Principled Bias Detection and Mitigation across the ML Data Lifecycle

Bruno Scarone
Supervised by Ricardo Baeza-Yates and Renée J. Miller
Northeastern University
Boston, MA, USA
scarone.b@northeastern.edu

ABSTRACT

Machine learning models increasingly drive high-stakes decisions, yet the data they learn from remains a primary and underexplored source of discriminatory outcomes. This dissertation treats data bias as a first-class data management problem and develops principled frameworks for detecting and correcting it across the ML data lifecycle. We introduce Uniform Bias (UB), an interpretable intersectional measure with formal guarantees, and an ILP-based mitigation framework that models coverage and budget constraints. We characterize the price of fairness as the minimum data modification cost as a function of fairness tolerance. We then study survival bias (a form of selection bias) under iterative model retraining in credit scoring, exposing a structural failure mode in which standard evaluation metrics reward strategies that amplify selection bias. We further propose controlled exploration as an assumption-free diagnostic. Together, these contributions position fairness not as a model-level constraint but as a data management problem amenable to the rigorous, guarantee-driven methods that characterize the database tradition.

VLDB Workshop Reference Format:
Bruno Scarone. Principled Bias Detection and Mitigation across the ML Data Lifecycle. VLDB 2026 Workshop: Ph.D. Workshop.

1 INTRODUCTION

Machine learning models are increasingly used to make high-stakes decisions — from loan approvals to hiring — and the quality of their training data has direct consequences for the people affected by these decisions [2]. In particular, biased training data is one of the most well-documented sources of discriminatory ML outcomes [8], and yet surprisingly few principled tools exist to detect and fix bias at the data level with formal guarantees. Existing bias measures are either not computable directly from the data, fail to handle intersectional groups or non-binary settings, or lack the formal foundations needed to derive mitigation strategies with correctness guarantees [8, 17]. This dissertation treats data bias as a first-class data management problem, by developing a unified framework for bias detection and mitigation across the ML data lifecycle — the iterative process through which data is collected, curated, and used to train and redeploy ML models. Concretely, we show that:
(1) Bias in tabular ML data with discrete sensitive attributes can be measured intersectionally in an interpretable and computable way.
(2) Bias can be mitigated optimally for any given fairness tolerance through addition and deletion of real data tuples. In the presence of coverage constraints, sufficient group representation across intersectional subgroups can be formally guaranteed.
(3) In iterative ML model retraining, standard evaluation metrics systematically mislead practitioners by rewarding strategies that amplify selection bias. This feedback loop can be mitigated through controlled exploration without assumptions about missing labels.
The rest of this paper is organized as follows. Section 2 covers related work. Section 3 introduces Uniform Bias. Section 4 presents bias mitigation under coverage constraints and the price of fairness. Section 5 studies survival bias in credit scoring models. Section 6 concludes and outlines directions for future work.

2 RELATED WORK

Bias measurement and intersectionality. Researchers have proposed a wide variety of bias measures for ML [8], but consolidating a common set of metrics remains an open challenge, particularly for multi-class problems with non-binary sensitive attributes [8]. Crenshaw [6] established that discrimination against Black women exceeds the sum of racism and sexism individually, motivating the need for intersectional analysis; Buolamwini and Gebru [4] demonstrated intersectional bias empirically in commercial facial recognition systems. In the data fairness setting, datasets can satisfy group fairness for gender and race separately while failing at their intersection [9]. Our work [15] introduces Uniform Bias (UB), an interpretable measure that handles multi-valued sensitive attributes and labels, addressing gaps identified by recent surveys [8].
Data management approaches to fairness. Nargesian et al. [10] address fairness-aware data integration, proposing to integrate data from multiple sources cost-effectively such that subgroups meet user-specified count distributions. Our framework complements this by deriving the target distribution from a principled fairness goal rather than requiring the user to specify it. Salimi et al. [12] connect database repair to algorithmic fairness using a causal reasoning framework, but their measure is restricted to the binary, single-attribute setting and their repair algorithm does not provide direct control over bias; our Uniform Bias measure handles multiple non-binary sensitive attributes and label values directly. Chen et al. [5] argue that unfairness stemming from inadequate sample sizes should be addressed through data collection rather than model constraints, finding that additional samples often suffice to improve

Proceedings of the VLDB Endowment. ISSN 2150-8097.

fairness without sacrificing accuracy; our coverage constraints formalize this principle, decoupling fairness tolerance from minimum representation requirements and providing a formal guarantee that the optimizer maintains sufficient data representation.

