Data-Centric Factors in Algorithmic Fairness

Nianyun Li, Naman Goel, Elliott Ash

Objectives

- Empirically understand and document the extent to which various data-centric factors affect the fairness and accuracy of machine learning algorithms.
- How these data-centric factors interact with fairness interventions and other algorithmic design choices.
- Contribute a new, bigger dataset in recidivism prediction for fairness research.
- Encourage critical discussion on designing more rigorous evaluation and benchmarking methods for fair machine learning algorithms.

Wisconsin Circuit Courts Dataset

- WCCA API indexes public case records and docket information from 72 county courts.
- Collected records of cases filed from 1970, through 2020. 11M records (2.5M criminal).
- *Constructed* a dataset for machine learning as follows:

Features (X): current offense, prior criminal history until judgment disposition date, gender, race, age at disposition date, age at the first offense, and local demographic attributes*

Target (Y): binary: 1 if defendant recidivates, 0 otherwise

Summary of the Dataset

	Full sample	Caucasian	African American	Hispanic	Native American	Asian
Sample size	1,476,967	964,922	333,036	101,607	63,862	13,540
Sample share		65.33%	22.55%	6.88%	4.32%	0.92%
Recidivism (if observed)	42.21%	40.34%	46.43%	38.76%	56.47%	37.80%
Sex						
Male	80.40%	79.05%	83.47%	88.88%	69.65%	87.57%
Age						
Below 30	51.38%	49.45%	54.13%	56.91%	53.71%	68.60%
30 to 60	47.44%	49.09%	45.17%	42.61%	45.58%	30.85%
Case type						
Felony	32.18%	30.76%	39.98%	21.09%	29.80%	36.39%
Misdemeanor	43.04%	43.67%	43.14%	34.12%	47.55%	40.89%
Criminal Traffic	24.78%	25.57%	16.88%	44.79%	22.66%	22.73%

