Ethics in Data Analytics

1 Introduction

Ethics in Data Analytics is an 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. The scale and ease with which analytics can be conducted today completely changes the ethical framework.

1.1 The Facebook-Cambridge Analytica data scandal

The Cambridge Analytica Files

'I made Steve Bannon's psychological warfare tool': meet the data war whistleblower

▲ Former Cambridge Analytica employee Christopher Wylie Photograph: Antonio Olmos/The Observer

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

In 2015, Cambridge Analytica harvested information from millions of 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. Cambridge Analytica harvested personal information on where users lived, what pages they liked, etc. These data were used to build psychological profiles, with traits like openness, agreeableness, IQ, gender, age and political views. Here psychological profiles are a fundamental piece of information for making micro-targeting successful.

In response, 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.

Some ethical issues involvesd in this example are manipulation of users, privacy and transparency. Other issues, not discussed here, pertain to the legal aspects of the scandal.

1.2 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.
- 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.
- 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.

Governing ethical principles. Ownership (Who owns the data?); Transaction transparency; Consent; Privacy; Openness; 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.

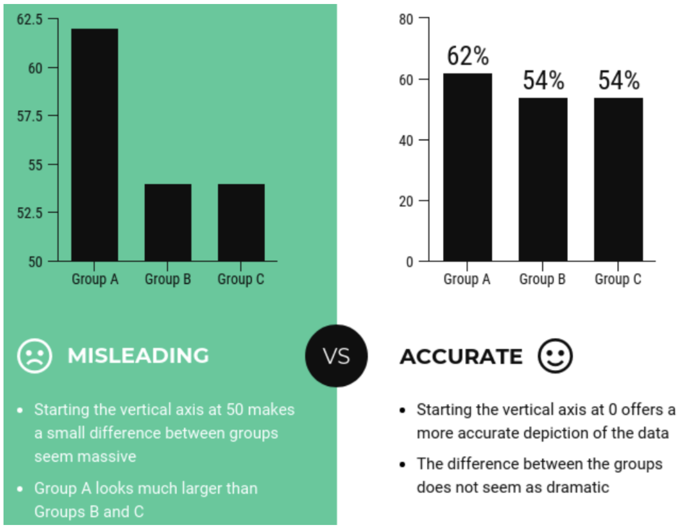
2 Using and Abusing Data Visualization

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).

2.1 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".



MISLEADING

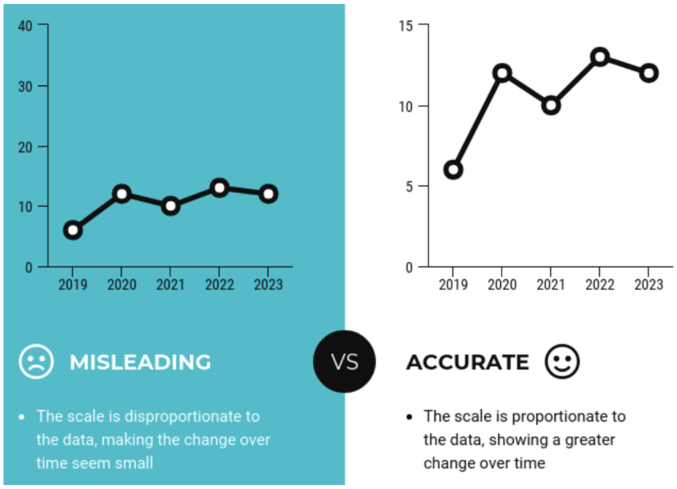
- Starting the vertical axis at 50 makes a small difference between groups seem massive
- Group A looks much larger than Groups B and C

ACCURATE

- Starting the vertical axis at 0 offers a more accurate depiction of the data
- The difference between the groups does not seem as dramatic

Figure 2: Issue 1. (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.



MISLEADING

- The scale is disproportionate to the data, making the change over time seem small

ACCURATE

- The scale is proportionate to the data, showing a greater change over time

Figure 3: Issue 2. (Source: `https://venngage.com/blog/misleading-graphs`)

A real example on the climate change in Fahrenheit is as follows, where the original plot and corrected plot are shown.

