

# CAN I TRUST MY FAIRNESS METRIC?

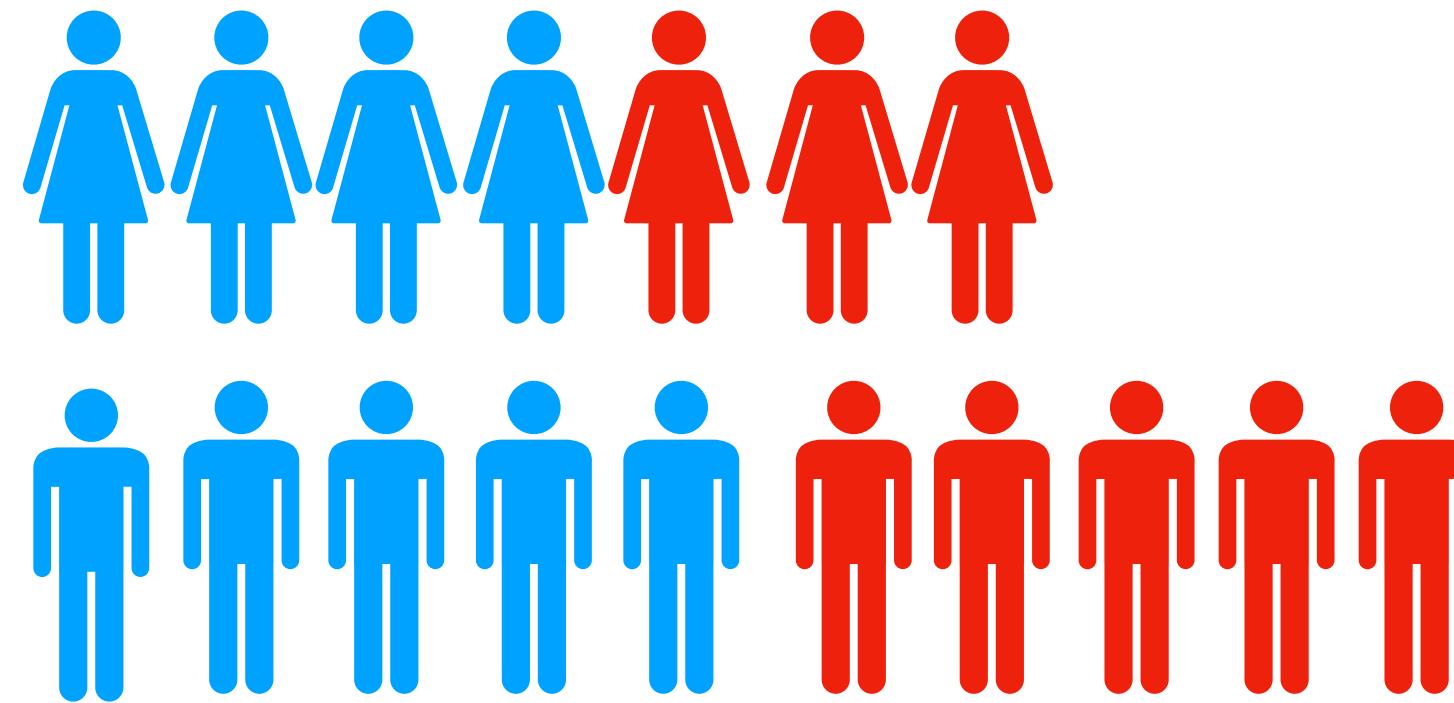
## ASSESSING FAIRNESS WITH UNLABELED DATA & BAYESIAN INFERENCE

**Disi Ji (UC Irvine)**

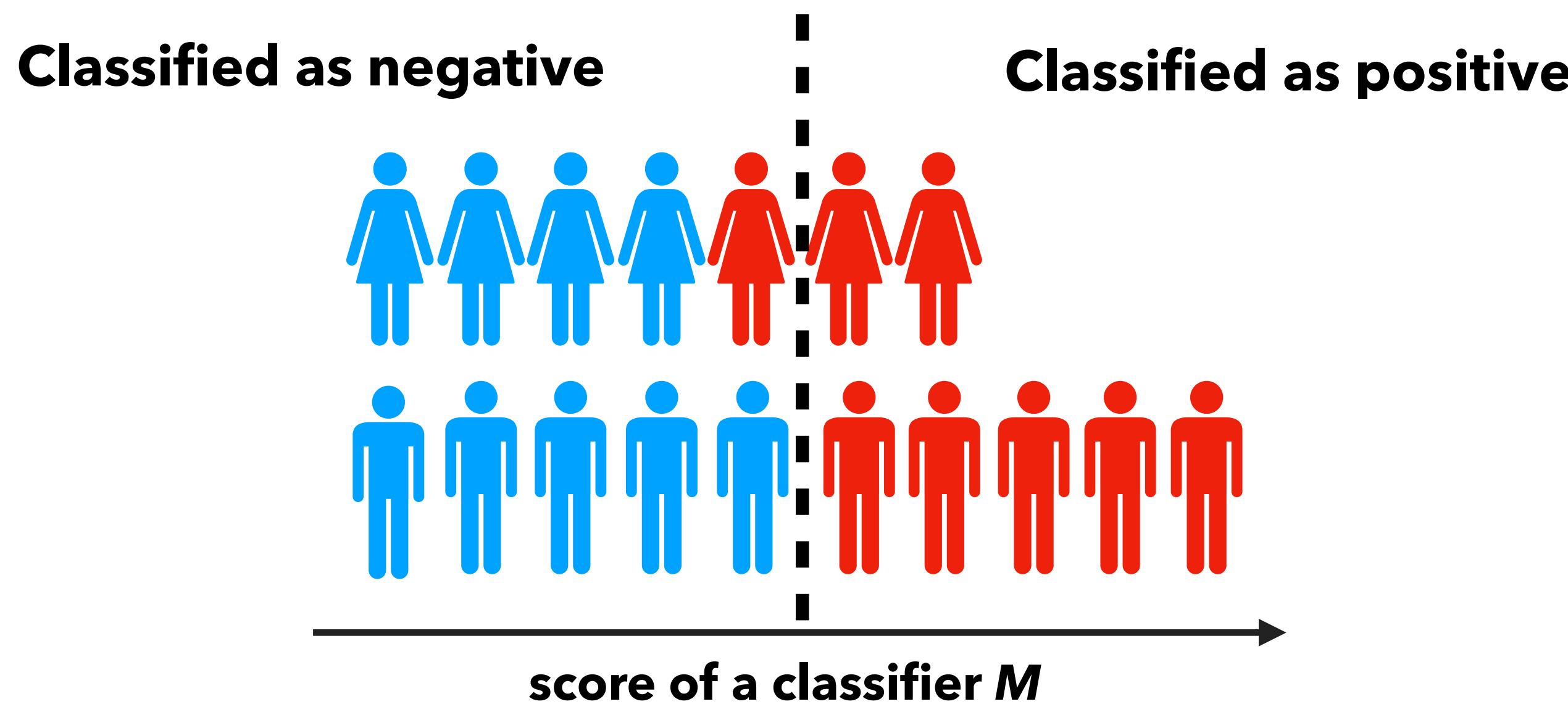


Joint work with **Padhraic Smyth** and **Mark Steyvers**  
(UC Irvine)

