Causal Fairness Analysis

(Causal Inference II - Lecture 4)

Elias Bareinboim

Drago Plecko

Columbia University
Computer Science

Reference:

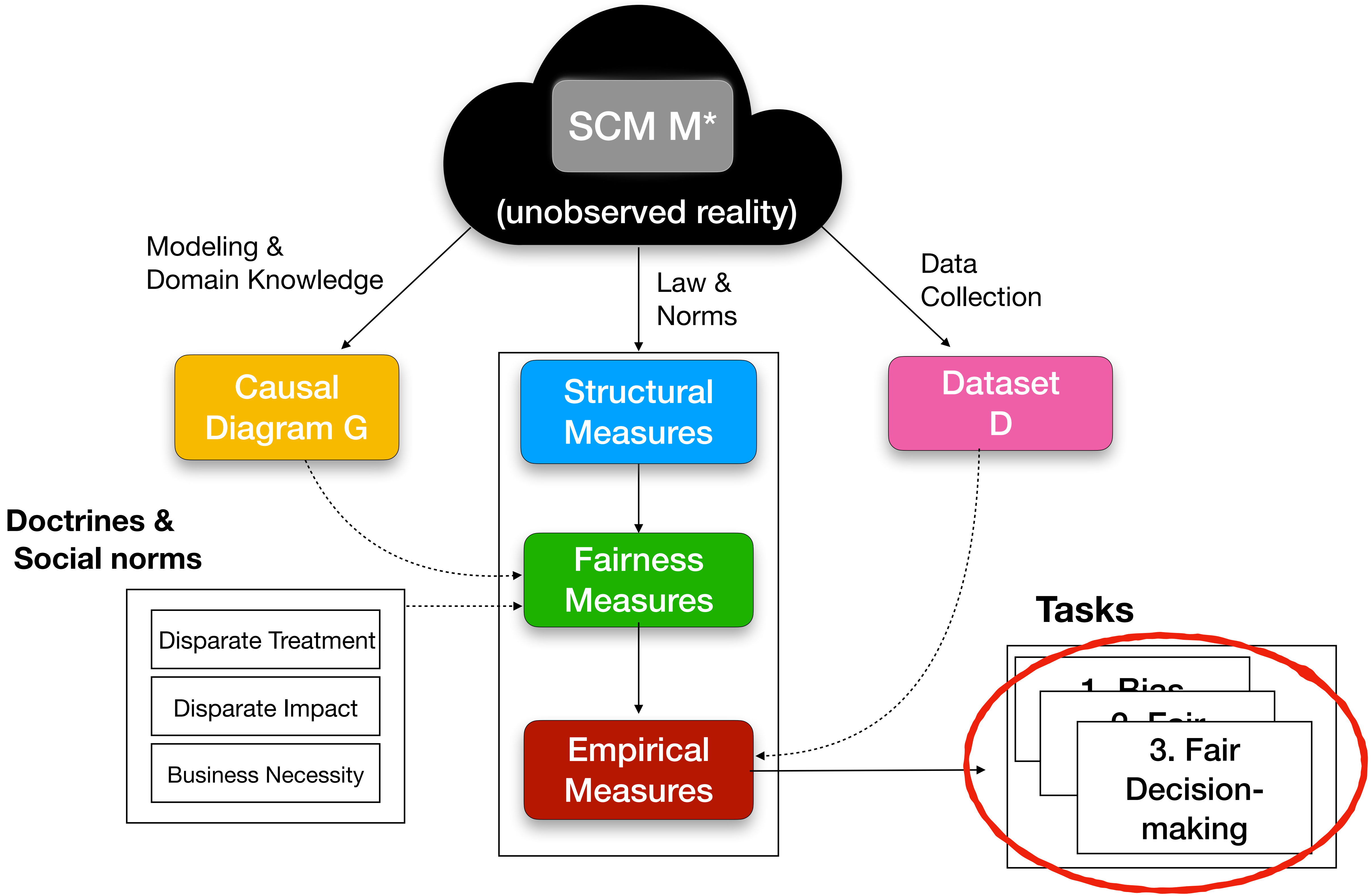
D. Plecko, E. Bareinboim.

