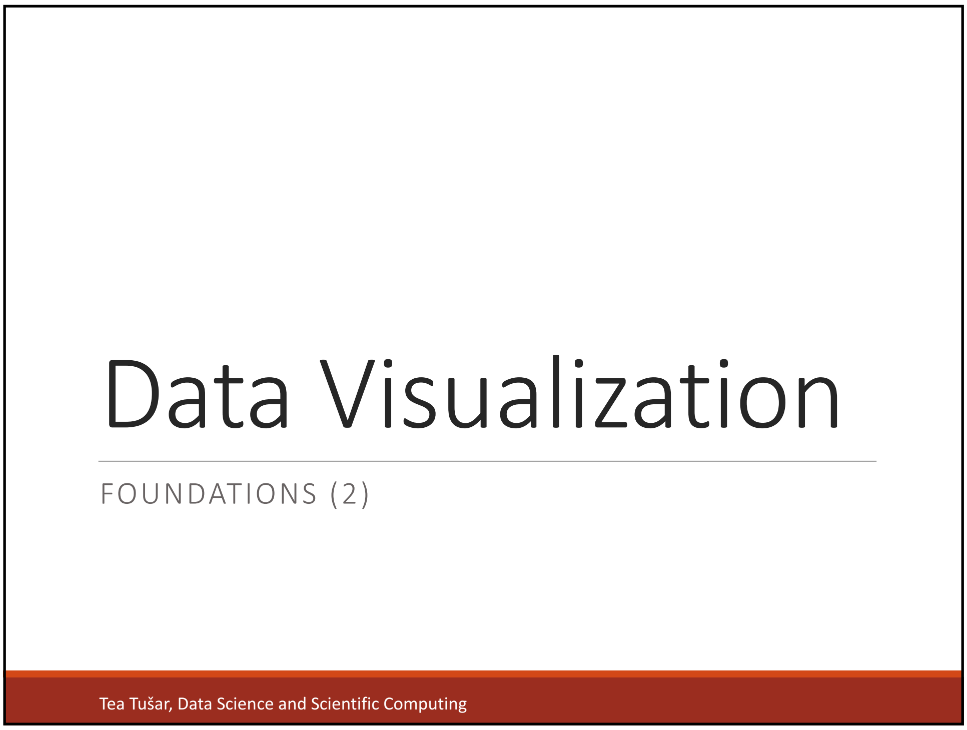
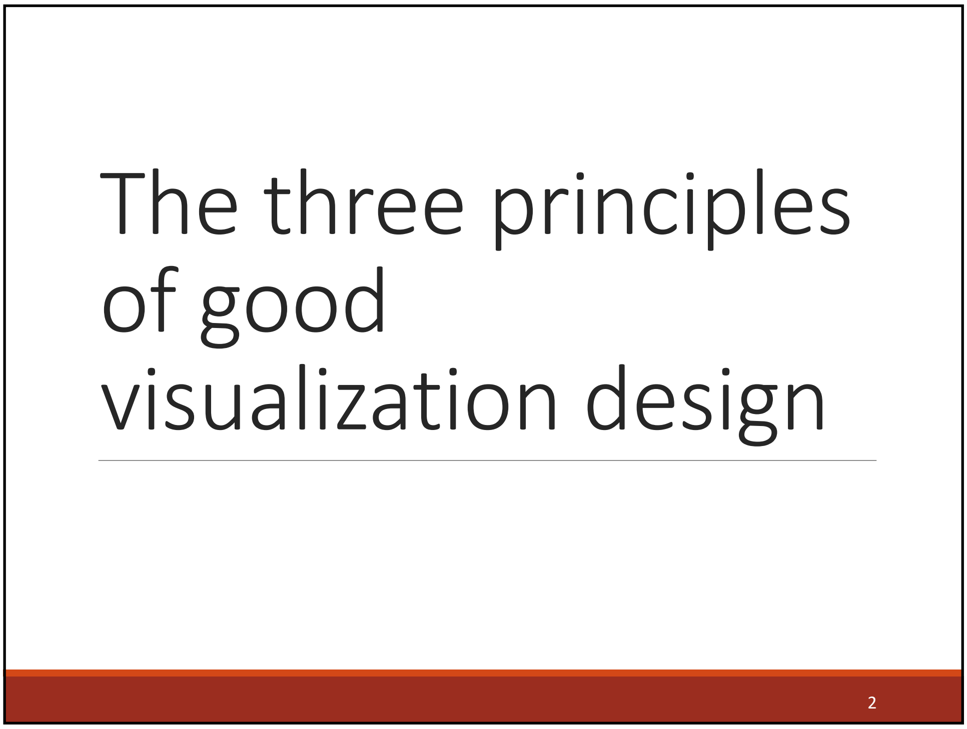


Data Visualization

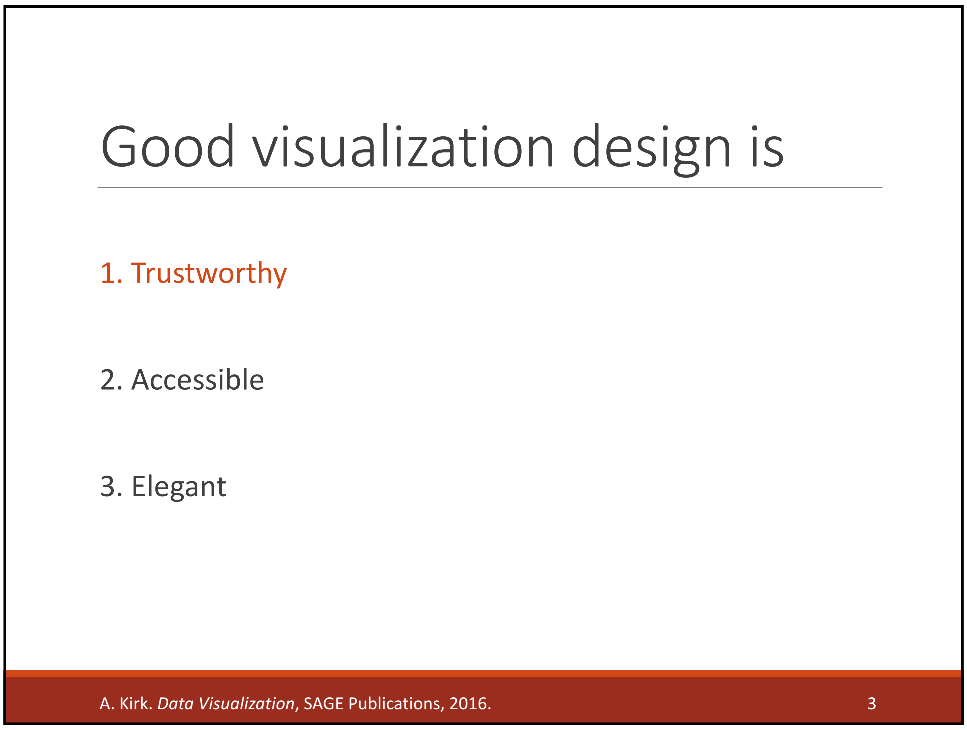
FOUNDATIONS (2)

Tea Tušar, Data Science and Scientific Computing



The three principles of good visualization design

2



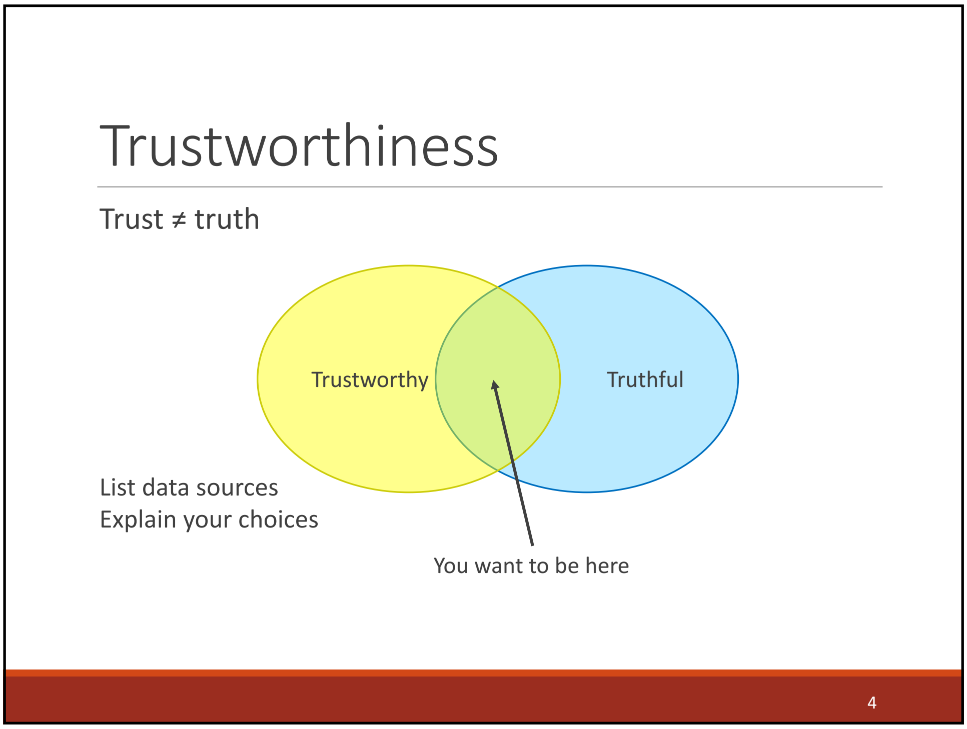
Good visualization design is

1. Trustworthy

2. Accessible

3. Elegant

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

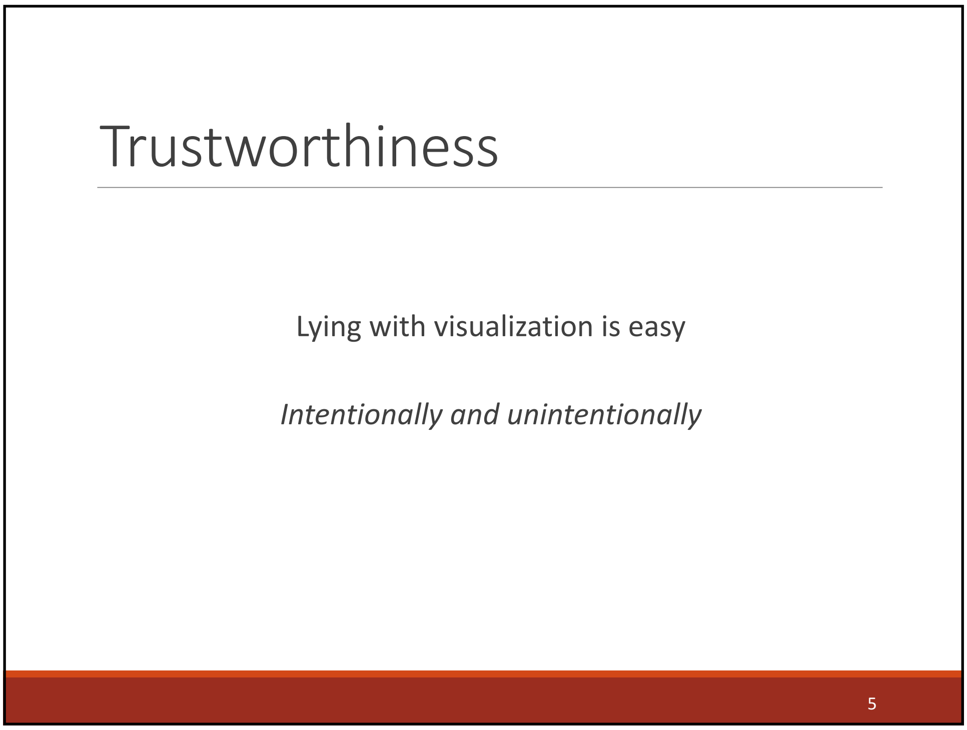


Trustworthiness

Trust ≠ truth

List data sources
Explain your choices

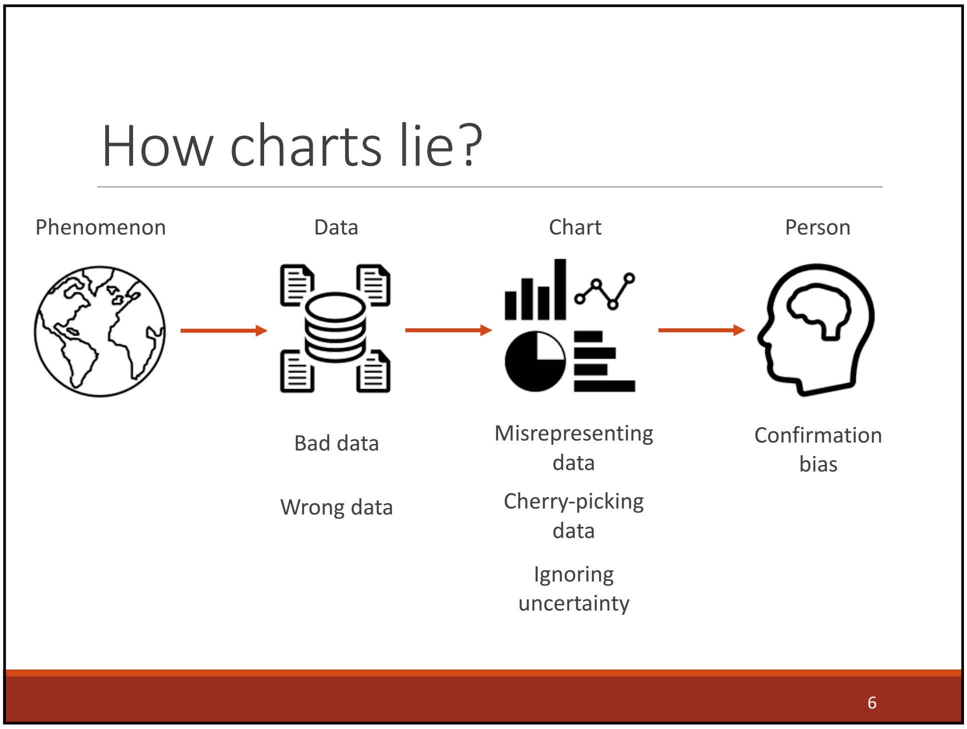
You want to be here



Trustworthiness

Lying with visualization is easy

Intentionally and unintentionally



How charts lie?

