Training Data and its Inherent Biases

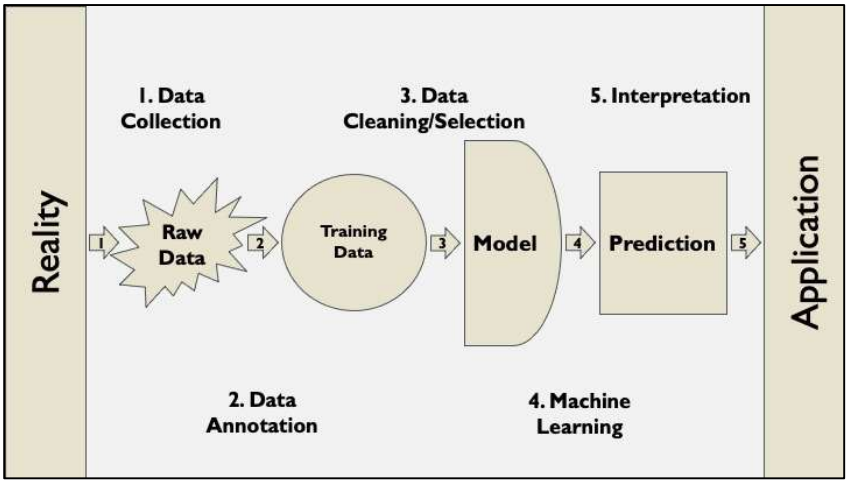
Griffin Holt

Introduction

In any machine learning model, there are five stages of action, each which carry data from one form to another. These five stages include the following: *data collection*, the gathering of data sampled from a real-world situation; *data annotation*, the proper labeling of that data in order to assist the model in its classifications or calculations; *data cleaning*, including variable selection, normalization, stratification, and the handling of incomplete data; the *machine learning algorithm* itself; and the *interpretation* and application of the algorithm's output model. The assumption of most machine learning projects is that this model, to some degree of accuracy, reflects the real world in some definable way.

Figure 1: The 5 Stages of Data Transformation in Machine Learning

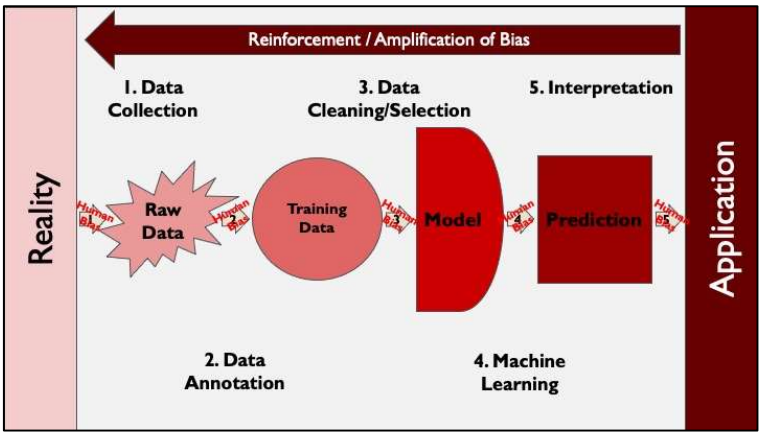
At each one of these five stages, there is potential for the introduction of bias caused by



humans. When it comes to social issues in particular, reality itself is usually already biased–due to human behavior–which means there is bound to be error before the data are even collected. Biases introduced at each stage of building a model will then result in an amplification of social

injustices[1] that exist in reality (Mitchell). In other words, the model can turn into a feedback loop of injustice that systematically discriminates against certain populations and systematically empowers others. In this paper, I will only be focusing on examples of bias in the first three stages–collection, annotation, and cleaning–although I would exhort readers to set aside time on their own to explore the causes and consequences of bias in the other two stages.

Figure 2: The "Snowball" Effect of Introduced Bias



Ethical Ramifications

From a utilitarian perspective, one needs to weigh the benefits of releasing a model constructed from biased training data versus the costs. The *benefits* of releasing such a model can be summarized as follows: even an inaccurate model can help an incredible amount of people. However, the scope, duration, intensity, and probability of the negative effects of the model could outweigh the model's usefulness. Imagine a machine learning model used by a bank to accept or reject loan applications with an accuracy rate of 95%; however, the 5% of applicants that this model incorrectly judges are *all* from a suburb in the middle of Louisiana. The scope of these negative effects could be long-lasting: one rejection by the model could influence the model to reject a second attempt by that same individual. If only 7% of all the applicants are from this suburb, then, over 70% of this suburb's applicants are being rejected. For some of these

[1]In *Figure 2*, this amplification of social injustice is illustrated by the addition of small amounts of red into the data at each stage, resulting in an even darker reality than we started with.

applicants, acquiring a loan could be the difference that feeds their children. However, a utilitarian may–according to the situation–decide that these losses are acceptable upon discovering that 80% of the granted loans are lifting 100,000's of other people out of poverty every year.

