

Psychology of Data Visualization: Course Overview

Michael Friendly
Psych 6135

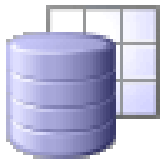


<http://euclid.psych.yorku.ca/www/psy6135/>

@datavisFriendly

Data, pictures, models & stories

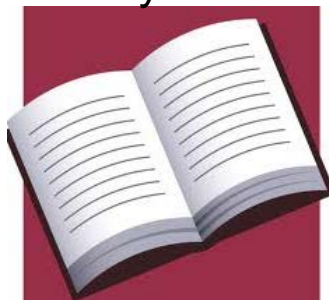
Goal: Tell a credible story about
some real data problem



data

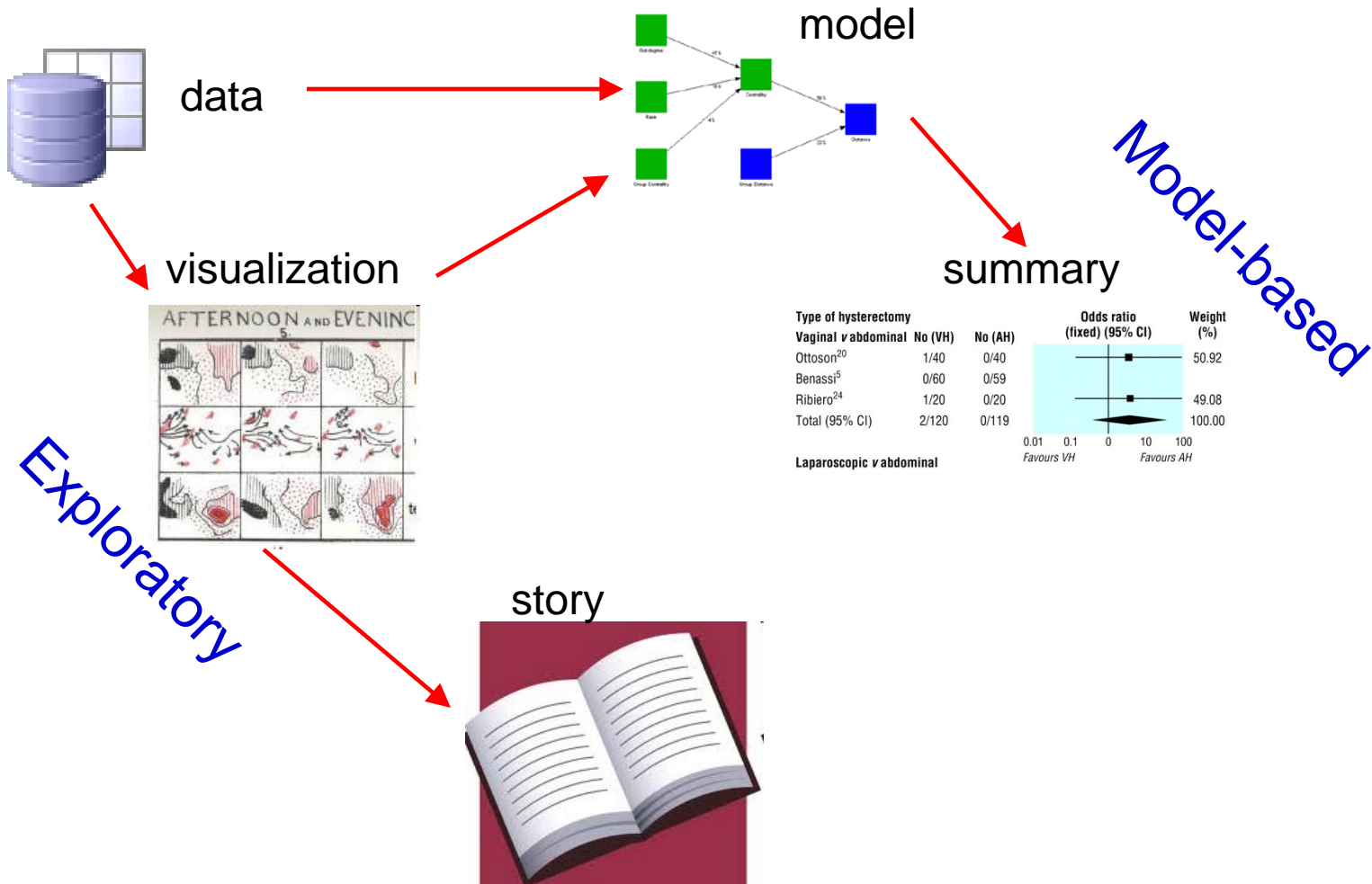
Measles vaccination
Global warming
...

story



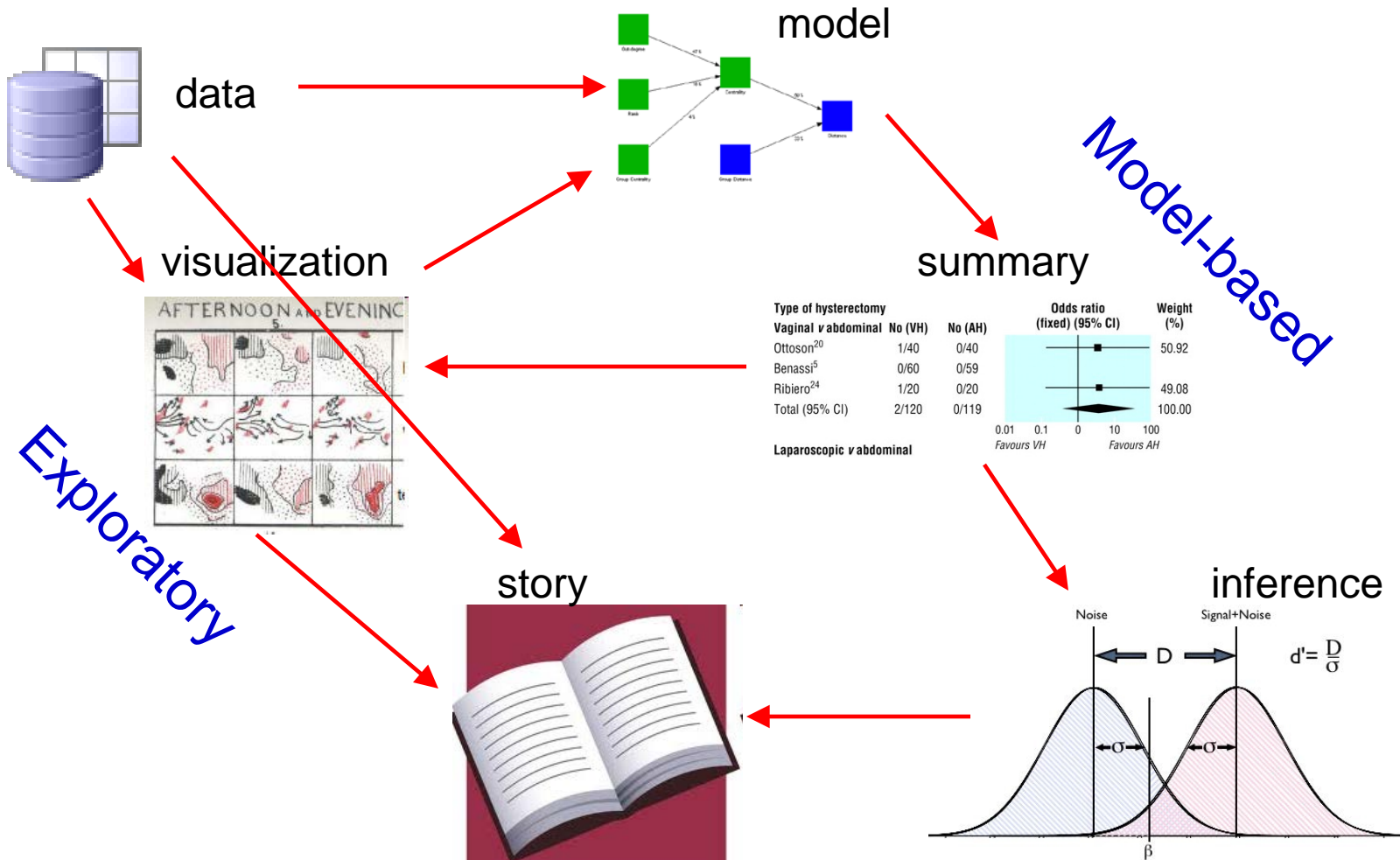
Data, pictures, models & stories

Two paths to enlightenment



Data, pictures, models & stories

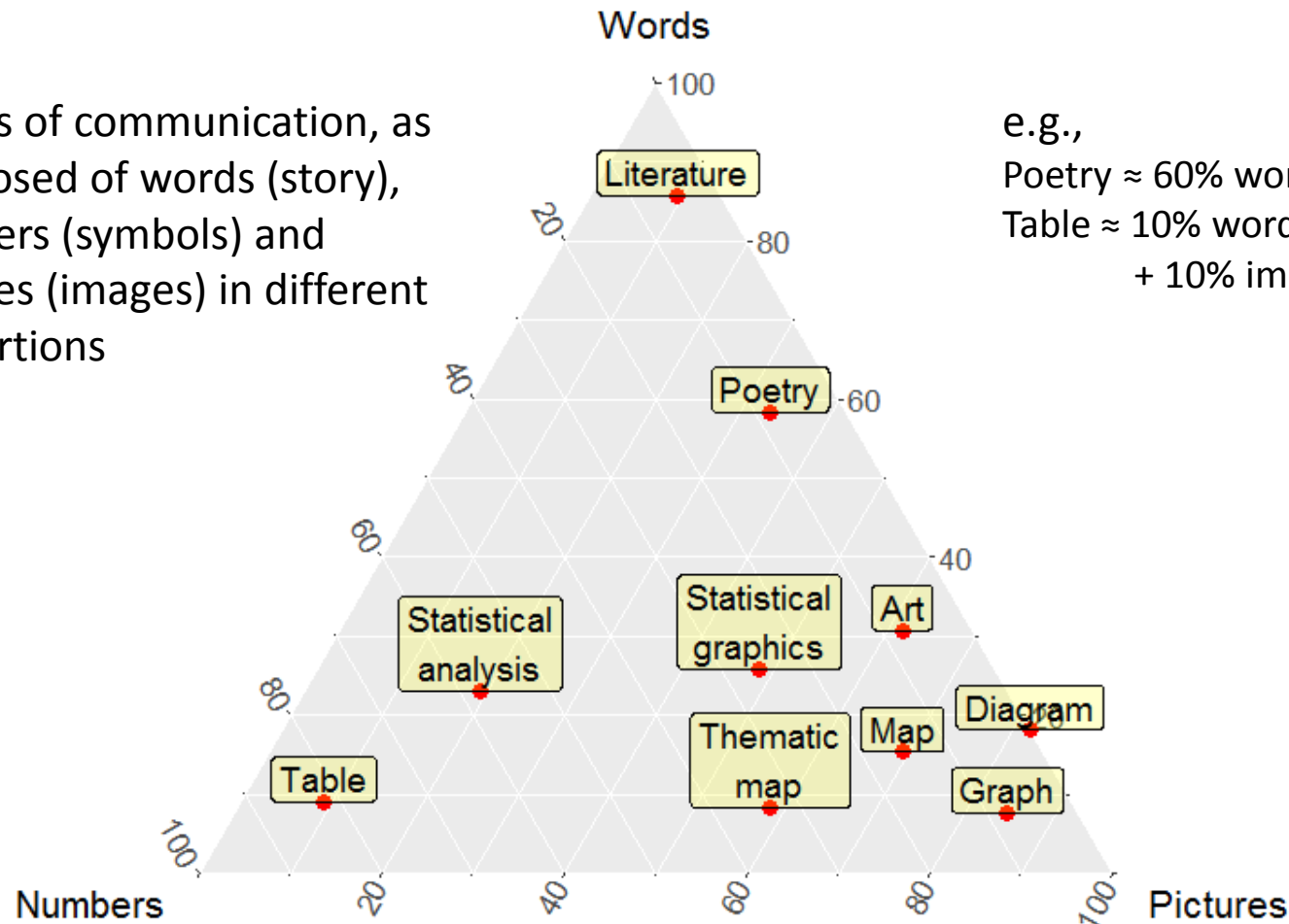
Now, tell the story!



