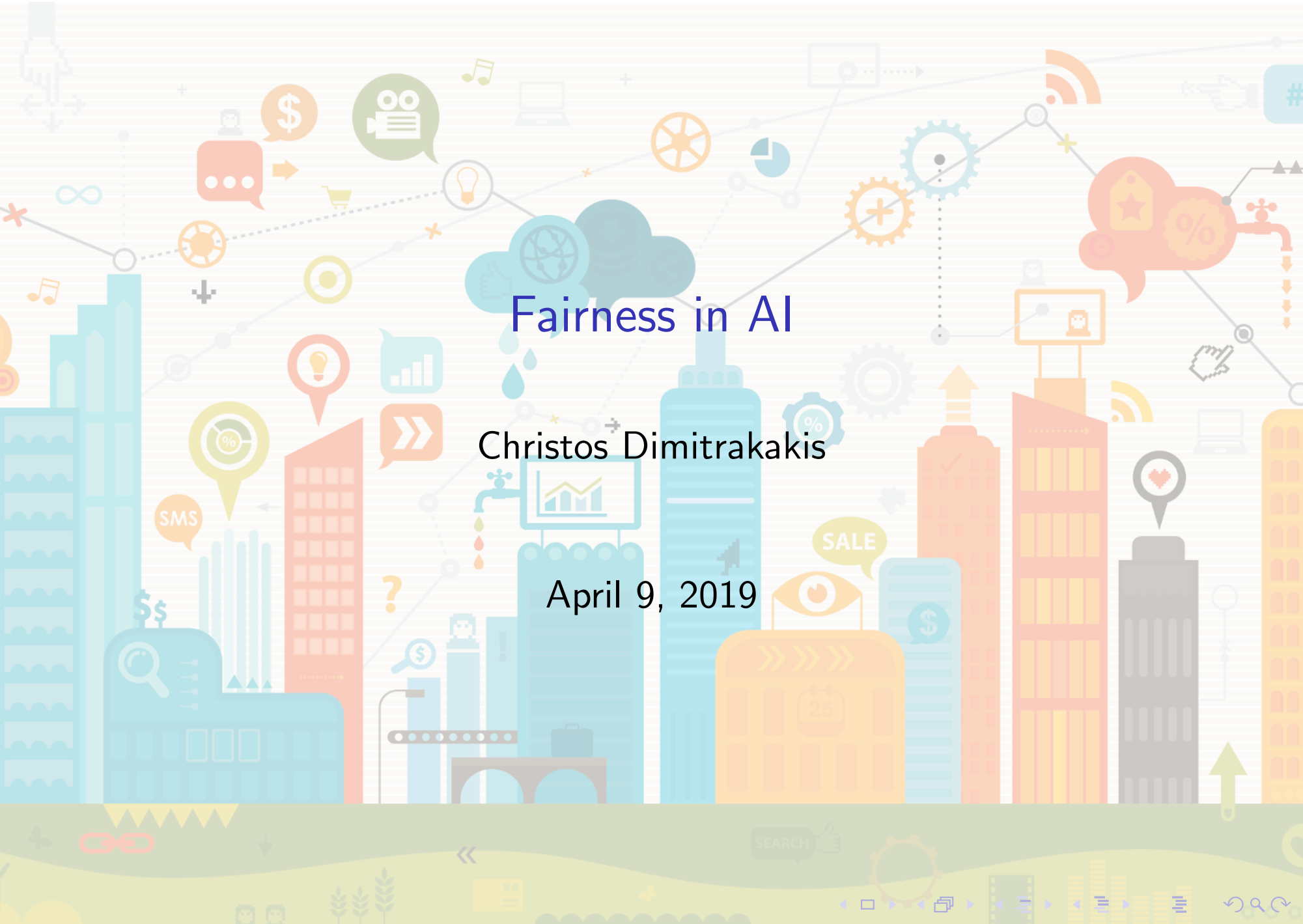
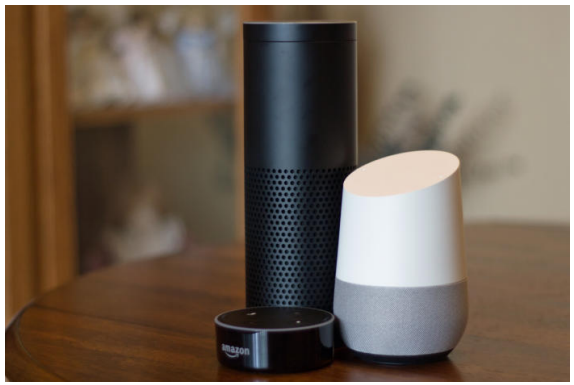


Fairness in AI

Christos Dimitrakakis

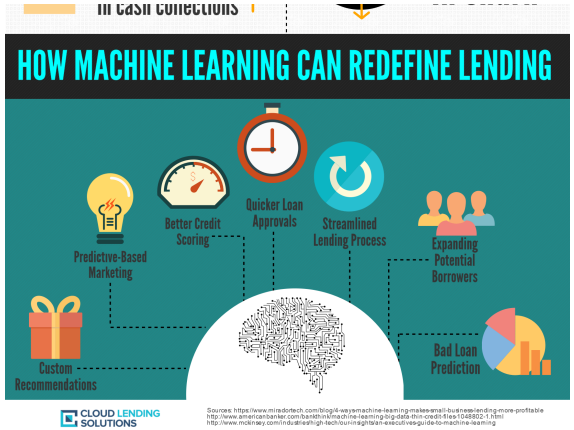
April 9, 2019

Pervasive "intelligent" systems



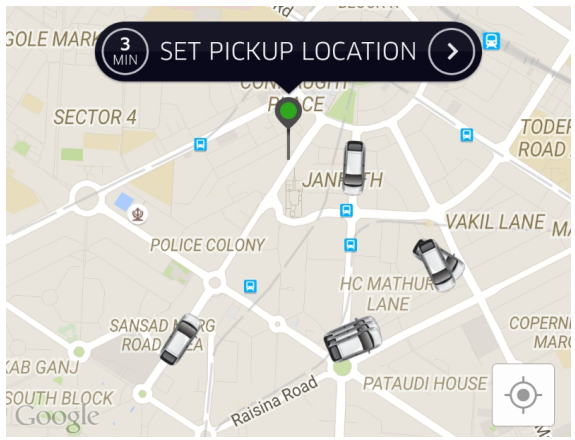
Home assistants

Web advertising



Lending

Autonomous vehicles

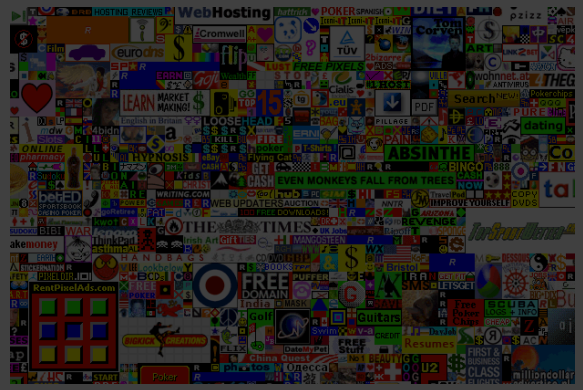


Ridesharing

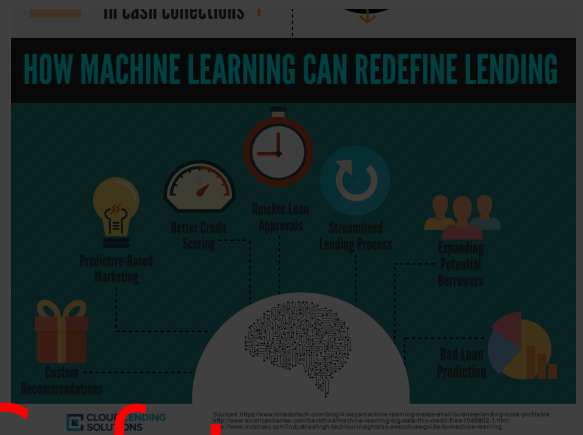
Public policy

Pervasive "intelligent" systems

Home assistants



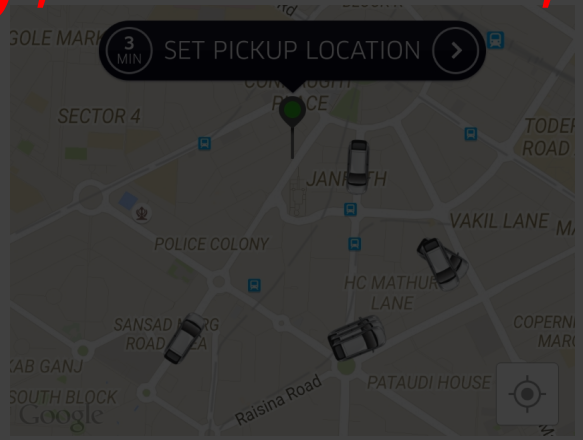
Web advertising



Lending

Privacy, Fairness, Safety

Autonomous vehicles



Ridesharing

Public policy

Fairness

Meritocracy

Meritocracy

Example 1 (College admissions)

- Student A has a grade 4/5 from Gota Highschool.
- Student B has a grade 5/5 from Vasa Highschool.

