**Data Ethics: Data Bias**

**Overall Word Count (reading and tasks) \_\_\_\_\_\_ (no more than 3000)**

Learning Goals

* **Define data bias and explain how it impacts results**
* **Analyze data for data bias**
* **Discuss how human bias infects ML**
* **Describe how to communicate bias to stakeholders**
* **Discuss ways to control for bias**

Introduction

Data analytics is a hot topic these days. The positive contributions it has made to many an industry in a relatively short period are well documented and have been significant. Without a doubt there are many more brilliant and profound discoveries to be made using data analytics. However, there is a cancer hiding within data analytics that threatens to corrupt the results of your hard work, lead to poor decision making, or cause us to adopt false beliefs. That cancer is Data Bias. Some may believe that data is neutral. Data is objective. Data is the raw truth and cannot be spun or misinterpreted. If you believe those things you would be wrong. Fortunately, there is a remedy for this problem and while it may not control for 100% of all the analytical pitfalls of bias, it can help us to see the truth we seek whenever we initiate a data analytics project. This remedy is Data Ethics. Application of ethics acts to prevent bias from creeping into the collection and analysis of data by mandating the sampling, testing, and processing of data are executed in a sound and consistent manner. This section will discuss some of the more common forms of bias, they can impact result and why a strong ethics foundation is necessary to minimize those impacts.

Sources of Bias

Collection Bias

When there are flaws in the data collection technique or process, biased data finds its way into the system and gives the consumer a distorted picture of reality. This is known as collection bias. For example, suppose you’re planning a vacation and are searching for a nice hotel for your stay. Of course, you realize that all hotels would have you believe they are the best hotel in the world if you ask them. Naturally we seek out the opinions of others who have stayed in the hotel so we can get a more informed perspective. You find Hotel A and it looks great from the pictures and information on the website. But as you read through the customer reviews, you might find that the majority of them are negative, with complaints ranging from poor room services to dirty bathrooms, to strange smells. You decide to go ahead and stay in Hotel A despite the warnings because it is most fairly priced. You are surprised to say after your vacation the Hotel A was not only not as bad as you expected, it was 5 star quality in your book! The flaw in the reviews lay in nature in which the data was collected, which lent itself to bias. It was only possible for you to read the review if the guest was motivated to take the time to write one. In other words, passively collected. Perhaps anger and disappointment are stronger motivating factors for writing a review. Perhaps those that take the time to write reviews are overly critical. Due to being completely reliant on volunteered data from unevaluated sources, the amount of data collected was not representative of the number of guests who have stayed at Hotel A. The website hosting the reviews had no controls in place to verify if those who left reviews actually even stayed at Hotel A. It’s very possible some reviews were maliciously submitted by competitors. It’s important to understand that your data collection process not rely on volunteered data which also leads to inadequate sample sizes. For ethical reasons, collection methods and processes should be transparent.

Sample Bias

Sample Bias is when the sample of a population or data source taken does not adequately represent the true distribution. While your collection technique can give you an adequately sized sample, your sample can still be biased due to lack of diversity within the sample. If you are politically active, you have probably been exposed to polling data which indicates the expected results of an upcoming election. We assume, sometimes mistakenly, that the sample size taken to project results is fair and includes sufficient diversity. Let’s examine a case study of the 1936 Literary Digest Poll. Literary Digest conducted a poll that sought to gain insight on who would be the next President of the United States in 1936. Data was collected from 2.4 million responders who were found through telephone directories, club associates, magazine subscribers, and other sources. Results indicated that Alfred Landon was expected to get 57% of the vote to incumbent President Roosevelt’s 43%. When the election was settled, Roosevelt had won 62% of the vote to Landon’s 38%. While 2.4 million people was an adequate sample size, the poll failed to be accurate due to poor diversity within the sample. Literary Digest failed to recognize that polling only those that were in the telephone directory (telephones were not as common in 1936), had club memberships and magazine subscriptions would give them sample that was skewed towards middle- and upper-class voters. The poll also fell victim to non-response bias, as only 2.4 million people of the 10 million polled responded. Needless to say, those with telephones, club memberships and magazine subscriptions and took time to respond to a poll conducted by mail did not constitute a diverse enough sample from which to draw sound conclusions. Literary Digest was left embarrassed and suffered some damage to their brand as a result.

Suppose some executives from a well-respected media outlet did know there were flaws in their sampling process and decided to present them anyway in the hopes of influencing the outcome of the election. What does that say about how thin the line is between a lack of data ethics and unintentional sample bias? What does this tell you about how much you should trust the results of a study such as this without knowing specifics of how the data was collected and sampled?

