When Shifts Happen — Confounding is to Blame

Gowtham Reddy Abbavaram
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Joint work with:

Celia Rubio-Madrigal

Rebekka Burkholz

Krikamol Muandet

ICLR

CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Empirical Observations and Open Questions

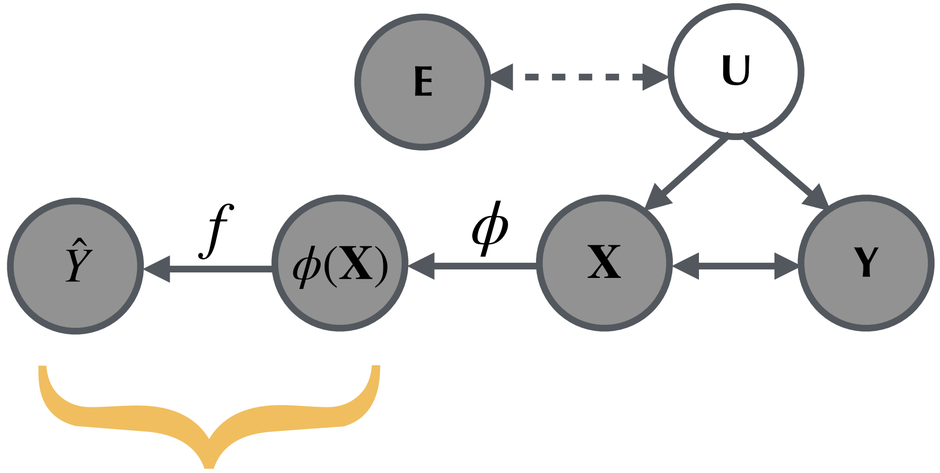
- Performance of MLP/Boosting $\approx$ performance of OOD generalization methods. Why?
- Performance with more variables > performance with only causal variables. Why?

Research Outcomes

- Hidden confounding shift is the reason
- Confounder-specific predictors perform better than invariance learning methods
- Using non-causal covariates help in approximating invariance

Causal Model

$\hat{Y} = (f \circ \phi)(\mathbf{X})$



Data Generating Process

- $\mathbf{X}$ is the set of covariates
- Y is the outcome
- U is a continuous (hidden) confounder
- E is a discrete environment variable
- E encodes a shift in $\mathbb{P}(U)$ across envs.

Computation Process

- $\phi(\mathbf{X})$ is a learned representation
- $\hat{Y}$ is the predicted outcome

Goal

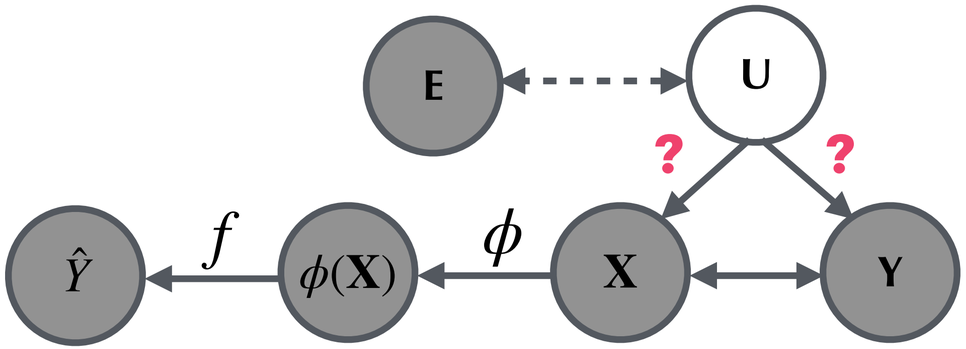
Achieve high predictive information $I(Y; \hat{Y})$

