Bridging Prediction and Intervention Problems in Societal Systems

Lydia T. Liu (Princeton University)
Inioluwa Deborah Raji (UC Berkeley)
Angela Zhou (University of Southern California)

Warm-up : Automated Decision Systems (ADS)

Slido Quiz

Which of these applications of ADS are you familiar with?

A. **Criminal justice risk assessment**

B. **Clinical alerts, Clinical decision making**

C. **Education advising / early warning systems**

D. **Housing or social-service triage**

E. **Child welfare screening**

F. **Hiring or workplace screening**

Automated Decision Systems (ADS) are everywhere…

Criminal justice risk assessment

"Over 60% of the US population lives in a jurisdiction that uses a pretrial RAT"

Sepsis prediction / clinical alerts

"More than 80% of clinicians in the United States alone use MDCalc…"

Education advising / early warning

"tracking 800 different risk factors for more than 40,000 students every day, …" (Georgia State University, 2024)

Housing or social-service triage

VI-SPDAT, WPRS.

Child welfare screening

Allegheny Family Screening Tool

Hiring or workplace screening

"More than one million job candidates around the world…" (Pymetrics)

But does individual prediction have the intended impact?

Sentiment check: where do you think have we seen improvements in domain-relevant *outcomes* from ADS deployment?

A. **Criminal justice risk assessment**

B. **Clinical alerts, Clinical decision making**

C. **Education advising / early warning systems**

D. **Housing or social-service triage**

E. **Child welfare screening**

F. **Hiring or workplace screening**

Mixed (or lack of) evidence across domains

Criminal justice risk assessment

Mixed evidence on **reducing detainment rates** while ensuring **public safety** (Green, 2020)

Sepsis prediction / clinical alerts

Increased subsequent treatment and guideline **compliance** (Sendak et al, 2020), but also causes **alert fatigue** (Wong et al., 2021), and **only minor reductions in mortality** (Boussina et al., 2024).

Education advising / early warning

- ***Georgia State University's GPS Advising:*** "… more than 250,000 one-on-one meetings between our advisers and students that were prompted by alerts generated by the system… **7 percentage point increase in graduation rates** " (GSU, 2024)
- ***DEWS in Wisconsin Public Schools:*** "…accurately sorts students by their dropout risk… However, these effect estimates are noisy and we **cannot rule out the possibility that the system has had no impact on student outcomes.**" (Perdomo et al., 2023)

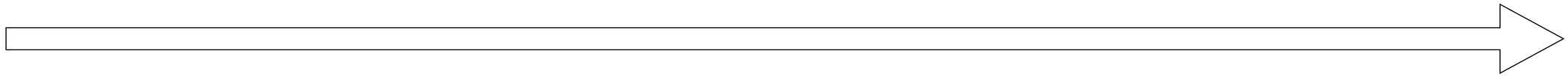
Goals of this tutorial

- Conceptualize **predictive Automated Decision Systems (ADS) as policy interventions** on institutional decision processes with broad consequences.
- Compare **model design** choices such as pure prediction, counterfactual baseline prediction, and individual treatment-effect targeting.
- **Evaluate** deployed ADS using experimental and quasi-experiment methods beyond benchmarking. Learn the limitations of benchmark evaluations.
- Understand the **implementation factors** including interaction design, intervention design, human discretion, organizational context, and governance, that impact deployment outcomes.

Road map

Overview: Predictions as Interventions

Model design	**Evaluation science**	**Implementation science**	**Tutorial Artifact: ADS deployment index**
What is being optimized?	What evidence shows deployment had the intended impact?	What implementation factors determine impact?	Apply prediction-as-intervention lens to deployments “in the wild”



As we move through the modules, we'll collect ingredients for today's community artifact.

A Multidisciplinary Community (2024-Present)

“Bridging Prediction and Intervention Problems in Social Systems”

Computer Science
Statistics
Operations
Business
Econ
HCI
Philosophy
Sociology
Law

arXiv 2507.05216

Organizers

Lydia Liu (Princeton University)

Inioluwa Deborah Raji (University of California - Berkeley)

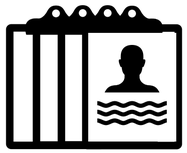
Angela Zhou (University of Southern California)

Arvind Narayanan (Princeton)

Bridging Prediction and Intervention Problems in Social Systems *

Lydia T. Liu†, Inioluwa Deborah Raji†, Angela Zhou†,
Luke Guerdan, Jessica Hullman, Daniel Malinsky, Bryan Wilder, Simone Zhang,
Hammaad Adam, Amanda Coston, Ben Laufer, Ezinne Nwankwo, Michael Zanger-Tishler,
Eli Ben-Michael, Solon Barocas, Avi Feller, Marissa Gerchick, Talia Gillis, Shion Guha,
Daniel Ho, Lily Hu, Kosuke Imai, Sayash Kapoor, Joshua Loftus, Razieh Nabi,
Arvind Narayanan, Ben Recht, Juan Carlos Perdomo, Matthew Salganik, Mark Sendak,
Alexander Tolbert, Berk Ustun, Suresh Venkatasubramanian, Angelina Wang, Ashia Wilson ‡

Status Quo: How We Think About Automated Decision Support



X

Covariates

R

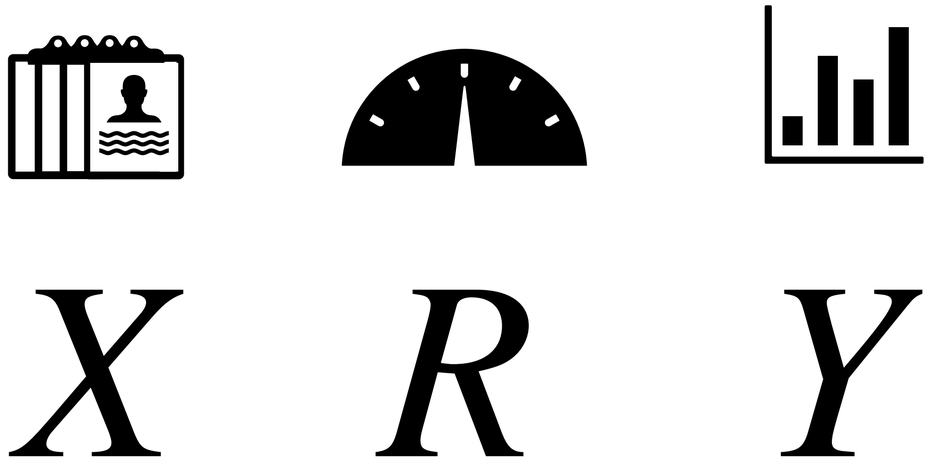
Risk Score

$E[Y \mid X]$

Y

Outcomes

Status Quo: How We Think About Automated Decision Support



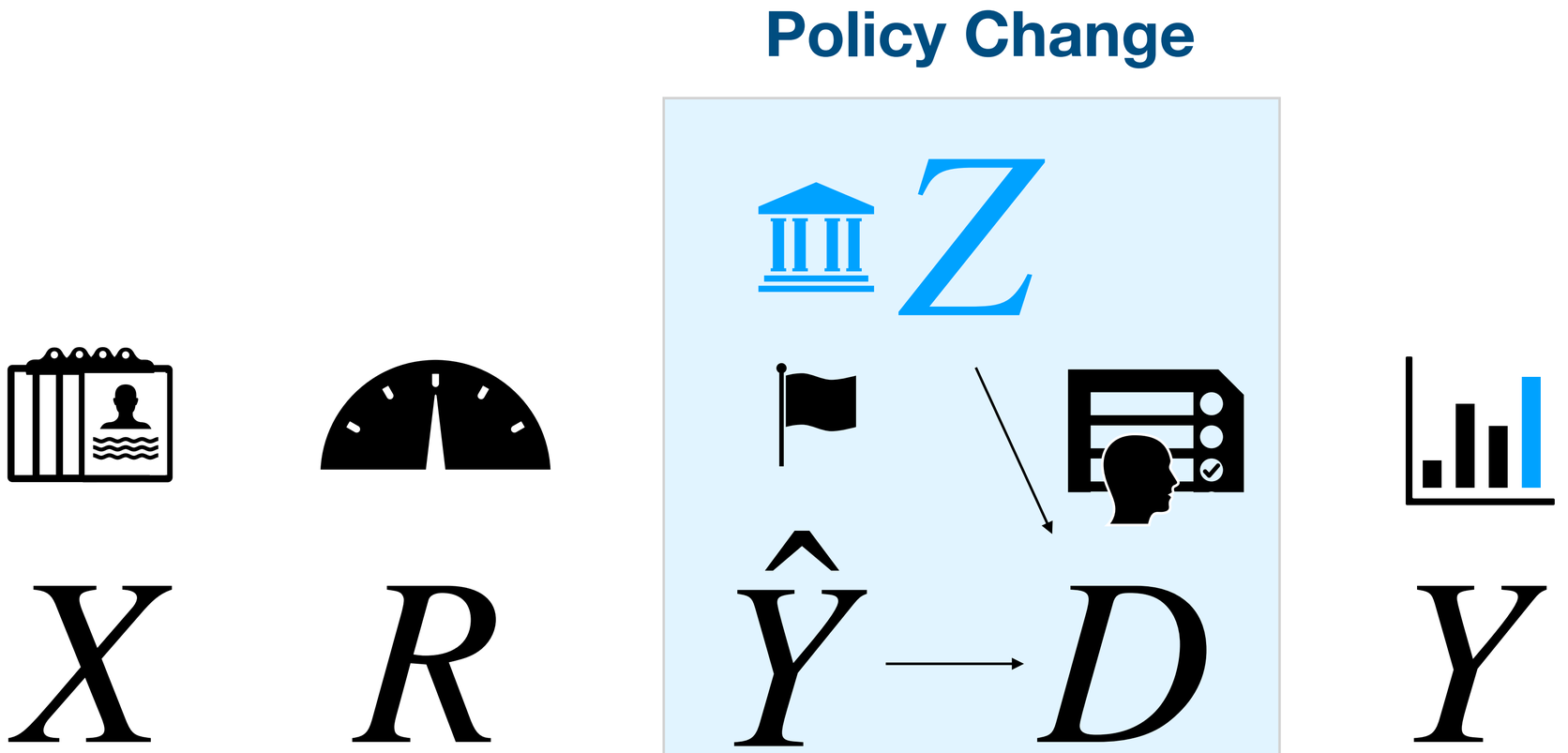
X R Y

Covariates Risk Score Outcomes

$$E[Y \mid X] = P(Y = 1 \mid X)$$

X	Y
Credit history	2-year default
Patient history, EHR data	Sepsis
Grades	2-year dropout
Background and prior criminal record	Failure to appear for trial

