Misrepresentation

Data Science in a Box

datasciencebox.org

Causality

TIME

Exercise Can Lower Risk of Some Cancers By 20%

People who were more active had on average a 20% lower risk of cancers of the esophagus, lung, kidney, stomach, endometrium and others compared with people who were less active.

Alice Park. Exercise Can Lower Risk of Some Cancers By 20%. Time Magazine. 16 May 2016.

Los Angeles Times

Exercising drives down risk for 13 cancers, research shows

[…] those who got the most moderate to intense exercise reduced their risk of developing seven kinds of cancer by at least 20%.

Melissa Healy. Exercising drives down risk for 13 cancers, research shows .
Los Angeles Times. 16 May 2016.

Original study

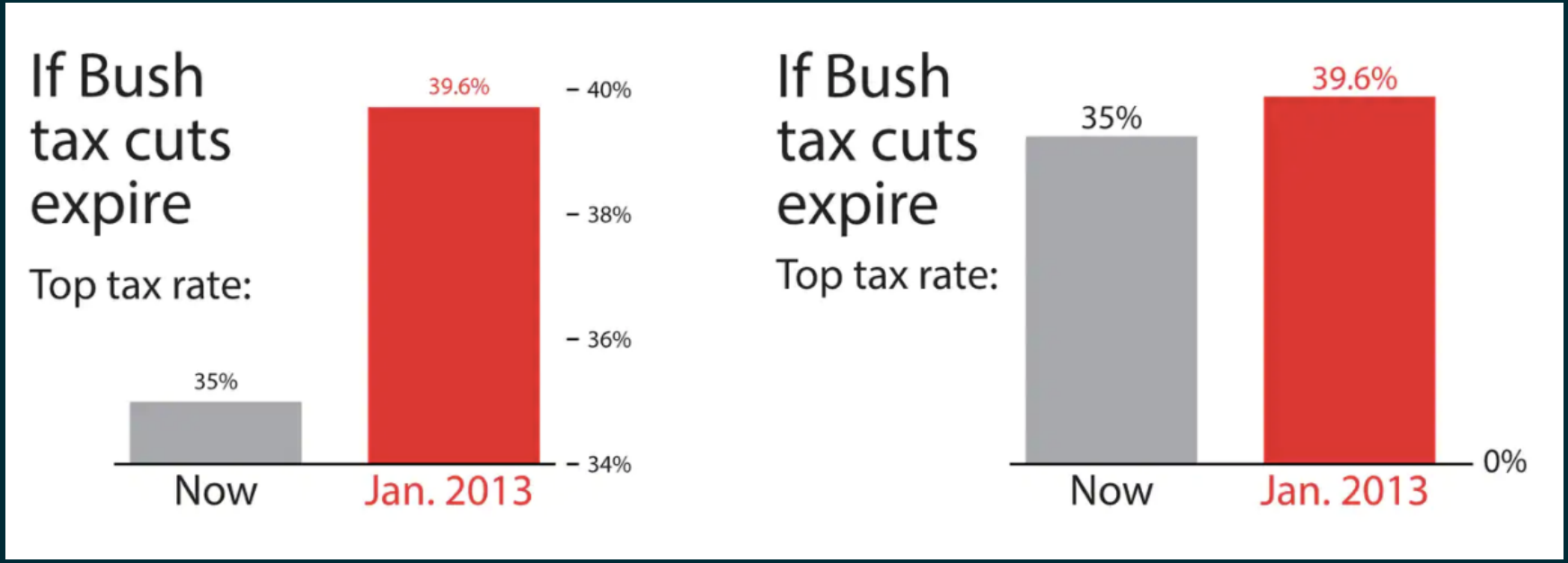
Moore, Steven C., et al. **"Association of leisure-time physical activity with risk of 26 types of cancer in 1.44 million adults."** JAMA internal medicine 176.6 (2016): 816-825.

- **Volunteers** were **asked** about their physical activity level over the preceding year.
- Half exercised less than about 150 minutes per week, half exercised more.
- Compared to the bottom 10% of exercisers, the top 10% had lower rates of esophageal, liver, lung, endometrial, colon, and breast cancer.
- Researchers found no association between exercising and 13 other cancers (e.g. pancreatic, ovarian, and brain).

Carl Bergstrom and Jevin West. Calling Bullshit: The art of skepticism in a data-driven world. Random House, 2020.
Sharon Begley. "Does exercise prevent cancer?". StatNews. 16 May 2016.

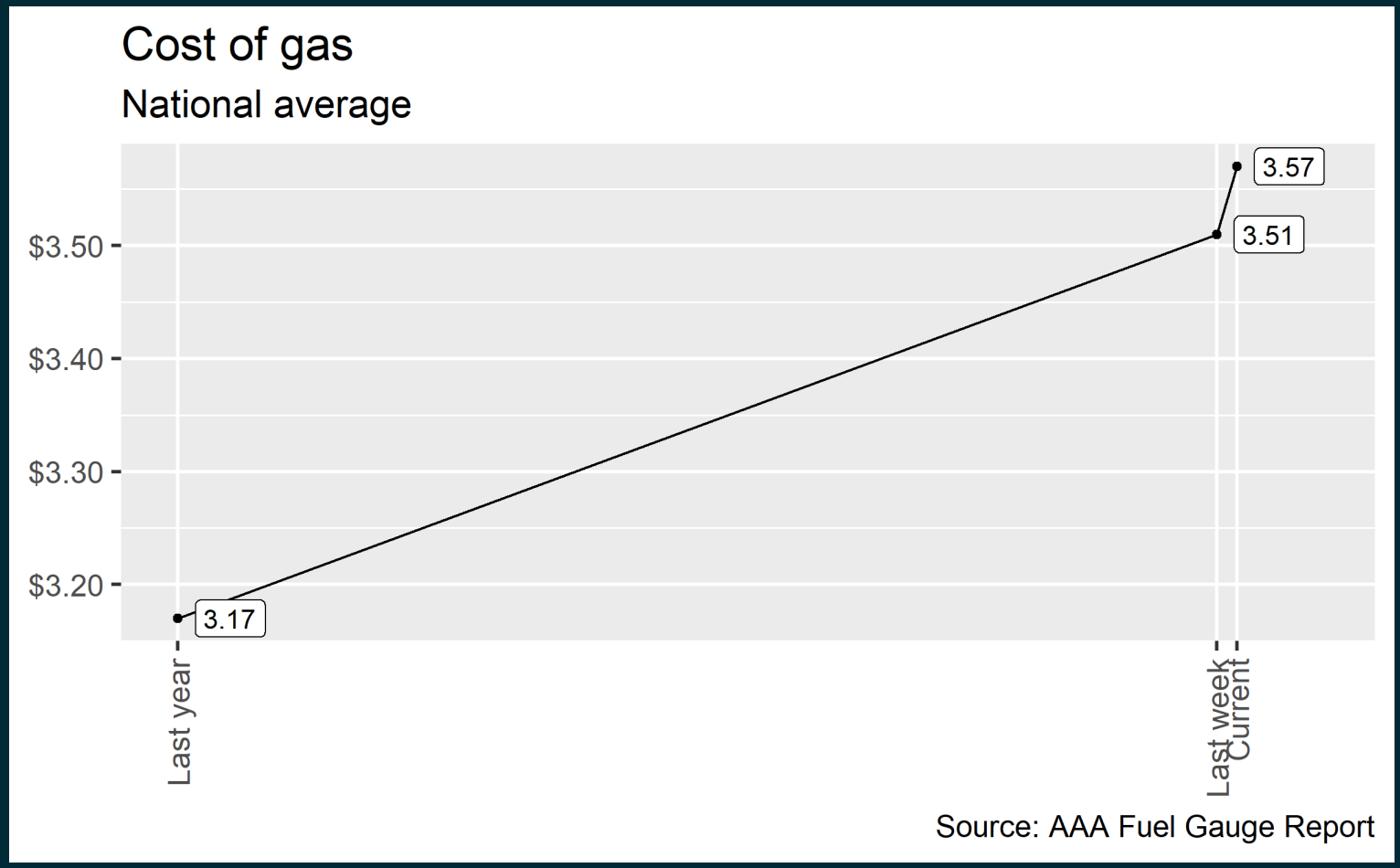
Axes and scale

What is the difference between these two pictures? Which presents a better way to represent these data?



