Counterfactually Comparing Abstaining Classifiers

Yo Joong Choe, Aditya Gangrade, and Aaditya Ramdas

37th Conference on Neural Information Processing Systems (NeurIPS 2023)

Yo Joong "YJ" Choe
University of Chicago
(work done at CMU)

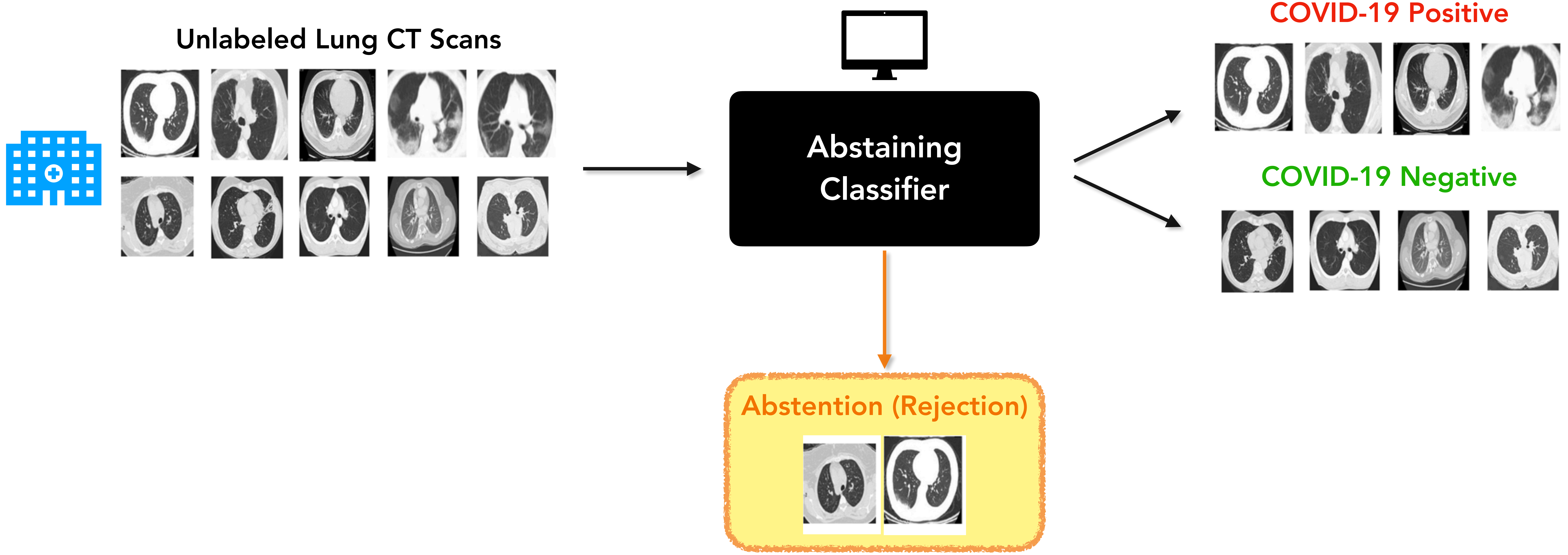
Aditya Gangrade
University of Michigan
(work done at CMU)

Aaditya Ramdas
Carnegie Mellon University

Abstaining Classifiers

Chow (1957)

a.k.a. Selective Classifiers; Classifiers with a Reject Option



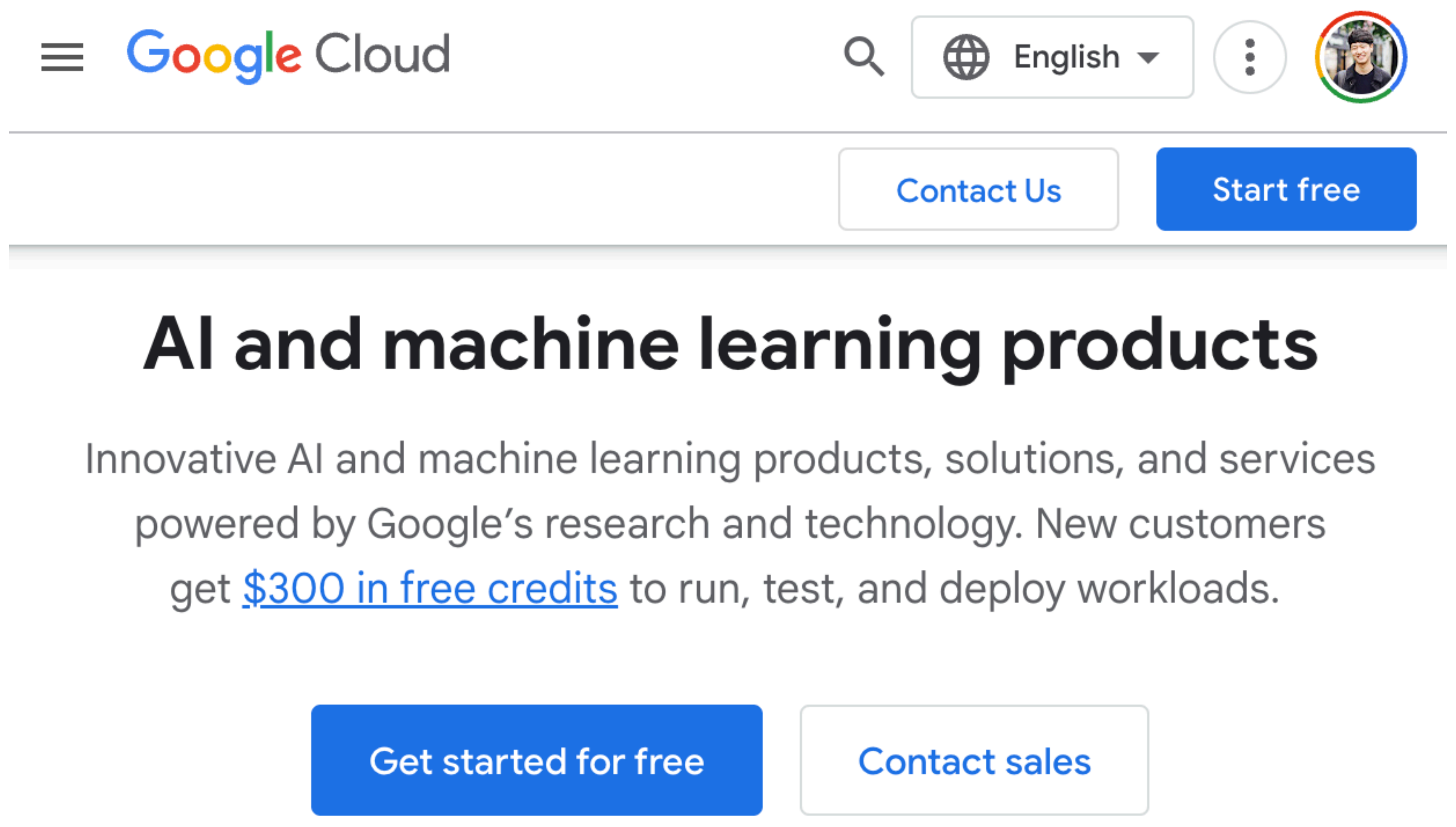
Sample lung CT scans of (non-)COVID patients from Ahuja et al. (2020).

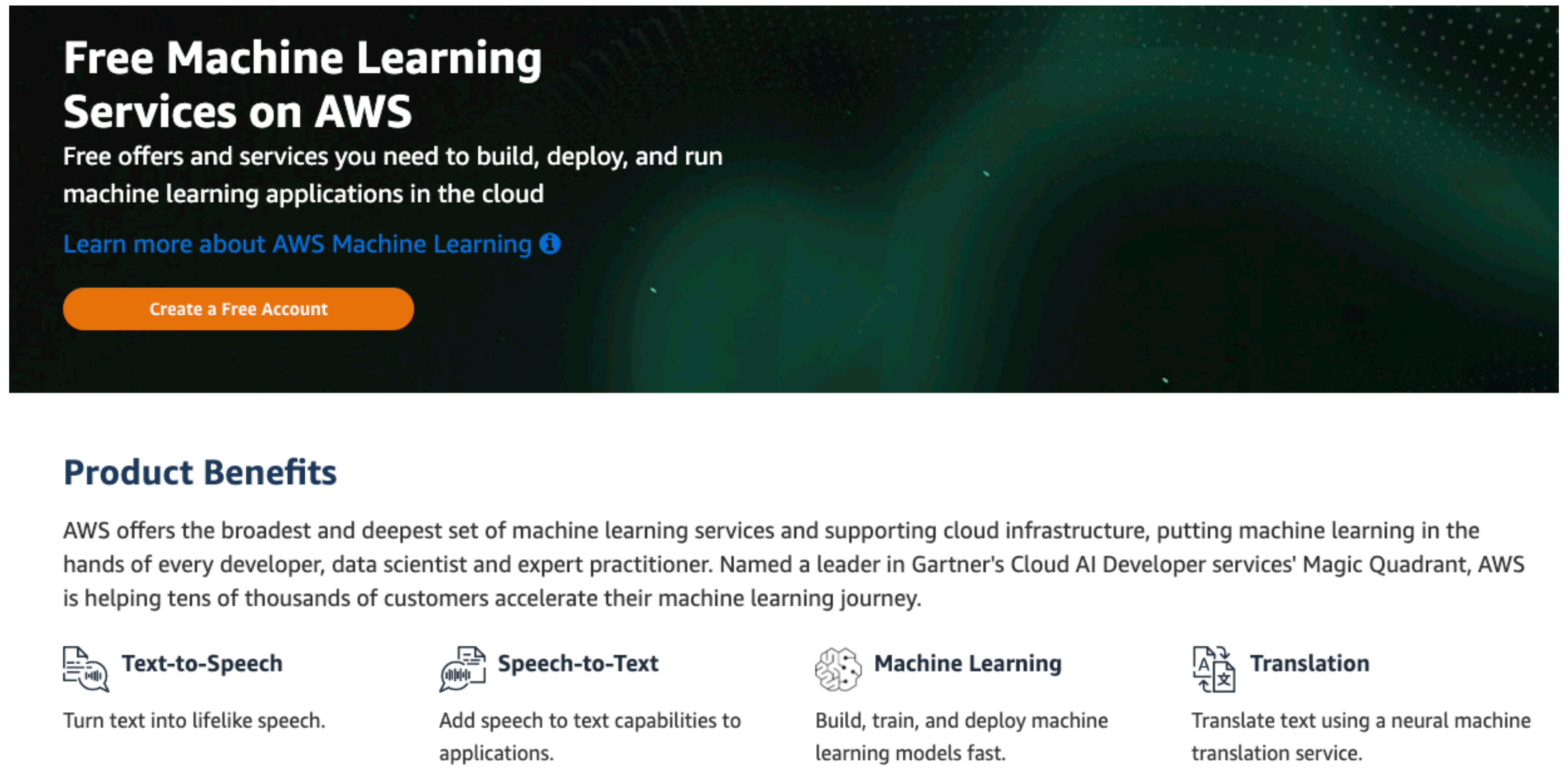
Motivation: Evaluating Free-Trial ML Services

- Suppose that we want to evaluate **black-box** ML prediction services for image classification.

- During the **free trial**, each service deploys an abstaining classifier, such that it only gives predictions on certain inputs and abstain on others.

- **The full (paid) versions do not abstain.** We want to compare the performance of the full versions.





Key Takeaway & Main Question

To the evaluator, abstentions are just **missing** predictions!

How do we compare black-box abstaining classifiers

while accounting for their missing predictions?

