Data Visualization

FOUNDATIONS

Outline

What is data visualization?

Why visualize data?

The three principles of good visualization design

- Trustworthiness
- Accessibility
- Elegance

What is data visualization?

Definition

The presentation of data in graphical form to facilitate understanding

https://en.wikipedia.org/wiki/Data_visualization

Distinctions in terminology

Data visualization ≈ information visualization

- Data + meaning = information
- When a distinction is made (we will not make it)
  - Data visualization is concerned with numerical data
  - Information visualization is concerned with abstract data structures

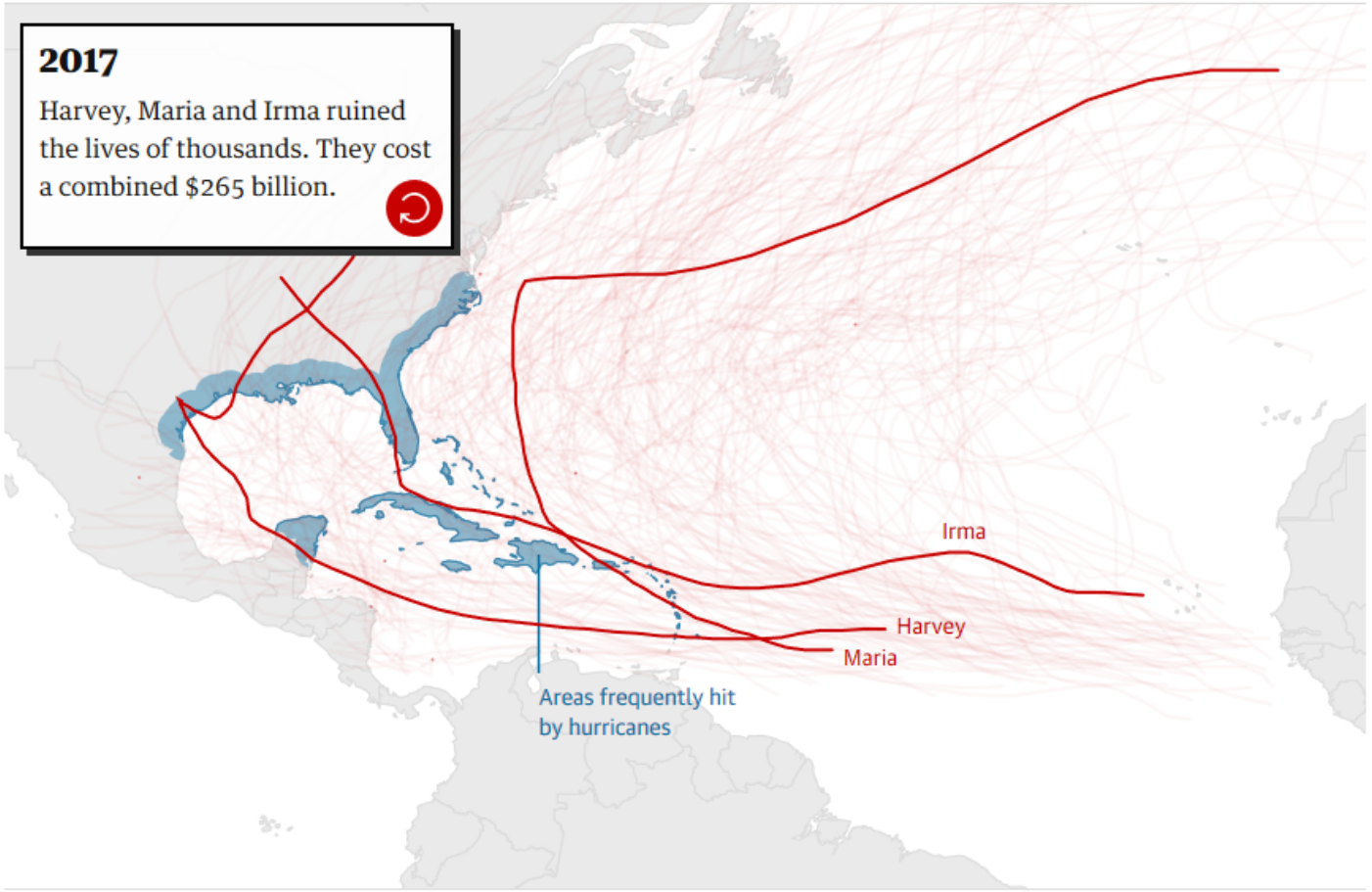
Scientific Visualization

- Visualization of 3-D phenomena for scientific purposes

Infographics

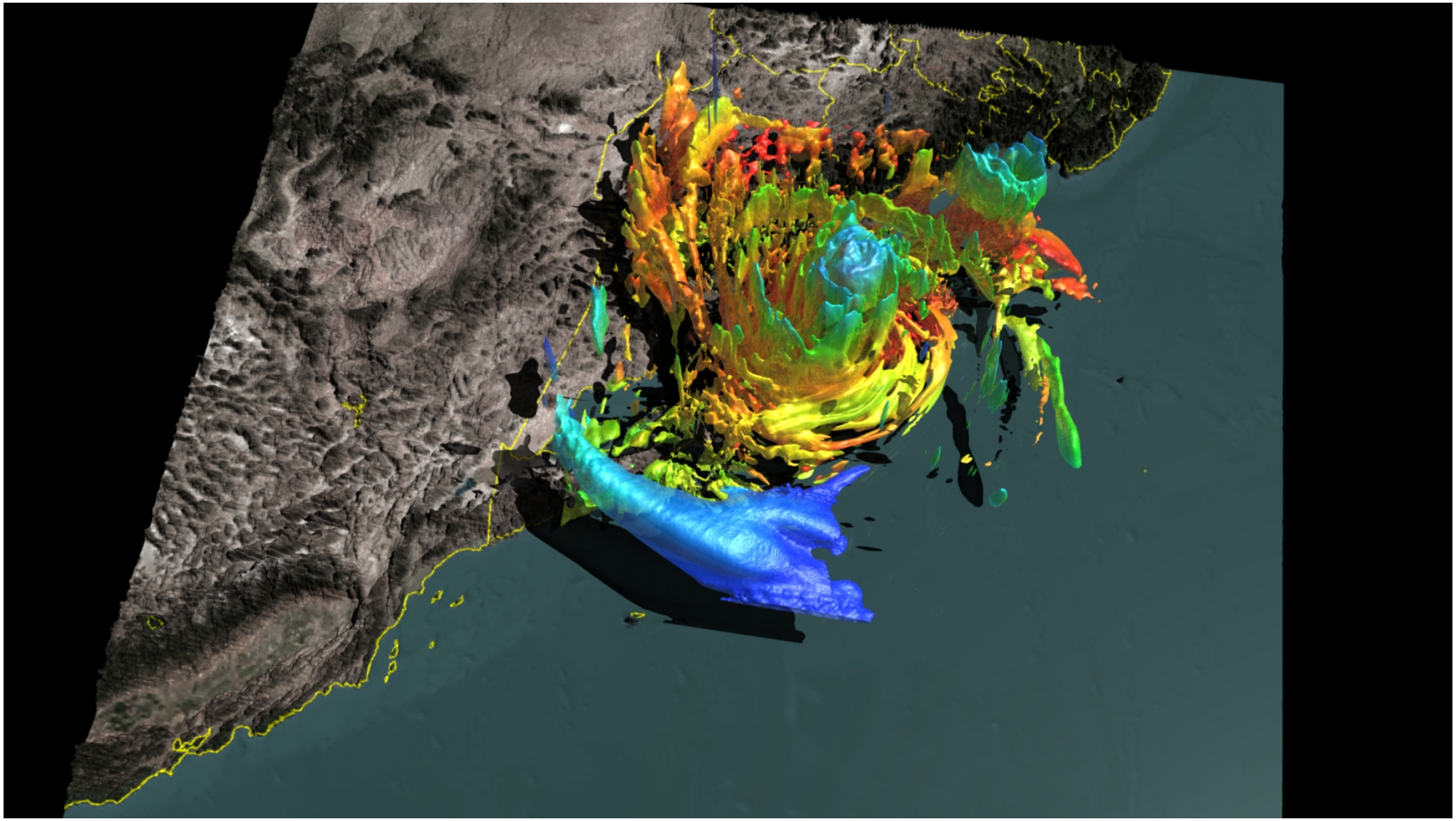
- Use different graphics for explanation (charts, illustrations, photo-imagery)
- Traditionally created for print consumption (static)
- Sometimes hard to discern from data visualization

