Causal Fairness Analysis

(Causal Inference II - Lecture 3)

Elias Bareinboim 

Drago Plecko 

Columbia University
Computer Science

Reference:

D. Plecko, E. Bareinboim.

Causal Fairness Analysis.

TR R-90, CausalAI Lab, Columbia University.

https://causalai.net/r90.pdf

Fairness Map (recap)

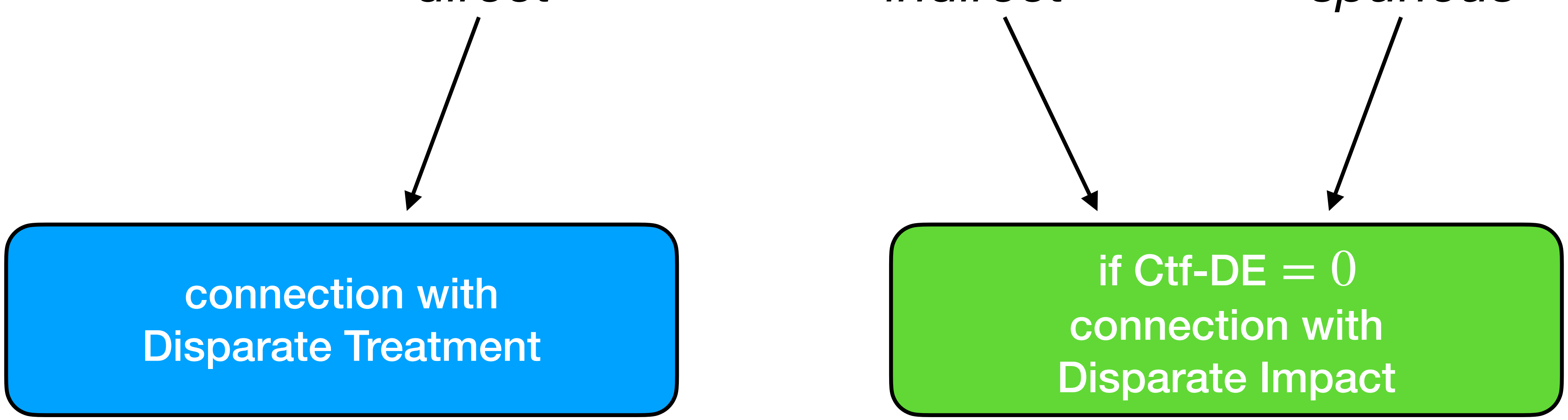
$$\text{TV} = E[Y \mid x_1] - E[Y \mid x_0]$$

		Direct	Indirect	Spurious
$P(u)$	TE	NDE	NIE	Exp-SE
$P(u \mid x)$	x-TE	x-DE	x-IE	x-SE
$P(u \mid z)$	z-TE	z-DE	z-IE	✘
$P(u \mid v)$	v-TE	v-DE	v-IE	✘
δ_u	u-TE	u-DE	u-IE	✘

Implications

Theorem (Zhang & Bareinboim, 2018). *The total variation (TV) measure admits a decomposition into counterfactual direct, indirect, and spurious effects*

$$TV_{x_0,x_1}(y) = \underbrace{\text{Ctf-DE}_{x_0,x_1}(y \mid x_0)}_{\text{direct}} - \underbrace{\text{Ctf-IE}_{x_1,x_0}(y \mid x_0)}_{\text{indirect}} + \underbrace{\text{Ctf-SE}_{x_1,x_0}(y)}_{\text{spurious}}.$$



Implications

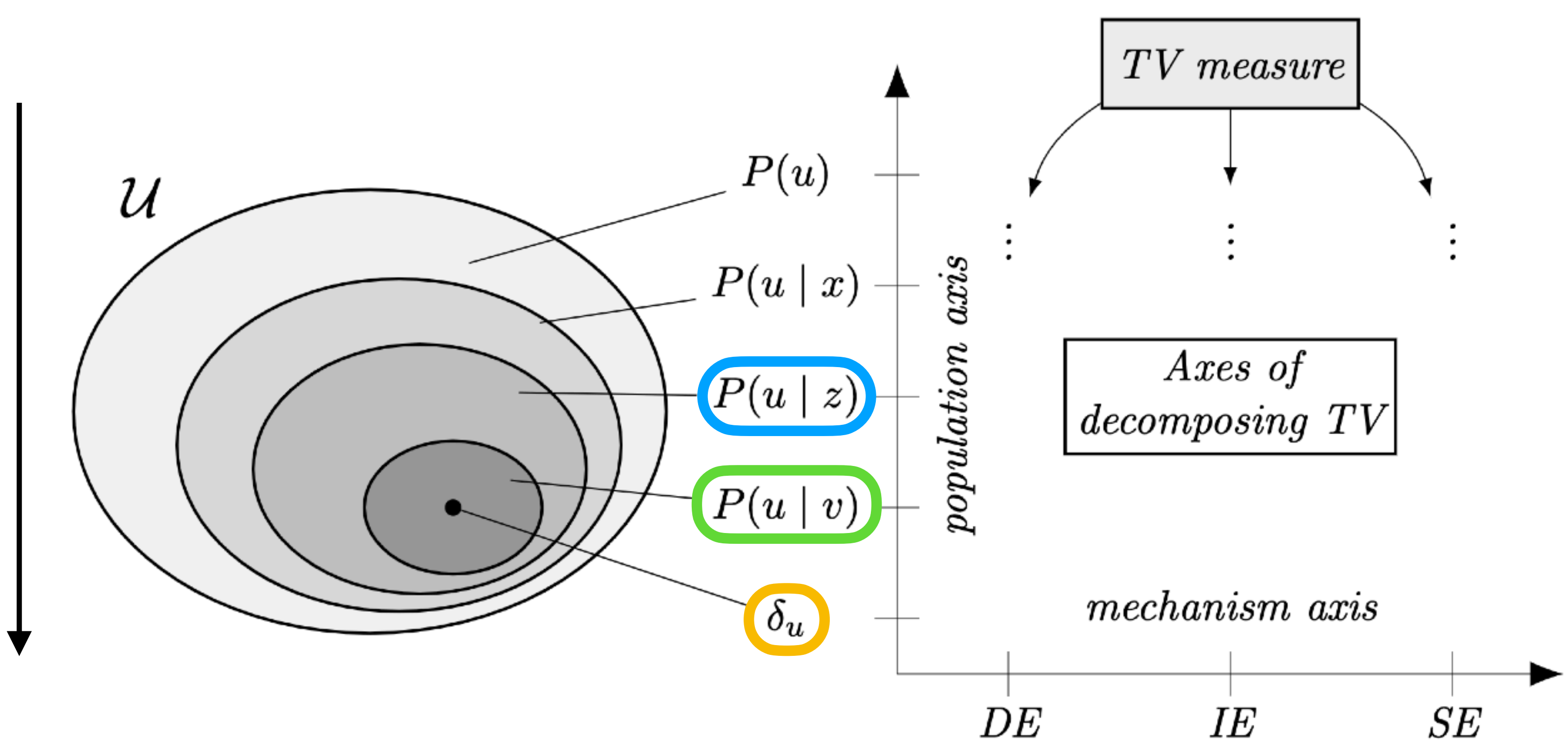
Theorem. The z-specific, v'-specific, and unit-level total effects admit a decomposition into direct and indirect effects:

$$z\text{-}TE_{x_0,x_1}(y \mid z) = z\text{-}DE_{x_0,x_1}(y \mid z) + z\text{-}IE_{x_1,x_0}(y \mid z)$$

$$v'\text{-}TE_{x_0,x_1}(y \mid v') = v'\text{-}DE_{x_0,x_1}(y \mid v') + v'\text{-}IE_{x_1,x_0}(y \mid v')$$

$$unit\text{-}TE_{x_0,x_1}(y(u)) = unit\text{-}DE_{x_0,x_1}(y(u)) + unit\text{-}IE_{x_1,x_0}(y(u))$$

decomposing works on different levels of precision in U



FPCFA (with Identification)

Definition. Let μ be a fairness measure defined over a space of SCMs Ω. Let $Q_1, \ldots, Q_k$ be a collection of structural fairness criteria. The Fundamental Problem of Causal Fairness Analysis is to find a collection of measures $\mu_1, \ldots, \mu_k$ s.t. the following properties are satisfied:

(i) μ is *decomposable* w.r.t. $\mu_1, \ldots, \mu_k$ **Decomposability**

(ii) $\mu_1, \ldots, \mu_k$ are *admissible* w.r.t. the structural fairness criteria $Q_1, Q_2, \ldots, Q_k$ **Admissibility**

(iii) $\mu_1, \ldots, \mu_k$ are as *powerful* as possible. **Power**

} **from before**

(iv) $\mu_1, \ldots, \mu_k$ are identifiable from the SFM and observational data. **Identifiability**

Example (Limitation of NDE). A new startup company is currently in hiring season. The hiring decision ($Y \in \{0,1\}$indicating whether the candidate is hired) is based on gender ($X \in \{0,1\}$, female and male, respectively), age ($Z \in \{0,1\}$, younger and older than 40 years, respectively), and education level ($W \in \{0,1\}$ which indicates whether the applicant has a Ph.D. degree). Following the legal guidelines, the startup is in this case obliged to avoid disparate treatment in hiring.

SCM M*
(unobserved)

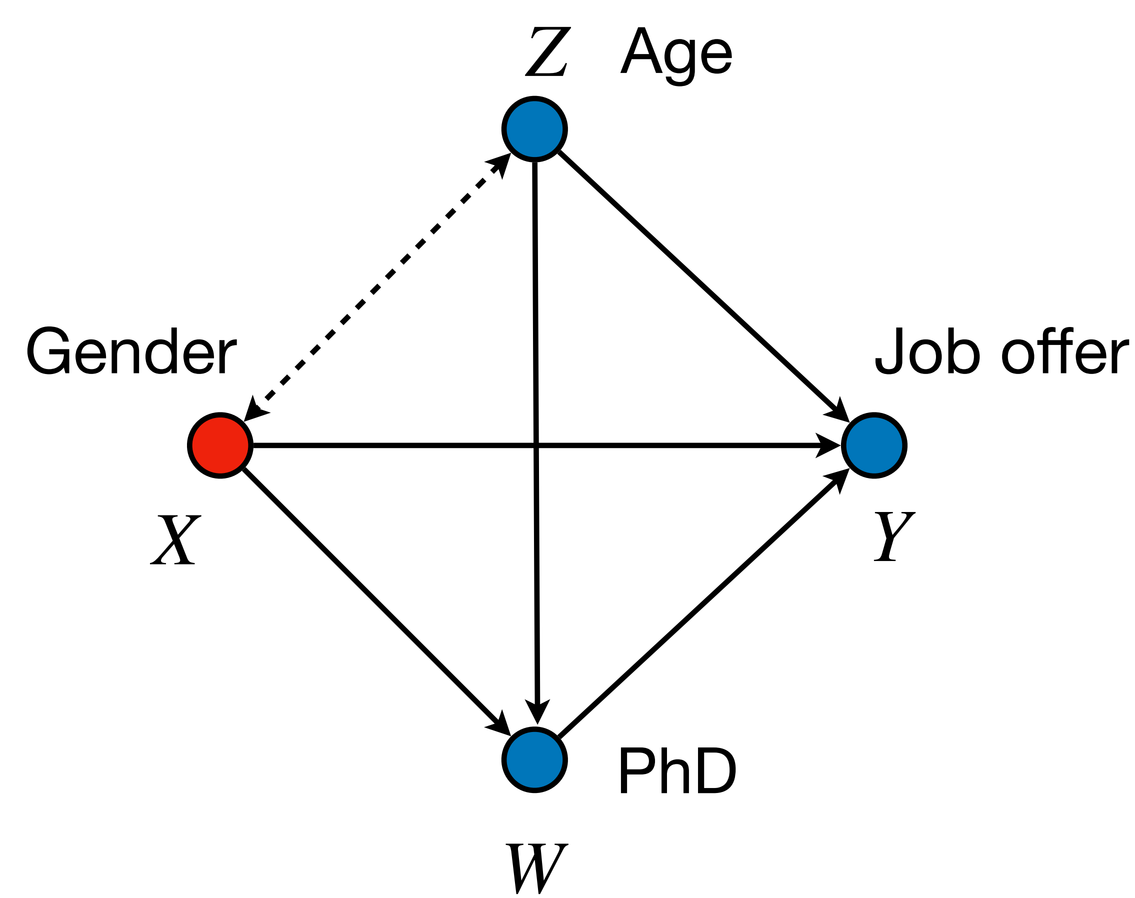
$U \leftarrow N(0,1)$

$X \leftarrow \text{Bernoulli}(expit(U))$

$Z \leftarrow \text{Bernoulli}(expit(U))$

$W \leftarrow \text{Bernoulli}(0.3)$

$Y \leftarrow \text{Bernoulli}(\frac{1}{5}(X + Z - 2XZ) + \frac{1}{6}W)$



Example (Limitation of NDE). A new startup company is currently in hiring season. The hiring decision ($Y \in \{0,1\}$ indicating whether the candidate is hired) is based on gender ($X \in \{0,1\}$, female and male, respectively), age ($Z \in \{0,1\}$, younger and older than 40 years, respectively), and education level ($W \in \{0,1\}$ which indicates whether the applicant has a Ph.D. degree). Following the legal guidelines, the startup is in this case obliged to avoid disparate treatment in hiring.

SCM M*

(unobserved)

$U \leftarrow N(0,1)$

$X \leftarrow \text{Bernoulli}(expit(U))$

$Z \leftarrow \text{Bernoulli}(expit(U))$

$W \leftarrow \text{Bernoulli}(0.3)$

$Y \leftarrow \text{Bernoulli}(\frac{1}{5}(X + Z - 2XZ) + \frac{1}{6}W)$

Admissibility

1) NDE admissible, but $\text{NDE}_{x_0,x_1}(y) = 0$

Power

2) x-DE admissible, and $x\text{-DE}_{x_0,x_1}(y) = 0.036$

Power

3) z-DE admissible, and $z\text{-DE}_{x_0,x_1}(y) = 0.2$

Which of these can be identified from observational data? (new part of FPCFA)

Soundness of the SFM

Theorem. Under the Standard Fairness Model (SFM) the orientation of edges within possibly multidimensional variable sets Z and W does not change any of general, x-specific, or z-specific measures.

