Künstliche Intelligenz

Herausgegeben von
dem Bundesministerium für Umwelt,
Naturschutz, nukleare Sicherheit und
Verbraucherschutz
und FRAUKE ROSTALSKI

Mohr Siebeck

Künstliche Intelligenz

Künstliche Intelligenz

Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa?

herausgegeben von

dem Bundesministerium
für Umwelt, Naturschutz, nukleare Sicherheit
und Verbraucherschutz

und

Frauke Rostalski

Mohr Siebeck

Frauke Rostalski, geboren 1985; Studium der Rechtswissenschaften in Marburg; 2011 Promotion Rechtswissenschaften; 2017 Promotion Philosophie; seit August 2018 Inhaberin des Lehrstuhls für Strafrecht, Strafprozessrecht, Rechtsphilosophie und Rechtsvergleichung an der Universität zu Köln.
orcid.org/0000-0002-5606-3639

Veröffentlicht mit Unterstützung des Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz.

ISBN 978-3-16-161298-5 / eISBN 978-3-16-161299-2
DOI 10.1628/978-3-16-161299-2

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über *http://dnb.dnb.de* abrufbar.

Das Buch wurde von Laupp & Göbel in Gomaringen gesetzt, auf alterungsbeständiges Werkdruckpapier gedruckt und von der Buchbinderei Spinner in Ottersweier gebunden.

Printed in Germany.

Inhaltsverzeichnis

What is "Good" Training Data?

Data Quality Dimensions that Matter for Machine Learning

Felix Neutatz / Ziawasch Abedjan

I. Introduction

Machine learning (ML) is becoming prevalent in almost all areas of our everyday life and enables artificial intelligence (AI) technologies that affect humans significantly with applications in personalized medicine,[1] automated credit rating,[2] and justice systems,[3] to name a few. As a result, it is vital to make sure that decisions and results that originate from ML are of high quality and explainable. One of the imminent problems in current AI systems is that they amplify societal bias, which harms minorities and other protected groups disproportionately. At this point, it is necessary to keep in mind that humans are not only at the receiving end of ML systems but also influence these systems at various major stages of their development and production life cycle. At each of these stages, there is a potential to spill over societal bias into the ML model. First, the data that is used to train these systems can originate from humans. For instance, Microsoft presented a chatbot that continuously learned from the interaction with its users. Some adversaries fed the bot with offensive content that in turn changed the behavior of the chatbot to be racist and sexist.[4] Second, humans develop and configure the embedded ML production pipelines. Technical choices of the developer can introduce technical bias into the entire process from data preparation to model creation.[5] Third, humans analyze and interpret the data and the ML model predictions. After the model returns the predictions, humans still have to decide how to design the thresholds

---

[1] *Emmert-Streib, F. / Dehmer, M.*, A machine learning perspective on personalized medicine: An automized, comprehensive knowledge base with ontology for pattern recognition, Mach. Learn. Knowl. Extr., vol. 1, no. 1, pp. 149–156, 2019.

[2] *Bono, T. / Croxson, K. / Giles, A.*, Algorithmic fairness in credit scoring, Oxford Review of Economic Policy, vol. 37, no. 3, pp. 585–617, 2021.

[3] *Mayson, S. G.*, Bias in, bias out, Yale LJ, vol. 128, p. 2218, 2018.

[4] *Lee, P.*, Learning from Tays introduction, 2016 [online]. Available: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

[5] *Schelter, S. / He, Y. / Khilnani, J. / Stoyanovich, J.*, FairPrep: promoting data to a first-class citizen in studies on fairness-enhancing interventions, EDBT, 2020, pp. 395–398; *Donini, M. / Oneto, L. / Ben-David, S. / Shawe-Taylor, J. / Pontil, M.*, Empirical risk minimization under fairness constraints, NeurIPS, 2018, pp. 2796–2806.