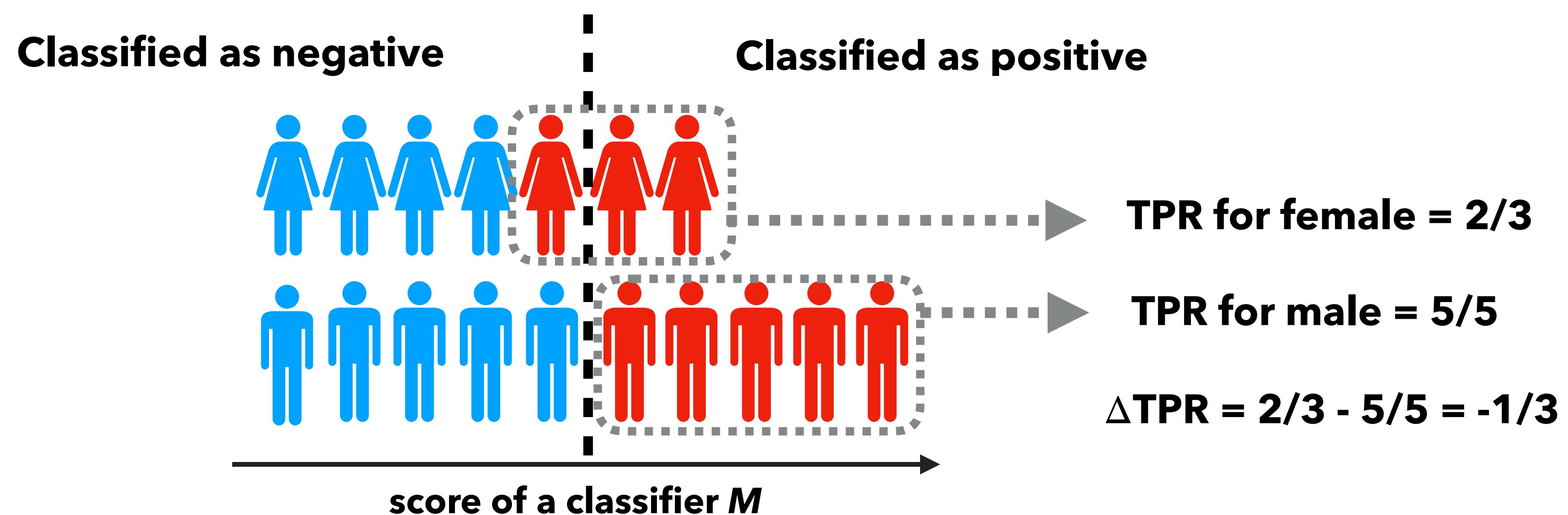
# IS THE CLASSIFIER REALLY UNFAIR?



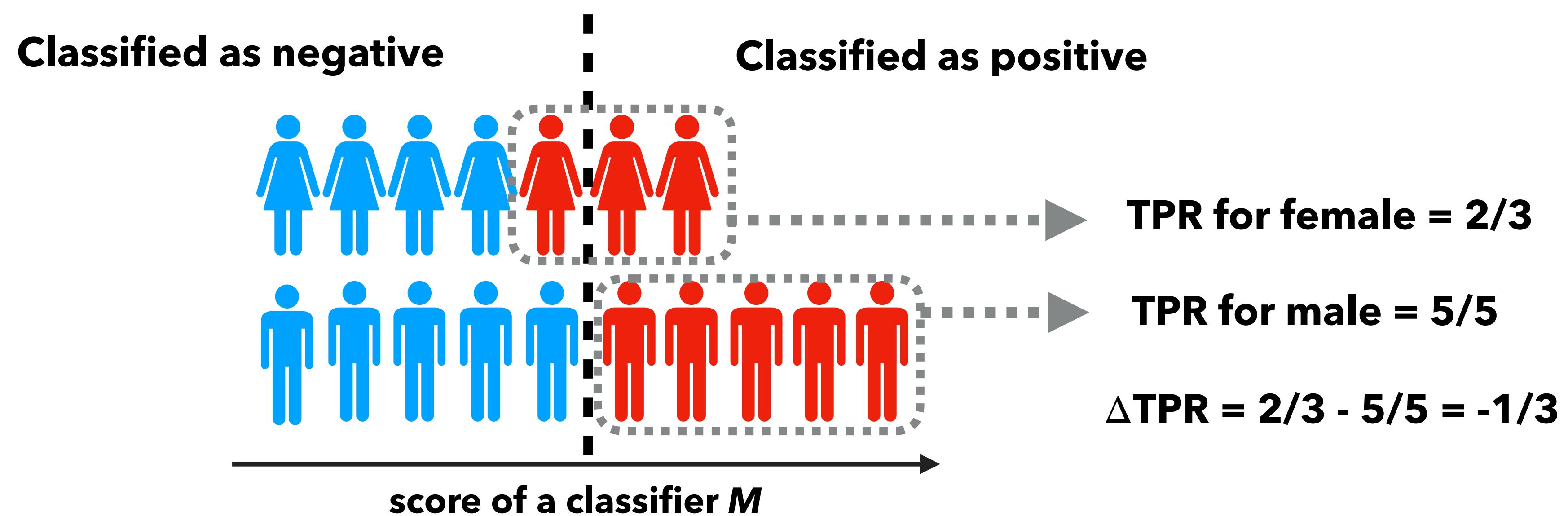
# IS THE CLASSIFIER REALLY UNFAIR?



