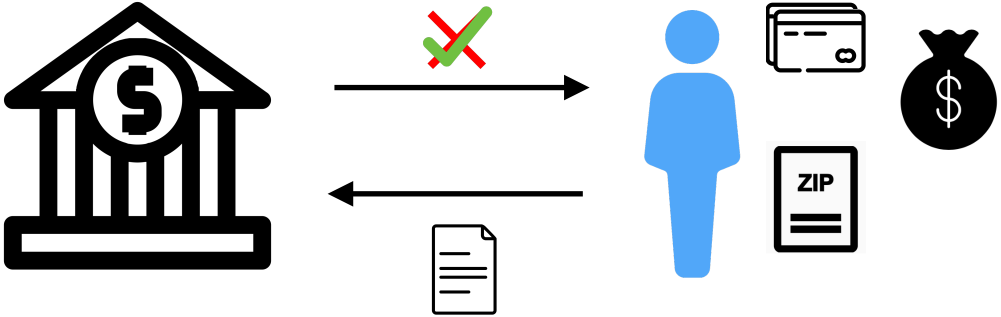
The Social Cost of Strategic Classification

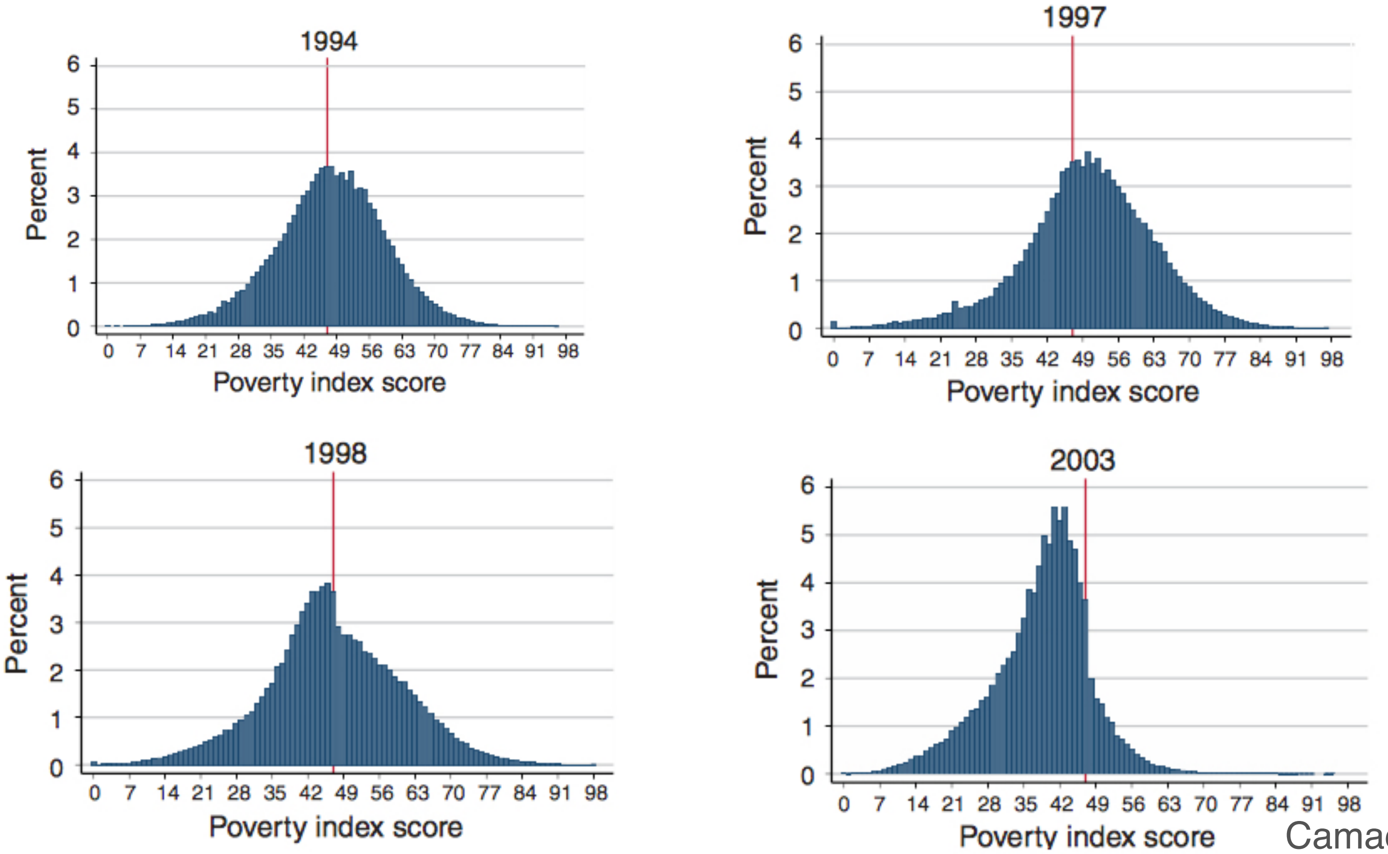
Smitha Milli, **John Miller**
Anca Dragan, Moritz Hardt

Strategic classification



Does Knowing Your FICO Score Change Financial Behavior? Evidence from a Field Experiment with Student Loan Borrowers

Means testing for social program eligibility



Camacho and Conover, 2012

When classification is used to allocate resources, individuals are incentivized to change to receive a positive classification.

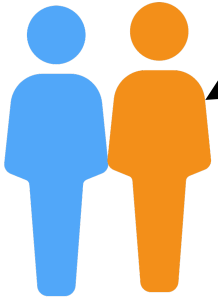
What we normally focus on

Find *Stackelberg equilibrium* which maximizes institution's accuracy after accounting for strategic behavior.

[Hardt et al, 2016, Brückner & Scheffer, 2011; Dong et al, 2018]

For Nash equilibria see [Brückner et al, 2012; Dalvi et al, 2004]

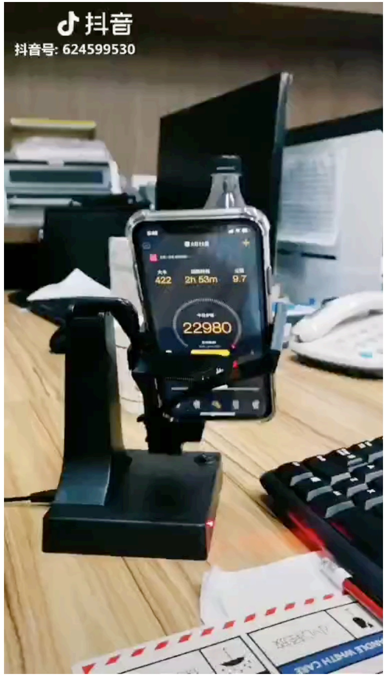
Our work



How do institutional efforts to improve strategy-robustness affect people being classified?

How should we think of strategic adaptation?

Insurance provider Oscar will reward you if you hit your step goal



Improvement or gaming?

