Measuring Model Biases in the Absence of Ground Truth

Osman Aka[*]
Google
U.S.A.
osmanaka@google.com

Ken Burke[*]
Google
U.S.A.
kenburke@google.com

Alex Bäuerle[†]
Ulm University
Germany

Christina Greer
Google
U.S.A.

Margaret Mitchell[‡]
Ethical AI LLC
U.S.A.

ABSTRACT

The measurement of bias in machine learning often focuses on model performance across identity subgroups (such as *man* and *woman*) with respect to groundtruth labels [15]. However, these methods do not directly measure the *associations* that a model may have learned, for example *between* labels and identity subgroups. Further, measuring a model’s bias requires a fully annotated evaluation dataset which may not be easily available in practice.

We present an elegant mathematical solution that tackles both issues simultaneously, using image classification as a working example. By treating a classification model’s predictions for a given image as a set of labels analogous to a “bag of words” [17], we rank the biases that a model has learned with respect to different identity labels. We use {*man*, *woman*} as a concrete example of an identity label set (although this set need not be binary), and present rankings for the labels that are most biased towards one identity or the other. We demonstrate how the statistical properties of different association metrics can lead to different rankings of the most “gender biased" labels, and conclude that normalized pointwise mutual information (*nPMI*) is most useful in practice. Finally, we announce an open-sourced *nPMI* visualization tool using TensorBoard.

CCS CONCEPTS

• **General and reference** → **Measurement**; **Metrics**; • **Applied computing** → *Document analysis.*

KEYWORDS

datasets, image tagging, fairness, bias, stereotypes, information extraction, model analysis

ACM Reference Format:
Osman Aka, Ken Burke, Alex Bäuerle, Christina Greer, and Margaret Mitchell. 2021. Measuring Model Biases in the Absence of Ground Truth. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3461702.3462557

1 INTRODUCTION

The impact of algorithmic bias in computer vision models has been well-documented [c.f., 5, 24]. Examples of the negative fairness impacts of machine learning models include decreased pedestrian detection accuracy on darker skin tones [29], gender stereotyping in image captioning [6], and perceived racial identities impacting unrelated labels [27]. Many of these examples are directly related to currently deployed technology, which highlights the urgency of solving these fairness problems as adoption of these technologies continues to grow.

Many common metrics for quantifying fairness in machine learning models, such as Statistical Parity [11], Equality of Opportunity [15] and Predictive Parity [8], rely on datasets with a significant amount of ground truth annotations for each label under analysis. However, some of the most commonly used datasets in computer vision have relatively sparse ground truth [19]. One reason for this is the significant growth in the number of predicted labels. The benchmark challenge dataset PASCAL VOC introduced in 2008 had only 20 categories [12], while less than 10 years later, the benchmark challenge dataset ImageNet provided hundreds of categories [22]. As systems have rapidly improved, it is now common to use the full set of ImageNet categories, which number more than 20,000 [10, 26].

While large label spaces offer a more fine-grained ontology of the visual world, they also increase the cost of implementing groundtruth-dependent fairness metrics. This concern is compounded by the common practice of collecting training datasets from multiple online resources [20]. This can lead to patterns where specific labels are omitted in a biased way, either through human bias (e.g., crowdsourcing where certain valid labels or tags are omitted systematically) or through algorithmic bias (e.g., selecting labels for human verification based on the predictions of another model [19]). If the ground truth annotations in a sparsely labelled dataset are potentially biased, then the premise of a fairness metric that “normalizes" model prediction patterns to groundtruth patterns may be incomplete. In light of this difficulty, we argue that it is important

[*] Equal contribution.
[†] Work conducted during internship at Google.
[‡] Work conducted while author was at Google.

AIES ’21, May 19–21, 2021, Virtual Event, USA.

ACM ISBN 978-1-4503-8473-5/21/05.
https://doi.org/10.1145/3461702.3462557

to develop bias metrics that do not explicitly rely on "unbiased" ground truth labels.

In this work, we introduce a novel approach for measuring problematic model biases, focusing on the associations between model predictions directly. This has several advantages compared to common fairness approaches in the context of large label spaces, making it possible to identify biases after the regular practice of running model inference over a dataset. We study several different metrics that measure associations between labels, building upon work in Natural Language Processing [9] and information theory. We perform experiments on these association metrics using the Open Images Dataset [19] which has a large enough label space to illustrate how this framework can be generally applied, but we note that the focus of this paper is on introducing the relevant techniques and do not require any specific dataset. We demonstrate that normalized pointwise mutual information (*nPMI*) is particularly useful for detecting the associations between model predictions and sensitive identity labels in this setting, and can uncover stereotype-aligned associations that the model has learned. This metric is particularly promising because:

- It requires no ground truth annotations.
- It provides a method for uncovering biases that the model itself has learned.
- It can be used to provide insight into per-label associations between model predictions and identity attributes.
- Biases for both low- and high-frequency labels are able to be detected and compared.

Finally we announce an open-sourced visualization tool in TensorBoard that allows users to explore patterns of label bias in large datasets using the *nPMI* metric.

2 RELATED WORK

In 1990, Church and Hanks [9] introduced a novel approach to quantifying associations between words based on mutual information [13, 23] and inspired by psycholinguistic work on word norms [28] that catalogue words that people closely associate. For example, subjects respond more quickly to the word *nurse* if it follows a highly associated word such as *doctor* [7, 14]. Church and Hanks' proposed metric applies mutual information to words using a *pointwise approach*, measuring co-occurrences of distinct word pairs rather than averaging over all words. This enables a quantification of the question, "How closely related are these words?" by measuring their co-occurrence rates relative to chance in the dataset of interest. In this case, the dataset of interest is a computer vision evaluation dataset, and the words are the labels that the model predicts.