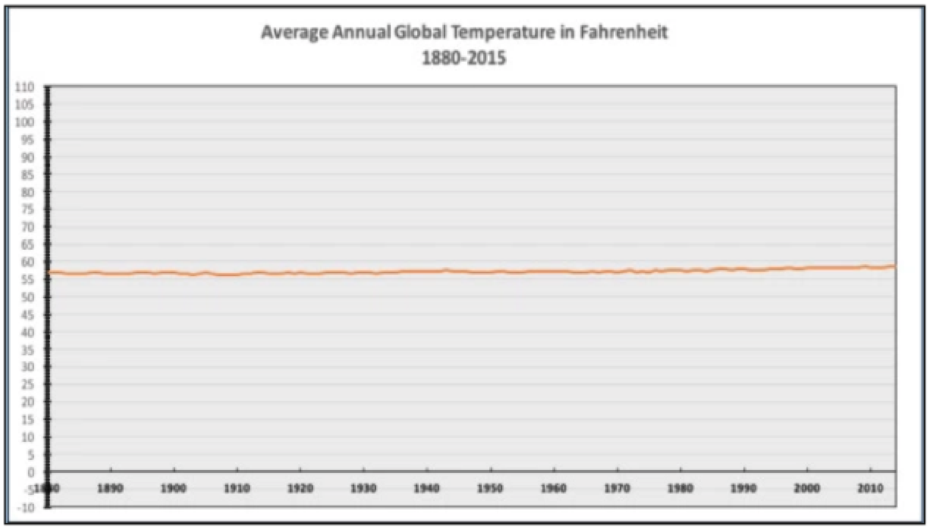


Figure 4: Original plot. (Source: `https://venngage.com/blog/misleading-graphs/`)

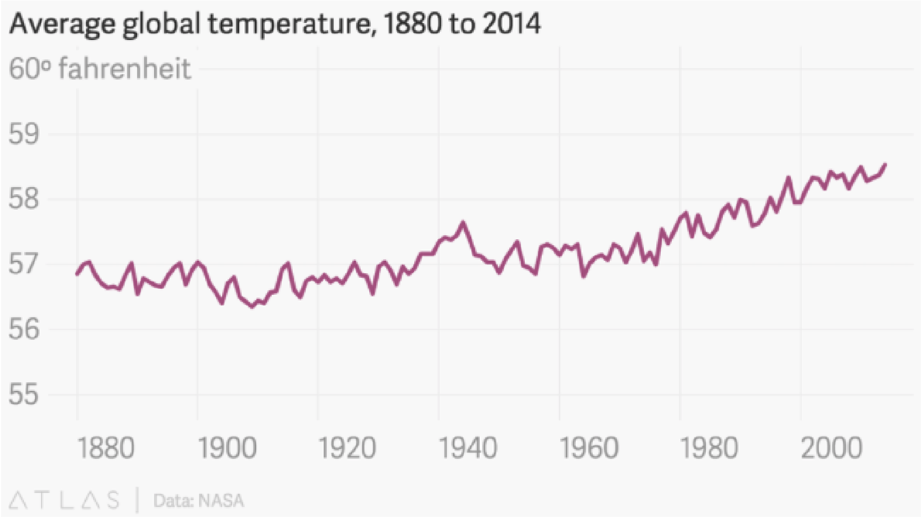


Figure 5: Modified / Correct plot. (Source: `https://venngage.com/blog/misleading-graphs/`)

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.

