

Data Visualization

FOUNDATIONS

Tea Tušar, Data Science and Scientific Computing, Information retrieval and data visualization

Outline

What is data visualization?

Why visualize data?

Historical visualizations

The three principles of good visualization design

- Trustworthiness
- Accessibility
- Elegance

What is data visualization?

3

Definition



The presentation of data in graphical form
to facilitate understanding

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

4

Distinctions in terminology

Data visualization \approx 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

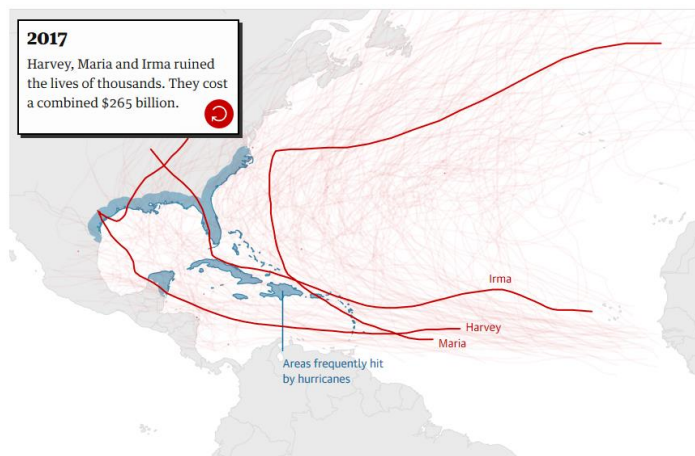
- Visualization of 3-D phenomena for scientific purposes

Infographics

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

5

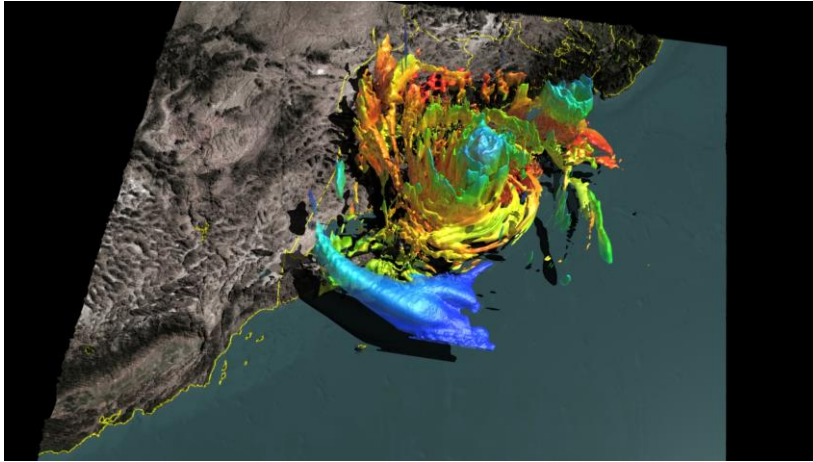
Data visualization example



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

6

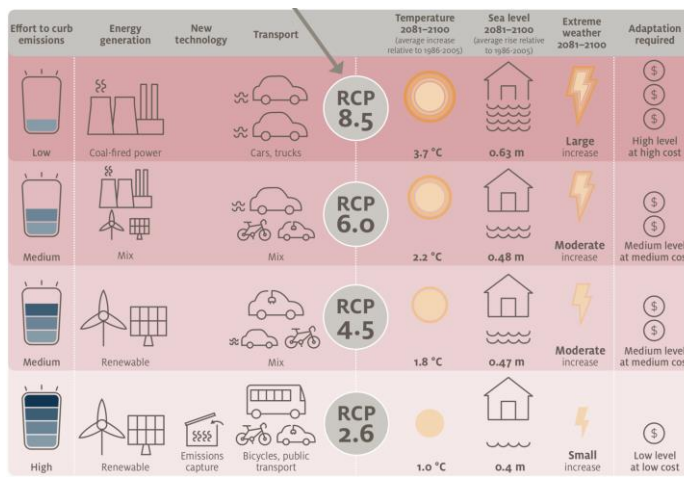
Scientific visualization example



<https://www2.cisl.ucar.edu/vislab>

7

Infographic example



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

8

Distinctions in terminology

Interchangeable use

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

9

Why visualize data?

