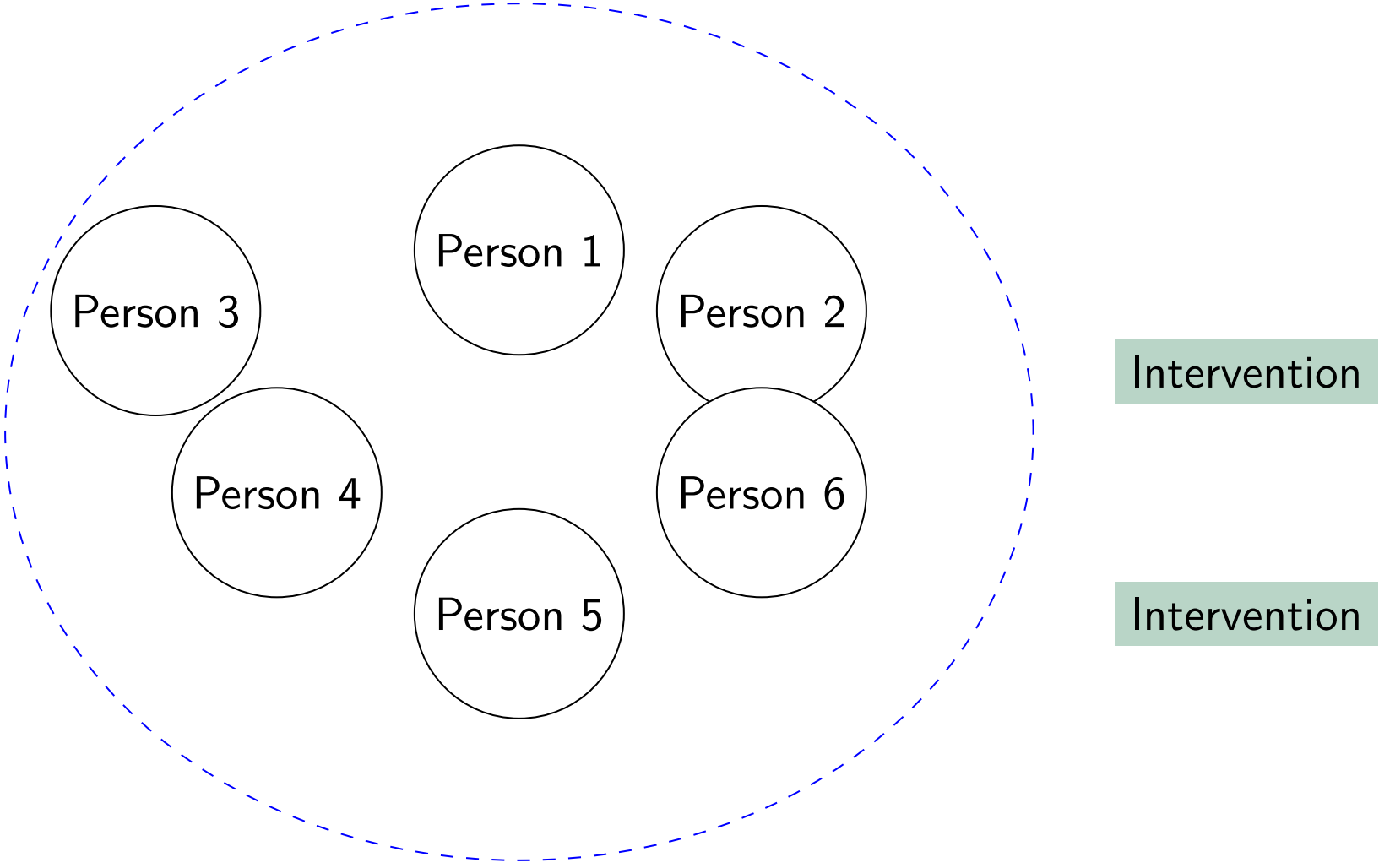
What is the Organizational Counterfactual? Categorical versus Algorithmic Prioritization in U.S. Social Policy

Rebecca Johnson [1,2] and Simone Zhang [1]

[1]Department of Sociology, Princeton University

[2]Quantitative Social Science, Dartmouth College

Background: public bureaucracies need to prioritize among when allocating scarce interventions



Individuals have many attributes/features that could indicate higher risk or need

Housing:

Person	*Income*	*Homeless*	*Dep. child*	...
1	10,000	1	1	...
2	0	0	0	...
⋮				
n	100,000	0	1	...

Schools:

Student	*Disability severity*	*English profic.*	*Parent income*	...
1	90	100	100,000	...
2	70	80	60,000	...
⋮				
n	10	30	30,000	...

Feature selection to decide between versus dimension reduction to summarize all

Housing:

Person	*Income*	*Homeless*	*Dep. child*	...
1	10,000	1	1	...
2	0	0	0	...
⋮				
n	100,000	0	1	...

Schools:

Student	*Disability severity*	*English profic.*	*Parent income*	...
1	90	100	100,000	...
2	70	80	60,000	...
⋮				
n	10	30	30,000	...