This information theoretic approach to uncovering word associations became a prominent method in the field of Natural Language Processing, with applications ranging from measuring topic coherence [1] to collocation extraction [4] to great effect, although often requiring a good deal of preprocessing in order to incorporate details of a sentence's syntactic structure. However, without preprocessing, this method functions to simply measure word associations regardless of their order in sentences or relationship to one another, treating words as an unordered set of tokens (a so-called "bag-of-words") [16].

As we show, this simple approach can be newly applied to an emergent problem in the machine learning ethics space: The identification of problematic associations that an ML model has learned. This approach is comparable to measuring correlations, although the common correlation metric of Spearman Rank [25] operates on assumptions that are not suitable for this task, such as linearity and monotonicity. The related correlation metric of the Kendall Rank Correlation [18] does not require such behavior, and we include comparisons with this approach.

Additionally, many potentially applicable metrics for this problem rely on simple counts of paired words, which does not take into consideration how the words are distributed with other words (e.g., sentence syntax or context); we will elaborate on how this information can be formally incorporated into a bias metric in the Discussion and Future Work sections.

This work is motivated by recent research on fairness in machine learning (e.g., [15]), which at a high level seeks to define criteria that result in equal outcomes across different subpopulations. The focus in this paper is complementary to previous fairness work, honing in on ways to identify and quantify the specific problematic associations that a model may learn rather than providing an overall measurement of a model's unfairness. It also offers an alternative to fairness metrics that rely on comprehensive ground truth labelling, which is not always available for large datasets.

Open Images Dataset. The Open Images dataset we use in this work was chosen because it is open-sourced, with millions of diverse images and a large label space of thousands of visual concepts (see [19] for more details). Furthermore, the dataset comes with pre-computed labels generated by a non-trivial algorithm that combines machine-generated predictions and human-verification; this allowed us to focus on analysis of label associations (rather than training a new classifier ourselves) and uncover the most common concepts related to sensitive characteristics, which in this dataset are *man* and *woman*.

We now turn to a formal description of the problem we seek to solve.

3 PROBLEM DEFINITION

We have a dataset $\mathcal{D}$ which contains image examples and labels generated by a image classifier. This classifier takes one image example and predicts "Is label y_i relevant to the image?" for each label in $\mathcal{L} = \{y_1, y_2, ..., y_n\}$[1]. We infer $P(y_i)$ and $P(y_i, y_j)$ from $\mathcal{D}$ such that for a given random image in $\mathcal{D}$, $P(y_i)$ is the probability of having y_i as positive prediction and $P(y_i, y_j)$ is the joint probability of having both y_i and y_j as positive predictions. We further assume that we have identity labels $x_1, x_2, ...x_n \in \mathcal{L}$ that belong to some sensitive identity group for which we wish to compute a bias metric (e.g., *man* and *woman* as labels for gender)[2]. These identity labels may even be generated by the same process that generates the labels y, as in our case where *man*, *woman* and all other y are elements of $\mathcal{L}$ predicted by the classifier. We then measure bias with respect

[1] For ease of notation, we use y and x rather than $\hat{y}$ and $\hat{x}$.

[2] For the rest of the paper, we focus on only two identity labels with notation x_1 and x_2 for simplicity, however the identity labels need not be binary (or one-dimensional) in this way. The remainder of this work is straightforwardly extended to any number of identity labels by using the pairwise comparison of all identity labels, or by using a one-vs-all gap (e.g., $A(x, y) - \mathbb{E}\left[A(x', y)\right]$ where x' is the set of all other x).