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



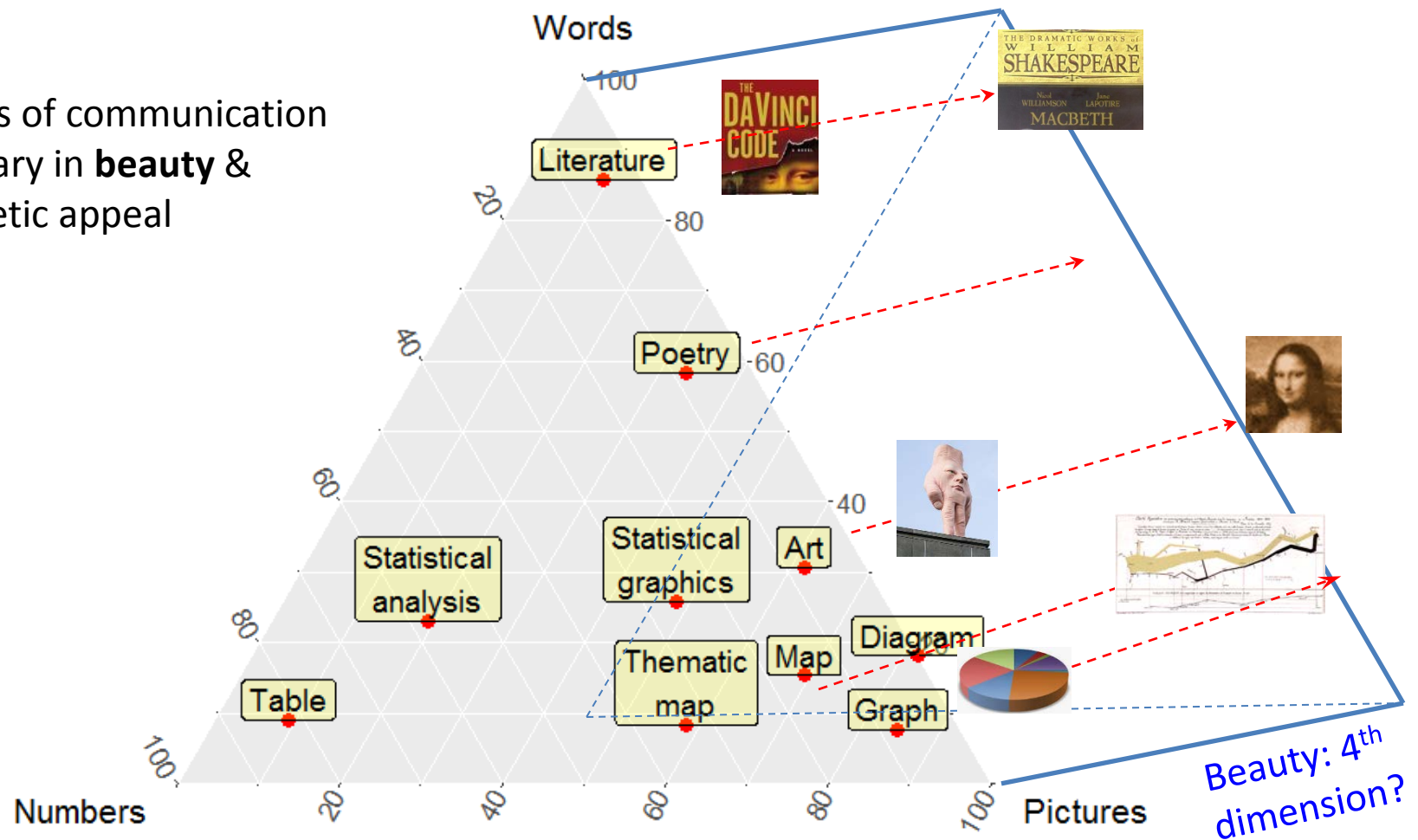
e.g.,

Poetry \approx 60% words + 40% images
Table \approx 10% words + 80% numbers
+ 10% images

Words, numbers and pictures

Beauty: The 4th dimension

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



Beauty: 4th dimension?

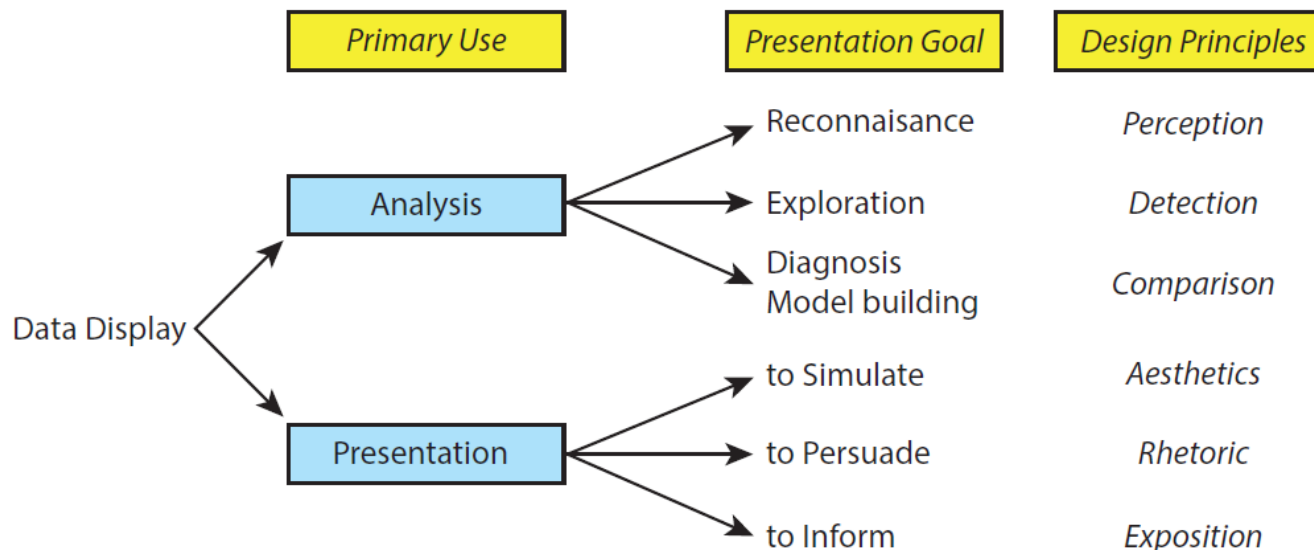
Roles of graphics in communication

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

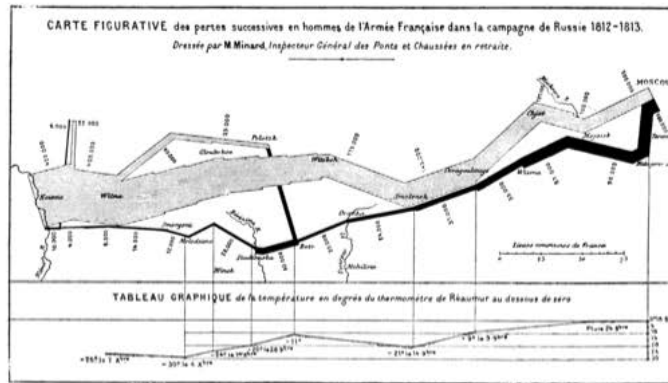
Analysis graphs: design to see patterns, trends, aid the process of data description, interpretation

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

Basic functions of data display



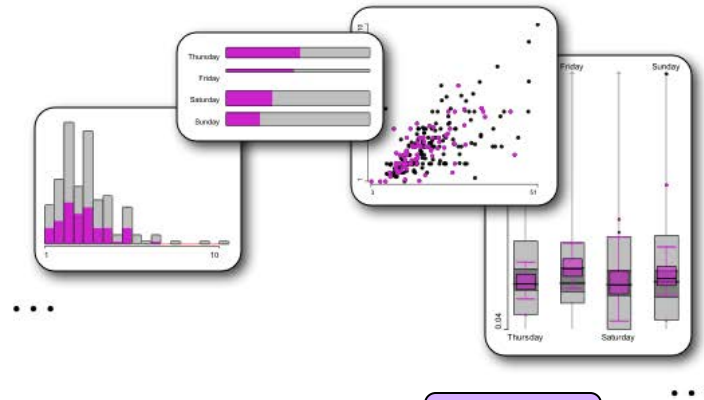
Different graphs for different purposes



Wow!

Presentation

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



Ah ha!

Exploration

Goal: the Ah ha! Experience
Many images, for a narrow audience
(you!), linked to analysis

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

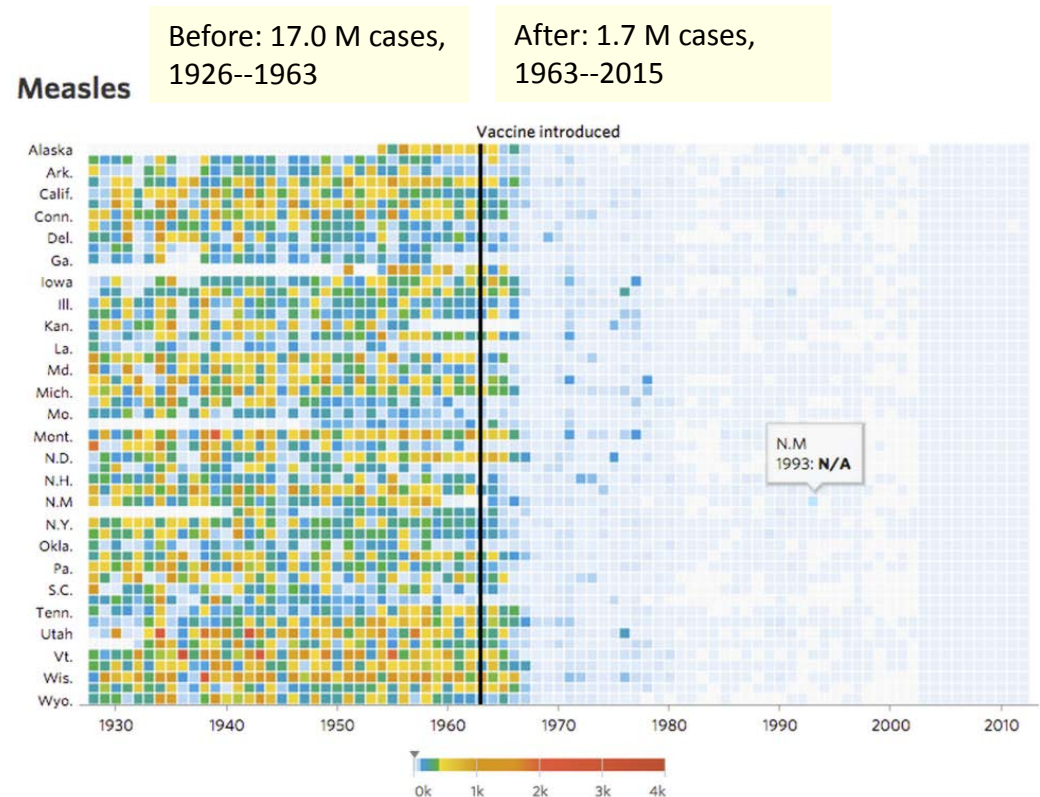
Measles was decimated!

The message hits you between the eyes!

Powerful graphs make comparison easy

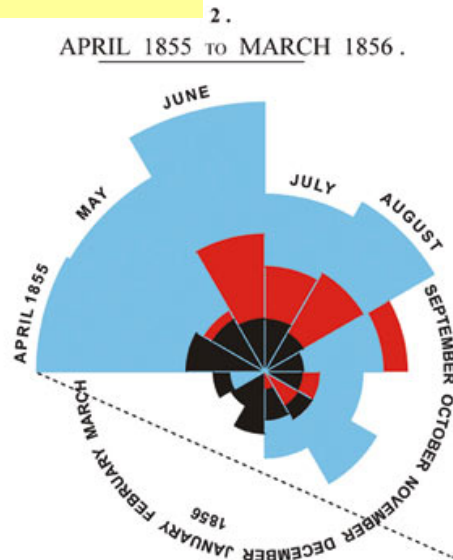
In 2014, vaccination rates declined and measles re-emerged in those areas

Effective graphs can cure ignorance, but not stupidity.



Presentation graph: Nightingale (1857)

After reform



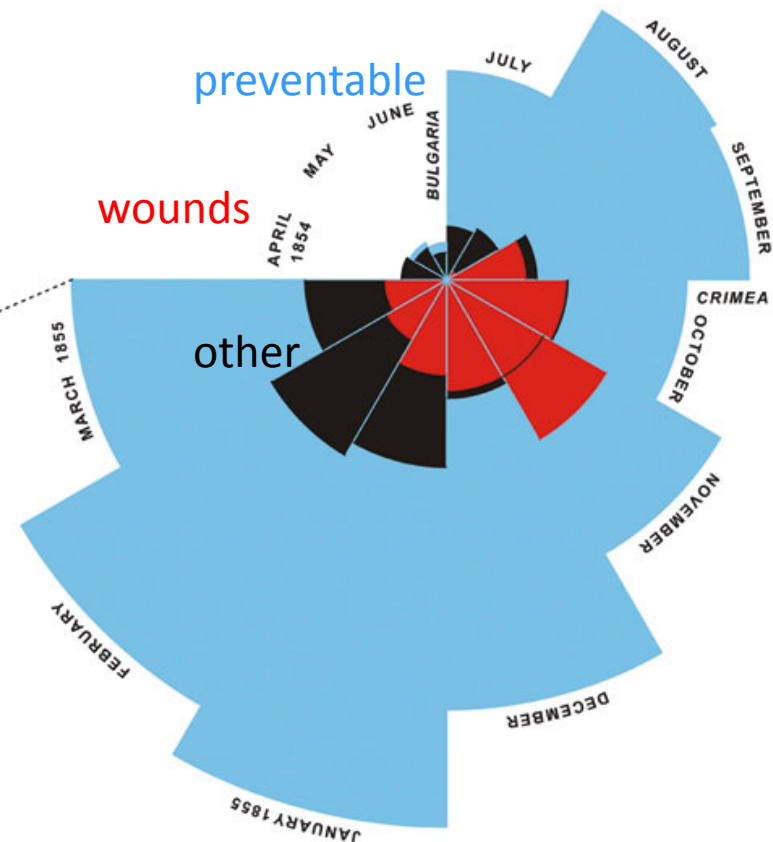
*The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic Diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes*

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

DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.

