# InfoFair: Information-Theoretical Intersectional Fairness



Jian Kang[1]


Tiankai Xie[2]


Xintao Wu[3]


Ross Maciejewski[2]

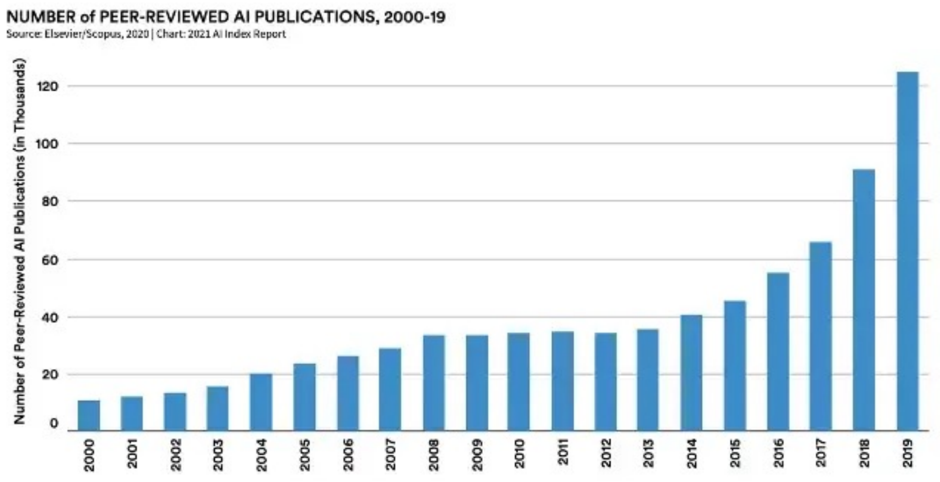

Hanghang Tong[1]

[1] University of Illinois at Urbana-Champaign

[2] Arizona State University

[3] University of Arkansas

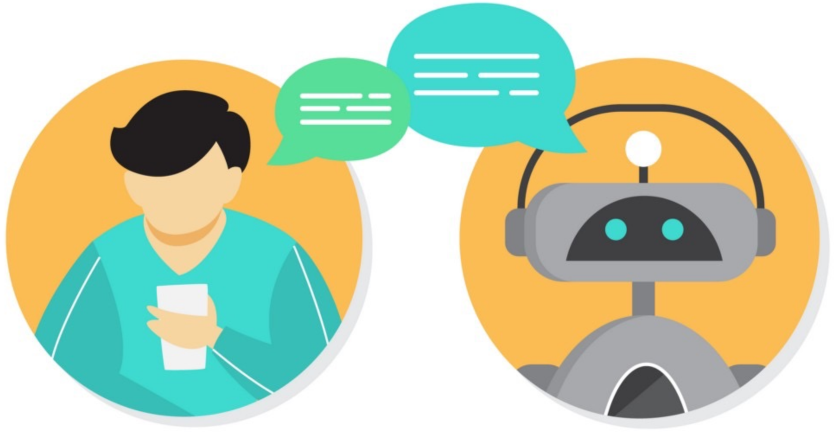
# Rise of Machine Learning




Number of publications in artificial intelligence/machine learning


E-commerce


Object detection

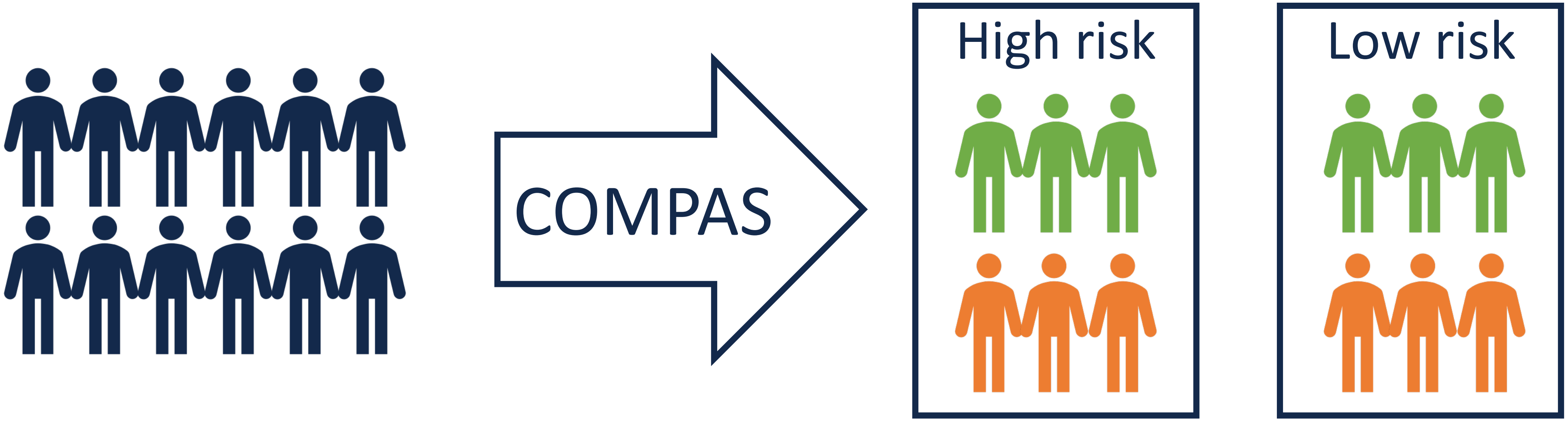


Question answering

[1] https://cekicbaris.medium.com/history-of-deep-learning-72144ebc9d44
[2] Wu, L., He, X., Wang, X., Zhang, K., & Wang, M.. A Survey on Neural Recommendation: From Collaborative Filtering to Content and Context Enriched Recommendation. arXiv 2021.
[3] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2022). YOLOv7: Trainable Bag-of-freebies Sets New State-of-the-art for Real-time Object Detectors. arXiv 2022.
[4] Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J.. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. NAACL 2021.