# IS THE CLASSIFIER REALLY UNFAIR?

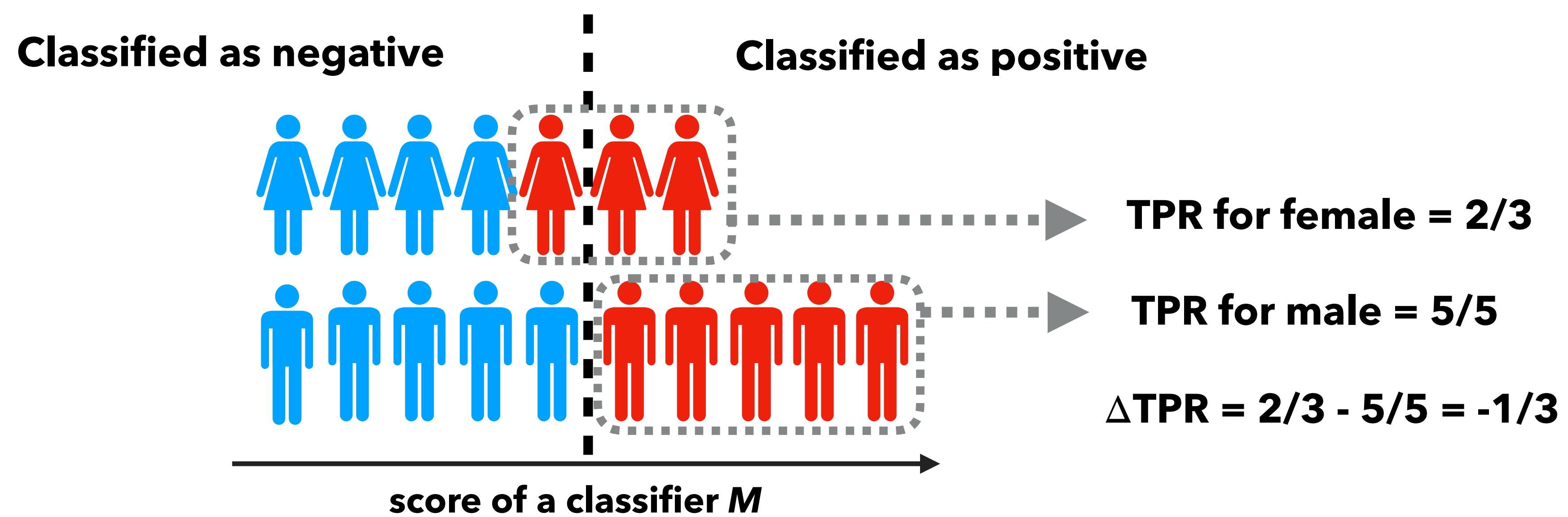


# IS THE CLASSIFIER REALLY UNFAIR?



- ▶ Equality of opportunity:
- ▶ equal TPR across different groups [Hardt, Price & Srebro, 2016]

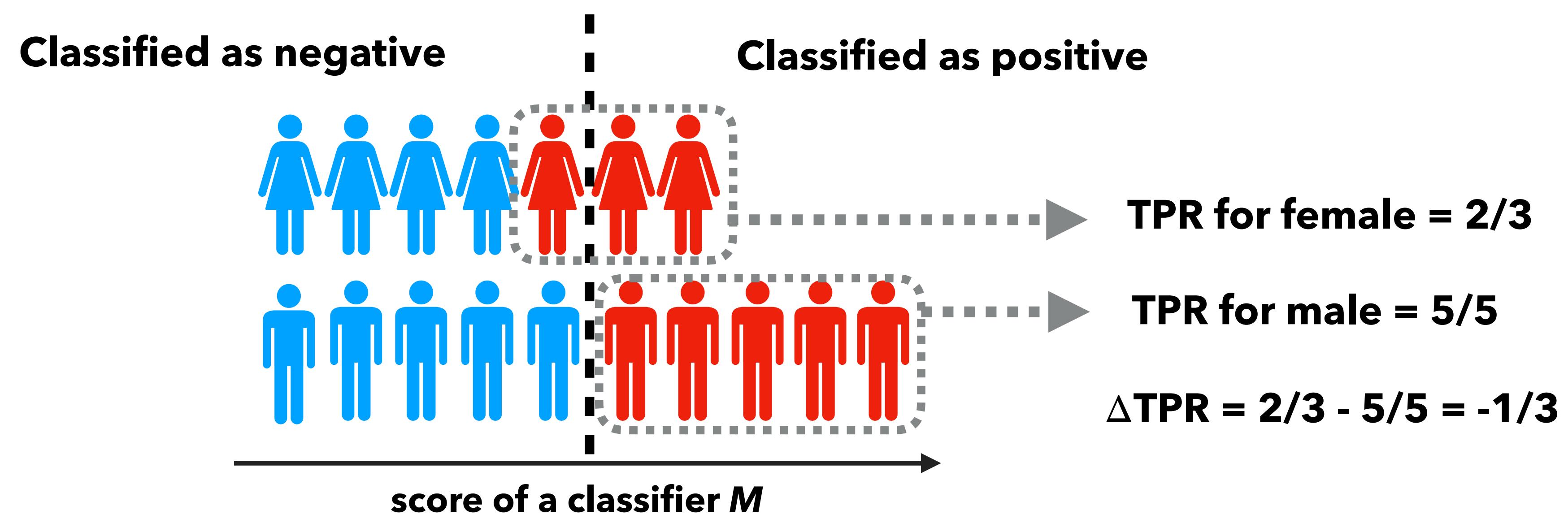
# IS THE CLASSIFIER REALLY UNFAIR?



- ▶ Equality of opportunity:

  - ▶ equal TPR across different groups [Hardt, Price & Srebro, 2016]
  - ▶ Due to small sample size, the estimated TPRs are noisy!

# IS THE CLASSIFIER REALLY UNFAIR?



- ▶ Equality of opportunity:

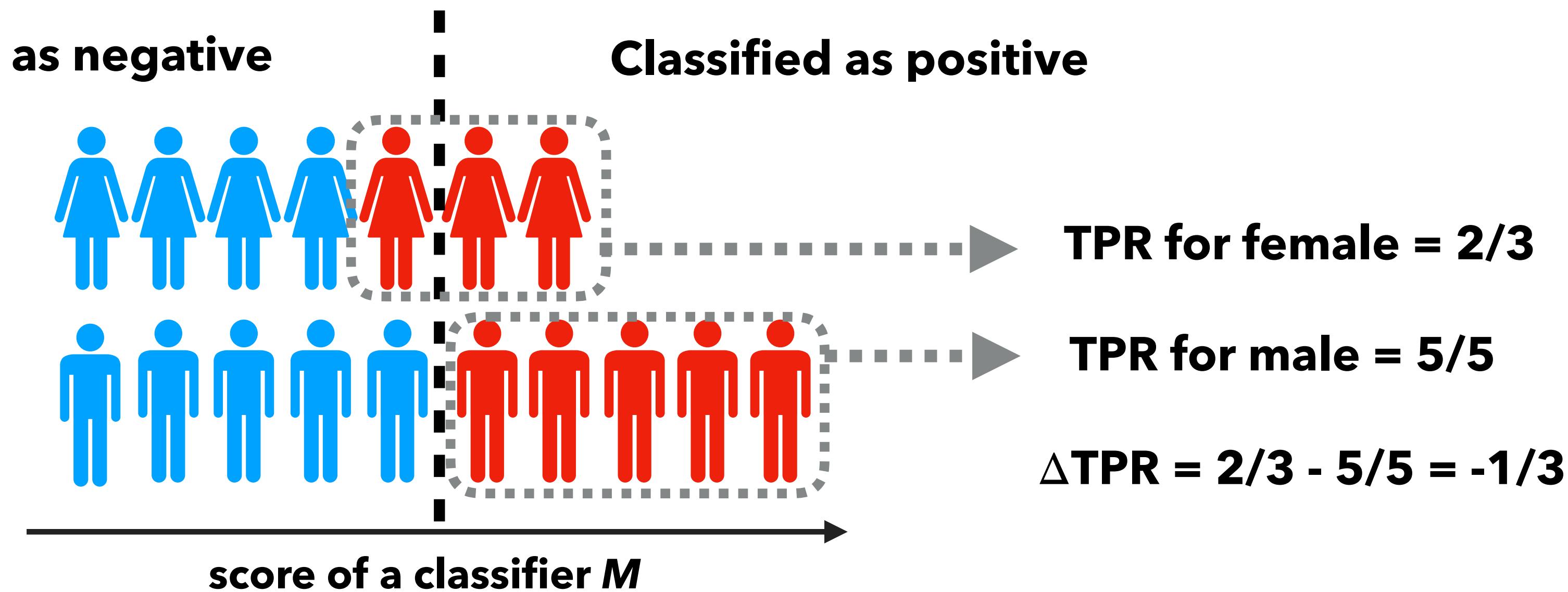
  - ▶ equal TPR across different groups [Hardt, Price & Srebro, 2016]

