2nd NJIT Workshop on Multimedia Intelligence (MMI 2026)

# Biases in machine learning and AI

Shin'ichi Satoh
National Institute of Informatics (NII)

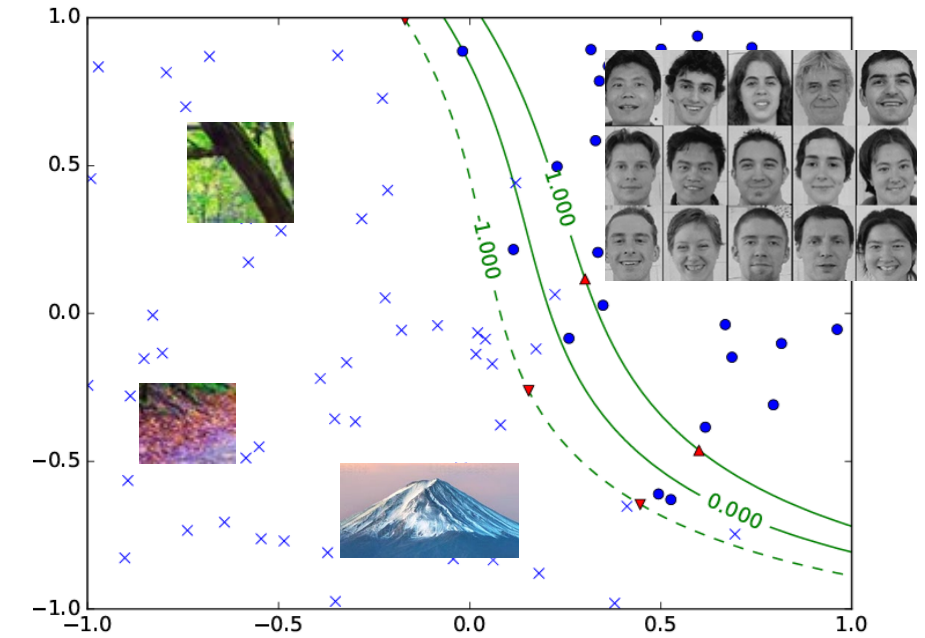
21 - 22 May, 2026

NJIT

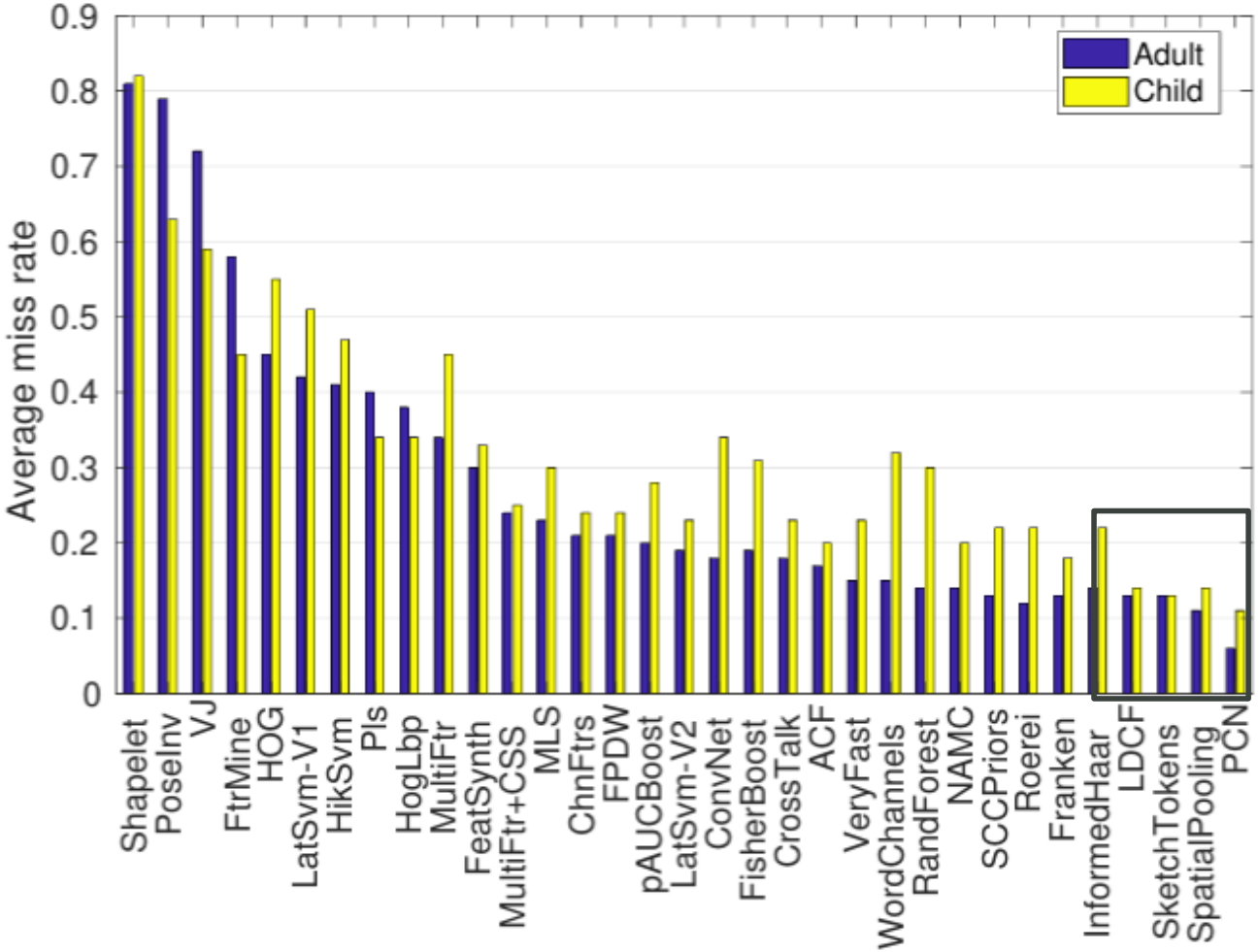
NII
Inter-University Research Institute Corporation /
Research Organization of Information and Systems
National Institute of Informatics

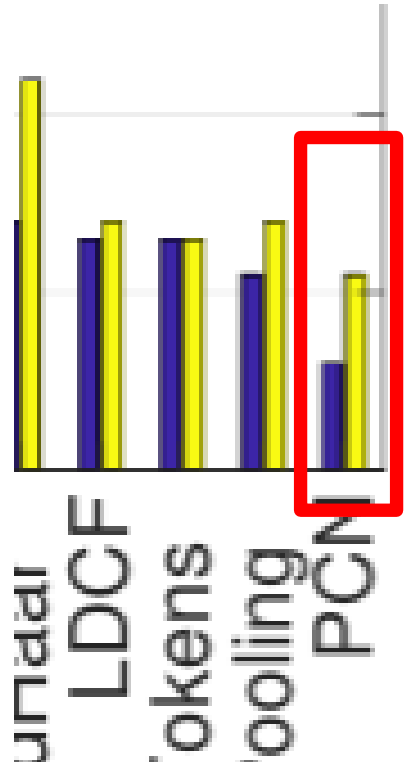
# Biases and Machine Learning

- Machine learning can effectively find “useful” patterns in training data
- Patterns are basically biases found in the feature space
- Inherently good machine learning (AI) is very good at finding biases in data
- However, because of this, machine learning can cause shortcut learning or spurious correlation
- Camel and desert, cow and grass…




# Biases in Machine Learning Systems




Martim Brandao, Age and gender bias in pedestrian detection algorithms, Workshop on Fairness Accountability Transparency and Ethics in Computer Vision at CVPR 2019

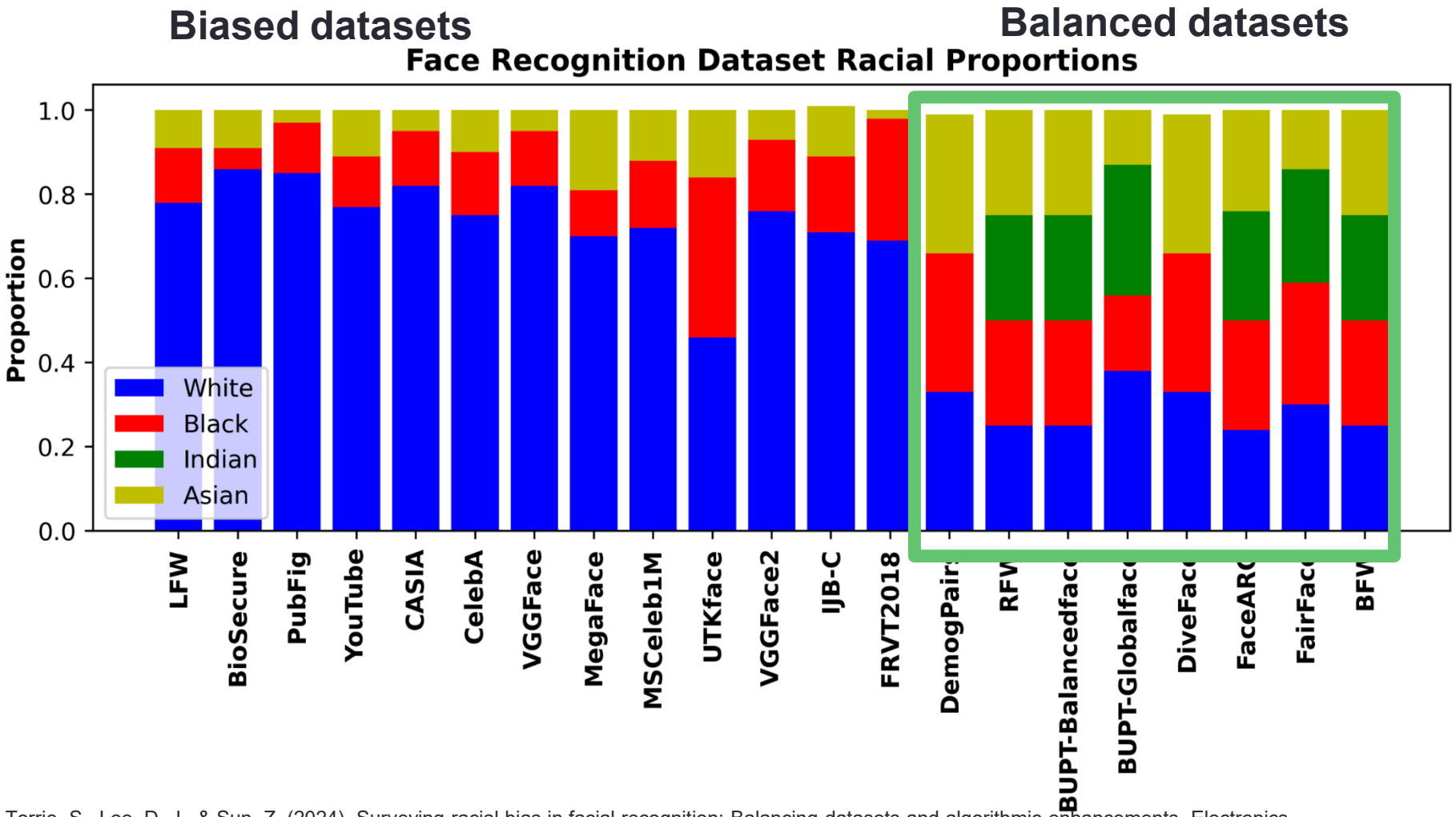
# Uncovering Gender Biases in Gender Identification Models for Japanese Data Analysis

- Workshop on Demographic Diversity in Computer Vision@ CVPR 2025
- Ana Manzano Rodríguez, Camille Guinaudeau, Shin'ichi Satoh

# Abstract


- NII has large-scale Japanese TV broadcast video archives
- Initially, we wanted to observe yearly changes of gender balances for some specific programs
- We then found that off-the-shelf state of the art face-based gender classifiers performed very pooly
- This paper studies the performance of gender classifier across couple of different datasets
- In summary, gender classifier works very well with well-known dataset with biases, but works very poorly on Asian faces
- Simple fine tuning can well mitigate this biased performance

