40.016: The Analytics Edge

Week 9 Lecture 2

Ethics in Data Analytics

Term 5, 2022

Outline

Outline

Needs for Ethics in Data Analytics

- Important aspect of the accreditation process
- The relevance of ethical issues in data analytics increases with the amount of data collected, stored, traded, and processed by both public and private sectors

$\rightarrow$ The scale and ease with which analytics can be conducted today completely changes the ethical framework

The Facebook-Cambridge Analytica data scandal

Figure: Source: `https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump`

The Facebook-Cambridge Analytica data scandal (cont'd)

What did Cambridge Analytica do?

- In 2015, Cambridge Analytica harvested information from ~ 87 million Facebook users
- Data were sourced through an app / personal quiz, "this is your digital life"
- The data were then used in political campaigns with a technique known as micro-targeting

The Facebook-Cambridge Analytica data scandal (cont'd)

What sort of data did Cambridge Analytica harvest?

- Personal information on where users lived, what pages they liked, etc.

How were these data useful?

- Data were used to build psychological profiles, with traits like openness, agreeableness, IQ, gender, age, political views ...
- Psychological profiles are a fundamental piece of information for making micro-targeting successful

The Facebook-Cambridge Analytica data scandal (cont'd)

Responses:

- Facebook CEO Mark Zuckerberg first apologized and then led the implementation of policies on data protection
- Governments worldwide took initiatives to understand (1) the role played by Facebook and Cambridge Analytica, and (2) the extent of the "data breach"
- In July 2019, the US Federal Trade Commission voted to approve fining Facebook around $5 billion USD to finally settle the investigation

The Facebook-Cambridge Analytica data scandal (cont'd)

(Some) Ethical issues:

- Manipulation of users
- Privacy
- Transparency

(Other issues, not discussed here, pertain to the legal aspects of the scandal.)

Ethical issues in data science

- Bias, discrimination, and exclusion: Algorithms and artificial intelligence can create biases, discrimination or even exclusion towards individuals and groups of people.
  - not necessarily by design
  - unintended consequence
- Algorithmic profiling: Personalizing versus collective benefits: Individuals have gained a great deal from profiling. This mindset of personalising can affect the key collective principles like democratic and cultural pluralism and risk-sharing in the realm of insurance.

Ethical issues in data science (cont'd)

- Preventing massive files while enhancing AI: Data protection laws are rooted in the belief that individuals rights regarding their personal data must be protected and thus prevent the creation of massive files.
- Quality, quantity, relevance: The acceptance of the existence of potential bias in datasets curated to train algorithms is of paramount importance.

(Some) Governing ethical principles

- Ownership (Who owns the data?)
- Transaction transparency
- Consent
- Privacy
- Openness

(Some) Governing ethical principles (cont'd)

- Fairness
- Justice
- Beneficence
- Non-maleficence

Guidelines

- American Statistical Association (Ethical Guidelines were updated in 2018)
- Association of Computing Machinery (“The Code” was updated in 2018)
- IEEE Code of Ethics

Relevant legislations

- The General Data Protection Regulation (GDPR) is a law on data protection and privacy for the European Economic Area. It became enforceable on May 2018.
- The GDPR is a model for other national laws adopted across the world.
- In Singapore, the Personal Data Protection Act 2012 sets out the law on data protection. It was amended on November 2, 2020.
- The Personal Data Protection Commission (PDPC) is the main authority in matters relating personal data.

Outline

Using and Abusing Data Visualization

- A misleading (or distorted) graph is a graph that misrepresents data
- It may be created intentionally to misguide the viewer
- A seminal work in this area is *How to Lie with Statistics*, by Darrell Huff (1954)

Example 1

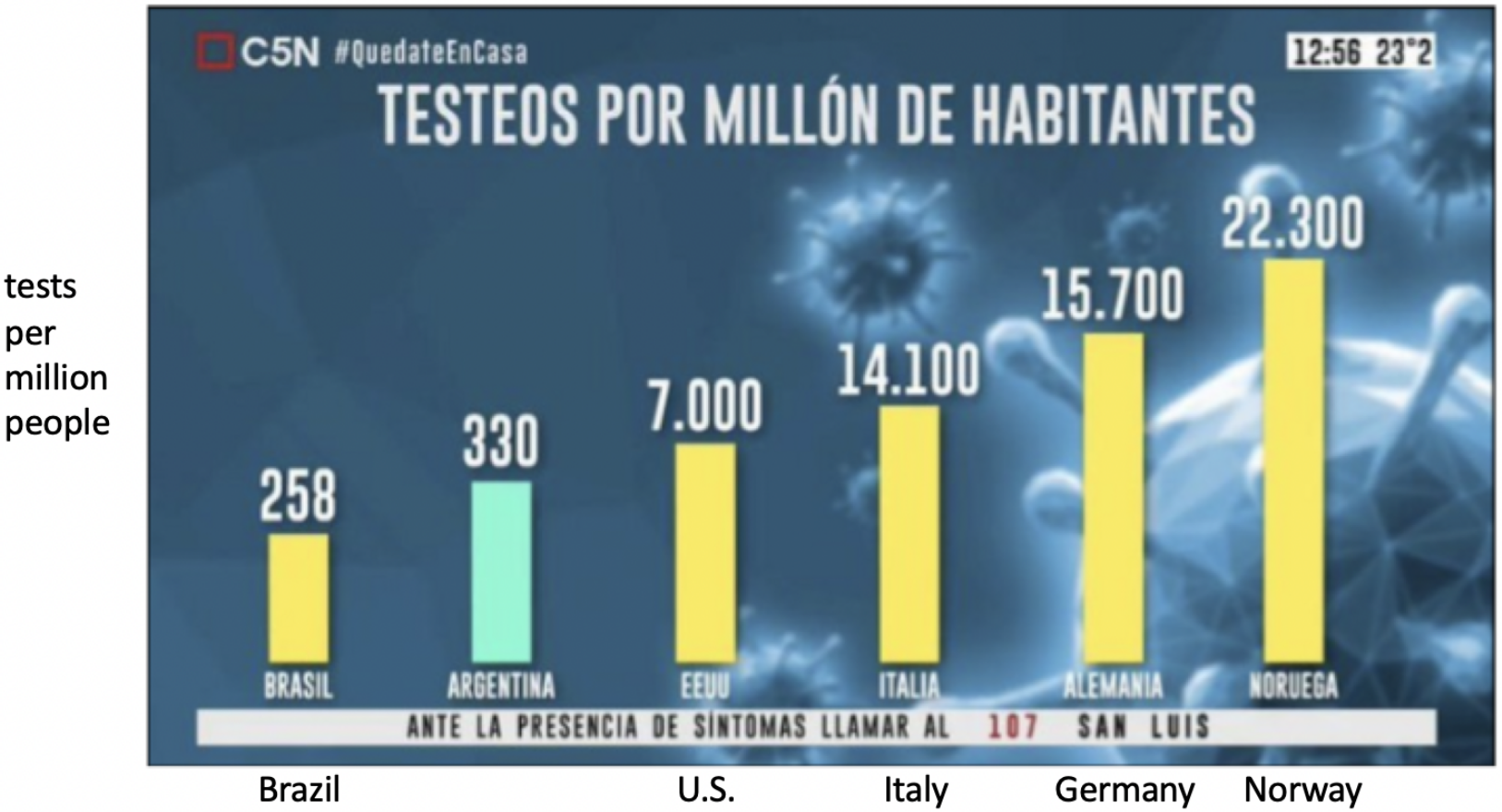


Figure: Argentinas number of COVID-19 tests, original plot. (Source: https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9)

Example 1 (cont'd)

