Understanding Data Bias

Types and sources of data bias

Prabhakar Krishnamurthy · Follow

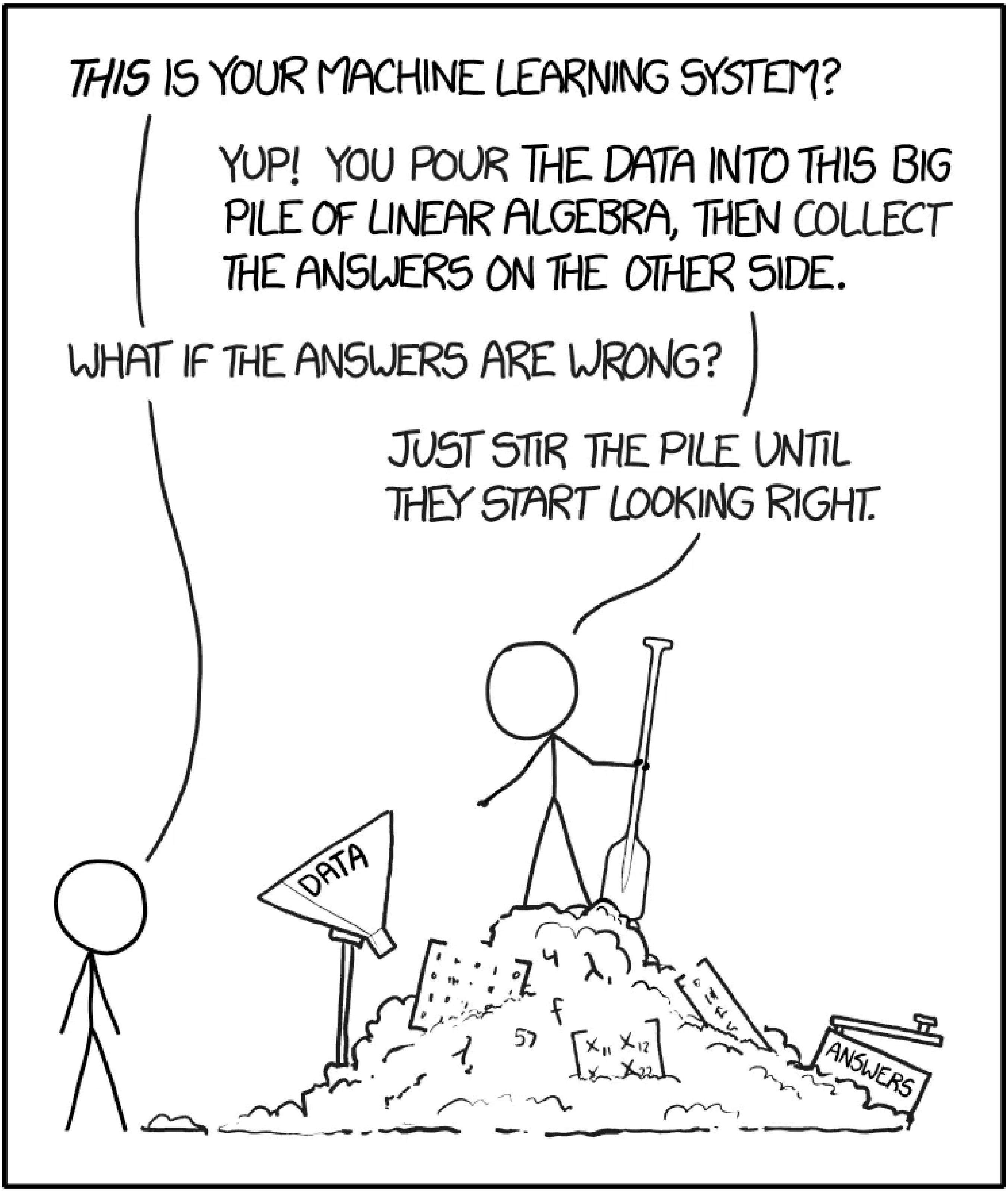
Published in Towards Data Science

13 min read · Sep 12, 2019

Listen Share

The huge success of machine learning (ML) applications in the past decade — in image recognition, recommendation systems, e-commerce and online advertising — has inspired its adoption in domains such as social justice, employment screening, and in smart interactive interfaces such as Siri, Alexa, and the like. Along with the proliferation of these applications, there has been an alarming rise in reports of gender, race and other types of bias in these systems. A widely discussed report in the periodical Propublica claimed serious bias against African Americans in a tool to score criminal defendants for recidivism risk [1]. Amazon shut down a model to score candidates for employment after they realized that it penalized women [2]. Predictive policing systems [3] have come under close scrutiny and their use has been curtailed due to discovered biases. Content personalization systems create filter bubbles [23] and ad ranking systems have been accused of racial and gender profiling [22].

