Fairness and Machine Learning

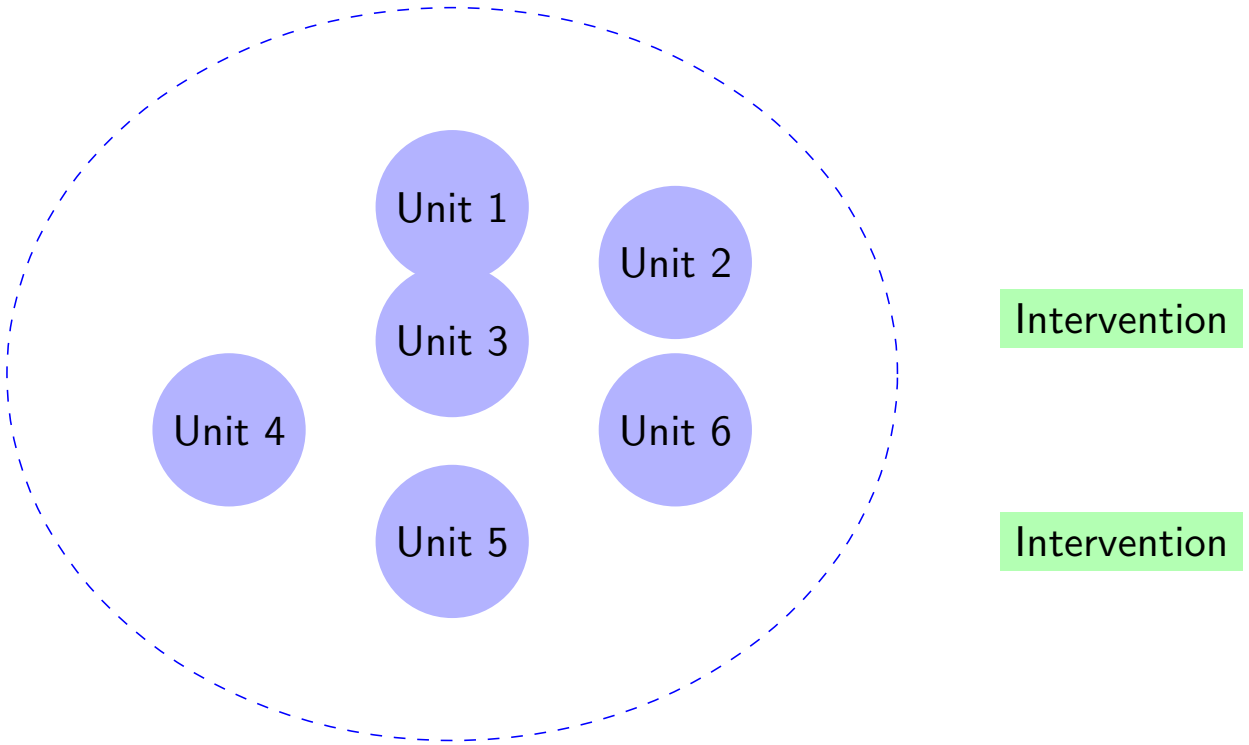
Rebecca Johnson

SocStats Reading Group. 10.25.2018

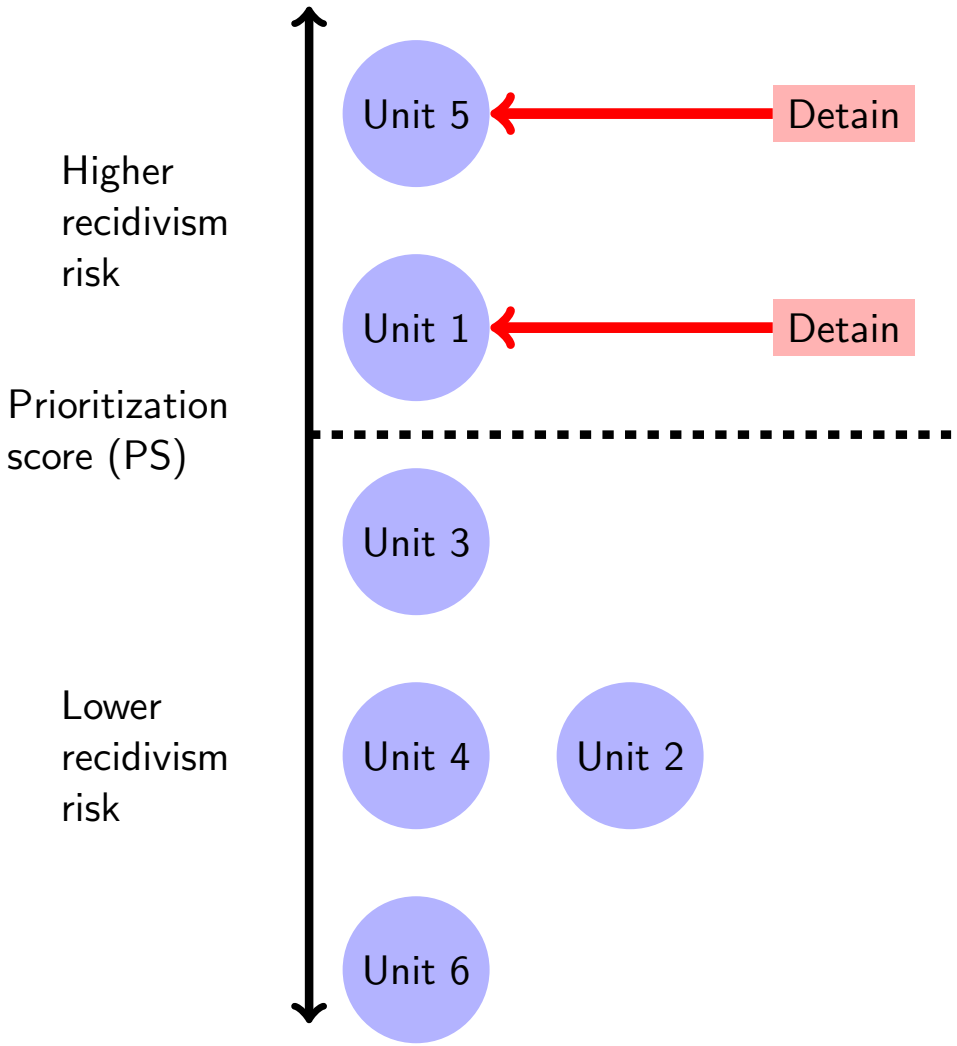
Papers covered

- Main focus (CDG): Corbett-Davies, S., Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv preprint arXiv:1808.00023.
- Others:
  - (LKLLM) Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., Mullainathan, S. (2017, August). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 275-284). ACM.
  - (CR): Chouldechova, A., Roth, A. (2018). The Frontiers of Fairness in Machine Learning. https://arxiv.org/abs/1810.08810

Background: prioritizing among units (people; school districts; etc.) when allocating interventions

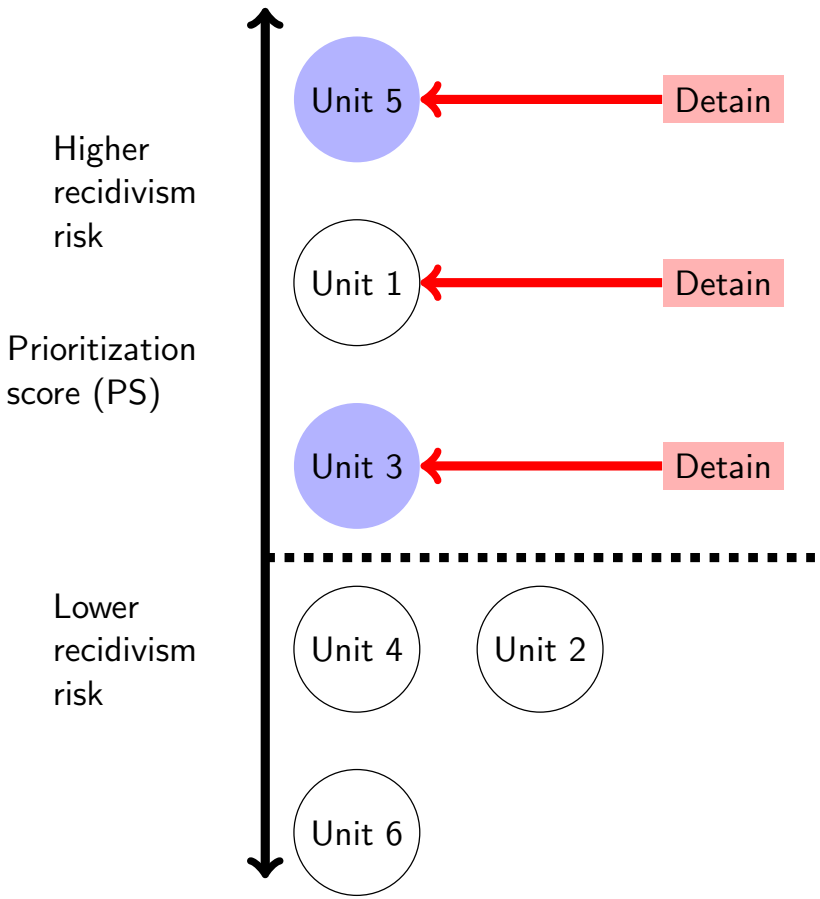


One way of deciding: rank individuals use a scalar prioritization score and use threshold

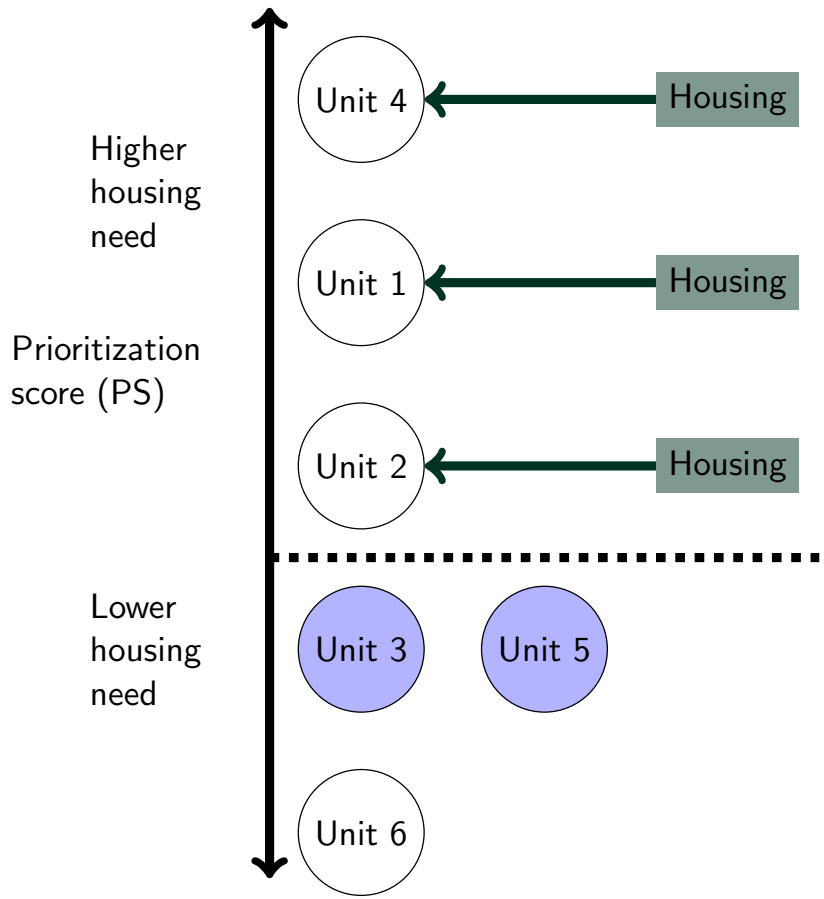


Two types of unfair allocations

Over-allocating a punitive resource to certain subgroups

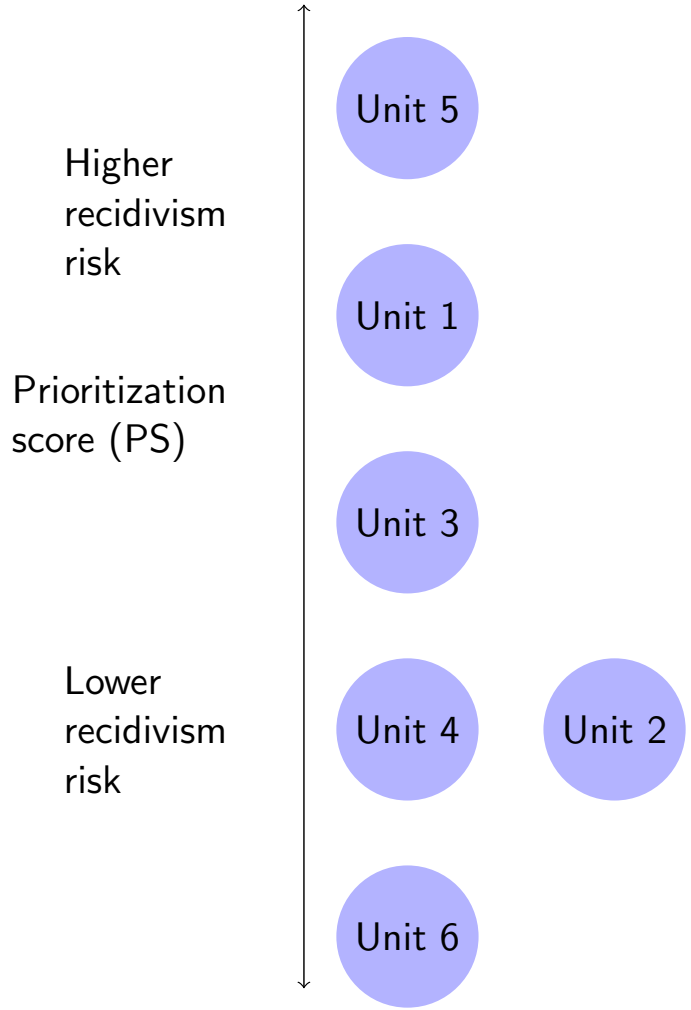


Under-allocating an assistive resource to certain subgroups



How do we generate the rankings? Preliminary notation

- i: individuals
- x: observed **features**/attributes/covariates of i
  - x_u: unprotected attributes we're willing to consider when prioritizing among individuals
  - x_p: protected attributes we're not willing to consider when prioritizing among individuals
- $y \in \{0, 1\}$: binary outcome outcome/target of prediction
- $r(x) = Pr(Y = 1|X = x)$: true risk as a function of observed features

