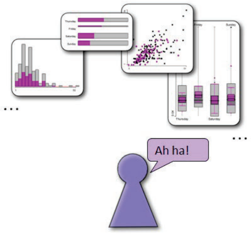
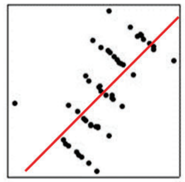
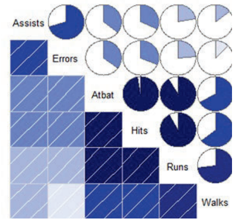




Lurking variable?



Psychology of Data Visualization: Course Overview

Michael Friendly
Psych 6135

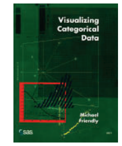
http://euclid.psych.yorku.ca/www/psy6135/
@datavisFriendly

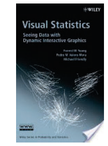
Introducing: me

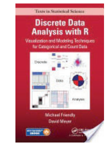
I wear two hats, both reflected on my license plate:

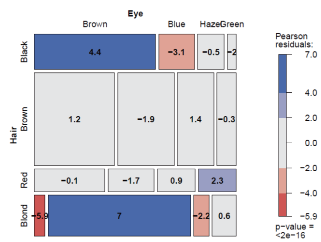
Yours to discover!

Statistical graphics developer (categorical & multivariate data analysis)

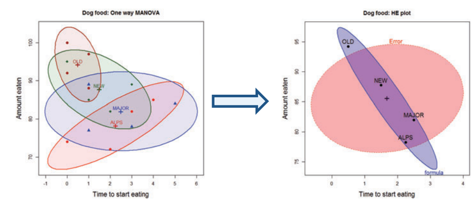








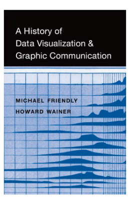
mosaic plots for frequency tables

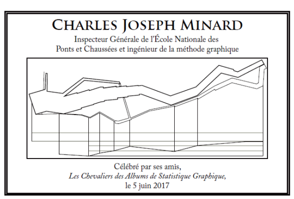


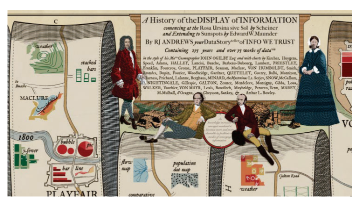
HE plots for MANOVA

Introducing: me

History of data visualization: *Les Chevaliers*; Friendly & Wainer (2021)

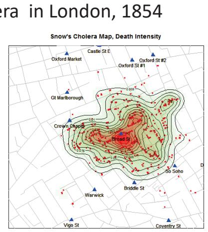






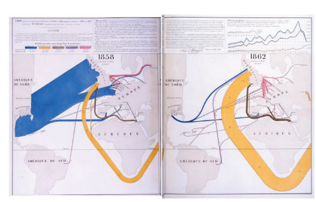
John Snow's map of cholera in London, 1854

original



modern enhancement

C. J. Minard: Flow maps of cotton trade



Visual explanation: What happened in the US Civil War?

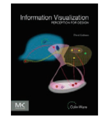
Course Topics

- Varieties of information visualization
  - Goals of visualization
  - Survey of graphic forms
- History of information visualization
- Psychological models, theories and results
  - What can people see, understand and remember from data displays?
  - Perceptual aspects, cognitive aspects
- Software tools for information visualization
- Visualization in statistics: case studies
  - Categorical data; High-D data; Dynamic and interactive methods
- Human factors research: how to tell what works

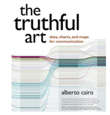
Your role

- Weekly readings – see the course web site for updates
- Discussion – no formal grade, but please contribute
- Discussion leader (20%)
  - Each week 1-2 of you will lead a brief discussion on one of the readings or sub-topics (~ 5 min.)
- Class presentation (40%)
  - In the last week, each person will give a ~ 20 min presentation on a topic of research, application or software related to data visualization
- Research proposal (40%)
  - Prepare a brief research proposal on a data visualization topic

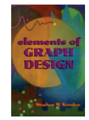
Books & Readings



Colin Ware, *Information Visualization*, 3rd Ed.
What perceptual science has to say about data visualization, from a bottom-up perspective
Course notes at: http://ccom.unh.edu/vislab/VisCourse/index.html



Alberto Cairo, *The Truthful Art*
Information graphics from a communication perspective
Blog: http://www.thefunctionalart.com/

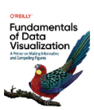


Steven Kosslyn, *Elements of Graph Design*
A cognitive psychologist looks at graphs and presents some do's and don'ts. There are better books like this today.

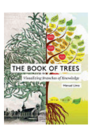
Hadley Wickham, *ggplot2: Elegant graphics for data analysis*, 2nd Ed.
1st Ed: Online, http://ggplot2.org/book/
ggplot2 Quick Reference: http://sape.inf.usi.ch/quick-reference/ggplot2/
Complete ggplot2 documentation: http://docs.ggplot2.org/current/

More books I like

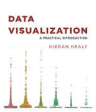
Tamara Munzner (2014), *Visualization Analysis & Design*
An attractive new book combining computer science and design perspectives
Web page: http://www.cs.ubc.ca/~tmm/vadbook/ with lots of illustrations & lectures



Claus Wilke(2019), *Fundamentals of Data Visualization*
A detailed, practically oriented book on data visualization methods
Online: https://clauswilke.com/dataviz/
Course notes: https://wilkelab.org/SDS375/



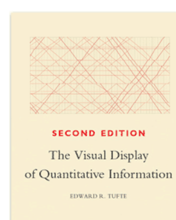
Manuel Lima, *The Book of Trees: Visualizing branches of knowledge*
A visual delight; an entire history of tree-type diagrams
Blog: http://www.visualcomplexity.com/vc/blog/

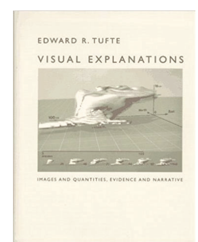


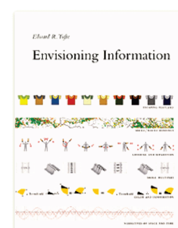
Keiran Healy, *Data Visualization: A Practical Introduction*
An accessible primer on how to create effective graphics from data using ggplot2
Online: http://socvis.co

Tufte Stufte

Four books by Edward Tufte largely defined the landscape for data visualization and information design



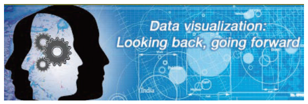




Concepts introduced:
- chart junk,
- data-ink ratio,
- small multiples,
- substance takes precedence over visual design

Web site:
https://www.edwardtufte.com

Blogs & Web resources



My web site, http://datavis.ca. Contains the Milestone Project on the history of data vis, Data Visualization gallery, links to books, papers and courses.

JUNK CHARTS
Recycling chartjunk as junk art

Kaiser Fung, http://junkcharts.typepad.com/. Fung discusses a variety of data displays and discusses how they can be improved.

FLOWINGDATA

Nathan Yau's blog, http://flowingdata.com. A large number of blog posts illustrating data visualization methods with tutorials on how to do these with R and other software.

Visiphilia
The love of plotting data

http://visiphilia.org/. Statisticians Di Cook and Heike Hofmann from Iowa State University blog about data visualization topics, using R

visual complexity

Manuel Lima's blog, http://www.visualcomplexity.com/vc/blog/, with hundreds of projects on all types of visualizations

Blogs & Web resources

DATA STORIES

http://datastori.es/. A podcast on data visualization with Enrico Bertini and Moritz Stefaner; interviews with over 100 graphic designers & developers.

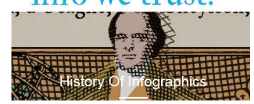
KANTAR Information is Beautiful Awards

Annual awards celebrate excellence and beauty in data visualizations, infographics, interactives & information art. https://www.informationisbeautifulawards.com

R-bloggers
R news and tutorials contributed by (750) R bloggers

https://www.r-bloggers.com/. A large collection of posts on R news and tutorials by over 750 R bloggers.

info we trust.



Raymond Andrews, http://infowetrust.com/. A visual storyteller delights with graphic stories from the history of data visualization

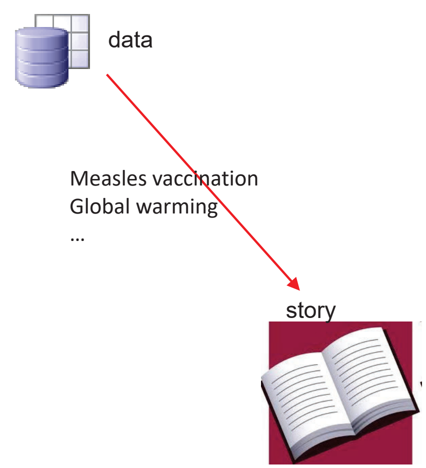
Psychology facts: Why visualization?

- ~90% of information about the environment is received by a person through the eyes.
- ~50% of our brain neurons are constantly involved in the processing of visual information.
- The presence of pictures increases desire to read the text by ~80%.
- A person remembers 10% of what he/she heard, 20% of what's read, and 80% of what's seen!
- People perceive 70% of the information if there are no illustrations. Add pictures there — the figure will increase up to 95%.