- ▶ Due to small sample size, the estimated TPRs are noisy!
- ▶ Contribution: quantify uncertainty in fairness metrics using Bayesian methods

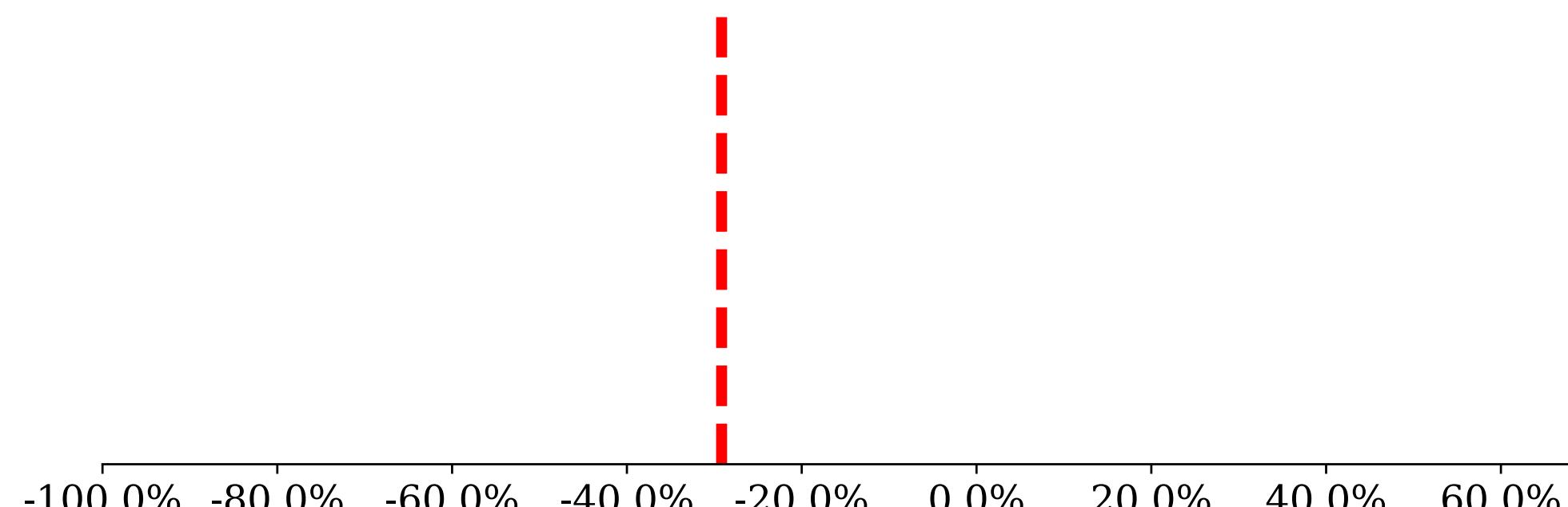
# MODEL FAIRNESS METRICS WITH UNCERTAINTY

**Classified as negative**

**Classified as positive**



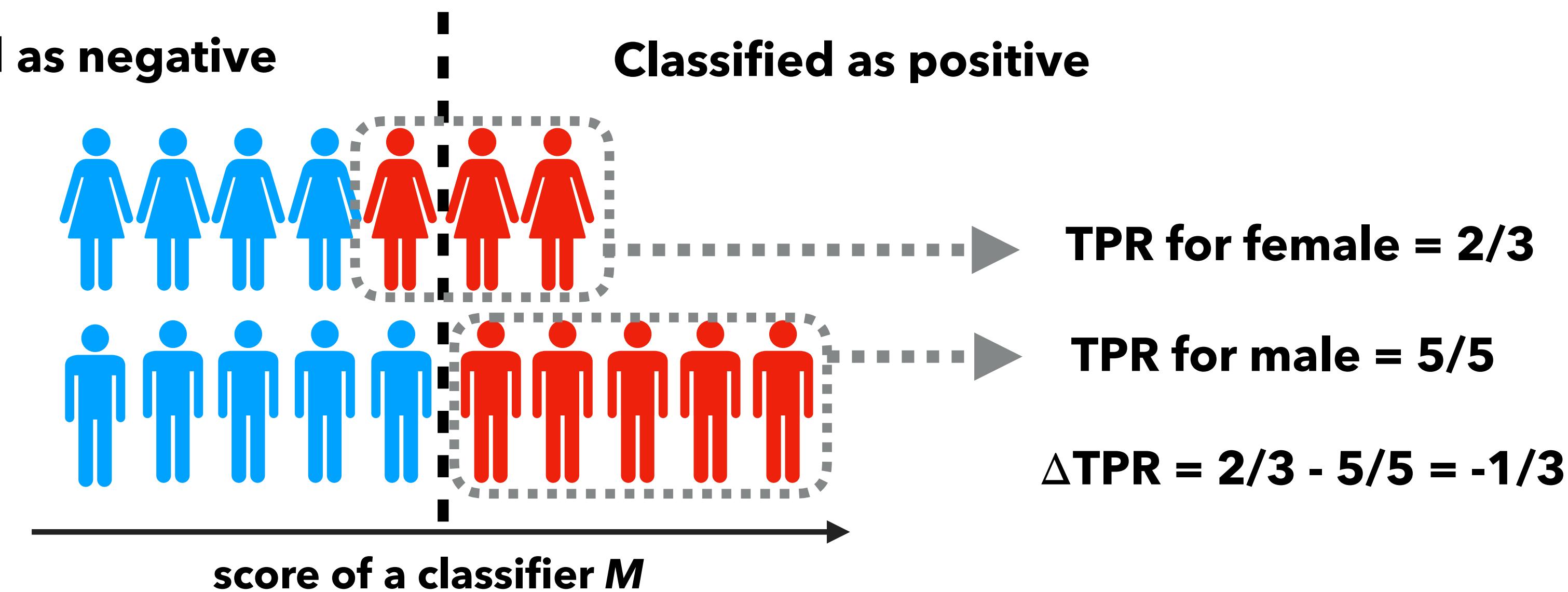
**Point estimation of  $\Delta\text{TPR}$**



