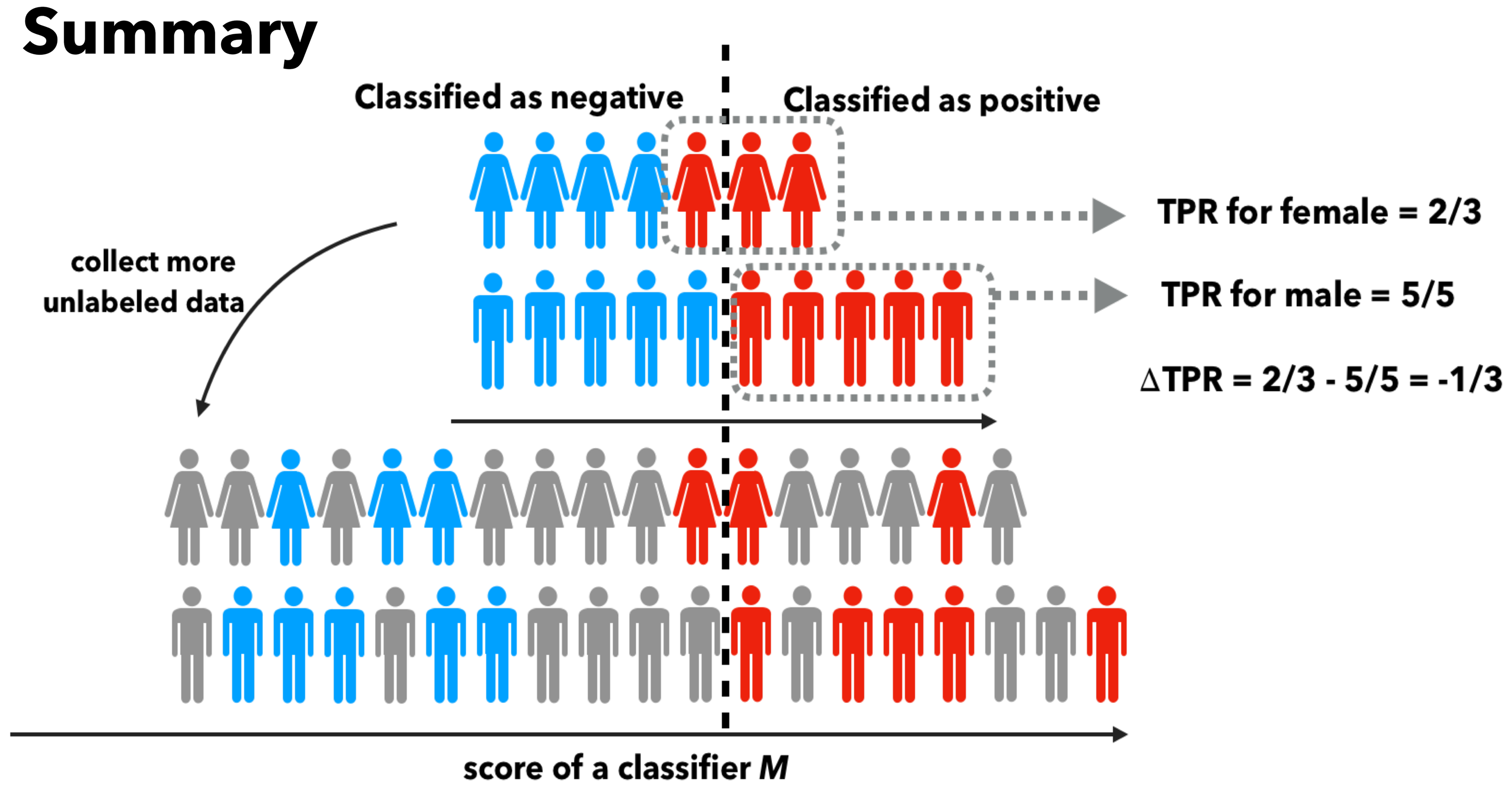
Can I Trust My Fairness Metric?
Assessing Fairness with Unlabeled Data and Bayesian Inference

Disi Ji[1], Padhraic Smyth[1], Mark Steyvers[2]
[1]Department of Computer Science, [2]Department of Cognitive Sciences
University of California, Irvine

