Correcting systemic biases for fairer and more replicable data science

Joshua Loftus

February 19th, 2020

1. Fairness and causal inference
2. Selection bias and the replication crisis
3. Conclusions and future directions

Fairness and causal inference

Joint work with

Matt Kusner

UCL

Chris Russell

Surrey

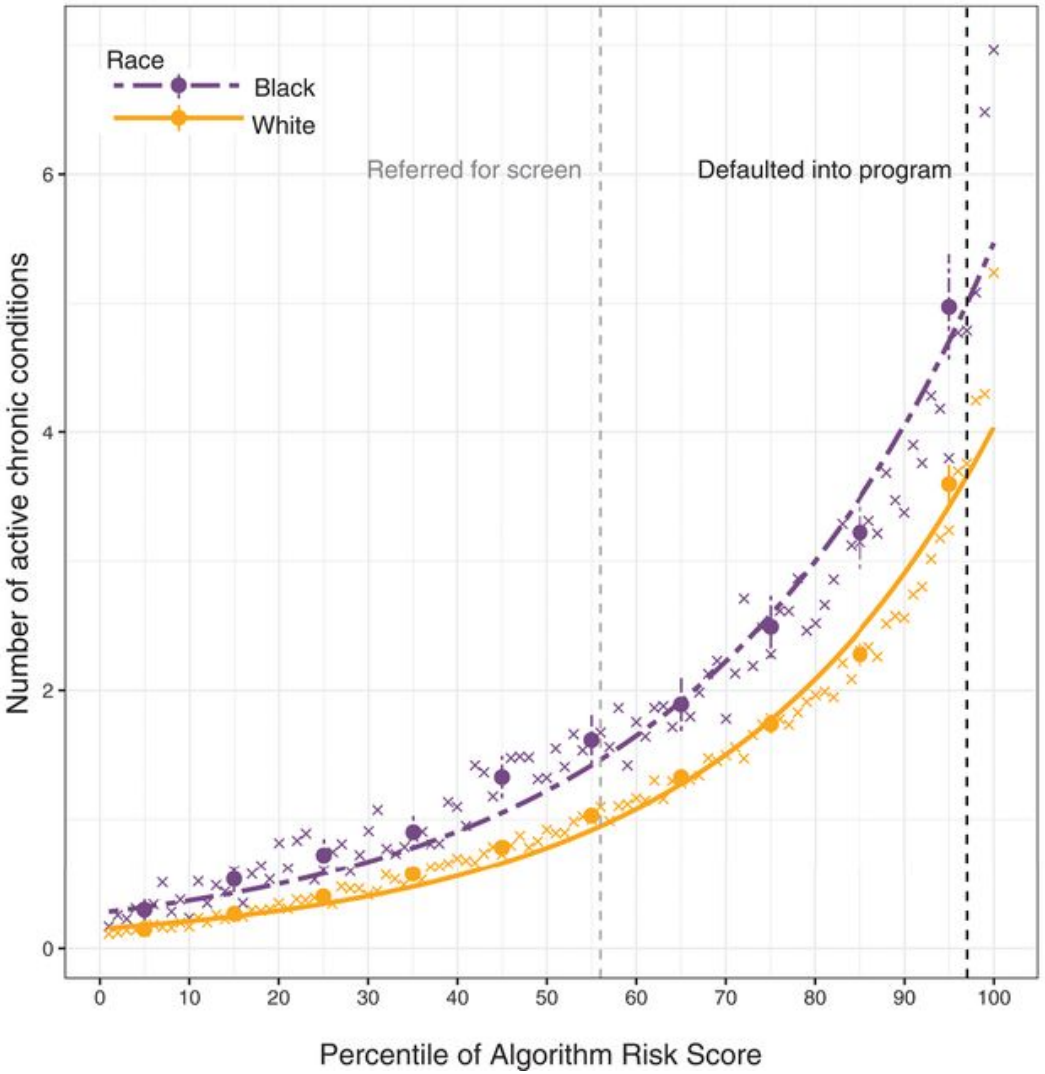
Ricardo Silva

UCL

The Alan Turing Institute

And my current student @ NYU Stern, Margarita Boyarskaya

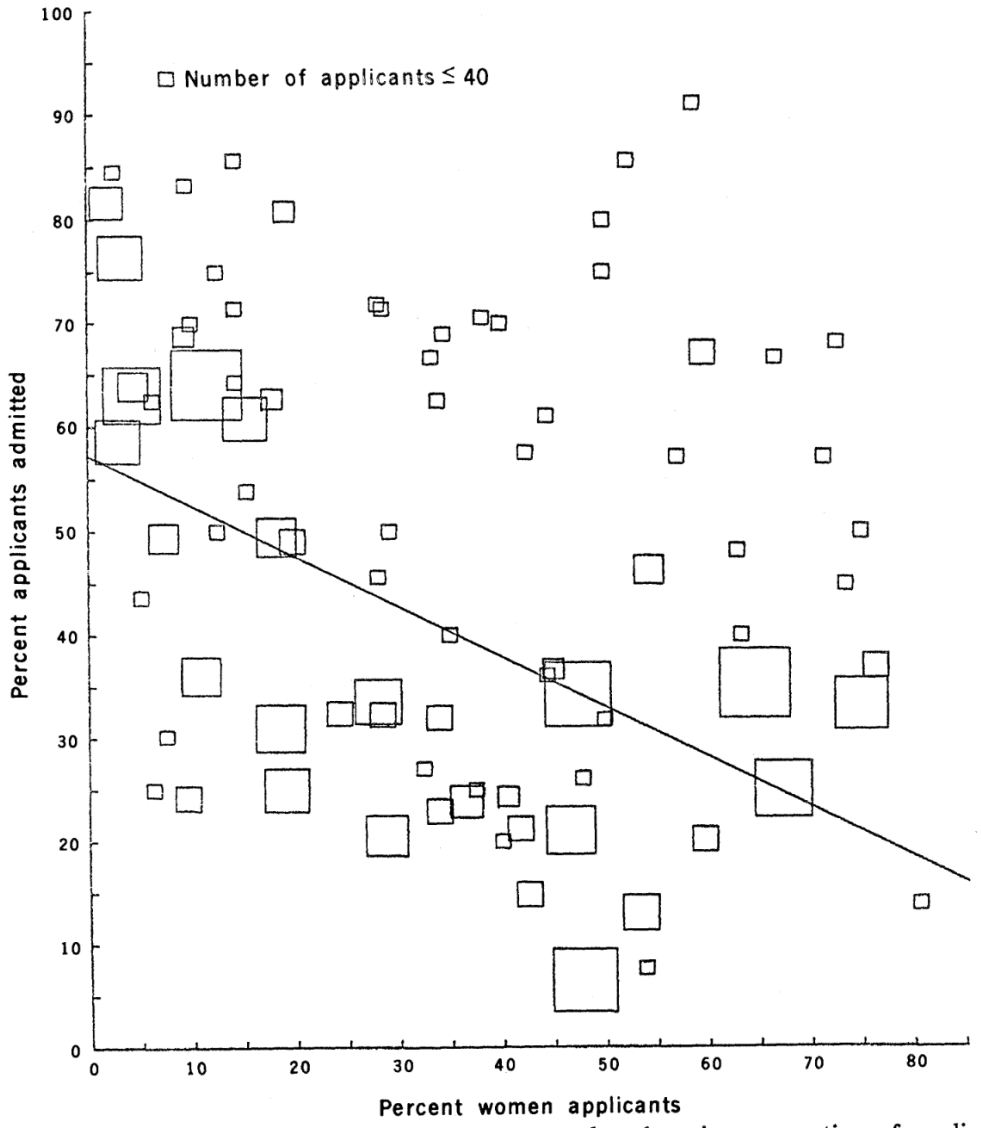
Risk scores and healthcare patient data



Obermeyer et al (2019)