Survival bias and iterative ML. The feedback loop we study is an empirical instance of performative prediction [11], where model deployment shifts the data distribution; analogous self-reinforcing effects have been documented in predictive policing [7]. Reject inference [1] refers to the family of techniques that attempt to recover missing outcome information for rejected applicants; we conduct a systematic evaluation of several such methods under iterative retraining [13].

3 UNIFORM BIAS

Existing bias measures such as the Impact Ratio (IR) and Mean Difference (MD) handle only single binary sensitive attributes and binary labels, and can assign identical values to datasets with substantially different label distributions [15, 17]. We introduce *Uniform Bias* (UB), a measure that is directly computable from table statistics in a single linear-time pass over the data, handles multiple non-binary categorical sensitive attributes and labels, and admits an interpretable quantification of how far a dataset is from the unbiased ideal [15]. We consider the classical setting of algorithmic fairness where tuples in a dataset T have a label attribute Y and multiple categorical sensitive attributes $\mathbf{S} = \langle S_1, \ldots, S_k \rangle$; both Y and each S_i may take non-binary values. We represent a sensitive group as a k-tuple over the k sensitive attributes, where each entry is either a specific value or a wildcard $*$ indicating marginalization over that attribute [14]. For example, with $\mathbf{S} = \langle \text{Gender}, \text{Race} \rangle$, the tuple $\langle \text{female}, * \rangle$ refers to all women regardless of race, while $\langle \text{female}, \text{non-white} \rangle$ is a fully specified intersectional group. For a group $\mathbf{s}$ and label y, we denote by $[\mathbf{s}y]$ the number of tuples with sensitive attributes $\mathbf{s}$ and label y. To simplify notation, when a value such as $\mathbf{s}$ or y appears in a formula, it denotes the count of tuples with that value (i.e., $|\mathbf{s}|$ and $|y|$ respectively). We define the relative frequencies $f_{\mathbf{s},y} = [\mathbf{s}y]/\mathbf{s}$ (the proportion of group $\mathbf{s}$ with label y) and $f_y = y/n$ (the global proportion with label y, where n is the total number of tuples). The intuition behind UB is as follows: in an unbiased dataset, the proportion of tuples with label y within any group $\mathbf{s}$ should match the global proportion f_y. The *Uniform Bias* of group $\mathbf{s}$ w.r.t. label y is defined as $b_{\mathbf{s},y} = 1 - f_{\mathbf{s},y}/f_y$ [15], measuring the percentage deviation from this ideal: a value of $b_{\mathbf{s},y} = 0.2$ means group $\mathbf{s}$ has 20% fewer outcomes with label y than expected under the unbiased condition ($f_{\mathbf{s},y} = 0.8 \cdot f_y$), while negative values indicate overrepresentation. UB ranges from $-\infty$ to 1 and is size-independent: two datasets with the same relative frequencies have the same UB regardless of their size. A dataset T is unbiased w.r.t. $\mathbf{S}$ and label y if $f_{\mathbf{s},y} = f_y$ for every $\mathbf{s} \in \text{Dom}(\mathbf{S})$ [15].

4 BIAS MITIGATION UNDER COVERAGE CONSTRAINTS & THE PRICE OF FAIRNESS

Measuring bias is only the first step: correcting or mitigating it requires modifying the dataset. We restrict modifications to additions of real tuples from external sources and deletions of existing ones, as arbitrarily altering attribute values could produce a dataset no longer corresponding to real individuals. The question is then: which real tuples should be added or deleted, and at what cost? Our original UB framework [15] gave a closed-form mitigation algorithm that adds real tuples to reach zero bias, and additionally proves that the global label frequencies f_y are preserved [15]—providing an additional guarantee that the mitigated dataset remains suitable for use in the same context as the original, since most ML methods rely on assumptions about the label distribution. However, it fixed a single mitigation strategy in advance—that is, a single choice of reference label per group in the underlying linear system—and did not consider deletions, coverage requirements, or acquisition costs—all of which we address as follows in our paper [14]:

Coverage constraints. Bias mitigation that only minimizes bias can collapse group-label counts to a handful of tuples, destroying predictive performance of ML models. We incorporate *coverage constraints* into the mitigation framework: requirements of the form $[\mathbf{s}y] + \Delta[\mathbf{s}y] \geq m_{\mathbf{s},y}$ specifying, for each group-label pair, the minimum number of tuples the mitigated dataset must retain [14]. In the linear system underlying the mitigation algorithm [15], each group $\mathbf{s}$ has a free variable $\Delta[\mathbf{s}y_i]$: intuitively, this is the number of additions or deletions for a chosen reference label y_i, whose value the user assigns directly, with all remaining $\Delta[\mathbf{s}y]$ values determined as a consequence. Coverage constraints translate into a range of admissible values for each free variable; the closed-form algorithm [14] selects the smallest admissible value, minimizing the total number of data modifications while satisfying all coverage requirements simultaneously. We prove tight error bounds on the residual bias introduced by integer rounding, showing that the approximation error is $O(1/(|\mathbf{s}|+\Delta_{\mathbf{s}}^{\text{ex}}))$; a number that is remarkably small in practice, since $|\mathbf{s}| + \Delta_{\mathbf{s}}^{\text{ex}}$ is the size of group $\mathbf{s}$ after exact mitigation.

