Information Discrepancy in Strategic Learning

International Conference on Machine Learning

Baltimore, USA, July 2022

Yahav Bechavod
Hebrew University

Chara Podimata
Harvard University

Steven Wu
Carnegie Mellon University

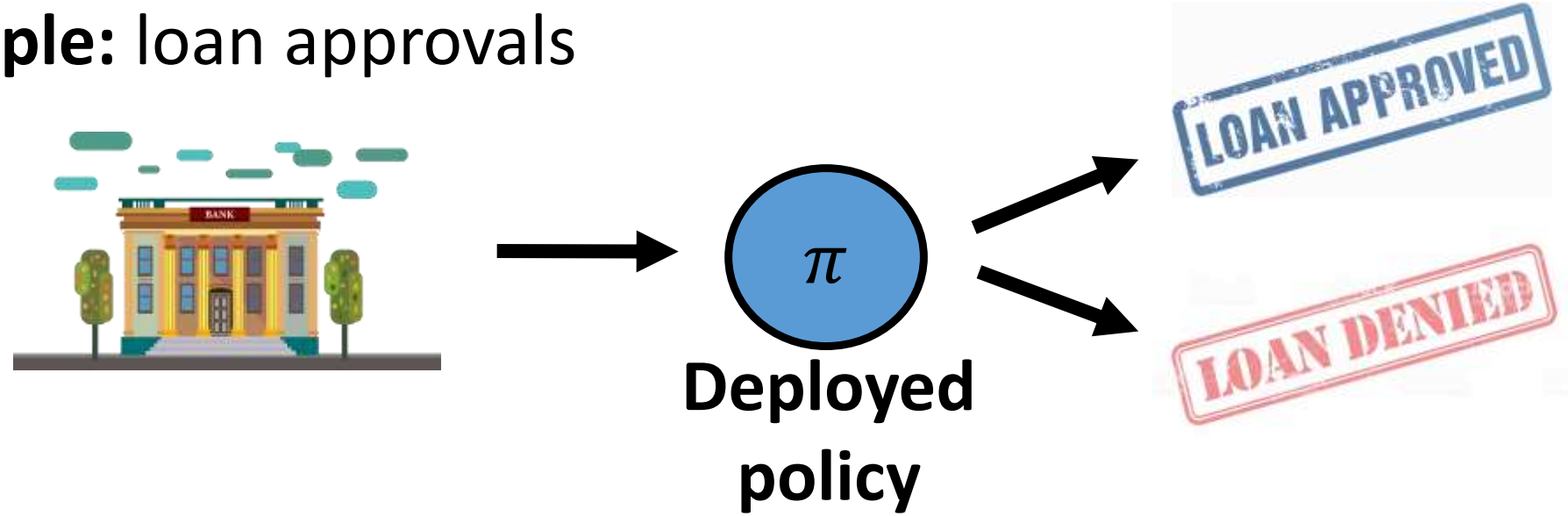
Juba Ziani
Georgia Institute of Technology

Strategic Learning

Settings which:

1. Involve decision making over human individuals.
2. Certain outcomes more desirable than others.

Example: loan approvals



Deployed policy

Strategic Learning

Individuals would like to receive
more favorable assessments
→ Act <u>strategically</u>
→ Strategic feature modifications

Example: loan approvals

π

Deployed policy

Strategic Feature Modification

π

Deployed policy

$x \in R^d$

$x' \in R^d$

$\pi(z)$=0 — **"Likely to default"**

$\pi(z)$=1 — **"Likely to repay"**

Ideally:

1. $\pi(x') >> \pi(x)$.
2. $Cost(x, x')$ is small.