Conversely, to the proponent of deontology, it may not matter that 95% of the population is benefitted–treating people *fairly* is more important. If our machine learning models are not *fair*, then they may not be worth the cost of our humanity.

Issues growing out of biases in training data represent a conflict between the prima facie duties of *justice, non-injury*, *veracity,* and *beneficence.* Justice demands that the model be fair towards all. Non-injury demands that the model not harm anyone or any particular group of people. Veracity demands that our machine learning model be *as truthful a representation of the world as possible*. Beneficence demands that we build these models to help rather than to hurt. It is the duty of beneficence that is less clearly for or against the deployment of a biased model, since, as I stated before, even a biased model can do much good.

To summarize, knowing whether to deploy a model is not a clear-cut issue. The Association for Computing Machinery's established Code of Ethics provides slightly clearer boundaries–the code emphasizes virtues, such as *fairness*, and the prima facie duties of *justice* and *non-injury* more than it emphasizes other principles (ACM, 4-12). This suggests that one of our highest priorities as machine learning engineers and scientists should be to ensure that our models do not unfairly perpetuate bias against any one individual or group of people.

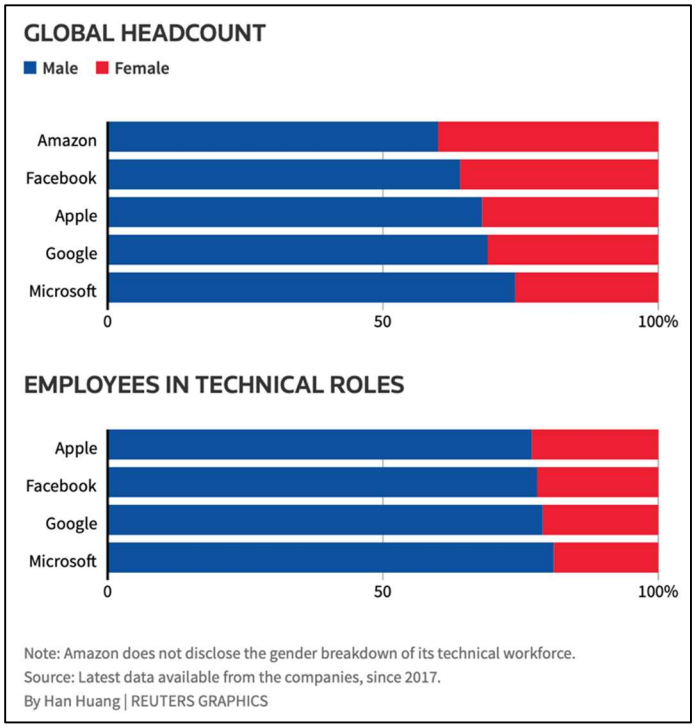
Case Study: Amazon's Recruiting Engine (2014 – 2018)

Starting in 2014, a team of machine learning engineers at Amazon set out to construct an algorithm that could predict the hiring potential of a job applicant. According to one of the scientists, Amazon "literally wanted it to be an engine where I'm going to give you 100 resumes,

it will spit out the top five, and we'll hire those" (Dastin). Unfortunately, the algorithm quickly began to display a bias against female applicants–the algorithm was penalizing phrases such as "women's chess club" and, more so, downgraded applicants that had attended one of two specific all-women's colleges. Even after attempting to remove gender-related terms from the applications, the researchers could not guarantee that the algorithm would not pick up on the applicant's gender through other subtle differences in word choice or experiences. Amazon stated that the tool "was never used by recruiters to evaluate candidates" (Dastin). Insiders familiar with the program claimed that although recruiters could view candidates' rankings, evaluations of candidates were never solely based on these algorithm results.

The researchers at Amazon were never able to eliminate the bias in the recruiting model because the only data they could feed it–the historical success of hired candidates at Amazon–was inherently biased against women. The technology industry is largely dominated by male professionals, as can be seen in the chart at left. Thus, when the machine learning model was trained on this data, the model learned–incorrectly–that maleness must be an indicator of a candidate's potential success because the majority of employees at Amazon are male. This error is an example of both *selection bias* and *association bias*. Selection bias is introduced during data

Figure 3: Proportions of Male and Female Employees at Large Tech Companies



collection when the collected sample does not accurately reflect the population it was collected from; association bias occurs when data fed into a model reinforces a pre-existing cultural bias. To understand the selection bias, we must first define the population and the sample in this modeling schema. If we define the population to be all past successful employees at Amazon, then the sample is representative and the model is accurate; the population of past successful Amazon employees is composed mostly of males, and the model therefore predicts that more males have been successful in the past at Amazon. However, this was not the goal. The researchers were actually trying to predict whether any future applicants would be successful at Amazon; this means that the population they should draw from is "all people who could be successful at Amazon," not "all people who have been successful at Amazon." In this sense, the sample of past Amazon employees is clearly unrepresentative of the larger population: "all people who could be successful at Amazon" most likely contains a more equal proportion of male and female candidates. The association bias is then much easier to identify–the unrepresentative sample data reinforced the historical pattern of male hiring dominance.