Data visualization example

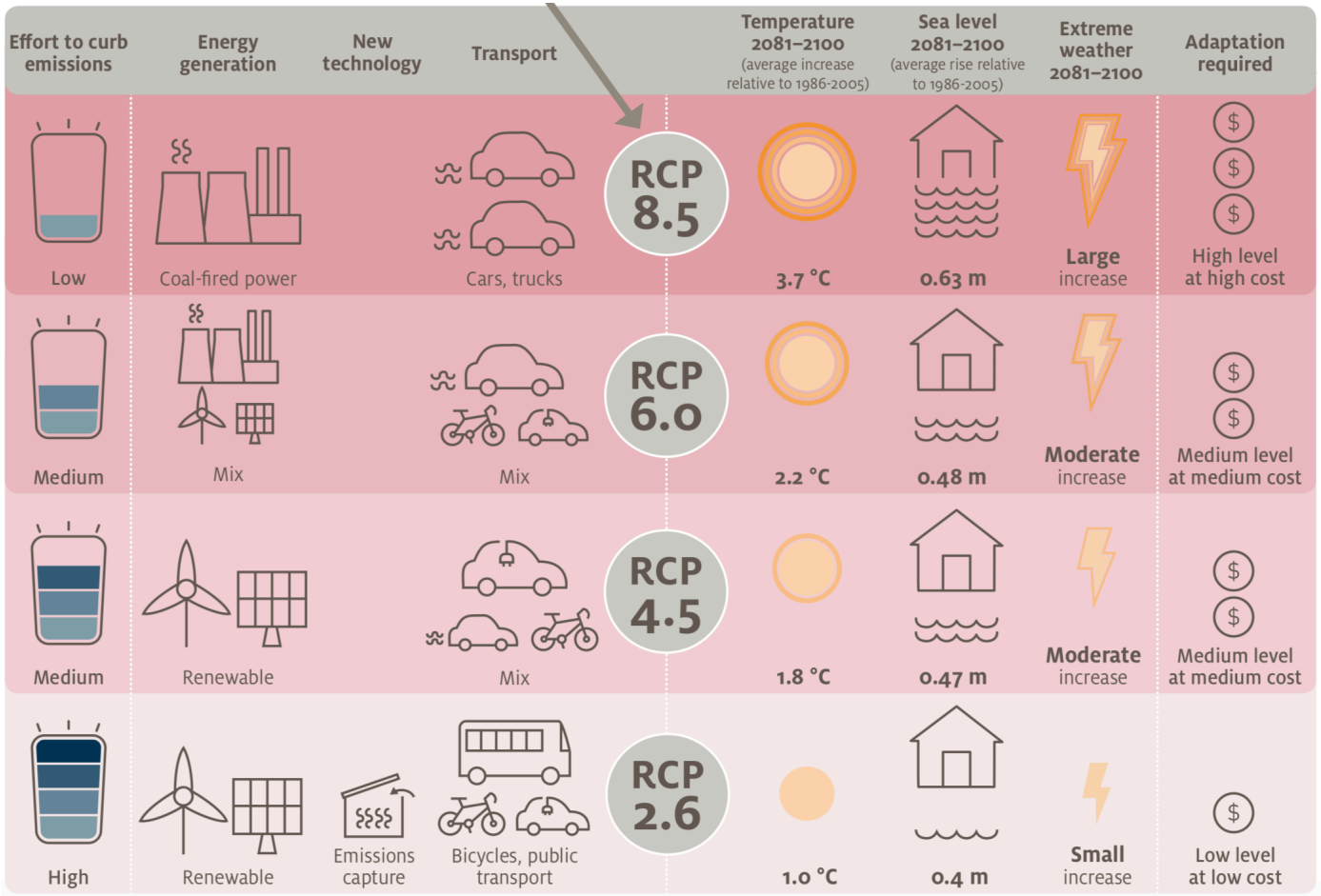


https://www.theguardian.com/weather/ng-interactive/2018/sep/11/atlantic-hurricanes-are-storms-getting-worse

Scientific visualization example



Infographic example



https://coastadapt.com.au/sites/default/files/infographics/15-117-NCCARFINFOGRAPHICS-01-UPLOADED-WEB%2827Feb%29.pdf

Distinctions in terminology

Interchangeable use

- Chart
- Graph
- Plot
- Diagram
- Map (sometimes!)

Why visualize data?

'A PICTURE IS WORTH A THOUSAND WORDS'

Anscombe's quartet

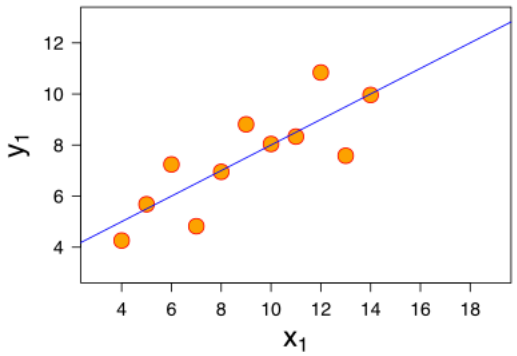
4 datasets with pairs of numbers (*x, y*) that have nearly identical simple descriptive statistics

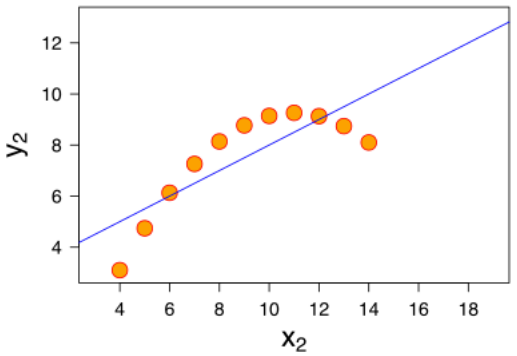
Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

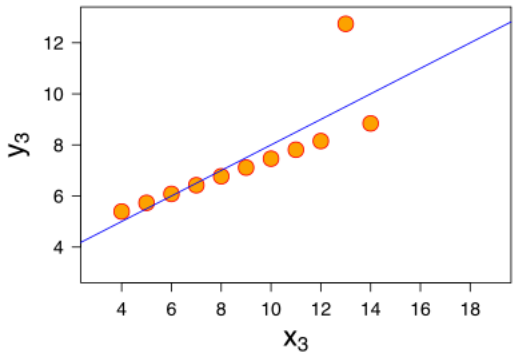
F. J. Anscombe. Graphs in Statistical Analysis. *American Statistician*, 27(1):17-21, 1973
https://en.wikipedia.org/wiki/Anscombe's_quartet