McLeod, S. A. (2018). Visual perception theory. Simply Psychology. www.simplypsychology.org/perception-theories.html

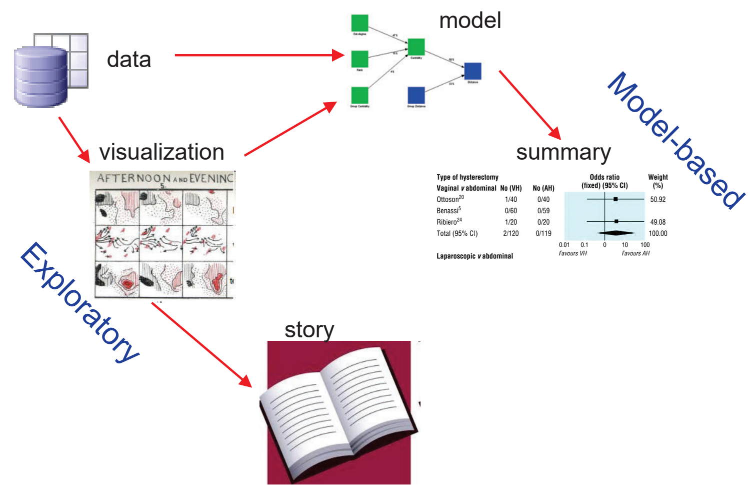
Data, pictures, models & stories

Goal: Tell a credible story about some real data problem



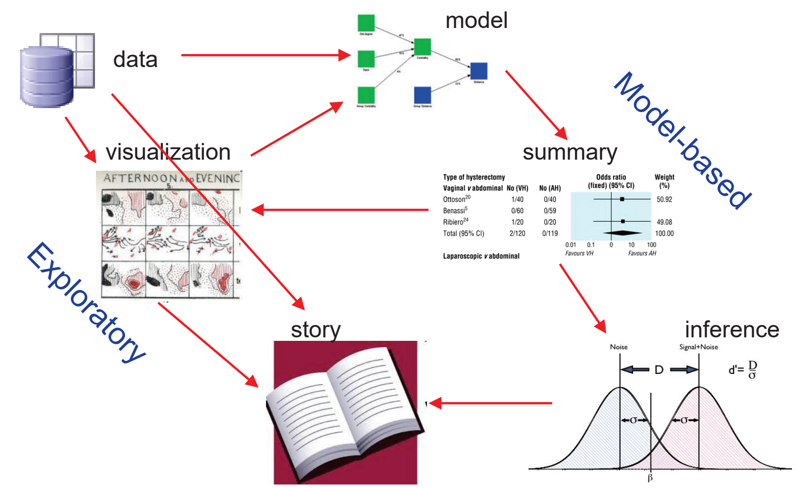
Data, pictures, models & stories

Two paths to enlightenment



Data, pictures, models & stories

Now, tell the story!

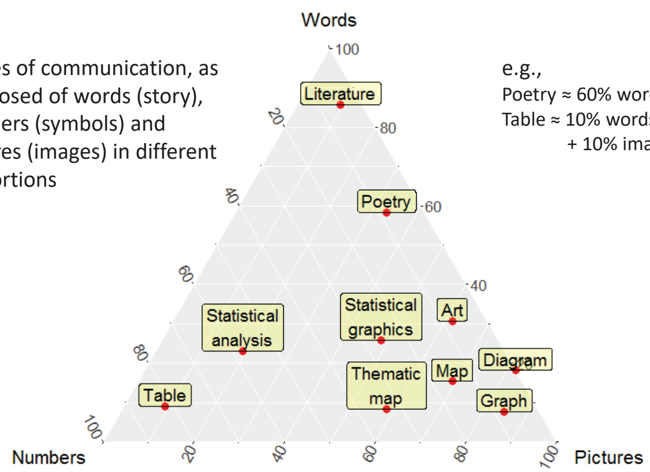


Words, numbers and pictures

Pictures and images in a wider context

Modes of communication, as composed of words (story), numbers (symbols) and pictures (images) in different proportions

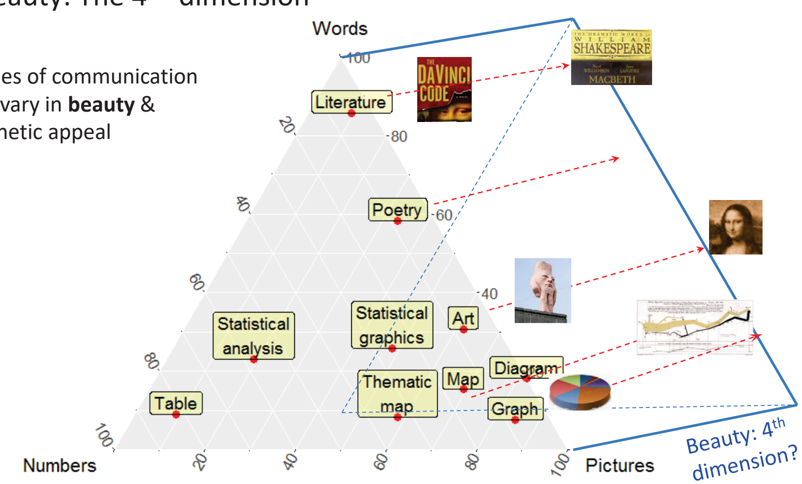
e.g.,
Poetry ≈ 60% words + 40% images
Table ≈ 10% words + 80% numbers + 10% images



Words, numbers and pictures

Beauty: The 4th dimension

Modes of communication also vary in **beauty** & aesthetic appeal



Roles of graphics in communication

- Graphs (& tables) are forms of communication:
  - What is the audience?
  - What is the message?

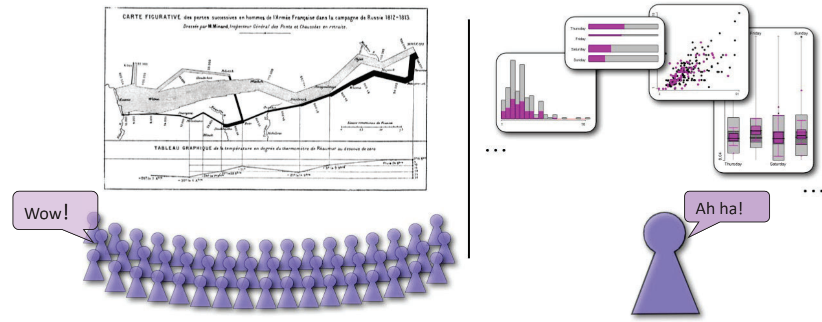
Analysis graphs: design to see patterns, trends, aid the process of data description, interpretation

Presentation graphs: design to attract attention, make a point, illustrate a conclusion

Basic functions of data display

	Primary Use	*Presentation Goal*	*Design Principles*
Data Display	Analysis	Reconnaisance	*Perception*
		Exploration	*Detection*
		Diagnosis Model building	*Comparison*
	Presentation	to Simulate	*Aesthetics*
		to Persuade	*Rhetoric*
		to Inform	*Exposition*

Different graphs for different purposes



Presentation

Goal: the Wow! experience
Single image for a large audience
Tells a clear story!

Exploration

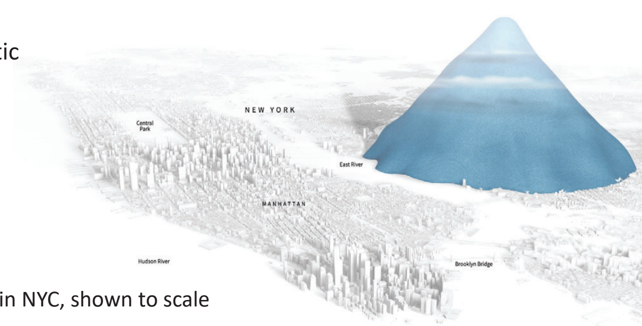
Goal: the Ah ha! Experience

Many images, for a narrow audience (you!), linked to analysis

Infographics

The best infographics tell a story, using numbers, but shown visually

Drowning in plastic



Plastic bottles sold in NYC, shown to scale

From: https://graphics.reuters.com/ENVIRONMENT-PLASTIC/0100B275155/index.html

Powerful graphs: Measels and vaccines

Visualizing the impact of health policy interventions

In 2015 Tynan DeBold & Dov Friedman in the *Wall Street Journal* show the effect of the introduction of vaccination programs in the US states on disease incidence, using color-coded heat maps for a variety of diseases

Measles was decimated!

The message hits you between the eyes!

Powerful graphs make comparison easy

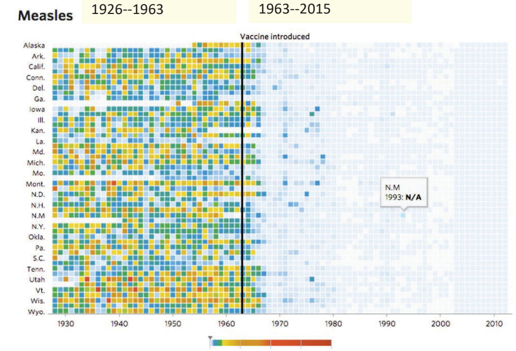
In 2014, vaccination rates declined and measles re-emerged in those areas

Effective graphs can cure ignorance, but not stupidity.

Measles

Before: 17.0 M cases, 1926--1963

After: 1.7 M cases, 1963--2015

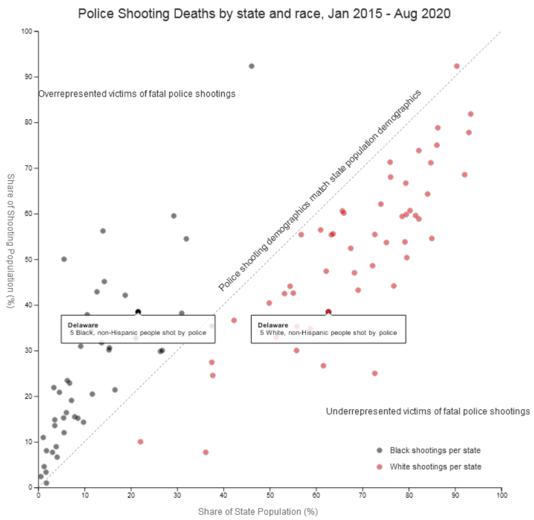


Source: http://graphics.wsj.com/infectious-diseases-and-vaccines/

Police shooting deaths

Analysis of Washington Post database on 5500 police shooting deaths for Blacks vs. Whites

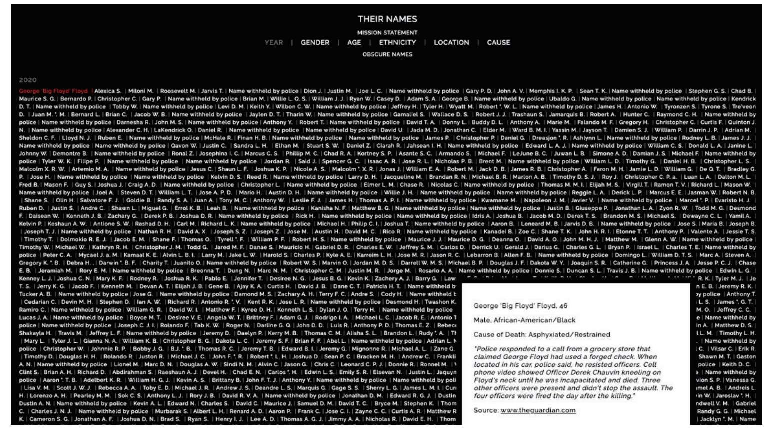
Plotting % of shooting vs. % of pop shows a clear & disturbing pattern



From: https://medium.com/two-n/visualizing-wapos-police-shooting-dataset-3792593f6be

