

Data Visualization

EXAMPLES (1)

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

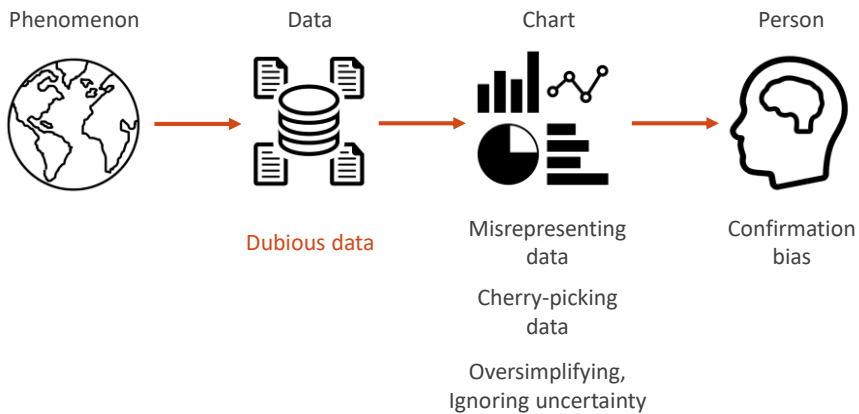
Good visualization design is

1. Trustworthy
2. Accessible
3. Elegant

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

2

How charts lie?



3

Dubious data

Unrepresentative data

- Missing data
- Polls on unrepresentative populations
- Measurements on unrepresentative samples

Biased data

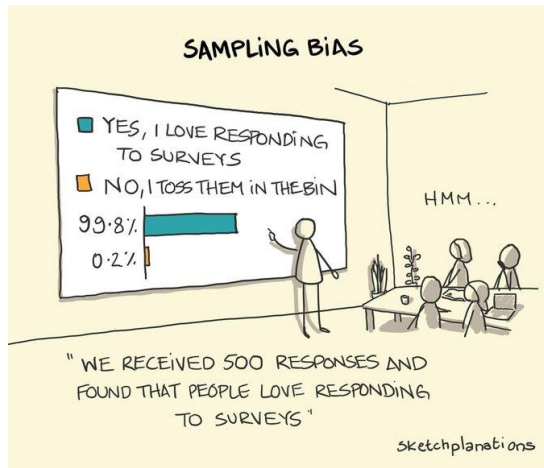
- Question framing in polls
- Choice of measures

Wrong comparisons

- Non-comparable data
- Absolute instead of cumulative data (and vice versa)
- Absolute instead of relative data

4

Unrepresentative samples



<https://sketchplanations.com/sampling-bias>

5

Unrepresentative samples



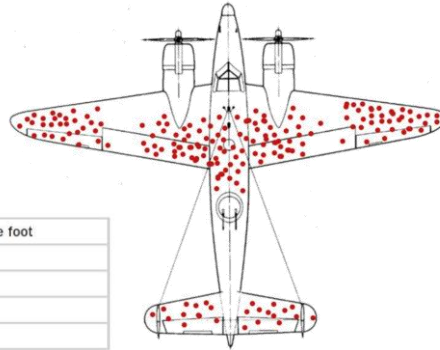
[https://www.reddit.com/r/dataisbeautiful/comments/9pkka4/all recorded meteorite impacts in the us from](https://www.reddit.com/r/dataisbeautiful/comments/9pkka4/all_recorded_meteorite_impacts_in_the_us_from)

6

Unrepresentative samples

Abraham Wald and the Missing Bullet Holes

Armour planes so that they don't get shot by enemy fighters. Armour is heavy, so use it only where is really needed.



| Section of plane | Bullet holes per square foot |
|-------------------|------------------------------|
| Engine | 1.11 |
| Fuselage | 1.73 |
| Fuel system | 1.55 |
| Rest of the plane | 1.8 |

<https://medium.com/@penguinpress/an-excerpt-from-how-not-to-be-wrong-by-jordan-ellenberg-664e708cfc3d>

7

Question framing in polls

Brexit referendum

- First proposal

“Should the United Kingdom remain a member of the European Union?”

yes/no

- Final question

“Should the United Kingdom remain a member of the European Union or leave the European Union?”

remain/leave

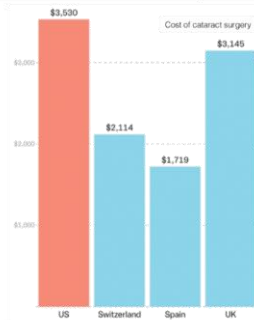
8

Non-comparable data used in comparisons

Vox

TWITTER SHARE

America's health care prices are out of control. These 11 charts prove it.



Two issues

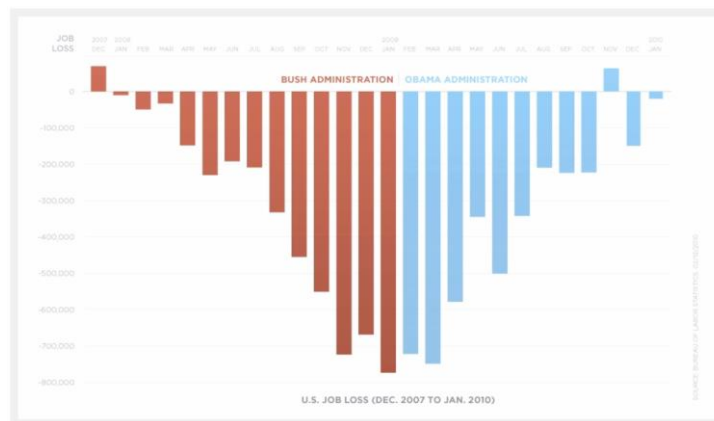
- Prices not adjusted for purchasing power
- Different sources of data

The data source specifically warns against using this data for comparison

https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

9

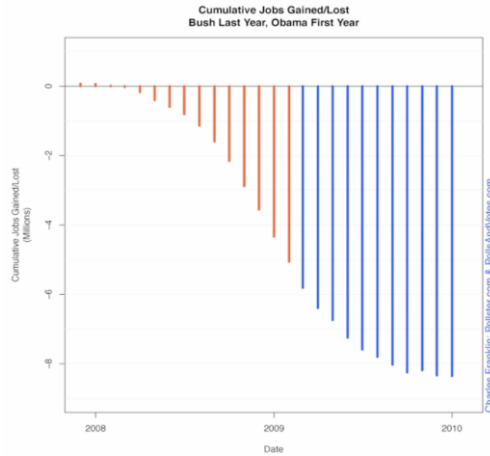
Absolute instead of cumulative data



<https://itunes.apple.com/us/course/data-literacy-and-data-visualization/id693097601>

10

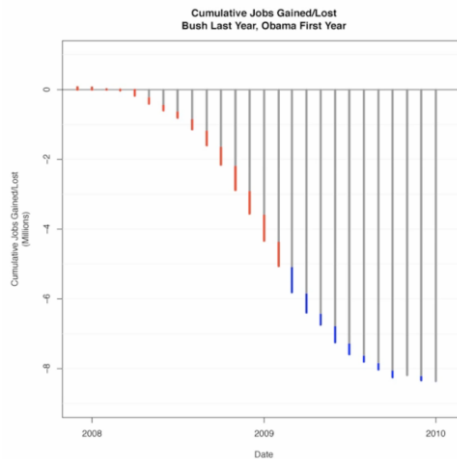
Cumulative data



<https://itunes.apple.com/us/course/data-literacy-and-data-visualization/id693097601>

11

Absolute and cumulative data



<https://itunes.apple.com/us/course/data-literacy-and-data-visualization/id693097601>