Meritocracy

Example 1 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Example 2 (Additional information)

- 70% of admitted Gota graduates with $4+$ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

Meritocracy

Example 1 (College admissions)

- Student A has a grade $4/5$ from Gota Highschool.
- Student B has a grade $5/5$ from Vasa Highschool.

Example 2 (Additional information)

- 70% of admitted Gota graduates with $4+$ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

We still don't know how a specific student will do!

Solutions

Meritocracy

Example 1 (College admissions)

- Student A has a grade 4/5 from Gota Highschool.
- Student B has a grade 5/5 from Vasa Highschool.

Example 2 (Additional information)

- 70% of admitted Gota graduates with 4+ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

We still don't know how a specific student will do!

Solutions

- Admit everybody?

Meritocracy

Example 1 (College admissions)

- Student A has a grade 4/5 from Gota Highschool.
- Student B has a grade 5/5 from Vasa Highschool.

Example 2 (Additional information)

- 70% of admitted Gota graduates with 4+ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

We still don't know how a specific student will do!

Solutions

- Admit everybody?
- Admit randomly?

Meritocracy

Example 1 (College admissions)

- Student A has a grade 4/5 from Gota Highschool.
- Student B has a grade 5/5 from Vasa Highschool.

Example 2 (Additional information)

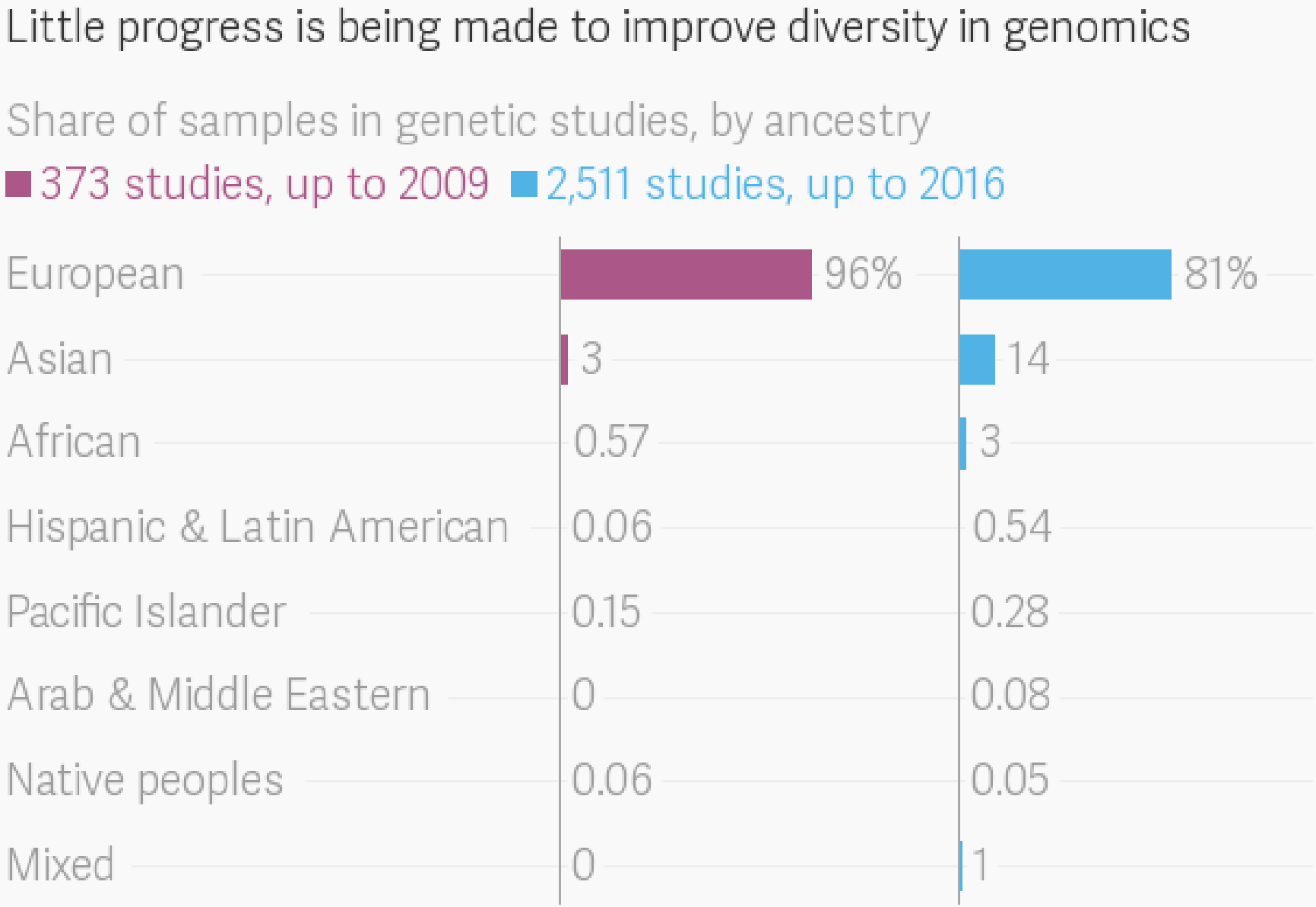
- 70% of admitted Gota graduates with 4+ get their degree.
- 50% of admitted Vasa graduates with 5 get their degree.

We still don't know how a specific student will do!

Solutions

- Admit everybody?
- Admit randomly?
- Use prediction of individual academic performance?

Proportional representation

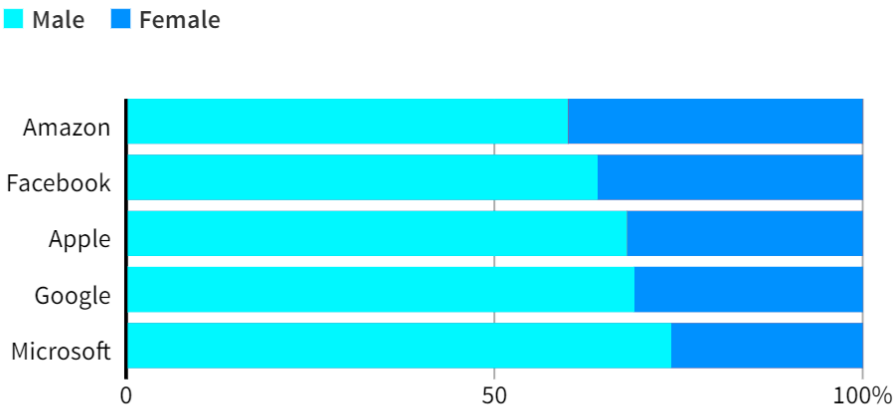


Hiring decisions

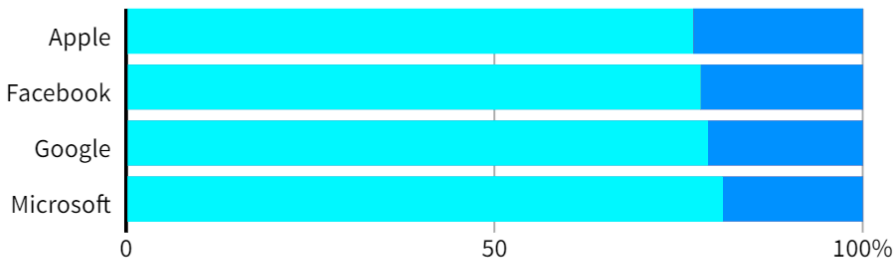
Dominated by men

Top U.S. tech companies have yet to close the gender gap in hiring, a disparity most pronounced among technical staff such as software developers where men far outnumber women. Amazon's experimental recruiting engine followed the same pattern, learning to penalize resumes including the word "women's" until the company discovered the problem.

