# Transportable Representations for Domain Generalization

Kasra Jalaldoust

Elias Bareinboim

Columbia University
Computer Science

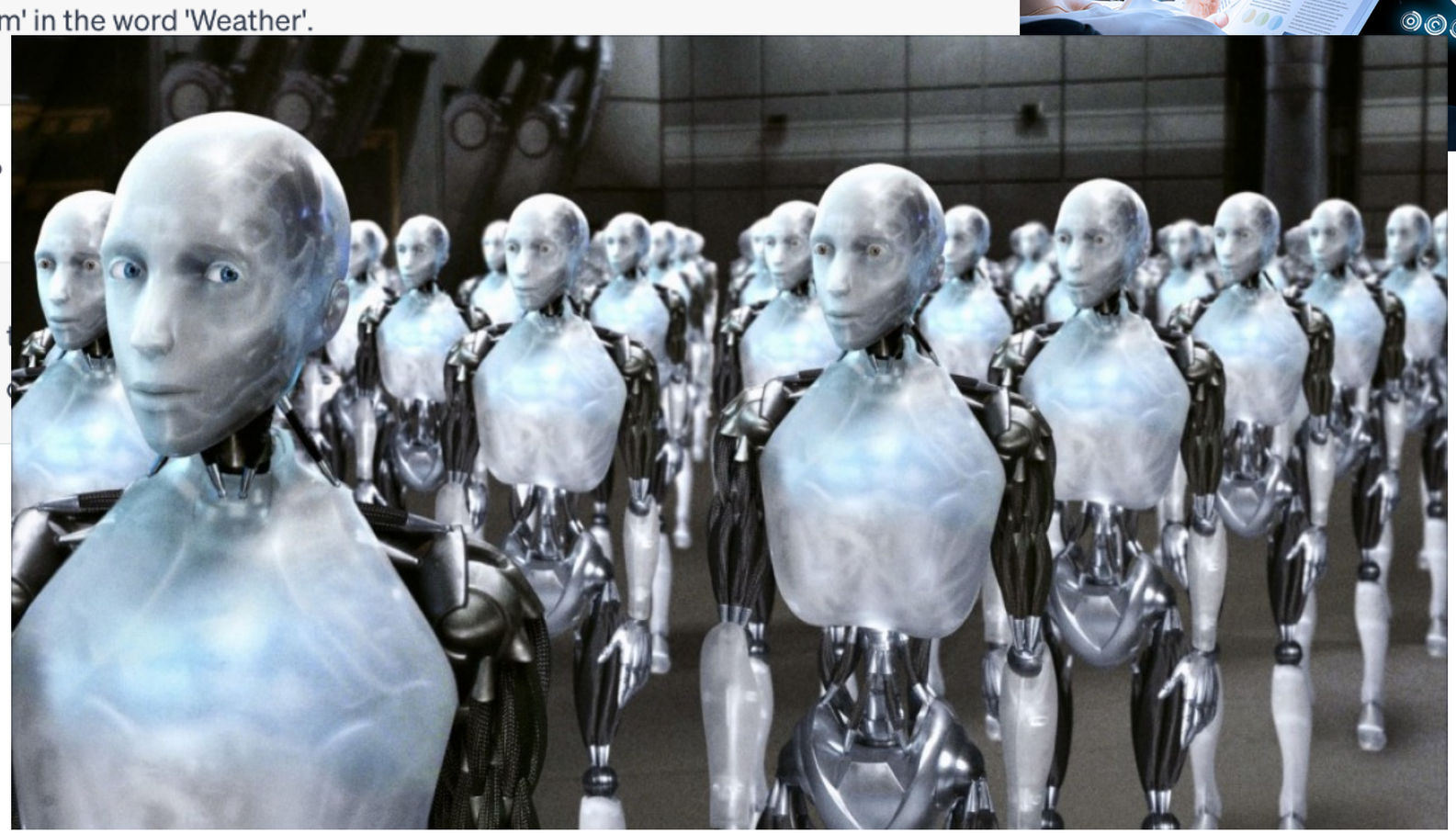
# Generalization Challenges


How many 'm's are in the word 'Weather'?

There is one 'm' in the word 'Weather'.

Are you sure?

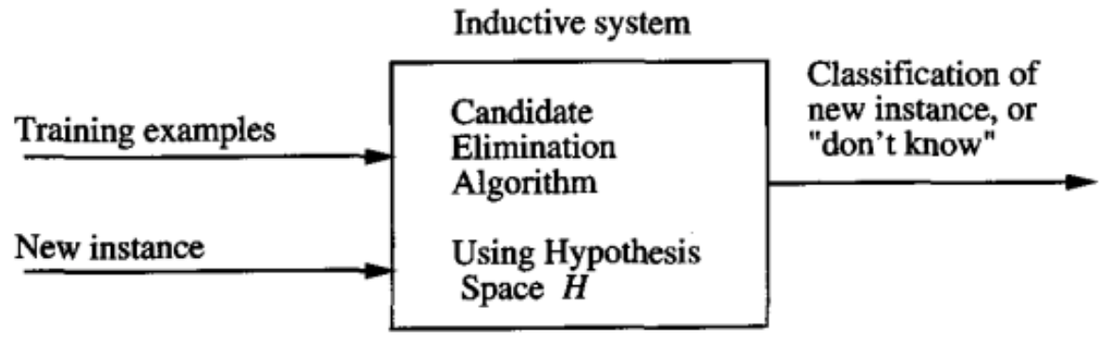
Apologies for
for pointing it



# Why Causality for Generalization?

- Tom Mitchell, one of the pioneers of the field, noted in his celebrated book:

  **"A learner that makes no a priori assumptions regarding the identity of the target concept has no rational basis for classifying any unseen instances."**




- Hume, the Scottish philosopher, argues that the ***inductive biases*** should be modeled explicitly, as it arises from the fundamental underspecification of the inductive problem.

- Immanuel Kant argued that certain concepts such as ***causality*** are prerequisite for making sense of experience.

- Causality is one of main's pillars of human's intelligence, and being agnostic to it can prolong development of **safe and reliable AI**, or even bring about catastrophic harm to society.

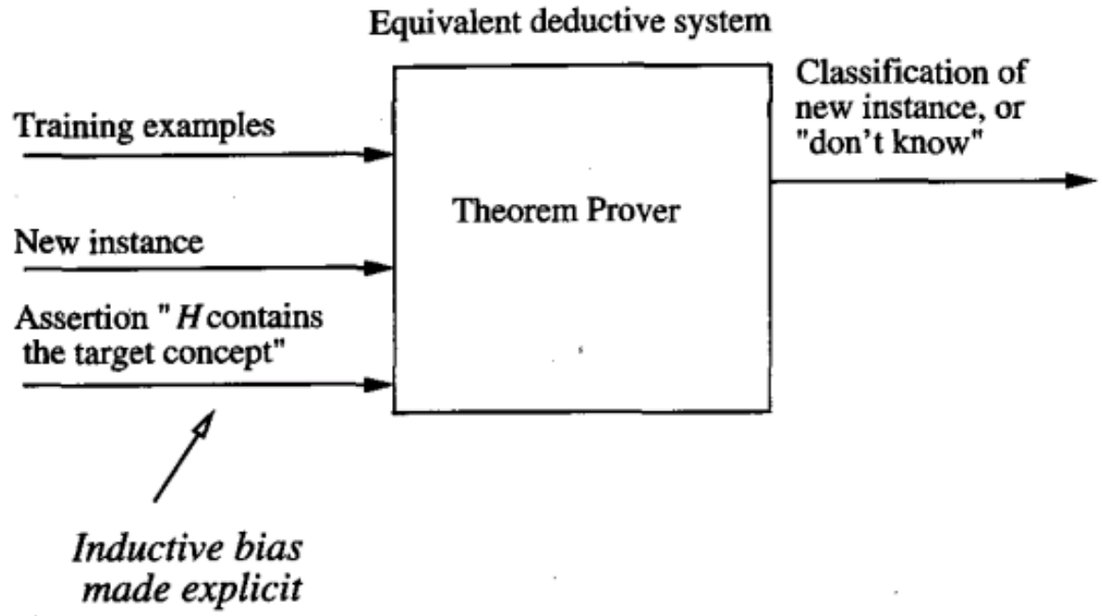



**FIGURE 2.8**

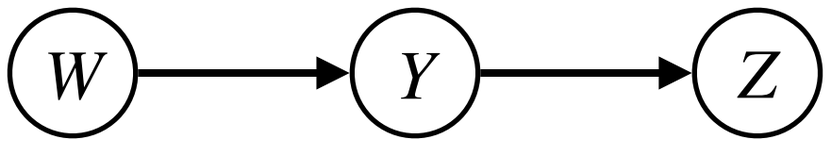
# Outline

- Basics about generalization in AI
    - Causal generalizability scheme
- Statistical Transportability
    - Direct transportability
    - Score-TR Algorithm
    - Transportable representations
- Causal Mechanistic Stability Assumption
- invariance learning and the generalization literature.

# A tale of two domains

**Domain** $\Pi$

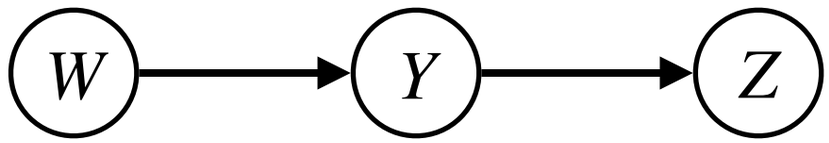
**SCM** $\mathscr{M}$




**Obs. dist.** $P(w, y, z)$

**Domain** $\Pi^*$

**SCM** $\mathscr{M}^*$




**Obs. dist.** $P^*(w, y, z)$

**Input.** Data from $P(w, y, z)$

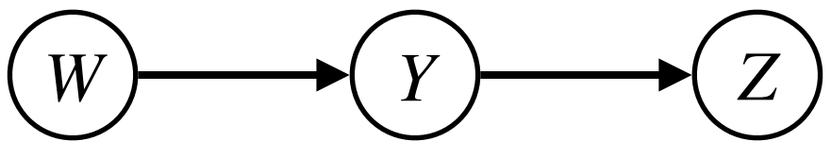
**Task.** Predict $Y$ based on $Z, W$ under $P^*$

**Query of interest** $P^*(y \mid w, z)$

# A tale of two domains

**Domain $\Pi$**

**SCM $\mathscr{M}$**




$$U_W, U_Y, U_Z \sim P(u_W, u_Y, u_Z)$$

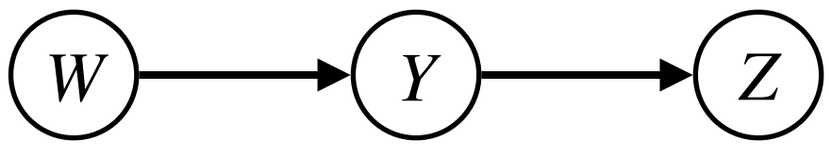
$$W \leftarrow f_W(U_W)$$

$$Y \leftarrow f_Y(X_1, U_Y)$$

$$Z \leftarrow f_Z(Y, U_Z)$$

**Domain $\Pi^*$**

**SCM $\mathscr{M}^*$**




$$U_W, U_Y, U_Z \sim P^*(u_W, u_Y, u_Z)$$

$$W \leftarrow f^*_W(U_W)$$

$$Y \leftarrow f^*_Y(X_1, U_Y)$$

$$Z \leftarrow f^*_Z(Y, U_Z)$$

# A tale of two domains

**Domain $\Pi$**

**SCM $\mathscr{M}$**

$\mathbf{U} \sim \text{Bern}(0.2)$

$W \leftarrow U_W$

$Y \leftarrow W \oplus U_Y$

$Z \leftarrow Y \oplus U_Z$

**Domain $\Pi^*$**

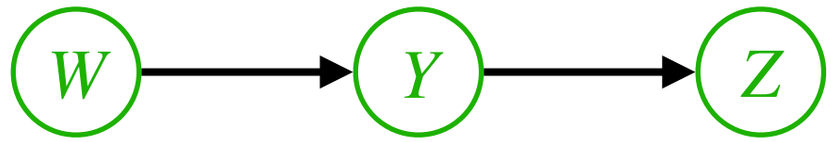
**SCM $\mathscr{M}_a^*$**

$\mathbf{U} \sim \text{Bern}(0.2)$

$W \leftarrow U_W$

$Y \leftarrow W \oplus U_Y$

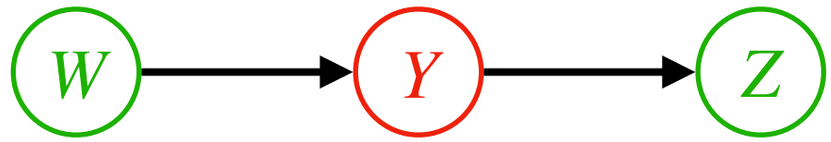
$Z \leftarrow Y \oplus U_Z$




$$P(z, y, w) = P^*(z, y, w) \Rightarrow P^*(y \mid w, z) = P(y \mid w, z)$$

# A tale of two domains

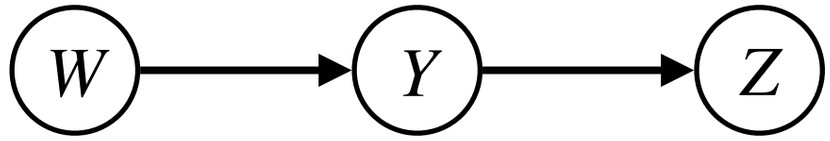
| Domain $\Pi$ | Domain $\Pi^*$ |
|---|---|
| SCM $\mathscr{M}$ | SCM $\mathscr{M}_b^*$ |
| $\mathbf{U} \sim \text{Bern}(0.2)$ | $\mathbf{U} \sim \text{Bern}(0.2)$ |
| $W \leftarrow U_W$ | $W \leftarrow U_W$ |
| $Y \leftarrow W \oplus U_Y$ | $Y \leftarrow \neg(W \oplus U_Y)$ |
| $Z \leftarrow Y \oplus U_Z$ | $Z \leftarrow Y \oplus U_Z$ |




$$P(z, y, w) = P^*(z, \neg y, w) \Rightarrow P^*(y \mid w, z) = P(\neg y \mid w, z)$$

# A tale of two domains

**Assumption.** The causal diagram




**Domain** $\Pi$

**SCM** $\mathscr{M}$

**Estimable** $P(y, w, z)$

Too pathological if $f_Y$ is different between $\Pi, \Pi^*$...