# Face datasets are biased

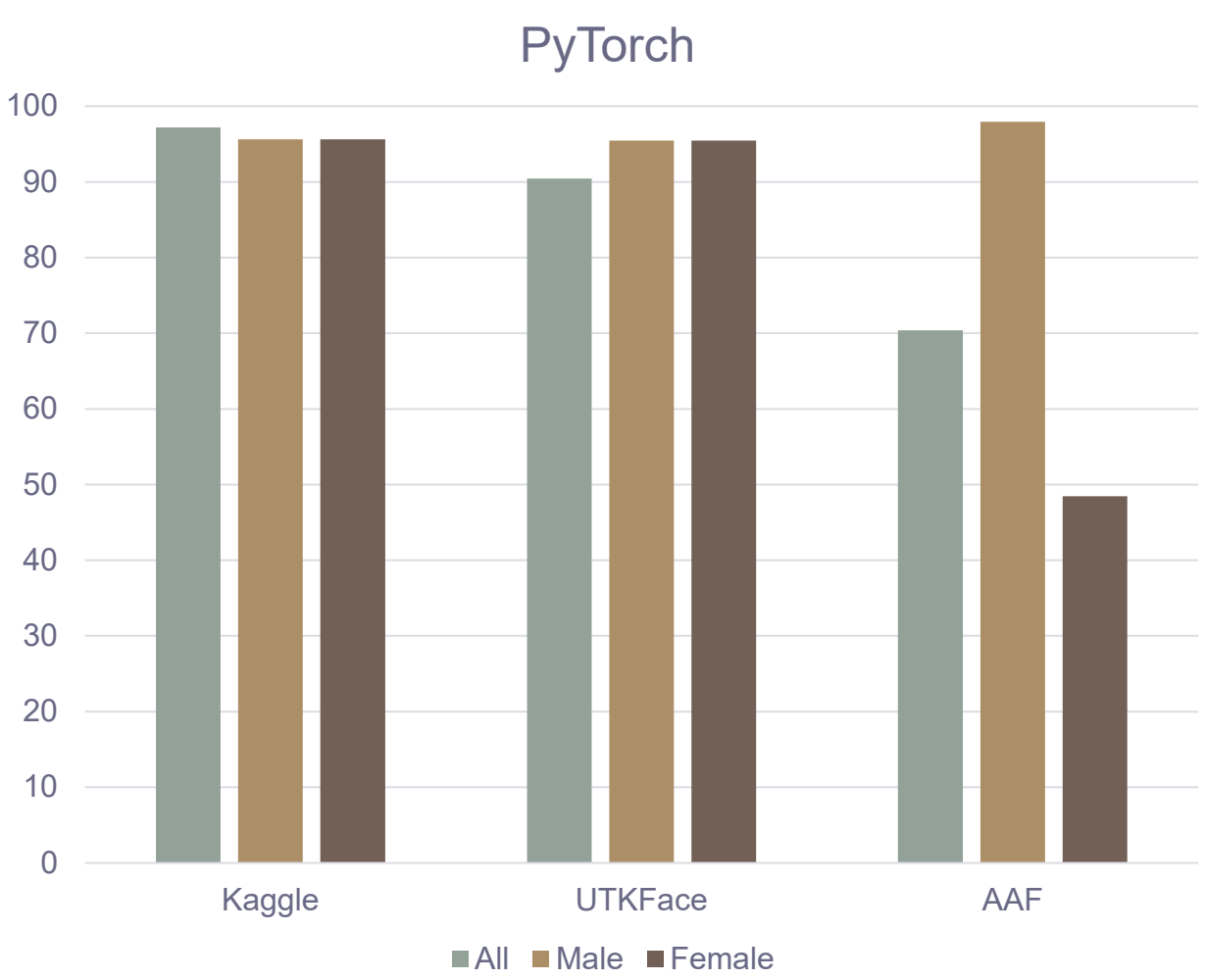


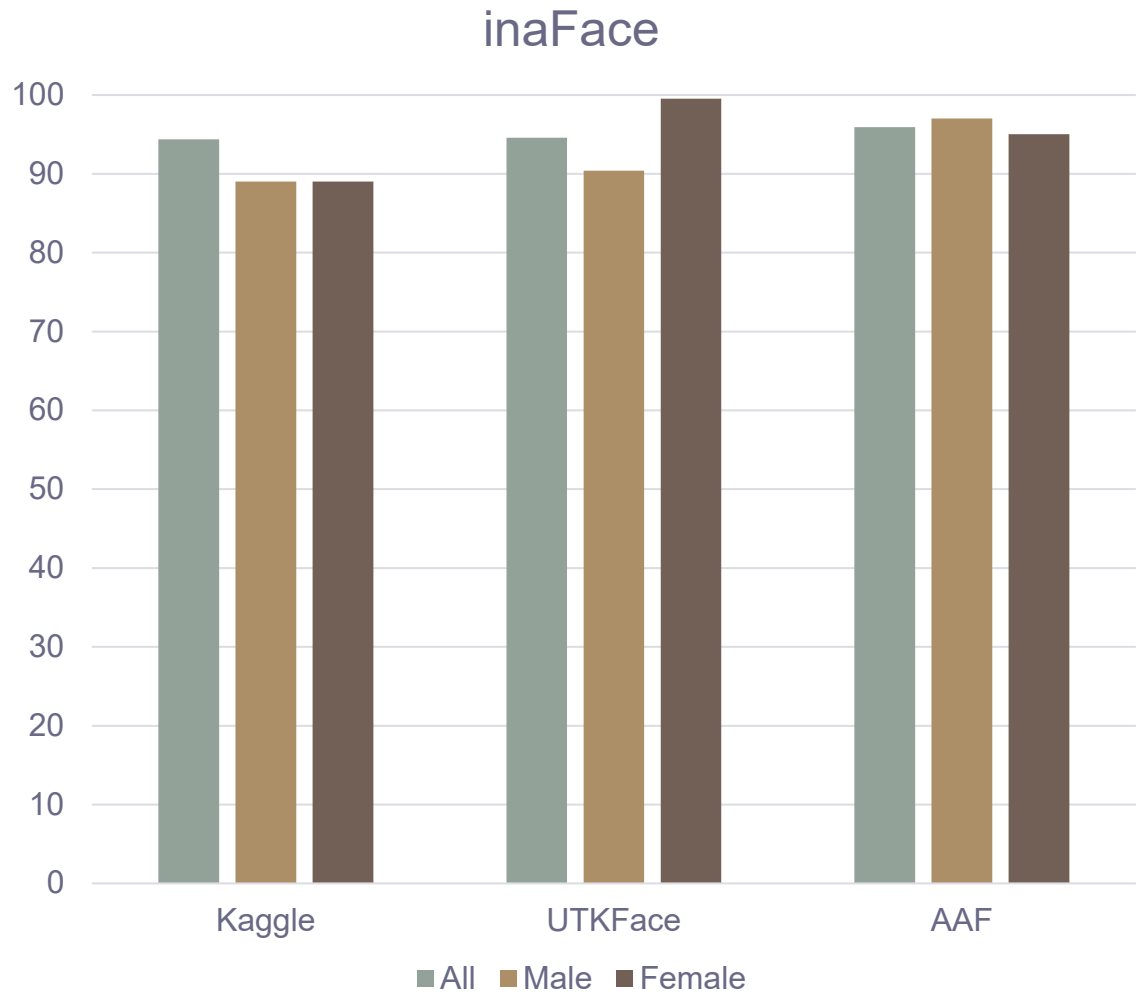

[4] Sumsion, A., Torrie, S., Lee, D.-J., & Sun, Z. (2024). Surveying racial bias in facial recognition: Balancing datasets and algorithmic enhancements. Electronics
[5] Cheng, J., Li, Y., Wang, J., Yu, L., & Wang, S. (2019). *Exploiting effective facial patches for robust gender recognition.* Tsinghua Sci. Technol., 24(3), 333–345.
[6] Karkkainen, K., & Joo, J. (2021). *FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation.* In WACV '21, pp. 1548–1558.

# Datasets, classifiers, …

- Datasets:
  - **Kaggle** dataset: 23k x 2 training, 5.5k x 2 validation, demographically diverse but lacks information on race and age
  - **UTKFace** dataset: 20k with age, gender, ethnicity.  Balanced gener
  - All-Age-Faces dataset (**AAF**): 13k, primarily Asian
  - **FairFace** dataset: 108k, balanced across seven racial groups, with labels
  - **NHK** dataset: 1.2k from NHK News 7
- Studied classifiers:
  - **PyTorch model**: ResNet18-based classifier trained with Kaggle dataset
  - https://github.com/ndb796/Face-Gender-Classification-PyTorch
  - **INA face analyzer**: ResNet50 trained on FairFace dataset
  - https://github.com/ndb796/Face-Gender-Classification-PyTorch
- We evaluated the gender classification performance with combinations of datasets and classifiers

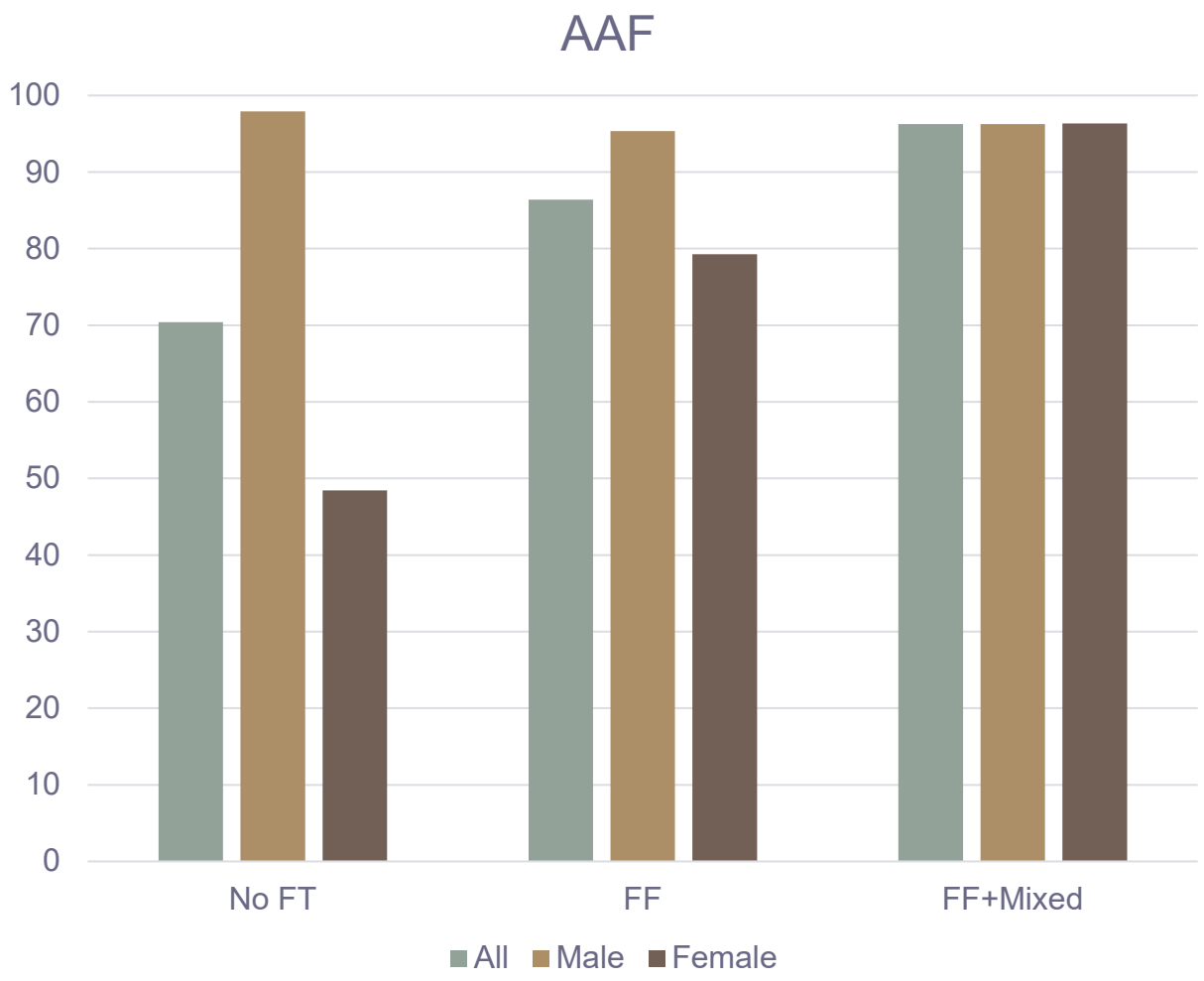
# Baseline gender classification performance