for different decisions and how to rate different types of mispredictions.[6] False negatives and false positives have different implications and societal cost. For example, detecting a cancer diagnosis by mistake (false positive) has a different impact than missing a cancer case (false negative). Humans have to understand these trade-offs and configure the model and the resulting decision support system, accordingly. One of the main pillars of responsible data management is to ensure good training data.[7] It is widely understood that "good" training data yields high ML model accuracy. Traditionally, good training data has the following characteristics: correct, complete, up-to-date. The data needs to be correct because if annotations are incorrect ML models learn incorrect patterns. Likewise, missing properties inside a dataset reduce the overall expressiveness of a model. Finally, data has to be up-to-date because if we train a model on data from 10 years ago and apply it to data of today, temporal concept shifts change the distributions that cause mispredictions. The larger the amount of such data, the easier it is for the model to generalize and differentiate noise from actual trends. With the maturity of ML algorithms and systems and their application on real-world use cases, good data also has to satisfy additional characteristics: it has to be representative of different groups of a population and free from historic bias. A representative sample of all data is important to ensure high model accuracy for all population groups. For example, in many image datasets, people with dark skin are underrepresented. This misrepresentation has led to poor prediction performance for this group. E. g. Twitter was focusing white people's faces while cropping the faces of people with dark skin[8] or Google[9] and Facebook's[10] models predicted people with black skin as gorillas. While representation requires explicit modeling of different population groups, for some use cases such information leads to amplification of discrimination based on historic biases towards certain groups. One example of such a case is a system for supporting hiring decisions. It turned out that the majority of the historic data contained male hires. Therefore, the model learned that gender was an accurate signal for the prediction task and its predictions discriminated against women.[11] This example shows that it is crucial to

[6] *Stoyanovich, J. / Howe, B. / Jagadish, H. V.*, Responsible data management, PVLDB, vol. 13, no. 12, pp. 3474–3488, 2020.

[7] *Stoyanovich / Howe / Jagadish* (Fn. 6).

[8] *Chowdhury, R.*, Sharing learnings about our image cropping algorithm, 2021 [online]. Available: https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm.

[9] *Vincent, J.*, Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech, 2018 [online]. Available: https://www.theverge.com/2018/1/12/16882408/google-racistgorillas-photo-recognition-algorithm-ai.

[10] *Mac, R.*, Facebook apologizes after a.i. puts 'primates' label on video of black men, 2021 [online]. Available: https://www.nytimes.com/2021/09/03/technology/facebook-ai-raceprimates.htm.

[11] *Dastin, J.*, Amazon scraps secret ai recruiting tool that showed bias against women, 2018 [online]. Available: https://www.reuters.com/article/usamazon-com-jobs-automation-insight-idUSKCN1MK08G.

consider concepts, such as equality of outcome through demographic parity, to protect minority groups. Often the problem is linked to so-called sensitive attributes that are historically correlated but causally irrelevant for the outcome of a prediction task, such as gender, race, or religion. A naive solution would be to simply drop those sensitive attributes. This way the algorithm is blind across these dimensions. Unfortunately, this approach does not work well in practice because the model might learn the affiliation of persons to minority groups from other attributes – so-called proxy attributes. For instance, *Selbst* showed that zip code is a proxy attribute for race in the US.[12] In this paper, we focus on the aspect of fair data engineering. We can approach fairness in two ways: from the individual perspective or the group perspective. Individual fairness ensures equal treatment for people with similar characteristics. Group fairness ensures equal treatment across groups – no group is disadvantaged. At the first glimpse, individual and group fairness might contradict each other in some cases. For example, a higher qualified person from a majority group misses an opportunity, which fell to the top candidate from a minority group despite lower qualification scores. However, Reuben makes the case that with careful consideration, we can avoid such dilemmas.[13] For instance, in the case of financial lending, we start by asking what is the purpose of the algorithm – here, which persons should get a credit? The only important point is that a person can pay back the credit in the end. The second question that has to be answered is what kind of assumptions do we have for the data. E. g. where does the data about creditworthiness come from and is this data meaningful or is it prejudicial? The next question is which characteristics should be used? To estimate creditworthiness, one can rely on income, securities, and assets. Finally, we need to understand whether there is historic or structural discrimination in the data. For example, maybe a structurally poor place of residence is associated with credit denial because despite having high potential to be creditworthy they lack the historic evidence that equals to candidates from more prosperous places. After answering all these questions, one can understand the domain-specific bias problem and can address it with a technical solution. The examples show that the original goal of good data to maximize model accuracy is not sufficient. Instead, one has to consider the problem as a multi-objective problem or more practically a maximization problem under constraints. Constraints are minimum thresholds on metrics that capture novel quality dimensions of ML. These quality dimensions prominently include fairness, privacy (e. g. GDPR compliance), or explainability. In this article, we will briefly discuss the implication of data quality with regard to traditional dimensions as well as novel population-level dimensions on AI and surface the current state-of-the-art technologies that reduce bias in ML technol-

[12] *Selbst, A. D.*, Disparate impact in big data policing, Georgia law review, vol. 52, p. 3373, 2017.