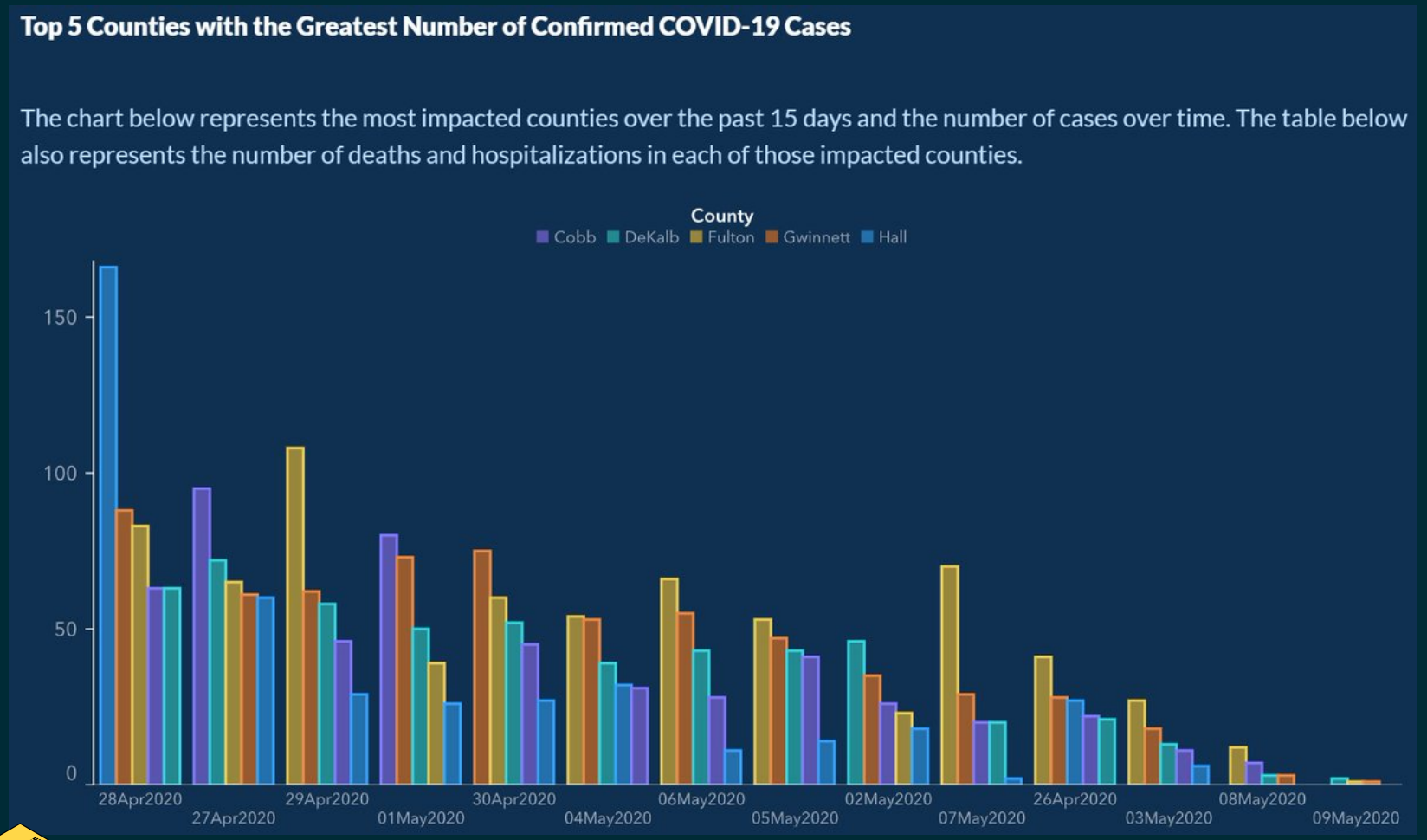
Christopher Ingraham. "You've been reading charts wrong. Here's how a pro does it.". The Washington Post. 14 October 2019.

What is wrong with this picture? How would you correct it?

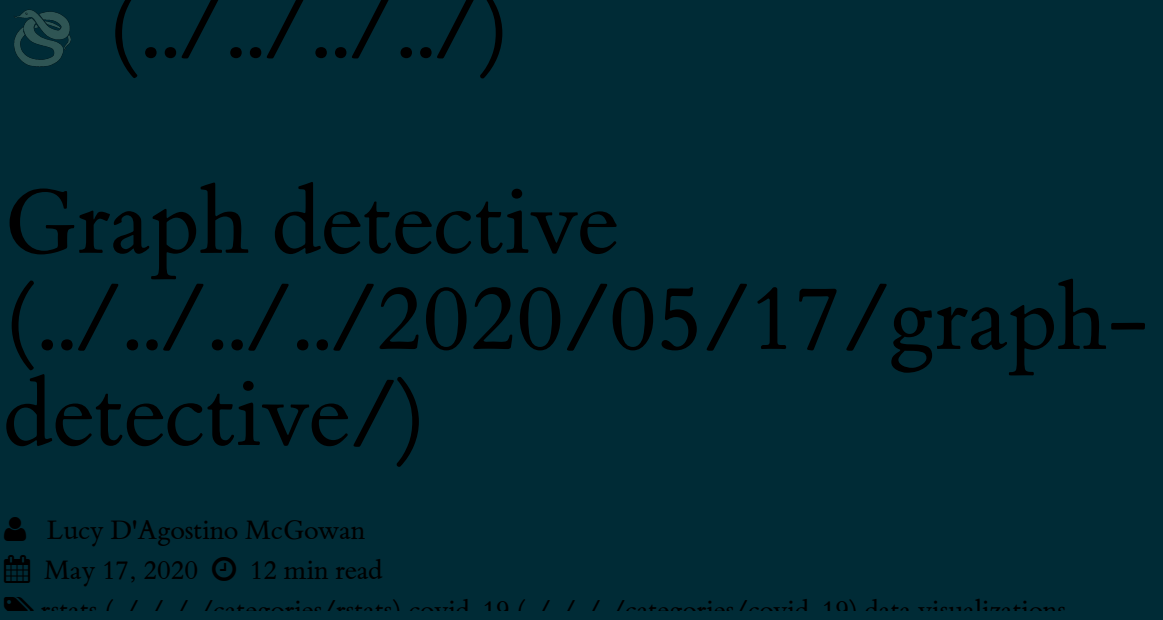

Cost of gas
National average
$3.50
$3.40
$3.30
$3.20
3.57
3.51
3.17
Last year
Last week
Current
Source: AAA Fuel Gauge Report


COST OF GAS
NATIONAL AVERAGE
$3.60
$3.50
$3.40
$3.30
$3.20
$3.10
$3.00
$2.90
$3.17
$3.51
$3.57
LAST YEAR
LAST WEEK
CURRENT
FOX NEWS channel
SOURCE: AAA FUEL GAUGE REPORT
PAY... THE STRIKE, WHICH BEGAN LAST WEEK, IS
DOW 45.79

What is wrong with this picture? How would you correct it?



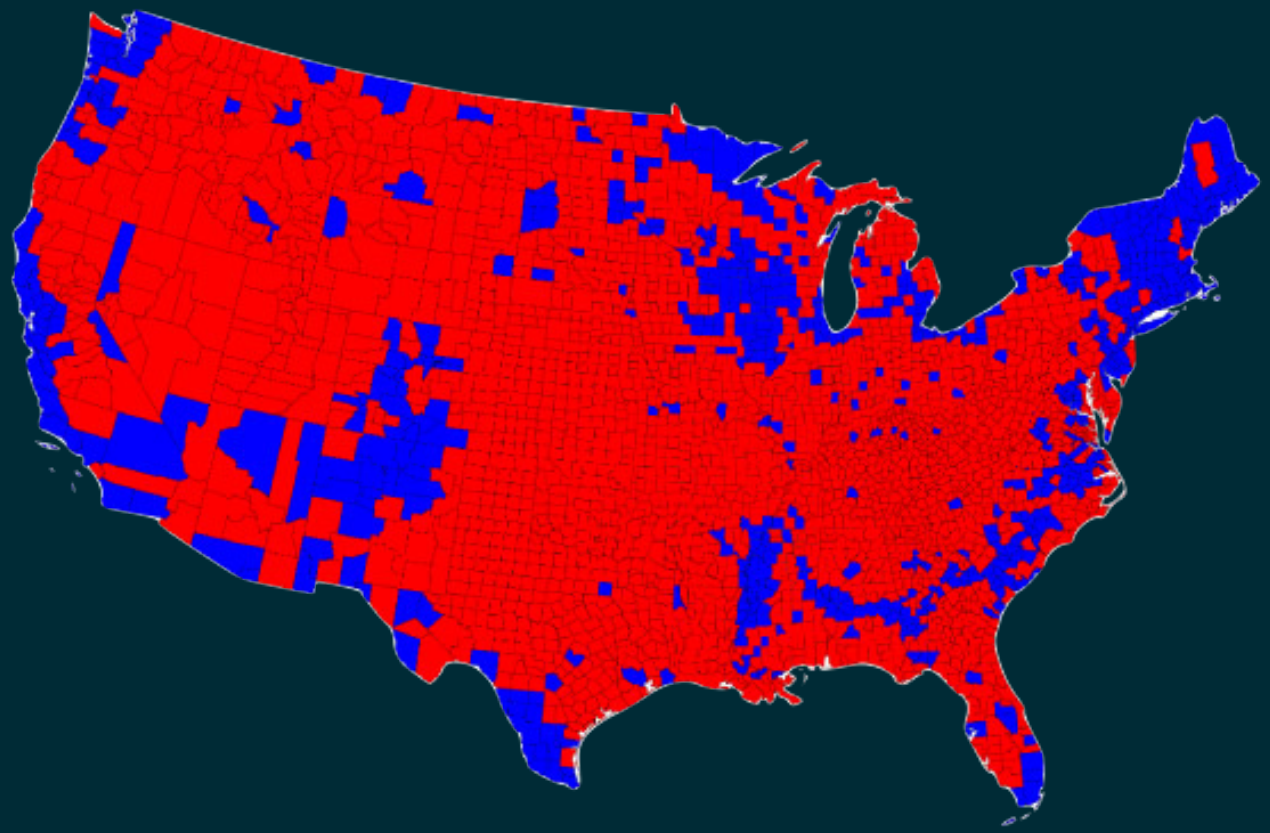
Graph detective



Lucy D'Agostino McGowan. Graph detective. Live Free or Dichotomize. 17 May 2020.

