Responsible Machine Learning through the Lens of Causal Inference

Amanda Coston

December 6

Operations, Information, & Decisions Department, Wharton School

University of Pennsylvania

Carnegie Mellon University

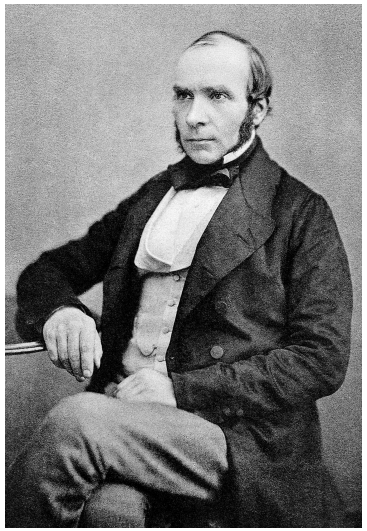
Carnegie Mellon University Heinz College

ML MACHINE LEARNING DEPARTMENT

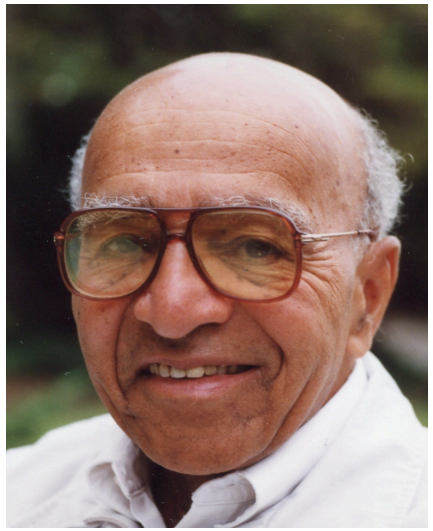
How can we use data to make better decisions?

How can we use data to make better decisions?

Florence Nightingale
Credit: Henry Hering (1814-1893) - National Portrait Gallery, London, Public Domain,



John Snow
Credit: Wikipedia



David Blackwell
Credit: George M. Bergman, CC BY-SA 4.0,

Russel Ackoff

Machine learning used to inform decisions

Lending

Criminal justice

Child welfare

Healthcare

Machine learning used to inform decisions

Lending

Criminal justice

Child welfare

Healthcare

Machine learning used to inform decisions

Lending

Criminal justice

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Child welfare

Healthcare

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE · 25 Oct 2019 · Vol 366, Issue 6464 · pp. 447-453 · DOI: 10.1126/science.aax2342

New York Regulator Probes UnitedHealth Algorithm for Racial Bias

Financial Services Department is investigating whether algorithm violates state antidiscrimination law

Machine learning used to inform decisions

Lending

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

Wells Fargo Rejected Half Its Black Applicants in Mortgage Refinancing Boom

Source: Bloomberg

Criminal justice

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Child welfare

Healthcare

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE · 25 Oct 2019 · Vol 366, Issue 6464 · pp. 447-453 · DOI: 10.1126/science.aax2342

New York Regulator Probes UnitedHealth Algorithm for Racial Bias

Financial Services Department is investigating whether algorithm violates state antidiscrimination law

Machine learning used to inform decisions

Lending

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

Wells Fargo Rejected Half Its Black Applicants in Mortgage Refinancing Boom

Source: Bloomberg

Criminal justice

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Child welfare

VIRGINIA EUBANKS BUSINESS 01.15.2018 08

A Child Abuse Prediction Model Fails Poor Families

Why Pittsburgh's predictive analytics misdiagnoses child maltreatment and prescribes the wrong solutions

Healthcare

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE · 25 Oct 2019 · Vol 366, Issue 6464 · pp. 447-453 · DOI: 10.1126/science.aax2342

New York Regulator Probes UnitedHealth Algorithm for Racial Bias

Financial Services Department is investigating whether algorithm violates state antidiscrimination law

Smartphone Location Data Can Leave Out Those Most Hit by Covid-19

Younger, white populations may be overrepresented in the data, researchers say

Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas With the Same Risk

Our analysis of premiums and payouts in California, Illinois, Texas and Missouri shows that some major insurers charge minority neighborhoods as much as 30 percent more than other areas with similar accident costs.

by Julia Angwin, Jeff Larson, Lauren Kirchner and Surya Mattu, ProPublica
April 5, 2017

This story was co-published with Consumer Reports.

Michigan's MiDAS Unemployment System: Algorithm Alchemy Created Lead, Not Gold › A case study into how to automate false accusations of fraud for more than 34,000 unemployed people

BY ROBERT N. CHARETTE | 24 JAN 2018 | 5 MIN READ

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations

SCIENCE · 25 Oct 2019 · Vol 366, Issue 6464 · pp. 447-453 · DOI: 10.1126/science.aax2342

VIRGINIA EUBANKS BUSINESS 01.15.2018 08

A Child Abuse Prediction Model Fails Poor Families

Why Pittsburgh's predictive analytics misdiagnoses child maltreatment and prescribes the wrong solutions

Wells Fargo Rejected Half Its Black Applicants in Mortgage Refinancing Boom

Source: Bloomberg

New York Regulator Probes UnitedHealth Algorithm for Racial Bias

Financial Services Department is investigating whether algorithm violates state antidiscrimination law

Court Rules Deliveroo Used 'Discriminatory' Algorithm

An Italian court determined that companies can be held liable even if an algorithm unintentionally discriminates against a protected group.

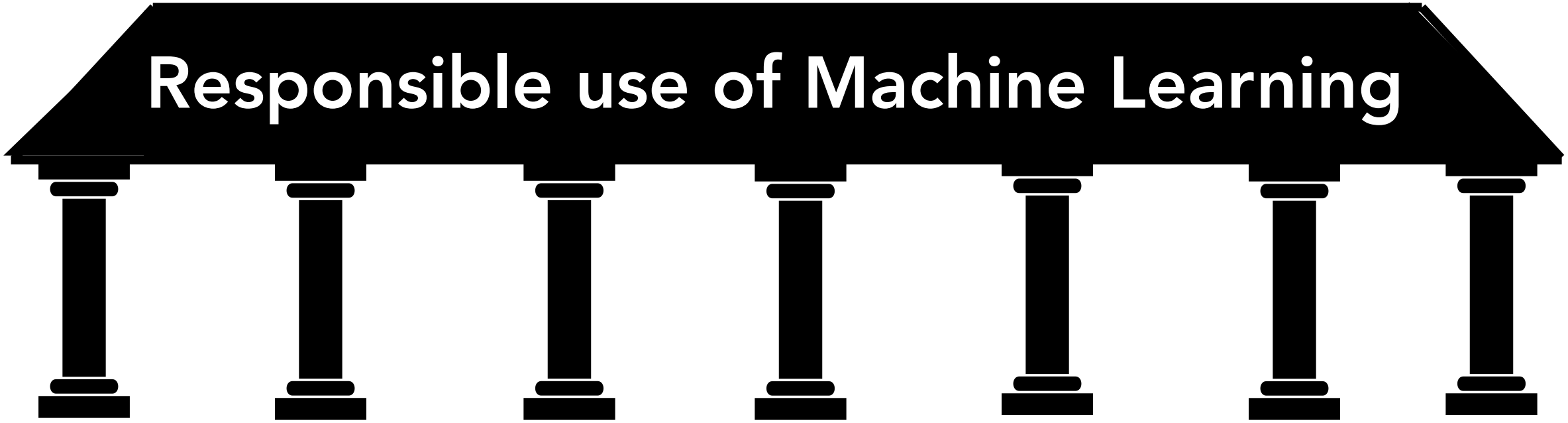
GG By Gabriel Geiger

Answer Sheet

The fundamental flaws of 'value added' teacher evaluation

By Valerie Strauss

December 23, 2012


Responsible use of Machine Learning
Transparency
Accountability
Equity
Validity
Oversight
Robustness
Privacy

Related work on responsible use

Equity & Accountability

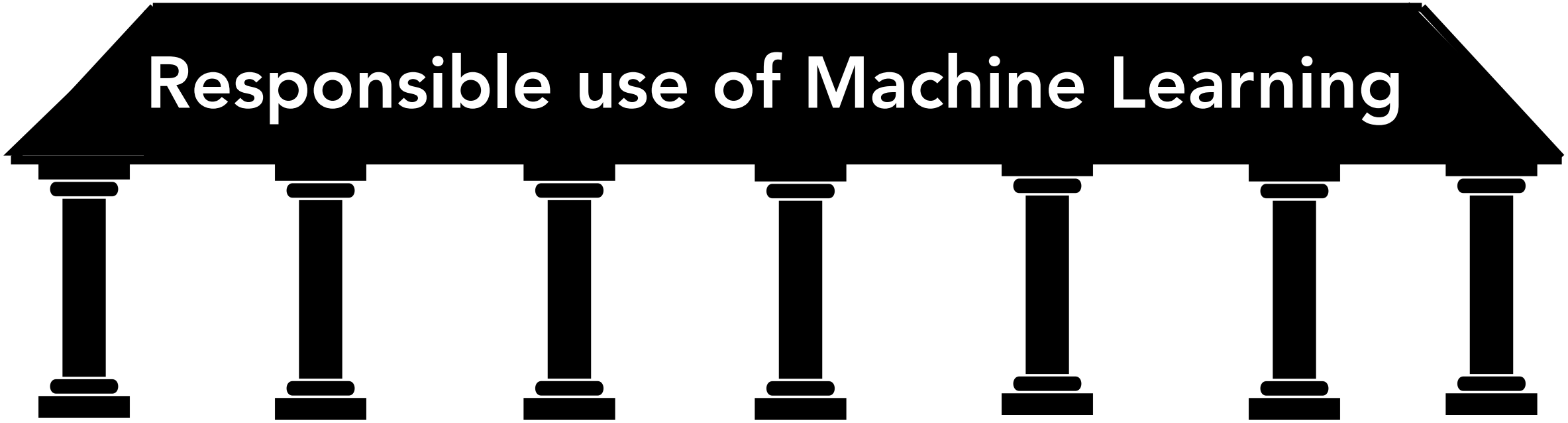
- Park, Ahn, Hosanagar, & Lee. CHI 2022.
- Lam, Gordon, Metaxa, Hancock, Landay, and Bernstein. CSCW 2022.
- Zeng, Dobriban, & Cheng. 2022.
- Cai, Encarnacion, Chern, Corbett-Davies, Bogen, Bergman, and Goel. FAccT 2022
- Metaxa, Gan, Goh, Hancock, and Landay. CSCW 2021.
- Berk, Heidari, Jabbari, Kearns, & Roth. SM&R 2021.
- Knox. Science 2021.
- Tambe, Cappelli, & Yakubovich, California Management Review. 2019.
- Kearns, Neel, Roth, & Wu. ICML 2018.
- Bastani, Kim, & Bastani. FATML 2017.

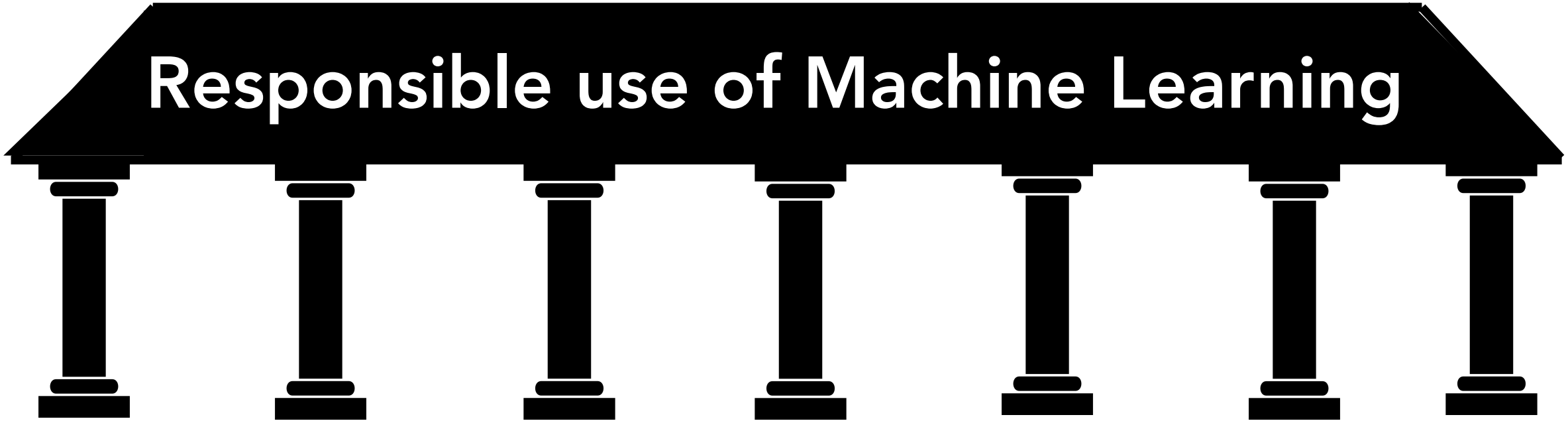
Oversight

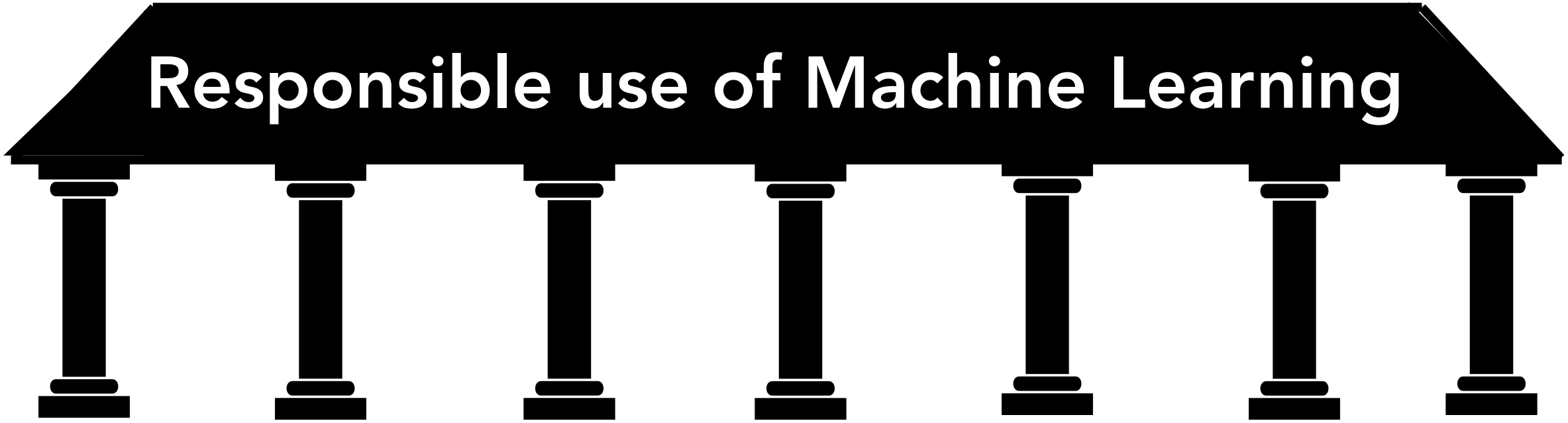
- Bastani et al. INFORMS JAA 2022.
- Hosanagar. Viking 2019.

Transparency

- Dietvorst, Simmons, Massey. Management Science 2018.
- Dietvorst, Simmons, Massey. JEP 2015


Responsible use of Machine Learning
Transparency
Accountability
Equity
Validity
Oversight
Robustness
Privacy


Responsible use of Machine Learning
Transparency
Accountability
Equity
Validity
Oversight
Robustness
Privacy


Responsible use of Machine Learning
Transparency
Accountability
Equity
Validity
Oversight
Robustness
Privacy

Inequity

Model unjustifiably advantages some over others.

Validity

"Amelia Bedelia, the sun will fade the furniture.
I asked you to draw the drapes," said Mrs. Rogers.
"I did! I did! See," said Amelia Bedelia.
She held up her picture.

Peggy Parish & Fritz Siebel

Validity

Model predicts the quantity we think it does.

"Amelia Bedelia, the sun will fade the furniture.
I asked you to draw the drapes," said Mrs. Rogers.
"I did! I did! See," said Amelia Bedelia.
She held up her picture.

Peggy Parish & Fritz Siebel

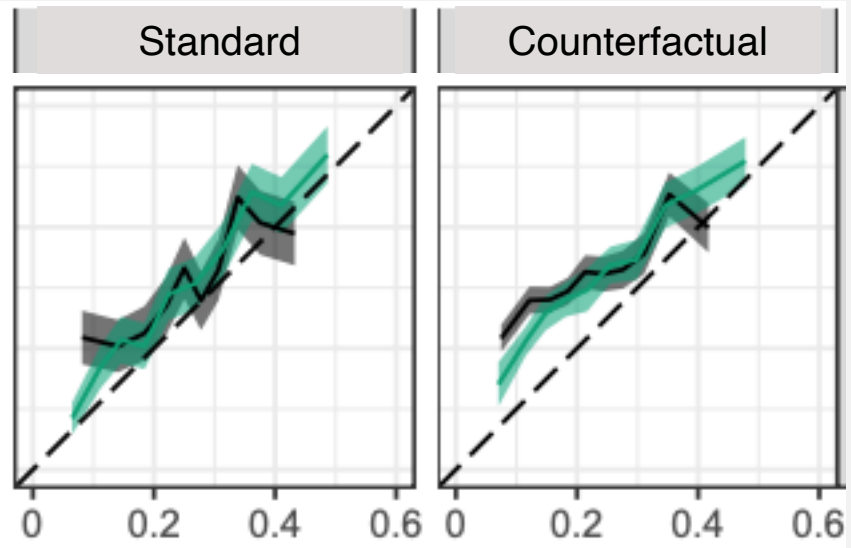
How do we assess the validity and equity of decision-making algorithms?

How do we assess the validity and equity of decision-making algorithms?

- Causal inference

How do we assess the validity and equity of decision-making algorithms?

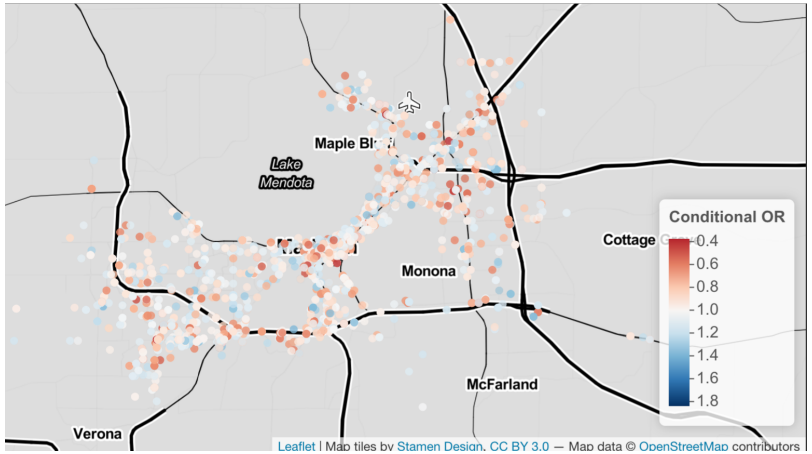
- Causal inference



Counterfactual risk assessments, evaluation, and fairness

FAccT 2020

Coston, Mishler, Kennedy, & Chouldechova



A counterfactual audit for racial bias in police traffic stops

ACIC 2022

Coston & Kennedy

Agenda

Agenda

1. Perform valid performance assessment under missing data via counterfactual evaluation

Agenda

1. Perform valid performance assessment under missing data via counterfactual evaluation

2. Address sociotechnical context via counterfactual equity audits

Agenda

1. Perform valid performance assessment under missing data via counterfactual evaluation

2. Address sociotechnical context via counterfactual equity audits

“Machine learning tool” = “algorithm” = “predictive model”

Algorithms used to inform decisions

- Predict the likelihood of an outcome if a particular decision is taken

Lending

Child welfare

Criminal justice

Algorithms used to inform decisions

- Predict the likelihood of an outcome if a particular decision is taken

Lending

Default if loan approved

Child welfare

Criminal justice

Algorithms used to inform decisions

- Predict the likelihood of an outcome if a particular decision is taken

Lending

Default if loan approved

Child welfare

Harm if not investigated

Criminal justice

Algorithms used to inform decisions

- Predict the likelihood of an outcome if a particular decision is taken

Lending

Default if loan approved

Child welfare

Harm if not investigated

Criminal justice

Recidivate if released

Algorithms used to inform decisions

- Predict the likelihood of an outcome if a particular decision is taken

Lending

Default if loan approved

Child welfare

Harm if not investigated

Criminal justice

Recidivate if released

Data

Algorithms used to inform decisions

- Predict the likelihood of an outcome if a particular decision is taken

	Lending	Child welfare	Criminal justice
			
	Default if loan approved	Harm if not investigated	Recidivate if released
Data	Commonwealth Bank of Australia	Allegheny County, PA child welfare hotline	

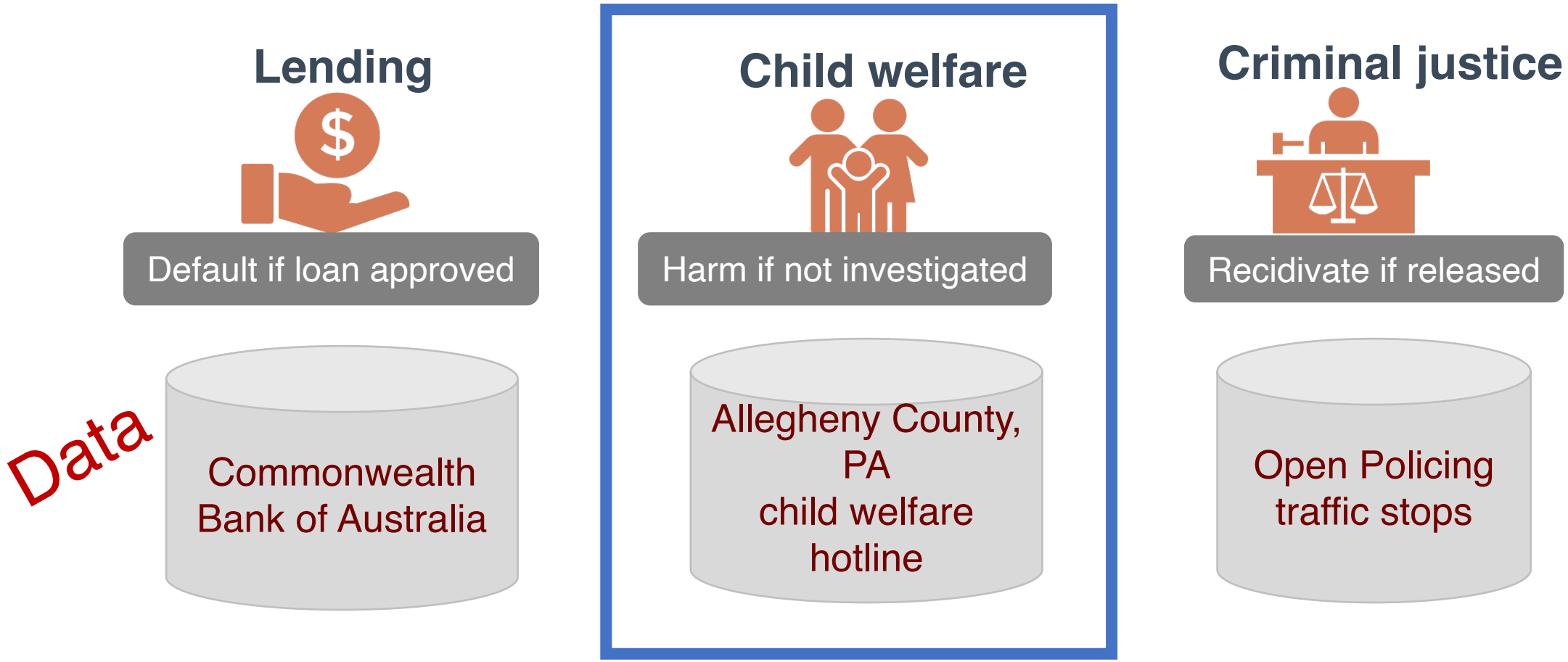
Algorithms used to inform decisions

- Predict the likelihood of an outcome if a particular decision is taken

	Lending	Child welfare	Criminal justice
			
	Default if loan approved	Harm if not investigated	Recidivate if released
Data	Commonwealth Bank of Australia	Allegheny County, PA child welfare hotline	Open Policing traffic stops

Algorithms used to inform decisions

- Predict the likelihood of an outcome if a particular decision is taken



Missing outcomes due to bandit feedback

Missing outcomes due to bandit feedback

Child welfare

Missing outcomes due to bandit feedback

Child welfare

Missing outcomes due to bandit feedback

Child welfare

Harm if not investigated (screened out)

Missing outcomes due to bandit feedback

Child welfare

Harm if not investigated (screened out)

Missing outcomes due to bandit feedback

Child welfare

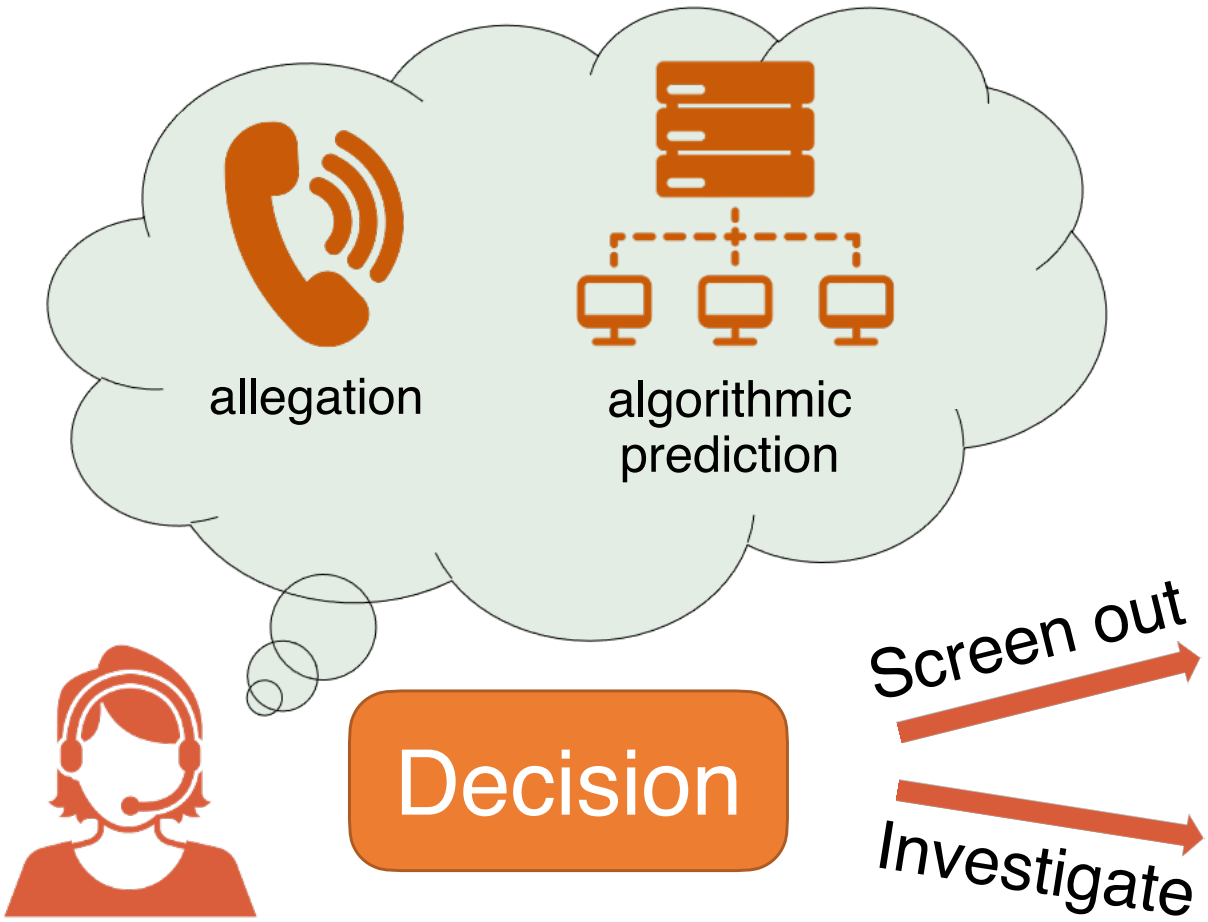
Harm if not investigated (screened out)

Missing outcomes due to bandit feedback

Child welfare

Harm if not investigated (screened out)

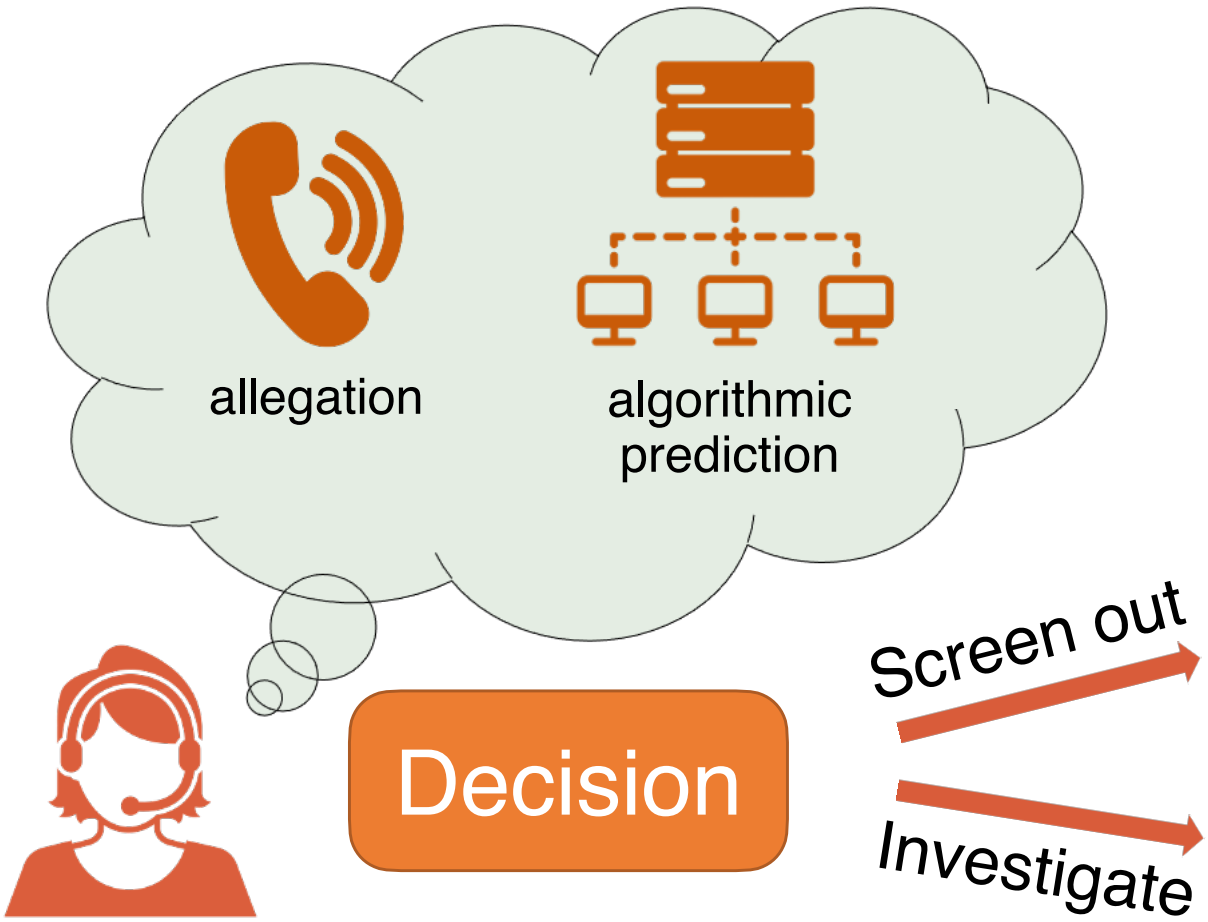
Missing outcomes due to bandit feedback



Child welfare

Harm if not investigated (screened out)

Missing outcomes due to bandit feedback



Outcomes

Child welfare

Harm if not investigated (screened out)

Missing outcomes due to bandit feedback

Outcomes

Child welfare

Harm if not investigated (screened out)

Missing outcomes due to bandit feedback

Outcomes

Child welfare

Harm if not investigated (screened out)

Missing outcomes due to bandit feedback

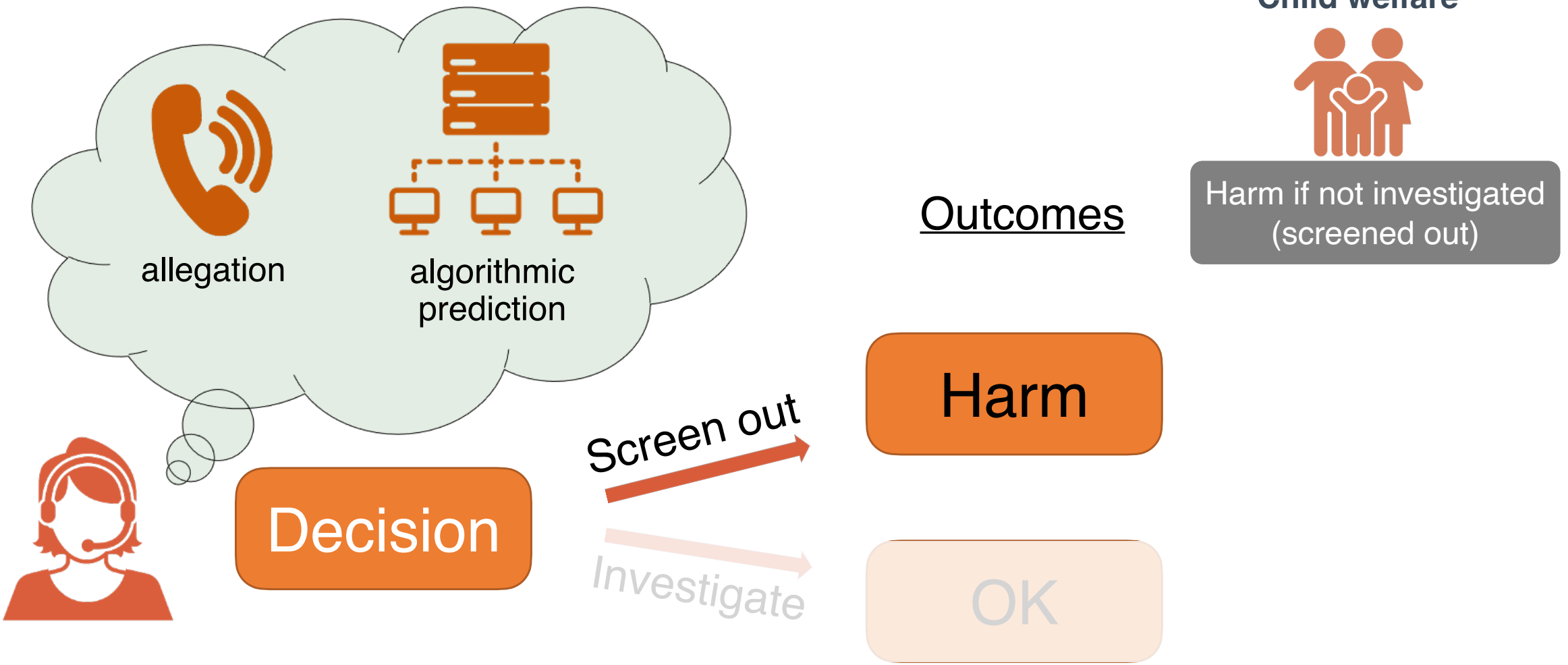
Outcomes

Child welfare

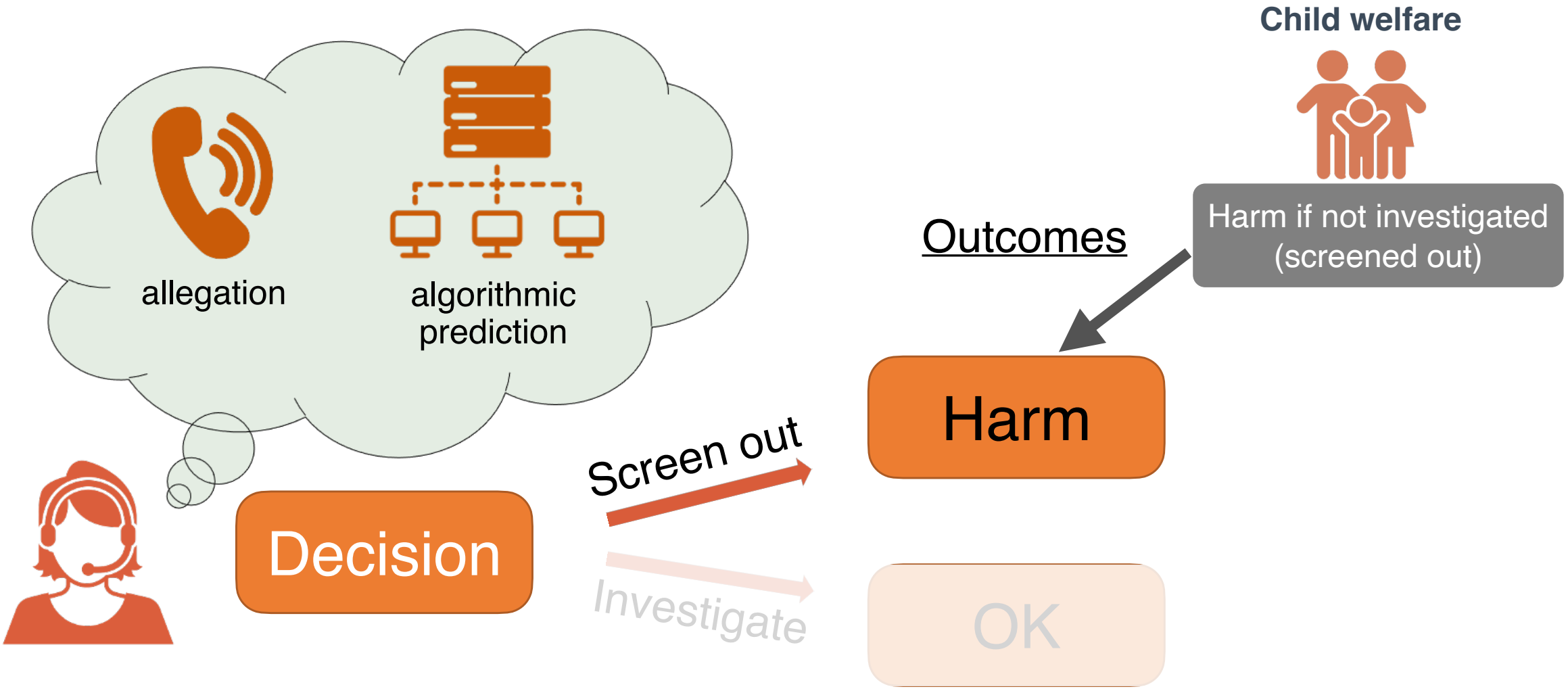
Harm if not investigated (screened out)

OK

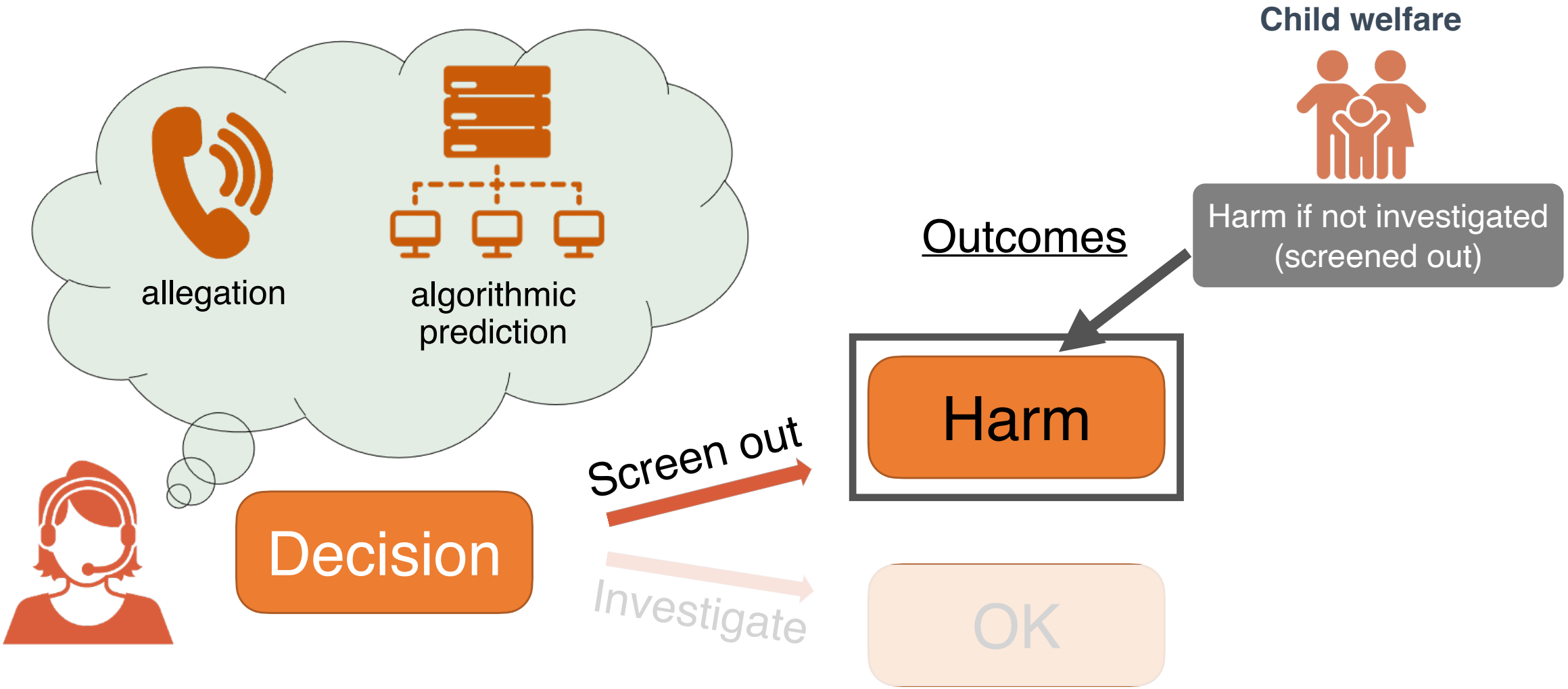
Missing outcomes due to bandit feedback



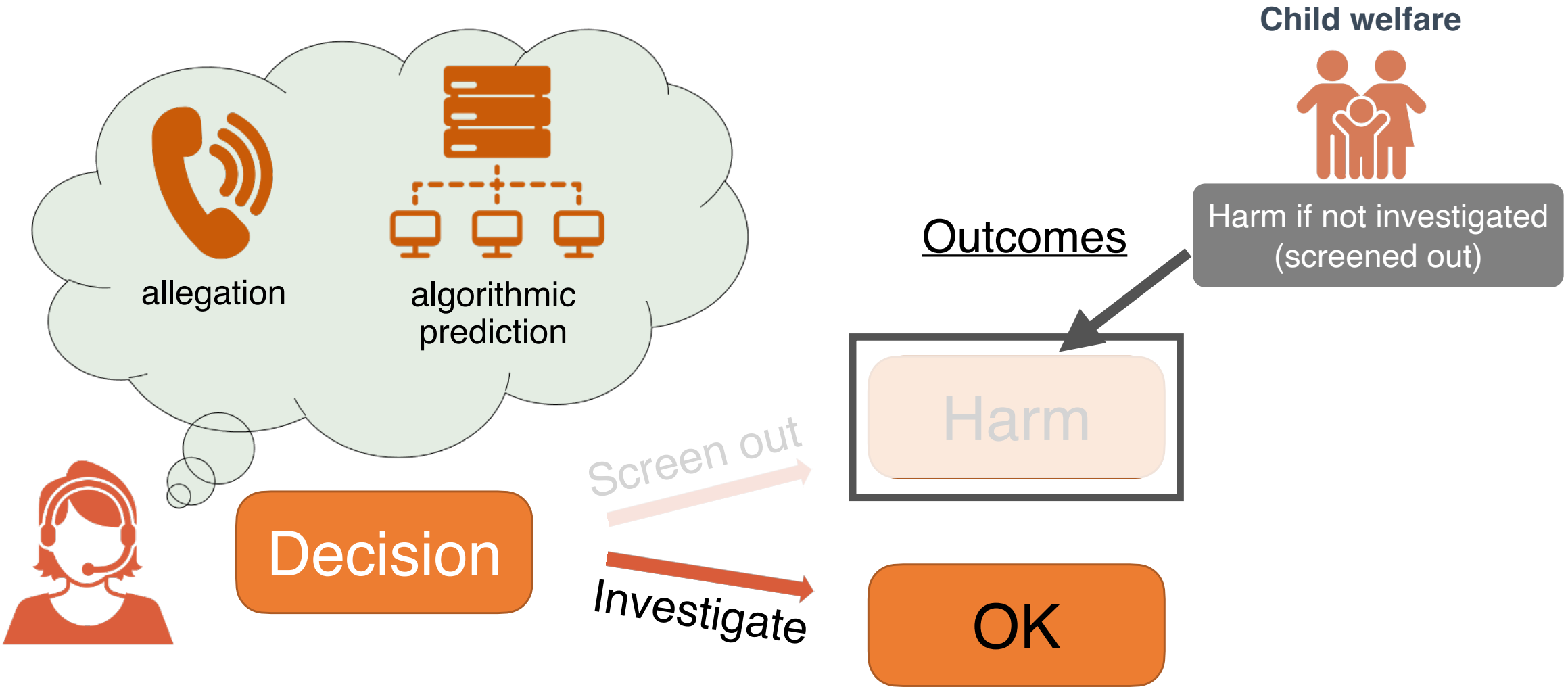
Missing outcomes due to bandit feedback



Missing outcomes due to bandit feedback



Missing outcomes due to bandit feedback



Problem

- Missing outcomes due to bandit feedback
- Standard approaches to evaluating algorithms are misleading

Problem

- Missing outcomes due to bandit feedback
- Standard approaches to evaluating algorithms are misleading

Research Question #1

- How do we evaluate predictive performance when we have missing outcomes?