Bad data

Wrong data

Misrepresenting data

Cherry-picking data

Ignoring uncertainty

Confirmation bias

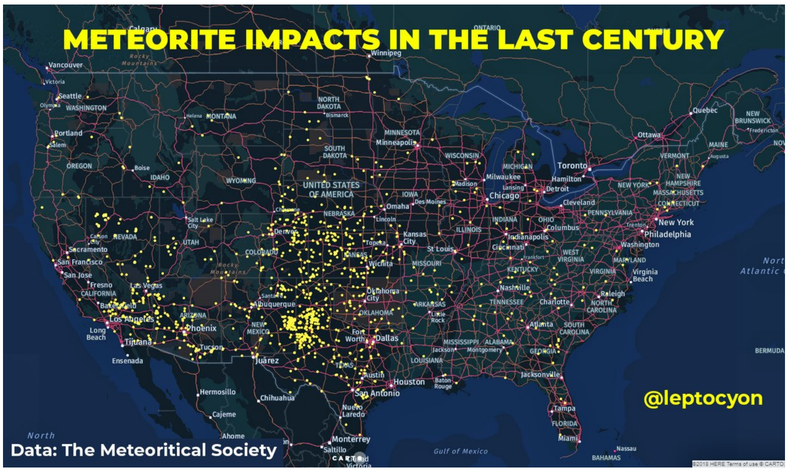
Bad data

Garbage in, garbage out

- Unrepresentative data
  - Polls on unrepresentative populations
  - Measurements on unrepresentative samples
  - Too much missing data
- Biased data
  - Question framing in polls
  - Choice of measures

This is a problem when it is not made clear and the data is used for analyses that are suitable for more ‘regular’ data

Unrepresentative data

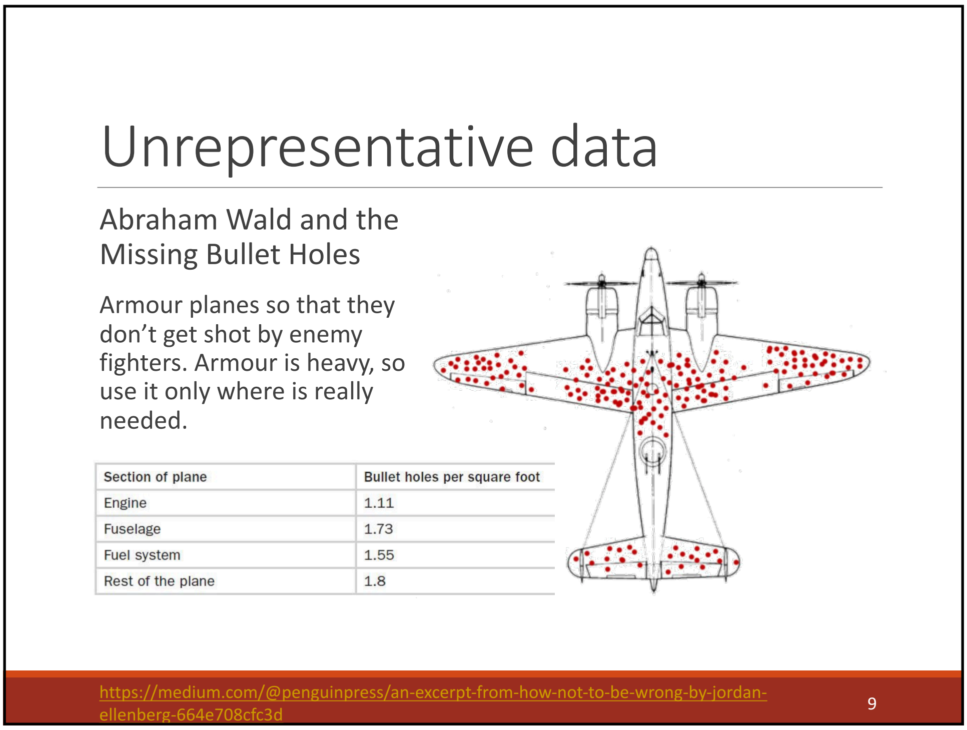


METEORITE IMPACTS IN THE LAST CENTURY

@leptocyon

Data: The Meteoritical Society

https://www.reddit.com/r/dataisbeautiful/comments/9pkka4/all_recorded_meteorite_impacts_in_the_us_from



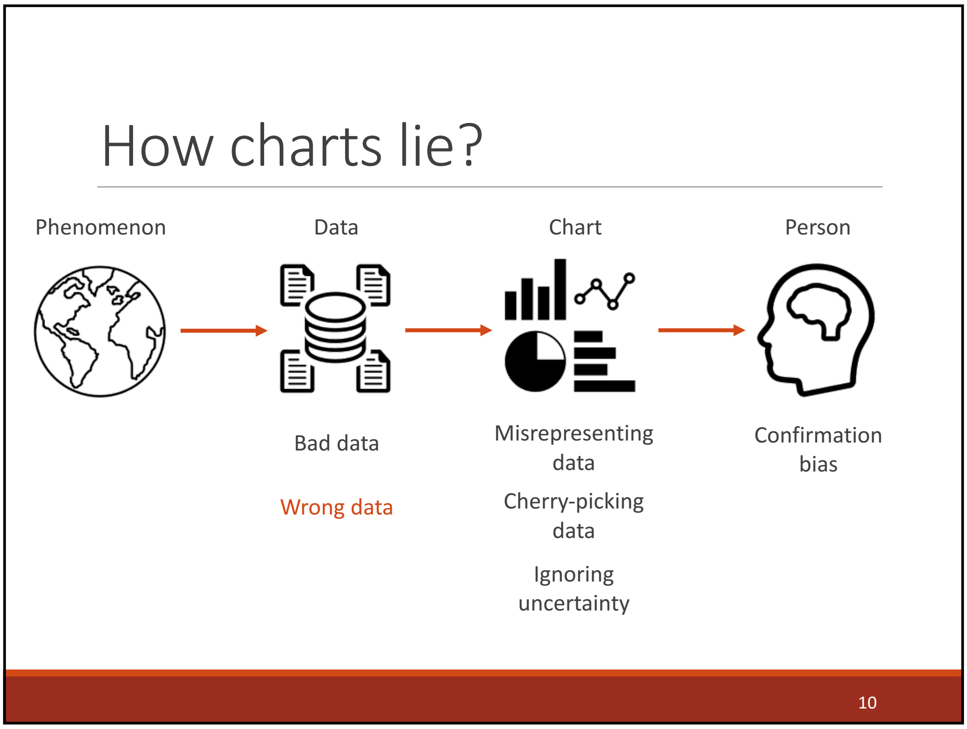
Unrepresentative data

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



How charts lie?

Phenomenon

Data

Chart

Person

Bad data

Wrong data

Misrepresenting data

Cherry-picking data

Ignoring uncertainty

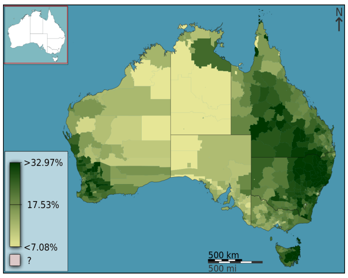
Confirmation bias

Wrong data

Comparisons using

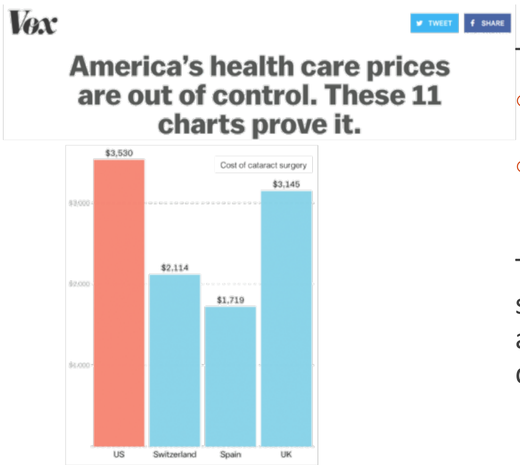
- Non-comparable data
- Absolute instead of cumulative data (and vice versa)
- Absolute instead of relative data (in charts and choropleth maps)

The fraction of Australians that identified as Anglican at the 2011 census



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

Non-comparable data used in comparisons



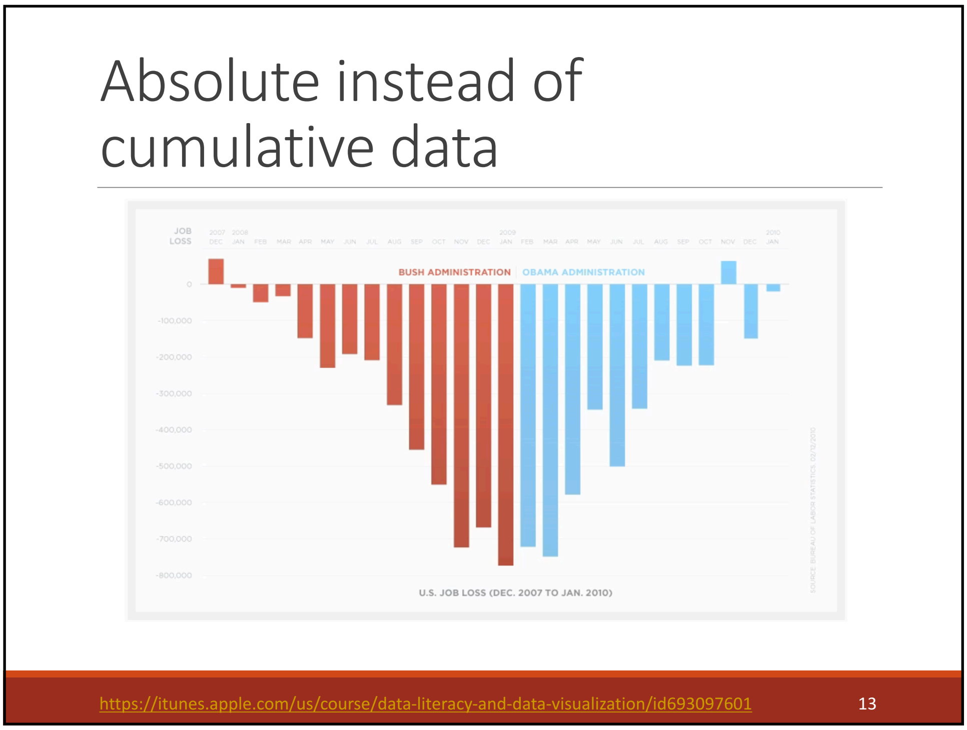
America's health care prices are out of control. These 11 charts prove it.

Two issues

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

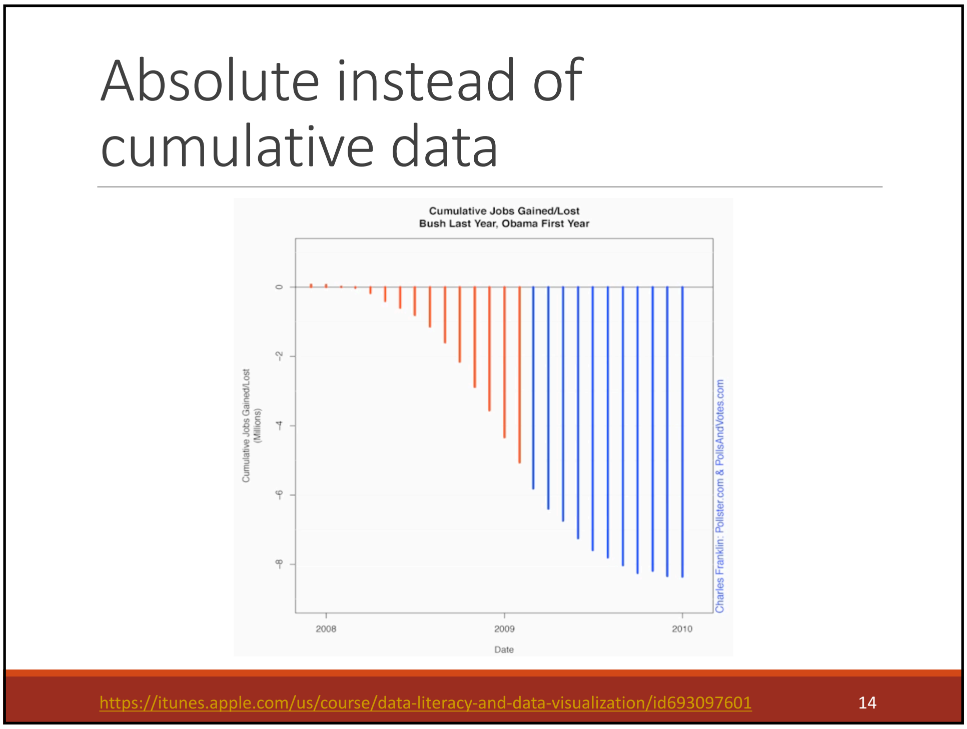
The data source specifically warns against using this data for comparison

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



Absolute instead of cumulative data

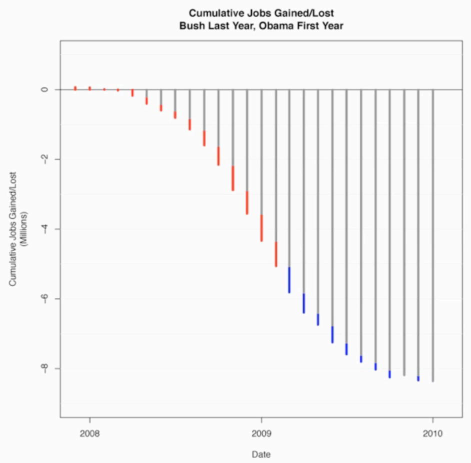
https://itunes.apple.com/us/course/data-literacy-and-data-visualization/id693097601 13



Absolute instead of cumulative data

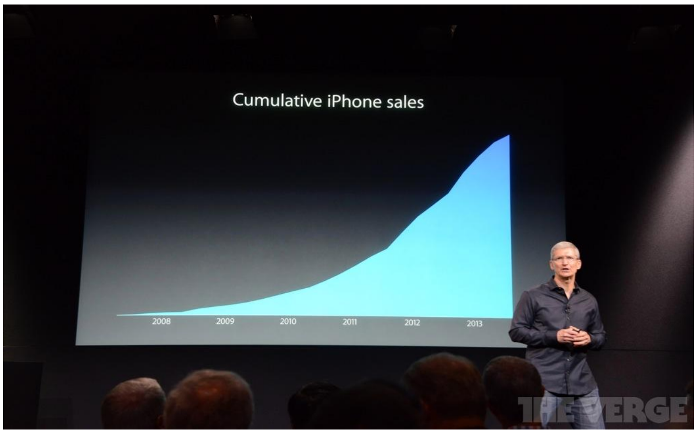
https://itunes.apple.com/us/course/data-literacy-and-data-visualization/id693097601 14

Absolute instead of cumulative data



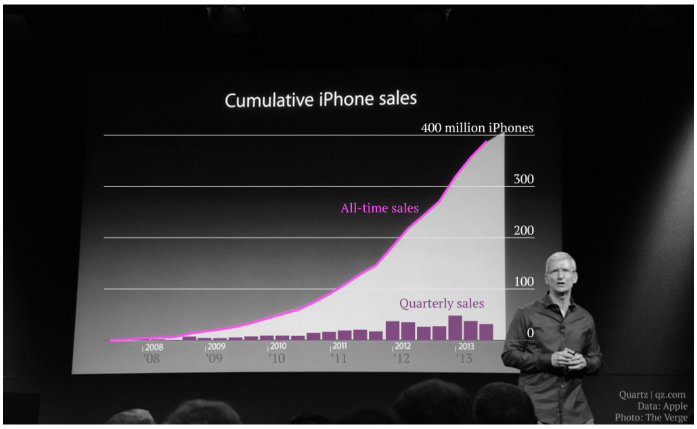
https://itunes.apple.com/us/course/data-literacy-and-data-visualization/id693097601

Cumulative instead of absolute data



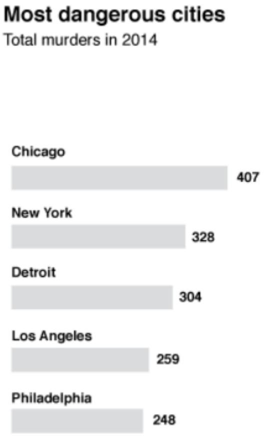
https://qz.com/122921/the-chart-tim-cook-doesnt-want-you-to-see/

Cumulative instead of absolute data



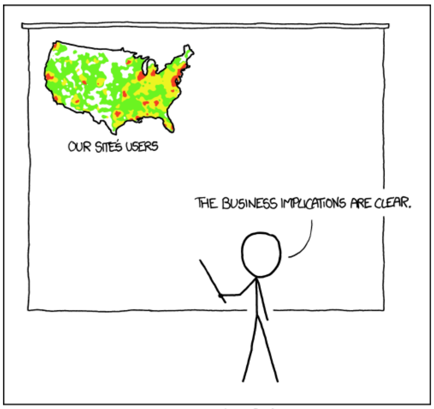
https://qz.com/122921/the-chart-tim-cook-doesnt-want-you-to-see/

Absolute data in comparisons



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

Absolute data in choropleth maps

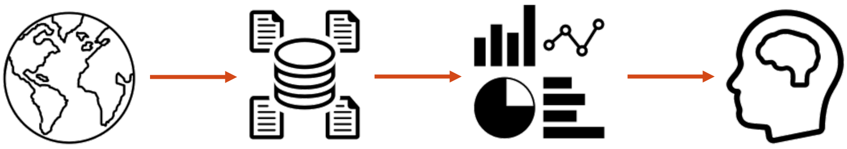


PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE
BASICALLY JUST POPULATION MAPS

https://xkcd.com/1138/

How charts lie?

