# Data Visualization

EXAMPLES (1)

# Good visualization design is

1. Trustworthy

2. Accessible

3. Elegant

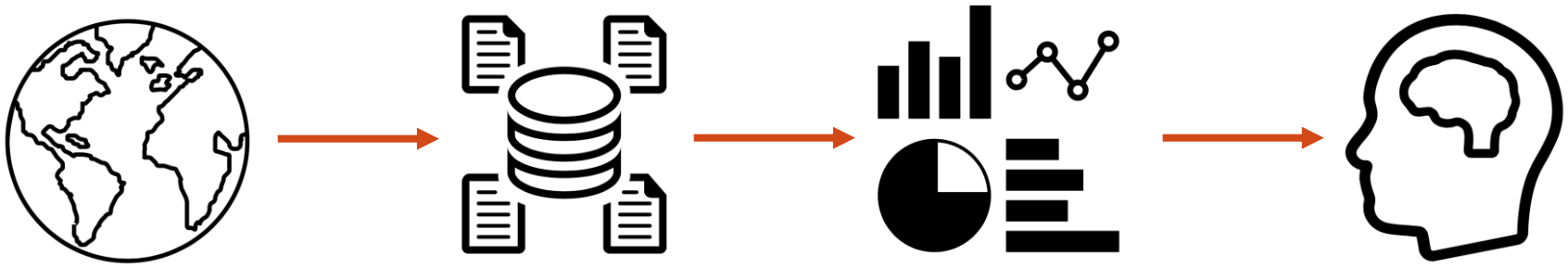
# How charts lie?

Phenomenon

Data

Chart

Person



Dubious data

Misrepresenting data

Cherry-picking data

Ignoring uncertainty

Confirmation bias

# Dubious data

## Unrepresentative data

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

## Biased data

- Question framing in polls
- Choice of measures

## Comparisons using

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

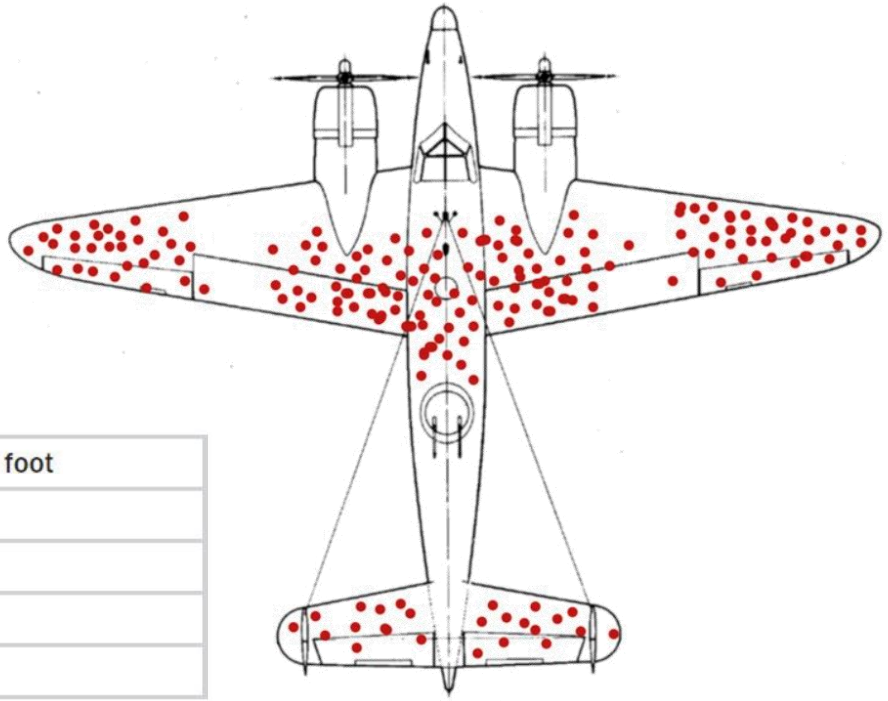
# Unrepresentative samples


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

# Missing 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

# Question framing in polls

Brexit referendum

- First proposal

*"Should the United Kingdom remain a member of the European Union?"*

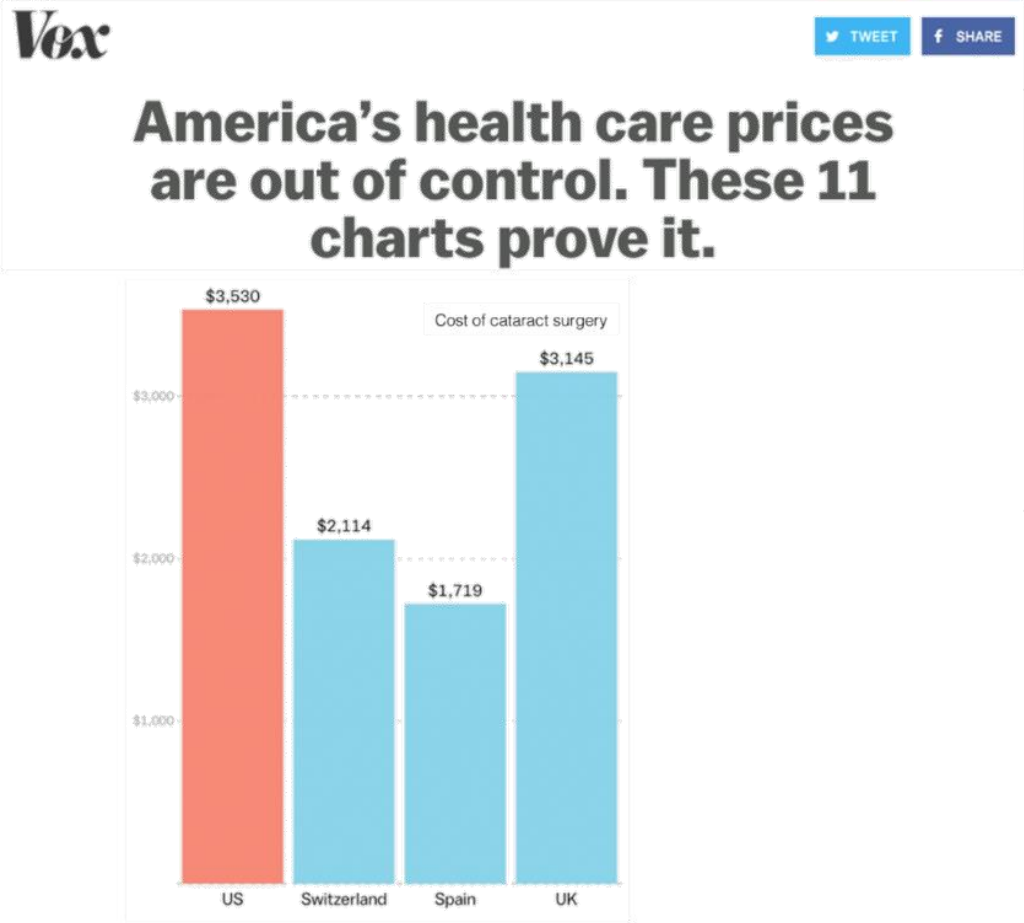
*yes/no*

- Final question

*"Should the United Kingdom remain a member of the European Union or leave the European Union?"*

*remain/leave*

# Non-comparable data used in comparisons



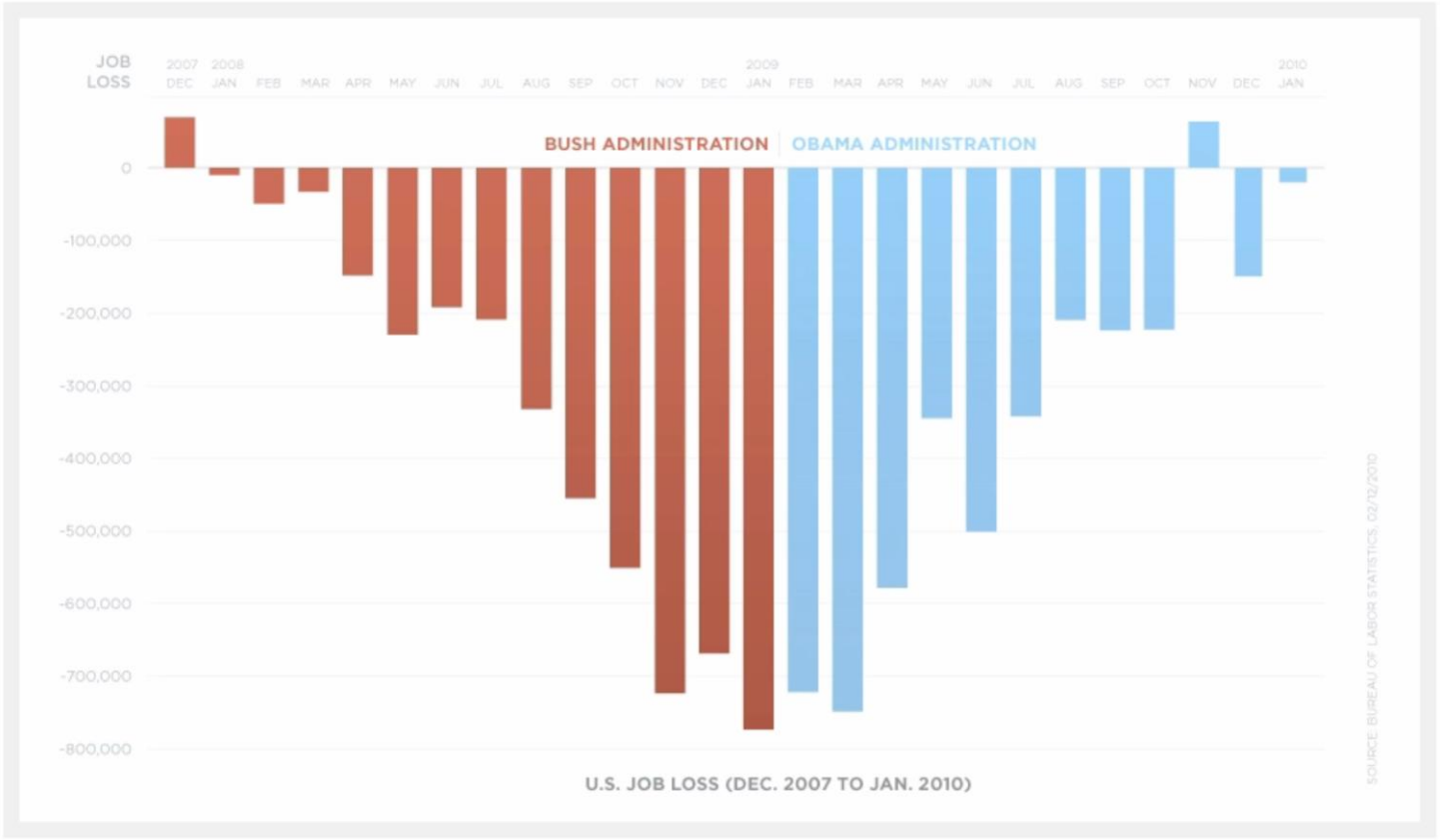

Two issues

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

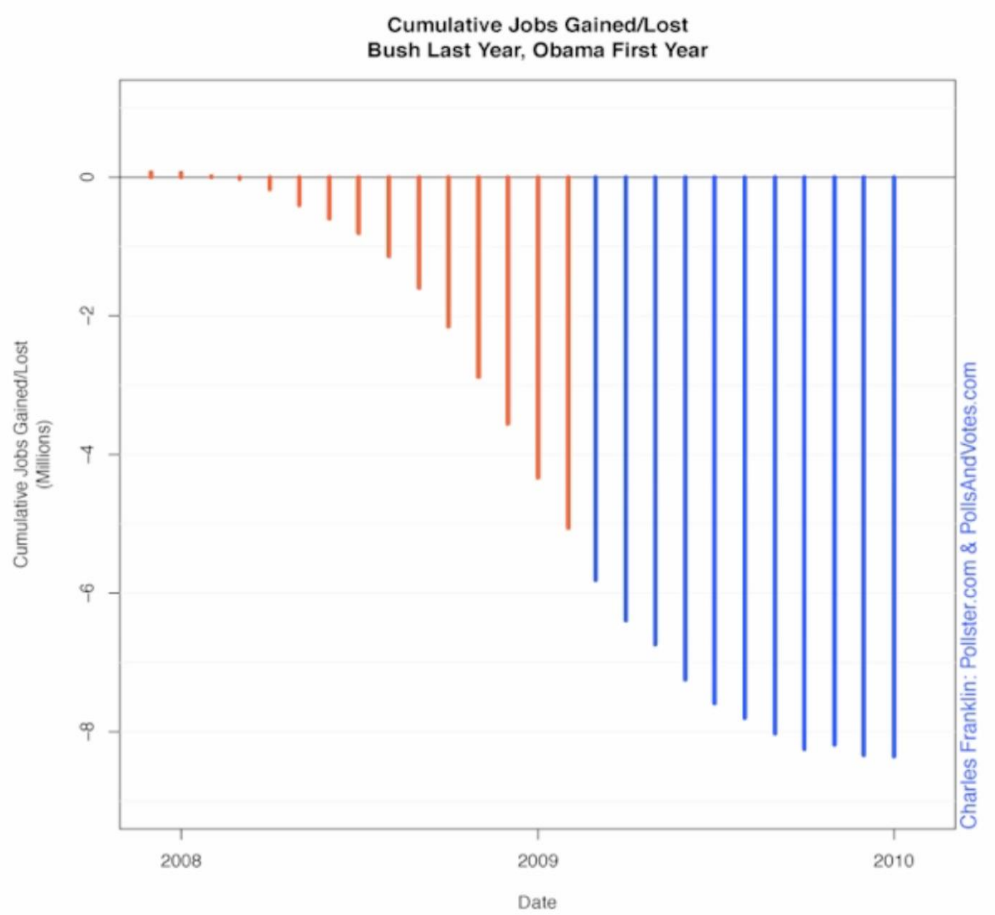
The data source specifically warns against using this data for comparison