# Machine Learning Could Be Unfair

- **Example:** COMPAS
  - A risk assessment system to evaluate whether an individual would re-offend a crime

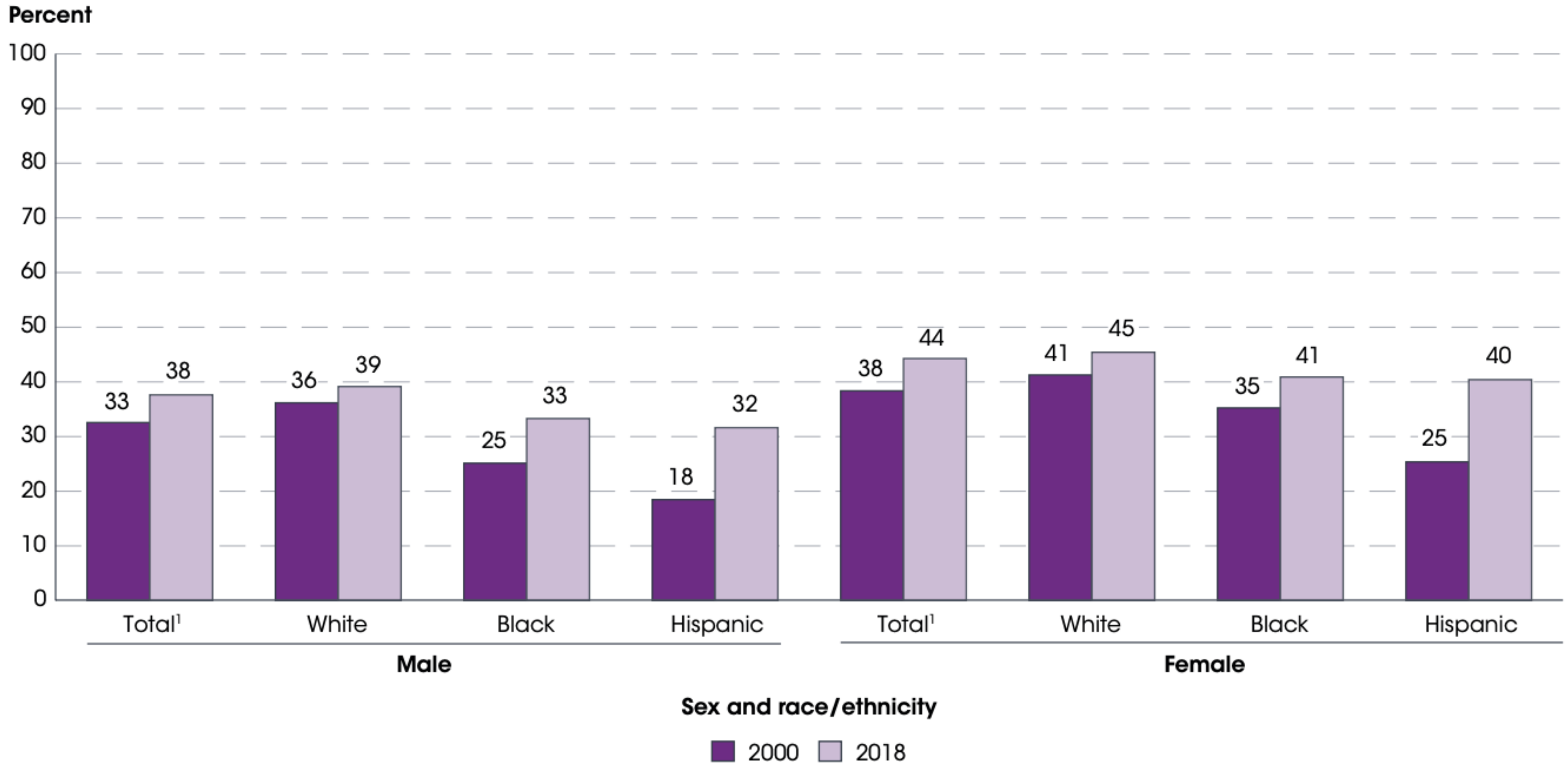



| | Orange | Green |
|---|---|---|
| **labeled high risk, but didn’t re-offend** | **23.5%** | **44.9%** |
| **labeled low risk, but did re-offend** | **47.4%** | **28.0%** |

* In this example, we use the imaginary race groups (green and orange) to avoid potential offenses.
[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

