Universal Adaptability

A Target-Independent Approach to Inference

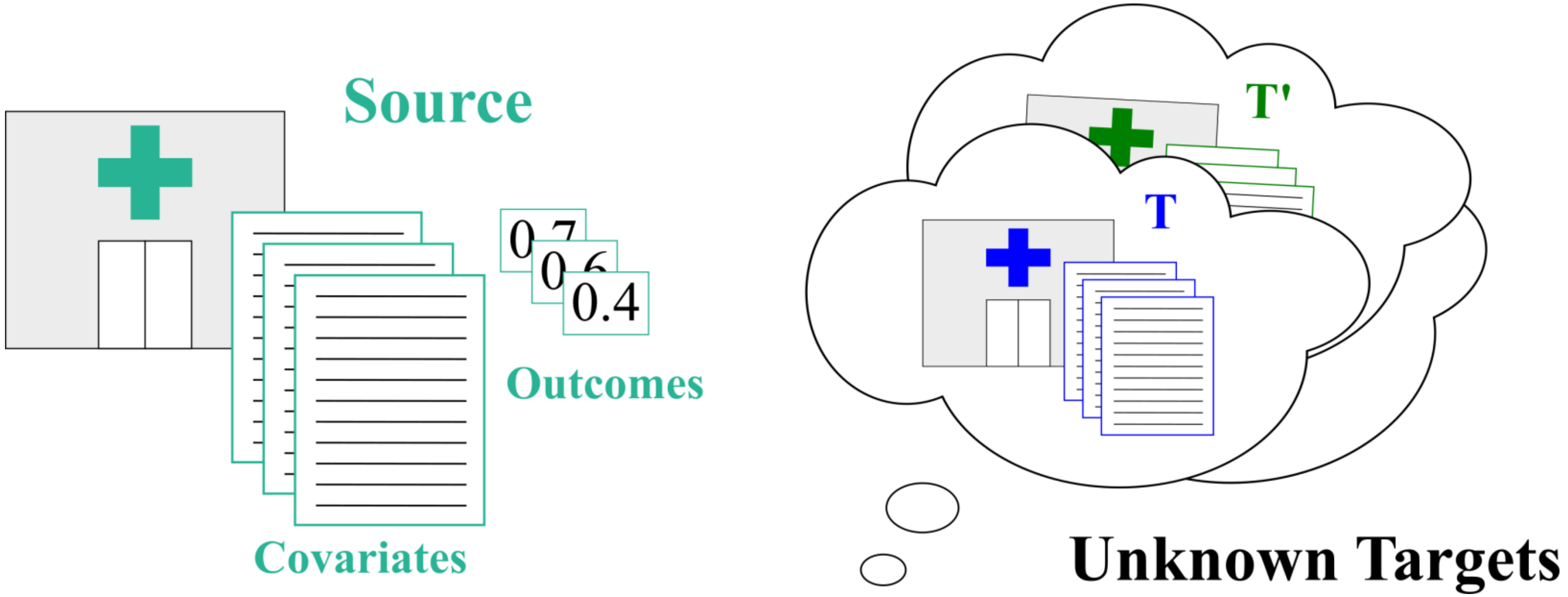
Michael P. Kim
Miller Institute for Basic Research in Science
UC Berkeley

Christoph Kern
School of Social Sciences, University of Mannheim
JPSM, University of Maryland

Joint work with Shafi Goldwasser, Frauke Kreuter, Omer Reingold

Proceedings of the National Academy of Sciences 119(4)

Inference Challenge



Data collected from source that differs from target population

Overview

- Common Approach: ***Propensity Score Reweighting***
- Key Challenge:
  - Single source, many targets
  - Proposal: ***Universal Adaptability***
- Multicalibrated Predictors are Universally Adaptable!
- MCBoost algorithm and applications

Inference Challenge

Goal: Given access to

- ***labeled*** source data $\{(X_i, Y_i)\} \sim s$
- ***unlabeled*** target data $\{(X_i, ?\,)\} \sim t$

estimate average outcome Y in target.

Challenge: source/target populations differ in composition

Classic guarantees of statistical validity fail!

