

# Conditional Learning of Fair Representations

---

Han Zhao

[han.zhao@cs.cmu.edu](mailto:han.zhao@cs.cmu.edu)

Machine Learning Department  
Carnegie Mellon University

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

# Potential Bias of Data in High-stakes Domains

---

# Potential Bias of Data in High-stakes Domains

The image shows a hand writing on a 'STUDENT LOAN APPLICATION' form. The form is titled 'STUDENT LOAN APPLICATION' in large, bold letters. It contains several sections with checkboxes and text input fields. The sections include:

- Applying For A:** ☐ Permit ☐ ID Card ☐ Renewal ☐ Replacement
- Your Personal:** Full Last Name, Full First Name, Date of birth, Nationality, ID card number and Details, Your Personal Details (Height, Eye color, Status: ☐ Single ☐ Married ☐ Divorced, Address where you live: Unit No., Street, State, Post Code, Town/City, Country).
- Identification Information:** Driver license? ☐ Yes ☐ No, Learner permit? ☐ Yes ☐ No, Non-driver ID Card? ☐ Yes ☐ No.
- Contact Details:** Home Phone, Mobile Phone, Fax, Address where you live (This address will be used for all correspondence), Has your mailing address changed? (What is the change and the reason for it?), Other change: (new license class, wrong date of birth, etc.).

A calculator is visible in the bottom left corner, showing the number 8980. There are also some papers and a small envelope in the background.

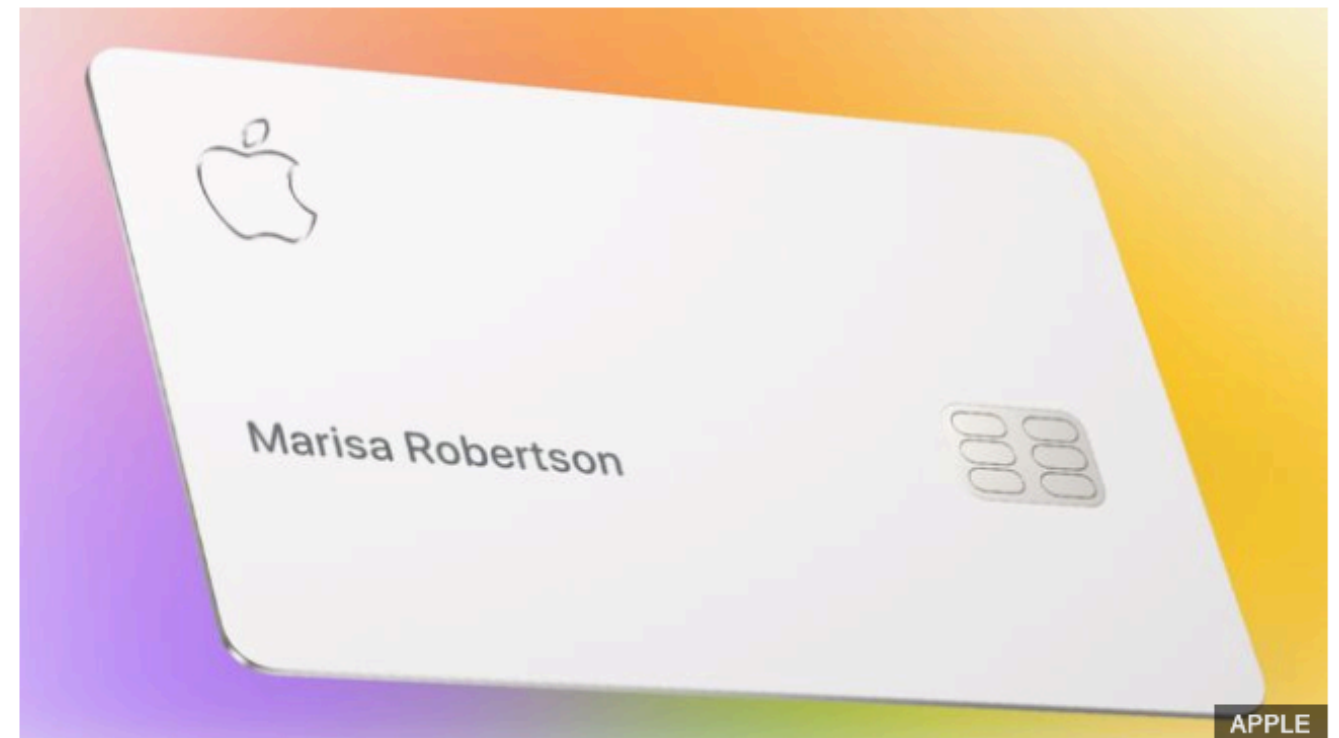
# Potential Bias of Data in High-stakes Domains



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

🕒 11 November 2019

[f](#) [m](#) [t](#) [e](#) [Share](#)



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



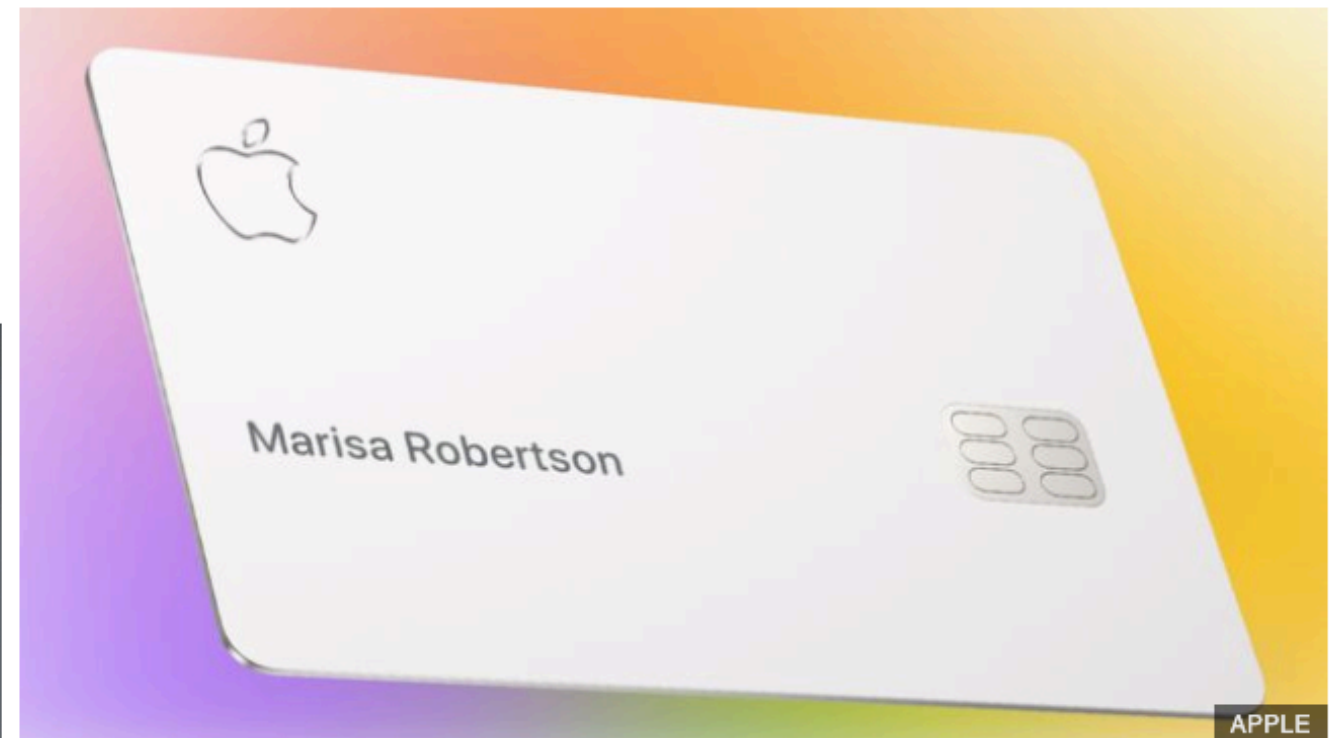
# Potential Bias of Data in High-stakes Domains



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

11 November 2019

f m t e Share



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



## Machine Bias

There's software used across the country to predict future criminals.  
And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016



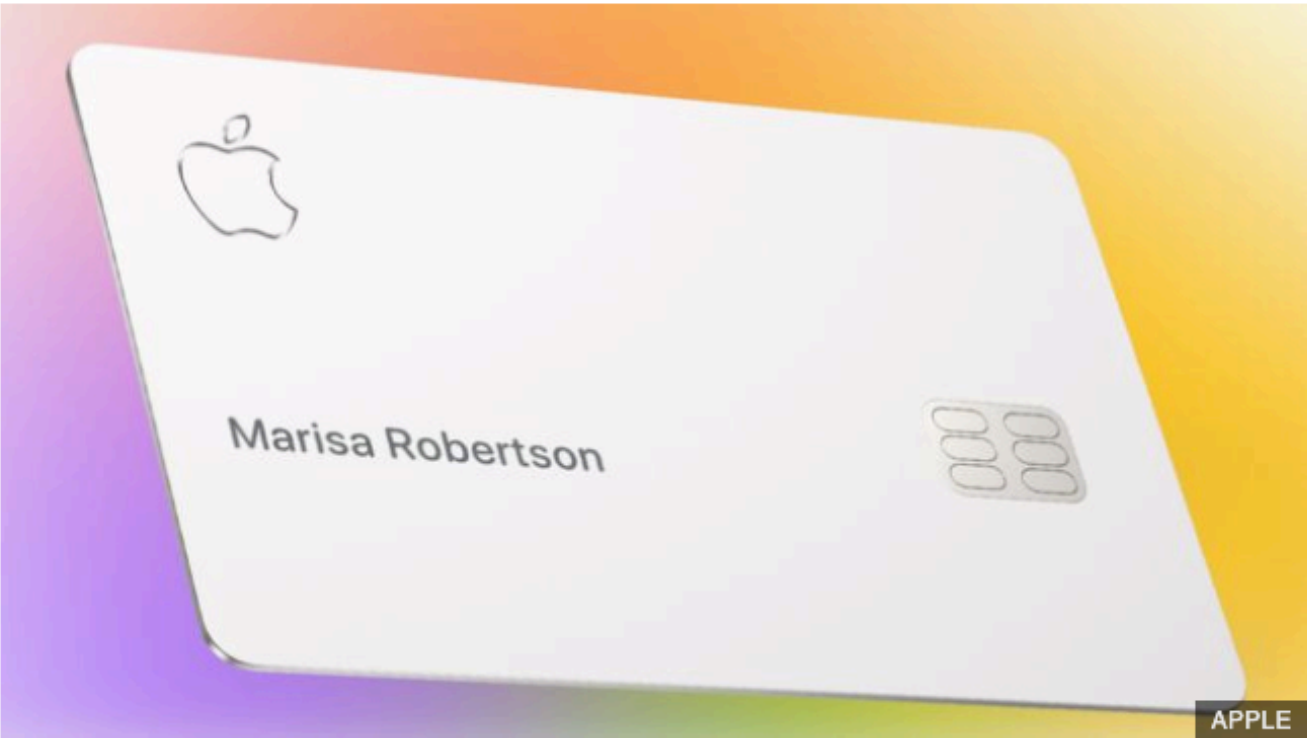
# Potential Bias of Data in High-stakes Domains



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

11 November 2019

[f](#) [m](#) [t](#) [e](#) [Share](#)



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



## Machine Bias

There's software used across the country to predict future crime. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica  
May 23, 2016

TECHNOLOGY NEWS OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO

## Amazon scraps secret AI recruiting tool that showed bias against women

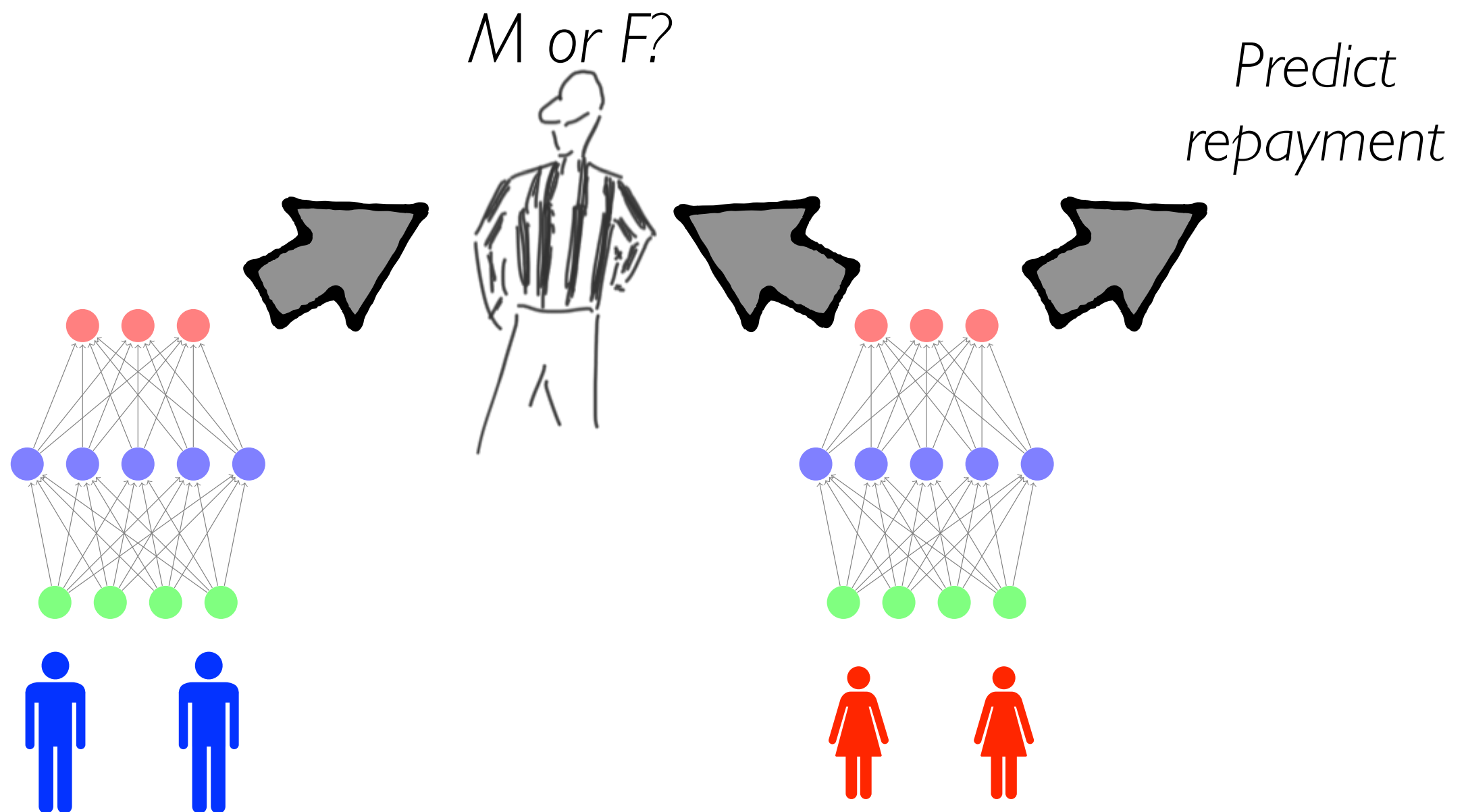
Jeffrey Dastin

8 MIN READ

[t](#) [f](#)