Maps and areas

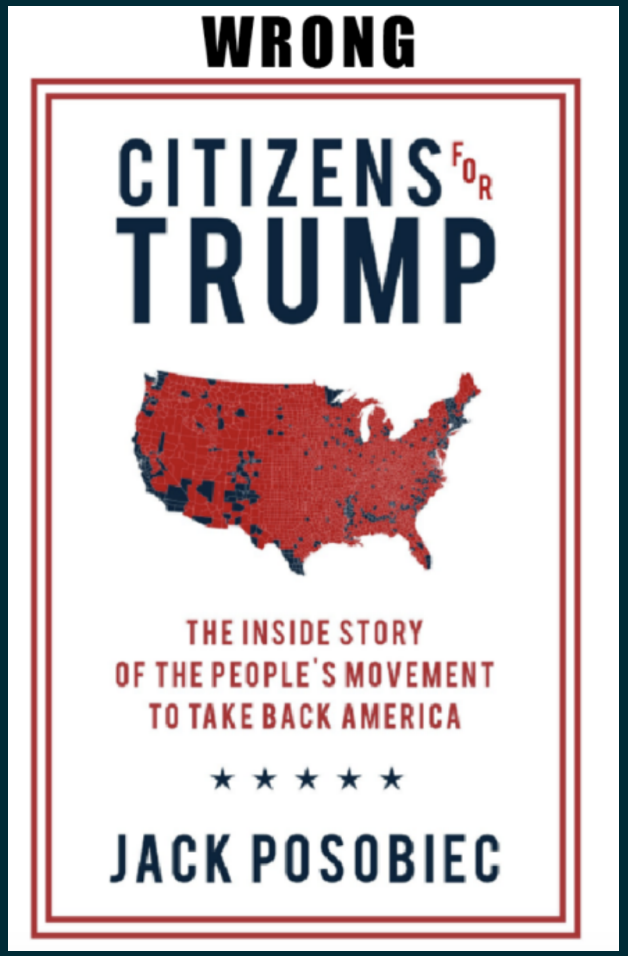
Do you recognize this map? What does it show?



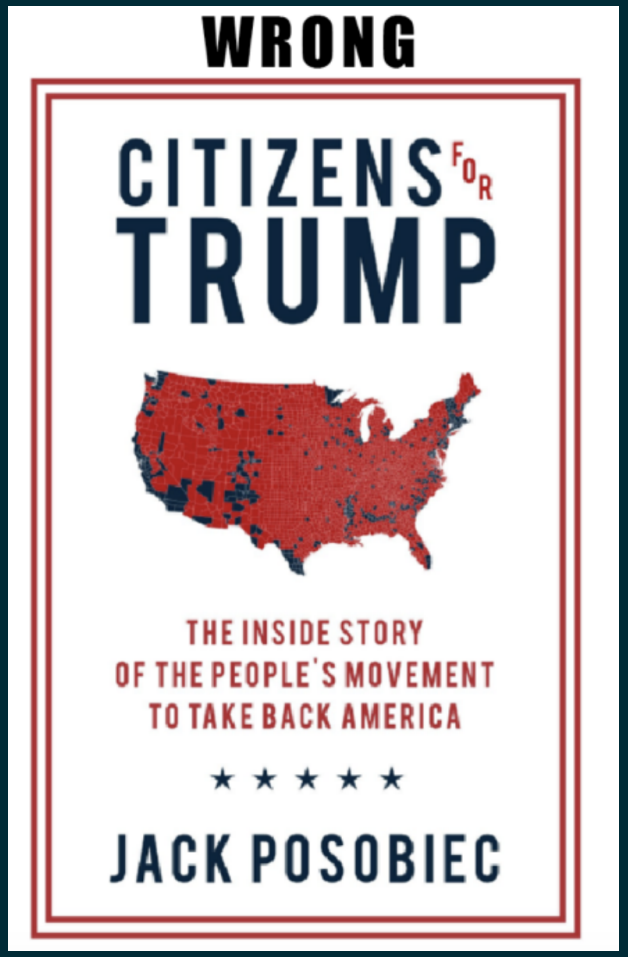
Do you recognize this map? What does it show?

Lazaro Gamio. "Election maps are telling you big lies about small things". The Washington Post. 1 Nov 2016.

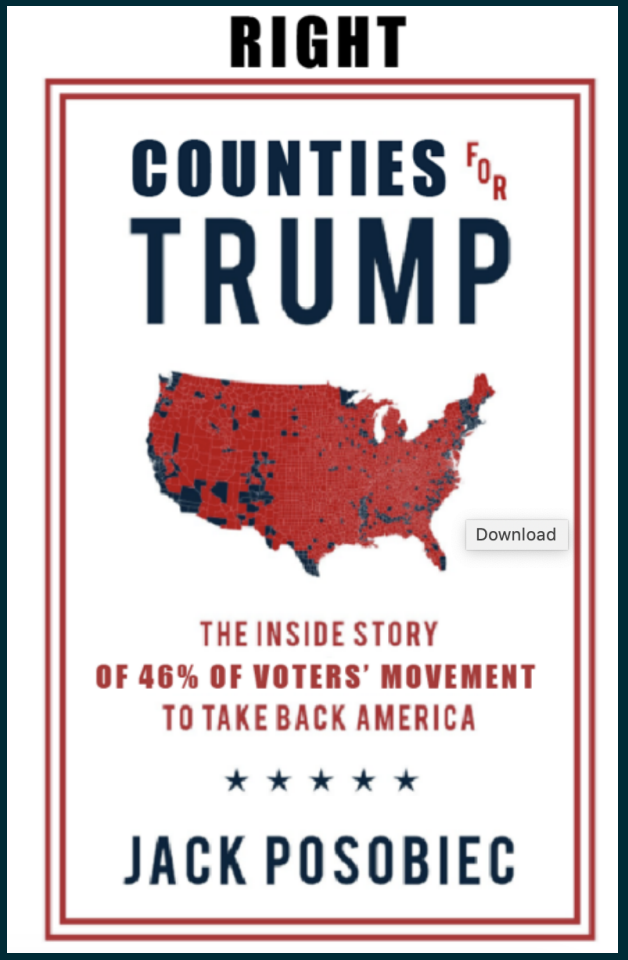
WRONG


CITIZENS FOR
TRUMP
THE INSIDE STORY
OF THE PEOPLE'S MOVEMENT
TO TAKE BACK AMERICA
★ ★ ★ ★ ★
JACK POSOBIEC

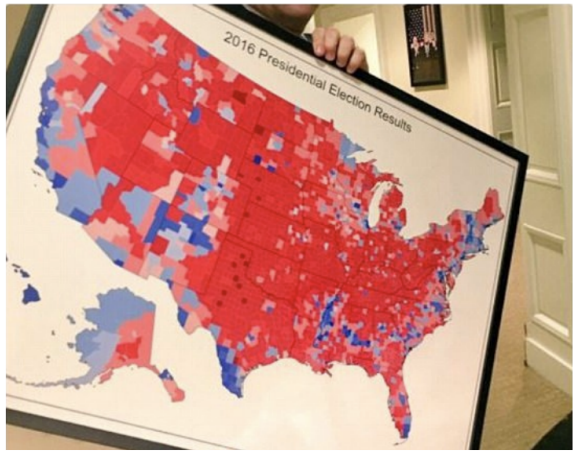
WRONG



RIGHT



Alberto Cairo. Visual Trumpery talk.