12

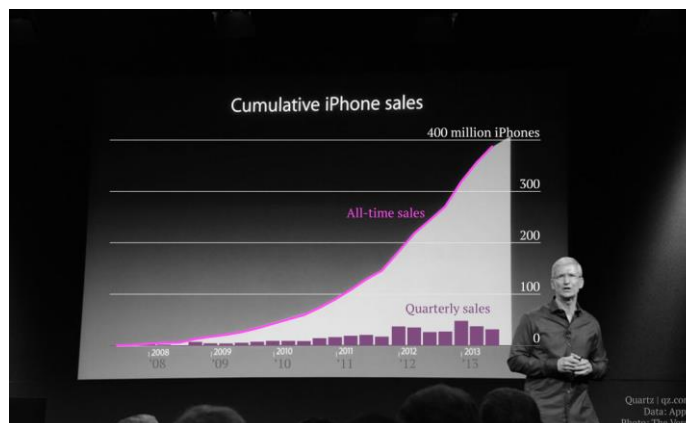
Cumulative instead of absolute data



<https://qz.com/122921/the-chart-tim-cook-doesnt-want-you-to-see/>

13

Cumulative and absolute data



<https://qz.com/122921/the-chart-tim-cook-doesnt-want-you-to-see/>

14

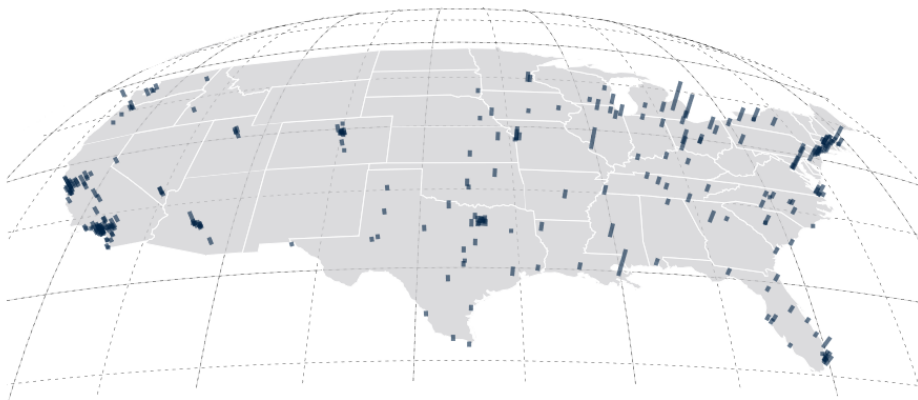
Absolute data



<https://kartograph.org/showcase/3d/>

15

Relative data



<https://kartograph.org/showcase/3d/>

16

Dubious data

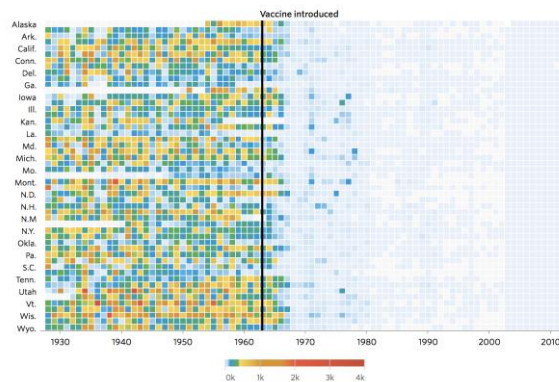
Garbage in, garbage out

Can still work if the issues are explained

17

Dubious data explained

Measles

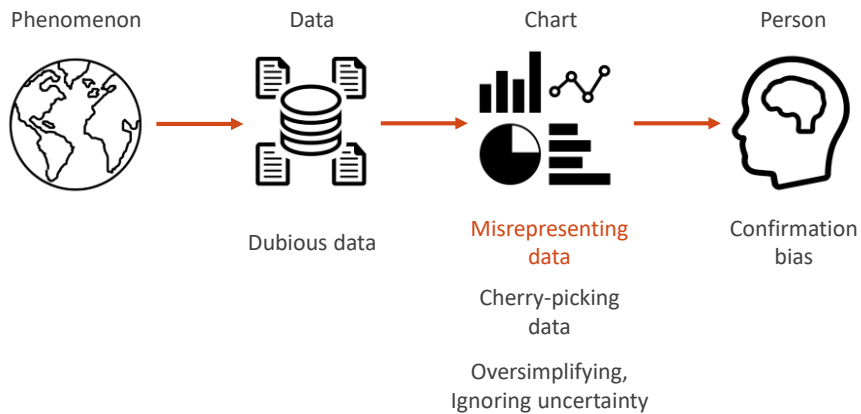


Note: CDC data from 2003-2012 comes from its Summary of Notifiable Diseases, which publishes yearly rather than weekly and counts confirmed cases as opposed to provisional ones.

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

18

How charts lie?



19

Misrepresenting data

Ignoring conventions

- Placement of dependent and independent variables
- Pie charts that do not add up to 100%

Abusing scales

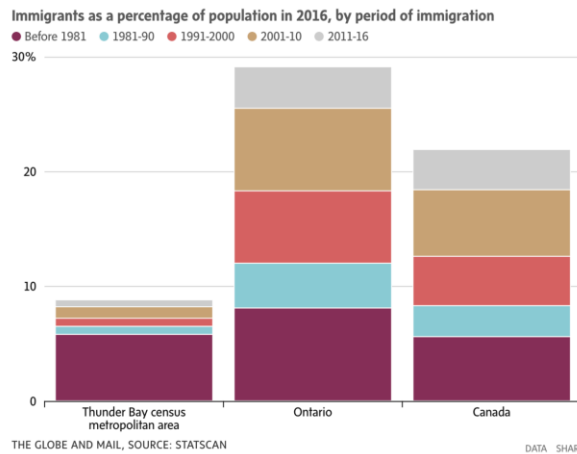
- Distorted axis
- Truncated/elongated axis
- Dual axes
- Improper scaling of areas and pictograms

Unnecessary 3-D

Improper categorization

20

Time not on an axis



<https://viz.wtf/post/187558414596/the-axis-choices-are-interesting-this-thing-is>

