GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 18: DATA VISUALIZATION

Data Visualization

Steve Marschner
Cornell CS 3220

unless noted, images are from
Tufte, *The Visual Display of Quantitative Information*
(these slides also indebted to Pat Hanrahan's slides for CS448B at Stanford)

Data

A lot of 3220 is about data

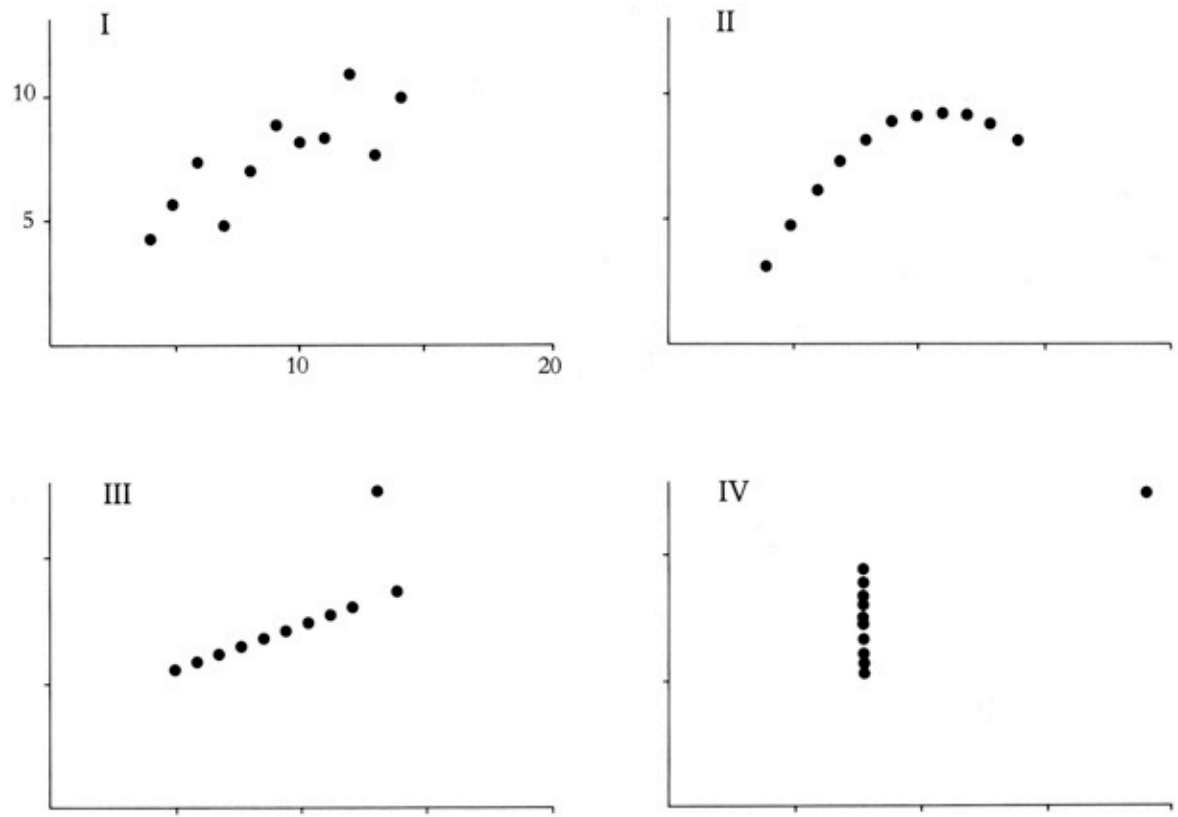
- input to fitting problems
- output of simulations

Understanding all but the simplest is not easy

- tables of numbers give little insight
- appropriate pictures are invaluable!

I		II		III		IV	
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

$N = 11$
mean of X's $= 9.0$
mean of Y's $= 7.5$
equation of regression line: $Y = 3 + 0.5X$
standard error of estimate of slope $= 0.118$
$t = 4.24$
sum of squares $X - \overline{X} = 110.0$
regression sum of squares $= 27.50$
residual sum of squares of $Y = 13.75$
correlation coefficient $= .82$
$r^2 = .67$


I
10
5
10
20
II
III
IV

Purposes of visualization

Organize and display data (for yourself)

- provide data in a form our brains & visual systems are able to use
- making pictures of data helps you understand it
- designing visualizations forces you to organize the data
- **a key part of the intellectual and creative process**

Present data (for others)

- data in support of arguments (scientific, policy, …)
- data for making decisions (funding, operational, …)
- good presentation of data is key to any good presentation of complex technical material
- **a part of informative & persuasive communication**

John C. Snow
(1854)

Purposes of visualization

Organize and display data (for yourself)

- provide data in a form our brains & visual systems are able to use
- making pictures of data helps you understand it
- designing visualizations forces you to organize the data
- **a key part of the intellectual and creative process**

Present data (for others)

- data in support of arguments (scientific, policy, …)
- data for making decisions (funding, operational, …)
- good presentation of data is key to any good presentation of complex technical material
- **a part of informative & persuasive communication**

HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

	SRM No.	Cross Sectional View: Erosion Depth (in.)	Cross Sectional View: Perimeter Affected (deg)	Cross Sectional View: Nominal Dia. (in.)	Top View: Length Of Max Erosion (in.)	Top View: Total Heat Affected Length (in.)	Clocking Location (deg)
61A LH Center Field**	22A	None	None	0.280	None	None	36°--66°
61A LH ~~CENTER~~ FIELD**	22A	NONE	NONE	0.280	NONE	NONE	338°-18°
51C LH Forward Field**	15A	0.010	154.0	0.280	4.25	5.25	163
51C RH Center Field (prim)***	15B	0.038	130.0	0.280	12.50	58.75	354
51C RH Center Field (sec)***	15B	None	45.0	0.280	None	29.50	354
41D RH Forward Field	13B	0.028	110.0	0.280	3.00	None	275
41C LH Aft Field*	11A	None	None	0.280	None	None	--
41B LH Forward Field	10A	0.040	217.0	0.280	3.00	14.50	351
STS-2 RH Aft Field	2B	0.053	116.0	0.280	--	--	90

*Hot gas path detected in putty. Indication of heat on O-ring, but no damage.
**Soot behind primary O-ring.
***Soot behind primary O-ring, heat affected secondary O-ring.

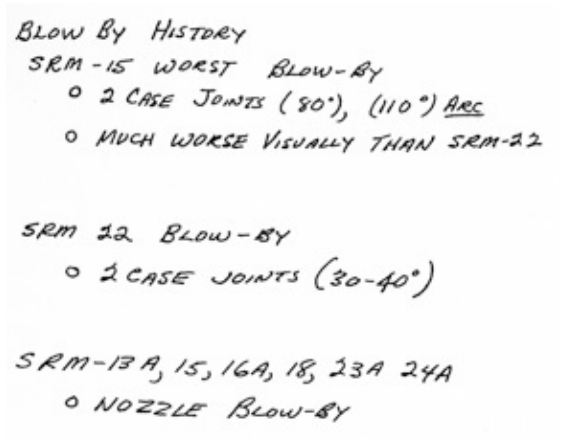
Clocking location of leak check port - 0 deg.

OTHER SRM-15 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY AND NO SOOT NEAR OR BEYOND THE PRIMARY O-RING.

SRM-22 FORWARD FIELD JOINT HAD PUTTY PATH TO PRIMARY O-RING, BUT NO O-RING EROSION AND NO SOOT BLOWBY. OTHER SRM-22 FIELD JOINTS HAD NO BLOWHOLES IN PUTTY.

[from Tufte, *Visual Explanations*]

data presented by rocket's manufacturer to argue for canceling the launch.



HISTORY OF O-RING TEMPERATURES (DEGREES - F)

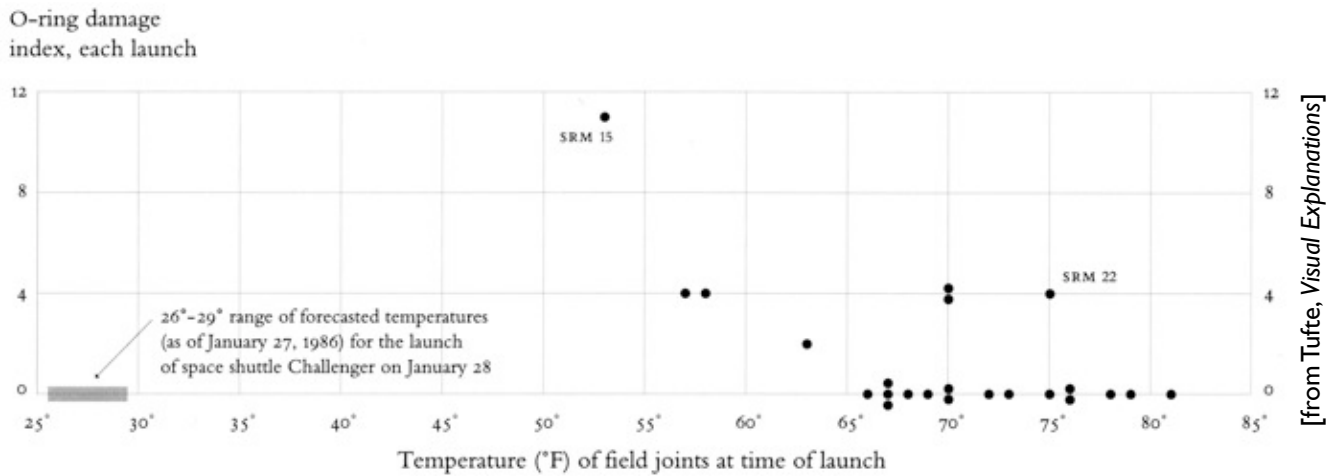
MOTOR	MBT	AMB	O-RING	WIND
DM-4	68	36	47	10 MPH
DM-2	76	45	52	10 MPH
QM-3	72.5	40	48	10 MPH
QM-4	76	48	51	10 MPH
SRM-15	52	64	53	10 MPH
SRM-22	77	78	75	10 MPH
SRM-25	55	26	29 27	10 MPH 25 MPH

[from Tufte, *Visual Explanations*]

data presented by rocket's manufacturer to argue for canceling the launch.

[NASA]

Space Shuttle mission STS-51-L, about 75 sec. after liftoff. 1986



[from Tufte, *Visual Explanations*]

Tufte's more convincing re-presentation of the same data. 1997

Data Mappings

Mapping data into a visual display

Datatypes

- programming: char, int, float, double, String, …
- scientific data has types too

Graphical information channels

- there are many ways to put the data into pictures
- good datatype-to-channel matches are important!

Datatypes

Nominal select from unorganized set (enumerated type, in C)

apples, oranges, tomatoes, …

Toyota, Ford, Subaru, …

Ordinal ordered set of values (< operator available)

January, February, March, …

Trial 1, Trial 2, Trial 3, …

12 Oak St., 125 Oak St., 129 Oak St., …

S. S. Stevens, *On the theory of scales of measurement* (1946)

Datatypes (quantitative)

Interval values are meaningful, but zero is arbitrary (+, – avail.)

- degrees Celsius
- position
- potential energy

Ratio values are meaningful, meaningful zero (×, ÷ avail.)

- degrees Kelvin
- length
- mass

S. S. Stevens, *On the theory of scales of measurement* (1946)

Graphical information channels

Spatial

- length
- position
- size (area, volume?)

Color

- value (lightness, black to white)
- saturation (colorfulness, gray to vivid)
- hue (color)
- texture (fill pattern)

Details

- shape
- orientation

Datatypes and channels

Pay attention to data semantics

Chose channel that carries the semantics well

		N	O	I	R
spatial	length				Y
	position	Y	Y	Y	
	size		Y	~	~
color	value		Y		~
	saturation		Y		
	hue	Y			
	texture	Y			
detail	shape	Y			
	orientation		Y	~	

