Conditional Learning of Fair Representations

Han Zhao
han.zhao@cs.cmu.edu

Machine Learning Department
Carnegie Mellon University

Joint work with A. Coston, T. Adel and G. Gordon

Carnegie Mellon University Microsoft Research UNIVERSITY OF CAMBRIDGE

Potential Bias of Data in High-stakes Domains

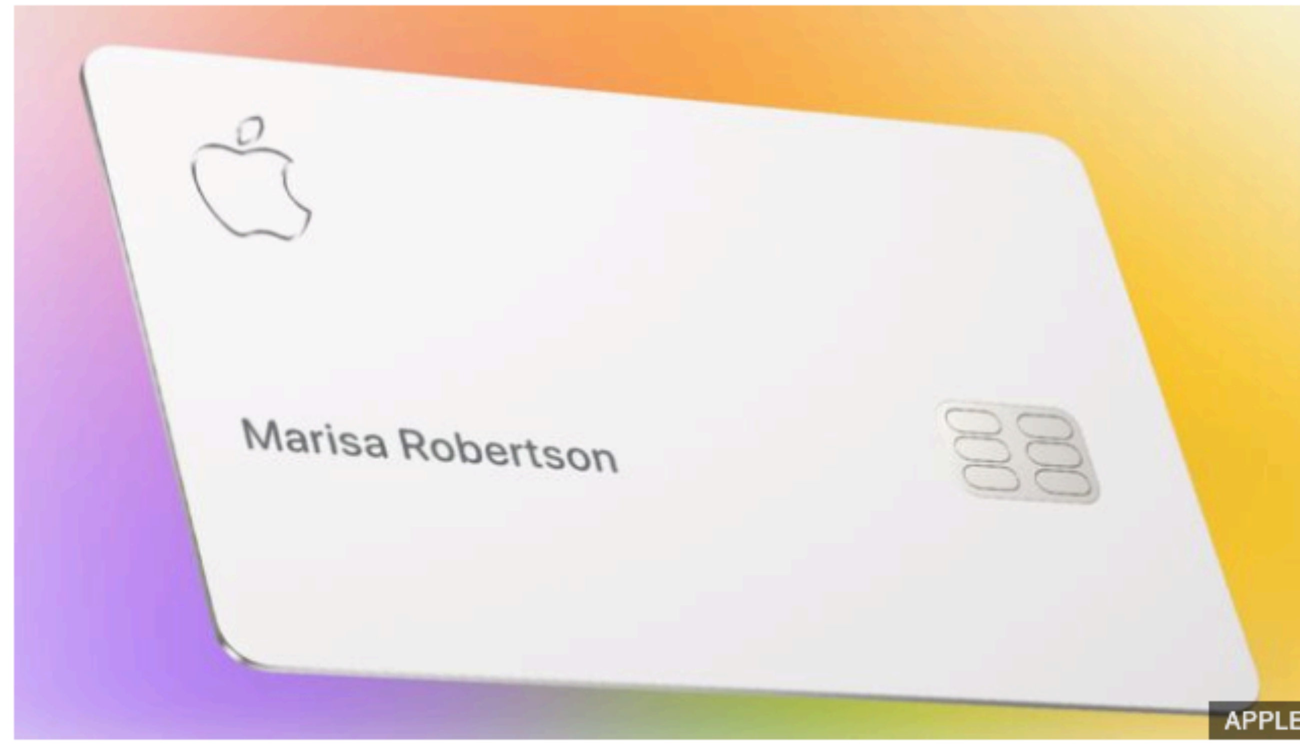
Potential Bias of Data in High-stakes Domains

Potential Bias of Data in High-stakes Domains

Apple's 'sexist' credit card investigated by US regulator

11 November 2019

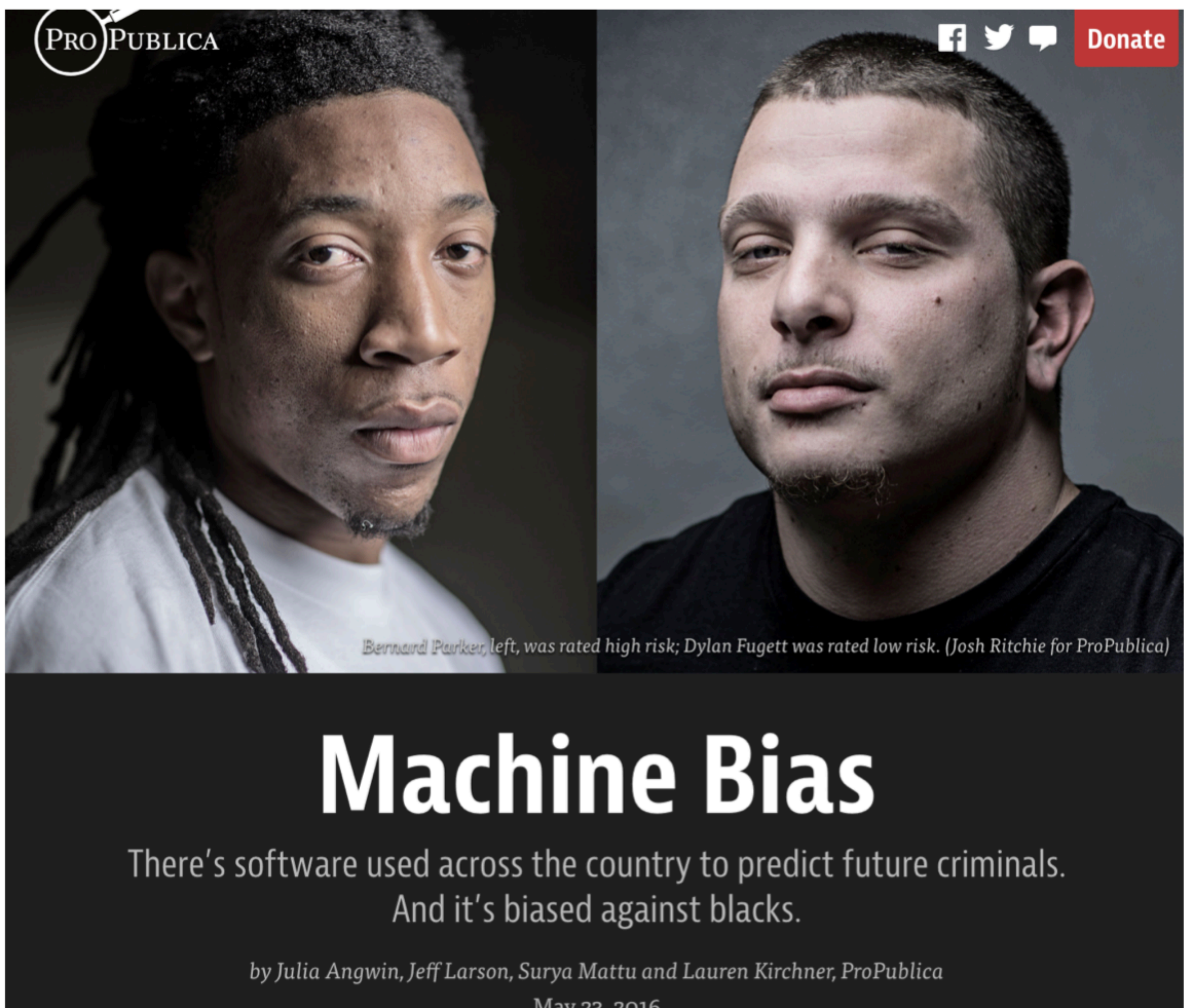
Share



APPLE

A US financial regulator has opened an investigation into claims Apple's credit card offered different credit limits for men and women.

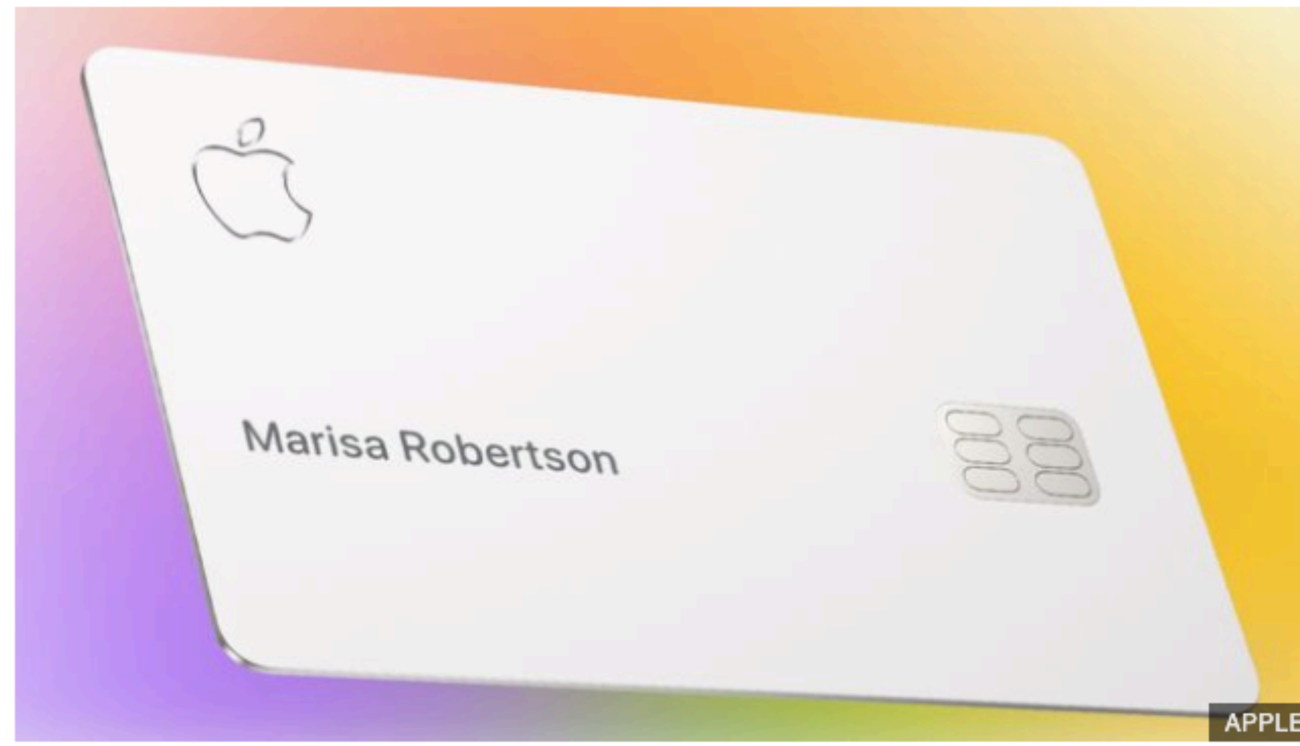
Potential Bias of Data in High-stakes Domains



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Apple's 'sexist' credit card investigated by US regulator