[13] *Binns, R.*, On the apparent conflict between individual and group fairness, FAccT, 2020, pp. 514–524.

ogies and their training data. We demonstrate one particular technology, which is designed to deal with sensitive attributes and their proxies and conclude with a set of suggestions for reconciling socio-technical aspects of algorithmic fairness.

II. How Data Quality impacts AI

ML relies on – and is "programmed" by – training data.[14] Thus, the quality of the training data is fundamental toward robust and accurate models, and ultimately toward useful and reliable ML-based applications.[15] Thus, there is no surprise that data and ML engineers report spending a tremendous amount of time on preparing datasets for ML applications.[16] Traditional dimensions of data quality include completeness, correctness, and freshness of datasets and are per se independent of the downstream ML application. Completeness of a dataset captures to which degree a dataset is populated with content. Incomplete datasets typically miss attribute values either because of negligence in the data creation phase or because certain data values are not known. As ML-based systems are required to understand associations between given properties and the target property, it is easy to see that missing values might reduce the performance. Similarly, it is well understood that incorrect values in the data might negatively impact the accuracy of an ML system. Research on data quality has so far led to a large number of methods, heuristics, and systems that support data quality improvement through data cleaning, which comprises the identification and correction of data quality problems, such as identifying missing values and imputing them. While most of the existing work focuses on cleaning independent of the application – here ML – there are novel sparks towards ML-dependent cleaning techniques. Traditional techniques typically rely on error models to detect duplicate or outlying values, external constraint information (e. g., business or integrity constraints), or human assessment and input (e. g., to recommend repairs or cleaning examples). The separation of data cleaning and the application can lead to several problems. First, it is hard for users to antic-

[14] *Neutatz, F. / Chen, B. / Abedjan, Z. / Wu, E.*, From cleaning before ml to cleaning for ml, IEEE Bulletin, 2021.

[15] *Lee* (Fn. 4); *Li, P. / Rao, X. / Blase, J. / Zhang, Y. / Chu, X. / Zhang, C.*, CleanML: a study for evaluating the impact of data cleaning on ml classification tasks, ICDE, 2021; *Baylor, D. / Breck, E. / Cheng, H.-T. / Fiedel, N. / Foo, C. Y. / Haque, Z. / Haykal, S. / Ispir, M. / Jain, V. / Koc, L.* et al., Tfx: A tensorflow-based production-scale machine learning platform, SIGKDD, 2017, pp. 1387–1395.

[16] *Deng, D. / Fernandez, R. C. / Abedjan, Z. / Wang, S. / Stonebraker, M. / Elmagarmid, A. K. / Ilyas, I. F. / Madden, S. / Ouzzani, M. / Tang, N.*, The data civilizer system, CIDR, 2017; *Agrawal, A. / Chatterjee, R. / Curino, C. / Floratou, A. / Godwal, N. / Interlandi, M. / Jindal, A. / Karanasos, K. / Krishnan, S. / Kroth, B. / Leeka, J. / Park, K. / Patel, H. / Poppe, O. / Psallidas, F. / Ramakrishnan, R. / Roy, A. / Saur, K. / Sen, R. / Weimer, M. / Wright, T. / Zhu, Y.*, Cloudy with high chance of DBMS: a 10-year prediction for enterprise-grade ML, CIDR, 2020; *Kandel, S. / Paepcke, A. / Hellerstein, J. M. / Heer, J.*, Enterprise data analysis and visualization: An interview study, TVCG, vol. 18, pp. 2917–2926, 2012.