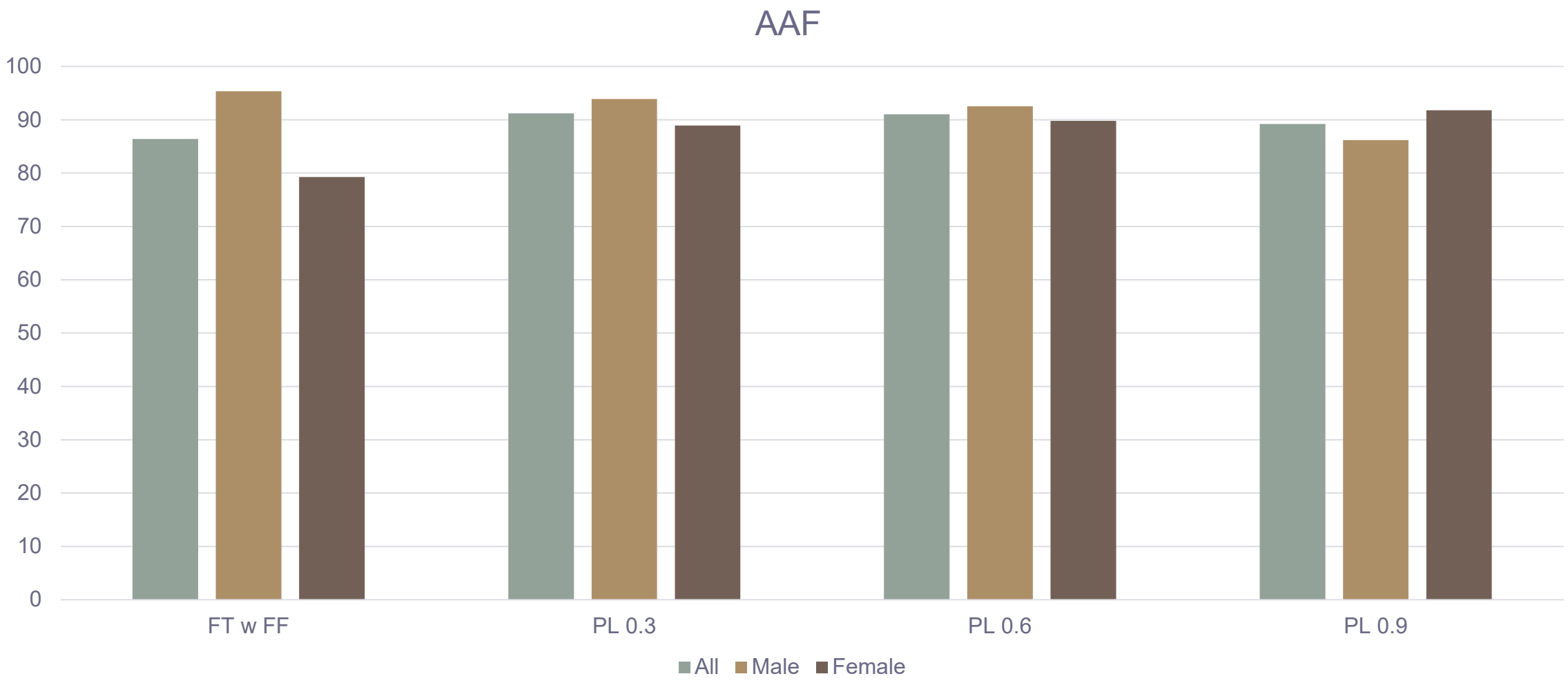



# Effect of fine-tuning

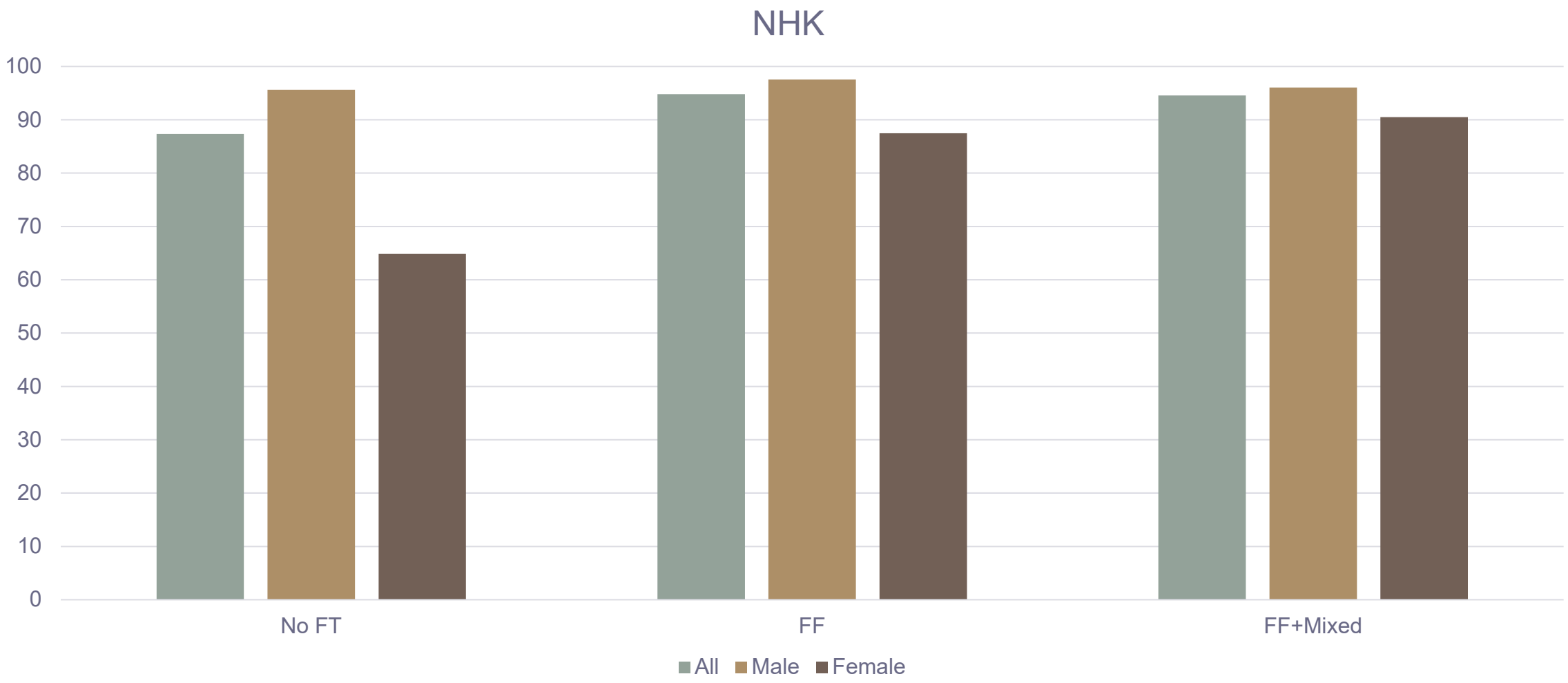
- Fine-tuning is very effective
- However, this is supervised fine-tuning
- Namely, we need ground truth information for the data to be used for fine tuning



# Pseudo labeling with NHK



# Performance on NHK



# Conclution

- Off-the-shelf gender classifiers are fairly biased
- Supervised fine tuning is confirmed to be very effective, but labeled data is needed
- Pseudo labeling is promising since this does not require labeled data
- However, we prepared balanced data (male vs female) and used it for pseudo labeling
- What happens if data to be used for unsupervised fine tuning is biased?

# Fairness Without Labels: Pseudo-Balancing for Bias Mitigation in Face Gender Classification

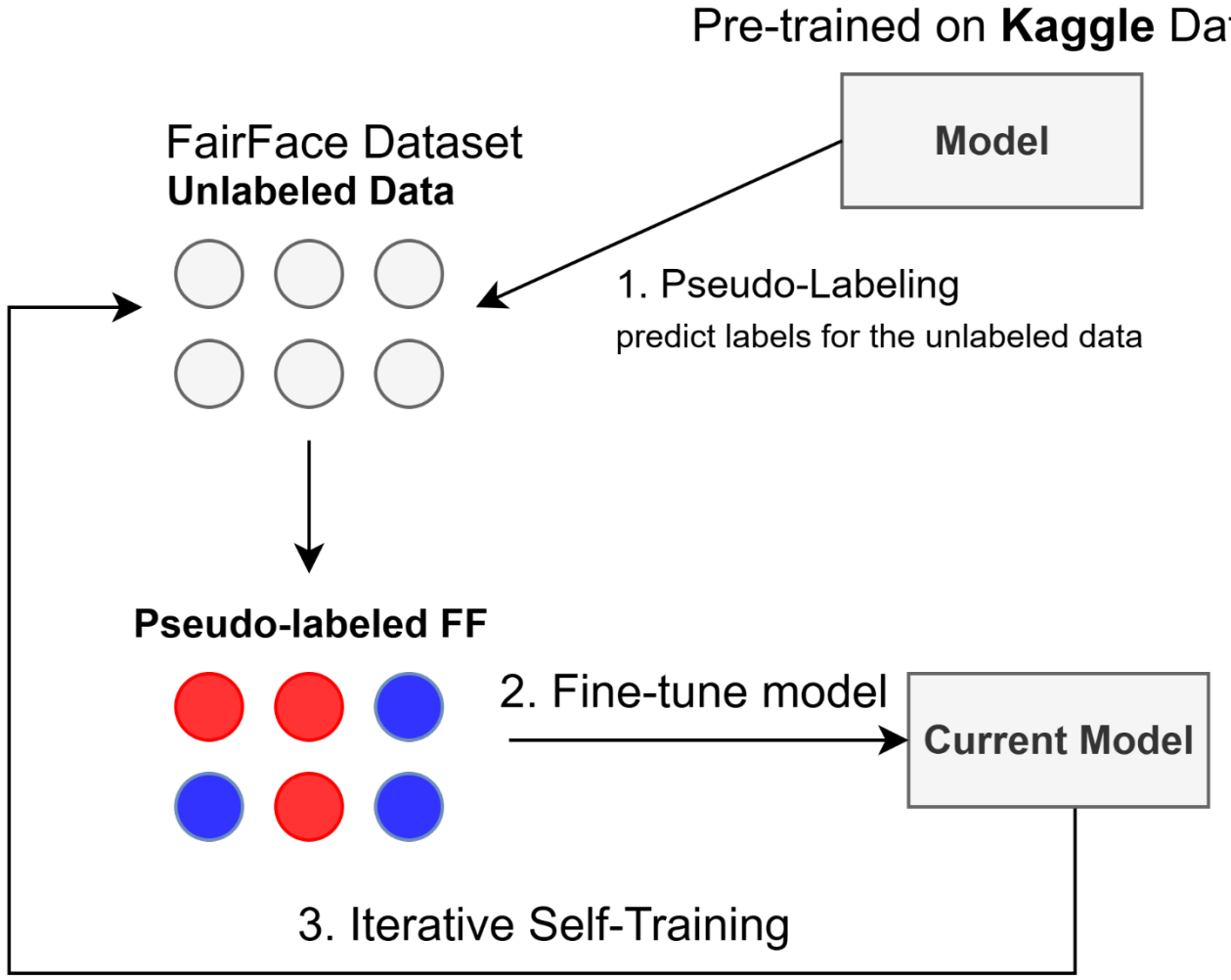
- Second workshop on Fairness and ethics towards transparent AI: facing the chalLEnge through model Debiasing (FAILED)
- ICCV 2025 Workshop
- Haohua Dong, Ana Manzano Rodríguez, Camille Guinaudeau, Shin'ichi Satoh




# Abstract


- Biased data, if used for unsupervised fine tuning, may even amplify the bias of the resultant classifier
- We propose very simple idea, pseudo balancing, on top of FixMatch, to mitigate biases in unsupervised fine tuning data

# Our Approach - Pseudo-Balancing




**Pseudo-Labeling**: Use model predictions on unlabeled data as training labels

- FixMatch: keep only high-confidence pseudo-labels (fixed-threshold)
- FlexMatch: dynamically adjust threshold depending on class learning difficulty

*How does this simple sampling technique mitigate bias in face gender classification?*

# Experimental Design

**Scenario 1 – Balanced Unlabeled Data**

- Pre-trained CNN on Kaggle Gender (biased)
- Self-train with unlabeled FairFace

**Scenario 2 – Biased Unlabeled Data**

- Stress-test with biased subsets of FairFace
  - East Asian
  - Black
  - Gender bias 80-20%
  - East Asian Female – this reflects the target domain's characteristics

