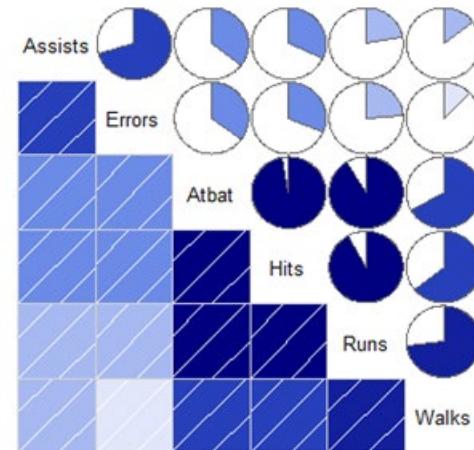
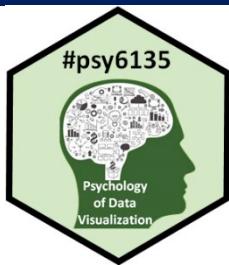


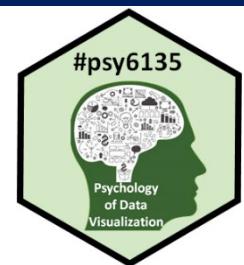
Lurking variable?



# Psychology of Data Visualization: Course Overview



Michael Friendly  
Psych 6135



<https://friendly.github.io/6135>

@datavisFriendly



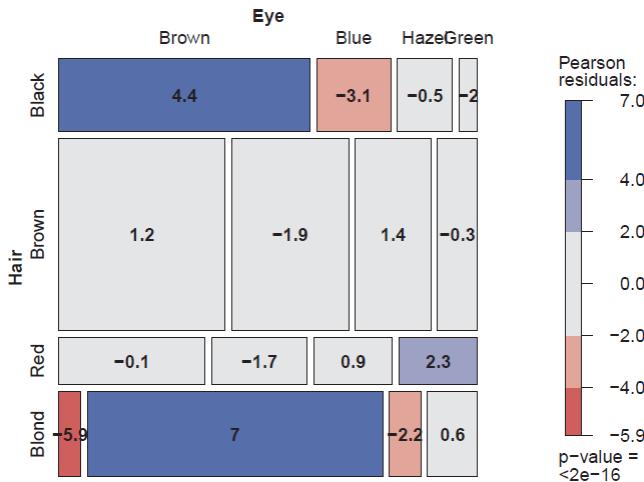
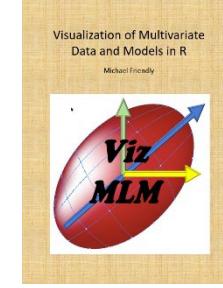
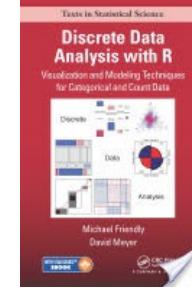
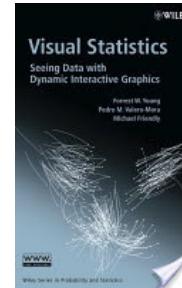
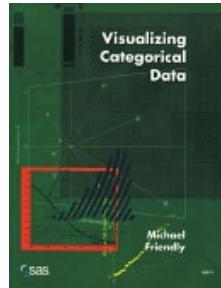
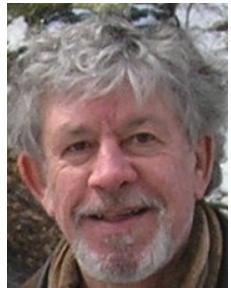
# Introducing: me

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

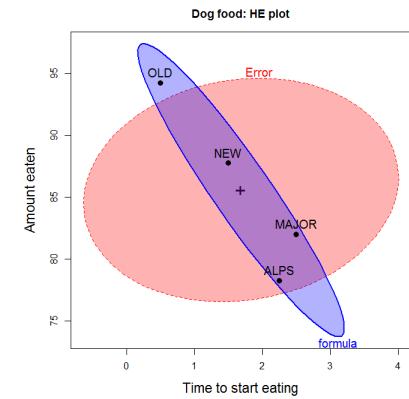
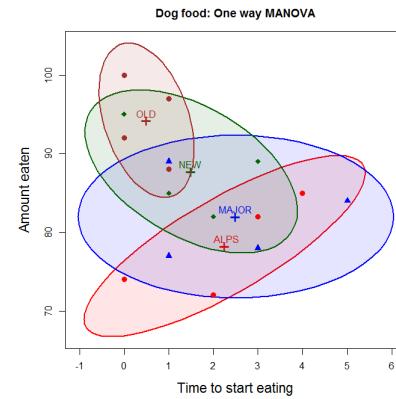
Statistical graphics developer (categorical & multivariate data analysis)



Yours to discover!



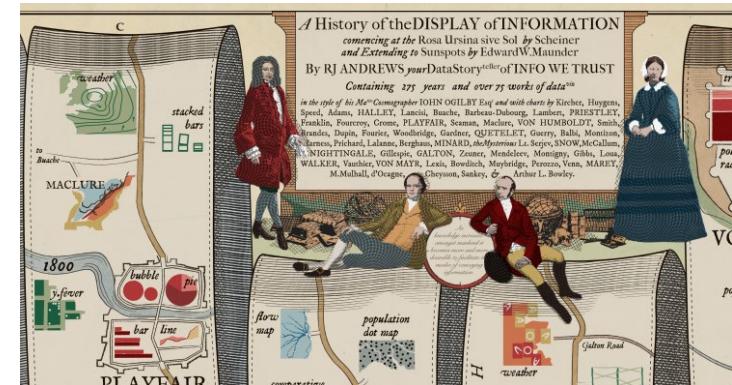
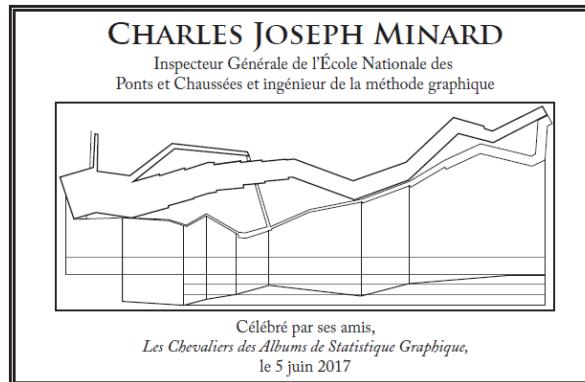
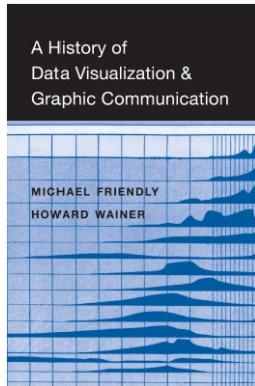
mosaic plots for frequency tables



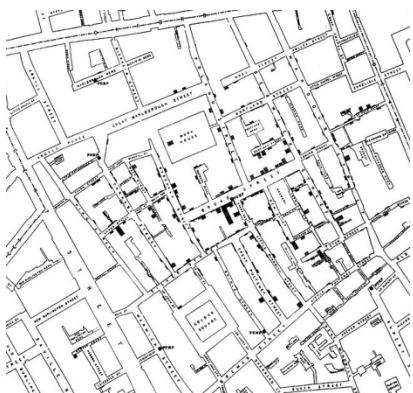
HE plots for MANOVA

# Introducing: me

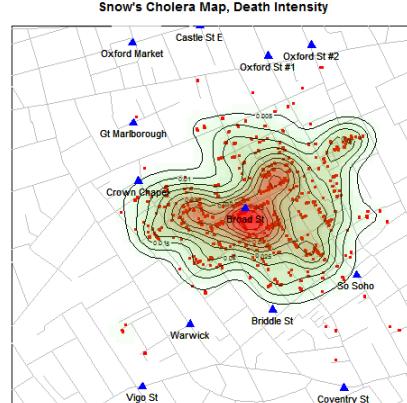
History of data visualization: *Les Chevaliers*; Friendly & Wainer (2021)



John Snow's map of cholera in London, 1854

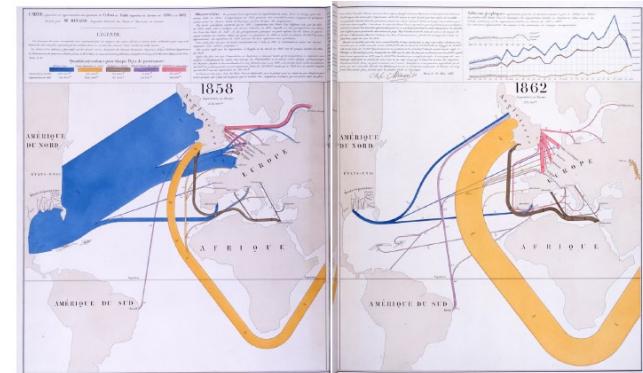


original



modern enhancement

C. J. Minard: Flow maps of cotton trade



Visual explanation: What happened in the US Civil War?

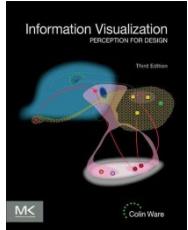
# Course Topics

- Varieties of information visualization
  - Goals of visualization
  - Survey of graphic forms
- History of information visualization
  - “Those who don’t know history are doomed to plagiarize it”
- Psychological models, theories and results
  - What can people see, **understand** and **remember** from data displays?
  - Perceptual aspects, cognitive aspects
- Human factors research
  - How to tell what works for different tasks?
- Software tools for information visualization (mainly R)
  - ggplot2; tidyverse
- Special topics
  - Visualizing uncertainty; Data journalism
  - Categorical data; High-D data; Dynamic and interactive methods

# Your role

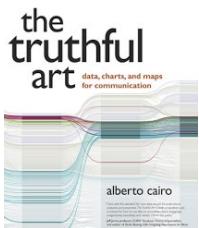
- Weekly readings – see the course [web](#) site for updates
- Discussion – no formal grade, but please contribute
- Discussion leader (20%)
  - Each week 1-2 of you will lead a brief discussion on one of the readings or sub-topics (~ 5-8 min.)
  - Please sign up on the Google sheet, <https://bit.ly/3S6o9pP>
- Class presentation (40%)
  - In the last 2 weeks, 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*, 3<sup>rd</sup> Ed.

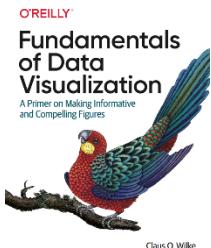
What perceptual science has to say about data visualization, from a bottom-up perspective  
Course notes at: <http://ccom.unh.edu/vislab/VisCourse/index.html>



Alberto Cairo, *The Truthful Art*

Information graphics from a communication perspective

Blog: <http://www.thefunctionalart.com/>



Claus Wilke, *Fundamentals of Data Visualization*

Well thought out, a wide range of topics, good practical advice, lots of examples.

Online version: <https://clauswilke.com/dataviz/>

Course notes: <https://wilkelab.org/SDS375/>



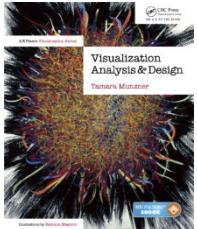
Hadley Wickham, *ggplot2: Elegant graphics for data analysis*, 2nd Ed.

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/>

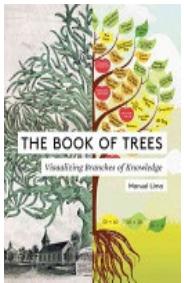
# More books I like



Tamara Munzner (2014), *Visualization Analysis & Design*

An attractive new book combining computer science and design perspectives

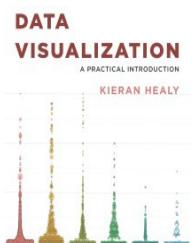
Web page: <http://www.cs.ubc.ca/~tmm/vadbook/> with lots of illustrations & lectures



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/>



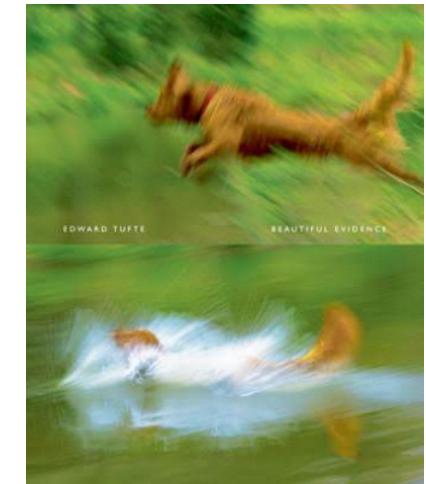
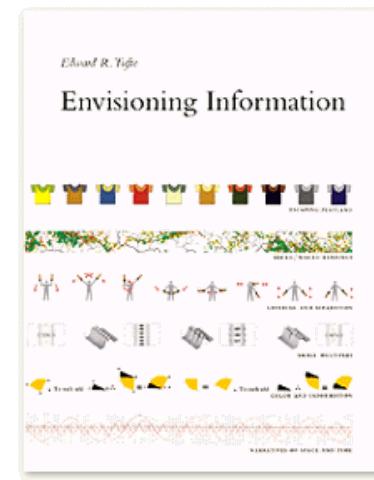
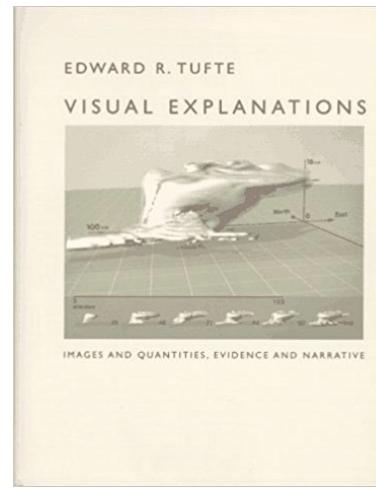
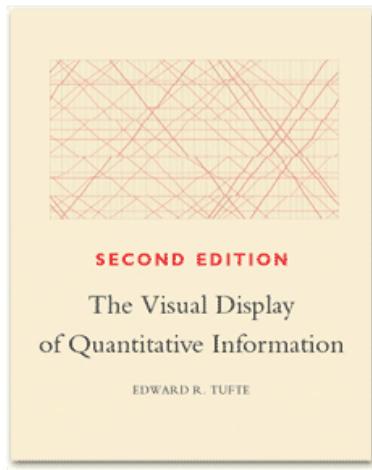
Keiran Healy, *Data Visualization: A Practical Introduction*

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

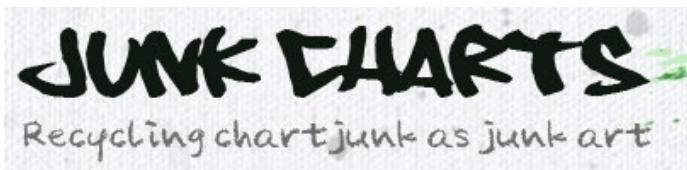
Web site:

<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.



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



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.



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



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

# Blogs & Web resources

## DATA STORIES



KANTAR  
Information is Beautiful  
Awards

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



R news and tutorials contributed by (750) R bloggers

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



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

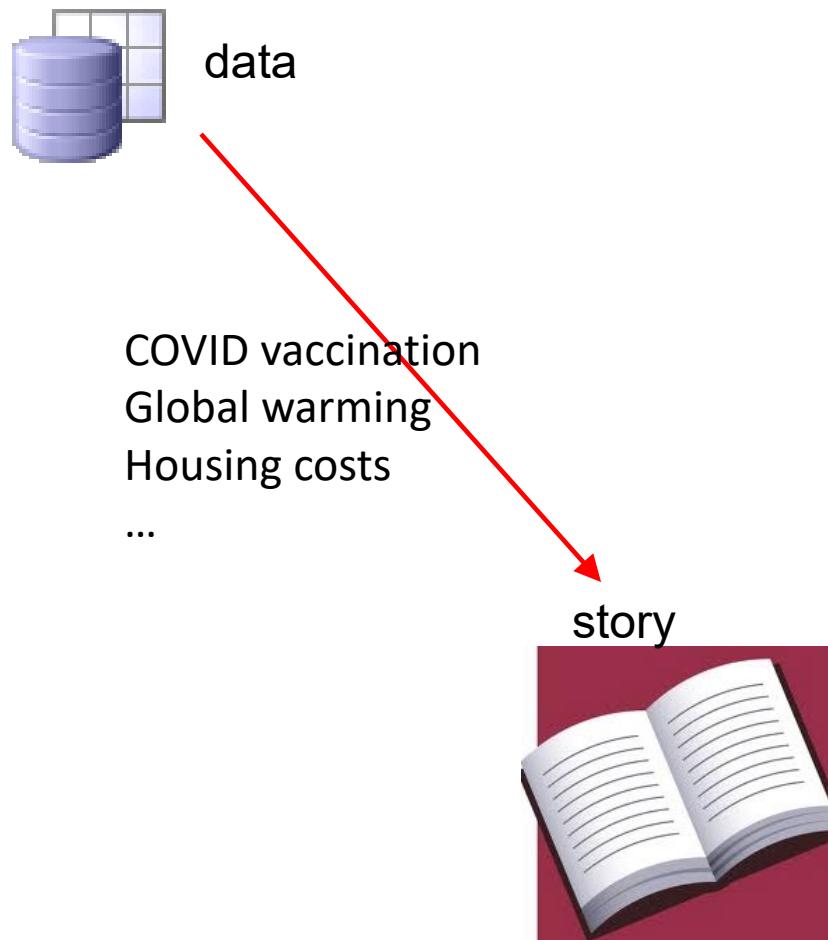
Raymond Andrews, <http://infowetrust.com/>. A visual storyteller delights with graphic stories from the history of data visualization

# Psychology facts: Why visualization?

- ~90% of information about the environment is received through the eyes.
- ~50% of our brain neurons are involved in the processing of visual information.
- The presence of pictures increases desire to read the text by ~80%.
- We remember 10% of what we 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%.

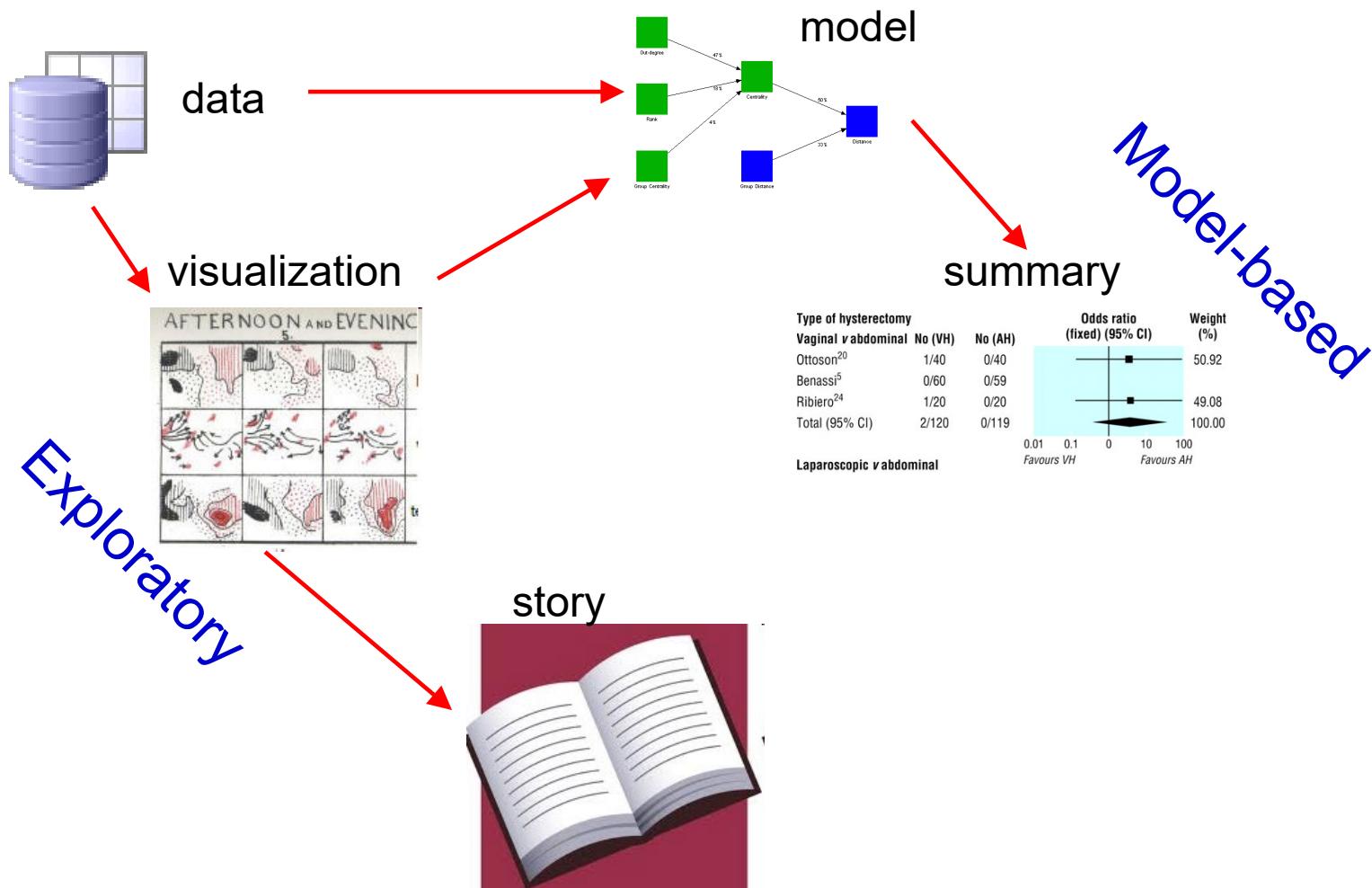
# Data, pictures, models & stories

Goal: Tell a **credible story** about  
some **real** data problem



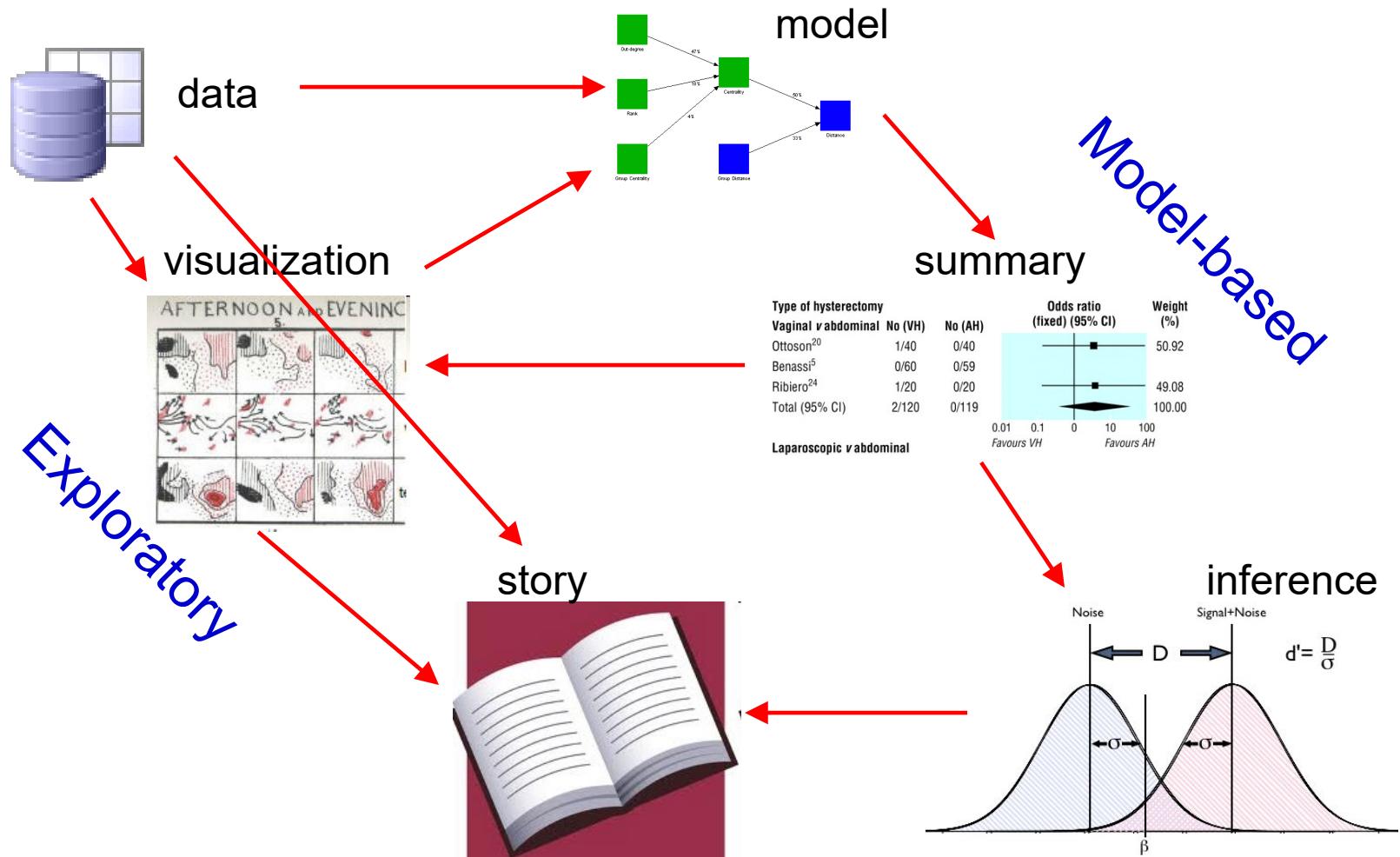
# Data, pictures, models & stories

## Two paths to enlightenment



# Data, pictures, models & stories

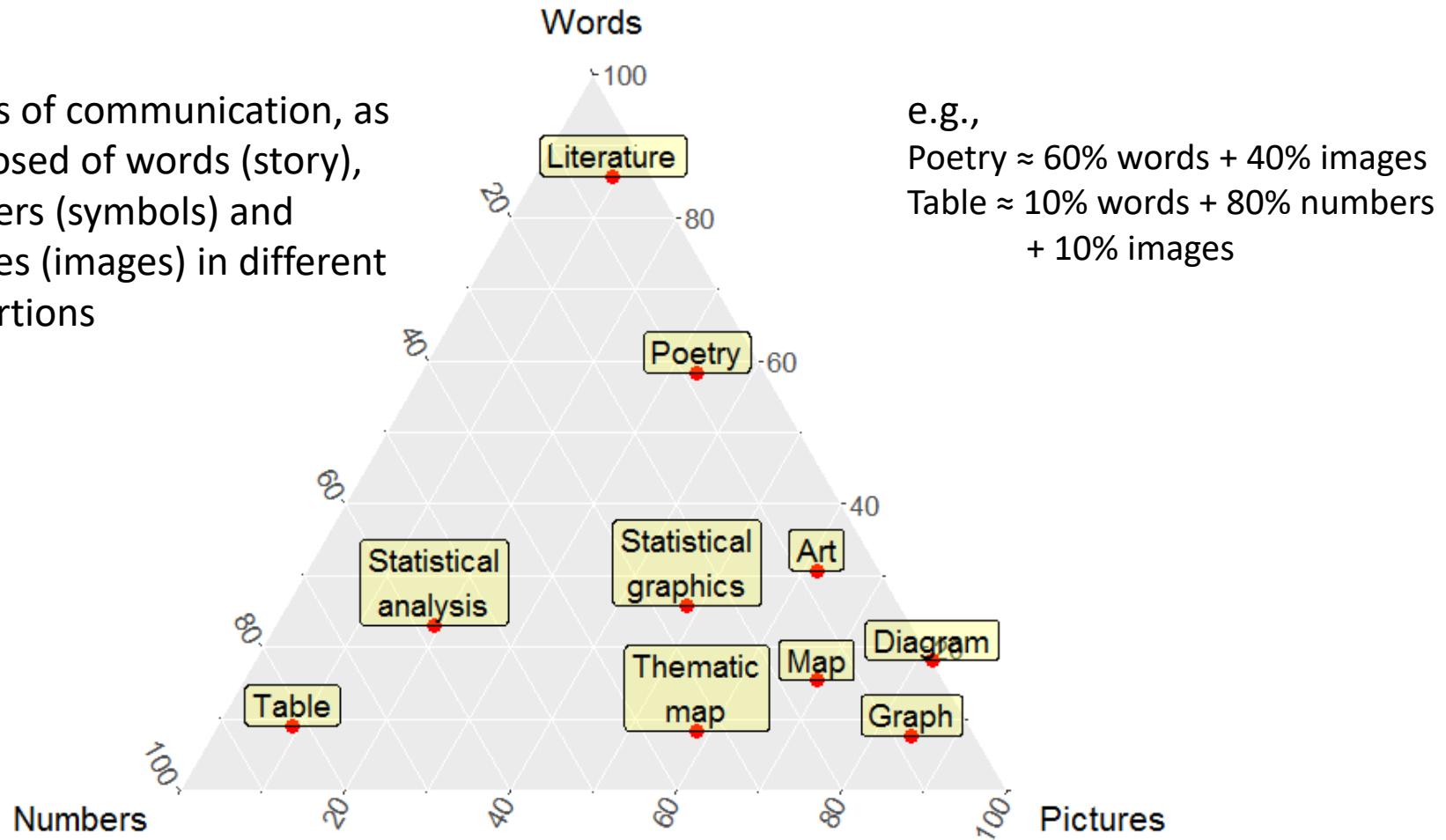
Now, tell the story!