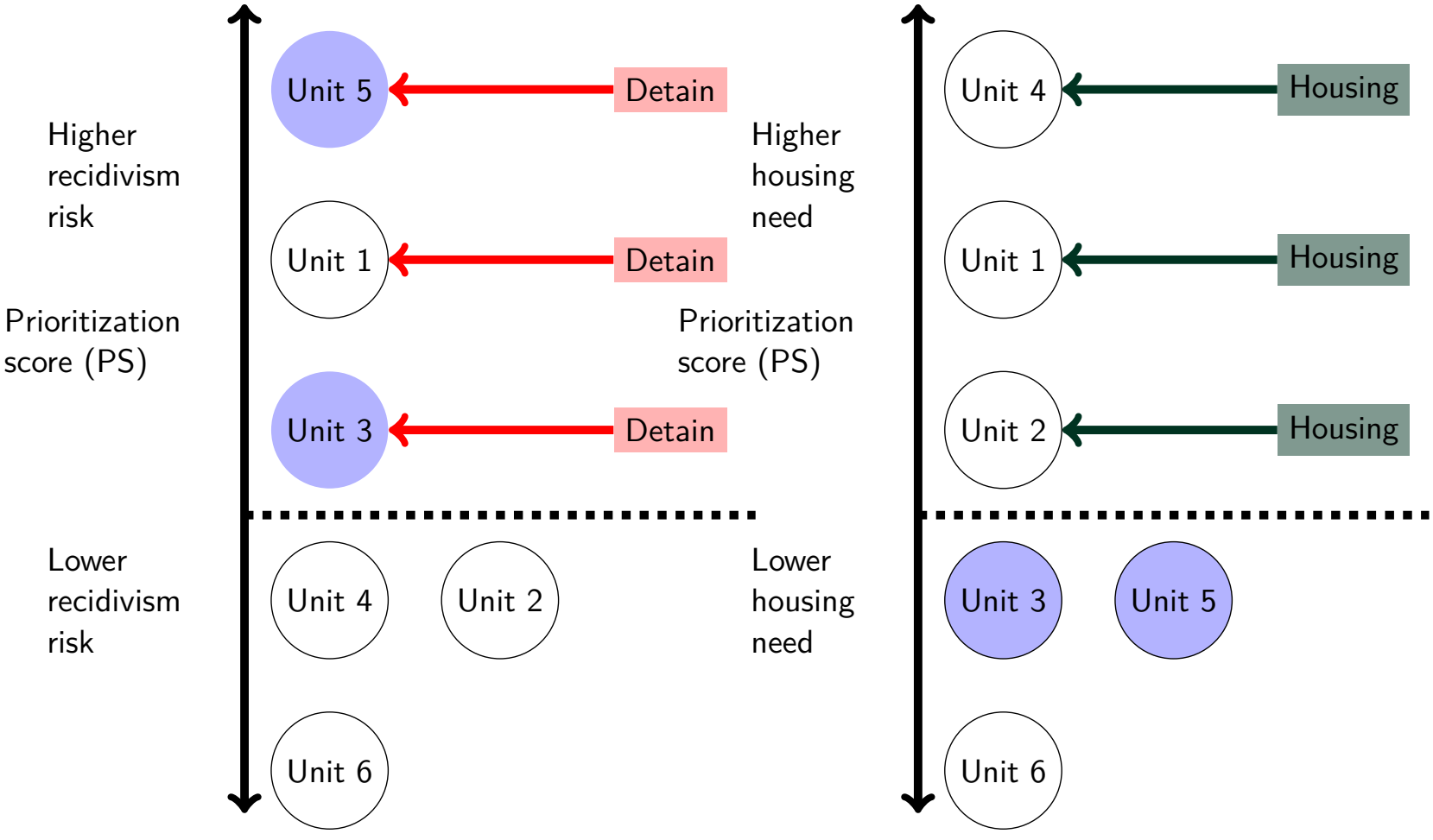
Present talk: contrasting two forms of feature selection/dimension reduction

1. **Algorithmic prioritization (AP):** using a model to weight different attributes
   - 1.1 Define a prioritization target that measures risk of bad outcome/need for resource/or other goal: e.g., $E[\text{Y} = \text{homeless in the next year}]$; $E[Y(\text{dropout}(\text{D} = \text{extra help}) - Y(\text{dropout}\ (\text{D} = \text{no extra help}))]$
   - 1.2 Use machine learning to learn that target as flexible function of many attributes about individuals/units
   - 1.3 Prioritize individuals with highest $\hat{y}$ (highest estimated need or risk of bad outcome as a function of many attributes)
2. **Categorical prioritization (CP):** manually deciding which attributes matter and how to weight them

Two types of unfair outcomes

Over-allocating a punitive resource to certain subgroups

Under-allocating an assistive resource to certain subgroups



Present talk

Focus today: concerns that algorithmic prioritization will result in *under-allocating* an assistive resource to certain subgroups

Outline

- **Relationship with Prior Work**
- Overview of Categorical Prioritization (CP) and two cases (school districts; local housing authorities)
- Source of unfairness that Algorithmic Prioritization could potentially remedy

Unfairness in Social Services Agencies using Algorithmic Prioritization

Which data to use?

Which model to map features to label?

How to evaluate model performance?

Loans (FH, 2013; 2016);
housing assistance (E, 2018);
school resources (BK), 2017);
social services (BK, 2017)

FH: Fourcade and Healy; E: Eubanks; BK: Bakalar and Zevenbergen

Present talk joins other work looking at "treatment effect" of algorithmic prioritization relative to some *counterfactual prioritization method*

Cowgill and Tucker (2017) treat algorithmic prioritization as a treatment ($T = 1$) whose fairness should be assessed relative to some other prioritization method ($T = 0$)

Decision	$T = 1$	$T = 0$	Citation(s)
Detain pre-trial or not	AP	Human judge	Lakkaraju et al. (2017); Kleinberg et al. (2017)
Invite to interview or not	AP	Human screener	Cowgill (2018)
Talk to about landlord harassment or not	AP	Human agency worker	Johnson et al. (2019)

Contribution

- Past work assessing fairness of algorithmic prioritization relative to some other method focuses on 1) humans as the decision-makers; and 2) those humans assessing individual cases to make a yes or no decision
- The present talk focuses on 1) the federal government or a social planner as the decision-maker; 2) that government agency deciding more broadly whom to target help to, and using either algorithms to decide whom to target or using manually-selected categories

Prevalence of and changes in manually-selected categories across social policies

Decision across policies: give assistance to or not; CP refers to categories prioritized

Policy	Example reform	Pre-reform CP	Post-reform CP
Cash welfare	1996 PRWORA	Income $<$ threshold *and* some work req.	Income $<$ threshold *and* stronger work req.
Medicaid	ACA	Income $<$ threshold *and* parent of dependent child *or* Disability	Income $<$ higher threshold and allowed to be childless
Supplemental Security Income (SSI)	*Sullivan v. Zebley*	Listed disability	Listed disability *or* individualized assessment

Outline

- Relationship with Prior Work
- **Overview of Categorical Prioritization (CP) and two cases (school districts; local housing authorities)**
- Source of unfairness that Algorithmic Prioritization could potentially remedy

General structure of categorical prioritization

Imagine each individual/household as a vector of binary and continuous attributes, e.g.:

- Individual 1: [income 500% of federal poverty line (FPL), disability severity of 100, veteran, mother, no felony conviction, ...]
- Individual 2: [income 150% of FPL, disability severity of 10, not veteran, not a mother, no felony conviction, ...]
- Individual 3: [income 50% of FPL, disability severity of 50, not veteran, not a mother, no felony conviction, ...]
- Individual 4: [income 50% of FPL, disability severity of 50, not veteran, not a mother, felony conviction, ...]

