MAKING DECISIONS THAT REDUCE DISCRIMINATORY IMPACT

Matt J. Kusner

Chris Russell

Joshua R. Loftus

Ricardo Silva

ML IS INVOLVED IN LIFE-CHANGING DECISIONS

Policing
[Ensign et al., 2017]

PredPol
Predict Crime in Real Time

Parole Sentencing
[Larson et al., 2016]

COMPAS

Advertising
[Sweeney, 2013]

Google

Insurance

H_2O.ai

Lending

underwrite.ai

Insert your favorite application here!

WE HAVE PROBLEMS

Page last updated at 10:35 GMT, Thursday, 24 December 2009

HP camera 'can't see' black faces

A YouTube video suggesting that face recognition cameras installed in HP laptops cannot detect black faces has had over one million views.

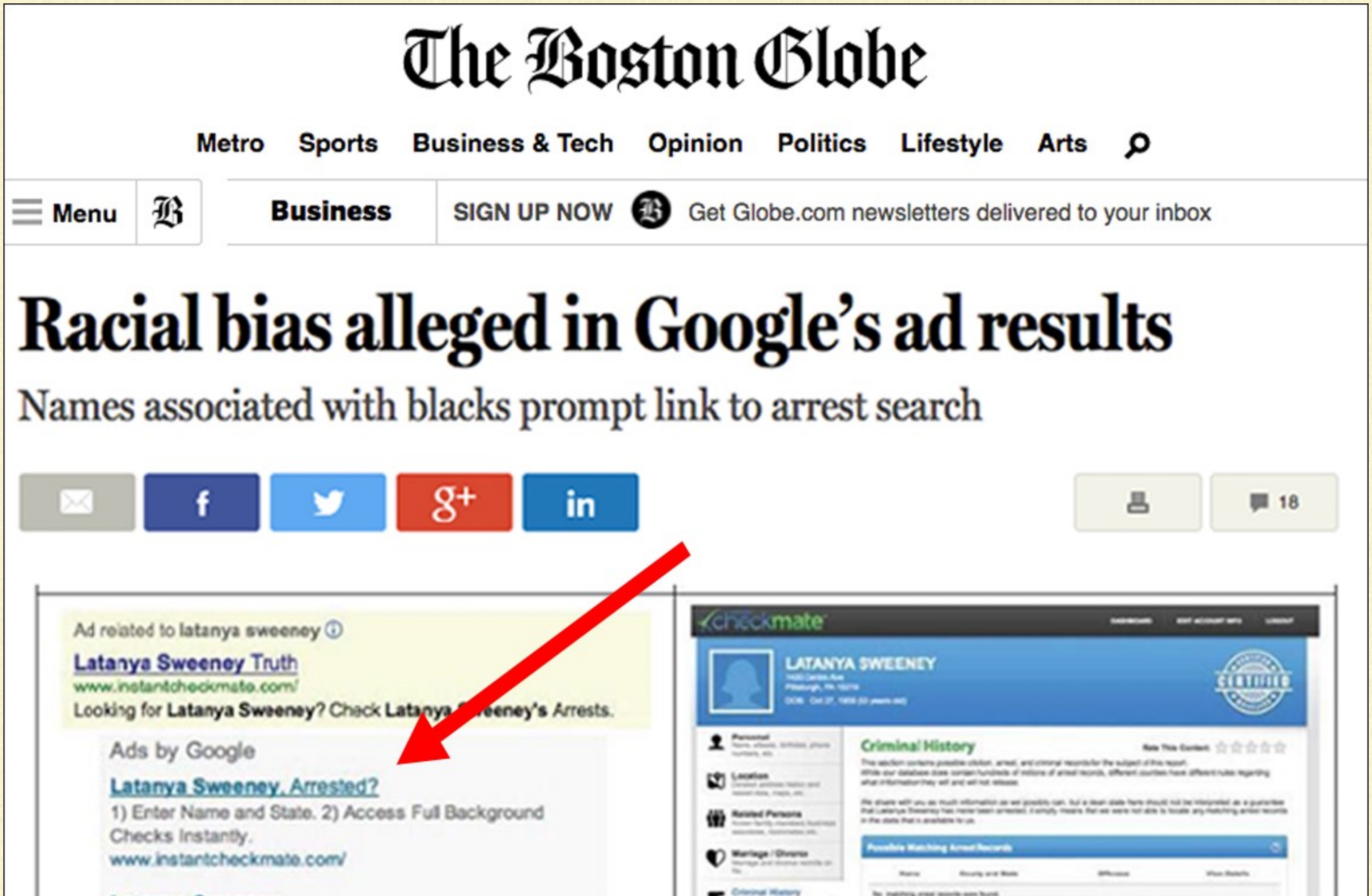
"Black Desi" in the YouTube video

The short movie, uploaded earlier this month, features "Black Desi" and his colleague "White Wanda".

When Wanda, a white woman, is in front of the screen, the camera zooms to her face and moves as she moves.

ML CAN BE RACIST…

[SWEENEY, 2013]



ML CAN BE SEXIST…

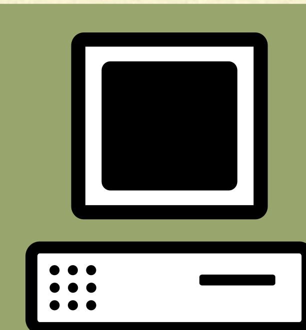
[BOLUKBASI ET AL. 2016]

DEFINITIONS OF FAIRNESS

Fairness Through Unawareness	Equality of Opportunity [Hardt et al., 2016]
Individual Fairness [Dwork et al., 2012]	Demographic Parity [Zemel et al., 2013; Zliobaite, 2015]
Fair Calibration [Pleiss et al., 2017]	Preference Fairness [Zafar et al., 2017]
Counterfactual Fairness [Kusner et al., 2017]	Path-Specific Fairness [Shpitser et al., 2017; Chiappa et al., 2018]

PROBLEM #1

The Discriminatory **Prediction** Problem



$$\hat{Y} : \mathcal{X} \longrightarrow Y$$

ML IS INVOLVED IN LIFE-CHANGING DECISIONS

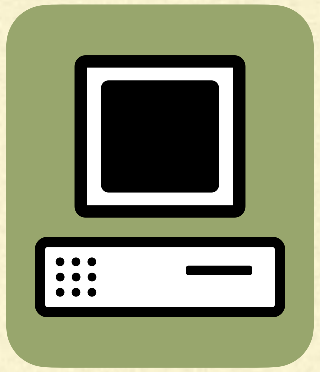
Policing
[Ensign et al., 2017]

PredPol
Predict Crime in Real Time

Parole Sentencing
[Larson et al., 2016]

COMPAS

Advertising



Insurance

H_2O.ai

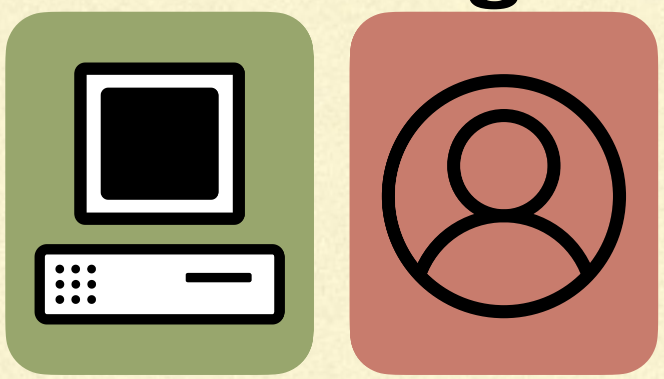
Lending

underwrite.ai

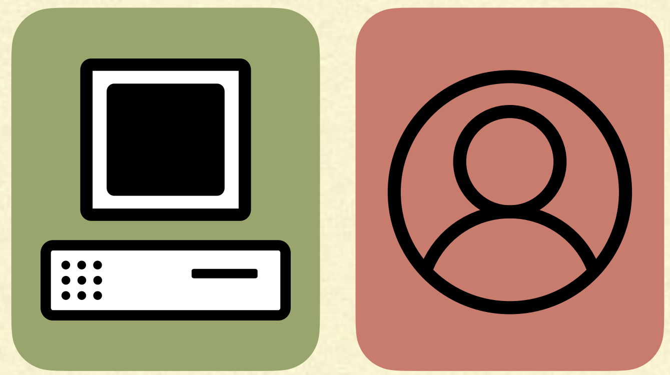
Insert your favorite application here!

ML IS INVOLVED IN LIFE-CHANGING DECISIONS

Policing

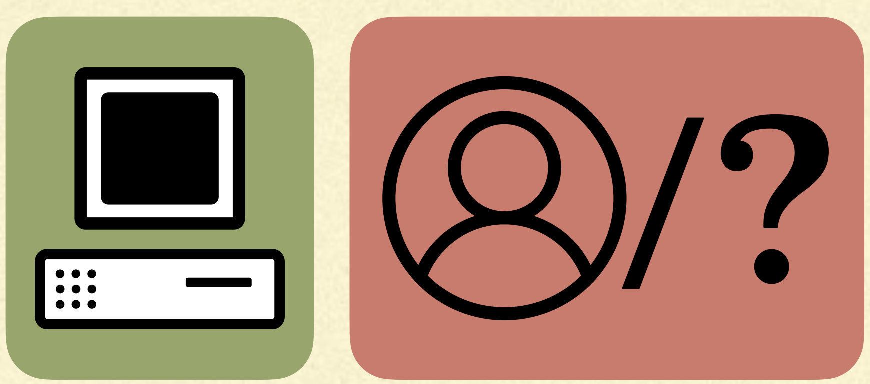


Parole Sentencing

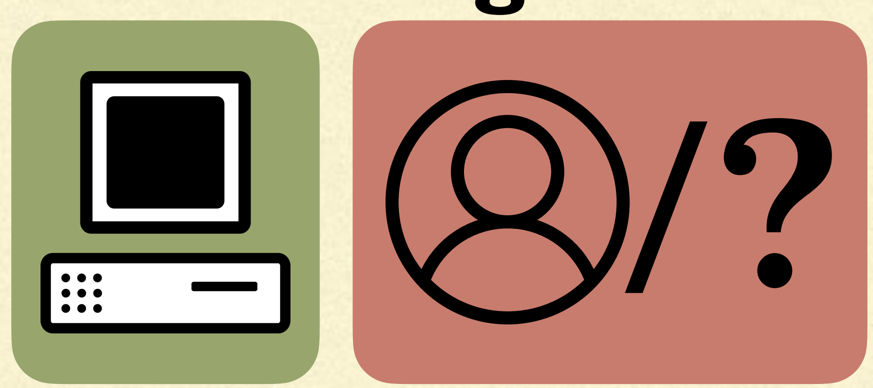


Advertising

Insurance



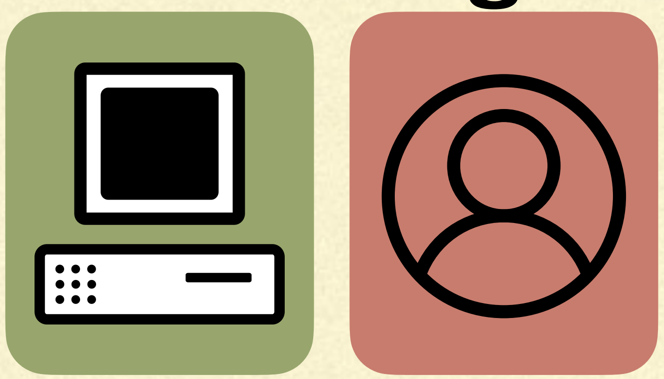
Lending



Insert your favorite application here!

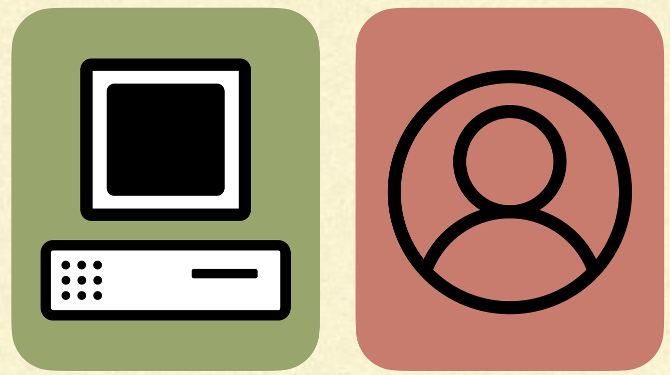
ML IS INVOLVED IN LIFE-CHANGING DECISIONS

Policing



impact: arrest

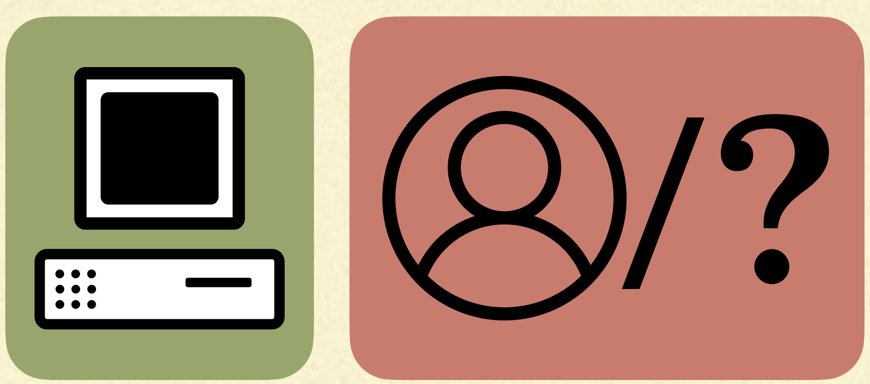
Parole Sentencing



impact: jail-time

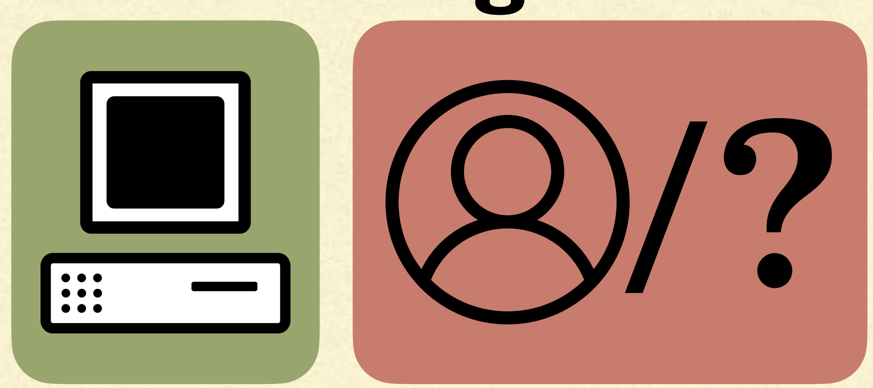
Advertising

Insurance



impact: improved health

Lending

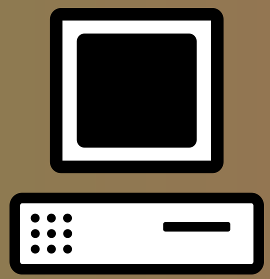


impact: pay off home loans

Insert your favorite application here!