Common types of visualizations

data maps

time series

relational plots

histograms

bar charts

polar plots

color maps

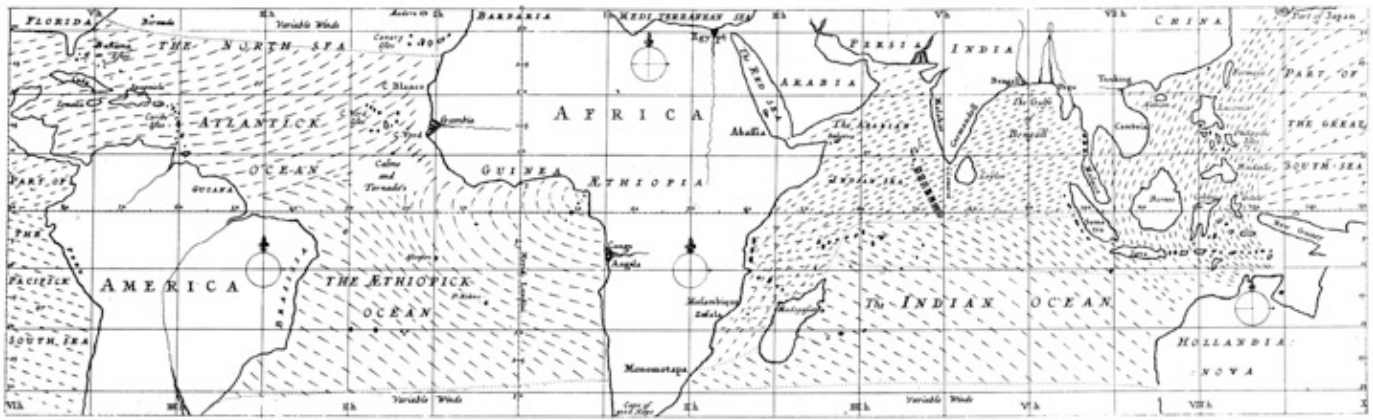
Data Maps

Position: position

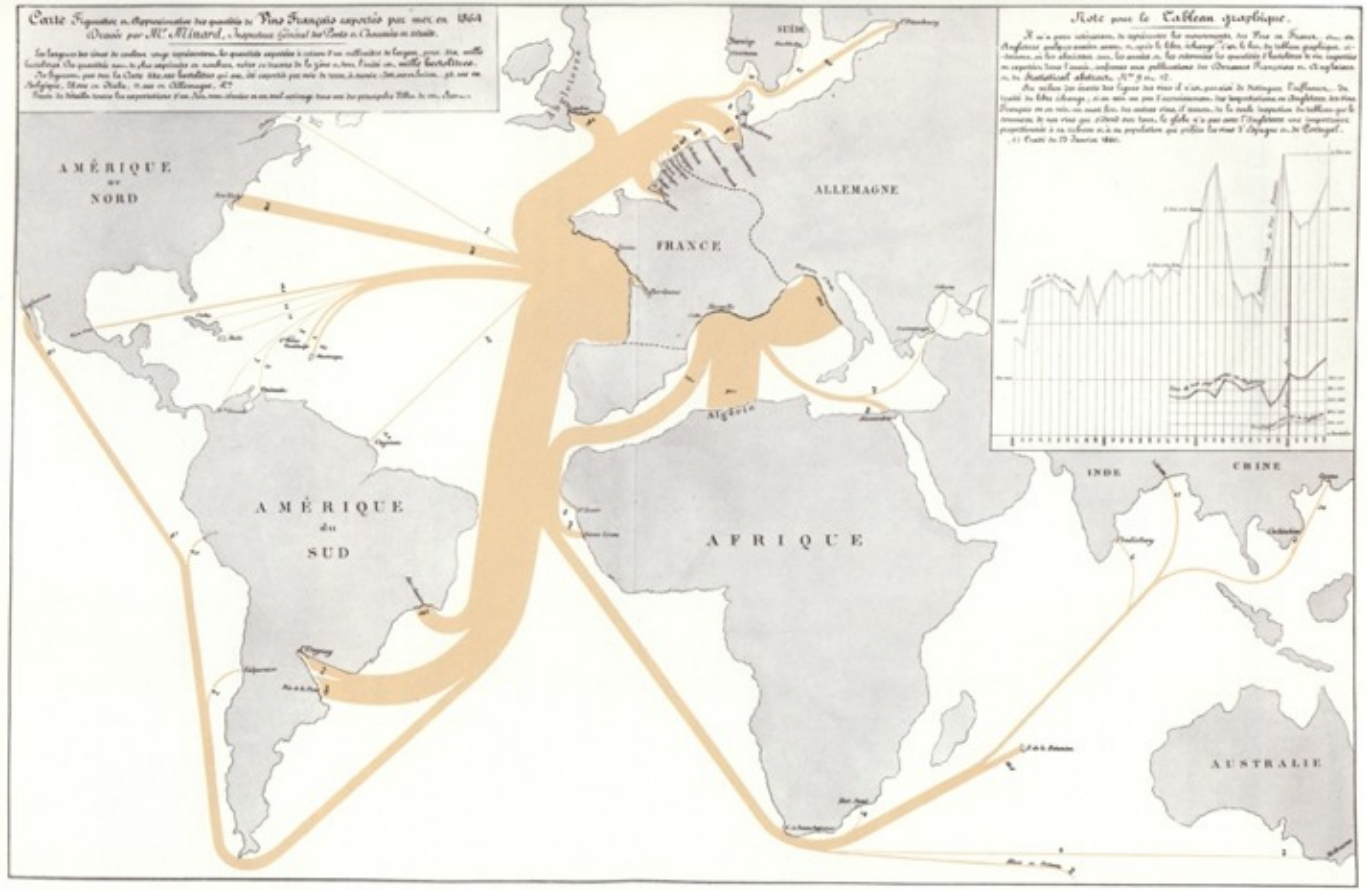
Symbols, colors: various variables (N, O, or Q)

very old form of data visualization

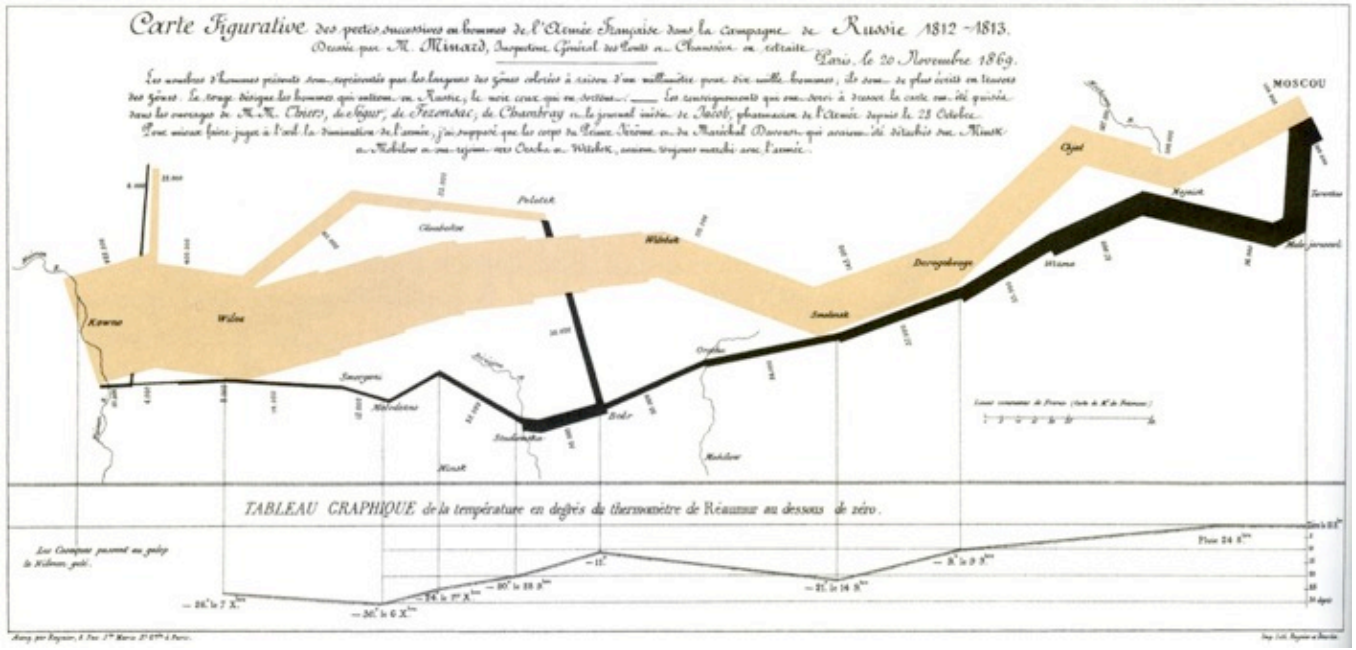
readily interpreted with little training or effort



E. Halley. Map illustrating trade winds. 1686



C. J. Minard. Map illustrating exports of French wine. 1864



J.C. Minard. Depiction of losses during French Army march to (and retreat from) Moscow, 1812–1813.

Time series

Horizontal axis: time (Interval—Position)

Vertical axis: some quantitative value (often money)

very old form of data visualization

readily interpreted with little training or effort

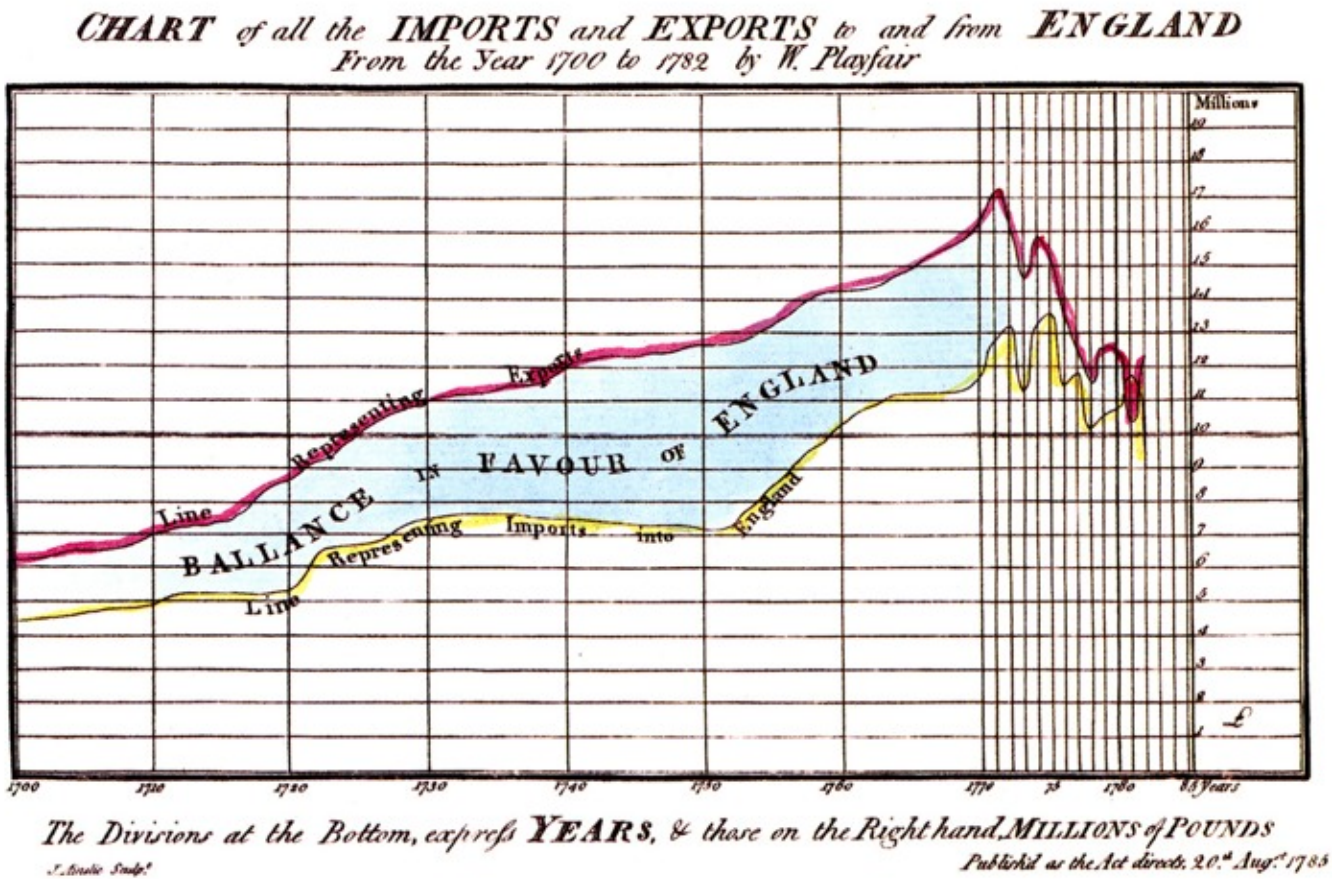
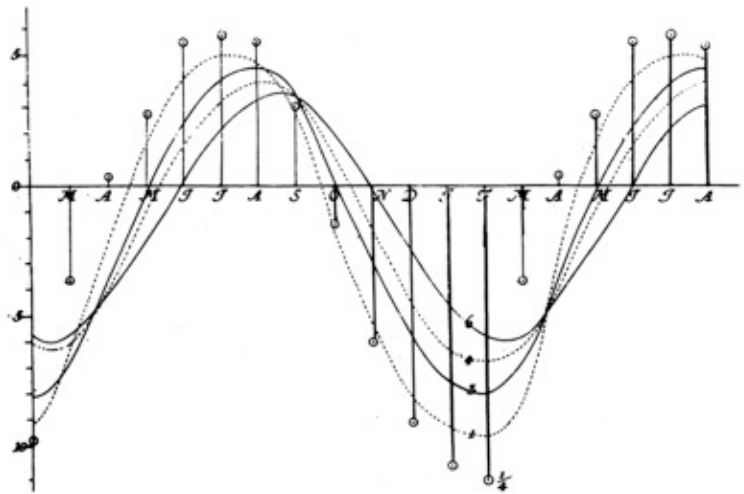
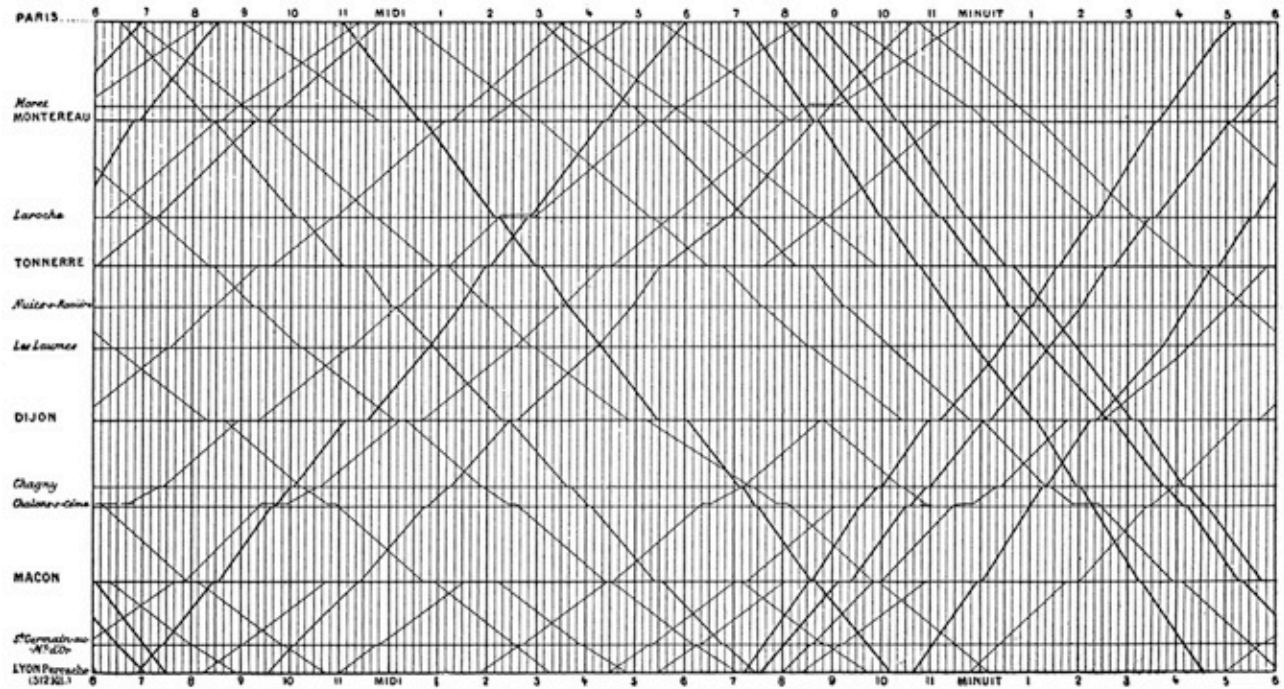


