FIELD GUIDE
to Address Bias in Datasets

Inspired by the Omidyar Network

Govind Nagubandi
Penn Law Policy Lab on AI and Bias

April 2021

Field Guide to Mitigating Bias in Machine Learning

The proliferation of artificial intelligence and machine learning models has reached near commoditization. There are minimal barriers for a person, scientist or a layperson, to construct a machine learning model and deploy it directly to consumers. Combining cheap or nearly costless cloud computing tools and APIs (AWS, GCP, Azure, and many others), with open-source packages, and YouTube tutorials, a person can quickly construct an app that scores individuals, ranks customers for offers or engagements, or prioritizes service. The flexibility of the cloud allows individuals to easily share and their models and quickly expand usage and influence. In many instances, cloud providers allow users to deploy pre-trained models built from various but unknown data or offers free datasets to jumpstart model training. While these are terrific ways to quickly build and test components, they may not be the best tools to use due to the threat of bias held in the offered data. Every data science practitioner should be well versed in the threat of data bias as well as armed with practical methods to identify bias. This field guide shares a framework of where to look for bias and will hopefully help users build a robust skepticism of data before blindly training models. It can be used by students, existing practitioners, business users, and others as a reference for where to look for bias.

Our goal is not to determine whether or not bias exists in a dataset; but to understand where it may exist and how it might affect your outcome in desirable and undesirable ways. Looking from another angle, before deploying a model, we want to understand *intended* outcomes versus the *unintended* consequences. For example, a model might help recruiters find qualified candidates, but it might fall short if it only recommends women for nursing positions. The intended outcome of finding candidates comes with unintended consequences of all female candidates. One level deeper, we might ask questions like, “what should the gender ratio be for candidates?”. This might be an unknowable question, but we address it briefly in the framework as a reference distribution. Although this is a simple example, it speaks to the competing forces at work in machine learning models.

It is also important to note that our goal is to find and measure bias, not to remove it from our data. Identifying existing relationships within our data, including outliers, is arguably the goal of all machine learning, where a model enables inference and prediction. Looking from another angle, regular society is biased and holds many inequalities; if we were to remove or reduce those inherent differences, we would be the ones introducing a new type of bias that does not reflect ‘ground truth’! We come back to our goal, as a practitioner, to identify, measure, and understand what biases might exist in our dataset. We humbly reserve the corresponding question “what really is fair?” to a conversation with our philosopher and legal scholar colleagues. However, we do think there is a seat at that table for data scientists!

There are many frameworks on the data science process. Most of these explain the full process of interrogating data, training models, and rejecting overfit models. However, many of these do not include a clear focus on identifying bias. The purpose of this guide is to supplement those frameworks with an in depth look at bias. We walk through different methods to identify bias, suggest questions to ask of data, and consider a deeper look at certain aspects of a dataset.

Table of Contents

Data Assessment

An initial data assessment begins with an understanding of data, which, in addition to metadata, should include an understanding of the attributes of data collection. In other words, a practitioner should have a good sense of why the dataset might exist, what is include/ excluded in the data, how long the data has been in existence, etc. These sets of qualitative and quantitative measurements will allow you to make a determination on the quality of data and if it should even be used for inference or prediction at all. Often these questions fall under a general category of "is the data of machine learning quality?", but we recommend carving out specific questions related to bias. Below we break the questions out by qualitative and quantitative assessments.

Qualitative Assessments of Bias

A practitioner should determine the usefulness of data by seeking answers to high level questions about the dataset itself, including metadata, provenance, collection method, and any transformations made to raw data. Here is a sample framework originally published as an NIH study [i] of clinical trials that look at types of bias entering at different stages. We encourage you to research the different types of bias and their effects. Not all types of bias will be present in every dataset, but it is important to measure the effect, if any.

Study or Collection Stage	**Type of Bias**
Trial planning and pre-trial	Flawed study design, Selection bias, Channeling bias
Trial implementation or during collection	Interview bias, Chronology bias, Recall bias, Transfer bias, Misclassification of exposure or outcome, Performance bias
Data Analysis, Publication, or After trial	Citation bias, Confounding

We may not consider these in modeling efforts, but, for example, knowing that Recall bias can enter a dataset during collection is an important theme. We also recommend expanding the framework to include questions specific to modeling. The proposed questions may not be related to an identified and studies bias, but contribute to understanding where a dataset can be over or underrepresented.