PROBLEM #2

The Discriminatory **Impact** Problem



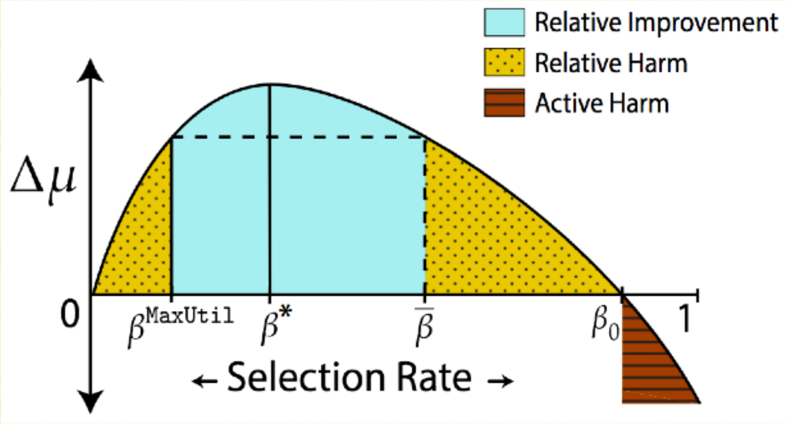
$$\hat{Y} : \mathcal{X} \longrightarrow Z \overset{/?}{\rightsquigarrow} Y$$

RELATED WORK

Introduction of discriminatory
impact problem

[Liu et al., ICML 2018]

[Green & Chen, FAT* 2019]

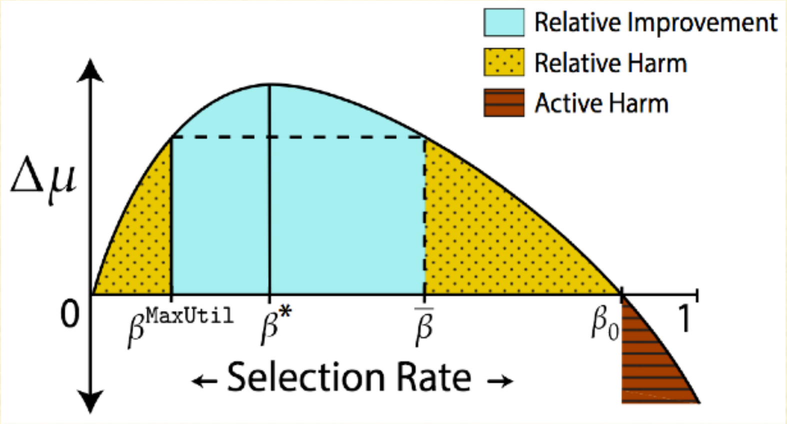


RELATED WORK

Introduction of discriminatory impact problem

[Liu et al., ICML 2018]

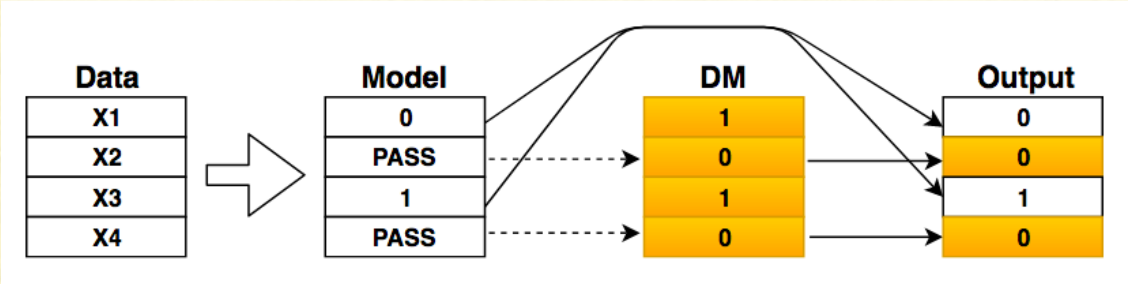
[Green & Chen, FAT* 2019]



Models for special cases

[Madras et al., NeurIPS 2018]

[Kannan et al., FAT* 2019]

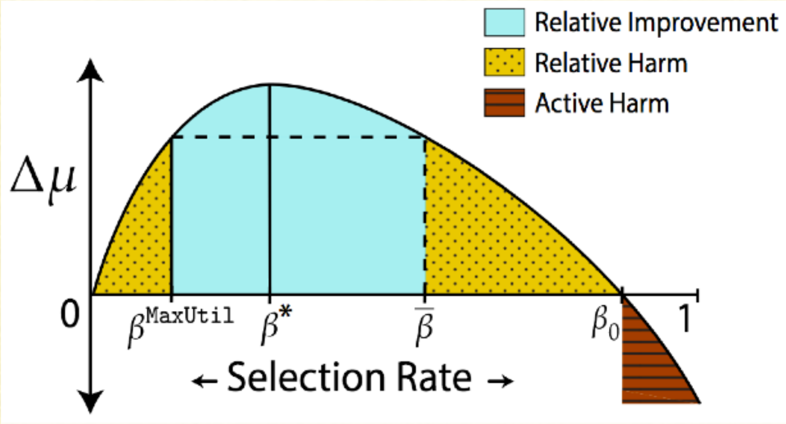


RELATED WORK

Introduction of discriminatory **impact** problem

[Liu et al., ICML 2018]

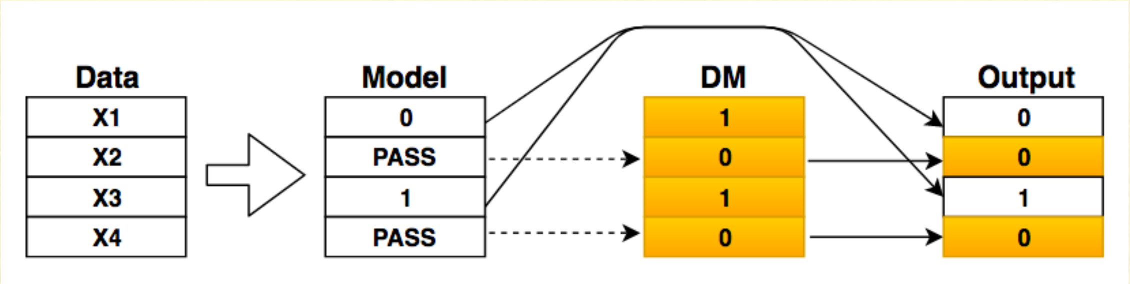
[Green & Chen, FAT* 2019]



Models for special cases

[Madras et al., NeurIPS 2018]

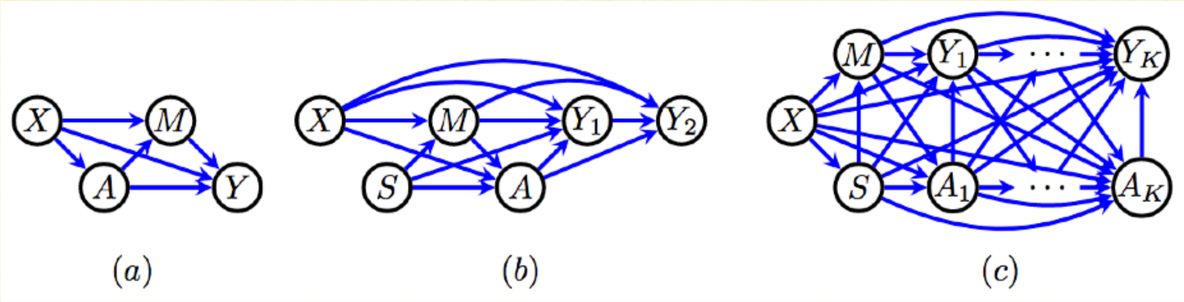
[Kannan et al., FAT* 2019]



Complimentary approaches (RL, social dynamics)

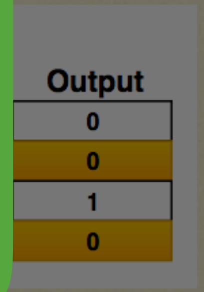
[Nabi et al., ICML 2019]

[Hedari et al., ICML 2019]



RELATED WORK

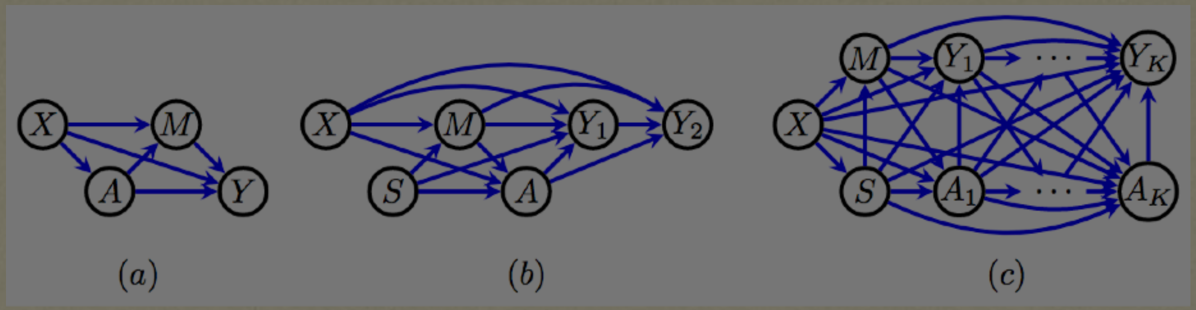
our work: a general framework for *reducing discriminatory impact* based on causal modeling and MILP



<u>Complimentary approaches (social dynamics, RL)</u>

[Nabi et al., ICML 2019]

[Hedari et al., ICML 2019]



high school 1

high school 2

high school 1

high school 2

intervention: fund advanced classes

impact: to increase college applications (% SAT/ACT-taking)