Problem

- Missing outcomes due to bandit feedback
- Standard approaches to evaluating algorithms are misleading

Research Question #1

- How do we evaluate predictive performance when we have missing outcomes?

Contribution

- Counterfactual techniques to impute missing outcomes

Problem

- Missing outcomes due to bandit feedback
- Standard approaches to evaluating algorithms are misleading

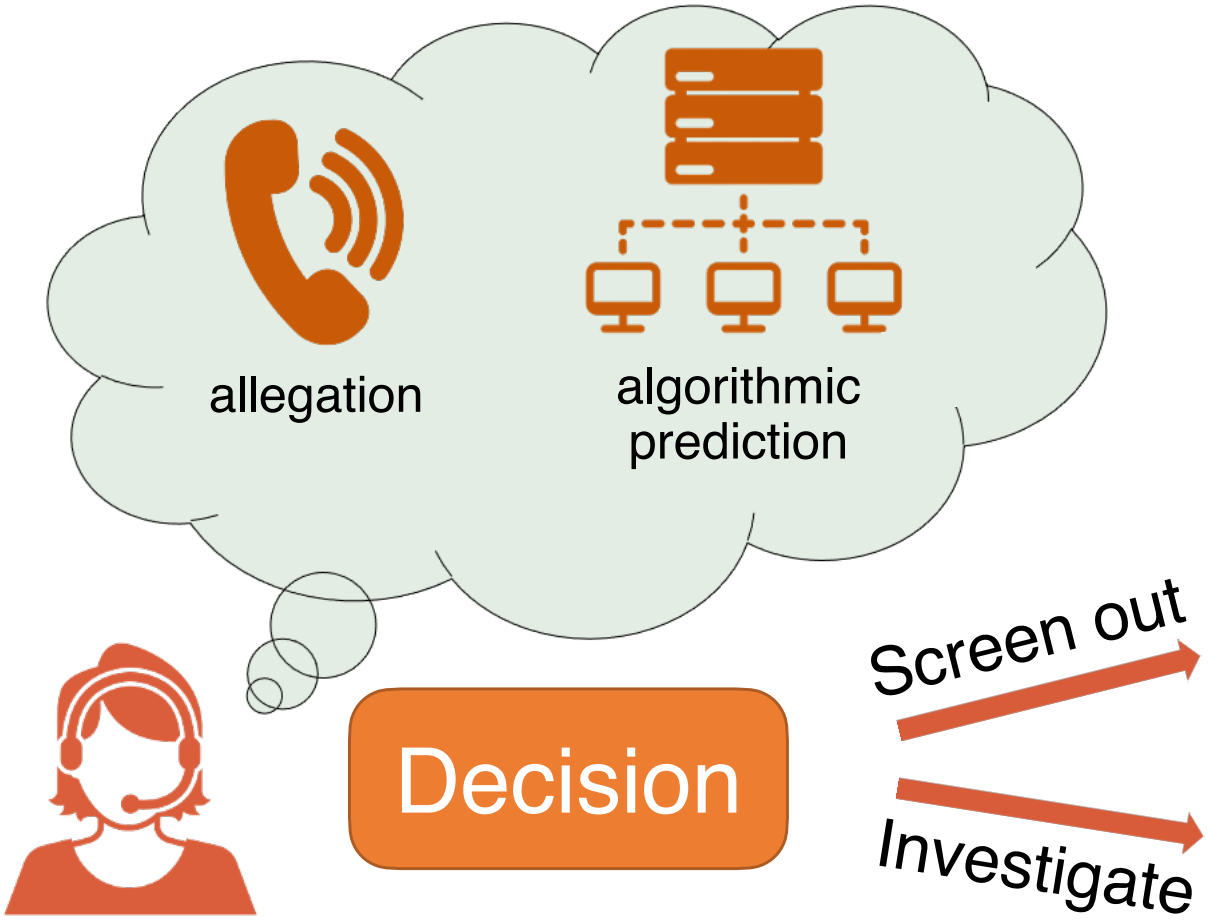
Research Question #1

- How do we evaluate predictive performance when we have missing outcomes?

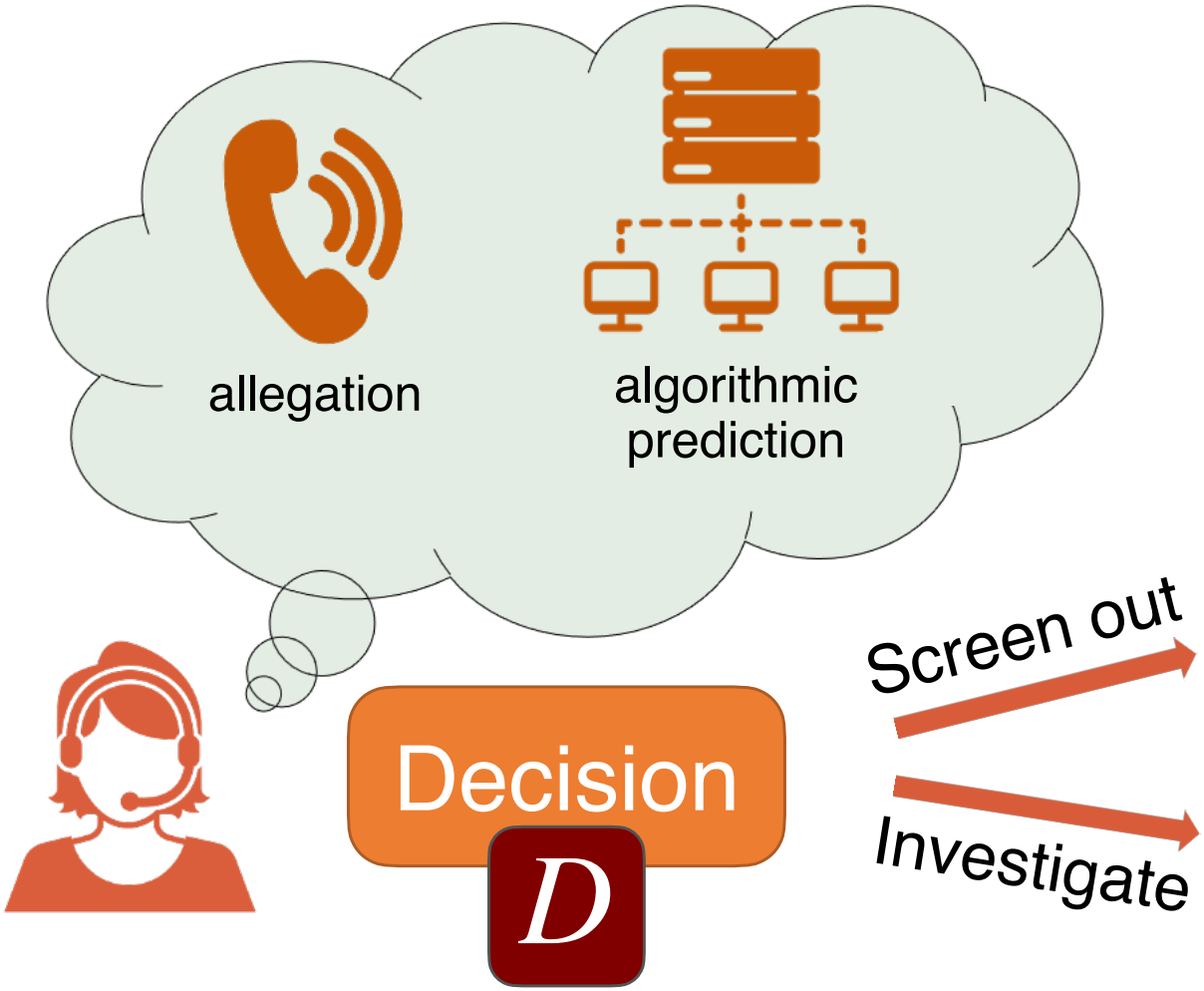
Contribution

- Counterfactual techniques to impute missing outcomes
- Valid evaluation under key conditions

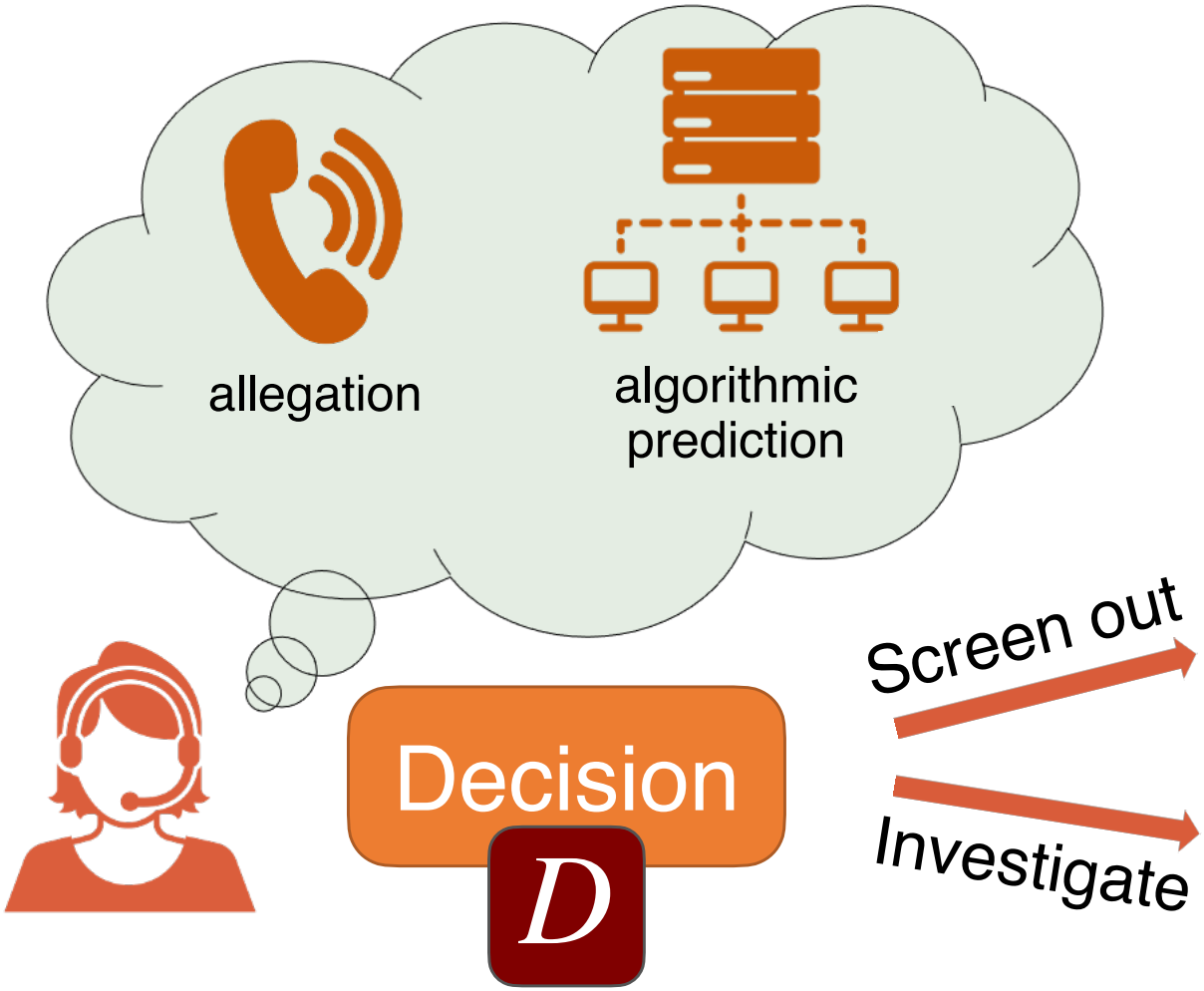
Notation


allegation
algorithmic prediction
Decision
Screen out
Investigate

Notation

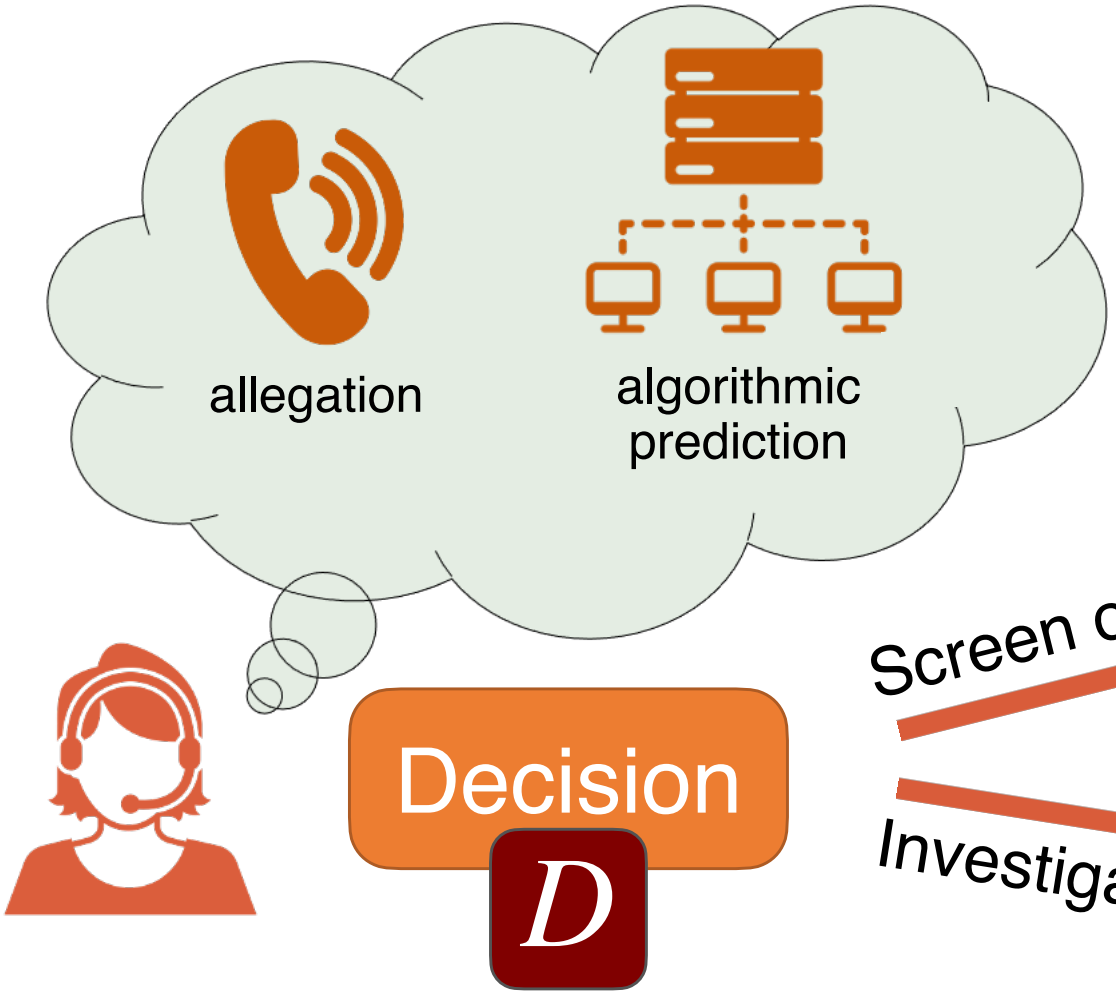

allegation
algorithmic prediction
Decision
D
Screen out
Investigate

Notation



outcomes Y

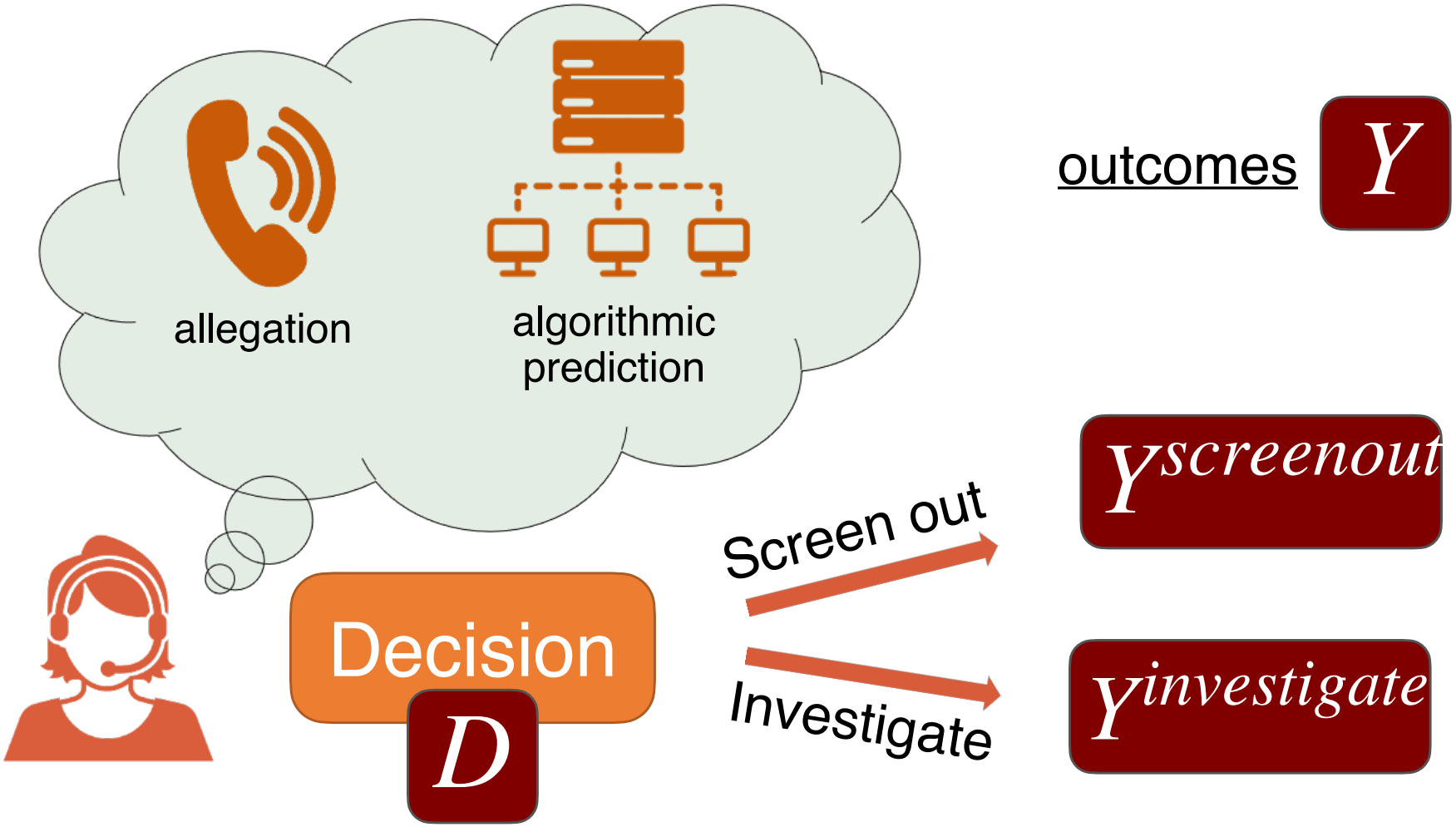
Notation



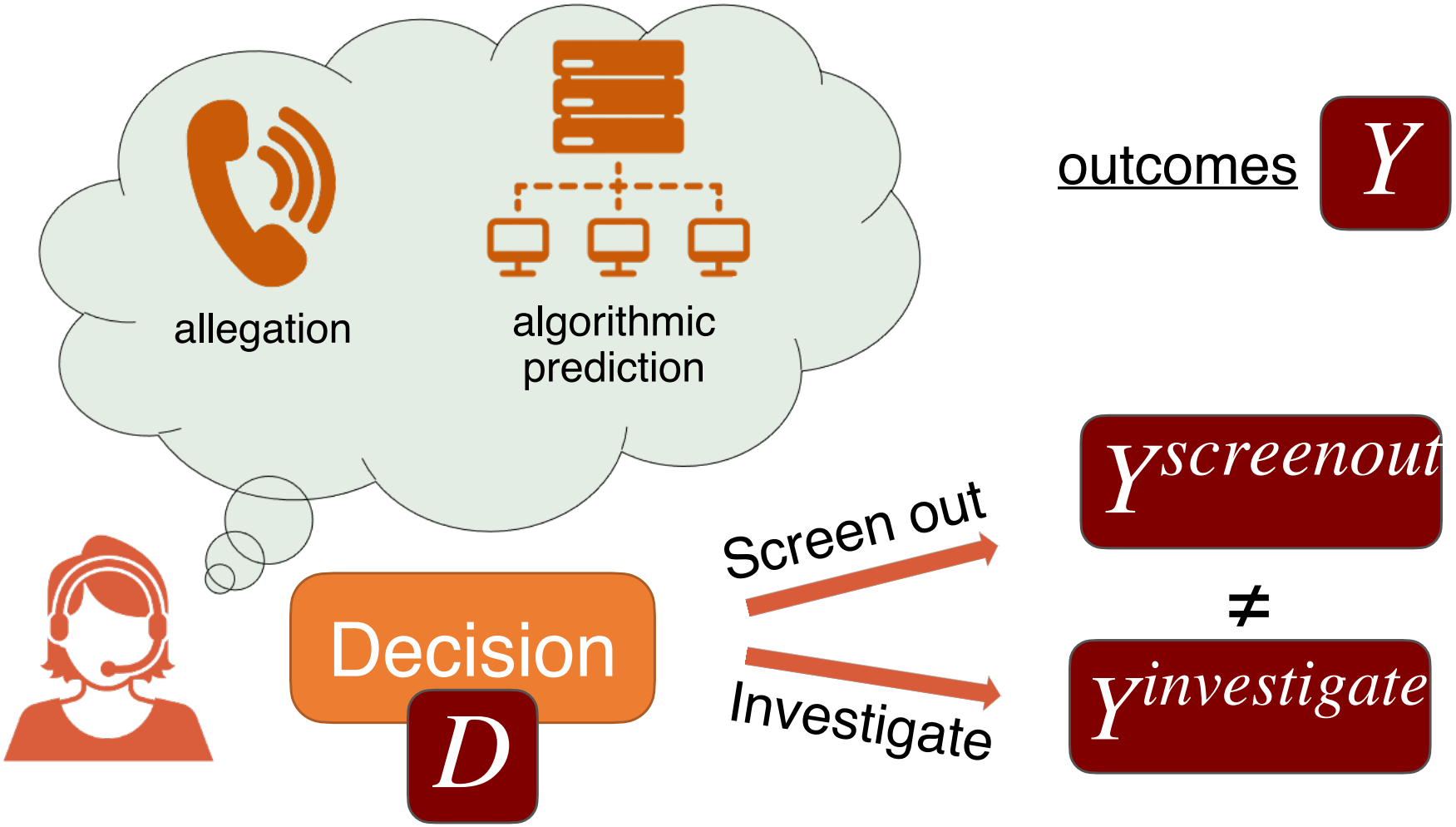
outcomes Y

$Y^{screenout}$

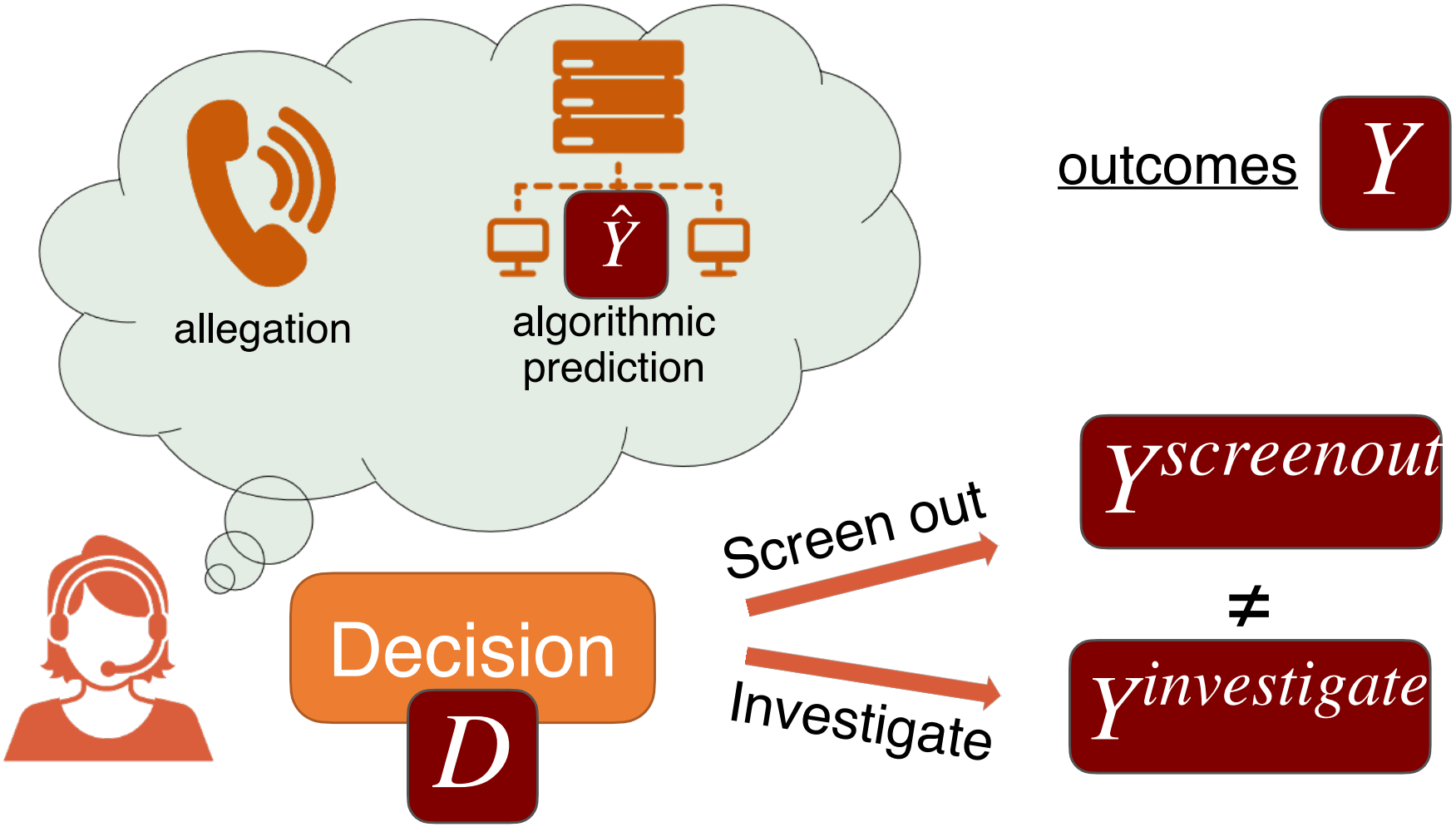
Notation


allegation
algorithmic prediction
Decision
D
Screen out
Investigate
outcomes
Y
Y^screenout
Y^investigate

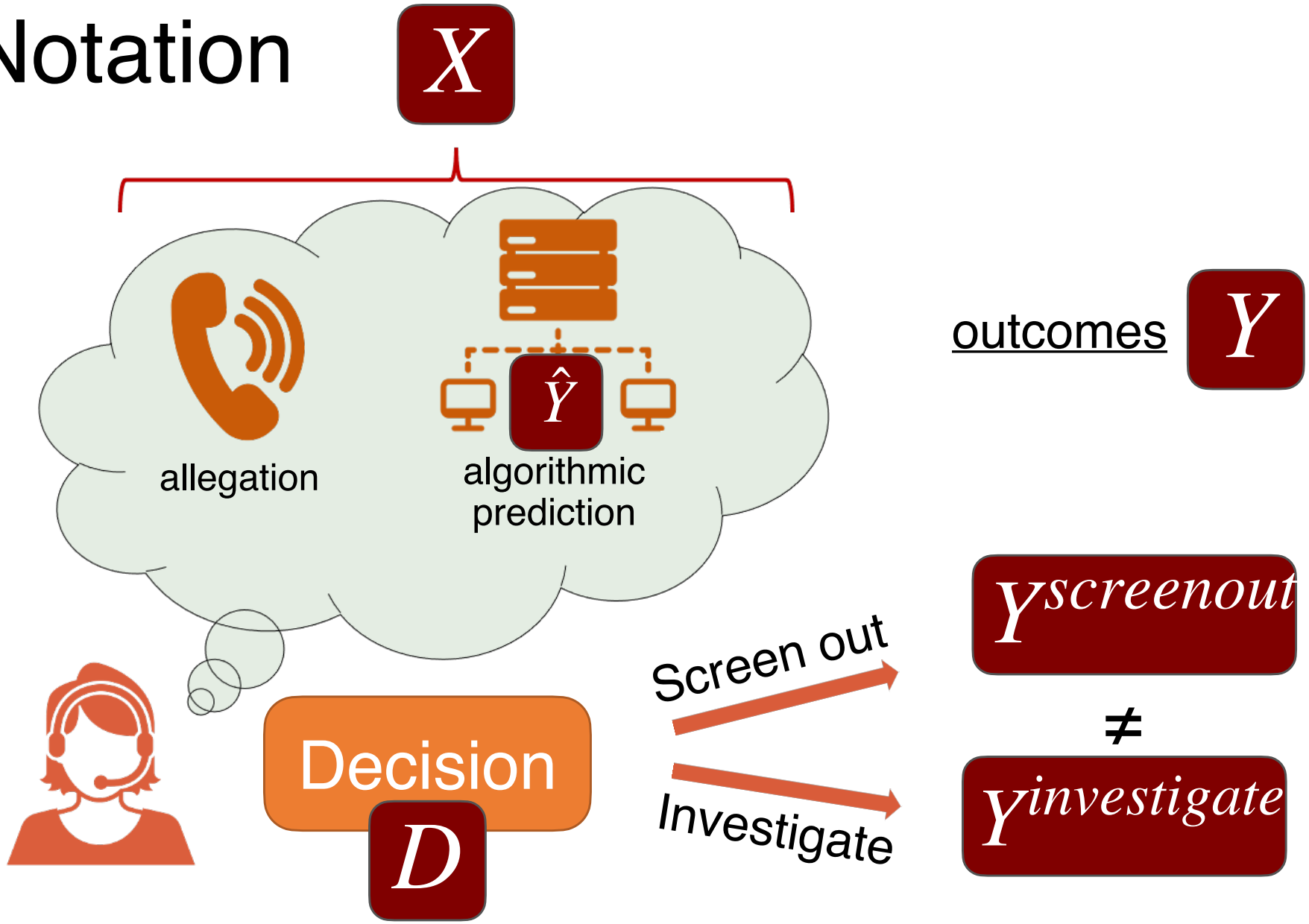
Notation


allegation
algorithmic prediction
Decision
D
Screen out
Investigate
outcomes Y
$Y^{screenout}$
≠
$Y^{investigate}$

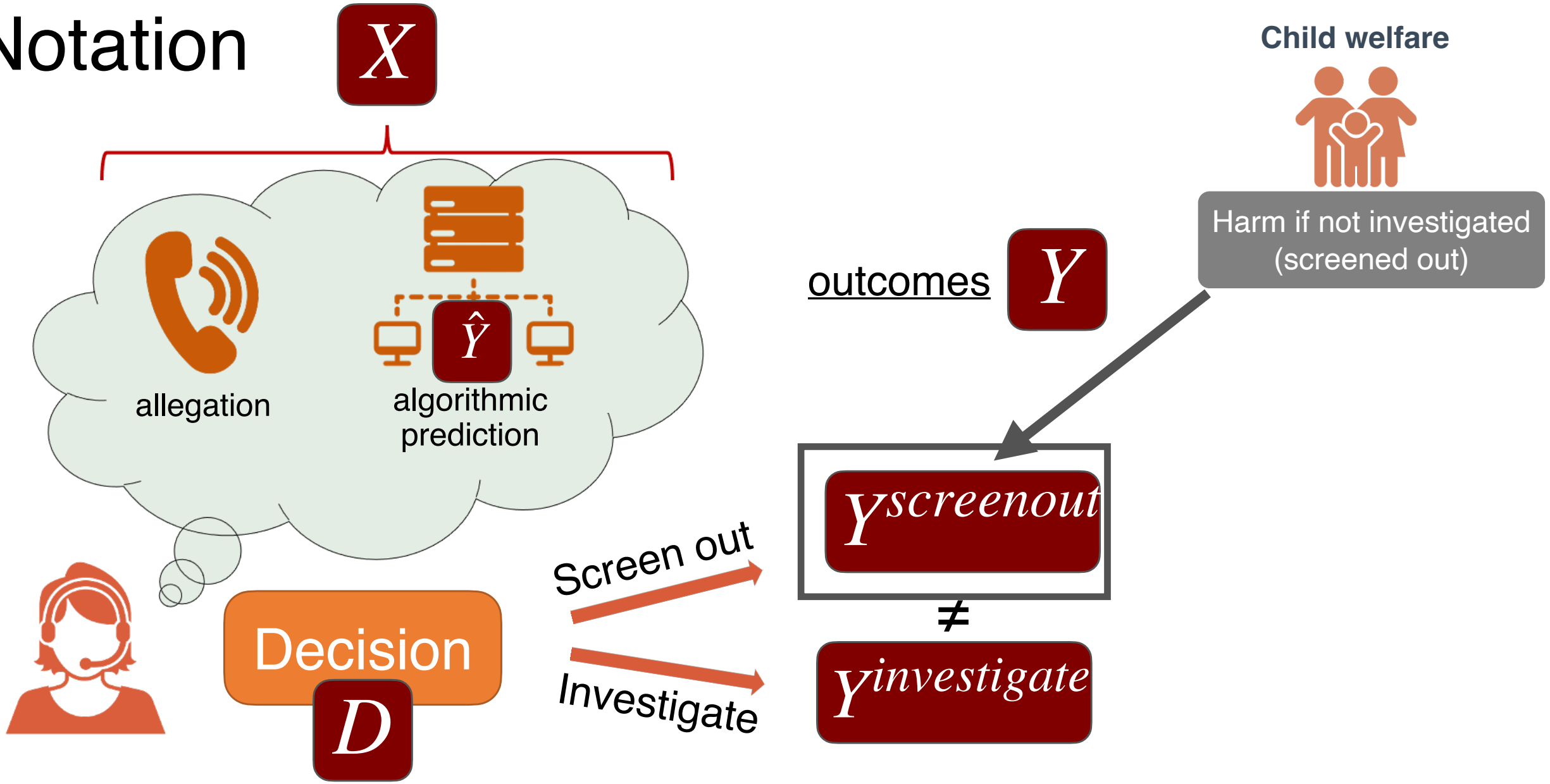
Notation


allegation
algorithmic prediction
$\hat{Y}$
Decision
D
Screen out
Investigate
outcomes Y
$Y^{screenout}$
≠
$Y^{investigate}$