Problem Setup

Chow (1957); El-Yaniv & Wiener (2010)

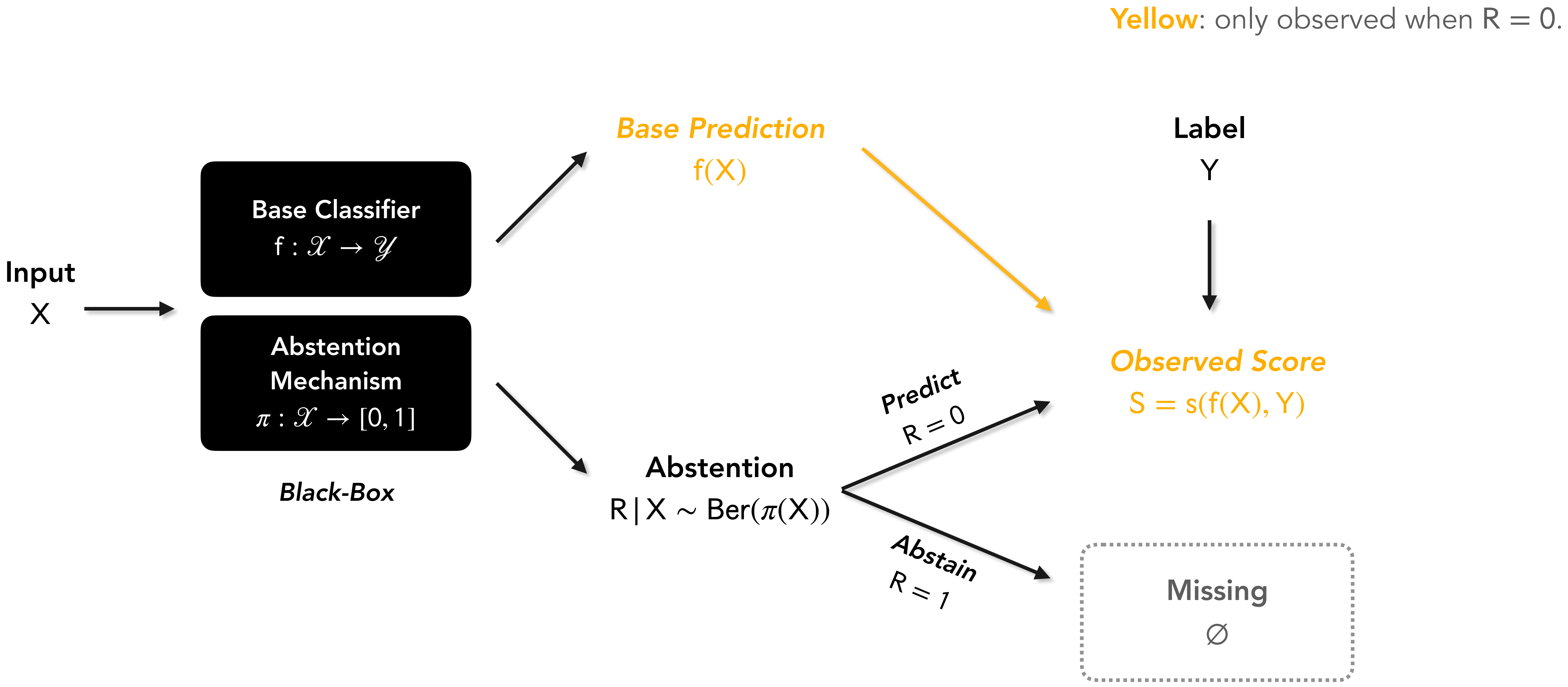
Definition. An **abstaining classifier** is a pair of functions (f, π), where

- $f : \mathscr{X} \to \mathscr{P}(\mathscr{Y})$ is the **base classifier**, which outputs a (probabilistic) prediction; and
- $\pi : \mathscr{X} \to [0, 1]$ is the **abstention mechanism**, which outputs the probability of abstention.

Evaluating a black-box abstaining classifier (f, π).

1. Classifier receives an input X.
2. Classifier decides whether or not it will abstain: $R \mid X \sim \mathrm{Ber}(\pi(X))$.
   - If $R = 0$, then Evaluator observes the prediction & score: $S = s(f(X), Y)$.
   - If $R = 1$ ("rejection"), then Evaluator does NOT see its prediction or score **(S is missing)**.

Illustration: Evaluating a Black-Box Abstaining Classifier



*Base classifier & abstention mechanism may be conjoined (e.g., via shared feature layers).

The 3-Step Approach To Nonparametric Causal Inference

Target Definition → **Identification** → **Estimation**

Target Definition

$$\psi \overset{\text{def}}{=} \mathbb{E}[S]$$

("Counterfactual Score")

<u>*Nuisance Functions (Learnable):</u>

Abstention Mechanism

$$\pi(X) \overset{\text{def}}{=} \mathbb{P}(R = 1 \mid X)$$

Selective Score Predictor

$$\mu_0(X) \overset{\text{def}}{=} \mathbb{E}[S \mid R = 0, X]$$

Identification

$$\psi = \mathbb{E}[\mu_0(X)]$$

<u>Conditions</u>

1. **Missing At Random:**
   $S \perp\!\!\!\perp R \mid X$
   Independent Evaluation Set
2. **Positivity:**
   $\pi(X) \leq 1 - \epsilon$
   (for some $\epsilon > 0$)
   Stochastic Abstentions

Estimation

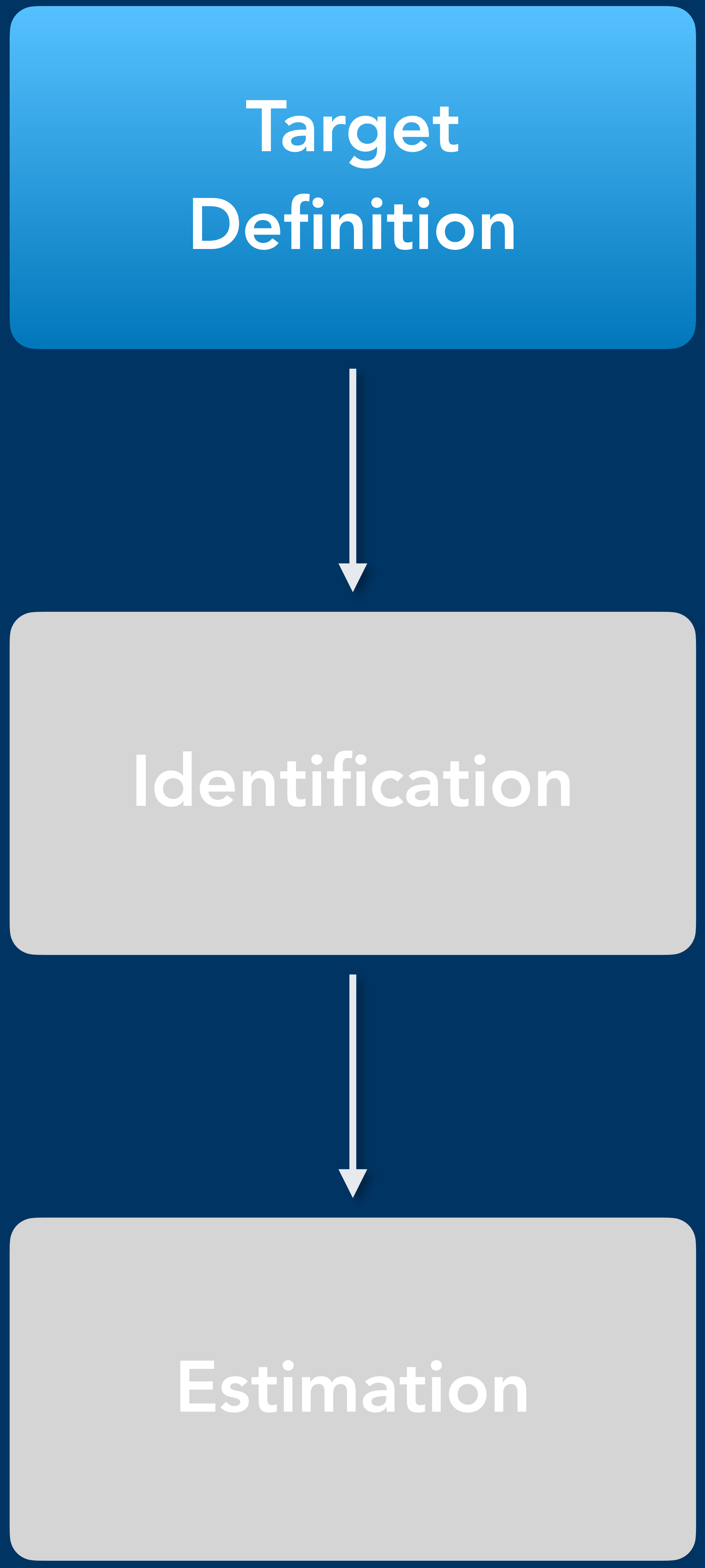
$$\sqrt{n}\left(\hat{\psi}_{dr} - \psi\right) \rightsquigarrow \mathcal{N}\left(0, \mathrm{Var}_{\mathbb{P}}[\mathrm{IF}]\right)$$

<u>Conditions</u>

1. **Double Robustness:**
   $\|\hat{\pi} - \pi\|_{L^2} \|\hat{\mu}_0 - \mu_0\|_{L^2} = o_{\mathbb{P}}(1/\sqrt{n})$
   Flexible Nuisance Learners (NN, RF, …)
2. **IF Consistency:**
   $\|\hat{\mathrm{IF}} - \mathrm{IF}\| = o_{\mathbb{P}}(1)$

cf. Rubin (1974); Robins et al. (1994); many others.

Target Definition



Our Target: The Counterfactual Score

Definition (Counterfactual Score): Given an abstaining classifier, we define the **counterfactual score** ψ as

$$\boxed{\psi = \mathbb{E}[S],}$$

NOTE:
S is missing when $R = 1$.

where $S = s(f(X), Y)$ for some scoring function s (e.g., accuracy & Brier score).
Expectation $\mathbb{E}$ is taken over (X, R, S). *No conditioning on non-abstentions ($R = 0$).*

Why the counterfactual score?

- Measures **how each classifier *would have performed,* had it not been allowed to abstain**.
- There exist efficient estimators that do not require parametric modeling assumptions.

The Popular Metric Does NOT Account for Missing Predictions

It is common to evaluate abstaining classifier using selective score & coverage (a two-dimensional metric):

- **Selective score** = expected score *only* on selections (non-abstentions) = $\mathbb{E}[S \mid R = 0]$.
- **Coverage** = expected rate of non-abstentions = $\mathbb{P}(R = 0)$.

Selective score + coverage do NOT capture the classifier's performance adequately,

particularly when the missing predictions matter.

For Comparison: The Counterfactual Score *Difference*

Definition (Counterfactual Score *Difference*): given two abstaining classifiers, A & B, we define their **counterfactual score *difference*** Δ as

$$\boxed{\Delta := \mathbb{E}[S^A - S^B],}$$

where $S^A := s(f^A(X), Y)$ and $S^B := s(f^B(X), Y)$ for some scoring function s.
Expectation $\mathbb{E}$ is taken over (X, R^A, S^A, R^B, S^B). *No conditioning on non-abstentions.*

Remark: The two classifiers can operate under their ***separate* abstention mechanisms**.