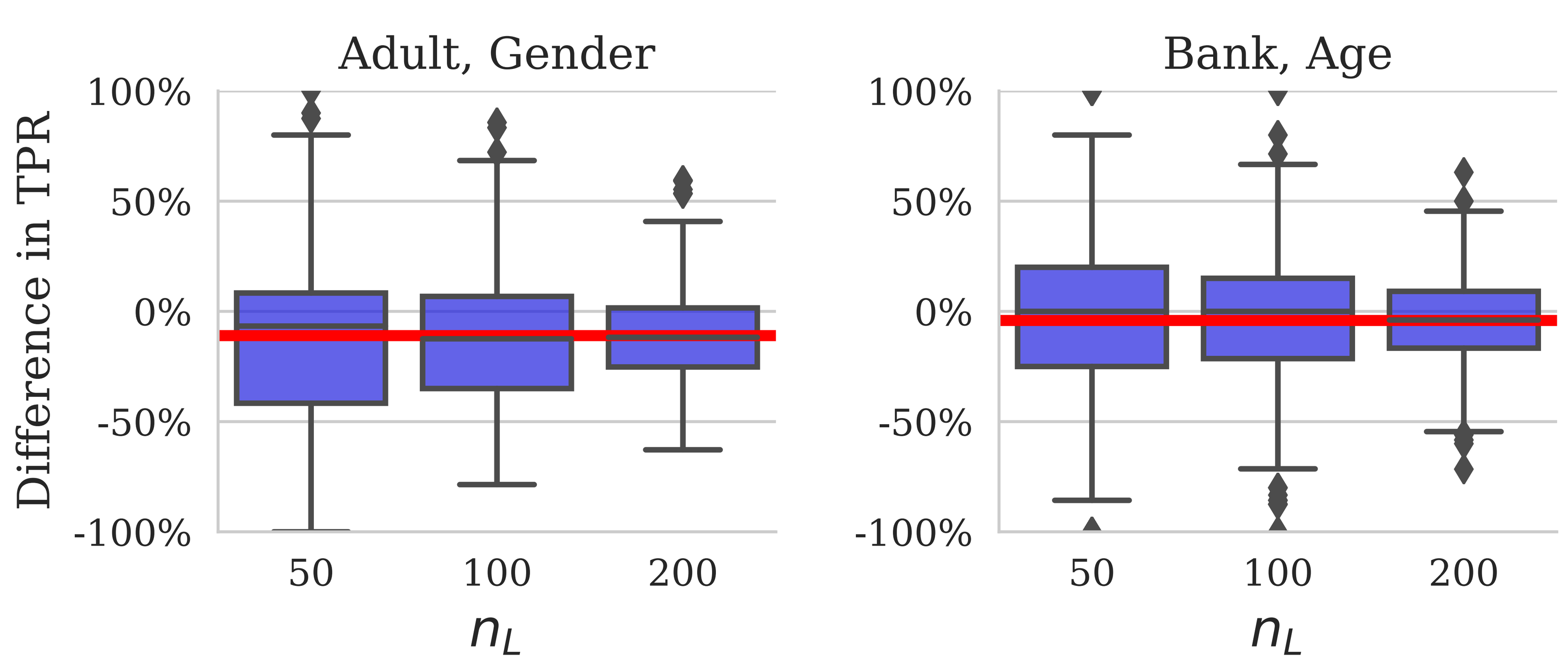
Summary



➢ **Equality of opportunity**: equal TPR across different groups
➢ Due to small sample size, the estimated TPRs are **noisy**!

➢ Contribution:
➢ 1. **Quantify uncertainty** in fairness metrics using Bayesian methods
➢ 2. **Reduce uncertainty** of fairness by leveraging unlabeled data

Frequentist-based Estimates Have High Variance



➢ High variability for the estimated TPRs relative to the true TPRs (shown in red) as a function of the number of labeled examples.
➢ In many cases the estimates are two or three or more times larger than the true difference.
➢ A relatively large percentage of the estimates have the opposite sign of the true difference, potentially leading to mistaken conclusions