A principal source of these biases is the data used to train the ML models . **The fact is almost all big data sets, generated by systems powered by ML/AI based models, are known to be biased**. However, most ML modelers are not aware of these biases and even if they are, they do not know what to do about it. Few ML publications discuss data — the details of what data was used, how it was generated and what was done to the data before modeling. Instead, authors seem intent on impressing readers with their prowess in constructing complex, esoteric models and blowing them away with the accuracy of their models. The cartoon below by xkcd captures this mindset well.



THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Cartoon by xkcd.com, creative commons license (https://xkcd.com/license.html)

Machine Learning scientists would do well to pay heed to Andrew Gelman's dictum: "The most important aspect of a statistical analysis is not what you do with the data, it's what data you use" [4]. Statistical analysts understand the importance of Exploratory Data Analysis (EDA) as systematized by John Tukey [16] decades ago. However most ML scientists have a different pedigree and do not appear to appreciate the importance of Exploratory Data Analysis.

My goal in this paper is to identify the main sources of data bias focusing mainly on web data and data generated by online systems such as ad serving and content ranking systems.

The takeaway messages of this paper are:

- Most (almost all) big datasets generated by ML powered systems are biased
- Bias in data produces biased models which can be discriminatory and harmful to humans
- A thorough evaluation of the available data and its processing to mitigate biases should be a key step in modeling

So what do we mean by “data bias”? The common definition of data bias is that the available data is not representative of the population or phenomenon of study. But I use it in a broader sense. Bias also denotes:

- Data does not include variables that properly capture the phenomenon we want to predict
- Data includes content produced by humans which may contain bias against groups of people

Based on this definition, except for data generated by carefully designed randomized experiments, most organically produced datasets are biased. Even curated reference datasets are biased as Torralba and Efros discuss in their paper [12] on image recognition models. They considered 12 widely used reference datasets of images. They trained a model on one of the datasets and tested its performance on the other 11. They saw a significant drop in prediction accuracy on these test datasets. Their lesson from this toy experiment is that, despite the best efforts of their creators, these references datasets have a strong built-in bias.

Data bias occurs due to structural characteristics of the systems that produce the data. Based on my analysis, the following are the most common types of data bias:

- Response or Activity Bias: This type of bias occurs in content generated by humans: reviews on Amazon, Twitter tweets, Facebook posts, Wikipedia entries,etc. The fact is only a small proportion of people contribute this content and their opinions and preferences are unlikely to reflect the opinions of the population as a whole.

- Selection bias due to feedback loops: This type of bias occurs when a model itself influences the generation of data that is used to train it. This occurs in systems that rank content such as in content and ad personalization, recommender systems which present or give priority to some items over others. Users' responses (which generates the labels for examples) to items presented are collected, responses to items not presented are unknown. User responses are also influenced by the position of the items on the page and the details of presentation such as font, media (does the item contain images?).
- Bias due to system drift: Drift refers to changes over time to the system generating the data. Changes include the definition of the attributes captured in the data (including the outcome) or the underlying model or algorithm that changes how users interact with the system. Examples are: the addition of new modes of user interaction such as like or share buttons, addition of search assist feature.
- Omitted variable bias: This type of bias occurs in data in which critical attributes that influence the outcome are missing. This typically happens when data generation relies on human input or the process recording the data does not have access to key attributes.
- Societal bias: This type of bias occurs in content produced by humans, whether it be social media content or curated news articles. Examples: the use of gender or race stereotypes. This type of bias can be considered a form of *label bias.*

Let's consider these types in more detail. Awareness of bias is the first step, mitigation is the next step. I do not discuss bias mitigation techniques in detail since techniques for bias mitigation depend on the particular data set and its application.