Classifiers Can Use Separate Abstention Mechanisms

Counterfactual Score Difference

$$\Delta = \mathbb{E}[S^A - S^B]$$

ID	X	R^A	S^A	R^B	S^B
1	●	●	●	●	●
2	●	○	?	○	?
3	●	●	●	●	●
4	●	○	?	●	●
⋮			⋮		⋮
⋮			⋮		⋮
N	●	●	●	○	?

may observe both, either, or neither

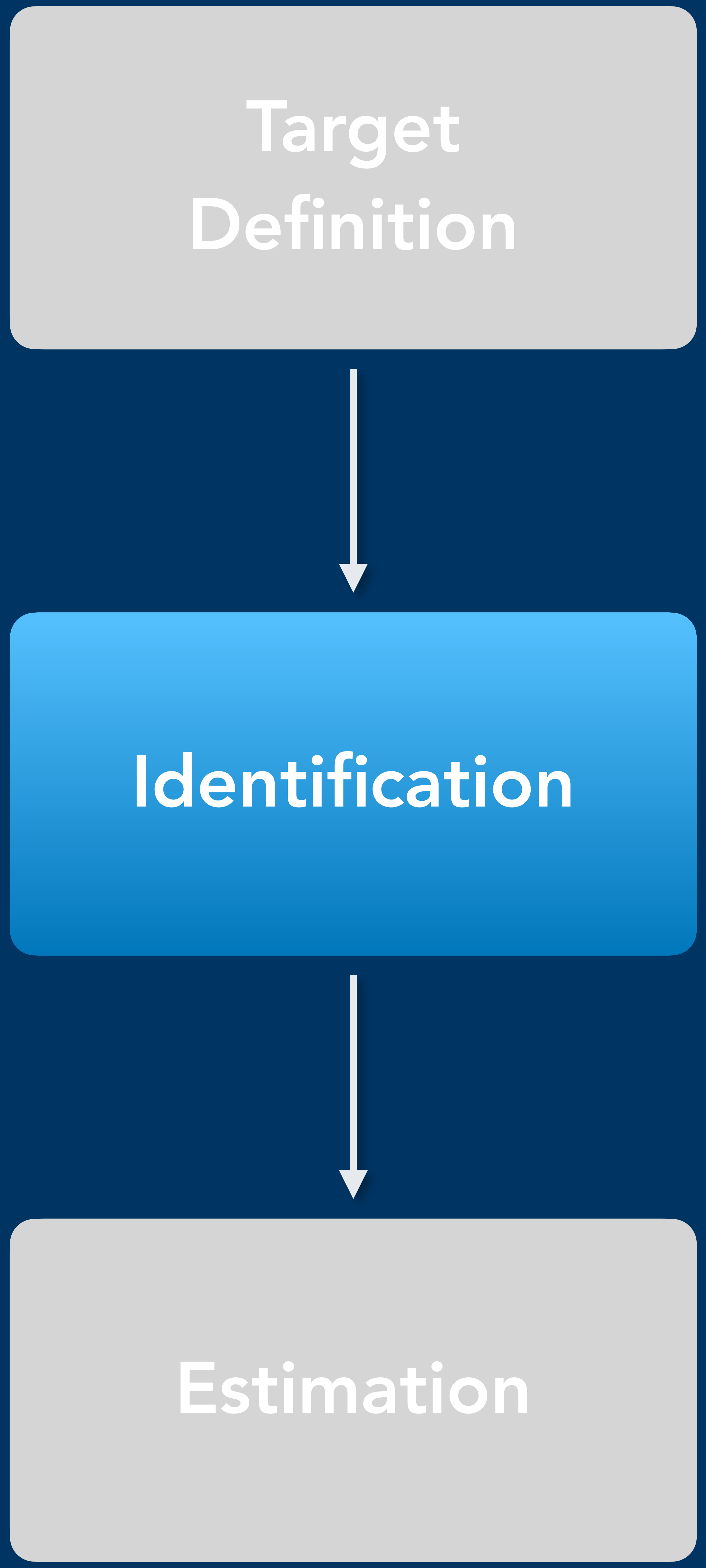
Average Treatment Effect

$$ATE = \mathbb{E}[Y^1 - Y^0]$$

ID	X	T	Y^1	Y^0
1	●	●	●	?
2	●	●	?	●
3	●	●	●	?
4	●	●	?	●
⋮			⋮	⋮
⋮			⋮	⋮
N	●	●	●	?

observe one or the other

Identification



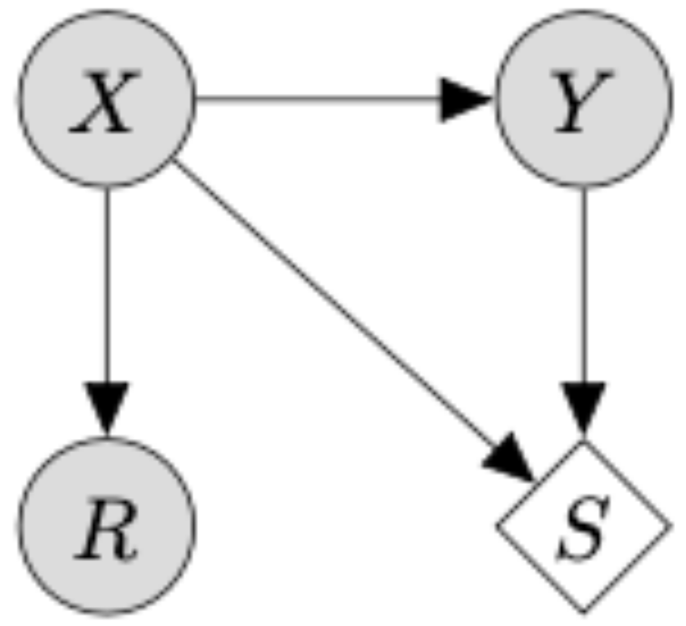
Identifying Condition #1: Missing-at-Random

(a.k.a. *ignorability* & *no unmeasured confounding*)

The **missing-at-random (MAR)** condition says that, given the input X, the decision to abstain R is independent of the base classifier's score $S = s(f(X), Y)$:

$$\boxed{S \perp\!\!\!\perp R \mid X\,.}$$

- MAR is satisfied as long as **the evaluation data** is independent from the classifier (Ppn. 4.1).
- Typically, predictions are **NOT** missing **completely** at random (MCAR), i.e., $S \not\perp\!\!\!\perp R$.

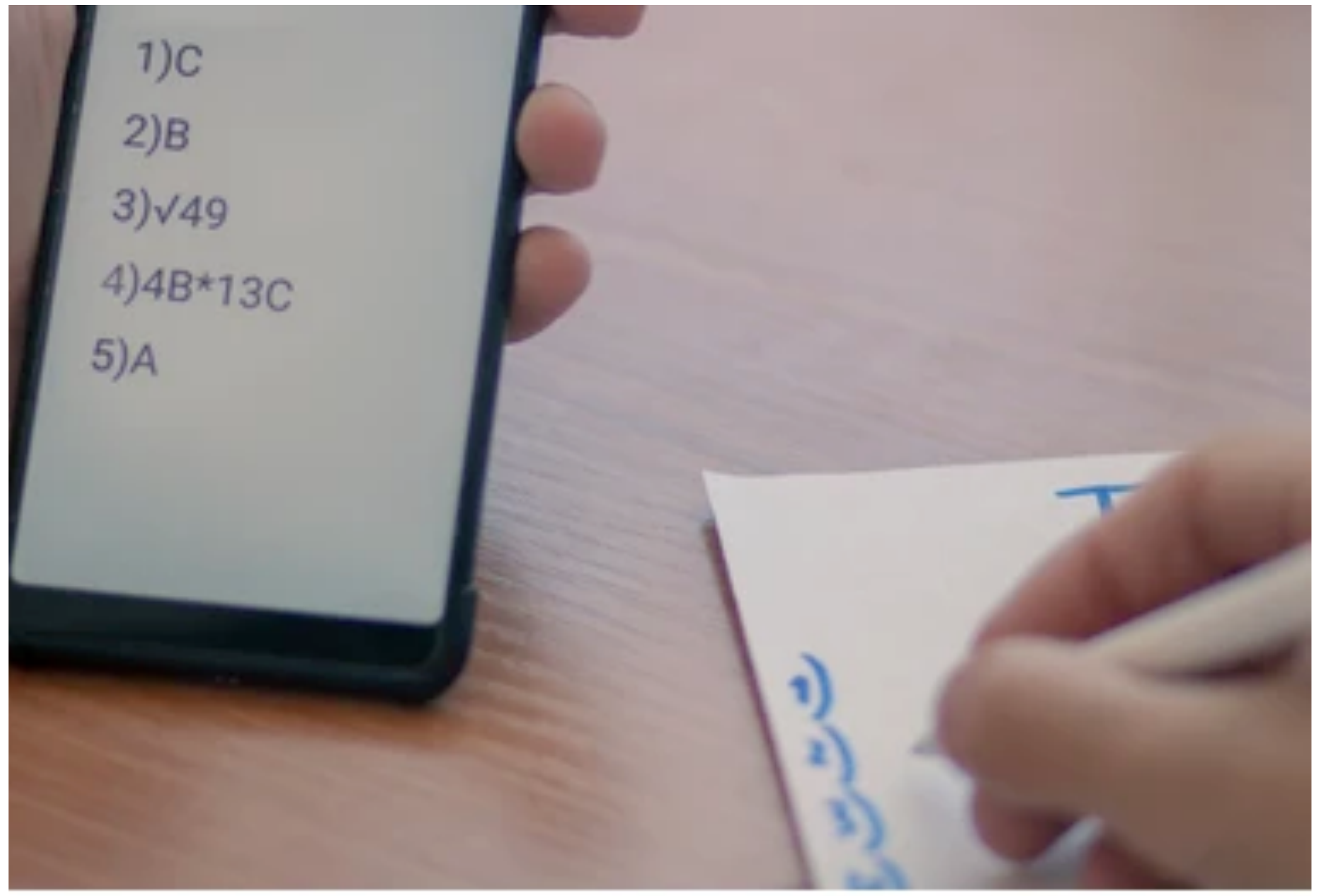


Conditioned on X,
S and R are d-separated.

Diamond <S> means partially observed.
(cf. missingness graphs by Mohan et al., 2013)

Can the MAR Condition Ever Be Violated?

- **The MAR condition is met as long as $(X, Y) \perp\!\!\!\perp \mathscr{D}_{\text{train}}$ (Ppn. 4.1), i.e., the classifier's training data is *independent* from the test data.**
  - This is expected in a typical setup for evaluating learning algorithms.
  - *If a classifier already saw the test data, then it would surely do better.*
- **Unfortunately, in a purely black-box setting, the evaluator may *not* know what training data was used by the classifier.**
  - E.g., large ML models pre-trained on publicly available datasets.
- **Practical suggestions for preventing/addressing MAR violations:**
  - **Use a test set that is not publicly available (e.g., patient data).**
  - Conduct sensitivity analysis, e.g., under a contamination model (Bonvini & Kennedy, 2022).



shutterstock.com · 1460735429

Identifying Condition #2: Positivity

The **positivity** condition for this problem requires that each abstaining classifier **cannot deterministically abstain** (on any meaningful input region):