Phenomenon → Data → Chart → Person



Data: Bad data

Wrong data

Chart: Misrepresenting data

Cherry-picking data

Ignoring uncertainty

Person: Confirmation bias

Misrepresenting data

Ignoring conventions
- Placement of dependent and independent variables
- Inverted y axis
- Unequal intervals
- Pie charts that do not add up to 100%

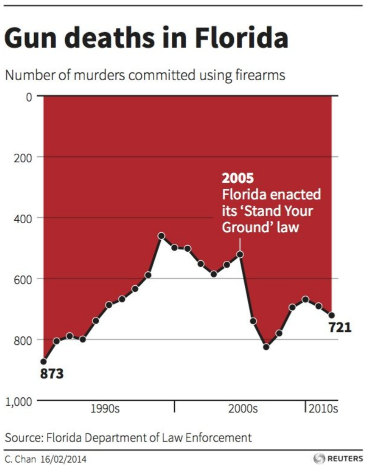
Abusing scales
- Bar charts with truncated axis
- Aspect ratio bias
- Dual axes
- Improper scaling of areas (and pictograms)

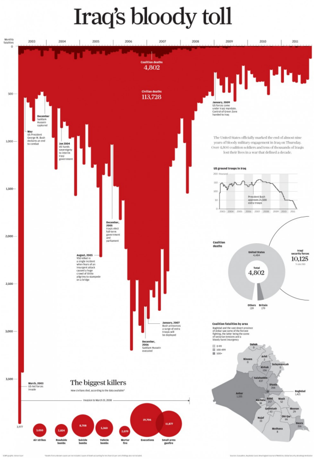
Unnecessary 3-D

Improper categorization

Oversimplifying

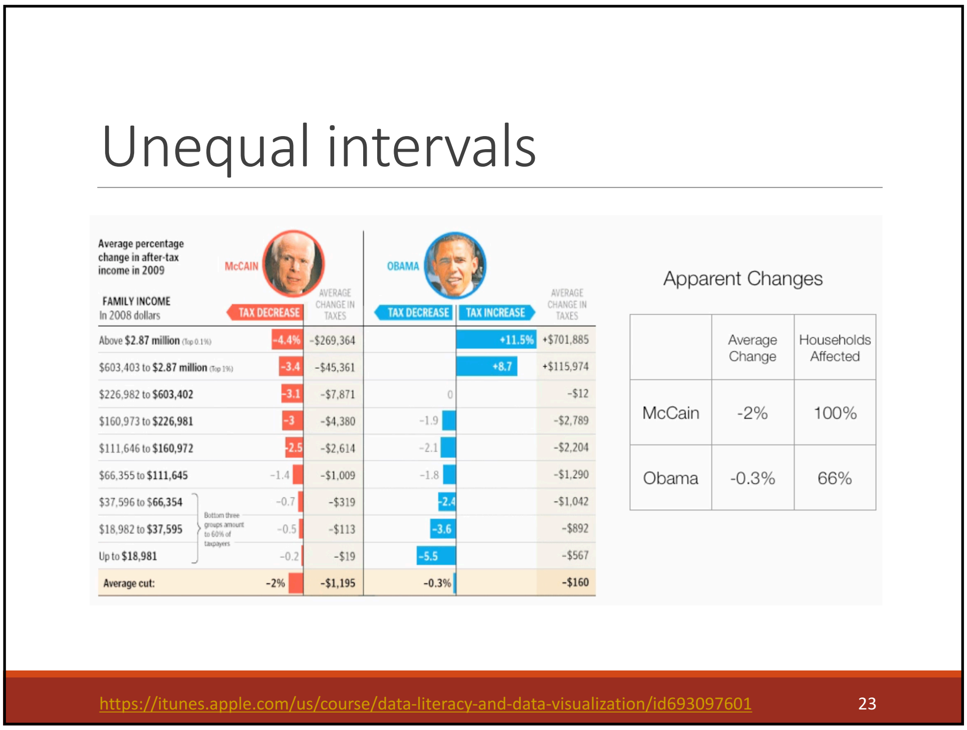
Inverted y axis





The United States officially marked the end of almost nine years of bloody military engagement in Iraq on Thursday. Over 4,800 coalition soldiers and tens of thousands of Iraqis lost their lives in a war that defined a decade.

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



Unequal intervals

Average percentage change in after-tax income in 2009

FAMILY INCOME In 2008 dollars	McCAIN TAX DECREASE	AVERAGE CHANGE IN TAXES	OBAMA TAX DECREASE	OBAMA TAX INCREASE	AVERAGE CHANGE IN TAXES
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**

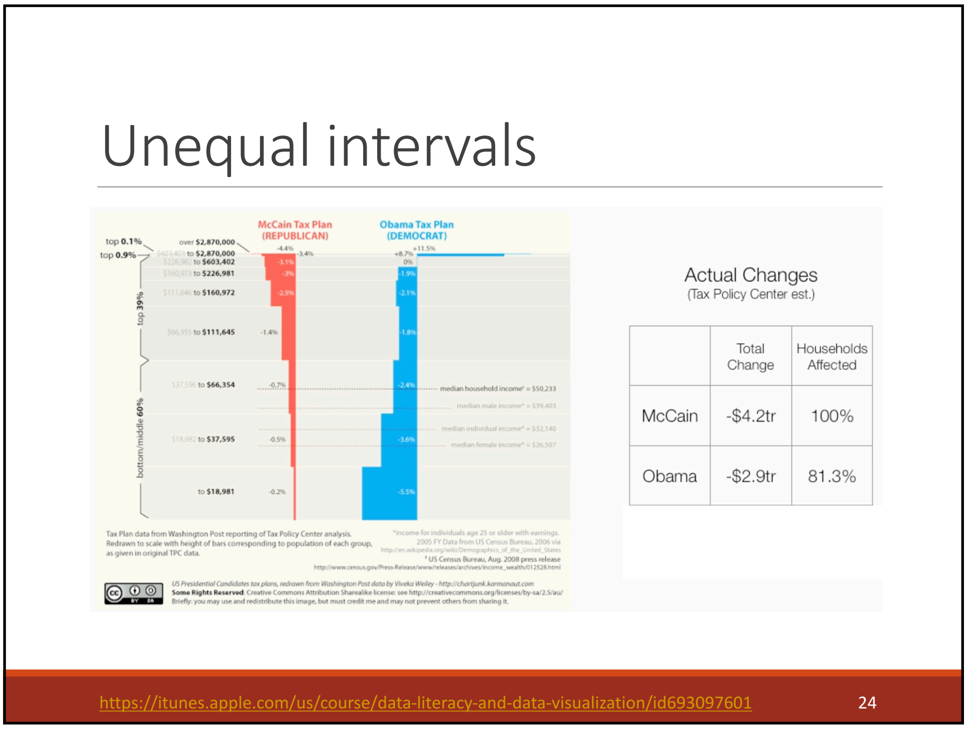
Bottom three groups amount to 60% of taxpayers

Apparent Changes

	Average Change	Households Affected
McCain	-2%	100%
Obama	-0.3%	66%

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

23



Unequal intervals

Tax Plan data from Washington Post reporting of Tax Policy Center analysis. Redrawn to scale with height of bars corresponding to population of each group, as given in original TPC data.

*income for individuals age 25 or older with earnings. 2005 FY Data from US Census Bureau, 2006 via http://en.wikipedia.org/wiki/Demographics_of_the_United_States

‡ US Census Bureau, Aug. 2008 press release http://www.census.gov/Press-Release/www/releases/archives/income_wealth/012528.html

US Presidential Candidates tax plans, redrawn from Washington Post data by Viveka Weiley - http://chartjunk.karmanaut.com

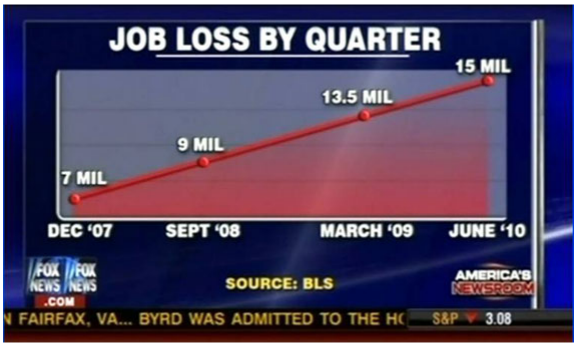
Actual Changes
(Tax Policy Center est.)

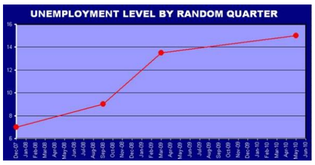
	Total Change	Households Affected
McCain	-$4.2tr	100%
Obama	-$2.9tr	81.3%

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

24

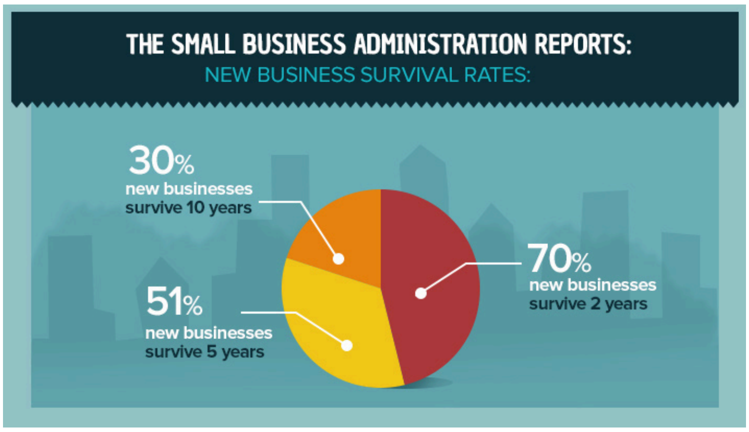
Unequal intervals





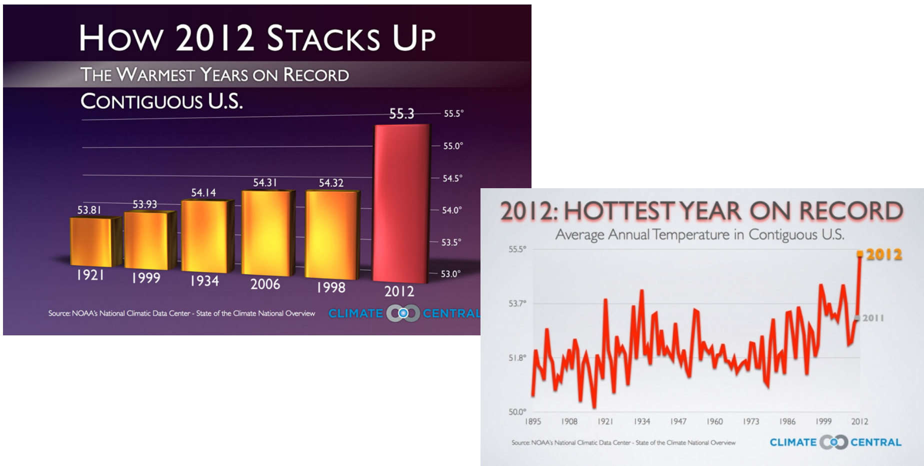
https://www.mediamatters.org/blog/2010/06/28/updated-worst-chart-ive-seen-all-day/166862

Over 100% pie chart



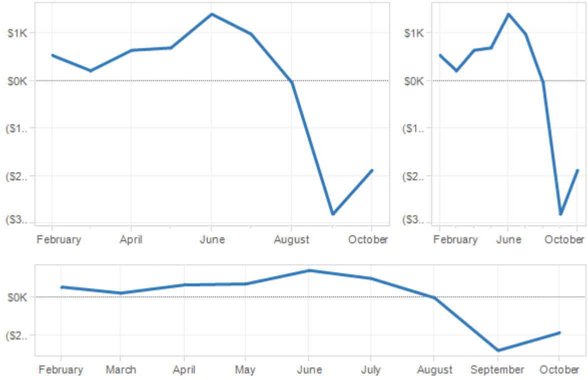
https://www.shinobicontrols.com/blog/6-common-mistakes-with-data-visualization

Bar chart with truncated axis



http://www.climatecentral.org/news/noaa-2012-was-warmest-and-second-most-extreme-year-on-record-15436-second-most-extreme-year-on-record-15436

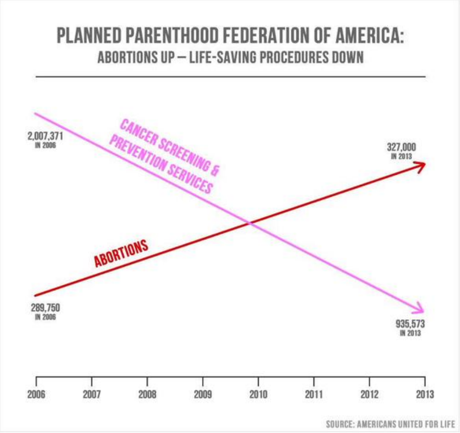
Aspect ratio bias



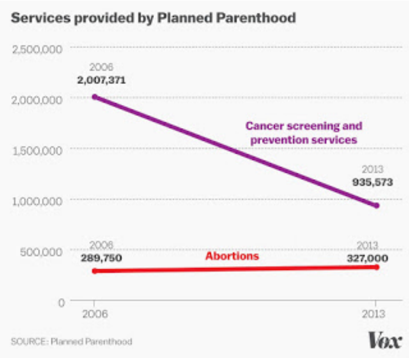
Banking to 45 Degrees

https://eagereyes.org/basics/banking-45-degrees

Dual axes



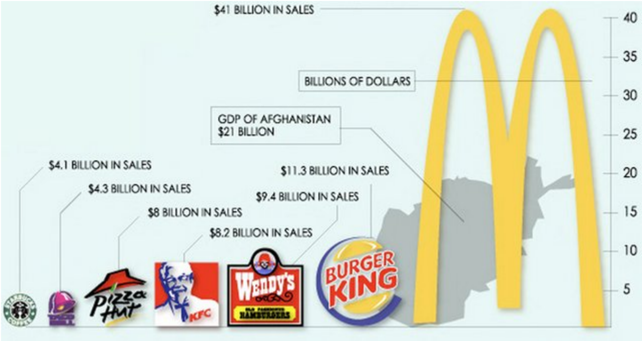
SOURCE: AMERICANS UNITED FOR LIFE

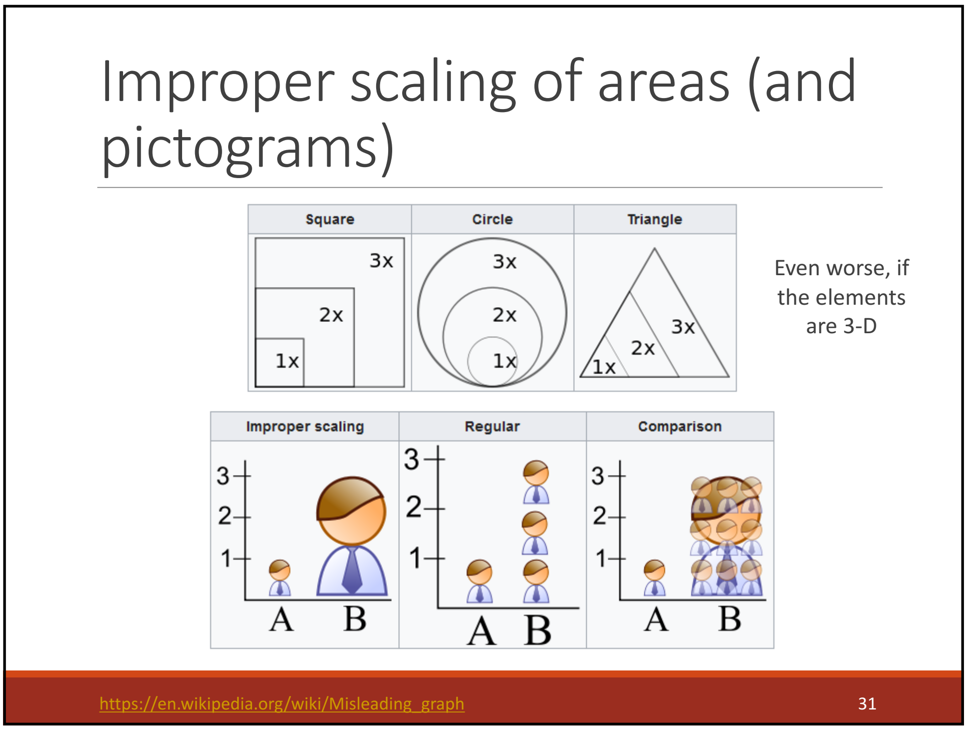


SOURCE: Planned Parenthood

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

Improper scaling of areas (and pictograms)