ipate which cleaning routines matter for the downstream ML routine, which will unequivocally lead to a waste of resources and user time. Interestingly, improving a dataset could in fact degrade the outcome of the downstream ML model.[17] For instance, it is not clear whether a partial improvement of the data quality might lead to other inconsistencies with systematic errors that had been exploited via downstream neural networks. In a prior study, we manually curated clean and dirty versions of the FAA Flights delay[18] and U.S. Census datasets[19]. Each dataset served a different prediction task. We showed that whether or not cleaning is beneficial heavily depends on the application. For the Flights dataset, cleaned training data improves the model accuracy on both clean and dirty test data. However, cleaning the Census training data actually degrades the model accuracy on dirty data. In fact, training and testing on dirty data is as accurate as training and testing on clean data, yet requires no effort. This experiment provides evidence that cleaning is not a local "one-and-done" process. In fact, the appropriate cleaning intervention is dependent on the type of error as well as the rest of the application, and should be approached from this perspective. Consequently, all of the complexities inherent in modern ML applications become complexities that affect how data is cleaned. In prior work, we argued that data cleaning needs to take an end-to-end application-driven approach that integrates cleaning throughout the ML application.[20] In contrast to traditional data cleaning, there are approaches that are embedded within the ML development phase, which focuses on model development, training, and evaluation. Although the ML community has developed a multitude of robust model designs and training techniques,[21] it is often better to directly address errors and biases in the data.[22] Methods that are applied in this phase leverage the ML models and validation data to identify both the data points for cleaning and the appropriate cleaning routine that directly improve the ML accuracy. So far the discussed methods and problems related to objective and factual problems with the data. However, data quality in the context of AI spans other dimensions as well. In particular, traditional means of cleaning do not count in population-level problems. A recent study found out that existing data imputation methods rely on the maximum likelihood of values skewing the result of imputation towards majority groups and amplifying misrepresentation of minority groups in the data.[23] In fact, the data quality dimension of fairness with regard to metrics on demographic par-

---

[17] *Amershi, S./Begel, A./Bird, C./DeLine, R./Gall, H. C./Kamar, E./Nagappan, N./Nushi, B./Zimmermann, T.*, Software engineering for machine learning: a case study, ICSE, 2019, pp. 291–300.

[18] *Mahdavi, M./Abedjan, Z.*, Baran: Effective error correction via a unified context representation and transfer learning, PVLDB, vol. 13, no. 11, pp. 1948–1961, 2020.

[19] *Li/Rao/Blase* et al. (Fn. 15).

[20] *Neutatz/Chen/Abedjan/Wu* (Fn. 14).

[21] *Li, J. Z.*, Principled approaches to robust machine learning and beyond, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, USA, 2018.

[22] *Li/Rao/Blase* et al. (Fn. 15).

[23] *Schelter/He/Khilnani/Stoyanovich* (Fn. 5).

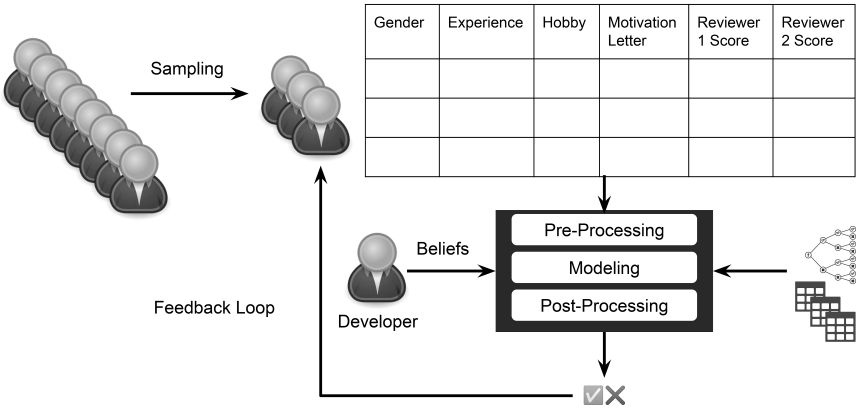

Figure 1: Fairness in the ML Workflow.

ity has gained more attention in recent years. In the next section, we discuss how existing work tries to assess and enforce fairness in ML applications.