Statistical Estimation Setup

- **Formal Setup**
  - $X \in \mathcal{X}$ — Covariates (features of individuals)
  - $Y \in \mathcal{Y} \subseteq [0,1]$ — Outcome of interest (real or discrete)
  - $Z \in \{s, t\}$ — source vs. target population
- **Goal:** target estimation $E[Y|Z = t]$
- **Assumptions:**
  - ***conditional independence***
  - ***positivity***

Handling Source→Target Shift

Single source → **single** target

Idea: control for membership in source/target populations

Reweight source population to “look like” target population

The Propensity Score

- Models the shift in covariates
- Odds of sampling a given individual x from source vs. target

Definition:
For source s and target t, the ***propensity score*** for given x is

$$e_{st}(x) = Pr[Z = s | X = x]$$

[Rosenbaum, Rubin '83]

Note: $1 - e_{st}(x) = Pr[Z = t | X = x]$

Valid Inference from Propensity Score

Fact: Assuming (1) conditional independence (2) $Pr[Z = s] = Pr[Z = t]$,

$$E[Y|Z = t] = E\left[\left(\frac{1 - e_{st}(X)}{e_{st}(X)}\right) \cdot Y|Z = s\right]$$

(Follows by iterated expectations and Bayes' rule.)

Valid Inference from Propensity Score

Fact: Assuming (1) conditional independence (2) $Pr[Z = s] = Pr[Z = t]$,

$$E[Y|Z = t] = E\left[\left(\frac{1 - e_{st}(X)}{e_{st}(X)}\right) \cdot Y|Z = s\right]$$

(Follows by iterated expectations and Bayes' rule.)

Approach for statistically-valid target inferences

1. Estimate e_{st} with ***combined*** samples $\{X_i\} \sim s$ and $\{X_i\} \sim t$
2. Average labeled source samples $\{(X_i, Y_i)\} \sim s$
   reweighted by propensity odds $(1 - e_{st}(X_i))/e_{st}(X_i)$

Target-Specific Inference

- Fit propensity score $\sigma \in \Sigma$ to minimize estimation error

Propensity Score Reweighting:

Given a score $\sigma : \mathcal{X} \to [0,1]$, estimate $E[Y|Z = t]$ as

$$PS_{st}(\sigma) = E\left[\left(\frac{1 - \sigma(X)}{\sigma(X)}\right) \cdot Y|Z = s\right]$$

For a class of propensity scores Σ, we measure the estimation error as:

$$\text{error}(PS_{st}(\Sigma)) = \min_{\sigma \in \Sigma} |PS_{st}(e_{st}) - PS_{st}(\sigma)|$$

Flavors of Propensity Score Methods

- **Inverse Propensity Score Weighting (IPSW)**
  - Direct use of propensity scores for reweighting [Elliott, Valliant 2017]