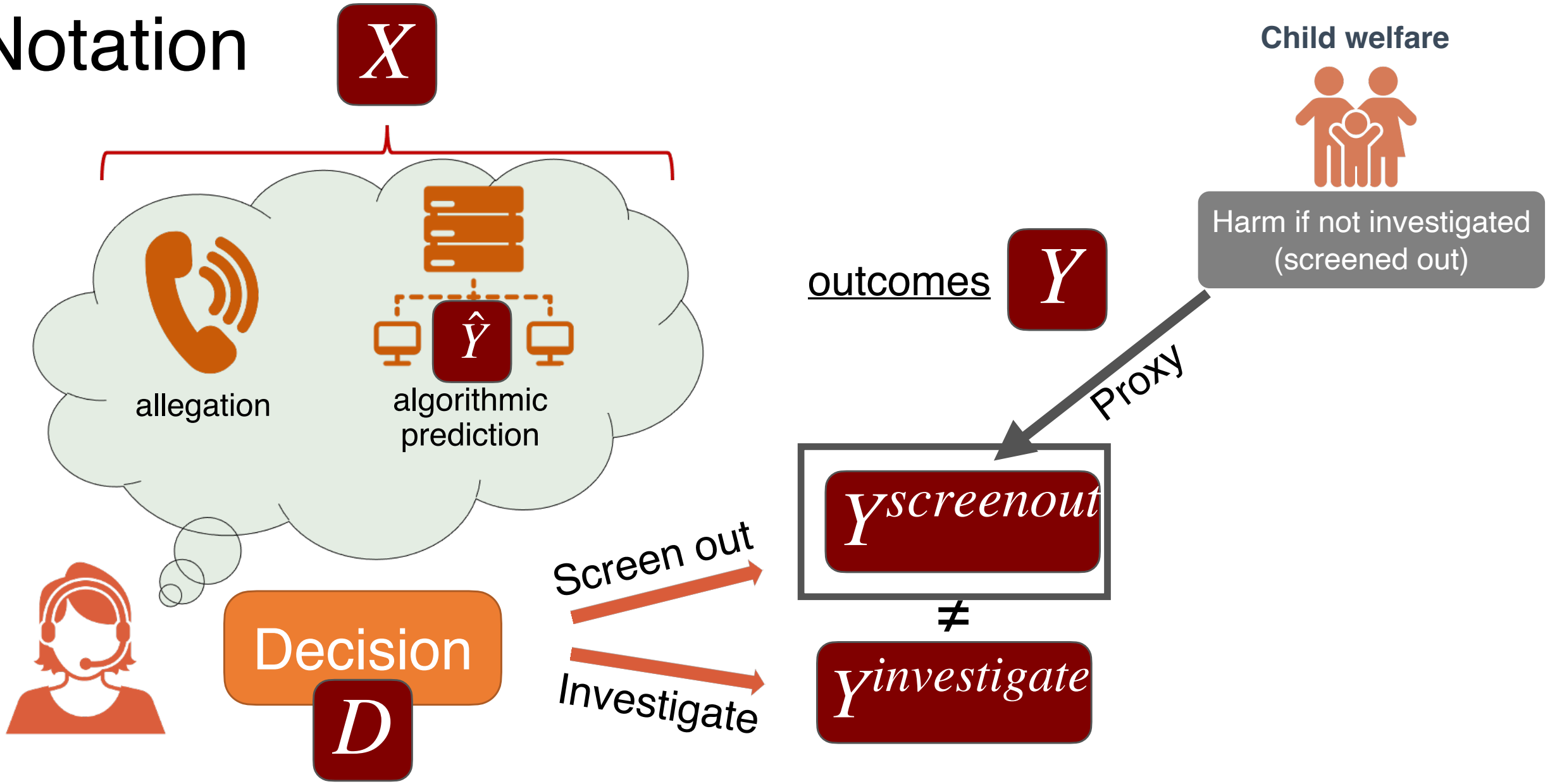
Notation


X
allegation
$\hat{Y}$
algorithmic prediction
Decision
D
Screen out
Investigate
outcomes Y
$Y^{screenout}$
≠
$Y^{investigate}$

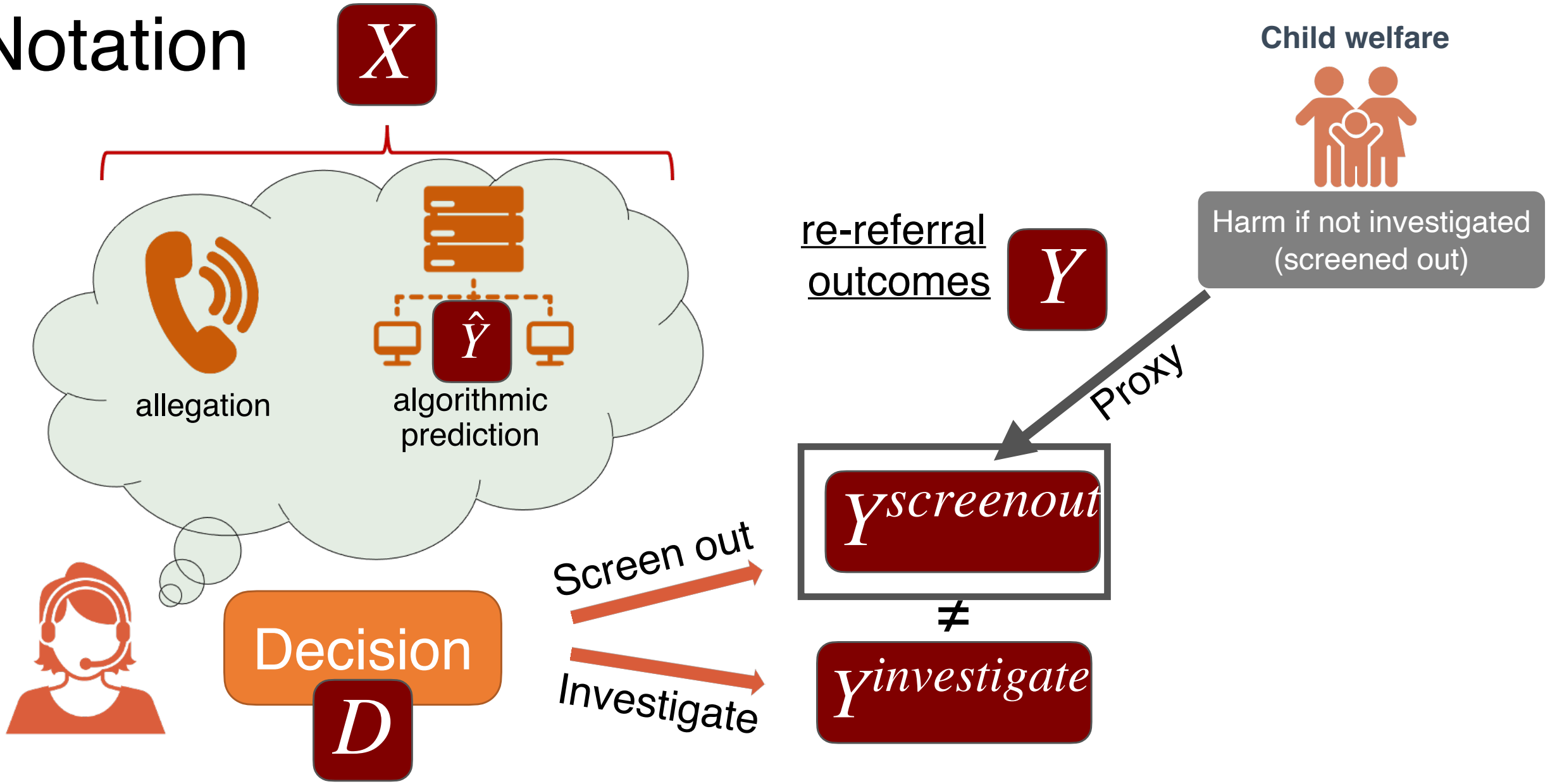
Notation


X
allegation
$\hat{Y}$
algorithmic prediction
Child welfare
Harm if not investigated (screened out)
outcomes Y
$Y^{screenout}$
≠
$Y^{investigate}$
Screen out
Investigate
Decision
D

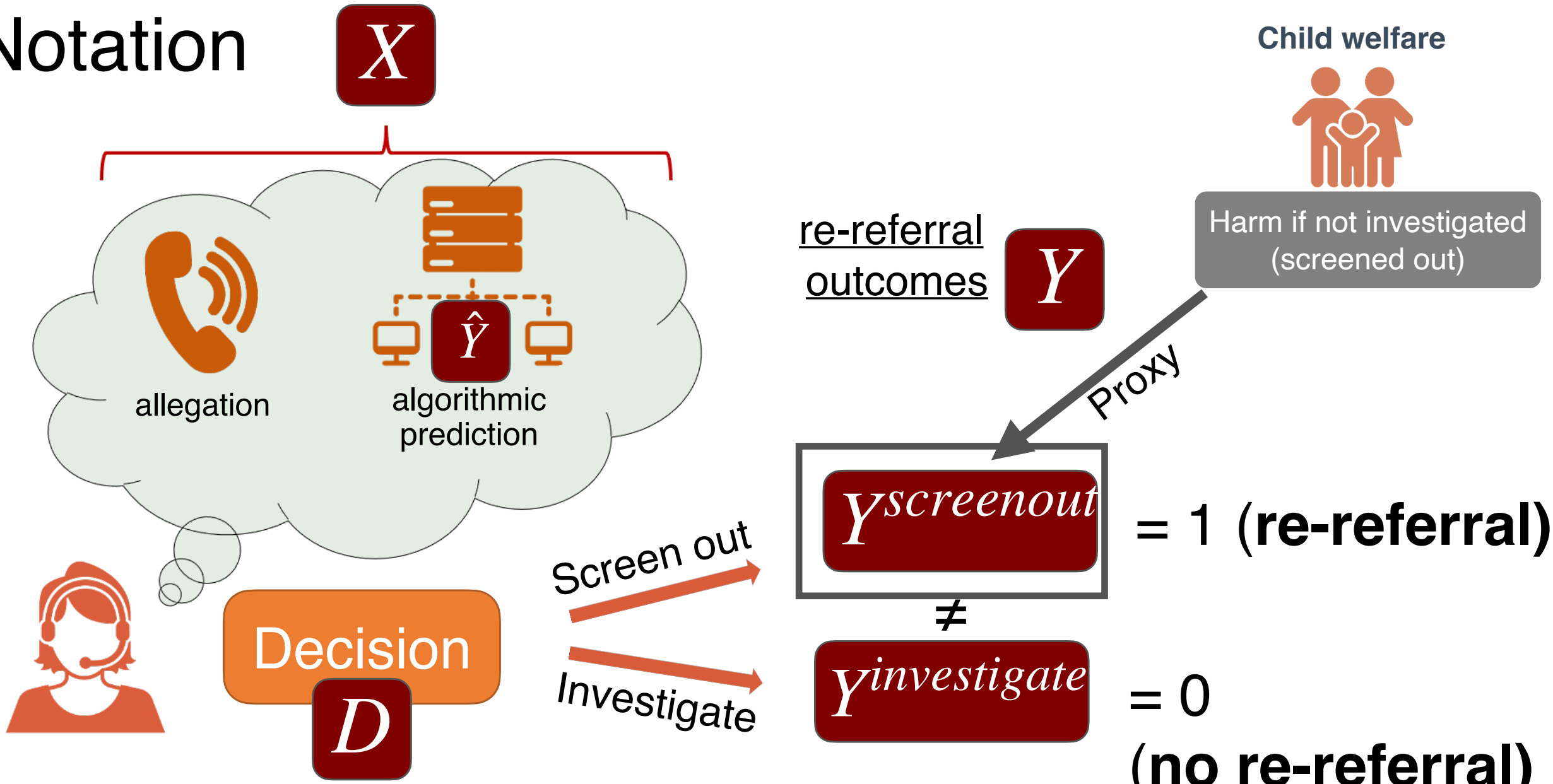
Notation


X
allegation
$\hat{Y}$
algorithmic prediction
Child welfare
Harm if not investigated (screened out)
outcomes Y
Proxy
$Y^{screenout}$
≠
$Y^{investigate}$
Screen out
Investigate
Decision
D

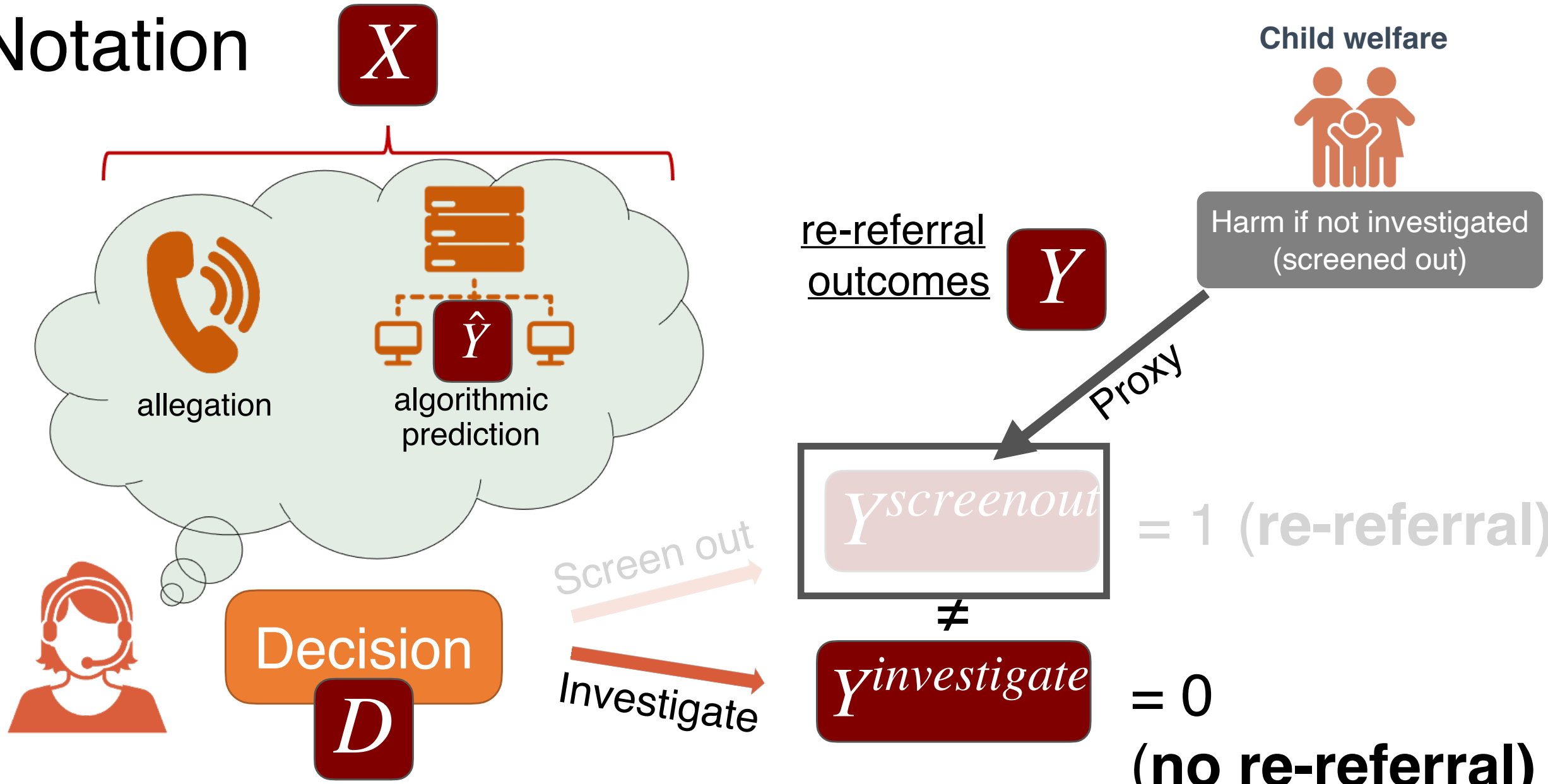
Notation


X
allegation
$\hat{Y}$
algorithmic prediction
Child welfare
Harm if not investigated (screened out)
re-referral outcomes
Y
Proxy
$Y^{screenout}$
≠
$Y^{investigate}$
Screen out
Investigate
Decision
D

Notation


X
allegation
$\hat{Y}$
algorithmic prediction
Decision
D
Screen out
Investigate
Child welfare
Harm if not investigated (screened out)
re-referral outcomes
Y
Proxy
$Y^{screenout}$
$\neq$
$Y^{investigate}$
= 1 (re-referral)
= 0 (no re-referral)

Notation


X
allegation
$\hat{Y}$
algorithmic prediction
Decision
D
Screen out
Investigate
Child welfare
Harm if not investigated (screened out)
re-referral outcomes
Y
Proxy
$Y^{screenout}$
= 1 (re-referral)
≠
$Y^{investigate}$
= 0
(no re-referral)

Research Question #1

How to assess performance when we have missing outcomes (bandit feedback)?

Standard approach

Evaluates on subsample with D = screen out

Schulam & Saria. Reliable decision support using counterfactual models. NeurIPS 2017.

Standard approach

Child welfare

Allegheny County

Evaluates on subsample with D = screen out

Covariates	**Decision** D	$Y^{screenout}$	**Algorithmic prediction** $\hat{Y}$
X_1	Screen out	0	0
X_2	Investigate		0
X_3	Screen out	0	0
X_4	Investigate		1

Schulam & Saria. Reliable decision support using counterfactual models. NeurIPS 2017.

Standard approach

Child welfare

Allegheny County

Evaluates on subsample with D = screen out

Covariates	**Decision** D	$Y^{screenout}$	**Algorithmic prediction** $\hat{Y}$
X_1	Screen out	0	0
X_2	Investigate		0
X_3	Screen out	0	0
X_4	Investigate		1

Schulam & Saria. Reliable decision support using counterfactual models. NeurIPS 2017.

Standard approach

Child welfare

Allegheny County

biased

Evaluates on subsample with D = screen out

Covariates	**Decision** D	$Y^{screenout}$	**Algorithmic prediction** $\hat{Y}$
X_1	Screen out	0	0
X_2	Investigate	1	0
X_3	Screen out	0	0
X_4	Investigate	1	1

Schulam & Saria. Reliable decision support using counterfactual models. NeurIPS 2017.

Standard approach

Child welfare

Allegheny County

biased

Evaluates on subsample with D = screen out

Covariates	**Decision** D	$Y^{screenout}$	**Algorithmic prediction** $\hat{Y}$
X_1	Screen out	0 ✓	0
X_2	Investigate	1	0
X_3	Screen out	0 ✓	0
X_4	Investigate	1	1

Schulam & Saria. Reliable decision support using counterfactual models. NeurIPS 2017.

Standard approach

Child welfare

Allegheny County

biased

Evaluates on subsample with D = screen out

Covariates	Decision D	$Y^{screenout}$	Algorithmic prediction $\hat{Y}$
X_1	Screen out	0 ✓	0
X_2	Investigate	1	0
X_3	Screen out	0 ✓	0
X_4	Investigate	1	1

100% accuracy on screened out

Schulam & Saria. Reliable decision support using counterfactual models. NeurIPS 2017.

Standard approach

Child welfare

Allegheny County

biased

Evaluates on subsample with D = screen out

Covariates	Decision D	$Y^{screenout}$	Algorithmic prediction $\hat{Y}$
X_1	Screen out	0 ✓	0
X_2	Investigate	1 ✗	0
X_3	Screen out	0 ✓	0
X_4	Investigate	1 ✓	1

100% accuracy on screened out

75% accuracy overall

Schulam & Saria. Reliable decision support using counterfactual models. NeurIPS 2017.

Evaluating four credit risk models

Lending

Default if loan approved

Evaluating four credit risk models

Evaluating four credit risk models

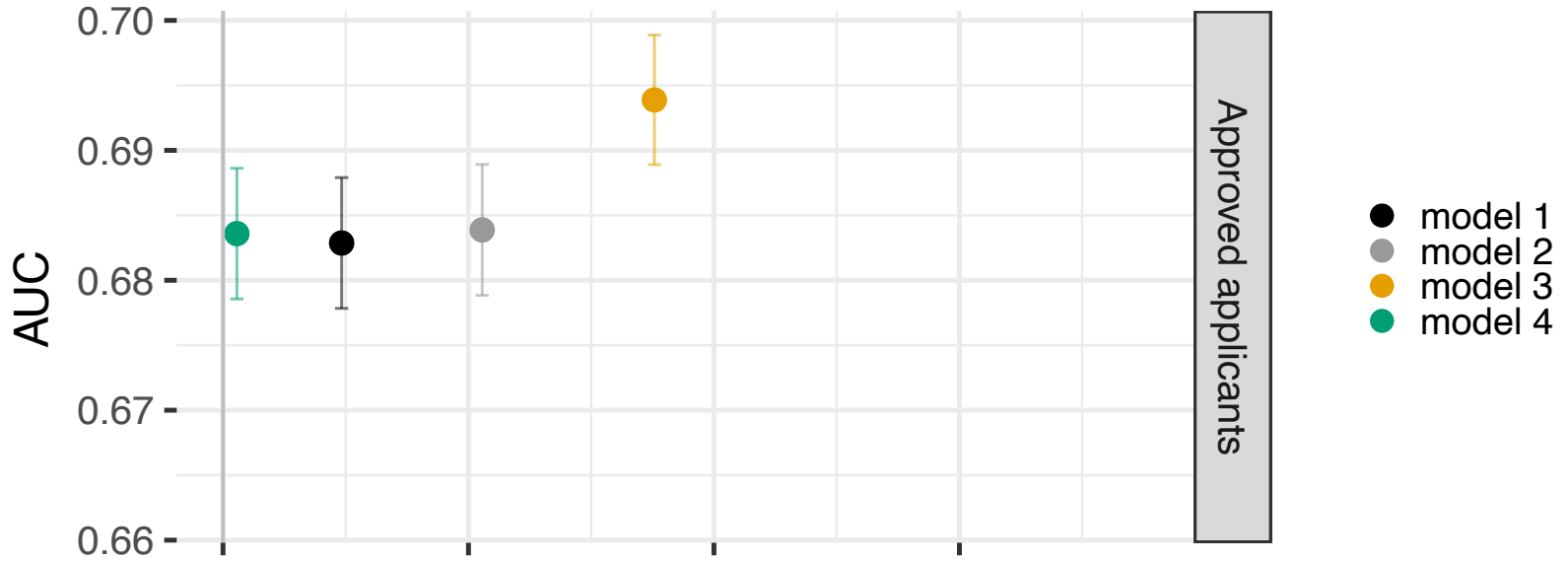
- model 1
- model 2
- model 3
- model 4

Evaluating four credit risk models

Lending

Default if loan approved

Commonwealth Bank of Australia

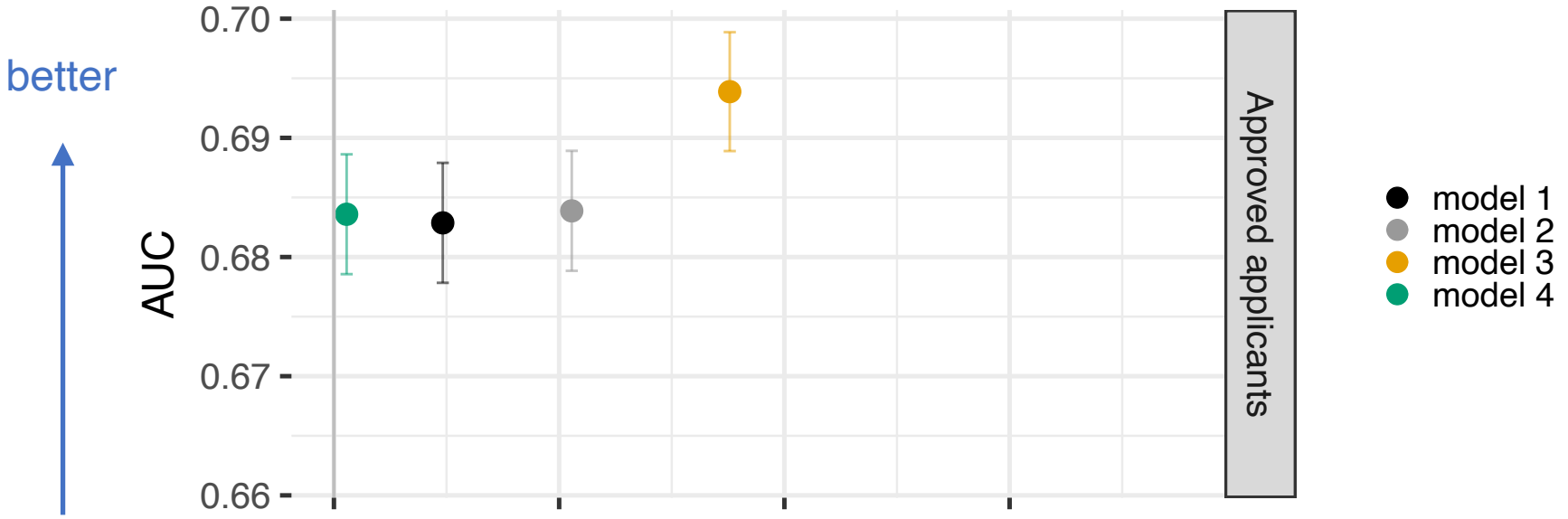


Evaluating four credit risk models

Lending

Default if loan approved

Commonwealth Bank of Australia

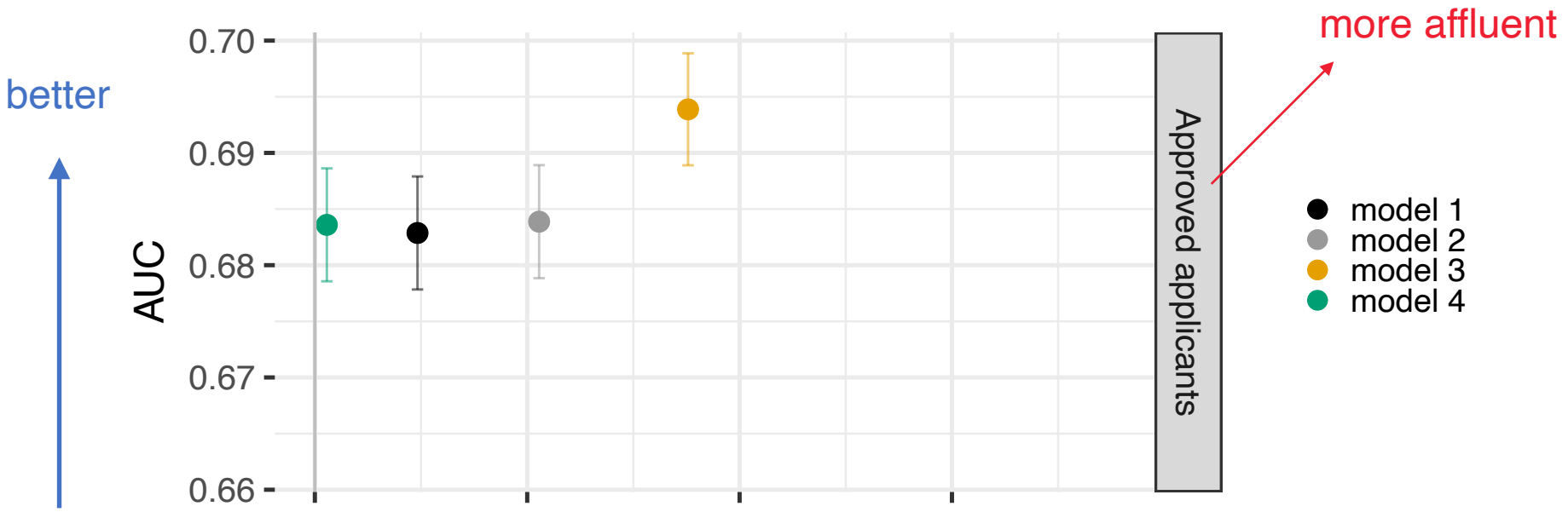


Evaluating four credit risk models

Lending

Default if loan approved

Commonwealth Bank of Australia



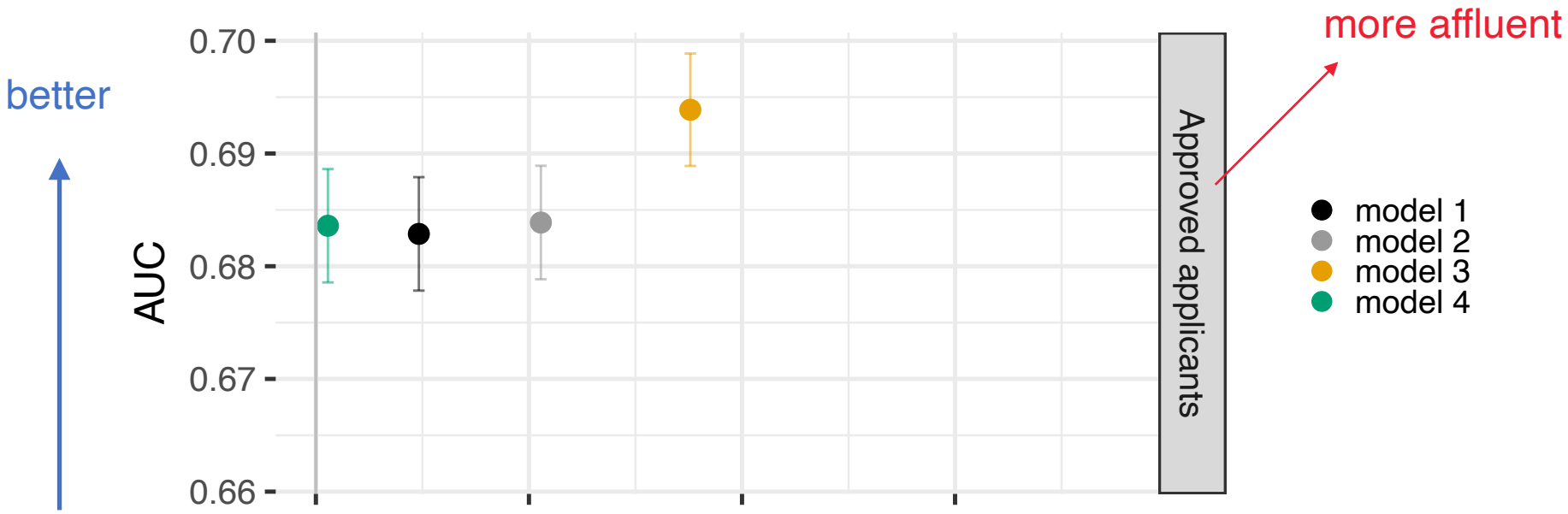
Evaluating four credit risk models

Semi-synthetic analysis

Commonwealth Bank of Australia

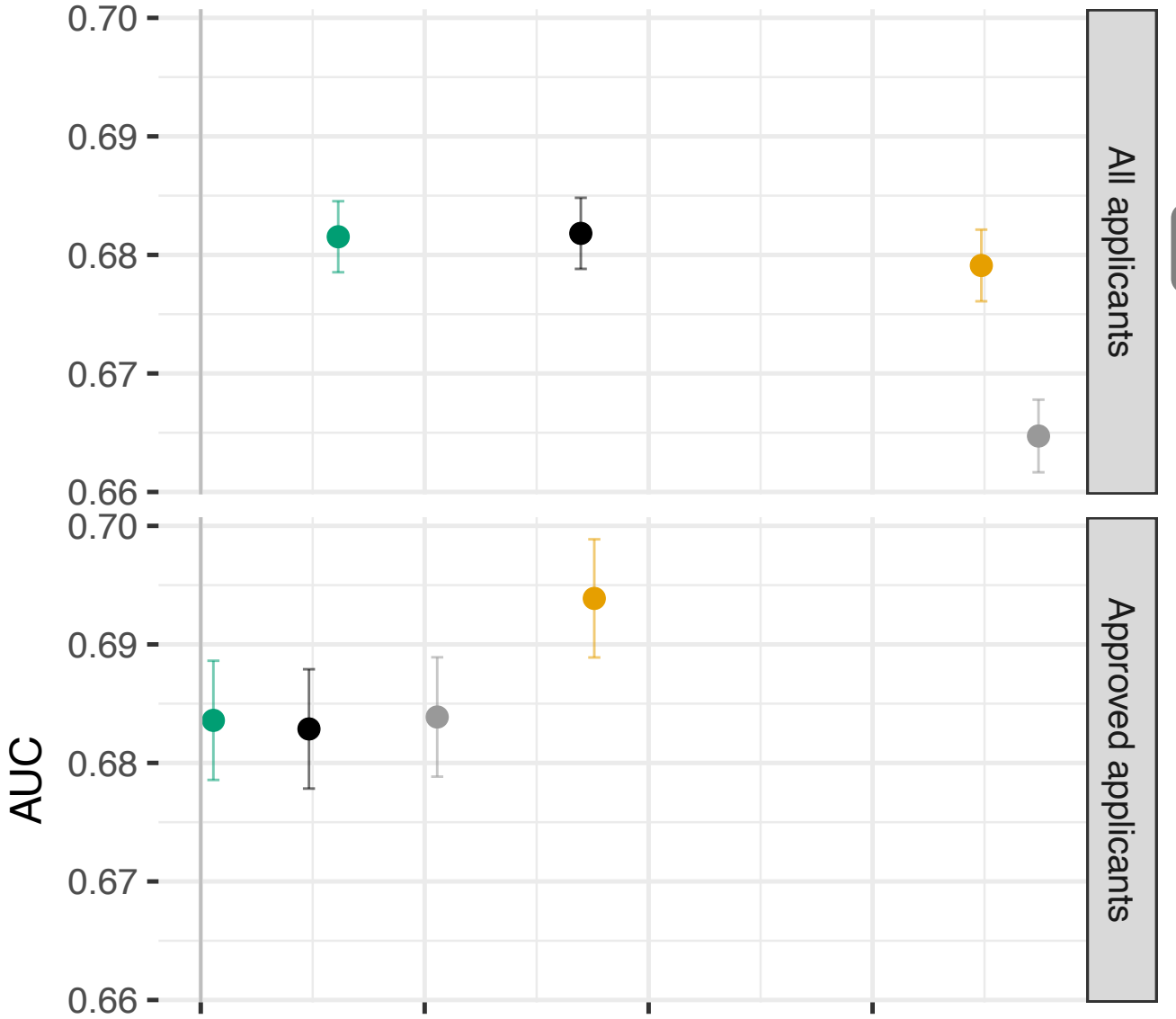
Lending

Default if loan approved



Evaluating four credit risk models [ICML 2021]

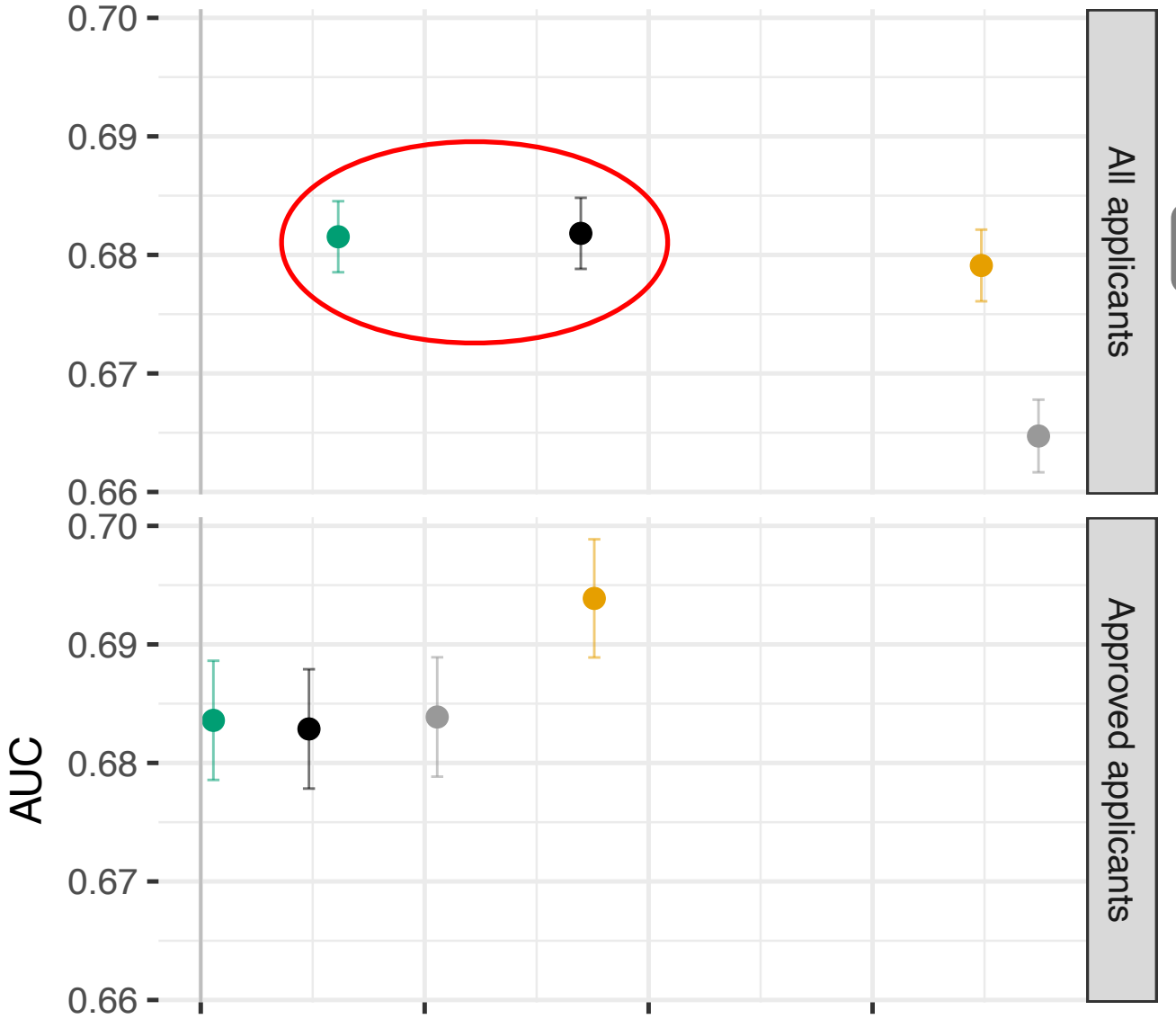
Semi-synthetic analysis



Default if loan approved

Evaluating four credit risk models [ICML 2021]

Semi-synthetic analysis

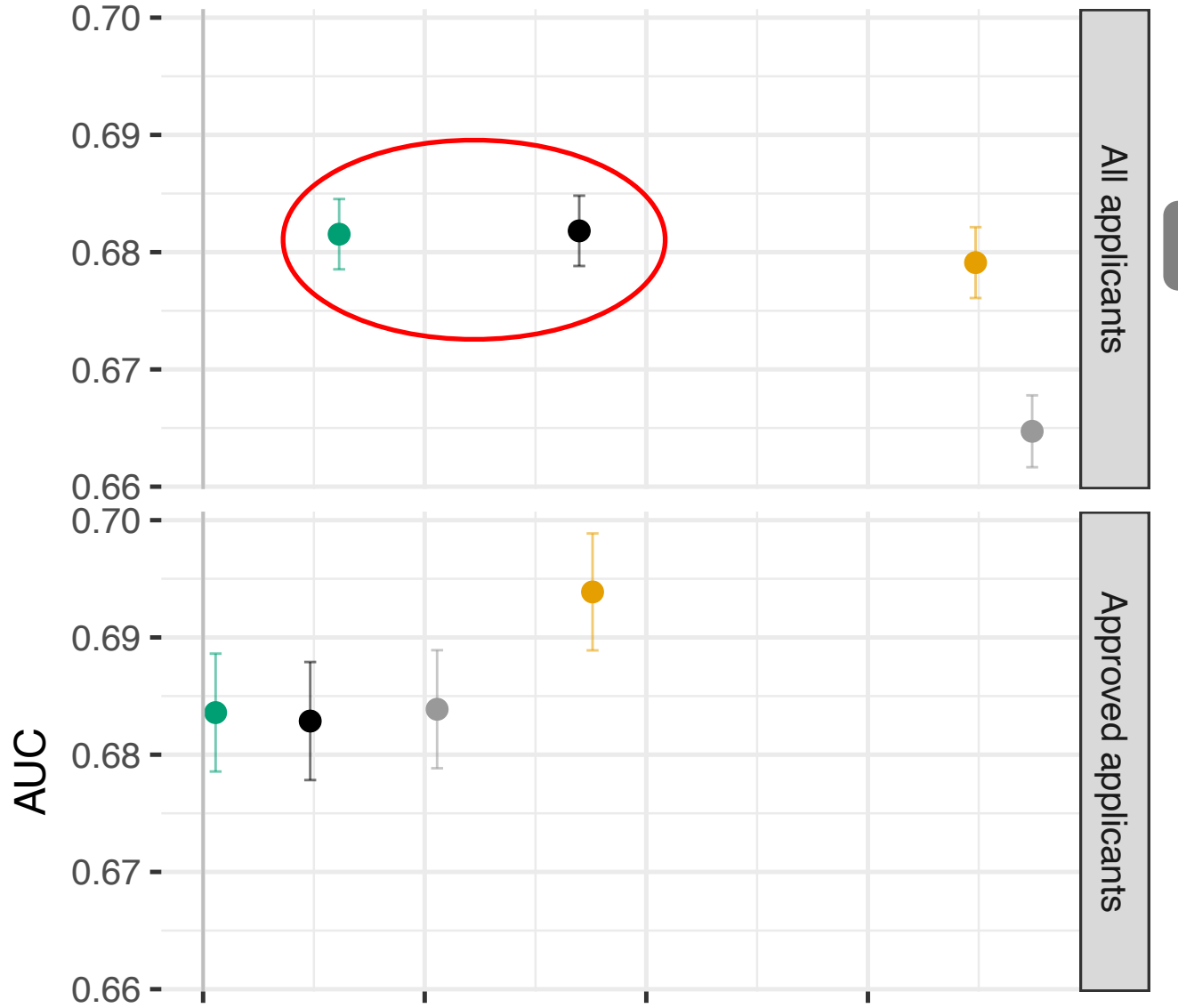


Lending

Default if loan approved

Evaluating four credit risk models [ICML 2021]

Semi-synthetic analysis



Lending

Default if loan approved

Problem

- Missing outcomes due to bandit feedback
- Selection bias may invalidate standard evaluations

Research Question #1

- How do we evaluate predictive performance when we have missing outcomes?

