MTTTS17
Dimensionality Reduction and Visualization

Spring 2020
Jaakko Peltonen

Lecture 4: Graphical Excellence

Slides originally by Francesco Corona

Outline

Information visualization

Edward Tufte

The visual display of quantitative information

Graphical excellence
- Data maps
- Time series
- Space-time narratives
- Relational graphics

Graphical integrity
- Distortion in data graphics
- Design and data variation
- Visual area and numerical measure

Information visualization

Data graphics visually display measured quantities by means of the combined use of *points, lines, a coordinate system, numbers, words, shading* and *colour*

The use of abstract, non-representational pictures to show numbers is a surprisingly recent invention, perhaps because of the diversity of skills required:

- visual-artistic, empirical-statistical, and mathematical

It was not until 1750-1800 that statistical graphics were invented, long after Cartesian coordinates, logarithms, the calculus, and the basics of probability theory

William Playfair (1759-1823) developed/improved upon (nearly) all fundamental graphical designs, seeking to replace conventional tables of numbers with systematic visual representations

- A Scottish engineer and a political economist
- The founder of graphical methods of statistics
- A pioneer of information graphics

Information visualization (cont.)

Modern data graphics can do much more than simply substitute for statistical tables

Graphics are instruments for reasoning about quantitative information

Often, the most efficient way to describe, explore, and summarize a set of numbers (even a very large set) is to look at pictures of those numbers

Of all methods for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and the most powerful

Edward Tufte

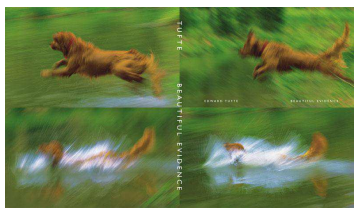
The first part of the course is about the design of statistical graphics but it is also about how to communicate information through the simultaneous presentation of

words, numbers, and pictures

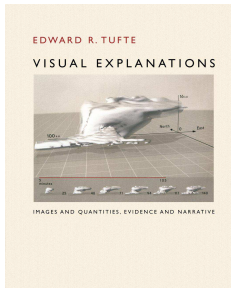
Edward Tufte (1942-), an American statistician and Yale University emeritus professor of political science, computer science and statistics

Today he is known as ET and is a sculptor

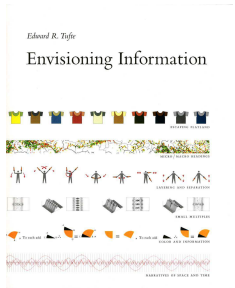
www.edwardtufte.com



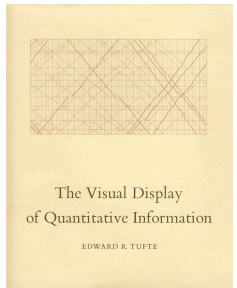
Beautiful evidence



Visual explanations



Envisioning information

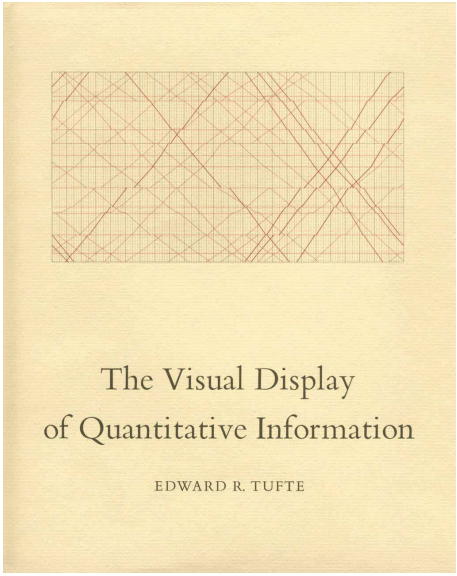


The visual display of quantitative information

The visual display of quantitative information

The visual display of quantitative 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



The visual display of quantitative information (cont.)

The first part is about reviewing Graphical practice of the two centuries since Playfair

Graphical excellence: Historical glories for the main types of graphical designs (Data maps, Time series, Space-time narrative and Relational graphics)

Graphical integrity: Lapses and lost opportunities (Distortion, Design and data variation, Visual area and numerical measure and Context)

Sources of integrity and sophistication

Excellence

Graphical excellence is all about the well-designed presentation of interesting data

- Excellence is a matter of substance, of statistics, and of design
- Complex ideas communicated with clarity, precision, and efficiency

Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space

- It is nearly always multivariate
- It tells the truth about the data

Graphical excellence is illustrated for fundamental graphical designs (Data maps, Time series, Space-time narrative and Relational graphics), and serve multiple purposes

- providing a set of high-quality graphics
- constructing a theory of data graphics
- helping to demonstrate a descriptive terminology
- telling about the history of graphical development
- seeing how good statistical graphics can be

Integrity

For many the first word that comes to mind when they think about infocharts is LIE

Some graphics do distort the data, making it hard for the viewer to learn the truth

Data graphics are no different from words in this regard, for any means of communication can be used to deceive

There is no reason to believe that graphics are especially vulnerable to exploitation by liars

Most of us have excellent graphical lie detectors that help us see through frauds

False graphics are still with us, deception must always be confronted and demolished

Graphical excellence begins with telling the truth about the data

The visual display of quantitative information (cont.)

The second part is about providing a simple language for the Theory of data graphics

Data-ink and redesign: Empirical measures of graphical performance and sequential improvement of graphics through revision and editing

Chartjunk: Bad graphical displays and why graphics do not become attractive and interesting through the addition of ornamental hatching

Data-ink maximization, Malfunctioning elements, High-resolution and Aesthetics

Data-ink and redesign

Five principles in the theory of data graphics produce substantial changes in design

- Above all show the data
- Maximize the data-ink ratio
- Erase non data-ink
- Erase redundant data-ink

These principles apply to many graphics and yield a large series of design options

- Revise and edit

Chartjunk and data-ink maximization

Ornamental hatching leading to chartjunk does not achieve the goals of its propagators

- Graphics do not become attractive and interesting through chartjunk and ducks
- Chartjunk can turn bores into disasters but it can never rescue a thin data set

The best designs are intriguing and curiosity-provoking, drawing the viewer into the wonder of data, sometimes by narrative power, sometimes by immense detail and sometimes by elegant presentation of simple but interesting data

- No information, no sense of discovery, and no wonder is generated by chartjunk

Most of the graphic's ink should vary only in response to data variation

- Maximizing data-ink is but a single dimension of a complex design task

Graphical practice

Graphical displays 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 place
- Make large datasets coherent
- Encourage the eye to compare different pieces of data
- Reveal the data at several levels of detail, from a 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 a dataset

Graphics must reveal data

Graphical excellence

Excellence in statistical graphics is all about communicating complex ideas with

- Clarity
- Precision
- Efficiency

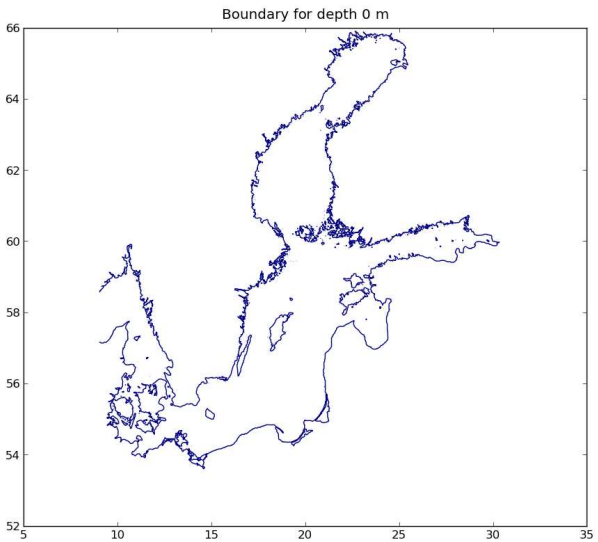
Graphical excellence is illustrated here for fundamental graphical designs

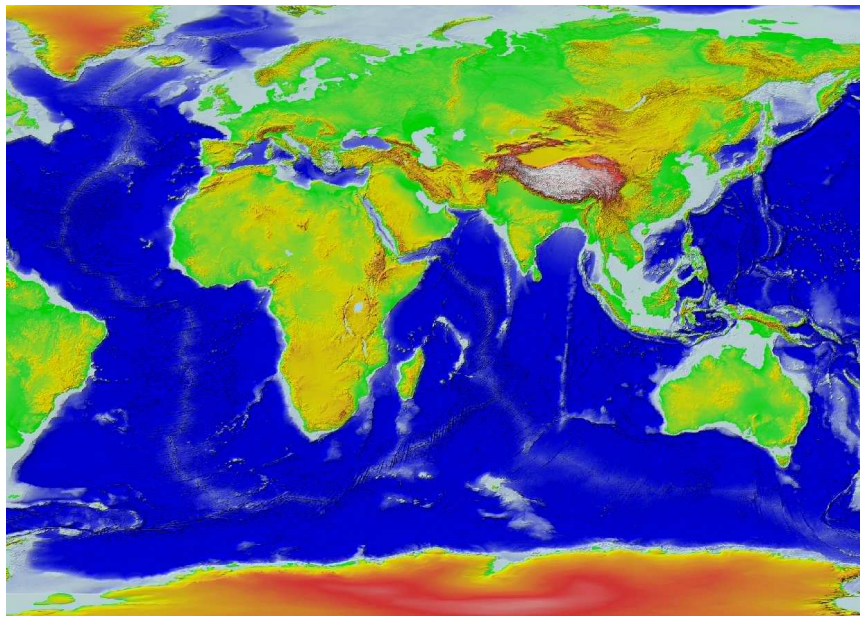
- Data-maps
- Time-series
- Space-time narrative graphics
- Relational graphics

Data maps

It was not until the seventeenth century that 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





... and many highly sophisticated geographic maps were produced centuries before the first map containing any statistical material was drawn

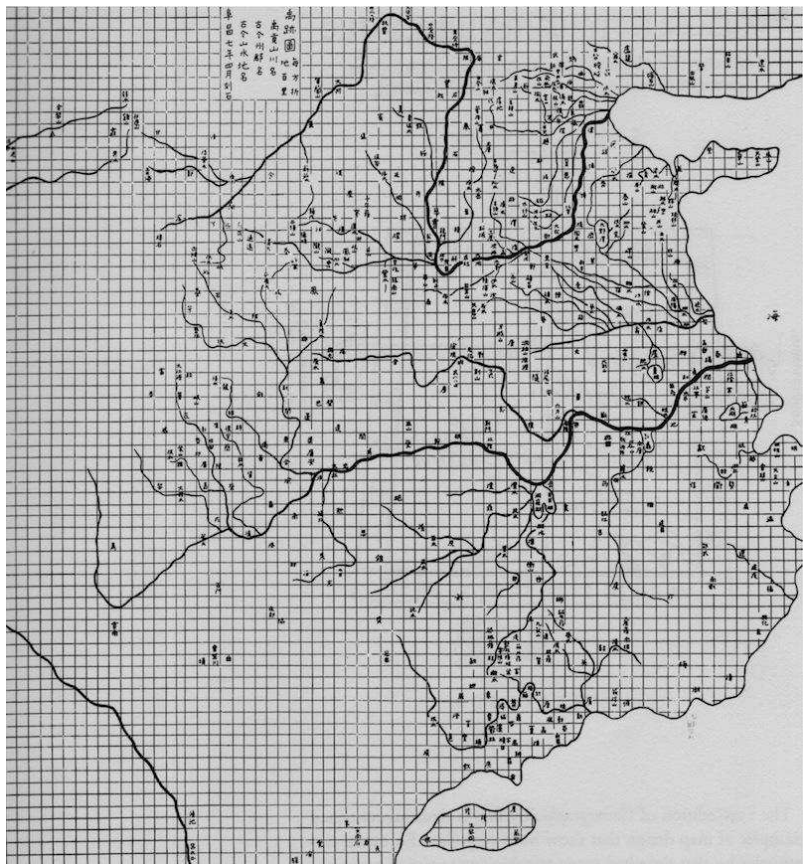
Data maps (cont.)

