Faking Fairness via Stealthily Biased Sampling

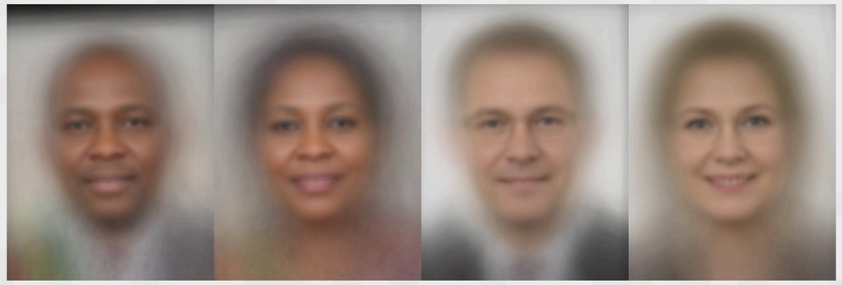
Feb 10, 2020. AAAI 2020 on AISI Track

Kazuto Fukuchi
University of Tsukuba / RIKEN AIP

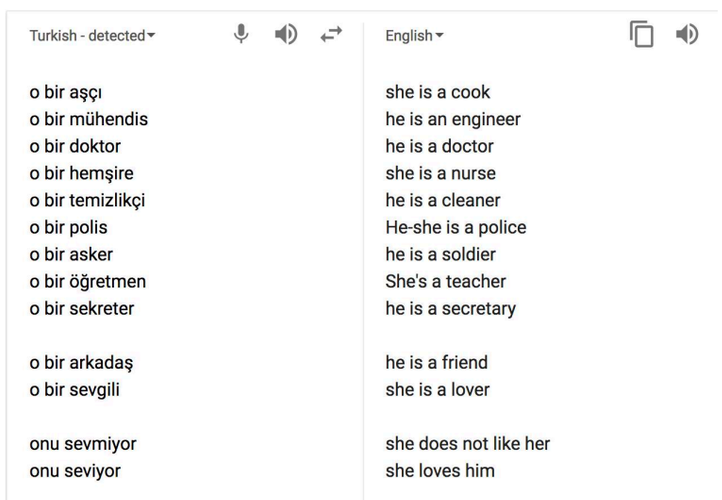
Joint work with
Satoshi Hara (Osaka University) and Takanori Maehara (RIKEN AIP)

Unfairness in Machine Learning

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Face recognition [Buolamwini+'18]



Machine translation
[Şarbak's facebook post]

BUSINESS NEWS OCTOBER 10, 2018 / 12:12 PM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

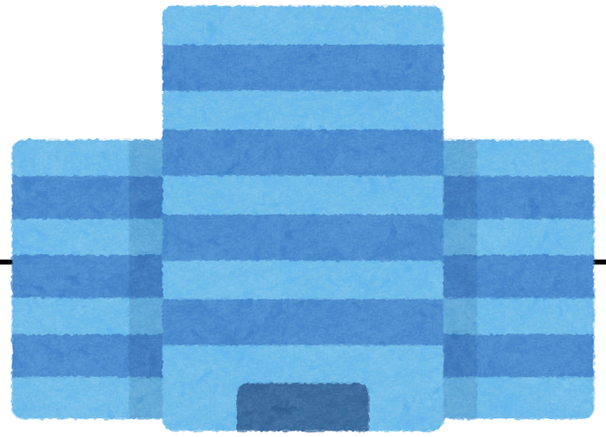
SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

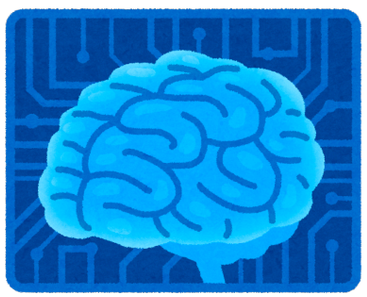
Hiring [Dastin'18]

Criminal risk assessment [Angwin+'16]