Questions related to data items:

- Which data items are included as original measurements, and which are derived?
  - If derived, are there any filters, assumptions, or aggregations used?
  - For example, marking rows 'incomplete' if absent email contact information might make sense for certain processes, but would exclude rows we would otherwise want for modeling
- What is the granularity of the data in time, identifiers, transaction, etc.?
- How much history is available? Is history based on the data collecting scheme or is history available in other places
  - For example, social media patterns may exist before collection began (or before the app or the internet existed). For example, the Billboard music charts existed

long before music streaming. What does it mean if current top charts and streaming ranking lists are uncorrelated?
    - Are there available reference points or studies done prior to- or adjacent to- the data collection?
- What do we know about the full universe of data this dataset is representing?
    - What is the full population size? Are there any population metrics that can be trusted?
    - Consider two example datasets which may require different statistical treatments, sample estimates, and confidence around predictions. The former is easy to apply statistics. The latter, you may even begin with the question "if I have enough people, is it even really a sample anymore?"
        - i) A sample of car dealerships in a region of the USA and number of cars sold each month
        - ii) A panel of credit card transactions in the past year
- Are there any known relationships to individuals, specific objects/items, locations, or other distinguishable features in the dataset?
    - If yes, do we have a similar assessment of the other related data, before we fold it in?
    - **In the author's experience, this is a prominent area for bias to enter datasets, when you are fatigued by assessing a single dataset and blindly accept 'more data is better'. Models are easily able to consume more and granular datasets without question, but practitioners should not!**

Qualitative questions related to data collection:

- How long has this type of data been collected? How old is the newest data? Is this a new mechanism or does it have a long history of existence? Is it complicated or tiring to submit data?
- How long do we think this data source will stay active in the future? Will it change in anyway, including adding more features, questions, or data?
- How long do we think participants will continue to use the data collection mechanism?
- How is the data collected, manually or automatically, passively or actively, digitally or through a model, opt-in or opt-out, is it mandated or volunteer? Is this method harder to use than other methods?
- Who has access to the collection mechanism? Is it a paywall or behind a paywall? Does it require a credit card, a car, a home, a bank account? Who would like to contribute but is excluded?
    - For example, if data is collected from app usage, even with large amounts of users, what is the geographic diversity? 100,000 users who all live in NY, LA, SF, is not very representative!
- Do participants know they are being monitored and collected? Who would like to contribute but can't?
- What does this source look like in the larger context of industry? How representative is the data collected?

Not all the questions might apply, and the list is not exhaustive. However, it should give you a general sense of how to ask questions and where to look for bias entering the dataset.

Case Study

We apply the framework of additional questions to a few *real* datasets. It is important to remember that the framework can be applied to any or all datasets, and answers don't necessarily exclude the dataset from usage.

Web History as an Indicator for Credit Worthiness

In the mid 2010s, a credit card company collects ancillary information during an online credit card application. Information collected includes browser type (e.g. Chrome, Safari, IE), computer type (Mac, PC), device type (computer, tablet, mobile), last page visited or referral page (i.e. google ad, search, or specific website), operating system version (newest or outdated), etc. Using all this data the company discovered individuals who used Macs, came through an ad on NYTimes.com, and had the latest operating system installed, were the most credit worthy. The company used this data as factors to approve/deny applications. However, after a few years, they found that these indicators were no longer as useful in predicting non-delinquent, high credit applications. One of the reasons cited was the success Apple had in selling Macs, and the standardization of web browser.

This is a dataset that may not have existed prior to 2010, also had a shelf life of only a few years.

Geographic Clues

A dataset of mobile phone users collects location information every time a user opens a specific set of apps. The data provider runs an ad network that developers can integrate in their apps. Data collection includes approximate location, time of day, category of app, session length of app, in-app payment flags, and a list of other known apps installed. For a specific set of users, the app also approximates home location to a census zone. More granular location data is available when a user connect to WiFi, and the data provider can determine, with high probability, the home location of the user. The data provider will not expose this information, but they do use it to partner with a video-on-demand set top box provider to create a richer profile of TV viewing habits. TV viewing habits are sold as an add-on package. Data is sold in aggregates of ~50 users, so no individual can be identified.