# Unfairness: Multiple Sensitive Attribute

- **Example:** college admission



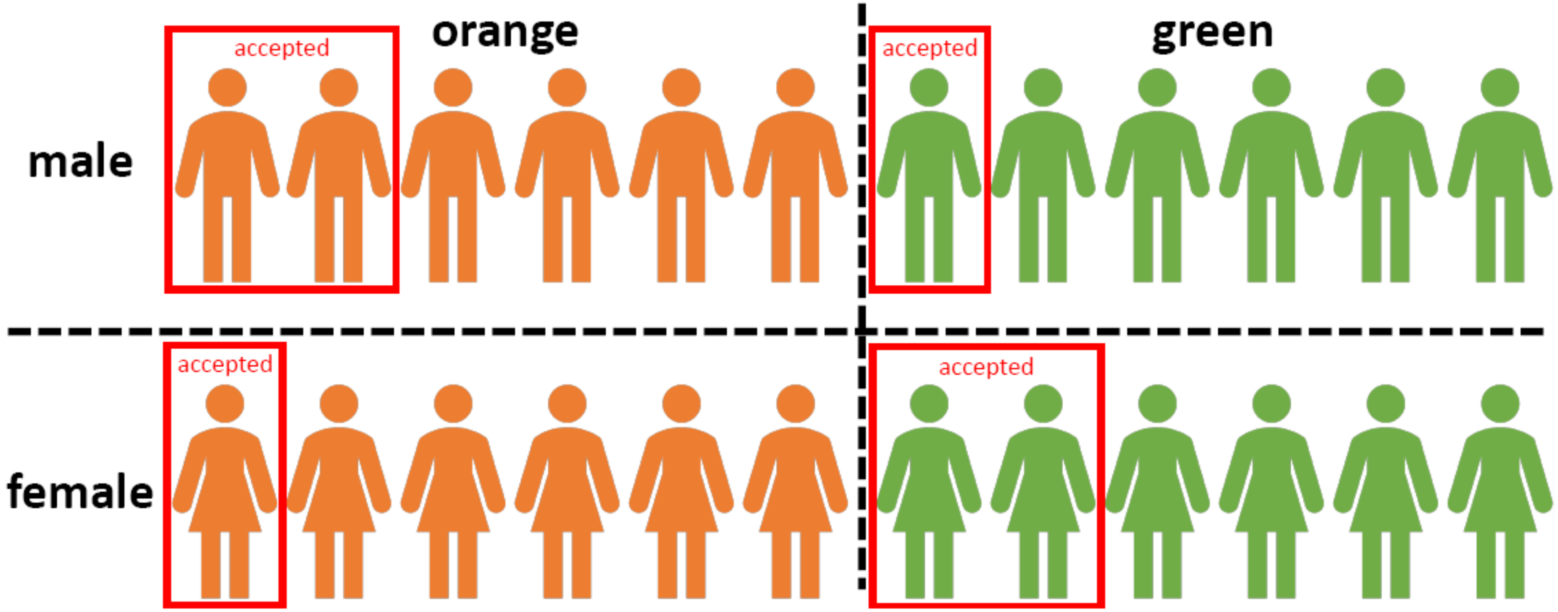

- **Observation:** the admission decision is unfair when we consider sex and race/ethnicity simultaneously

* In this example, we consider the binary biological sex. However, the gender identity of an individual could be non-binary.


[1] Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., ... & Dilig, R.. The Condition of Education 2020. NCES 2020.

# Existing Works: What to Debias

- **What to debias**
  - **Key idea:** debias multiple distinct sensitive attribute
  - **Examples:** compositional fairness
  - **Limitation:** fail to guarantee fairness on the fine-grained groups formed by multiple sensitive attributes
- **Examples**




* In this example, we consider the binary biological sex. However, the gender identity of an individual could be non-binary.


[1] Bose, A., & Hamilton, W.. Compositional Fairness Constraints for Graph Embeddings. ICML 2019.

# Existing Works: How to Debias

- **How to debias**
  - **Key idea:** optimize a surrogate constraints of group fairness
  - **Examples:** adversarial debiasing, linear correlation optimization
  - **Limitation:** achieve fairness unless the well-trained module that mitigates the bias could perfectly learn the mapping between sensitive attribute and model outcomes
- **Question:** can we achieve group fairness
  - With respect to multiple sensitive attributes simultaneously
  - Without optimizing a surrogate constraint

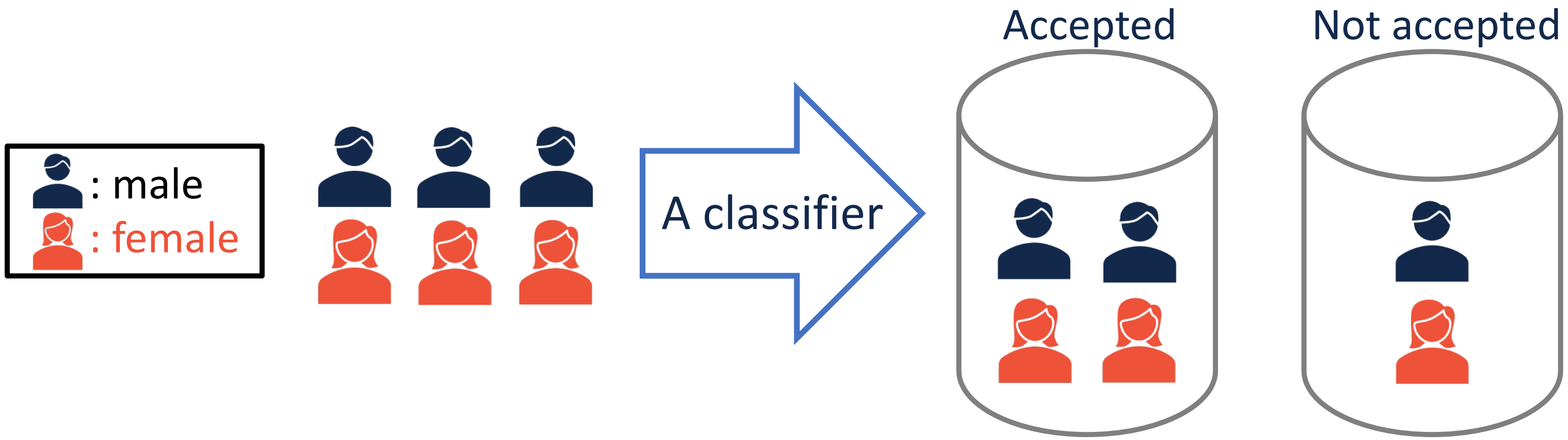
[1] Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P.. Fairness Constraints: Mechanisms for Fair Classification. AISTATS 2017.