Causal Fairness Analysis.

TR R-90, CausalAI Lab, Columbia University.

https://causalai.net/r90.pdf

Fairness Tasks
(Big Picture)



Fairness Map

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	✘

Task 1.
Bias Detection & Quantification

Fairness Cookbook

Fairness Cookbook

Section 5.1
Algorithm 1

1) Obtain data on past decisions $\mathscr{D}$.

2) Determine the (possibly simplified) causal diagram $\mathscr{G}$ (w.r.t. underlying $\mathscr{M}^*$).

3) Determine the **Business Necessity** (BN) set ($\emptyset$, $\{Z\}$, $\{W\}$, $\{Z, W\}$).

4) Consider existence of **Disparate Treatment:**

$$H_0^{(x\text{-DE})} : x\text{-DE}_{x_0,x_1}(y \mid x_0) = 0.$$

- rejected → evidence of disparate treatment (population level)
- not rejected → no evidence of disparate treatment (population level)

5) Consider existence of **Disparate Impact:**

5a) Indirect effect:

if($W \notin$ BN-set) $H_0^{(x\text{-IE})} : x\text{-IE}_{x_0,x_1}(y \mid x_0) = 0.$

- rejected → evidence of disparate impact
- not rejected → go to next step

5b) Spurious effect:

if($Z \notin$ BN-set) $H_0^{(x\text{-SE})} : x\text{-SE}_{x_0,x_1}(y) = 0.$

- rejected → evidence of disparate impact
- not rejected → no evidence of disparate impact