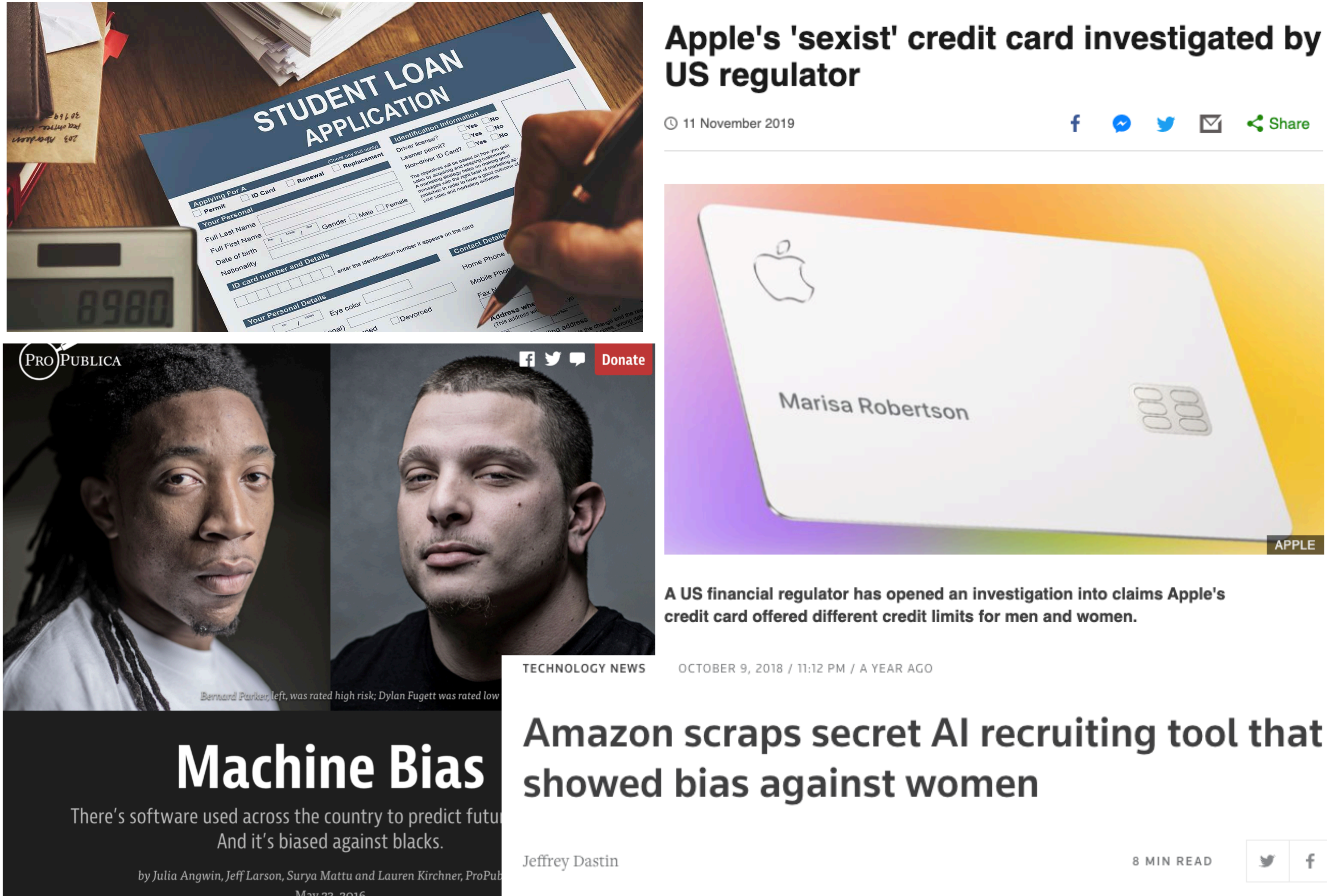
11 November 2019 · Share



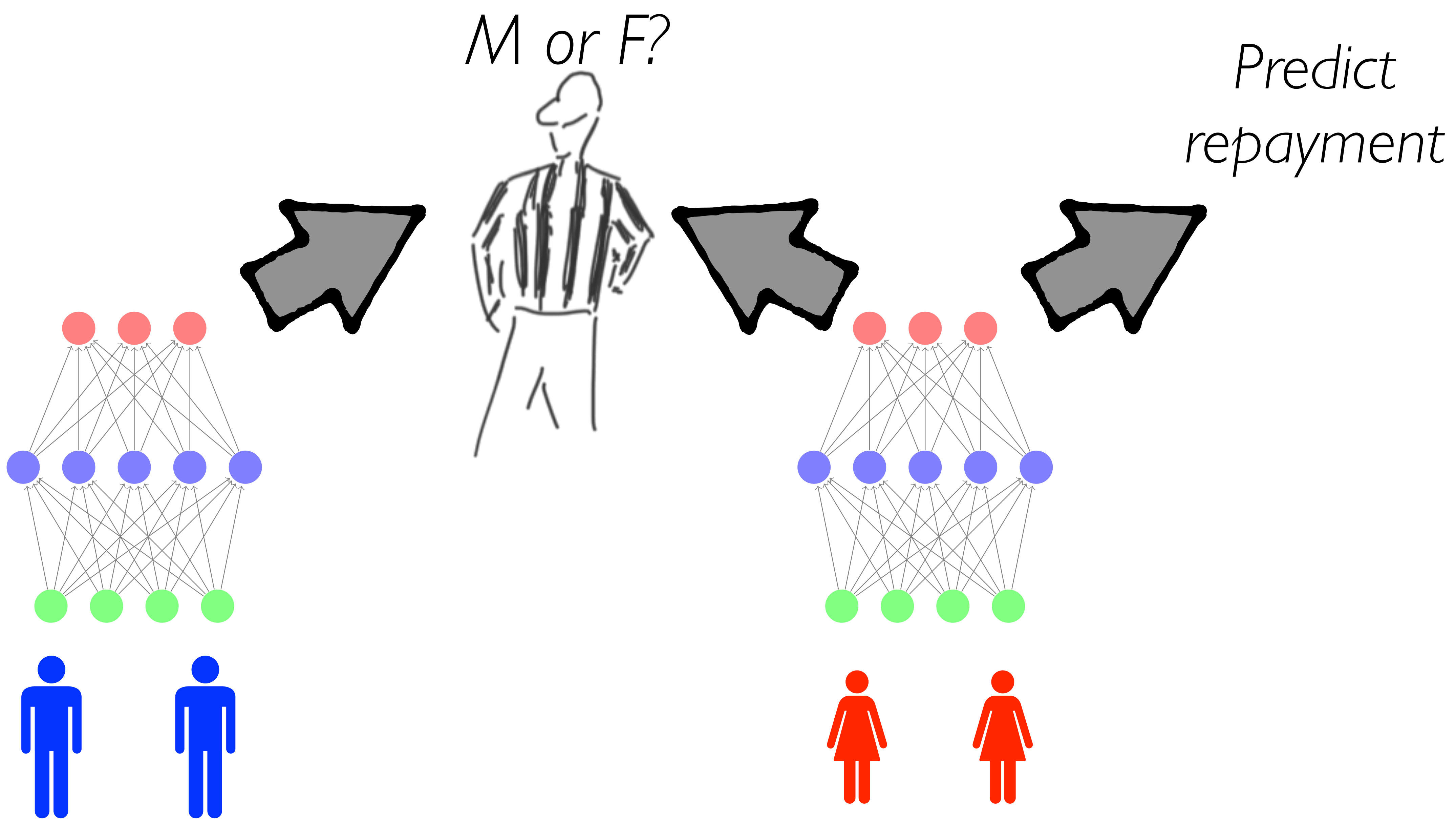
APPLE

A US financial regulator has opened an investigation into claims Apple's credit card offered different credit limits for men and women.

Potential Bias of Data in High-stakes Domains

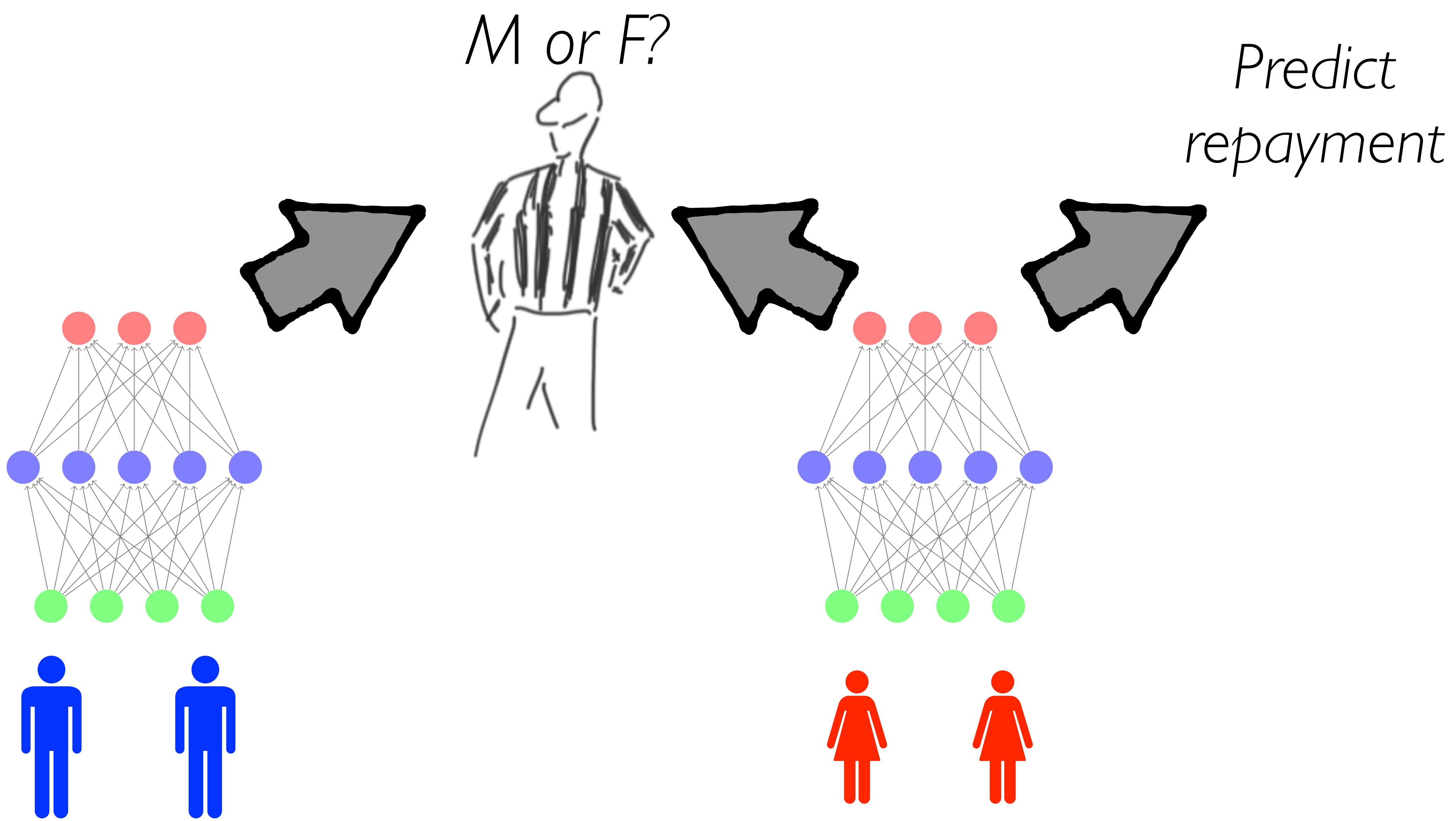


This Work



This Work

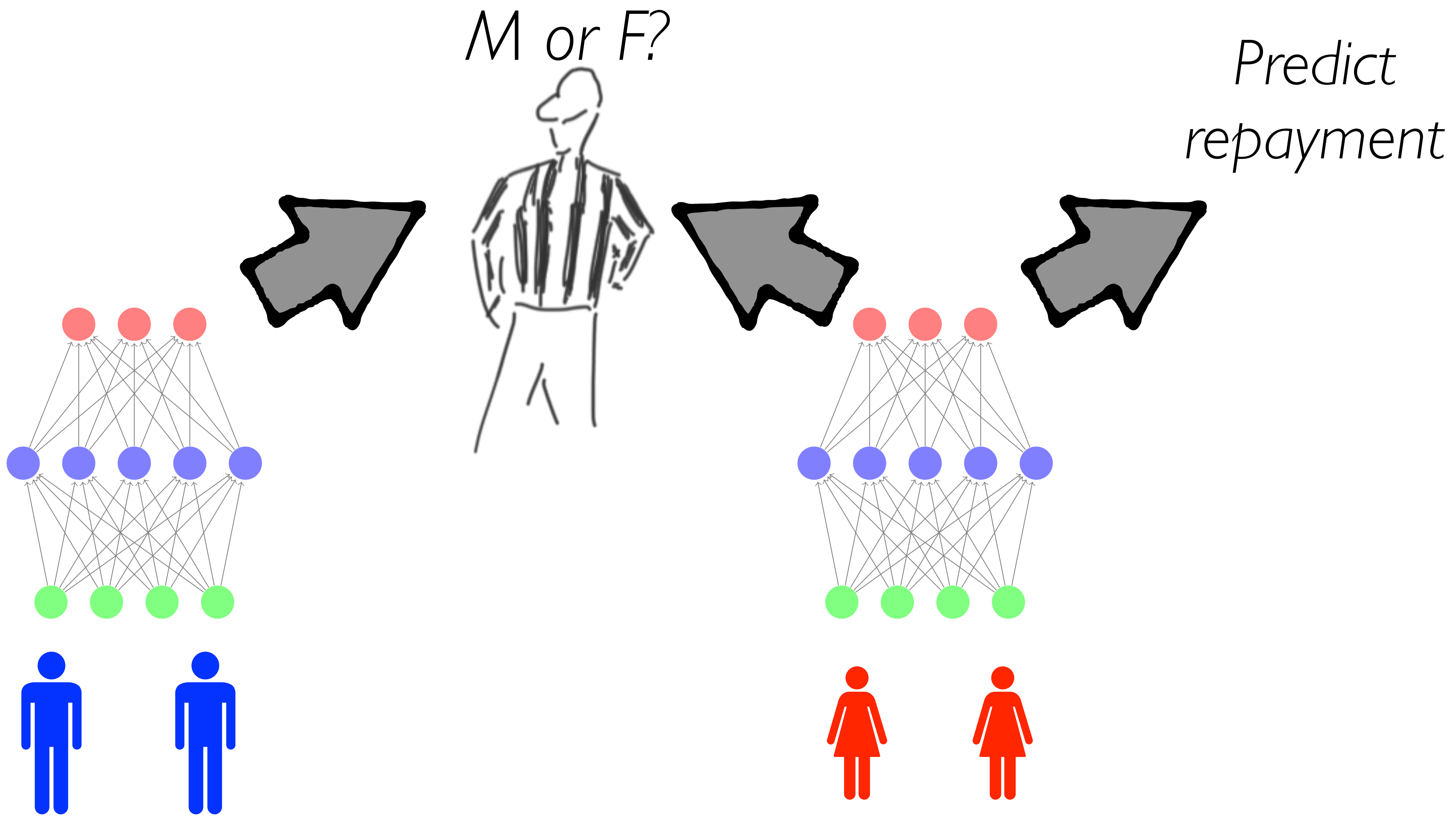
From a representation learning perspective, design algorithmic intervention to



This Work

From a representation learning perspective, design algorithmic intervention to

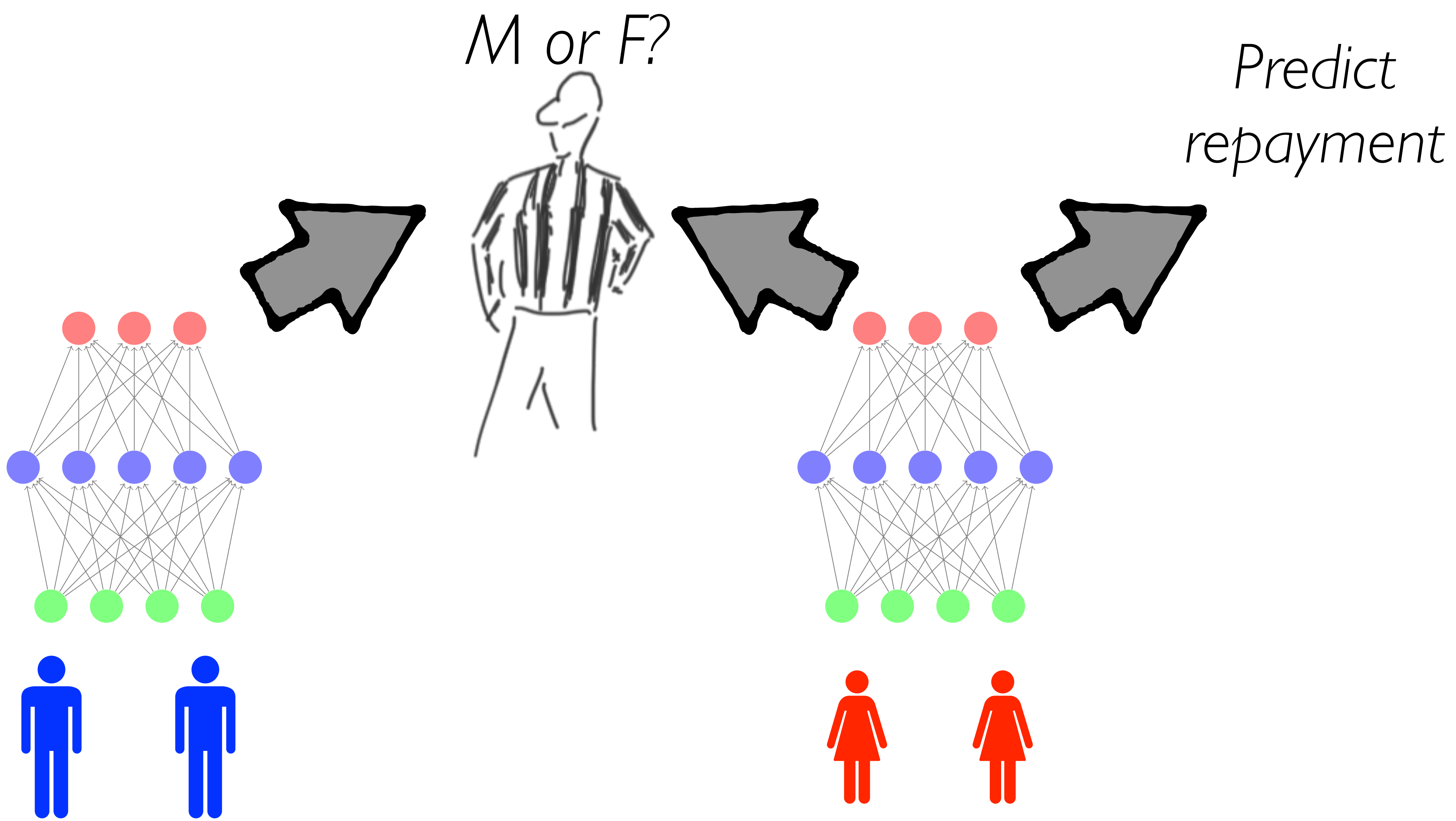
- Seek for equalized odds and accuracy parity simultaneously



