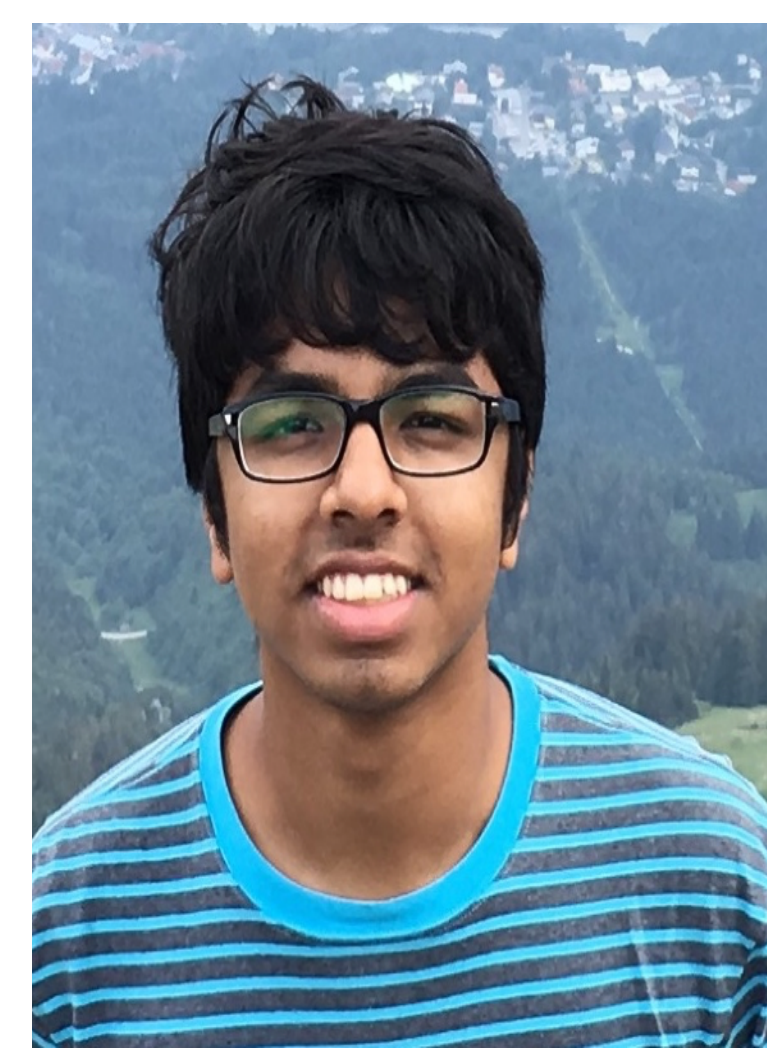


Treatment Effect Estimation Using Invariant Risk Minimization



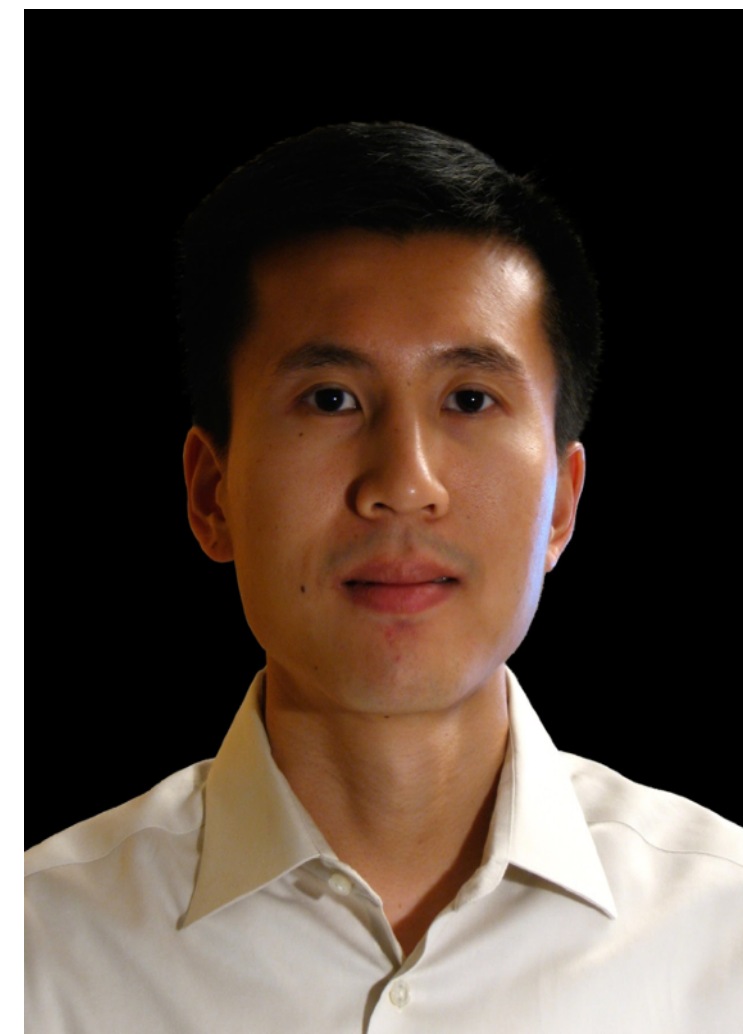
Abhin Shah



Kartik Ahuja



Karthikeyan Shanmugam



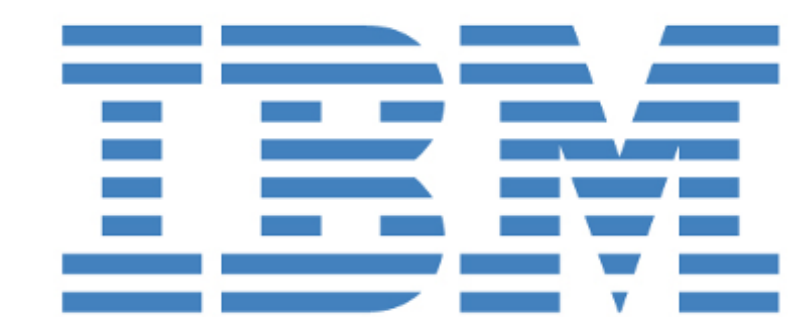
Dennis Wei



Kush R. Varshney



Amit Dhurandhar

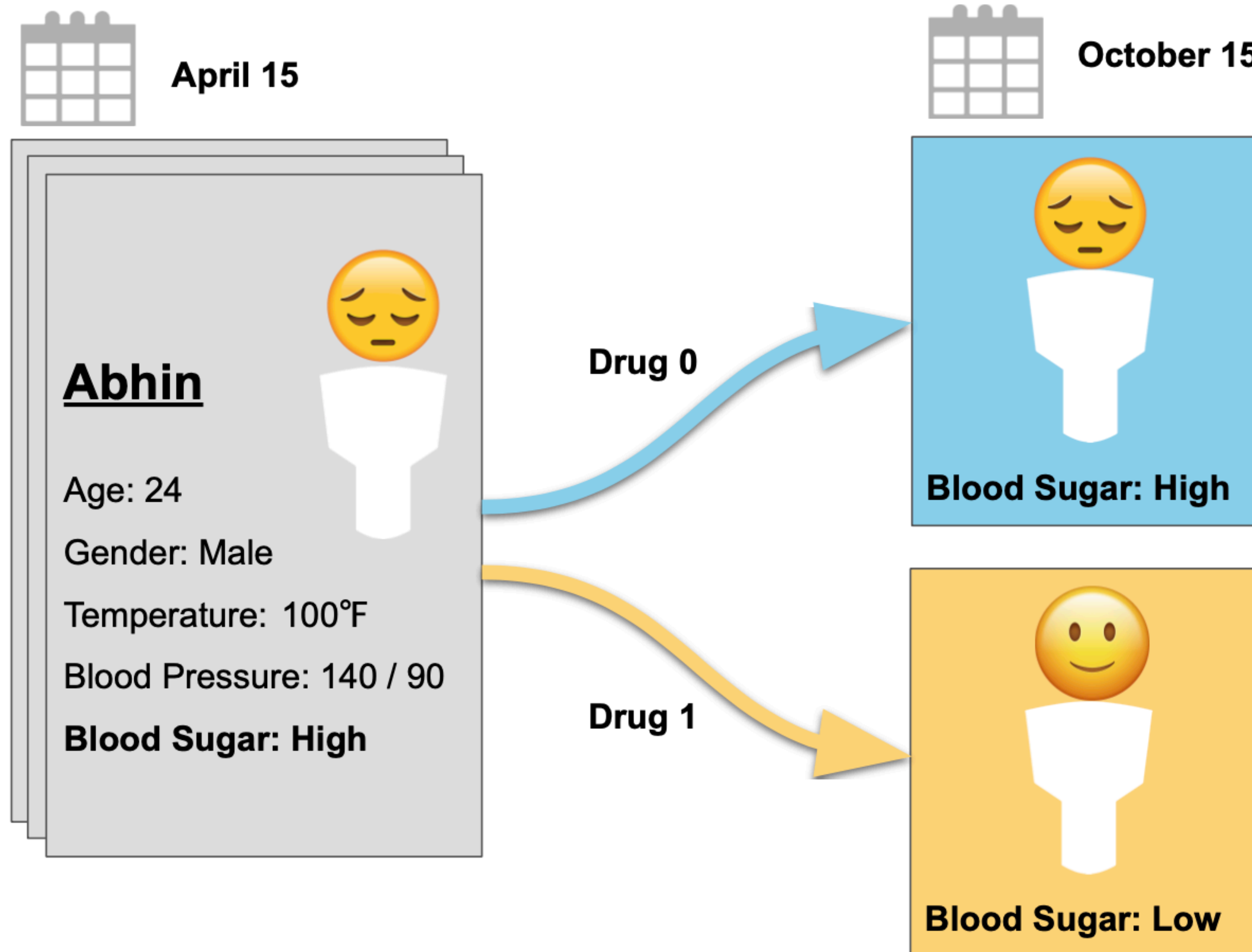


ICASSP 2021



Treatment Effect Estimation

Goal: Understand the causal effect of a treatment on an individual