III. How Data leads to Discrimination

Machine learning finds patterns in the data to make predictions for new unseen data. If this data is biased, the extracted patterns are likely to be biased, too. These biased patterns lead to predictions that discriminate. Biases sneak into an ML application across the entire ML workflow, which consists of data collection, pre-processing, modeling, and post-processing. *Friedman* and *Nissenbaum* identified three main types of bias: pre-existing, technical, and emergent.[24] To describe these types of bias in more detail, we explain them with the help of the workflow for an ML hiring application as shown in Figure 1. Pre-existing bias has its roots in the current beliefs of society and is generally independent of the ML application. For instance, the fraction of women in the German parliament in 2021 is only 35 %,[25] or the well-known phenomenon – the gender pay gap – that women earn significantly less than men.[26] So, even if we would be able to gather data on all people in the world, an accurate representation of existing circumstances in our society would let our ML application to reproduce discrimination. For instance, if a model leverages the salary to predict whether a person can get a credit or not,

[24] *Friedman, B. / Nissenbaum, H.*, Bias in computer systems, ACM Trans. Inf. Syst., vol. 14, no. 3, pp. 330–347, 1996.

[25] Wikipedia, Frauenanteil im Deutschen Bundestag seit 1949, 2021 [online]. Available: https://de.wikipedia.org/wiki/Frauenanteil im Deutschen Bundestag seit 1949.

[26] *Blau, F. D. / Kahn, L. M.*, Understanding international differences in the gender pay gap, Journal of Labor economics, vol. 21, no. 1, pp. 106–144, 2003.

the gender pay gap makes the salary a proxy attribute for gender and therefore, the model will discriminate against women. Consider our running example of an ML-driven hiring application. To address pre-existing bias, the data scientists have to first assess the types of risks with regard to discrimination and how existing circumstances influence the outcome. For instance, one could follow the concept of substantive equality of opportunity that only compares people to other people with the same circumstances.[27] Next, the data scientist has to define the characteristics that should be allowed to judge the applicants. Assume, the available data is the gender, the experience, the hobby, the motivation letter, and two reviewers' scores. The data scientist would carefully remove attributes that are not supposed to influence the outcome. The application should be blind with regard to attributes, such as gender. Ideally, the scientist would also identify proxies of such attributes that strongly correlate with their values. For example, the data scientists discover that the hobby is a proxy attribute for gender. So, they remove this attribute. Furthermore, they realize that reviewer 2 prefers applicants from certain universities. Therefore, they group the score by the university and normalize it. This way, they make sure that they can use the reviews without introducing bias. Finally, the data scientists have to make sure that, for all specified features, the necessary data is available across all groups because missing values might introduce new bias. This concept is known as feature equity.[28] After formulating the beliefs and identifying the attributes, one might think that, now, we can develop any algorithm and do not need to worry about fairness anymore because the data does not contain any sensitive or proxy attributes. This blindness approach is quite common. However, in many cases, we still have to measure the bias based on the formulated beliefs because the underlying data or the following algorithm might introduce bias. Technical bias is introduced or enforced by the developed ML application. Technical bias can occur at any stage of the ML workflow, ranging from sampling, preprocessing, modeling, and post-processing. In the sampling phase, the data scientists have to choose from a large number of previous applicants who they choose as a foundation to learn who to hire. An example of bias in sampling is an Amazon ML application, which "learned" that female candidates are less likely to succeed at the company because their data contained mainly male candidates.[29] So, we have to ensure to draw a representative sample across all demographic groups. This ensurance is known as representation equity.[30] So, the data scientists might access

---

[27] *Zehlike, M./Yang, K./Stoyanovich, J.*, Fairness in ranking: A survey, CoRR, vol. abs/2103.14000, 2021.

[28] *Jagadish, H. V./Stoyanovich, J./Howe, B.*, The many facets of data equity, Proceedings of the Workshops of the EDBT/ICDT 2021 Joint Conference, Nicosia, Cyprus, March 23, 2021, ser. CEUR Workshop Proceedings, C. Costa and E. Pitoura, Eds., vol. 2841. CEUR-WS.org, 2021 [online]. Available: http://ceur-ws.org/Vol-2841/PIE+Q 6.pdf.

[29] *Dastin* (Fn. 11).