Exclusion Bias

When certain features or immeasurable factors are excluded or not included in your dataset because they are believed to be irrelevant or are never considered, your results will be tainted with Exclusion Bias. Whenever the subject of shark attacks is discussed, you may hear people site scientific studies that claim your odds of being attacked by a shark are lower than being struck by lightning. Data from the Florida Museum of Natural History reveals about 82 people per year suffer an unprovoked attack by a shark, with 8 of those dying. They claim your odds of dying from a shark attack are 1 in 3,748,067. In an effort to prevent the proliferation of irrational fear of sharks, in 2017 USA Today published an article citing the data from The Florida Museum of Natural History and contrasted it with the probability of being killed is several other terrible fashions to include being killed by a dog (1 in 112,400) and being struck by lightning (1 in 161,856). The reader may be left with the impression that swimming in the ocean is safer than they might otherwise believe. But the flaw in these statistics is that they exclude data that would most likely show that on a yearly basis, there are far more people exposed to lightning storms for far more time than there are who recreate in the ocean. In order to make an honest comparison, we would need the data for both.

Some ways to help control against Exclusion bias are to get input from a colleague. Having that neutral perspective play devil’s advocate can help you see the bias in the data. Challenge the assumptions your statistics are based on and how they are presented. A wise person once said there are lies, outright lies, and statistics.

Measurement Bias

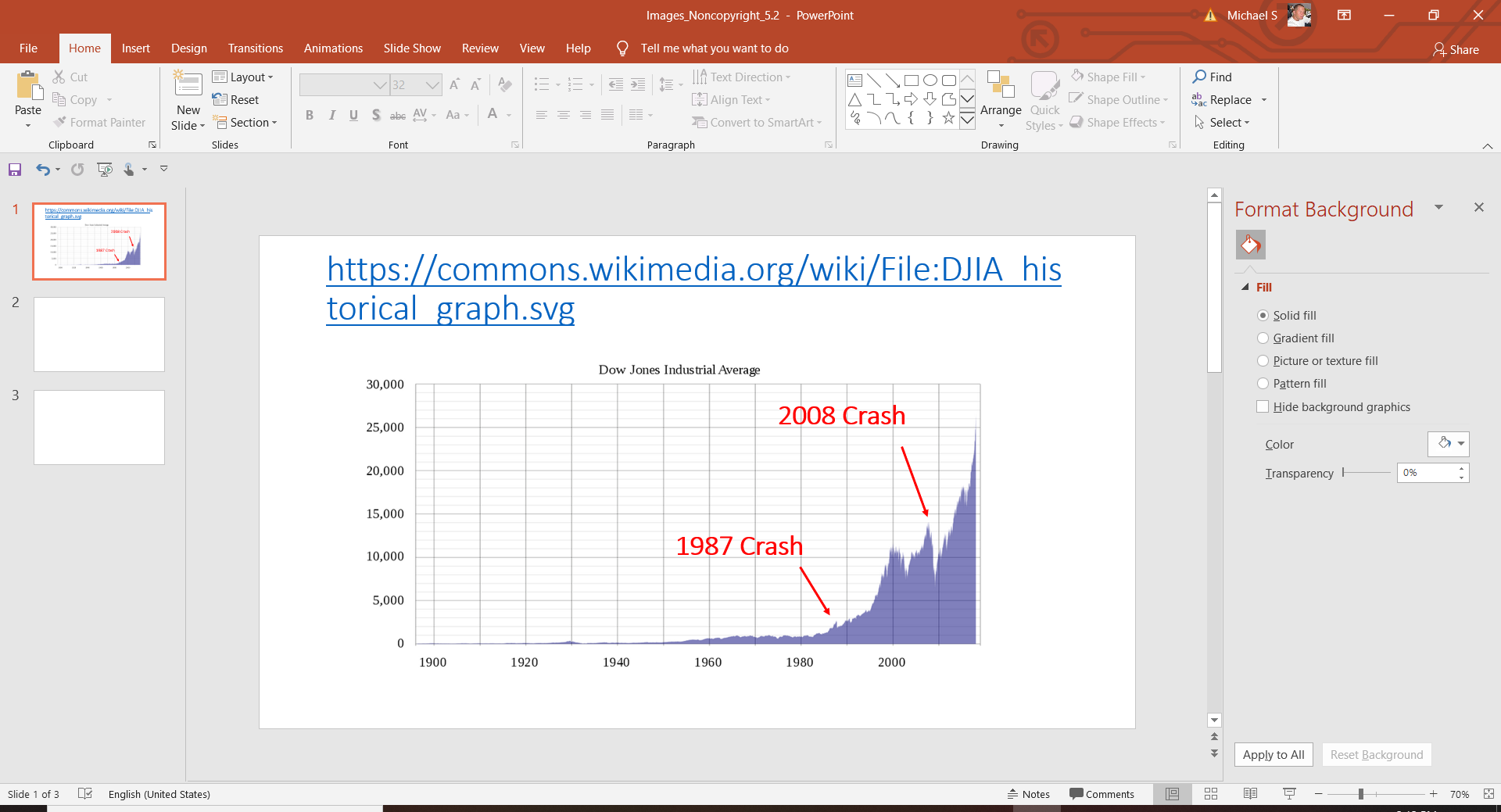
Measurement Bias occurs when there is a problem with the machines or humans doing the measuring or observing. Ensuring that there are multiple measuring devices and humans are properly trained to observe the outputs of devices/models is the best way to control for Measurement Bias.