The map of the tracks of Y¬u the great described by Joseph Needham in Science and civilization of China (1959) is a detailed map engraved during the 11th century A.D.

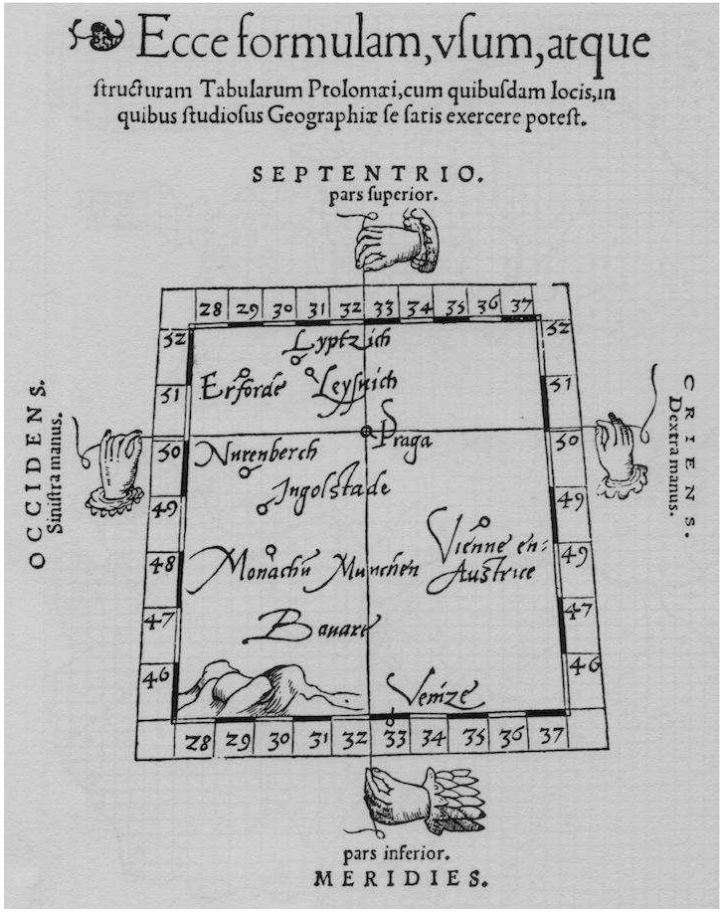
... the most remarkable cartographic work of its age, in any culture

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 A.D.



Data maps (cont.)



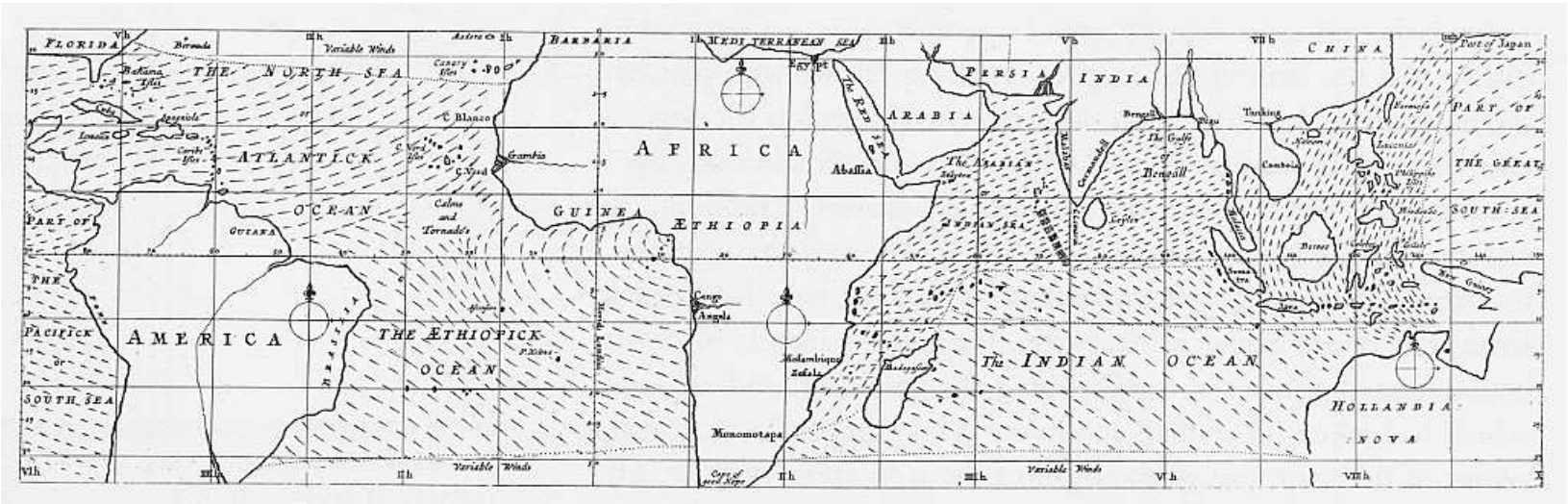
By that time, European cartography had come close to achieving statistical graphicacy, even approaching scatterplots

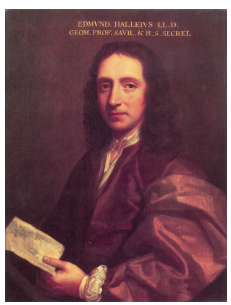
No one had made the quantitative abstraction of placing a measured quantity on the map's surface

Let alone the more difficult abstraction of replacing latitude and longitude with some other dimensions

Data maps (cont.)

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





Edmond Halley (1656-1742) was an English astronomer, geophysicist, mathematician, meteorologist, and physicist

Best known for computing the orbit of a comet

Data maps (cont.)

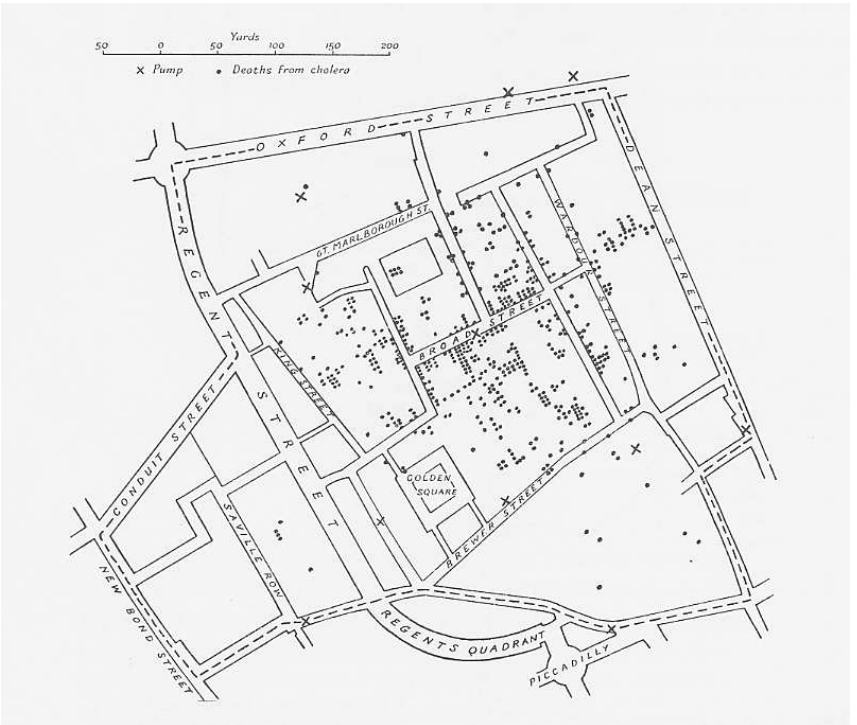
An early and worthy use of a map to chart non-geographical patterns is the famous

Dot map of Cholera deaths in London

- On August 31, 1854, cholera broke out in the Broad Street area (Soho) of Central London
- Over 500 people died
- At that time such diseases were believedto becaused by bad air

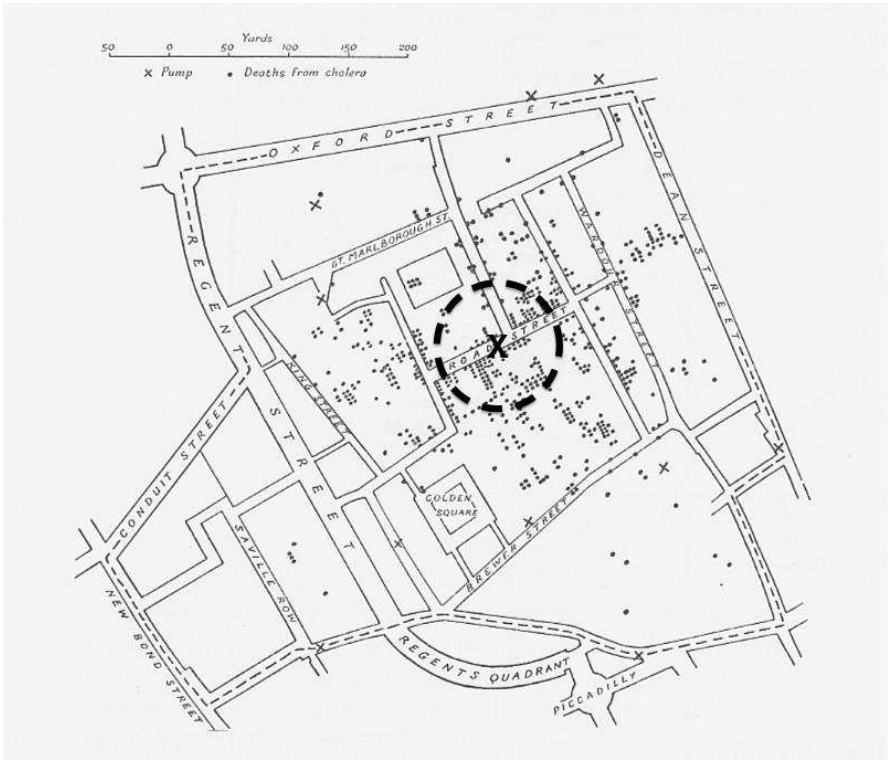
Aphysician, Dr. John Snow (1813-1858), was skeptical about this theory

- On a map of the area, he marked deaths by dots
- The area's eleven water pumps were located by crosses



Data maps (cont.)

From the scatter over the surface of the map, Snow observed that cholera occurred almost entirely among those who lived near (and drank from) the Broad St. pump



He thought he had discovered a probable cause for the epidemics

Chemical and microscope analysis could not conclusively prove it

His argument on pattern of the disease was convincing enough to persuade the local council to disable the pump by removing its handle

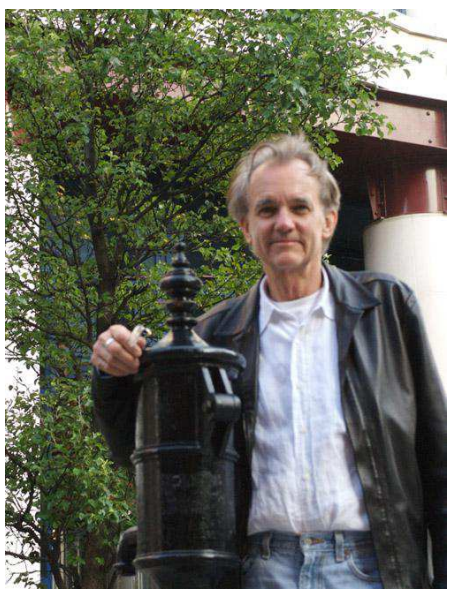
Data maps (cont.)