Categorical prioritization: 1) manually select attributes

- Individual 1: [income 500% of federal poverty line (FPL), disability severity of 100, ~~veteran~~, ~~mother~~, no felony conviction, ...]
- Individual 2: [income 150% of FPL disability severity of 10, ~~not veteran~~, ~~not mother~~, no felony conviction, ...]
- Individual 3: [income 50% of FPL, disability severity of 50, ~~not veteran~~, ~~not mother~~, no felony conviction, ...]
- Individual 4: [income 50% of FPL, disability severity of 50, ~~not veteran~~, ~~not mother~~, felony conviction, ...]

Categorical prioritization: 2) manually decide thresholds to convert continuous attributes to binary

- Individual 1: [~~income 500% of federal poverty line (FPL)~~ not low-income, ~~disability severity of 100~~ has disability, no felony conviction, ...]
- Individual 2: [~~income 150% of FPL~~ not low-income, ~~disability severity of 10~~ no disability, no felony conviction, ...]
- Individual 3: [~~income 50% of FPL~~ low-income, ~~disability severity of 50~~ no disability, no felony conviction, ...]
- Individual 4: [~~income 50% of FPL~~ low-income, ~~disability severity of 50~~ no disability, felony conviction, ...]

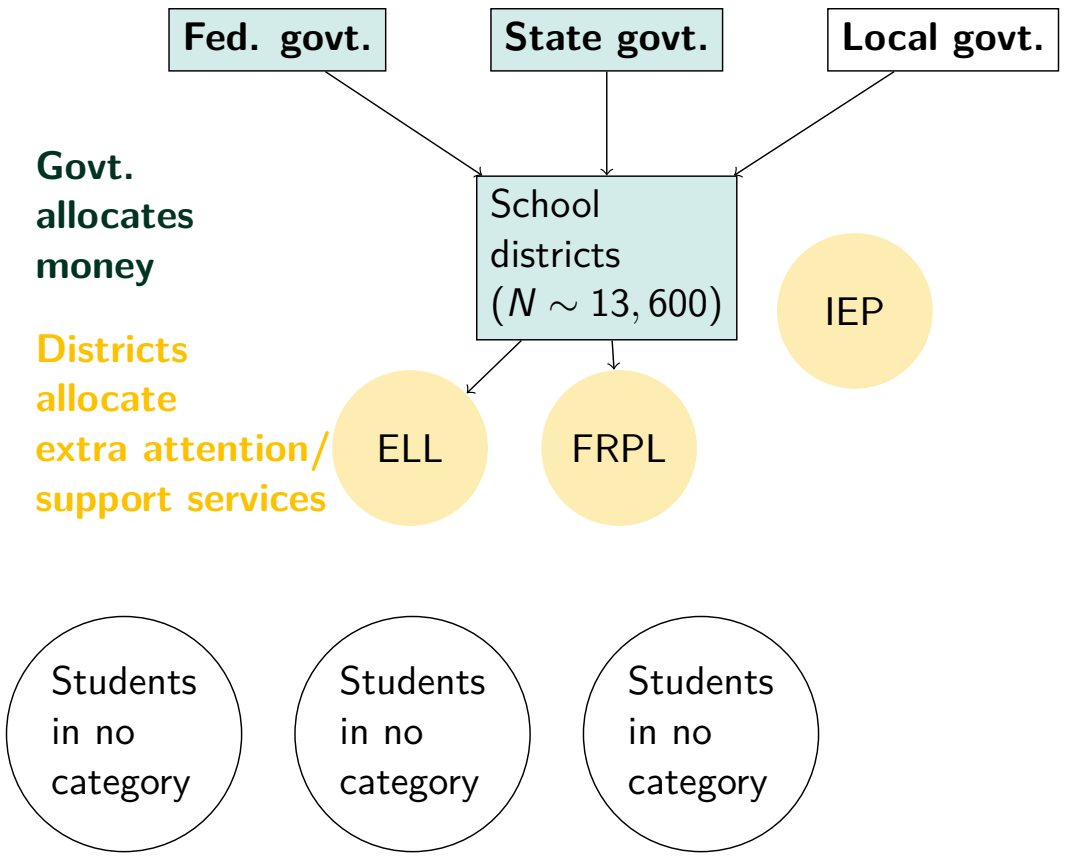
Categorical prioritization: 3) decide aggregation logic for binary attributes

1. **And** logic: need to possess both attributes
   - *E.g.*: SSI requiring both low-income and having disability
2. **Or** logic: need to possess either attribute
   - *E.g.*: can qualify for Medicaid on basis of income *or* disability
3. **Unless** logic: if have an attribute, then excluded (even if you display all other attributes)
   - *E.g.:* exclusions of individuals with felony convictions or legal citizenship from certain social policies

First case of categorical prioritization: school districts

Who's doing the prioritization; Which categories they're prioritizing

(ELL = English Language Learner students; FRPL = Free or Reduced Price Lunch; IEP = Individualized Education Plan for disability)