Before reform

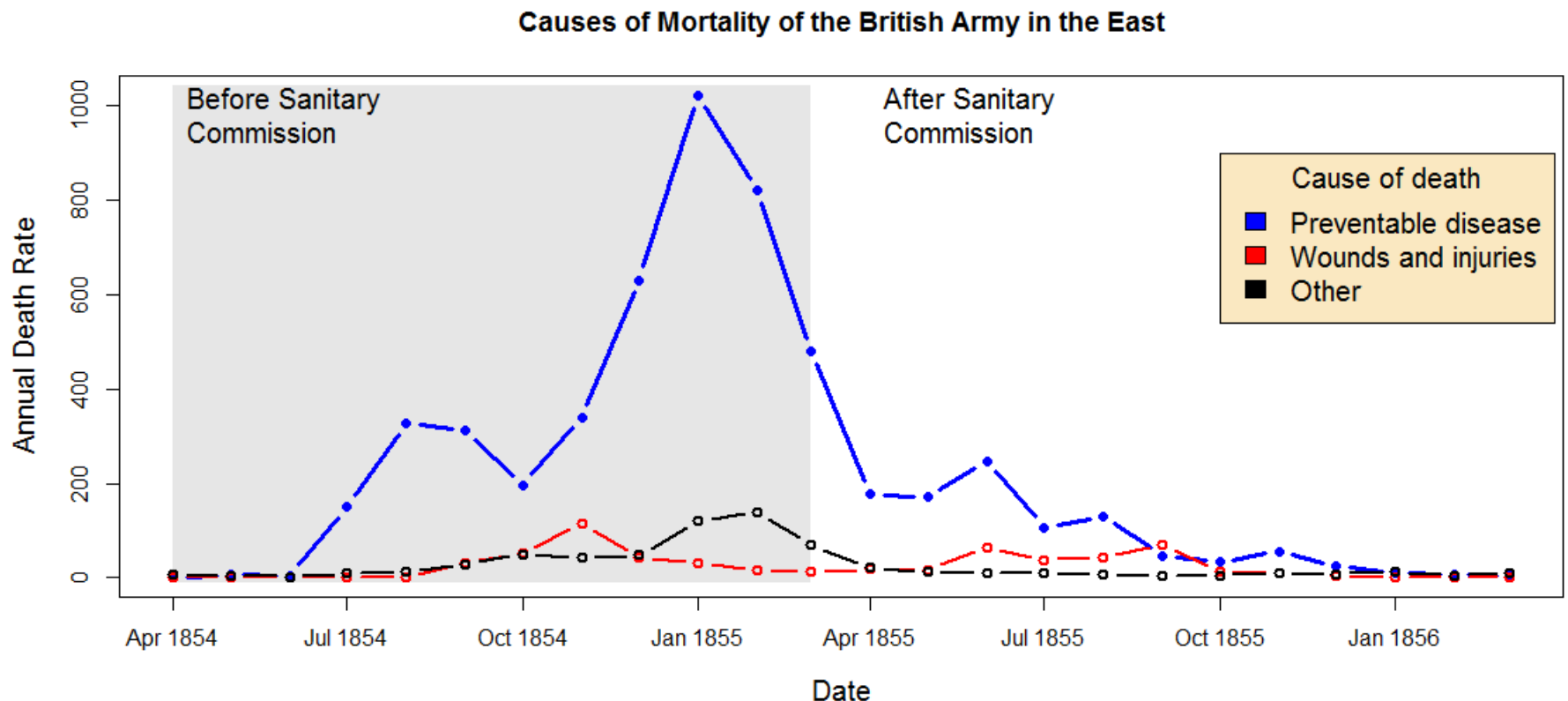
APRIL 1854 TO MARCH 1855.



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.



Analysis graph: Deaths vs. Income

Scatterplot of deaths vs. income

- Loess smooth + CI band
- Labels: year
- Color: party in power

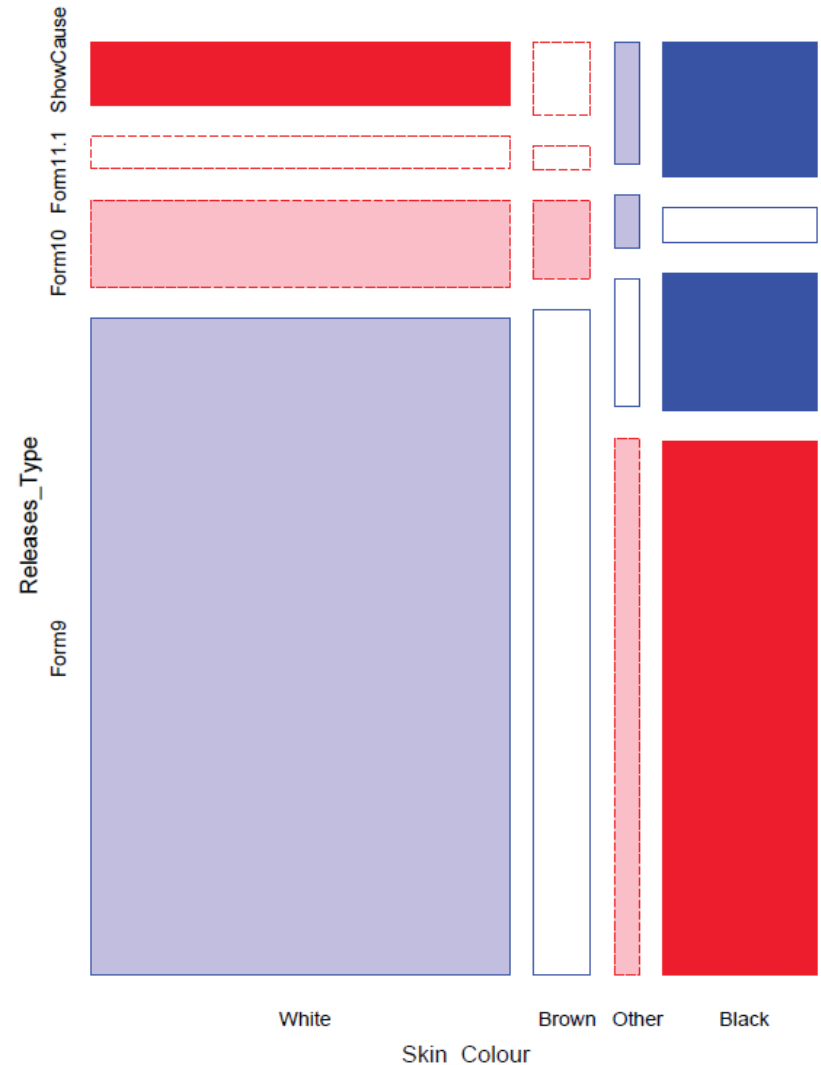
The message here is interesting, but it lacks the power and eloquence of the original graph

As well, the relationship of deaths to time & party is lost



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 \sim frequency
 - shading: \sim residual
 - Obs > Expected in blue
 - Obs < Expected in red



Racial profiling: Presentation graphic

Together, we created this **self-explaining** infographic

Title gives the main conclusion

Text description gives details

Bar width ~ charges
Divided by % release

numbers shown in the cells

Legend gives a layman's description of shading levels

Same charge, different treatment

Statistical analysis of single drug possession charges shows that blacks are much less likely to be released at the scene and much more likely to be held in custody for a bail hearing. Darker colours represent a stronger statistical link between skin colour and police treatment.

Degree of likelihood

- Much less* likely to occur
- Much more* likely to occur
- More likely* to occur

Whites are more likely to be released at the scene

6,662
charges
laid

78%
released at the scene

14.5%
released
at station

7.5%
held
for
bail

Blacks are much more likely to be held for bail hearings

2,446
charges
laid

64%
released at the scene

20%
released at station

16% held
for bail

0% 10 20 30 40 50 60 70 80 90 100

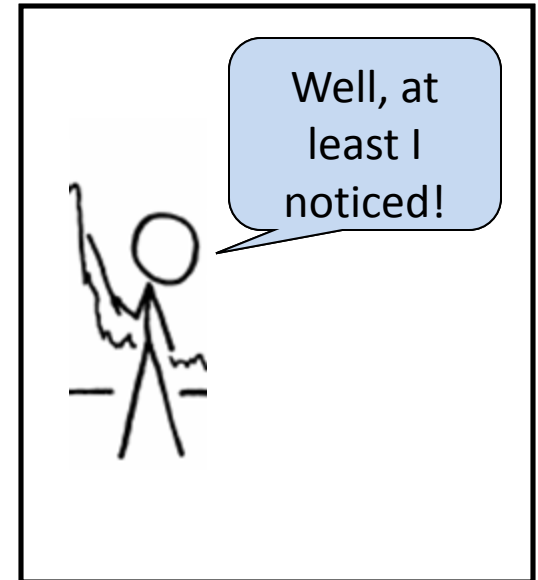
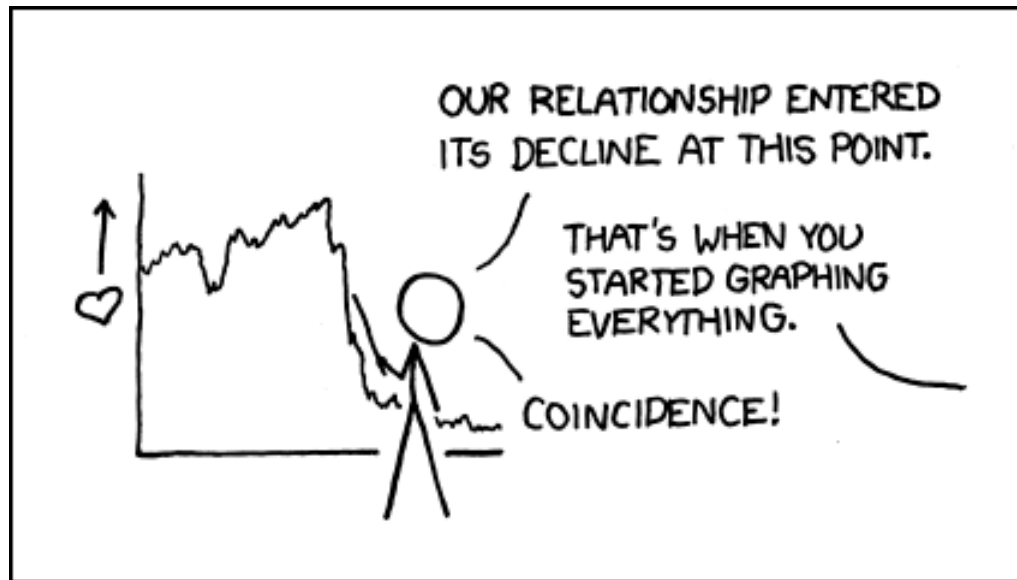
SOURCE: Toronto police arrest records 1996-2002

Why plot your data?

Graphs help us to see

patterns, trends, anomalies and other features

not otherwise easily apparent from numerical summaries.

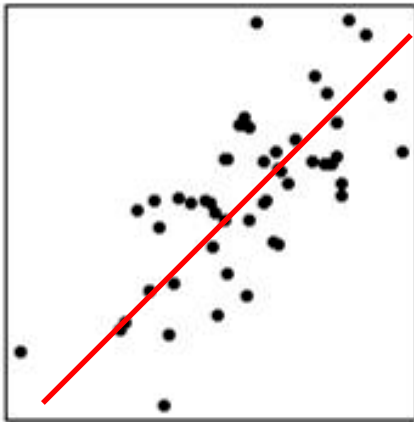


Source: <http://xkcd.com/523/>

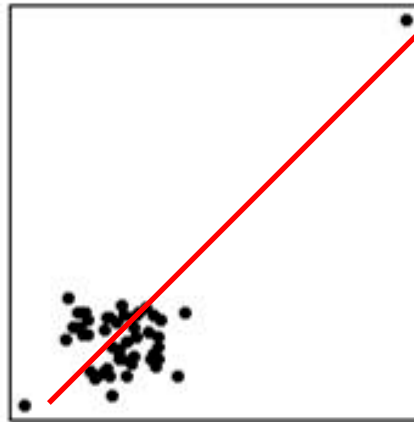
Why plot your data?

Three data sets with exactly the same bivariate summary statistics:

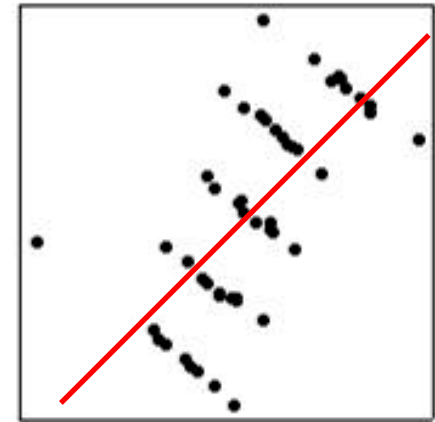
- Same correlations, linear regression lines, etc
- Indistinguishable from standard printed output
- Totally different interpretations!



Standard data



$r=0$ but + 2 outliers



Lurking variable?