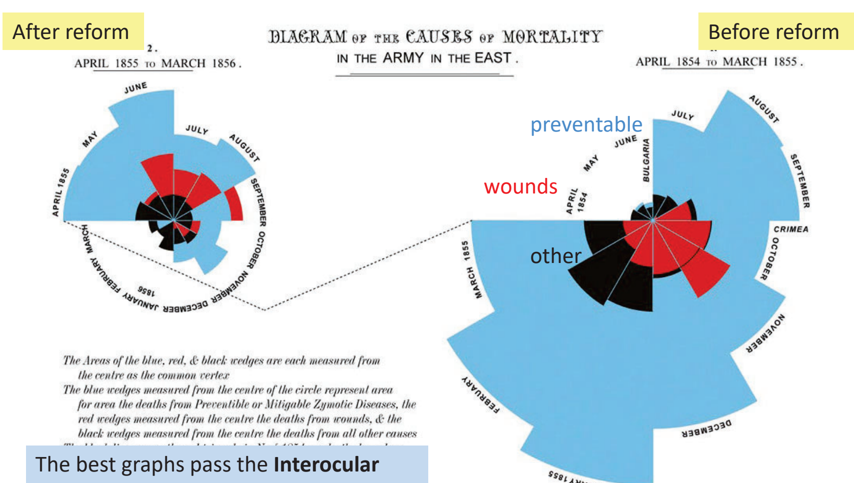
Their names: Interactive graphic

As powerful as Yad Vashem & the Washington D.C. Vietnam memorial, this list of 28,000 US fatal encounters with police commands attention.
Each one is linked to a story or description. Classified by {Gender, Age, Cause}



https://theirnames.org/

Presentation graph: Nightingale (1857)

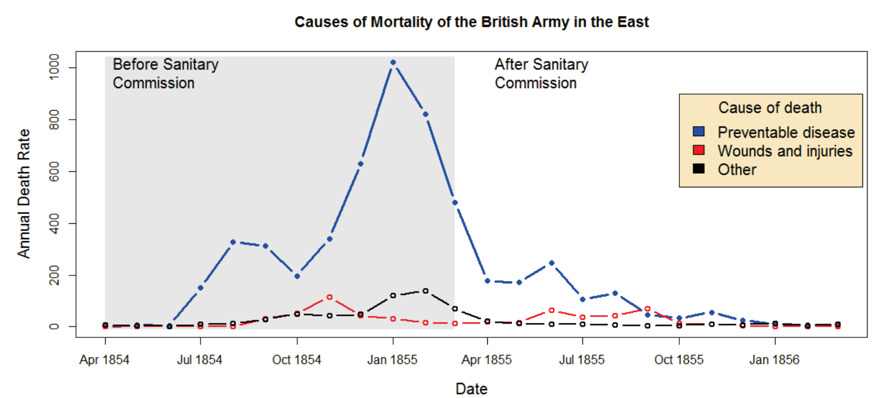


The best graphs pass the **Interocular Traumatic Test**: the message hits you between the eyes!

Data graph: Nightingale (1857)

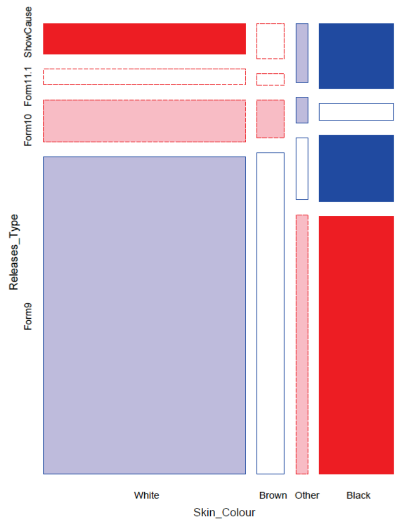
The same, as a data graph, using time-series line plots

Many statisticians might prefer this today, but it doesn't draw attention or interest as Flo's original did.



Racial profiling: Analysis graph

- Toronto Star (2002) study of police actions on a charge of simple possession of marijuana
  - release with a summons (Form 9) vs. hold for bail (Show cause)
  - Evidence for racial bias?
- First graph: mosaic display
  - area ~ frequency
  - shading: ~ residual
    - Obs > Expected in blue
    - Obs < Expected in red



Racial profiling: The process

How to communicate these results most effectively?

- What is the message? What features are directly comprehensible to the audience?

Graphic designer's early attempts

My early attempts

Racial profiling: Presentation graphic

Together, we created this self-explaining infographic

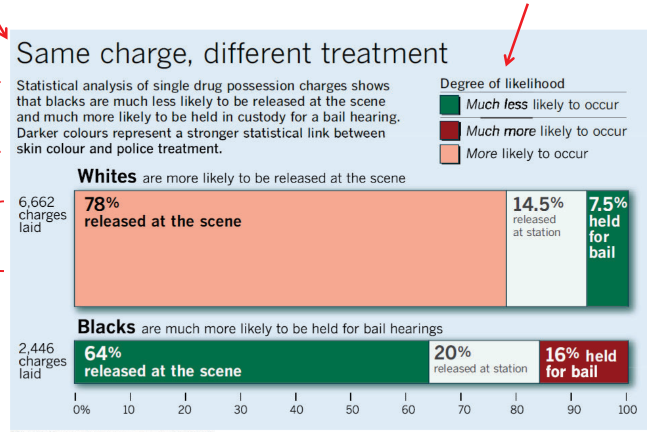
Title gives the main conclusion

Legend gives a layman's description of shading levels

Text description gives details

Bar width ~ charges
Divided by % release

numbers shown in the cells

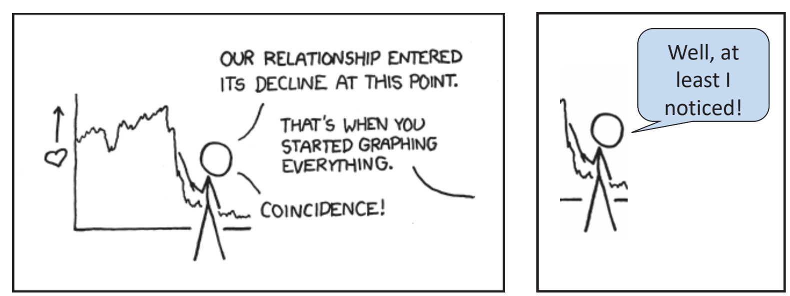


Why plot your data?

Graphs help us to see

patterns, trends, anomalies and other features

not otherwise easily apparent from numerical summaries.

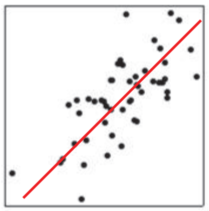


Source: http://xkcd.com/523/

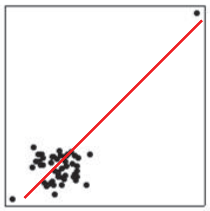
Why plot your data?

Three data sets with exactly the same bivariate summary statistics:

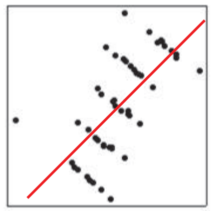
- Same correlations, linear regression lines, etc
- Indistinguishable from standard printed output
- Totally different interpretations!



Standard data



r=0 but + 2 outliers

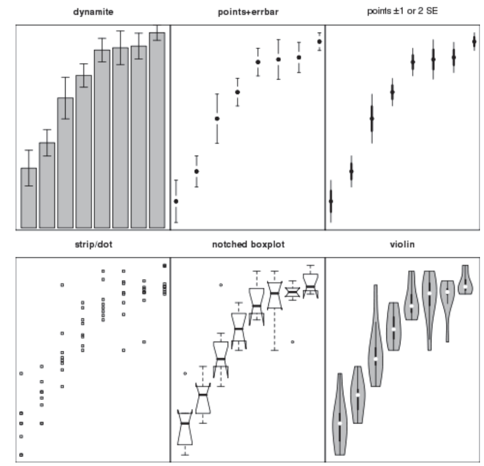


Lurking variable?

Comparing groups: Analysis vs. Presentation graphs

Six different graphs for comparing groups in a one-way design

- which group means differ?
- equal variability?
- distribution shape?
- what do error bars mean?
- unusual observations?



Never use dynamite plots

Always explain what error bars mean

Consider tradeoff between summarization & exposure

Presentation: Turning tables into graphs

Table 2 from Stevens (2006): Determinants of Authoritarian Aggression

Variable	Coefficient (Standard Error)
Constant	.41 (.93)
Countries	
Argentina	1.31 (.33)$^{\#\#\#\ B,M}$
Chile	.93 (.32)$^{\#\#\#\ B,M}$
Colombia	1.46 (.32) $^{\#\#\#\ B,M}$
Mexico	.07 (.32)^{A,CH,CO,V}
Venezuela	.96 (.37)$^{\#\#\ B,M}$
Threat	
Retrospective egocentric economic perceptions	.20 (.13)
Prospective egocentric economic perceptions	.22 (.12)$^{\#}$
Retrospective sociotropic economic perceptions	-.21 (.12)$^{\#}$
Prospective sociotropic economic perceptions	-.32 (.12)$^{\#\#}$
Ideological Distance from president	
Ideology	
Ideology	.23 (.07) $^{\#\#\#}$
Individual Differences	
Age	.00 (.01)
Female	-.03 (.21)
Education	.13 (.14)
Academic Sector	.15 (.29)
Business Sector	.31 (.25)
Government Sector	-.10 (.27)
R^2	.15
Adjusted R^2	.12
n	500

$^{\#\#\#}p < .01$, $^{\#\#}p < .05$, $^{\#}p < .10$ (two-tailed)

^{A} Coefficient is significantly different from Argentina's at p < .05;

^{B} Coefficient is significantly different from Brazil's at p < .05;

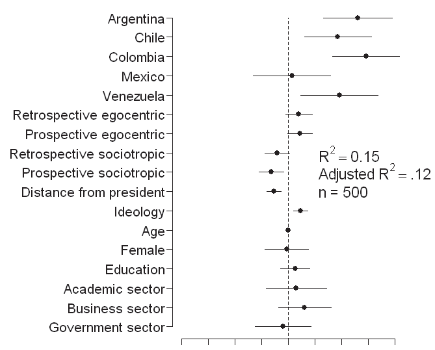
^{CH} Coefficient is significantly different from Chile's at p < .05;

^{CO} Coefficient is significantly different from Colombia's at p < .05;

^{M} Coefficient is significantly different from Mexico's at p < .05;

^{V} Coefficient is significantly different from Venzeluela's at p < .05

Graphs of model coefficients are often clearer than tables



Source: tables2graphs.com

Graphs & Information design

- Graphs & info displays should be viewed in relation to communication goals & audience
- Some criteria for assessing:
  - **comprehensibility**: does it make information as easy to understand as possible?
  - **attention**: does the audience take notice?
  - **aesthetics**: is it visually appealing?
  - **memorability**: will they remember it?
  - **behavior**: does it result in some desired action?