'A PICTURE IS WORTH A THOUSAND WORDS'

10

Anscombe's quartet

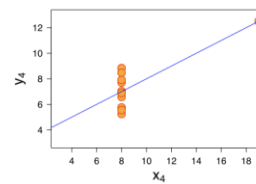
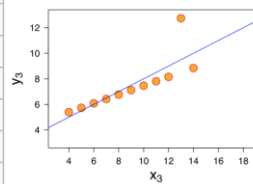
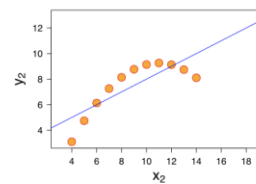
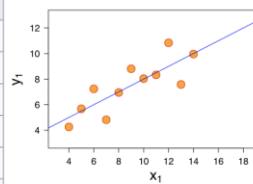
4 datasets with pairs of numbers (x, y) that have nearly identical simple descriptive statistics

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

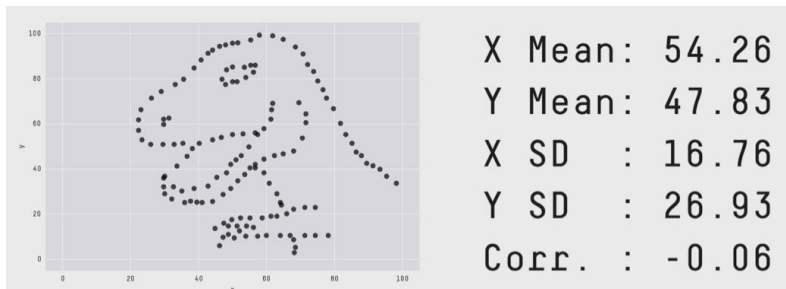


F. J. Anscombe. *Graphs in Statistical Analysis. American Statistician*, 27(1):17-21, 1973
https://en.wikipedia.org/wiki/Anscombe's_quartet

Datasaurus

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

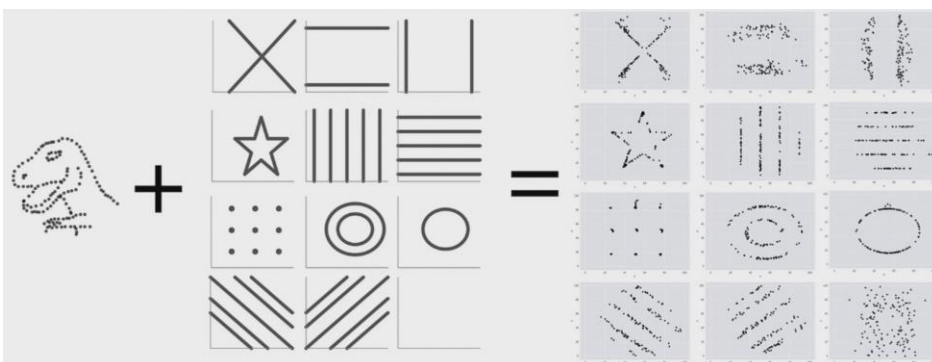
Datasaurus by Alberto Cairo



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

13

Datasaurus dozen



Never trust summary statistics alone, always visualize your data

<https://www.autodeskresearch.com/publications/samestats>

14

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' (actually short line)
 - Denoted the locations of the water pumps