The outbreak ended when the handle was removed

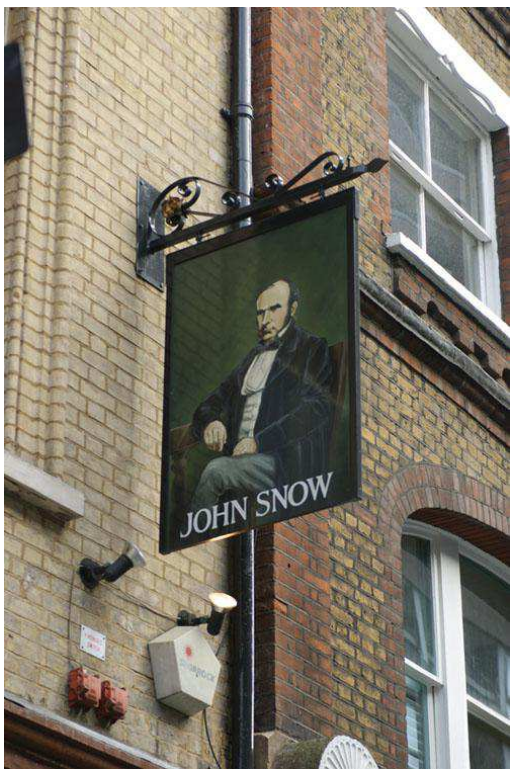
Now they sell it as a souvenir

Why was the mystery solved and why was Snow taken seriously?

- Data were placed in an appropriate context
- The relation between cause and effect highlighted



Data maps (cont.)



Data maps (cont.)

Computerized cartography and modern photographic techniques have increased the density of information some 5K-fold in current data maps compared to Halley's

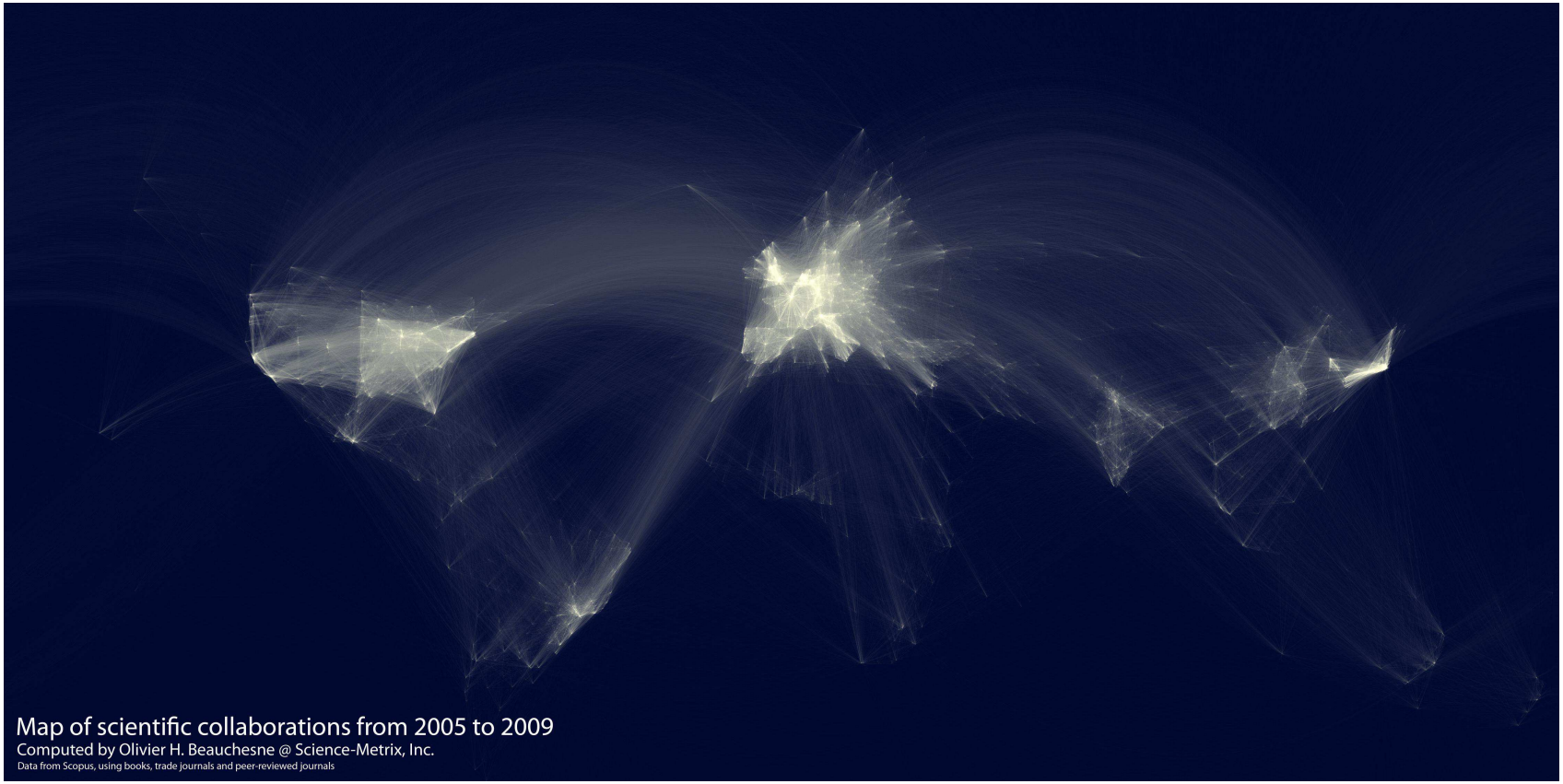
The most extensive data maps, place millions of bits of information on a single page

No other method for the display of statistical information is so powerful

Data maps (cont.)

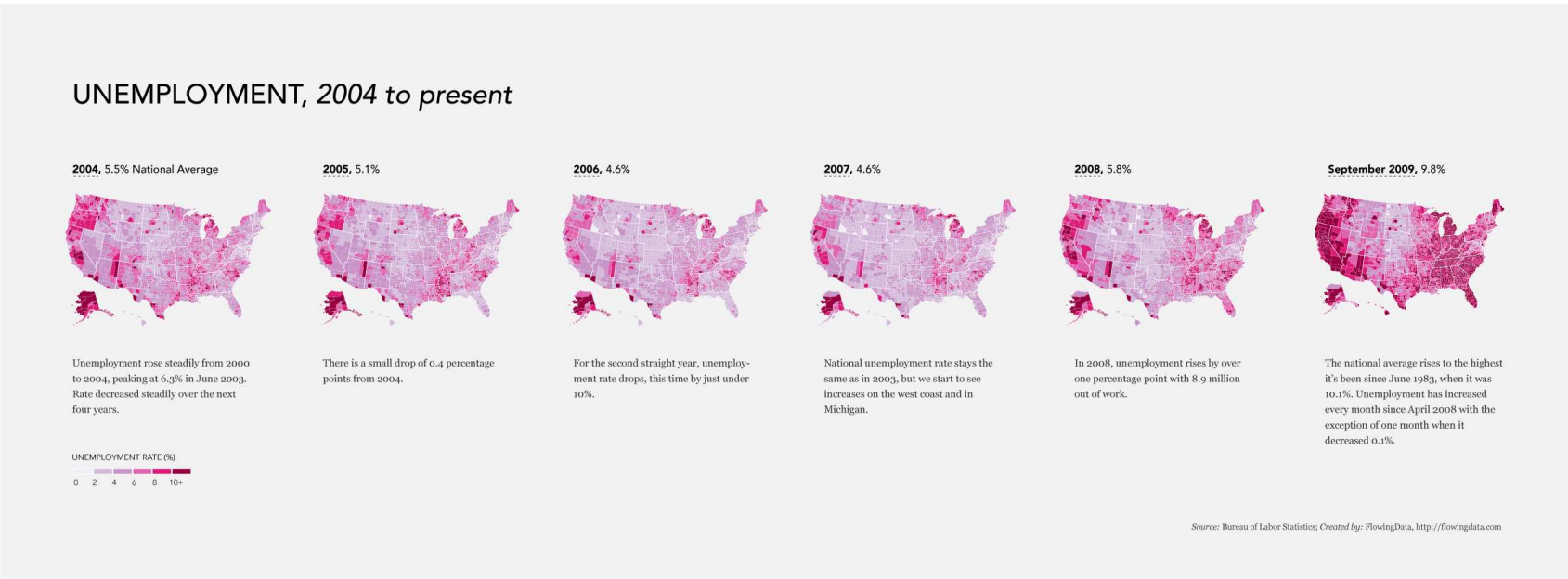
Paul Butler: Visualizing friendships (2010),
http://paulbutler.org/archives/visualizing-facebook-friends/

Data maps (cont.)



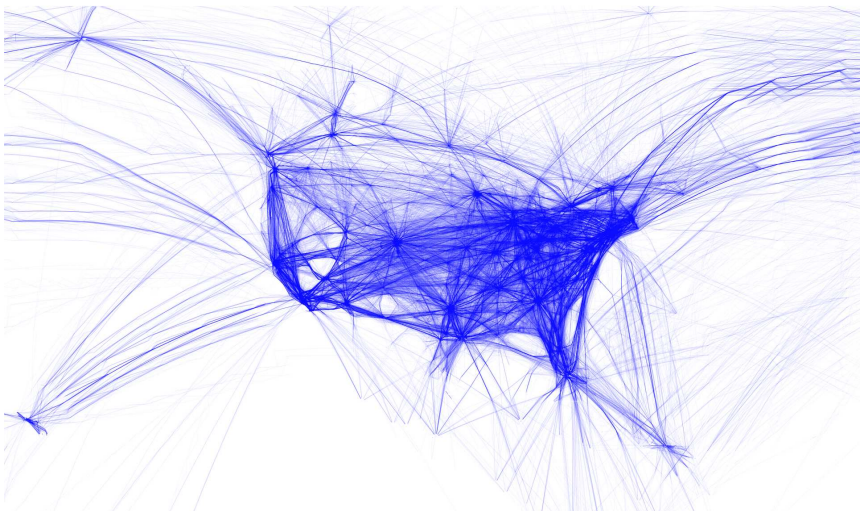
Olivier H. Beauchesne: Map of scientific collaboration between researchers (2011)
http://olihb.com/2011/01/23/map-of-scientific-collaboration-between-researchers/

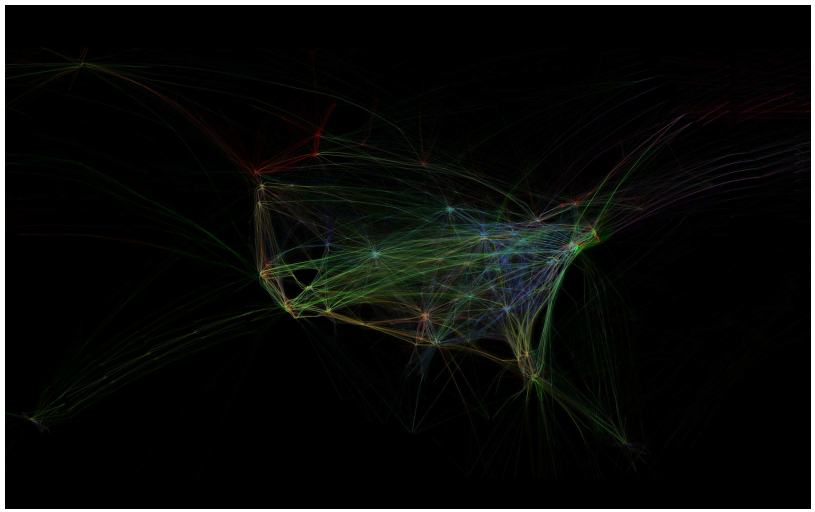
Data maps (cont.)