- Algorithm assigns risk scores by predicting healthcare **costs** from patient records
- Underestimates risk of health conditions for black patients compared to white patients
- Adjusting algorithm to close the gap results in 2.5x black patients receiving more care

Statistics and discrimination



Sex Bias in Graduate Admissions: Data from Berkeley (P. J. Bickel, E. A. Hammel, J. W. O'Connell, 1975)

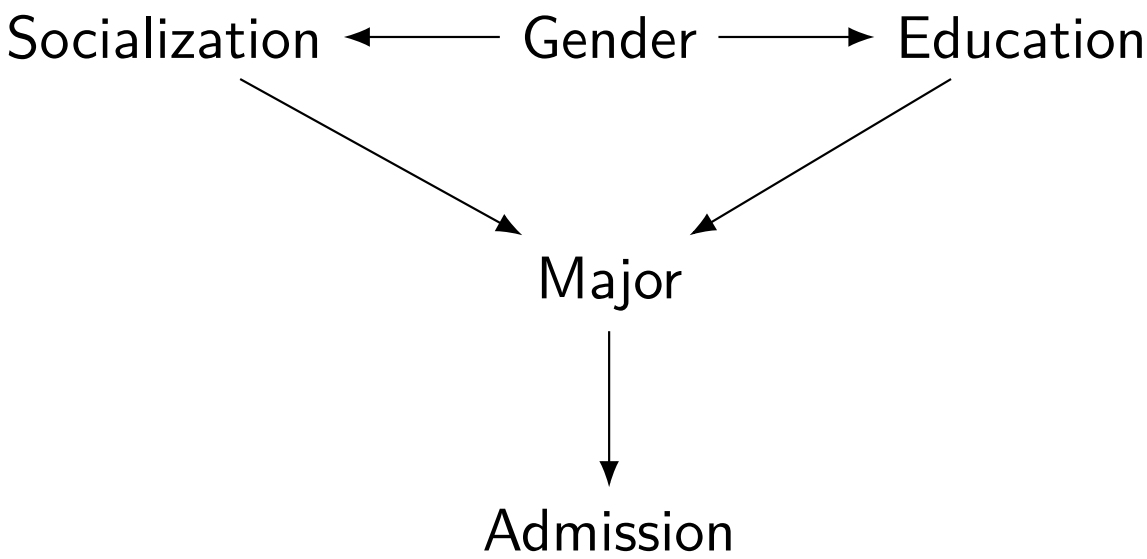
Classical example of "Simpson's paradox"

Focus on causes

The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seem quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

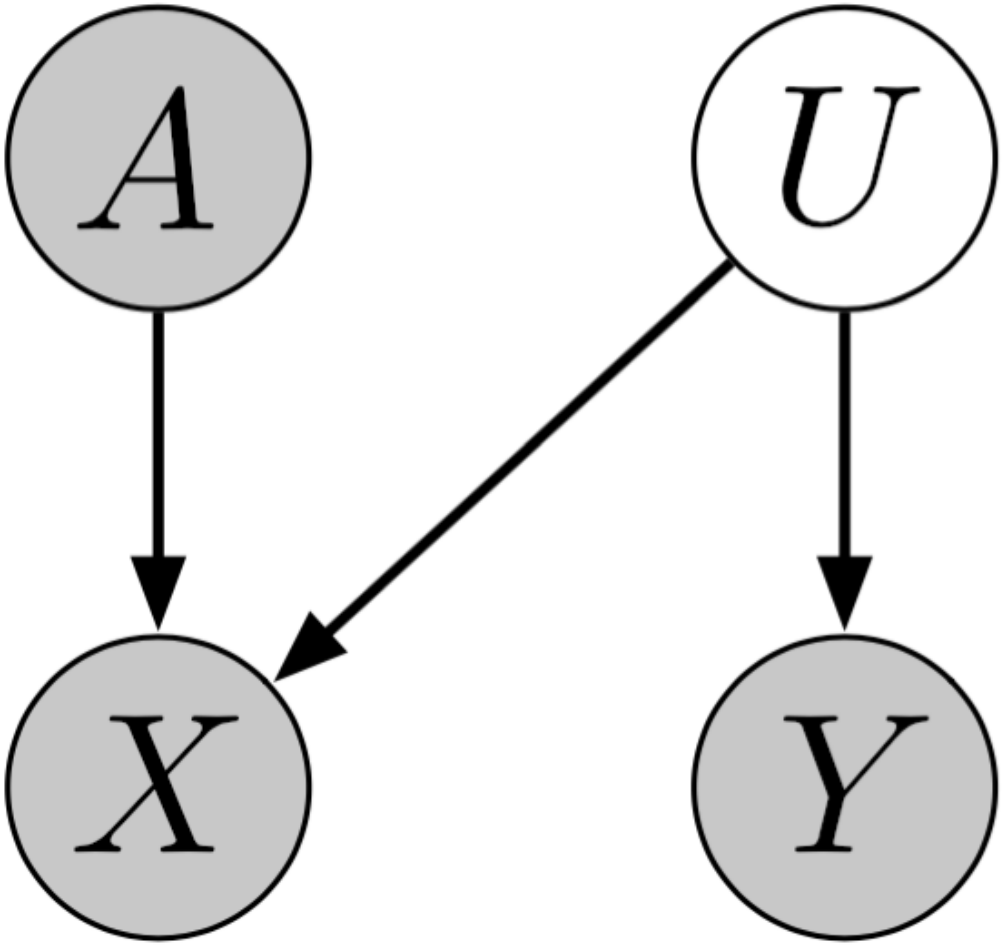
- from the final paragraph of Bickel et al (1975)

Causal diagram



Imagine the counterfactual: when Joshua showed enjoyment of math as a student, what if his teachers, family, etc, reacted to that by encouraging him toward a career in nursing?