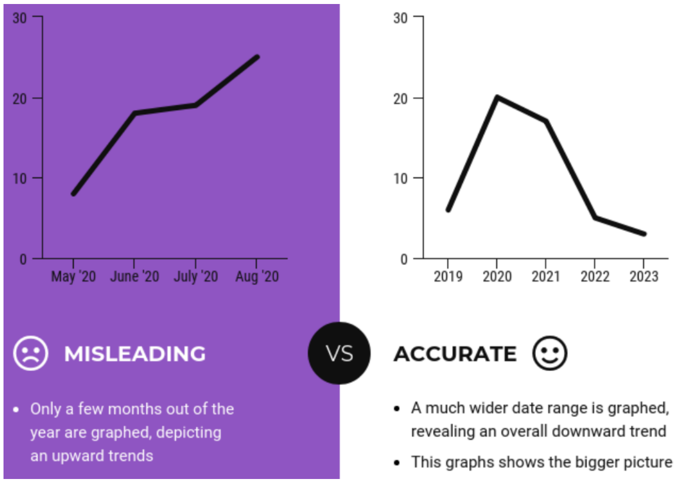


Figure 6: Issue 3. (Source: `https://venngage.com/blog/misleading-graphs`)

A real example on Arctic Ice Area is shown in the following figure.

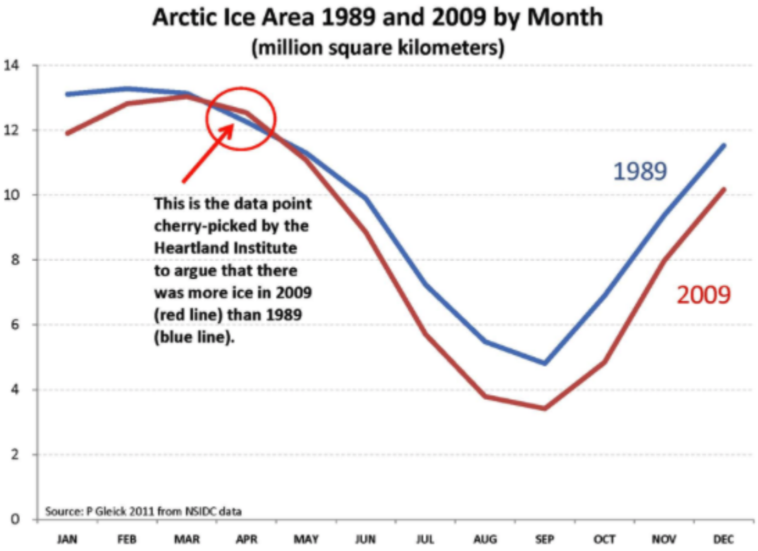
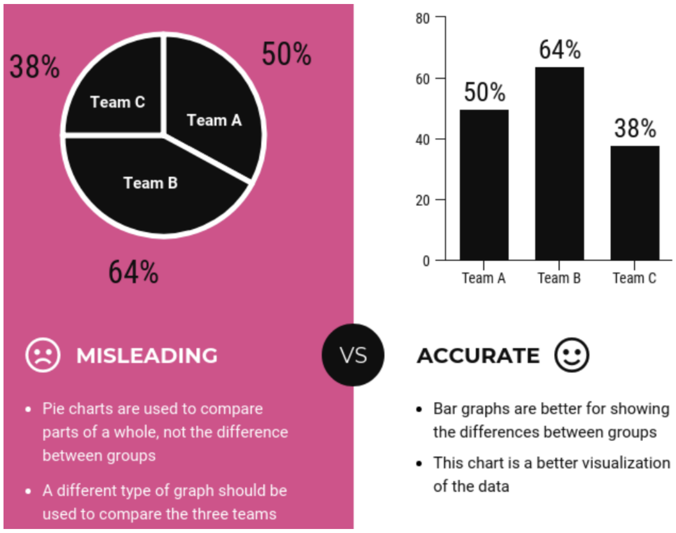


Figure 7: Original plot. (Source: `https://venngage.com/blog/misleading-graphs/`)

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.