- **Improvement:** Altering the decision by manipulation that changes the underlying label (Hand 1997)
    - Increasing net worth improves creditworthiness
    - *Positive effects*

- **Gaming**: Altering the decision by manipulating proxy features without changing the underlying label (Goodhart)
    - Opening unnecessary credit cards unrelated to repayment
    - *Unjustifiable or pointless effort*
    - **Default case in machine learning**

A causal perspective

- Partitions features into two types: *causal* and *anti-causal*

- **Improvement:** Manipulating causal features
  - Preserves classifier performance
- **Gaming:** Manipulating anti-causal features
  - Degrades classifier performances

Machine learning invites gaming

- Most machine learning systems are *anti-causal* *(Scholkopf et al. 2012)*

- Decision rules degrade under manipulation pressure
  - Goodhart's Law:
    - *"Once a measure becomes a target, it's no longer a good measure."*

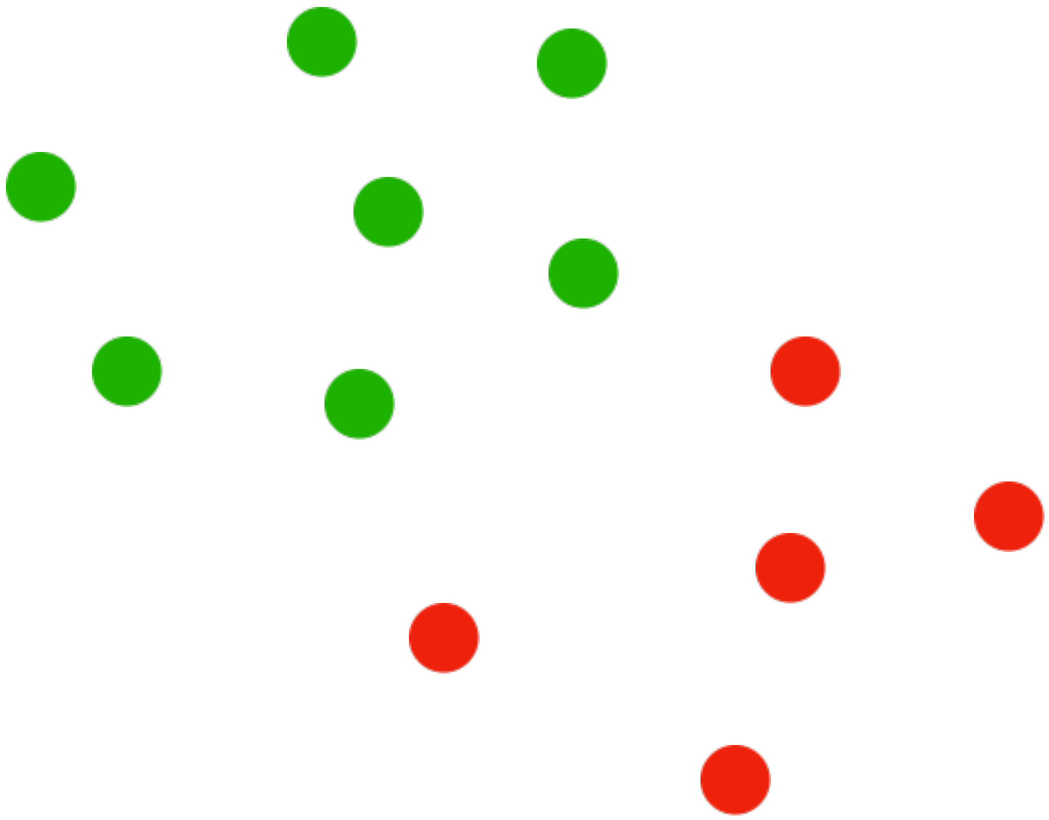
- **Strategic adaptation to machine learning systems is often gaming, not improvement.**

Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment

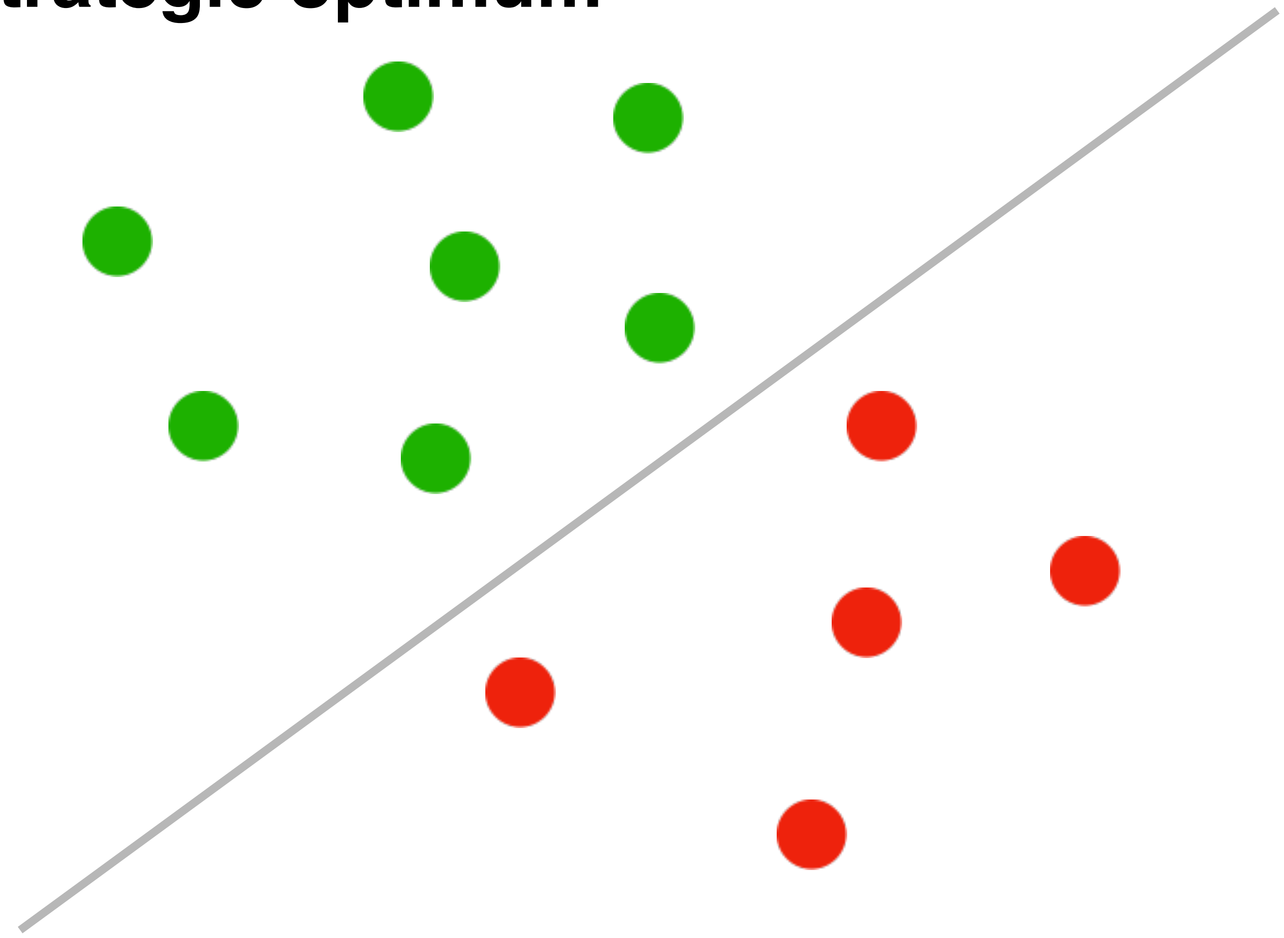
Daniel Björkegren[1] and Darrell Grissen[2]

