**Definition of fairness in Multiacc:**

We investigate a **notion of fairness – multiaccuracy** – originally proposed in [HKRR18], and develop a framework for auditing and post-processing for multiaccuracy.

Our notion of multiaccuracy **differs** from parity-based notions of fairness, and is reasonable in settings such as gender detection where we would like to boost the classifier’s accuracy across many subgroups.

Indeed, weaknesses of group notions of fairness were discussed in [DHP+12] (for a somewhat related notion called statistical parity), as a motivation for introducing an individual notion of fairness (see further discussion and comparisons below).

**Other important details:**

Our algorithm, Multiaccuracy Boost, works in any setting where we have black-box access to a predictor and a relatively small set of labeled data for auditing; importantly, this black-box framework allows for improved fairness and accountability of predictions, even when the predictor is minimally transparent.

our work focuses on a **setting**, **adapted from [HKRR18]**, that is common in practice but distinct from much of the other literature on fairness in classification.

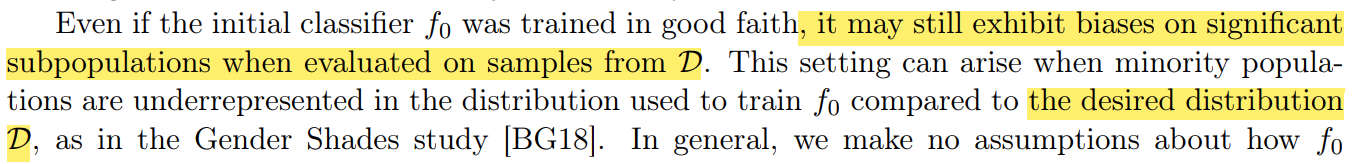
**The setting**

文本, 信件

描述已自动生成

We are given black-box access to a classifier, f0, and a relatively small “validation set" of labeled samples drawn from **some representative distribution D**; our goal is to audit f0 to determine whether the predictor satisfies a strong notion of subgroup fairness, *multiaccuracy*.

**Multiaccuracy requires** (in a sense that we make formal in Section 2) that predictions be unbiased, not just overall, **but on every identifiable subpopulation**. If auditing reveals that the predictor does not satisfy multiaccuracy, we aim to post-process f0 to produce a new classifier f that is multiaccurate, **without adversely affecting the subpopulations where f0 was already accurate at.**



**D should have desired distribution..**

Indeed, the influential work on “Fairness Through Awareness” [DHP+12], followed by [KNRW17,HKRR18], demonstrated the **weakness of statistical notions of fairness** (such as statistical parity, **equalized odds**, and calibration), showing that prediction systems **can exhibit material forms of discrimination against protected populations, even though they satisfy statistical fairness conditions.** Left unaddressed, such forms of discrimination may discourage the participation of minority populations, leading to further underrepresentation of these groups. Thus, our goal will be to mitigate systematic biases broadly enough to handle inadvertent and malicious forms of discrimination