# Preliminary: Statistical Parity

- **Given**
  - $s$: a binary sensitive attribute
  - $\mathcal{D} = \{(\mathbf{x}_i, s_i, y_i) | i = 1, \ldots, n\}$: a dataset of $n$ data points
    - $\mathbf{x}_i, s_i, y_i$: feature vector, sensitive attribute value and a binary label of the $i$-th data point
- **Definition:** the predicted labels $\tilde{\mathcal{Y}} = \{\tilde{y}_i | i = 1, \ldots, n\}$ satisfies statistical parity iff.

$$\underbrace{\Pr(\tilde{y} = 1 | s = 0) = \Pr(\tilde{y} = 1 | s = 1)}_{\text{Probabilistic perspective}} \Leftrightarrow \underbrace{I(\tilde{y}; s) = 0}_{\text{Information-theoretic perspective}}$$

- **Example:** loan approval




$\Pr(\tilde{y} = \text{accepted} \mid \quad) = 2/3$

$\Pr(\tilde{y} = \text{accepted} \mid \quad) = 2/3$

**Statistical parity satisfied**

Same acceptance rate for male and female

[1] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S.. Certifying and Removing Disparate Impact. KDD 2015.

# Problem Definition

- **Input**
  - $\mathcal{S} = \{s^{(1)}, \dots, s^{(k)}\}$: a set of $k$ sensitive attributes
    - $s^{(j)}$: $j$-th sensitive attribute
  - $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{s}_i, y_i) | i = 1, \dots, n\}$: a set of $n$ data points
    - $\mathbf{s}_i = [s_i^{(1)}, \dots, s_i^{(k)}]$: the vectorized sensitive feature of the $i$-th data point that includes all interested sensitive attribute
  - $l(\mathbf{x}; \mathbf{s}; y; \tilde{\mathbf{y}}; \boldsymbol{\theta})$: a loss function to be minimized by a learning algorithm
    - $\tilde{\mathbf{y}}^* = \mathrm{argmin}_{\tilde{\mathbf{y}}} l(\mathbf{x}; \mathbf{s}; y; \tilde{\mathbf{y}}; \boldsymbol{\theta})$: the optimal learning outcome w.r.t. the input data
- **Output:** a set of revised learning outcomes $\{\tilde{\mathbf{y}}_i{}^* | i = 1, \dots, n\}$ that minimizes
  - Empirical loss $\mathbb{E}_{(\mathbf{x}, \mathbf{s}, y) \sim \mathcal{D}}[l(\mathbf{x}; \mathbf{s}; y; \tilde{\mathbf{y}}; \boldsymbol{\theta})]$
  - Mutual information between the learning outcomes and sensitive attribute $I(\tilde{\mathbf{y}}; \mathbf{s})$

# Roadmap

- Motivation ☑
- Proposed method: InfoFair
- Experiments
- Conclusion

# Problem Formulation

- **Optimization problem**

$$\min_{\boldsymbol{\theta}} \quad J = \mathbb{E}_{(\mathbf{x},s,y)\sim\mathcal{D}}[l(\mathbf{x};\mathbf{s};y;\tilde{\mathbf{y}};\boldsymbol{\theta}) + \alpha \boxed{I(\tilde{\mathbf{y}};\mathbf{s})}]$$

Key term to optimize

  - $\alpha$: regularization hyperparameter, non-negative

- **Common approach:** adversarial learning
  - **Key idea:** predicting one random variable (e.g., $\mathbf{s}$) using another one (e.g., $\tilde{\mathbf{y}}$)
  - **Limitation:** requiring perfect modeling of distribution between two variables

$$p(\mathbf{s}|\tilde{\mathbf{y}}) = q(\mathbf{s}|\tilde{\mathbf{y}})$$

    - $p(\mathbf{s}|\tilde{\mathbf{y}})$, $q(\mathbf{s}|\tilde{\mathbf{y}})$: probability density functions of $\mathbf{s}$ given $\tilde{\mathbf{y}}$
    - $q(\mathbf{s}|\tilde{\mathbf{y}})$ is modeled by an adversary with some learnable parameters

- **Question:** how to minimize mutual information when $p(\mathbf{s}|\tilde{\mathbf{y}}) = q(\mathbf{s}|\tilde{\mathbf{y}})$ does not hold?

# Mutual Information: A Variational Representation

- **Mutual information**

$$I(\tilde{\mathbf{y}};\mathbf{s}) = H(\mathbf{s}) - H(\mathbf{s}|\tilde{\mathbf{y}})$$

  - $H(\mathbf{s}) = -\mathbb{E}_{\mathbf{s}}[\log p(\mathbf{s})]$: entropy of $\mathbf{s}$
  - $H(\mathbf{s}|\tilde{\mathbf{y}}) = -\mathbb{E}_{\mathbf{s},\tilde{\mathbf{y}}}[\log p(\mathbf{s}|\tilde{\mathbf{y}})]$: conditional entropy of $\mathbf{s}$ given $\tilde{\mathbf{y}}$

- **A variational representation**

$$I(\tilde{\mathbf{y}};\mathbf{s}) = H(\mathbf{s}) + \mathbb{E}_{\mathbf{s},\tilde{\mathbf{y}}}\Big[\underbrace{\boxed{\log q(\mathbf{s}|\tilde{\mathbf{y}})}}_{\text{Key term \#1}}\Big] + \mathbb{E}_{\mathbf{s},\tilde{\mathbf{y}}}\left[\underbrace{\boxed{\log \frac{p(\tilde{\mathbf{y}};\mathbf{s})}{p(\tilde{\mathbf{y}})q(\mathbf{s}|\tilde{\mathbf{y}})}}}_{\text{Key term \#2}}\right]$$

  - $q(\mathbf{s}|\tilde{\mathbf{y}})$: a variational distribution of $p(\mathbf{s}|\tilde{\mathbf{y}})$
  - $H(\mathbf{s})$: a constant (our assumption), $\mathbf{s}$ relates to demographic information which is commonly unchanged

- **Question:** how to calculate these key terms?