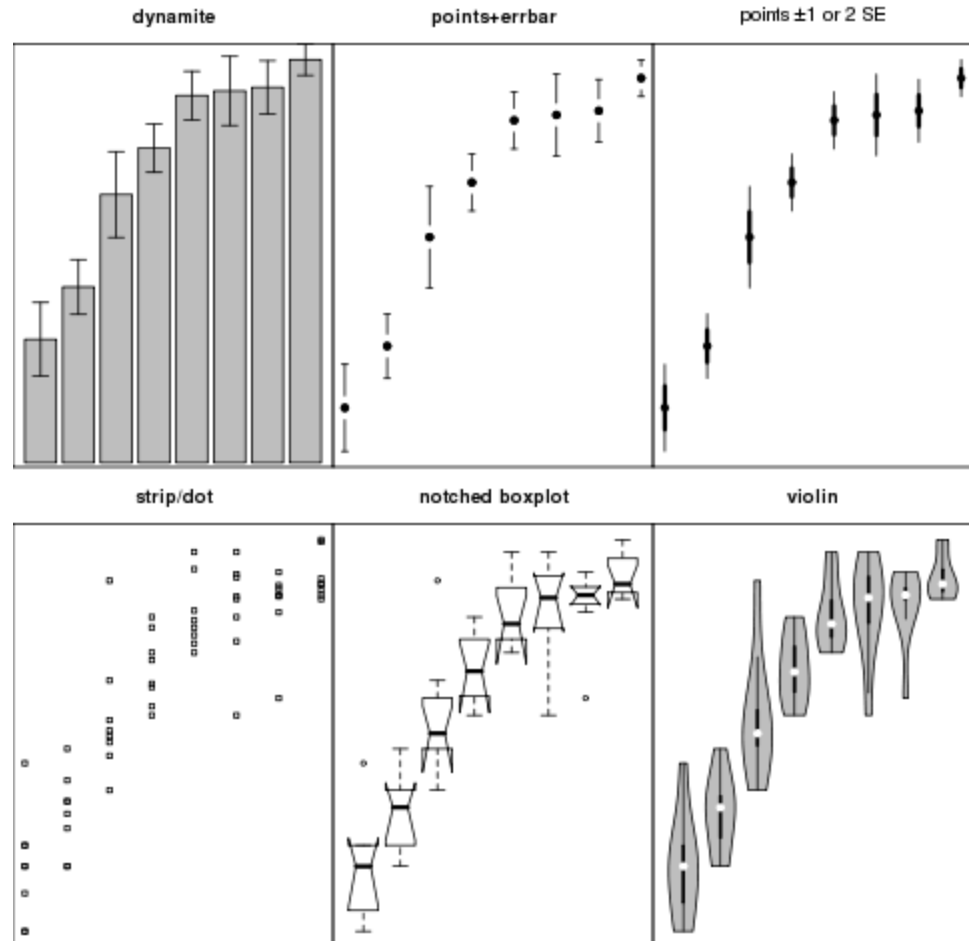
Effective data display

- 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

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

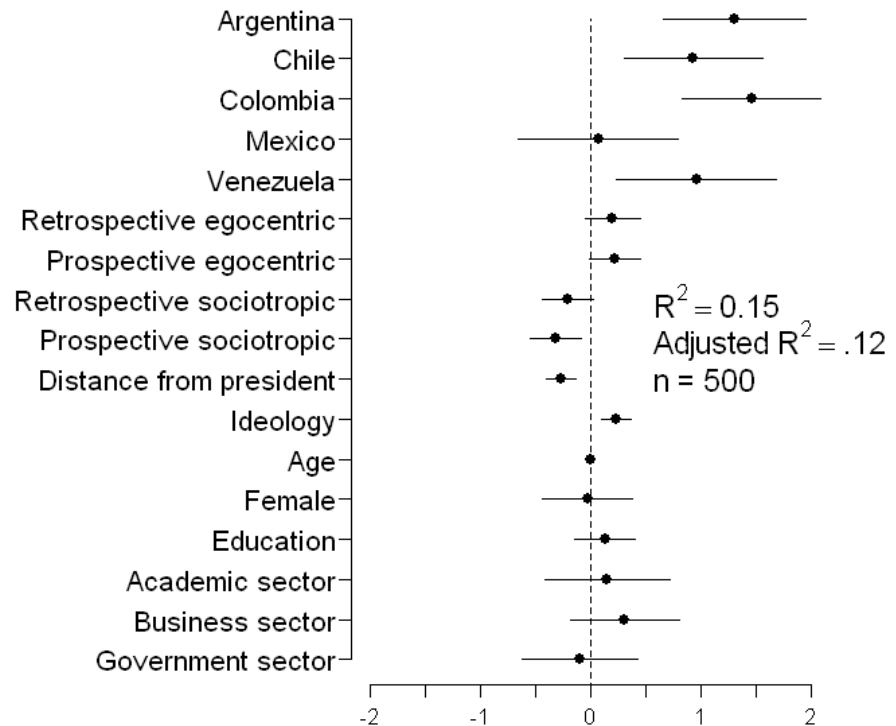
Consider tradeoff between
summarization & exposure

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 Venezuela's at p < .05	

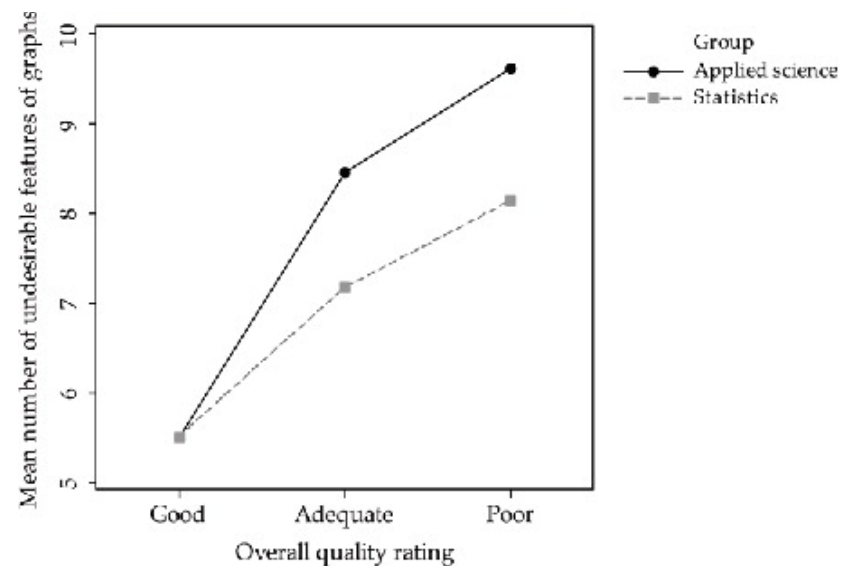
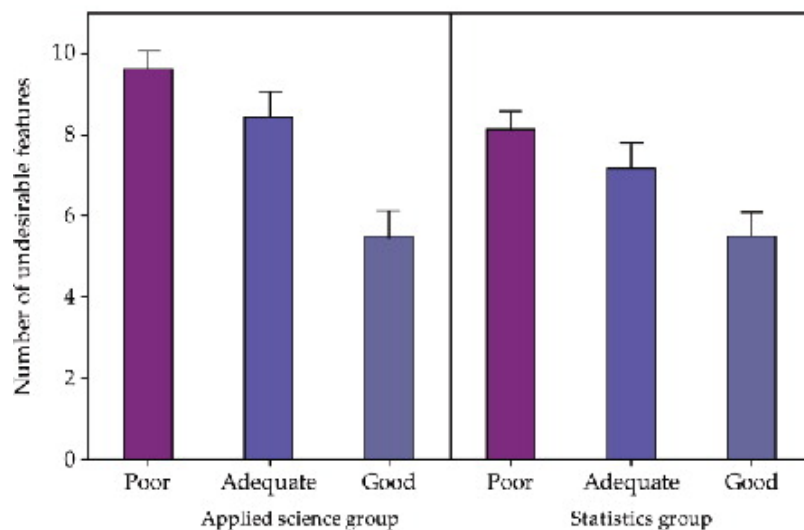
Graphs of model coefficients are often clearer than tables



Source: tables2graphs.com

Make comparisons *direct*

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



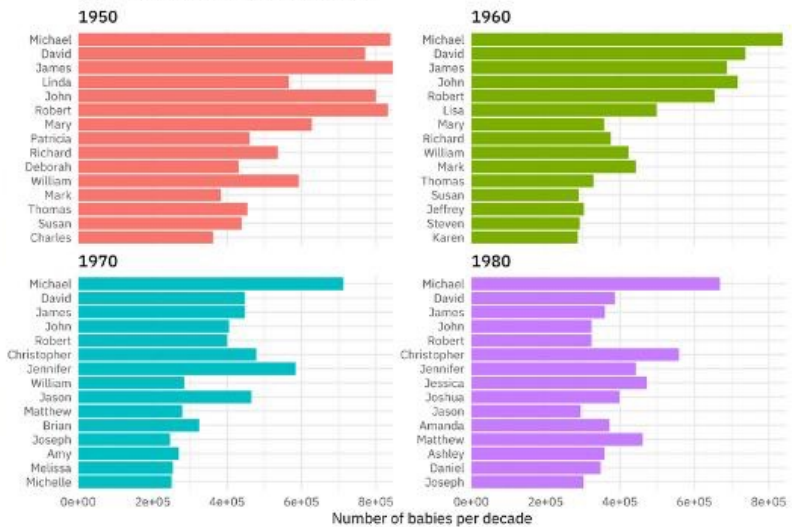
Effect ordering

- 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



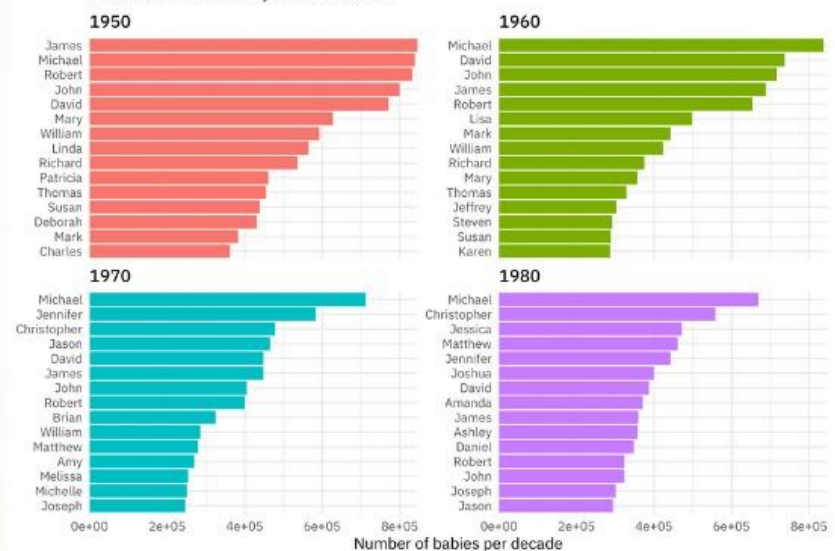
What were the most common baby names in each decade?

Via US Social Security Administration



What were the most common baby names in each decade?

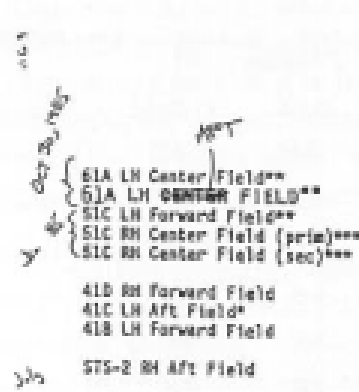
Via US Social Security Administration



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, 1986, 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



SRM No.	Cross Sectional View			Top View		Clocking Location (deg)
	Erosion Depth (in.)	Perimeter Affected (deg)	Horizontal Dia. (in.)	Length Of Max Erosion (in.)	Total Heat Affected Length (in.)	
20A	None	None	0.280	None	None	35° - 68°
22A	NONE	NONE	0.280	NONE	NONE	338° - 18°
15A	0.040	154.0	0.280	4.25	5.25	163
15B	0.038	130.0	0.280	12.50	58.75	154
15B	None	45.0	0.280	None	29.50	154
13B	0.028	110.0	0.280	3.00	None	275
11A	None	None	0.280	None	None	--
10A	0.040	217.0	0.280	3.00	14.50	161
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

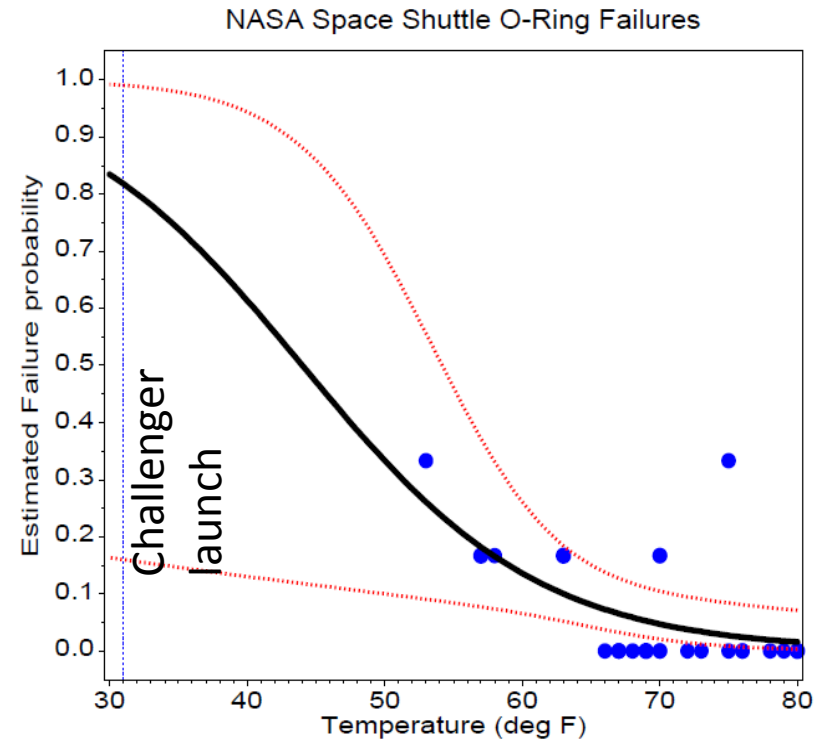


Visual explanation: Graphics

- 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.

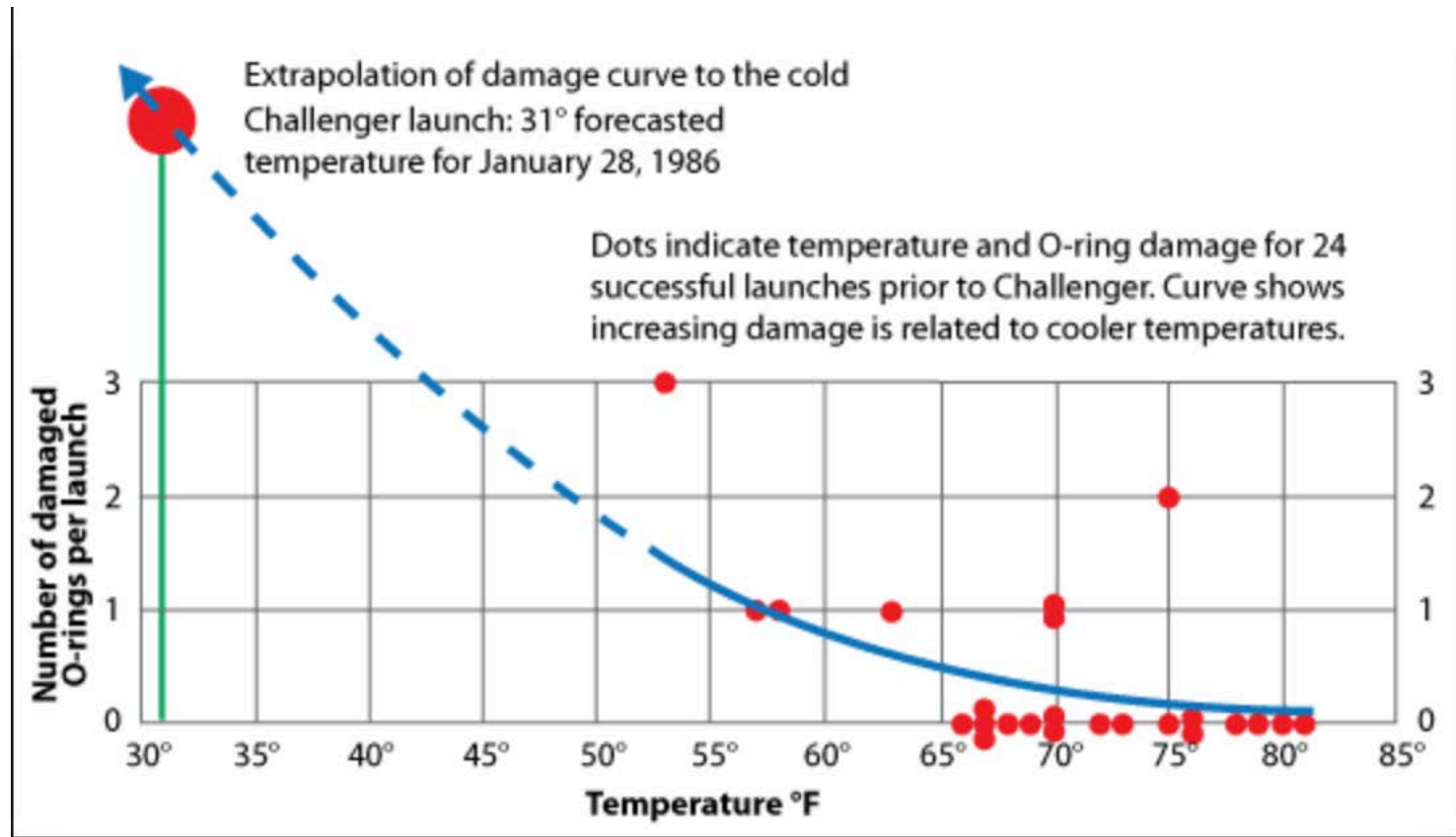
The graph shown here is the result of a statistical model fit to the data