Quantify Uncertainty of Fairness Assessment

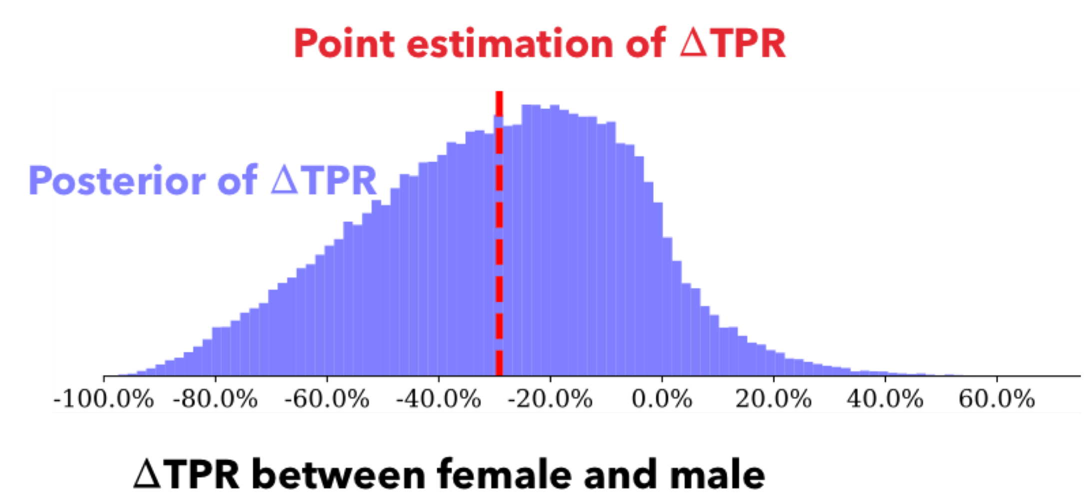
For two groups $g = 0{,}1$, and n_L labeled data D_L:
➢ **Groupwise performance metric** $\theta_g = P(\hat{y} = 1 | y = 1, g)$

$$\theta_g \sim Beta(\alpha_g, \beta_g)$$

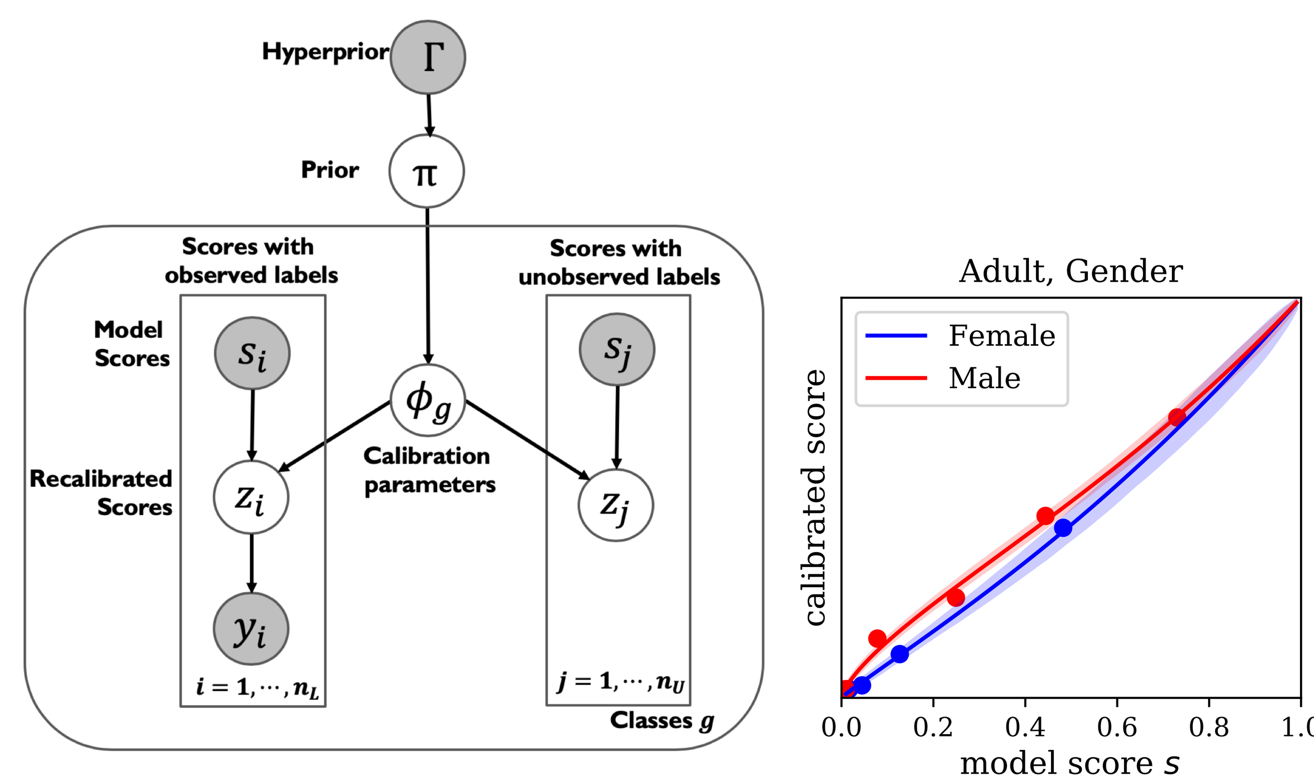
➢ **Correctness** of the prediction model for i: $I_i = I(\hat{y}_i = y_i),\ 1 \le i \le n_L$:

$$I_i \sim Bernoulli(\theta_g)$$

➢ **Group fairness metric**: $\Delta = \theta_1 - \theta_0$
➢ Obtain posterior distribution $P(\Delta | D_L)$ via Monte Carlo samples