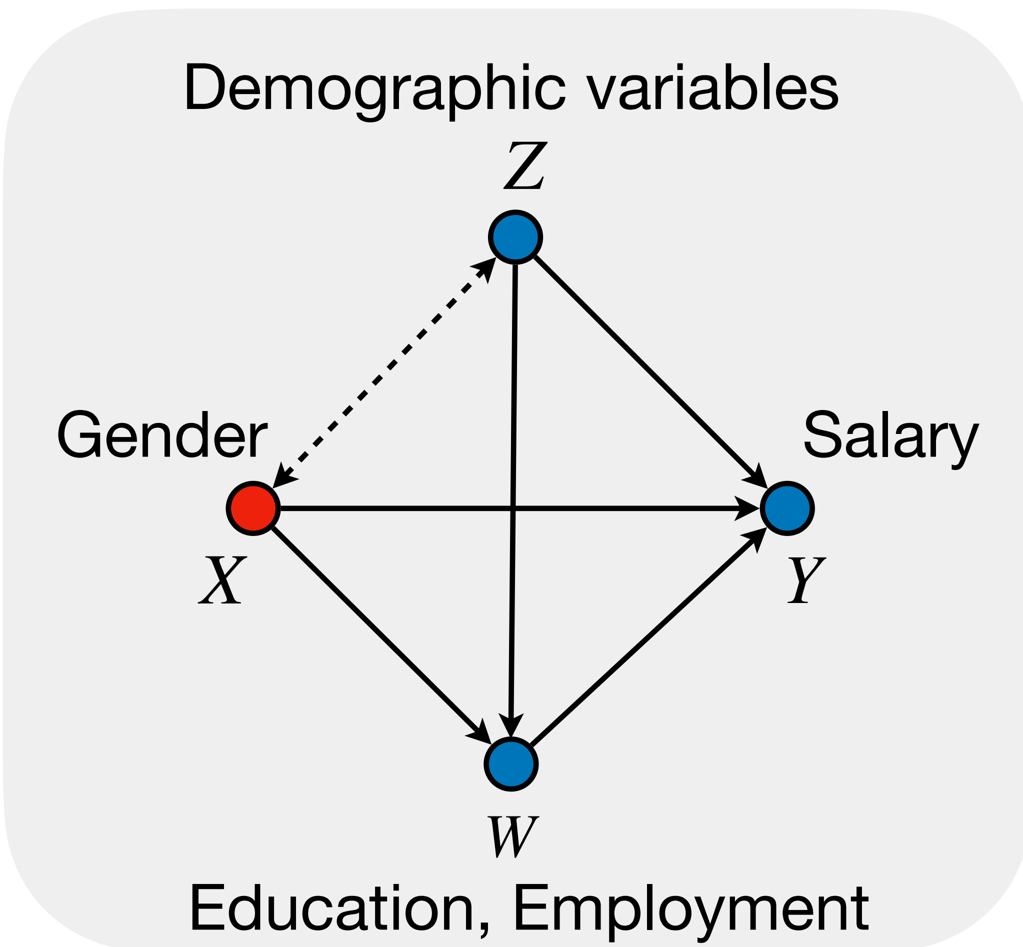
Spectrum of Fairness Notions

Business Necessity Considerations

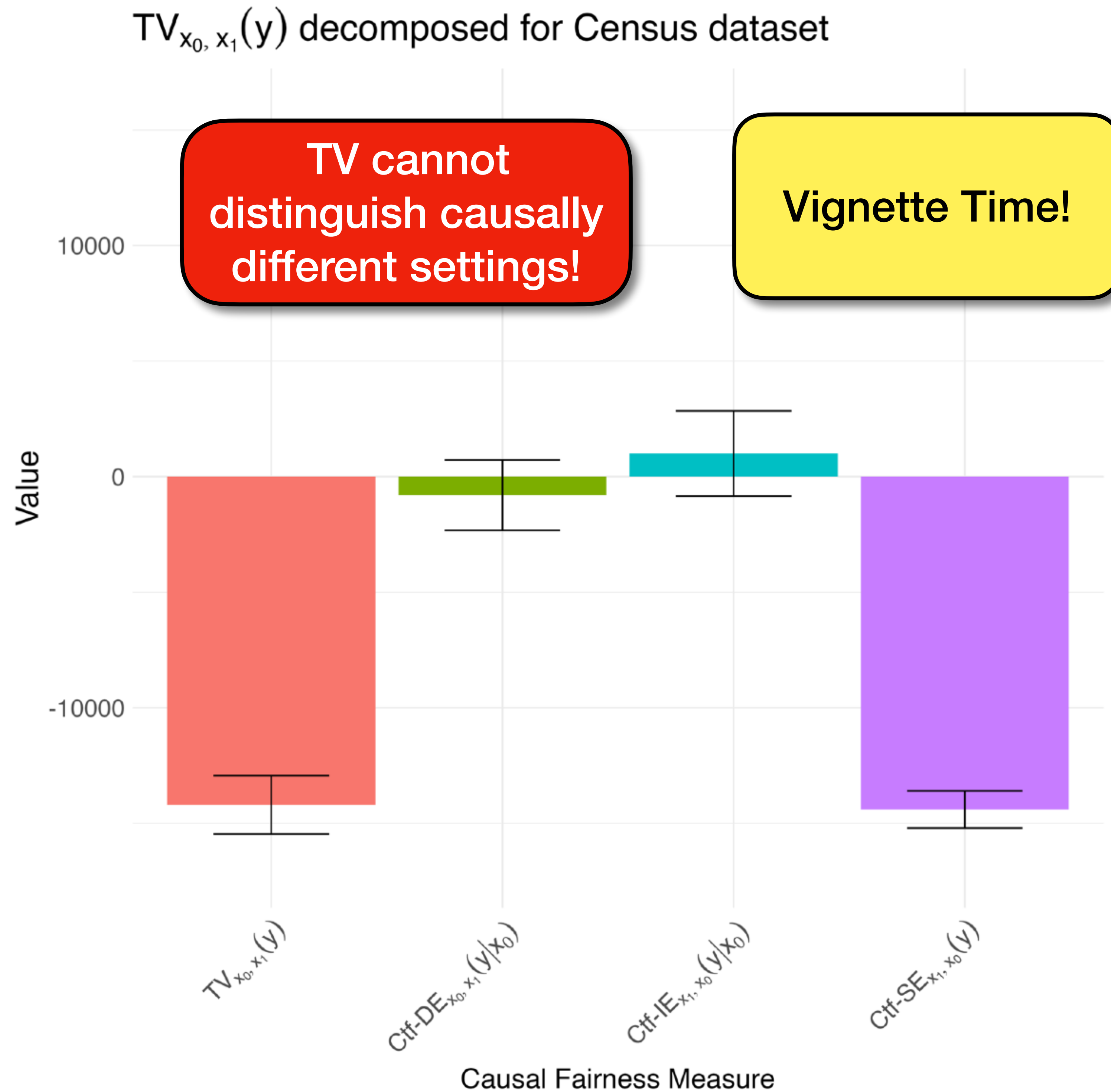
Disparate Treatment

Disparate Impact
(extreme case)

Task 1A: One Step Bias Quantification (Census 2018)



- Observed disparity: $\text{TV}_{x_0,x_1}(y) = \$14{,}000/\text{year}$



Task 1B: Multi-step Bias Quantification (College Admissions)