Contribution

- Counterfactual techniques to impute missing outcomes
- Valid evaluation under key conditions
- Doubly-robust techniques to achieve fast rates

Counterfactual evaluation

Counterfactual evaluation

Precision $P(Y^{screenout} = 1 \mid \hat{Y} = 1)$

Counterfactual evaluation

Precision $P(Y^{screenout} = 1 \mid \hat{Y} = 1)$

Covariates	Decision D	$Y^{screenout}$	Algorithmic prediction $\hat{Y}$
X_1	Screen out	0	0
X_2	Investigate		0
X_3	Screen out	0	0
X_4	Investigate		1
X_5	Screen out	0	1
X_6	Screen out	1	1

Counterfactual evaluation

Precision $P(Y^{screenout} = 1 \mid \hat{Y} = 1)$

Covariates	Decision D	$Y^{screenout}$	Algorithmic prediction $\hat{Y}$
X_1	Screen out	0	0
X_2	Investigate		0
X_3	Screen out	0	0
X_4	Investigate		1
X_5	Screen out	0	1
X_6	Screen out	1	1

Counterfactual evaluation

Precision $P(Y^{screenout} = 1 \mid \hat{Y} = 1)$

Covariates	Decision D	$Y^{screenout}$	Algorithmic prediction $\hat{Y}$
X_1	Screen out	0	0
X_2	Investigate		0
X_3	Screen out	0	0
X_4	Investigate		1
X_5	Screen out	0	1
X_6	Screen out	1	1

50% precision
on screened out

Counterfactual evaluation [FAccT 2020]

Counterfactual evaluation [FAccT 2020]

Precision $P(Y^d = 1 \mid \hat{Y} = 1)$

Counterfactual evaluation [FAccT 2020]

d = screenout

Precision $P(Y^d = 1 \mid \hat{Y} = 1)$

Counterfactual evaluation [FAccT 2020]

d = screenout

Precision $P(Y^d = 1 \mid \hat{Y} = 1) = \mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

where $\mu(x) := P(Y^d = 1 \mid X = x)$

Counterfactual evaluation [FAccT 2020]

d = screenout

Precision $P(Y^d = 1 \mid \hat{Y} = 1) = \mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

where $\mu(x) := P(Y^d = 1 \mid X = x)$ outcome regression

"nuisance" function

Counterfactual evaluation [FAccT 2020]

d = screenout

Precision $P(Y^d = 1 \mid \hat{Y} = 1) = \mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

where $\mu(x) := P(Y^d = 1 \mid X = x)$ outcome regression

"nuisance" function

Identifying assumptions

Counterfactual evaluation [FAccT 2020]

d = screenout

Precision $P(Y^d = 1 \mid \hat{Y} = 1) = \mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

where $\mu(x) := P(Y^d = 1 \mid X = x)$ outcome regression

"nuisance" function

Identifying assumptions

1. No unmeasured confounding $D \perp Y^d \mid X$

Counterfactual evaluation [FAccT 2020]

d = screenout

Precision $P(Y^d = 1 \mid \hat{Y} = 1) = \mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

where $\mu(x) := P(Y^d = 1 \mid X = x)$

Identifying assumptions

1. No unmeasured confounding $D \perp Y^d \mid X$

Counterfactual evaluation [FAccT 2020]

d = screenout

Precision $P(Y^d = 1 \mid \hat{Y} = 1) = \mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

where $\mu(x) := P(Y^d = 1 \mid X = x)$

Identifying assumptions

1. No unmeasured confounding $D \perp Y^d \mid X$

Counterfactual evaluation [FAccT 2020]

d = screenout

Precision $P(Y^d = 1 \mid \hat{Y} = 1) = \mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

where $\mu(x) := P(Y^d = 1 \mid X = x)$

Identifying assumptions

1. No unmeasured confounding $D \perp Y^d \mid X$
2. Overlap $P(\pi(X) > 0) = 1$

where $\pi(x) = P(D = d \mid X = x)$

Counterfactual evaluation [FAccT 2020]

d = screenout

Precision $P(Y^d = 1 \mid \hat{Y} = 1) = \mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

where $\mu(x) := P(Y^d = 1 \mid X = x)$

Identifying assumptions

1. No unmeasured confounding $D \perp Y^d \mid X$
2. Overlap $P(\pi(X) > 0) = 1$

where $\pi(x) = P(D = d \mid X = x)$

$$\Longrightarrow \qquad \mu(x) = P(Y = 1 \mid D = d, X = x)$$

Counterfactual evaluation [FAccT 2020]

d = screenout

Precision $P(Y^d = 1 \mid \hat{Y} = 1) = \mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

where $\mu(x) := P(Y^d = 1 \mid X = x)$

Identifying assumptions

1. No unmeasured confounding $D \perp Y^d \mid X$
2. Overlap $P(\pi(X) > 0) = 1$

where $\pi(x) = P(D = d \mid X = x)$

$$\implies \quad \mu(x) = P(Y = 1 \mid D = d, X = x)$$

Key result: doubly-robust estimators for counterfactual evaluation

Estimator for Precision $\mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

Key result: doubly-robust estimators for counterfactual evaluation

Estimator for Precision $\mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

Estimate $\hat{\mu}(X) = \hat{\mathbb{P}}(Y = 1 \mid D = d, X)$ and $\hat{\pi}(x) = \hat{\mathbb{P}}(D = d \mid X = x)$

Key result: doubly-robust estimators for counterfactual evaluation

Estimator for Precision $\mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

Estimate $\hat{\mu}(X) = \hat{\mathbb{P}}(Y = 1 \mid D = d, X)$ and $\hat{\pi}(x) = \hat{\mathbb{P}}(D = d \mid X = x)$

Let n be # with $\hat{Y} = 1$. Take average over these:

Key result: doubly-robust estimators for counterfactual evaluation

Estimator for Precision $\mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

Estimate $\hat{\mu}(X) = \hat{\mathbb{P}}(Y = 1 \mid D = d, X)$ and $\hat{\pi}(x) = \hat{\mathbb{P}}(D = d \mid X = x)$

Let n be # with $\hat{Y} = 1$. Take average over these:

$$\frac{1}{n} \sum_{i:\hat{Y}_i=1}^{n} \hat{\mu}(X_i)$$

Key result: doubly-robust estimators for counterfactual evaluation

Estimator for Precision $\mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

Estimate $\hat{\mu}(X) = \hat{\mathbb{P}}(Y = 1 \mid D = d, X)$ and $\hat{\pi}(x) = \hat{\mathbb{P}}(D = d \mid X = x)$

Let n be # with $\hat{Y} = 1$. Take average over these:

$$\frac{1}{n} \sum_{i:\hat{Y}_i=1}^{n} \hat{\mu}(X_i) + \frac{D_i}{\hat{\pi}(X_i)}(Y_i - \hat{\mu}(X_i))$$

Key result: doubly-robust estimators for counterfactual evaluation

Estimator for Precision $\mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

Estimate $\hat{\mu}(X) = \hat{\mathbb{P}}(Y = 1 \mid D = d, X)$ and $\hat{\pi}(x) = \hat{\mathbb{P}}(D = d \mid X = x)$

Let n be # with $\hat{Y} = 1$. Take average over these:

$$\frac{1}{n} \sum_{i:\hat{Y}_i=1}^{n} \hat{\mu}(X_i) + \frac{D_i}{\hat{\pi}(X_i)}(Y_i - \hat{\mu}(X_i))$$

Bias-correction term

Key result: doubly-robust estimators for counterfactual evaluation

Estimator for Precision $\mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

Estimate $\hat{\mu}(X) = \hat{\mathbb{P}}(Y = 1 \mid D = d, X)$ and $\hat{\pi}(x) = \hat{\mathbb{P}}(D = d \mid X = x)$

Let n be # with $\hat{Y} = 1$. Take average over these:

$$\frac{1}{n} \sum_{i:\hat{Y}_i=1}^{n} \hat{\mu}(X_i) + \frac{D_i}{\hat{\pi}(X_i)}(Y_i - \hat{\mu}(X_i))$$

Bias-correction term

Key result: doubly-robust estimators for counterfactual evaluation

Estimator for Precision $\mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

Estimate $\hat{\mu}(X) = \hat{\mathbb{P}}(Y = 1 \mid D = d, X)$ and $\hat{\pi}(x) = \hat{\mathbb{P}}(D = d \mid X = x)$

Let n be # with $\hat{Y} = 1$. Take average over these:

$$\frac{1}{n} \sum_{i:\hat{Y}_i=1}^{n} \hat{\mu}(X_i) + \frac{D_i}{\hat{\pi}(X_i)}(Y_i - \hat{\mu}(X_i))$$

→ Bias-correction term

$\sqrt{n}$-consistent & asymptotically normal under sample splitting and $\sqrt[4]{n}$ convergence in μ and π estimation

Key result: doubly-robust estimators for counterfactual evaluation

Estimator for Precision $\mathbb{E}[\mu(X) \mid \hat{Y} = 1]$

Estimate $\hat{\mu}(X) = \hat{\mathbb{P}}(Y = 1 \mid D = d, X)$ and $\hat{\pi}(x) = \hat{\mathbb{P}}(D = d \mid X = x)$

Let n be # with $\hat{Y} = 1$. Take average over these:

$$\frac{1}{n} \sum_{i:\hat{Y}_i=1}^{n} \hat{\mu}(X_i) + \frac{D_i}{\hat{\pi}(X_i)}(Y_i - \hat{\mu}(X_i))$$

→ Bias-correction term

$\sqrt{n}$-consistent & asymptotically normal under sample splitting and $\sqrt[4]{n}$ convergence in μ and π estimation

➡ Estimate μ and π flexibly & get valid CI on precision estimate

Doubly-robust estimate of precision

Covariates	Decision D	$Y^{screenout}$	Pseudo outcomes	Algorithmic prediction $\hat{Y}$
X_1	Screen out	0		0
X_2	Investigate			0
X_3	Screen out	0		0
X_4	Investigate		$\hat{\mu}(X_4)$	1
X_5	Screen out	0	$\hat{\mu}(X_5) - \frac{\hat{\mu}(X_5)}{\hat{\pi}(X_5)}$	1
X_6	Screen out	1	$\hat{\mu}(X_6) + \frac{1 - \hat{\mu}(X_6)}{\hat{\pi}(X_6)}$	1

Doubly-robust estimate of precision

Covariates	Decision D	$Y^{screenout}$	Pseudo outcomes	Algorithmic prediction $\hat{Y}$
X_1	Screen out	0		0
X_2	Investigate			0
X_3	Screen out	0		0
X_4	Investigate		$\hat{\mu}(X_4)$	1
X_5	Screen out	0	$\hat{\mu}(X_5) - \frac{\hat{\mu}(X_5)}{\hat{\pi}(X_5)}$	1
X_6	Screen out	1	$\hat{\mu}(X_6) + \frac{1 - \hat{\mu}(X_6)}{\hat{\pi}(X_6)}$	1

Average to get precision estimate

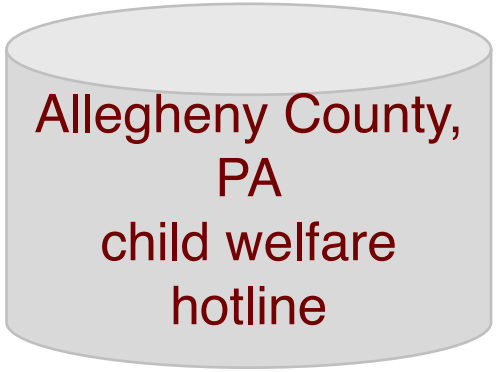
Doubly-robust estimate of precision

Covariates	Decision D	$Y^{screenout}$	Pseudo outcomes	Algorithmic prediction $\hat{Y}$
X_1	Screen out	0		0
X_2	Investigate			0
X_3	Screen out	0		0
X_4	Investigate		$\hat{\mu}(X_4)$	1
X_5	Screen out	0	$\hat{\mu}(X_5) - \frac{\hat{\mu}(X_5)}{\hat{\pi}(X_5)}$	1
X_6	Screen out	1	$\hat{\mu}(X_6) + \frac{1 - \hat{\mu}(X_6)}{\hat{\pi}(X_6)}$	1

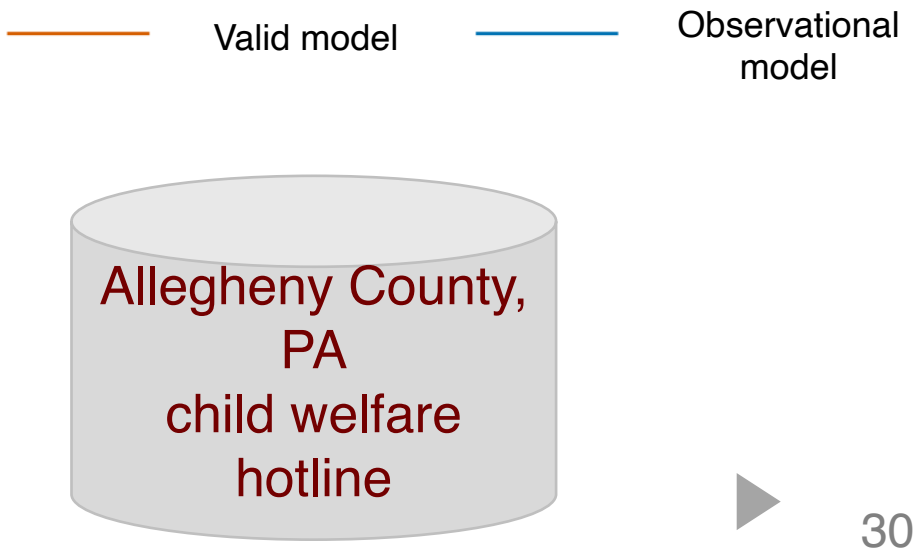
Average to get precision estimate

Doubly-robust estimators for classification metrics in **Counterfactual risk assessments, evaluation, and fairness** (FAccT 2020) and for regression metrics in **Counterfactual predictions under runtime confounding** (NeurIPS 2020)

Results



Results


Valid model
Observational model
Allegheny County, PA
child welfare hotline

Results

Covariate	$Y^{screenout}$	$Y^{investigate}$	Decision *D*	Observed Y
X_1	0	0	Screen out	0
X_2	1	0	Investigate	0
X_3	0	0	Screen out	0

Allegheny County, PA
child welfare hotline

Results

Covariate	$Y^{screenout}$	$Y^{investigate}$	Decision *D*	Observed Y
X_1	0	0	Screen out	0
X_2	1	0	Investigate	0
X_3	0	0	Screen out	0

Valid model

Observational model

Allegheny County, PA
child welfare hotline

Results

Covariate	$Y^{screenout}$	$Y^{investigate}$	Decision *D*	Observed Y
X_1	0	0	Screen out	0
X_2	1	0	Investigate	0
X_3	0	0	Screen out	0

Valid model

Observational model

Allegheny County, PA
child welfare hotline

Results

Covariate	$Y^{screenout}$	$Y^{investigate}$	Decision *D*	Observed Y
X_1	0	0	Screen out	0
X_2	1	0	Investigate	0
X_3	0	0	Screen out	0

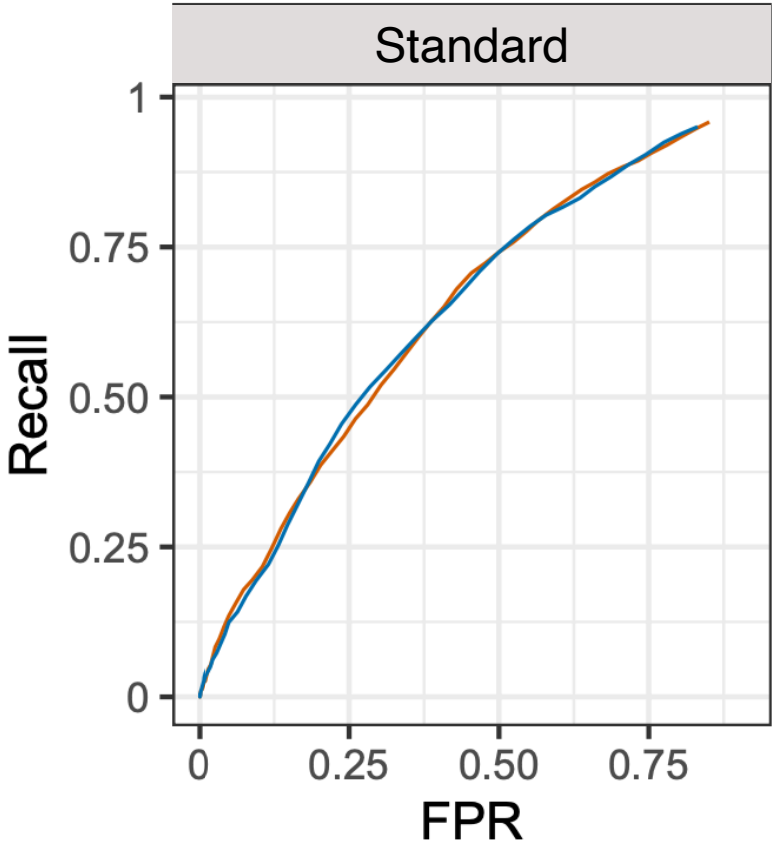
Valid model

Observational model

Not valid!

Allegheny County, PA
child welfare hotline

Results



Covariate	$Y^{screenout}$	$Y^{investigate}$	Decision *D*	Observed Y
X_1	0	0	Screen out	0
X_2	1	0	Investigate	0
X_3	0	0	Screen out	0

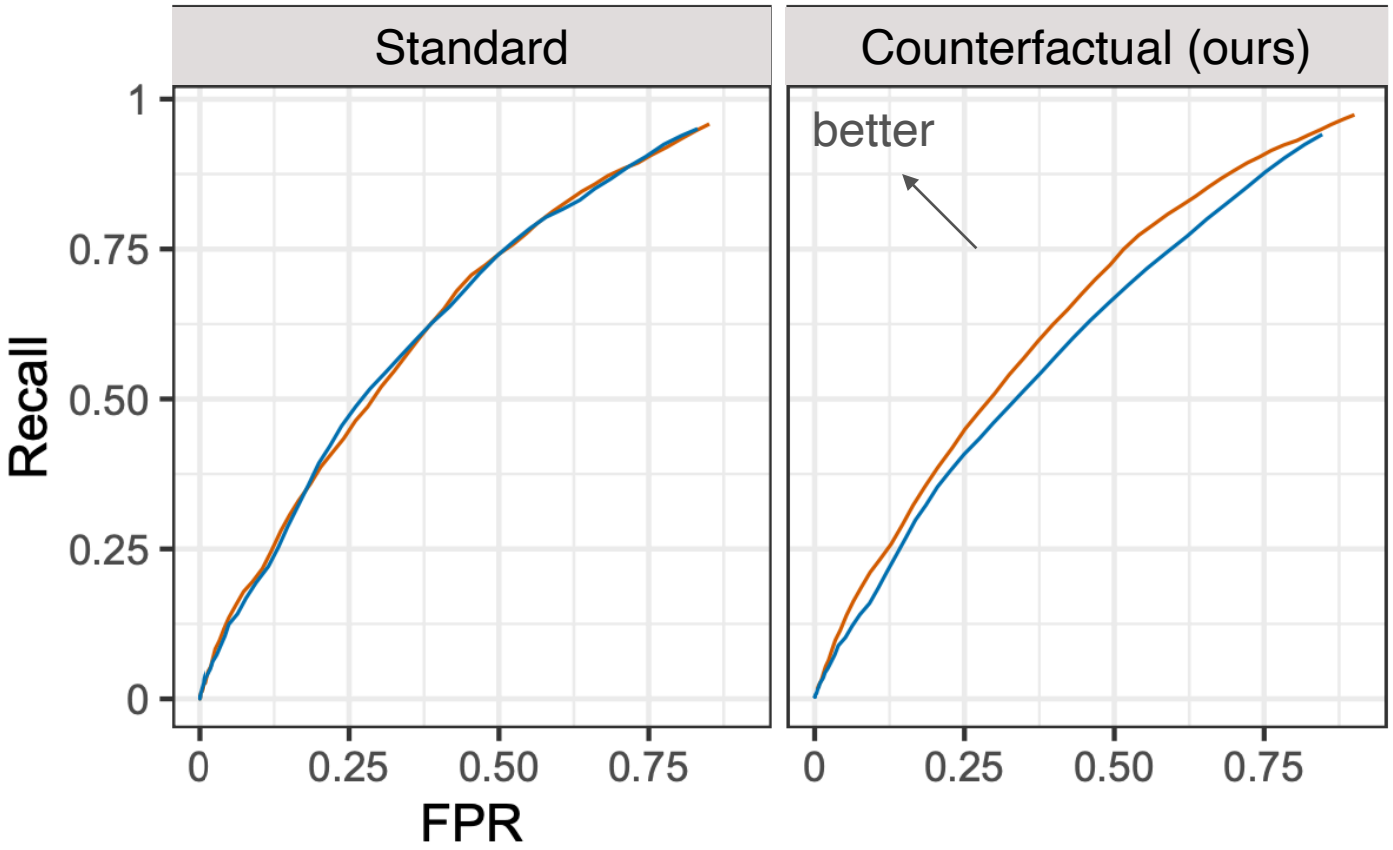
Valid model

Observational model

Not valid!

Allegheny County, PA
child welfare hotline

Results



Covariate	$Y^{screenout}$	$Y^{investigate}$	Decision *D*	Observed Y
X_1	0	0	Screen out	0
X_2	1	0	Investigate	0
X_3	0	0	Screen out	0

Valid model

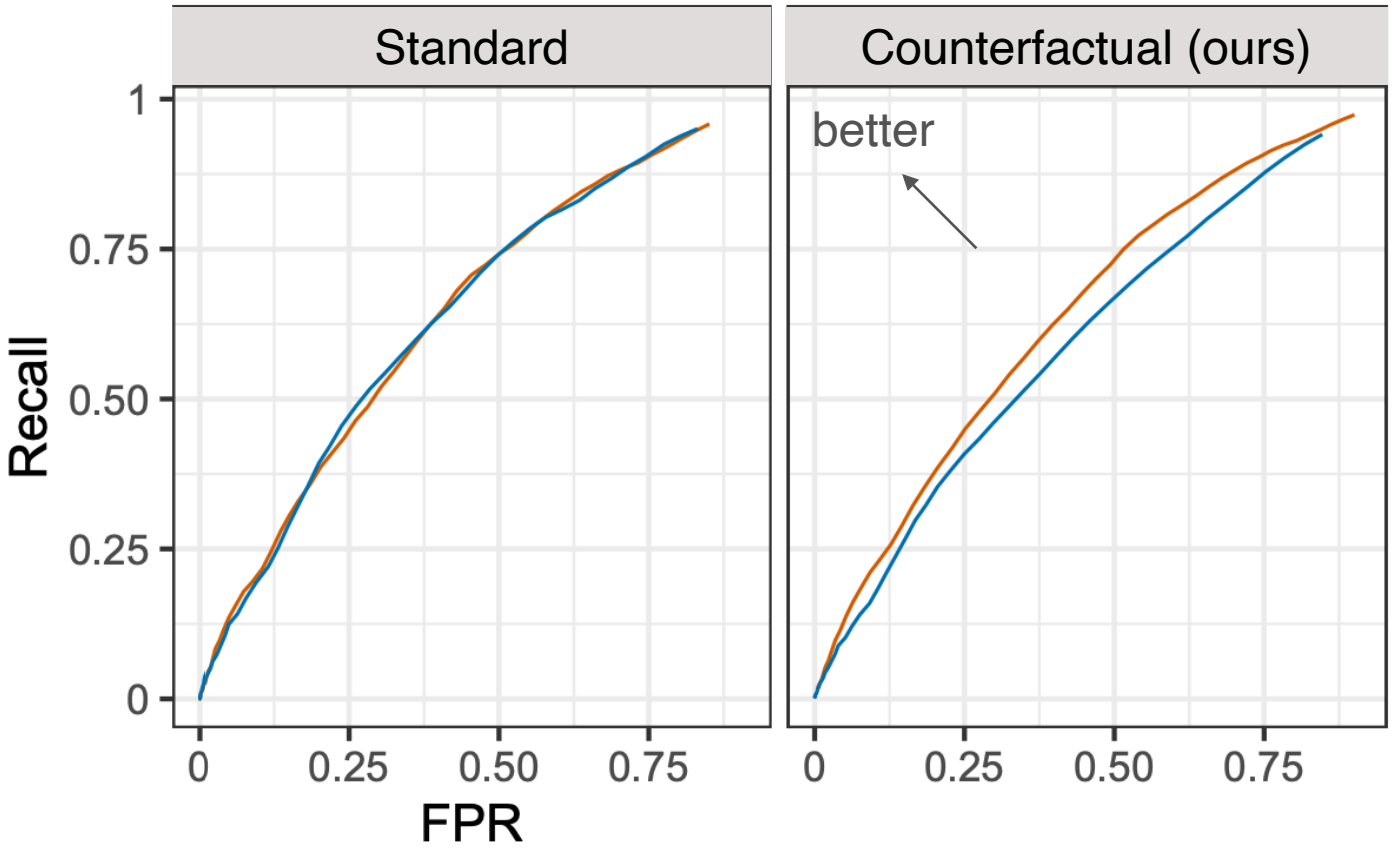
Observational model

Not valid!

Allegheny County, PA
child welfare hotline

Results

Counterfactual eval distinguishes
valid model from invalid model



Covariate	$Y^{screenout}$	$Y^{investigate}$	Decision *D*	Observed Y
X_1	0	0	Screen out	0
X_2	1	0	Investigate	0
X_3	0	0	Screen out	0

Valid model

Observational model

Not valid!

Allegheny County, PA
child welfare hotline

Problem

- Missing outcomes due to bandit feedback

Research Question #1

- How do we evaluate predictive performance when we have missing outcomes?

Contribution

- Counterfactual evaluation framework that imputes missing outcomes
- Doubly-robust techniques to achieve fast rates
- Valid confidence intervals & flexible ML methods

Problem

- Missing outcomes due to bandit feedback

Research Question #1

- How do we evaluate predictive performance when we have missing outcomes?

Contribution

- Counterfactual evaluation framework that imputes missing outcomes
- Doubly-robust techniques to achieve fast rates
- Valid confidence intervals & flexible ML methods

➡ Applicable to fairness assessments

Equity

Common approach to algorithmic fairness

Common approach to algorithmic fairness

Parity in performance metric btw demographic groups, e.g.,

Common approach to algorithmic fairness

Parity in performance metric btw demographic groups, e.g.,

- Accuracy is same for men & women

Common approach to algorithmic fairness

Parity in performance metric btw demographic groups, e.g.,

- Accuracy is same for men & women
- Precision is same across race

Common approach to algorithmic fairness

Parity in performance metric btw demographic groups, e.g.,

- Accuracy is same for men & women
- Precision is same across race

Common approach to algorithmic fairness

Parity in performance metric btw demographic groups, e.g.,

- Accuracy is same for men & women
- Precision is same across race

- Possible to achieve parity in training data but still have disparities in the target population due to selection bias

Common approach to algorithmic fairness

Parity in performance metric btw demographic groups, e.g.,

- Accuracy is same for men & women
- Precision is same across race

- Possible to achieve parity in training data but still have disparities in the target population due to selection bias

Kallus & Zhou ICML 2018

Common approach to algorithmic fairness

Parity in performance metric btw demographic groups, e.g.,

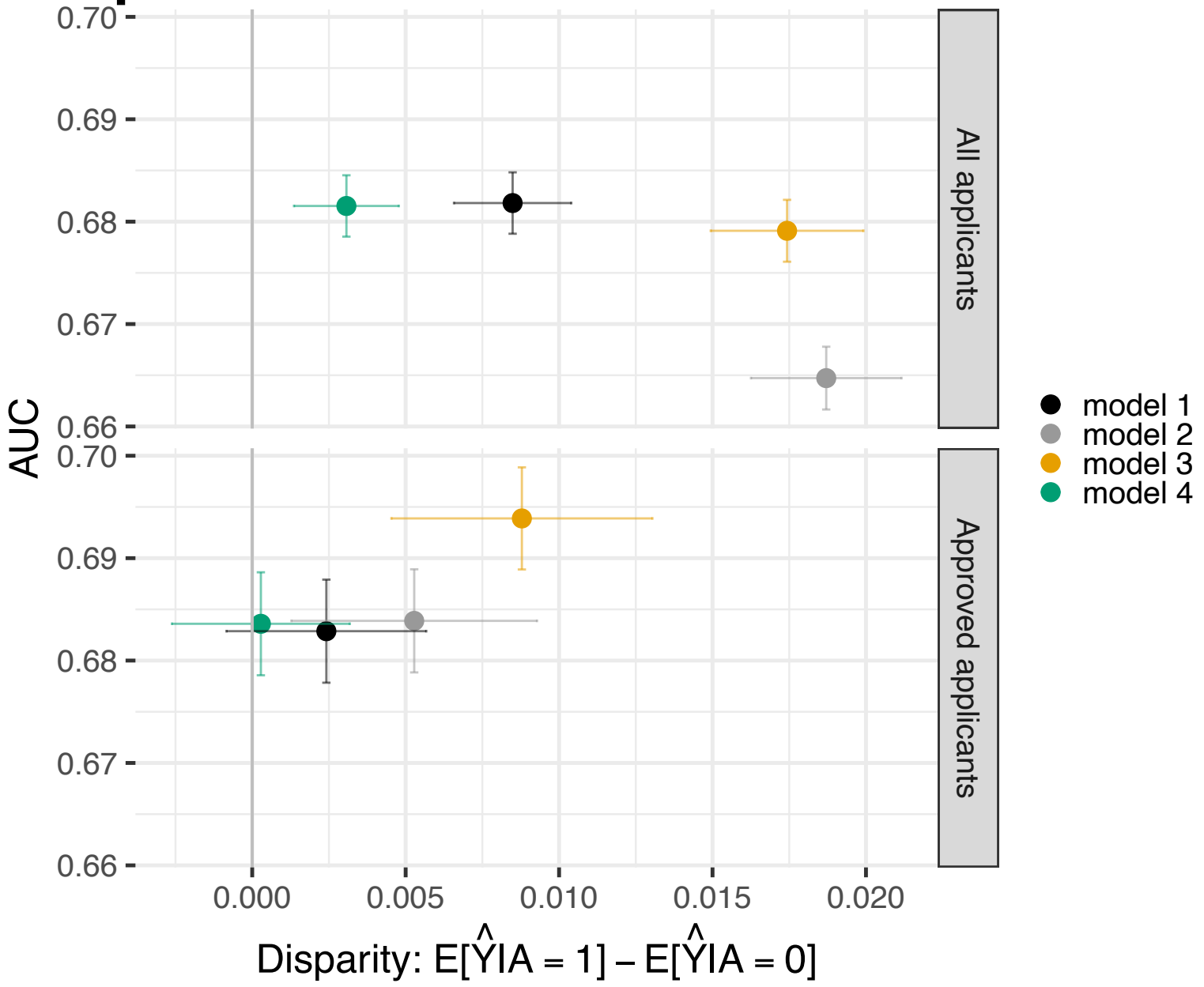
- Accuracy is same for men & women
- Precision is same across race

- Possible to achieve parity in training data but still have disparities in the target population due to selection bias

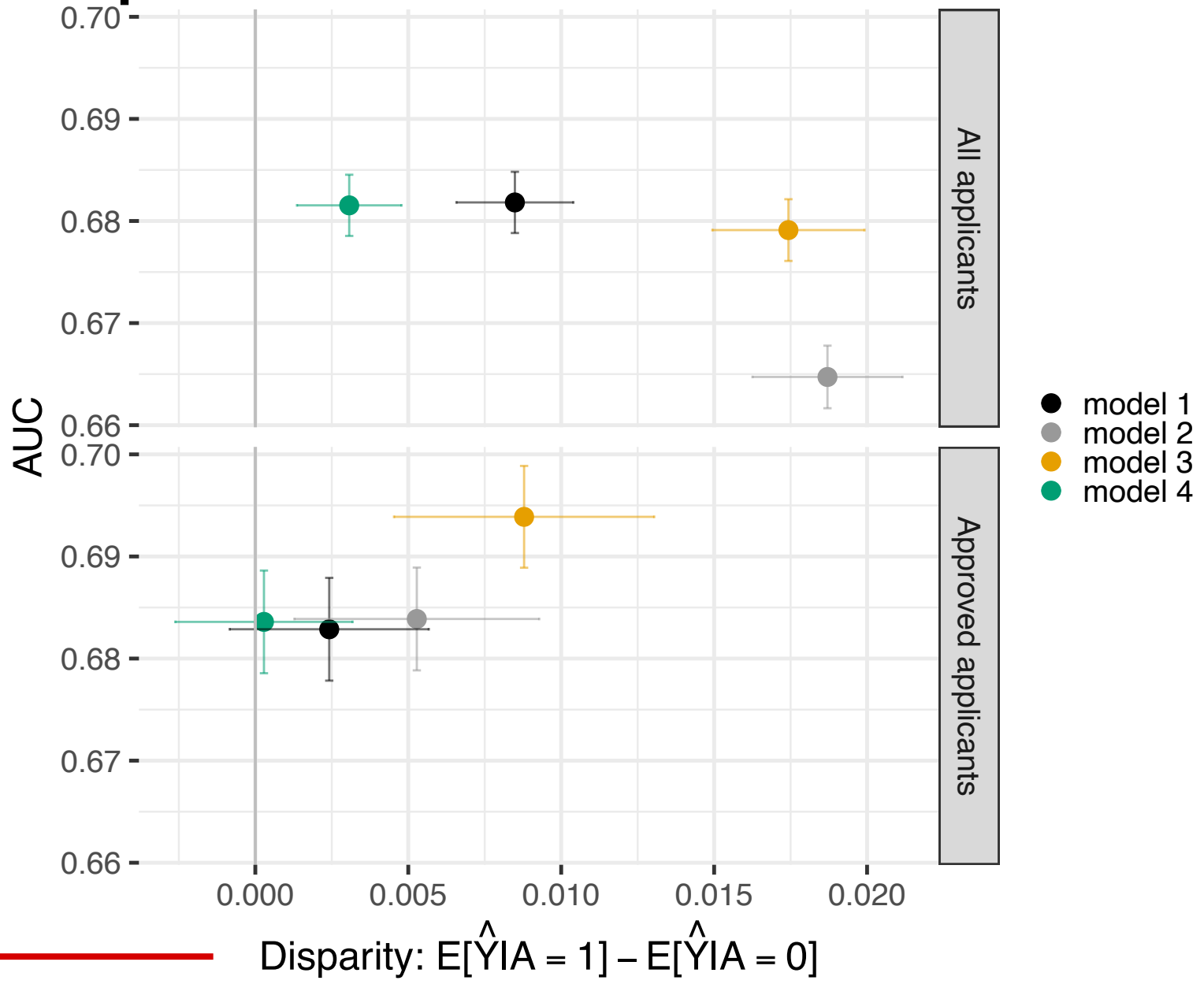
Coston, Rambachan, & Chouldechova ICML 2021

Kallus & Zhou ICML 2018

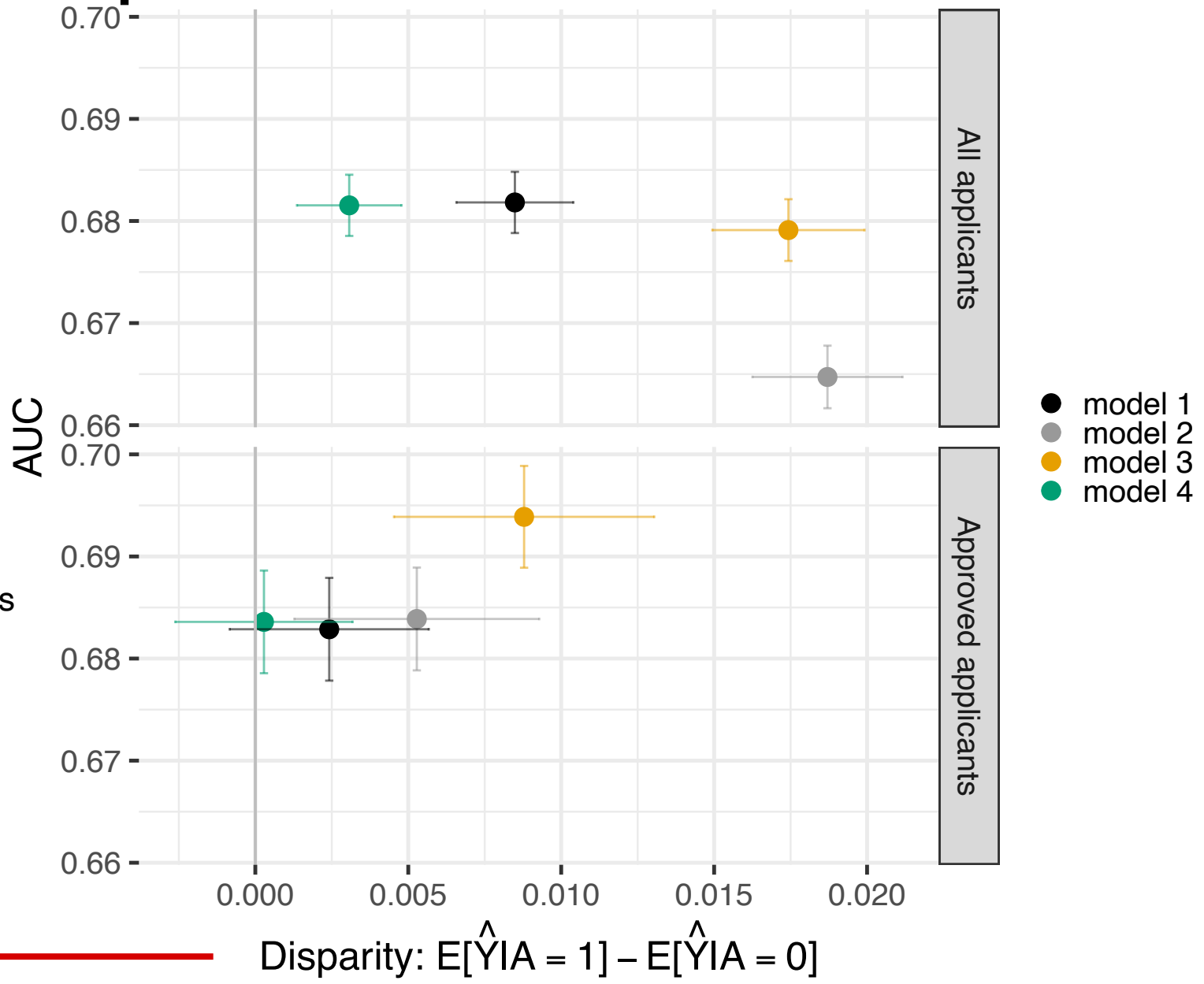
Predictive disparities under selection bias