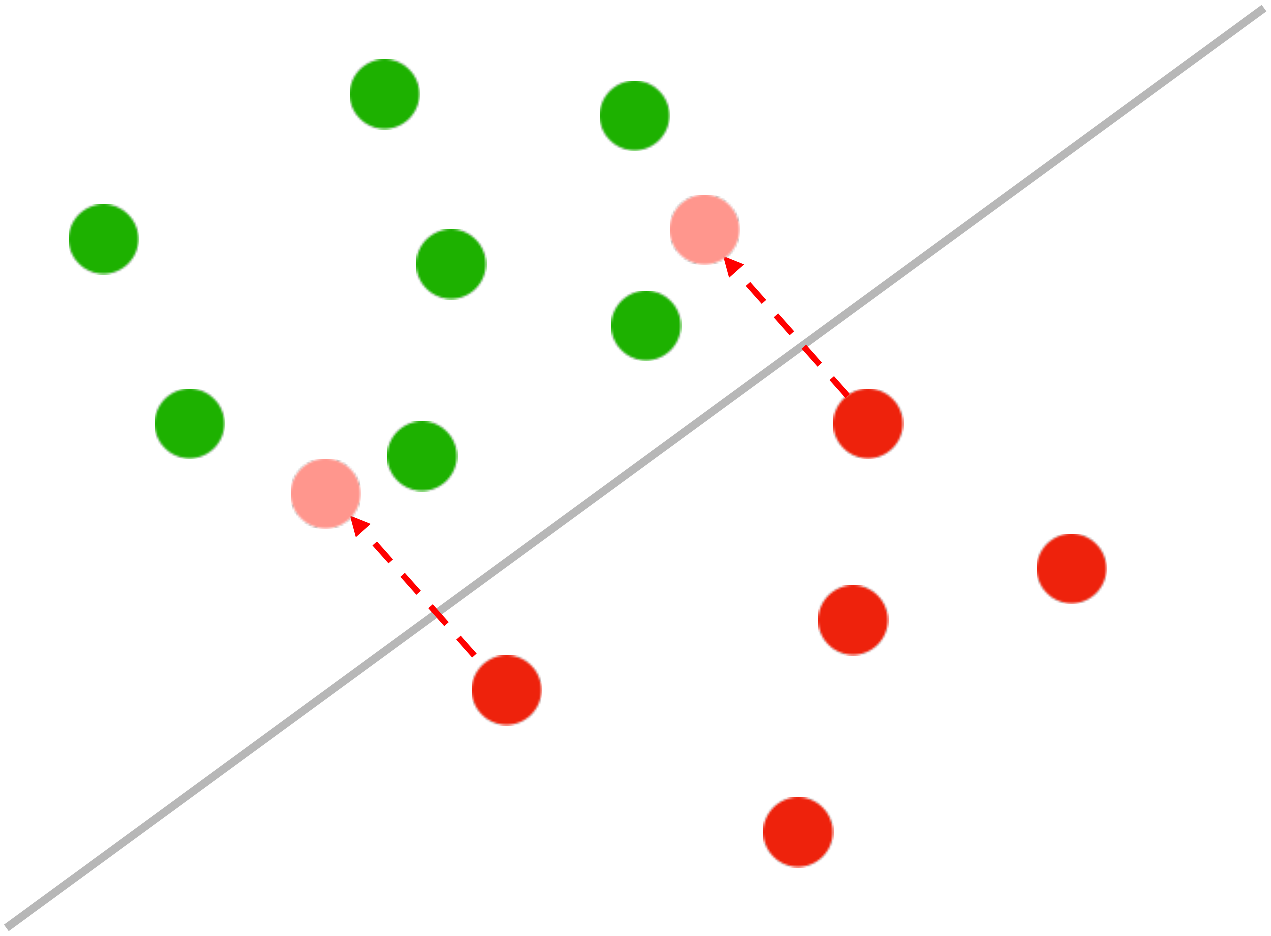
Features
Number of outgoing calls
Text response rate
Average airtime balance
Entropy of GPS coordinates

How does the institution defend against strategic behavior?