Surface on the county-level map:

Red: 80%

Blue: 20%

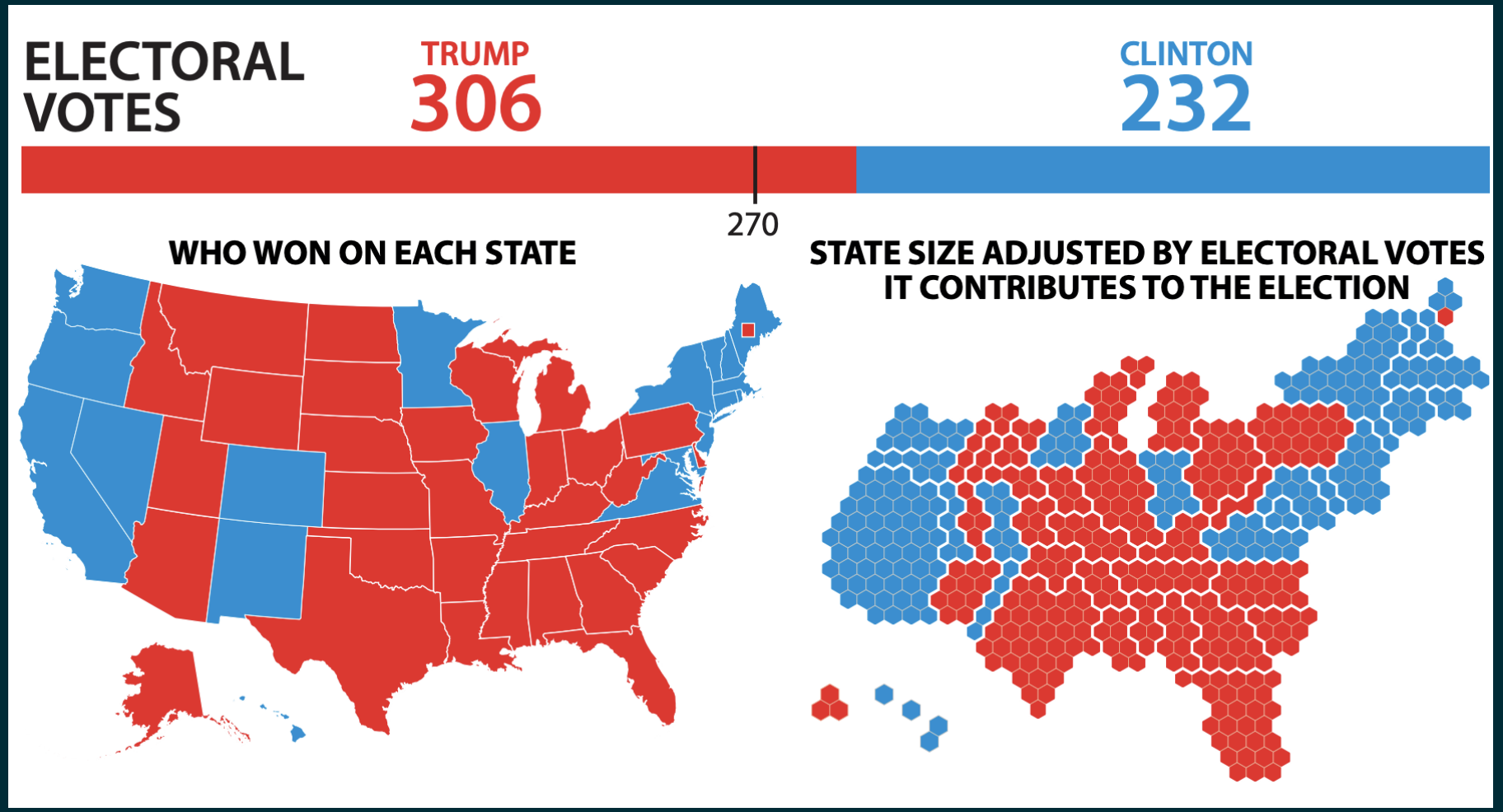
SHARE OF THE POPULAR VOTE IN THE 2016 PRESIDENTIAL ELECTION

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%

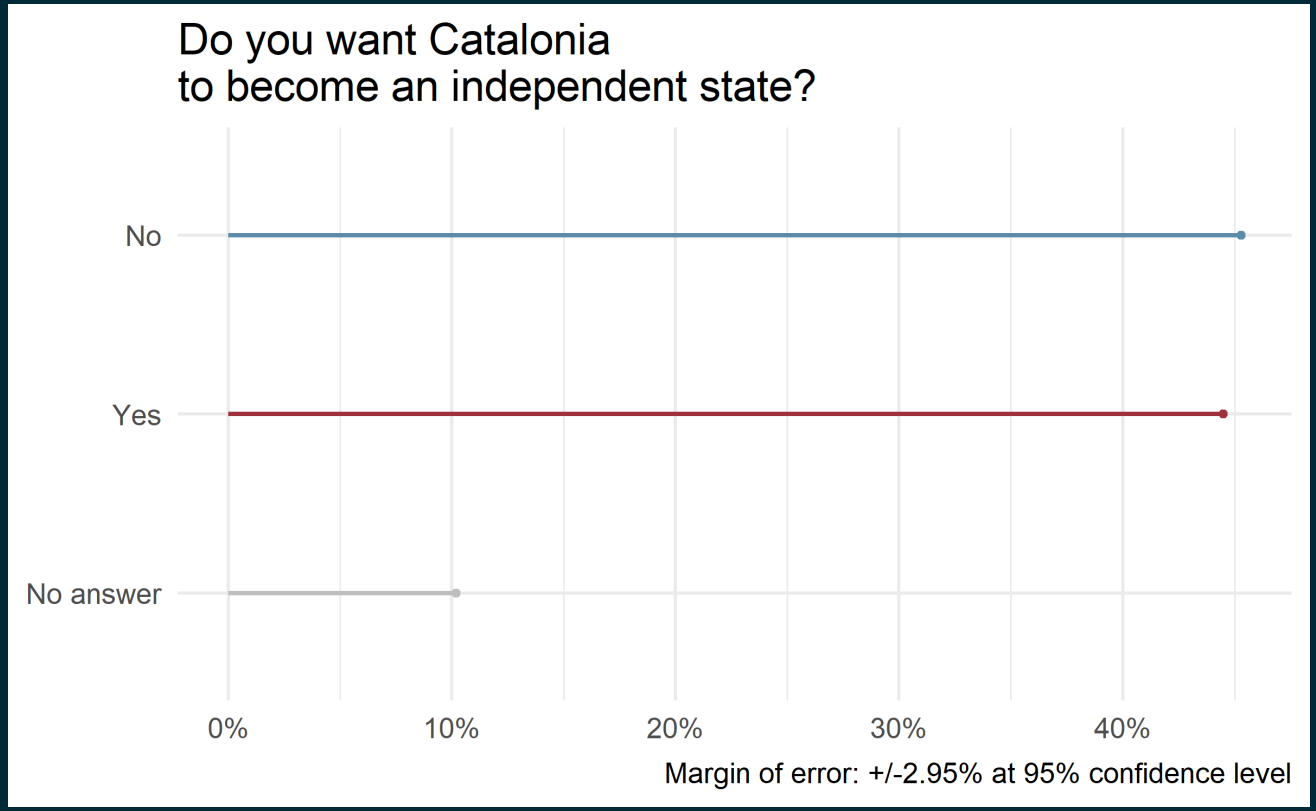
Alberto Cairo. Visual Trumpery talk.



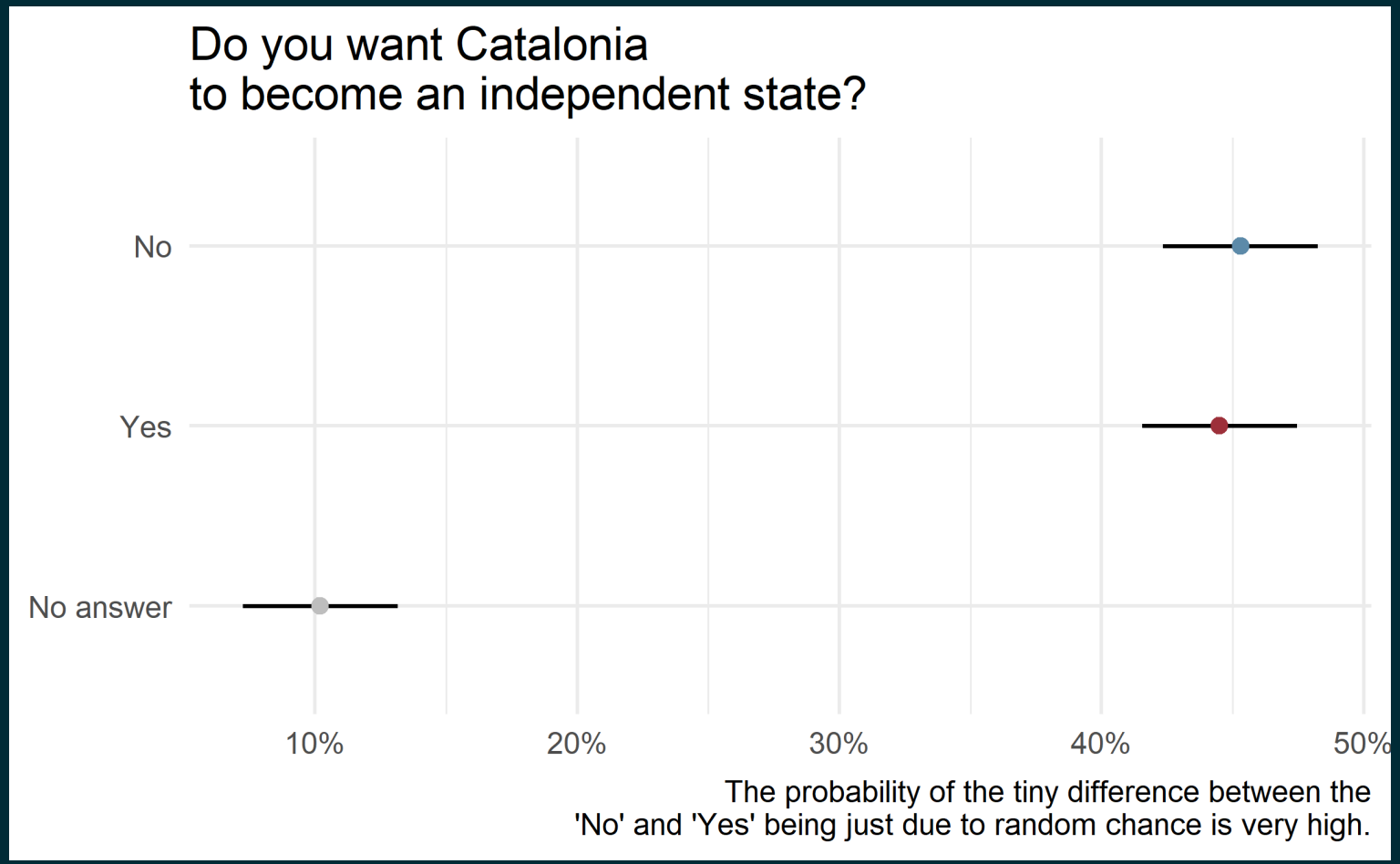
Alberto Cairo. Visual Trumpery talk.