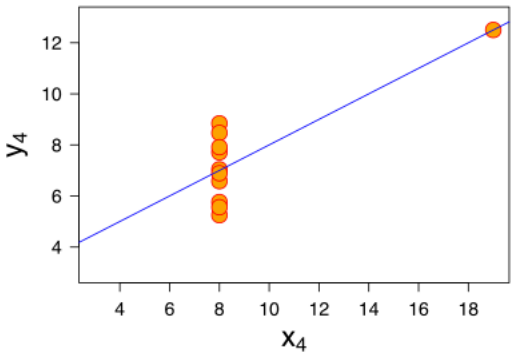
Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89







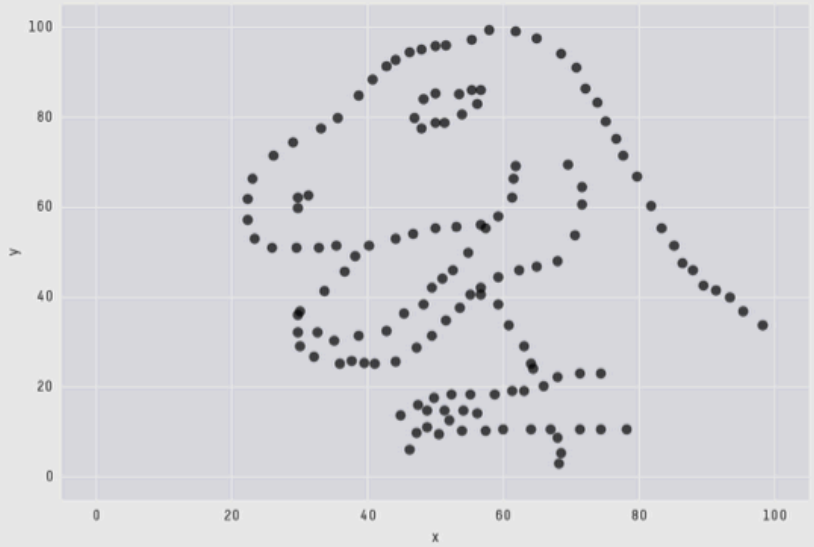


F. J. Anscombe. Graphs in Statistical Analysis. *American Statistician*, 27(1):17-21, 1973
https://en.wikipedia.org/wiki/Anscombe's_quartet

Datasaurus

DrawMyData tool for teaching stats and data science by Robert Grant: http://robertgrantstats.co.uk/drawmydata.html

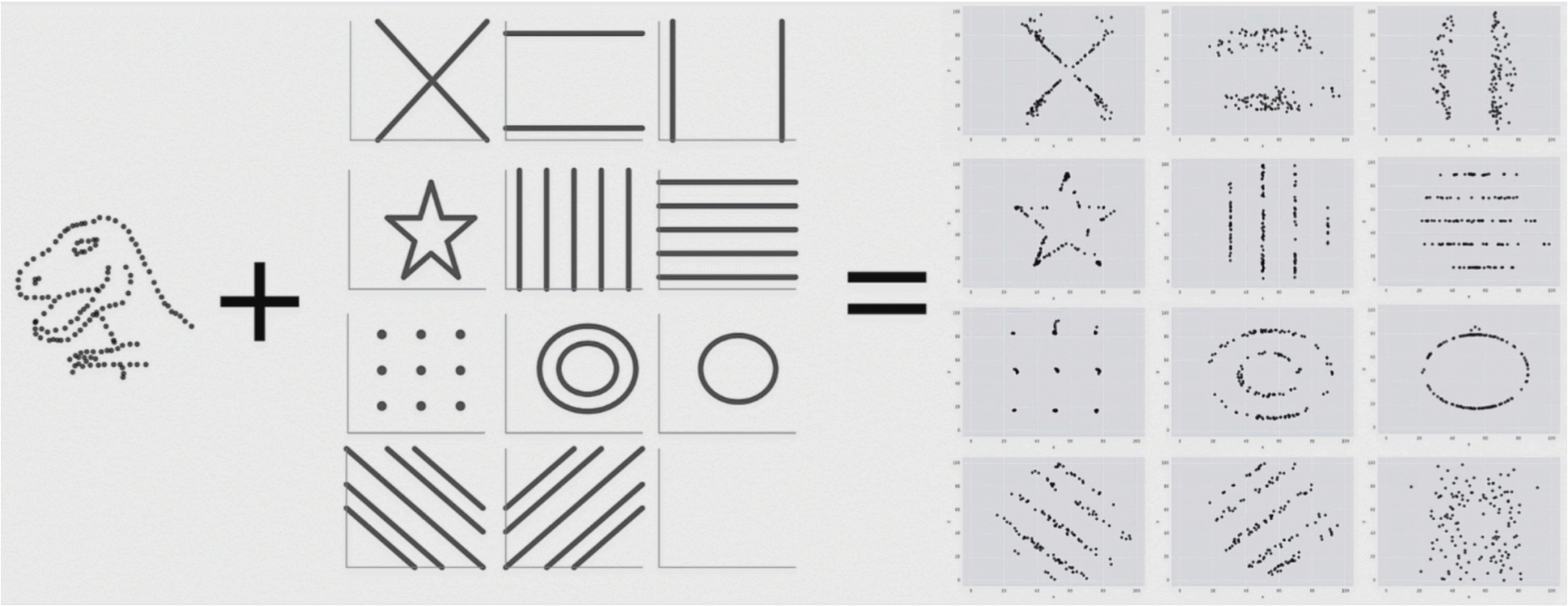
Datasaurus by Alberto Cairo



X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html
https://www.autodeskresearch.com/publications/samestats