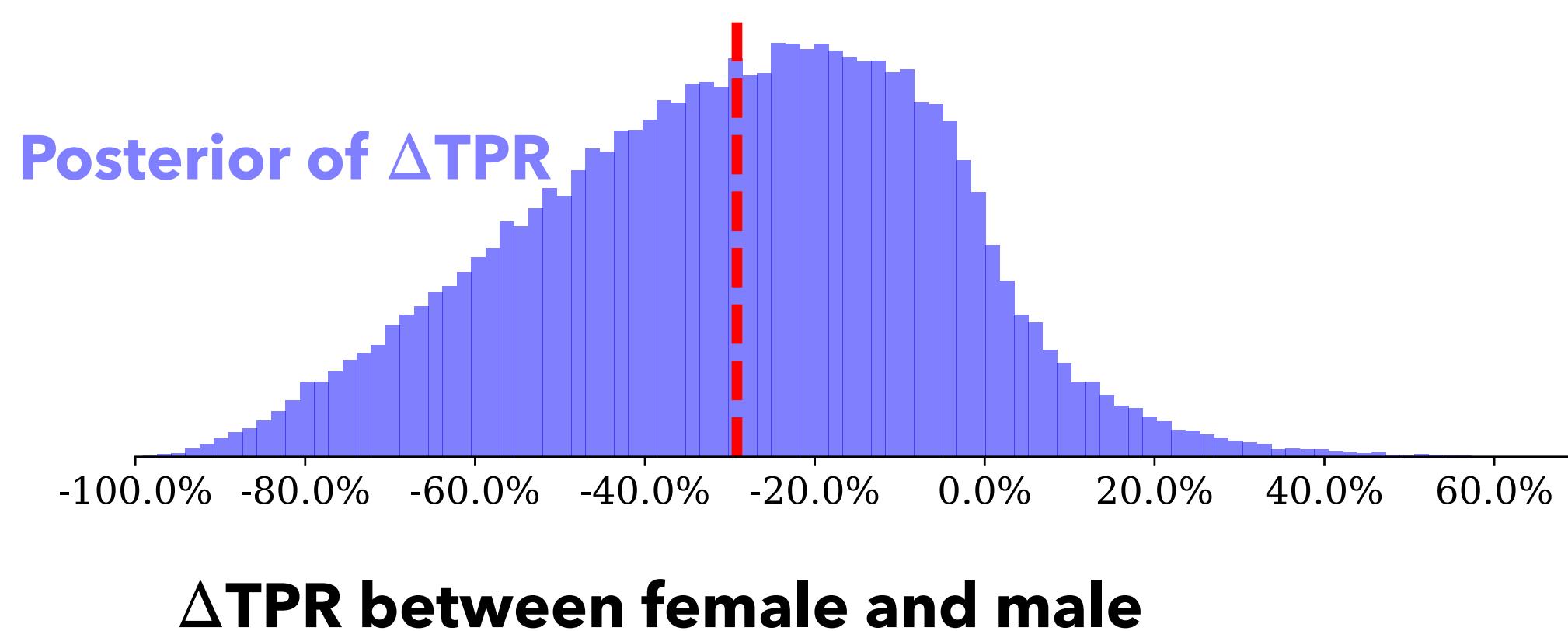
# MODEL FAIRNESS METRICS WITH UNCERTAINTY

**Classified as negative**

**Classified as positive**



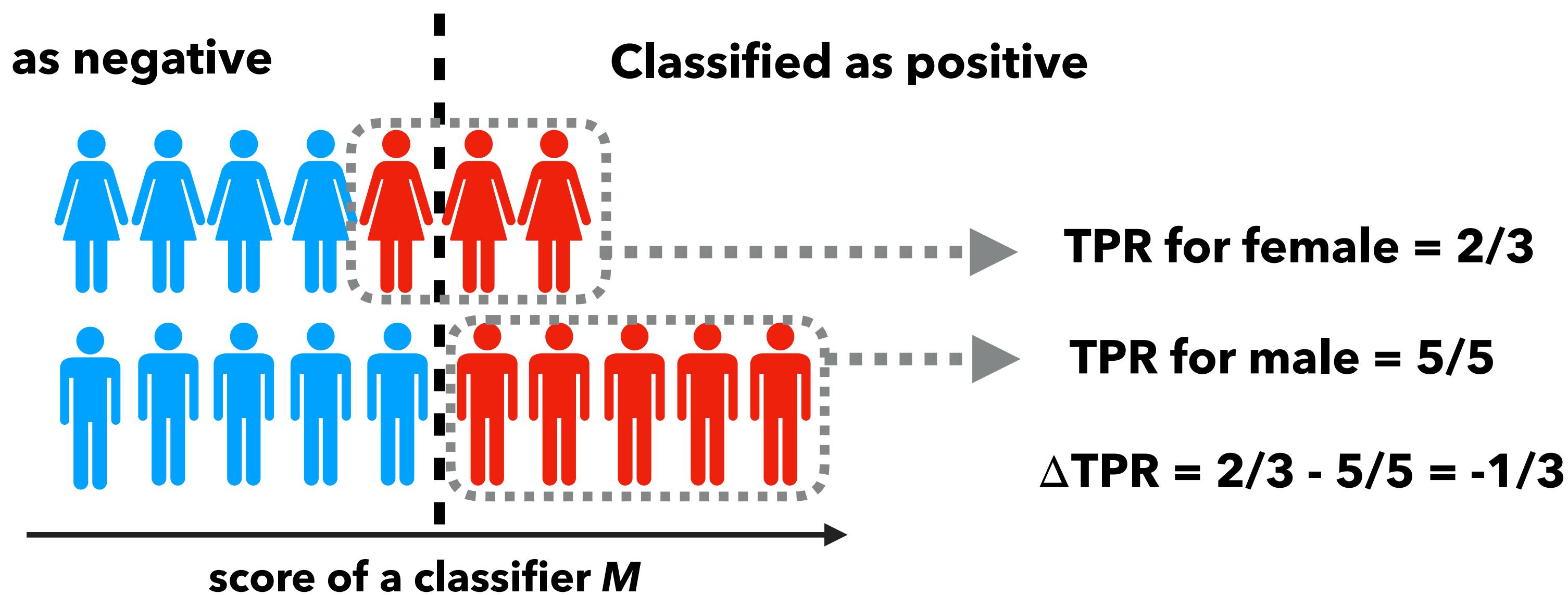
**Point estimation of  $\Delta\text{TPR}$**