CHART of all the IMPORTS and EXPORTS to and from ENGLAND
From the Year 1700 to 1782 by W. Playfair
Millions
Line Representing Exports
BALLANCE IN FAVOUR OF ENGLAND
Line Representing Imports into England
1700
1710
1720
1730
1740
1750
1760
1770
78
1780
83 Years
The Divisions at the Bottom, express YEARS, & those on the Right hand MILLIONS of POUNDS
Published as the Act directs. 20th Augt 1785



J.H. Lambert. Soil temperature over time at various depths. 1779



E.J. Marey. Train schedule for Paris–Lyon line. 1885

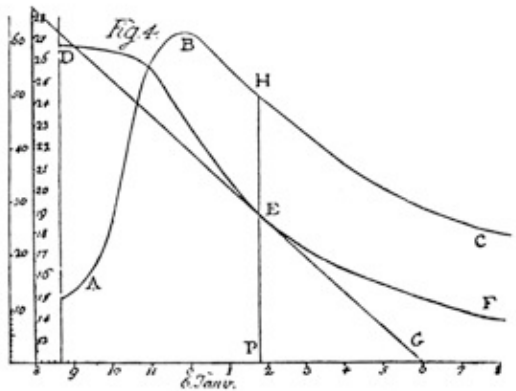
Relational plots

Horizontal axis: alleged "cause"

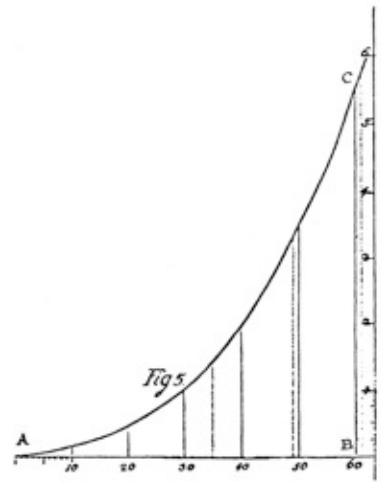
Vertical axis: alleged "effect"

very powerful tool to investigate relationships

scatter plot for unordered set of points;
connected line for ordered sequence of points
or to emphasize functional "law"

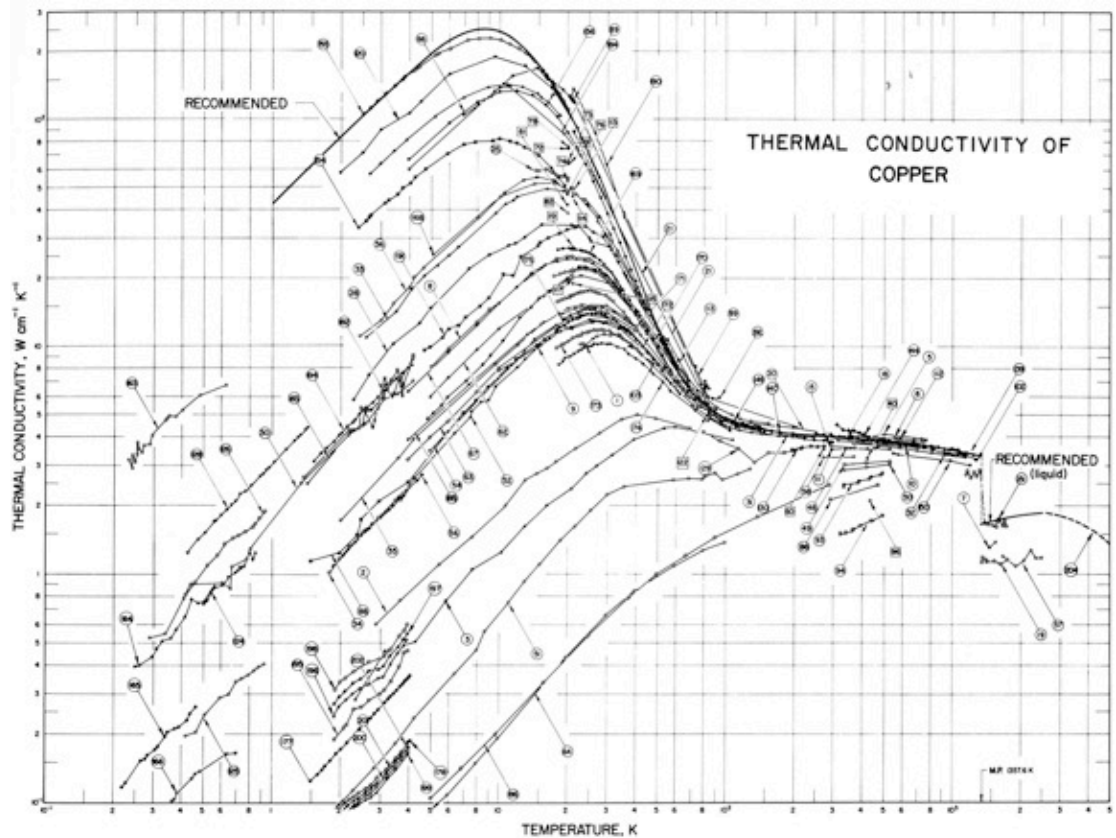


ABC: temperature over time
DEF: height of water over time

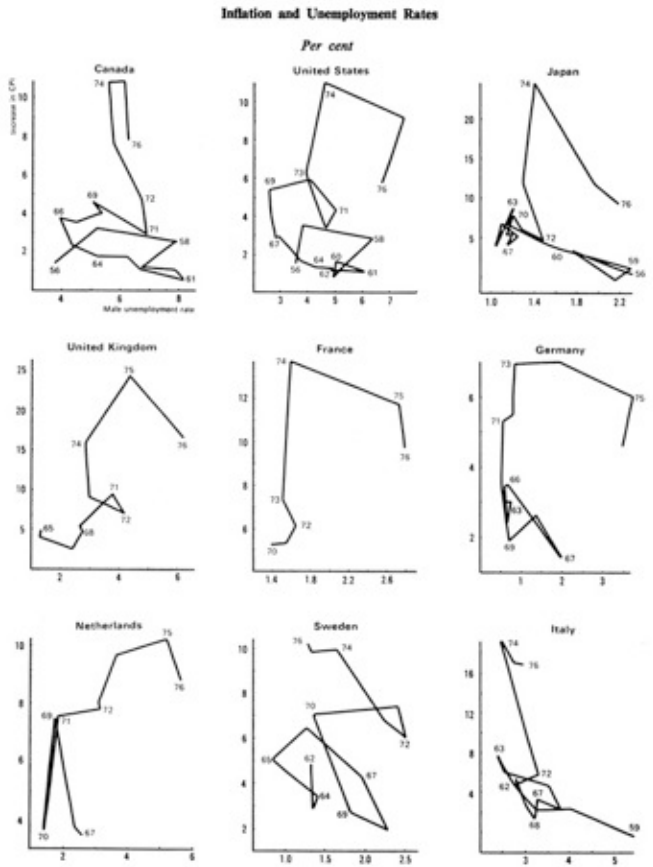


evaporation rate
vs. temperature

J.H. Lambert: influence of temperature on evaporation. 1769



C.Y. Ho et al. Review of thermal conductivity data. 1974



P. McCracken et al. Phillips curves. 1977

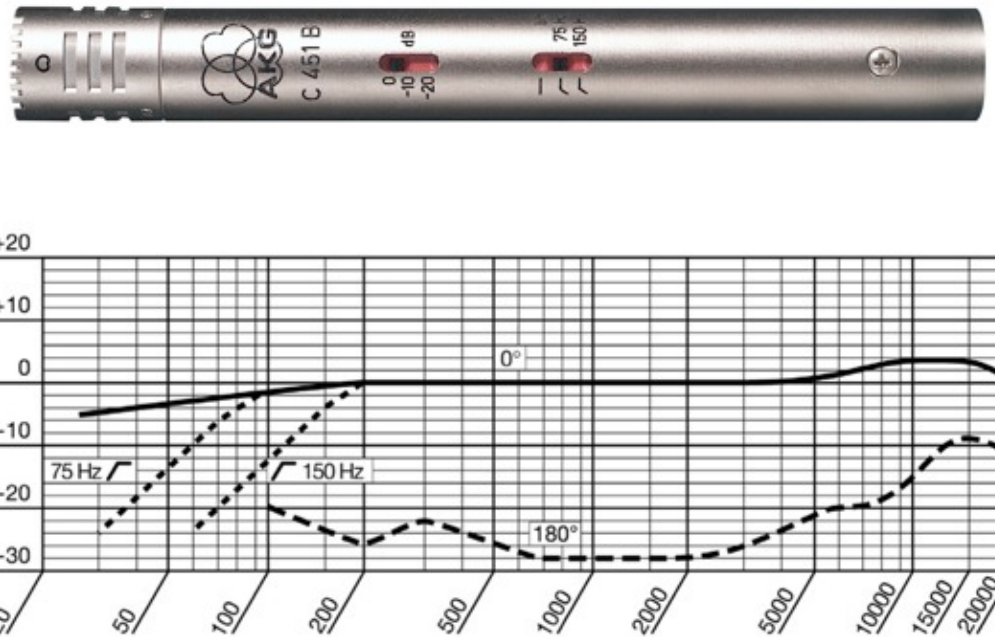
Logarithmic plots

For one or both axes, replace direct (linear) data–position mapping with logarithmic mapping

Useful for data with high dynamic range

Useful for exponential and power-law relationships

Caution: converts type from ratio to interval



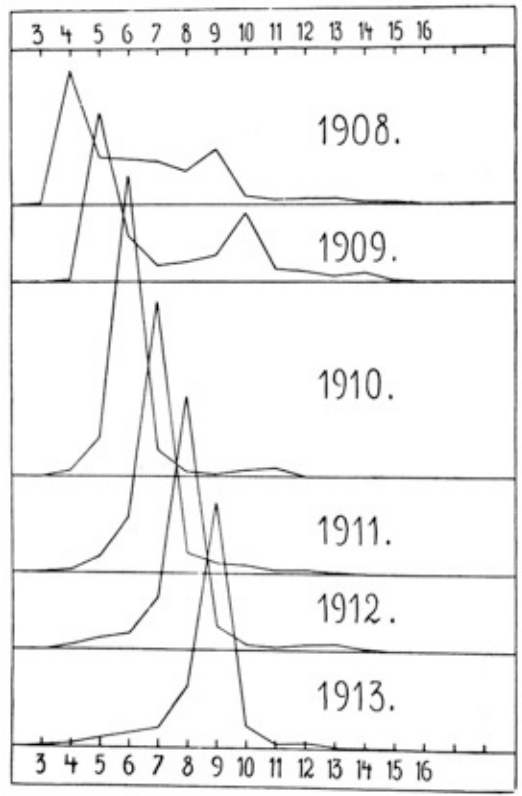
AKG Acoustics. Performance data for C451B microphone. 1973

Histograms

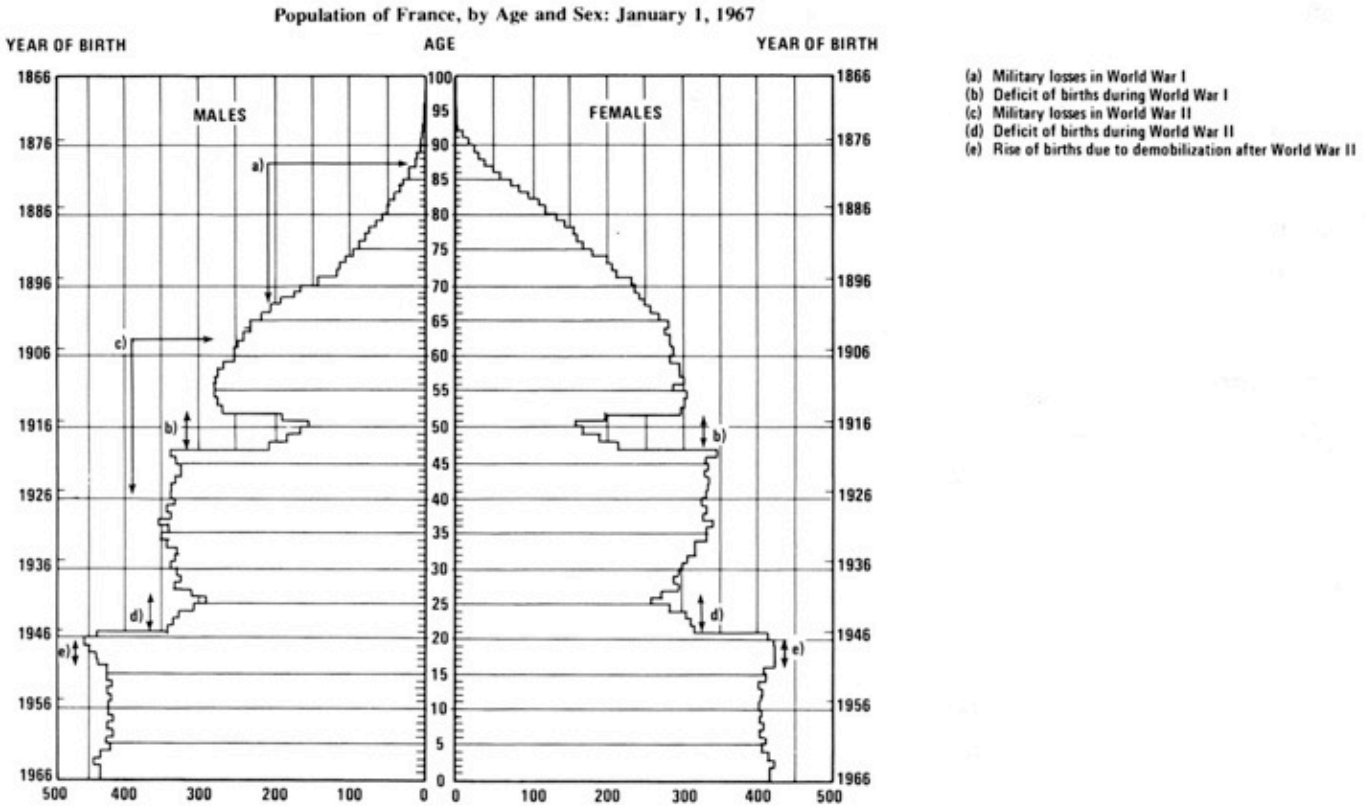
First axis (oft. horiz.): Nominal or Ordinal variable

Second axis: count of something (ratio)

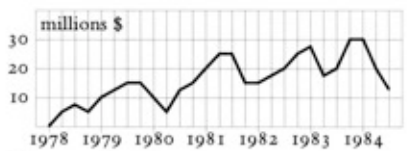
often convert Quantitative to Ordinal by binning (danger!)