Promotion of Fairness

Company



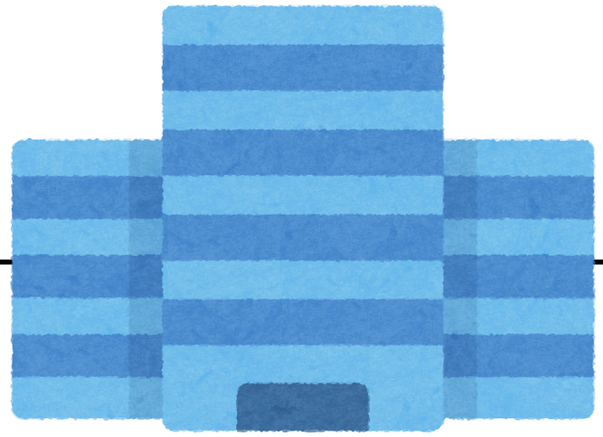


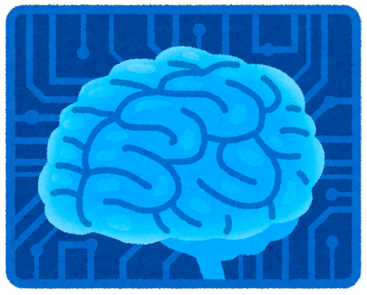
ML model

Service

Promotion of Fairness

Company





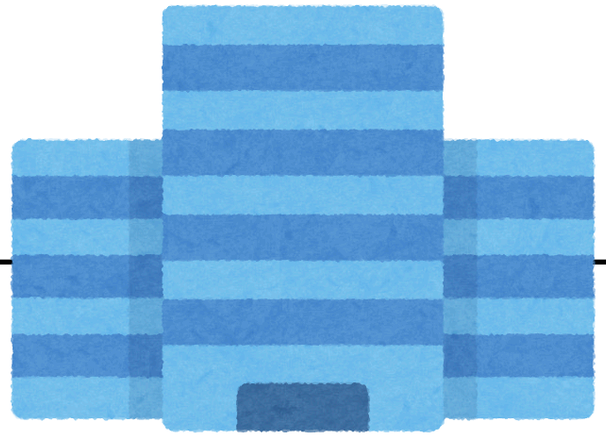
ML model

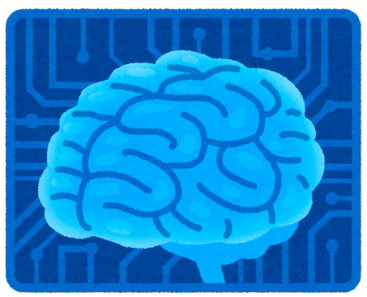
Service

Is ML Model fair?

Promotion of Fairness

Company





ML model

Service

Ok! Model is fair!

Evidence of fairness

Example: Score based evidence

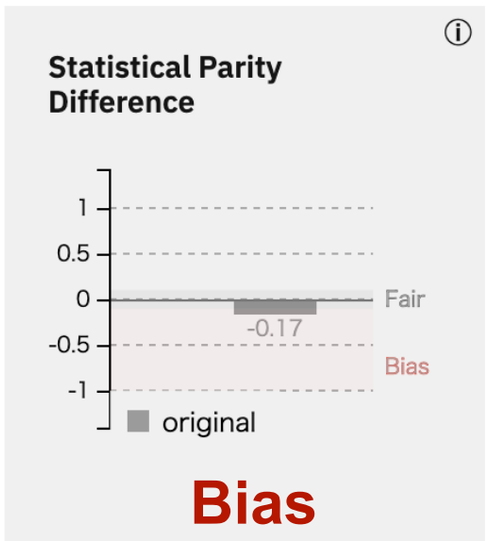
- Fairness score: a level of fairness
- Many tools for auditing fairness score have developed.
  - E.g., FairML, AI Fairness 360 [Bellamy+'18], Aequitas [Saleiro+'18]