# InfoFair: Sensitive Feature Reconstruction

- **Goal:** practical computation of $\log q(\mathbf{s}|\tilde{\mathbf{y}})$
- **Key idea:** reconstruction of sensitive feature $\mathbf{s}$ given $\tilde{\mathbf{y}}$
- **Solution:** a decoder $f$

$$\log q(\mathbf{s}|\tilde{\mathbf{y}}) = \log f(\tilde{\mathbf{y}}; \mathbf{s}; \mathbf{W})$$

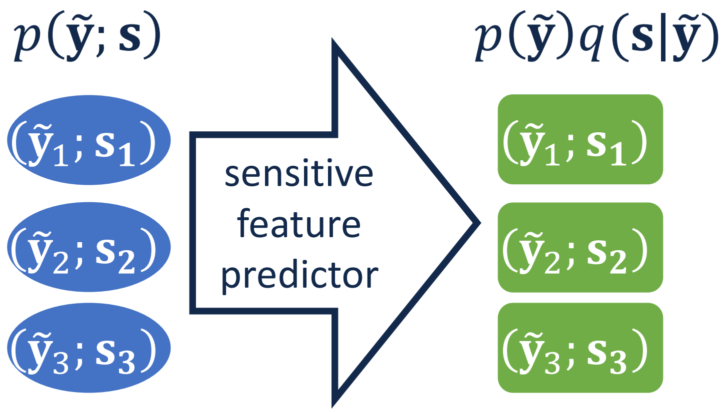
  – **Input:** $\tilde{\mathbf{y}}$ = the learning outcome of a data point, $\mathbf{s}$ = the sensitive feature of a data point, $\mathbf{W}$ = learnable parameters
  – **Output:** $f(\tilde{\mathbf{y}}; \mathbf{s}; \mathbf{W})$ = output of the decoder

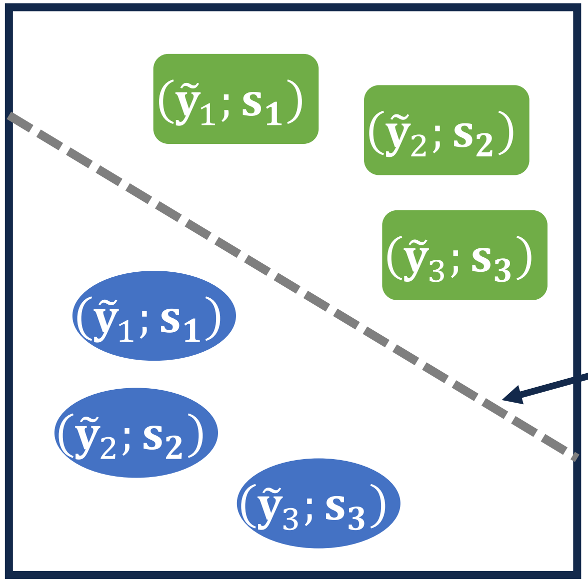
- **Examples of sensitive feature predictor**
  – **Categorical sensitive feature $\mathbf{s}$:** $f(\tilde{\mathbf{y}}; \mathbf{s}; \mathbf{W})$ = log-likelihood $\log \Pr(\mathbf{s}|\tilde{\mathbf{y}})$
  – **Continuous sensitive feature $\mathbf{s}$:** $f(\tilde{\mathbf{y}}; \mathbf{s}; \mathbf{W})$ = output of some probabilistic generative model (e.g., variational autoencoders)

[1] Bose, A., & Hamilton, W.. Compositional Fairness Constraints for Graph Embeddings. ICML 2019.
[2] Zhang, B. H., Lemoine, B., & Mitchell, M.. Mitigating Unwanted Biases with Adversarial Learning. AIES 2018.

# InfoFair: Density Ratio Estimation

- **Goal:** practical computation of $\log \frac{p(\tilde{\mathbf{y}};\mathbf{s})}{p(\tilde{\mathbf{y}})q(\mathbf{s}|\tilde{\mathbf{y}})}$
- **Key idea:** density ratio estimation
- **Solution:** class probability estimation (originally developed for covariate shift)
  - **Intuition:** predict the probability that a pair $(\tilde{\mathbf{y}};\mathbf{s})$ is drawn from the true distribution $p$
- **Example**




Decision boundary of a classifier

- **Goal:** predict how possible a pair $(\tilde{\mathbf{y}};\mathbf{s})$ is $(\tilde{\mathbf{y}};\mathbf{s})$

[1] Bickel, S., Brückner, M., & Scheffer, T.. Discriminative Learning under Covariate Shift. JMLR 2009.