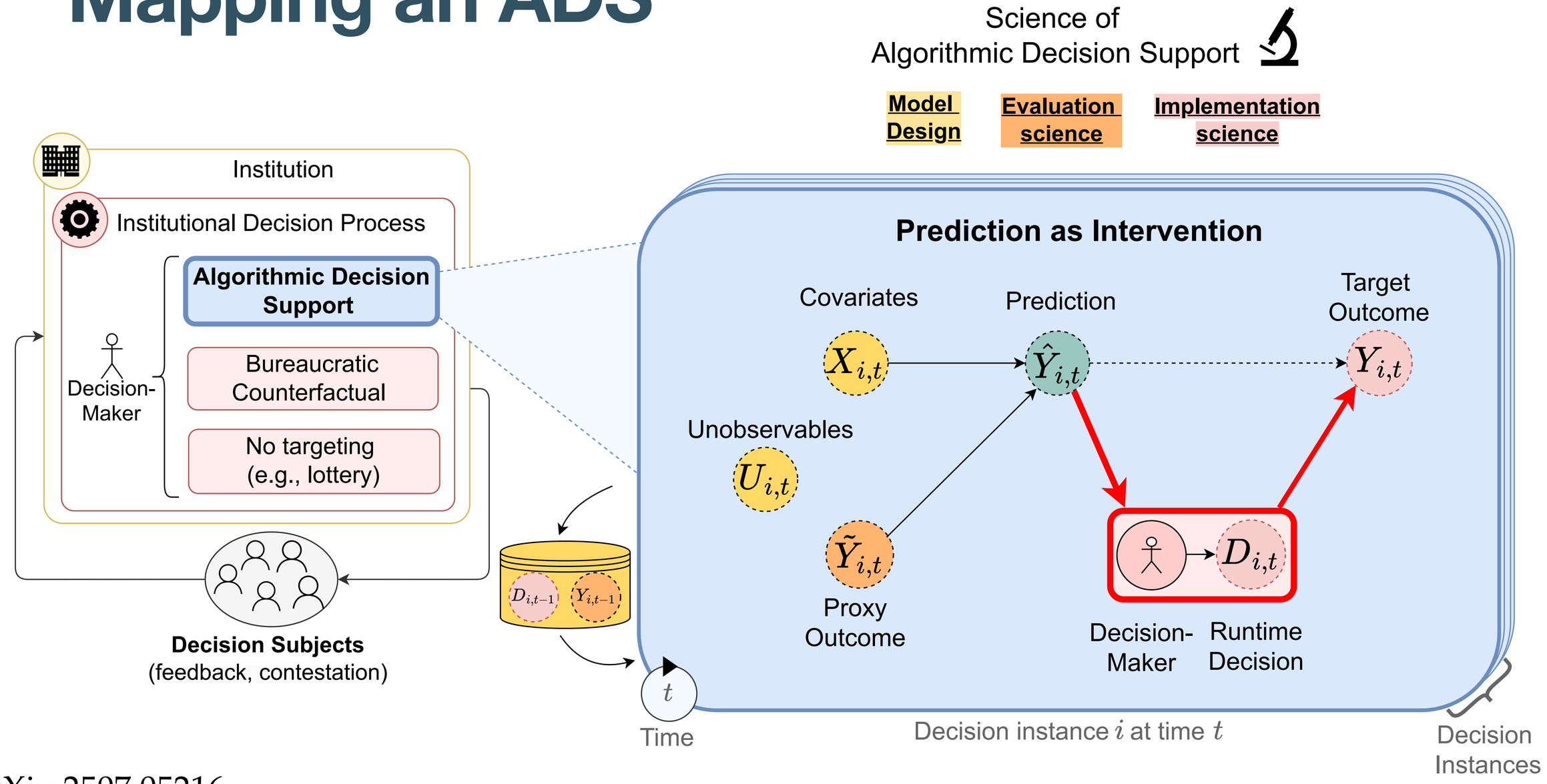
ADS Deployments as Policy **Interventions**



ADS Deployments as Policy **Interventions**

	X	D *(Decision)*	Y
Criminal justice	Background and prior criminal record	Pretrial detention? Supportive services?	Failure to appear for trial
Lending	Credit history	Approved? Interest rate?	2-year default
Health	Patient history, EHR data	Bed allocation Early attention	Sepsis
Education - Dropout prediction	Grades	Advising hours, attendance monitoring	2-year dropout

Mapping an ADS



arXiv 2507.05216

Build the Prediction-as-Intervention Index

At the end, we'll have some time to build the Prediction-as-Intervention index
ADS "aporia" (contradictions, gaps, issues…)

tinyurl.com/ads-pred-int-index

Contributor(s)	ADS / System	Domain	Covariates (X)	Decision (D)	Proxy Outcome (~Y)
seed	Pretrial risk assessment / PSA	Criminal justice	Failure to appear / new criminal activity risk	Release, detention, bail, or conditions (e.g. supervised release). Supportive services?	Failure to appear for trial
seed	Epic Sepsis Model	Healthcare	Patient history, EHR data	Alert clinicians, early attention, bed allocation	Recorded sepsis / clinical labels
seed	Wisconsin Dropout Early Warning System (DEWS)	Education	Grades, …	Educators identify students for intervention or support?	On-time graduation / dropout
seed	MAAPS data-driven advising protocol	Education	Grades, …	Advising appointments, downstream interventions?	Graduation, retention, academic progress?
seed	Credit Score	Financial Services	Credit history	Pretrial detention? Supportive services?	2-year default

Scan me~

1: Model Design

Problem formulations change from predictions to decisions.

What, when, how?

Model Design:

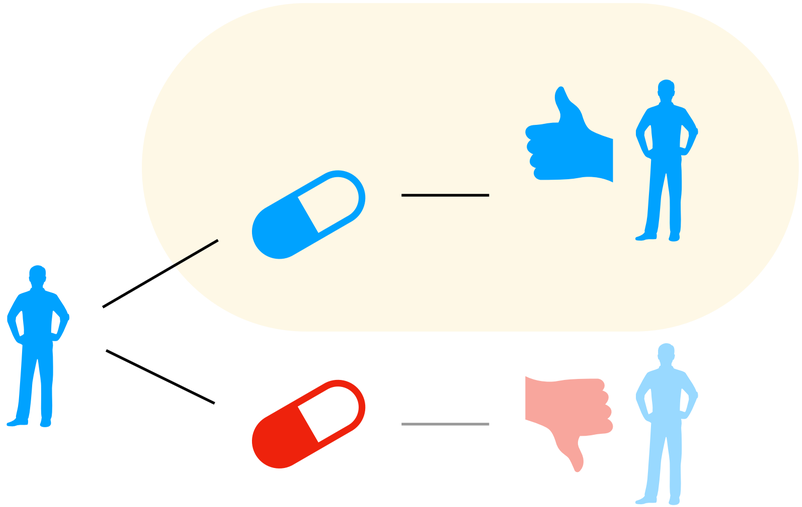
Predictive risk models prioritize baseline risk, not optimal downstream decisions (when decisions have unknown effects).

What is predictive vs. interventional targeting?

Predictive Targeting

- Predictive risk model
  $R(X) = P(Y = 1 \mid X)$
- Treat if $E[Y \mid X] > t$
  (for some threshold t)

Interventional Framework



- $Y(D), D \in \{0,1\}$ Potential outcomes
- Fundamental problem of causal inference
- Average treatment effect:
  $ATE = E[Y(1) - Y(0)]$
- Conditional average treatment effect
  $CATE(X) = E[Y(1) - Y(0) \mid X]$
- Under assumptions (omitted),
  $ATE = E[CATE(X)]$

Predictive Targeting

- Predictive risk model
  $R(X) = P(Y = 1 \mid X)$
- Treat if $E[Y \mid X] > t$
  (for some threshold t)

  Treat if this person is high risk.

Interventional Targeting

- Treat if
  $E[Y(1) - Y(0) \mid X] > t$
  (for some threshold t)

Treat if treatment improves outcomes for this person.

Also called: optimal treatment regime, etc ..

Counterfactual risk assessments

- Treat if $E[Y(0) \mid X] > t.$

Predictive Targeting

- Predictive risk model $R(X) = P(Y = 1 \mid X)$
- Treat if $E[Y \mid X] > t$ (for some threshold t)

  Treat if this person is high risk.

Interventional Targeting

- Treat if $E[Y(1) - Y(0) \mid X] > t$ (for some threshold t)

Treat if treatment improves outcomes for this person.

Counterfactual risk assessments

- Treat if $E[Y(0) \mid X] > t.$

Differ when historical data includes prior interventions/decisions

Predictive Targeting

- Predictive risk model $R(X) = P(Y = 1 \mid X)$
- Treat if $E[Y \mid X] > t$ (for some threshold t)

Treat if this person is high risk.

Interventional Targeting

- Treat if $E[Y(1) - Y(0) \mid X] > t$ (for some threshold t)

Treat if treatment improves outcomes for this person.

Case study: Supervised release (Example 4)

$$P(Y(0) = 1 \mid X) \uparrow$$

Pretrial risk assessment in criminal justice

$\hat{Y}$: Recommendation for *supervised release*

$D(\hat{Y})$: Electronic monitoring (EM) - does judge assign EM if recommended by PSA-DMF?

$Y(D)$: Failure to appear (FTA)
- does defendant fail to appear (bad!!) for court if released under EM?

Exhibit 2: Decision Making Framework

Decision Making Framework Matrix

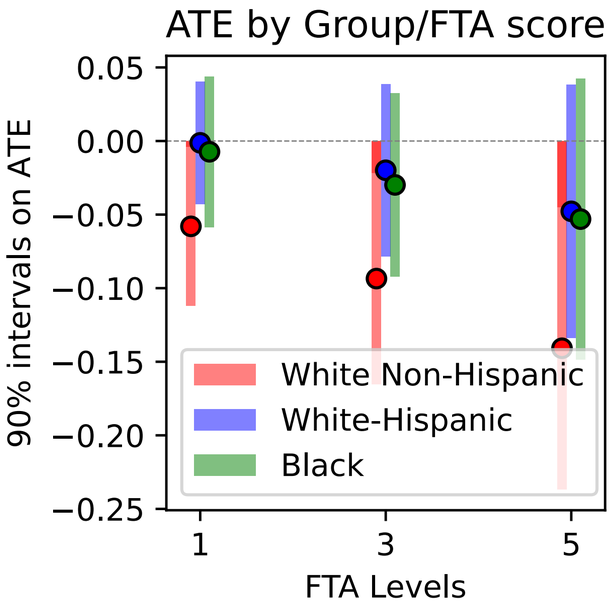
	NCA 1	NCA 2	NCA 3	NCA 4	NCA 5	NCA 6
FTA 1	Release with No Conditions	Release with No Conditions				
FTA 2	Release with No Conditions	Release with PM	Release with		PSL II	
FTA 3		Release w PM			L III with Curfew	Release Not Recommended
FTA 4		PSL I			lease with eriff's EM	Release Not Recommended
FTA 5		PSL I	PSL III	Release with Sheriff's EM	Release Not Recommended	Release Not Recommended
FTA 6				Release Not Recommended	Release Not Recommended	Release Not Recommended