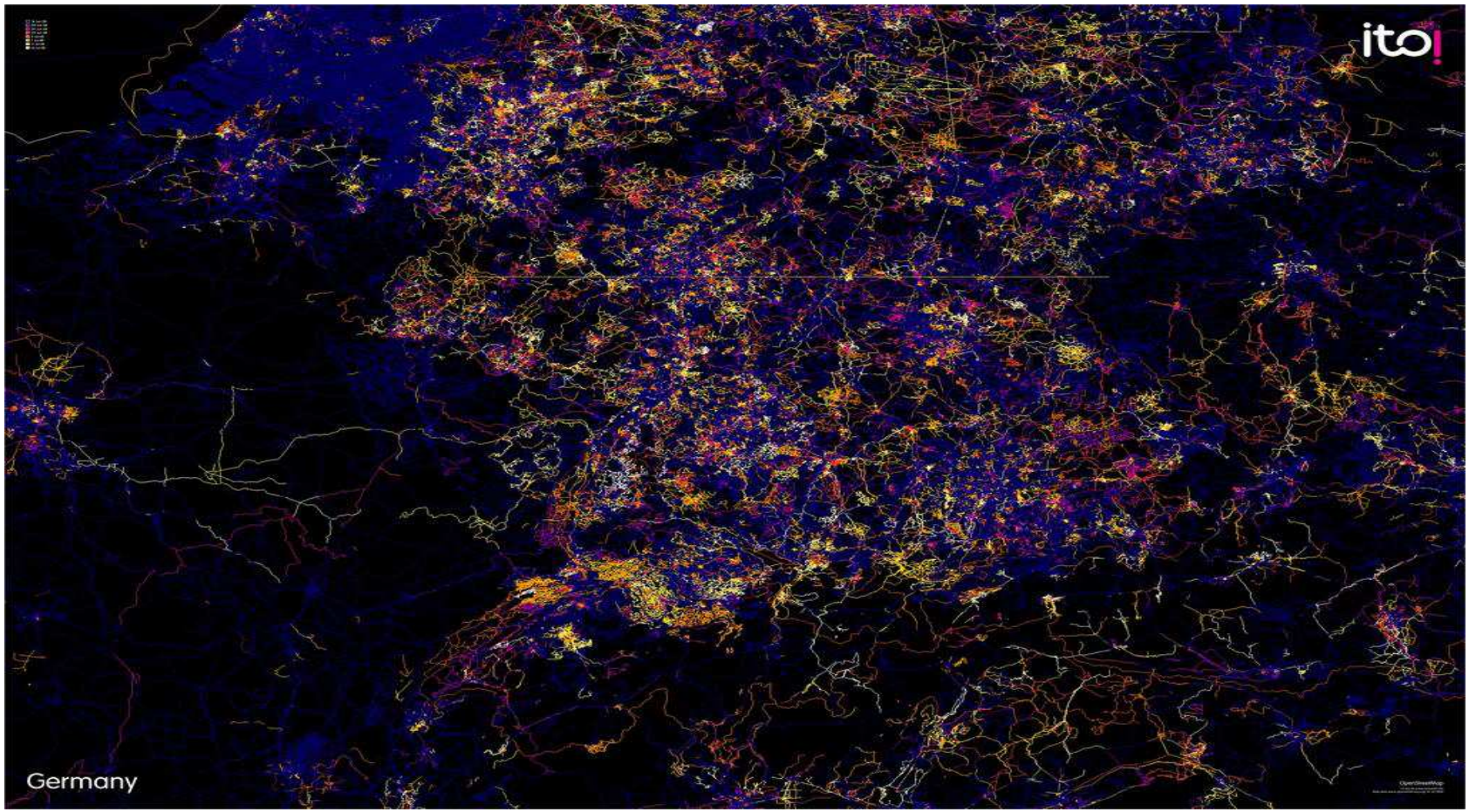
Nathan Yau: Unemployment in the United States, 2004 to present (2009)
http://projects.flowingdata.com/america/unemployment/

Data maps (cont.)





Aaron Koblin: 24-hour flight patterns over the US (2009)
http://www.aaronkoblin.com/work/flightpatterns/index.html



Peter Miller: A year of edits of OpenStreetMap (2009-)
http://www.itoworld.com/static/openstreetmap_year_of_edits.html

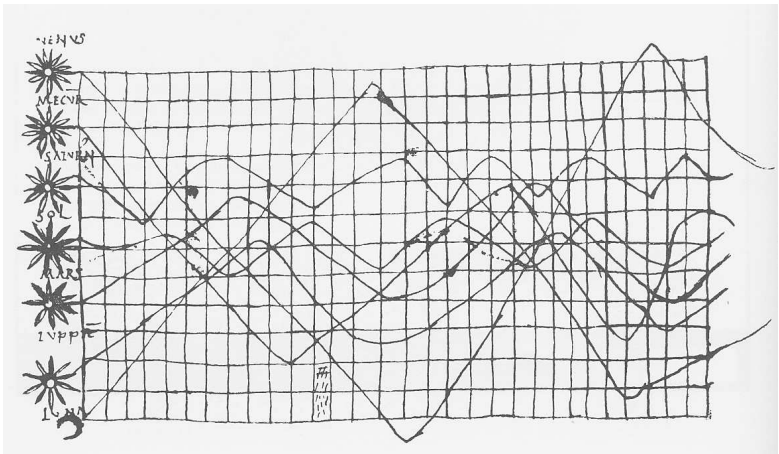
BBC: Britain from above, London taxis (2008)
http://www.bbc.co.uk/britainfromabove/stories/visualisations/taxis.shtml

Time series

The time series plot is the most frequently used form of graphic design

With one dimension marching along the regular rhythm of seconds, minutes, hours, days, weeks, months, years, centuries, or millennia

The natural ordering of the time scale gives this design a strength and efficiency of interpretation found in no other graphic arrangement



Also time series plots started appearing in the 10th or 11st century

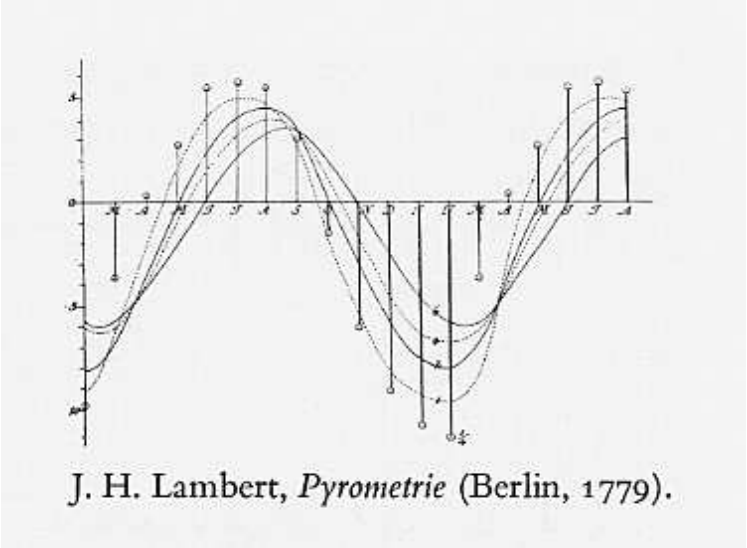
Time series (cont.)

Not until the late 1700s that time series charts began to appear in scientific writings

This drawing of Johann Lambert 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 graphic designs showing time series differ little from those of Lambert, although the data bases are far larger



J. H. Lambert, *Pyrometrie* (Berlin, 1779).

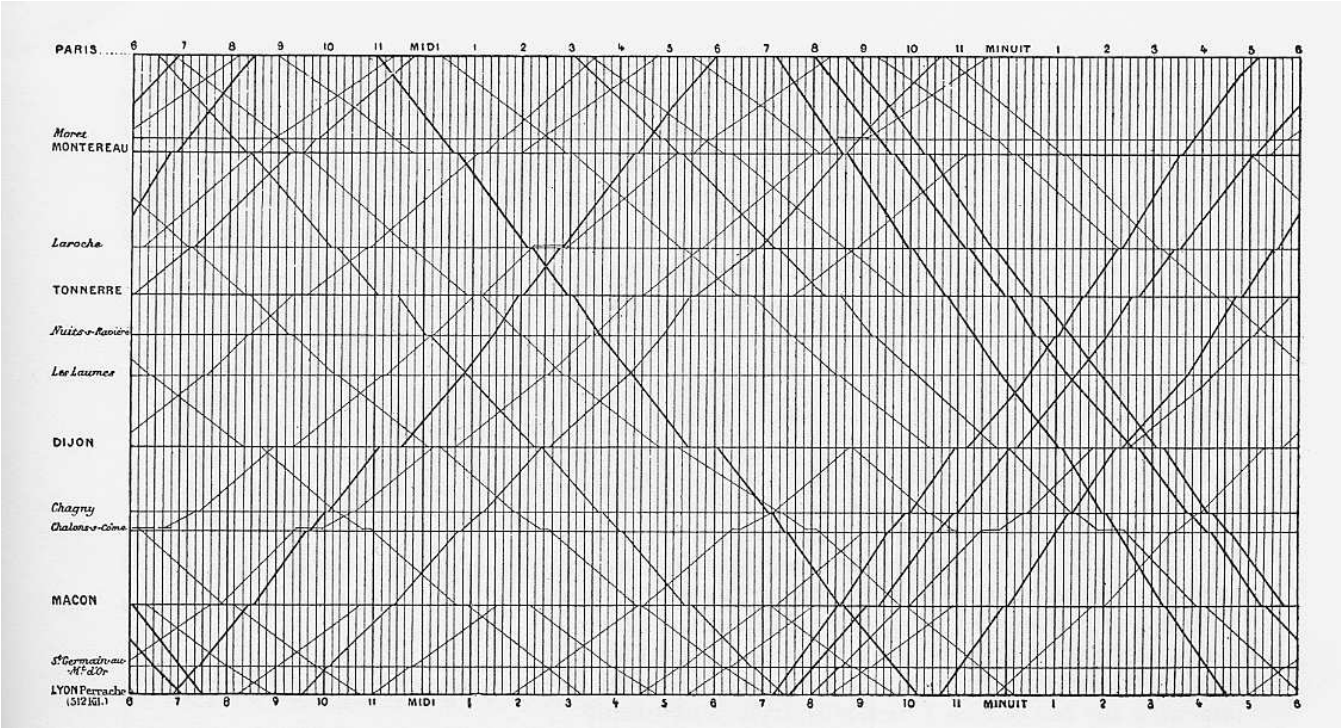
Time series displays are always at their best for big data sets with proper variability

Why waste the power of data graphics on simple linear changes, which can usually be better summarized in one or two numbers?

Time series (cont.)

The train schedule from Paris to Lyon

Èti enne- Jules M arey (1830-1904) was a French scientist and chronophotographer



- Arrivals and departures from a station are located along the horizontal
- Length of stop at a station is indicated by the length of the horizontal line
- Stations are separated in proportion to their actual distance apart
- The slope of the line reflects the speed of the train

Time series (cont.)

Johann Heinrich Lambert (1727-1777)

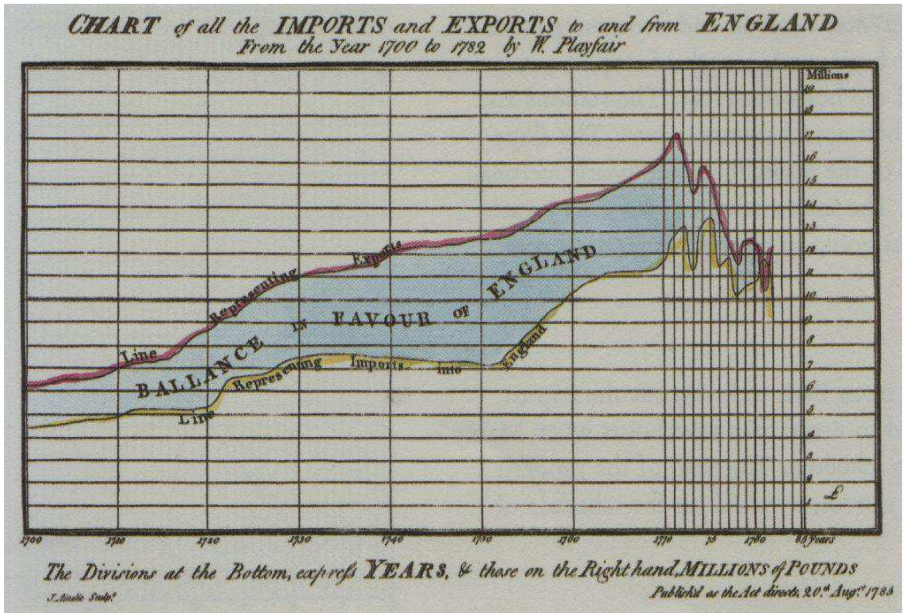
- A Swiss mathematician, physicist, astronomer, ...
- Hyperbolic geometry and properties of map projections
- The law of light absorption (Beer-Lambert Law)

William Playfair (1759-1823)

- A Scottish engineer and a political economist
- The founder of graphical methods of statistics
- A pioneer of information graphics

Time series (cont.)