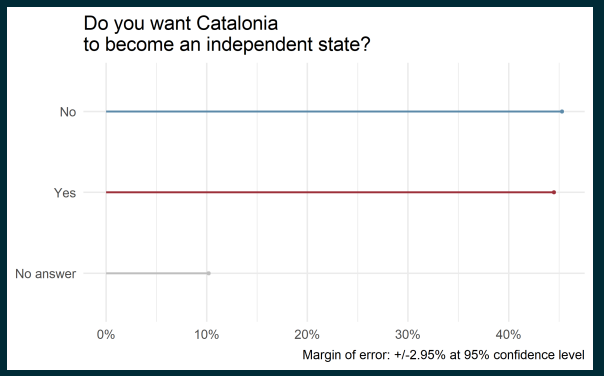
Visualising uncertainty

On December 19, 2014, the front page of Spanish national newspaper El País read *"Catalan public opinion swings toward 'no' for independence, says survey"*.



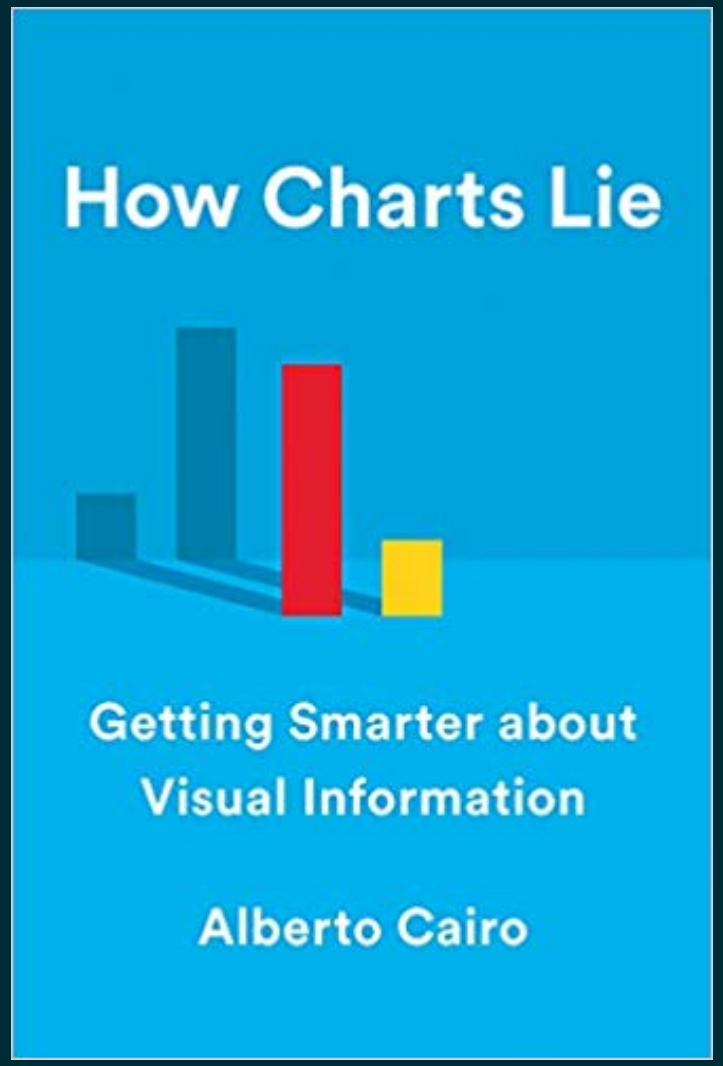
Alberto Cairo. The truthful art: Data, charts, and maps for communication. New Riders, 2016.





Alberto Cairo. "Uncertainty and Graphicacy: How Should Statisticians Journalists and Designers Reveal Uncertainty in Graphics for Public Consumption?", Power from Statistics: Data Information and Knowledge, 2017.

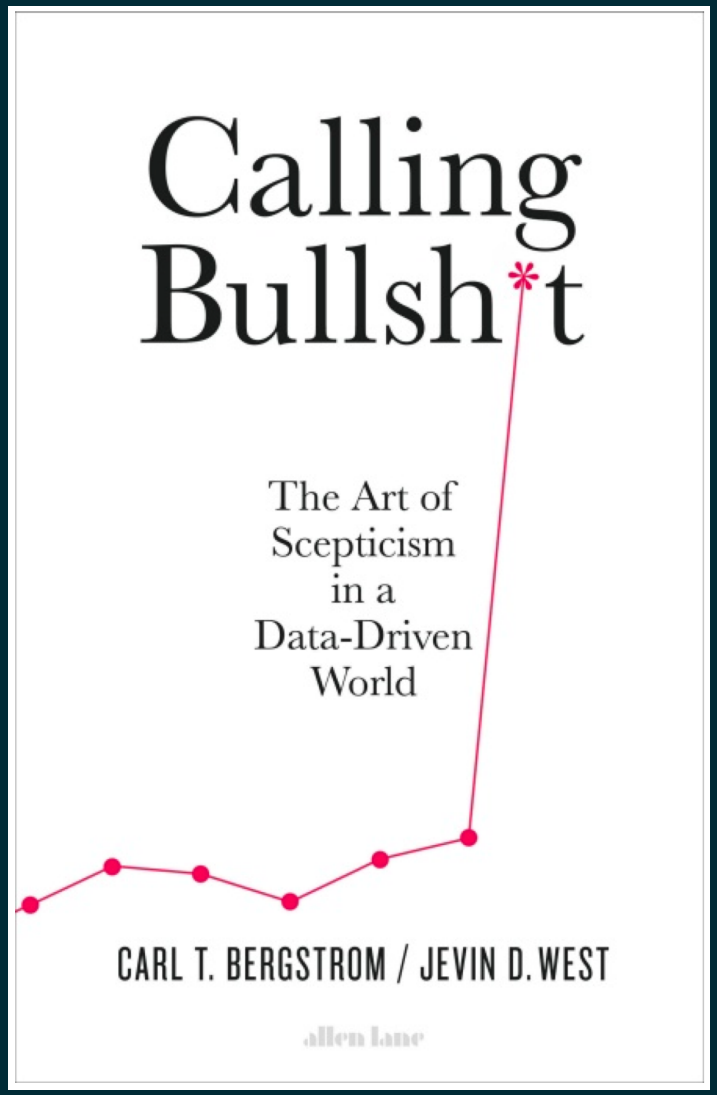
Further reading



How Charts Lie

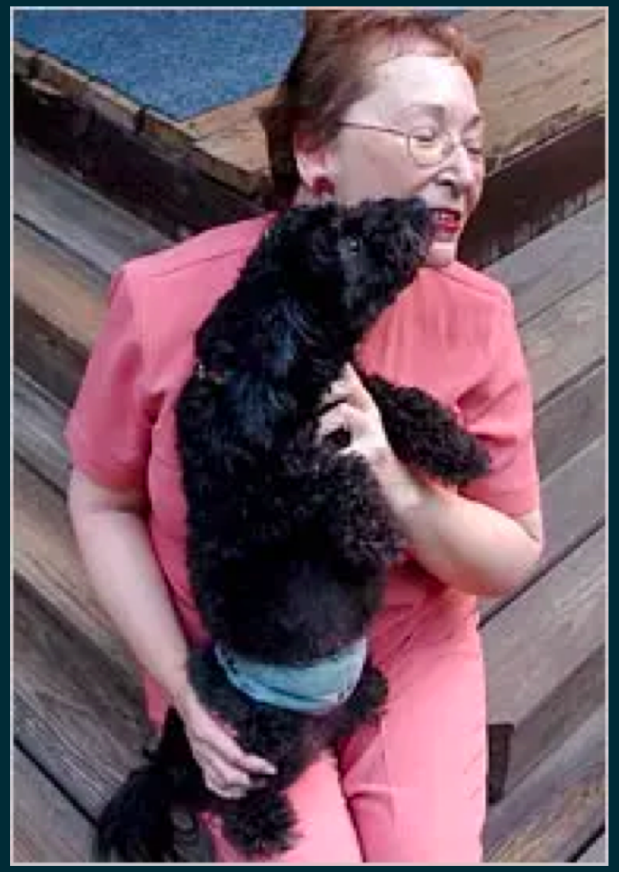
Getting Smarter about Visual Information

by Alberto Cairo



Calling Bullshit
The Art of Skepticism in a Data-Driven World

by Carl Bergstrom and Jevin West



The New York Times

A Face Is Exposed for AOL Searcher No. 4417749

Ms. [Thelma] Arnold, who agreed to discuss her searches with a reporter, said she was shocked to hear that AOL had saved and published three months' worth of them. "My goodness, it's my whole personal life," she said. "I had no idea somebody was looking over my shoulder."