# This Work

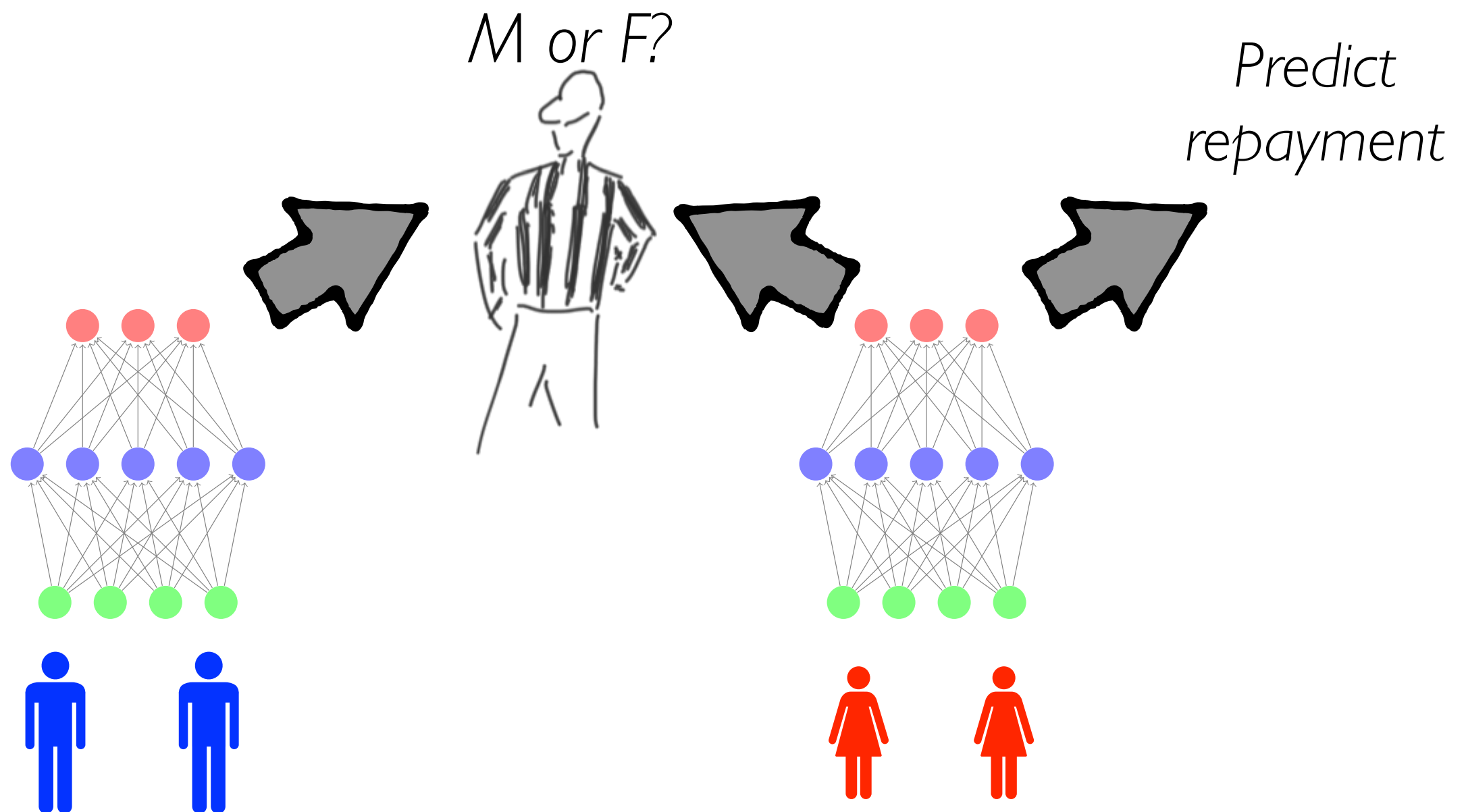
---



# This Work

---

From a representation learning perspective, design algorithmic intervention to



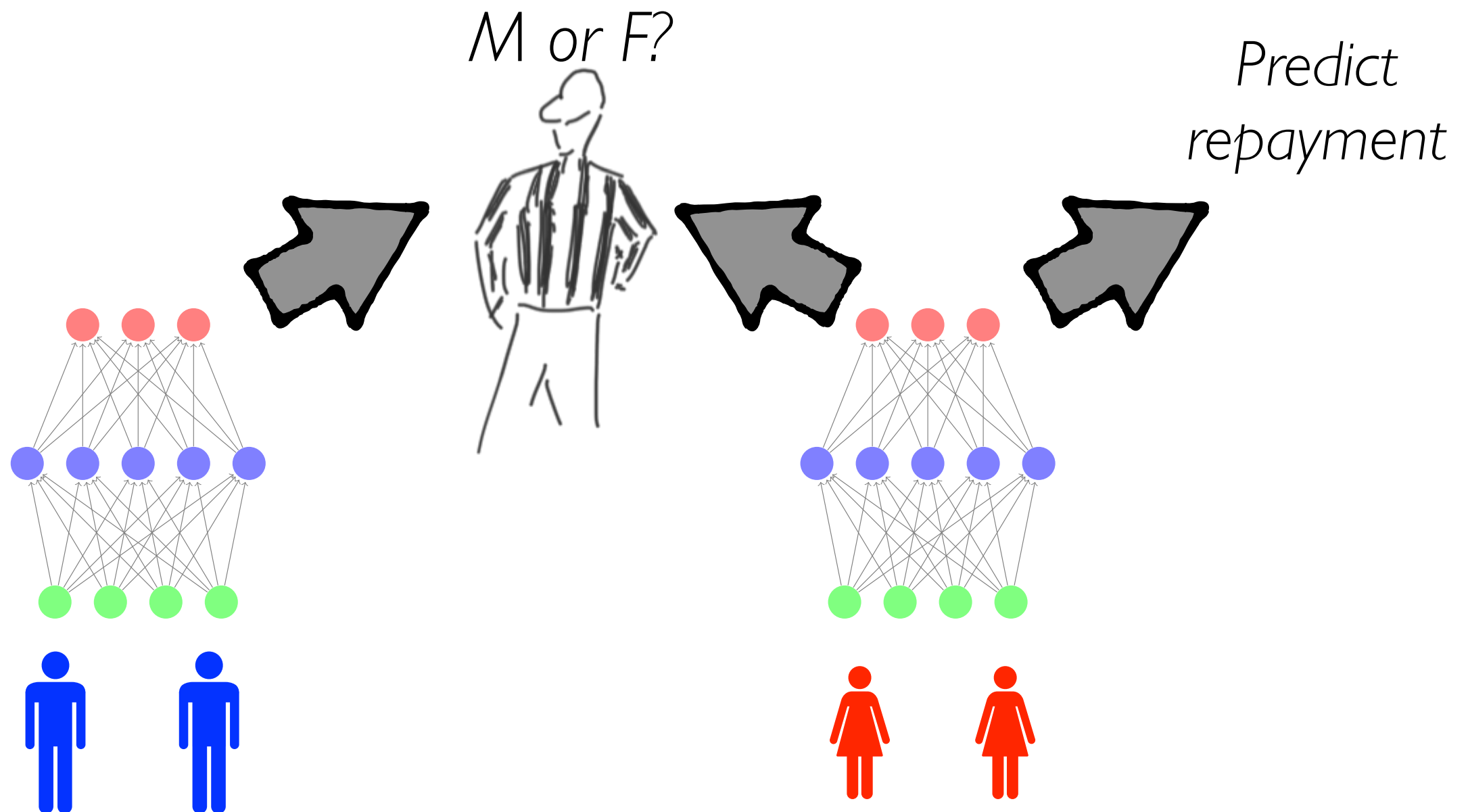


# This Work

---

From a representation learning perspective, design algorithmic intervention to

- Seek for equalized odds and accuracy parity simultaneously

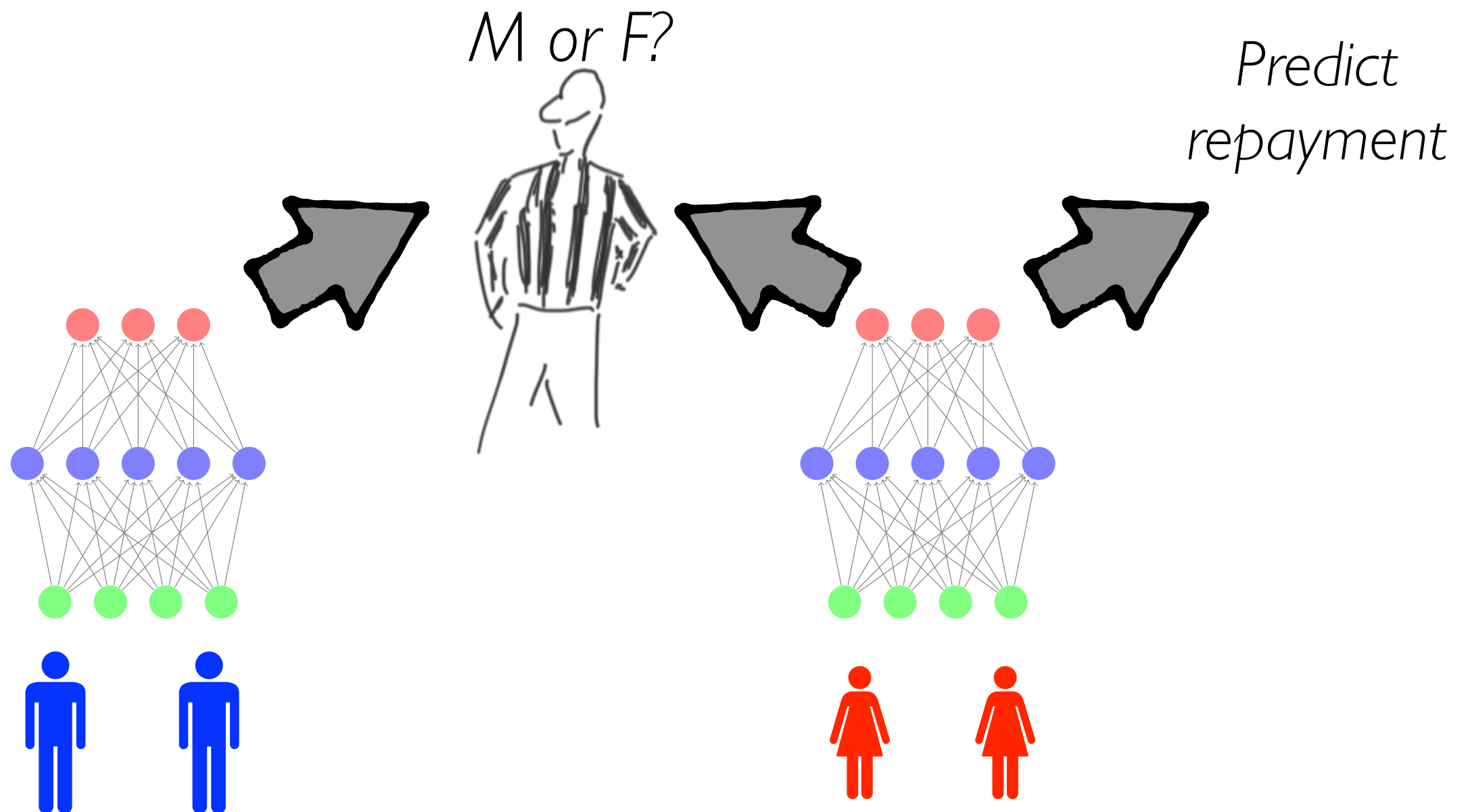


# This Work

---

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?



**Arvind Narayanan** ✓

@random\_walker

Follow

I wrote up a 2-pager titled "21 fairness definitions and their politics" based on the tweetstorm below and it was accepted at a tutorial for the Conference on Fairness, Accountability, and Transparency!

Here it is (with minor edits):

[docs.google.com/document/d/1bn ...](https://docs.google.com/document/d/1bn...)

See you on Feb 23/24.

**Arvind Narayanan** ✓ @random\_walker

When I tell my computer science colleagues that there are so many fairness definitions, they are often surprised and/or confused. [Thread]

[twitter.com/random\\_walker/...](https://twitter.com/random_walker/...)

