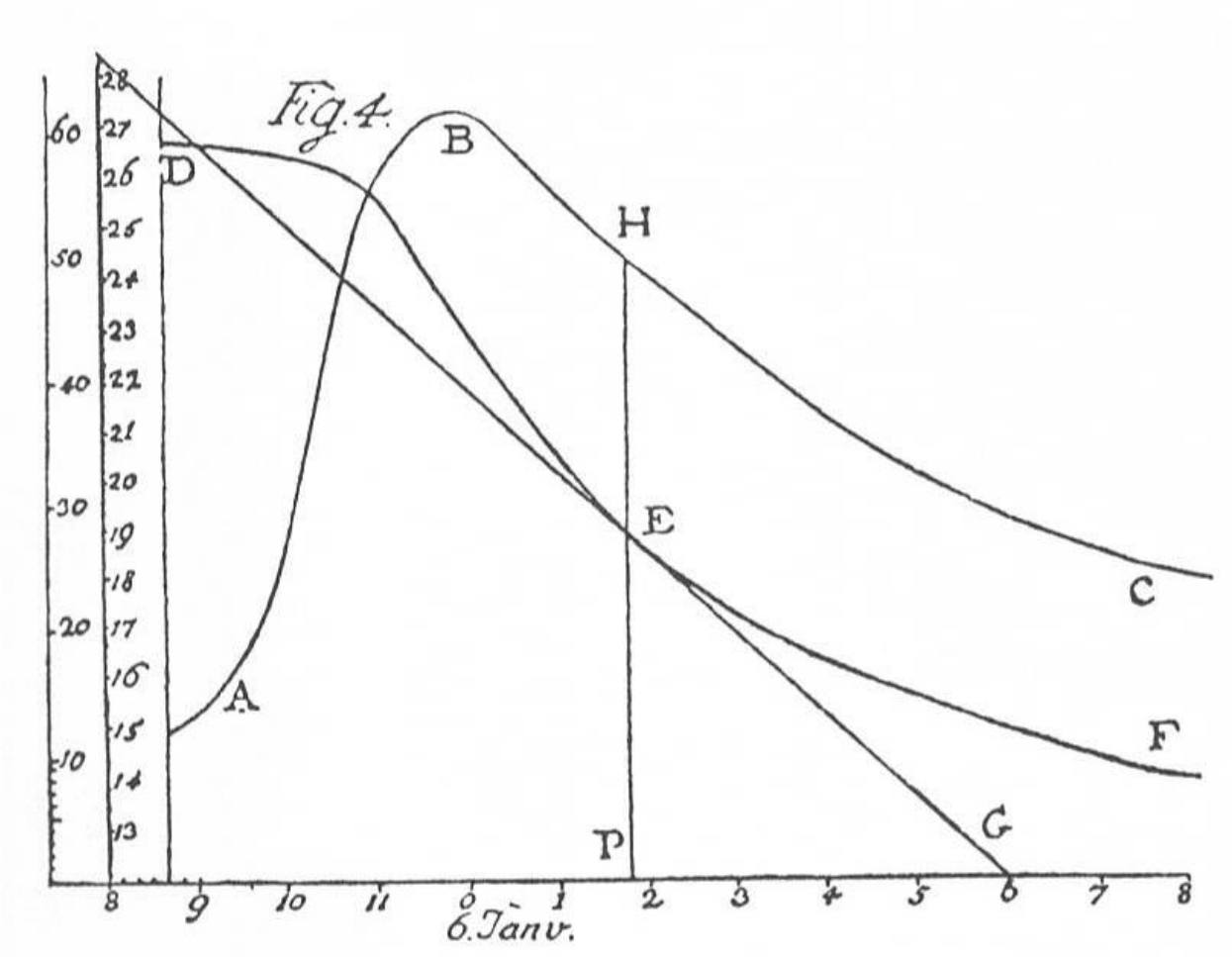


un.org

Relational graphics

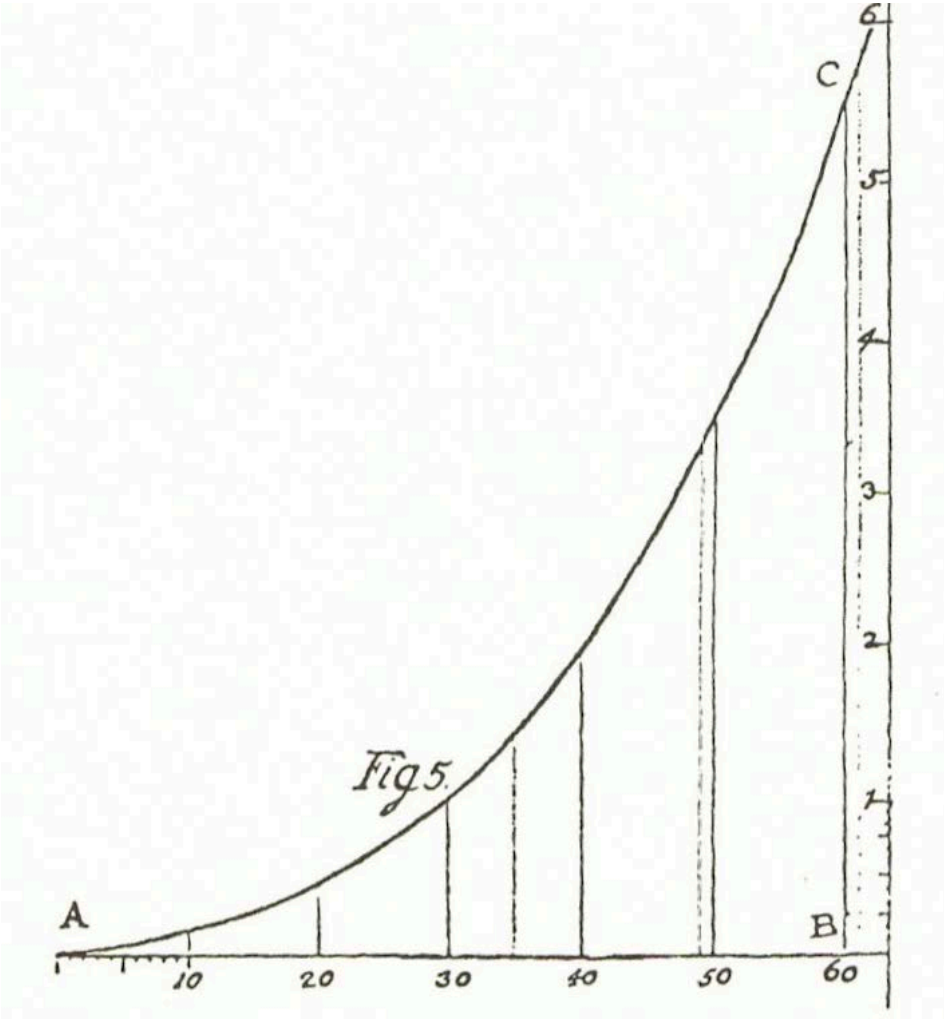
- The invention of data graphics required replacing the coordinates of the map with more abstracts measures, not based on geographical analogy
  - moving from maps and time series to fully abstract statistical graphics was a **big** step
  - thousands of years passed before this step was taken
  - Lambert,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

From overlapping time series...



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

...to relational graphics



X = temperature

Y = measured rate

Graphical excellence

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

Next lectures

- Graphical practice (cont.)
- Theory of data graphics
- Something to watch meanwhile
  - Martin **Krzywinski**’s lecture(s)
    https://youtu.be/M-rTAr3pj5g
  - Hans **Roesling**'s TED talks
    https://youtu.be/hVimVzgtD6w