Levels of Supervised release

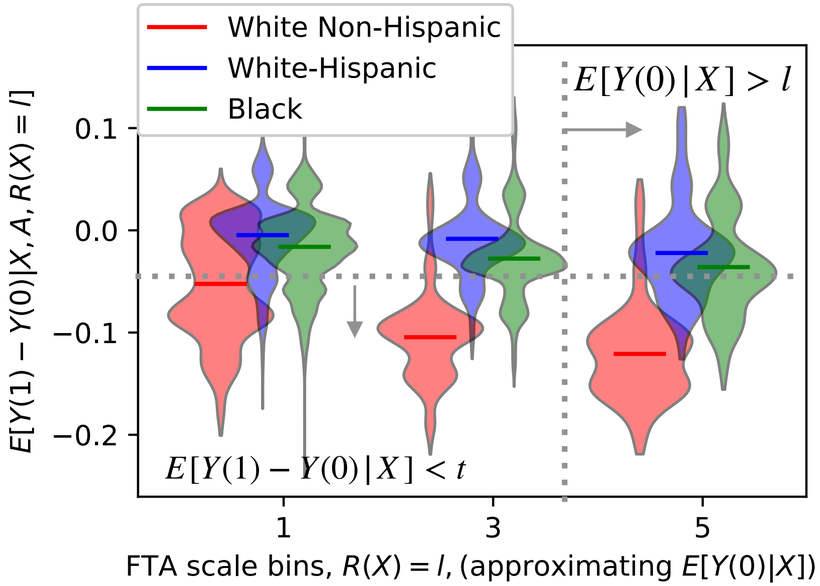
Source: Cook County Sheriff's Office, Sheriff's Justice Institute, *Central Bond Court Report*, April 2016, p. 7, https://www.chicagoreader.com/pdf/20161026/Sheriff_s-Justice-Institute-Central-Bond-Court-Study-070616.pdf

PSA-DMF
*(Coarsened) data from Chicago

Causal impact of ***electronic monitoring*** (EM) on reducing FTA
Heterogeneity: EM could mean *burdensome surveillance* (paying for EM, losing job)



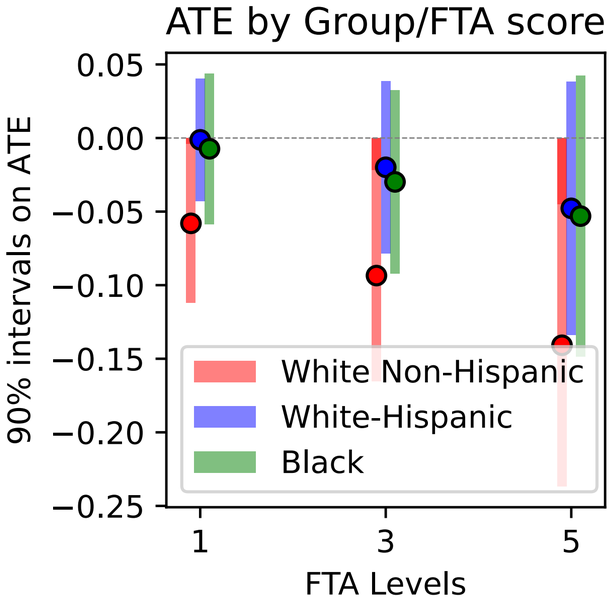
Average treatment effect (90% confidence intervals) for different groups, stratified by FTA risk.



Distribution of heterogeneous treatment effects by group, stratified by FTA risk.

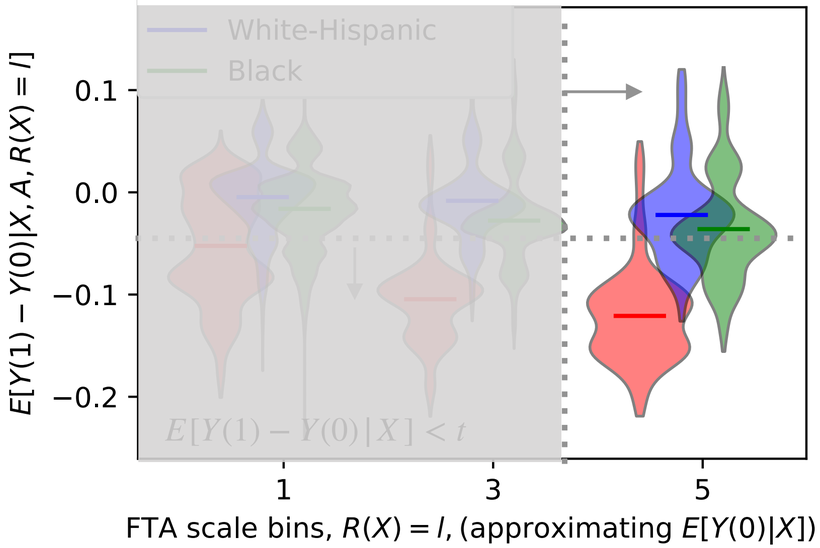
*Treatment appears more beneficial (under assumptions) for white vs. black/Hispanic

* Black/Hispanic receives supervision at higher rates

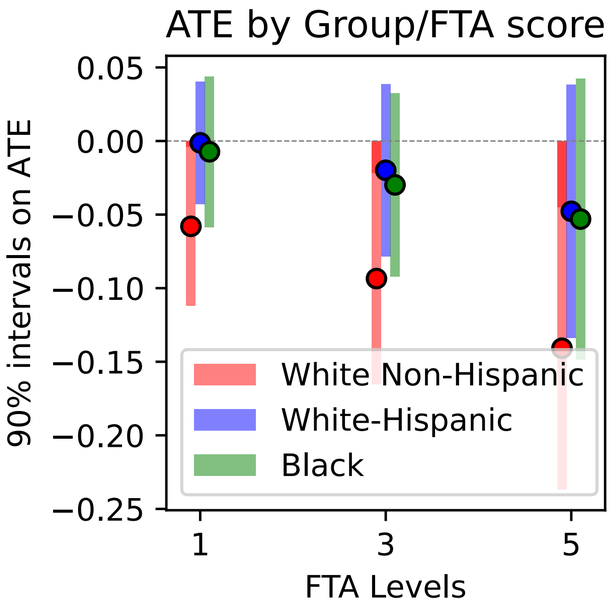


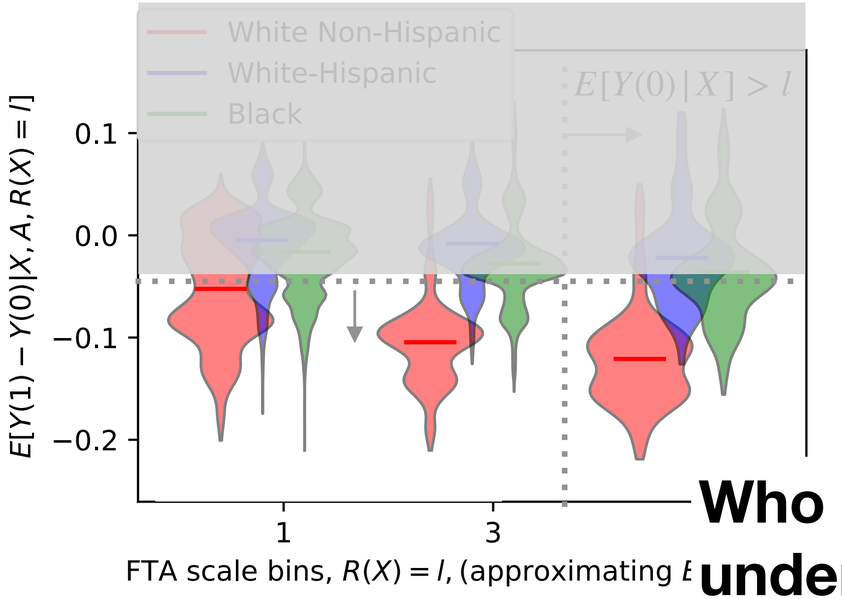
Average treatment effect (90% confidence intervals) for different groups, stratified by FTA risk.

Who is recommended for SR under predictive targeting for FTA risk $(\hat{Y} = 1)$



Distribution of heterogeneous treatment effects by group, stratified by FTA risk.



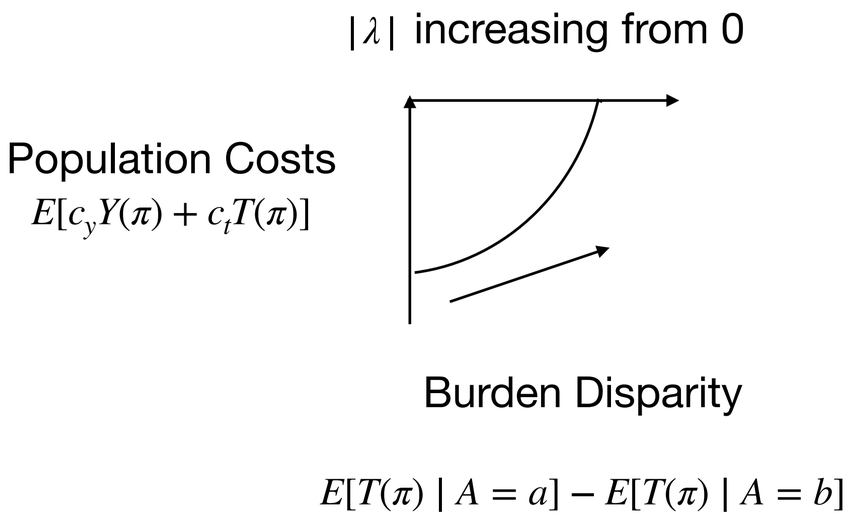


Who is rec. for SR under interventional targeting

*Treatment appears more beneficial (under assumptions) for white vs. black/Hispanic

*** Black/Hispanic receives supervision at higher rates (Surveillance disparities that are suboptimal for population outcomes)**

Optimizing for causal improvement can find Pareto-improving fairness improvements upon status quo