Show this thread

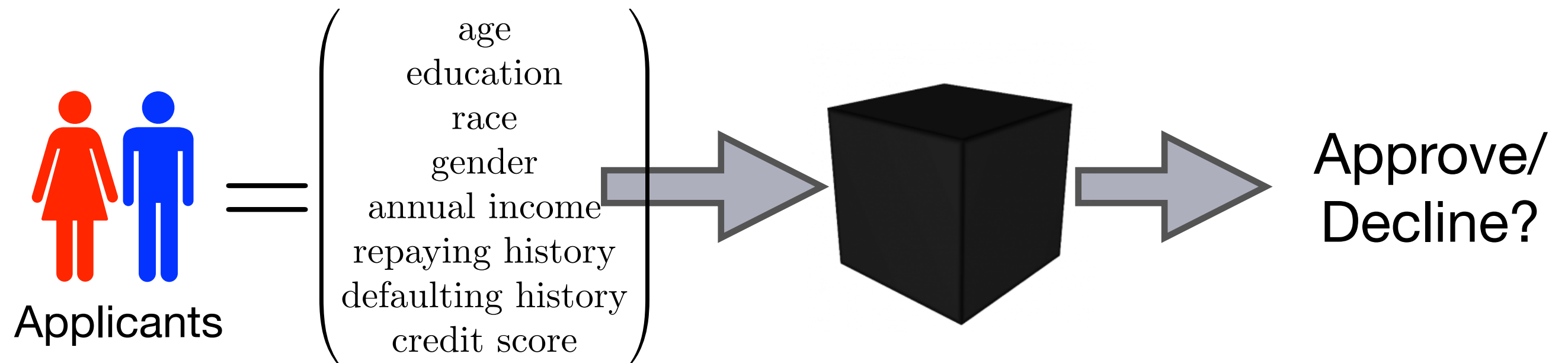
Definition	Paper	Citation #
Group fairness or statistical parity	[12]	208
Conditional statistical parity	[11]	29
Predictive parity	[10]	57
False positive error rate balance	[10]	57
False negative error rate balance	[10]	57
Equalised odds	[14]	106
Conditional use accuracy equality	[8]	18
Overall accuracy 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