DAG/SEM causal model framework



- Nodes: variables (U unobserved, Y outcome, A sensitive, X other predictors)
- Arrows: causal relationships / conditional (in)dependence
- Structural equations: functional forms of (arrow) relationships

A causal definition of fairness

- Compute model-based counterfactual values for variables downstream of **A**
- Define fairness using a causal model
- Train algorithms to satisfy fairness constraints

Counterfactual fairness

An estimator $\widehat{\mathbf{Y}}$ is **counterfactually fair** if

$$\mathbb{P}(\widehat{\mathbf{Y}}_{a}|\mathbf{X} = x, \mathbf{A} = a) = \mathbb{P}(\widehat{\mathbf{Y}}_{a'}|\mathbf{X} = x, \mathbf{A} = a)$$

for all x, a, a'.

My work in this area

Counterfactual fairness

M. J. Kusner, J. R. Loftus, C. Russell, R. Silva. *Counterfactual fairness*. NeurIPS, 2017.
C. Russell, M. J. Kusner, J. R. Loftus, R. Silva. *When worlds collide: integrating different counterfactual assumptions in fairness*. NeurIPS, 2017.

Fair interventions / causal interference

M. J. Kusner, C. Russell, J. R. Loftus, R. Silva. *Making Decisions that Reduce Discriminatory Impact*. ICML, 2019.

Ongoing / future work

Unfair sample selection bias (with PhD student), interpretable models/counterfactuals, relation to other notions (justification, responsibility, agency)

Selection bias and the replication crisis

Based on work with my coauthors

Jonathan Taylor

Stats@Stanford

Rob Tibshirani

Stats@Stanford

Ryan Tibshirani

Stats@CMU

Xiaoying Tian

Farallon

And my current student @ NYU Stern, Weichi Yao

Motivating example: regression coefficients

- 19th century artificial intelligence: Legendre, Gauss, Quetelet
- 1908-1922 Gosset and Fisher discover and prove probability distribution result for coefficients, enabling statistical hypothesis tests and intervals
- Computer age (Efron and Hastie, 2016) - larger datasets, more variables, exponentially many possible models
- Use algorithms to try many models and pick the best one

The candy ranking dataset

```
candy <- fivethirtyeight::candy_rankings
head(candy[, c(1:2, 11:13)])
```

```
## # A tibble: 6 x 5
##   competitorname chocolate sugarpercent pricepercent winpercent
##   <chr>          <lgl>             <dbl>        <dbl>      <dbl>
## 1 100 Grand      TRUE              0.732        0.860       67.0
## 2 3 Musketeers   TRUE              0.604        0.511       67.6
## 3 One dime       FALSE             0.011        0.116       32.3
## 4 One quarter    FALSE             0.011        0.511       46.1
## 5 Air Heads      FALSE             0.906        0.511       52.3
## 6 Almond Joy     TRUE              0.465        0.767       50.3
```

Model selection: forward stepwise with AIC

```
# Forward stepwise with AIC
model <- step(lm(winpercent ~ . - competitorname, candy),
              k = 2, trace = 0)
# Significance tests for selected model
print(summary(model)$coefficients, digits = 2)
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)           76        6.3    12.1  9.6e-20
## chocolateTRUE        -18        7.7    -2.3  2.2e-02
## hardTRUE             -16        8.1    -2.0  5.3e-02
## barTRUE               12        8.8     1.4  1.7e-01
## pricepercent         -27       12.4    -2.2  3.0e-02
```

What's wrong with this?

- Not what Gosset and Fisher considered...
- Tests, without novel data or "bold" hypotheses?

Look behind the curtain:

```
candy$winpercent <- 100 * runif(nrow(candy))
```

(If you know `rmd`, I hid this line using the option `include=FALSE` in an earlier code chunk)

One example of a problem with major implications

Replication crisis in science

> *We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. [. . .] Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size;* ***39% of effects were subjectively rated to have replicated the original result***

From: *Estimating the reproducibility of psychological science* (Open Science Collaboration, 2015).

See also: *Why most published research findings are false* (Ioannidis, 2005).

One possible solution: selective (conditional) inference

- $M = M(y)$ model selection (e.g. forward stepwise)
- Test $H_0(m)$ with statistic T conditioned on $\{y : M(y) = m\}$

Selective type 1 error (Fithian, Sun, Taylor, 2014)

Reject $H_0(m)$ if $T > c_\alpha(m)$, where

$$\mathbb{P}_{H_0(m)}(T > c_\alpha(m) \mid M = m) \leq \alpha$$

- Reduces to usual type 1 error in several cases

Marginal screening example

Consider independent effects under a global null hypothesis

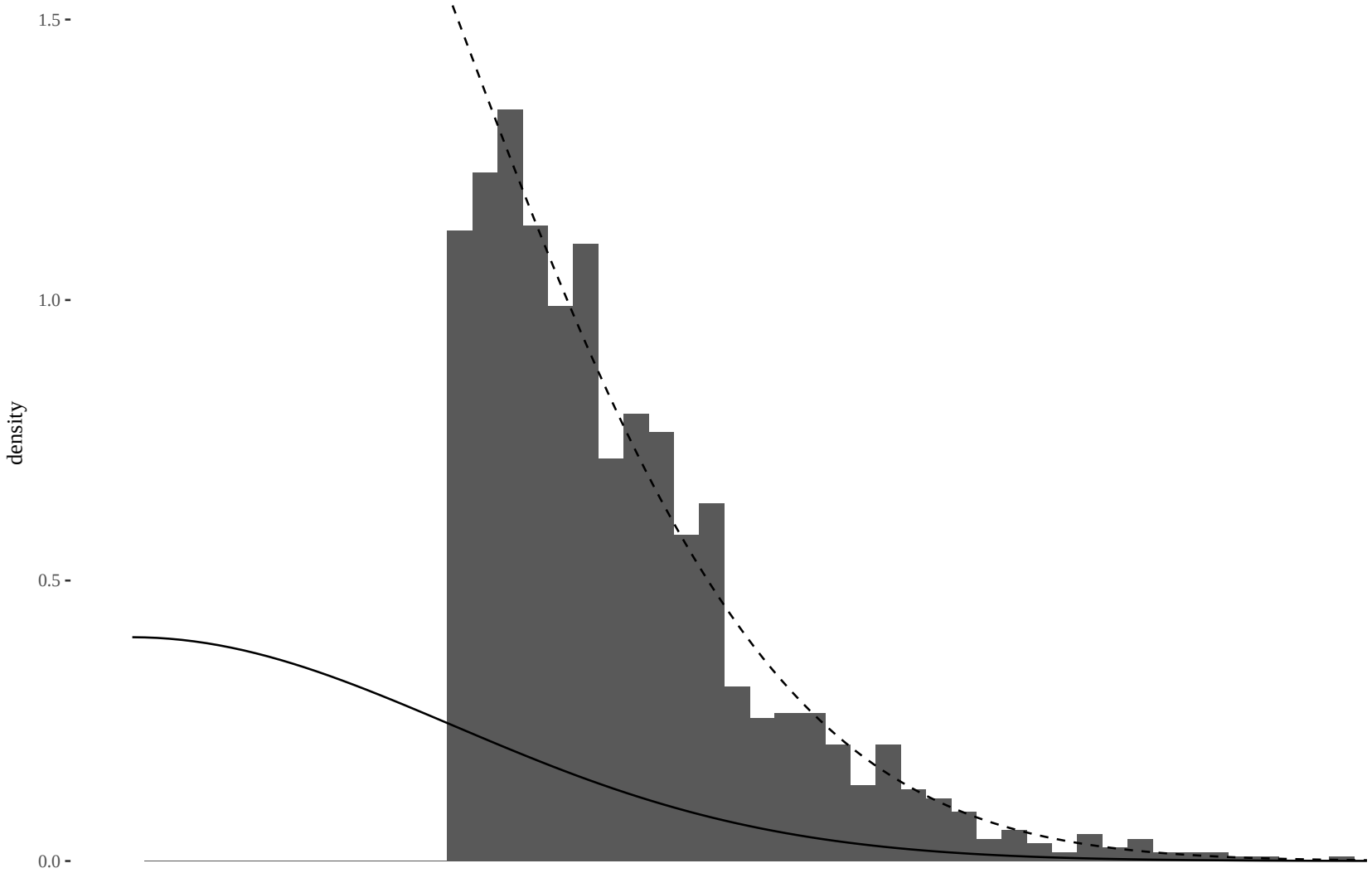
$$Z_i \sim N(\mu_i, 1) \quad \text{under} \quad \bigcap_i \{H_{0,i} : \mu_i = 0\}$$

```
Z <- rnorm(10000)
selected_Z <- data.frame(Z = Z[Z > 1])
```