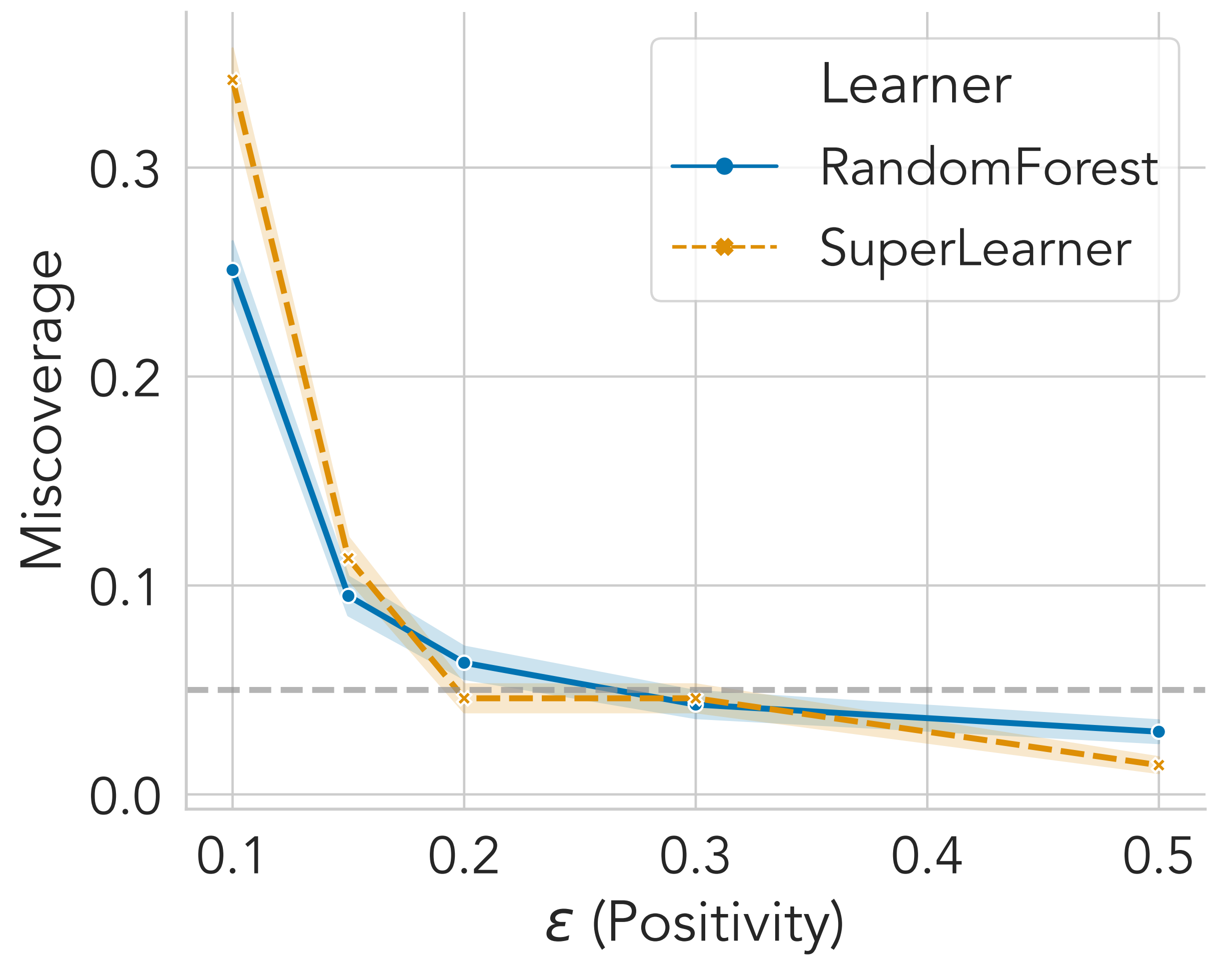
$$\exists \epsilon > 0 : \quad \pi(X) = \mathbb{P}(R = 1 \mid X) \leq 1 - \epsilon \,.$$

This is a **necessary** condition:

- If a classifier deterministically abstains on some nontrivial part of the input space, then there is no way of knowing what it would have done in that region.

How Can We Address the Positivity Condition?

- **Positivity violations can affect the validity and efficiency of the estimator.**
  - Yet, in practice, classifiers may abstain deterministically on certain inputs.
- **Argument 1: unidentifiability & a need for a policy-level approach.**
  - If a governing body seeks to audit commercial softwares for safety-critical tasks, then they must **require** vendors to match a level of positivity.
- **Argument 2: stochastic abstentions can improve performances.**
  - Kalai & Kanade (2021) showed that stochastic abstentions can improve **out-of-distribution (OOD) performance** of abstaining classifiers.
  - Schreuder & Chzhen (2021) derived a stochastically abstaining classifier that achieves good performance **subject to a fairness constraint**.



For $\epsilon < 0.2$,
the miscoverage rate of a 95% CI
rises above the intended level.

Identification

<u>Proposition.</u> Under the MAR and positivity conditions, we can identify the counterfactual score as an expectation over observables:

$$\boxed{\psi = \mathbb{E}[S] = \mathbb{E}[\mu_0(X)],}$$

where μ_0 is the score regression function: $\mu_0(x) = \mathbb{E}[S \mid R = 0, X = x]$.

In other words, **the target parameter can now be estimated with observed data.**

The rest of the problem is purely that of *functional estimation* (nothing causal).

One-Line Proof of Identification

Proposition. Under the MAR and positivity conditions, we can identify the counterfactual score as following:

$$\boxed{\psi = \mathbb{E}[S] = \mathbb{E}[\mu_0(X)],}$$

where μ_0 is the score regression function: $\mu_0(x) = \mathbb{E}[S \mid R = 0, X = x]$.

One-line proof using standard arguments: μ_0 is well-defined by positivity; then,

$$\psi = \mathbb{E}[S] = \mathbb{E}[\mathbb{E}[S \mid X]] \overset{\text{(MAR)}}{=} \mathbb{E}[\mathbb{E}[S \mid X, R = 0]] = \mathbb{E}[\mu_0(X)].$$

Identification for Δ^{AB}

Under the identifying assumptions, we have that:

$$\boxed{\Delta^{AB} = \mathbb{E}[S^A - S^B] = \mathbb{E}[\mu_0^A(X) - \mu_0^B(X)],}$$

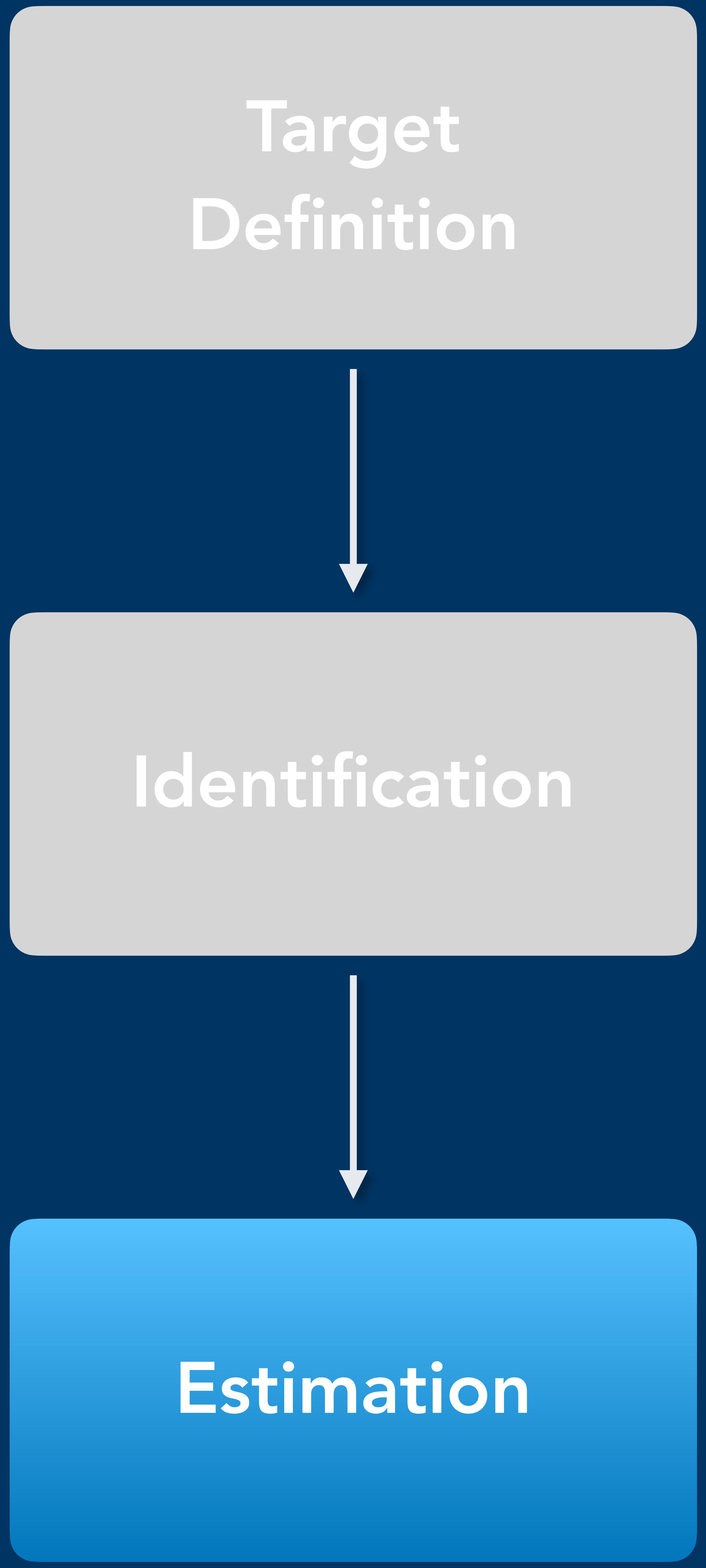
where

$$\mu_0^A(x) = \mathbb{E}[S^A \mid R^A = 0, X = x] \quad \text{and} \quad \mu_0^B(x) = \mathbb{E}[S^B \mid R^B = 0, X = x].$$

As before, **the target parameter can now be estimated with observed data!**

The rest of the problem is purely that of *function estimation* (and not causal).

Estimation



Other names: augmented IPW (Robins et al., 1994); targeted MLE (van der Laan & Rubin, 2006); double ML (Chernozhukov et al., 2018)

The Doubly Robust Estimator $\hat{\psi}_{\mathrm{dr}}$

Given an *i.i.d.* data of potentially missing predictions, $\{(X_i, R_i, (1 - R_i)S_i)\}_{i=1}^{n} \sim \mathbb{P}$, the **doubly robust (DR) estimator for ψ** is defined as:

$$\hat{\psi}_{\mathrm{dr}} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_0(X_i) + \frac{1 - R_i}{1 - \hat{\pi}(X_i)}\big(S_i - \hat{\mu}_0(X_i)\big)\right].$$

The summand is the **influence function** for $\mathbb{E}[\mu_0(X)]$ (a first-order bias correction).

For comparison, we can simply take the difference between the two classifiers ($\hat{\psi}^{A}_{\mathrm{dr}} - \hat{\psi}^{B}_{\mathrm{dr}}$).

*The nuisance functions, $\hat{\mu}_0$ and $\hat{\pi}$, are estimated via **cross-fitting** (K-fold sample splitting).

DR Estimator is Asymptotically Normal & Efficient

Theorem (DR estimation of the counterfactual score). Assume the identifying conditions hold & that the nuisance functions are estimated at a parametric rate *in product*:

$$\|\hat{\pi} - \pi\|_{L^2(\mathbb{P})}\|\hat{\mu}_0 - \mu_0\|_{L^2(\mathbb{P})} = o_{\mathbb{P}}(1/\sqrt{n}).$$

Then, assuming $\|\hat{IF} - IF\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(1)$, the DR estimator is asymptotically normal, and its variance matches the nonparametric (and locally minimax) efficiency bound:

$$\sqrt{n}\left(\hat{\psi}_{dr} - \psi\right) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}_{\mathbb{P}}[IF]\right).$$

An asymptotic CI for ψ can be constructed using the empirical estimate of $\mathrm{Var}_{\mathbb{P}}[IF]$.

cf. Robins et al. (1994); Bang & Robins (2005); many others later.