**Unseen Domain** $\Pi^*$

**SCM** $\mathscr{M}^*_a$ **SCM** $\mathscr{M}^*_b$ **...**

$$P^*(y \mid w, z) = \begin{cases} 0.95 \\ P(y \mid w, z) \\ 1 - P(y \mid w, z) \\ 0.05 \\ \dots \end{cases}$$

# A tale of two domains (v2)

**Domain $\Pi$**

**SCM $\mathscr{M}$**

$\mathbf{U} \sim \text{Bern}(0.2)$

$W \leftarrow U_W$

$Y \leftarrow W \oplus U_Y$

$Z \leftarrow Y \oplus U_Z$

**Domain $\Pi^*$**

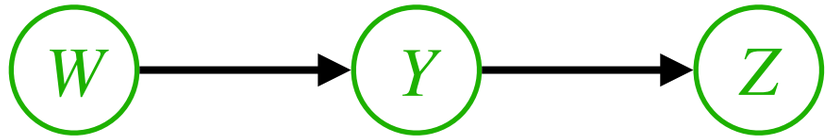
**SCM $\mathscr{M}_a^*$**

$\mathbf{U} \sim \text{Bern}(0.2)$

$W \leftarrow U_W$

$Y \leftarrow W \oplus U_Y$

$Z \leftarrow Y \oplus U_Z$

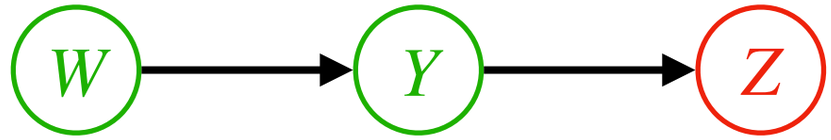



**Assumption.** the mechanism of $Y$ remains "Invariant" between domains $\Pi, \Pi^*$

$$P(z, y, w) = P^*(z, y, w) \Rightarrow P^*(y \mid w, z) = P(y \mid w, z)$$

# A tale of two domains (v2)

| **Domain $\Pi$** | **Domain $\Pi^*$** |
|---|---|
| **SCM $\mathscr{M}$** | **SCM $\mathscr{M}_c^*$** |
| $\mathbf{U} \sim \text{Bern}(0.2)$ | $\mathbf{U} \sim \text{Bern}(0.2)$ |
| $W \leftarrow U_W$ | $W \leftarrow U_W$ |
| $Y \leftarrow W \oplus U_Y$ | $Y \leftarrow W \oplus U_Y$ |
| $Z \leftarrow Y \oplus U_Z$ | $Z \leftarrow \neg(Y \oplus U_Z)$ |

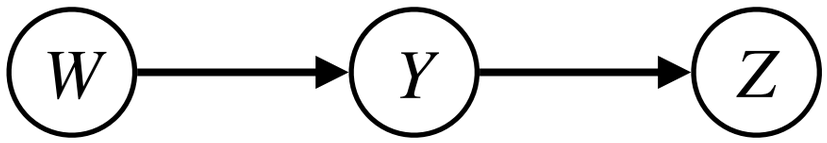



**Assumption.** the mechanism of $Y$ remains "Invariant" between domains $\Pi, \Pi^*$

$$P(z, y, w) = P^*(z, \neg y, w) \rightarrow P^*(y \mid w, z) = P(y \mid w, \neg z)$$

# A tale of two domains (v2)

**Assumption.** The causal diagram + the mechanism of $Y$ remains "Invariant" between domains $\Pi, \Pi^*$




**Domain** $\Pi$

**SCM** $\mathcal{M}$

**Estimable** $P(y, w, z)$

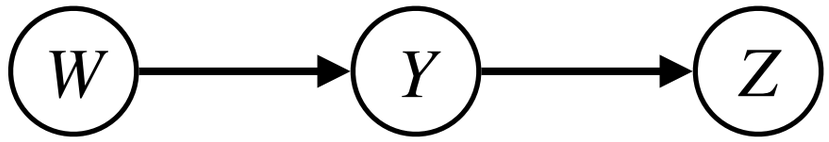
**Unseen Domain** $\Pi^*$

**SCM** $\mathcal{M}_a^*$ **SCM** $\mathcal{M}_c^*$ **...**

$$P^*(y \mid w, z) = \begin{cases} P(y \mid w, z) & 0.95 \\ P(y \mid w, \neg z) & 0.5 \\ \dots \end{cases}$$

# A tale of two domains (v3)

Objective is to predict $Y$ based on $Z, W$.
What if we just use $W$ and drop $Z$?




$$P^*(y \mid w) = P^*(y \mid do(w))$$

**rule 2:** $Y \perp W$ **in** $G_{\underline{W}}$

$$= \sum_{u_Y} 1_{\{f_Y^*(w,u_Y)=y\}} \cdot P^*(u_Y)$$

**prob. rules**

$$= \sum_{u_Y} 1_{\{f_Y(w,u_Y)=y\}} \cdot P(u_Y) \qquad f_Y^* = f_Y, P^*(u_Y) = P(u_Y)$$

$$= P(y \mid do(w))$$

**prob. rules**

$$= P(y \mid w)$$

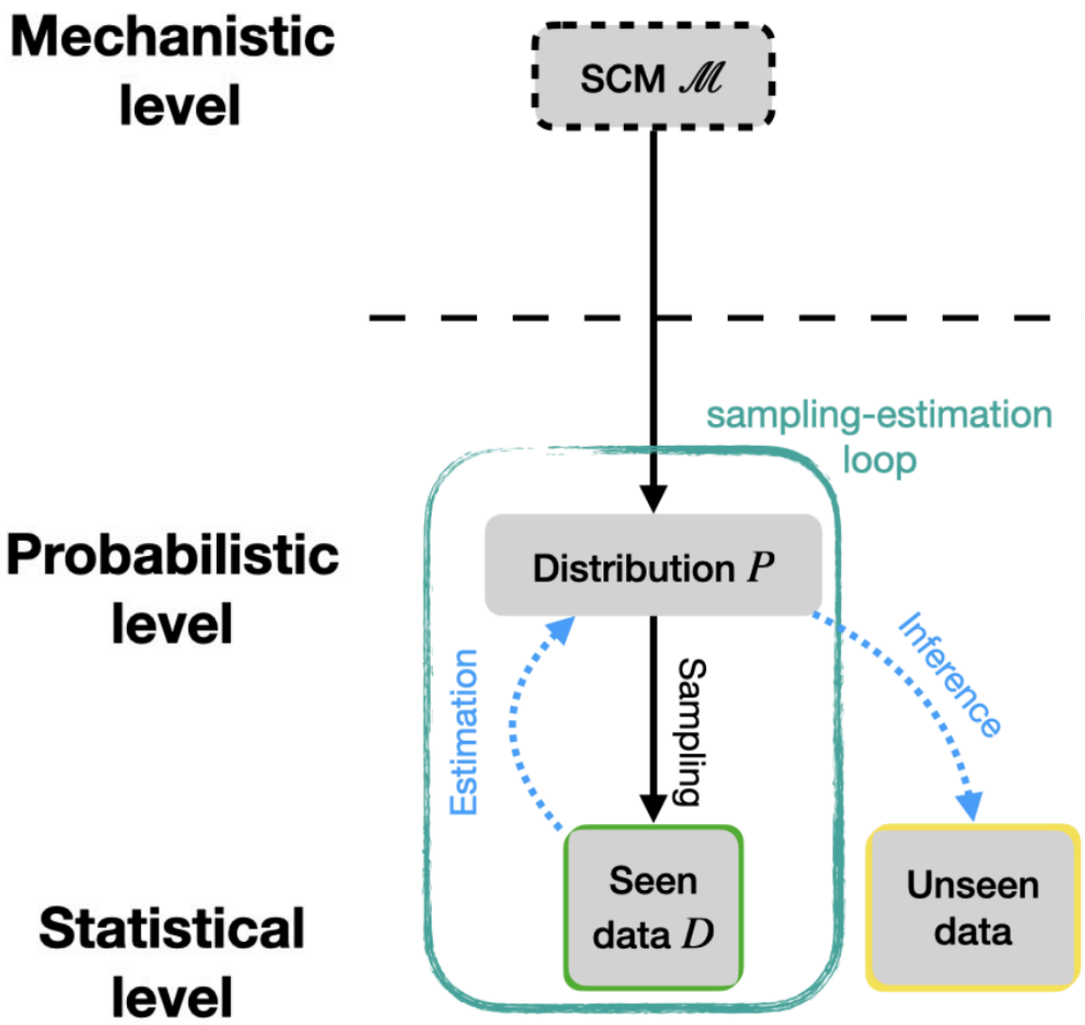
**rule 2:** $Y \perp W$ **in** $G_{\underline{W}}$

# Observations

- V1. Only access to causal diagrams of source and target does not give us an attack to the DG task.
- V2. Not all queries can be uniquely computed even under *reasonable* structural invariances.
- V3. DG task can be address through assessment of transportability of various different queries.

- Question. What is the bigger picture?

# Generalizability Schema




(a) Sampling-estimation loop (in-distribution learning)

# Cross-population Learning Tasks (A structural taxonomy)

- Domain generalization (DG)

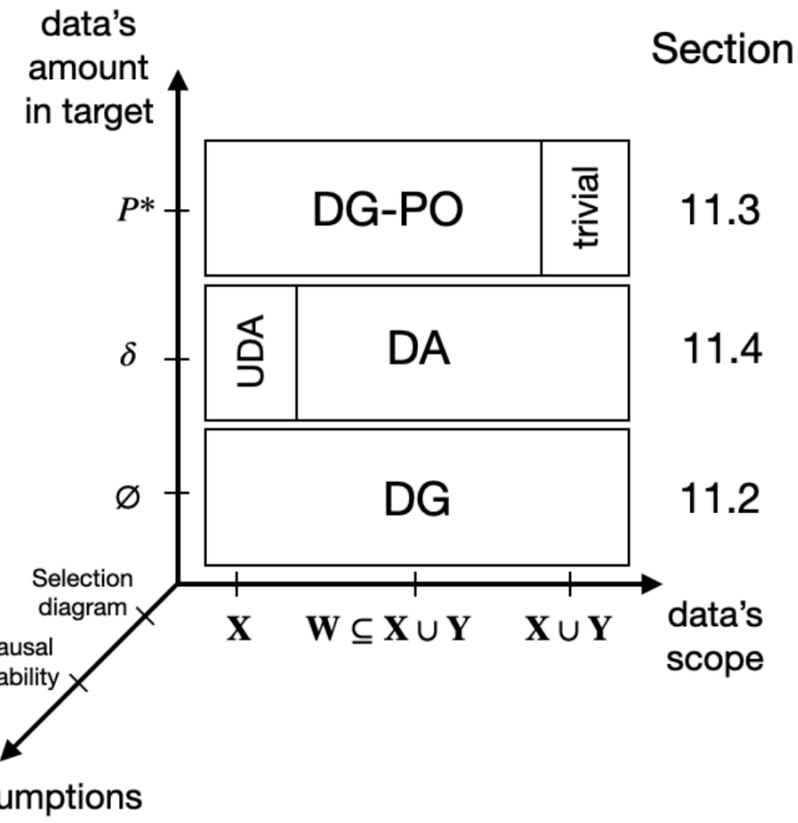
$$P(\mathbf{x}, y)$$

- DG with partially observed target domain

$$P(\mathbf{x}, y) + P^*(\mathbf{W})$$

- Domain adaptation

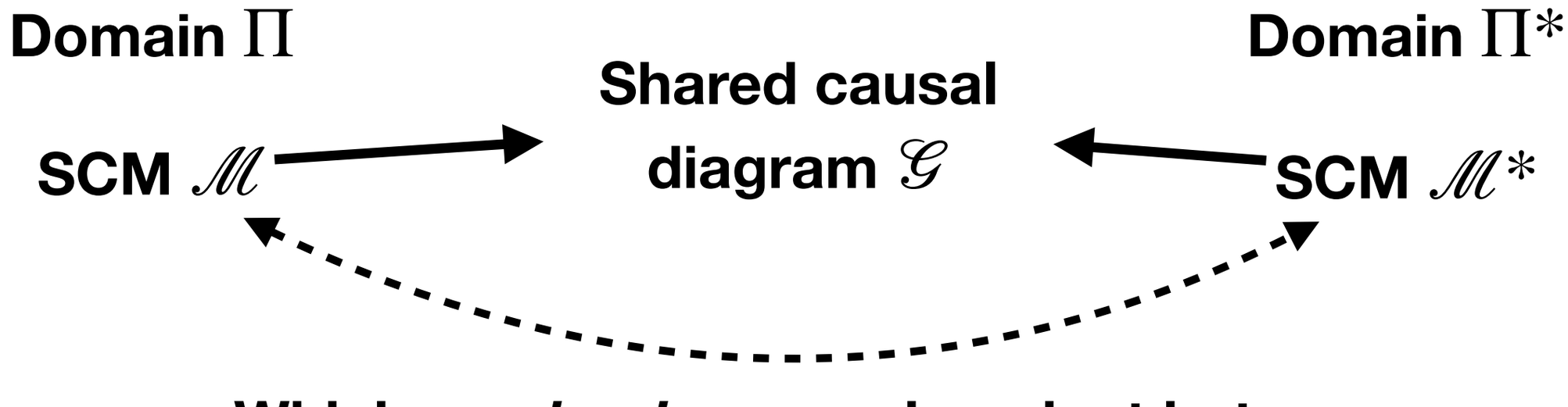
$$P(\mathbf{x}, y) + \text{data}^*$$




In this lecture, our metric for *learning* is the classification risk in the target domain, denoted as:

$$\mathscr{R}_{P^*}(h) := \mathbb{E}_{P^*}[1_{\{Y=h(\mathbf{X})\}}] = P^*(Y \neq h(\mathbf{X}))$$

# A Formal Approach



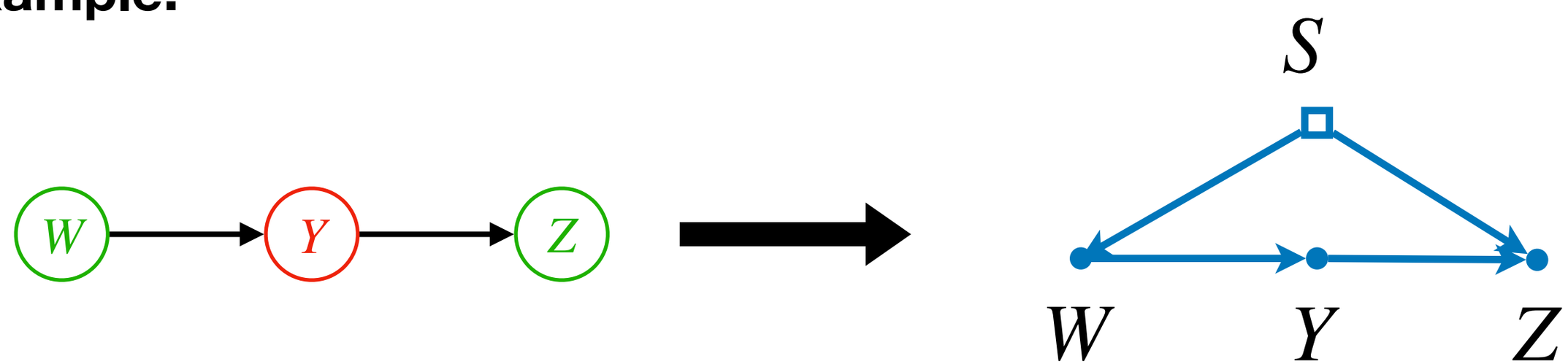

**Which *mechanisms* are invariant between the source and target domains?**

**Definition 10.1.1. (Domain discrepancy sets)** Let $\Pi^a, \Pi^b$ be domains with SCMs $\mathscr{M}^a, \mathscr{M}^b$. The subset of variables $\Delta_{a,b} \subset \mathbf{V}$ is called a domain discrepancy set, and indicates that there ***might*** be a mechanism discrepancy for every $V \in \Delta_{a,b}$, that is possibility of $f_V^a \neq f_V^a$ and/or $P^a(u_V) \neq P^b(u_V)$.

# Selection Diagrams

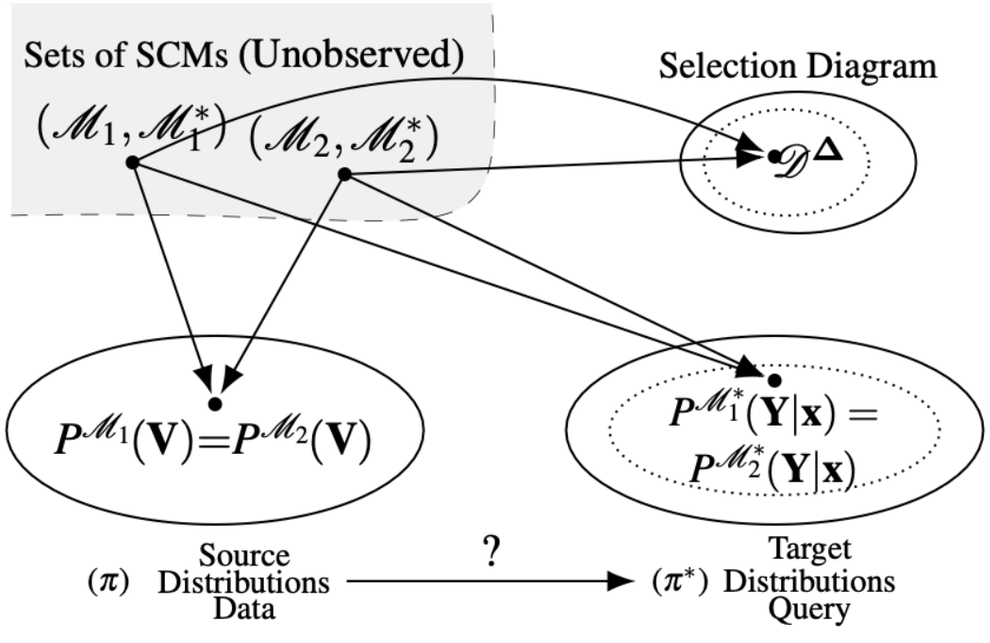
**Definition 10.1.2. (Selection diagram)** Given a causal diagram $G$ and a domain discrepancy set $\Delta_{a,b}$, we define the selection diagram $\mathcal{G}^{\Delta_{a,b}}$ by augmenting $\mathcal{G}$ with a new node $S_{a,b}$ (denoted with a square for distinction with other variables), and adding edges $S_{a,b} \rightarrow V$ for every $V \in \Delta_{a,b}$.