GLOBAL HEADCOUNT



EMPLOYEES IN TECHNICAL ROLES



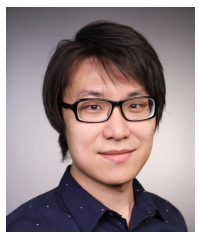
Fairness and information

Example 3 (College admissions data)

School	Male	Female
A	62%	82%
B	63%	68%
C	37%	34%
D	33%	35%
E	28%	24%
F	6%	7%
Average	*45%*	*38%*

Bayesian Fairness[1]

Fairness under partial information.


[1] *C. Dimitrakakis, Y. Liu, G. Radanovic, D. C. Parkes*, AAAI 2019.

Bail decisions

x, z

$\searrow \pi$

- x - features
- z - sensitive features
- a - action
- y - outcome

Bail decisions

x, z

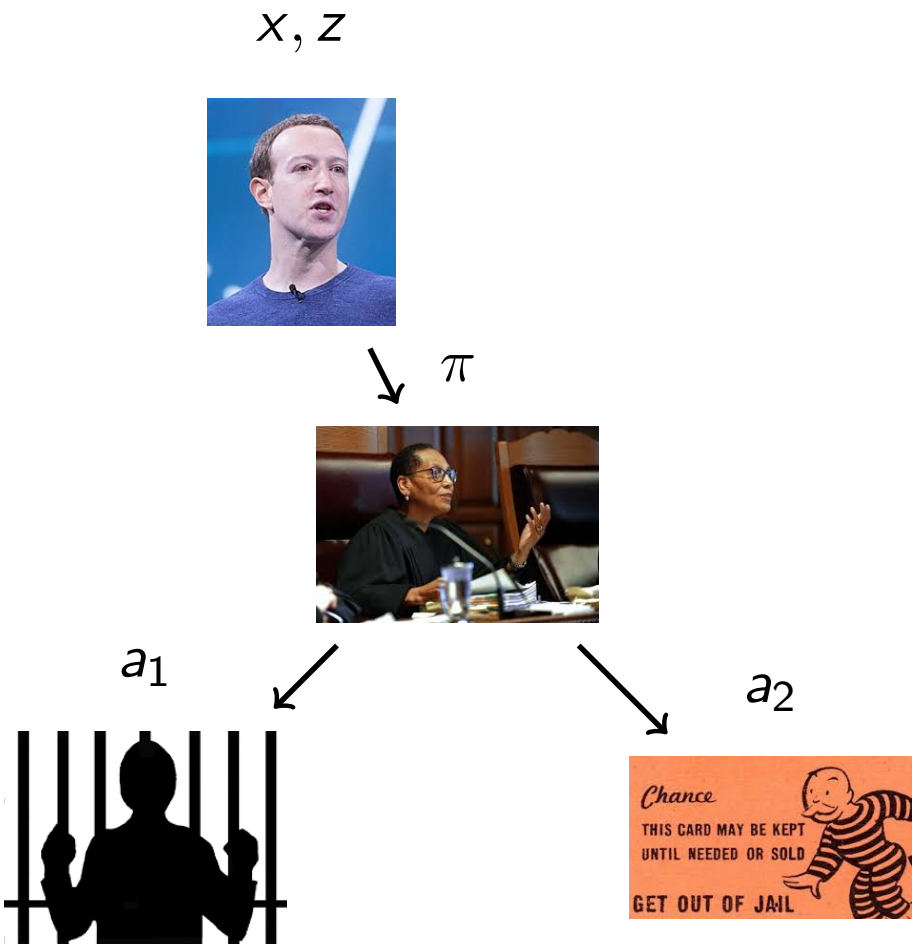
π

a_1

- x - features
- z - sensitive features
- a - action