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



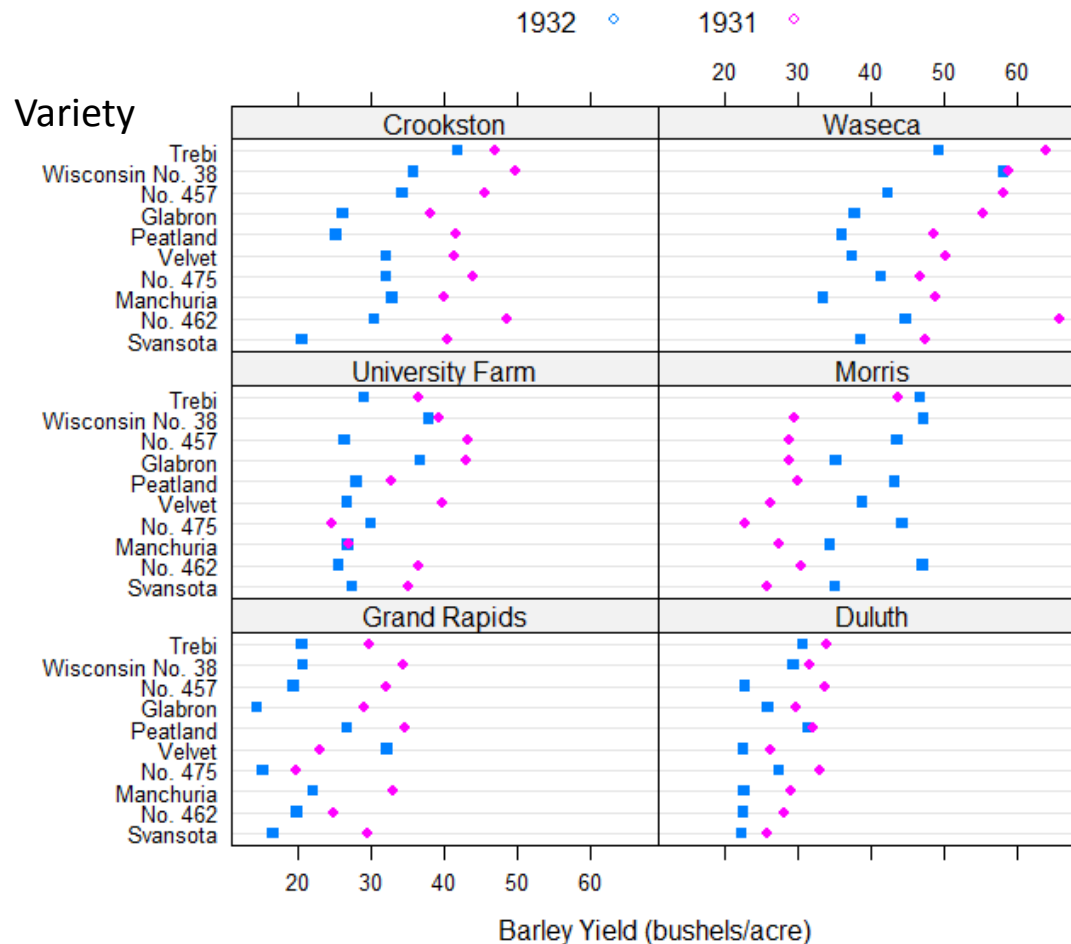
Presentation graphic

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						<i>Mean</i>
	Crookston	Duluth	Grand Rapids	Morris	University Farm	Waseca	
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
<i>Mean</i>	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 $\text{Yield} = \text{Variety} + \text{Site}$

Variety	Site						Mean
	Grand Rapids	Duluth	University Farm	Morris	Crookston	Waseca	
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, $\Delta y_{ij} = 1931 - 1932$ 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

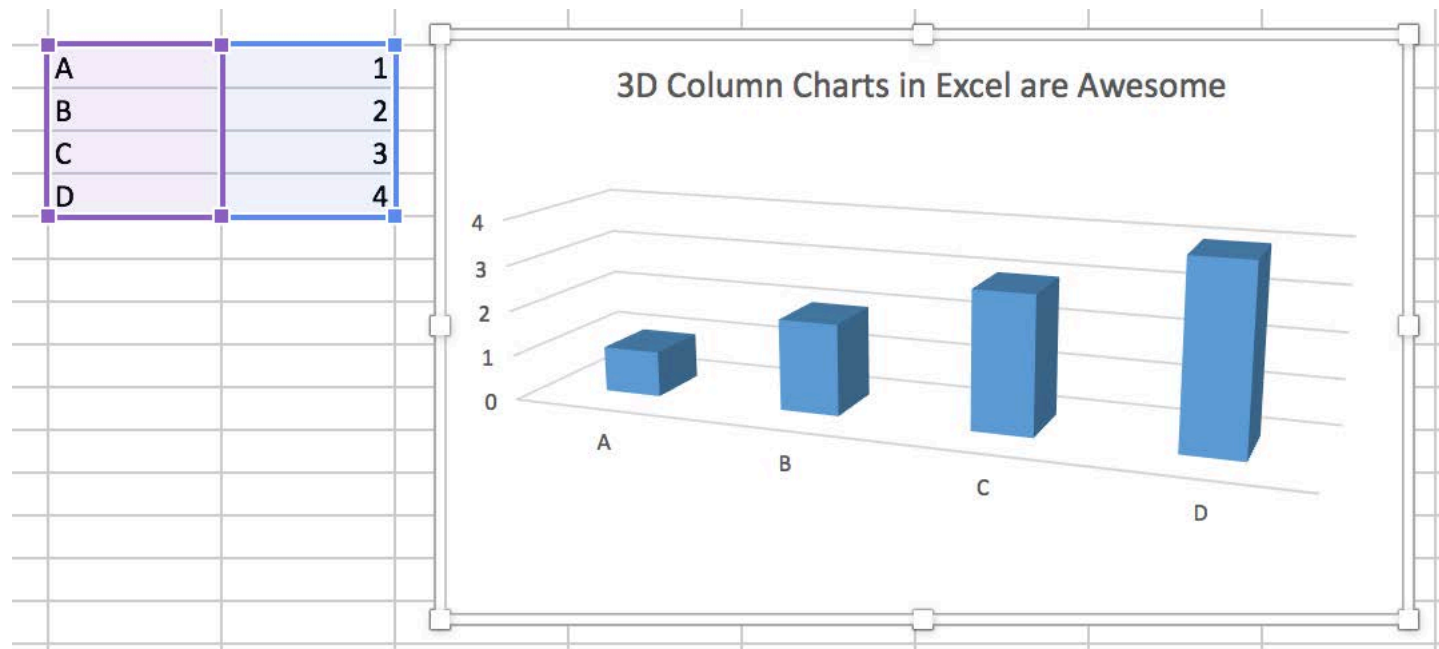
Variety	Site						Mean
	Morris	Duluth	University Farm	Grand Rapids	Waseca	Crookston	
No. 475	-22	6	-5	4	6	12	0.1
Wisconsin No. 38	-18	2	1	14	1	14	2.4
Velvet	-13	4	13	-9	13	9	2.9
Peatland	-13	1	5	8	13	16	4.8
Manchuria	-7	6	0	11	15	7	5.5
Trebi	-3	3	7	9	15	5	6.1
Svansota	-9	3	8	13	9	20	7.3
No. 462	-17	6	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

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

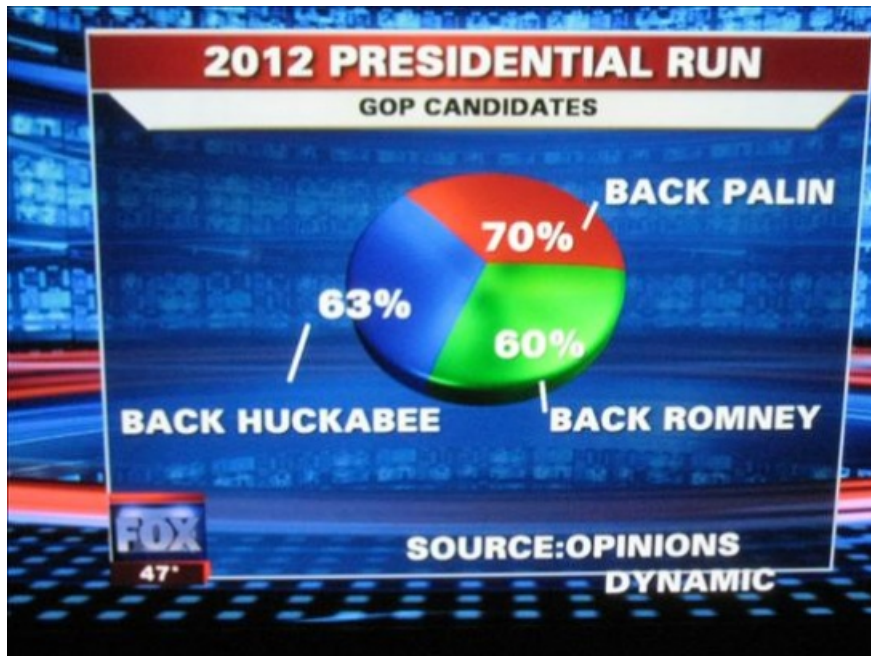
Friends don't let friends use Excel for data visualization or statistics



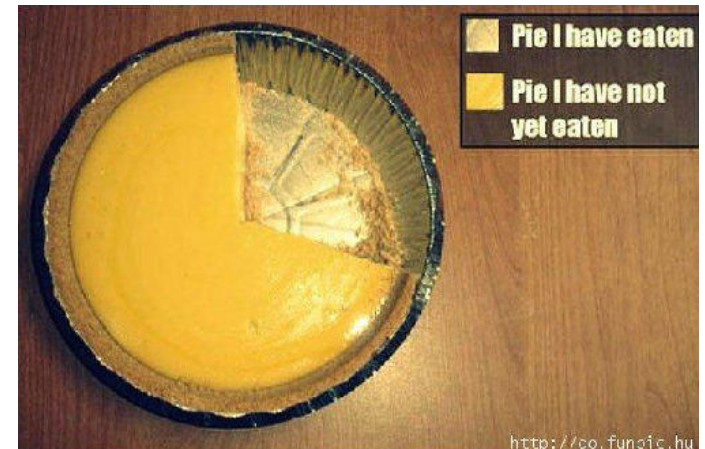
How many things are wrong with this graph?

Pie charts are easy to abuse

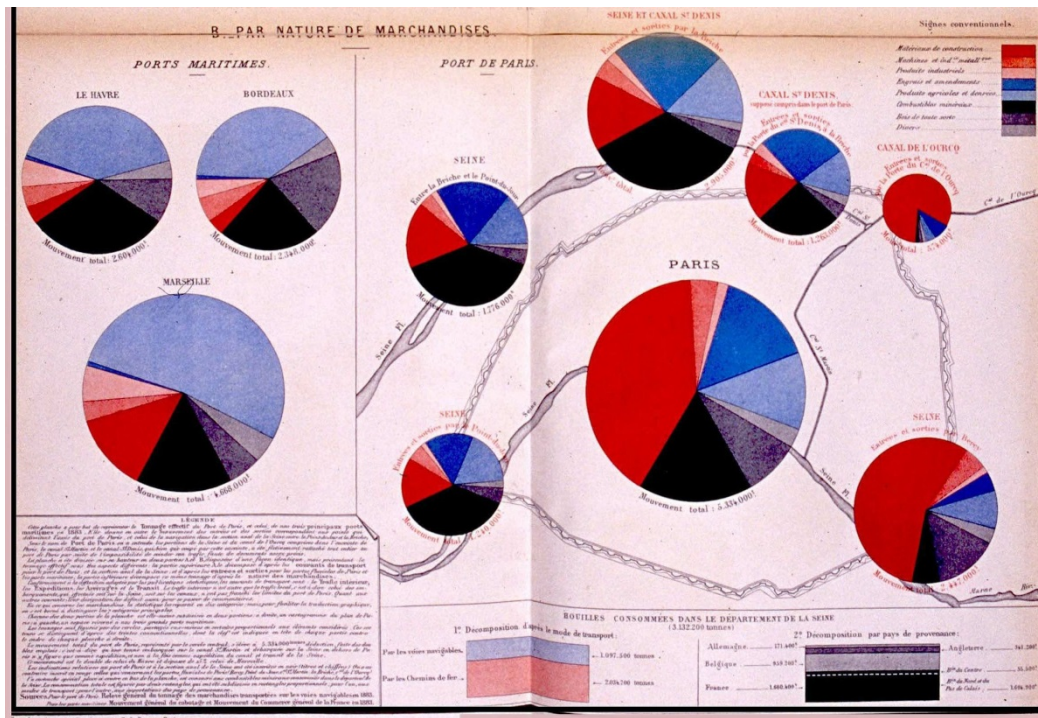
What's wrong with this picture?