How can we test significance **of the selected** Z_i?

Conditional null distribution is truncated: $Z|Z > 1$

```
truncated_Z_pdf <- function(z) dnorm(z)/pnorm(1, lower.tail = F)
# plot code hidden
```



Interpretation

- Keep model selection and inference compartmentalized
- The procedure/algorithm selects model m, this is a sort of evidence in favor of the model, but not a hypothesis test
- If we want hypothesis tests as additional evidence, we must do something to correct the selection bias
- Data must "surprise" us twice: once to select a model, and then again to pass test(s)

Growing literature (including e.g. details of truncation regions) for various selection methods in various settings

My work in this area

Kac-Rice test

Taylor, Loftus, and Tibshirani. *Inference in adaptive regression via the Kac-Rice formula*. Annals of Statistics, 2016

Dissertation work

Quadratic selection events: groups of variables, cross-validation, F-tests. `selectiveInference` R package.

Square-root lasso

Selection with unknown variance. Tian, Loftus, Taylor. *Selective inference with unknown variance via the square-root LASSO*. Biometrika, 2018

Ongoing / future work

Goodness of fit (with PhD student), follow-up studies, other diagnostics, interpretation/application

Conclusions and future directions

Interpretability

- Relative strength of statistics within data science
- Simpler models, orientation toward hypotheses about the real world, causation
- Fairness is closely connected to interpretability (equal treatment?)
- Interpretation and replicability: which hypotheses/generalizations are actually being tested?

Thank you!

Reading for fairly general audiences, with references for additional details

- Kusner and Loftus: *The long road to fairer algorithms*. Nature, 2020
- Taylor and Tibshirani: *Statistical learning and selective inference*. PNAS, 2015

joshualoftus.com

Affine selection events

For many commonly used regression model selection methods (choosing a subset of columns m so that $E[y] \approx X_m \beta_m$), the selection event $M(y) = m$ can be characterized via affine inequalities

Model selection events as quadratic events

$$\{M(y) = m\} = \{A_m y \leq b_m\}$$

Selective (conditional) inference: truncate multivariate normal to this subset of $\mathbb{R}^n$

Examples: forward stepwise, Lasso with fixed (not data-dependent) λ, etc.

Quadratic selection events

Wider class of selection methods, including e.g. stepwise with groups of variables (Loftus & Taylor 2015), lasso with cross-validation (Loftus 2015), ...

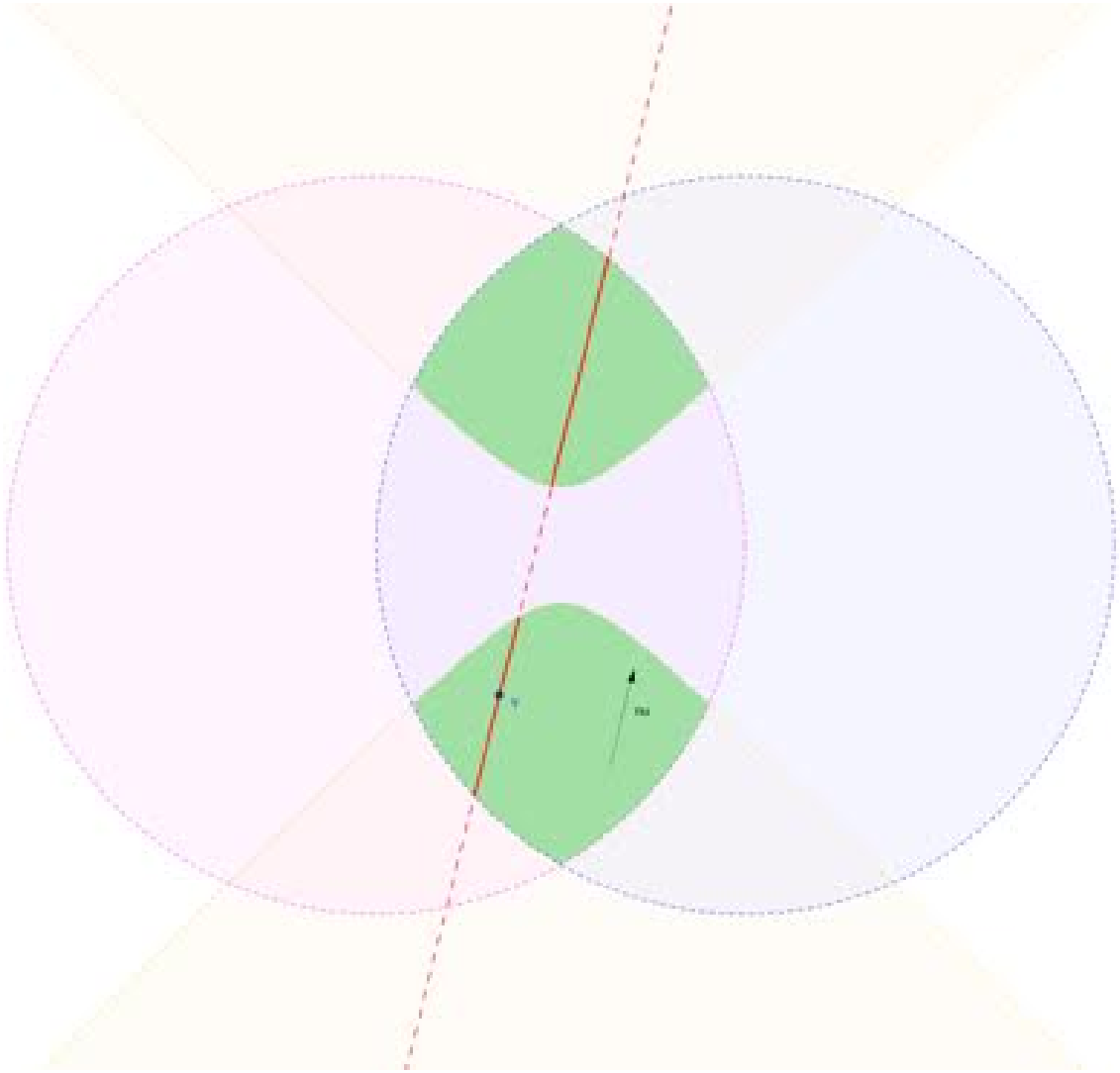
Model selection events as quadratic events