This is an exclusive dataset requiring a smart phone and apps. Though it is not required, it is suggested that the cohort that includes TV viewing will be the most interrogated. This subset is more exclusive, requiring an identifiable home (and likely not an apartment, though barometer readings can measure altitude), and a video on demand subscription.

Suggesting Reviews

An online platform allowed individuals to rate their employer anonymously, leave feedback on the CEO, and comment on different aspects of employment including overall direction, benefits, and employee morale. The reviews had individuals volunteer their current employment status, job title, and tenure. Online review systems are known to have Selection bias, where more people with negative experience are likely to leave a review out of frustration. However, the online platform is popular in the United States and companies grew sensitive to reviews and feared they negatively impact recruitment. While the online platform remained neutral, the reviewed companies sought ways to 'game the system' within the bounds of the rules. They suggested current employees with high morale write reviews, such as those who were recently promoted, those who live nearby the office (studies show this has a tendency of higher satisfaction), and new employees. The reviewed companies effectively boosted their scores and had better reviews on display!

This is an example of a highly skewed dataset that is available for public viewing. Companies 'in the know' have vastly different populations writing reviews than companies who don't actively manage. Reviews are inherently bias, but perhaps this dataset combines a few other biases as well! The important part is knowing when this data should be relied upon in a training set; perhaps predicting company outlook based on new reviews is not a great use case.

Quantitative Assessment of Bias

After a robust qualitative assessment of the dataset, a practitioner should continue with a quantitative interrogation of data. This is the most common starting area when building a model, so we will focus on assessment of bias. It is important to remember the eventual goal is not to identify and remove outliers, nor to clean and normalize the data, but to better understand the limitations of the data.

Basic Parameter Bias Measurements

- Histograms and Skewed Distributions – histograms and frequency distributions hold more information than simple statistics. It is easy to automate and identify skewness either by limits or scanning charts. Be sure to include a plot of transformed non-normal data! For example, plot the log or square root of a non-normal series
- Variable Correlation – measure the level of correlation between variables to quickly identify which variables might represent biased variables. Additionally, measuring correlation or variance against known bias (e.g., gender, race, ethnicity, zip code) can reveal which other variables carry this information
- Serial Correlation – if you are measuring a time series, run a serial correlation (or correlation versus prior values). This will help you understand normal values and thresholds; for example, running serial correlation on credit card spends and plotting by gender, might reveal hidden information in your data
- Parameter Variance – simply measuring the variance of a parameter can help us determine if it is worth including in a model; frequently we remove variables with zero or near-zero variance; we recommend splitting the dataset by known biased variables (e.g. gender) to understand if there is near-zero variance for a specific category

Quantitatively Analyzing Protected Classes

Under United States law there are a set of classes that are protected from discrimination. In simple terms, this means that individuals are not allowed to be treated differently based on any aspect of these classes, termed disparate treatment. The federally protected classes include Race, Religion, National Origin, Age, Sex, Pregnancy, Familial Status, Disability, Veteran Status, and Genetic Information. When constructing a model, it might not be simple enough to remove these variables from the model because they could be reconstructed from other variables. In such case, the simple absence of the variables would not preclude disparate treatment. We highlight a few methods to determine if a model may be improperly impacted

- ANOVA/Tukey – a Tukey procedure will help identify factors (of a categorical variable) that have significantly different means. For example, measuring the mean or distribution against the variable 'gender'. The output of this test can show side-by-side comparisons of test outputs against different categories of protected classes.
- Proxy test for Hidden Bias – as the name implies, it is difficult to detect all types of bias including those in categories not explicitly labeled. Researchers have created frameworks to try to detect "hidden" bias. Two suggested patterns
    - Pattern 1: Holdout variables
        - Remove or holdout all variables known to hold bias
        - Create the model with regular methods
        - Bring back the held-out variables and run a set of tests (histograms, correlation, means) to check for differences in outcomes based on held-out data
        - Measure how aligned outcomes are to the held-out categories
    - Pattern 2: Randomly re-assign biased variables
        - Remove or holdout all variables known to hold bias
        - Create a new dummy variable which randomly re-assigns the held-out variable --- i.e. the dummy variable should not match exactly the held-out variable
        - Create the model with regular methods
        - Evaluate the influence of the dummy variable on outcome
        - This method also works well when you are comparing multiple models and already setup champion-challenger tests

One Level Deeper

Understanding Data

We covered some qualitative and quantitative methods to interrogate a dataset. This includes questions to research on why the dataset was created and what it might hold. It is important for a data scientist to understand that bias might already be in data that is assumed to be "clean". Sometimes, it is important to investigate 'one level deeper' and question accepted practice. Here, we delve into two common data items.