Predictive disparities under selection bias



Predictive disparities under selection bias



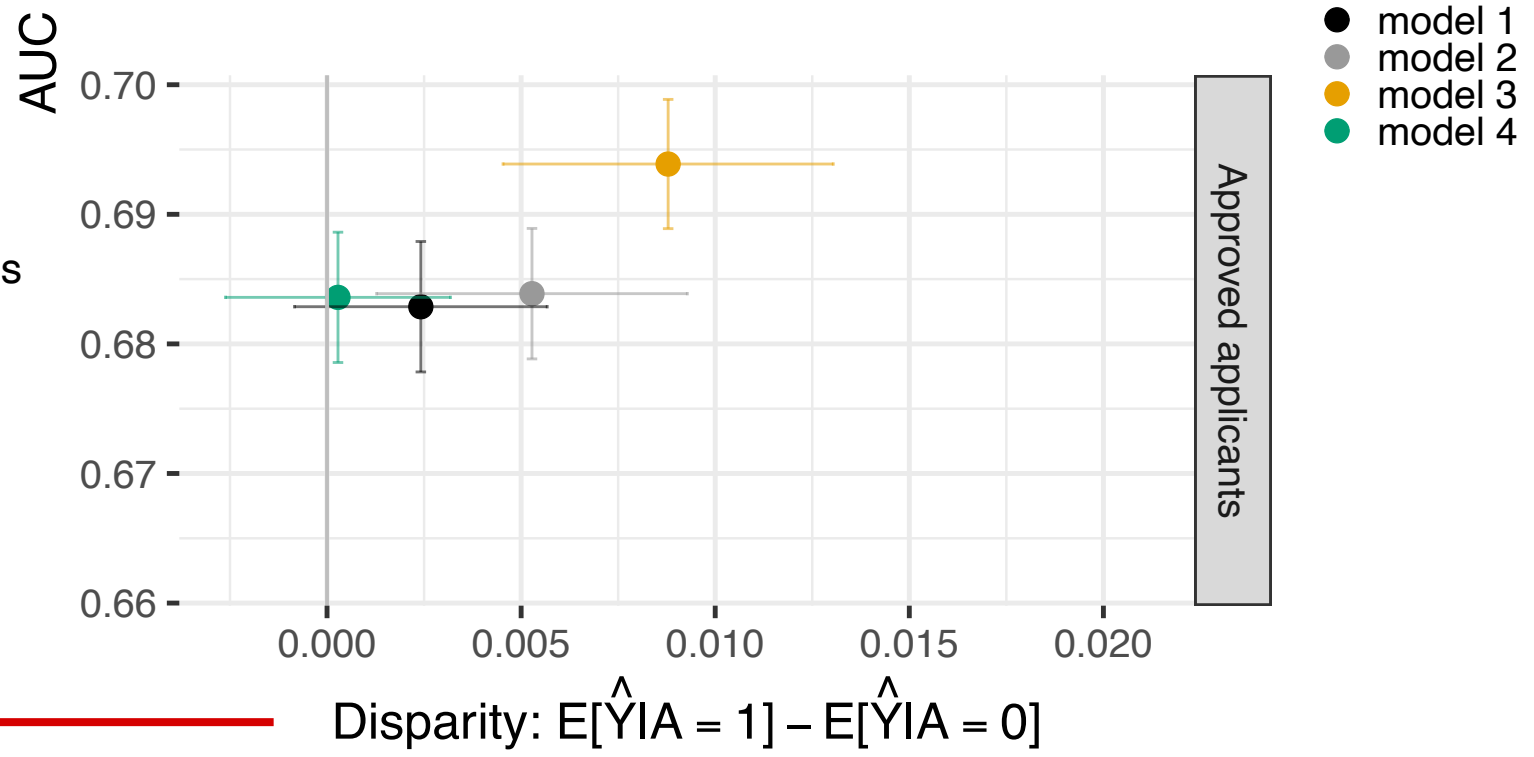
A=1: historically disadvantaged regions

A=0: privileged regions

Predictive disparities under selection bias

A=1: historically disadvantaged regions

A=0: privileged regions

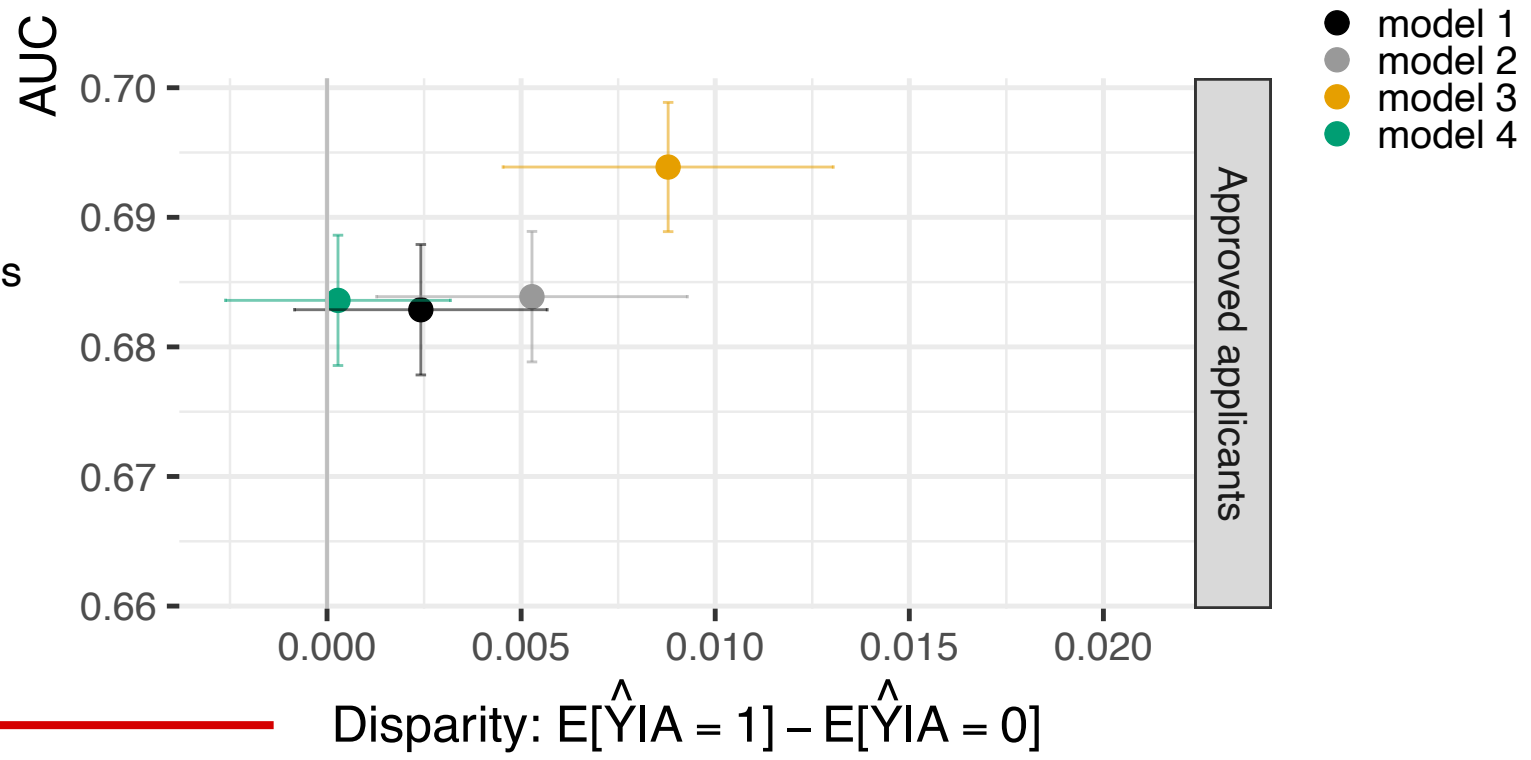


Predictive disparities under selection bias

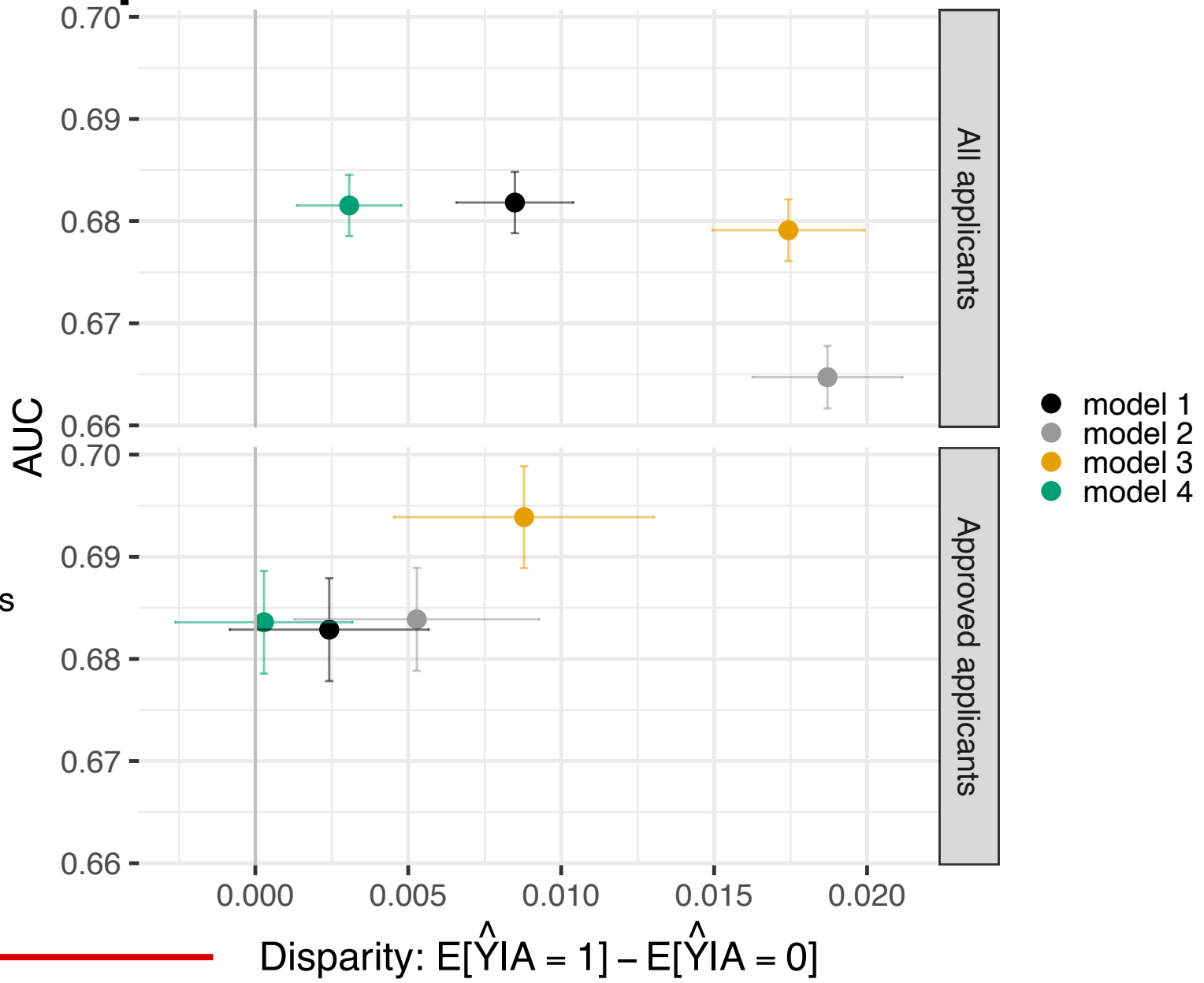
Standard eval suggests tradeoff btw. AUC & disparities

A=1: historically disadvantaged regions

A=0: privileged regions



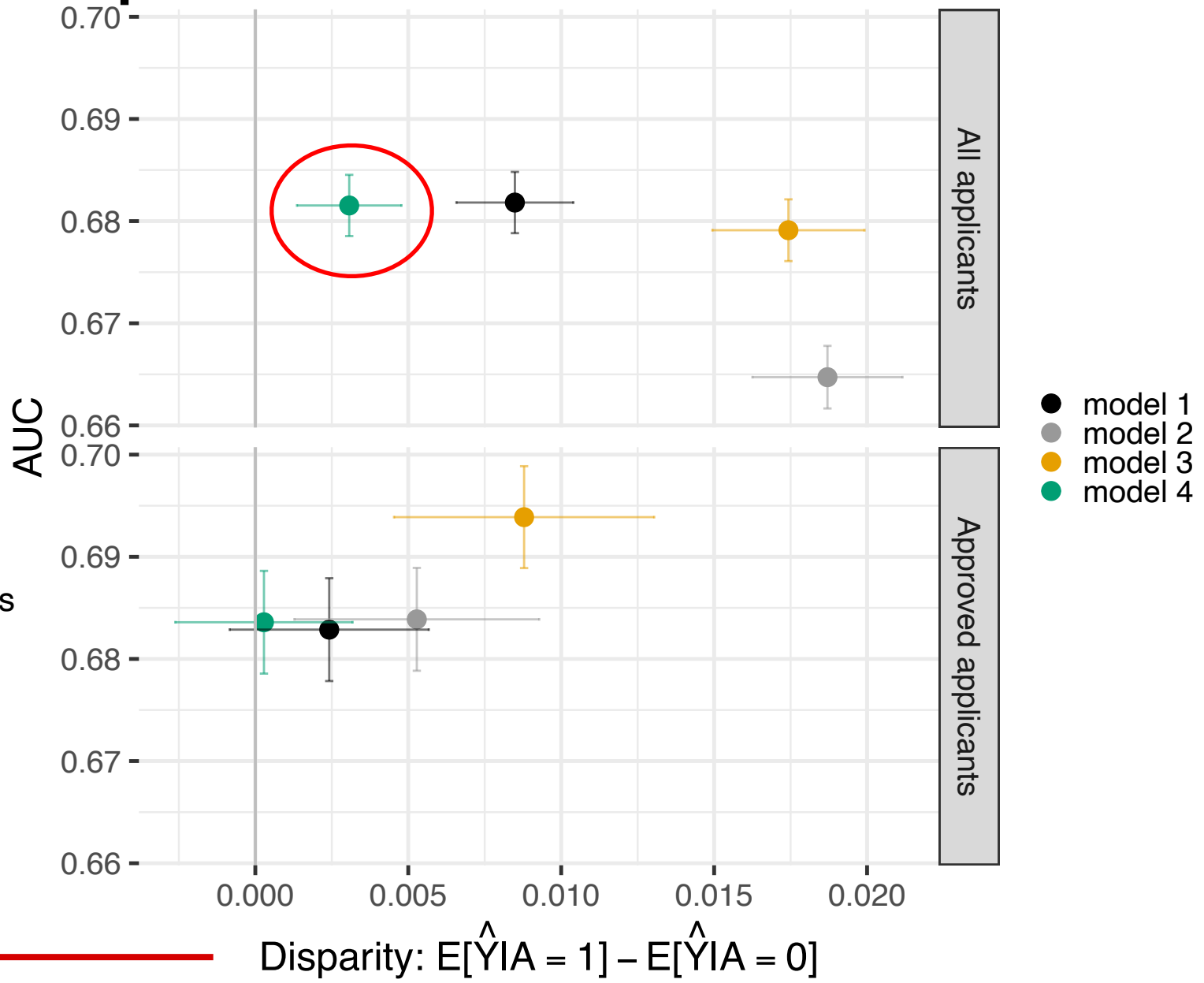
Predictive disparities under selection bias



A=1: historically disadvantaged regions

A=0: privileged regions

Predictive disparities under selection bias



A=1: historically disadvantaged regions

A=0: privileged regions

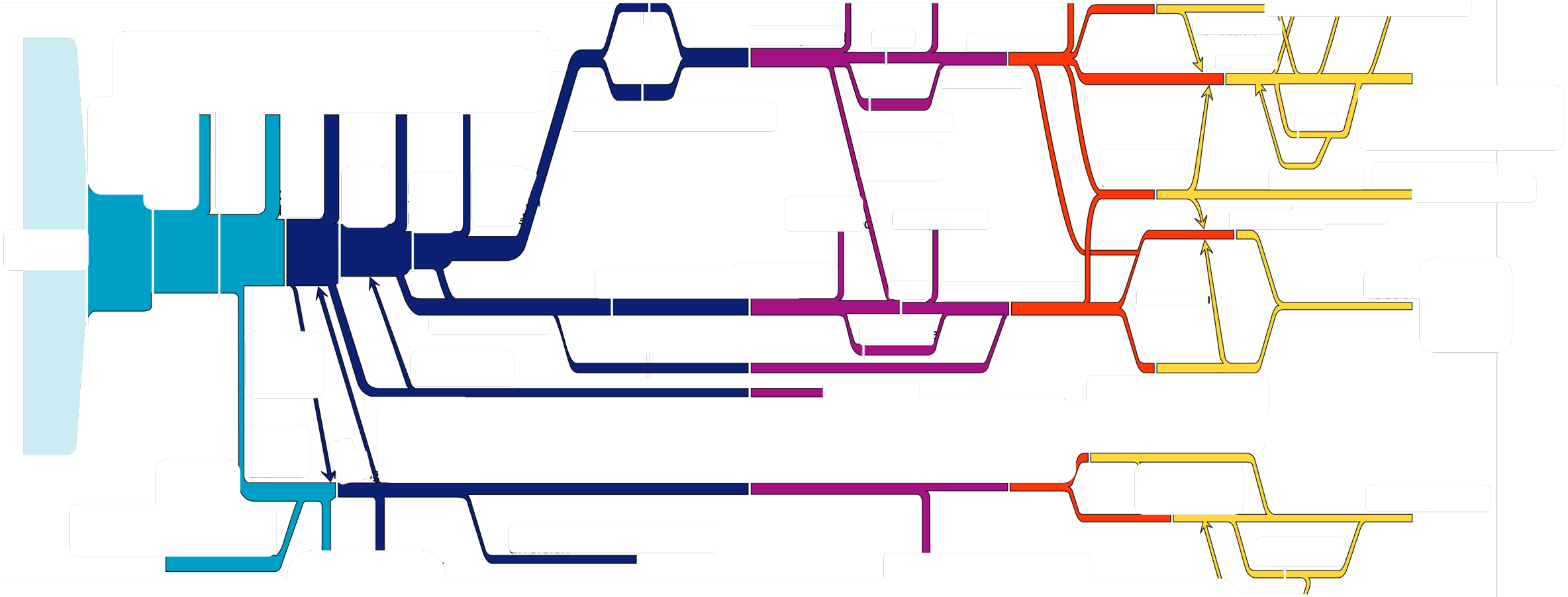

HALFTIME

Equity

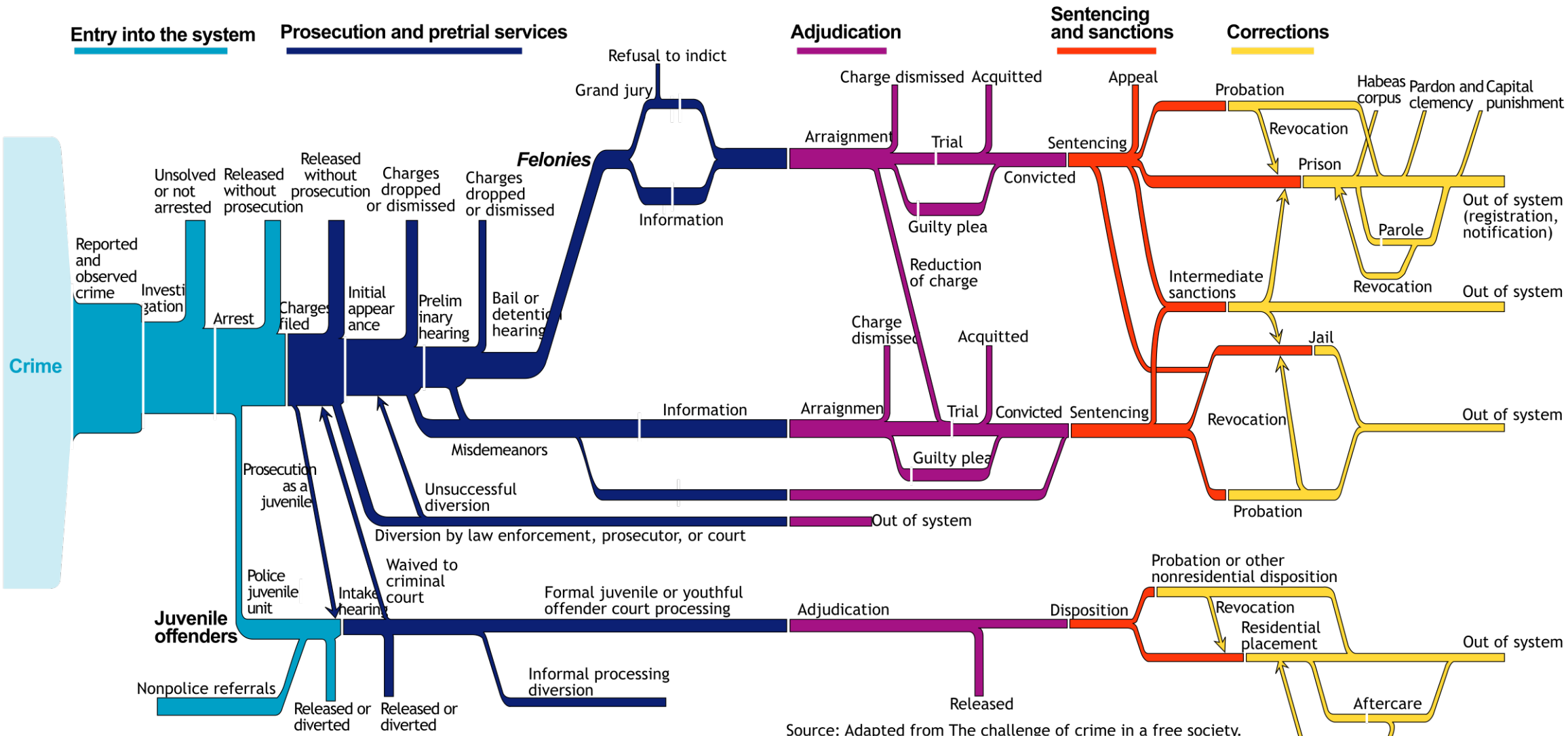
Equity

Human decisions

Algorithms are embedded in complex systems



What is the sequence of events in the criminal justice system?

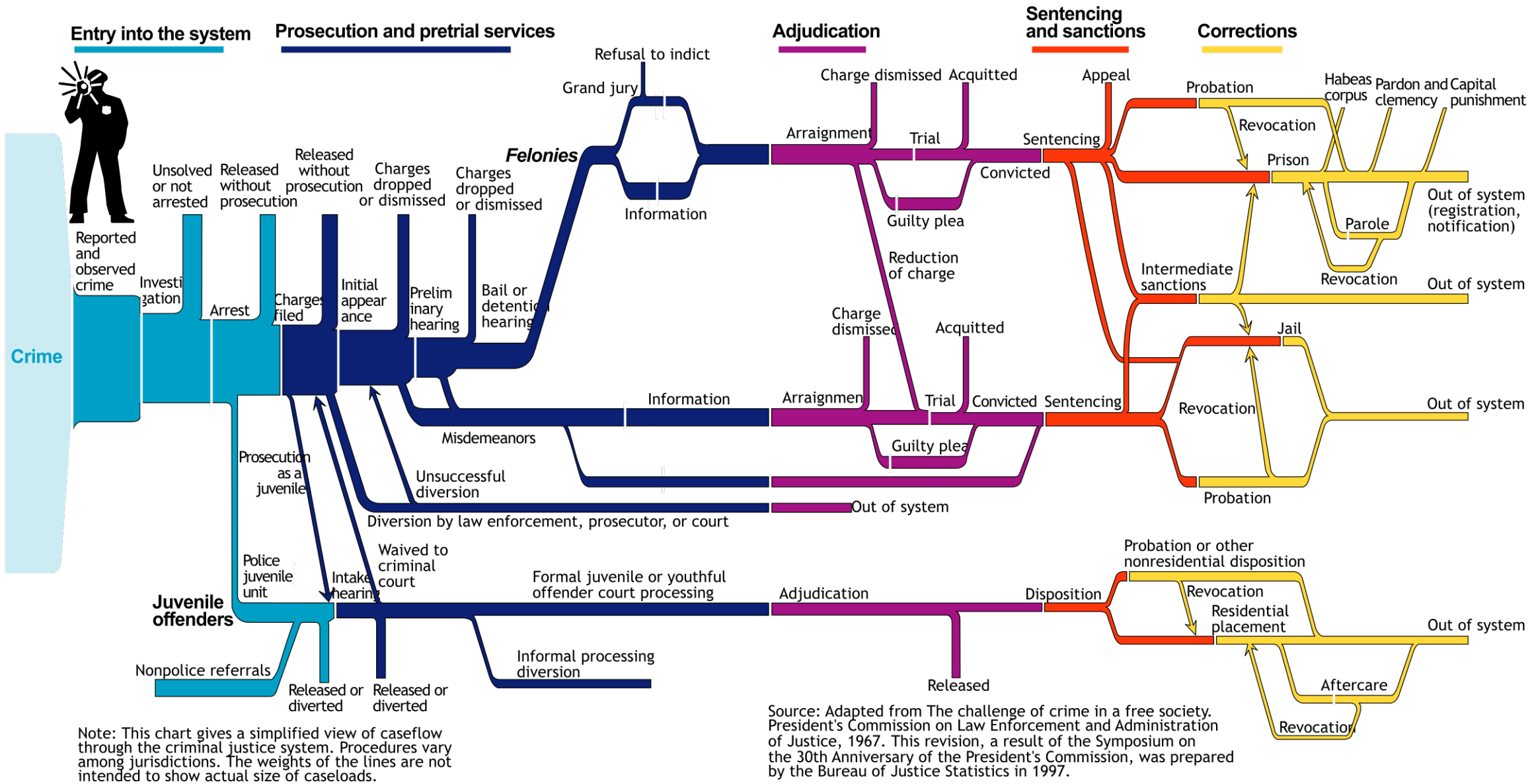


Note: This chart gives a simplified view of caseflow through the criminal justice system. Procedures vary among jurisdictions. The weights of the lines are not intended to show actual size of caseloads.

Source: Adapted from The challenge of crime in a free society. President's Commission on Law Enforcement and Administration of Justice, 1967. This revision, a result of the Symposium on the 30th Anniversary of the President's Commission, was prepared by the Bureau of Justice Statistics in 1997.

Bureau of Justice Statistics

What is the sequence of events in the criminal justice system?

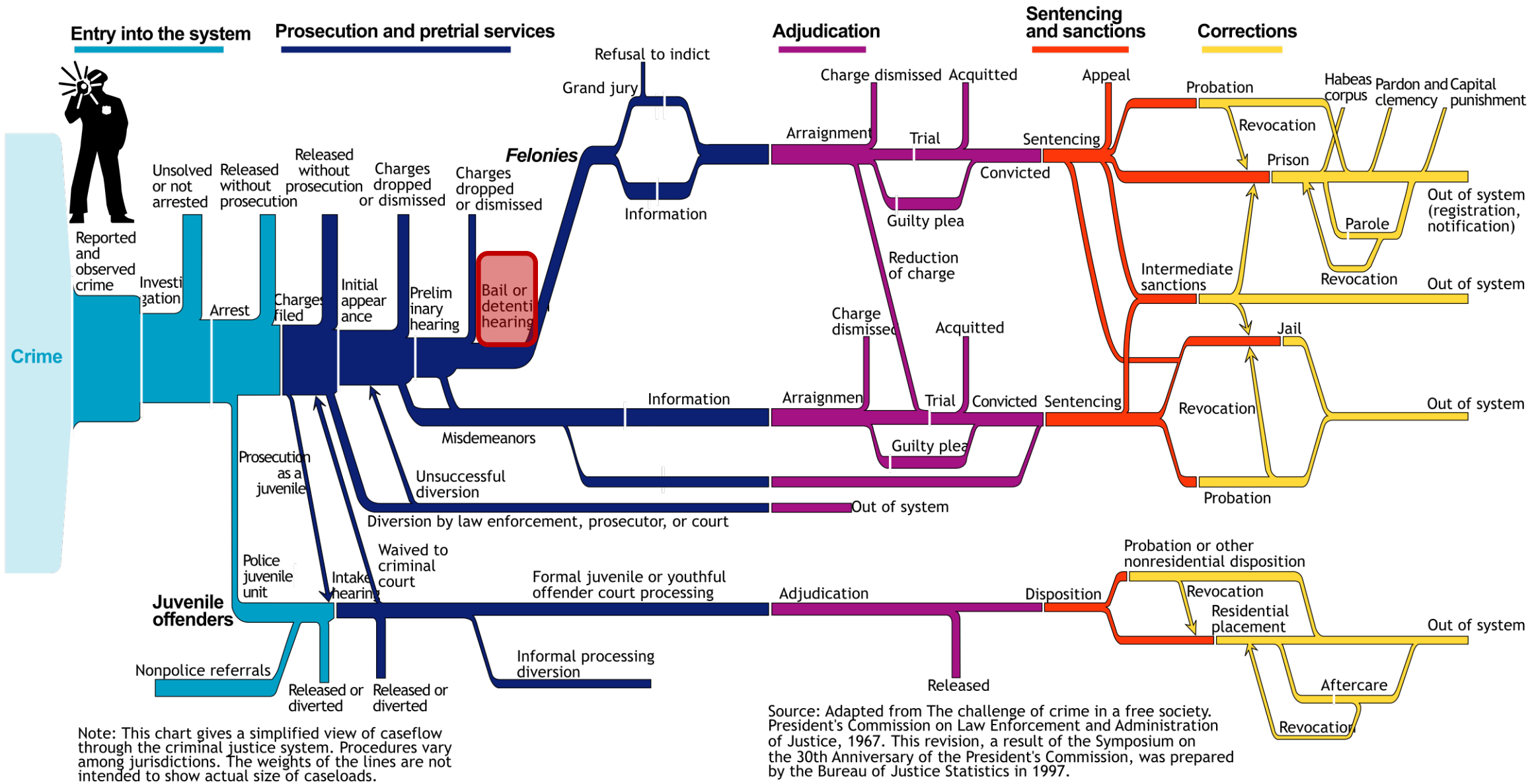


Note: This chart gives a simplified view of caseflow through the criminal justice system. Procedures vary among jurisdictions. The weights of the lines are not intended to show actual size of caseloads.

Source: Adapted from The challenge of crime in a free society. President's Commission on Law Enforcement and Administration of Justice, 1967. This revision, a result of the Symposium on the 30th Anniversary of the President's Commission, was prepared by the Bureau of Justice Statistics in 1997.

Bureau of Justice Statistics

What is the sequence of events in the criminal justice system?



Note: This chart gives a simplified view of caseflow through the criminal justice system. Procedures vary among jurisdictions. The weights of the lines are not intended to show actual size of caseloads.

Source: Adapted from The challenge of crime in a free society. President's Commission on Law Enforcement and Administration of Justice, 1967. This revision, a result of the Symposium on the 30th Anniversary of the President's Commission, was prepared by the Bureau of Justice Statistics in 1997.

Bureau of Justice Statistics

Are disparities justified?

Are disparities justified?

- reflect underlying crime patterns?

Are disparities justified?

- reflect underlying crime patterns?

Challenge: lack ground truth data

Do police officers make traffic stops based on race?

Do police officers make traffic stops based on race?

Lower Merion, PA

Do police officers make traffic stops based on race?

Lower Merion, PA

- Population: 84% white, 5% black

Do police officers make traffic stops based on race?

Lower Merion, PA

- Population: 84% white, 5% black
- 8000 police stops in 2020

➢53% white

➢36% black

source

Do police officers make traffic stops based on race?

Lower Merion, PA

- Population: 84% white, 5% black
- 8000 police stops in 2020

➢53% white

➢36% black

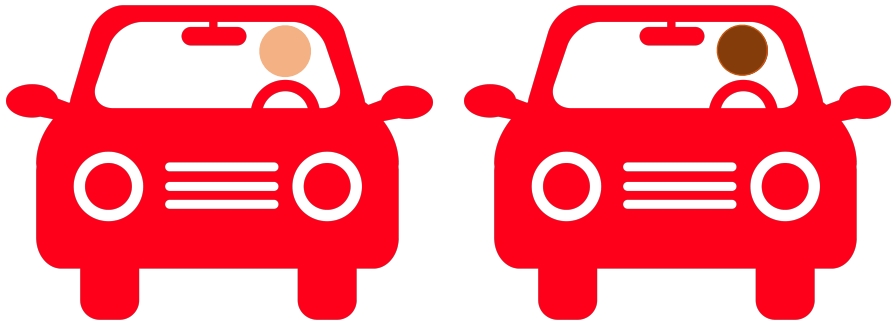
➢75% non-resident of Lower Merion

source

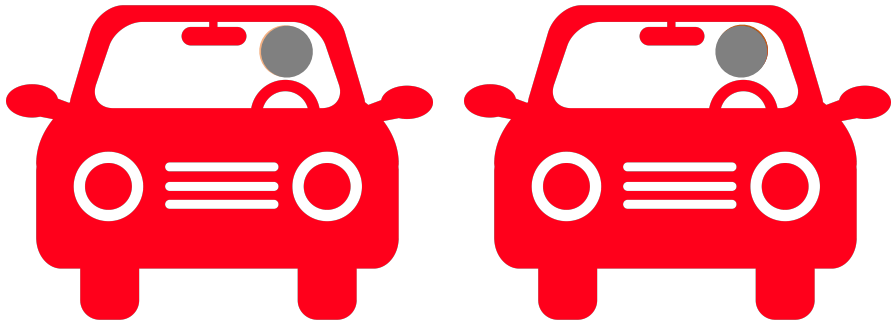
Research Question #2

How to audit for bias in police traffic stops?

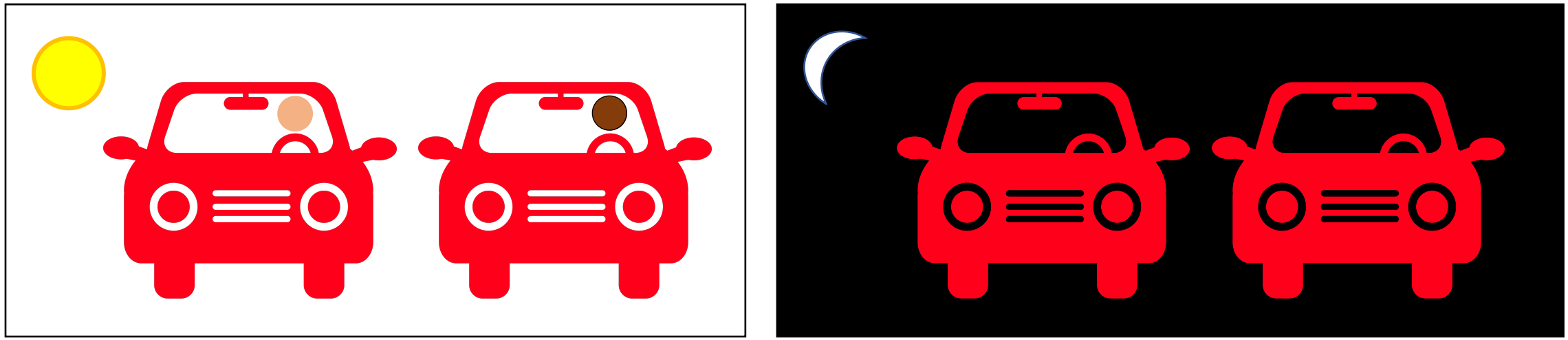
Do police officers make traffic stops based on race?



Do police officers make traffic stops based on race?

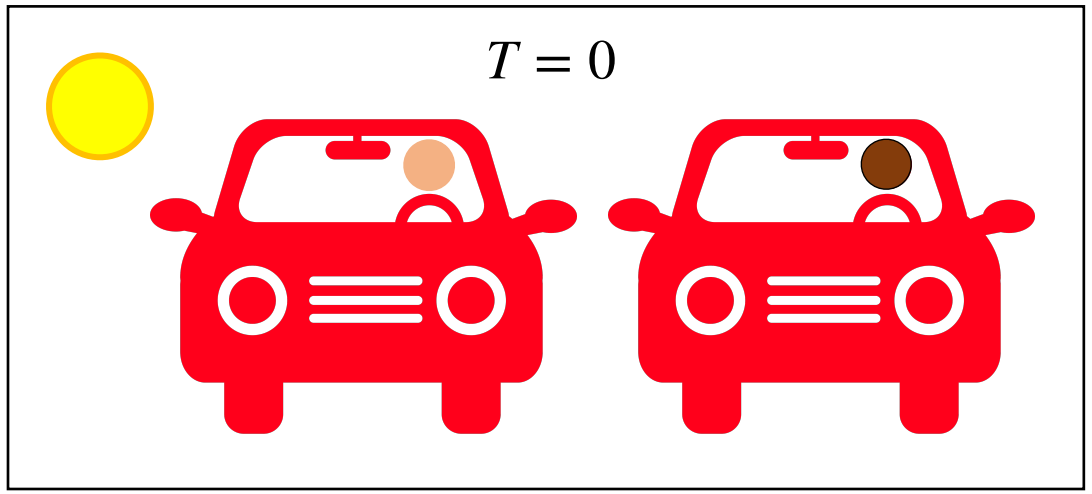


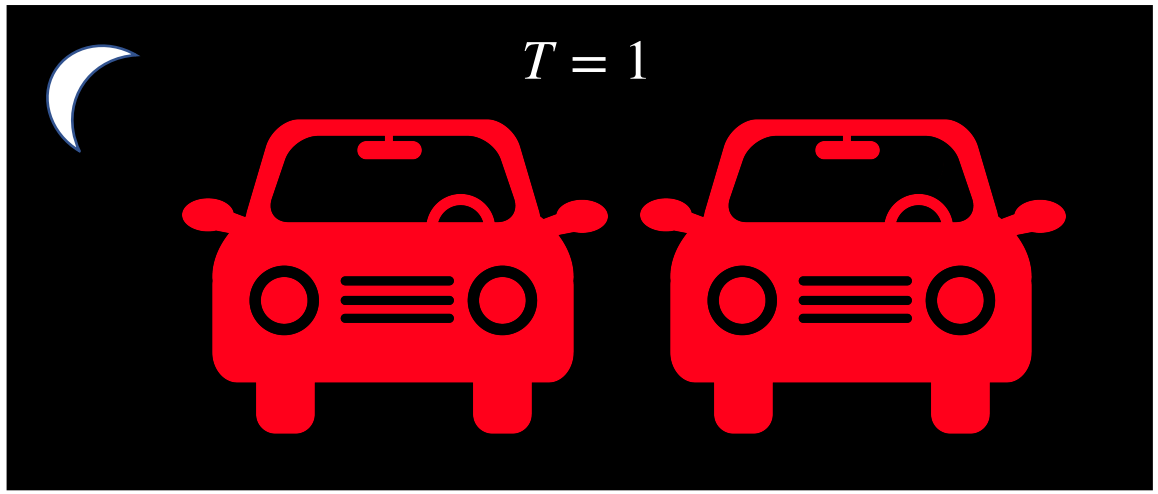
Veil of darkness



Grogger & Ridgeway. Testing for racial profiling in traffic stops from behind a veil of darkness. *JASA* 2006.

Veil of darkness





$$\text{odds}(\text{black} \mid \text{stopped}, T, X) \sim \exp(\beta^T X + \alpha T)$$

Grogger & Ridgeway. Testing for racial profiling in traffic stops from behind a veil of darkness. *JASA* 2006.

Research Question #2

- How do we audit for bias in police traffic stops?

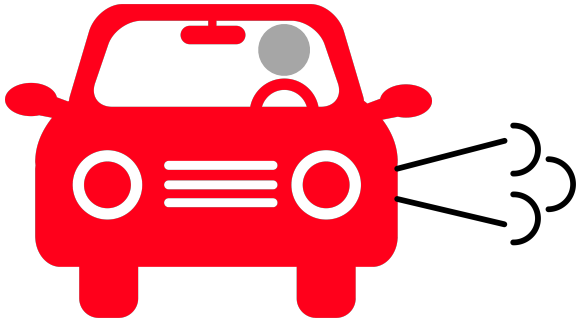
Contribution

- Counterfactual audit clarifies measure of bias & assumptions needed
- Addresses missing potential outcomes & missingness from sampling design

Ideal experiment

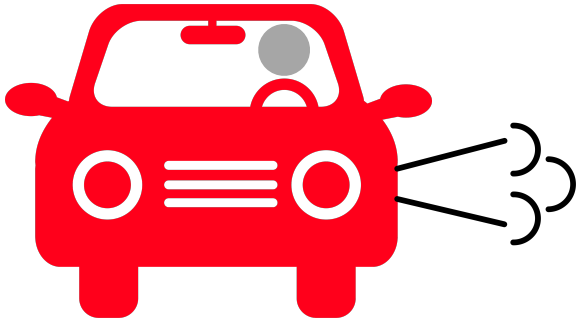
Ideal experiment

Ideal experiment



Ideal experiment

$S^0 = 1$



$S^1 = 0$

- T indicates the intervention that obfuscates race
- S^t indicate the driver is stopped under t

Feasible experiment

Feasible experiment

- T indicates dark of night
- S^t indicates the driver is stopped under t
- B indicates perceived driver race is black

Feasible experiment

- T indicates dark of night
- S^t indicates the driver is stopped under t
- B indicates perceived driver race is black

What effects could measure racial bias?