Treatment Effect Estimation

Applications

- Understanding how a certain medication affects a patient's health
- Understanding how Yelp ratings influence a potential restaurant customer
- Evaluating the effect of a policy on unemployment rates
- Estimating the influence of individuals in a social network

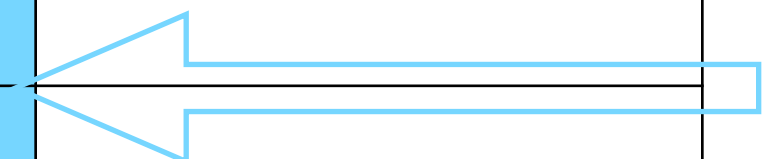
Observational Data

Contains past actions and their responses

Patient	Age	Blood Pressure	Drug	Blood Sugar
A	22	145/95	0	Low
B	26	135/80	0	Low
C	58	130/70	1	Low
D	50	145/80	1	High
E	24	150/85	1	Low

Observational Data

Patient	Age	Blood Pressure	Drug	Blood Sugar
A	22	145/95	0	Low
B	26	135/80	0	Low
C	58	130/70	1	Low
D	50	145/80	1	High
E	24	150/85	1	Low



Control Group

Observational Data

Patient	Age	Blood Pressure	Drug	Blood Sugar
A	22	145/95	0	Low
B	26	135/80	0	Low
C	58	130/70	1	Low
D	50	145/80	1	High
E	24	150/85	1	Low