FairFace

All-Age-Faces

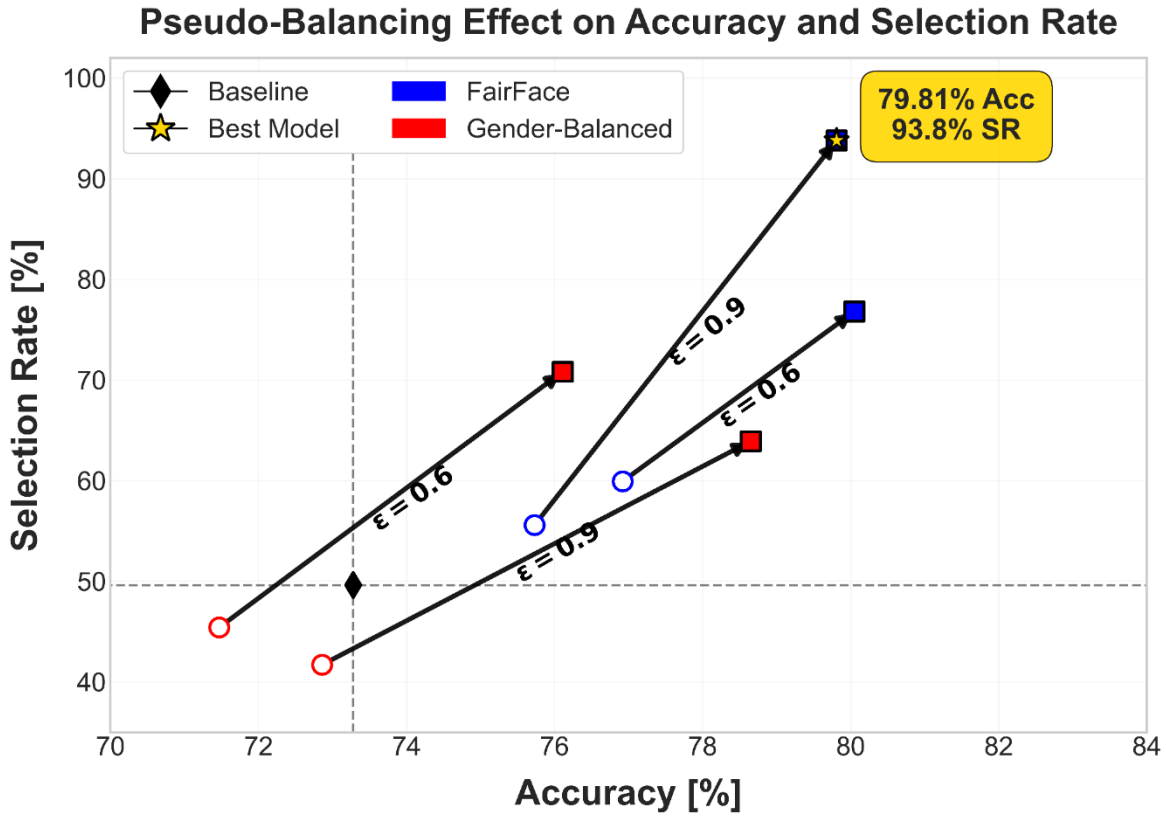
**Benchmark**: All-Age-Faces (mostly East Asians)

**Metrics**: Accuracy + Selection Rate (accuracy gap across genders)

[5] Cheng, J., Li, Y., Wang, J., Yu, L., & Wang, S. (2019). *Exploiting effective facial patches for robust gender recognition.* Tsinghua Sci. Technol., 24(3), 333–345.
[6] Karkkainen, K., & Joo, J. (2021). *FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation.* In WACV '21, pp. 1548–1558

# Key Results

| Model | PB | Accuracy [%] (Male / Female) | Selection Rate [%] |
|---|---|---|---|
| Baseline | - | 73.28<br>97.94 / 48.61 –<br>**49% gap** | 49.63 |
| FF<br>$\epsilon = 0.6$ | yes | 80.05<br>91.12 / 69.98 | 76.80 |
| **FF**<br>**$\epsilon = 0.9$** | **yes** | **79.81**<br>**82.37 / 77.26** | **93.80** |
| East-Asian Female<br>FF<br>$\epsilon = 0.9$ | no | 81.30<br>87.99 / 74.60 | 84.78 |




- +6.53% accuracy over baseline
- 44.17% reduction in gender gap

- **Improves Accuracy and SR in most cases**



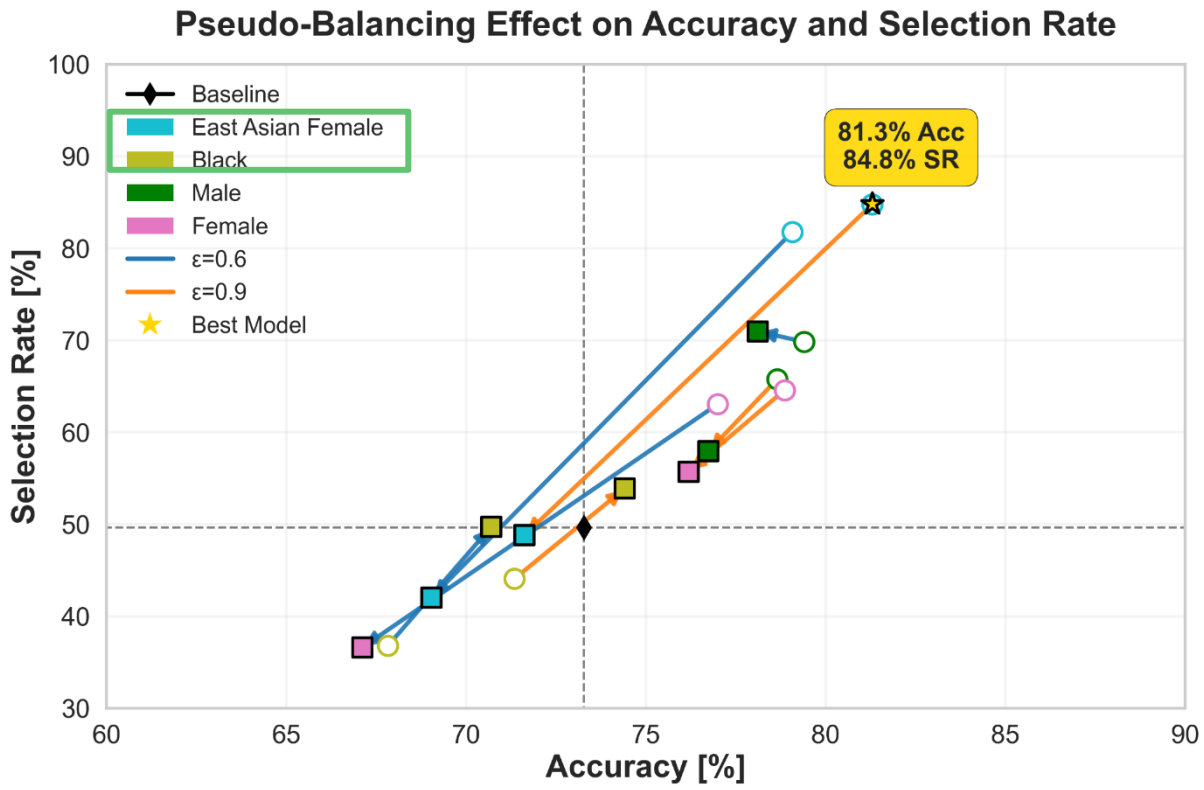
[7] Dong, H., Manzano Rodríguez, A., Guinaudeau, C., & Satoh, S. (2025). Fairness without labels: Pseudo-balancing for bias mitigation in face gender classification.

# When Does Pseudo-Balancing Work?

- Works best with **balanced or moderately biased** unlabeled data

- Fails if unlabeled data is **severely biased** and/or aligned with the **baseline model bias** (e.g., East Asian Female)

- Depends on:
    1. Initial baseline model bias
    2. Demographic composition of unlabeled data
    3. Alignment with target domain




✓ ***Pseudo-labeling with PB*** *improves accuracy and fairness when* the unlabeled data provides enough demographic diversity, but it breaks down if the data is too biased or narrow.

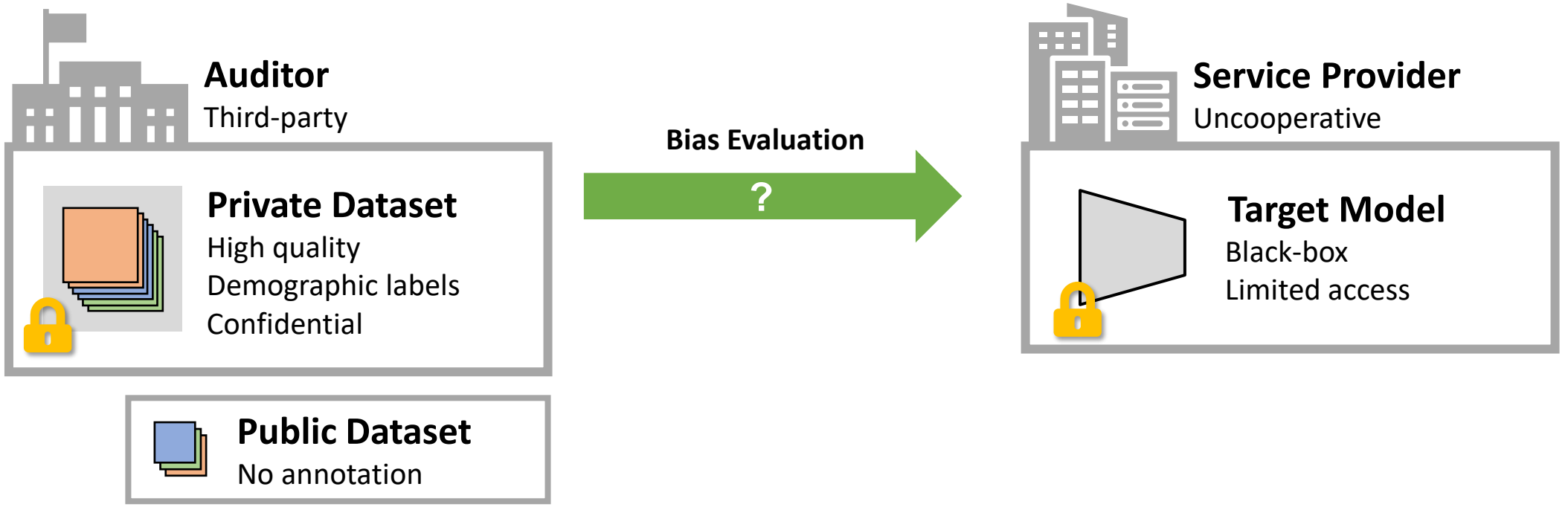


[7] Dong, H., Manzano Rodríguez, A., Guinaudeau, C., & Satoh, S. (2025). Fairness without labels: Pseudo-balancing for bias mitigation in face gender classification.

# Fairness Auditing of Limited-Access Models Using Protected Datasets

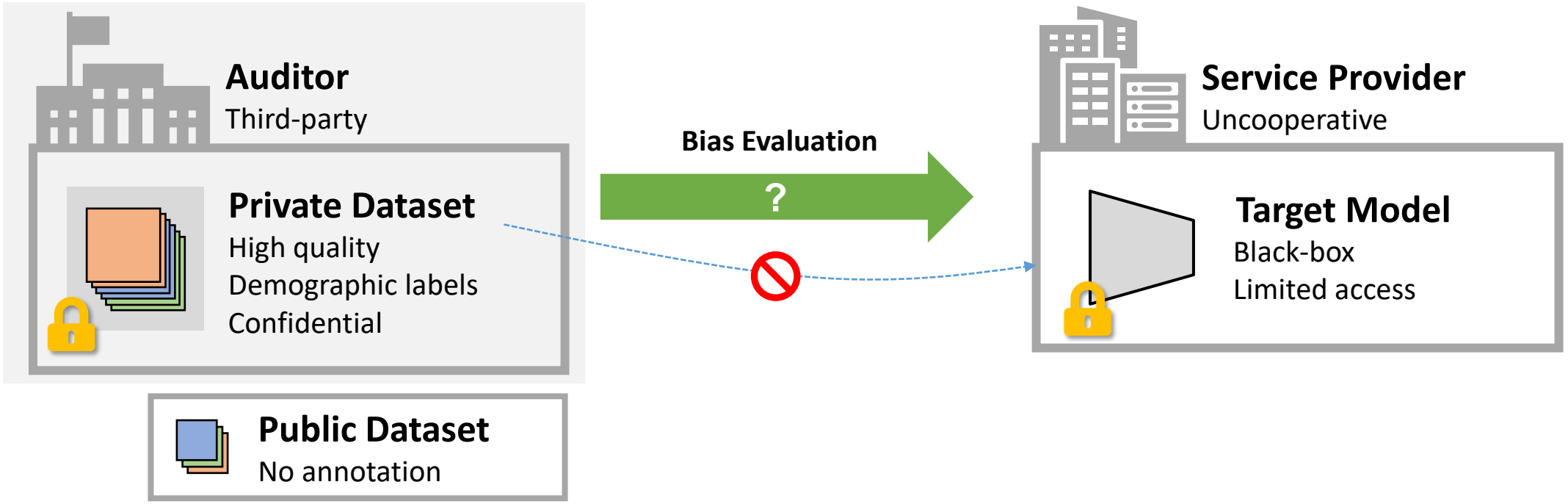
- IEEE Access, 2025
- Zhaohui Zhu, Marc A Kastner, and Shin’ichi Satoh