Playfair published the first known time series chart using economic data (1786)



Playfair contrasted his graphical method with the tabular presentation of data

Graphics were preferable to tables because graphics showed the shape of the data in a comparative perspective

Note the graphical arithmetic showing the shifting balance of trade by the difference between the import export time series

Time series (cont.)

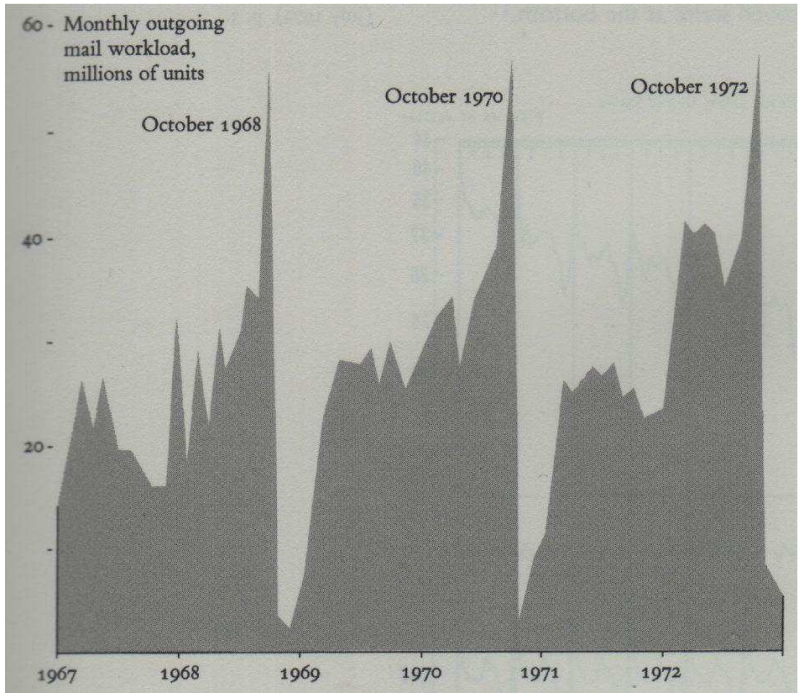
The problem with time series is that the simple passage of time is not a good explanatory variable: Descriptive chronology is not causal explanation

There are exceptions, especially when there is a clear mechanism that drives the Y-variable

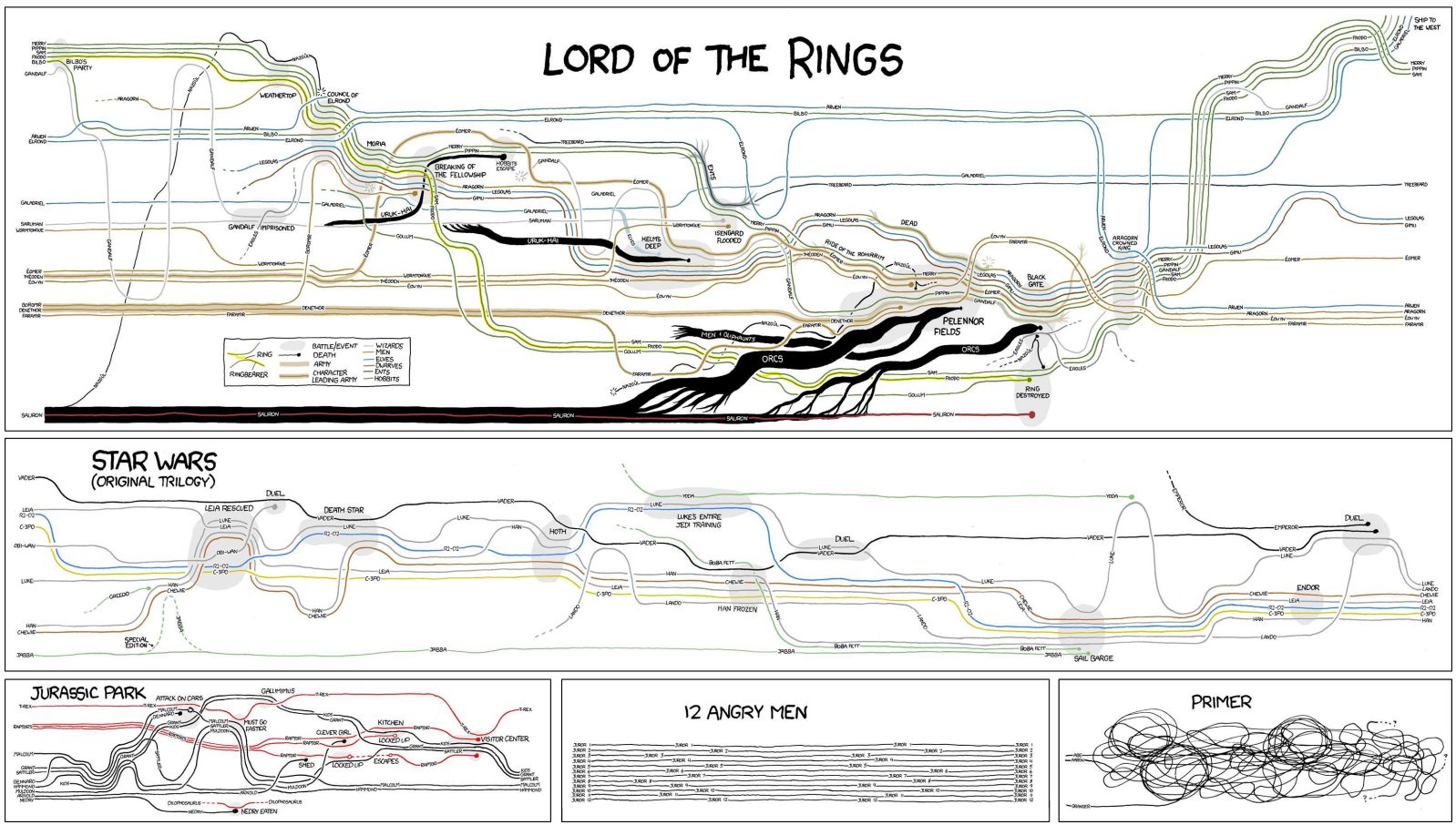
The outgoing mail of the US House of Representative peaks every two years, just before the election days

This time series does testify about causality

The graphic is worth at least 700 words



Time series (cont.)



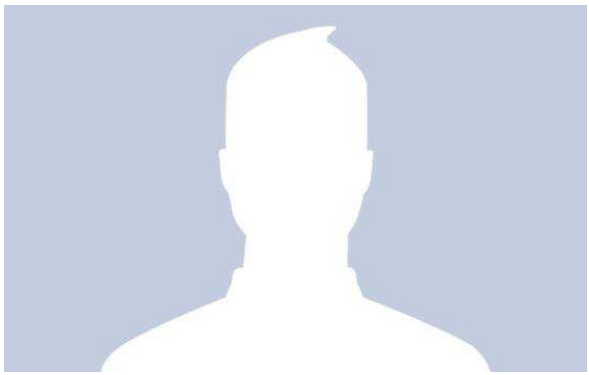
Space-time narratives

An especially effective device for enhancing the explanatory power of time series displays is to add spatial dimensions to the design of the graphic

The data are moving over space (in 2 or 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



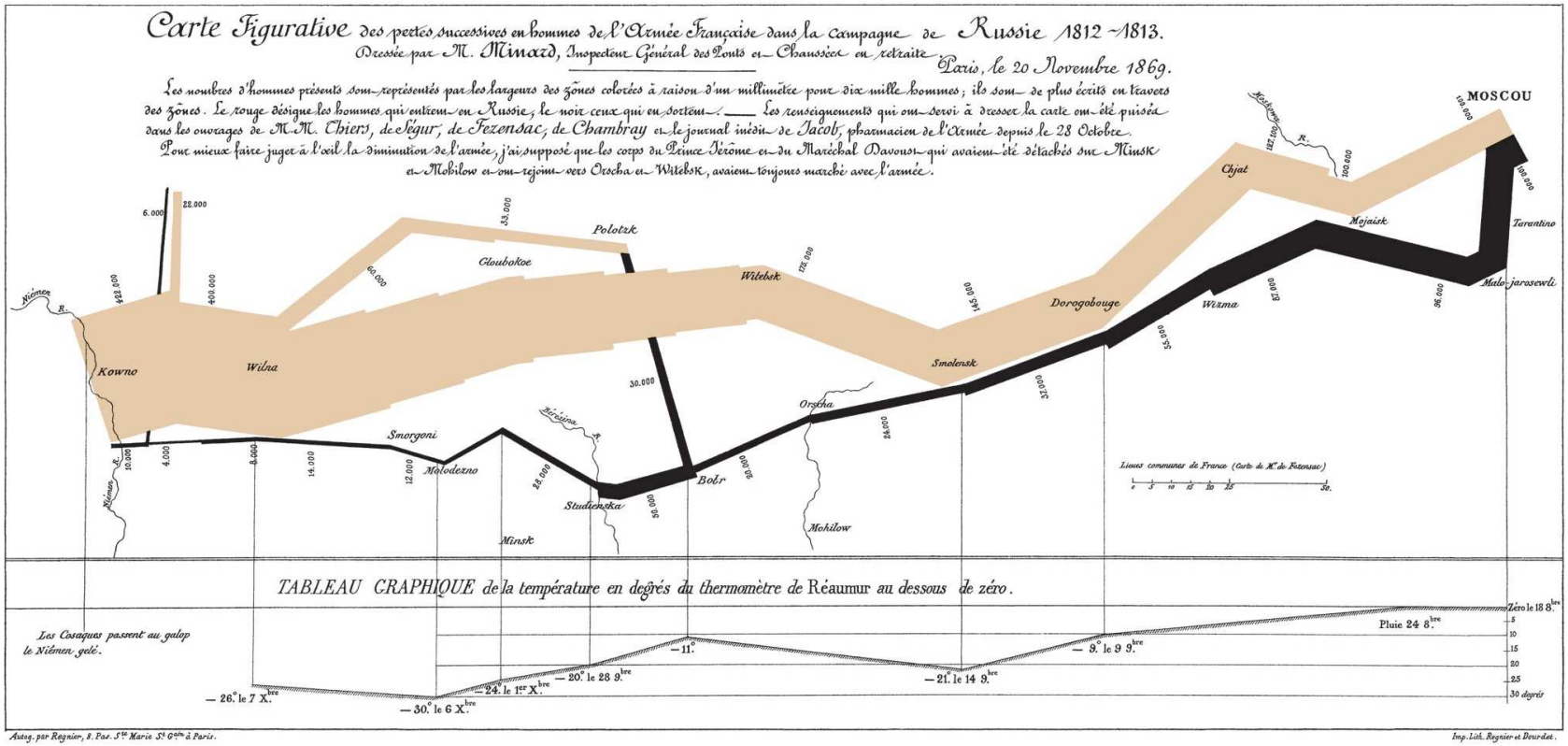
Charles Joseph Minard (1781-1870)

A French civil engineer

A graphic of the terrible fate of Napoleon's army in the Russian campaign (1869)

Space-time narratives (cont.)

Polish-Russian border, Niemen river (June 1812) Invasion →



← Retreat The sack of Moscow (September 1812)

Space-time narratives (cont.)