# Words, numbers and pictures

Pictures and images in a wider context

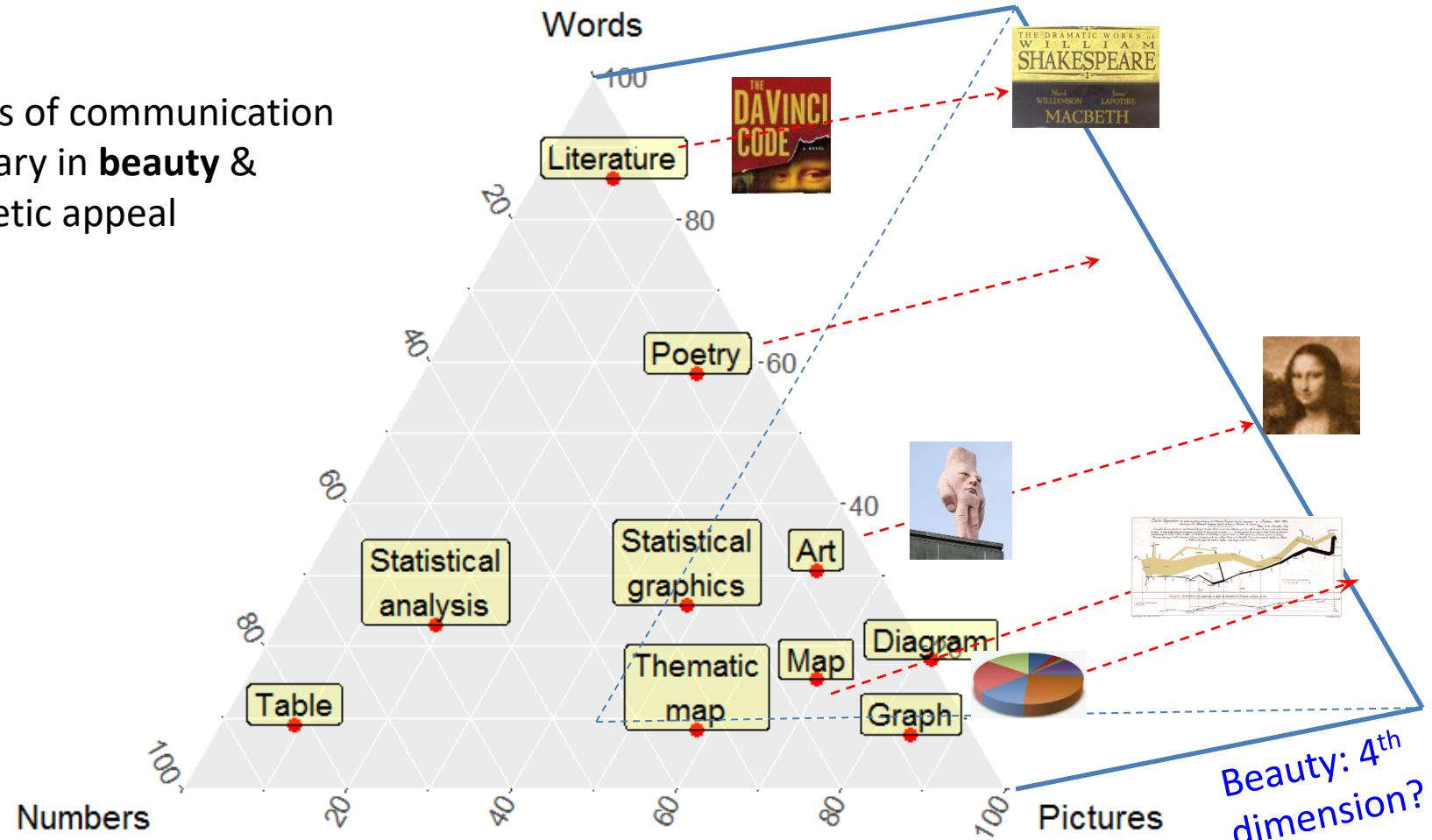
Modes of communication, as composed of words (story), numbers (symbols) and pictures (images) in different proportions



# Words, numbers and pictures

Beauty: The 4<sup>th</sup> dimension

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

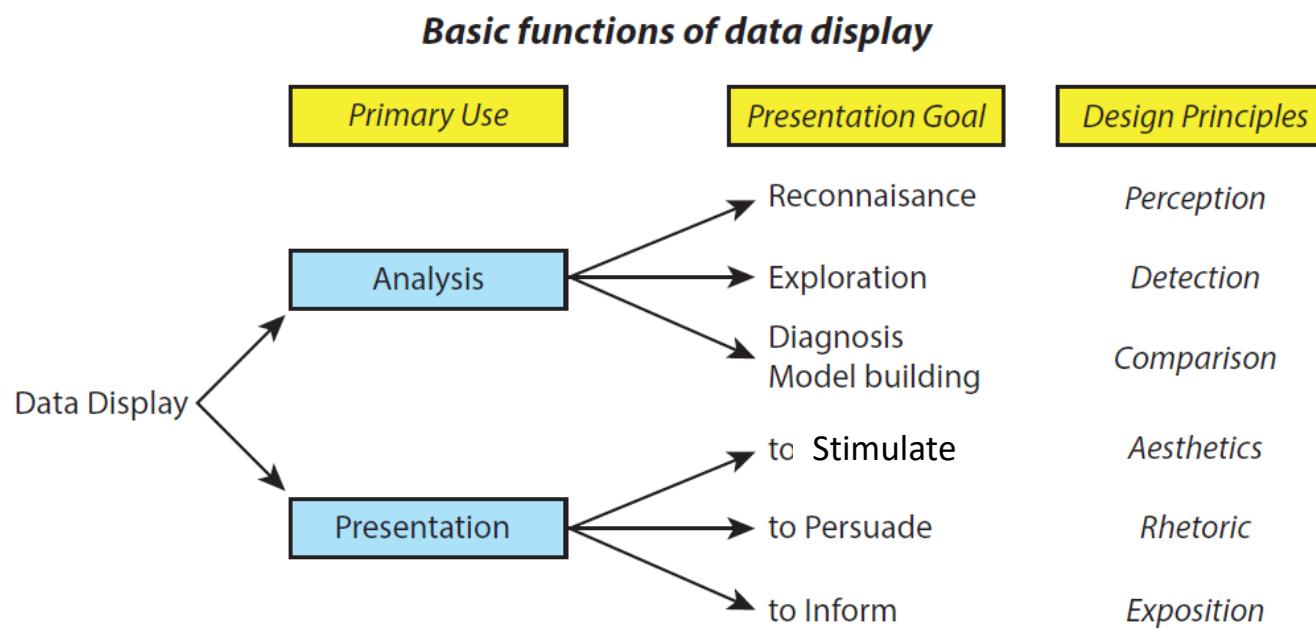


# Roles of graphics in communication

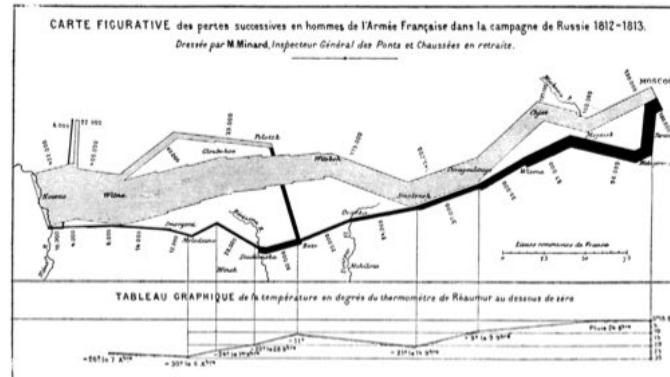
- Graphs (& tables) are forms of communication:
  - What is the audience?
  - What is the message?

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

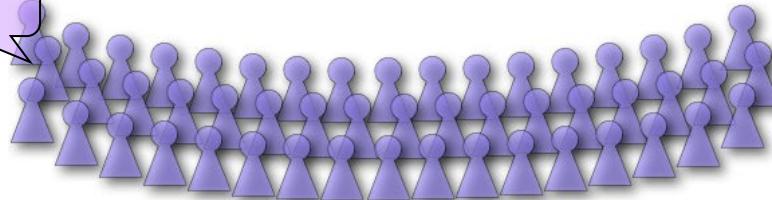
**Presentation graphs:** design to attract attention, make a point, illustrate a conclusion



# Different graphs for different purposes

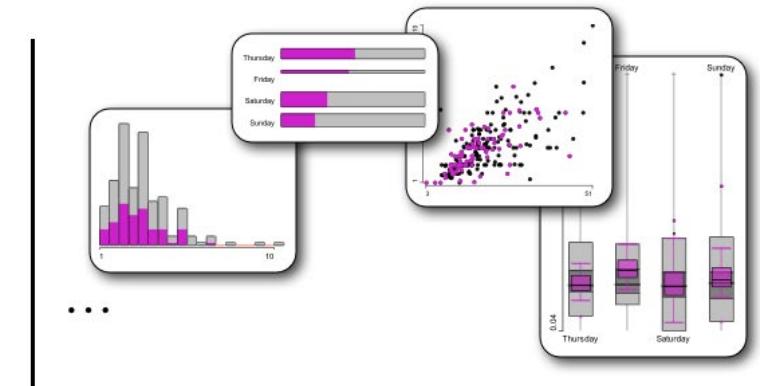


Wow!

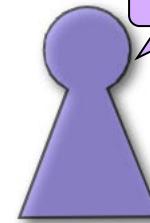


Presentation

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



Ah ha!



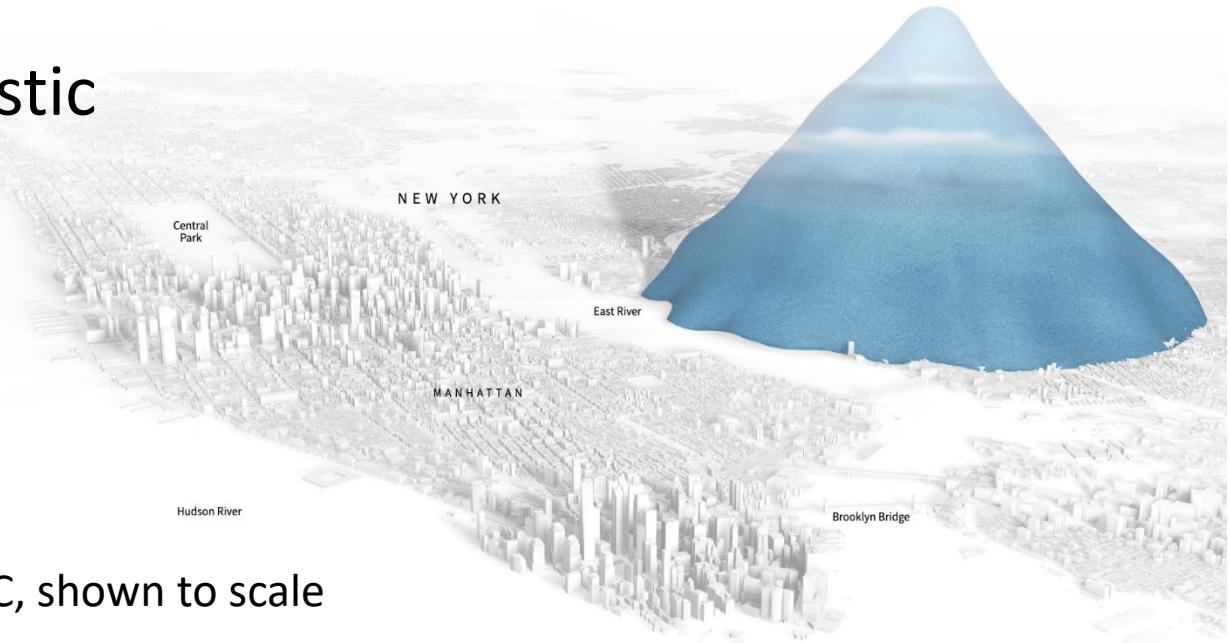
Exploration

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

# 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: Measles 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

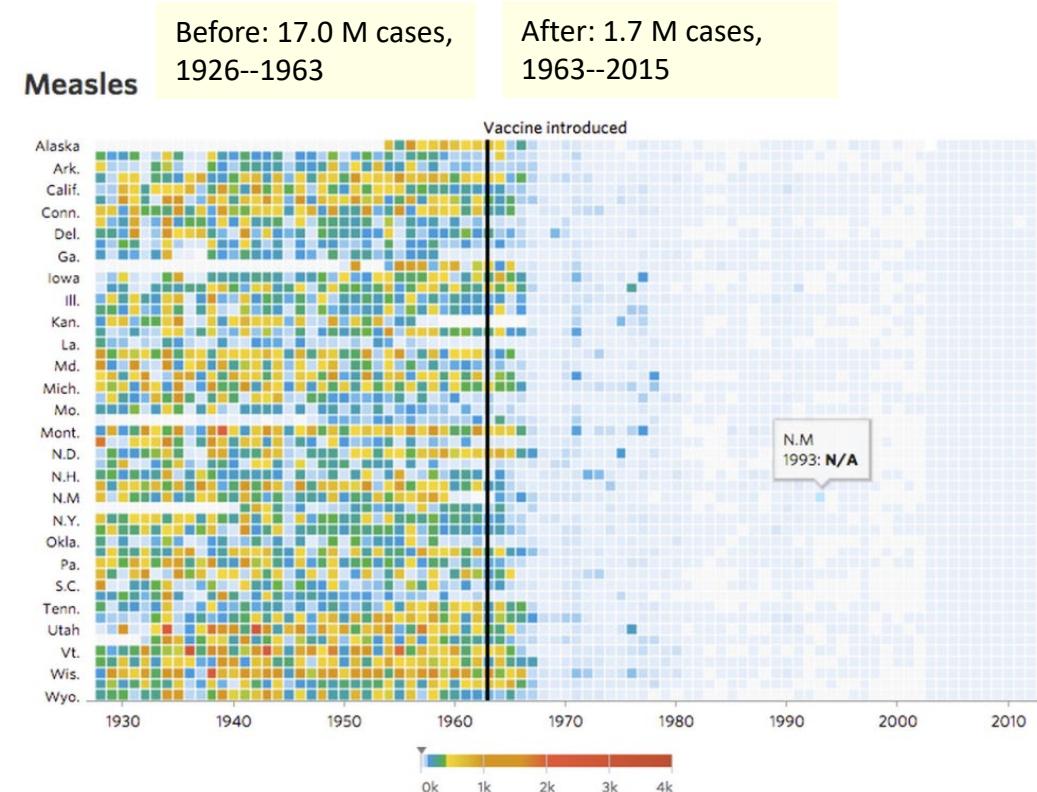
Measles was decimated!

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.

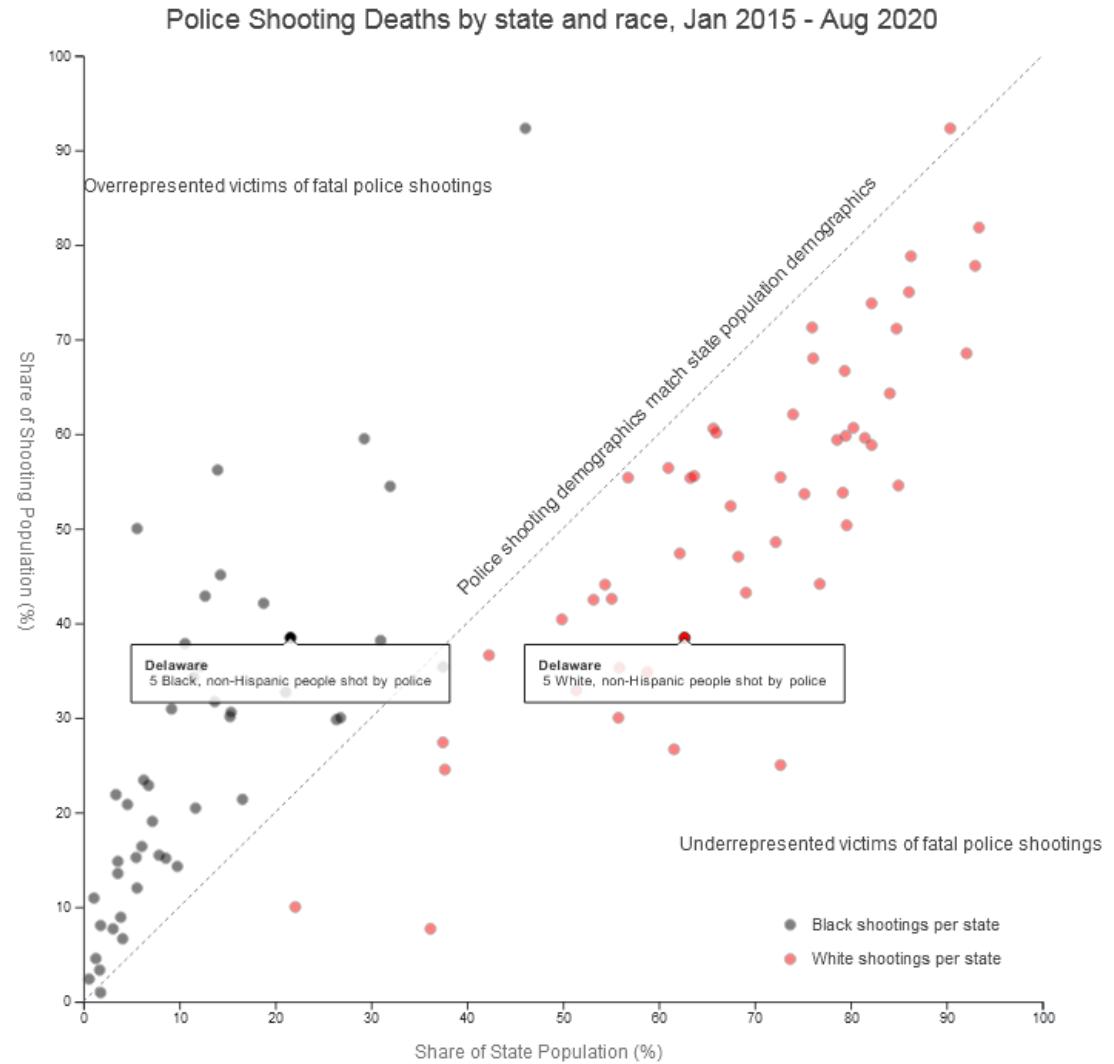


# Police shooting deaths

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

Annotations help to tell the story





# Presentation graph: Nightingale (1857)

After reform

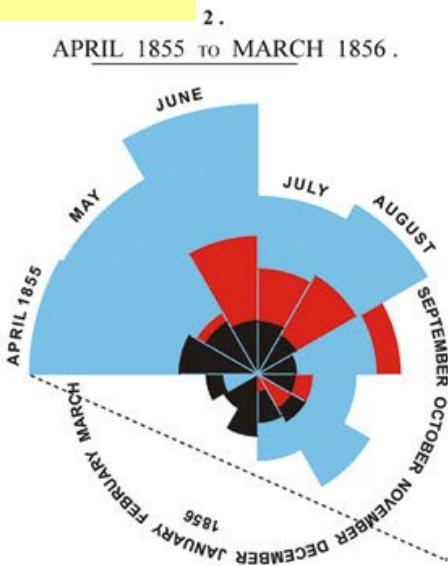
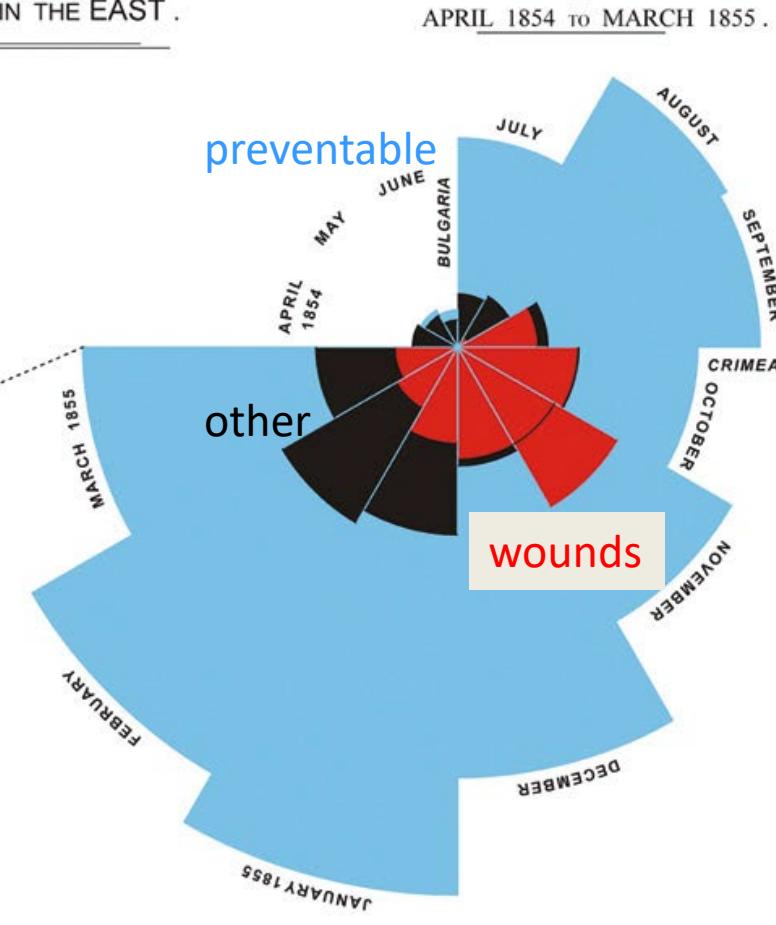


DIAGRAM OF THE CAUSES OF MORTALITY  
IN THE ARMY IN THE EAST.

Before reform



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex

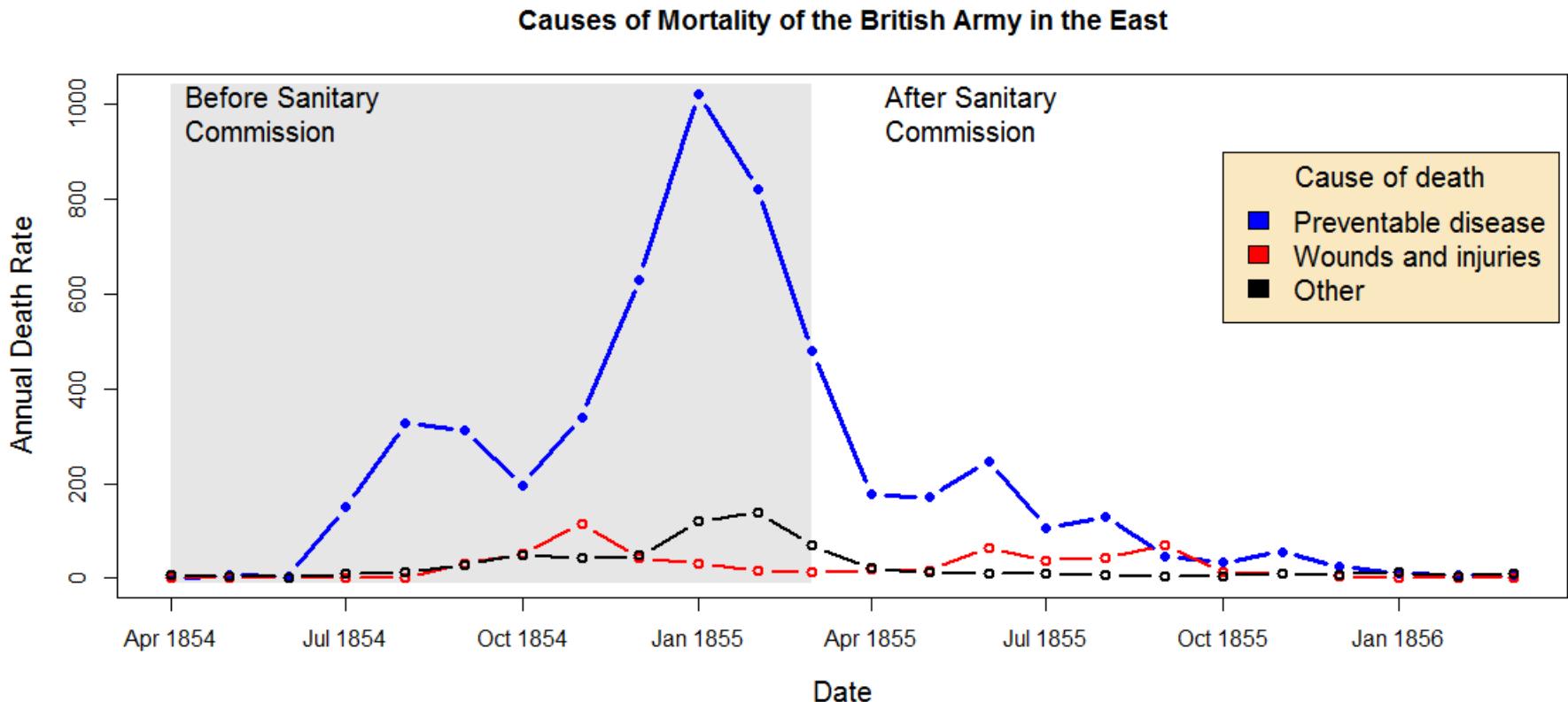
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic Diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes

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

# Data graph: Nightingale (1857)

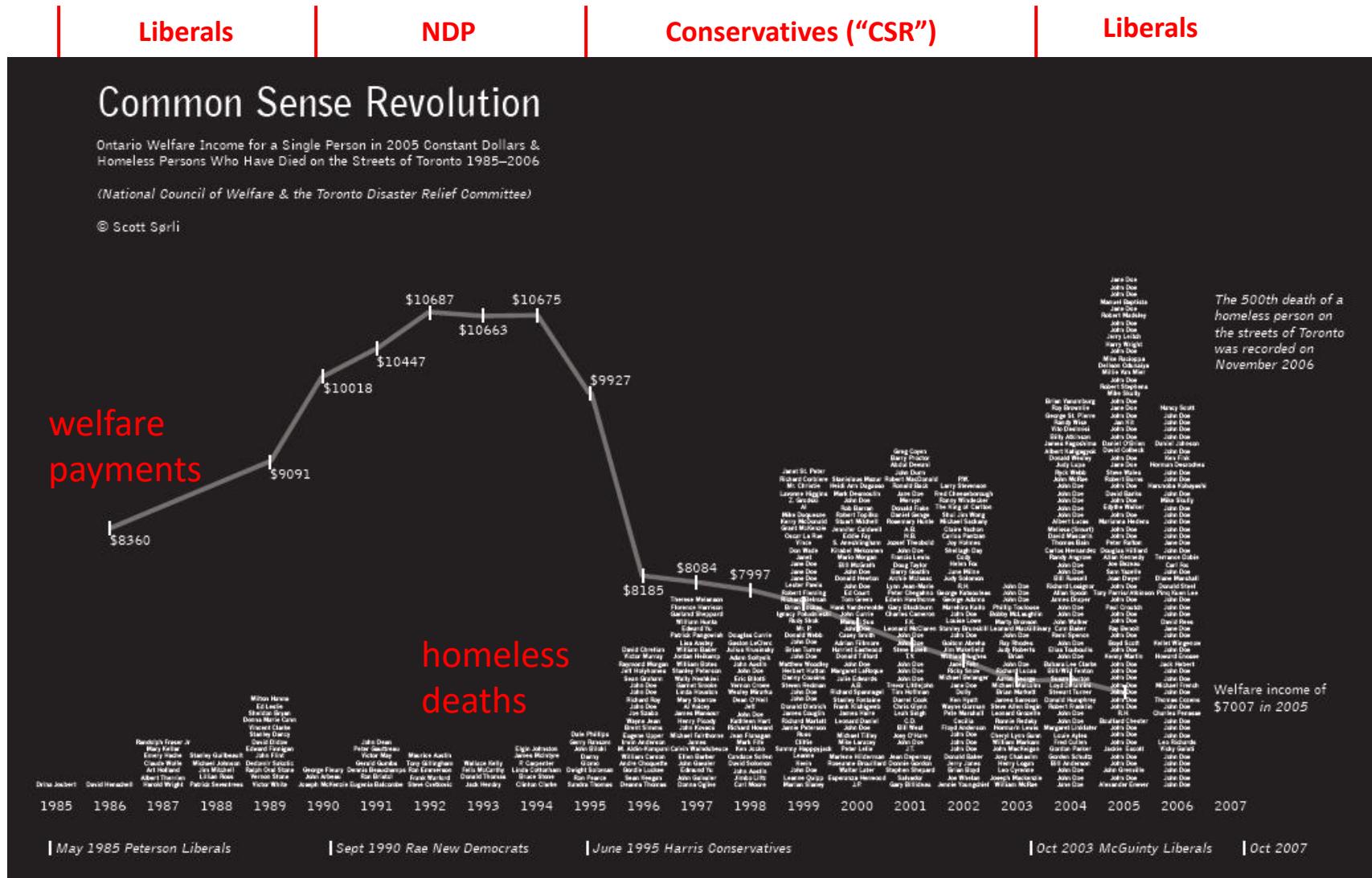
The same, as a data graph, using time-series line plots