A General Decomposition of Predictive Information $I(Y; \hat{Y})$

Theorem-1

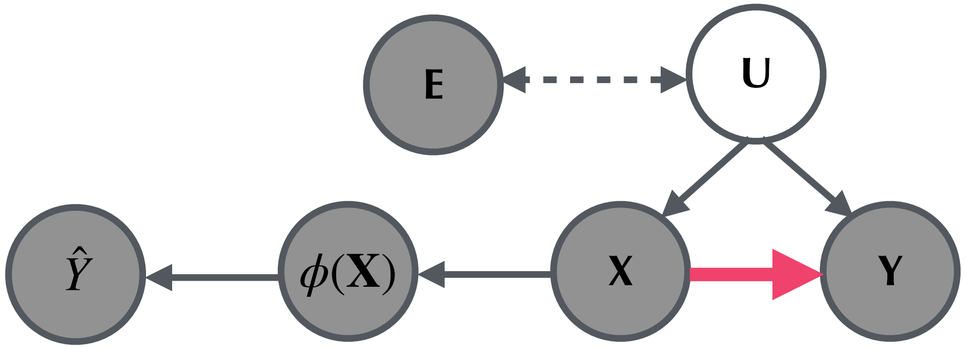
If we don't know how U influences $\mathbf{X}$, Y and if $\mathbf{X} \leftrightarrow Y$

$$\hat{Y} = (f \circ \phi)(\mathbf{X})$$



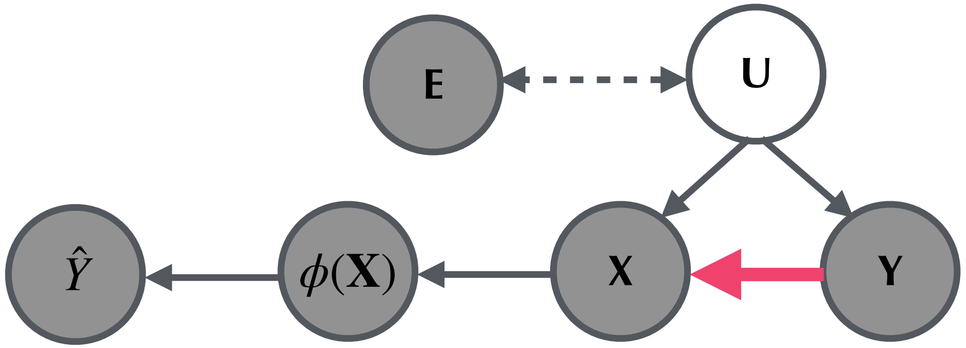
$$I(Y; \hat{Y}) = \overbrace{I(\phi(\mathbf{X}); Y \mid E)}^{\text{Cond. Inform.}} - \frac{\overbrace{I(\phi(\mathbf{X}); E \mid Y)}^{\text{Variation}}}{2} + \frac{\overbrace{I(Y; E)}^{\text{Label shift}}}{2} + \frac{\overbrace{I(\phi(\mathbf{X}); E)}^{\text{Feature shift}}}{2} - \frac{\overbrace{I(Y; E \mid \phi(\mathbf{X}))}^{\text{Concept shift}}}{2} - \overbrace{I(\phi(\mathbf{X}); Y \mid \hat{Y})}^{\text{Residual}}$$

Predictive Information Under Hidden Confounding Shift



$$\overbrace{I(\phi(\mathbf{X}); E \mid Y)}^{\text{Variation}} \geq \overbrace{I(\phi(\mathbf{X}); E)}^{\text{Feature shift}}$$

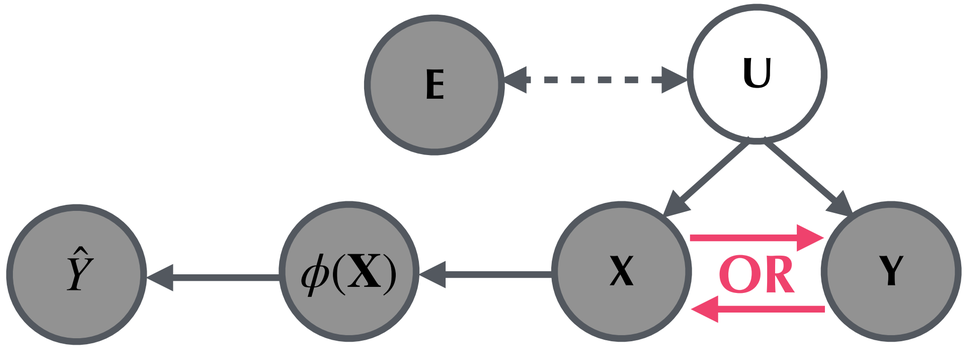
$$\overbrace{I(Y; E)}^{\text{Label shift}} \geq \overbrace{I(Y; E \mid \phi(\mathbf{X}))}^{\text{Concept shift}}$$