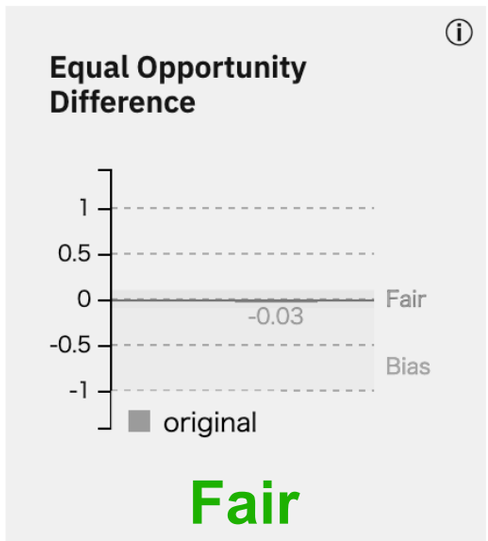
Protected Attribute: Race

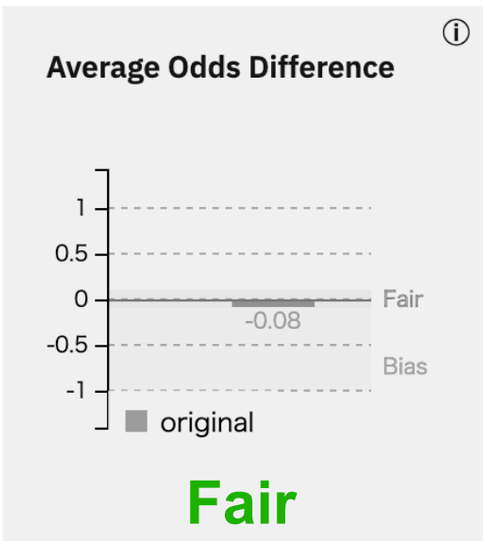
Privileged Group: ***White***, Unprivileged Group: ***Non-white***

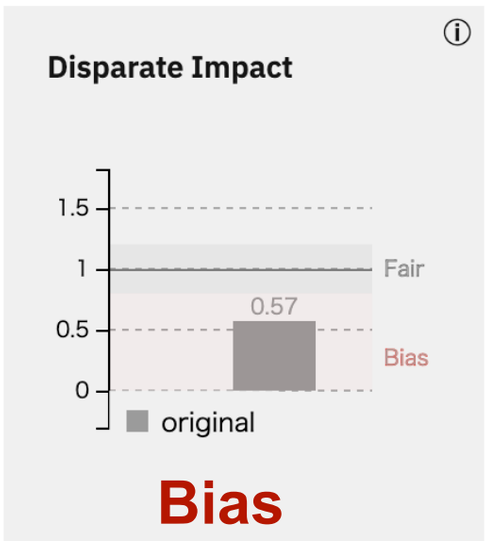
Accuracy with no mitigation applied is 82%

With default thresholds, bias against unprivileged group detected in 2 out of 5 metrics





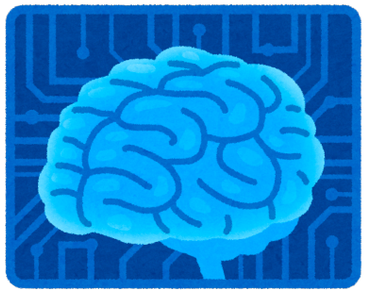




AI Fairness 360, Demo, https://aif360.mybluemix.net/data

Fake Fairness of Model

Company



Unfair ML model

Service

Is this fake evidence?

Fake evidence of fairness

Evidence of Fairness

	Pros	Cons
Score	ML model is in private	We cannot detect fake
Benchmark dataset	ML model is in private	We can detect fake(?)
Model	No chance to fake	Leakage of confidential information

Evidence of Fairness

	Pros	Cons
Score	ML model is in private	We cannot detect fake
Benchmark dataset	ML model is in private	**We can detect fake(?)**
Model	No chance to fake	Leakage of confidential information

Contributions

Fake in benchmark dataset is almost impossible to detect!