Datasaurus dozen



Never trust summary statistics alone, always visualize your data

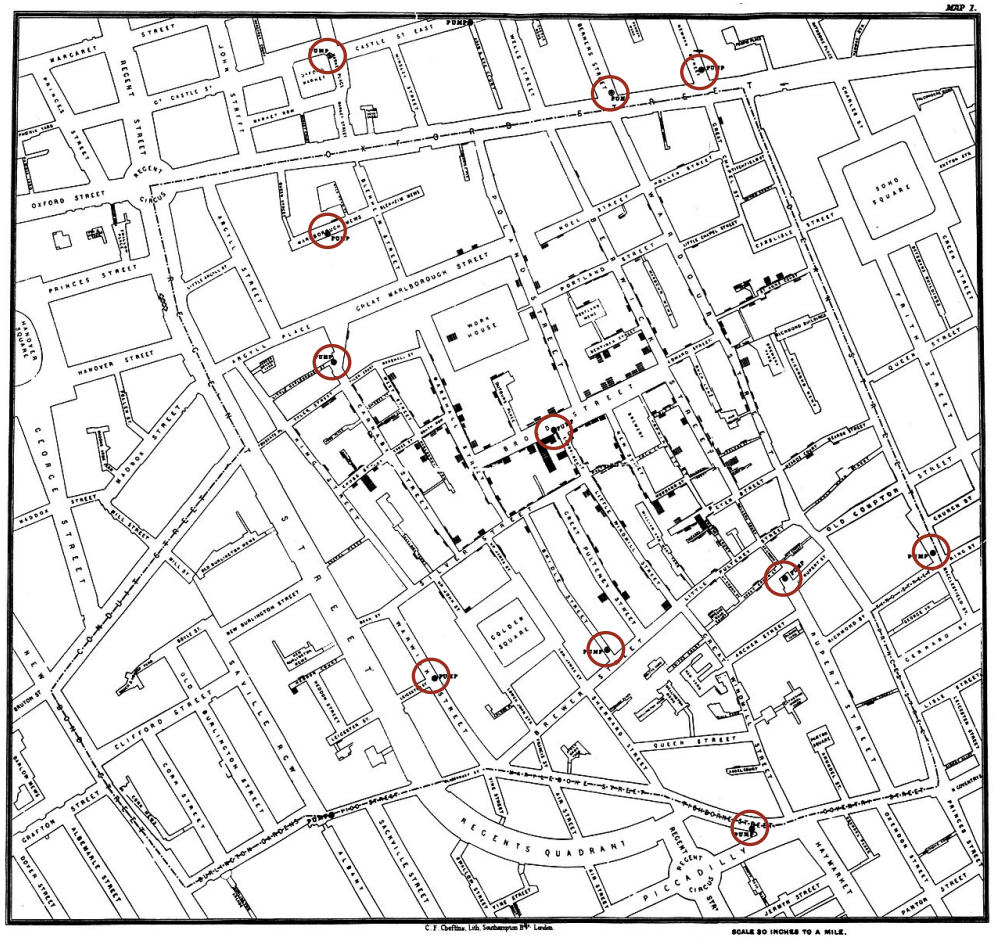
Cholera outbreak in London

- In 1854, more than 600 people died of cholera in London's Soho district
- Cause of the disease was unknown at the time
- Two competing theories
  - Cholera is spread by air (predominant)
  - Cholera is spread by water
- Physician John Snow gathered patient data and found the infected water pump
- To convince authorities to close the water pump, he drew a dot distribution map
  - One infected person = one 'dot'
  - Denoted the locations of the water pumps

Cholera outbreak in London

Cholera cases clustered around a public water pump on Broad Street

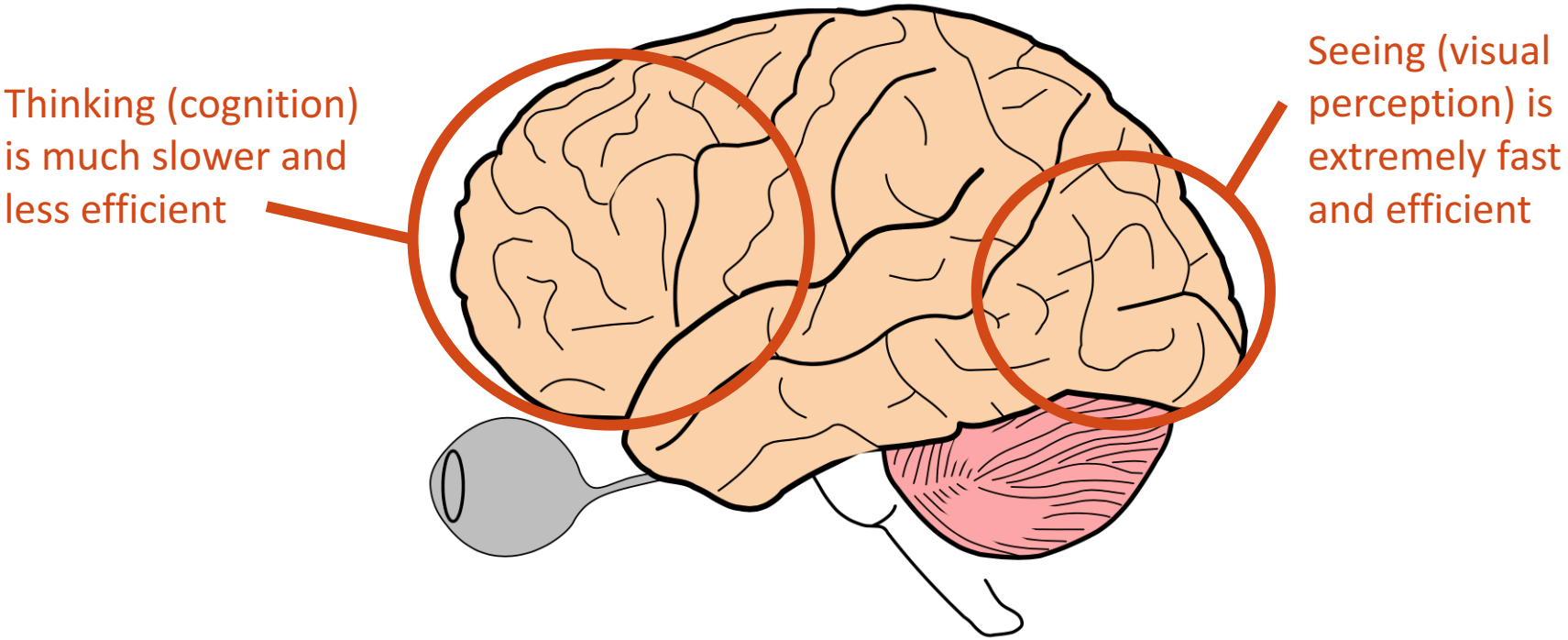
Jo(h)n Snow saved the day!



https://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg

Why does visualization work?

Perception vs. cognition



Data visualization is effective because it shifts the balance between perception and cognition to take fuller advantage of the brain's abilities

Purposes of data visualization

Analyze data to support reasoning

- Develop and assess hypotheses
- Discover errors in data
- Find patterns and correlations

