Causes & impacts of data bias (Answers)

Worksheet section	Contents
1	Types of data bias
2	Causes & mitigation of bias

Version: 1.0

This lesson has been created by effini in partnership with Data Education in Schools and Skills Development Scotland.

If you require this document in an alternative format, such as large print or a coloured background, please contact

hello@effini.com

or

4th Floor, The Bayes Centre
47 Potterrow
Edinburgh
EH8 9BT

1. Types of data bias

Section 1.1 (recall)

1) Which of these definitions of data ethics is correct?

a) "When data is used to prejudice against a person"

b) "When data is used to prejudice for or against one person or group"

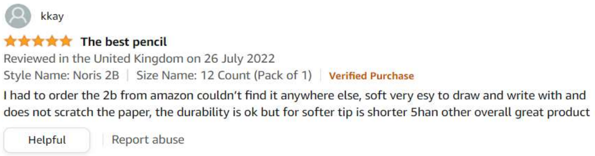
c) "When data is used to prejudice for a group of people"

B

Section 1.2 (apply)

2) Many websites such as Amazon ask customers to provide reviews of the products. Not all customers fill in reviews for the products they buy. How could this cause bias in the reviews of the products?

People tend to fill in reviews if they are really happy or unhappy with a product and/or have time to fill in the reviews. However they may not be representative of everyone who buys the product which could led to bias in the dataset of reviews.



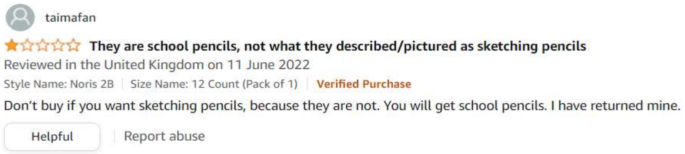
kkay

★★★★★ **The best pencil**

Reviewed in the United Kingdom on 26 July 2022

Style Name: Noris 2B | Size Name: 12 Count (Pack of 1) | **Verified Purchase**

I had to order the 2b from amazon couldn't find it anywhere else, soft very esy to draw and write with and does not scratch the paper, the durability is ok but for softer tip is shorter 5han other overall great product



taimafan

★☆☆☆☆ **They are school pencils, not what they described/pictured as sketching pencils**

Reviewed in the United Kingdom on 11 June 2022

Style Name: Noris 2B | Size Name: 12 Count (Pack of 1) | **Verified Purchase**

Don't buy if you want sketching pencils, because they are not. You will get school pencils. I have returned mine.

Section 1.3 (active)

3) Search engines such as Google can show examples of algorithmic bias.

Reminder: Algorithmic bias is a computer system that makes decisions that are unfair to certain individuals or groups

Use a search engine to find images for these phrases. Describe what the images look like and whether they appear to be biased in any way.

Teacher style

Google search Aug 2022: Nearly every image is of a women, even when c.30% of teachers are men.
https://www.gov.scot/publications/summary-statistics-schools-scotland-9-2018/pages/3/

Plumber person

Based on Google search Aug 2022: All the images are of men but not all plumbers are male.

Nurse person

Based on Google search Aug 2022: All the images are of women but not all nurses are female.

Scottish person

Based on Google search Aug 2022: Nearly all the images are of men, but c.50% of the population of Scotland is female.

1. Types of data bias

If you would like any further reading on how search engines can show algorithmic bias, please read this article from July 2022, "Gender bias in search algorithms has effect on users"

https://www.nyu.edu/about/news-publications/news/2022/july/gender-bias-in-search-algorithms-has-effect-on-users--new-study-.html

Now you are going to play a game called **Survival of the best fit*** that will show you more about algorithmic bias. Click on the link below to start the game.

https://www.survivalofthebestfit.com/game/

4) Were you able to create an unbiased automated solution for hiring new employees?

Most likely the answer will be no.

5) If a company wanted to create an automated recruitment process like the one in the game, how could they try to reduce the impact of bias on the system,

a) When the data is collected?

Make sure the dataset contains a mix of people from different backgrounds.

b) When the data is being analysed/system is being created?

The system is checked that there is no prejudice for or against any groups of people or individuals.

c) When the system is being used to hire new employees?

The system is regularly monitored to make sure it is not prejudiced for or against any groups of people or individuals.