- Construct an attack algorithm, stealthily biased subsampling attack.
- Show the generated fake dataset is almost impossible to detect in theoretical and experimental ways.

Stealthily biased subsampling attack

- Two goals:
  - **Fairness**: S looks fair
  - **Stealthiness**: Distribution of S is similar to that of D

Stealthily biased subsampling attack

- Two goals:
  - **Fairness**: S looks fair
  - **Stealthiness**: Distribution of S is similar to that of D

Target contingency table

	Female	Male
Positive	40%	40%
Negative	10%	10%

Fairness

$$\min_S W(S, D)$$

Stealthiness

Optimization

Minimize Wasserstein distance

Stealthiness

Fairness

$$\min_S W(S, D)$$

$$\text{sub to } C(S) = C_T$$

Contingency table of S is equivalent to target

Optimization

Minimize Wasserstein distance

Stealthiness $\min_S W(S, D)$

Fairness $\text{sub to } C(S) = C_T$

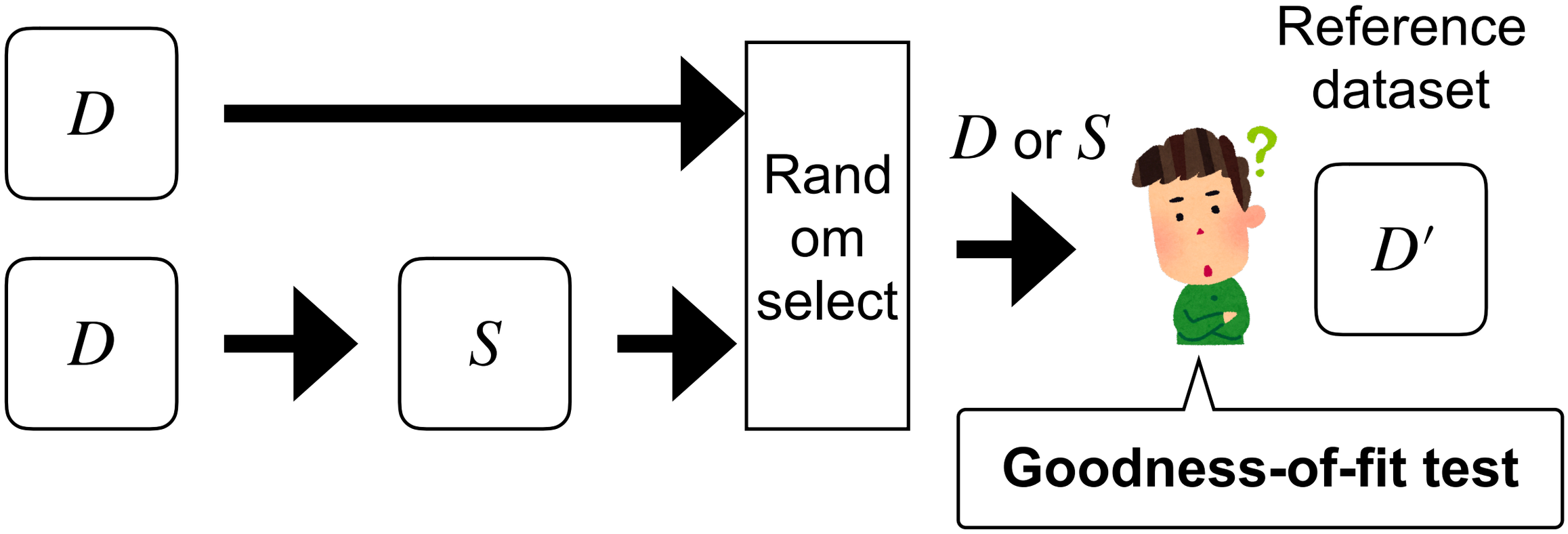
Contingency table of S is equivalent to target