Feasible experiment

- T indicates dark of night
- S^t indicates the driver is stopped under t
- B indicates perceived driver race is black

What effects could measure racial bias?

$$\mathbb{E}[S^1 - S^0 \mid B = 1]$$

Feasible experiment

- T indicates dark of night
- S^t indicates the driver is stopped under t
- B indicates perceived driver race is black

What effects could measure racial bias?

$$\mathbb{E}[S^1 - S^0 \mid B = 1]$$

Feasible experiment

- T indicates dark of night
- S^t indicates the driver is stopped under t
- B indicates perceived driver race is black

What effects could measure racial bias?

$$\mathbb{E}[S^1 - S^0 \mid B = 1] \quad \text{vs} \quad \mathbb{E}[S^1 - S^0 \mid B = 0]$$

Feasible experiment

- T indicates dark of night
- S^t indicates the driver is stopped under t
- B indicates perceived driver race is black

What effects could measure racial bias?

$$\mathbb{E}[S^1 - S^0 \mid B = 1] \quad \text{vs} \quad \mathbb{E}[S^1 - S^0 \mid B = 0]$$

$$\frac{\mathbb{E}[S^1 \mid B = 1]}{\mathbb{E}[S^0 \mid B = 1]} \quad \text{vs} \quad \frac{\mathbb{E}[S^1 \mid B = 0]}{\mathbb{E}[S^0 \mid B = 0]}$$

Feasible experiment

- T indicates dark of night
- S^t indicates the driver is stopped under t
- B indicates perceived driver race is black

What effects could measure racial bias?

$$\mathbb{E}[S^1 - S^0 \mid B = 1] \quad \text{vs} \quad \mathbb{E}[S^1 - S^0 \mid B = 0]$$

$$\frac{\mathbb{E}[S^1 \mid B = 1]}{\mathbb{E}[S^0 \mid B = 1]} \quad \text{vs} \quad \frac{\mathbb{E}[S^1 \mid B = 0]}{\mathbb{E}[S^0 \mid B = 0]}$$

Feasible experiment

- T indicates dark of night
- S^t indicates the driver is stopped under t
- B indicates perceived driver race is black

Feasible experiment

- T indicates dark of night
- S^t indicates the driver is stopped under t
- B indicates perceived driver race is black

Driver	Race	S^1	S^0	T
1	Black	1	0	1
2	White	1	0	0
3	Black	0	1	0

Feasible experiment

- T indicates dark of night
- S^t indicates the driver is stopped under t
- B indicates perceived driver race is black

Driver	Race	S^1	S^0	T
1	Black	1	0	1
2	White	1	0	0
3	Black	0	1	0

"Fundamental problem of causal inference"
(Holland 1986)

Feasible experiment

- T indicates dark of night
- S^t indicates the driver is stopped under t
- B indicates perceived driver race is black
- We don't observe $S^1 = 0$ or $S^0 = 0$

Driver	Race	S^1	S^0	T
1	Black	1	0	1
2	White	1	0	0
3	Black	0	1	0

"Fundamental problem of causal inference" (Holland 1986)

Feasible experiment

- T indicates dark of night
- S^t indicates the driver is stopped under t
- B indicates perceived driver race is black
- We don't observe $S^1 = 0$ or $S^0 = 0$

Driver	Race	S^1	S^0	T
1	Black	1	0	1
2	White	1	0	0
3	Black	0	1	0

"Fundamental problem of causal inference" (Holland 1986)

+ outcome dependent sampling

Available data

- Observe samples $\sim P(T, B, X \mid S = 1)$
- S indicates stop
- T indicates dark of night
- B indicates perceived driver race is black
- X describes features like location & time

Risk ratio measure of racial bias

$$\frac{P(S^1 = 1 \mid B = 1)}{P(S^0 = 1 \mid B = 1)} \Big/ \frac{P(S^1 = 1 \mid B = 0)}{P(S^0 = 1 \mid B = 0)}$$

Ratio of risk ratios

Risk ratio measure of racial bias

$$\frac{P(S^1 = 1 \mid B = 1)}{P(S^0 = 1 \mid B = 1)} \Big/ \frac{P(S^1 = 1 \mid B = 0)}{P(S^0 = 1 \mid B = 0)}$$

Ratio of risk ratios

$$= \frac{P(B = 1 \mid S^1 = 1)}{P(B = 0 \mid S^1 = 1)} \Big/ \frac{P(B = 1 \mid S^0 = 1)}{P(B = 0 \mid S^0 = 1)}$$

Odds ratio

Risk ratio measure of racial bias

$$\frac{P(S^1 = 1 \mid B = 1)}{P(S^0 = 1 \mid B = 1)} \Big/ \frac{P(S^1 = 1 \mid B = 0)}{P(S^0 = 1 \mid B = 0)}$$

Ratio of risk ratios

$$= \frac{P(B = 1 \mid S^1 = 1)}{P(B = 0 \mid S^1 = 1)} \Big/ \frac{P(B = 1 \mid S^0 = 1)}{P(B = 0 \mid S^0 = 1)}$$

Odds ratio

Key result: Identification of X-conditional measure

Key result: Identification of X-conditional measure

Assumptions

- $T \perp B \mid S^t = 1, X$

Key result: Identification of X-conditional measure

Assumptions

- $T \perp B \mid S^t = 1, X$
- $P(0 < \pi(X) < 1) = 1$ where $\pi(x) = P(T = 1 \mid X = x, S = 1)$

Key result: Identification of X-conditional measure

Assumptions

- $T \perp B \mid S^t = 1, X$
- $P(0 < \pi(X) < 1) = 1$ where $\pi(x) = P(T = 1 \mid X = x, S = 1)$

$$\psi(X) := \frac{\text{odds}(B = 1 \mid X = x, S^1 = 1)}{\text{odds}(B = 1 \mid X = x, S^0 = 1)} = \frac{\text{odds}(B = 1 \mid X = x, T = 1, S = 1)}{\text{odds}(B = 1 \mid X = x, T = 0, S = 1)}$$

Key result: Identification of X-conditional measure

Assumptions

- $T \perp B \mid S^t = 1, X$
- $P(0 < \pi(X) < 1) = 1$ where $\pi(x) = P(T = 1 \mid X = x, S = 1)$

$$\psi(X) := \frac{\text{odds}(B = 1 \mid X = x, S^1 = 1)}{\text{odds}(B = 1 \mid X = x, S^0 = 1)} = \frac{\text{odds}(B = 1 \mid X = x, T = 1, S = 1)}{\text{odds}(B = 1 \mid X = x, T = 0, S = 1)}$$

Target of prior veil of darkness work

Our flexible estimator for X-conditional measure

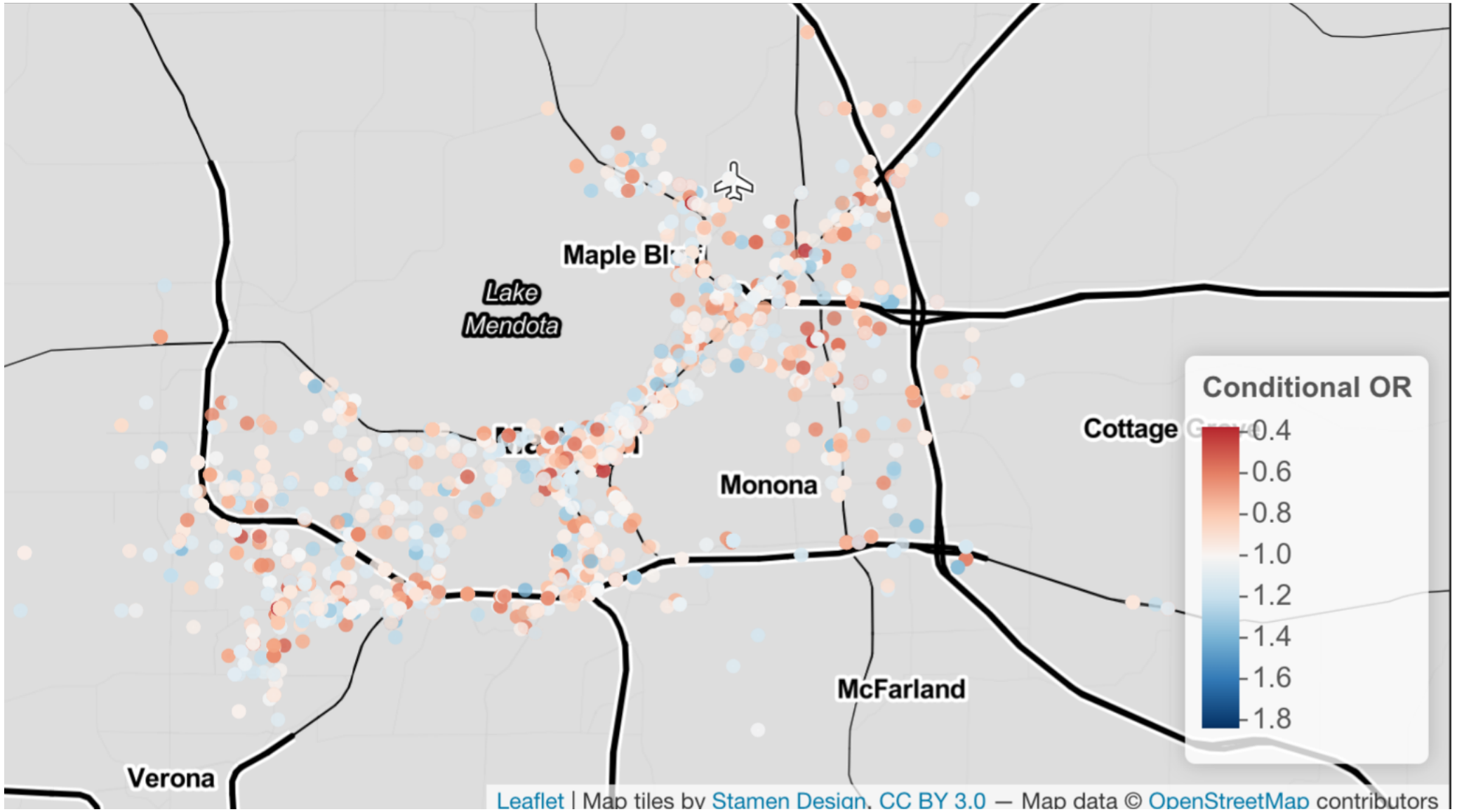
$$\mu_1(x) := P(B = 1 \mid X = x,\ T = 1,\ S = 1)$$
$$\mu_0(x) := P(B = 1 \mid X = x,\ T = 0,\ S = 1)$$

$$\hat{\psi}(x) = \frac{\hat{\mu}_1(X) \ / \ \left(1 - \hat{\mu}_1(x)\right)}{\hat{\mu}_0(X) \ / \ \left(1 - \hat{\mu}_0(x)\right)}$$

Empirical analysis

- Data from Stanford Open Policing Project[1] over 10-year period
- Random variation: daylight saving time
- $X = \{$time of day, day of week, fall/spring$\}$

[1] Pierson et al. (2020). A large-scale analysis of racial disparities in police stops across the United States. Nature human behaviour, 4(7), 736-745.


Lake
Mendota
Monona
McFarland
Verona
Cottage
Conditional OR
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap contributors

Our bias-corrected estimator for aggregated effect

$$P_n(\hat{\phi}) := \frac{1}{n} \sum_{i=1}^{n} \hat{\phi}(T_i, B_i, X_i)$$

Our bias-corrected estimator for aggregated effect

$$P_n(\hat{\phi}) := \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}(T_i, B_i, X_i)$$

$$\hat{\phi}(T, B, X) =$$

$$\text{logit}(\hat{\mu}_1(X)) - \text{logit}(\hat{\mu}_0(X)) + \frac{T(B - \hat{\mu}_1(X))}{\hat{\mu}_1(X)(1 - \hat{\mu}_1(X))\hat{\pi}(X)} - \frac{(1 - T)(B - \hat{\mu}_0(X))}{\hat{\mu}_0(X)(1 - \hat{\mu}_0(X))(1 - \hat{\pi}(X))}$$

Our bias-corrected estimator for aggregated effect

$$P_n(\hat{\phi}) := \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}(T_i, B_i, X_i)$$

$$\hat{\phi}(T, B, X) =$$

$$\text{logit}(\hat{\mu}_1(X)) - \text{logit}(\hat{\mu}_0(X)) + \frac{T(B - \hat{\mu}_1(X))}{\hat{\mu}_1(X)(1 - \hat{\mu}_1(X))\hat{\pi}(X)} - \frac{(1 - T)(B - \hat{\mu}_0(X))}{\hat{\mu}_0(X)(1 - \hat{\mu}_0(X))(1 - \hat{\pi}(X))}$$

Plug-in

Our bias-corrected estimator for aggregated effect

$$P_n(\hat{\phi}) := \frac{1}{n}\sum_{i=1}^{n} \hat{\phi}(T_i, B_i, X_i)$$

$$\hat{\phi}(T, B, X) =$$

$$\text{logit}(\hat{\mu}_1(X)) - \text{logit}(\hat{\mu}_0(X)) + \frac{T(B - \hat{\mu}_1(X))}{\hat{\mu}_1(X)(1 - \hat{\mu}_1(X))\hat{\pi}(X)} - \frac{(1 - T)(B - \hat{\mu}_0(X))}{\hat{\mu}_0(X)(1 - \hat{\mu}_0(X))(1 - \hat{\pi}(X))}$$

Plug-in

Bias correction

Key result: our estimator has 2nd-order error

Theorem. Our estimator has error $P_n(\hat{\phi}) - \log(\Psi) =$

$$(P_n - P)(\phi) + o_P\left(\frac{1}{\sqrt{n}}\right) + O_P\left(\sum_{a=0}^{1} \|\hat{\pi} - \pi\| \, \|\hat{\mu}_a - \mu_a\| + \|\hat{\mu}_a - \mu_a\|^2\right)$$

assuming

- $\left\|\phi - \hat{\phi}\right\| = o_P(1)$
- sample-splitting & strong overlap
- $P(\mu_a(X)(1 - \mu_a(X)) > \epsilon) = 1$ & $P(\hat{\mu}_a(X)(1 - \hat{\mu}_a(X)) > \epsilon) = 1$

Key result: our estimator has 2nd-order error

Theorem. Our estimator has error $P_n(\hat{\phi}) - \log(\Psi) =$

CLT

$$(P_n - P)(\phi) + o_P\left(\frac{1}{\sqrt{n}}\right) + O_P\left(\sum_{a=0}^{1} \|\hat{\pi} - \pi\| \, \|\hat{\mu}_a - \mu_a\| + \|\hat{\mu}_a - \mu_a\|^2\right)$$

assuming

- $\left\|\phi - \hat{\phi}\right\| = o_P(1)$
- sample-splitting & strong overlap
- $P(\mu_a(X)(1 - \mu_a(X)) > \epsilon) = 1$ & $P(\hat{\mu}_a(X)(1 - \hat{\mu}_a(X)) > \epsilon) = 1$

Key result: our estimator has 2nd-order error

Theorem. Our estimator has error $P_n(\hat{\phi}) - \log(\Psi) =$

CLT small error

$$(P_n - P)(\phi) + o_P\left(\frac{1}{\sqrt{n}}\right) + O_P\left(\sum_{a=0}^{1} \|\hat{\pi} - \pi\| \, \|\hat{\mu}_a - \mu_a\| + \|\hat{\mu}_a - \mu_a\|^2\right)$$

assuming

- $\left\|\phi - \hat{\phi}\right\| = o_P(1)$
- sample-splitting & strong overlap
- $P(\mu_a(X)(1 - \mu_a(X)) > \epsilon) = 1$ & $P(\hat{\mu}_a(X)(1 - \hat{\mu}_a(X)) > \epsilon) = 1$

Key result: our estimator has 2nd-order error

Theorem. Our estimator has error $P_n(\hat{\phi}) - \log(\Psi) =$

CLT small error 2nd-order nuisance error

$$(P_n - P)(\phi) + o_P\left(\frac{1}{\sqrt{n}}\right) + O_P\left(\sum_{a=0}^{1} \|\hat{\pi} - \pi\| \, \|\hat{\mu}_a - \mu_a\| + \|\hat{\mu}_a - \mu_a\|^2\right)$$

assuming

- $\left\|\phi - \hat{\phi}\right\| = o_P(1)$
- sample-splitting & strong overlap
- $P(\mu_a(X)(1 - \mu_a(X)) > \epsilon) = 1$ & $P(\hat{\mu}_a(X)(1 - \hat{\mu}_a(X)) > \epsilon) = 1$

Key result: our estimator has 2nd-order error

Theorem. Our estimator has error $P_n(\hat{\phi}) - \log(\Psi) =$

CLT small error 2nd-order nuisance error

$$(P_n - P)(\phi) + o_P\left(\frac{1}{\sqrt{n}}\right) + O_P\left(\sum_{a=0}^{1} \|\hat{\pi} - \pi\| \, \|\hat{\mu}_a - \mu_a\| + \|\hat{\mu}_a - \mu_a\|^2\right)$$

vs error of basic plug-in:

1st-order nuisance error

$$O_P\left(\sum_{a=0}^{1} \|\hat{\mu}_a - \mu_a\|\right)$$

Key result: our estimator has 2nd-order error

Theorem. Our estimator has error $P_n(\hat{\phi}) - \log(\Psi) =$

CLT small error 2nd-order nuisance error

$$(P_n - P)(\phi) + o_P\left(\frac{1}{\sqrt{n}}\right) + O_P\left(\sum_{a=0}^{1} \|\hat{\pi} - \pi\| \, \|\hat{\mu}_a - \mu_a\| + \|\hat{\mu}_a - \mu_a\|^2\right)$$

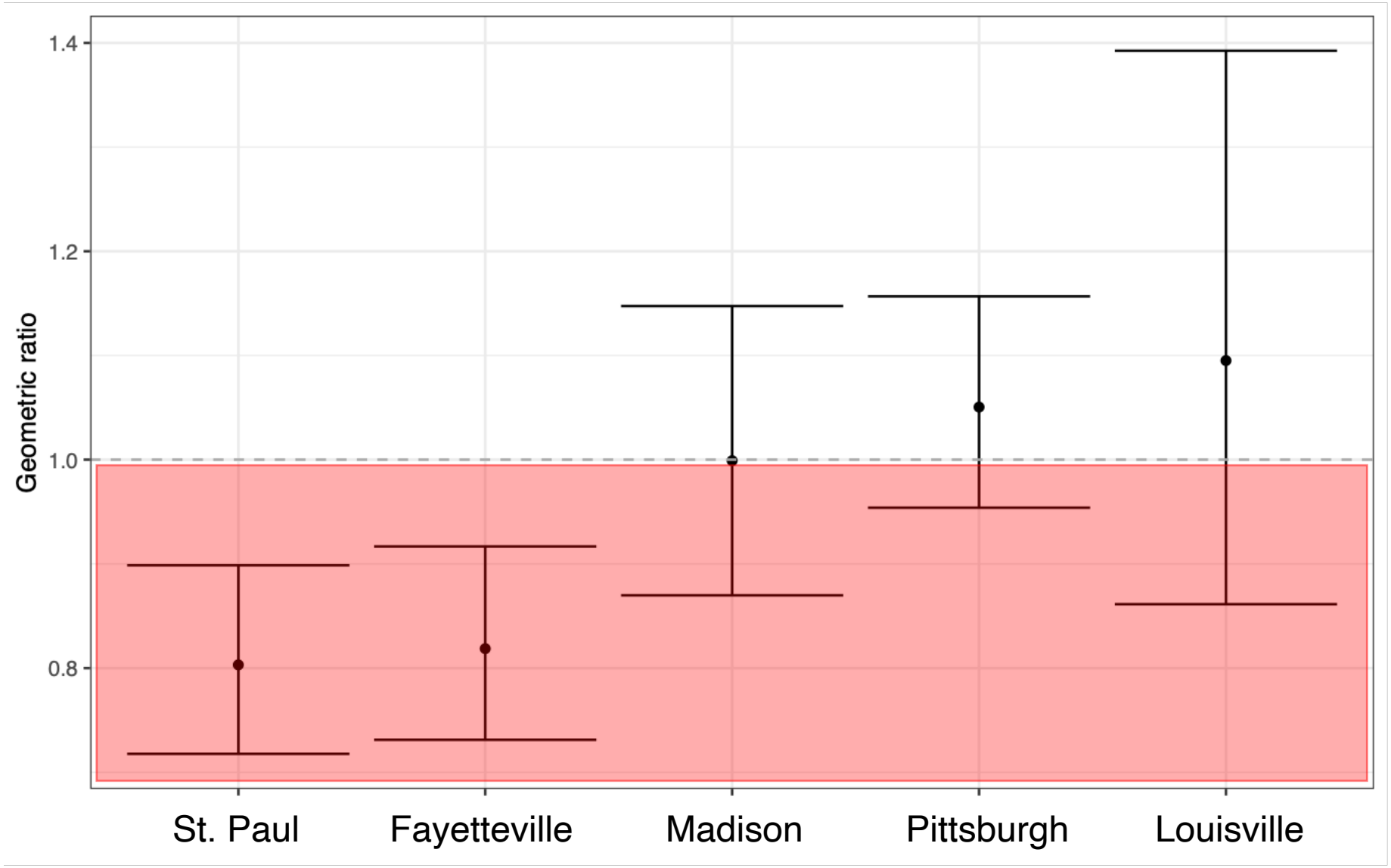
arXiv > stat > arXiv:2207.09016

Statistics > Methodology

[Submitted on 19 Jul 2022]

The role of the geometric mean in case-control studies

Amanda Coston, Edward H. Kennedy



darkness decreases risk of stop for black drivers

How do we assess the validity and equity of algorithms?

Evaluating predictive models

- **Problem:** Missing data invalidates standard approaches
- **Solution:** Counterfactual evaluation

Auditing for biased decision-making

- **Problem:** No ground truth data to assess if disparities are justified
- **Solution:** Counterfactual audit

Validity

Counterfactual risk assessments, evaluation, & fairness
FAccT 2020

Counterfactual predictions under runtime confounding
NeurIPS 2020

Counterfactual prediction under unmeasured confounding
ACIC 2022

Equity

Fair transfer learning with missing protected attributes
AIES 2019

Leveraging administrative data for bias audits
FAccT 2021

Characterizing fairness over set of good models under selective labels
ICML 2021

Counterfactual audit for racial bias in police traffic stops
ACIC 2022

Validity

Counterfactual risk assessments, evaluation, & fairness
FAccT 2020

Counterfactual predictions under runtime confounding
NeurIPS 2020

Counterfactual prediction under unmeasured confounding
ACIC 2022

Equity

Fair transfer learning with missing protected attributes
AIES 2019

Leveraging administrative data for bias audits
FAccT 2021

Characterizing fairness over set of good models under selective labels
ICML 2021

Counterfactual audit for racial bias in police traffic stops
ACIC 2022

Oversight

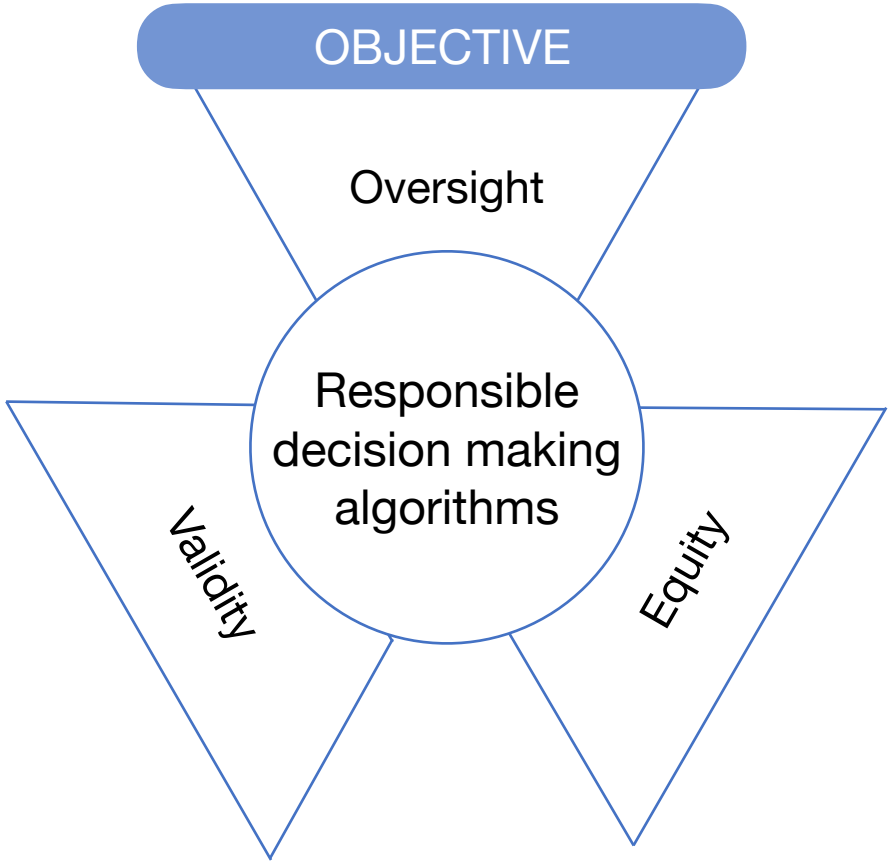
Validity Perspective on Evaluating Justified Use of Algorithms
SATML 2023

My research

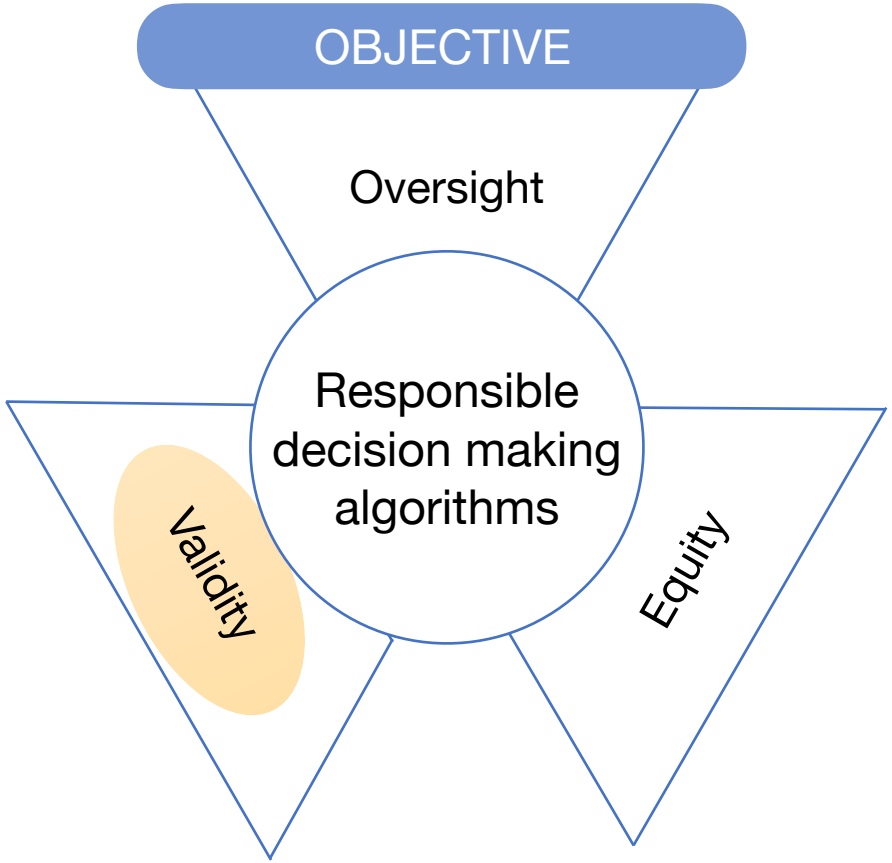
OBJECTIVE

Responsible decision making algorithms

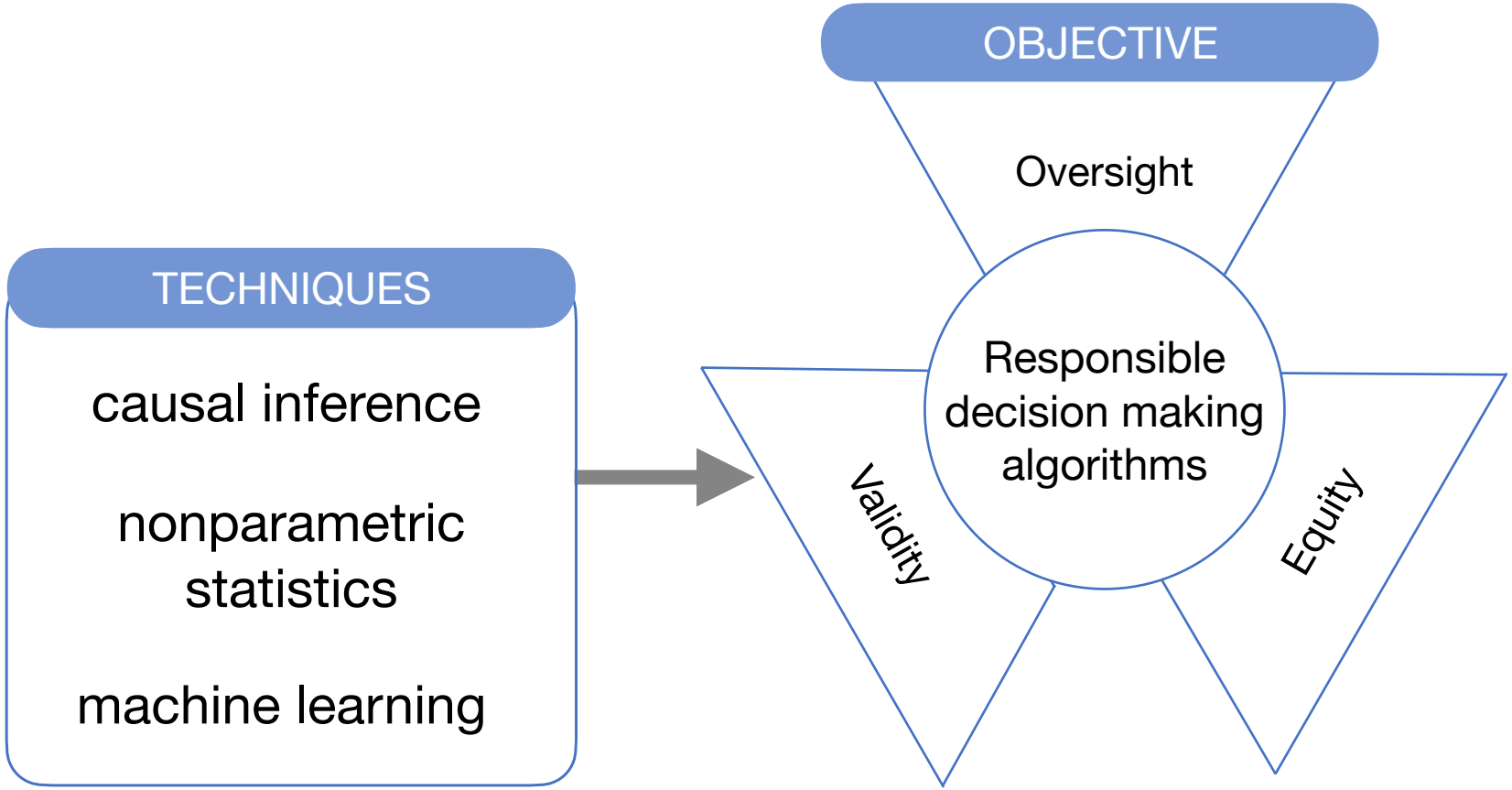
My research


OBJECTIVE
Oversight
Responsible decision making algorithms
Validity
Equity

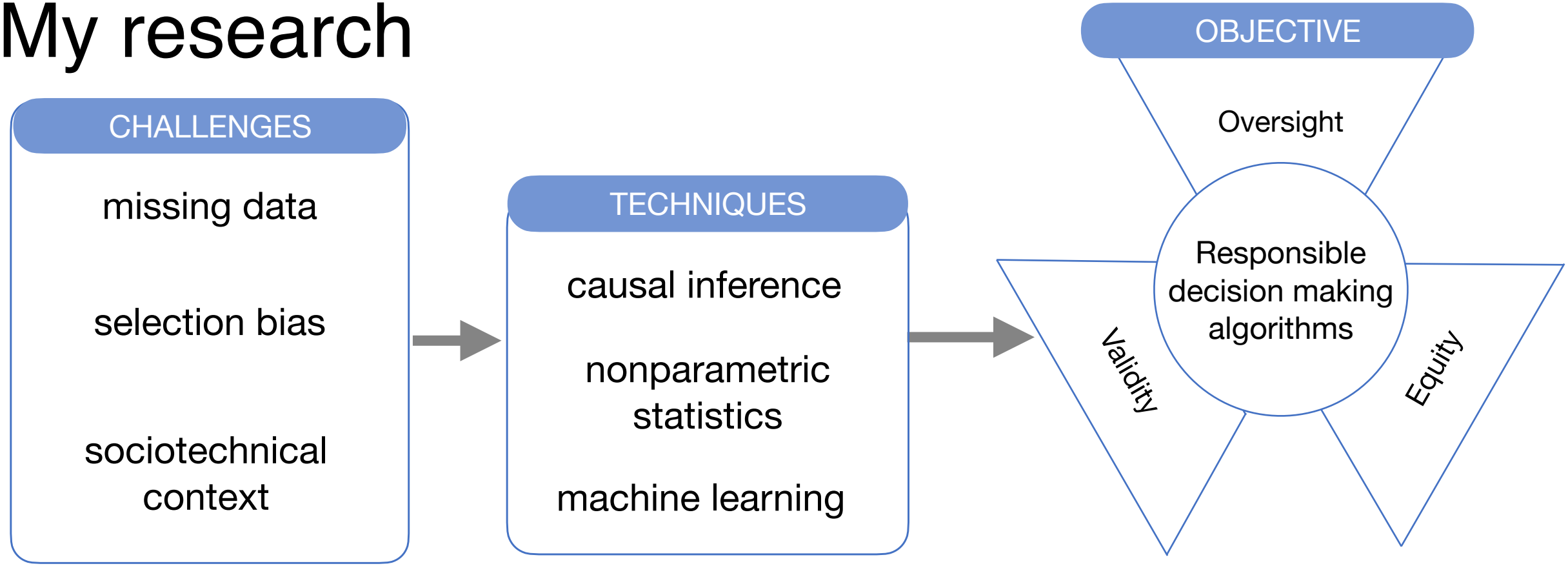
My research


OBJECTIVE
Oversight
Responsible decision making algorithms
Validity
Equity

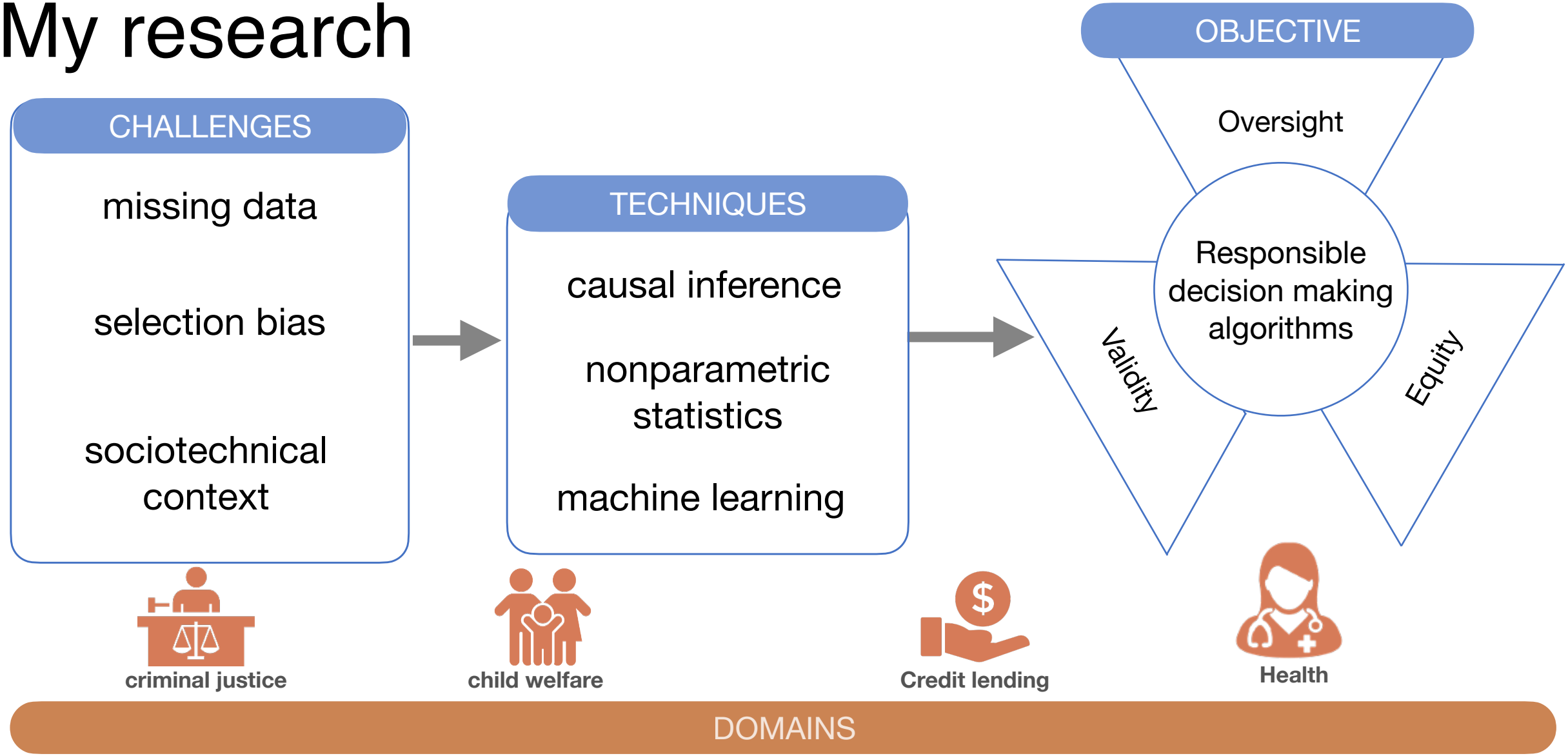
My research


TECHNIQUES
causal inference
nonparametric statistics
machine learning
OBJECTIVE
Oversight
Responsible decision making algorithms
Validity
Equity