Understanding Double Robustness

The DR assumption says that the nuisance functions are estimated at a parametric rate *in product*:

$$\|\hat{\pi} - \pi\|_{L^2(\mathbb{P})}\|\hat{\mu}_0 - \mu_0\|_{L^2(\mathbb{P})} = o_{\mathbb{P}}(1/\sqrt{n}).$$

In particular,

- **Both nuisance functions can be learned at a *nonparametric* rate, say, $o_{\mathbb{P}}(n^{-1/4})$, such that the product of their rates of convergence is $o_{\mathbb{P}}(1/\sqrt{n})$.**
    - Allows complex nuisance learners, such as the super learner (stacking) and additive models. (In practice, random forests & deep neural nets can also work.)

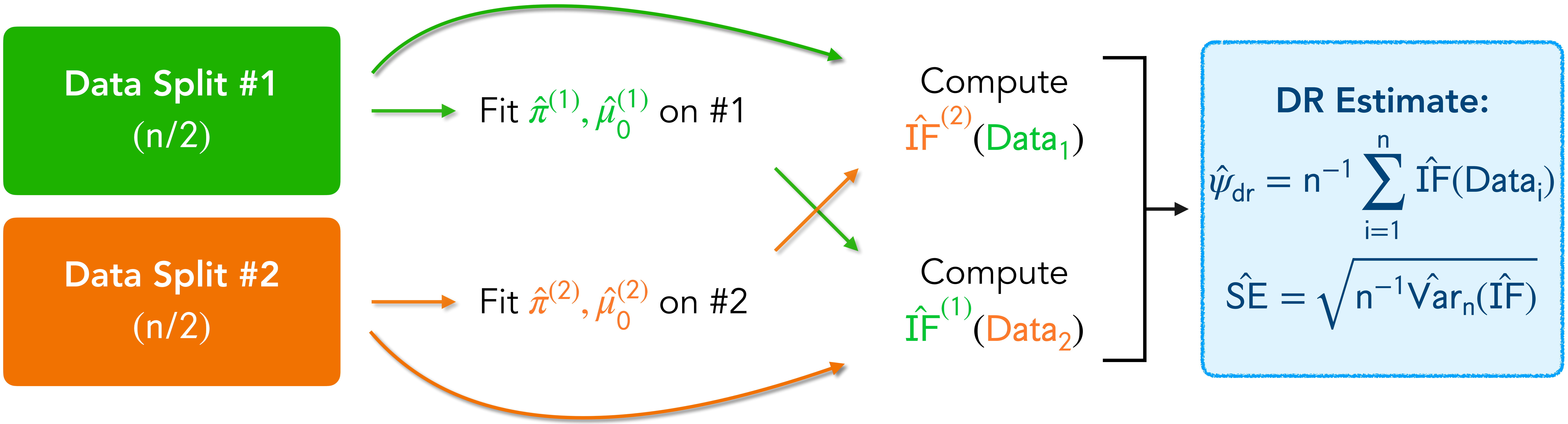
cf. Horowitz (2009); Kennedy (2018)

Learning the Nuisance Functions via Cross-Fitting

Computing $\hat{\text{IF}}$ requires learning the **nuisance functions** π and μ_0 from data:

$$\hat{\pi}(x) = \hat{\mathbb{P}}(R = 1 \mid X = x); \; \hat{\mu}_0(x) = \hat{\mathbb{E}}[S \mid R = 0, X = x].$$

Cross-fitting (Robins et al., 2008) allows us to learn them without losing sample efficiency.



cf. Robins et al. (2008); Zheng & van der Laan (2011); Chernozhukov et al. (2018)

DR Estimator for the Counterfactual Score Difference

<u>Theorem (DR estimation of the CF score *difference*).</u> Assume the identifying conditions hold & that the nuisance functions are estimated at a parametric rate *in product*:

$$\|\hat{\pi}^A - \pi^A\|_{L^2(\mathbb{P})}\|\hat{\mu}_0^A - \mu_0^A\|_{L^2(\mathbb{P})} + \|\hat{\pi}^B - \pi^B\|_{L^2(\mathbb{P})}\|\hat{\mu}_0^B - \mu_0^B\|_{L^2(\mathbb{P})} = o_{\mathbb{P}}(1/\sqrt{n}).$$

Let $IF^{AB} = IF^A - IF^B$. Assuming $\|\hat{IF}^{AB} - IF^{AB}\| = o_{\mathbb{P}}(1)$, the DR estimator is asymptotically normal, and its variance matches the nonparametric (and locally minimax) efficiency bound:

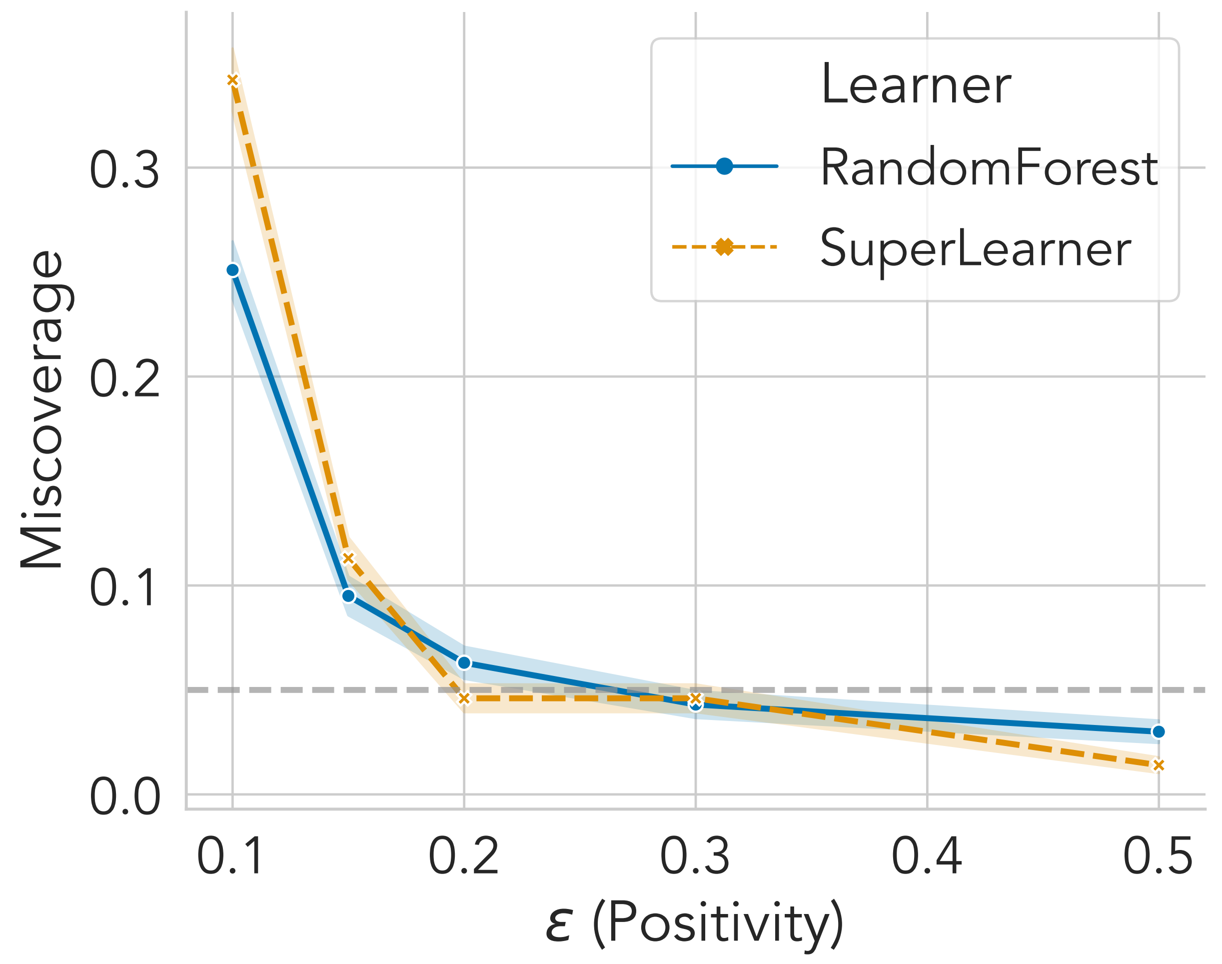
$$\sqrt{n}\left(\hat{\Delta}_{dr}^{AB} - \Delta^{AB}\right) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}_{\mathbb{P}}[IF^{AB}]\right).$$

An asymptotic CI for Δ^{AB}, or a hypothesis test for $H_0 : \psi^A = \psi^B$, can be constructed.

cf. Robins et al. (1994); Bang & Robins (2005); many others later.

How Can We Address the Positivity Condition?

- **Positivity violations can affect the validity and efficiency of the estimator.**
  - Yet, in practice, classifiers may abstain deterministically on certain inputs.
- **Argument 1: unidentifiability & a need for a policy-level approach.**
  - If a governing body seeks to audit commercial softwares for safety-critical tasks, then they must **require** vendors to match a level of positivity.
- **Argument 2: stochastic abstentions can improve performances.**
  - Kalai & Kanade (2021) showed that stochastic abstentions can improve **out-of-distribution (OOD) performance** of abstaining classifiers.
  - Schreuder & Chzhen (2021) derived a stochastically abstaining classifier that achieves good performance **subject to a fairness constraint**.

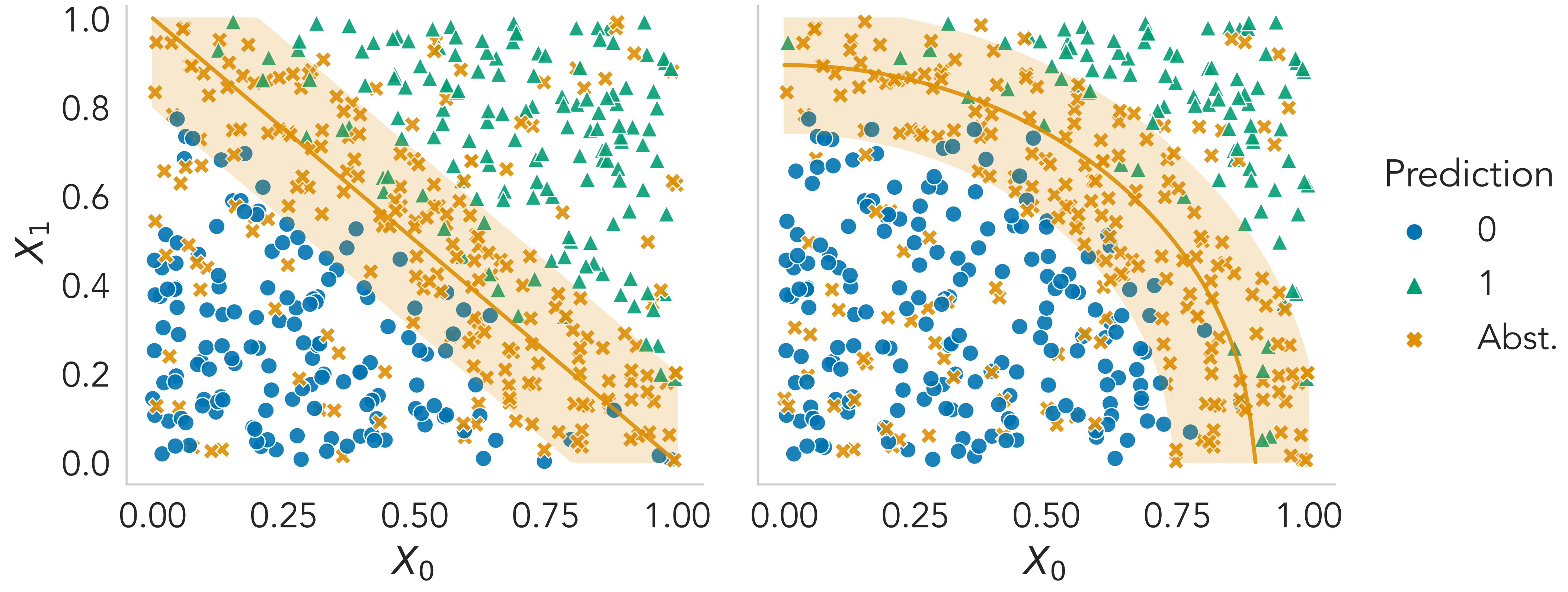


For $\epsilon < 0.2$,
the miscoverage rate of a 95% CI rises above the intended level.