**Example.**

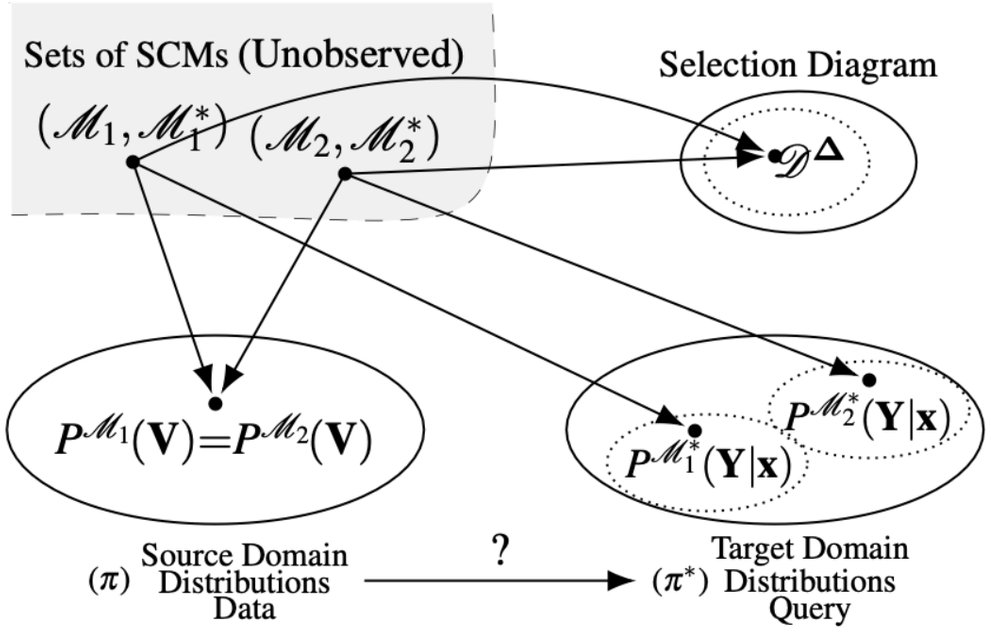


# Direct Statistical Transportability

**Definition 11.2.1. (Direct statistical transportability)** Consider a selection diagram $\mathcal{G}^{\Delta}$ encoding the commonalities and disparities across two domains represented by SCMs $\mathcal{M}$ and $\mathcal{M}^*$, a target variable $Y$, and a set of variables $\mathbf{W}$. The conditional distribution $P^*(y \mid \mathbf{w})$ is said *directly transportable* if every pair of models compatible with $\mathcal{G}^{\Delta}$, it holds that $P^*(y \mid \mathbf{w}) = P(y \mid \mathbf{w})$.




(a) Transportable




(b) Non-transportable

# S-admissibility

**Definition 11.2.2. (S-admissibility)** Consider the SCMs $\mathscr{M}^i$, $\mathscr{M}^j$ and the sets of variables $\mathbf{Z}$, $\mathbf{A}$. The set $\mathbf{A}$ is said to be S-admissible conditioned on $\mathbf{Z}$ w.r.t. the domains $\mathscr{M}^i$, $\mathscr{M}^j$ whenever $\mathbf{A}$ is d-separated from $S_{i,j}$ given $\mathbf{Z}$ in $\mathcal{G}^{\Delta_{i,j}}$. The conditional distribution of $\mathbf{A}$ given $\mathbf{Z}$ is invariant across the two domains.
More formally:

$$S_{i,j} \perp \mathbf{A} \mid \mathbf{Z} \text{ in } \mathcal{G}^{\Delta_{i,j}} \implies P^i(\mathbf{a} \mid \mathbf{z}) = P^j(\mathbf{a} \mid \mathbf{z})$$

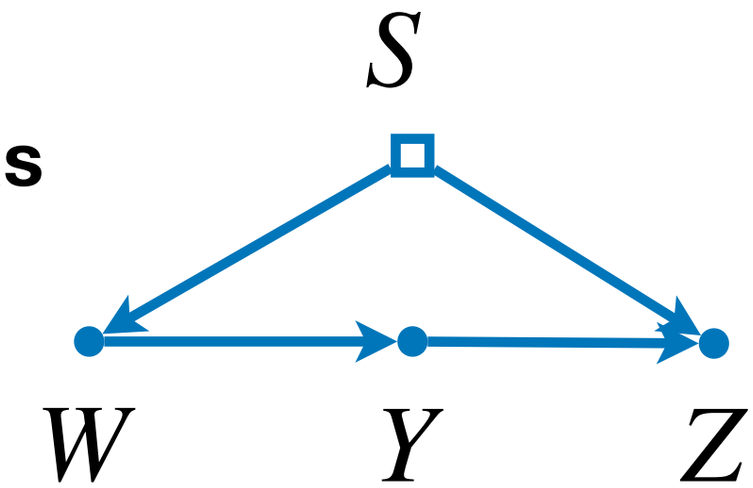
**Theorem 11.2.1. (Direct statistical-TR)** Consider the SCMs $\mathscr{M}$, $\mathscr{M}^*$ over $\mathbf{X} \cup \{Y\}$. For a subset $\mathbf{Z} \subseteq \mathbf{X}$, the query $P^*(y \mid \mathbf{z})$ is direct-transportable **if and only if** $Y$ is S-admissible conditioned on $\mathbf{Z}$ w.r.t. $\mathscr{M}$, $\mathscr{M}^*$.

# Direct-transportability

**Corollary 11.2.2. (Direct non-TR)** If there exists an open path between $S$ and $Y$ conditional on $\mathbf{Z}$, then $P^*(y \mid \mathbf{z})$ can not be uniquely computed from $P(\mathbf{v})$.

**Example.**

$P^*(y \mid w, z)$ **is direct non-transportable, as shown earlier.**




**Idea.** Let's connect transportability with generalization in ML!

# Covariate Shift

**Definition 11.2.3. (Structural Covariate Shift Assumption)** Consider SCMs $\mathscr{M}^1, \mathscr{M}^2, \ldots, \mathscr{M}^*$ over the variables $\mathbf{X} \cup \{Y\}$. SCS holds If:

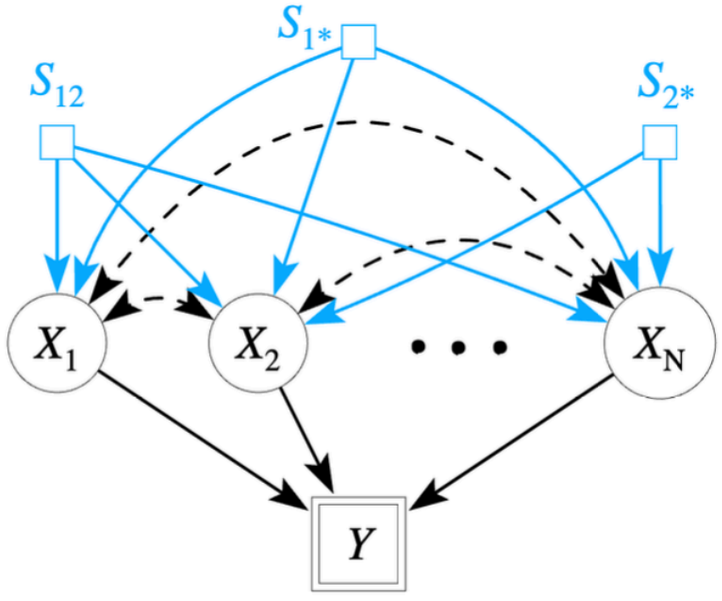
1. All variables in $\mathbf{X}$ are direct cause (i.e., parent) to $Y$.

2. $Y \notin \Delta$

3. $Y$ is not confounded with any other variable, i.e., $\mathbf{U}_Y \cap \mathbf{U}_{\mathbf{X}} = \varnothing$.

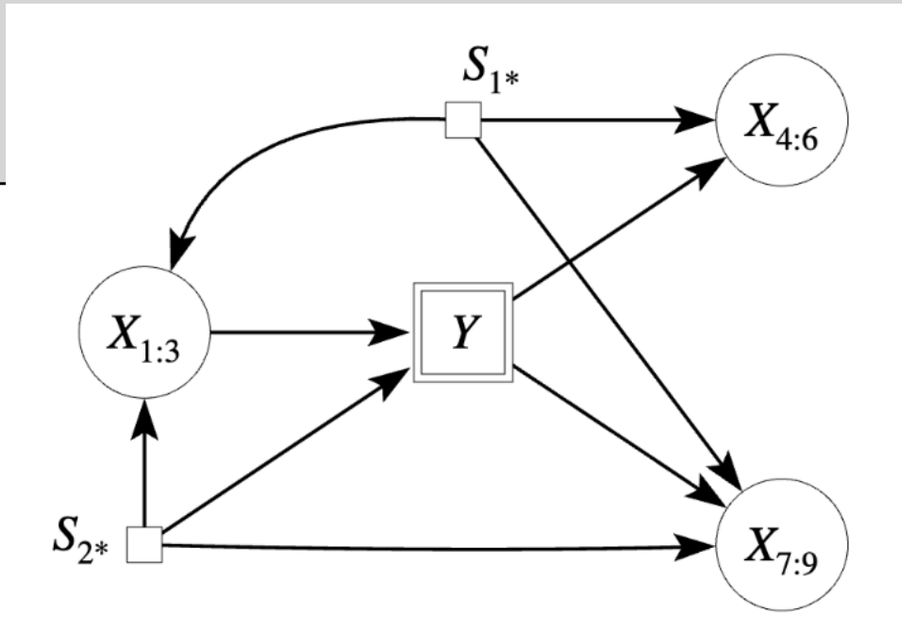
## Example 11.6.

Let $h_{OBC}(\mathbf{x}) = 1_{\{P(Y=1|\mathbf{x})>\frac{1}{2}\}}$ be the optimal Bayes classifier in domain $\mathscr{M}^i$. This classifier can be approximated from large source data, and can be safely deployed to the target domain (why?)



# Beyond SCS

**Definition 11.2.4. (Transportable feature sets)** A subset of variables $\mathbf{Z} \subseteq \mathbf{X}$ is called a feature set, and its score function is defined as $l_{\mathbf{Z}}(\mathbf{z}) = \mathbb{E}[Y \mid \mathbf{z}]$. A feature set is said to be transportable from a collection of distributions $\mathbb{P} = \langle P^1, P^2, \ldots, P^K \rangle$ given $\mathcal{G}^{\Delta}$, if for every tuple of source and target SCMs $\langle \mathcal{M}_a^1, \mathcal{M}_a^2, \ldots, \mathcal{M}_a^K, \mathcal{M}_a^* \rangle$ and $\langle \mathcal{M}_b^1, \mathcal{M}_b^2, \ldots, \mathcal{M}_b^K, \mathcal{M}_b^* \rangle$ that induce $\mathcal{G}^{\Delta}$ and entail $\mathbb{P}$, it holds that $\mathbb{E}_{P\mathcal{M}_a^*}[Y \mid \mathbf{z}] = \mathbb{E}_{P\mathcal{M}_b^*}[Y \mid \mathbf{z}]$




**Example 11.8.**

$$\mathbb{E}_{P*}[Y \mid \mathbf{x}_{1:9}] = !!$$

$$\mathbb{E}_{P*}[Y \mid \mathbf{x}_{1:3}] = \mathbb{E}_{P^{1,2}}[Y \mid \mathbf{x}_{1:3}]$$

$$\mathbb{E}_{P*}[Y \mid \mathbf{x}_{1:6}] \propto \mathbb{E}_{P^{1,2}}[Y \mid \mathbf{x}_{1:3}] \cdot P^1(\mathbf{x}_{4:6} \mid y)$$

# Score-TR algorithm

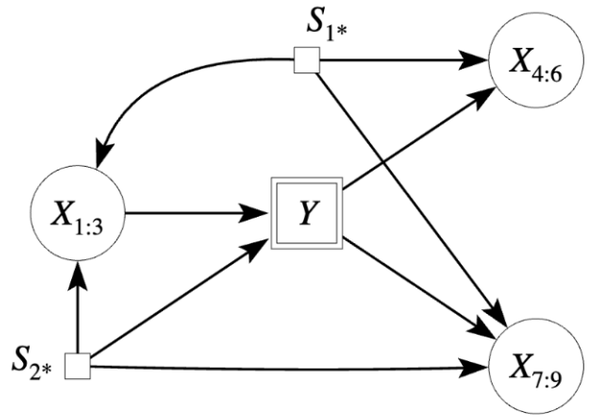
**Theorem 11.2.4.** A feature set $\mathbf{Z} \subseteq \mathbf{X}$ is transportable from source distributions if and only if the Score-TR algorithm returns a transportability formula for it.

**Algorithm 31** Score-TR$(\mathbf{Z}, \mathcal{G}, \Delta)$

$\tilde{\mathbf{Z}} \leftarrow \emptyset$
**while** $\exists \quad Z_0 \in \mathbf{Z}$ such that $Y \perp\!\!\!\perp Z_0 \mid \mathbf{Z} \setminus \{Z_0\}$ in $\mathcal{G}_{\mathbf{X}\setminus\tilde{\mathbf{Z}}\cup\{Y\},\underline{Z_0}}$ **do**
    $\mathbf{Z} \leftarrow \mathbf{Z} \setminus \{Z_0\}$
    $\tilde{\mathbf{Z}} \leftarrow \tilde{\mathbf{Z}} \cup \{Z_0\}$
**end while**
$\mathbf{W} \leftarrow An\big(\mathbf{Z} \setminus \tilde{\mathbf{Z}} \cup \{Y\}\big)_{\mathcal{G}_{\mathbf{X}\setminus\tilde{\mathbf{Z}}\cup\{Y\}}}$.
Let $\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_T$ be the C-components of $\mathbf{W}$ in $\mathcal{G}_{\mathbf{X}\setminus\tilde{\mathbf{Z}}\cup\{Y\}}$.
prod $\leftarrow 1$
**for** $t$ from 1 to $T$ **do**
    **if** $\exists$ domain $\mathcal{M}^i$ such that $\mathbf{C}_t \cap \Delta_{i,*} = \emptyset$ **then**
        $\text{prod} \leftarrow \text{prod} \cdot \frac{\prod_{V\in\mathbf{X}\cup\{Y\}} P^i(V|\mathbf{Pa}_V)}{\prod_{V\in\mathbf{X}\cup\{Y\}\setminus\mathbf{C}_t} P^i(V|\mathbf{Pa}_V)}$
    **else**
        **return** NOT TRANSPORTABLE
    **end if**
**end for**
**return** $\frac{\sum_{\mathbf{w}\setminus(\{y\}\cup\mathbf{z}\setminus\tilde{\mathbf{z}})} \text{prod}}{\sum_{\mathbf{w}\setminus\{y\}} \text{prod}}$