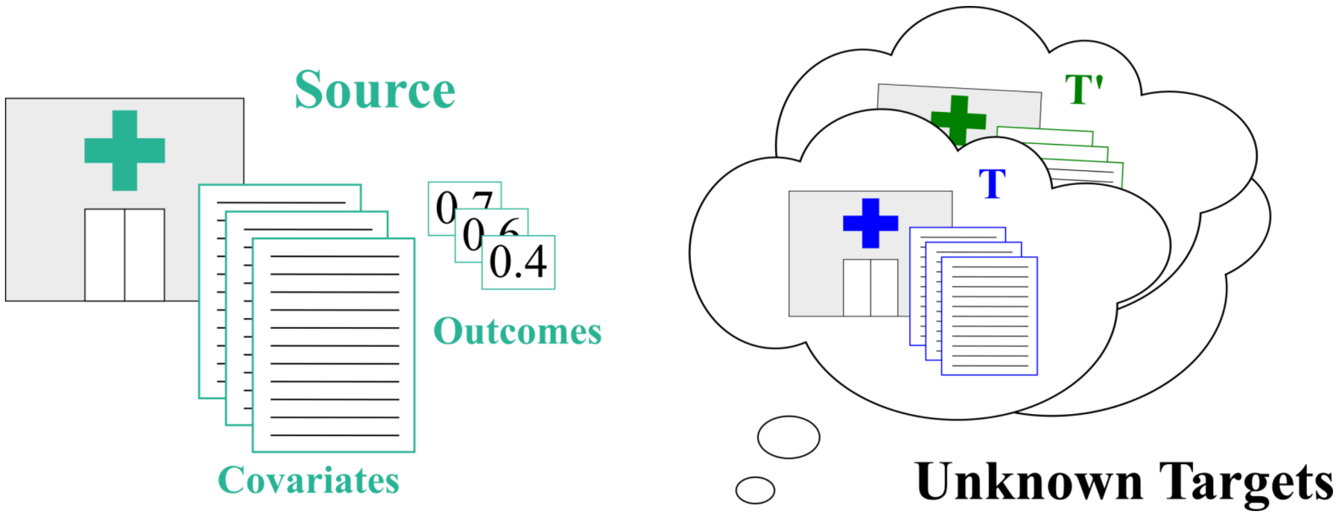
- Propensity Score Adjustment by Subclassification/ Kernel Weighting
  - Use of propensity scores as a similarity measure [Lee, Valliant 2009], [Wang, Graubard, Katki, Li 2020], [Kern, Li, Wang 2020]

- Doubly-robust approaches
  - Combined use of propensity score weighting and imputation [Chen, Li, Wu 2020]

Key Challenge

Single source → many different targets!

- *s: large medical study run by UMD*
- *t: different hospital populations across the country*



Key Challenge

Single source → many different targets!

- s: *large medical study run by UMD*
- t: *different hospital populations across the country*

Challenge: Reweighting for every target is costly

Insight from study requires target-specific propensity score

Burden lies with target communities to reweight

Key Challenge

Single source → many different targets!

- s: *large medical study run by UMD*
- t: *different hospital populations across the country*

Challenge: Reweighting for every target is costly

Insight from study requires target-specific propensity score

Burden lies with target communities to reweight

Goal: Provide insights in a "universal" format

Reorient responsibility to reweight at the source

Target-Independent Inference?

Target-Specific Inference
e.g., propensity scoring

Training Time:
unlabeled samples from s
unlabeled samples from t
Evaluation Time:
labeled samples from s

Target-Independent Inference

Training time:
labeled samples from s

Evaluation Time:
unlabeled samples from t

Target-Independent Inference?

- Learn an ***outcome predictor*** $p: \mathcal{X} \to \mathcal{Y}$ from source data
- Average the "imputed" value in target distribution

Imputation:

Given a predictor $p: \mathcal{X} \to [0,1]$, estimate $E[Y|Z = t]$ as

$$\hat{\mu}_t(p) = E[p(X)|Z = t]$$

We measure the imputation error as:

$$\text{error}(\hat{\mu}_t(p)) = |E[Y|Z = t] - \hat{\mu}_t(p)|$$

Universal Adaptability

- Predictor trained on source may give bad predictions on target

Again, classic guarantees of validity fail!

Universal Adaptability

- Predictor trained on source may give bad predictions on target

Again, classic guarantees of validity fail!

Definition: For a fixed source s, and a class of propensity scores Σ, a predictor $\tilde{p}$ is (Σ, β)-***universally adaptable***, if for ***any*** target t,

$$\text{error}(\hat{\mu}_t(\tilde{p})) \leq \text{error}(PS_{st}(\Sigma)) + \beta$$

Possibility of Universal Adaptability?

- Do universally-adaptable predictors exist?

Fact: For every class of scores Σ, the ***optimal predictor*** $p^*(x) = E[Y|X = x]$ is $(\Sigma, 0)$-universally adaptable.