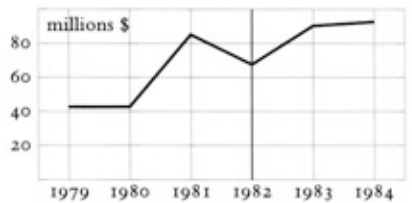
J. Hjort. Age composition of herring catches. 1914



H.S. Shyrock & J.S. Siegel. Rendering of French government population data. 1973



Above, this chart shows *quarterly* revenue data in a financial graphic for a legal case. Several dips in revenue are visible.



Aggregating the quarterly data into years, this chart above shows revenue by *fiscal year* (beginning July 1, ending June 30). Note the dip in 1982, the basis of a claim for damages.

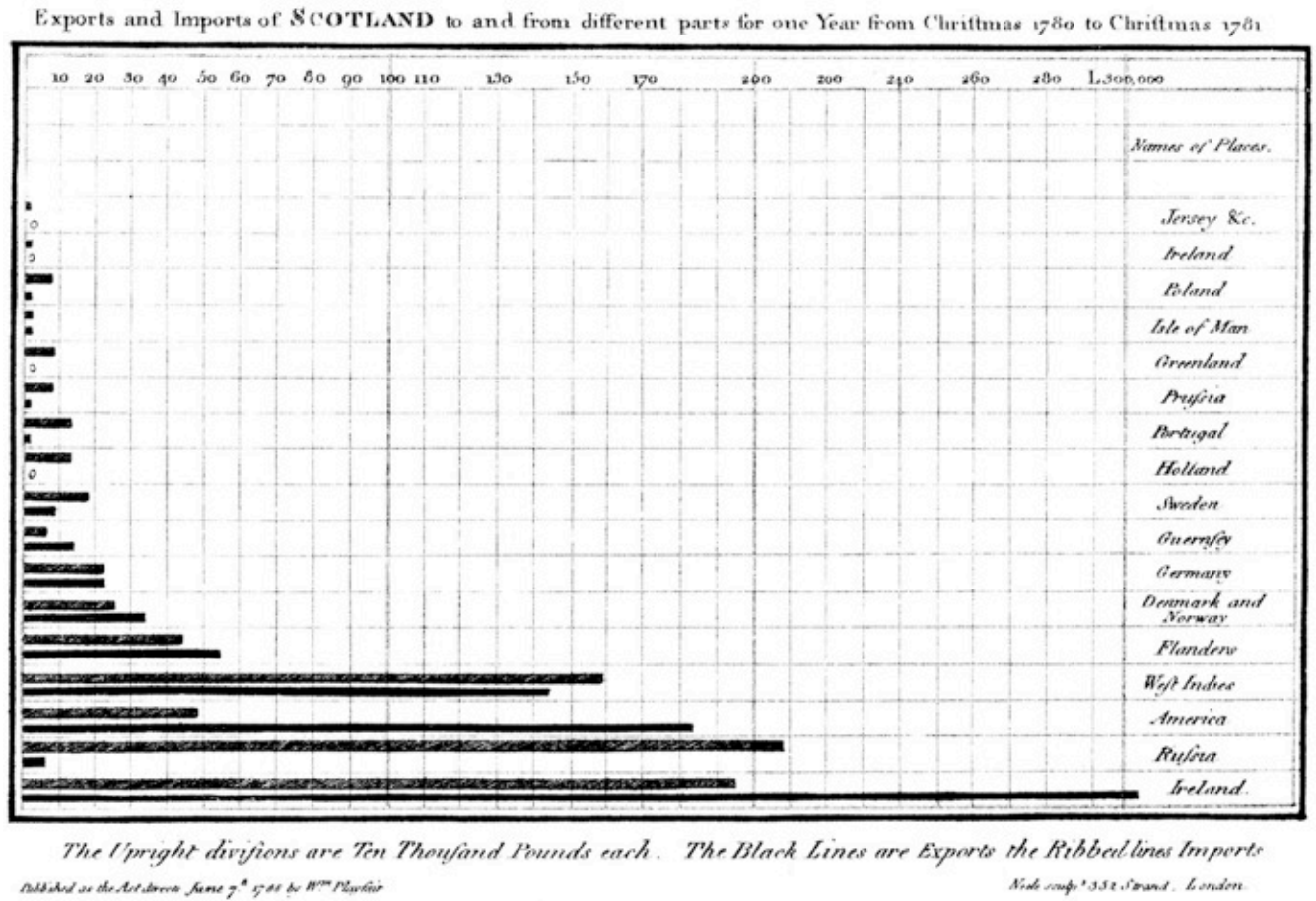
Shown above are the same quarterly revenue data added up into *calendar years*. The 1982 dip has vanished.

Bar charts

First axis (oft. horiz.): Nominal or Ordinal variable

Second axis: ratio quantity (ratio—length)

less appropriate for non-ratio quantities
(implied meaningful zero)


Exports and Imports of SCOTLAND to and from different parts for one Year from Christmas 1780 to Christmas 1781
10 20 30 40 50 60 70 80 90 100 110 130 150 170 190 200 240 260 280 L.300,000
Names of Places.
Jersey &c.
Ireland
Poland
Isle of Man
Greenland
Prussia
Portugal
Holland
Sweden
Guernsey
Germany
Denmark and Norway
Flanders
West Indies
America
Russia
Ireland
The Upright divisions are Ten Thousand Pounds each. The Black Lines are Exports the Ribbed lines Imports
Published as the Act directs June 7th 1785 by Wm Playfair
Neele sculp. 352 Strand. London

Polar plots

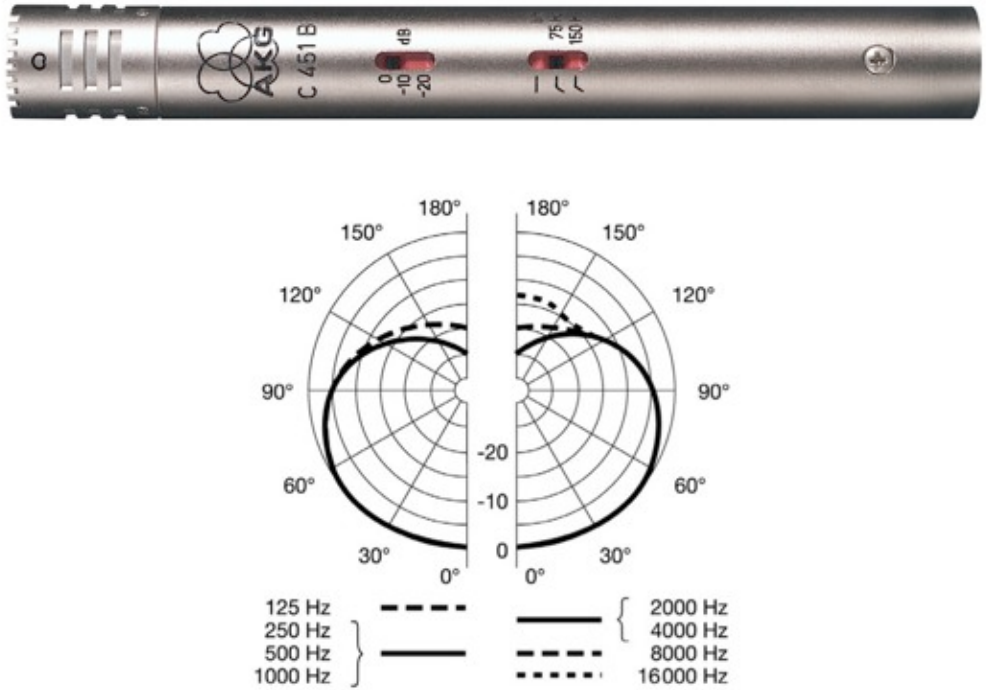
Angle: some relevant angle

Radius: ratio quantity (ratio—length)

not appropriate for non-angular quantities

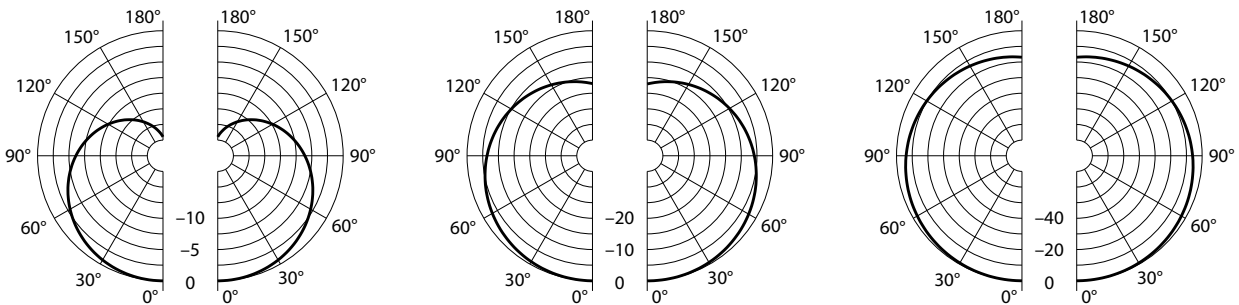
less appropriate for non-ratio quantities

beware of area exaggeration



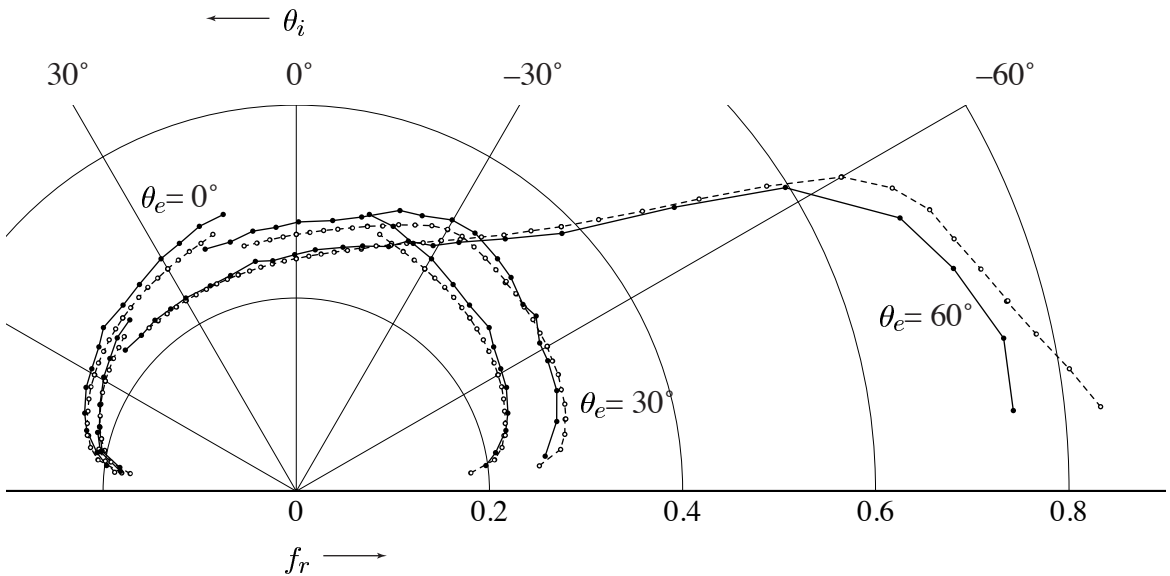
AKG Acoustics. Performance data for C451B microphone. 1973

Danger of polar plots with interval scales



Same data, 3 choices of logarithmic scale:
leads to very different shapes

Ratio quantity in polar plot: set shape



S.R. Marschner. Light scattering data for paper. 1998

Color maps

Position: position, direction, or more abstract mapping

Color: interval, ratio, or nominal quantity

be careful to map color attributes appropriately!