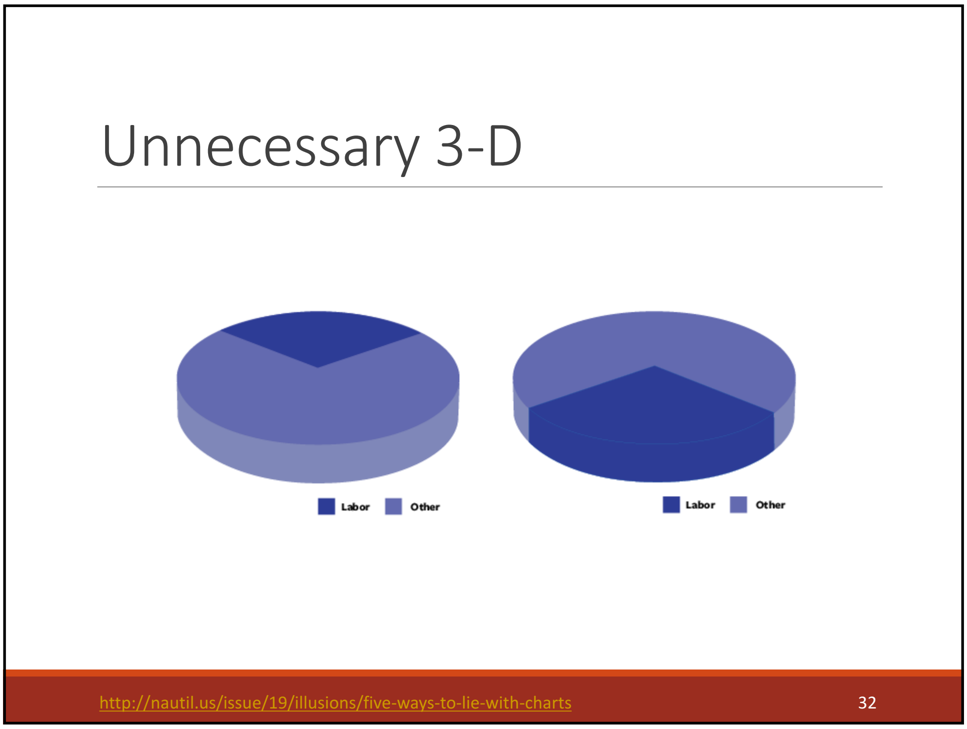


Improper scaling of areas (and pictograms)

Even worse, if the elements are 3-D

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

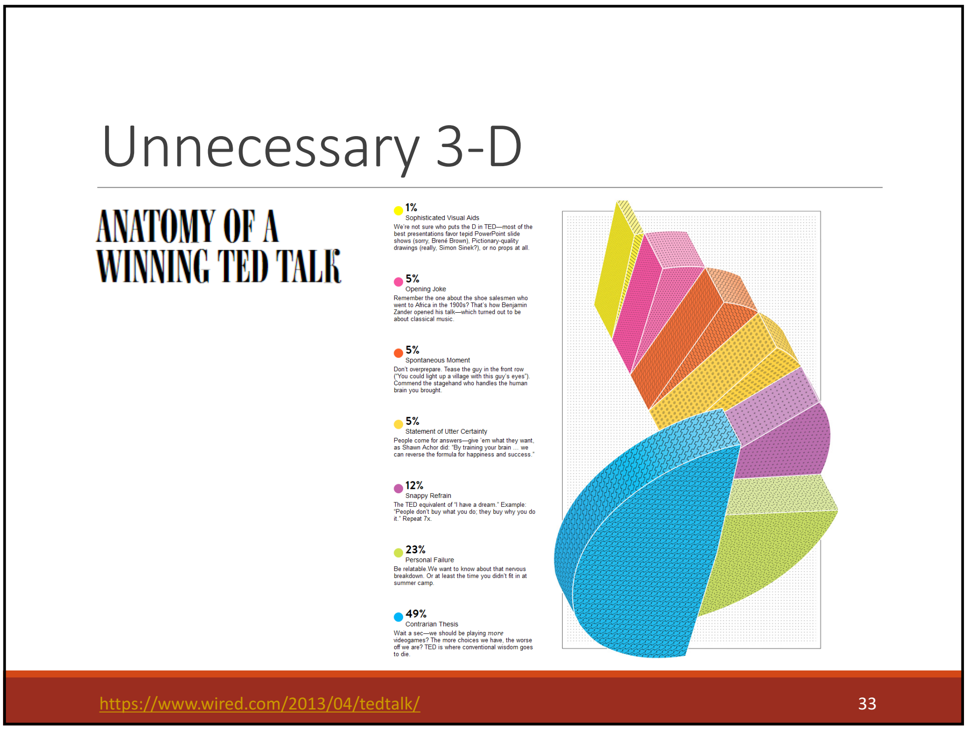
31



Unnecessary 3-D

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

32



Unnecessary 3-D

ANATOMY OF A WINNING TED TALK

● **1%**
Sophisticated Visual Aids
We're not sure who puts the D in TED—most of the best presentations favor tepid PowerPoint slide shows (sorry, Brené Brown), Pictionary-quality drawings (really, Simon Sinek?), or no props at all.

● **5%**
Opening Joke
Remember the one about the shoe salesmen who went to Africa in the 1900s? That's how Benjamin Zander opened his talk—which turned out to be about classical music.

● **5%**
Spontaneous Moment
Don't overprepare. Tease the guy in the front row ("You could light up a village with this guy's eyes"). Commend the stagehand who handles the human brain you brought.

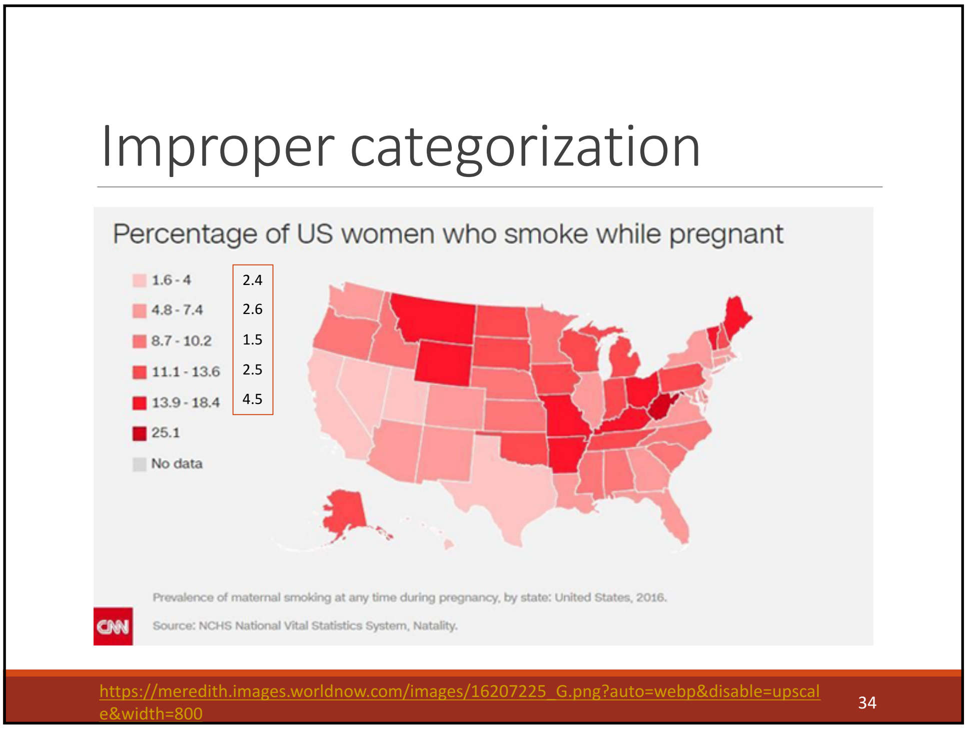
● **5%**
Statement of Utter Certainty
People come for answers—give 'em what they want, as Shawn Achor did: "By training your brain ... we can reverse the formula for happiness and success."

● **12%**
Snappy Refrain
The TED equivalent of "I have a dream." Example: "People don't buy what you do; they buy why you do it." Repeat 7x.

● **23%**
Personal Failure
Be relatable.We want to know about that nervous breakdown. Or at least the time you didn't fit in at summer camp.

● **49%**
Contrarian Thesis
Wait a sec—we should be playing *more* videogames? The more choices we have, the worse off we are? TED is where conventional wisdom goes to die.

https://www.wired.com/2013/04/tedtalk/ 33



Improper categorization

Percentage of US women who smoke while pregnant

1.6 - 4	2.4
4.8 - 7.4	2.6
8.7 - 10.2	1.5
11.1 - 13.6	2.5
13.9 - 18.4	4.5
25.1	
No data	

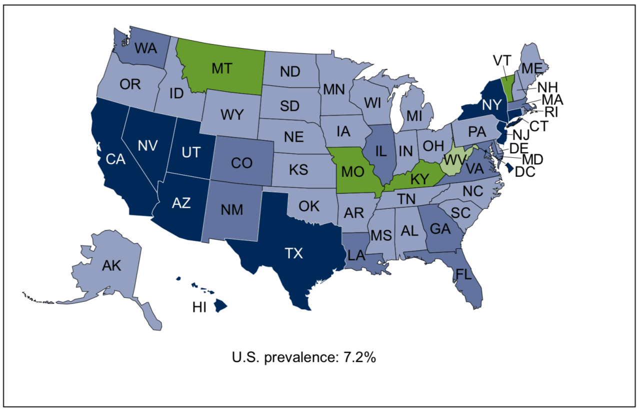
Prevalence of maternal smoking at any time during pregnancy, by state: United States, 2016.

Source: NCHS National Vital Statistics System, Natality.

https://meredith.images.worldnow.com/images/16207225_G.png?auto=webp&disable=upscale&width=800 34

Improper categorization

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



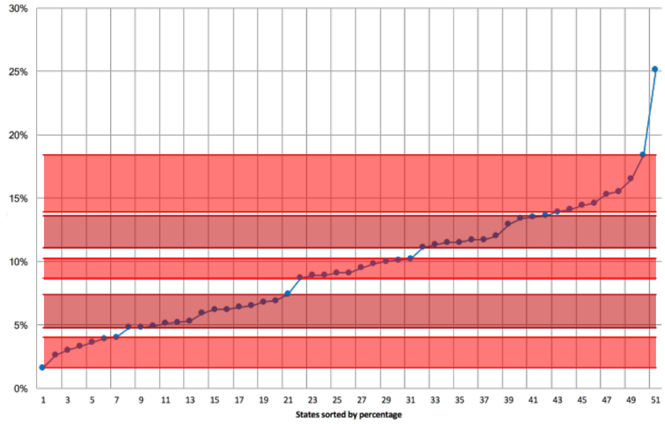
U.S. prevalence: 7.2%

4.9
2.1
7.7
4.9

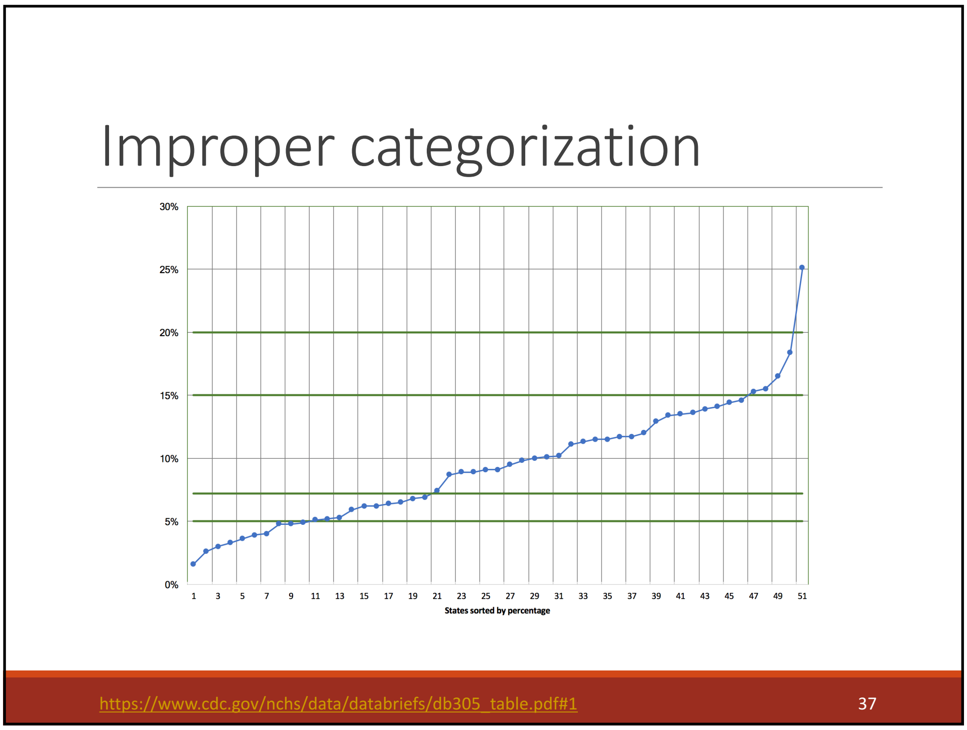
NOTE: Access data table for Figure 1 at: https://www.cdc.gov/nchs/data/databriefs/db305_table.pdf#1.
SOURCE: NCHS National Vital Statistics System, Natality.

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

Improper categorization



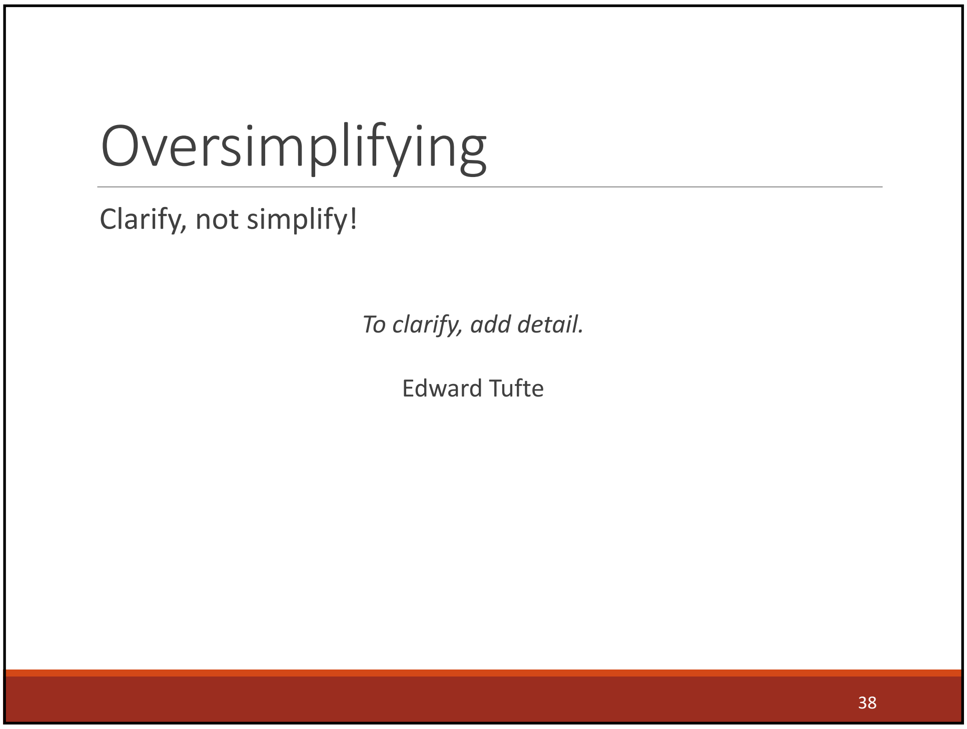
https://www.cdc.gov/nchs/data/databriefs/db305_table.pdf#1



Improper categorization

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

37



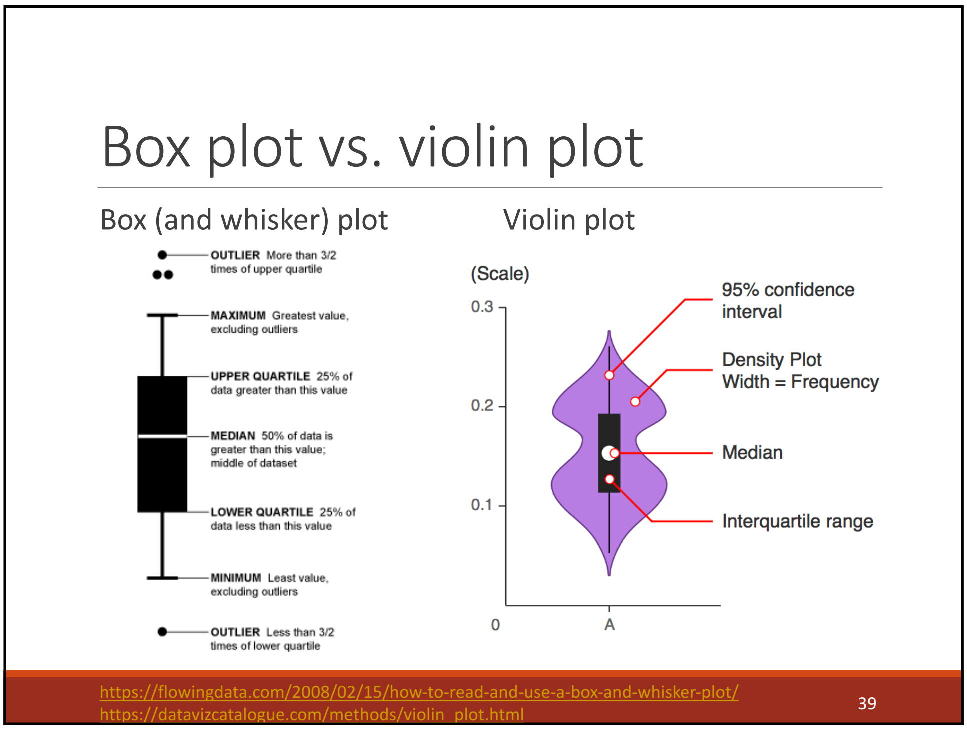
Oversimplifying

Clarify, not simplify!