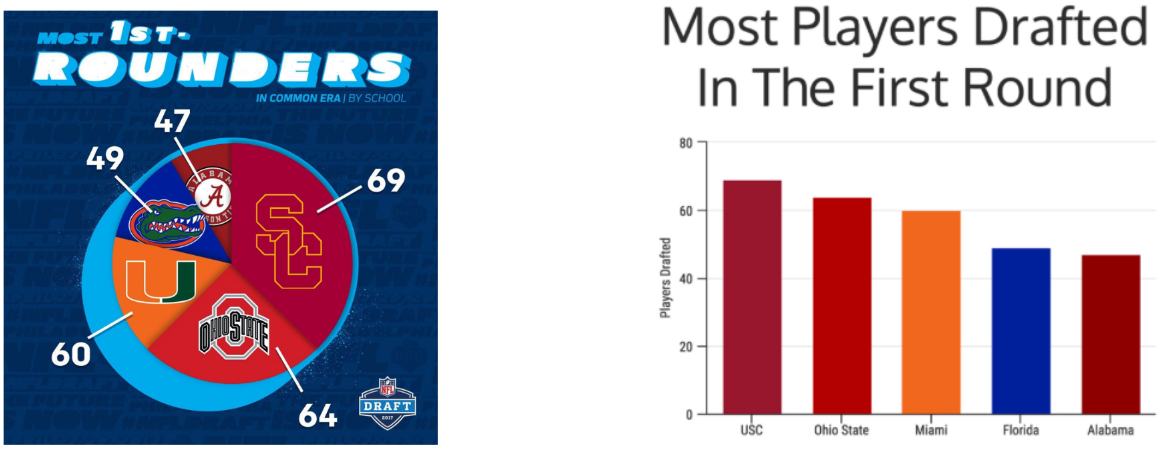


- Pie charts are used to compare parts of a whole, not the difference between groups
- A different type of graph should be used to compare the three teams

- Bar graphs are better for showing the differences between groups
- This chart is a better visualization of the data

Figure 8: Issue 4. (Source: `https://venngage.com/blog/misleading-graphs/`)

A real example on players in the first round is as follows, where the original plot and corrected plot are shown.

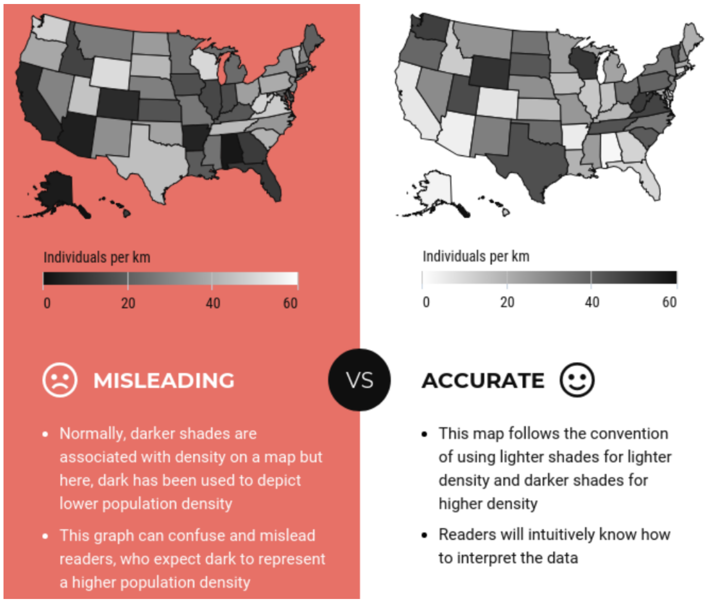


(a) Original plot.

(b) Modified / Correct plot.

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

Issue 5: Going against conventions. Over time, we have developed standards for how data is visualized. Flipping those conventions can make a graph confusing or misleading to readers.



- Normally, darker shades are associated with density on a map but here, dark has been used to depict lower population density
- This graph can confuse and mislead readers, who expect dark to represent a higher population density

- This map follows the convention of using lighter shades for lighter density and darker shades for higher density
- Readers will intuitively know how to interpret the data

Figure 10: Issue 5. (Source: `https://venngage.com/blog/misleading-graphs/`)

There are also some other common issues: improper intervals or units, omitting data, extrapolation unnecessary complexity and so on.