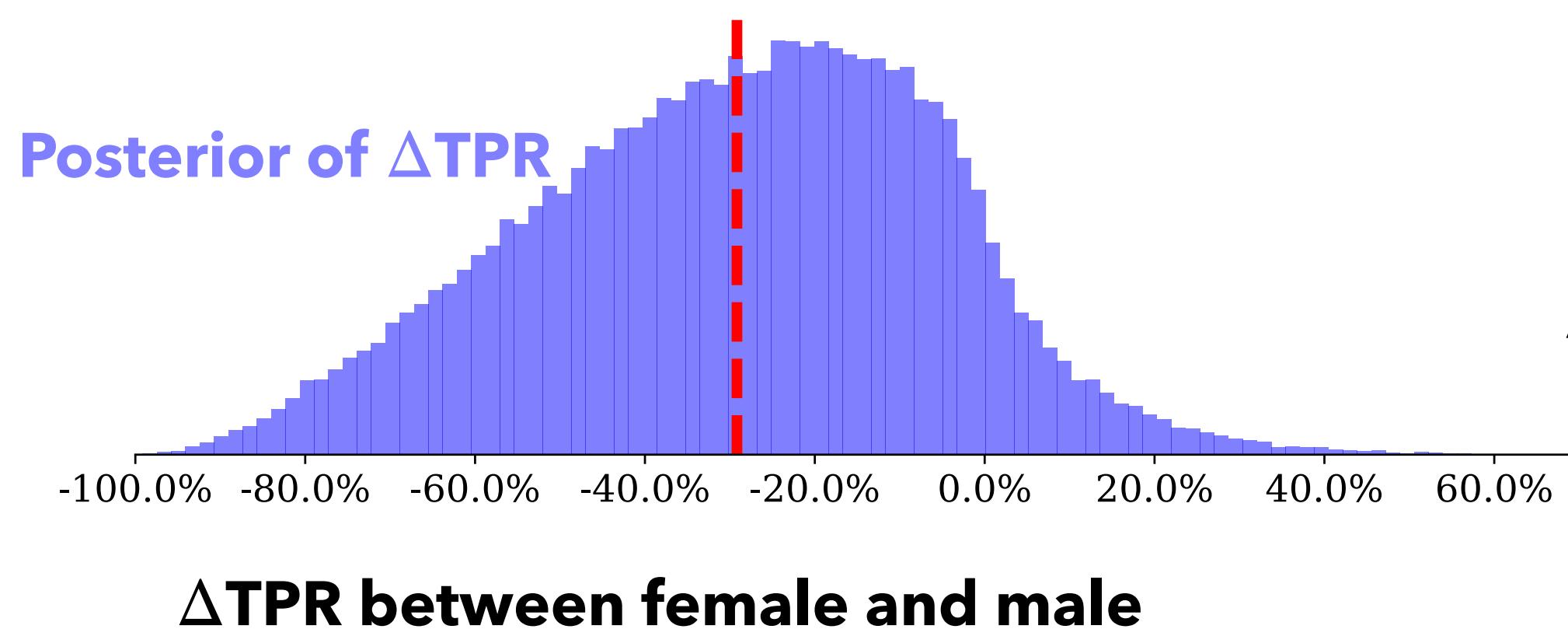
# MODEL FAIRNESS METRICS WITH UNCERTAINTY