Some statisticians might prefer this today, but it doesn't draw attention or interest as Flo's original did. **It likely would not have roused British Parliament to act!**



# Rhetorical graph: Welfare income and Homeless deaths after the “Common Sense Revolution”

Scott Sorli (2007)



# Analysis graph: Deaths vs. Income

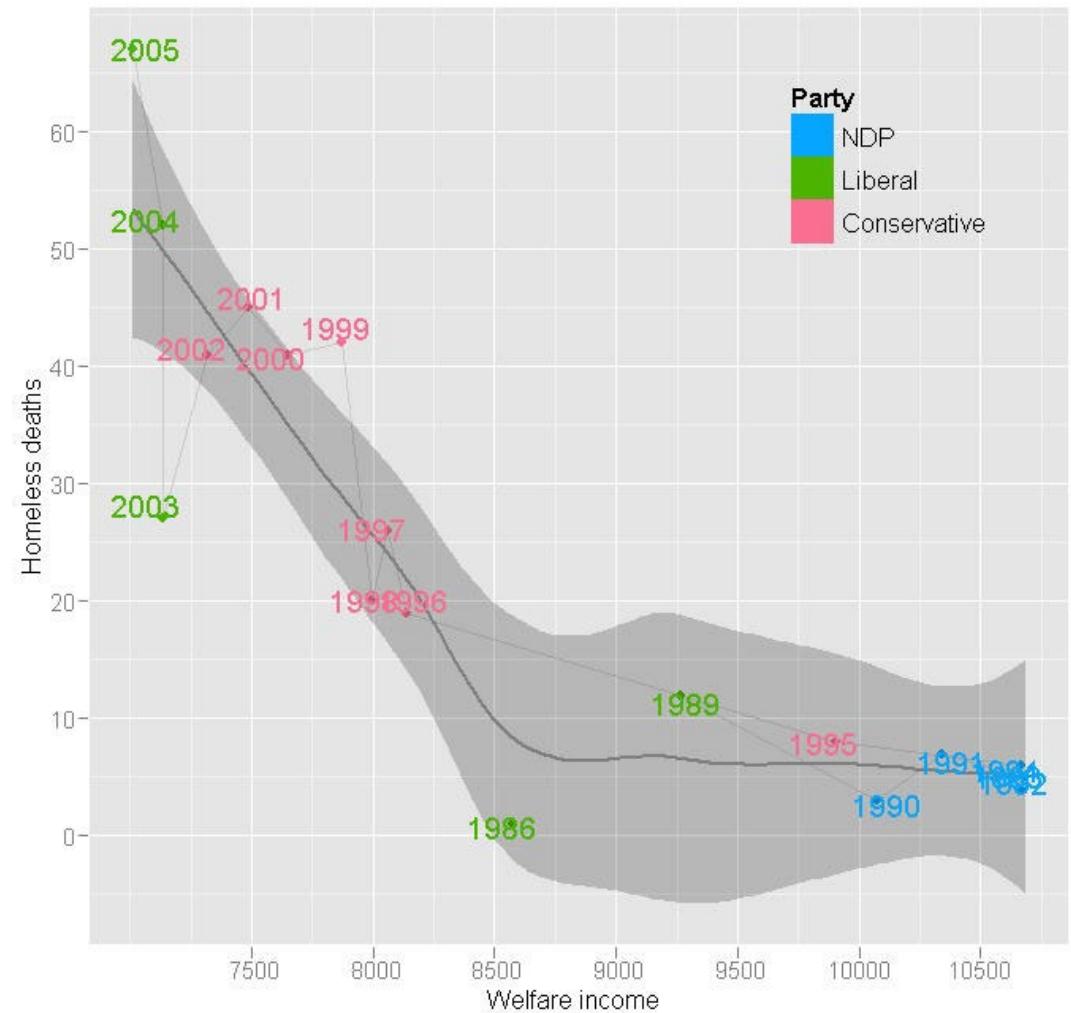
Scatterplot of deaths vs. income

- 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

Common Sense Revolution

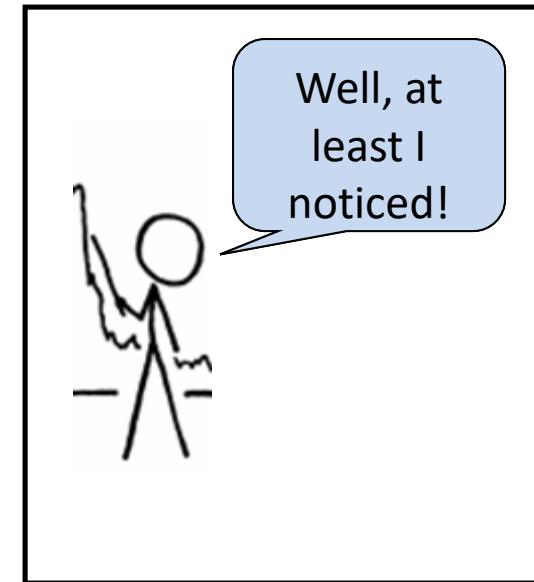
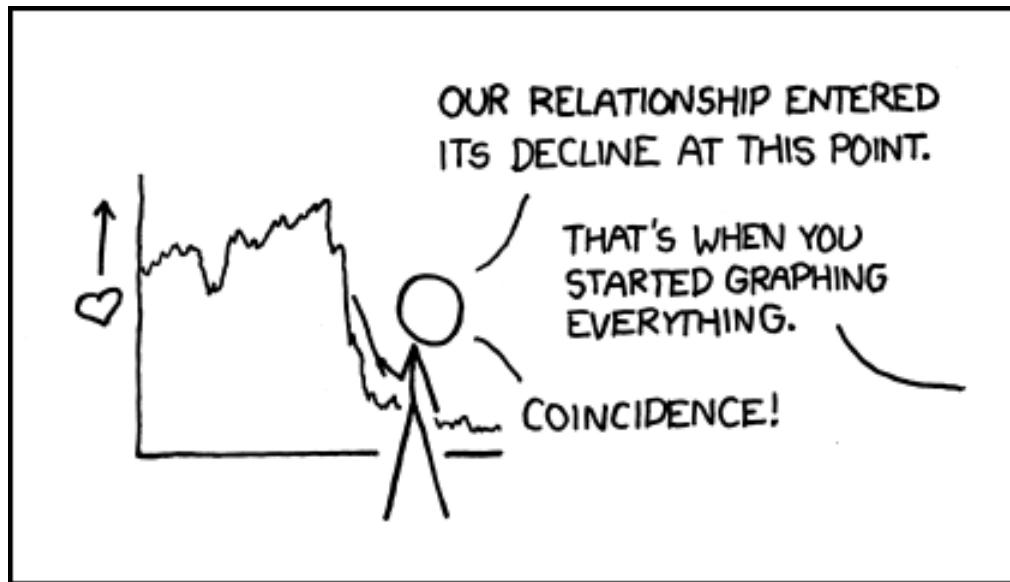


# Why plot your data?

Graphs help us to see

**patterns, trends, anomalies and other features**

not otherwise easily apparent from numerical summaries.

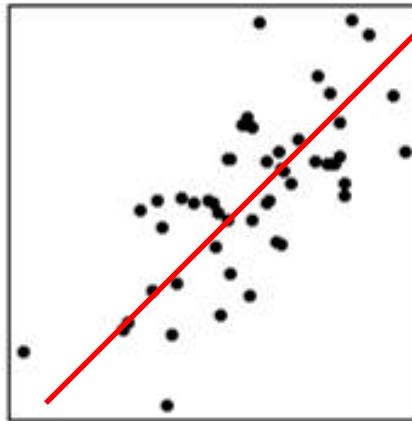


Source: <http://xkcd.com/523/>

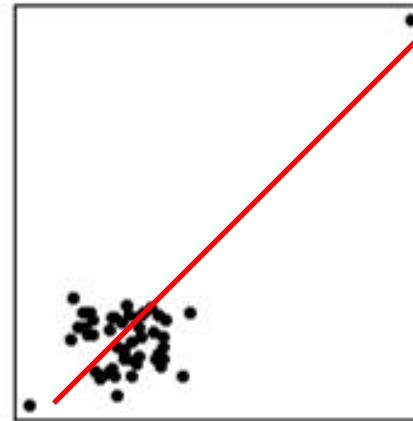
# Why plot your data?

Three data sets with exactly the same bivariate summary statistics:

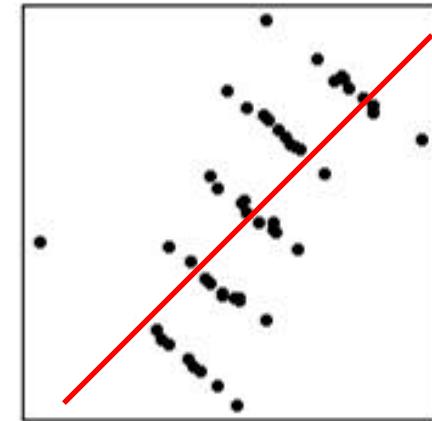
- Same correlations, linear regression lines, etc
- Indistinguishable from standard printed output
- Totally different interpretations!



Standard data



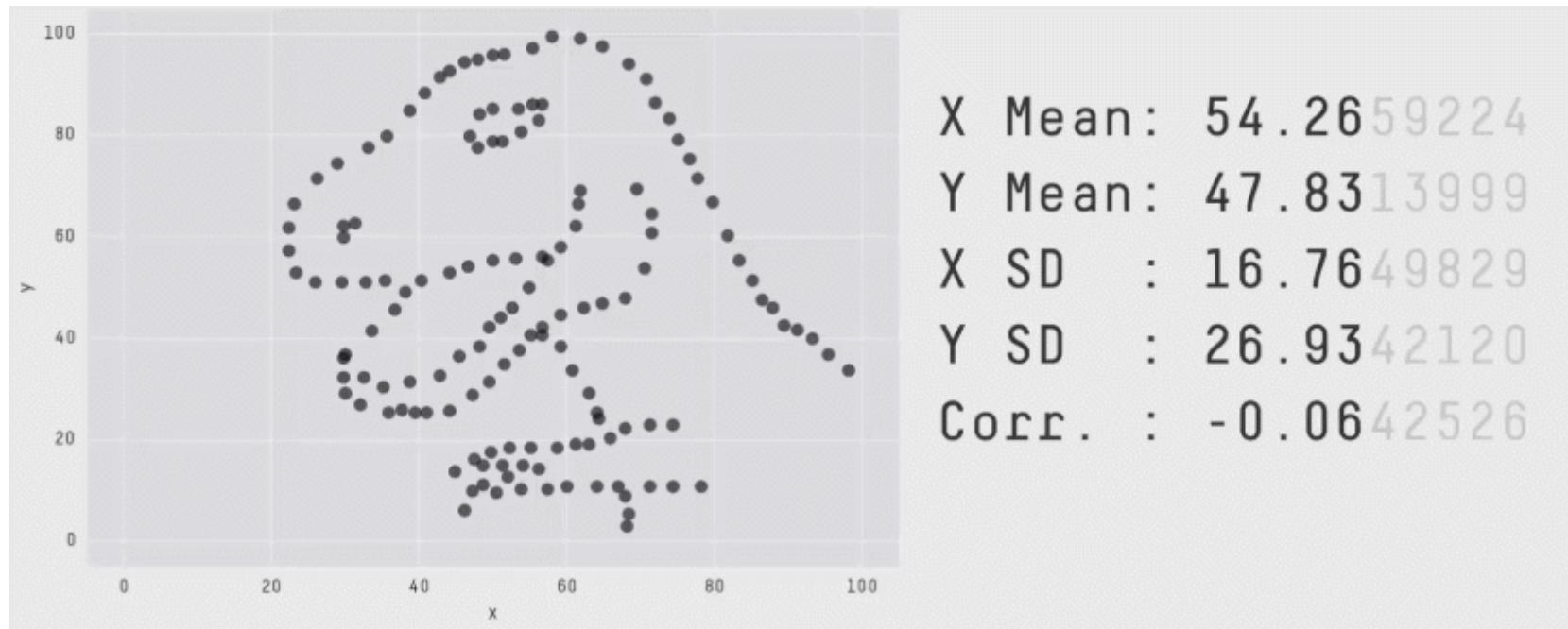
$r=0$  but + 2 outliers



Lurking variable?

# Datasaurus Dozen

As it happens, it is also possible to create any number of datasets with the same means, standard deviations and correlations with nearly any shape you like — even a dinosaur!



Method: Matejka & Fitzmaurice (2017). Same stats, different graphs. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* <https://doi.org/10.1145/3025453.3025912>  
Image source: [https://youtu.be/lI4UA75z\\_KQ](https://youtu.be/lI4UA75z_KQ)

# Comparing groups: Analysis vs. Presentation graphs

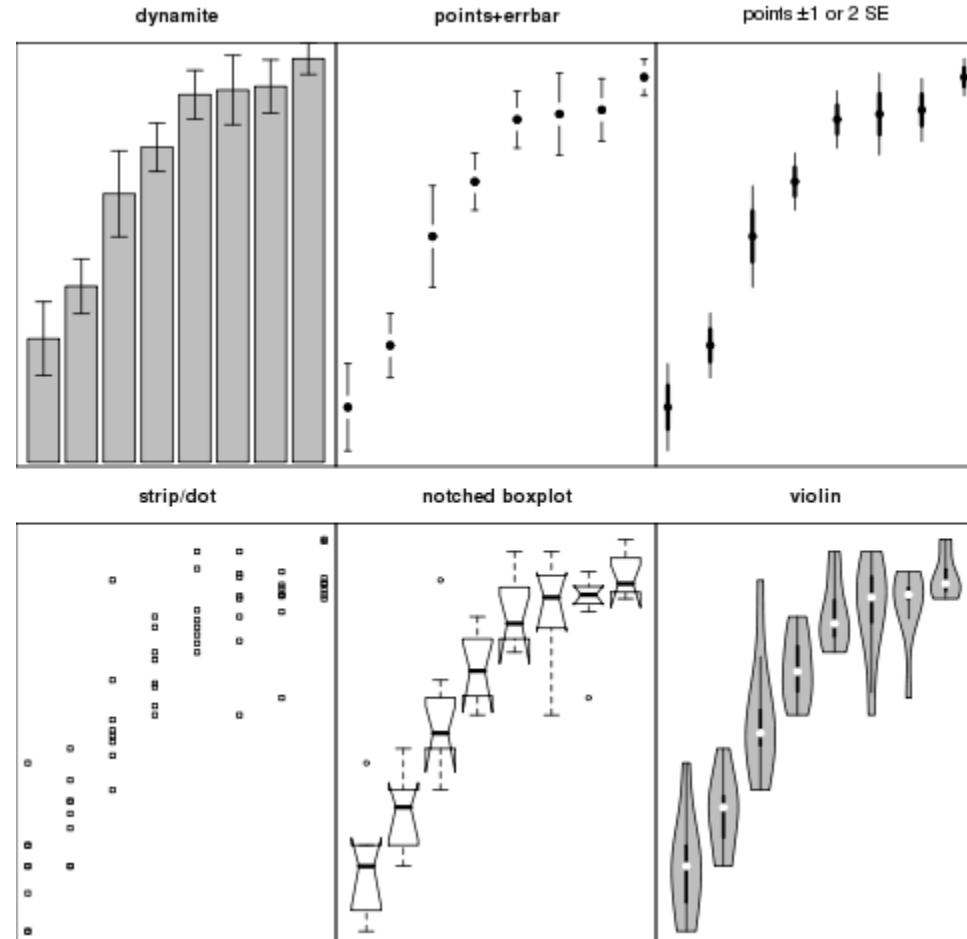
Six different graphs for comparing groups in a one-way design

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

Never use dynamite plots

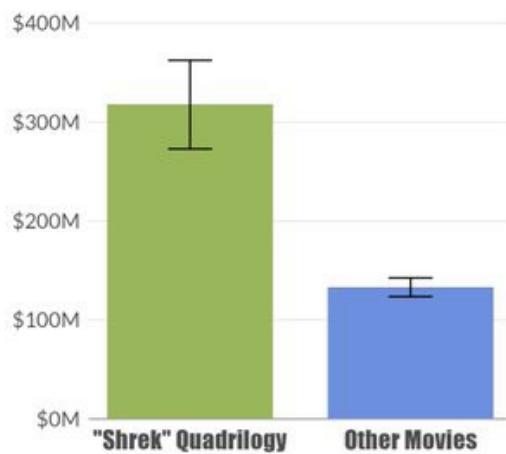
Always explain what error bars mean

Consider tradeoff between  
summarization & exposure

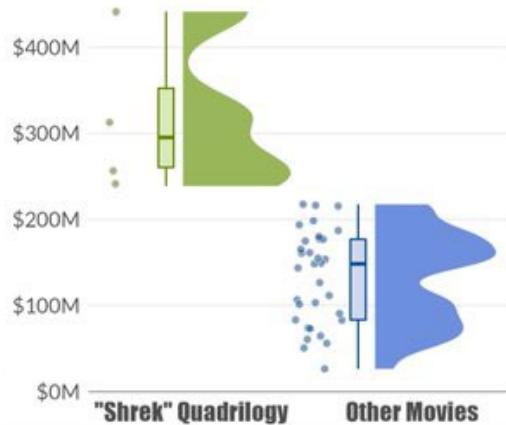




## Domestic Box Office of DreamWorks Movies



## Domestic Box Office of DreamWorks Movies



Dynamite plots – barcharts with error bars provide little useful information

Raincloud plots:  
Boxplots, violin plots show the important features

- Center
- Spread
- Shape

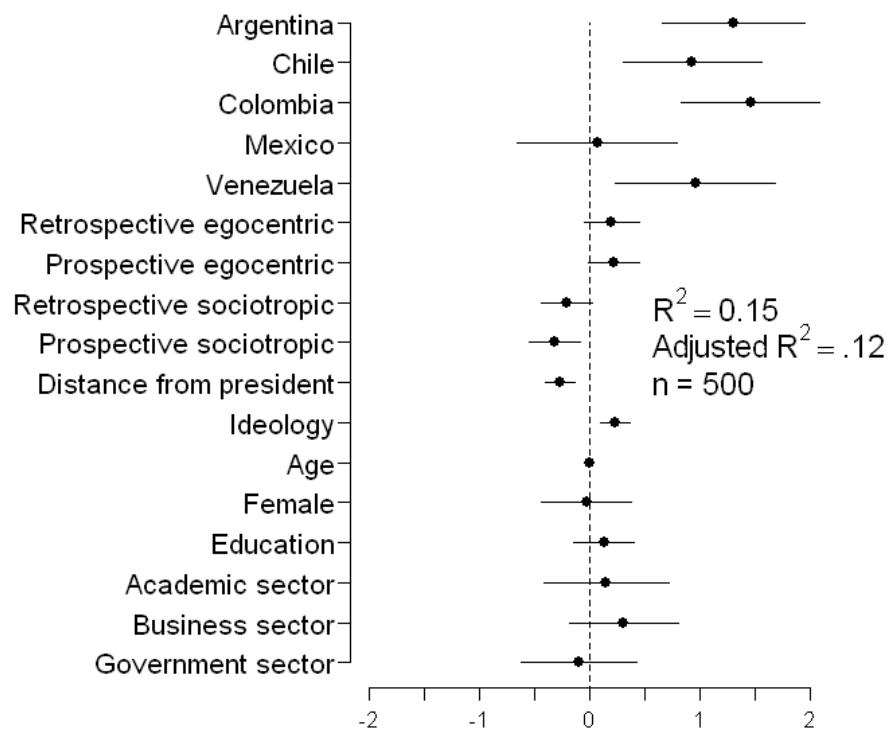
Dot plots show the data

# Presentation: Turning tables into graphs

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

Variable	Coefficient (Standard Error)
Constant	.41 (.93)
<b>Countries</b>	
Argentina	1.31 (.33)### B,M
Chile	.93 (.32)### B,M
Colombia	1.46 (.32) ### B,M
Mexico	.07 (.32) <sup>A,CH,CO,V</sup>
Venezuela	.96 (.37)## B,M
<b>Threat</b>	
Retrospective egocentric economic perceptions	.20 (.13)
Prospective egocentric economic perceptions	.22 (.12) <sup>#</sup>
Retrospective sociotropic economic perceptions	-.21 (.12) <sup>#</sup>
Prospective sociotropic economic perceptions	-.32 (.12) <sup>##</sup>
Ideological Distance from president	
<b>Ideology</b>	
Ideology	.23 (.07) ###
<b>Individual Differences</b>	
Age	.00 (.01)
Female	-.03 (.21)
Education	.13 (.14)
Academic Sector	.15 (.29)
Business Sector	.31 (.25)
Government Sector	-.10 (.27)
R <sup>2</sup>	.15
Adjusted R <sup>2</sup>	.12
n	500
### p < .01, ## p < .05, # p < .10 (two-tailed)	
<sup>A</sup> Coefficient is significantly different from Argentina's at p < .05;	
<sup>B</sup> Coefficient is significantly different from Brazil's at p < .05;	
<sup>CH</sup> Coefficient is significantly different from Chile's at p < .05;	
<sup>CO</sup> Coefficient is significantly different from Colombia's at p < .05;	
<sup>M</sup> Coefficient is significantly different from Mexico's at p < .05;	
<sup>V</sup> Coefficient is significantly different from Venezuela's at p < .05	

Graphs of model coefficients are often clearer than tables



From: Kastellec & Leoni, 2007. "Using Graphs Instead of Tables in Political Science." *Perspectives on Politics* 5 (4): 755–71.  
See: The R dotwhisker pkg, <https://fsolt.org/dotwhisker/>

# Semi-graphic tables

Tables can be made more visual by including mini-graphics in some table cells  
The original idea— [sparklines](#) (Tufte), is now more general

A good solution for the [Table 1 problem](#) – description of your sample

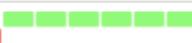
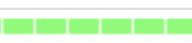
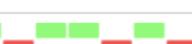
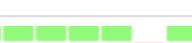
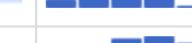
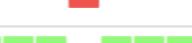
[13]:	index	id	name	line	bar	tristate	box
0	1	Oli Bob					
1	2	Mary May					
2	3	Christine Lobowski					
3	4	Brendon Philips					
4	5	Margret Marmajuke					
5	6	Frank Harbours					
6	7	Jamie Newhart					
7	8	Gemma Jane					
8	9	Emily Sykes					
9	10	James Newman					

Image from: <https://discourse.holoviz.org/t/add-sparkline-formatters-to-tabulator/3197>

# 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?

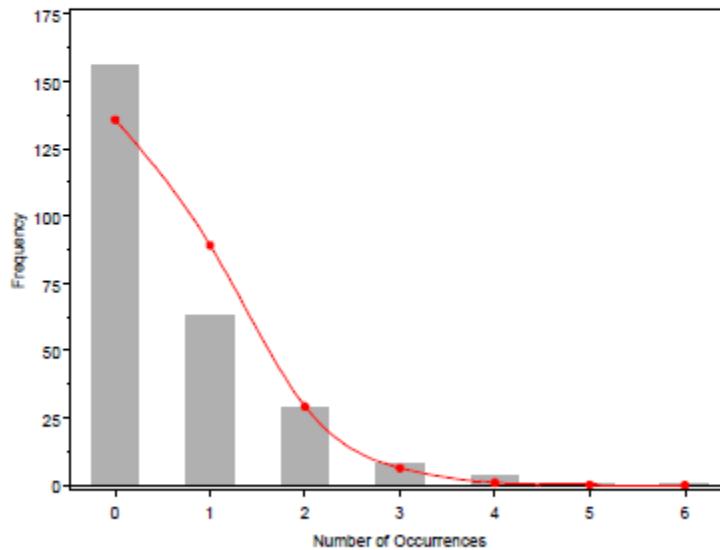
From: Ben Jones, To Optimize or to Satisfice in Data Visualization? <https://dataremixed.com/2016/01/optimize-or-satisfice-in-dataviz/>

# Effective data display

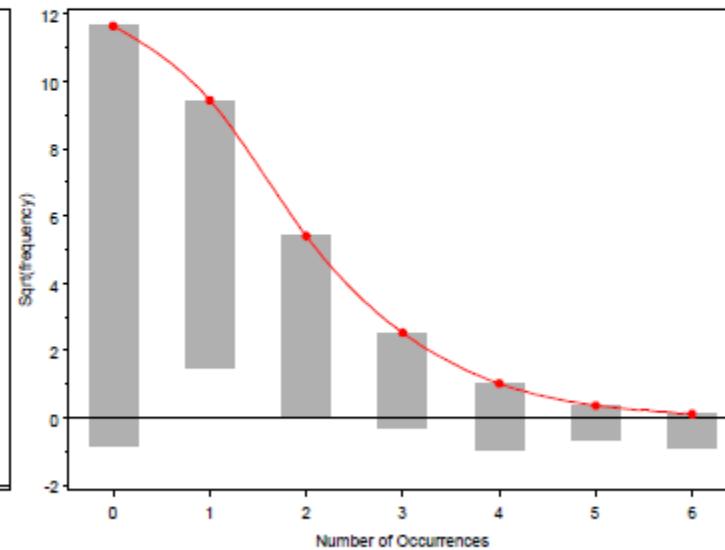
- Make the data stand out
  - Fill the data region (axes, ranges)
  - Use visually distinct symbols (shape, color) for different groups
  - Avoid chart junk, heavy grid lines that detract from the data
- Facilitate comparison
  - Emphasize the important comparisons visually
  - Side-by-side easier than 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

# 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



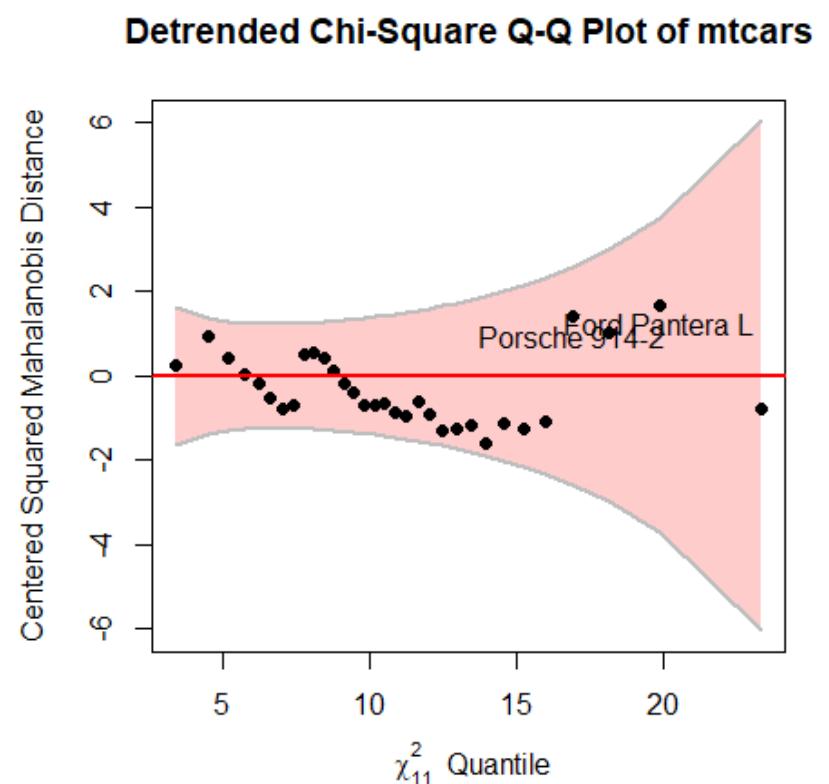
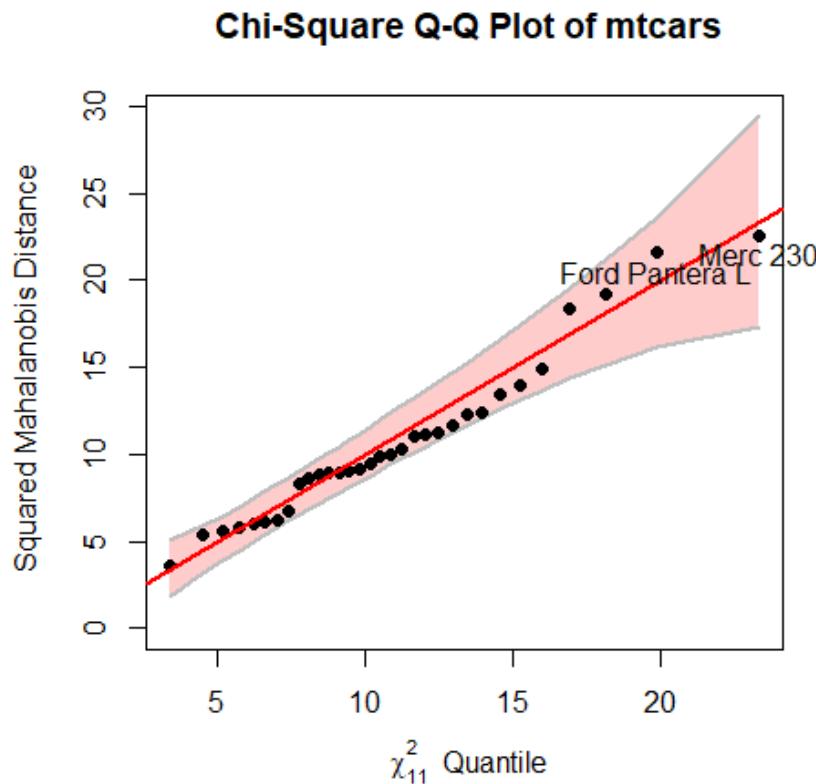
Standard histogram with fit