That is, if two causal diagrams G_1 and G_2 have the same projection to the Standard Fairness Model, i.e.,

$$\Pi_{\text{SFM}}(G_1) = \Pi_{\text{SFM}}(G_2)$$

Section 4.4
Theorem 4.12

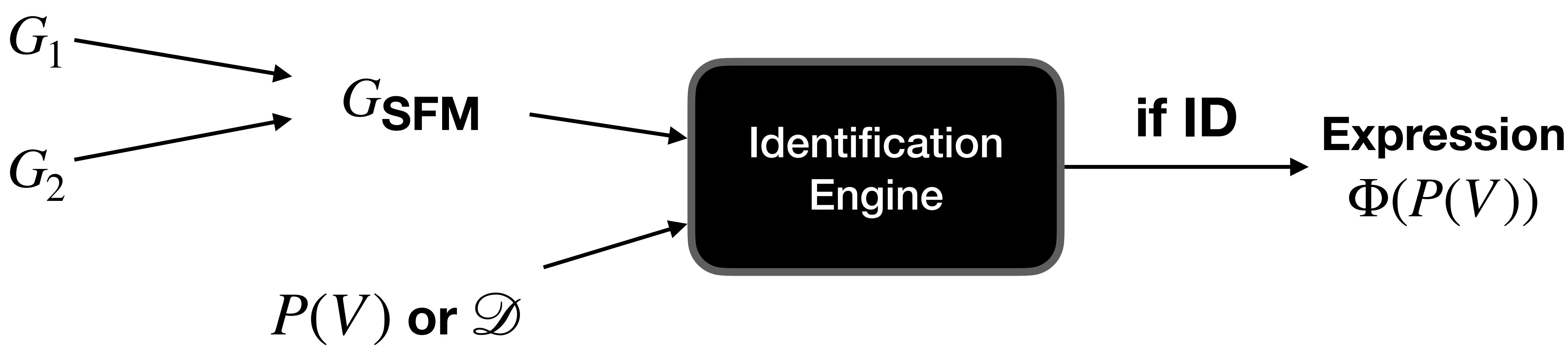
then any measure $M(P(v), G)$ will satisfy

$$M(P(v), \mathcal{G}_1) = M(P(v), \mathcal{G}_2) = M(P(v), \mathcal{G}_{\text{SFM}}),$$

where $M(P(v), \mathcal{G})$ means that the measures are computed based on the observational distribution $P(v)$ and the causal diagram G.

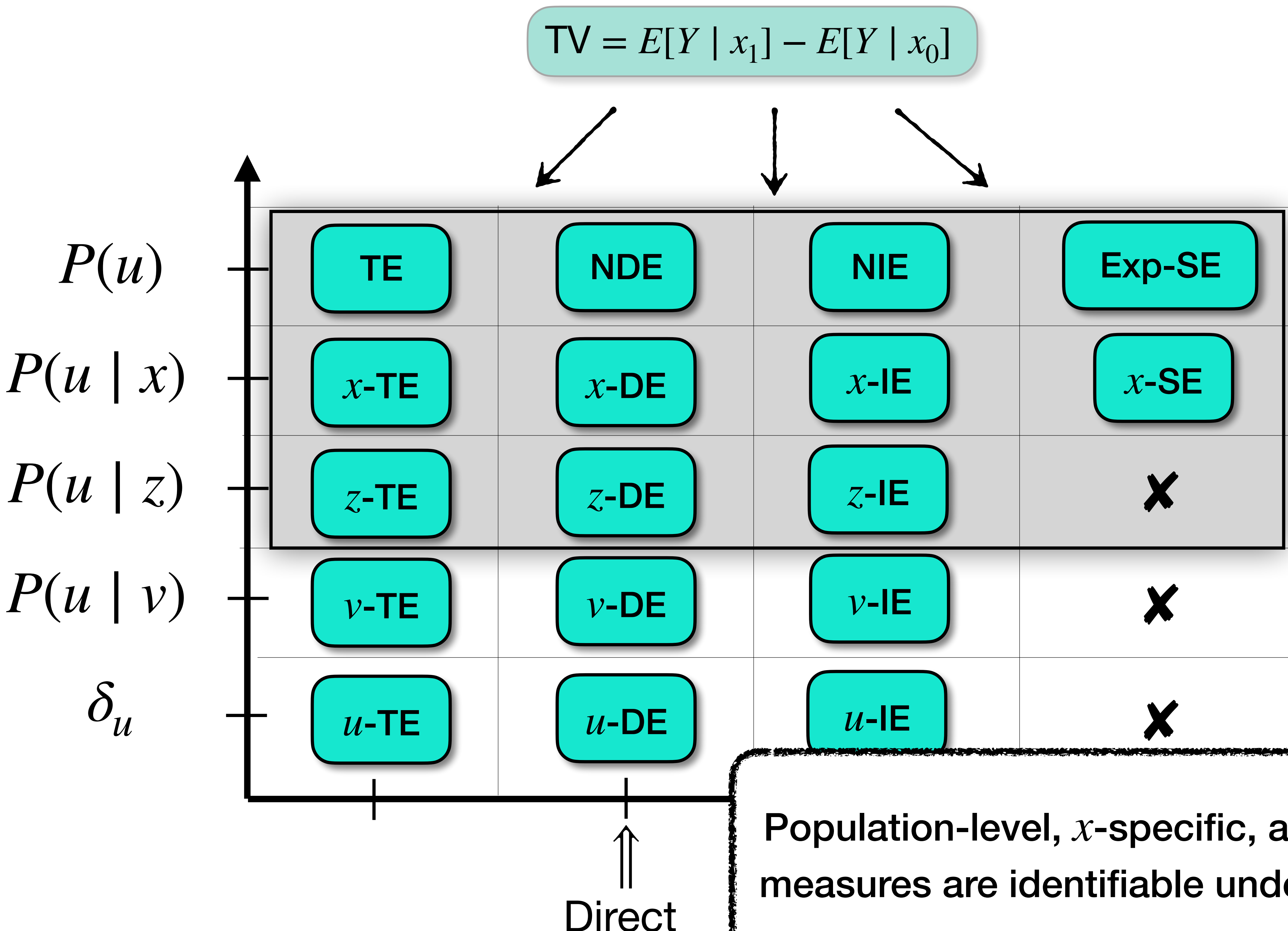
Proof sketch

Since
$\Pi_{\text{SFM}}(G_1) = \Pi_{\text{SFM}}(G_2)$



$\Longrightarrow$ **Any quantity ID just from the SFM is the same for** G_1, G_2

Identifiability in the Fairness Map



Population-level, x-specific, and z-specific measures are identifiable under the G_{SFM}.

SFM's Identification

- In words, identification in our context means that *L_2 and L_3 quantities can be computed using obs. (L_1) data:*

	Measure	ID expression
general	$\text{TE}_{x_0,x_1}(y)$	$\sum_z [P(y \mid x_1, z) - P(y \mid x_0, z)]P(z)$
	$\text{Exp-SE}_x(y)$	$\sum_z P(y \mid x, z)[P(z) - P(z \mid x)]$
	$\text{NDE}_{x_0,x_1}(y)$	$\sum_{z,w} [P(y \mid x_1, z, w) - P(y \mid x_0, z, w)]P(w \mid x_0, z)P(z)$
	$\text{NIE}_{x_0,x_1}(y)$	$\sum_{z,w} P(y \mid x_0, z, w)[P(w \mid x_1, z) - P(w \mid x_0, z)]P(z)$
x-specific	$\text{ETT}_{x_0,x_1}(y \mid x)$	$\sum_z [P(y \mid x_1, z) - P(y \mid x_0, z)]P(z \mid x)$
	$\text{Ctf-SE}_{x_0,x_1}(y)$	$\sum_z P(y \mid x_0, z)[P(z \mid x_0) - P(z \mid x_1)]$
	$\text{Ctf-DE}_{x_0,x_1}(y \mid x)$	$\sum_{z,w} [P(y \mid x_1, z, w) - P(y \mid x_0, z, w)]P(w \mid x_0, z)P(z \mid x)$
	$\text{Ctf-IE}_{x_0,x_1}(y \mid x)$	$\sum_{z,w} P(y \mid x_0, z, w)[P(w \mid x_1, z) - P(w \mid x_0, z)]P(z \mid x)$
z-specific	$z\text{-TE}_{x_0,x_1}(y \mid x)$	$P(y \mid x_1, z) - P(y \mid x_0, z)$
	$z\text{-DE}_{x_0,x_1}(y \mid x)$	$\sum_w [P(y \mid x_1, z, w) - P(y \mid x_0, z, w)]P(w \mid x_0, z)$
	$z\text{-IE}_{x_0,x_1}(y \mid x)$	$\sum_w P(y \mid x_0, z, w)[P(w \mid x_1, z) - P(w \mid x_0, z)]$

Contrasts & Identification - recap

	Measure	C_0	C_1	E_0	E_1
general	TV_{x_0,x_1}	$\emptyset$	$\emptyset$	x_0	x_1
	TE_{x_0,x_1}	x_0	x_1	$\emptyset$	$\emptyset$
	Exp-SE_x	x	x	$\emptyset$	x
	NDE_{x_0,x_1}	x_0	x_1, W_{x_0}	$\emptyset$	$\emptyset$
	NIE_{x_0,x_1}	x_0	x_0, W_{x_1}	$\emptyset$	$\emptyset$
$X = x$	ETT_{x_0,x_1}	x_0	x_1	x	x
	$\text{Ctf-SE}_{x_0,x_1}$	x_0	x_0	x_0	x_1
	$\text{Ctf-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	x	x
	$\text{Ctf-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	x	x
$Z = z$	$z\text{-TE}_{x_0,x_1}$	x_0	x_1	z	z
	$z\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	z	z
	$z\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	z	z
$V' \subseteq V$	$v'\text{-TE}_{x_0,x_1}$	x_0	x_1	v'	v'
	$v'\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	v'	v'
	$v'\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	v'	v'
unit	$\text{unit-TE}_{x_0,x_1}$	x_0	x_1	u	u
	$\text{unit-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	u	u
	$\text{unit-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	u	u

Contrasts that are identifiable under the SFM (without additional assumptions)

Contrasts that are not identifiable under the SFM (without additional assumptions)

Contrasts & Identification - recap

	Measure	C_0	C_1	E_0	E_1
general	TV_{x_0,x_1}	$\emptyset$	$\emptyset$	x_0	x_1
	TE_{x_0,x_1}	x_0	x_1	$\emptyset$	$\emptyset$
	Exp-SE_x	x	x	$\emptyset$	x
	NDE_{x_0,x_1}	x_0	x_1, W_{x_0}	$\emptyset$	$\emptyset$
	NIE_{x_0,x_1}	x_0	x_0, W_{x_1}	$\emptyset$	$\emptyset$
$X = x$	ETT_{x_0,x_1}	x_0	x_1	x	x
	$\text{Ctf-SE}_{x_0,x_1}$	x_0	x_0	x_0	x_1
	$\text{Ctf-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	x	x
	$\text{Ctf-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	x	x
$Z = z$	$z\text{-TE}_{x_0,x_1}$	x_0	x_1	z	z
	$z\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	z	z
	$z\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	z	z
$V' \subseteq V$	$v'\text{-TE}_{x_0,x_1}$	x_0	x_1	v'	v'
	$v'\text{-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	v'	v'
	$v'\text{-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	v'	v'
unit	$\text{unit-TE}_{x_0,x_1}$	x_0	x_1	u	u
	$\text{unit-DE}_{x_0,x_1}$	x_0	x_1, W_{x_0}	u	u
	$\text{unit-IE}_{x_0,x_1}$	x_0	x_0, W_{x_1}	u	u

Contrasts that are

It's understood which contrasts are computable from the data, and which ones are harder.

But how can they be estimated in practice?

identifiable under the SFM (without additional assumptions)

Estimation

Recall from CI1: Inverse Probability Weighting (IPW) Derivation

Holds true for the SFM!

- If Z is a back-door set for X, Y, then

$$P(\mathbf{y}_x) = \sum_{\mathbf{z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z})$$

$$= \sum_{\mathbf{z}} \frac{P(\mathbf{y}, \mathbf{x}, \mathbf{z})}{P(\mathbf{x}, \mathbf{z})} P(\mathbf{z})$$

$$= \sum_{\mathbf{z}} \frac{P(\mathbf{y}, \mathbf{x}, \mathbf{z})}{P(\mathbf{x} \mid \mathbf{z}) P(\mathbf{z})} P(\mathbf{z})$$

$$= \sum_{\mathbf{z}} \frac{P(\mathbf{y}, \mathbf{x}, \mathbf{z})}{P(\mathbf{x} \mid \mathbf{z})}$$

Entries of the joint distribution

Fit a function $g(z)$ that approximates this probability

Recall from CI1: Inverse Probability Weighting (IPW) Derivation

- Assuming we have N samples, we can compute

$$P(\mathbf{y}_x) = \sum_{\mathbf{z}} \frac{P(\mathbf{y}, \mathbf{x}, \mathbf{z})}{P(\mathbf{x} \mid \mathbf{z})}$$

$$= \sum_{\mathbf{z}} \frac{\frac{1}{N} \sum_{i=1}^{N} 1_{\mathbf{Y}_i=\mathbf{y}, \mathbf{X}_i=\mathbf{x}, \mathbf{Z}_i=\mathbf{z}}}{g(\mathbf{z})}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{\mathbf{z}} \frac{1_{\mathbf{Y}_i=\mathbf{y}, \mathbf{X}_i=\mathbf{x}, \mathbf{Z}_i=\mathbf{z}}}{g(\mathbf{z})}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \frac{1_{\mathbf{Y}_i=\mathbf{y}, \mathbf{X}_i=\mathbf{x}, \mathbf{Z}_i=\mathbf{z}}}{g(\mathbf{z})}$$

→ Requires time proportional to the number of samples N

Inverse Probability Weighting (IPW)

- Thus, a typical way to compute $E[y_x]$ is to use inverse propensity weighting (IPW) and an estimator of the form

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1(X_i = x)Y_i}{\hat{p}(X_i \mid Z_i)}.$$

- The assumptions that we need (on top of the SFM)

Assumption (Positivity). The positivity assumption holds if $\forall\ x, z$**,** $P(X = x \mid Z = z)$ **is bounded away from** 0**, that is**

$$\delta < P(X = x \mid Z = z) < 1 - \delta,$$

for some $\delta > 0$**.**

Beyond IPW

$\Longrightarrow$ Double Machine Learning

IPW term:

$$\frac{1(X_i = x)Y_i}{\widehat{p}(X_i \mid Z_i)}$$

$<<$

DML term:

$$\frac{1(X_i = x)(Y_i - \widehat{\mu}(Y_i \mid Z_i, X_i))}{\widehat{p}(X_i \mid Z_i)} + \widehat{\mu}(Y_i \mid Z_i, X_i)$$

+ sample splitting!