21

Over 100% pie chart

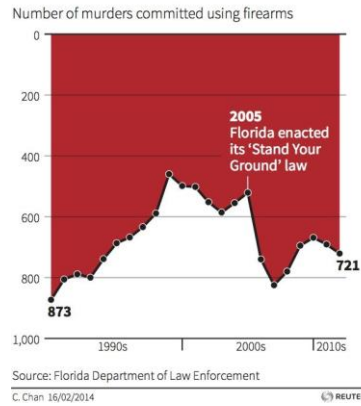


<https://twitter.com/dergigi/status/1243315176000180224/>

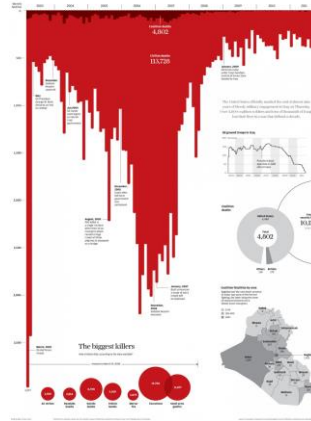
22

Distorted axis: Inverted y axis

Gun deaths in Florida



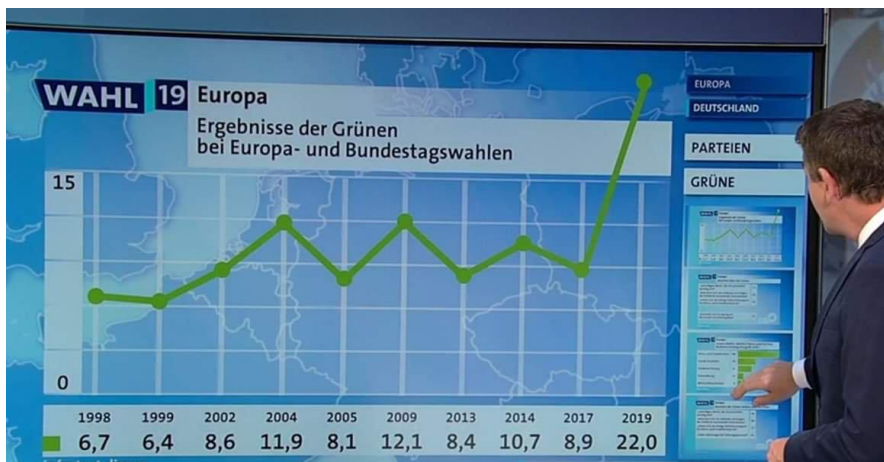
Iraq's bloody toll



<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>

23

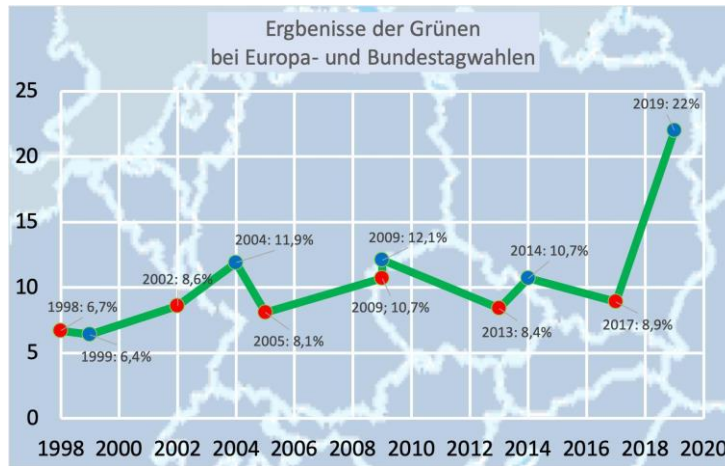
Distorted axis



<https://twitter.com/maartenzam/status/1132961446592172032?s=12>

24

Fixed axis



<https://twitter.com/maartenzam/status/1132961446592172032?s=12>

25

Unequal intervals

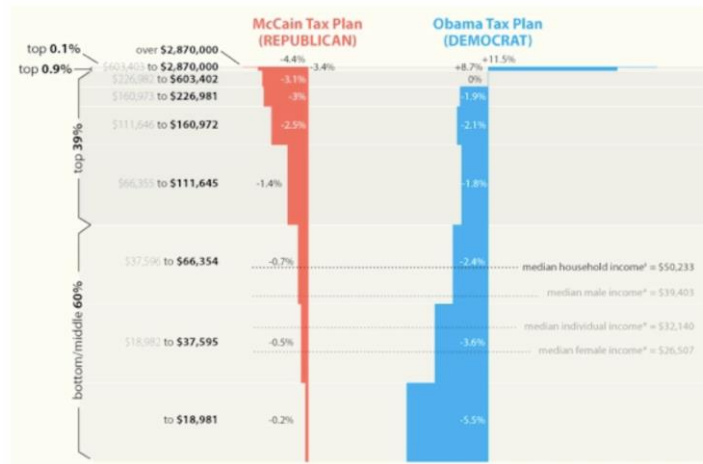
| FAMILY INCOME In 2008 dollars | McCain | | Obama | |
|--------------------------------------|--------------|-------------------------|--------------|---------------|
| | TAX DECREASE | AVERAGE CHANGE IN TAXES | TAX DECREASE | TAX INCREASE |
| Above \$2.87 million (top 0.1%) | -4.4% | -\$269,364 | +11.5% | +\$701,885 |
| \$603,403 to \$2.87 million (top 1%) | -3.4 | -\$45,361 | +8.7 | +\$115,974 |
| \$226,982 to \$603,402 | -3.1 | -\$7,871 | 0 | -\$12 |
| \$160,973 to \$226,981 | -3 | -\$4,380 | -1.9 | -\$2,789 |
| \$111,646 to \$160,972 | -2.5 | -\$2,614 | -2.1 | -\$2,204 |
| \$66,355 to \$111,645 | -1.4 | -\$1,009 | -1.8 | -\$1,290 |
| \$37,596 to \$66,354 | -0.7 | -\$319 | -2.4 | -\$1,042 |
| \$18,982 to \$37,595 | -0.5 | -\$113 | -3.6 | -\$892 |
| Up to \$18,981 | -0.2 | -\$19 | -5.5 | -\$567 |
| Average cut: | -2% | -\$1,195 | -0.3% | -\$160 |

Bottom three groups amount to 60% of taxpayers

<https://itunes.apple.com/us/course/data-literacy-and-data-visualization/id693097601>

26

Fixed intervals



<https://itunes.apple.com/us/course/data-literacy-and-data-visualization/id693097601>

27

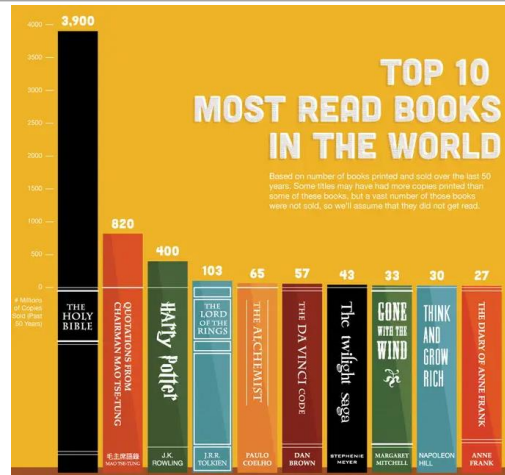
Bar chart with truncated axis



<https://medium.com/@stephen.tracy/how-to-lie-with-charts-97396e4642>

28

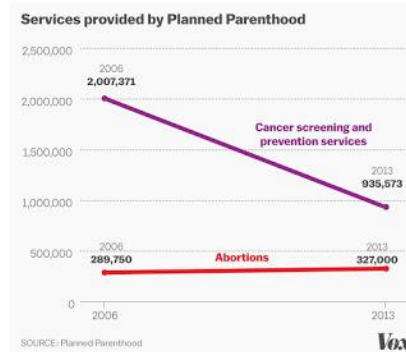
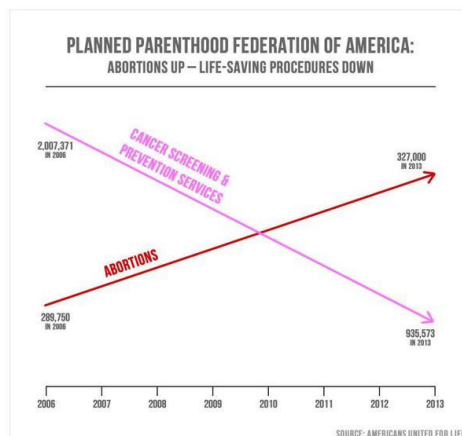
Bar chart with elongated axis



<https://www.businessinsider.com/infographic-the-top-10-most-read-books-in-the-world-2012-5>

29

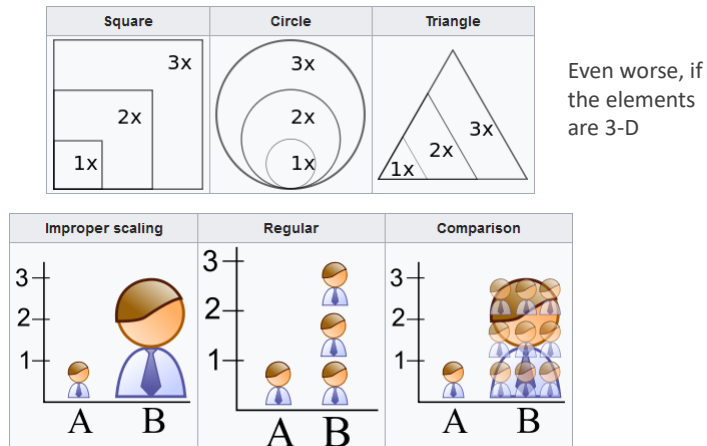
Dual axes



<http://www.thefunctionalart.com/2015/10/if-you-see-bullshit-say-bullshit.html>

30

Improper scaling of areas/pictograms



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

31

Improper scaling of areas/pictograms



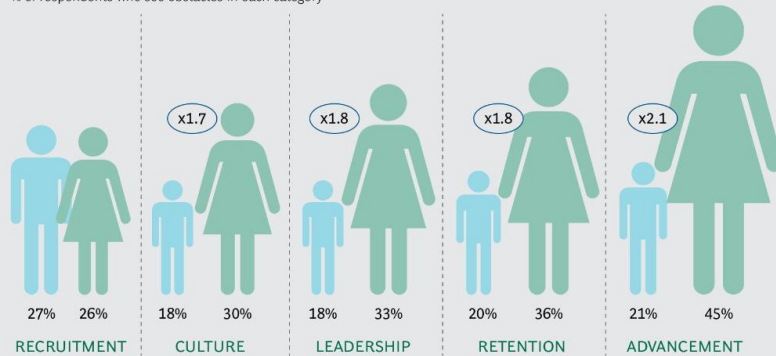
<https://gizmodo.com/why-i-paid-150-to-import-a-tamagotchi-smartwatch-i-can-1848094289>