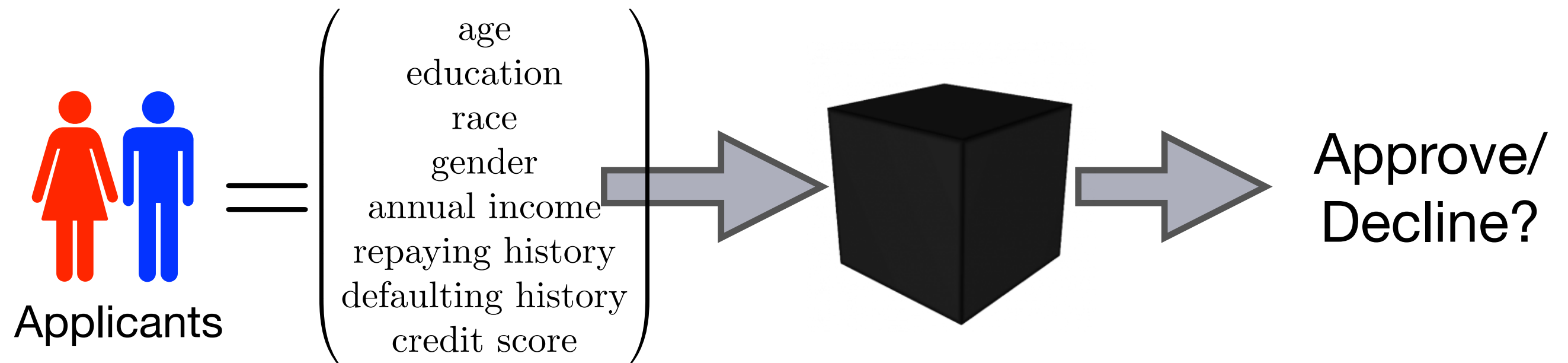




# Statistical Definition of Fairness

---

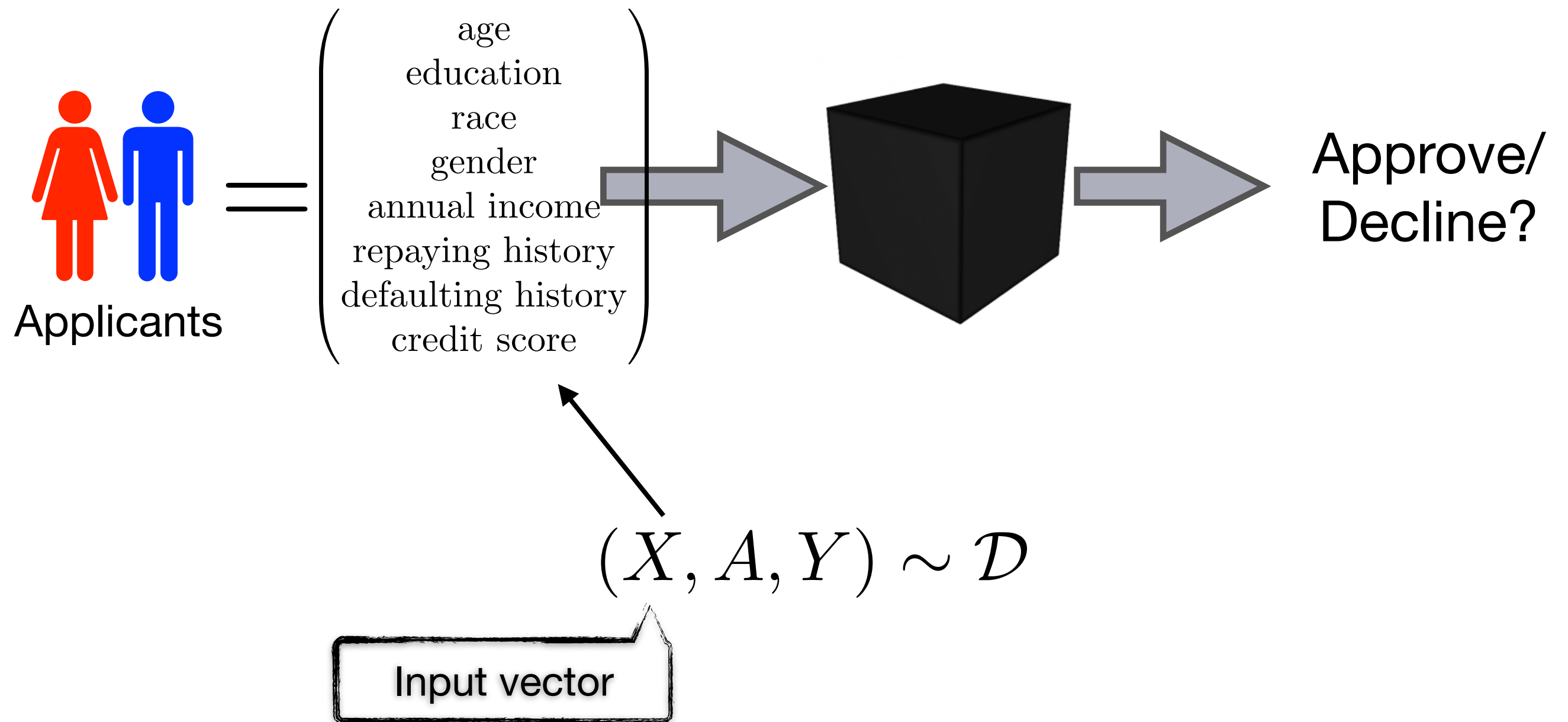
## Example in loan application



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

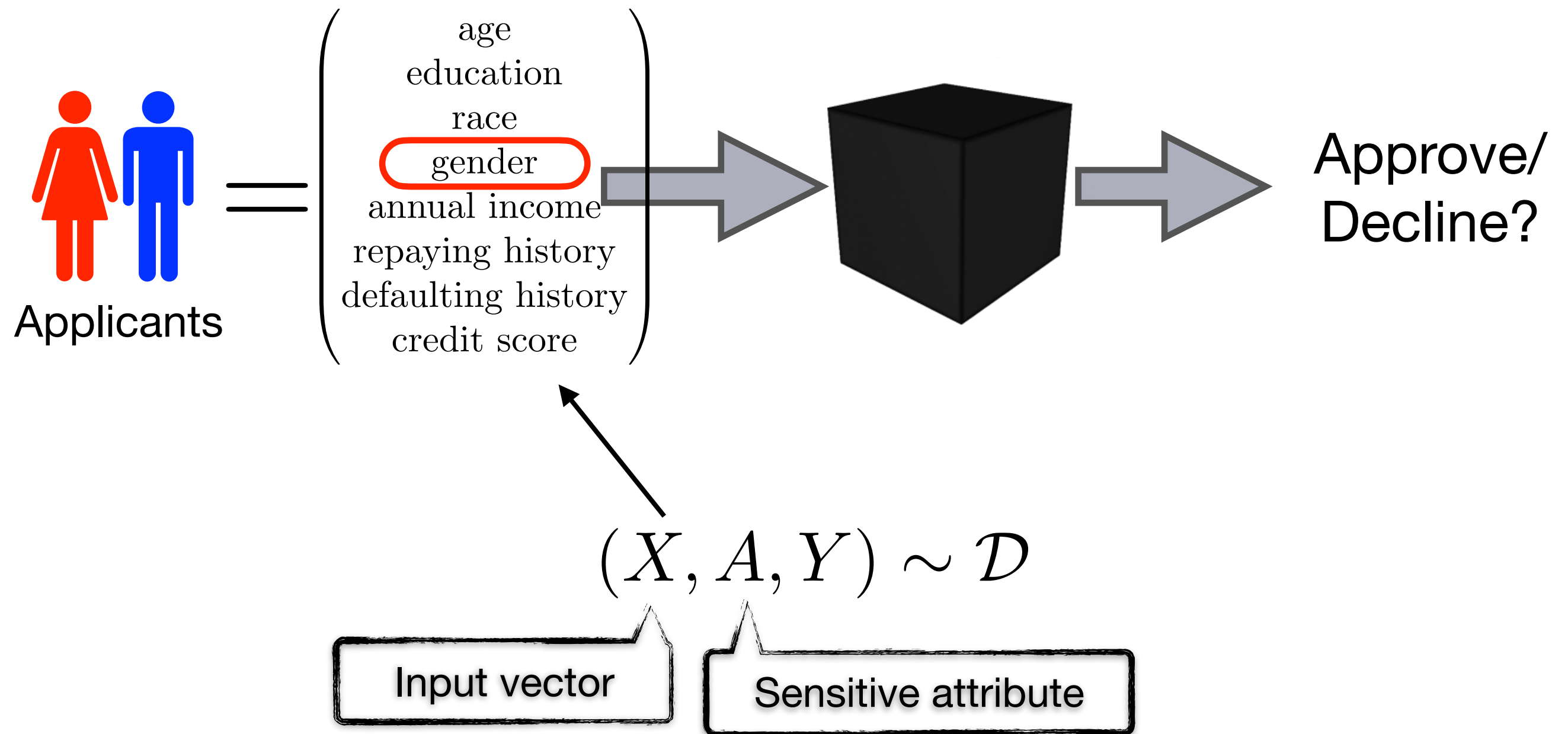
# Statistical Definition of Fairness

## Example in loan application



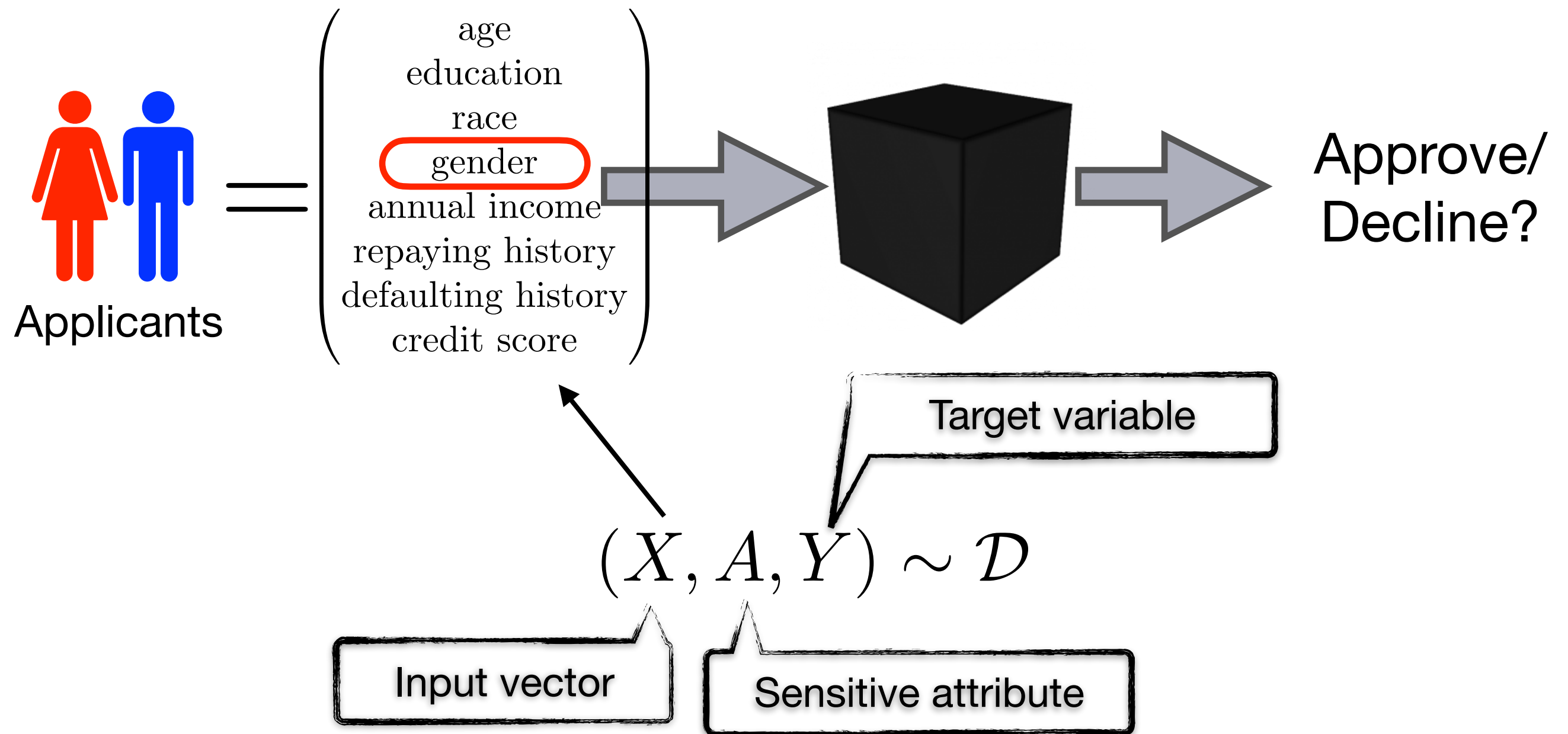
# Statistical Definition of Fairness

## Example in loan application



# Statistical Definition of Fairness

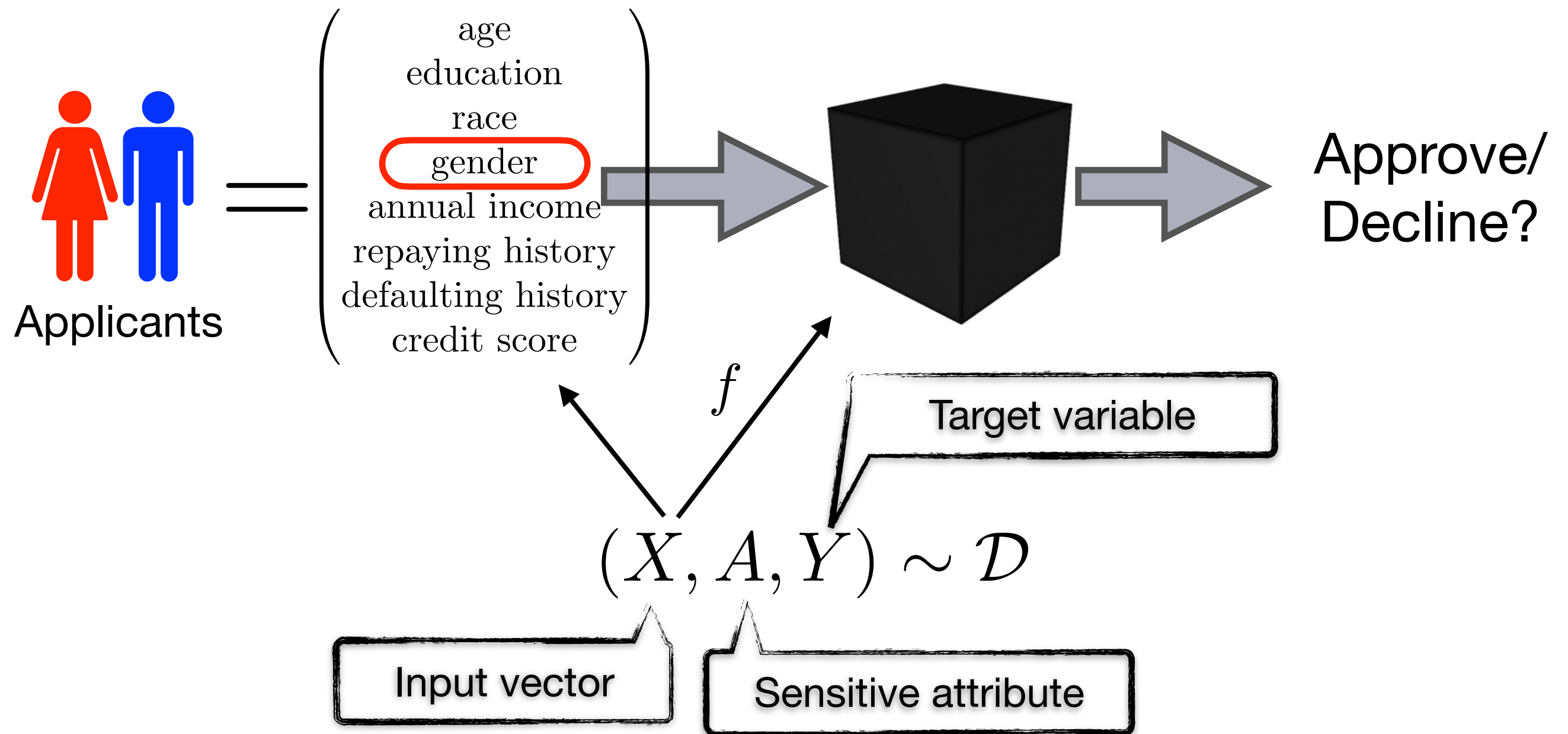
## Example in loan application





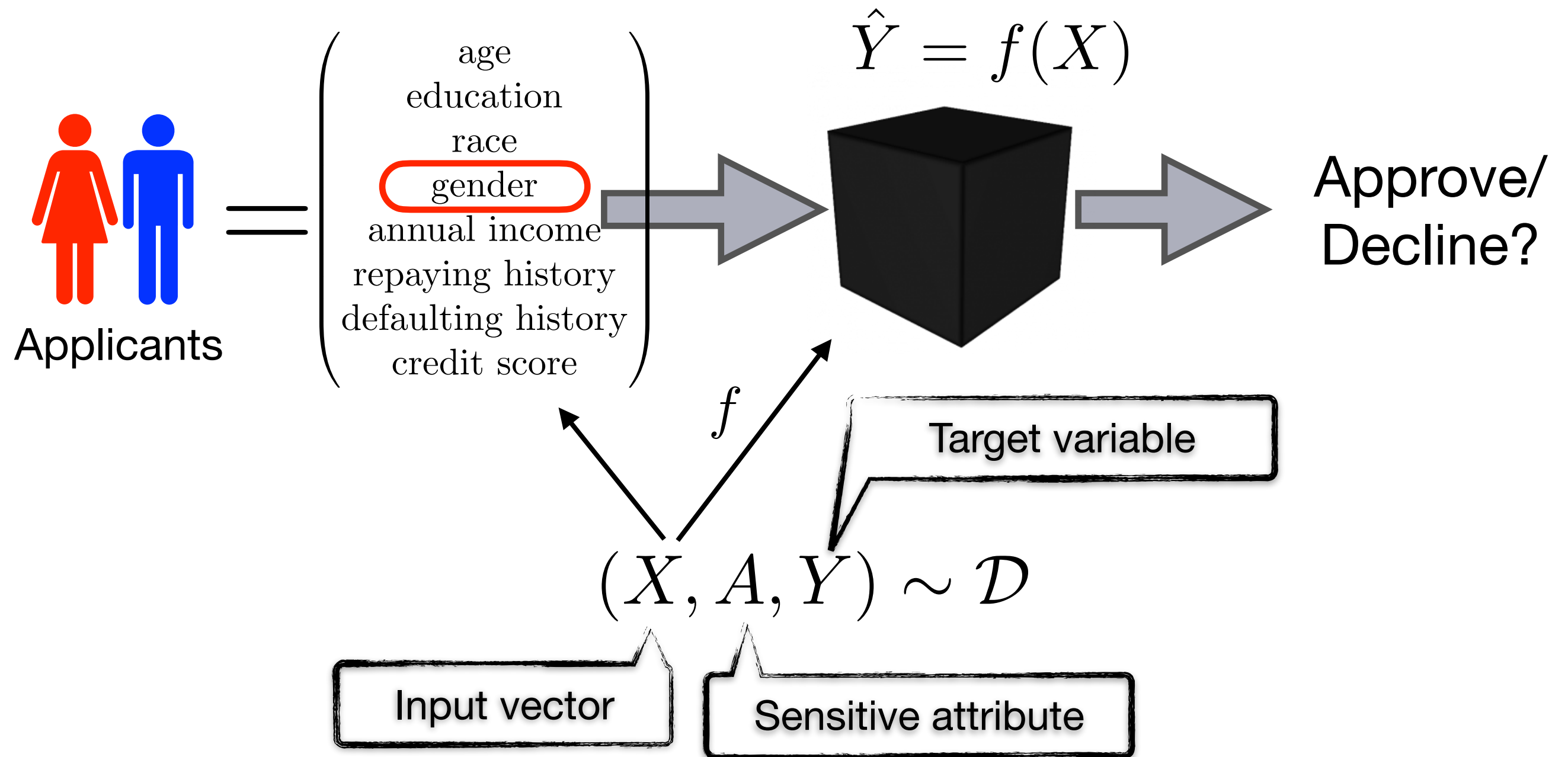
# Statistical Definition of Fairness

## Example in loan application



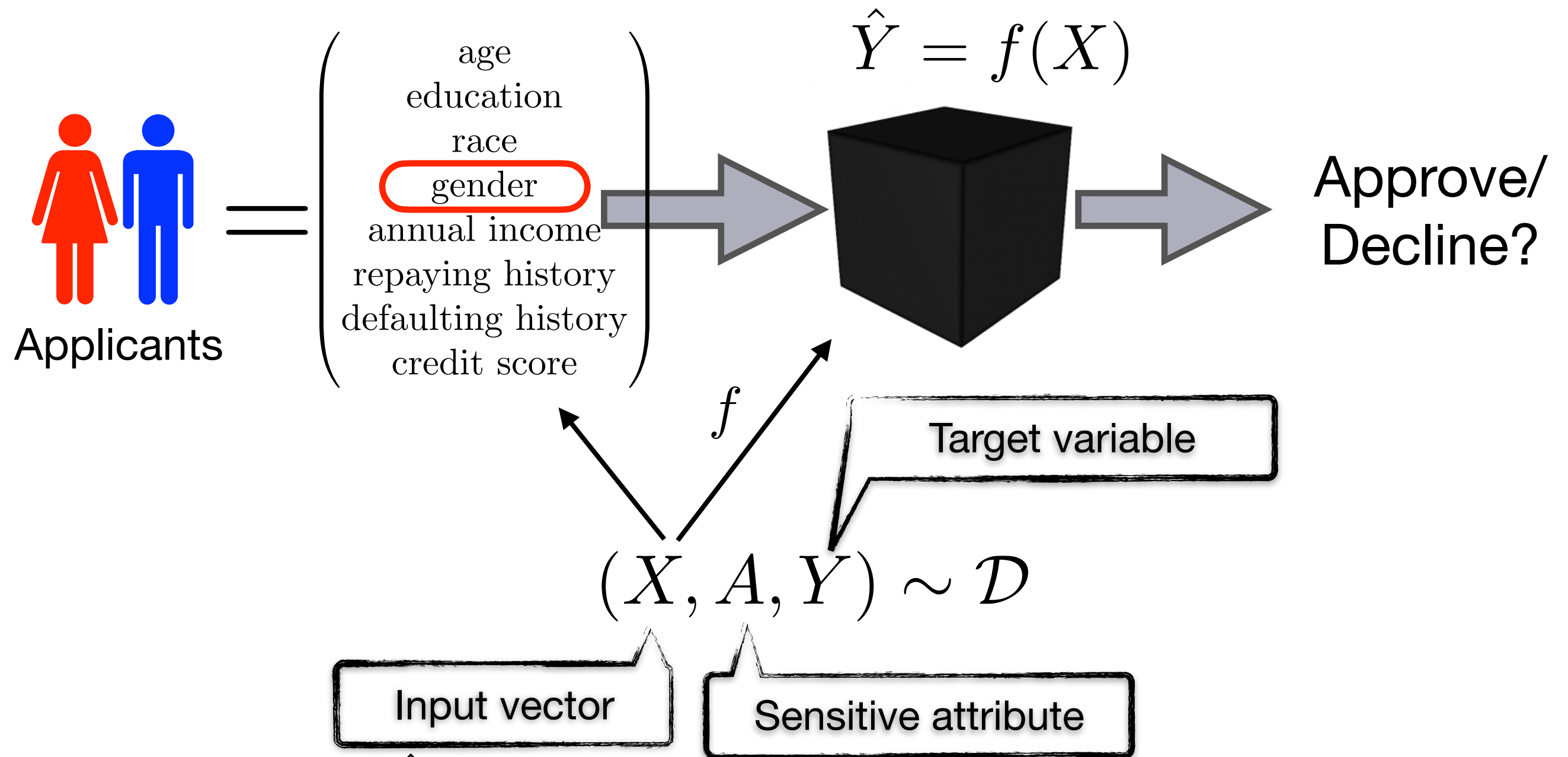
# Statistical Definition of Fairness

## Example in loan application



# Statistical Definition of Fairness

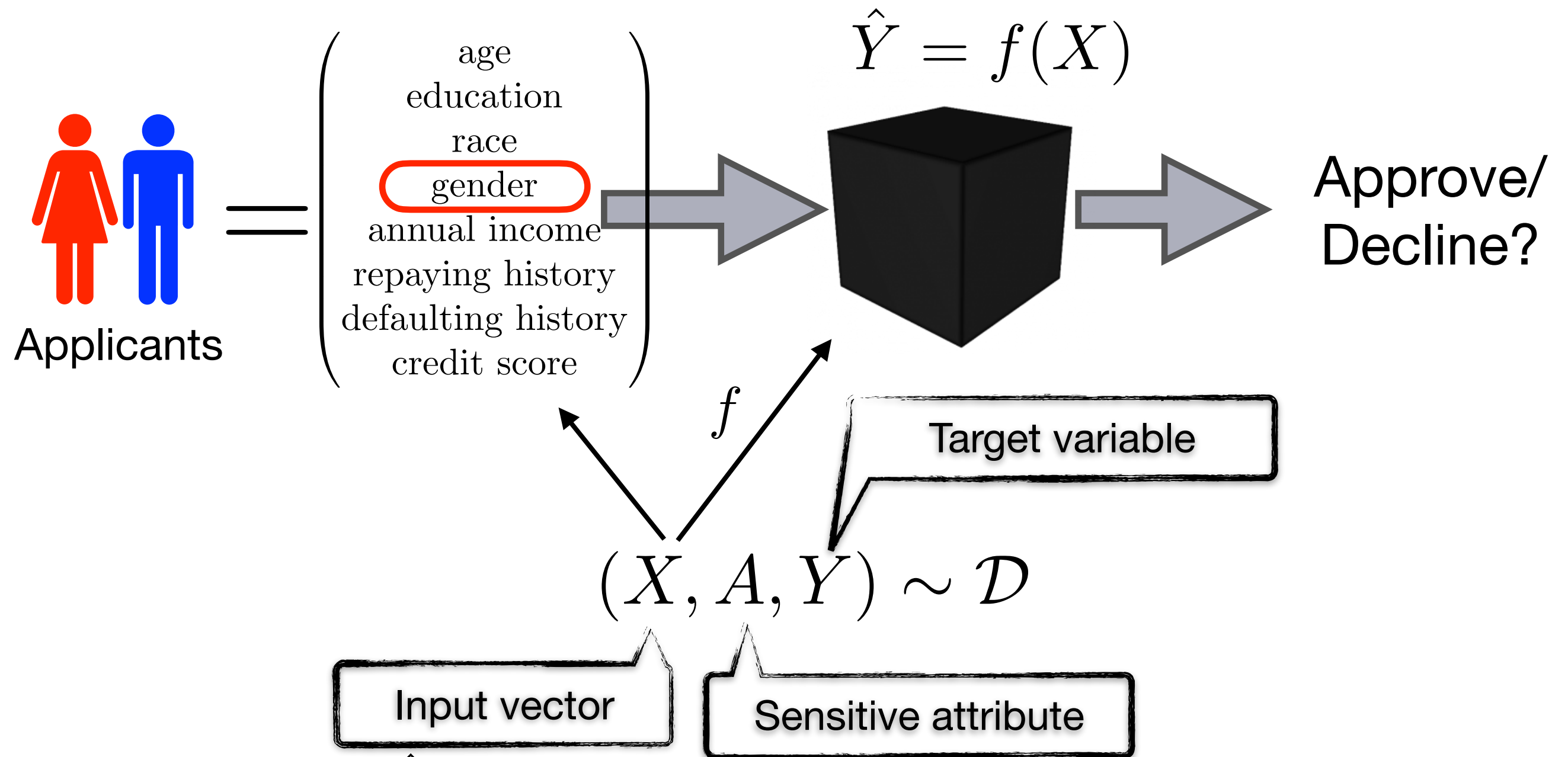
## Example in loan application



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

# Statistical Definition of Fairness

## Example in loan application

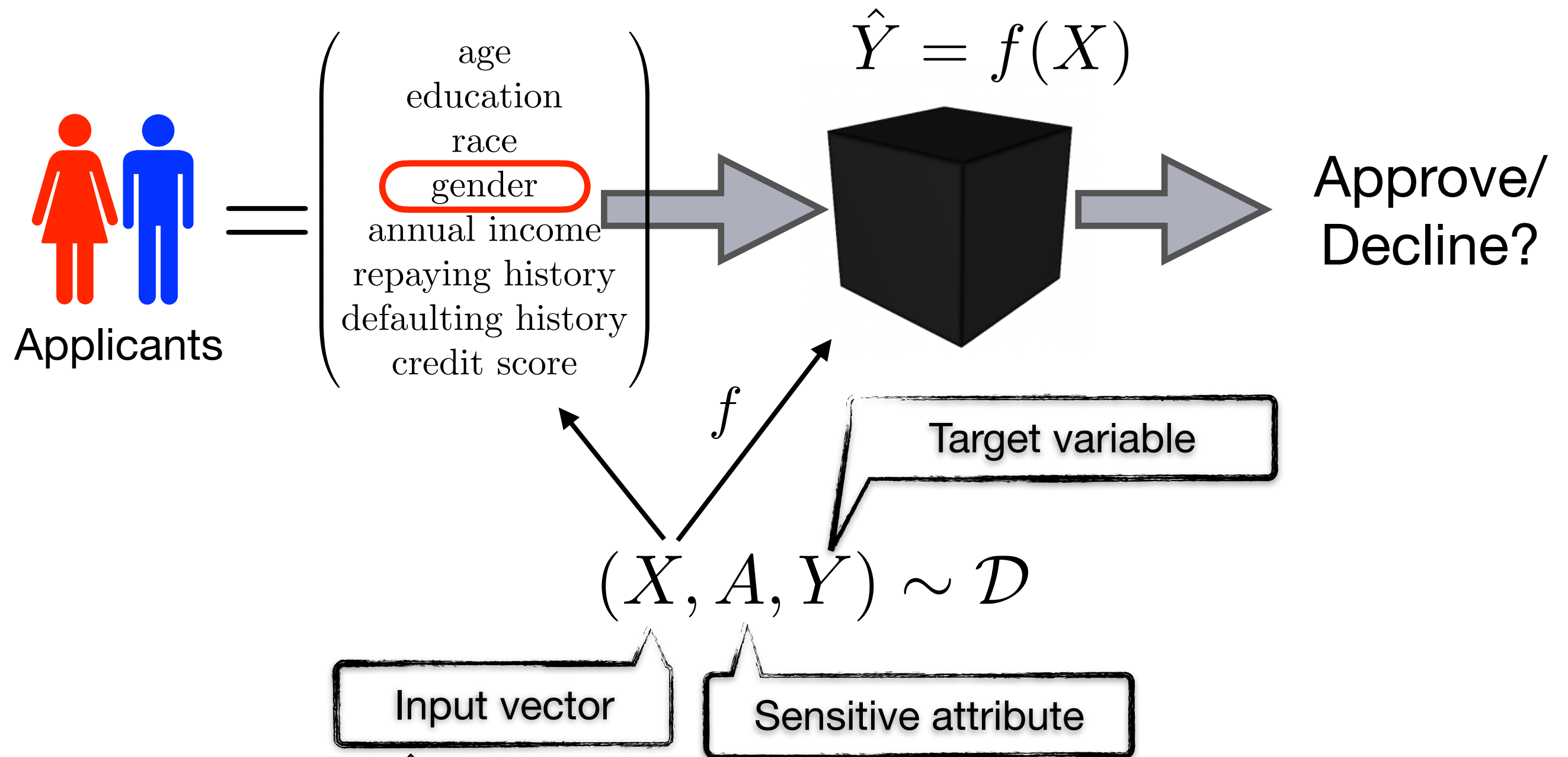


Statistical parity:  $\hat{Y} \perp A$

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

# Statistical Definition of Fairness

## Example in loan application



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

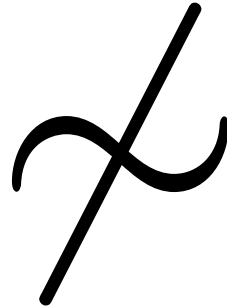
Equalized odds (Hardt et al. 16):  $\hat{Y} \perp A \mid Y$

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