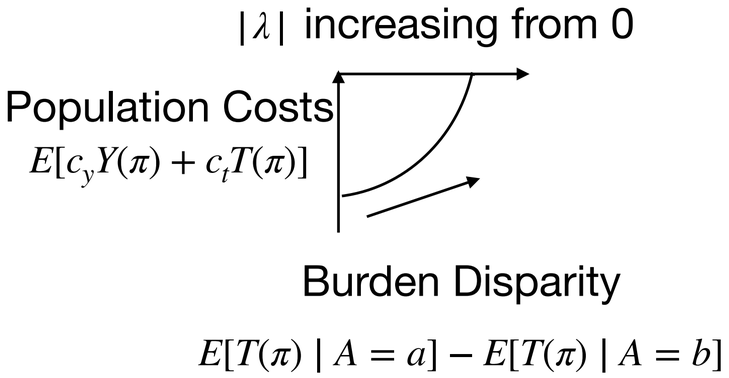
Minimize population costs

$$\min \quad E[c_y Y(\pi) + c_t T(\pi)]$$

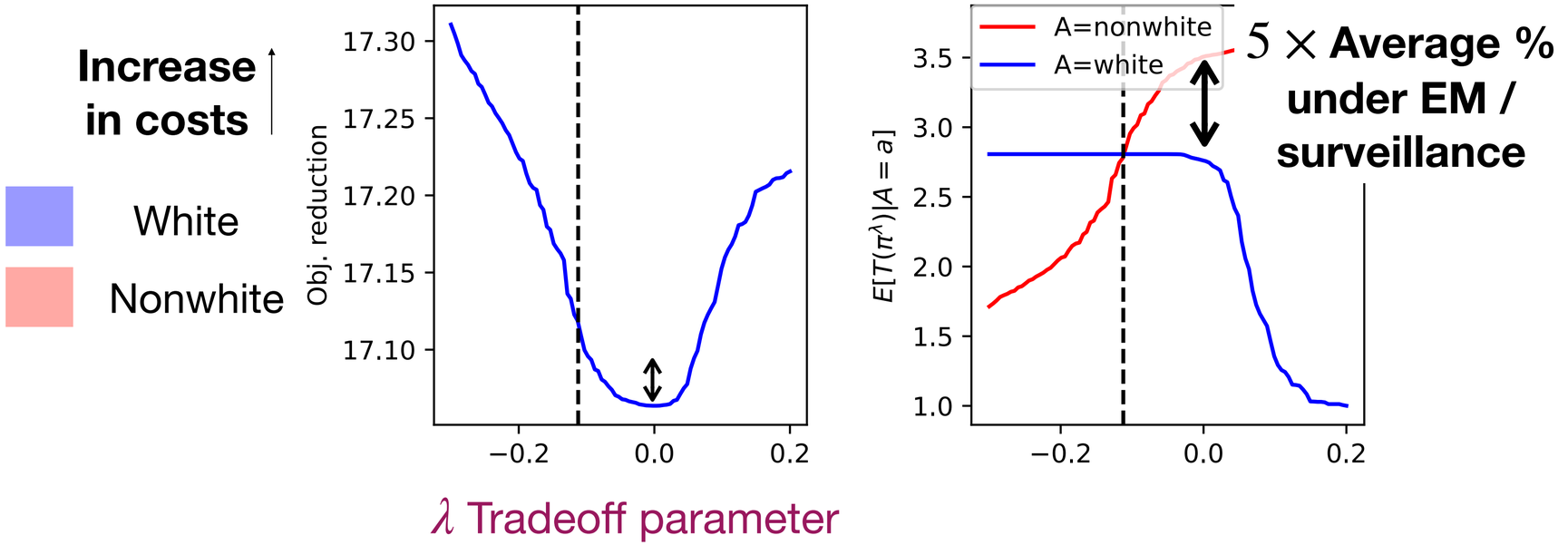
$$\textbf{s.t.} \quad E[D(\pi) \mid A = a] - E[D(\pi) \mid A = b] = 0$$

Eliminate surveillance disparities

Optimizing for causal improvement can find Pareto-improving fairness improvements upon status quo

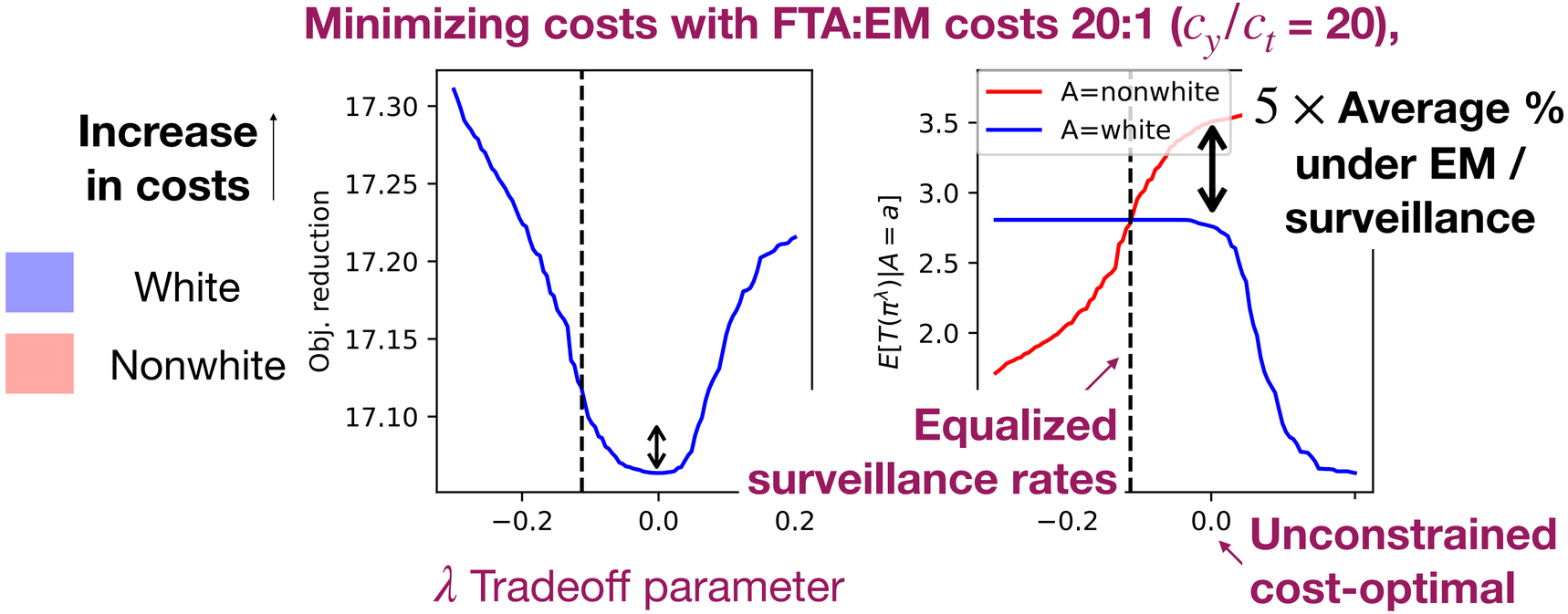


Minimizing costs with FTA:EM costs 20:1 (c_y/c_t = 20),



20% relative reduction in EM (surveillance) disparities for 0.05% relative increase in FTA

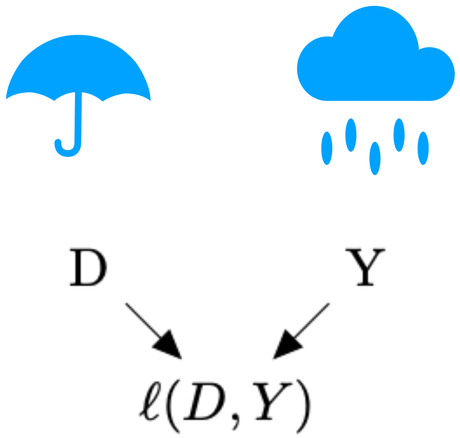
Optimizing for causal improvement can find Pareto-improving fairness improvements upon status quo



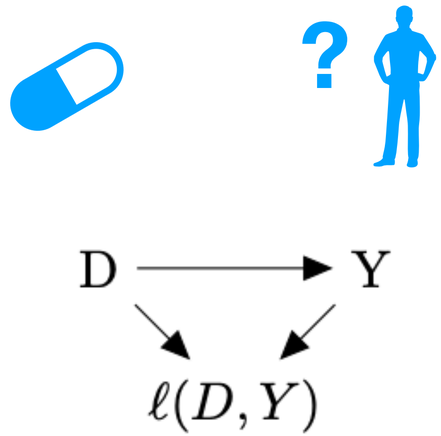
20% relative reduction in EM (surveillance) disparities for 0.05% relative increase in FTA

"When do we use which problem formulation?"

Depends on how organizations benefit/suffer from how predictions inform decisions and outcomes



(a) Can re-evaluate loss function on historical data under counterfactual decisions.



(b) Causal estimation required. Corresponds to learning eq. (optimal treatment policy) if $\ell(D, Y) = Y(D)$, i.e., loss/welfare is the potential outcome.

Your turn

Where have you seen predictive vs. interventional opportunities or gaps?

tinyurl.com/ads-pred-int-index

	X	D	Y	*ADS aporia*
Housing	Housing history, demographics, Health / social factors	Which level of supportive housing, if any	Re-entry within two years	- VI-SPDAT adopted because HUD mandated prioritization guidelines & it was available - Efficiency vs. equity

Scan me~

2: Evaluation Science

“How do we evaluate a predictive ADS in a broader decision-making system?”

- Evaluating **predictions** (e.g. Benchmarking)
- Evaluating **decisions** (e.g. Lab Studies)
- Evaluating **actions** & **interventions** (e.g. Field Experiments)

Evaluation Science:
What does it mean for machine learning to “work”?

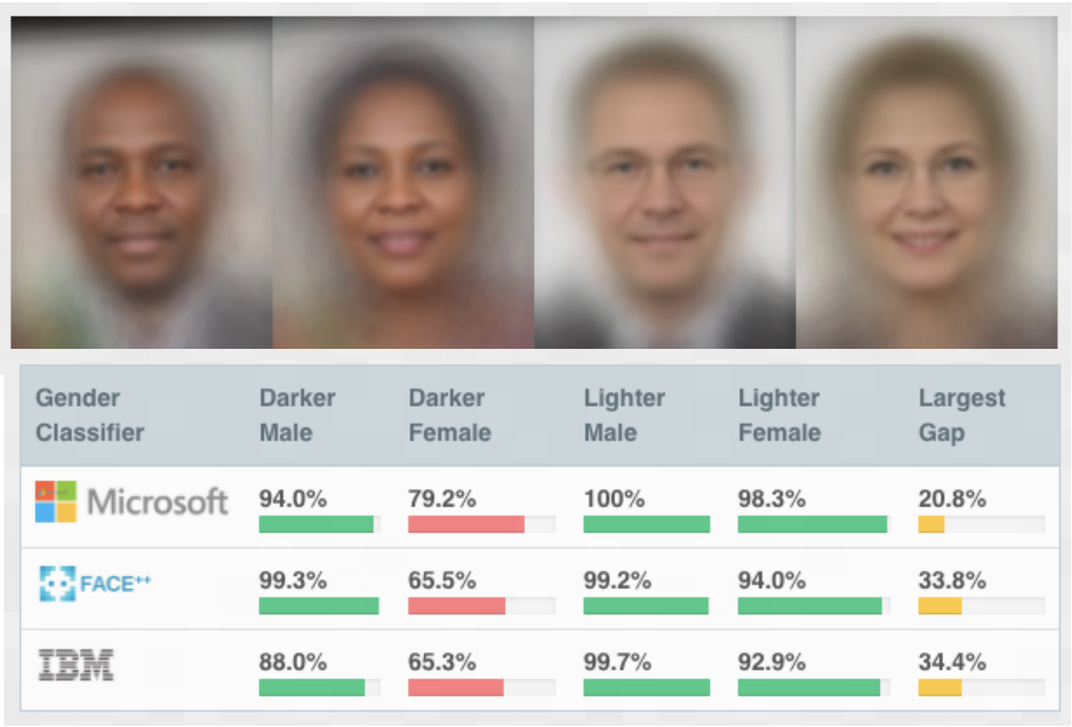
Tempting answer: any model that accurately “solves” its prediction problem!