Example. A university in the United States admits applicants every year. The university is interested in quantifying discrimination in the admission process and track it over time, between 2010 and 2020. The data generating process changes over time, and can be described as follows. Let X denote gender (x_0 female, x_1 male). Let Z be the age at time of application ($Z = 0$ under 20 years, $Z = 1$ over 20 years) and let W denote the department of application ($W = 0$ for arts & humanities, $W = 1$ for sciences). Finally, let Y denote the admission decision.

SCM $\langle \mathcal{F}_t, P_t(U) \rangle$

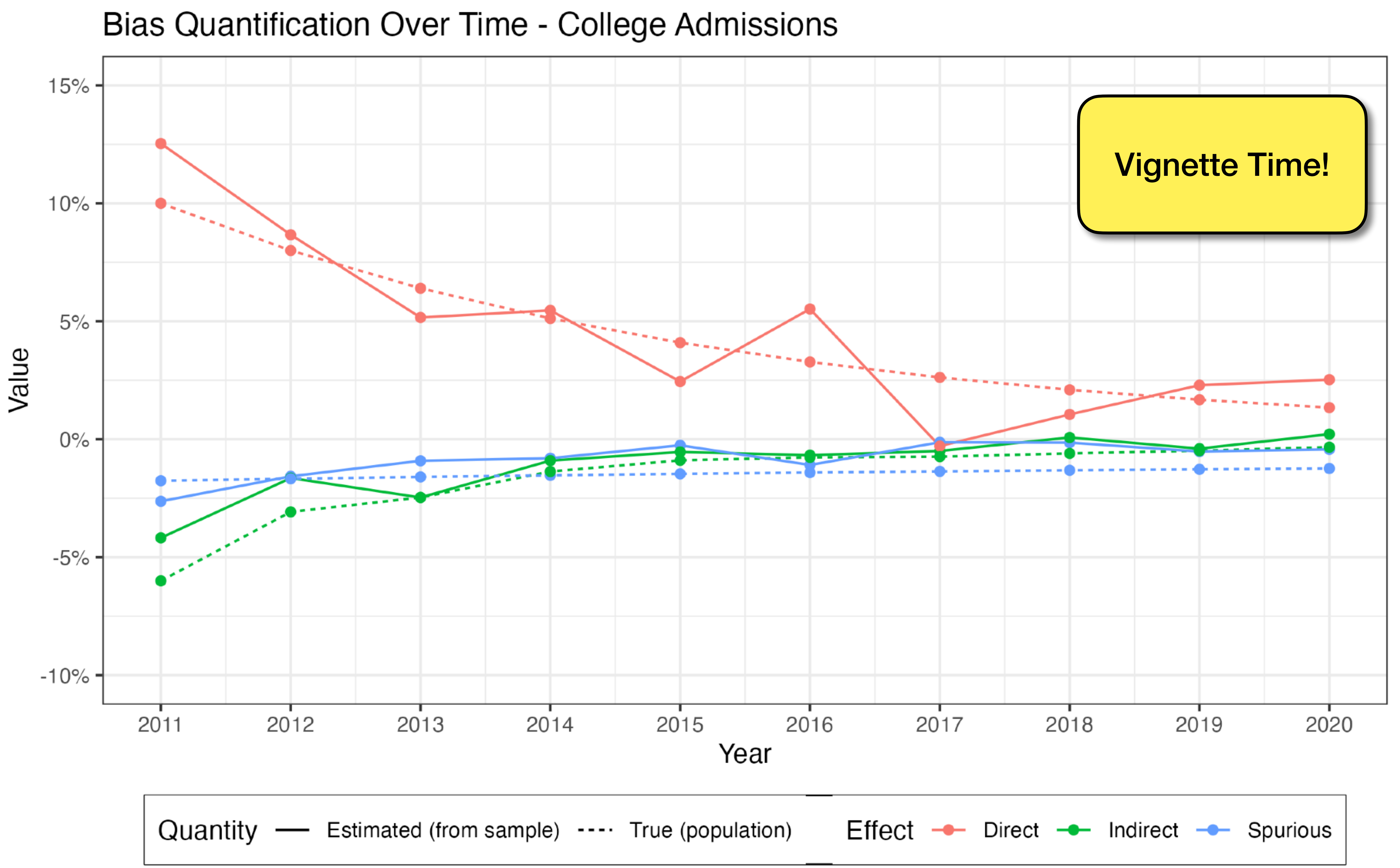
$$
\begin{aligned}
X &\leftarrow 1(U_X < 0.5 + 0.1 U_{XZ}) \\
Z &\leftarrow 1(U_Z < 0.5 + \kappa(t) U_{XZ}) \\
W &\leftarrow 1(U_W < 0.5 + \lambda(t) X) \\
Y &\leftarrow 1(U_Y < 0.1 + \alpha(t) X + \beta(t) W + 0.1 Z)
\end{aligned}
$$