$$\{M(y) = m\} = \left\{ \bigcap_j y^T Q_{m,j} y + A_{m,j} y + b_{m,j} \geq 0 \right\}$$

Intersection of quadratic regions... can be very complicated!

Dimension reduction: condition on everything but the test statistic to reduce to one dimension

Cartoon example with $n = 2$



May be computationally expensive, loss of power from over-conditioning

Examples: groups of variables, cross-validation

- Group selection problems (categorical, structure, etc)
- Cross-validation: let f, g index CV test folds. On fold f, model $m_{f,s}$ at sparsity s, and $-f$ denoting the training set (complement of f).

Blockwise quadratic form of cv-RSS (Loftus 2015)

Define $P_{f,s} := X^f_{m_f,s}(X^{-f}_{m_f,s})^\dagger$, $Q^s_{ff} := \sum_{g \neq f}(P_{g,s})^T_f (P_{g,s})_f$ and

$$Q^s_{fg} := -(P_{f,s})_g - (P_{g,s})^T_f + \sum_{\substack{h=1 \\ h \notin \{f,g\}}}^{K} (P_{h,s})^T_f (P_{h,s})^T_g$$

Then with y_K denoting the observations ordered by CV-folds,

$$\text{cv-RSS}(s) = y_K^T Q^s y_K$$

Selective unbiasedness

Goodness-of-fit tests can be (conditionally) worse at detecting the wrong model has been selected than just tossing an α-coin

Selective unbiasedness

We say that a test is selectively unbiased if for any selected model m and alternative hypothesis $H_1(m)$,

$$\mathbb{P}_{H_1(m)}(\text{reject } H_0 | M = m) \geq \alpha$$

As with selective type 1 error, we achieve this by using the conditional (truncated) distribution of the test statistic.

Regression example: F-tests (of unselected variables)

- Regression models $E[Y] = X_A \beta_A$ for some subset A of columns of a matrix X.
- With nested subsets $A \subsetneq A'$, we'll conduct an F-test and consider this as a goodness-of-fit test for the model with variables A.
- In `R` we just use the `anova` function with these two linear models.

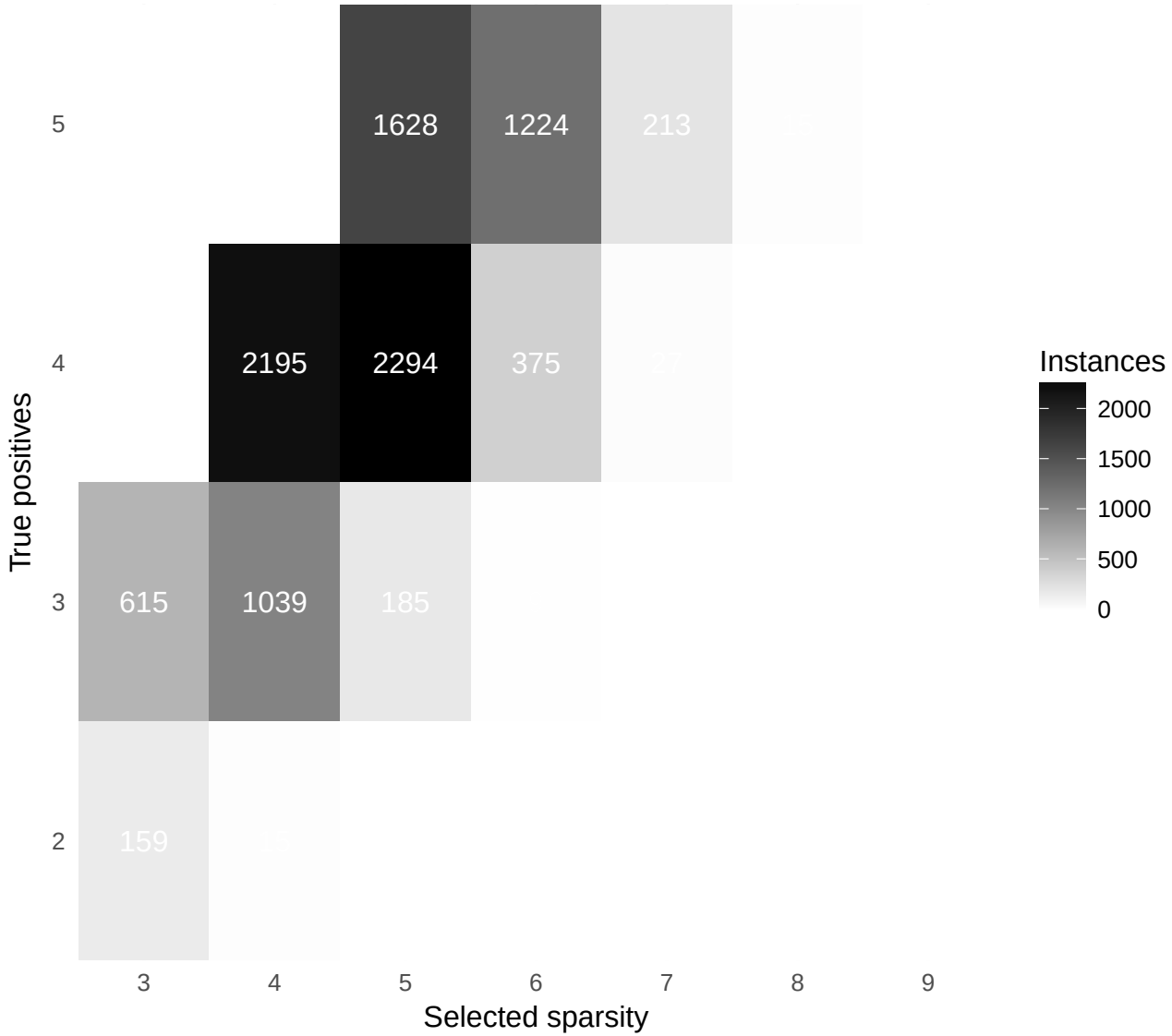
The distribution of the F-statistic is derived, of course, under the assumption that A and A' have been chosen *a priori*...