Color mappings

lightness (brightness, value)

strongly ordered, high resolution
quantitative variables

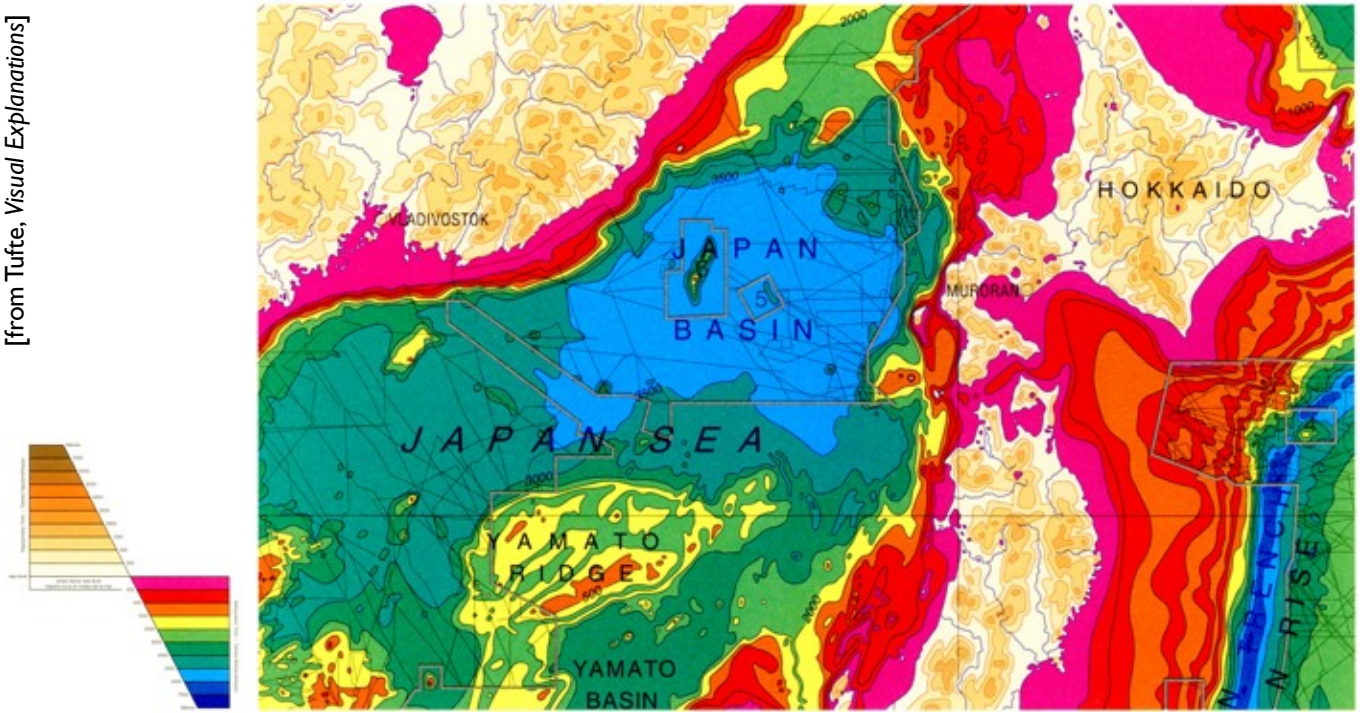
hue (what kind of color)

circular, weakly ordered, identifiable
nominal variables, or as secondary feature

saturation (colorfulness, vividness)

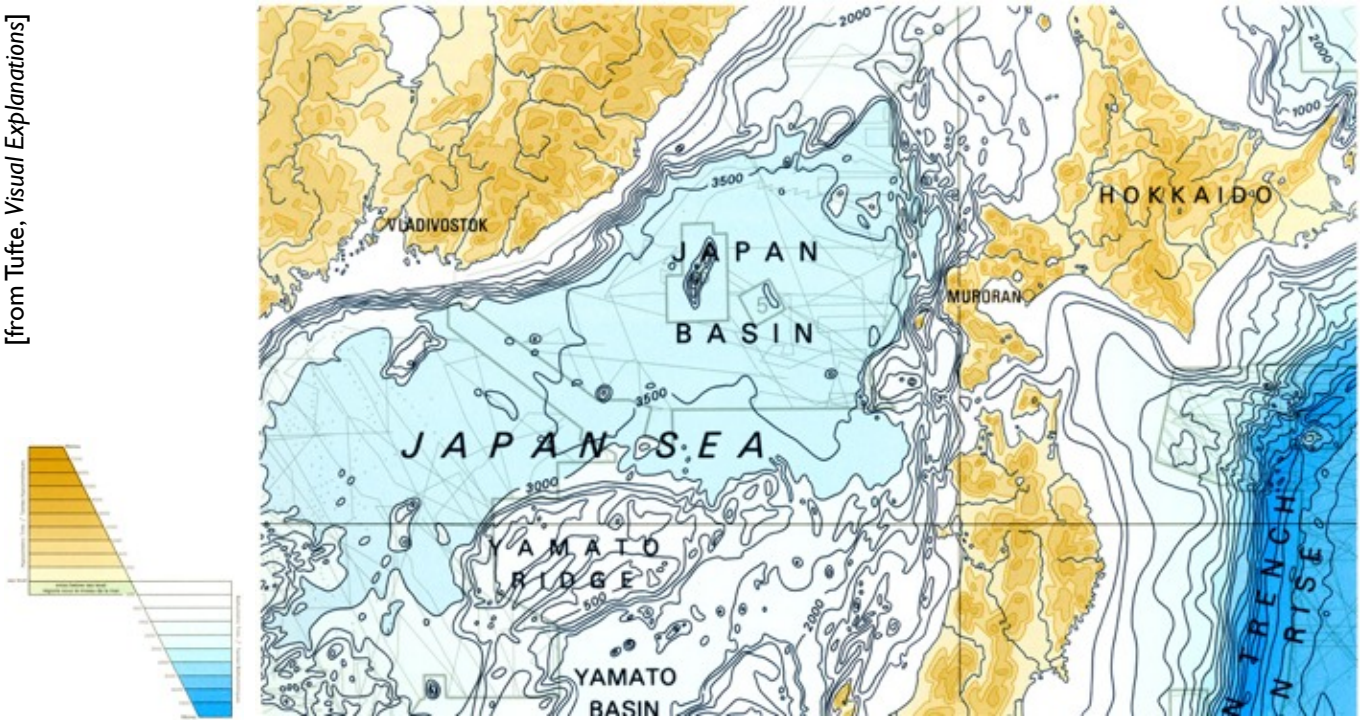
ordered, low resolution
minor quantitative variables, or combined with saturation for nominal

[from Tufte, *Visual Explanations*]

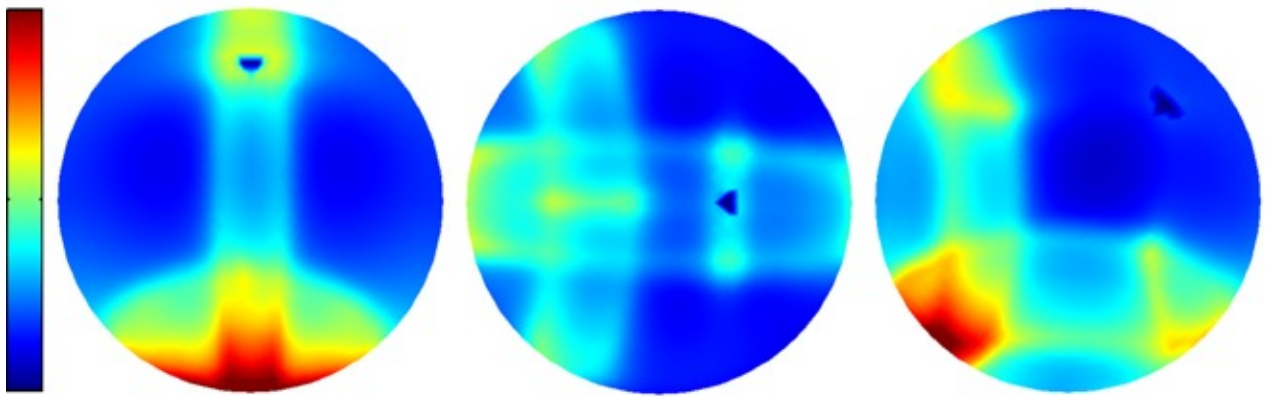


International Hydrographic Organization, 1984
(as deliberately corrupted by Tufte)

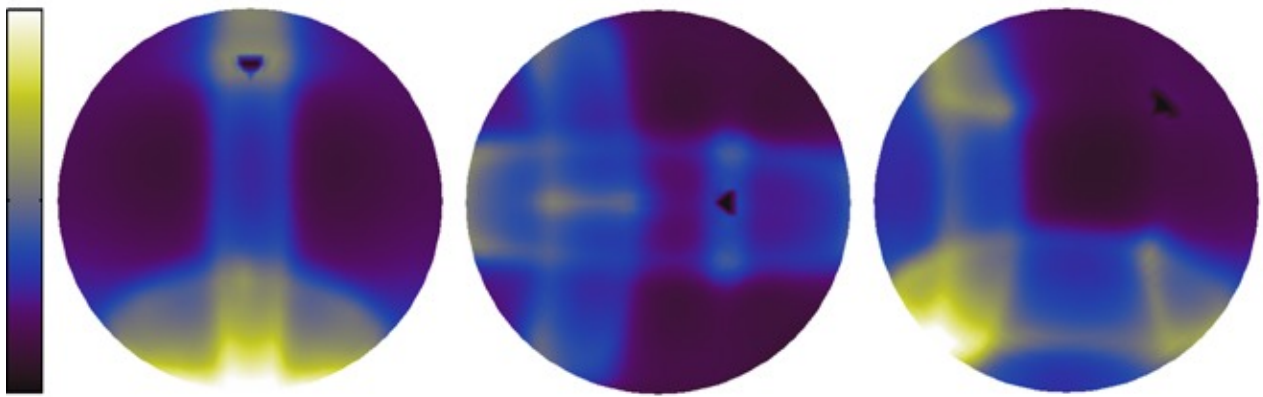
[from Tufte, *Visual Explanations*]



International Hydrographic Organization, 1984



P. Irawan & S. Marschner. Scattering data for polyester cloth. 2007
(Matlab default colormap)



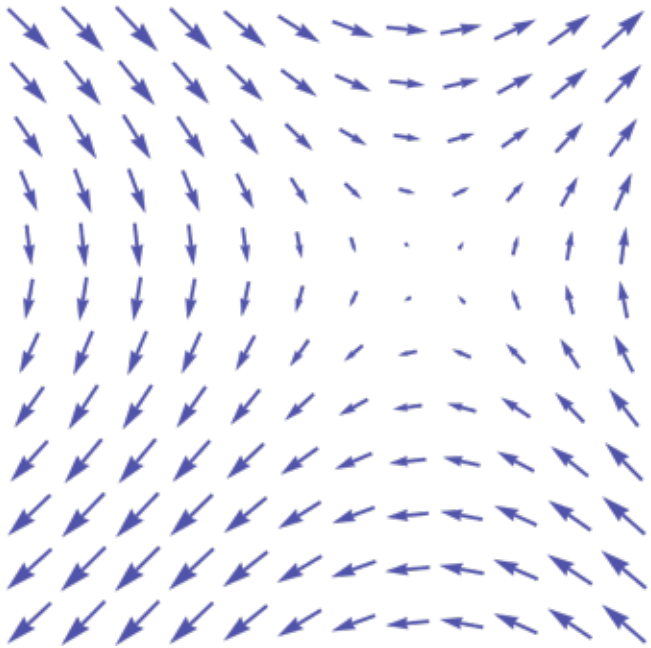
P. Irawan & S. Marschner. Scattering data for polyester cloth. 2007
(increasing value colormap)

Vector fields

Vectors are 2 (or more)-D ratio quantities

Often mapped to a textural representation

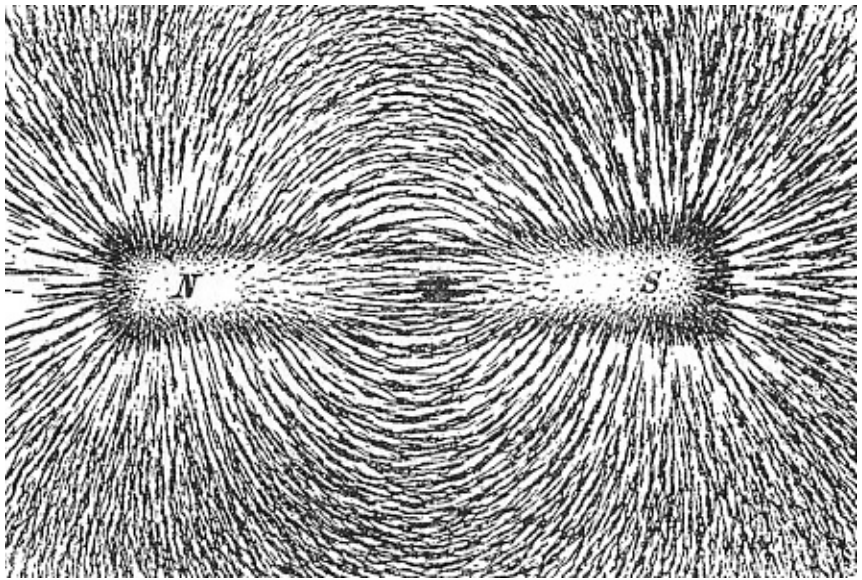
Vector fields as repeated oriented glyphs



[Jim Belk]

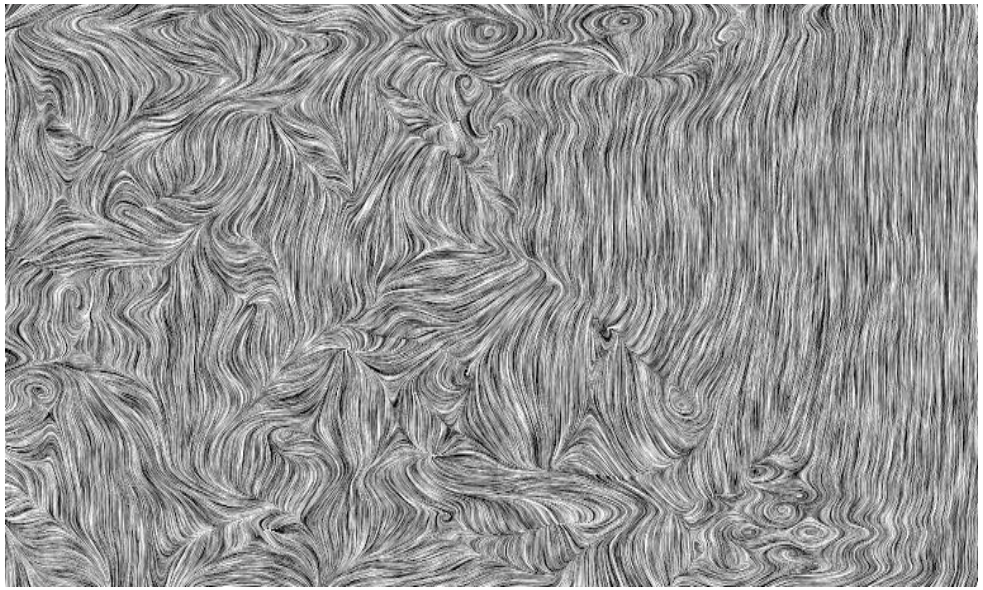
Magnitude maps to size; direction maps to direction
(note arrows are centered at grid points)