Proof:

$$\hat{\mu}_t(p^*) = E[p^*(X)|Z = t] = E[E[Y|X = x]|Z = t] = E[Y|Z = t]$$

Thus, for any class of propensity scores Σ

$$\text{error}(\hat{\mu}_t(p^*)) = 0 \leq \text{error}(PS_{st}(\Sigma))$$

Anticipating Covariate Shifts

- Learning optimal predictions p^* is infeasible :(

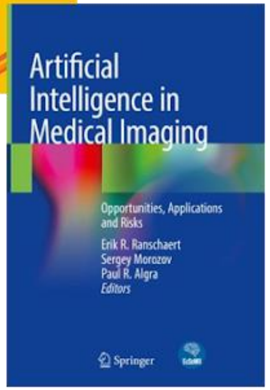
Too strong! p^* is valid under ***every*** possible shift

- Universal Adaptability: validity under ***restricted class*** of shifts
  - For each $\sigma \in \Sigma$, imagine source data under target shift σ
  - Ensure predictor $\tilde{p}: \mathcal{X} \to [0,1]$ valid on shifted data

A Detour: Algorithmic Fairness

- Predictive algorithms are everywhere we look


"...and if anyone here suspects that the algorithm that put these two together might be flawed, speak now..."

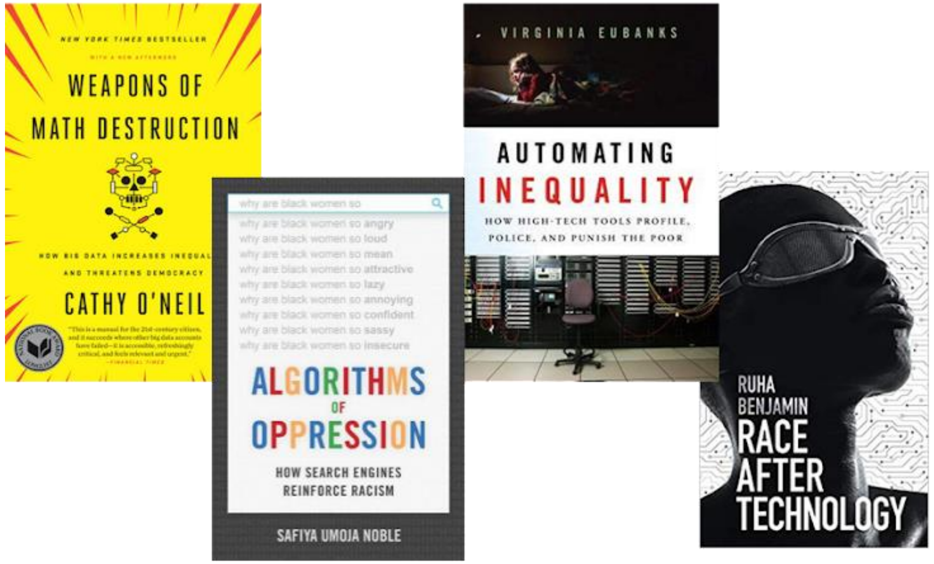


A Detour: Algorithmic Fairness

- Concern: these algorithms may be biased!

LOUISE MATSAKIS BUSINESS 04.06.19 07:00 AM

FACEBOOK'S AD SYSTEM MIGHT BE HARD-CODED FOR DISCRIMINATION



Who's a CEO? Google image results can shift gender biases

UNIVERSITY OF WASHINGTON

SHARE PRINT E-MAIL

IMAGE: PERCENTAGE OF WOMEN IN TOP 100 GOOGLE IMAGE SEARCH RESULTS FOR CEO IS: 11 PERCENT. PERCENTAGE OF US CEOS WHO ARE WOMEN IS: 27 PERCENT. view more ›

Amazon reportedly scraps internal AI recruiting tool that was biased against women

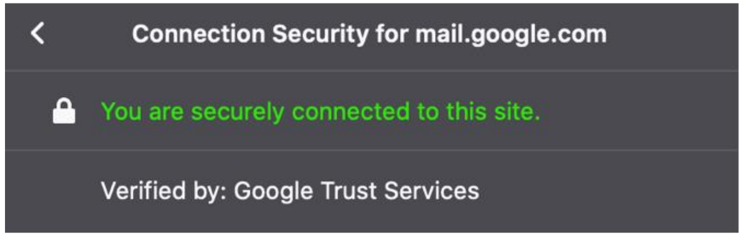
The secret program penalized applications that contained the word "women's"