Example: Boston Public Schools

Example: Student A

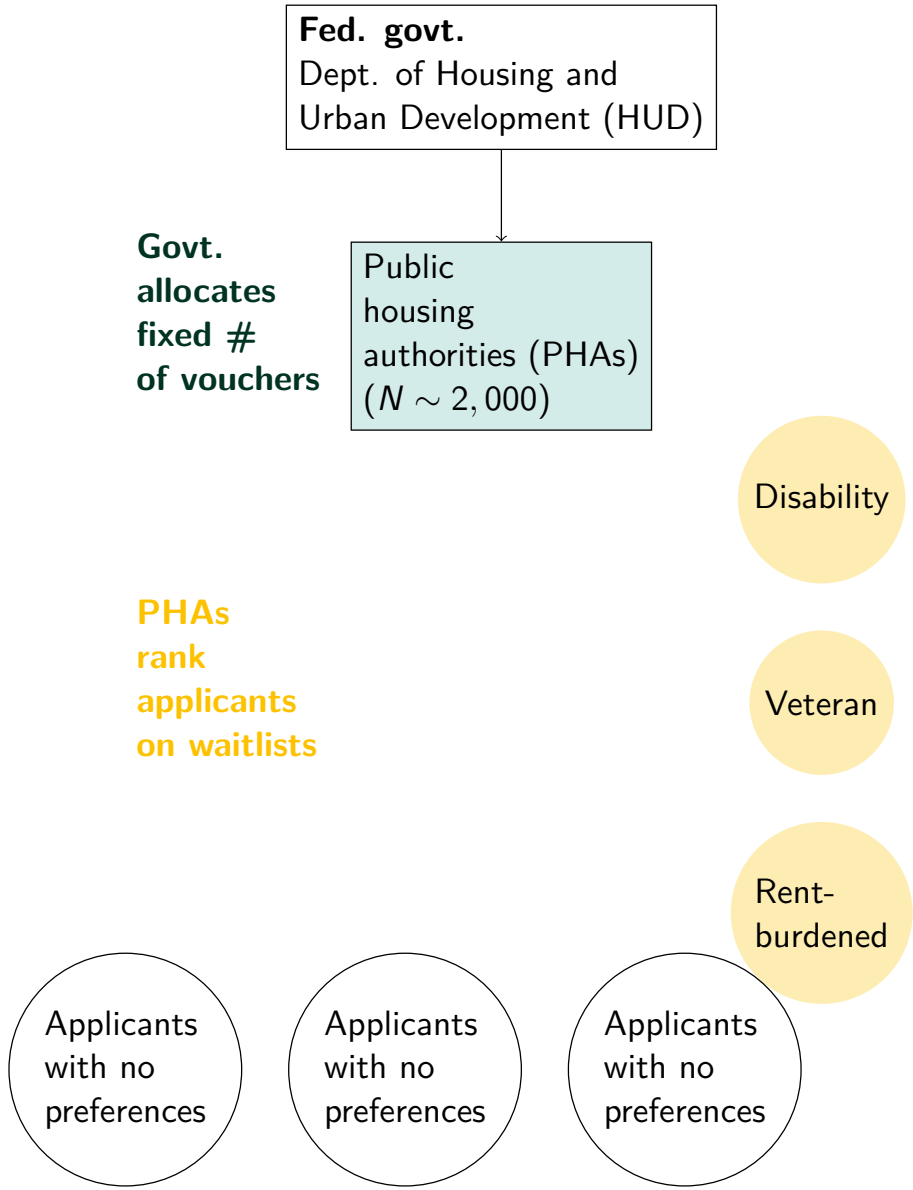
Type	Weight	$ Amount
Grade 6	1.4	$5,121
Poverty	0.1	$366
ELL	0.25	$915
SUBTOTAL		$6,402

Example: Student B

Type	Weight	$ Amount
Grade 4	1.3	$4,755
Autism	4.3	$15,730
SUBTOTAL		$20,485

- Principles of weighted student funding: equity, transparency, student focused and differentiated based on need
  - Everyone could see how much every school received and knew why within minutes
  - The "little lady with curlers in her hair in South Boston" knew how much her child received
- Developed weights with central office staff and cross functional group ("Group of 60")

Second context for categorical prioritization: local housing authorities (PHAs)



Example: housing authority

The Jeffersonville Housing Authority will select families based on the following preferences based on our local housing needs and priorities:

A. Substandard Housing as determined by local housing code / Involuntarily Displaced by Government Action, Declared Disaster at the local level or sale/loss of property by landlord, or Victims of Domestic Violence (20 points)

B. Elderly/Handicapped/Disabled OR working a minimum of 20 hours weekly (20 points)

C. Applicants who are working and/or living within the City Limits of Jeffersonville/Clarksville (30 points)

E. Veterans (honorably discharged) (5 points)

A family may qualify for more than one (1) preference. The family with the most cumulative preferences will be offered housing first based upon availability.

Outline

- Relationship with Prior Work
- **Overview of Categorical Prioritization (CP) and two cases (school districts; local housing authorities)**
- Source of unfairness that Algorithmic Prioritization could potentially remedy

***Manually* select a few attributes**

+	-	Schools	Housing
Transparent *who* is prioritized	Non-transparent *why* they are prioritized; lack of transparency may obscure de-prioritization of those who are needy but perceived as more blameworthy/less deserving	BPS: $15,000 for each child with autism; $900 for each English Language Learner	...