32

Improper scaling of areas/pictograms

EXHIBIT 2 | Men and Women Rank Obstacles to Gender Diversity Differently

% of respondents who see obstacles in each category



Source: BCG Global Gender Diversity Survey 2017.

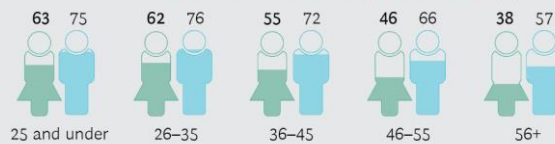
<https://www.bcg.com/publications/2017/people-organization-behavior-culture-getting-the-most-from-diversity-dollars.aspx?linkId=59621326&redir=true>

33

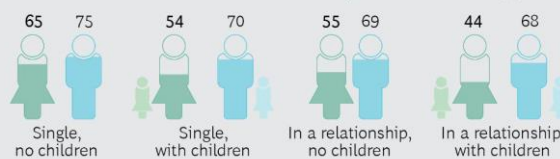
Proper scaling of areas/pictograms

Women Show The Greatest Willingness to Move When They Are Young

WILLINGNESS TO MOVE ABROAD, BY AGE GROUP (%)



WILLINGNESS TO MOVE ABROAD, BY FAMILY STATUS (%)

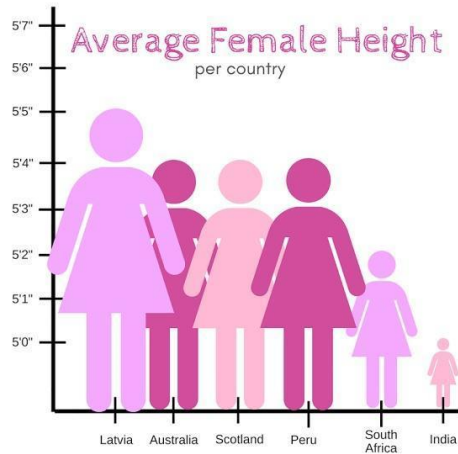


Source: BCG analysis.

<https://www.bcg.com/en-us/publications/2017/people-organization-leadership-talent-women-on-the-move.aspx?linkId=58878371&redir=true>

34

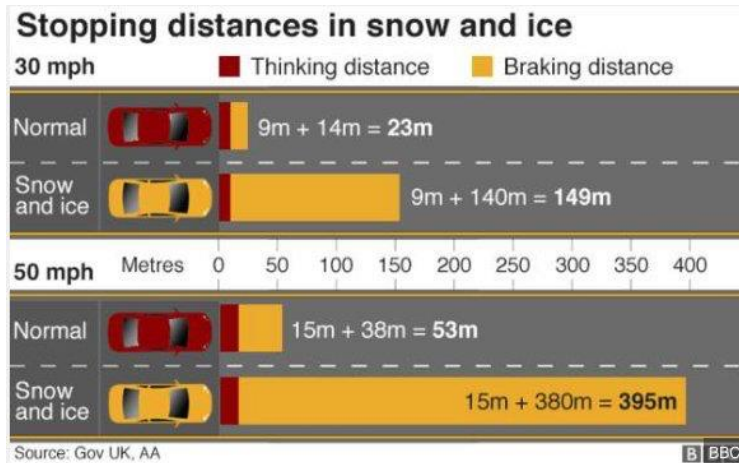
Improper scaling of areas/pictograms AND axis



<https://twitter.com/graphcrimes/status/1448190784239554563>

35

Improper scaling of areas/pictograms



<https://twitter.com/bengoldacre/status/1091326000384929793>

36

Improper scaling of areas/pictograms – fixed

Stopping distances in snow and ice



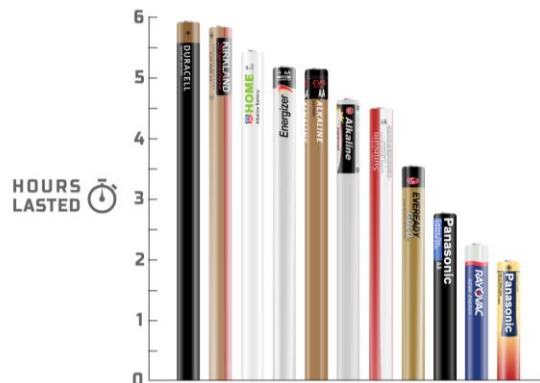
<https://t.co/dMvGp5R8VH>

37

Proper scaling of areas/pictograms

WHICH BATTERIES LAST LONGEST?

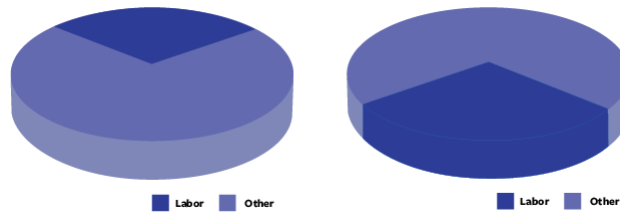
11 different brands of AA batteries, tested in identical flashlights.



https://www.reddit.com/r/dataisbeautiful/comments/855y7m/11_different_brands_of_aa_batteries_tested_in/

38

Unnecessary 3-D



<http://nautil.us/issue/19/illusions/five-ways-to-lie-with-charts>

39

Unnecessary 3-D

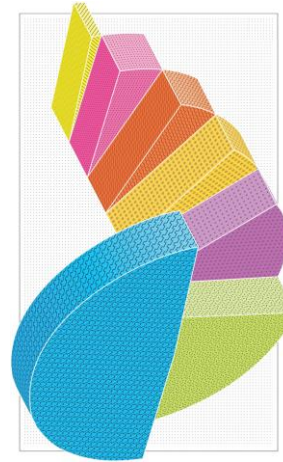


40

Unnecessary 3-D

ANATOMY OF A WINNING TED TALK

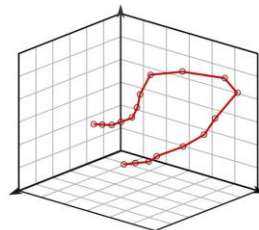
- **1%**
Specialized Visual Aids
We're not sure who gets the D in TED—most of the best presentations have used PowerPoint slide shows (only Steve Jobs). Pulmonary-artery diameter (only Simon Sinek), or no props at all.
- **5%**
Opening jokes
Remember the joke about the shoe salesman who went to Africa in the '60s? That's how Benjamin Zander opened his talk—which turned out to be about classical music.
- **5%**
Spontaneous Moment
Don't overprepare. Take the guy in the first row. ("You could light up a stage with the guy's name.") Comments on "laughter" who handles the human brain you brought.
- **5%**
Statement of Other Certainty
People come for answers—give 'em what they want, as Steven Adler did. "By taking your brain, we can reverse the formula for happiness and success."
- **12%**
Shaggy Dogfish
The TED equivalent of "I have a dream." Example: "People don't think you do, they buy why you do it." Richard P.
- **23%**
Personal Failure
Be vulnerable. You want to know about that tennis bracelet? Or at least the time you didn't fit in at summer camp.
- **49%**
Conclusion Thesis
Wait a sec—we should be playing more videogames! The most obvious way to have the secret of life isn't TED; it's where conventional wisdom goes to die.