# Density Ratio Estimation: Detailed Steps

- **Key steps**
  - Assign positive label ($c = 1$) for $\tilde{\mathbf{y}}$ and the ground-truth sensitive features
  - Assign negative label ($c = -1$) for $\tilde{\mathbf{y}}$ and its reconstructed sensitive features
  - Apply a classifier to predict $c$ for a given pair of $\tilde{\mathbf{y}}$ and ground-truth/reconstructed sensitive feature

$$p(\tilde{\mathbf{y}}; \mathbf{s}) = \Pr(c = 1 | \tilde{\mathbf{y}}, \mathbf{s}) \qquad p(\tilde{\mathbf{y}}) q(\mathbf{s} | \tilde{\mathbf{y}}) = \Pr(c = -1 | \tilde{\mathbf{y}}, \mathbf{s})$$

  - Calculate the density ratio

$$\log \frac{p(\tilde{\mathbf{y}}; \mathbf{s})}{p(\tilde{\mathbf{y}}) q(\mathbf{s} | \tilde{\mathbf{y}})} = \log \frac{\Pr(c = 1 | \tilde{\mathbf{y}}, \mathbf{s})}{1 - \Pr(c = 1 | \tilde{\mathbf{y}}, \mathbf{s})} = \operatorname{logit}(\Pr(c = 1 | \tilde{\mathbf{y}}, \mathbf{s}))$$

- **Classifier = logistic regression classifier**

$$\log \frac{p(\tilde{\mathbf{y}}; \mathbf{s})}{p(\tilde{\mathbf{y}}) q(\mathbf{s} | \tilde{\mathbf{y}})} = \operatorname{logit}(\Pr(c = 1 | \tilde{\mathbf{y}}, \mathbf{s})) = \mathbf{w}_1^T \tilde{\mathbf{y}} + \mathbf{w}_2^T \mathbf{s}$$

  - $\mathbf{w}_1$: learnable parameters corresponding to $\tilde{\mathbf{y}}$
  - $\mathbf{w}_2$: learnable parameters corresponding to $\mathbf{s}$

# InfoFair: Optimization Problem

- **Practical computation of the variational representation**
  - Sensitive attribute reconstruction with decoder
  - Density ratio estimation as class probability estimation
- **Optimization problem**

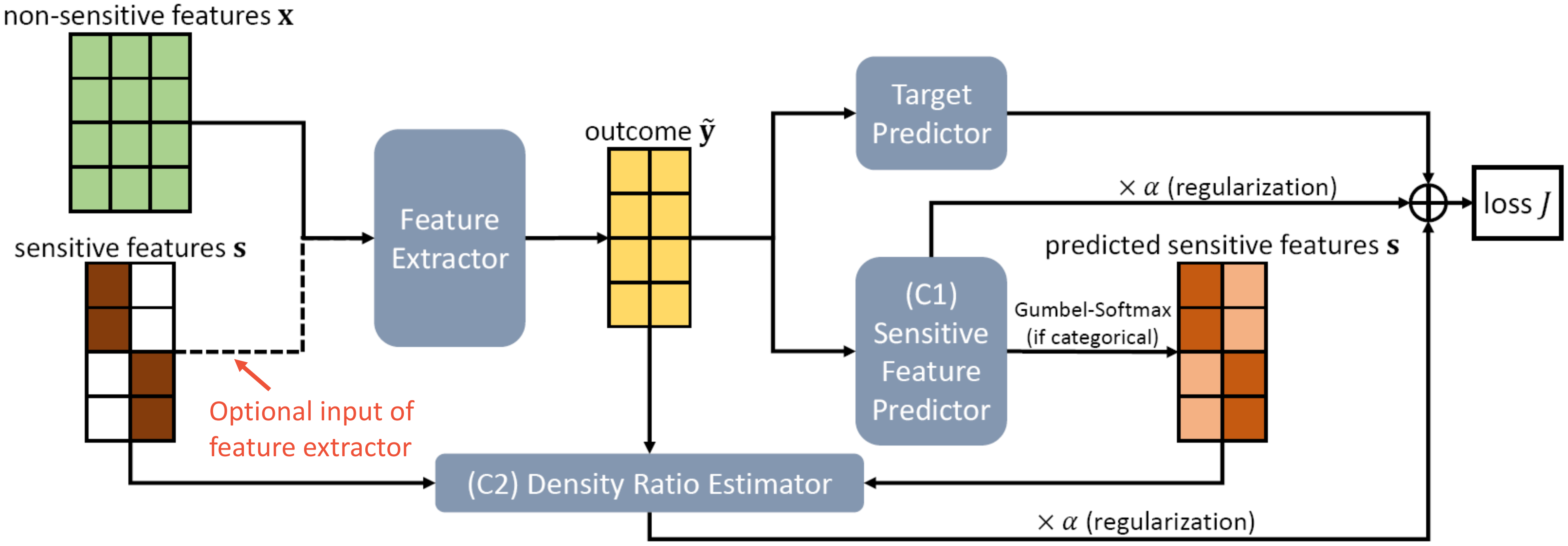
Sensitive attribute reconstruction

$$\min_{\boldsymbol{\theta},\mathbf{w}_1,\mathbf{w}_2} \quad J = \mathbb{E}_{(\mathbf{x},s,y)\sim\mathcal{D}}\left[l(\mathbf{x};\mathbf{s};y;\tilde{\mathbf{y}};\boldsymbol{\theta}) + \alpha \boxed{\log q(\mathbf{s}|\tilde{\mathbf{y}})}\right] + \boxed{\mathbb{E}_{\{(\tilde{\mathbf{y}},\mathbf{s})\sim p(\tilde{\mathbf{y}},\mathbf{s})\}\cup\{(\tilde{\mathbf{y}},\mathbf{s})\sim p(\tilde{\mathbf{y}})q(\mathbf{s}|\tilde{\mathbf{y}})\}}\left[\mathbf{w}_1^T\tilde{\mathbf{y}} + \mathbf{w}_2^T\mathbf{s}\right]}$$