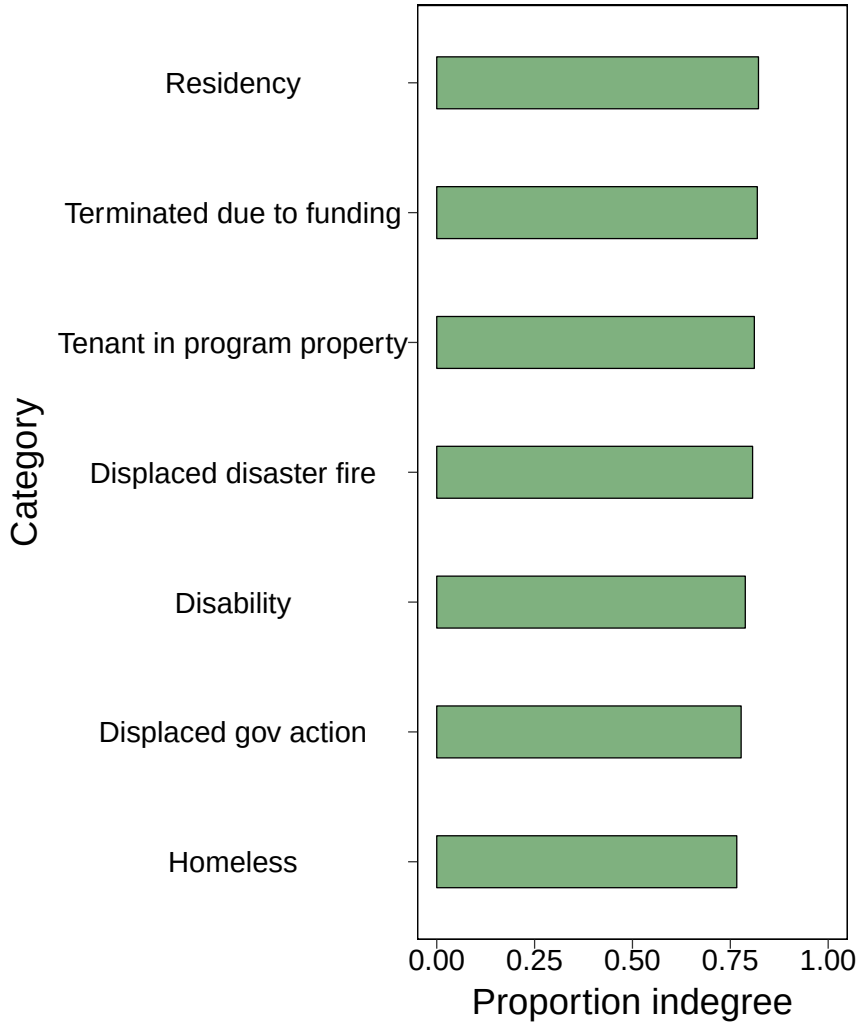
Local housing authorities

- **(Some) transparency about *which* categories are prioritized but little transparency about *why* those categories are prioritized**:

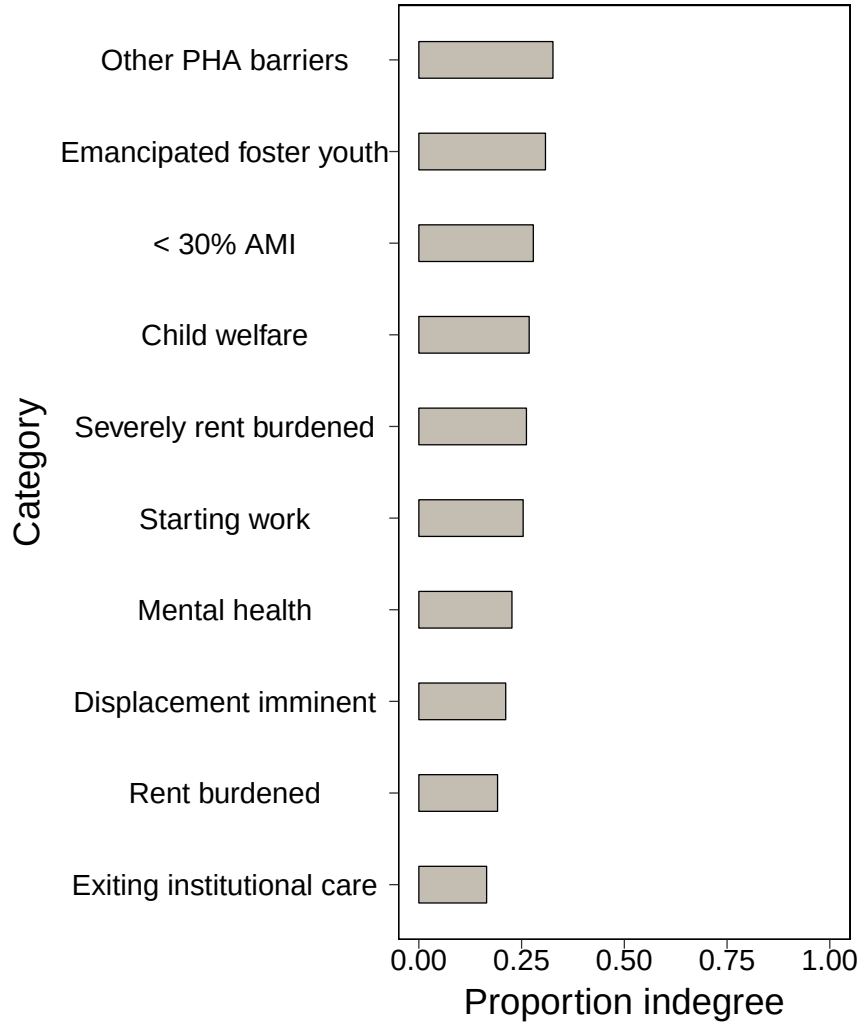
  > *The PHA system of selection preferences must be based on local housing needs and priorities as determined by the PHA. In determining such needs and priorities, the PHA shall use generally accepted data sources. Such sources include public comment on the PHA plan (as received pursuant to § 903.17 of this chapter), and on the consolidated plan for the relevant jurisdiction (as received pursuant to part 91 of this title). (24 CFR § 960.206)*

- "Black-boxing" the moral rationale can facilitate inequality *within prioritized categories* in how they are ranked and inequality in which attributes receive explicit priority
- **Data and Method**: use email outreach and website visits to collect $\sim$ 650 plans that list categories/ranks; present results with random sample of $\sim$ 270; treat each category as a node in a directed network, with edges pointing upwards from lower-ranked categories to higher-ranked ones; use centrality measures to priority

Top-ranked:



Lower-ranked:



How might algorithmic prioritization alter these forms of inequality?

- Need to choose a prioritization target to model reduces "black-boxing" of moral rationale for *why* certain individuals are prioritized:
  - School districts, e.g.:
    - $\mathbf{E}[Y = \text{dropout} | X = x]$
    - $\mathbf{E}[Y = \text{no adult indep.} | X = x]$
    - $\mathbf{E}[\text{Dropout}(D = \text{help}) - \text{Dropout}(D = \text{no help}) | X = x]$
  - Housing authorities, e.g.:
    - $\mathbf{E}[Y = \text{homelessness during next year} | X = x]$
    - $\mathbf{E}[Y = \text{Rent}{>}30\% \text{ of income} | X = x]$
- If organizations are then committed to prioritizing those with the highest $\hat{Y}$, this prioritization may end up favor individuals–for instance, those with substance dependence issues who face high risks of homelessness–who would struggle to gain priority under categorical prioritization

***Manually* select a few attributes**

+	-	Schools	Housing
Transparent *who* is prioritized	Non-transparent *why* they are prioritized; lack of transparency may obscure de-prioritization of those who are needy but perceived as more blameworthy/less deserving	...	...