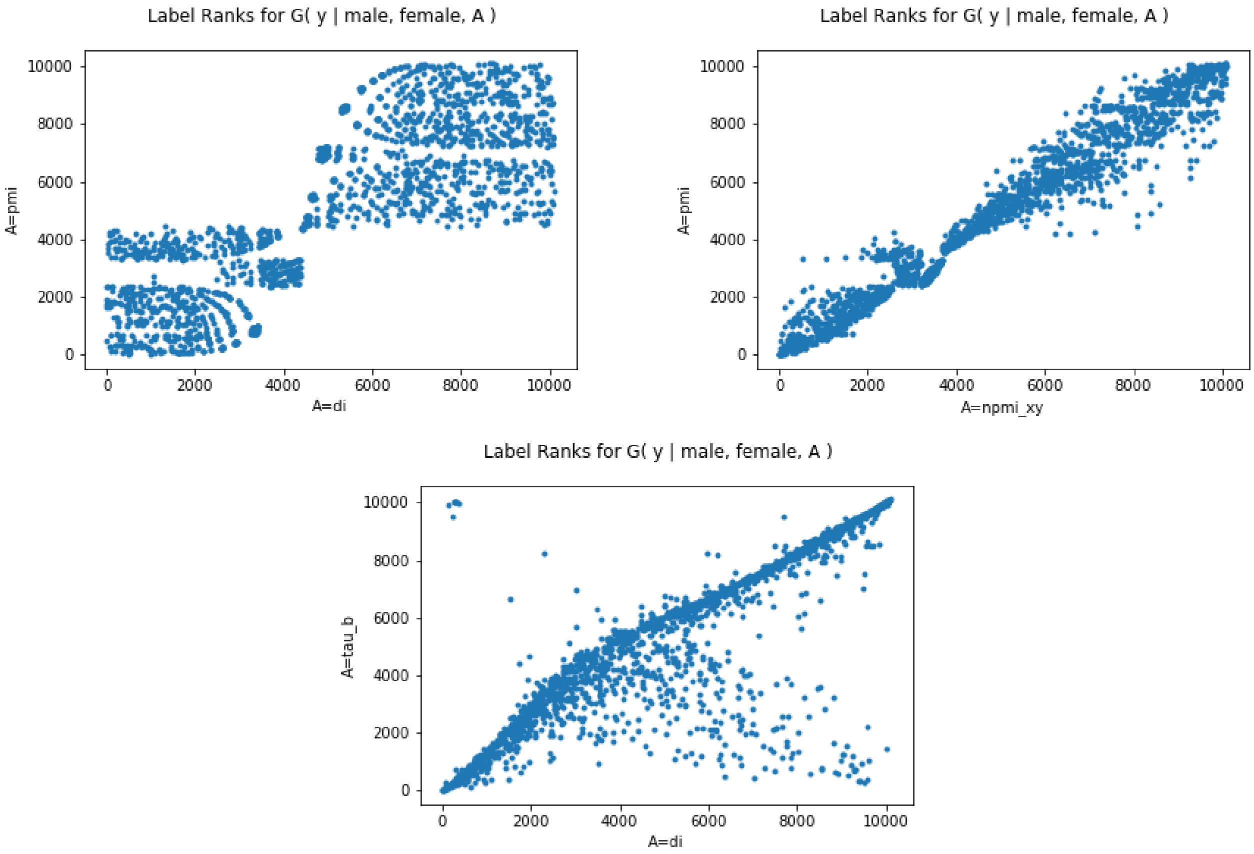


Figure 1: Label ranking shifts for metric-to-metric comparison. Each point represents the rank of a single label when sorted by $G(y|x_1, x_2, A(\cdot))$ (i.e., the label rank by gap). The coordinates represent the rankings by gap for different association metrics $A(\cdot)$ on the x and y axes. A highly-correlated plot along $y = x$ would imply that the two metrics lead to very similar bias rankings according to $G(\cdot)$.

to these identity labels for all other labels $y \in \mathcal{L}$. As the size of this label space $|\mathcal{L}|$ approaches tens of thousands and increases year by year for modern machine learning models, it is important to have a simple bias metric that can be computed and reasoned about at scale.

For ease of discussion in the rest of this paper, we denote any generic association metric $A(x_j, y)$, where x_j is an identity label and y is any other label. We define an *association gap* for label y between two identity labels $[x_1, x_2]$ with respect to the association metric $A(x_j, y)$ as $G(y|x_1, x_2, A(\cdot)) = A(x_1, y) - A(x_2, y)$. For example, the association between the labels *woman* and *bike*, $A(woman, bike)$ can be compared to the association between the labels *man* and *bike*, $A(man, bike)$. The difference between them is the **association gap** for the label *bike*:

$$G(bike|woman, man, A(\cdot)) = \\ A(woman, bike) \text{ - } A(man, bike)$$

We use this association gap $G(\cdot)$ as a measurement of "bias" or "skew" of a given label across sensitive identity subgroups.

The first objective we are interested in is the question "Is the prediction of label y biased towards either x_1 or x_2?". The second objective is then ranking the labels y by the size of this bias. If x_1 and x_2 both belong to the same identity group (e.g., *man* and *woman* are both labels for "gender"), then one may consider these measurements to approximate the *gender bias*.

We choose this gender example because of the abundance of these specific labels in the Open Images Dataset, however this choice should not be interpreted to mean that gender representation is one-dimensional, nor that paired labels are required for the general[1] approach. Nonetheless, this simplification is important because it allows us to demonstrate how a single per-label approximation of "bias" can be measured between paired labels, and we leave details of further expansions, including calculations across multiple sensitive identity labels, to the Discussion section.

3.1 Association Metrics

We consider several sets of related association metrics $A(\cdot)$ that can be applied given the constraints of the problem at hand – limited groundtruth, non-linearity, and limited assumptions about the underlying distribution of the data. All of these metrics share in common the general intuition of measuring how labels associate with each other in a dataset, but as we will demonstrate, they yield very different results due to differences in how they quantify this notion of "association".

We first consider fairness metrics as types of association metrics. One of the most common fairness metrics, Demographic (or Statistical) Parity [3, 11, 15], a quantification of the legal doctrine of Disparate Impact [2], can be applied directly for the given task constraints.[3] Other metrics that are possible to adopt for this task include those based on Intersection-over-Union (IOU) measurements, and metrics based on correlation and statistical tests. We next describe these metrics in further detail and their relationship to the task at hand. In summary, we compare the following families of metrics:

- Fairnesss: Demographic Parity (*DP*)
- Entropy: Pointwise Mutual Information (*PMI*), Normalized Pointwise Mutual Information (*nPMI*).
- IOU: Sørensen-Dice Coefficient (*SDC*), Jaccard Index (*JI*).
- Correlation and Statistical Tests: Kendall Rank Correlation (τ_b), Log-Likelihood Ratio (*LLR*), t-test.

One of the important aspects of our problem setting is the counts of images with labels and label intersections, i.e., $C(y)$, $C(x_1, y)$, and $C(x_2, y)$. These values can span a large range for different labels y in the label set $\mathcal{L}$, depending on how common they are in the dataset. Some metrics are theoretically more sensitive to the frequencies/counts of the label y as determined by their nonzero partial derivatives with respect to $P(y)$ (see Table 2). However, as we further discuss in the Experiments and Discussion sections, our experiments indicate that in practice, metrics with non-zero partial derivatives are surprisingly better able to capture biases across a range of label frequencies than metrics with a zero partial derivative. Differential sensitivity to label frequency could be problematic in practice for two reasons:

(1) It would not be possible to compare $G(y|x_1, x_2, A(\cdot))$ *between* different labels y with different marginal frequencies (counts) $C(y)$. For example, the ideal bias metric should be able to capture gender bias equally well for both *car* and *Nissan Maxima* even though the first label is more common than the second.
(2) The alternative, bucketizing labels by marginal frequency and setting distinct thresholds per bucket, would add significantly more hyperparameters and essentially amount to manual frequency-normalization.

The following sections contain basic explanations of these metrics for a general audience, with the running example of *bike*, *man*, and *woman*. We leave further mathematical analyses of the metrics to the Appendix. However, integral to the application of *nPMI* in this task is the choice of normalization factor, and so we discuss this in further detail in the Normalizing *PMI* subsection.

