Paradoxes in Fair Machine Learning

Paul Gölz, Anson Kahng, and Ariel Procaccia

NeurIPS 2019

Research Question

What is the relationship between
fairness in machine learning and **fairness in fair division**?

Research Question

What is the relationship between
fairness in machine learning and **fairness in fair division**?

↑ Statistical notions of fairness
(e.g., equalized odds)

↑ Axioms of fair division
(e.g., resource monotonicity, population monotonicity)

RESEARCH QUESTION

What is the relationship between
fairness in machine learning and **fairness in fair division**?

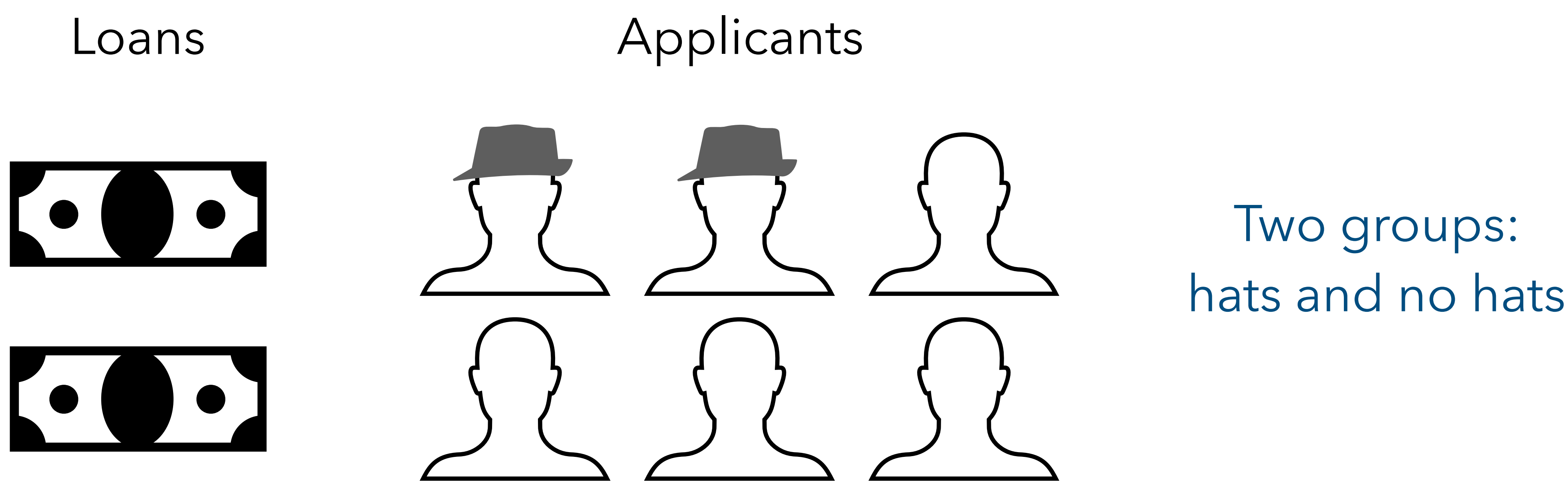
↑ Statistical notions of fairness
(e.g., equalized odds)

↑ Axioms of fair division
(e.g., resource monotonicity, population monotonicity)

In order to compare these, we need the right setting.

