

Data Visualization

EXAMPLES (1)

Good visualization design is

1. Trustworthy

2. Accessible

3. Elegant

How charts lie?

Phenomenon



Data



Dubious data

Chart



Misrepresenting data

Cherry-picking data

Ignoring uncertainty

Person



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

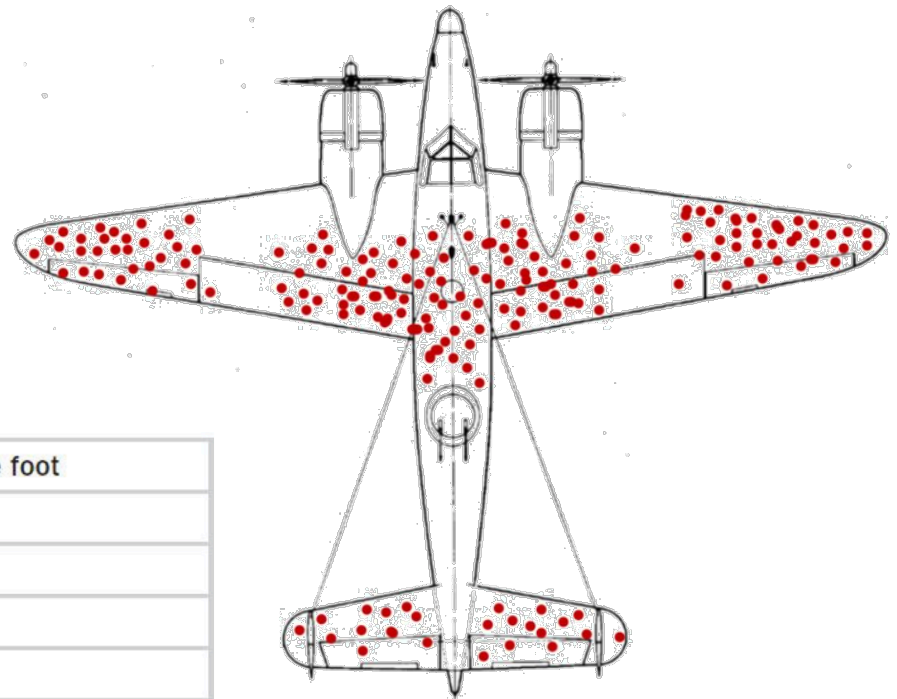


https://www.reddit.com/r/dataisbeautiful/comments/9pkka4/all_recorded_meteorite_impacts_in_the_us_from

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

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?”

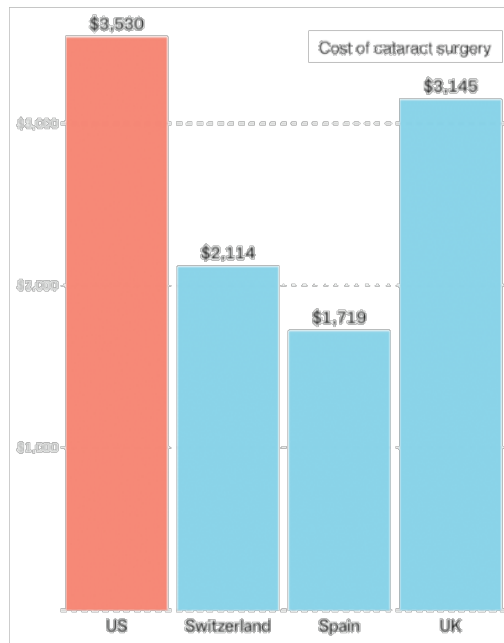
remain/leave

Non-comparable data used in comparisons

Vox

TWEET SHARE

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

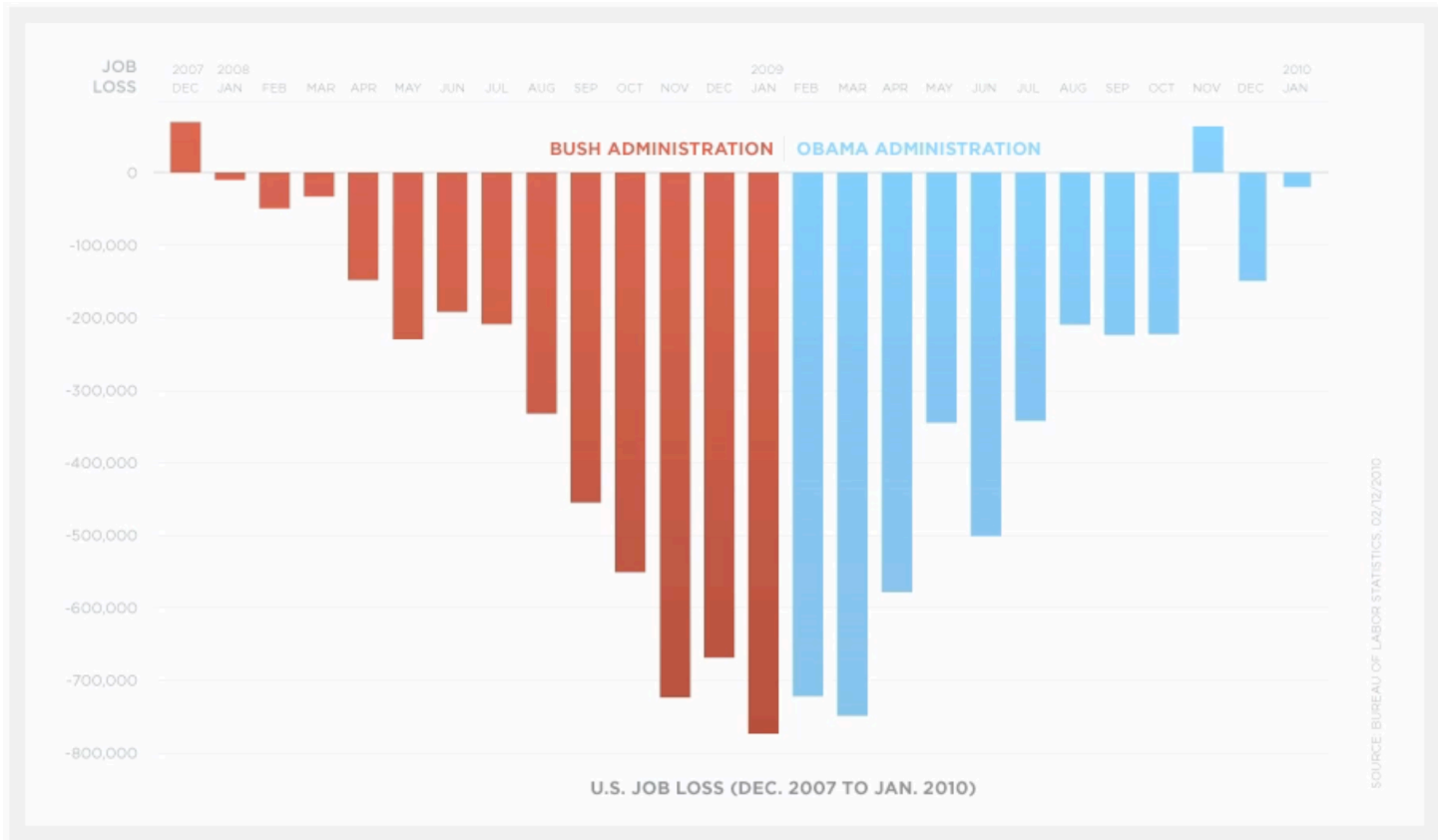


Two issues

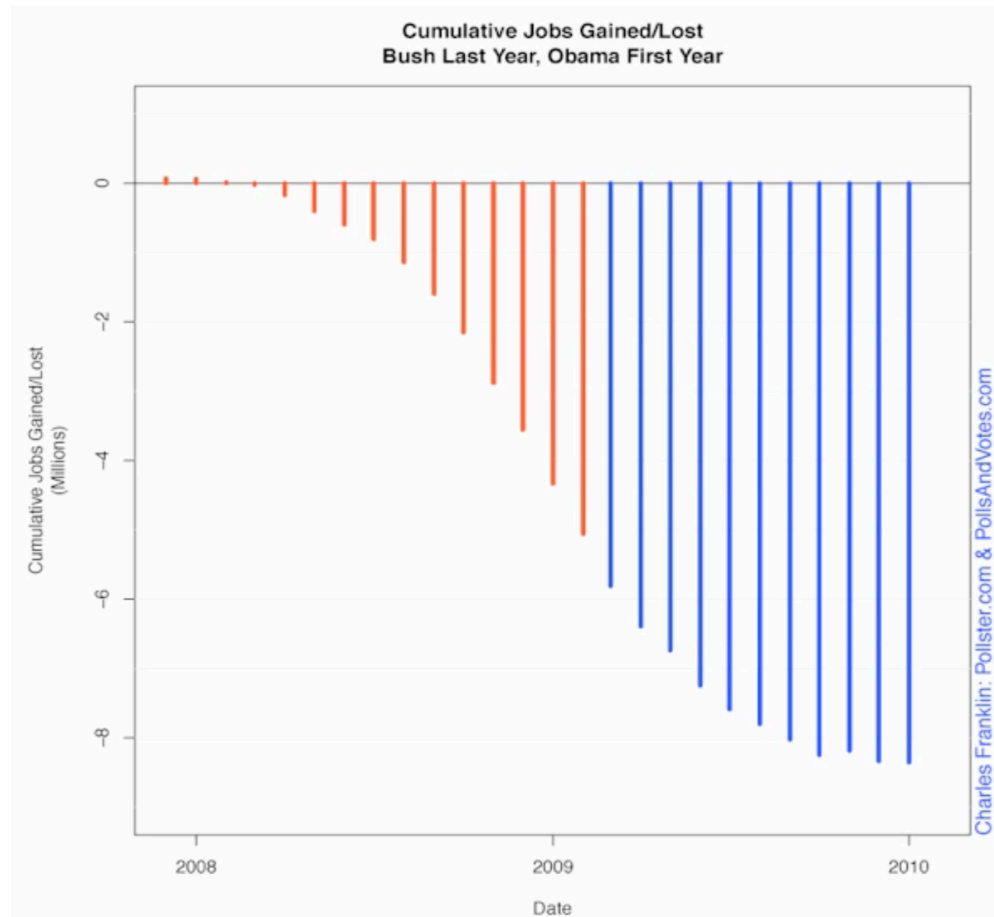
- Prices not adjusted for purchasing power
- Different sources of data