Facial Recognition → Mis-identification

Public Benefits Risk Assessment (SIS) → Mis-classification

Large Language Models → “hallucination”, mis-representation

Evaluating Predictions: Benchmarking



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

(Boulamwini & Gebru, 2018; Raji & Boulamwini, 2019)

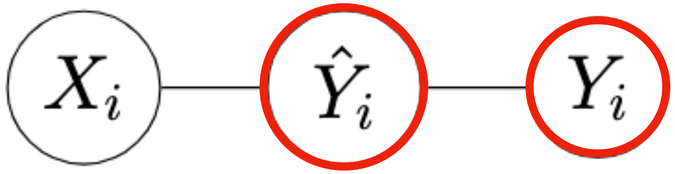


Figure 1: Benchmark Evaluation Paradigm

(Raji & Liu, 2025)

* What makes a good benchmark?

* How can we identify and benchmark for minority populations of x_i ?

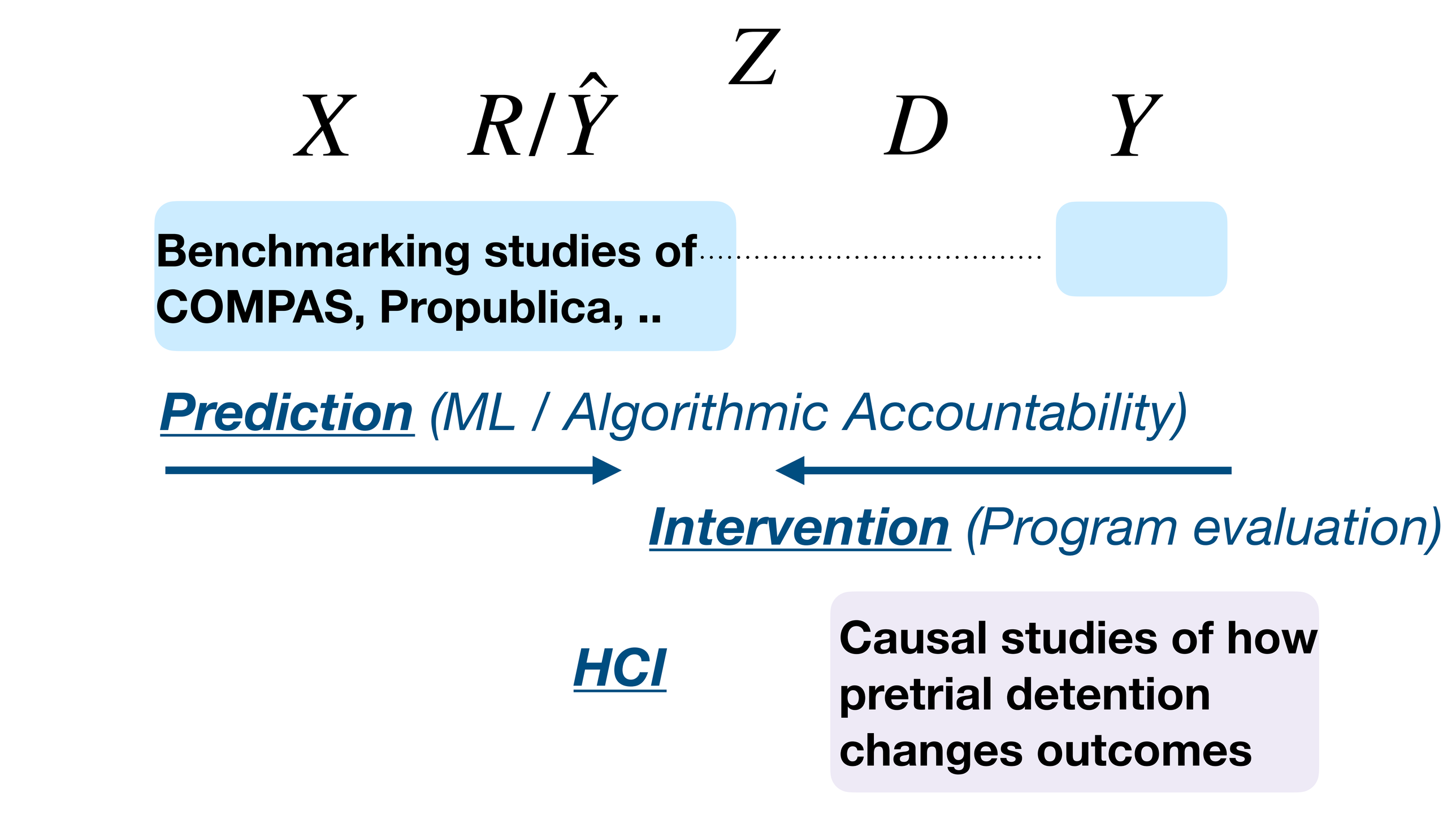
Evaluation Science:

Existing studies imply our integrated framework

$$X \quad R/\hat{Y} \quad Z \quad D \quad Y$$

Benchmarking studies of COMPAS, Propublica, ..

Causal studies of how pretrial detention changes outcomes

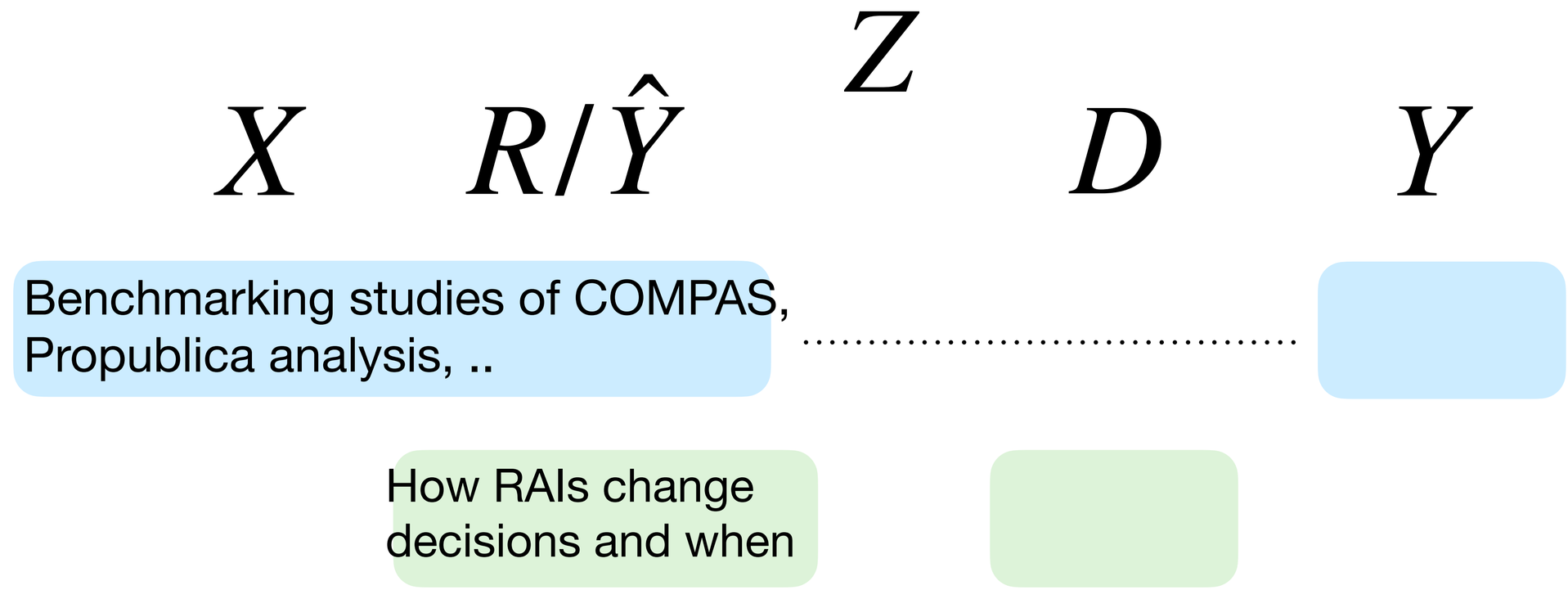

X
$R / \hat{Y}$
Z
D
Y
Benchmarking studies of COMPAS, Propublica, ..
Prediction (ML / Algorithmic Accountability)
Intervention (Program evaluation)
HCI
Causal studies of how pretrial detention changes outcomes

$$X \qquad R/\hat{Y} \qquad Z \qquad D \qquad Y$$

Benchmarking studies of COMPAS, Propublica analysis, ..

Selection bias:
$D = 1$

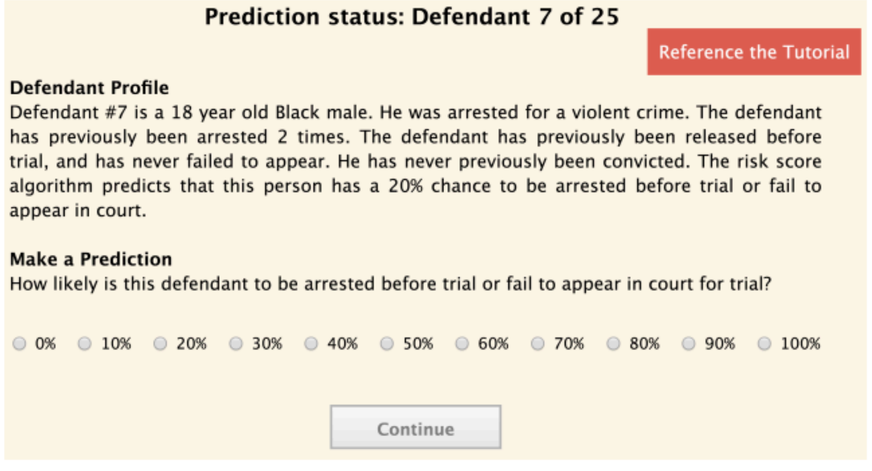
Label bias: Arrests or crime?