- y - outcome

$$\pi(a \mid x) \qquad \text{(policy)}$$

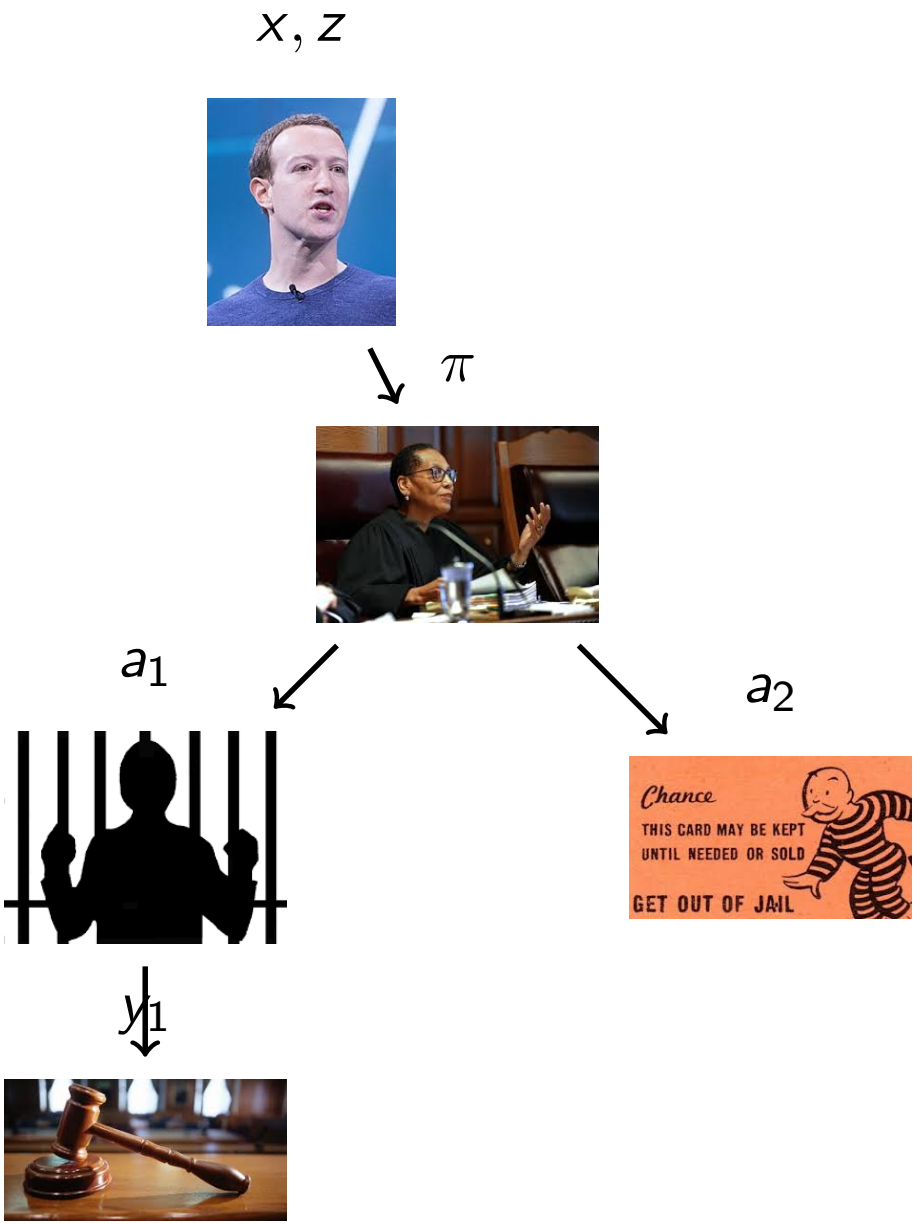
Bail decisions



- x - features
- z - sensitive features
- a - action
- y - outcome

$$\pi(a \mid x) \qquad \text{(policy)}$$

Bail decisions

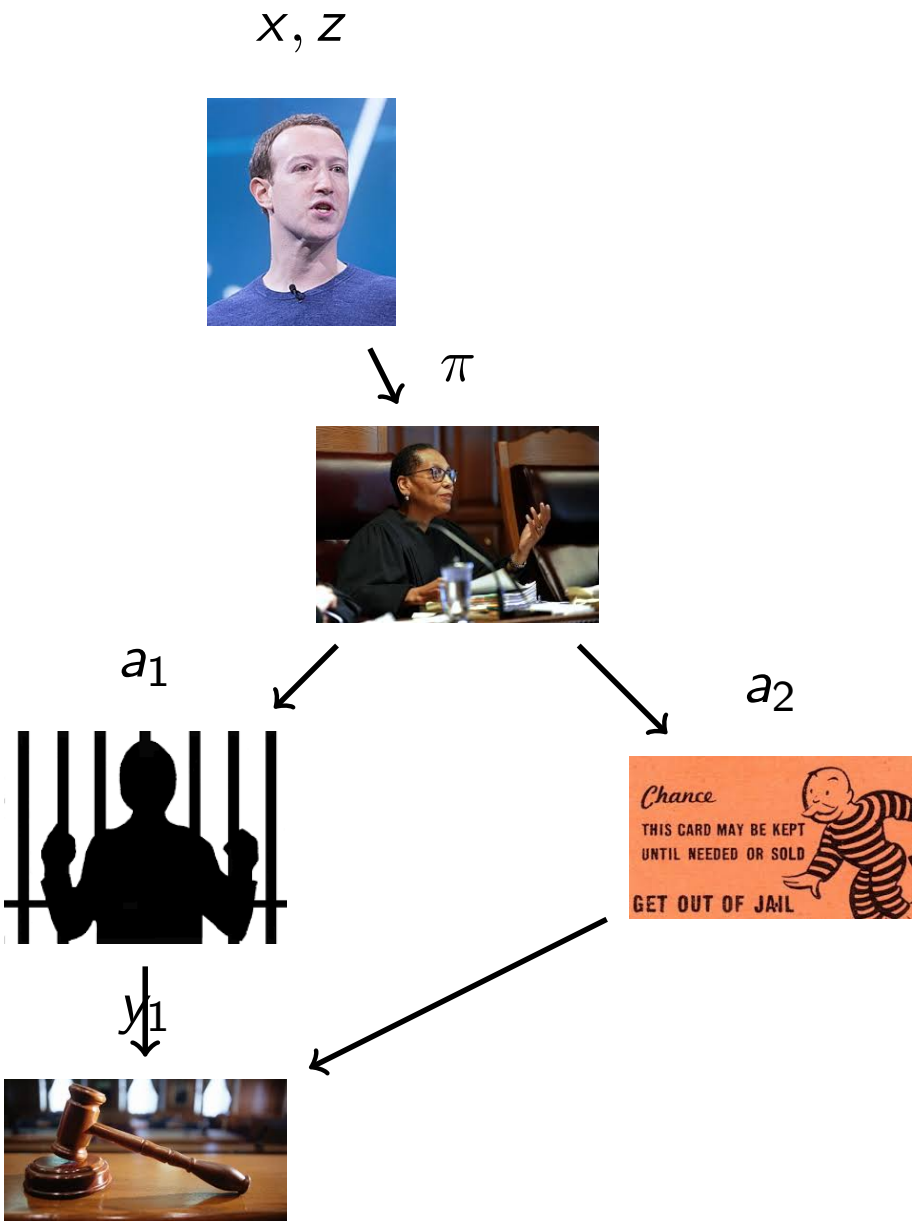


- x - features
- z - sensitive features
- a - action
- y - outcome

$$\pi(a \mid x) \qquad \text{(policy)}$$

$$P_\theta(x, y, z) \qquad \text{(model)}$$

Bail decisions

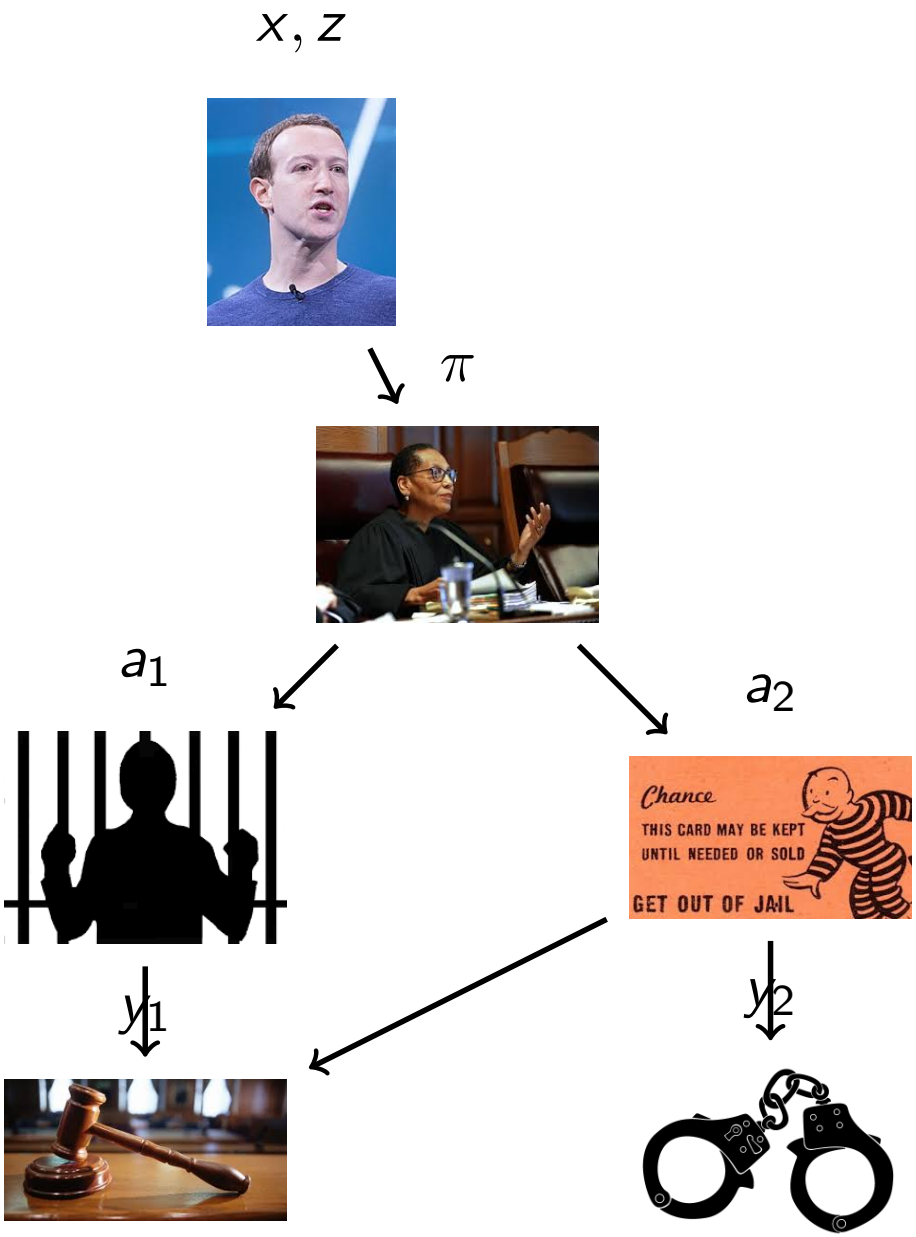


- x - features
- z - sensitive features
- a - action
- y - outcome

$$\pi(a \mid x) \qquad \text{(policy)}$$