high school 1

high school 2

race distribution

$A^{(1)}$

race distribution

$A^{(2)}$

$X^{(1)}$

counselors

$Y^{(1)}$

% SAT/ACT-taking

$Y^{(2)}$

% SAT/ACT-taking

$X^{(2)}$

counselors

high school 1

race distribution

$A^{(1)}$

intervention: calculus classes

$Z^{(1)}$

$X^{(1)}$

counselors

$Y^{(1)}$

% SAT/ACT-taking

high school 2

intervention: calculus classes

$Z^{(2)}$

race distribution

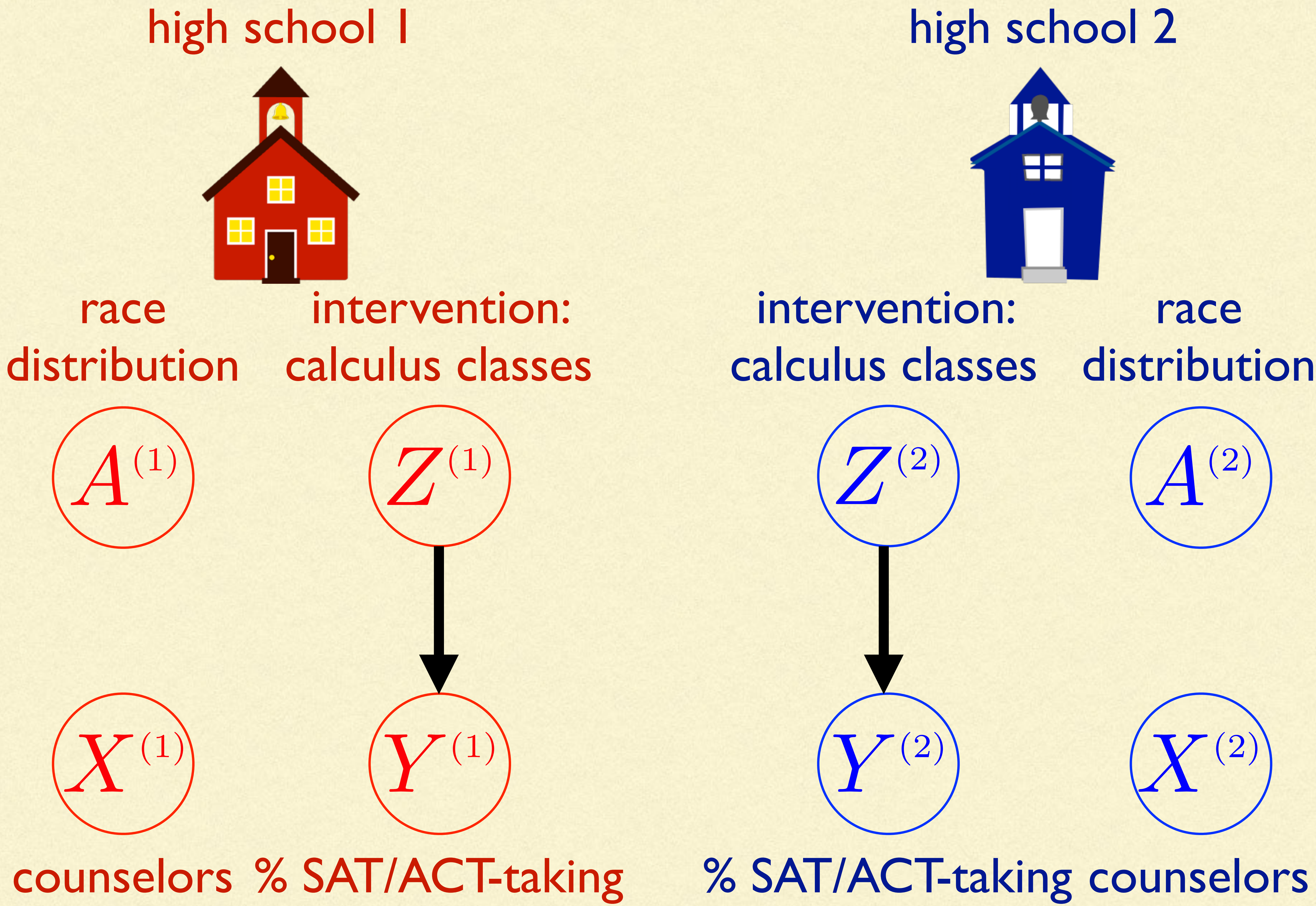
$A^{(2)}$

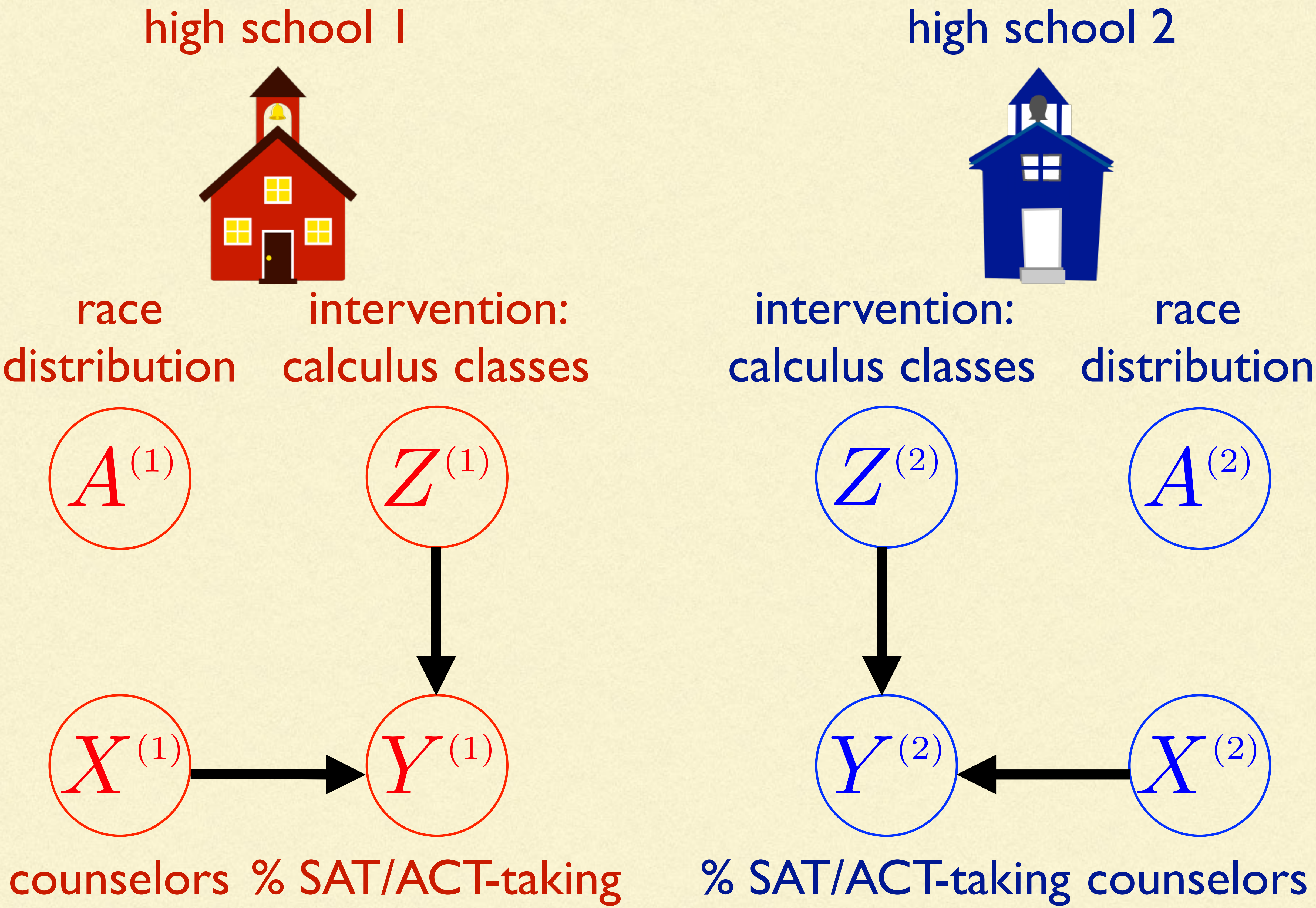
$Y^{(2)}$

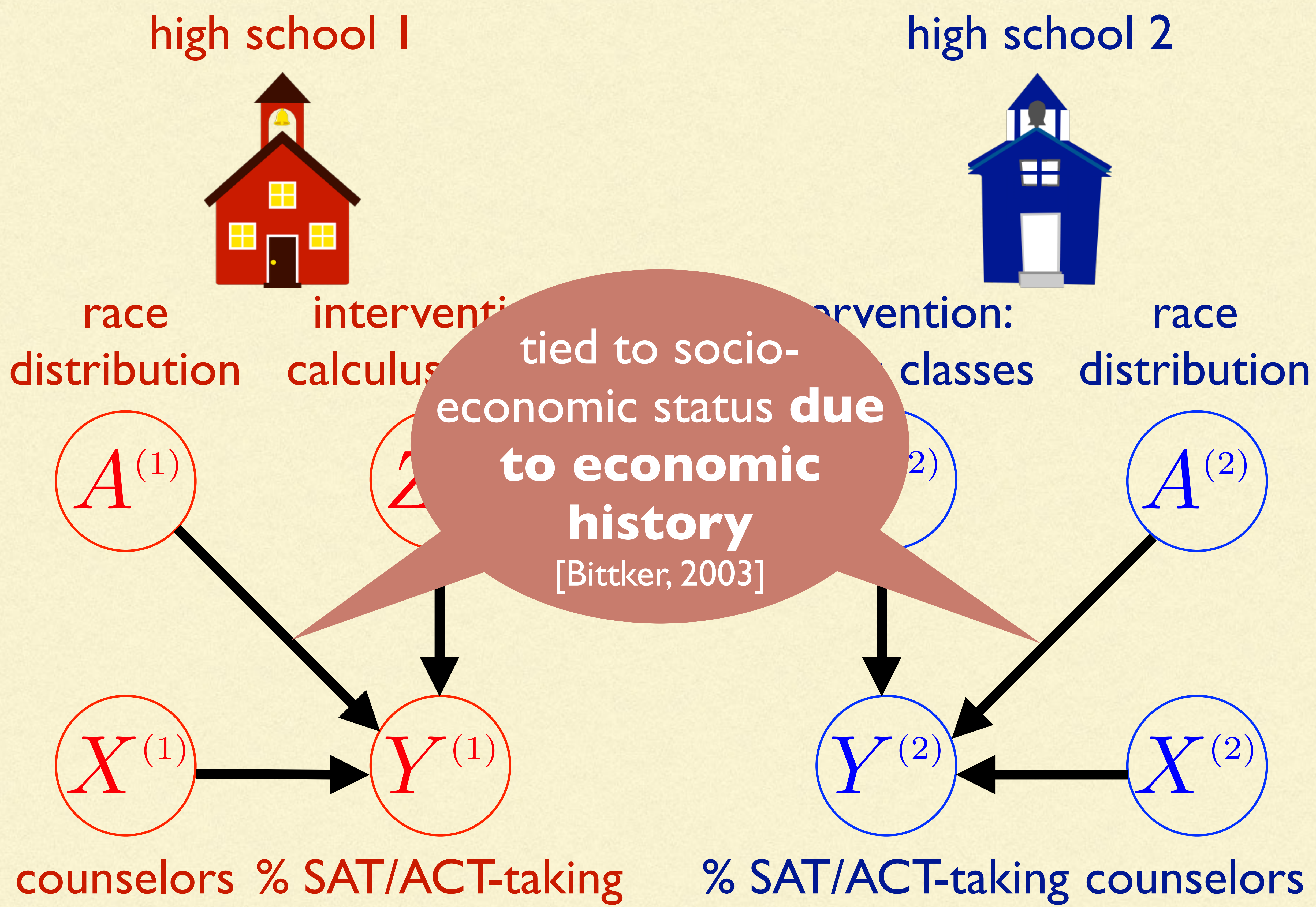
% SAT/ACT-taking

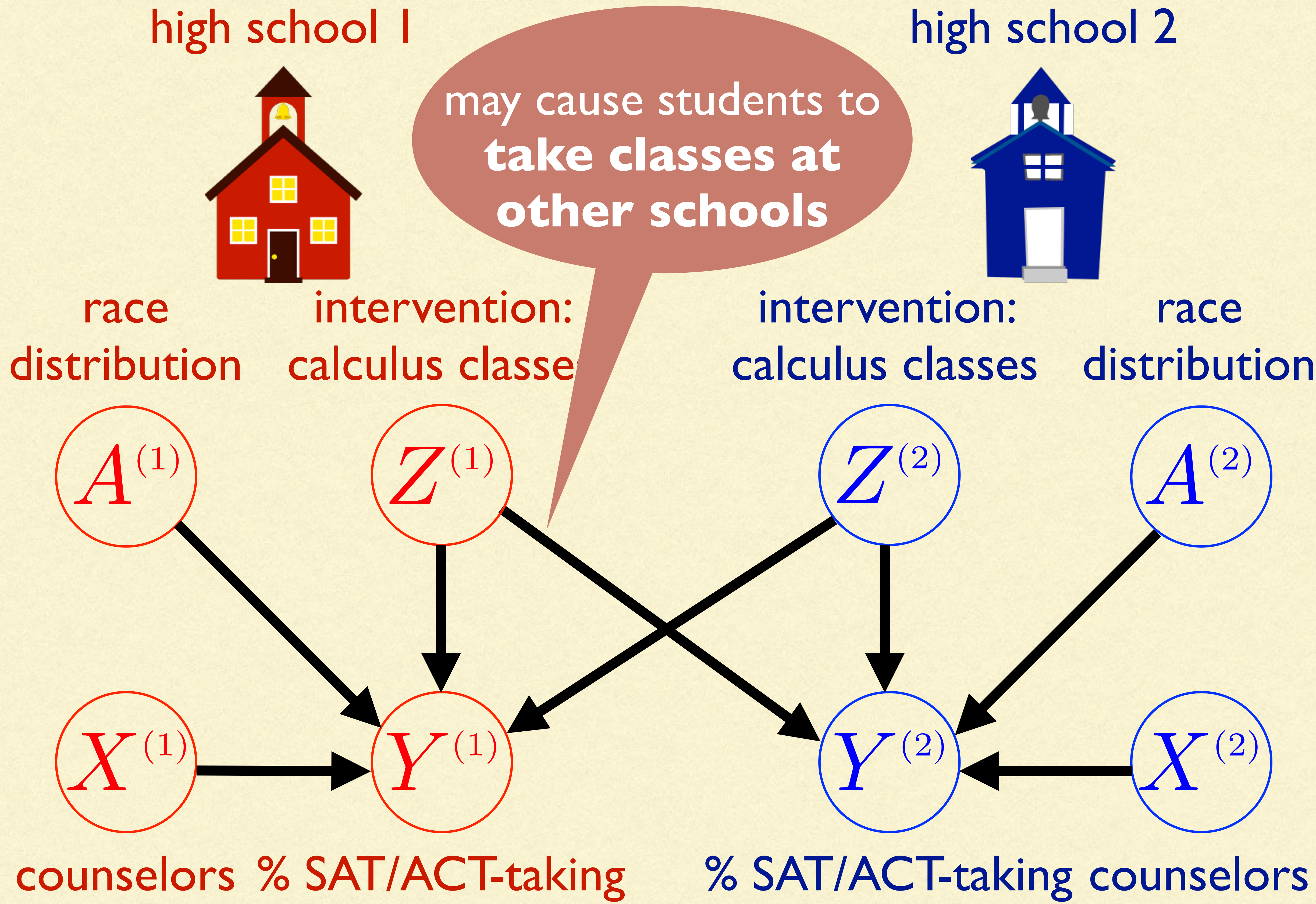
$X^{(2)}$

counselors

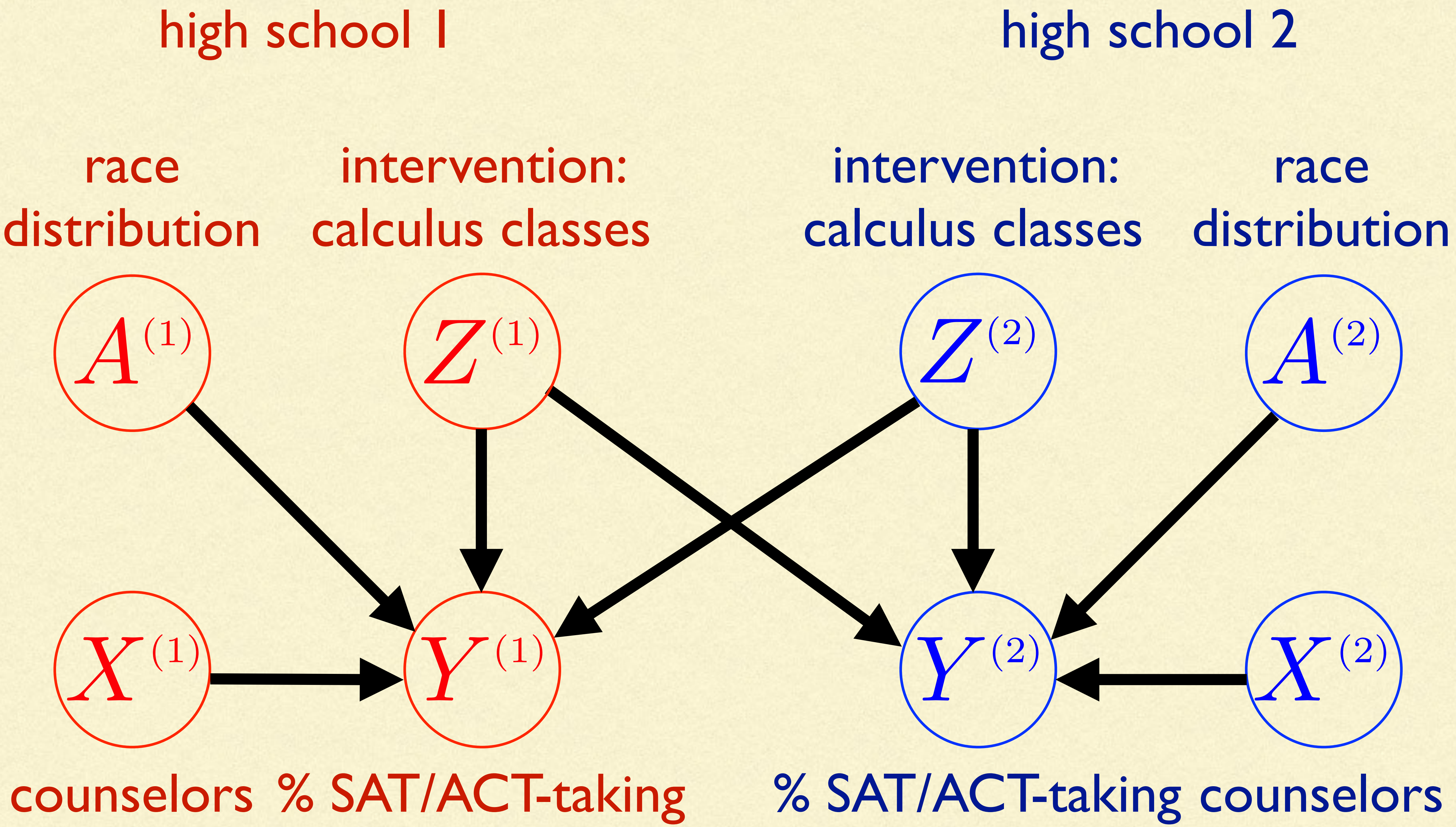

high school 1
high school 2
race distribution
intervention: calculus classes
intervention: calculus classes
race distribution
$A^{(1)}$
$Z^{(1)}$
$Z^{(2)}$
$A^{(2)}$
$X^{(1)}$
$Y^{(1)}$
$Y^{(2)}$
$X^{(2)}$
counselors
% SAT/ACT-taking
% SAT/ACT-taking
counselors