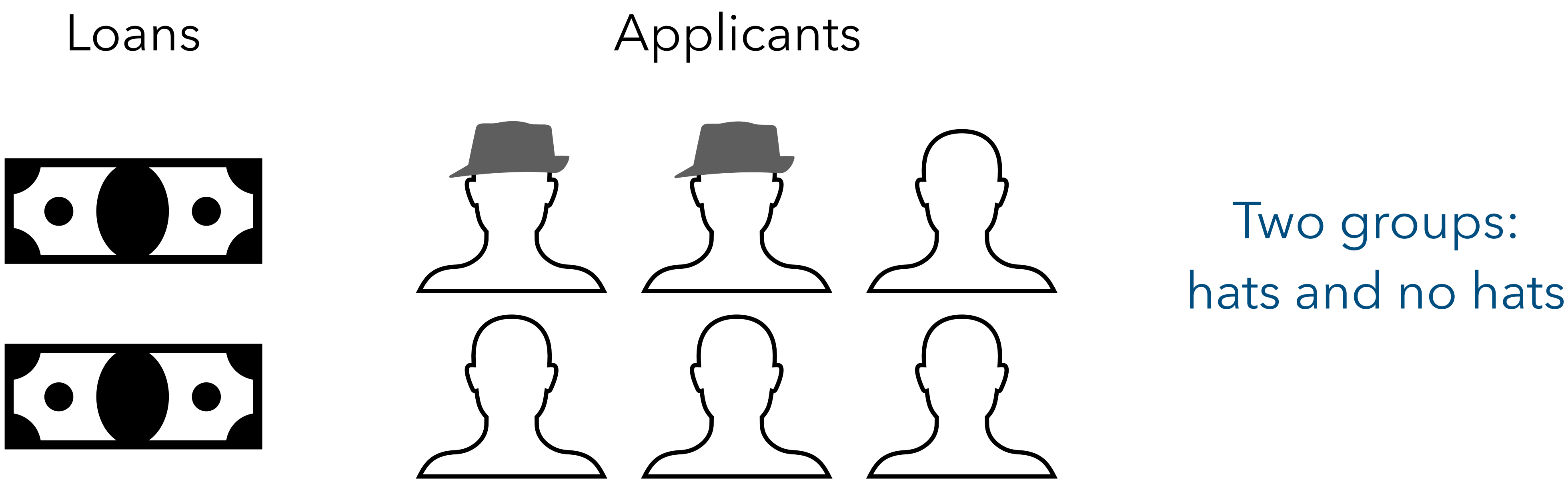
CLASSIFICATION WITH CARDINALITY CONSTRAINTS

Classification problem with a fixed budget of available resources to distribute: e.g., **financial aid**.



Goal: maximize **efficiency** (fraction of loans repaid)

CLASSIFICATION WITH CARDINALITY CONSTRAINTS



As a **classification** problem: label *k* applicants positively

As a **fair division** problem: divide *k* loans among applicants

What does it mean to be fair in each setting?

Fairness Concepts

Statistical Fairness	Fair Division Axioms
Equalized odds	Resource monotonicity
Demographic parity	Population monotonicity
	Consistency

Research question (rephrased):

How much does efficiency suffer if we must satisfy both equalized odds and various fair division axioms?

Statistical Fairness

Equalized Odds (EO):

"A predictor $\hat{Y}$ satisfies equalized odds with respect to a protected attribute A and outcome Y if $\hat{Y}$ and A are independent conditional on Y." (Hardt et al. 2016)

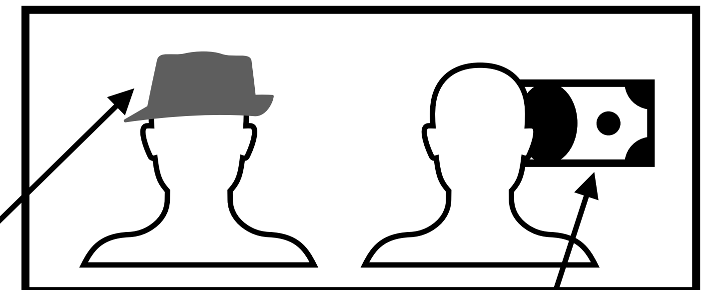
$$\Pr(\hat{Y} = 1 \,|\, A = 1, Y = 1) = \Pr(\hat{Y} = 1 \,|\, A = 0, Y = 1)$$

$$\Pr(\hat{Y} = 1 \,|\, A = 1, Y = 0) = \Pr(\hat{Y} = 1 \,|\, A = 0, Y = 0)$$

"True positive and false positive rates are equal across groups"

STATISTICAL FAIRNESS

Equalized Odds (EO):



"A predictor $\hat{Y}$ satisfies equalized odds with respect to a protected attribute A and outcome Y if $\hat{Y}$ and A are independent conditional on Y." (Hardt et al. 2016)

$$\Pr(\hat{Y} = 1 \mid A = 1, Y = 1) = \Pr(\hat{Y} = 1 \mid A = 0, Y = 1)$$

$$\Pr(\hat{Y} = 1 \mid A = 1, Y = 0) = \Pr(\hat{Y} = 1 \mid A = 0, Y = 0)$$

"True positive and false positive rates are equal across groups"

Fair Division Axioms

Resource monotonicity:

“Adding more resources makes everyone better off.”