# Fairness Auditing Scenario



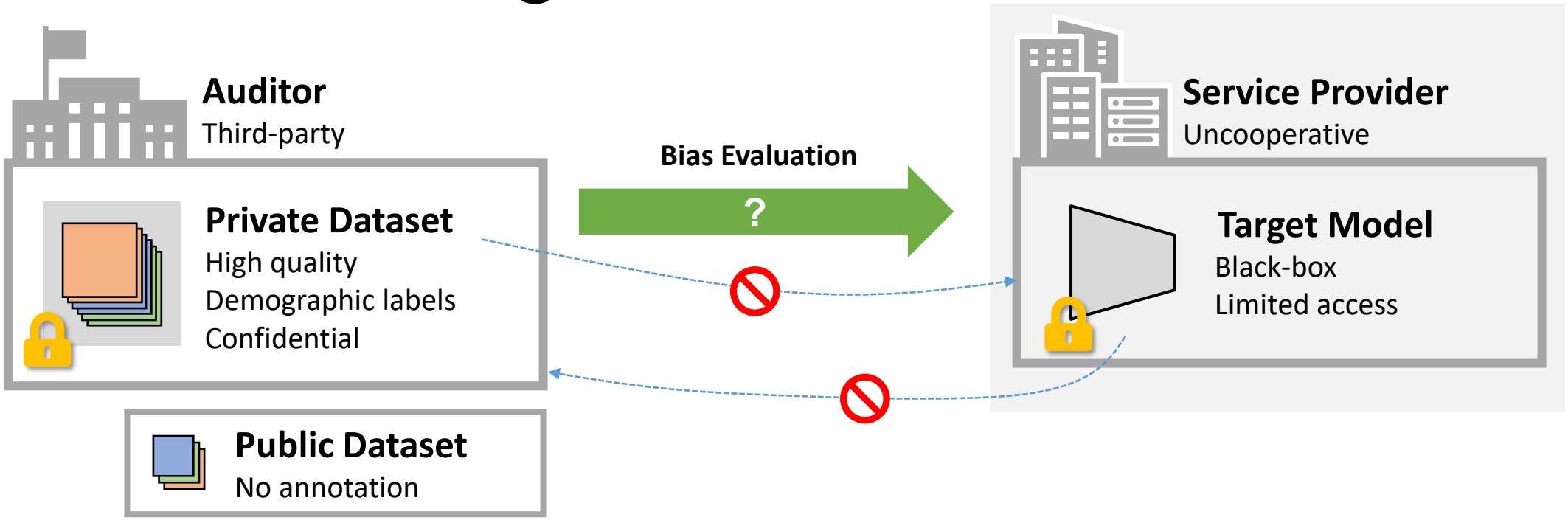

- Practical yet underexplored scenario

# Fairness Auditing Scenario



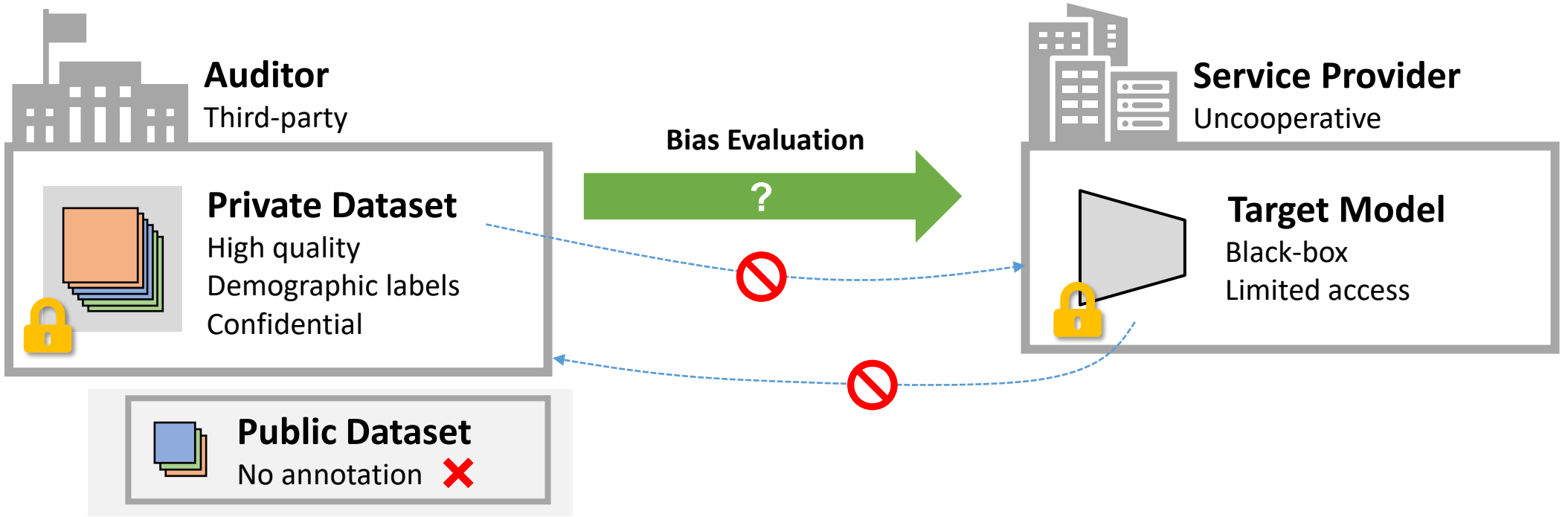

- **Auditor:** regulatory agencies / certification bodies (NIST, FDA, …)
- **Private Dataset:** demographic annotations, balanced
  - **Confidential**: Avoid test data leakage; Sensitive domains
- ***Objective:*** evaluate the bias of target model on the private dataset

# Fairness Auditing Scenario



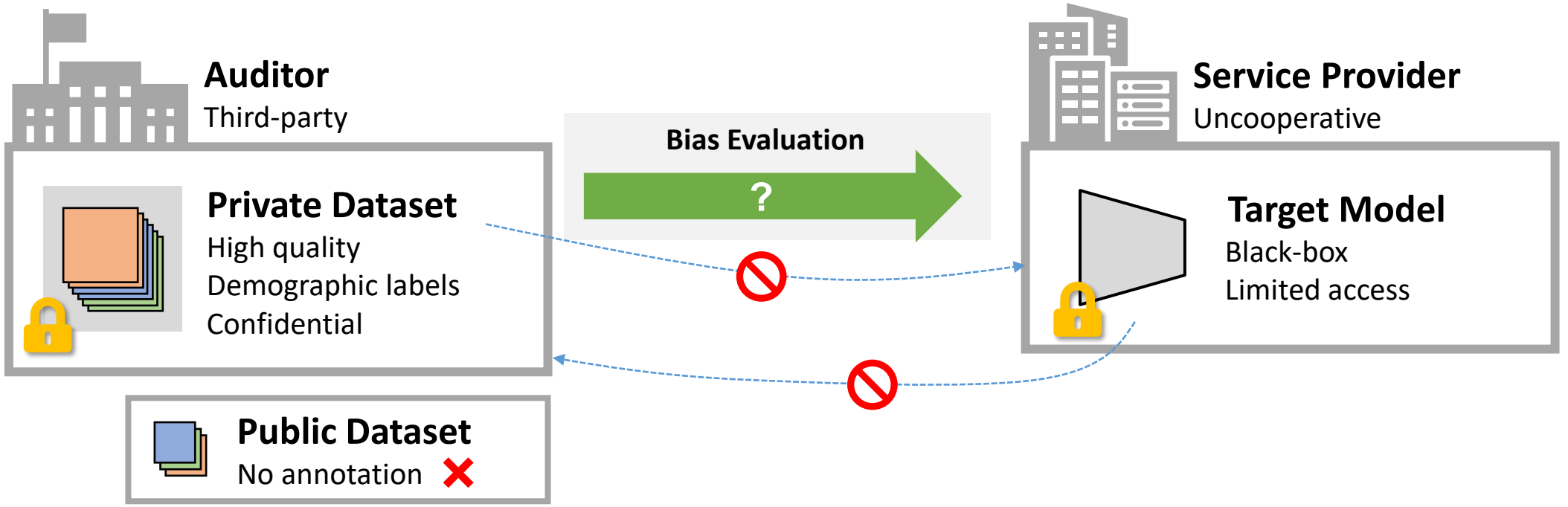

- **Service Provider:** companies / developers (uncooperative)
- **Target Model:**
  - Black-box: APIs only; no external deployment (privacy, IP)
  - Limited access: no massive queries (prevent reverse engineering; cost)

# Fairness Auditing Scenario




- **Public Dataset:**
  - No demographic annotation
  - Not balanced
    ⇏ fairness evaluation

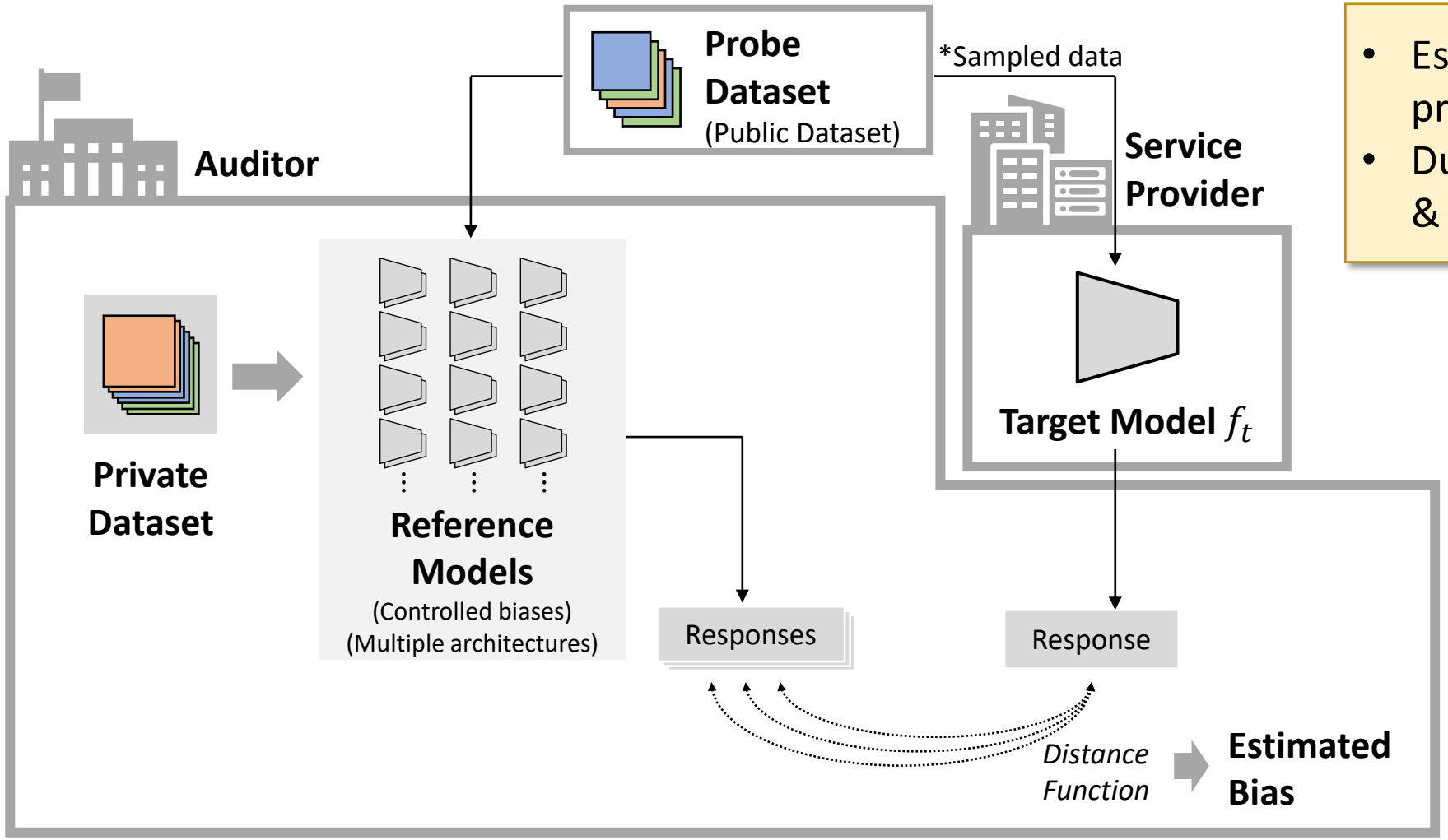
# Fairness Auditing Scenario




- Conventional approach does not work: mutual untrusted
  - ✗ Send the private dataset ✗ Use the public dataset ✗ Test inside the auditor
- Other techniques (model extraction, cryptographic, …)
  - Limitations: need significant cooperation, …
- **Challenge**: evaluate/estimate the bias under dual constraints