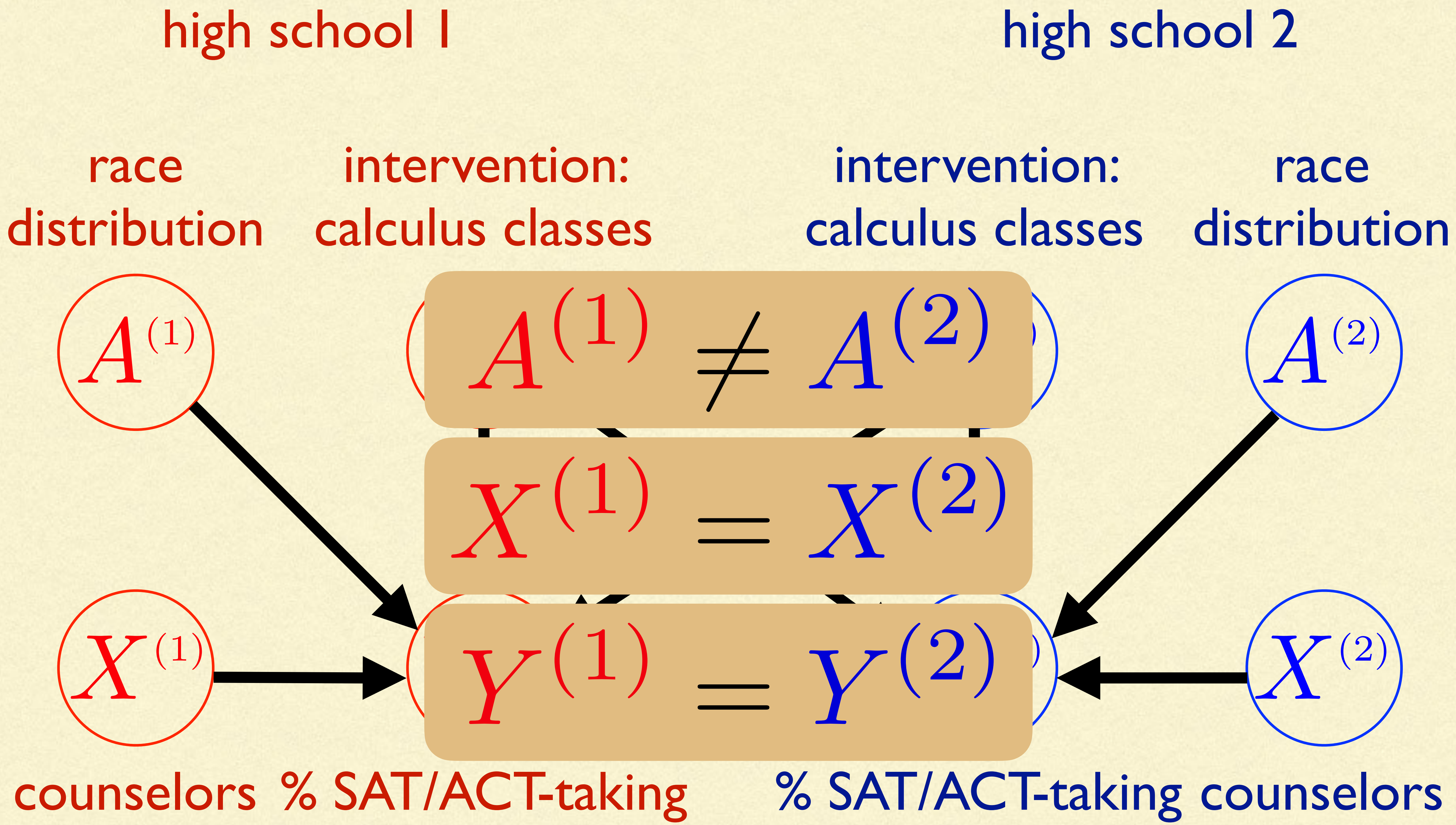

high school 1
high school 2
race distribution
intervention: calculus classes
intervention: calculus classes
race distribution
$A^{(1)}$
$Z^{(1)}$
$Z^{(2)}$
$A^{(2)}$
$X^{(1)}$
$Y^{(1)}$
$Y^{(2)}$
$X^{(2)}$
counselors
% SAT/ACT-taking
% SAT/ACT-taking
counselors


high school 1
high school 2
race distribution
interventi
calculus
rvention:
classes
race distribution
$A^{(1)}$
$A^{(2)}$
$X^{(1)}$
$Y^{(1)}$
$Y^{(2)}$
$X^{(2)}$
counselors
% SAT/ACT-taking
% SAT/ACT-taking
counselors
tied to socio-economic status **due to economic history**
[Bittker, 2003]



high school 1
high school 2
may cause students to **take classes at other schools**
race distribution
intervention: calculus classes
intervention: calculus classes
race distribution
$A^{(1)}$
$Z^{(1)}$
$Z^{(2)}$
$A^{(2)}$
$X^{(1)}$
$Y^{(1)}$
$Y^{(2)}$
$X^{(2)}$
counselors
% SAT/ACT-taking
% SAT/ACT-taking
counselors

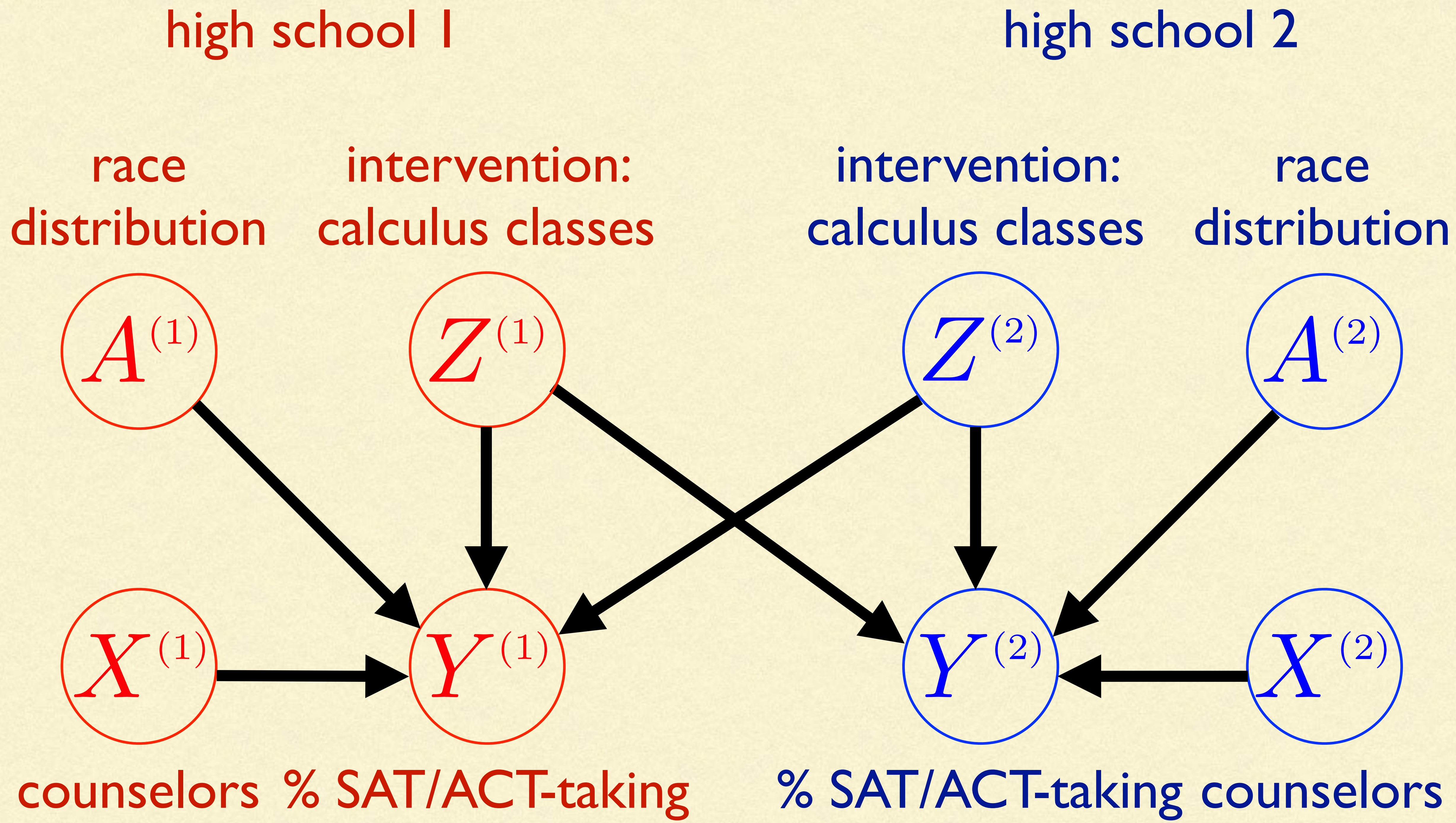

high school 1
high school 2
race distribution
intervention: calculus classes
intervention: calculus classes
race distribution
$A^{(1)}$
$Z^{(1)}$
$Z^{(2)}$
$A^{(2)}$
$X^{(1)}$
$Y^{(1)}$
$Y^{(2)}$
$X^{(2)}$
counselors
% SAT/ACT-taking
% SAT/ACT-taking
counselors



$$A^{(1)} \neq A^{(2)}$$

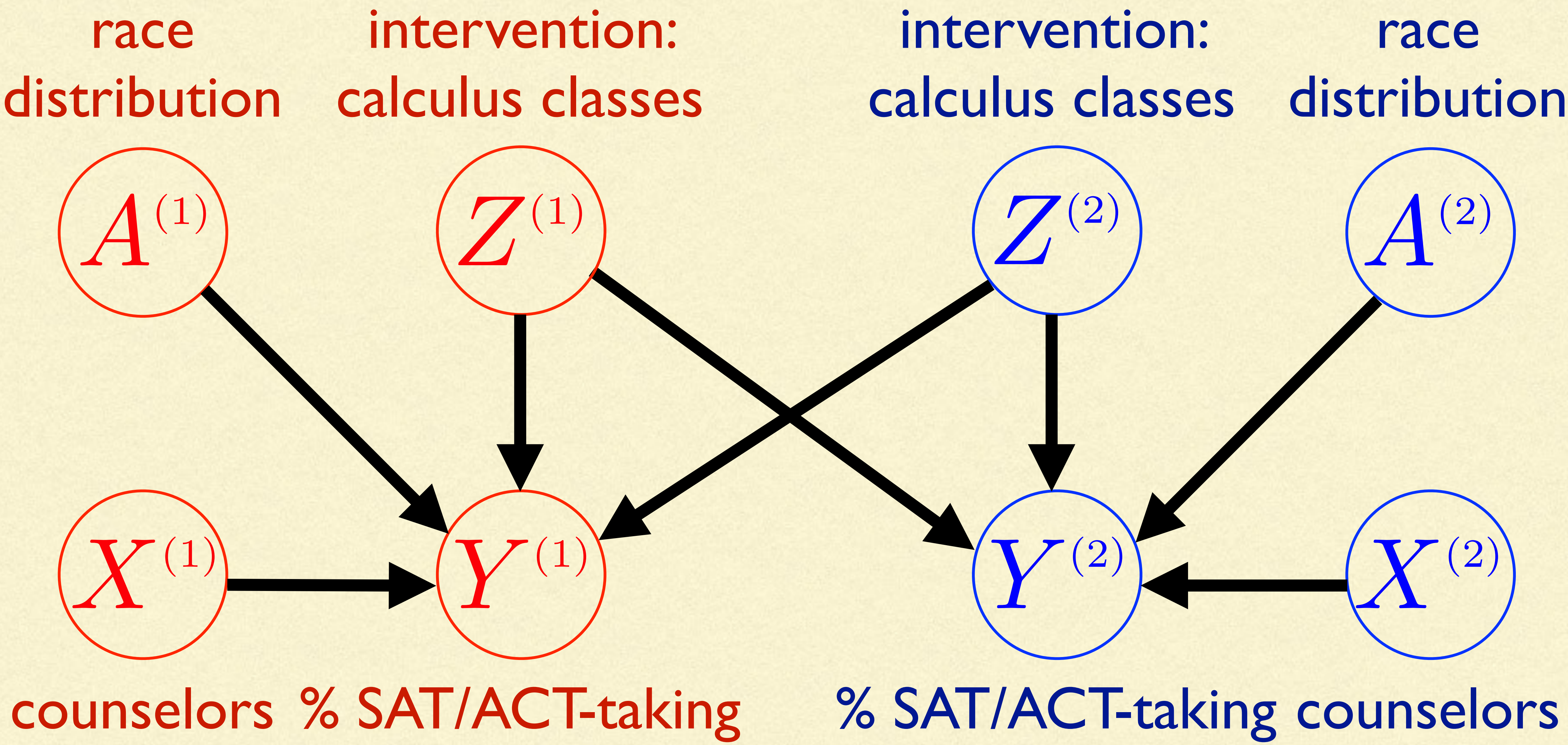
$$X^{(1)} = X^{(2)}$$

$$Y^{(1)} = Y^{(2)}$$


high school 1
high school 2
race distribution
intervention: calculus classes
intervention: calculus classes
race distribution
$A^{(1)}$
$Z^{(1)}$
$Z^{(2)}$
$A^{(2)}$
$X^{(1)}$
$Y^{(1)}$
$Y^{(2)}$
$X^{(2)}$
counselors
% SAT/ACT-taking
% SAT/ACT-taking
counselors