By James Vincent on October 10, 2018 7:09 am

A theory of "fair" predictions

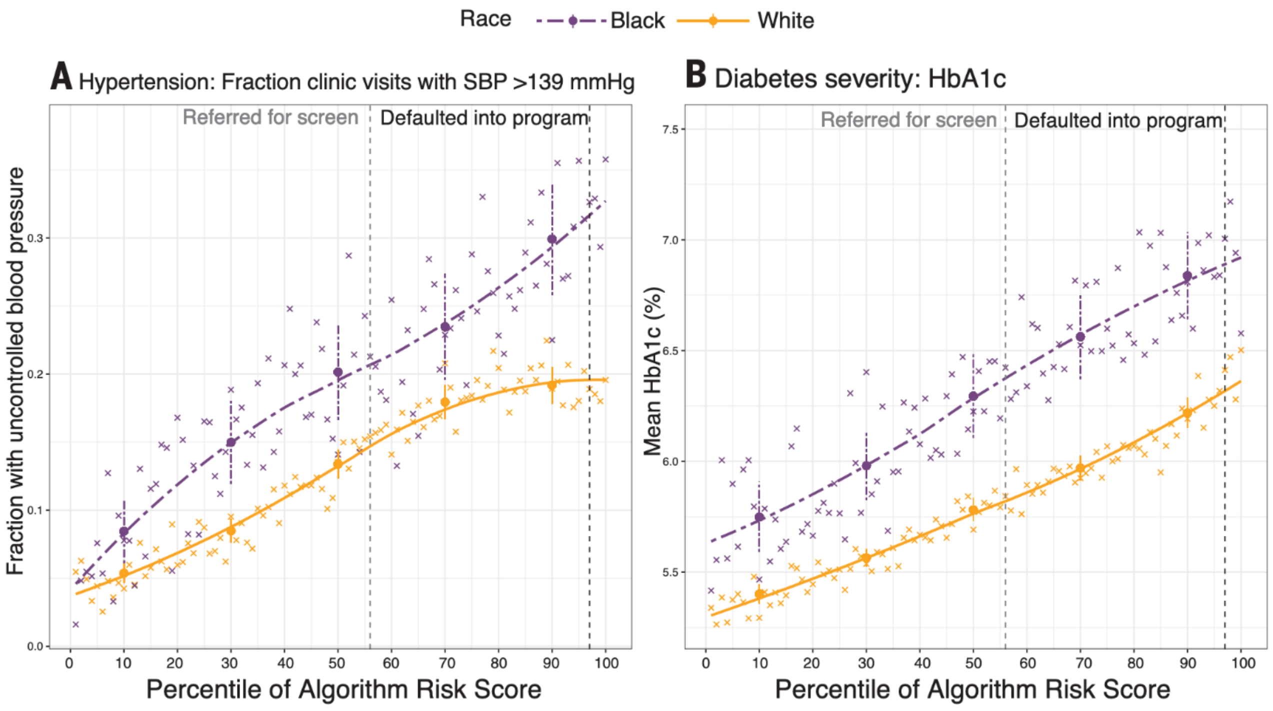
- Formalize goals in the language of CS Theory / Statistics
- Emphasis on Definitions and Abstractions

- **Eventual goal:** provable certificates of fair predictions
  - Analogy to modern cryptography



Miscalibration leads to unfair decisions

- Predictions mean different things in different groups



[Obermeyer, Powers, Vogeli, Mullainathan 2019]

Preventing Systematic Bias

- Objection: predictions miscalibrated across groups

Definition: A predictor p is ***calibrated***, if for every $v \in [0,1]$

$$E[\,Y \mid p(X) = v\,] \approx v$$

Calibration often necessary, but insufficient for fairness

Protecting subpopulations

- Group-wise calibration insufficient
- Protect subpopulations!

Protect Black women, who live in Baltimore and wear glasses

Not simply by race and gender marginally!

Calibration for every “computationally-identifiable” group?