Demographic Parity

$$G(y|x_1, x_2, DP) = P(y|x_1) - P(y|x_2)$$

Demographic Parity focuses on differences between the conditional probability of y given x_1 and x_2: How likely *bike* is for *man* vs *woman*.

Entropy

$$G(y|x_1, x_2, PMI) = ln\left(\frac{P(x_1, y)}{P(x_1)P(y)}\right) - ln\left(\frac{P(x_2, y)}{P(x_2)P(y)}\right)$$

Pointwise Mutual Information, adapted from information theory, is the main entropy-based metric studied here. In this form, we are analyzing the entropy difference between $[x_1, y]$ and $[x_2, y]$. This essentially examines the dependence of, for example, the *bike* label distribution on two other label distributions: *man* and *woman*.

Remaining Metrics

We use the Sørensen-Dice Coefficient (*SDC*), which has the commonly-used F1-score as one of its variants; the Jaccard Index (*JI*), a common metric in Computer Vision also known as Intersection Over Union (IOU); Log-Likelihood Ratio (*LLR*), a classic flexible comparison approach; Kendall Rank Correlation, which is also known as τ_b-correlation, and is the particular *correlation* method that can be reasonably applied in this setting; and the t-$test$, a common statistical significance test that can be adapted in this setting [17]. Each of these metrics have different behaviours, however, we limit our mathematical explanation to the Appendix, as we found these metrics are either less useful in practice or behave similarly to other metrics in this use case.

3.2 Normalizing PMI

One major challenge in our problem setting is the sensitivity of these association metrics to the frequencies/counts of the labels in $\mathcal{L}$. Some metrics are weighted more heavily towards common labels (i.e., large marginal counts, $C(y)$) in spite of differences in their joint probabilities with identity labels ($P(x_1, y), P(x_2, y)$). The opposite is true for other metrics, which are weighted towards rare labels with smaller marginal frequencies. In order to compensate for this problem, several different normalization techniques have been applied to *PMI* [4, 21]. Common normalizations include:

- $nPMI_y$: Normalizing each term by $P(y)$.
- $nPMI_{xy}$: Normalizing the two terms by $P(x_1, y)$ and $P(x_2, y)$, respectively.
- PMI^2: Using $P(x_1, y)^2$ and $P(x_2, y)^2$ instead of $P(x_1, y)$ and $P(x_2, y)$, the normalization effects of which are further illustrated in the Appendix.

Each of these normalization methods have different impacts on the *PMI* metric. The main advantage of these normalizations is the ability to compare association gaps *between* label pairs $[y_1, y_2]$ (e.g., comparing the gender skews of two labels like *Long Hair* and *Dido Flip*) even if $P(y_1)$ and $P(y_2)$ are very different. In the Experiments section, we discuss which of these is most effective and meaningful for the fairness and bias use case motivating this work.

4 EXPERIMENTS

In order to compare these metrics, we use the Open Images Dataset (OID) [19] described above in the Related Works section. This

[3] Other common fairness metrics, such as Equality of Opportunity, require both a model prediction and a groundtruth label, which makes the correct way to apply them to this task less clear. We leave this for further work.

dataset is useful for demonstrating realistic bias detection use cases because of the number of distinct labels that may be applied to the images (the dataset itself is annotated with nearly 20,000 labels). The label space is also diverse, including objects (*cat*, *car*, *tree*), materials (*leather*, *basketball*), moments/actions (*blowing out candles*, *woman playing guitar*), and scene descriptors and attributes (*beautiful*, *smiling*, *forest*). We seek to measure the gender bias as described above for each of these labels, and compare these bias values directly in order to determine which of the labels are most skewed towards associating with the labels *man* and *woman*.

In our experiments, we apply each of the association metrics $A(x_j, y)$ to the machine-generated predictions for identity labels $x_1 = man, x_2 = woman$ and all other labels y in the Open Images Dataset. We then compute the gap value $G(\cdot)$ between the identity labels for each label y and sort, providing a ranking of labels that are most biased towards x_1 or x_2. As we will show, sorting labels by this *association gap* creates different rankings for different association metrics. We examine which labels are ranked within the top 100 for the different association metrics in the Top 100 Labels by Metric Gaps subsection.

4.1 Label Ranks

The first experiment we performed is to compute the association metrics and the gaps between them for different labels – $A(x_1, y)$, $A(x_2, y)$ and $G(y|x_1, x_2, A(\cdot))$ – over the OID dataset. We then sorted the labels by $G(\cdot)$ and studied how the ranking by gap differed between different association metrics $A(\cdot)$. Figure 1 shows examples of these metric-to-metric comparisons of label rankings by gap (all other metric comparisons can be found in the Appendix). We can see that a single label can have quite a different ranking depending on the metric.

When comparing metrics, we found that they grouped together in a few clusters based on similar ranking patterns when sorting by $G(y|man, woman, A(\cdot))$. In the first cluster, pairwise comparisons between *PMI*, *PMI*^{2} and *LLR* show linear relationships when sorting labels by $G(\cdot)$. Indeed, while some labels show modest changes in rank between these metrics, they share *all* of their top 100 labels, and > 99% of label pairs maintain the same relative ranking between metrics. By contrast, there are only 7 labels in common between the top 100 labels of *PMI*^{2} and *SDC*. Due to the similar behavior of this cluster of metrics, we chose to focus on *PMI* as representative of these 3 metrics moving forward (see the Appendix for further details on these relationships).

Similar results were obtained for another cluster of metrics: *DP*, *JI*, and *SDC*. All pairwise comparisons generated linear plots, with about 70% of the top 100 labels shared in common between these metrics when sorted by $G(\cdot)$. Furthermore, about 95% of pairs of those overlapping labels maintained the same relative ranking between metrics. Similar to the *PMI* cluster, we chose to focus on Demographic Parity (*DP*) as the representative metric from its cluster, due to its mathematical simplicity and prominence in fairness literature.