$$Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) = 1.0$$

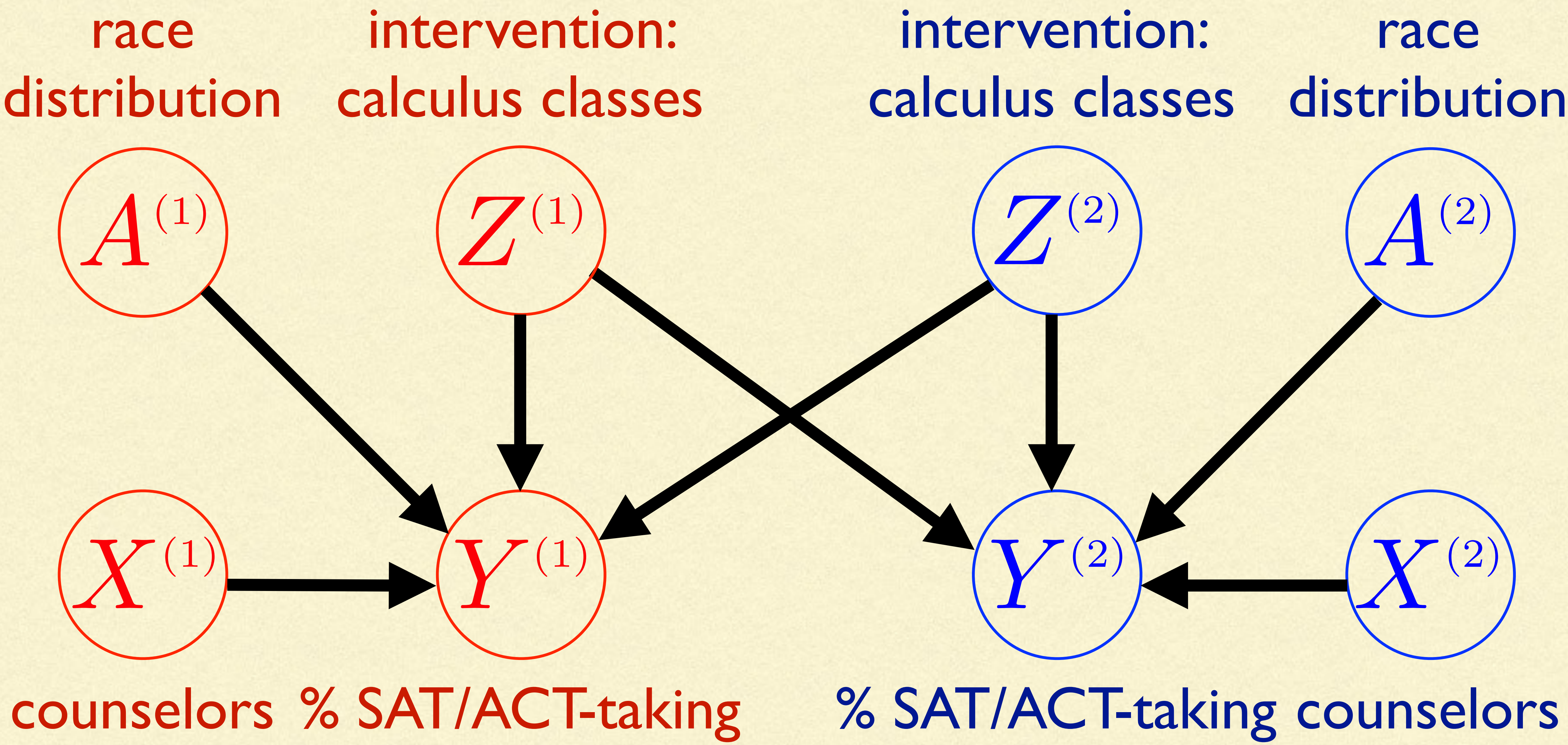
high school 1 | high school 2



$Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) = 1.0$

$Y^{(2)}([z^{(1)} = 1, z^{(2)} = 0]) = 0.1$

high school 1 high school 2

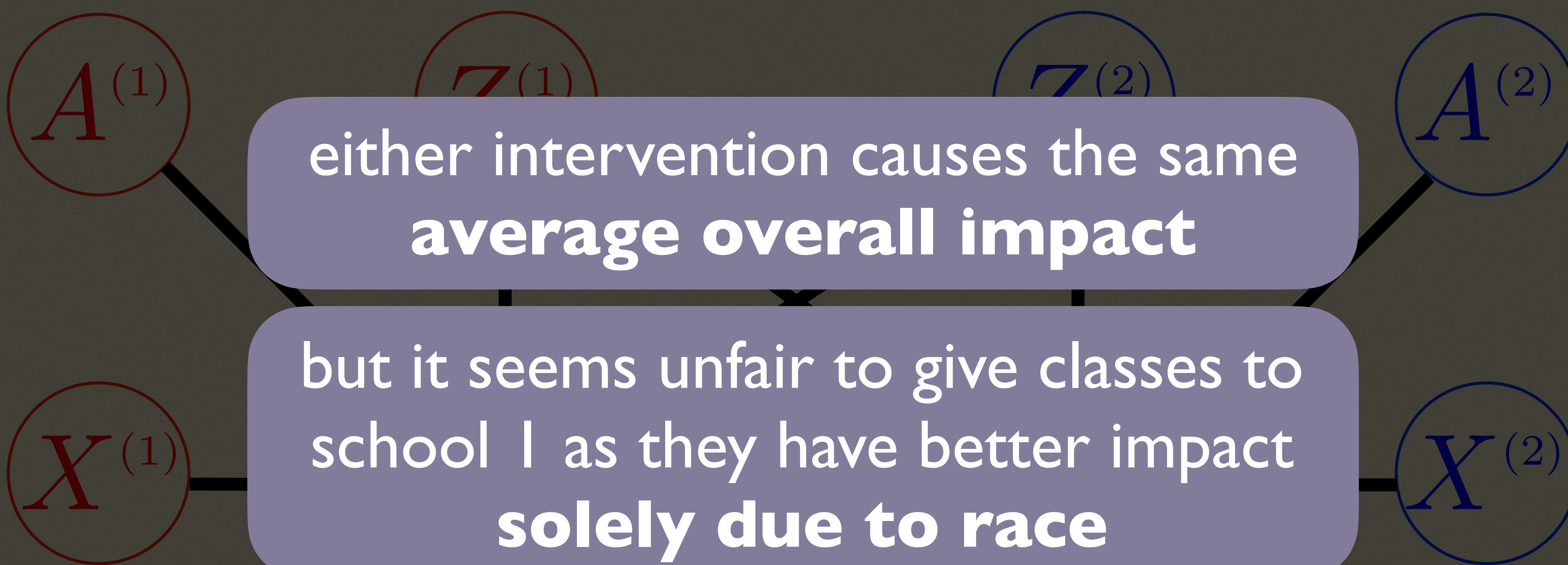


$Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) = 1.0$ $Y^{(2)}([z^{(1)} = 1, z^{(2)} = 0]) = 0.1$

$Y^{(1)}([z^{(1)} = 0, z^{(2)} = 1]) = 0.6$ $Y^{(2)}([z^{(1)} = 0, z^{(2)} = 1]) = 0.5$

high school 1 | high school 2

race distribution | decision: calculus classes | decision: calculus classes | race distribution



either intervention causes the same **average overall impact**

but it seems unfair to give classes to school 1 as they have better impact **solely due to race**

counselors | % SAT/ACT-taking | % SAT/ACT-taking | counselors

$Y^{(1)}([z^{(1)}=1, z^{(2)}=0]) = 1.0$

$Y^{(1)}([z^{(1)}=0, z^{(2)}=1]) = 0.6$

$Y^{(2)}([z^{(1)}=1, z^{(2)}=0]) = 0.1$

$Y^{(2)}([z^{(1)}=0, z^{(2)}=1]) = 0.5$

MORE FORMALLY

Maximizing overall impact

$$\max_{\mathbf{z} \in \{0,1\}^n} \sum_{i=1}^{n} \mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}],$$

$$s.t., \sum_{i=1}^{n} z^{(i)} \leq b$$

MORE FORMALLY

whether to grant advanced classes

# of schools

Maximizing overall impact

$$\max_{\mathbf{z}\in\{0,1\}^n} \sum_{i=1}^{n} \mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)}=a^{(i)}, X^{(i)}=\mathbf{x}^{(i)}]$$

$$s.t., \sum_{i=1}^{n} z^{(i)} \leq b$$

government budget

QUANTIFYING PRIVILEGE

$$\mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}]$$

$$\mathbb{E}[Y^{(i)}(a', \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}]$$

QUANTIFYING PRIVILEGE

$$\mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}]$$
$$\mathbb{E}[Y^{(i)}(a', \mathbf{z}) \quad \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}]$$

a counterfactual:
impact school i *would have gotten* for interventions **z** if race distribution was a’

QUANTIFYING PRIVILEGE

$$\mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}]$$
$$-\mathbb{E}[Y^{(i)}(a', \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}] < \tau$$

a counterfactual: impact school i *would have gotten* for interventions **z** if race distribution was a’

QUANTIFYING PRIVILEGE

constraint on **counterfactual privilege**

$$\mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}] \\ - \mathbb{E}[Y^{(i)}(a', \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}] < \tau$$

a counterfactual: impact school i *would have gotten* for interventions **z** if race distribution was a'

MORE FORMALLY

Maximizing impact with privilege constraints

$$\max_{\mathbf{z}\in\{0,1\}^n} \sum_{i=1}^{n} \mathbb{E}[Y^{(i)}(a^{(i)},\mathbf{z}) \mid A^{(i)}=a^{(i)}, X^{(i)}=\mathbf{x}^{(i)}]$$

$$s.t., \sum_{i=1}^{n} z^{(i)} \leq b$$

$$c_{ia'} \leq \tau \quad \forall a' \in \mathcal{A},\ i \in \{1,\ldots,n\},$$

$$\begin{aligned} &\mathbb{E}[Y^{(i)}(a^{(i)},\mathbf{z}) \mid A^{(i)}=a^{(i)}, X^{(i)}=\mathbf{x}^{(i)}] \\ -&\mathbb{E}[Y^{(i)}(a',\mathbf{z}) \mid A^{(i)}=a^{(i)}, X^{(i)}=\mathbf{x}^{(i)}] \end{aligned}$$

$$\mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}]$$
$$- \mathbb{E}[Y^{(i)}(a', \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}] < \tau$$

$$\tau = 0$$

$$Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) = 1.0$$
$$Y^{(1)}([z^{(1)} = 0, z^{(2)} = 1]) = 0.6$$

high school 1

$$Y^{(2)}([z^{(1)} = 1, z^{(2)} = 0]) = 0.1$$
$$Y^{(2)}([z^{(1)} = 0, z^{(2)} = 1]) = 0.5$$

high school 2

$$\mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}]$$
$$-\mathbb{E}[Y^{(i)}(a', \mathbf{z}) \quad \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}] < \tau$$

$$\tau = 0$$

essentially counterfactuals!

$Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) = 1.0$

$Y^{(1)}([z^{(1)} = 0, z^{(2)} = 1]) = 0.6$

high school 1

$Y^{(2)}([z^{(1)} = 1, z^{(2)} = 0]) = 0.1$

$Y^{(2)}([z^{(1)} = 0, z^{(2)} = 1]) = 0.5$

high school 2

$$\mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}]$$
$$-\mathbb{E}[Y^{(i)}(a', \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}] < \tau$$

$$\tau = 0$$

school 1
gets classes

$$Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) - Y^{(2)}([z^{(1)} = 1, z^{(2)} = 0]) = 0.9$$

$$Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) = 1.0$$

$$Y^{(1)}([z^{(1)} = 0, z^{(2)} = 1]) = 0.6$$

high school 1

$$Y^{(2)}([z^{(1)} = 1, z^{(2)} = 0]) = 0.1$$

$$Y^{(2)}([z^{(1)} = 0, z^{(2)} = 1]) = 0.5$$

high school 2

$$\mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}]$$
$$-\mathbb{E}[Y^{(i)}(a', \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}] < \tau$$

$$\tau = 0$$

school 1
gets classes

$$Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) - Y^{(2)}([z^{(1)} = 1, z^{(2)} = 0]) = 0.9$$

school 2
gets classes

$$Y^{(2)}([z^{(1)} = 0, z^{(2)} = 1]) - Y^{(1)}([z^{(1)} = 0, z^{(2)} = 1]) = -0.1$$

$$Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) = 1.0$$

$$Y^{(1)}([z^{(1)} = 0, z^{(2)} = 1]) = 0.6$$

high school 1

$$Y^{(2)}([z^{(1)} = 1, z^{(2)} = 0]) = 0.1$$

$$Y^{(2)}([z^{(1)} = 0, z^{(2)} = 1]) = 0.5$$

high school 2

$$
\begin{aligned}
&\mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}] \\
-&\mathbb{E}[Y^{(i)}(a', \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}] < \tau
\end{aligned}
$$

$$\tau = 0$$

school 1 gets classes

$$
\begin{aligned}
&Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) \\
-\,&Y^{(2)}([z^{(1)} = 1, z^{(2)} = 0]) = 0.9
\end{aligned}
$$

✗

school 2 gets classes

$$
\begin{aligned}
&Y^{(2)}([z^{(1)} = 0, z^{(2)} = 1]) \\
-\,&Y^{(1)}([z^{(1)} = 0, z^{(2)} = 1]) = -0.1
\end{aligned}
$$

✓

$$Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) = 1.0$$

$$Y^{(1)}([z^{(1)} = 0, z^{(2)} = 1]) = 0.6$$

high school 1

$$Y^{(2)}([z^{(1)} = 1, z^{(2)} = 0]) = 0.1$$

$$Y^{(2)}([z^{(1)} = 0, z^{(2)} = 1]) = 0.5$$

high school 2

VS. COUNTERFACTUAL FAIRNESS

constraint on **counterfactual privilege**

$$\mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}] - \mathbb{E}[Y^{(i)}(a', \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}] < \tau$$

vs.

$$P(\hat{Y}^{(i)}(a^{(i)}) = y \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}) = P(\hat{Y}^{(i)}(a') = y \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)})$$

counterfactual fairness

[Kusner et al., 2017]

OPTIMIZATION: MILP

$$\max_{\substack{\mathbf{z}\in\{0,1\}^n \\ \mathbf{H}\in[0,1]^{(n,2^K)}}} \sum_{i=1}^{n}\sum_{j=1}^{2^K} h_{ij}\xi^{ij}(a^{(i)})$$

$$s.t., \sum_{j=1}^{2^K} h_{i,j}\left[\xi^{ij}_{\prec}(a^{(i)}) - \xi^{ij}_{\prec}(a')\right] < \tau, \quad \forall a', i$$

$$\mathbb{I}[\mathbf{e}_j = 1]h_{ij} \leq \mathbf{z}^{N(i)}, \qquad \forall i, j$$

$$\mathbb{I}[\mathbf{e}_j = 0]h_{ij} \leq 1 - \mathbf{z}^{N(i)}, \qquad \forall i, j$$

$$\sum_{j=1}^{2^K} h_{ij} = 1, \qquad \forall i$$

$$\sum_{i=1}^{n} z^{(i)} \leq b.$$

OPTIMIZATION: MILP

can accommodate any formulation of impact

$$\xi^{ij}(a^{(i)}) := \mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}]$$

$$\max_{\substack{\mathbf{z}\in\{0,1\}^n \\ \mathbf{H}\in[0,1]^{(n,2^K)}}} \quad \sum_{i=1}^{n}\sum_{j=1}^{2^K} h_{ij}\xi^{ij}(a^{(i)})$$

$$s.t., \sum_{j=1}^{2^K} h_{i,j}\left[\xi^{ij}_{\prec}(a^{(i)}) - \xi^{ij}_{\prec}(a')\right] < \tau, \quad \forall a', i$$

$$\mathbb{I}[\mathbf{e}_j = 1]h_{ij} \le \mathbf{z}^{N(i)}, \qquad \forall i, j$$

$$\mathbb{I}[\mathbf{e}_j = 0]h_{ij} \le 1 - \mathbf{z}^{N(i)}, \qquad \forall i, j$$

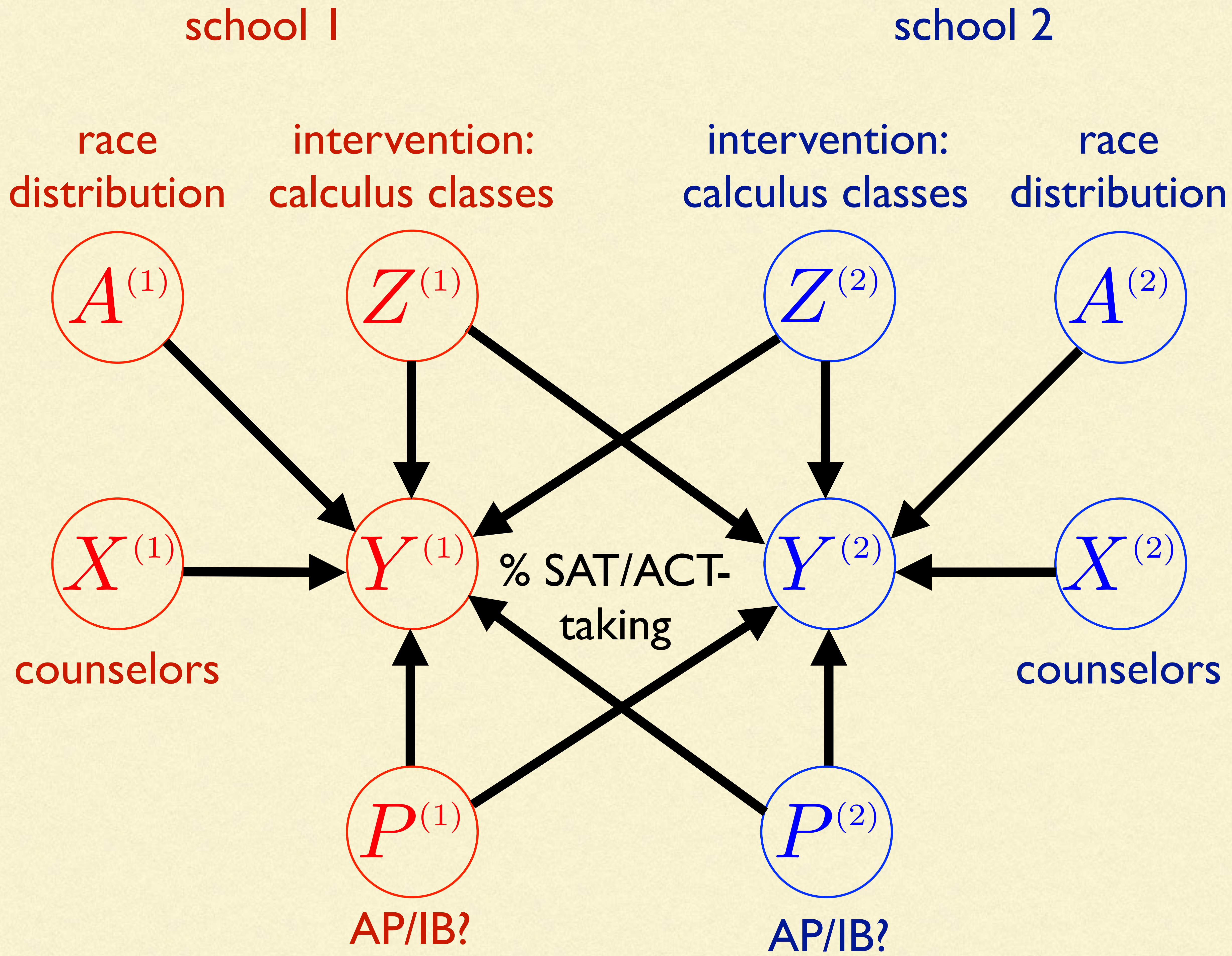
$$\sum_{j=1}^{2^K} h_{ij} = 1, \qquad \forall i$$