15

Cholera outbreak in London

Cholera cases clustered around a public water pump on Broad Street



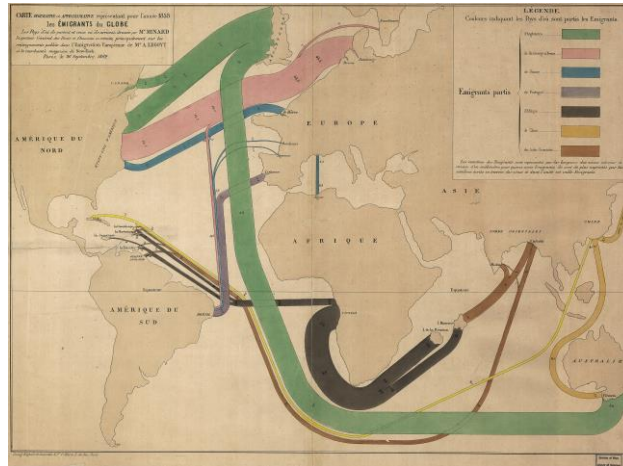
Jo(h)n Snow
saved the day!



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

16

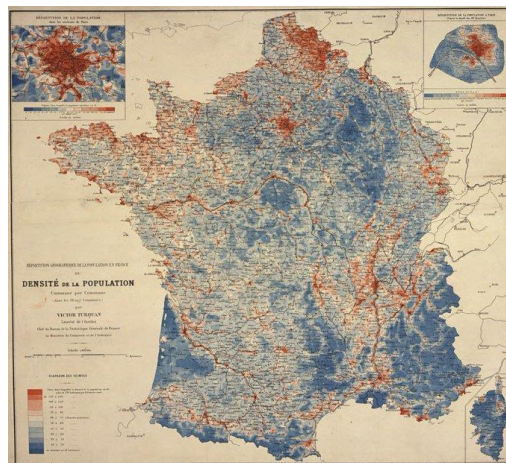
Immigration patterns



C. J. Minard. *Les Émigrants du Globe*, 1858.

19

Population density

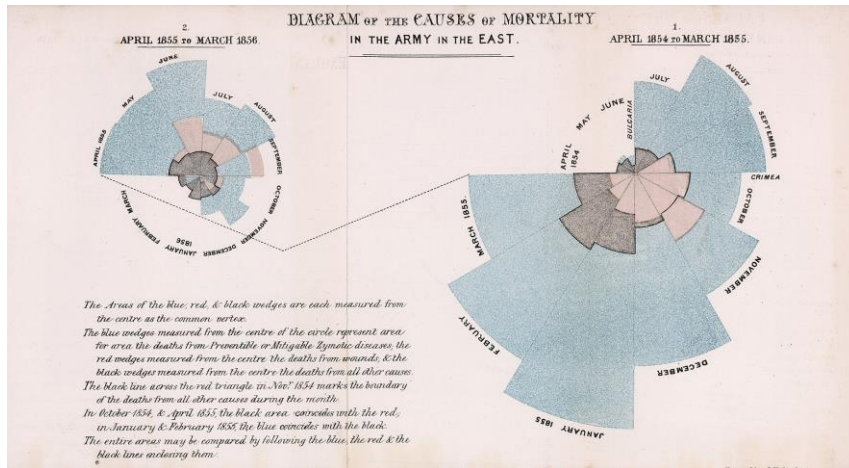


1887

<https://twitter.com/LOCMaps/status/1461337145059291143>

20

Causes of mortality in the army in the East



21

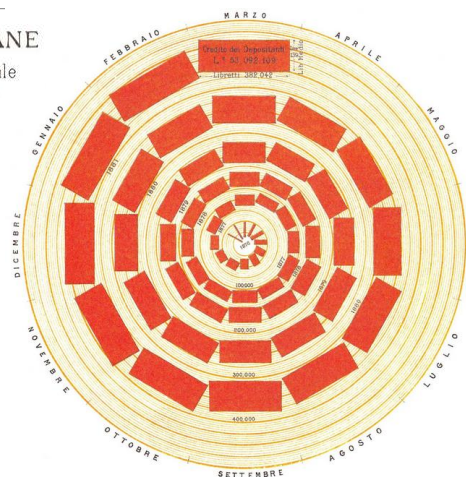
Italian postal savings

CASSE POSTALI DI RISPARMIO ITALIANE

Numero dei Libretti, Libretto medio e Deposito totale
al fine di ogni mese



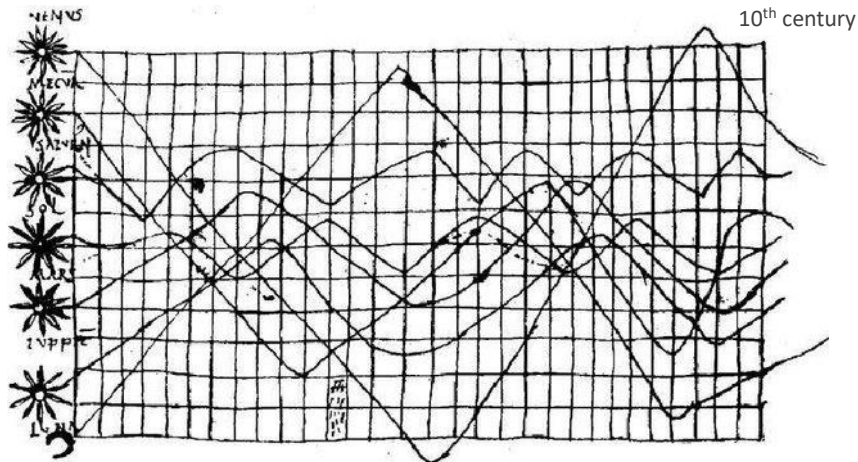
Data from 1876 to 1881



A. Gabaglio. *Teoria Generale della Statistica*, 1888.

22

Planetary movements



M. Friendly. A Brief History of Data Visualization, *Handbook of Computational Statistics: Data Visualization*, 2006.

23

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
- Educate (explain why data is important)
- Inspire