Z
X
$R/\hat{Y}$
D
Y
Benchmarking studies of COMPAS, Propublica analysis, ..
How RAIs change decisions and when

$$X \quad R/\hat{Y} \quad Z \quad D \quad Y$$

Before-after comparisons of how *deploying RAIs* changes decisions

Albright 2019,
Green and Chen 2019



$$X \quad R/\hat{Y} \quad Z \quad D \quad Y$$

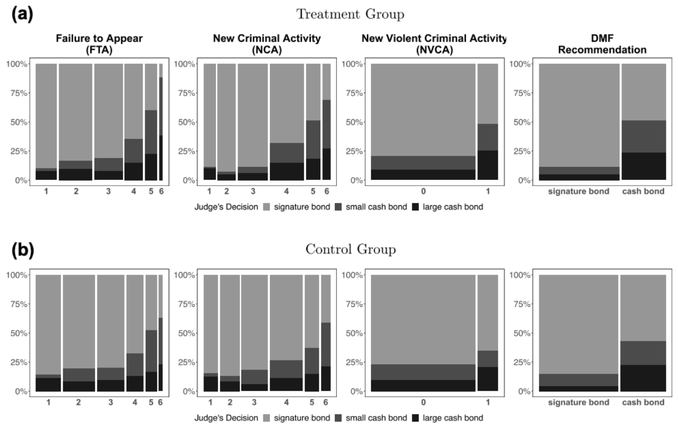


Figure 1. The distribution of the Judge's Decisions given the Pretrial Public Safety Assessment (PSA) among the Cases in the Treatment (Top Panel) and Control (Bottom Panel) Groups. There are three PSA scores, two of which are ordinal—FTA and NCA—while the other is dichotomous—NVCA. The judge's decision is coded as a three-category ordinal variable based on the type and amount of bail: a signature bond, a small cash bond (less than $1,000), and a large cash bond (greater than or equal to $1,000). The DMF recommendation is presented as a binary variable: signature or cash bond. The width of each bar is proportional to the number of cases for each value of the corresponding PSA score. There exists a positive correlation between PSA scores and the severity of the judge's decisions in both treatment and control groups. (a) Treatment group. (b) Control group.

RCTs of how RAIs change decisions and outcomes

Ben-Michael et al 2025, Imai et al 2023

$$X \quad R/\hat{Y} \quad Z \quad D \quad Y$$

Benchmarking studies of COMPAS, Propublica analysis, ..

How RAIs change decisions and when

Before-after comparisons of how *deploying RAIs* changes decisions

RCTs of how RAIs change decisions and outcomes

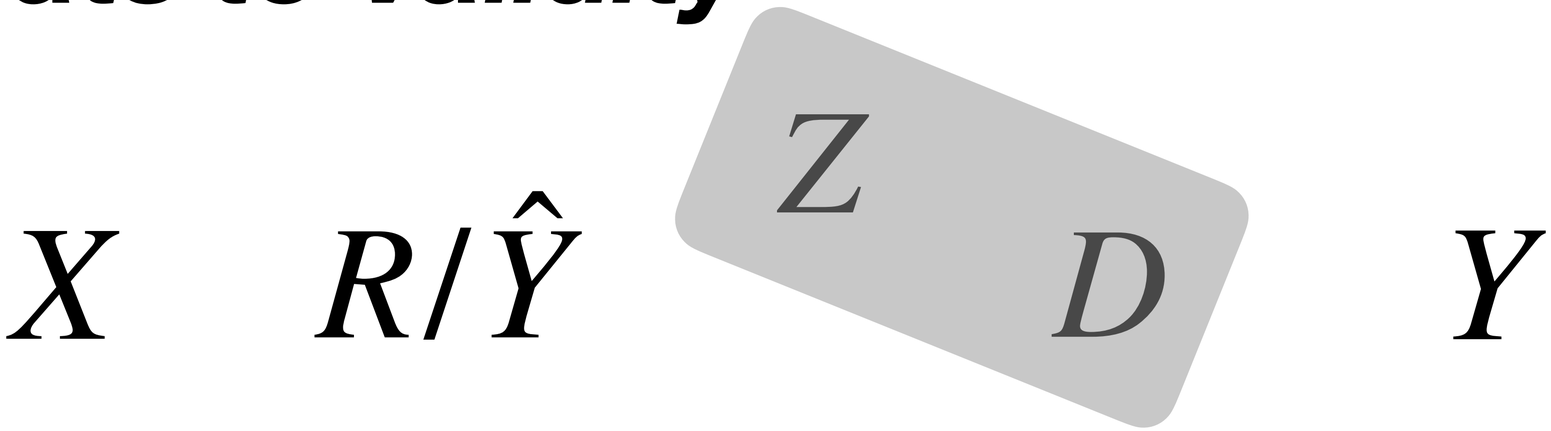
Causal studies of how pretrial detention changes outcomes

Koppel et al, 2024

Evaluation Science:

Looking backwards,
***Historical* decisions threaten evaluation validity**

Threats to Validity



- Organizations are always trying to improve outcomes
  $\Longrightarrow$ distribution of D, $Y(D)$ shifts (zombie predictions)

- Decisions censor outcomes: Selective labels (Lakkaraju et al.)

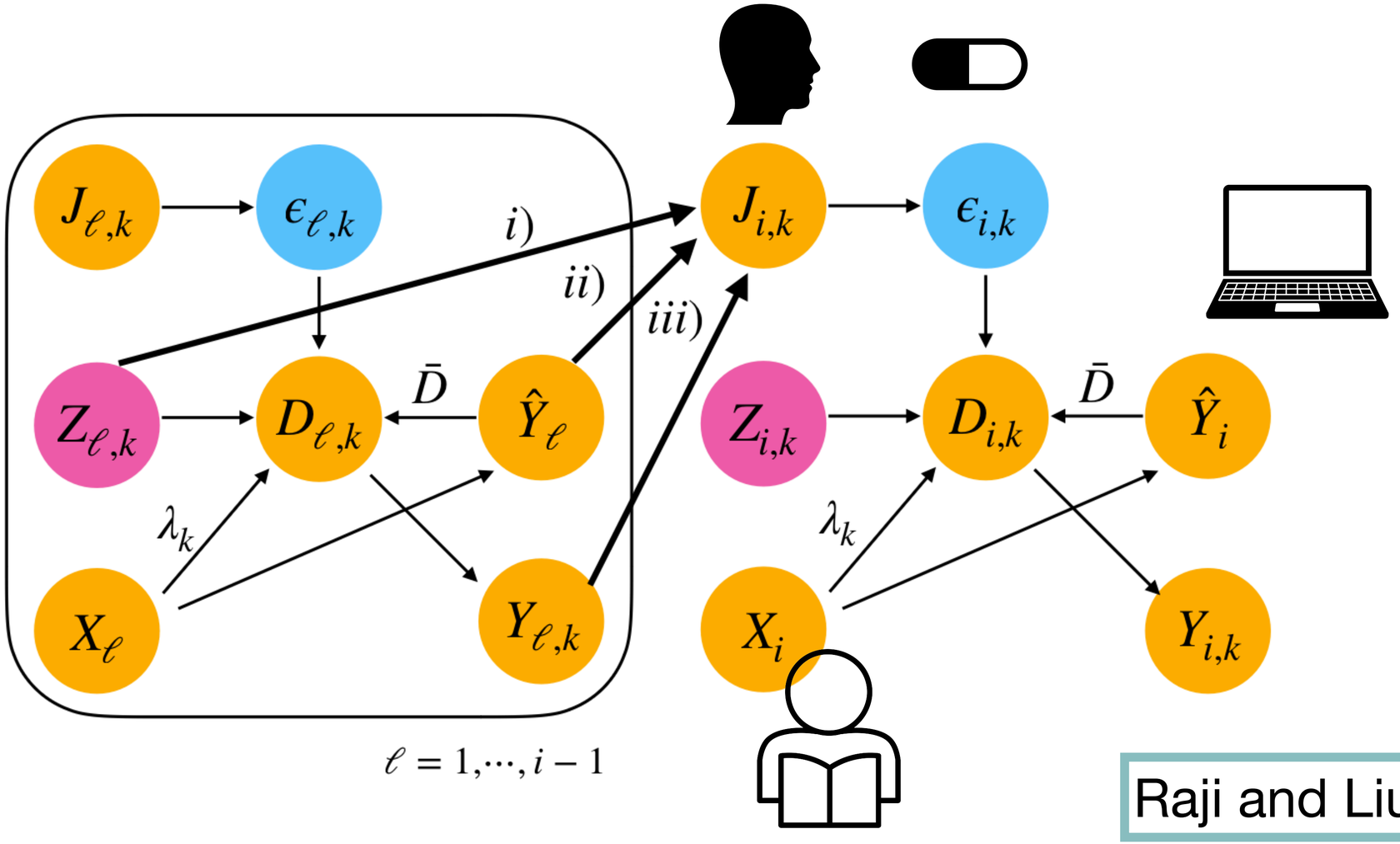
- Organizational decision-making introduces “spurious correlations”
  E.g. in healthcare, Caruana et al.: asthma correlated with *lower risk pneumonia*, but because asthmatic patients patients admitted to ICU and receive better care (D)

Evaluation Science:

Looking forwards,
ADS should not be evaluated on predictive accuracy alone,
but how they impact downstream decisions and outcomes

Experimental evaluations with human decision makers

i),ii), iii) result in interference between units (decision subjects)