In the privacy of her four-bedroom home, Ms. Arnold searched for the answers to scores of life's questions, big and small. How could she buy "school supplies for Iraq children"? What is the "safest place to live"? What is "the best season to visit Italy"?

Michael Barbaro and Tom Zeller Jr. A Face Is Exposed for AOL Searcher No. 4417749. New York Times. 9 August 2006.

Case study: OK Cupid

OK Cupid data breach

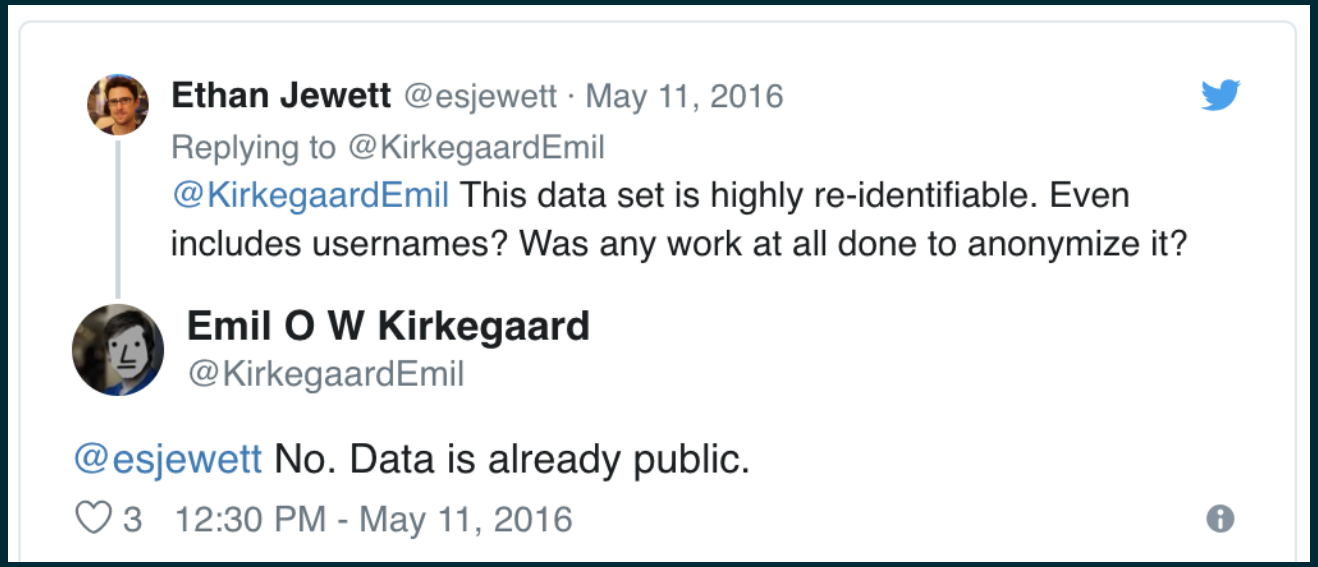
- In 2016, researchers published data of 70,000 OkCupid users—including usernames, political leanings, drug usage, and intimate sexual details
- Researchers didn't release the real names and pictures of OKCupid users, but their identities could easily be uncovered from the details provided, e.g. usernames

OK Cupid data breach

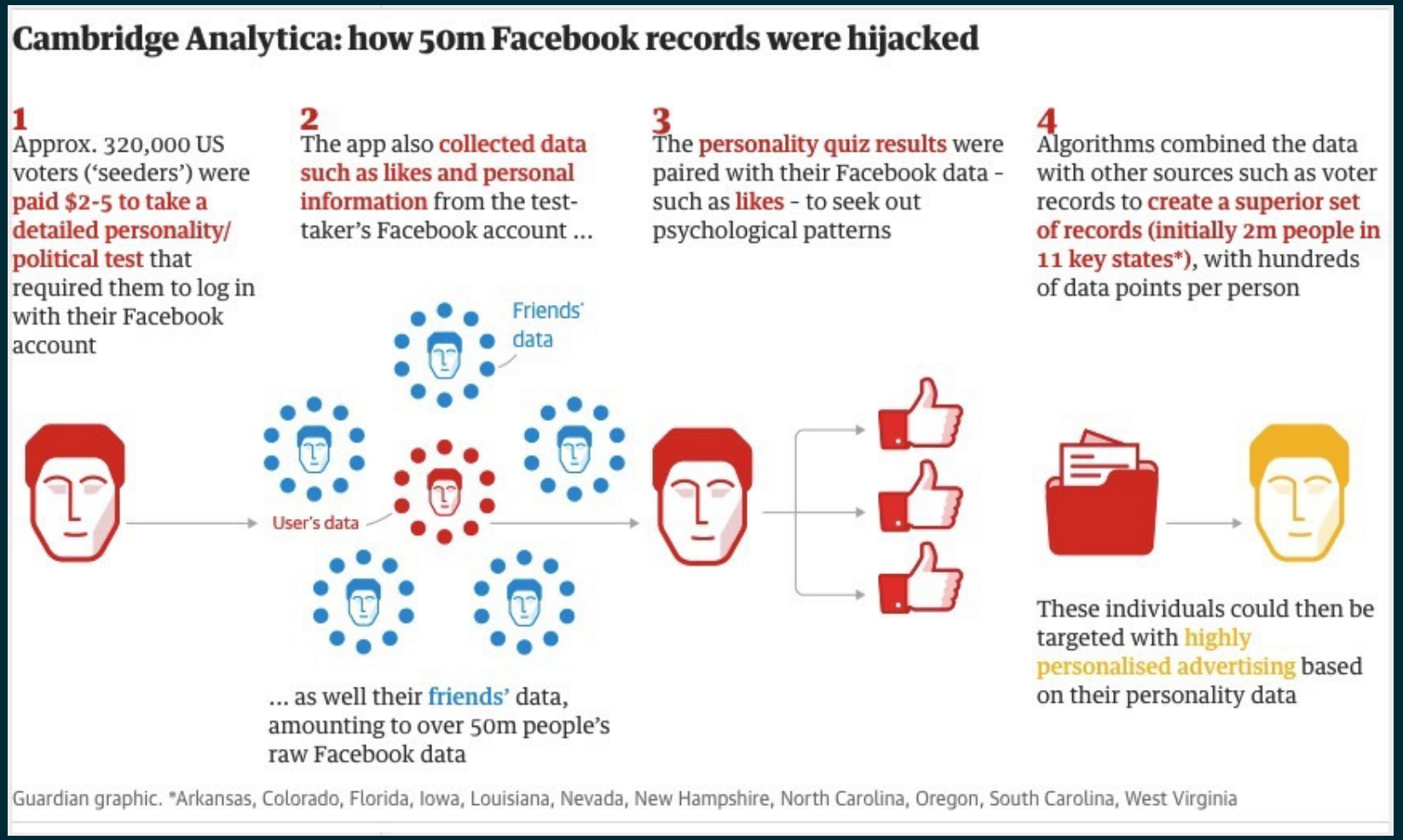
- In 2016, researchers published data of 70,000 OkCupid users—including usernames, political leanings, drug usage, and intimate sexual details
- Researchers didn't release the real names and pictures of OKCupid users, but their identities could easily be uncovered from the details provided, e.g. usernames

> Some may object to the ethics of gathering and releasing this data. However, all the data found in the dataset are or were already publicly available, so releasing this dataset merely presents it in a more useful form.
>
> Researchers Emil Kirkegaard and Julius Daugbjerg Bjerrekær

In analysis of data that individuals willingly shared publicly on a given platform (e.g. social media), how do you make sure you don't violate reasonable expectations of privacy?



