40.016 The Analytics Edge

Ethics in Data Analytics

Stefano Galelli, Bikramjit Das

SUTD

Term 6, 2020

Outline

- Introduction
- Using and Abusing Data Visualization
- Causality / Statistical traps (with R)
- Spotting Fake News (with R)

Introduction

Needs for Ethics in Data Analytics

1. Important aspect of the accreditation process
2. 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

Introduction

Needs for Ethics in Data Analytics

1. Important aspect of the accreditation process
2. 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

–> The scale and ease with which analytics can be conducted today completely changes the ethical framework

Introduction

The Facebook–Cambridge Analytica data scandal

Figure 1: Source: https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump

Introduction

The Facebook–Cambridge Analytica data scandal

What did Cambridge Analytica do?

- In 2015, Cambridge Analytica harvested information from $\sim$ 87 million Facebook users
- Data were sourced through an app / personal quiz, "thisisyourdigitallife"
- The data were then used in political campaigns with a technique known as *micro-targeting*

Introduction

The Facebook–Cambridge Analytica data scandal

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

Introduction

The Facebook–Cambridge Analytica data scandal

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

Introduction

The Facebook–Cambridge Analytica data scandal

(Some) Ethical issues:

- Manipulation of users
- Privacy
- Transparency

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

Introduction

Ethical issues in data science

1. *Bias, discrimination, and exclusion:* Algorithms and artificial intelligence can create biases, discrimination or even exclusion towards individuals and groups of people.

Introduction

Ethical issues in data science

1. *Bias, discrimination, and exclusion:* Algorithms and artificial intelligence can create biases, discrimination or even exclusion towards individuals and groups of people.
2. *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.

Introduction

Ethical issues in data science

3. *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.

Introduction

Ethical issues in data science

3. *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.
4. *Quality, quantity, relevance:* The acceptance of the existence of potential bias in datasets curated to train algorithms is of paramount importance.

Introduction

(Some) Governing ethical principles

1. *Ownership* (Who owns the data?)
2. *Transaction transparency*
3. *Consent*
4. *Privacy*
5. *Currency*
6. *Openness*

Introduction

(Some) Governing ethical principles

7. *Fairness*
8. *Justice*
9. *Beneficence*
10. *Non-maleficence*

Introduction

Guidelines

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

Introduction

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.

Introduction

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.

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)

Using and Abusing Data Visualization

Example 1

Figure 2: Argentina's number of COVID-19 tests, original plot. (Source: https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9)

Using and Abusing Data Visualization

Example 1

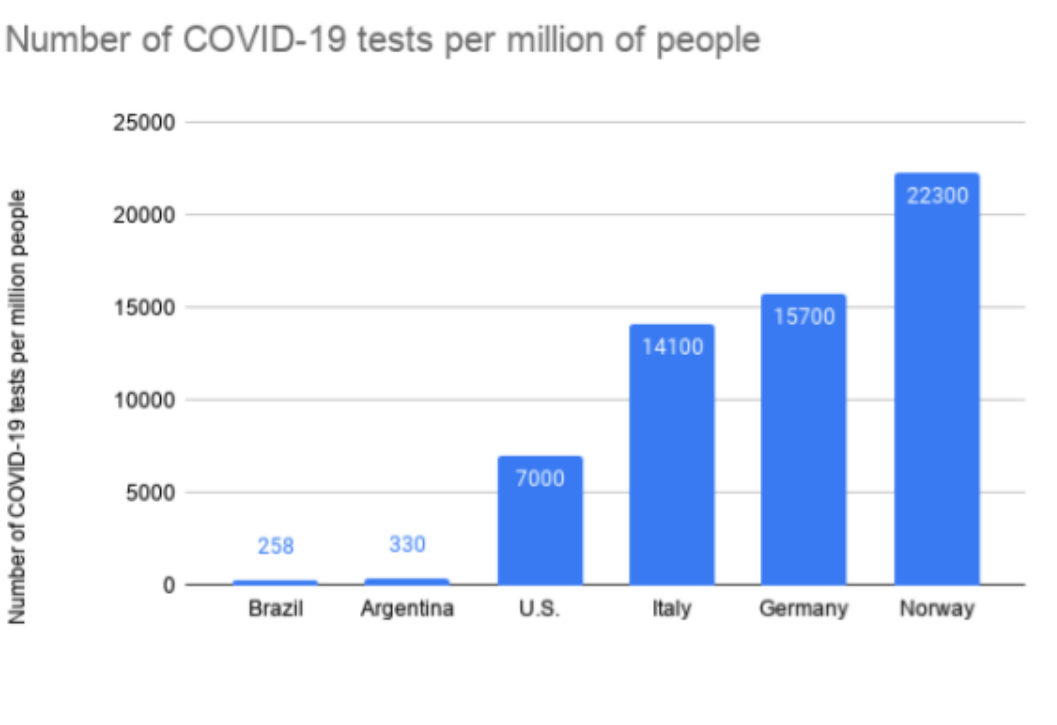


Figure 3: Argentina's number of COVID-19 tests, modified / correct plot. (Source: https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9)