$$
U_{XZ} \in \{0,1\}, P(U_{XZ} = 1) = 0.5, \\
U_X, U_Z, U_W, U_Y \sim \textbf{Unif}[0,1] .
$$

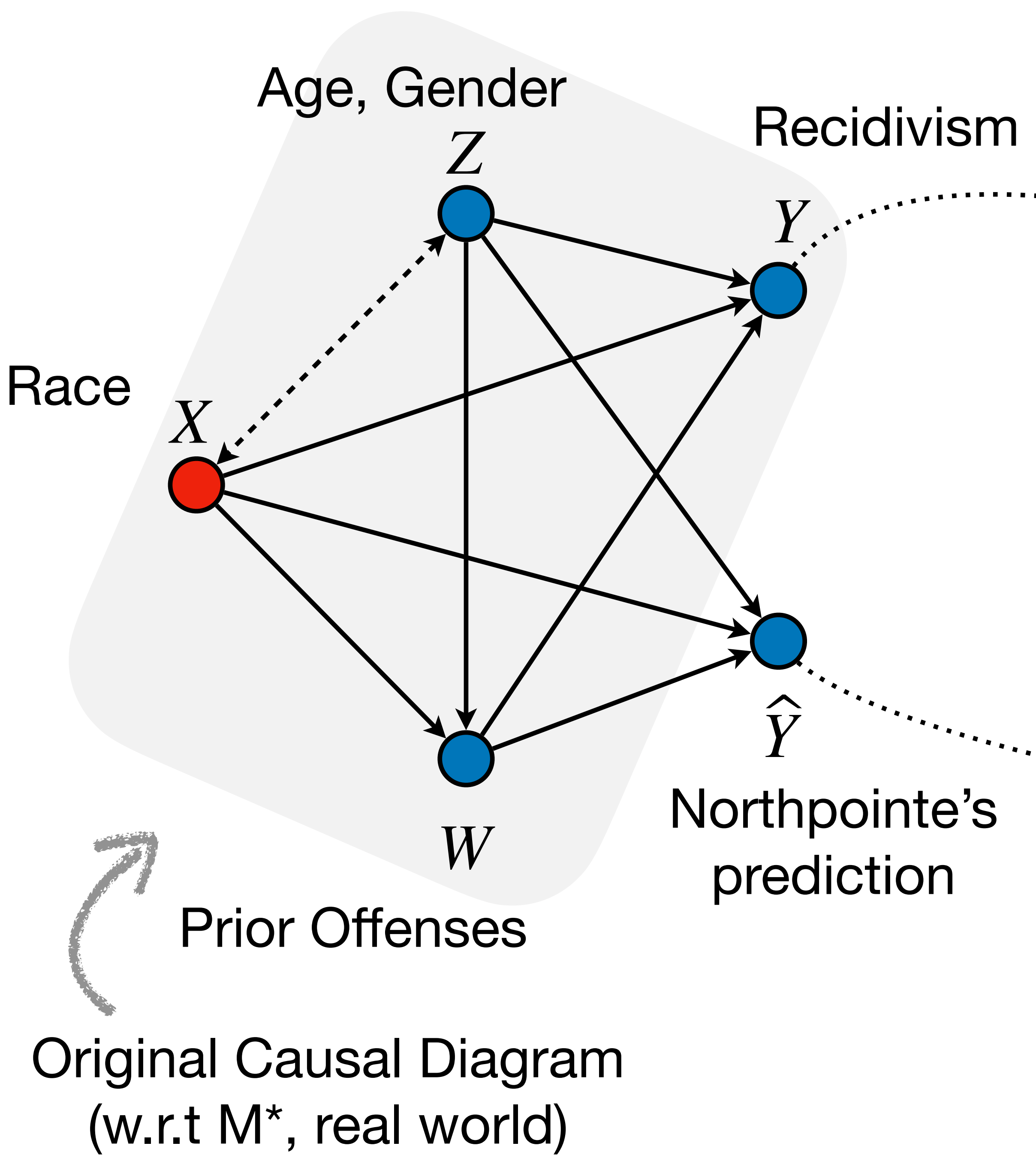
Time Evolution $\theta_{t \to t+1}$

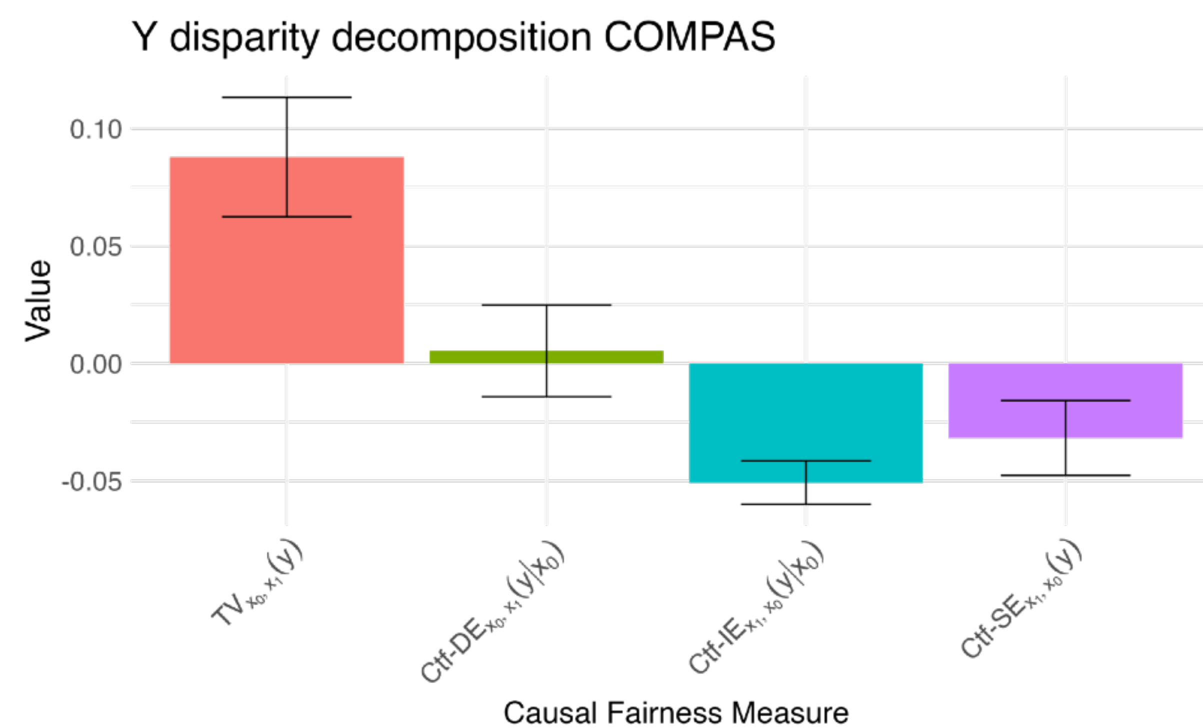
$$
\begin{aligned}
\kappa(t+1) &= 0.9 \kappa(t) \\
\lambda(t+1) &= \lambda(t)(1 - \beta(t)) \\
\beta(t+1) &= \beta(t)(1 - \lambda(t)) f(t), \\
& \qquad f(t) \sim \textbf{Unif}[0.8, 1.2] \\
\alpha(t+1) &= 0.8 \alpha(t)
\end{aligned}
$$