Multicalibration

- Calibration for every "computationally-identifiable" group

Definition: For a class of functions $C \subseteq \{c: \mathcal{X} \to \mathbb{R}^+\}$, a predictor $\tilde{p}$ is (C, α)-***multicalibrated***, if for every $c \in C$

$$E[c(X) \cdot (Y - \tilde{p}(X))] \leq \alpha$$

[Hébert-Johnson, **Kim**, Reingold, Rothblum '18]

- Think of C as:
  - A collection of demographic subpopulations
  - A learnable hypothesis class (e.g., decision trees, linear functions, etc.)

End of Detour: Universal Adaptability

- Predictor trained on source may give bad predictions on target

Again, classic guarantees of validity fail!

Definition: For a fixed source s, and a class of propensity scores Σ, a predictor $\tilde{p}$ is (Σ, β)-***universally adaptable***, if for ***any*** target t,

$$\mathrm{error}(\hat{\mu}_t(\tilde{p})) \leq \mathrm{error}(PS_{st}(\Sigma)) + \beta$$

Mitigating Bias Across Subpopulations

Analogy between two goals

Fairness goal: protect subpopulations from miscalibrated predictions

Statistical goal: ensure unbiased estimates on downstream targets

Multicalibration Guarantees Universal Adaptability

- Given a class of propensity scores Σ, consider the class of functions $C(\Sigma)$ defined as

$$C(\Sigma) = \{c_\sigma : \sigma \in \Sigma\} \qquad \text{where} \quad c_\sigma(x) = \frac{1 - \sigma(x)}{\sigma(x)}$$

Theorem [**K**KGKR'22]**:** If $\tilde{p}$ is $(C(\Sigma), \alpha)$-multicalibrated over source s, then $\tilde{p}$ is (Σ, β)-universally adaptable for $\beta \leq \alpha + \delta_{st}(\Sigma)$.

where $\delta_{st}(\Sigma)$ is a constant (independent of $\tilde{p}$) that captures how well Σ fits the true propensity score e_{st}

Mitigating Bias Across Subpopulations

Analogy between two goals

Fairness goal: protect subpopulations from miscalibrated predictions

Statistical goal: ensure unbiased estimates on downstream targets

The role of concept class C for multicalibration

C identifies ***qualified minority*** subpopulations

C identifies ***potential shifts*** in covariate distribution

Not a Panacea

- Multicalibration cannot create information!
  - If a target subpopulation t not represented in source s, then cannot anticipate shifts toward t

But, neither can propensity scoring!

MCBoost: Post-Processing for Multicalibration

[Hébert-Johnson, **Kim**, Reingold, Rothblum ’18]

Given:

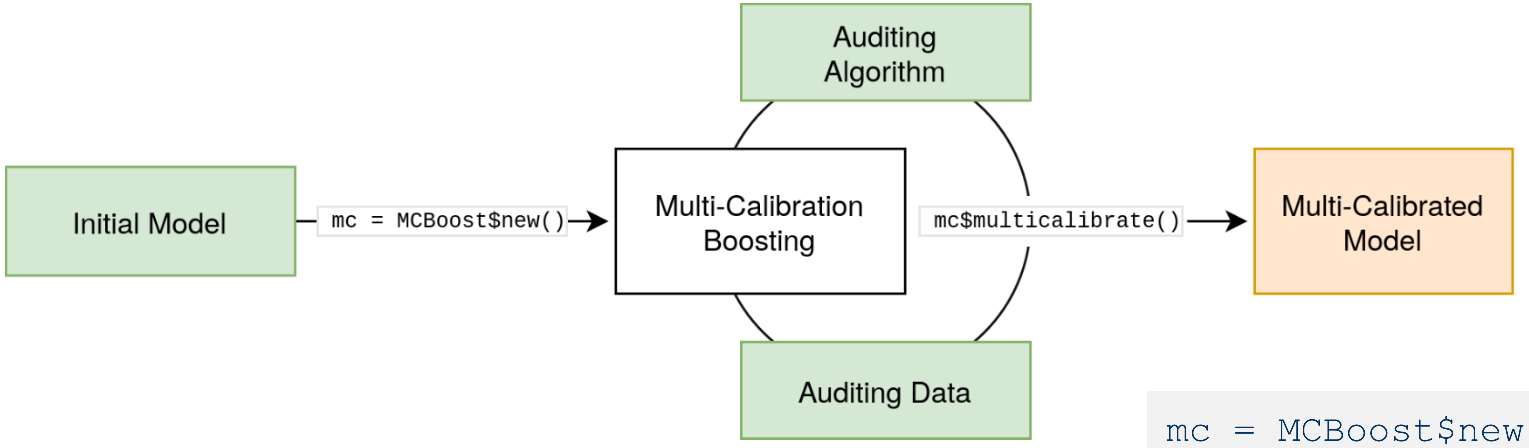
- Initial predictor $\tilde{p}$
- Validation data D
- An auditor to search for subpopulations c
  - Find largest residuals
  - e.g. ridge regression, decision tree (auditor defines collection C)