Suspended rootogram

# Make visual comparisons easy

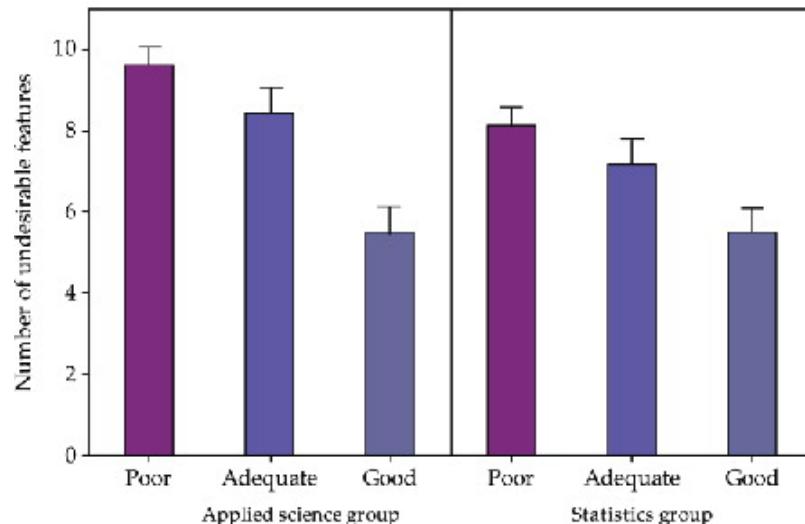
QQ plots assess the degree to which data follows a particular distribution (normal,  $\chi^2$ )  
The reference line of equality is usually a  $45^\circ$  line  
It is easier to see any departure in relation to a [horizontal line](#) at 0



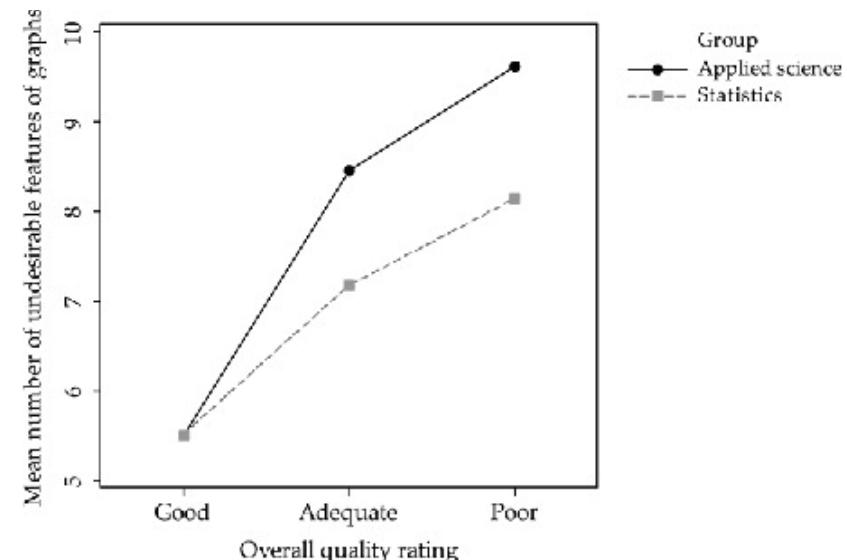
# Make comparisons *direct*

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

Is there evidence of an interaction here?



???

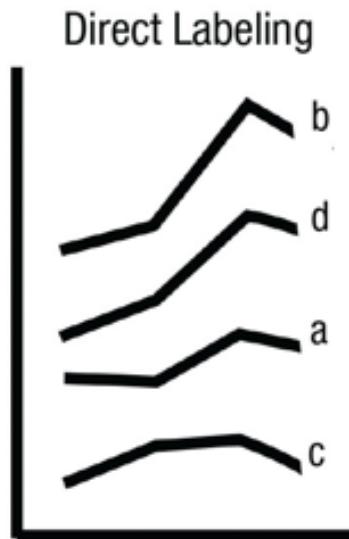


YES

# Direct labels vs. legends

Direct labels for points, lines and regions are usually easier and faster than legends

- Give the names of the four groups shown in the line graph at left in top-to-bottom order.  
(Answer: b, d, a, c.)

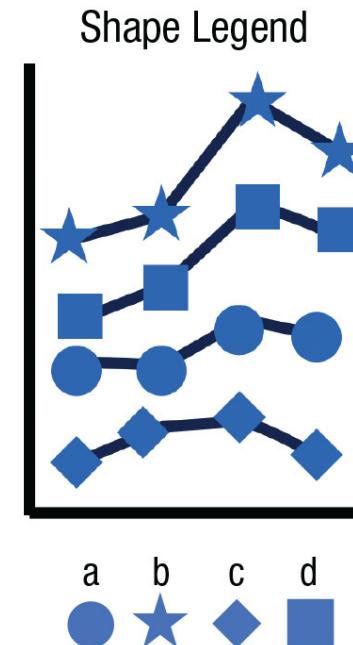
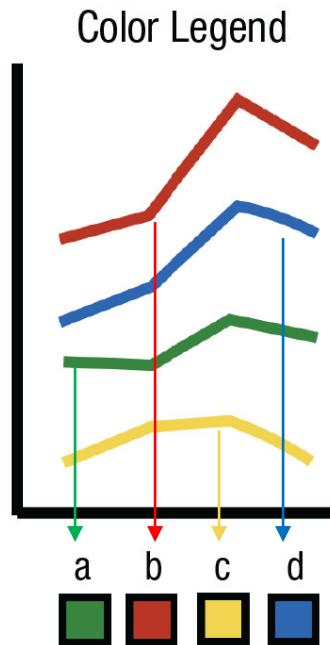
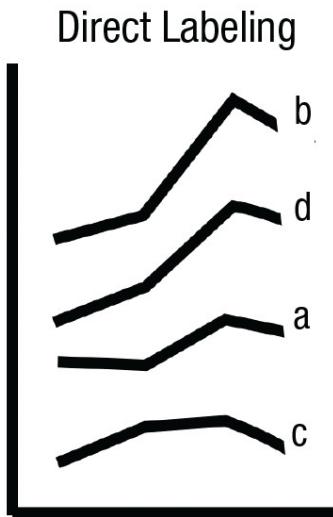


Source: Franconeri et al. DOI:  
[10.1177/15291006211051956](https://doi.org/10.1177/15291006211051956)

# Direct labels vs. legends

Direct labels for points, lines and regions are usually easier and faster than legends

- Give the names of the four groups shown in the line graph at left in top-to-bottom order.  
(Answer: b, d, a, c.)
- Now do the same for the graphs using **color** or **shape** legends
- You need to look back and forth between the graph and legend

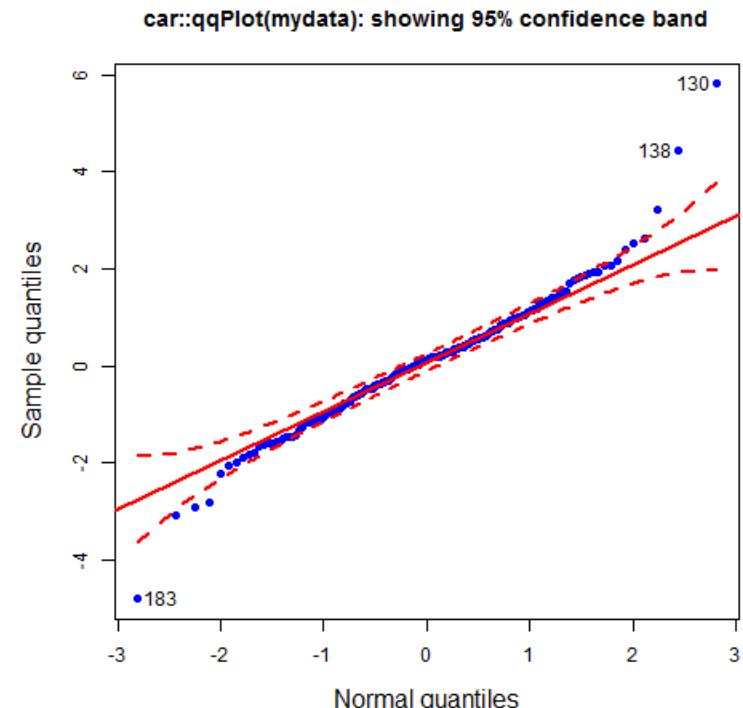
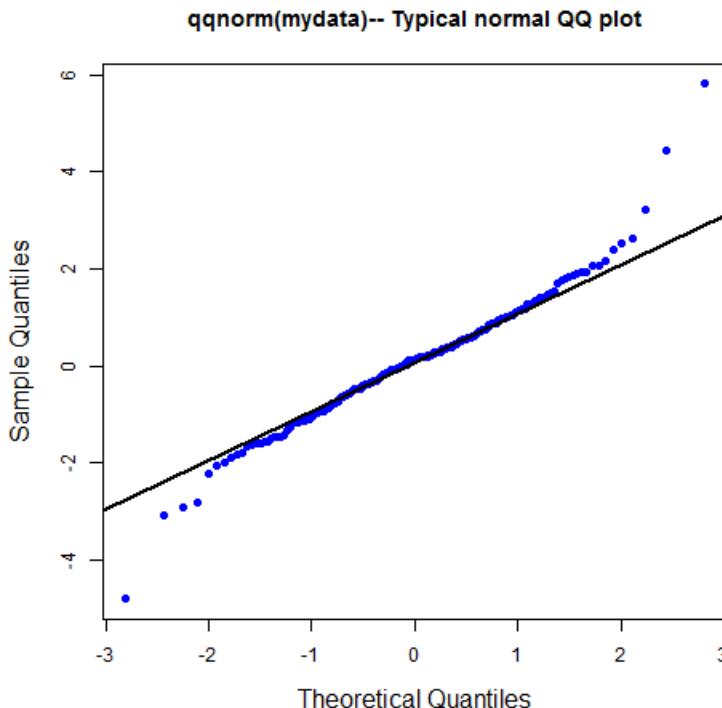


Source: Franconeri et al. DOI:  
[10.1177/15291006211051956](https://doi.org/10.1177/15291006211051956)

# Visualizing uncertainty

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

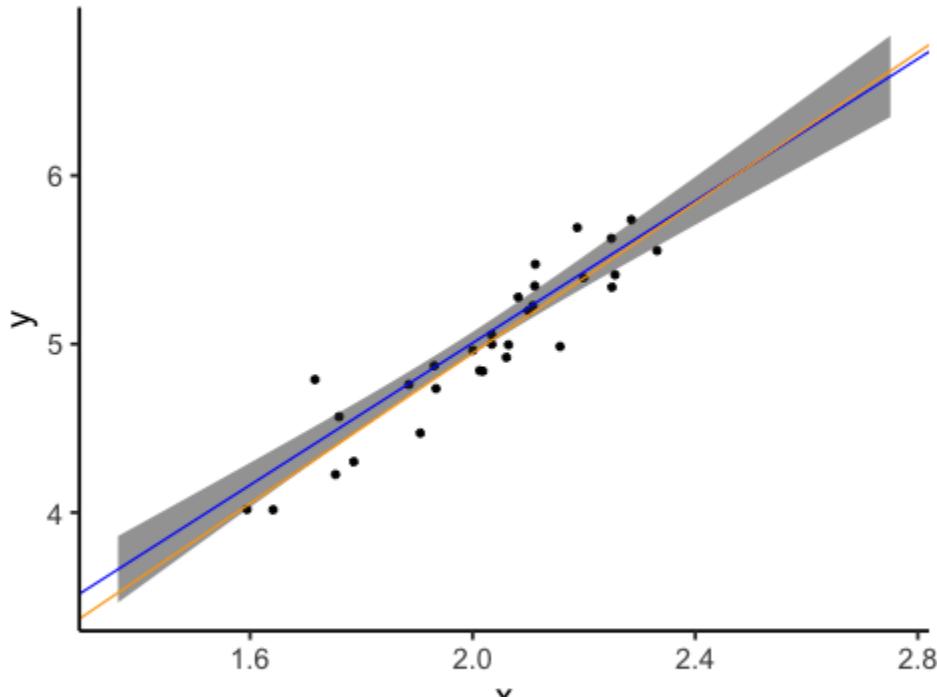
Example: Normal QQ plots used to assess normality of data



# Uncertainty: Theory & simulation

## Effect of CI under simulations

Simulation 1



blue: single estimated linear regression line,  
orange: 20 simulated linear regression lines,  
gray: 95% CI for the blue line

In most cases, standard theory allows calculations of uncertainty intervals (confidence bands)

Simulation provides a theory-free alternative

Viewed as an animation gives a better sense of variation from sample to sample

# 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

Friendly & Kwan (2003). Effect Ordering for Data Displays. *Computational Statistics and Data Analysis*, **43**, 509–539. DOI: [10.1016/S0167-9473\(02\)00290-6](https://doi.org/10.1016/S0167-9473(02)00290-6)

# 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							
Launch No.	Erosion Depth [in.]	Cross Sectional View			Top View		Clocking Location [deg]
		Perimeter Affected [deg]	HORIZONTAL Dist. [in.]	Length Of Max Erosion [in.]	Total Heel Affected Length [in.]		
61A LH Center Field**	228	None	None	0-160	None	None	36°--60°*
61A LH CENTER FIELD**	22A	None	None	0-280	None	None	33°-18°
SIC LH Forward Field**	15A	0-010	154.0	0-260	4.25	5.75	162
SIC RH Center Field (prim)***	15B	0-038	130.0	0-280	12.50	58.75	254
SIC RH Center Field (sec)***	15B	None	45.0	0-280	None	29.50	304
41B RH Forward Field	13B	0-028	110.0	0-280	3.00	None	275
41C LH Aft Field*	11A	None	0-280	None	None	--	--
41B LH Forward Field	10A	0-040	217.0	0-280	3.00	14.50	361
STS-2 RH Aft Field	2B	0-053	115.0	0-280	--	--	10

# 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

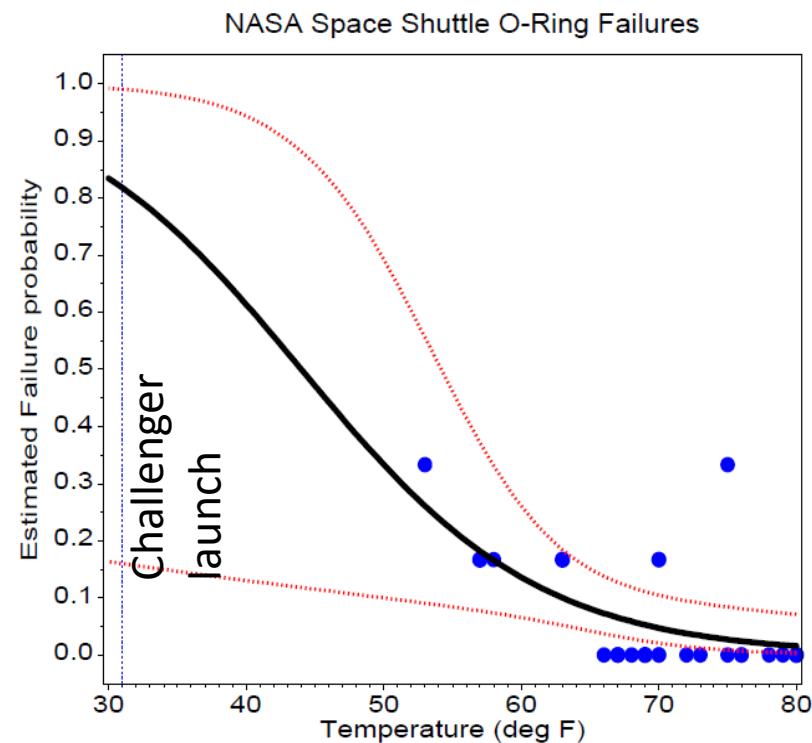


# Visual explanation: Graphics

- 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.

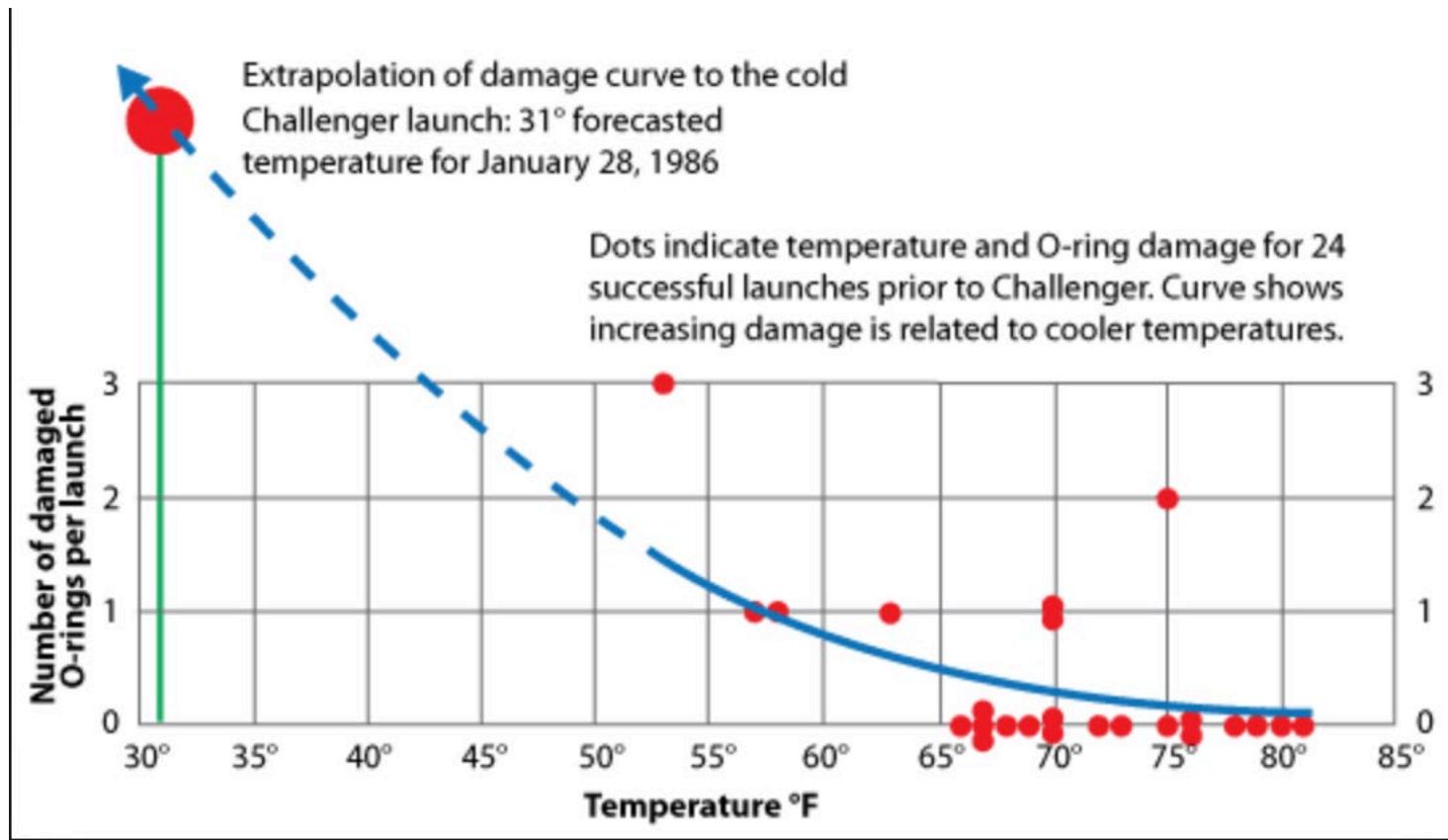
The graph shown here is the result of a statistical model fit to the data

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



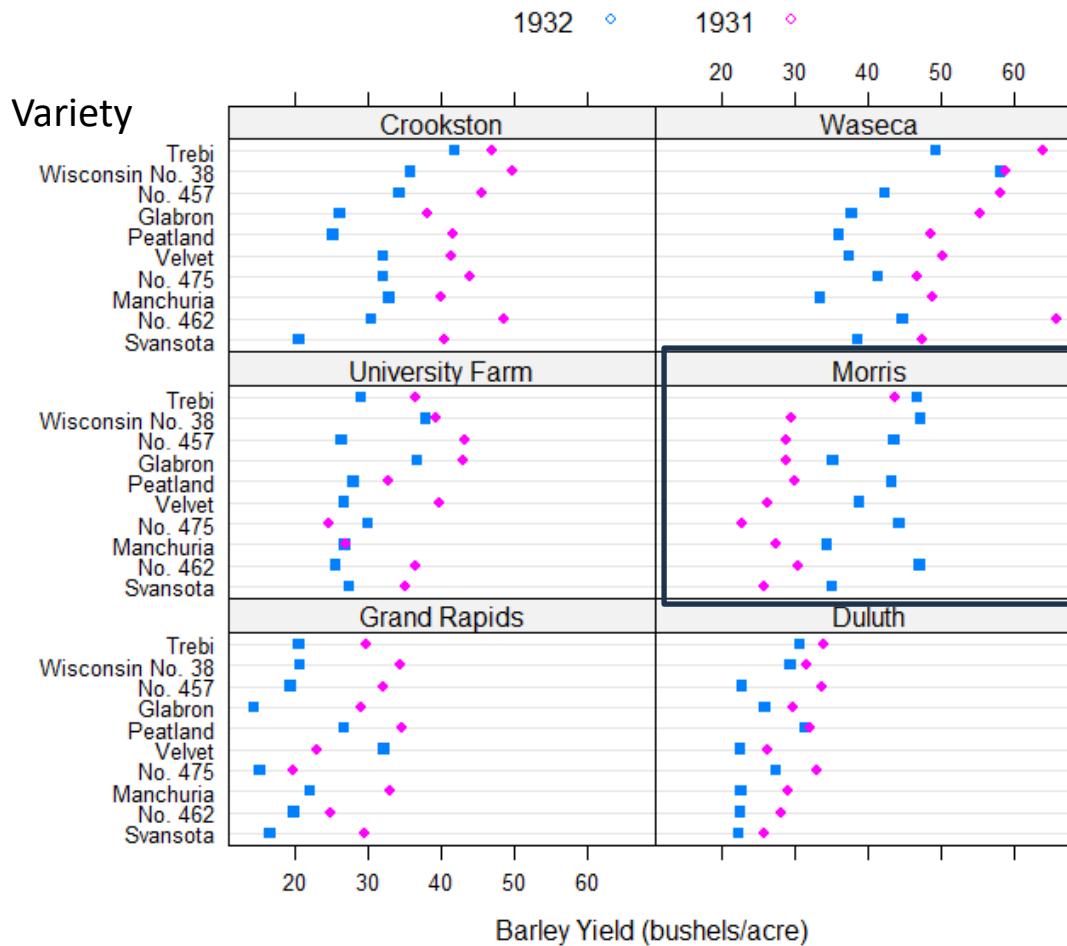
# Presentation graphic

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						Mean
	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
<b>Mean</b>	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						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

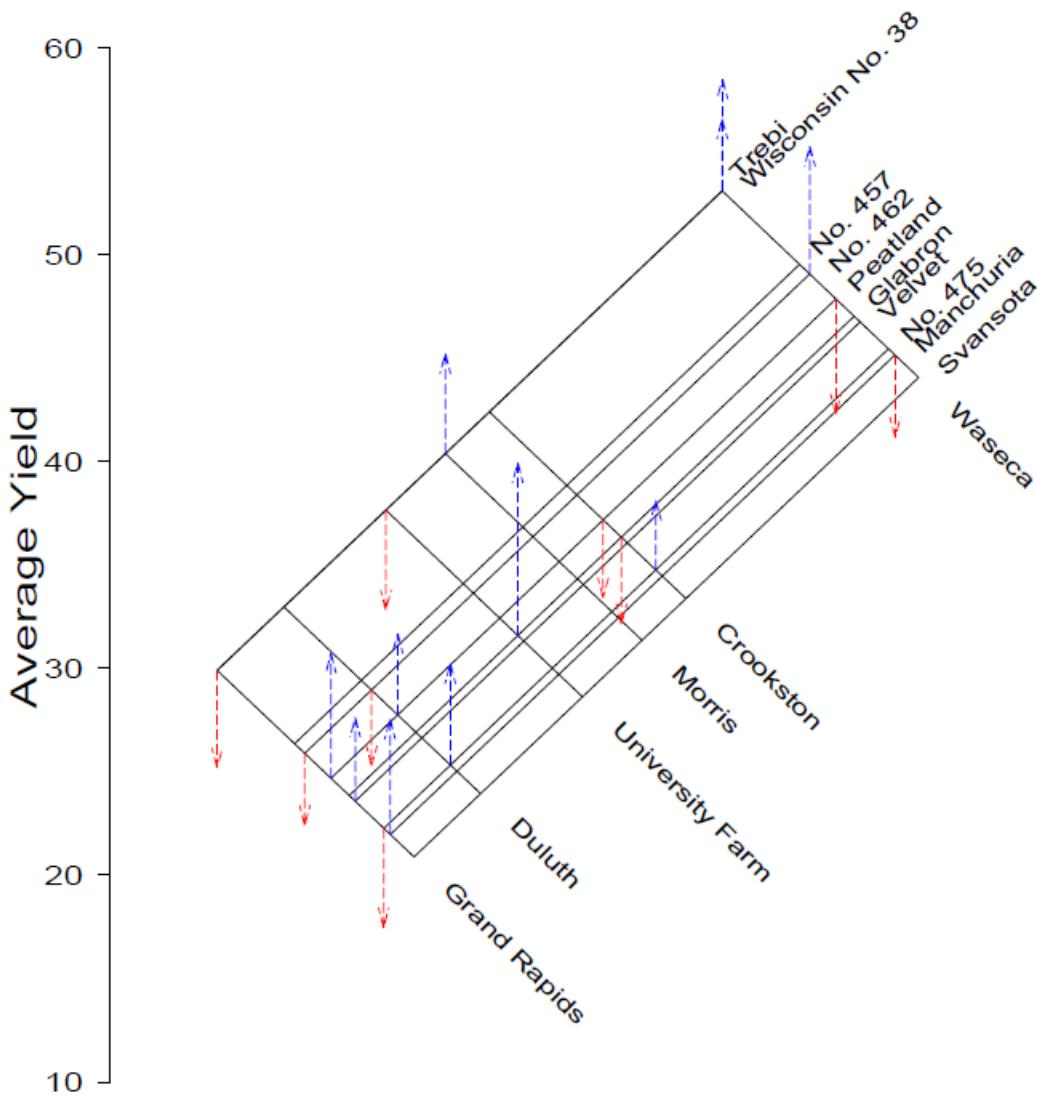
Ordered: by row and column means; shaded: by value ( $|\Delta y_{ij}| > \{2,3\} \times \sigma(\Delta y_{ij})$ )

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
<b>Mean</b>	-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,

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

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

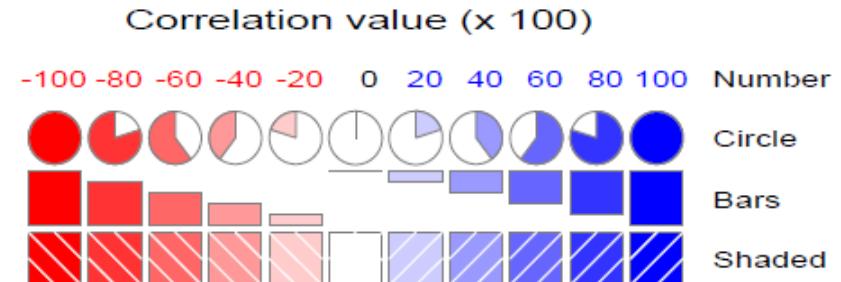
```
> cor(bb)
      Assists Atbat   Errors   Hits   Homer logSal Putouts   RBI   Runs   Walks   Years
Assists  1.0000 0.3421  0.70350 0.3040 -0.16160  0.0500 -0.0434  0.0629  0.179  0.1025 -0.0851
Atbat    0.3421 1.0000  0.32558 0.9640  0.55510  0.4149  0.3096  0.7960  0.900  0.6244  0.0127
Errors   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
```