$$P_\theta(x, y, z) \qquad \text{(model)}$$

Bail decisions

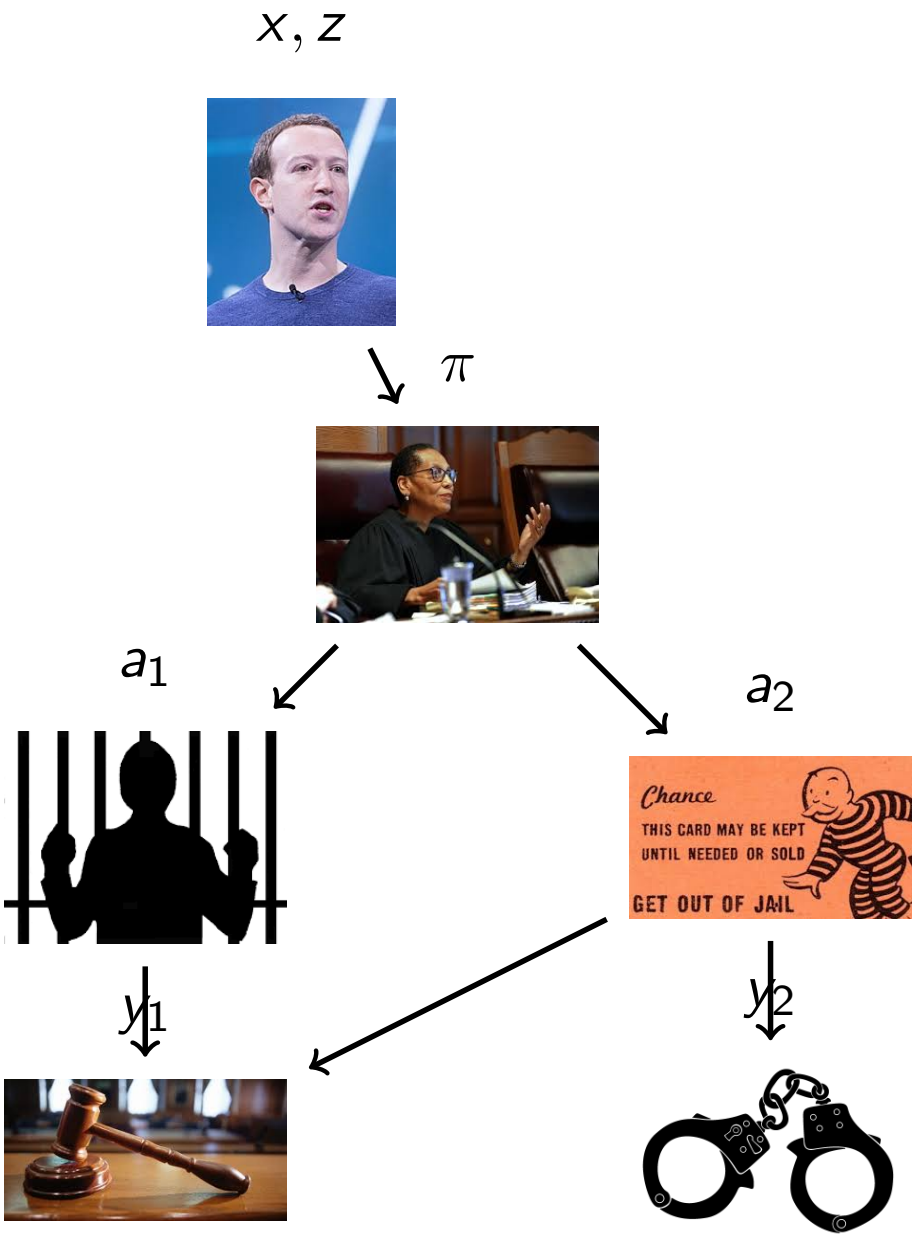


- x - features
- z - sensitive features
- a - action
- y - outcome

$$\pi(a \mid x) \qquad \text{(policy)}$$

$$P_\theta(x, y, z) \qquad \text{(model)}$$

Bail decisions



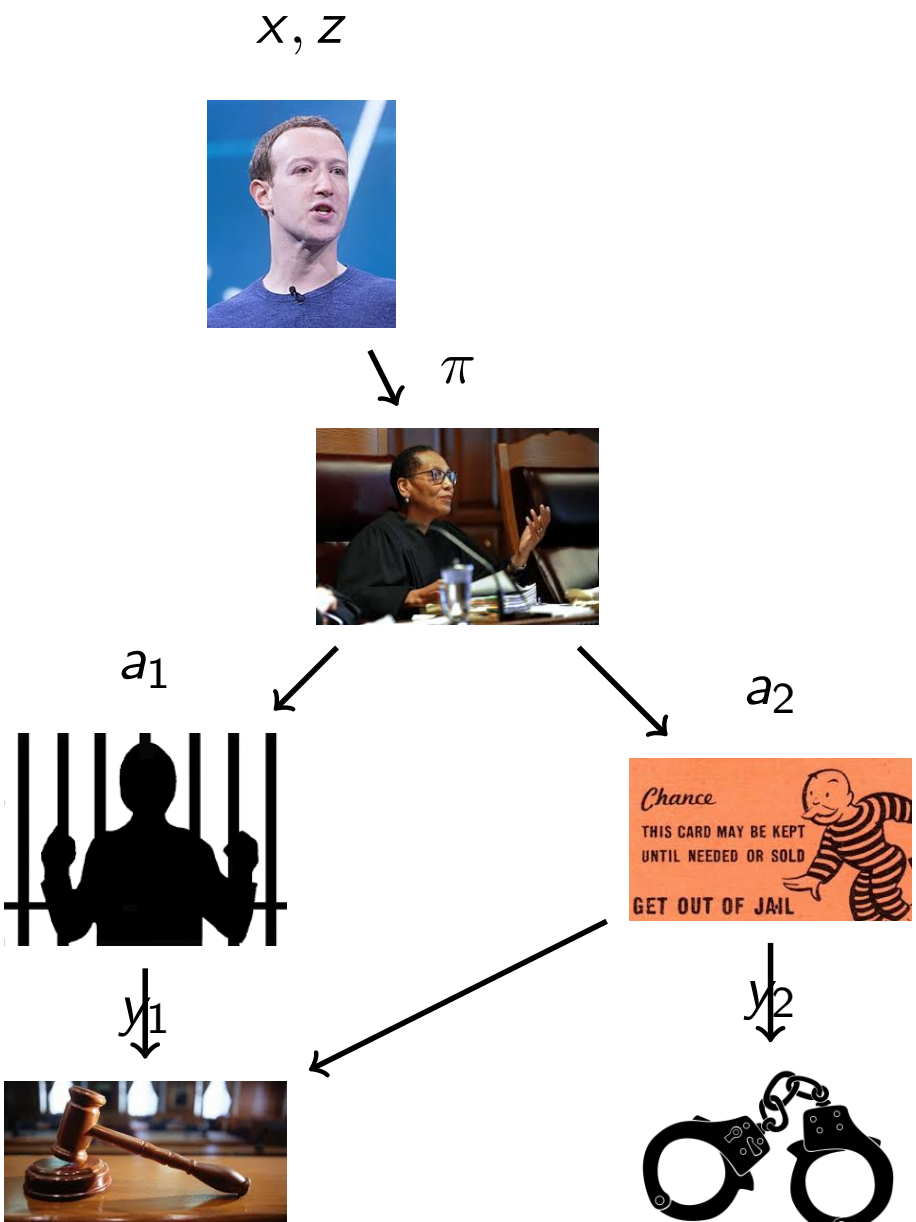
- x - features
- z - sensitive features
- a - action
- y - outcome

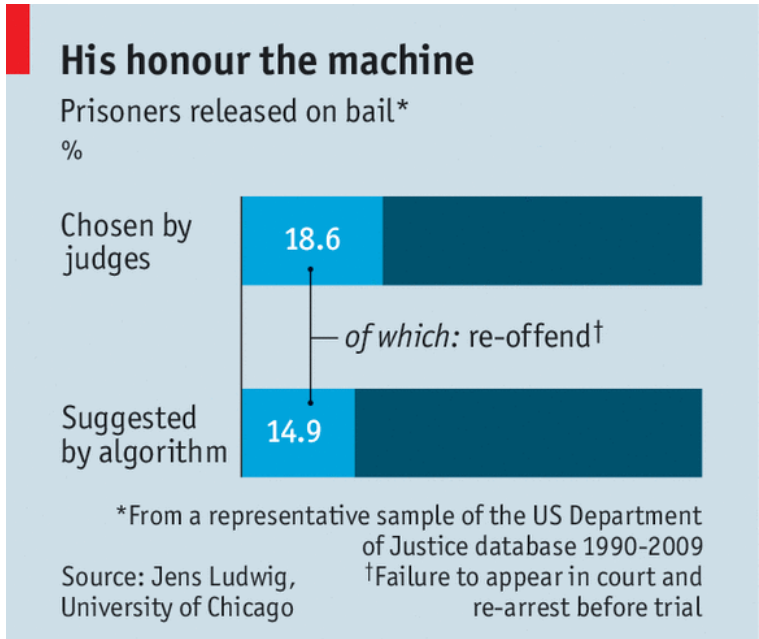
$$\pi(a \mid x) \qquad \text{(policy)}$$

$$P_\theta(x, y, z) \qquad \text{(model)}$$

$$U(a, y) \qquad \text{(utility)}$$

Bail decisions





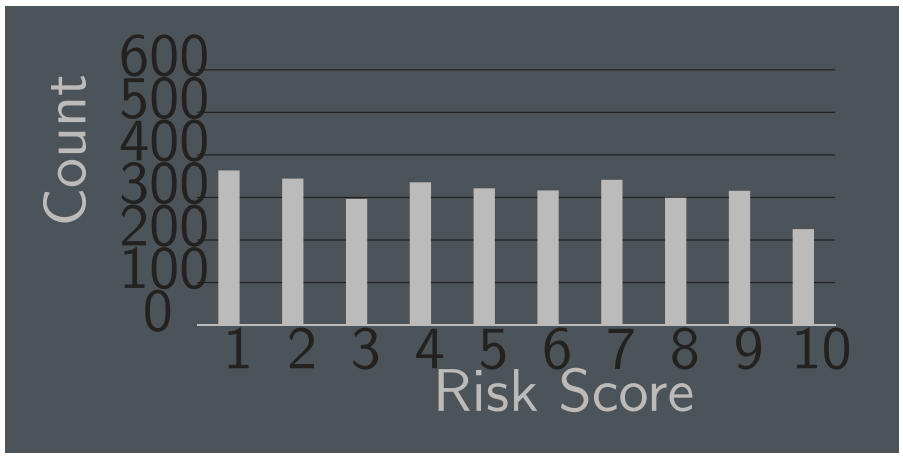
$\pi(a \mid x)$ (policy)