**Classified as negative**

**Classified as positive**

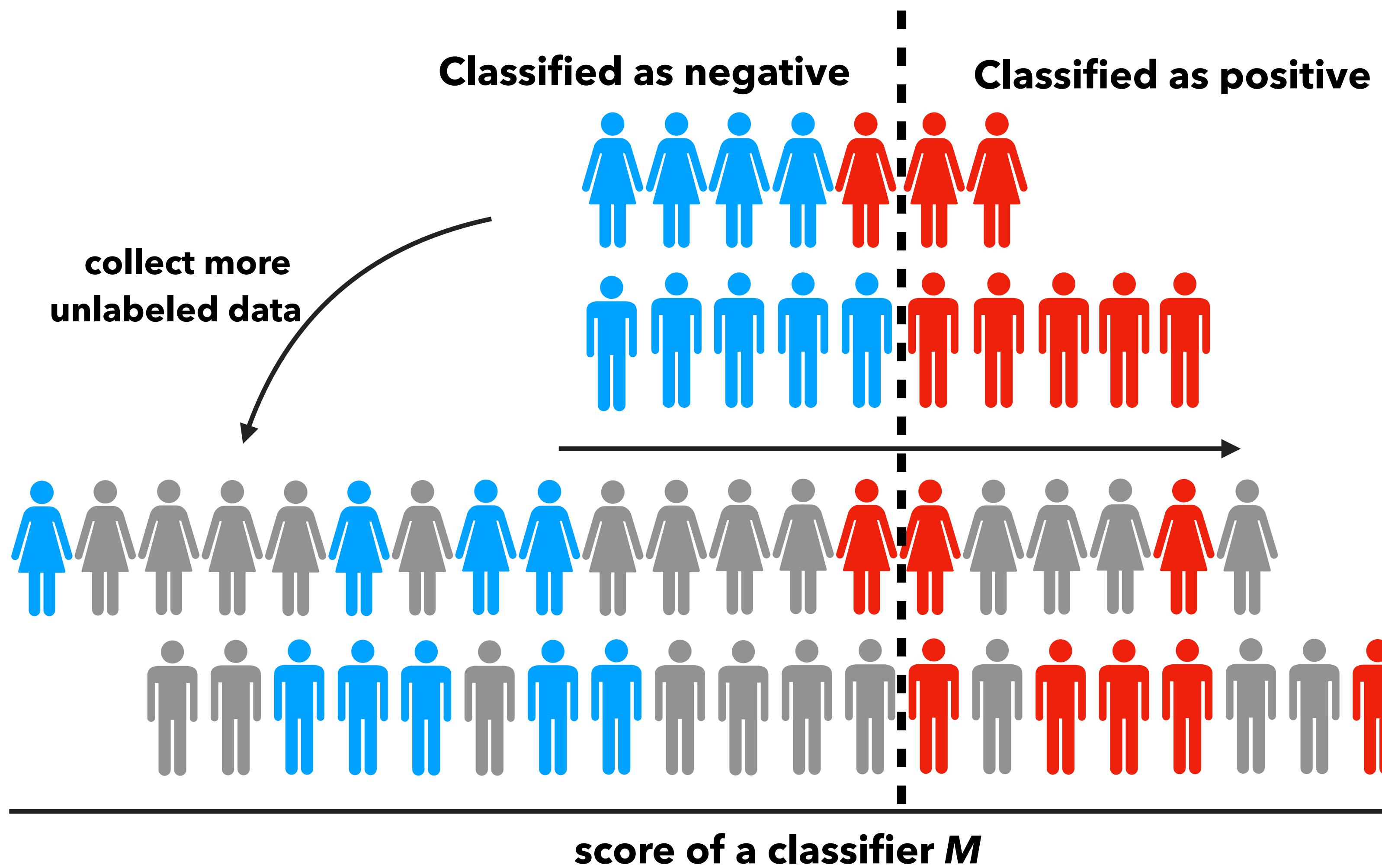


**Point estimation of  $\Delta\text{TPR}$**

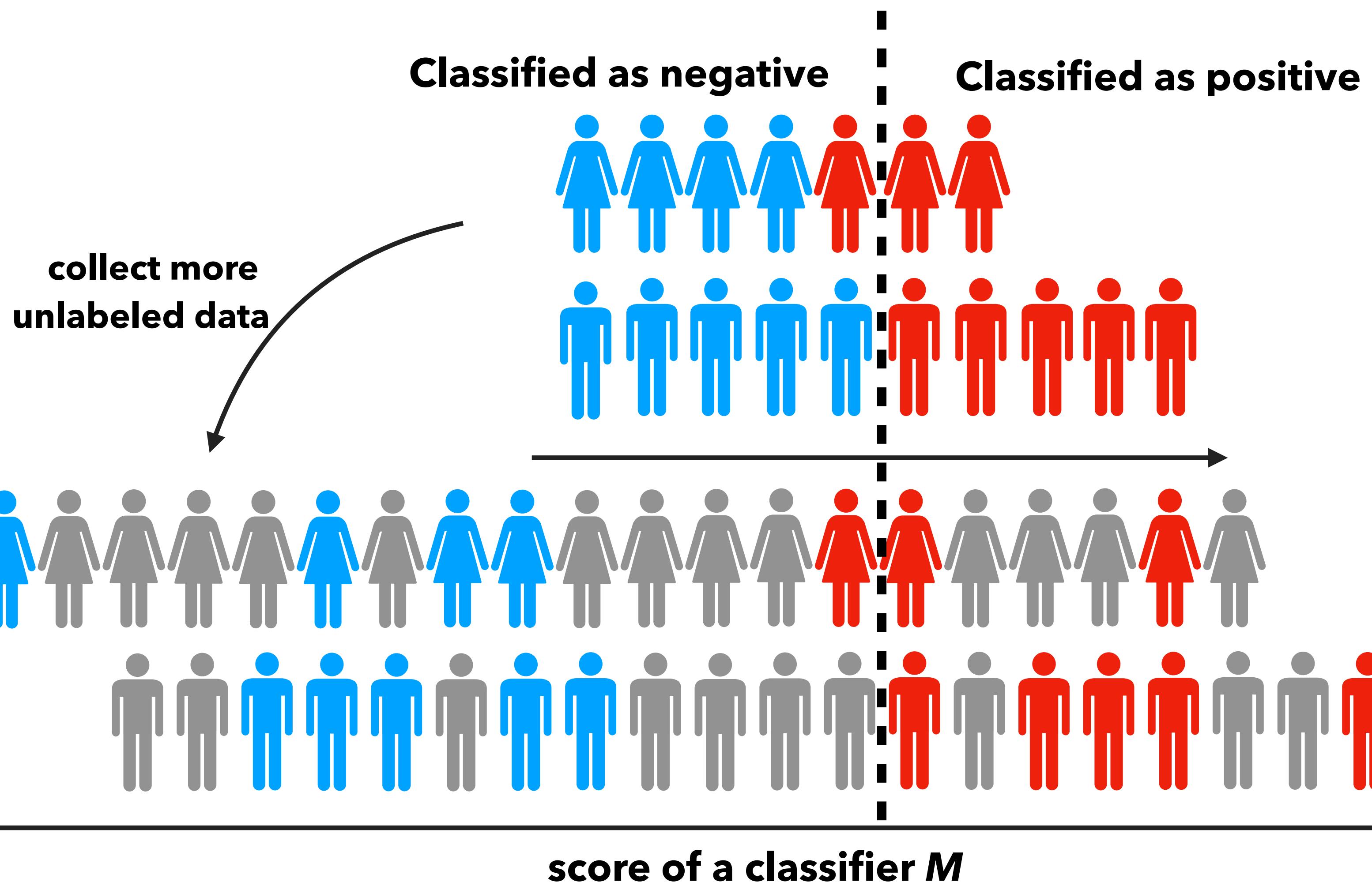


**Q:** The uncertainty is high! How to reduce it?  
**A:** Collect more data! Labeled or **unlabeled!**

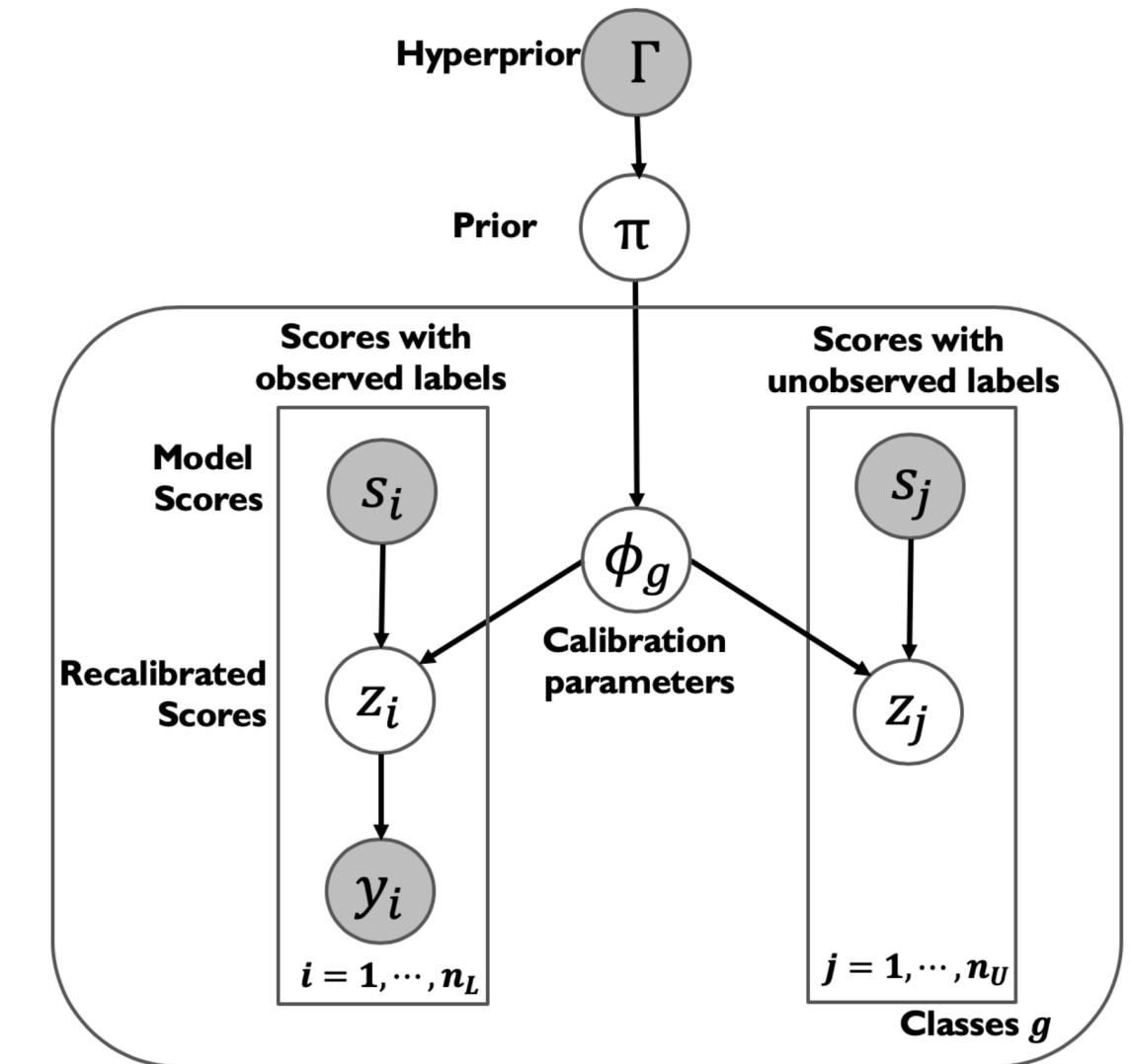
# REDUCE UNCERTAINTY OF FAIRNESS WITH MORE UNLABELED DATA<sup>4</sup>



# REDUCE UNCERTAINTY OF FAIRNESS WITH MORE UNLABELED DATA<sup>4</sup>

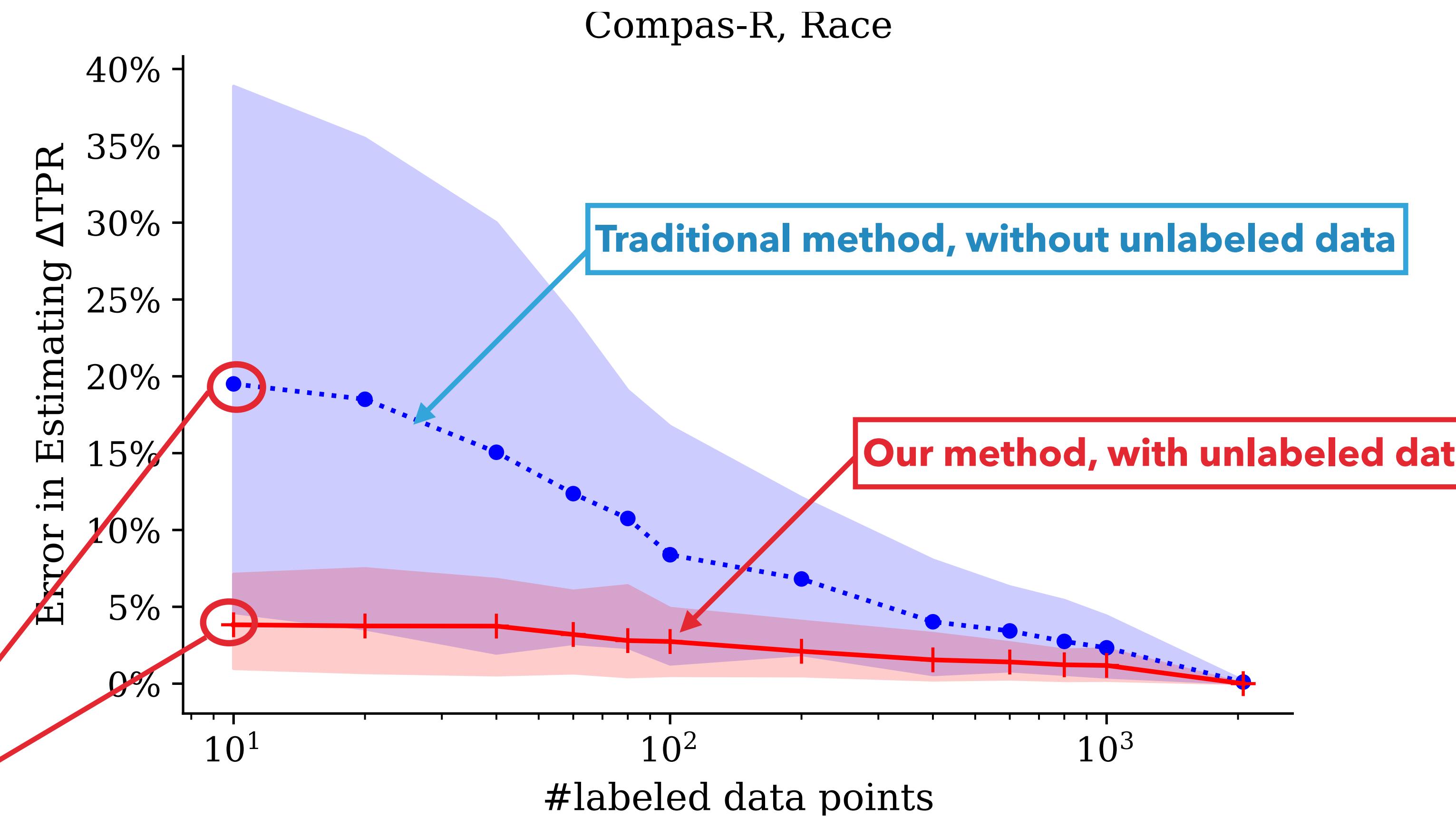


**Method:** train a hierarchical Bayesian calibration model to predict the model performance on unlabeled data



# EXAMPLE: ASSESS DELTA TPR OF COMPAS RECIDIVISM

5



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

# SO, CAN I TRUST MY FAIRNESS METRIC?

- ▶ **Be aware of uncertainty in fairness assessment:** especially when test sizes are relatively small (as is often the case in practice)
- ▶ **Collect more data, labeled or unlabeled, to make the assessment more reliable**
  - ▶ a new Bayesian methodology that uses calibration to leverage information from both unlabeled and labeled examples