24

The three principles of good visualization design

25

Good visualization design is

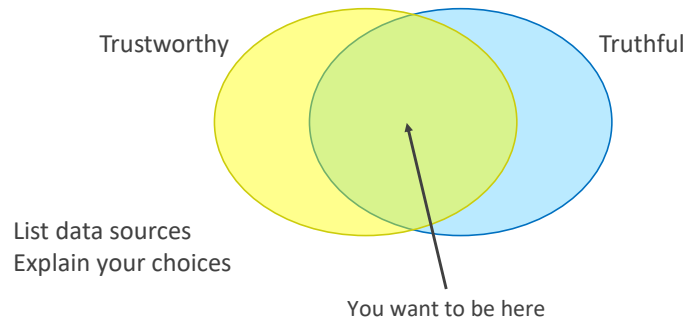
1. Trustworthy
2. Accessible
3. Elegant

A. Kirk. *Data Visualization*, SAGE Publications, 2016.

26

Trustworthiness

Trust \neq truth



27

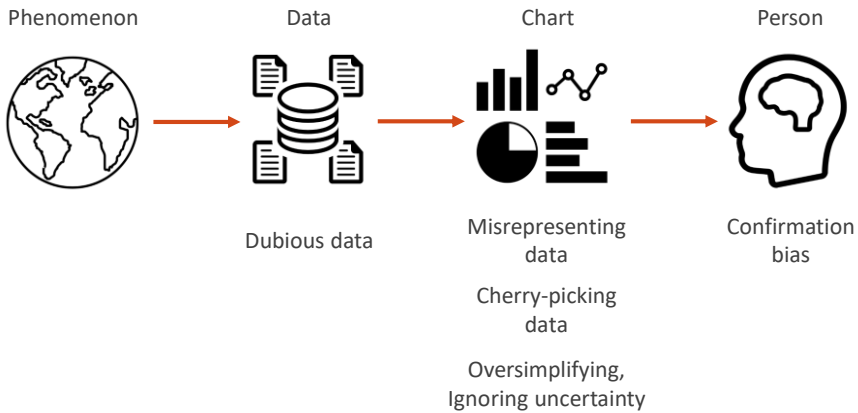
Trustworthiness

Lying with visualization is easy

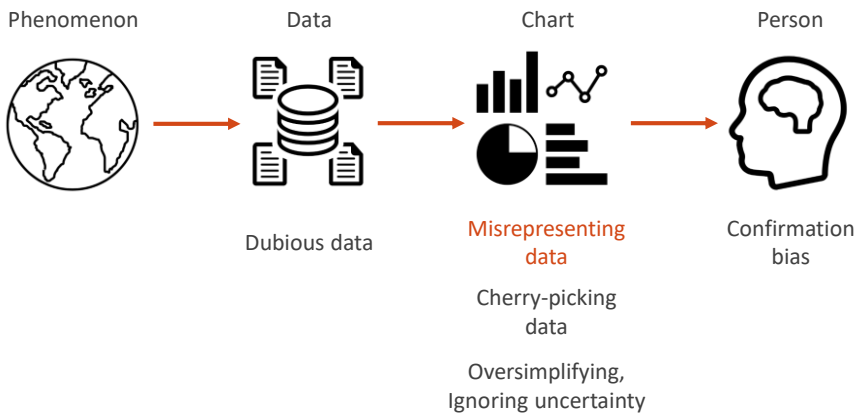
Intentionally and unintentionally

28

How charts lie?

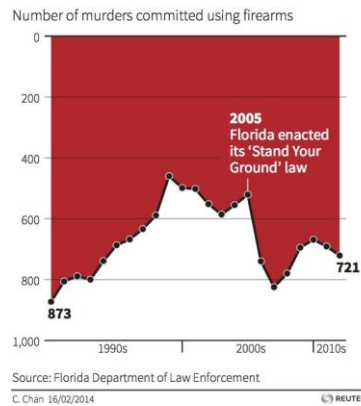


How charts lie?

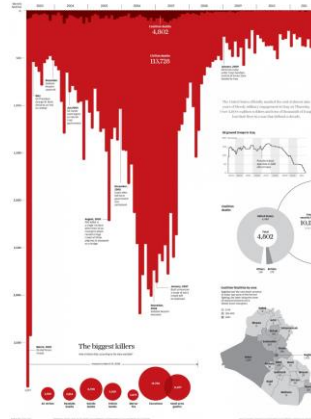


Inverted y axis

Gun deaths in Florida



Iraq's bloody toll



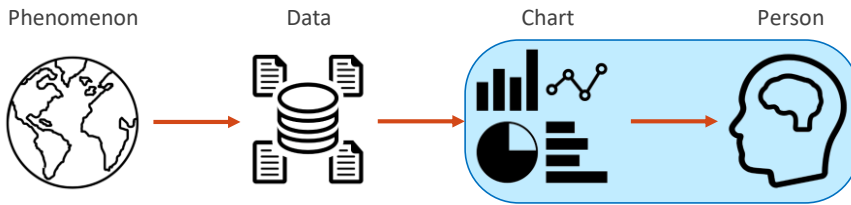
<http://www.businessinsider.com/gun-deaths-in-florida-increased-with-stand-your-ground-2014-2>
<http://www.scmp.com/infoGraphics/article/1284683/iraqs-bloody-toll>

31

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

33

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.

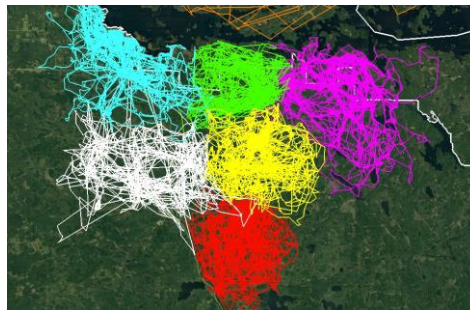
Zach Gemignani, CEO/Founder of Juice Analytics

34

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)

Movement of wolves



<https://earthlymission.com/gps-tracking-shows-how-much-wolf-packs-avoid-each-others-range/>

35

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 specific task (or tasks)
- Contains the appropriate amount of detail (clarity, not simplicity)
- Takes into account human visual processing abilities
 - Is mindful of the choice of color (and other channels)
 - Uses annotations
- Minimizes clutter ('chart junk')