# Incompatibility between Fairness Notions

---

Statistical Parity



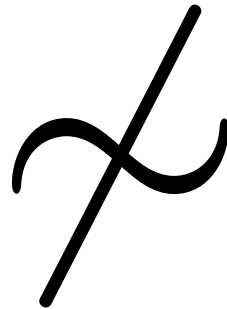
Equalized Odds



# Incompatibility between Fairness Notions

---

Statistical Parity

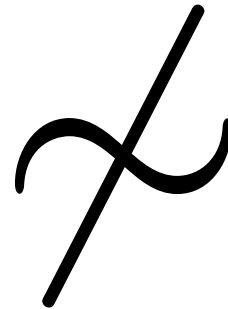


Equalized Odds

# Incompatibility between Fairness Notions

---

Statistical Parity



Equalized Odds

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

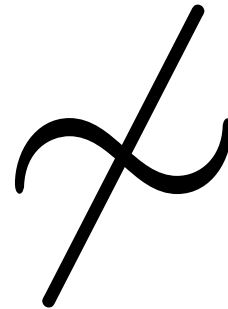
Theorem [ZG, NeurIPS 19]:

$$\varepsilon_{A=0}(h) + \varepsilon_{A=1}(h) \geq \Delta_{\text{BR}} \quad 6$$

# Incompatibility between Fairness Notions

---

Statistical Parity



Equalized Odds

Accuracy Parity

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

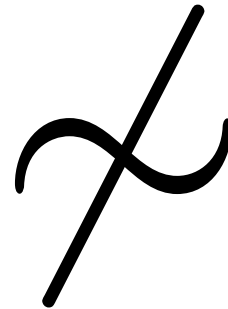
Theorem [ZG, NeurIPS 19]:

$$\varepsilon_{A=0}(h) + \varepsilon_{A=1}(h) \geq \Delta_{\text{BR}} \quad 6$$

# Incompatibility between Fairness Notions

---

Statistical Parity



Equalized Odds



Accuracy Parity

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

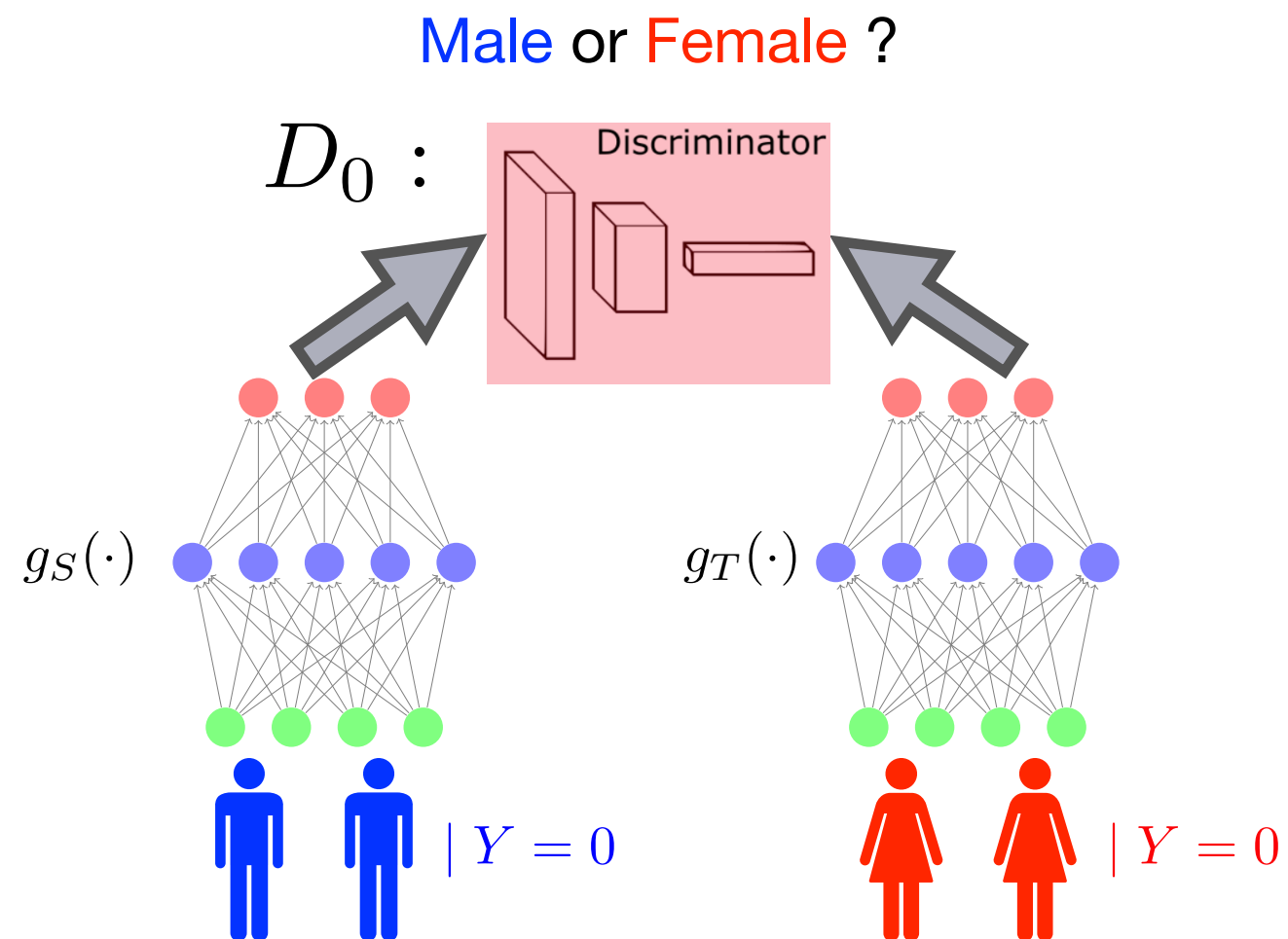
Theorem [ZG, NeurIPS 19]:

$$\varepsilon_{A=0}(h) + \varepsilon_{A=1}(h) \geq \Delta_{\text{BR}} \quad 6$$

# Conditional Learning of Fair Representations

---

# Conditional Learning of Fair Representations

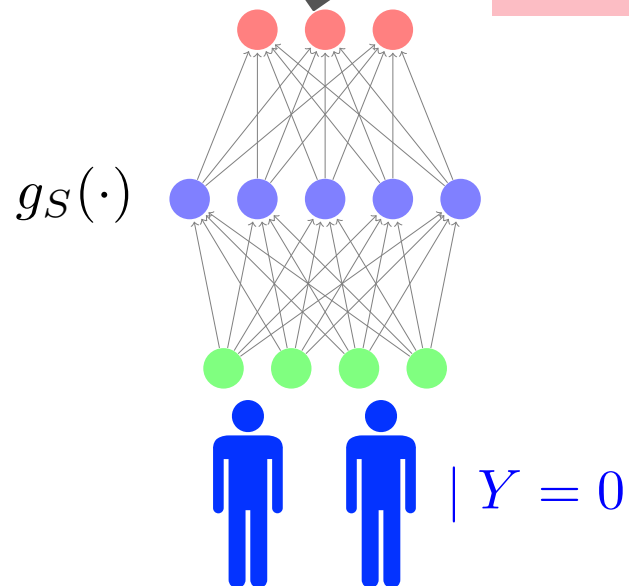
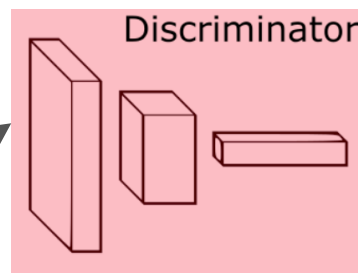




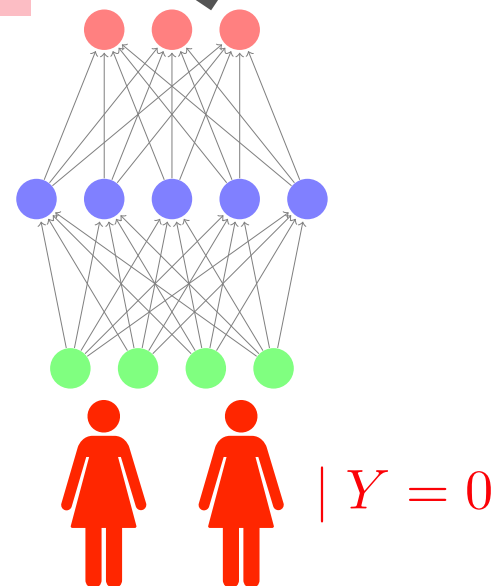
# Conditional Learning of Fair Representations

Male or Female ?