Population monotonicity:

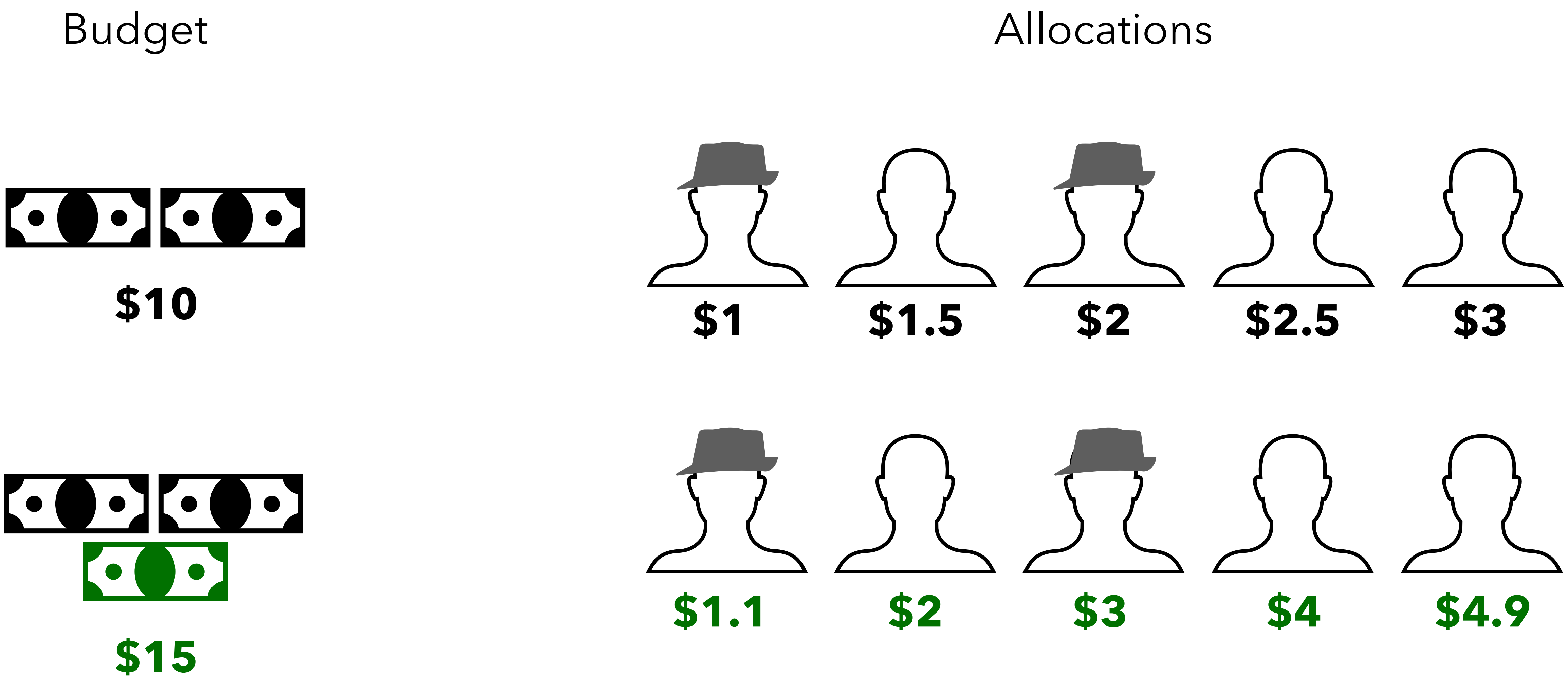
“Adding more people makes everyone worse off.”

Think of these axioms as preclusions of paradoxes.