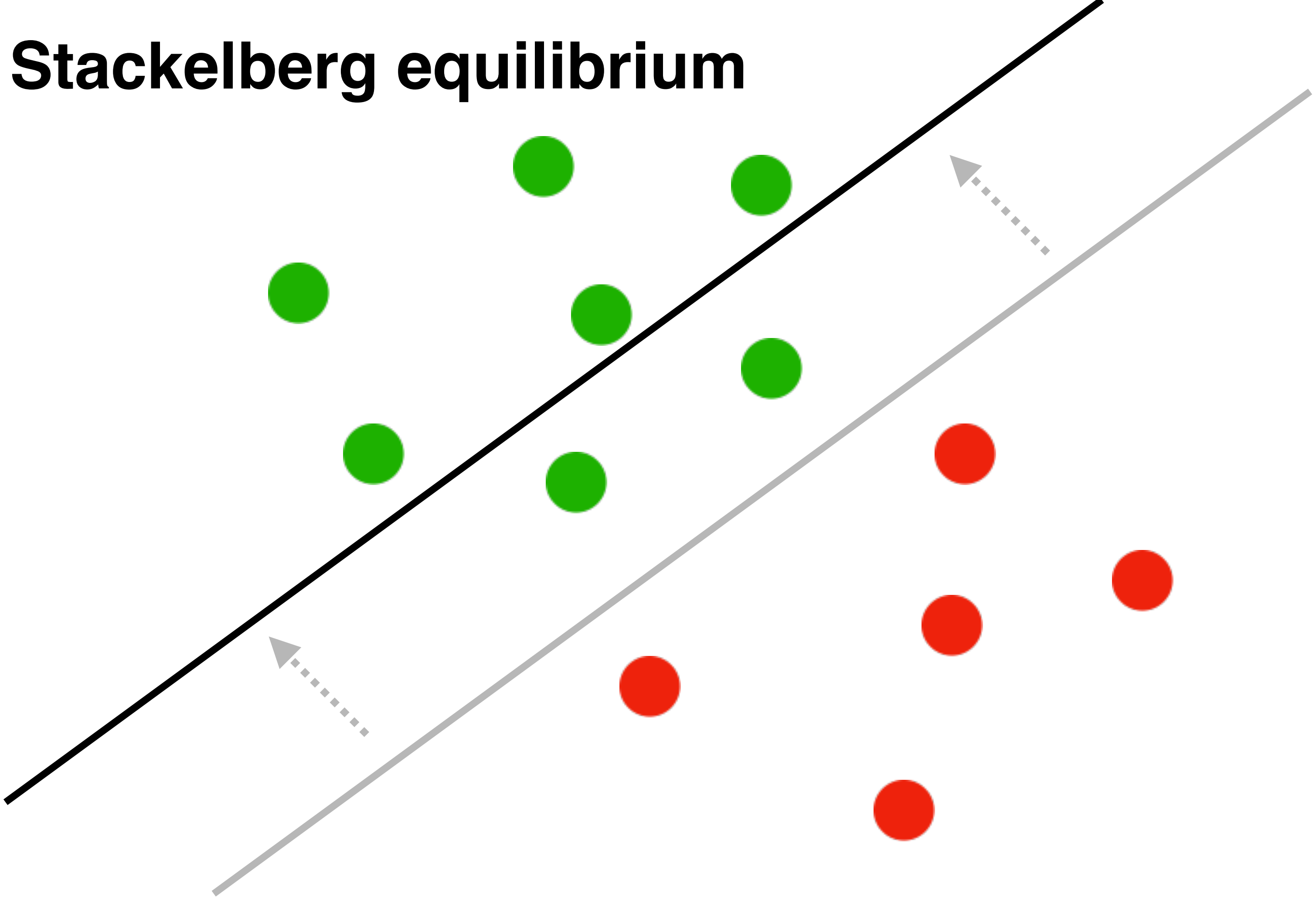


Non-strategic optimum

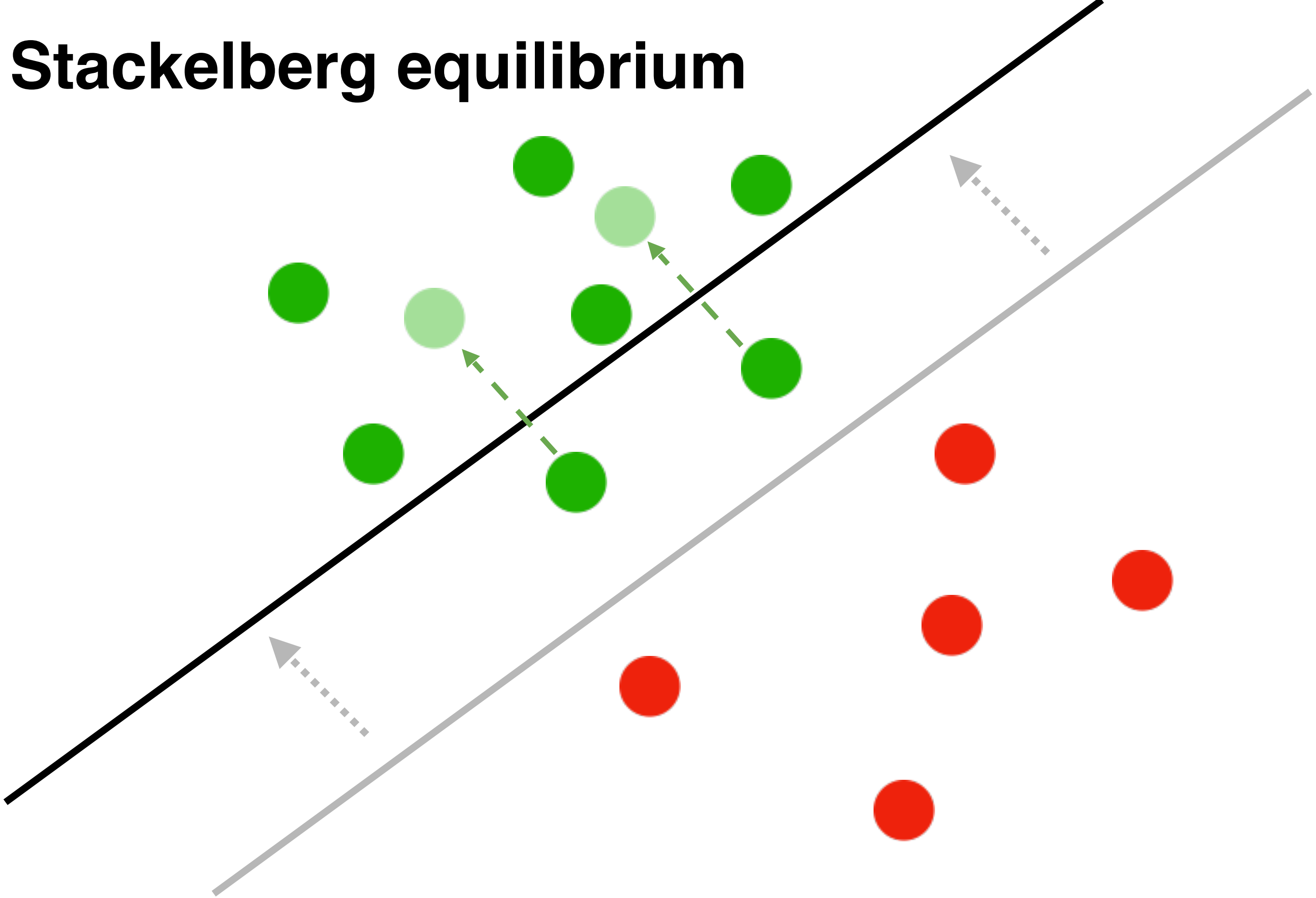




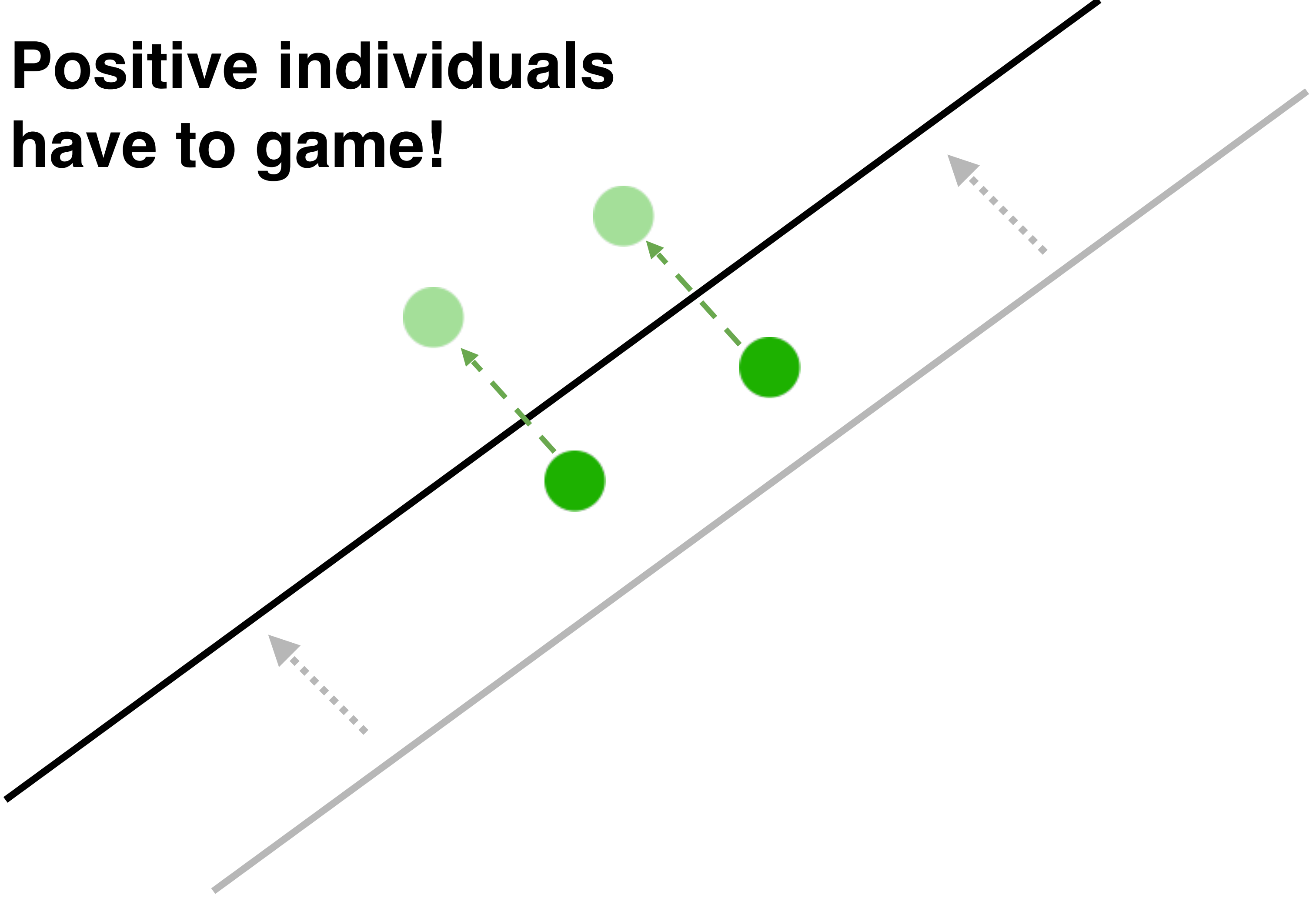
Stackelberg equilibrium



Stackelberg equilibrium



Positive individuals have to game!



Accuracy and social cost trade-offs

A formal model

classifier

$f : \mathcal{X} \to \{0, 1\}$

initial x

Label $y \in \{0, 1\}$

$\text{cost}(\text{initial } x, \text{new } x)$

$$\text{Utility}(f) = \mathbb{P}(f(\text{initial } x) = y)$$

$$\text{BR}(\text{initial } x; f) = \arg\max_{\text{new } x} f(\text{new } x) - \text{cost}(\text{initial } x, \text{new } x)$$

$$\text{StrategicUtility}(f) = \mathbb{P}(f(\text{BR}(\text{initial } x)) = y)$$

Social burden

- **Individual burden**: Minimum cost to be classified positively

$$\text{Burden}(\text{initial } x) = \min_{\substack{\text{new } x \\ \text{accepted}}} \text{cost}(\text{initial } x, \text{new } x)$$

- **Social burden:** Expected cost for positive individual to be classified positively

$$\text{SocialBurden} = \mathbb{E}\left[\text{Burden}(\text{initial } x) \mid Y = 1\right]$$

- *Gaming costs for positive individuals to receive the correct outcome*

Alternative measures of social cost

- $\text{SocialBurden} = \mathbb{E}\left[\text{Burden}(\text{initial } x) \mid Y = 1\right]$

- Expected **cost of recourse:** $\mathbb{E}\left[\text{Burden}(\text{initial } x)\right]$ (Ustun et al. 2019)
    - Cf. Delayed Impact of Fair Machine Learning

- Expected **cost of strategy** (Braverman and Garg 2019)

$$\mathbb{E}\left[\text{AgentUtility}(\text{BR}(\text{initial } x)) \mid Y = 1\right]$$

- **False positive/false negative rates** (Hu et al. 2019)

How does institutional utility trade-off against social burden?