<https://www.wired.com/2013/04/tedtalk/>

41

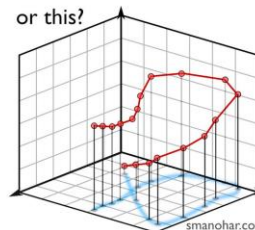
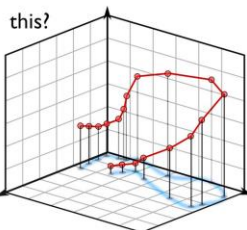
When 3-D is necessary



Please never do this.

3D plots are ambiguous without a projection.
Each point has a whole line of possible 3D locations.

Do you mean...



smanohar.com

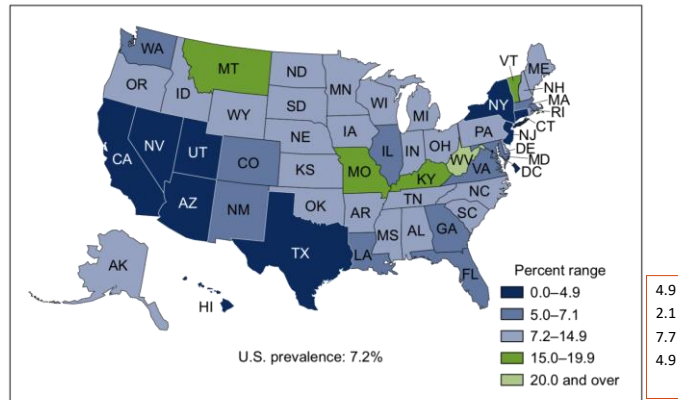
<https://twitter.com/BrainInTheMind/status/1517543839833243649?s=20&t=etXt5ldl3Q6hxVQqJiaMA>

42

Improper categorization

Figure 1. Prevalence of maternal smoking at any time during pregnancy, by state: United States, 2016

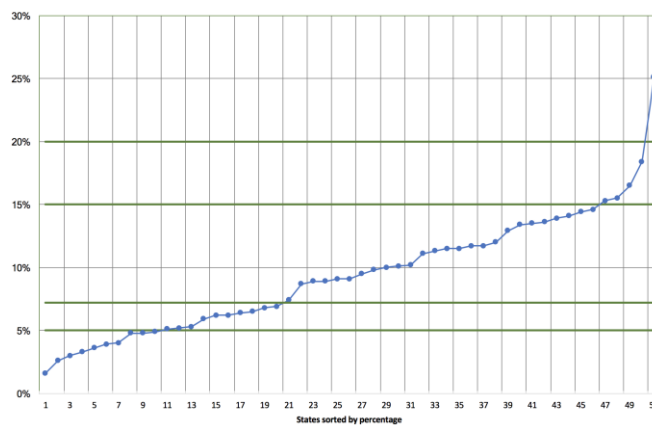
Also bad choice of color!



NOTE: Access data table for Figure 1 at: https://www.cdc.gov/nchs/data/databriefs/db305_table.pdf#1.
SOURCE: NCHS National Vital Statistics System, Natality.

<https://www.cdc.gov/nchs/data/databriefs/db305.pdf>

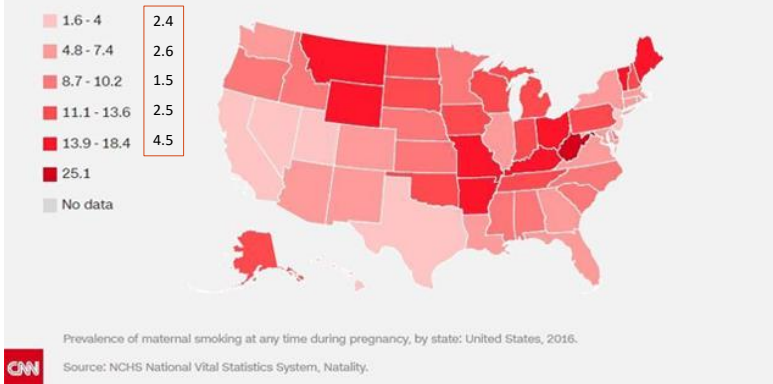
Improper categorization



https://www.cdc.gov/nchs/data/databriefs/db305_table.pdf#1

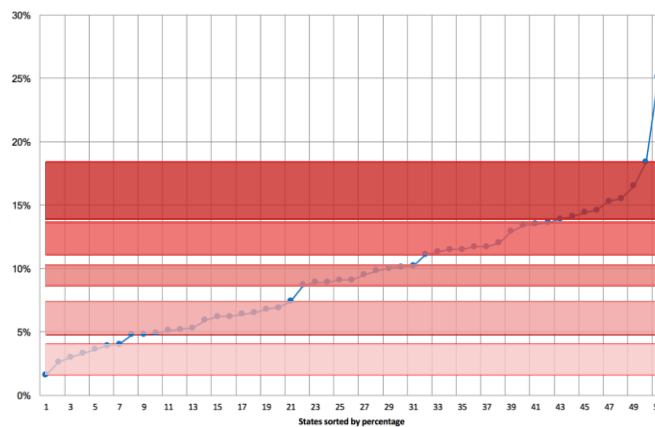
Improper categorization

Percentage of US women who smoke while pregnant



https://meredith.images.worldnow.com/images/16207225_G.png?auto=webp&disable=upscale&width=800

Improper categorization

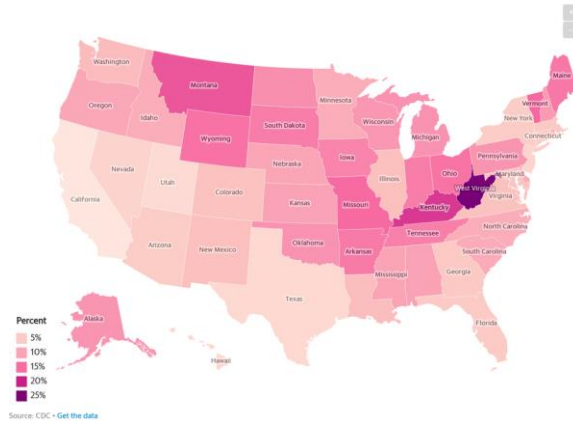


https://www.cdc.gov/nchs/data/databriefs/db305_table.pdf#1

Improper categorization

SMOKING DURING PREGNANCY

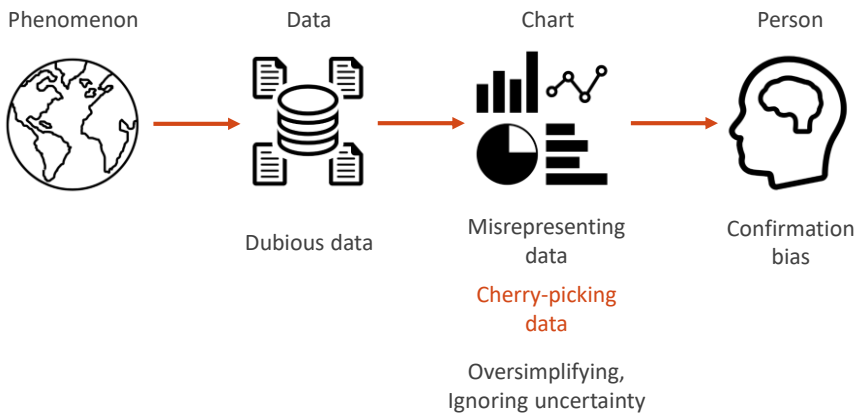
Percentage of women who smoked during pregnancy, 2016 ...



https://www.reddit.com/r/dataisbeautiful/comments/9svpwd/smoking_during_pregnancy_by_us_state_oc/

47

How charts lie?