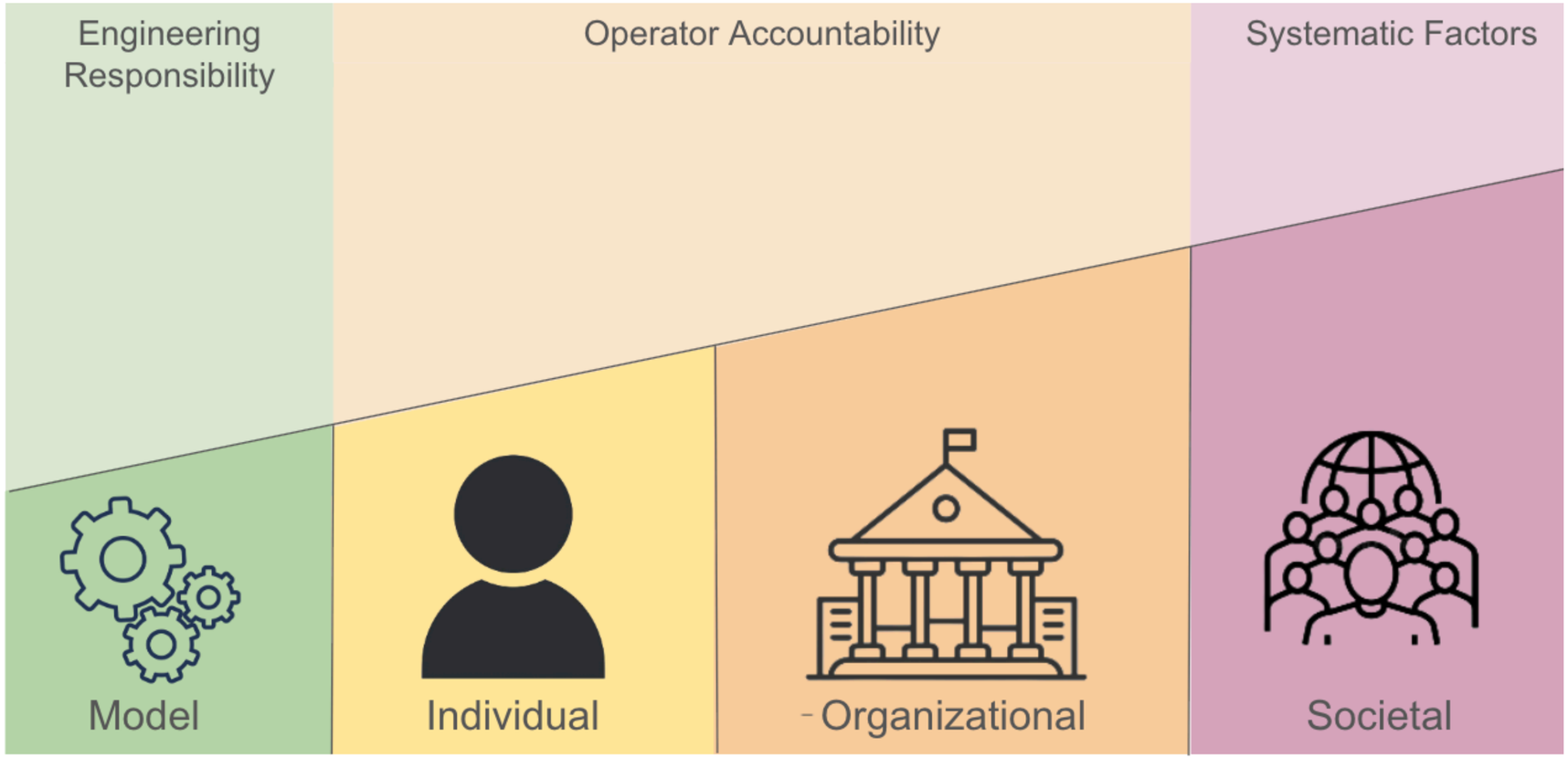


Raji and Liu 2024

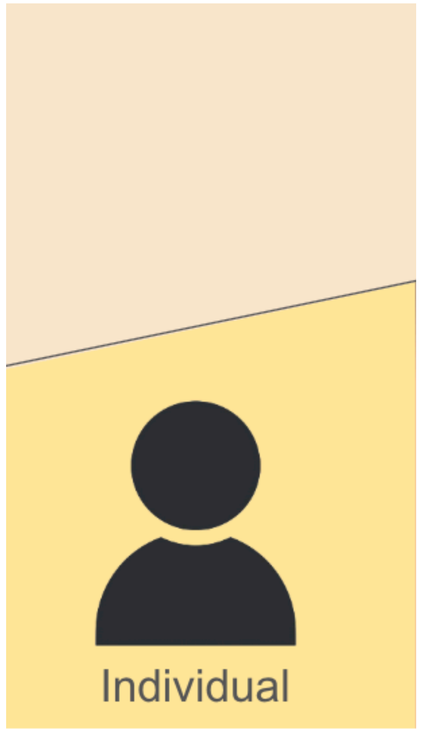
3: Implementation Science

Towards a systematic study of methods and strategies that support the integration and sustained use of evidence-based ADS in real-world settings.

Implementation is part of the intervention



Individual Deployment Context



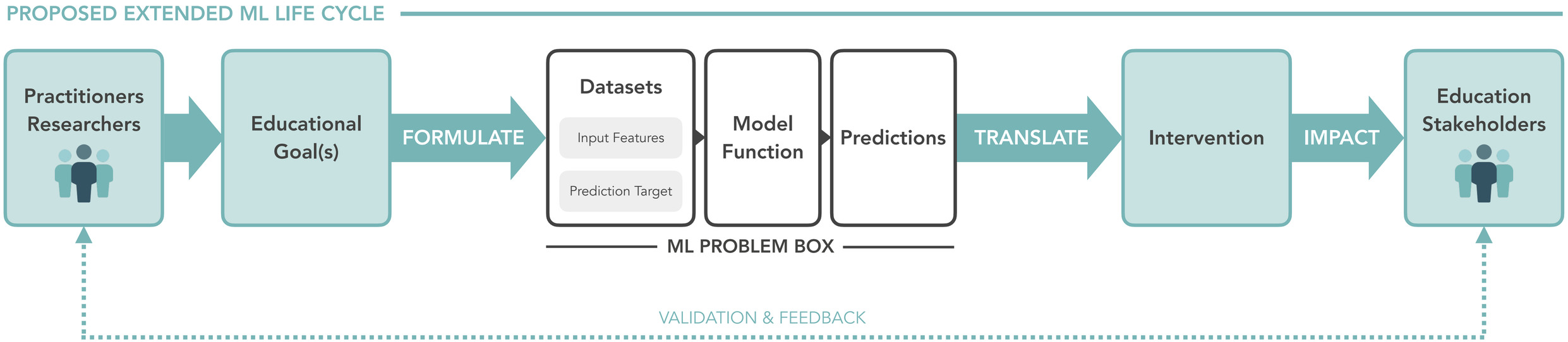
Interaction Design

- What information does the human decision maker have (vs. model)?
- Does human need to retain control of final decision?

Intervention Design

- Are the interventions effective?
- Are the interventions personalized by the human or standardized?

Interventions in the AI/ML life cycle in Education



“Reimagining the Machine Learning Life Cycle to Improve Educational Outcomes of Students”
L.T. Liu, Serena Wang, Tolani Britton, Rediet Abebe. PNAS, 2023.

AI can Predict, but Advisors Guide and Intervene

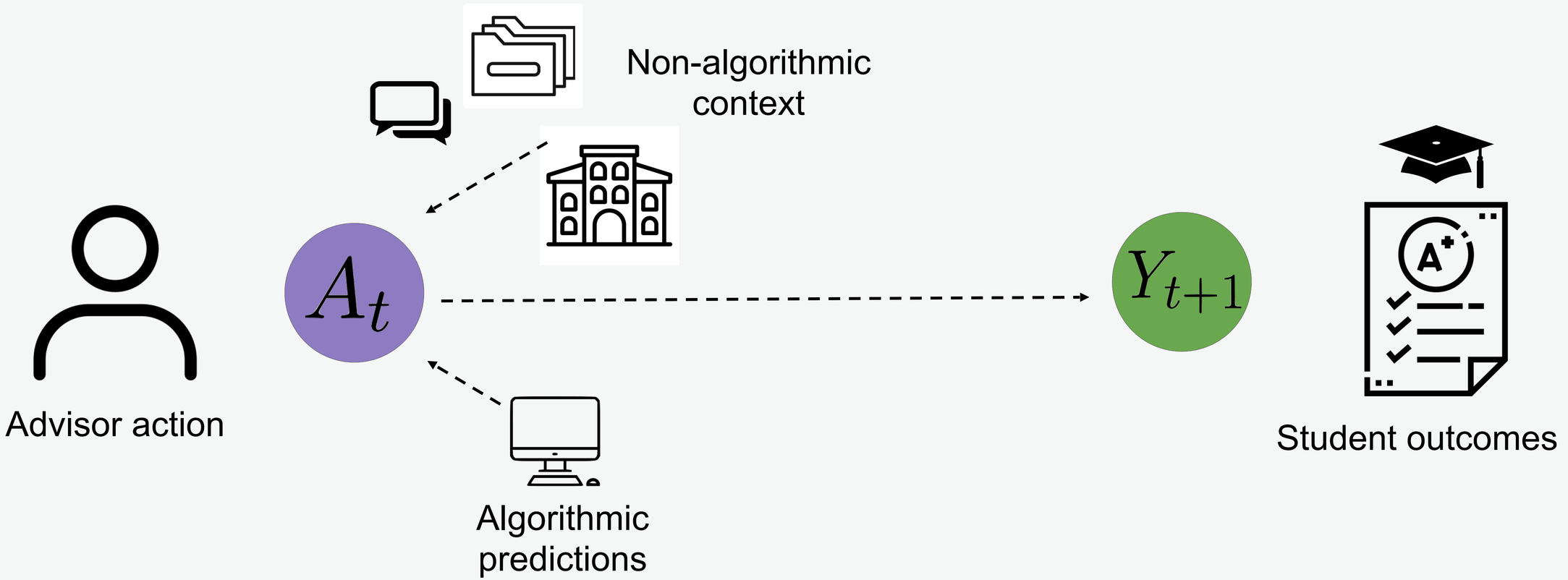
- Risk prediction systems could add value if they were designed with the teacher's (or advisor's) partnership in mind.
- Human expertise: Advisors “fill in the gaps for students” to help them “envision a pathway” from a relatively vague conversation.

P9: “That sounds great. I had no idea what an occupational therapist even was.”

P9: “We have respiratory therapy, or we have nutrition, ...all these other healthcare disciplines that might allow you to help people, to work in healthcare, and to get a job, which [are] [...] three boxes you said you wanted to check.”

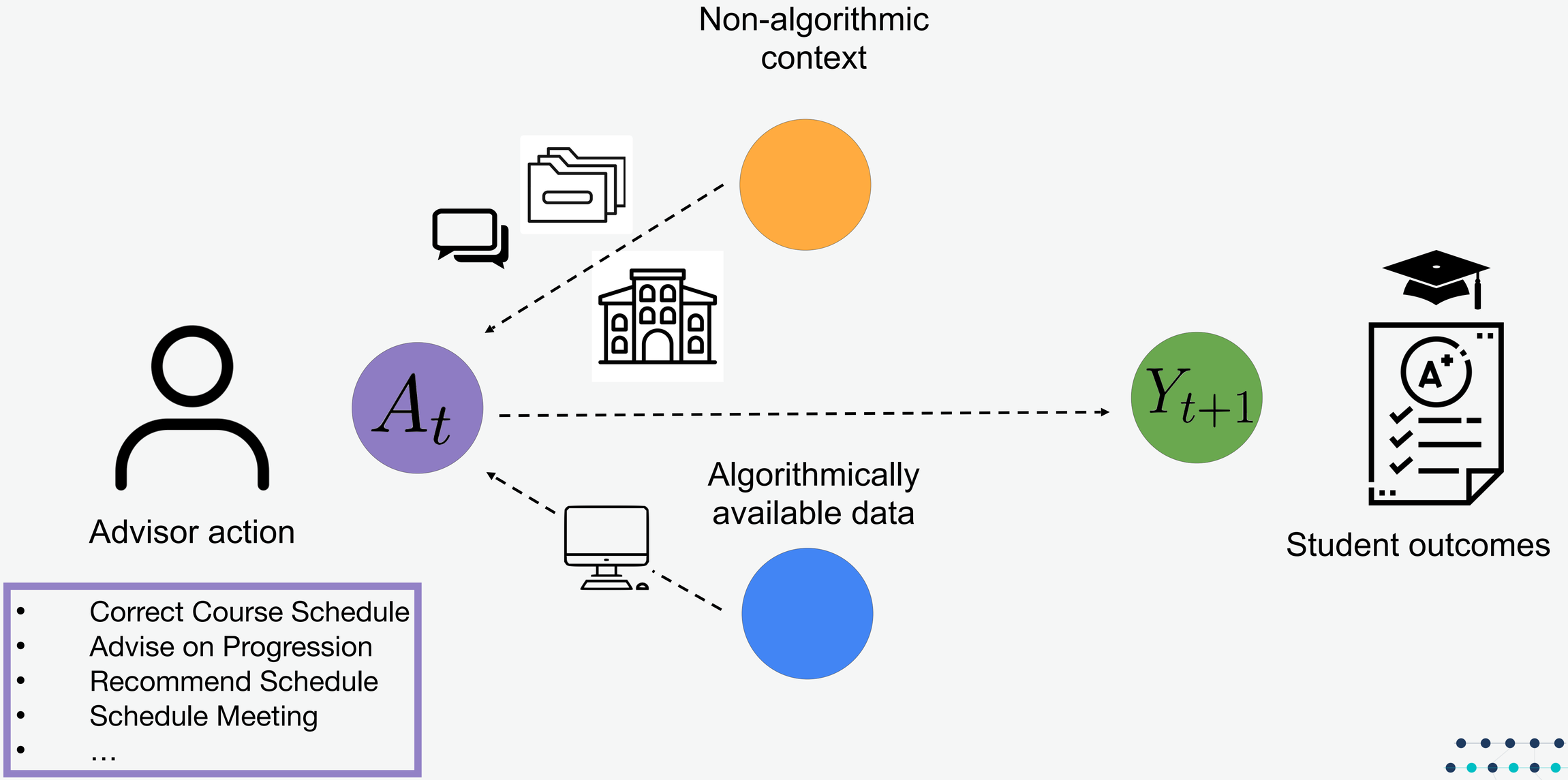
P9: “You're not going to be a nurse, sorry, like, good luck”