Using and Abusing Data Visualization

Example 2

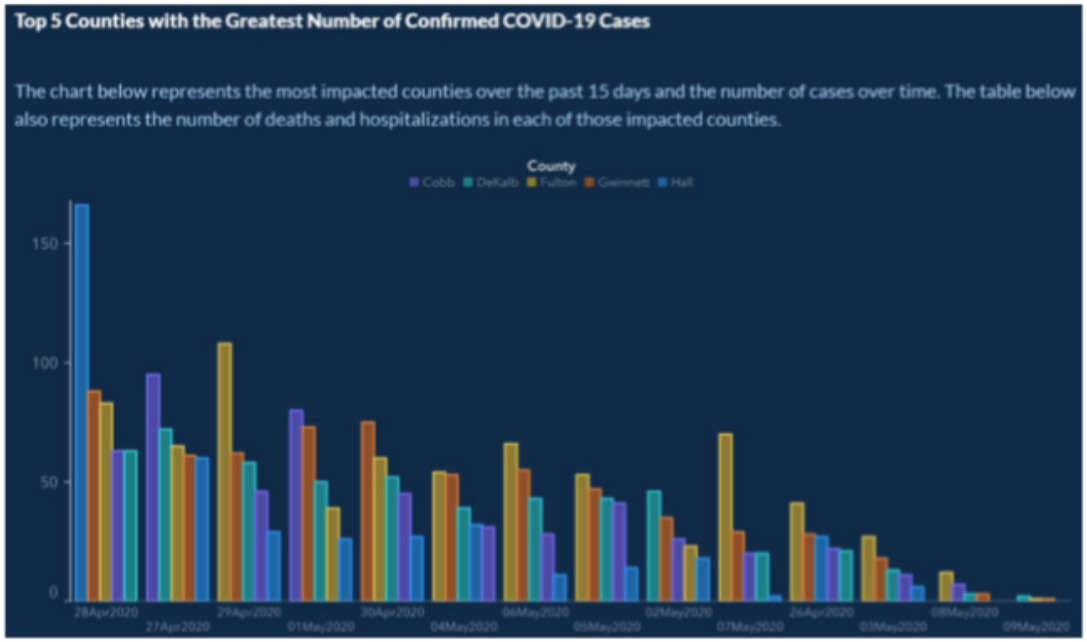


Figure 4: Number of COVID-19 cases in five counties in Georgia. The plot was retrieved from the Georgia Department of Public Health! (Source: https://towardsdatascience.com/stopping-covid-19-with-misleading-graphs-6812a61a57c9)

Using and Abusing Data Visualization

Issue 1: Omitting the baseline

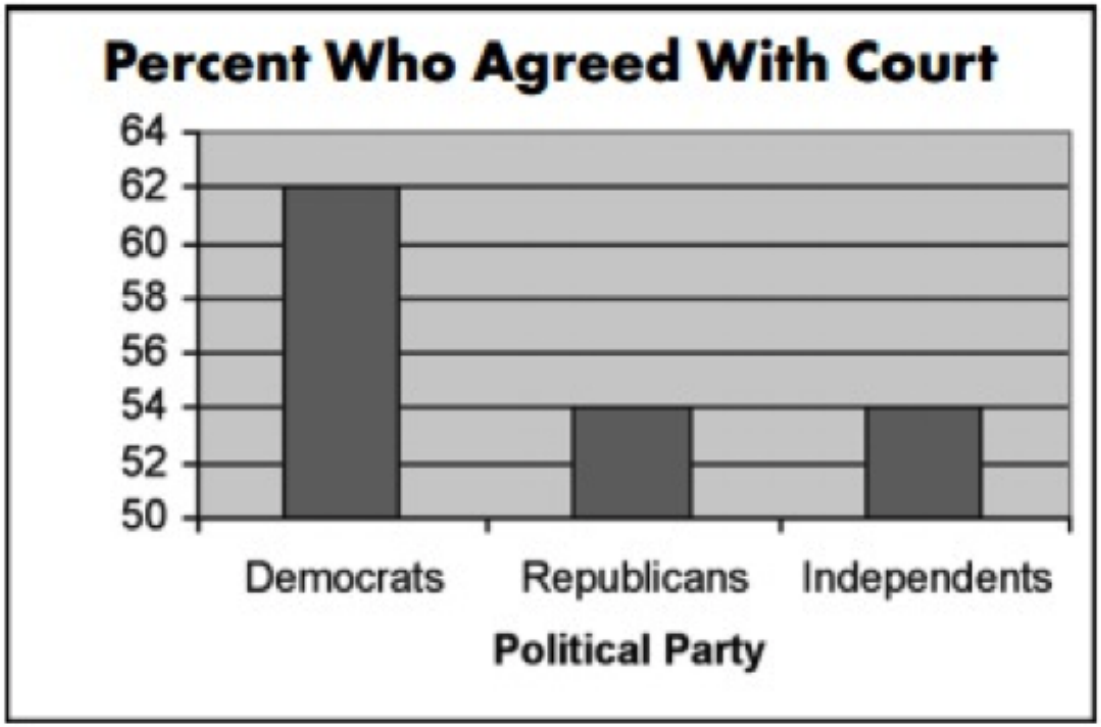


Figure 5: Original plot. (Source: https://venngage.com/blog/misleading-graphs/)