My research

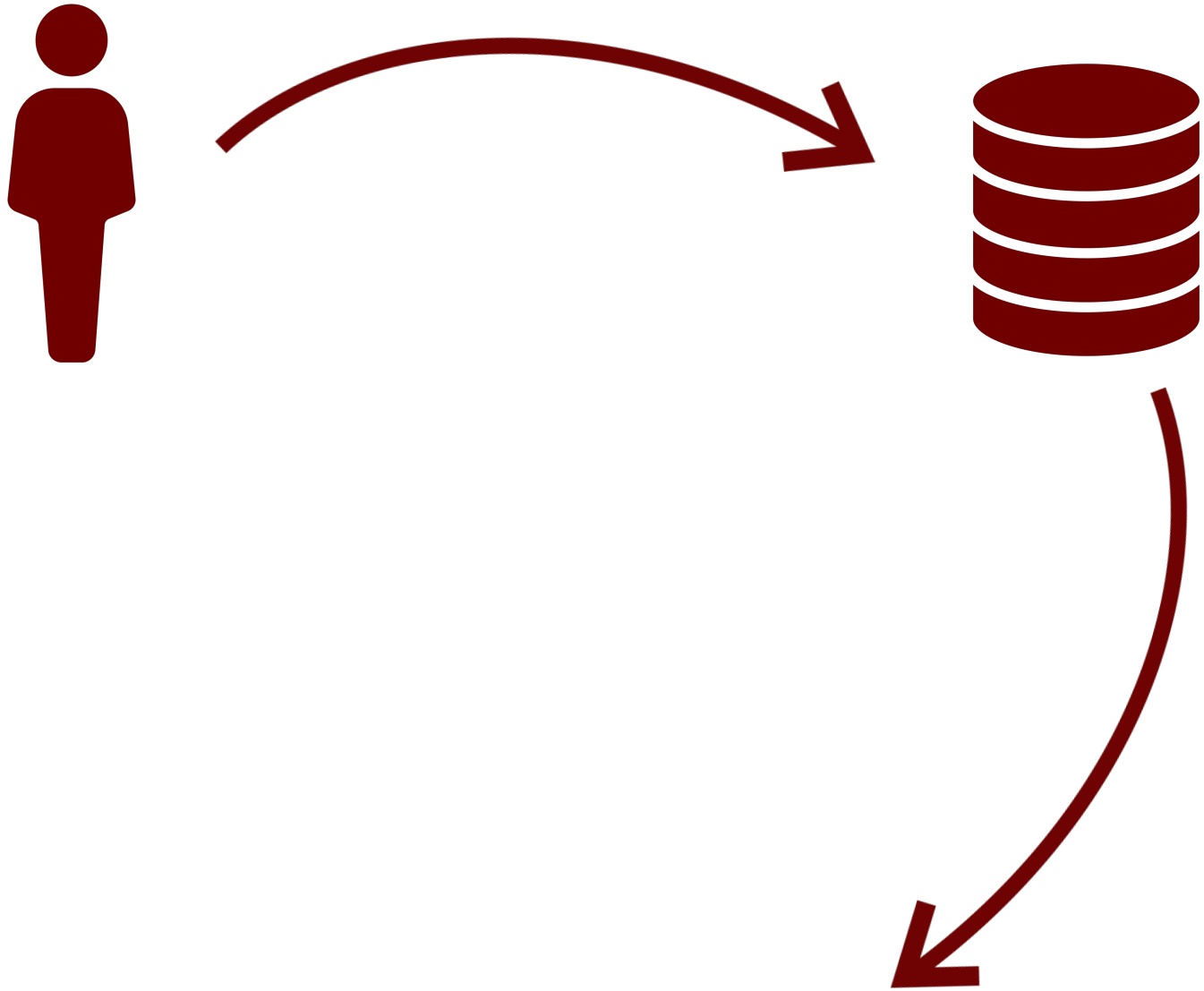

CHALLENGES
missing data
selection bias
sociotechnical context
TECHNIQUES
causal inference
nonparametric statistics
machine learning
OBJECTIVE
Oversight
Responsible decision making algorithms
Validity
Equity

My research

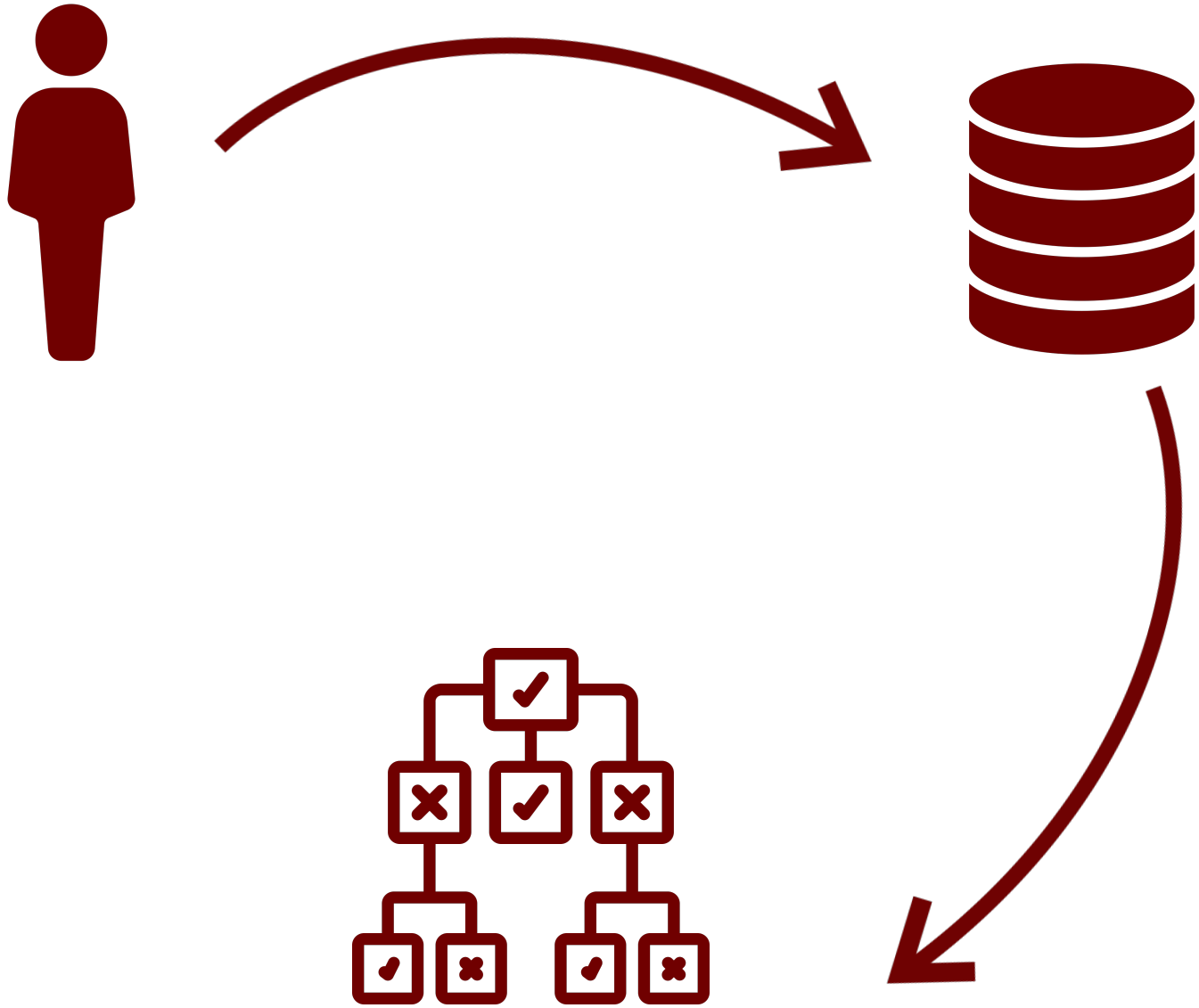

CHALLENGES
missing data
selection bias
sociotechnical context
TECHNIQUES
causal inference
nonparametric statistics
machine learning
OBJECTIVE
Oversight
Responsible decision making algorithms
Validity
Equity
criminal justice
child welfare
Credit lending
Health
DOMAINS

Future work

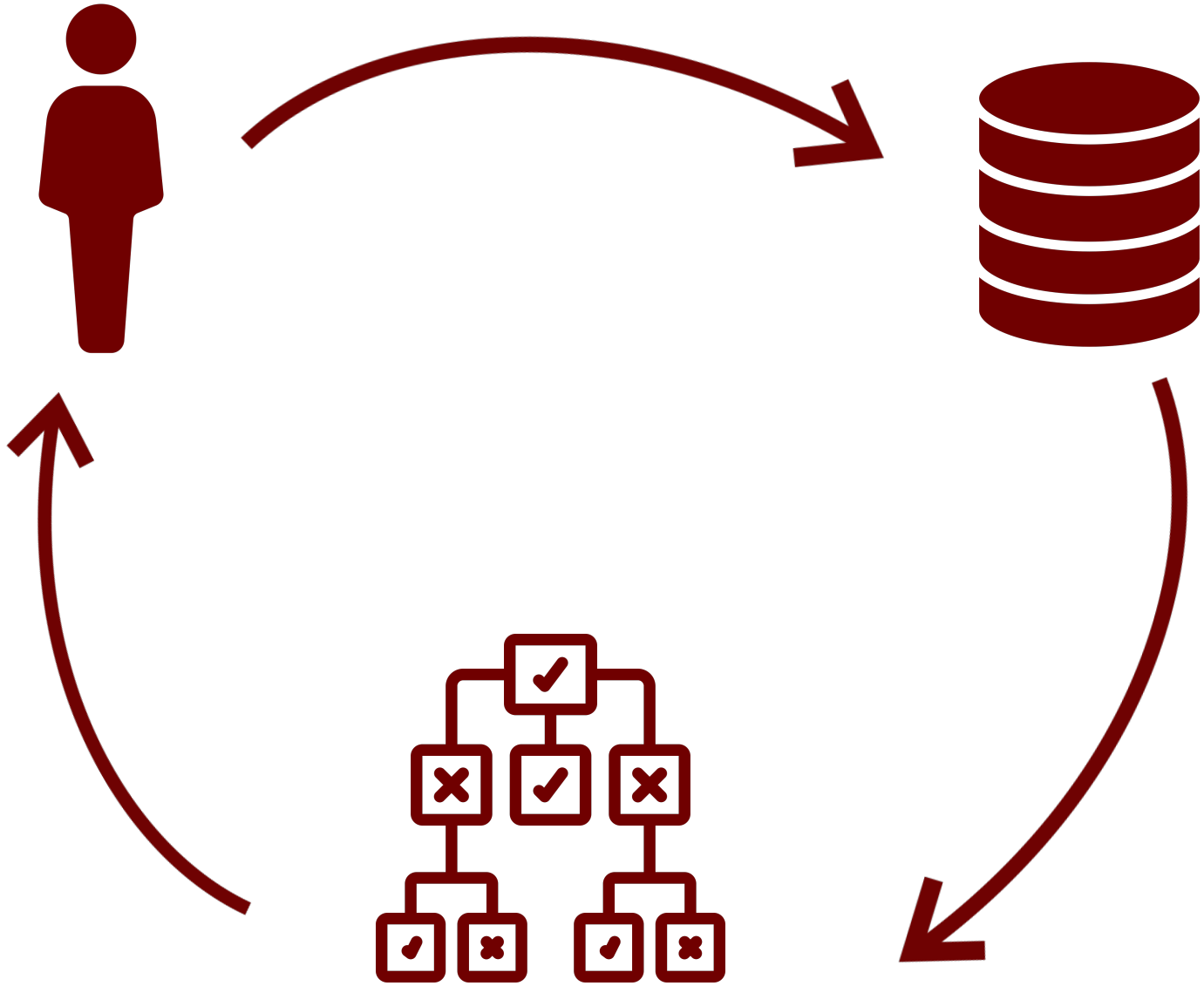
Future work



Future work

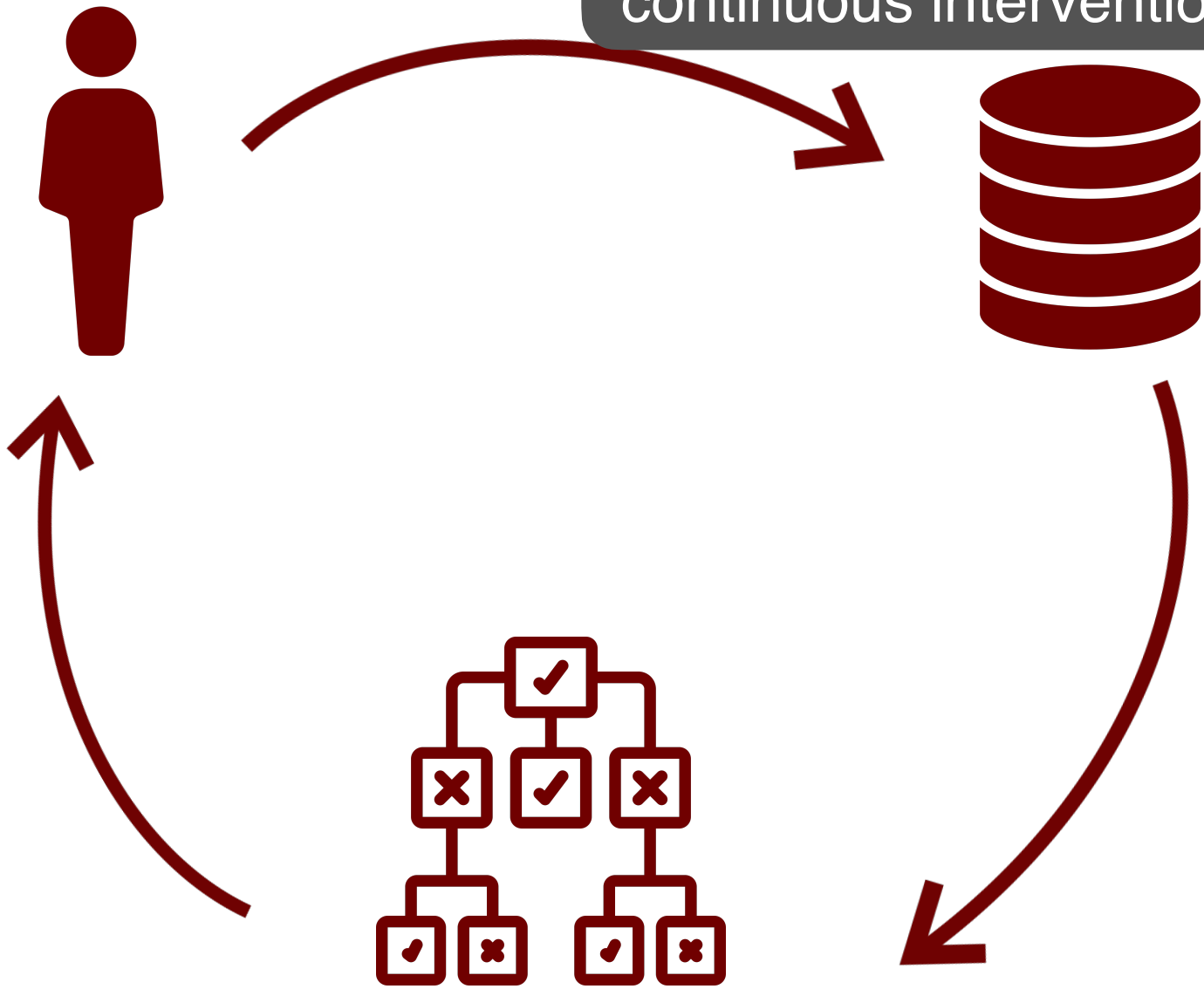


Future work



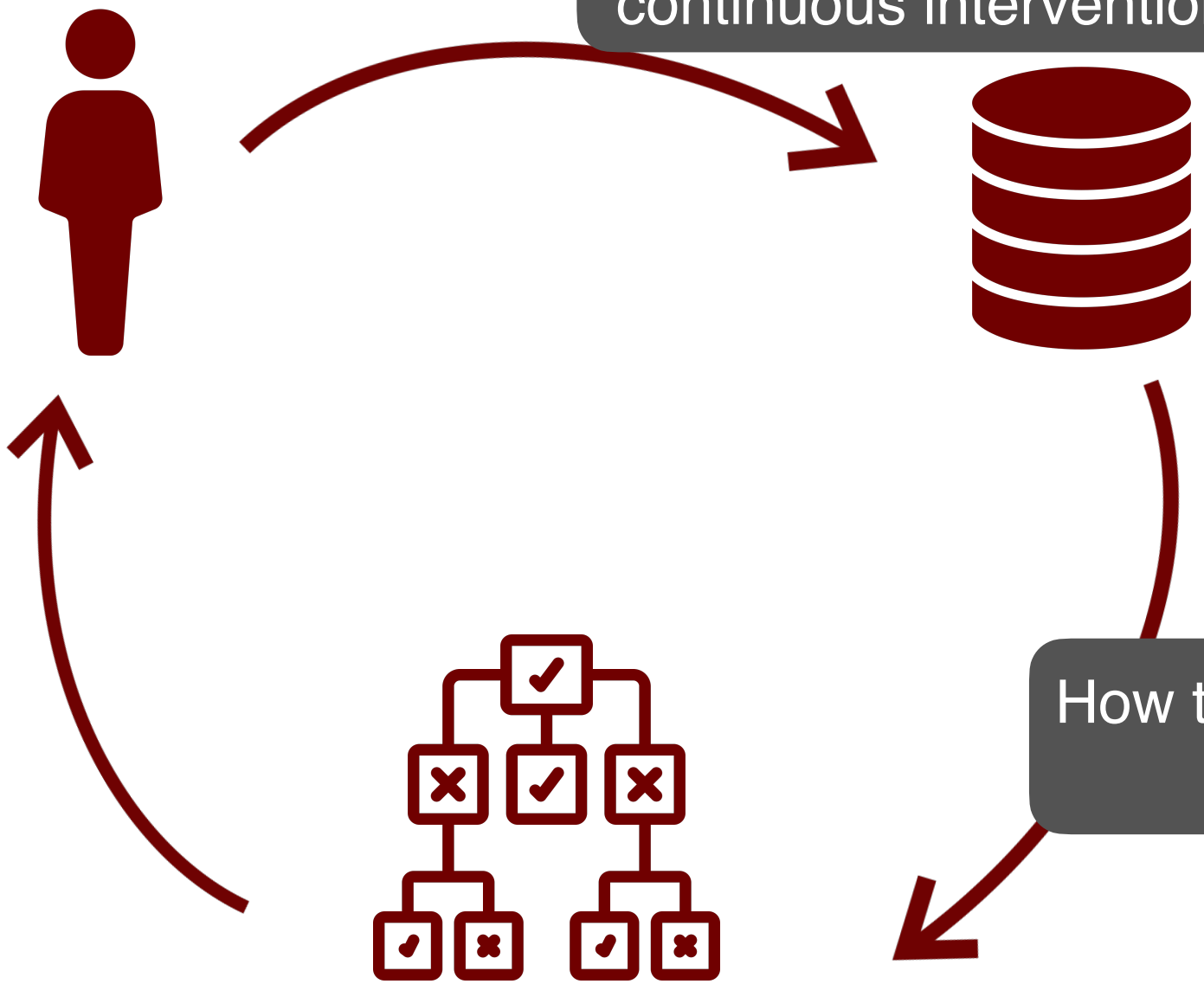
Future work

How to audit for bias under high-dimensional or continuous interventions?



Future work

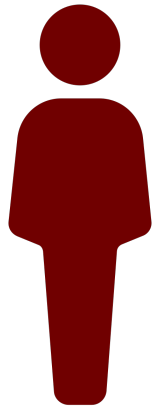
How to audit for bias under high-dimensional or continuous interventions?

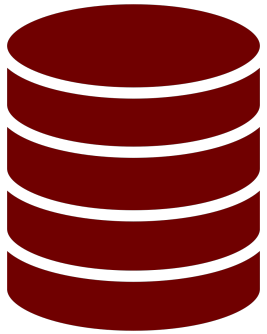


How to proceed given bias audit results?

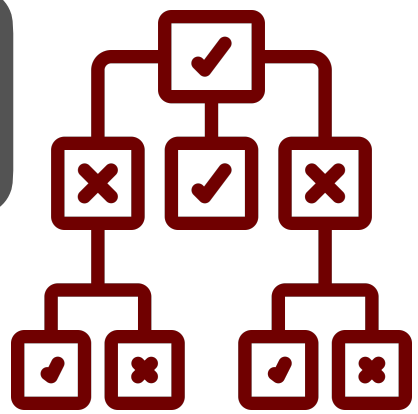
Future work

How to audit for bias under high-dimensional or continuous interventions?



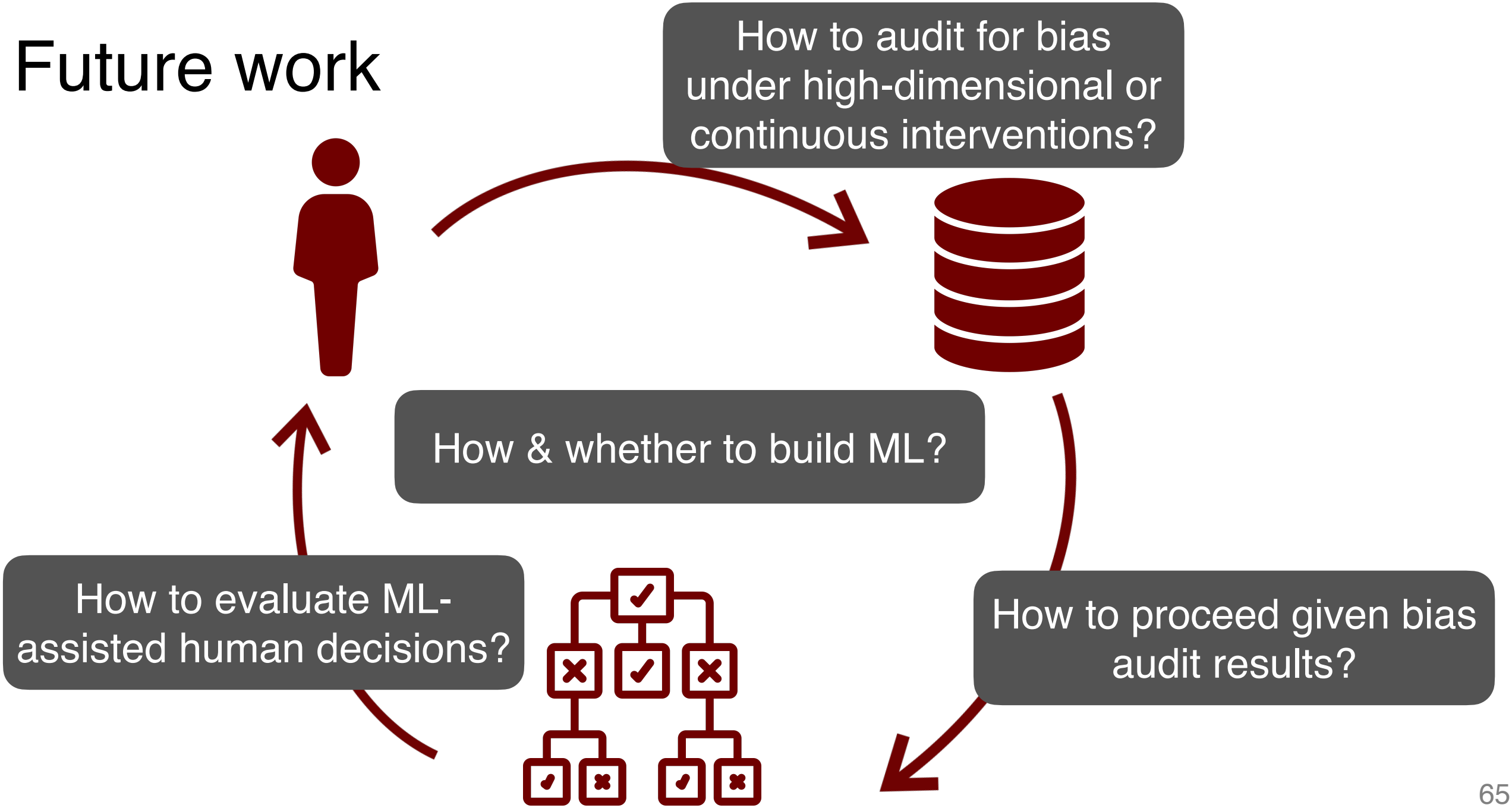


How to proceed given bias audit results?

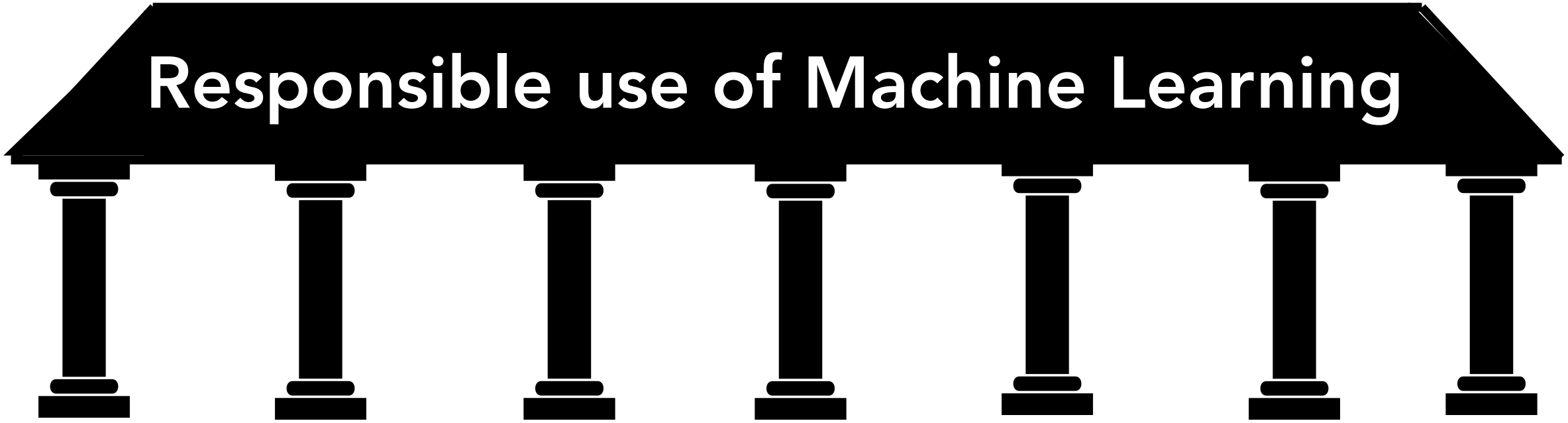


How to evaluate ML-assisted human decisions?

Future work

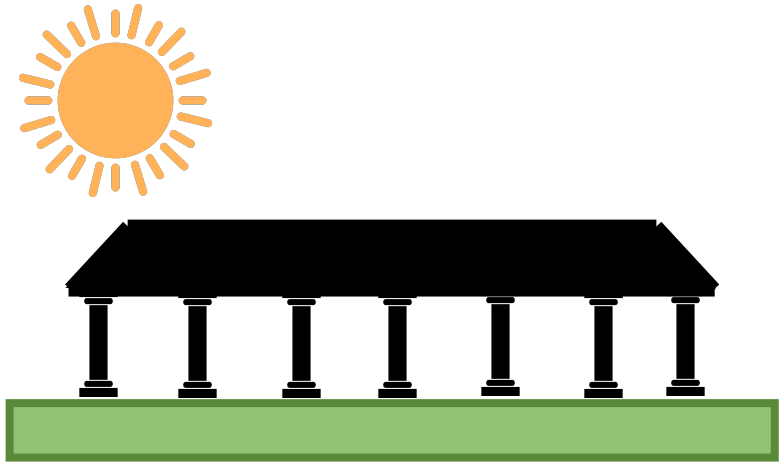

How to audit for bias under high-dimensional or continuous interventions?
How & whether to build ML?
How to proceed given bias audit results?
How to evaluate ML-assisted human decisions?

Vision

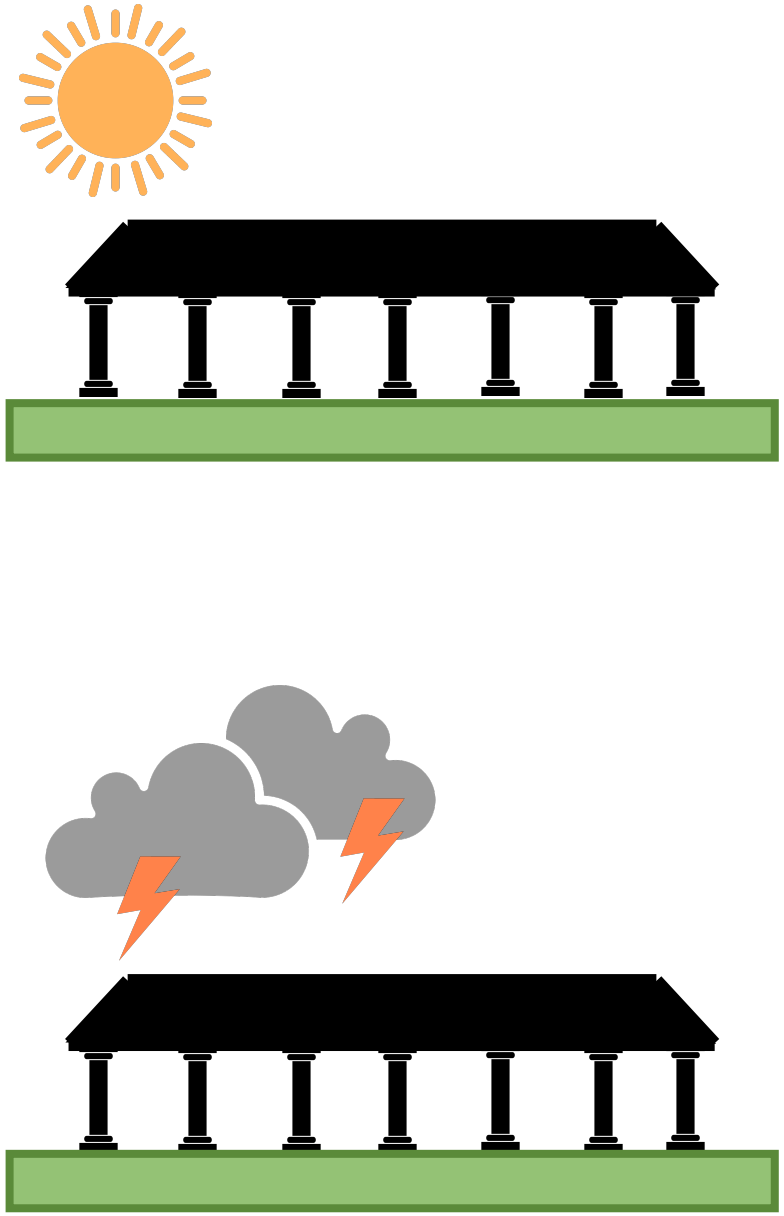

Responsible use of Machine Learning
Transparency
Accountability
Equity
Validity
Oversight
Robustness
Privacy