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

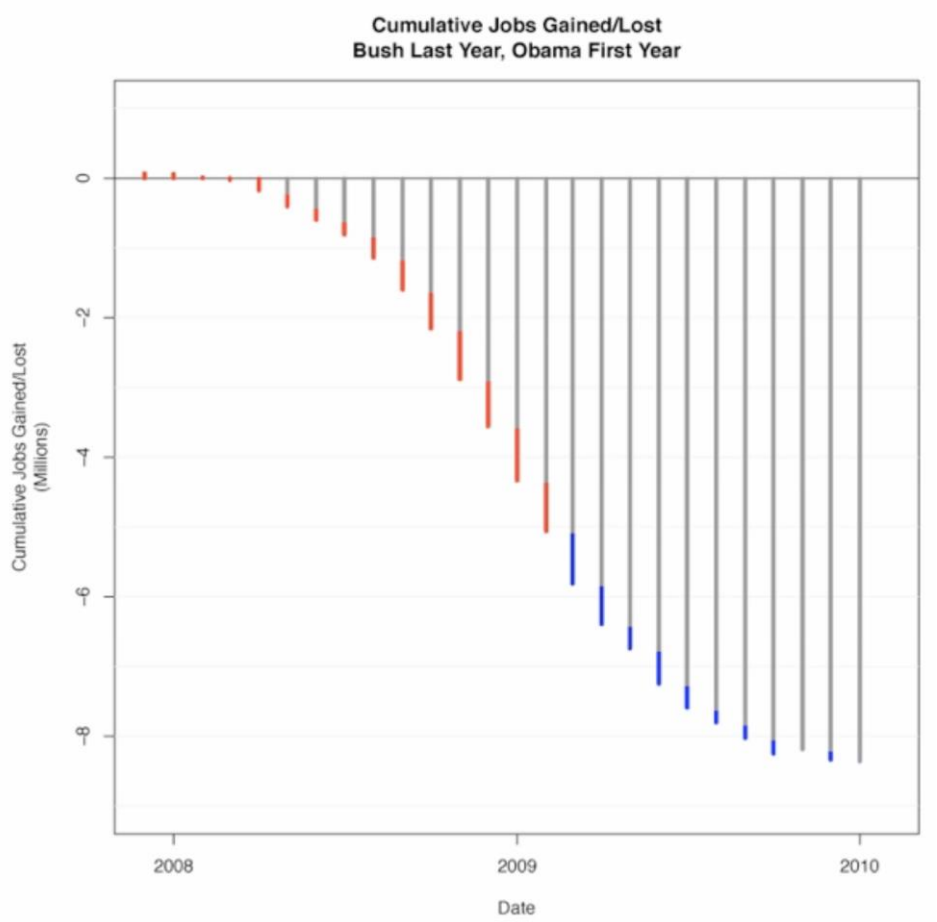
# Absolute instead of cumulative data



# Cumulative data

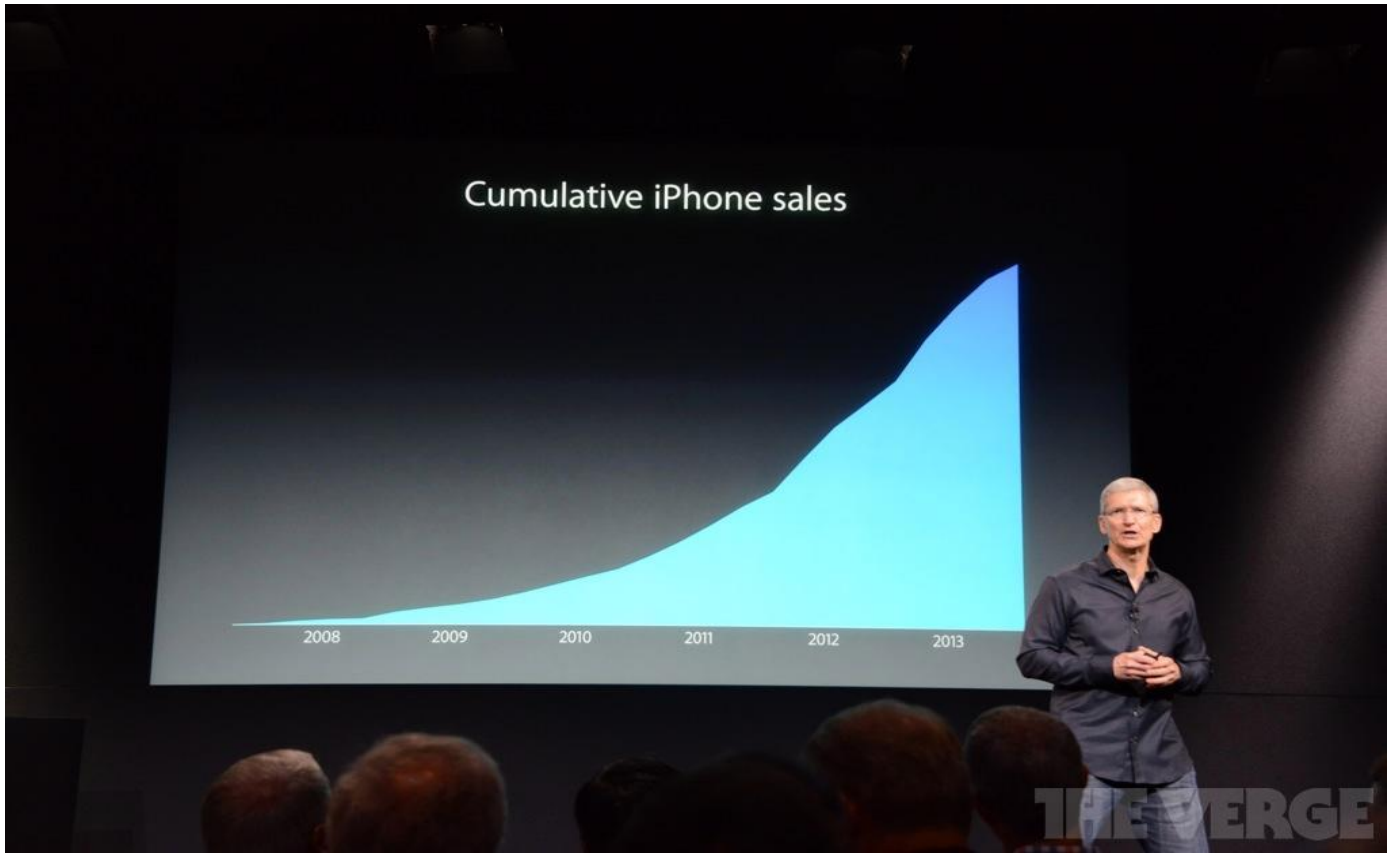


# Absolute and cumulative data



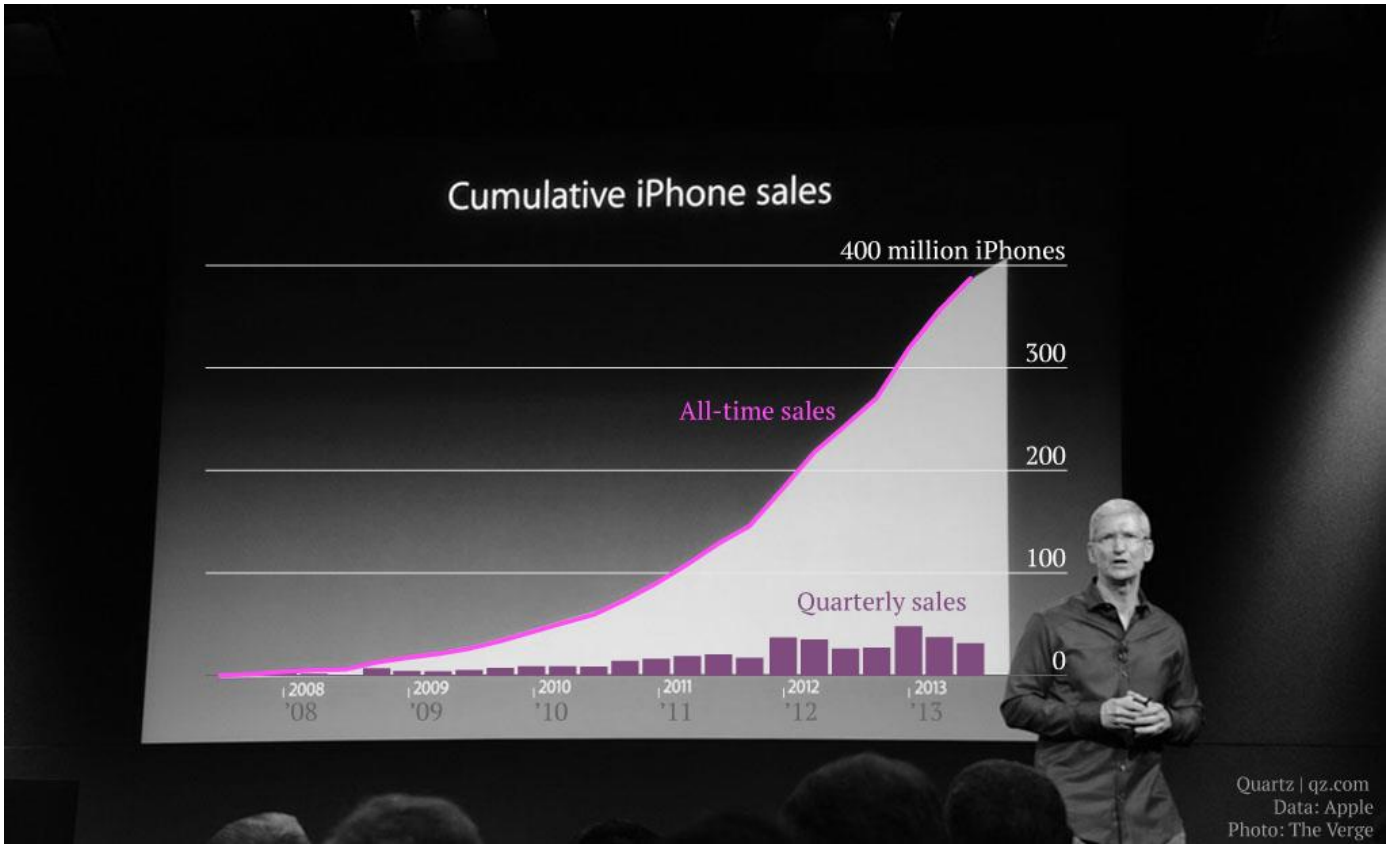

https://itunes.apple.com/us/course/data-literacy-and-data-visualization/id693097601

# Cumulative instead of absolute data



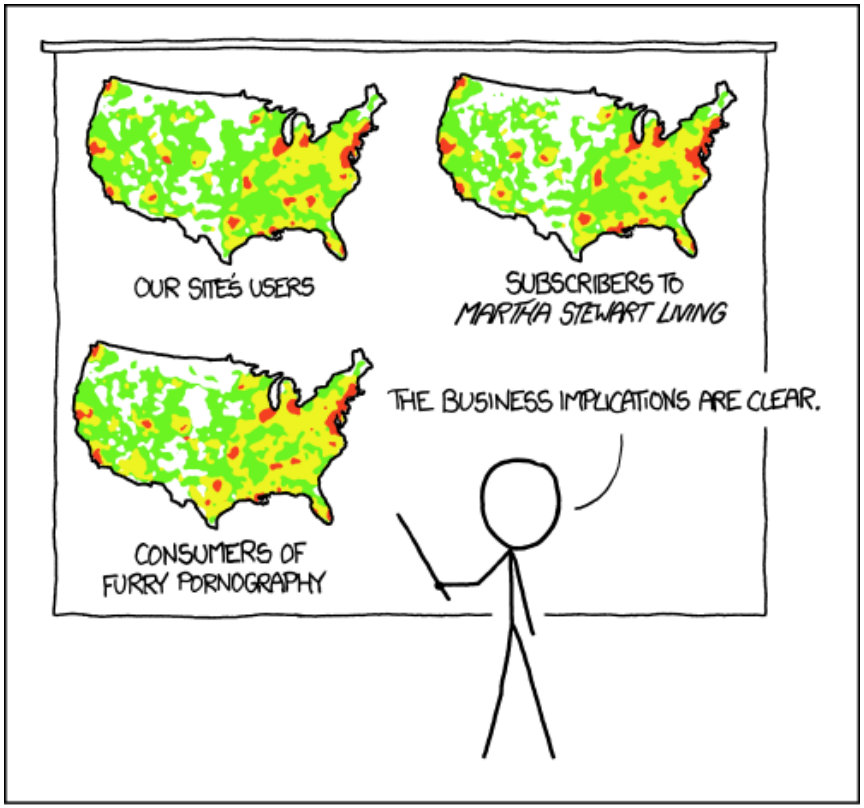

https://qz.com/122921/the-chart-tim-cook-doesnt-want-you-to-see/

# Cumulative and absolute data




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

# Absolute instead of relative data



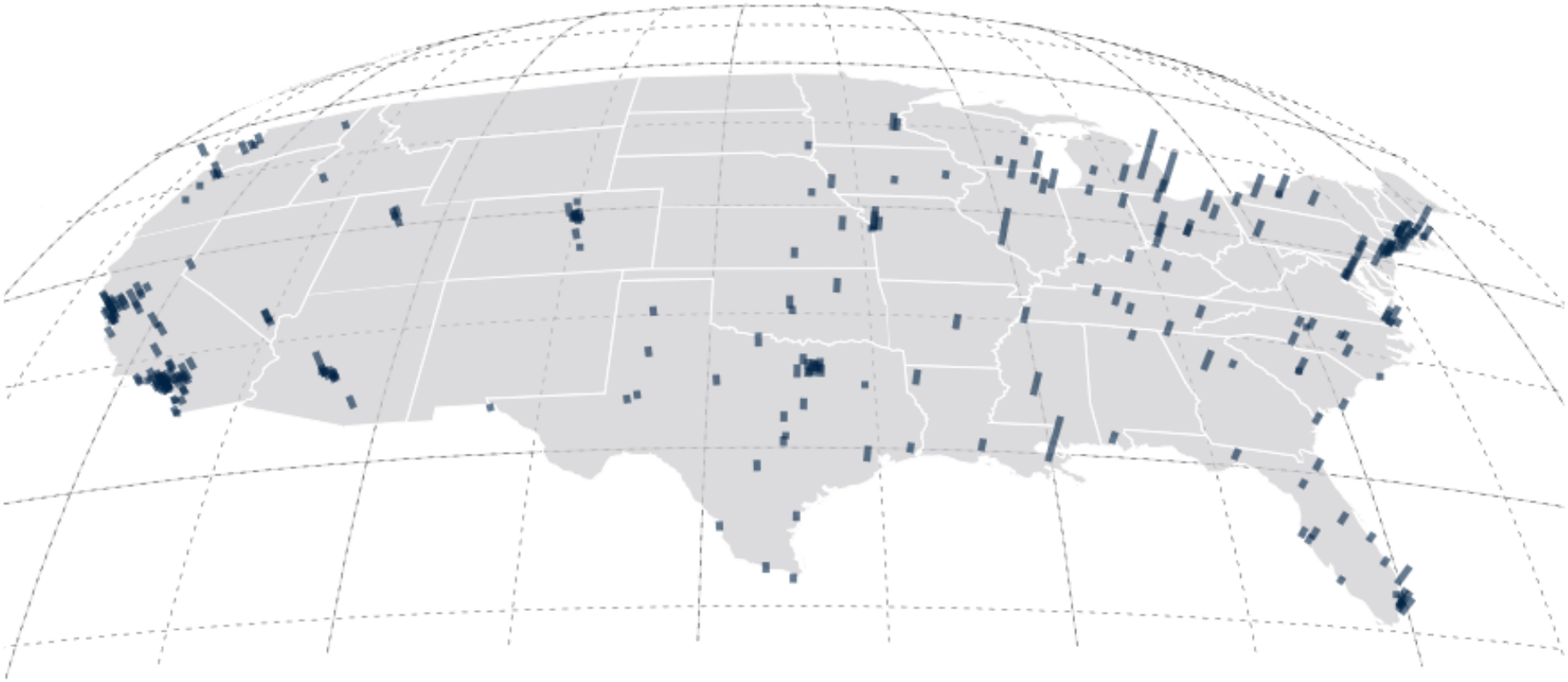

PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

https://xkcd.com/1138/

# Absolute data

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

# Relative data



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

# Dubious data

*Garbage in, garbage out*

A problem when

- It is not made clear
- The data is used for visualizations that are suitable for more ‘regular’ data

# How charts lie?

Phenomenon

Data

Dubious data

Chart

Misrepresenting data

Cherry-picking data

Ignoring uncertainty

Person

Confirmation bias

# Misrepresenting data

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

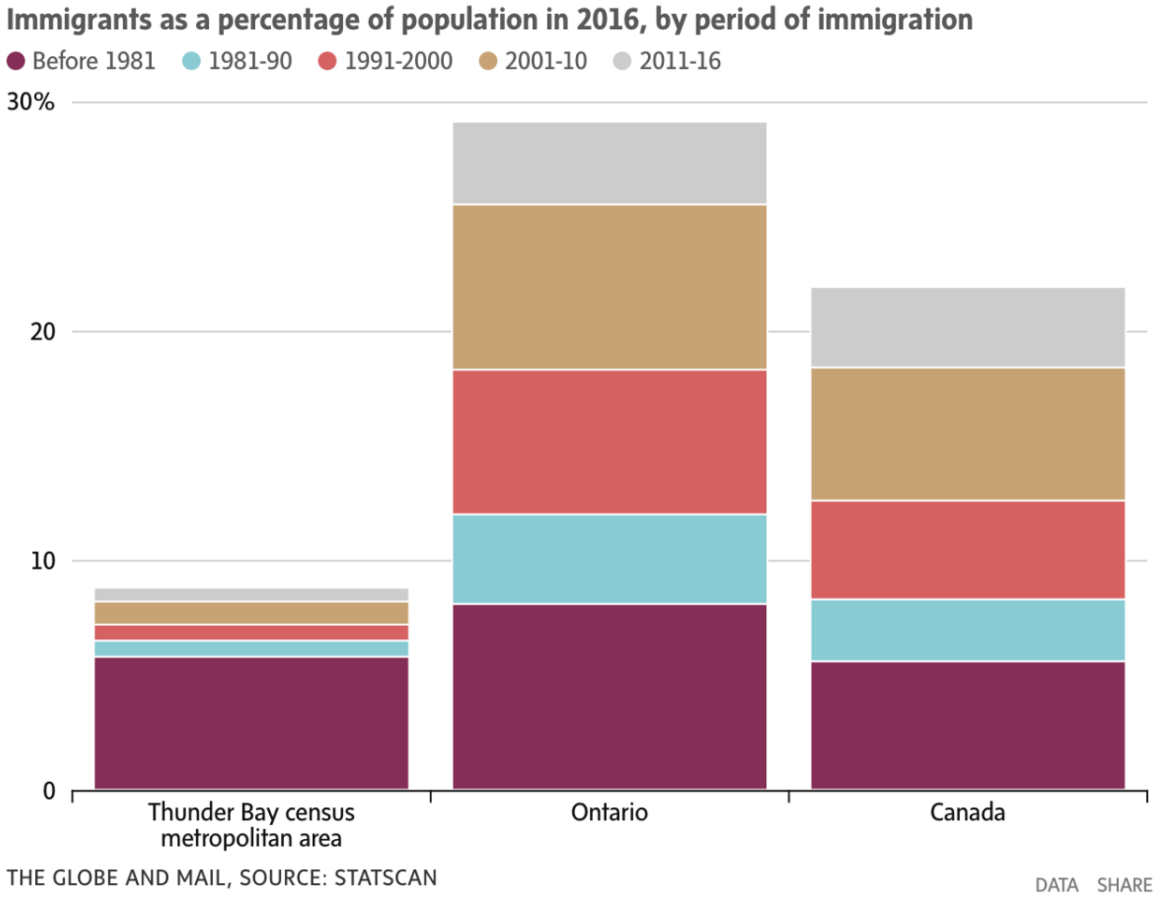
Abusing scales
- Truncated axis
- Aspect ratio bias
- Dual axes
- Improper scaling of areas and pictograms