This Work

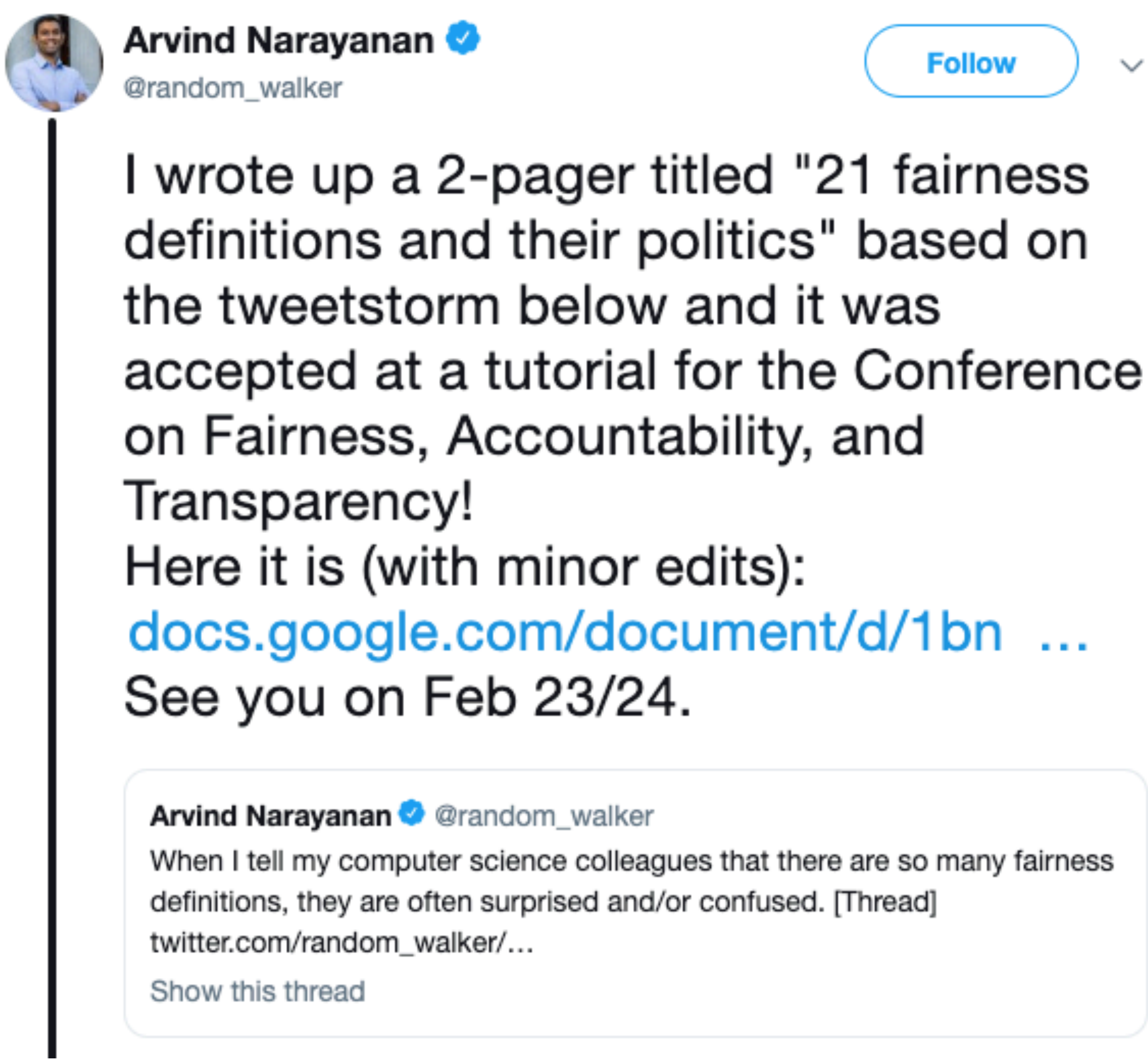
From a representation learning perspective, design algorithmic intervention to

- Seek for equalized odds and accuracy parity simultaneously
- Not harm the existing statistical parity gap



Statistical Definition of Fairness

But, what's fairness in an algorithmic context?

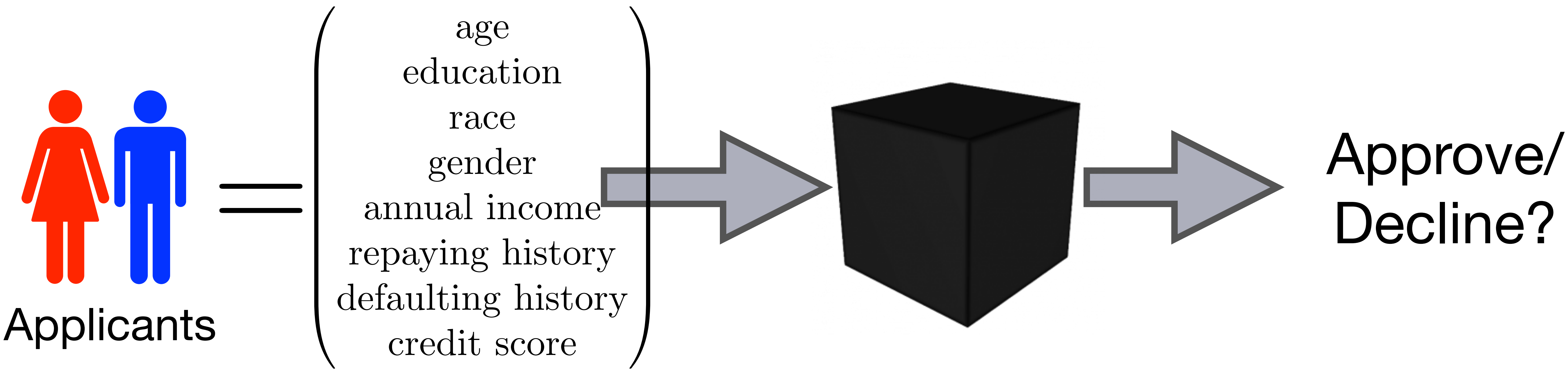


Definition	Paper	Citation #
Group fairness or statistical parity	[12]	208
Conditional statistical parity	[11]	29
Predictive parity	[10]	57
False positive error rate balance	[10]	57
False negative error rate balance	[10]	57
Equalised odds	[14]	106
Conditional use accuracy equality	[8]	18
Overall accuracy equality	[8]	18
Treatment equality	[8]	18
Test-fairness or calibration	[10]	57
Well calibration	[16]	81
Balance for positive class	[16]	81
Balance for negative class	[16]	81

[Verma et al. 18]