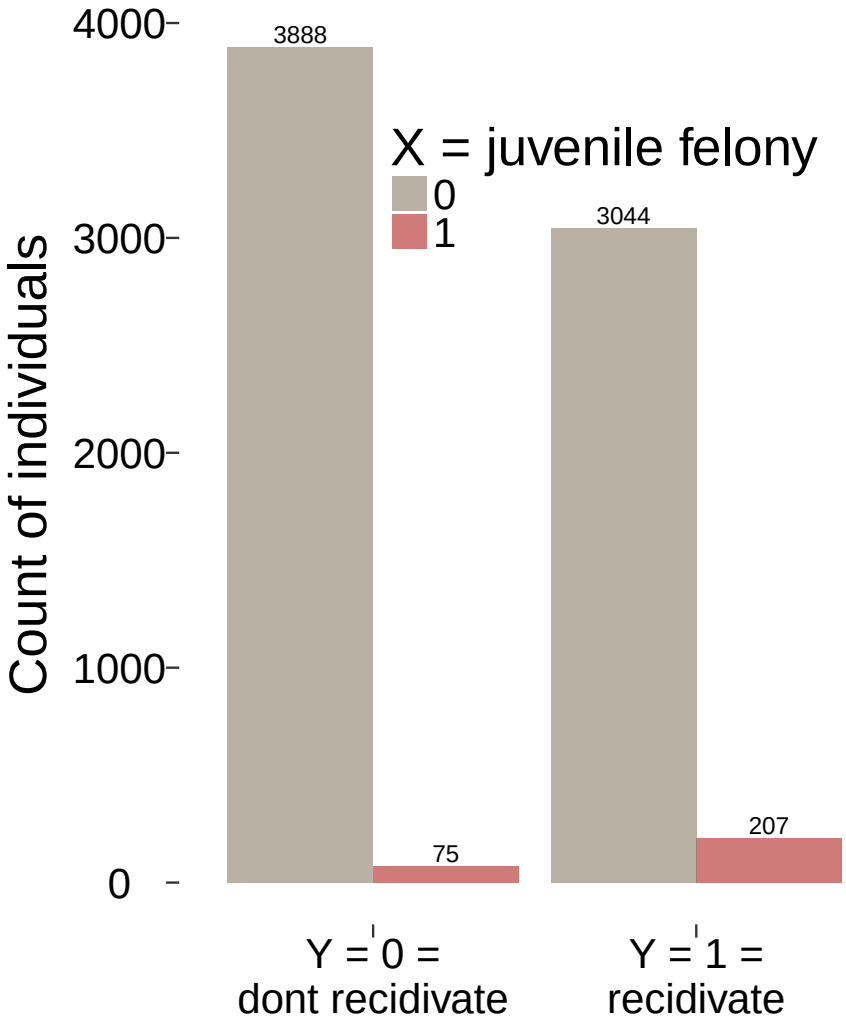


r(x) as a function of one feature

Person	Recidivate?	Juv. felony
	Y	**X**
1	1	1
2	1	1
3	1	0
4	0	0
⋮		
n	0	0

r(x) as a function of one feature

$$\hat{r}(x) = Pr(\text{recidivate} = 1 | \text{juv. felony})$$



$$Pr(recid = 1 | jv = 1) = \frac{Pr(jv = 1 | recid = 1) * Pr(recid = 1)}{Pr(jv = 1)}$$

$$Pr(recid = 1 | \text{jv} = 1) = \frac{\frac{207}{207+3044} * 0.45}{0.039}$$

$$Pr(recid = 1 | \text{jv} = 1) = 0.72$$

$$Pr(recid = 1 | \text{jv} = 0) = 0.43$$

r(x) as a function of one feature: problems with prediction

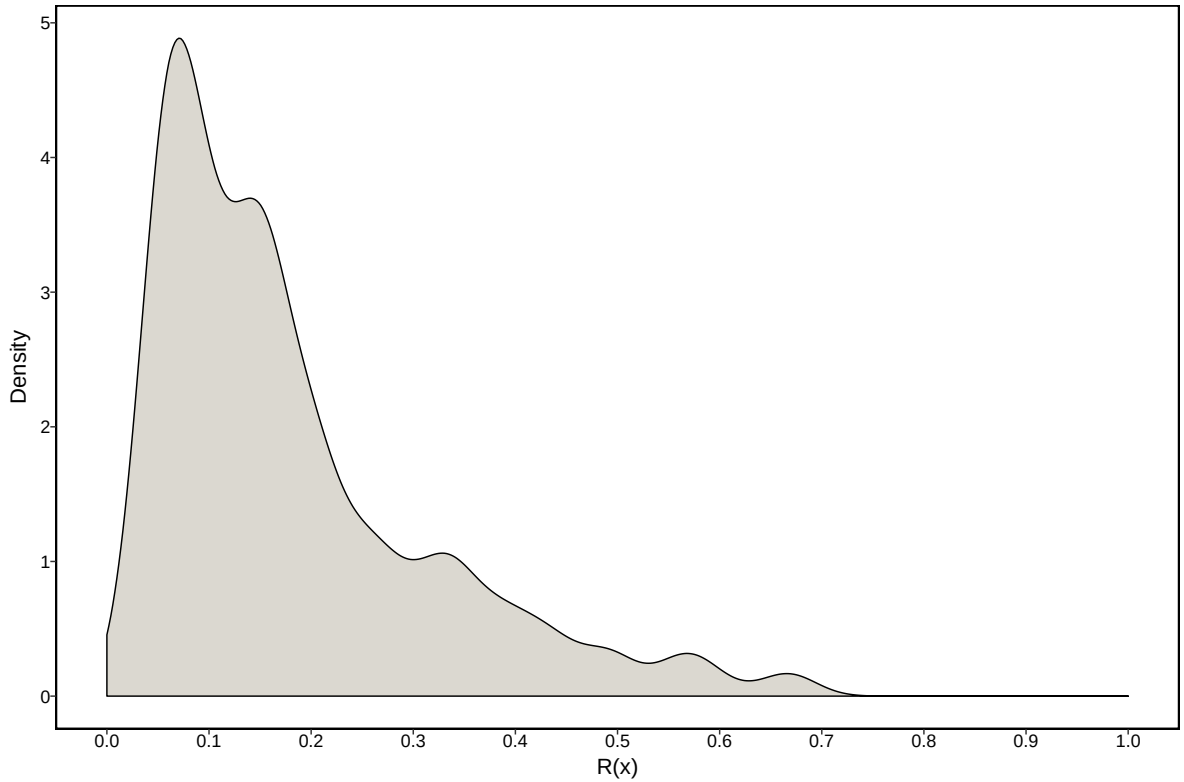
Person	Recidivate?	Juv. felony	$\hat{r}(x)$
	Y	**X**	
1	1	1	0.72
2	1	1	0.72
3	1	0	0.43
4	0	0	0.43
⋮			
n	0	0	0.43

Usually, want to calculate r(x) as a function of many features/complex interactions between features...

id	sex	age	race	Juv. fel.	Juv. misd.	Priors	Charge
8670	Female	31	White	0	0	2	Assault
7898	Male	35	Black	0	0	2	Trespassing/Construction Site
4390	Male	66	White	0	0	0	Grand Theft in the 3rd Degree
8613	Male	33	White	0	0	2	DUI Level 0.15 Or Minor In Veh
6107	Female	24	White	0	0	0	Uttering a Forged Instrument
5449	Male	32	White	0	0	1	Burglary Structure Unoccup
4615	Female	24	Black	0	0	2	Unlaw LicTag/Sticker Attach
1850	Male	32	Black	0	0	4	arrest case no charge
8174	Male	27	Black	0	0	1	Grand Theft (Motor Vehicle)
8759	Female	36	Black	0	0	4	arrest case no charge

How to use r(x) to make decisions

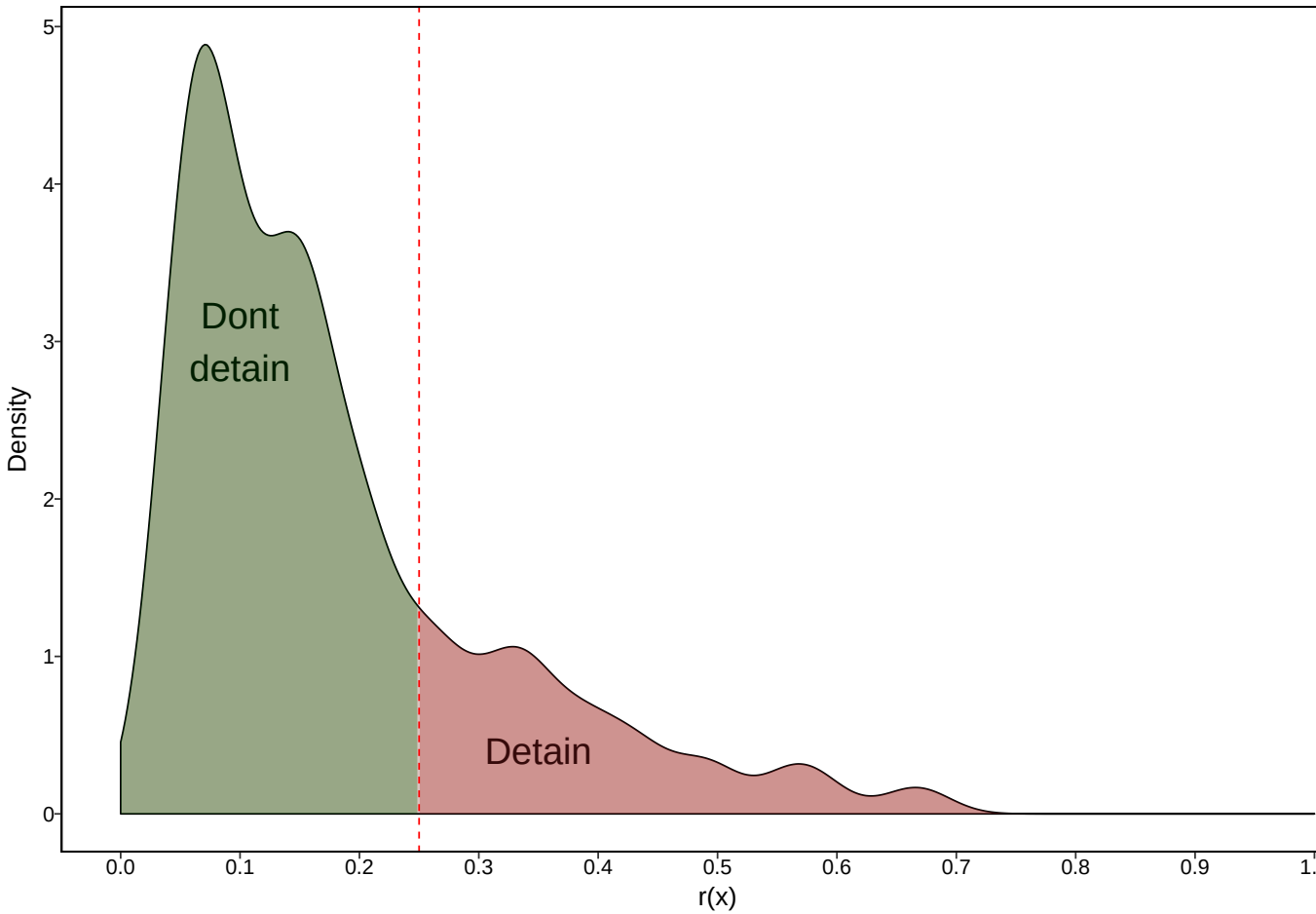
1. Start with a distribution of r(x) that represents an individual's risk of an outcome conditional on some combination of features (X)