Unnecessary 3-D

Improper categorization

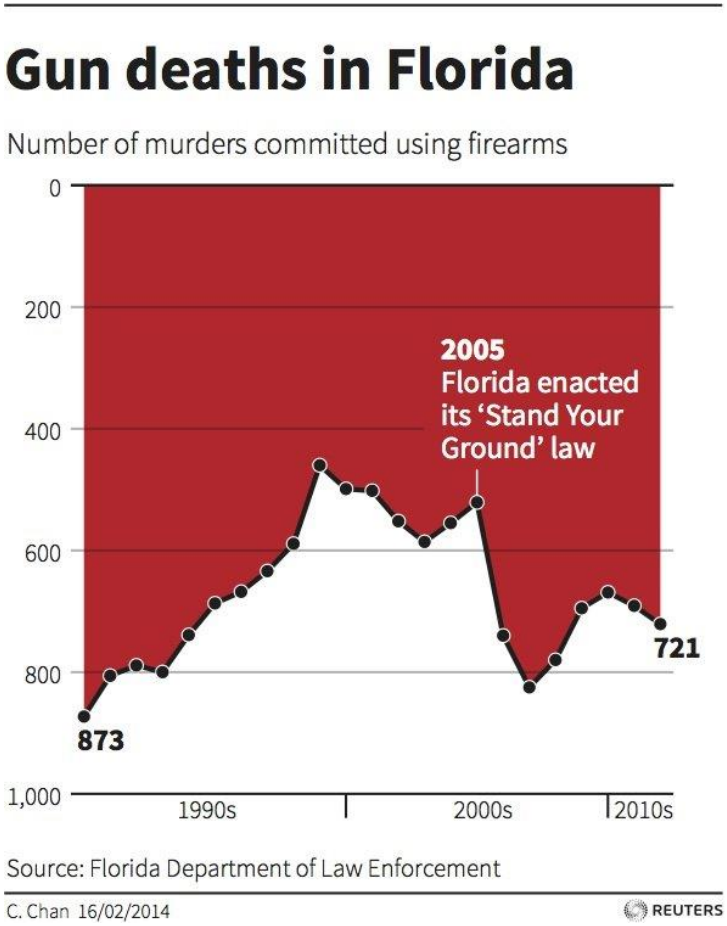
Oversimplifying

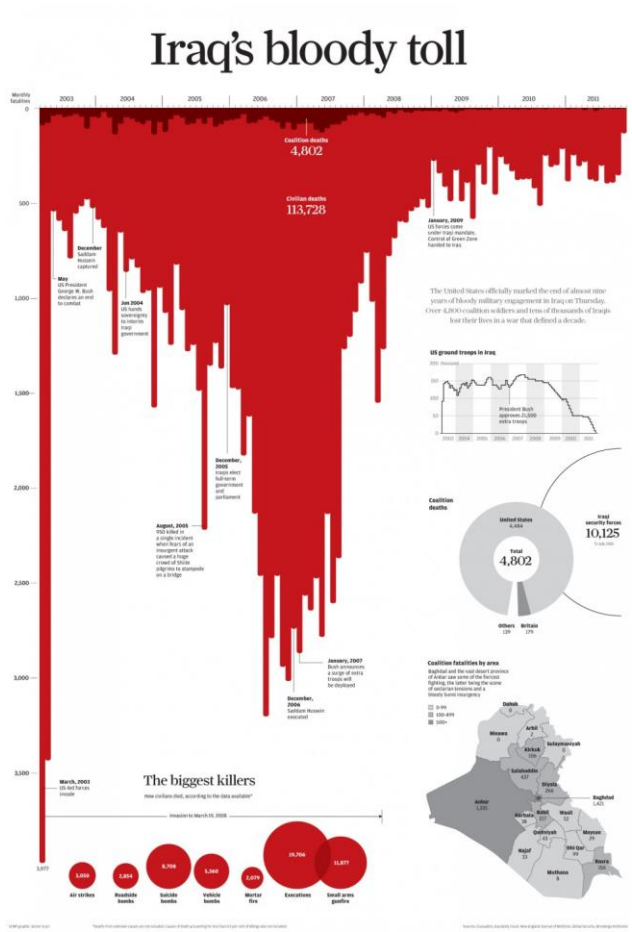
# Time not on an axis




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

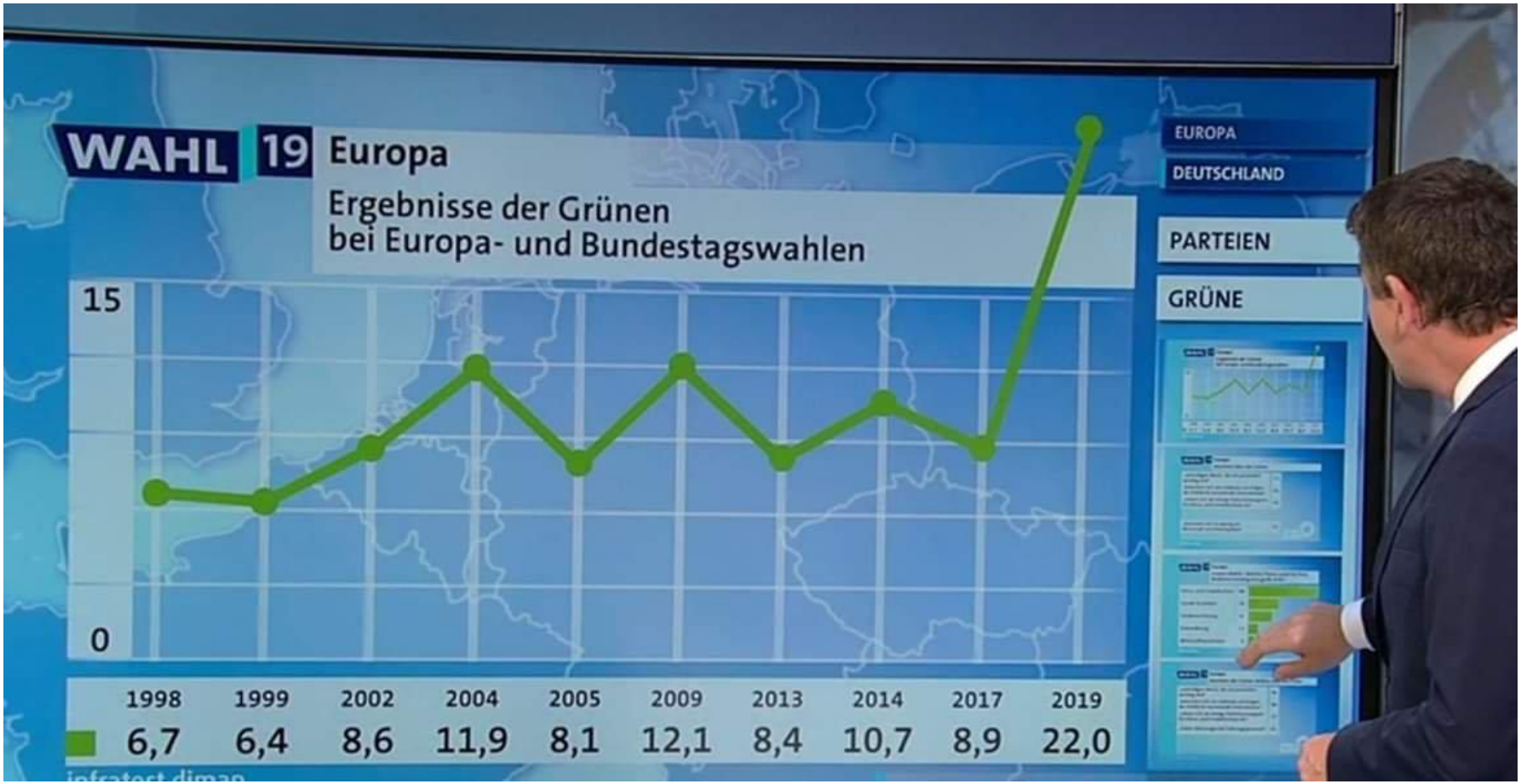
# Distorted axis: Inverted y axis



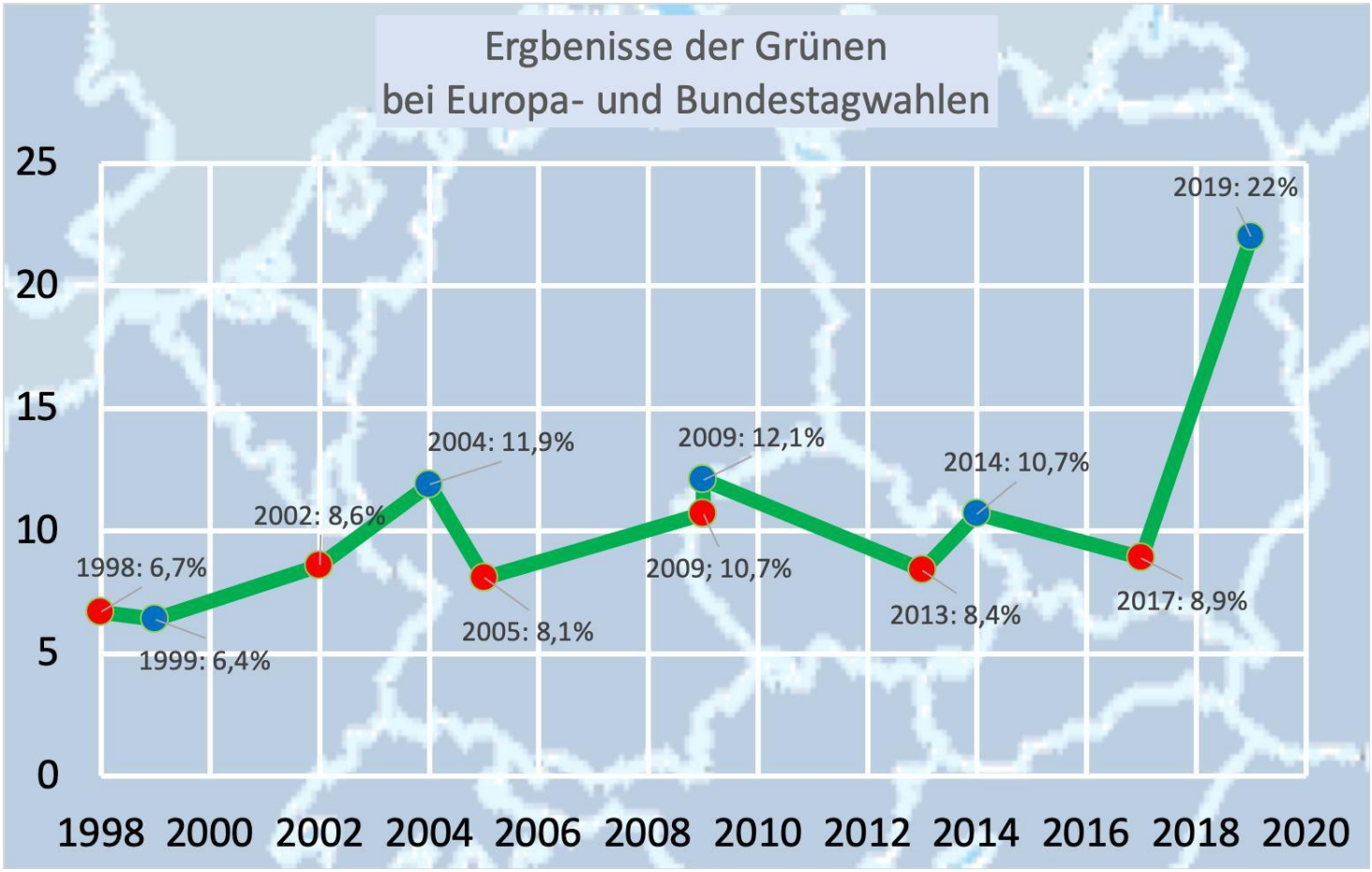

http://www.businessinsider.com/gun-deaths-in-florida-increased-with-stand-your-ground-2014-2
http://www.scmp.com/infographics/article/1284683/iraqs-bloody-toll