Communicate information to others

- Present an argument or tell a story
- Inspire

The three principles of good visualization design

Good visualization design is

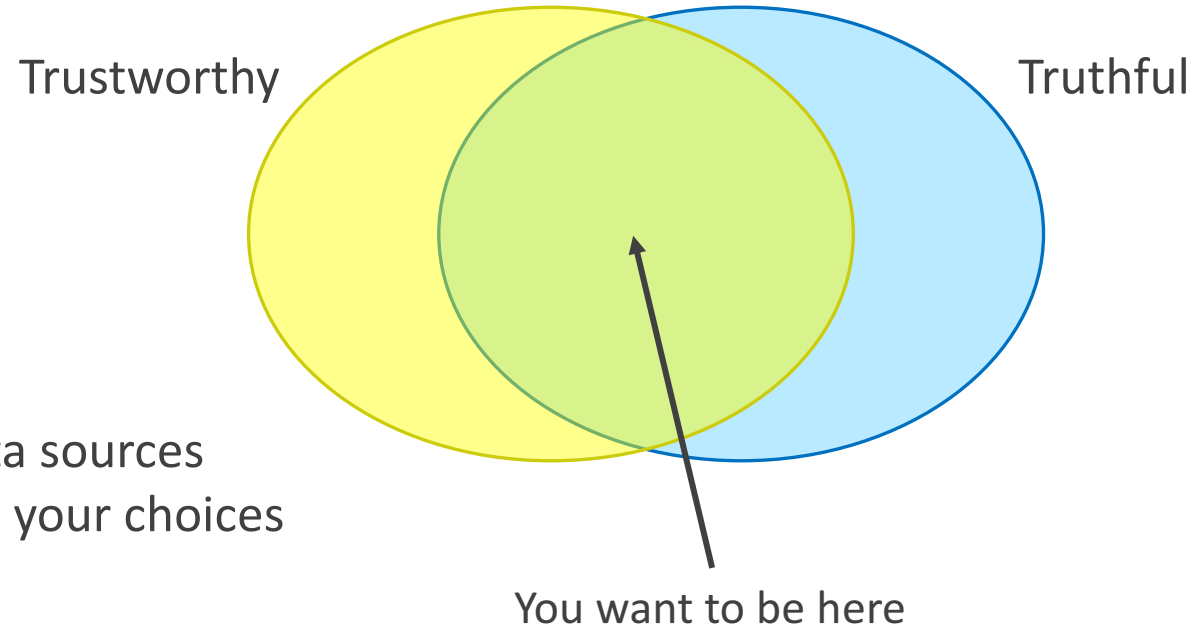
1. Trustworthy

2. Accessible

3. Elegant

Trustworthiness

Trust ≠ truth



List data sources
Explain your choices

Trustworthiness

Lying with visualization is easy

Intentionally and unintentionally

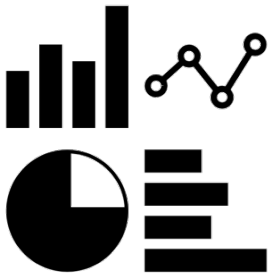
How charts lie?

Phenomenon

Data

Dubious data

Chart



Misrepresenting data

Cherry-picking data

Ignoring uncertainty

Person

Confirmation bias

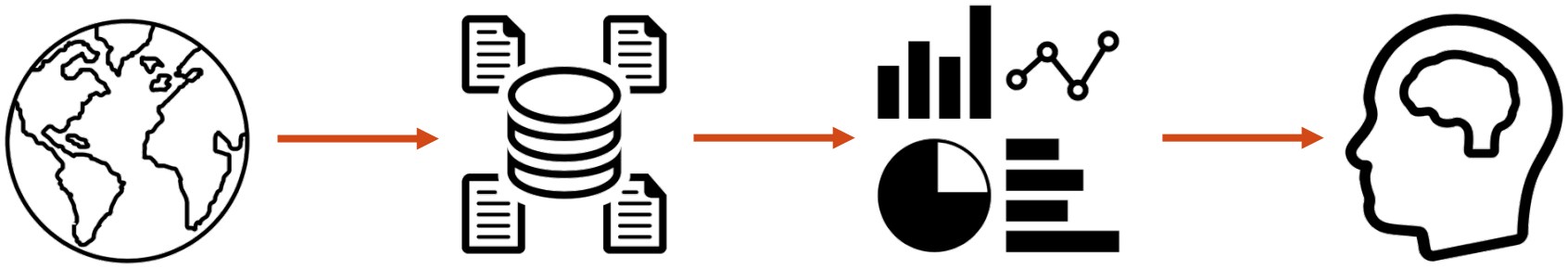
How charts lie?

Phenomenon

Data

Chart

Person



Dubious data

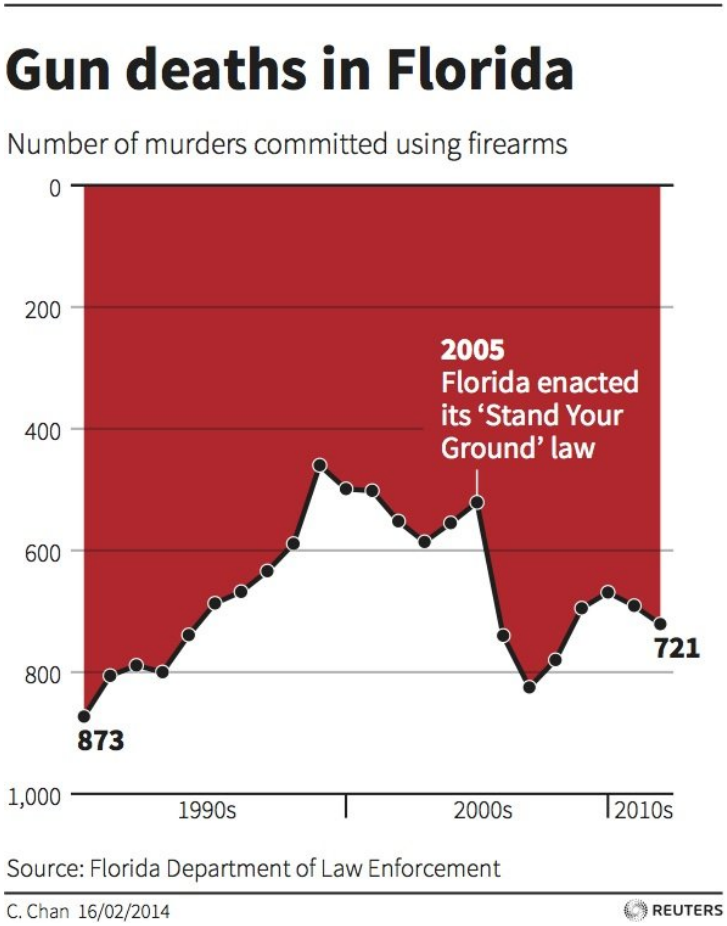
Misrepresenting data

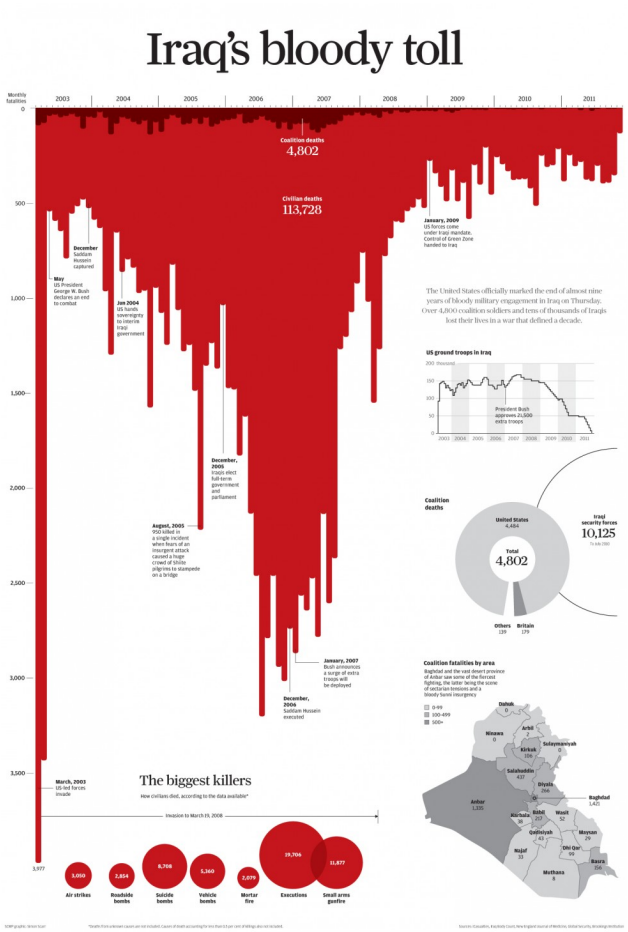
Cherry-picking data

Ignoring uncertainty

Confirmation bias

Inverted y axis





http://www.businessinsider.com/gun-deaths-in-florida-increased-with-stand-your-ground-2014-2
http://www.scmp.com/infographics/article/1284683/iraqs-bloody-toll