2.2 Closure

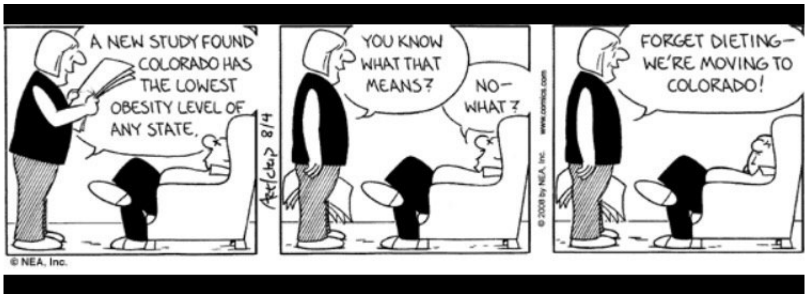
A hippocratic oath for visualization (by Jason Moore from US Air Force Research Laboratory): *"I shall not use visualization to intentionally hide or confuse the truth which it is intended to portray. I will respect the great power visualization has in garnering wisdom and misleading the uninformed. I accept this responsibility willfully and without reservation, and promise to defend this oath against all enemies, both domestic and foreign."*

3 Causality / Statistical traps (with R)

3.1 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. Some examples of spurious correlations are shown in `https://tylervigen.com`.

Example: Milton Friedman's thermostat. Details can be found in `https://justinhohn.typepad.com/blog/2013/01/milton-friedmans-thermostat-analogy.html`.

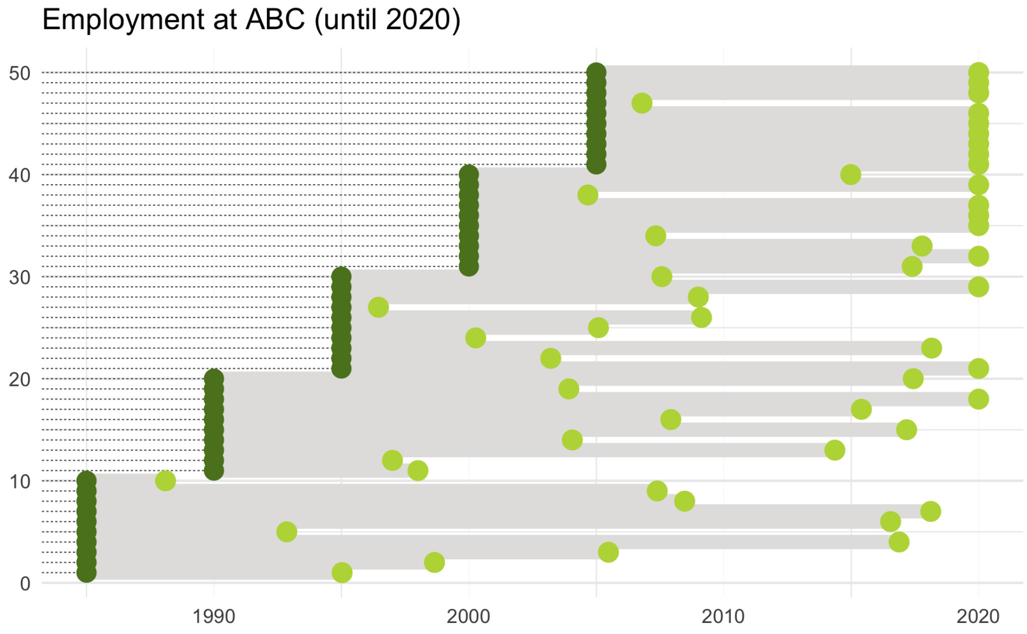

A NEW STUDY FOUND COLORADO HAS THE LOWEST OBESITY LEVEL OF ANY STATE.
YOU KNOW WHAT THAT MEANS?
NO— WHAT?
FORGET DIETING— WE'RE MOVING TO COLORADO!
8/4
© 2008 by NEA, Inc.
www.comics.com
© NEA, Inc.