Amazon neutralized its responsibility of the error by claiming *denial of victim*: no candidates were hurt because recruiters did not actually use the tool to evaluate candidates (a statement that is difficult to guarantee if the recruiters did indeed have access to the rankings). Amazon's duty to justice, veracity, and non-harm outweighs any beneficence that could have been produced by the recruitment engine. Especially when recruiting candidates, fairness in evaluation towards each candidate is critical for both the success of the company and the welfare of each candidate. Fortunately, after four years of struggling to correct the biases, Amazon ultimately scrapped the project before it could affect too many more people.

Conclusion: Combating Bias in Data

Machine learning engineers and scientists are ethically obligated to combat bias in training data so that their models may be fair, non-harmful, and beneficent to all people. By analyzing the case studies above and reviewing leading research papers and industry standards, including Google's "Responsible AI Practices," I synthesized a few guidelines that will assist engineers and scientists in this effort.

First, seek to expose your own point-of-view to diversity. The more diversity you encounter throughout your life, the more capable you become as a researcher to identify and mitigate the various types of data biases[2] that can occur when creating a model. Second, stay on top of the latest research and modeling techniques. There are great strides being made towards developing mathematical techniques that can assist scientists in identifying and even removing bias from models. Third, when building a model, establish concrete goals and measures for fairness and inclusion before beginning the modelling process; use representative and accurate sample data; think about the edge cases–often, people already in the margins of society can be the ones that are the most marginalized by algorithms; and be willing to scrap your model–like Amazon did–if fairness cannot be achieved. Finally, do what you can to diversify the field of artificial intelligence itself. For example, if the technology industry continues to be dominated by males, then the technology

Table 1: Six Basic Types of Biases in Data

Types of Bias	Explanation
Selection Bias	The collected sample doesn't reflection the population
Measurement Bias	Faulty measurements when collecting / annotating data
Recall Bias	The data is inconsistently annotated
Association Bias	Data fed to the model reinforces a cultural bias
Exclusion Bias	Valuable data thought to be unimportant is deleted in the cleaning process
Observer Bias	Researchers go into a project with subjective thoughts about the study (conscious/unconscious)

[2] *Table 1* gives an overview of the basic types of data biases that can occur in data collection, data annotation, and data cleaning.

industry will continue to suffer from the myopia that accompanies such a lack of diversity. We need the opinions and contributions of a variety of people in order to minimize bias in data and build constructive, accurate machine learning models.

Works Cited

"AI4ALL Home Page." AI4ALL, 2020, ai-4-all.org/.

Arena Analytics. "Machine Learning Removes Bias from Algorithms and the Hiring Process." PR Newswire: News Distribution, Targeting and Monitoring, Cision, 6 Nov. 2020, www.prnewswire.com/news-releases/machine-learning-removes-bias-from-algorithms-and-the-hiring-process-301167669.html.

The Association for Computing Machinery, Inc. ACM Code of Ethics and Professional Conduct, The Association for Computing Machinery, Inc., 2018, www.acm.org/binaries/content/assets/about/acm-code-of-ethics-booklet.pdf.

Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women." Reuters, Thomson Reuters, 10 Oct. 2018, www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

"Fairness: Types of Bias." Machine Learning Crash Course, Google, 2020, developers.google.com/machine-learning/crash-course/fairness/types-of-bias.

Goldfain, Cristina. "Sources of Unintended Bias in Training Data." Towards Data Science, Medium, 19 Aug. 2020, towardsdatascience.com/sources-of-unintended-bias-in-training-data-be5b7f3347d0.

Horev, Rani. "Identifying and Correcting Label Bias in Machine Learning." Towards Data Science, Medium, 9 Feb. 2019, towardsdatascience.com/identifying-and-correcting-label-bias-in-machine-learning-ed177d30349e.

Lim, Hengtee. "7 Types of Data Bias in Machine Learning." Lionbridge AI, 20 July 2020, lionbridge.ai/articles/7-types-of-data-bias-in-machine-learning/.

Mitchell, Margaret. “Bias in the Vision and Language of Artificial Intelligence.” Stanford CS224N: NLP with Deep Learning. 2020, Stanford, California, web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/slides/cs224n-2019-lecture19-bias.pdf.

Mitchell, Margaret. “Margaret Mitchell, Senior Research Scientist.” Margaret Mitchell, 2019, www.m-mitchell.com/.

“Responsible AI Practices.” Google AI, Google, 2020, ai.google/responsibilities/responsible-ai-practices/?category=fairness.

Silberg, Jake, et al. “What Do We Do About the Biases in AI?” Harvard Business Review, Harvard University, 25 Oct. 2019, hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai.

Varshney, Kush R. “Introducing AI Fairness 360, A Step Towards Trusted AI - IBM Research.” IBM Research Blog, IBM, 12 Feb. 2019, www.ibm.com/blogs/research/2018/09/ai-fairness-360/.