Experiments

Simulated Experiment: CI Miscoverage & Width

A: linear classifier with the *optimal* decision boundary. **B:** *biased* classifier with a curved boundary.



Two abstaining classifiers, depicted using their decision boundary (orange), predictions (●/▲), and abstentions (x).

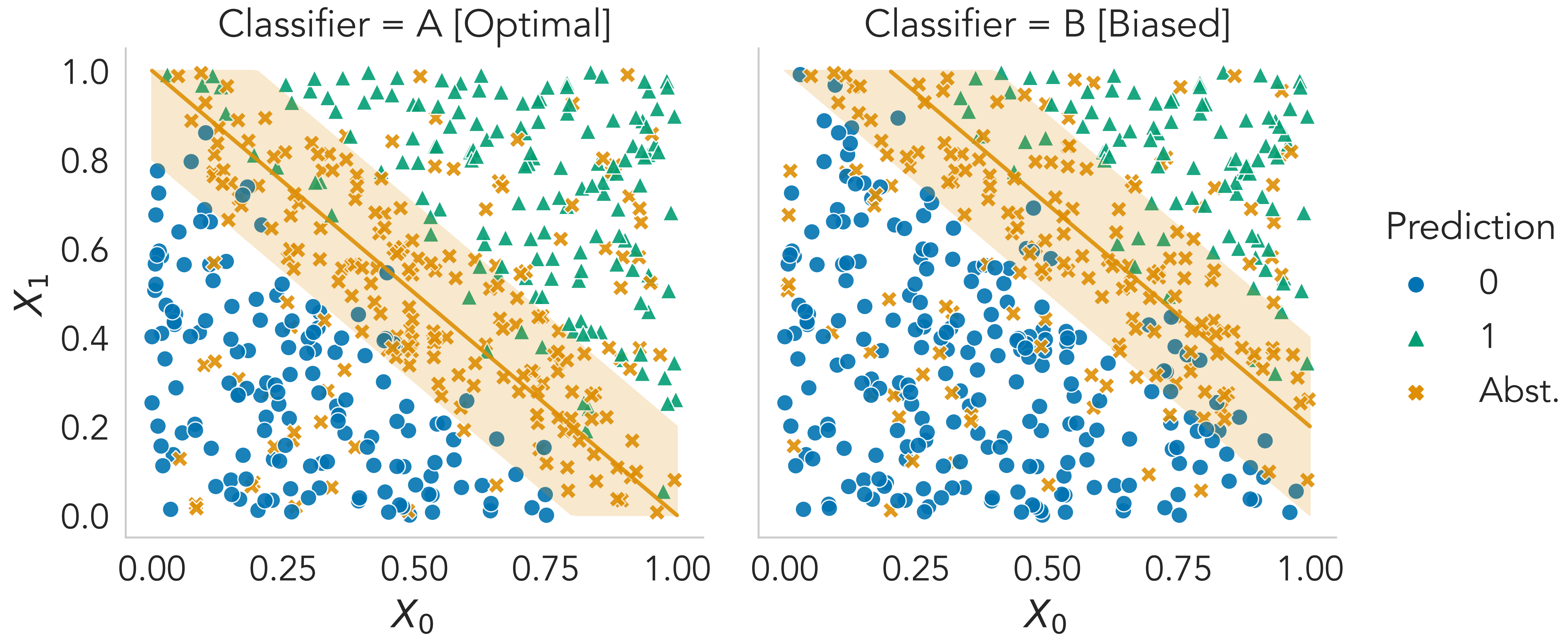
$\hat{\pi}$ / $\hat{\mu}_0$	*95% CI's*	Plug-in	IPW	DR
Random Forest	*Miscoverage*	0.64	0.14	**0.05**
	Width	*0.02*	*0.13*	*0.07*
Super Learner	*Miscoverage*	0.91	**0.03**	**0.05**
	Width	*0.01*	*0.12*	*0.06*

CI Miscoverage: rate of the 95% CI not covering the true Δ^{AB}, based on *accuracy*. (**Blue**: valid miscoverage.)
Width: upper minus lower confidence bound.
Both averaged over 1,000 repeated simulations.

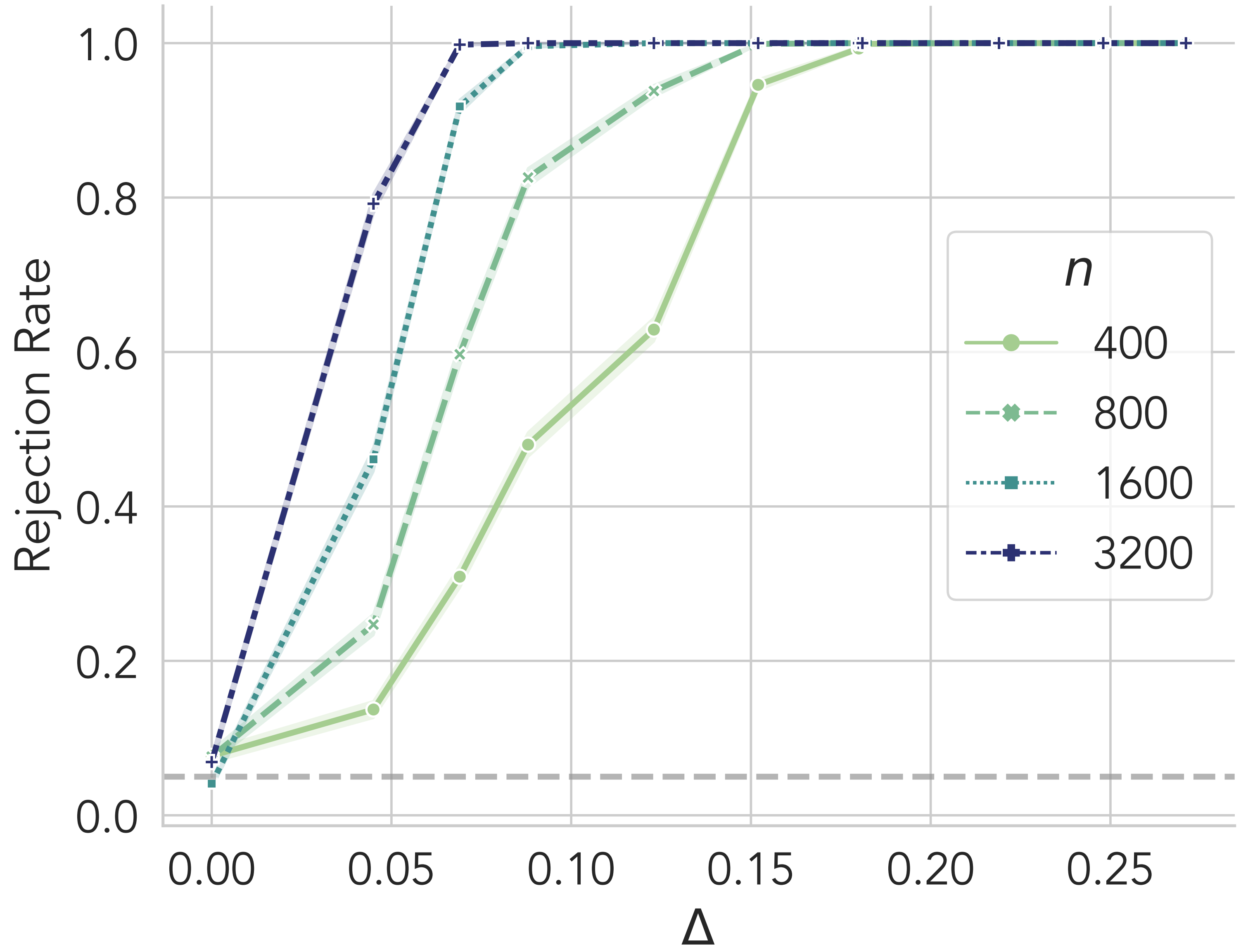
DR CI achieves the correct miscoverage rate **(small bias)**, and its width is half the width of the IPW CI **(small variance)**.

Simulated Experiment: Power Analysis

Setup: score difference (Δ^{AB}) grows larger as the suboptimal classifier (B) become more biased.



Two abstaining classifiers, depicted using their decision boundary (orange), predictions (●/▲), and abstentions (x). Δ^{AB}=0.1.



Power (rejection rate) of the hypothesis test for $H_0 : \Delta^{AB} = 0$.

Δ^{AB} increases as B shifts farther away from A.

Real Data Experiment: Comparing VGG-16 Classifiers on CIFAR-100

- **Setup:** We compare abstaining classifiers based off of a pre-trained VGG-16 deep convolutional neural network* for the CIFAR-100 dataset. Evaluation set size is 5,000.

- Nuisance functions $(\hat{\pi}^{A}, \hat{\mu}_0^{A}, \hat{\pi}^{B}, \hat{\mu}_0^{B})$ are learned on top of the pre-trained VGG-16 network, but they each use a different output layer (learned via cross-fitting).

Scenarios	**Base Classifier**	**Abstention Rule**	$\bar{\Delta}^{AB}$	**95% DR CI**	**Reject H_0?**
I	Same	Different	0.000	(-0.005, 0.018)	No
II	Same	Different	0.000	(-0.014, 0.008)	No
III	Different	Same	−0.029	(-0.051, -0.028)	Yes

Comparing VGG-16-Based Abstaining Classifiers on CIFAR-100 (n=5,000) using the Brier score.
Estimation target: $\Delta^{AB} := \psi^{A} - \psi^{B}$; null hypothesis $H_0 : \Delta^{AB} = 0$.

*reproduced version from https://github.com/chenyaofo/pytorch-cifar-models.

Summary of Contributions

- We propose the ***counterfactual score***, a novel evaluation metric for black-box abstaining classifiers that assess the expected score had the classifier not been allowed to abstain.

- The score and its framework reveals an **underexplored connection** between abstaining classifiers, black-box evaluation, and missing data / causal inference.

- We formalize the **identifying assumptions (MAR and positivity)** for the score and give examples of settings in which they can be justified.