The data source specifically warns against using this data for comparison

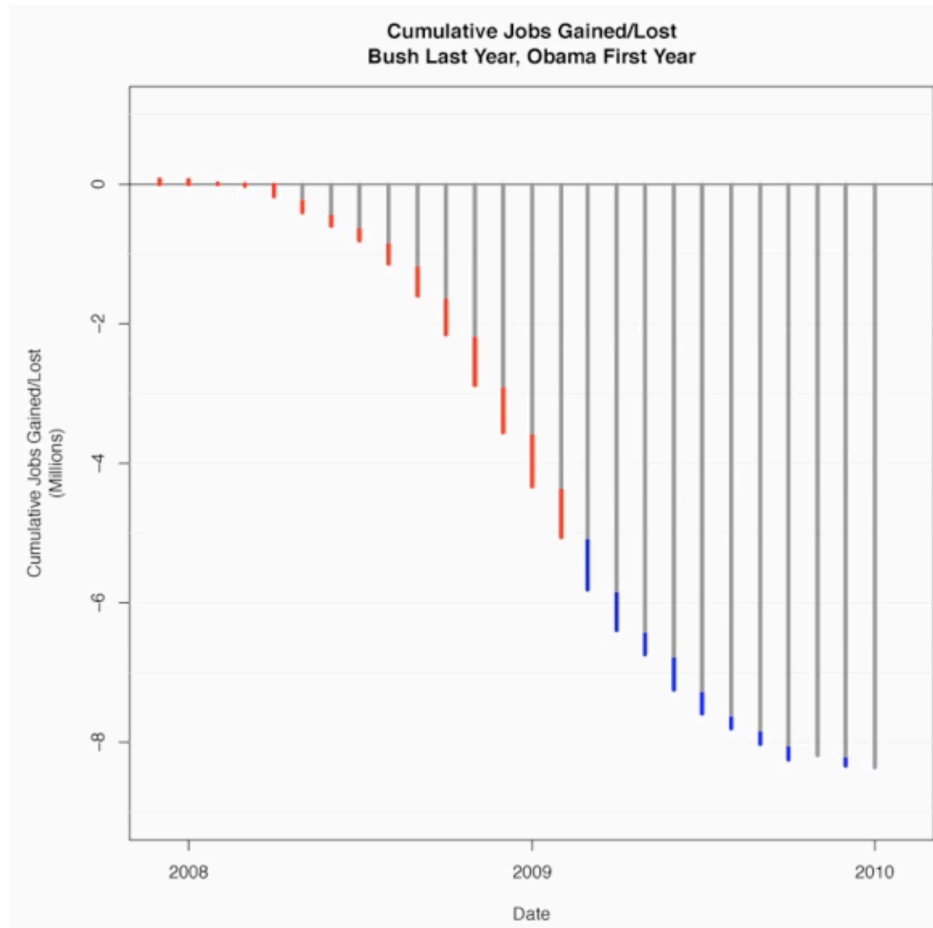
Absolute instead of cumulative data



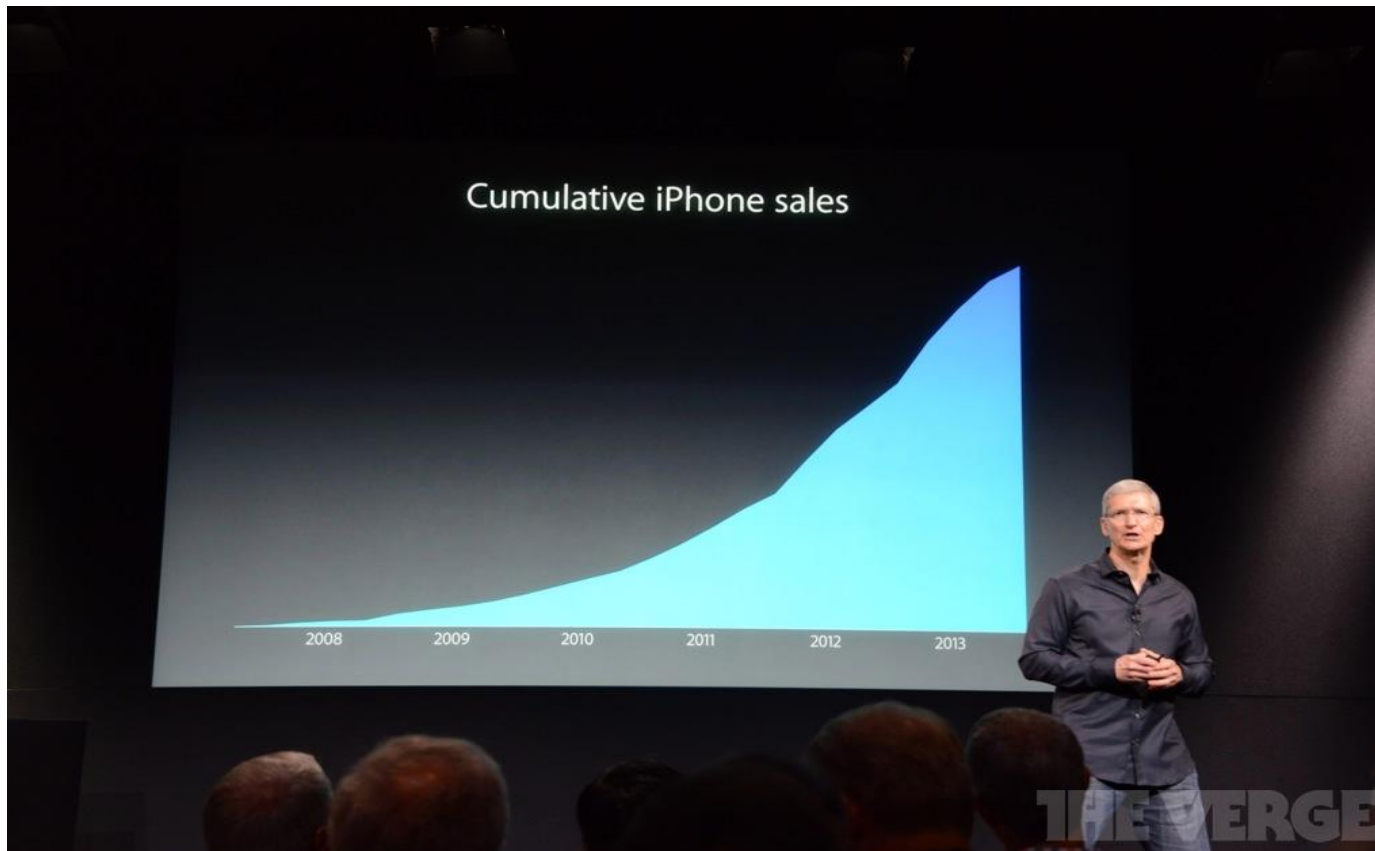
Cumulative data



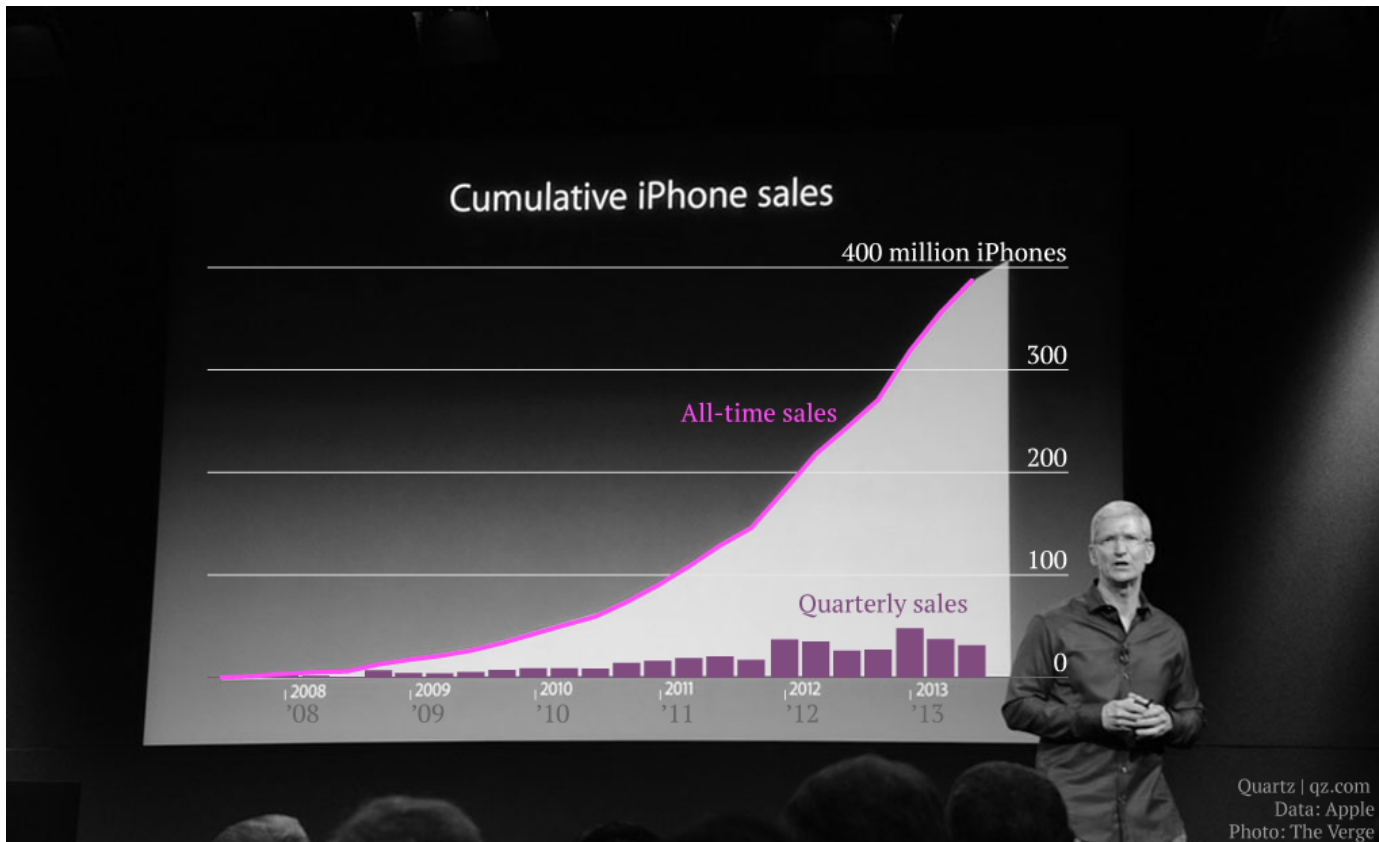
Absolute and cumulative data



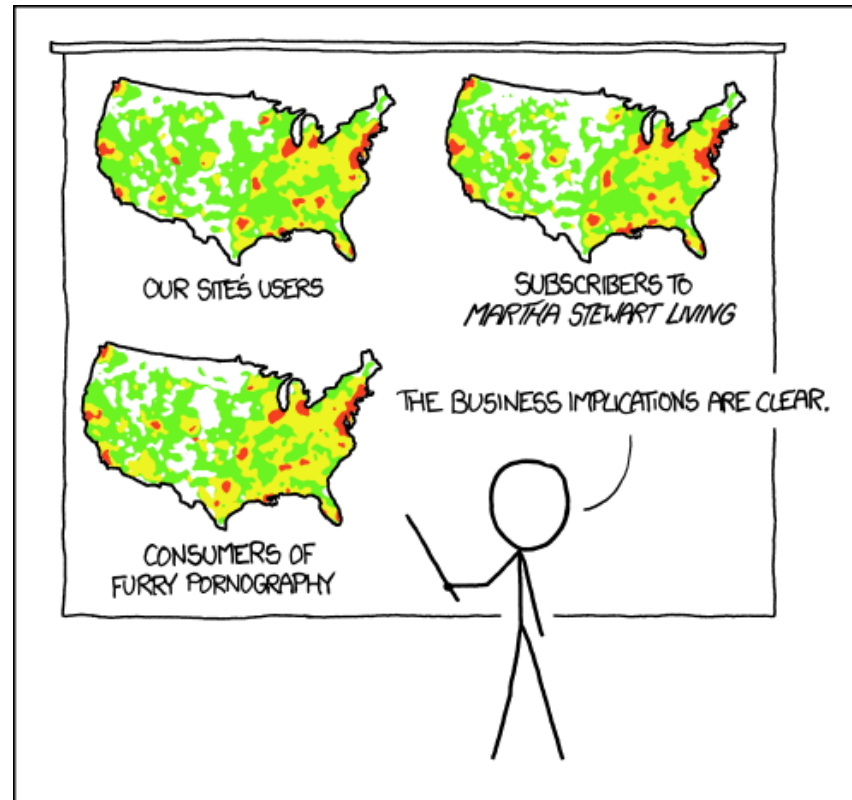
Cumulative instead of absolute data



Cumulative and absolute data

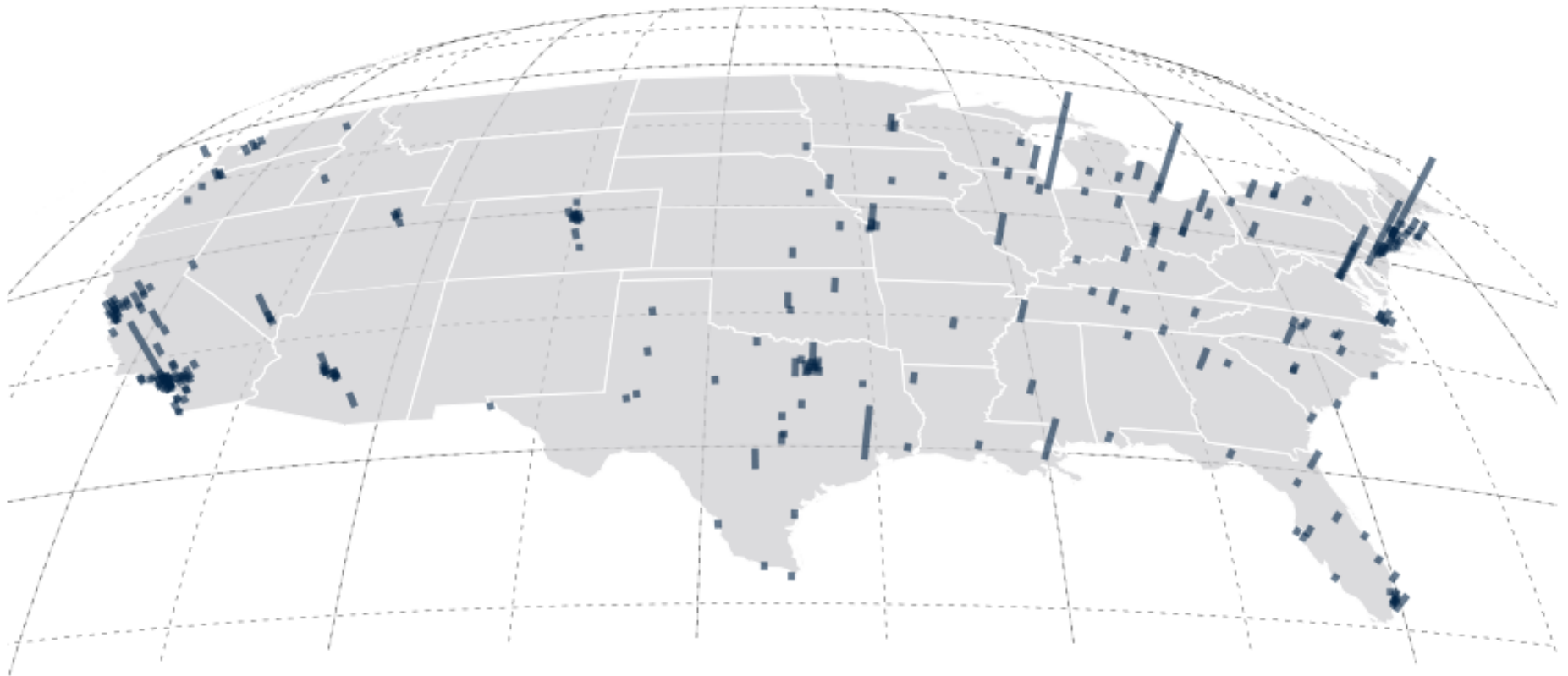


Absolute instead of relative data

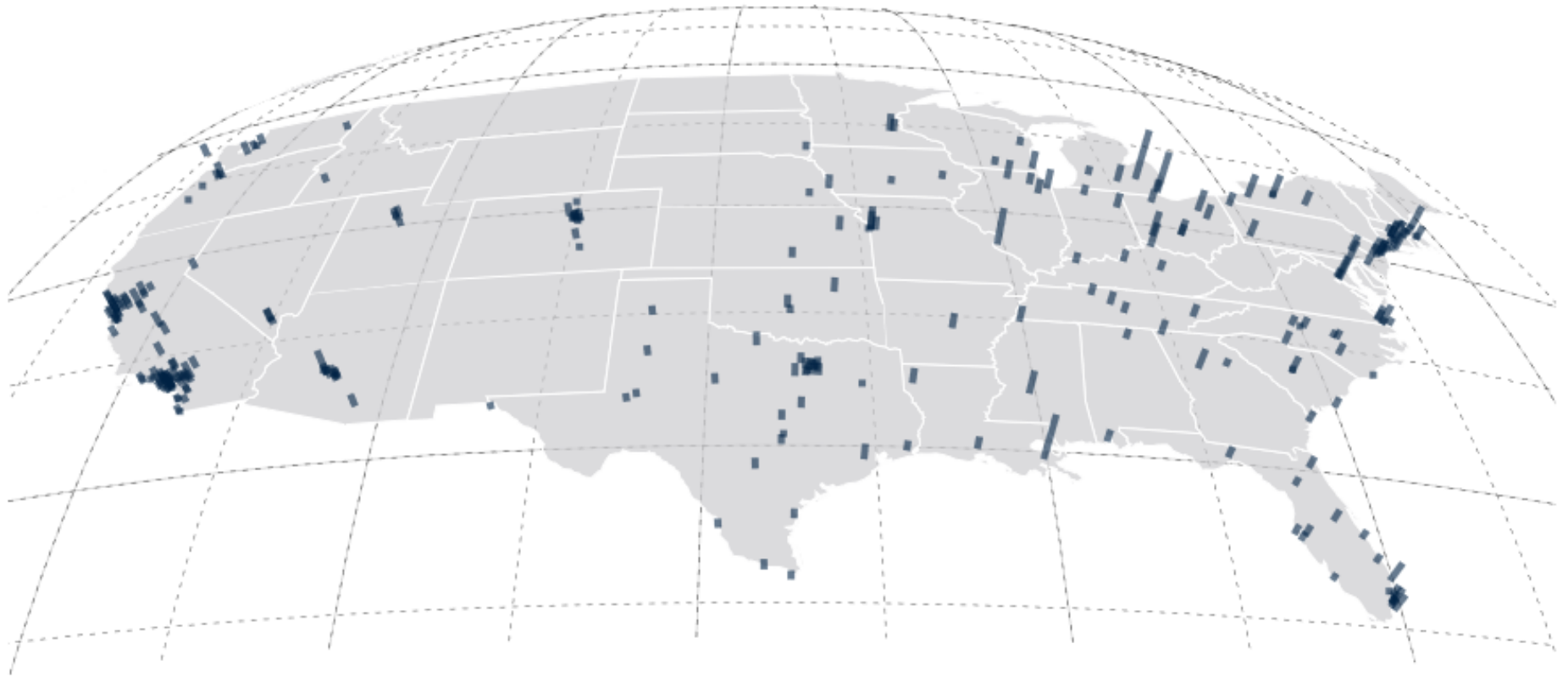


PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE
BASICALLY JUST POPULATION MAPS

Absolute data



Relative data



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

Chart

Person



Dubious data

Misrepresenting
data

Cherry-picking
data

Ignoring
uncertainty

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 of different scales
- Improper scaling of areas and pictograms