# Distorted axis



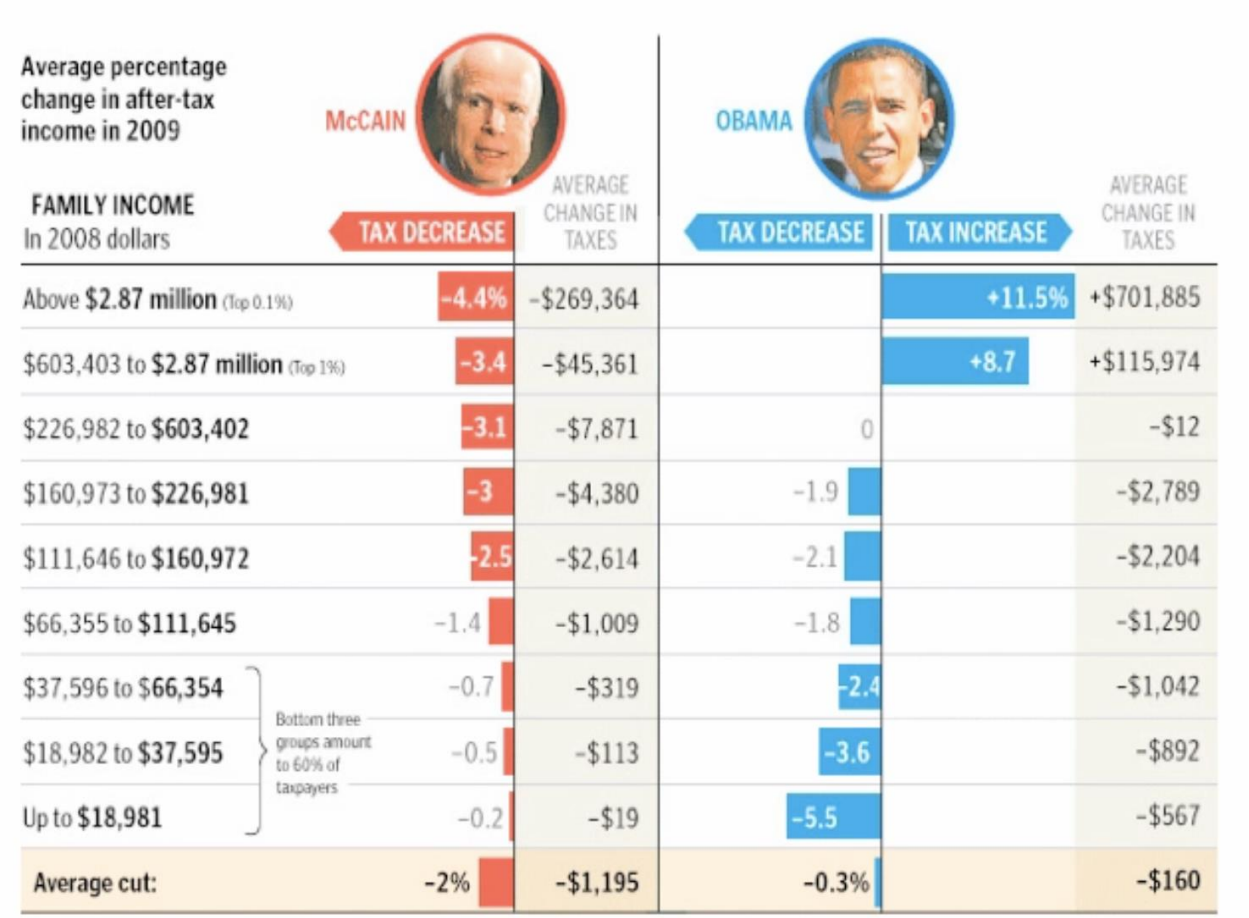

https://twitter.com/maartenzam/status/1132961446592172032?s=12

# Fixed axis

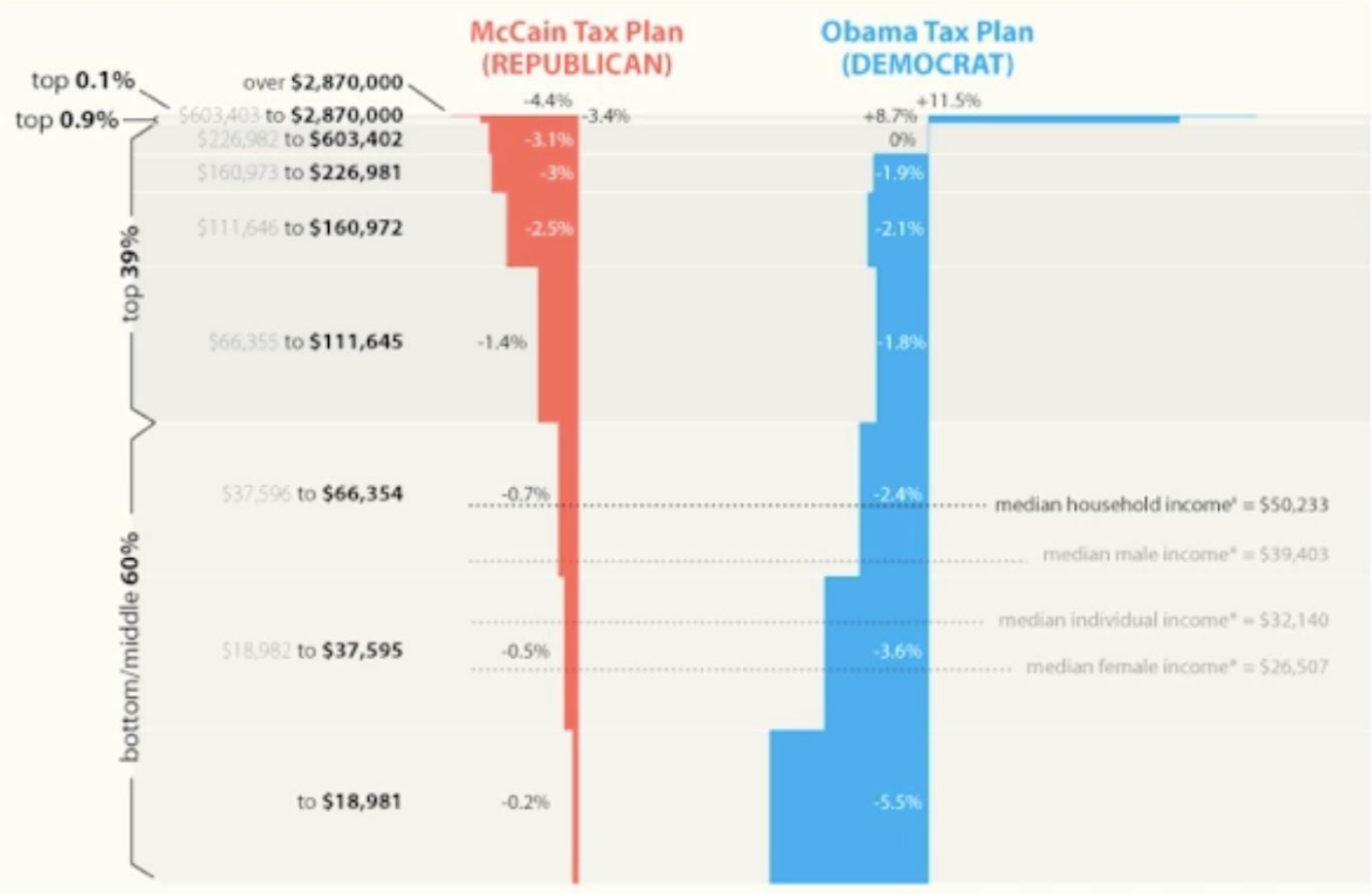



https://twitter.com/maartenzam/status/1132961446592172032?s=12

# Unequal intervals



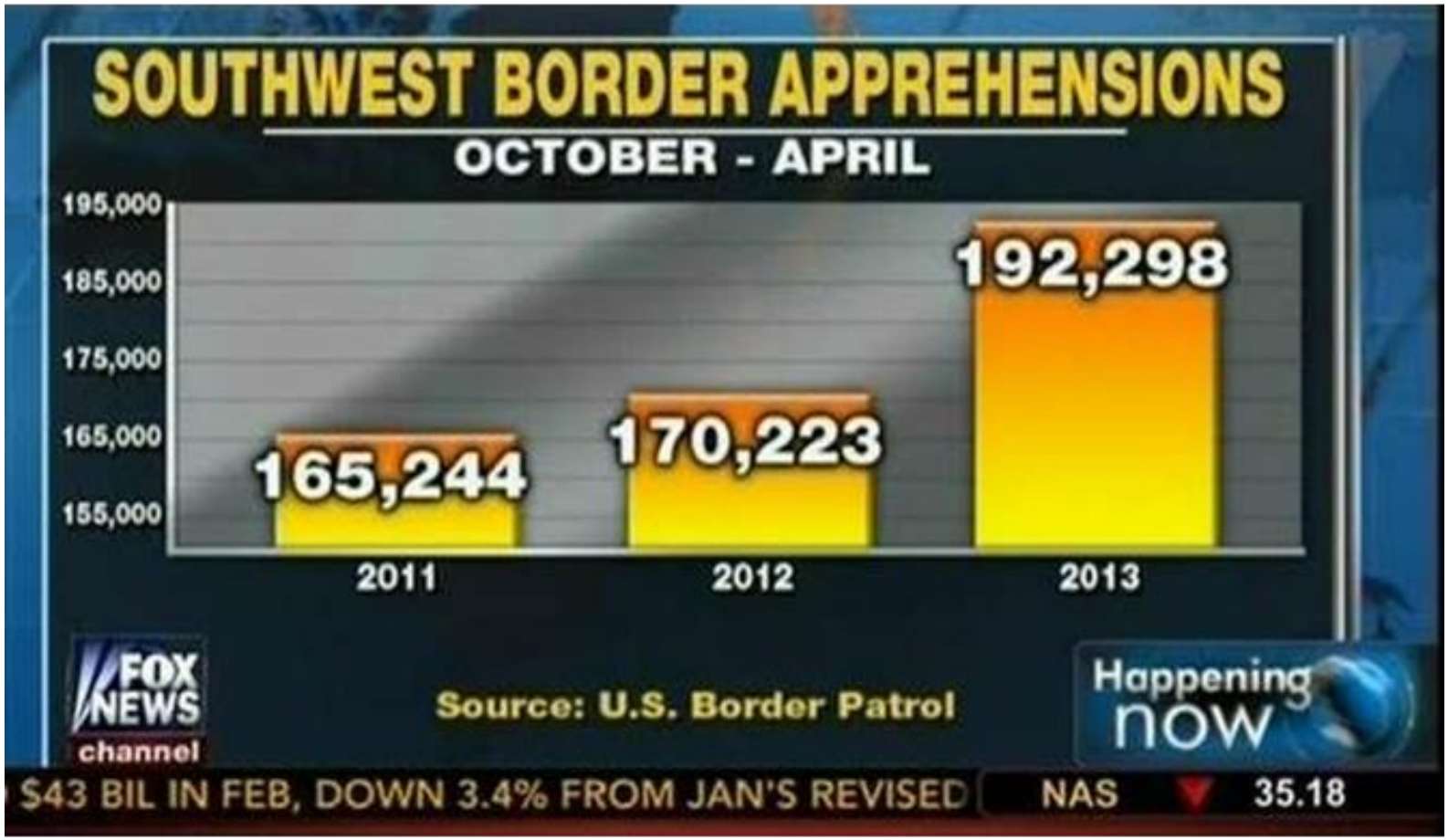
# Fixed intervals



# Over 100% pie chart

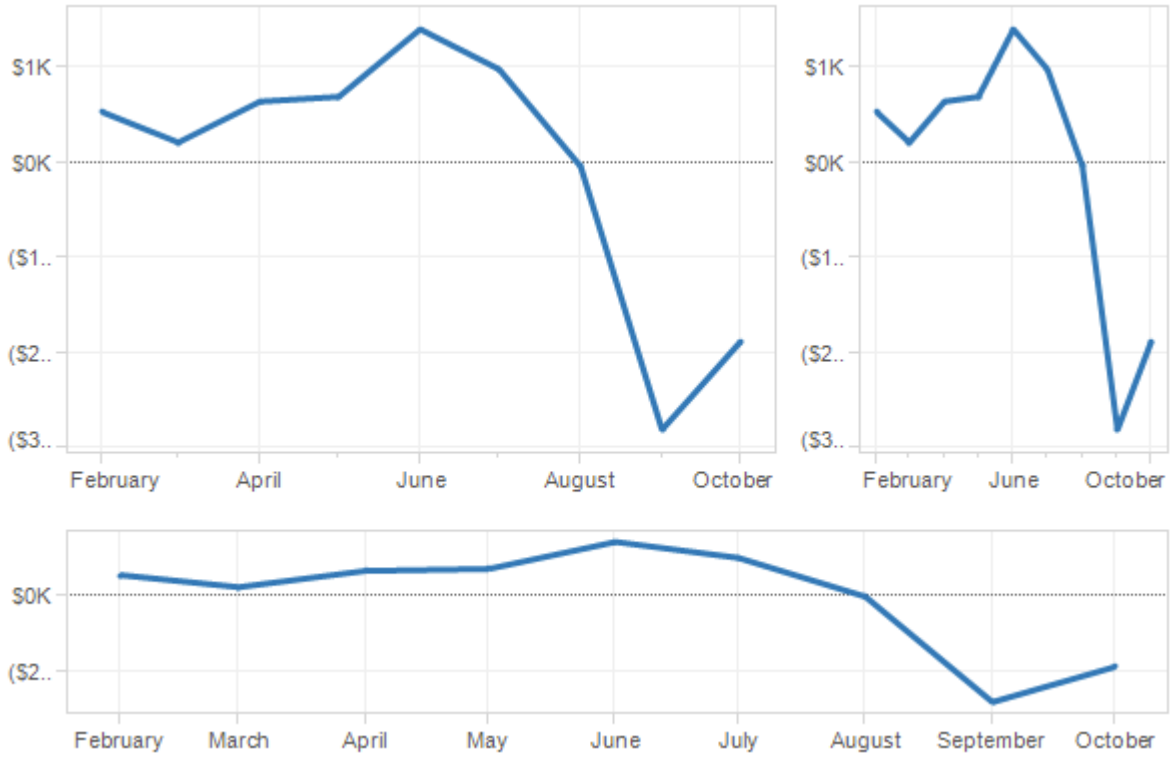

https://twitter.com/dergigi/status/1243315176000180224/

# Bar chart with truncated axis




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

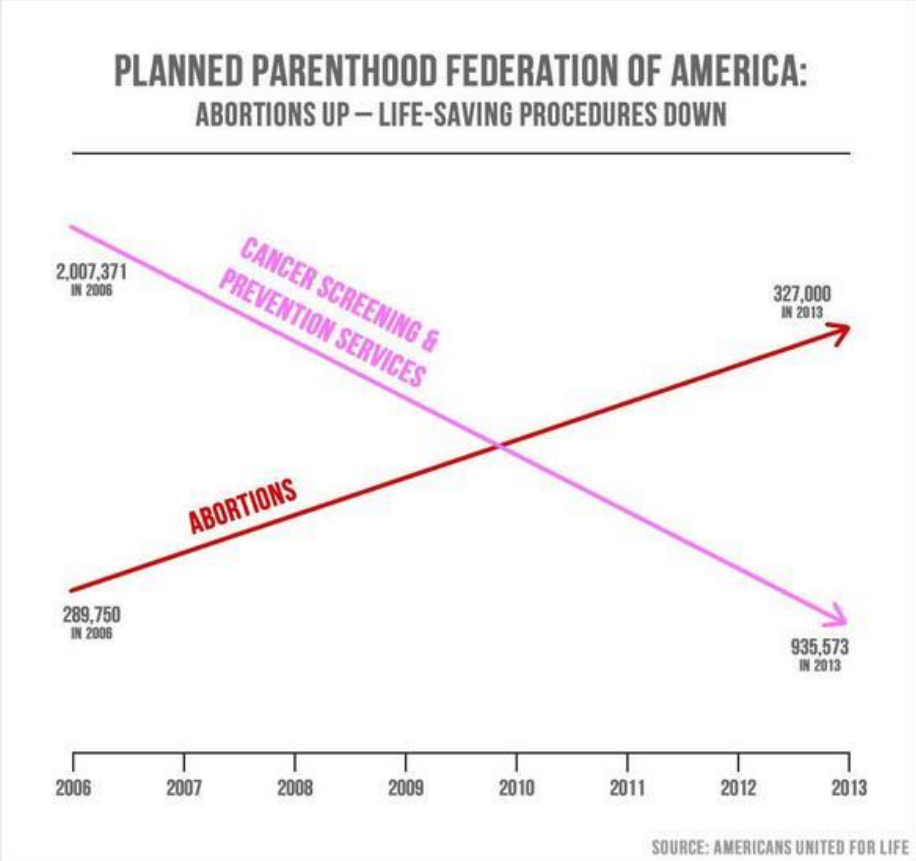
# Aspect ratio bias

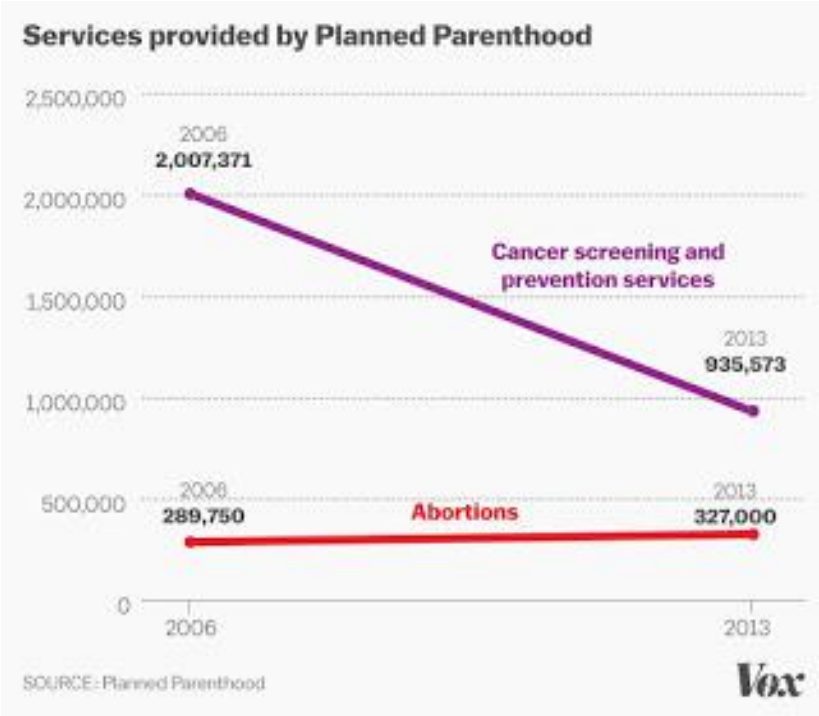



Banking to 45 Degrees

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

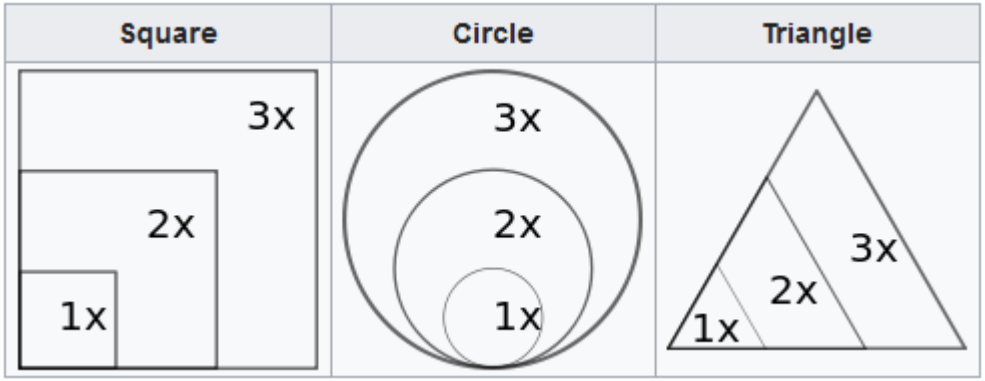
# Dual axes

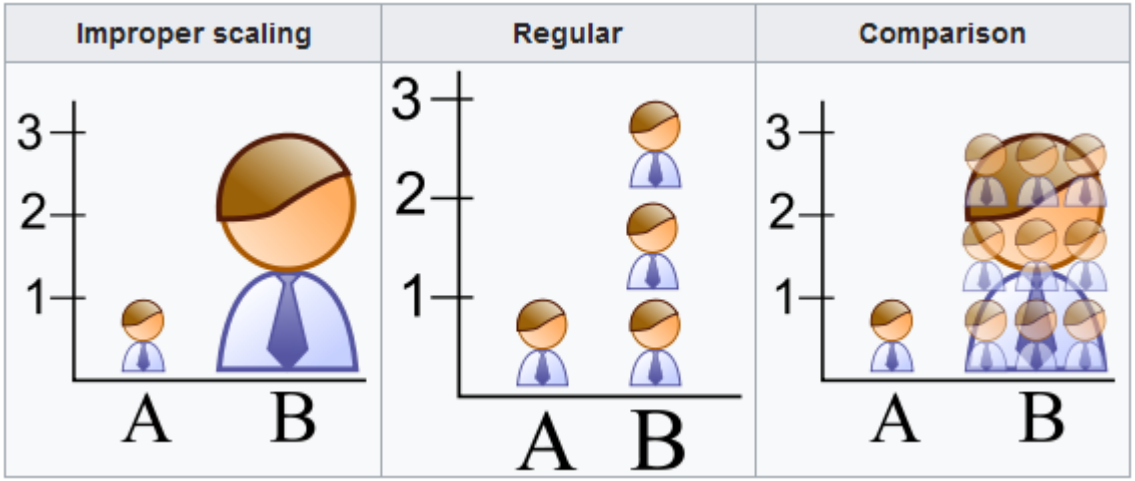



http://www.thefunctionalart.com/2015/10/if-you-see-bullshit-say-bullshit.html

# Improper scaling of areas/pictograms



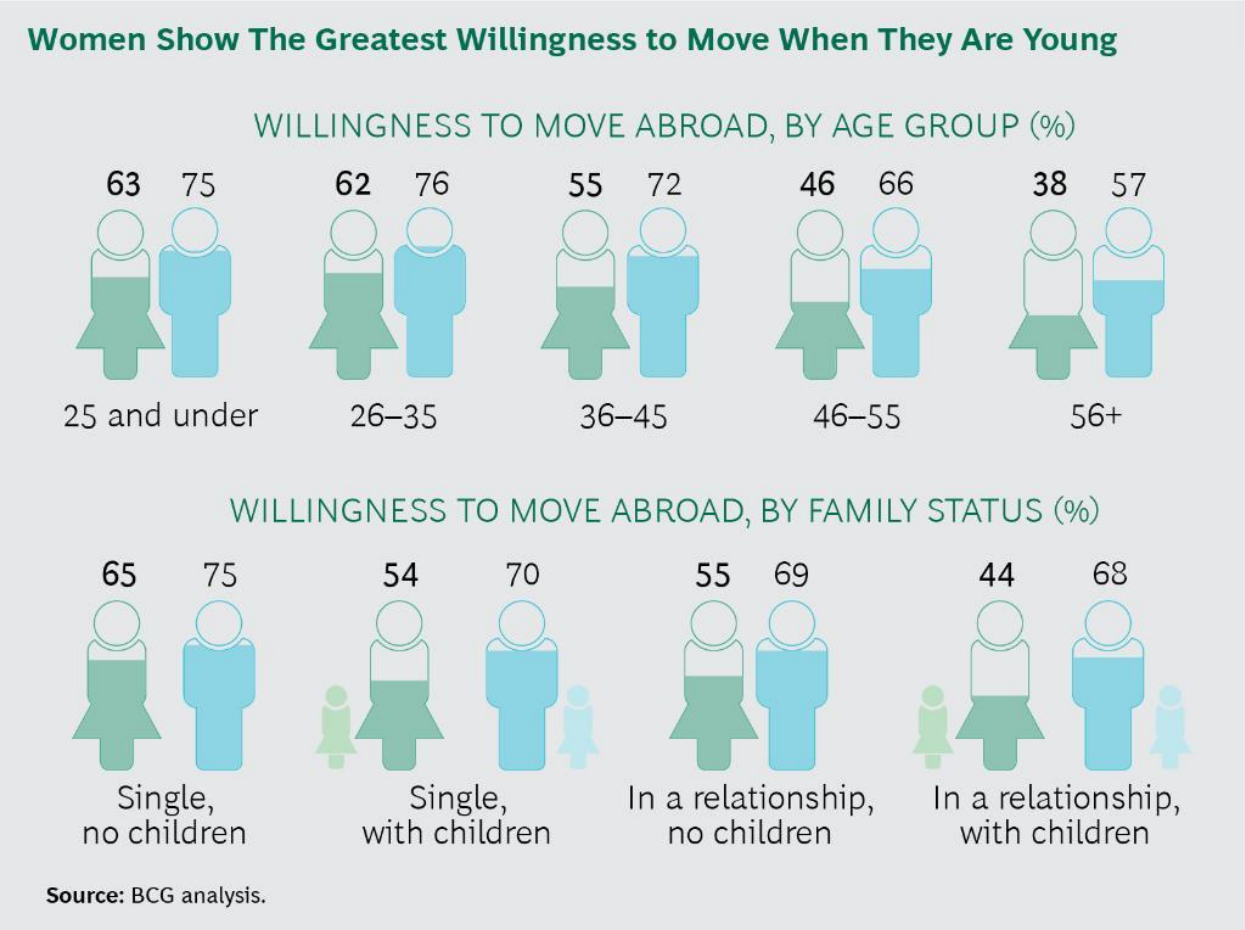

Even worse, if the elements are 3-D

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

# Improper scaling of areas/pictograms

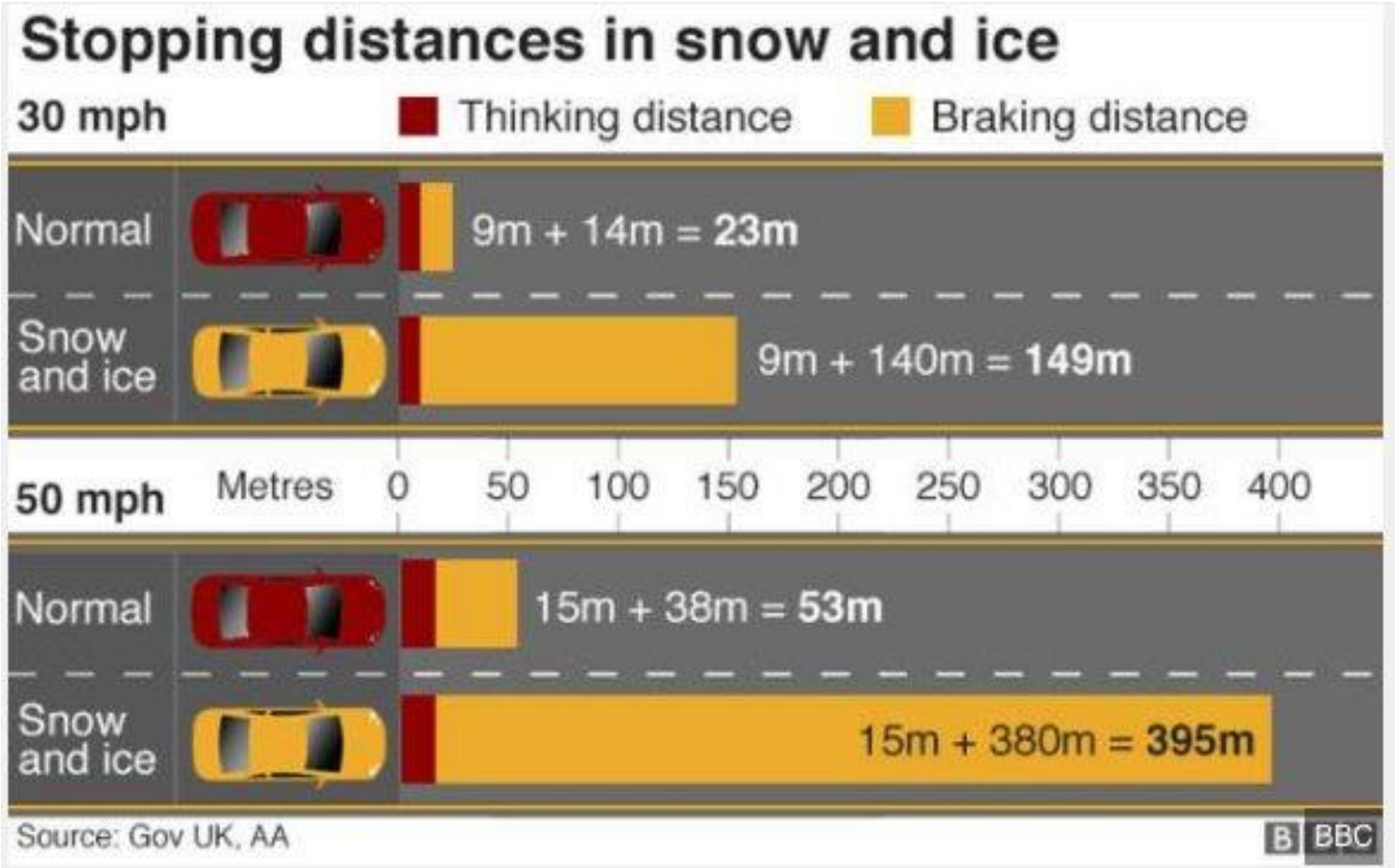

https://www.bcg.com/publications/2017/people-organization-behavior-culture-getting-the-most-from-diversity-dollars.aspx?linkId=59621326&redir=true

# Proper scaling of areas/pictograms



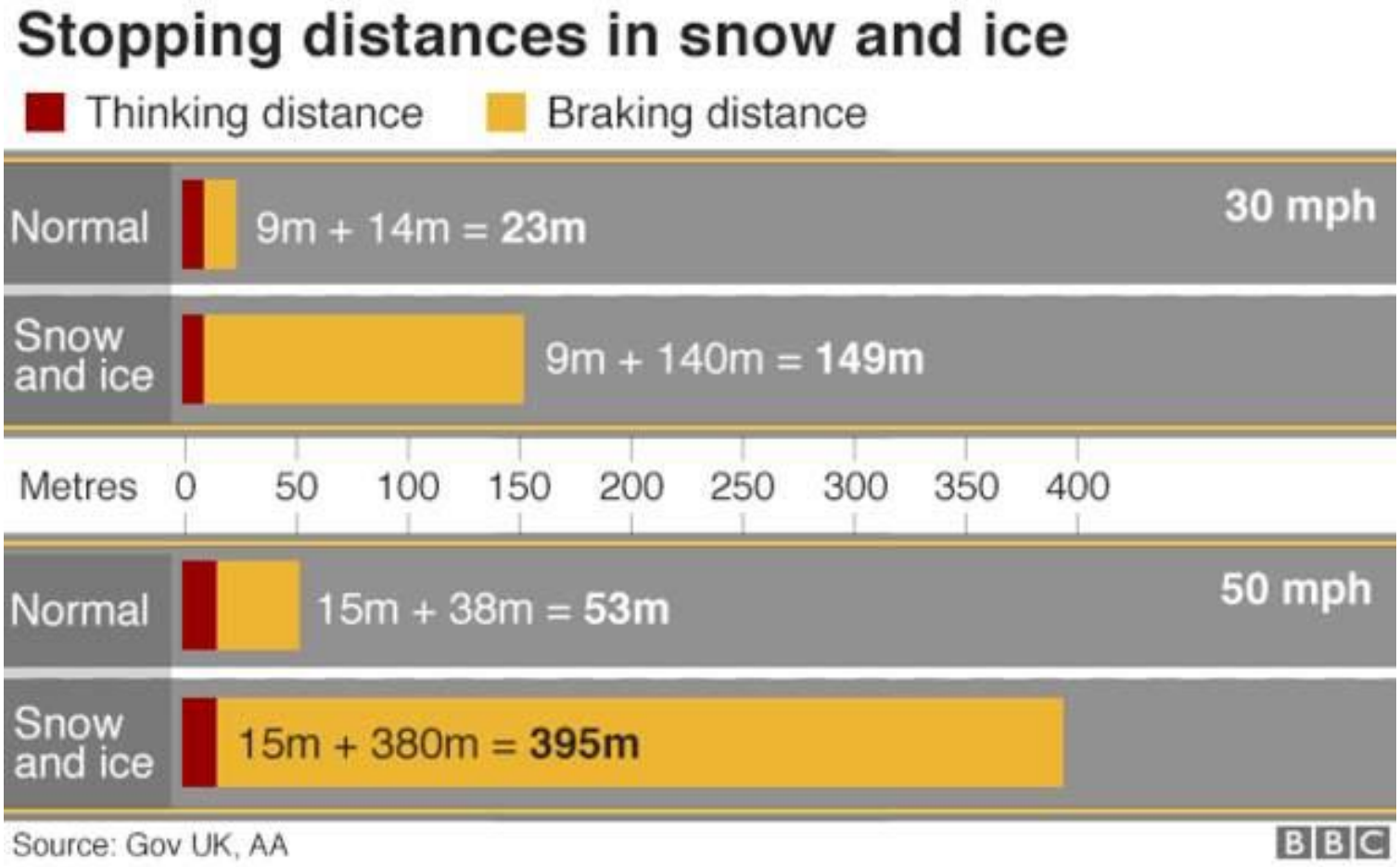

https://www.bcg.com/en-us/publications/2017/people-organization-leadership-talent-women-on-the-move.aspx?linkId=58878371&redir=true