Strategic feature modification

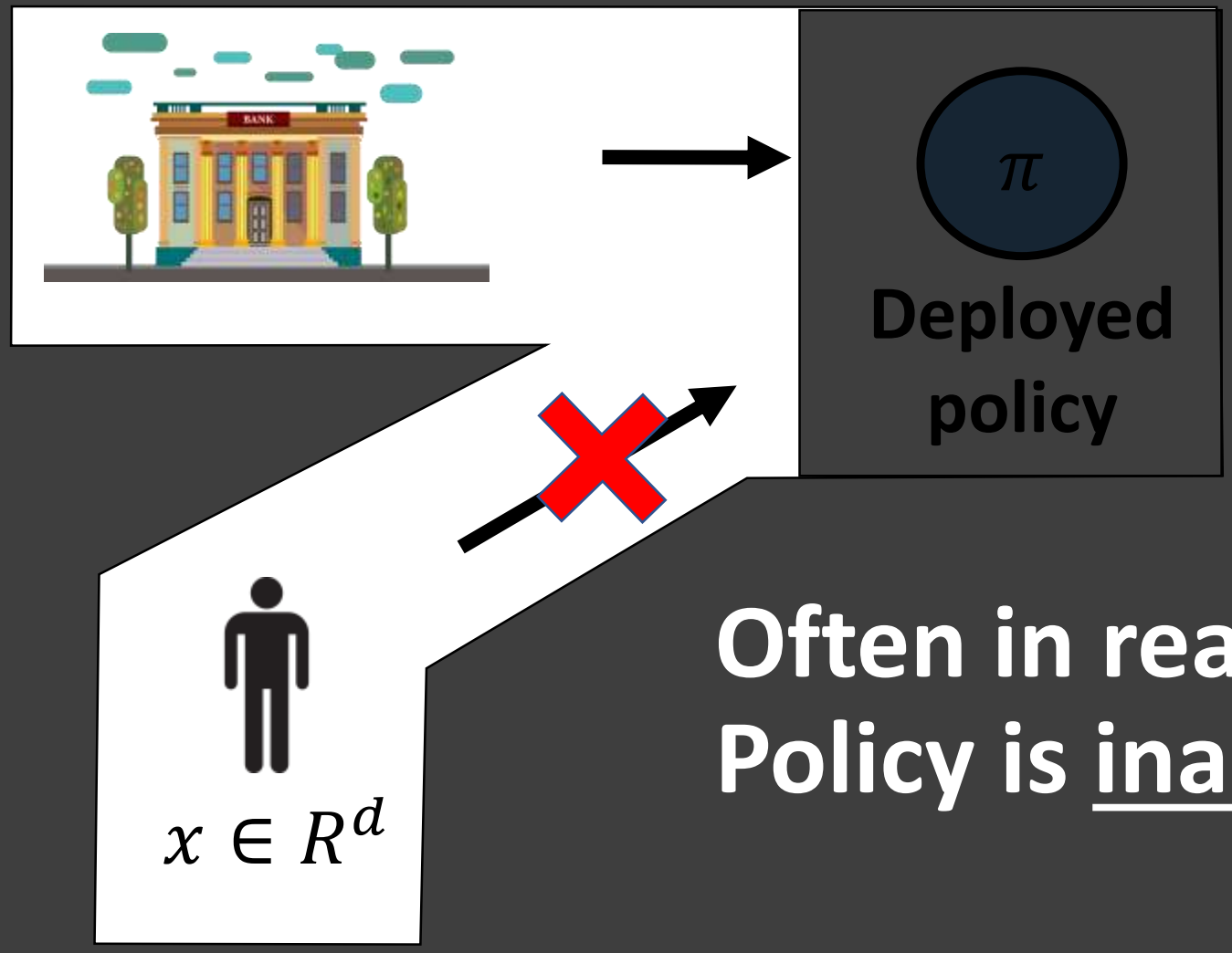
π

Deployed policy

$x \in R^d$

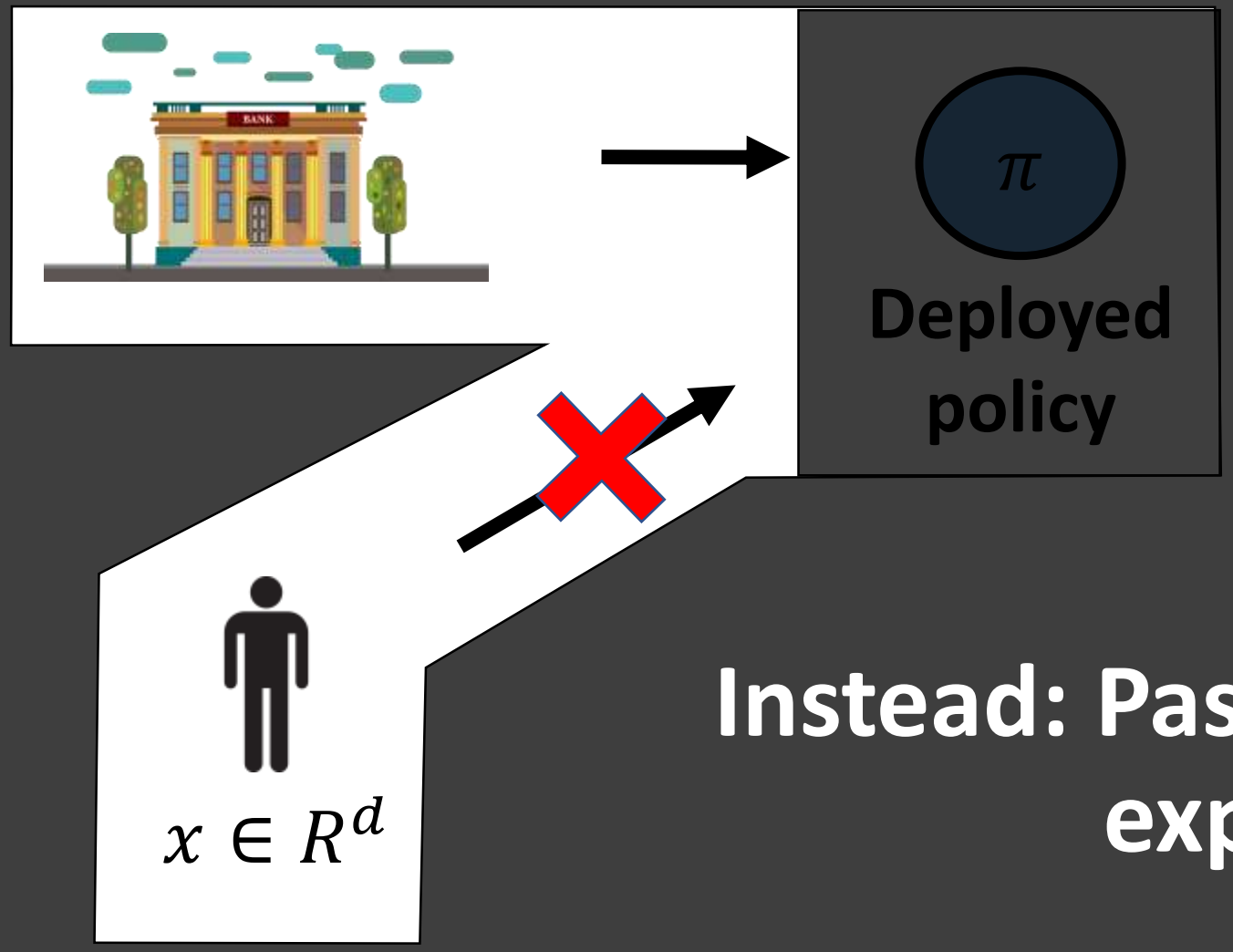
Key assumption: deployed policy is accessible.

Strategic feature modification



Often in reality:
Policy is <u>inaccessible</u>.

Strategic feature modification



Instead: Past labelled examples, explanations, …

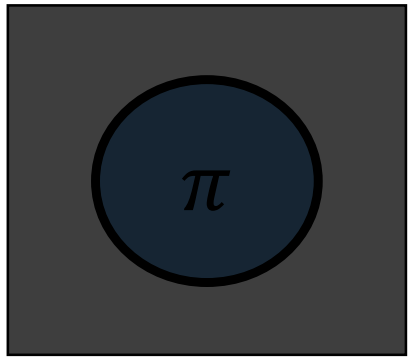
Our Work

Focus on strategic learning when decision rules are **inaccessible**.