Unnecessary 3-D

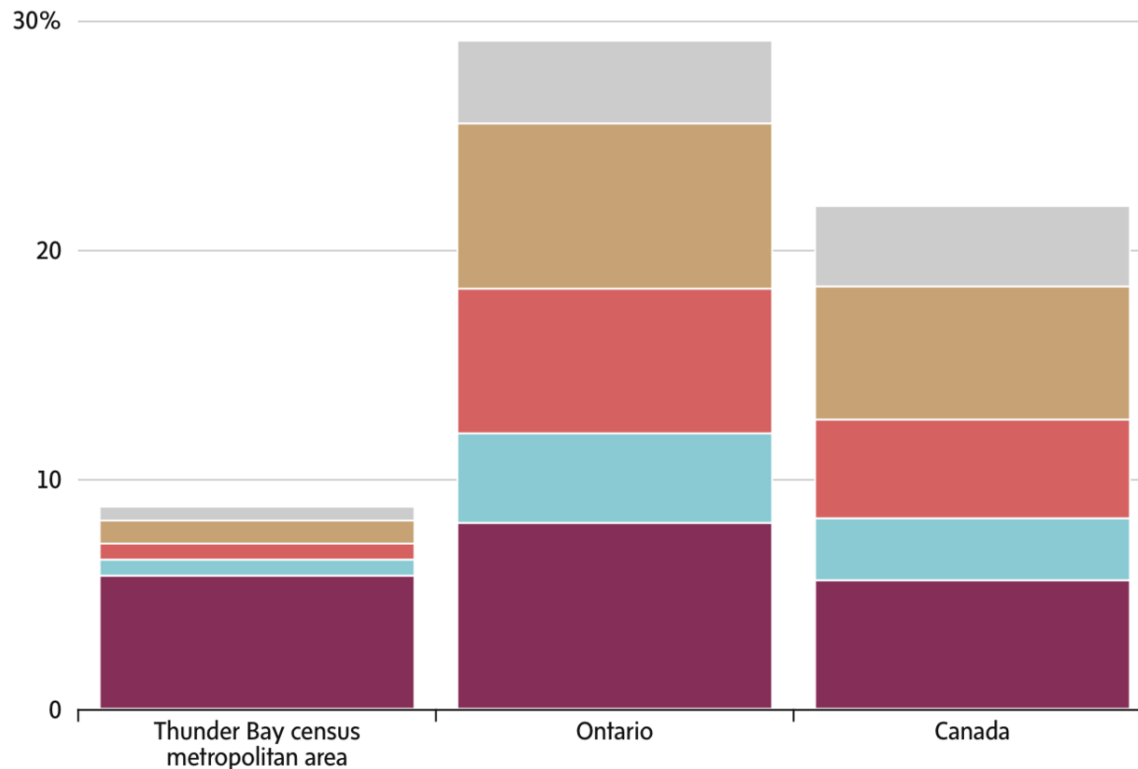
Improper categorization

Oversimplifying

Time not on an axis

Immigrants as a percentage of population in 2016, by period of immigration

● Before 1981 ● 1981-90 ● 1991-2000 ● 2001-10 ● 2011-16



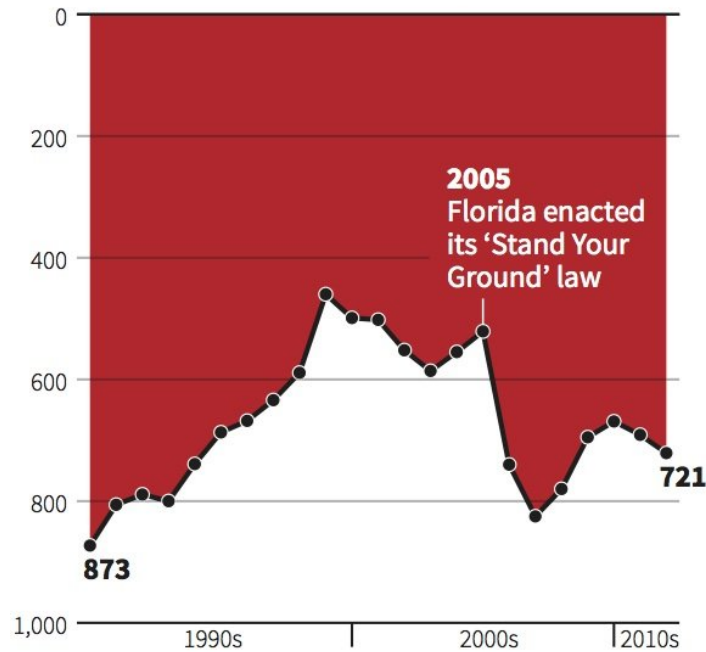
THE GLOBE AND MAIL, SOURCE: STATSCAN

DATA SHARE

Distorted axis: Inverted y axis

Gun deaths in Florida

Number of murders committed using firearms

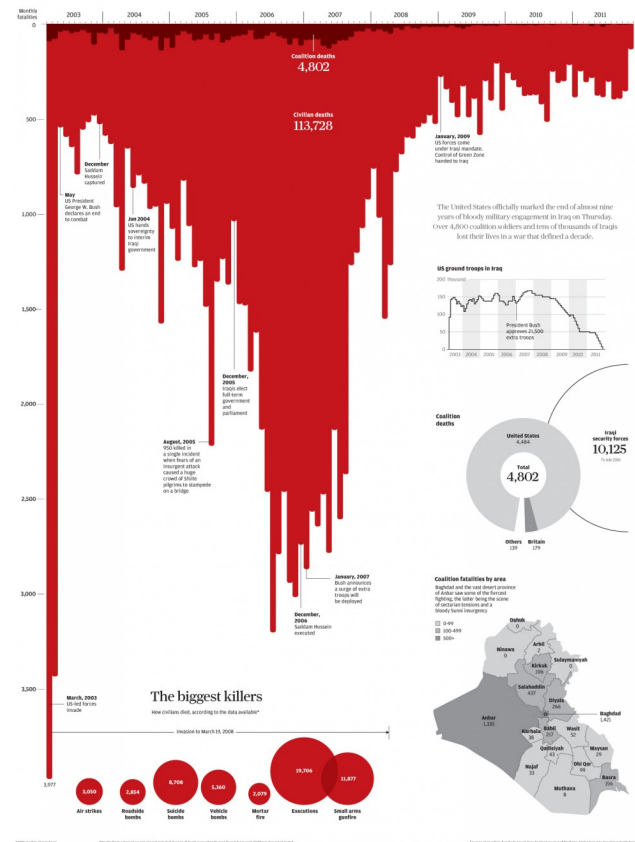


Source: Florida Department of Law Enforcement

C. Chan 16/02/2014



Iraq's bloody toll



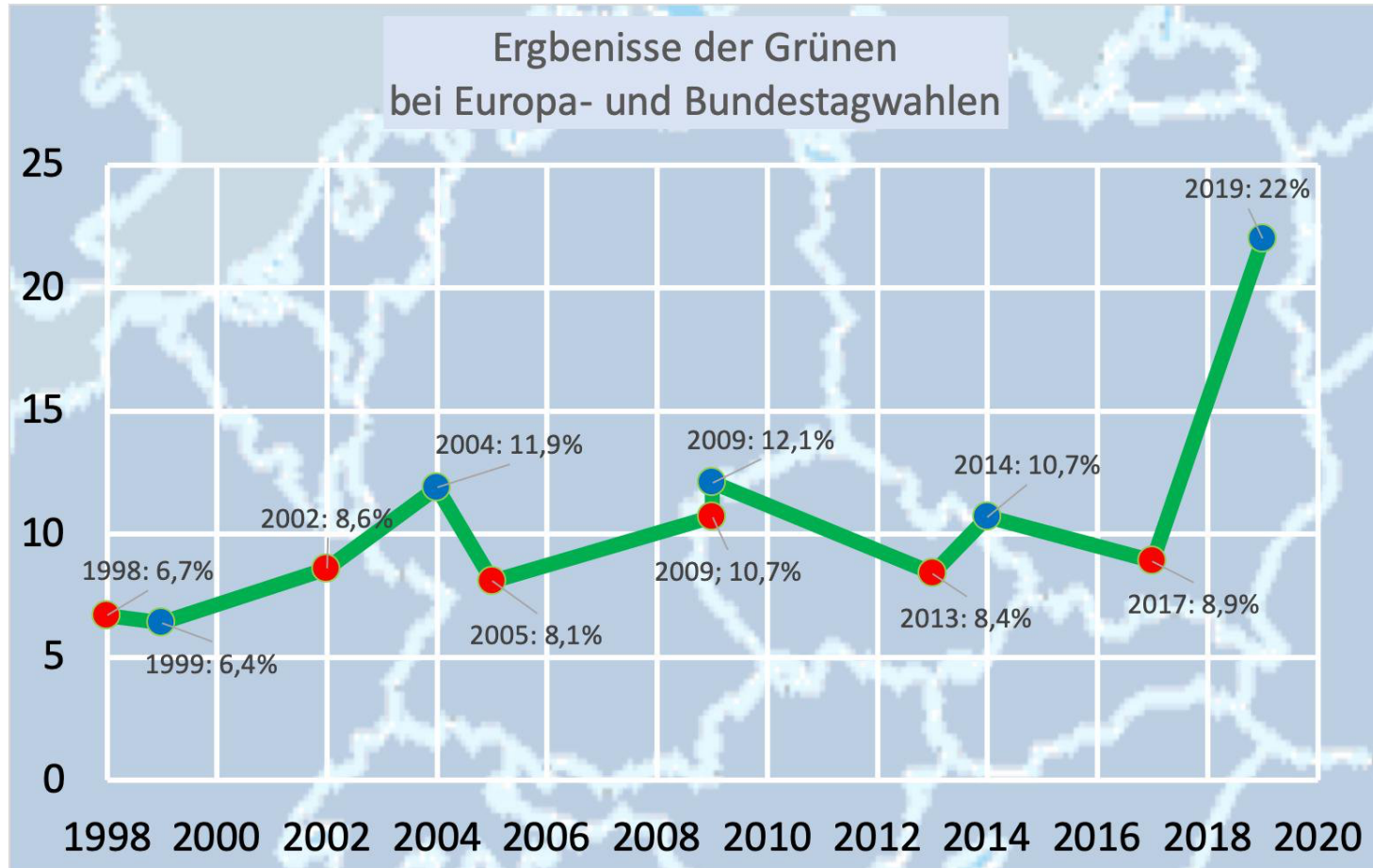
Distorted axis



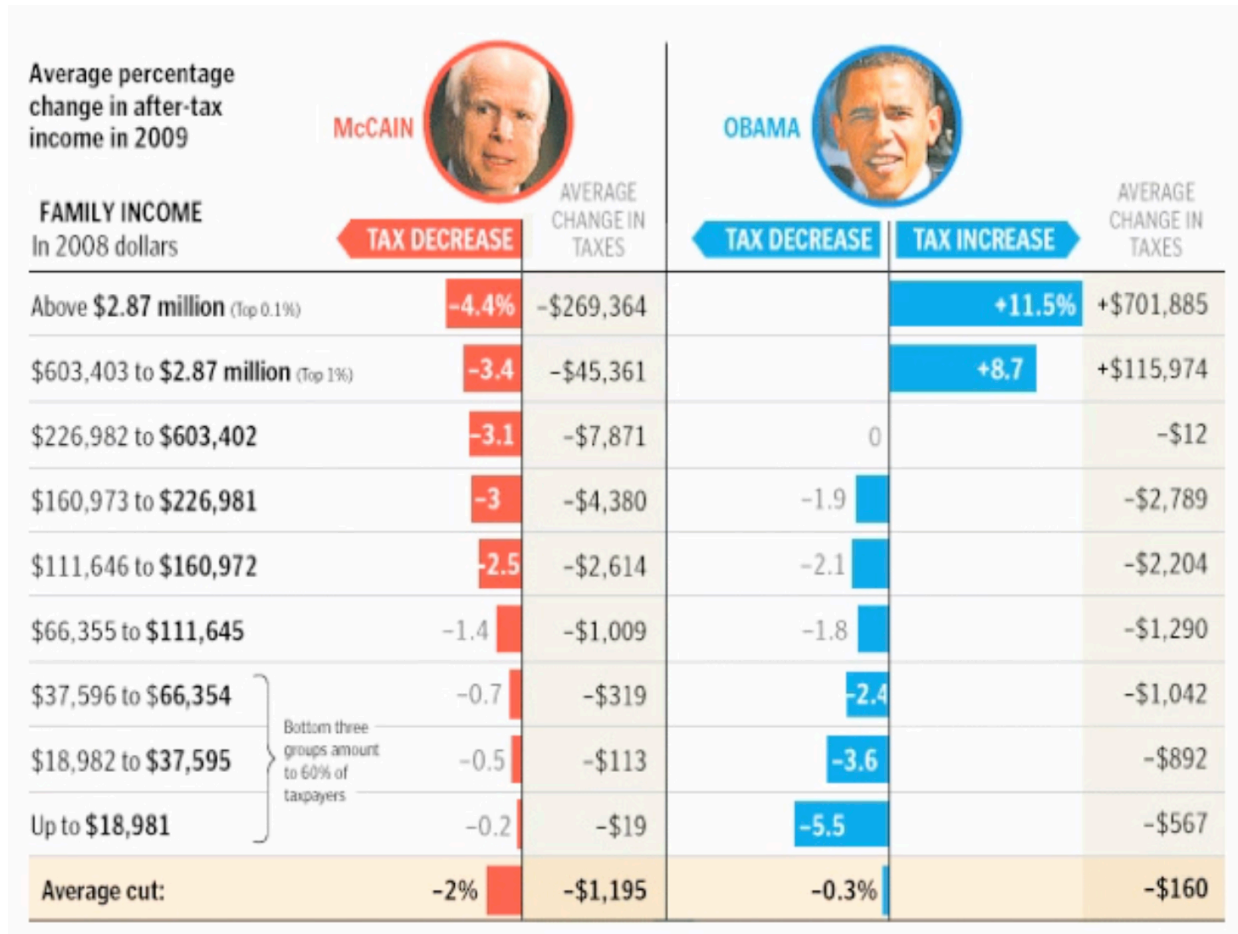
Distorted axis



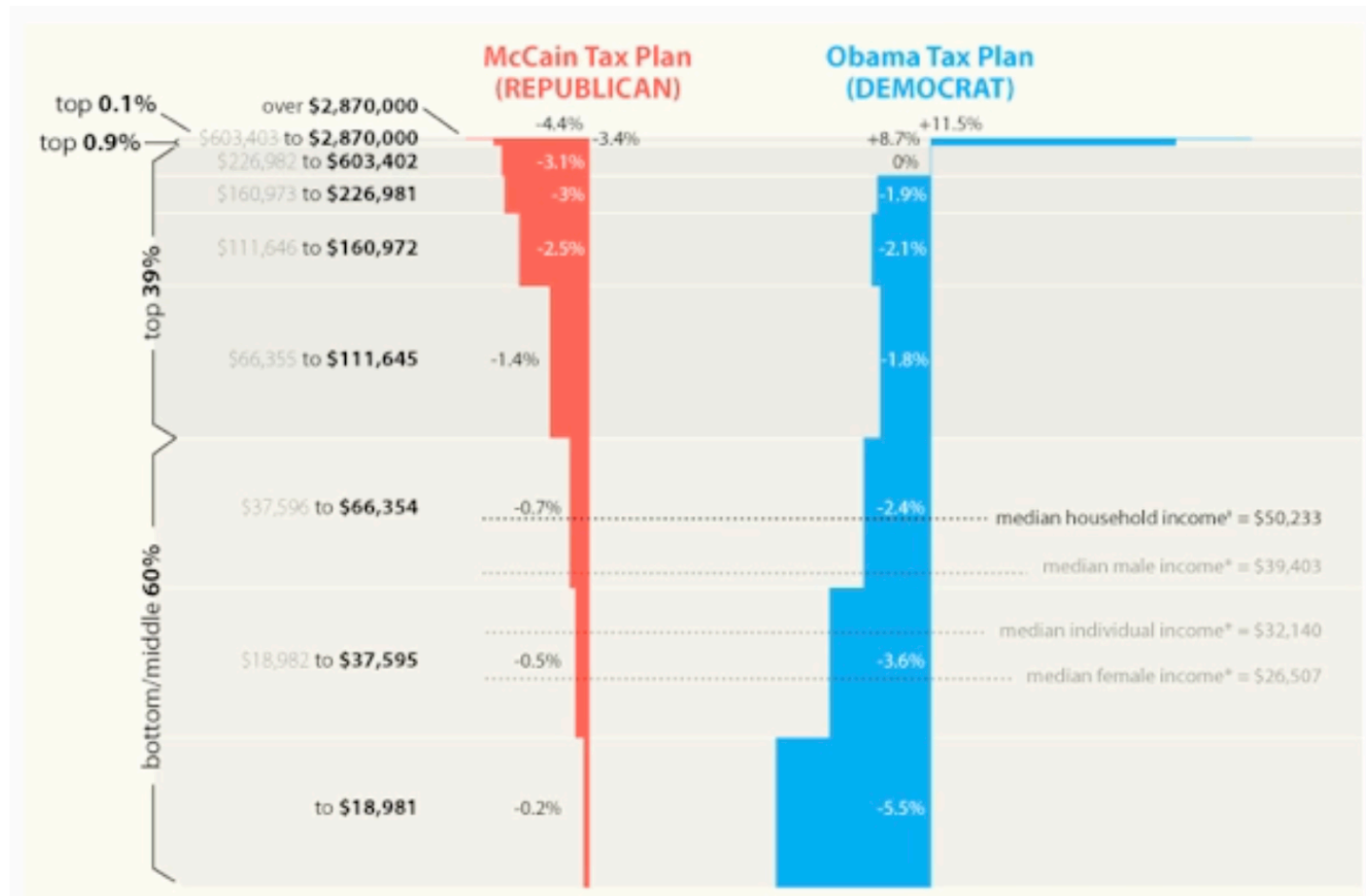
Fixed axis



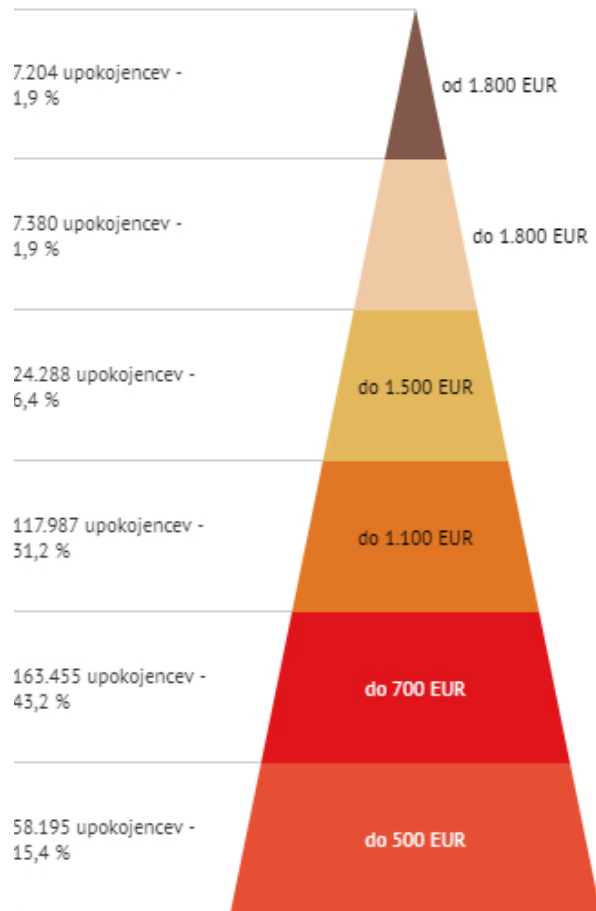
Unequal intervals



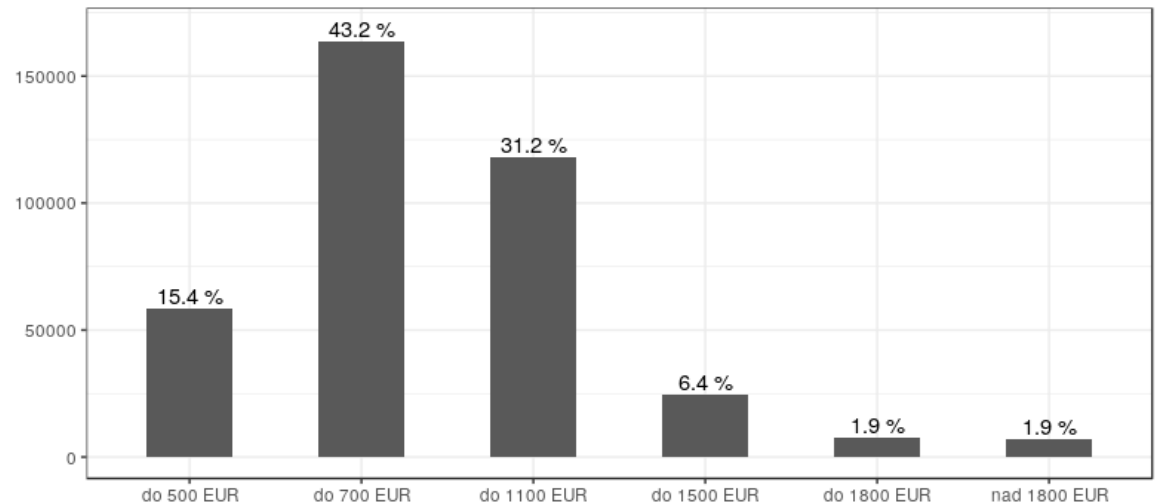
Fixed intervals



Unequal intervals (x2)