Repeat:

- Search over $c \in C$
- If $\left|E_{x\sim D}\left[c(x) \cdot \left(y - \tilde{p}(x)\right)\right]\right| > \alpha$
  - update as $\tilde{p}(x) \leftarrow \tilde{p}(x) - \eta \cdot c(x)$

MCBoost: Post-Processing for Multicalibration

R package available on CRAN – https://github.com/mlr-org/mcboost




```r
mc = MCBoost$new(
    init_predictor = init_predictor,
    auditor_fitter = "TreeAuditorFitter")
mc$multicalibrate(train_data, train_labels)
```

Empirical Evaluation

- Setting
  - Source: US National Health and Nutrition Examination Survey
  - Target: US National Health Interview Survey (weighted)
  - Estimate 15-year mortality rate across demographic groups
- Inference Methods
  - **IPSW-Overall**: Reweighting with global propensity scores (PS)
  - **IPSW-Subgroup**: Reweighting with subgroup-specific PS
  - **RF-Naive**: Mortality prediction with random forest
  - **RF-MCBoost**: Mortality prediction with multicalibrated RF

Mortality Estimation – Results

	IPSW		RF	
	Overall	Subgroup	Naive	MC-Boost
Overall	2.37 (13.5%)	—	1.11 (6.3%)	**0.52 (3.0%)**
Male	2.51 (13.4)	0.91 (4.9)	-0.34 (1.8)	**0.11 (0.6)**
Female	2.40 (14.6)	3.99 (24.2)	2.43 (14.8)	**0.90 (5.4)**
Age 18-24	**0.00 (0.1)**	-0.39 (17.5)	6.03 (270.2)	1.76 (79.0)
Age 25-44	**-0.20 (5.2)**	-0.41 (10.6)	0.82 (21.2)	0.66 (17.2)
Age 45-64	-0.75 (4.2)	-0.41 (2.3)	0.86 (4.8)	-0.29 (1.6)
Age 65-69	-4.23 (9.3)	-5.23 (11.5)	**-3.52 (7.7)**	**-1.99 (4.4)**
Age 70-74	-1.36 (2.3)	**0.47 (0.8)**	-3.02 (5.0)	**0.61 (1.0)**
Age 75+	3.53 (4.1)	2.85 (3.3)	0.51 (0.6)	2.19 (2.5)
White	3.53 (18.9)	0.75 (4.0)	1.03 (5.5)	0.69 (3.7)
Black	-4.00 (21.1)	**-0.48 (2.5)**	**-0.66 (3.5)**	**-0.52 (2.7)**
Hispanic	1.73 (17.0)	**0.48 (4.7)**	2.91 (28.6)	1.55 (15.2)
Other	**-0.02 (0.2)**	-3.54 (39.5)	3.52 (39.3)	-2.06 (23.0)