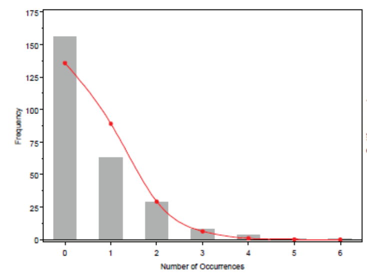
From: Ben Jones, To Optimize or to Satisfice in Data Visualization? https://dataremixed.com/2016/01/optimize-or-satisfice-in-dataviz/

Effective data display

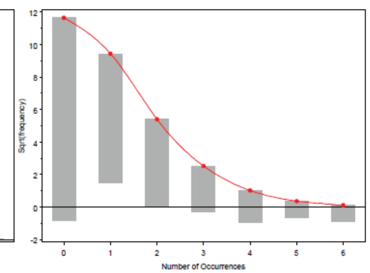
- Make the data stand out
  - Fill the data region (axes, ranges)
  - Use visually distinct symbols (shape, color) for different groups
  - Avoid chart junk, heavy grid lines that detract from the data
- Facilitate comparison
  - Emphasize the important comparisons visually
  - Side-by-side easier than in separate panels
  - “data” vs. a “standard” easier against a horizontal line
  - Show uncertainty where possible
- Effect ordering
  - For variables and unordered factors, arrange them according to the effects to be seen

Comparisons— Make visual comparisons easy

- Visual grouping— connect with lines, make key comparisons contiguous
- Baselines— compare *data* to *model* against a line, preferably horizontal
- Frequencies often better plotted on a square-root scale



Standard histogram with fit

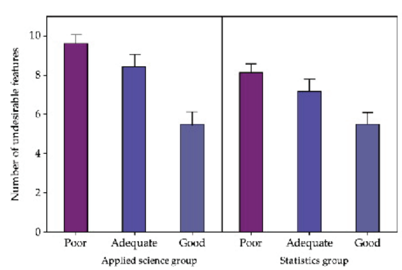


Suspended rootogram

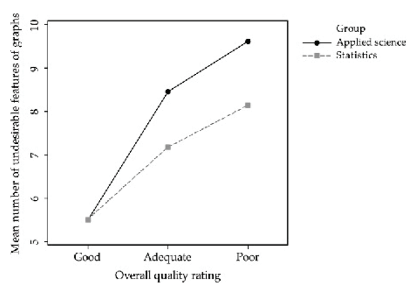
Make comparisons *direct*

- Use points not bars
- Connect similar by lines
- Same panel rather than different panels

Is there evidence of an interaction here?



???



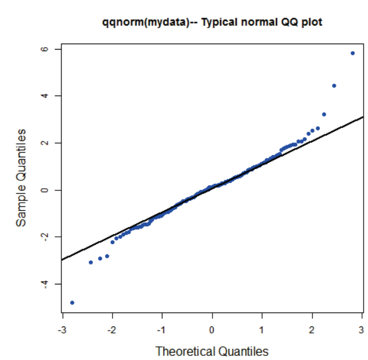
YES

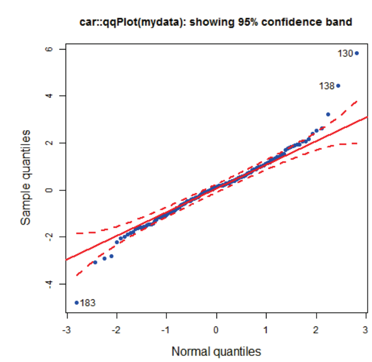
Published in: Ian Gordon; Sue Finch; *Journal of Computational and Graphical Statistics* **2015,** 24, 1210-1229.
DOI: 10.1080/10618600.2014.989324

Showing uncertainty

- Standard plots of observed vs. predicted lack a basis for assessment of uncertainty
- Confidence envelopes indicate extent of deviation
- Identify “noteworthy” observations to track them down

Example: Normal QQ plots used to assess normality of data



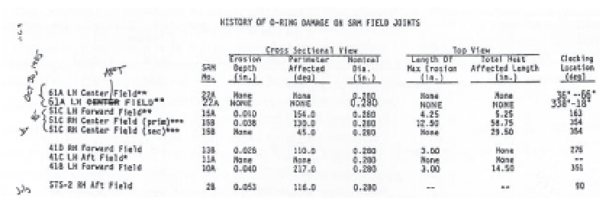


Effect ordering

- Information presentation is always ordered
  - in time or sequence (a talk or written paper)
  - in space (table or graph)
  - Constraints of time & space are dominant– can conceal or reveal the important message
- Effect ordering for data display
  - Sort the data by the effects to be seen
  - Order the data to facilitate the task at hand
    - lookup – find a value
    - comparison – which is greater?
    - detection – find patterns, trends, anomalies

Effect order failure: the *Challenger* disaster

- Few events in history provide as compelling illustration of importance of appropriate ordering and display of information
  - On January 28, 1968, the space shuttle Challenger exploded on take-off.
  - The cause was later determined to be that rubber O-rings failed due to cold weather
- Tables and charts presented to NASA by Thiokol engineers showed data from prior launches ordered by time (launch number), rather than by temperature—the crucial factor.
- The engineers' charts were also remarkable for information obfuscation: "erosion depth" (O-ring damage), "blow-by" (soot on O-rings), …

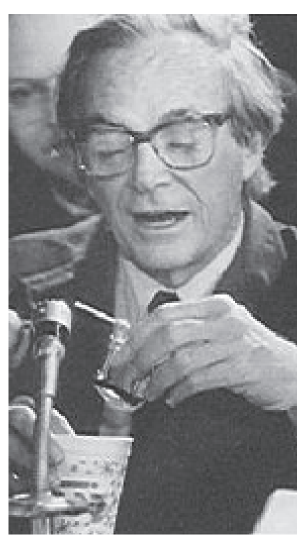


HISTORY OF O-RING DAMAGE ON SRM FIELD JOINTS

		Cross Sectional View			Top View		
	SRM No.	Erosion Depth (in.)	Perimeter Affected (deg)	Nominal Dia. (in.)	Length Of Max Erosion (in.)	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

Visual explanation: Physics

- NASA appointed members of the Rogers Commission to investigate the cause of the disaster
- the noted physicist Richard Feynman discovered the cause: at low temperature, O-rings became brittle and were subject to failure
- in his testimony, he demonstrated the effect by plunging a rubber O-ring into a cup of ice water

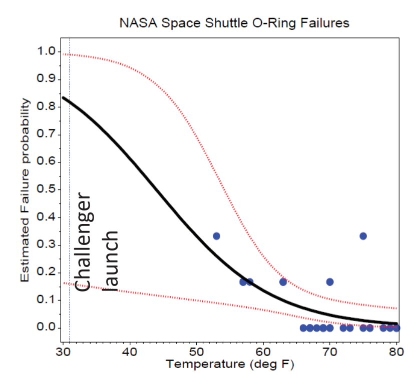


Visual explanation: Graphics

- Subsequent statistical analysis showed the relationship between launch temperature and O-ring failures
- As Tufte (1997) notes: the fatal flaw was in the ordering of the data.

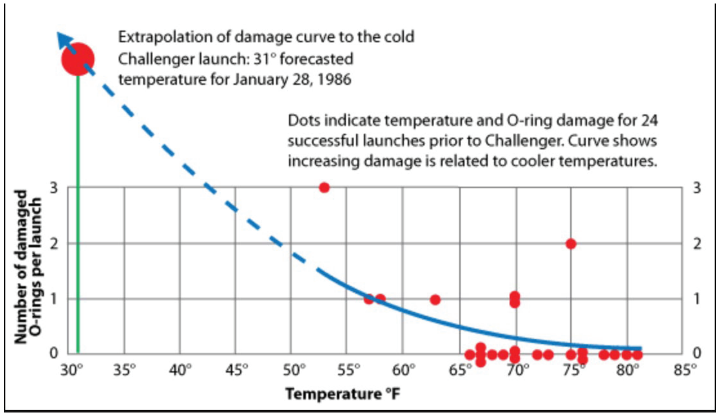
The graph shown here is the result of a statistical model fit to the data

- The **thick** line shows the predicted value of failure vs. temperature
- The red dotted lines show uncertainty of the predicted values



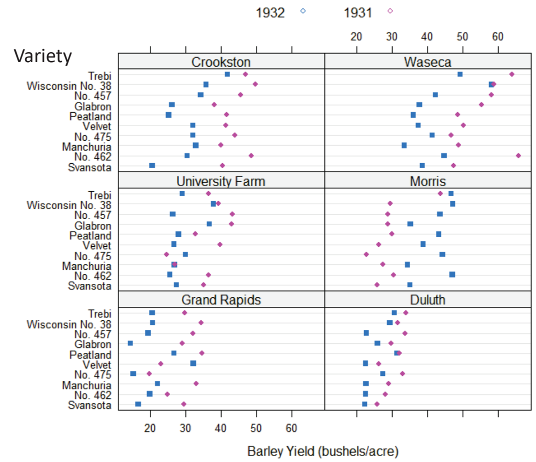
Presentation graphic

A presentation version of the previous graph alters the scales and describes the story in text annotations



Graphic displays: Main effect ordering

- To see trends, patterns, anomalies: **Sort unordered factors by means or medians**



Data on barley yields
10 varieties x 6 sites x 2 years

3 way dot plot, sorted by main effect means

- Which site has the highest yield?
- Which variety is highest on average?
- Which site stands out in pattern over year?