Lemma: Reduction to threshold classifiers

- $\text{Likelihood}(x) = \mathbb{P}(Y = 1 \mid X = x)$

- Key assumption: **Outcome monotonic costs**
  - Cost to move to higher likelihood points increases monotonically with likelihood
  - No cost to move to points with lower likelihood

$$\text{Likelihood}(x) > \text{Likelihood}(z) \implies \text{cost}(a, x) > \text{cost}(a, z)$$

- Lemma: If costs are outcome monotonic, every classifier has an equivalent **threshold classifier** *with the same institutional utility and social burden.*

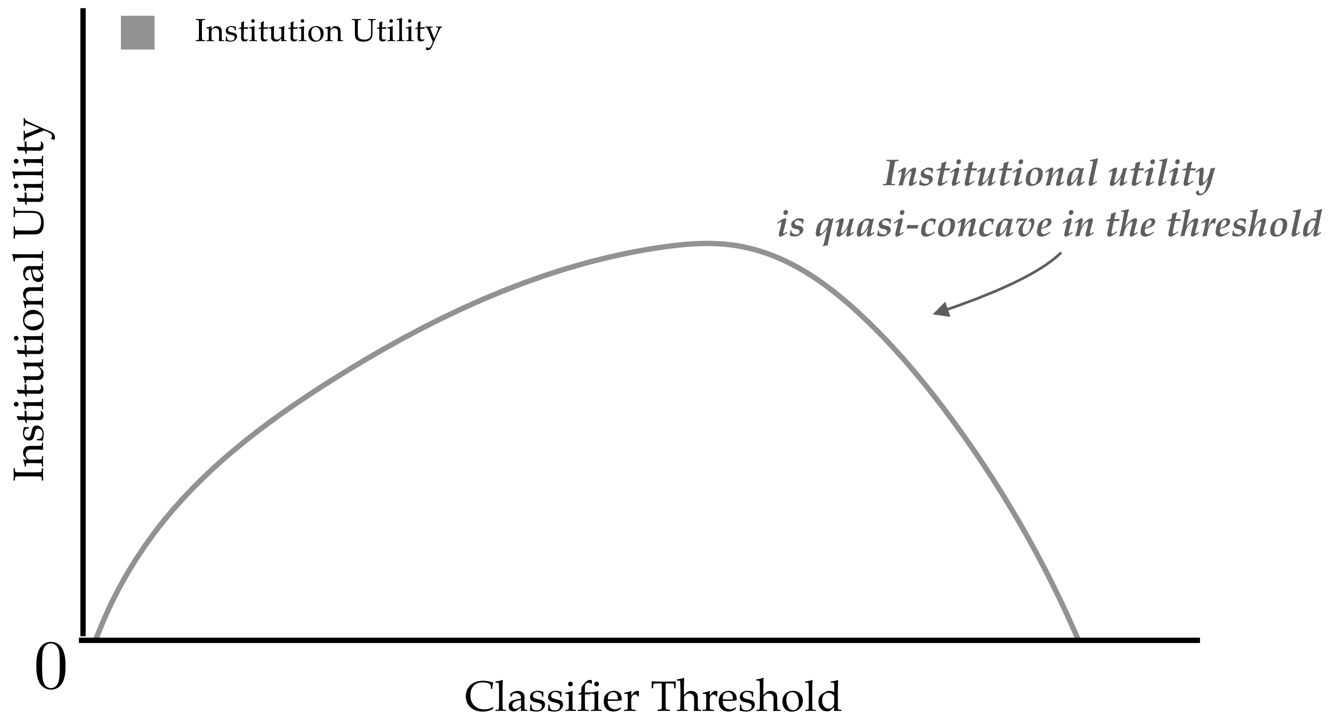
$$f(x) = \mathbb{I}\{\text{Likelihood}(x) > \tau\}$$