Only for $E[y_x]$

Need also $E[y_{x',W_x}]$

For $E[y_{x',W_x}]$:

Mediation DML term:

$$\frac{1(X = x_1)p_{x_0}(Z, W)}{p_{x_1}(Z, W)p_{x_0}(Z)}[Y - \mu(x_1, W, Z)]$$

$$+\frac{1(X = x_0)}{p_{x_0}(Z)}[\mu(x_1, W, Z) - E[\mu(x_1, W, Z) \mid X = x_0, Z]]$$

$$+E[\mu(x_1, W, Z) \mid X = x_0, Z]$$

Relationship to previous literature

- How does the presented framework of Causal Fairness Analysis relate to previous literature?
- In particular, we discuss

(i) Counterfactual Fairness (Kusner et. al., 2017)

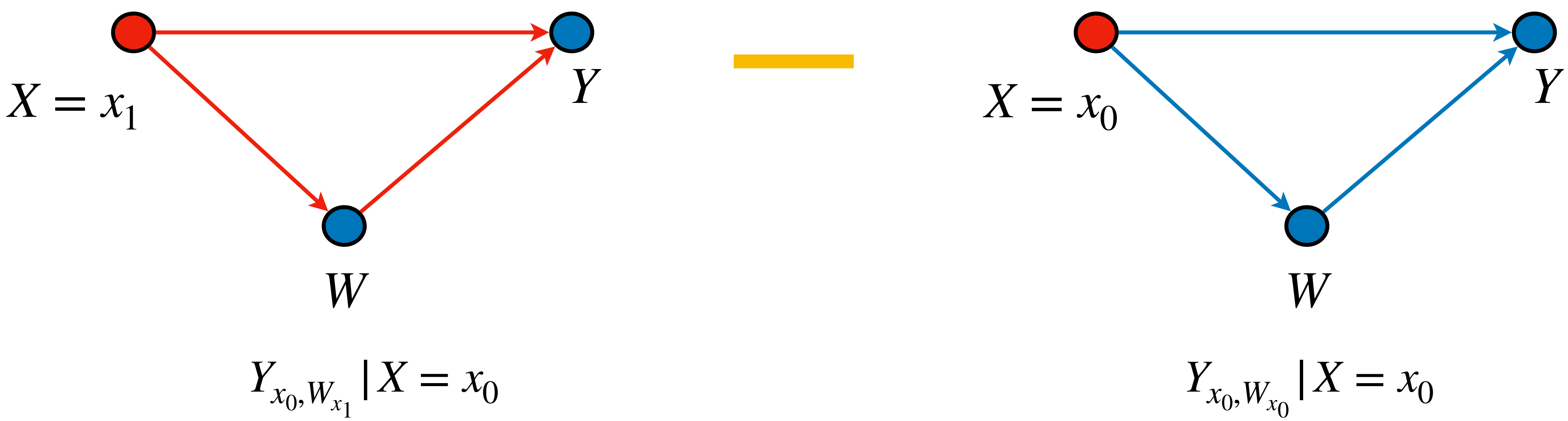
(ii) Individual Fairness (Dwork et. al., 2012)

(iii) Predictive Parity (Chouldechova, 2017)

Counterfactual Fairness

Definition (Counterfactual Fairness, Kusner et. al., 2017).
An outcome Y is said to be counterfactually fair if and only if

$$P(y_x(u) \mid X = x, W = w) = P(y_{x'}(u) \mid X = x, W = w), \quad \forall x, x', w \, .$$



Note: if the u is fixed, there are no probabilistic statements involved.

Note: if the u is not fixed, averaging over posterior $P(u \mid X = x, W = w)$.

Counterfactual Fairness

Definition (Counterfactual Fairness, Kusner et. al., 2017).
An outcome Y is said to be counterfactually fair if and only if

$$P(y_x(u) \mid X = x, W = w) = P(y_{x'}(u) \mid X = x, W = w), \ \ \forall x, x', w .$$

$Y_{x_0, W_{x_1}} \mid X = x_0$ $\qquad$ $Y_{x_0, W_{x_0}} \mid X = x_0$

Intuition: granular measure of <u>total</u> effect.

Counterfactual Fairness

Definition (Counterfactual Fairness, Kusner et. al., 2017).
An outcome Y is said to be *counterfactually fair* if and only if

$$P(y_x(u) \mid X = x, W = w) = P(y_{x'}(u) \mid X = x, W = w), \;\; \forall x, x', w \,.$$

the paper leaves space for ambiguity in interpretation

unit-level

$$y_x(u) - y_{x'}(u) = 0, \;\; \forall x, x', u \in \mathcal{U}$$

consistent with authors' claim:
"*emphasize that counterfactual fairness is an individual-level definition, which is substantially different from comparing different individuals that happen to share the same "treatment" $X = x$ and coincide on the values of $W = w$*"

across units

$$P(y_x \mid X = x, W = w) = P(y_{x'} \mid X = x, W = w)$$

also consistent with authors' claim:
"*the distribution over possible predictions for an individual should remain unchanged in a world where an individual's protected attributes had been different*"

Counterfactual Fairness

Definition (Counterfactual Fairness, Kusner et. al., 2017).
An outcome Y is said to be *counterfactually fair* if and only if

$$P(y_x(u) \mid X = x, W = w) = P(y_{x'}(u) \mid X = x, W = w), \;\; \forall x, x', w \, .$$

the paper leaves space for ambiguity in interpretation

$y_x(u)$ …

"empha…
individua…
differen…
that hap…
$X = x$ and coincide on the values of $W = w$"

…$, W = w)$

…s for
… in a
…tributes
had been different"

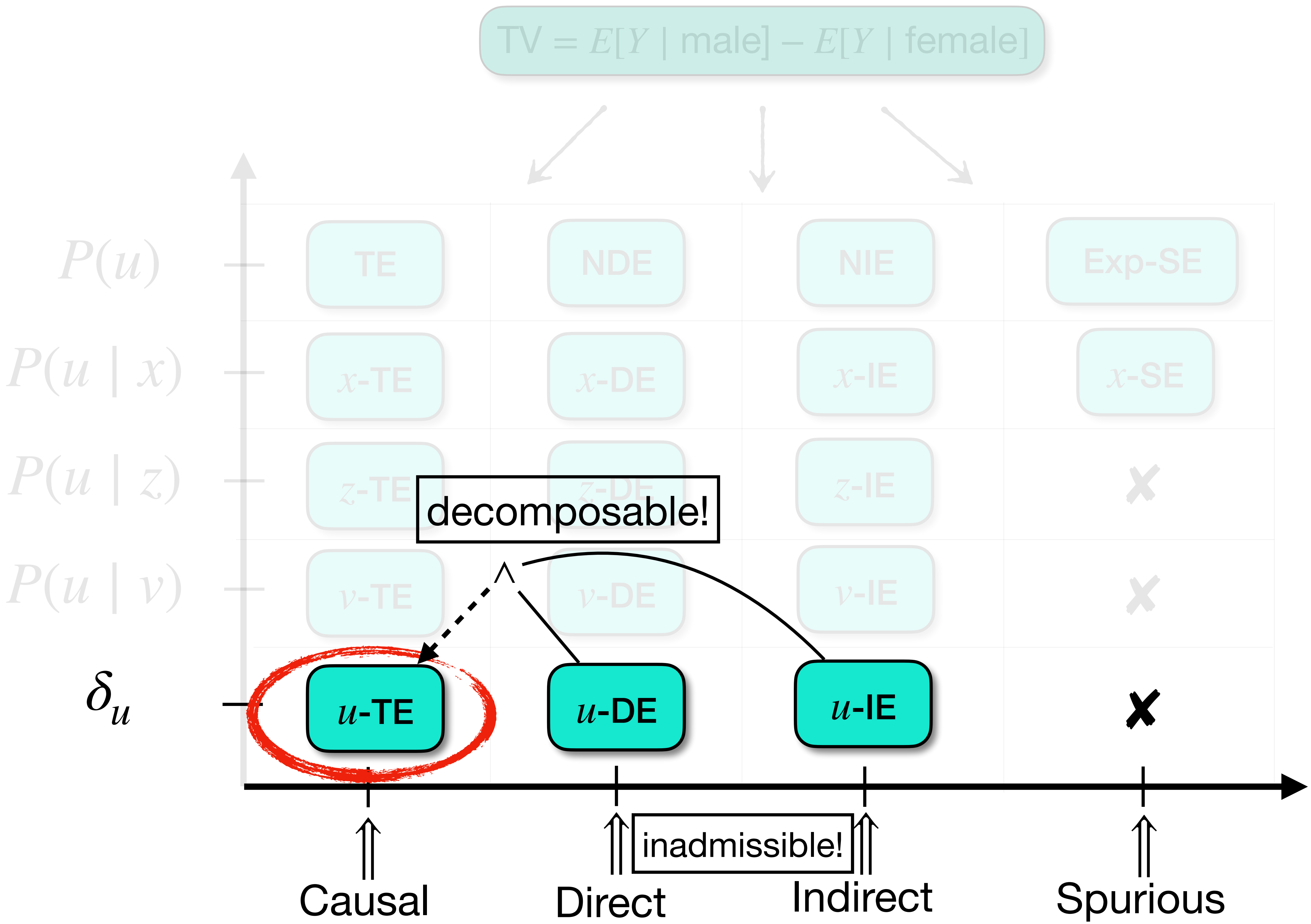
Luckily, both of these measures are covered by the Fairness Map!

Counterfactual fairness (Kusner et. al., 2017)

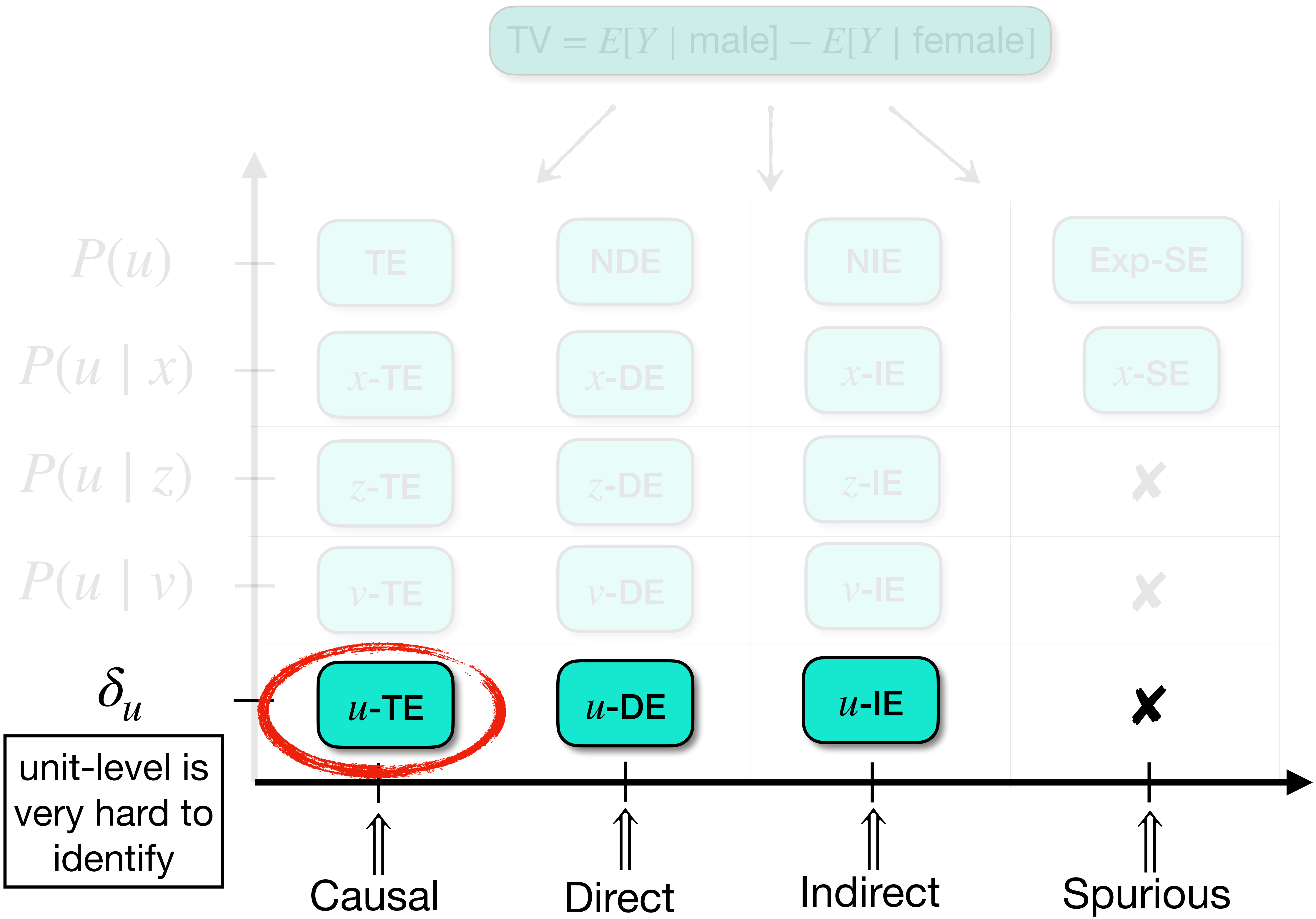
TV = $E[Y \mid \text{male}] - E[Y \mid \text{female}]$

	Causal	Direct	Indirect	Spurious
$P(u)$	TE	NDE	NIE	Exp-SE
$P(u \mid x)$	x-TE	x-DE	x-IE	x-SE
$P(u \mid z)$	z-TE	z-DE	z-IE	✘
$P(u \mid v)$	v-TE	v-DE	v-IE	✘
δ_u	u-TE	u-DE	u-IE	✘

Counterfactual fairness (Kusner et. al., 2017)



Counterfactual fairness (Kusner et. al., 2017)