As data analyst working in the Anti-Money Laundering (AML) compliance department of a large global bank, you must create models that predict transaction activity that poses a risk to the bank and have them independently validated by subject matter experts before they can be implemented. The Investigations department loans you 4 investigators to perform the validation of your latest model’s output. You distribute an equal amount of work items to the investigators that your model generated. After reviewing the aggregate of their scores, you determine they were just high enough to necessitate implementation. One year later you and your team are reviewing the efficiency of your model and unfortunately it has not met expectations. Upon re-analyzing the individual investigator scores you find one investigator, John, had given much higher risk scores to the output than the other three, which threw off your results to the point the aggregate score indicated the model would be more efficient in production. What you failed to account for is the individual bias between different investigators due to their varying levels of experience and training. John was less experienced and more risk averse and it reflected in his scores. His bias stemming from his lack of training and experience caused him to see risk where there wasn’t, which led you to believe your model would be effective.

This is an example of how human bias can corrupt machine learning. We can inadvertently program our own biases into our models and cause them to produce shoddy results that can have far worse affects. To account for measurement bias, insure you design and communicate your testing procedures clearly and be sure to insist your validators have a level of experience and training that better reflects the Investigations Unit as a whole. Review your scoring process and determine if it is lending itself to measurement bias. Perhaps you could have designed your output to require 2 of 3 investigators to agree on the risk level in order to hedge against individual bias. As a data analyst, you must always be prepared to communicate your project’s susceptibility to bias and how you control for bias in your collection, measurements and analysis to stakeholders.

Biased Presentation of Data

Choosing the wrong visual to represent your data and the message you wish to convey can corrupt your presentation with bias. Let’s take a look at the line chart below that shows the DOW Jones Industrial Average historical performance going back to March, 1982.



The purple shaded area represents the DOW index at a given point in time. The stock market crash in 2008 is represented by a clear, steep, long drop. The 1987 crash is just a small blip and if one didn’t know any better, you wouldn’t even know a crash occurred. In terms of showing the historical index score based only on the points, there is nothing wrong with this visual. But if you intended to create a visual that communicated the risks of investing in the stock market to stakeholders, this visual is biased. In terms of percentage drop, the crash in 1987 was much more severe that the above visual indicates. If we were to build a visual based on percentage decline, it would look much different than the one above.

Summary

In this achievement we have covered some of the more common types of bias and demonstrated why controlling for bias is a critical part of an ethical approach to data analysis. Failing to control for bias in how data is collected and analyzed can result in models that fail to drive sound business decision making. At worst they can result machines that learn from biased data and reinforce stereotypes and exclude relevant variables. Everyone has bias because our perspectives of the world around us are limited to our own. As you design and execute your projects from start to finish, remember to be transparent in your processes and invite your colleagues to provide input to best minimize the affects of bias in your work.

Resources

<http://web.mit.edu/simester/Public/Papers/Deceptive_Reviews.pdf>

<https://www.nytimes.com/2018/06/13/smarter-living/trust-negative-product-reviews.html>

<https://link.springer.com/chapter/10.1007/978-3-211-77280-5_4>

<https://towardsdatascience.com/5-types-of-bias-how-to-eliminate-them-in-your-machine-learning-project-75959af9d3a0>

<https://www.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html>

<https://www.floridamuseum.ufl.edu/shark-attacks/yearly-worldwide-summary/>

<https://www.usatoday.com/story/news/nation-now/2017/07/24/shark-week-7-things-way-more-likely-kill-you-than-sharks/506115001/>

<https://www.msn.com/en-us/money/indexdetails/dji-dji/fi-a6qja2>

<https://www.forbes.com/sites/greatspeculations/2020/03/16/market-crashes-compared28-coronavirus-crash-vs-4-historic-market-crashes/#6cf8c363fa45>

Images

<https://commons.wikimedia.org/wiki/File:DJIA_historical_graph.svg>