Density ratio estimation

# InfoFair: Overall Framework

- **Key components**
  - **Feature extractor + target predictor:** predict target for downstream tasks
  - **Sensitive feature predictor:** reconstruct sensitive feature
  - **Density ratio estimator:** calculate the density ratio



# InfoFair: Generalizations and Variants

- **InfoFair with equal opportunity**
  - **Solution:** calculate the variational representation of mutual information for samples with specific label only
- **Relationship to adversarial debiasing**
  - **Solution:** (1) merge feature extractor and target predictor to one module and (2) remove the density ratio estimator
- **Relationship to information bottleneck**
  - **Solution:** set the loss function to be the negative mutual information between ground truth and learning outcomes
- **Fairness for continuous-valued sensitive attributes**
  - **Solution:** utilize a probabilistic generative model to reconstruct sensitive feature
- **Fairness for non-i.i.d. graph data**
  - **Solution:** change the feature extractor to a graph neural network

[1] Hardt, M., Price, E., & Srebro, N.. Equality of opportunity in supervised learning. NeurIPS 2016.
[2] Zhang, B. H., Lemoine, B., & Mitchell, M.. Mitigating Unwanted Biases with Adversarial Learning. AIES 2018.
[3] Tishby, N., Pereira, F. C., & Bialek, W.. The Information Bottleneck Method. arXiv 2000.
[4] Kipf, T. N., & Welling, M.. Semi-supervised Classification with Graph Convolutional Networks. ICLR 2017.

# Roadmap

- Motivation ✓
- Proposed method: InfoFair ✓
- Experiments
- Conclusion

# Experiments: Settings

- **Task:** binary classification
- **Sensitive attribute:** binary attribute, non-binary attribute, multiple attributes
- **Benchmark datasets**

| Datasets | # Samples | # Attributes | # Classes |
|---|---|---|---|
| COMPAS | 6,172 | 52 | 2 |
| Adult Income | 45,222 | 14 | 2 |
| Dutch Census | 60,420 | 11 | 2 |

- **Baseline methods**
  - **Vanilla model:** Vanilla
  - **Fairness-aware models:** LFR, MinDiff, DI, Adversarial, FCFC, GerryFair, GDP
- **Metrics**
  - **Utility:** micro F1 and macro F1 (Micro/Macro F1)
  - **Fairness:** statistical imparity (Imparity) and relative reduction (Reduction)

# Experiments: Effectiveness Results

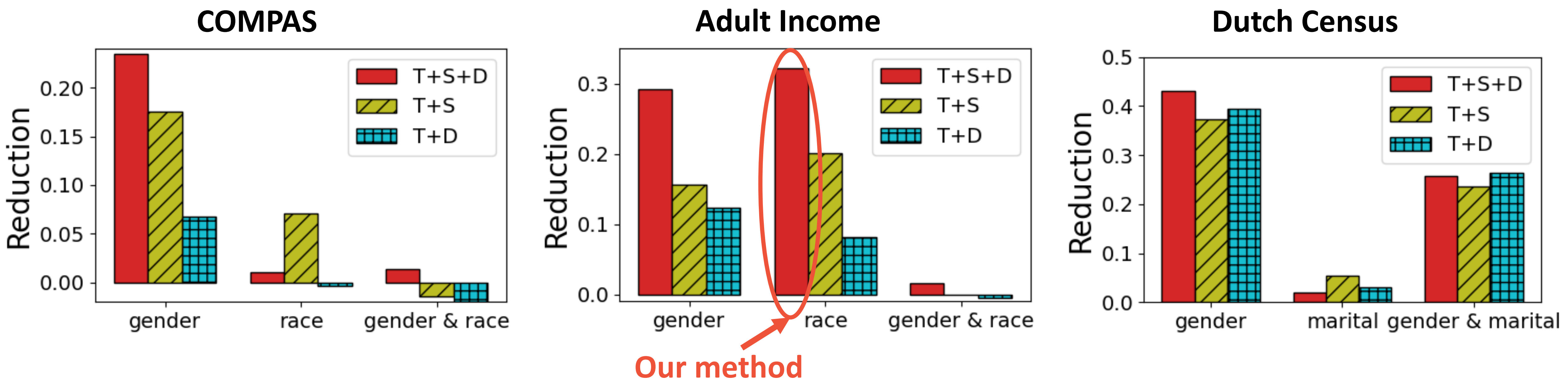
- **Observation:** InfoFair (red box) consistently mitigates the most bias while maintaining accuracy
  - Mitigating more bias = lower imparity, higher reduction
  - LFR, Adversarial and FCFC achieves 100% bias reduction by predicting all data points to one class
  - Similar observation on COMPAS and Dutch Census dataset