# Contribution: Fairness Auditing Framework

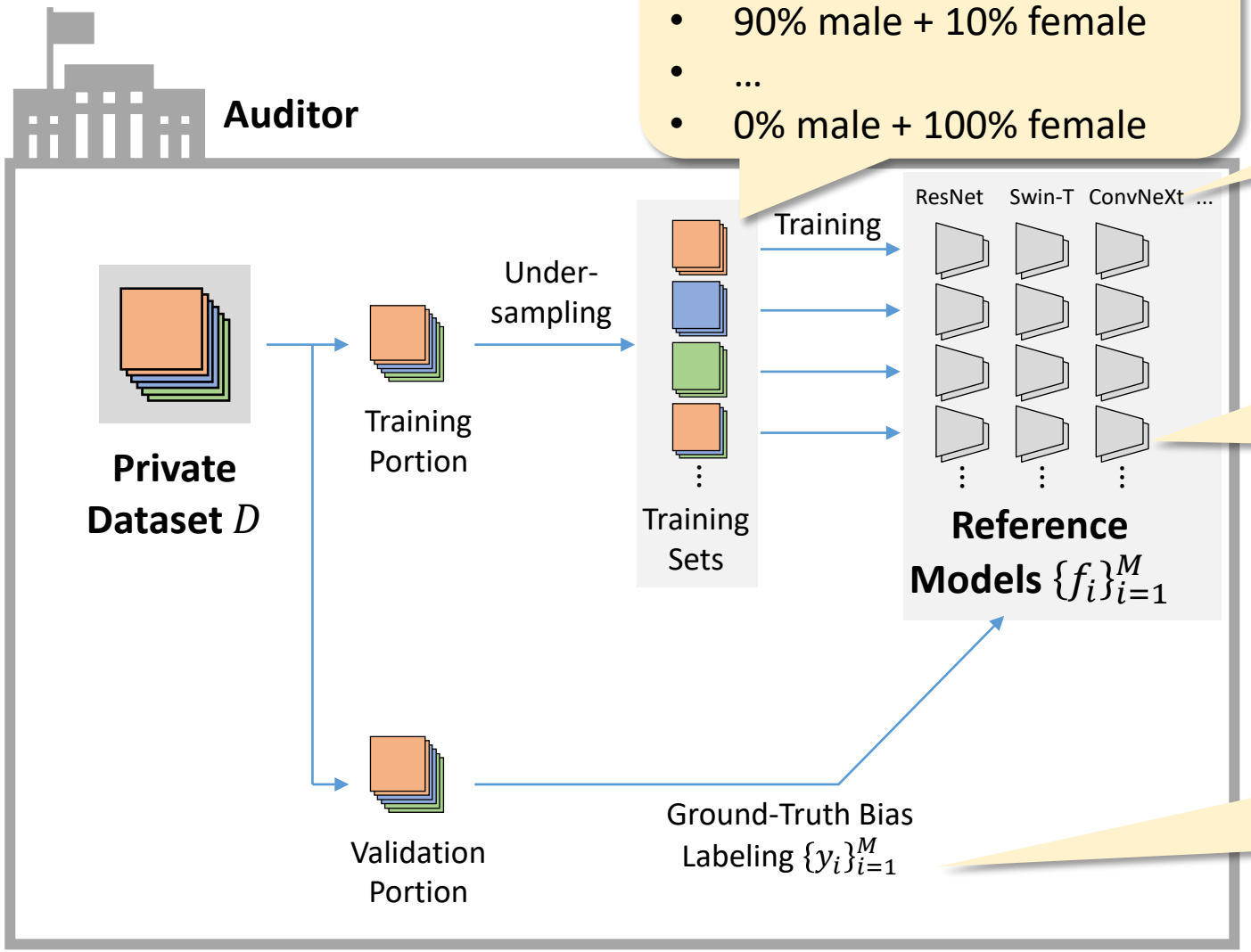
- Goal: fairness metric of target model (e.g., facial age estimation models) on private dataset
  - E.g., gender MAE gap: $y_t = \Delta_{\mathrm{MAE},t} = \mathrm{MAE}_{\mathrm{male},t} - \mathrm{MAE}_{\mathrm{female},t}$




- Estimate without exposing private dataset
- Dual constraints: protected data & limited access

# Fairness Auditing Framework

- **Step 1: Reference Models**




**Controlled bias,** E.g.:
- 100% male + 0% female
- 90% male + 10% female
- …
- 0% male + 100% female

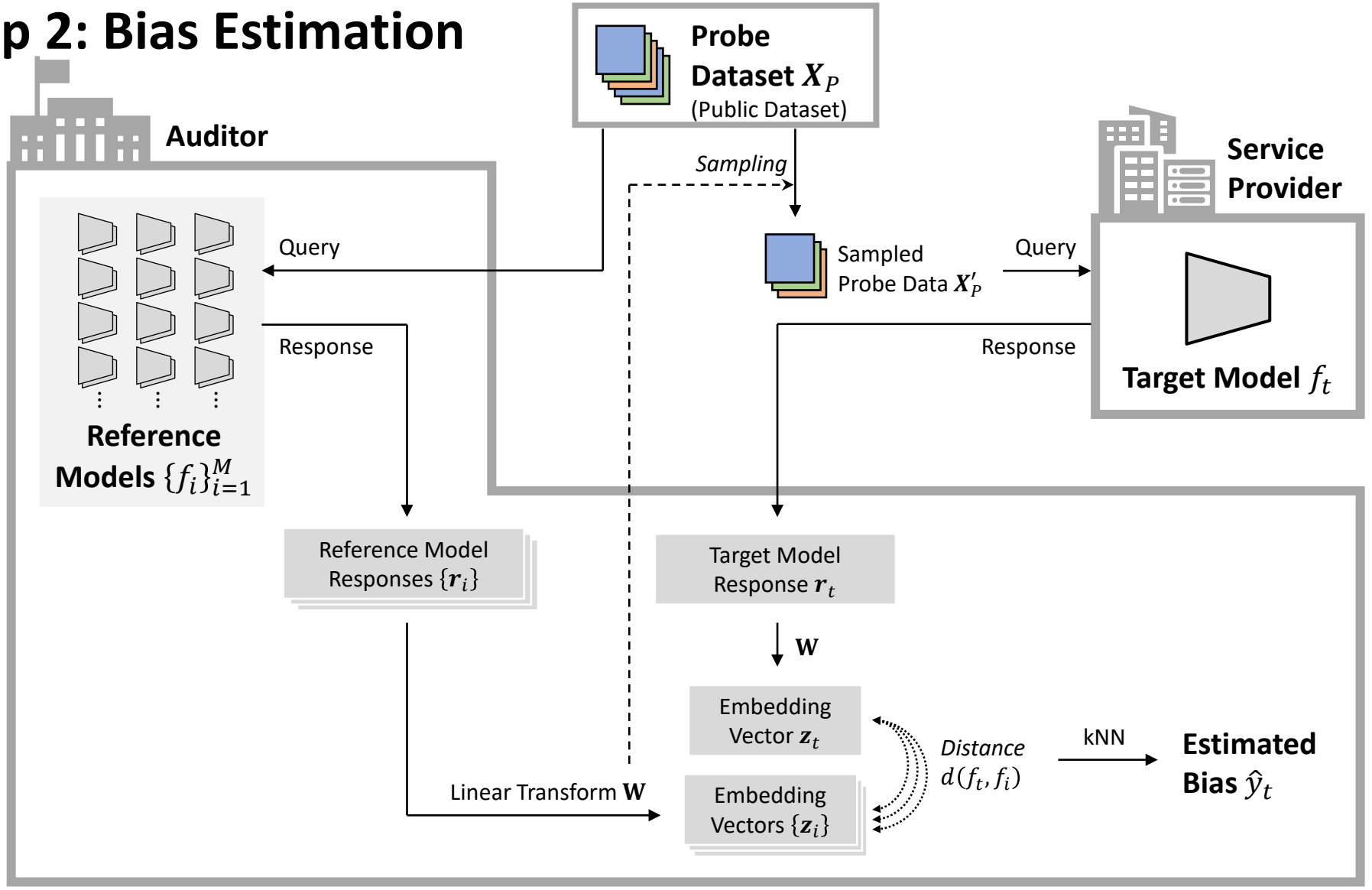
**Multiple model architectures**
- CNN-based
- Transformer-based

**Random variations**
- Random label shuffling
- Random subsets
- Random seeds

**Ground-truth bias label**
E.g.: gender MAE gap
$y_i^g = \Delta_{\mathrm{MAE},i}^g = \mathrm{MAE}_{\mathrm{male},i} - \mathrm{MAE}_{\mathrm{female},i}$