To clarify, add detail.

Edward Tufte

38



Box plot vs. violin plot

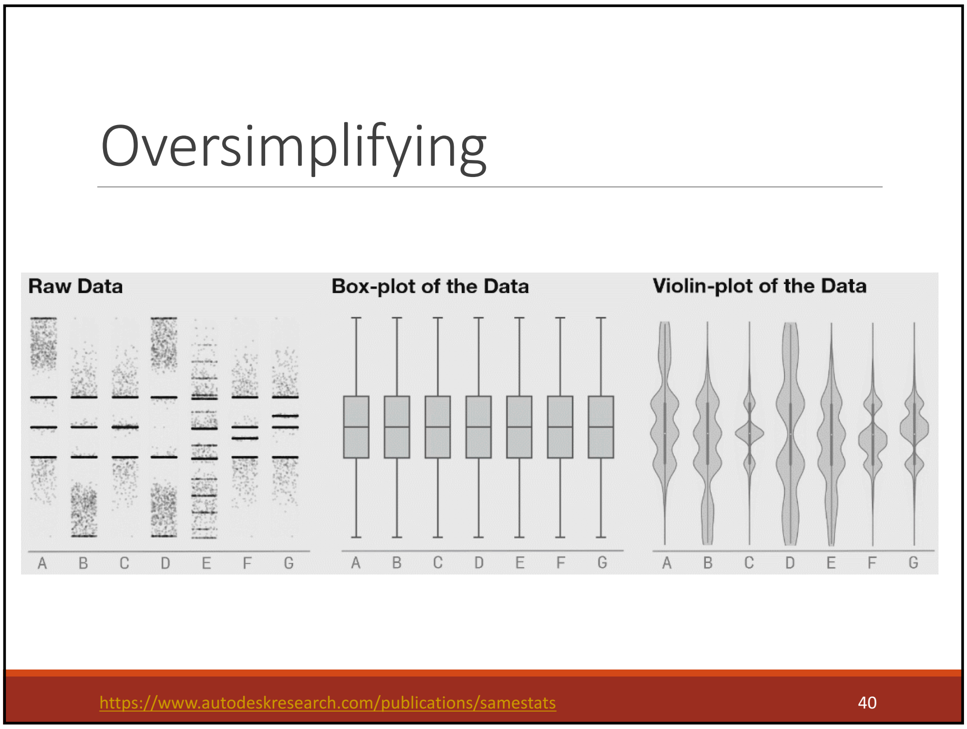
Box (and whisker) plot

Violin plot

OUTLIER More than 3/2 times of upper quartile
MAXIMUM Greatest value, excluding outliers
UPPER QUARTILE 25% of data greater than this value
MEDIAN 50% of data is greater than this value; middle of dataset
LOWER QUARTILE 25% of data less than this value
MINIMUM Least value, excluding outliers
OUTLIER Less than 3/2 times of lower quartile
(Scale)
0.3
0.2
0.1
0
A
95% confidence interval
Density Plot Width = Frequency
Median
Interquartile range

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

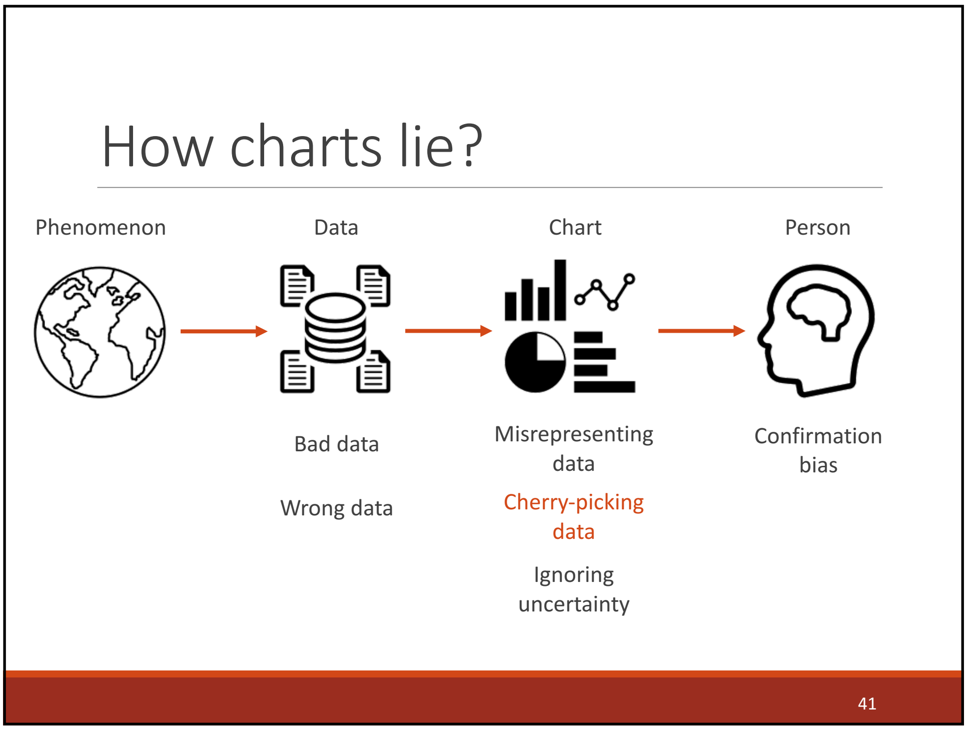
39



Oversimplifying

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

40



How charts lie?

Phenomenon → Data → Chart → Person

Data:

- Bad data
- Wrong data

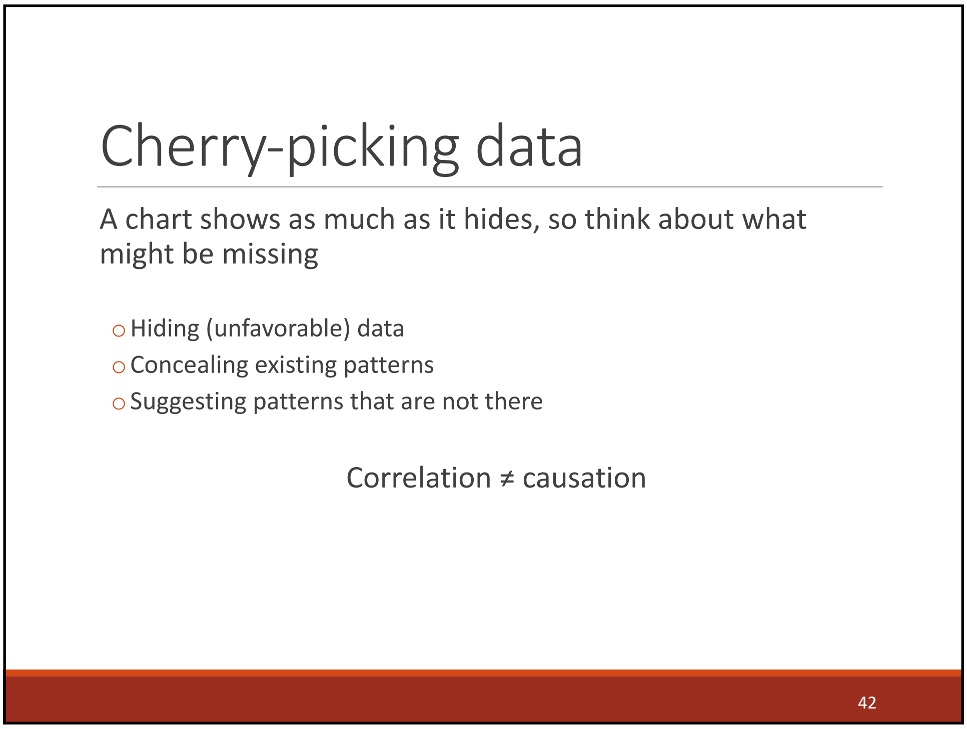
Chart:

- Misrepresenting data
- Cherry-picking data
- Ignoring uncertainty

Person:

- Confirmation bias

41



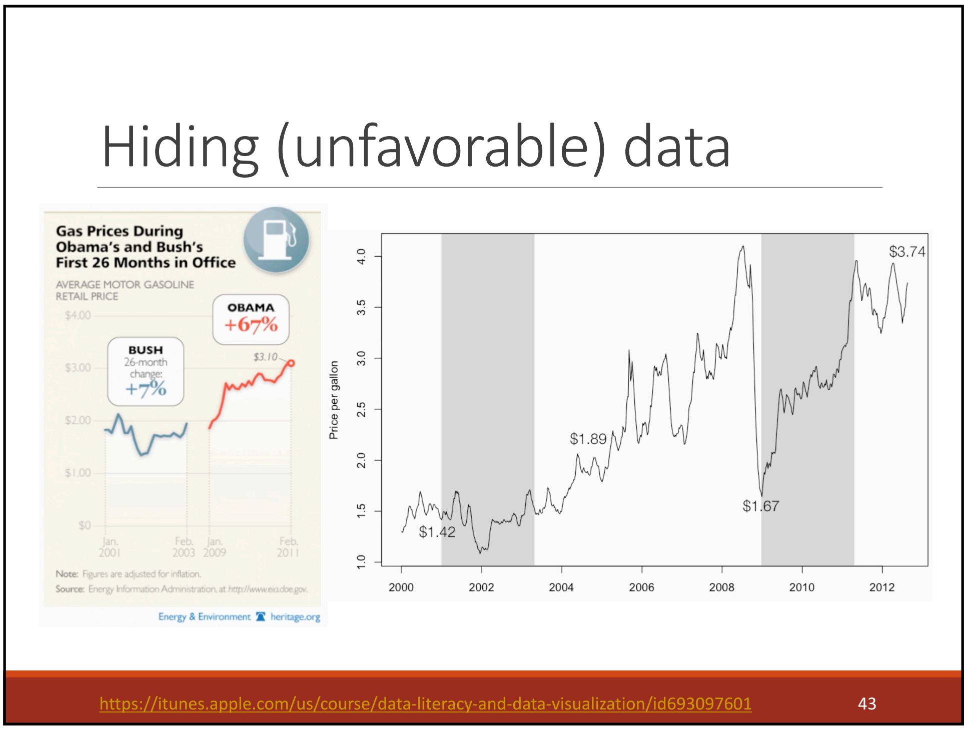
Cherry-picking data

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

- Hiding (unfavorable) data
- Concealing existing patterns
- Suggesting patterns that are not there