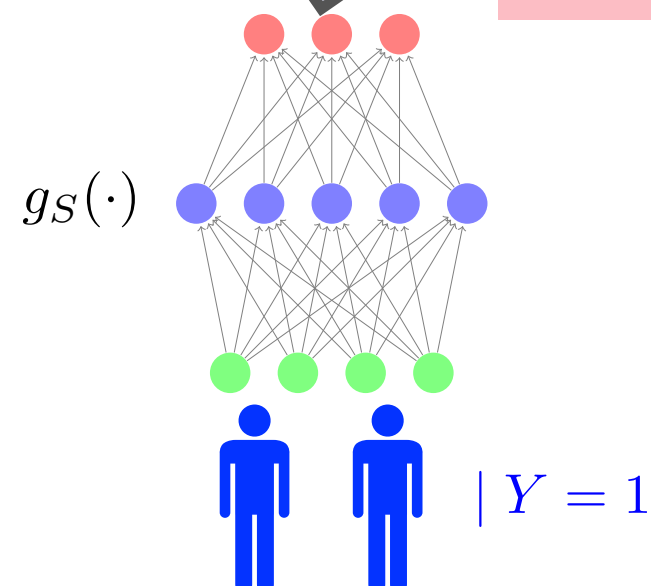
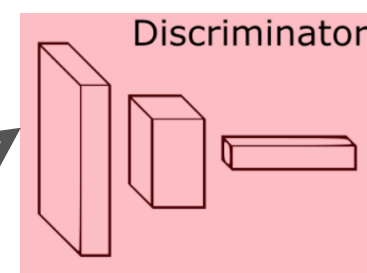
$D_0 :$



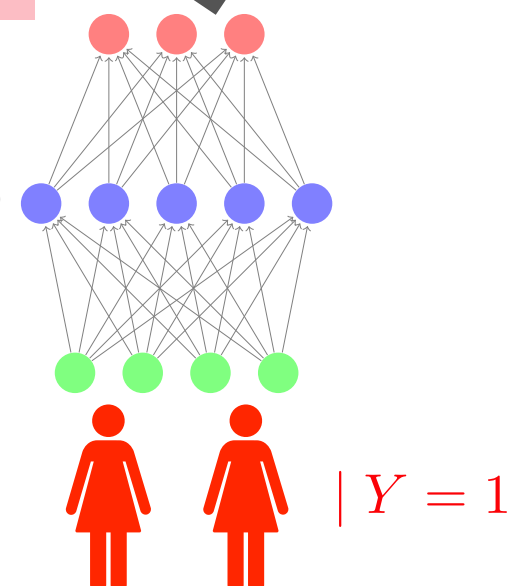
$g_T(\cdot)$



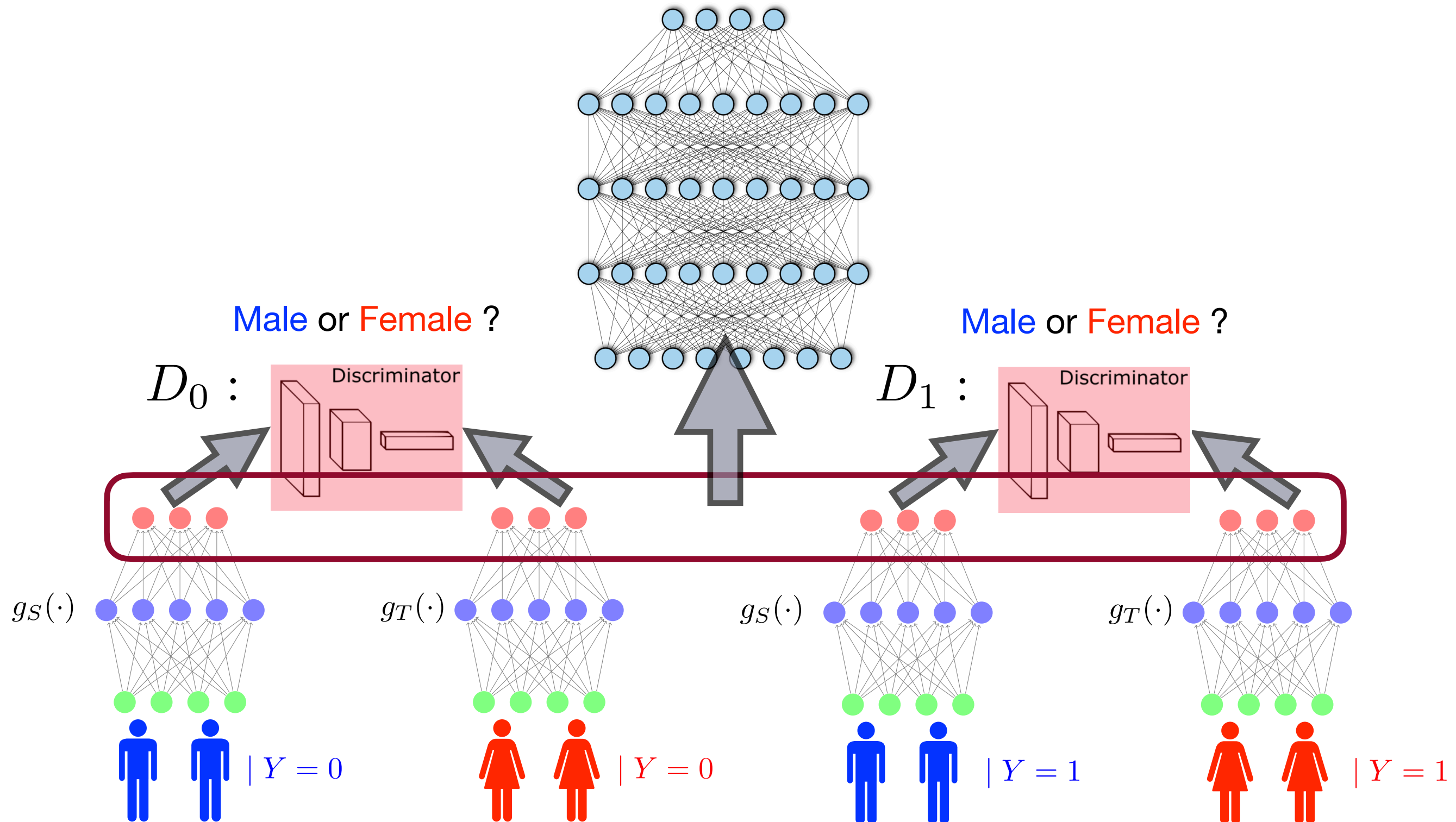
$D_1 :$



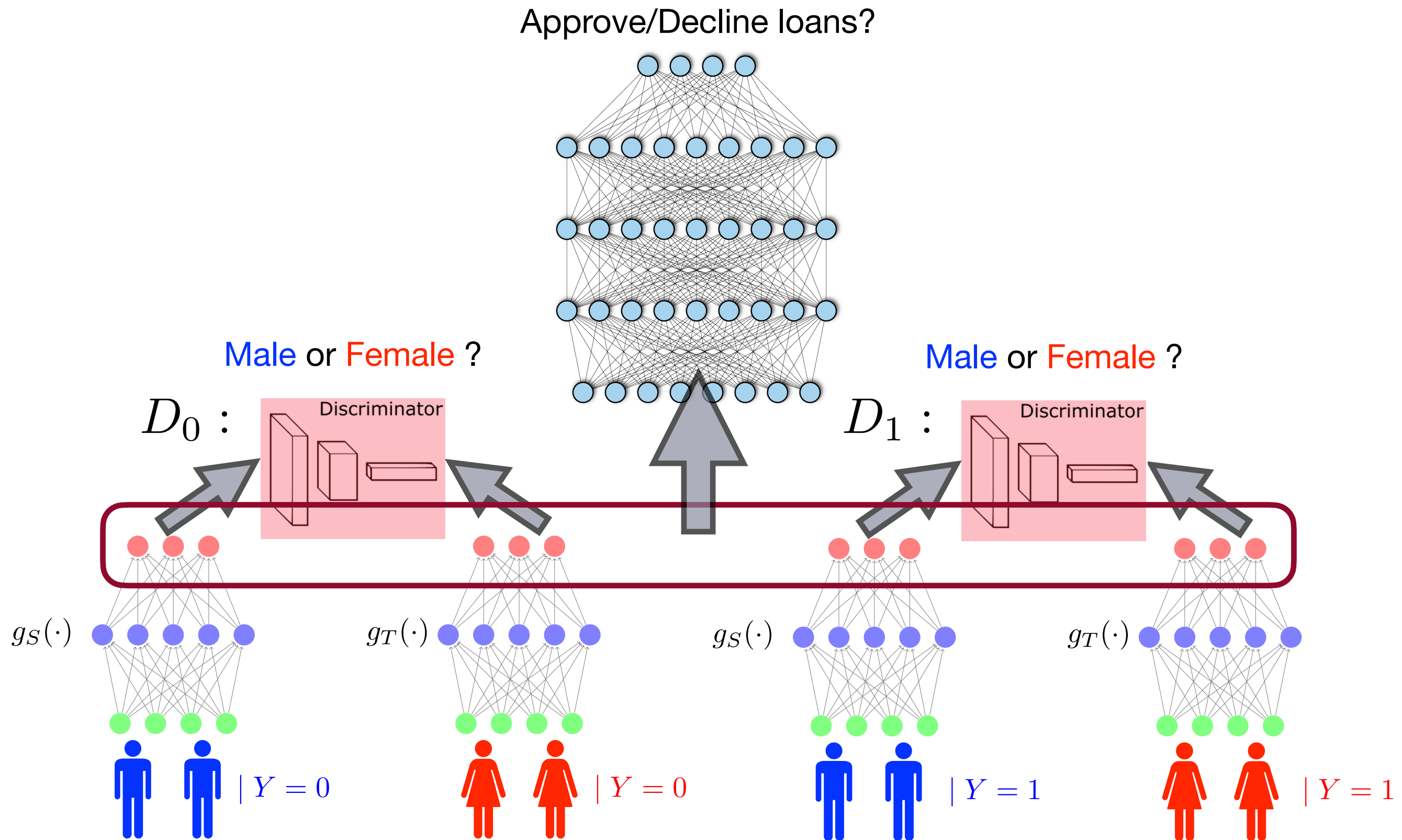
$g_T(\cdot)$



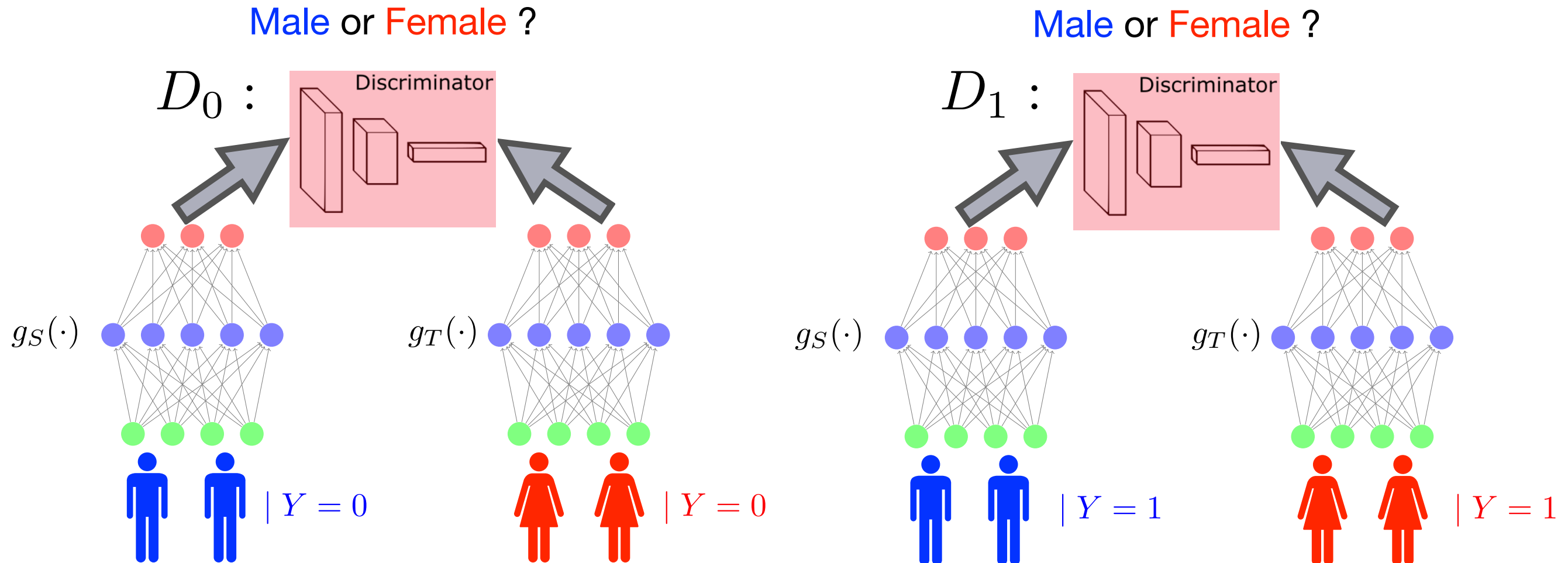
# Conditional Learning of Fair Representations