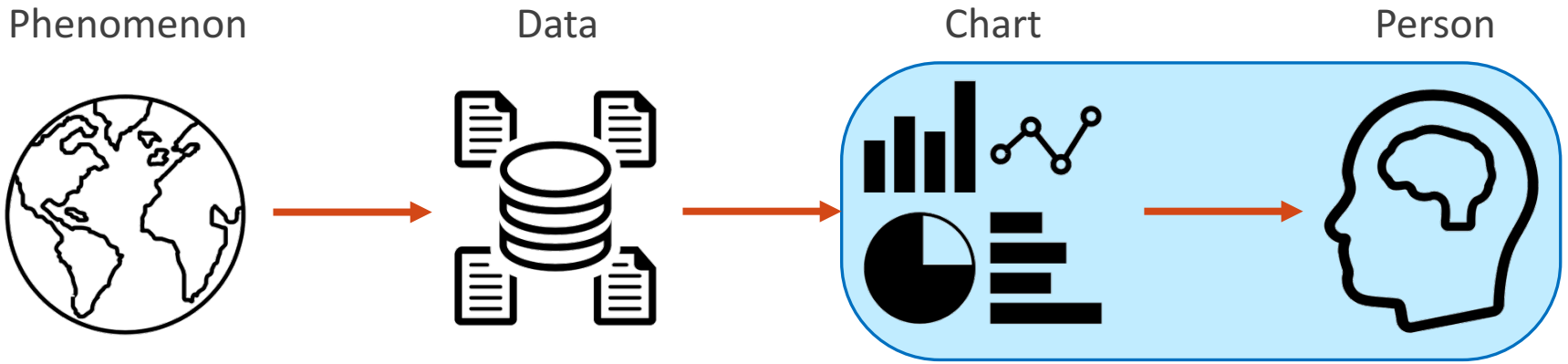
Good visualization design is

1. Trustworthy

2. Accessible

3. Elegant

Accessibility



There should be no obstacles between the visualization and the person that tries to understand it

Make design choices that facilitate understanding

An accessible visualization

- Is tailored to the audience (their needs, expectations, expertise)

Data visualization is like family photos. If you don't know the people in the picture, the beauty of the composition won't keep your attention.

Zach Gemignani, CEO/Founder of Juice Analytics

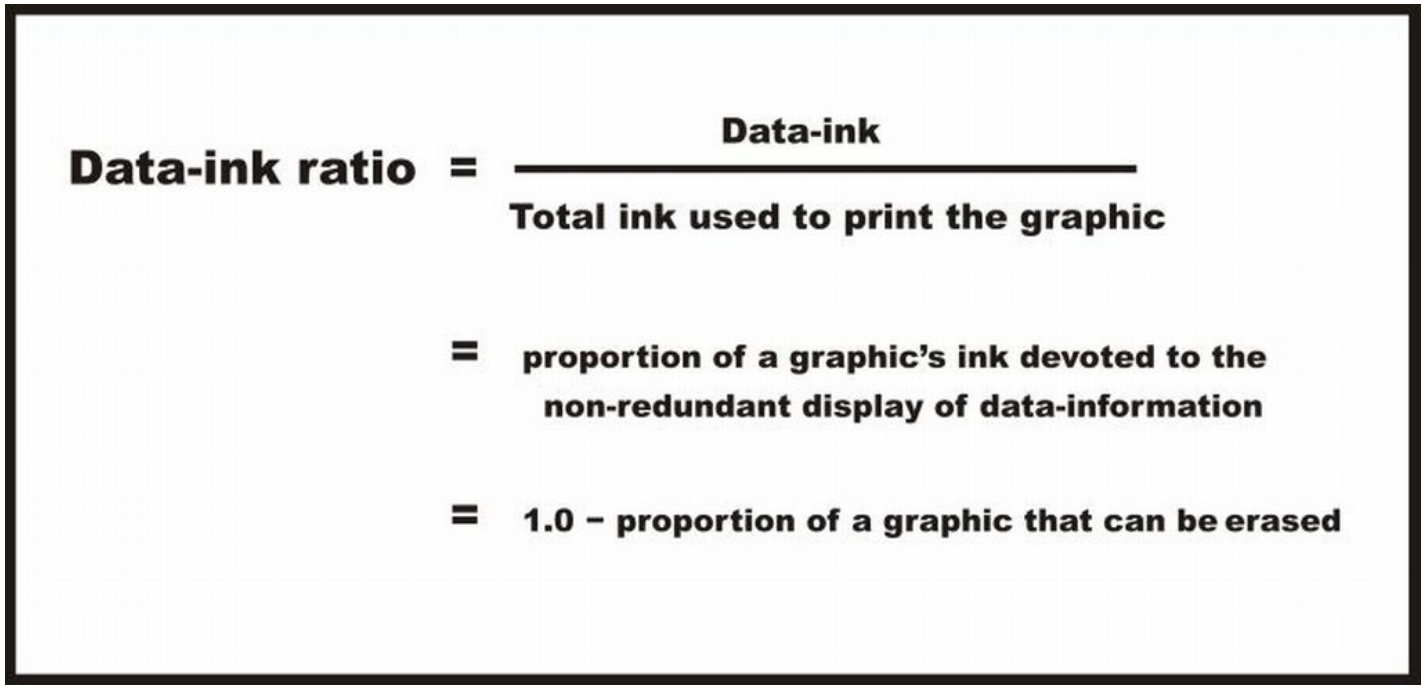
An accessible visualization

- Is tailored to the audience (their needs, expectations, expertise)
- Is appropriate for the given format (print, presentation, online, …)
- Is appropriate for the given data (type and values)
- Addresses a task (or tasks)
- Contains the appropriate amount of detail (clarity, not simplicity)
- Minimizes clutter (‘chart junk’)
- Takes into account human visual processing abilities
  - Is mindful of the choice of color (and other channels)
  - Uses annotations