The price of fairness. To find the globally optimal mitigation strategy rather than fixing one in advance, we formulate bias mitigation as an *integer linear program* (ILP) [14]. The ILP minimizes an objective (final dataset size, modifications, or weighted cost) subject to an ϵ-approximate fairness constraint for every group-label pair, together with coverage and budget constraints. Varying ϵ traces out the *price of fairness* curve: the minimum data modification cost as a function of fairness tolerance. This is directly actionable for regulatory compliance: regulations such as the EU AI Act implicitly mandate a fairness threshold, and the curve tells practitioners exactly what data acquisition budget that threshold requires.

The role of deletions. A key and non-obvious insight is that permitting strategic deletions of overrepresented tuples substantially reduces the number of additions required. Deletions shift the global label frequencies f_y, lowering the target that underrepresented groups must reach and thereby reducing the required additions—which matters because acquiring new external data is costly and may introduce distribution shifts between the external source and the original dataset [14]. The following example, drawn from our work [14], illustrates this concretely.

Example 4.1. Consider a property management company that is training a rental screening model using the Adult dataset [3] (n = 48,842 tuples, binary income label, binary sensitive attributes sex and race), a standard benchmark in ML fairness research. Bias

analysis reveals that the dataset disadvantages women, particularly non-white women: letting fnw = ⟨female, non-white⟩, the Uniform Bias $b_{\text{fnw},+} = 0.70$ for the high-income label—meaning 70% fewer high-income outcomes than expected under the unbiased condition ($f_{\text{fnw},+} = 0.30 \cdot f_+$). To mitigate all group-label biases to within $\epsilon = 0.05$ (the fairness tolerance) using only additions, the ILP requires 4,201 new tuples. Allowing 1,059 strategic deletions of overrepresented high-income male tuples reduces the required additions to 1,599—a 62% reduction in data acquisition cost, with total modifications dropping from 4,201 to 2,658. This example[1] shows why modeling deletions is essential: without it, a data scientist would over-purchase external data when a modest reduction of the majority group would suffice.

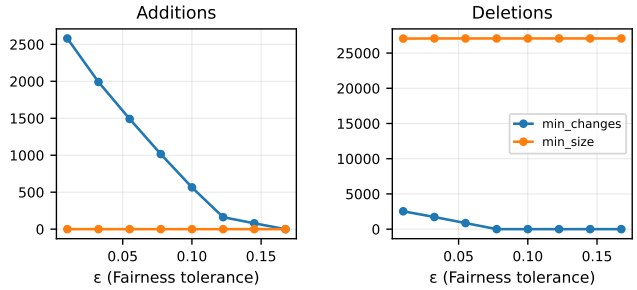

Figure 1: Price of fairness on the Adult dataset ($m_{s,y} = 1$, equal addition and deletion costs). Minimizing changes balances both operations; minimizing size deletes aggressively and adds nothing, regardless of ϵ.

ML-based evaluation. We show empirically that bias mitigation via the ILP preserves predictive accuracy across five classifiers and three datasets. Crucially, the choice of optimization objective matters more than the fairness tolerance itself: under the objective of minimizing the resulting dataset size without coverage constraints, the optimizer can reduce the Default dataset to as few as 24 observations, causing accuracy to collapse. As shown in Figure 1, the two objectives take fundamentally different paths: minimizing changes balances additions and deletions, implicitly preserving data representation, while minimizing size deletes aggressively and adds nothing. This is why explicit coverage constraints are essential to prevent mitigation algorithms from discarding data that is informative for prediction but not required for fairness.

Extensions. The framework supports two further generalizations. First, the unbiased condition $f_{\mathbf{s},y} = f_y$ can be relaxed by introducing constants $K_{\mathbf{s},y} > 0$ representing the desired ratio of the success rate of group $\mathbf{s}$ to the population rate, with the standard unbiased condition recovered by setting $K_{\mathbf{s},y} = 1$ [15]. This allows domain experts to encode normative judgments about acceptable disparities directly into the mitigation target. Second, both the closed-form algorithm and the ILP support upper bound constraints $[\mathbf{s}y] + \Delta[\mathbf{s}y] \leq M_{\mathbf{s},y}$, useful when over-collection of data from a specific group raises ethical or practical concerns [14].