* Survival of the best fit is an open source game built by Built by Gabor Csapo, Jihyun Kim, Miha Klasinc, and Alia ElKattan.
Supported by the Creative Media Award from Mozilla Foundation.
You can follow them on Twitter @sotbf_
For more details, please see https://www.survivalofthebestfit.com/

2. Causes and mitigation of data bias

Section 2.1 (recall)

1) Fill in the missing words in these causes of data bias.

Sample bias	The dataset used for data analysis does not represent the population
Exclusion bias	Predictive variables are removed prior to the data analysis.
Measurement bias	Arises from a systematic issue with the collection of the data.
Confirmation bias	Arises from cherry-picking data that confirms a human's existing beliefs.
Stereotype bias	Cultural or gender stereotypes may led to an uneven distribution of data.
Survivorship bias	Arises from concentrating on the successful output of the process and ignoring those who don't survive.

Section 2.2 (rephase)

2) Explain why this data bias problem has been caused by **confirmation bias.**

A business is looking to launch a new product. They conduct a survey of potential customers and most of the people don't like the new product. However, when the new product is shown to a friend of the product developer their friend really likes it. The product developer places more importance on the views of their friend than the survey of potential customers.

The product developer is cherry-picking the data that confirms their belief that the new product is a good idea.

3) Explain why this data bias problem has been caused by **survivorship bias.**

A TV company is trying to understand what makes a successful TV show. They look at variables related to successful shows such as length of the show, type of show, what time it is broadcast etc to try and predict whether future shows will be successful.

The TV company is only focusing on successful shows and ignoring those that failed.

4) Explain what are the best ways to mitigate against data bias.

1. Review findings with team members and subject matter experts
2. Include multiple data sources from as many providers as possible
3. Have multiple people with differing backgrounds.

2. Causes and mitigation of data bias

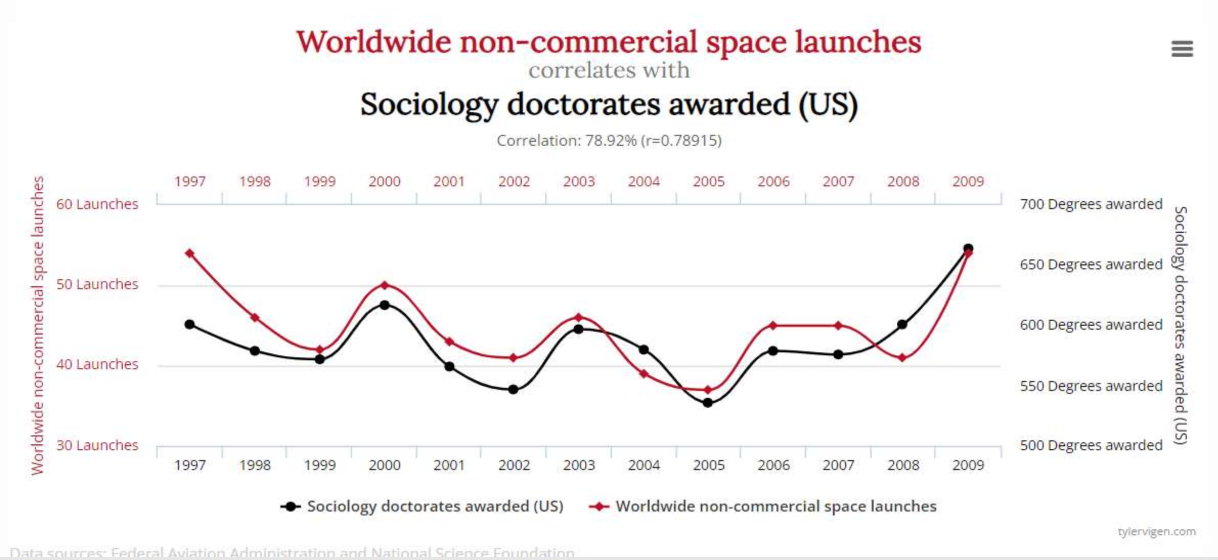
Section 2.3 (active)

5) Correlation bias is when there appears to be a relationship between two data items but they are not actually linked.

The website below contains examples of data items that show the same patterns but are not related.
https://www.tylervigen.com/spurious-correlations

Review the website and copy and paste an interesting example below.

Example of graph from the website



Data sources: Federal Aviation Administration and National Science Foundation