# Score-TR algorithm

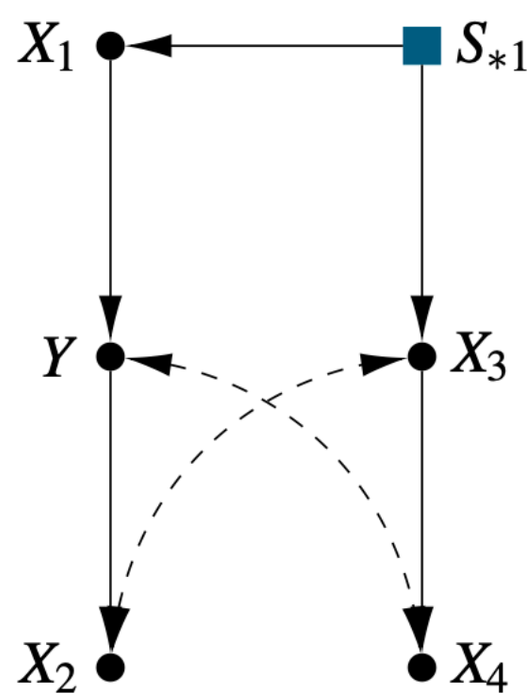
**Example 11.9.**




$$
\begin{aligned}
\mathbb{E}_{P^*}[Y \mid \mathbf{X}_{1:6}] &= P^*(Y = 1 \mid \mathbf{X}_{1:3}, \mathbf{X}_{4:6}) && \text{(definition)} && (11.44)\\
&= \frac{P^*(Y = 1, \mathbf{X}_{4:6} \mid \mathbf{X}_{1:3})}{P^*(\mathbf{X}_{4:6} \mid \mathbf{X}_{1:3})} && \text{(conditioning)} && (11.45)\\
&= \frac{P^*(Y = 1 \mid \mathbf{X}_{1:3}) \cdot P^*(\mathbf{X}_{4:6} \mid Y = 1, \mathbf{X}_{1:3})}{\sum_{y=0}^{1} P^*(Y = y \mid \mathbf{X}_{1:3}) \cdot P^*(\mathbf{X}_{4:6} \mid Y = y, \mathbf{X}_{1:3})} && \text{(factorization)} && (11.46)\\
&= \frac{P^1(Y = 1 \mid \mathbf{X}_{1:3}) \cdot P^*(\mathbf{X}_{4:6} \mid Y = 1, \mathbf{X}_{1:3})}{\sum_{y=0}^{1} P^1(Y = y \mid \mathbf{X}_{1:3}) \cdot P^*(\mathbf{X}_{4:6} \mid Y = y, \mathbf{X}_{1:3})} && (S_{1*} \perp\!\!\!\perp Y \mid \mathbf{X}_{1:3} \text{ in } \mathcal{G}^{\Delta}_{\text{aug}}) && \\
& && && (11.47)\\
&= \frac{P^1(Y = 1 \mid \mathbf{X}_{1:3}) \cdot P^2(\mathbf{X}_{4:6} \mid Y = 1, \mathbf{X}_{1:3})}{\sum_{y=0}^{1} P^1(Y = y \mid \mathbf{X}_{1:3}) \cdot P^2(\mathbf{X}_{4:6} \mid Y = y, \mathbf{X}_{1:3})} && (S_{2*} \perp\!\!\!\perp \mathbf{X}_{4:6} \mid \mathbf{X}_{1:3}, Y \text{ in } \mathcal{G}^{\Delta}), &&
\end{aligned}
$$

# Maximal TR feature sets

**Example 11.10.**




$$\mathbb{E}_{P*}[Y \mid x_1] \stackrel{?}{=} \mathbb{E}_{P}[Y \mid x_1] \quad ✅$$

$$\mathbb{E}_{P*}[Y \mid x_1, x_2] \stackrel{?}{=} \mathbb{E}_{P}[Y \mid x_1, x_2] \quad ✅$$

$$\mathbb{E}_{P*}[Y \mid x_1, x_3, x_4] \stackrel{?}{=} \mathbb{E}_{P}[Y \mid x_1, x_3, x_4] \quad ✅$$

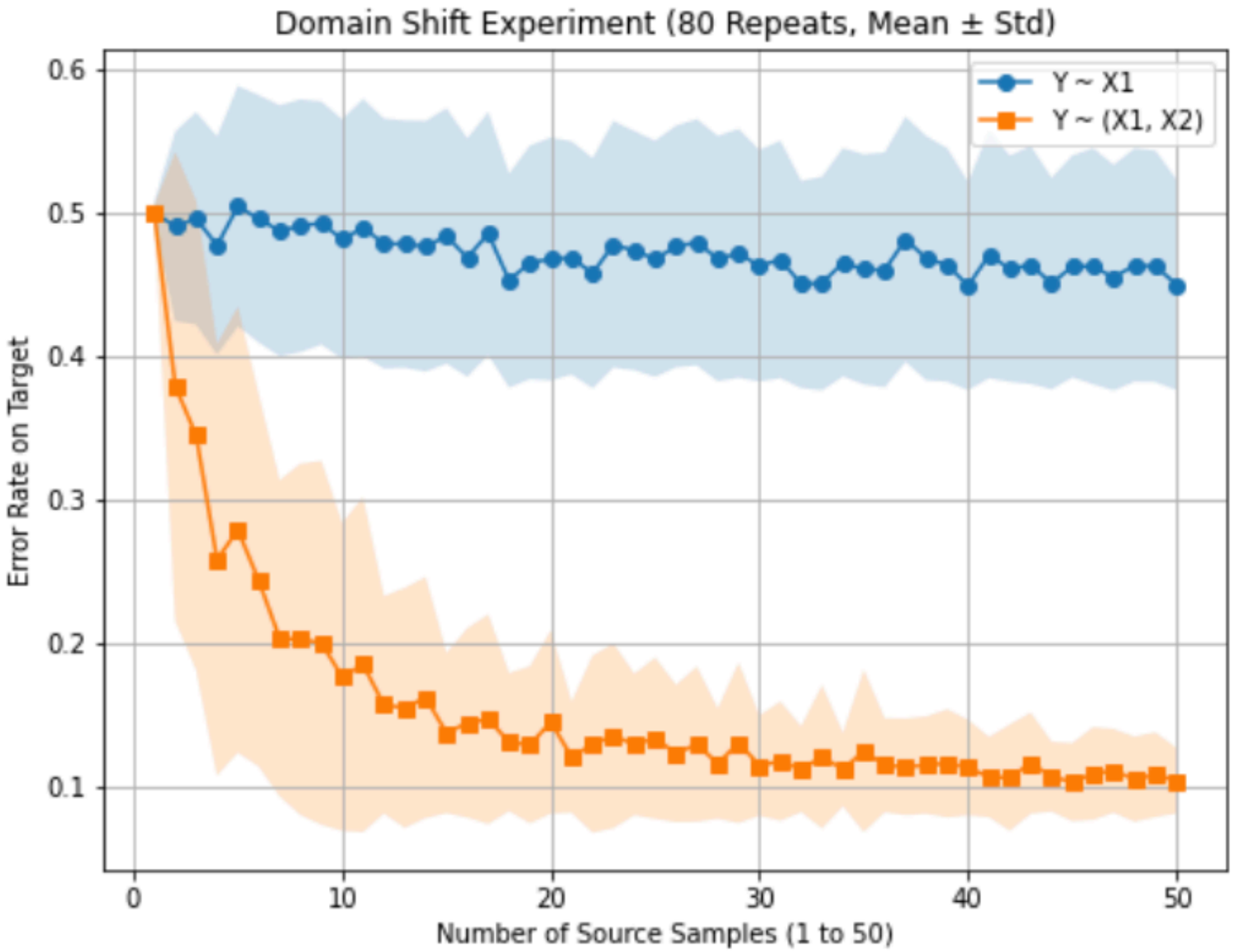
$$\mathbb{E}_{P*}[Y \mid x_1, x_2, x_3, x_4] \stackrel{?}{=} \mathbb{E}_{P}[Y \mid x_1, x_2, x_3, x_4] \quad ❌$$

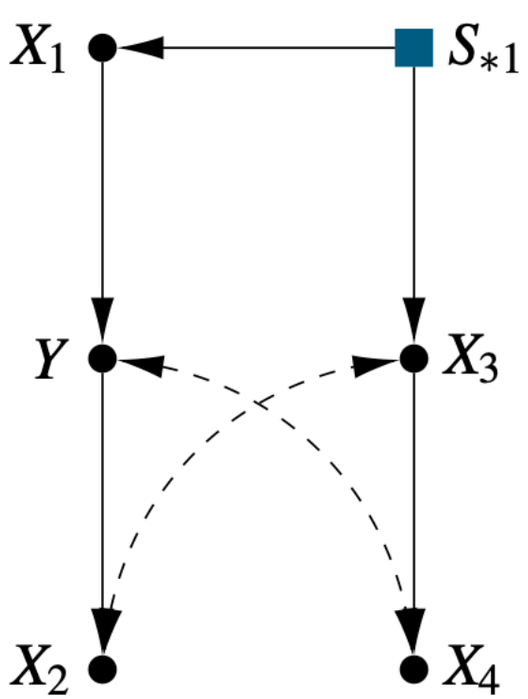
**(Causal)** **(TR)** **(Full)**

$$X_1 \quad \preceq \quad \begin{matrix} X_1, X_2 \\ X_1, X_3, X_4 \end{matrix} \quad \preceq \quad X_1, X_2, X_3, X_4$$

# Maximal TR feature sets

**Definition 11.2.5.** A transportable feature set (FS) $\mathbf{Z} \subseteq \mathbf{X}$ is *maximal* if it is not contained in any other transportable FS.




$X_1 \not\supseteq X_1, X_2$

$X_1 \not\supseteq X_1, X_3, X_4$

$X_1 \not\supseteq X_1, X_2, X_3, X_4$

# Many maximal TR FS!

**Example 11.12.**

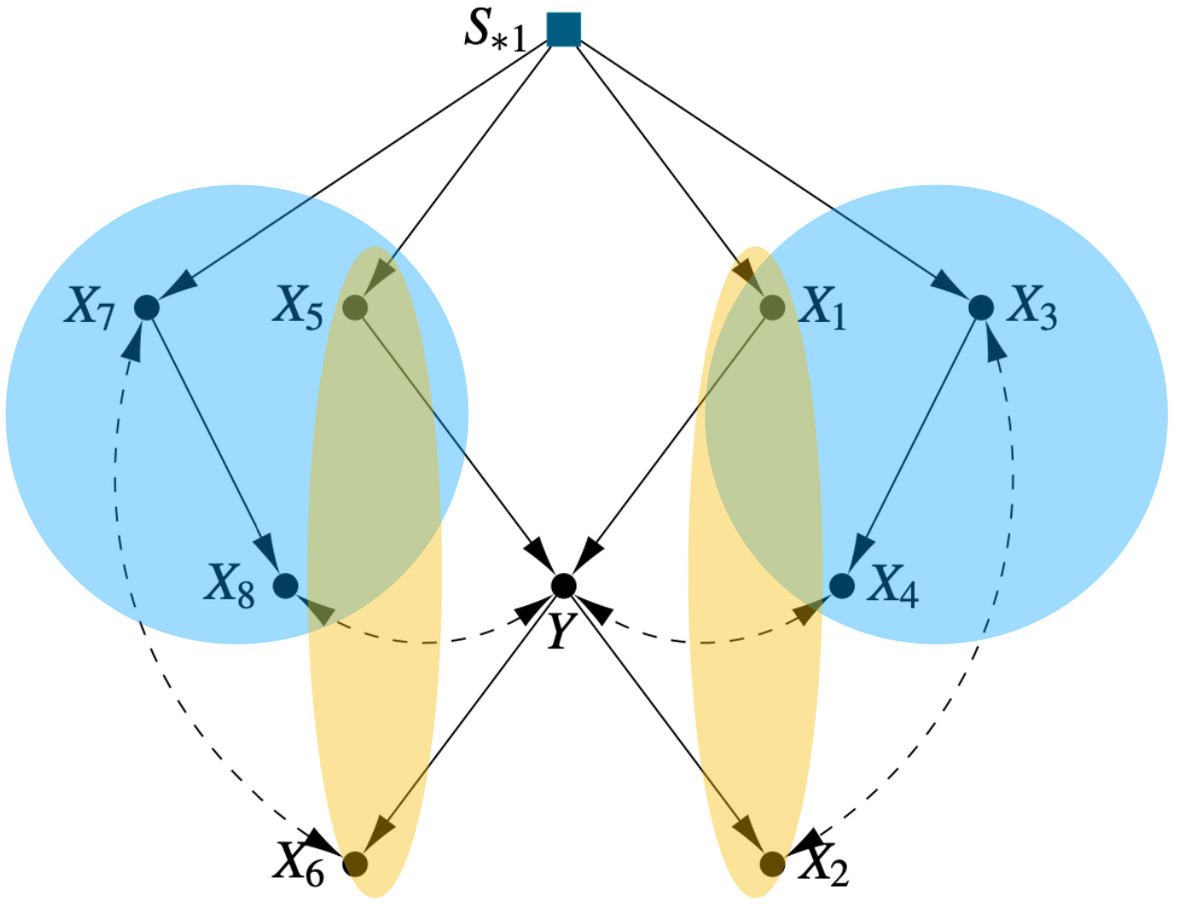
This graph is constructed by overlaying two instances of Ex. 11.10. In that case, we had two maximal TR FS, and turns out that every combination of those sets yields a maximal TR FS in this case,

$$\{X_1, X_2\} \cup \{X_5, x_6\}$$

$$\{X_1, X_3, X_4\} \cup \{X_5, X_6\}$$

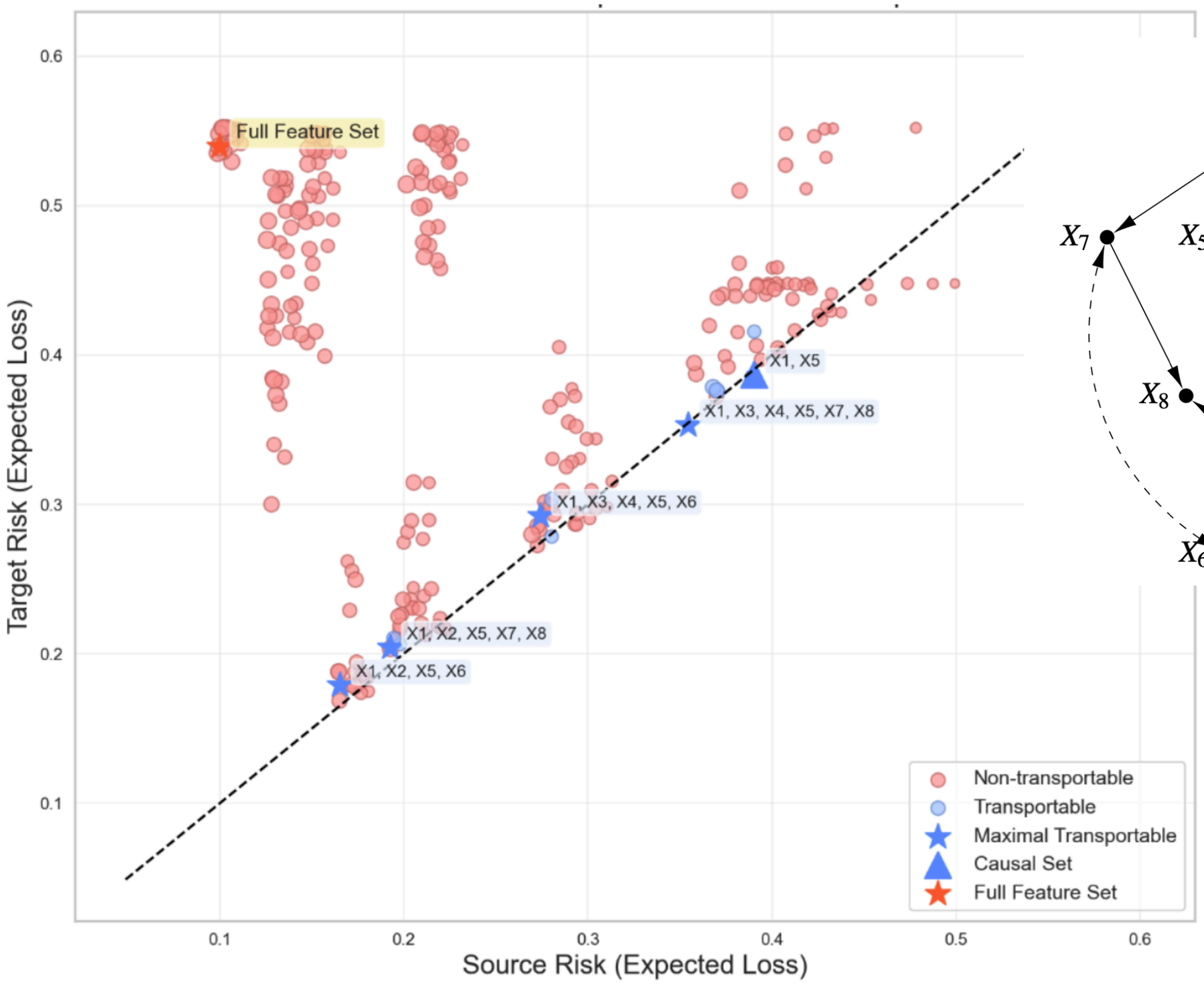
$$\{X_1, X_2\} \cup \{X_5, X_7, X_8\}$$

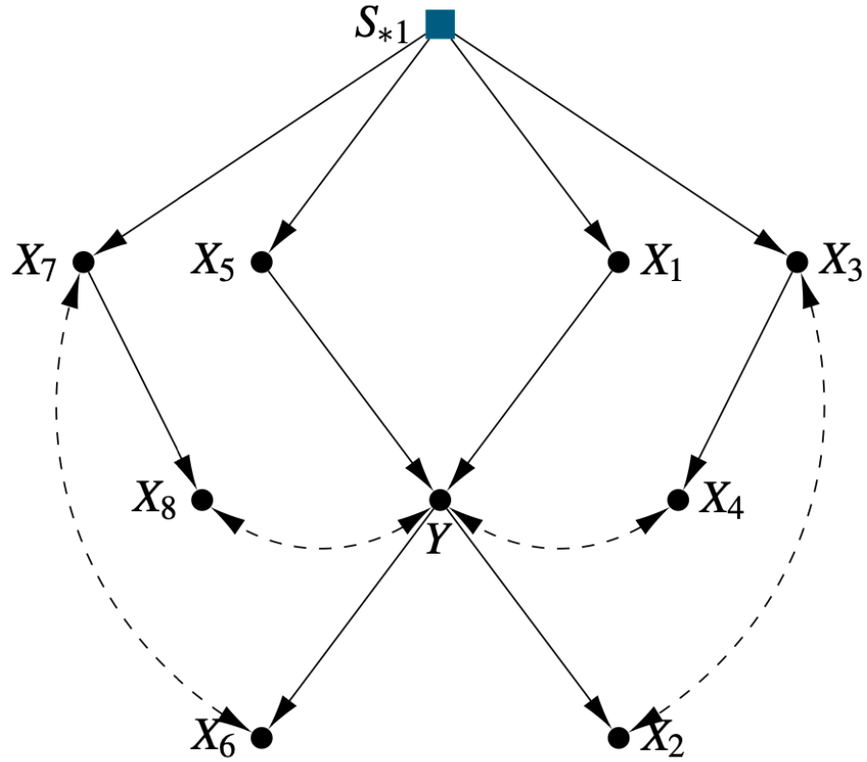
$$\{X_1, X_3, X_4\} \cup \{X_5, X_7, X_8\}$$




Same pattern can create $2^m$ maximal TR FS with only $4m + 1$ variables in the SCMs.