Using and Abusing Data Visualization

Issue 1: Omitting the baseline

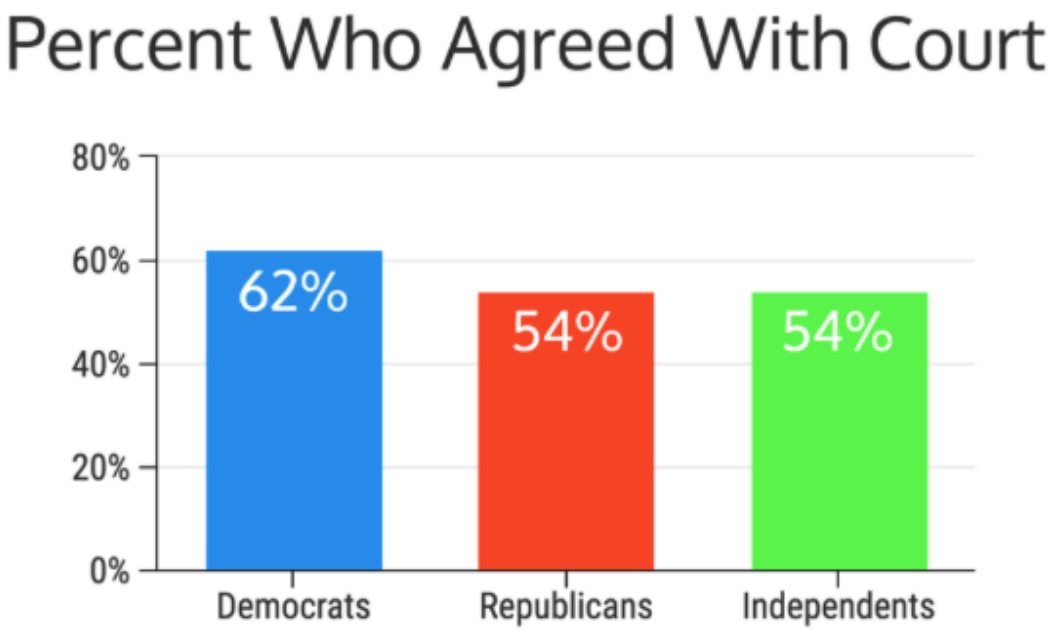


Figure 6: Modified / Correct plot. (Source: https://venngage.com/blog/misleading-graphs/)

Using and Abusing Data Visualization

Issue 2: Manipulating the y-axis

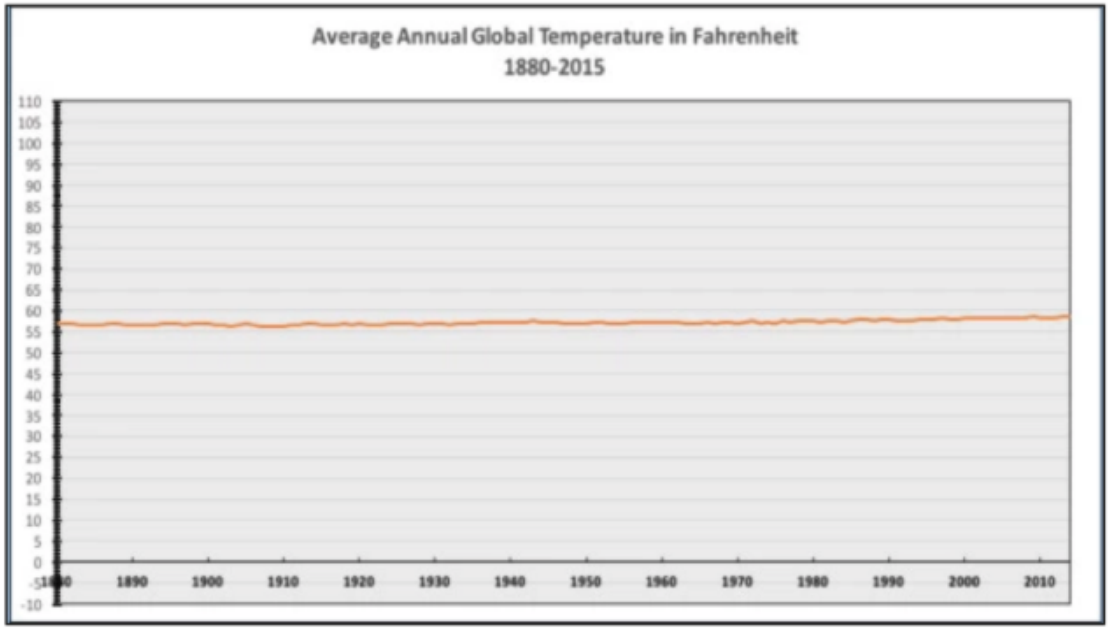


Figure 7: Original plot. (Source: https://venngage.com/blog/misleading-graphs/)