We next sought to understand how incremental changes to the counts of the labels, $C(y)$, affect these association gap rankings in a real dataset (see Appendix Section ??). To achieve this, we added a fake label to the real labels of OID, setting initial values for its counts and co-occurences in the dataset, $P(y)$, $P(x_1, y)$, and $P(x_2, y)$. Then we incrementally increased or decreased the count of label y, and measured whether its bias ranking in G would change relative to the other labels in OID. We repeated this procedure for different orders of magnitude of label count $C(y)$ while maintaining the ratio $P(x_1, y)/P(x_2, y)$ as constant.

Metrics	Min/Max $C(y)$	Min/Max $C(x_1, y)$	Min/Max $C(x_2, y)$
PMI	15 / 10,551	1 / 1,059	8 / 7,755
PMI^2	15 / 10,551	1 / 1,059	8 / 7,755
LLR	15 / 10,551	1 / 1,059	8 / 7,755
DP	6,104 / 785,045	628 / 239,950	5,347 / 197,795
JI	4,158 / 562,445	399 / 144,185	3,359 / 183,132
SDC	2,906 / 562,445	139 / 144,185	2,563 / 183,132
$nPMI_y$	35 / 562,445	1/144,185	9 / 183,132
$nPMI_{xy}$	34 / 270,748	1 / 144,185	20 / 183,132
τ_b	6,104 / 785,045	628 / 207,723	5,347 / 183,132
t-$test$	960 / 562,445	72 / 144,185	870 / 183,132

Table 1: Minimum and maximum counts $C(y)$ of the top 100 labels with the largest association gaps for each metric. Note that these min/max values for $C(y)$ vary by orders of magnitude for different metrics. A larger version of this table is available in Appendix, Table ??.

This experiment allowed us to determine whether the theoretical sensitivities of each metric to label frequency $P(y)$, as determined by partial derivatives $\partial A(\cdot)/\partial P(y)$ (see Table 2), would hold in the context of real-world data, where the underlying distribution of label frequencies may not be uniform. If certain subregions of the label distribution are relatively sparse, for example, then the gap ranking of our hypothetical label may not change even if $\partial A(\cdot)/\partial C(y) \neq 0$. However, in practice we do not observe this behavior in the tested settings (see Appendix for plots of these experiments), where label rank moves with label count roughly as predicted by the partial derivatives in Table 2. In fact, we observed that metrics with larger partial derivatives for x_1, x_2, or y often led to a larger change in rank. For example, slightly increasing $P(x_1, y)$ when y always co-occurs with x_1, $P(y) = P(x_1, y)$ affects ranking more for $A = nPMI_y$ compared to $A = PMI$ (see Appendix).

4.2 Top 100 Labels by Metric Gaps

When applying these metrics to fairness and bias use cases, model users may be most interested in surfacing the labels with the largest association gaps. If one filters results to a "top K" label set, then the normalization chosen could lead to vastly different sets of labels

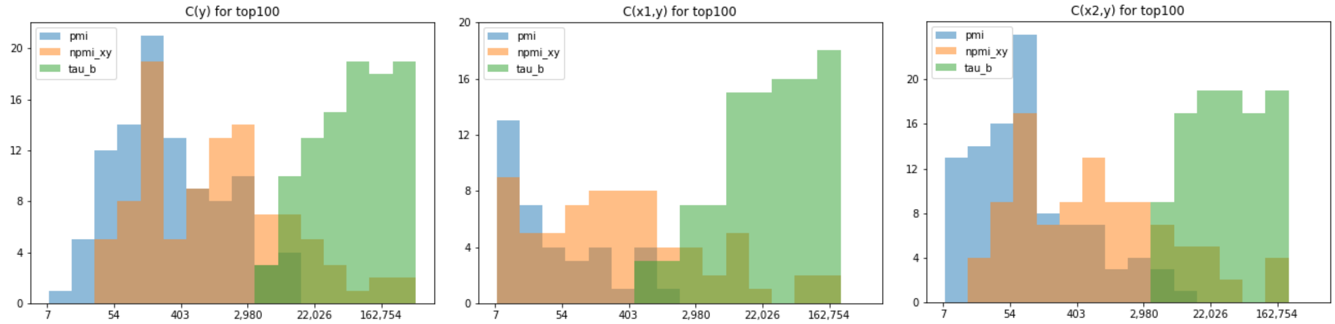


Figure 2: Top 100 count distributions for PMI, $nPMI_{xy}$, and τ_b

The distribution of $C(y)$, $C(x_1, y)$, and $C(x_2, y)$ for the top 100 labels sorted by gap $G(y|x_1, x_2, A(\cdot))$ for PMI, $nPMI_{xy}$, and τ_b. The x-axis is the logarithmic-scaled bins and the y-axis is the number of labels which have the corresponding count values in that bin.

	$\partial p(y)$	$\partial p(x_1, y)$	$\partial p(x_2, y)$
∂DP	0	$\frac{1}{p(x_1)}$	$\frac{-1}{p(x_2)}$
∂PMI	0	$\frac{1}{p(x_1,y)}$	$\frac{-1}{p(x_2,y)}$
$\partial nPMI_y$	$\frac{ln(\frac{p(x_2\|y)}{p(x_1\|y)})}{ln^2(p(y))p(y)}$	$\frac{1}{ln(p(y))p(x_1,y)}$	$\frac{-1}{ln(p(y))p(x_2,y)}$
$\partial nPMI_{xy}$	$\frac{1}{ln(p(x_1,y))p(y)} - \frac{1}{ln(p(x_2,y))p(y)}$	$\frac{ln(p(y))-ln(p(x_1))}{ln^2(p(x_1,y))p(x_1,y)}$	$\frac{ln(p(x_2))-ln(p(y))}{ln^2(p(x_2,y))p(x_2,y)}$
∂PMI^2	0	$\frac{2}{p(x_1,y)}$	$\frac{-2}{p(x_2,y)}$
∂SDC	*see Appendix* **??**	$\frac{1}{p(x_1)+p(y)}$	$\frac{-1}{p(x_2)+p(y)}$
∂JI	*see Appendix* **??**	$\frac{p(x_1)+p(y)}{(p(x_1)+p(y)-p(x_1,y))^2}$	$\frac{p(x_2)+p(y)}{(p(x_2)+p(y)-p(x_2,y))^2}$
∂LLR	0	$\frac{1}{p(x_1,y)}$	$\frac{-1}{p(x_2,y)}$
$\partial\tau_b$	*see Appendix* **??**	$\frac{(2-\frac{4}{n})}{\sqrt{(p(x_1)-p(x_1)^2)(p(y)-p(y)^2)}}$	$\frac{(\frac{4}{n}-2)}{\sqrt{(p(x_2)-p(x_2)^2)(p(y)-p(y)^2)}}$
$\partial t\text{-}test\_gap$	$\frac{\sqrt{p(x_2)}-\sqrt{p(x_1)}}{2\sqrt{p(y)}}$	$\frac{1}{\sqrt{p(x_1)p(y)}}$	$\frac{-1}{\sqrt{p(x_2)p(y)}}$