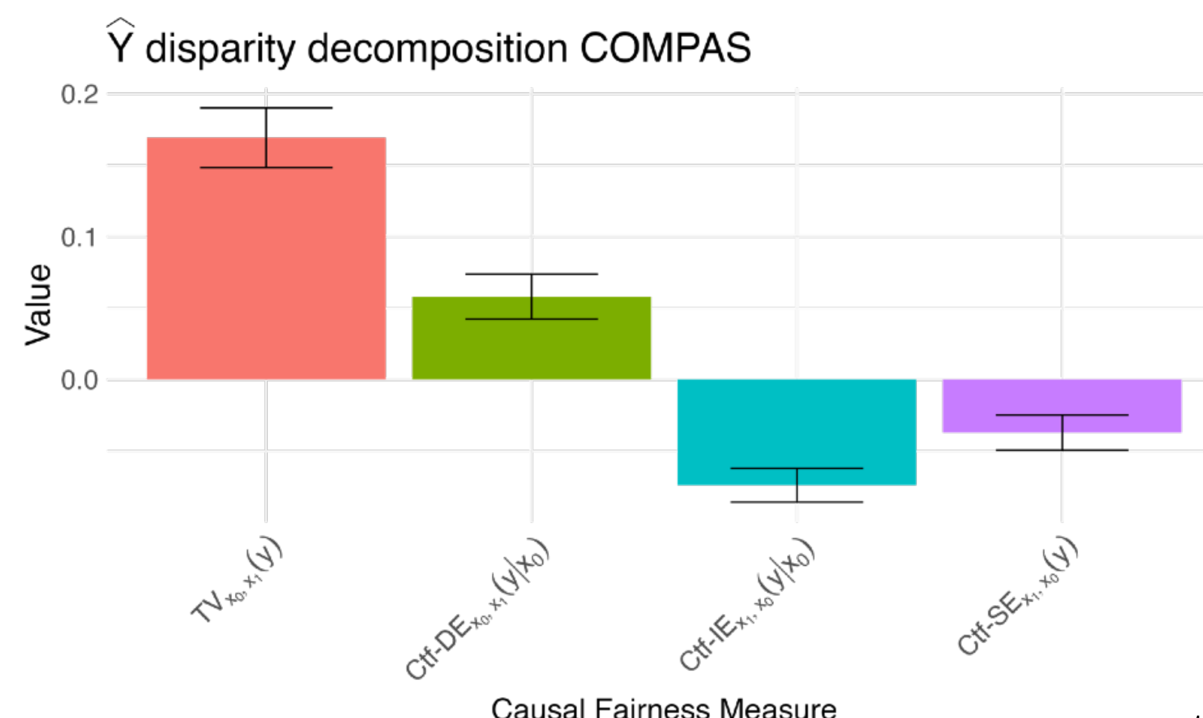
Bias Quantification over time



Task 1C: Bias Detection for Y and $\widehat{Y}$ (COMPAS)