How to use $r(x)$ to make decisions

2. Choose a threshold t to separate individuals into those who receive the intervention ($d = 1$) and those who do not ($d = 0$):

$$d(x) = \begin{cases} 1 & \text{if } r(x) \geq t \\ 0 & \text{if } r(x) < t \end{cases}$$



How do we choose (or end up with) a specific t? (scarce resource)

Scarce resource: rather than choosing t *a priori*, have a fixed number of interventions to allocate (e.g., 20 housing vouchers) so may:

1. Rank units by $r(x)$ (estimated in a model as s(x) or $\hat{r}(x)$)
2. Take top-20 units
3. t is just observed risk for the 20th-unit

How do we choose (or end up with) a specific t? (non-scarce resource)

- May define threshold with reference to four cells:

	(Counterfactual) Outcome	
Decision	*Recidivate*	*Not recidivate*
Detain	b_{11}	c_{10}
Not detain	c_{01}	b_{00}

- In words:
  - b_{11}: Benefits of detaining an individual who would have gone on to recidivate (true positive; public safety)
  - b_{00}: Benefits of releasing people who would have gone on to not recidivate (true negative; preventing unjust detention)
  - c_{10}: Costs of detaining people who would have gone on to not recidivate (false positive; preventing unjust detention)
  - c_{01}: Costs of releasing people who would have gone on to recidivate (false negative; public safety)

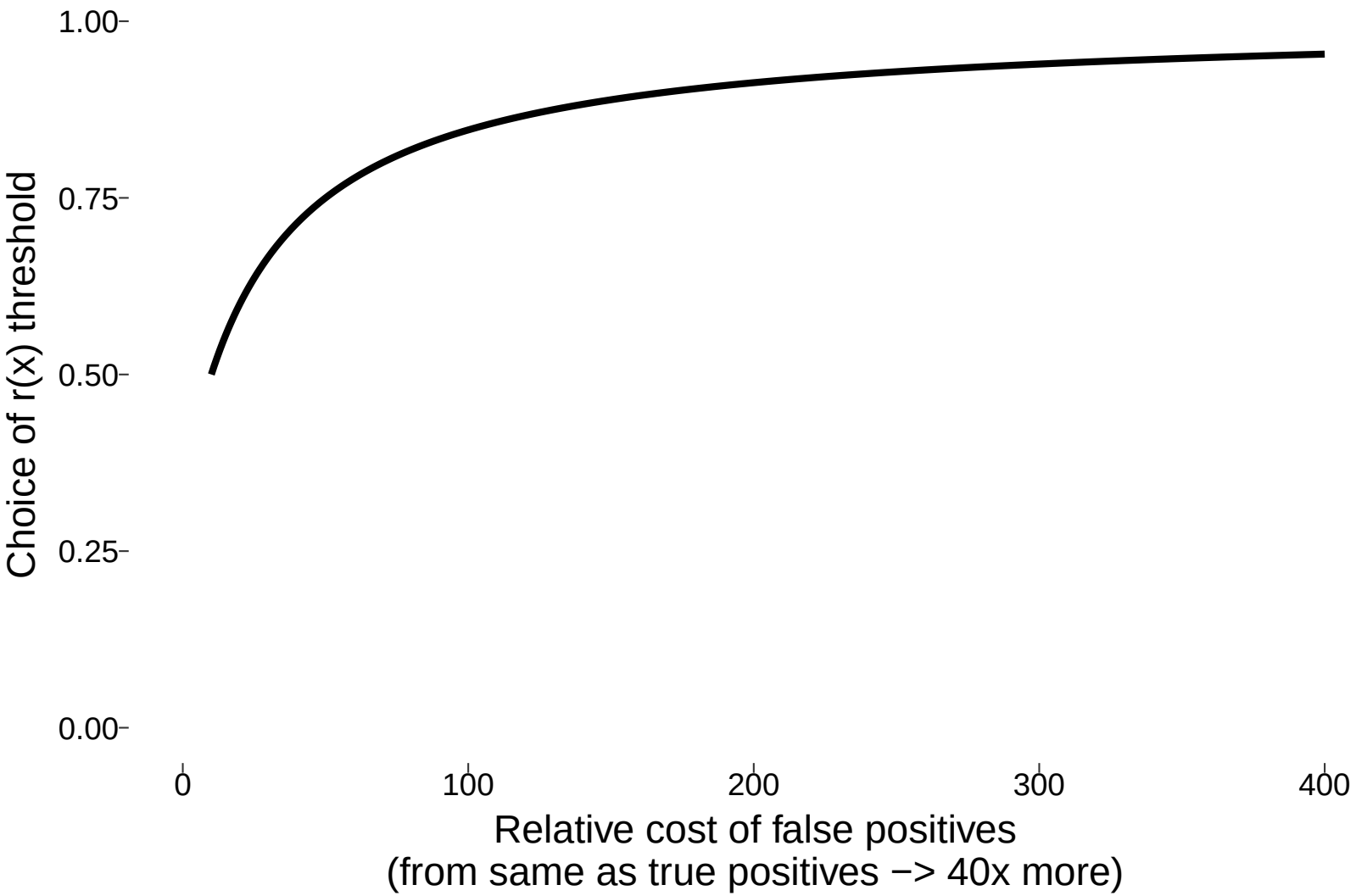
Different weighting of these priorities yields different thresholds

Shows that even before examining fairness issues with reference to advantaged/disadvantaged subgroups, choice of threshold can reflect substantive ideas about fairness

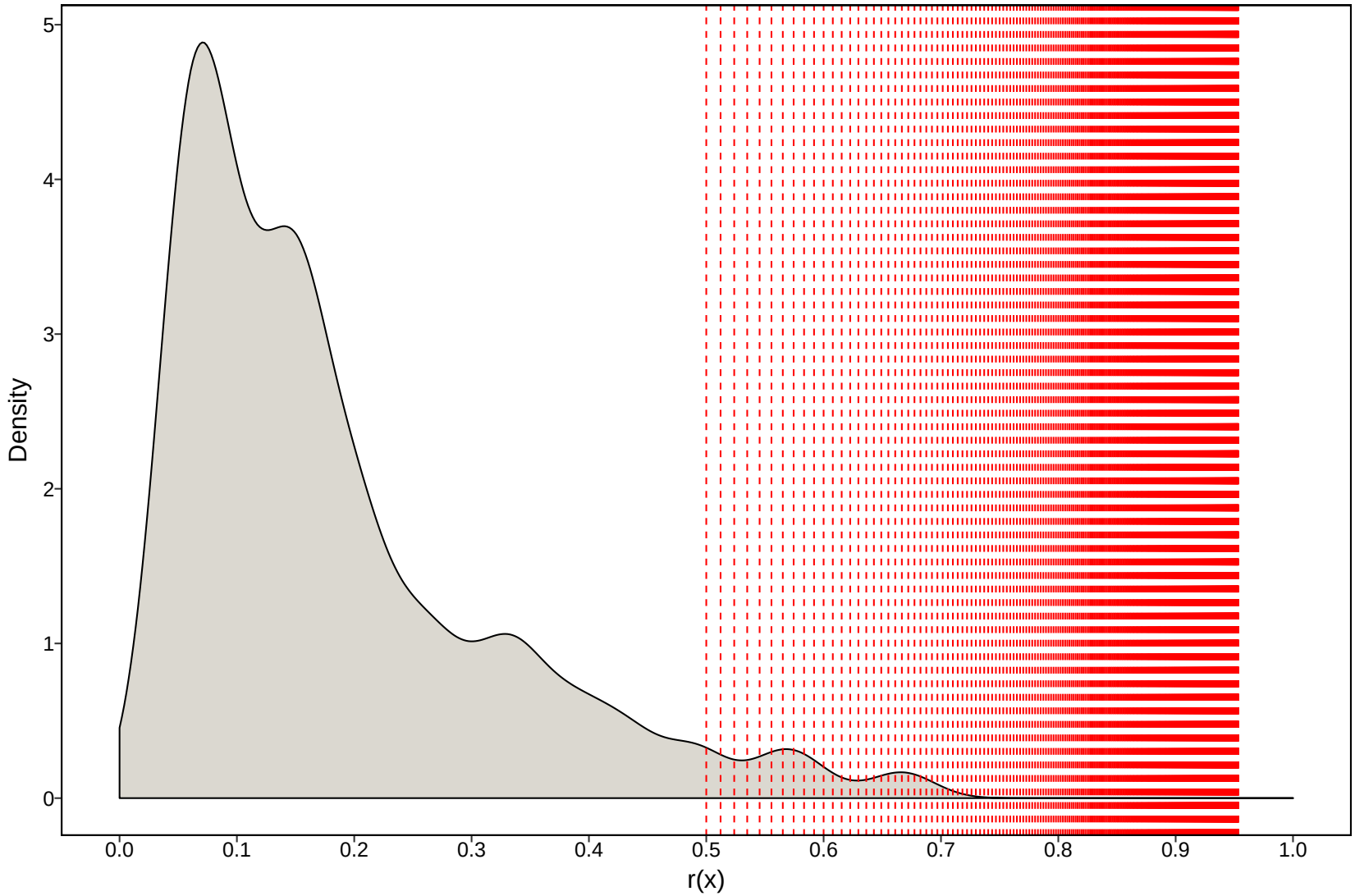
$$r(x) \geq \frac{b_{00} + c_{10}\ \text{(preventing unjust detention)}}{b_{00} + c_{10} + b_{11} + c_{01}\ \text{(preventing crime)}}$$

$$r(x) \geq \frac{10 + c_{10}\ \text{(preventing unjust detention)}}{10 + c_{10} + 10 + 10\ \text{(preventing crime)}}$$

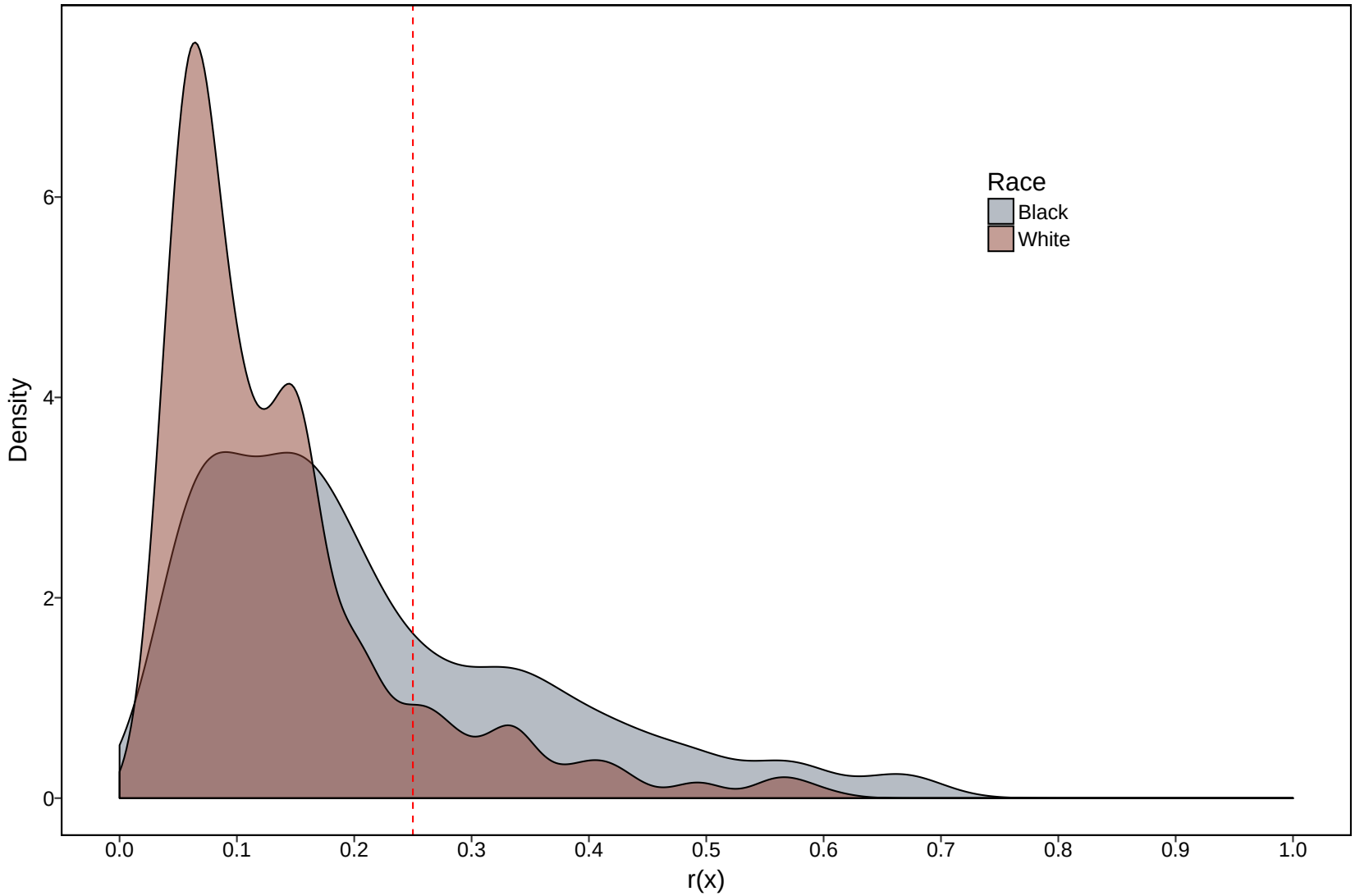
Different weighting of these priorities yields different thresholds