# Some experiments



# Conclusions

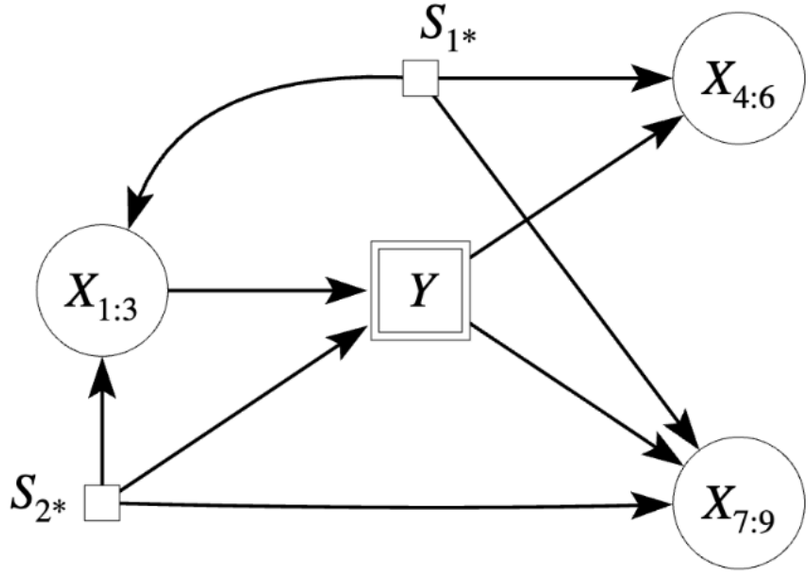
- A prediction rule may be effective in the source domain but perform poorly in the target domain.
- Empirical risk minimization (ERM) does not necessarily yield generalizable prediction rules.
- The target's optimal prediction rules, when based on transportable feature sets, can be estimated using only source data under the causal structural assumptions.
- Larger feature sets yield greater predictive power, although the number of maximally predictive transportable feature sets can grow exponentially.

# Representations

**Definition 11.2.6. (Representations and score functions).** The r.v. $\mathbf{R}$ with support $\text{supp}(\mathbf{R})$ is said to be a representation (of $\mathbf{X}$) if there exists a mapping $\phi : \text{supp}(\mathbf{X}) \to \text{supp}(\mathbf{R})$ such that $\mathbf{R} = \phi(\mathbf{X})$. Further, the corresponding score function is defined as $l_\phi(\mathbf{r}) := \mathbb{E}_{P*}[Y \mid \mathbf{R} = \mathbf{r}]$.

**Example 11.12.**

$$\mathbf{R} = \phi(\mathbf{X}) := \langle \underbrace{\beta^\top \cdot \mathbf{X}_{1:3}}_{R_1}, \underbrace{\sum_{i=1}^{6} X_i}_{R_2} \rangle$$




**Where** $\beta \sim \mathcal{N}(0, \mathbf{I}_3)$

# Transportable Representations

**Definition 11.2.7. (Transportable representation).** The rep. $\mathbf{R} = \phi(\mathbf{X})$ is said to be transportable from the collection of dists. $\mathbb{P}$ given $\mathcal{G}^{\Delta}$ if for every pair of tuples of SCMs that induce $\mathcal{G}^{\Delta}$ and entail $\mathbb{P}$, the quantity $\mathbb{E}_{P*}[Y \mid \phi(\mathbf{X}) = \mathbf{r}]$ evaluates to a unique value, for all $\mathbf{r} \in \text{supp}(\mathbf{R})$.

All transportable feature sets $\mathbf{Z} \subseteq \mathbf{X}$ correspond to a transportable representation.

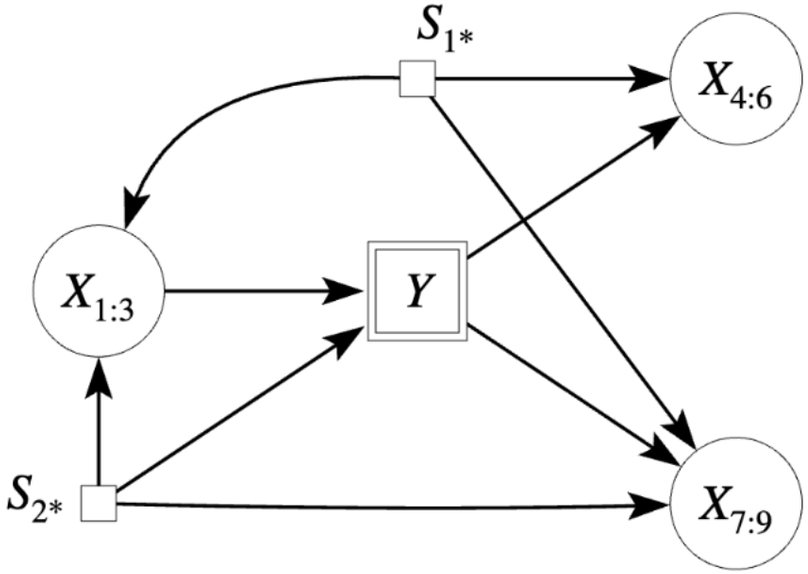
What other summarizations of $\mathbf{X}$ are transportable?

# Solving $\mathbf{R} = \phi(\mathbf{X})$

**Example 11.12.**

$$\mathbf{R} = \begin{cases} R_1 = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 \\ R_2 = X_1 + X_2 + \ldots + X_6 \end{cases}$$

$\in \mathbb{R} \times \{0,1,\ldots,6\}$




Since $\beta_i$ are random and independent, the support of $R_1$ contains exactly $2^3 = 8$ points with positive mass, thus the mapping is one-to-one ---> $x_1, x_2, x_3 = \phi_1^{-1}(R_1)$

$$\phi(\mathbf{X}) = \langle r_1, r_2 \rangle \iff \begin{cases} \langle X_1, X_2, X_3 \rangle = \langle x_1, x_2, x_3 \rangle \\ \phi^{\dagger}(\mathbf{X}_{4:6}) = r_2 - x_1 - x_2 - x_3 \end{cases}$$

# Det/cons/free variables

**Definition 11.2.8. (Determined, constrained, and free variables).**

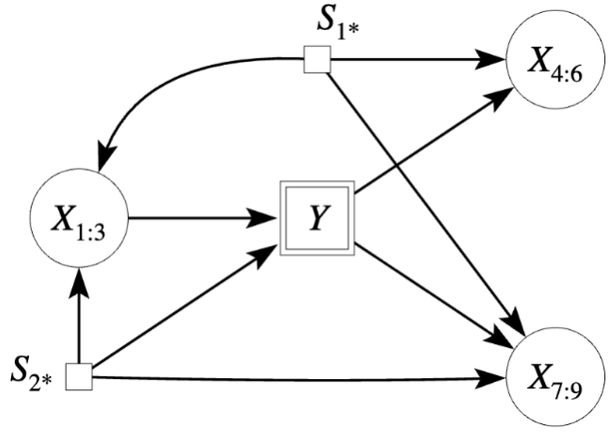
- $\det(\phi) = \mathbf{Z} \subseteq \mathbf{X}$ are said to be *determined* by $\phi$ if for every $\mathbf{r} \in \text{supp}(\mathbf{R})$, a single value for $\mathbf{Z}$ can be derived from the system of equation $\mathbf{R} = \phi(\mathbf{X})$.
- $\text{cons}(\phi) = \mathbf{Z}^\dagger \subseteq \mathbf{X} \backslash \mathbf{Z}$ are said to be *constrained by* $\phi$ if for at least one value of $\mathbf{r} \in \text{supp}(\mathbf{R})$ and at least one value $\mathbf{z}^\dagger \in \text{supp}(\mathbf{Z}^\dagger)$, the system of equation $\phi(\mathbf{X} \backslash \mathbf{Z}^\dagger, \mathbf{z}^\dagger) = \mathbf{r}$ is inconsistent.
- The rest of the variables are called *free from* $\phi$.

**Example 11.12.** $\mathbf{R} = \phi(\mathbf{X}) := \langle \beta^\top \cdot \mathbf{X}_{1:3}, \sum_{i=1}^{6} X_i \rangle$

$$\det(\phi) = \{X_1, X_2, X_3\}$$

$$\text{cons}(\phi) = \{X_4, X_5, X_6\}$$

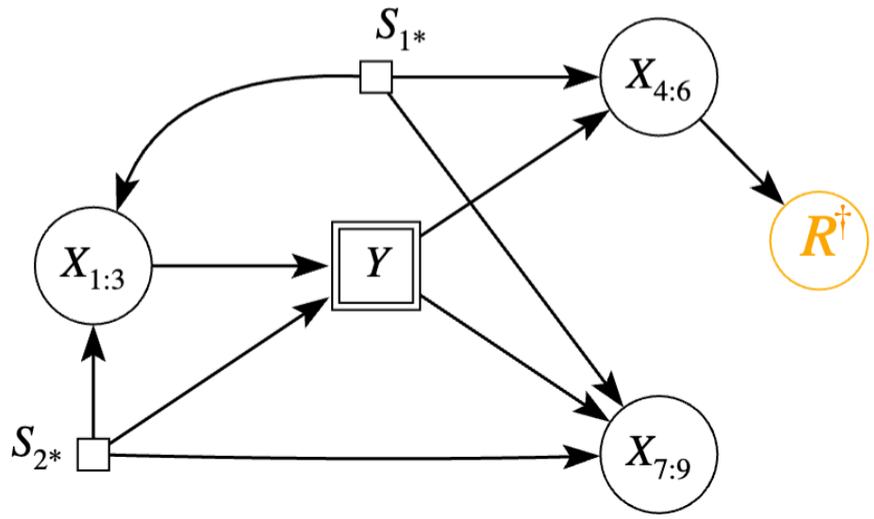
$$\text{free}(\phi) = \{X_7, X_8, X_9\}$$



# Augmented SD

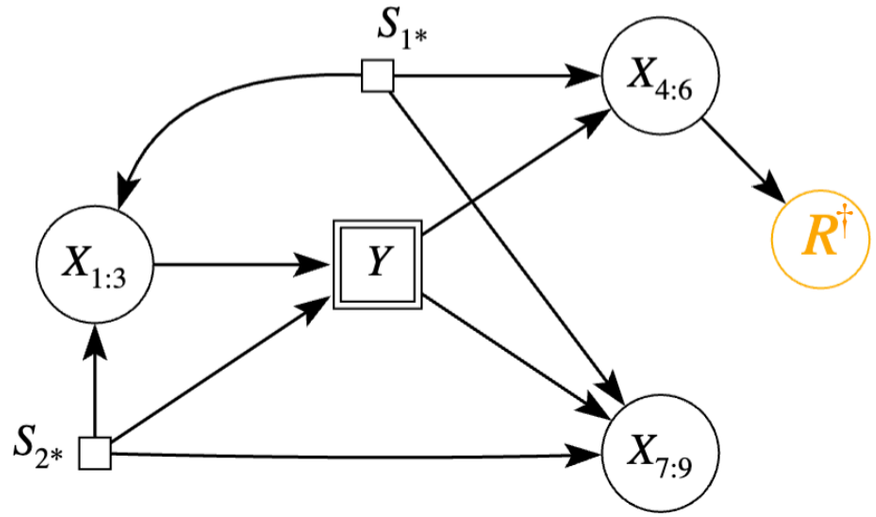
**Definition 11.2.8. (Augmented selection diagram).** Let $\mathcal{G}^{\Delta}$ be a selection diagram over $\mathbf{X}$, $Y$ and $\mathbb{P}$ denote a set of source distributions, and $\phi(\mathbf{X})$ be a representation. We augment the variable set with $\mathbf{R}^{\dagger}$ as a child of $\mathbf{Z}^{\dagger} = \text{cons}(\phi)$ variables. We set the mechanism $\mathbf{R}^{\dagger} = \phi^{\dagger}(\mathbf{Z}^{\dagger})$, and the new graph is called the augmented selection diagram denoted by $\mathcal{G}^{\Delta}_{\text{aug}}$.

$$\phi(\mathbf{X}) = \langle r_1, r_2 \rangle \iff \begin{cases} \langle X_1, X_2, X_3 \rangle = \langle x_1, x_2, x_3 \rangle \\ \phi^{\dagger}(\mathbf{X}_{4:6}) = \underbrace{r_2 - x_1 - x_2 - x_3}_{\mathbf{R}^{\dagger}} \end{cases}$$



# Transport a representation (I)

$$\phi(\mathbf{X}) = \langle r_1, r_2 \rangle \iff \begin{cases} \langle X_1, X_2, X_3 \rangle = \langle x_1, x_2, x_3 \rangle \\ \phi^\dagger(\mathbf{X}_{4:6}) = \underbrace{r_2 - x_1 - x_2 - x_3}_{\mathbf{R}^\dagger} \end{cases}$$




$$\begin{aligned}
l_\phi(\mathbf{R}) &= \mathbb{E}_{P^*}[Y \mid r_1, r_2] && \text{(Def. 11.2.6)} \\
&= P^*(Y = 1 \mid \mathbf{X}_{1:3}, r^\dagger) && \text{(Eq. 11.94)} \\
&= \frac{P^*(Y = 1 \mid \mathbf{X}_{1:3}) \cdot P^*(r^\dagger \mid Y = 1, \mathbf{X}_{1:3})}{\sum_{y=0}^{1} P^*(Y = y \mid \mathbf{X}_{1:3}) \cdot P^*(r^\dagger \mid Y = y, \mathbf{X}_{1:3})} && \text{(conditioning \& factorization)} \\
&= \frac{P^1(Y = 1 \mid \mathbf{X}_{1:3}) \cdot P^*(r^\dagger \mid Y = 1, \mathbf{X}_{1:3})}{\sum_{y=0}^{1} P^1(Y = y \mid \mathbf{X}_{1:3}) \cdot P^*(r^\dagger \mid Y = y, \mathbf{X}_{1:3})} && (S_{1*} \perp\!\!\!\perp Y \mid \mathbf{X}_{1:3} \text{ in } \mathcal{G}^\Delta_{\text{aug}}) \\
&= \frac{P^1(Y = 1 \mid \mathbf{X}_{1:3}) \cdot P^2(r^\dagger \mid Y = 1, \mathbf{X}_{1:3})}{\sum_{y=0}^{1} P^1(Y = y \mid \mathbf{X}_{1:3}) \cdot P^2(r^\dagger \mid Y = y, \mathbf{X}_{1:3})} && (S_{2*} \perp\!\!\!\perp R^\dagger \mid \mathbf{X}_{1:3}, Y \text{ in } \mathcal{G}^\Delta_{\text{aug}})
\end{aligned}$$

# Transport a representation (II)

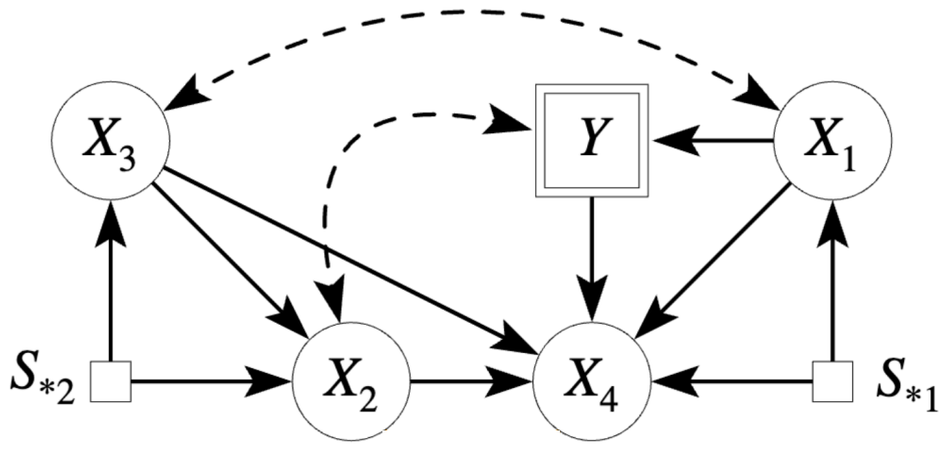
**Theorem 11.2.10. (Graphical TR of Rep.)** For a representation $\mathbf{R} = \phi(\mathbf{X})$, the score function can be expressed as

$$l_\phi(\mathbf{r}) = \mathbb{E}_{P*}[Y \mid \mathbf{r}] = P*(Y = 1 \mid \mathbf{z}, \mathbf{r}^\dagger)$$

The latter can be transported using the augmented selection diagram $\mathcal{G}^{\Delta}_{\text{aug}}$ using the Score-TR algorithm; in case of non-TR, the representation is non-TR too.