Using and Abusing Data Visualization

Issue 2: Manipulating the y-axis

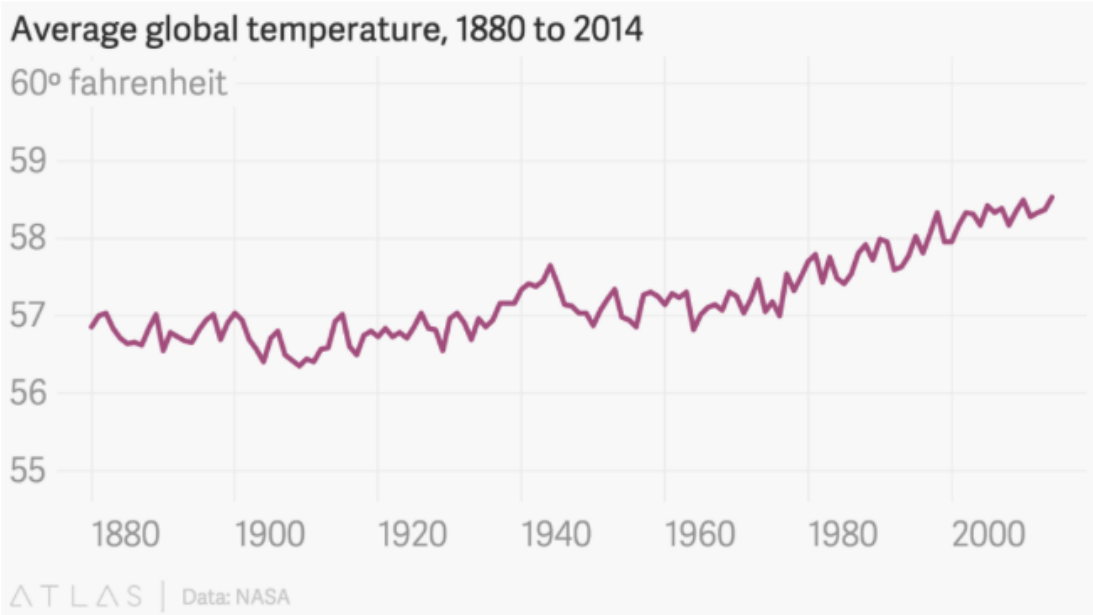


Figure 8: Modified / Correct plot. (Source: https://venngage.com/blog/misleading-graphs/)

Using and Abusing Data Visualization

Issue 3: Cherry picking data

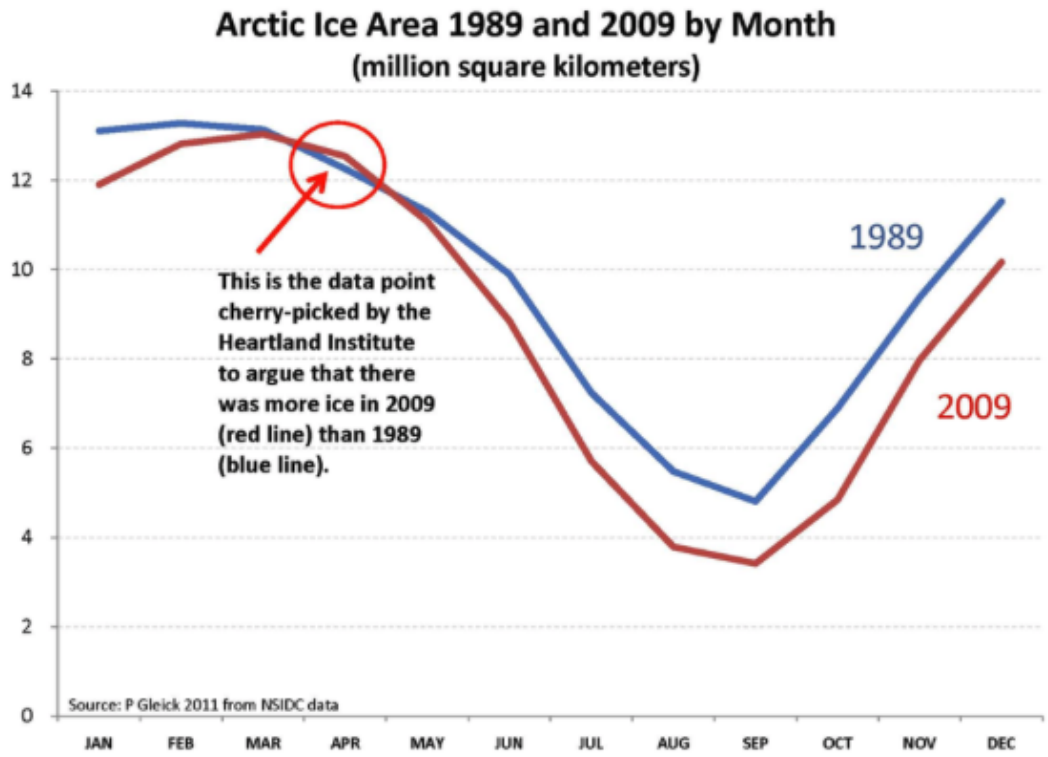


Figure 9: Original plot. (Source: https://venngage.com/blog/misleading-graphs/)

Using and Abusing Data Visualization

Issue 4: Using the wrong graph

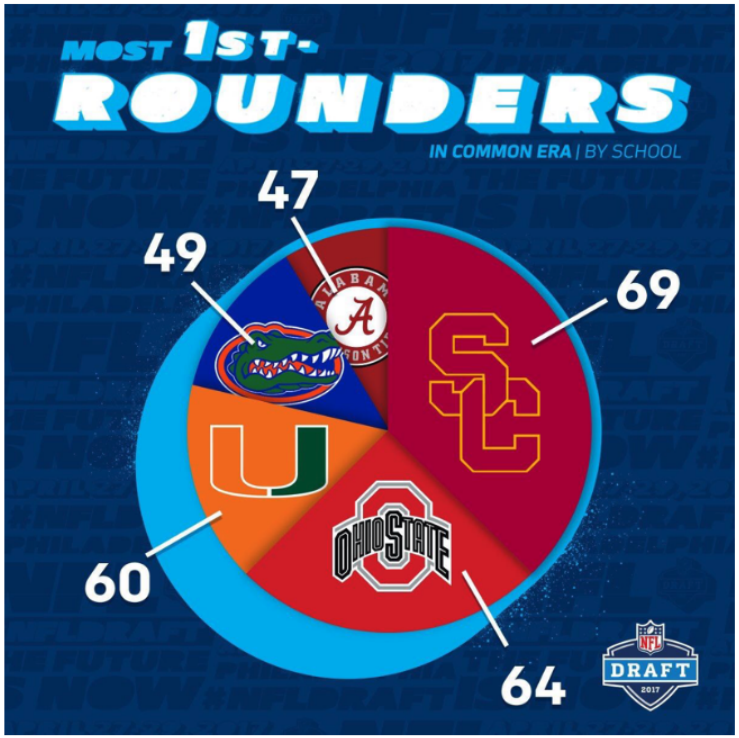


Figure 10: Original plot. (Source: https://venngage.com/blog/misleading-graphs/)