At a high enough threshold, detain no indviduals

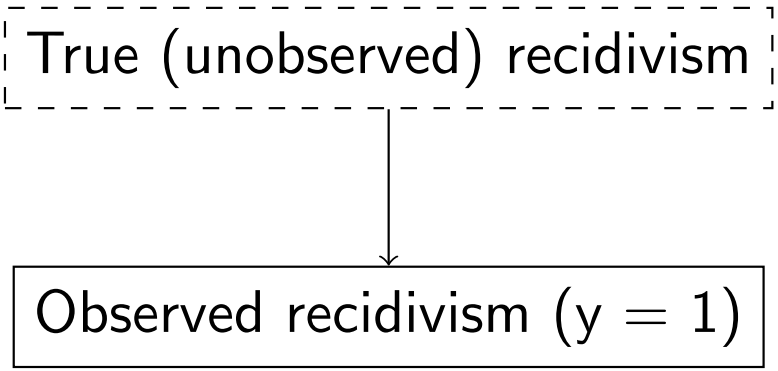


Introducing subgroups



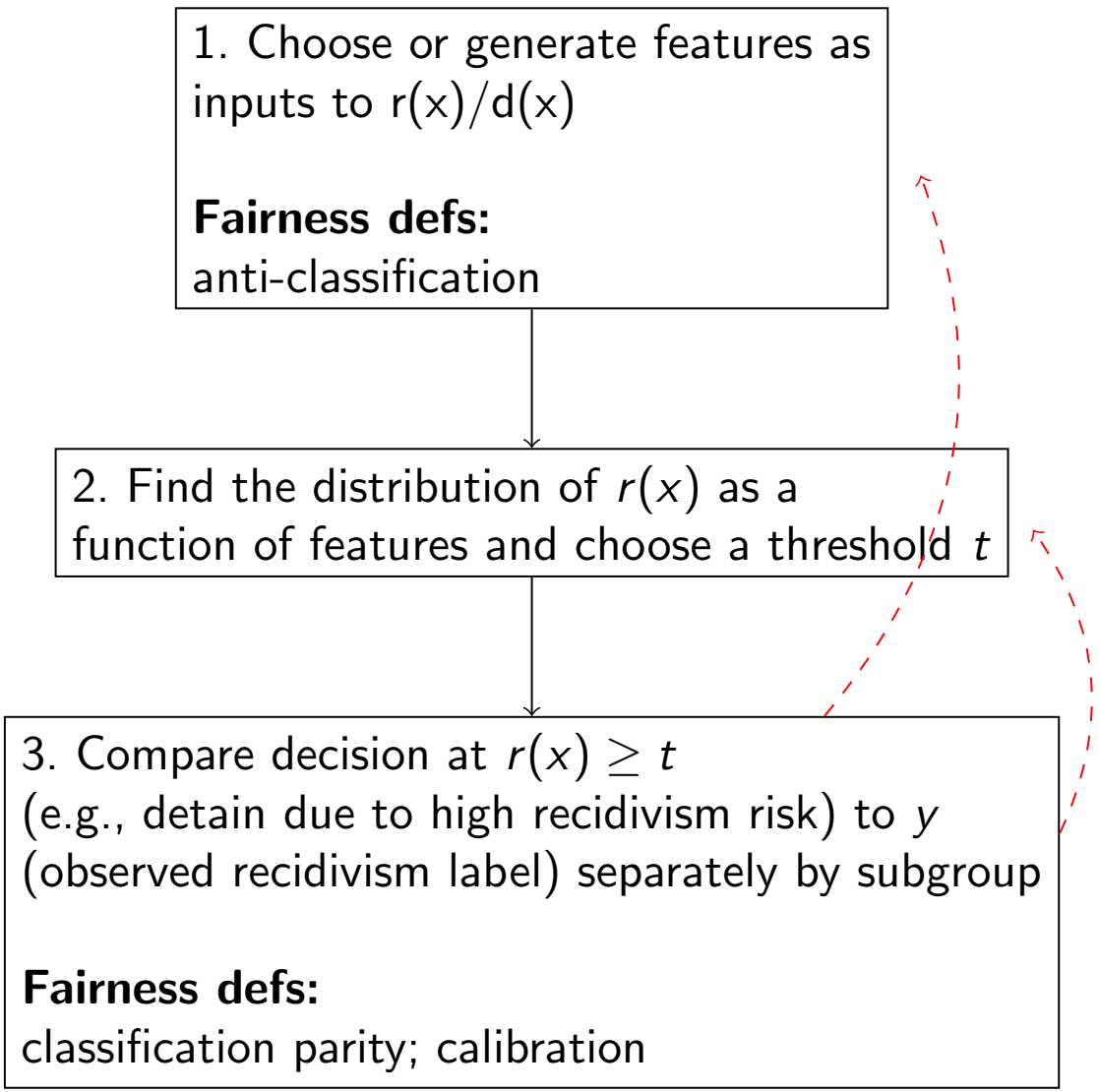
Assumptions for now (later, we'll cover problems with these assumptions!)

- Label (recidivism) captures construct of interest (commission of repeated crime) equally well for both groups



- Have enough of each subgroup in the training data to accurately learn relationships between features and label; likewise, have enough in test data to accurately evaluate predictions
- We know $r(x)$ (the true distribution of risk in each subgroup) so differences stem from true differences in that distribution/rather than differences that emerge through bad estimation of that true risk

Three definitions of algorithmic fairness vs. stage of the process in using a model to learn $r(x)$



Anti-classification: definition and problems

- *In math*: $d(x) = d(x')$ for all x, x' such that $x_u = x_u^{'}$
- *In words*: two individuals with the same values of unprotected features have the same decision; that is, a protected feature like race doesn't change the score/change the decision
- Problems:
  - Suppose:
    - Original model (OM): $r(X_u, X_p)$
    - Anti-classification model (AM): $r(X_u)$
    - Error: $\sum_{i=1}^{n}(y_i - \hat{r}_i)^2$
  - Importance of (protected-attribute) representative training data suggests #2 is more prevalent than #1
    1. $error_{am} \leq error_{om}$
    2. $error_{am} > error_{om}$
  - $cor(X_p, X_u)$: proxies for protected attribute

Classification parity

- *Previous steps:* allow both X_u and X_p into the model that estimates $r(x)$ and generate $\hat{r}(x)$
- Return to the 2×2 table for judging costs and benefits:

Decision	**Outcome**	
	Recidivate	*Not recidivate*
Detain $(r(x) \geq t)$	b_{11} (TP)	c_{10} (FP)
Not detain $(r(x) < t)$	c_{01} (FN)	b_{00} (TN)

- Calculate table separately by protected attribute
- Compare chosen metric that you think should be equal between the two groups
- If metrics are unequal, three choices:
  1. Change feature set
  2. Use same feature set but different model
  3. Use same feature set + same model, but different t (either for whole sample or $t(X_p = 1)$; $t(X_p = 0)$

Choice of metric is linked to values about relative harms of different types of classification errors

- For a punitive intervention like detention, likely:
  - Care more about a higher false positive rate discrepancies (higher rate of inappropriate detention of those who don't recidivate in disadvantaged subgroup)...
  - Than false negative rate (higher rate of inappropriate release)
- For a helpful intervention like housing, vice versa

Classification parity: focus on false positive rate by group

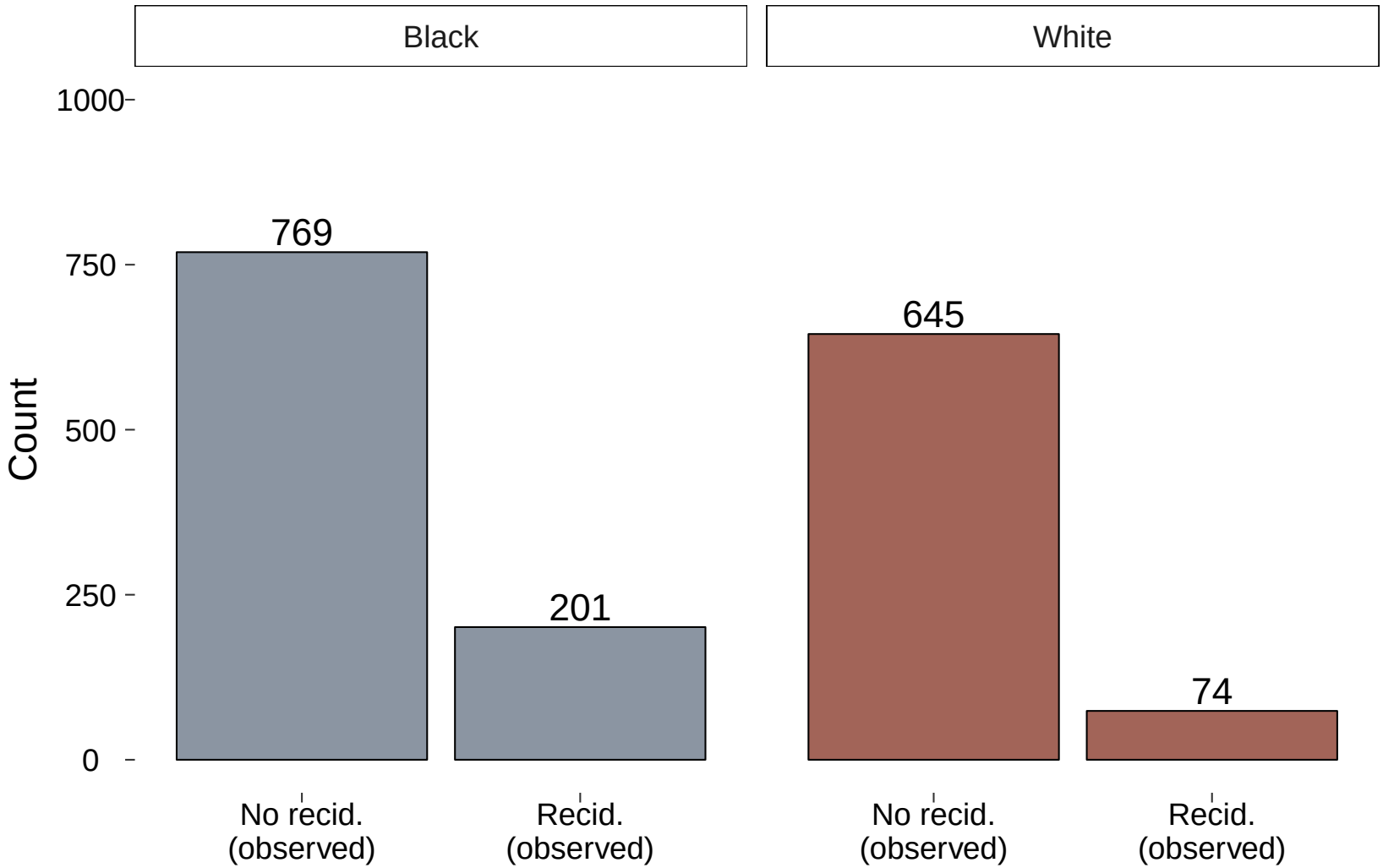
- Confusion matrix:

	Outcome	
Decision	*Recidivate*	*Not recidivate*
Detain ($r(x) \geq t$)	b_{11} (TP)	c_{10} (FP)
Not detain ($r(x) < t$)	c_{01} (FN)	b_{00} (TN)