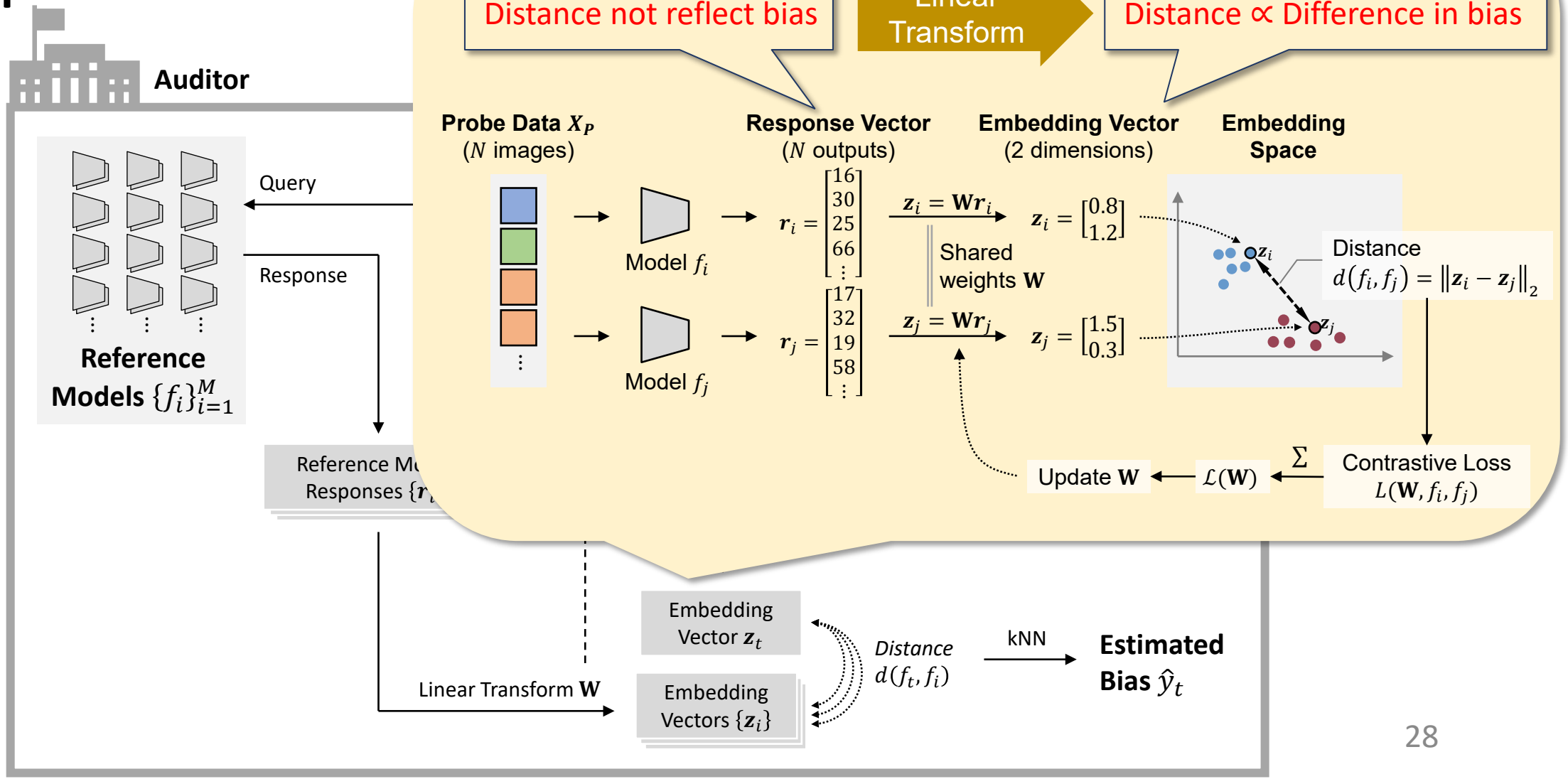
# Fairness Auditing Framework

- **Step 2: Bias Estimation**



# Fairness Auditing Framework

- **Step 2: Bias Estimation**



# Fairness Auditing Framewor

- **Step 2: Bias Estimation**

**Static sampling**

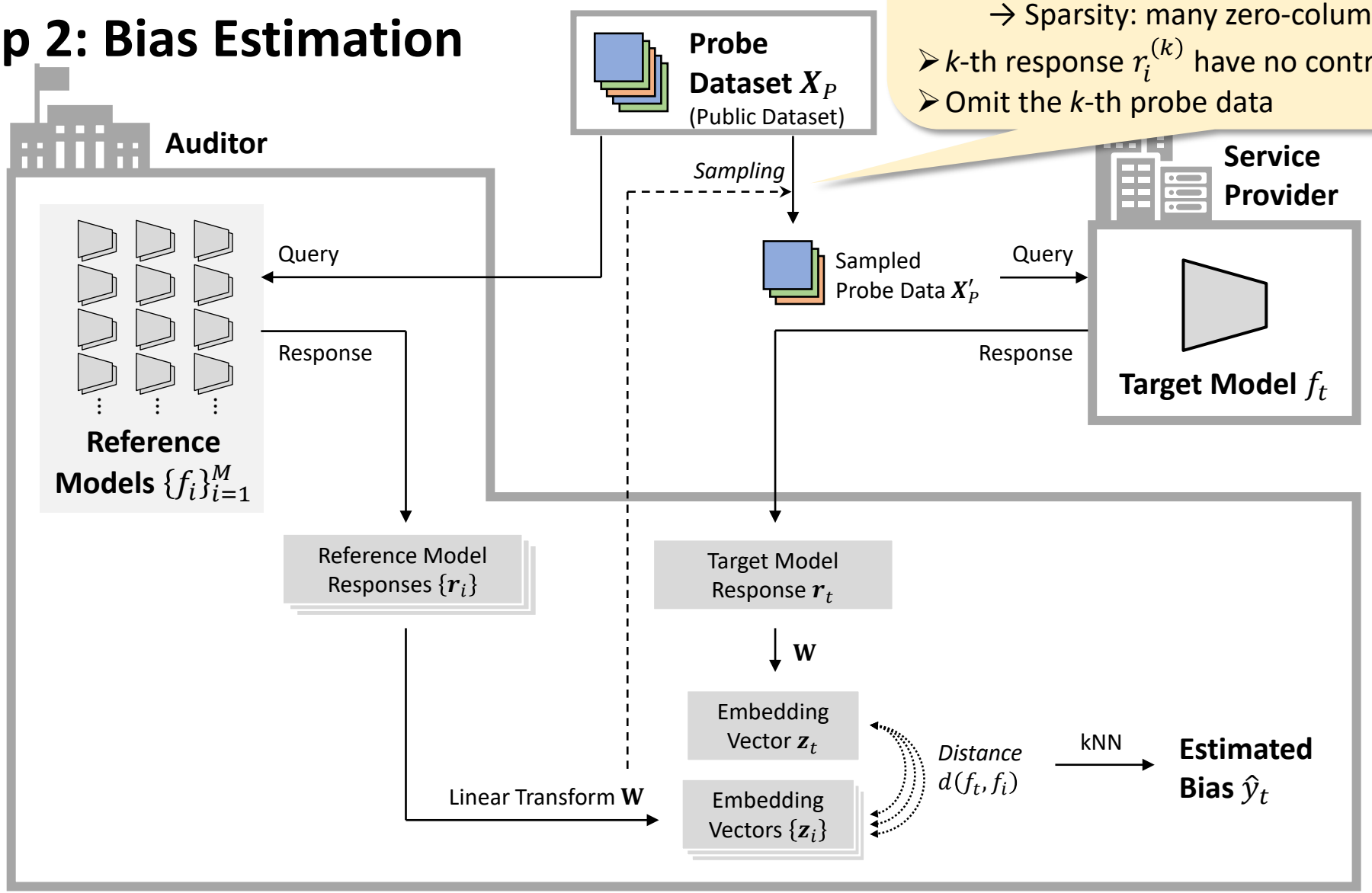
- L1 regularization (entry-wise $L_{2,1}$ norm):

$$\|\mathbf{W}\|_{2,1} = \sum_{k=1}^{N} \|\boldsymbol{w}_k\|_2 = \sum_{k=1}^{N} \sqrt{w_{1,k}^2 + w_{2,k}^2}$$

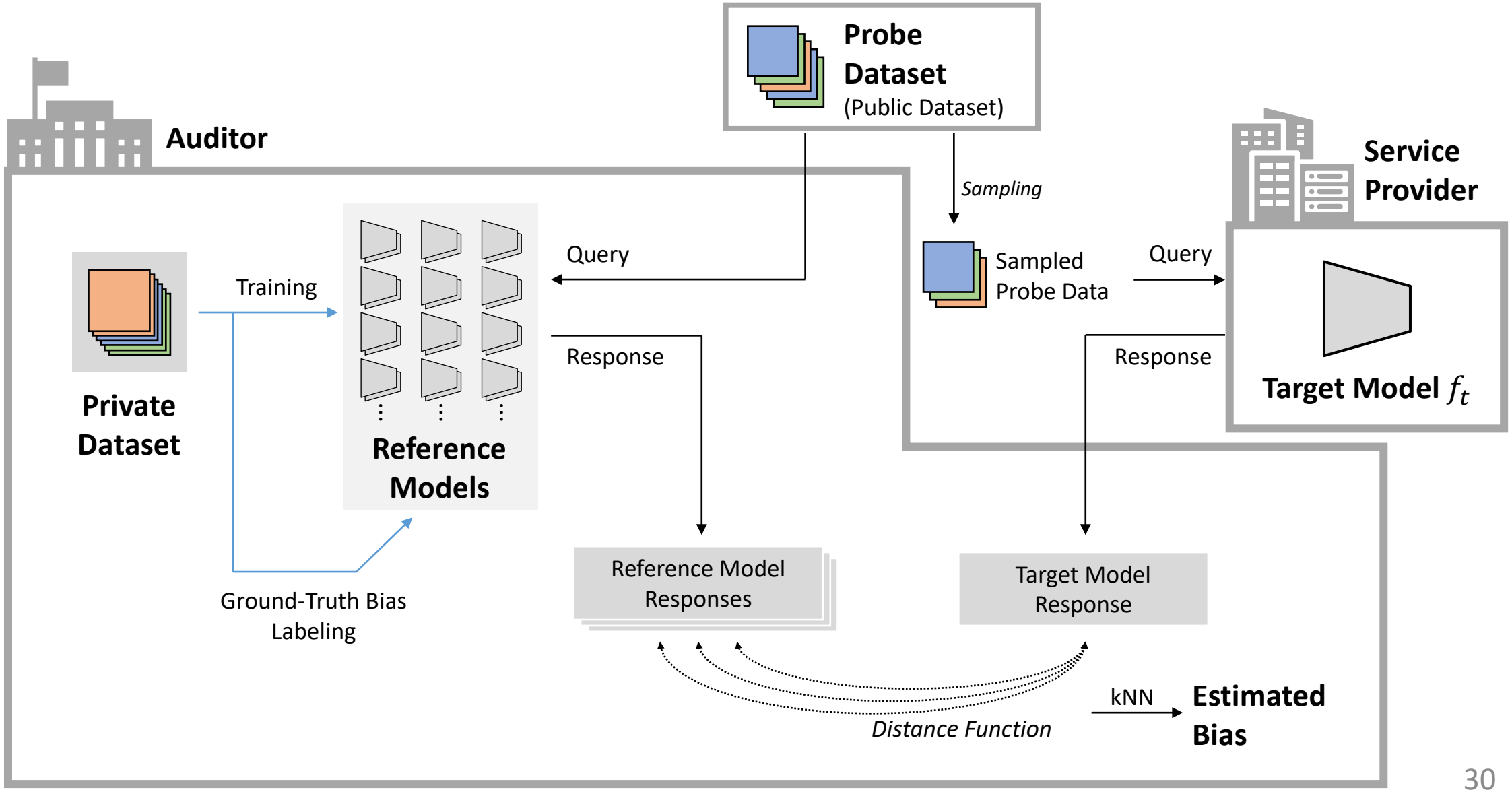
→ Sparsity: many zero-columns $\boldsymbol{w}_k \approx \mathbf{0}$

➢ $k$-th response $r_i^{(k)}$ have no contribution to $\boldsymbol{z}_i = \mathbf{W}\boldsymbol{r}_i$

➢ Omit the $k$-th probe data



# Fairness Auditing Framework



# Evaluation

- Experimental Setup
  1. **Gender bias and racial bias of facial age estimation models**
  2. Gender bias and racial bias of facial gender classification models

  - **Gender Bias:** *gender MAE gap*

$$\Delta^{g}_{\mathrm{MAE}} = \mathrm{MAE}_{\mathrm{male}} - \mathrm{MAE}_{\mathrm{female}}.$$

    - Positive value: Biased against males
    - Negative value: Biased against females

  - **Racial Bias:** *racial MAE gap*

    - Bias towards Black $\Delta^{r,B}_{\mathrm{MAE}} = \mathrm{MAE}_{\mathrm{Black}} - \mathrm{MAE}_{\mathrm{all}},$
    - Bias towards East Asian $\Delta^{r,E}_{\mathrm{MAE}} = \mathrm{MAE}_{\mathrm{East\ Asian}} - \mathrm{MAE}_{\mathrm{all}},$

# Evaluation

- Datasets
  - Private dataset: simulated by FairFace
  - Probe dataset: simulated by UTKFace

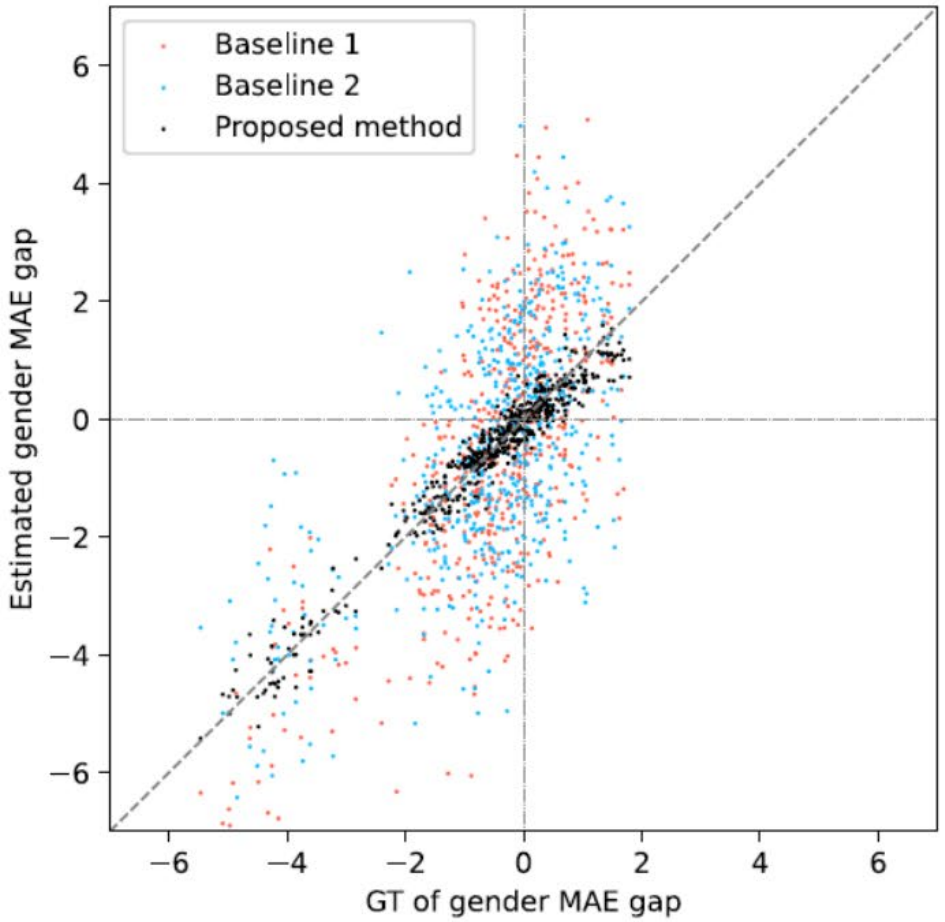
- Baselines: conventional ways with compromise
  - Baseline 1: assume a **public dataset** with annotations (test on UTKFace)
  - Baseline 2: leak a subset of **private dataset** (test on FairFace)

- Experimental Setup
  1. Cross validation on internal models
     - Internal models: trained on private dataset with controlled bias (2,200 models for gender bias; 3,200 models for racial bias)
     - Four model architectures: ResNet-18, EfficientNet-B1, ConvNeXt-T, Swin-T
     - Leave-one-architecture-out cross-validation