$P_\theta(x, y, z)$ (model)

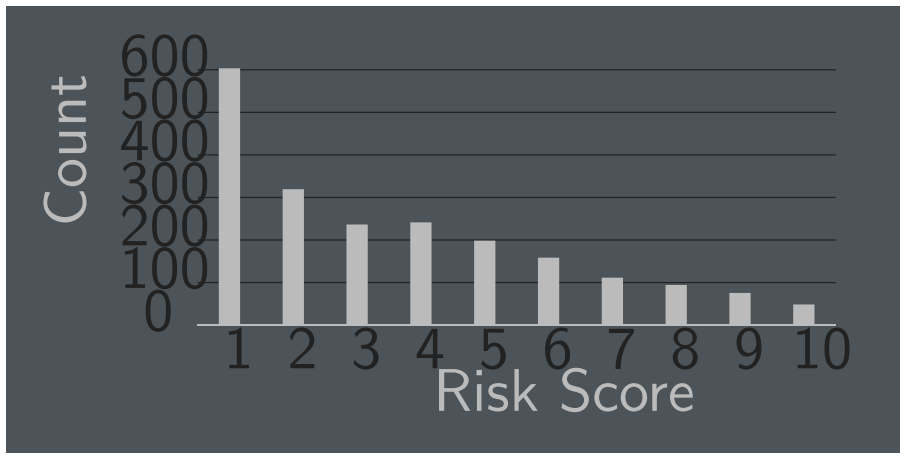
$U(a, y)$ (utility)

The USA COMPAS System[2]

Assigns a risk score a_t to defendant x_t.



Black

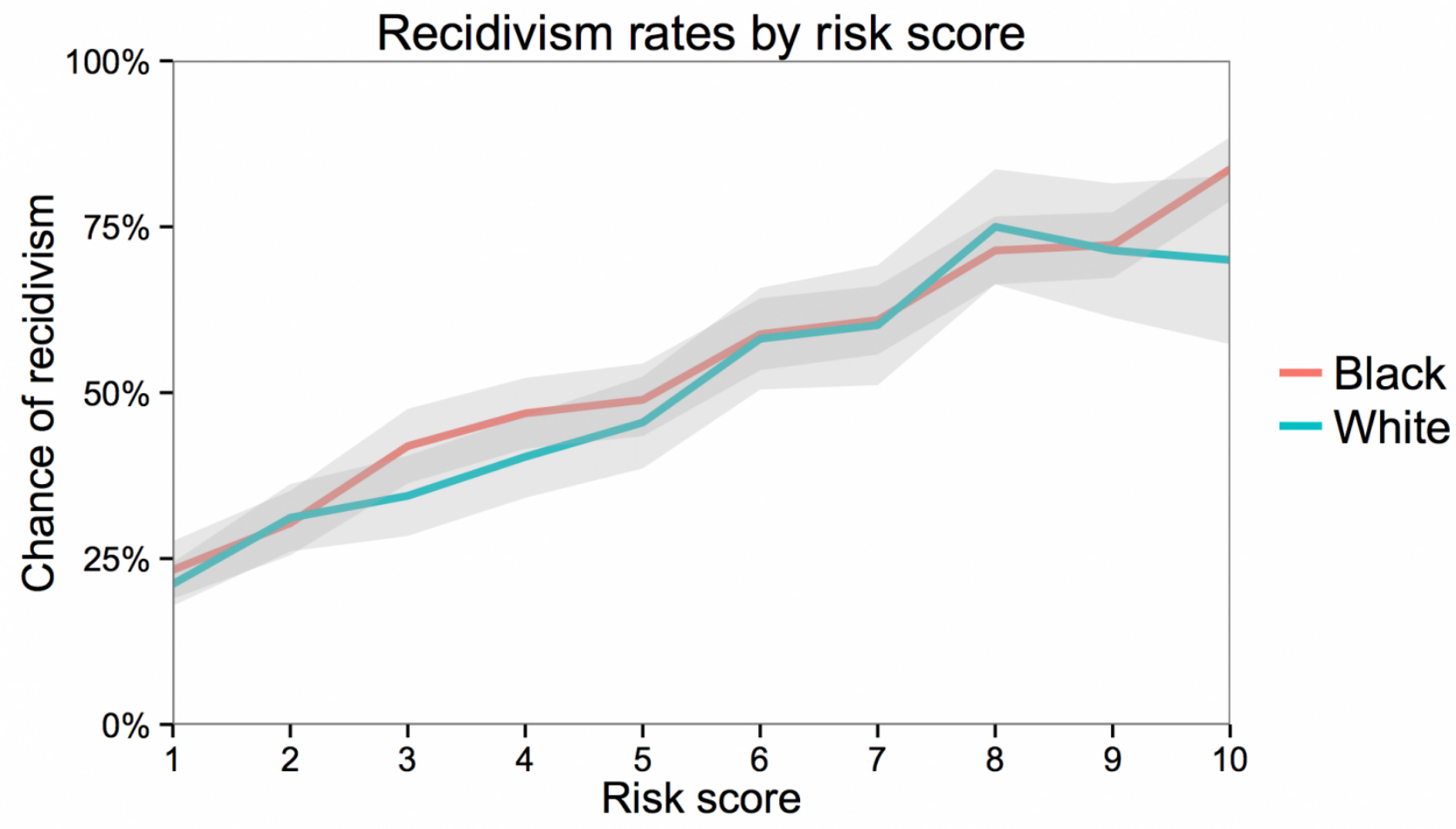


White

Figure : Apparent bias in risk scores towards black versus white defendants.

$$\mathbb{P}^{\pi}_{\theta}(a \mid z) = \mathbb{P}^{\pi}_{\theta}(a) \qquad \text{(equal treatment)}$$