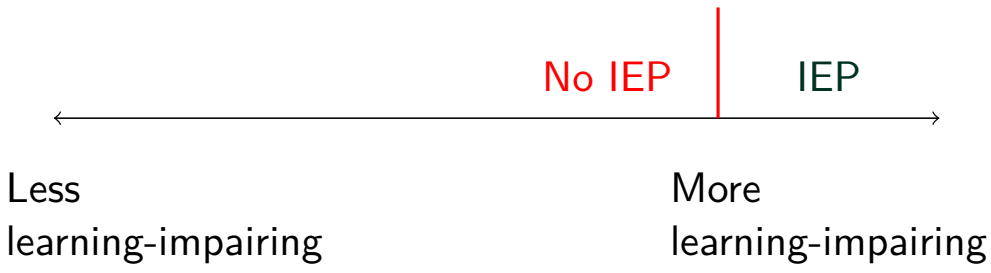
Make those *few attributes* binary

Less complexity (e.g., Starr, 1992)	High stakes of threshold means policymakers to implement highly burdensome procedures to root out false positives	...	Extensive proof required for domestic violence category

School districts: the high stakes of category membership

- Students might have a range of medically-caused learning challenges that mean the student would benefit from extra help:

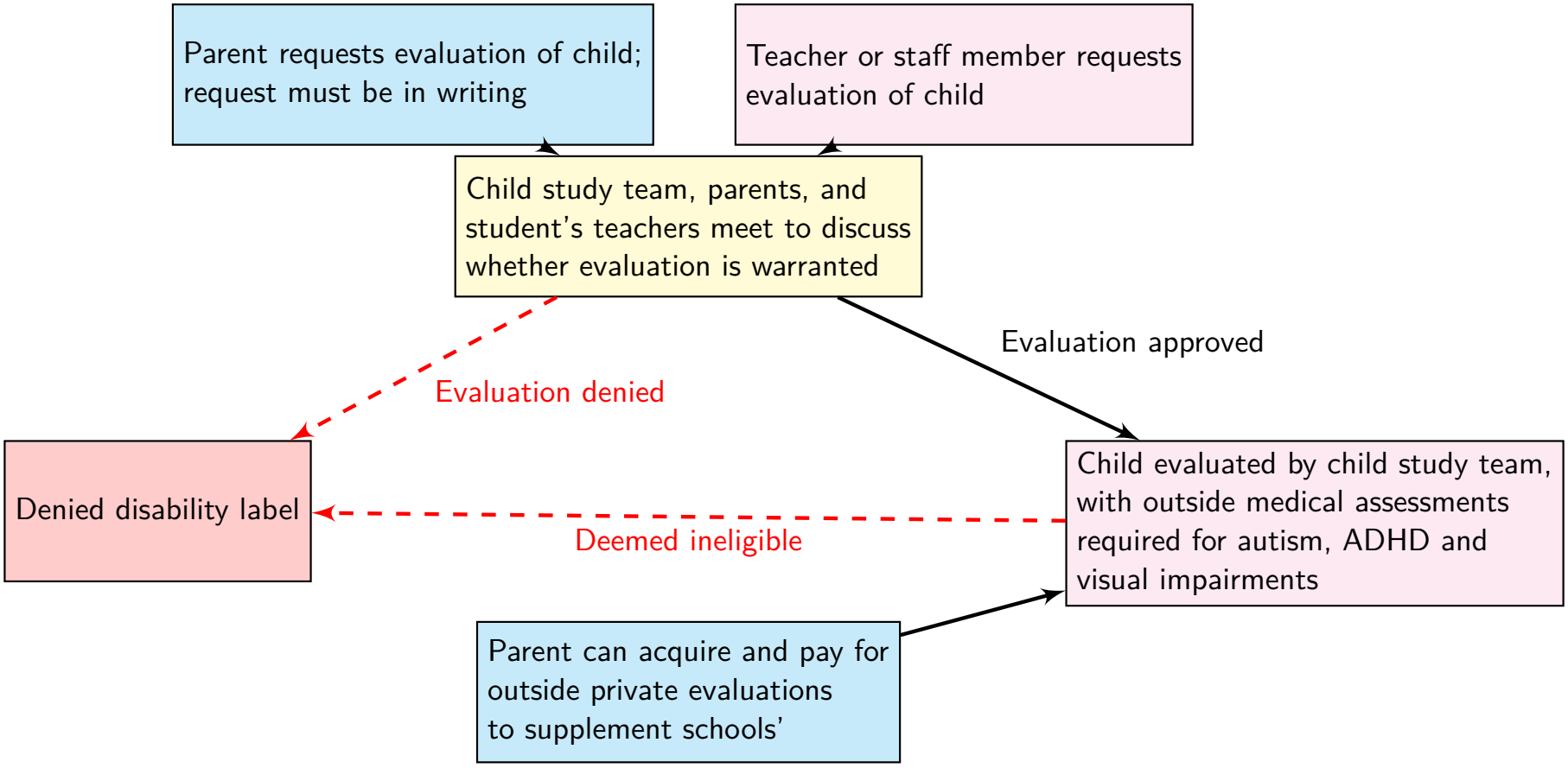
- But school districts draw an implicit threshold along that and group students into a binary category:



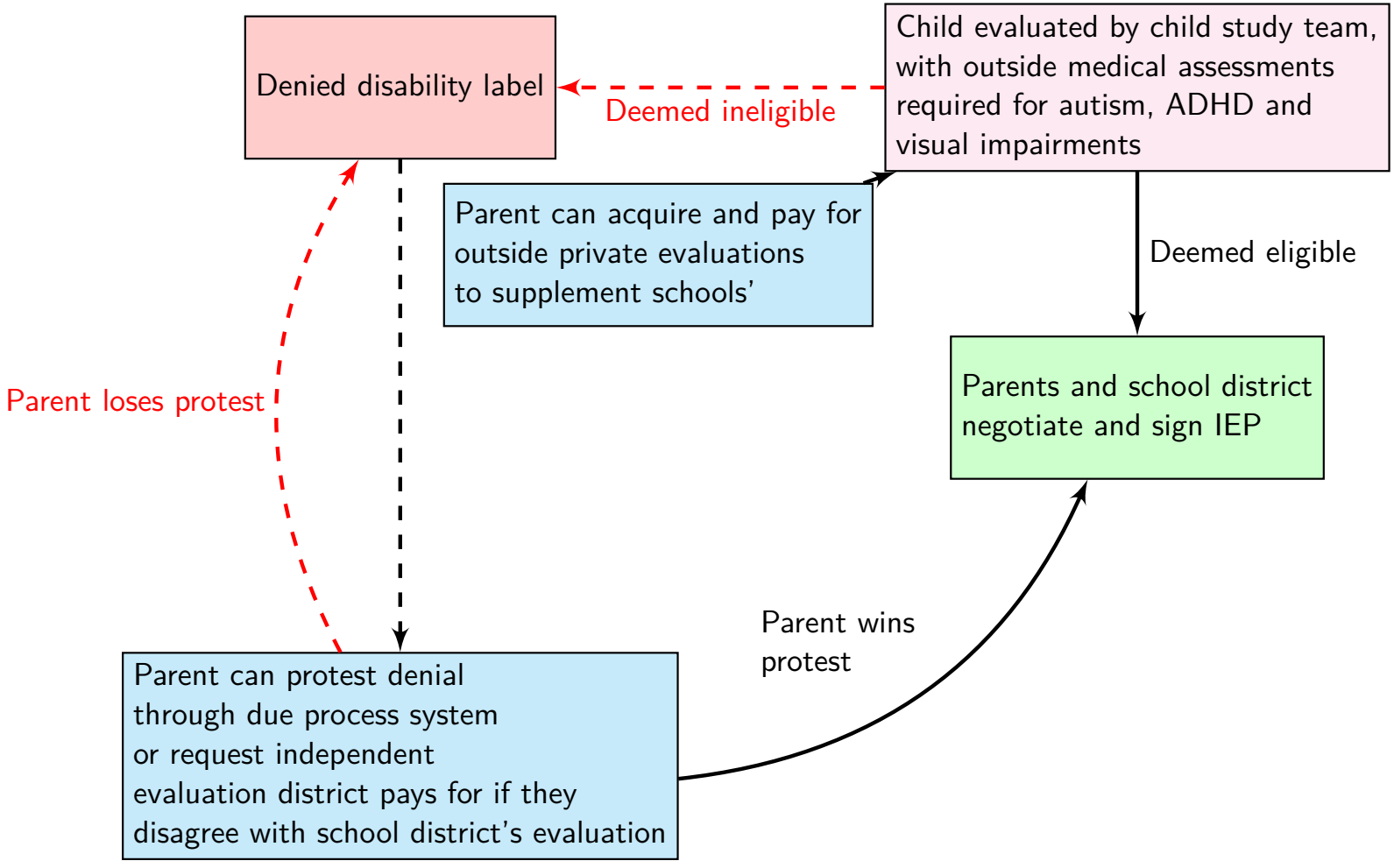
Concern about false positives leads policymakers to require extensive procedures that likely increase false negatives

	School categorizes as having disability	School categorizes as having no disability
No actual disability	False positive	True negative
Actual disability	True positive	False negative

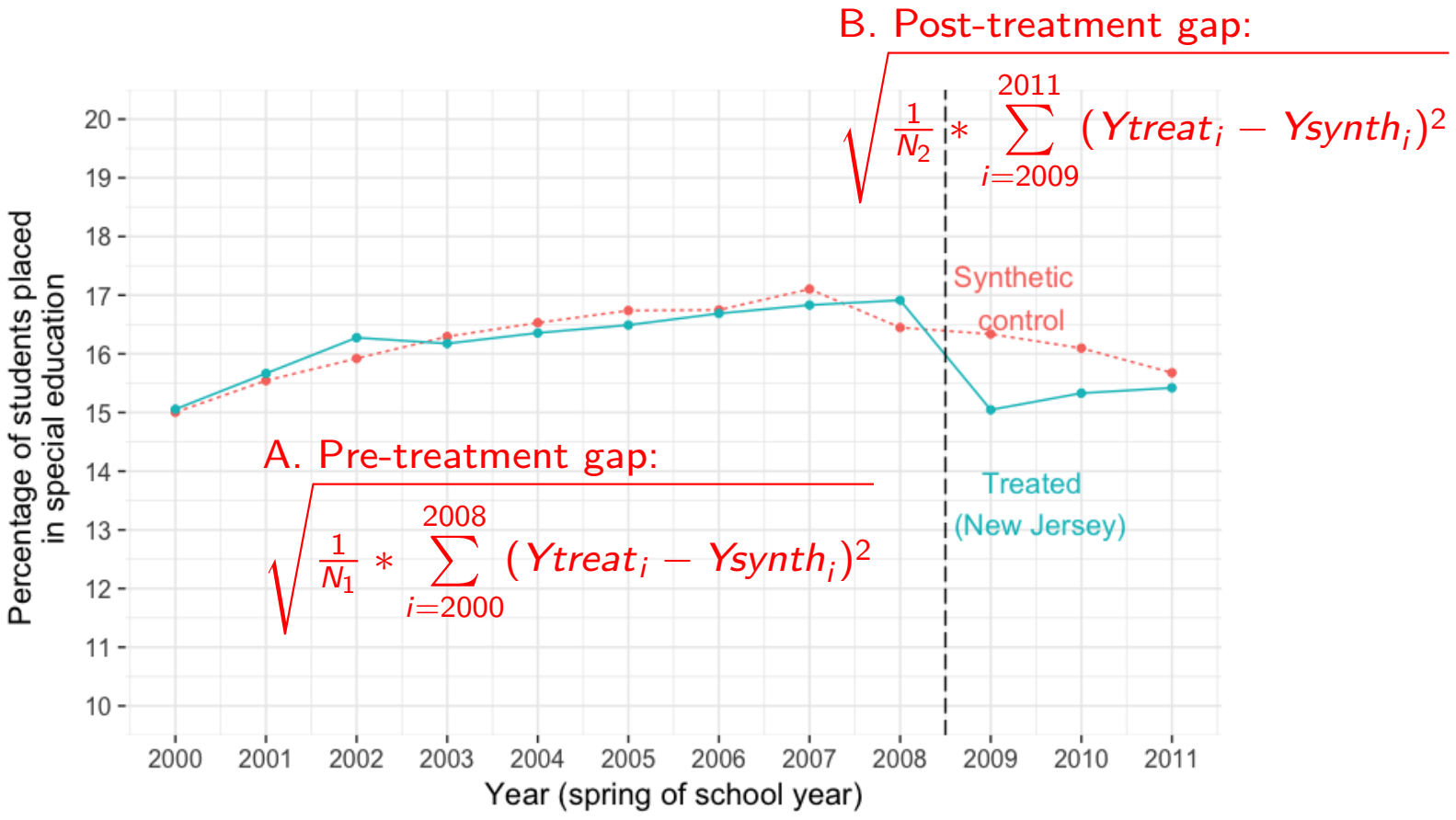
The potential for inequality in parents' ability to navigate those procedures