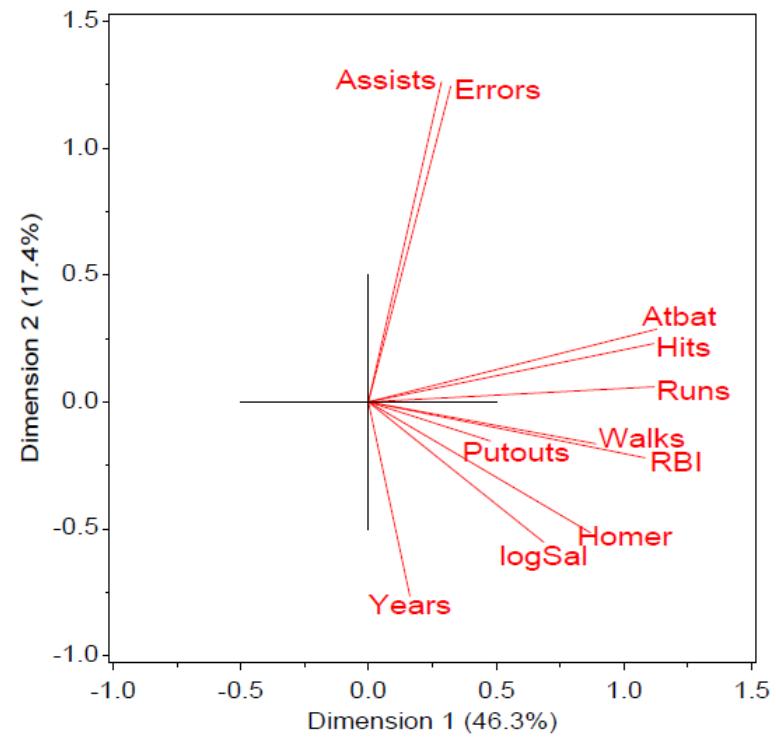
# Correlation ordering: corrgrams

**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

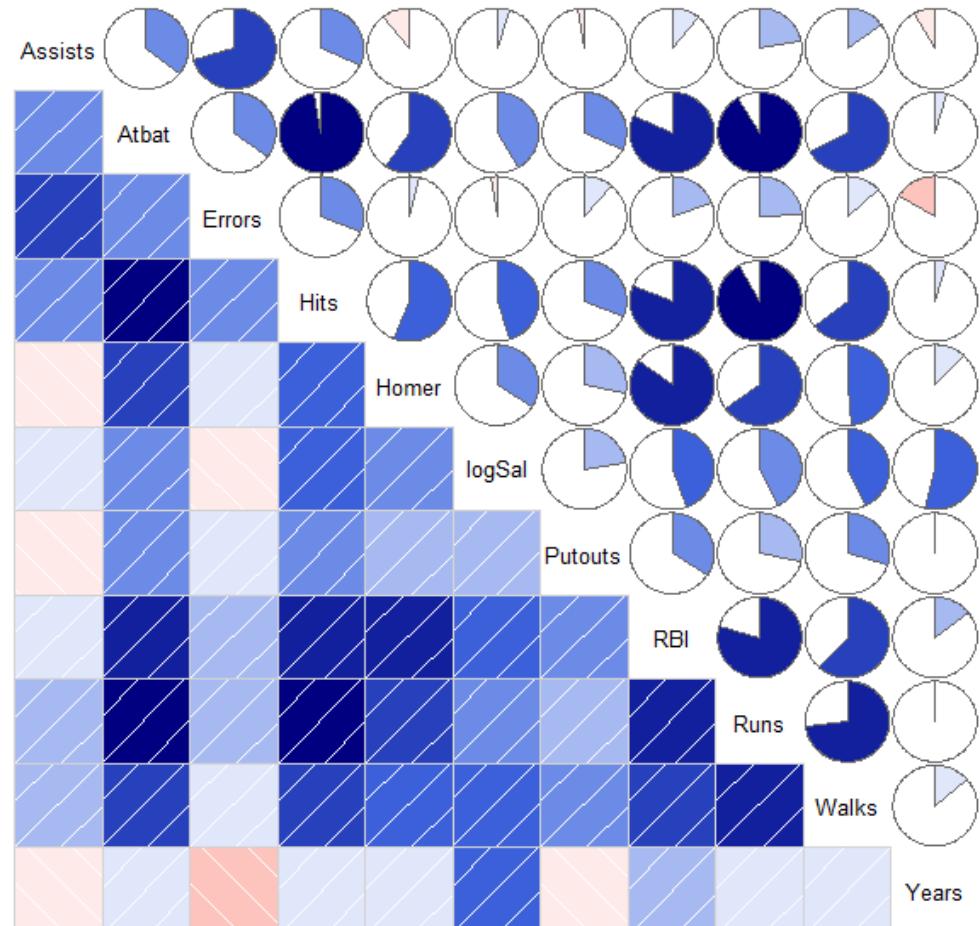


Friendly (2002), *Corrgrams: Exploratory displays for correlation matrices*, American Statistician.

# Baseball data

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

Baseball data alphabetic order

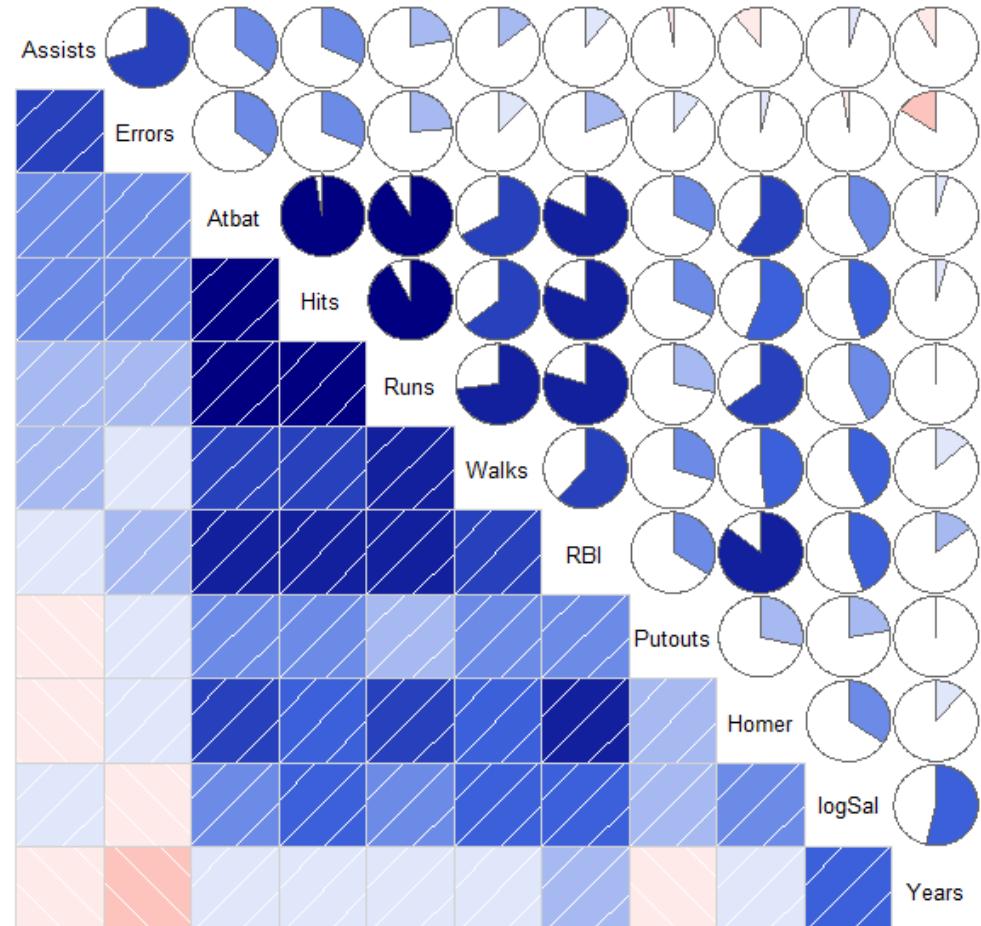


# Baseball data

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

Baseball data PC2/PC1 order





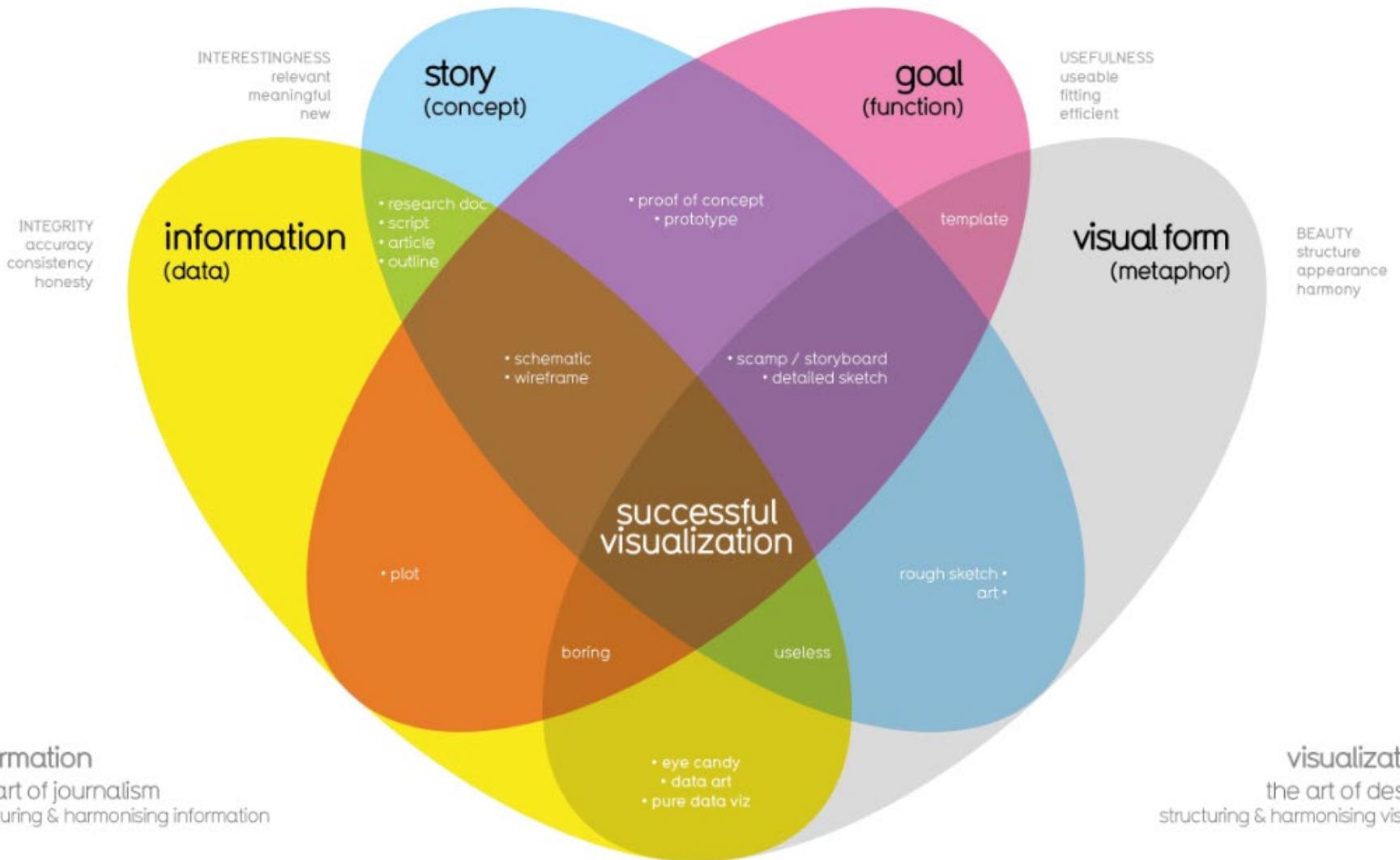
# Good/excellent graphs



- 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**
- What makes a good graph?
  - Integrity (quality of information)
  - Story (interesting, meaningful)
  - Goal (usefulness)
  - Visual form (beauty)

# What Makes a Good Visualization?

explicit (implicit)



information  
the art of journalism  
structuring & harmonising information

visualization  
the art of design  
structuring & harmonising visuals

David McCandless  
[InformationisBeautiful.net](http://InformationisBeautiful.net)

taken from new book  
*Knowledge is Beautiful*

find out more  
[bit.ly/KIB\\_Books](http://bit.ly/KIB_Books)

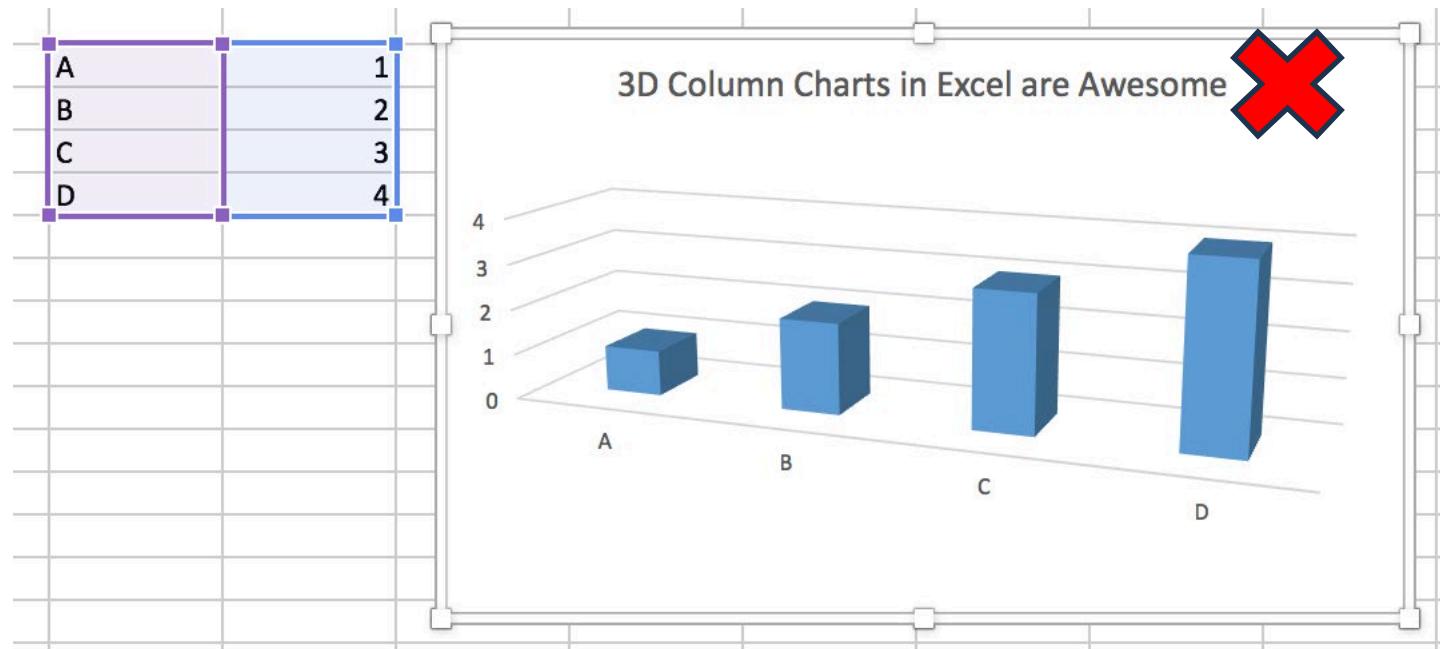
# Bad/evil graphs



- 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

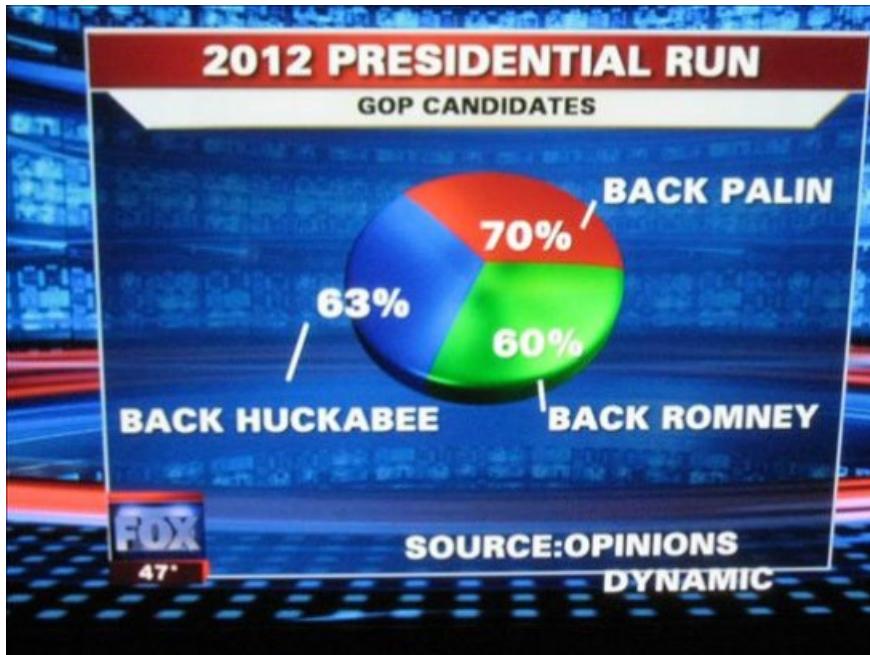
Friends don't let friends use Excel for data visualization or statistics



How many things are wrong with this graph?

# Pie charts are easy to abuse

What's wrong with these pictures?



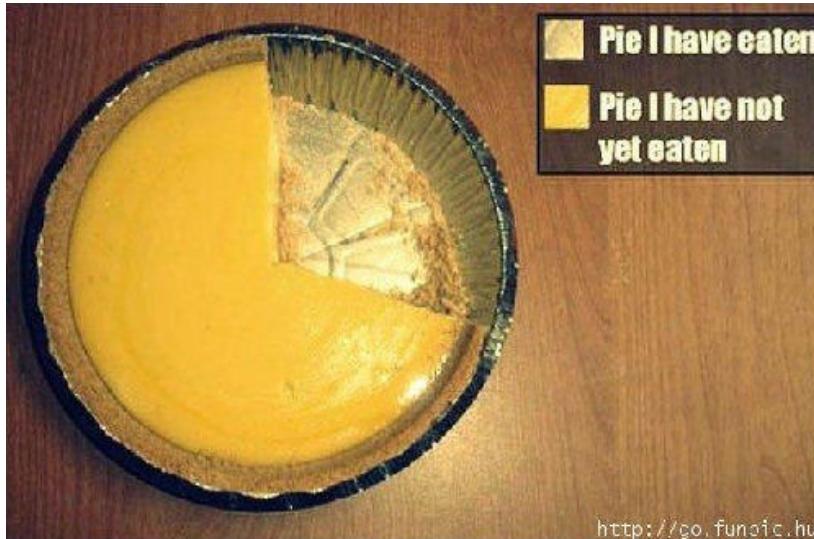
$$1 \pi = 193\%$$



$$1 \pi = 178\%$$

Why do graphic designers so often get this wrong?

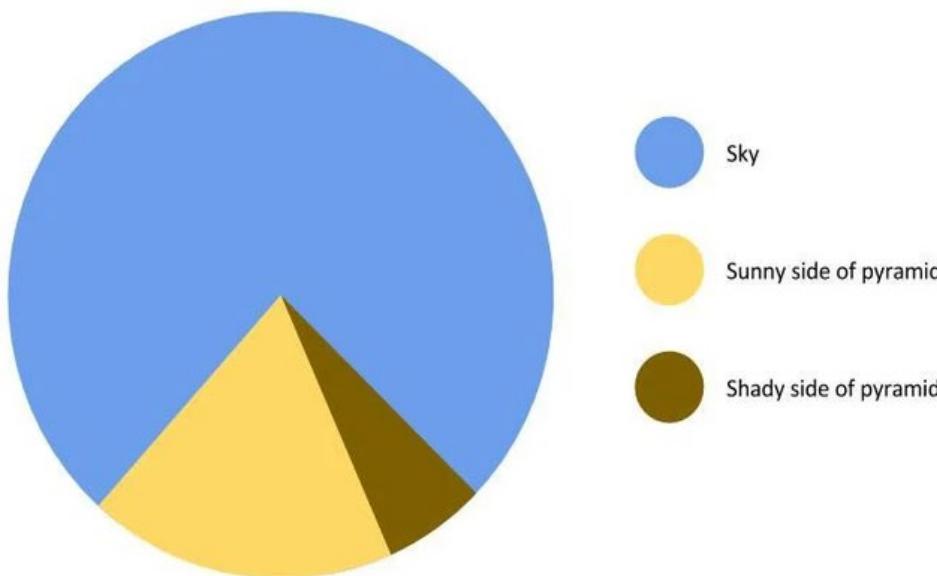
# Pie chart merriment



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

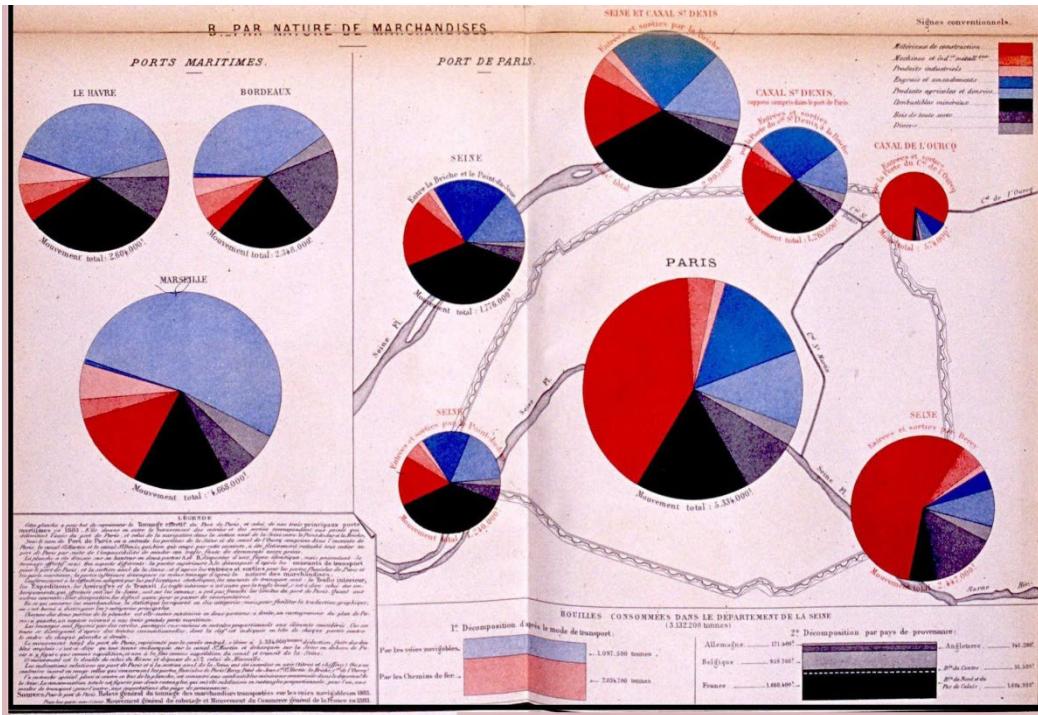
But: how much  $\pi$  have I eaten?

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

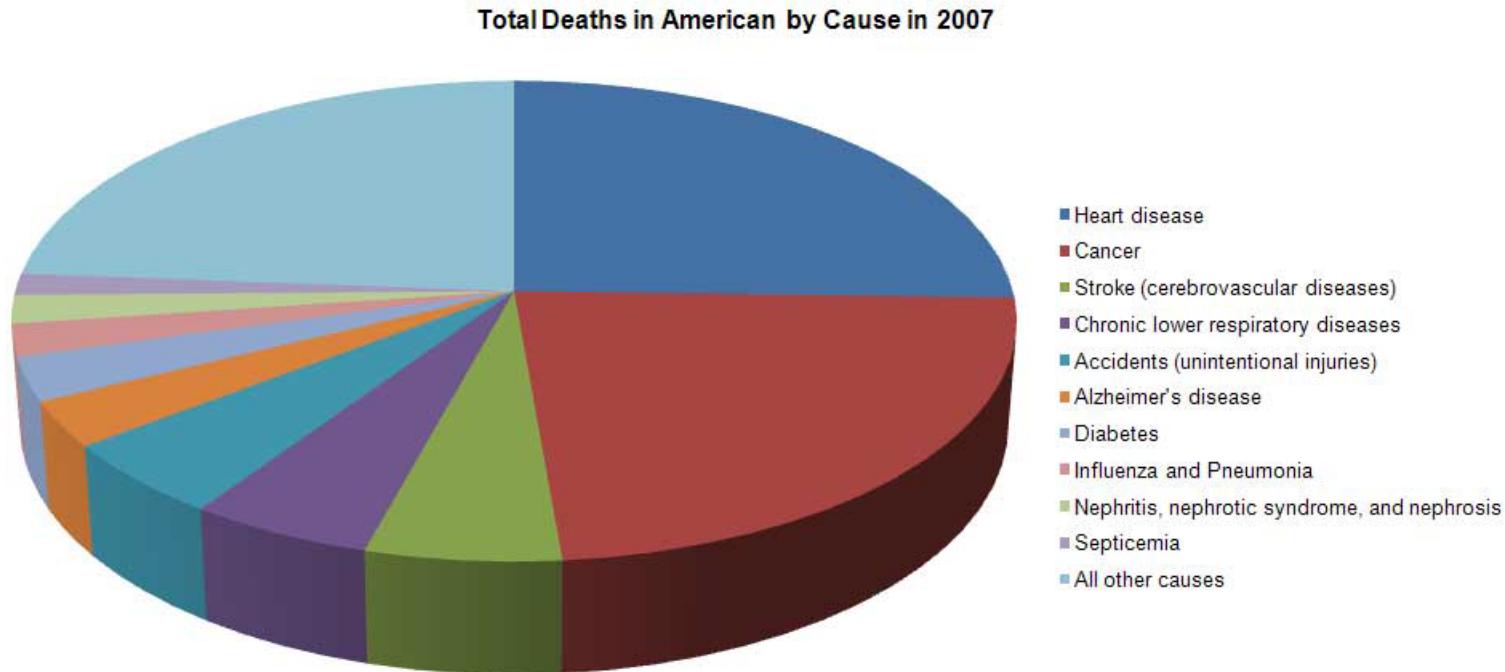


This graphic uses pie charts to show the transport of different kinds of goods to the ports of Paris and the principal maritime ports