Reduce Uncertainty with Unlabeled Data



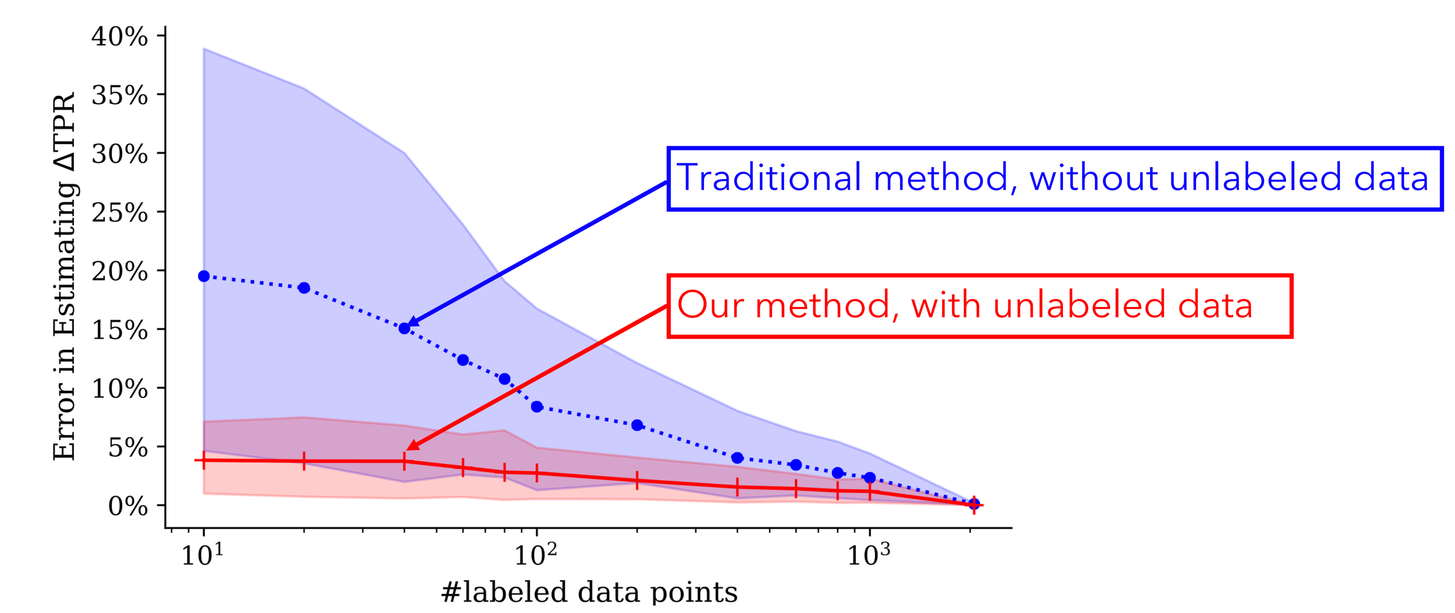
We treat each $z_j, j = 1, \ldots, n_U$ as a latent variable per example. The high level steps of the approach are as follows:

1. Use the n_L labeled examples to **estimate groupwise calibration functions with parameters φ_g**, that transform the (potentially) uncalibrated scores s of the model to calibrated scores. More specifically, we perform Bayesian inference to obtain posterior samples from $P(\varphi_g | D_L)$ for the groupwise calibration parameters φ_g.
2. **Obtain posterior samples of recalibrated scores** from $P\varphi_g(z_j | D_L, s_j)$ for each unlabeled example $j = 1, \ldots, n_U$, conditioned on posterior samples of the φ_g's.
3. Use posterior samples from the z_j's, combined with the labeled data, to **generate estimates of the groupwise metrics θ_g and the difference in metrics Δ.**