36

Data-ink ratio

Above all else, show the data

Edward Tufte

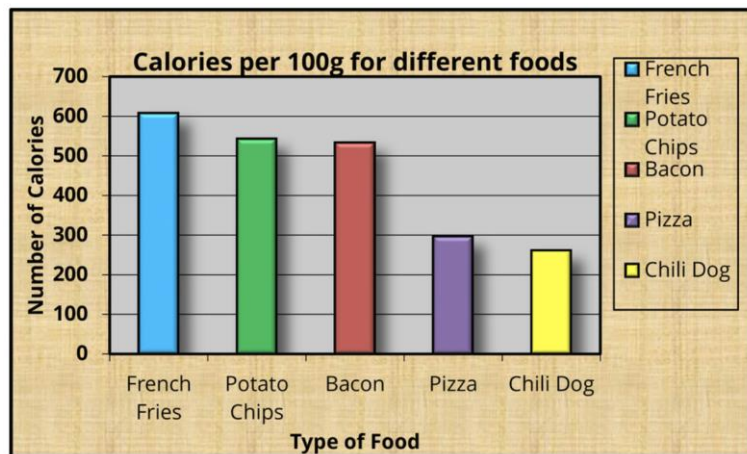
Data-ink ratio =
$$\frac{\text{Data-ink}}{\text{Total ink used to print the graphic}}$$

- = proportion of a graphic's ink devoted to the non-redundant display of data-information
- = 1.0 - proportion of a graphic that can be erased

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

37

Remove 'chart junk'



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

38

Remove 'chart junk'

Remove
to improve
(the **data-ink** ratio)