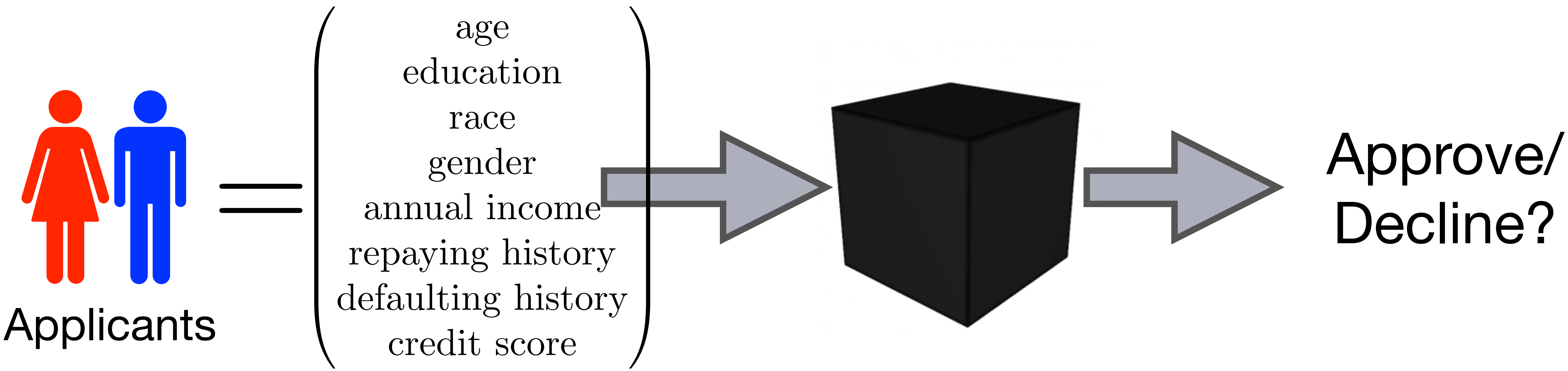
Statistical Definition of Fairness

Example in loan application



Statistical Definition of Fairness

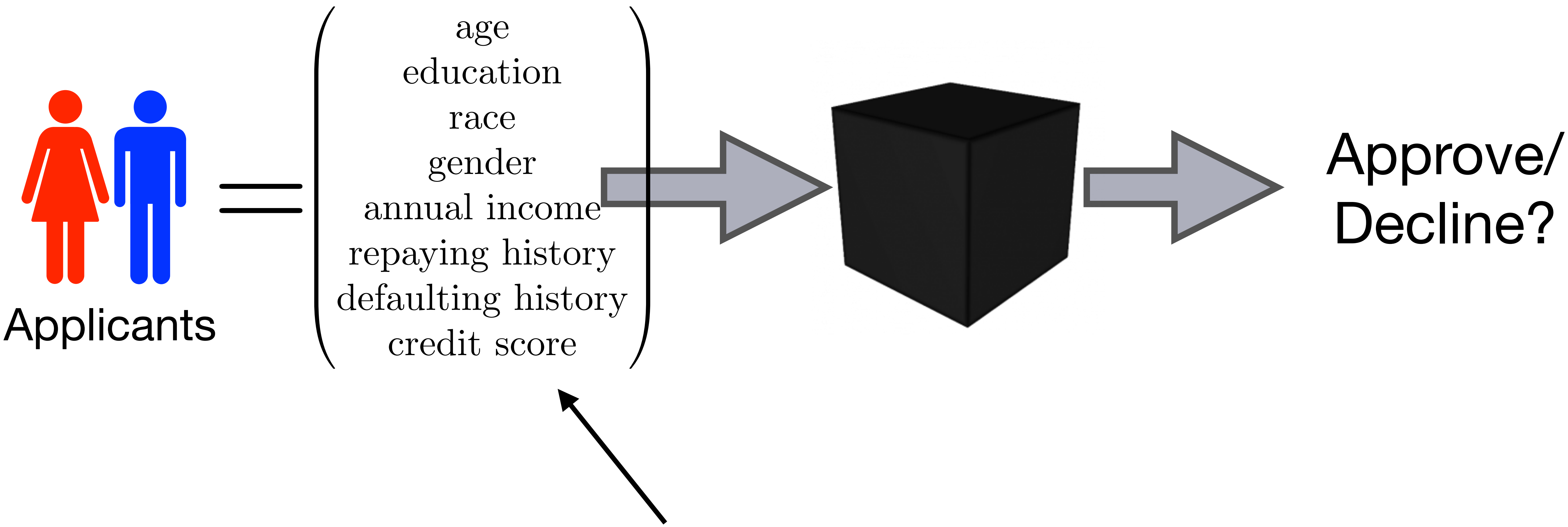
Example in loan application



$$(X, A, Y) \sim \mathcal{D}$$

Statistical Definition of Fairness

Example in loan application

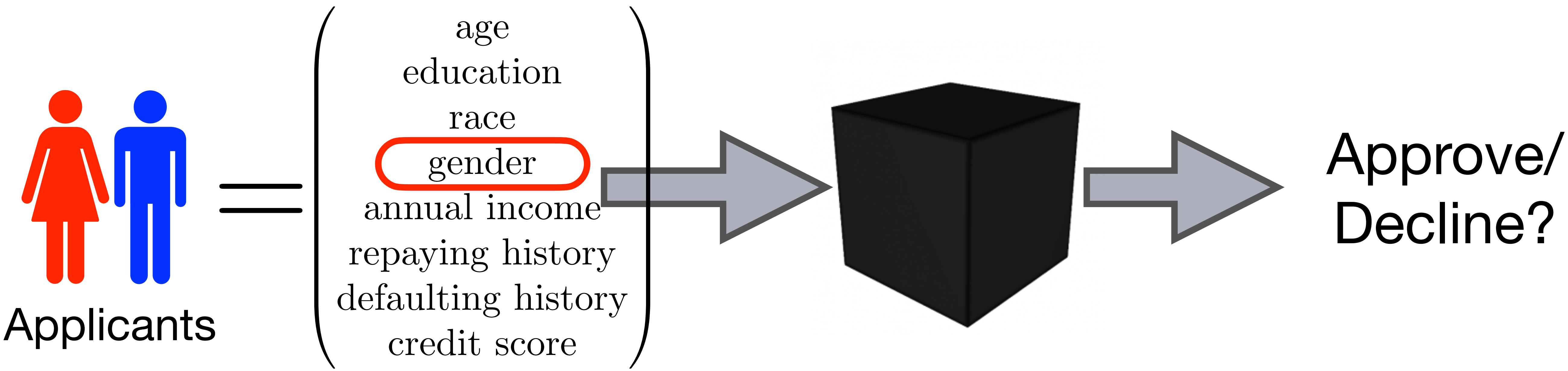


$$(X, A, Y) \sim \mathcal{D}$$

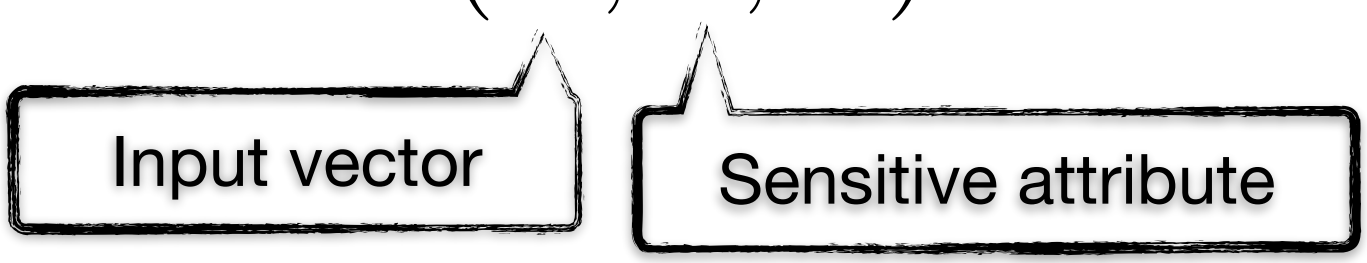
Input vector

Statistical Definition of Fairness

Example in loan application

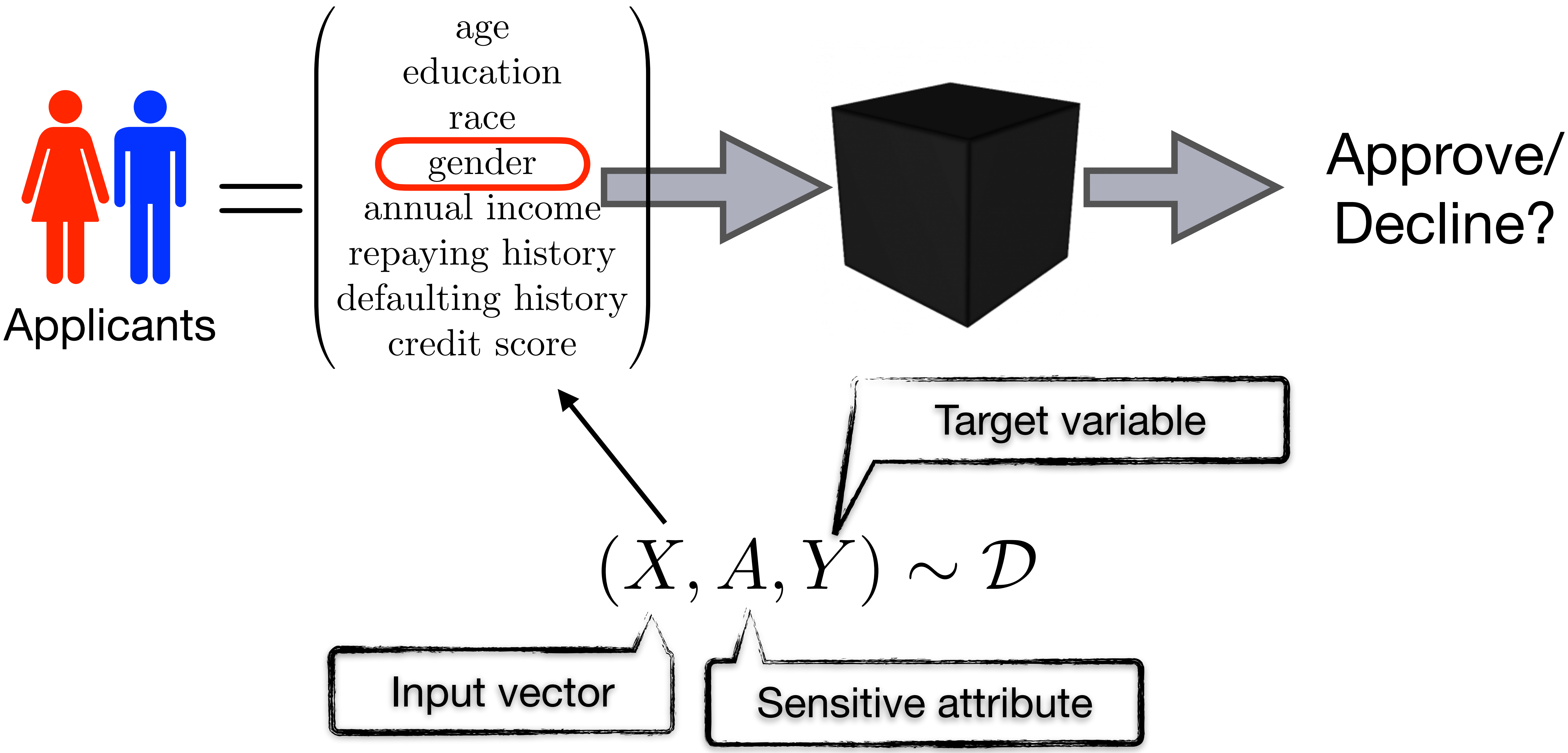


$$(X, A, Y) \sim D$$



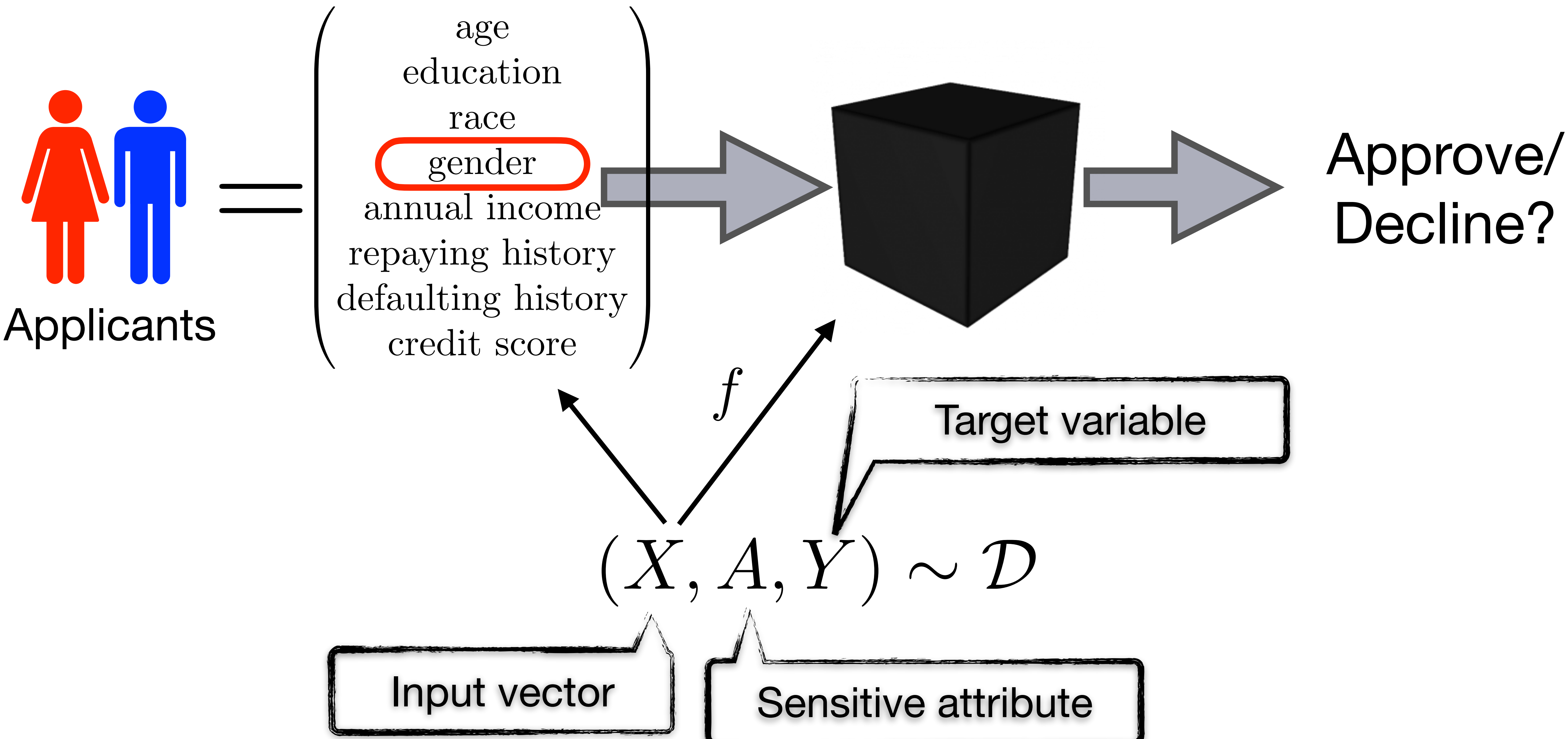
Statistical Definition of Fairness

Example in loan application



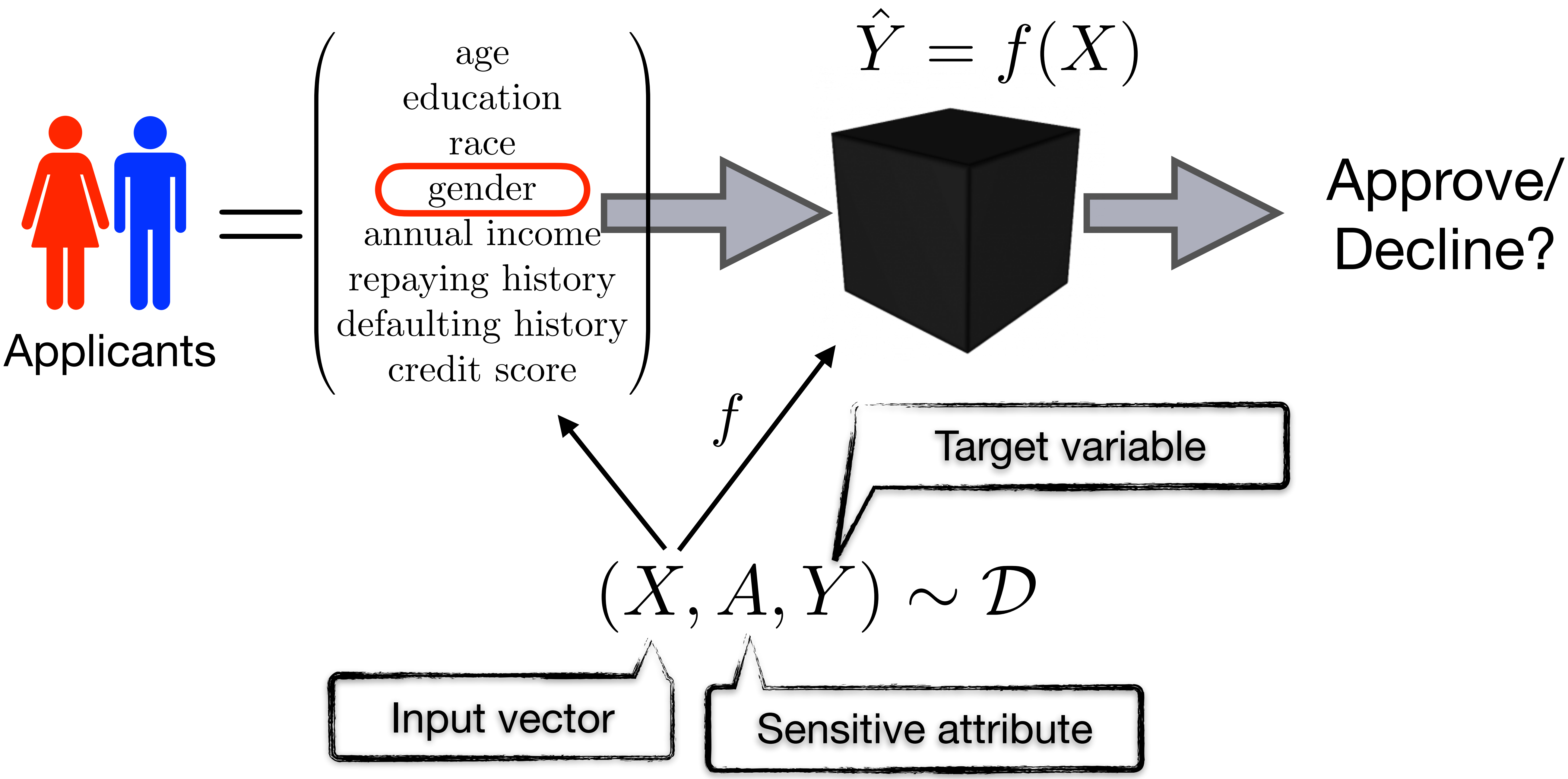
Statistical Definition of Fairness

Example in loan application



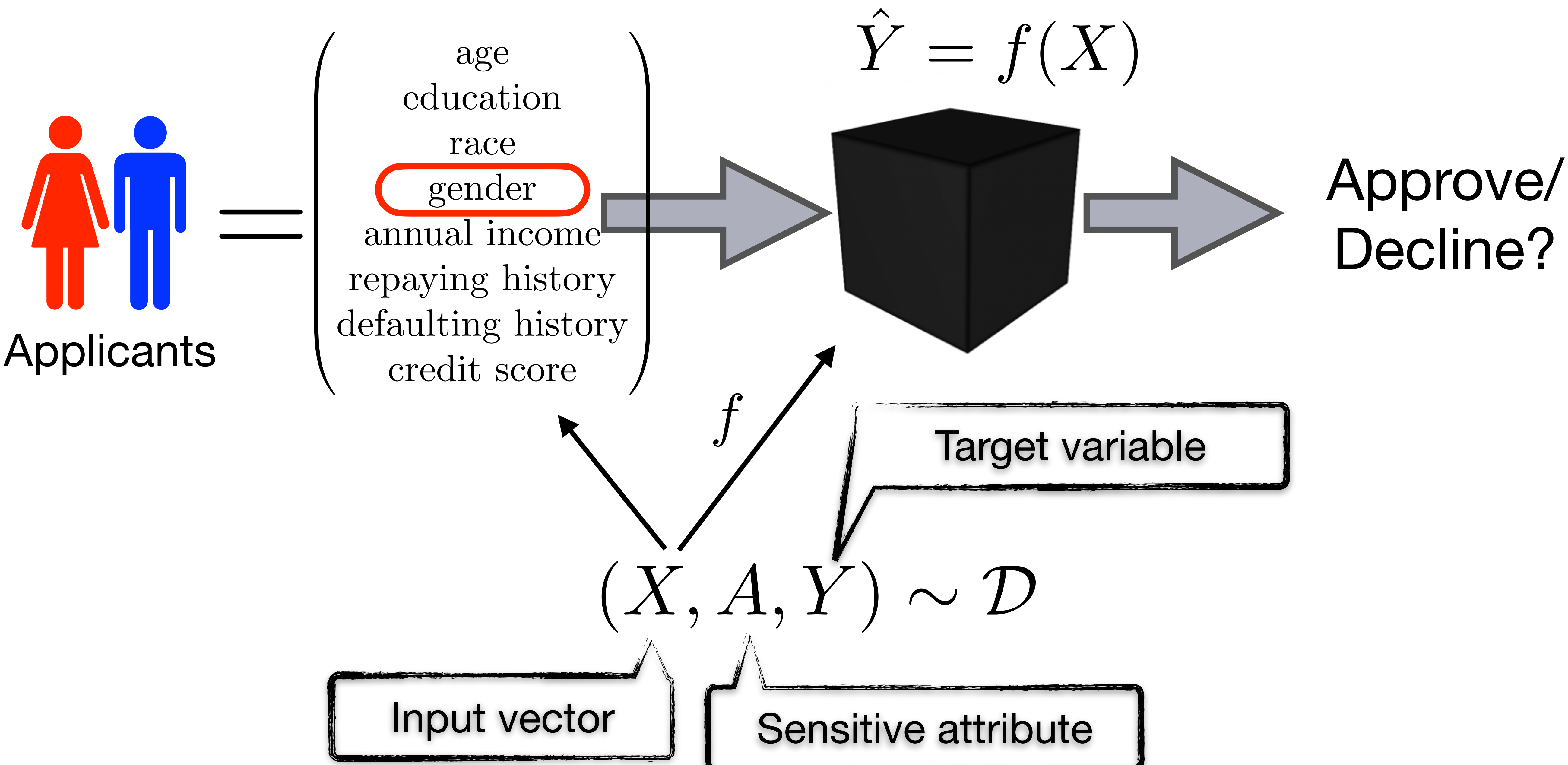
Statistical Definition of Fairness

Example in loan application



Statistical Definition of Fairness

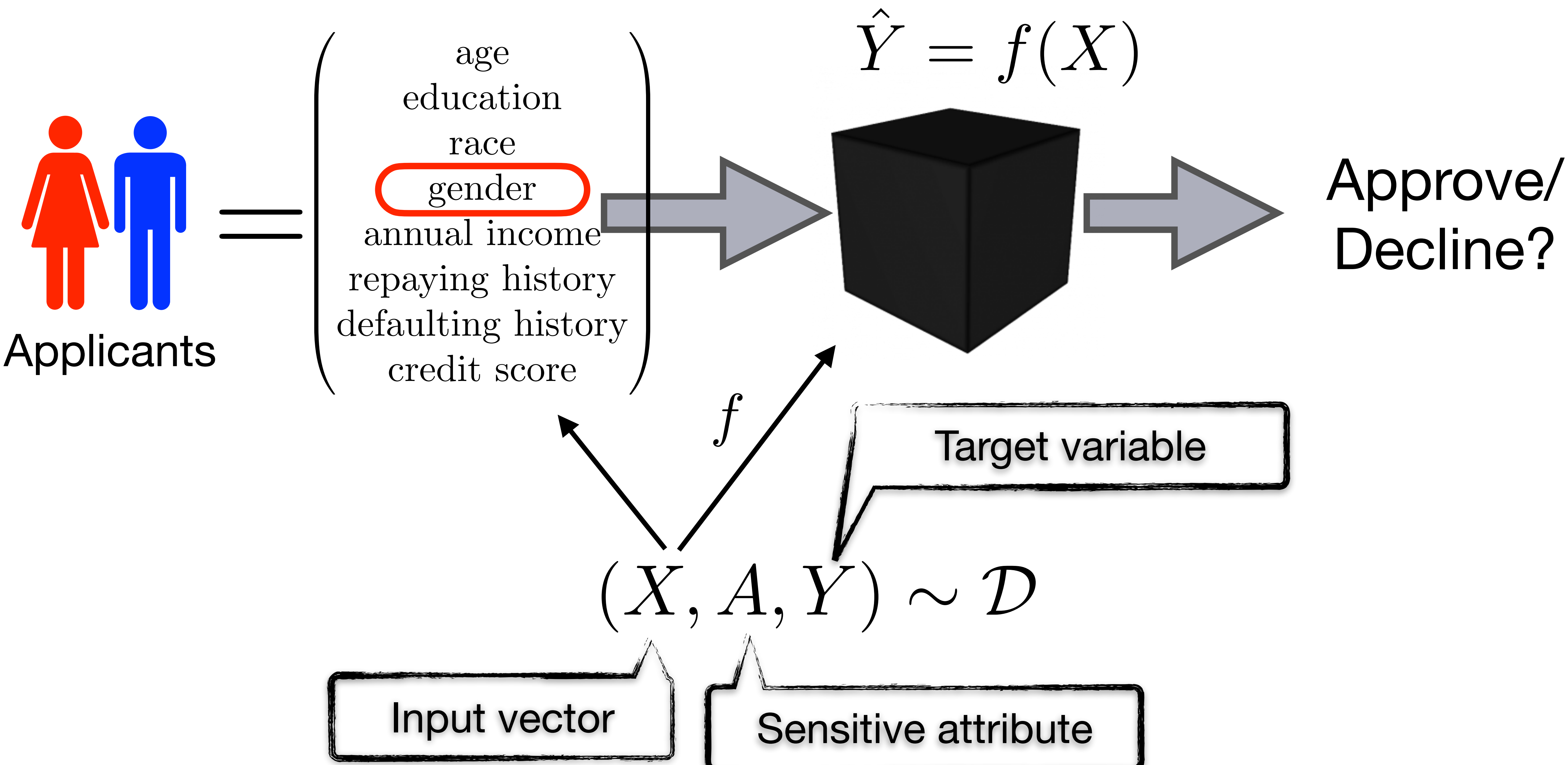
Example in loan application



Statistical parity: $\hat{Y} \perp A$

Statistical Definition of Fairness

Example in loan application

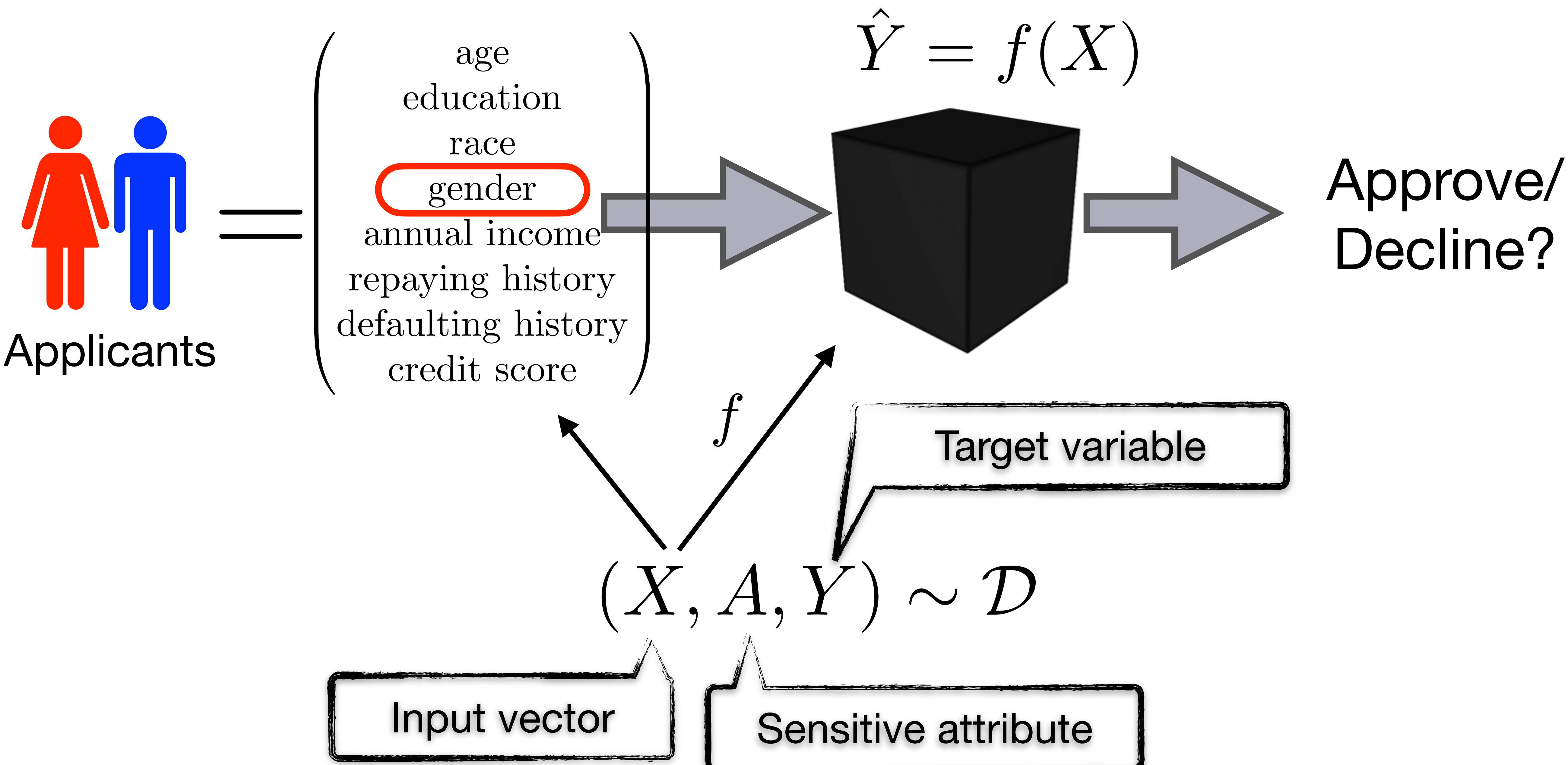


Statistical parity: $\hat{Y} \perp A$

Equalized odds (Hardt et al. 16): $\hat{Y} \perp A \mid Y$