(We briefly mentioned an idea like this in Tian et al. (2018))

Regression variable selection

- For concreteness, consider selecting variables using forward stepwise with BIC, i.e. in `R` with `step(..., k = log(n))`.
- Simulation with $n = 100$ observations of $p = 10$ variables, the first two coefficients are larger than the next 3, and the last 5 are all 0.
- In this low-dimensional example, we'll take $A' = \{1, \ldots, 10\}$ for simplicity.
- Consider the F-test as a goodness-of-fit test for the selected $A \subset A'$, and compute both unadjusted (classical) and adjusted (selective) p-values.

Profile of model selection events



Distributions of p-values for full-model F-tests

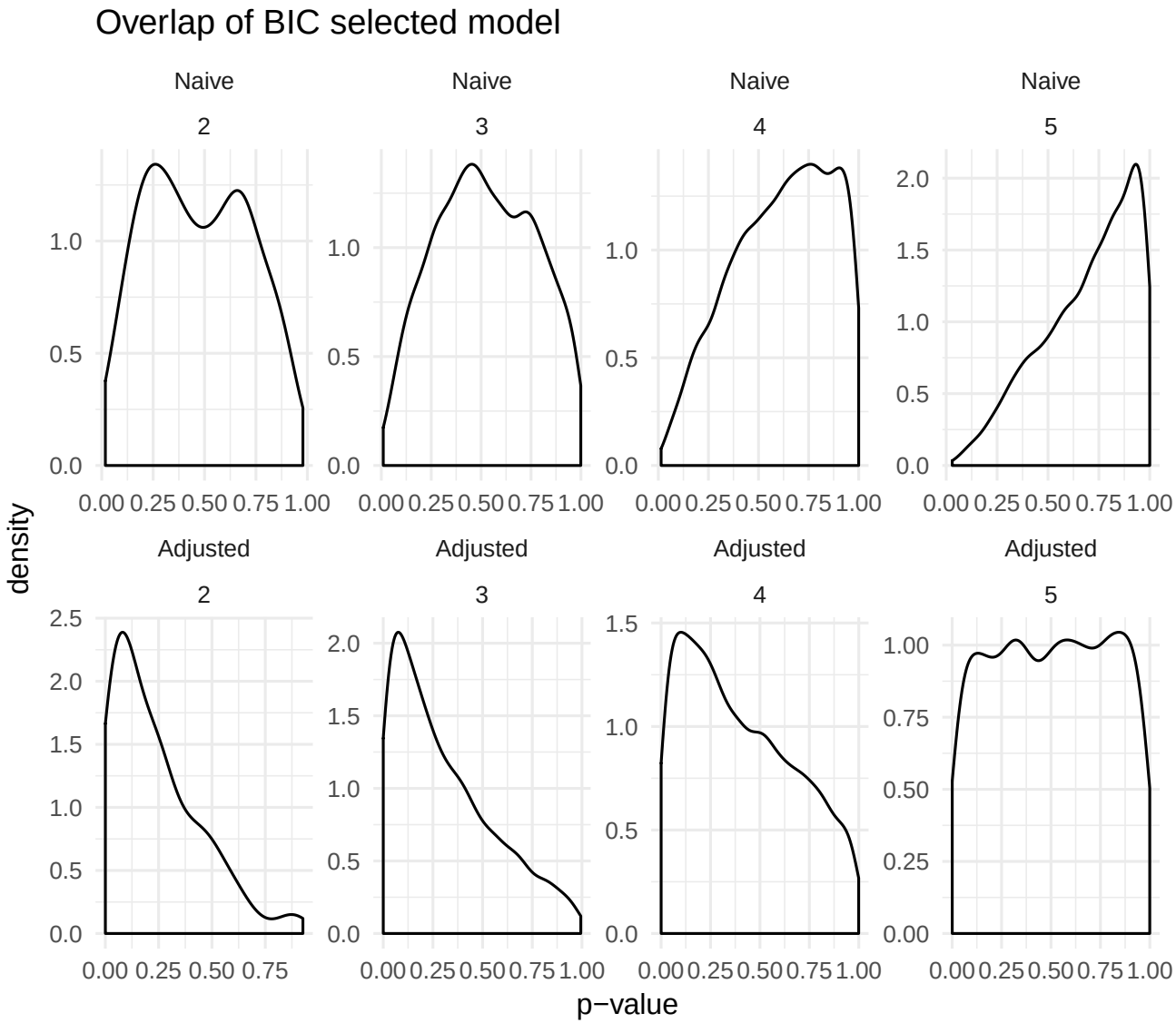


Figure 1: Top: unadjusted p-values. Bottom: adjusted for selection

Probability of rejection

Table 1: Probability of rejection at level 0.1, conditional on size of overlap

pvalue	overlap	Pr(reject)
Naive	2	0.056
Naive	3	0.032
Naive	4	0.017
Naive	5	0.005
Adjusted	2	0.328
Adjusted	3	0.251
Adjusted	4	0.156
Adjusted	5	0.101

Conditional power

Table 2: Probability of rejection at level 0.1, conditional on overlap less than 5

pvalue	Pr(reject)
Naive	0.022
Adjusted	0.186

Follow-up study after Benjamini-Hochberg

Idea: after using the BH(q) procedure to select a subset of hypotheses while controlling the false discovery rate, decide whether to conduct a follow-up study of the hypotheses that were *not* selected

Use the conditional distribution of the non-rejected $p_{(j)}$ to adjust
e.g. Fisher's combination test

Interesting meta-analysis implications

Power of Fisher's combination test

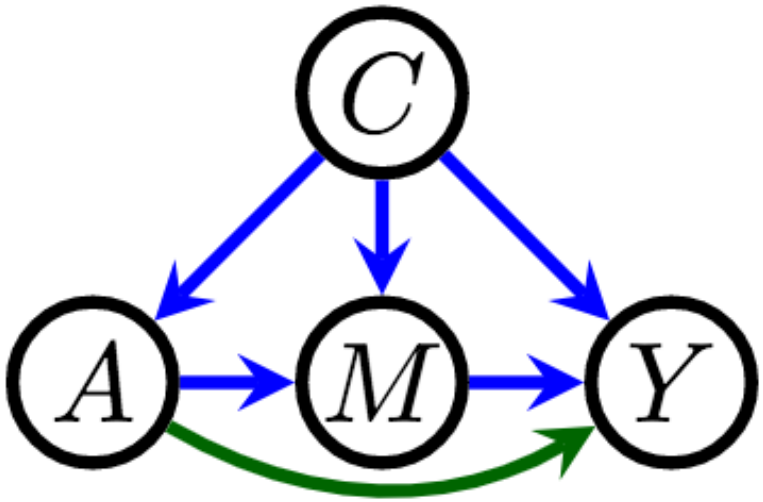
Table 3: Probability of rejection at level 0.05 with different Beta$(1, \mu)$