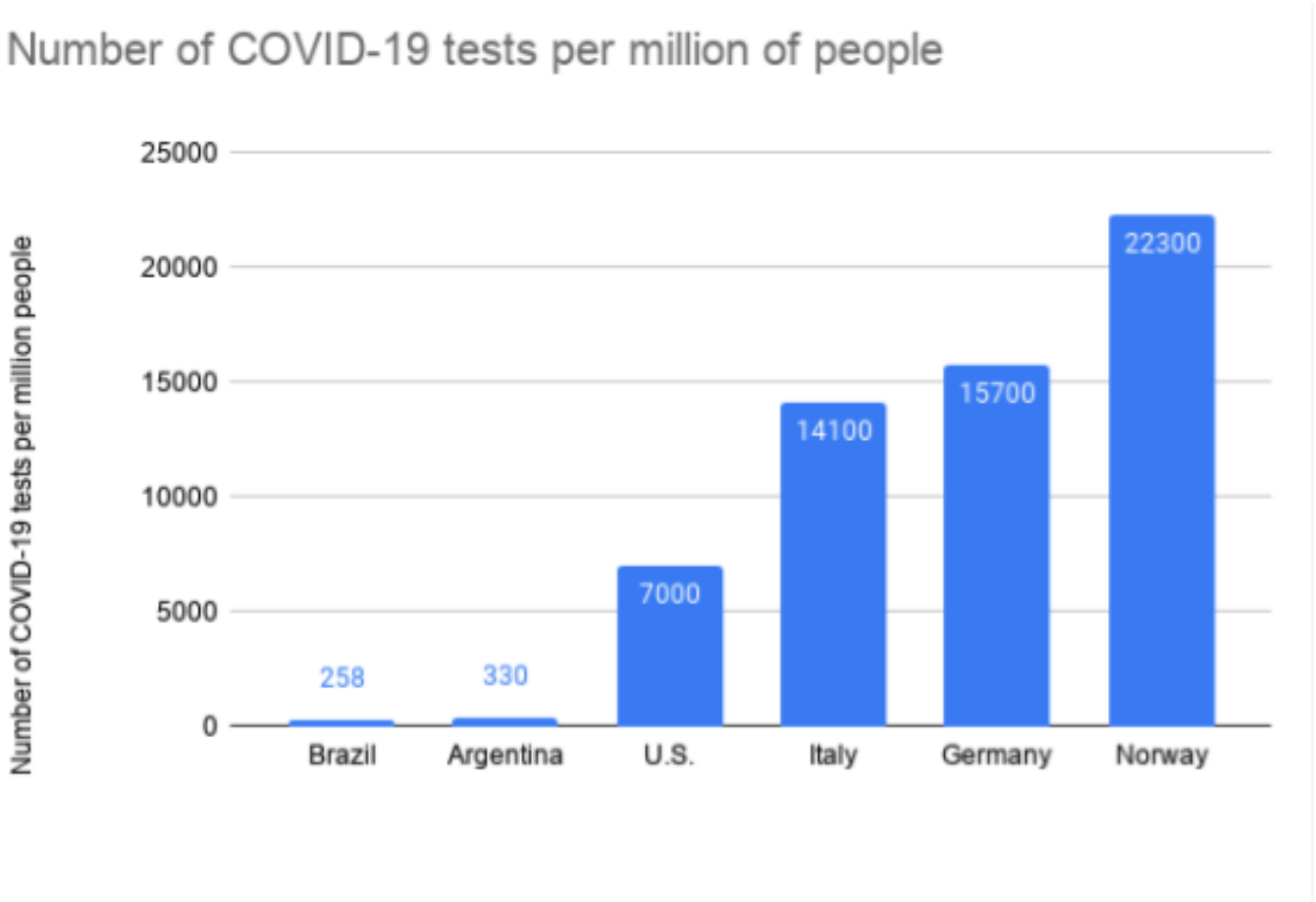
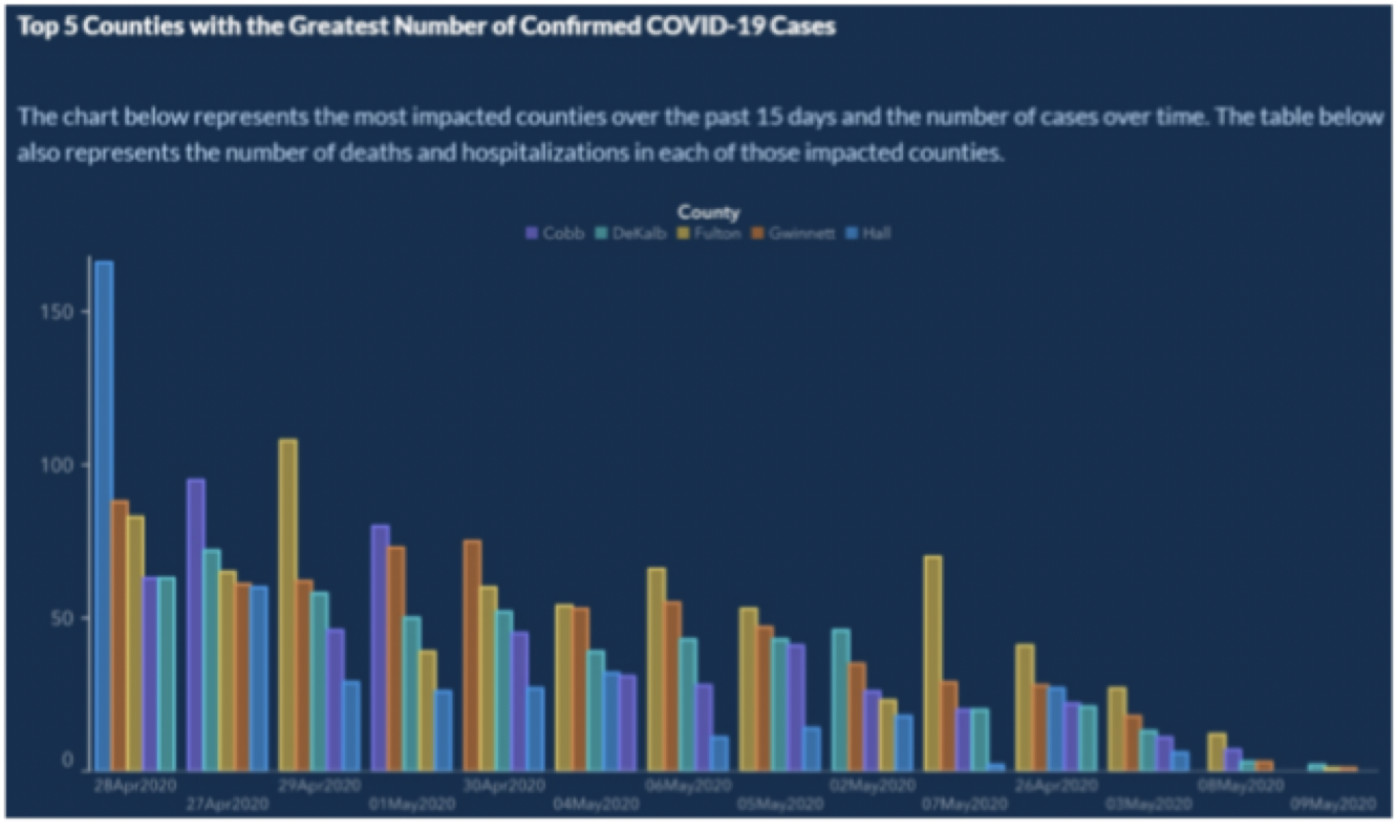


Figure: Argentinas number of COVID-19 tests, modified / correct plot.

(Source: `https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9`)

Example 2

Not in chronological order!



28Apr;27Apr;29Apr;01May;30Apr...

Figure: Number of COVID-19 cases in five counties in Georgia. The plot was retrieved from the Georgia Department of Public Health! (Source: `https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9`)

5 ways writers use graphs to mislead you

- Writers use graphs to make their information seem credible
- But graphs should be read with a critical eye
- There are ways that writers will misrepresent and skew data to support their narratives
- Here are 5 of the most common ways writers use graphs to mislead readers

Issue 1: Omitting the baseline

- In most cases, the baseline for a graph is 0.
- But writers can skew how data is perceived by making the baseline a different number.
- This is known as a “truncated graph”.

Issue 1: Omitting the baseline (cont'd)