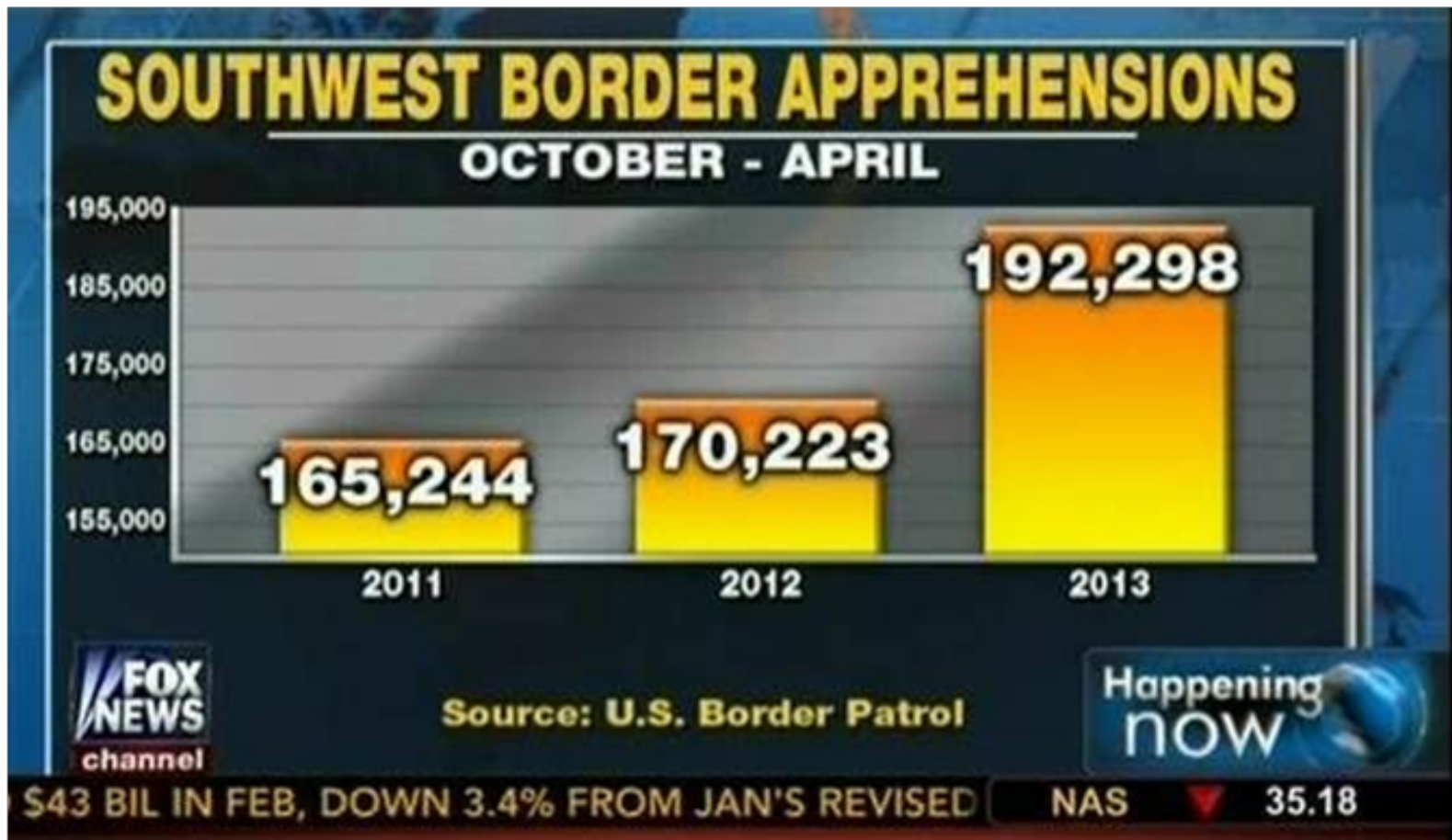
Pensions in Slovenia



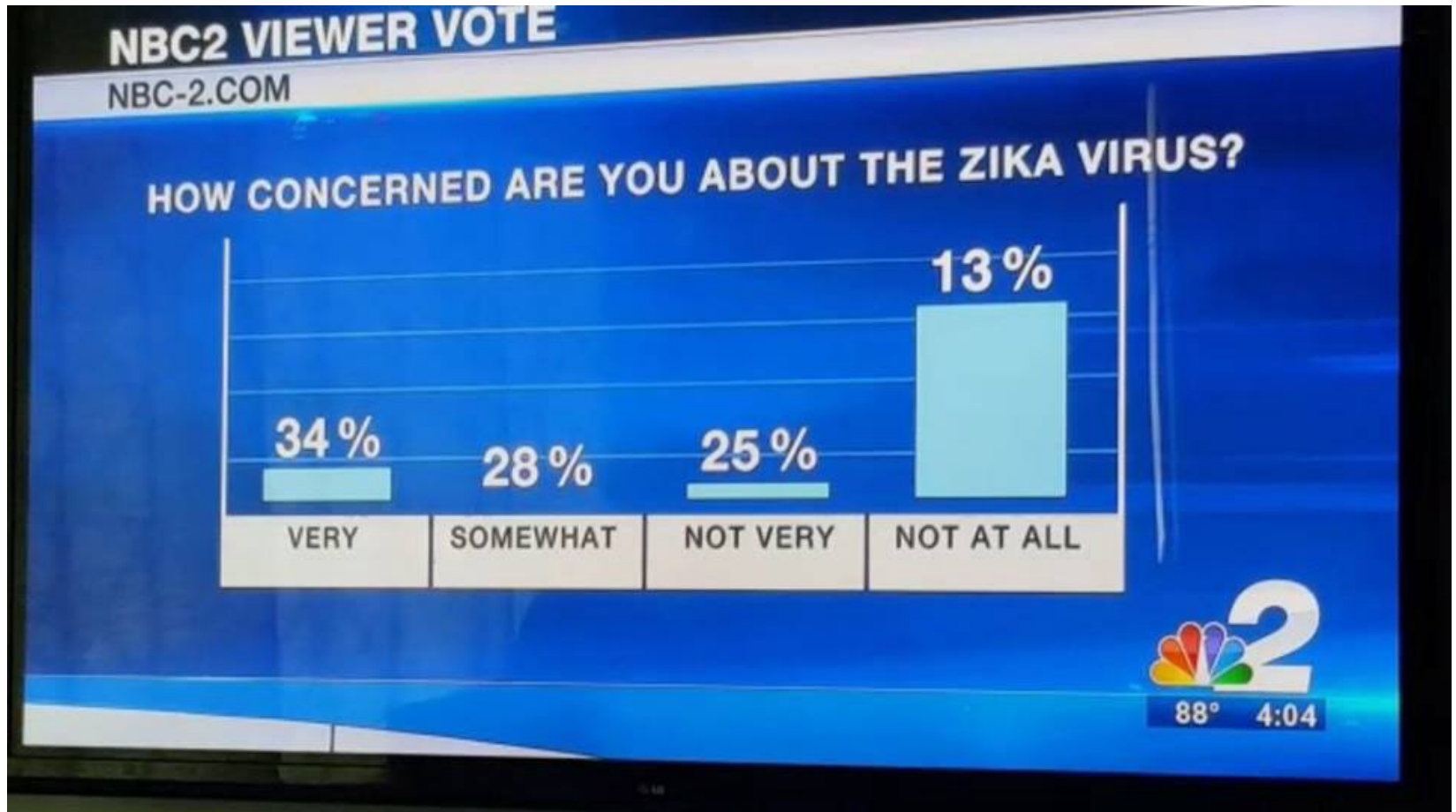
Over 100% pie chart



Bar chart with truncated axis

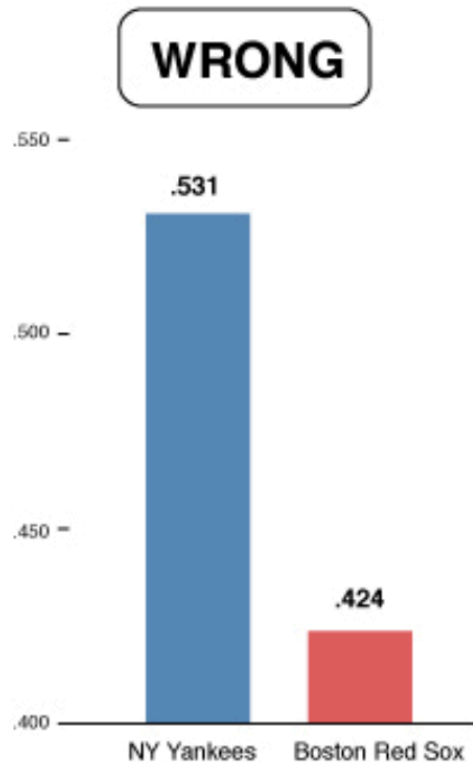


Bar chart with errors

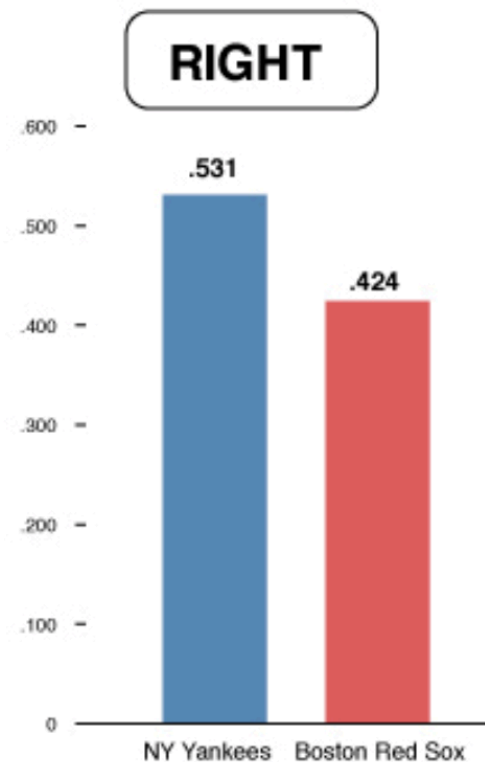


Bar chart with truncated axis

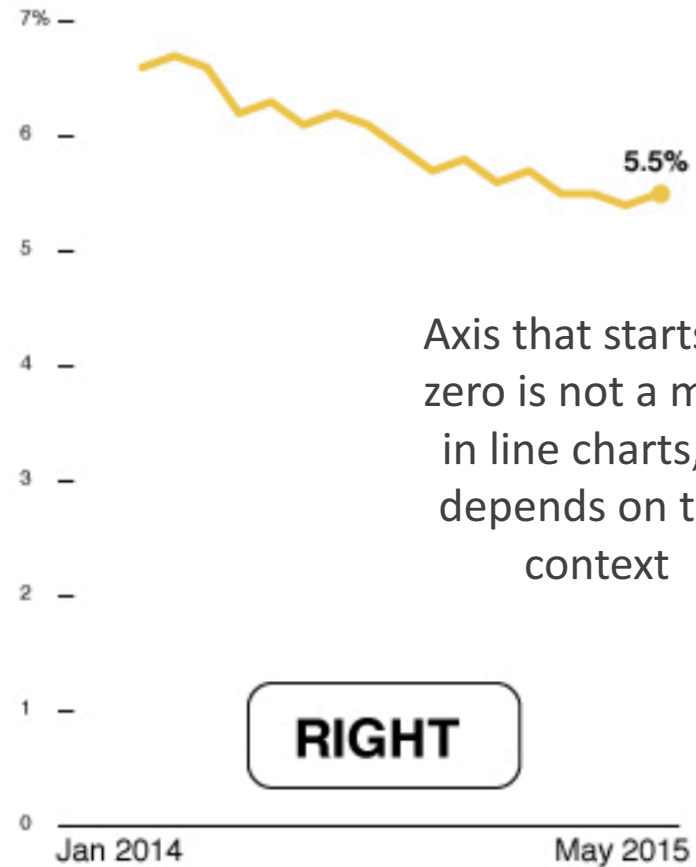
Percentage of victories



Percentage of victories

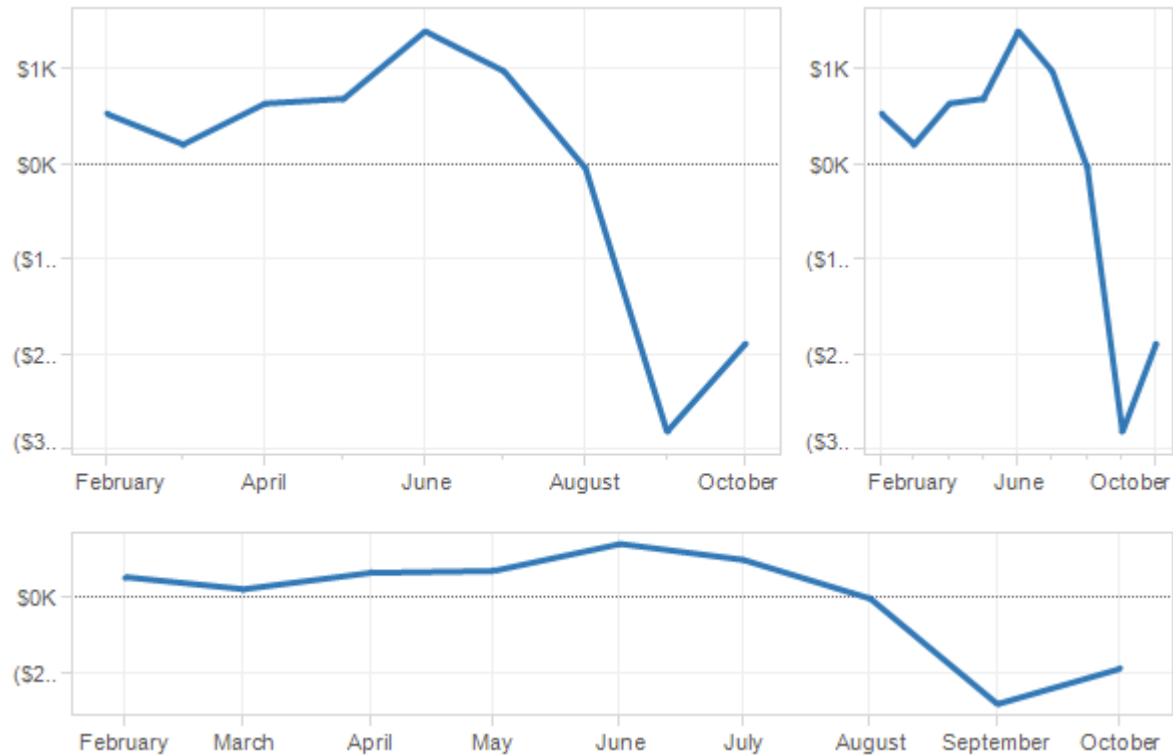


Line chart with truncated axis



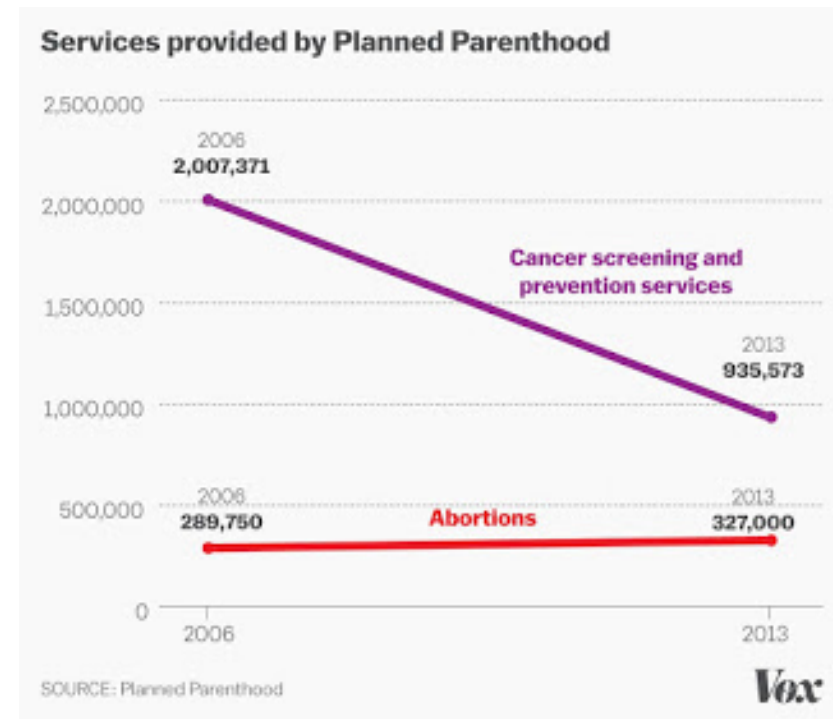
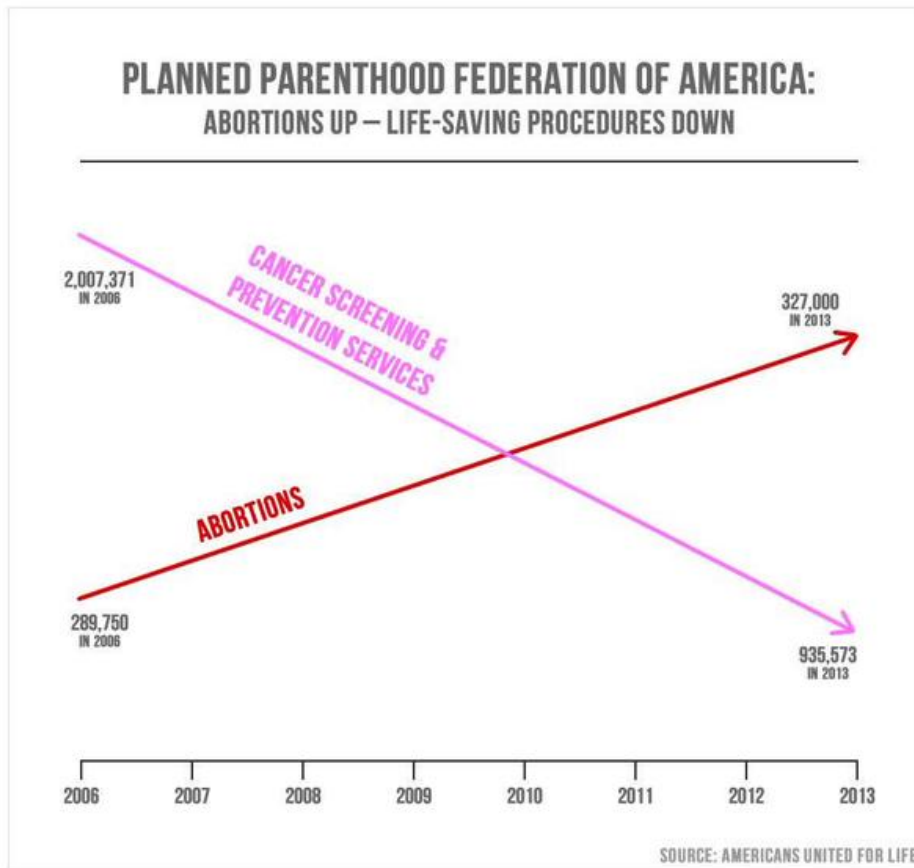
Axis that starts at zero is not a must in line charts, it depends on the context

Aspect ratio bias

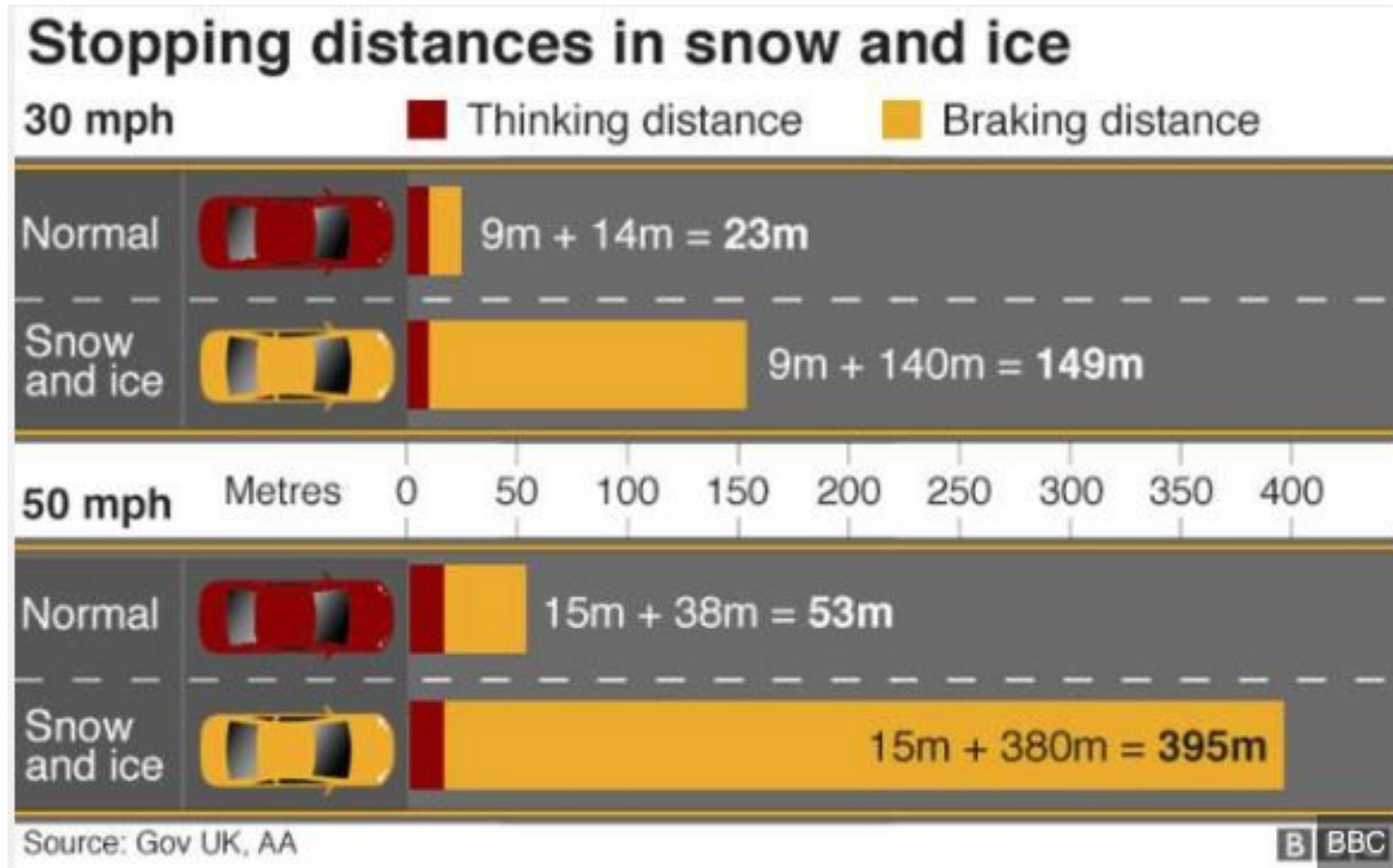


Banking to 45 Degrees

Dual axes with different scale



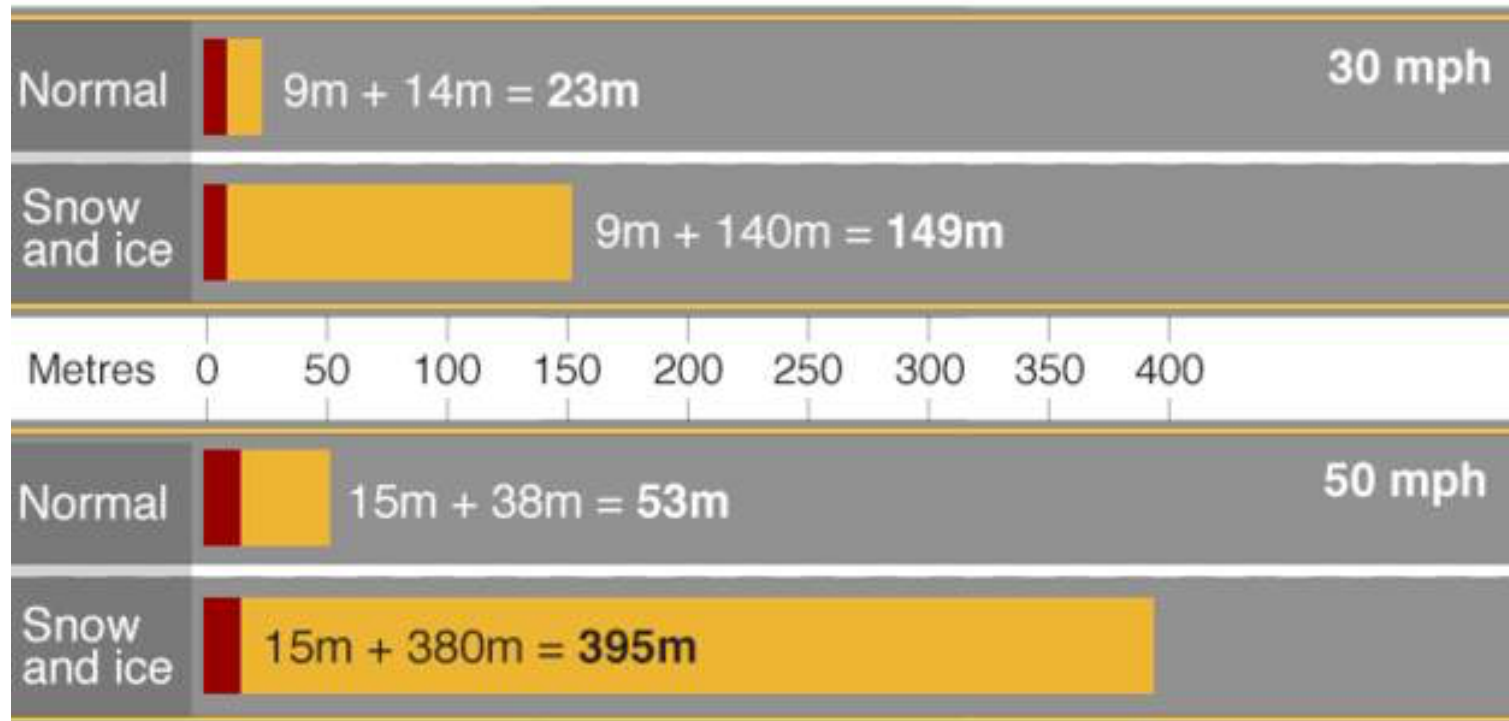
Improper scaling of areas/pictograms



Improper scaling of areas/pictograms – fixed

Stopping distances in snow and ice

■ Thinking distance ■ Braking distance



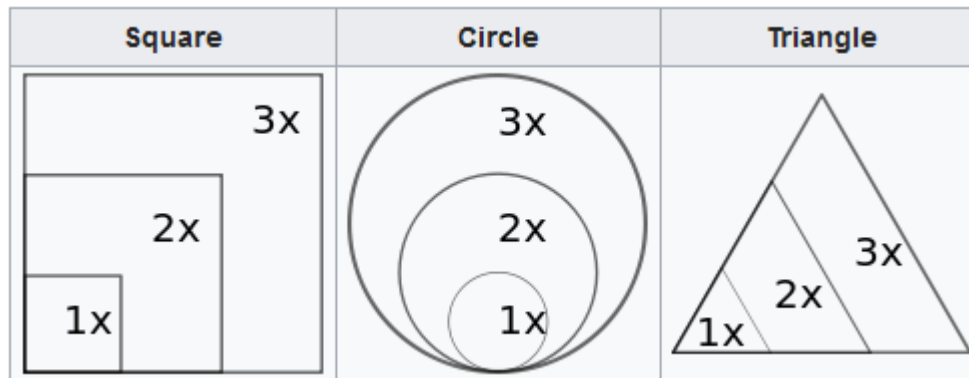
Source: Gov UK, AA



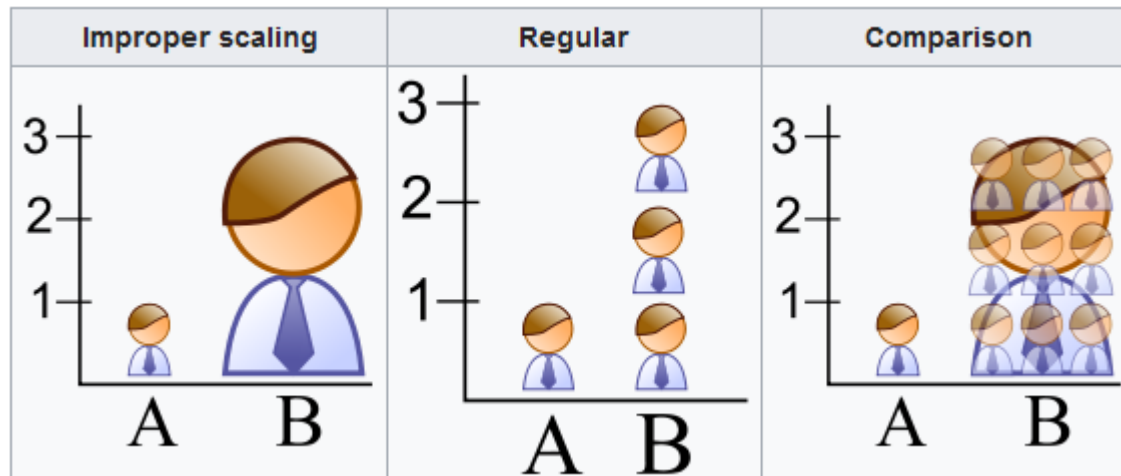
Improper scaling of areas/pictograms



Improper scaling of areas/pictograms



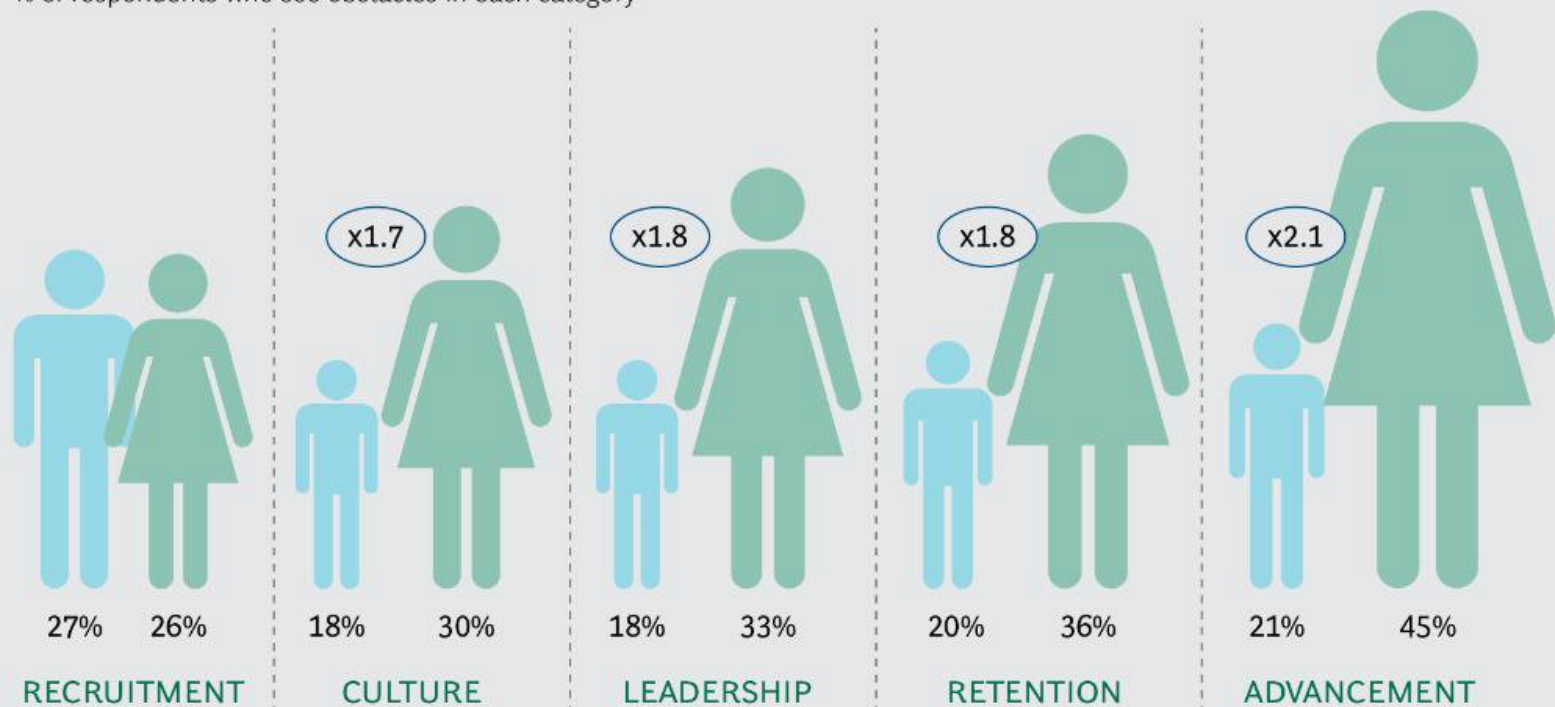
Even worse, if the elements are 3-D



Improper scaling of areas/pictograms

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

% of respondents who see obstacles in each category

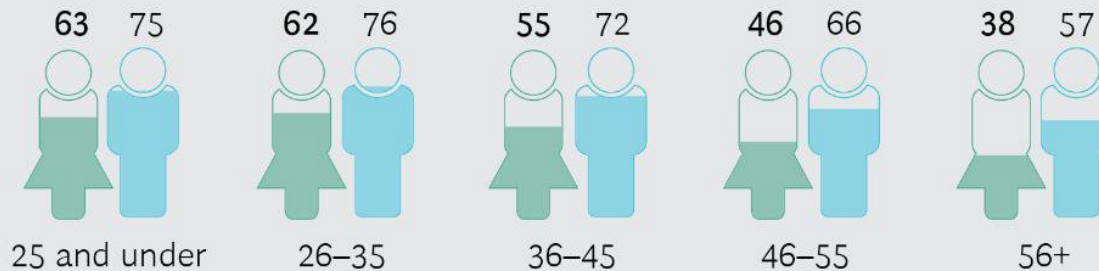


Source: BCG Global Gender Diversity Survey 2017.