Statistical Definition of Fairness

Example in loan application



Statistical parity: $\hat{Y} \perp A$

Equalized odds (Hardt et al. 16): $\hat{Y} \perp A \mid Y$

Accuracy parity: $\mathrm{err}(\hat{Y}) \perp A$

Incompatibility between Fairness Notions

Statistical Parity

$$\not\sim$$

Equalized Odds

Incompatibility between Fairness Notions

Statistical Parity

$\not\sim$

Equalized Odds

[Chouldechova. Big data 16]
[Kleinberg et al. ITCS 16]
[Hardt et al. NeurIPS 17]

Incompatibility between Fairness Notions

Statistical Parity

$$\not\sim$$

Equalized Odds

[Chouldechova. Big data 16]
[Kleinberg et al. ITCS 16]
[Hardt et al. NeurIPS 17]

Theorem [ZG, NeurIPS 19]:

$$\varepsilon_{A=0}(h) + \varepsilon_{A=1}(h) \geq \Delta_{\mathrm{BR}}$$

Incompatibility between Fairness Notions

Statistical Parity

$\not\sim$

Equalized Odds

Accuracy Parity

[Chouldechova. Big data 16]
[Kleinberg et al. ITCS 16]
[Hardt et al. NeurIPS 17]

Theorem [ZG, NeurIPS 19]:

$$\varepsilon_{A=0}(h) + \varepsilon_{A=1}(h) \geq \Delta_{\mathrm{BR}}$$

Incompatibility between Fairness Notions

Statistical Parity

$\not\sim$

Equalized Odds

?

Accuracy Parity

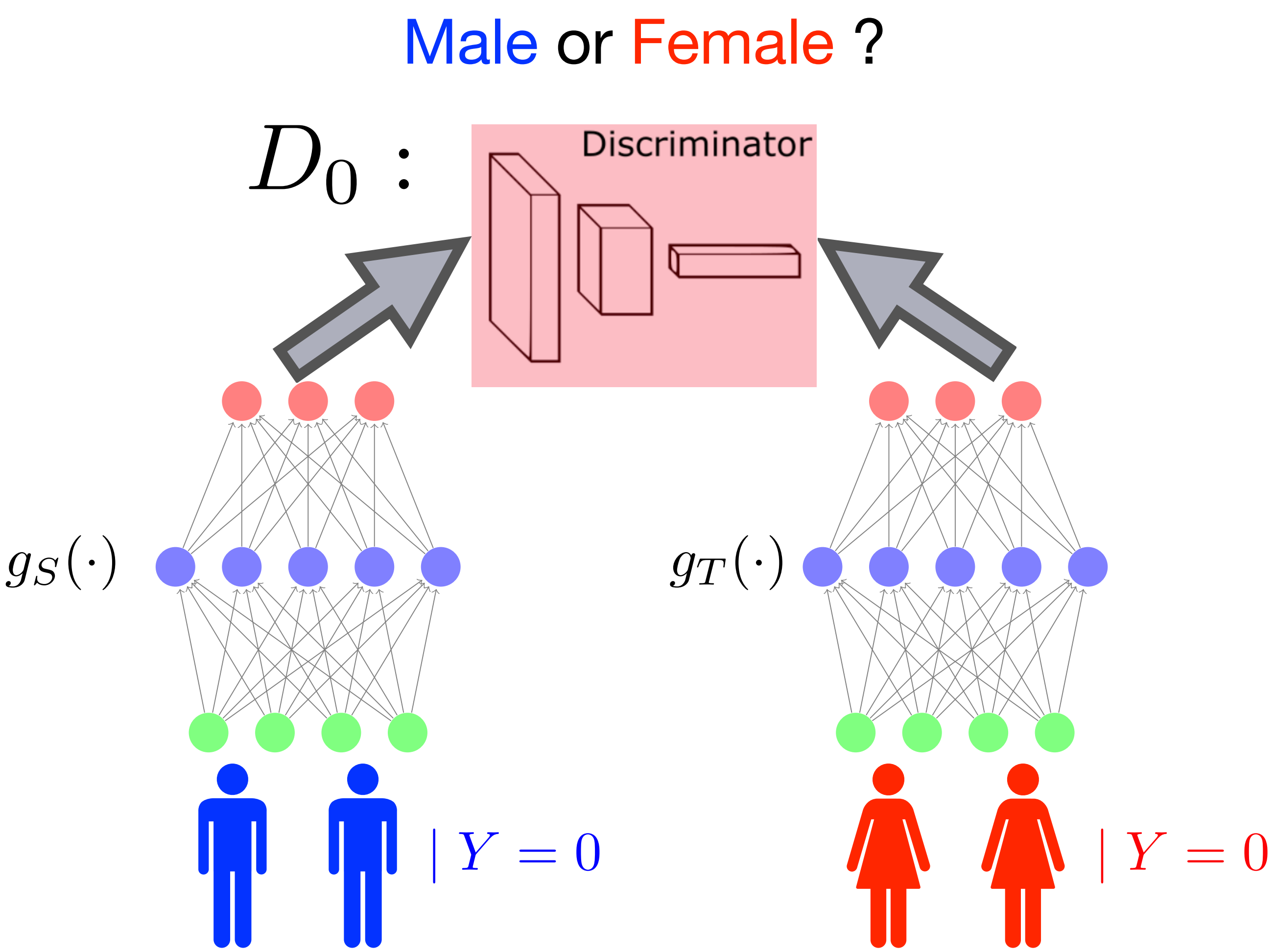
[Chouldechova. Big data 16]
[Kleinberg et al. ITCS 16]
[Hardt et al. NeurIPS 17]