Table 2: Metric orientations.

This table shows the partial derivatives of the metrics with respect to $P(y)$, $P(x_1, y)$, and $P(x_2, y)$. This provides a quantification of the theoretical sensitivity of the metrics for different probability values of $P(y)$, $P(x_1, y)$, and $P(x_2, y)$. We see similar experimental results for the metrics with similar orientations (see Appendix).

(e.g., as mentioned earlier, PMI^2 and SDC only shared 7 labels in their top 100 set for OID).

To further analyze this issue, we calculated simple values for each metric's top 100 labels sorted by $G(\cdot)$: minimum and maximum values of $C(y)$, $C(x_1, y)$ and $C(x_2, y)$ as shown in Table 1. The most salient point is that the clusters of metrics from the Label Ranks subsection also appear to hold in this analysis as well; PMI, PMI^2, and LLR have low $C(y)$, $C(x_1, y)$ and $C(x_2, y)$ ranges, whereas DP, JI, and SDC have relatively high ranges. Another straightforward observation we can make is that the $nPMI_y$ and $nPMI_{xy}$ ranges are much broader than the first two clusters, and include the other metrics' ranges especially for the joint co-occurrences, $C(x_1, y)$ and $C(x_2, y)$.

To demonstrate this point more clearly, we plot the distributions of these counts for PMI, $nPMI_{xy}$ and τ_b skews in Figure 2 (all other combinations can be found in the Appendix). These three metric distributions show that gap calculations based on PMI (blue distribution) exclusively rank labels with low counts in the top 100 most skewed labels, where τ_b calculations (green distribution) almost exclusively rank labels with much higher counts. The exception is $nPMI_y$ and $nPMI_{xy}$ (orange distribution); these two metrics are capable of capturing labels across a range of marginal frequencies. In other words, ranking labels by PMI gaps is likely to highlight *rare labels*, ranking by τ_b will highlight *common labels*, and ranking by $nPMI_{xy}$ will highlight *both rare and common labels*.

An example of this relationship between association metric choice and label commonality can be seen in Table 3 (note, here we use Demographic Parity instead of τ_b because they behave similarly in this respect). In this table, we show the "Top 15 labels" most heavily skewed towards *woman* relative to *man* according to DP, unnormalized PMI, and $nPMI_{xy}$. DP almost exclusively highlights common labels predicted for over 100,000 images in the dataset (e.g., *Happiness* and *Fashion*), whereas PMI largely highlights rarer labels predicted for less than 1,000 images (e.g., *Treggings* and *Boho-chic*). By contrast, $nPMI_{xy}$ highlights both common and rare labels (e.g., *Long Hair* as well as *Boho-chic*).

5 DISCUSSION

In the previous section, we first showed that some association metrics behave very similarly when ranking labels from the Open Images Dataset (OID). We then showed that the mathematical orientations and sensitivity of these metrics align with experimental results from OID. Finally, we showed that the different normalizations affect whether labels with high or low marginal frequencies are likely to be detected as having a significant bias according to $G(y|x_1, x_2, A(\cdot))$ in this dataset. We arrive at the conclusion that the $nPMI$ metrics are preferable to other commonly used association metrics in the problem setting of detecting biases without groundtruth labels.

What is the intuition behind this particular association metric as a bias metric? All of the studied entropy-based metrics (gaps in $nPMI$ as well as PMI) approximately correspond to whether one identity label x_1 co-occurs with the target label y more often than another identity label x_2 **relative to chance levels**. This chance-level normalization is important because even completely unbiased labels would still co-occur at some baseline rate by chance alone.

The further normalization of PMI by either the marginal or joint probability ($nPMI_y$ and $nPMI_{xy}$, respectively) takes this one step further in practice by surfacing labels with larger marginal counts at higher ranks in $G(y|x_1, x_2, nPMI)$ alongside labels with smaller marginal counts. This is a somewhat surprising result, because in theory PMI should already be independent of the marginal frequency of $P(y)$ (because $\partial PMI/\partial p(y) = 0$), whereas this derivative for $nPMI$ is non-zero. When we examined this pattern in practice, the labels with smaller counts can achieve very large $P(x_1, y)/P(x_2, y)$ ratios (and therefore their bias rankings can get very high) merely by reducing the denominator to a single image example. PMI is unable to compensate for this noise, whereas the normalizations we use for $nPMI$ allow us to capture a significant amount of common labels in the top 100 labels by $G(y|x_1, x_2, nPMI)$ in spite of this pattern. This result is indicated by the ranges in Table 1 and Figure 2, as well as the set of both common and rare labels for $nPMI$ in Table 3.