Result

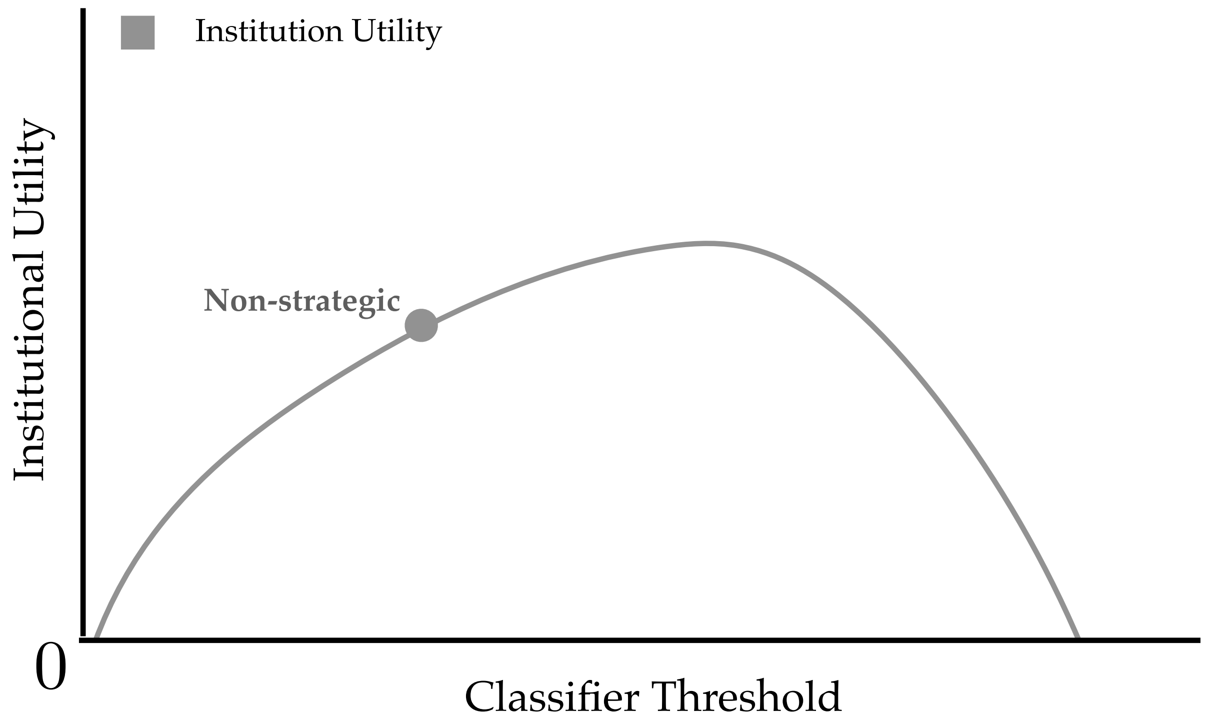
0

Classifier Threshold

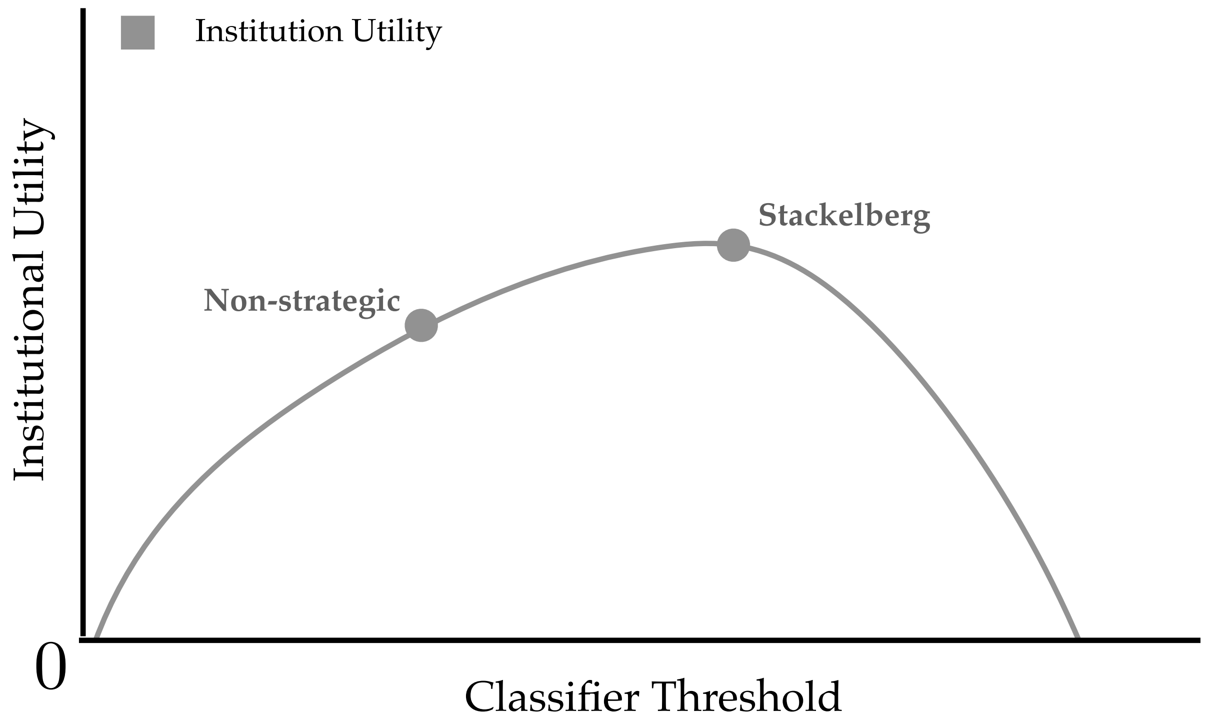
Result


Institution Utility
Institutional Utility
Institutional utility
is quasi-concave in the threshold
0
Classifier Threshold

Result

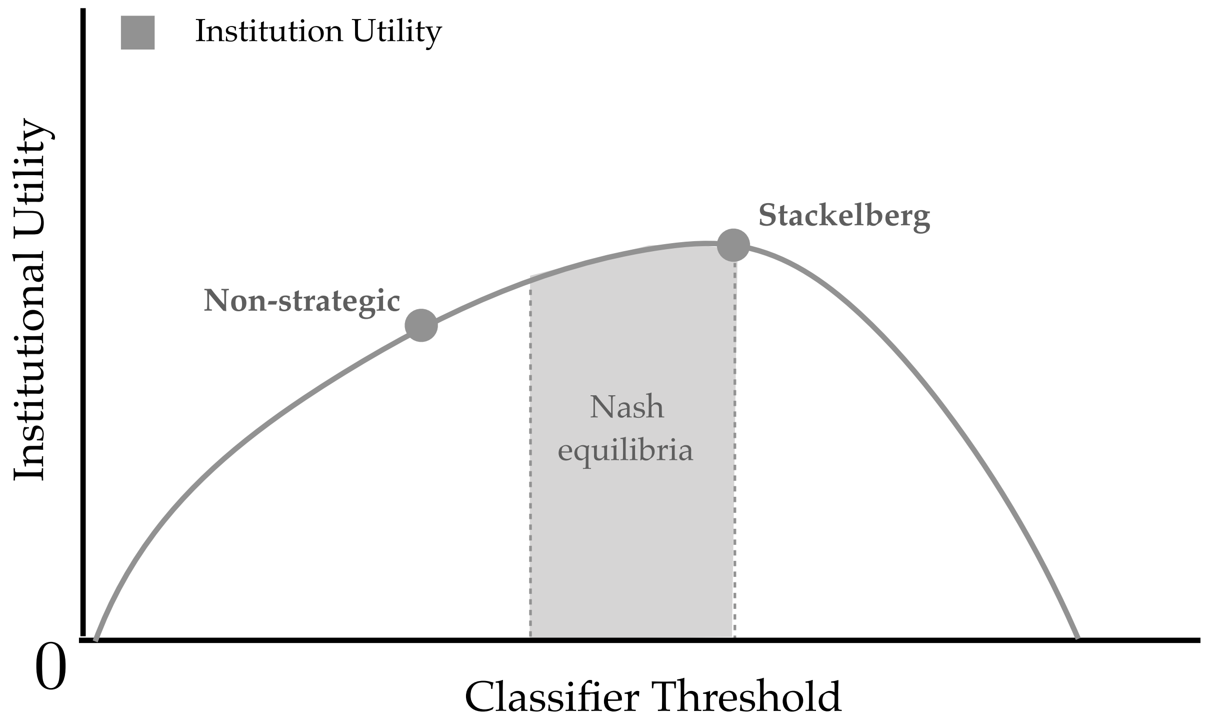

Institution Utility
Non-strategic
Institutional Utility
0
Classifier Threshold

Result

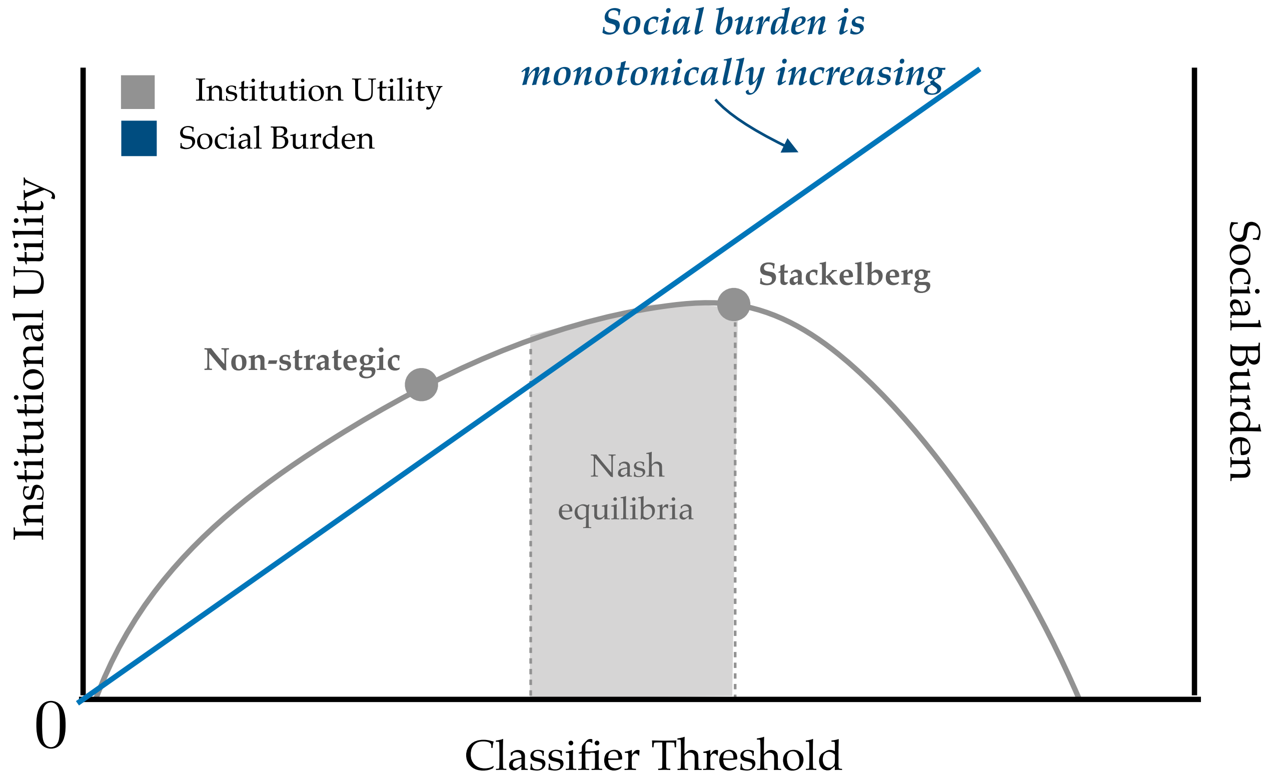


Institution Utility
Institutional Utility
Stackelberg
Non-strategic
0
Classifier Threshold

Result


Institution Utility
Institutional Utility
Stackelberg
Non-strategic
Nash
equilibria
0
Classifier Threshold