Vision

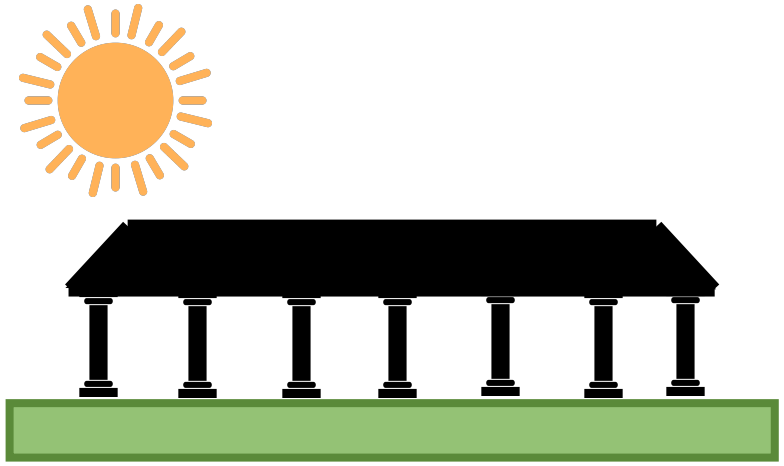
Vision

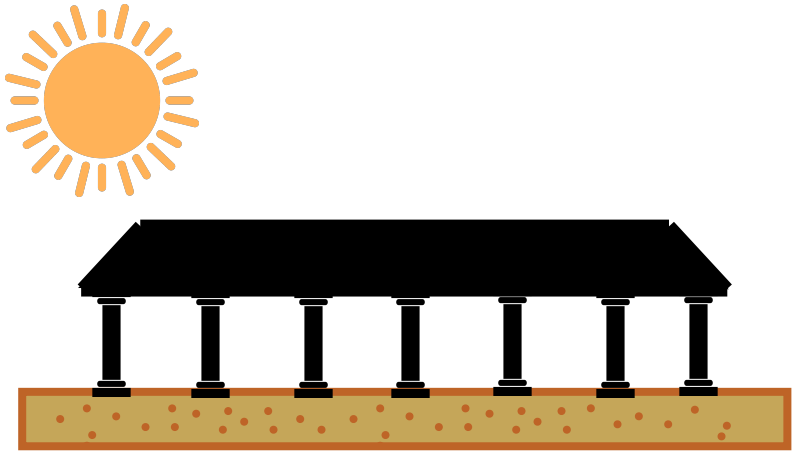


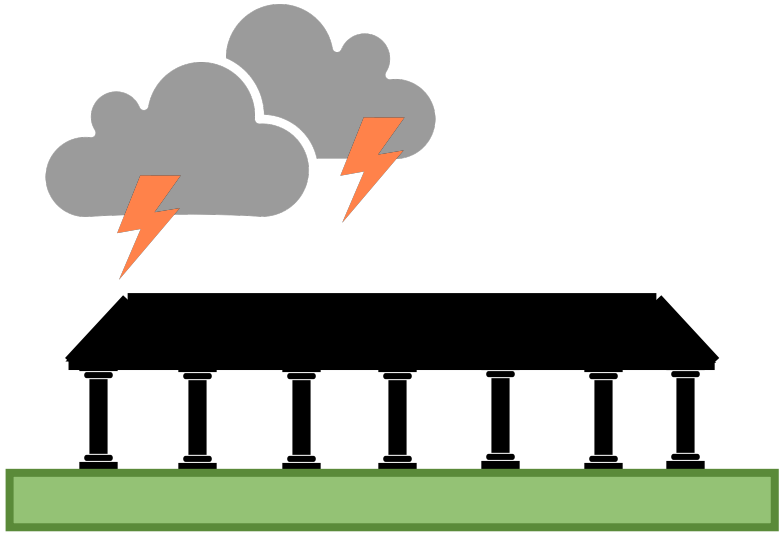
Vision



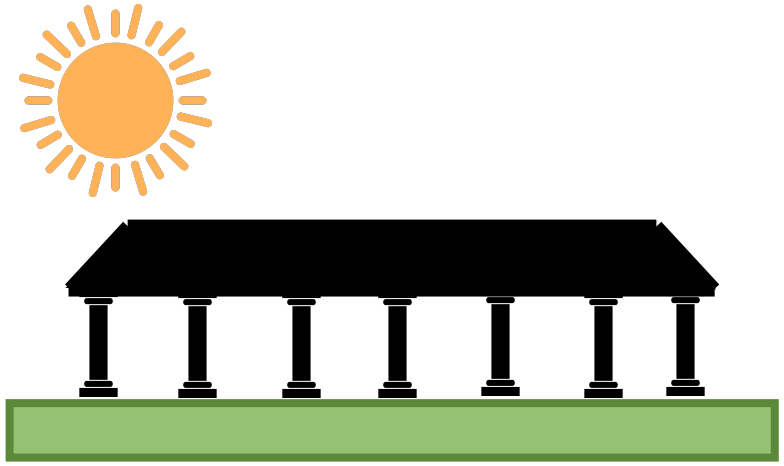
Vision

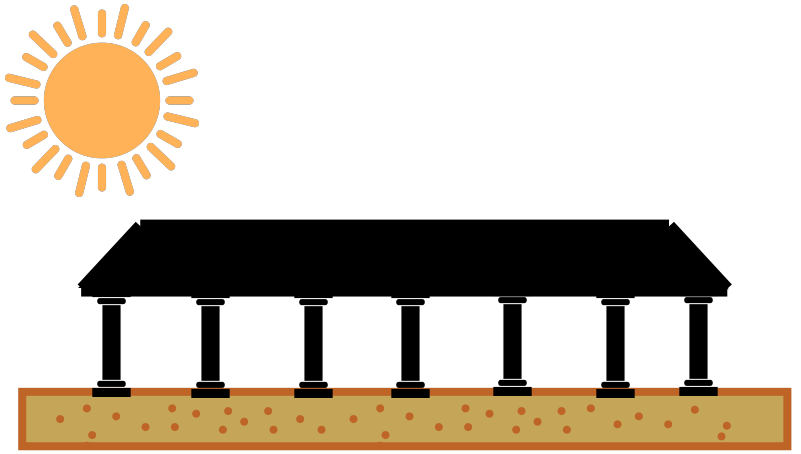


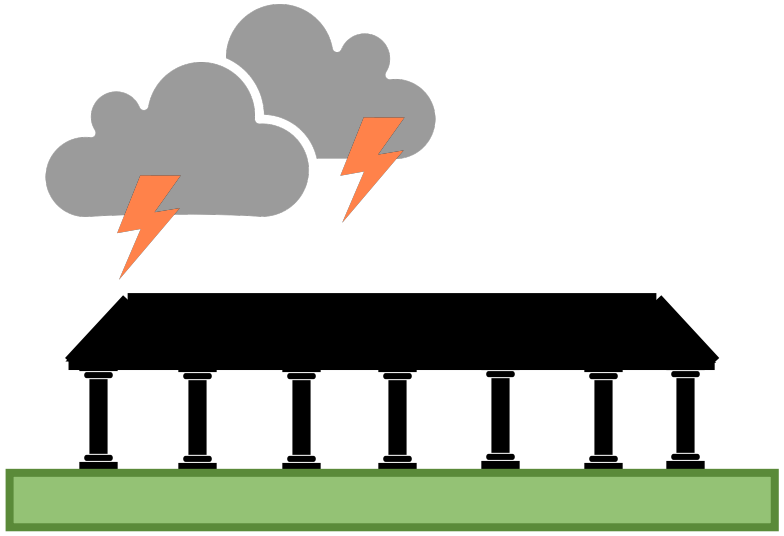




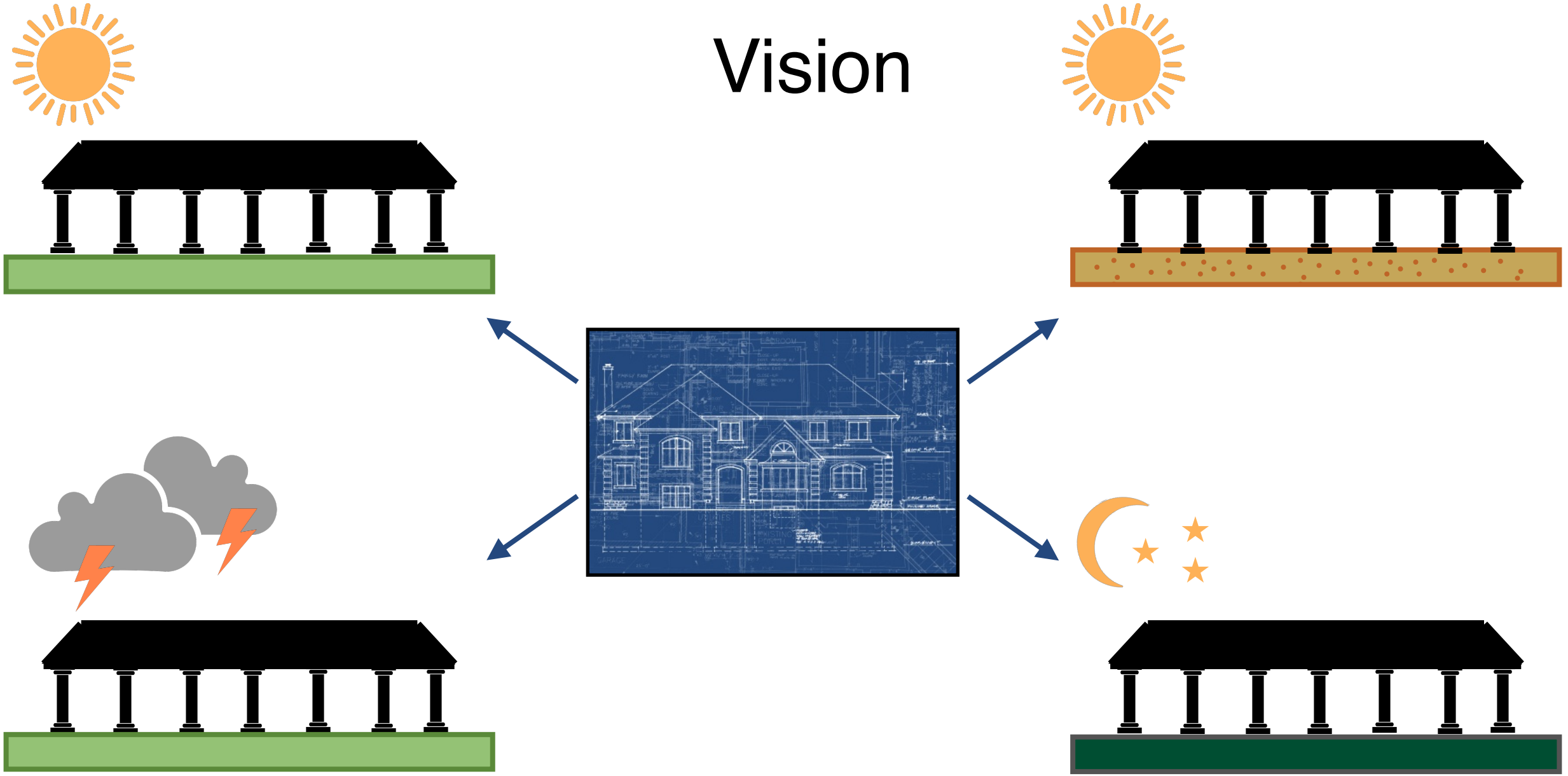
Vision



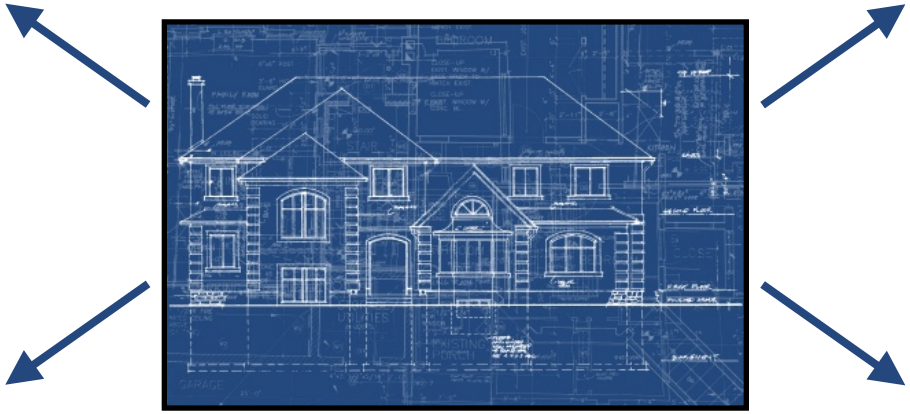




Vision



Vision

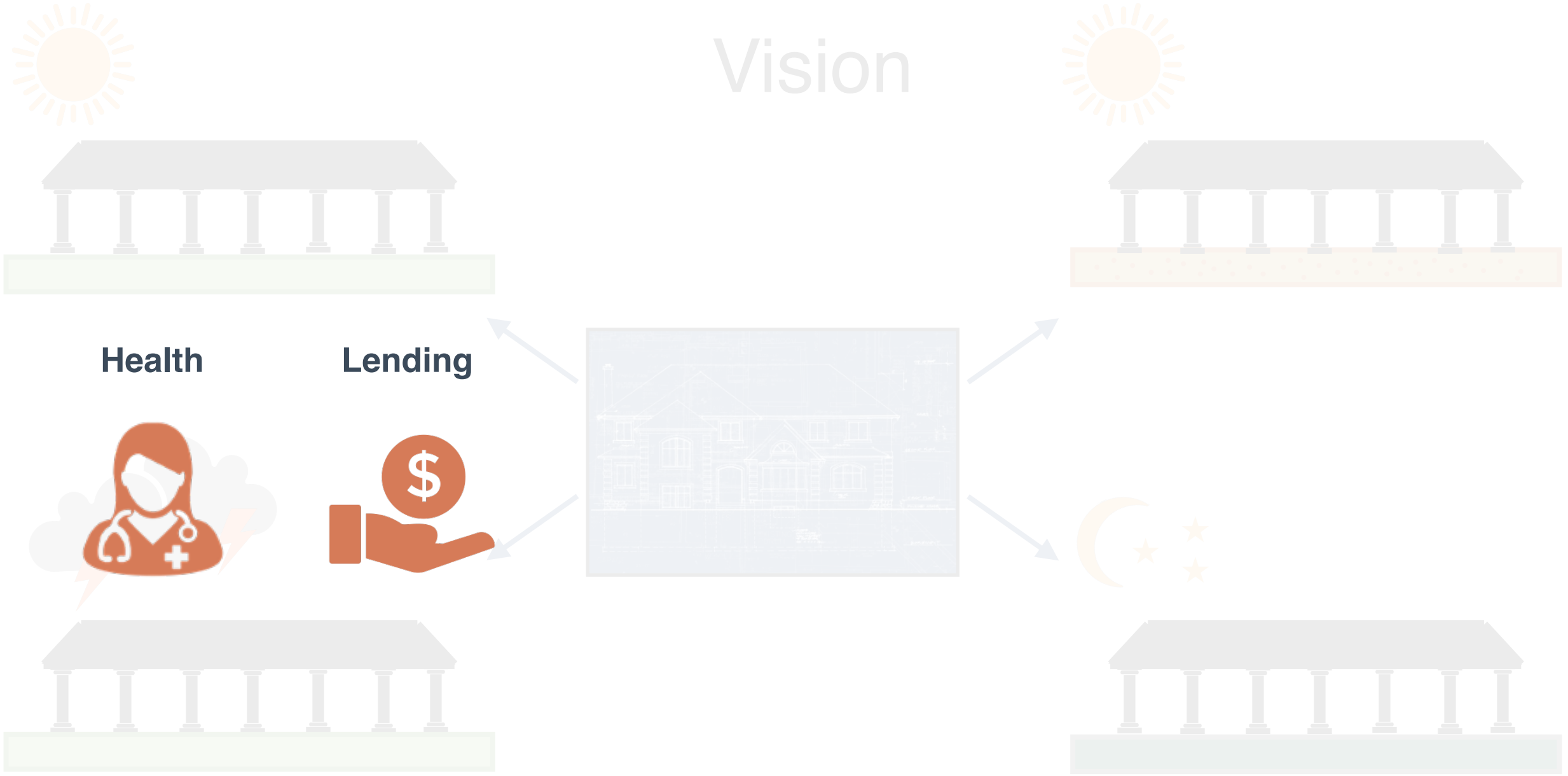


Vision

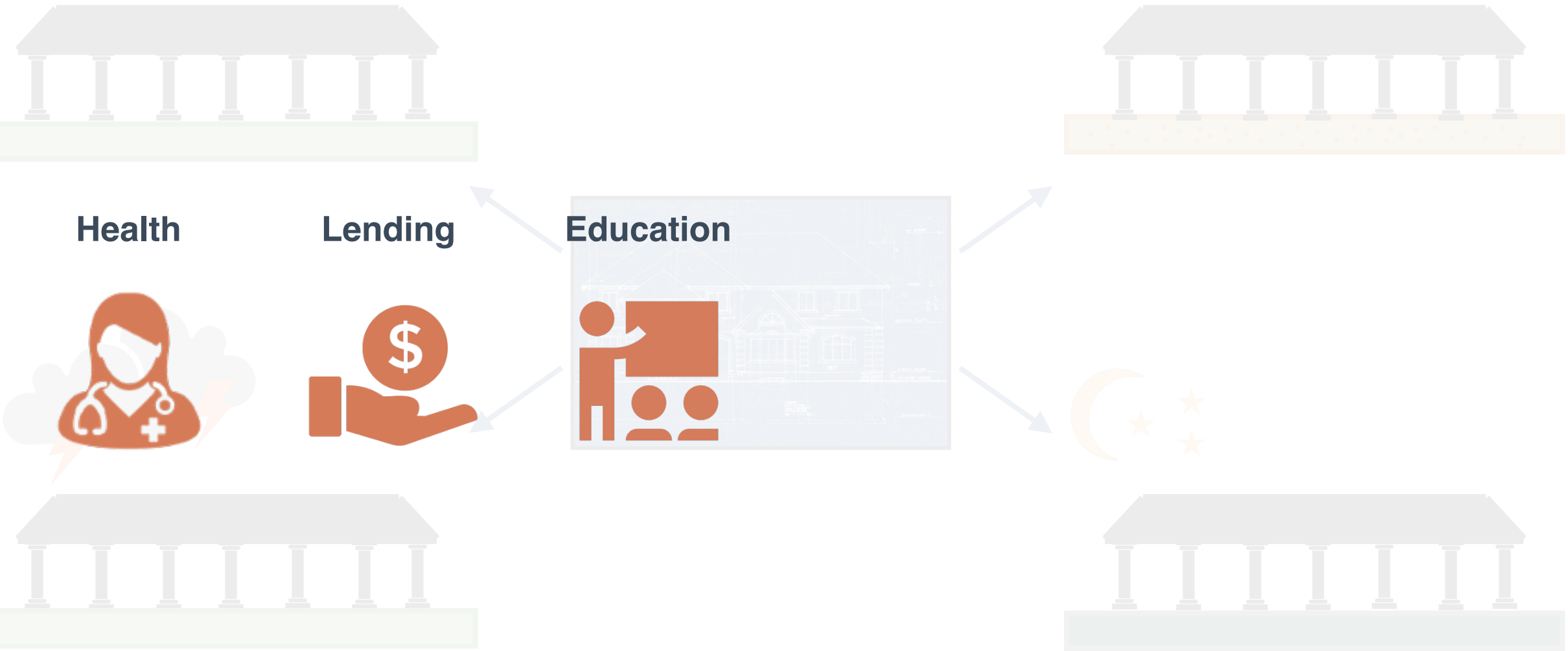
Vision

Health

Vision



Vision



Vision

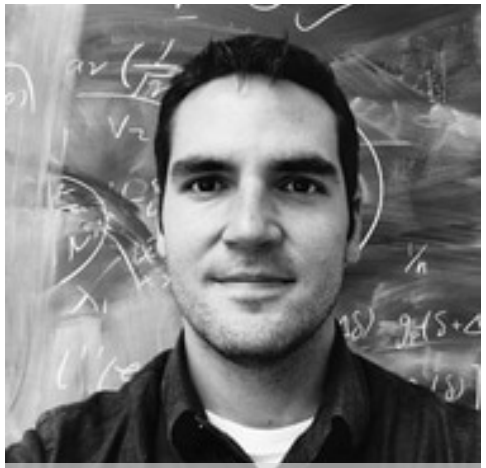
Vision

Vision

Thank you!

Acknowledgments

Alex Chouldechova



Edward H. Kennedy

Ashesh Rambachan

Alan Mishler


CommonwealthBank


Allegheny County
Department of Human Services

Questions?

Evaluating predictive models

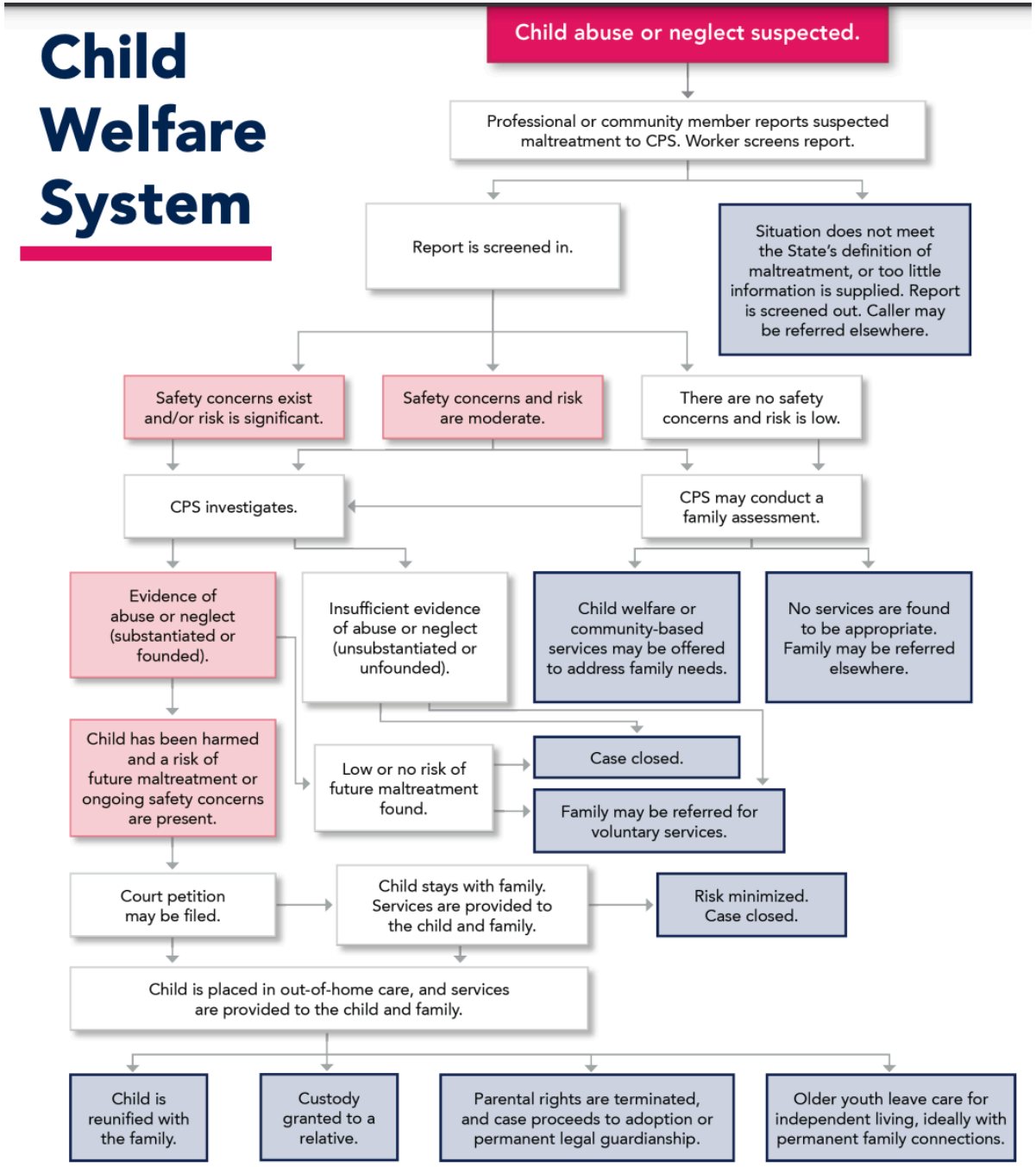
- **Problem:** Missing data threatens validity
- **Solution:** Counterfactual evaluation

Auditing for biased decision-making

- **Problem:** No ground truth data to assess if disparities are justified
- **Solution:** Counterfactual audit

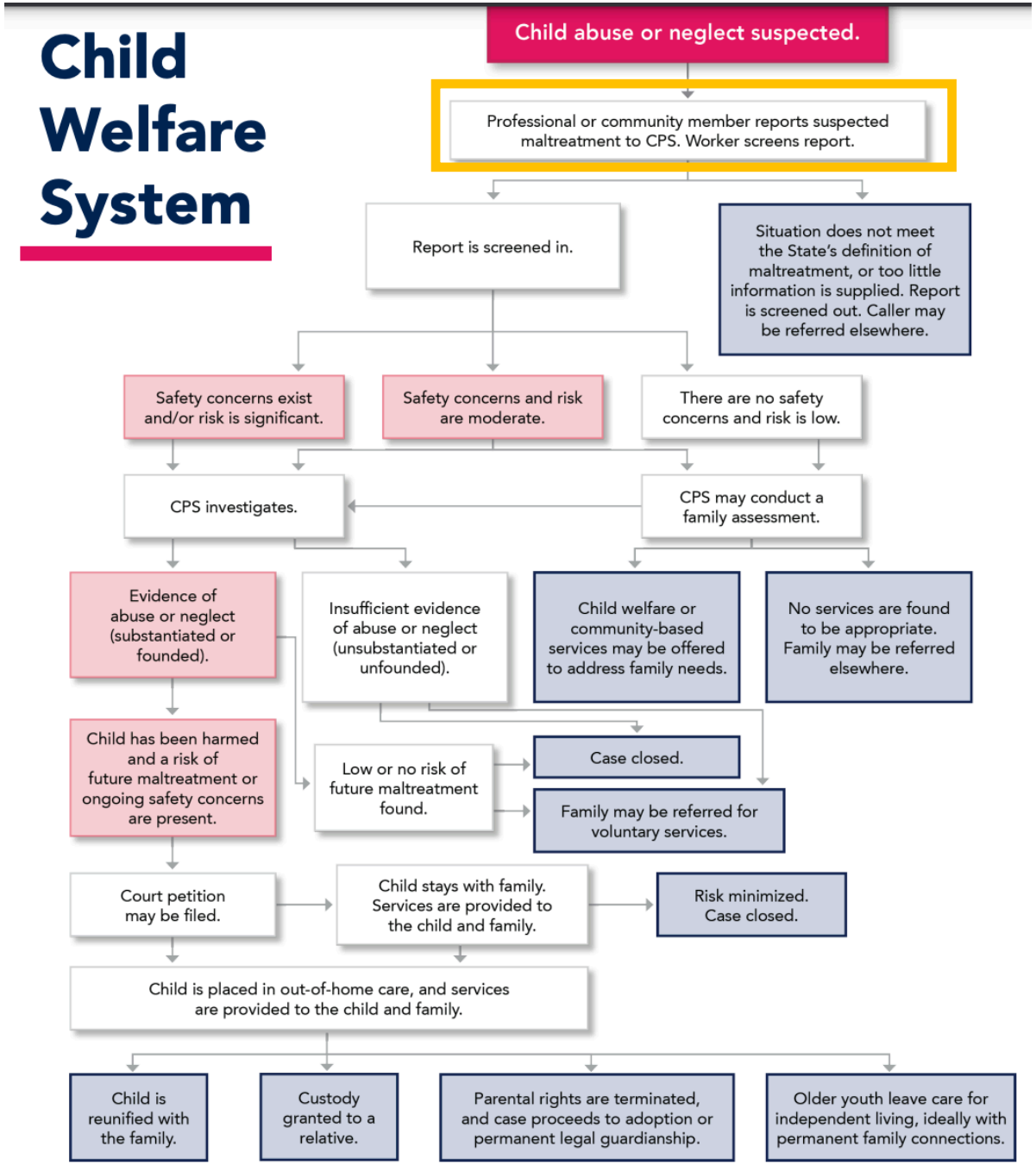
Appendix

Child Welfare System



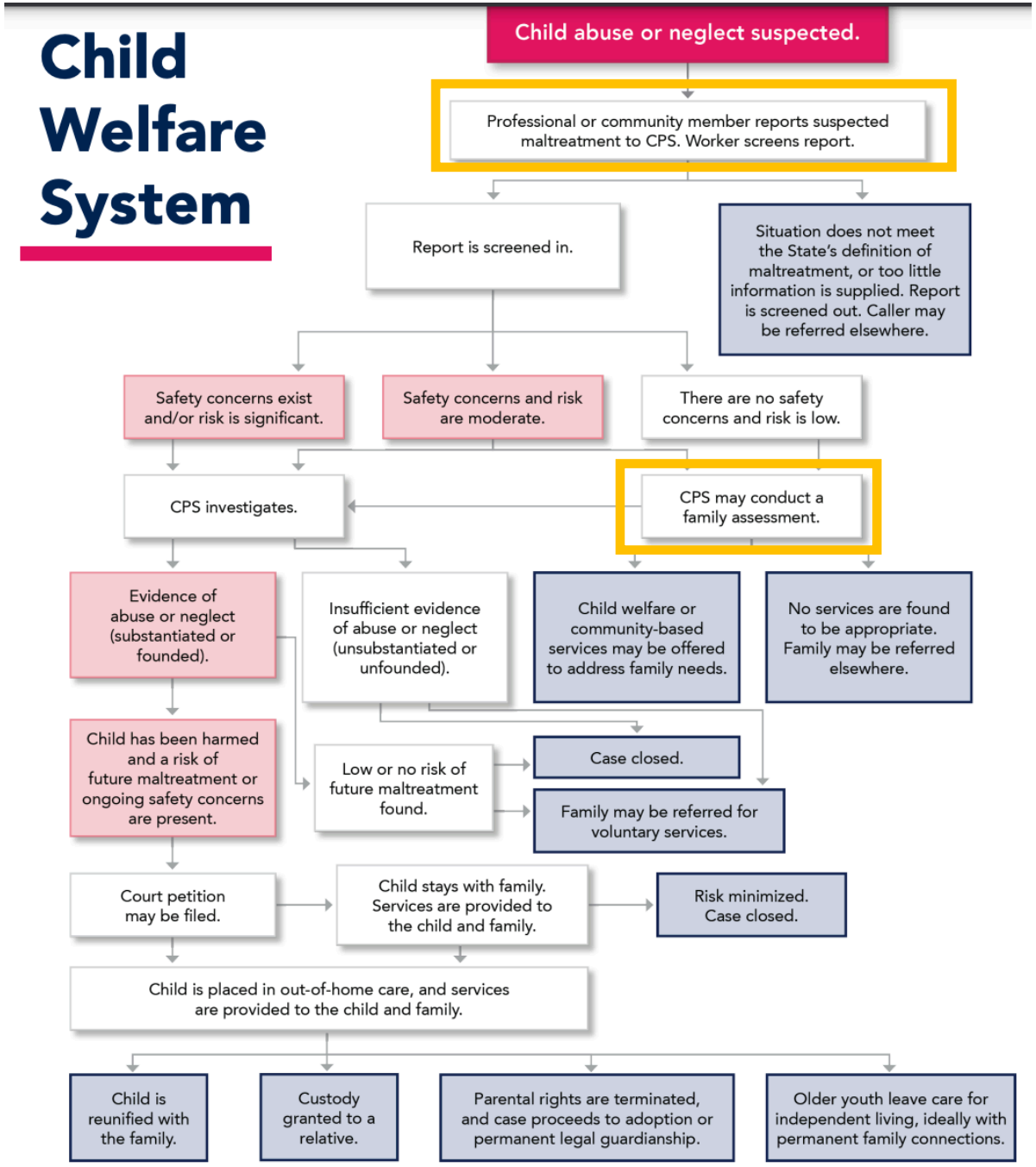
https://www.childwelfare.gov/pubpdfs/cpswork.pdf

Child Welfare System



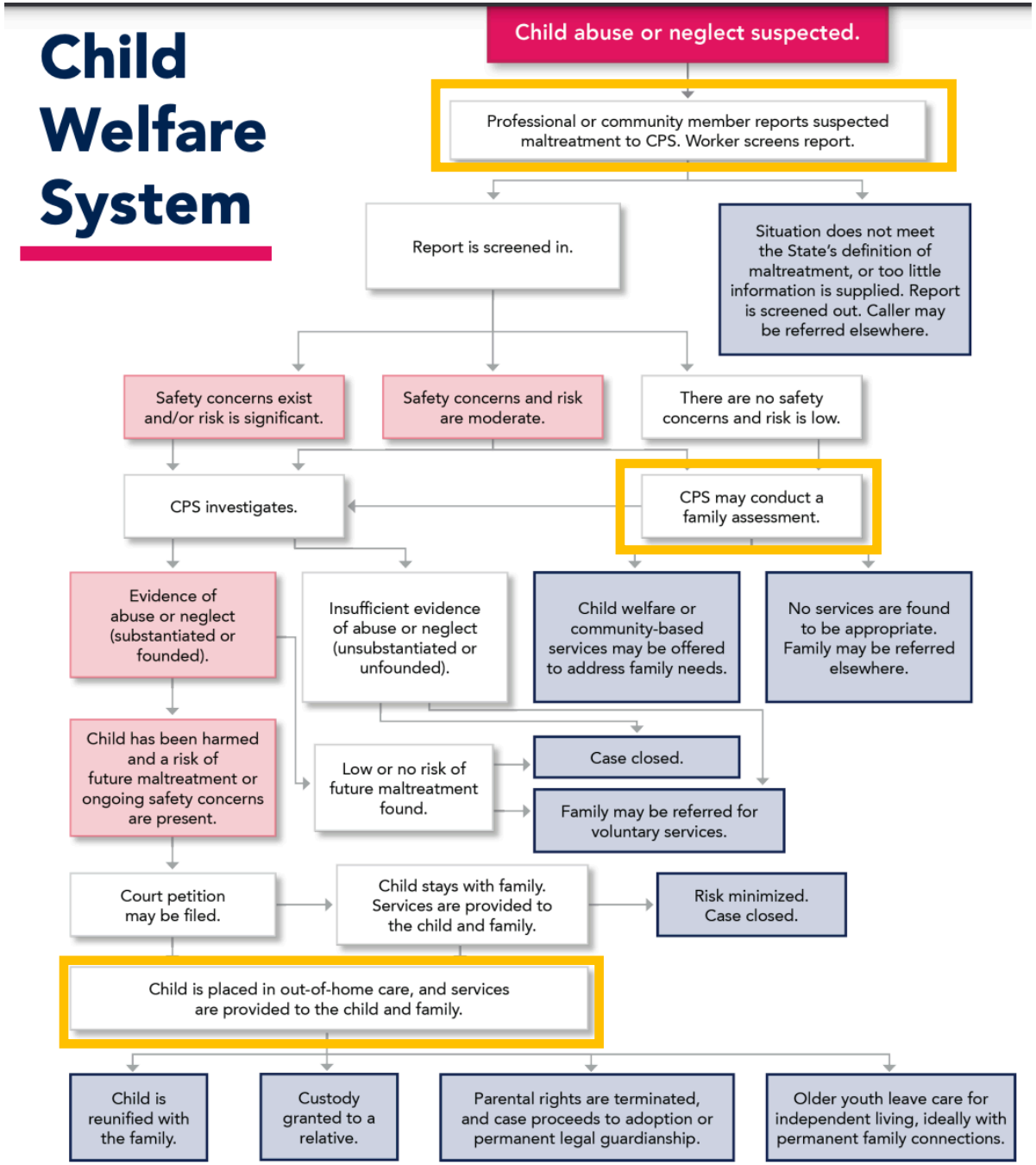
https://www.childwelfare.gov/pubpdfs/cpswork.pdf

Child Welfare System



https://www.childwelfare.gov/pubpdfs/cpswork.pdf

Child Welfare System



https://www.childwelfare.gov/pubpdfs/cpswork.pdf

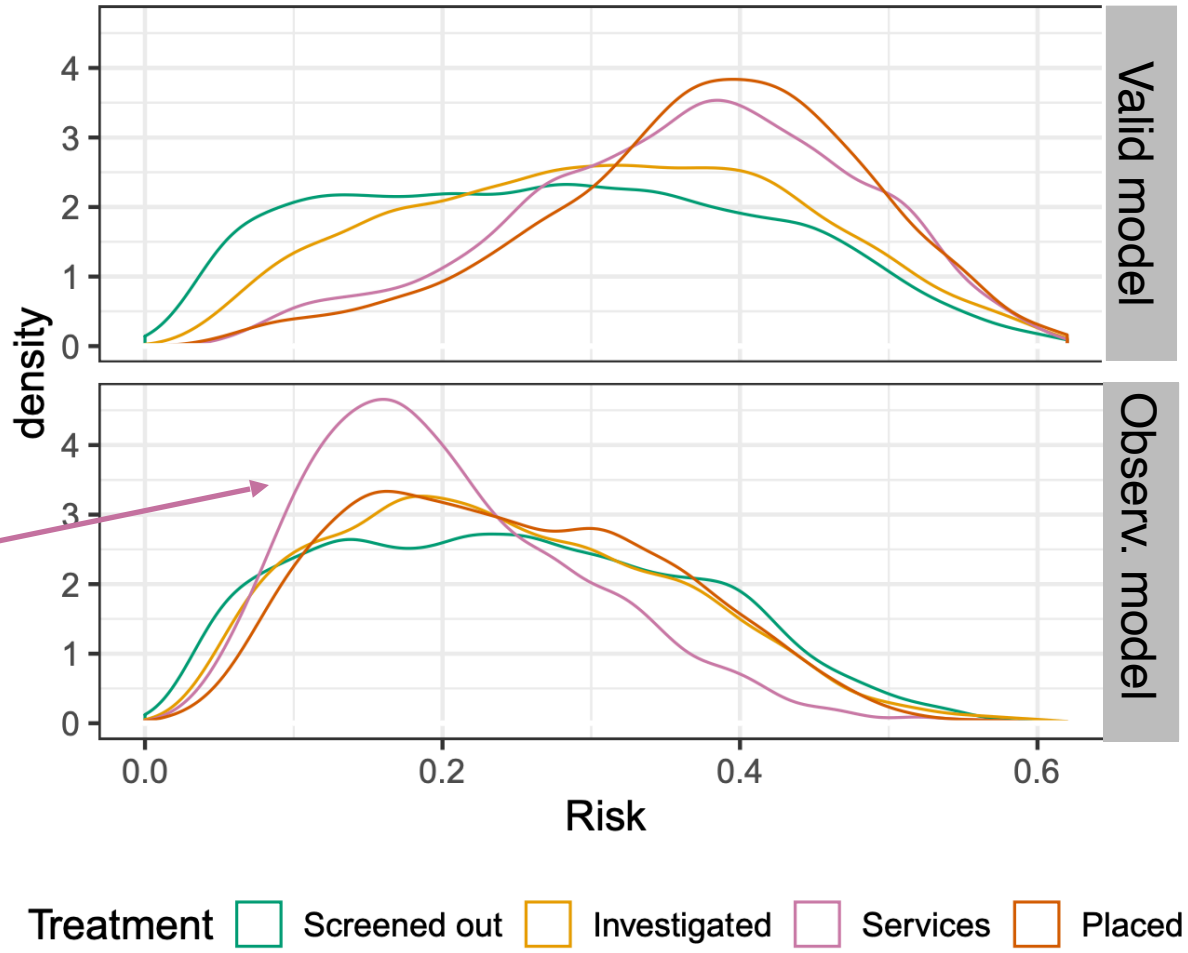
Child welfare hotline data

- 35,885 calls in Allegheny County, PA, from 2010-2014
- Risk and danger ratings based on allegation
- Criminal justice & child welfare history, substance abuse, demographic information, allegation type (e.g. neglect, malnutrition, sexual assault)
- Outcome $Y \in \{0,1\}$: re-referral in a six month period
- Observational re-referral rate $E[Y]$= 0.25
- Counterfactual re-rereferral rate $\mathrm{E}\left[Y^0\right]$= 0.30
- 62.4% black cases are investigated
- 52.5% white cases are investigated

Child welfare hotline data

- 35,885 calls in Allegheny County, PA, from 2010-2014
- Risk and danger ratings based on allegation
- Criminal justice & child welfare history, substance abuse, demographic information, allegation type (e.g. neglect, malnutrition, sexual assault)
- Outcome $Y \in \{0,1\}$: re-referral in a six month period
- Observational re-referral rate $E[Y]$= 0.25
- Counterfactual re-rereferral rate $\mathrm{E}\left[Y^0\right]$= 0.30 if no one had been investigated.
- 62.4% black cases are investigated
- 52.5% white cases are investigated

Predictive models for child welfare screening

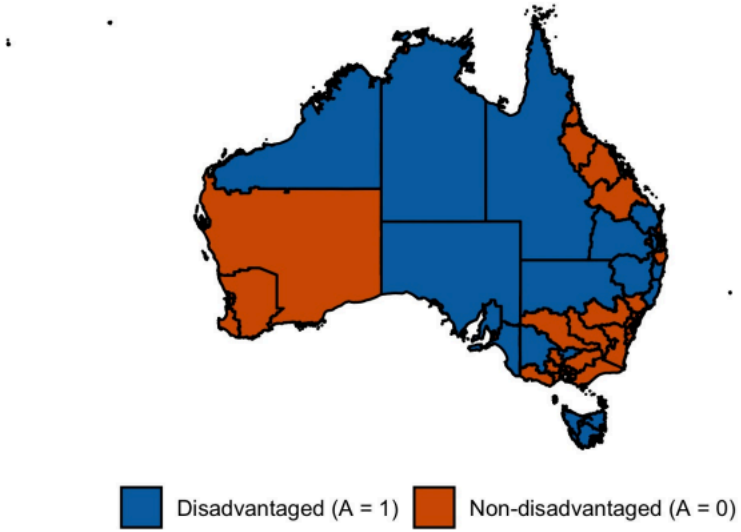


Those who received (beneficial, risk-mitigating) services in the past have the lowest predicted "risk"

Similar results shown for healthcare setting in
Caruana, R., et al. Intelligible models for healthcare: Predicting pneumonia risk. *KDD* 2015.

Commonwealth Bank of Australia

- 7414 personal loan applications submitted from July 2017 to July 2019
- credit score, reported income, approval decision, terms offered, repayment
- 44.9% of applications were funded
- 2.0% of funded loans defaulted within 5 months.



- 21.7% of loan applications from disadvantaged regions
- 19.7 % of funded loans from disadvantaged regions

Figure 2: SA4 regions in Australia. We classify SA4 regions as being "socioeconomically disadvantaged" (red) and "non-socioeconomically disadvantaged" (blue) based on the Index of Relative Socioeconomic Disadvantage (IRSD).

Stanford Open Policing traffic stop data

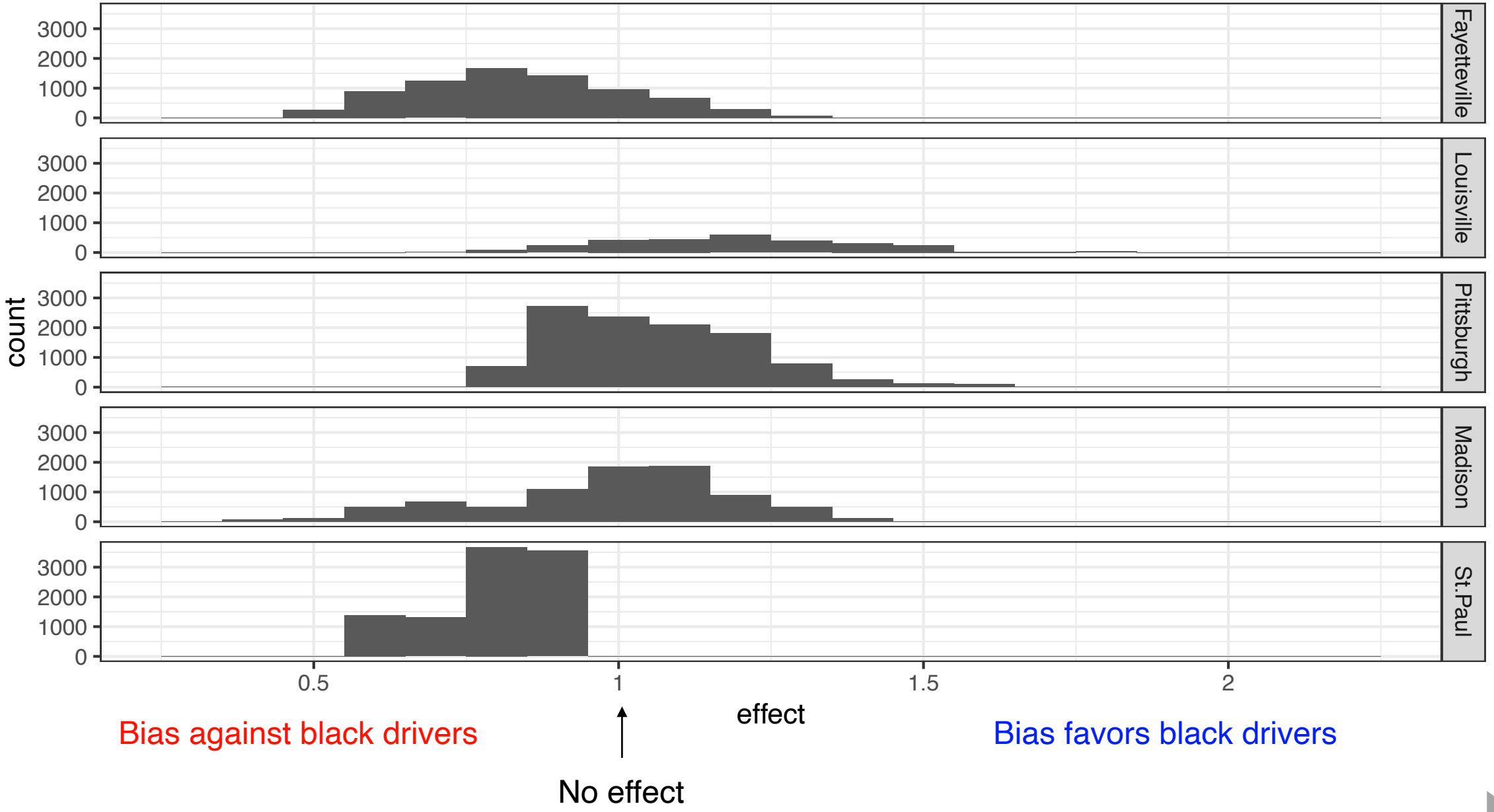
- 100 million police traffic stops occurring between 1999-2020
- From 21 state patrol agencies and 35 municipal police departments
- Driver & officer demographics
- Time & place of stop
- Reason for stop, citation issued, arrest made
- Missingness in rows over time and in features (e.g., place not recorded for some jurisdictions)

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., ... & Goel, S. (2020). A large-scale analysis of racial disparities in police stops across the United States. *Nature human behaviour*, *4*(7), 736-745.

Empirical analysis

- Data from Stanford Open Policing Project[1] over 10-year period
- Random variation: whether stop occurs before/after DST starts (spring) or ends (fall)
- $X =$ {time of day, day of week, fall/spring}
- $B = 1$ indicates perceived race is black; otherwise white
- 5 p.m.- 8 p.m., excluding stops between sunset & dusk

[1] Pierson et al. (2020). A large-scale analysis of racial disparities in police stops across the United States. Nature human behaviour, 4(7), 736-745.


Fayetteville
Louisville
Pittsburgh
Madison
St.Paul
count
3000
2000
1000
0
0.5
1
1.5
2
effect
Bias against black drivers
Bias favors black drivers
No effect

Notation

- Outcome $Y \in \{0,1\}$ indicates a re-referral to hotline
- Decision D
- Counterfactual target outcome Y^d
  - outcome we would observe if we set $D = d$
  - $Y^{screenout}$
- Algorithmic prediction $\hat{Y}$
- Covariates X

Notation

- Functional L_2 norm $\|f\|^2 := \int (f(x))^2 dP(x).$

Non-collapsibility of odds-ratio

X	$P(B = 1 \mid X, S^1 = 1)$	$P(B = 1 \mid X, S^0 = 1)$	OR(X)	OR
Urban	$\frac{1}{6}$	$\frac{9}{10}$	$\frac{1}{45}$	$\frac{3}{2} \times \frac{1}{45}$
Rural	$\frac{1}{26}$	$\frac{9}{14}$	$\frac{1}{45}$	

Aggregation via geometric mean

$$\Psi := \prod_{x \in \mathcal{X}} \left\{ \psi(x) \right\}^{dP(x|S=1)}$$

- Odds ratio is collapsible wrt geometric mean

X	$P(B = 1 \mid X, S^1 = 1)$	$P(B = 1 \mid X, S^0 = 1)$	OR(X)	OR	Geometric OR
Urban	$\frac{1}{6}$	$\frac{9}{10}$	$\frac{1}{45}$	$\frac{3}{2} \times \frac{1}{45}$	$\frac{1}{45}$
Rural	$\frac{1}{26}$	$\frac{9}{14}$	$\frac{1}{45}$		

X-conditional measure

$$\psi(X) := = \frac{\text{odds}(B = 1 \mid X = x, T = 1, S = 1)}{\text{odds}(B = 1 \mid X = x, T = 0, S = 1)}$$

Two types of counterfactual fairness

Ours: Counterfactual of the proposed decision $Y^{investigate}$ vs $Y^{screenout}$

- Purpose: address bandit feedback
- Intervention clearly defined

Kusner, Kilbertus, Chiappa, etc: counterfactual of the sensitive attribute Y^{female} vs Y^{male}

- Purpose: address pathways of gender bias in society
- Intervention is hard to define
- Challenge: identification

main

Validity

Method estimates the intended quantity we intended.

"Amelia Bedelia, the sun will fade the furniture.
I asked you to draw the drapes," said Mrs. Rogers.
"I did! I did! See," said Amelia Bedelia.
She held up her picture.

Peggy Parish & Fritz Siebel

Oversight

- Transparency
- Credible scrutiny

References

Coston, A., Mishler, A., Kennedy, E.H. and Chouldechova, A., 2020, January. Counterfactual risk assessments, evaluation, and fairness. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 582-593).

Coston, A. And Kennedy, E.H. 2022, May. Counterfactual audit for racial bias in police traffic stops. In American Causal Inference Conference (ACIC).

Coston, A., Rambachan, A. and Chouldechova, A., 2021, July. Characterizing fairness over the set of good models under selective labels. In International Conference on Machine Learning (pp. 2144-2155). PMLR.

Coston, A. and Kennedy, E.H., 2022. The role of the geometric mean in case-control studies. arXiv preprint arXiv:2207.09016.

References con’t

Schulam, Peter, and Suchi Saria. "Reliable decision support using counterfactual models." Advances in neural information processing systems 30 (2017).

Kallus, Nathan, and Angela Zhou. "Residual unfairness in fair machine learning from prejudiced data." In International Conference on Machine Learning, pp. 2439-2448. PMLR, 2018.

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R. and Goel, S., 2020. A large-scale analysis of racial disparities in police stops across the United States. Nature human behaviour, 4(7), pp.736-745.

Grogger, J. and Ridgeway, G., 2006. Testing for racial profiling in traffic stops from behind a veil of darkness. Journal of the American Statistical Association, 101(475), pp.878-887.