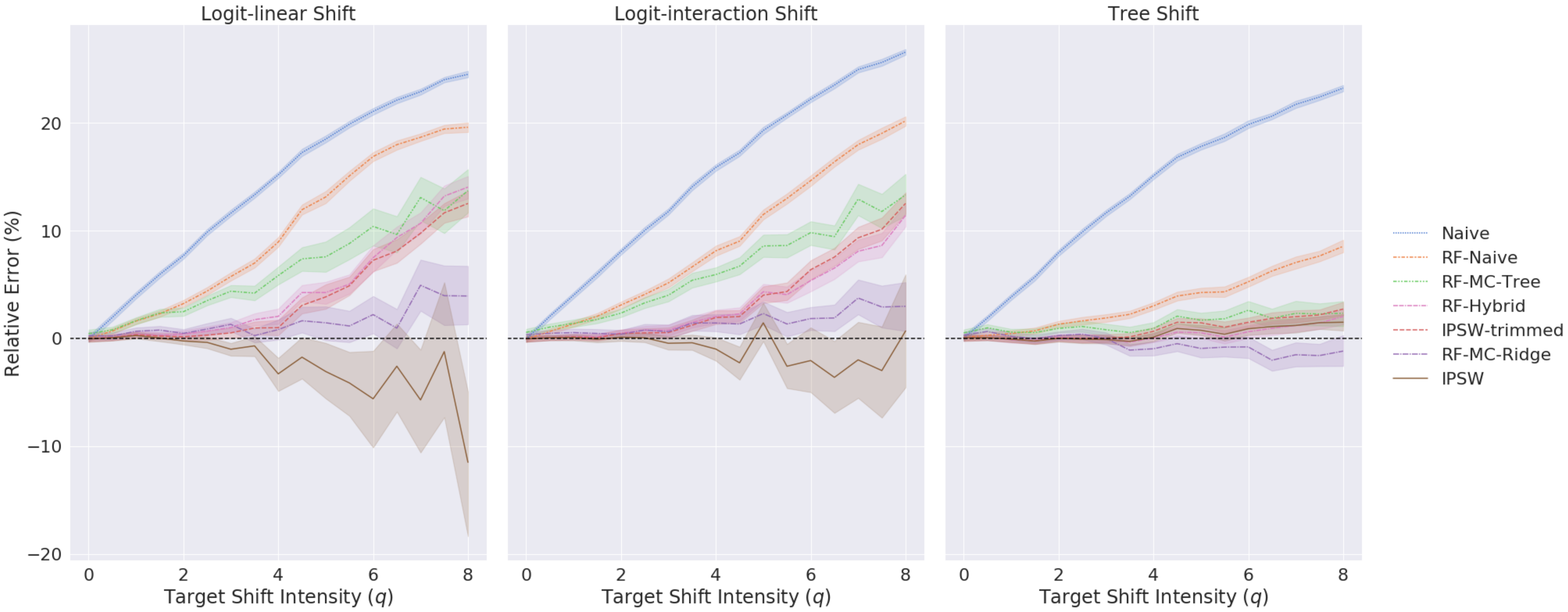
Semi-synthetic Simulation

- Setting
  - A "non-probability" sample, denoted D_{np}, based on 31,319 online opt-in panel interviews
  - A "reference population", D_p, with 20,000 observations that combines information from multiple high quality surveys
  - Estimate voting rates for the 2014 midterm election across *different degrees of covariate shift*
    1. We estimate the propensity score between D_{np} and D_p using different techniques (**Logit-linear**, **Logit-interaction**, **Tree**)
    2. For each propensity model, we generate synthetic data of various shift intensity (q) by sampling from D_{np} with weights

Semi-synthetic Simulation

- Inference Methods
  - **Naive**: Unweighted source mean as proxy for target mean
  - **IPSW**: Inverse propensity score weighting with logistic regression
  - **IPSW-trimmed**: IPSW with trimmed weights
  - **RF-Naive**: Prediction with random forest
  - **RF-Hybrid**: Prediction with random forest trained on IPSW-source
  - **RF-MC-Ridge**: Prediction with ridge-regression-multicalibrated RF
  - **RF-MC-Tree**: Prediction with tree-multicalibrated RF

Semi-synthetic Simulation – Results



Takeaways and Musings

Universal Adaptability

Valid inferences across a rich class of targets

General Result

Multicalibration persists under covariate shift

Meta-Takeaway

Algorithmic fairness useful beyond "fairness"

Thanks!