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

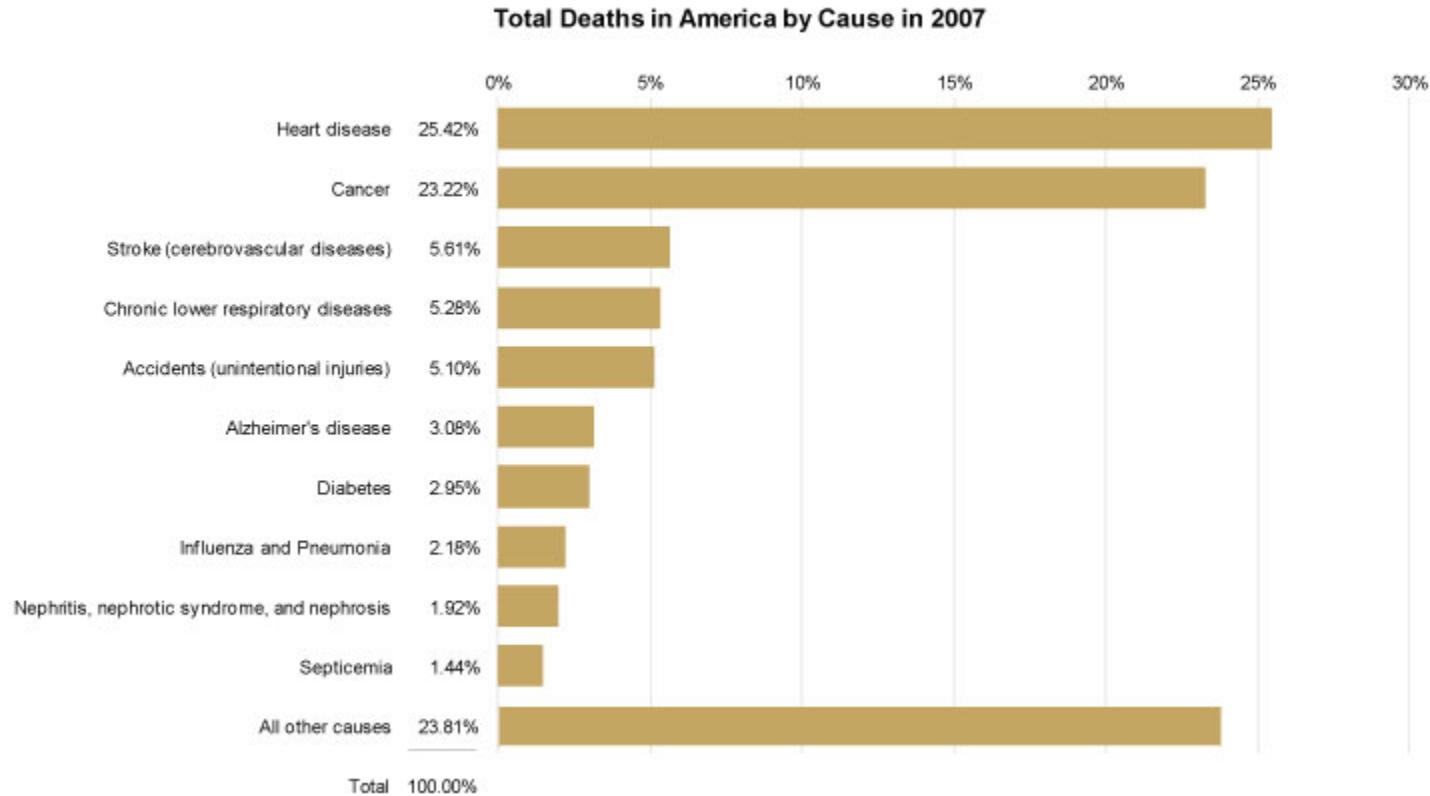
Émile Cheysson, Mouvement de la Navigation Dans le Port de Paris et les Principaux Ports Français en 1883.  
Album de Statistique Graphique, 1885, plate 17. <https://bit.ly/47DwnKU>

# 3D pie charts are usually evil

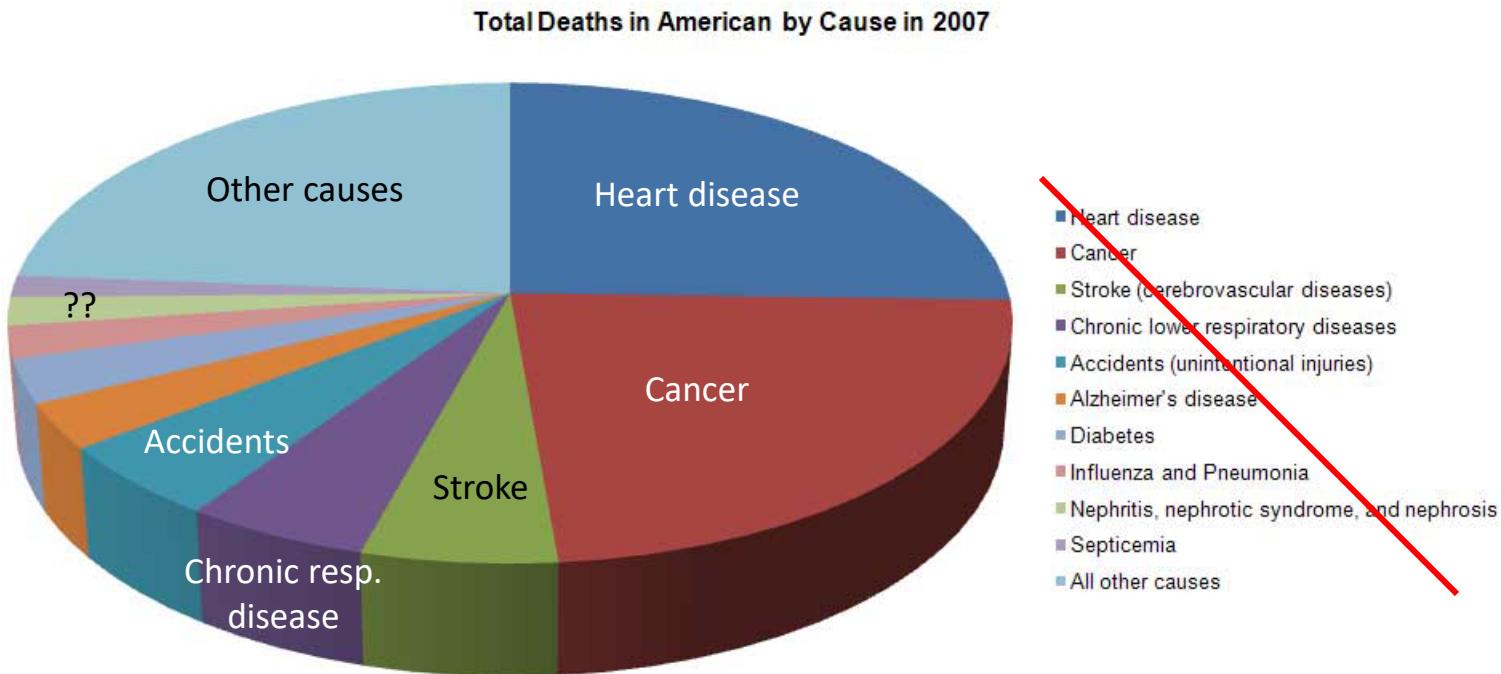


What was the intent of the designer of this graphic?  
Which category led to the greatest total deaths?  
What was the proportion of deaths due to strokes?  
Did more people die from strokes vs. accidents?

# Simple re-design makes it clearer



# Or, try to use direct labels



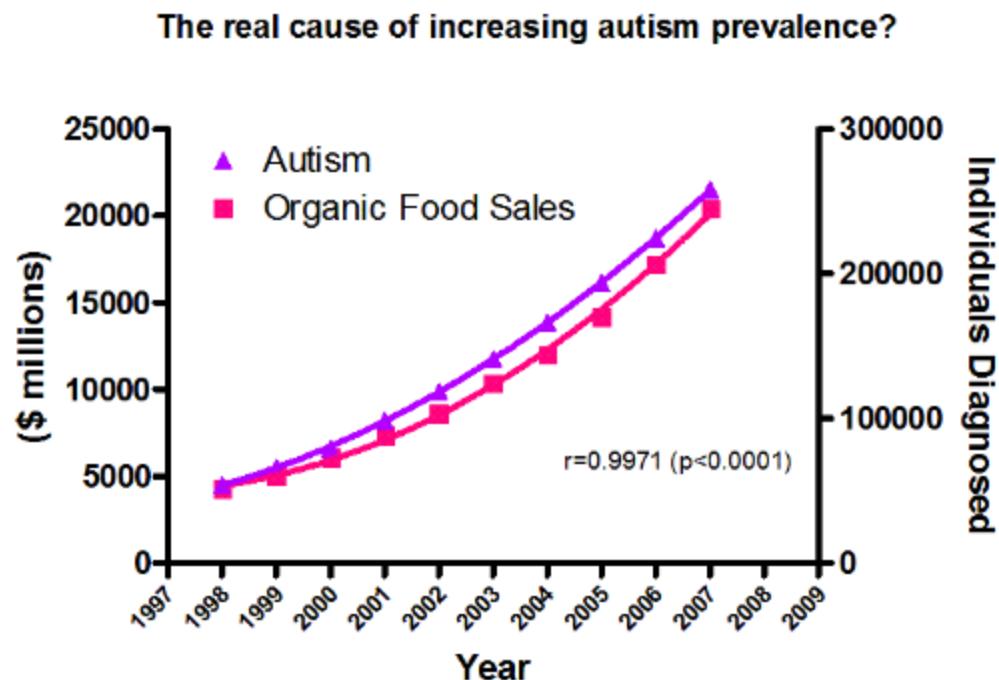
As a rule of thumb, pie charts max out at ~ 5-7 sectors

# 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!



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

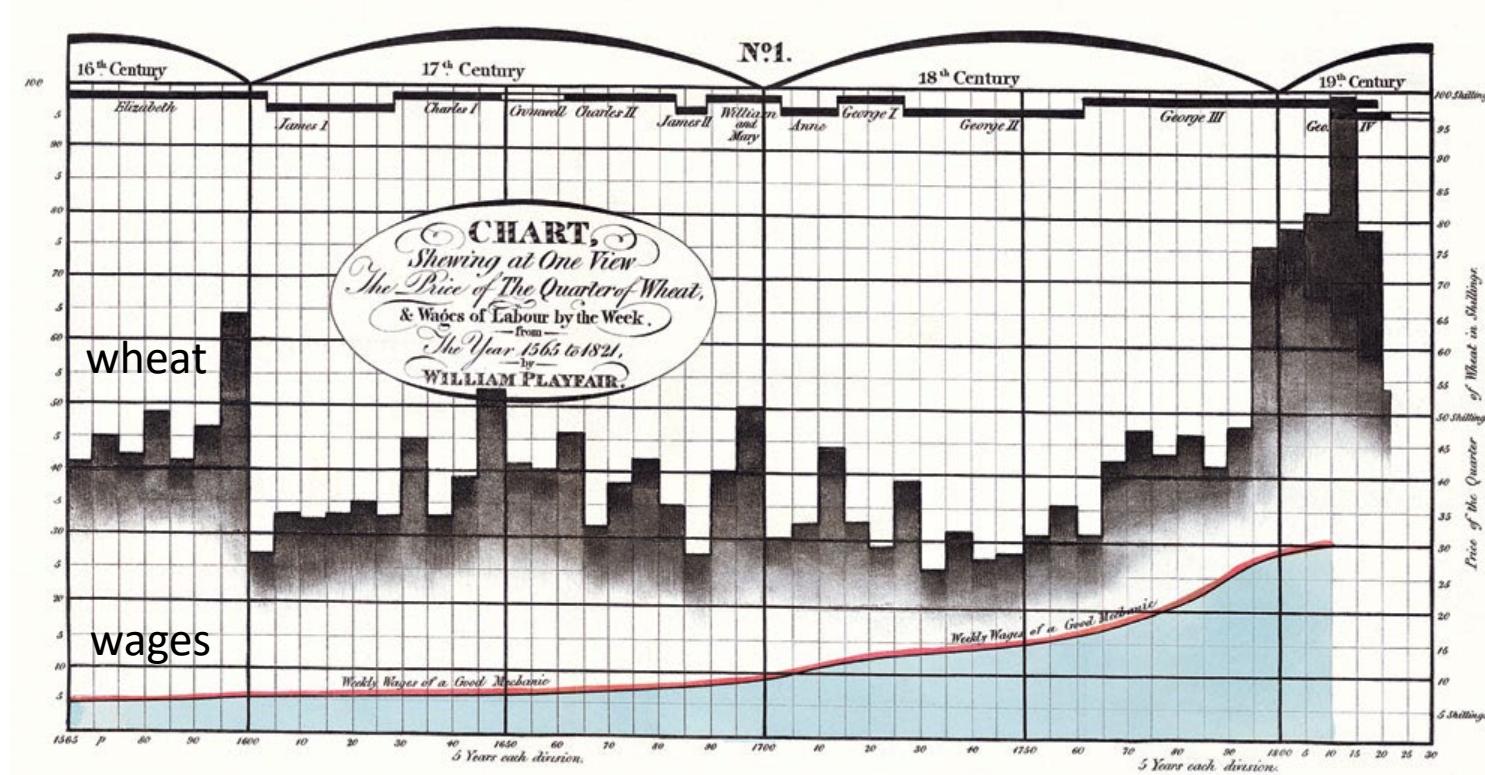
William Playfair invented the pie chart, line chart and bar chart.

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

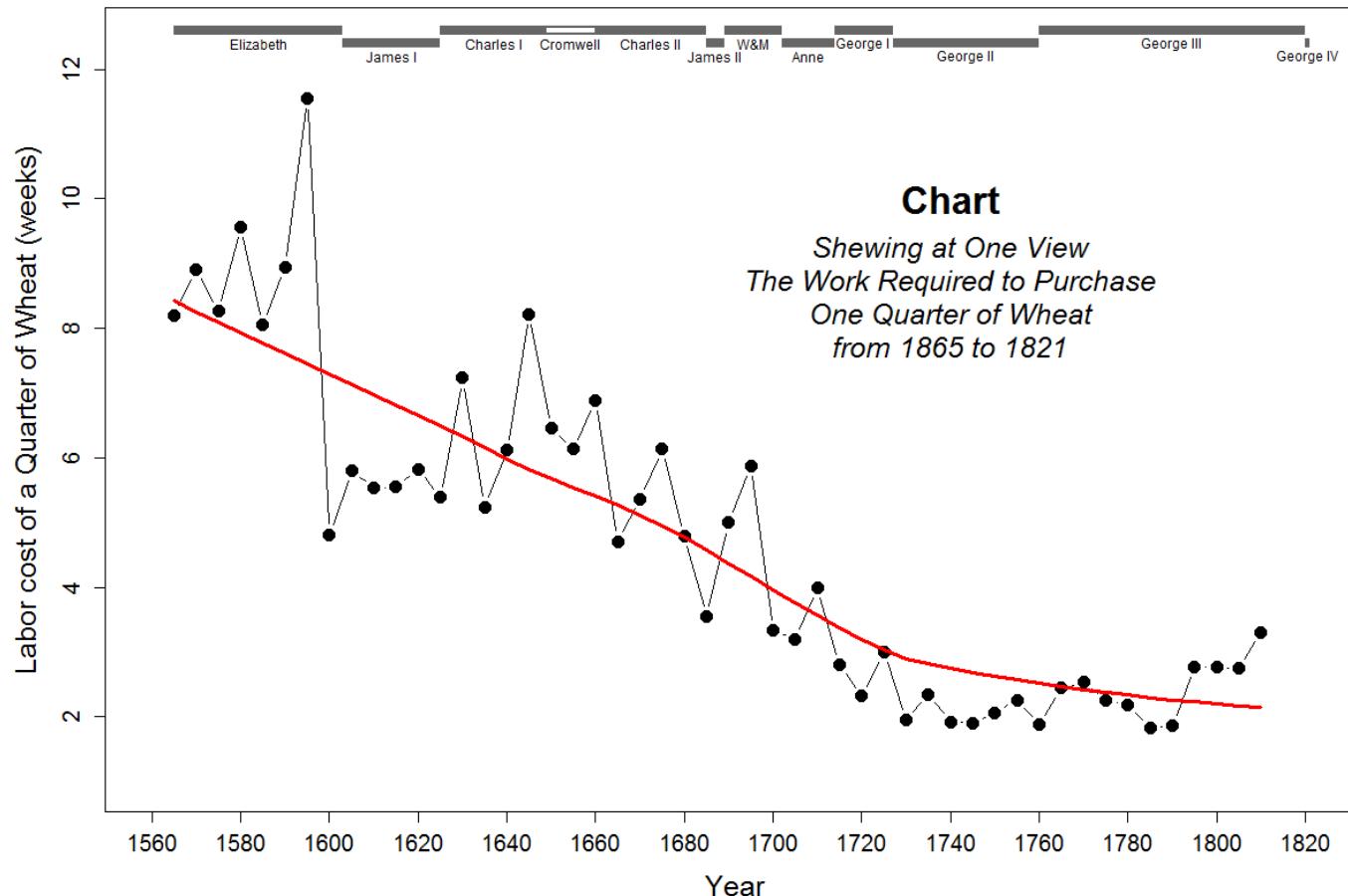
**Goal:** show that workers were better off most recently (1810) than in the past.

Is this what you see?



# Or, another graph would have been better

A modern **re-vision** plots the **ratio** of price of wheat to wages directly  
Makes Playfair's point more directly, but less beautifully



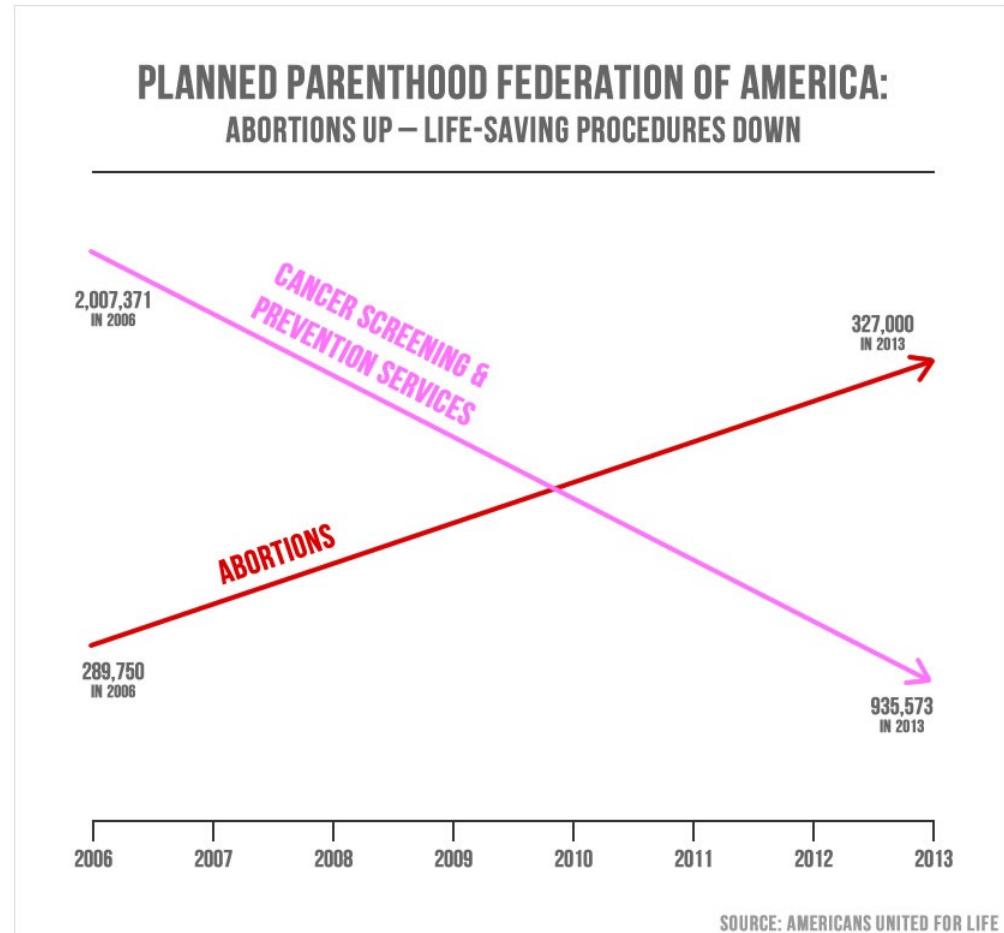
# Even more evil: No scales, no data

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?



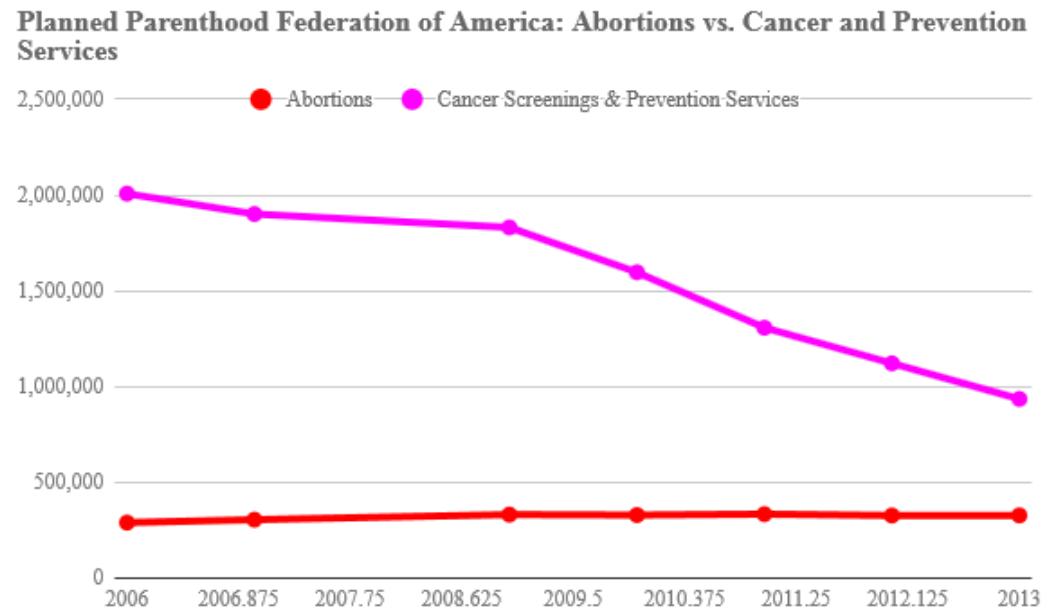
See: <http://www.politifact.com/truth-o-meter/statements/2015/oct/01/jason-chaffetz/chart-shown-planned-parenthood-hearing-misleading-/>

# 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.



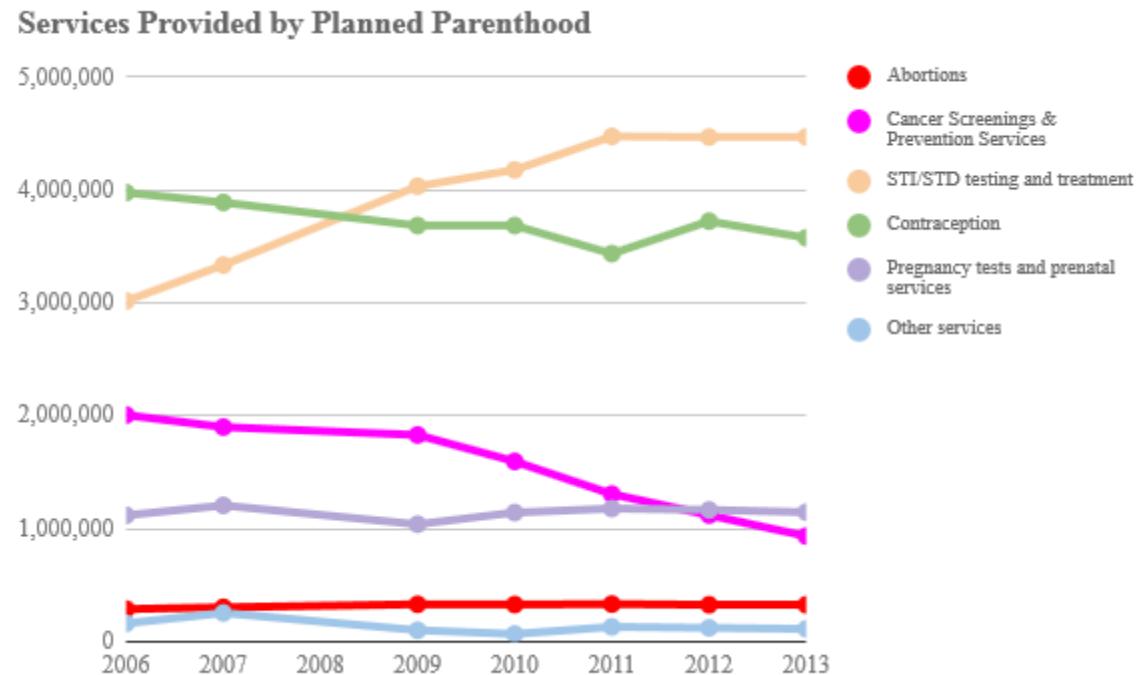
What improvements can be made to this graph?