**Example 11.20.**

$$\mathbf{R} = \phi(\mathbf{X}) := \langle \underbrace{\frac{X_1 \cdot X_2 \cdot X_3}{X_4}}_{R_1}, \underbrace{\frac{X_1 \cdot X_2}{X_3 \cdot X_4}}_{R_2}, \underbrace{\frac{X_2 \cdot X_3}{X_1 \cdot X_4}}_{R_3} \rangle.$$

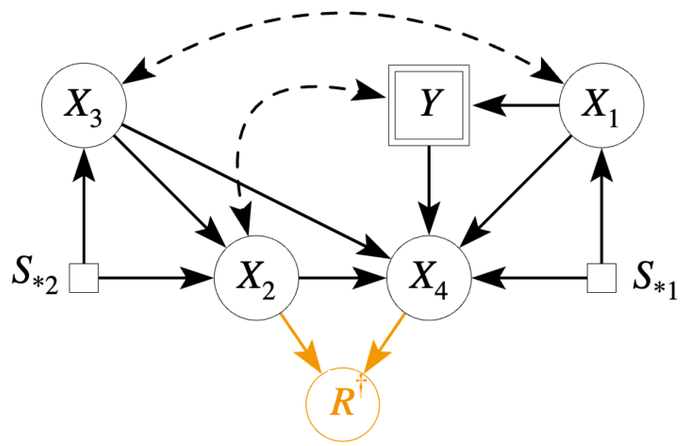



Compute $l_\phi(\mathbf{r}) = \mathbb{E}^*_P[Y \mid r_1, r_2, r_3]$

# Transport a representation (II)

**Example 11.20.**

$$\mathbf{R} = \phi(\mathbf{X}) := \langle \underbrace{\frac{X_1 \cdot X_2 \cdot X_3}{X_4}}_{R_1}, \underbrace{\frac{X_1 \cdot X_2}{X_3 \cdot X_4}}_{R_2}, \underbrace{\frac{X_2 \cdot X_3}{X_1 \cdot X_4}}_{R_3} \rangle.$$




$$\det(\phi) \begin{cases} X_1 = \sqrt{\frac{R_1}{R_3}}, \\ X_3 = \sqrt{\frac{R_1}{R_2}}, \end{cases} \quad \text{cons}(\phi) \quad \frac{X_2}{X_4} = \sqrt{R_2 \cdot R_3}. \quad (R^\dagger)$$

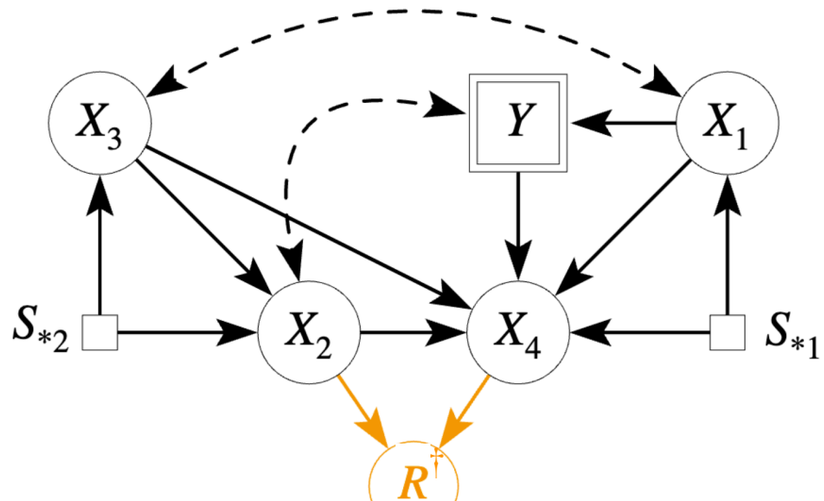
$$\Rightarrow \quad x_1^{\mathbf{R}} := \sqrt{\frac{r_1}{r_3}}, x_3^{\mathbf{R}} = \sqrt{\frac{r_1}{r_2}}, r^{\dagger \mathbf{R}} = \sqrt{r_2 \cdot r_3}$$

$$\Downarrow$$

$$Q : l_\phi(\mathbf{R}) = \mathbb{E}_{P^*}[Y \mid \mathbf{R} = \mathbf{r}] = \underbrace{P^*(Y = 1 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}}, r^{\dagger \mathbf{R}})}_{Q'}.$$

# Transport a representation (II)

**Example 11.20.**




$$Q : l_\phi(\mathbf{R}) = \mathbb{E}_{P^*}[Y \mid \mathbf{R} = \mathbf{r}] = \underbrace{P^*(Y = 1 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}}, r^{\dagger\mathbf{R}})}_{Q'}.$$

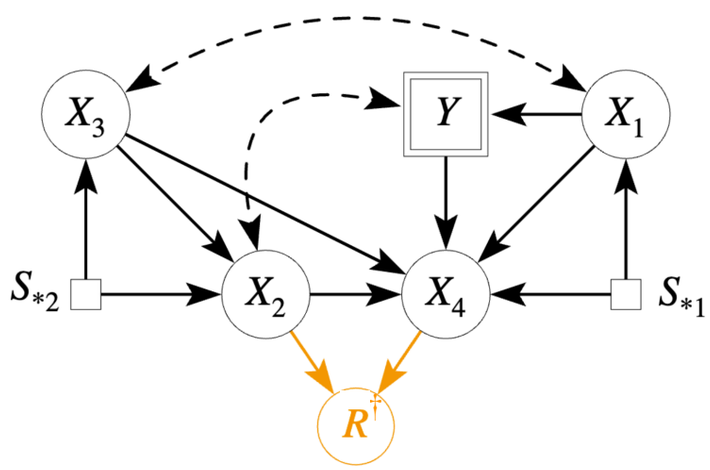
$$\begin{aligned} Q' &= P^*(y \mid r^{\dagger\mathbf{R}}, x_1^{\mathbf{R}}, x_3^{\mathbf{R}}) \\ &= \frac{\sum_{x_2,x_4} P^*(y, r^{\dagger\mathbf{R}}, x_2, x_4 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}})}{\sum_{y,x_2,x_4} \underbrace{P^*(y, r^{\dagger\mathbf{R}}, x_2, x_4 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}})}_{Q''}}. \end{aligned}$$

$$\left.\begin{aligned} Q'' &= P^*(r^{\dagger\mathbf{R}} \mid y, x_1^{\mathbf{R}}, x_2, x_3^{\mathbf{R}}, x_4) \cdot P^*(y, x_2, x_4 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}}) \\ &= P^*(r^{\dagger\mathbf{R}} \mid x_2, x_4) \cdot P^*(y, x_2, x_4 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}}) \\ &= 1_{\{r^{\dagger\mathbf{R}} = \frac{x_2}{x_4}\}} \cdot \underbrace{P^*(y, x_2, x_4 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}})}_{Q'''}, \\ Q''' &= P^*(y, x_2 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}}) \cdot P^*(x_4 \mid y, x_2, x_1^{\mathbf{R}}, x_3^{\mathbf{R}}) \\ &= P^1(y, x_2 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}}) \cdot P^2(x_4 \mid y, x_2, x_1^{\mathbf{R}}, x_3^{\mathbf{R}}), \end{aligned}\right\} \quad l_\phi(\mathbf{R}) = \frac{\sum_{x_2,x_4} P^1(y, x_2 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}}) \cdot P^2(x_4 \mid y, x_2, x_1^{\mathbf{R}}, x_3^{\mathbf{R}})}{\sum_{y,x_2,x_4} P^1(y, x_2 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}}) \cdot P^2(x_4 \mid y, x_2, x_1^{\mathbf{R}}, x_3^{\mathbf{R}})}$$

# Transport a representation (II)

**Example 11.20.**

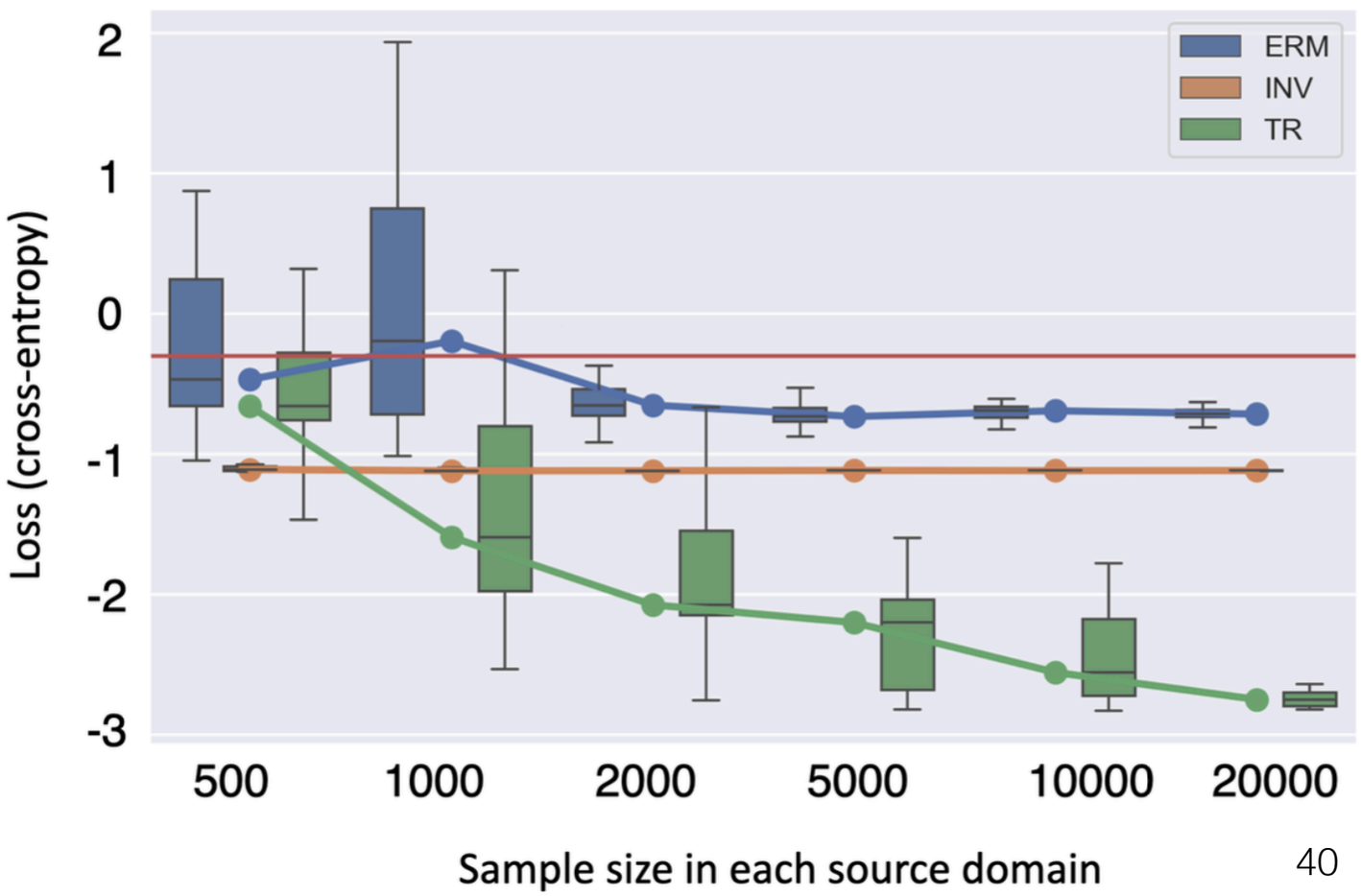
$$\mathbf{R} = \phi(\mathbf{X}) := \langle \underbrace{\frac{X_1 \cdot X_2 \cdot X_3}{X_4}}_{R_1}, \underbrace{\frac{X_1 \cdot X_2}{X_3 \cdot X_4}}_{R_2}, \underbrace{\frac{X_2 \cdot X_3}{X_1 \cdot X_4}}_{R_3} \rangle.$$




TR $\quad l_\phi(\mathbf{R}) = \dfrac{\sum_{x_2,x_4} P^1(y, x_2 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}}) \cdot P^2(x_4 \mid y, x_2, x_1^{\mathbf{R}}, x_3^{\mathbf{R}})}{\sum_{y,x_2,x_4} P^1(y, x_2 \mid x_1^{\mathbf{R}}, x_3^{\mathbf{R}}) \cdot P^2(x_4 \mid y, x_2, x_1^{\mathbf{R}}, x_3^{\mathbf{R}})}$

INV $\quad l_{X_1}(x_1) = \mathbb{E}_{P^1}[Y \mid x_1] = \mathbb{E}_{P^2}[Y \mid x_1]$

ERM $\quad \frac{1}{2}\mathbb{E}_{P^1}[Y \mid \mathbf{x}] + \frac{1}{2}\mathbb{E}_{P^2}[Y \mid \mathbf{x}]$



# Data-driven TR

**Challenge.** Can we define/compute transportability without graphical assumptions?

**Idea.** Instead of explicit access to $\mathcal{G}^{\Delta}$, we assume a particular structure to it.

**Assumption 11.2.11 -- Causal Mechanistic Stability (CMS).** We assume that the causal diagrams of all source domains is the same as that of the target domain. Moreover, we assume that if a variable does not belong to any cross-source domain discrepancy set, it does not belong to a source-target domain discrepancy set either, i.e., $V \notin \bigcup_{i,j=1}^{K} \Delta_{i,j} \implies V \notin \bigcup_{k=1}^{K} \Delta_{k,*}$

# Example of CMS holding

**Example 11.21. Pictures of cats and dogs in different seasons**

**Sources**

**Target**

Summer Winter Spring Fall

$$\text{lightOn} \leftarrow \neg\text{isSunny} \oplus U$$

$$P^{1=2=3}(\text{LightOn} \mid \text{isSunny}) = P^*(\text{LightOn} \mid \text{isSunny})$$