μ	$p_{1\|0}$	unadjusted	adjusted
10	97	1.0000	1.0000
20	87	0.9614	1.0000
30	60	0.5091	0.9838
40	33	0.0900	0.8580
50	19	0.0040	0.6080
60	11	0.0000	0.3320
70	6	0.0000	0.2004
80	4	0.0000	0.1245
90	3	0.0000	0.0995

$p_{1|0}$ denotes the average proportion of true nonnulls that are not rejected

Pathway analysis / decomposition

- Kusner et al (2017): path-dependent counterfactual fairness (supplement)
- Kilbertus et al (2017): proxies and resolving variables
- Nabi and Shpitser (2018): path specific effects, mediators, constrain parameters
- Chiappa and Gillam (2018): more flexible modeling, modify features
- Zhang and Bareinboim (2017): (counterfactual) direct, indirect, and spurious



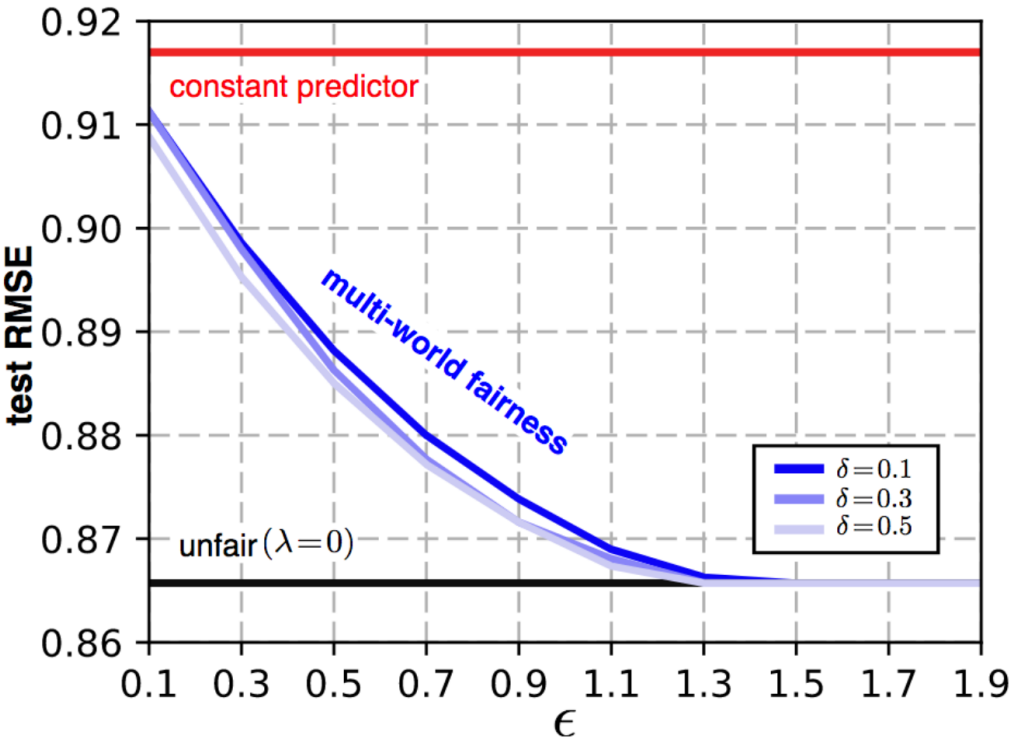
Model for crime data where the mediator can be, e.g., prior convictions

Multi-world consensus

When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness (Russell et al, *NeurIPS 2017*)

- Competing causal models
- Approximate counterfactual fairness (relax equality constraint)
- Predictions approximately satisfy fairness across both (all) models
- Limitation: the more contradictory are the competing models, the more trivial the predictions (constant)
- Causal framing of fundamental contradiction

Multi world fairness



At least $1-\delta$ probability of satisfying ϵ-relaxation of counterfactual fairness, λ is an optimization tuning parameter which weights the multi-world fairness penalty

Resolving the contradiction

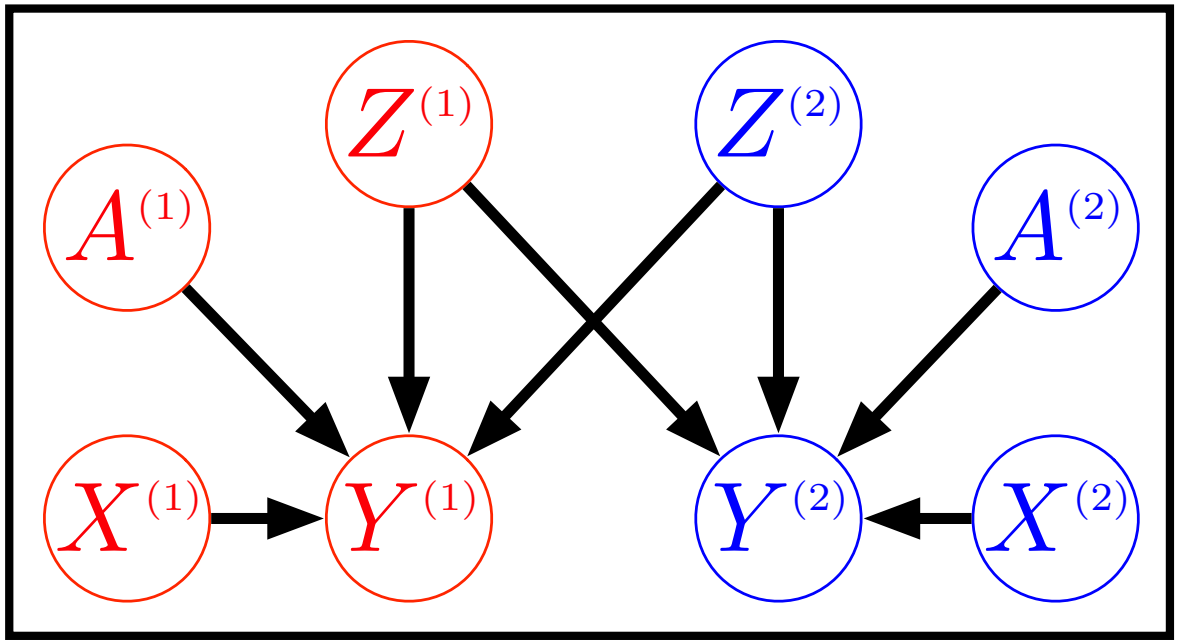
I think this is the right *path*. It's now about understanding the causes of unfairness well enough to reach consensus.