- False positive rate (note: depends on t!):

$$FPR = \frac{FP}{FP + TN}$$

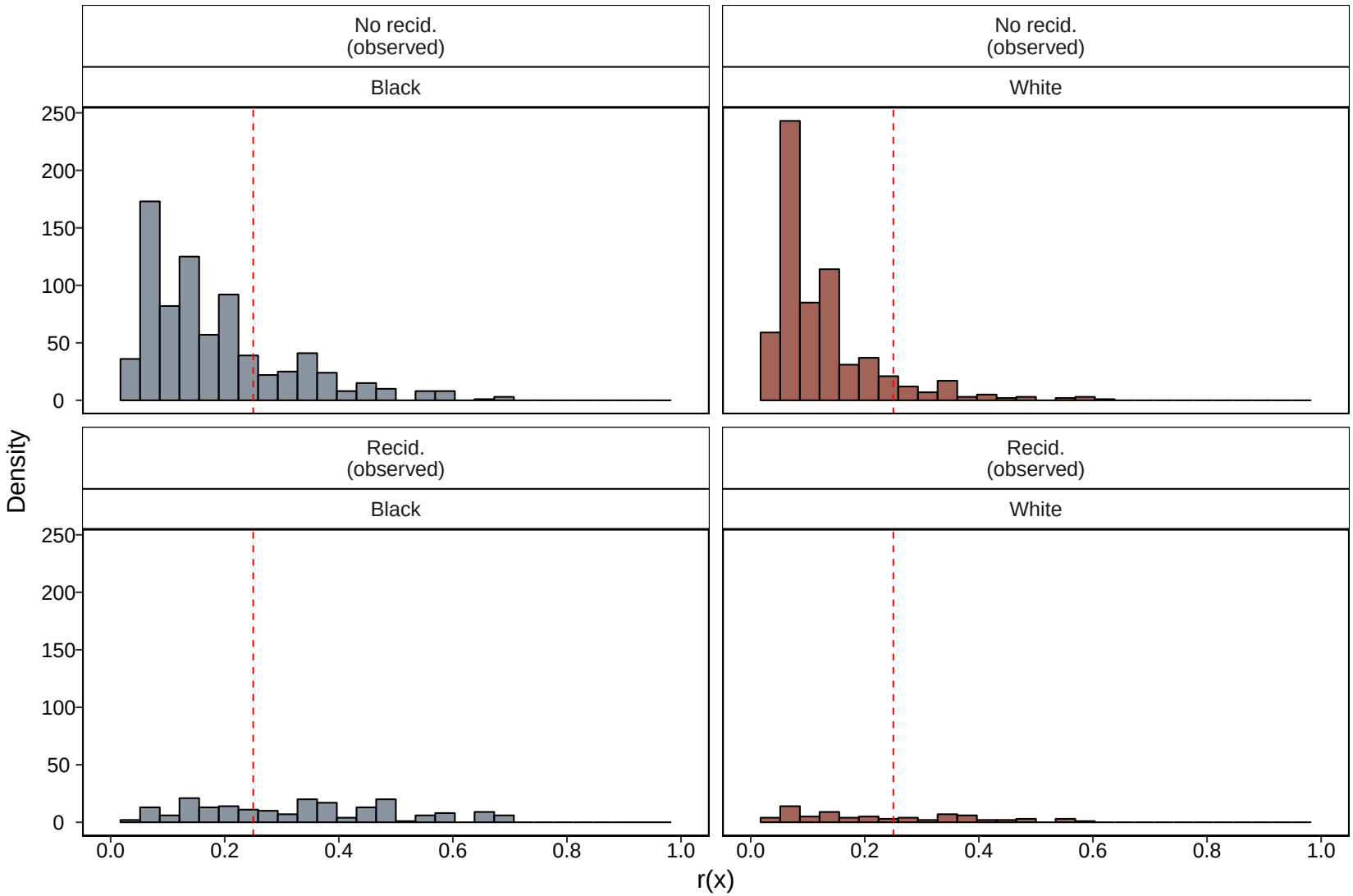
Example of FPR by group at one t: (observed) base rate differs between groups



When the base rate differs, the distribution of risk ($r(x)$), which incorporates information about other features, will also differ

Example of FPR by group at one t

$t = 0.25$; risk scores learn those labels

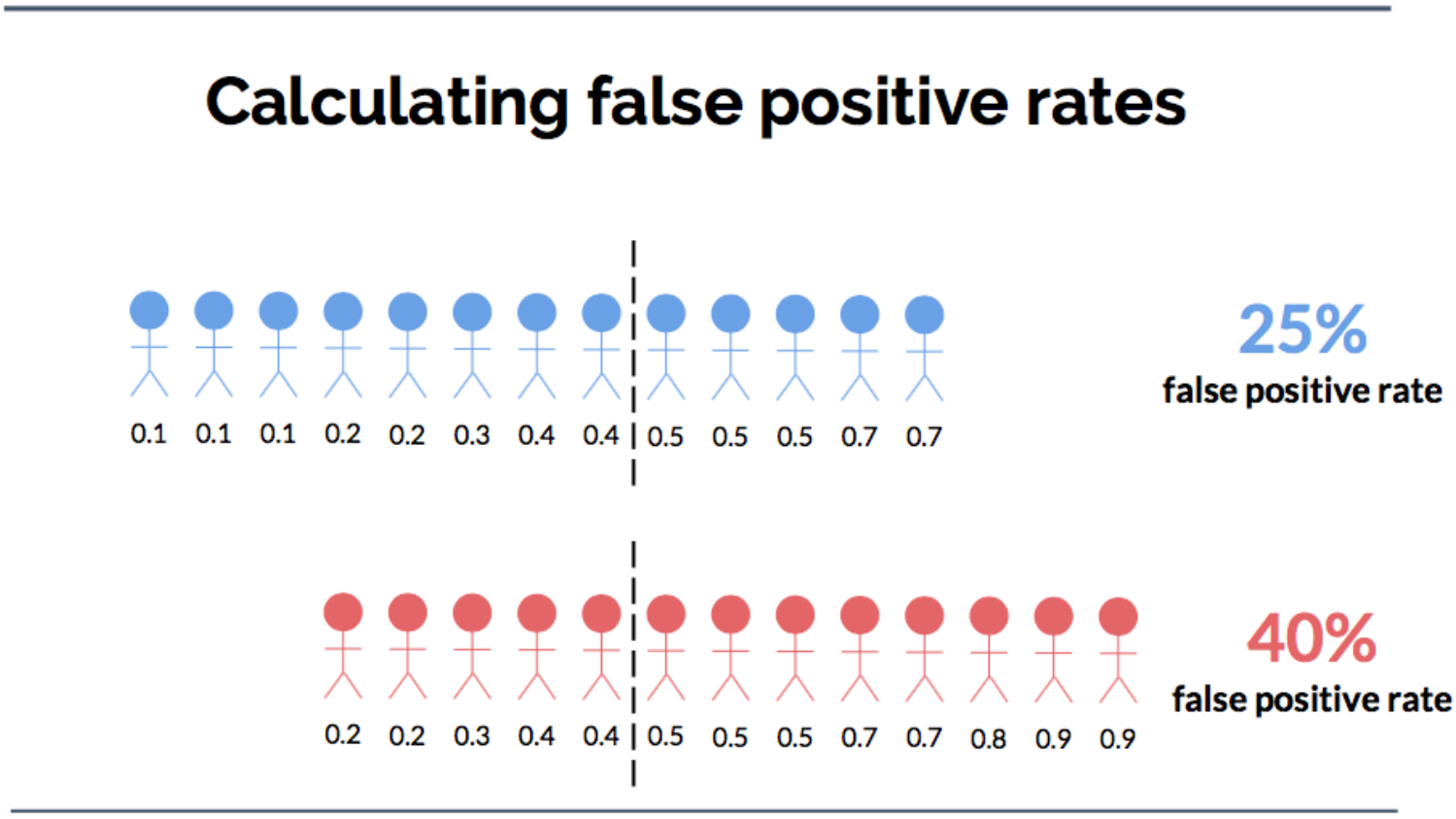


Example of FPR by group at one t (0.25)

```r
metrics_bygroup <- function(threshold,
                            data_withpred,
                            score_col,
                            label_name,
                            decision_name,
                            group_name){

  ## decision for each obs based on threshold
  data_withpred[decision_name] = ifelse(data_withpred[[score_col]] >= threshold,
                                        1, 0)

  ## classify error for each obs
  data_withpred['fp'] = ifelse(data_withpred[[decision_name]] == 1 &
                               data_withpred[[label_name]] == 0,
                                     1, 0)
  data_withpred['tp'] = ifelse(data_withpred[[decision_name]] == 1 &
                                 data_withpred[[label_name]] == 1,
                               1, 0)
  data_withpred['fn'] = ifelse(data_withpred[[decision_name]]  == 0 &
                                 data_withpred[[label_name]] == 1,
                               1, 0)
  data_withpred['tn'] = ifelse(data_withpred[[decision_name]]  == 0 &
                                 data_withpred[[label_name]] == 0,
                               1, 0)

  ## df with various metrics
  metric_summary = data_withpred %>% group_by_(.dots = group_name) %>%
                   summarise(fpr = sum(fp)/sum(fp + tn),
                             fnr = sum(fn)/sum(fn + tp),
                             precision = sum(tp)/sum(tp + fp)) %>%
                   mutate(threshold = threshold)

  ## return
  return(metric_summary)

}

metrics_byrace_onethres = metrics_bygroup(threshold = 0.25,
                          data_withpred = compas_score_better,
                          score_col = 'score',
                          label_name = 'two_year_recid',
                          decision_name = 'detain',
                          group_name = 'race')
```

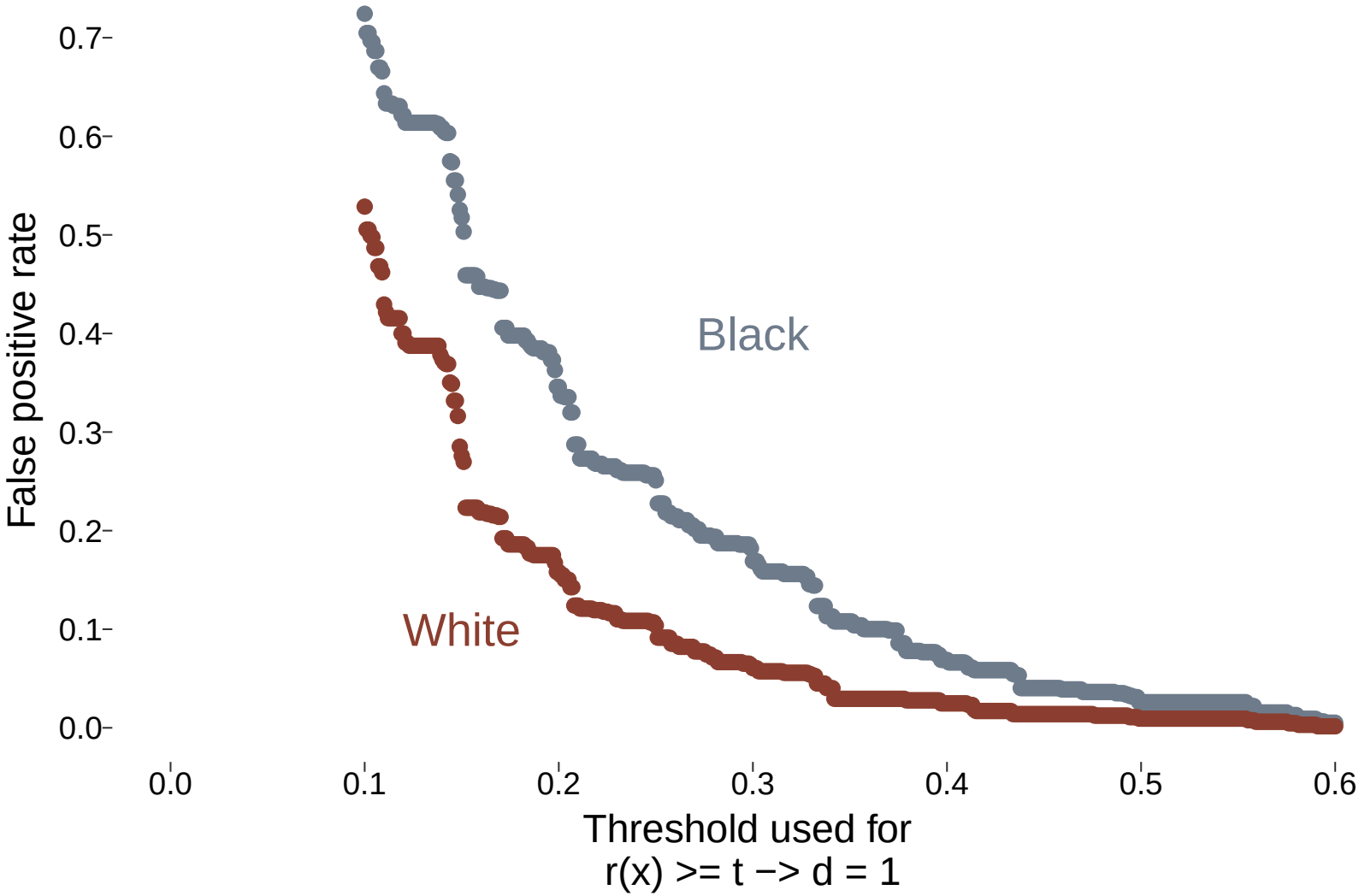
Race	FPR	FNR	Precision
Black	0.25	0.36	0.40
White	0.10	0.55	0.33

Or as the authors put it...