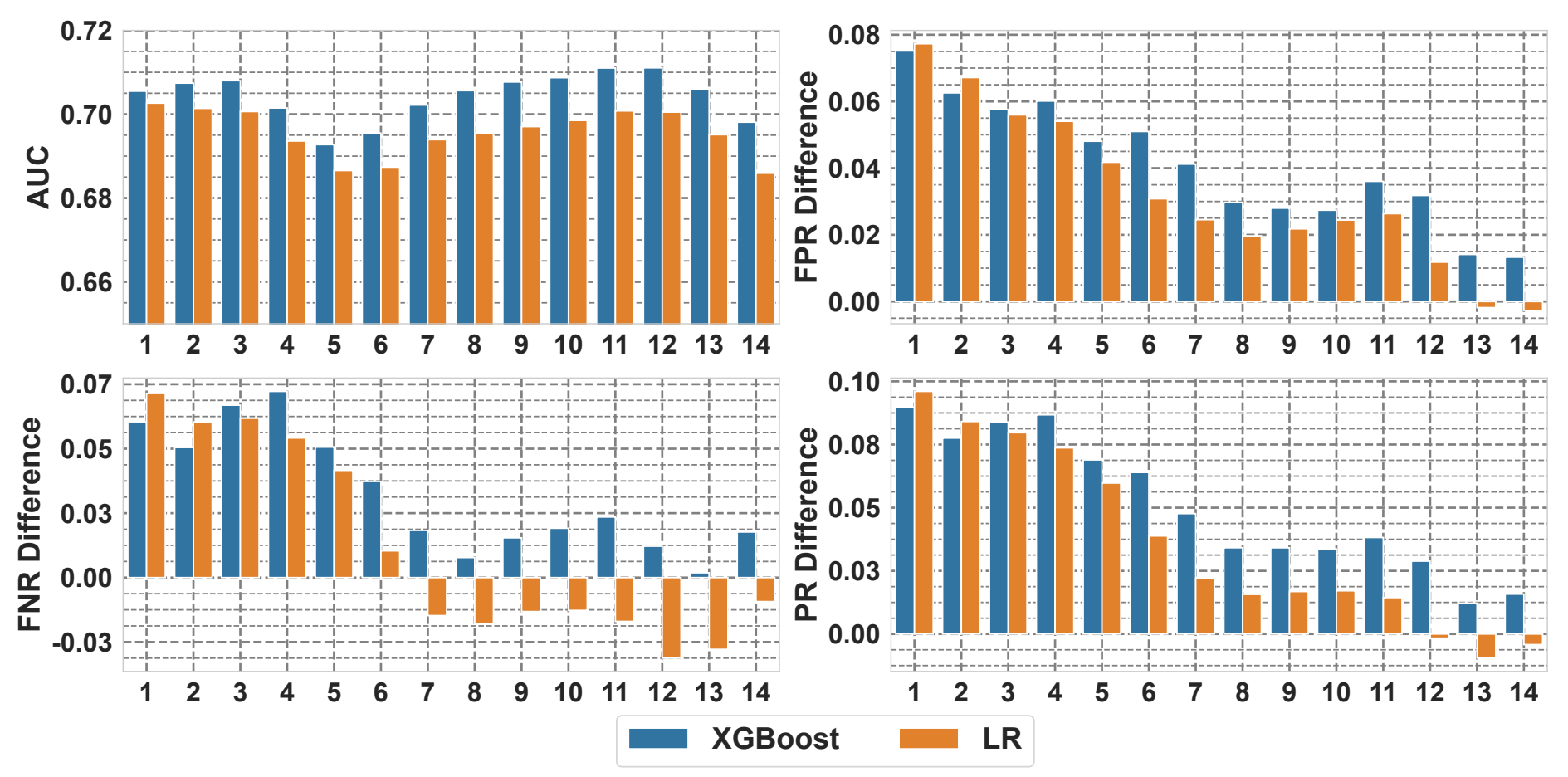
Preliminaries

- Classifiers: Logistic Regression (LR) and XGBoost
- Performance Metrics: Accuracy and AUC
- Fairness Metrics: FPR Difference, FNR Difference, PR difference, Bias Amplification.
- The largest FPR difference is around 13%, the largest FNR difference is around 18%, and the largest PR difference is around 20%; all between Native American and Hispanic groups with XGBoost classifier.
- Hispanic and Caucasian groups receive the most favorable decisions in the dataset, followed by Asian, African American and Native American groups (in that order).
- The overall accuracy of XGBoost and LR is not very different but the FNR, FPR and PR differences between groups is higher for XGBoost.