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.

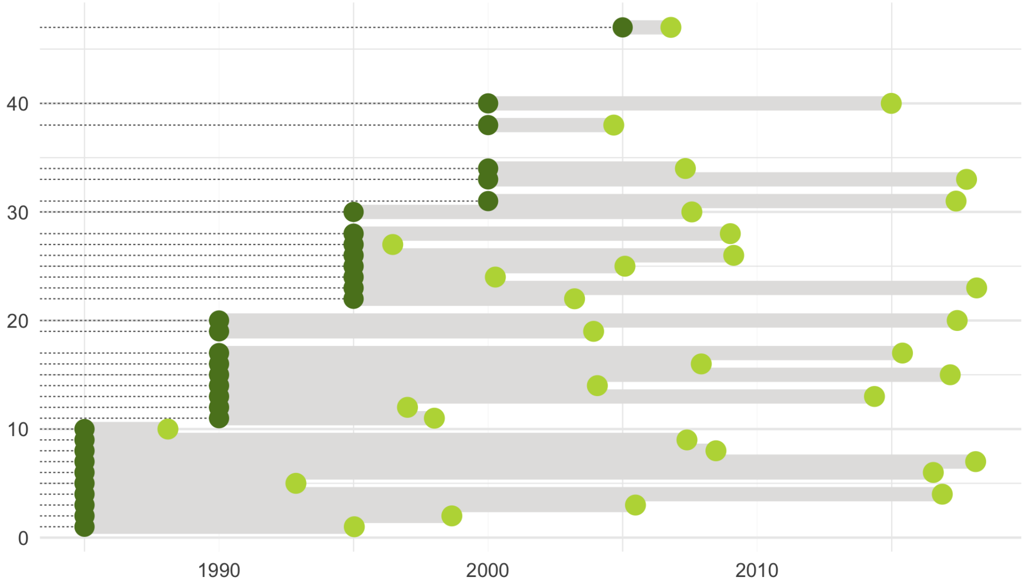
In financial economics, good fiscal policies are supposed to maintain a constant inflation rate.

3.2 Statistical traps

Example 1. Are employee tenures shorter than earlier at ABC? (This is a simulated dataset.)

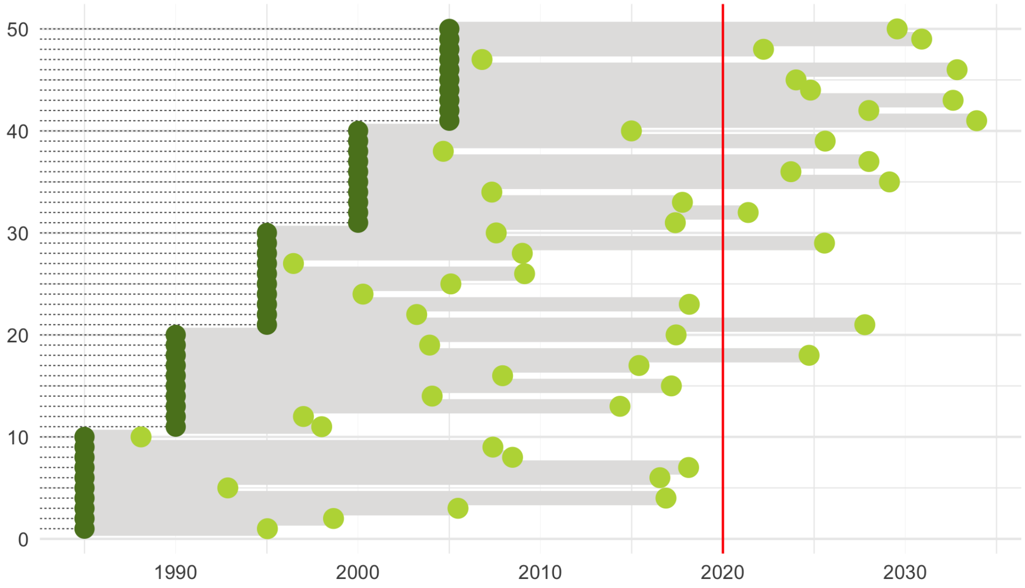


Employment at ABC (employees who left)