- Zip Codes – zip codes are useful data since they are mutually exclusive and collectively exhaustive. They can also link other data like demographics, education, home values, and many others. However, it is commonly known that zip code, or geographic features in general, effectively capture prior behaviors of segregation. In addition, zip codes are also, in this author's opinion, faulty, since they are not normalized. They represent

different areas by size, different populations by number and percentage, and they change (the digit refer to post office locations)! Census lines are usually better features, and, for experts, creating your own polygons based on an important dimension is recommended.
- Designated Marketing Areas (DMA) – geographic marketing areas have been used for decades to direct television and radio advertising as broadcast markets. Although the total amount of advertising spending has reduced in recent years, it is important to know that DMAs are actually created by the Nielsen company to measure TV ratings. There is a published process of how counties might move into/out of a DMA. Although Nielsen creates DMAs, they started out by closely matching the Television Marketing Areas (TMAs) created by the Federal Communications Commission (FCC). Since the DMA was deemed to be a "better measure", as Nielsen also measures online video, the FCC actually defaults to Nielsen's DMAs. A good practitioner might wonder if 'better' still really means they are any good. We leave further investigation up to the practitioner!

Models

Using open-source models is incredibly useful. However, a practitioner should make sure they know what an open-source package of model is actually doing. Many times, packages add convenience factors, which provide for quick initial output, but hidden assumptions. Here are some examples highlighting areas to look in different open-source packages:

- Package defaults – providing default parameters for packages is convenient, but practitioners should be careful to understand where and what defaults are set. For example, in the popular R 'caret' package, you can switch between many types of models quickly (i.e. glmnet vs. gbm), but if you do not carefully switch the tuning grid, you may not be comparing nearly the same amount of trials! This can lead to non-optimal models since the default tuning grid changes depending on model type!
- Differences in packages – expanding on the default values in packages, it is also important to be aware of differences in how packages handle similar calculations and functions. In a simple python example, a pandas dataframe is easily transmuted to-from a numpy array (or ndarray). However, when calculating single parameter metrics like standard deviation, there a small but meaningful difference in the default of each package. Pandas defaults to the unbiased estimator using (N-1) and numpy defaults to the unbiased estimator (N). When you are filtering, merging, and sampling data, you might accidentally compare different calculations!
- Default package behaviors – many packages define default behavior to quickly and conveniently produce initial results and outputs. In the author's opinion, this is meant to quickly produce a *directional* output meant for continual iteration. However, without examining these conveniences closely, we can make costly mistakes:
    - XGBoost – extreme gradient boosting has been a very successful technique in machine learning prediction, and it has even been the winning method for Kaggle competitions. XGBoost is a popular cross language framework, and the current packages provide a lot of convenience functions to get a model running quickly. However, some of the convenience allow you to input any type of data

format, namely either a dense or sparse matrix. XGBoost, and trees in general, are very good at handling missing data, however, the algorithm is affected by type of data. In other words, if you use the same dataset in a sparse or dense formation, you likely will get different model results! An implication for handling biased data is that transforming categorical variables into binary columns (i.e. one hot encoding) is not always optimal; but transforming data with multiple categories, may be useful for training. An example would be if a practitioner transforms a variable 'Gender', which holds two values 'Male' and 'Female'. Theoretically, a transformed binary variable Male [0,1] and Female [0,1], will hold the same information since the options are mutually exclusive. However, transforming all categorial variables to binary will affect how the tree is built. It's also useful to know that XGBoost defaults error measurements to squared error but changes the error model based on the input data shape. This should not be a surprise as model evaluation methods should consider model choice; but in the above example, a practitioner might not know they are choosing the model with their dataset shape.