Minard's graphics tells a rich and yet very coherent story with its multivariate data

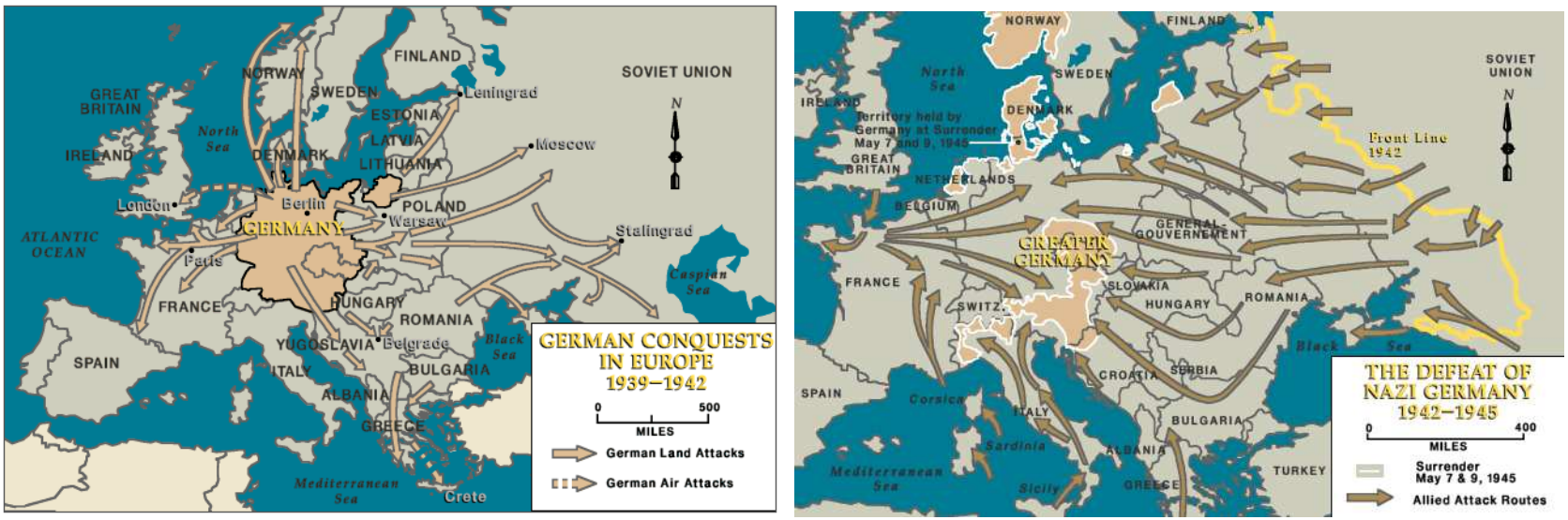
... far far far more enlightening that just a number bouncing over time

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

It may well be the best statistical graphic ever drawn

Space-time narratives (cont.)



(from the United States Holocaust Memorial Museum)

... and animated maps aren't much better!

Relational graphics

The invention of data graphics required replacing the latitude-longitude coordinates of the map with more abstracts measures not based on geographical analogy

- Moving from maps to statistical graphics was a BIG step
- Thousands of years passed before this step was taken
- Lambert and Playfair again (and others) in the 18th century

Here Playfair, who lacked mathematical skills, had a forerunner, for Lambert could think more clearly about the abstract problems of graphical design than he did

Relational graphics (cont.)

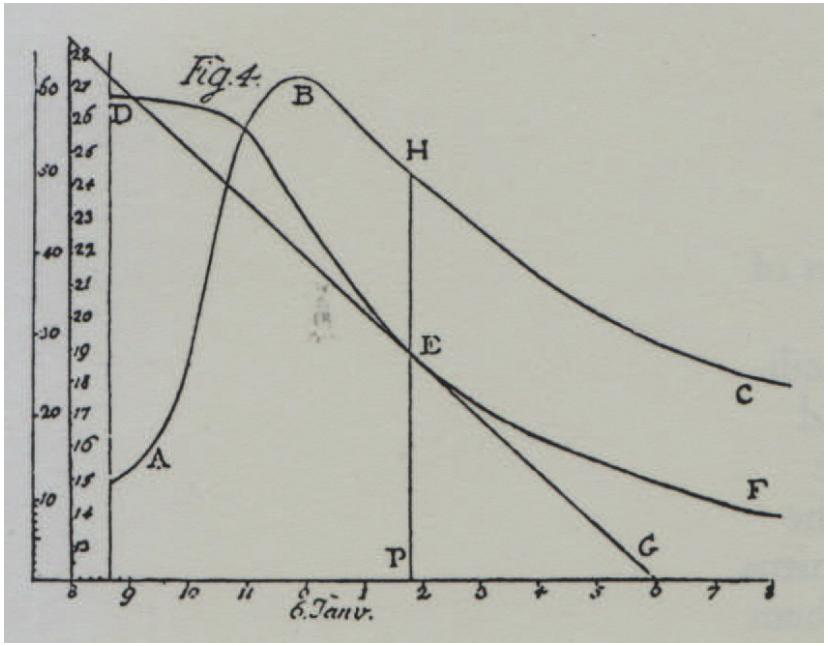
A graphical derivation of the evaporation rate of water as a function of temperature

Lambert's analysis begins with two time-series

DEF, the decreasing height of water in a capillary tube

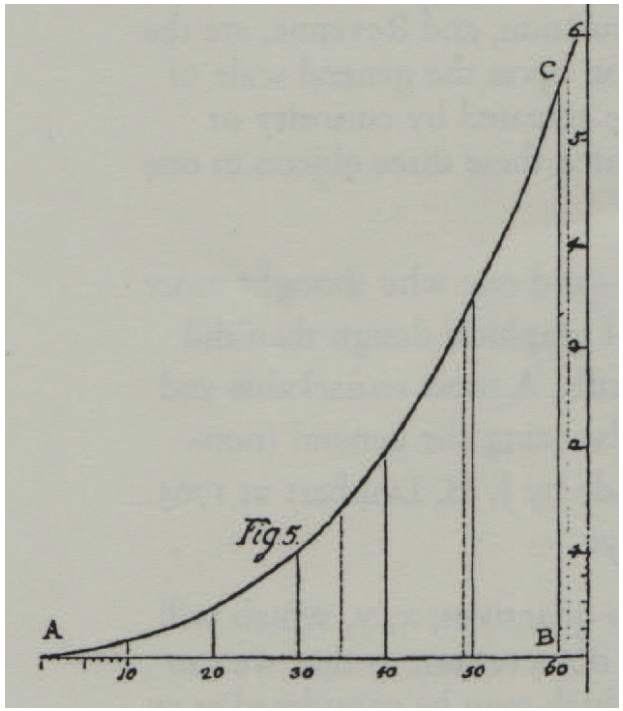
ABC, the corresponding behavior of temperature

The slope of the curve DEF is then taken (note the tangent DEG) at a number of places, yielding the rate of evaporation



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

Relational graphics (cont.)



To complete the graphical calculation, the measured rate is plotted against the corresponding temperature in this relational graphics

Relational graphics (cont.)

Lambert in Beyträge zum Gerbrauche der Mathematik und deren Anwendung (1765)

> *We have in general two variable quantities X and Y, which will be collated to one other by observation, so that we can determine for each value of X, which may be considered as an abscissa, the corresponding ordinate Y*
>
> *Were the experiments or observations completely accurate, these coordinates would give a number of points through which a straight or a curved line should be drawn*
>
> *But as this is not so, the line deviates to a greater or lesser extent from the observational points. It must therefore be drawn in such a way that it comes as near as possible to its true position and goes, as it were, through the middle of the given points*

Graphical excellence, again

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, if you like, it is what gives to the viewer

- the greatest number of ideas
- in the shortest time
- with the least ink
- in the smallest space

Graphical integrity

The first word that comes to mind when they think about statistical charts is LIE

- It is very true that some graphics do distort the underlying data
- No doubt that that makes it hard for the viewer to learn the truth

Data graphics are not different from words

- There is no reason to believe that graphics are especially vulnerable
- Most of us have good graphical lie detectors that help us to see the truth

Graphical integrity (cont.)

Much of the 20th century thinking about graphics has focused on the question of how charts might fool a viewer

- The use of graphics for serious data analysis was largely ignored

At the core of the preoccupation was the assumption that data graphics were mainly devices for showing the obvious to the ignorant, which led to two fruitless paths

- The graphics had to be alive, communicatively dynamic, overdecorated and exaggerated (otherwise, the dullards would fall asleep)
- The main task of graphical analysis was to detect and denounce deception (because the dullards could not protect themselves)

Graphical integrity (cont.)

It was only in the 1960s that John Tukey (1915-2000) started making statistical charts respectable, putting an end to the view that graphics were only for decorating numbers

American mathematician and a world-class data analyst

New designs and their effective use in the exploration of complex data

A forerunner of interactive graphics

Not a word about deception, just beautiful graphics as instruments for reasoning

Graphical integrity (cont.)

False graphics are still with us, deception must always be confronted and demolished

Lie detection is no longer at the fore front of research

Excellence begins with telling the truth about the data

Distortion in data graphics

A graphic does not distort if the visual and numerical representations are consistent

What then is the visual representation of the data?

As physically measured on the surface of the graphics?

or, as perceived visual effect?

How do we know that the visual image really represents the underlying numbers?

Distortion in data graphics (cont.)

Many experiments on visual perception of graphics have been conducted

Having people look at lines of varying length, circles of different areas and then recording their assessments of the numerical quantities

Experiments revealed *very* approximated power laws relating the numerical measures to the perceived measure

E.G., the perceived area of a circle grows more slowly than the actual measured area

$$\text{Perceived area} = (\text{Actual area})^{\alpha}, \quad \text{with} \quad \alpha = 0.8 \pm 0.3$$

However, different persons see the same ares somewhat differently

Perceptions change with experience

Perceptions are context-based

Distortion in data graphics (cont.)

The best we can hope for is some uniformity in graphics (if not in the perceivers) and some assurance that the perceiver has some fair chance to get the numbers right

Two principles lead towards those goals and thus try to enhance Graphical integrity

1. The representation of numbers, as physically measured on the surface of the graphics itself, should be directly proportional to the numerical quantities represented
2. Clear, detailed and thorough labeling should be used to defeat graphical distortion and ambiguity (write out explanations and label important events in the data)

Distortion in data graphics (cont.)

Violations of the first principle are one form of graphic misinterpretation

It is measured by the Lie factor, LF

$$\text{Lie factor} = \frac{\text{Size of effect shown in graphics}}{\text{Size of effect in data}}$$

If the Lie factor equals one, then the graphic may be doing a reasonable job

- Accurate representation of the underlying data

If the Lie factor is greater than 1.05 or smaller than 0.95, then graphics distort

- Beyond the unavoidable accuracies in plotting

The logarithm of the LF is usually taken to compare errors

Overstating (log LF $>$ 0)

Understating (log LF $<$ 0)

Overstating is the most common distortion, with Lie factors of two to five

Distortion in data graphics (cont.)