Response Bias

Response bias is common on the web, most data comes from a few sources. Baeza-Yates [5] provides several examples of bias on the web and its causes. He points out that:

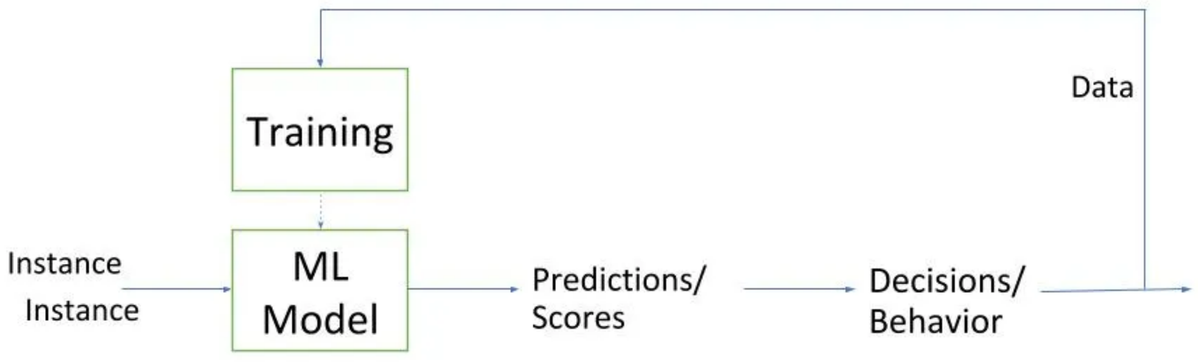
- 7% of users produce 50% of the posts on Facebook.
- 4% of users produce 50% of the reviews on Amazon

- 0.04% of Wikipedia's registered editors (about 2000 people) produced the first version of half the entries of English Wikipedia.

Even more concerning is that a majority of these users represent only a few demographic groups and geographic areas. Clearly, the data cannot be used to make inferences about all users. As one example, Crawford [6] reports that Twitter tweets were used to study people's behavior after Hurricane Sandy struck the US northeast. Hoping to understand the behavior of users in the worst hit areas, the researchers later discovered that most of the data came from Manhattan. Very few tweets came from the more severely affected regions in New York. She points out that over time, power blackouts set in, phone batteries drained even fewer tweets came from the worst hit areas.

Bias due to Feedback loops

Systems for online advertising, content personalization, recommendations, all have built-in feedback loops. These systems embed ML models that influences the data generated, which in turn feeds back into the system as training data for the model. ML models direct user attention to a small subset of items and record user actions on these items (see below for a diagrammatic representation). User preferences are inferred by tracking their views, clicks and scrolls. Selection bias occurs due to the non-random subset of items presented to users. Presenting these items in a ranked list introduces *position bias* — since users scan items from left to right and top down (based on experiments performed on US users). *Presentation bias* is introduced if fonts and media types (image vs text vs. video) vary across items. One impact of these biases is that model evaluation using holdout samples from this data tends to favor models that generated the data [8]. These systems can also change users' long-term content consumption behavior. Cheney et al [8] discuss how recommendation algorithms encourage similar users to interact with the same set of items, therefore homogenizing their behavior, relative to the same platform without recommended content. For example, popularity-based systems represent all users in the same way; this homogenizes all users. Social recommendation systems homogenize connected users or within cliques, and matrix factorization homogenizes users along learned latent factors.



Feedback loop in ad and content personalization systems

Bias Due To System Drift

System drift denotes system changes that change how the user interacts with the system or the nature of the data generated by the system. Examples of drift include:

- The definition of the concept or target being learned could change. In a fraud prediction system the definition of fraud changes. This type of drift is often referred to as “concept drift”.
- The user interaction model changes. For example a content personalization system may add a share or like button under articles. A web search interface may add query suggestions which can change the composition of user queries. This type of drift is often referred to as “model drift”.

In the presence of drift a static model can degrade in performance over time, a regularly trained model may not perform well if data from multiple time periods are used to train the model. Harel et al [10] propose a resampling method to detect concept drift.