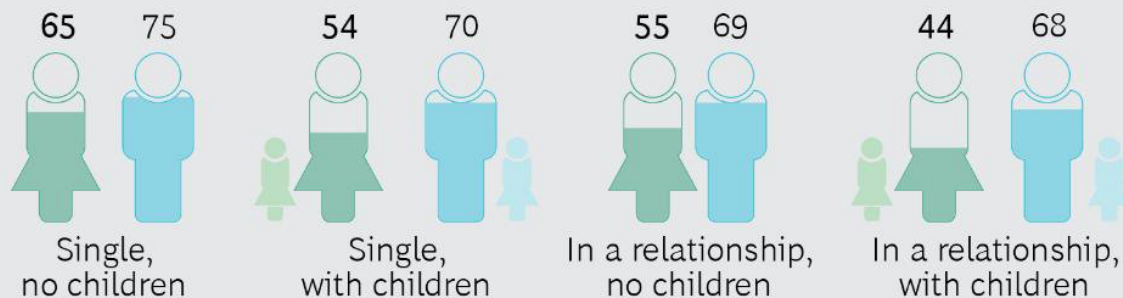
Proper scaling of areas/pictograms

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 (%)

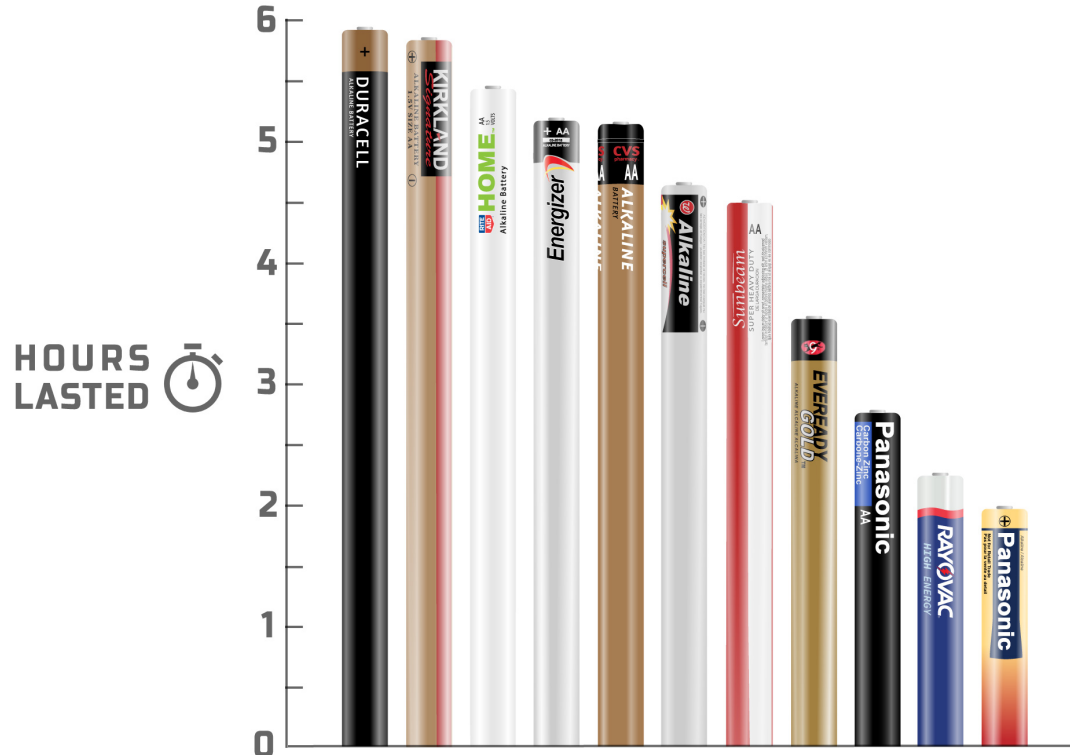


Source: BCG analysis.

Proper scaling of areas/pictograms

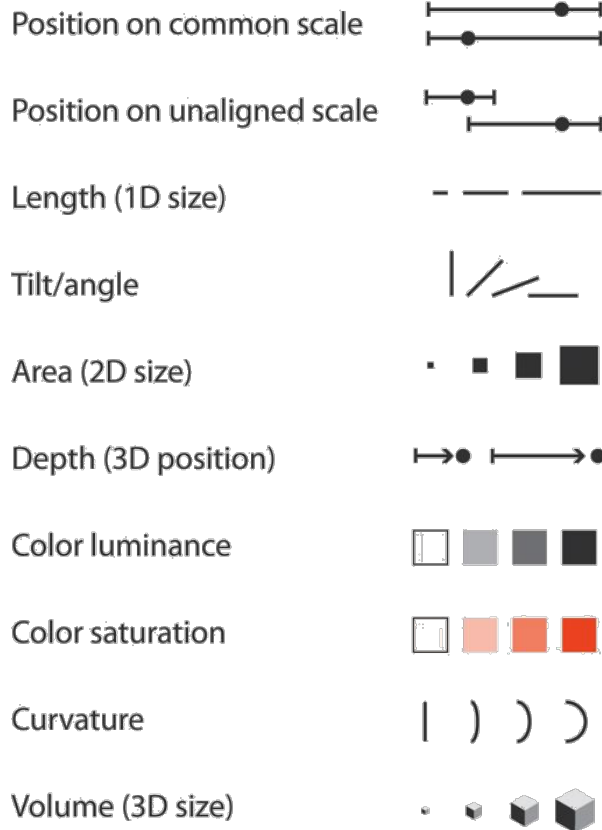
WHICH BATTERIES LAST LONGEST?

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



Accuracy of visual channels

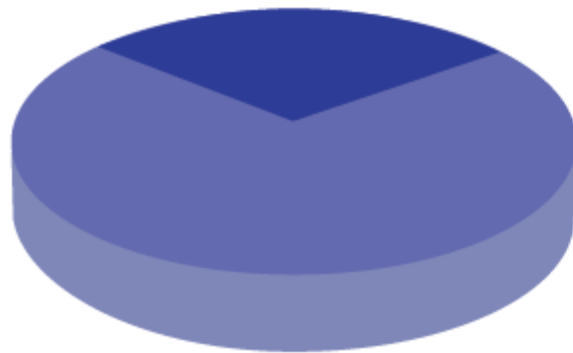
➔ Magnitude Channels: Ordered Attributes



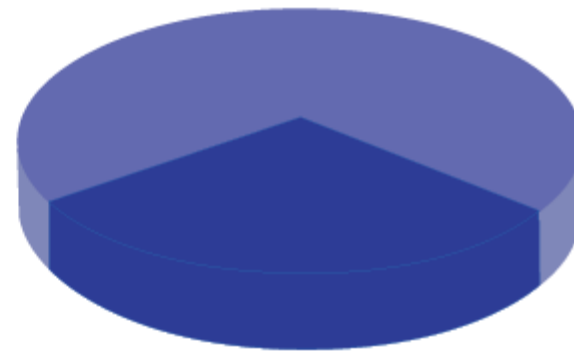
➔ Identity Channels: Categorical Attributes



Unnecessary 3-D



■ Labor ■ Other



■ Labor ■ Other

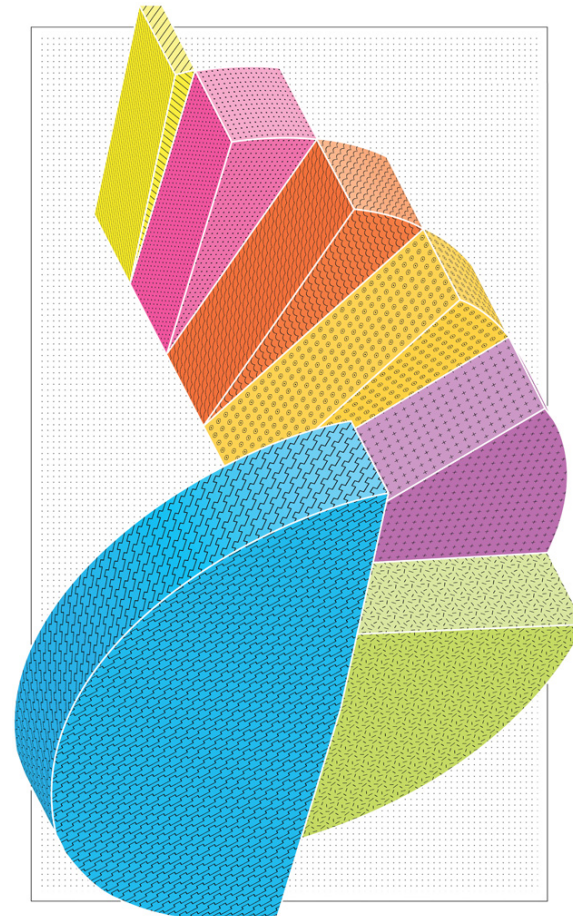
Unnecessary 3-D



Unnecessary 3-D

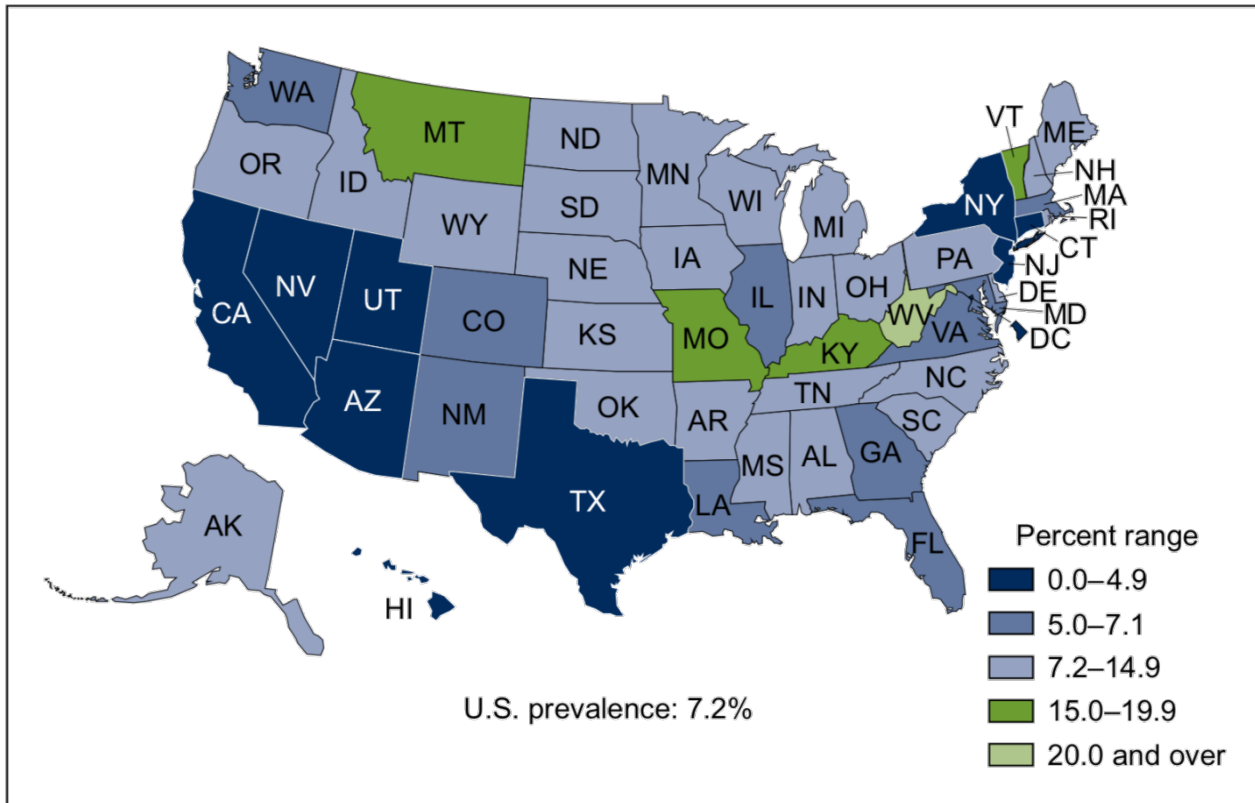
ANATOMY OF A WINNING TED TALK

- 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 salesman 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.



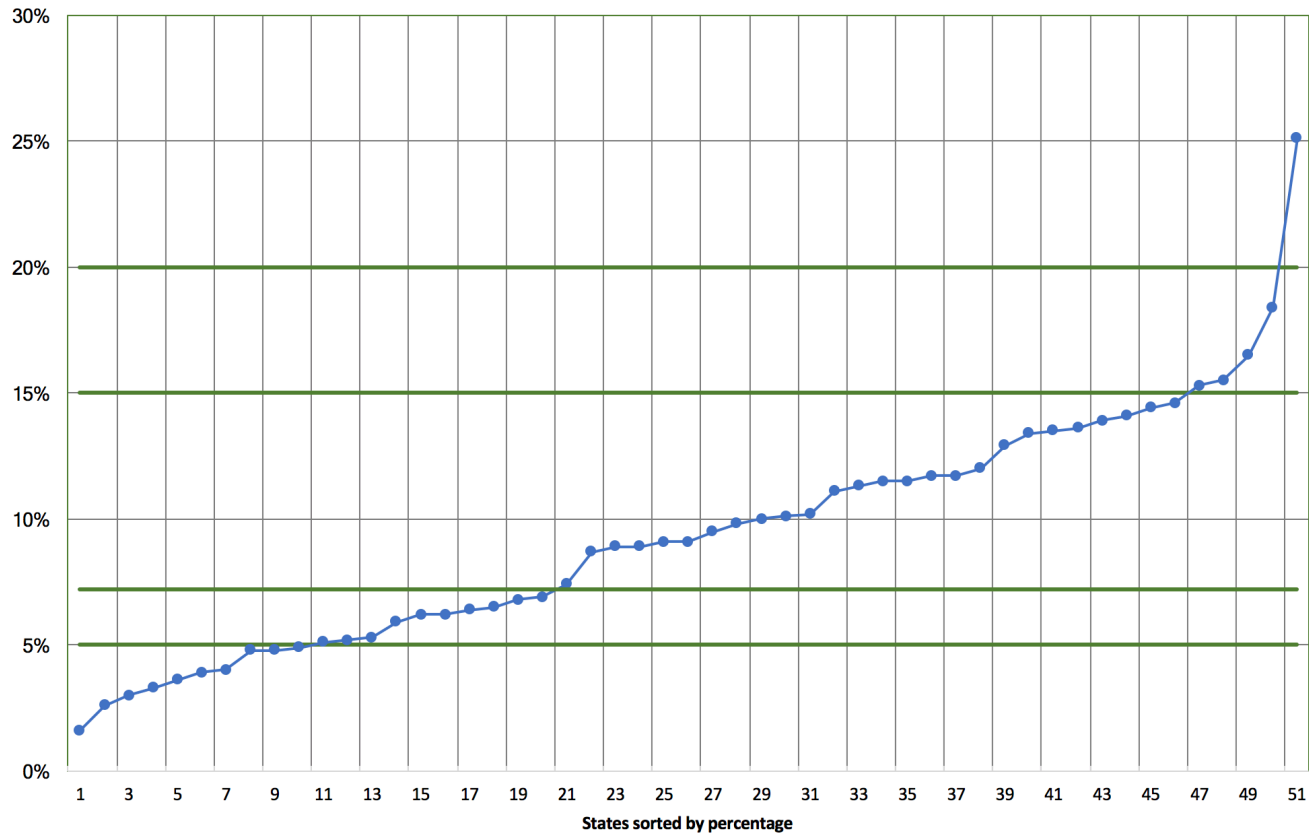
Improper categorization (and color choice)

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

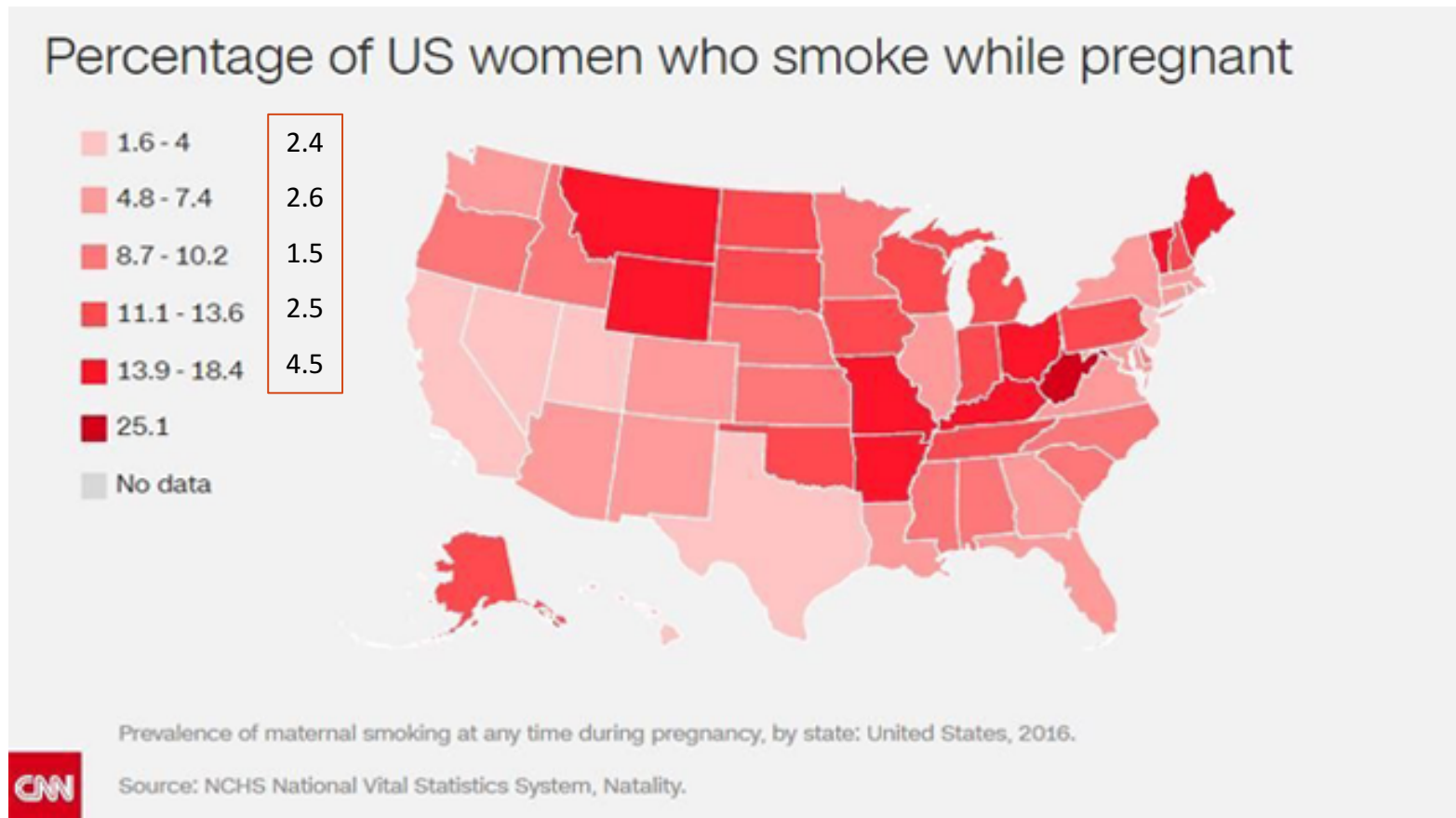


NOTE: Access data table for Figure 1 at: https://www.cdc.gov/nchs/data/databriefs/db305_table.pdf#1.
 SOURCE: NCHS National Vital Statistics System, Natality.