Ctf-fair, Issue 1: Inadmissibility

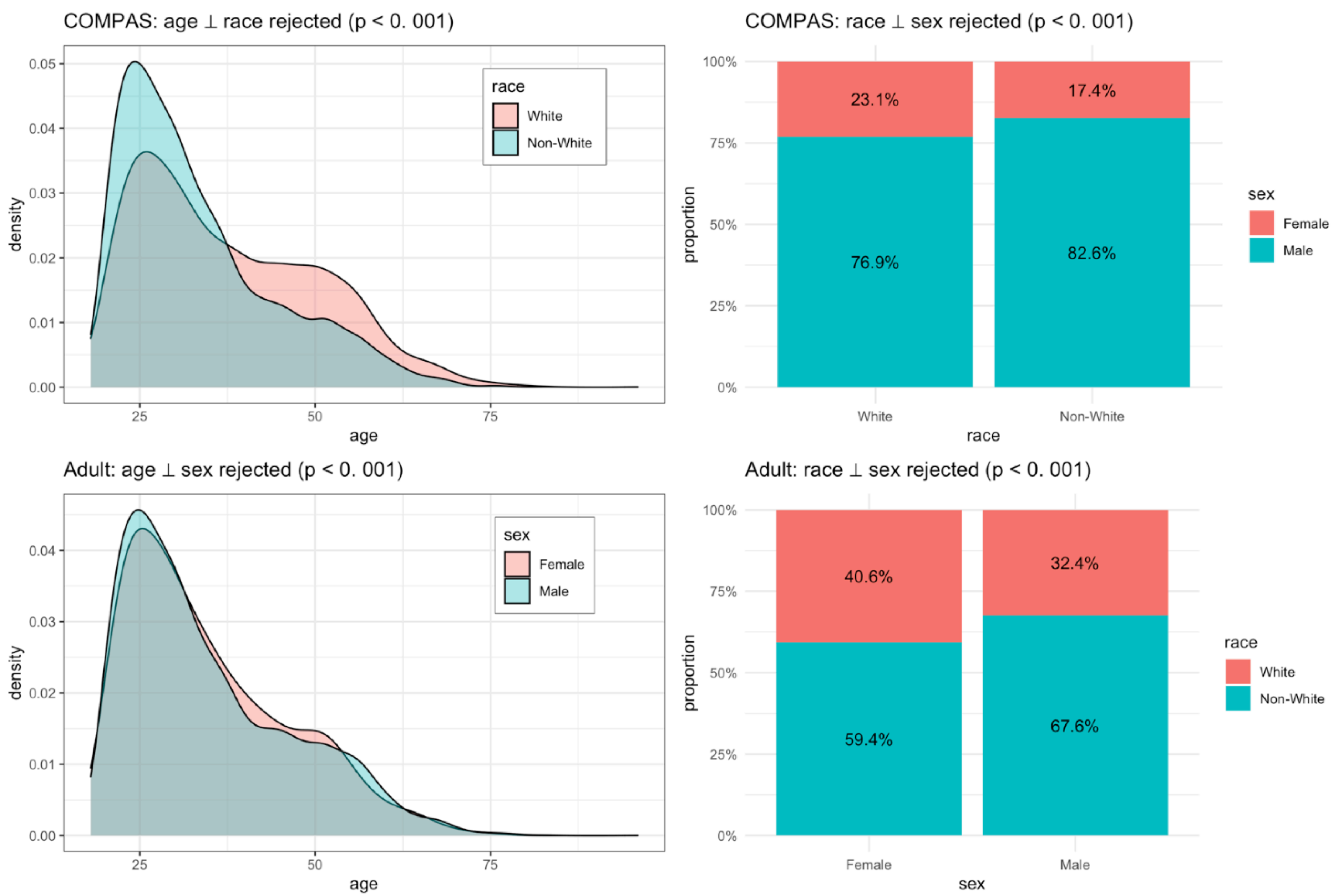
Proposition. The unit-level total effect (unit-TE$_{x_0,x_1}(y)$) and the (x, w)-specific total effect ((x, w)-TE$_{x_0,x_1}(y \mid x, w)$) are not admissible w.r.t. the structural direct, indirect, and spurious criteria. Formally, we write

$$\text{Str-DE-fair} \not\Longrightarrow \text{unit-TE-fair}, \quad \text{Str-DE-fair} \not\Longrightarrow (x, w)\text{-TE-fair}$$
$$\text{Str-IE-fair} \not\Longrightarrow \text{unit-TE-fair}, \quad \text{Str-IE-fair} \not\Longrightarrow (x, w)\text{-TE-fair}$$
$$\text{Str-SE-fair} \not\Longrightarrow \text{unit-TE-fair}, \quad \text{Str-SE-fair} \not\Longrightarrow (x, w)\text{-TE-fair} .$$

Counterfactual Fairness is inadmissible, therefore not suitable to reason about direct, indirect, or spurious effects.

Ctf-fair, Issue 2: Spurious Effects

Assumption: ancestral closure of set X.



redlining

religious segregation

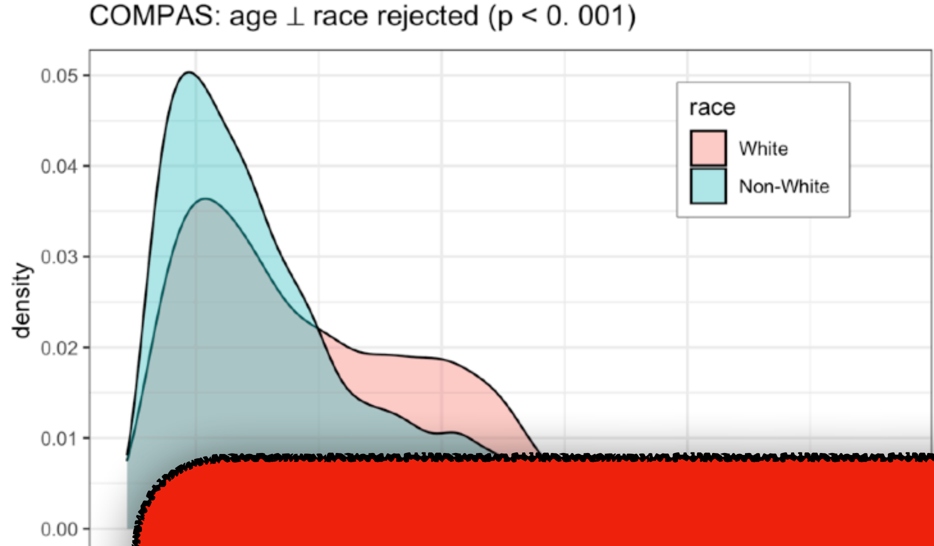
rural/urban balance of genders in China

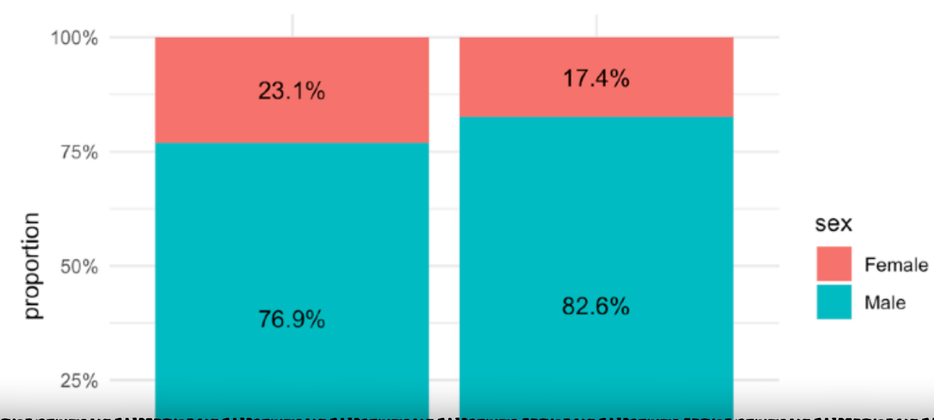
Ctf-fair, Issue 2: Spurious Effects

Assumption: ancestral closure of set X.

However, is this a realistic assumption?

Vignette Time!



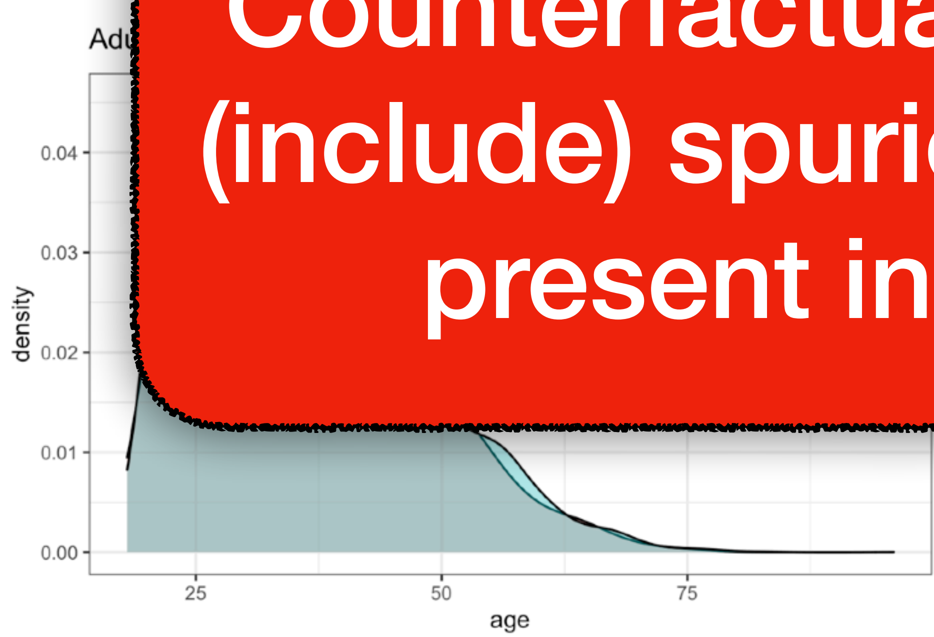


redlining

religious segregation

ce

na



Counterfactual Fairness does not account (include) spurious variations, which may be present in some practical settings.

Ctf-fair, Issue 3: Identifiability

Proposition. Suppose that $\mathscr{M}$ is a Markovian model and that $\mathscr{G}$ is the associated causal diagram. Assume that the set of mediators between X and Y is non-empty, $W \neq \emptyset$. Then, the measures unit-TE$_{x_0,x_1}(y)$ and (x, w)-TE$_{x_0,x_1}(y \mid x, w)$ are not identifiable from observational data, even if the fully specified diagram $\mathscr{G}$ is known.

Counterfactual Fairness requires strong assumptions for identification.

Ctf-fair, Issue 3: Identifiability (Example)

Example. The startup company from our previous example has closed the hiring season. In the hiring process, the company achieved demographic parity, which means in this context that 50% of new hires were female. Now, the company needs to decide on each employee's salary. In order to be "fair", each employee is evaluated on how well they perform their tasks. The salary Y is then determined based on this information, but, due to a possibly subconscious bias of the executives while determining employees' salaries, gender may also affects how salaries are determined.

SCM $\langle \mathcal{F}_1, P_1(U) \rangle$

$$X \leftarrow U_X$$
$$W \leftarrow -X + U_W$$
$$Y \leftarrow X + W + U_Y\,.$$
$$U_X \in \{0,1\}, P(U_X = 1) = 0.5,$$
$$U_W, U_Y \sim N(0,1)\,.$$

SCM $\langle \mathcal{F}_2, P_2(U) \rangle$

$$X \leftarrow U_X$$
$$W \leftarrow -X + (-1)^X U_W$$
$$Y \leftarrow X + W + U_Y\,.$$
$$U_X \in \{0,1\}, P(U_X = 1) = 0.5,$$
$$U_W, U_Y \sim N(0,1)\,.$$

Ctf-fair, Issue 3: Identifiability (Example)

SCM $\langle \mathcal{F}_1, P_1(U) \rangle$

$$X \leftarrow U_X$$
$$W \leftarrow -X + U_W$$
$$Y \leftarrow X + W + U_Y .$$
$$U_X \in \{0,1\}, P(U_X = 1) = 0.5,$$
$$U_W, U_Y \sim N(0,1) .$$

$$y_{x_1}(u) - y_{x_0}(u) = \underbrace{(1 + (-1 + u_w) + u_y)}_{y_{x_1}(u)} - \underbrace{(0 + (-0 + u_w) + u_y)}_{y_{x_0}(u)} = 0.$$

same graph $\mathcal{G}$ + same observational distribution $P(V)$

SCM $\langle \mathcal{F}_2, P_2(U) \rangle$

$$X \leftarrow U_X$$
$$W \leftarrow -X + (-1)^X U_W$$
$$Y \leftarrow X + W + U_Y .$$
$$U_X \in \{0,1\}, P(U_X = 1) = 0.5,$$
$$U_W, U_Y \sim N(0,1) .$$

$$y_{x_1}(u) - y_{x_0}(u) = \underbrace{(1 + (-1 - u_w) + u_y)}_{y_{x_1}(u)} - \underbrace{(0 + (-0 + u_w) + u_y)}_{y_{x_0}(u)} = -2u_w \neq 0$$

whenever $u_w \neq 0$.

Counterfactual Fairness Summary

In summary, counterfactual fairness is:

- decomposable & inadmissible (w.r.t DE, IE, SE),

- not identifiable in general, and

- oblivious to spurious effects (and corresponding business necessity requirements).