[2] Pro-publica, 2016

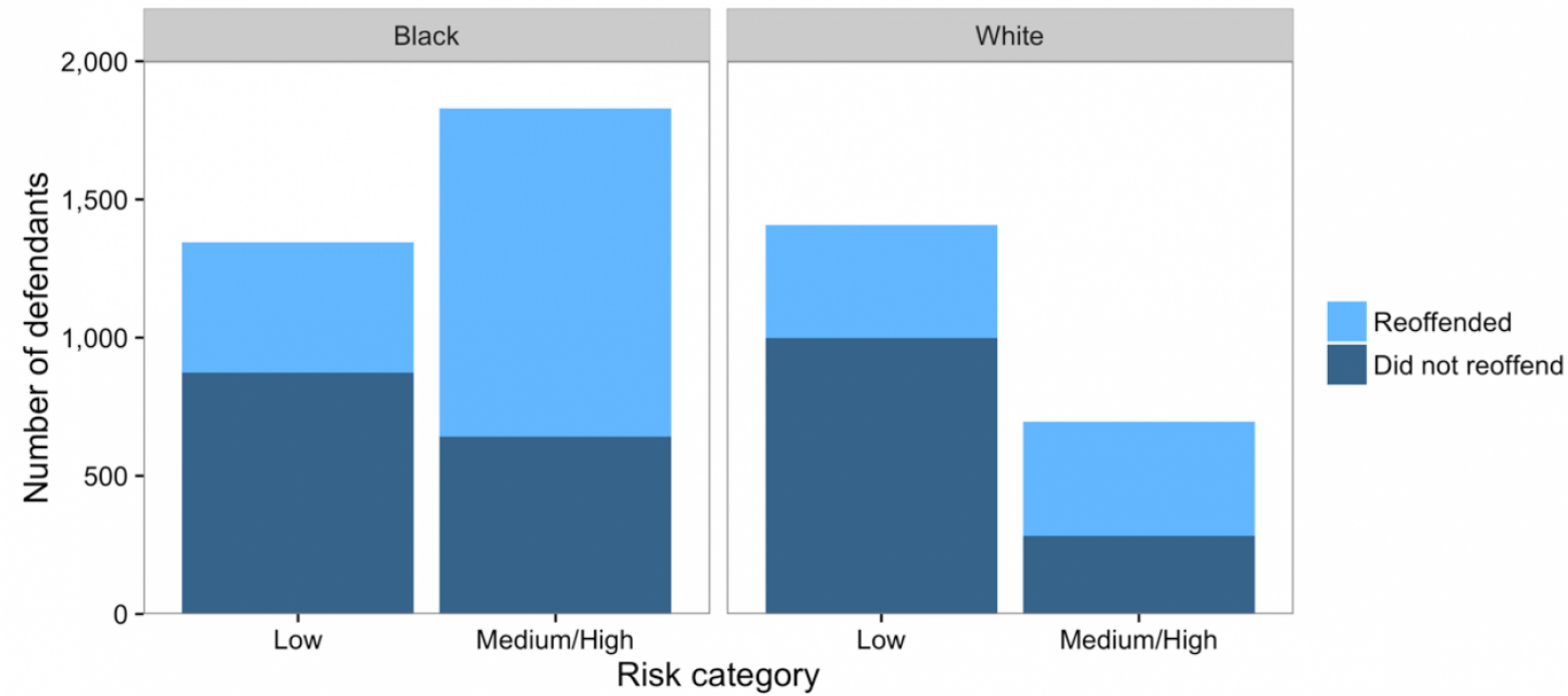


Washington Post, 2016

- y Result.
- a Assigned score.
- z Race.

$$\mathbb{P}^{\pi}_{\theta}(y \mid a, z) = \mathbb{P}^{\pi}_{\theta}(y \mid a) \qquad \text{(calibration)}$$

$$\mathbb{P}^{\pi}_{\theta}(a \mid y, z) = \mathbb{P}^{\pi}_{\theta}(a \mid y) \qquad \text{(balance)}$$



Pro-publica, 2016

- y Result.
- a Assigned score.
- z Race.

$$\mathbb{P}^{\pi}_{\theta}(y \mid a, z) = \mathbb{P}^{\pi}_{\theta}(y \mid a) \qquad \text{(calibration)}$$

$$\mathbb{P}^{\pi}_{\theta}(a \mid y, z) = \mathbb{P}^{\pi}_{\theta}(a \mid y) \qquad \text{(balance)}$$

The value of a policy

Fairness metrics: balance

$$F_{\text{balance}}(\theta, \pi) \triangleq \sum_{y,z,a} |\, \mathbb{P}^{\pi}_{\theta}(a \mid y, z) - \mathbb{P}^{\pi}_{\theta}(a \mid y)|^2 \tag{1.1}$$

The value of a policy

Fairness metrics: balance

$$F_{\text{balance}}(\theta, \pi) \triangleq \sum_{y,z,a} |\, \mathbb{P}_\theta^\pi(a \mid y, z) - \mathbb{P}_\theta^\pi(a \mid y)|^2 \tag{1.1}$$

Utility: Classification accuracy

$$U(\theta, \pi) = \mathbb{P}_\theta^\pi(y_t = a_t)$$

The value of a policy

Fairness metrics: balance

$$F_{\text{balance}}(\theta, \pi) \triangleq \sum_{y,z,a} |\, \mathbb{P}_\theta^\pi(a \mid y, z) - \mathbb{P}_\theta^\pi(a \mid y)|^2 \tag{1.1}$$

Utility: Classification accuracy

$$U(\theta, \pi) = \mathbb{P}_\theta^\pi(y_t = a_t)$$

Use λ to trade-off utility and fairness

$$V(\lambda, \theta, \pi) = (1-\lambda) \overbrace{U(\theta, \pi)}^{\text{utility}} - \lambda \underbrace{F(\theta, \pi)}_{\text{unfairness}} \tag{1.2}$$

Model uncertainty

θ is unknown

Theorem 4

A decision rule in the form of a lottery, i.e.

$$\pi(a \mid x) = p_a$$

can be the only way to satisfy balance for all possible θ*.*

Possible solutions

- Marginalize over θ ("expected" model)
- Use Bayesian reasoning

The setting

Bayesian modelling of uncertainty

- Model family $\{P_\theta(x, y, z) \mid \theta \in \Theta\}$
- Prior belief $\xi_0(\theta)$.
- Data $D \sim P_\theta(x, y, z)$.
- Posterior belief $\xi(\theta) \triangleq \xi_0(\theta \mid D)$.

The setting

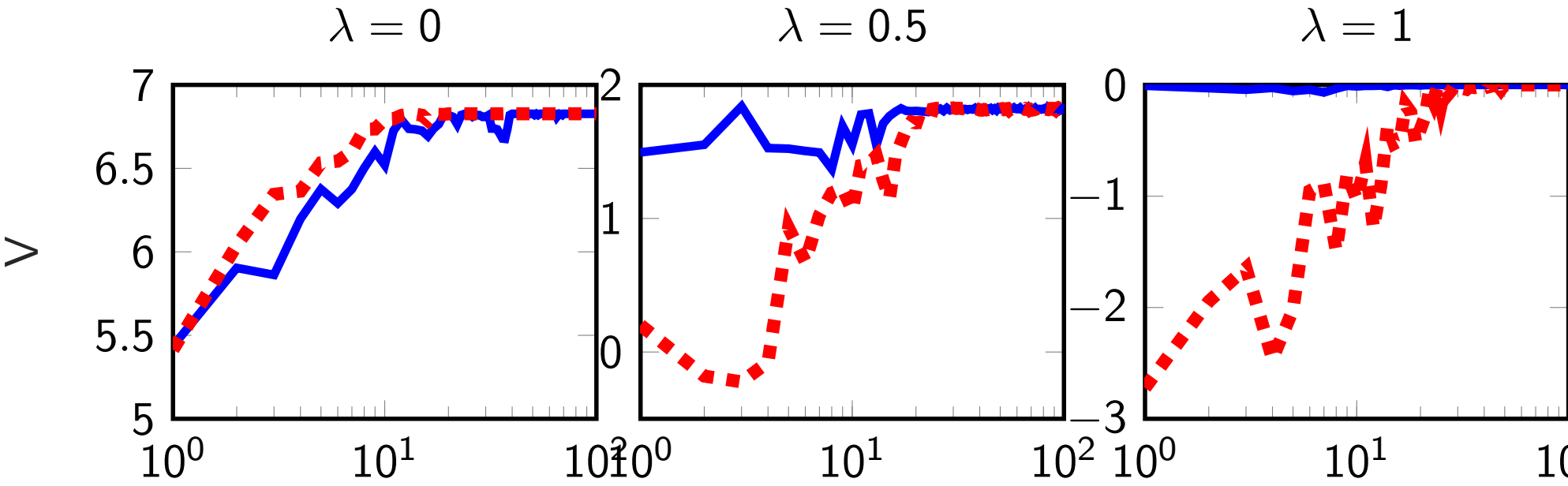
Bayesian modelling of uncertainty

- Model family $\{P_\theta(x, y, z) \mid \theta \in \Theta\}$
- Prior belief $\xi_0(\theta)$.
- Data $D \sim P_\theta(x, y, z)$.
- Posterior belief $\xi(\theta) \triangleq \xi_0(\theta \mid D)$.