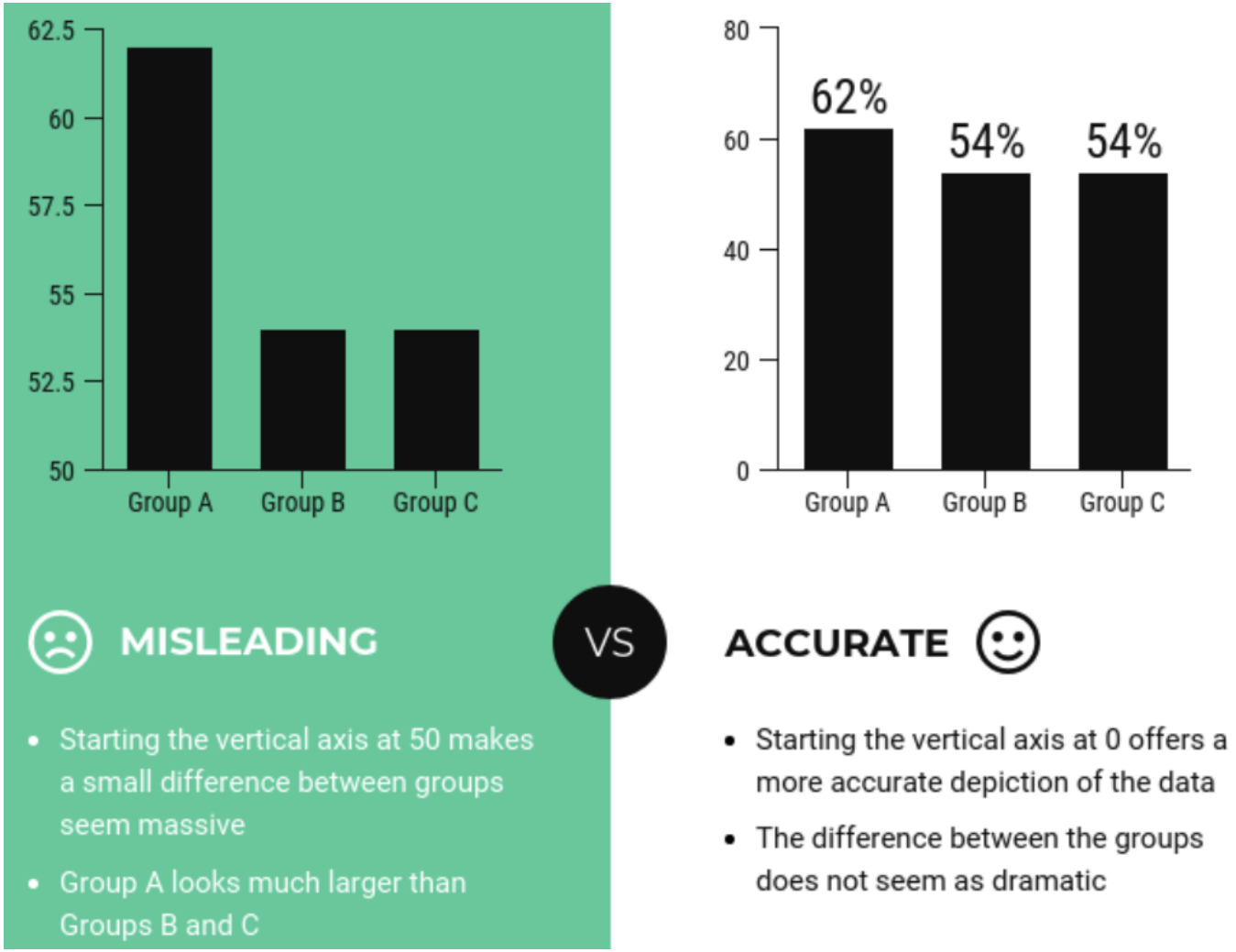


Figure: (Source: `https://venngage.com/blog/misleading-graphs`)

Issue 2: Manipulating the y-axis

- Expanding or compressing the scale on a graph can make changes in data seem more or less significant than they actually are.

Issue 2: Manipulating the y-axis (cont'd)

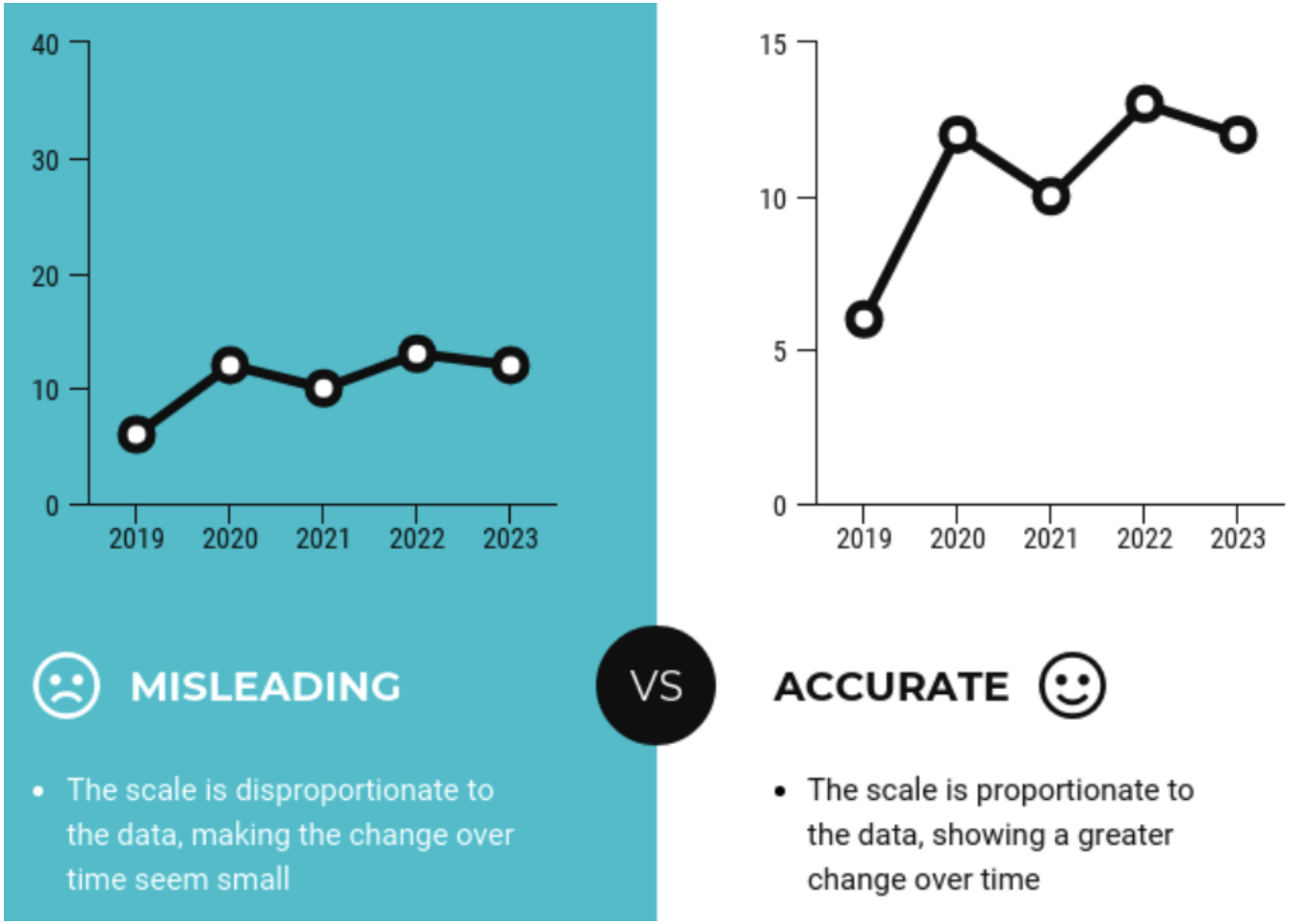


Figure: (Source: `https://venngage.com/blog/misleading-graphs/`)

Issue 2: Manipulating the y-axis (cont'd)

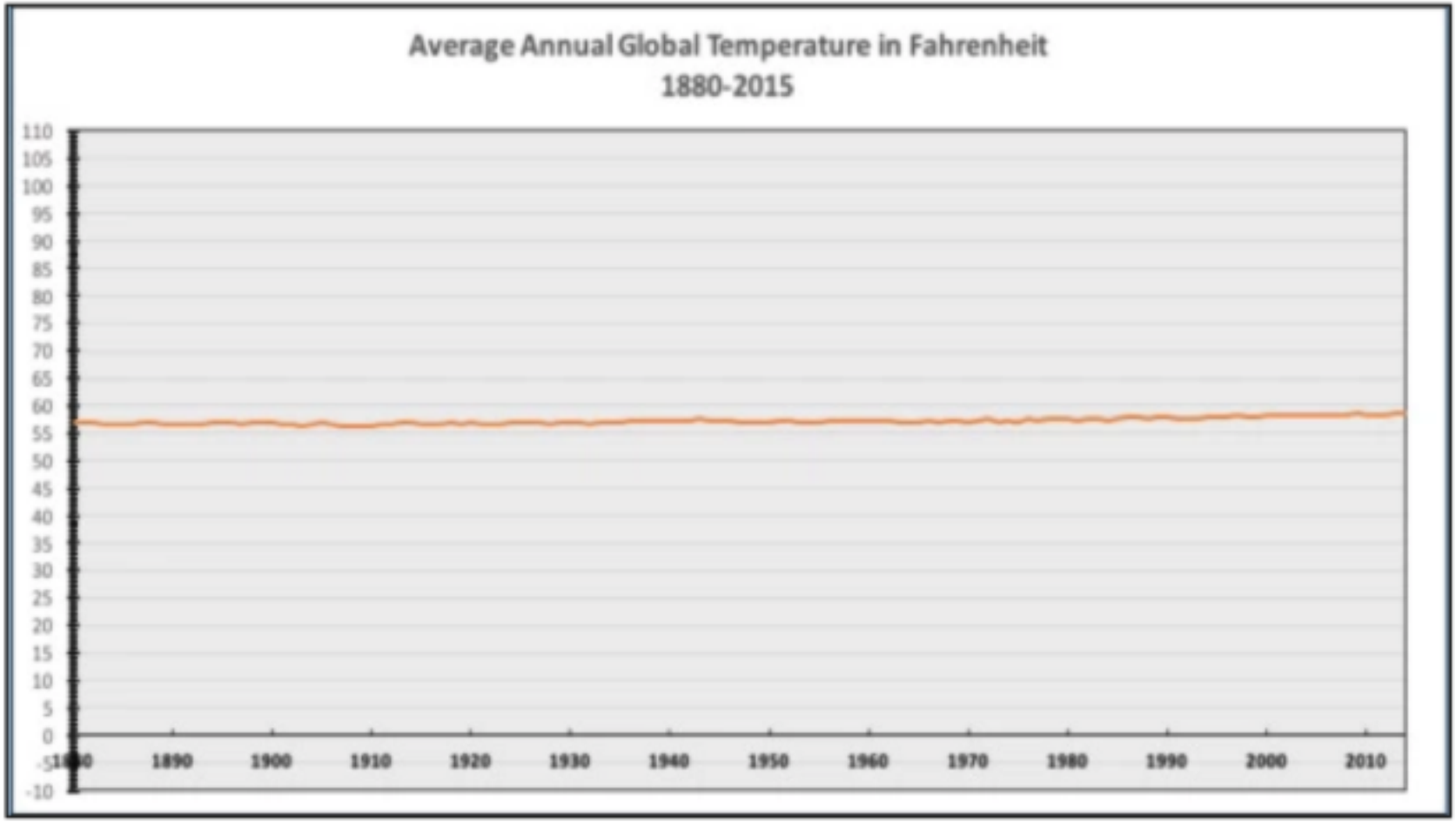


Figure: Original plot. (Source:
`https://venngage.com/blog/misleading-graphs/`)