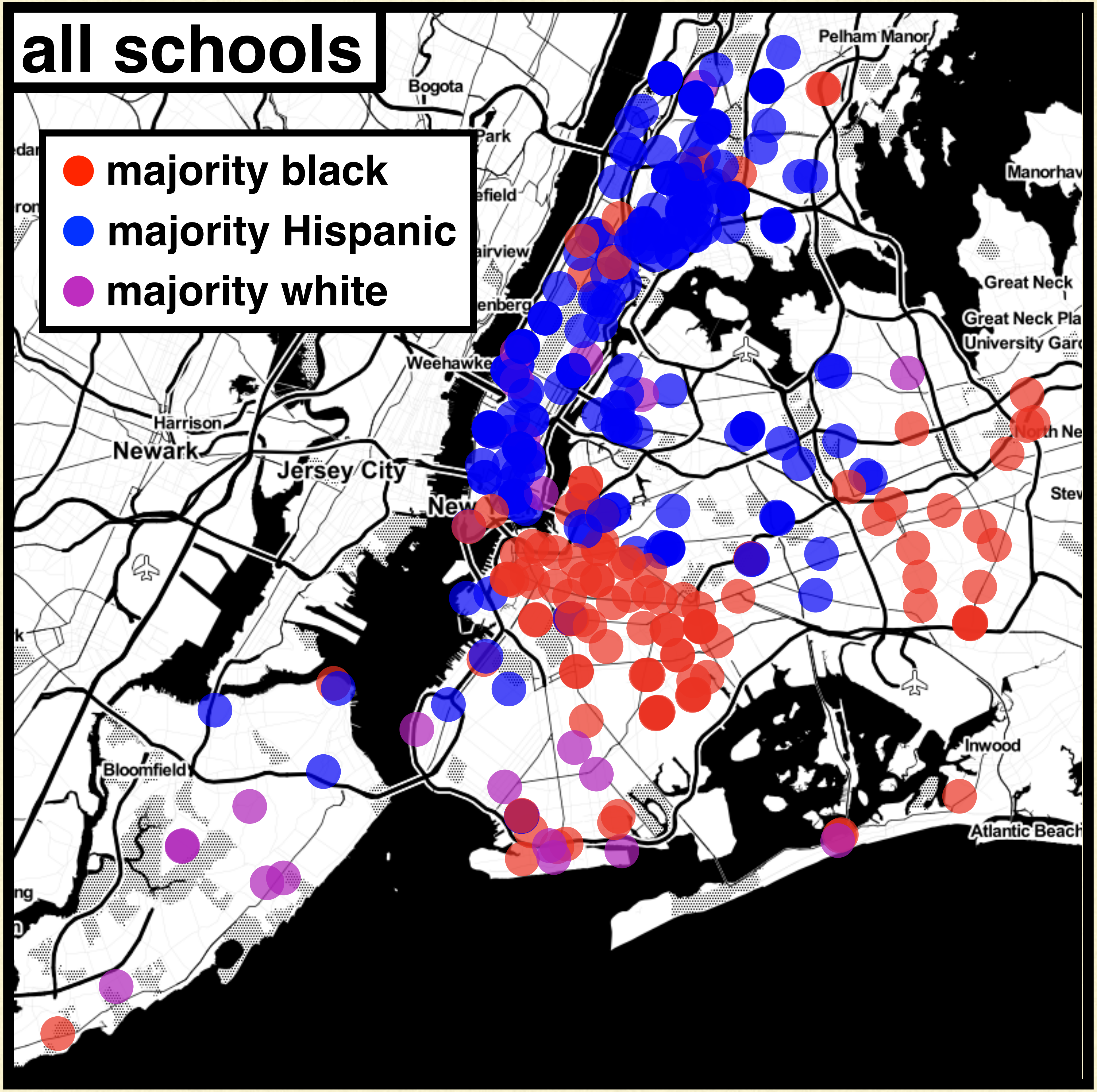
$$\sum_{i=1}^{n} z^{(i)} \le b.$$

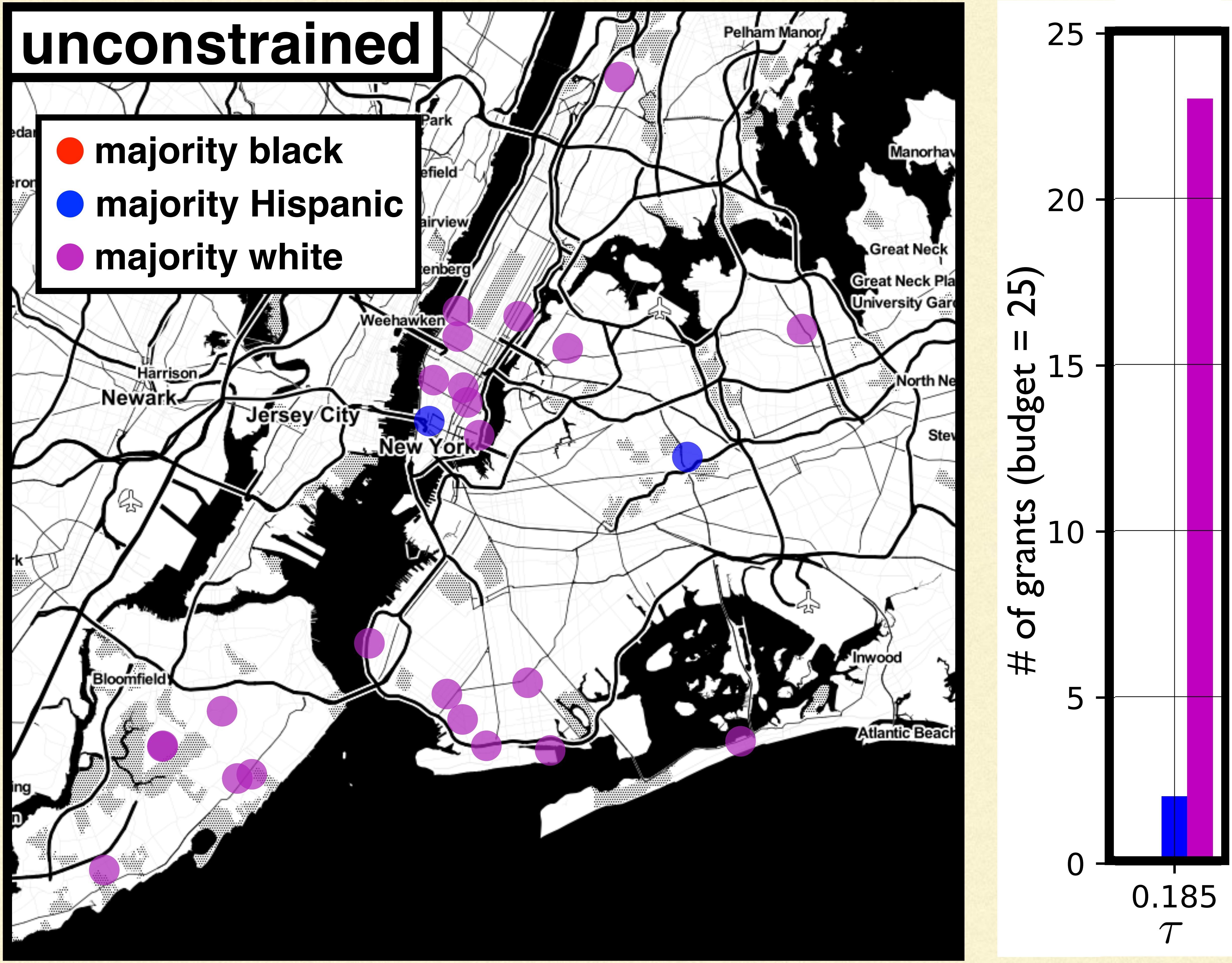
NYC PUBLIC SCHOOL FUNDING

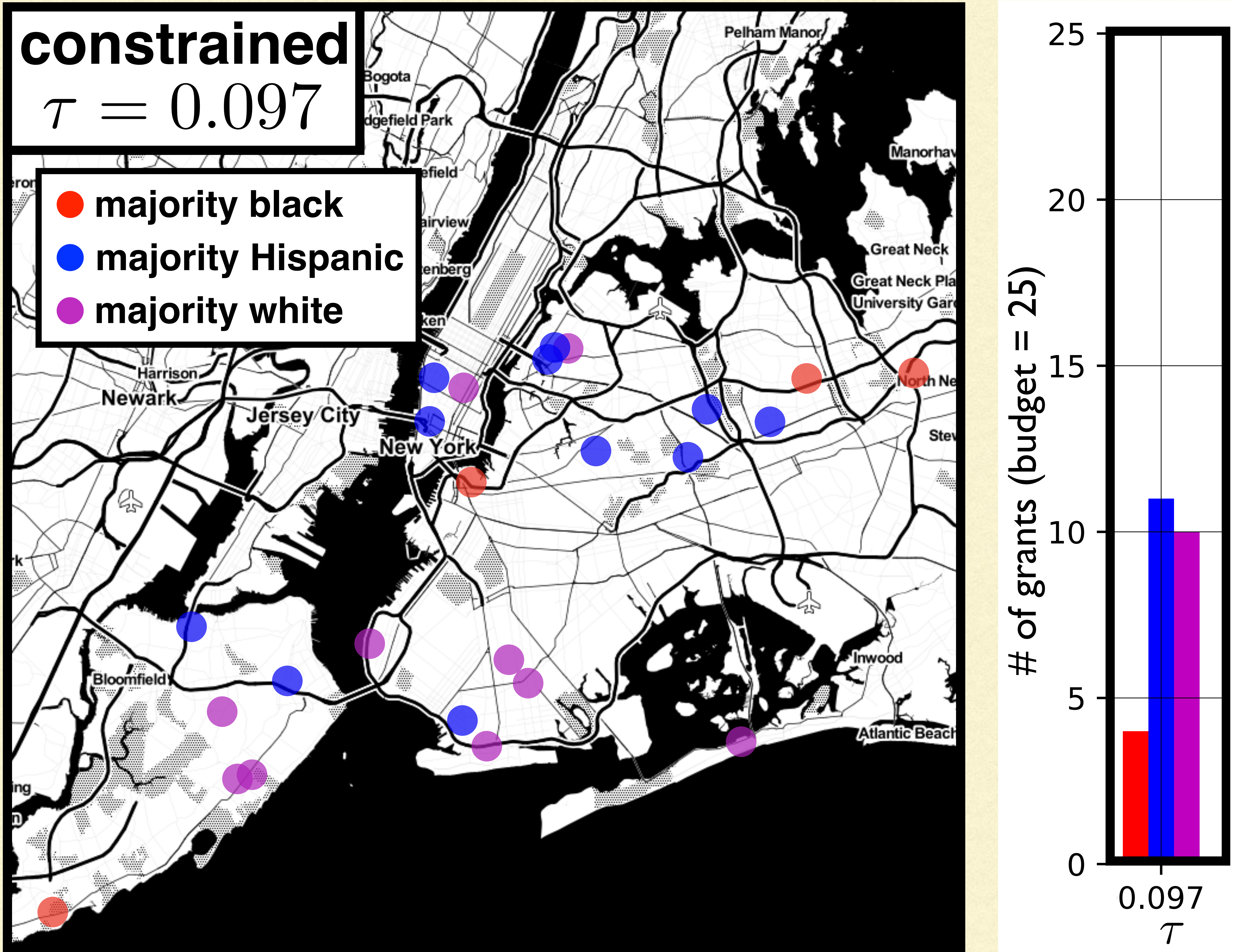
[CRDC, HTTPS://OCRDATA.ED.GOV/]