The Bayesian value of a policy

$$V(\lambda, \xi, \pi) \triangleq \sum_{\theta \in \Theta} V(\lambda, \theta, \pi)\xi(\theta). \tag{1.3}$$

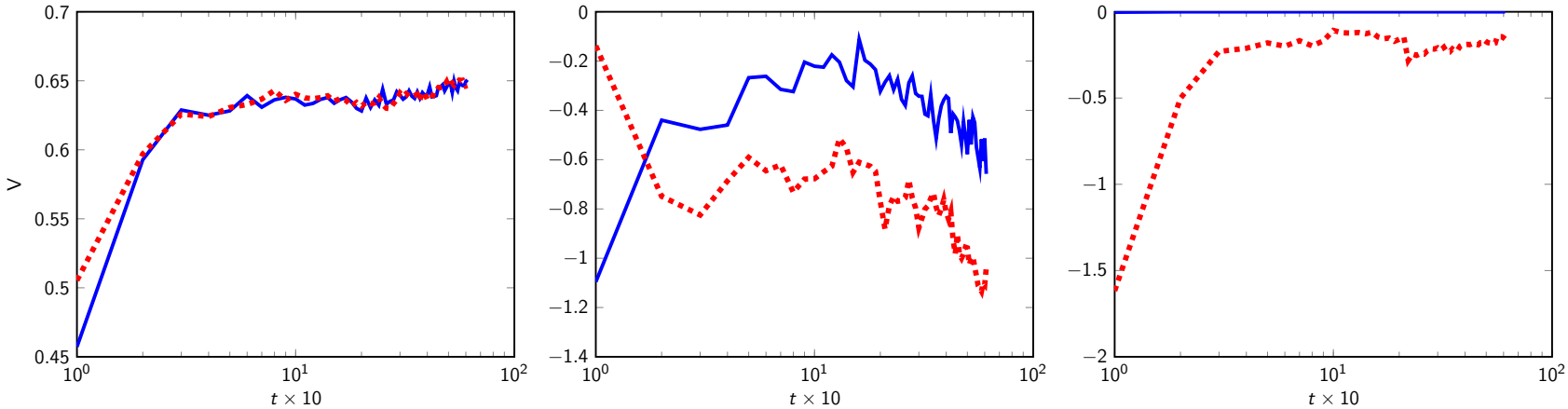
Synthetic data



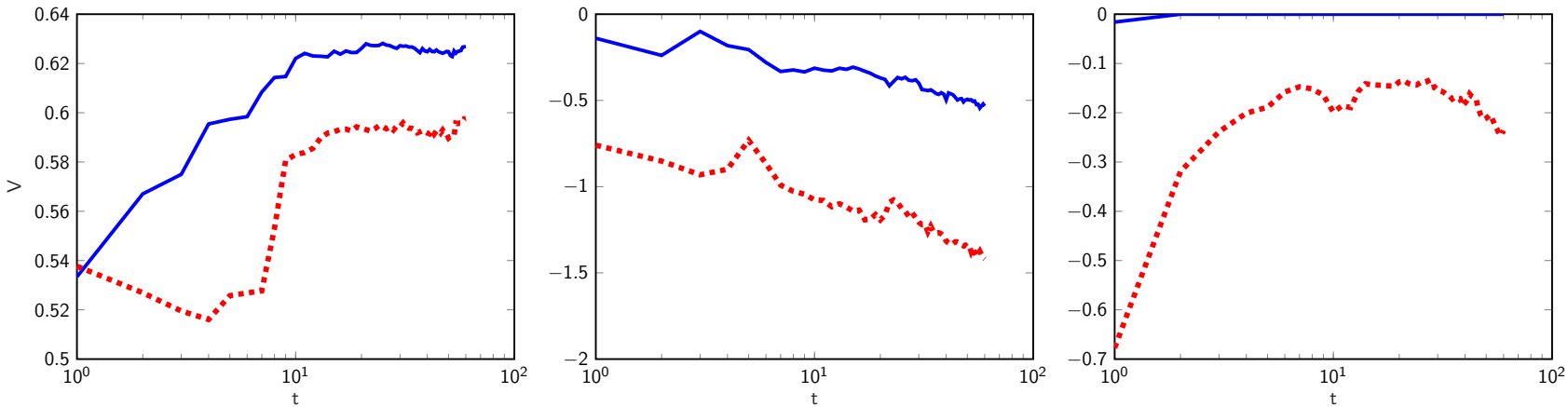
- Observe x_t.
- Take action a_t.
- Observe y_t.

Bayes
Marginal

Offline setting on COMPAS data

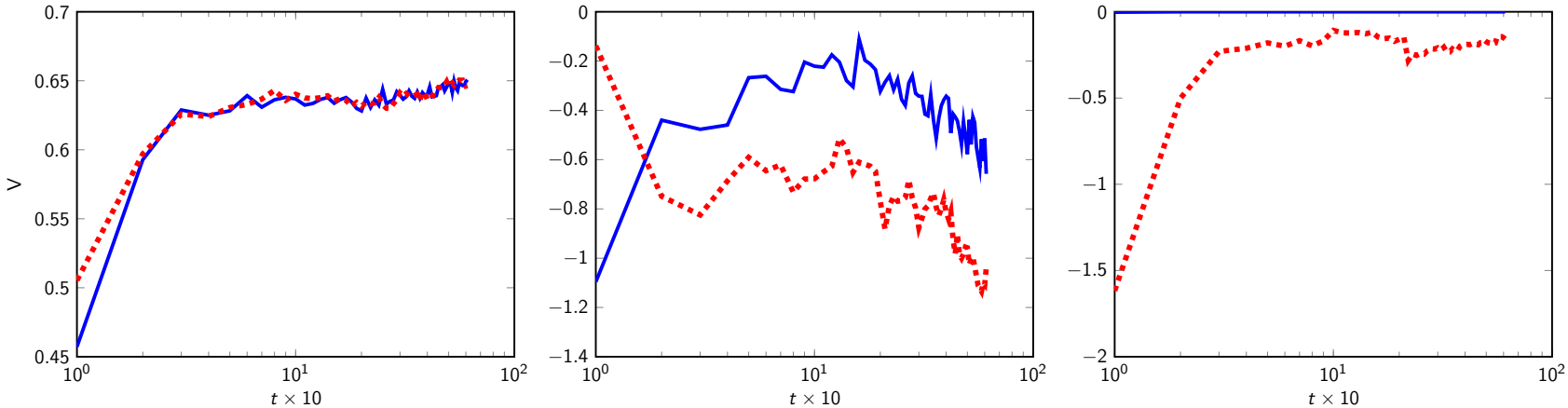


Online setting: data seen depends on decisions

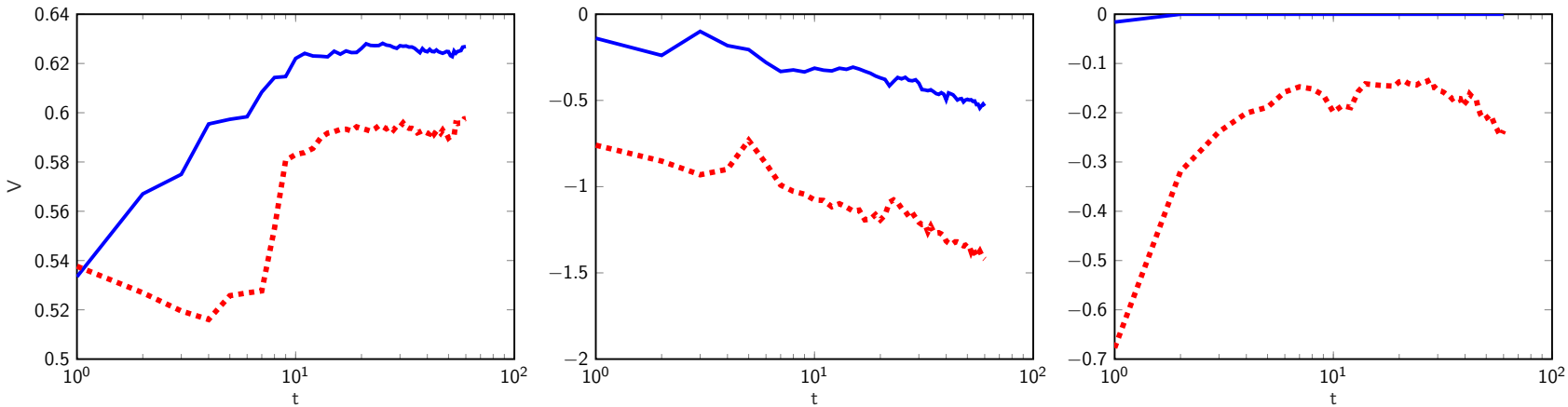


- Observe x_t.
- Take action a_t.
- Observe y_t.

Offline setting on COMPAS data



Online setting: data seen depends on decisions



- Observe x_t.
- Take action a_t.
- Observe y_t if $a_t = 1$.

Summary

- Existing criteria hard to satisfy.
- Bayesian framework:
  - Incorporates model uncertainty.
  - Considers informational aspects.
  - Lower fairness violation.

Future work

- Non-myopic sequential decision making.
- Approximate Bayesian methods.






Summary

- Existing criteria hard to satisfy.
- Bayesian framework:
  - Incorporates model uncertainty.
  - Considers informational aspects.
  - Lower fairness violation.

Future work

- Non-myopic sequential decision making.
- Approximate Bayesian methods.