48

Cherry-picking data

A chart shows as much as it hides, so think about what might be missing

- Hiding (unfavorable) data
- Concealing existing patterns
- Simpson's paradox
- Suggesting patterns that are not there

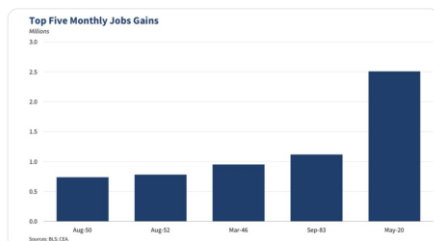
Correlation \neq causation

Hiding (unfavorable) data



Donald J. Trump
@realDonaldTrump

Greatest Top Five Monthly Jobs Gains in HISTORY. We are #1!



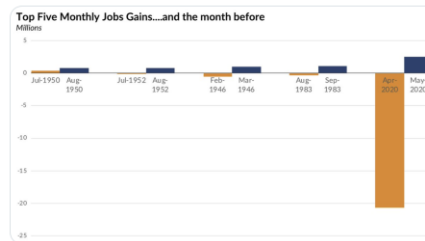
12:31 PM · Jun 5, 2020 · Twitter for iPhone

21.8K Retweets 72K Likes

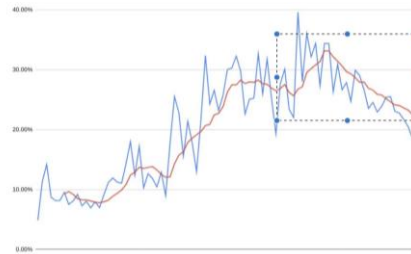


Jon Schwabish
@jschwabish · 4h

I fixed this graph for you Mr. President. So, you know, it's not a complete misrepresentation of the facts..



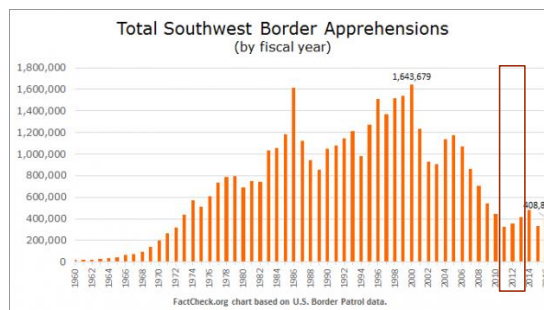
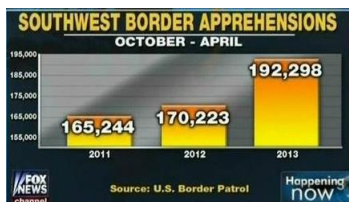
Hiding (unfavorable) data



<https://twitter.com/medel2020/status/1280676453768736768>

51

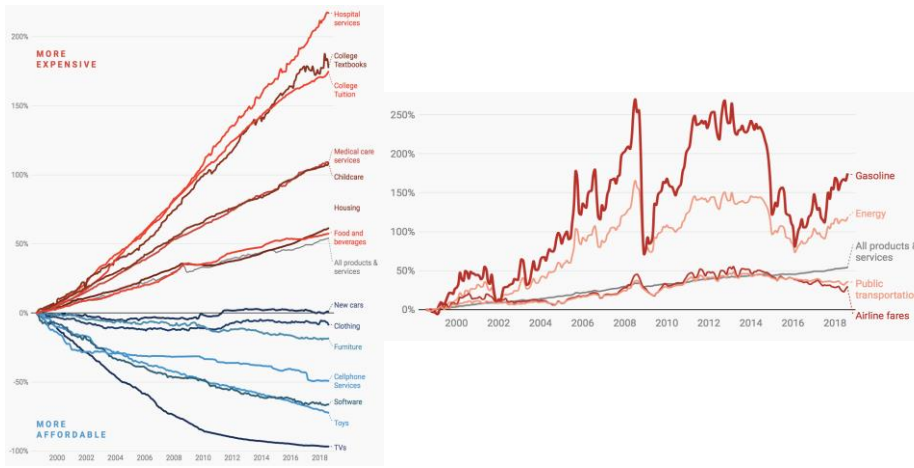
Concealing existing patterns



<https://www.businessinsider.com/the-27-worst-charts-of-all-time-2013-6#welcome-to-fox-where-the-line-graphs-are-made-up-and-the-points-dont-matter-12>

52

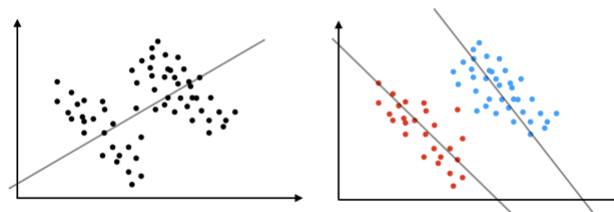
Concealing existing patterns



<https://blog.datawrapper.de/weekly47-cpi-dollars-for-college/>

53

Simpson's paradox



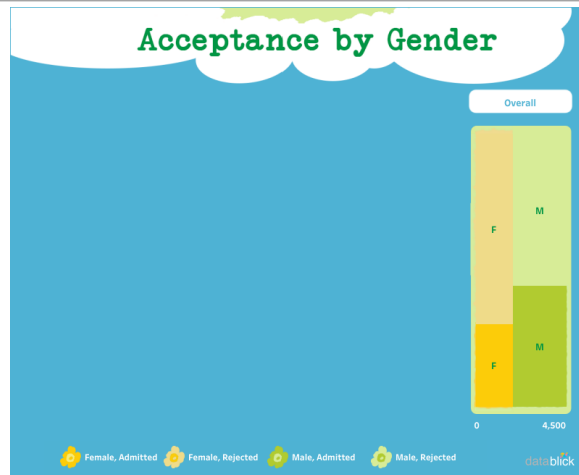
The overall trend reverses when data is grouped by categories

<https://towardsdatascience.com/simpsons-paradox-and-interpreting-data-6a0443516765>

54

Simpson's paradox

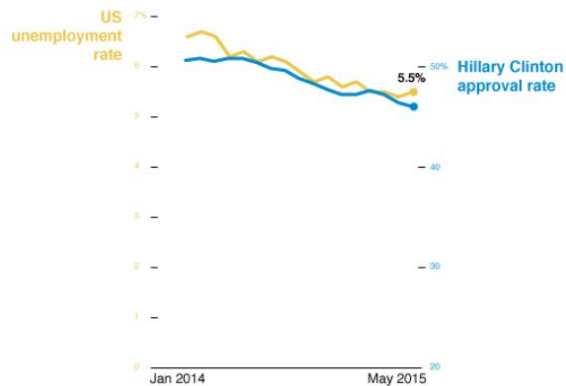
Admission of applicants to UC Berkeley



<https://public.tableau.com/app/profile/jonathan.drummey/viz/marimekko-mosaicplot/MarimekkoGT>

55

Suggesting patterns that are not there

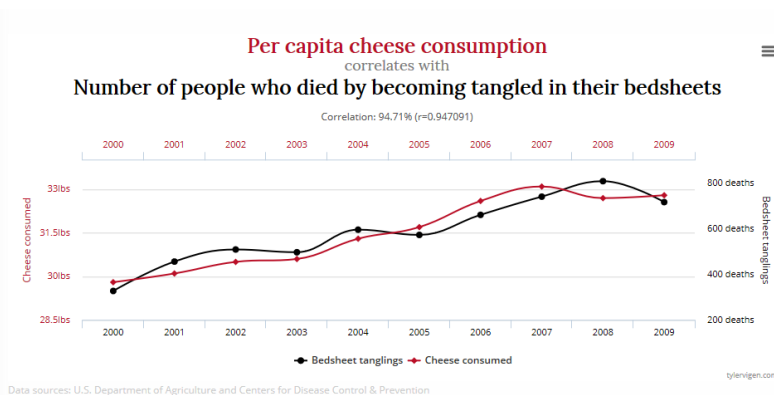


<https://news.nationalgeographic.com/2015/06/150619-data-points-five-ways-to-lie-with-charts/>