Issue 2: Manipulating the y-axis (cont'd)

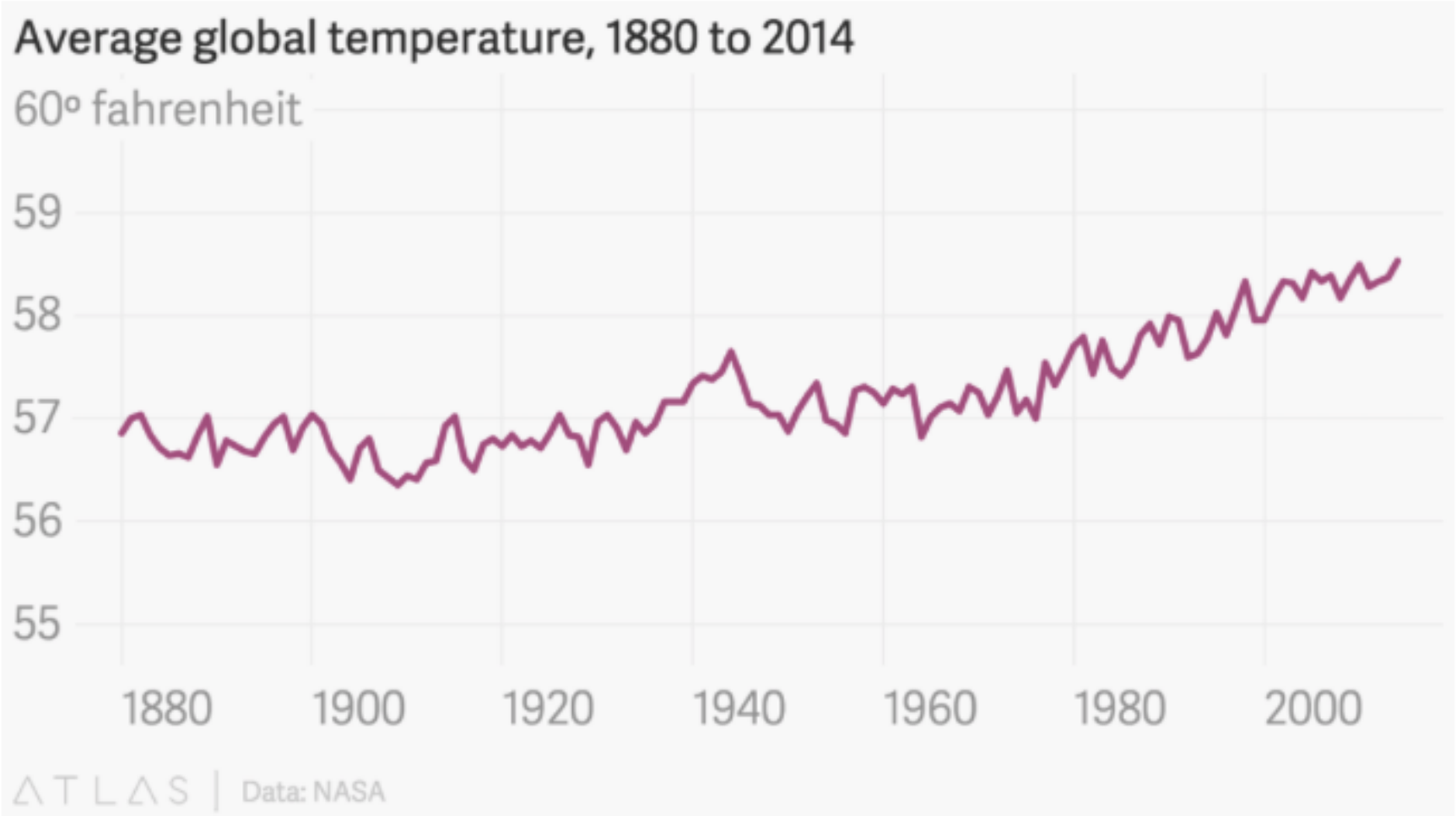


Figure: Modified / Correct plot. (Source:
`https://venngage.com/blog/misleading-graphs/`)

Issue 3: Cherry picking data

- Writers may only include certain data points on their graphs to reinforce their narratives.
- This can create a false impression of the data.

Issue 3: Cherry picking data (cont'd)

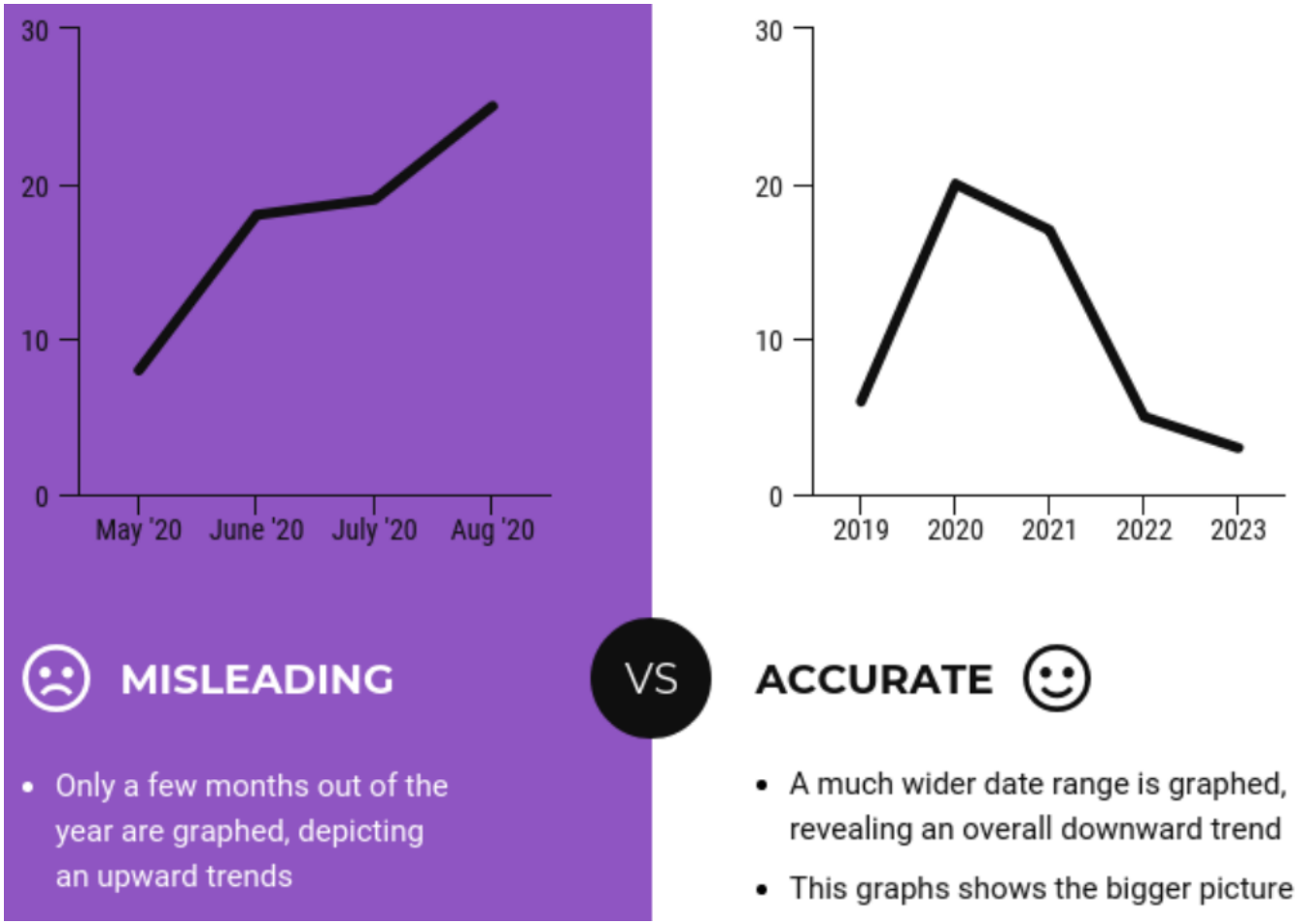


Figure: (Source: `https://venngage.com/blog/misleading-graphs/`)

Issue 3: Cherry picking data (cont'd)