Relationship to previous literature

- How does the presented framework of Causal Fairness Analysis relate to previous literature?
- In particular, we discuss

(i) Counterfactual Fairness (Kusner et. al., 2017)

(ii) Individual Fairness (Dwork et. al., 2012)

(iii) Predictive Parity (Chouldechova, 2017)

Individual Fairness

Individual Level

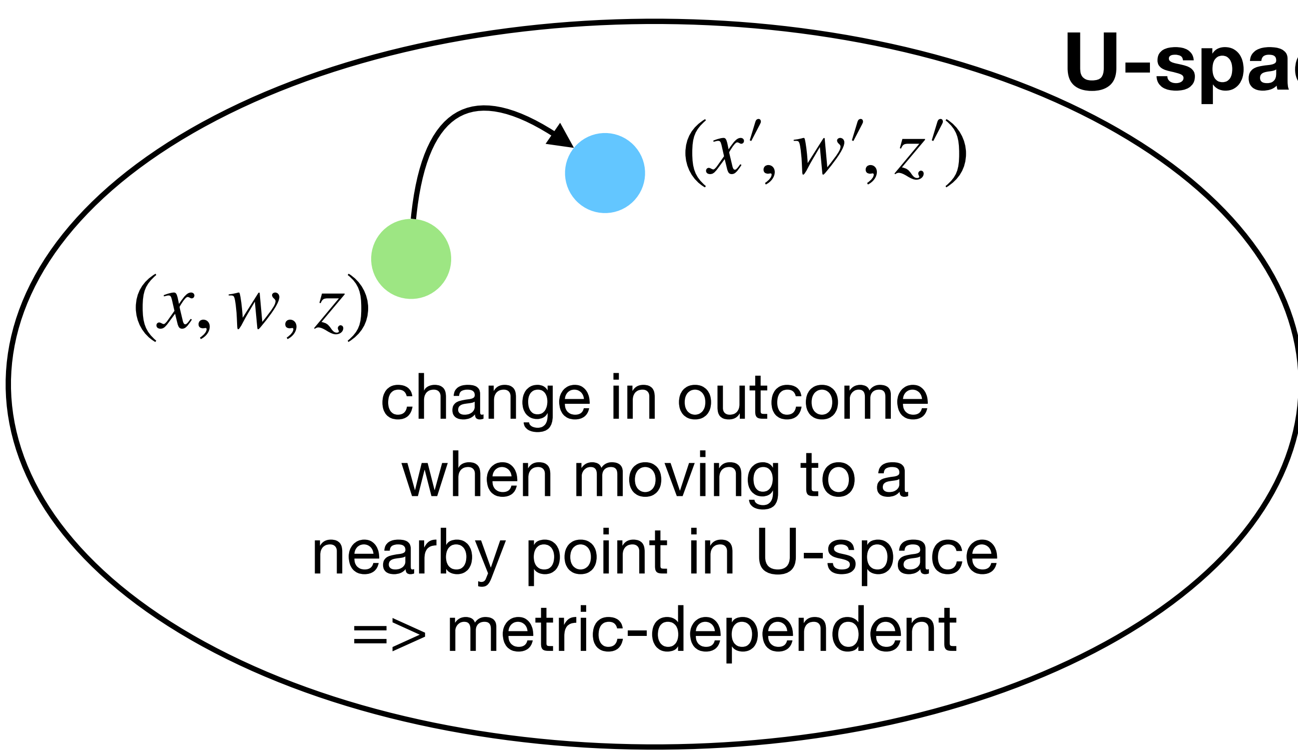
Definition (Individual Fairness, Dwork et. al., 2012).
Let d be a fairness metric on $\mathscr{X} \times \mathscr{Z} \times \mathscr{W}$. An outcome Y is said to satisfy individual fairness if

$$|P(y \mid x, z, w) - P(y \mid x', z', w')| \leq d((x, z, w), (x', z', w')) \quad \forall\ x, x', w, w', z, z'$$

Intuition: individuals similar w.r.t d should have similar outcomes.

we call this ***IF condition***

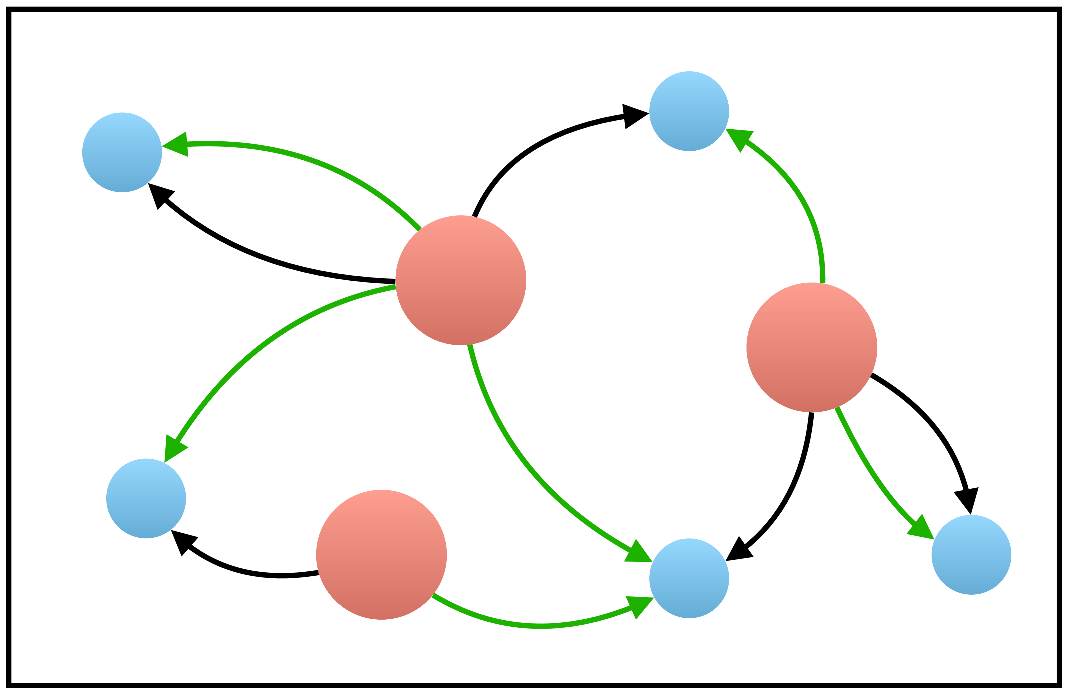
U-space



$$d((x, w, z), (x', w', z')) \text{ small}$$
$$\Longrightarrow$$
$$P(y \mid x, z, w) - P(y \mid x', z', w') \text{ small}$$

Quick Detour: Optimal Transport

- What is optimal transport?



piles of rubble

empty pits

Monge (1781): how do we optimally transport the rubble into the pits?

- How do we define OT formally?

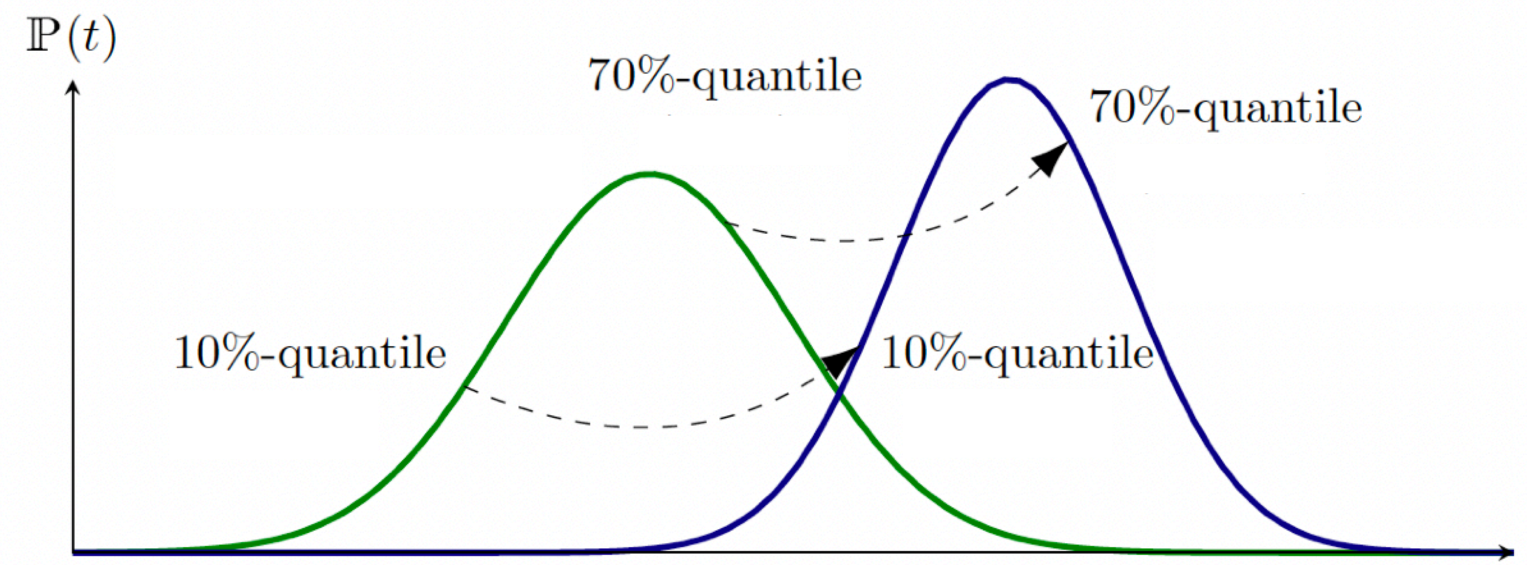
Given a measure μ over X and ν over Y the optimal transport problem is given by

$$\min \int_{X \times Y} c(x, y) d\pi(x, y)$$

where $c(x, y)$ is the cost function (L_1, L_2) and π a transport plan with marginals μ, ν .

Quick Detour: Optimal Transport

- What do optimal transport plans look like?



- In general, dimension $d > 1$, OT plans are not easy to find!

Summary:
Optimal Transport gives an intuitive way of measuring a distance between distributions which has been shown as useful in many sciences (mathematics, physics, statistics, etc.)

Individual Fairness: Local to Global

Proposition (OT bounds TV, Dwork et. al., 2012).

Global Level

Suppose that the IF condition holds. Let the optimal transport cost between $Z, W \mid X = x_1$ and $Z, W \mid X = x_0$ be denoted by $\text{OTC}^d_{x_0,x_1}((Z, W))$. Then, it holds that

$$|\text{TV}_{x_0,x_1}(y)| \leq C_d * \text{OTC}^d_{x_0,x_1}((Z, W)) .$$

(1) IF criterion

(2) Small $\text{OTC}^d_{x_0,x_1}((Z, W))$

$\Longrightarrow$

Small TV ✓

DE ?

IE ?

SE ?

Local to Global: Intuition

IF condition + OTC $Z, W \mid x_1 \rightarrow Z, W \mid x_0$ small = TV small

take $(x_1, z, w), (x_0, z, w)$

IF condition yields

$P(y \mid x_1, z, w) - P(y \mid x_0, z, w) = 0.$

i.e., observational direct effect is 0.

distribution of attributes Z, W same for x_0, x_1 groups

small disparity

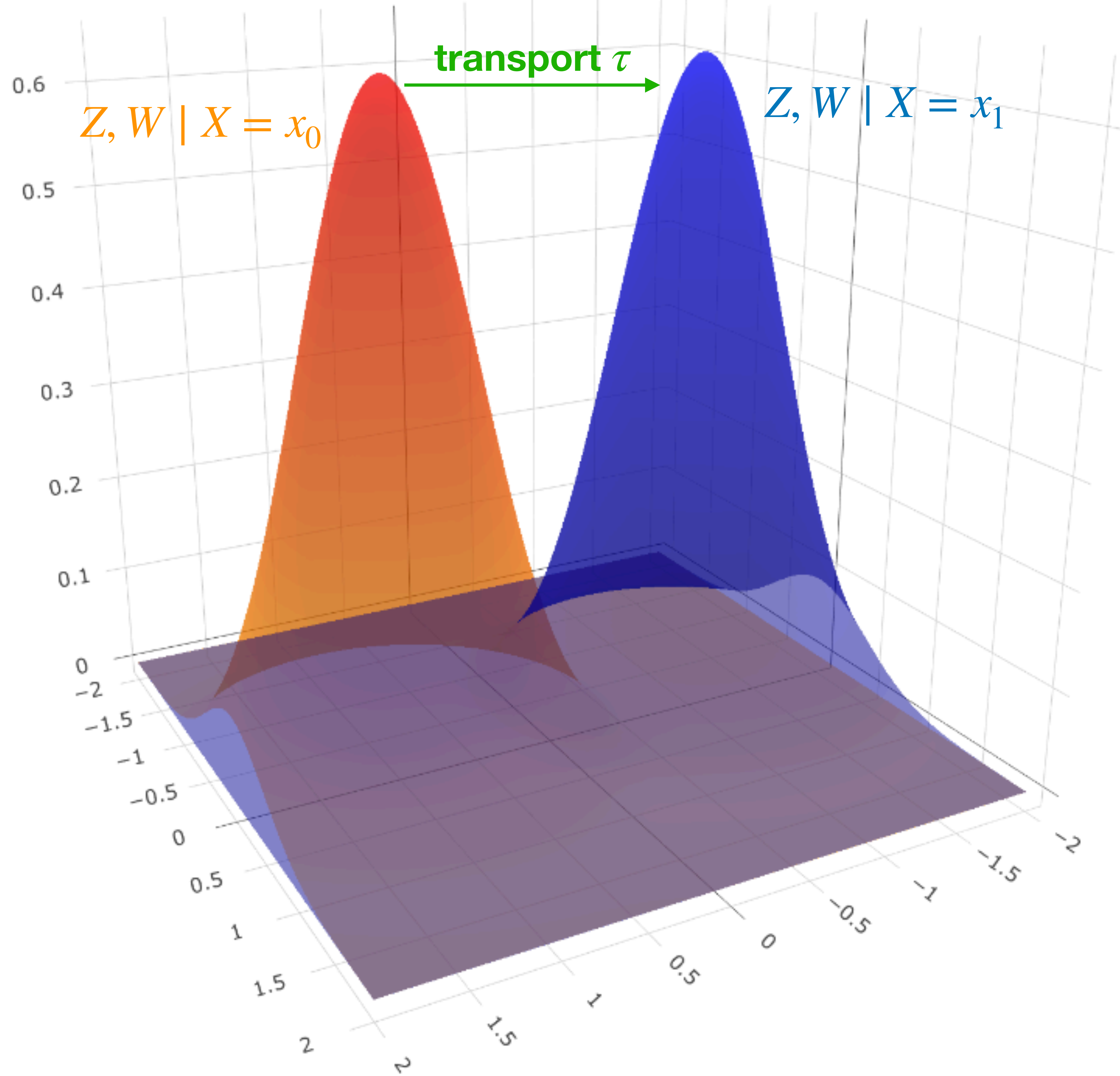
$$P(y \mid x_1) = \sum_{z,w} P(y \mid x_1, z, w) P(z, w \mid x_1)$$

$$= \sum_{z_w} P(y \mid x_0, z, w) P(z, w \mid x_1) = \sum_{z,w} P(y \mid x_0, z, w) P(z, w \mid x_0) = P(y \mid x_0)$$

IF condition

OTC = 0

i.e., TV = 0


transport τ
$Z, W \mid X = x_0$
$Z, W \mid X = x_1$
0.6
0.5
0.4
0.3
0.2
0.1
0
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2

Individual Fairness
(Dwork. et. al., 2012)

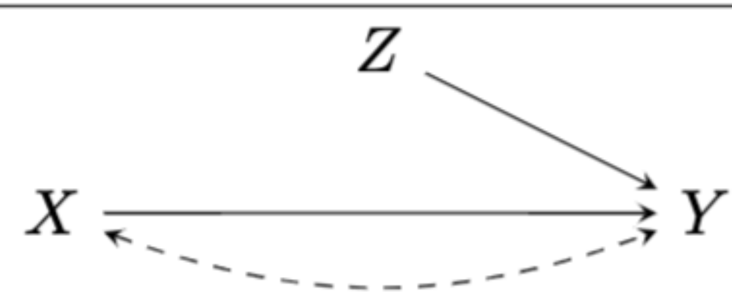
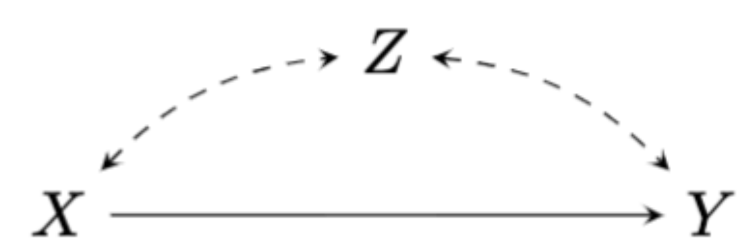
Causal Fairness Analysis implications on IF:

- IF is oblivious to the underlying causal mechanisms.
- IF captures the direct effect only under the SFM.
- IF with a sparse metric d is not admissible.
- IF with a complete metric d doesn't account for business necessity.