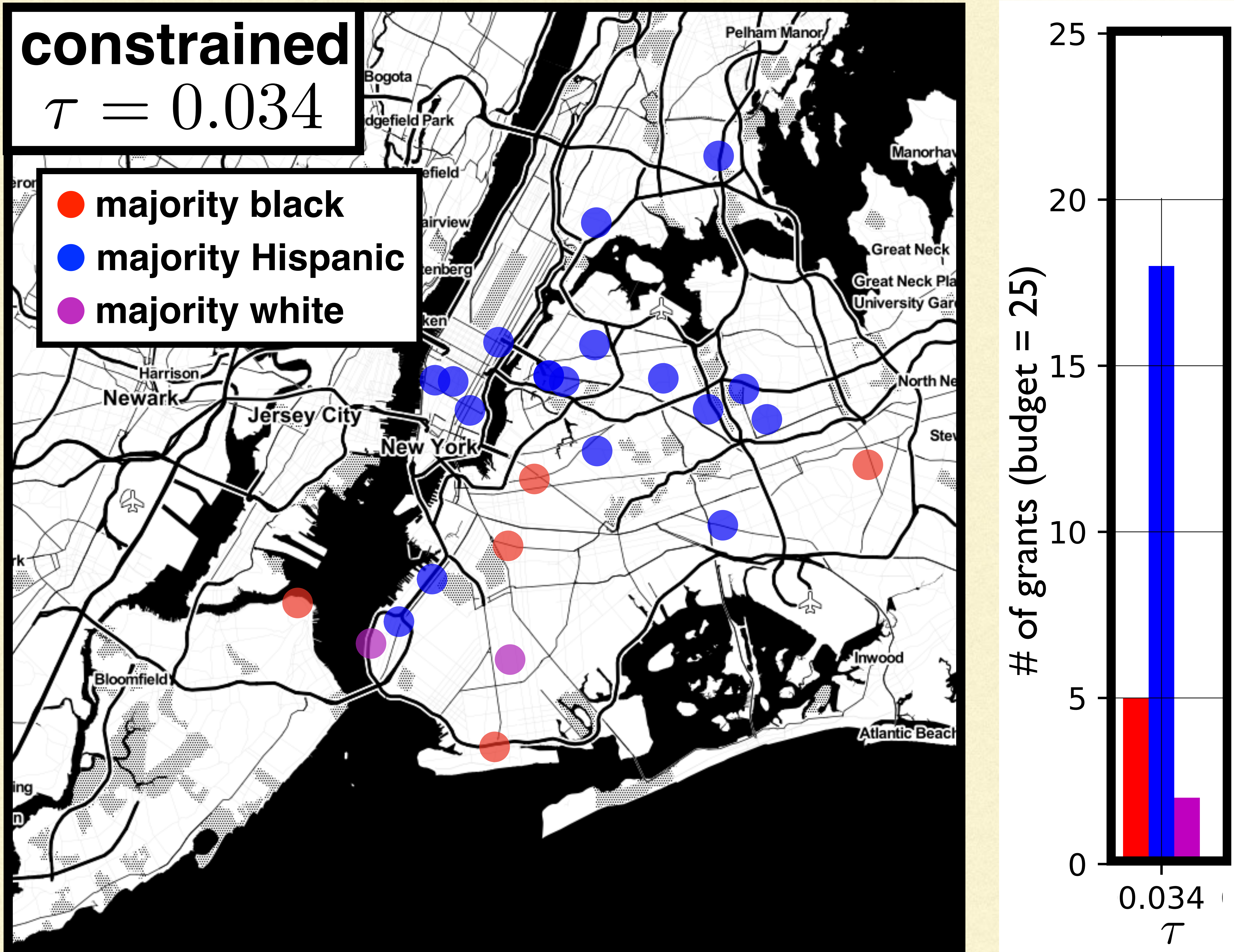
345 schools

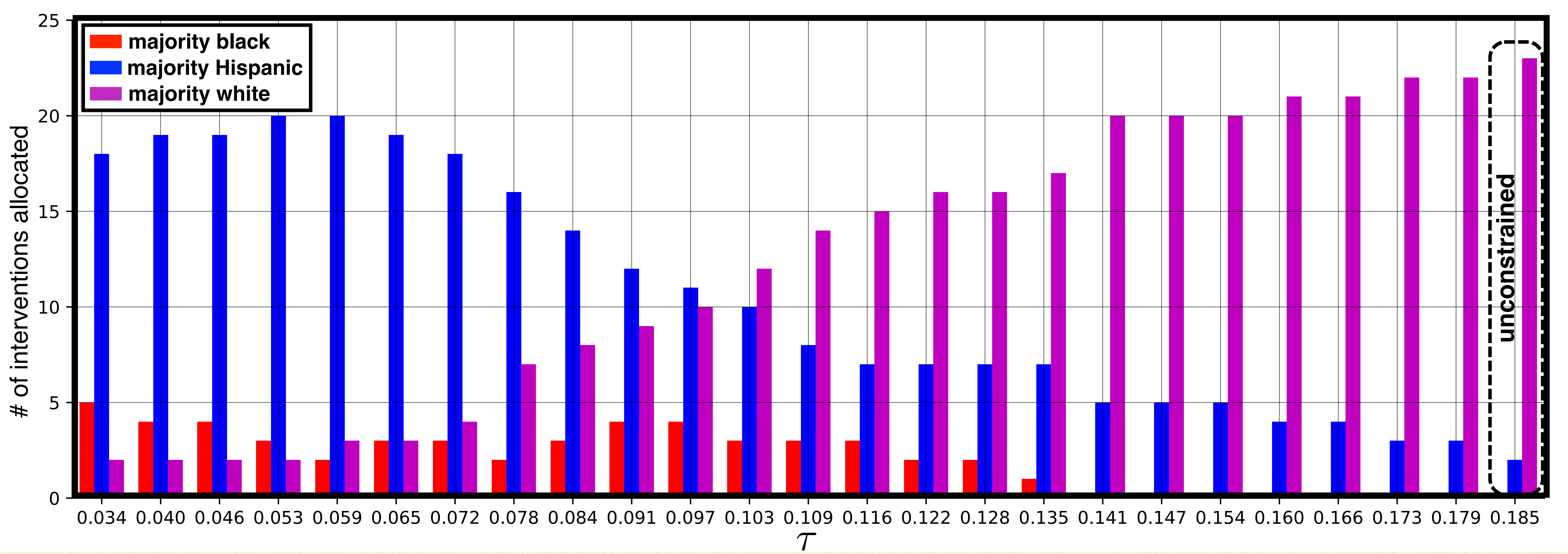

school 1
school 2
race distribution
intervention: calculus classes
intervention: calculus classes
race distribution
$A^{(1)}$
$Z^{(1)}$
$Z^{(2)}$
$A^{(2)}$
$X^{(1)}$
$Y^{(1)}$
% SAT/ACT-taking
$Y^{(2)}$
$X^{(2)}$
counselors
counselors
$P^{(1)}$
$P^{(2)}$
AP/IB?
AP/IB?


all schools
majority black
majority Hispanic
majority white
Pelham Manor
Bogota
Manorhav
Great Neck
Great Neck Pla
University Gar
Weehawke
Harrison
Newark
Jersey City
Inwood
Bloomfield
Atlantic Beach

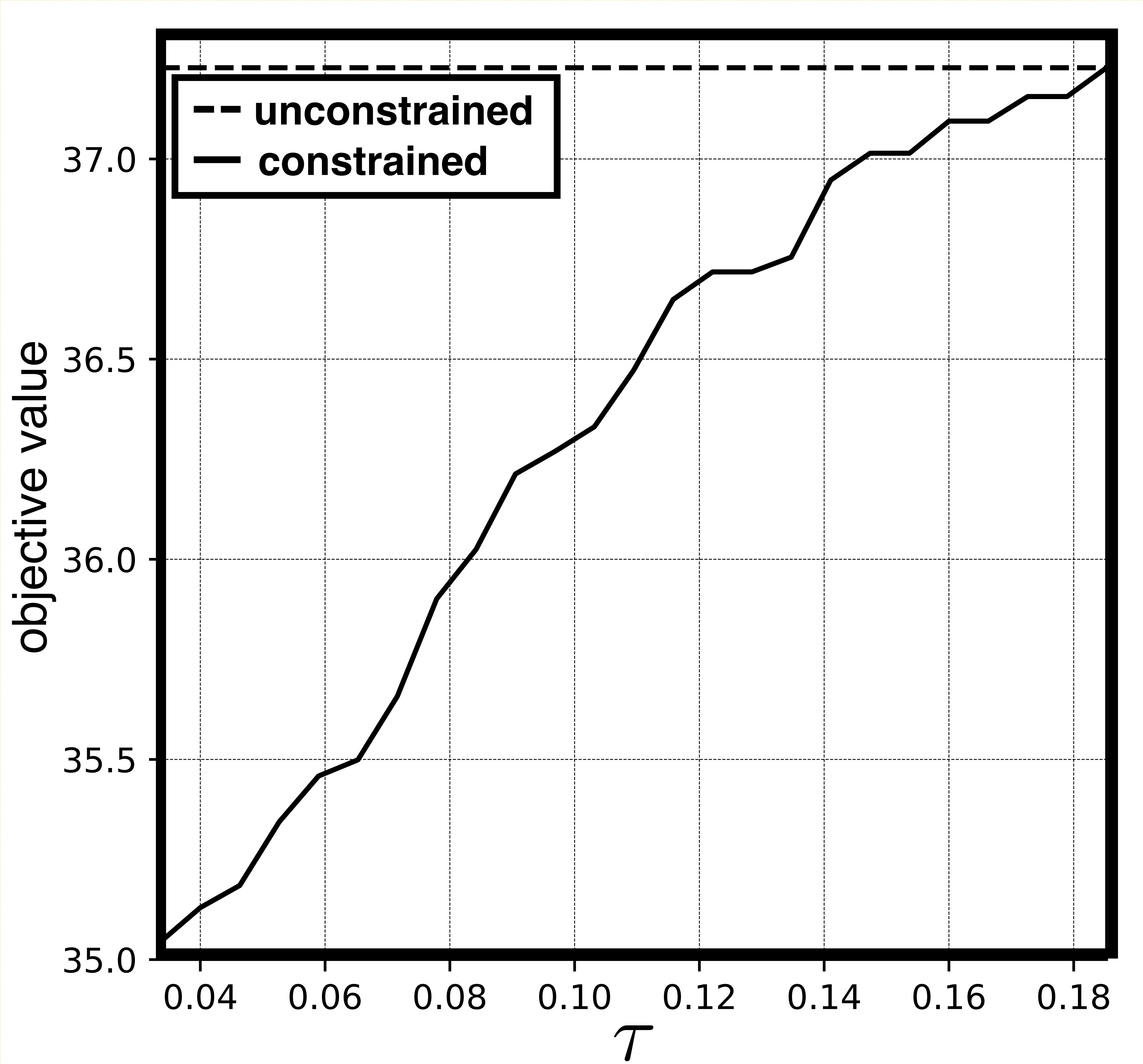

unconstrained
majority black
majority Hispanic
majority white
Pelham Manor
Park
Manorhav
efield
airview
Great Neck
Great Neck Pla
University Gar
enberg
Weehawken
Harrison
Newark
Jersey City
New York
North Ne
Stev
Bloomfield
Inwood
Atlantic Beach
25
20
15
10
5
0
# of grants (budget = 25)
0.185
τ

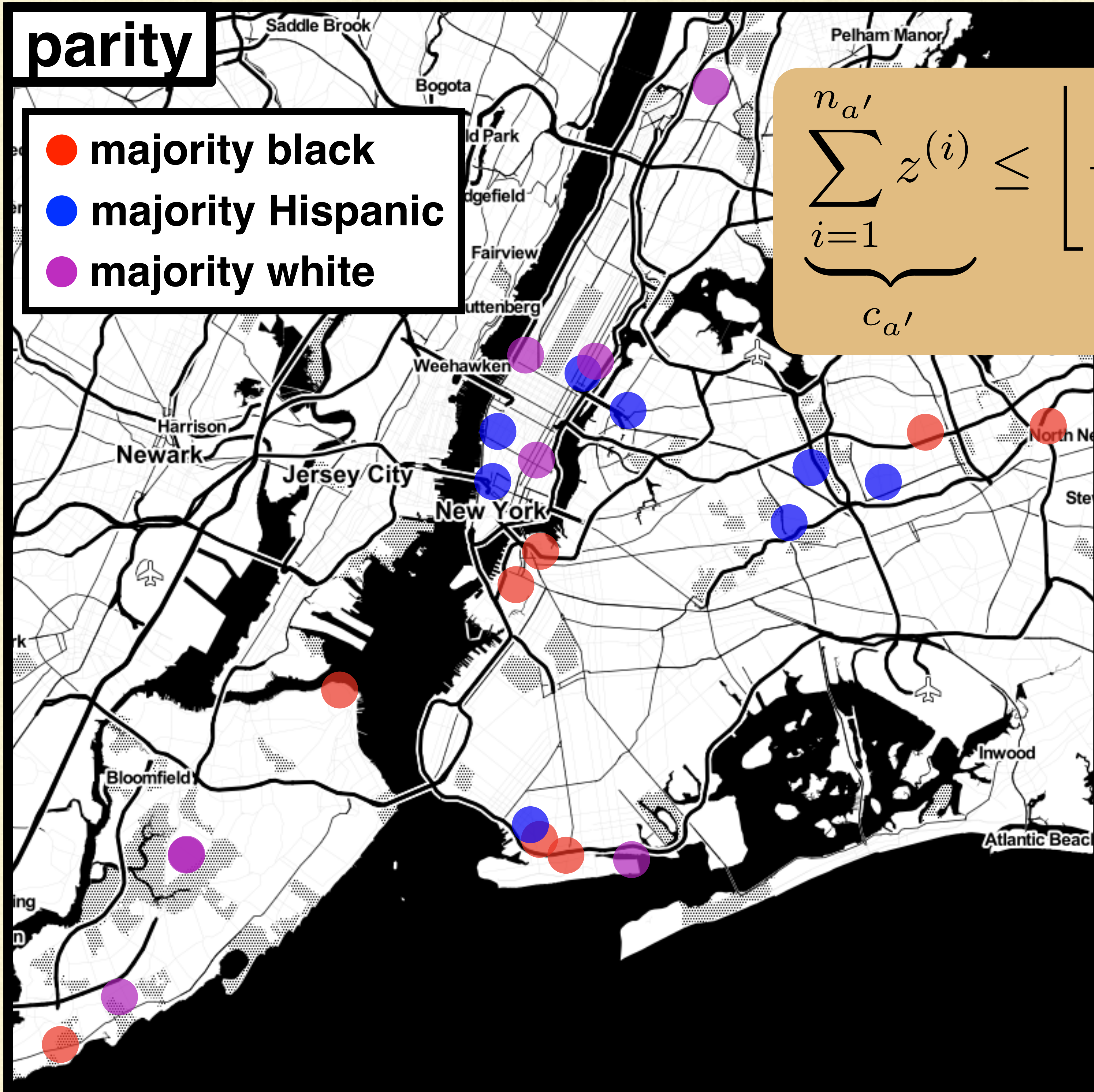

constrained
$\tau = 0.097$
majority black
majority Hispanic
majority white
Pelham Manor
Bogota
Ridgefield Park
Manorhav
Great Neck
Great Neck Pla
University Gar
Harrison
Newark
Jersey City
New York
North Ne
Bloomfield
Inwood
Atlantic Beach
# of grants (budget = 25)
0
5
10
15
20
25
0.097
τ


constrained
$\tau = 0.034$
majority black
majority Hispanic
majority white
Pelham Manor
Bogota
Manorhav
Great Neck
Great Neck Pla
University Gar
North Ne
Harrison
Newark
Jersey City
New York
Stev
Inwood
Bloomfield
Atlantic Beach
# of grants (budget = 25)
0
5
10
15
20
25
0.034
τ

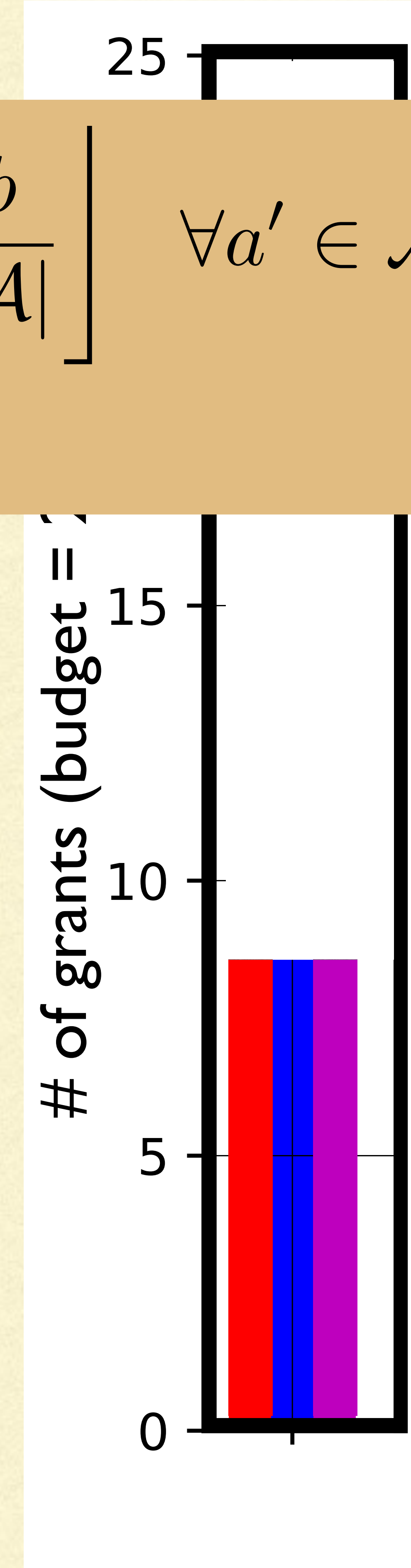


majority black
majority Hispanic
majority white
unconstrained
# of interventions allocated
τ
0
5
10
15
20
25
0.034 0.040 0.046 0.053 0.059 0.065 0.072 0.078 0.084 0.091 0.097 0.103 0.109 0.116 0.122 0.128 0.135 0.141 0.147 0.154 0.160 0.166 0.173 0.179 0.185