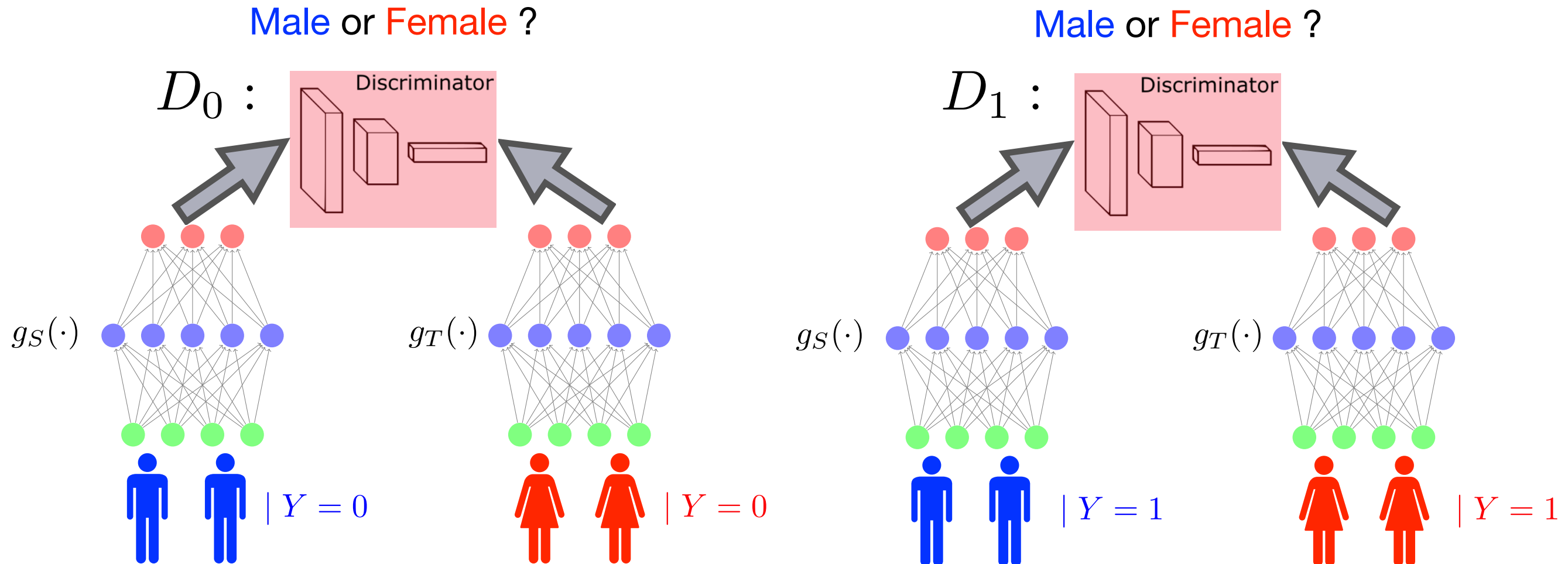
# Conditional Learning of Fair Representations



# Conditional Learning of Fair Representations



# Conditional Learning of Fair Representations

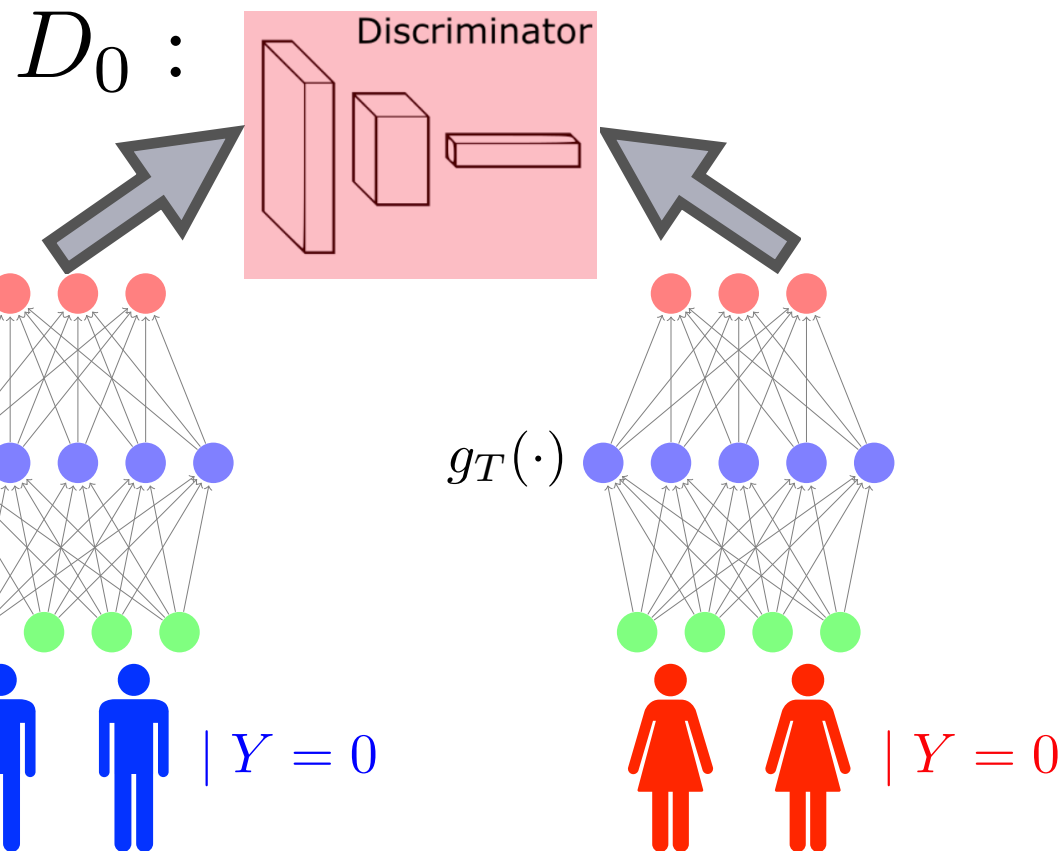


$$\Pr_{A=0}(\hat{Y} = 1 \mid Y = 0) \approx \Pr_{A=1}(\hat{Y} = 1 \mid Y = 0)$$

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

# Conditional Learning of Fair Representations

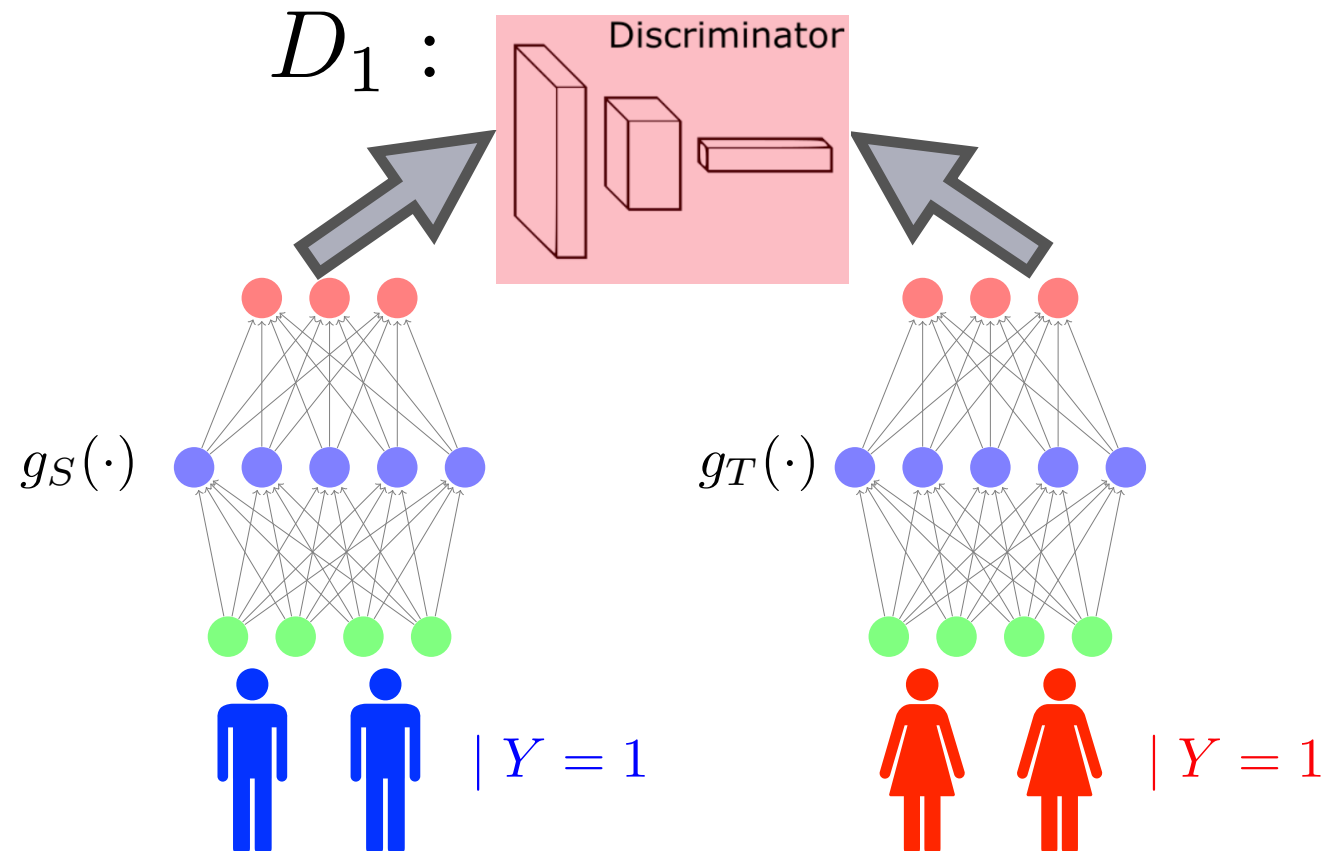
Male or Female ?



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

Equalized False Positive Rate (FPR)

Male or Female ?



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

Equalized False Negative Rate (FNR)



# An Error Decomposition Theorem

---

Approximately equal error rates across groups:

$$\begin{aligned} |\varepsilon_{A=0}(\hat{Y}) - \varepsilon_{A=1}(\hat{Y})| &\leq \Delta_{\text{BR}} \cdot (\text{FPR}(\hat{Y}) + \text{FNR}(\hat{Y})) \\ &\quad + 2 \max \left\{ d \left( \mathcal{D}_{A=0}^{Z|Y=0}, \mathcal{D}_{A=1}^{Z|Y=0} \right), d \left( \mathcal{D}_{A=0}^{Z|Y=1}, \mathcal{D}_{A=1}^{Z|Y=1} \right) \right\} \end{aligned}$$

$$\Delta_{\text{BR}} := |\Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1)| \quad 9$$

# An Error Decomposition Theorem

---

Approximately equal error rates across groups:

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

distance between  
marginal label distributions

# An Error Decomposition Theorem

---

Approximately equal error rates across groups:

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

distance between  
marginal label distributions

distance between  
conditional feature distributions

# An Error Decomposition Theorem

---

Approximately equal error rates across groups:

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

distance between  
marginal label distributions

distance between  
conditional feature distributions



# An Error Decomposition Theorem

---

Approximately equal error rates across groups:

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

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  $Y$

# An Error Decomposition Theorem

---

Approximately equal error rates across groups:

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

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  $Y$

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

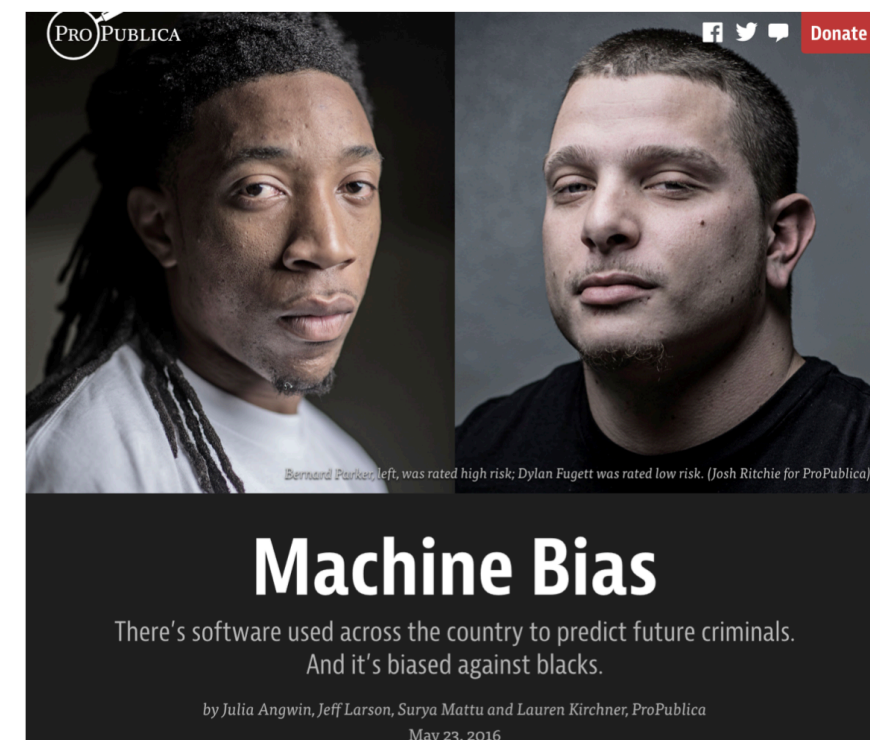
$$\Delta_{\text{DP}}(\hat{Y}) := \left| \Pr(\hat{Y} = 1 \mid A = 0) - \Pr(\hat{Y} = 1 \mid A = 1) \right|$$

$$\Delta_{\text{BR}} := \left| \Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1) \right| \quad 9$$