# Example of CMS violation

**Example 11.22. The isSunny feature is not visible anymore.**

**Sources**

Summer

Winter

Spring

**Target**

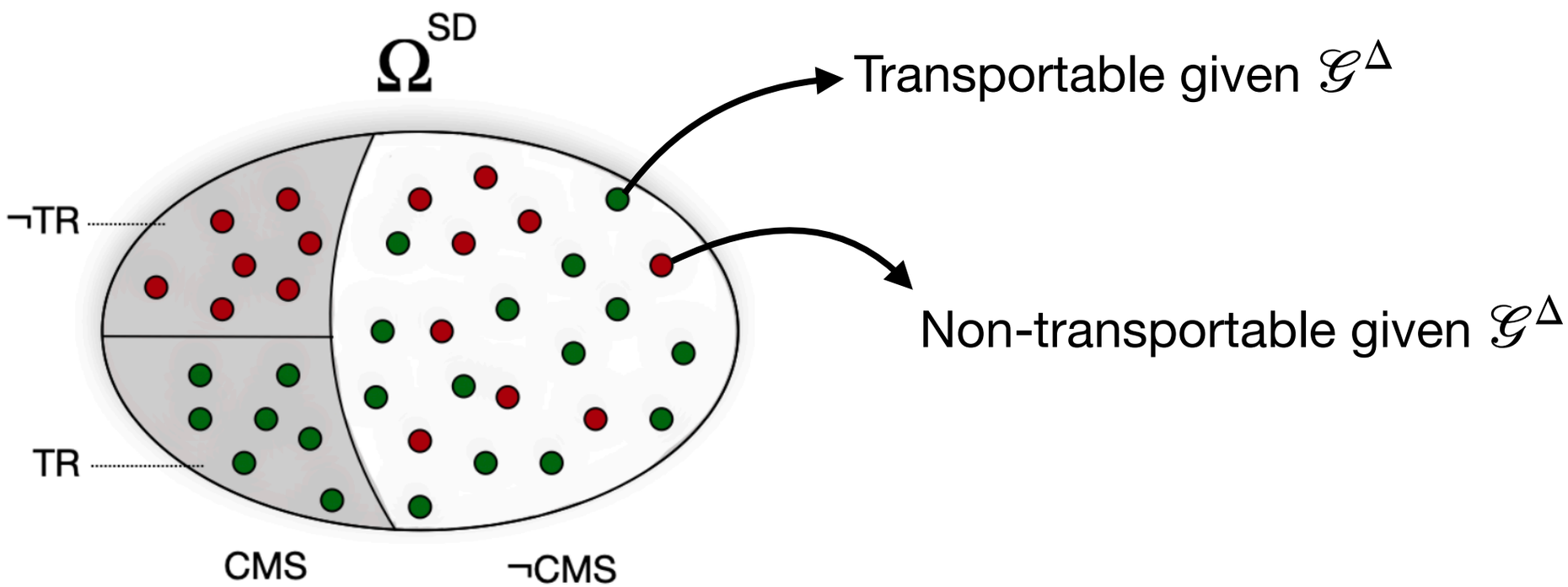
Fall

$\text{lightOn} \leftarrow \neg \text{i}$  $\text{ny} \oplus U$

$$P^1(\text{LightOn}) \neq P^2(\text{LightOn}) \neq P^3(\text{LightOn}) \neq P^*(\text{LightOn})$$

# Data-driven TR

**Definition 11.2.10 -- TR representation: Data-driven.** The representation $\mathbf{R} = \phi(\mathbf{X})$ is transportable from the source distributions $\mathbb{P}$ under CMS assumption if for every tuple of source and target SCMs $\langle \mathcal{M}_a^1, \mathcal{M}_a^2, \ldots, \mathcal{M}_a^K, \mathcal{M}_a^* \rangle$ and $\langle \mathcal{M}_b^1, \mathcal{M}_b^2, \ldots, \mathcal{M}_b^K, \mathcal{M}_b^* \rangle$ that satisfy CMS and entail $\mathbb{P}$, it holds that $\mathbb{E}_{P^{\mathcal{M}_a^*}}[Y \mid \mathbf{r}] = \mathbb{E}_{P^{\mathcal{M}_b^*}}[Y \mid \mathbf{r}]$ for all $\mathbf{r} \in \text{supp}(\mathbf{R})$.



# Exchangeable domains

**Lemma 11.2.12 -- Exchangeable domains.** Under the assumption of Causal Mechanistic Stability, there exists a possible target SCM that entails $P^*(\mathbf{v}) = P^i(\mathbf{v})$ for each source distribution $P^i \in \mathbb{P}$.

Having exchangeable domains under CMS indicates a drawback of CMS in comparison to the graphical assumptions, as elaborated in Ex. 11.23.

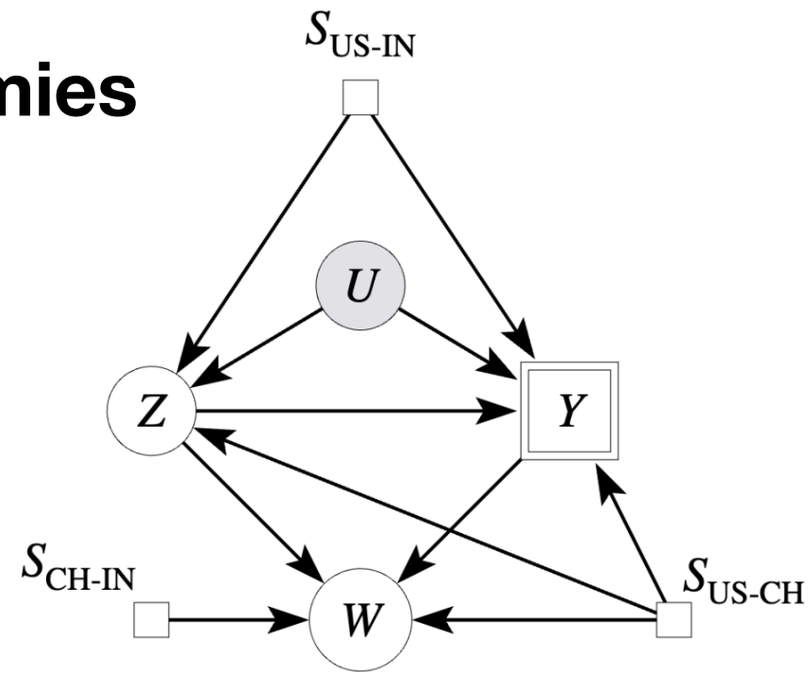
## Example 11.23. Generalizing across economies

$U$ : Government policy efficiency

$Z$ : Technological innovations

$Y$ : Decision on significant economic reform

$W$ : Economic growth rate



# Exchangeable domains violated

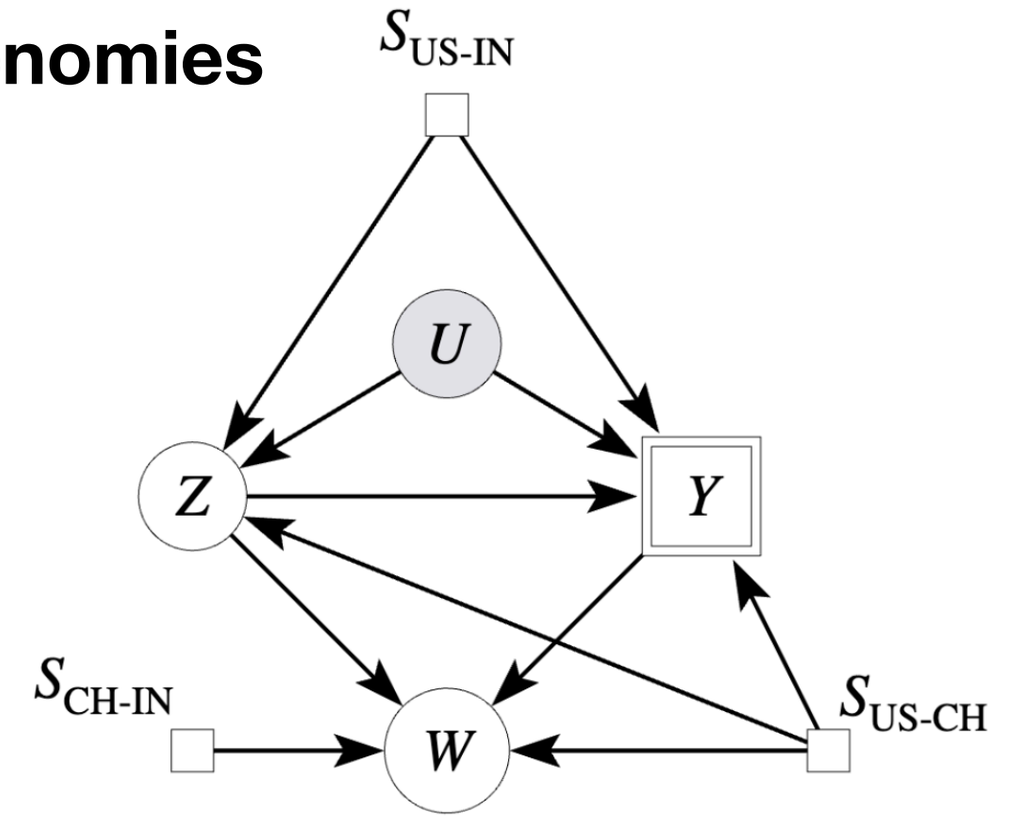
**Example 11.23. Generalizing across economies**

$U$ : Government policy efficiency

$Z$ : Technological innovations

$Y$ : Decision on significant economic reform

$W$ : Economic growth rate




Representation $\mathbf{R} = \phi(z, w) = \langle z, w \rangle$

$$
\left.
\begin{aligned}
l_\phi(z,w) &= \mathbb{E}_{P^{\text{IN}}}[Y \mid z,w] \\
&= \frac{P^{\text{IN}}(Y=1,z,w)}{\sum_{y=0}^{1} \underbrace{P^{\text{IN}}(y,z,w)}_{Q}}. \\
Q &= P^{\text{IN}}(y,z) \cdot P^{\text{IN}}(w \mid y,z) \\
&= P^{\text{CH}}(y,z) \cdot P^{\text{US}}(w \mid y,z),
\end{aligned}
\right\}
\quad
\mathbb{E}_{P^{\text{IN}}}[Y \mid z,w] = \frac{P^{\text{CH}}(Y=1,z) \cdot P^{\text{US}}(w \mid Y=1,z)}{\sum_{y=0}^{1} P^{\text{CH}}(y,z) \cdot P^{\text{US}}(w \mid y,z)}.
$$

# Invariance

**Corollary 11.2.13 -- Necessity of invariance.** If a representation $\mathbf{R} = \phi(\mathbf{X})$ is transportable from source distributions $\mathbb{P}$ given CMS, then the score function is invariant across all domains, i.e.,

$$\mathbb{E}_{P1}[Y \mid \mathbf{r}] = \mathbb{E}_{P2}[Y \mid \mathbf{r}] = \ldots = \mathbb{E}_{PK}[Y \mid \mathbf{r}] = \mathbb{E}_{P*}[Y \mid \mathbf{r}]$$

**Def. 11.2.11 -- Source invariance property.** If the score function matches between two sources $\mathscr{M}^i$, $\mathscr{M}^j$, we denote it as $\text{INV}_{i,j}[\phi] : \mathbb{E}_{P^i}[Y \mid \mathbf{r}] = \mathbb{E}_{P^j}[Y \mid \mathbf{r}]$. The source invariance property is defined as $\bigwedge_{i,j=1}^{k} \text{INV}_{i,j}(\phi)$, and in this case $\phi$ is called an *invariant representation*.

# r-faithfulness

**Assumption 11.2.14 -- r-faithfulness.** The collection of source distributions $\mathbb{P} = \langle P^1, P^2, \ldots, P^k \rangle$ is r-faithful to the selection diagram $\mathcal{G}^{\Delta}$ if for all representations $\mathbf{R} = \phi(\mathbf{X})$ and for every $i, j \in \{1,2,...,K\}$,

$$\text{INV}_{i,j}(\phi) \implies S_{i,j} \perp_d Y \mid \mathbf{Z}, \mathbf{R}^{\dagger} \text{ in } \mathcal{G}^{\Delta}_{\text{aug}}$$

Where $\mathbf{Z} = \det(\phi)$ and $\mathbf{R}^{\dagger} = \phi^{\dagger}(\mathbf{Z}^{\dagger})$ denotes the constraints over the constrained variables $\mathbf{Z}^{\dagger} = \text{cons}(\phi)$.

**Theorem 11.2.15 -- Data-driven transportability.** For a representation $\mathbf{R} = \phi(\mathbf{X})$, under r-faithfulness, CMS, the score function $l_{\phi} : \mathbb{E}_{P^*}[Y \mid \mathbf{r}]$ can be transported from source distributions $\mathbb{P}$ if and only if $\phi$ satisfies the source invariance property (sound and complete criterion for TR under CMS).

# Summary

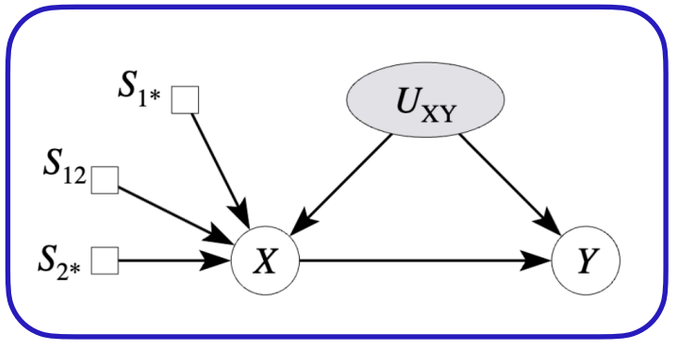
- One can extend the notion of transportability beyond just subsets of features by considering arbitrary mappings called representations.
- On can use the selection diagrams to decide (algorithmically) whether a representation is useful for generalization, i.e., the score function is computable uniquely.
- We can relax the assumptions encoded in the selection diagram, and instead use an implicit causal assumption we call Causal Mechanistic Stability (CMS).
- CMS is a strong assumption, since every representation that is TR under CMS is invariant across the source domains.
- Additional regularity assumption (r-faithfulness) guarantees that source invariance is the data-driven criterion for deciding TR under CMS.

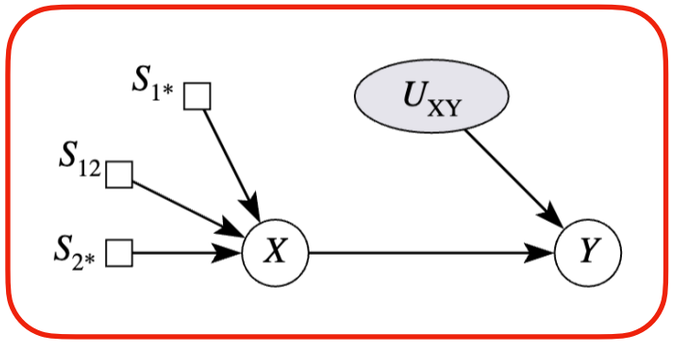
**How do we find invariant representations?**

# Searching for invariant representations

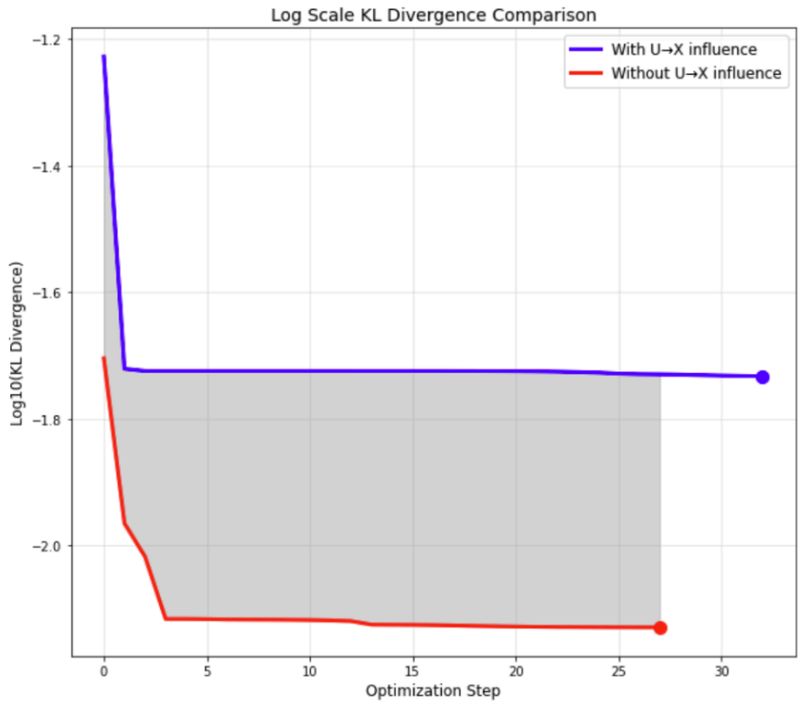
We showed that under CMS the only transportable representations are invariant representations, but it remains a challenge to find such representations.