- CEM – Coarsened Exact Matching[ii] is a technique for non-parametric pre-processing of data. It can be used to create balanced experiment groups based on multiple variables and is useful to estimate causal effects. The R package, CEM, employs a robust process to create groups, filter matching sets, and estimating causal effects. *Note: The author highly recommends the practitioner review at the work by Prof. Gary King[iii].* If we follow along the vignette for CEM, it suggests running a matching process, a filtering process, and an estimation process. However, if you follow the guide and steps, you might end up creating estimates (linear regression) from sets of too few observations. You might ask, "how might this happen and why is it not described in the package?". The answer is that it *is described* in the documentation, but it is almost off-hand with a comment suggesting to only use filtering if you 'have a large dataset'. It doesn't tell you where or how to check your sizes. In this author's experience with the package, the filtering method (k2k) will reduce the universe of matching sets to create strata of equal sizes, however, it does not put a minimum limit on the number of observations required for a stratum pass the filter. Using the filtered set of strata, you might accidentally build a set of linear models on sets of only a few data points! The point of this section is two-fold first, to share the CEM package and the terrific work by Gary King, and second, to surface instances even trusted processes need to be checked for your own use case.

Other Items to Consider

The term 'data science' is relatively new, but the study of data and the application of insights have a long and interdisciplinary history. There are many topics with historical perspective, similar to zip-codes, that are worth exploring and further worth questioning. In many cases, assumptions that worked in the past, may not be the best to keep in the present.

- Student’s T-test – we can dedicate an entire guide to the history of statistics and improper uses of statistics, but we focus only on the t-test. The t-test was developed to measure quality of manufacturing on small scale systems (brewing beer!). The t-test effectively tracks how sample sizes effect statistical significance. The test itself tries to measure how far away a sample mean is from the population mean. Perhaps an inherent assumption needed is that we a) we have information on the entire population, not just our sample, and b) we have trustworthy information about the same population. The t-test, to be performed correctly, also requires random sampling, large samples, a normal data distribution, and equal variance. Often times, people use the t-test to compare unlike, non-random, non-normal samples. The idea here is not to be critical of people who are using statistics improperly, we applaud their consideration of statistics, but to help them understand that proper use of statistics requires exact requirements.
- Reference Distributions – expanding a little further on statistics, it is important to understand the full distribution of a population, or the reference distribution. A reference distribution can be created from empirical data, but the point here is to make sure to measure changes in the full distribution instead of just a single parameter (e.g., mean). It can be tough to measure total impact on the distribution, but interventions tend to affect different sub-populations in different ways.
- Representation and Census/ACS – it is difficult or impossible to measure the change or effect on every member of a population, so we often use representative samples. One good reference point is comparing your population demographics to overall country demographics measured in the Census or American Community Survey (ACS). Both the Census and ACS have their own flaws, but they are often the best references. Comparing a representative sample to the overall population, and then to the Census, is good practice.
- Order of Operations in Model Building – expanding on representative datasets, it is also important to consider the order of operations when building a model. For example, marketing surveys typically ask for demographic information, but this data needs to be withheld before an analysis to find insights. If you don’t withhold demographics, the analysis might find significant relationships based on demographic characteristics *before* survey results. In other words, the modeling will weigh demographics too heavily and suggest there are different demographics in the sample! A better way might be to with-hold demographics and then add them back to the created model. This is one way to conduct a data audit.

Existing Toolkits

The fact of bias in machine learning is not novel, and many other data scientists have researched the issue. Many large corporations have devoted resources to help educate the field and also to help practitioners. A simple search for ‘bias in machine learning’ will produce many references to great work. Here are a few toolkits that might also be useful guides

- FairLearn – a python package/notebook developed my Microsoft that will assess model ‘fairness’. It relies upon Reduction-based algorithms to measure which groups may be negatively impacted in a model. https://fairlearn.org/
- LiFT – LinkedIn released a Fairness Toolkit for Scale/Spark to measure fairness in large scale machine learning models. The framework can measure biases in training data and score fairness metrics from models. https://github.com/linkedin/LiFT
- IBM AI Fairness 360 – a toolkit in python and R to help you examine, report, and mitigate bias in data and models. The toolkit packages a set of functions and examples to mitigate bias in datasets. http://aif360.mybluemix.net/

[i] Pannucci CJ, Wilkins EG. Identifying and avoiding bias in research. *Plast Reconstr Surg*. 2010;126(2):619-625. doi:10.1097/PRS.0b013e3181de24bc

[ii] CEM: Coarsened Exact Matching Software, https://gking.harvard.edu/cem

[iii] Quantitative Social Science Research by Gary King https://gking.harvard.edu/