Data-ink ratio

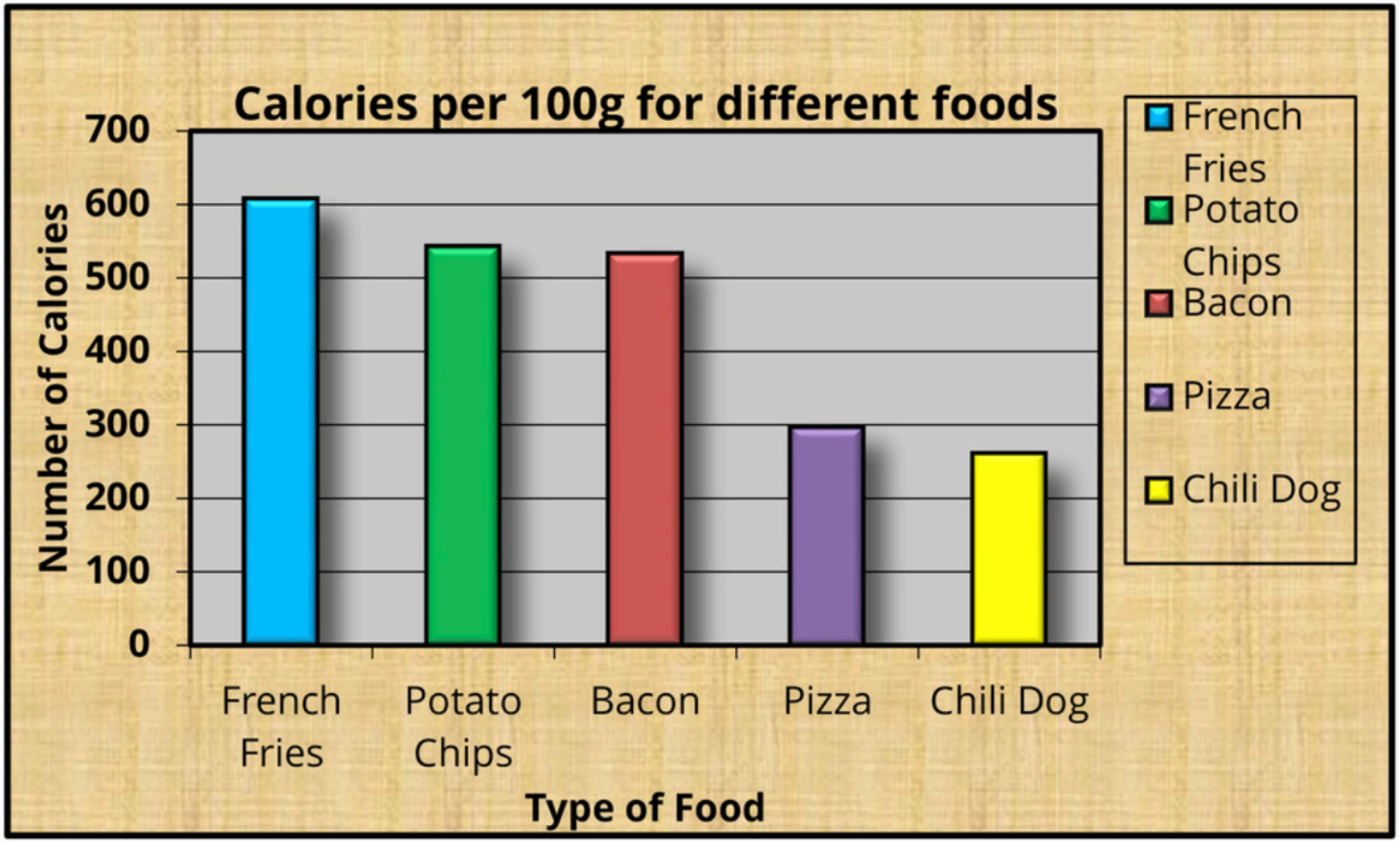
Above all else, show the data

Edward Tufte



https://infovis-wiki.net/wiki/Data-Ink_Ratio

Remove ‘chart junk’