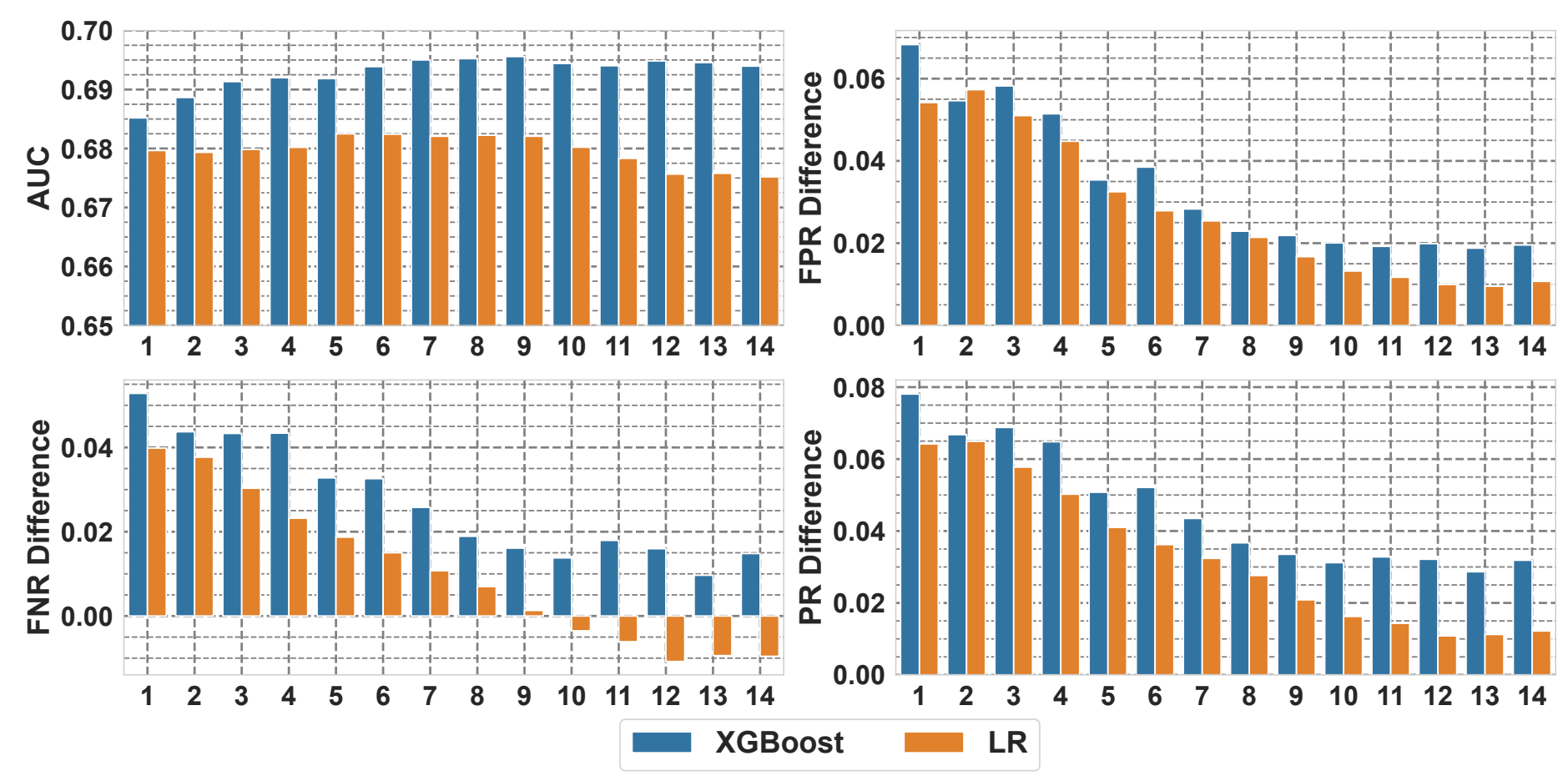
Summary of Key Observations

- More training data does not necessarily lead to a fairer model.
- Base rates and group sizes are not the only determinants of unfairness; the disparity does not necessarily decrease when we balance these between races.
- Depending on the time of training data and when the model is applied, fairness evaluation varies significantly.
- Adding race as an attribute may increase unfairness without increasing accuracy, but adding neighborhood characteristics increases fairness in our experiments.
- For some types of offense, fairness is much worse than other types of offense.
- Training separate models for different races is not always favorable for the minority.
- Data-centric interventions often affect fairness metrics but not accuracy metrics.
- Fairness and accuracy estimates often vary significantly under distribution shift.