Section 4.5.2

IF, Issue 1: Ignoring Causal Structure

Examples.

		A	B
SCM $\mathcal{M}$	$\mathcal{F}$	$X \leftarrow U_{XY}$ $Z \leftarrow U_Z$ $Y \leftarrow X - U_{XY} + Z + U_Y$	$X \leftarrow U_{XZ}$ $Z \leftarrow U_{XZ} + U_{ZY}$ $Y \leftarrow U_{ZY} + U_Y$
	$P(u)$	$U_{XY} \sim \text{Bernoulli}(0.5)$, $U_Z, U_Y \sim N(0,1)$	$U_{XZ} \sim \text{Bernoulli}(0.5)$, $U_{ZY}, U_Y \sim N(0,1)$
diagram	$\mathcal{G}$		

metric

$$d((x, z), (x', z')) = |z - z'|$$

IF, Issue 1: Insensitive to the Causal Structure

Example A: We can compute that

$$
\begin{aligned}
E^{\mathscr{M}_A}[y \mid x, z] &= E^{\mathscr{M}_A}[X - U_{XY} + Z + U_Y \mid x, z] \\
&= \underbrace{E^{\mathscr{M}_A}[X - U_{XY} \mid x, z]}_{=0 \text{ as } X = U_{XY}} + E^{\mathscr{M}_A}[Z \mid x, z] + \underbrace{E^{\mathscr{M}_A}[U_Y \mid x, z]}_{=0} \\
&= z \,.
\end{aligned}
$$

IF holds, but direct effect still exists

$$
\implies \left| E^{\mathscr{M}_A}[y \mid x_1, z] - E^{\mathscr{M}_A}[y \mid x_0, z'] \right| = |z - z'|
$$

Example B: We can compute that

$$
\begin{aligned}
E^{\mathscr{M}_B}[y \mid x, z] &= E^{\mathscr{M}_B}[U_{ZY} + U_Y \mid x, z] \\
&= E^{\mathscr{M}_B}[Z - U_{XZ} \mid x, z] + \underbrace{E^{\mathscr{M}_B}[U_Y \mid x, z]}_{=0} \\
&= E^{\mathscr{M}_B}[Z - X \mid x, z] = z - x \,.
\end{aligned}
$$

IF does not hold, but direct effect does not exist

$$
\implies \left| E^{\mathscr{M}_B}[y \mid x_1, z] - E^{\mathscr{M}_B}[y \mid x_0, z'] \right| = |z - 1 - z'|
$$

IF, Issue 1: Insensitive to the Causal Structure

Example A: We can compute that

$$E^{\mathscr{M}_A}[y \mid x, z] = E^{\mathscr{M}_A}[X - U_{XY} + Z + U_Y \mid x, z]$$

$$= \underbrace{E^{\mathscr{M}_A}[X - U_{XY} \mid x, z]}_{=0 \text{ as } X=U_{XY}} + E^{\mathscr{M}_A}[Z \mid x, z] + \underbrace{E^{\mathscr{M}_A}[U_Y \mid x, z]}_{=0}$$

$$= z\,.$$

IF holds, but direct effect still exists

IF is oblivious to the underlying causal structure, which translates in lack of both necessity and sufficiency w.r.t. DE.

Exam

$E^{\mathscr{M}_B}[$

$$= E^{\mathscr{M}_B}[Z - U_{XZ} \mid x, z] + \underbrace{E^{\mathscr{M}_B}[U_Y \mid x, z]}_{=0}$$

IF does not hold, but direct effect does not exist

$$= E^{\mathscr{M}_B}[Z - X \mid x, z] = z - x\,.$$

$$\Longrightarrow \left| E^{\mathscr{M}_B}[y \mid x_1, z] - E^{\mathscr{M}_B}[y \mid x_0, z'] \right| = |z - 1 - z'|$$

IF, Issue 2: Direct Effect (under suitable assumptions)

Proposition. Suppose that the metric d does not depend on the X variable, that is,

$$d((x, z, w), (x', z', w')) = d((z, w), (z', w')) \, .$$

Then, under the assumptions of the Standard Fairness Model, the IF criterion implies that Ctf-DE equals 0, that is

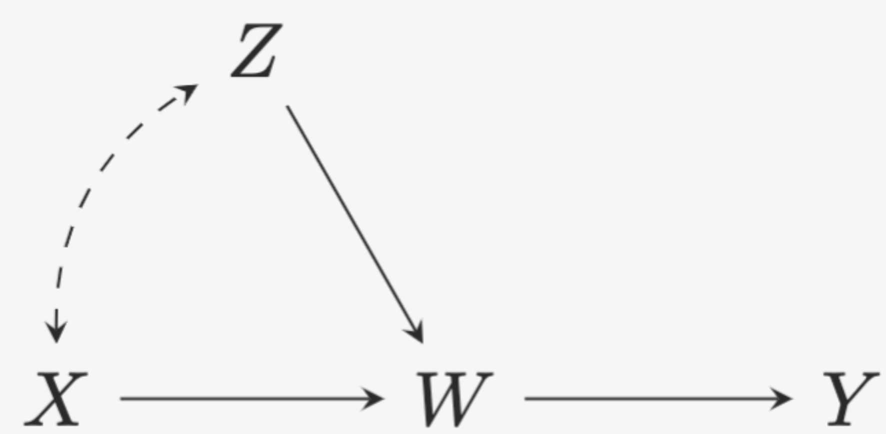
$$\text{IF} \implies \text{Ctf-DE}_{x_0,x_1}(y \mid x) = 0.$$

IF captures the direct effect - but under the assumptions entailed by the SFM

IF, Issue 3: Sparse metrics d suffer from decomposability issue

Example.

$$\mathcal{F}^*, P^*(U) := \begin{cases} X \leftarrow U_{XZ} \\ Z \leftarrow -U_{XZ} + U_Z \\ W \leftarrow X + Z + U_W \\ Y \leftarrow 1(U_Y < \textbf{expit}(W)), \\ \\ U_{XZ} \in \{0,1\}, P(U_{XZ} = 1) = 0.5, \\ U_Z, U_W, U_Y \sim \textbf{Unif}[0,1], \end{cases}$$



metric $d((x, z, w), (x', z', w')) = |w - w'|$.

We can compute that $|P(y \mid x, z, w) - P(y \mid x', z', w')| = |\text{expit}(w) - \text{expit}(w')|$

$$\leq \frac{1}{4}|w - w'| \implies \text{IF holds!}$$

However:

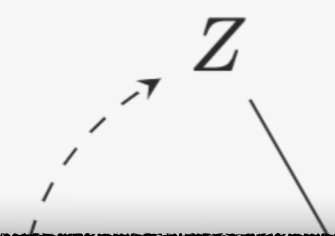
$$\text{TV}_{x_0,x_1}(y) = x\text{-DE}_{x_0,x_1}(y \mid x_0) - x\text{-IE}_{x_1,x_0}(y \mid x_0) - x\text{-SE}_{x_1,x_0}(y)$$

$$= \underbrace{(0\%)}_{direct} - \underbrace{(14\%)}_{indirect} - \underbrace{(-14\%)}_{spurious},$$

IF, Issue 3: Sparse metrics d suffer from decomposability issue

Example.

$$\begin{cases} X \leftarrow U_{XZ} \\ Z \leftarrow -U_{XZ} + U_Z \\ W \leftarrow X + Z + U_W \end{cases}$$



$\mathscr{F}^*, F$ ... $w'|$.

We can ...

IF can be decomposed whenever the metric d is sparse (complete metrics d addressed later)

4

However:

$$\begin{aligned} \text{TV}_{x_0,x_1}(y) &= x\text{-DE}_{x_0,x_1}(y \mid x_0) - x\text{-IE}_{x_1,x_0}(y \mid x_0) - x\text{-SE}_{x_1,x_0}(y) \\ &= \underbrace{(0\%)}_{direct} - \underbrace{(14\%)}_{indirect} - \underbrace{(-14\%)}_{spurious}, \end{aligned}$$

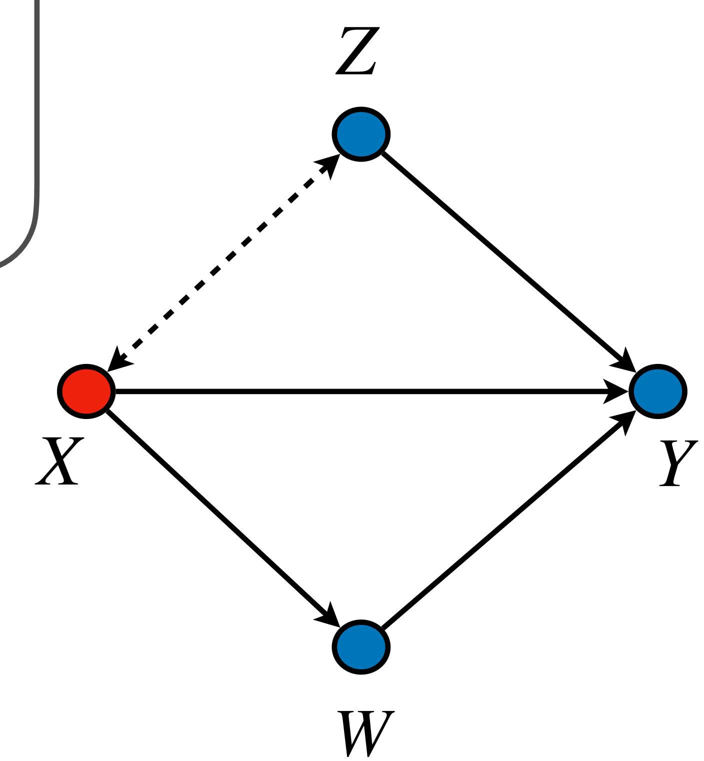
IF, Issue 4: complete metric d does not allow for business necessity

Part I. If $d((x, z, w), (x', z', w')) = \|z - z'\| + \|w - w'\|$

then $\text{OTC}^{d}_{x_0,x_1}((Z, W)) = 0 \implies X \perp\!\!\!\perp Z, W.$

Part II. If IF condition holds, then for Y binary

$$X \perp\!\!\!\perp Y \mid Z, W.$$



Part III (I + II). $X \perp\!\!\!\perp Z, W \quad \wedge \quad X \perp\!\!\!\perp Y \mid Z, W \quad \implies \quad X \perp\!\!\!\perp Z, W, Y$

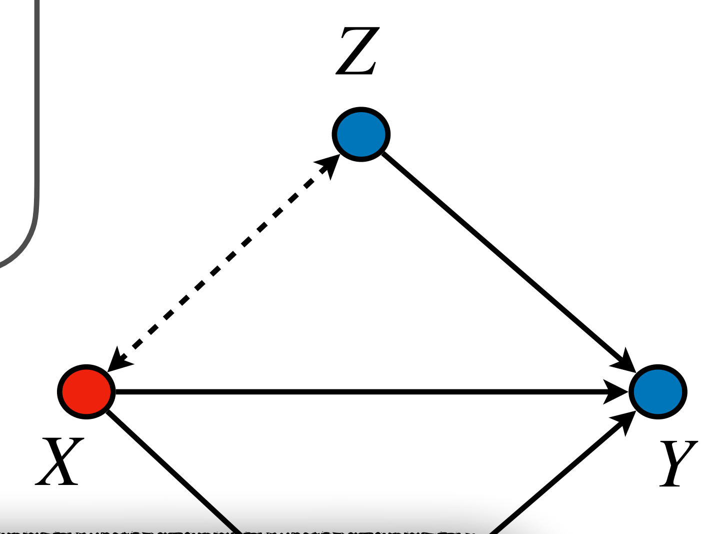
IF, Issue 4: complete metric d does not allow for business necessity

Part I. If $d((x,z,w),(x',z',w')) = \|z - z'\| + \|w - w'\|$

then $\text{OTC}^d_{x_0,x_1}((Z,W)) = 0 \implies X \perp\!\!\!\perp Z, W.$

Part II. If IF condition holds, then for Y binary

Par ... Y



A complete metric d implies X is independent of all other attributes, which is a strict requirement.