Result



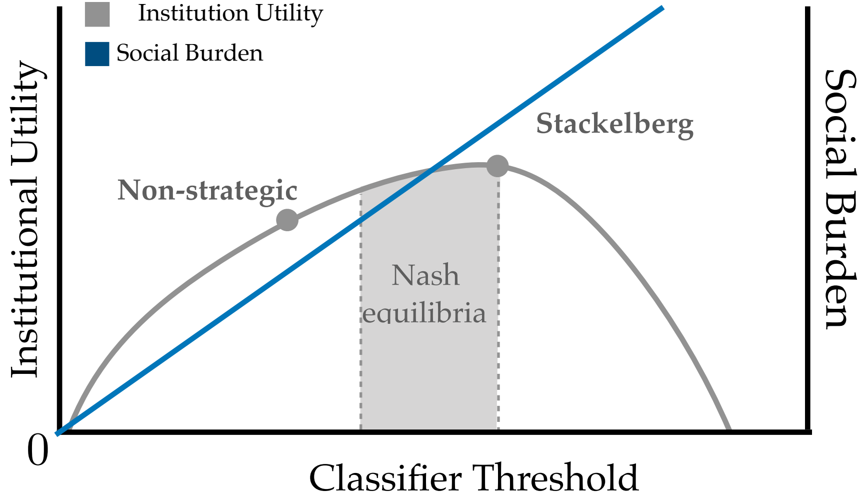
Social burden is monotonically increasing
Institution Utility
Social Burden
Institutional Utility
Social Burden
Stackelberg
Non-strategic
Nash equilibria
0
Classifier Threshold

Institutions should utilize different trade-offs

- Choice of operating point is *context-dependent*

- Institutional accuracy is often a misspecified objective
  - Small accuracy gains may not outweigh social burden

- Nash equilibria:
  - Lower social burden

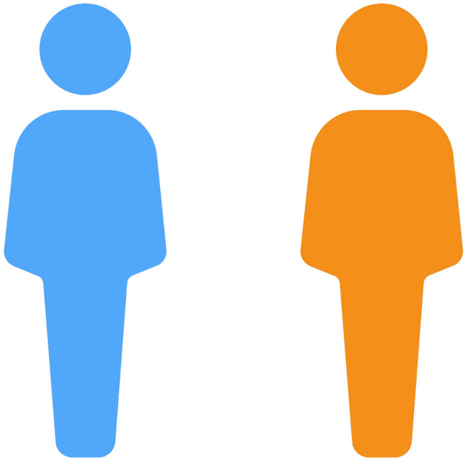


Fairness concerns

How is the social burden distributed across subpopulations?

Measure of unequal burden

Advantaged group G=a



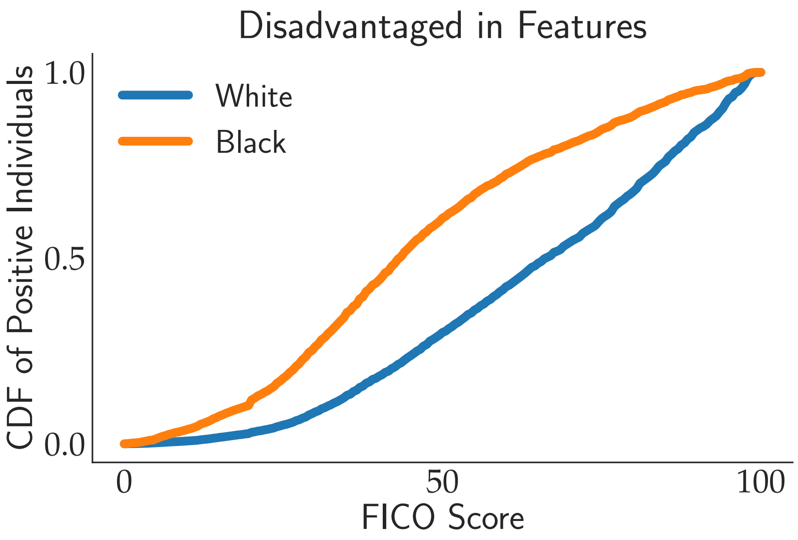
Disadvantaged group G=b

$$\mathrm{SocialGap} = \mathrm{SocialBurden}(b) - \mathrm{SocialBurden}(a)$$

Disadvantage 1: Disadvantaged in features

- Positive individuals from Group B have *lower outcome likelihoods* than Group A

$$\mathbb{P}(\text{Likelihood} \leq \ell \mid Y = 1, G = a) \leq \mathbb{P}(\text{Likelihood} \leq \ell \mid Y = 1, G = b)$$



Disadvantage 2: Disadvantaged in costs

- Positive individuals from Group B have *higher manipulation costs* than Group A
  - Similar to Hu et al. 2019

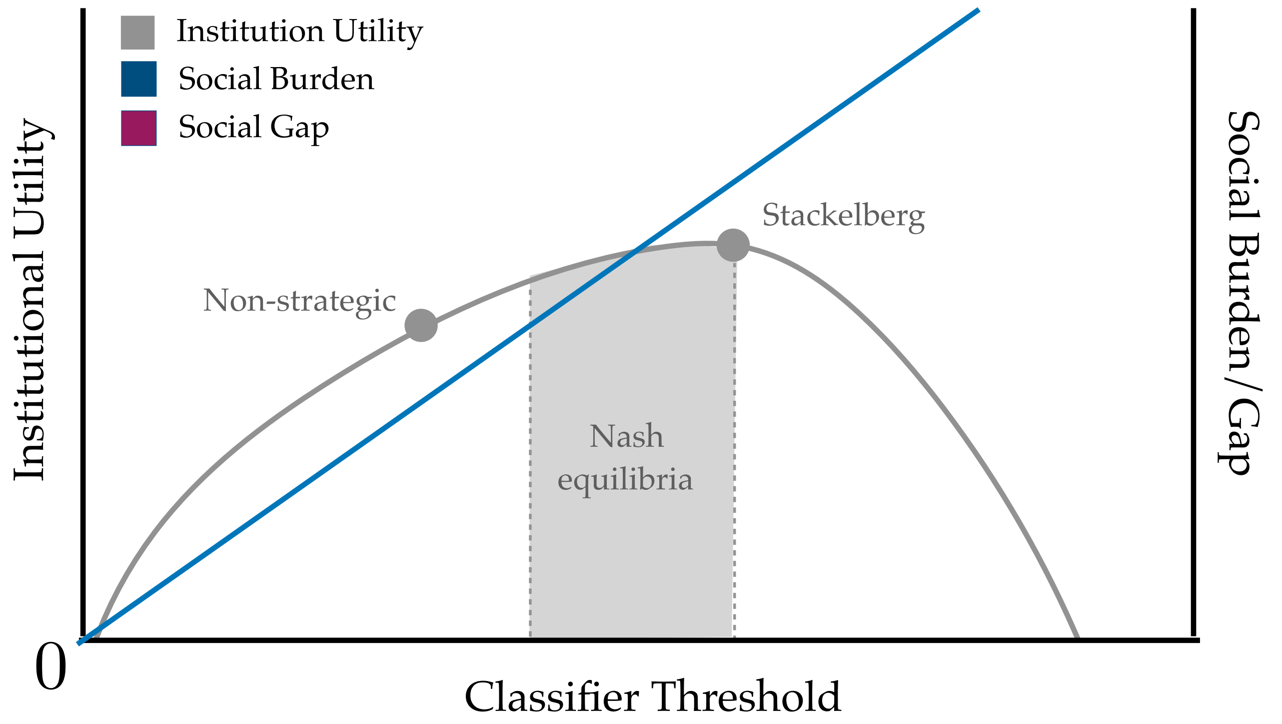
$$\mathrm{cost}_{\text{Group B}}(\cdot,\cdot) \geq \mathrm{cost}_{\text{Group A}}(\cdot,\cdot)$$

How can differences in cost arise?