An extreme example in 1978: A series of fuel economy standards to be met by automobile manufacturers

From 18 mpg (in 1978)

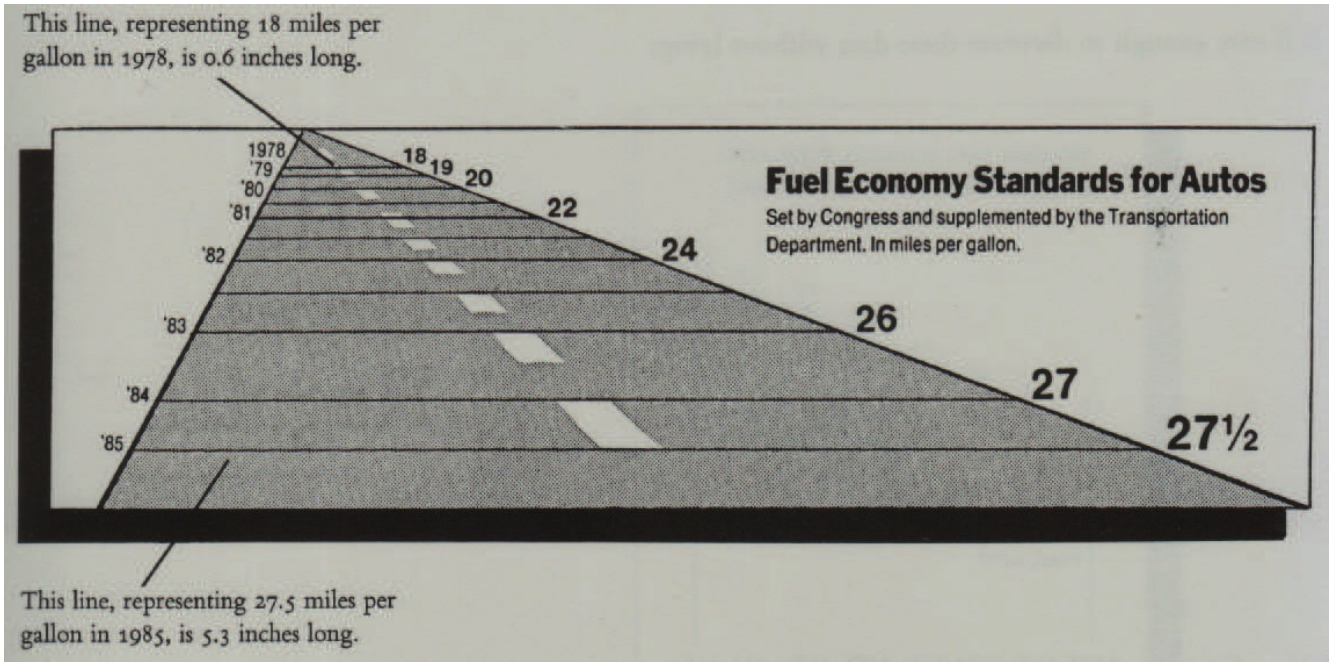
To 27.5 mpg (by 1985)

These standards and the dates for their attainment were plotted

Lie Factor = 14.8

53% increase in numb ers

783% increase in length



Distortion in data graphics (cont.)

Other peculiarities of the display, ...

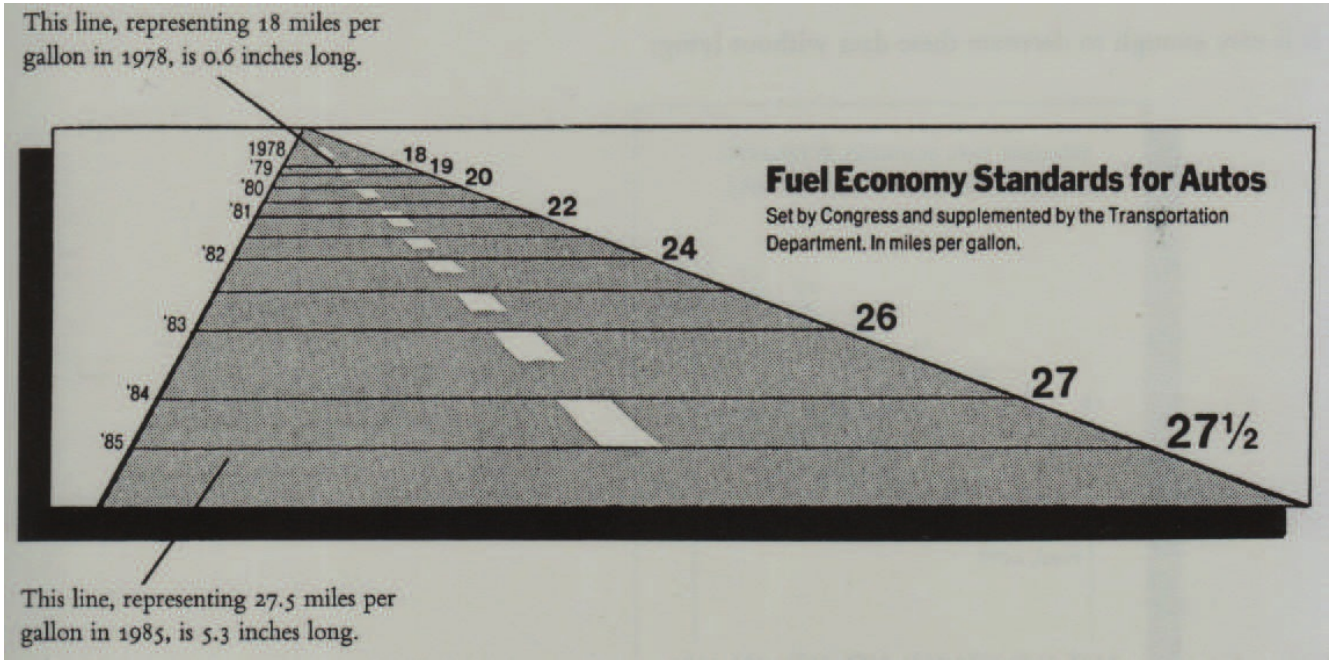
In most roads, the future is in front of us, toward the horizon

The dates remain of constant size, as we move toward the horizon

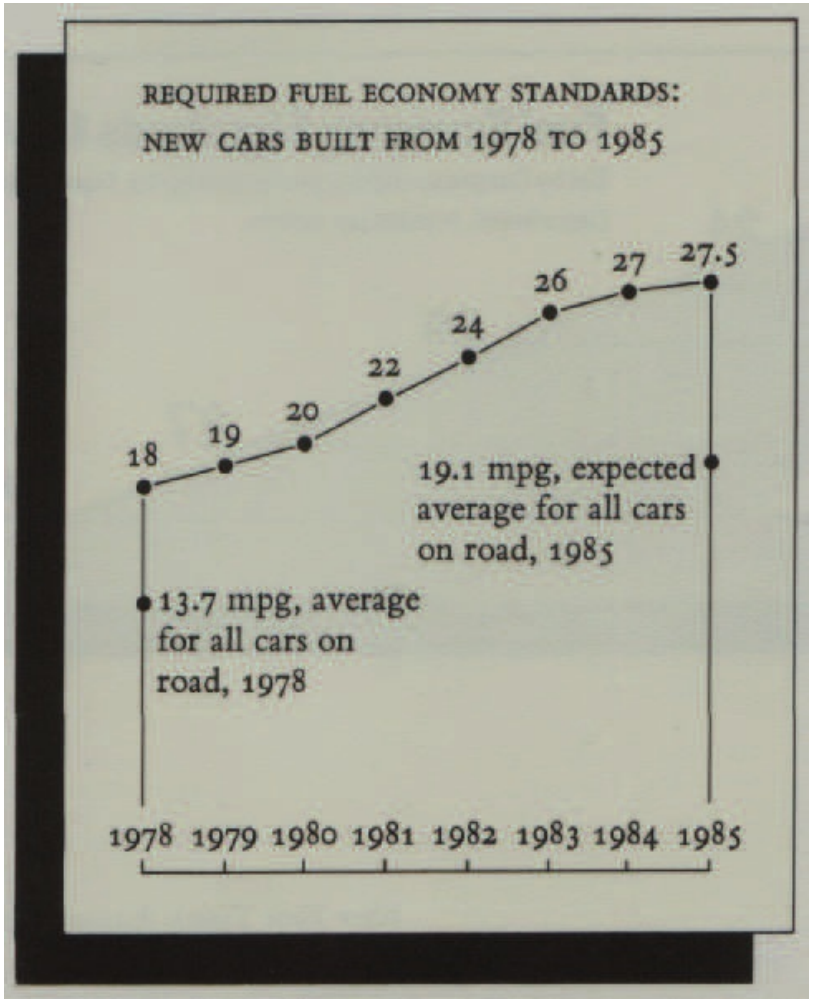
The numbers (as well as the width of the road) shrink because of two simultaneous effects

changes in values

perspective



Distortion in data graphics (cont.)



The non-lying version, with proper context

New cars standards compared with average cars on the road

Notice the revealed dynamics of fuel economy (slow startup, fast growth and final stabilization)

The New York Times (1978)

Design and data variation

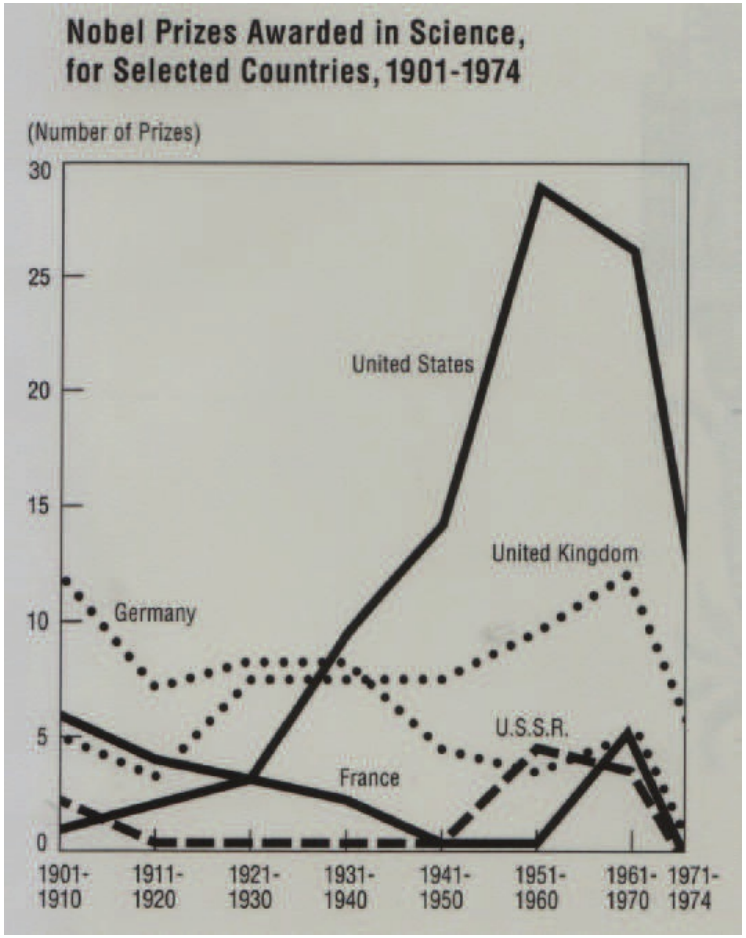
Each part of a graphic generates visual expectations about its other parts

- These expectations determine, very often, what the eyes actually see
- An incorrect extrapolation of visual expectations, generated at one place on the graphics, leads to deceptive results in other places

A typical example is a scale moving at regular intervals, ... it is also expected to move to the very end in a coherent fashion

Let’s see what happens with the muddling or trickery of non-uniform changes ...

Design and data variation (cont.)