**Example 11.25. Finding invariant rep.**




$$\min_{\beta} D_{KL}\left(P^1(y \mid \beta^\top \cdot x), P^2(y \mid \beta^\top \cdot x)\right)$$




**Corollary 11.2.16 -- impossibility of transport.** Under CMS and r-faithfulness assumptions, if $\phi(\mathbf{X}) = \mathbf{Pa}_Y$ is not an invariant representation, then there exists no transportable representation.

# $\alpha$-loss & $\alpha$-risk

**Definition 11.2.12 -- $\alpha$-loss.** For parameter $\alpha \in (0,1)$, we construct a class-sensitive loss function as follows:

$$\mathscr{L}^{\alpha}(y,\hat{y}) := \frac{1}{\alpha} \cdot 1[y=0,\hat{y}=1] + \frac{1}{1-\alpha} \cdot 1[y=1,\hat{Y}=0]\,.$$

For a classifier $h : \text{supp}(\mathbf{X}) \to \{0,1\}$, the expected $\alpha$-loss under $P(\mathbf{x}, y)$ is called $\alpha$-risk, and is denoted by,

$$\mathscr{R}_P^{\alpha}(h) = \mathbb{E}_P[\mathscr{L}^{\alpha}(Y, h(\mathbf{X}))].$$

For example, $\frac{1}{2}$-loss is the regular symmetric loss.

For $\alpha > \frac{1}{2}$ we incur larger loss for **false negatives** and for $\alpha < \frac{1}{2}$ we incur larger loss for **false positives.**

# Uniform invariant risk minimization

## Definition 11.2.13 -- The optimization scheme

**Idea:** Find a representation such that for every level of class-sensitivity $\alpha$, the optimal classifier matches across the source domains.

$$\min_{\phi, \{h^\alpha\}_{\alpha \in [0,1]}} \quad \sum_{i=1}^{K} \mathscr{R}_{Pi}^{\frac{1}{2}}(h^{\frac{1}{2}} \circ \phi)$$

$$\text{s.t.} \quad \forall \alpha \in [0,1] \quad \forall i \in \{1, 2, \ldots, K\} : h^\alpha \in \underset{\tilde{h}:\text{supp}(\mathbf{R}) \to \{0,1\}}{\arg\min} \quad \mathscr{R}_{Pi}^{\alpha}(\tilde{h} \circ \phi)$$

Lagrangian, i.e., the corresponding penalized likelihood program:

$$\min_{\phi, \{h_\theta^\alpha\}_{\alpha \in [0,1]}} \quad \sum_{i=1}^{K} \mathscr{R}_{Pi}^{\frac{1}{2}}(h_\theta^{\frac{1}{2}} \circ \phi) + \lambda \cdot \int_0^1 \|\nabla_\theta \mathscr{R}_{Pi}^{\alpha}(h_\theta^\alpha \circ \phi)\|^2 \cdot d\alpha$$

# UIRM finds invariant reps

**UIRM constraints:**

$$\min_{\phi,\{h^\alpha\}_{\alpha\in[0,1]}} \quad \sum_{i=1}^{K} \mathscr{R}_{P^i}^{\frac{1}{2}}(h^{\frac{1}{2}} \circ \phi)$$
$$\text{s.t.} \quad \forall \alpha \in [0,1] \quad \forall i \in \{1,2,\ldots,K\} : h^\alpha \in \underset{\tilde{h}:\text{supp}(\mathbf{R})\to\{0,1\}}{\arg\min} \mathscr{R}_{P^i}^{\alpha}(\tilde{h} \circ \phi)$$

**Proposition 11.2.17 -- Source invariance & UIRM.** Under CMS and r-faithfulness, a representation satisfies the source invariance property <u>if and only if</u> it satisfies the UIRM constraints. Furthermore, any solution to UIRM satisfies the source invariance property, thus, generalizes to the target domain under CMS and r-faithfulness.

**Question:** Can we use UIRM (penalized) for checking if there exists an invariant representation?

**Yes, tend $\lambda \to \infty$ and see if the following diverges.**

$$\min_{\phi,\{h_\theta^\alpha\}_{\alpha\in[0,1]}} \quad \sum_{i=1}^{K} \mathscr{R}_{P^i}^{\frac{1}{2}}(h_\theta^{\frac{1}{2}} \circ \phi) + \lambda \cdot \int_0^1 \|\nabla_\theta \mathscr{R}_{P^i}^{\alpha}(h_\theta^{\alpha} \circ \phi)\|^2 \cdot d\alpha$$

# Maximal transportable reps

**Definition 11.2.17 -- Maximal TR reps.** A representation $\phi_a$ is *preferable* to $\phi_b$ if there exists a tuple of source and target SCMs $\langle \mathscr{M}^1, \mathscr{M}^2, \ldots, \mathscr{M}^K, \mathscr{M}^* \rangle$ that respect CMS and entails $\mathbb{P}$, where,

$$\min_{h_a} \mathscr{R}_{P^{\mathscr{M}^*}}(h_a \circ \phi_a) \leq \min_{h_a} \mathscr{R}_{P^{\mathscr{M}^*}}(h_a \circ \phi_a).$$

**A maximal TR rep is one that is preferable to all other transportable reps.**

There may be many maximal TR reps (akin to maximal TR feature sets).

**Theorem 11.2.18 -- Maximal TR & UIRM.** Any solution to UIRM is a maximal transportable representation.

# Not all invariances are created equal: Failure of DANN

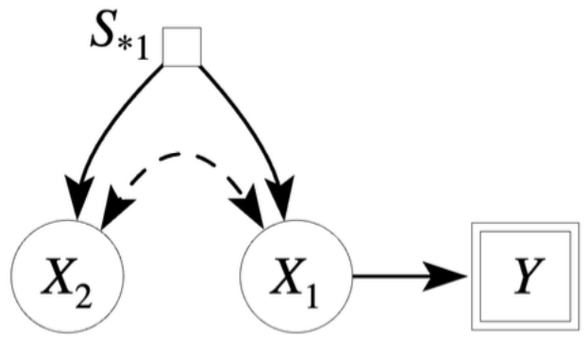
Domain adversarial training of neural networks by Ganin et al. (2016) is a practical implementation of a criterion proposed by Ben-David et al. (2006) that summarizes as follows:

**Theorem 2 -- Ben-David et al. (2006) (informal).** A representation $\phi$ generalizes well from a single source to a target if $P(\phi(\mathbf{X})) \approx P^*(\phi(\mathbf{X}))$.

## Example 11.26 -- A failure case of DANN.

$$
\begin{aligned}
U &\sim \text{unif}(\{1,2,\ldots,100\}) \\
U_Y &\sim \mathcal{N}(0,1) \\
X_1 &\leftarrow \alpha_1 \cdot U \\
X_2 &\leftarrow \alpha_2 \cdot U \\
Y &\leftarrow \begin{cases} 1 \text{ if } X + U_Y \geq c \\ 0 \text{ otherwise.} \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
U &\sim \text{unif}(\{1,2,\ldots,100\}) \\
U_Y &\sim \mathcal{N}(0,1) \\
X_1 &\leftarrow \alpha_1 \cdot U + \delta_1 \\
X_2 &\leftarrow \alpha_2 \cdot U + \delta_2 \\
Y &\leftarrow \begin{cases} 1 \text{ if } X + U_Y \geq c \\ 0 \text{ otherwise.} \end{cases}
\end{aligned}
$$



# Not all invariances are created equal: Failure of REx

Out-of distribution generalization via Risk Extrapolation by Krueger et al. (2020) proposes matching the risk across the source domains, as expressed below:

**Definition 11.2.17 — Variance Risk Extrapolation (V-REx)..** The V-REx predictor is the solution to the following optimization problem,

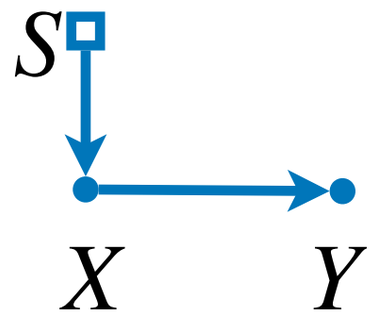
$$h^{\text{VREx}} \in \underset{h:\text{supp}(\mathbf{X})\to\{0,1\}}{\arg\min} \; \beta \cdot \text{Var}(\{\mathscr{R}_{P1}(h), \mathscr{R}_{P2}(h) \ldots, \mathscr{R}_{PK}(h)\}) + \sum_{i=1}^{T} \mathscr{R}_{Pi}(h), \qquad (11.184)$$

where $0 < \beta \leq \infty$ penalizes the variation in the risk value across the source domains.

## Example 11.27.

$$\mathscr{M}^1\text{:} \quad \begin{aligned} U_X &\sim \text{Bern}(0.9) \\ U_{Y1} &\sim \text{Bern}(0.9) \\ U_{Y0} &\sim \text{Bern}(0.6) \\ X &\leftarrow U_X \\ Y &\leftarrow \begin{cases} U_{Y1} & \text{if } X = 1 \\ U_{Y0} & \text{otherwise.} \end{cases} \end{aligned}$$

$$\mathscr{M}^2\text{:} \quad \begin{aligned} U_X &\sim \text{Bern}(0.1) \\ U_{Y1} &\sim \text{Bern}(0.9) \\ U_{Y0} &\sim \text{Bern}(0.6) \\ X &\leftarrow U_X \\ Y &\leftarrow \begin{cases} U_{Y1} & \text{if } X = 1 \\ U_{Y0} & \text{otherwise.} \end{cases} \end{aligned}$$




$$\mathscr{R}_{P1}(h) = 0.1 \cdot 0.9 + 0.4 \cdot 0.1 = \mathbf{0.15},$$
$$\mathscr{R}_{P2}(h) = 0.1 \cdot 0.1 + 0.4 \cdot 0.9 = \mathbf{0.37}.$$

# Not all invariances are created equal: Causal balanced-rate classifiers

Matching other notions of error across the source domains yields transportability.

False omission rate: $\mathrm{FOR}_P(h) := P(Y = 1 \mid h(\mathbf{X}) = 0),$

False discovery rate: $\mathrm{FDR}_P(h) := P(Y = 0 \mid h(\mathbf{X}) = 1).$

**Definition 11.2.18 -- balanced rate classifiers**

$$h^{\infty}_{\mathbf{R}} \in \min_{h:\mathrm{supp}(\mathbf{X})\to\{0,1\}} \sum_{i=1}^{K} \mathbf{R}_{P^i}(h)$$

$$\text{s.t.} \quad \mathrm{FOR}_{P^i}(h) = \mathrm{FOR}_{P^j}(h), \ \ \forall P^i \in \mathbb{P} \text{and} \mathrm{FDR}_{P^i}(h) = \mathrm{FDR}_{P^j}(h), \quad \forall P^i \in \mathbb{P}.$$

**Theorem 11.2.22.** Balanced rate classifiers generalize to the target domain under CMS and r-faithfulness.

# Not all invariances are created equal: Validity of Multi-Calibration for DG

**Definition 11.2.19 — Multi-Calibrated score.** A representation $\psi(\mathbf{X})$ with the support $[0,1]$ is called a score function. It is calibrated w.r.t. the distribution $P$ if,

$$\forall e \in [0,1] : \mathbb{E}_P[Y \mid \psi(\mathbf{X}) = e] = e. \tag{11.205}$$

It is called multi-calibrated if it is calibrated w.r.t. all source distributions.

**Lemma 11.2.23 — Source invariance & multi-calibration.** If any representation $\mathbf{R} = \phi(\mathbf{X})$ satisfies the source invariance property, which is,

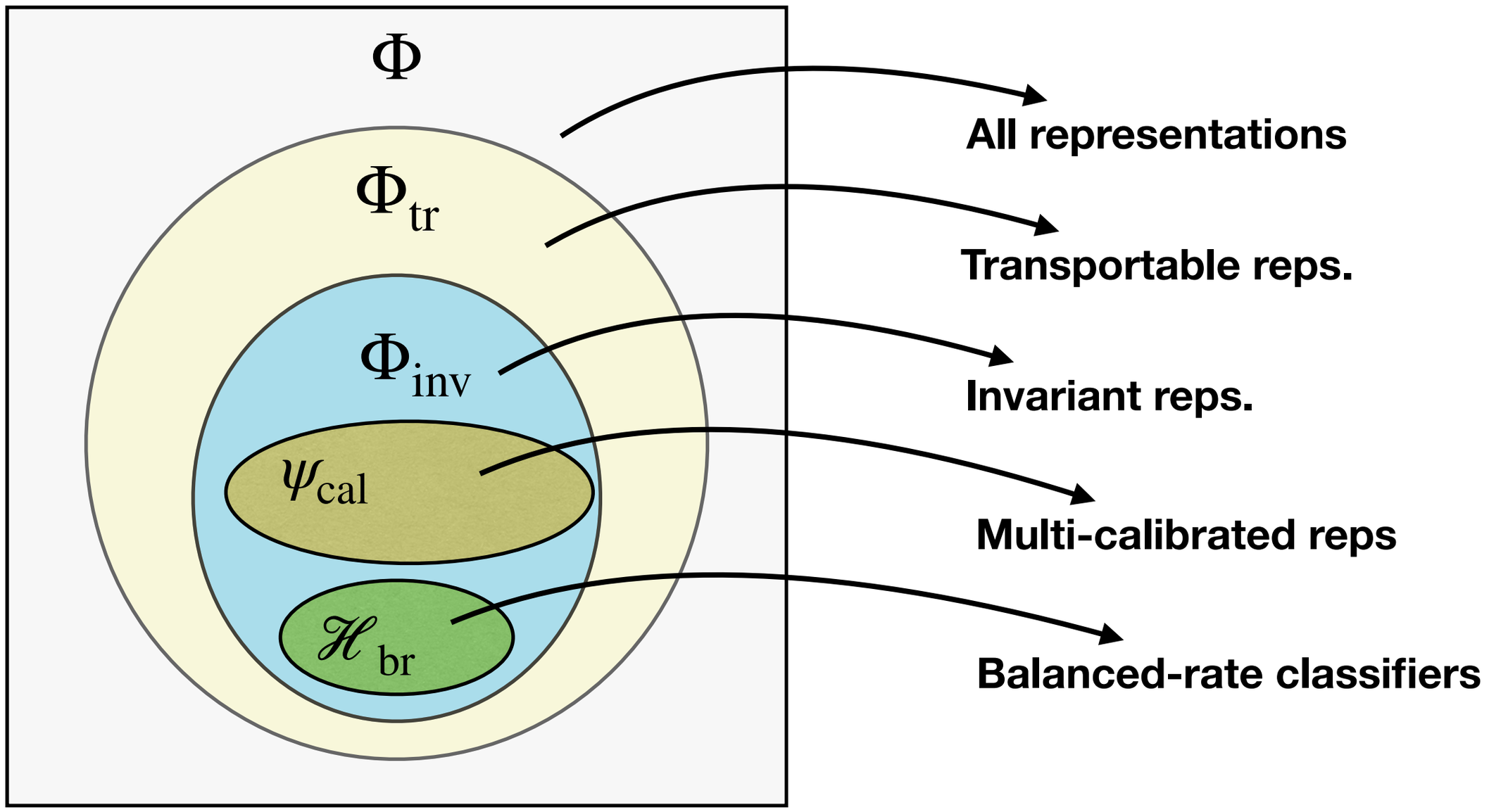
$$\text{INV}_{ij}(\phi) : \mathbb{E}_{P^i}[Y \mid \phi(\mathbf{X}) = \mathbf{r}] = \mathbb{E}_{P^j}[Y \mid \phi(\mathbf{X}) = \mathbf{r}], \quad \forall P^i, P^j \in \mathbb{P}, \tag{11.220}$$

then the score $\psi(\mathbf{x}) := \mathbb{E}_{P^i}[Y \mid \mathbf{R} = \phi(\mathbf{x})]$ (for any of the source distributions $P^i \in \mathbb{P}$) is multi-calibrated. Moreover, for every $P^i \in \mathbb{P}$,

$$I_{P^i}(Y; \phi(\mathbf{X})) = I_{P^i}(Y; \psi(\mathbf{X})), \tag{11.221}$$

meaning that $\phi$ and $\psi$ have equivalent prediction power on all source domains.

# Taxonomy of representations



# Summary

- Various notions of invariance have been explored for DG task.
- Not all of them are valid in the causal transportability framework.
- Even if an optimization scheme is valid, it is not easy to solve them or arrive at good local minima using gradient-based methods.
- Transportability under Causal Mechanistic Stability unifies some of the existing work in DG, and extends their theoretical scope beyond specific graphs and linear assumptions.