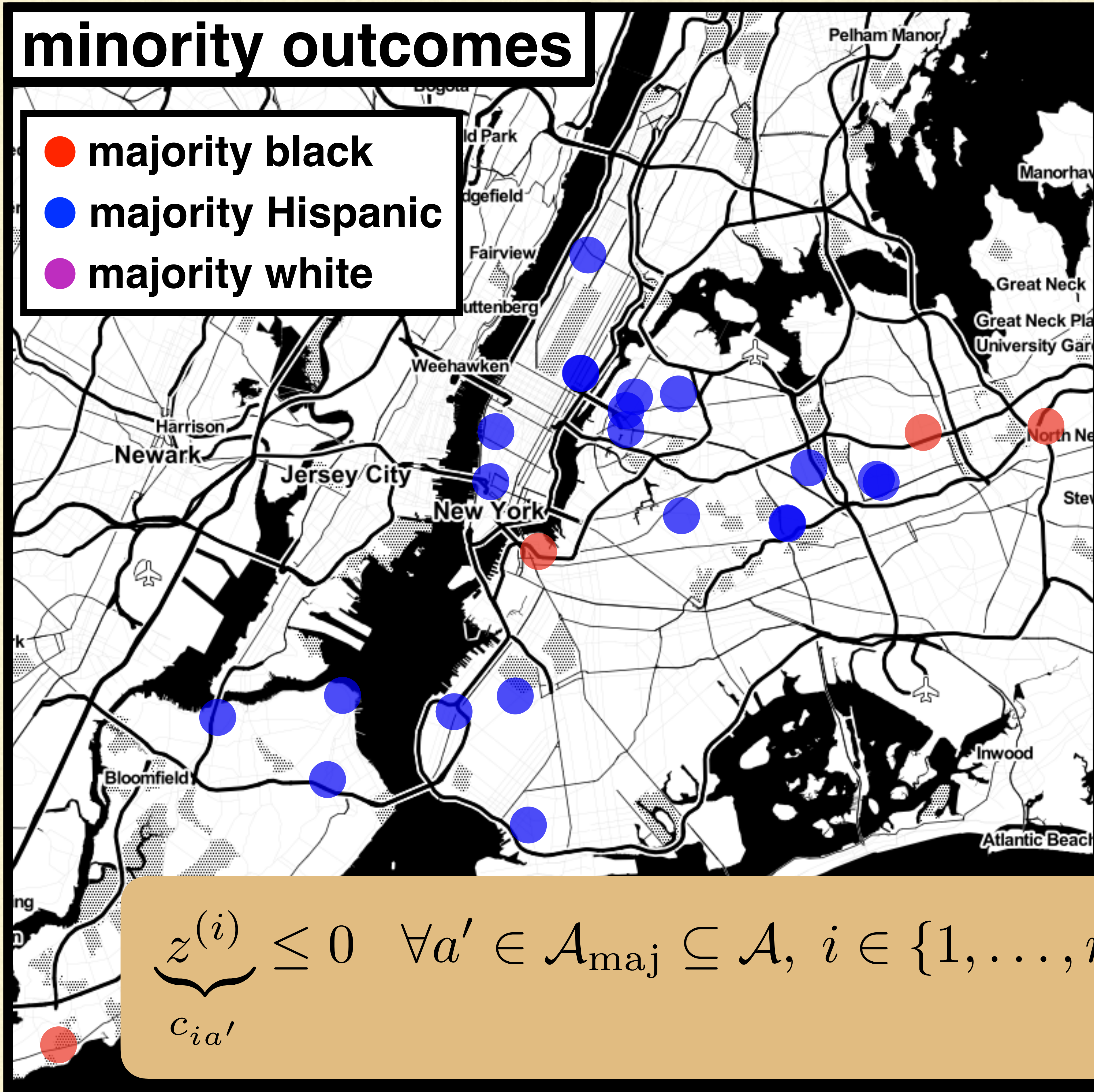

unconstrained
constrained
37.0
36.5
36.0
35.5
35.0
objective value
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
τ



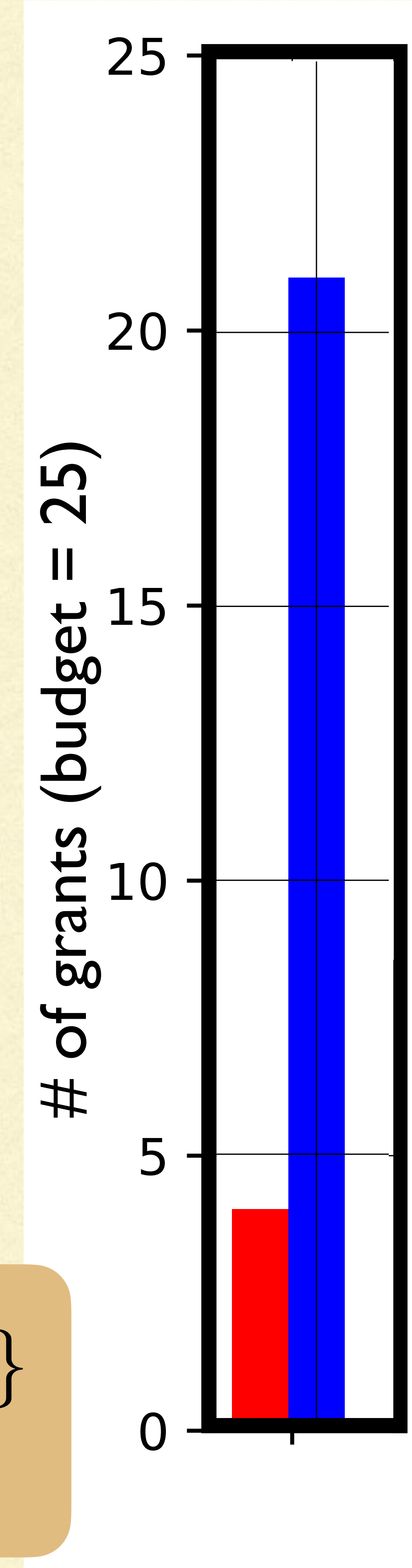
$$\underbrace{\sum_{i=1}^{n_{a'}} z^{(i)}}_{c_{a'}} \leq \left\lfloor \frac{b}{|\mathcal{A}|} \right\rfloor \quad \forall a' \in \mathcal{A}$$

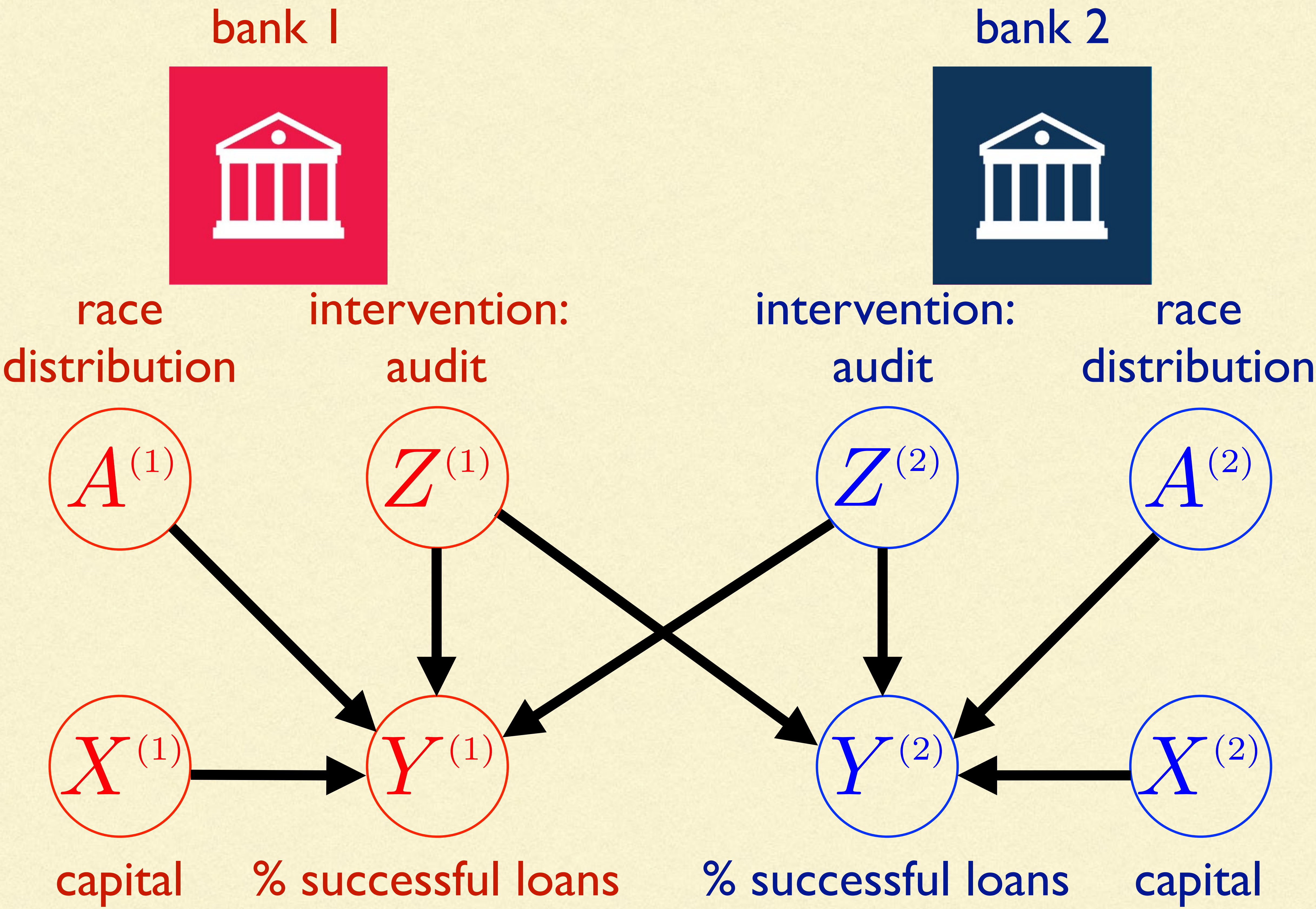


minority outcomes



$$\underbrace{z^{(i)}}_{c_{ia'}} \le 0 \quad \forall a' \in \mathcal{A}_{\text{maj}} \subseteq \mathcal{A},\ i \in \{1, \ldots, n\}$$



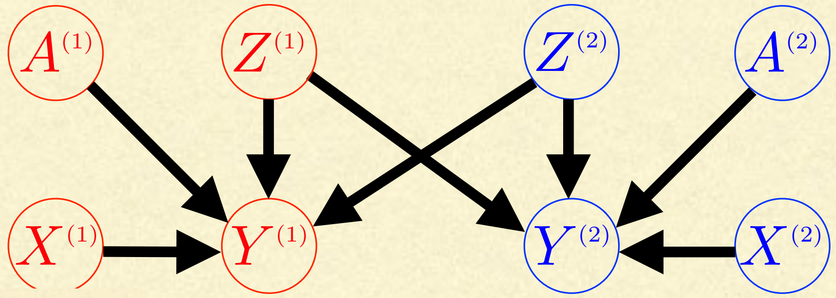

bank 1
bank 2
race distribution
intervention: audit
intervention: audit
race distribution
$A^{(1)}$
$Z^{(1)}$
$Z^{(2)}$
$A^{(2)}$
$X^{(1)}$
$Y^{(1)}$
$Y^{(2)}$
$X^{(2)}$
capital
% successful loans
% successful loans
capital

TAKE-AWAYS

- Many cases where ML algorithms decide only **part of an impact**

TAKE-AWAYS

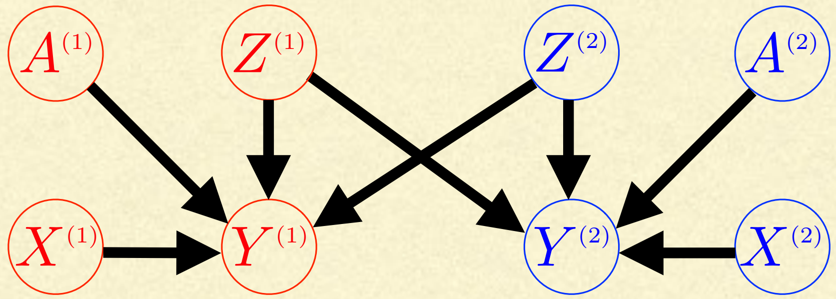
- Many cases where ML algorithms decide only **part of an impact**



- Idea: formalize algorithmic decisions within society **using causal models**

TAKE-AWAYS

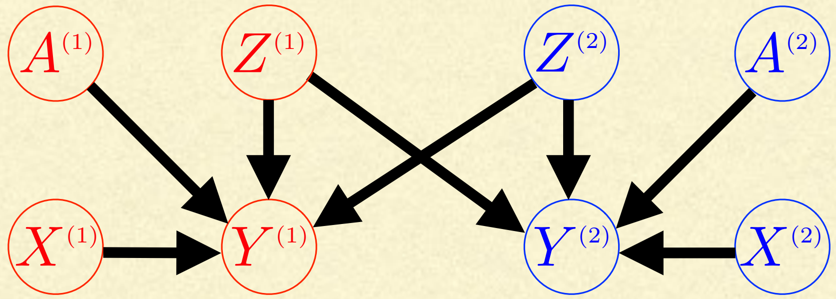
- Many cases where ML algorithms decide only **part of an impact**
- Idea: formalize algorithmic decisions within society **using causal models**
- **Counterfactuals** allow us to formulate **discriminatory privilege**



$$Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) - Y^{(2)}([z^{(1)} = 1, z^{(2)} = 0]) = 0.9$$

TAKE-AWAYS

- Many cases where ML algorithms decide only **part of an impact**



- Idea: formalize algorithmic decisions within society **using causal models**
- **Counterfactuals** allow us to formulate **discriminatory privilege**

$$Y^{(1)}([z^{(1)} = 1, z^{(2)} = 0]) - Y^{(2)}([z^{(1)} = 1, z^{(2)} = 0]) = 0.9$$

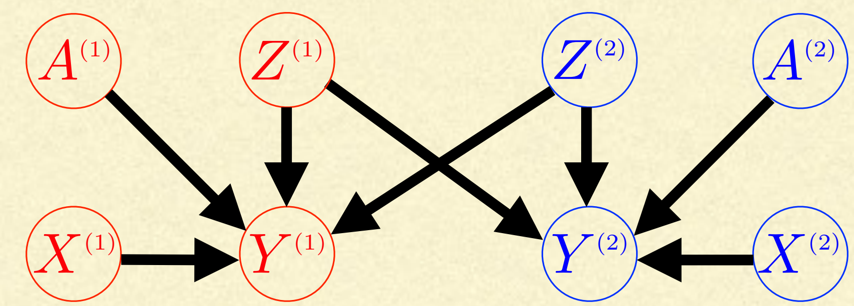
- We propose a constrained optimization problem that maximizes overall impact while **reducing privileged impact**

$$\max_{\mathbf{z} \in \{0,1\}^n} \sum_{i=1}^{n} \mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)} = a^{(i)}, X^{(i)} = \mathbf{x}^{(i)}]$$
$$s.t., \sum_{i=1}^{n} z^{(i)} \leq b$$
$$c_{ia'} \leq \tau \quad \forall a' \in \mathcal{A},\ i \in \{1, \ldots, n\},$$

TAKE-AWAYS

- Many cases where ML algorithms decide only **part of an impact**

- Idea: formalize algorithmic decisions within society **using causal models**



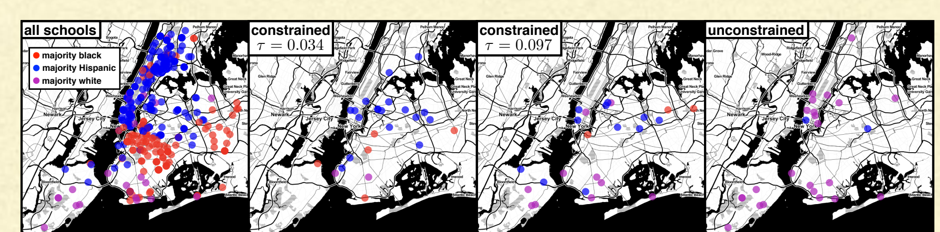
- **Counterfactuals** allow us to formulate **discriminatory privilege**

$$Y^{(1)}([z^{(1)}=1, z^{(2)}=0]) - Y^{(2)}([z^{(1)}=1, z^{(2)}=0]) = 0.9$$

- We propose a constrained optimization problem that maximizes overall impact while **reducing privileged impact**

$$\max_{\mathbf{z}\in\{0,1\}^n} \sum_{i=1}^{n} \mathbb{E}[Y^{(i)}(a^{(i)}, \mathbf{z}) \mid A^{(i)}=a^{(i)}, X^{(i)}=\mathbf{x}^{(i)}]$$
$$s.t., \sum_{i=1}^{n} z^{(i)} \le b$$
$$c_{ia'} \le \tau \quad \forall a' \in \mathcal{A},\ i \in \{1,\dots,n\},$$

- Allows one to make less discriminatory policy decisions for school funding



MY COAUTHORS

Chris Russell

The Alan Turing Institute

Joshua R Loftus

NYU

Ricardo Silva

UCL

The Alan Turing Institute