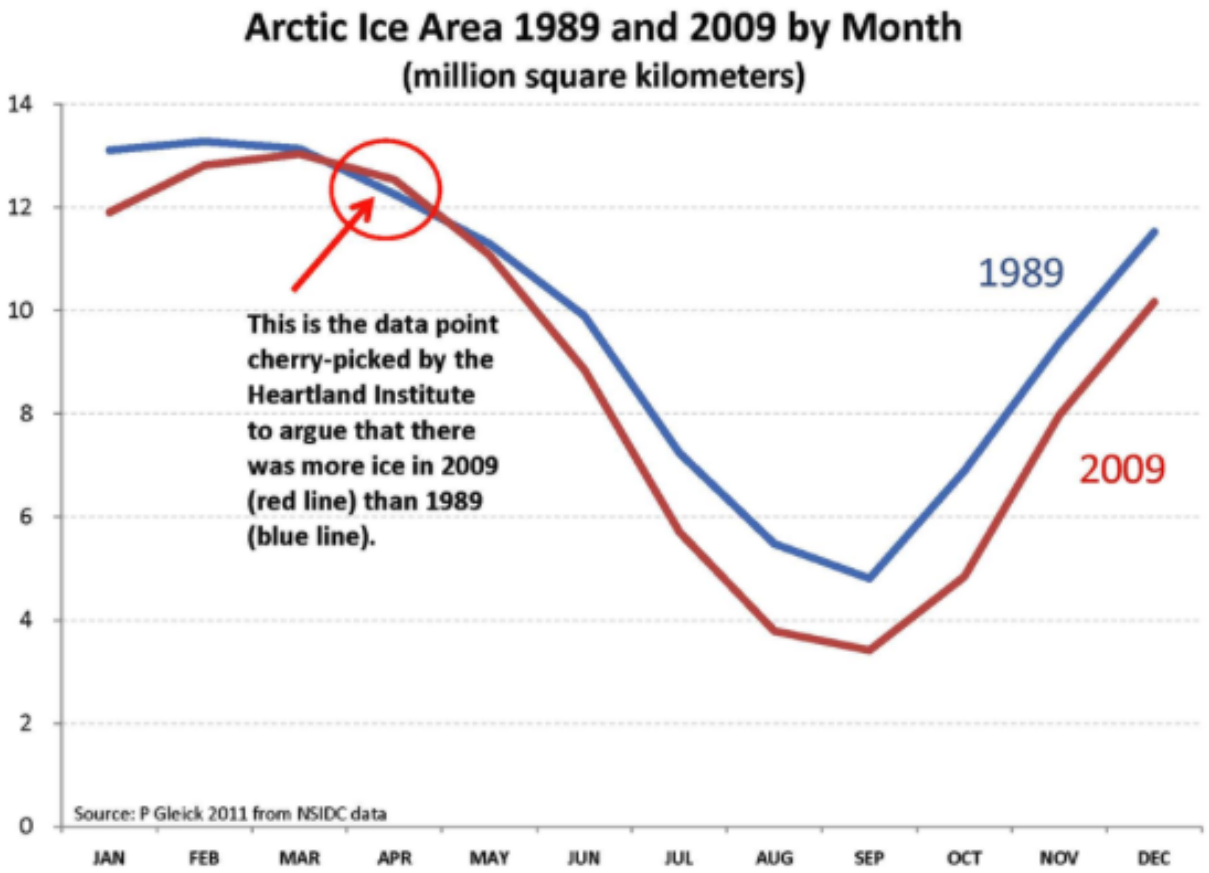


Figure: Original plot. (Source: `https://venngage.com/blog/misleading-graphs/`)

Issue 4: Using the wrong graph

- The type of graph you use should depend on the type of data you want to visualize.
- Using the wrong type of graph can skew the data.
- Writers will sometimes use the wrong type of graph on purpose.

Issue 4: Using the wrong graph (cont'd)

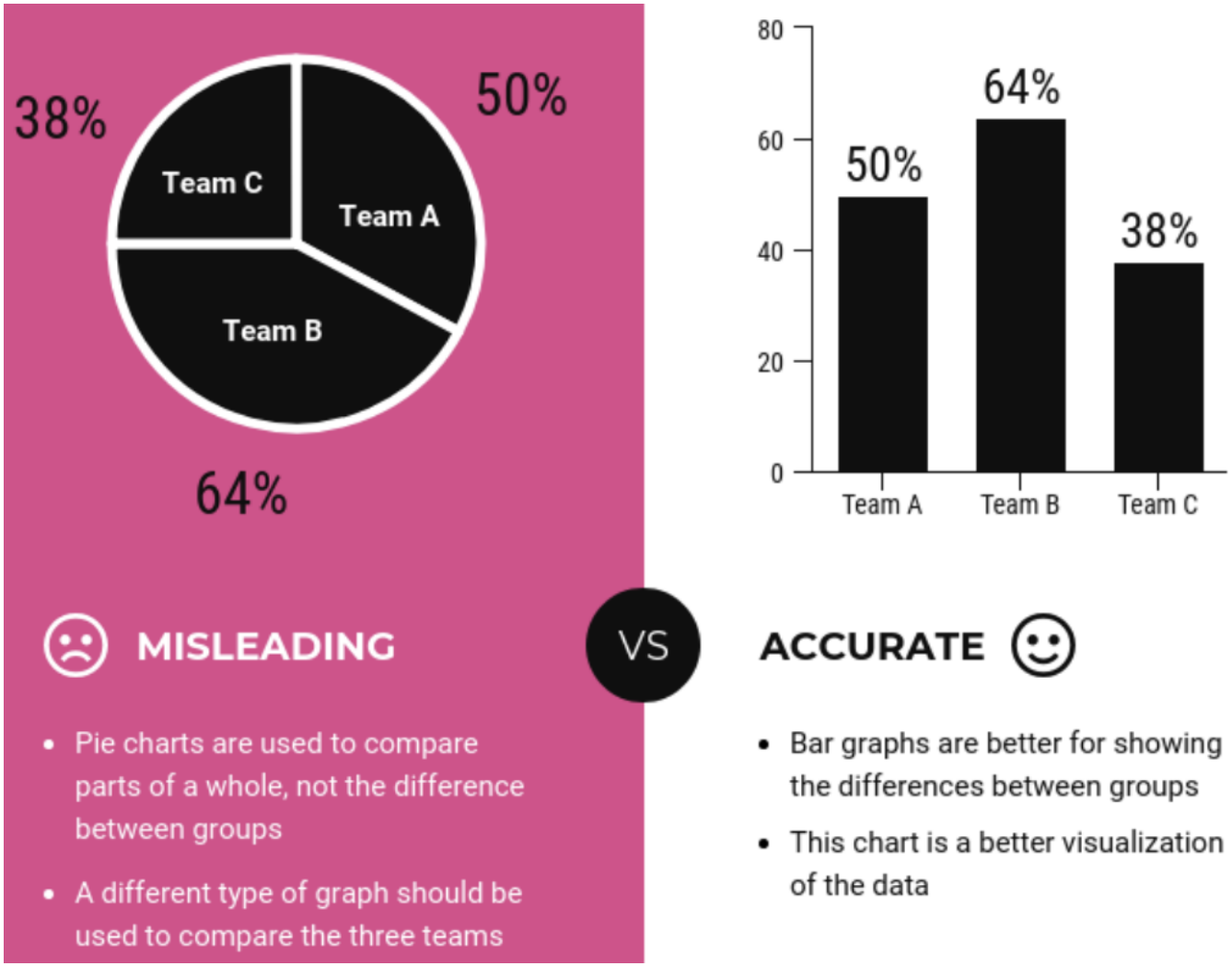


Figure: (Source: `https://venngage.com/blog/misleading-graphs/`)

Issue 4: Using the wrong graph (cont'd)

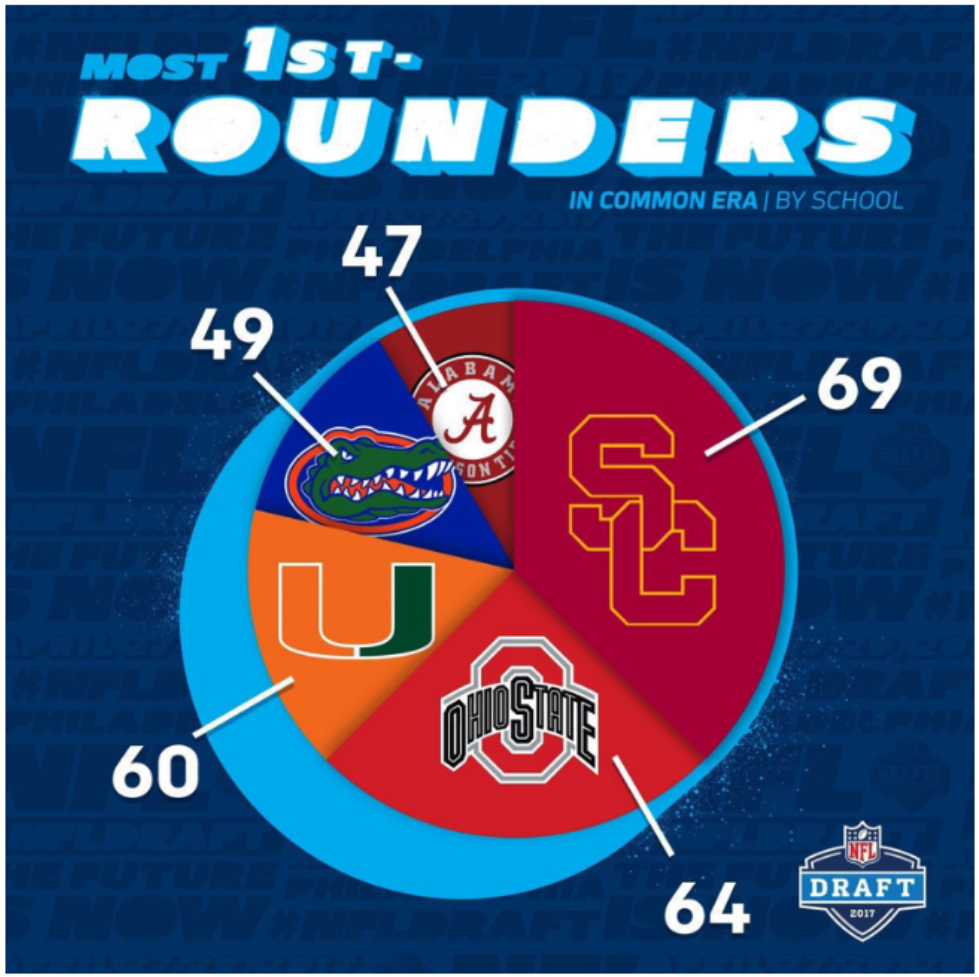


Figure: Original plot. (Source:
`https://venngage.com/blog/misleading-graphs/`)