Theorem [ZG, NeurIPS 19]:

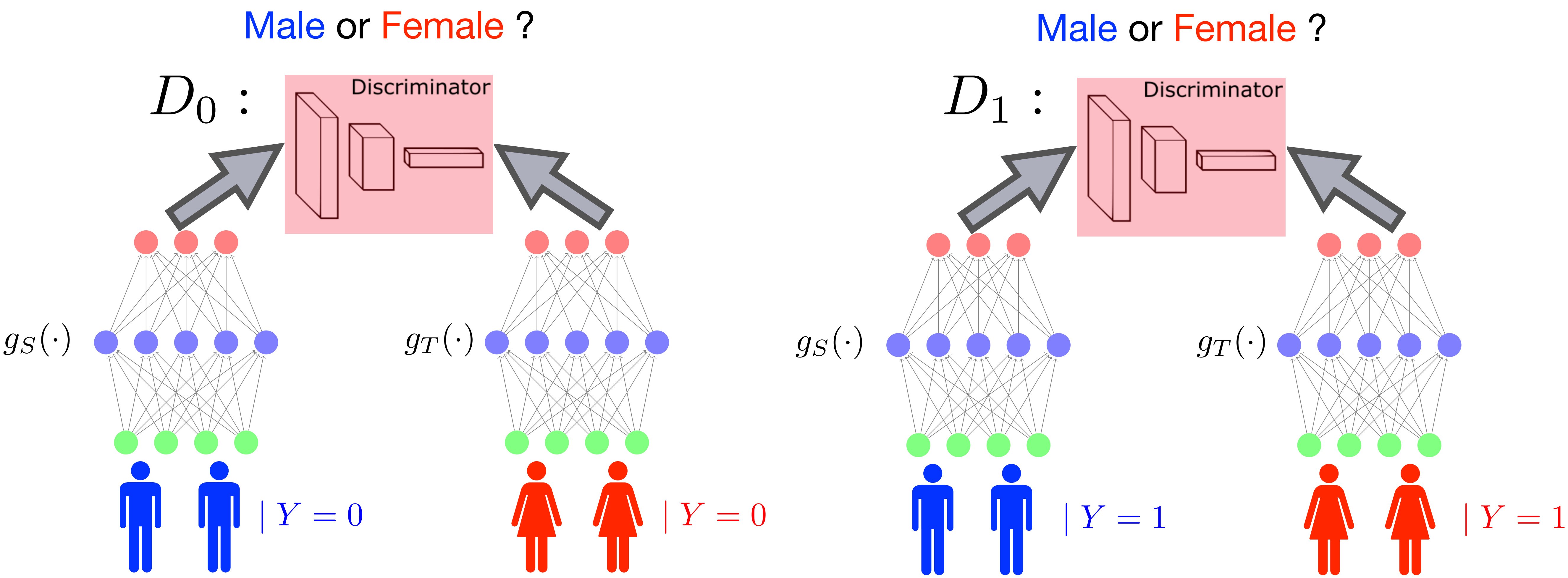
$$\varepsilon_{A=0}(h) + \varepsilon_{A=1}(h) \geq \Delta_{\mathrm{BR}}$$

Conditional Learning of Fair Representations

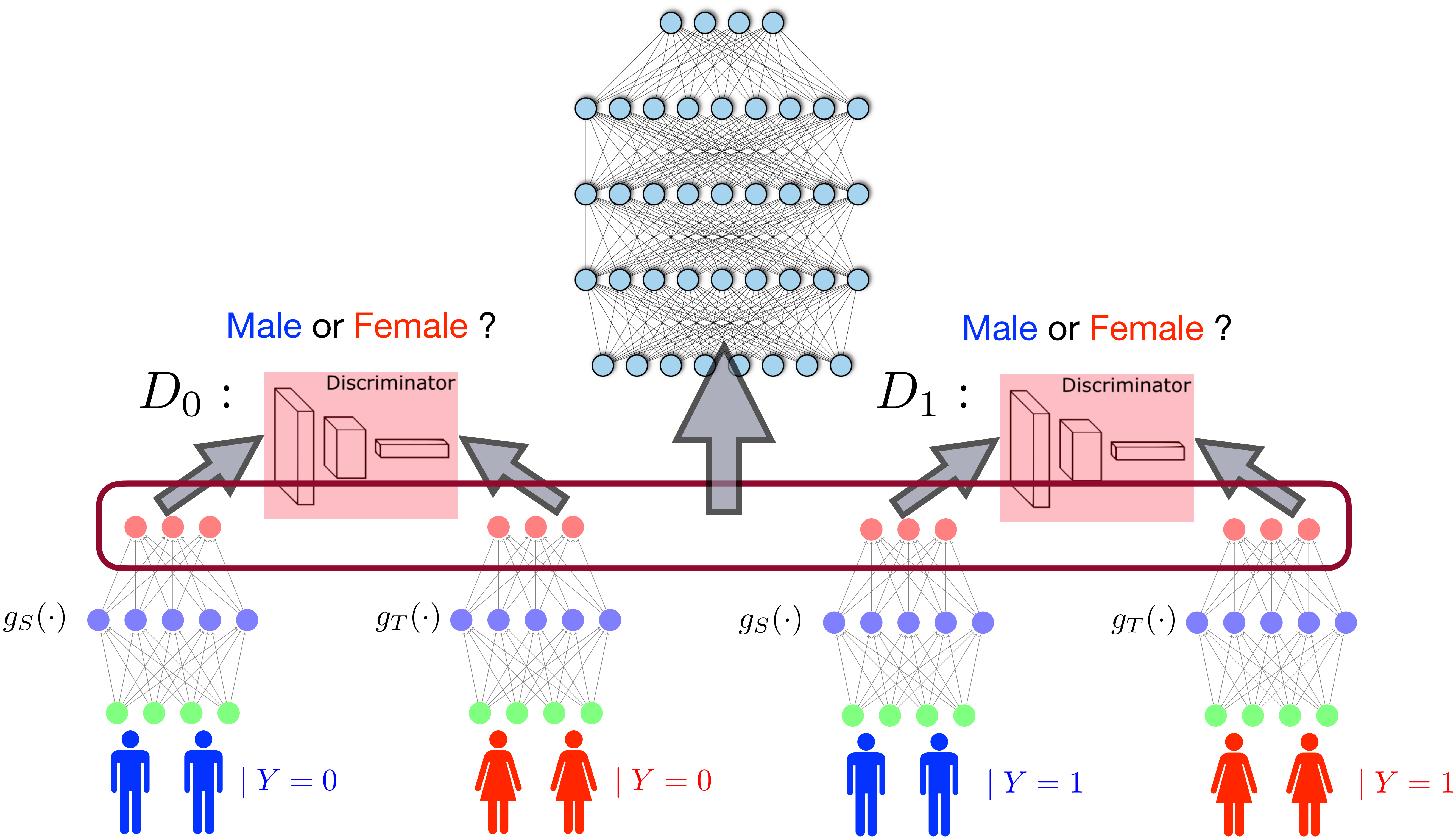
Conditional Learning of Fair Representations



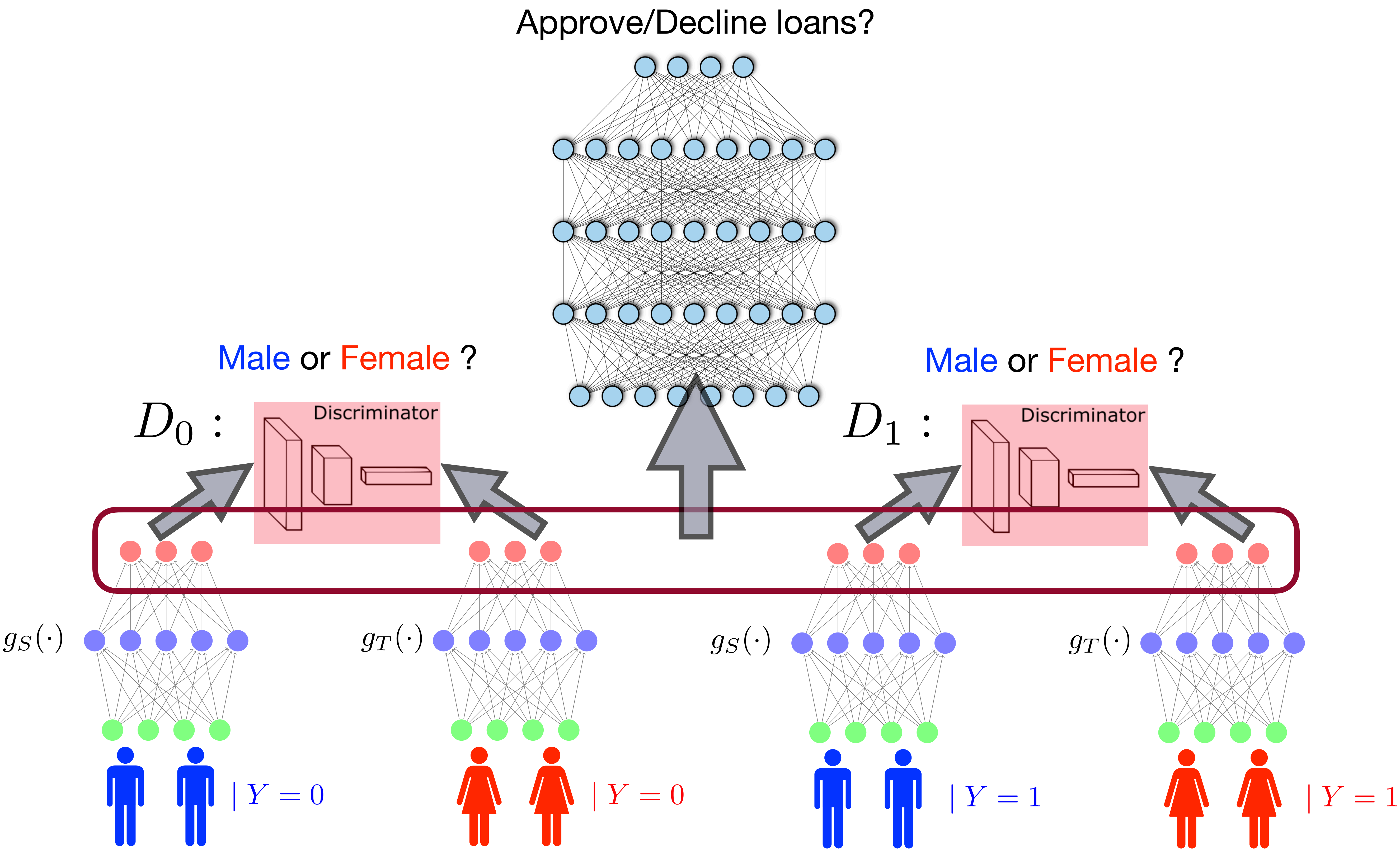
Conditional Learning of Fair Representations



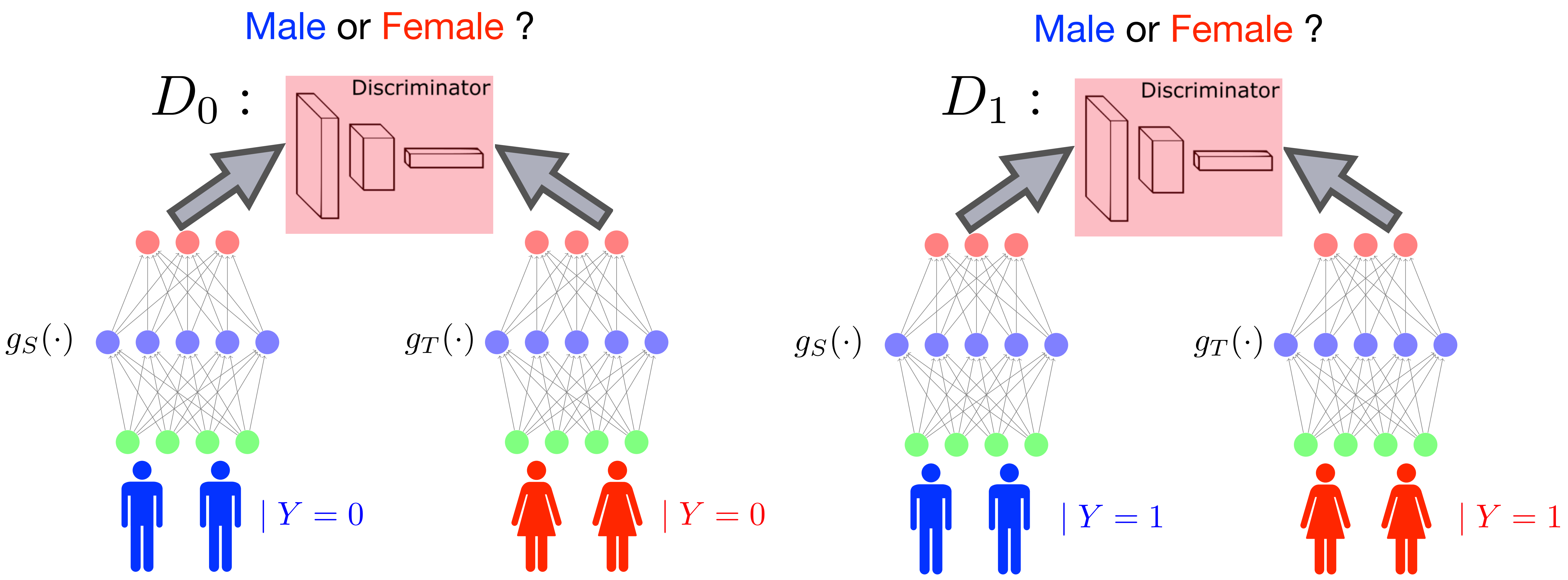
Conditional Learning of Fair Representations