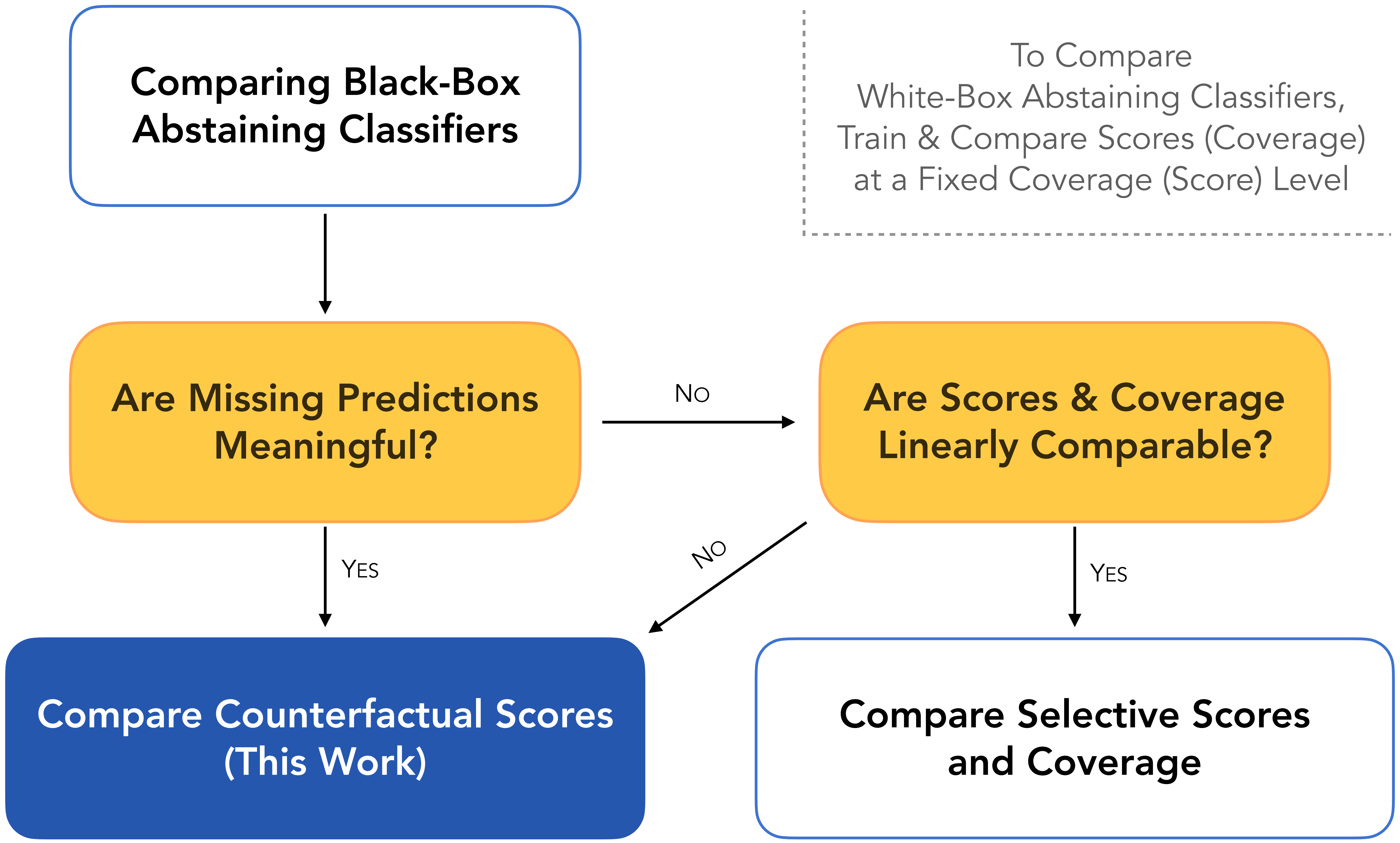
- We develop **nonparametrically efficient** estimators for the counterfactual score (difference), and empirically show their validity & efficiency on simulated/real datasets.

Thank You

Paper: https://arxiv.org/abs/2305.10564
Code: https://github.com/yjchoe/ComparingAbstainingClassifiers
NeurIPS Link: https://neurips.cc/virtual/2023/poster/72515
YJ's Webpage (for links to slides & poster): https://yjchoe.github.io/

Appendix

How Should We Compare Black-Box Abstaining Classifiers?

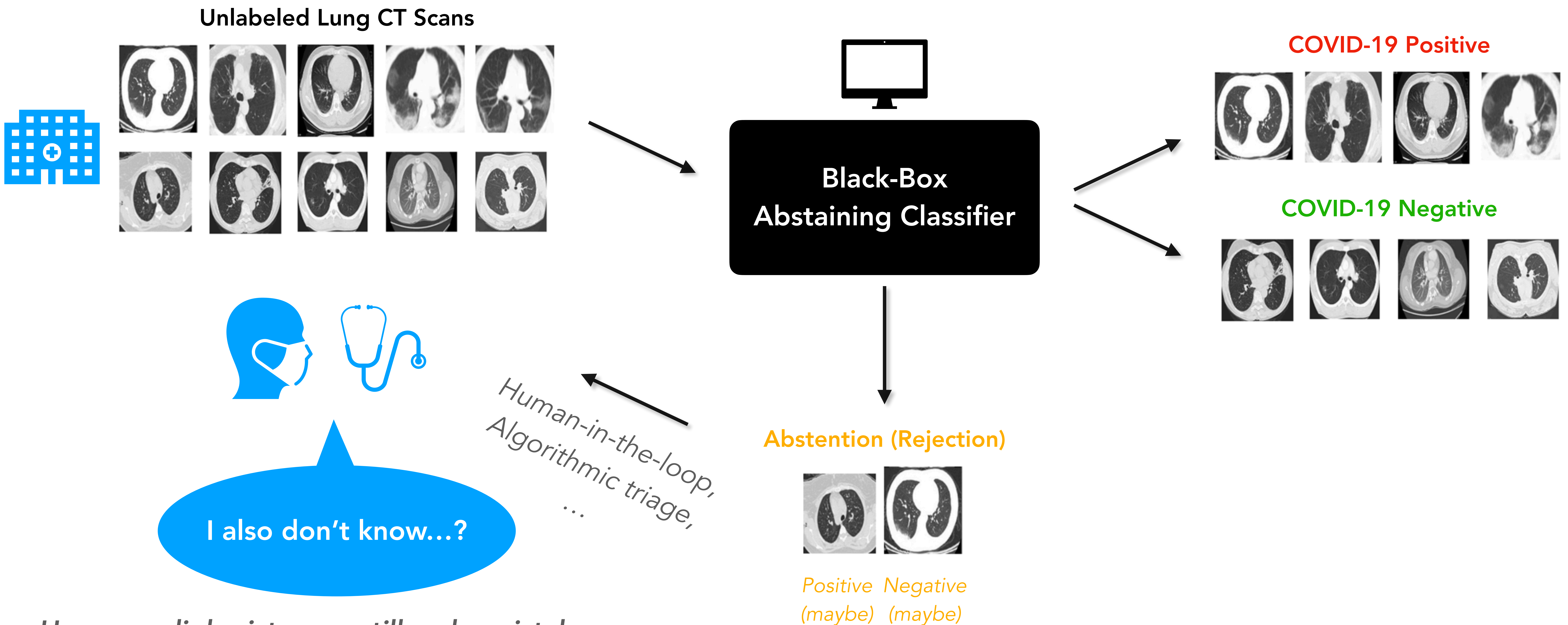


Summary of Problem Formulations & Approaches

Table 1: Problem formulation for the counterfactual evaluation of black-box abstaining classifiers. The proposed approaches are applicable to settings where the evaluator has no access to the underlying abstention mechanisms or base predictors, and they do not rely on parametric modeling assumptions.

	Evaluation	**Comparison**
Classifier(s)	(f, π)	$(f^{\mathsf{A}}, \pi^{\mathsf{A}})$ & $(f^{\mathsf{B}}, \pi^{\mathsf{B}})$
Target	$\psi = \mathbb{E}[S]$	$\Delta^{\mathsf{AB}} = \mathbb{E}\left[S^{\mathsf{A}} - S^{\mathsf{B}}\right]$
Identification	MAR & positivity	
Estimation	Doubly robust CI	
Optimality	Nonparametrically efficient	

Example #2: Secondary Diagnosis



Human radiologists may still make mistakes or possess cognitive biases (Busby et al., 2018).

Sample lung CT scans of COVID/non-COVID patients from Ahuja et al. (2020).

Example #3: Self-Driving Cars

<u>Example (Alert the Driver):</u> Let's say a semi-autonomous car deploys an abstaining image classifier that aids its driving decisions.

When it abstains, the car alerts the driver to take back the control. But…

- Unfortunately, NHSTA* reports that Tesla Autopilot can alert the driver during the very last seconds before a crash.
- Sometimes, the driver is just asleep or inattentive**. We'd still want to avoid accidents.

Can we evaluate the abstaining classifier *while accounting for* its performance even on its abstentions?

"Tesla Driver Caught On Camera Apparently Asleep At The Wheel" - NBC Nightly News (Sep 9, 2019)

*NHSTA: National Highway Traffic Safety Administration (U.S.)

**Research shows that the lack of active involvement correlates with tardy responses to takeover requests (Vogelpohl et al., 2019).

Abstentions as Missing Predictions

These examples illustrate cases where abstentions are really **missing predictions** that we'd like to know.

- The free-trial service example shows how the missing predictions have **direct uses in the future.**
- The self-driving car & secondary diagnosis examples shows how the missing predictions may be **used under a failure mode.**
- The LLM example shows how missing predictions may be utilized for the **assessment of internal biases**.

Comparison with Existing Evaluation Metrics

The counterfactual score $\psi = \mathbb{E}[\mathsf{S}]$ can be decomposed in the following way:

$$\boxed{\psi = \mathbb{E}[\mathsf{S} \mid \mathsf{R} = 0]\mathbb{P}(\mathsf{R} = 0) + \mathbb{E}[\mathsf{S} \mid \mathsf{R} = 1]\mathbb{P}(\mathsf{R} = 1)\,.}$$

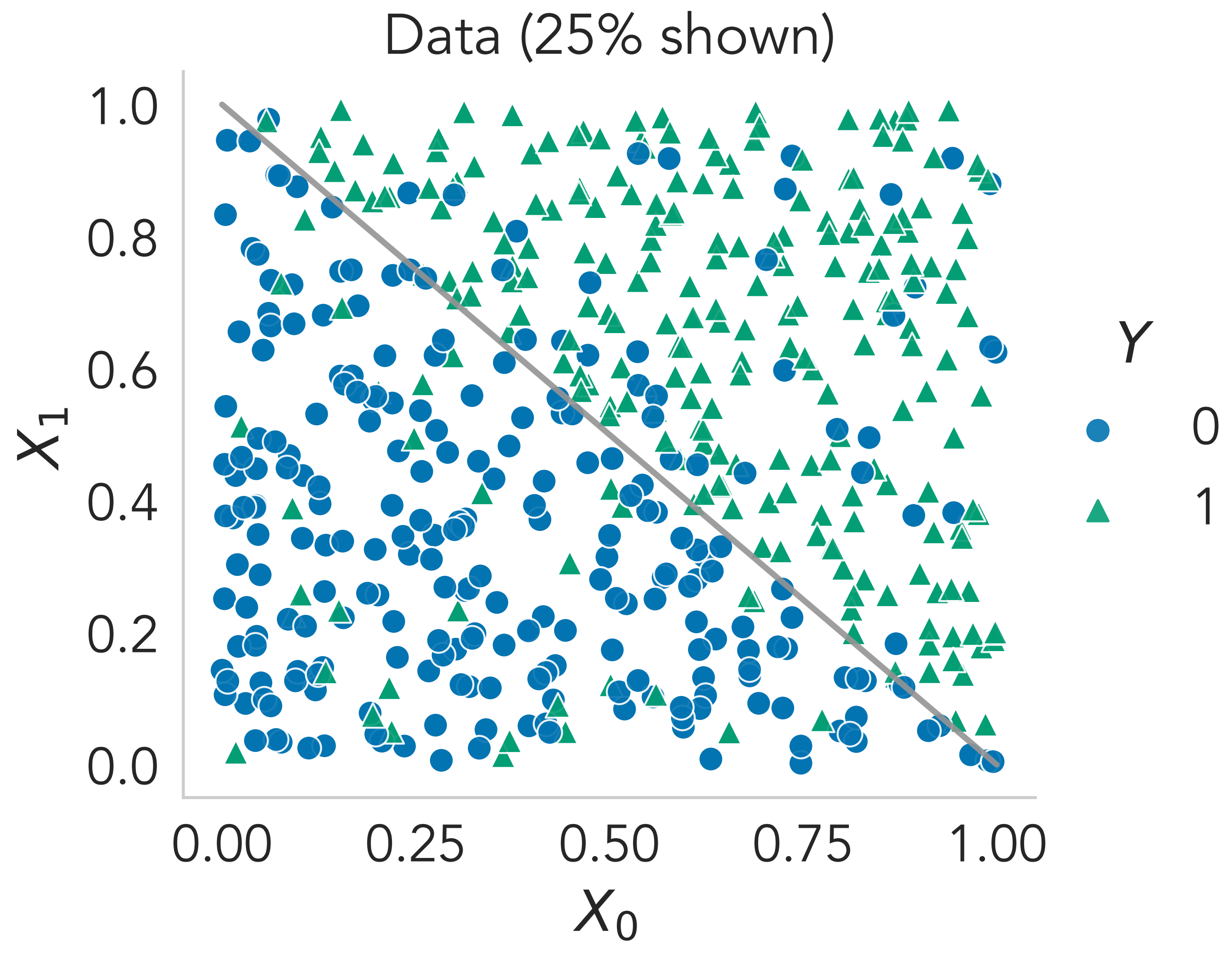
The first term is a product of the selective score and coverage (second term is ignored).