Issue 4: Using the wrong graph (cont'd)

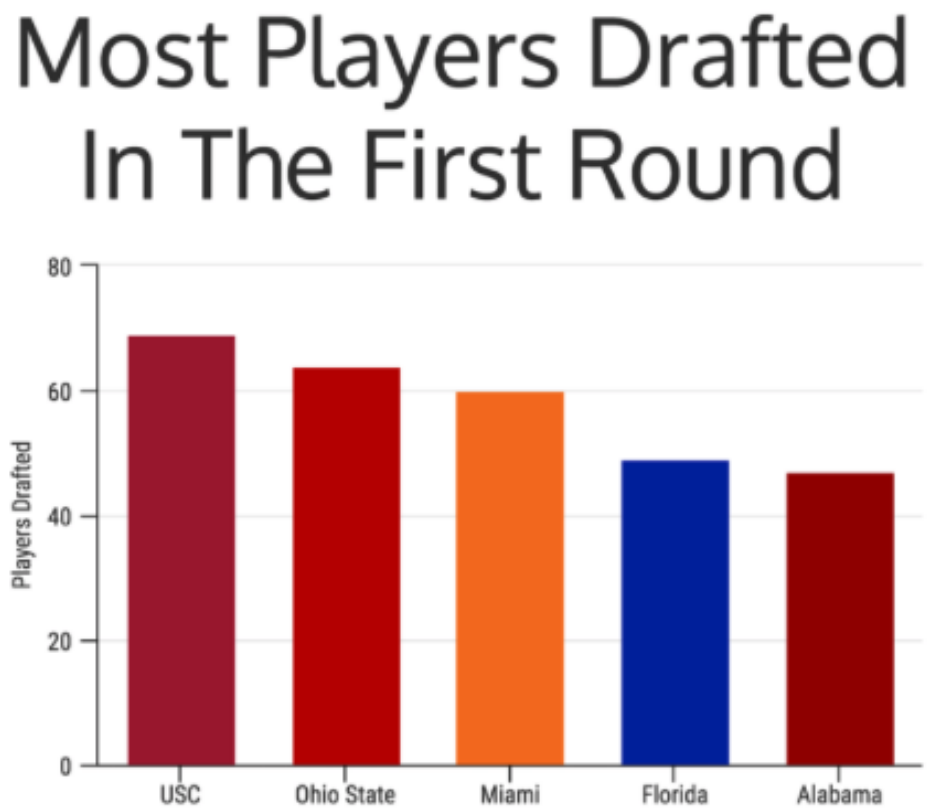


Figure: Modified / Correct plot. (Source: `https://venngage.com/blog/misleading-graphs/`)

Issue 5: Going against conventions (cont'd)

- Over time, we have developed standards for how data is visualized.
- Flipping those conventions can make a graph confusing or misleading to readers.

Issue 5: Going against conventions (cont'd)

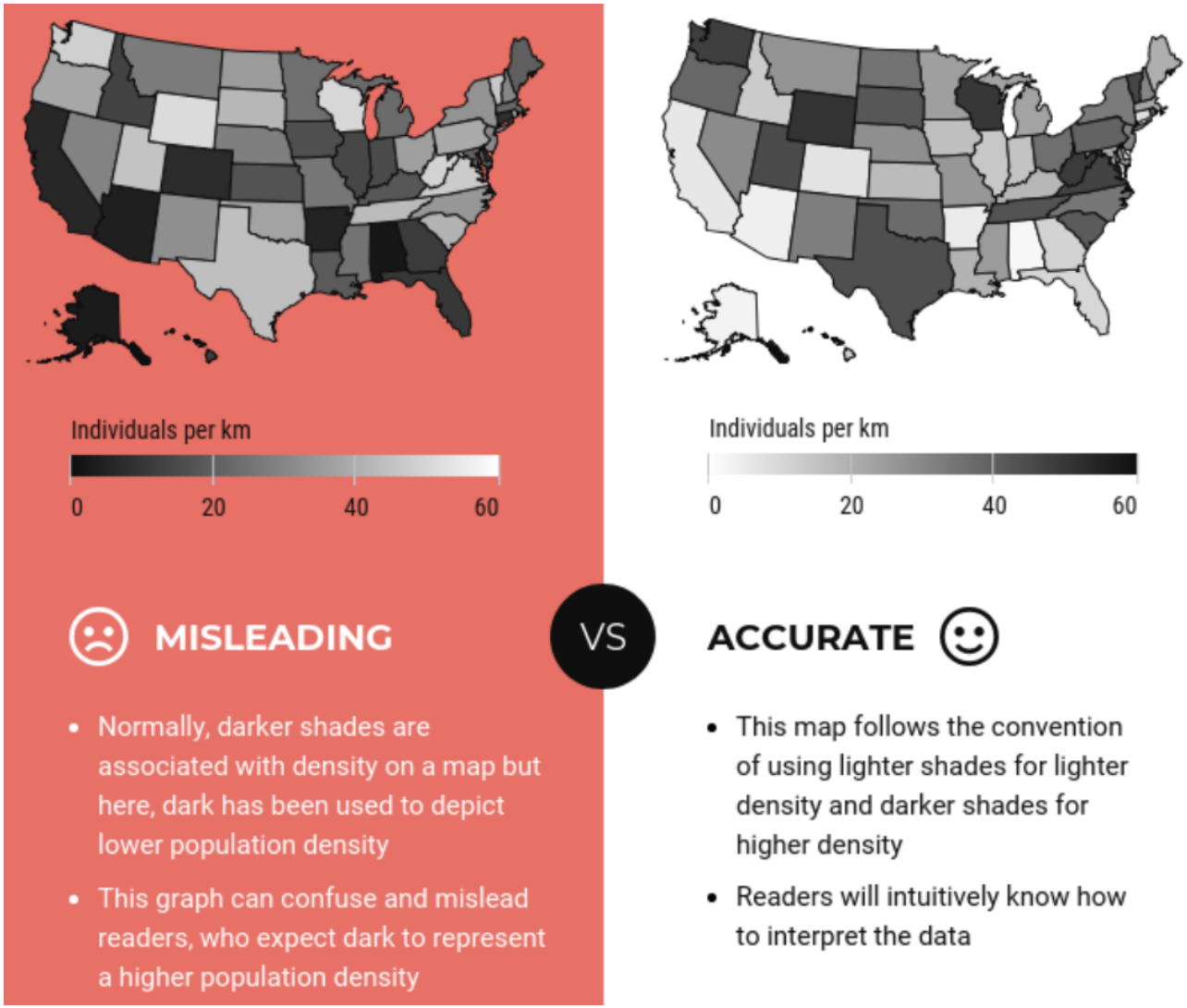


Figure: (Source: `https://venngage.com/blog/misleading-graphs/`)

Other common issues

- Improper intervals or units
- Omitting data
- Extrapolation
- Unnecessary complexity
- ...