“Reimagining the Machine Learning Life Cycle to Improve Educational Outcomes of Students” L.T. Liu, Serena Wang, Tolani Britton, Rediet Abebe. PNAS, 2023.



Schechtman, K., Brandon, B., Stafford, J., Li, H., & **Liu,** L. T. 2025. Discretion in the Loop: Human Expertise in Algorithm-Assisted College Advising. In Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization.

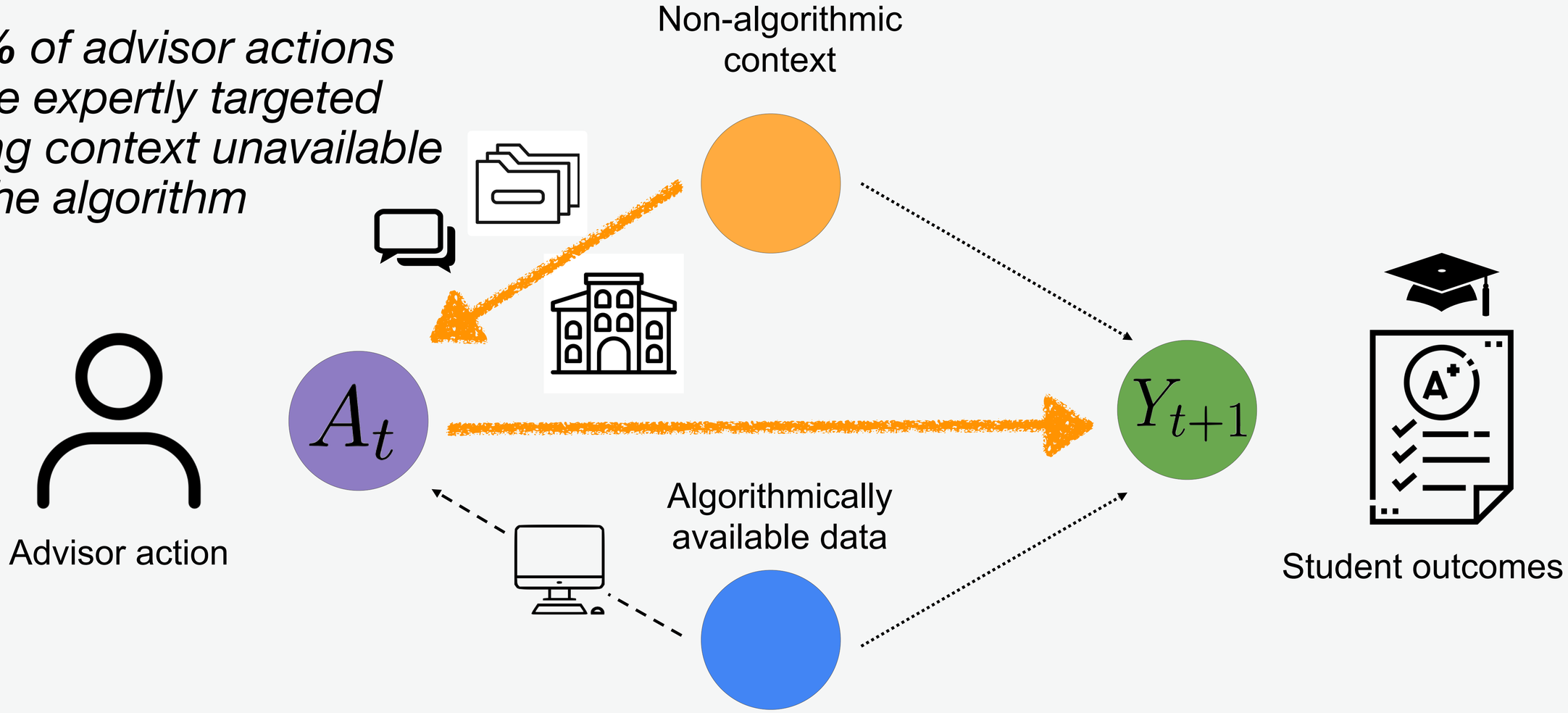
PRINCETON HAIKU Lab



PRINCETON HAIKU Lab

Schechtman, K., Brandon, B., Stafford, J., Li, H., & **Liu,** L. T. 2025. Discretion in the Loop: Human Expertise in Algorithm-Assisted College Advising. In Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization.

***66%** of advisor actions were expertly targeted using context unavailable to the algorithm*

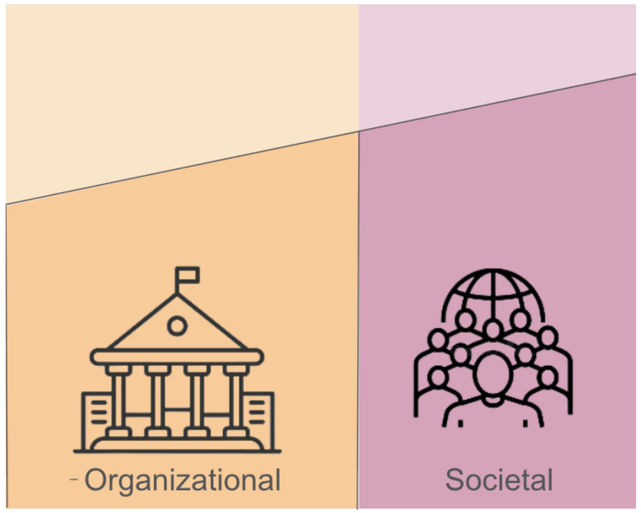


Findings: Human Expertise Drives the Impact of AI-Supported Advising

Schechtman, K., Brandon, B., Stafford, J., Li, H., & **Liu,** L. T. 2025. Discretion in the Loop: Human Expertise in Algorithm-Assisted College Advising. In Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization.

PRINCETON HAIKU Lab

Discuss: Institutional Factors

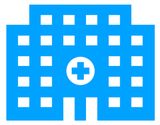


Integration & change management, Personnel guidance & training, Bureaucratic counterfactual, Implementation factors

Application domain norms, Regulatory Compliance & Legal Compatibility

“What is the difference between predicting risk in (A) criminal recidivism and (B) hospital re-admission?”

These two tasks have almost identical technical formulations—input characteristics predict the probability of a negative outcome and inform interventions to minimize it.



But the **decision stakes, stakeholder incentives, capacity constraints, regulatory environment, availability of recourse**…differently determine the outcomes of deployment.

Further Resources on Implementation Science

- **Implementation science foundations** (Bauer et al. 2015; Wensing 2015; Moir 2018) **Statistical prediction vs. clinical experts** (Meehl 1954; Grove et al. 2000)
- **Public-sector ADS and street-level discretion** (Saxena et al. 2021; Saxena and Guha 2024; Johnson and Zhang 2022) **ADS interventions and expertise in education** (Liu et al. 2023; McConvey et al. 2023; Schechtman et al. 2025)
- **Human-AI Interaction through a decision-theoretic lens** (Hullman et al. 2025a; Hullman et al. 2025b; Guo et al 2024, 2025, Bhatt et al. 2021) **Compliance, over-reliance, and trust** (Green and Chen 2019, 2021; Bansal et al. 2021; Buccinca et al. 2020, 2021; Mozannar et al. 2022, 2023) **Capacity constraints** (Boutilier et al. 2024; Raji and Liu 2025)
- **Deployment complexity and change management** (Boag et al. 2024; Sendak et al. 2020; Klein 2017) **Governance and accountability** (Raji et al. 2020, 2022; Xiang and Raji 2019; Ojewale et al. 2024)

4: Prediction-as-Intervention Index

Connecting across our disciplinary and “field” experiences

Community Poll: Who should receive the intervention?

A jurisdiction has resources (e.g. temporary housing) for only 20% of cases. Which targeting rule should it use?

Highest predicted risk
Prioritize people most likely to have the bad outcome.

Highest expected benefit
Prioritize people most likely to improve because of action.

Lottery among eligible cases
Avoid false precision and distribute scarce resources procedurally.

Categorical eligibility
Use a transparent rule set before any model is considered.

Build the Prediction-as-Intervention Index

Use the Googlesheets live archive. Add a new ADS, or start with one seeded row. Discuss proxy-target gap, one eval, one implementation factor. Note sources at the end.

Contributor(s)	ADS / System	Domain	Covariates (X)	Decision (D)	Proxy Outcome (~Y)
seed	Pretrial risk assessment / PSA	Criminal justice	Failure to appear / new criminal activity risk	Release, detention, bail, or conditions (e.g. supervised release). Supportive services?	Failure to appear for trial
seed	Epic Sepsis Model	Healthcare	Patient history, EHR data	Alert clinicians, early attention, bed allocation	Recorded sepsis / clinical labels
seed	Wisconsin Dropout Early Warning System (DEWS)	Education	Grades, ...	Educators identify students for intervention or support?	On-time graduation / dropout
seed	MAAPS data-driven advising protocol	Education	Grades, ...	Advising appointments, downstream interventions?	Graduation, retention, academic progress?
seed	Credit Score	Financial Services	Credit history	Pretrial detention? Supportive services?	2-year default

Scan me~

tinyurl.com/ads-pred-int-index

In-room: work in pairs or your table. Zoom: breakout groups/ add directly

Paths forward for further research and collaboration

1. Interrogate prediction
When is prediction the right abstraction?

2. Better engineering
Design tools around actions, uncertainty, and constraints.

3. Deployment assessment

Make causal impact evaluation routine.

4. Sustainable governance

Build monitoring, recourse, and accountability into institutions.