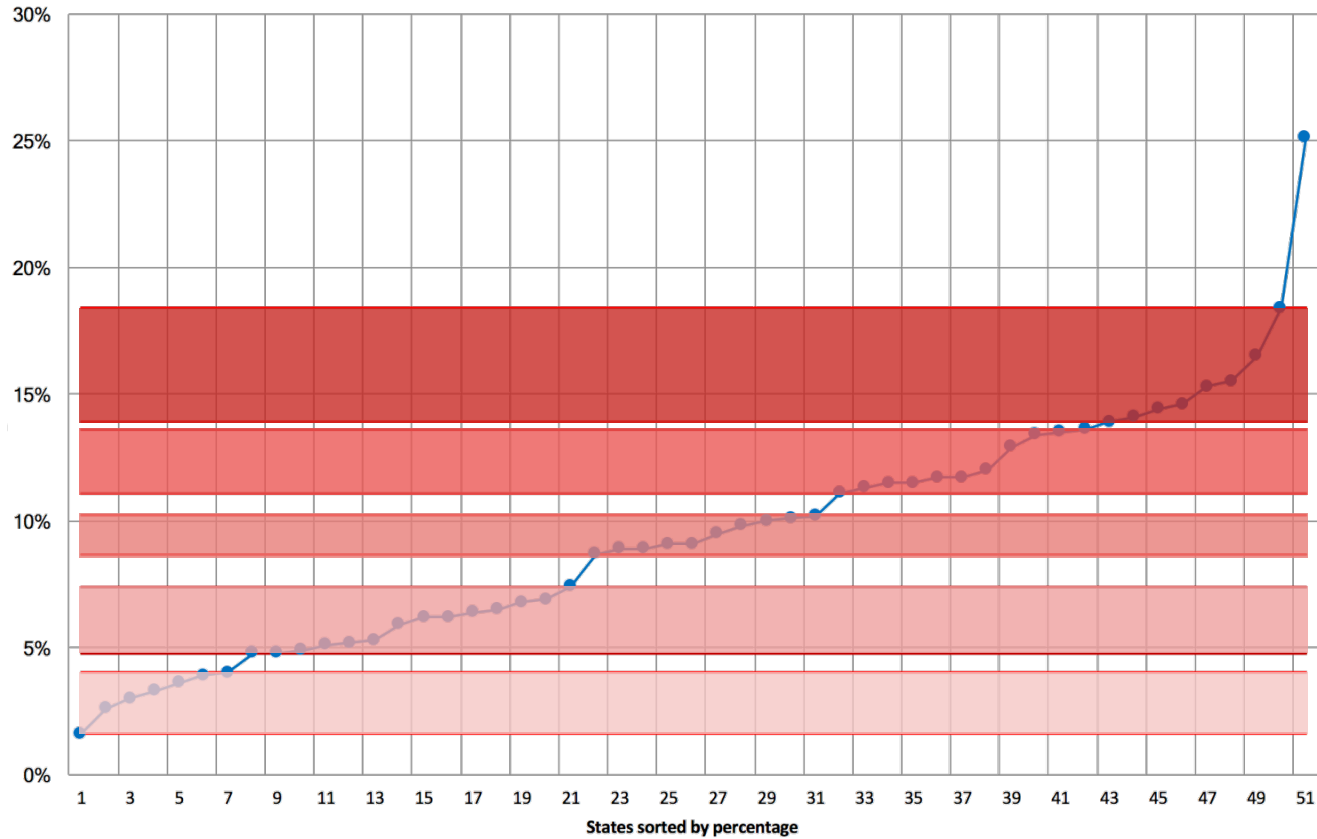
Improper categorization (and color choice)



Improper categorization



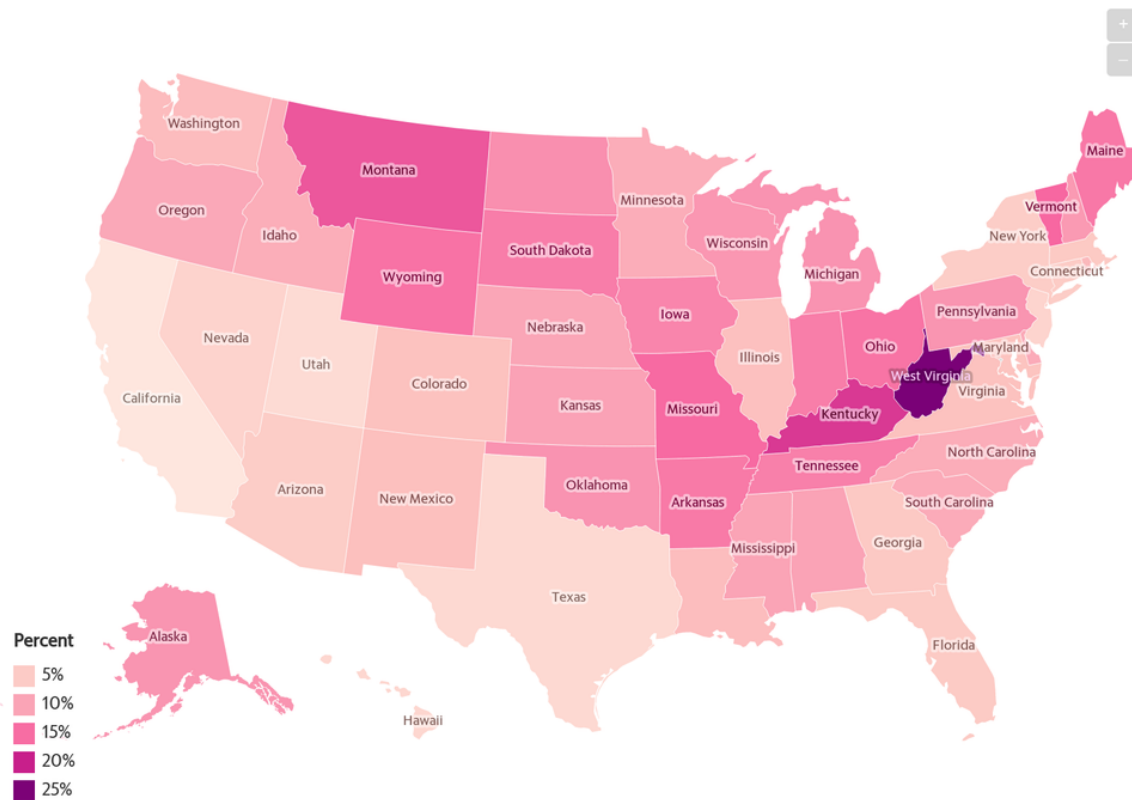
Improper categorization



Fixed categorization

SMOKING DURING PREGNANCY

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



Source: CDC • [Get the data](#)

Oversimplifying

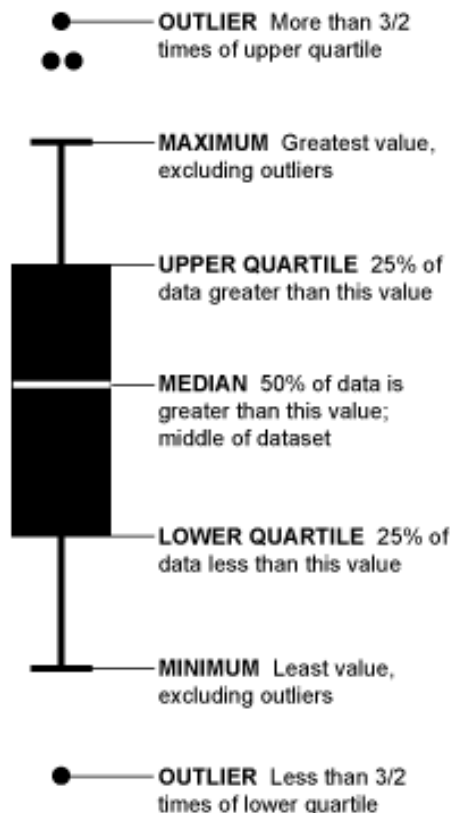
Clarify, not simplify!

To clarify, add detail.

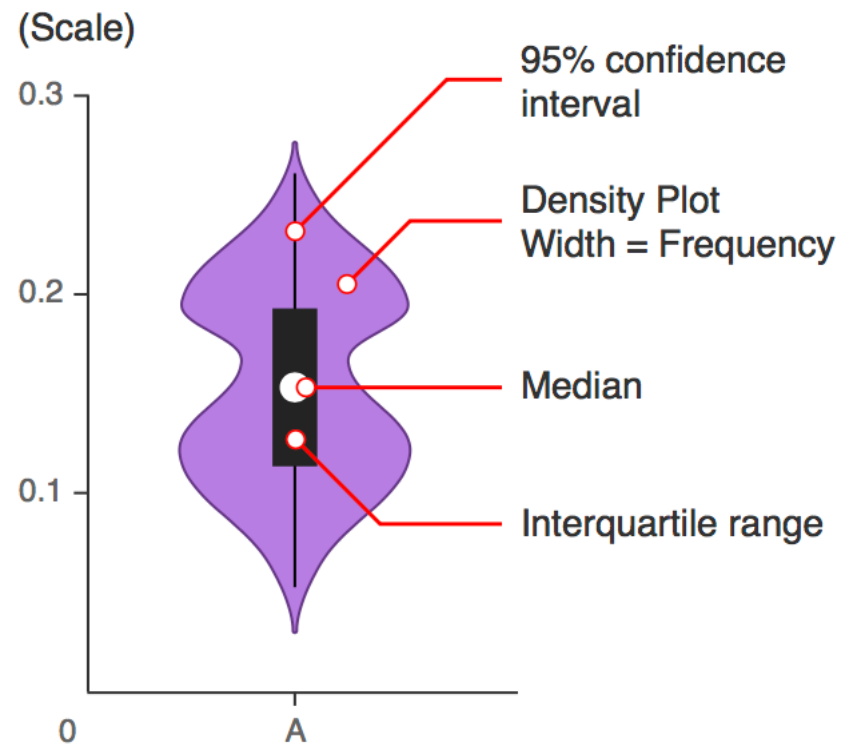
Edward Tufte

Box plot vs. violin plot

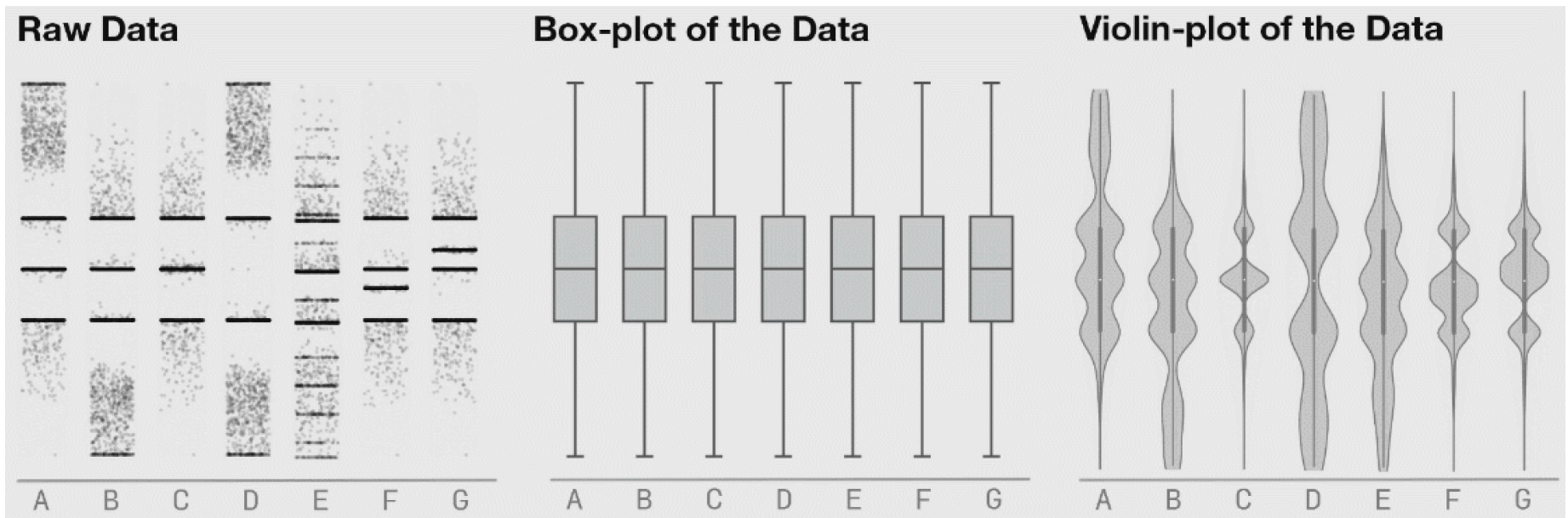
Box (and whisker) plot



Violin plot



Oversimplifying



How charts lie?

Phenomenon



Data



Dubious data

Chart



Misrepresenting data

Cherry-picking data

Ignoring uncertainty

Person



Confirmation bias

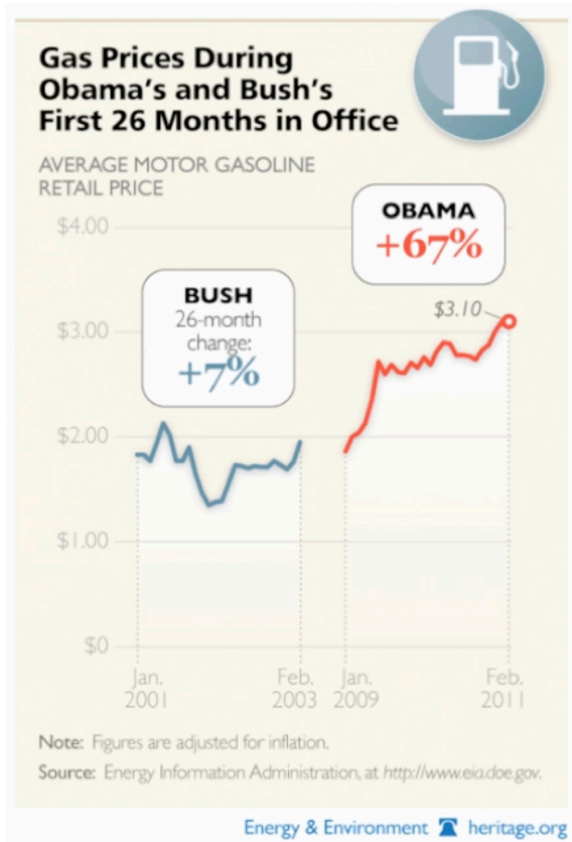
Cherry-picking data

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

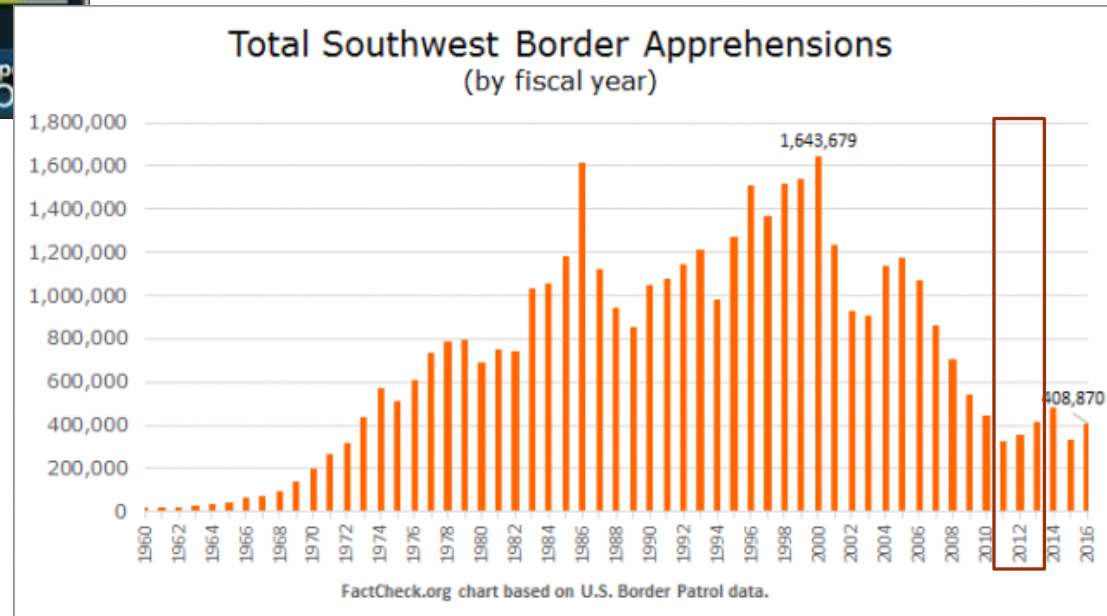
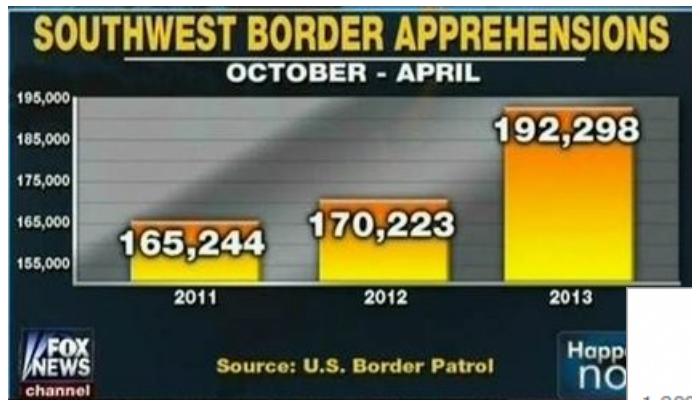
- Hiding (unfavorable) data
- Concealing existing patterns
- Suggesting patterns that are not there

Correlation \neq causation

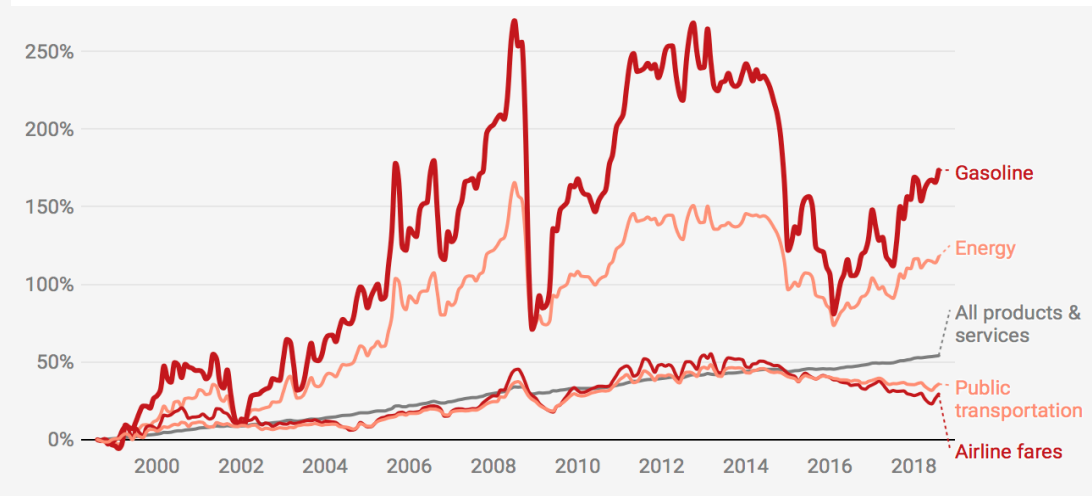
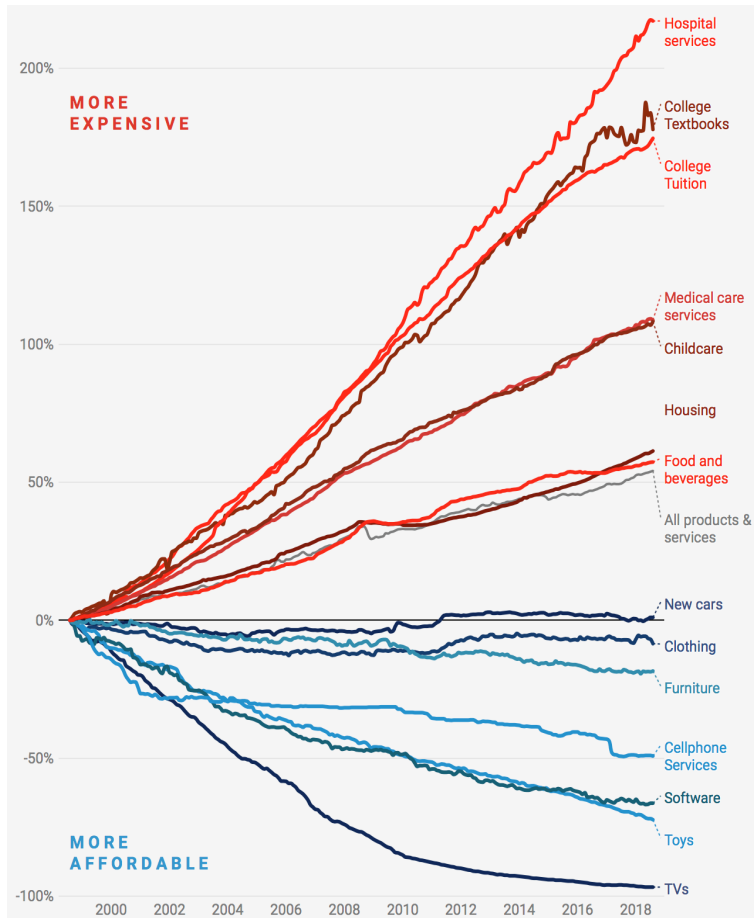
Hiding (unfavorable) data