Correlation ≠ causation

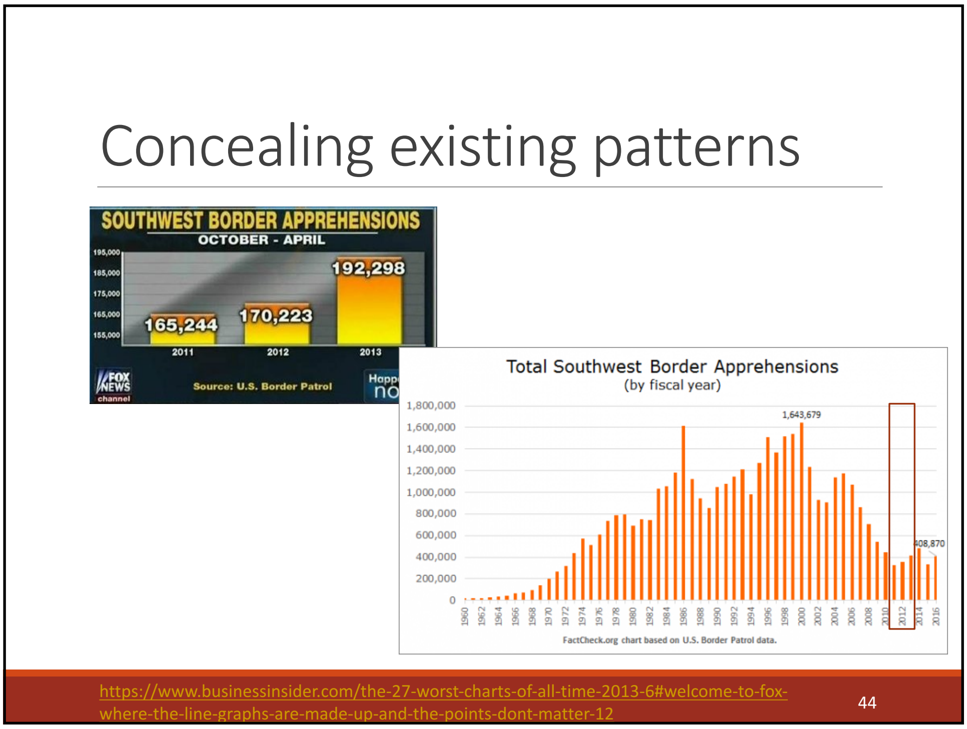
42



Hiding (unfavorable) data

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

43

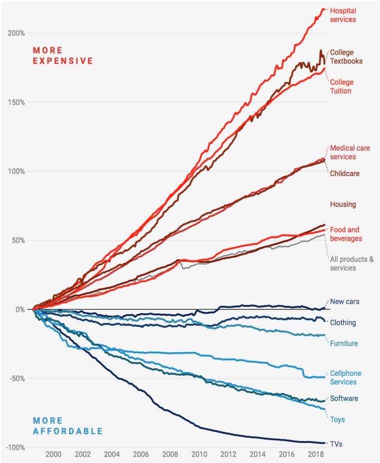


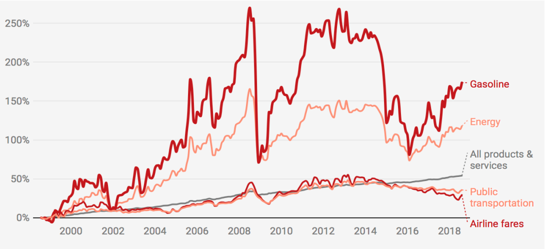
Concealing existing patterns

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

44

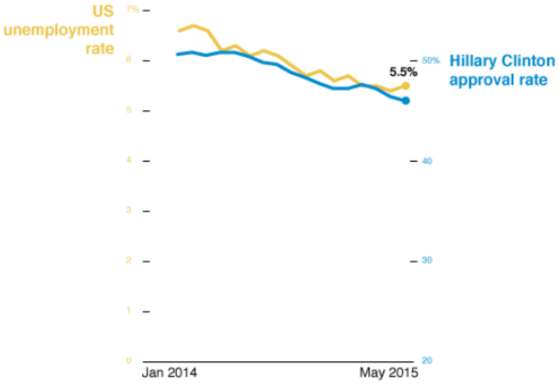
Concealing existing patterns



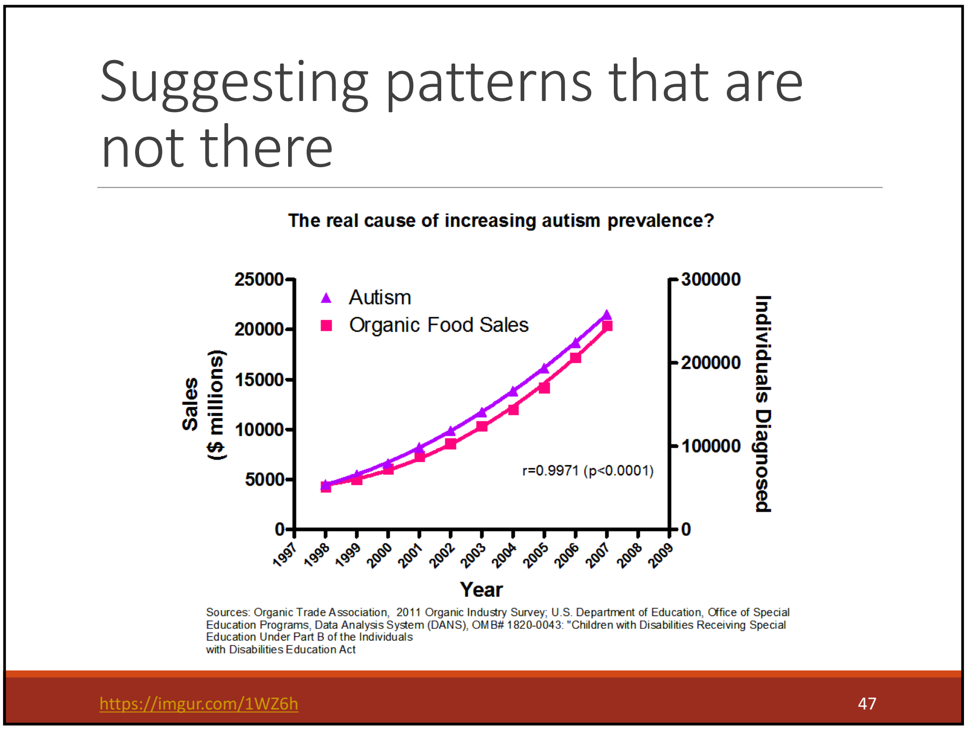


https://blog.datawrapper.de/weekly47-cpi-dollars-for-college/

Suggesting patterns that are not there



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



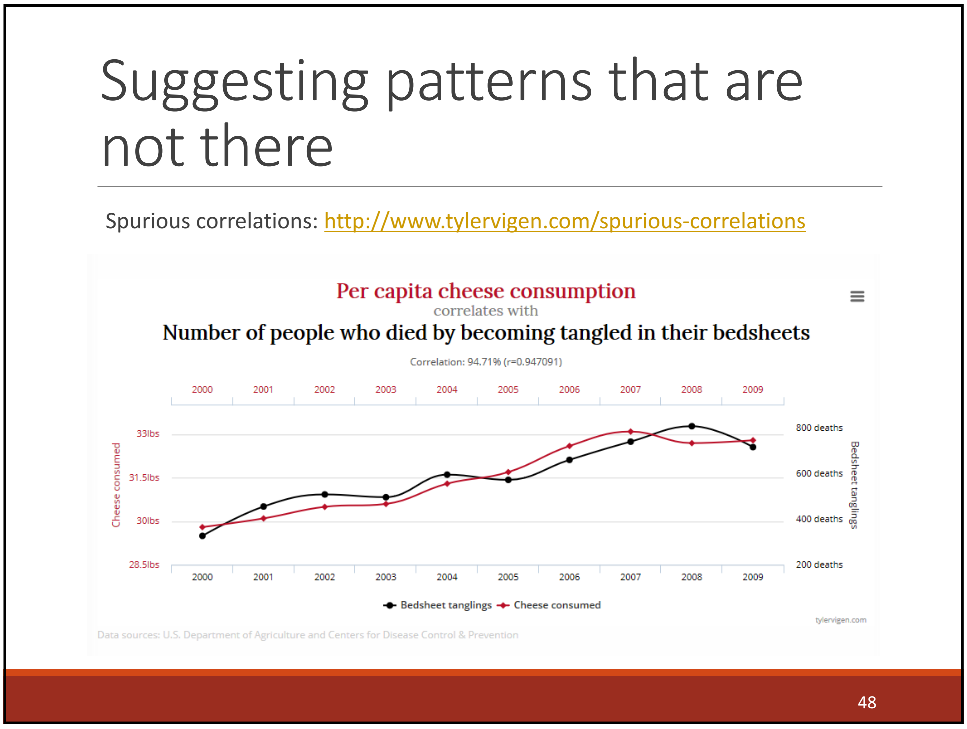
Suggesting patterns that are not there

The real cause of increasing autism prevalence?

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

https://imgur.com/1WZ6h

47



Suggesting patterns that are not there

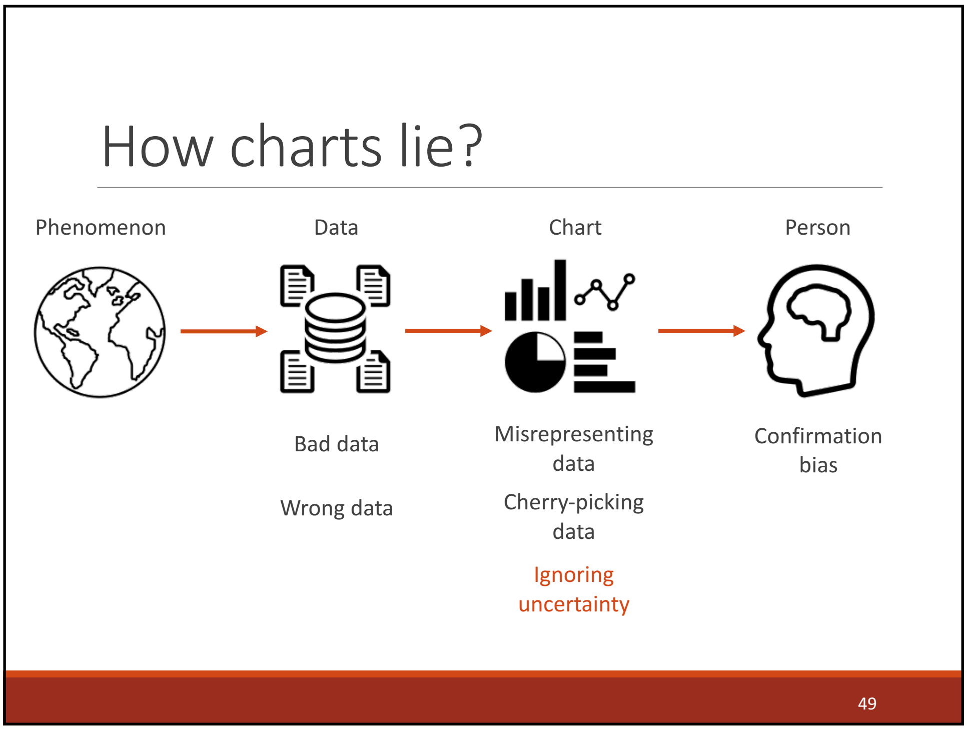
Spurious correlations: http://www.tylervigen.com/spurious-correlations

Per capita cheese consumption
correlates with
Number of people who died by becoming tangled in their bedsheets

Correlation: 94.71% (r=0.947091)

Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

48



How charts lie?