Debiasing results on *Adult Income* dataset

| Method | gender | | | race | | | gender & race | | |
|---|---|---|---|---|---|---|---|---|---|
| | Micro/Macro F1 | Imparity | Reduction | Micro/Macro F1 | Imparity | Reduction | Micro/Macro F1 | Imparity | Reduction |
| Vanilla | 0.830/0.762 | 0.066 | 0.000% | 0.830/0.762 | 0.062 | 0.000% | 0.830/0.762 | 0.083 | 0.000% |
| LFR | 0.743/0.426 | 0.000 | 100.0% | N/A | N/A | N/A | N/A | N/A | N/A |
| MinDiff | 0.828/0.746 | 0.058 | 12.06% | N/A | N/A | N/A | N/A | N/A | N/A |
| DI | 0.823/0.730 | 0.053 | 19.85% | 0.825/0.743 | 0.056 | 10.62% | 0.823/0.736 | 0.081 | 2.276% |
| Adversarial | 0.743/0.426 | 0.000 | 100.0% | 0.743/0.426 | 0.000 | 100.0% | 0.743/0.426 | 0.000 | 100.0% |
| FCFC | 0.257/0.204 | 0.000 | 100.0% | 0.257/0.204 | 0.000 | 100.0% | 0.257/0.204 | 0.000 | 100.0% |
| GerryFair | 0.833/0.752 | 0.056 | 15.70% | 0.833/0.752 | 0.067 | −7.664% | 0.797/0.710 | 0.215 | −158.3% |
| GDP | 0.825/0.744 | 0.055 | 16.73% | 0.827/0.749 | 0.059 | 6.351% | 0.824/0.740 | 0.075 | 9.246% |
| INFOFAIR | 0.816/0.721 | 0.047 | 29.24% | 0.810/0.686 | 0.042 | 32.11% | 0.818/0.714 | 0.082 | 1.532% |

# Experiments: Ablation Study

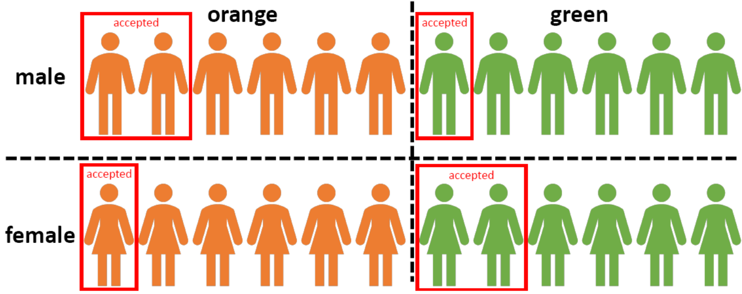
- **Observation:** InfoFair (red bar) mitigates the most bias compared to its ablated variants



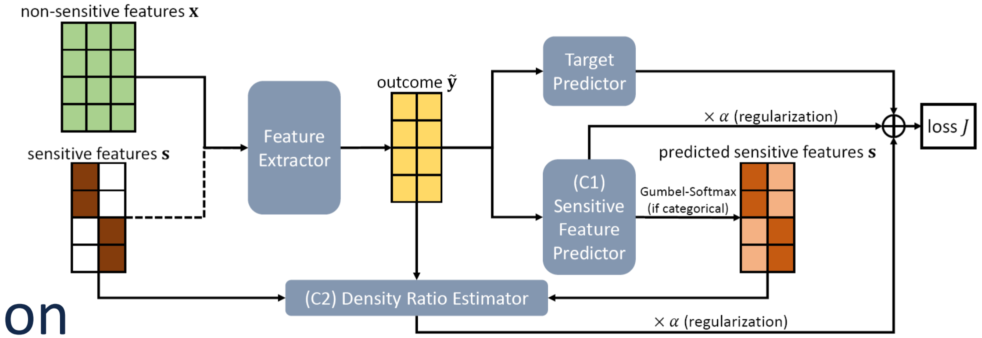
# Roadmap

- Motivation ✓
- Proposed method: InfoFair ✓
- Experiments ✓
- Conclusion

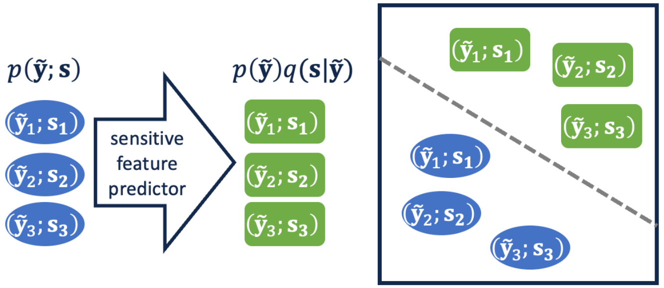
# Takeaways

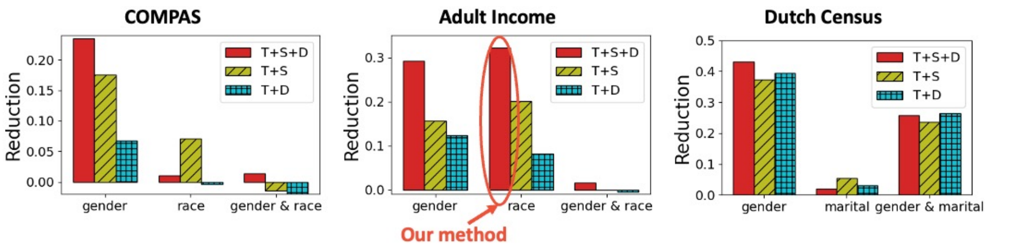



- **Problem:** information-theoretic intersectional fairness
  - **Intersectional fairness:** joint variable of all interested sensitive attribute
  - **Information-theoretic perspective:** mutual information minimization
- **Solution:** InfoFair
  - Variational representation of mutual information
  - Sensitive attribute reconstruction with autoencoder
  - Density ratio estimation as class probability estimation




- **Results:** effectiveness in bias mitigation while maintaining accuracy
- More details in the paper
  - Mathematical analysis
  - Detailed experiments





**Title:** InfoFair: Information-Theoretic Intersectional Fairness
**Authors:** Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, Hanghang Tong
**Website:** http://jiank2.web.illinois.edu/
**Email:** jiank2@illinois.edu