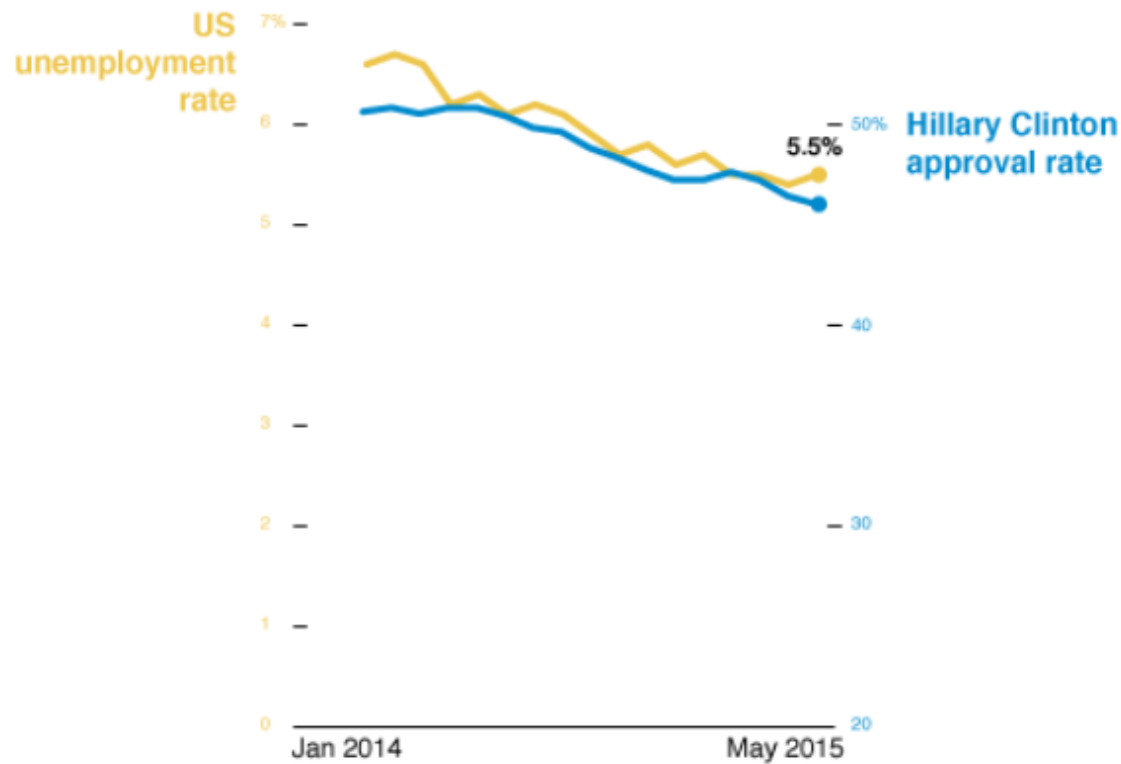
Concealing existing patterns



Concealing existing patterns

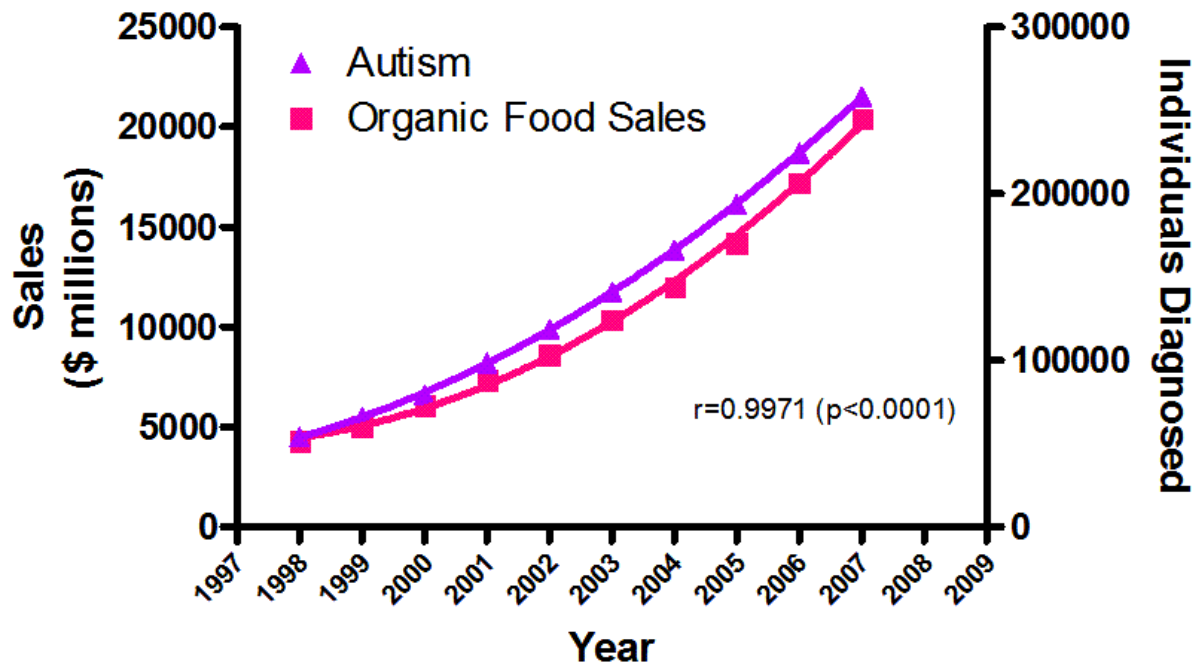


Suggesting patterns that are not there



Suggesting patterns that are not there

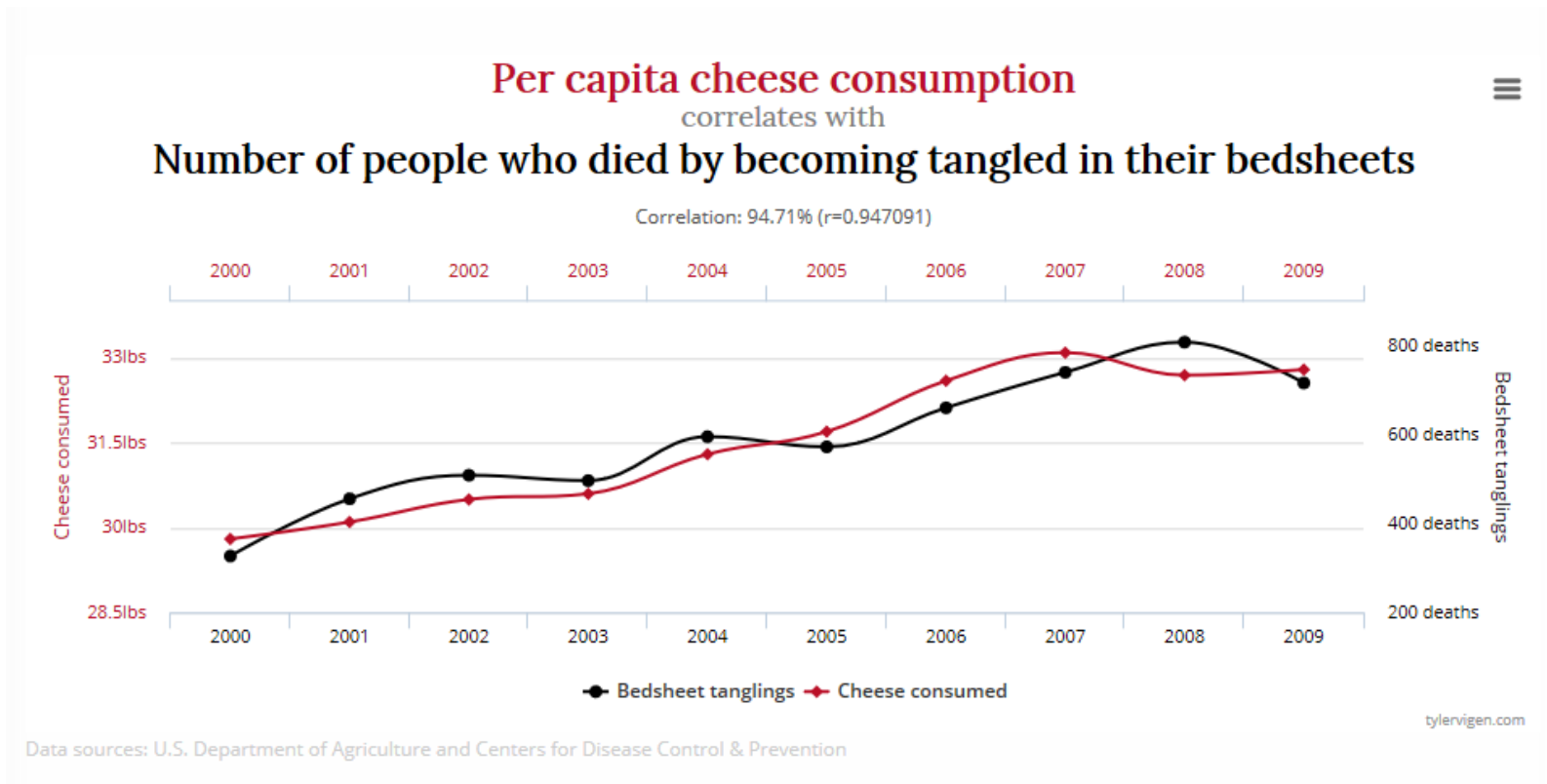
The real cause of increasing autism prevalence?



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"

Suggesting patterns that are not there

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



How charts lie?

Phenomenon



Data



Dubious data

Chart



Misrepresenting data

Cherry-picking data

Ignoring uncertainty

Person



Confirmation bias

Ignoring uncertainty

- Misrepresenting uncertainty
- Concealing uncertainty

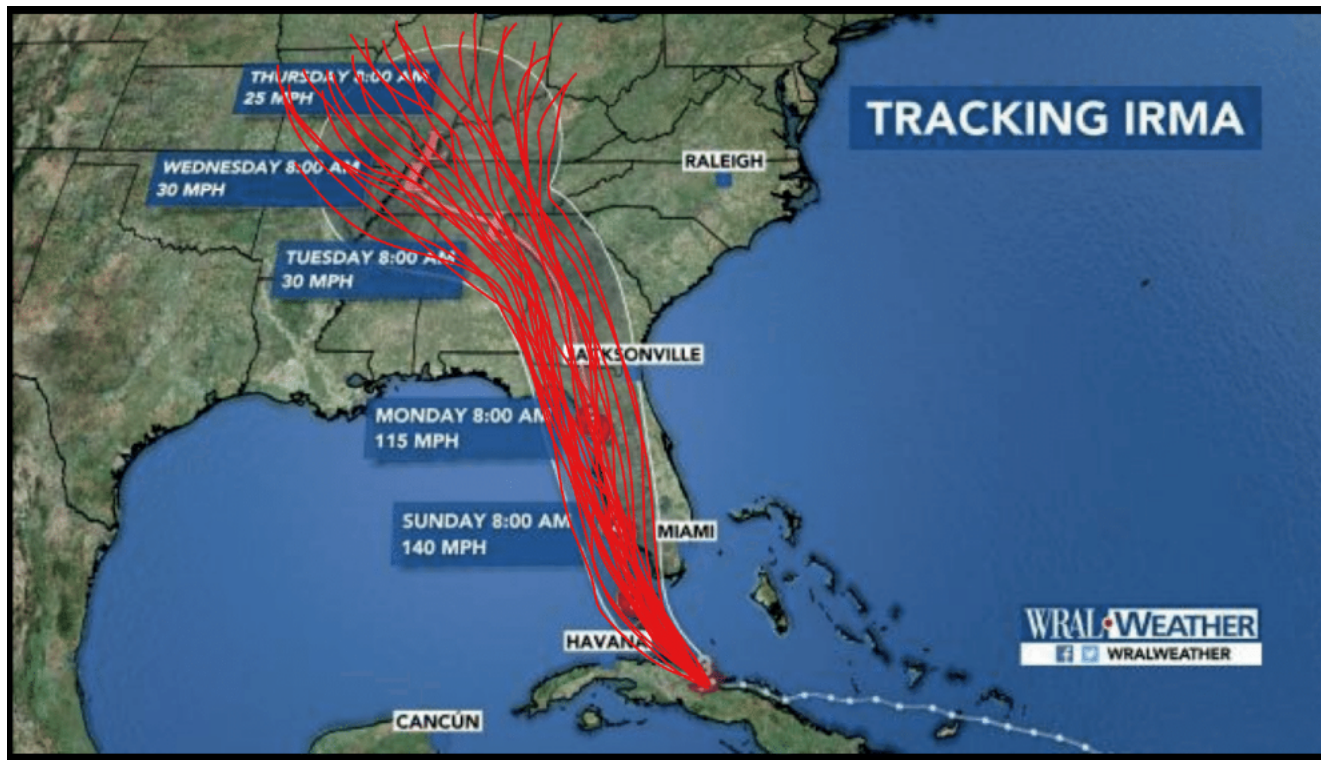
Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted



Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted



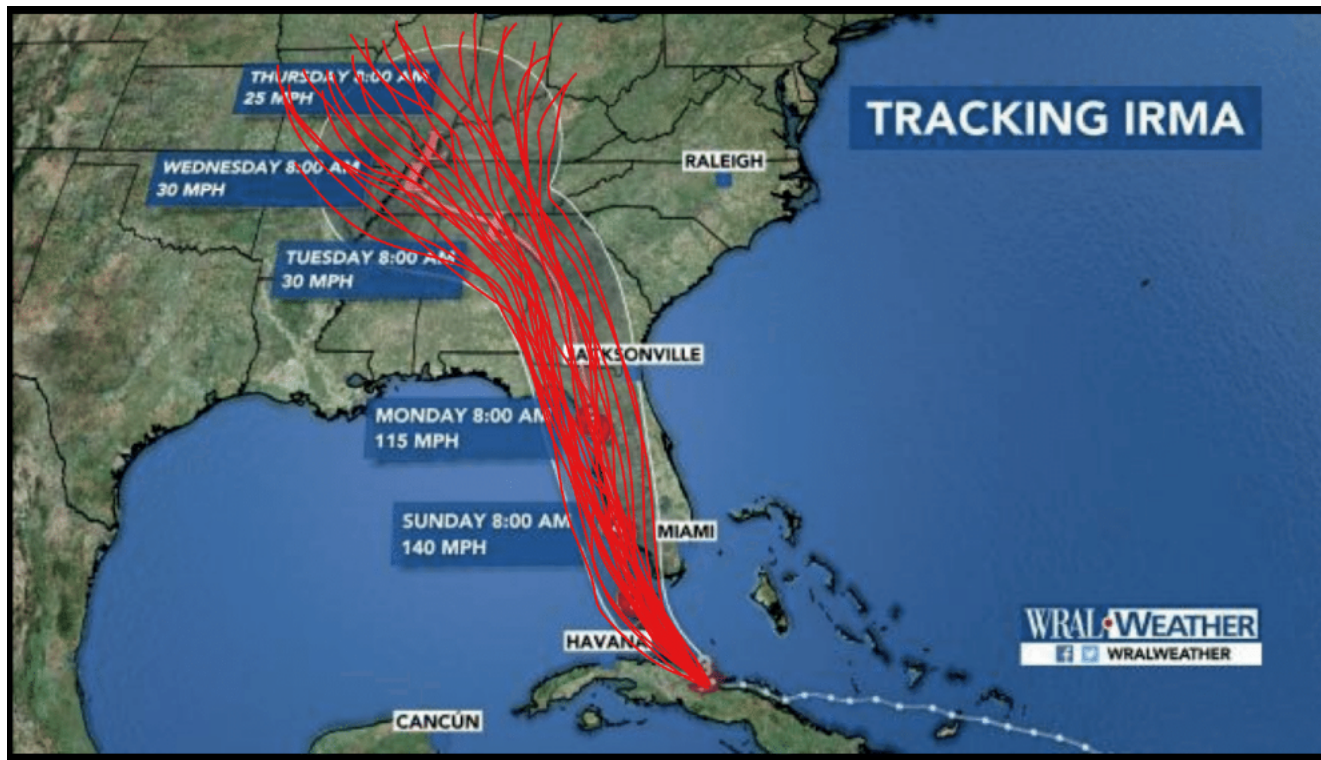
Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted



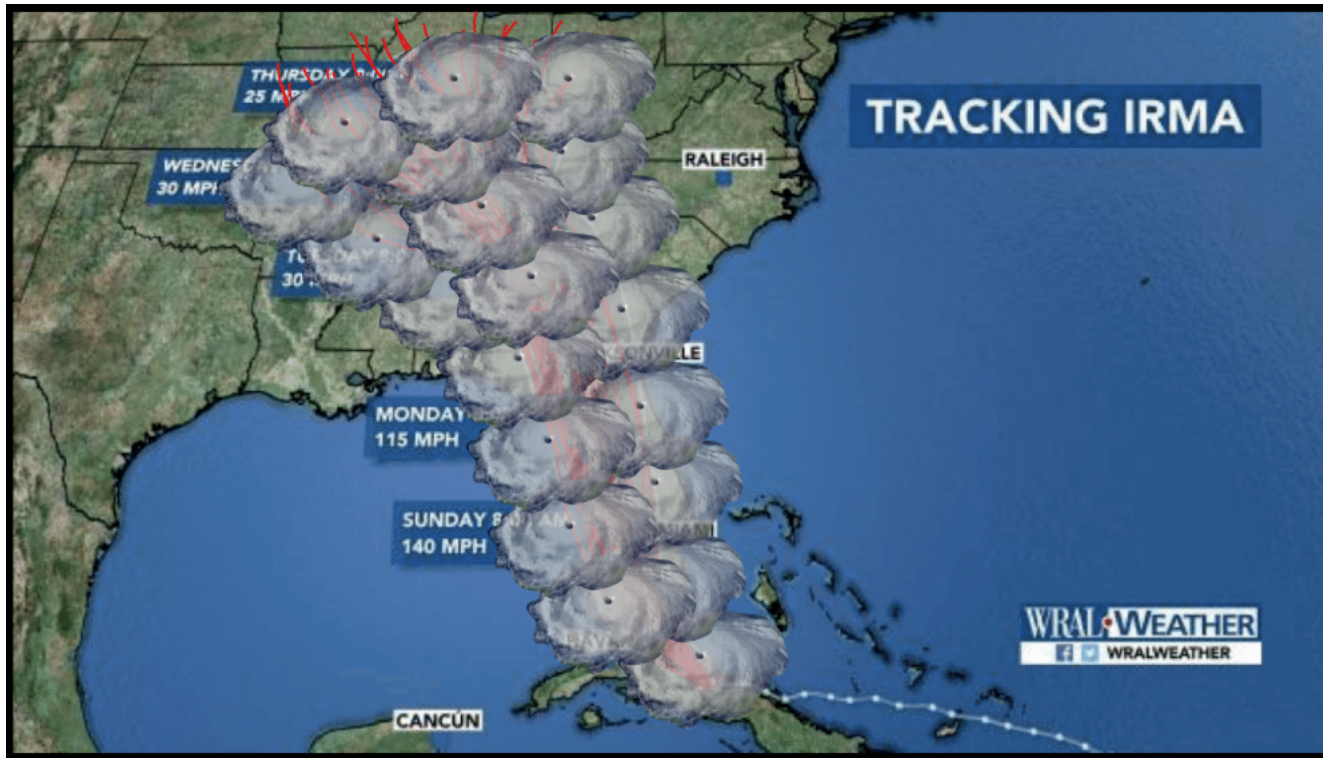
Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted



Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted



Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted

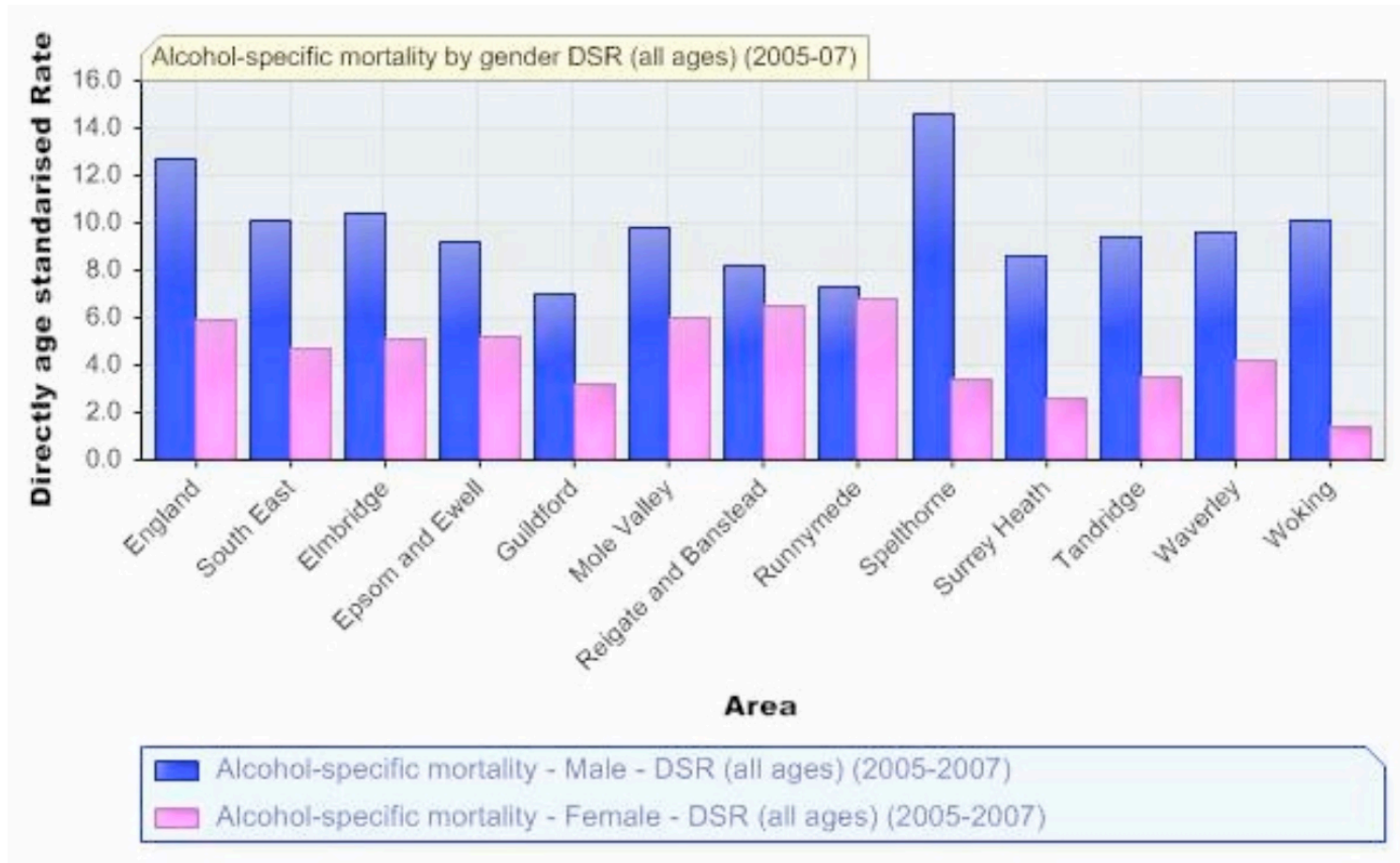


Misrepresenting uncertainty

The cone of uncertainty is widely misinterpreted

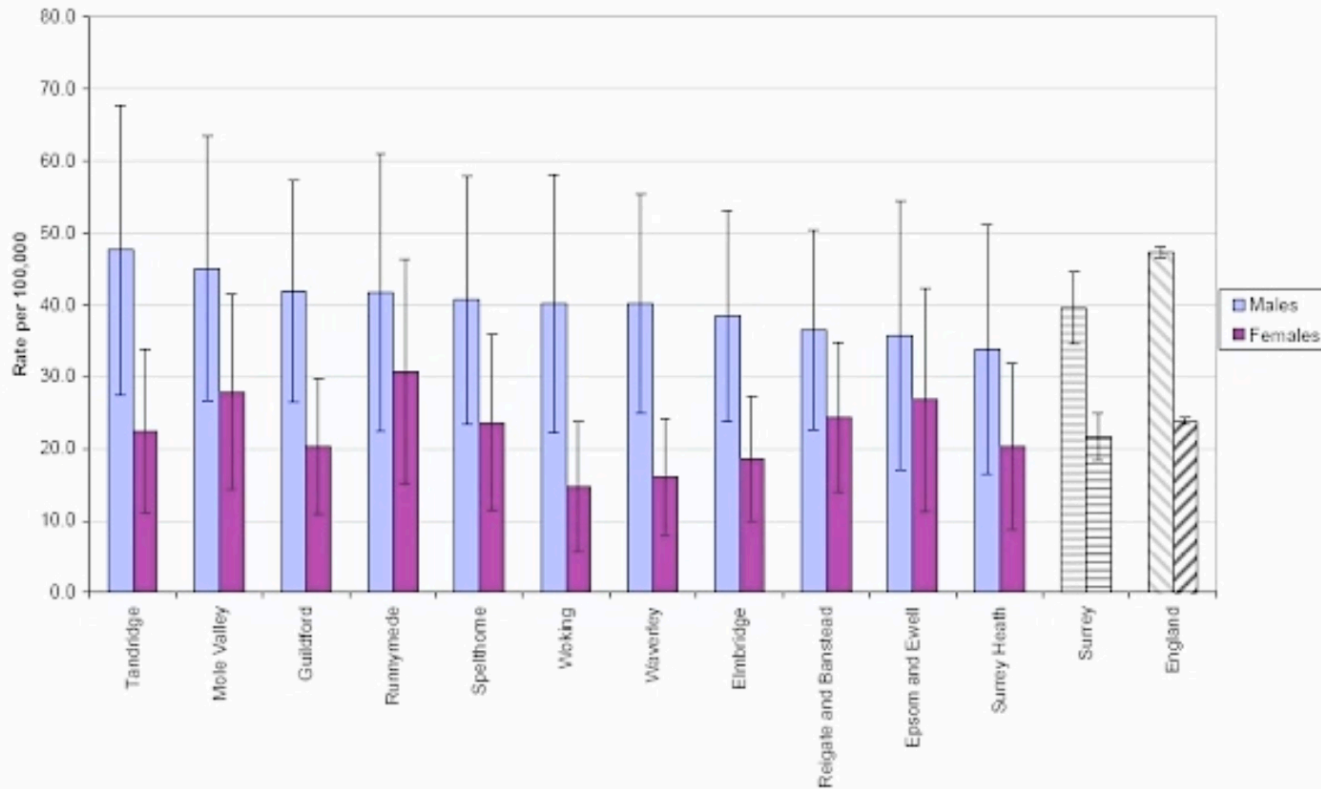


Concealing uncertainty



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).



How charts lie?

Phenomenon



Data



Dubious data

Chart



Misrepresenting data

Cherry-picking data

Ignoring uncertainty

Person

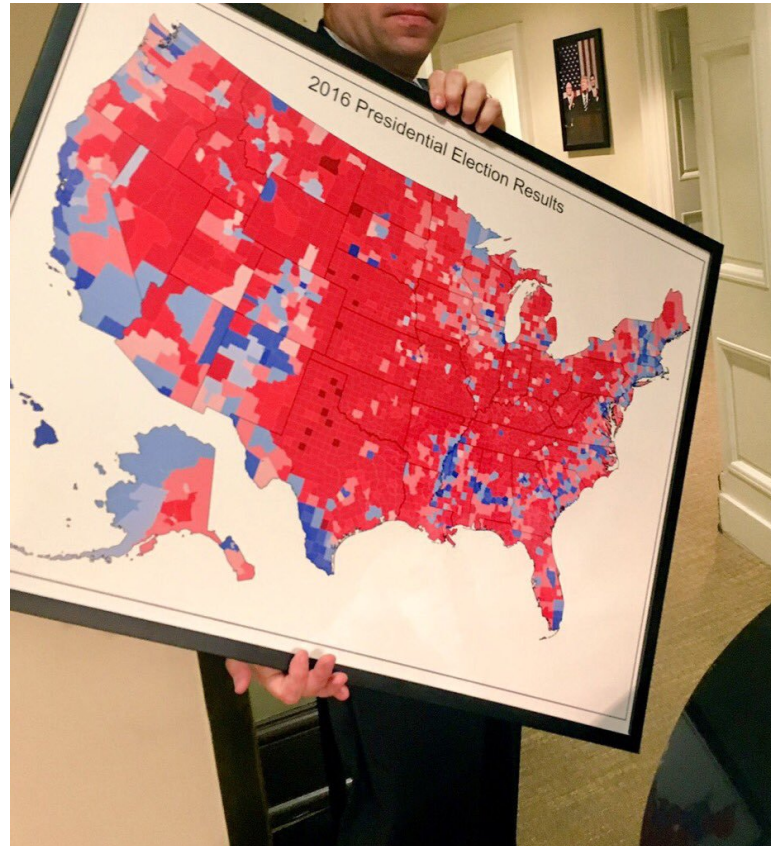


Confirmation bias

Confirmation bias

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

Confirmation bias



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

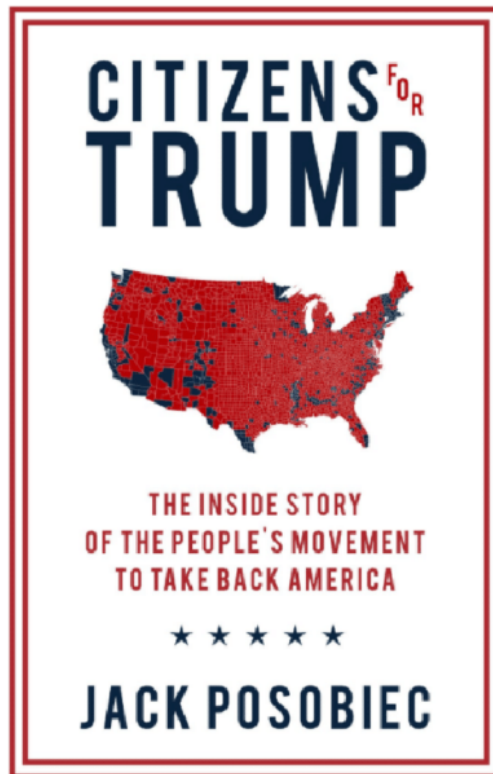
Nancy Steffan @steffan_nancy [Follow](#)

Replying to @trevorHartje @realDonaldTrump

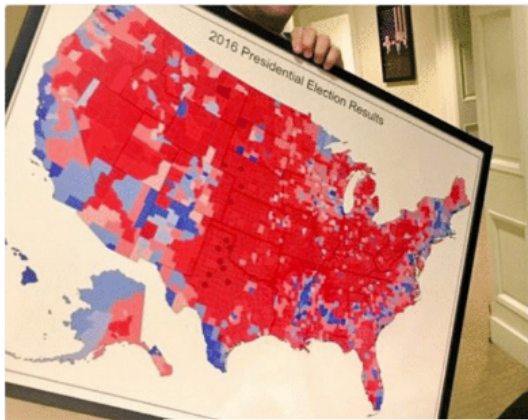
Really?



Confirmation bias



Confirmation bias

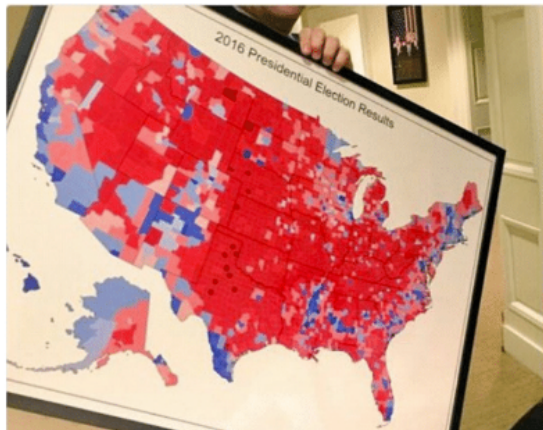


Surface on the
county-level map:

Red: 80%

Blue: 20%

Confirmation bias



Surface on the county-level map:

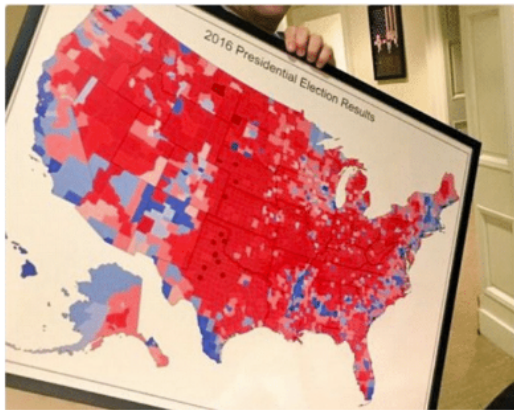
Red: 80%

Blue: 20%

SHARE OF THE POPULAR VOTE IN THE 2016 PRESIDENTIAL ELECTION



Confirmation bias

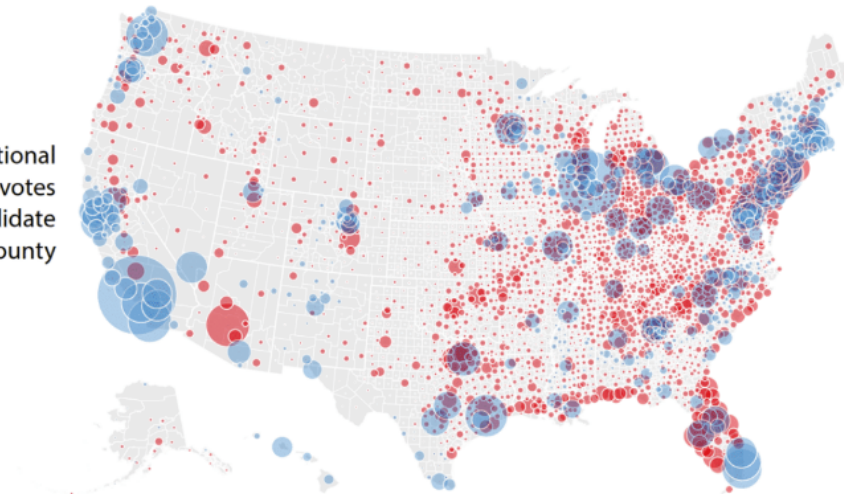


Surface on the county-level map:

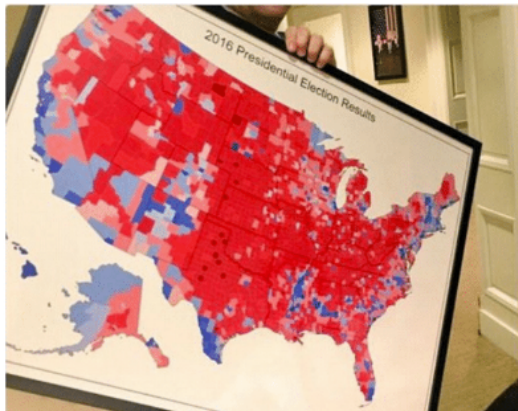
Red: 80%

Blue: 20%

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



Confirmation bias



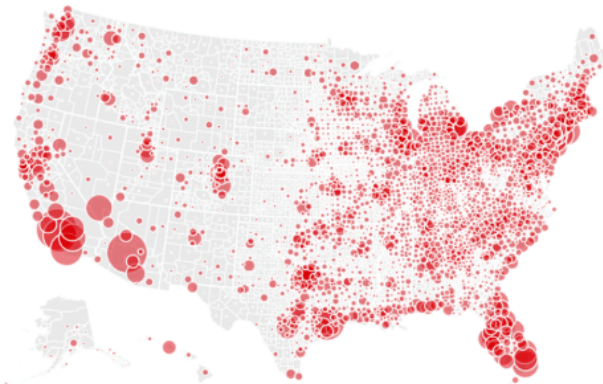
SHARE OF THE POPULAR VOTE IN THE 2016 PRESIDENTIAL ELECTION



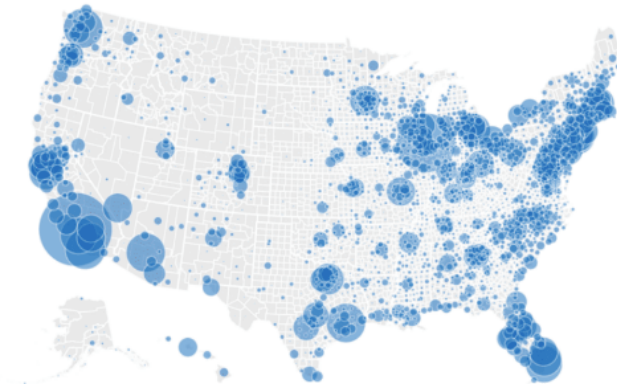
PERCENTAGE OF ELIGIBLE VOTERS



VOTES FOR DONALD TRUMP



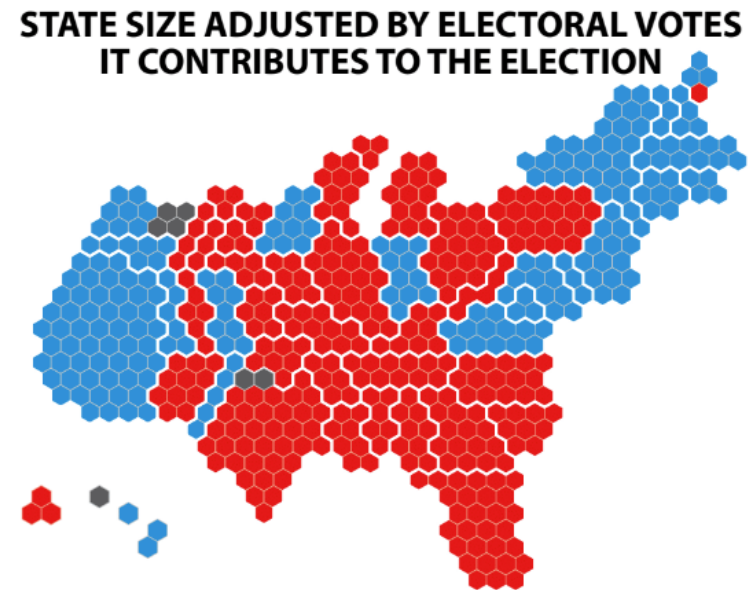
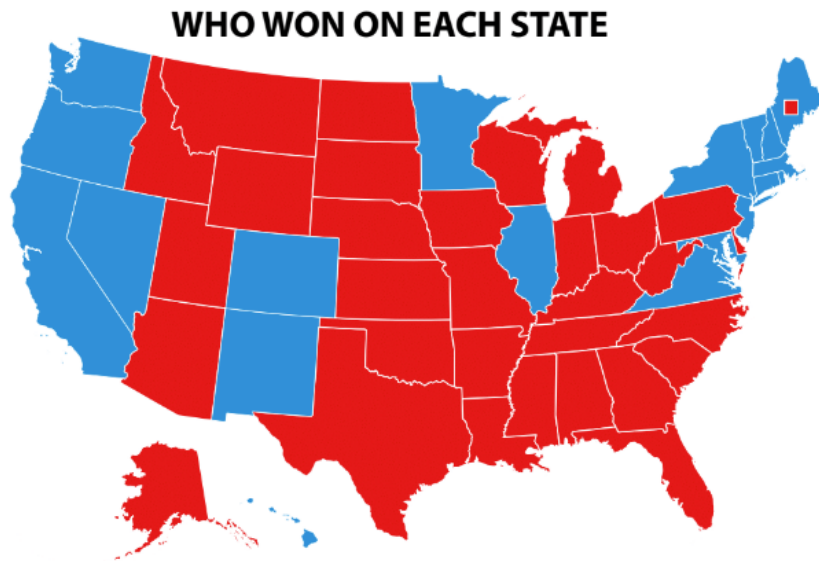
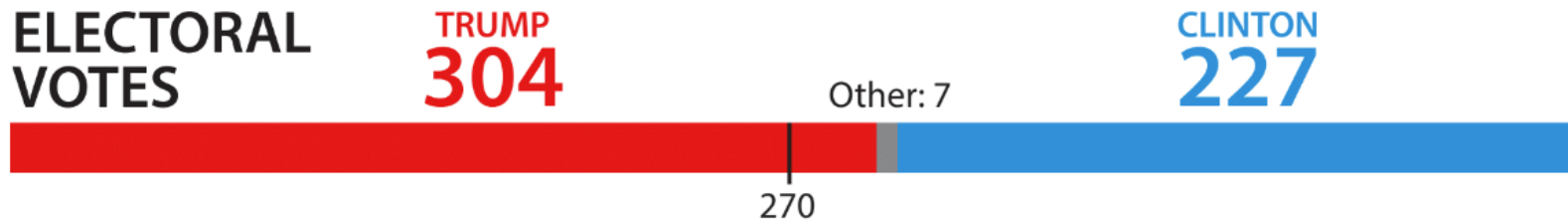
VOTES FOR HILLARY CLINTON



Bubble size is proportional to the number of votes per county

Confirmation bias

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



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)