[1] Binary attributes used for illustration only; UB, the closed-form algorithm, and the ILP all generalize to non-binary attributes.

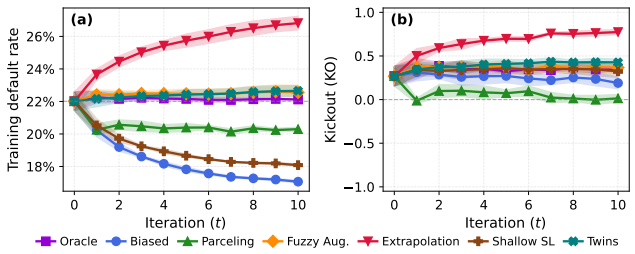

Figure 2: TDR (left) and Kickout (right) over ten lending iterations for all strategies on Default [16] (GBDT). Dashed line: population default rate (22.0%). The value $KO = 1$ is best (every rejection is a true defaulter); -1 is worst.

5 BIAS IN ITERATIVE ML MODEL RETRAINING

Static bias mitigation assumes a fixed dataset. In practice, credit scoring models are *iteratively retrained*: a model selects which applicants to approve, outcomes are observed only for approved applicants, and the model is retrained on this filtered data. This induces *survival bias* —only the outcomes of individuals who pass a selection criterion are observed— since rejected applicants never receive credit and their true repayment behavior remains unknown. Our forthcoming ECML 2026 paper [13] studies this feedback loop and uncovers two structural results.

Temporal lending simulation. We simulate iterative model-based lending using three credit scoring datasets, including Default [16]. An initial training set of size n_0 is drawn uniformly from the complete dataset, modeling a bank's initial data collection step; n_0 is selected per dataset–model pair via learning curve analysis. At each subsequent lending cycle $t \in \{1, \ldots, T\}$, a uniform random sample of new applicants arrives; the current model approves those predicted to repay and rejects the rest, observing true outcomes only for approved applicants. We evaluate seven strategies, each defining a different rule for how rejected applicants are handled when retraining the next model. The BIASED baseline trains only on accepted applicants, representing the full extent of survival bias accumulation; the ORACLE baseline observes true outcomes for all applicants, providing a theoretical upper bound. We then evaluate five *reject inference* (RI) strategies—methods that attempt to impute the missing labels of rejected applicants and include them in the training set—namely PARCELING, FUZZY AUGMENTATION, SIMPLE EXTRAPOLATION, SHALLOW SELF-LEARNING, and TWINS, using GBDT and RF classifiers.

Training default rate as a diagnostic. A key contribution is the *training default rate* (TDR) as a diagnostic for feedback-loop severity: the proportion of defaulters in the training set at each iteration, which should remain near the population default rate in the absence of survival bias. TDR reveals a pattern invisible to standard metrics. On Default [16]/GBDT by iteration $t = 10$, SIMPLE EXTRAPOLATION inflates TDR to 26.8% (a +22% relative distortion above the population rate of 22.0%), while the BIASED baseline deflates it to 17.1% (−22% relative)—a symmetric distortion of opposite sign. SIMPLE EXTRAPOLATION does not mitigate survival bias; it reverses its sign, substituting one distributional distortion for another (Figure 2, left).

In contrast, Fuzzy Augmentation and Twins track the population rate as closely as Oracle, making them the only strategies that genuinely preserve training set representativeness—a property captured by the *Kickout* metric, which measures rejection quality as the difference between the fraction of rejections that are true defaulters and the fraction that are not, and that standard accuracy entirely fails to capture.

The illusion of improvement. Standard metrics—accuracy, precision, recall—systematically reward strategies that amplify selection bias. The clearest illustration is that Simple Extrapolation, which labels rejected applicants using the model's own predictions and retrains on the augmented data, achieves the highest accuracy and recall of all seven evaluated strategies across all six experimental configurations, yet it has no access to true labels. Meanwhile, the Oracle strategy—which observes true outcomes for all applicants and represents the theoretical optimum—achieves *lower* accuracy than the Biased baseline, because retaining genuine label diversity makes the learning task harder. A practitioner relying on standard metrics would therefore select the worst strategy and conclude the model is improving when it is deteriorating [13].