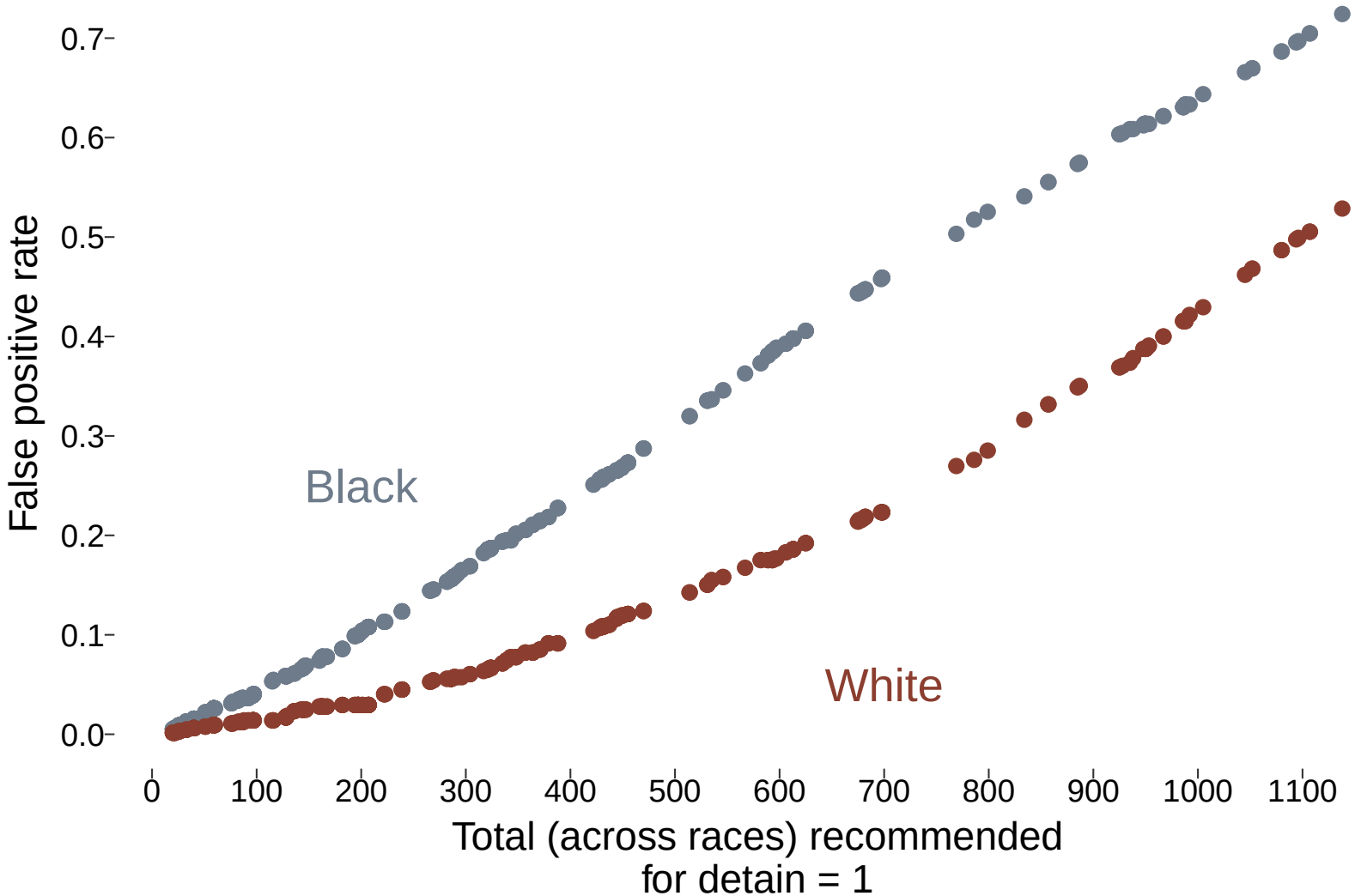


Source: https://policylab.stanford.edu/projects/defining-and-designing-fair-algorithms.html

How does this changes at different t (but t remaining equal for both groups)?



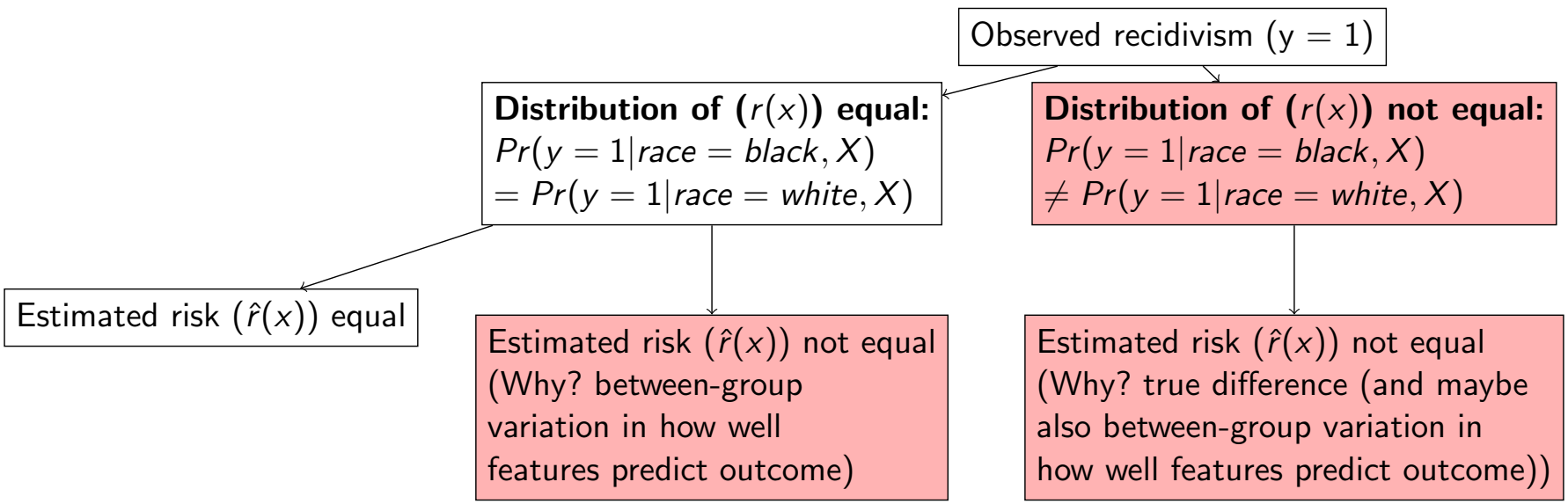
How does this changes at different t (but t remaining equal for both groups)?



Summarizing where we're going next:

- Previous slides/paper show:
  1. Distribution of $r(x)_b \neq r(x)_w$
  2. Because of #1, using the same threshold across the two groups leads to different error rates
- Leaves two options if we want to equalize error rates across groups:
  1. **Change the distribution** of $r(x)$ by changing the features so that the distribution is (more) equal between the two groups
  2. **Keep the distribution but change the threshold** used to translate $r(x)$ into a decision

First option: can we change the distribution of $r(x)$? Do we want to? Two pathways to between-group differences in $r(x)$

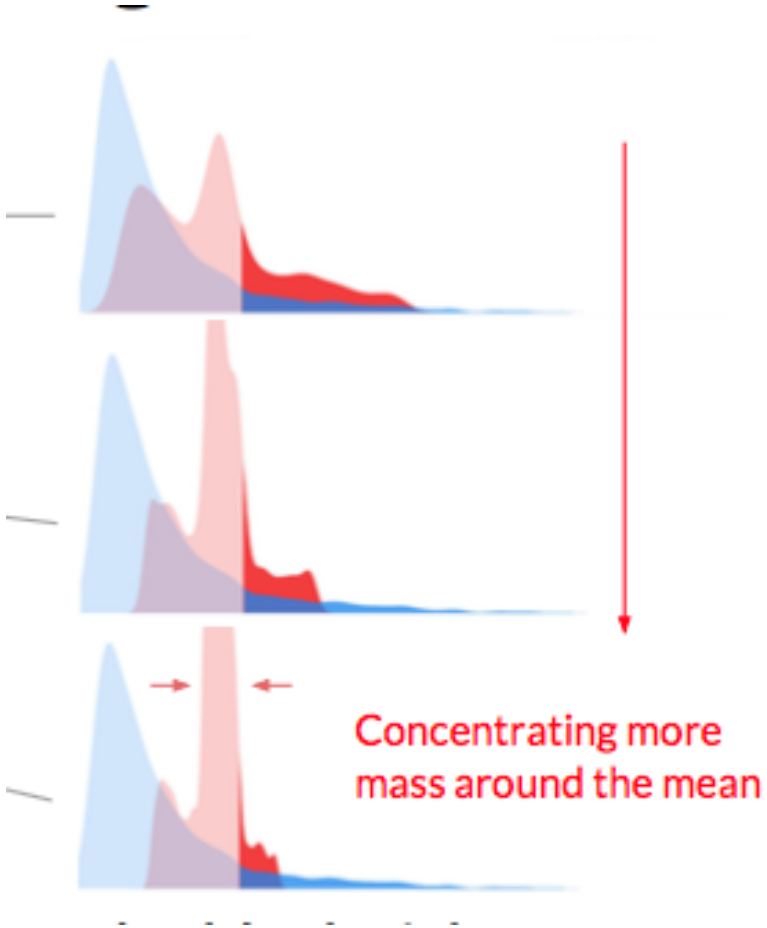


In both cases, $r(x)$ of each group is a function of...

- Things we can't change:
  - y (label)
- Things we can do to change $r(x)$ (true $Pr(recid|race, \text{other features})$):
  - Which features other than race we condition on
- Things we can do to change $\hat{r}(x)/s(x)$ (model-estimated risk)
  - Which model we use to estimate $r(x)$ (e.g., basic logistic regression v. SVM)

But there are sometimes perverse consequences from changing the feature set to change $r(x)$

- One way to get more similar risk distributions is to remove informative features in ways that lead to a less informative risk distribution for some subgroup
- E.g., in bottom distribution, features predict the outcome less well (more mass near mean) and model has worst performance, but it has the lowest FPR for blacks (smallest part of the distribution is above the threshold)



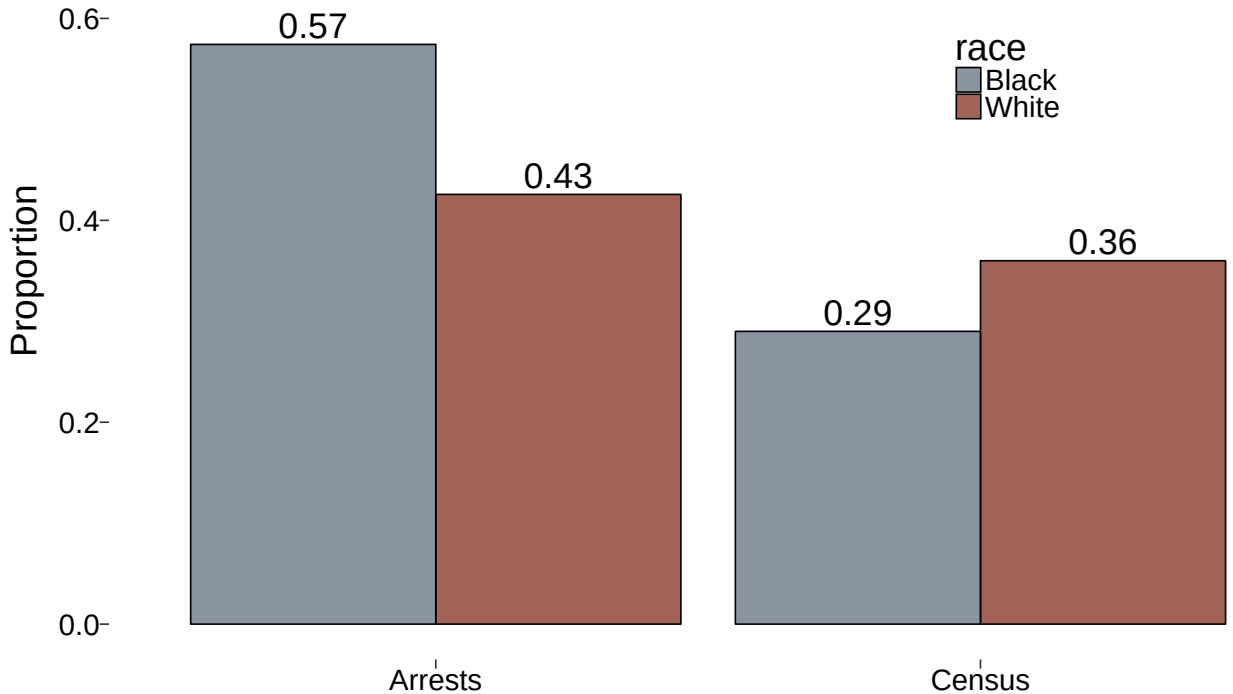
Source: https://policylab.stanford.edu/projects/defining-and-designing-fair-algorithms.html