Control Group

Treatment Group

Observational Data

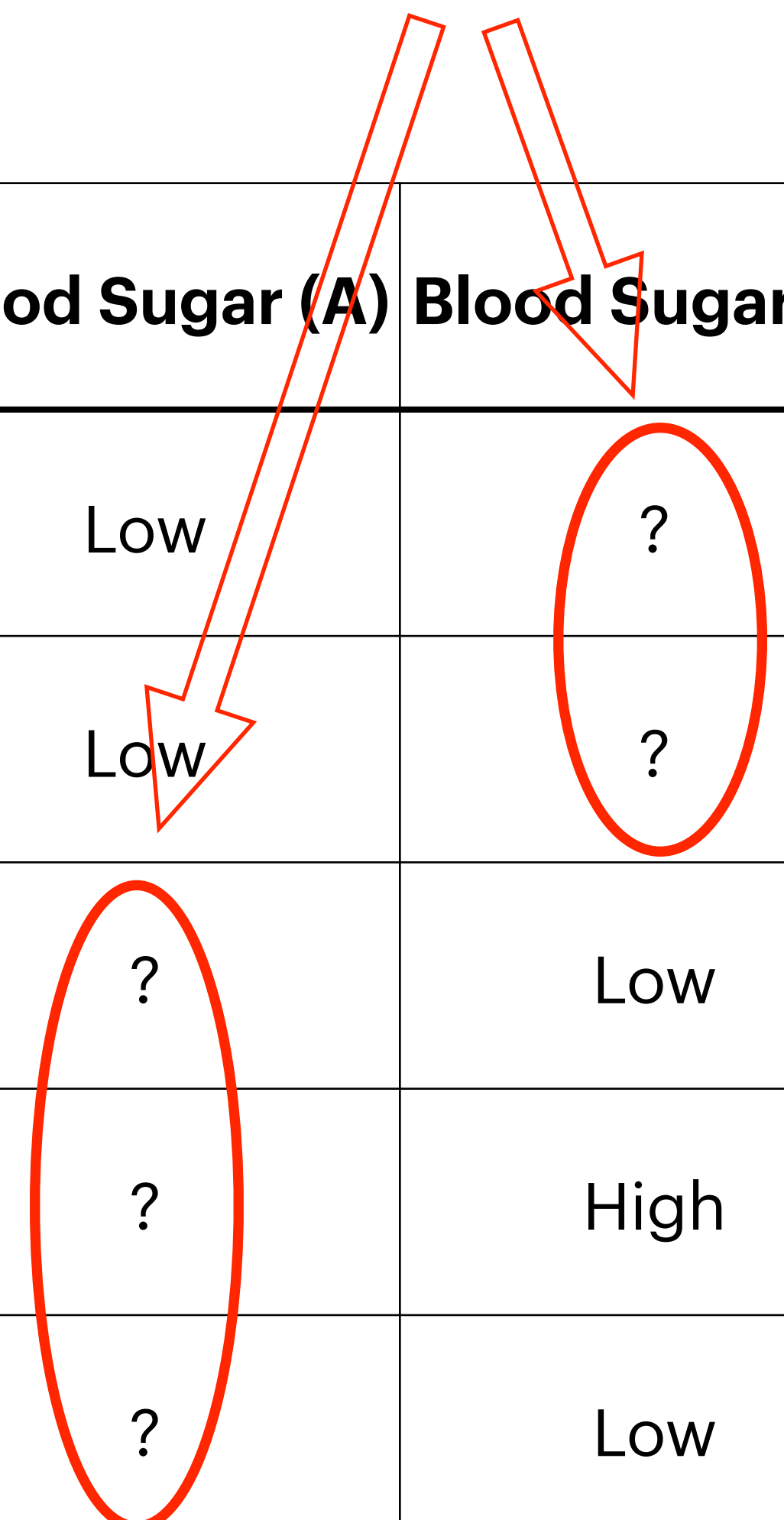
Observed factual outcomes

Patient	Age	Blood Pressure	Drug	Blood Sugar (0)	Blood Sugar (1)
A	22	145/95	0	Low	?
B	26	135/80	0	Low	?
C	58	130/70	1	?	Low
D	50	145/80	1	?	High
E	24	150/85	1	?	Low

Observational Data

Unobserved counterfactuals
Missing not at random!