Using and Abusing Data Visualization

Issue 4: Using the wrong graph

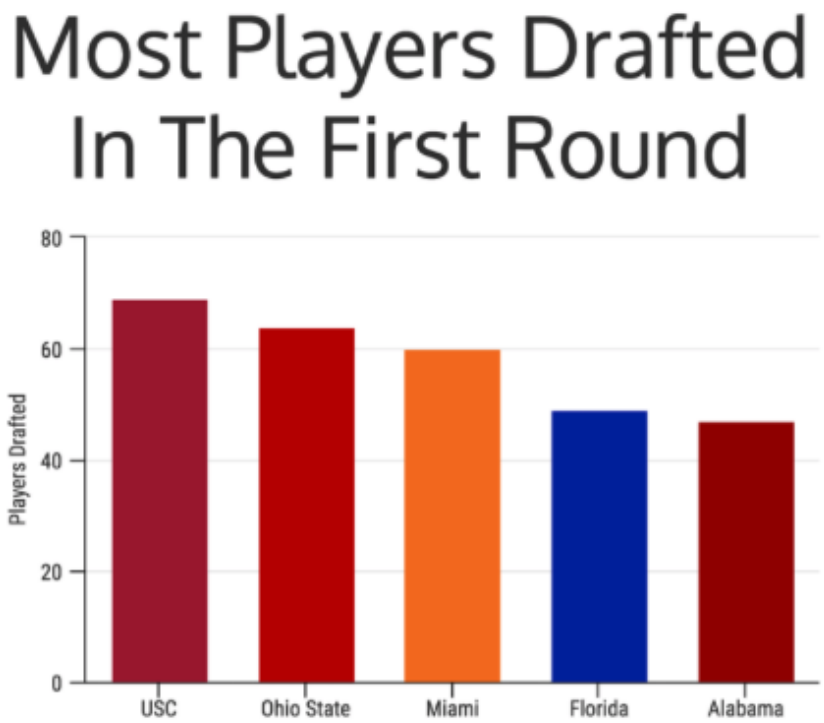


Figure 11: Modified / Correct plot. (Source: https://venngage.com/blog/misleading-graphs/)

Using and Abusing Data Visualization

Issue 5: Going against conventions

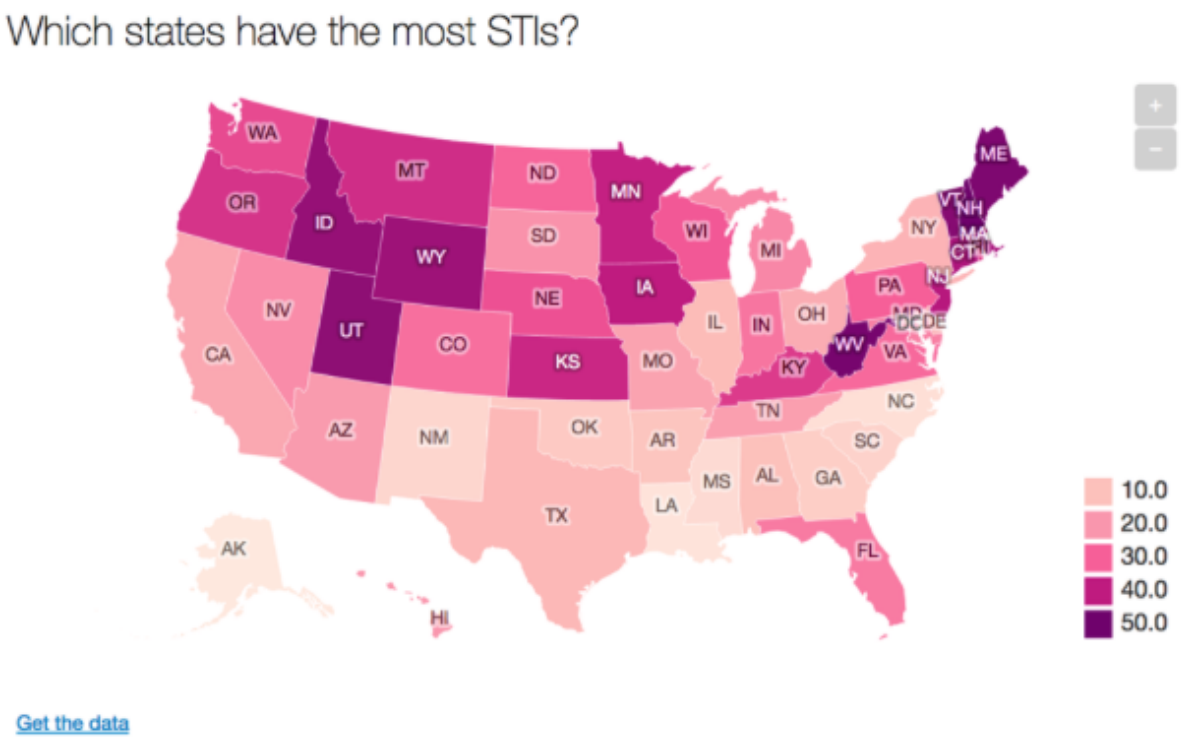


Figure 12: Original plot. (Source: https://venngage.com/blog/misleading-graphs/)

Using and Abusing Data Visualization

Other common issues:

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

Using and Abusing Data Visualization

Closure

A hippocratic oath for visualization (by Jason Moore): *"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."*

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.*
- Open `storks.csv` and `CvsC.Rmd` files in R.