Potentially more promising...second option: keep each group's $r(x)$ the same but apply group-specific thresholds

- Way to reduce FPR disparities (or disparities in some other metric): apply *higher threshold* for detention to group with *higher base rate/more rightward risk distribution*
- Possible ways to choose:
  - Threshold to equalize FPR
  - Threshold based on some other criteria that (likely) leads to smaller difference between $FPR_{black} - FPR_{white}$

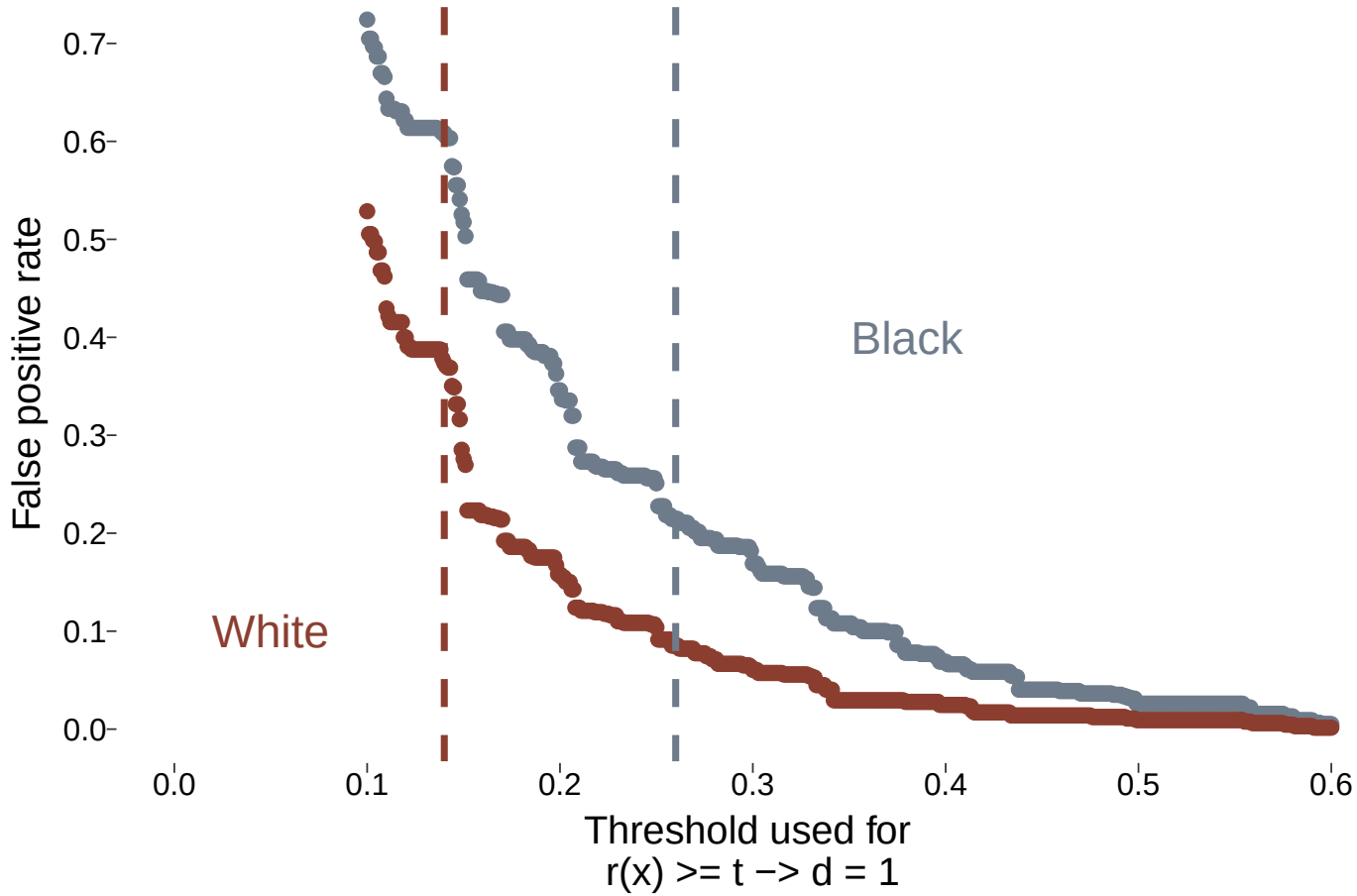
One way to choose group-specific thresholds if we're skeptical of real differences in base rate: proportionality

1. Rank all individuals together from highest to lowest $r(x)$
2. Decide on a fraction of each group to detain that's a reflection of some other proportion. E.g.:



Results of external *proportion*-driven thresholds

Example: stop detaining members of group when $\frac{\text{count detained}}{\text{all arrested}} \sim$ to that group's proportion in population; threshold is where you stop



Arguments for group-specific threshold versus same threshold

$$r(x) \geq \frac{b_{00} + c_{10} \text{ (preventing unjust detention)}}{b_{00} + c_{10} + b_{11} + c_{01} \text{ (preventing crime)}}$$

- *Allow thresholds to vary by group*:
  - *Errors are not equally costly across all individuals*: assign higher costs to errors in a group
  - *Biased labels*: justification of using the same threshold is based on assumption that observed differences in base rate reflect true differences in base rate
  - For assistive interventions, limited resources and prioritize certain groups
- *Keep thresholds the same*:
  - Allow distribution of interventions to reflect observed distribution of risk

Other challenges with classification parity and potential solutions (CR)

- Can't optimize for equality on all metrics
- Fairness between groups versus fairness between individuals (CR):
  - Classification parity focuses on between-group averages: e.g.:
    - Black v. white
    - Male v. female
    - Poverty v. not
  - We might think it's fairer to do more granular comparisons (e.g., Black male in poverty versus white male in poverty)
  - This intuition suggests group-level comparisons are rough proxy for actual goal of *treating similar individuals similarly*

Calibration

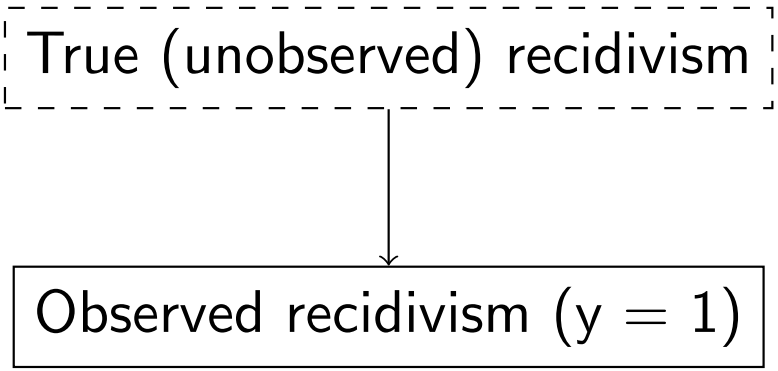
- In math:

$$Pr(Y = recid|s(x), race) = Pr(Y = recid|s(x))$$

- Similar to classification parity– two individuals with the same risk score should have the same observed rates of recidivism

Returning to earlier assumptions, and adding other concerns...

- Label (recidivism) captures construct of interest (commission of repeated crime) equally well for both groups



- Have enough of each subgroup in the training data to accurately learn relationships between features and label; likewise, have enough in test data to compare predictive accuracy by subgroup
  - *Solution*: more attention to generalizability of models fit/evaluated on a specific sample (e.g., convenience sample of facial images) when we use model to generate predictions for larger population
- Model interpretability (Kleinberg and Mullainathan, 2017)

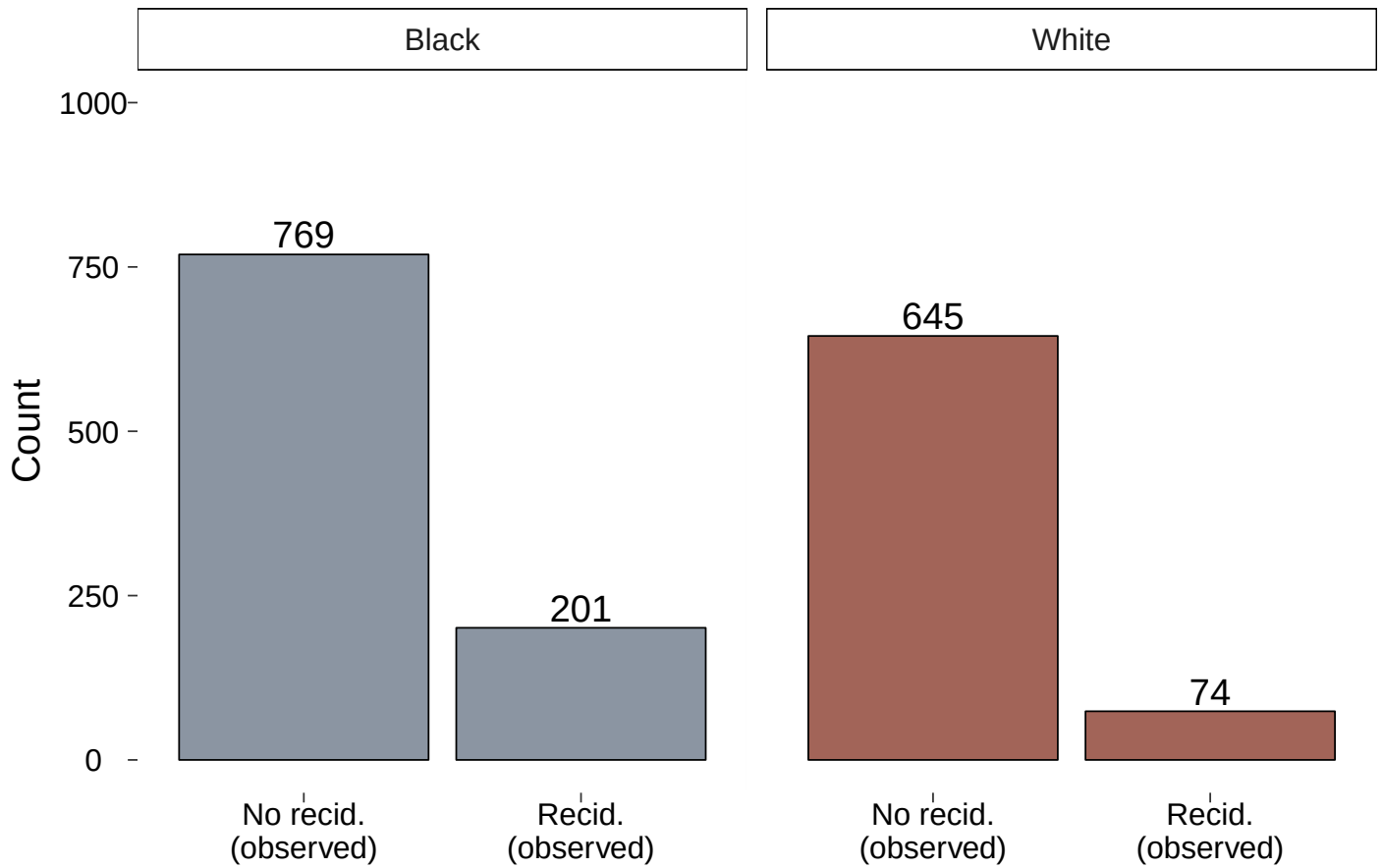
Focus on label bias, which can become feature bias in temporal contexts when y_{t-1} may be a feature that is highly predictive of y_t

First form of label bias: observe labels for everyone, but labels are imperfect proxies for true construct of interest

Person	Observed recid.	Actual recid.	Race	Corrupt alderman
	Y	**Y***	X_{obs}	X_{unobs}
1	1	0	B	0
2	1	1	B	0
3	1	1	W	0
4	0	1	B	1
5	0	1	W	0
$\vdots$				
n	0	0	B	0