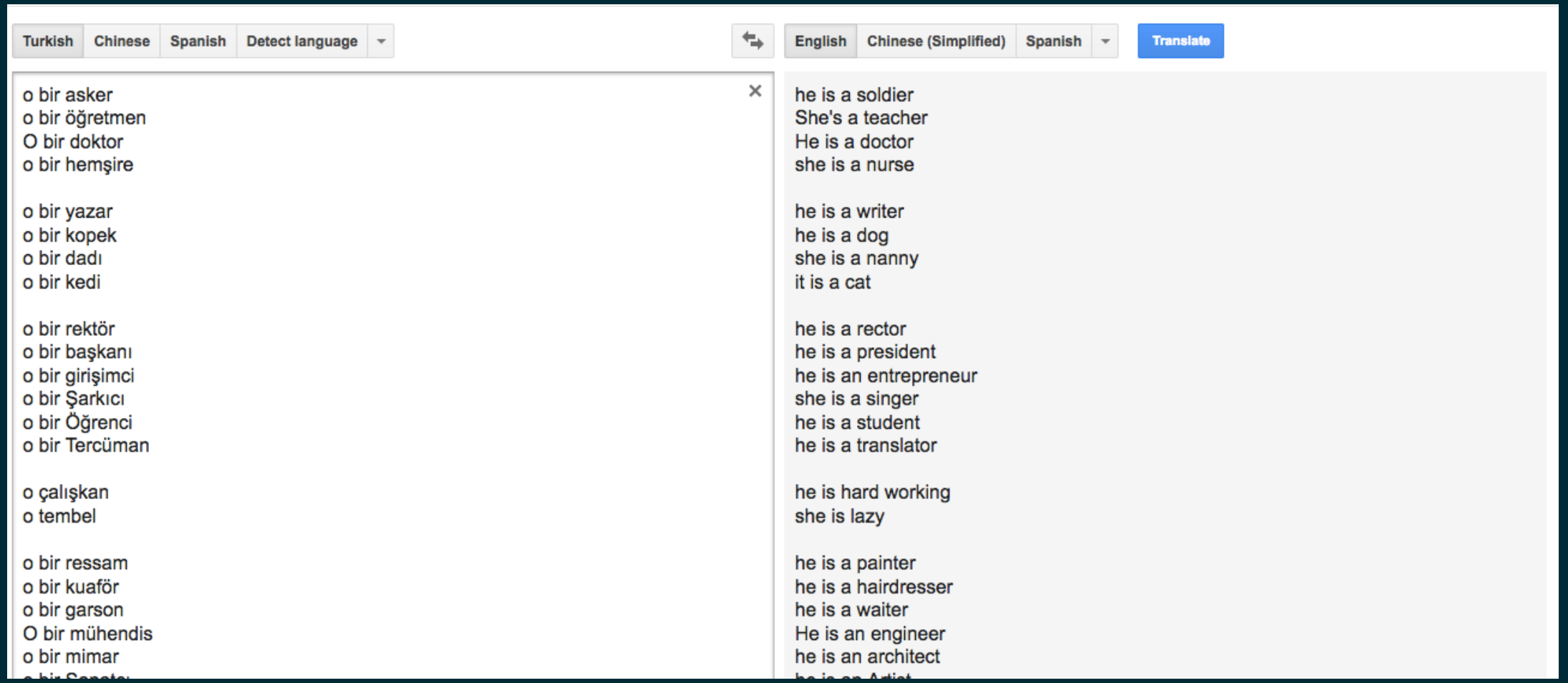
Case study:
Facebook & Cambridge Analytica



Carole Cadwalladr and Emma Graham-Harrison. How Cambridge Analytica turned Facebook 'likes' into a lucrative political tool. The Guardian. 17 March 2018.

Algorithmic bias and gender

Google Translate



Amazon's experimental hiring algorithm

- Used AI to give job candidates scores ranging from one to five stars -- much like shoppers rate products on Amazon
- Amazon's system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way; it taught itself that male candidates were preferable

> Gender bias was not the only issue. Problems with the data that underpinned the models' judgments meant that unqualified candidates were often recommended for all manner of jobs, the people said.

Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. 10 Oct 2018.

Algorithmic bias and race

Facial recognition

Interview

'A white mask worked better': why algorithms are not colour blind

Ian Tucker

When Joy Buolamwini found that a robot recognised her face better when she wore a white mask, she knew a problem needed fixing

Sun 28 May 2017 13.27 BST

Joy Buolamwini is a graduate researcher at the MIT Media Lab and founder of the Algorithmic Justice League - an organisation that aims to challenge the biases in decision-making software. She grew up in Mississippi, gained a Rhodes scholarship, and she is also a Fulbright fellow, an Astronaut scholar and a Google Anita Borg scholar. Earlier this year she won a $50,000 scholarship funded by the makers of the film *Hidden Figures* for her work fighting coded discrimination.

Ian Tucker. 'A white mask worked better': why algorithms are not colour blind. The Guardian. 28 May 2017.

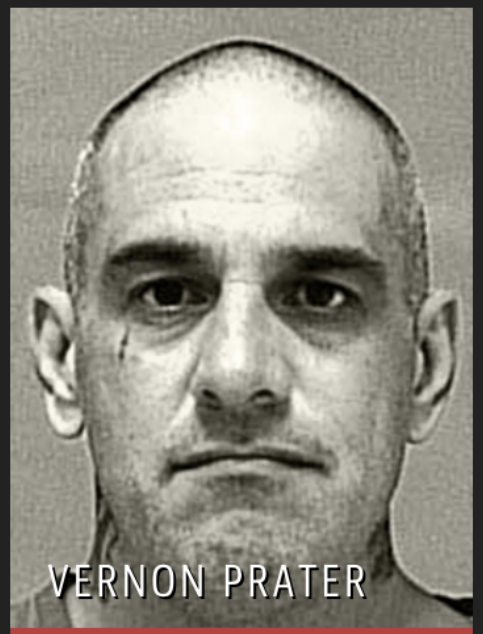
Criminal Sentencing

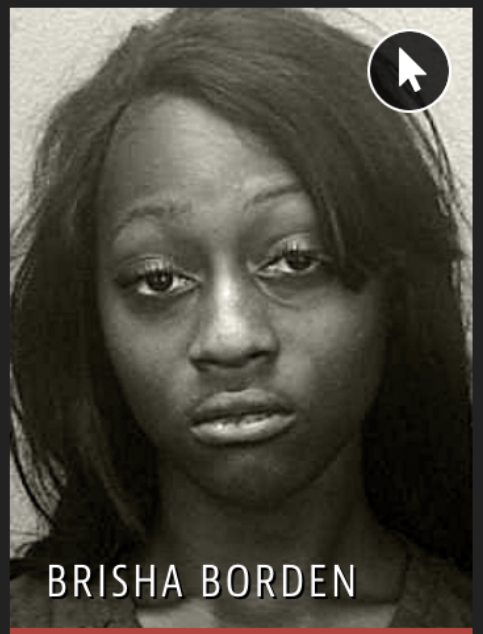
There's software used across the country to predict future criminals.
And it's biased against blacks.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. 23 May 2016. ProPublica.

A tale of two convicts

Two Petty Theft Arrests

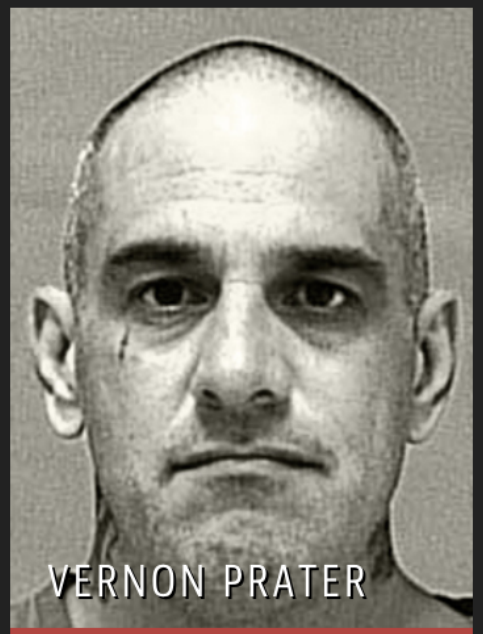




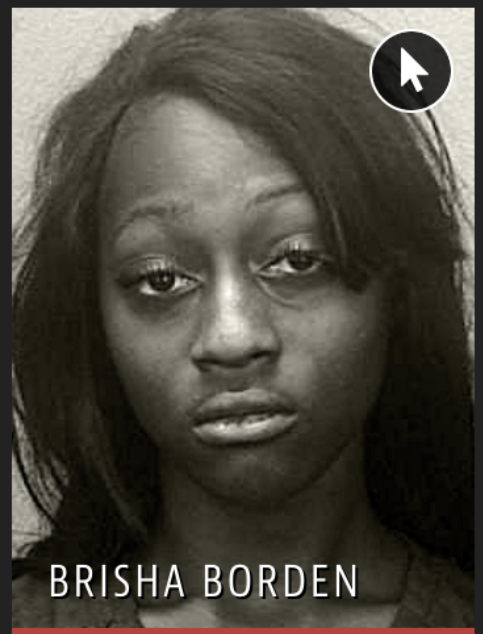
Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

A tale of two convicts

Two Petty Theft Arrests



LOW RISK 3



HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two Petty Theft Arrests

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

> “Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice,” he said, adding, “they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.”
>
> Then U.S. Attorney General Eric Holder (2014)

ProPublica analysis

Data:

Risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 + whether they were charged with new crimes over the next two years

ProPublica analysis

Results:

- 20% of those predicted to commit violent crimes actually did
- Algorithm had higher accuracy (61%) when full range of crimes taken into account (e.g. misdemeanors)

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

- Algorithm was more likely to falsely flag black defendants as future criminals, at almost twice the rate as white defendants
- White defendants were mislabeled as low risk more often than black defendants

How to write a racist AI without trying

How to write a racist AI in R without really trying

5 min read

2018/09/27

Last year, Robyn Speer wrote a really great post How to make a racist AI without really trying. Go read it.