Indeed, if the evaluation set is properly designed to match the distribution of use cases of a classification model "in the wild", then we argue more common labels that have a smaller $P(x_1, y)/P(x_2, y)$ ratio are still critical to audit for biases. Normalization strategies must be titrated carefully to balance this simple ratio of joint probabilities with the label's rarity in the dataset.

An alternative solution to this problem could be bucketing labels by their marginal frequency. We argue this is a suboptimal solution for two reasons. First, determining even a single threshold hyperparameter is a painful process for defining fairness constraints. Systems that prevent models from being published if their fairness discrepancies exceed a threshold would then be required to titrate this threshold for every bucket. Secondly, bucketing labels by frequency is essentially a manual and discontinuous form of normalization; we argue that building normalization into the metric directly is a more elegant solution.

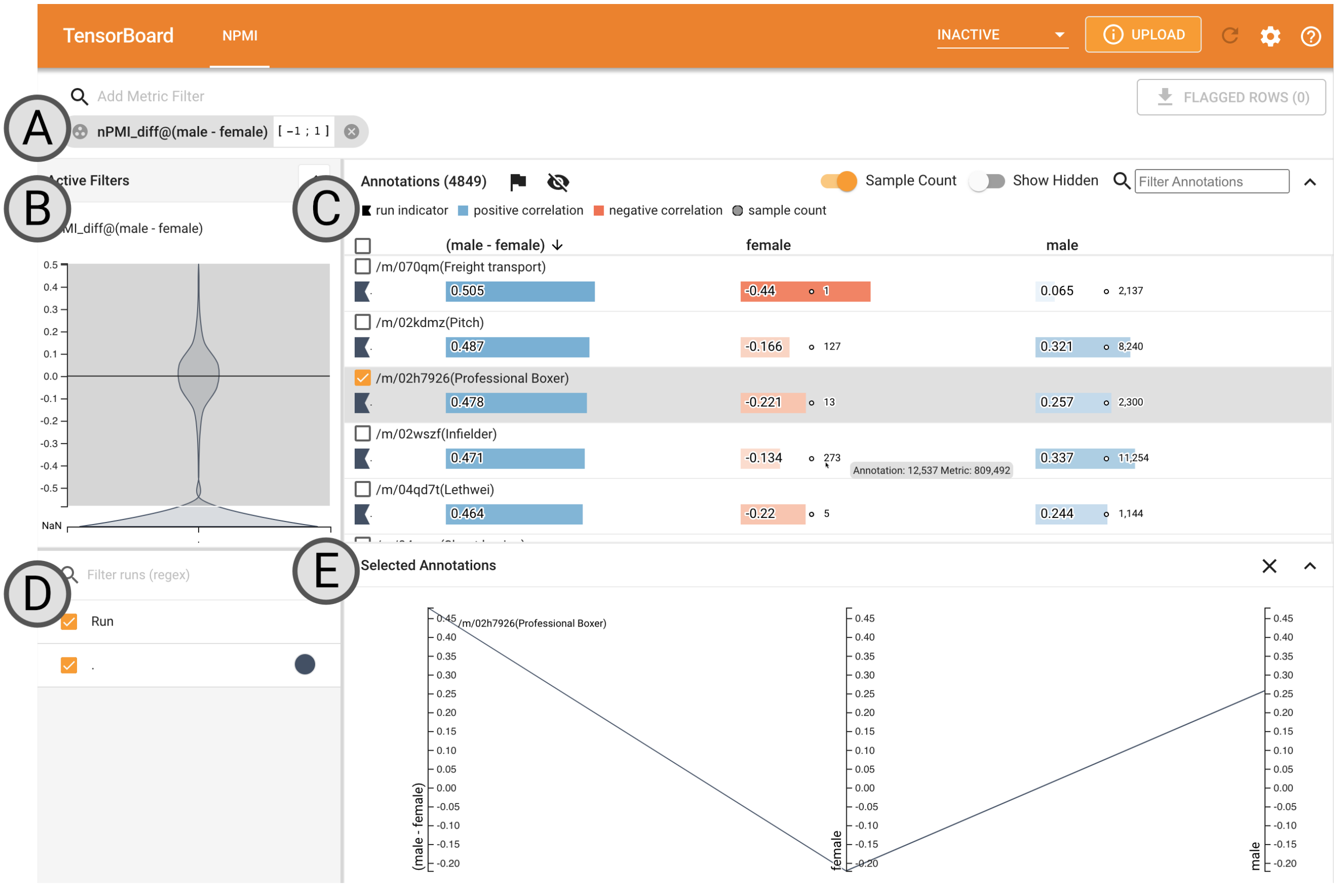
Finally, to enable detailed investigation of the model predictions, we implemented and open-sourced a tool to visualize $nPMI$ metrics as a TensorBoard plugin for developers[4] (see Figure 3). It allows users to investigate discrepancies between two or more identity labels and their pairwise comparisons. Users can visualize probabilities, image counts, sample and label distributions, and filter, flag, and download these results.

[4] https://github.com/tensorflow/tensorboard/tree/master/tensorboard/plugins/npmi

Metric *A*	*DP*		*PMI*		$nPMI_{xy}$	
Ranks	Label *y*	Count	Label *y*	Count	Label *y*	Count
0		265,853	Dido Flip	140		610
1		270,748	Webcam Model	184	Dido Flip	140
2		221,017	Boho-chic	151		2,906
3		166,186		610	Eye Liner	3,144
4	Beauty	562,445	Treggings	126	Long Hair	56,832
5	Long Hair	56,832	Mascara	539	Mascara	539
6	Happiness	117,562		145	Lipstick	8,688
7	Hairstyle	145,151	Lace Wig	70	Step Cutting	6,104
8	Smile	144,694	Eyelash Extension	1,167	Model	10,551
9	Fashion	238,100	Bohemian Style	460	Eye Shadow	1,235
10	Fashion Designer	101,854		78	Photo Shoot	8,775
11	Iris	120,411	Gravure Idole	200	Eyelash Extension	1,167
12	Skin	202,360		165	Boho-chic	460
13	Textile	231,628	Eye Shadow	1,235	Webcam Model	151
14	Adolescence	221,940		156	Bohemian Style	184

Table 3: Top 15 labels skewed towards *woman*.