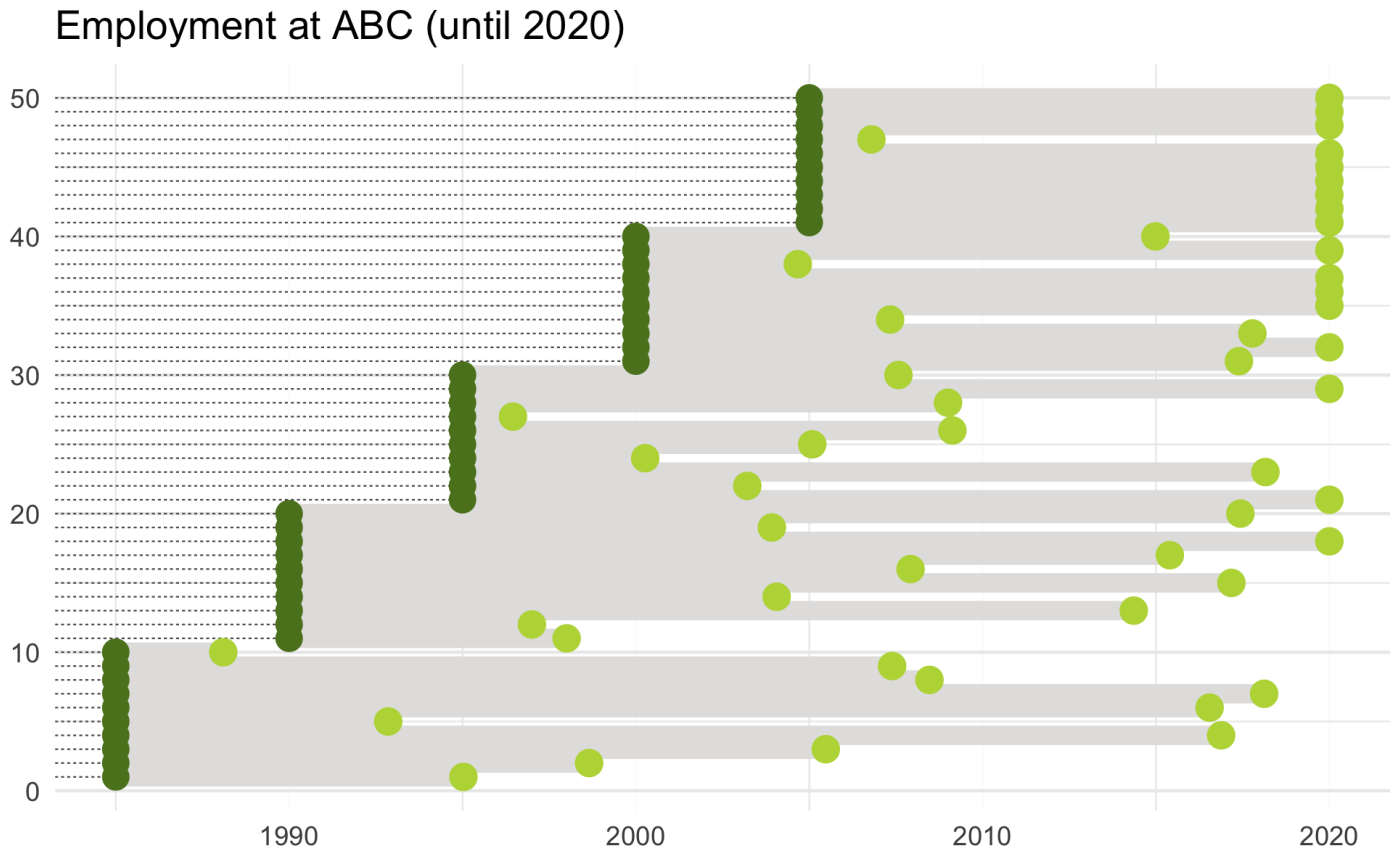
Correlation vs. Causation

Example: *Milton Friedman's thermostat*

- 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.
- Since you find neither O nor E has an influence on indoor temperature, you decide to switch your A/C off.
- In financial economics, good fiscal policies are supposed to maintain a constant inflation rate.