Patient	Age	Blood Pressure	Drug	Blood Sugar (A)	Blood Sugar (B)
A	22	145/95	0	Low	?
B	26	135/80	0	Low	?
C	58	130/70	1	?	Low
D	50	145/80	1	?	High
E	24	150/85	1	?	Low

The image contains red annotations on the table. Two red arrows point from the text 'Unobserved counterfactuals' to the 'Blood Sugar (A)' cells for patients A and B. Another two red arrows point from the text 'Missing not at random!' to the 'Blood Sugar (B)' cells for patients A and B. Additionally, a red oval encircles the 'Blood Sugar (A)' cells for patients C, D, and E, and another red oval encircles the 'Blood Sugar (B)' cells for patients A and B.

Setup

- $x \in \mathbb{R}^d$ — Covariates / features
- $t \in \{0,1\}$ — Treatment assignment
- $y_0, y_1 \in \mathbb{R}$ — Potential outcomes under $t = 0$ and $t = 1$
- $y_f = t \times y_1 + (1 - t) \times y_0$ — Factual outcome

Individual Treatment Effect

Inference Task : Estimate Individual Treatment Effect (ITE)

$$\tau(x) = y_1(x) - y_0(x)$$

- Estimating ITE from observational data differs from classical supervised learning because we never observe the ITE in our training data.

Treatment assignment bias

Observational data is often prone to treatment assignment bias

- Patients receiving drug '0' may have a higher natural tendency (due to their age) to have low blood sugar than patients receiving drug '1'.
- The control group and the treatment group can have very different distributions.
- Traditional supervised learning model trained to predict the effect of treatment would fail to generalize well to the entire population.

Empirical Risk Minimization (ERM)

Minimize the average loss on the training data

1. OLS/LR1 - fit a single linear model to estimate $\mathbb{P}(y | x, t)$
2. OLS/LR2 - fit separate linear models to estimate $\mathbb{P}(y | x, t = 0)$ and $\mathbb{P}(y | x, t = 1)$

Invariant Risk Minimization (IRM)

A recent Domain Generalization framework

- ‘Invariant’ features : features whose predictive power is invariant across domains.
- ‘Spurious’ features : features whose predictive power varies across domains.
- **Goal** : identifies which properties of the training data describe spurious correlations and which properties represent phenomenon of interest.

Correlation vs Causation



Train distribution

≠

Test distribution

- Minimizing training error leads machines into recklessly absorbing **all** the correlations found in training data.
- However, *spurious* correlations stemming from data biases are unrelated to the *causal* explanation of interest.

Causation \Rightarrow invariance

- Assume that the training data is collected into distinct, separate *environments*.
- *Spurious* correlations **do not** appear to be *stable* properties across training environments.
- Promote learning correlations that are **stable** across training environments, as these should also hold in novel testing environments.

Contributions

- We propose an IRM-based ITE estimator aimed at tackling treatment assignment bias when there is little support overlap between the control group and the treatment group.
- We accomplish this by creating *diversity*: given a single dataset, we split the data into multiple domains artificially.
 - ⇒ These diverse domains are then exploited by IRM to more effectively generalize regression-based models to data regions that lack support overlap.
- We show gains over classical regression approaches to ITE estimation in settings when support mismatch is more pronounced.