An example of system drift that caused a high profile ML application to fail, is Google Flu Trends (GFT). For a few years Google Flu Trends was held up as an innovative use of search data to “forecast” the expected number of flu cases for a season. However in February 2013, Nature reported [18] that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States. This happened despite the fact that GFT was built to predict CDC reports [11].

According to Lazer et al [11], one of the key reasons behind the failure is due to changes Google periodically makes to its search interface. Google introduced "recommended searches", based on similar searches by other users. This increased the relative magnitude of certain searches. Because GFT uses the relative prevalence of search terms in its model, improvements in the search algorithm adversely affected GFT's estimates. GFT baked in an assumption that relative search volume for certain terms is statically related to external events, but search behavior is not just exogenously determined, it is also endogenously cultivated by the service provider.

Omitted Variable Bias

This type of bias typically happens in systems where data is generated by humans manually inputting the data or in online systems, where certain events or actions are not recorded due to privacy concerns or lack of access. This implies that a key predictor variable may not be available to include in the model.

Two conditions must hold true for omitted-variable bias to exist in regression models:

- The omitted variable must be correlated with the ***dependent (target)*** variable.
- The omitted variable must be correlated with ***one or more other explanatory*** (or ***predictor***) variables.

Here's an example of a scenario where omitted variable bias can occur: A laptop manufacturer has an online chat system which its customers can use for support requests or to ask questions. The manufacturer wants to use the opportunity to cross-sell products and has developed a model to score users on how likely they are to buy additional products. The score is intended help agents working the chat system to allocate their time efficiently. When they are busy agents put in more effort (and time) trying to cross-sell to users with high-scores and less effort on users with lower scores. However, the time (and effort) expended by the agents is not recorded. Without this data it will appear that the scoring model is performing very well, whereas the time spent by agents might be much better explanation for user purchase decisions.

Caruana et al [14] advocate the use of intelligible models for high-risk applications such as healthcare to identify bias due to omitted variables. Modeling approaches such as neural networks, random forests, boosted trees, etc often produce more accurate models than approaches such as logistic regression and decision trees, but the latter are intelligible and therefore are more trustworthy. They discuss a case where the goal was to predict the probability of death (POD) for patients with pneumonia so that high-risk patients could be admitted to the hospital while low-risk patients are treated as outpatients. Their ML model learnt a pattern in the data where pneumonia patients who were asthmatic had lower risk of dying from pneumonia compared to patients who were not asthmatic. Realizing that this seemed anomalous, the researchers investigated further and learned that patients with a history of asthma who exhibited symptoms of pneumonia usually were admitted not only to the hospital but directly to the ICU (Intensive Care Unit). The aggressive care received by asthmatic pneumonia patients was so effective that it lowered their risk of dying from pneumonia compared to the non-asthmatic patients. However, the fact that these patients received aggressive care in the ICU was not recorded in the data. The authors point out that they would not have learned about the anomalous pattern in unintelligible models.

Societal bias

Human generated content on the web and in social media abound in biases. Two high profile cases will serve to illustrate this point. Bolukbasi et al [15] show that word embeddings trained on even Google News articles exhibit female/male gender stereotypes. For example: females were associated with professions of nurse and nanny whereas males were associated with professions of doctor and financier. They propose debiasing strategies to mitigate these biases. In another case [2], Amazon tried to build an AI tool to screen candidates until management discovered that it had learned to penalize women candidates. The problem is that in most companies today, technical roles are filled by men and this bias creeps into any models that use current employee data to train models. Unless detected societal biases in data can lead to algorithmic that discriminate on gender, race and other categories.

How to identify and correct for bias

The first key step in identifying bias is to understand how the data was generated. As I have discussed above, once the data generation process has been mapped the types of bias can be anticipated and one can design interventions to either pre-process data or obtain additional data. Another key step is to perform comprehensive Exploratory Data Analysis (EDA) [16]. EDA techniques are discussed in several textbooks and papers.

Identifying and dealing with societal bias requires special techniques, some of which I will reference below. For bias due to feedback loops, one approach would be to design the system to randomize on a small sample of queries. For instance, on a small fraction of requests (say 0.1%) the items are selected and presented in random fashion. Only data for these requests are used in model training and evaluation. Sometimes, this is not feasible due to user experience concerns. In these cases, propensity weighting techniques proposed in [9] is one possible approach.