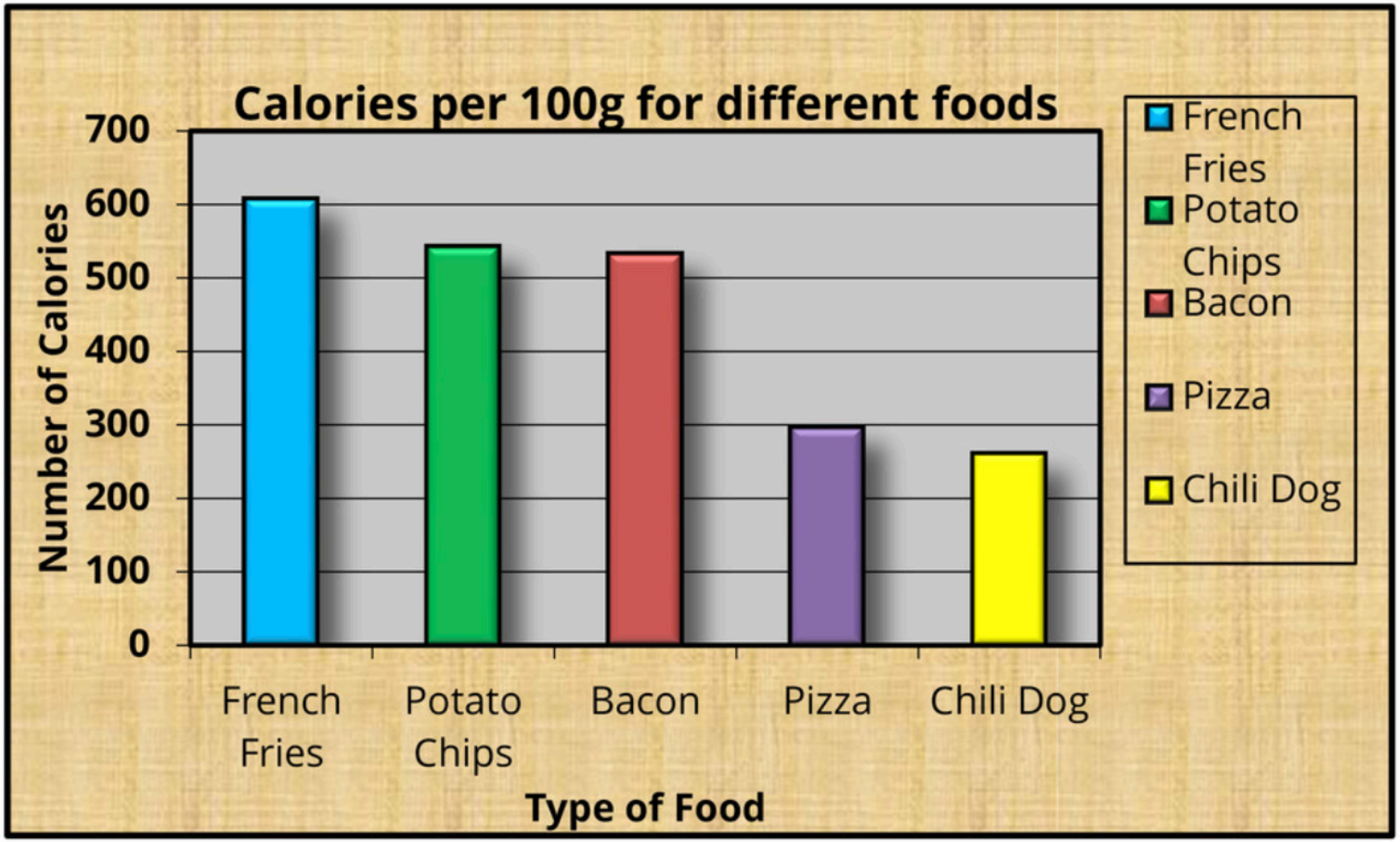


https://www.darkhorseanalytics.com/blog/data-looks-better-naked

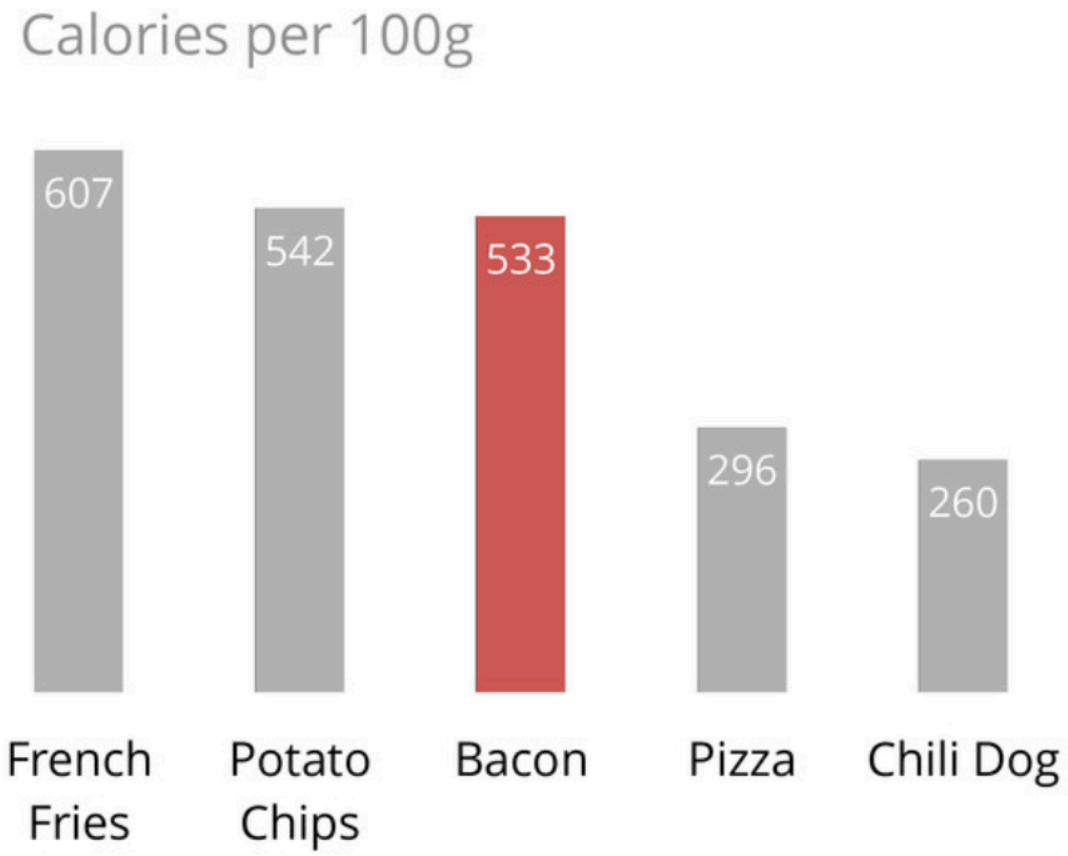
Remove ‘chart junk’

https://www.darkhorseanalytics.com/blog/data-looks-better-naked

Remove ‘chart junk’ – before

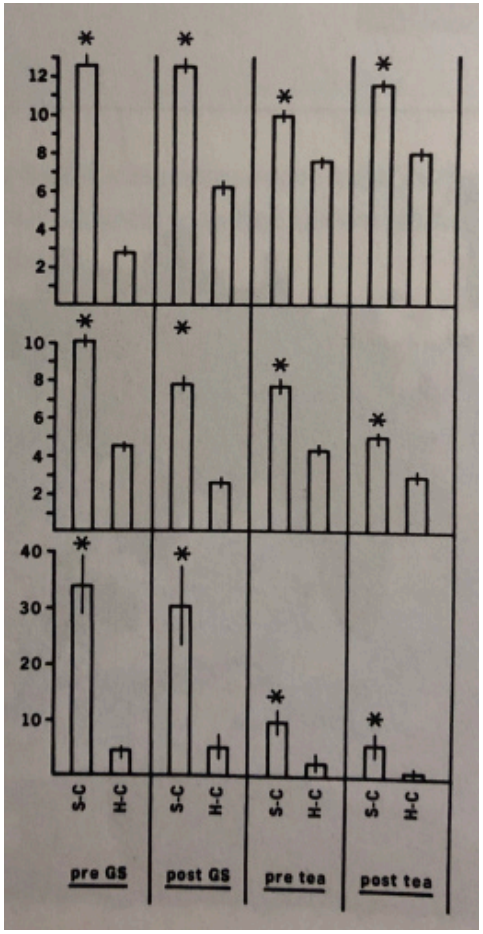


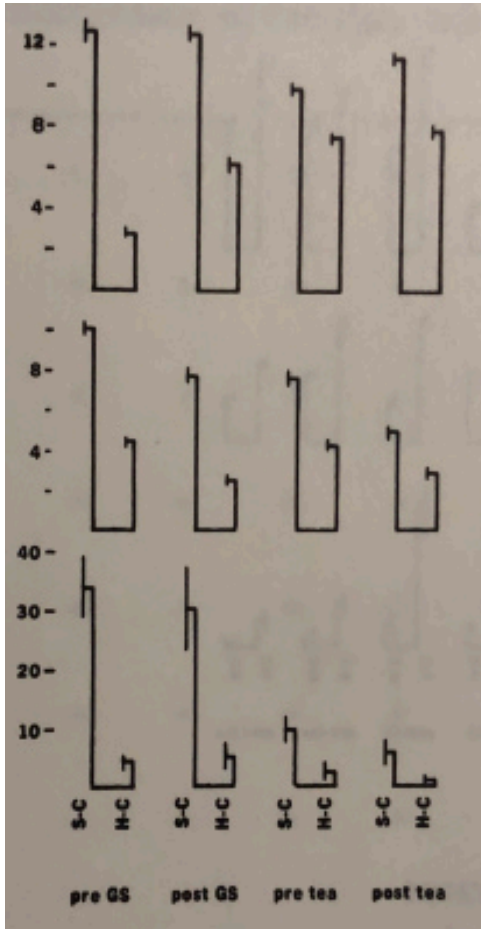
Remove ‘chart junk’ – after



https://www.darkhorseanalytics.com/blog/data-looks-better-naked

Going too far?



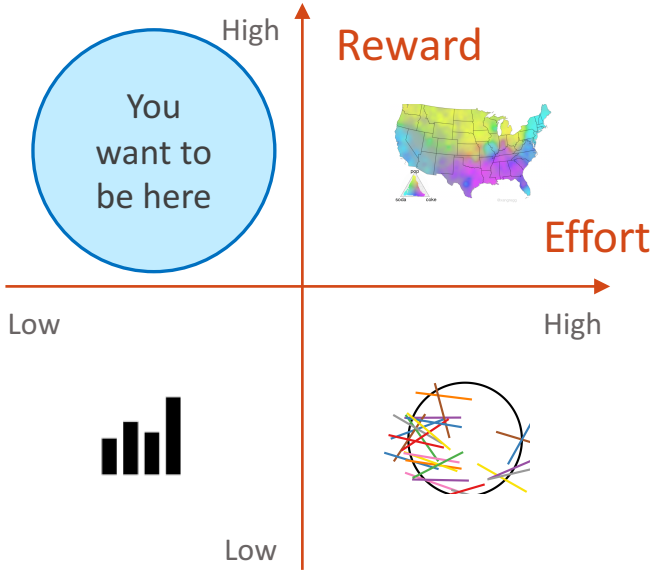


Minimalism relies on some familiarity of the concepts used (previous knowledge)

Using uncommon charts

Use an uncommon chart only if it shows something that the more common ones cannot

Always have in mind the trade-off between getting the message through and spending time to explain the more 'complex' chart



Good visualization design is

1. Trustworthy

2. Accessible

3. Elegant

Elegance

Don't make something unless it is both made necessary and useful; but if it is both necessary and useful, don't hesitate to make it beautiful.

Shaker dictum

Good design is as little design as possible

Rams' principle

Be inspired

Information is beautiful awards

Visualizing data (best of …)

New York Times' Upshot

Washington Post

Guardian's interactives

FiveThirtyEight

r/dataisbeautiful subreddit

Don’t get overwhelmed

The best visualizations take weeks of effort by multiple people – you are not expected to perform at that level

Keep in mind what is important:

1. Trustworthiness
2. Accessibility
3. Elegance (if there’s time)