- Fix formatting of axis labels
- Choose better colors
- Use direct labels

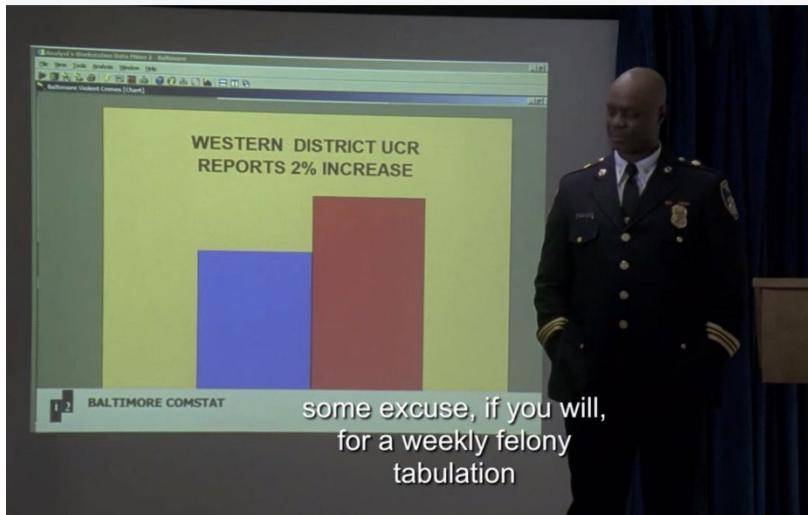
# Corrected graph, in context

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

PP activities were far higher for contraception and STD testing



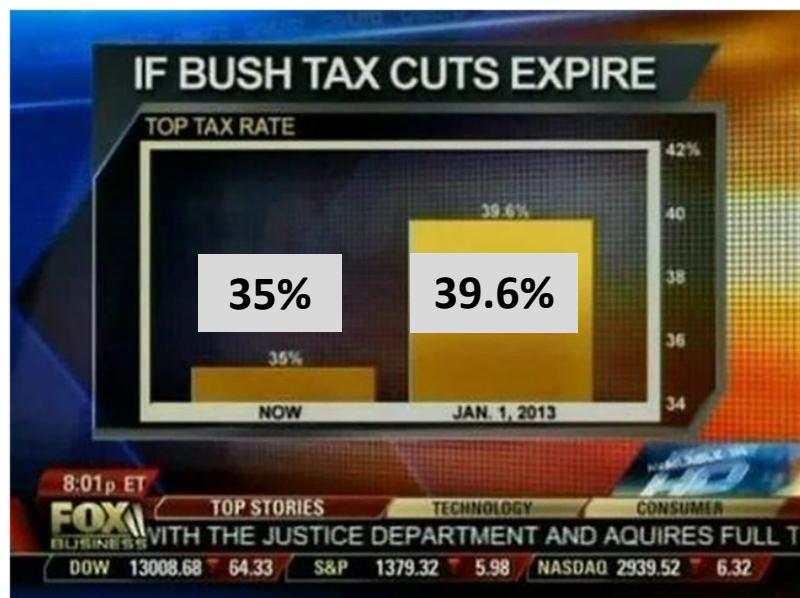
# Evil Bars



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

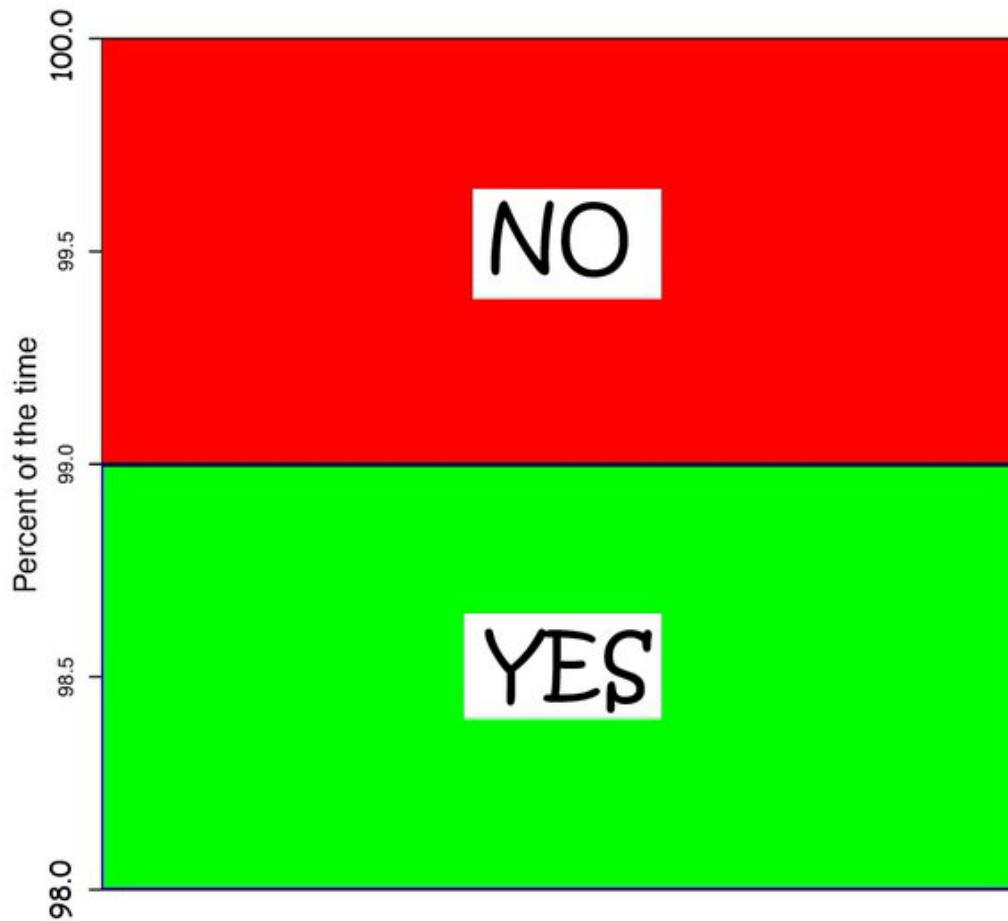


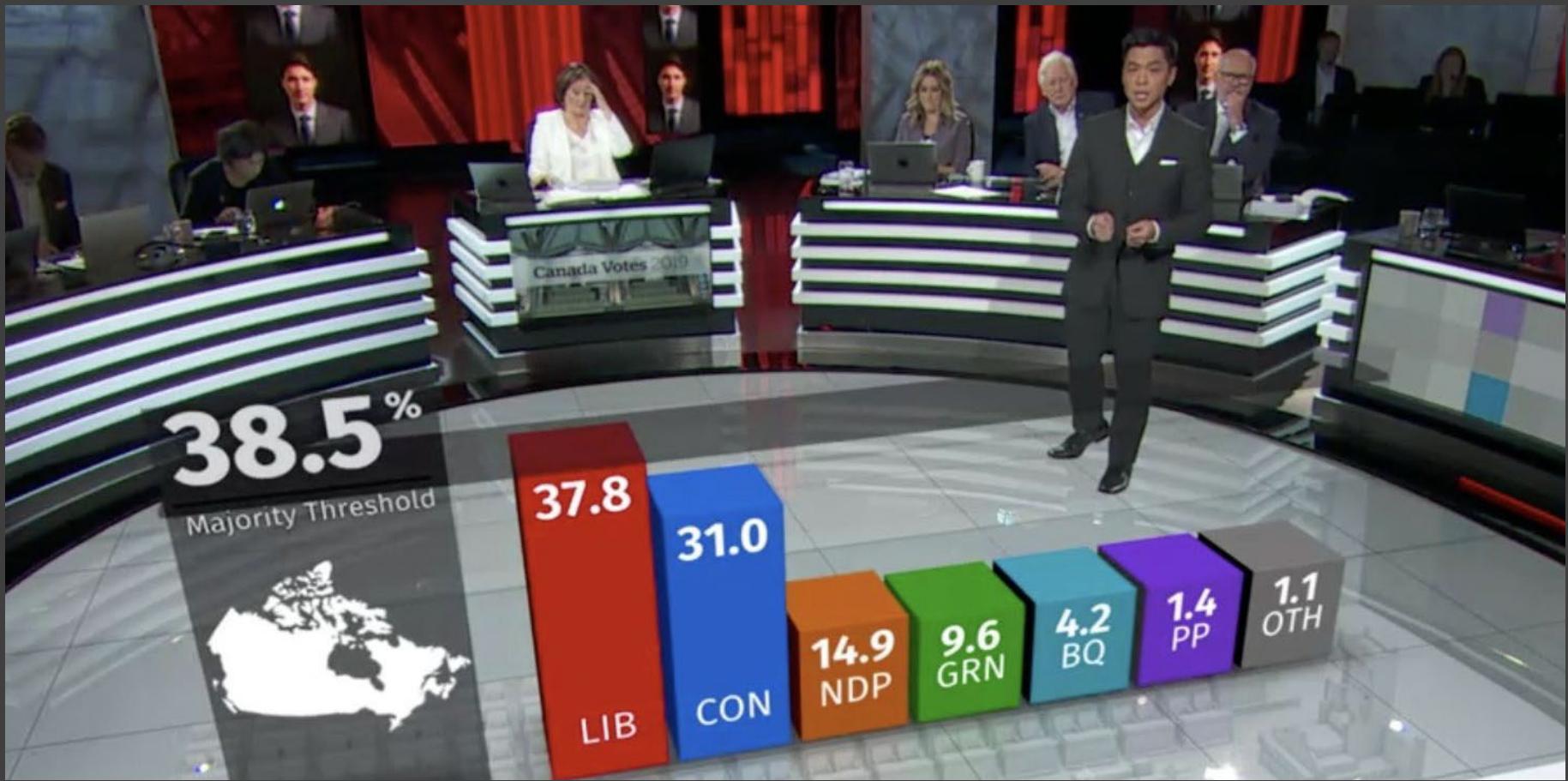
You can greatly distort the perception of difference or ratio by truncating the Y axis.

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

# Evil bars: A source of merriment?

Is truncating the y-axis misleading?





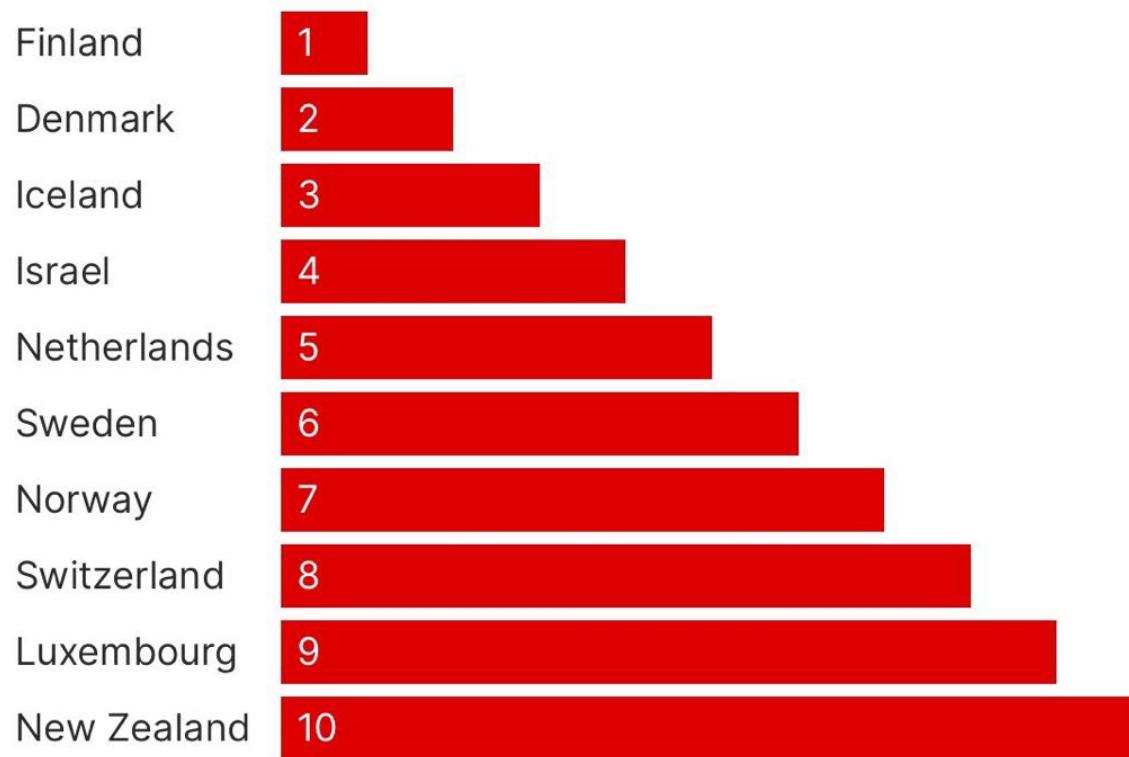
## More evil: 3D bars

- 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?

# How to make a confusing bar chart

## Top 10 Happiest Countries in 2023

This chart shows the top 10 happiest countries according to the 2023 World Happiness Report.



Why did they make this a bar chart?

What re-design would avoid the confusion?

# 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	Scale
Roads, streetscapes	30	35	54	83	15	0 - 10
Culture, community, recreation	26	50	59	47	16	11 - 20
Parks	27	41	41	52	20	21 - 30
Affordable housing	17	26	38	56	11	31 - 40
Public art	26	25	41	32	4	41 - 50
Heritage	16	13	26	18	3	51 - 60
Transit	11	7	10	20	3	61 - 70
Libraries	6	2	5	11	1	71 - 80
Other	3	6	7	8	3	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?

# 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?

- Grid lines obscure the data
- What do the grey bars mean?

Figure 18: Measuring Perceptions of Uncertainty

STATEMENT

Almost Certainly

Highly Likely

Very Good Chance

Probable

Likely

Probably

We Believe

Better Than Even

About Even

We Doubt

Improbable

Unlikely

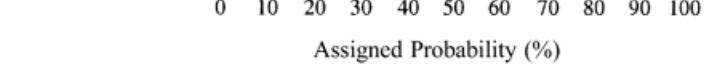
Probably Not

Little Chance

Almost No Chance

Highly Unlikely

Chances Are Slight



# Graphical excellence

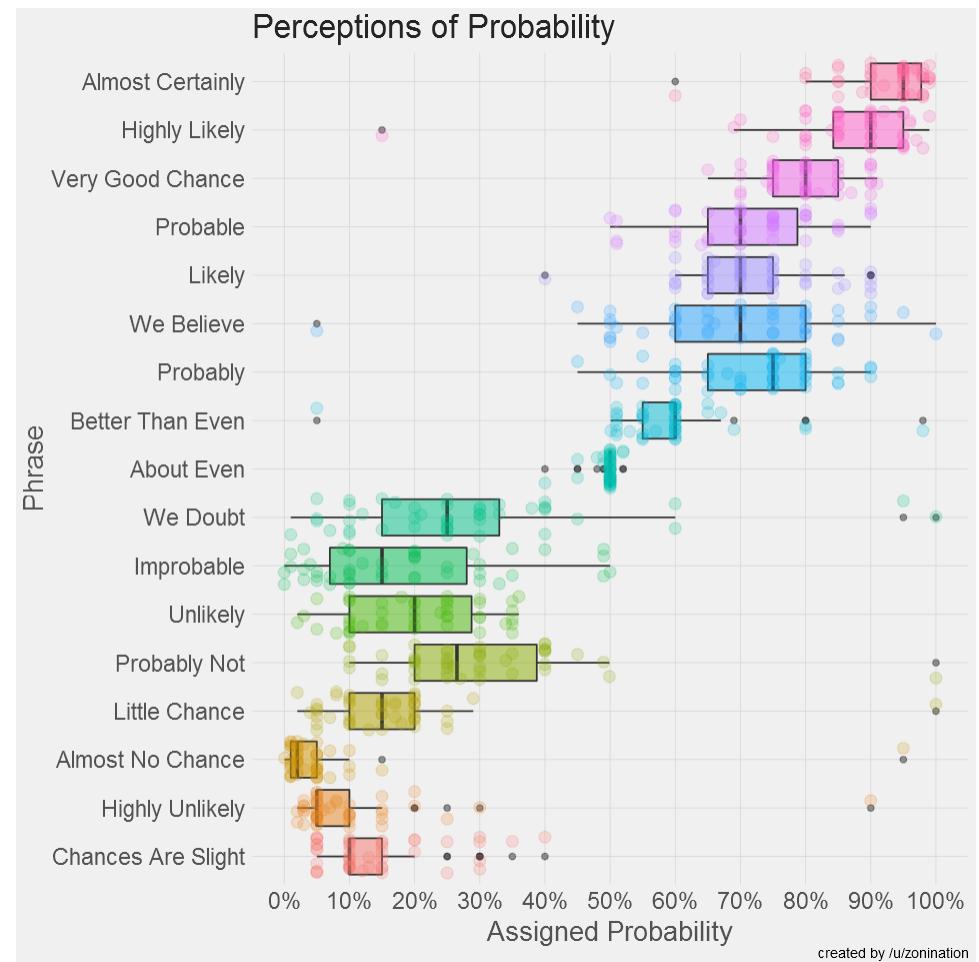
This graph shows the same data, as both dotplots & boxplots

We can see a lot more:

- “about even” has very low variability
- the bottom 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>

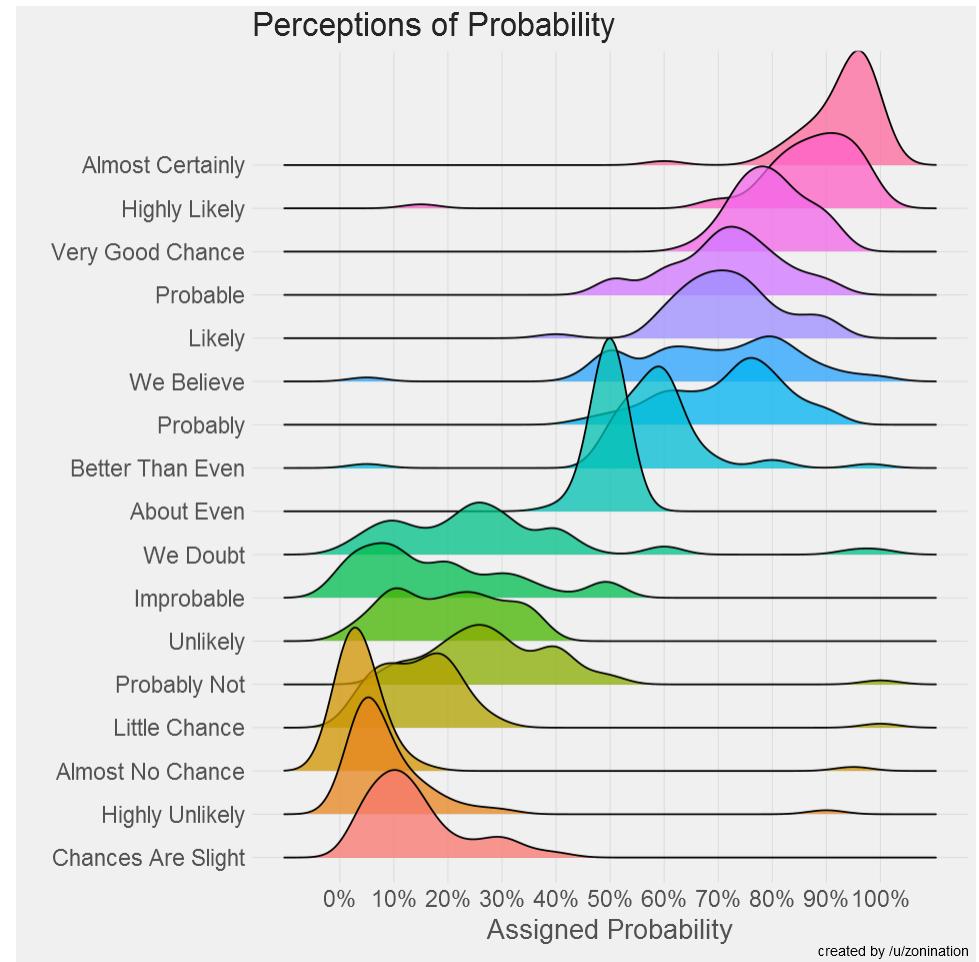
# Graphical excellence

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.

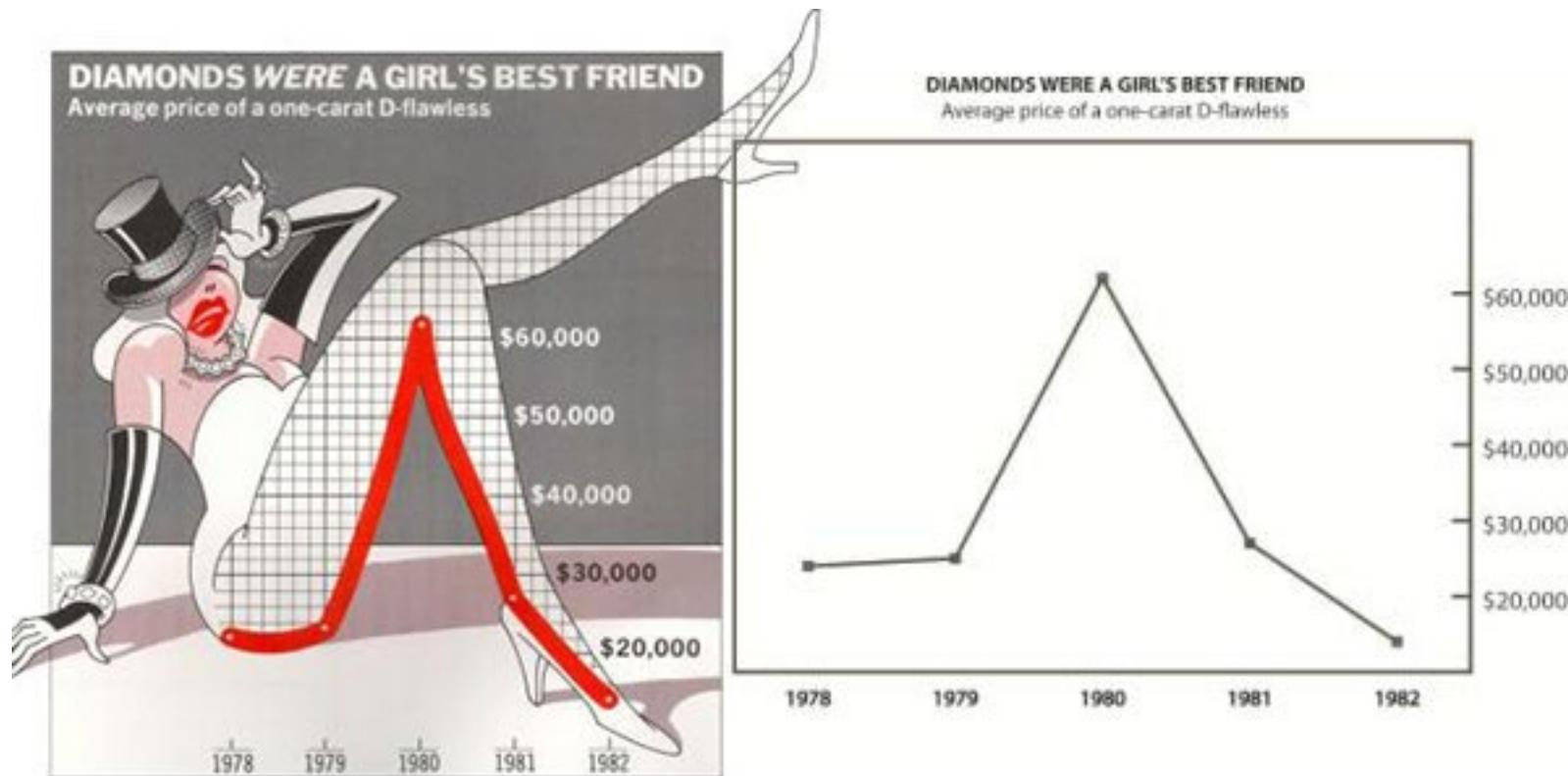
Color & transparency are used effectively



# Chart junk or effective info vis?

Charts can be offensive  
and/or effective

What is the message?  
Who is the audience?



NB: Info designers now consider the term “chart junk” itself offensive

# 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?

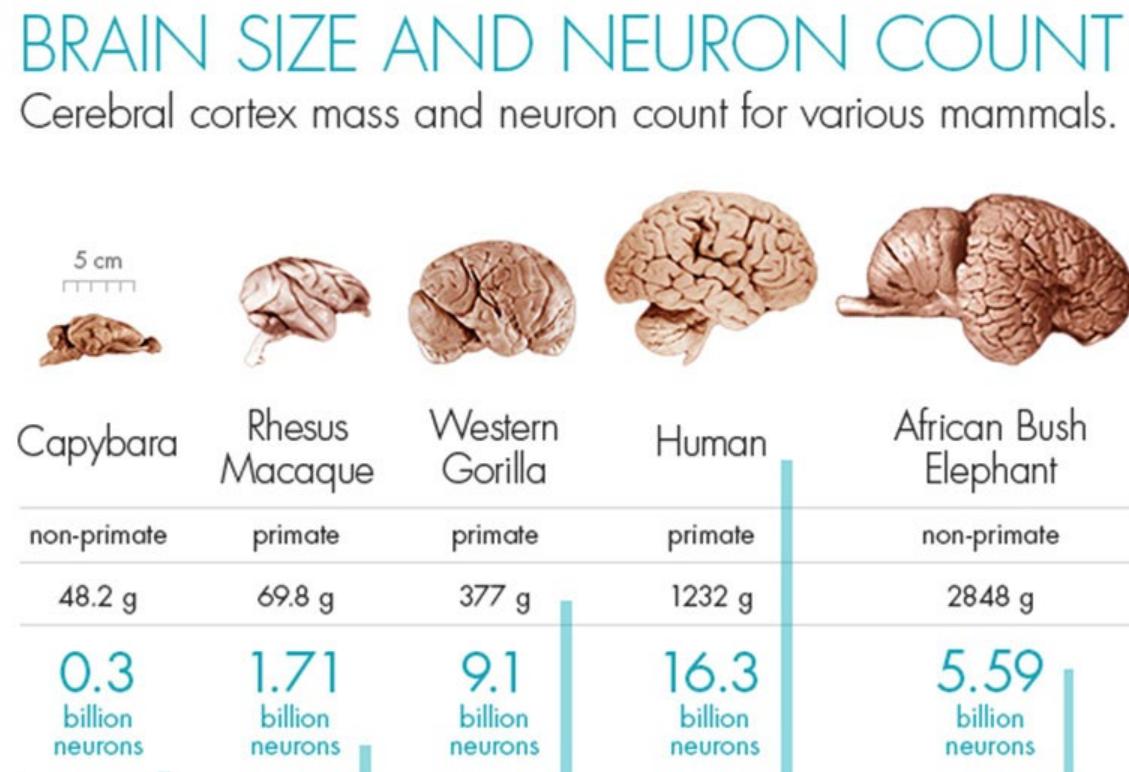
Goal: compare mammal species brain size and cortical neuron count

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>

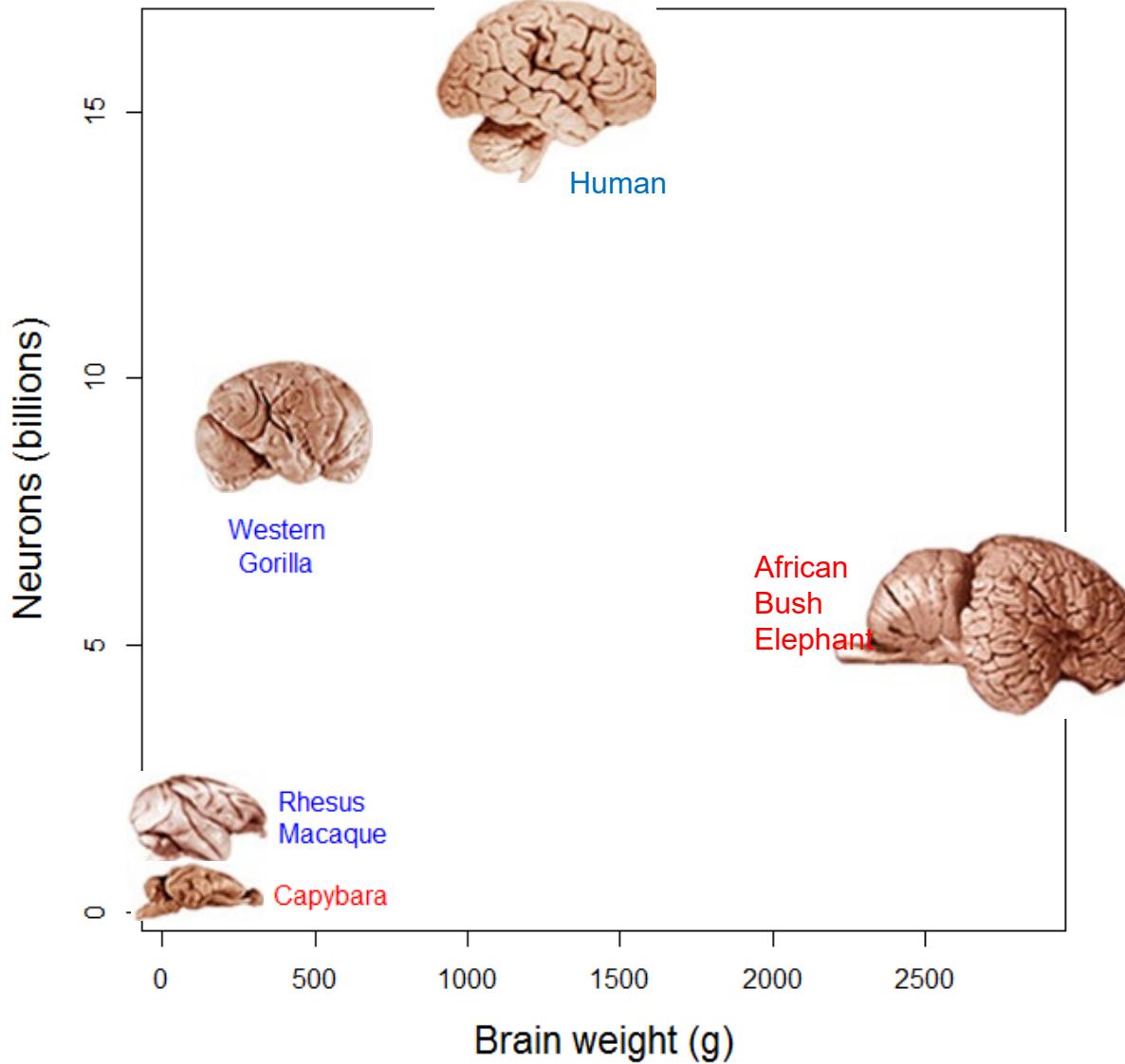
# As a scatterplot?

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?

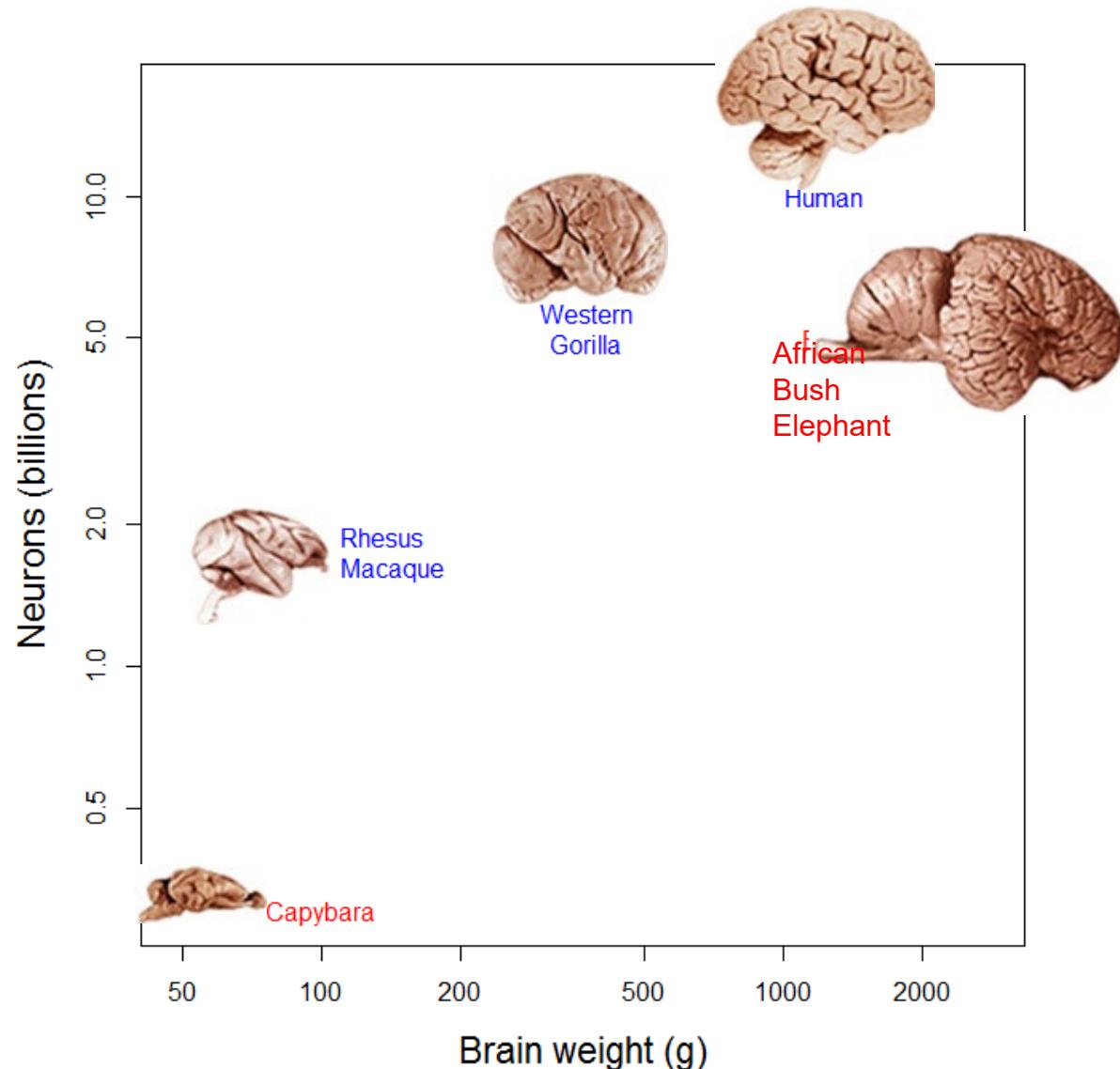


# As a scatterplot – log scale

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<sub>2</sub> emissions as the main agent.