The idea is to do sentiment analysis with obvious, off-the-shelf tools. As the post says

> So that's what we're going to do here, following the path of least resistance at every step, obtaining a classifier that should look

Thomas Lumley. How to write a racist AI in R without really trying.
Biased and Inefficient. 27 September 2018.

Further reading

Machine Bias

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

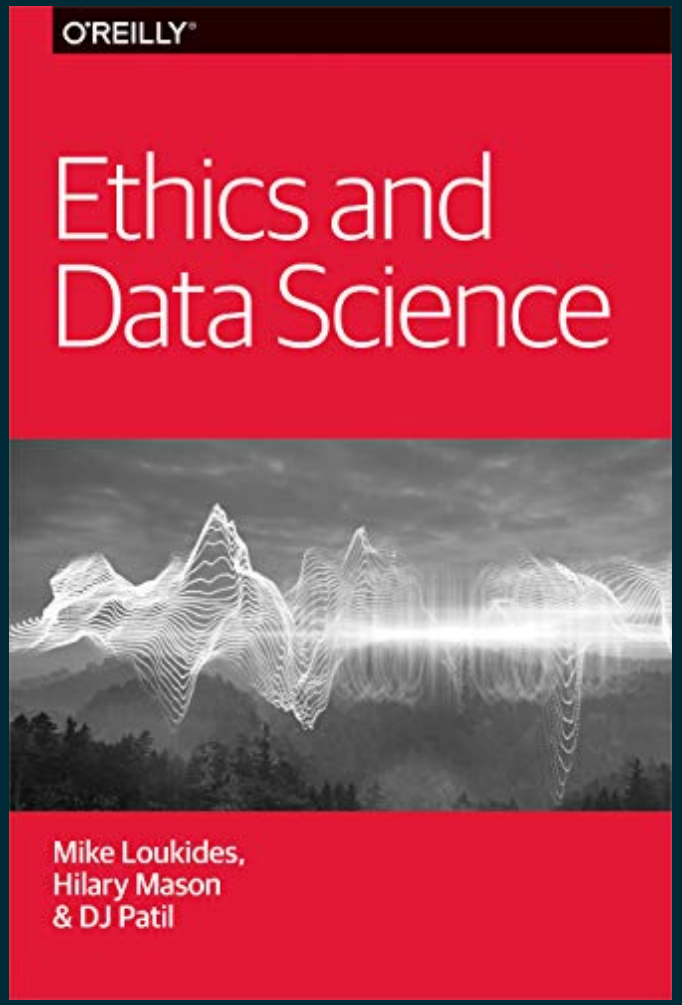
Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Machine Bias

by Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner

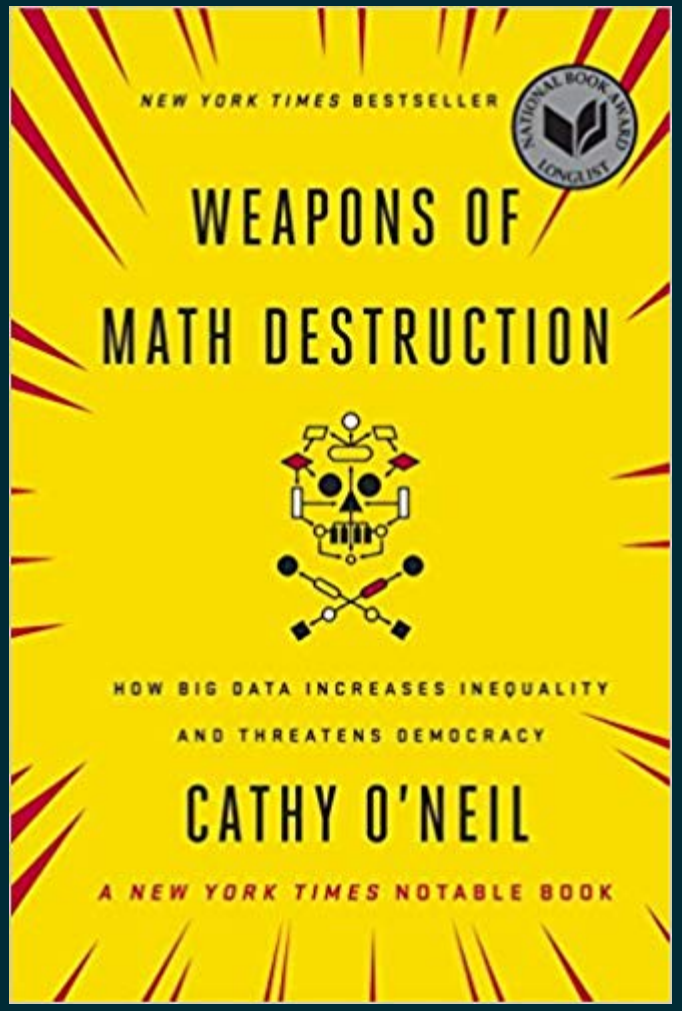
Ethics and Data Science



Ethics and Data Science

by Mike Loukides, Hilary Mason, DJ Patil
(Free Kindle download)

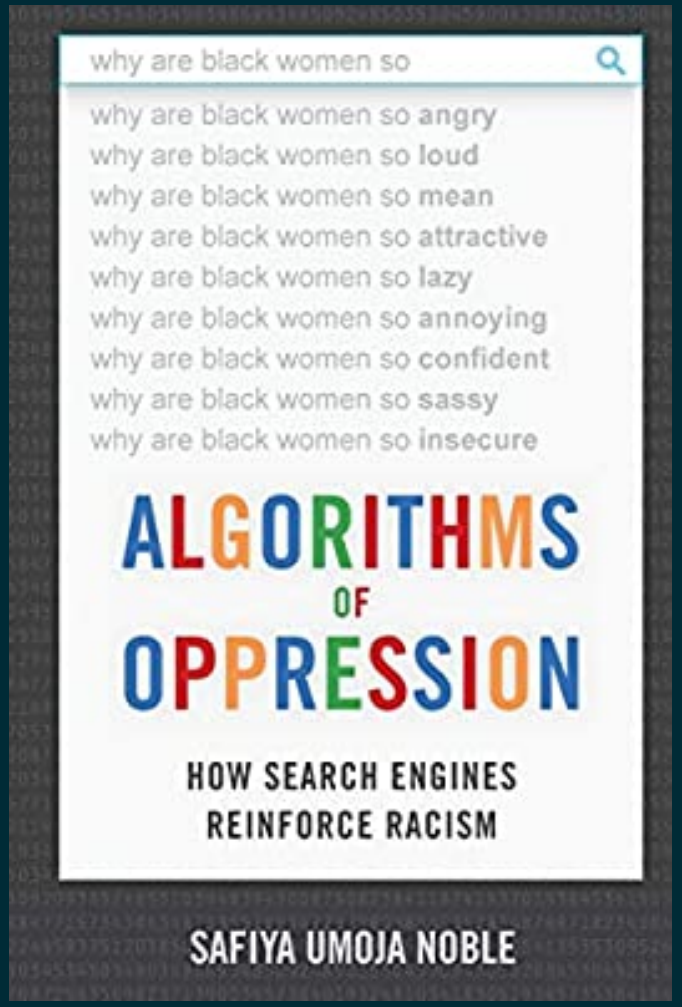
Weapons of Math Destruction



Weapons of Math Destruction
How Big Data Increases Inequality and Threatens Democracy

by Cathy O'Neil

Algorithms of Oppression



Algorithms of Oppression

How Search Engines Reinforce Racism

by Safiya Umoja Noble

Parting thoughts

- At some point during your data science learning journey you will learn tools that can be used unethically
- You might also be tempted to use your knowledge in a way that is ethically questionable either because of business goals or for the pursuit of further knowledge (or because your boss told you to do so)

> How do you train yourself to make the right decisions (or reduce the likelihood of accidentally making the wrong decisions) at those points?

Do good with data

- Data Science for Social Good:
  - The Alan Turing Institute
  - University of Chicago
- DataKind: DataKind brings high-impact organizations together with leading data scientists to use data science in the service of humanity.
- Sign the Manifesto for Data Practices: datapractices.org/manifesto

Further watching

Julien Cornebise. AI for Good in the R and Python ecosystems. useR 2019.