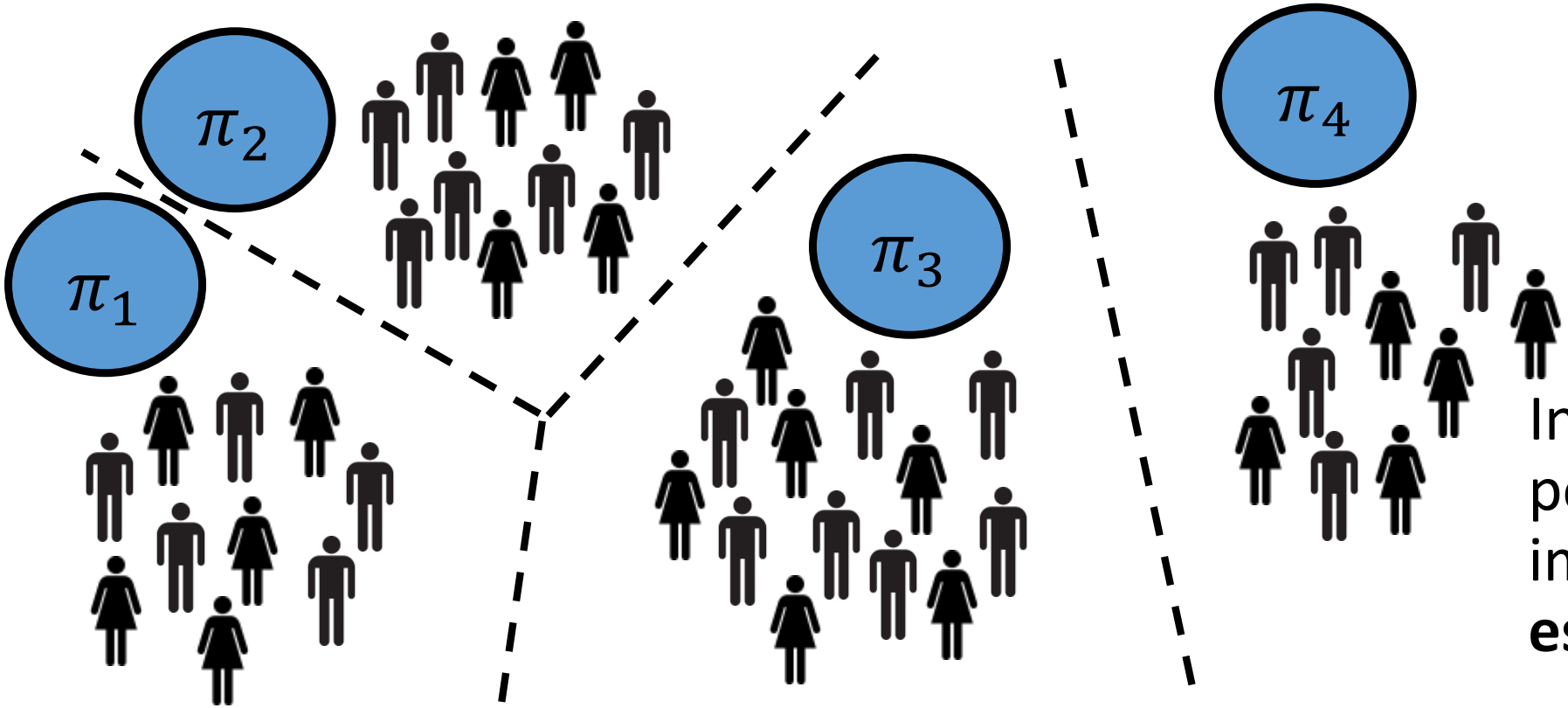
Model based on **learning from peers**.

Instead of observing the decision rule, individuals try to learn about it from friends, acquaintances **who applied previously**.

Strategic Learning with Inaccessible rules



Due to information discrepancy, different peer-networks may form **different estimates** of the deployed rule.



Individuals in each sub-population attempt to improve using **its own estimation** of the rule.

Strategic Learning with Inaccessible rules

Q: What are the effects of information discrepancy between different sub-populations on the ability of individuals to improve?

Adult Dataset

Publicly available at UCI repository.