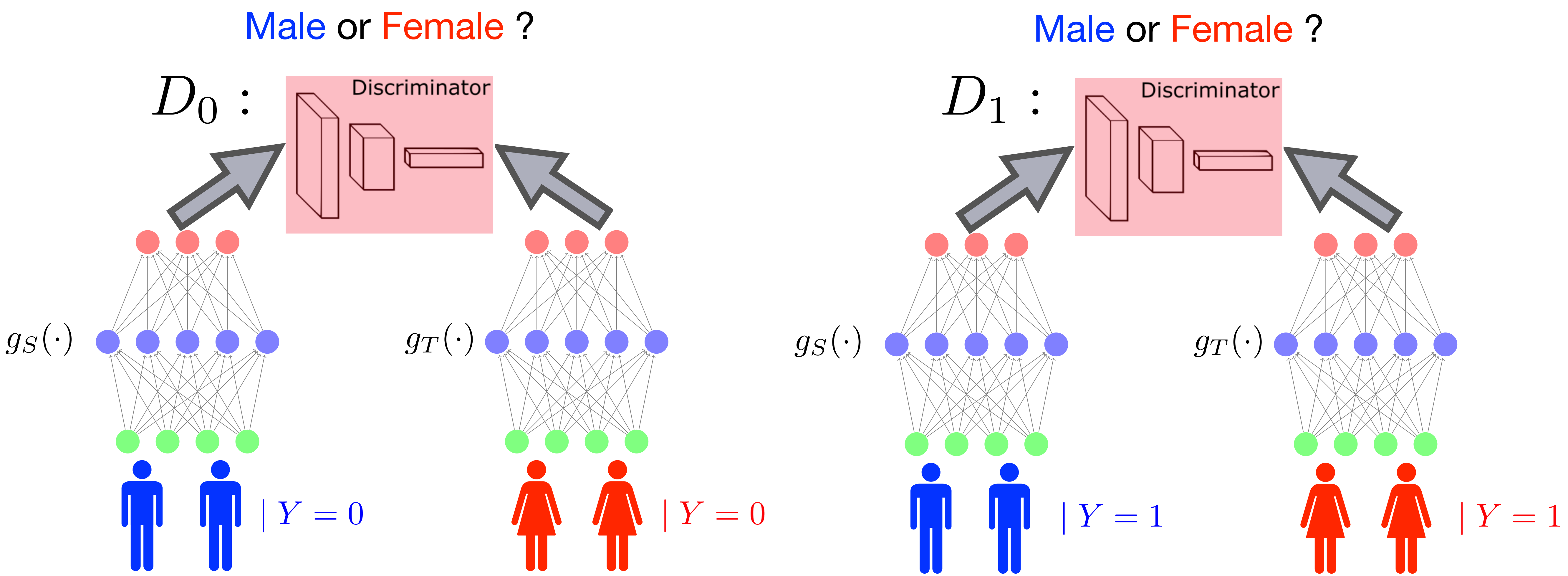
Conditional Learning of Fair Representations



Conditional Learning of Fair Representations

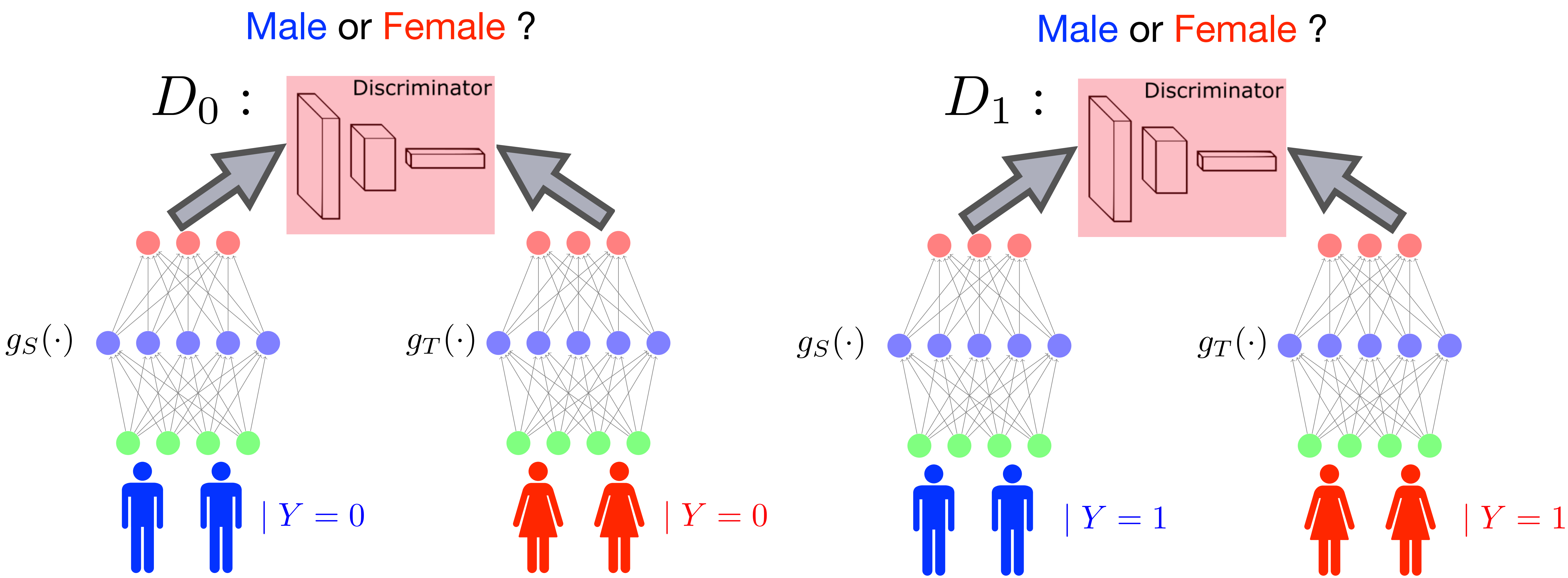


Conditional Learning of Fair Representations



$$\Pr_{A=0}(\widehat{Y} = 1 \mid Y = 0) \approx \Pr_{A=1}(\widehat{Y} = 1 \mid Y = 0) \qquad \Pr_{A=0}(\widehat{Y} = 0 \mid Y = 1) \approx \Pr_{A=1}(\widehat{Y} = 0 \mid Y = 1)$$

Conditional Learning of Fair Representations



$$\Pr_{A=0}(\widehat{Y} = 1 \mid Y = 0) \approx \Pr_{A=1}(\widehat{Y} = 1 \mid Y = 0)$$

$$\Pr_{A=0}(\widehat{Y} = 0 \mid Y = 1) \approx \Pr_{A=1}(\widehat{Y} = 0 \mid Y = 1)$$

Equalized False Positive Rate (FPR)

Equalized False Negative Rate (FNR)

An Error Decomposition Theorem

Approximately equal error rates across groups:

$$|\varepsilon_{A=0}(\widehat{Y}) - \varepsilon_{A=1}(\widehat{Y})| \leq \Delta_{\mathrm{BR}} \cdot (\mathrm{FPR}(\widehat{Y}) + \mathrm{FNR}(\widehat{Y})) + 2\max\left\{d\left(\mathcal{D}_{A=0}^{Z|Y=0}, \mathcal{D}_{A=1}^{Z|Y=0}\right), d\left(\mathcal{D}_{A=0}^{Z|Y=1}, \mathcal{D}_{A=1}^{Z|Y=1}\right)\right\}$$

$\Delta_{\mathrm{BR}} := |\Pr(Y=1 \mid A=0) - \Pr(Y=1 \mid A=1)|$

An Error Decomposition Theorem

Approximately equal error rates across groups:

$$|\varepsilon_{A=0}(\widehat{Y}) - \varepsilon_{A=1}(\widehat{Y})| \leq \Delta_{\mathrm{BR}} \cdot (\mathrm{FPR}(\widehat{Y}) + \mathrm{FNR}(\widehat{Y})) + 2\max\left\{d\left(\mathcal{D}_{A=0}^{Z|Y=0}, \mathcal{D}_{A=1}^{Z|Y=0}\right), d\left(\mathcal{D}_{A=0}^{Z|Y=1}, \mathcal{D}_{A=1}^{Z|Y=1}\right)\right\}$$

distance between
marginal label distributions

$$\Delta_{\mathrm{BR}} := |\Pr(Y=1 \mid A=0) - \Pr(Y=1 \mid A=1)|$$

An Error Decomposition Theorem

Approximately equal error rates across groups:

$$|\varepsilon_{A=0}(\widehat{Y}) - \varepsilon_{A=1}(\widehat{Y})| \leq \Delta_{\mathrm{BR}} \cdot (\mathrm{FPR}(\widehat{Y}) + \mathrm{FNR}(\widehat{Y})) + 2\max\left\{d\left(\mathcal{D}_{A=0}^{Z|Y=0}, \mathcal{D}_{A=1}^{Z|Y=0}\right), d\left(\mathcal{D}_{A=0}^{Z|Y=1}, \mathcal{D}_{A=1}^{Z|Y=1}\right)\right\}$$

distance between marginal label distributions

distance between conditional feature distributions

$\Delta_{\mathrm{BR}} := |\Pr(Y=1 \mid A=0) - \Pr(Y=1 \mid A=1)|$

An Error Decomposition Theorem

Approximately equal error rates across groups:

$$|\varepsilon_{A=0}(\widehat{Y}) - \varepsilon_{A=1}(\widehat{Y})| \leq \Delta_{\mathrm{BR}} \cdot (\mathrm{FPR}(\widehat{Y}) + \mathrm{FNR}(\widehat{Y})) + 2\max\left\{d\left(\mathcal{D}_{A=0}^{Z|Y=0}, \mathcal{D}_{A=1}^{Z|Y=0}\right), d\left(\mathcal{D}_{A=0}^{Z|Y=1}, \mathcal{D}_{A=1}^{Z|Y=1}\right)\right\}$$

distance between
marginal label distributions

distance between
conditional feature distributions

$\Delta_{\mathrm{BR}} := |\Pr(Y=1 \mid A=0) - \Pr(Y=1 \mid A=1)|$

An Error Decomposition Theorem

Approximately equal error rates across groups:

$$|\varepsilon_{A=0}(\widehat{Y}) - \varepsilon_{A=1}(\widehat{Y})| \leq \Delta_{\mathrm{BR}} \cdot (\mathrm{FPR}(\widehat{Y}) + \mathrm{FNR}(\widehat{Y})) + 2\max\left\{d\left(\mathcal{D}_{A=0}^{Z|Y=0}, \mathcal{D}_{A=1}^{Z|Y=0}\right), d\left(\mathcal{D}_{A=0}^{Z|Y=1}, \mathcal{D}_{A=1}^{Z|Y=1}\right)\right\}$$

distance between marginal label distributions

distance between conditional feature distributions

Theorem (informal): Furthermore, if $Z \perp A \mid Y$, then the gap of SP for any $\hat{Y} = h(Z)$ is smaller than the gap of the optimal classifier Y

$\Delta_{\mathrm{BR}} := |\Pr(Y=1 \mid A=0) - \Pr(Y=1 \mid A=1)|$

An Error Decomposition Theorem

Approximately equal error rates across groups:

$$|\varepsilon_{A=0}(\widehat{Y}) - \varepsilon_{A=1}(\widehat{Y})| \leq \Delta_{\mathrm{BR}} \cdot (\mathrm{FPR}(\widehat{Y}) + \mathrm{FNR}(\widehat{Y})) + 2\max\left\{d\left(\mathcal{D}_{A=0}^{Z|Y=0}, \mathcal{D}_{A=1}^{Z|Y=0}\right), d\left(\mathcal{D}_{A=0}^{Z|Y=1}, \mathcal{D}_{A=1}^{Z|Y=1}\right)\right\}$$

distance between marginal label distributions

distance between conditional feature distributions

Theorem (informal): Furthermore, if $Z \perp A \mid Y$, then the gap of SP for any $\hat{Y} = h(Z)$ is smaller than the gap of the optimal classifier Y

$$\Delta_{\mathrm{DP}}(\hat{Y}) \leq \Delta_{\mathrm{DP}}(Y)$$

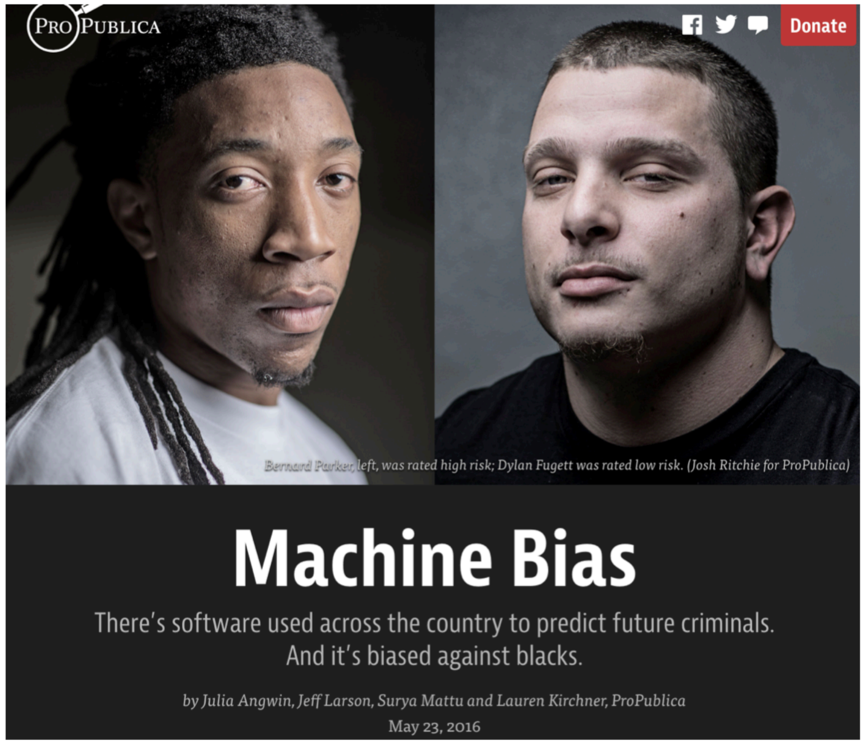
$$\Delta_{\mathrm{DP}}(\hat{Y}) := \left|\Pr(\hat{Y} = 1 \mid A = 0) - \Pr(\hat{Y} = 1 \mid A = 1)\right|$$

$$\Delta_{\mathrm{BR}} := |\Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1)|$$