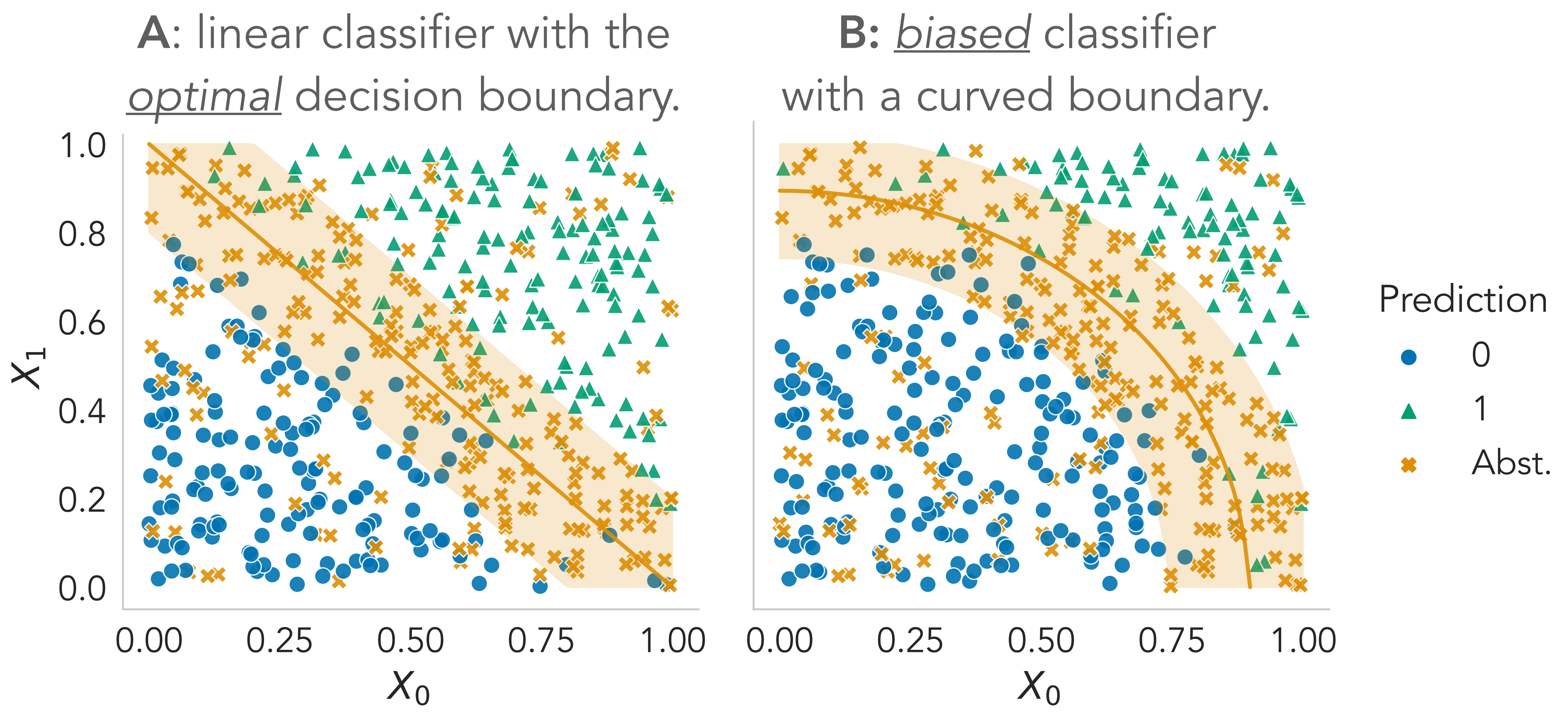
Condessa et al. (2017) proposes the **classification quality score** θ, assuming $\mathsf{S} \in [0, 1]$:

$$\theta = \mathbb{E}[\mathsf{S} \mid \mathsf{R} = 0]\mathbb{P}(\mathsf{R} = 0) + \mathbb{E}[1 - \mathsf{S} \mid \mathsf{R} = 1]\mathbb{P}(\mathsf{R} = 1)\,.$$

This would ***penalize*** abstaining on good predictions, which is not ideal in our applications. But it can also be estimated given our tools, as $\theta + \psi$ is an observable quantity.

Simulated Experiment: Data & Predictions



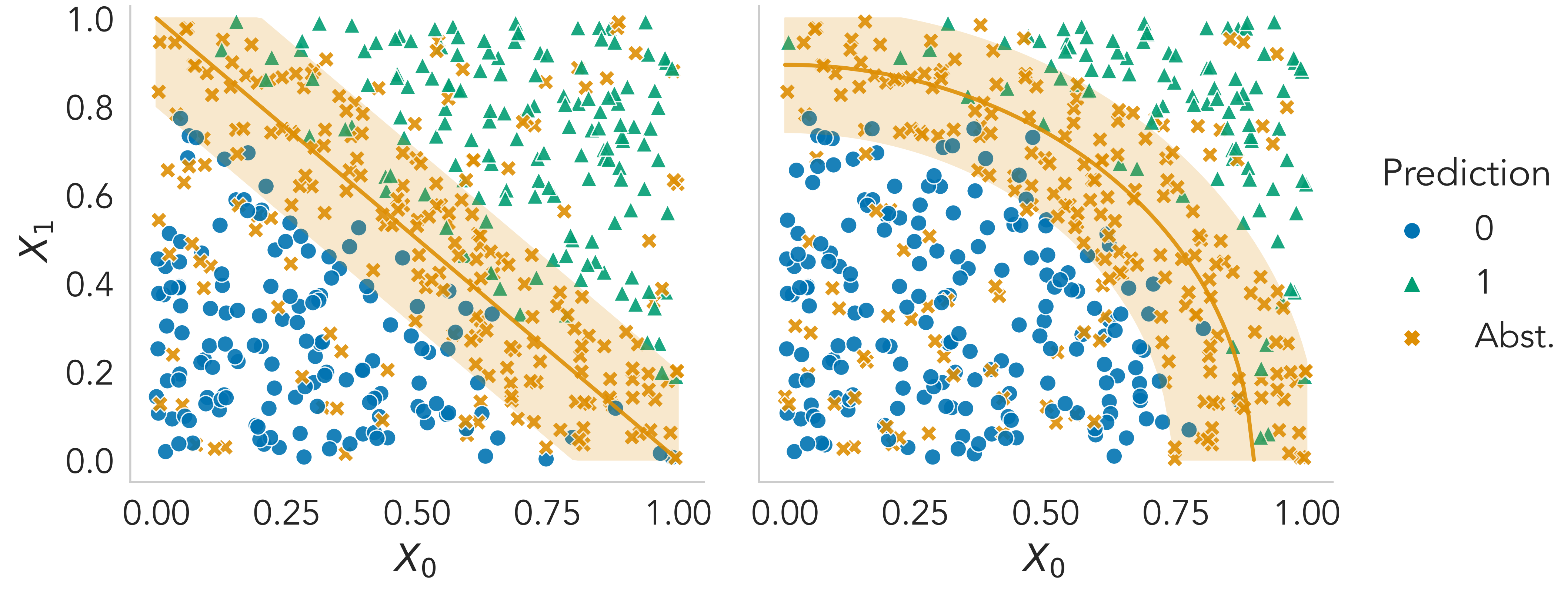


Two abstaining classifiers, depicted using their decision boundary (orange), predictions (●/▲), and abstentions (x).

Simulated Experiment: Full Results

A: linear classifier with the *optimal* decision boundary.

B: *biased* classifier with a curved boundary.



Two abstaining classifiers, depicted using their decision boundary (orange), predictions (●/▲), and abstentions (x).

$\hat{\pi}$ / $\hat{\mu}_0$	95% CI's	Plug-in	IPW	DR
Linear/Logistic	*Miscoverage*	1.00	0.76	1.00
	Width	*0.00*	*0.09*	*0.04*
Random Forest	*Miscoverage*	0.64	0.14	**0.05**
	Width	*0.02*	*0.13*	*0.07*
Super Learner	*Miscoverage*	0.91	**0.03**	**0.05**
	Width	*0.01*	*0.12*	*0.06*

CI Miscoverage*: rate of the 95% CI not covering the true Δ^{AB}, based on *accuracy*. (**Blue**: valid miscoverage.) **Width:** upper minus lower confidence bound. Both averaged over 1,000 repeated simulations.

With sufficiently flexible nuisance learners,
DR CI achieves the correct miscoverage rate **(small bias)**, and
its width is half the width of the IPW CI **(small variance)**.

**Not to be confused with the coverage of an abstaining classifier.*

Details for the CIFAR-100 Experiment

- **Scenario I: same** base classifiers (pre-trained VGG-16) & **different** thresholds for the softmax response (SR) (at 0.8 vs. 0.5).

$$\mathrm{SR}(\mathbf{p}) = \max_{c \in [C]} p_c.$$

  - *Note that these are deterministic abstention rules (still works, as the two happen to abstain on similar examples* and *their scores on abstentions happen to be similar).*

- **Scenario II: same** base classifiers & **different** stochastic abstention rules (SR vs. Gini).

- **Scenario III: different** base classifiers (1 vs. 2 output layers) & **same** abstention rules (SR).

- *Note: First half (5,000) of the "test set" is used to train the output layers.*

Asymptotic Confidence Sequences for Counterfactual Scores

- Leveraging the recent results by Waudby-Smith et al. (2021), we can further estimate the counterfactual scores of abstaining classifiers in an *anytime-valid* manner (i.e., at arbitrary stopping times).
- Informally, an **asymptotic confidence sequence (AsympCS)** refers to a sequence of intervals that is an arbitrarily precise approximation to a non-asymptotic CS, as $n \to \infty$.

Theorem. Let $\psi = \mathbb{E}[S]$ be the counterfactual score of an abstaining classifier. Assume an (i.i.d.) test set $\{(X_i, Y_i)\}_{i=1}^{n}$. Also, let $\hat{\psi}_{dr}$ be the DR estimator. If the nuisance functions for $\hat{\psi}_t$ are estimated at a product $o_{a.s.}(\sqrt{n^{-1}\log\log n})$ rate, then, for each $\alpha \in (0,1)$,

$$C_n := \left(\hat{\psi}_{dr} \pm \sqrt{\widehat{\mathrm{Var}}_n(\widehat{\mathrm{IF}})} \sqrt{n^{-2}(2n\hat{\sigma}_n^2 + 1) \cdot \log\left(\alpha^{-1}\sqrt{n\hat{\sigma}_n^2 + 1} \right)} \right)$$ forms a $(1-\alpha)$-level AsympCS for ψ.

where $\hat{\sigma}_n^2$ is the variance estimate of $\hat{\psi}_n$.

End of Slides