Controlled exploration. We propose *controlled exploration* as an assumption-free alternative to reject inference [13]: the lender deliberately approves a fraction r of rejected applicants and observes their true repayment outcomes, where $r = 0$ corresponds to Biased (maximum exploitation) and $r = 1$ to Oracle (maximum exploration). Varying $r \in [0, 1]$ shows that accuracy and Kickout move in opposite directions: on Default [16]/gbdt, accuracy drops by only 1.5% at full exploration [13] while Kickout rises by 16% (Figure 2, right). Even minimal exploration ($r = 2$–5%) is sufficient to diagnose the severity of the feedback loop at near-zero cost, defining a *diagnostic zone* that allows a lender to determine whether its model is operating on a representative sample or on one corrupted by its own past decisions [13].

6 CONCLUSIONS & FUTURE WORK

This dissertation establishes that data bias in tabular ML data can be measured and mitigated with principled mathematical frameworks and formal guarantees, and that survival bias under iterative retraining can be diagnosed, exposing how standard evaluation metrics mislead practitioners. Uniform Bias measures bias interpretably and intersectionally; the ILP mitigation framework finds globally optimal corrections under coverage and cost constraints. In the iterative setting, the controlled exploration strategy breaks the feedback loop without assumptions about missing labels. Together, these contributions treat fairness not as a model-level constraint but as a data management problem: bias mitigation as constrained data repair, coverage requirements as integrity constraints on group representation, and the price of fairness as the cost of data acquisition under those constraints — problems amenable to the rigorous, guarantee-driven methods that characterize the database tradition. In addition to the published results, we are currently pursuing three directions of work described next.

Formal intersectional bias decomposition. The UB measure can be evaluated independently for every group-label pair, including intersectional groups such as ⟨female, non-white⟩. What it cannot yet do is quantify how much of the bias observed for an intersectional group is *non-additive*—that is, genuinely intersectional rather than a superposition of marginal biases. Formalizing this decomposition would give practitioners a tool to distinguish additive discrimination from intersectional effects.

Bias mitigation under incomplete data. The current mitigation framework assumes all sensitive attribute values are observed. Real datasets frequently contain null values in sensitive attributes, which introduces uncertainty about group membership. Extending the framework to handle missing sensitive attributes is a natural and practically important next step.

Social cost of bias. The iterative credit scoring work quantifies the feedback loop in terms of model metrics. A forthcoming work operationalizes the *social cost* of survival bias: the aggregate harm imposed on rejected applicants who would have repaid but were excluded from credit access.

REFERENCES

[1] John Banasik, Jonathan Crook, and Lyn Thomas. 2003. Sample selection bias in credit scoring models. *Journal of the Operational Research Society* 54, 8 (2003), 822–832.

[2] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671.

[3] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. doi: https://doi.org/10.24432/C5XW20.

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, USA, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[5] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory?. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc., Montréal, Quebec, Canada. https://proceedings.neurips.cc/paper_files/paper/2018/file/1f1baa5b8edac74eb4eaa329f14a0361-Paper.pdf

[6] Kimberlé Crenshaw. 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist Legal Theories*. Routledge, London, UK, 23–51.

[7] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency*. PMLR, 160–171.

[8] Max Hort, Zhenpeng Chen, Jie M Zhang, Mark Harman, and Federica Sarro. 2024. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing* 1, 2 (2024), 1–52.

[9] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An Empirical Study of Rich Subgroup Fairness for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 100–109. https://doi.org/10.1145/3287560.3287592

[10] Fatemeh Nargesian, Abolfazl Asudeh, and HV Jagadish. 2021. Tailoring data source distributions for fairness-aware data integration. *Proceedings of the VLDB Endowment* 14, 11 (2021), 2519–2532.

[11] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.

[12] Babak Salimi, Bill Howe, and Dan Suciu. 2020. Database repair meets algorithmic fairness. *ACM SIGMOD Record* 49, 1 (2020), 34–41.

[13] Bruno Scarone and Ricardo Baeza-Yates. 2026. The Illusion of Improvement: Reject Inference Strategies in Credit Scoring. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*. Springer, Naples, Italy.

[14] Bruno Scarone, Alfredo Viola, and Renée J. Miller. 2026. Data Bias Mitigation under Coverage Constraints & The Price of Fairness. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26)*. ACM, New York, NY, USA, 4935–4968. https://doi.org/10.1145/3805689.3812359

[15] Bruno Scarone, Alfredo Viola, Renée J. Miller, and Ricardo Baeza-Yates. 2025. A Principled Approach for Data Bias Mitigation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, 3 (Oct. 2025), 2273–2283. https://doi.org/10.1609/aies.v8i3.36712

[16] I-Cheng Yeh. 2016. Default of credit card clients. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55S3H.

[17] Indrė Žliobaitė. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31, 4 (2017), 1060–1089.