- This is a linear programming but its general solver is slow :(

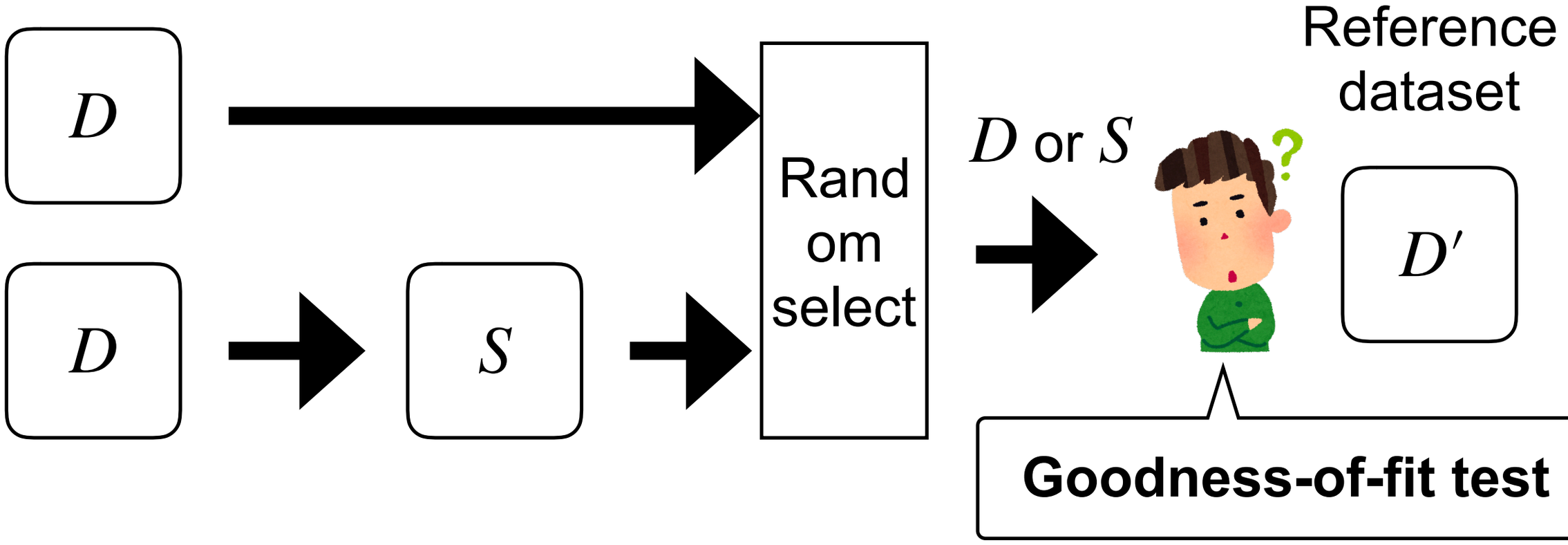
Develop fast optimization technique

with complexity $O(|D|^{2.5})$

Does Wasserstein distance actually work?



Does Wasserstein distance actually work?



(Theorem) For KS-test detector,
Detectability $\leq O(K^{1/s}W(\mu^K, \nu^K)) + o(1)$.