Tabular displays: Main effect ordering

- Tables are often presented with rows/cols ordered alphabetically
  - good for lookup
  - bad for seeing patterns, trends, anomalies

Table 1: Average Barley Yields (rounded), Means by Site and Variety

Variety	Site: Crookston	Duluth	Grand Rapids	Morris	University Farm	Waseca	*Mean*
Glabron	32	28	22	32	40	46	33.3
Manchuria	36	26	28	31	27	41	31.5
No. 457	40	28	26	36	35	50	35.8
No. 462	40	25	22	39	31	55	35.4
No. 475	38	30	17	33	27	44	31.8
Peatland	33	32	31	37	30	42	34.2
Svansota	31	24	23	30	31	43	30.4
Trebi	44	32	25	45	33	57	39.4
Velvet	37	24	28	32	33	44	33.1
Wisconsin No. 38	43	30	28	38	39	58	39.4
Mean	37.4	28.0	24.9	35.4	32.7	48.1	34.4

Tabular displays: Main effect ordering

- Better: sort rows/cols by means/medians
- Shade cells according to residual from additive model

Table 2: Average Barley Yields, sorted by Mean, shaded by residual from the model `Yield = Variety + Site`

Variety	Site: Grand Rapids	Duluth	University Farm	Morris	Crookston	Waseca	*Mean*
Svansota	23	24	31	30	31	43	30.4
Manchuria	28	26	27	31	36	41	31.5
No. 475	17	30	27	33	38	44	31.8
Velvet	28	24	33	32	37	44	33.1
Glabron	22	28	40	32	32	46	33.3
Peatland	31	32	30	37	33	42	34.2
No. 462	22	25	31	39	40	55	35.4
No. 457	26	28	35	36	40	50	35.8
Wisconsin No. 38	28	30	39	38	43	58	39.4
Trebi	25	32	33	45	44	57	39.4
Mean	24.9	28.0	32.7	35.4	37.4	48.1	34.4

Tabular displays: Main effect ordering

Yield difference, Δy_{ij} = 1931 – 1932 by Variety & Site

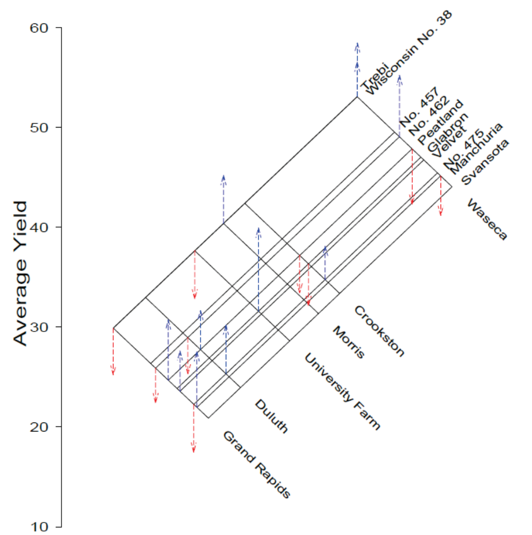
Ordered: by row and column means; **shaded:** by value ($| \Delta y_{ij} | > \{2,3\} \times \sigma(\Delta y_{ij})$)

What features stand out?

Table 3: Yield Differences, 1931-1932, sorted by mean difference, and shaded by value

Variety	Site						*Mean*
	Morris	Duluth	University Farm	Grand Rapids	Waseca	Crookston	
No. 475	-22	6	-5	4	6	12	0.1
Wisconsin No. 38	-18	2	1	14	1	14	2.4
Velvet	-13	4	13	-9	13	9	2.9
Peatland	-13	1	5	8	13	16	4.8
Manchuria	-7	6	0	11	15	7	5.5
Trebi	-3	3	7	9	15	5	6.1
Svansota	-9	3	8	13	9	20	7.3
No. 462	-17	6	11	5	21	18	7.4
Glabron	-6	4	6	15	17	12	8.0
No. 457	-15	11	17	13	16	11	8.8
Mean	-12.2	4.6	6.3	8.2	12.5	12.5	5.3

Graphical display: Two-way tables



Tukey two-way plot of average barley yield

If there is no interaction,

$y_{ij} = \mu + \alpha_{site} + \beta_{variety}$

Site & variety effects sorted automatically

Effects are spaced by fitted values

More variation among sites than varieties

Waseca best, by a wide margin

Multivariate data: correlation ordering

- Arrange ***variables*** so that:
  - Similar variables are contiguous
  - Ordered to show patterns of relations
- Arrange ***observations*** so that:
  - Similar variables are contiguous
  - Ordered to show patterns of relations

Correlation matrices

Baseball data: Batting, fielding and (log) Salary

Nobody wants to see all those decimals

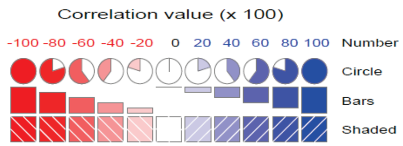
```
> cor(bb)
         Assists  Atbat   Errors   Hits    Homer  logSal Putouts    RBI   Runs  Walks   Years
Assists  1.0000 0.3421  0.70350 0.3040 -0.16160  0.0500 -0.0434 0.0629  0.179 0.1025 -0.0851
Atbat    0.3421 1.0000  0.32558 0.9640  0.55510  0.4149  0.3096 0.7960  0.900 0.6244  0.0127
Errors   0.7035 0.3256  1.00000 0.2799 -0.00974 -0.0208  0.0753 0.1502  0.193 0.0819 -0.1565
Hits     0.3040 0.9640  0.27988 1.0000  0.53063  0.4496  0.2997 0.7885  0.911 0.5873  0.0186
Homer   -0.1616 0.5551 -0.00974 0.5306  1.00000  0.3398  0.2509 0.8491  0.631 0.4405  0.1135
logSal   0.0500 0.4149 -0.02080 0.4496  0.33983  1.0000  0.2245 0.4441  0.426 0.4324  0.5374
Putouts -0.0434 0.3096  0.07531 0.2997  0.25093  0.2245  1.0000 0.3121  0.271 0.2809 -0.0200
RBI      0.0629 0.7960  0.15015 0.7885  0.84911  0.4441  0.3121 1.0000  0.779 0.5695  0.1297
Runs     0.1793 0.8998  0.19261 0.9106  0.63108  0.4256  0.2712 0.7787  1.000 0.6970 -0.0120
Walks    0.1025 0.6244  0.08194 0.5873  0.44045  0.4324  0.2809 0.5695  0.697 1.0000  0.1348
Years   -0.0851 0.0127 -0.15651 0.0186  0.11349  0.5374 -0.0200 0.1297 -0.012 0.1348  1.0000
```

If you are going to present the numbers, round a lot

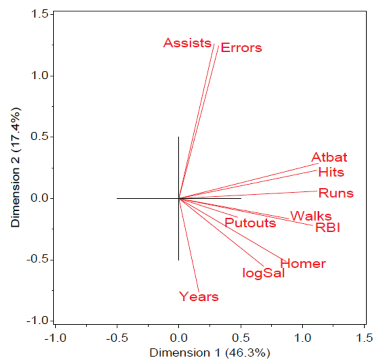
```
> round(100*cor(bb))
        Assists Atbat Errors Hits Homer logSal Putouts RBI Runs Walks Years
Assists     100    34     70   30   -16      5      -4   6   18    10    -9
Atbat        34   100     33   96    56     41      31  80   90    62     1
Errors       70    33    100   28    -1     -2       8  15   19     8   -16
Hits         30    96     28  100    53     45      30  79   91    59     2
Homer       -16    56     -1   53   100     34      25  85   63    44    11
logSal        5    41     -2   45    34    100      22  44   43    43    54
Putouts      -4    31      8   30    25     22     100  31   27    28    -2
RBI           6    80     15   79    85     44      31 100   78    57    13
Runs         18    90     19   91    63     43      27  78  100    70    -1
Walks        10    62      8   59    44     43      28  57   70   100    13
Years        -9     1    -16    2    11     54      -2  13   -1    13   100
```

Correlation ordering: corrgrams

Rendering: a correlation value can be displayed in different ways, for different tasks

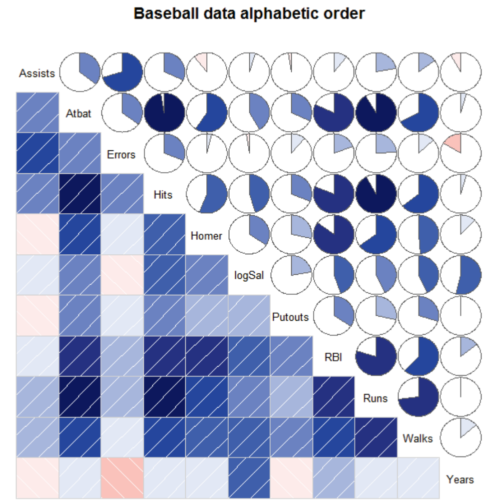


Correlation ordering:
- A PCA finds weighted sums of variable to maximize variance accounted for
- Angles between vectors reflect the correlations
- → Arrange variables in the order of their angles



Baseball data

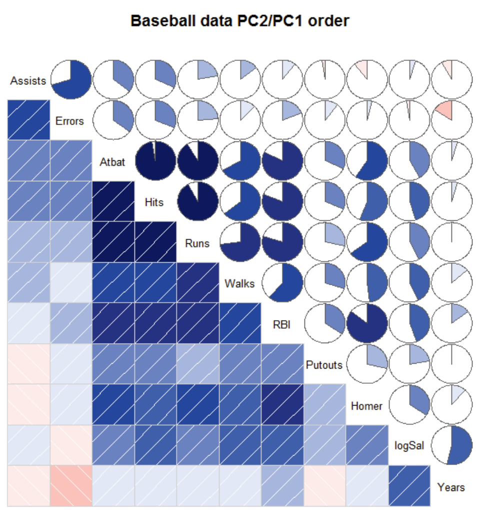
This is a corrgram display of the correlations among the baseball statistics, with the variables ordered alphabetically



Baseball data

The same display, with the variables sorted according to the angles between vectors in the PCA

Not that dramatic, but it isolates the positive & negative correlations

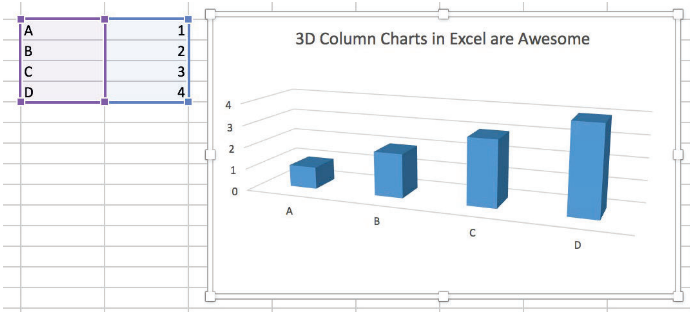


Graphs: Good/Bad, Excellent/Evil

- Like good writing, good graphical displays of data communicate ideas with:
  - clarity,
  - precision, and
  - efficiency– avoids graphic clutter
  - Even better: excellent graphs make the message obvious
- Like poor writing, bad graphical displays:
  - distort or obscure the data,
  - make it harder to understand or compare, or
  - thwart the communicative effect the graph should convey.
  - Even worse: evil graphs distort, or mislead.