56

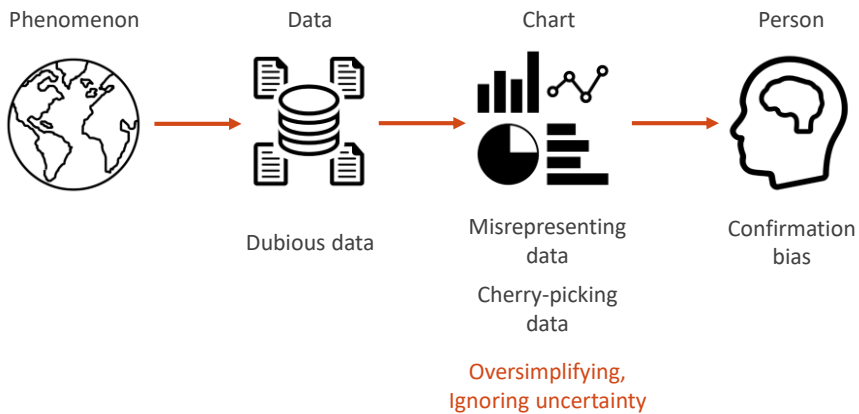
Suggesting patterns that are not there

Spurious correlations: <http://www.tylervigen.com/spurious-correlations>



57

How charts lie?



58

Oversimplifying, Ignoring uncertainty

- Oversimplifying
- Misrepresenting uncertainty
- Concealing uncertainty

59

Oversimplifying

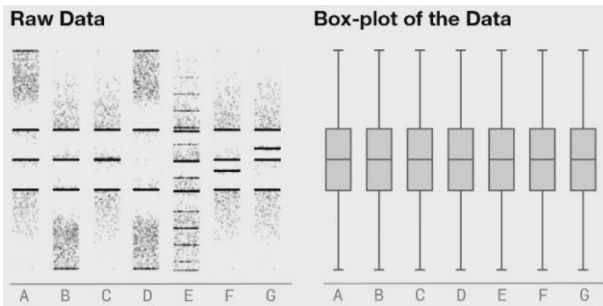
Clarify, not (over)simplify!

To clarify, add detail.

Edward Tufte

60

Oversimplifying

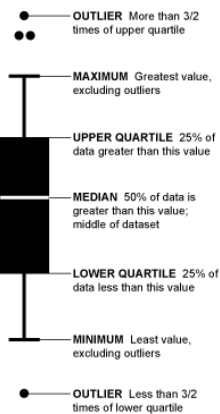


<https://www.autodeskresearch.com/publications/samestats>

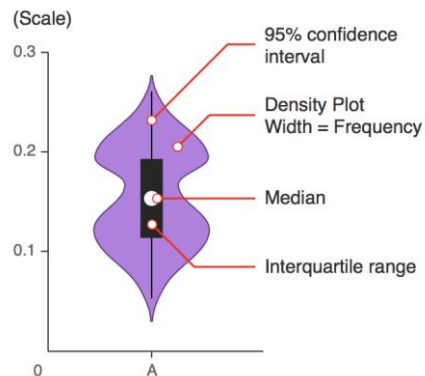
61

Box plot vs. violin plot

Box (and whisker) plot



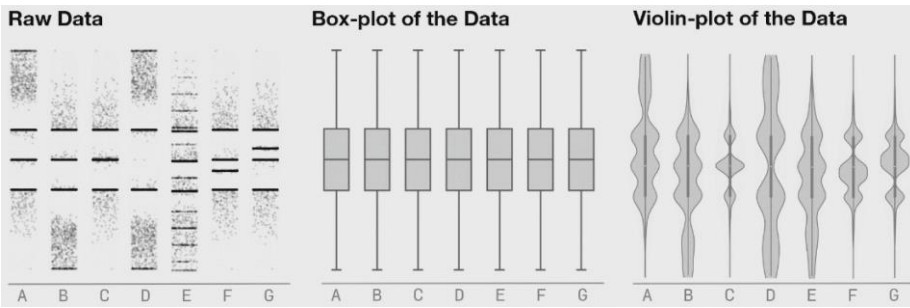
Violin plot



<https://flowingdata.com/2008/02/15/how-to-read-and-use-a-box-and-whisker-plot/>
https://datavizcatalogue.com/methods/violin_plot.html

62

Oversimplifying

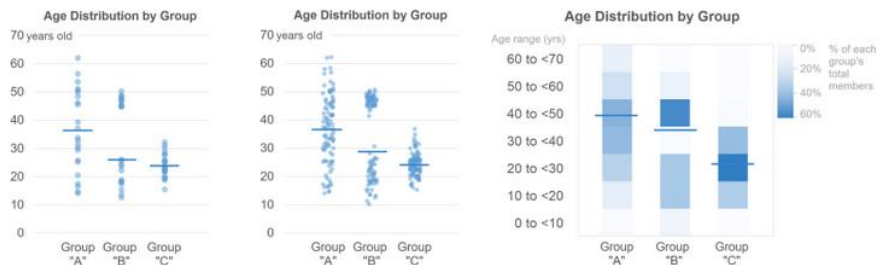


<https://www.autodeskresearch.com/publications/samestats>

63

Oversimplifying

Other alternatives to boxplots

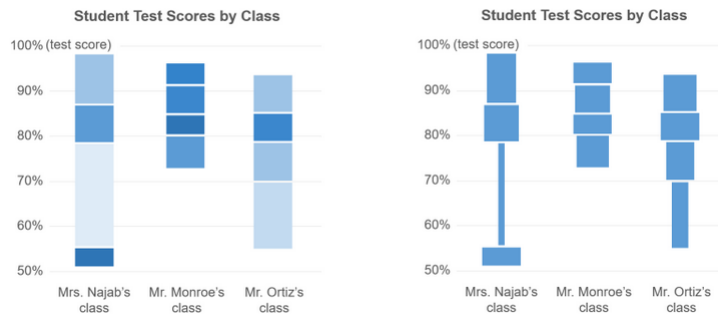


<https://nightingaledvs.com/ive-stopped-using-box-plots-should-you/>

64

Oversimplifying

Other alternatives to boxplots



<https://nightingaledvs.com/ive-stopped-using-box-plots-should-you/>

65

Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted



https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

66

Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted



https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

67

Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted



https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

68

Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted



https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

69

Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted



https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

70

Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted



https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

71

Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted

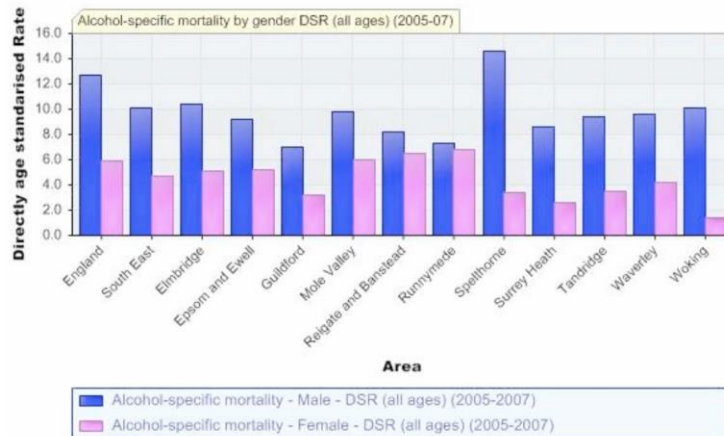


1/3 of the time,
the path of the eye of
the storm will go
either way beyond the
boundaries of the cone

https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

72

Concealing uncertainty

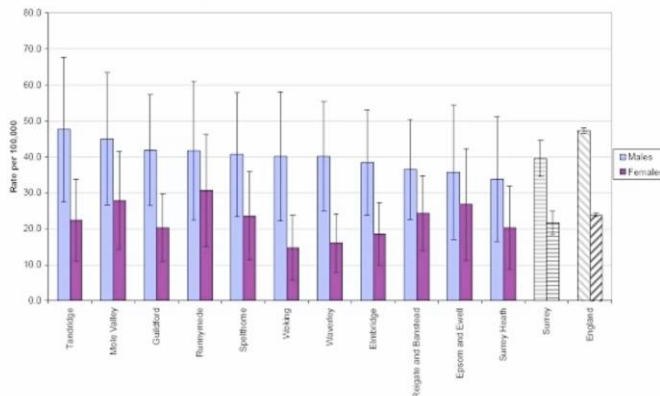


<https://itunes.apple.com/us/course/data-literacy-and-data-visualization/id693097601>