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



But, can be used to great effect

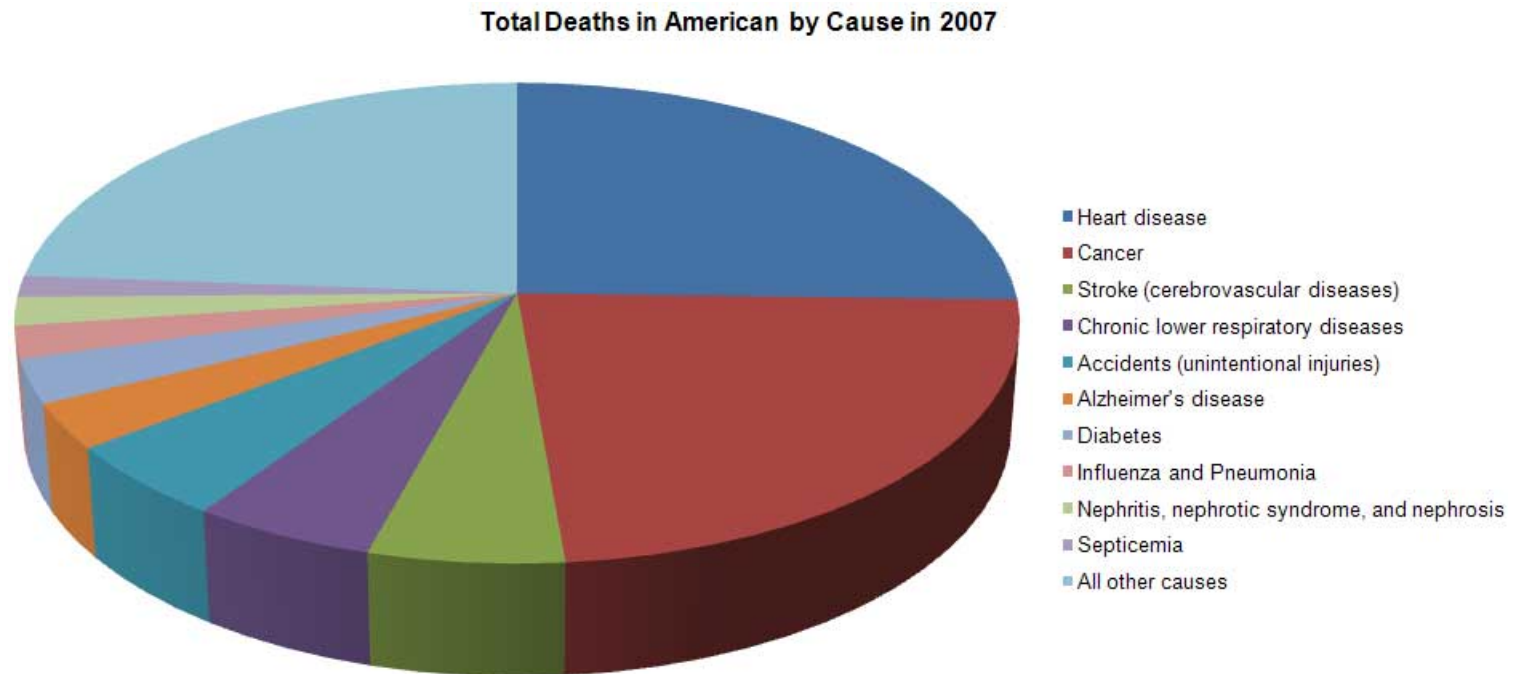


This graphic uses pie charts to show the transport of different kinds of goods to the ports of Paris and the principal maritime ports

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

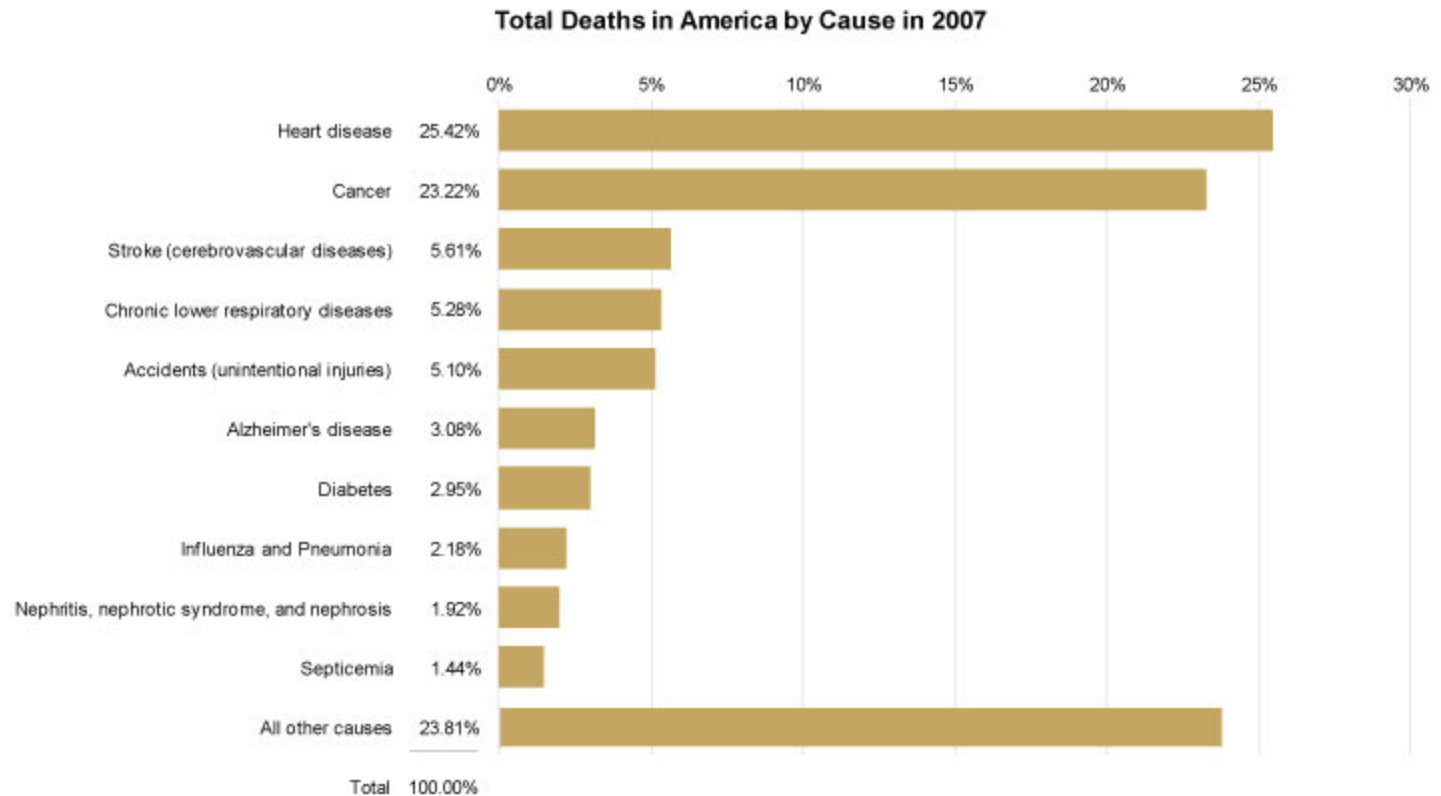
Album de Statistique Graphique, 1885, plate 17.

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?

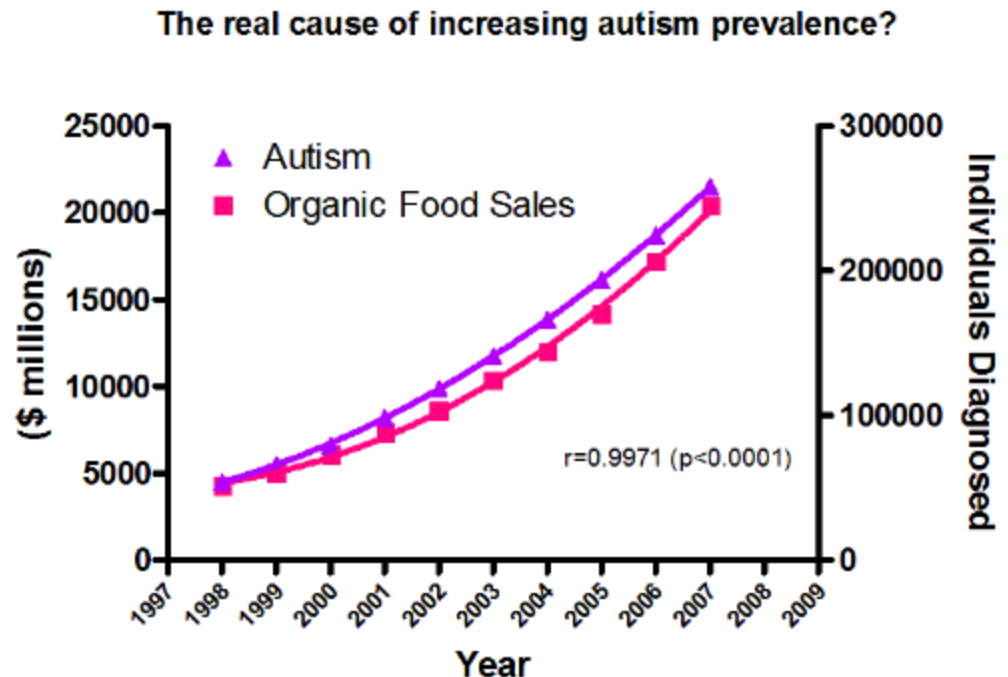
Simple re-design makes it clearer



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!



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043; "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act"