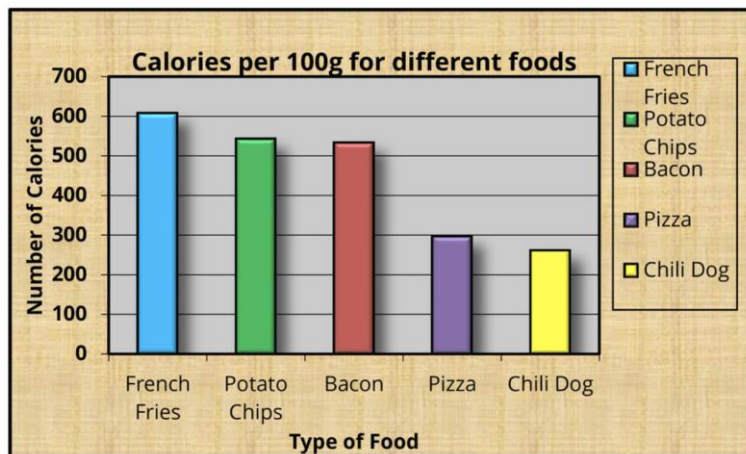
Created by Darkhorse Analytics

www.darkhorseanalytics.com

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

39

Remove 'chart junk' – before

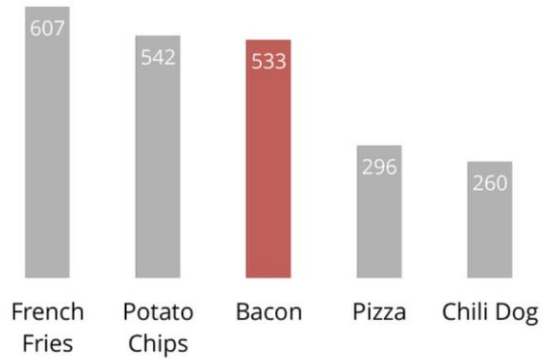


<https://www.darkhorseanalytics.com/blog/data-looks-better-naked>

40

Remove 'chart junk' – after

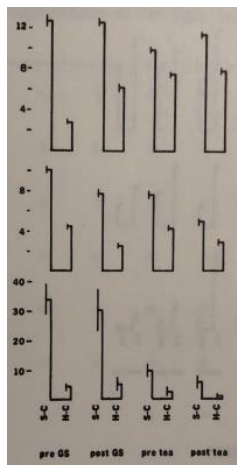
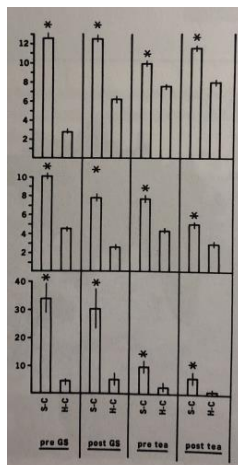
Calories per 100g



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

41

Going too far?



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

E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, 2015

42

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