73

Concealing uncertainty

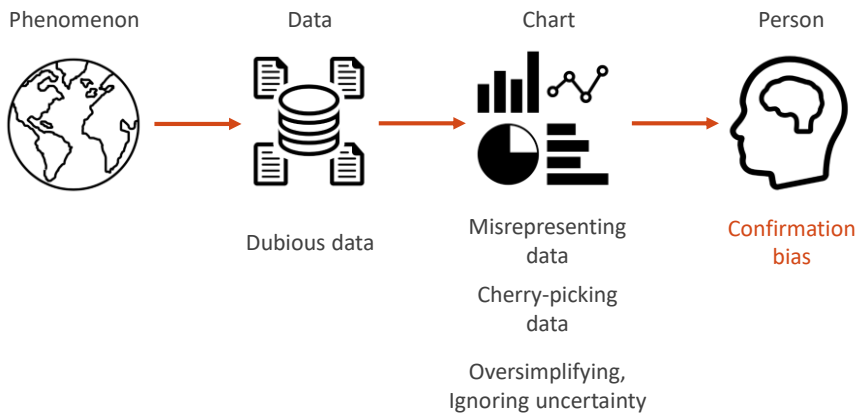
Directly age-standardised mortality from alcohol attributable conditions for men and women by borough in Surrey, rate per 100,000 people (2005/06).



<https://itunes.apple.com/us/course/data-literacy-and-data-visualization/id693097601>

74

How charts lie?



75

Confirmation bias

Confirmation bias is the tendency to search for, interpret, favor, and recall information in a way that confirms or supports one's prior beliefs or values

Charts lie because we lie to ourselves – we see what we want to see

The bias blind spot

- We think only others are biased
- This makes us more susceptible to bias

Confirmation bias does not affect only chart interpretation, but also visualization design, data analysis and even data collection

76

Confirmation bias



<https://twitter.com/TreyYingst/status/862669407868391424>

77

Confirmation bias

Donald J. Trump @realDonaldTrump · Feb 27
I want to encourage all of my many Texas friends to vote in the primary for Governor Greg Abbott, Senator Ted Cruz, Lt. Gov. Dan Patrick, and Attorney General Ken Paxton. They are helping me to Make America Great Again! Vote early or on March 6th.

20K 27K 99K

Trevor @trevorHartje · Feb 27
You have no friends

553 223 4.9K

Nancy Steffan @steffan_nancy

Follow

Replying to @trevorHartje @realDonaldTrump

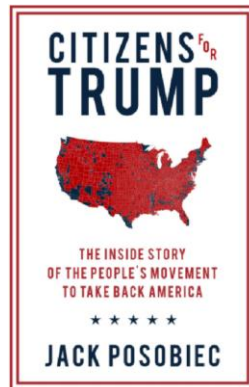
Really?



https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

78

Confirmation bias



https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

79

Confirmation bias



Surface on the
county-level map:
Red: 80%
Blue: 20%

https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

80

Confirmation bias



Surface on the
county-level map:
Red: 80%
Blue: 20%

SHARE OF THE POPULAR VOTE IN THE 2016 PRESIDENTIAL ELECTION

| | | | |
|------------------|---|-------|------------------|
| Donald Trump |  | 46.1% | 62,984,825 votes |
| Hillary Clinton |  | 48.2% | 65,853,516 votes |
| Other candidates |  | 5.7% | |

https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

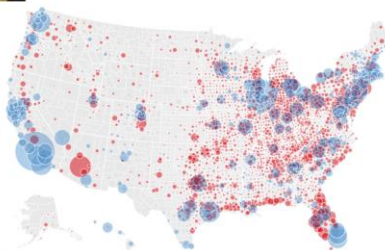
81

Confirmation bias



Surface on the
county-level map:
Red: 80%
Blue: 20%

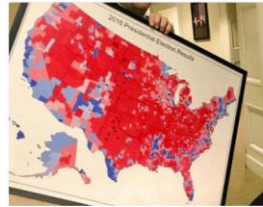
Bubble size is proportional
to the number of votes
received just by the candidate
who won on each county



https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

82

Confirmation bias



SHARE OF THE POPULAR VOTE IN THE 2016 PRESIDENTIAL ELECTION

| | | |
|------------------|-------|------------------|
| Donald Trump | 46.1% | 62,994,825 votes |
| Hillary Clinton | 48.2% | 65,853,516 votes |
| Other candidates | 5.7% | |

PERCENTAGE OF ELIGIBLE VOTERS

| | |
|----------------------------|-------|
| Didn't vote | 40.0% |
| Voted for Donald Trump | 27.7% |
| Voted for Hillary Clinton | 28.9% |
| Voted for other candidates | 3.4% |

VOTES FOR DONALD TRUMP



VOTES FOR HILLARY CLINTON



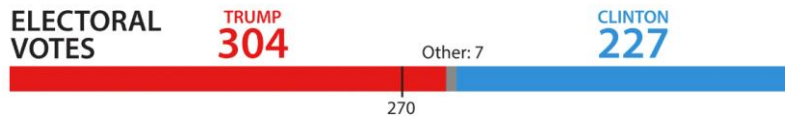
Bubble size is proportional to the number of votes per county

https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

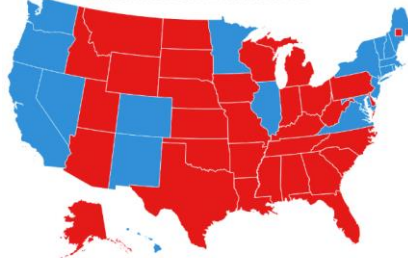
83

Confirmation bias

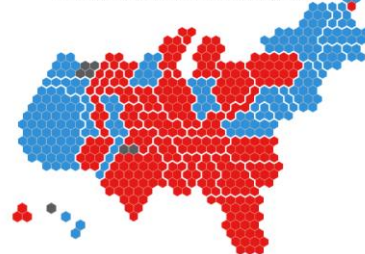
These are the numbers that truly matter in a U.S. Presidential Election



WHO WON ON EACH STATE



STATE SIZE ADJUSTED BY ELECTORAL VOTES IT CONTRIBUTES TO THE ELECTION



https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

84

To achieve trustworthiness (1)

- List the source(s) of data
- Show representative and unbiased data (or clearly denote and explain why this is not the case)
- Compare only data that can be meaningfully compared
- Be mindful of the choice between absolute and cumulative values
- Use relative instead of absolute data in comparisons
- Follow conventions
- Do not abuse scales

85

To achieve trustworthiness (2)

- Do not use 3-D representations for non 3-D data
- Choose categories mindfully
- Do not oversimplify
- Present the entire relevant data
- Do not suggest patterns that are not there
- Show uncertainty
- Be wary of confirmation bias

However... some rules can be bent (as long as you know what you are doing)

86