$$\overbrace{I(\phi(\mathbf{X}); E \mid Y)}^{\text{Variation}} \leq \overbrace{I(\phi(\mathbf{X}); E)}^{\text{Feature shift}}$$

$$\overbrace{I(Y; E)}^{\text{Label shift}} \leq \overbrace{I(Y; E \mid \phi(\mathbf{X}))}^{\text{Concept shift}}$$

Predictive Information Under Hidden Confounding Shift



Theorem 2

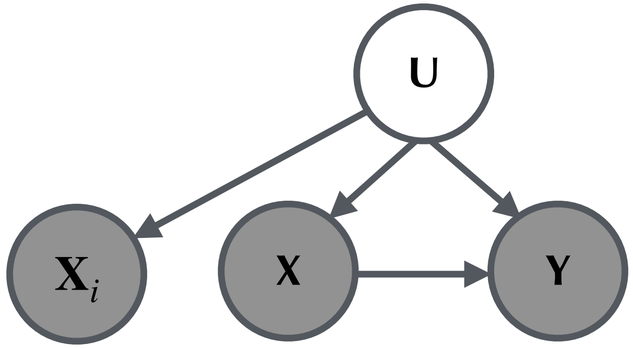
If U influences both $\mathbf{X}$, Y and if either (i) $\mathbf{X} \rightarrow Y$ or (ii) $Y \rightarrow \mathbf{X}$

$$I(Y;\hat{Y}) = \overbrace{I(\phi(\mathbf{X});Y \mid E)}^{\text{Cond. Inform.}} - \overbrace{I(\phi(\mathbf{X});Y \mid \hat{Y})}^{\text{Residual}}$$

Reddy, Abbavaram Gowtham, et al. "Boosting for Predictive Sufficiency." *The Fourteenth International Conference on Learning Representations*.