The potential for inequality in parents' ability to navigate those procedures

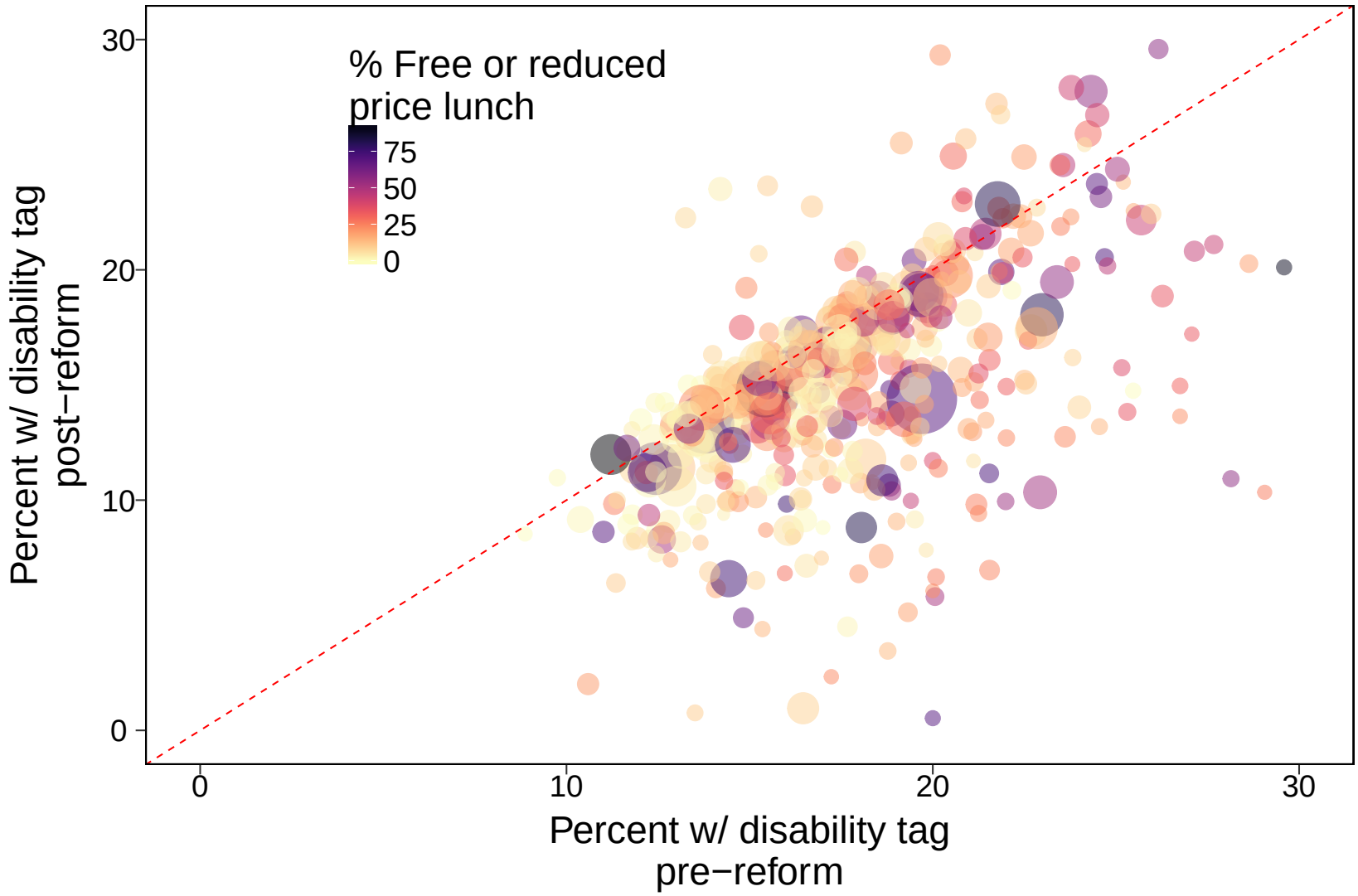


Districts facing cost pressure decide that fewer students fall into the category (potential false negatives)



RMSPE ratio: $\frac{B}{A} = 3.58$ (larger ratio is better; 2/33)

Burden falls disproportionately on students whose parents have less ability to navigate the procedures designed to prevent false positives



How might algorithmic prioritization alter these forms of inequality?

- Two aspects that might lead policymakers to require less extensive procedures to verify category:
  - *Increase in attributes from a few to many*: since estimated need/risk is a function of many attributes, might decrease pressure for orgs to verify each individual attribute
  - *When using algorithm, can use tools to make strategy-proof/resistant or selectively non-transparent*: e.g., Ghani (2016) discussing decreasing gameability of algorithms predicting police officer risk of adverse citizen interaction; market design work on strategy-proof algorithms
- Might result in much more attention to integrity of label being modeled

Concluding

Categorical prioritization		Algorithmic prioritization
***Manually* select a few attributes**		
+	-	+
Transparent *who* is prioritized	Non-transparent *why* they are prioritized	Defining label to model makes moral rationale more transparent and may make inequalities in sympathy for specific categories less salient
Make those *few attributes* binary		
Less complexity	High stakes of threshold means policymakers to implement highly burdensome procedures to root out false positives	Replace extensive procedures to root out false positives with focus on making algorithm in general more strategy-resistant

Thanks!

http://scholar.princeton.edu/rebeccajohnson/

raj2@princeton.edu