First form of label bias: attempts to address

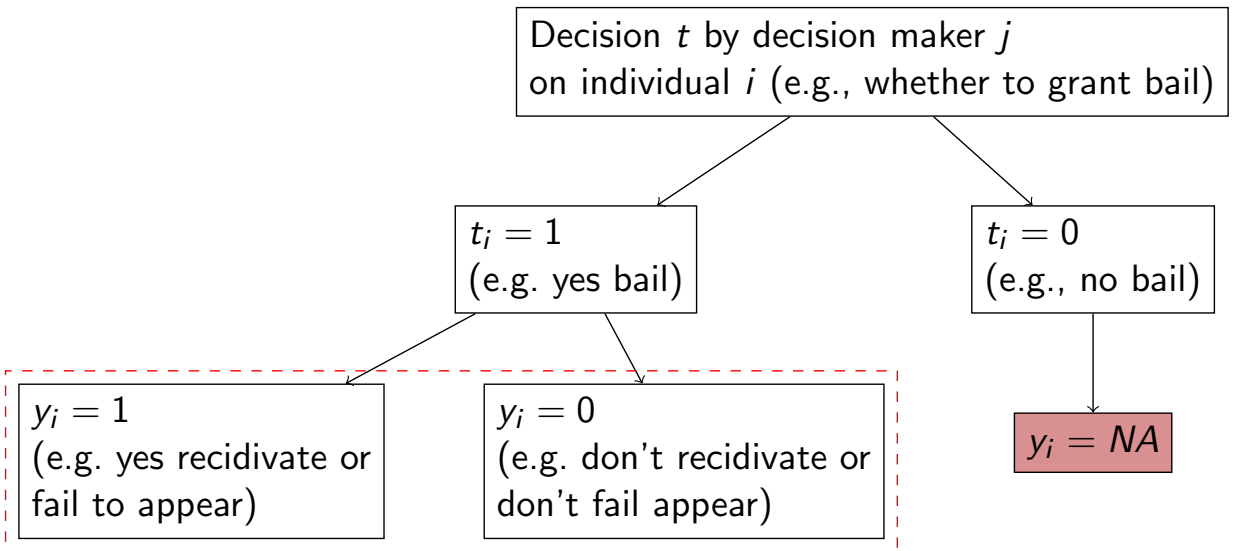
- Assume equal distribution of label across subgroups (so race is not informative for learning the risk of an individual) (CR discuss; Johndrow and Lum, 2017)– want to equalize the following:



First form of label bias: attempts to address

- Discussed earlier why removing race from the model may lead to *higher estimated risk* among a racial subgroup than leaving risk in
- Instead, start with goal of $Pr(y|race) = Pr(y)$
- Then, transform other X_u covariates (roughly, by finding their empirical distribution of race and altering in a way that minimizes the distance between X_u (raw covariate) and $\tilde{X}_u$ (statistically independent from race covariate))

Second (related) form of label bias: absence or presence of a label depends on earlier decision (and that earlier decision is at least partly a function of unobserved characteristics)



Problem of causal inference/counterfactual y: for $t_i = 0$, don't know whether $y_i = 1$ or $y_i = 0$ if that individual had the label-generating decision of $t_i = 1$

Missing labels are a problem because to evaluate our model, we need to compare $\hat{r}(x)$ to true y

Purposes of evaluation: to choose best-performing model; to assess whether it's performing better than human decisions; to compare performance by subgroup; etc.

Person	**Released**	**Observed recid.**	$\hat{r}(x)$
	T	**Y**	
1	1	1	0.8
2	0	*NA*	0.7
3	0	*NA*	0.5
4	1	0	0.3
5	1	0	0.2
⋮			
n	0	*NA*	0.1

Way one: assume individuals are missing labels completely at random and listwise deletion from test set

Assumption: $Pr(t = 1|X_{obs}, X_{unobs}) = Pr(t = 1)$ (e.g., likelihood of judge granting bail does not depend on observed or unobserved characteristics of the individual)

Person	**Released**	**Observed recid.**	$\hat{r}(x)$
	T	**Y**	
1	1	1	0.8
~~2~~	~~0~~	~~*NA*~~	~~0.7~~
~~3~~	~~0~~	~~*NA*~~	~~0.5~~
4	1	0	0.3
5	1	0	0.2
⋮			
~~*n*~~	~~0~~	~~*NA*~~	~~0.1~~

Way two: assume individuals are missing labels at random, perform some form of label imputation, and retain those individuals for test set evaluation

Assumption: $Pr(t = 1|X_{obs}, X_{unobs}) = Pr(t = 1|X_{obs})$ (e.g., likelihood of judge granting bail depends on observed characteristics of the individual but not unobserved characteristics)

Person	**Released**	**Observed recid.**	$\hat{r}(x)$	X_{obs}	**Imputed recid.**
	T	**Y**			**Y**
1	1	1	0.8	*White*, *F*, *HS*	
2	0	*NA*	0.7	*Black*, *M*, *HS*	0
3	0	*NA*	0.5	*White*, *F*, *HS*	1
4	1	0	0.3	*White*, *M*, *Coll.*	
5	1	0	0.2	*Black*, *M*, *HS*	
⋮					
n	0	*NA*	0.1	*Black*, *M*, *HS*	0

Way two: label imputation to form pessimistic and optimistic bounds on evaluation metrics

Optimistic bounds: `ifelse`$(\hat{r}(x) \geq t, y == 1, 0)$

Person	Released	Recid.	$\hat{r}(x)$
	T	Y	
1	1	1	0.8
2	0	1	0.7
3	0	1	0.5
4	1	0	0.3
5	1	0	0.2
$\vdots$			
n	0	0	0.1

Pessimistic bounds: `ifelse`$(\hat{r}(x) \geq t, y == 0, 1)$

Person	Released	Recid.	$\hat{r}(x)$
	T	Y	
1	1	1	0.8
2	0	0	0.7
3	0	0	0.5
4	1	0	0.3
5	1	0	0.2
$\vdots$			
n	0	1	0.1

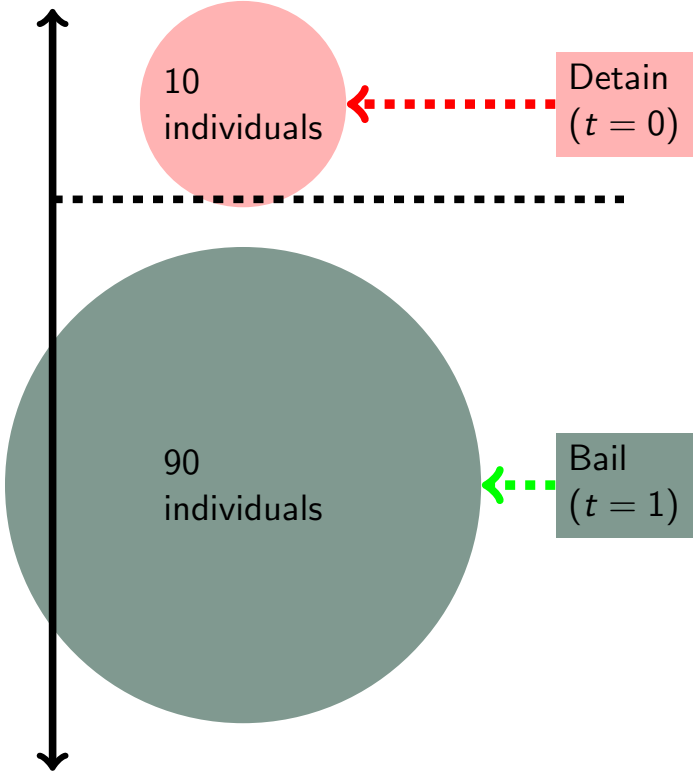
Way three: find some source of random variation in t (Lakkaraju et al., 2017)

Their contraction technique relies on features of the bail setting/assumptions that (may) generalize to other contexts:

1. Multiple j assigning t (and we know which j was paired with which i): rather than a single decision-maker, there are multiple decision-makers who make the label-generating decision (e.g., multiple judges who form the overall pool of bail decisions and we know which defendant saw which judge)
2. Random assignment of i to j: e.g., it's not that certain judges assess the cases of defendants with unobserved features that make them a worse candidate for bail
3. Variation across j in $\frac{t=1}{t}$
4. The variation stems from differences in leniency rather than differences in the ability to correctly assess unobservables correlated with the outcome

Step one in contraction: find decision-maker with highest fraction of observed labels

Step one: start with decision-maker who has the highest acceptance rate (highest proportion of defendants they saw who they gave a "yes" decision to that then generates a label)– call him/her q; better if q evaluates higher N



Step two: use all individuals to predict decision to deny bail (t = 0)

id	t (bail)	y (recidivate)	$\hat{t}$
1	0	NA	0.8
2	0	NA	0.7
3	0	NA	0.9
4	0	NA	0.2
5	0	NA	0.3
⋮			
10	0	NA	0.4
11	1	0	0.6
12	1	1	0.5
13	1	1	0.6
14	1	0	0.8
⋮			
100	1	1	0.7

Step three: detain the same 10 as the judge (regardless of their $\hat{t}$ and rank remaining 90 in order of predicted prob. of detention

rank	$\hat{t}(x)$	y (recid/failure)	bail
1	0.89	1	M + Q det.
2	0.72	1	M + Q det.
3	0.91	1	M + Q det.
4	0.22	0	M + Q det.
⋮			
10	0.46	0	M + Q det.
11	0.8	0	M det.
12	0.79	1	M det.
13	0.75	1	M det.
14	0.73	1	M det.
⋮			
30	0.65	1	M det.
31	0.58	0	M rel.
32	0.48	1	M rel.
33	0.46	0	M rel.
⋮			
100	0.10	0	M rel.

- Quantities: $\mathcal{D}_q = 100$ individuals; $\mathcal{R}_q = 90$ individuals; $\mathcal{R}_\mathcal{B} = 70$ individuals
- Model's failure rate at threshold r if $y = 1 = recid/fail$ (*Note*: y =1 and y = 0 reversed from their paper)

$$\sum_{i=1}^{70} \frac{\mathbf{1}(y_i = 1)}{100}$$

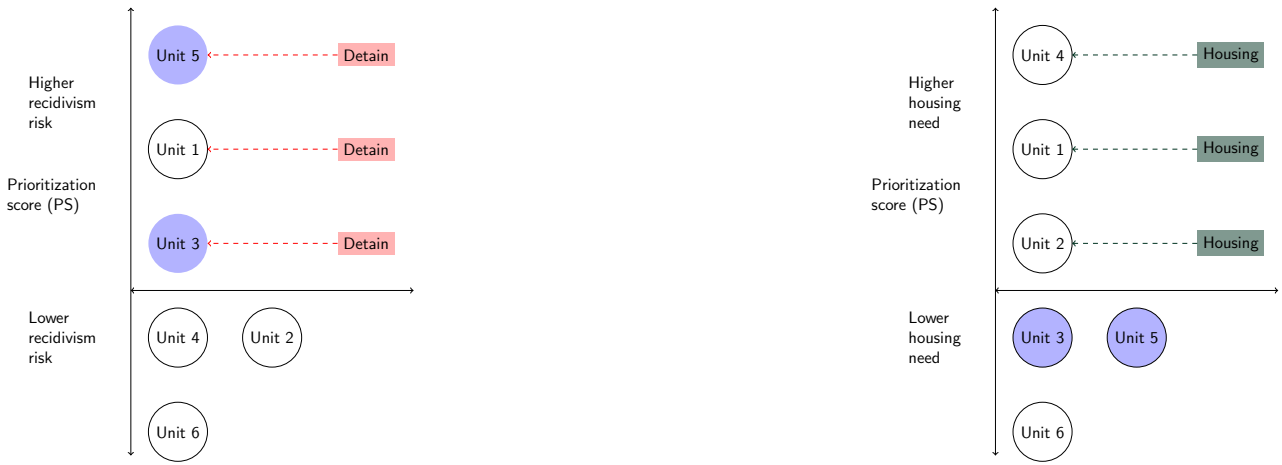
- Human failure rate at threshold r: bin humans and look at observed rate of y = 1 relative to all considered

Comments/points of confusion

- Which judges' observations are used to model t as you lower the acceptance threshold? Two options:
  - Continue to use the most lenient decision-maker
  - Pool all decision-makers with leniency $>$ acceptance threshold
    - If this, a bit weird because as they show in simulations, it's better to have higher N to model t; if you change threshold and change N, two moving pieces
- Generalizability to other cases:
  - Difference between features important for t(x)(label-generating decision) and features important for y(x)(outcome conditional on having a label)
  - Presence of identifiers in decision data

Concluding

- Much of what we've reviewed focuses on quantifying what counts as over or under-allocation of a resource on the basis of model predictions (e.g., criticisms of defining over-allocation with respect to higher FPR in a disadvantaged group when translating estimated risk into a decision)



- This maps onto definitions of fairness that look at the *outcomes* of an allocation
- Some attention in bail paper and Eubanks (2018), but perhaps role for sociologists is to think more about fair *processes* of allocations and how score-informed allocation processes differ from existing processes in organizations

Less interpretable weights on inputs means a lower potential for organizations to deliberately (and unequally) manipulate those inputs?

- *Example from my own research*: weights for presence/absence of group membership; summed within a school

Weighted Student Funding (WSF) allocates a pool of resources to schools by assigning weights to all students based on need

Base weight: What all students receive (differentiated by grade)

Need weights: Additional support to meet needs of certain populations. Applied to all eligible students. Examples include:

- Programs (e.g., ELL / LEP, SPED, Vocational)
- Student characteristics (e.g., poverty)
- Academic performance (e.g., off track)

Weights equal a proportional dollar amount, based on total pool of resources

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

Less interpretable weights on inputs means a lower potential for organizations to deliberately (and unequally) manipulate those inputs?

- Chosen by a committee and framed as more interpretable and hence, more fair

> - 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")

- Same transparency allows organizations to manipulate inputs; would be more difficult to manipulate inputs to allocation based on district-level aggregation of predicted risk of some bad outcome (e.g., dropout; not progressing to college)