[30] *Jagadish/Stoyanovich/Howe* (Fn. 28).

data about previous applicants in the company and whether they perform well or poorly later on. Needless to say that it is critical that the HR department ensures that the pool of applicants is as diverse as possible and that the interview process is as fair as possible. Otherwise, the first step of sampling is likely to introduce bias already. Pre-processing is another step that might introduce bias. It transforms the data into a machine-understandable format. These transformations include data cleaning and augmentation. For instance, *Schelter* et al. showed that missing value imputation such as mean value imputation can enforce bias, especially because minority groups are more likely to avoid disclosing sensitive information.[31] Therefore, developers have to keep in mind that any data transformation that joins or filters the data might change the data distribution and introduce or amplify bias. An example of how data augmentation can introduce bias can be explained by the transformation that translates the textual motivation letter into numeric form. For instance, one can leverage a neural network that models the language based on the data extracted from Wikipedia to transform the textual motivation letter into a numerical vector. However, 87 % of the contributors of Wikipedia are men and therefore consciously or unconsciously introduce bias.[32] Therefore, data scientists have to check all components and libraries that they include into the workflow. In the post-processing phase, data scientists modify the thresholds when to hire a candidate or not. Depending on the company, the policy allows for more false positives or false negatives. False negatives are applicants that are not hired but would have been great employees if they would have had a chance to prove themselves. False positives are applicants that are hired but turn out to not fit the company. Modifying these thresholds, the data scientists always keep track of fairness because different thresholds might affect minorities in different ways. After some time with the company, the hired applicant's data can be used to train a new model. However, the data scientists keep monitoring to avoid any reinforcing biases in this feedback loop. For instance, they should be careful to avoid survivorship bias. In the Second World War, the US army analyzed where planes were shot to strengthen the parts that are more likely to be in danger. Abraham Wald realized that it was best to strengthen the parts that in the analysis were unscathed because all the planes that they analyzed actually made it back to safety.[33] So, in our case, we run into the danger of only including data points about applicants that were actually hired. While one can remove preexisting bias and technical bias in a system during implementation, emergent bias arises only in a context of use, e. g. by incorrect use or distribution shifts over time. For instance, after finishing a hiring application for the German branch of a company, a manager wants to use

[31] *Schelter / He / Khilnani / Stoyanovich* (Fn. 5).

[32] *Torres, N.*, Why do so few women edit wikipedia, Harvard Business Review, vol. 2, 2016.

[33] *Wald, A.*, A method of estimating plane vulnerability based on damage of survivors, Center for Naval Analyses, 1980.

the same application internationally. However, the school systems differ significantly and might discriminate against some nationalities. Therefore, it is important that whenever an application is used in a new context, one has to analyze all potential biases again. This concept is also known as output equity.[34] We showed that bias can be introduced at multiple stages of any ML application. As far as the ML and data science community is aware of these problems there have been attempts and technical solutions for several of the aforementioned pitfalls.

IV. State of the Art

Algorithmic bias reduction has been approached from different perspectives. In this section, we briefly survey existing approaches for fair ML and discuss fundamentals of feature engineering, which will be important to present our take on bias reduction in ML. Algorithmic bias reduction can be categorized threefold: by where in the ML pipeline they address the issue, how they measure bias, and what types of bias they address. Algorithmic bias reduction can be implemented at three stages: during preprocessing, in-processing, or post-processing. Pre-processing approaches[35] reduce the bias by modifying the data, e. g. shifting the probability distributions in the data,[36] selecting features,[37] or weighting features[38]. In-processing approaches[39] reduce the bias in the model, e. g. by modifying

---

[34] *Jagadish/Stoyanovich/Howe* (Fn. 28).

[35] *du Pin Calmon, F./Wei, D./Vinzamuri, B./Ramamurthy, K. N./Varshney, K. R.*, Optimized pre-processing for discrimination prevention, Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), 2017, pp. 3992–4001; *Feldman, M.* et al., Certifying and removing disparate impact, Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), 2015, pp. 259–268; *Zhang, L./Wu, Y./Wu, X.*, A causal framework for discovering and removing direct and indirect discrimination, in Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2017, pp. 3929–3935; *Asudeh, A./Jagadish, H. V./Stoyanovich, J./Das, G.*, Designing fair ranking schemes, in Proceedings of the International Conference on Management of Data (SIGMOD), 2019, p. 1259–1276.

[36] *du Pin Calmon/Wei/Vinzamuri/Ramamurthy/Varshney* (Fn. 35); *Feldman* et al. (Fn. 35); *Zhang/Wu/Wu* (Fn. 35).