Relationship to previous literature

- How does the presented framework of Causal Fairness Analysis relate to previous literature?
- In particular, we discuss

(i) Counterfactual Fairness (Kusner et. al., 2017)

(ii) Individual Fairness (Dwork et. al., 2012)

(iii) Predictive Parity (Chouldechova, 2017)

Predictive Parity (PP)

Definition. Let $\widehat{Y}$ be the predictor of Y. We say that $\widehat{Y}$ satisfies predictive parity (PP) with respect to X, Y if

$$P(y \mid x_1, \widehat{y}) = P(y \mid x_0, \widehat{y}) \quad \forall \widehat{y}\ .$$

Alternatively, the PP criterion can also be written as a conditional independence statement

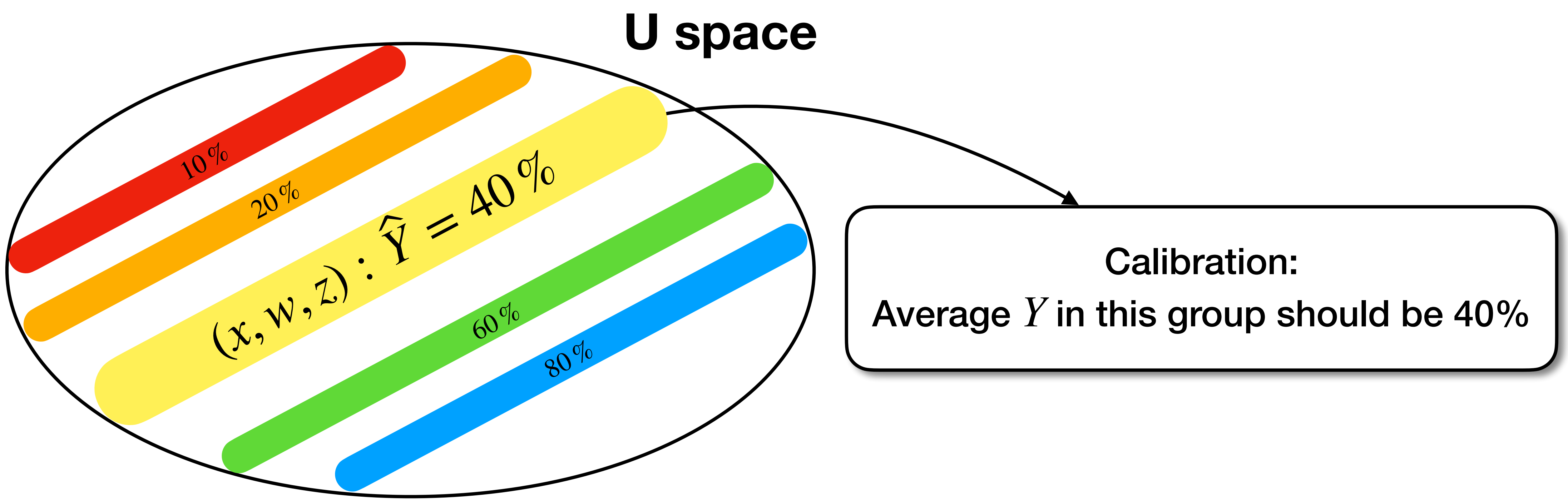
$$Y \perp\!\!\!\perp X \mid \widehat{Y}\ .$$

X has no more information about Y once we know $\widehat{Y}$

Finally, define the predictive parity measure to be

$$\text{PPM}_{x_0,x_1}(y \mid \widehat{y}) = P(y \mid x_1, \widehat{y}) - P(y \mid x_0, \widehat{y})\ .$$

PP Intuition



i.e.

$$\sum_{x,z,w:\hat{Y}(x,z,w)=\hat{y}} P(y \mid x, z, w) P(x, z, w \mid \hat{y}) = \hat{y}$$

Two key results on PP

Proposition 1 (PP & Efficient Learning). Suppose that the predictor $\widehat{Y}$ is based on the features X, Z, W. Suppose also that $\widehat{Y}$ is an efficient learner, meaning that:

$$\widehat{Y}(x, z, w) = P(y \mid x, z, w)\,.$$

Then, it follows that $\widehat{Y}$ satisfies predictive parity w.r.t

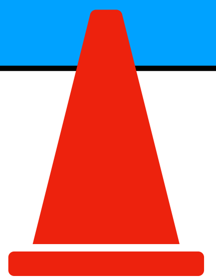
PP happens "naturally" for good learners!

Proposition 2 (PP & DP Impossibility). The fairness criteria of predictive parity and demographic parity,

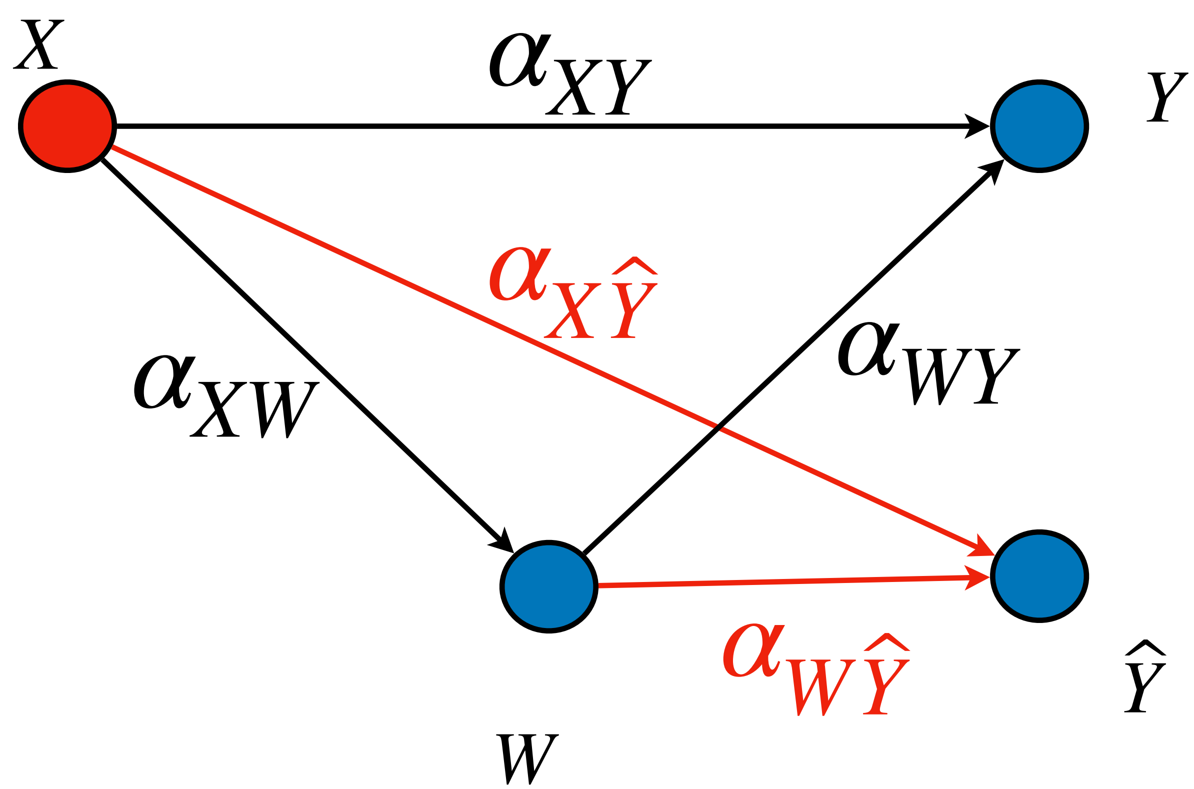
$$Y \perp\!\!\!\perp X \mid \widehat{Y},$$
$$\widehat{Y} \perp\!\!\!\perp X,$$

are mutually exclusive except for in degenerate cases, when $Y \perp\!\!\!\perp X$.

PP and DP are from different planets!



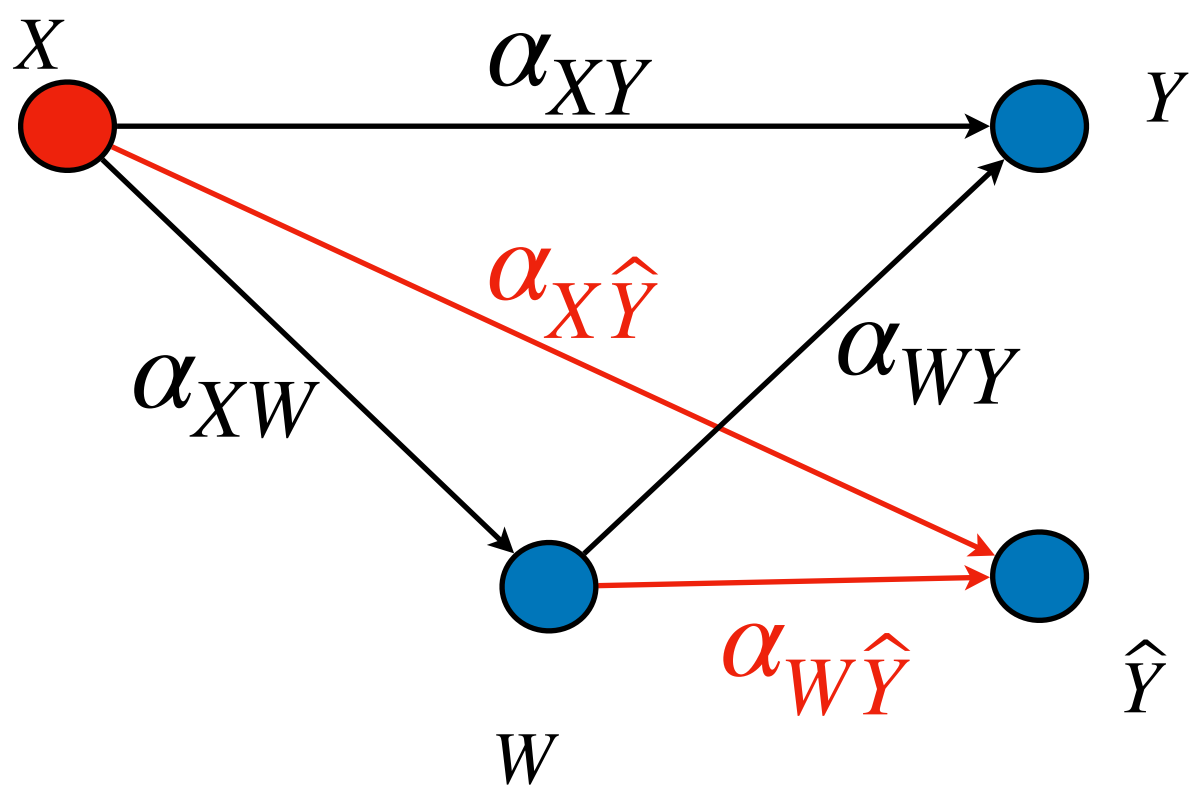
What is PP really doing?



$$E(y_{x_1} \mid x_1, \hat{y}) - E(y_{x_0} \mid x_1, \hat{y}) = \alpha_{XW}\alpha_{WY} + \alpha_{XY}$$

$$E(y_{x_0} \mid x_1, \hat{y}_{x_1}) - E(y_{x_0} \mid x_1, \hat{y}_{x_0}) = -(\alpha_{XW}\alpha_{WY} + \alpha_{XY}),$$

What is PP really doing?



$$E(y_{x_1} \mid x_1, \hat{y}) - E(y_{x_0} \mid x_1, \hat{y}) = \alpha_{XW}\alpha_{WY} + \alpha_{XY}$$

$$E(y_{x_0} \mid x_1, \hat{y}_{x_1}) - E(y_{x_0} \mid x_1, \hat{y}_{x_0}) = -(\alpha_{XW}\alpha_{WY} + \alpha_{XY}),$$

Not in control of the decision-maker!

Just the 2nd term is!

Causal Predictive Parity (CPP)

Definition. Let $\widehat{Y}$ be a predictor of the outcome Y, and let X be the protected attribute. Then we say that $\widehat{Y}$ satisfies causal predictive parity (CPP) with respect to a counterfactual contrast (C_0, C_1, E, E) if

$$E[y_{C_1} \mid E] - E[y_{C_0} \mid E] = E[\widehat{y}_{C_1} \mid E] - E[\widehat{y}_{C_0} \mid E] .$$

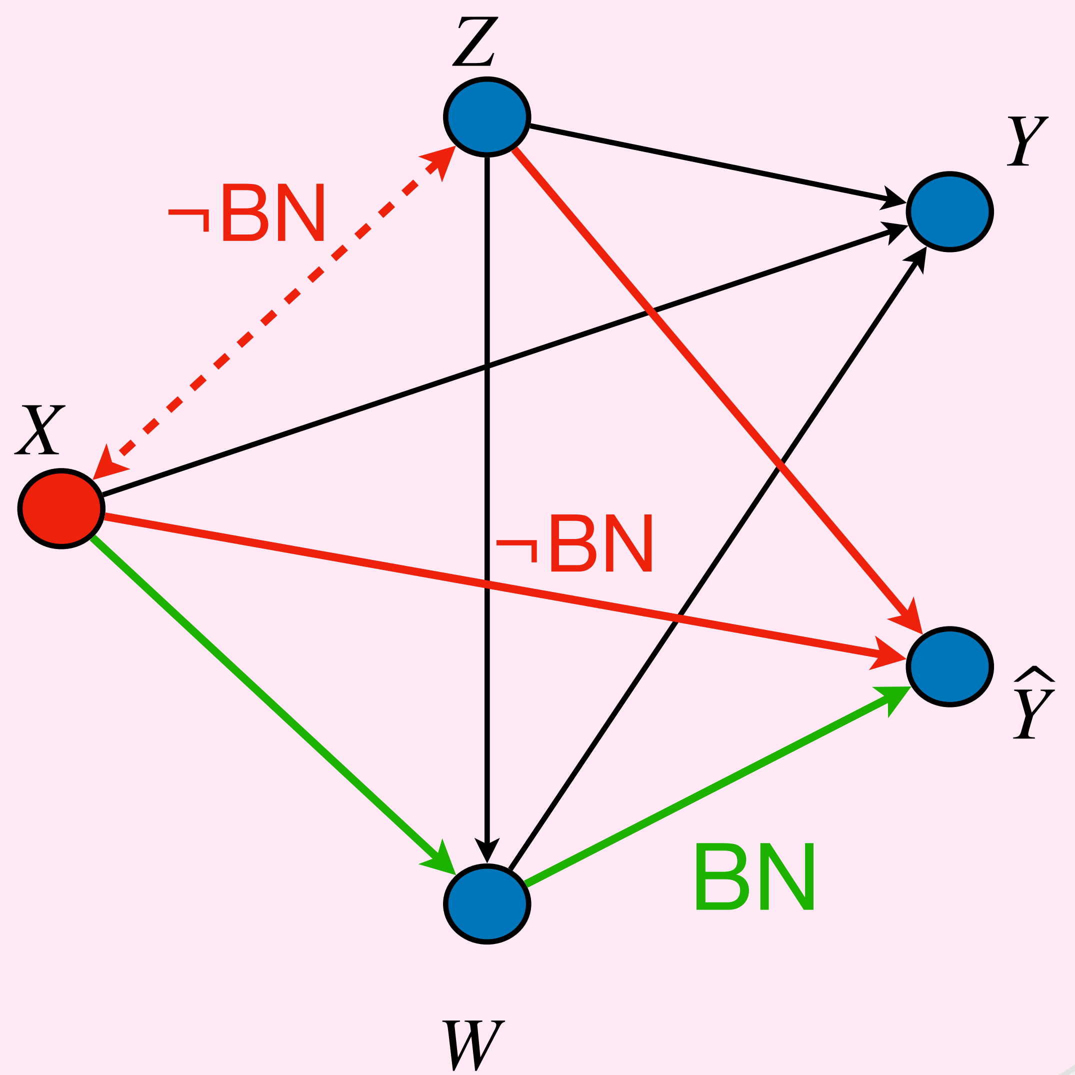
Furthermore, we say that $\widehat{Y}$satisfies CPP with respect to a factual contrast (C, C, E_0, E_1) if

$$E[y_C \mid E_1] - E[y_C \mid E_0] = E[\widehat{y}_C \mid E_1] - E[\widehat{y}_C \mid E_0] .$$

CPP implications?