Natural visualization of magnetic field



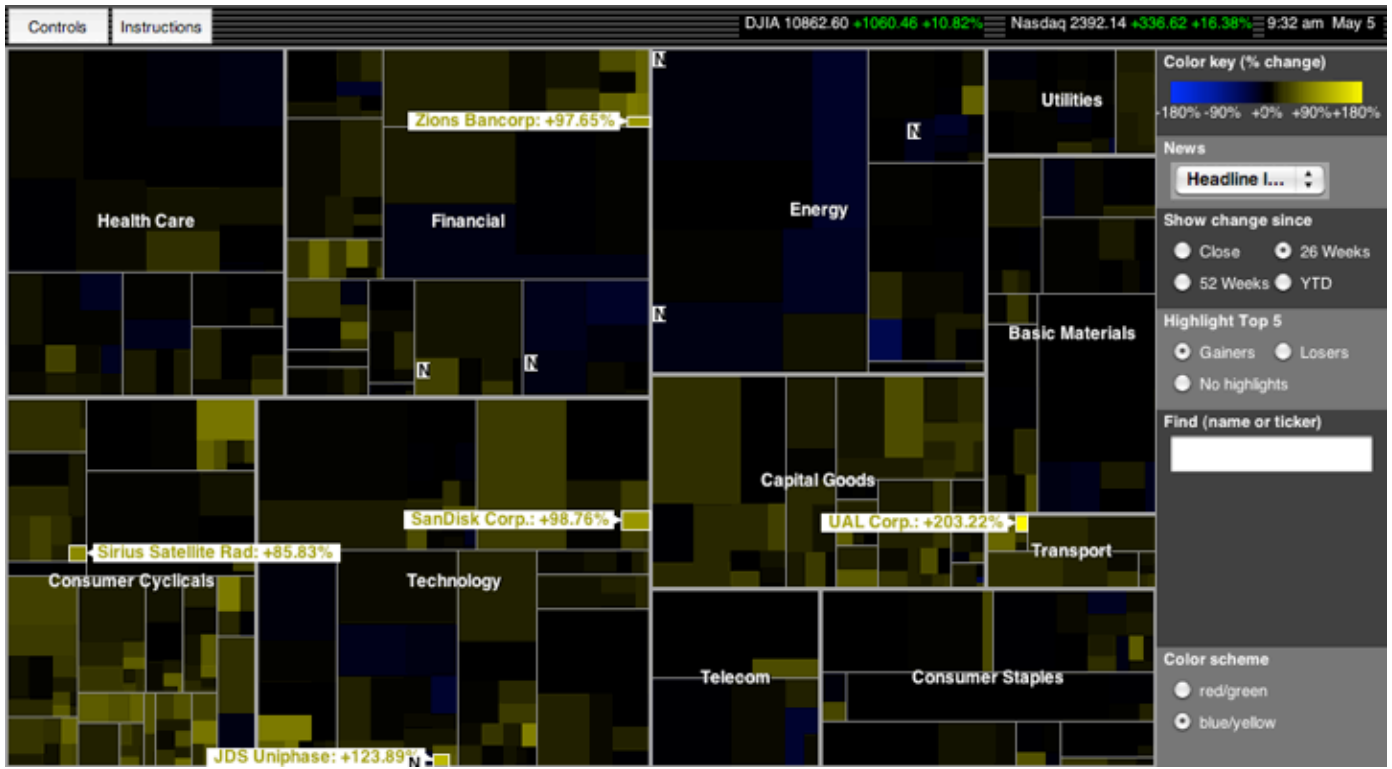
Black & Davis, *Practical Physics*. 1922

Line Integral Convolution for vector fields



Cabral and Leedom, SIGGRAPH 1993

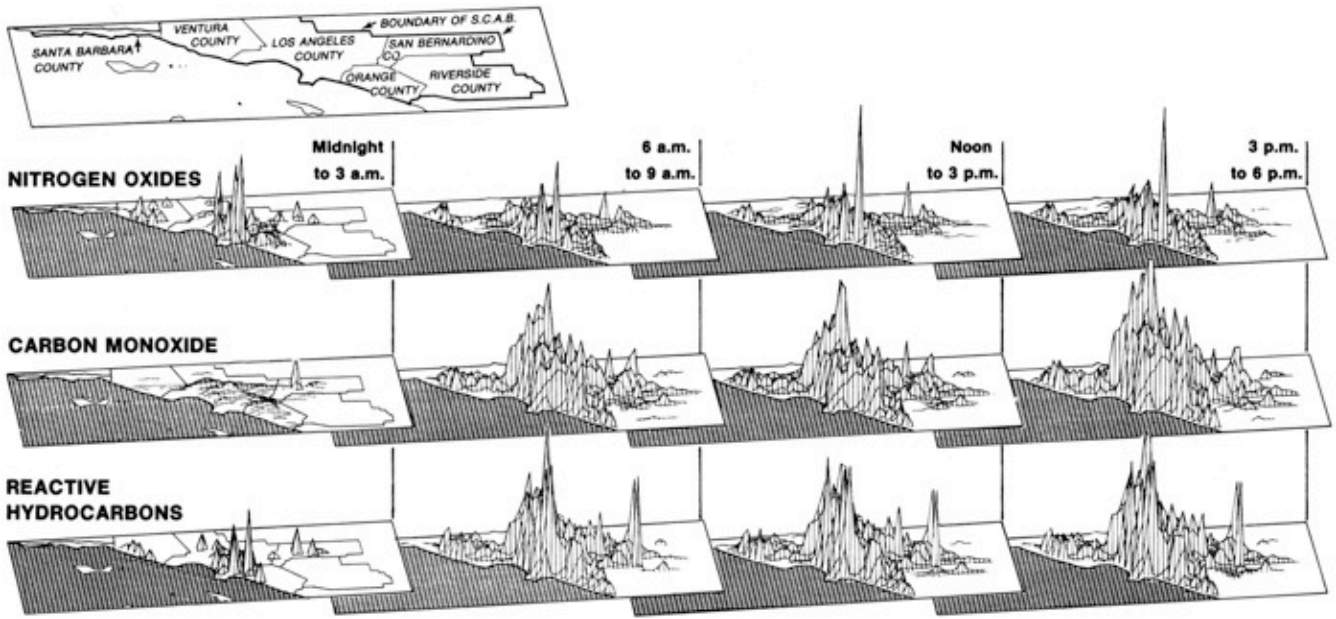
Treemaps



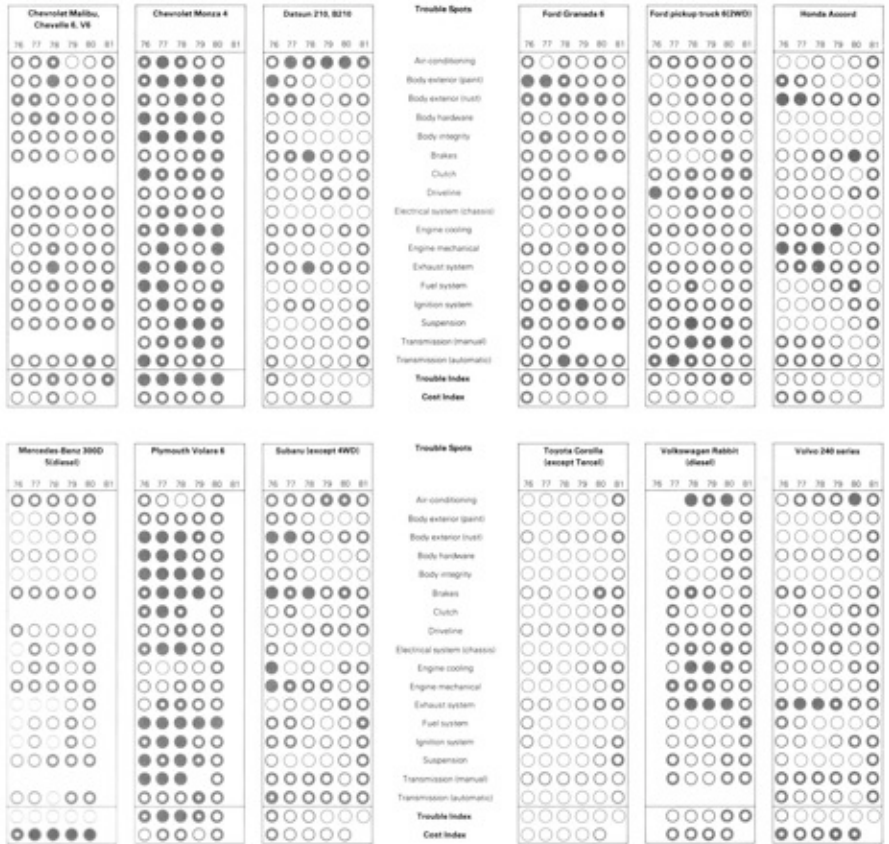
Martin Wattenberg (SmartMoney), *Map of the Market*, 1998

Small Multiples

A set of small figures following a common design that can be readily compared



Los Angeles Times / G.J. McRae. 1979



Consumer Reports. Display of historical automobile reliability data. 1982

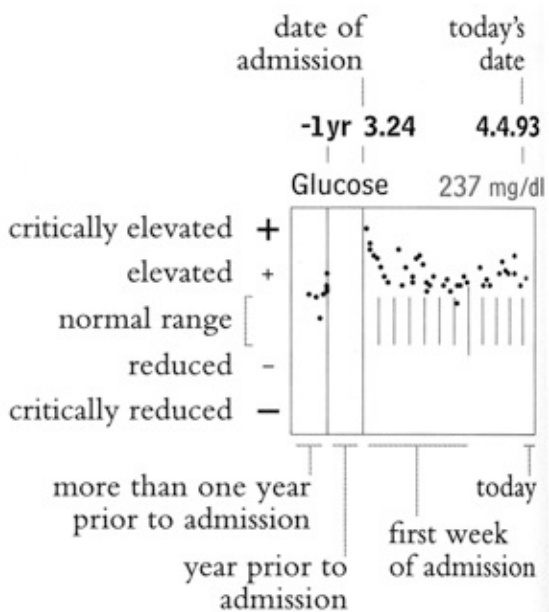
Popular mutual funds, based on assets under management.

ASSETS (MIL.)	FUND	RETURN 4 WKS.	2003	3-YR.	5-YR.
$64,368	Vanguard Index 500 Index	– 2.0%	+12.2%	– 11.7%	– 0.8%
62,510	Fidelity Magellan	– 2.1	+11.3	– 12.9	– 0.2
50,329	Amer A Invest Co of Am	– 1.2	+09.4	– 3.9	+ 4.0
47,355	Amer A WA Mutual Inv	– 1.5	+09.9	+ 00.8	+ 3.0
40,500	PIMCO Instl Tot Return	– 2.3	+02.4	+ 09.4	+ 7.6
37,641	Amer A Grow Fd of Amer	– 2.9	+14.1	– 11.0	+ 7.4
31,161	Fidelity Contrafund	– 1.0	+10.7	– 6.5	+ 3.0
28,296	Fidelity Growth & Inc	– 1.8	+ 8.2	– 8.7	– 0.1
25,314	Amer A Inc Fund of Amer	– 0.5	+ 9.9	+ 05.5	+ 5.4
24,155	Vanguard Instl Index	– 2.0	+12.3	– 11.6	– 0.7

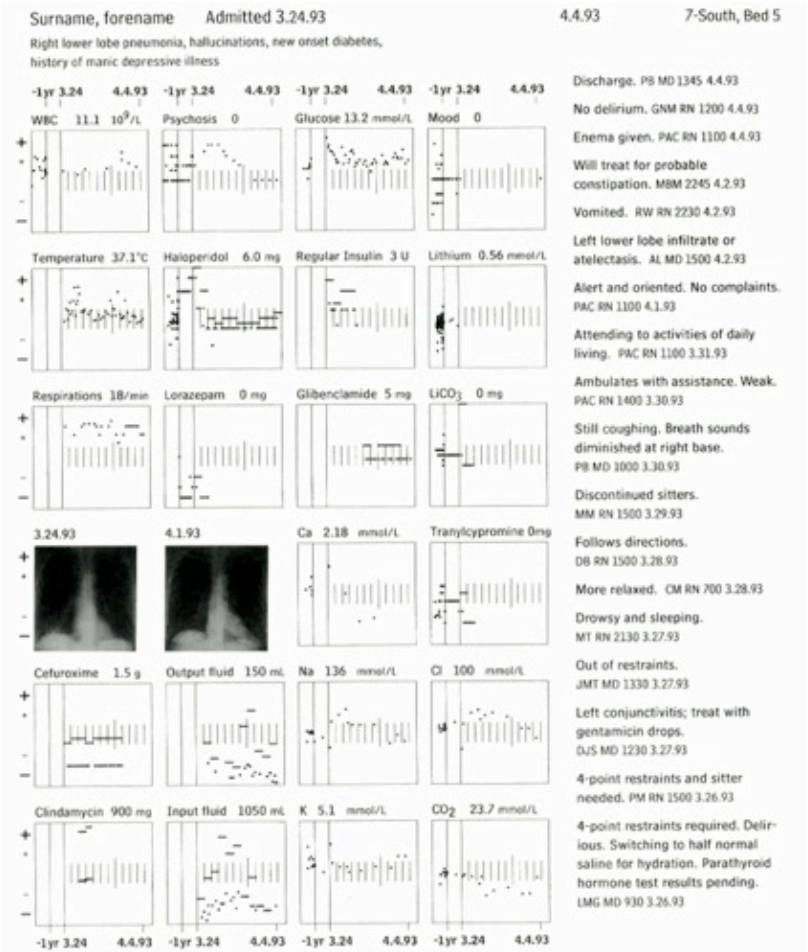
$64,368	Vanguard 500 Index	−2.0%	+12.2%	−11.7%	−0.8%
62,510	Fidelity Magellan	−2.1	+11.3	−12.9	−0.2
50,329	Amer A Invest Co Am	−1.2	+09.4	−03.9	+4.0
47,355	Amer A WA Mutual Inv	−1.5	+09.9	+00.8	+3.0
40,500	PIMCO Instl Tot Return	−2.3	+02.4	+09.4	+7.6
37,641	Amer A Grow Fd Amer	−2.9	+14.1	−11.0	+7.4
31,161	Fidelity Contrafund	−1.0	+10.7	−06.5	+3.0
28,296	Fidelity Growth & Inc	−1.8	+08.2	−08.7	−0.1
25,314	Amer A Inc Fund Amer	−0.5	+09.9	+05.5	+5.4
24,155	Vanguard Instl Index	−2.0	+12.3	−11.6	−0.7