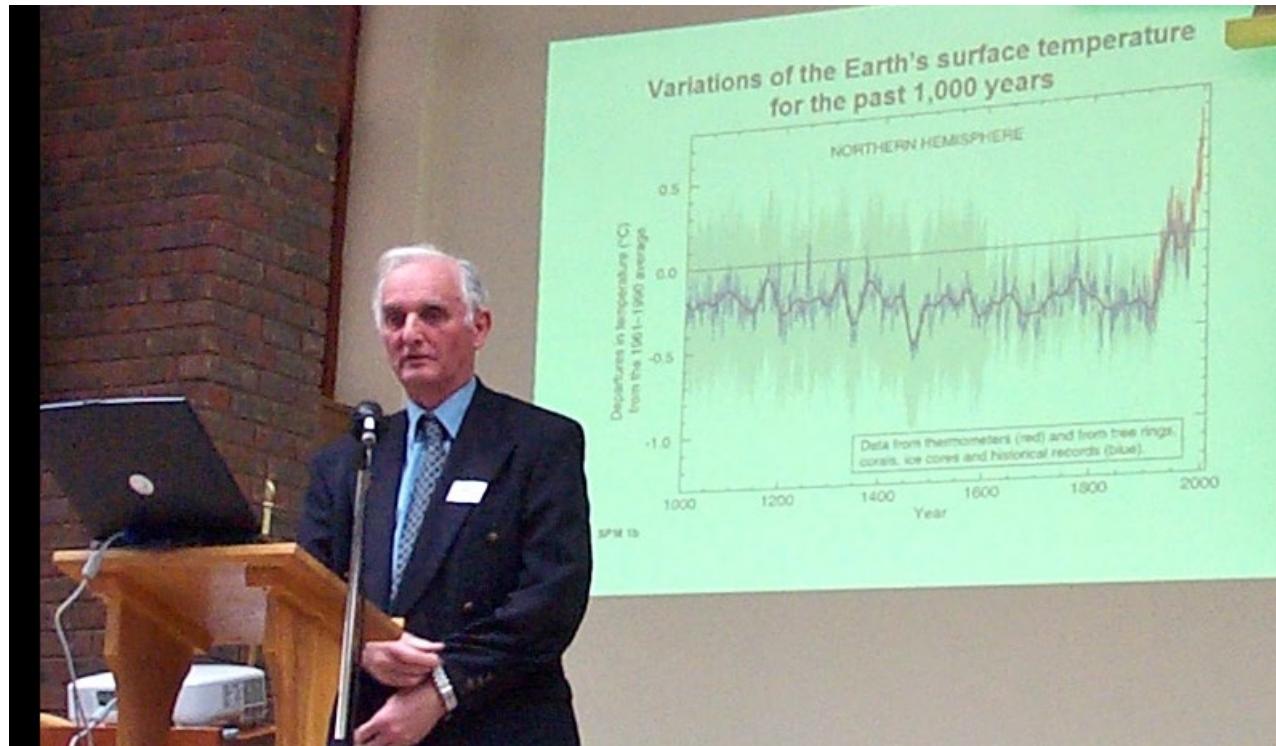


Movie: <https://www.youtube.com/watch?v=8ZUoYGAI5i0>; <http://www.imdb.com/title/tt0497116/>

# 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.

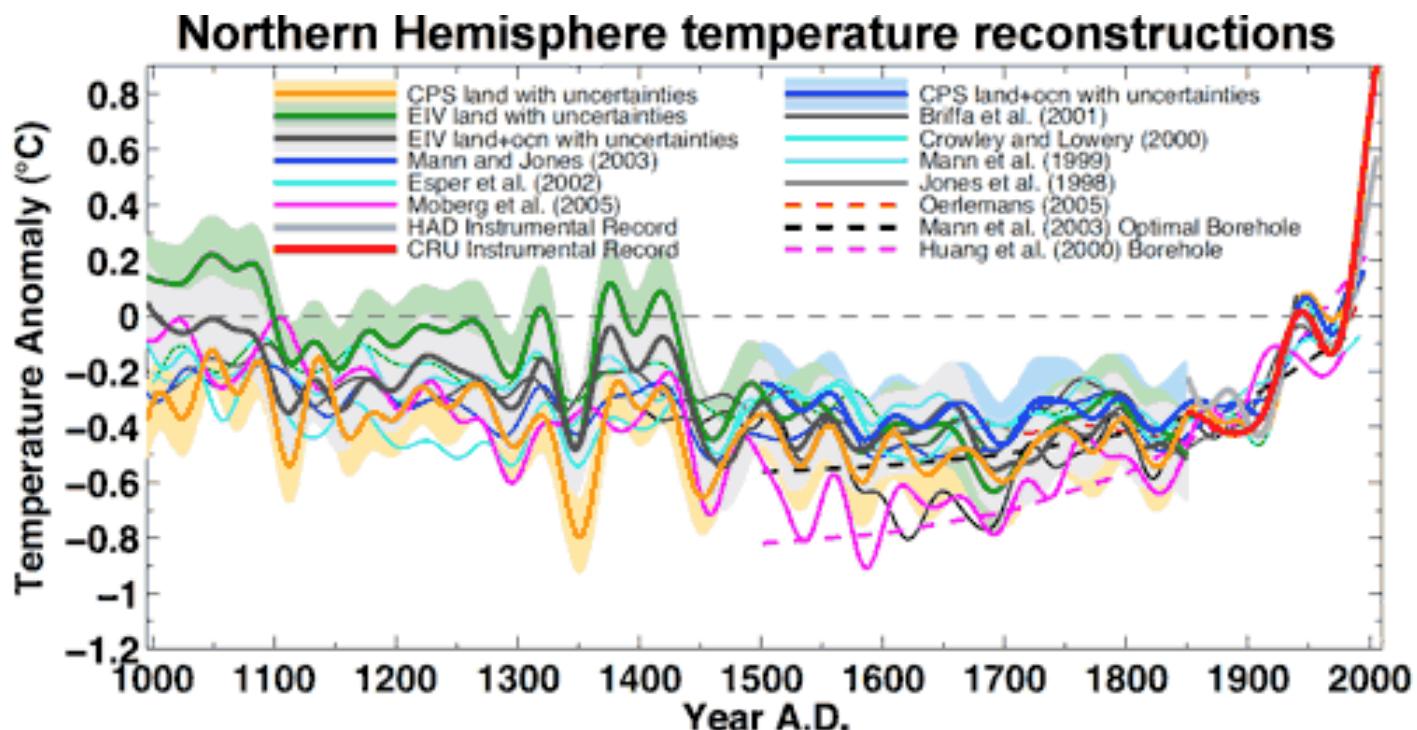


# Climate change: data sources

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.

**TOO MUCH INFO??**



See: [https://en.wikipedia.org/wiki/Hockey\\_stick\\_controversy](https://en.wikipedia.org/wiki/Hockey_stick_controversy) for details

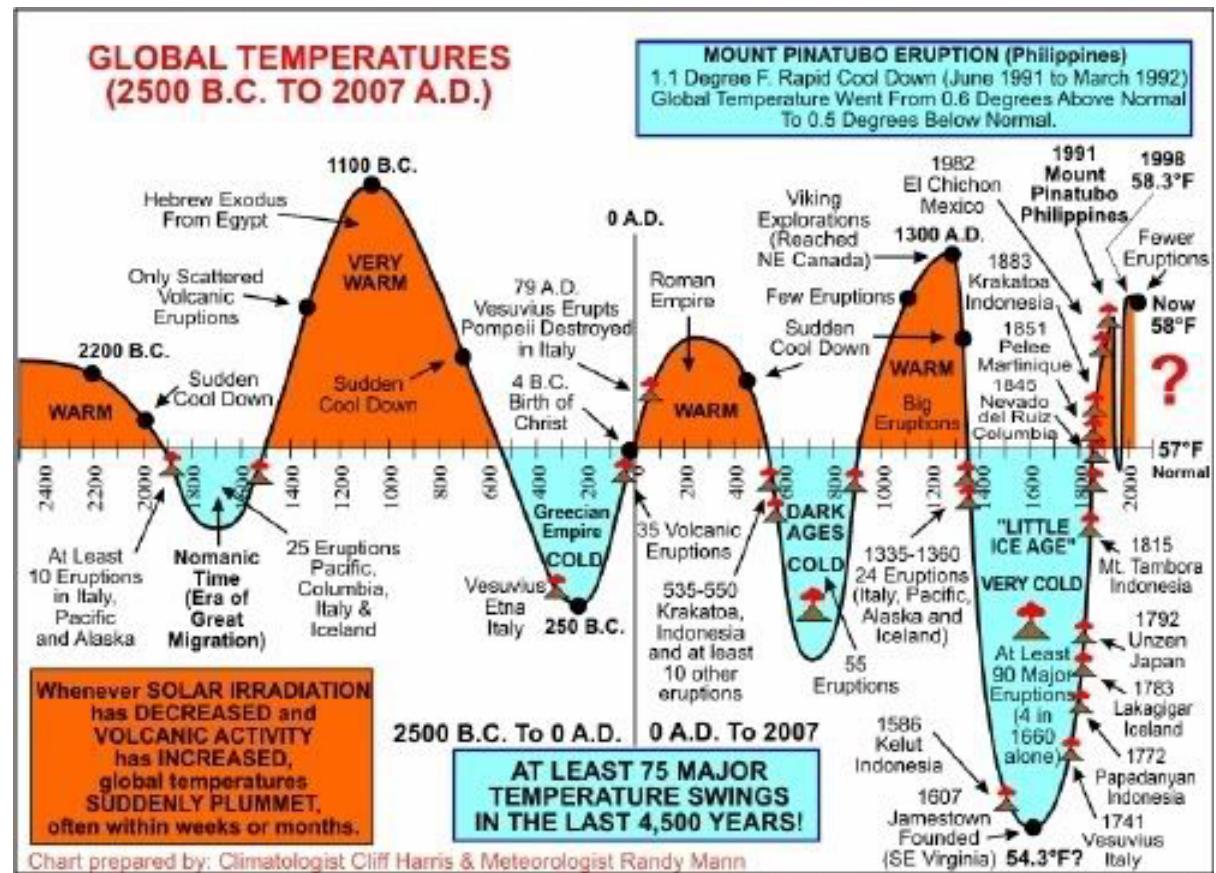
# Counteracting climate change

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

## Solar radiation & 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

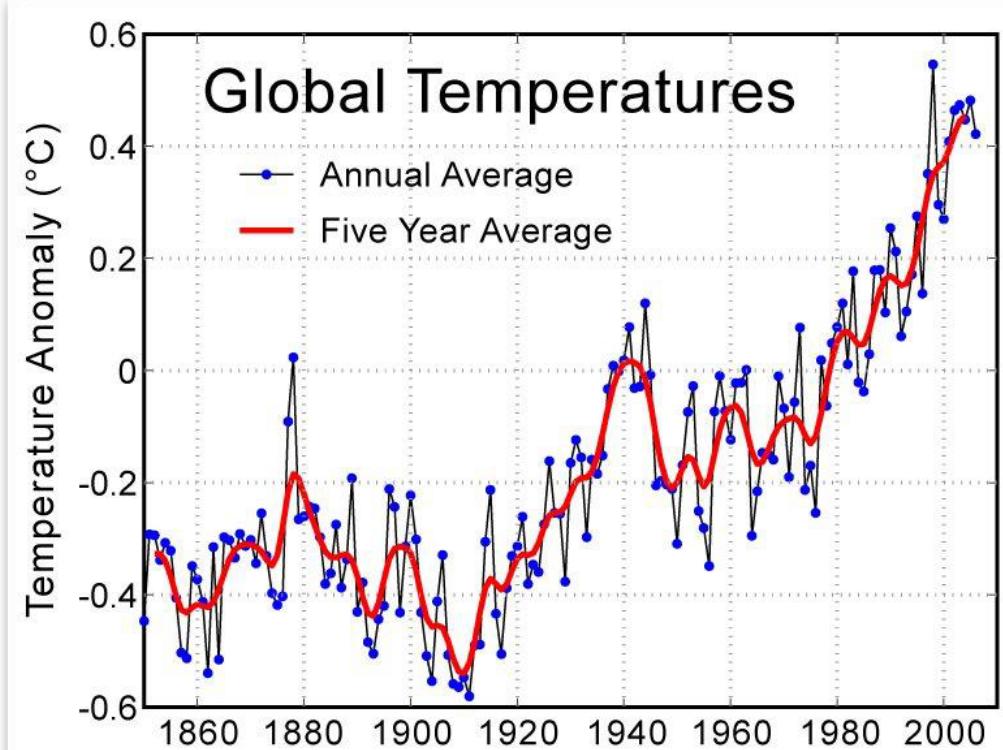


# Time scale

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?

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



# Climate change: Infographic

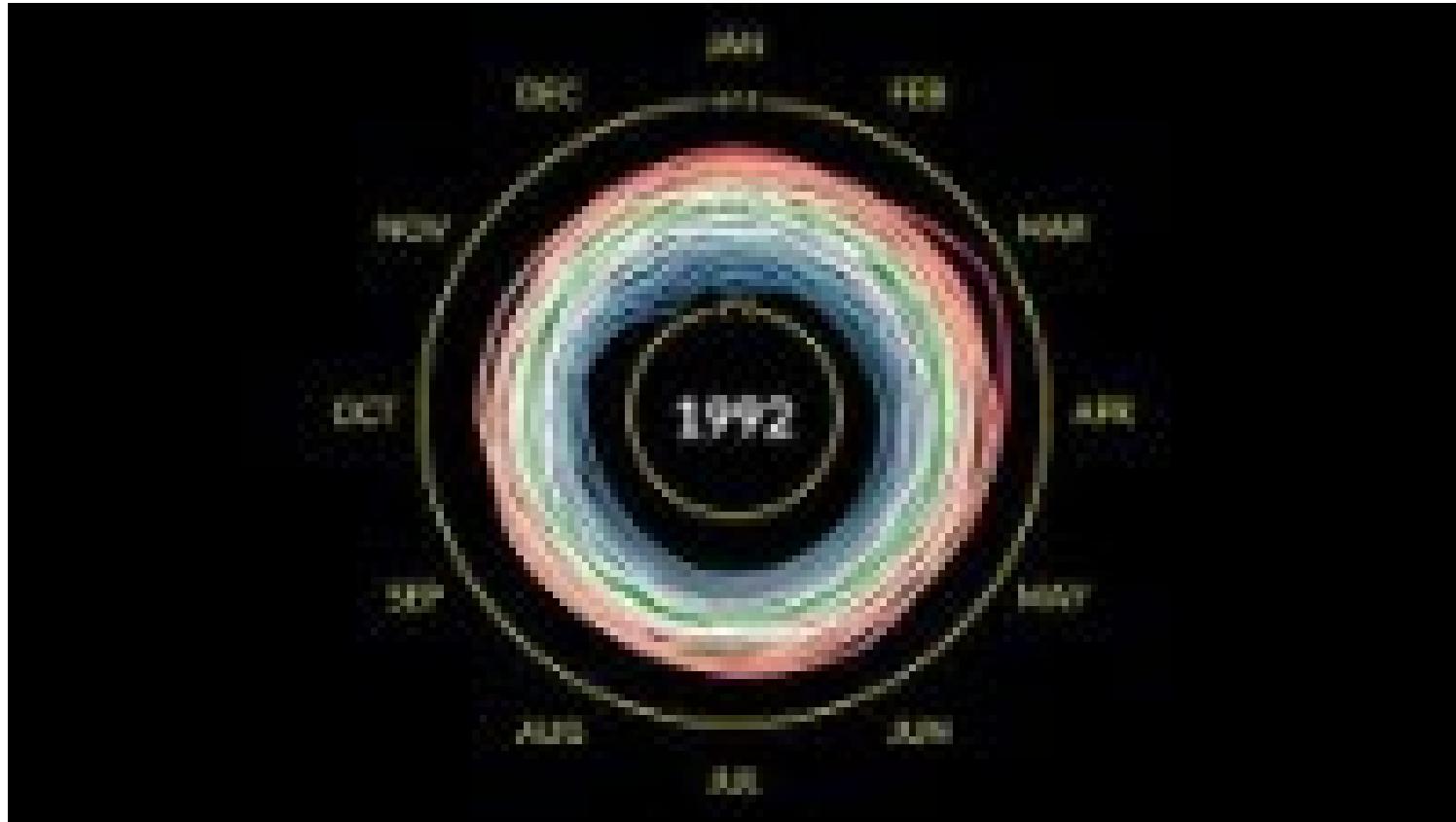
A politically-incorrect graphic shows very clearly the effect of global warming on panty size



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

# Climate change: animation

This NASA video by Ed Hawkins shows monthly global temperature anomalies, 1890-2021 as an increasing spiral whose color reflects temperature.

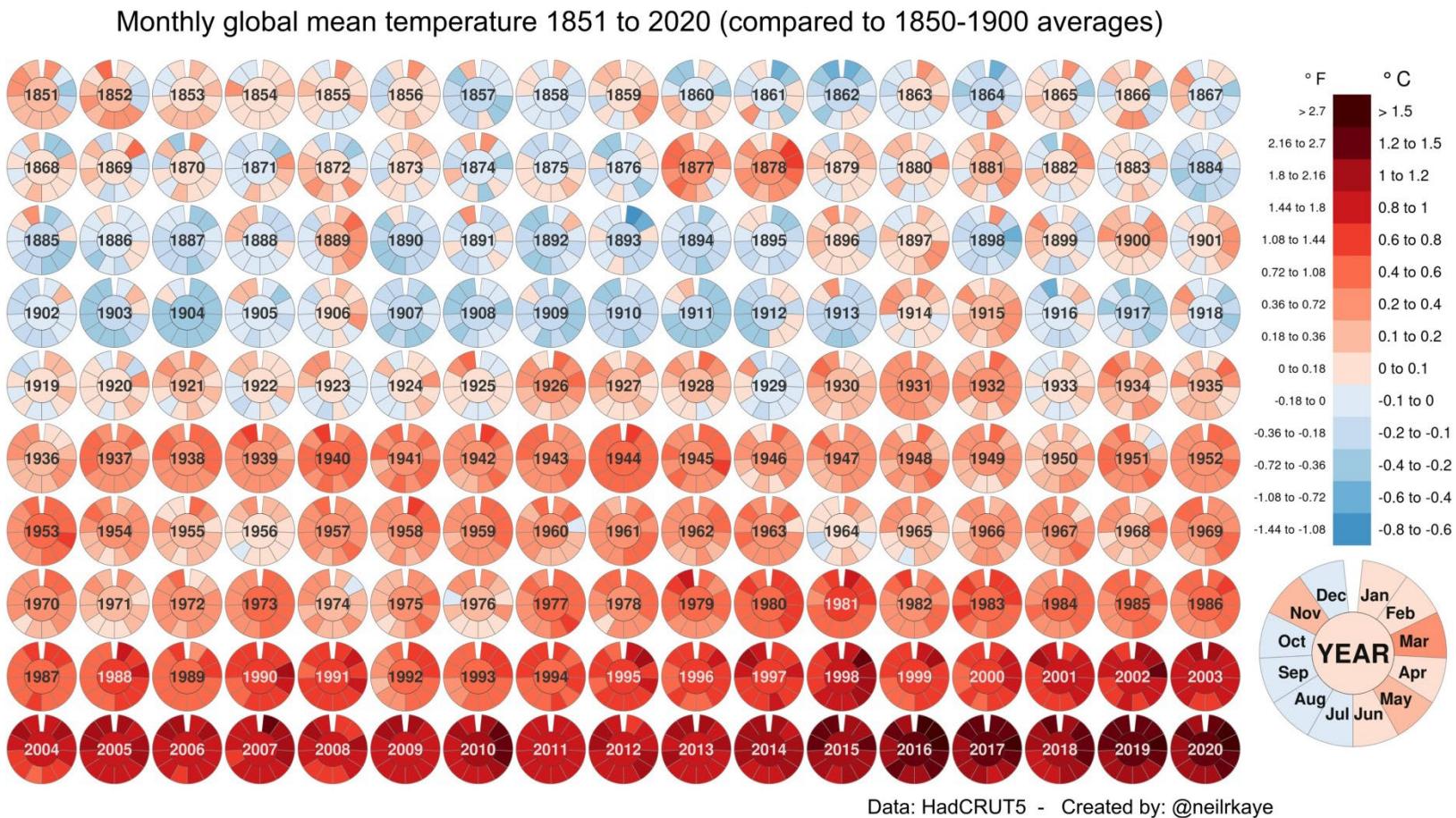


Video: <https://www.youtube.com/watch?v=jWoCXLuTIkI>

See also: <https://www.climate-lab-book.ac.uk/spirals/>

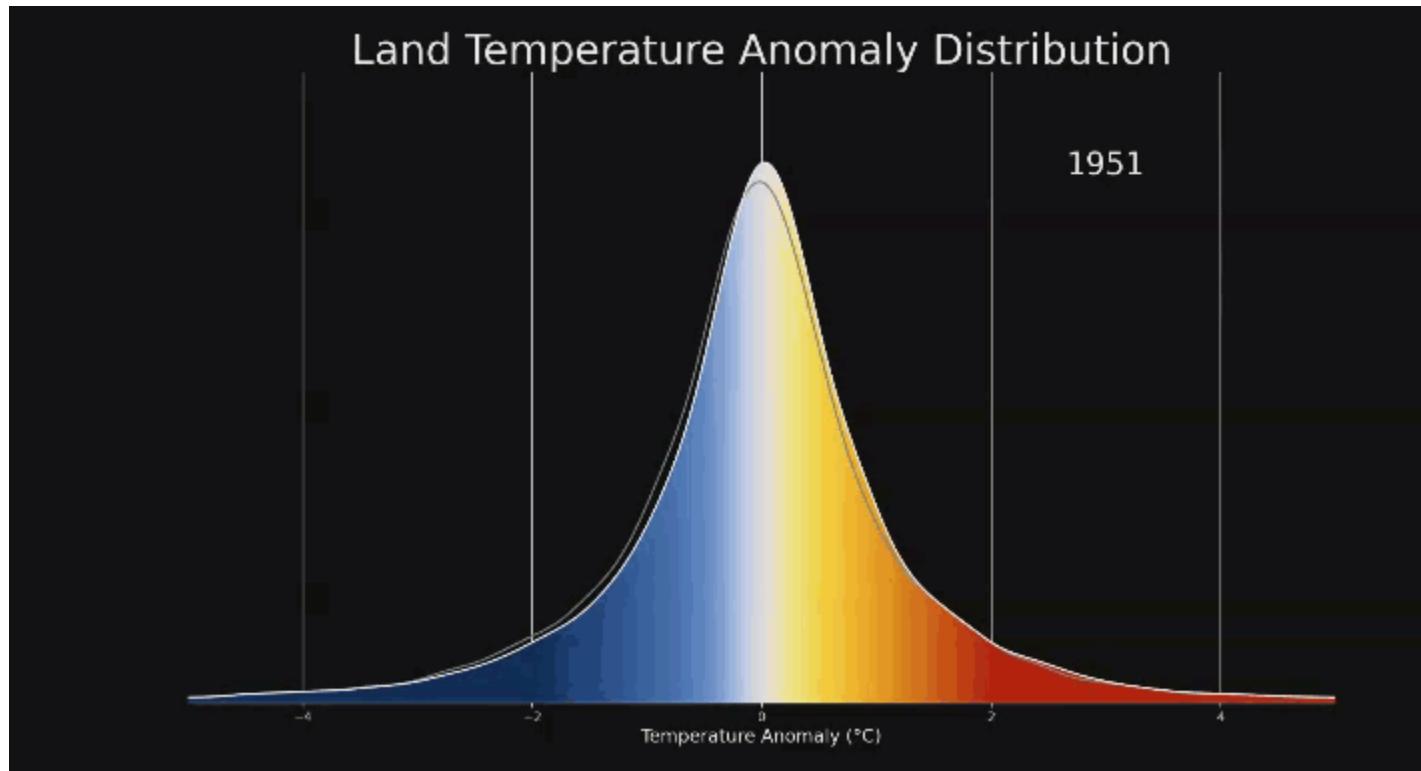
# Circle graphs

What features makes this graph effective?



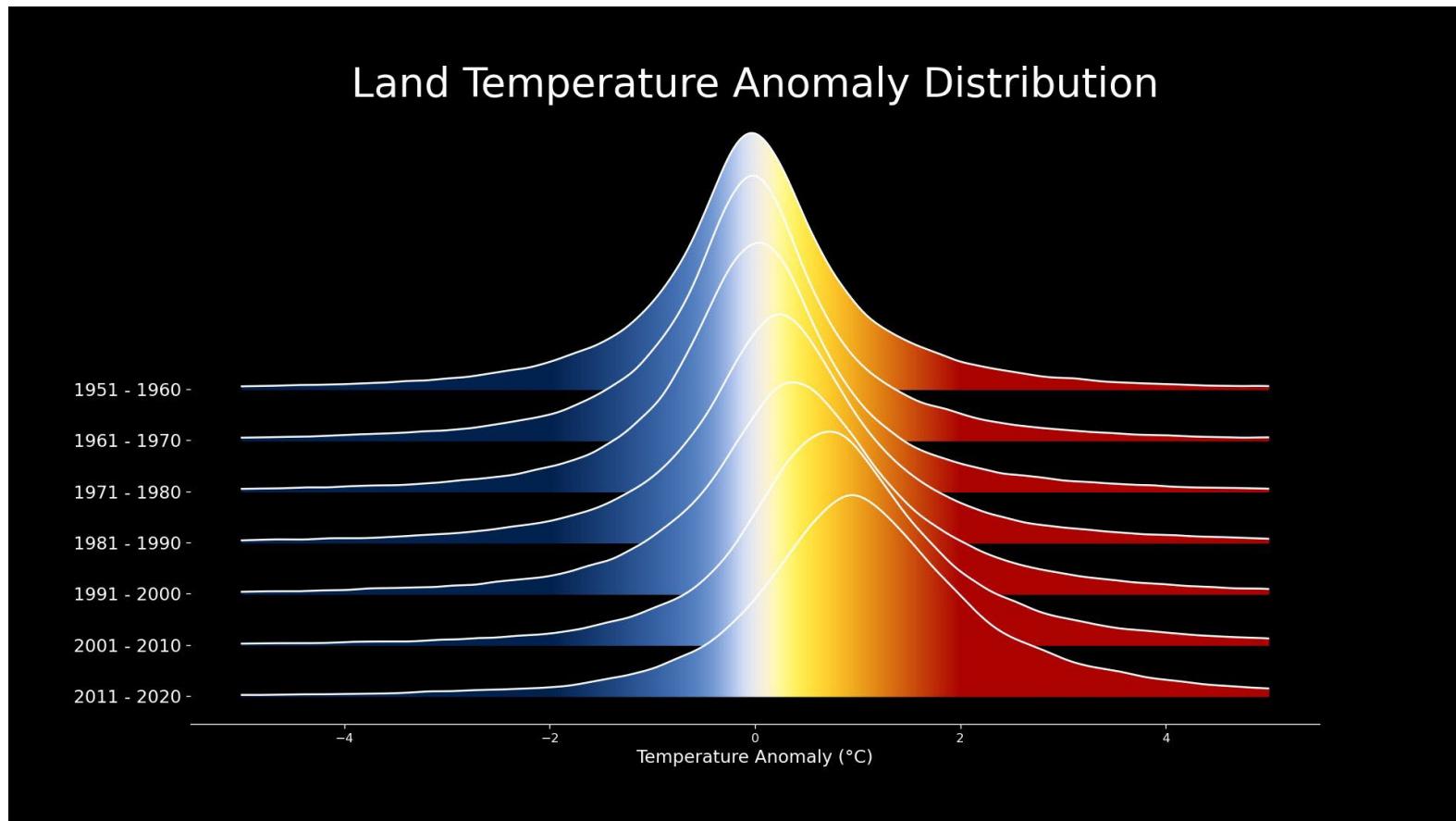
# Animation

Shifting Distribution of Land Temperature Anomalies, 1951-2020



# Animation -> Ridgeline plot

A ridgeline plot (stacking traces) turns the animation into a static graph



# 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 → Stories
- Goal: tell a **credible** story

Analysis graphs vs. presentation graphs

- Know your audience
- Target info to your message
- Avoid doing evil

Some principles of effective data display

- Make the data stand out
- Facilitate comparisons
- Effect ordering
- Direct labeling