# Improper scaling of areas/pictograms



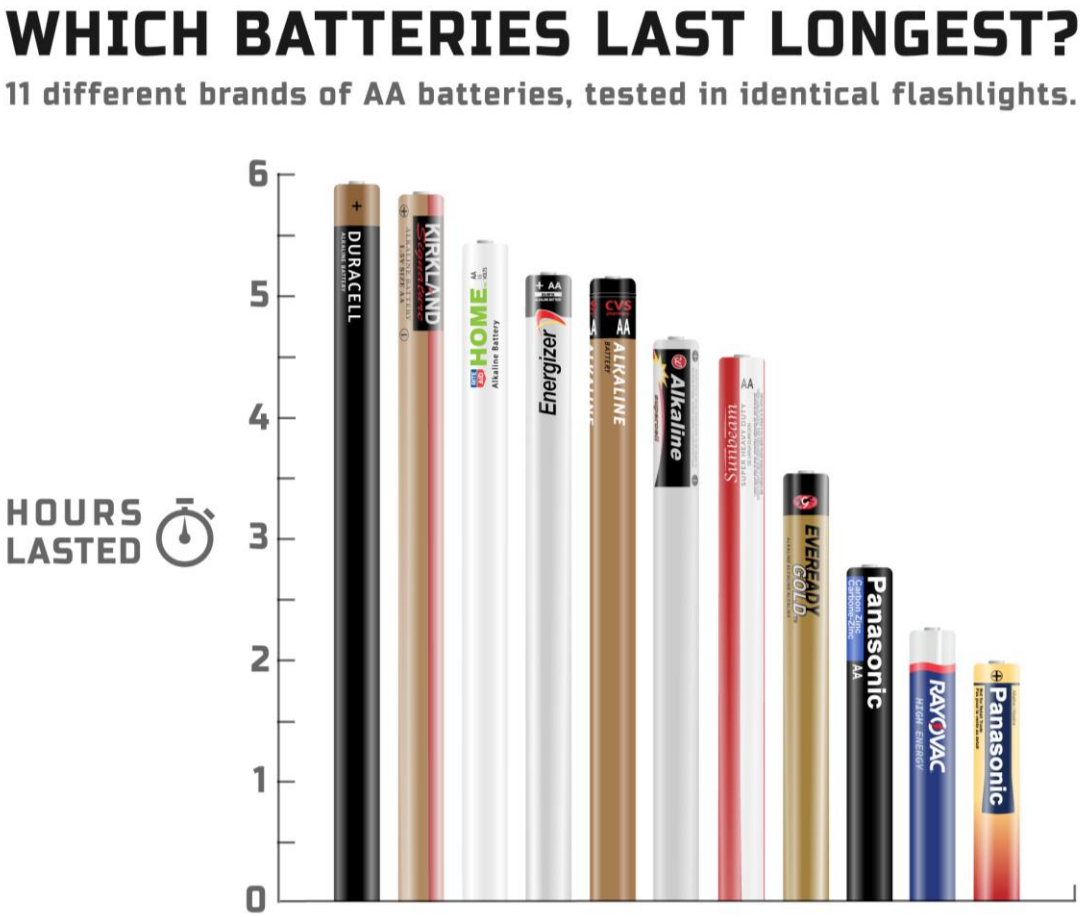

https://twitter.com/bengoldacre/status/1091326000384929793

# Improper scaling of areas/pictograms – fixed



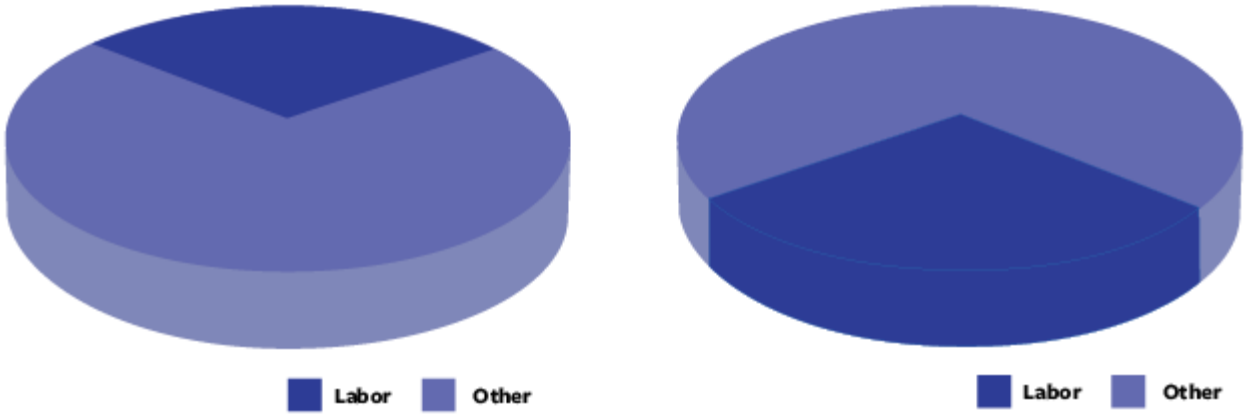

https://t.co/dMvGp5R8VH

# Proper scaling of areas/pictograms




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

# Unnecessary 3-D




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

# Unnecessary 3-D

# 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.

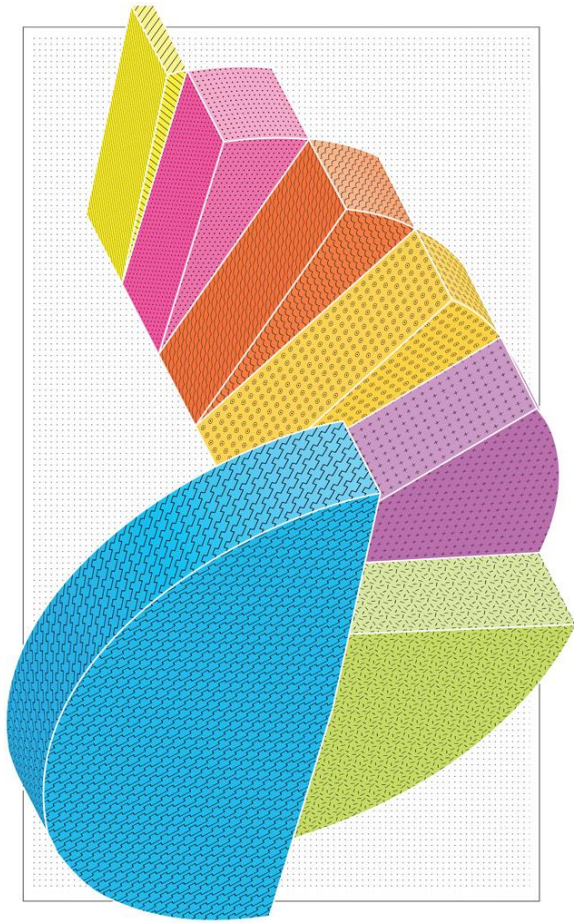
**5%**
Spontaneous Moment
Don't overprepare. Tease the guy in the front row ("You could light up a village with this guy's eyes"). Commend the stagehand who handles the human brain you brought.

**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.

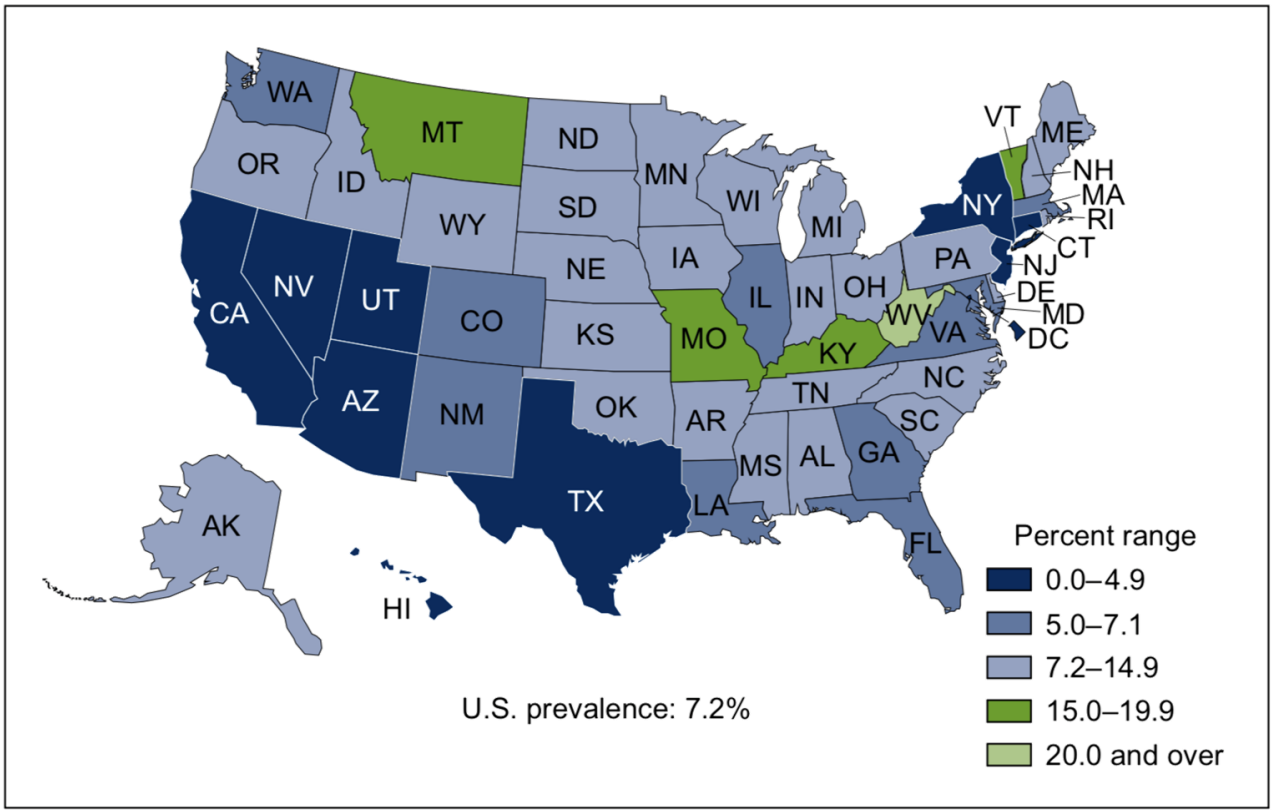
**49%**
Contrarian Thesis
Wait a sec—we should be playing *more* videogames? The more choices we have, the worse off we are? TED is where conventional wisdom goes to die.



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

# Improper categorization (and color choice)

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

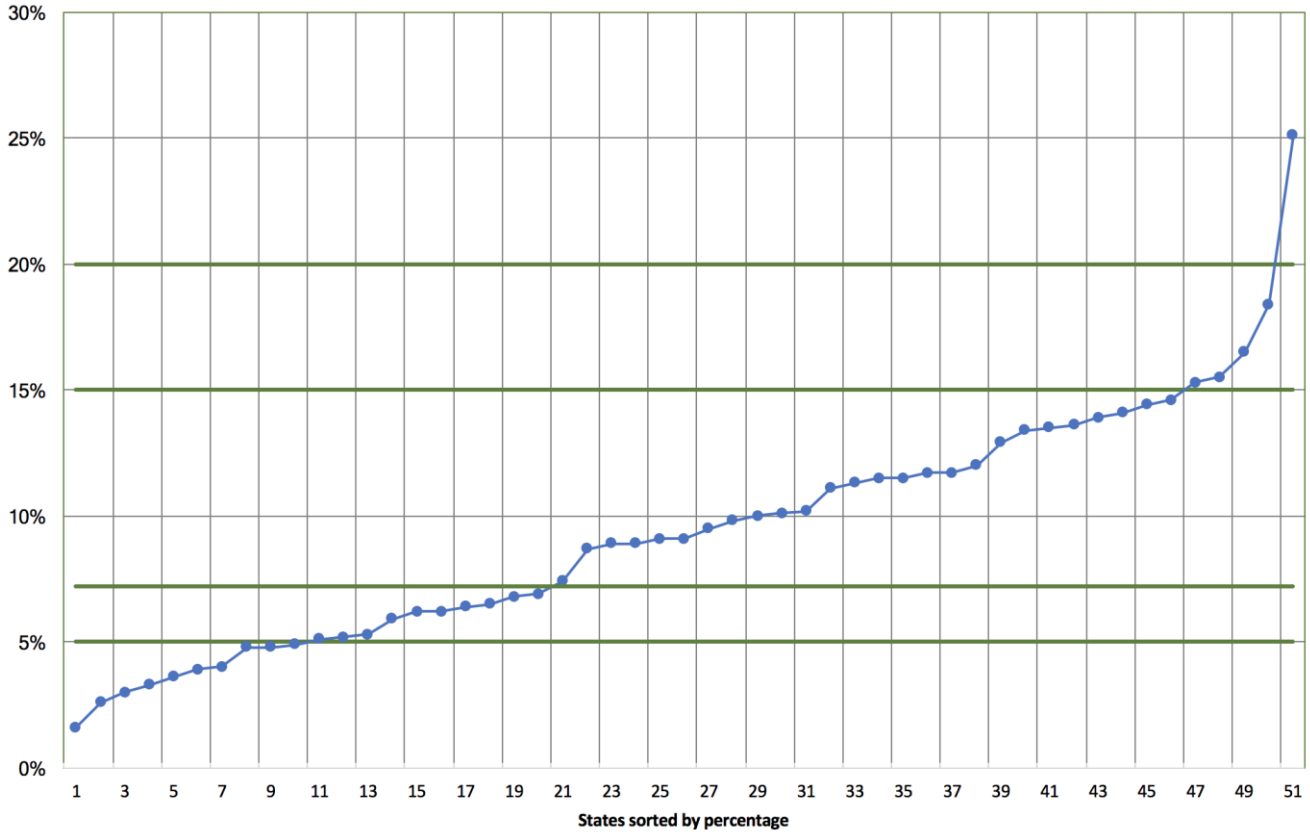



4.9
2.1
7.7
4.9

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.

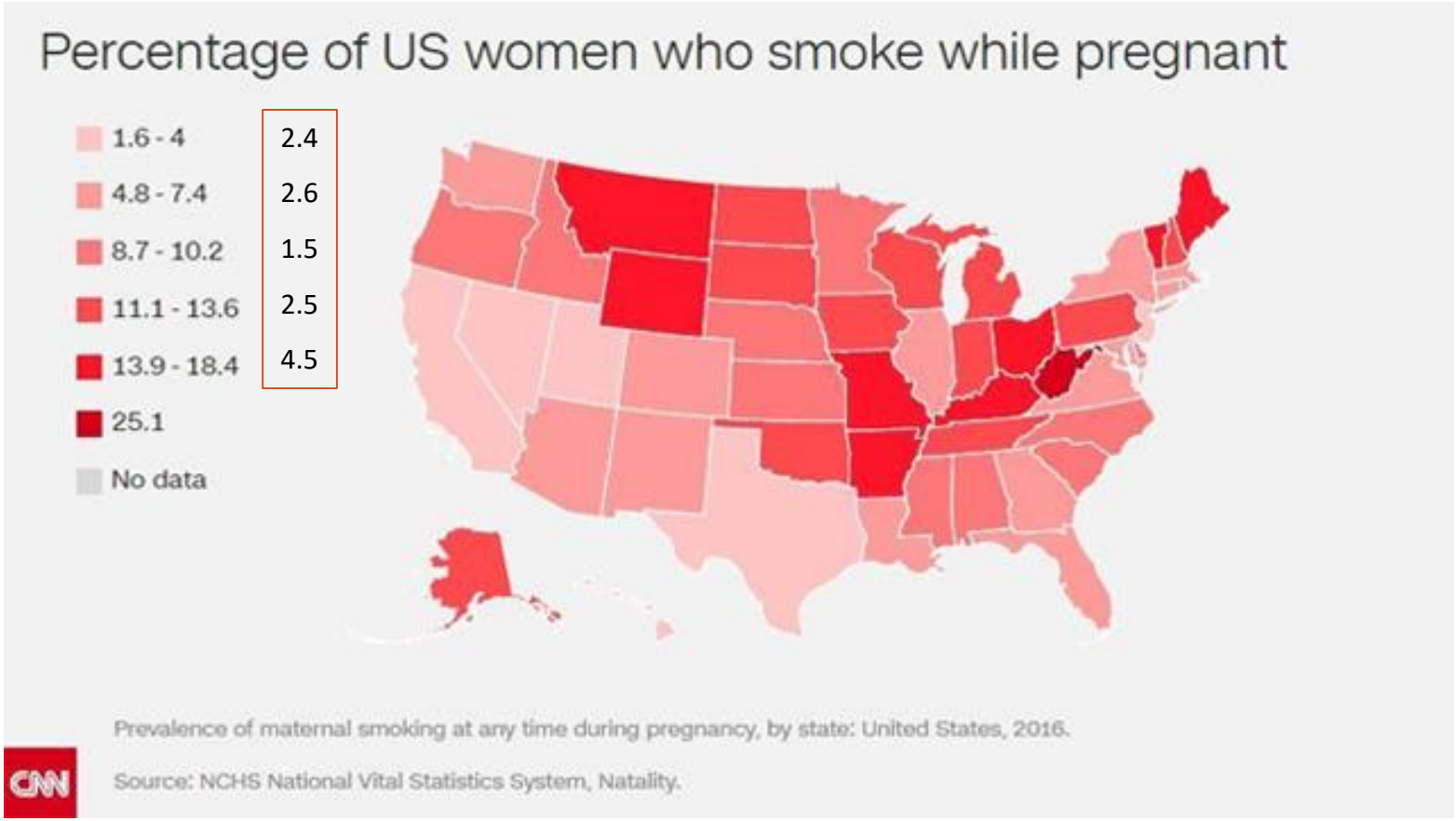
https://www.cdc.gov/nchs/data/databriefs/db305.pdf