Informative Covariates

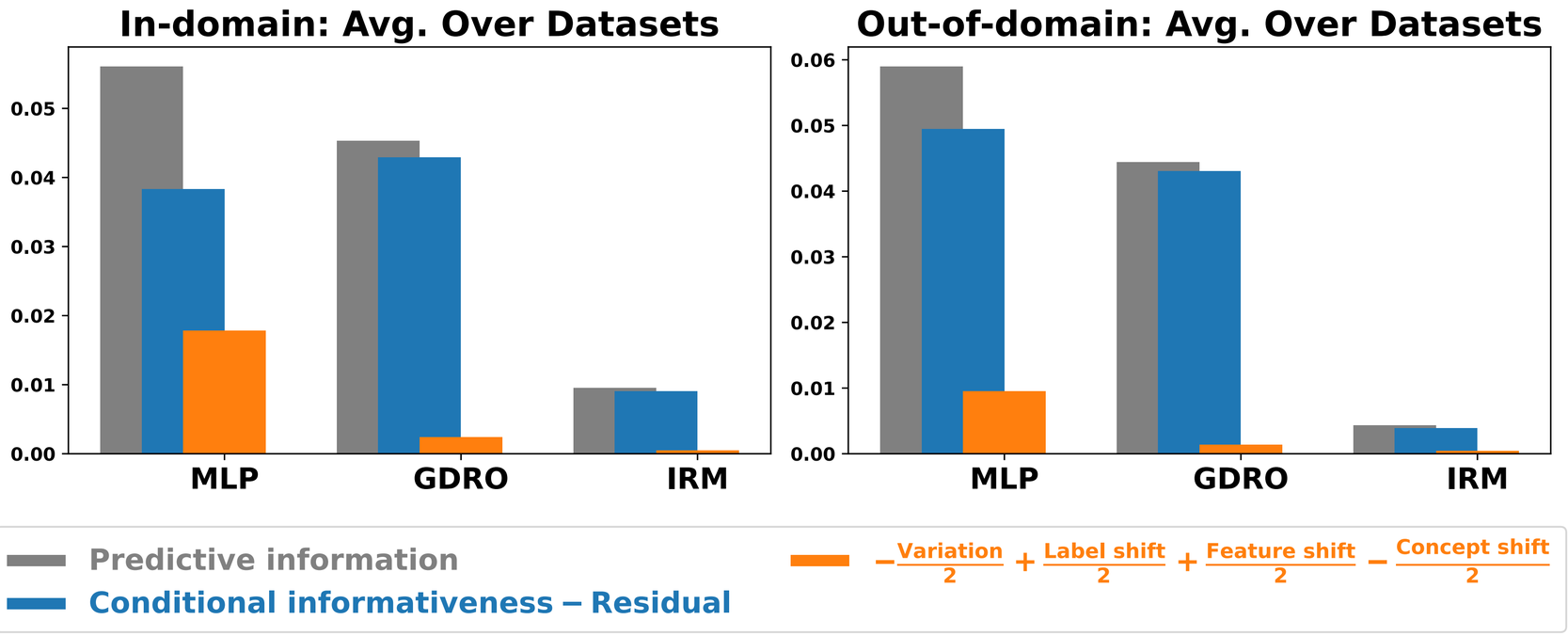
Definition: $\mathbf{X}_I$ that are neither causal nor anti-causal are informative to Y if $\mathbf{X}_I \not\perp Y \mid \mathbf{X}$



Proposition: Adding $\mathbf{X}_I$ to $\mathbf{X}$ results in the following

- $I(\phi_2(\{\mathbf{X} \cup \mathbf{X}_I\}); Y \mid E) > I(\phi_1(\mathbf{X}); Y \mid E)$ conditional informativeness increases
- $I(\phi_2(\{\mathbf{X} \cup \mathbf{X}_I\}); E) > I(\phi_1(\mathbf{X}); E)$ feature shift increases
- $I(Y; E \mid \phi_2(\{\mathbf{X} \cup \mathbf{X}_I\})) < I(Y; E \mid \phi_1(\mathbf{X}))$ concept shift decreases
- $I(\phi_2(\{\mathbf{X} \cup \mathbf{X}_I\}); E \mid Y) > I(\phi_1(\mathbf{X}); E \mid Y)$ variation increases

Experiments - Predictive Information



Method	ID Test Acc.	OOD Test Acc.
MLP	83.67	80.89
GDRO	82.48	79.64
IRM	67.65	65.48

Experiments - Sign Consistency

	Sign Consistency Metric (↑)					ID Test Accuracy (↑)			OOD Test Accuracy (↑)		
Method	**CI**	**Var**	**FS**	**CS**	**Res**	**C**	**AC**	**A**	**C**	**AC**	**A**
XGB	**0.89**	0.61	0.36	**0.89**	0.17	**81.24**	**84.84**	**85.22**	**69.92**	**76.72**	**77.00**
MLP	0.75	0.64	0.28	**0.89**	0.33	80.51	81.93	83.64	69.70	73.84	73.29
GDRO	0.69	0.67	0.19	**0.94**	0.36	80.92	83.76	78.40	69.61	73.49	72.28
IRM	0.78	0.61	0.25	**0.92**	0.25	67.91	68.22	71.17	61.21	62.96	64.89
VREX	0.64	0.61	0.25	**0.89**	0.44	54.39	62.32	59.74	56.62	63.53	63.06

Causal (C) $\subseteq$ Arguably Causal (AC) $\subseteq$ All (A)[1]

[1]Nastl, Vivian, and Moritz Hardt. "Do causal predictors generalize better to new domains?." *NeurIPS* 2024.