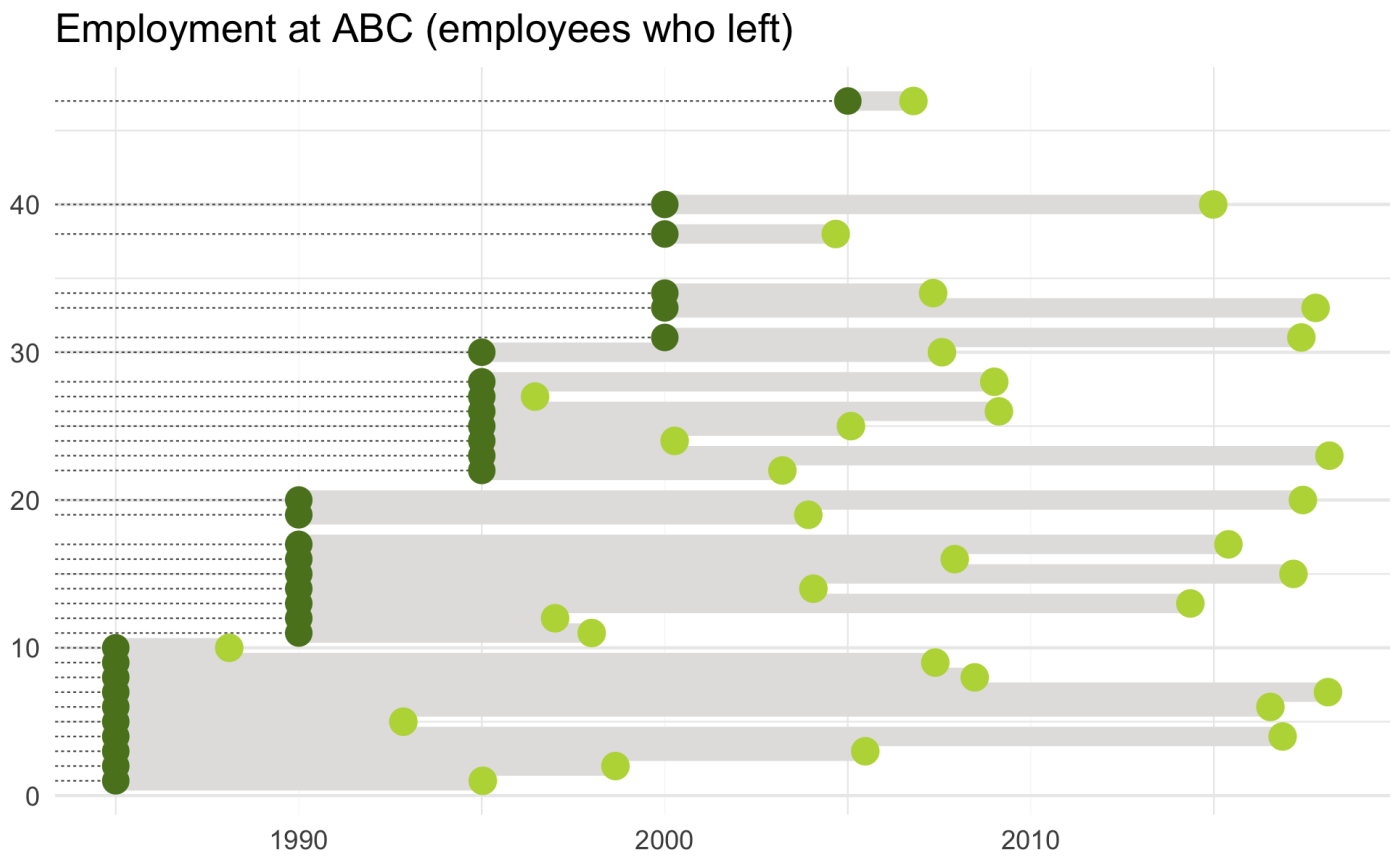
Statistical traps

- Are employee tenures shorter than earlier at ABC? [1]



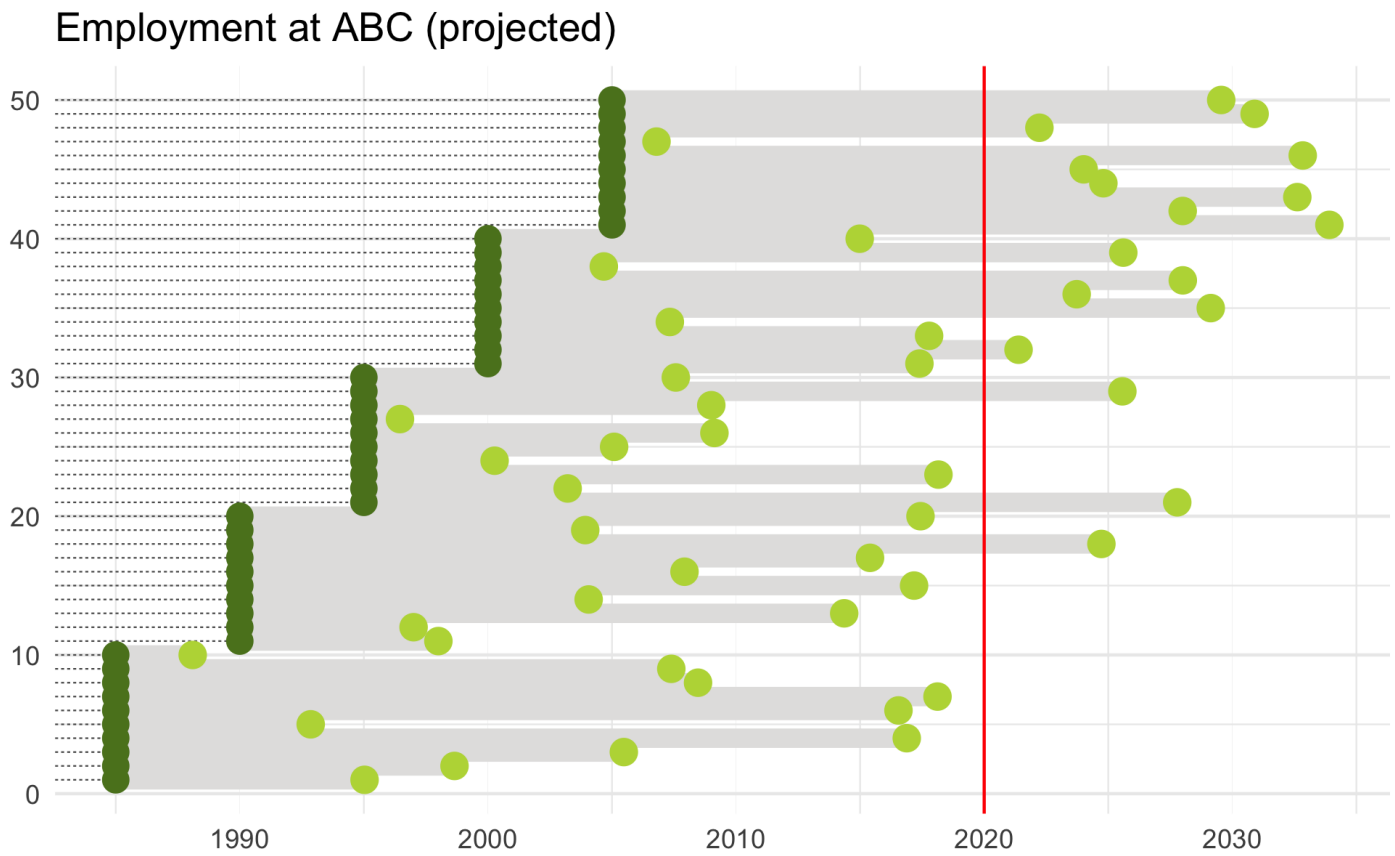
[1] This is a simulated dataset.