Minimizing WD =>
Minimizing upper bound on detectability

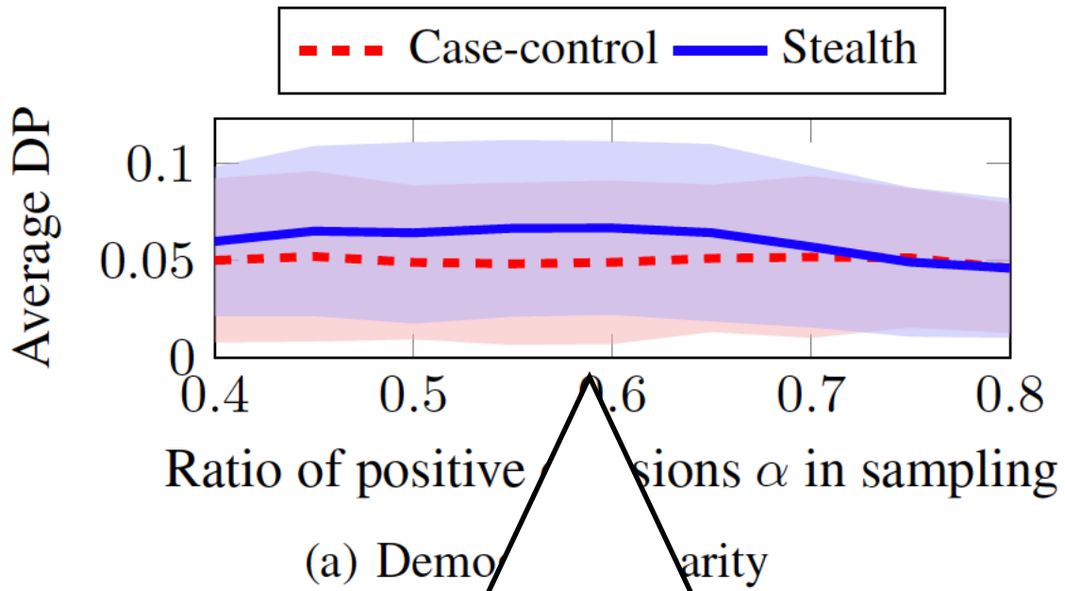
Synthetic dataset: Settings

- Evaluation criteria
  - Fairness: $\mathrm{DP} = |\mathbb{P}(y = 1 \,|\, s = 1) - \mathbb{P}(y = 1 \,|\, s = 0)|$
  - Stealthiness: Power of KS test with significance 0.05.
- Attacker made subsamples so that

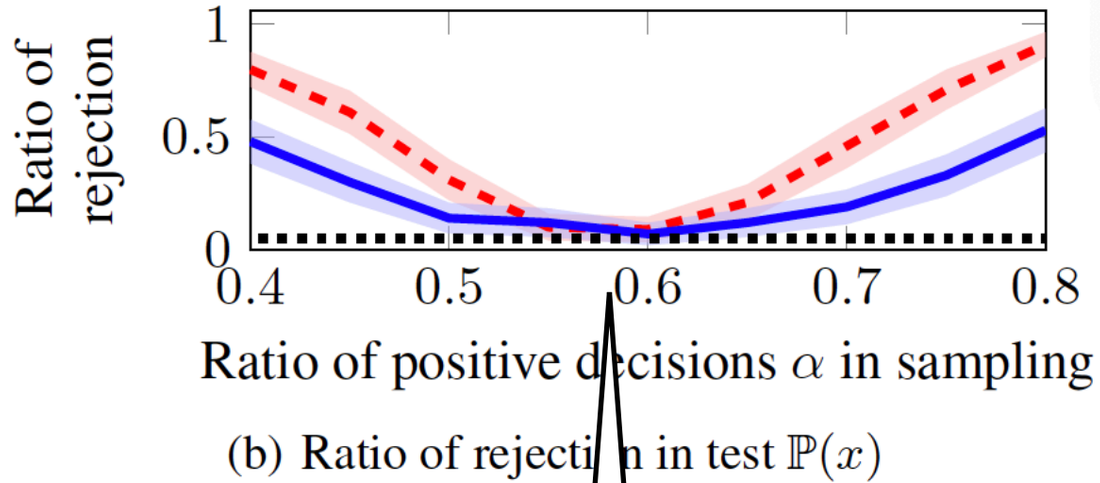
$$\mathbb{P}(y = 1 \,|\, s = 1) \approx \mathbb{P}(y = 1 \,|\, s = 0) \approx \alpha.$$