Bad graphs are easy in Excel

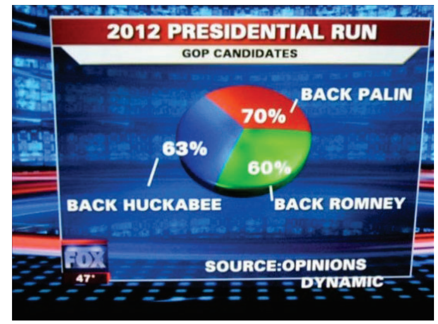
Friends don't let friends use Excel for data visualization or statistics



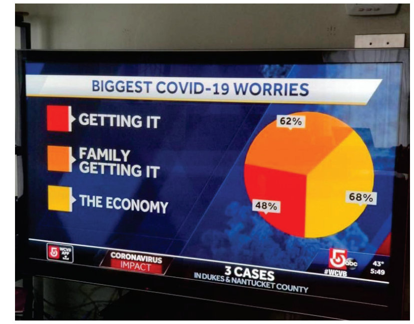
How many things are wrong with this graph?

Pie charts are easy to abuse

What's wrong with these pictures?



$1\ \pi = 193\%$



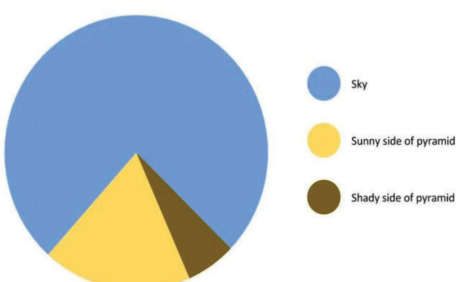
$1\ \pi = 178\%$

Why do graphic designers so often get this wrong?

Pie chart merriment

On the other hand, pie charts are a great source of merriment for people interested in graphics

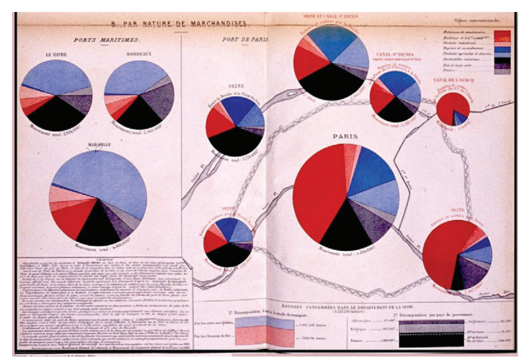
But: how much π have I eaten?



What perceptual ideas make this a great joke and lesson?

Once you see the pyramid, it's hard to see the pie.

But, can be used to great effect

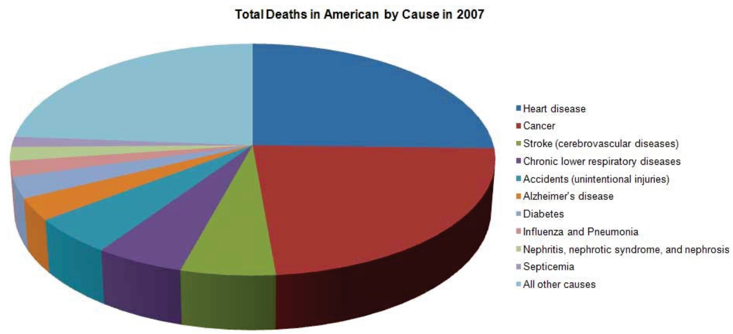


This graphic uses pie charts to show the transport of different kinds of goods to the ports of Paris and the principal maritime ports

- the size of each pie reflects total
- the sectors reflect relative %
- location places them in context

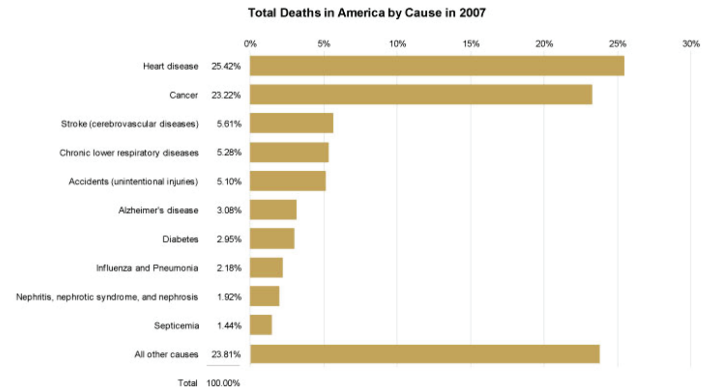
Album de Statistique Graphique, 1885, plate 17.

3D pie charts are usually evil



What was the intent of the designer of this graphic?
Which category led to the greatest total deaths?
What was the proportion of deaths due to strokes?
Did more people die from strokes vs. accidents?

Simple re-design makes it clearer

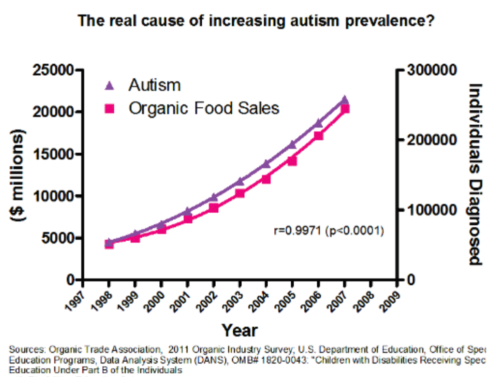


Double Y-axis: Really evil graphs

After pie charts, double Y-axis graphs have caused more trouble than almost any other

OMG, autism has been increasing directly with sales of organic food!

BAN ORGANIC FOOD!

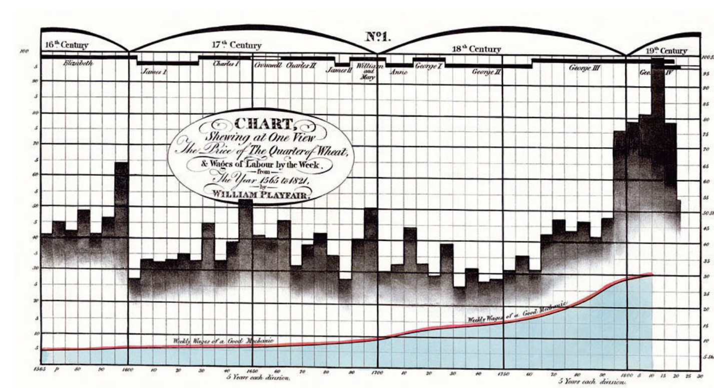


But, can be used to great effect

William Playfair invented the pie chart, line chart and bar chart.
In this figure, he shows 3 parallel time series over a 250-year period, 1560--1810

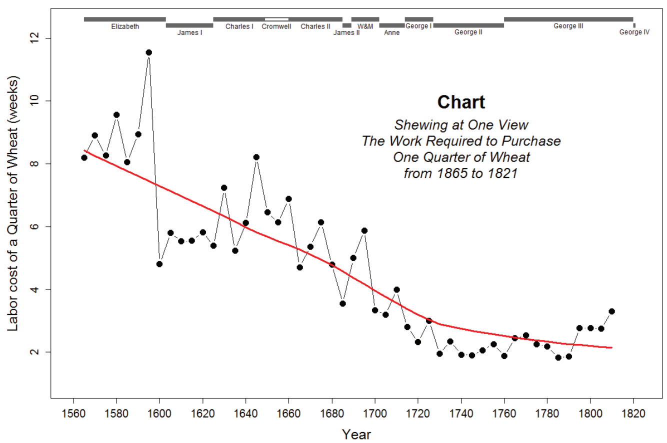
- weekly wages of a good mechanic
- price of wheat
- reigning monarch

Goal: show that workers were better off most recently (1810) than in the past



Or, another graph would have been better

A modern re-vision plots the ratio of price of wheat to wages directly



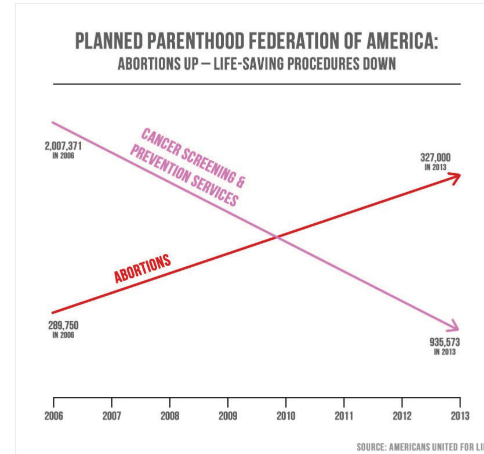
Even more evil: No scales, no data

Rep. Jason Chaffetz, R-Utah, sparred with Planned Parenthood president Cecile Richards during a high-profile hearing on Sept. 29, 2015 and presented this graph.

"*In pink, that's the reduction in the breast exams, and the red is the increase in the abortions. That's what's going on in your organization.*"

Created by an anti-abortion group it is a deliberate attempt to mislead.

Can you see why?



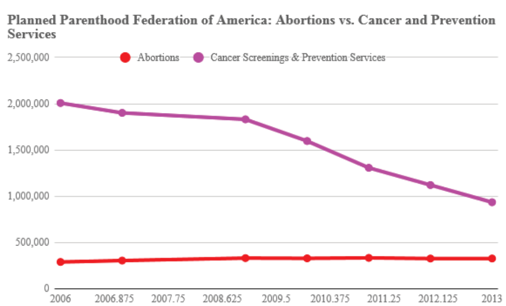
See: http://www.politifact.com/truth-o-meter/statements/2015/oct/01/jason-chaffetz/chart-shown-planned-parenthood-hearing-misleading-/

Corrected graph

This graph shows the actual data from the Planned Parenthood reports used by Americans United for Life

The number of abortions was relatively steady.

Some services like pap smears, dropped due to changing medical standards about who should be screened and how often.