E. Tufte "sparklines"

Visualization for medical records

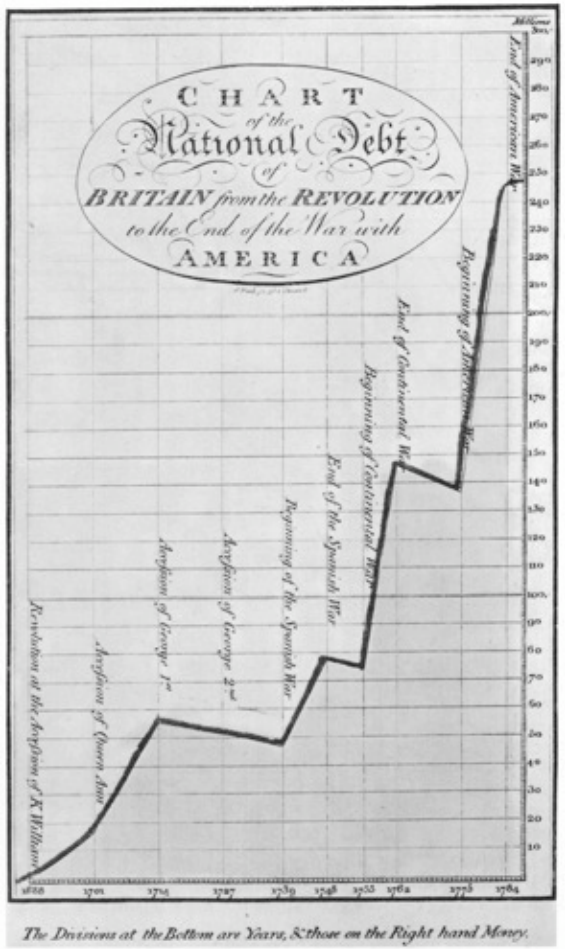


S.M. Powsner & E.R. Tufte, *The Lancet* **344**:6 1994



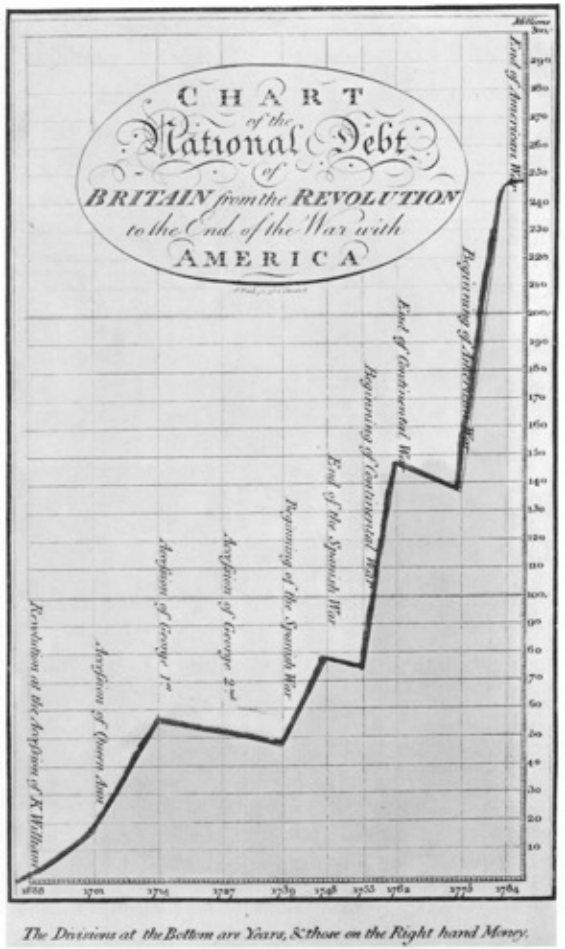
S.M. Powsner & E.R. Tufte, *The Lancet* **344**:6 1994

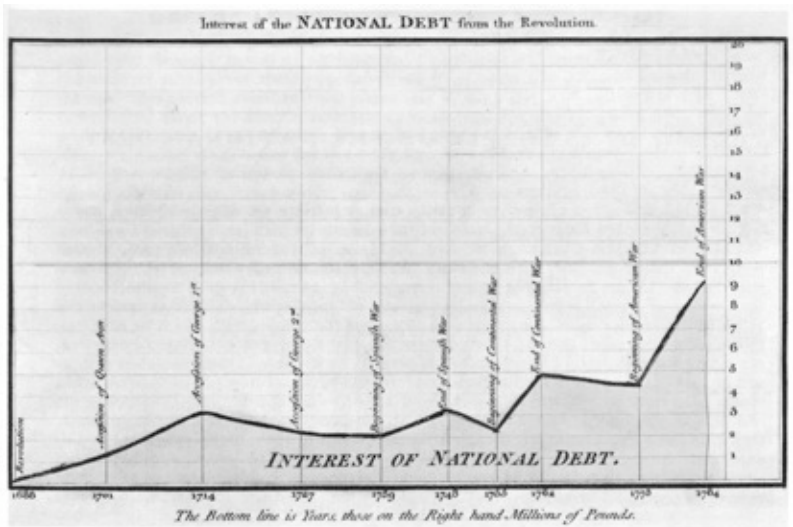
Graphical integrity

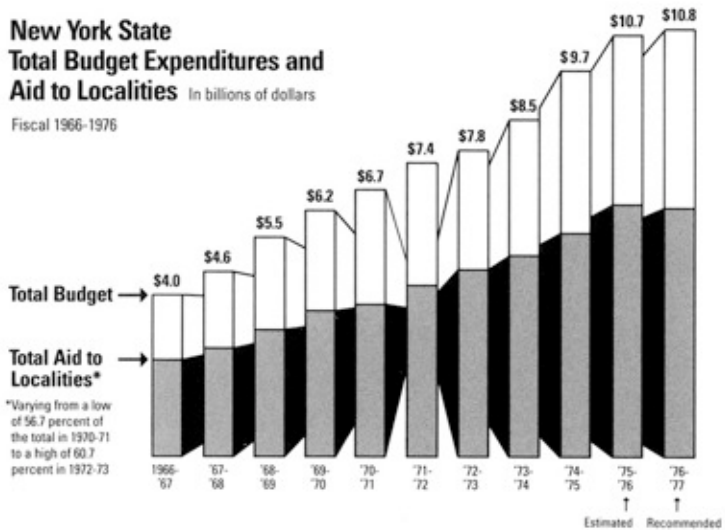


To emphasize growth, use
tall scale and don't adjust for inflation
W. Playfair, 1786

To emphasize growth, use
tall scale and don't adjust for inflation
W. Playfair, 1786

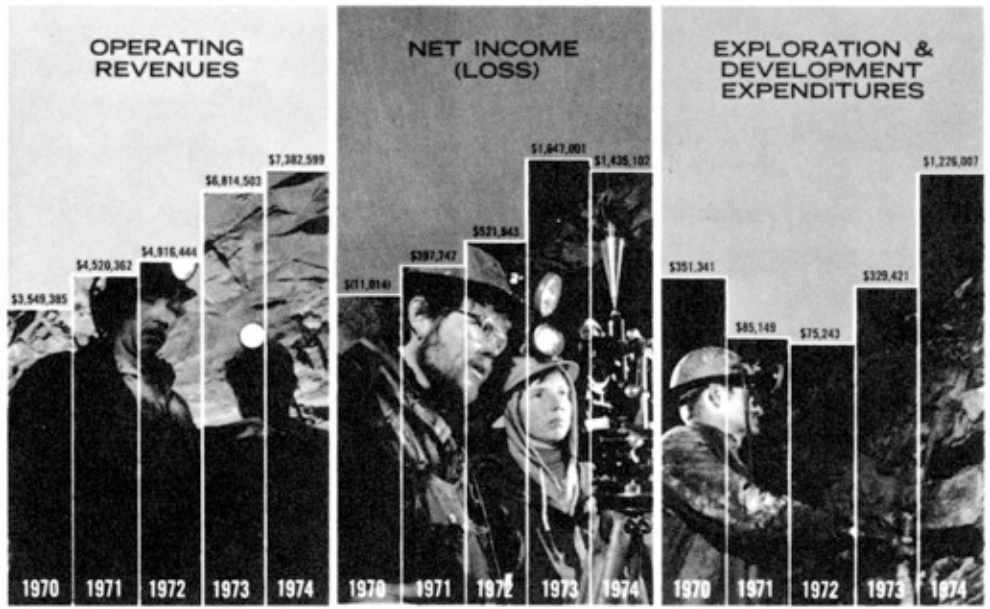




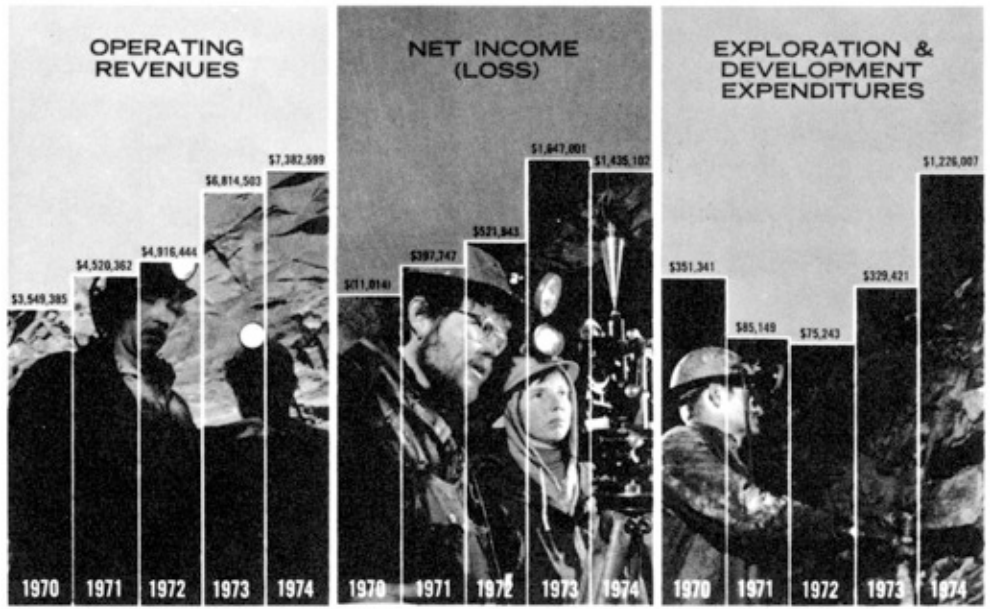


New York Times. 1976

E. R. Tufte. Fair presentation of the same data. 1983



Day Mines, Inc. 1974



Day Mines, Inc. 1974

–$4.2e6

Washington Post, 1978

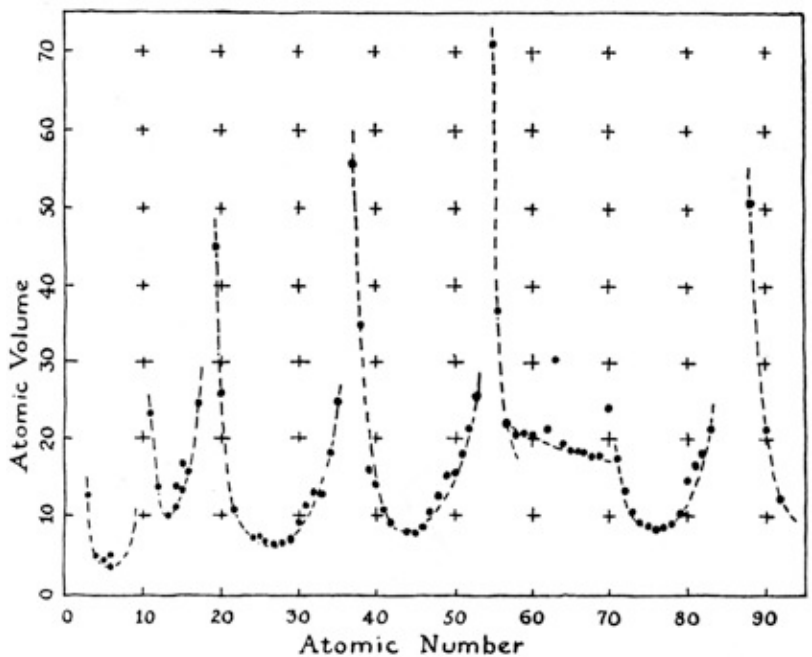
Graphical makeovers

Maximizing data:ink ratio

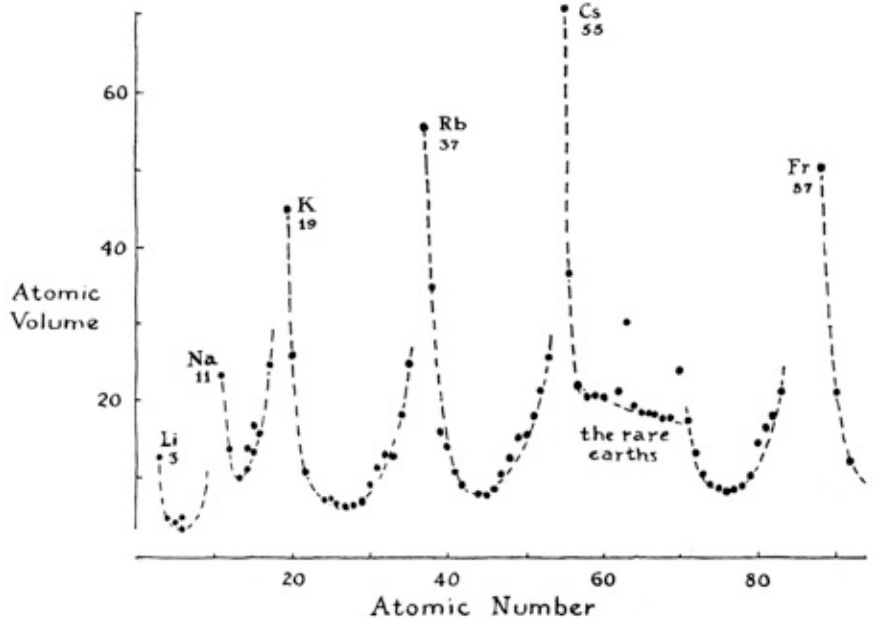
"A sentence should contain no unnecessary words, a paragraph no unnecessary sentences, for the same reason that a drawing should have no unnecessary lines and a machine no unnecessary parts."