Phenomenon → Data → Chart → Person

Data: Bad data

Wrong data

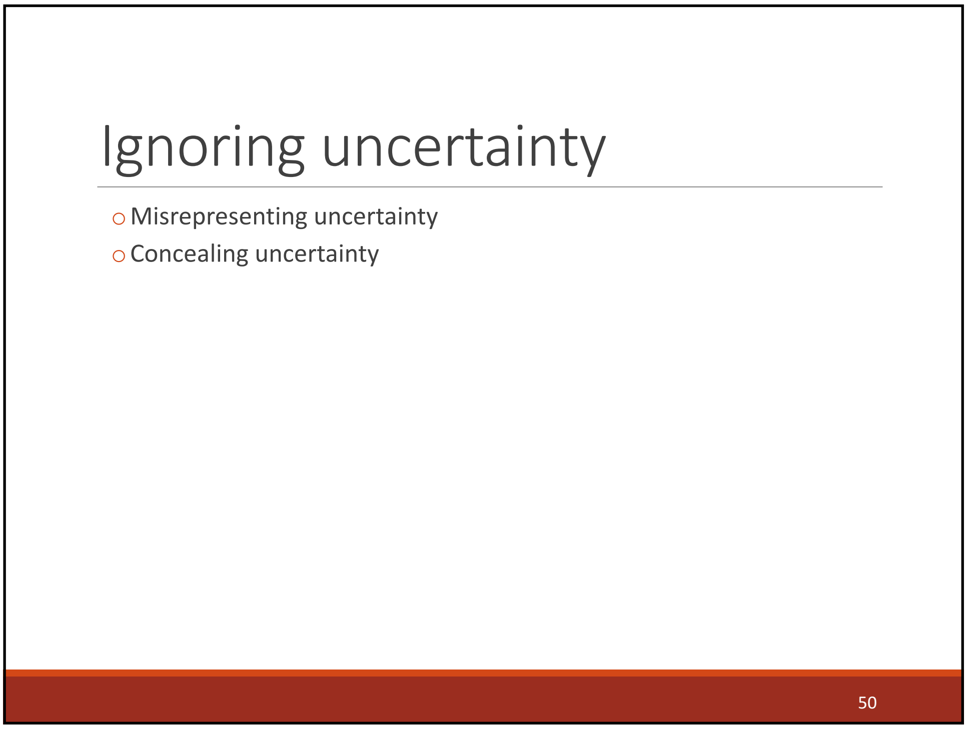
Chart: Misrepresenting data

Cherry-picking data

Ignoring uncertainty

Person: Confirmation bias

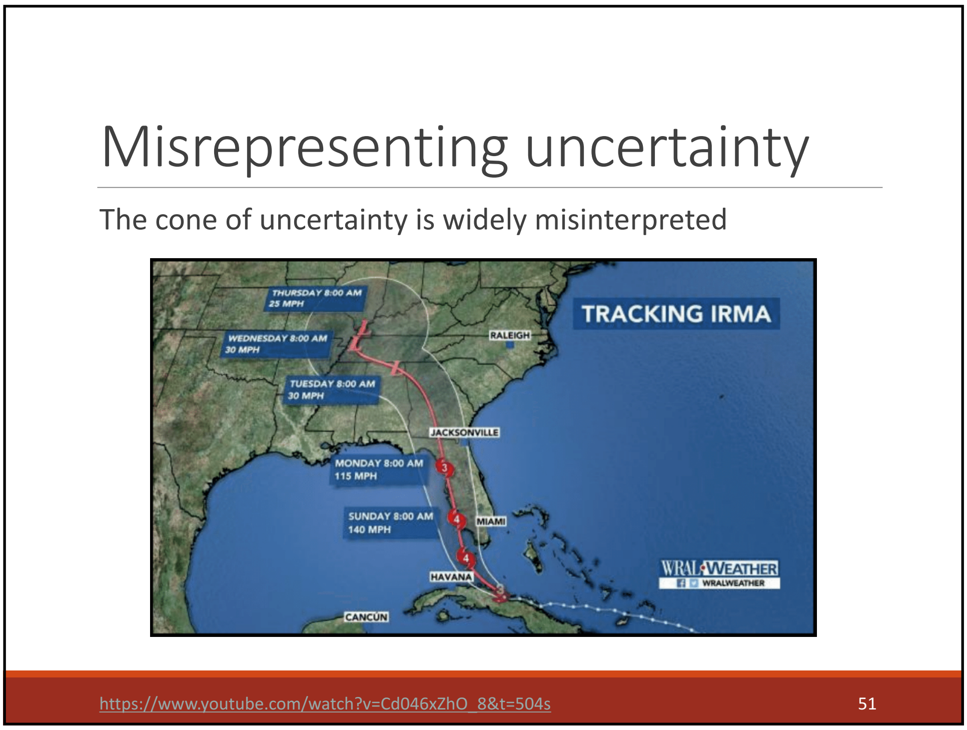
49



Ignoring uncertainty

- Misrepresenting uncertainty
- Concealing uncertainty

50

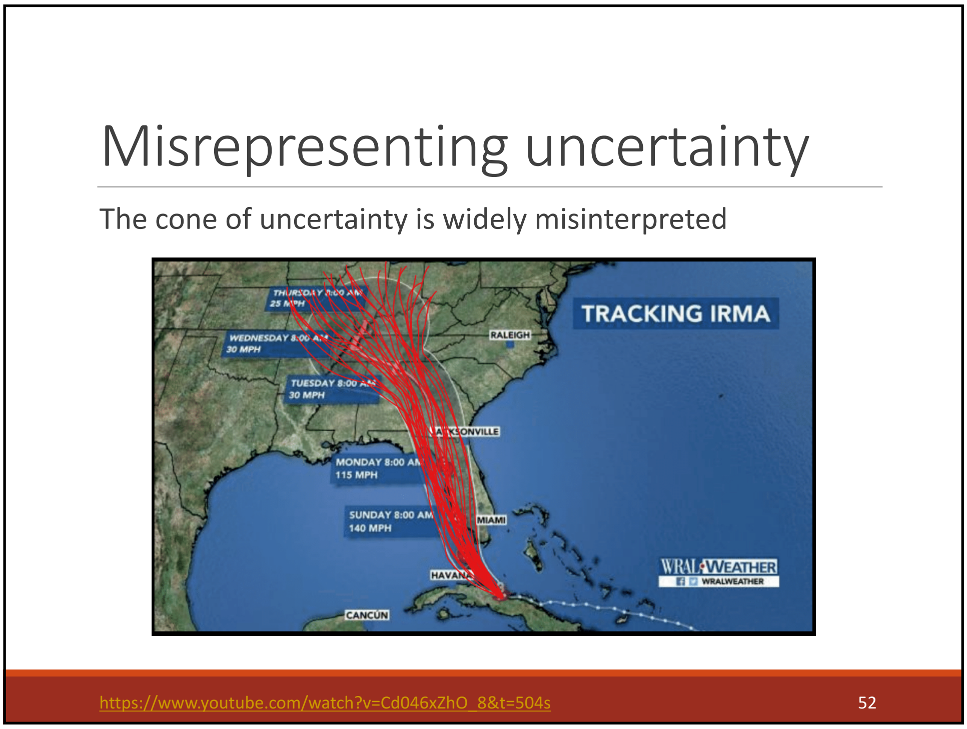


Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted

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

51

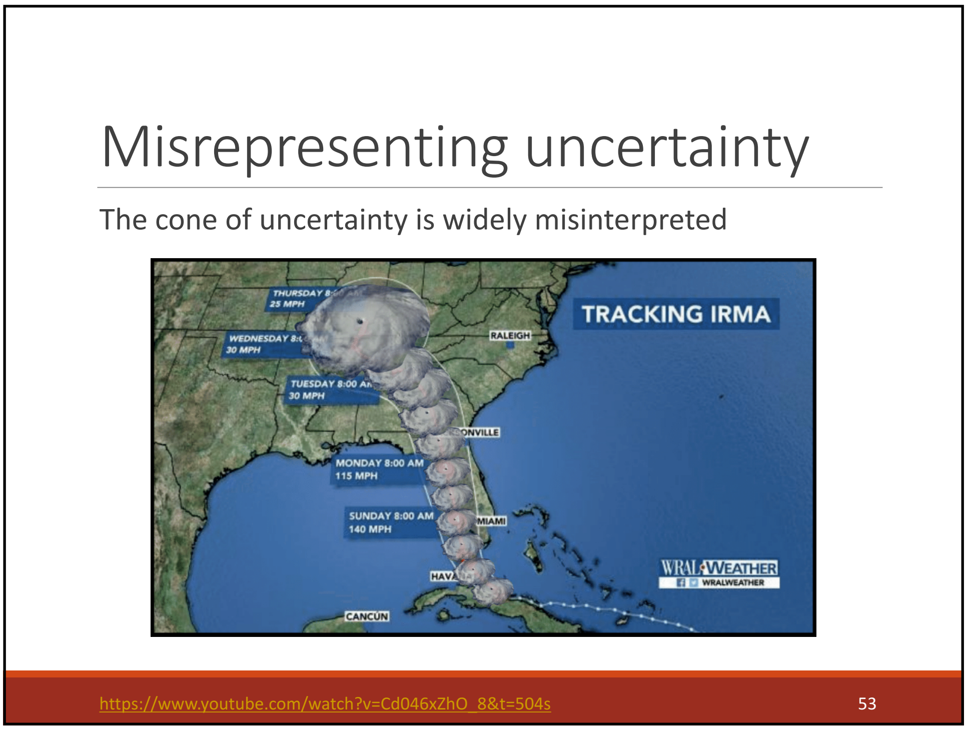


Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted

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

52

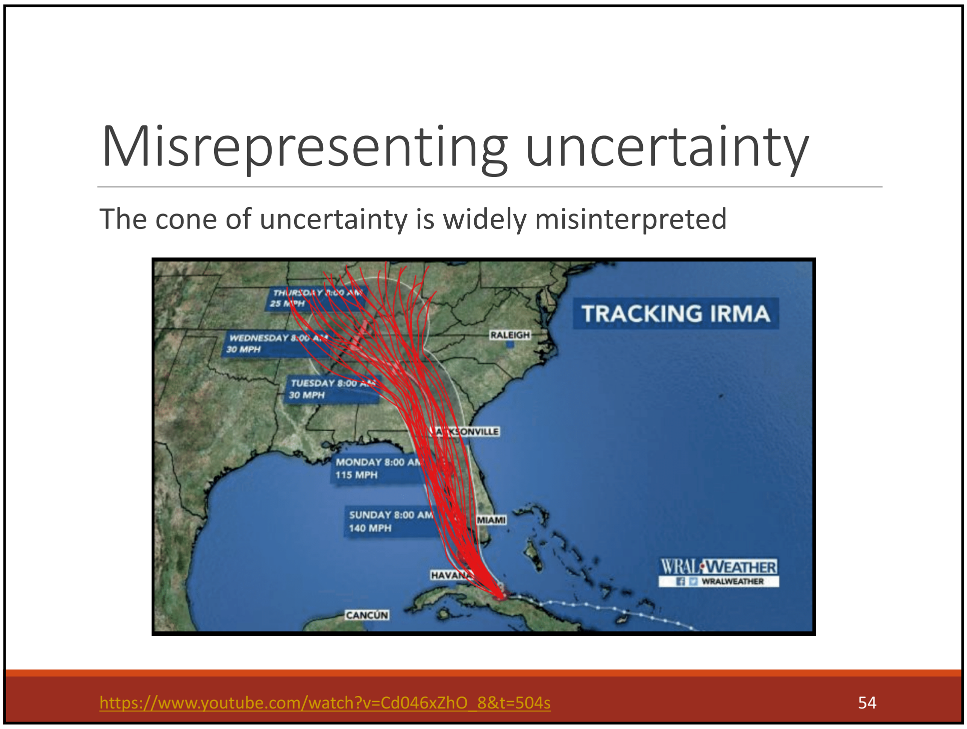


Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted

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

53



Misrepresenting uncertainty

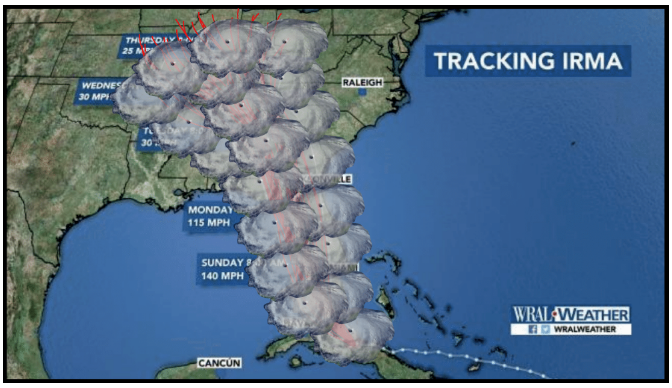
The cone of uncertainty is widely misinterpreted

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

54

Misrepresenting uncertainty

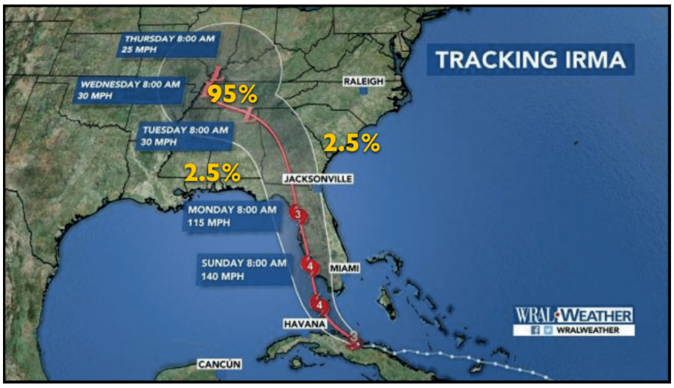
The cone of uncertainty is widely misinterpreted



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

Misrepresenting uncertainty

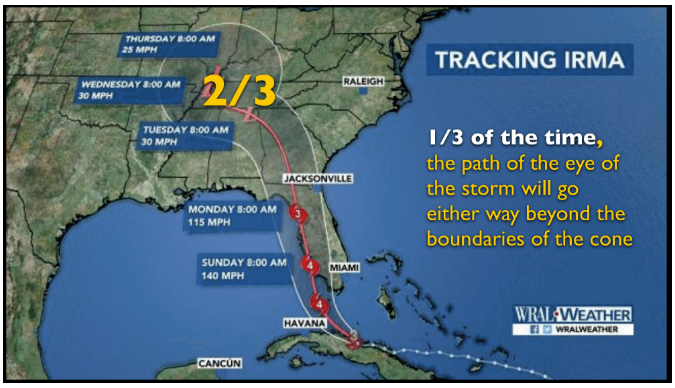
The cone of uncertainty is widely misinterpreted



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

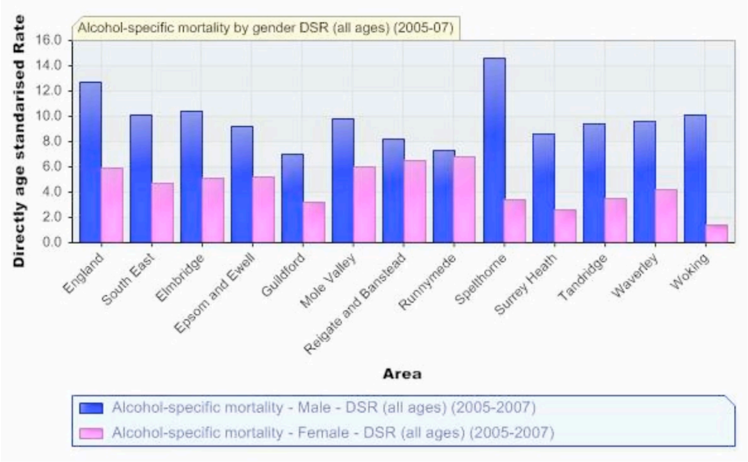
Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted

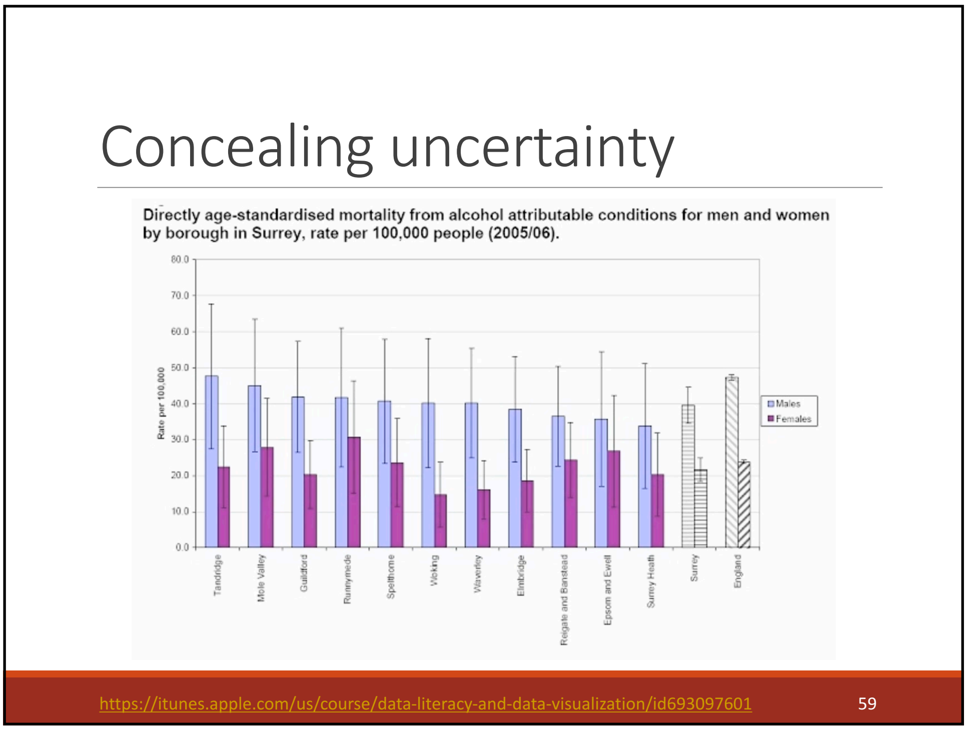


https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

Concealing uncertainty



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

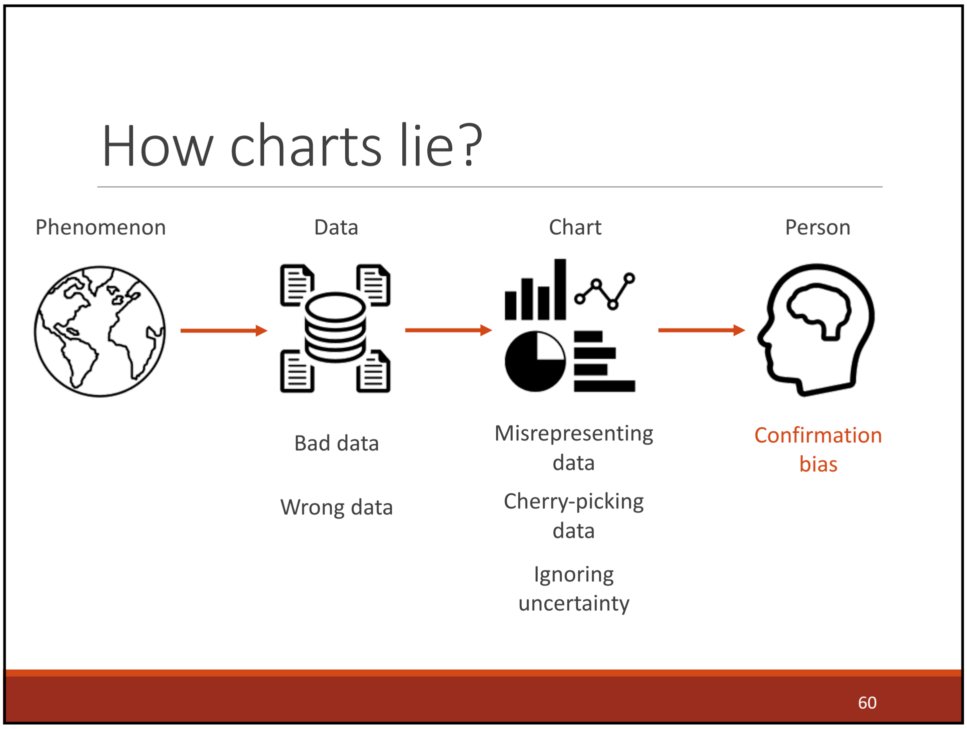


Concealing uncertainty

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

59



How charts lie?

Bad data

Wrong data

Misrepresenting data

Cherry-picking data

Ignoring uncertainty

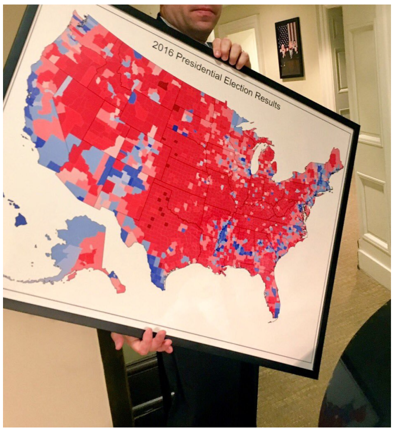
Confirmation bias

60

Confirmation bias

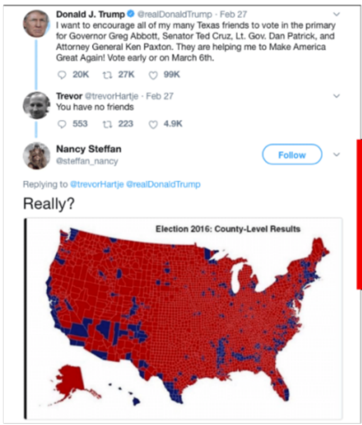
Charts lie because we lie to ourselves – we see what we want to see

Confirmation bias



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

Confirmation bias

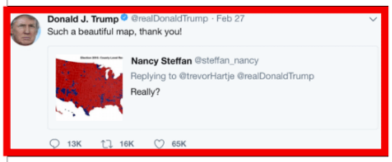


Donald J. Trump @realDonaldTrump · Feb 27
I want to encourage all of my many Texas friends to vote in the primary for Governor Greg Abbott, Senator Ted Cruz, Lt. Gov. Dan Patrick, and Attorney General Ken Paxton. They are helping me to Make America Great Again! Vote early or on March 6th.
20K 27K 99K

Trevor @trevorHartje · Feb 27
You have no friends
553 223 4.9K

Nancy Steffan @steffan_nancy

Replying to @trevorHartje @realDonaldTrump
Really?



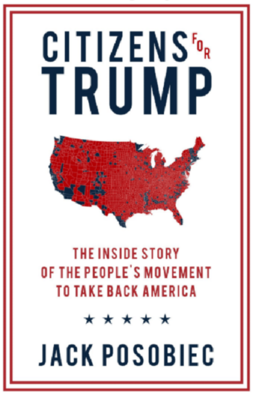
Donald J. Trump @realDonaldTrump · Feb 27
Such a beautiful map, thank you!