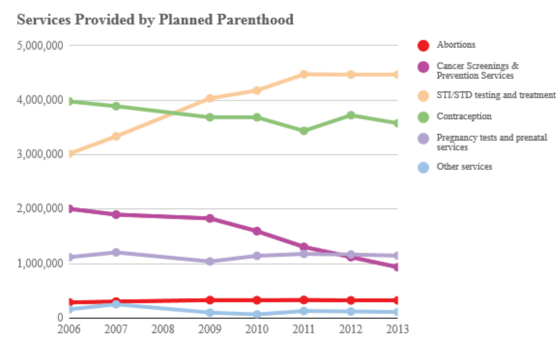


What are a few improvements that could be made to this graph?

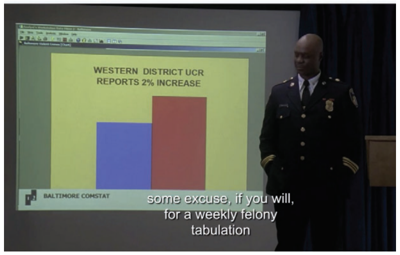
Corrected graph, in context

Showing a wider range of PP activities puts these data in context

PP activities were far higher for contraception and STD testing



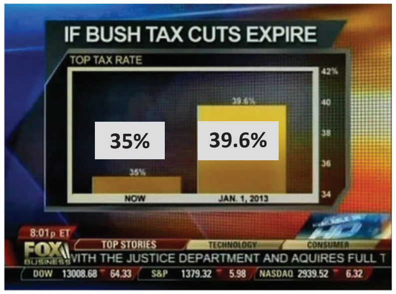
Evil Bars



You can say anything you want if you don't show a scale for the vertical axis

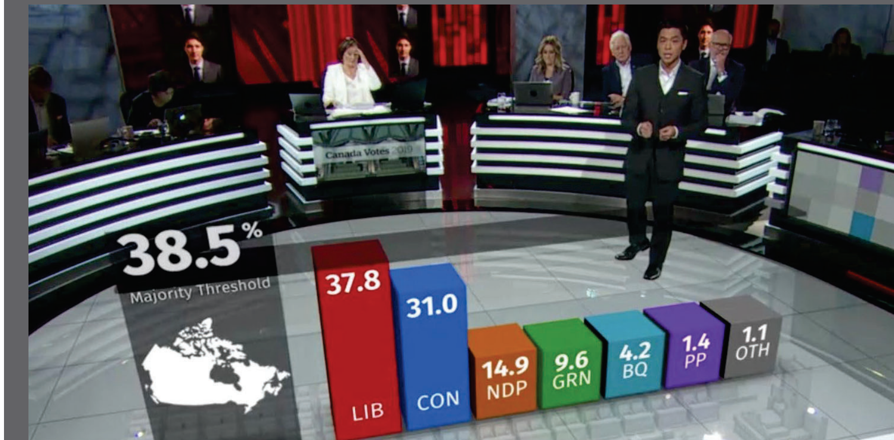
Q: Do people judge the difference in heights or the ratio of heights?

A: It depends on the question



You can greatly distort the perception of difference or ratio by truncating the Y axis.

Y-axis truncation is/was the default in Excel!



More evil: 3D bars

- CBC found it irresistible to make 3D bars to show the 2015 election projection.
- Why do you think the smallest 5 bars are all the same height?

Graphical Excellence: Tables

A study by Abigail Friendly (2017) wanted to show the use of benefits afforded to Toronto developers for their contributions of different types over time

Figure 9: Section 37 benefits by type (1998–2015)

	1998–2002	2003–2005	2006–2009	2010–2013	2014–2016
Roads, streetscapes	30	35	54	83	15
Culture, community, recreation	26	50	59	47	16
Parks	27	41	41	52	20
Affordable housing	17	26	38	56	11
Public art	26	25	41	32	4
Heritage	16	13	26	18	3
Transit	11	7	10	20	3
Libraries	6	2	5	11	1
Other	3	6	7	8	3

Scale
0 - 10
11 - 20
21 - 30
31 - 40
41 - 50
51 - 60
61 - 70
71 - 80
81 - 90

Color background scale from light to dark highlights the largest values

Most frequent benefits appear at the top

Can see overall trends and anomalies

What happened in 2014-2016?

Source: Friendly, A. R. (2017). *Land Value Capture and Social Benefits: Toronto and São Paulo Compared*. IMFG Papers on Municipal Finance and Governance, No 33, University of Toronto, https://munkschool.utoronto.ca/imfg/

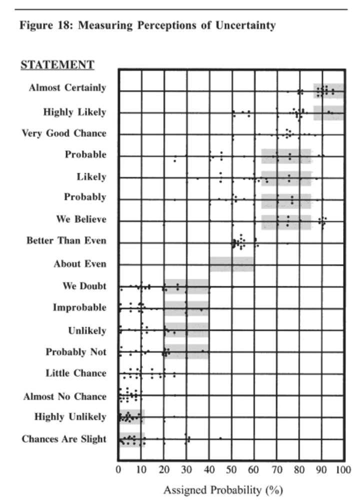
Graphical failure

This graph reports the results of a survey by Sherman Kent for the CIA with the question:

What [probability/number] would you assign to the phrase "[phrase]"

The goal was to contribute to an understanding of how intelligence analysts use these terms

Why can this be considered a graphical failure?



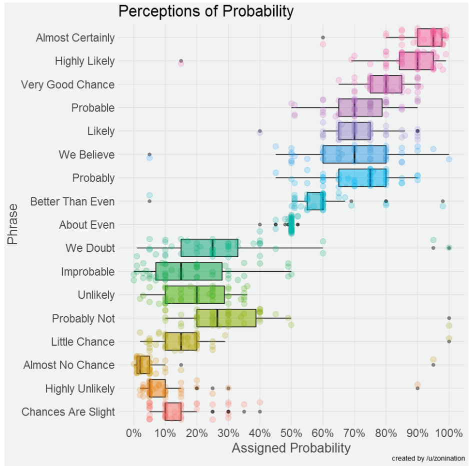
https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/psychology-of-intelligence-analysis/fig18.gif/image.gif

Graphical excellence

This graph shows the same data, as both dotplots & boxplots

We can see a lot more:
- "about even" has very low variability
- the last 3 categories are listed out of order
- the extreme outliers stand out
- skewness is – for high probability, + for low probability

Technical notes:
- software: ggplot2
- design: faint grid lines
- color: points use transparent color & jittering; outliers also shown in black



From: https://github.com/zonination/perceptions

Graphical excellence

This graph uses "ridgeline" plots to show the same data

Each one is a small version of a density plot showing a smoothed version of the distribution

Stacking them in this way allows center, variability, shape and other features to be readily compared.

Color & transparency are used effectively

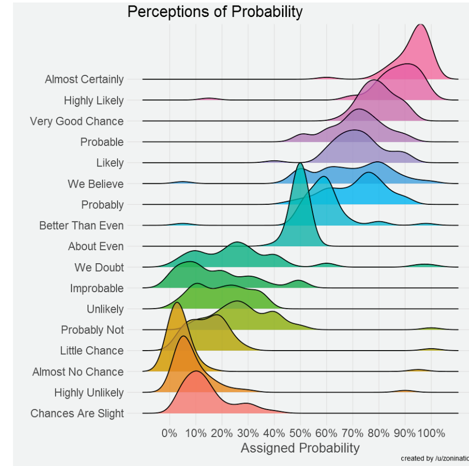


Chart junk or effective info vis?

Charts can be offensive and/or effective

What is the message?
Who is the audience?

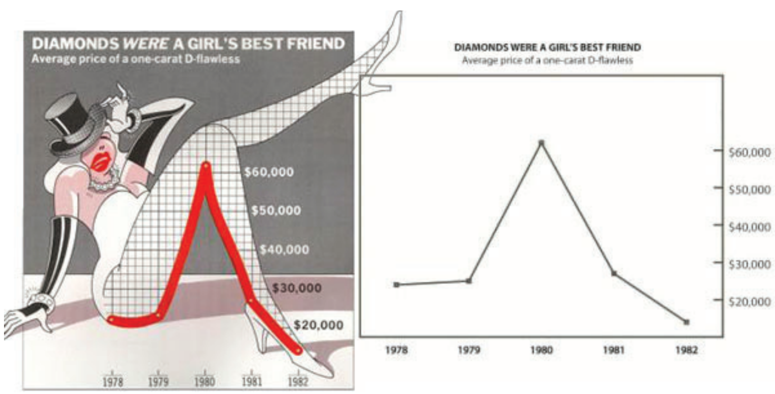


Chart junk or effective info vis?

Suzana Herculano-Houzel has a new method for determining counts of cortical neurons across different species. How to present this effectively?

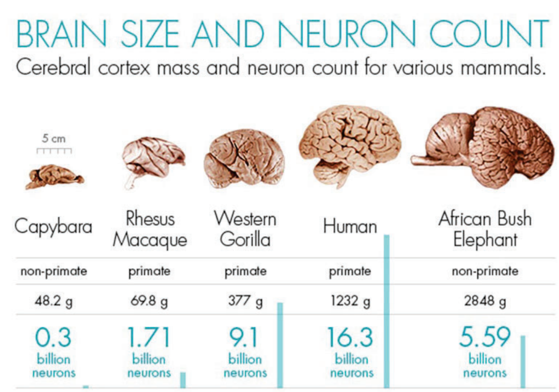
Goal: compare mammal species brain size and cortical neuron count

Neuron count is shown both as numbers and bars

Claim: Human brain is ~ linear of primate brains

What do you think?

How could this be made better?



From: Herculano-Houzel, "The human brain in numbers: a linearly scaled-up primate brain"
https://www.frontiersin.org/articles/10.3389/neuro.09.031.2009/full

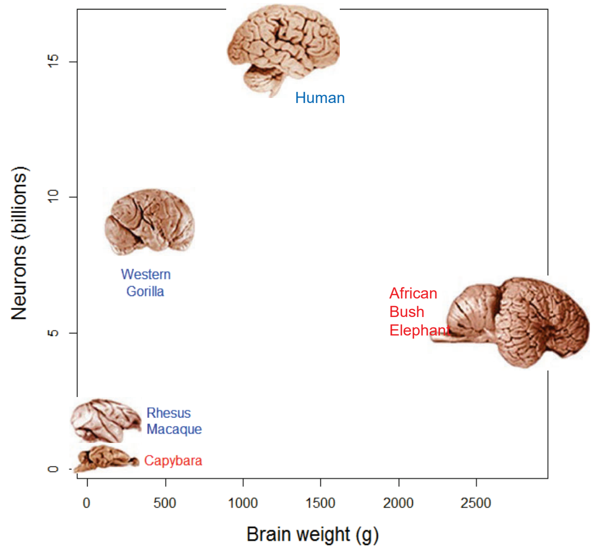
As a scatterplot

A scatterplot makes clear how humans differ from other species

- Using scaled images as point symbols also conveys brain size
- Primates are distinguished from non-primates by text color

This is arguably a more effective display.