Experimental Results

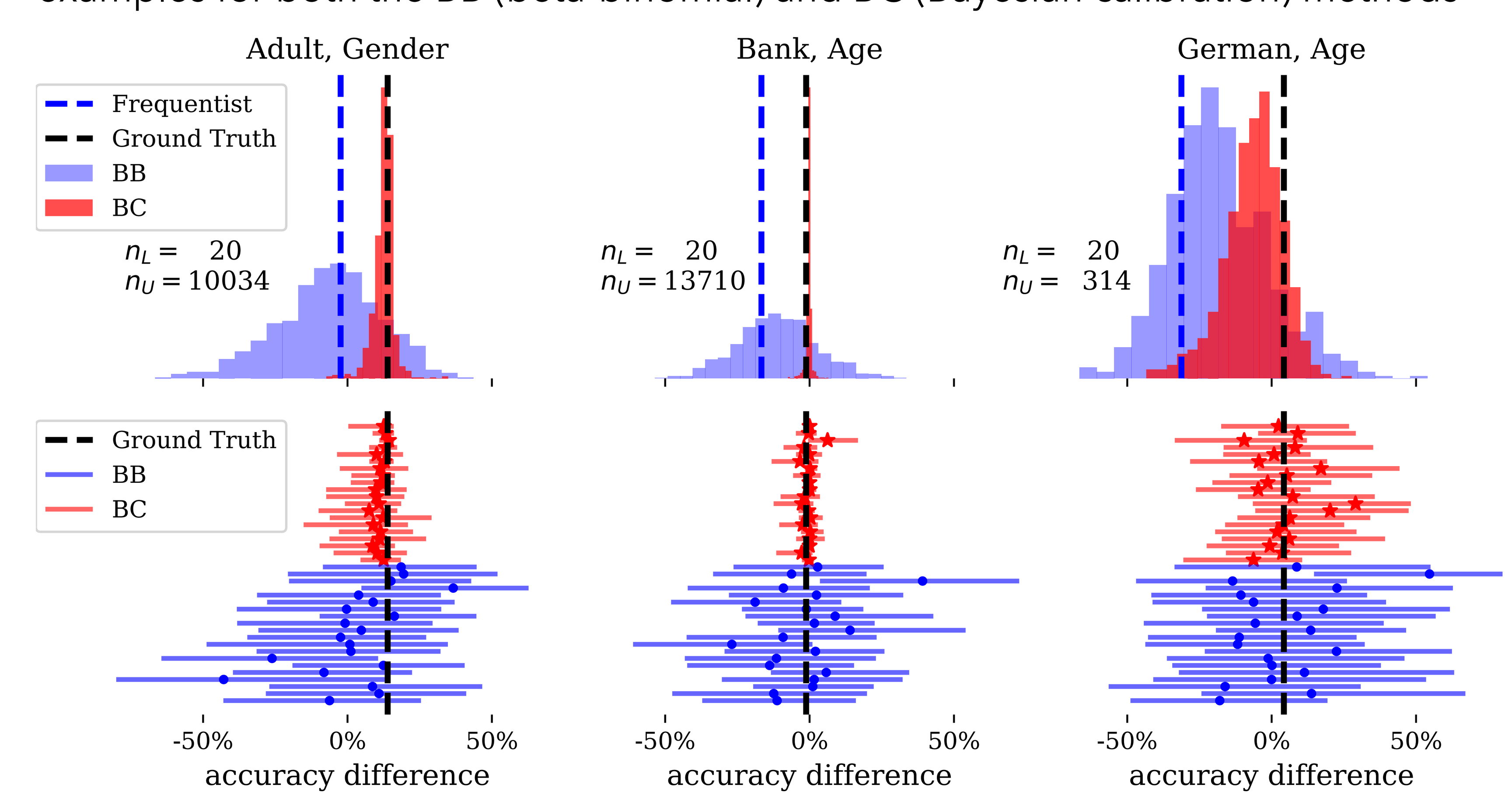
Dataset	Test Size	G	$P(g=0)$	$P(y=1)$
Adult	10054	gender, race	0.68, 0.86	0.25
Bank	13730	age	0.45	0.11
German	334	age, gender	0.79, 0.37	0.17
Compas-R	2056	gender, race	0.7, 0.85	0.69
Compas-VR	1337	gender, race	0.8, 0.34	0.47
Ricci	40	race	0.65	0.50

Example: assess Δ TPR of COMPAS- Recividism, Race



With **10** labeled data and **~2000** unlabeled data, error in estimating TPR is **5%** for our method versus 20% with only labeled data

Illustrative Results: Posterior density (samples) and frequentist estimates (dotted vertical blue lines) for the difference in group accuracy with **20** random labeled examples for both the BB (beta-binomial) and BC (Bayesian calibration) methods



Acknowledgements
This material is based upon work supported in part by the National Science Foundation under grants number 1900644 and 1927245 and by a Qualcomm Faculty Award (PS). In addition this work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at the University of California, Irvine.