Resource Monotonicity

"Adding more resources makes everyone weakly better off"

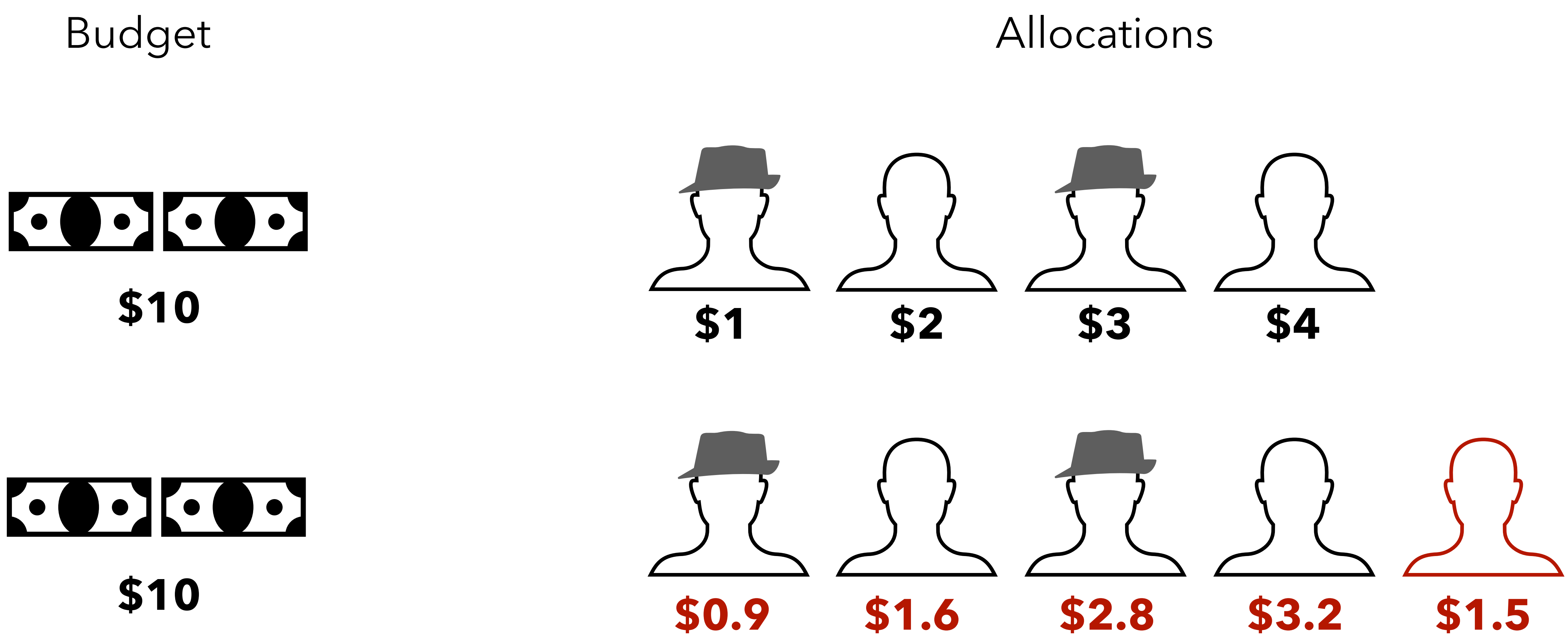
If the school gets more money, no one gets less allocated to them.



POPULATION MONOTONICITY

"Adding more people makes everyone weakly worse off"

If someone turns down aid, this can't hurt anyone else's allocation.



Results (Partial List)

1. In the cardinality-constrained model, we characterize the optimal allocation rule that satisfies equalized odds

2. Equalized odds and **resource monotonicity** are achievable with no loss to optimal EO efficiency

3. Any rule that satisfies equalized odds and **population monotonicity** cannot achieve a constant-factor approximation to optimal EO efficiency

RESULTS (PARTIAL LIST)

1. In the cardinality-constrained model, we characterize the optimal allocation rule that satisfies equalized odds

2. Equalized odds and **resource monotonicity** are achievable with no loss to optimal EO efficiency

3. Any rule that satisfies equalized odds and **population monotonicity** cannot achieve a constant-factor approximation to optimal EO efficiency

Thank you! Please come find me at poster #83.