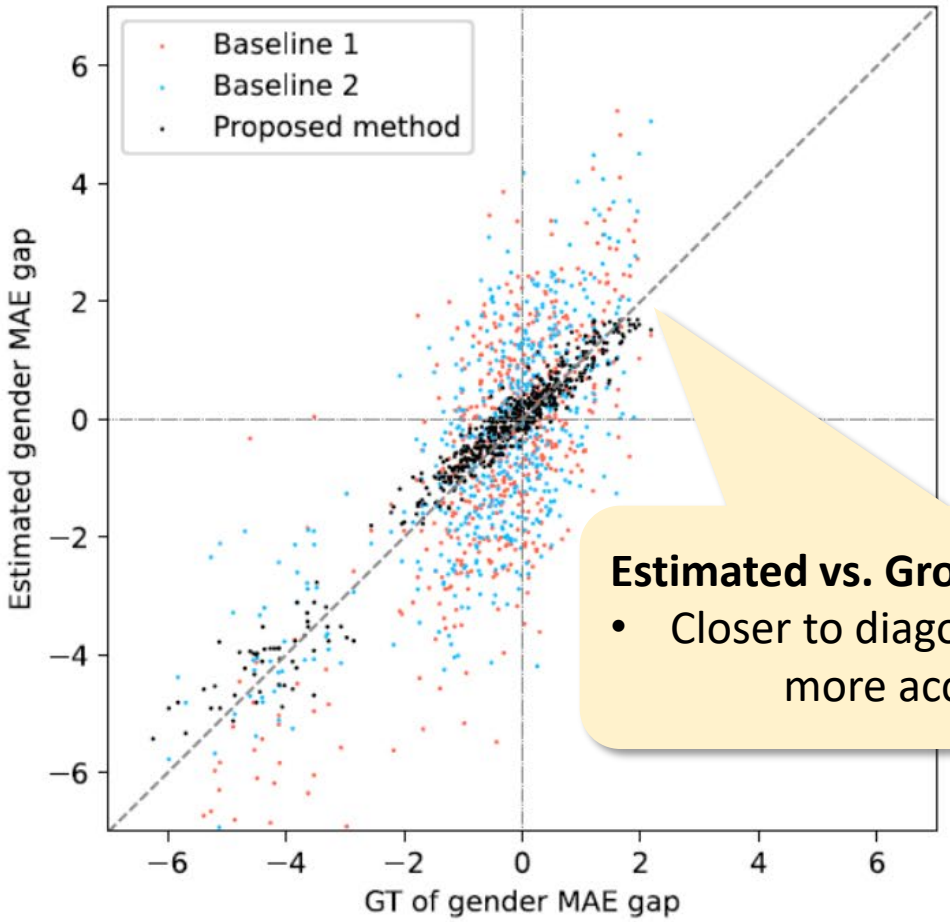
  2. Evaluation on external models
     - Target models: pre-trained models from external sources (4 models)
     - Reference models: Internal models (all 2,200/3,200 models of 4 architectures)

# Evaluation

- Results: **Gender Bias Estimation** (Internal models)




(a) Target Models: ResNet-18




(b) Target Models: EfficientNet-B1

**Estimated vs. Ground-truth**

- Closer to diagonal:
  more accurate than baselines

Other architectures:
similar results

# Evaluation

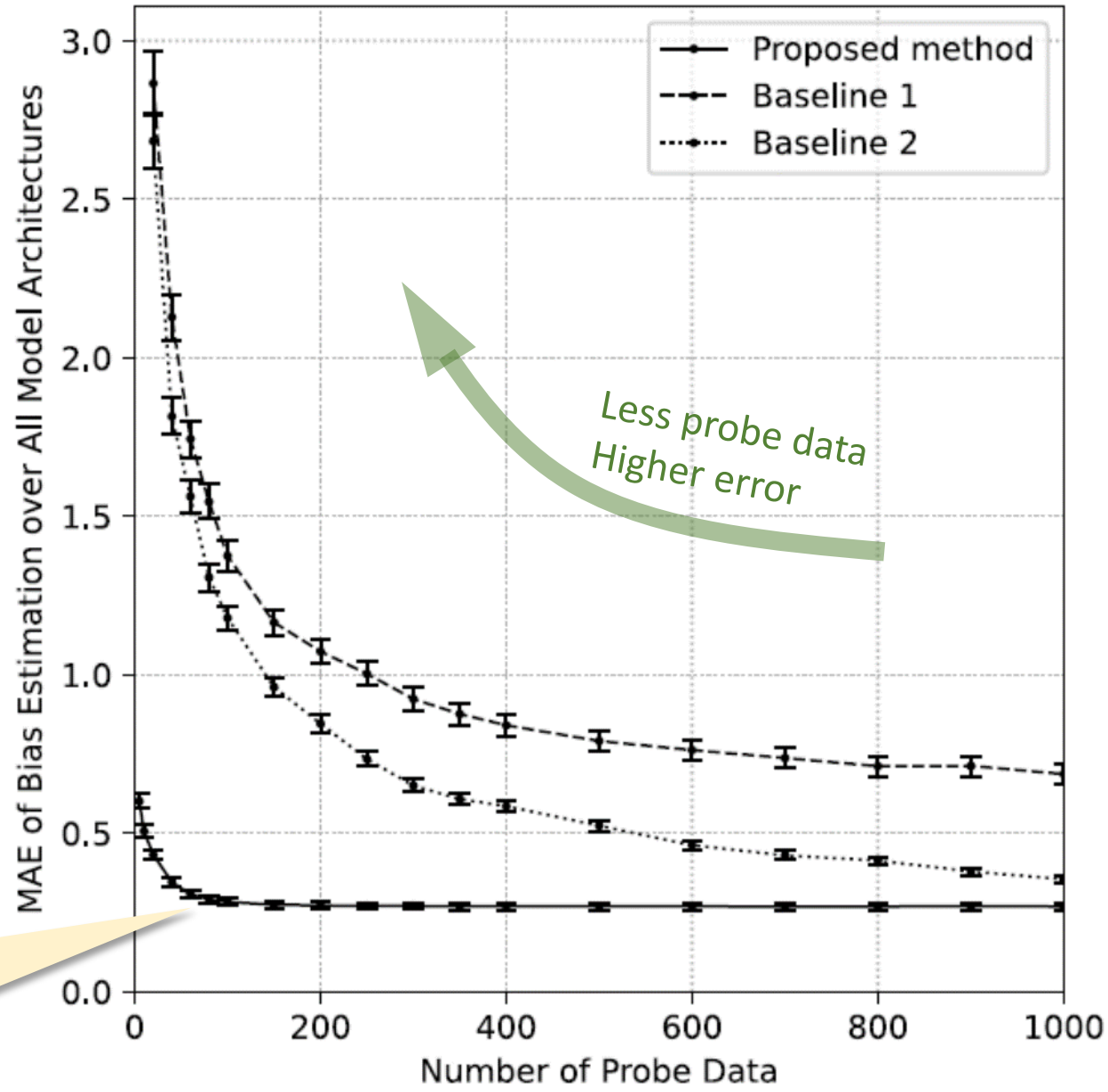
- Results: **Gender Bias** and **Racial Bias** Estimation(Internal models)
  - Metric: MAE of estimated MAE gap $\mathrm{MAE}_{\Delta^g_{\mathrm{MAE}}} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\Delta}^g_{\mathrm{MAE},i} - \Delta^g_{\mathrm{MAE},i}\right|$

| **Fairness Metrics** | **Probe Data** | Proposed Method | Baseline 1 | ($p$-value) | Baseline 2 | ($p$-value) |
|---|---|---|---|---|---|---|
| Gender Bias | 100 images | **0.28** $\pm$ 0.01 | 1.38 $\pm$ 0.05 | $p < 0.001$ | 1.18 $\pm$ 0.04 | $p < 0.001$ |
| | 1000 images | **0.27** $\pm$ 0.01 | 0.69 $\pm$ 0.03 | $p < 0.001$ | 0.35 $\pm$ 0.01 | $p < 0.001$ |
| Racial Bias Towards Black | 100 images | **0.23** $\pm$ 0.01 | 1.30 $\pm$ 0.04 | $p < 0.001$ | 1.51 $\pm$ 0.04 | $p < 0.001$ |
| | 1000 images | **0.22** $\pm$ 0.01 | 0.62 $\pm$ 0.02 | $p < 0.001$ | 0.46 $\pm$ 0.01 | $p < 0.001$ |
| Racial Bias Towards East Asian | 100 images | **0.24** $\pm$ 0.01 | 2.04 $\pm$ 0.05 | $p < 0.001$ | 1.53 $\pm$ 0.04 | $p < 0.001$ |
| | 1000 images | **0.20** $\pm$ 0.01 | 1.63 $\pm$ 0.02 | $p < 0.001$ | 0.48 $\pm$ 0.01 | $p < 0.001$ |

✓ Outperforms baselines

# Evaluation

- Results: **Probe Data Sampling**
  (Gender bias, internal models)




Proposed method:
Keep low error until
**60 – 80 probe data**

# Evaluation

- Experimental Setup
  2. Evaluation on external models
     - Reference models: Internal models (all 2,200/3,200 models of 4 architectures)
     - Target models: pre-trained models from external sources (4 models)

| Model | Architecture | Training Data |
|---|---|---|
| MiVOLO-D1 [53] | VOLO (Transformer) | IMDB-cleaned [59] |
| Levi-CNN [56] | Shallow Custom CNN | Adience [25] |
| MIVIA-MobileNet [29] | MobileNet | VMAGE |
| MIVIA-SENet [29] | SENet | VMAGE |

- Results

| Fairness Metrics | Proposed Method | Baseline 1 | Baseline 2 |
|---|---|---|---|
| Gender | **0.37** | 0.92 | 1.11 |
| Racial (Black) | **0.36** | 0.81 | 1.63 |
| Racial (East Asian) | **0.52** | 1.42 | 1.37 |

✓ Outperforms baselines

# Evaluation

- Additional task
  1. Gender bias and racial bias of facial age estimation models
  2. Gender bias and racial bias of **facial gender classification models**
- Results: Internal models

| Fairness Metrics | Probe Data | Proposed Method | Baseline 1 | ($p$-value) | Baseline 2 | ($p$-value) |
|---|---|---|---|---|---|---|
| Gender Bias | 100 images | **0.050** ± 0.002 | 0.073 ± 0.003 | $p < 0.001$ | 0.053 ± 0.002 | $p = 0.081$ |
| | 1000 images | 0.040 ± 0.002 | 0.057 ± 0.002 | $p < 0.001$ | **0.016** ± 0.001 | $p < 0.001$ |
| Racial Bias Towards Black | 100 images | **0.012** ± 0.000 | 0.084 ± 0.002 | $p < 0.001$ | 0.073 ± 0.002 | $p < 0.001$ |
| | 1000 images | **0.012** ± 0.000 | 0.077 ± 0.001 | $p < 0.001$ | 0.022 ± 0.001 | $p < 0.001$ |
| Racial Bias Towards East Asian | 100 images | **0.011** ± 0.000 | 0.075 ± 0.002 | $p < 0.001$ | 0.067 ± 0.002 | $p < 0.001$ |
| | 1000 images | **0.010** ± 0.000 | 0.056 ± 0.001 | $p < 0.001$ | 0.020 ± 0.001 | $p < 0.001$ |

- Results: External models

| Model | Architecture | Training Data |
|---|---|---|
| MiVOLO-D1 [53] | VOLO (Transformer) | IMDB-cleaned [59] |
| Kaggle-ResNet | ResNet-18 | Kaggle [40] |
| DeepFace [91] | VGG-Face | IMDB-WIKI [88, 87] |

| Fairness Metrics | Proposed Method | Baseline 1 | Baseline 2 |
|---|---|---|---|
| Gender | 0.074 | 0.147 | **0.038** |
| Racial (Black) | **0.004** | 0.127 | 0.076 |
| Racial (East Asian) | **0.019** | 0.136 | 0.042 |

- Consistent results: generalizability
- Baseline 2 sometimes better (private data leakage)
- External models: challenging due to domain gap

# Summary

- A novel fairness auditing framework
  - **Effective, accurate, robust**

- Tackle with dual constraints → Regulatory and certification scenario

- Insights
  - Diversity of reference models
  - Representativeness of probe data
  - Domain shift

# Conclusion

- Machine Learning and AI are good at finding patterns, i.e., biases, in training dataset
- This is why ML and AI may have danger to be trapped into biases
- There are obviously biases causing negative impact, in terms of poor performance as well as societally negative impact
- But “biases” can be useful information
- We have to be very careful to know the nature of ML and AI in handling biases