# Improper categorization (and color choice)



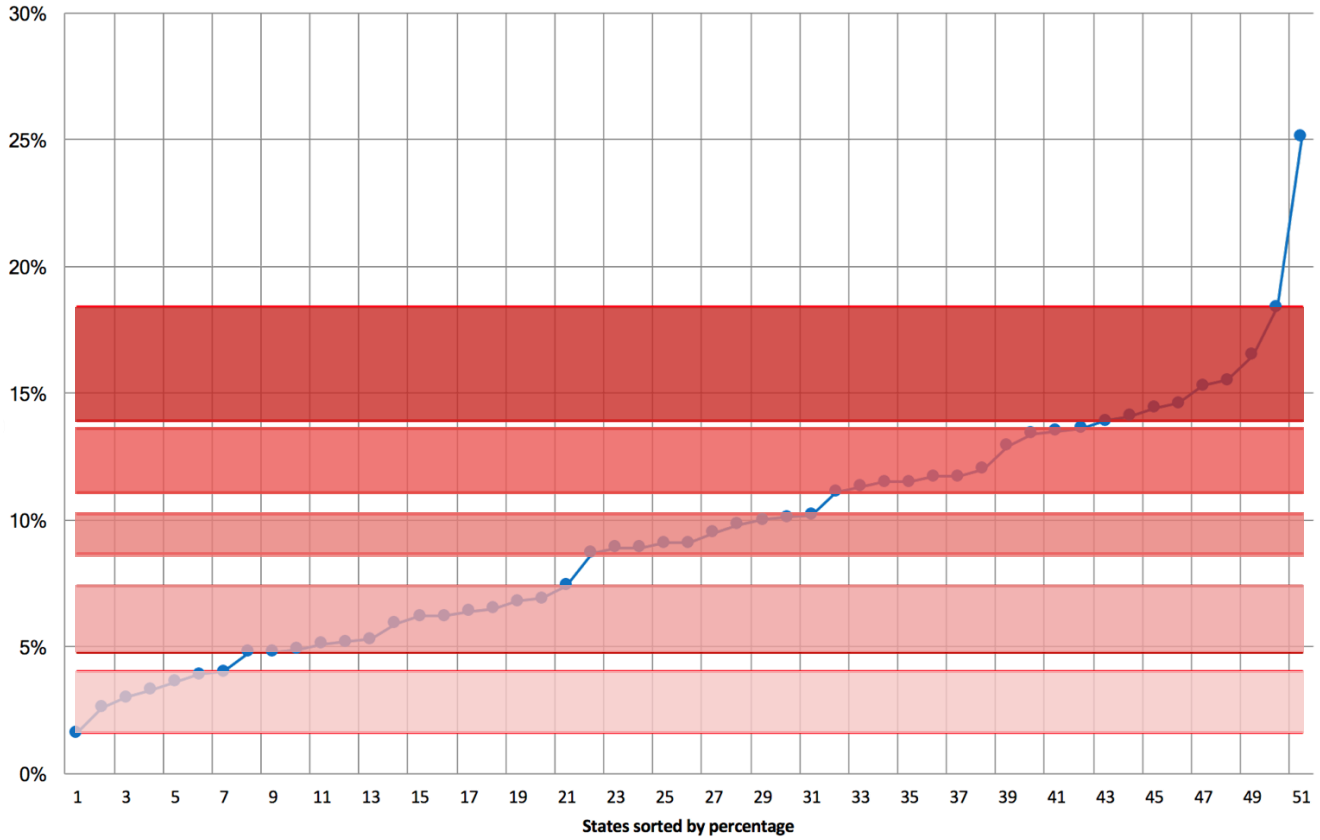

https://www.cdc.gov/nchs/data/databriefs/db305_table.pdf#1

# Improper categorization



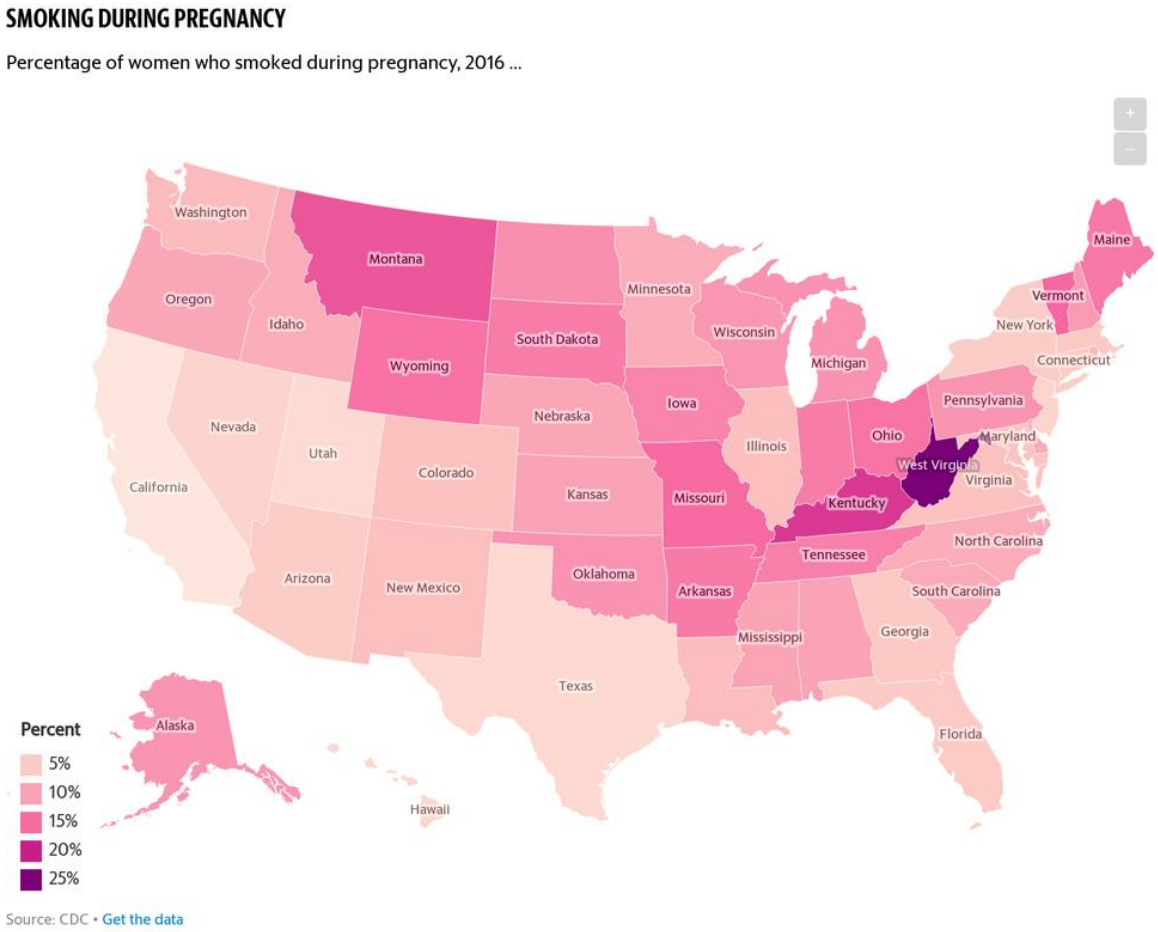

https://meredith.images.worldnow.com/images/16207225_G.png?auto=webp&disable=upscale&width=800

# Improper categorization




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

# Improper categorization




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

# Oversimplifying

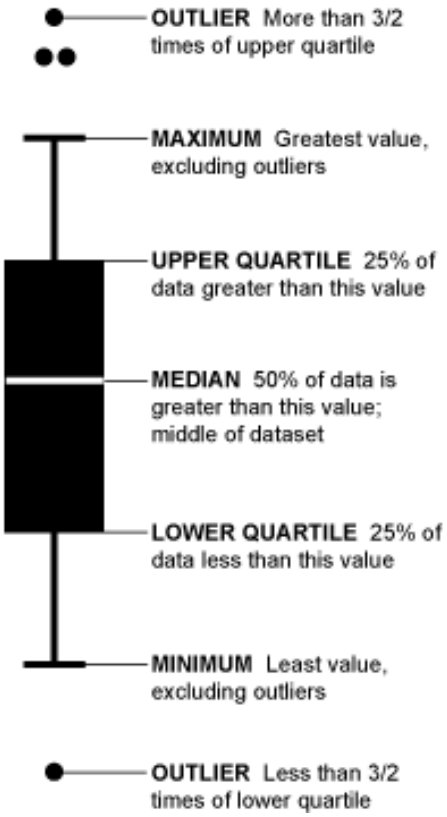
Clarify, not simplify!

*To clarify, add detail.*

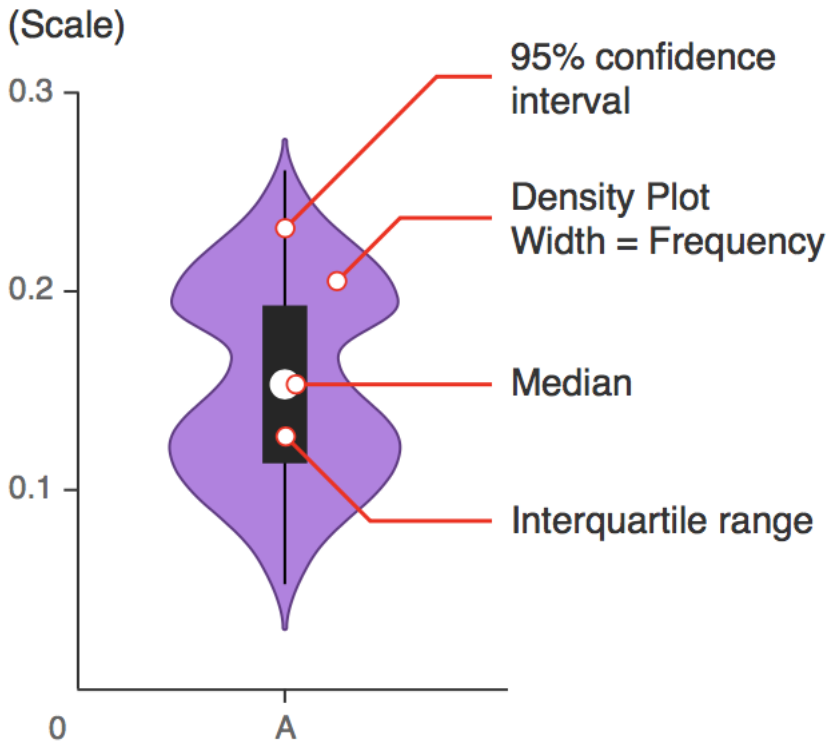
Edward Tufte

# Box plot vs. violin plot

Box (and whisker) plot

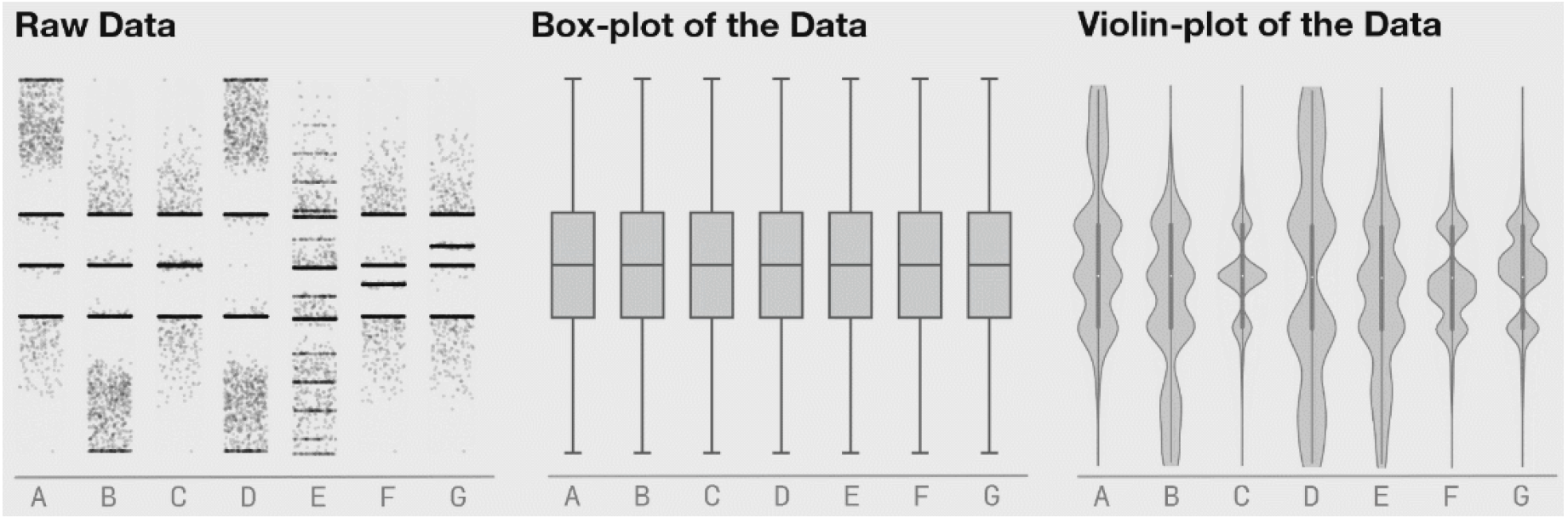



Violin plot




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

# Oversimplifying




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

# How charts lie?

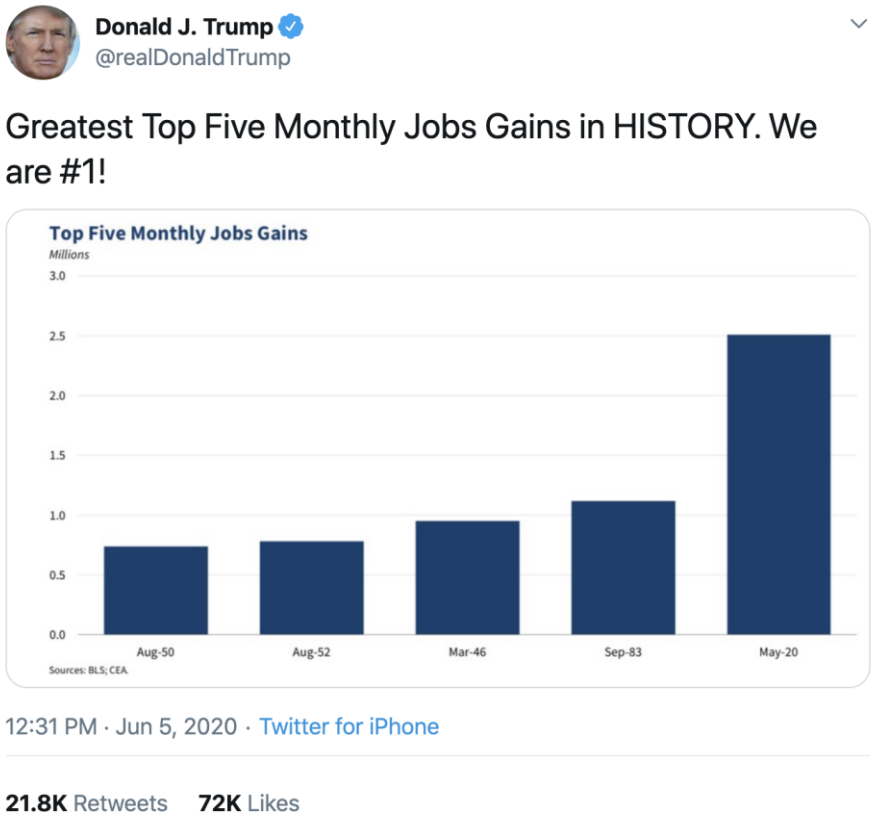
# Cherry-picking data

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

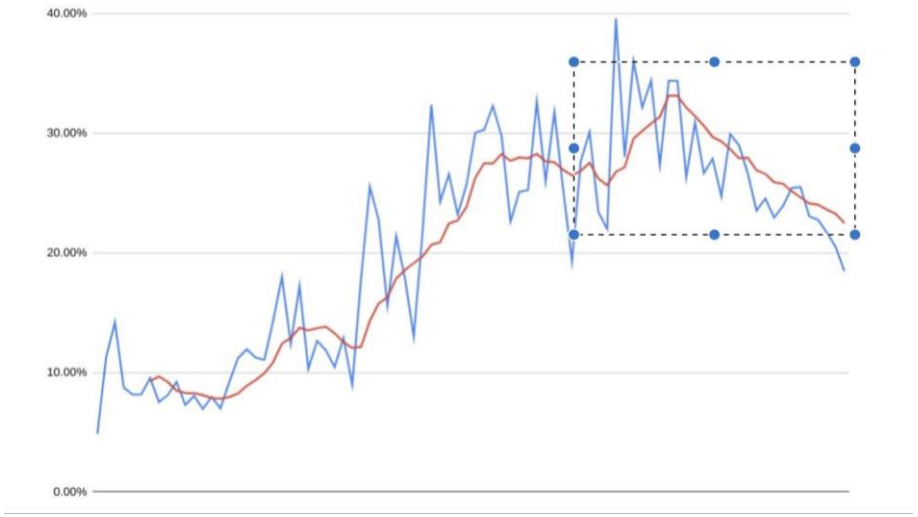
- Hiding (unfavorable) data
- Concealing existing patterns
- Suggesting patterns that are not there

Correlation ≠ causation

# Hiding (unfavorable) data

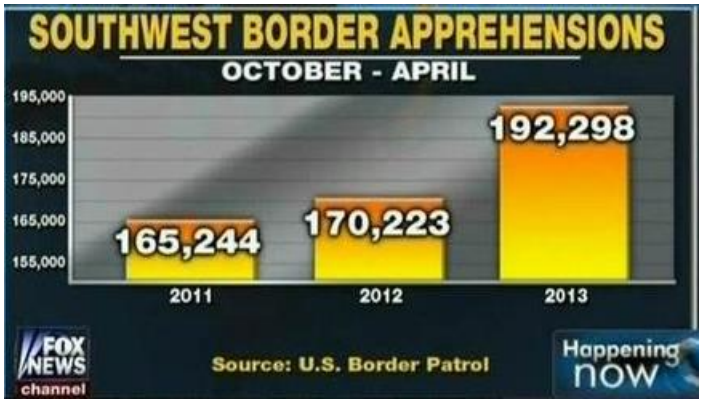


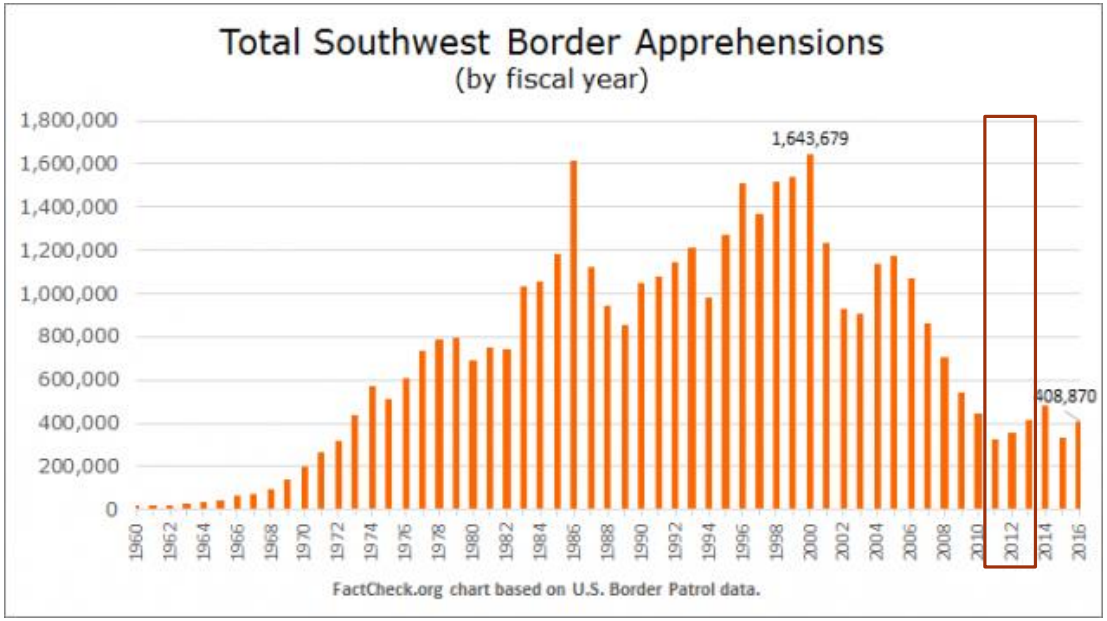

# Hiding (unfavorable) data


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

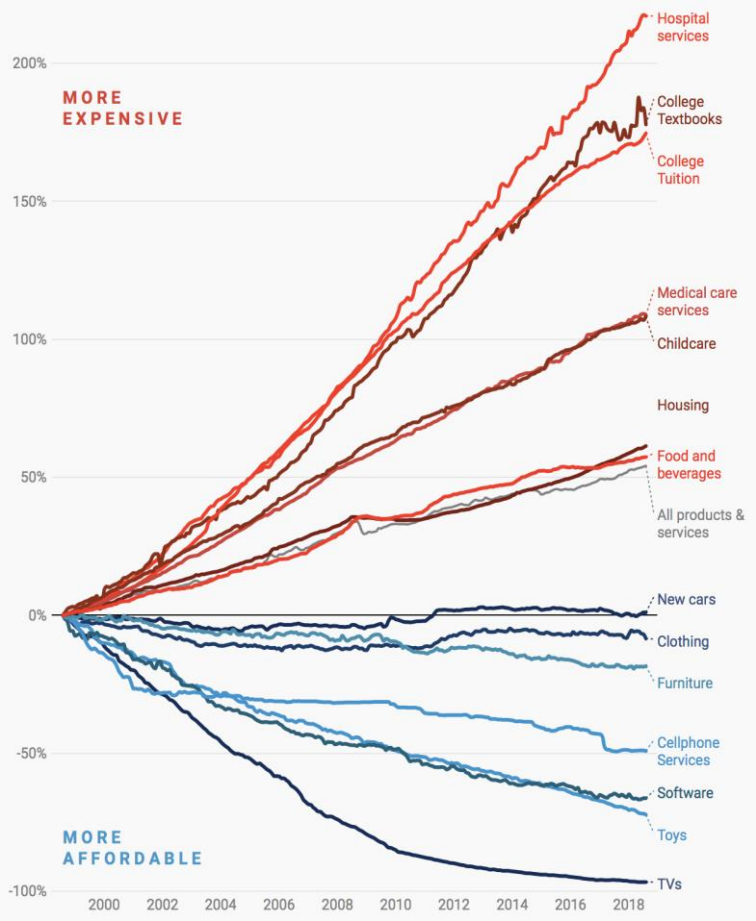
# Concealing existing patterns

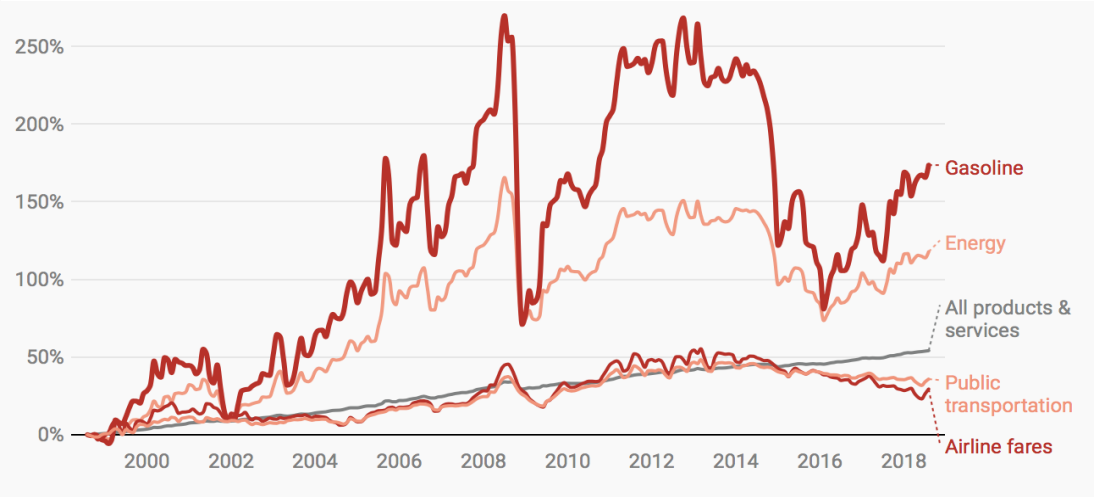



https://www.businessinsider.com/the-27-worst-charts-of-all-time-2013-6#welcome-to-fox-where-the-line-graphs-are-made-up-and-the-points-dont-matter-12