Here an irregular scale is used to concoct a pseudo-decline

- The first seven intervals are 10-yr long
- The rightmost is only 4-yr long

As a result, a conspicuous feature of the chart is an apparent fall of all the curves

American National Science Foundation (1976)

A sole effect of design variation and nothing but a big lie (both in extrapolation and reality), as the US curve turned up sharply in the 1971-1980 decade

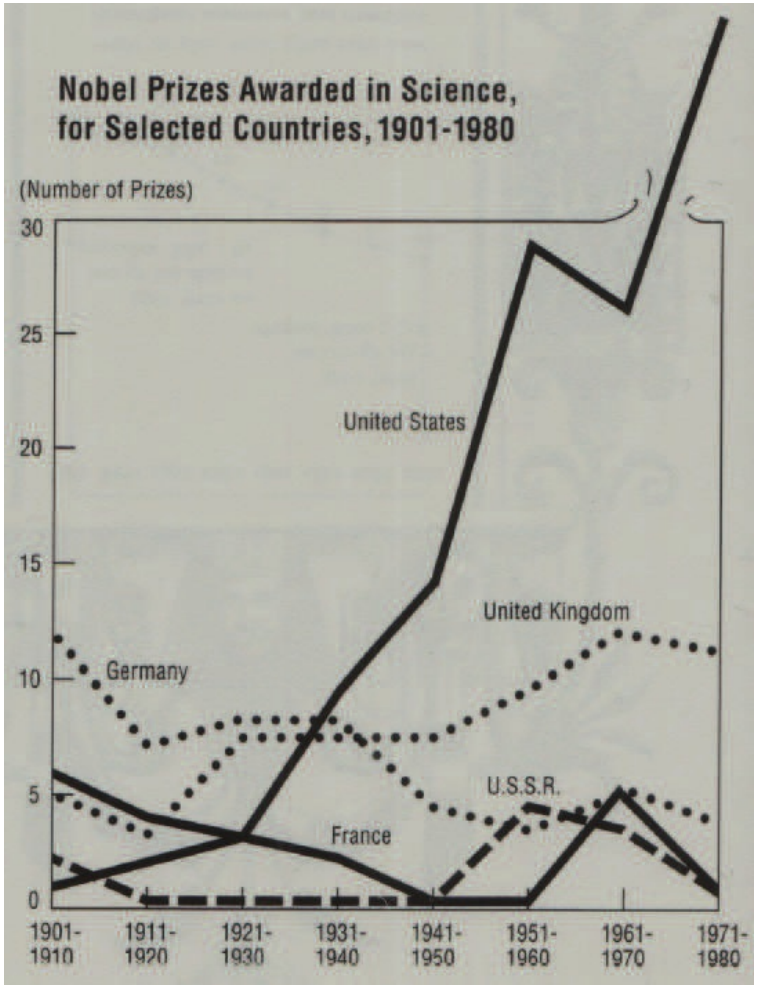
Design and data variation (cont.)

A correction with the actual data

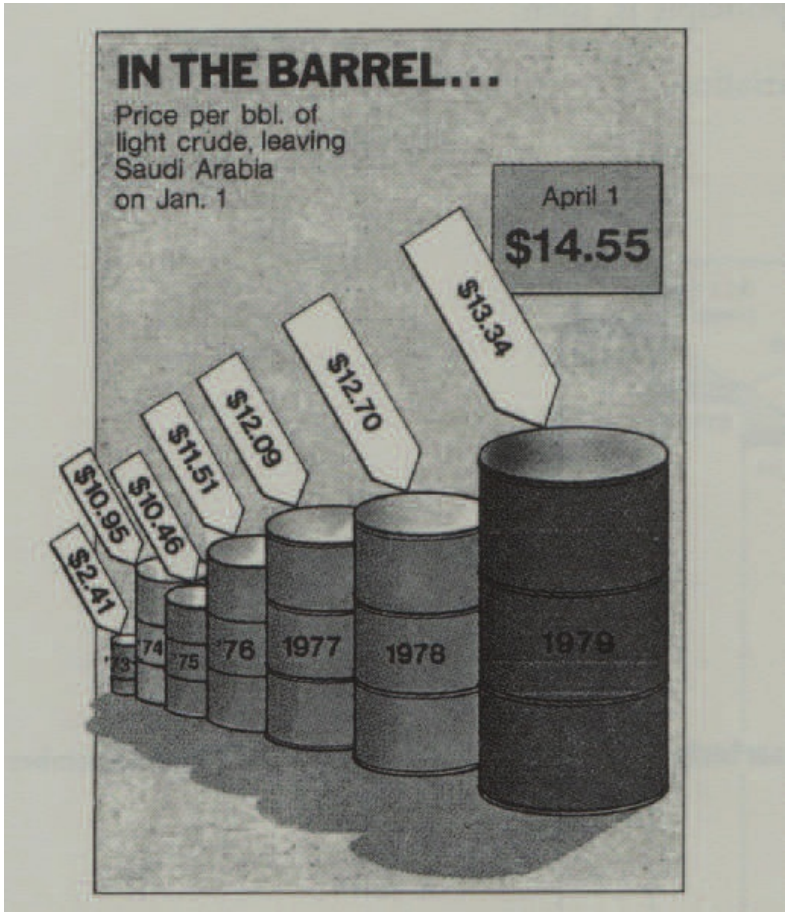
The confounding of design variation with actual data variation generated ambiguity and deception

The eye can mix up changes in the design with changes in the data

Show data variation, not design variation!



Design and data variation (cont.)



Another example where design variation infected a graphic

An increase of 454% is depicted as an increase of 4280%

Lie Factor?

The viewer gets mixed up by the fact that area (2D) is used to show 1D data, to confuse data variation with design variation

The Time magazine (1979)

Visual area and numerical measure

The problem is that many published works using areas to show magnitudes make the elementary mistake to change both dimensions simultaneously

Again, there are considerable ambiguities in how people perceive a 2D surface and then convert that perception into a number

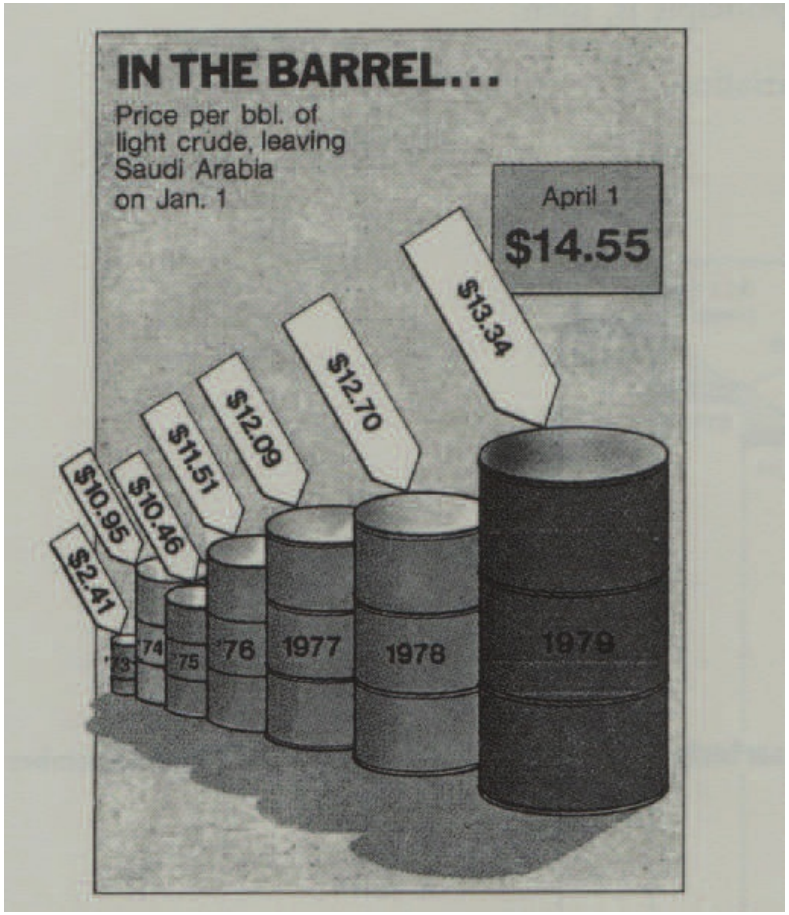
Changes in physical area on the surface of a graphic do not reliably produce appropriately proportional changes in the perceived areas

Visual area and numerical measure (cont.)

When areas are tricked up into volumes (3D), then the problem is all the worse

Try and take the barrel metaphor more seriously and, assume that the Volume of the barrel , not its Area, represents the actual price changes

The 1973-1979 increase (454%) is shown as 27K%, for a Lie Factor of 59.4



Visual area and numerical measure (cont.)

The use of 2 (or 3) varying dimensions to show 1D data is weak and inefficient

It generates errors in design and ambiguity in perception

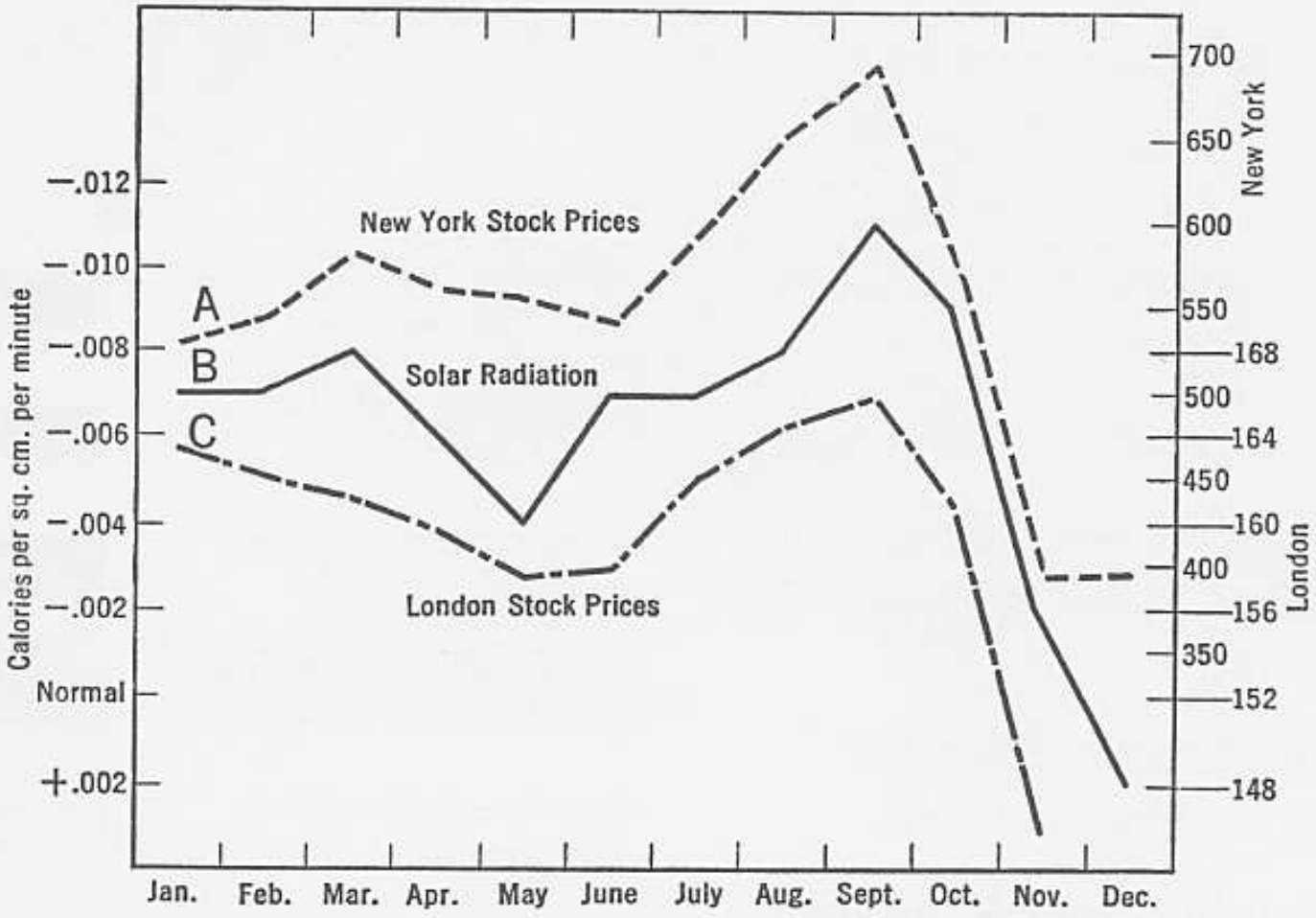
The use of this technique causes so many problems that it should be avoided

Graphical integrity, again

Graphical integrity is more likely to result if

- Numbers measured on the graphics are directly proportional to the numerical quantities represented
- Clear and detailed labeling is used to defeat distortion and ambiguity
- Data variation is shown, instead of design variation
- Deflated and standardized units are used
- The number of information carrying dimensions does not exceed the number of dimensions in the data

Graphics must not quote data out of context



SOLAR RADIATION AND STOCK PRICES

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

..., because a silly theory means a silly graphic, always!