- Original dataset: $\mathrm{DP} = 0.2$, sample size = 1000, and $\alpha \approx 0.6$.
- Reference sample size: 200

Synthetic dataset: Result



(a) Demographic parity



(b) Ratio of rejection in test $\mathbb{P}(x)$

DP is much smaller than original (0.2)

Power is same as significance (0.05)

Real datasets: Settings

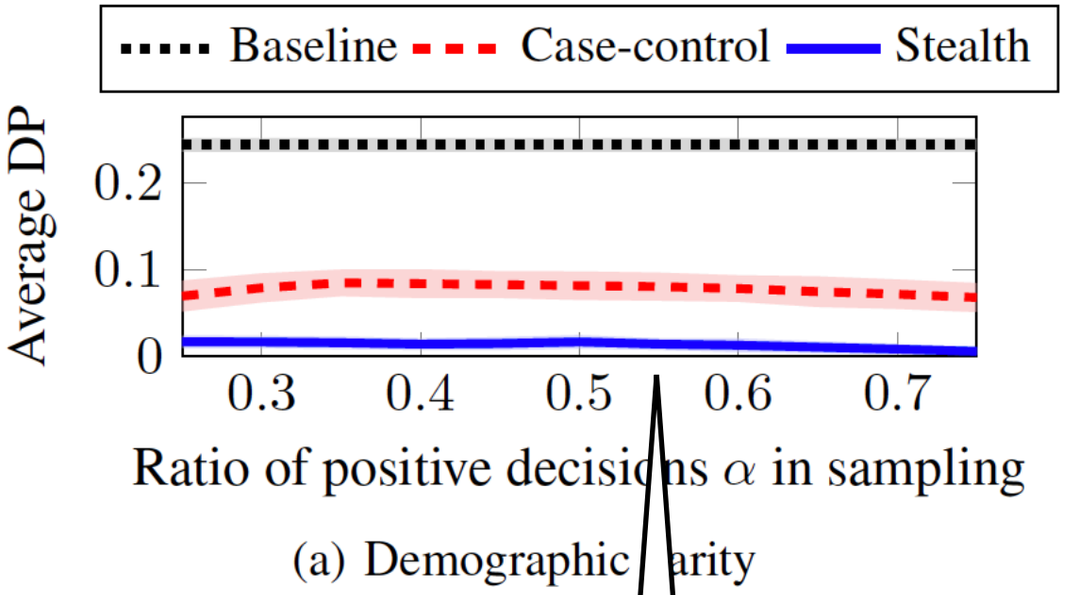
- Evaluation criteria
  - Fairness: $\mathrm{DP} = |\,\mathbb{P}(y = 1 \,|\, s = 1) - \mathbb{P}(y = 1 \,|\, s = 0)\,|$
  - Stealthiness: $W(S, D')$

- Attacker made 2000 subsamples so that

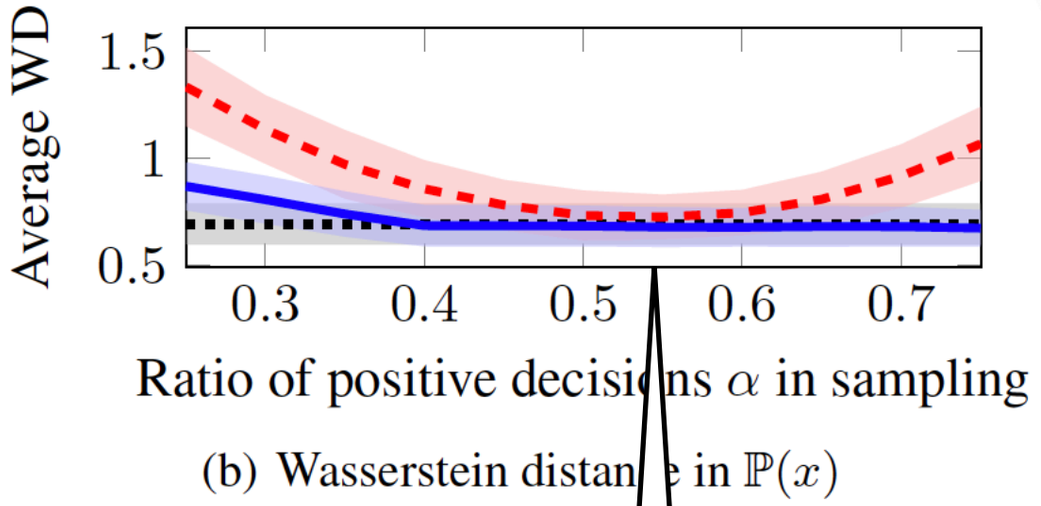
$$\mathbb{P}(y = 1 \,|\, s = 1) \approx \mathbb{P}(y = 1 \,|\, s = 0) \approx \alpha.$$