Temporal Factors

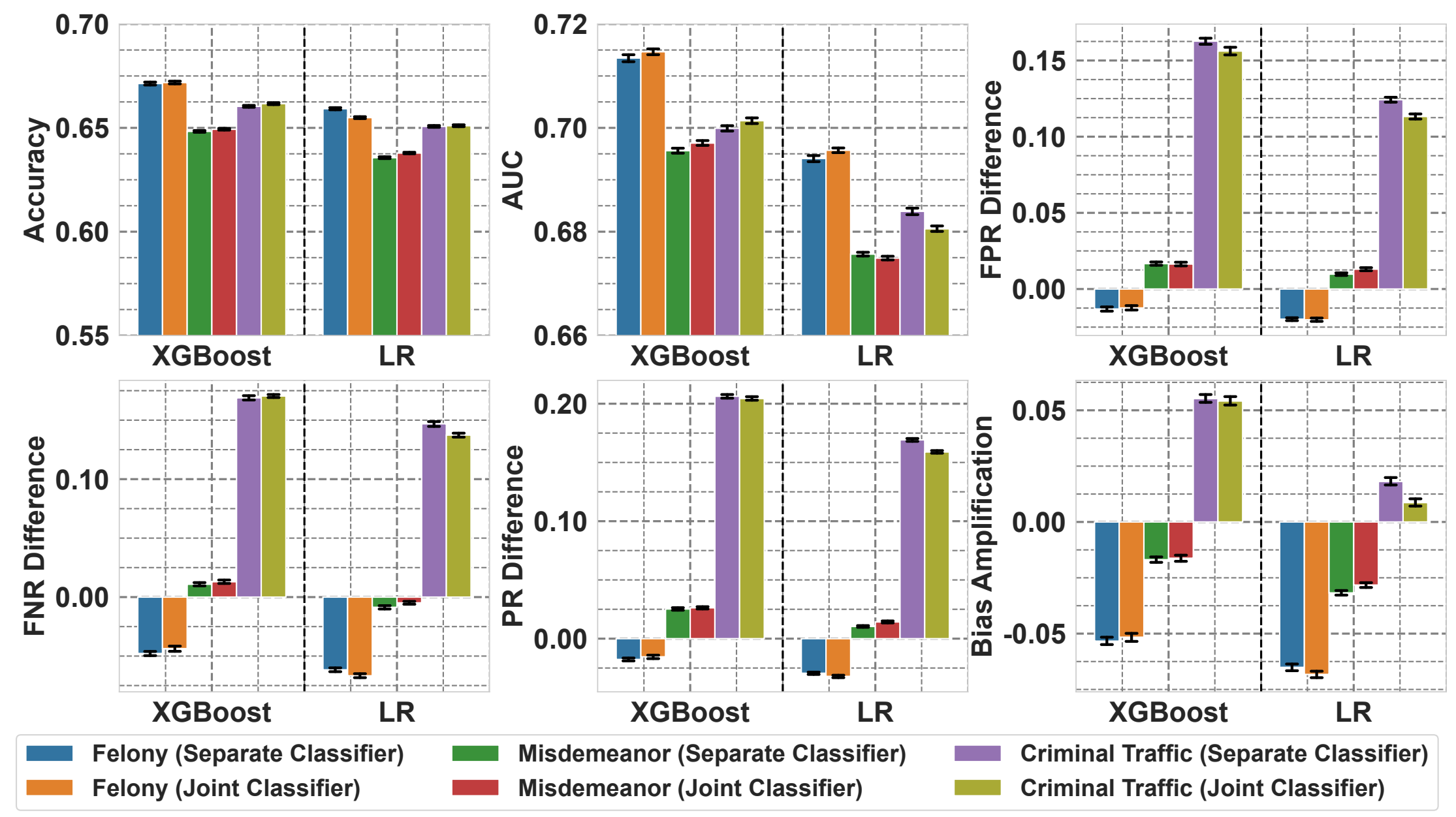


The Role of Time: X axis corresponds to training datasets from two consecutive years between 2000 and 2018 (e.g., "1" on x axis denotes the training data from the years 2000 & 2001, "2" denotes the training data from the years 2001 & 2002 and so on. The test data comes from the next two years after a two year gap (e.g. if training data is from 2000 & 2001, test comes from 2003 & 2004.)