# Experiment: Recidivism Prediction

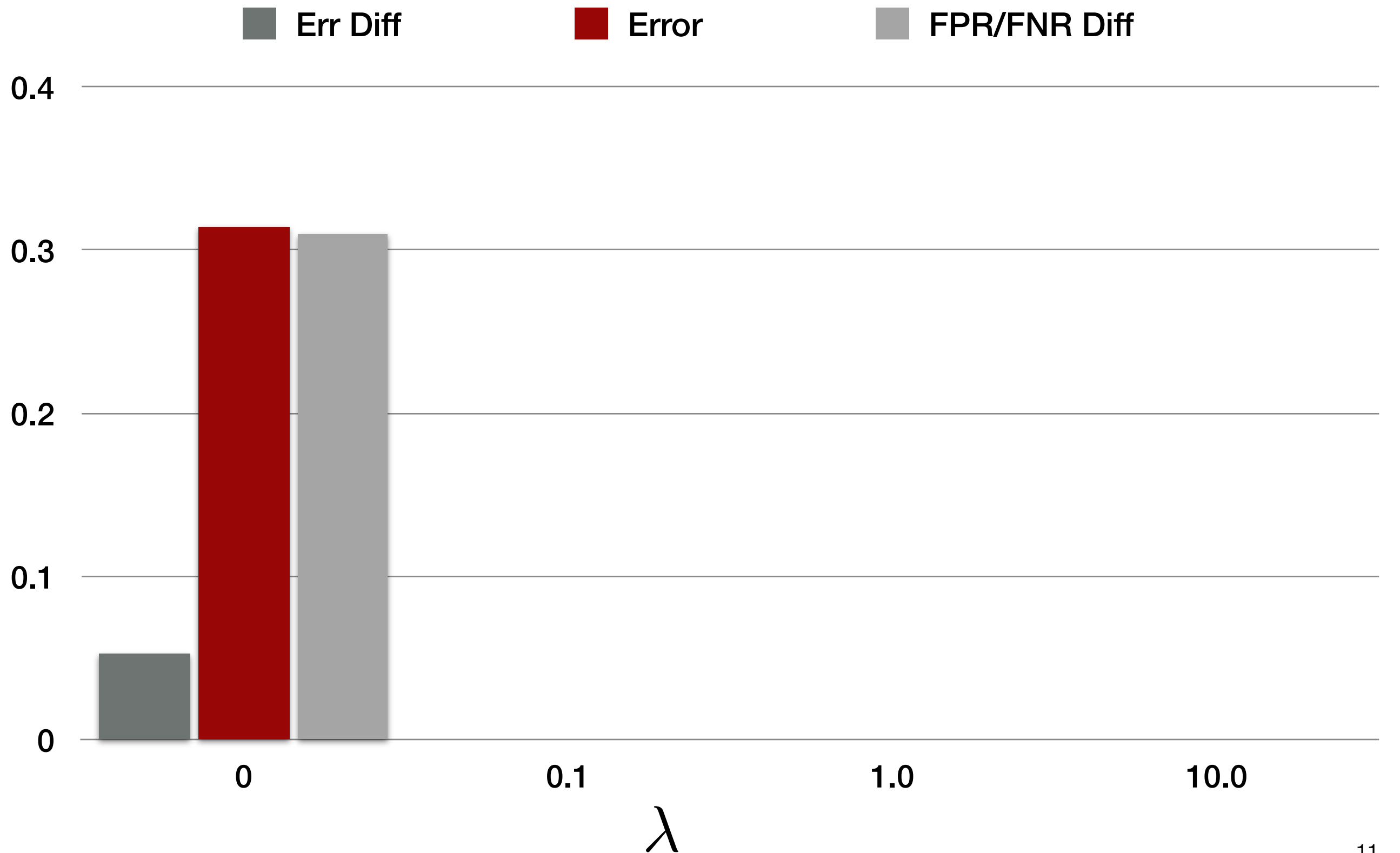
## COMPAS

- 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:  $\Delta_{BR} = 0.129$



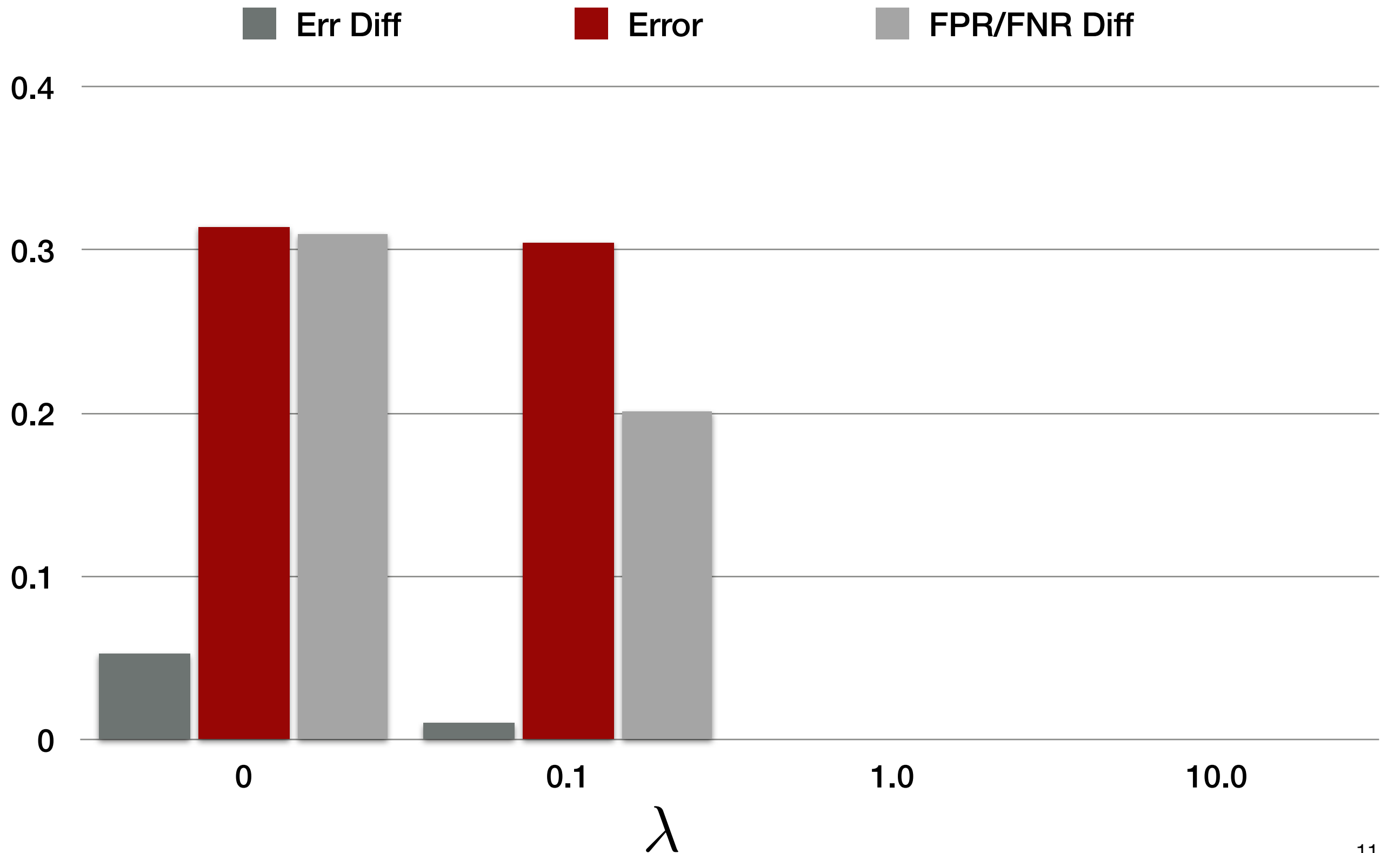
# Experiment: Recidivism Prediction

---



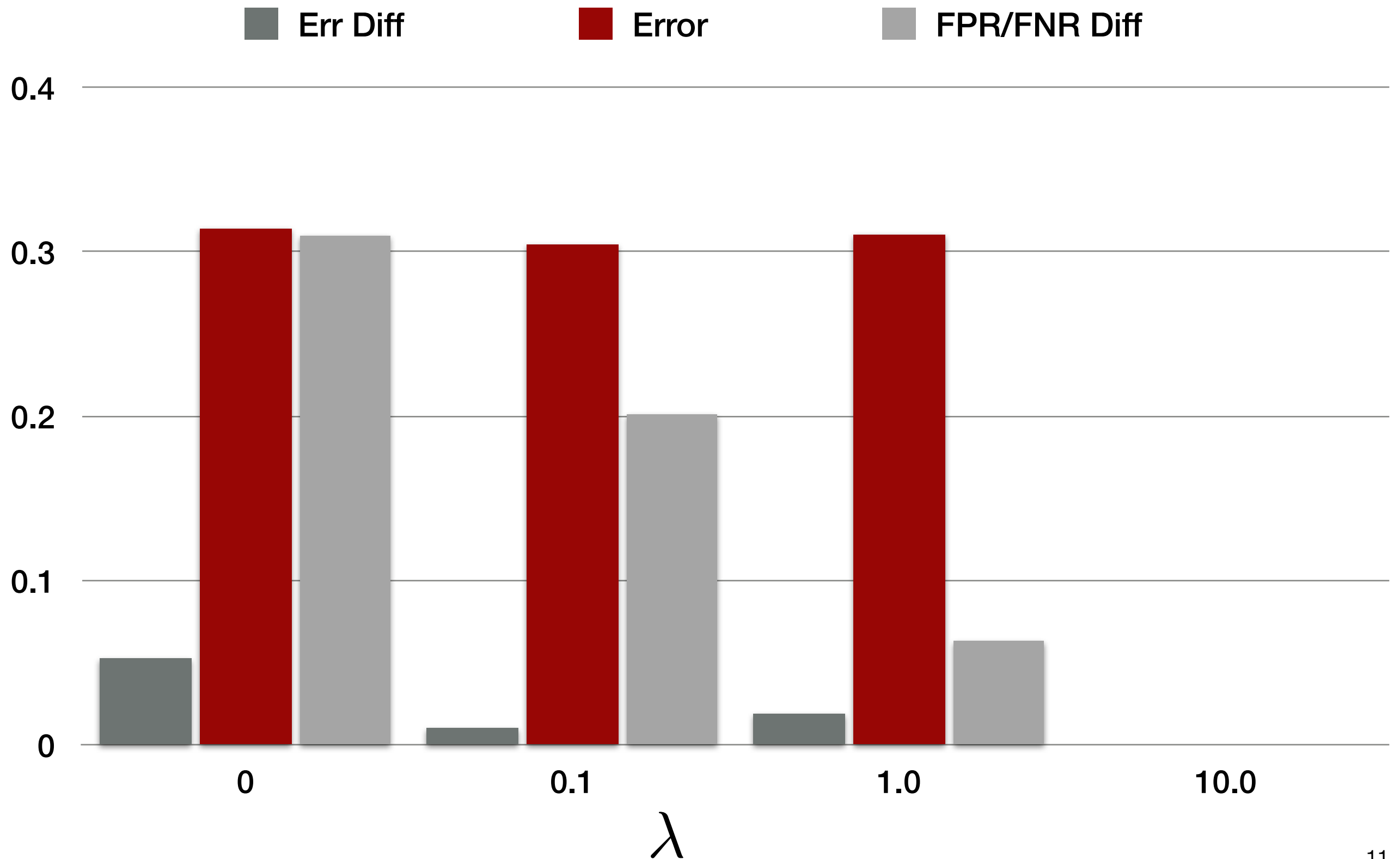
# Experiment: Recidivism Prediction

---



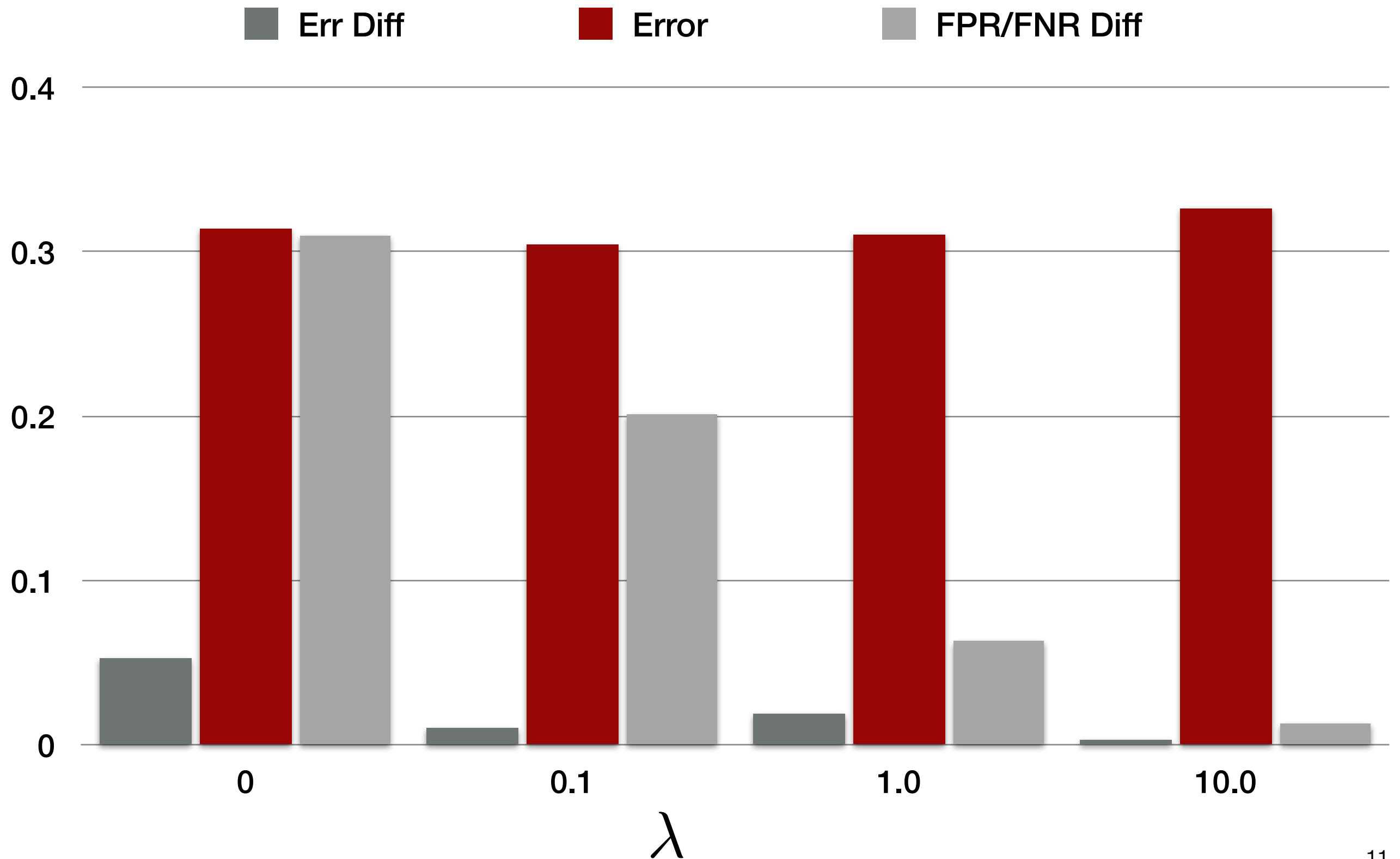
# Experiment: Recidivism Prediction

---





# 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