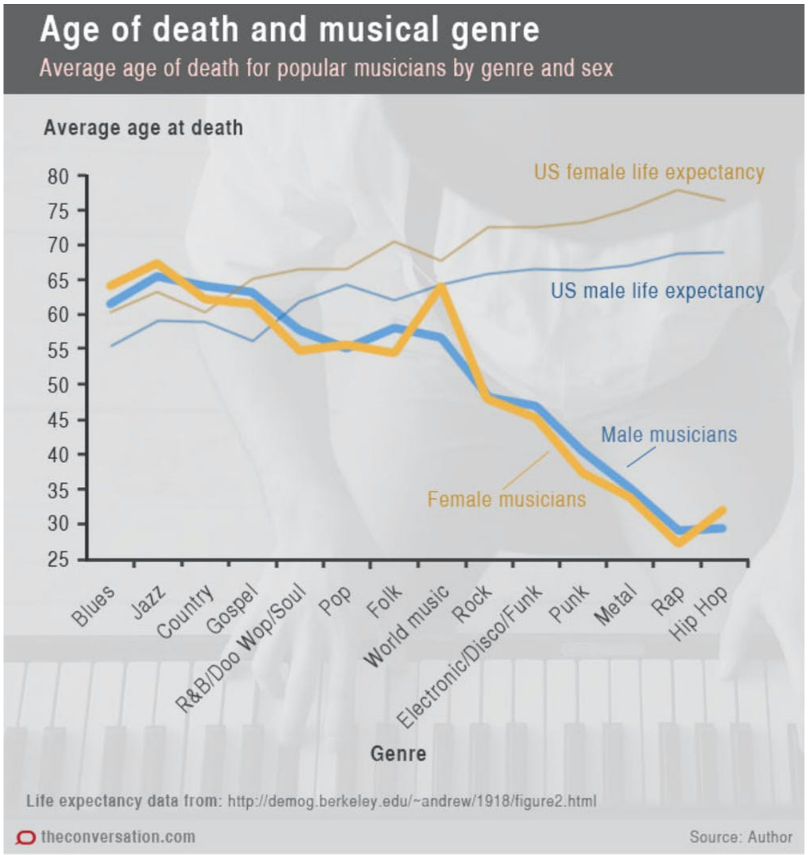
Year	1985	1990	1995	2000	2005
Average	19.69	13.26	12.37	8.65	8.86
Std. Dev.	10.61	7.63	6.34	6.42	3.74

Employment at ABC (projected)



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

Example 2. 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`)



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.

3.3 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.
- ...

4 Spotting Fake News (with R)

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.

4.1 A supervised learning approach

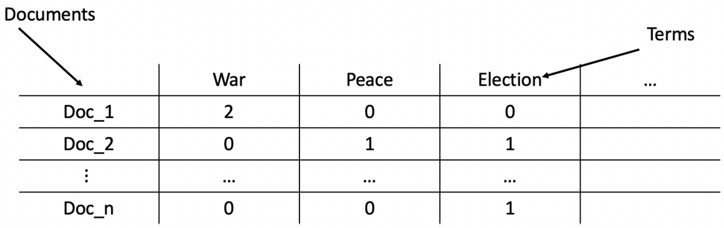
This is a **text classification problem**, where we want to classify articles / posts as reliable (0) and fake (1). To this purpose, we learn a classifier defined as

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

where a is the text of the article we want to verify. Here, the unknown function $f(\cdot)$ can be estimated via: logistic regression, CART, Bagging, Random forests...

4.2 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. In this lecture, 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
- 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/)
- *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)

Reference

[1] James et al. (2014) *An Introduction to Statistical Learning with Applications in R*, Springer, 2014. Chapter 5.2 and 8.2.