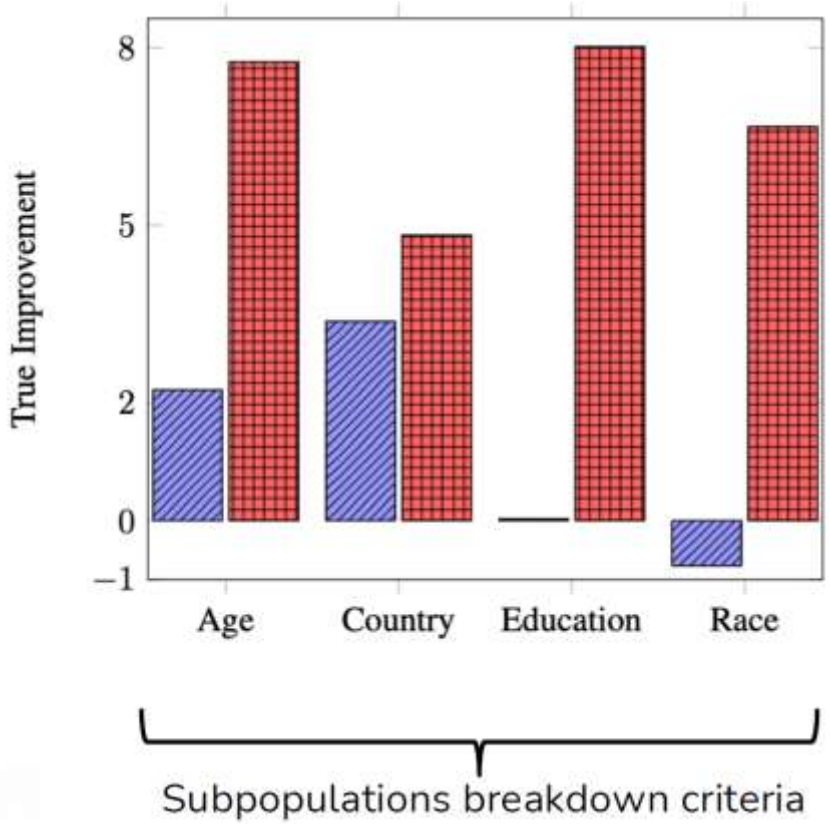
- ~50K datapoints
- 14 attributes including Age, Country, Workclass, Education, Race, etc.
- Label (annual income): = 50K Our process:
- 4 experiments separating subpopulations based on:

Characteristic	Subpopulation 1	Subpopulation 2
Age	<35 yrs old	>=35 yrs old
Country	All others	Western countries
Education	All others	Above high school
Race	All others	White

Predict **income improvement** (final income – original income) for each sub-population.

Results Snapshot: Adult Dataset

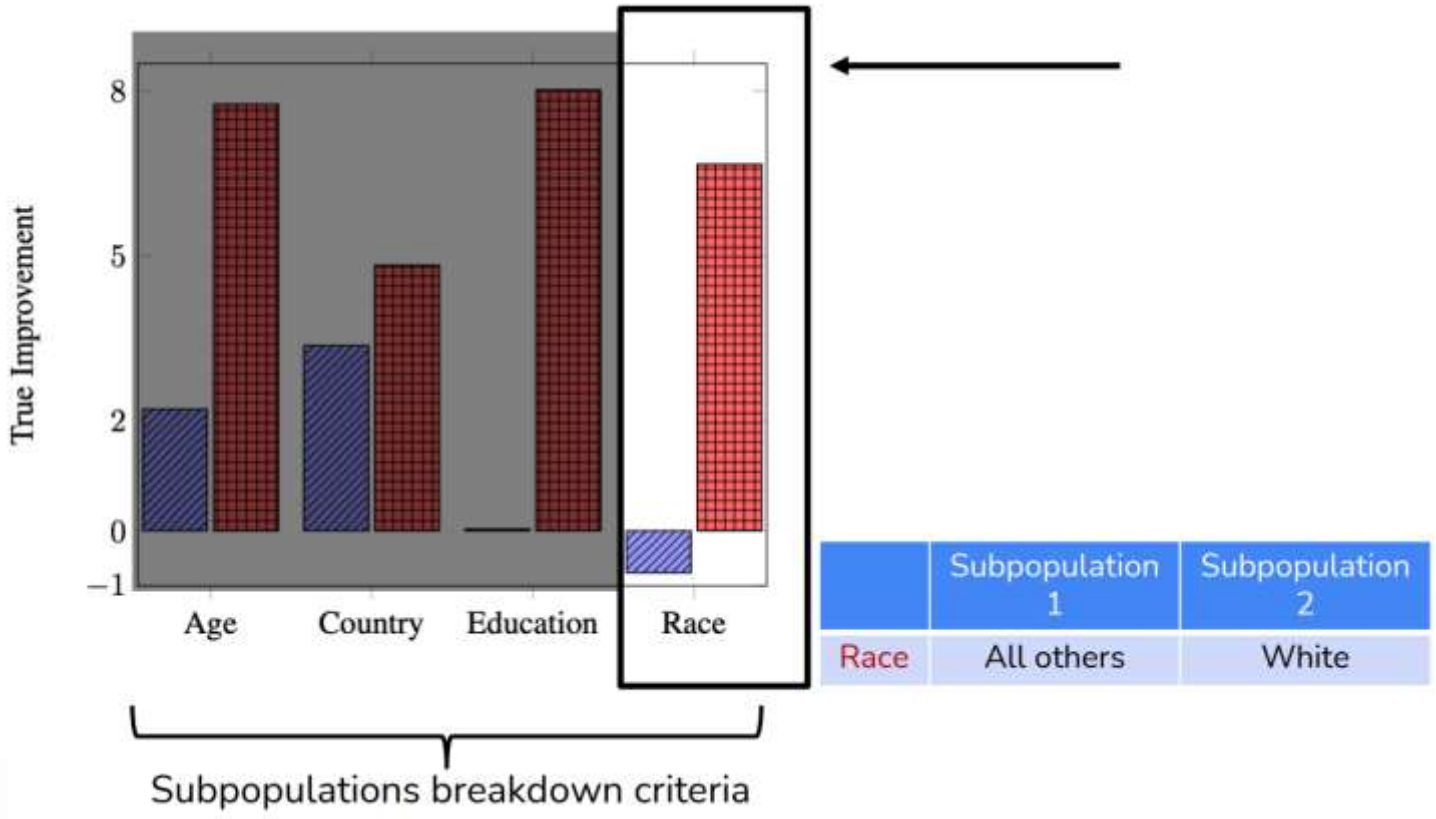
- Total income improvement currently subpopulation 1
- Total income improvement currently subpopulation 2