[37] *Galhotra, S./Shanmugam, K./Sattigeri, P./Varshney, K. R.*, Fair data integration, CoRR, vol. abs/2006.06053, 2020.

[38] *Asudeh/Jagadish/Stoyanovich/Das* (Fn. 35).

[39] *Schelter/He/Khilnani/Stoyanovich* (Fn. 5); *Kilbertus, N./Rojas-Carulla, M./Parascandolo, G./Hardt, M./Janzing, D./Scholkopf, B.*, Avoiding discrimination through causal reasoning, Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), 2017, pp. 656–666; *Nabi, R./Shpitser, I.*, Fair inference on outcomes, Proceedings of the Conference on Artificial Intelligence (AAAI), 2018, pp. 1931–1940; *Russell, C./Kusner, M. J./Loftus, J. R./Silva, R.*, When worlds collide: Integrating different counterfactual assumptions in fairness, Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), 2017, pp. 6414–6423; *Perrone, V./Donini, M./Kenthapadi, K./Archambeau, C.*, Fair bayesian optimization, CoRR, vol. abs/2006.05109, 2020.

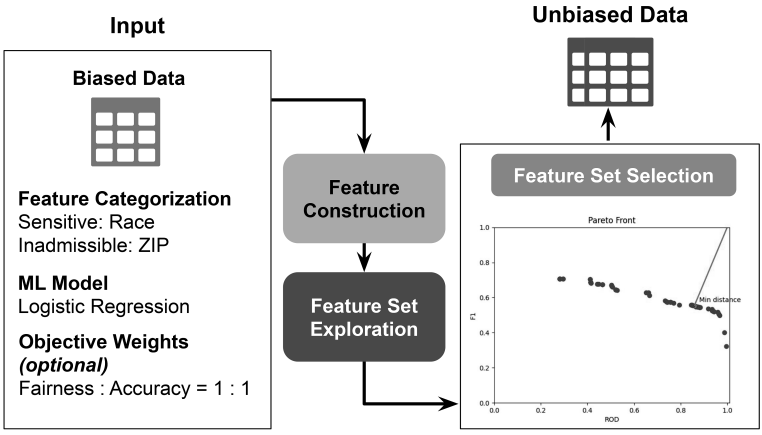

Figure 2: Architecture of the Fairness Explorer.

the model's loss.[40] Another example is to optimize the model's hyperparameters, which describe the configurations of a model for a setting, based on a fairness metric.[41] Post-processing approaches[42] reduce the bias in the final model predictions by applying transformations.[43] The second way to differentiate bias reduction approaches is to compare their fairness metrics. The two main approaches to measuring fairness are associational and causal. Associational approaches measure fairness in the model's predictions between groups of the sensitive feature. Three representative fairness criteria that are used by such approaches are independence, separation, and sufficiency.[44] Independence describes the concept of returning the same prediction for two similar individuals that only differ with respect to their sensitive attribute, such as religion, race, or gender. However, this metric ignores potential correlations between the group and the prediction target. Separation allows the score and the sensitive attribute to correlate to the extent that is justified

[40] *Kamishima, T./Akaho, S./Asoh, H./Sakuma, J.*, Fairness-aware classifier with prejudice remover regularizer, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), ser. Lecture Notes in Computer Science, vol. 7524, 2012, pp. 35–50.

[41] *Perrone/Donini/Kenthapadi/Archambeau* (Fn. 39).

[42] *Hardt, M./Price, E./Srebro, N.*, Equality of opportunity in supervised learning, Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS), 2016, pp. 3315–3323; *Woodworth, B. E./Gunasekar, S./Ohannessian, M. I./Srebro, N.*, Learning non-discriminatory predictors, Proceedings of the Conference on Learning Theory (COLT), vol. 65, 2017, pp. 1920–1953.

[43] *Hardt/Price/Srebro* (Fn. 42); *Corbett-Davies, S./Pierson, E./Feller, A./Goel, S./Huq, A.*, Algorithmic decision making and the cost of fairness, Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), 2017, pp. 797–806.

[44] *Barocas, S./Hardt, M./Narayanan, A.*, Fairness and Machine Learning, 2019, http://www.fairmlbook.org.