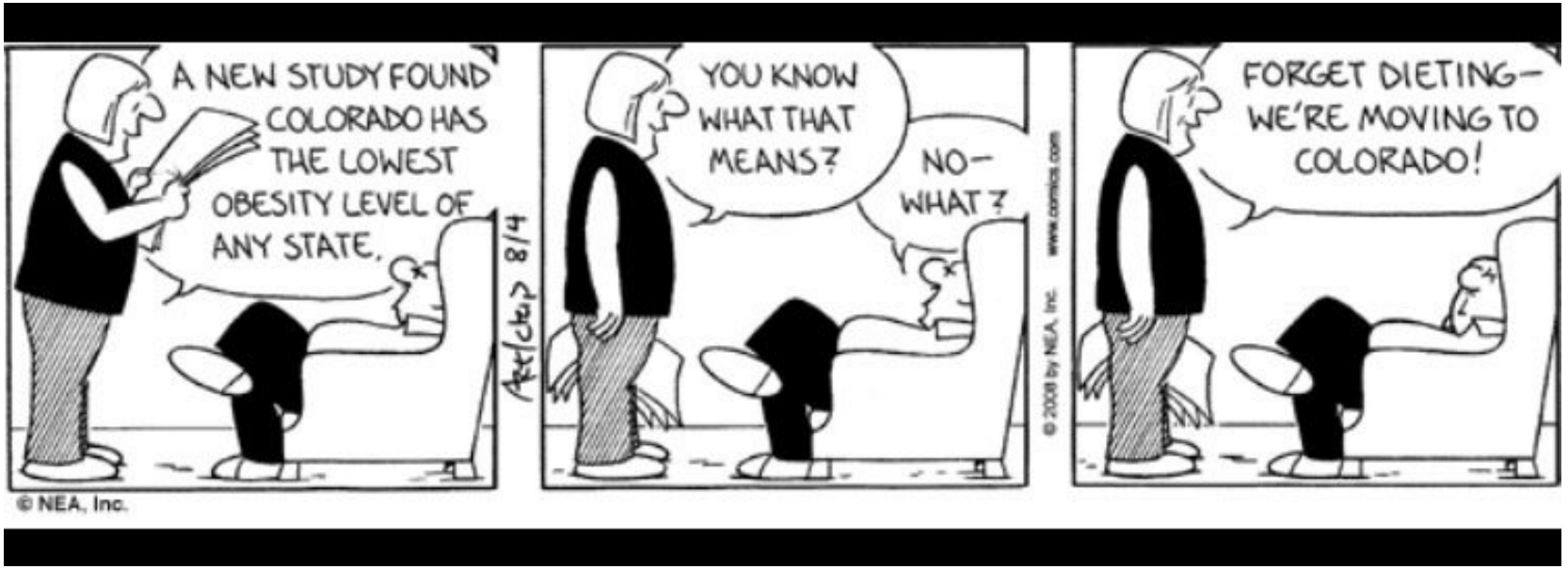
Closure

A hippocratic oath for visualization (by Jason Moore from US Air Force Research Laboratory):

"I shall not use visualization to intentionally hide or confuse the truth which it is intended to portray. I will respect the great power visualization has in garnering wisdom and misleading the uninformed. I accept this responsibility willfully and without reservation, and promise to defend this oath against all enemies, both domestic and foreign."

Outline

Correlation vs. Causation



Correlation vs. Causation

Correlation is a statistical measure of (linear) relationship between variables.

Causation indicates that one event is the consequence of another event.

- Correlation (or a strong mathematical relationship) between two data sets does not necessarily imply causation, even in a perfectly executed study.
- *Spurious correlations* (`https://tylervigen.com`)
- Open `storks.csv` and `CvsC.Rmd` files in R.

Example

Milton Friedman's thermostat (`https://justinhohn.typepad.com/blog/2013/01/milton-friedmans-thermostat-analogy.html`)

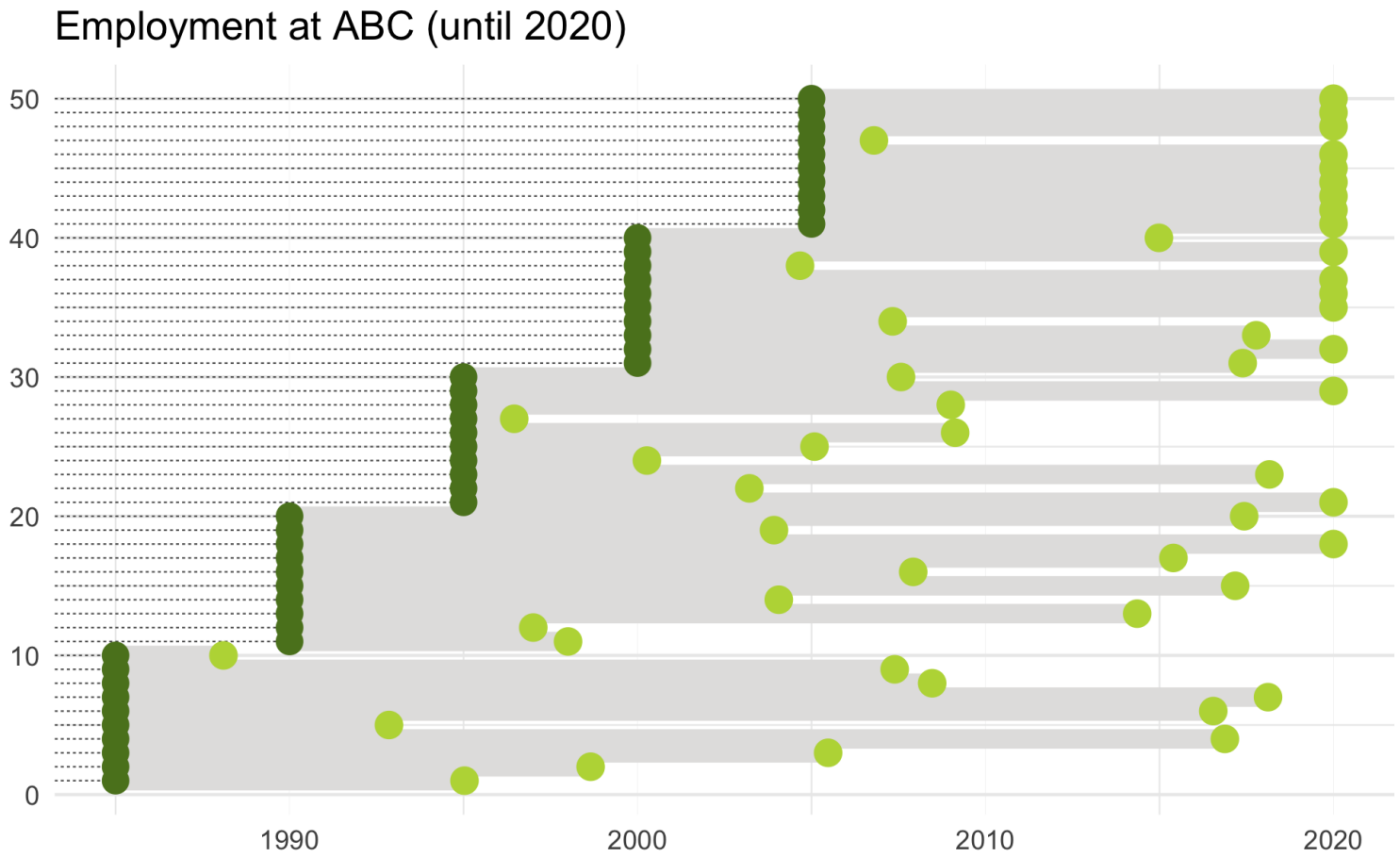
- Imagine a house in Singapore with proper air-conditioning.
  - Outside temperature (O) is positively correlated with energy consumption (E).
  - No correlation between indoor temperature (I), and E.
  - No correlation between I and O.
- Data analysis here is problematic.

Example (cont'd)

- Since you find neither O nor E has an influence on indoor temperature, you decide to switch your A/C off.
  - First you conclude: energy consumption has no effect on indoor temperature.
  - Then you conclude: outside temperature also has no effect indoor temperature.
  - Finally, you switch off the A/C since you found that indoor temperature isn't affected by either.
- In financial economics, good fiscal policies are supposed to maintain a constant inflation rate.

Statistical traps

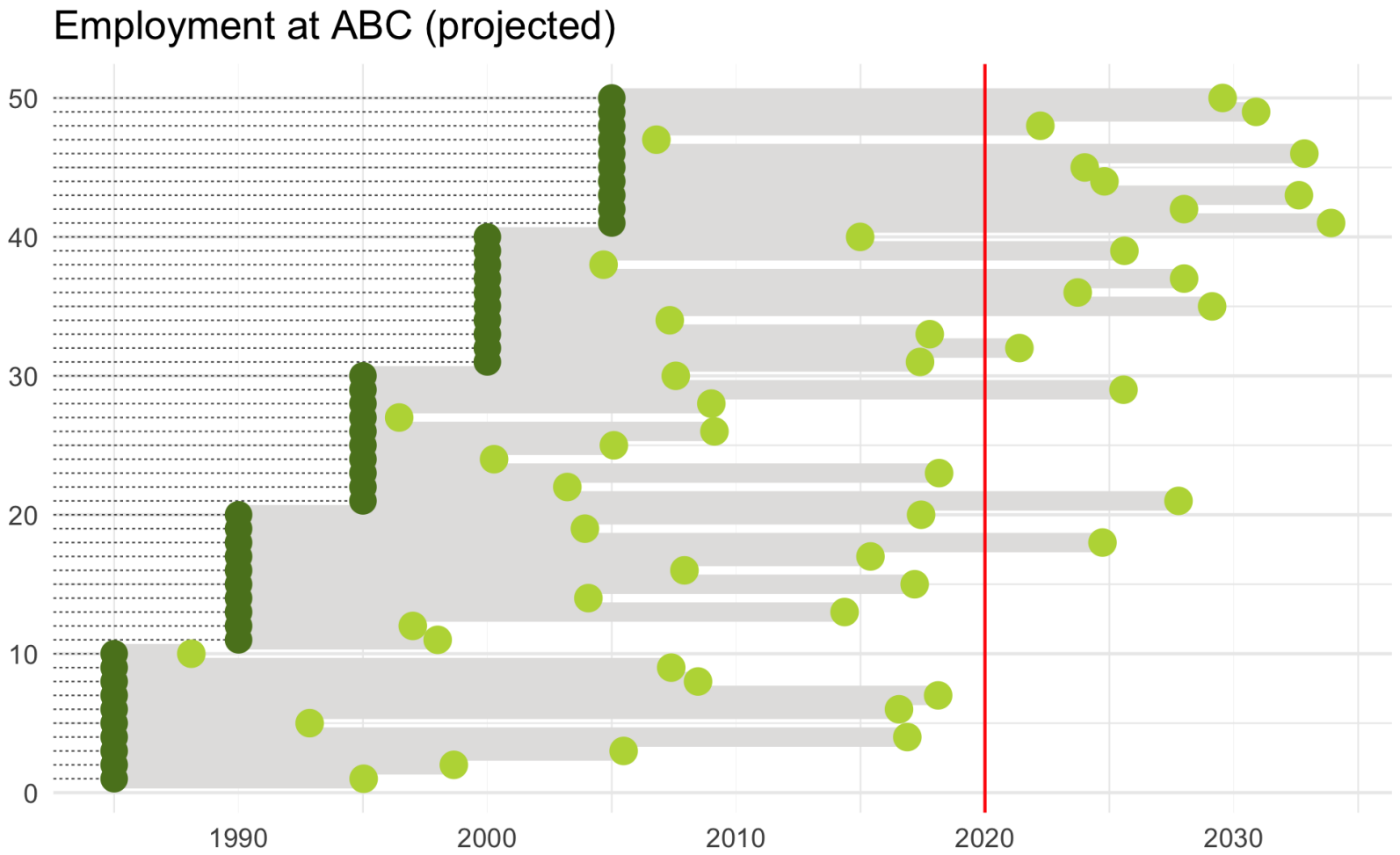
- Are employee tenures shorter than earlier at ABC?
  (This is a simulated dataset.)