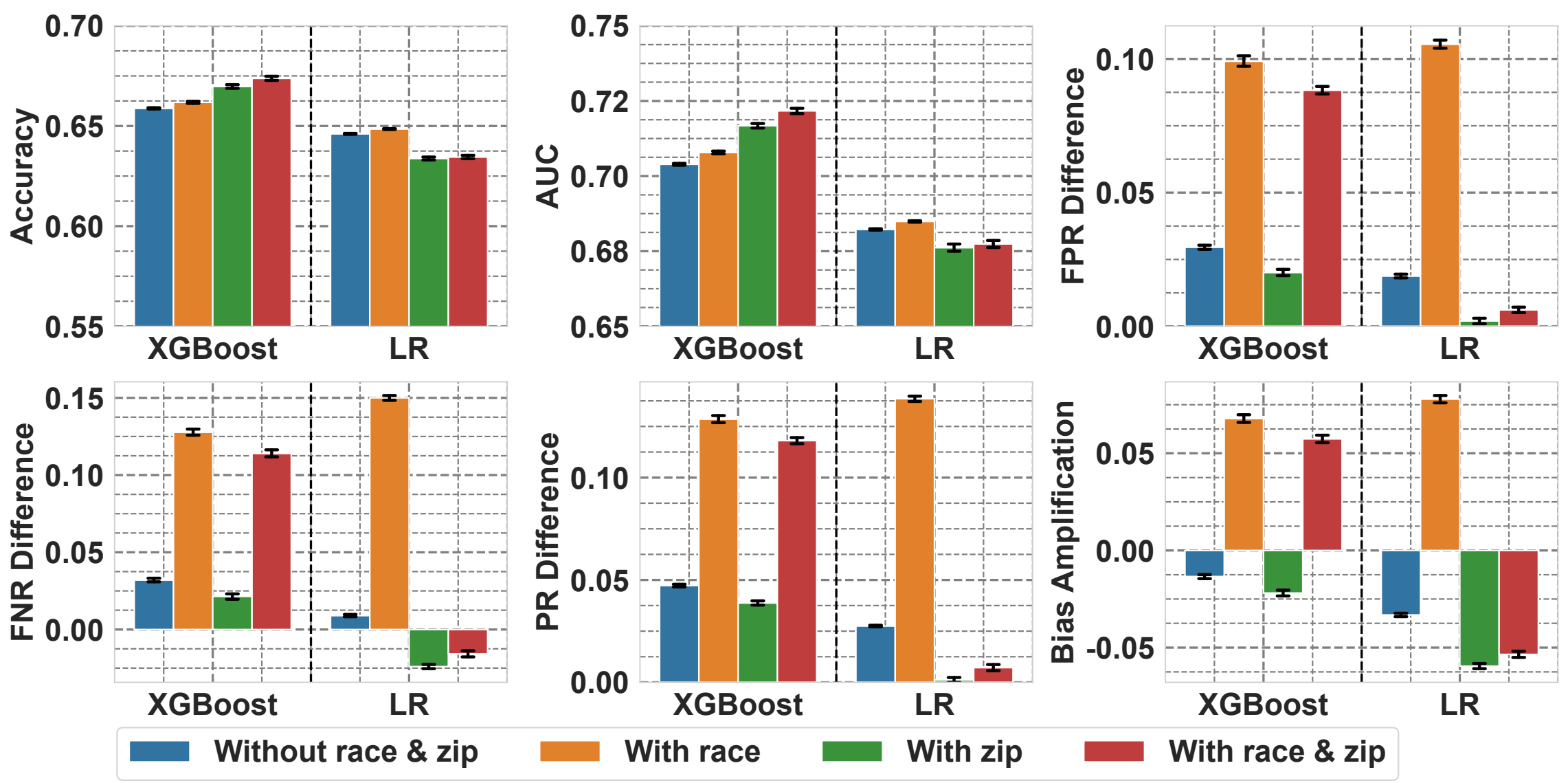
The Role of Time: X axis corresponds to training datasets, as described in the caption of Figure 4 caption. Main difference is that the test data in this figure is the reserved data from all the years between 2000 and 2018.

Type of Offense



A separate classifier for each offense type is trained on data from that offense type. The performance of the classifiers are then observed on respective offense types. For comparison, the performance of a joint classifier, that is trained on all the data and uses offense type as a predictor, is also shown by offense type.

Race and Zipcode Demographic Data



* Zipcode level demographic data (from census):

· Population density · Proportion who attended college · Proportion eligible for food stamp · African American population share · Hispanic population share · Proportion of male · Proportions who live in rural and urban area · Median household income

Concluding Remarks

- Simulation based study of data-centric factors in AI/ML fairness on a new, bigger and more diverse dataset in recidivism prediction.
- Data-centric factors in the context of fairness deserve independent attention in research and in practice. Even when we observed no effect on overall accuracy with different data-centric factors, there still was significant variation in fairness measures.
- The simulation methodology can also be used to study:
  - several other data-centric factors, e.g., biases due to selective labeling, label noise, geographical factors, judge characteristics in past decisions.
  - how various data-centric factors interact with common algorithmic design choices, for example, other types of classifiers, whether the classifier is a deterministic or a randomized one, the type of loss function, type of fairness metric, fairness enforcing techniques etc.
- Other questions for further research:
  - meaning of *good data quality* in the context of AI/ML fairness.
  - appropriate evaluation and benchmarking methods for fairness algorithms.
  - complementary theoretical analysis of data-centric factors.
- For more details and limitations, please read the full paper. Scan the QR Code below to download the full paper in pdf.

ETH zürich