—William Strunk, Jr.

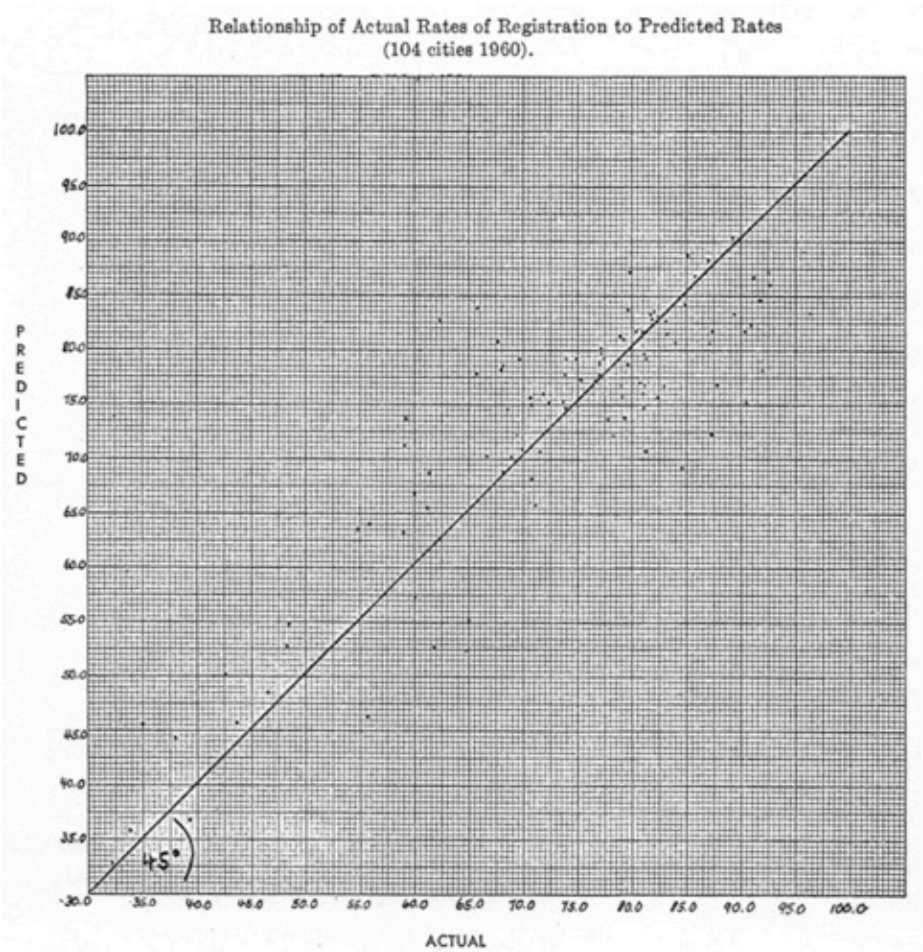
“Chart-junk”

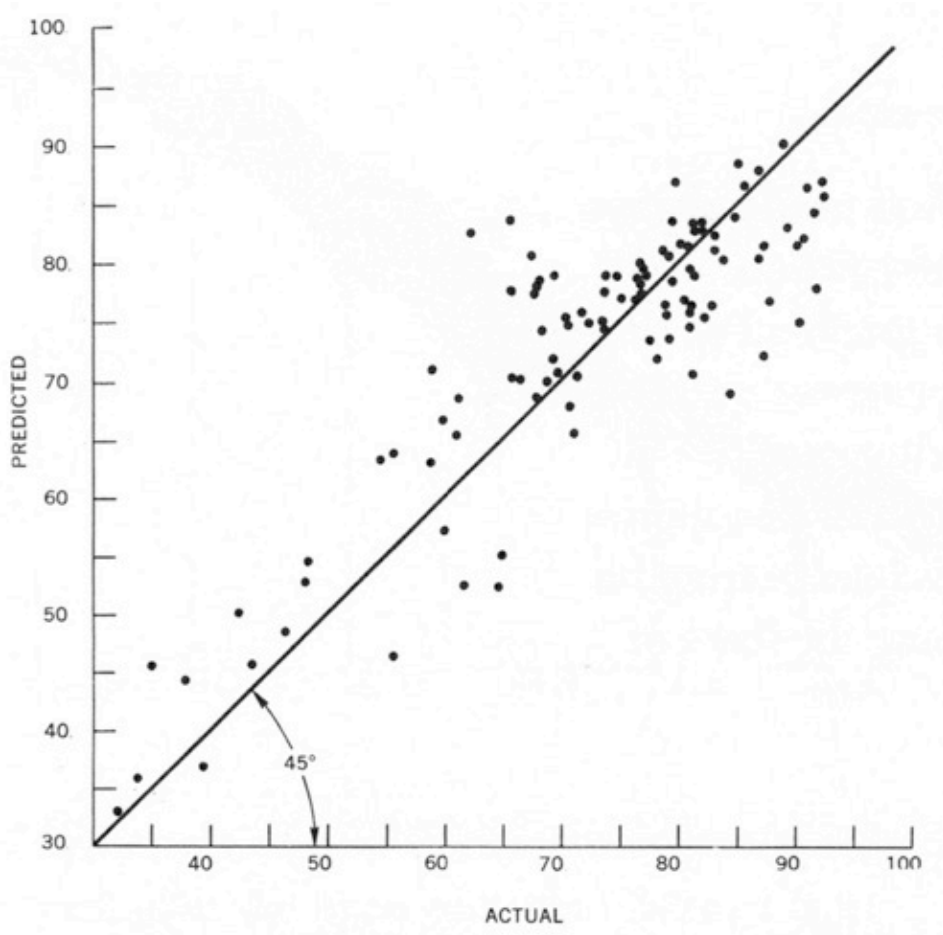


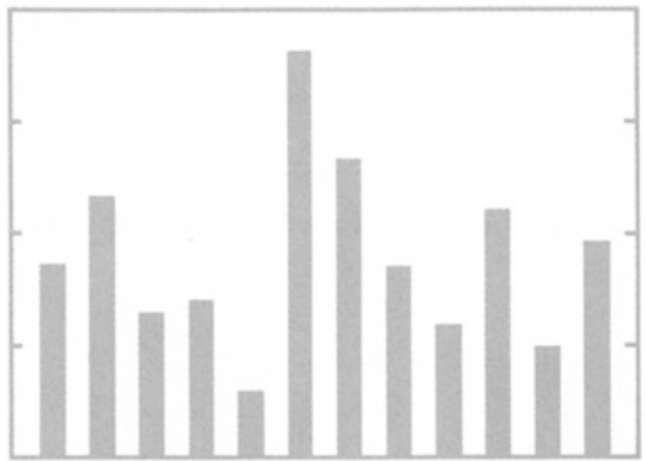
R. Hayward. From L. Pauling, *General Chemistry*. 1947

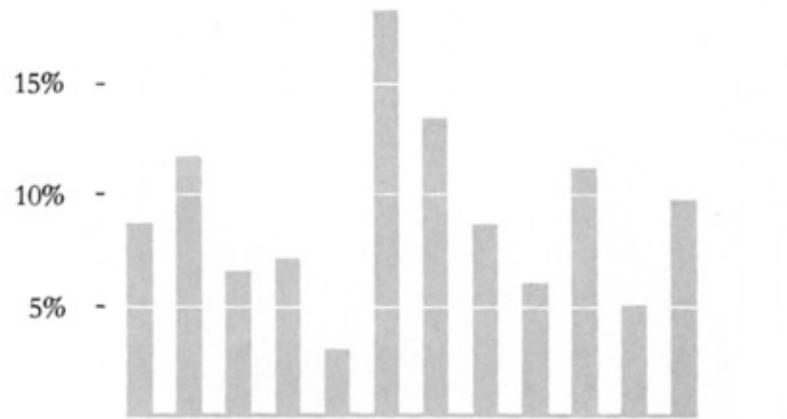


as modified by Tufte


Relationship of Actual Rates of Registration to Predicted Rates
(104 cities 1960).
PREDICTED
ACTUAL
45°
100.0
95.0
90.0
85.0
80.0
75.0
70.0
65.0
60.0
55.0
50.0
45.0
40.0
35.0
30.0
35.0
40.0
45.0
50.0
55.0
60.0
65.0
70.0
75.0
80.0
85.0
90.0
95.0
100.0


100
90
80
70
60
50
40
30
PREDICTED
45°
40
50
60
70
80
90
100
ACTUAL




15%
10%
5%

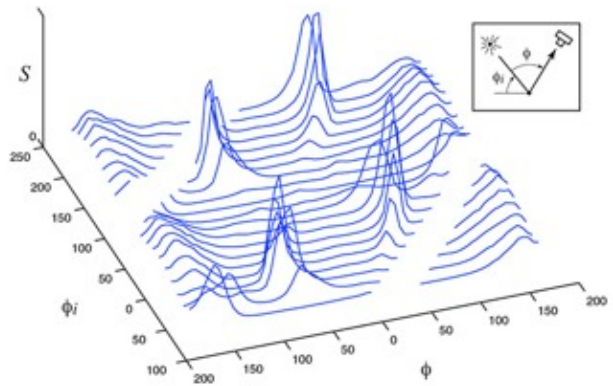


Figure 6: A measurement of scattering in the normal plane from a hair with substantial eccentricity. Bright glints appear whose location and strength depend on the orientation of the hair [subject HM].

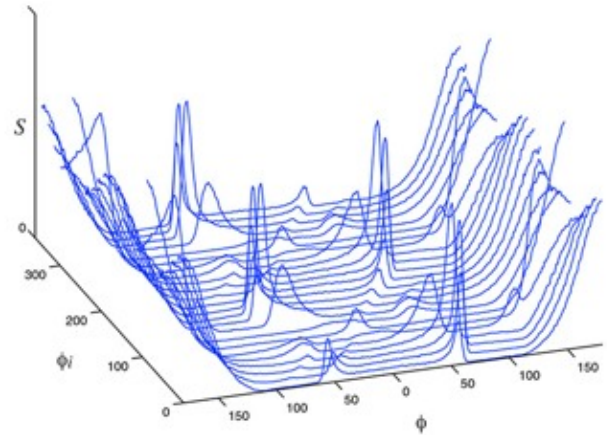
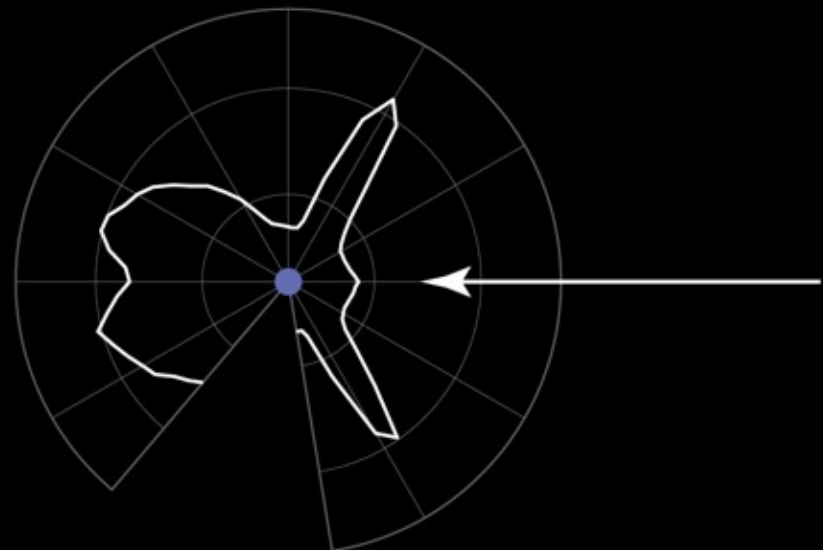


Figure 7: A photon-tracing simulation of scattering from a rough elliptical fiber. The axes are the same as in Figure 6.

S.R. Marschner. Presentation of fiber scattering data using default MATLAB plots. 2002



S.R. Marschner. Re-presentation using polar coordinates and small multiples. 2003
(thanks to François Guimbretière)

Marschner, Jensen, Cammarano, Worley, and Hanrahan. "Light Scattering from Human Hair Fibers," *SIGGRAPH 2003*.

hair
rotation
illumination

hair
rotation
illumination

hair
rotation
illumination

Displaying Quantitative Information

An exploration of Edward R. Tufte's *The Visual Display of Quantitative Information*

Jeffrey Nichols
Programming Usable Interfaces
May 2, 2003

Tufte's Principles

- Graphical Integrity
  - The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.
  - Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.
  - Show data variation, not design variation.
  - In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.
  - The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.
  - Graphics must not quote the data out of context.
- Theory of Data Graphics
  - Above all else, show the data.
  - Maximize the data-ink ratio
  - Erase non-data-ink
  - Erase redundant data-ink
  - Revise and edit
- Other comments
  - Graphical elegance is often found in simplicity of design and complexity of data
  - Data graphics are paragraphs about data and should be treated as such

Conclusions

- Show data variation, not design variation
- Avoid using ink for non-data items
- Avoid redundancy
- Clear and detailed labeling should be used to defeat graphical distortion
- Revise and Edit