Statistical traps (cont'd)

Year	1985	1990	1995	2000	2005
Average	19.69	13.26	12.37	8.65	8.86
Std. Dev.	10.61	7.63	6.34	6.42	3.74

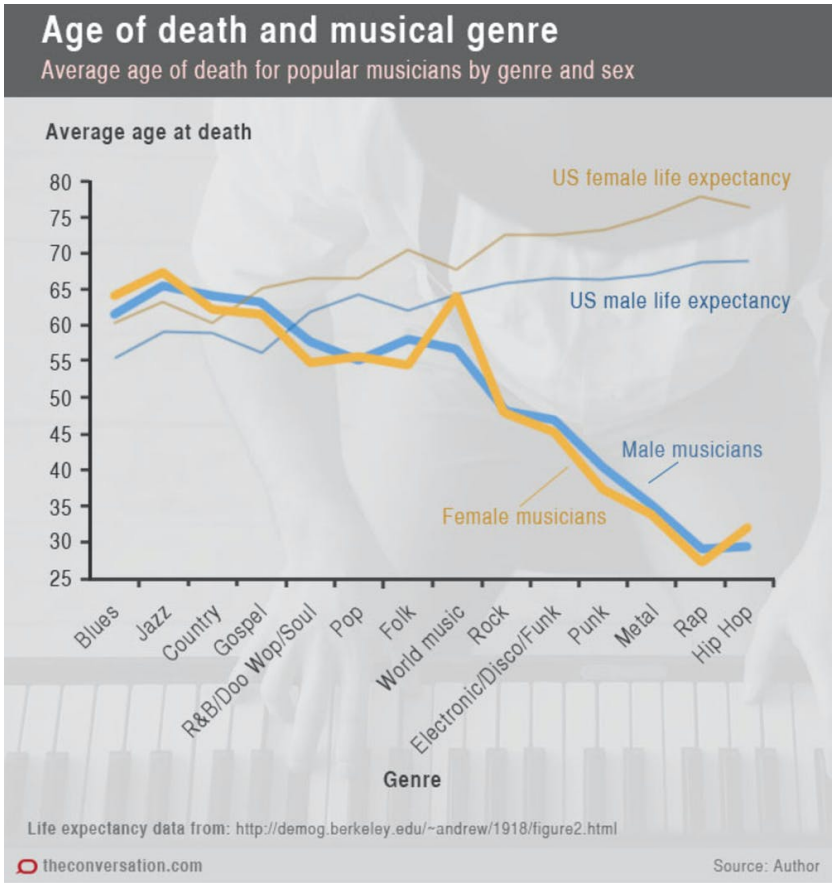
Statistical traps (cont'd)



Year	1985	1990	1995	2000	2005
Average	19.69	17.03	16.21	17.80	21.17
Std. Dev.	10.61	10.46	9.90	10.84	11.76

Statistical traps (cont'd)

- Will joining a punk rock band reduce your life expectancy?
  (The conversation: Music to die for
  `https://theconversation.com/music-to-die-for-how-genre-affects-popular-musicians-life-expectancy-36660`)



Statistical traps (cont'd)

Right censoring (Right censored data is data for items that have not yet failed.)

- Employment data: we actually right censored the data.
- Musician death by genre: implicitly right censored.
  - most rappers and hip-hop artists are still relatively young; deaths are premature, accidental.
  - A rapper is perhaps relatively younger than a gospel/country singers (on an average) ...
  - A major factor is perhaps not the musical genre but the age of musicians in a genre.
- Again, correlation is misinterpreted as a causal influence.

Further explorations: causality and statistical traps

- Bayes rule and conditional probability.
- Simpson's paradox
- Average: mean or median.
- Survivor-ship bias, Length time/lead-time bias.
- ...

Outline

Spotting Fake News

What is a fake news? Deliberately distorted information created to deceive and manipulate the audience.

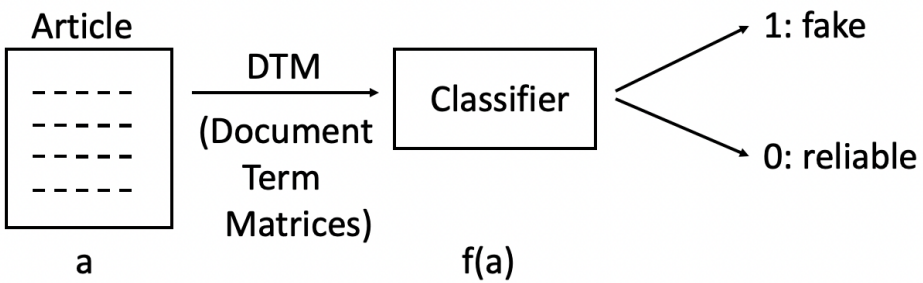
Challenge: Manual fact checking is a daunting task in the era of social media.

A supervised learning approach

- This is a text classification problem, where we want to classify articles / posts as reliable (0) and fake (1)
- To this purpose, we learn a classifier defined as

$$f(a) = \begin{cases} 1, & \text{if } a \text{ is a piece of fake news} \\ 0, & \text{otherwise} \end{cases}$$

where a is the text of the article we want to verify. $f(\cdot)$: logistic regression, CART, Bagging, Random forests.



Preparing the data

- We first need to transform each news article into a numerical representation in the form of a vector, known in this field as Document-Term Matrix (DTM)

Documents

Terms

	War	Peace	Election	...
Doc_1	2	0	0	
Doc_2	0	1	1	
⋮	...	...	...	
Doc_n	0	0	1	

- This is a long and complex process, on which we will focus in Week 10
- Today, we will work with existing DTMs and try to spot fake news with Random Forests

Blogs and Articles

- https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal#cite_note-:10-9
- https://www.businessinsider.com/cambridge-analytica-a-guide-to-the-trump-linked-data-firm-that-harvested-50-million-facebook-profiles-2018-3#where-did-it-come-from-3
- https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump
- https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html

Blogs and Articles (cont'd)

- Ursula Garzcarek and Detlef Steuer. "Approaching Ethical Guidelines for Data Scientists." *Applications in Statistical Computing*. Springer, Cham, 2019. 151-169. (`https://link.springer.com/chapter/10.1007/978-3-030-25147-5_10`)
- `https://www.fatml.org`
- Alberto Cairo. "Ethical infographics In data visualization, journalism meets engineering." 2014. (`https://www.dropbox.com/s/pqgmg02yz0pgju4/EthicalInfographics.pdf`)

Courses

- *Fairness in Machine Learning*, at UC Berkeley
  (`https://fairmlclass.github.io`)
- *Data Science Ethics*, at Yale
  (`https://datascienceethics.wordpress.com/`)
- *Responsible Data Science*, at NYU
  (`https://dataresponsibly.github.io/courses/spring19/`)
- *Applied Data Ethics*, at fast.ai
  (`https://www.fast.ai/2020/08/19/data-ethics/`)

Courses

- *Data Privacy and Ethics*, at Stanford University (`https://web.stanford.edu/group/msande234/cgi-bin/wordpress/`)
- *Ethics in Data Science*, at University of Utah (`https://utah.instructure.com/courses/462398/assignments/syllabus`)
- *Calling Bullshit*, at University of Washington (`https://www.callingbullshit.org`)