For biases in human generated content, there has been a lot of research recently on quantifying discrimination and also in debiasing techniques to improve fairness. Recently, benchmarks quantifying discrimination [20] and even datasets designed to evaluate the fairness of these algorithms [21] have emerged. The challenge of course is to improve fairness without sacrificing performance. However, these techniques are typically dataset and application specific. Generally, these techniques fall into one of three categories:

- those that use data pre-processing before training
- in-processing during training
- post-processing after training

However, the problem of severely imbalanced training datasets and the question of how to integrate debiasing capabilities into AI algorithms still remain largely unsolved [19]. One recently published research [19] uses post-processing after training and have attempted the integration of debiasing capabilities directly into a model training process that adapts automatically and without supervision to the shortcomings of the training data. Their approach features an end-to-end deep learning algorithm that simultaneously learns the desired task as well as the underlying latent structure of the training data. Learning the latent distributions in an unsupervised manner enables to uncover hidden or implicit biases within the training data. Bolukbasi et al [15] describe a post-processing approach to de-bias word embeddings to mitigate gender bias.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. “Machine bias: There’s software used across the country to predict future criminals, and it’s biased against blacks”. ProPublica (May 23, 2016).

[2] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[3] Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice, *New York University Law Review Online, Forthcoming (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423*)

[4] https://statmodeling.stat.columbia.edu/2018/08/07/important-aspect-statistical-analysis-not-data-data-use-survey-adjustment-edition/

[5] Ricardo Baeza-Yates, “Bias on the Web”. Communications of the ACM, June 2018, Vol. 61 №6, Pages 54–61.

[6] https://hbr.org/2013/04/the-hidden-biases-in-big-data

[7] Olteanu, Alexandra and Castillo, Carlos and Diaz, Fernando and Kiciman, Emre, Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries (December 20, 2016). Frontiers in Big Data 2:13. doi: 10.3389/fdata.2019.00013. Available at SSRN: https://ssrn.com/abstract=2886526or http://dx.doi.org/10.2139/ssrn.2886526

[8] Cheney, Allison J.B., Brandon M. Stewart, and Barbara E. Engelhardt, “How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility,” Proceedings of the 12th ACM Conference on Recommender Systems, Pages 224–232

[9] Schnabel, Tobias., Adith Swaminathan, Ashudeep Singh, Navin Chandak, Thorsten Joachims, “Recommendations as treatments: Debiasing learning and evaluation”, arXiv preprint arXiv:1602.05352.

[10] http://proceedings.mlr.press/v32/harel14.pdf

[11] David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani, “The Parable of Google Flu: Traps in Big Data Analysis,” SCIENCE VOL 343, 14 MARCH 2014

[12] A. Torralba, A. Efros. Unbiased Look at Dataset Bias. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[13] https://5harad.com/papers/included-variable-bias.pdf

[14] Rich Caruana, Paul Koch, Yin Lou, Marc Sturm, Johannes Gehrke, Noemie Elhadad., "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission", *KDD'15, August 10–13, 2015, Sydney, NSW, Australia* | August 2015

[15] (https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf

[16] John Tukey, Exploratory Data Analysis, Pearson Modern Classics.

[17] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard and Ed Snelson: Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising, Journal of Machine Learning Research, 14(Nov):3207–3260, 2013.

[18] Declan Butler, "When Google got Flu wrong," Nature, Vol 484, Issue 7436, Feb 2013.

[19] Alexander Amini, Ava Soleimany, Wilko Schwarting, Sangeeta Bhatia and Daniela Rus., "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure", AAAI/ACM Conference on Artificial Intelligence Ethics and Society, 2019.

[20] M Hardt, E Price, N Srebro, "Equality of opportunity in supervised learning", Advances in neural information processing systems, 2016

[21] Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency, 77–91.

[22] Sweeney, Latanya, Discrimination in Online Ad Delivery (January 28, 2013). Available at SSRN: https://ssrn.com/abstract=2208240 or http://dx.doi.org/10.2139/ssrn.2208240

[23] Eli Pariser, The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think, Penguin Books, 2012.