Recognizing separate design elements of predictive ML

- Model the **intervention** (that predictions will be used for) as its own separate variable **Z**
- Causal model including **Z**, explicitly model changes in the world due to use of ML
- Potentially multiple objectives to represent interests of different stakeholders

Recent work: Making Decisions that Reduce Discriminatory Impacts (Kusner et al, *ICML 2019*)

Example: fair intervention under interference



Counterfactual privilege

When designing an optimal, fair intervention $\mathbf{Z}$, instead of enforcing the equality in the definition of counterfactual fairness, we can also use an asymmetric bound on **counterfactual privilege**, for $\tau \geq 0$

$$\mathbb{E}[\widehat{\mathbf{Y}}(a, \mathbf{Z})] - \mathbb{E}[\widehat{\mathbf{Y}}(a', \mathbf{Z})] \leq \tau$$

- In practice these *asymmetric constraints* will only be active for privileged values of a (actual, left term), and inactive otherwise
- Intervention should not allocate resources in a way that helps people (in expectation) become more than τ (in terms of the outcome) units better than they would be if they were not privileged

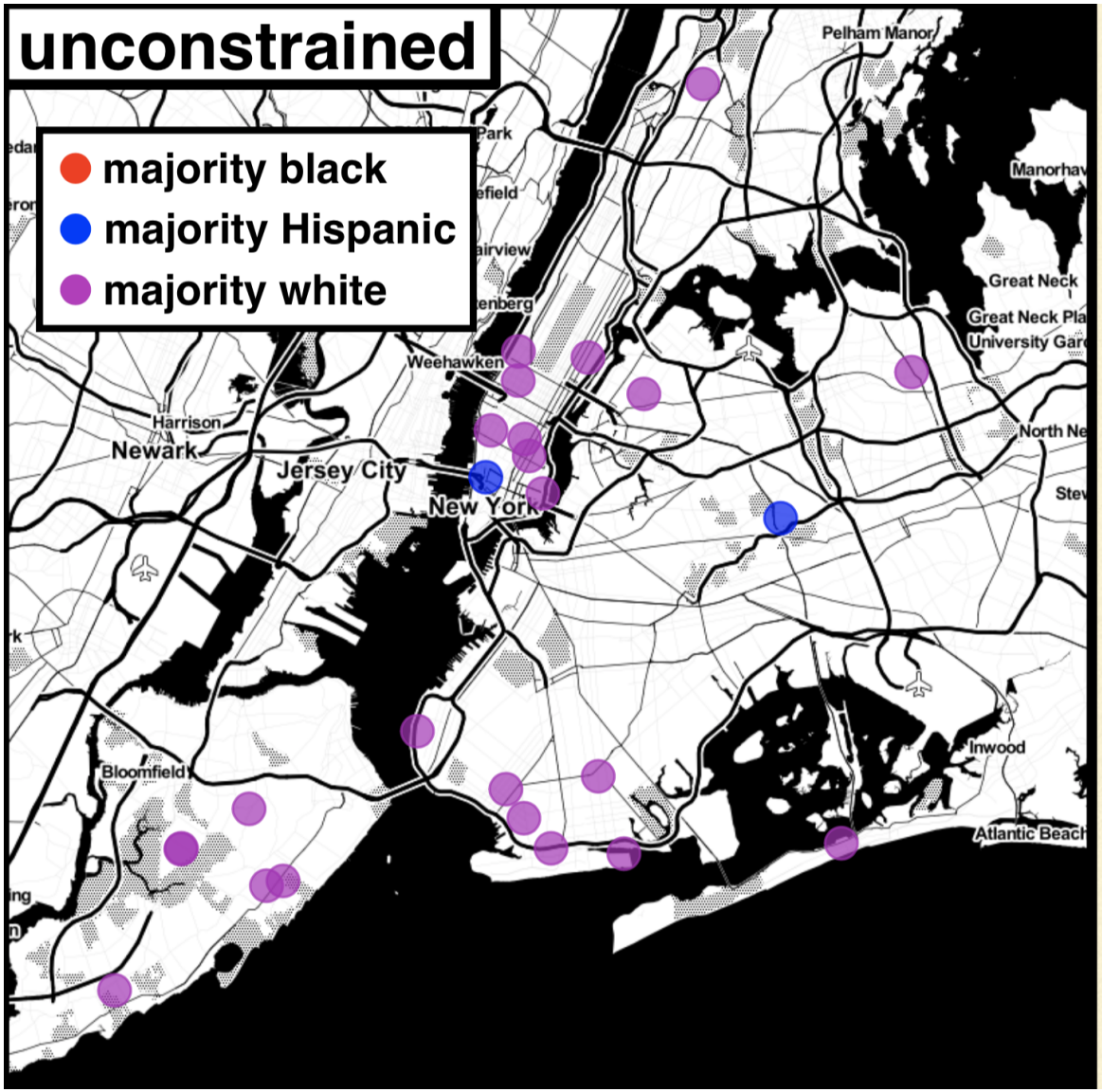
Optimal intervention under interference

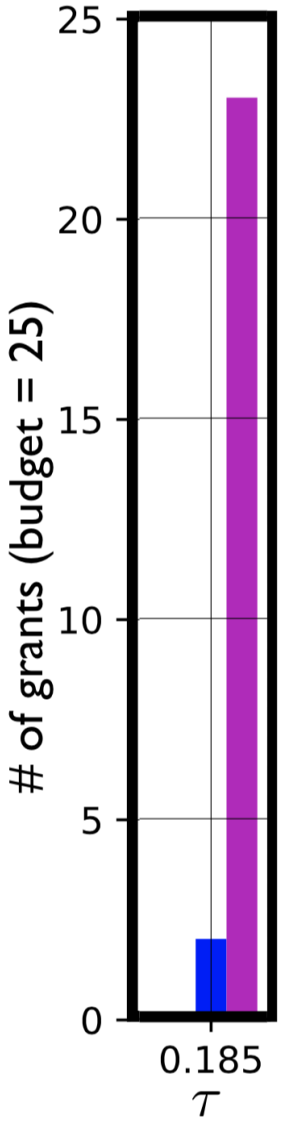
- Our goal is to design optimal interventions or policies $\mathbf{Z}$ subject to a budget constraint, e.g.

$$\mathbf{Z} = \arg\max \sum_i \mathbb{E}\left[\widehat{\mathbf{Y}}^{(i)}(a^{(i)}, \mathbf{Z}) | \mathbf{A}^{(i)}, \mathbf{X}^{(i)}\right] \quad s.t. \quad \sum_i \mathbf{Z}^{(i)} \leq b$$

- Interference means $\mathbf{Y}^{(i)}$ is potentially a function of all of $\mathbf{Z}$ and not just $\mathbf{Z}^{(i)}$
- Next two slides: optimal interventions with and without counterfactual privilege constraint

School resource allocation without fairness constraint





School resource allocation bounded counterfactual privilege