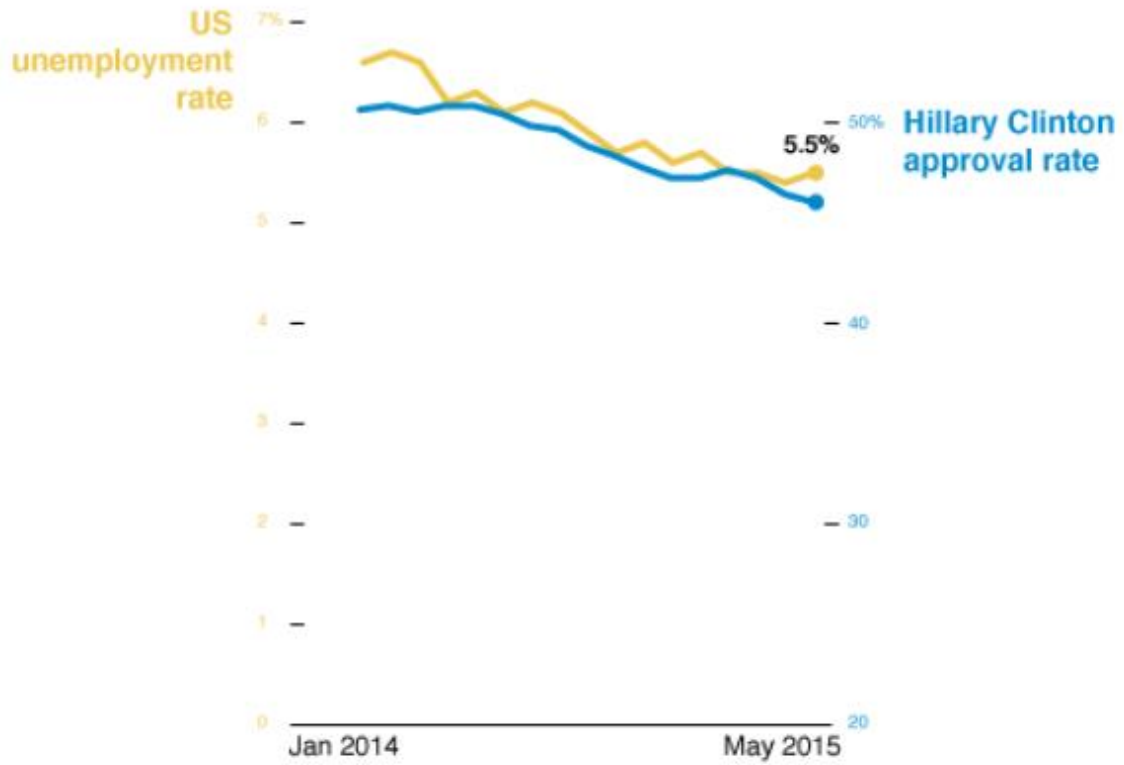
# Concealing existing patterns




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

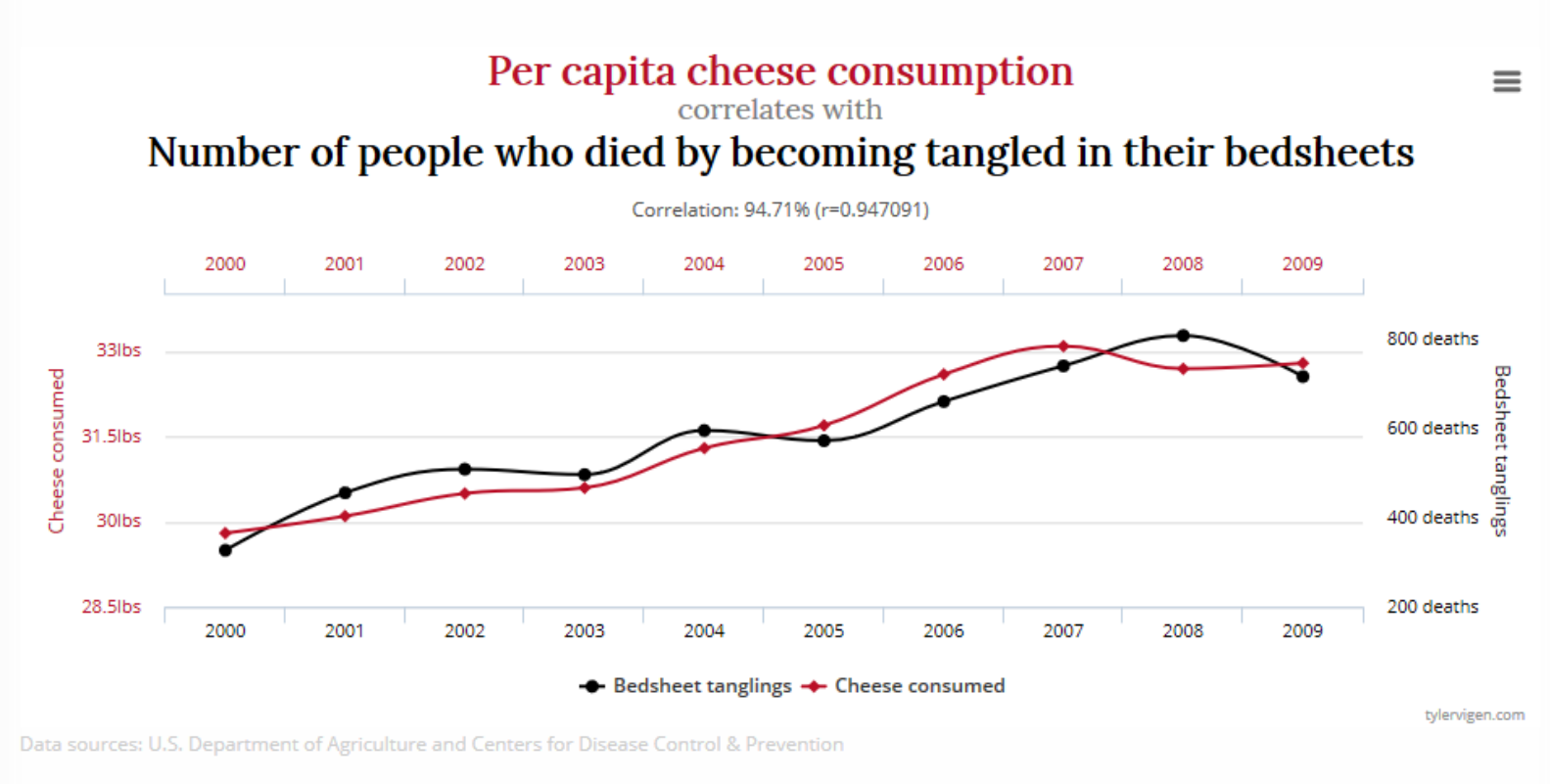
# Suggesting patterns that are not there




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

# Suggesting patterns that are not there

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



# How charts lie?

# Ignoring uncertainty

- Misrepresenting uncertainty
- Concealing uncertainty

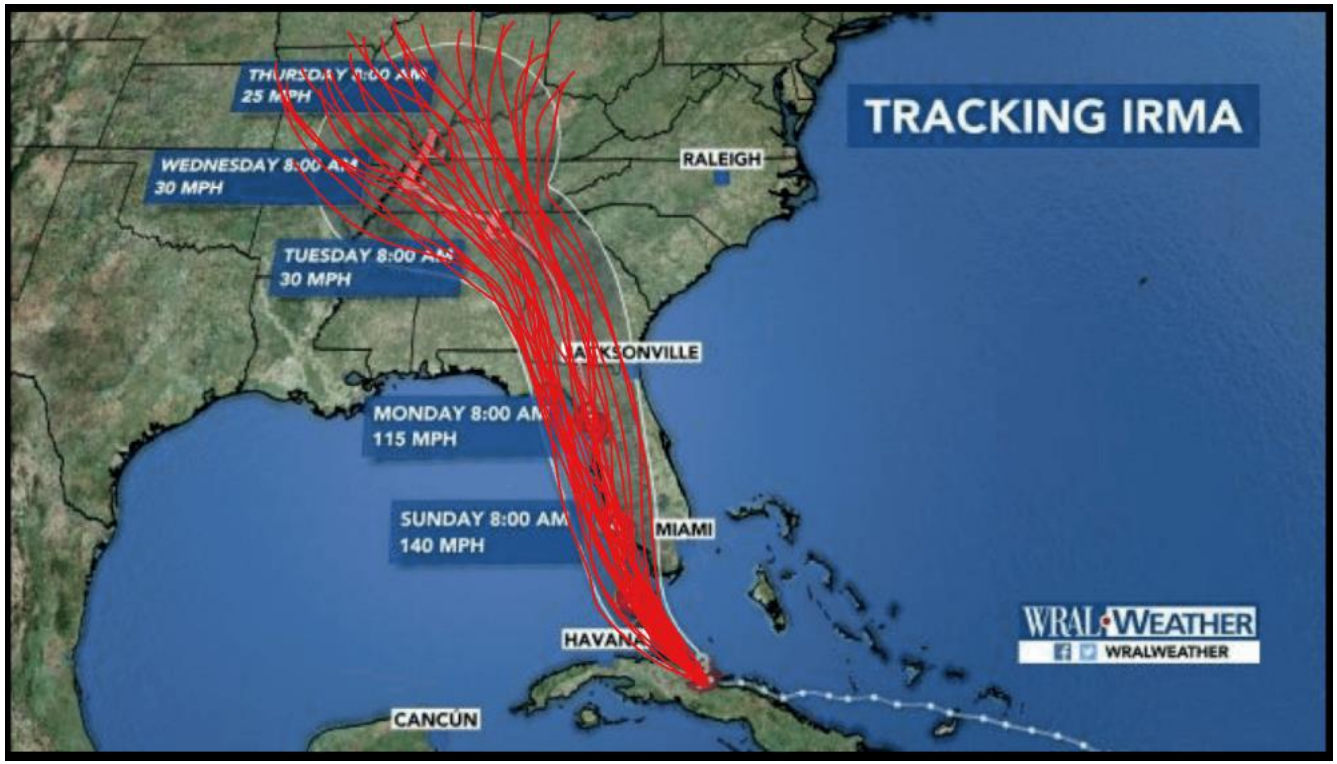
# Misrepresenting uncertainty

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

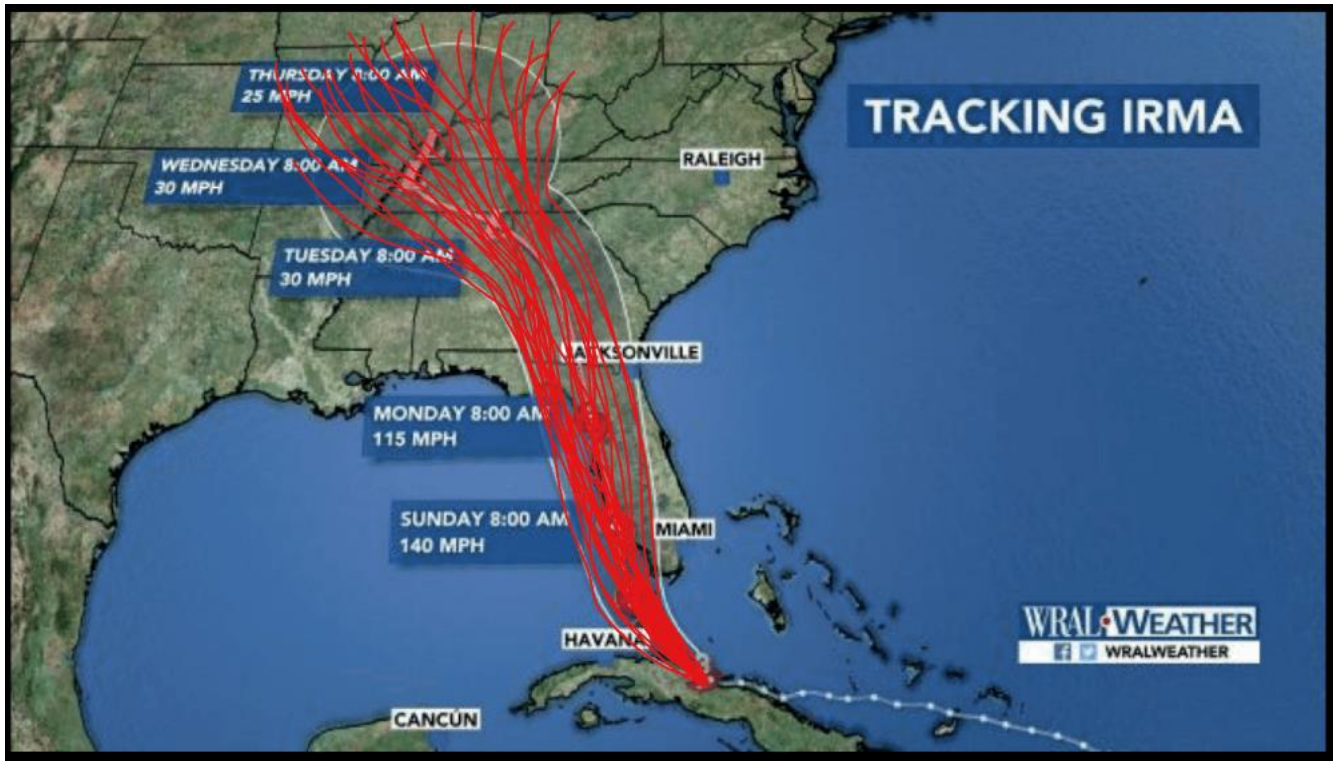
# Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted


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

# 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

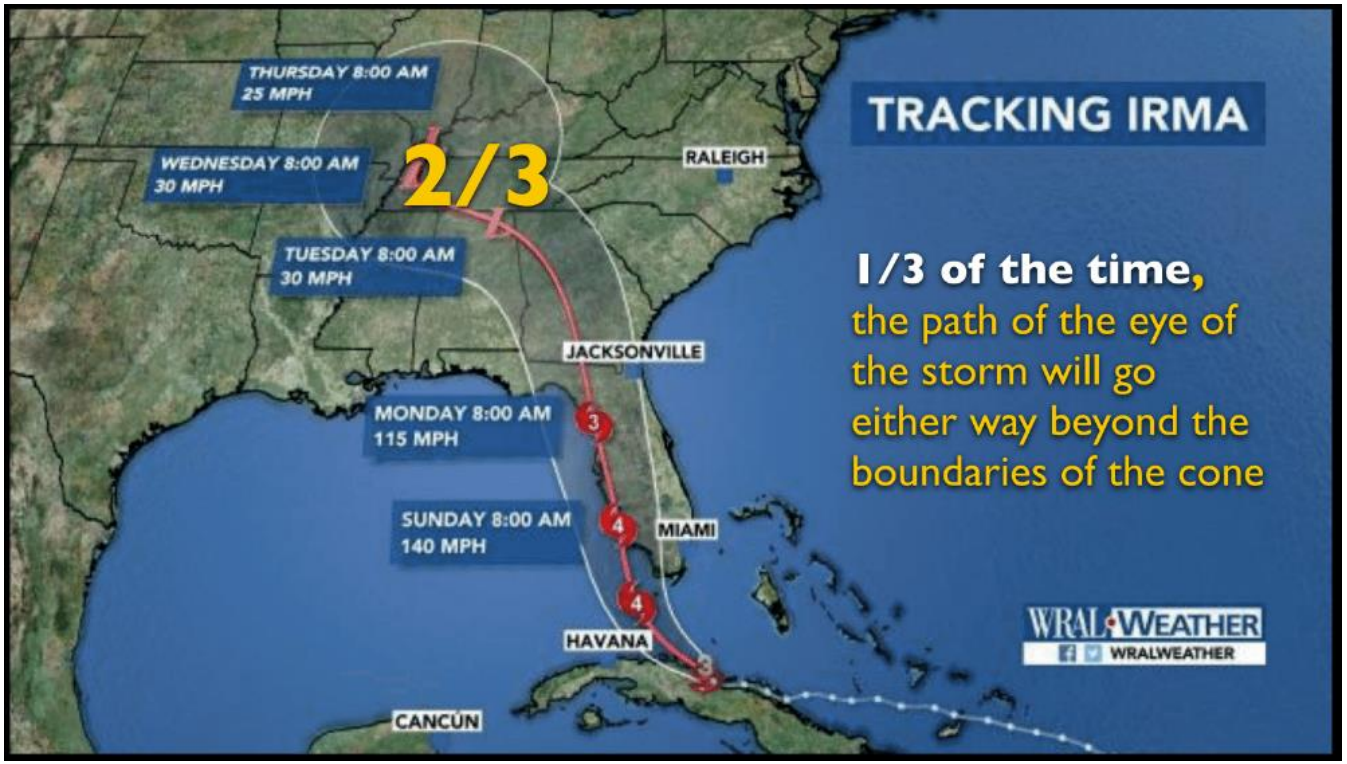
# Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted


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

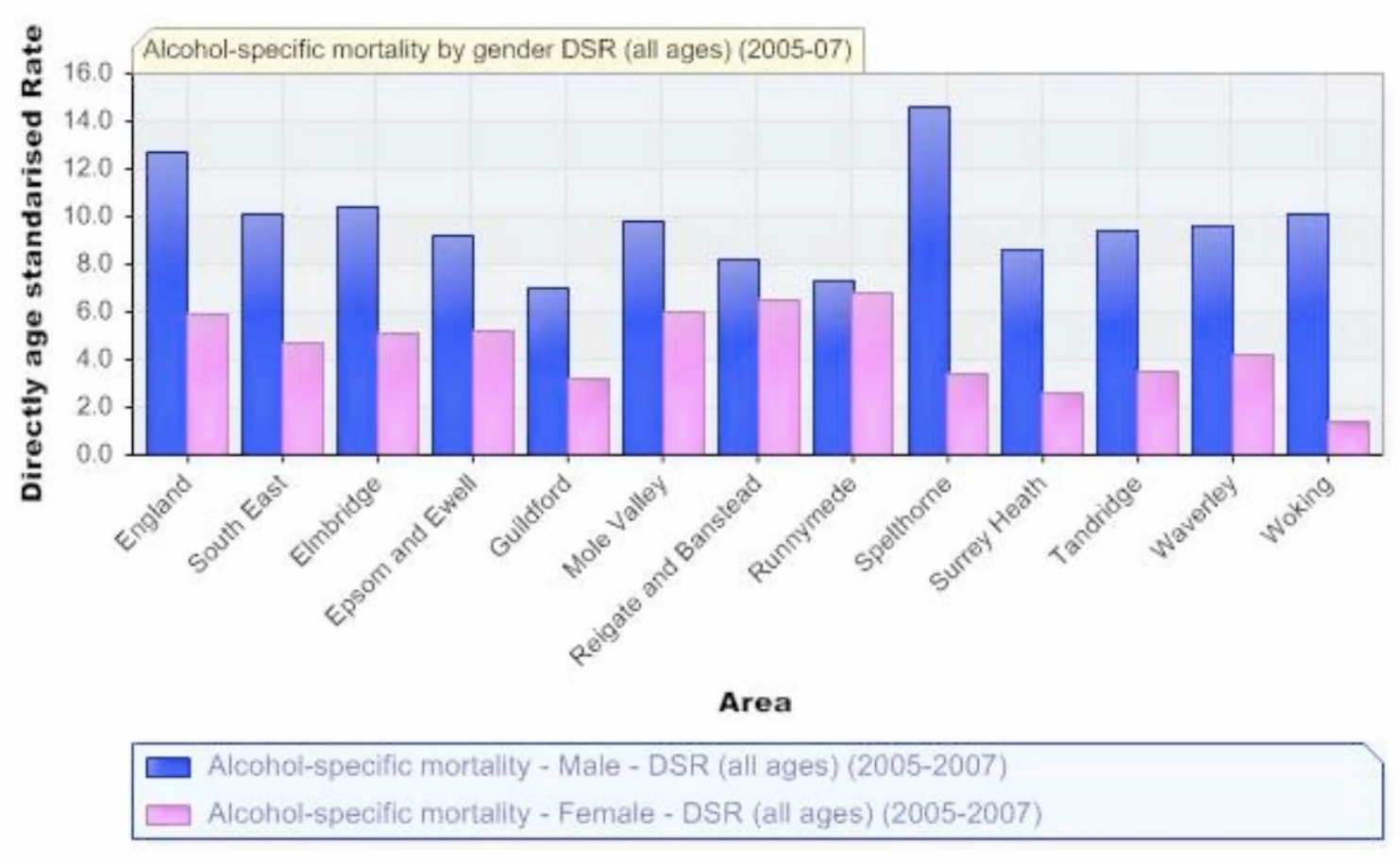
# Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted

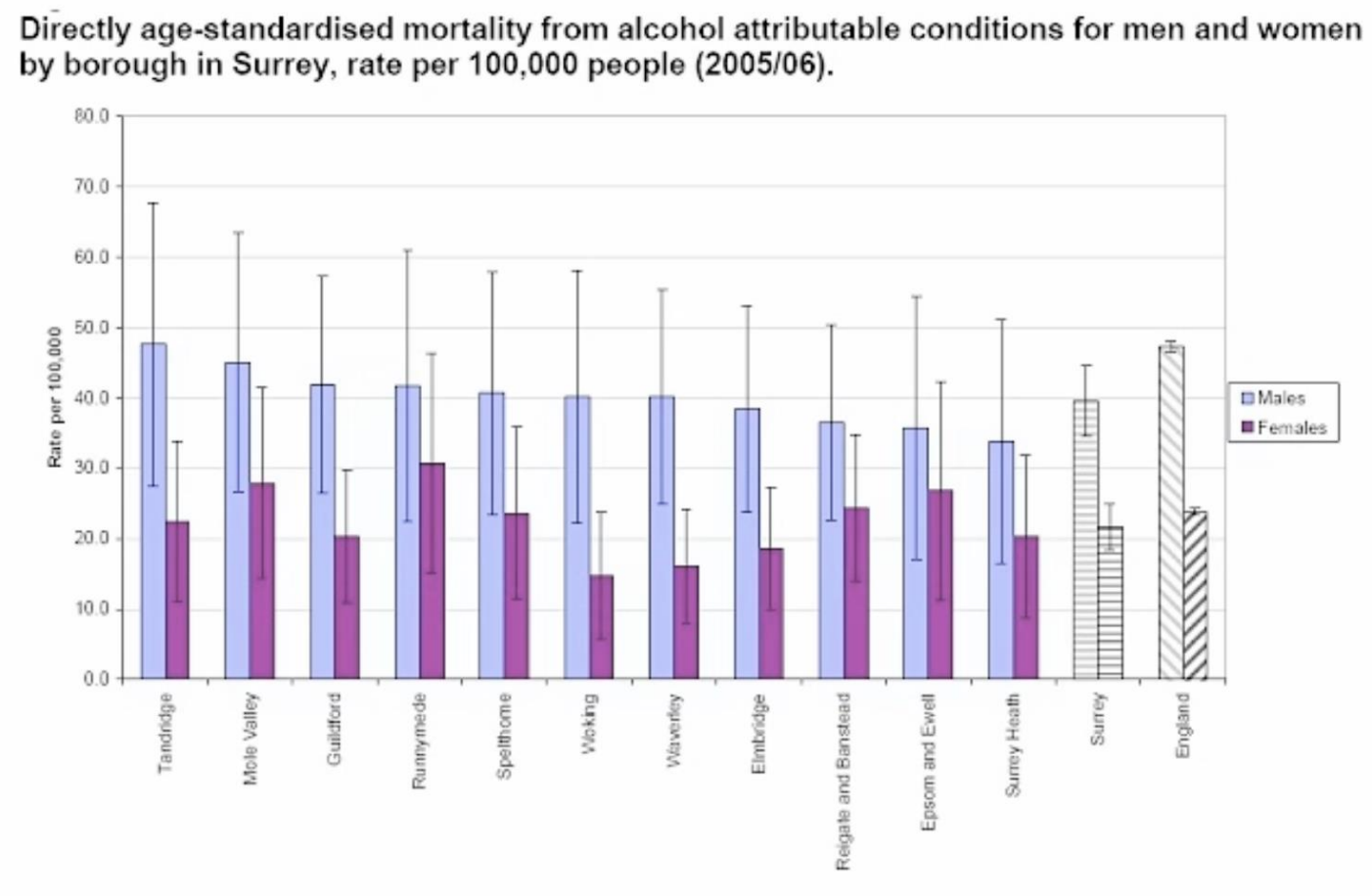



https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

# Concealing uncertainty



# Concealing uncertainty



# How charts lie?


Dubious data

Misrepresenting data

Cherry-picking data

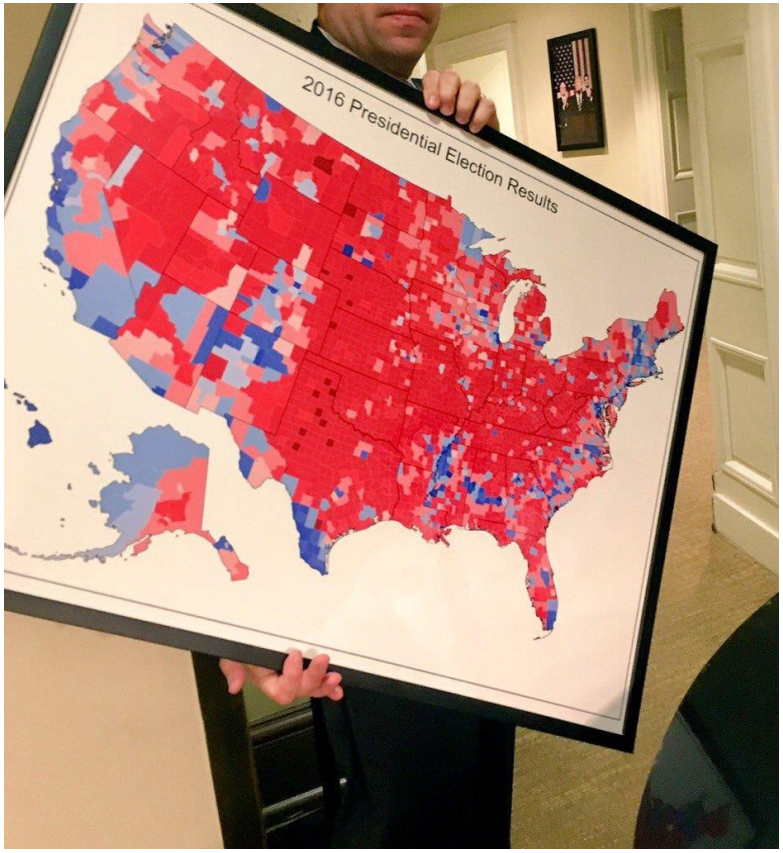
Ignoring uncertainty

Confirmation bias

# Confirmation bias

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

# Confirmation bias

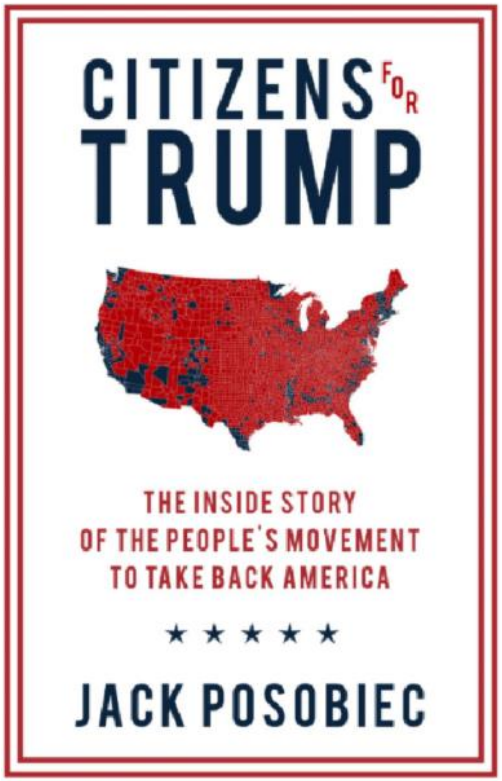



https://twitter.com/TreyYingst/status/862669407868391424

# Confirmation bias

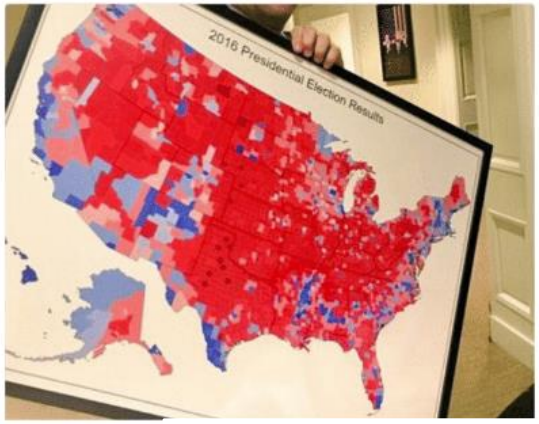

https://www.youtube.com/watch?v=Cd046xZhO_8&t=504s

# Confirmation bias




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

# Confirmation bias



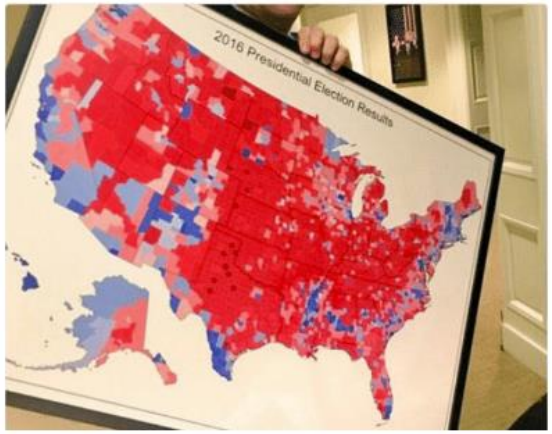

Surface on the county-level map:

**Red: 80%**

**Blue: 20%**

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

# Confirmation bias




Surface on the county-level map:

**Red: 80%**

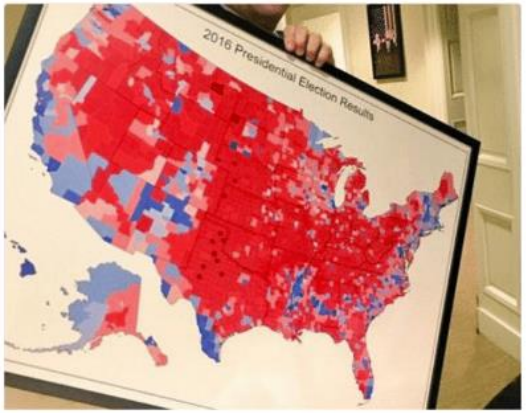
**Blue: 20%**

**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

# Confirmation bias

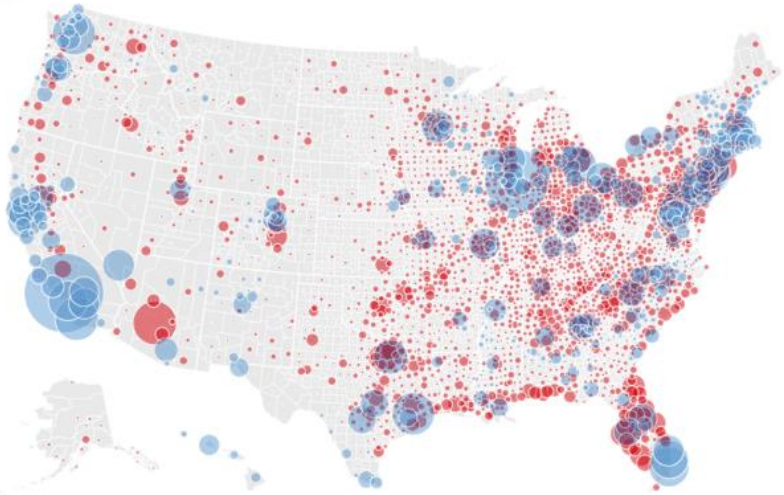



Surface on the county-level map:

**Red: 80%**

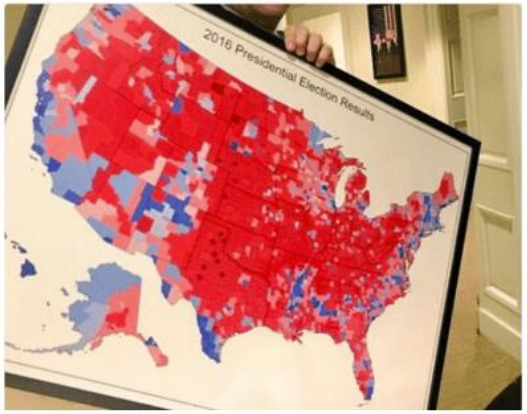
**Blue: 20%**

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



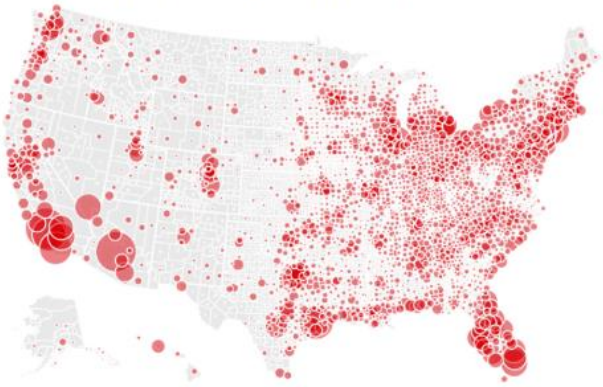

**SHARE OF THE POPULAR VOTE IN THE 2016 PRESIDENTIAL ELECTION**

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

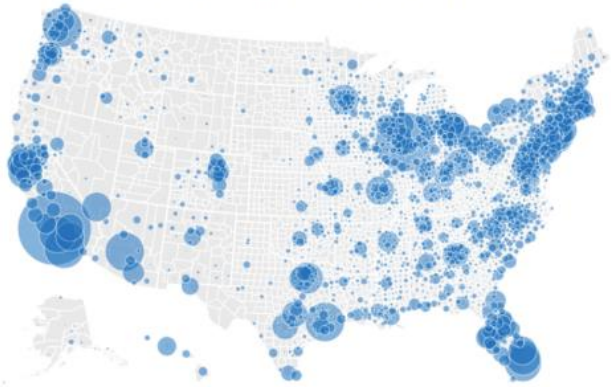
**PERCENTAGE OF ELIGIBLE VOTERS**

| | |
|---|---|
| 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**



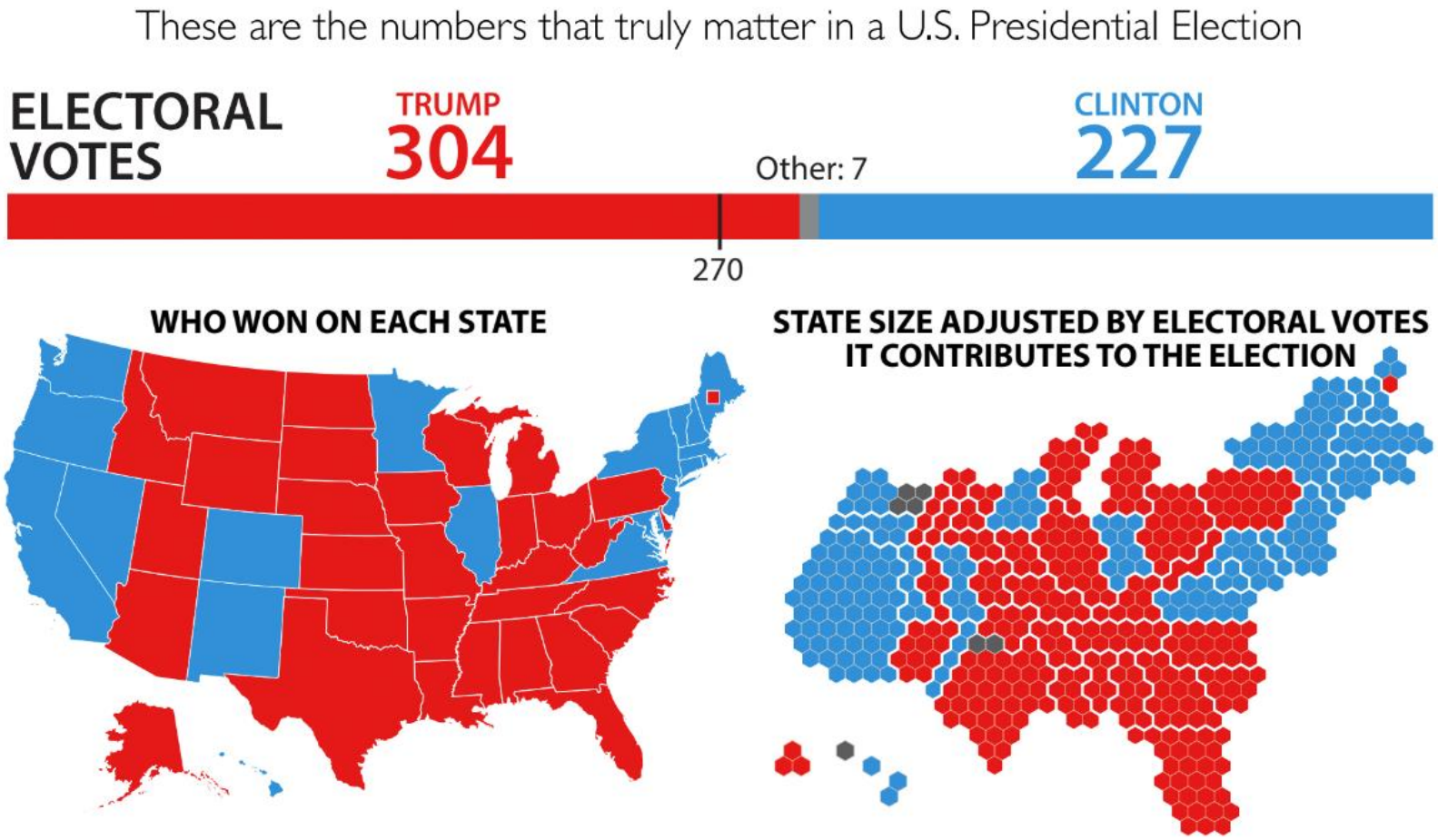
**VOTES FOR HILLARY CLINTON**



Bubble size is proportional to the number of votes per county

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

# Confirmation bias




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

# 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

# Trustworthiness

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