What do you think?

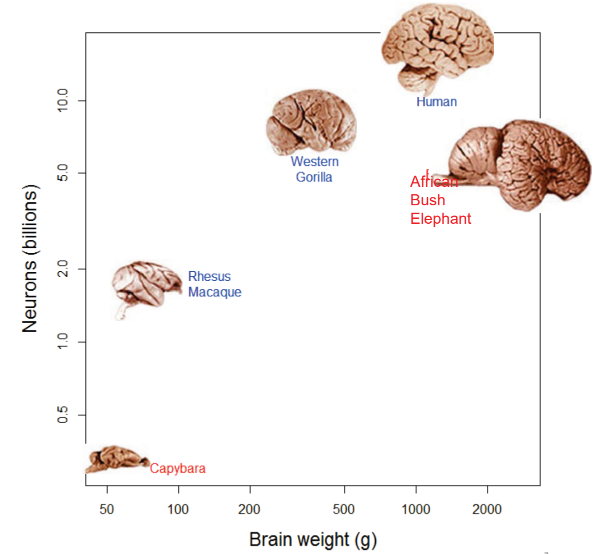


As a scatterplot – log scale

Perhaps even better is to make the plot using log scales for both axes

The relationship is now easier to see, but only approx. linear

The argument for neurons ~ brain weight needs more work.



Why graphs matter: Climate change

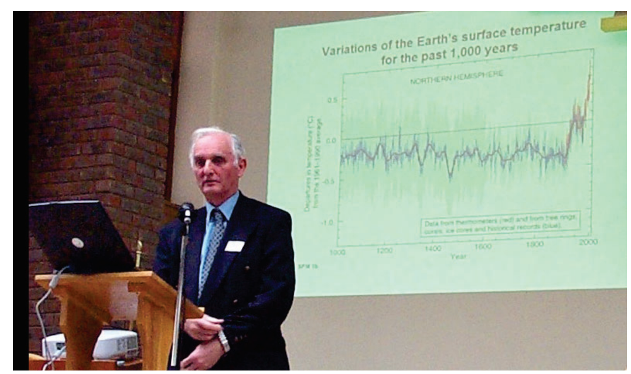
In the movie, *An Inconvenient Truth* (2006), Al Gore used the now-famous "hockey stick" graph to show that human activities had greatly increased the degree of global warming over the recent past

The goal was to raise public awareness and call for action to curb environmental effects: CO_2 emissions as the main agent.

Movie: https://www.youtube.com/watch?v=8ZUoYGAI5i0; http://www.imdb.com/title/tt0497116/

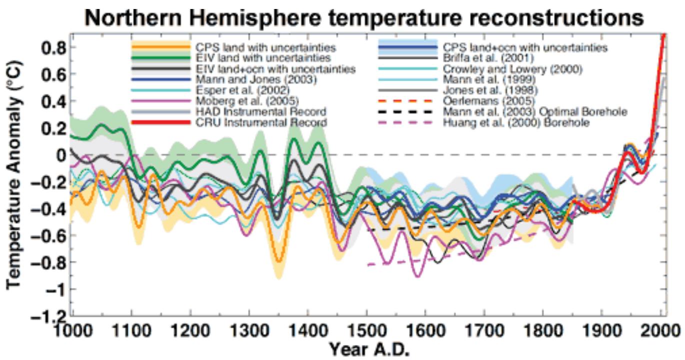
Climate change: Original graph

Sir John Houghton presents the original Northern Hemisphere hockey stick graph to the Intergovernmental Panel on Climate Change (IPCC) in 2005.
It is based on an analysis by Mann, Bradley & Hughes (1990), with a smoothed curve and uncertainty intervals.



Climate change: data sources

The MBH (1999) paper had used a wide variety of data sources. They were combined using a novel statistical technique, the first eigenvector-based climate field reconstruction (CFR).
Climate scientists understood this; the sceptics did not.



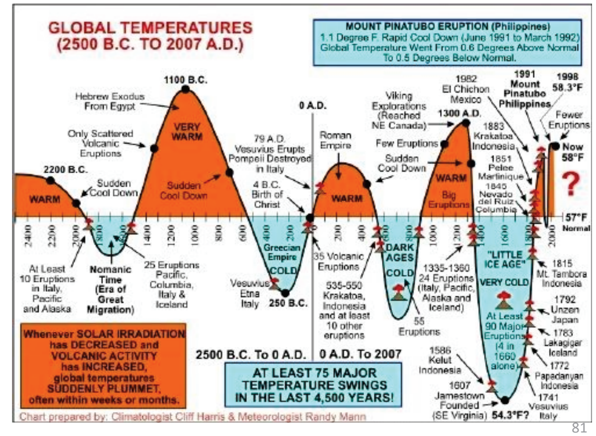
See: https://en.wikipedia.org/wiki/Hockey_stick_controversy for details

Countering climate change

Taking a longer view, and adding a lot of extraneous historical details, climate sceptics were easily able to mount alternative explanations: Solar irradiation & volcanoes

How to mislead:

- Show no temp. scale
- Draw smoothed curve
- Suggest that all is due to "swings" in temp.
- Compress recent history into the end of the time scale

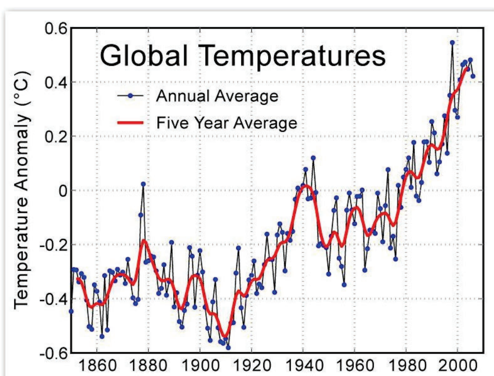


Time scale

Perhaps one fault with the original graphs was trying to show noisy data, from many sources, over too wide a time span.

What could you do to make this graph even more convincing?

- Add an overall smooth curve
- De-emphasize the annual data



Climate change: Infographic

A politically-incorrect graphic shows very clearly the effect of global warming on panty size

Source: http://www.politically-incorrect-humor.com/2010/03/positive-proof-of-global-warming

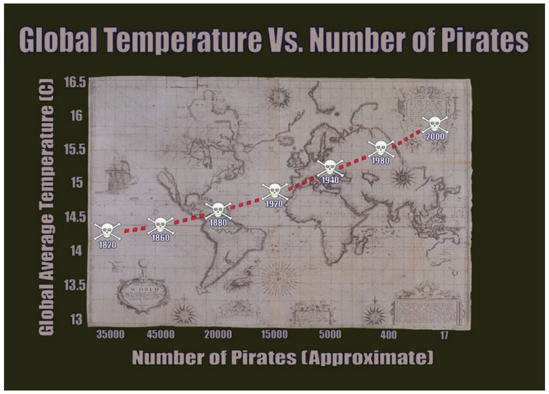
Climate change: other explanations

This infographic attempts to relate global warming to the decrease in pirates

Aside from the substance, how many things are wrong about this graphic?

Simple explanation:
Lack of pirates causes global warming!

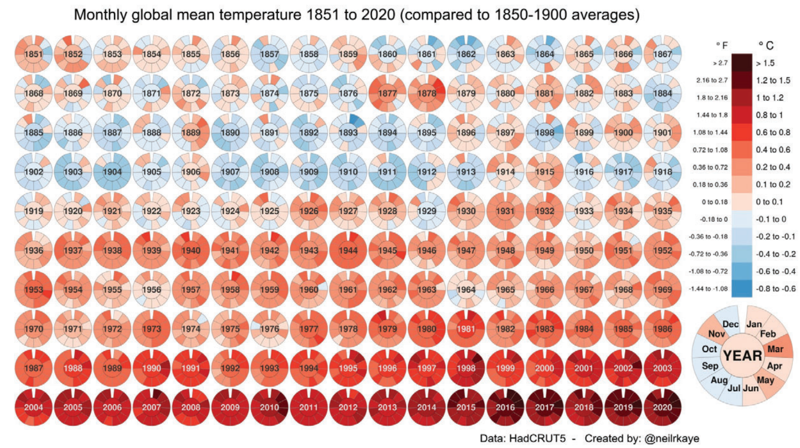
Conclusion:
To stop global warming, become a pirate!



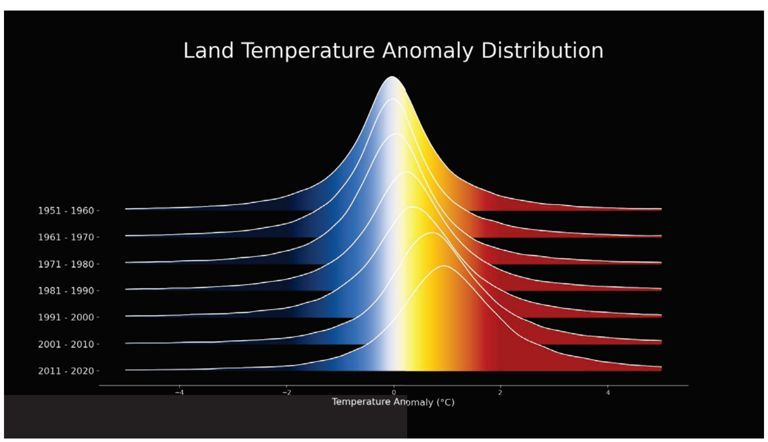
Source: http://www.forbes.com/sites/erikaandersen/2012/03/23/true-fact-the-lack-of-pirates-is-causing-global-warming

Circle graphs

What features makes this graph effective?



Ridgeline plot



Some graphic rules

- Bars
  - Don’t cut off their feet
  - Don’t add dynamite fuses
- Pies
  - Generally, best preserved for dessert
  - Well used for part-whole relations with a small number of categories
  - Better used as a graphic form in larger displays
- Axes
  - Avoid double Y-axes
  - Don’t truncate without considerable thought
- 3D
  - Avoid for useless “glitz”

Summary

- Graphs as a form of communication
  - Data (numbers), words, images $\rightarrow$ Stories
- Analysis graphs vs. presentation graphs
- Some principles of effective data display
  - Make the data stand out
  - Facilitate comparisons
  - Effect ordering