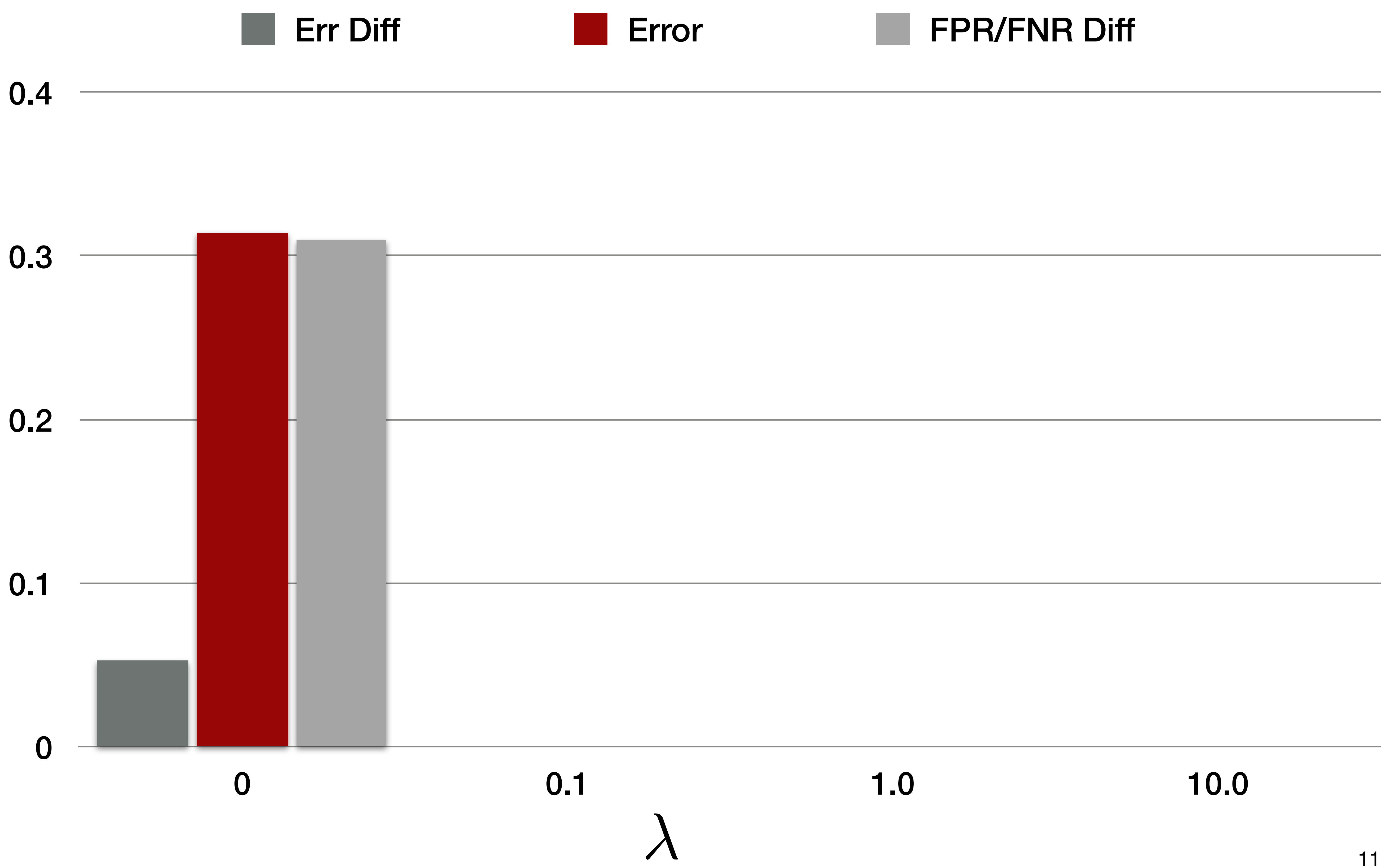
Experiment: Recidivism Prediction

COMPAS

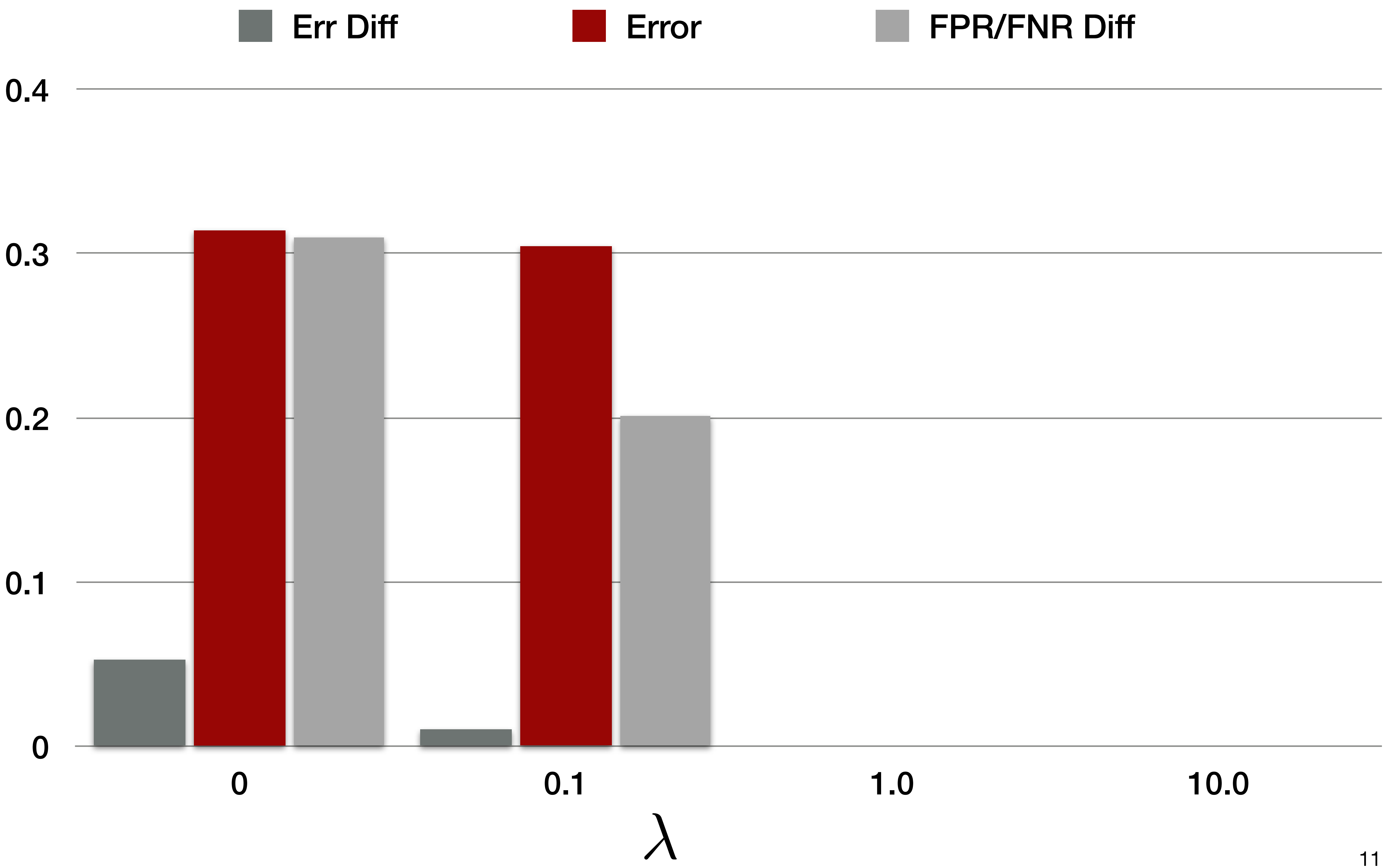
- Train/Test: 4,320/1,852 instances from the Northpointe
- Target task: 0/1 classification (recidivism?)
- Sensitive attribute: race (Black/White)
- Other attributes: gender, education, prior arrest history, … (12 total)
- Difference of base rate: $\Delta_{\mathrm{BR}} = 0.129$



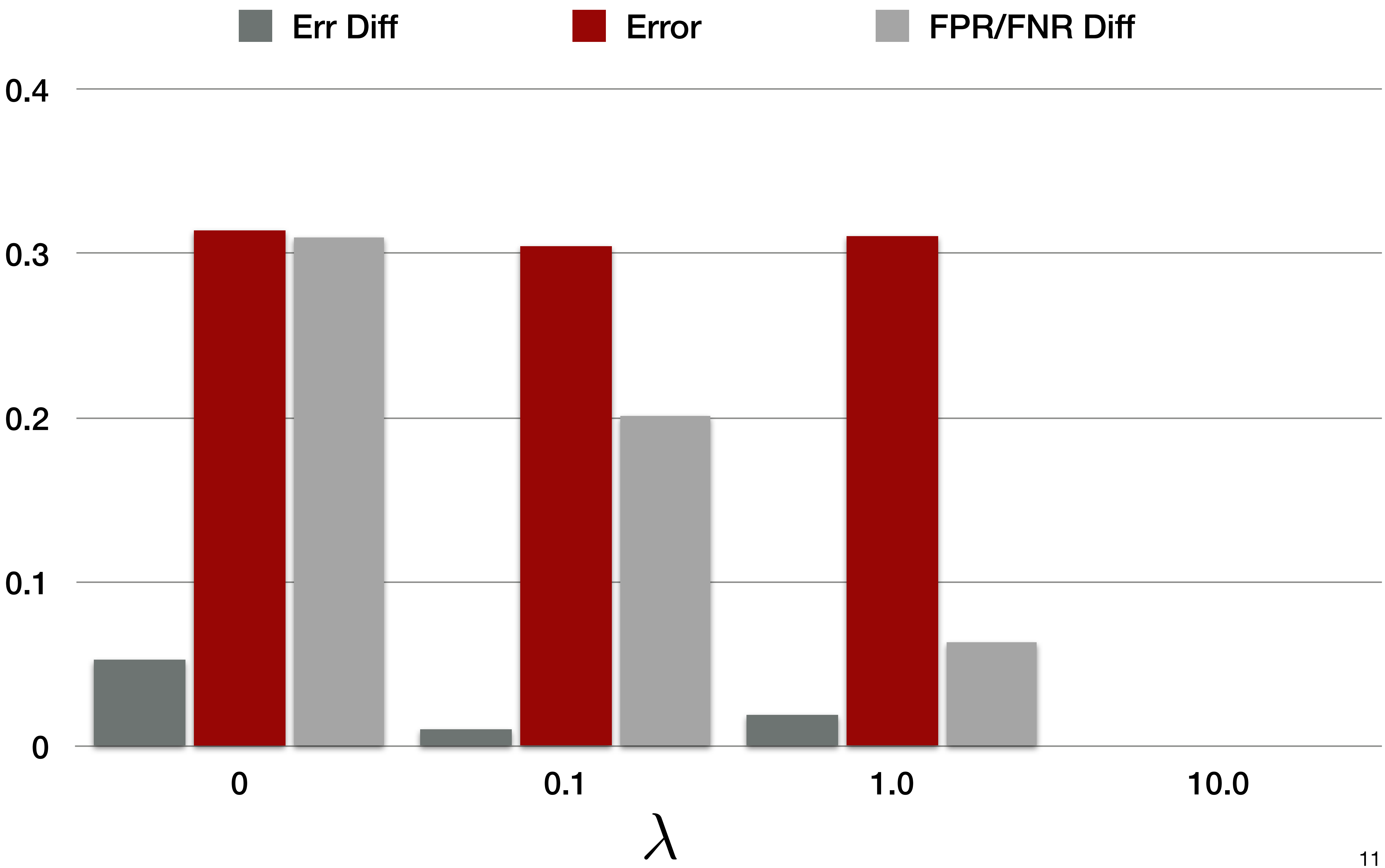
Experiment: Recidivism Prediction



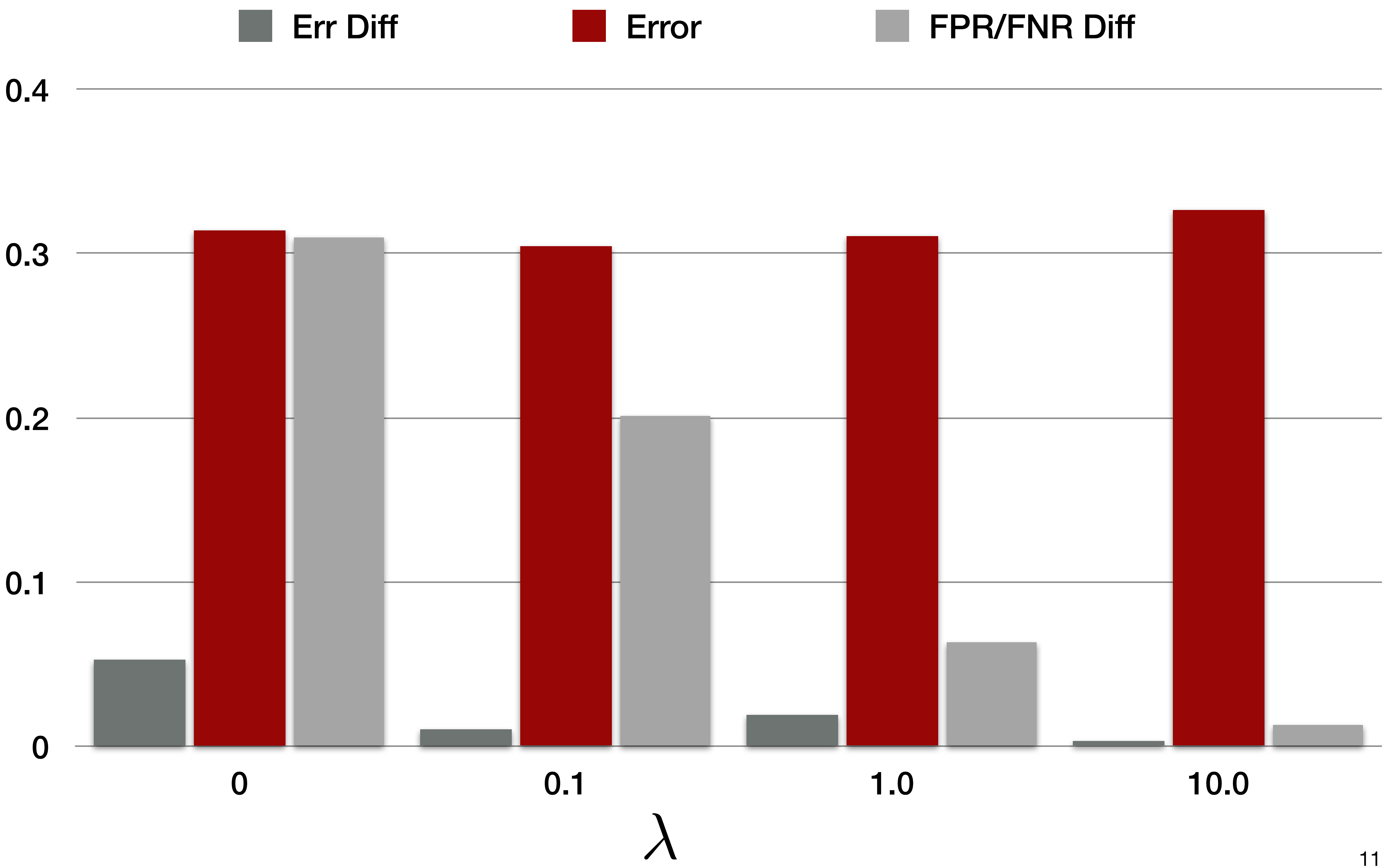
Experiment: Recidivism Prediction



Experiment: Recidivism Prediction



Experiment: Recidivism Prediction



Conclusion

From a representation learning perspective, design algorithmic intervention to

- Seek for equalized odds and accuracy parity simultaneously
- Not harm the existing statistical parity gap
- Practical implementation using adversarial training with two auditor networks