- **Economic differences**
  - *College admissions*: Families with means can access test-prep services and extracurriculars

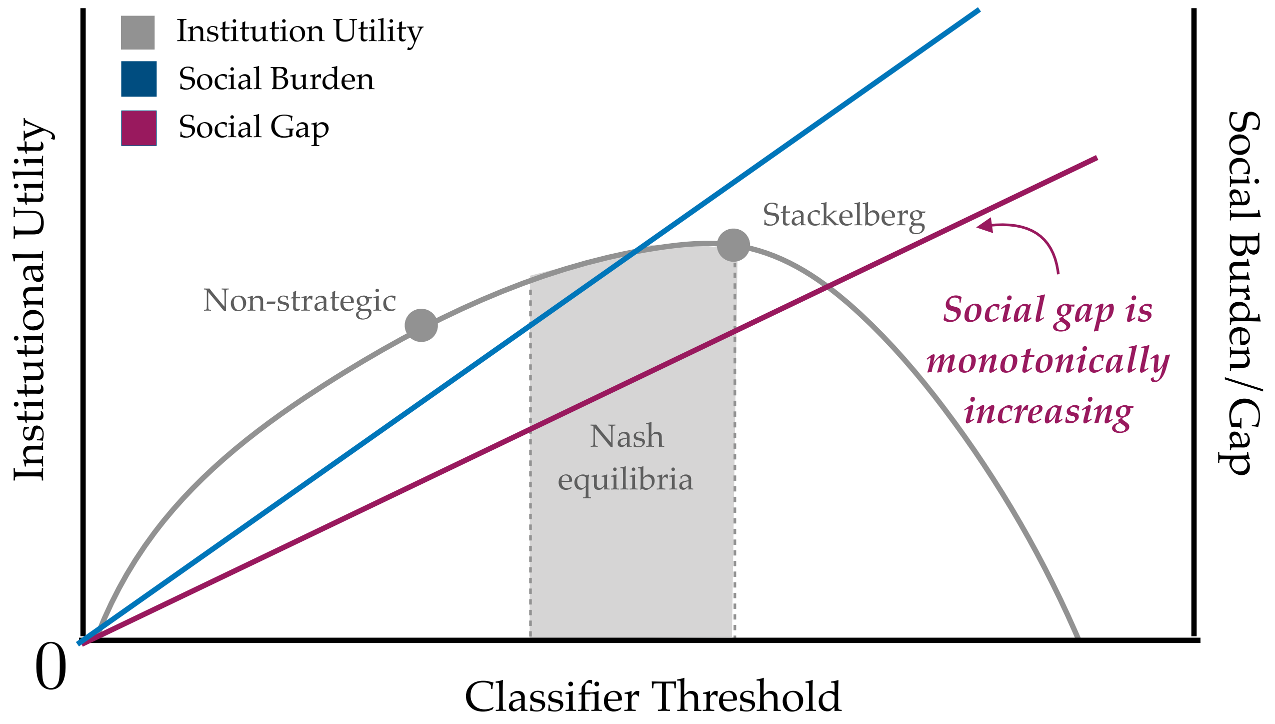
- **Information asymmetry**
  - *Social program targeting*: Families with political connections vs. those without

Result



Institution Utility
Social Burden
Social Gap
Institutional Utility
Social Burden/Gap
Stackelberg
Non-strategic
Nash
equilibria
0
Classifier Threshold

Result

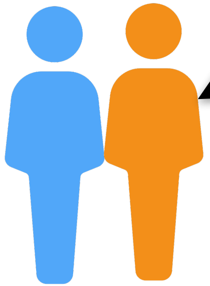


Institution Utility
Social Burden
Social Gap
Stackelberg
Non-strategic
Nash equilibria
Social gap is monotonically increasing
Institutional Utility
Social Burden/Gap
0
Classifier Threshold

Recap

What we normally focus on

Our work



↑ Institutional Accuracy

↑ Social Gap

↑ Social Burden

Understanding how our ML models affect the people who *adapt* to those models

Questions?

Smitha Milli

Anca D. Dragan

Moritz Hardt

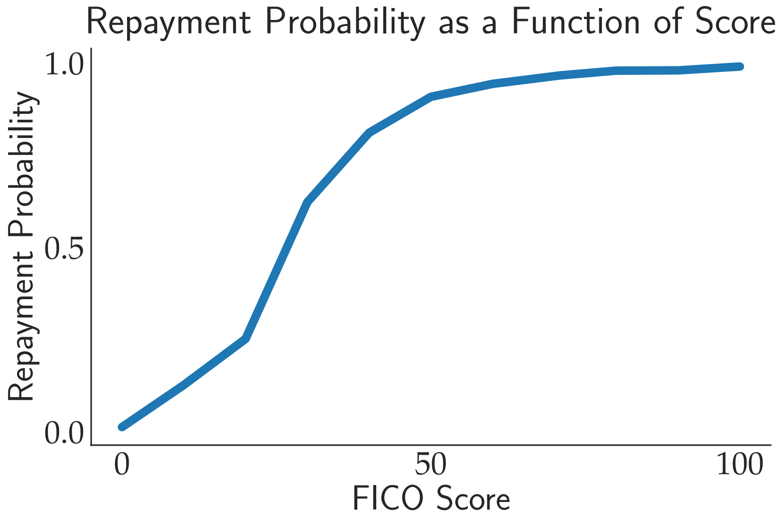
Thank you!

Experiments on FICO Data

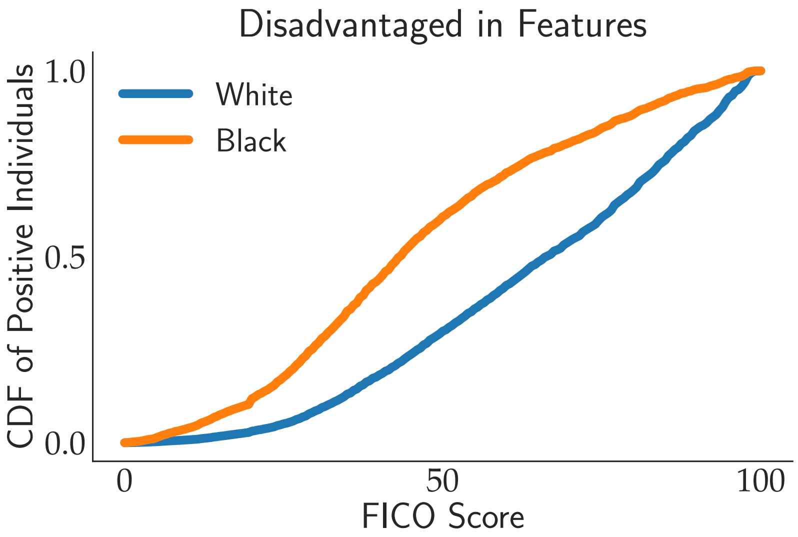
Experiments on FICO credit scores

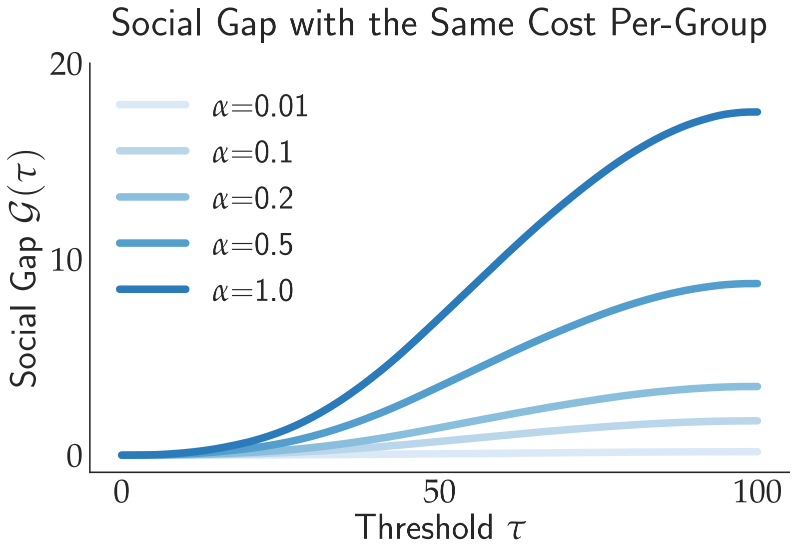
- 300,000+ TransUnion Risk Scores from 2003 (Hardt et al. 2016)
- Threshold classifiers on the FICO score
- Costs are linear function of score

$$\text{cost} = \alpha(\text{new score} - \text{initial score})_+$$



Measurement bias





Disadvantage in costs