Ranking by the gap of label y between *woman* x_1 and *man* x_2, according to $G(y|x_1, x_2, A(\cdot))$. Identity label names are omitted.



□	(male - female) ↓	female		male	
□ /m/070qm(Freight transport)	0.505	-0.44	1	0.065	2,137
□ /m/02kdmz(Pitch)	0.487	-0.166	127	0.321	8,240
☑ /m/02h7926(Professional Boxer)	0.478	-0.221	13	0.257	2,300
□ /m/02wszf(Infielder)	0.471	-0.134	273	0.337	11,254
□ /m/04qd7t(Lethwei)	0.464	-0.22	5	0.244	1,144

Annotation: 12,537 Metric: 809,492

Figure 3: Open-sourced tool we implemented for bias investigation.

In **A**, annotations are filtered by an association metric, in this case the *nPMI* difference (*nPMI_gap*) between *male* and *female*. In **B**, distribution of association values. In **C**, the filtered annotations view. In **D**, different models or datasets can be selected for display. In **E**, a parallel coordinates visualization of selected annotations to assess where association differences come from.

6 CONCLUSION AND FUTURE WORK

In this paper we have described association metrics that can measure *gaps* – biases or skews – towards specific labels in the large label space of current computer vision classification models. These metrics do not require ground truth annotations, which allows them to be applied in contexts where it is difficult to apply standard fairness metrics such as Equality of Opportunity [15]. According to our experiments, Normalized Pointwise Mutual Information (*nPMI*) is a particularly useful metric for measuring specific biases in a real-world dataset with a large label space, e.g., the Open Images Dataset.

This paper also introduces several questions for future work. The first is whether *nPMI* is also similarly useful as a bias metric in small label spaces (e.g., credit and loan applications). Second, if we were to have exhaustive ground truth labels for such a dataset, how would the sensitivity of *nPMI* in detecting biases compare to ground-truth-dependent fairness metrics? Finally, in this work we treated the labels predicted for an image as a flat set. However, just like sentences have rich syntactic structure beyond the "bag-of-words" model in NLP, images also have rich structure and relationships between objects that are not captured by mere rates of binary co-occurrence. This opens up the possibility that *within-image label relationships* could be leveraged to better understand how concepts are associated in a large computer vision dataset. We leave these questions for future work.

REFERENCES

[1] Nikolaos Aletras and Mark Stevenson. 2013. Evaluating Topic Coherence Using Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*. Association for Computational Linguistics, Potsdam, Germany, 13–22. https://www.aclweb.org/anthology/W13-0102

[2] Solon Barocas and Andrew D. Selbst. 2014. Big Data's Disparate Impact. *SSRN eLibrary* (2014).

[3] Richard Berk. 2016. A primer on fairness in criminal justice risk assessments. *The Criminologist* 41, 6 (2016), 6–9.

[4] G. Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction.

[5] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification *(Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[6] Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. 2018. Women also Snowboard: Overcoming Bias in Captioning Models. *CoRR* abs/1803.09797 (2018). arXiv:1803.09797 http://arxiv.org/abs/1803.09797

[7] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. https://doi.org/10.1126/science.aal4230 arXiv:https://science.sciencemag.org/content/356/6334/183.full.pdf

[8] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv:1610.07524 [stat.AP]

[9] Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16, 1 (1990), 22–29. https://www.aclweb.org/anthology/J90-1003

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2011. Fairness Through Awareness. *CoRR* abs/1104.3913 (2011). arXiv:1104.3913 http://arxiv.org/abs/1104.3913

[12] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision* 111, 1 (Jan. 2015), 98–136. https://doi.org/10.1007/s11263-014-0733-5

[13] Robert M Fano. 1961. Transmission of information: A statistical theory of communications. *American Journal of Physics* 29 (1961), 793–794.

[14] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74 (1998). Issue 6.

[15] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. arXiv:1610.02413 [cs.LG]

[16] Zellig Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162. https://doi.org/10.1007/978-94-009-8467-7_1

[17] Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition.* Pearson Prentice Hall, Upper Saddle River, N.J. http://www.amazon.com/Speech-Language-Processing-2nd-Edition/dp/0131873210/ref=pd_bxgy_b_img_y

[18] M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1-2 (06 1938), 81–93. arXiv:https://academic.oup.com/biomet/article-pdf/30/1-2/81/423380/30-1-2-81.pdf https://doi.org/10.1093/biomet/30.1-2.81

[19] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *CoRR* abs/1811.00982 (2018). arXiv:1811.00982 http://arxiv.org/abs/1811.00982

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR* abs/1405.0312 (2014). arXiv:1405.0312 http://arxiv.org/abs/1405.0312

[21] FranÇois Role and Mohamed Nadif. 2011. Handling the Impact of Low Frequency Events on Co-Occurrence Based Measures of Word Similarity - A Case Study of Pointwise Mutual Information. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval - KDIR, (IC3K 2011)*. SciTePress, 218–223.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[23] Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 3 (1948), 379–423. http://dblp.uni-trier.de/db/journals/bstj/bstj27.html#Shannon48

[24] Jacob Snow. 2018. Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots. (2018).

[25] C. Spearman. 1904. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology* 15 (1904), 88–103.

[26] Stanford Vision Lab. 2020. ImageNet. *http://image-net.org/explore* (2020). accessed 6.Oct.2020.

[27] Pierre Stock and Moustapha Cisse. 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 498–512.

[28] M. P. Toglia and W. F. Battig. 1978. *Handbook of semantic word norms.* Lawrence Erlbaum.

[29] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive Inequity in Object Detection. *CoRR* abs/1902.11097 (2019). arXiv:1902.11097 http://arxiv.org/abs/1902.11097