"Modelling"

BN considerations:



"Implementing"

Requirements:

DE $= 0$

SE $= 0$

~~IE = arbitrary?~~

IE$(\widehat{y})$ = IE(y)!

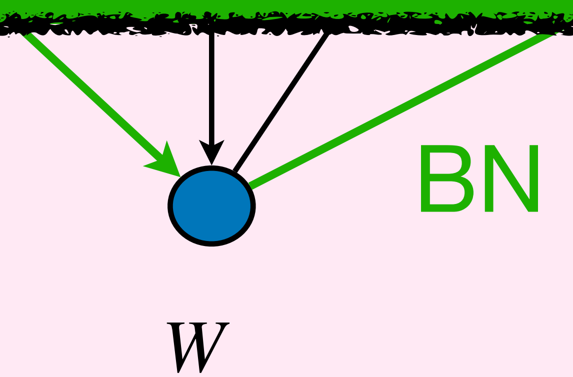
Causal PP

CPP implications?

"Modelling"

"Implementing"

BN considerations:



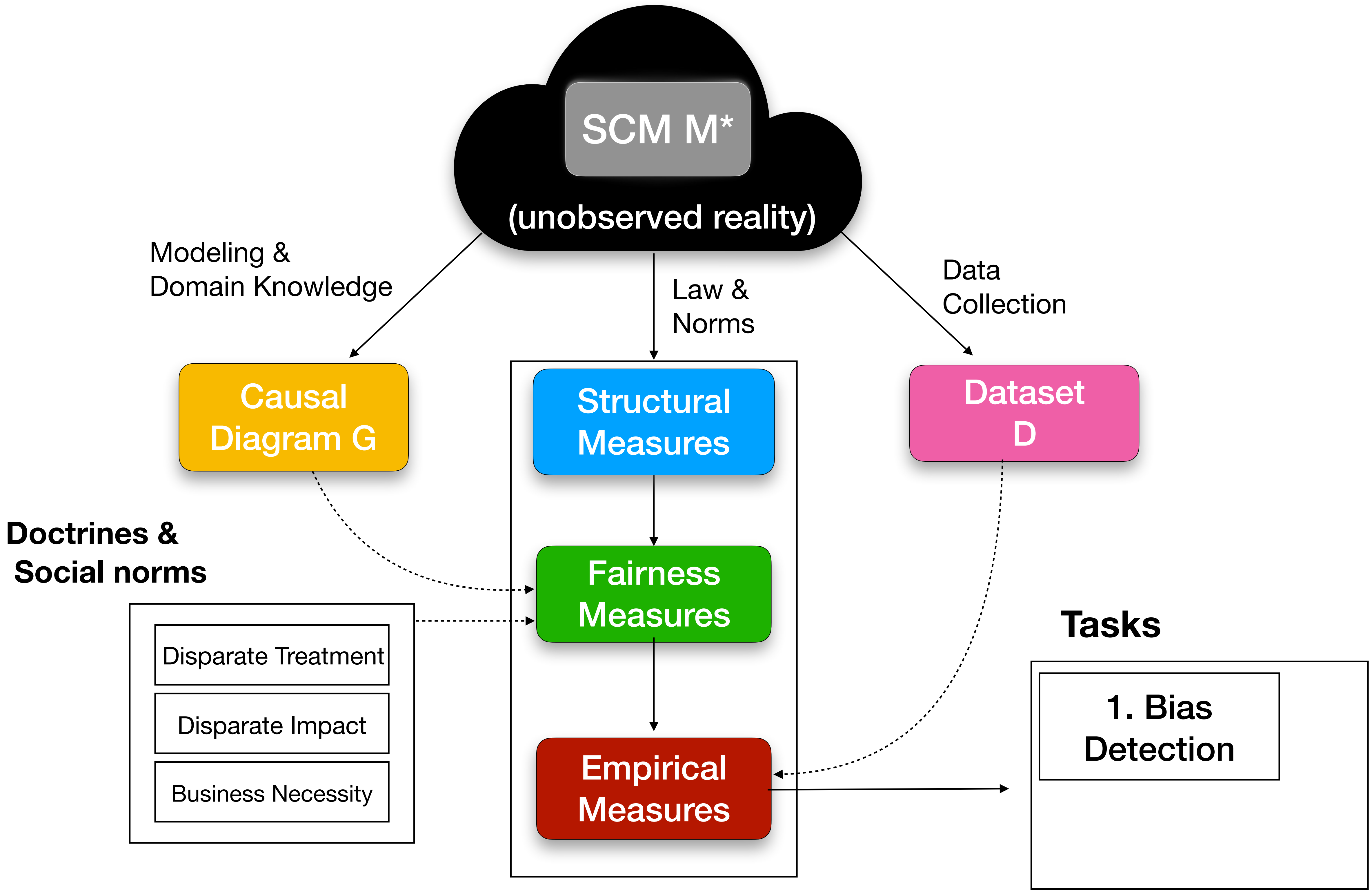
Requirements:

Completes the picture on Business Necessity!

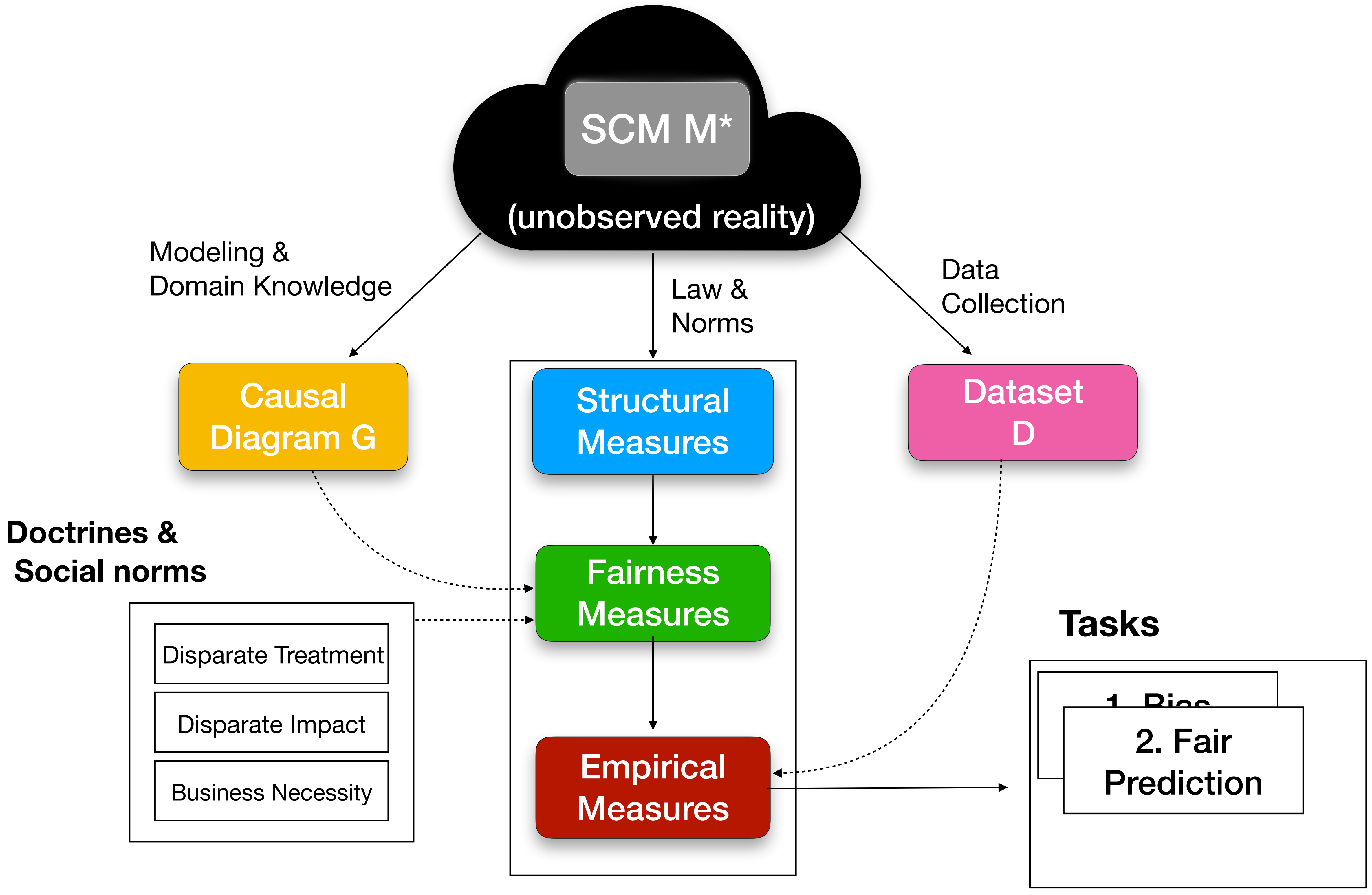
$\mathbb{E}(y) = \mathbb{E}(y)!$

Causal PP

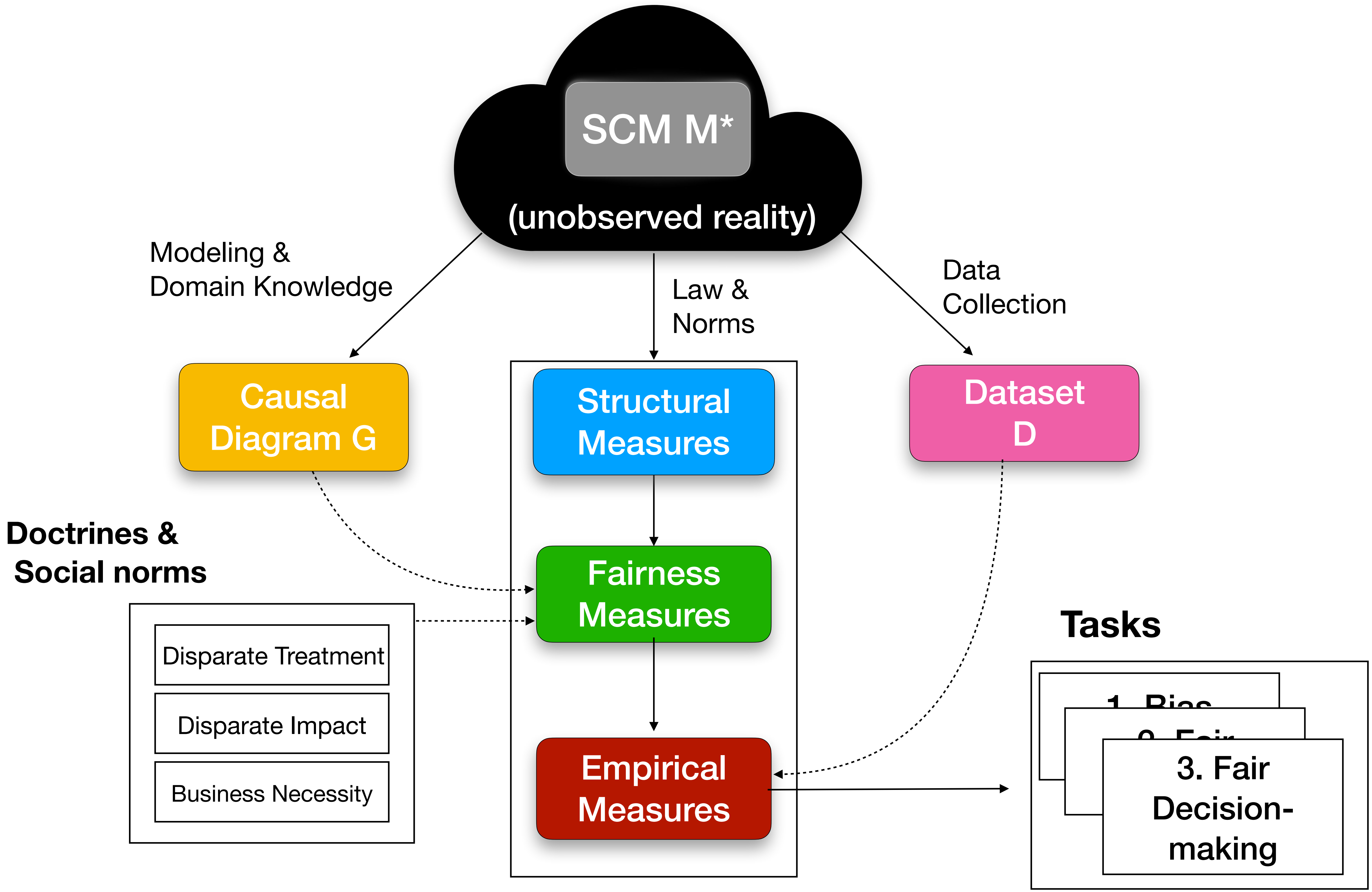
Fairness Tasks
(Big Picture)



Fairness Tasks
(Big Picture)



Fairness Tasks
(Big Picture)



Fairness Tasks
(Big Picture)