Note: difficult-to-interpret charts increase active processing, so we remember them better

However, viewers often prefer understanding something quickly to remembering it for a long time

<https://blog.datawrapper.de/in-defense-of-simple-charts/>

43

Peak time of day for sports and leisure

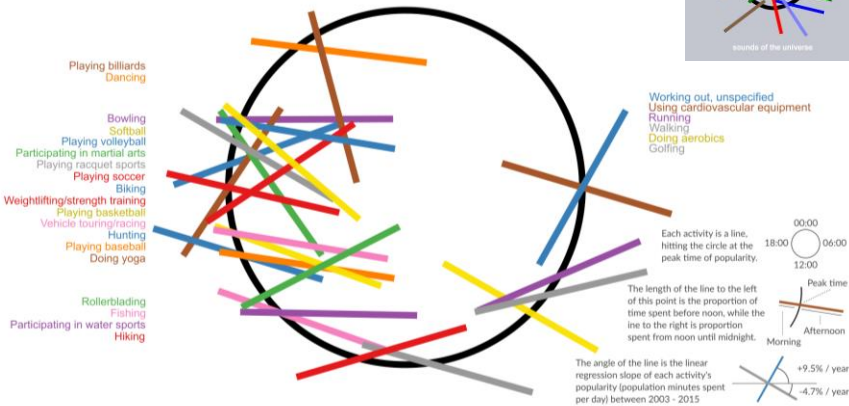
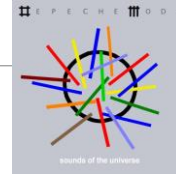


<https://github.com/halhen/viz-pub/tree/master/sports-time-of-day>

44

Peak time of day for sports and leisure

@mrkhdng | Source: American Time Use Survey

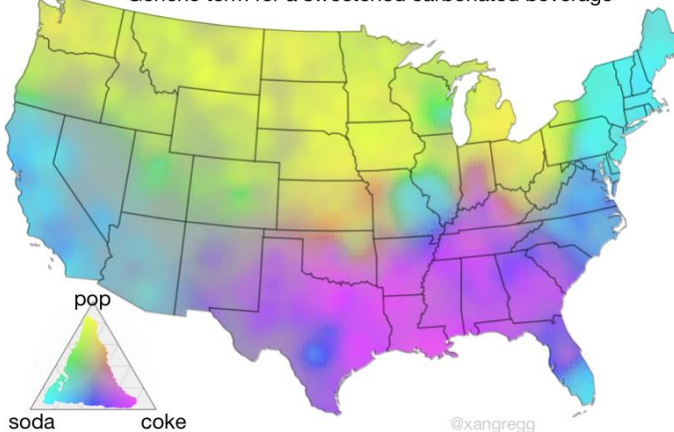


<https://github.com/halhen/viz-pub/tree/master/sports-depeche-plot>

45

Soda/coke/pop map

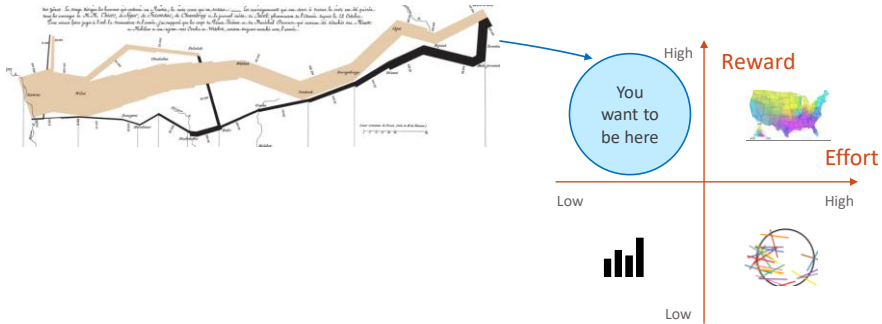
Generic term for a sweetened carbonated beverage