- Data
  - COMPAS (4000) and Adult (20000)
  - Reference sample size: 2000
  - $\alpha \approx 0.6$.

Real dataset: COMPAS



(a) Demographic parity

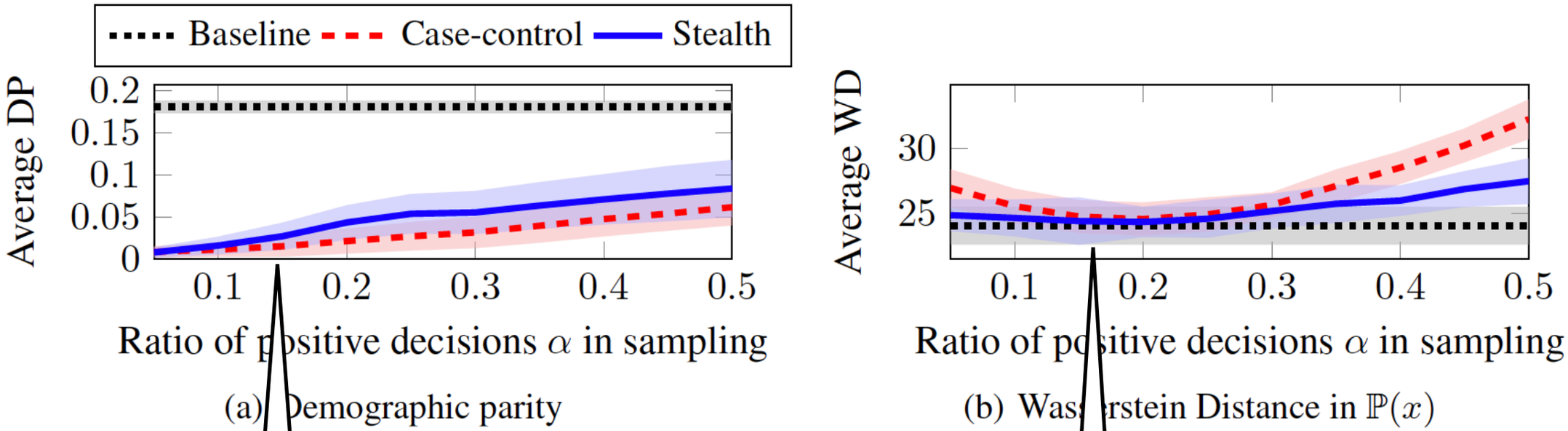


(b) Wasserstein distance in $\mathbb{P}(x)$

DP is much smaller than original (dotted line)

WD is same as baseline

Real dataset: Adult



DP is much smaller than original (dotted line)

WD is same as baseline (dotted line)

Conclusions

Summary:

- An evil company can deceive people by publishing fake evidence of fairness.
- We **CANNOT** detect fake in benchmark dataset.

We're facing a risk of fake fairness.

Paper: https://arxiv.org/abs/1901.08291

Code: https://github.com/sato9hara/stealthily-biased-sampling

Thank you!