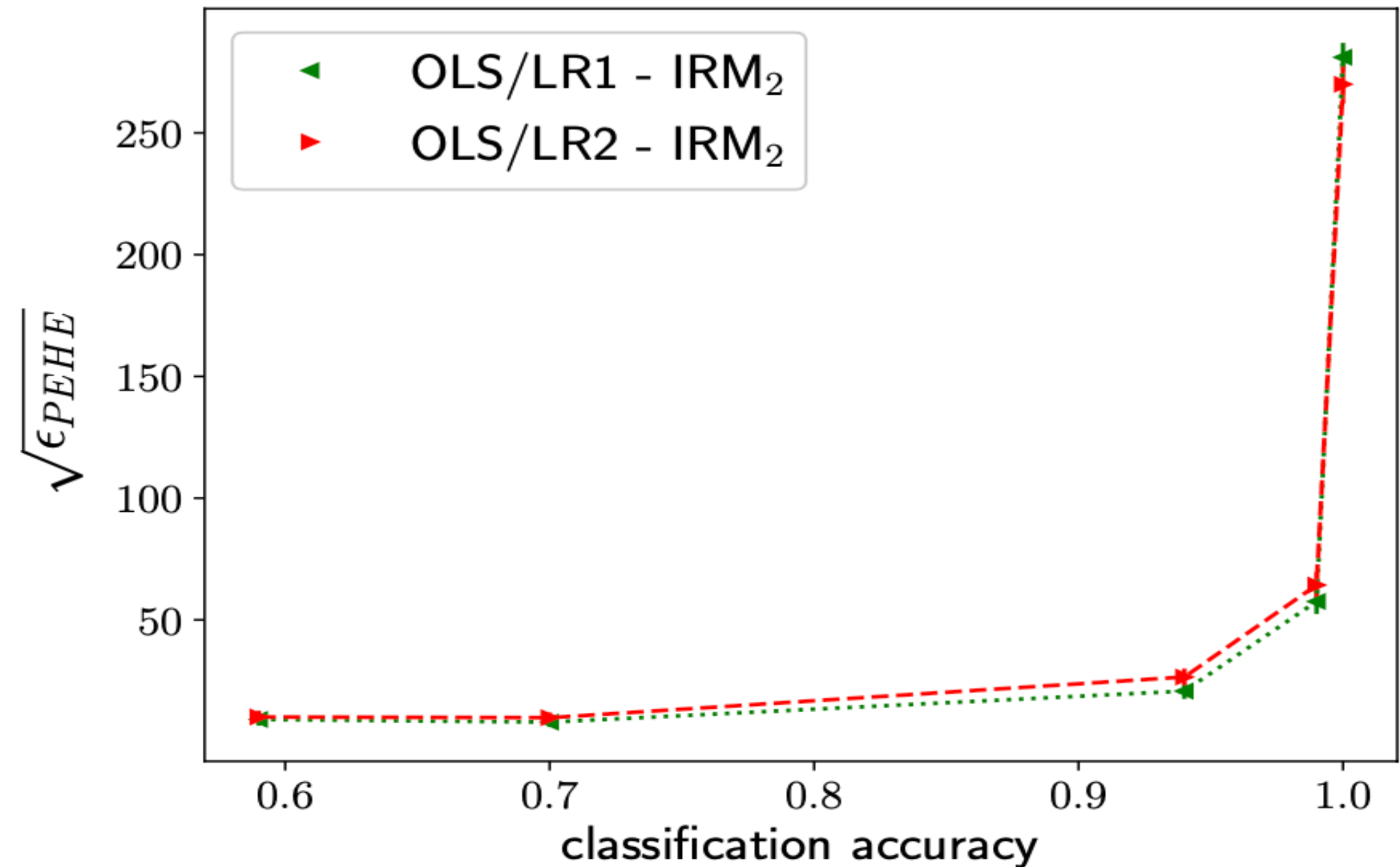
Synthetic Data

- $t \sim \text{Bernoulli}(0.5)$
- $x | t \sim N(\mu_t, \Sigma)$
- $y_t | x, t \sim N(x^T A_t x + x^T b_t + c_t, \sigma^2)$
- $e \sim \text{Bernoulli}(0.5)$

Performance Metric

$$\epsilon_{PEHE} = \frac{1}{n} \sum_{i=1}^n (\tau^{(i)} - \hat{\tau}^{(i)})^2$$

$\sqrt{\epsilon_{PEHE}}$ difference vs treatment group classification accuracy



Thank you!