Nancy Steffan @steffan_nancy
Replying to @trevorHartje @realDonaldTrump
Really?
13K 16K 65K

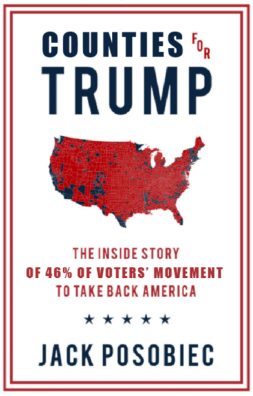
https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

Confirmation bias

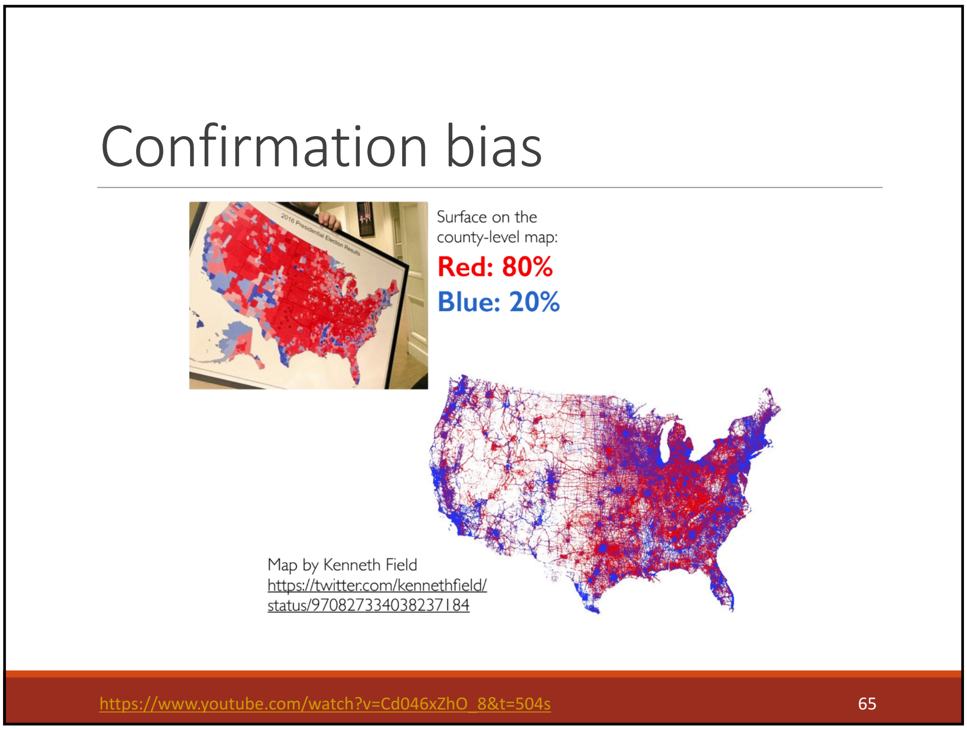
WRONG



RIGHT



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



Confirmation bias

Surface on the county-level map:

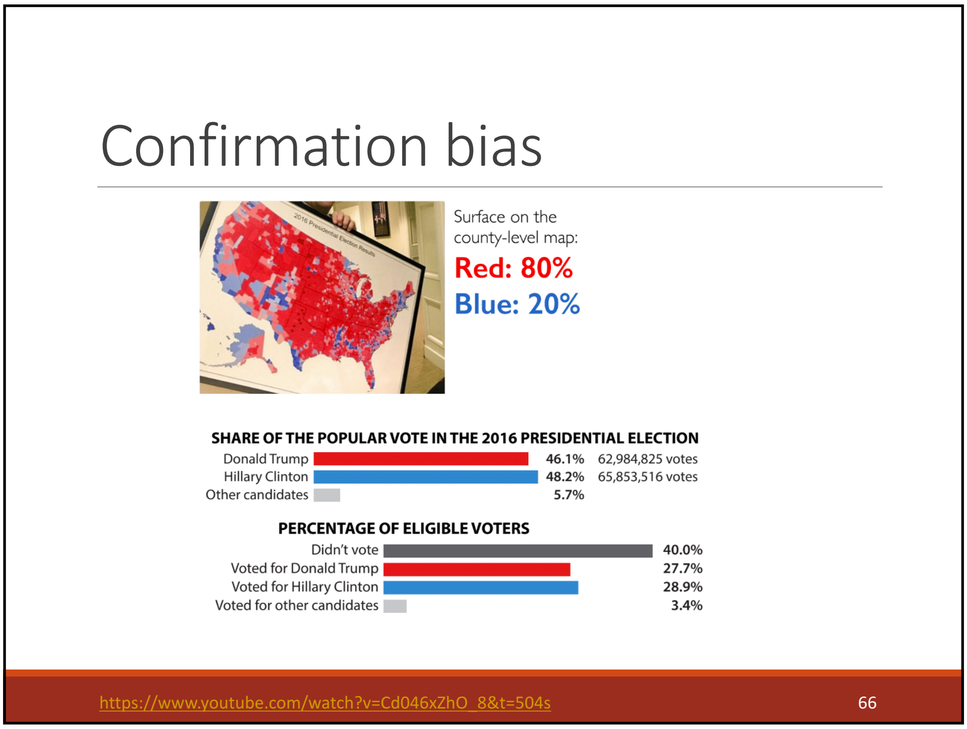
Red: 80%

Blue: 20%

Map by Kenneth Field
https://twitter.com/kennethfield/status/970827334038237184

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

65



Confirmation bias

Surface on the county-level map:

Red: 80%

Blue: 20%

SHARE OF THE POPULAR VOTE IN THE 2016 PRESIDENTIAL ELECTION

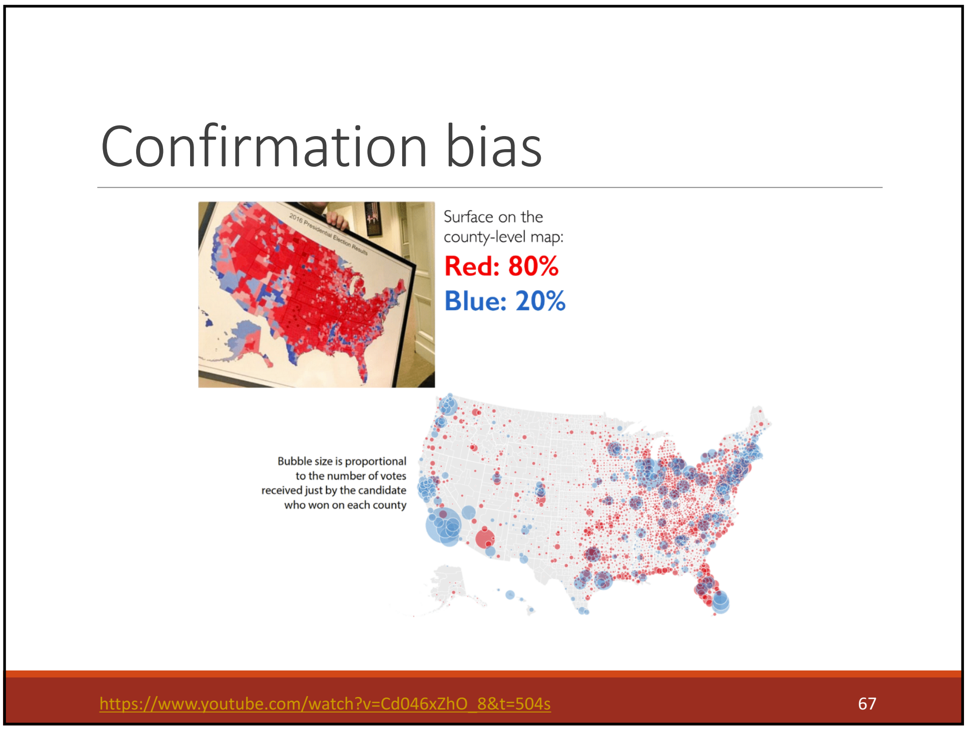
Donald Trump	**46.1%**	62,984,825 votes
Hillary Clinton	**48.2%**	65,853,516 votes
Other candidates	**5.7%**	

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%**

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

66



Confirmation bias

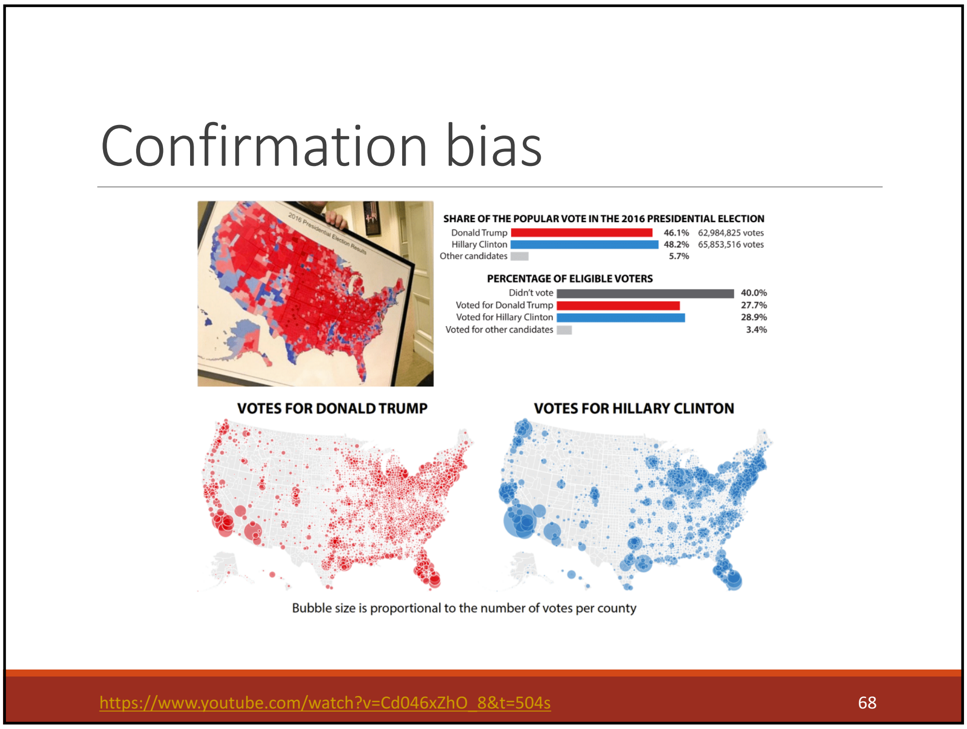
Surface on the county-level map:

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



Confirmation bias

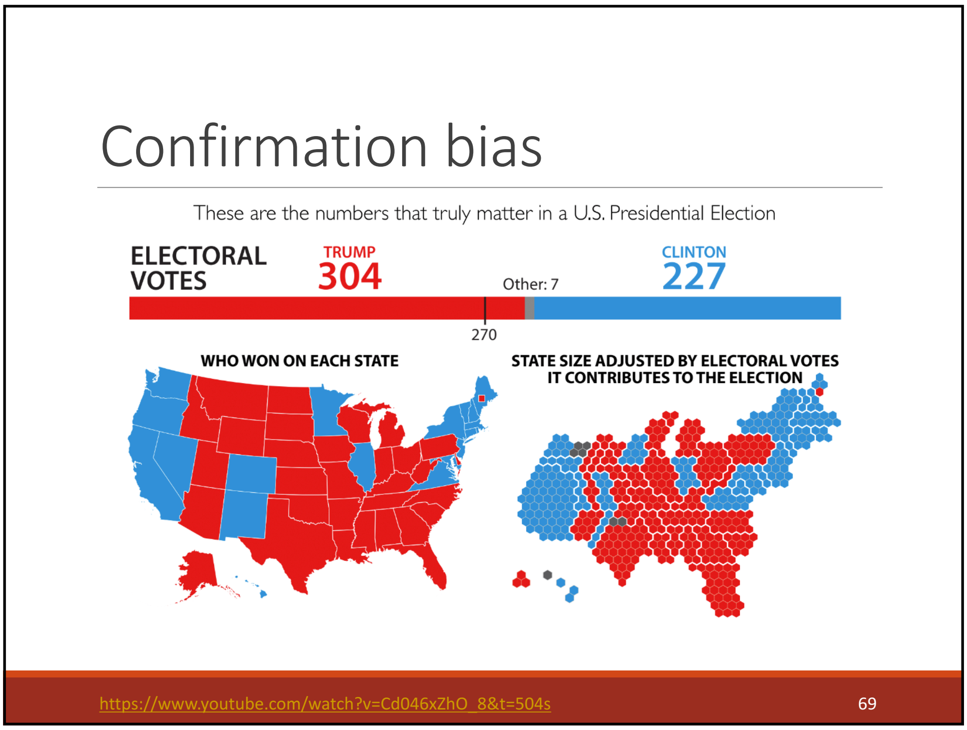
SHARE OF THE POPULAR VOTE IN THE 2016 PRESIDENTIAL ELECTION
Donald Trump 46.1% 62,984,825 votes
Hillary Clinton 48.2% 65,853,516 votes
Other candidates 5.7%
PERCENTAGE OF ELIGIBLE VOTERS
Didn't vote 40.0%
Voted for Donald Trump 27.7%
Voted for Hillary Clinton 28.9%
Voted for other candidates 3.4%

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



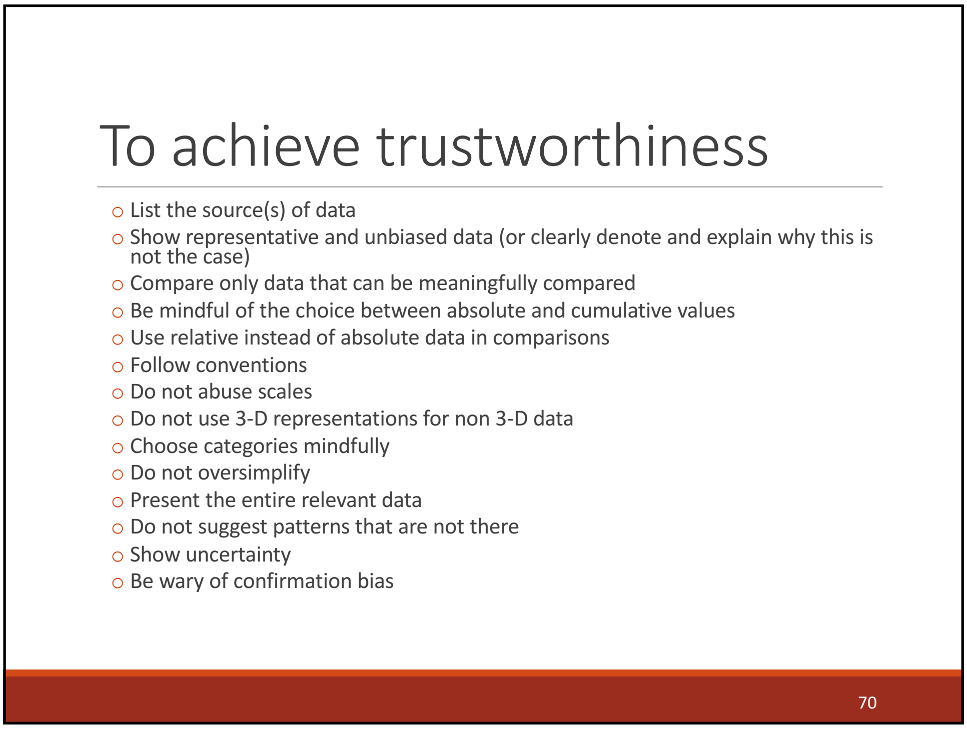
Confirmation bias

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

ELECTORAL VOTES
TRUMP
304
Other: 7
CLINTON
227
270
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

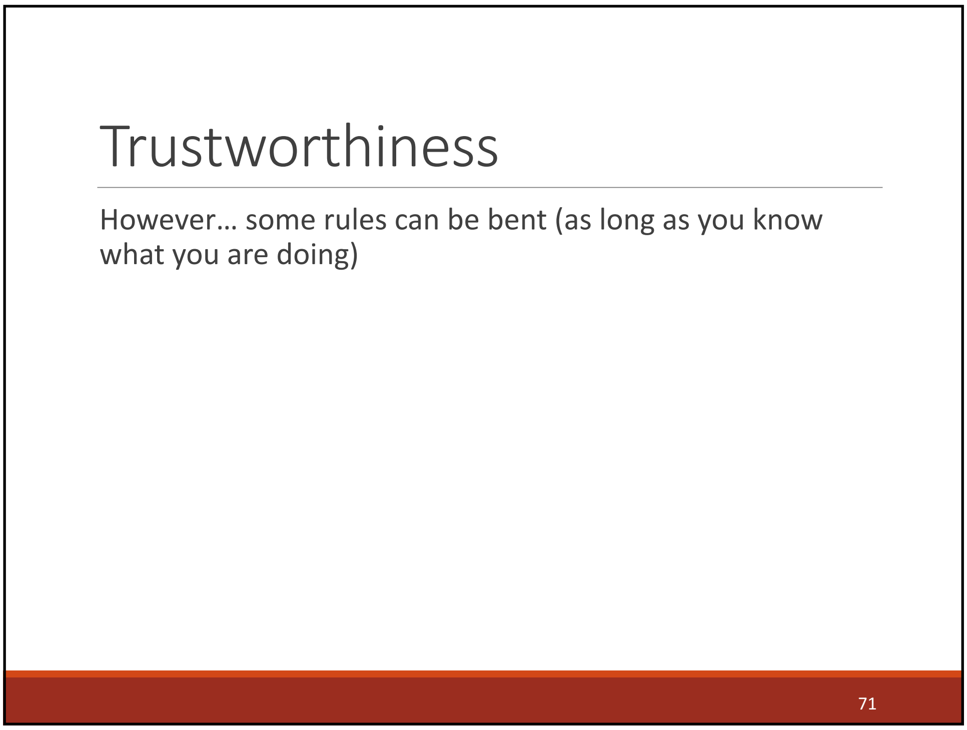
69



To achieve trustworthiness

- 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
- 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

70



Trustworthiness

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