Statistical traps



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

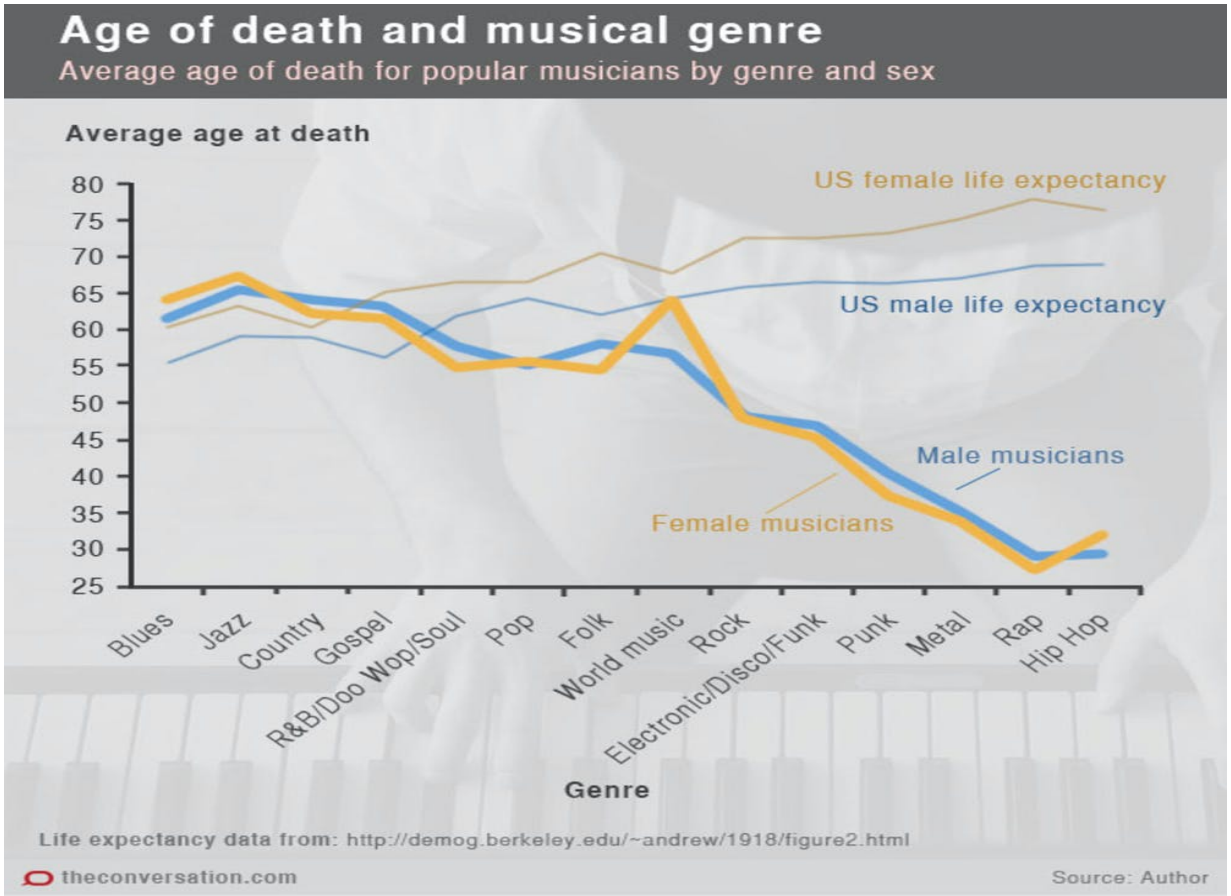
Statistical traps



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

- Will joining a punk rock band reduce your life expectancy? [2]



[2] *The conversation: Music to die for*

Statistical traps

Right censoring

- 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.
- ...

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.

Spotting Fake News

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.

Spotting Fake News

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)**

Spotting Fake News

- 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)**
- This is a long and complex process, on which we will focus in Week 9
- Today, we will work with existing DTMs and try to spot fake news with Random Forests

Additional Material

Blogs and Articles

- https://en.wikipedia.org/wiki/Facebook\T1\textendashCambridge_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

Additional Material

Blogs and Articles

- Ursula Garzcarek and Detlef Steuer. "Approaching Ethical Guidelines for Data Scientists." *Applications in Statistical Computing*. Springer, Cham, 2019. 151-169. Link
- https://www.fatml.org
- Alberto Cairo. "Ethical infographics In data visualization, journalism meets engineering." 2014. Link

Additional Material

Courses

- *Fairness in Machine Learning*, at UC Berkeley
- *Data Science Ethics*, at Yale
- *Responsible Data Science*, at NYU
- *Applied Data Ethics*, at fast.ai
- *Data Privacy and Ethics*, at Stanford University
- *Ethics in Data Science*, at University of Utah
- *Calling Bullshit*, at University of Washington