https://www.reddit.com/r/dataisbeautiful/comments/99rgnm/sodapopcoke_map_with_a_trivariate_color_encoding/

46

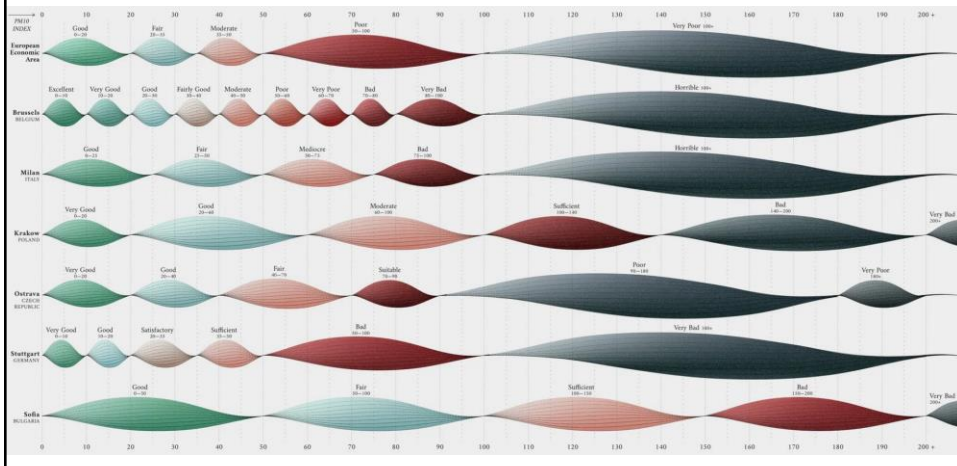
Using uncommon charts



Good visualization design is

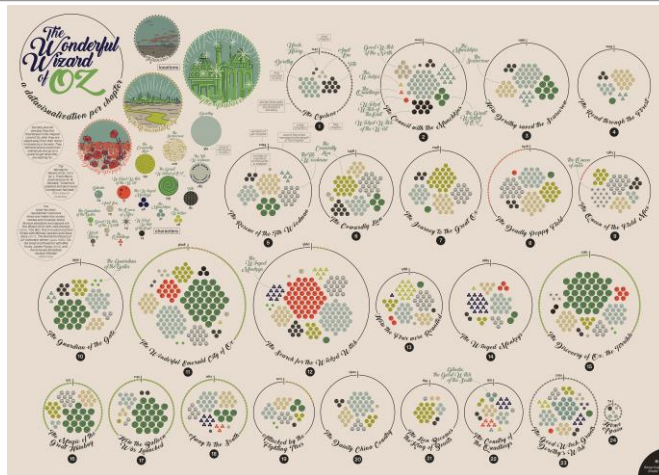
1. Trustworthy
2. Accessible
3. Elegant

Elegance



<https://nightingaledvs.com/what-makes-a-data-visualisation-elegant/>

Elegance



<https://www.domestika.org/en/courses/3069-effective-data-visualization-transform-information-into-art>

How to define elegance?

- Aesthetically nourishing (charms, creates a positive first impression)
- Stylistically coherent (matches the subject)
- Refined (simple, concise and to the point)
- Accessible (clearly understandable)
- Functionally delightful (usable, efficient)
- Noticeable when absent

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' The Upshot](#)

[Guardian's interactives](#)

[FiveThirtyEight](#)

[r/dataisbeautiful subreddit](#)

53

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)

54