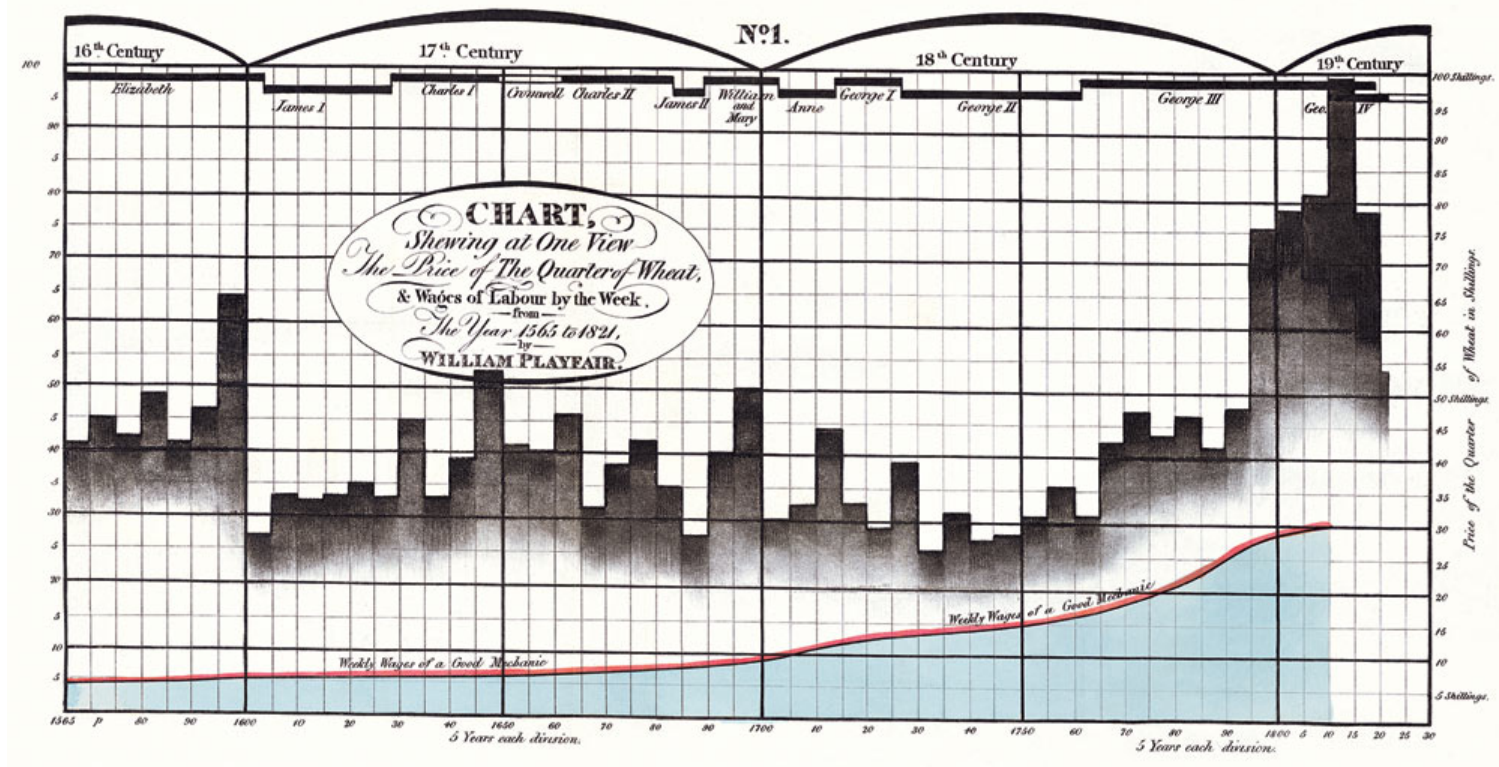
But, can be used to great effect

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



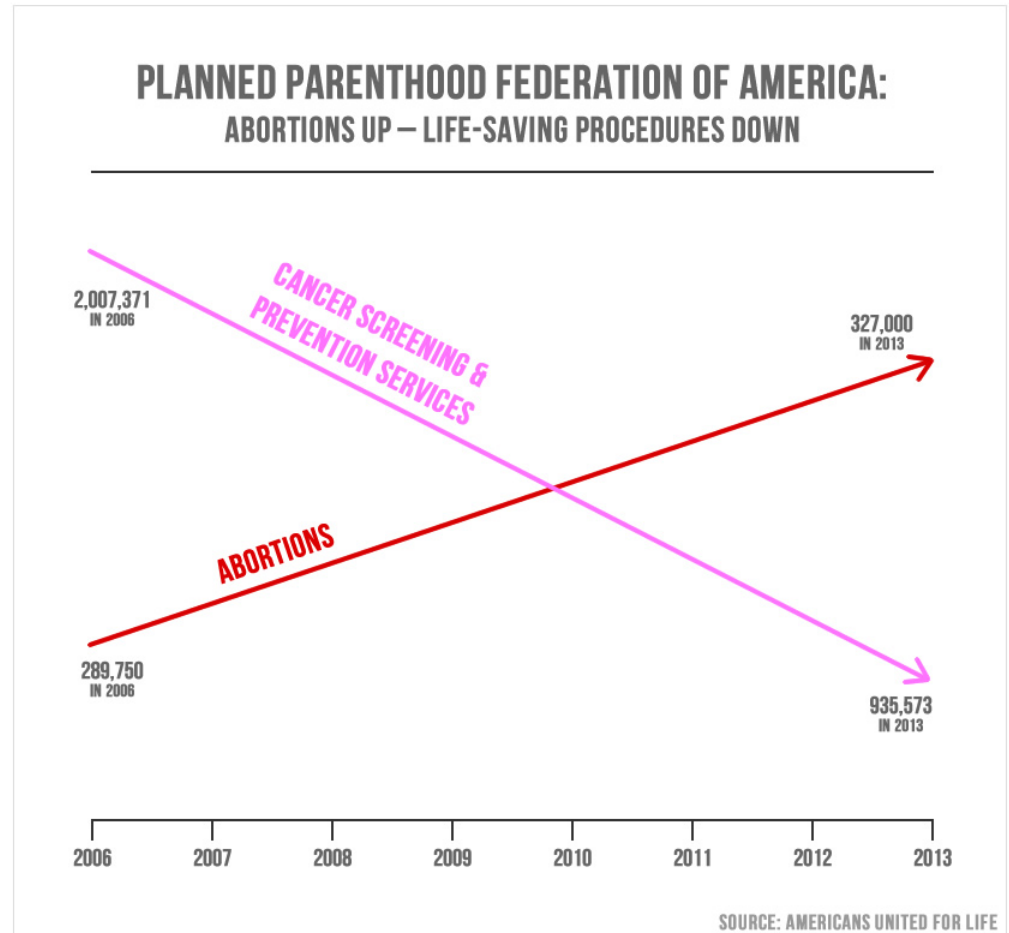
Even more evil: No scales, no data

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?



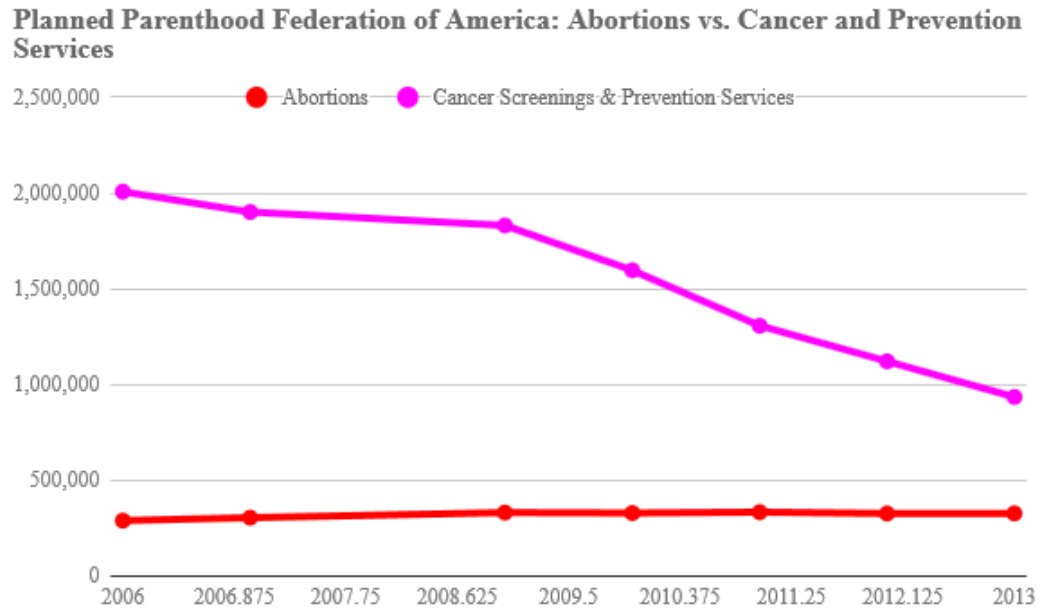
See: <http://www.politifact.com/truth-o-meter/statements/2015/oct/01/jason-chaffetz/chart-shown-planned-parenthood-hearing-misleading-/>

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.



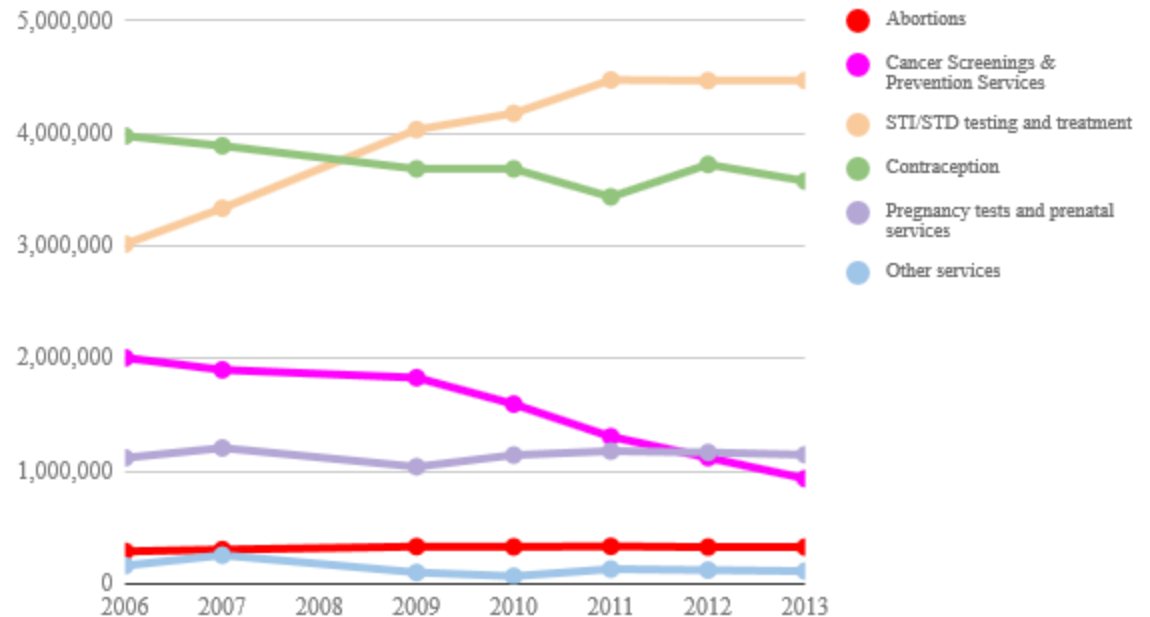
What are a few improvements that could be made to this graph?

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

Services Provided by Planned Parenthood



Graphical failure

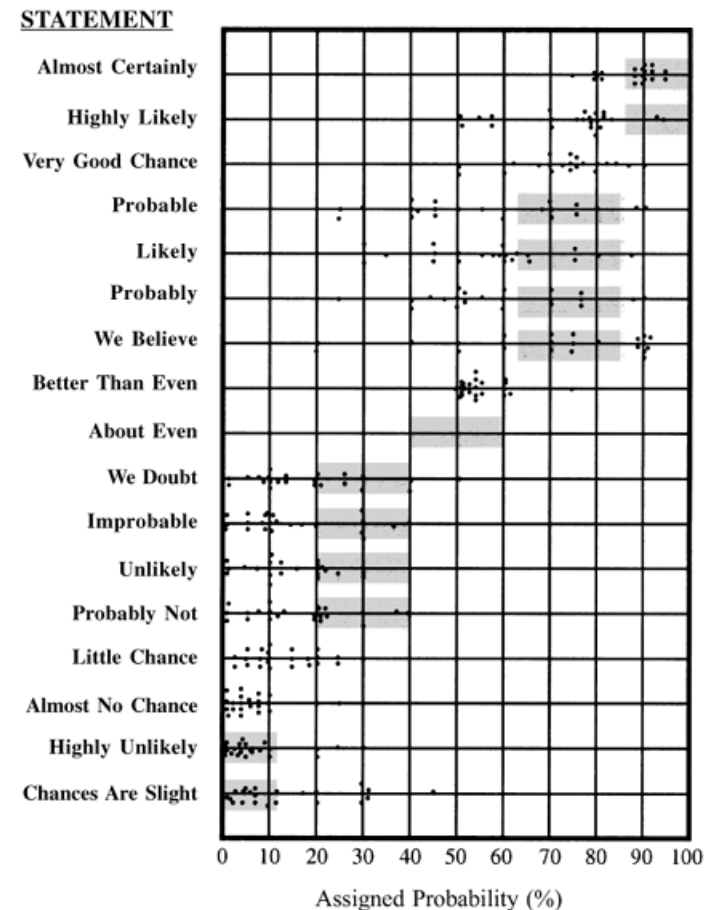
This graph reports the results of a survey by Sherman Kent for the CIA with the question:

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?

Figure 18: Measuring Perceptions of Uncertainty



Graphical excellence

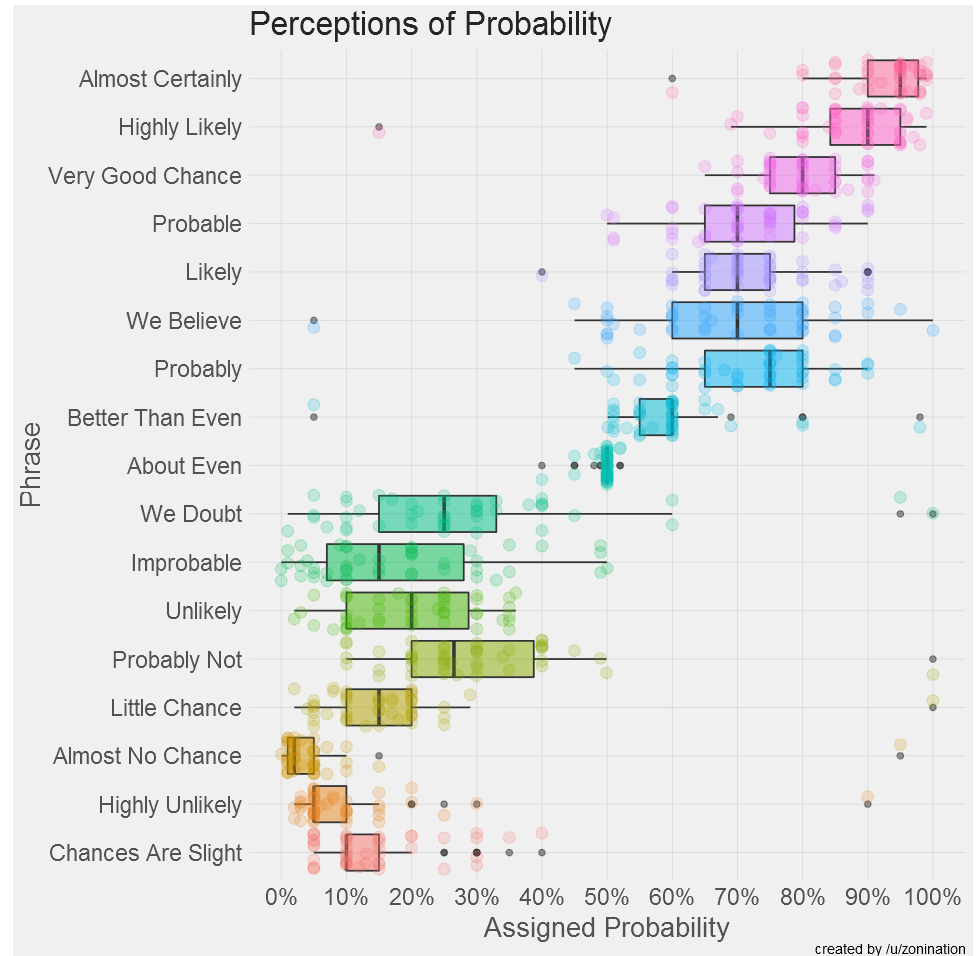
This graph shows the same data, as both dotplots & boxplots

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

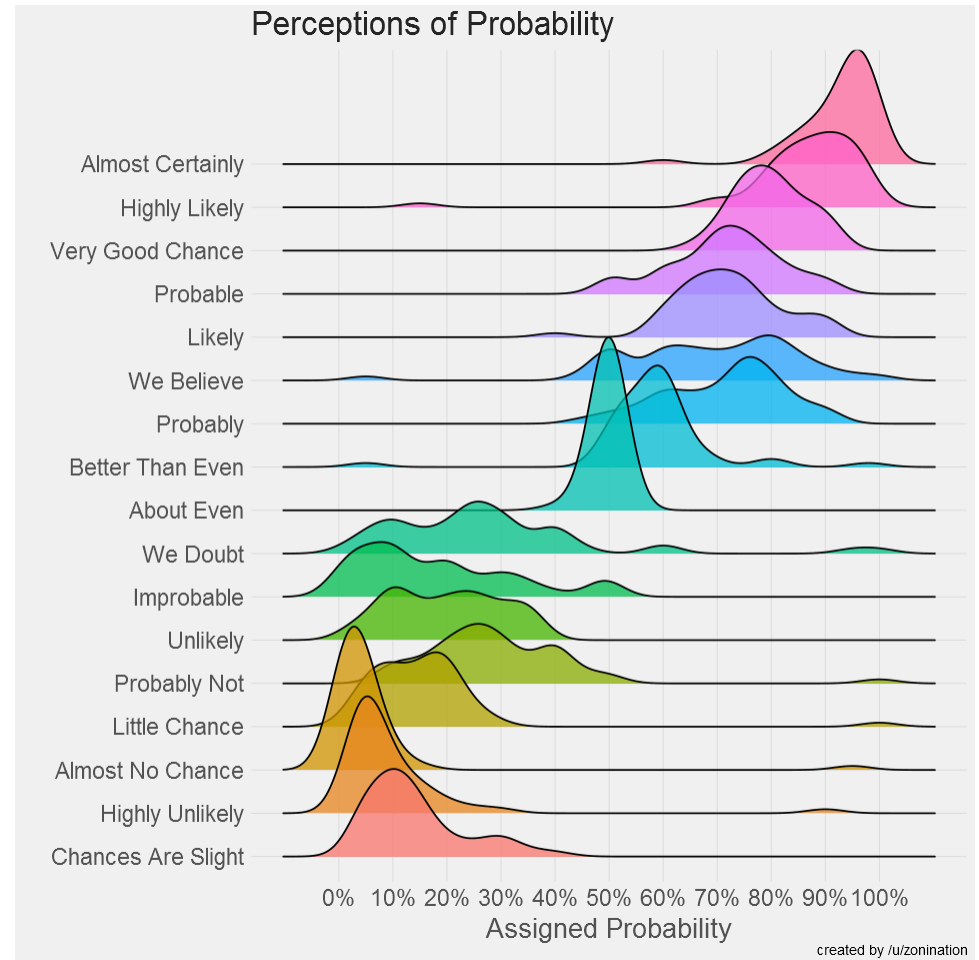


Graphical excellence

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.