Results Snapshot: Adult Dataset

Sub-populations may end up **worse off**.

- Total income improvement currently subpopulation 1
- Total income improvement currently subpopulation 2

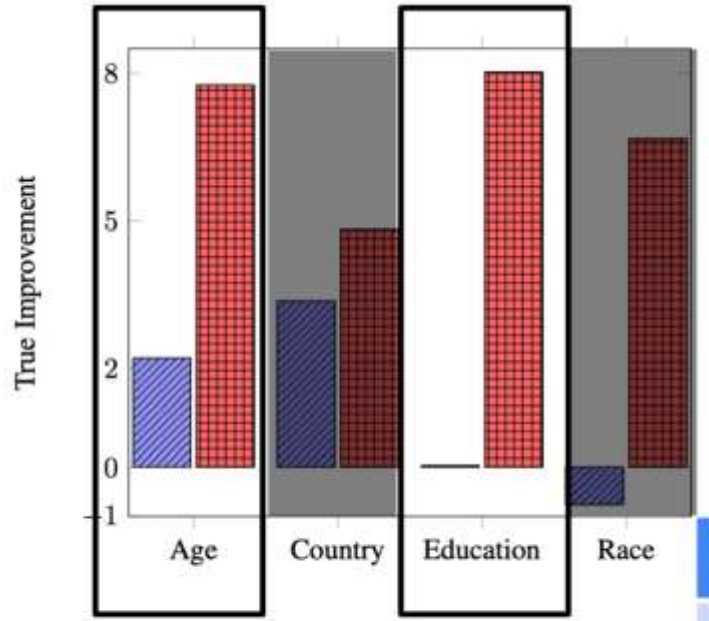


	Subpopulation 1	Subpopulation 2
Race	All others	White

Results Snapshot: Adult Dataset

Total improvement may be **very unequal** across sub-populations.

- Total income improvement currently subpopulation 1
- Total income improvement currently subpopulation 2



	Subpopulation 1	Subpopulation 2
Age	<35 yrs old	>=35 yrs old
Education	All others	Above HS

Results

We make explicit a connection between:

1. Information available to different sub-populations.
2. Ability of individuals to improve.

Theoretical characterizations for when, across all sub-populations:

1. Do-no-harm.
2. Equal improvements.
3. Effort is exerted optimally.

Information Discrepancy in Strategic Learning

Thank you!

Yahav Bechavod
Hebrew University

Chara Podimata
Harvard University

Steven Wu
Carnegie Mellon University

Juba Ziani
Georgia Institute of Technology