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: CO₂ emissions as the main agent.

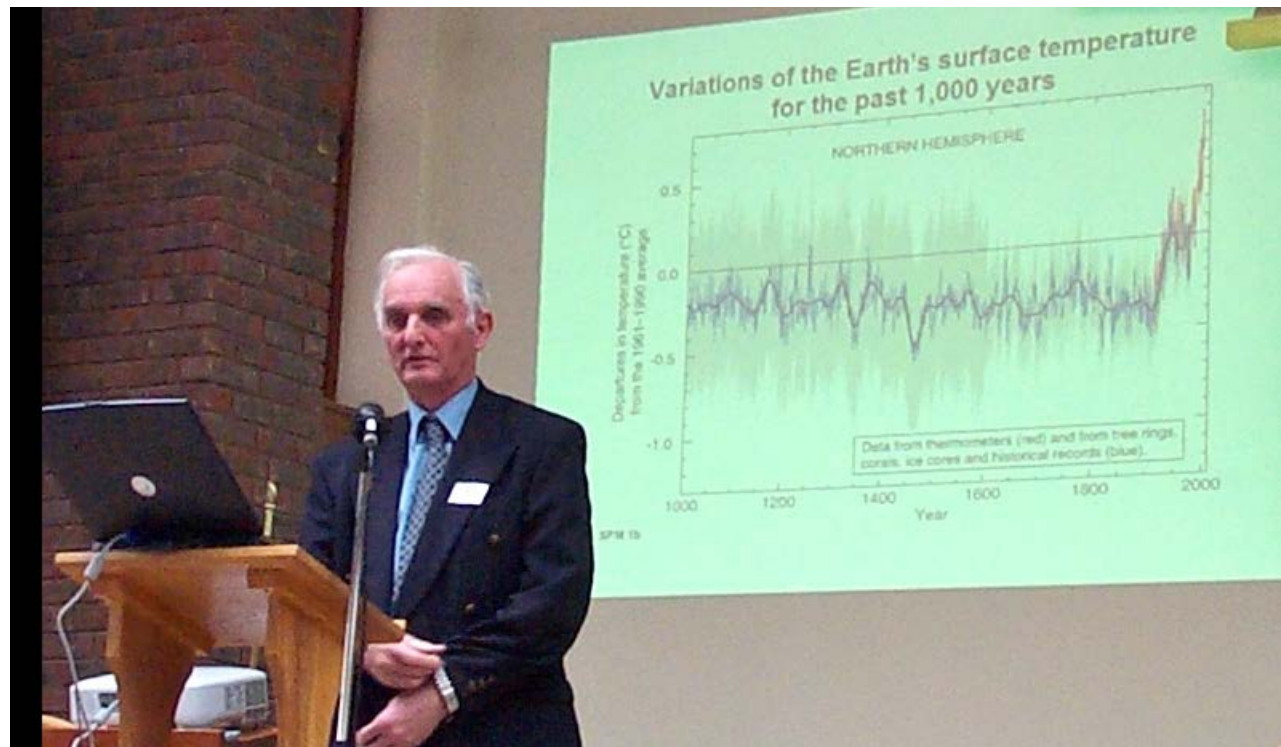


Movie: <https://www.youtube.com/watch?v=8ZUoYGAI5i0>; <http://www.imdb.com/title/tt0497116/>

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.

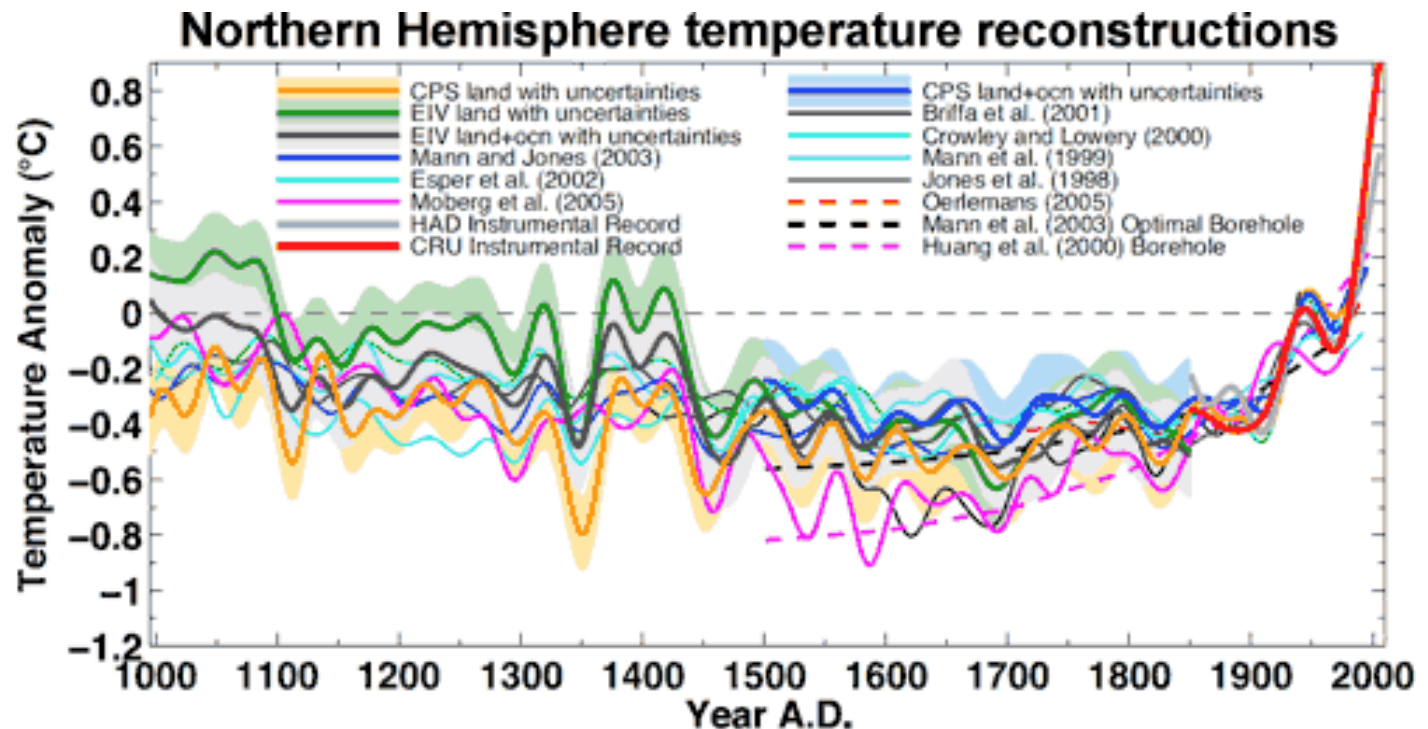
It is based on an analysis by Mann, Bradley & Hughes (1990), with a smoothed curve and uncertainty intervals.



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).

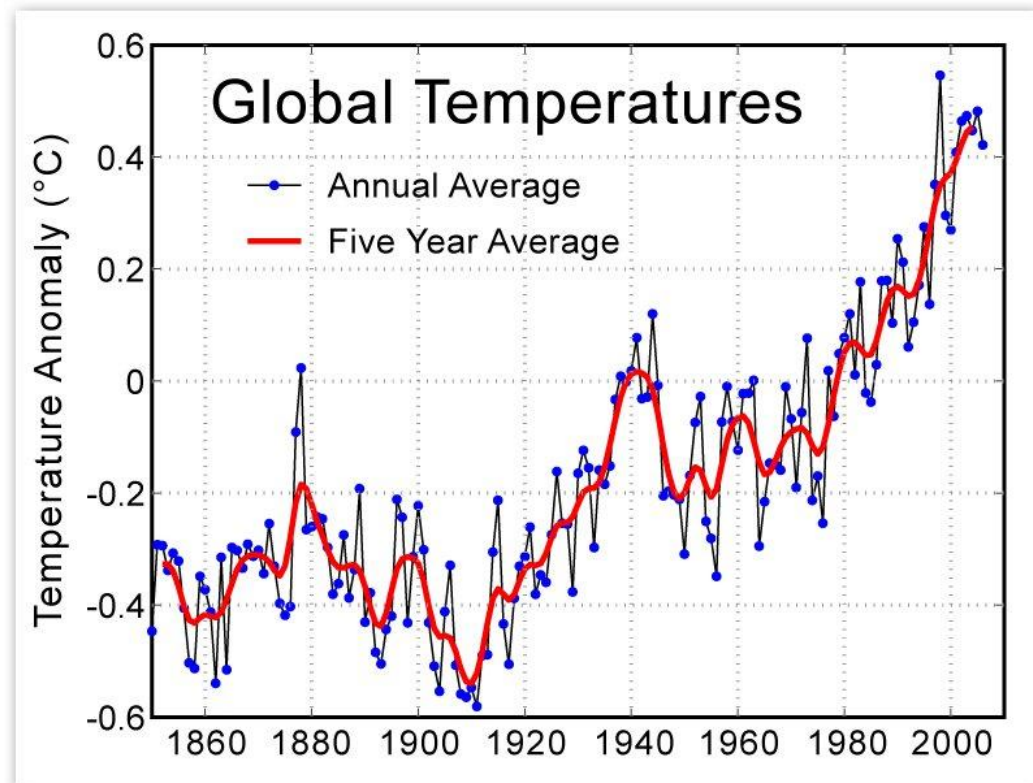
Climate scientists understood this; the sceptics did not.



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

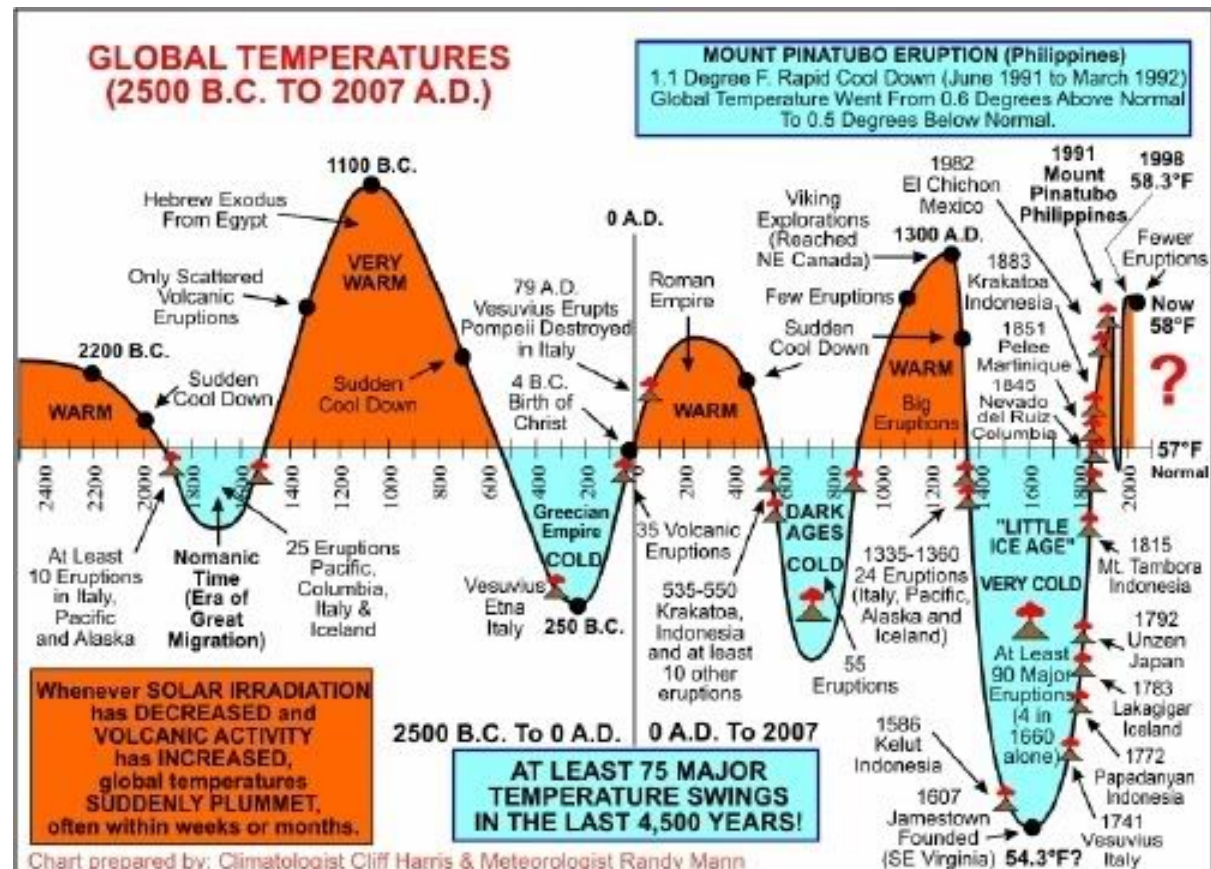
Time scale

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



Countering climate change

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



Summary

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