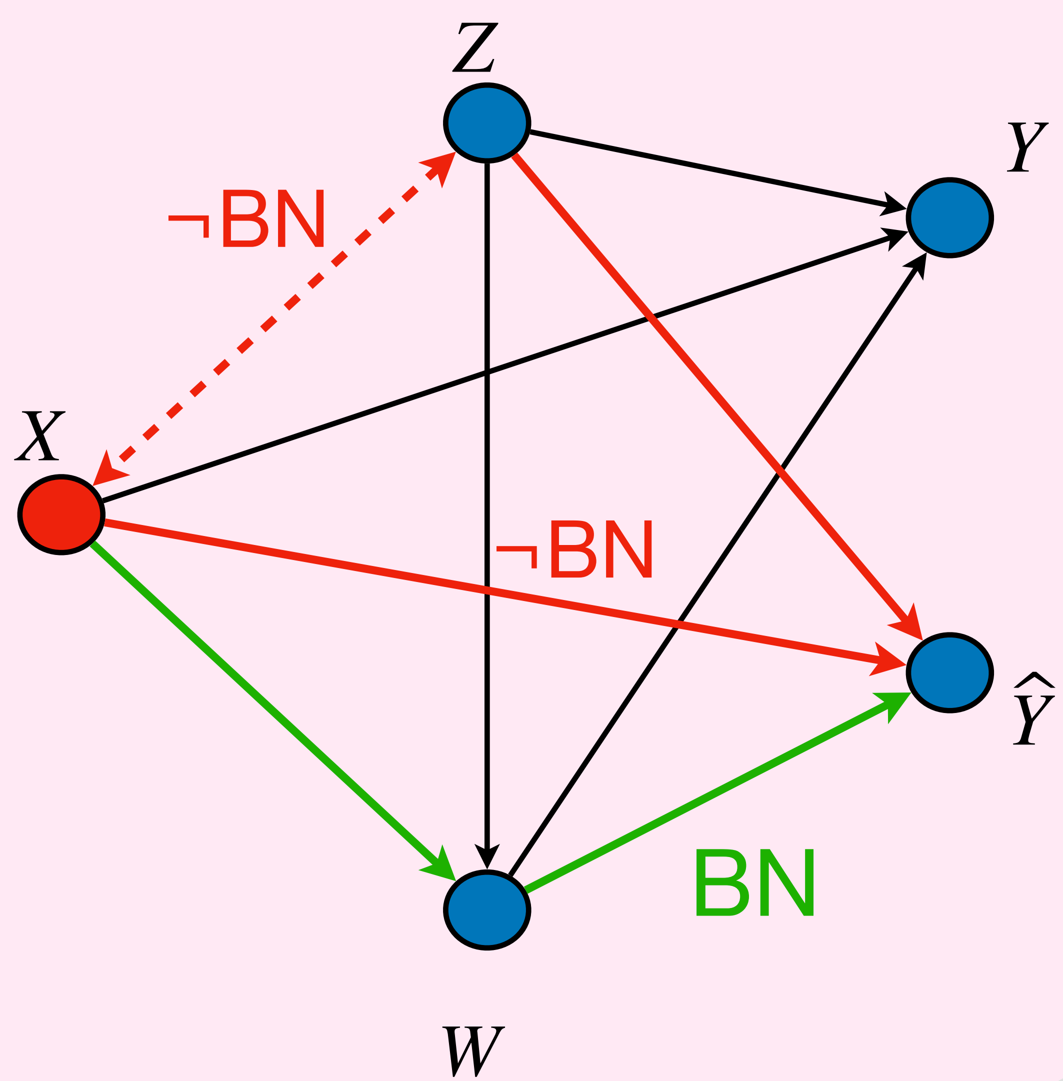




Recall: CPP Implications

"Modelling"

BN considerations:



"Implementing"

Requirements:

$\text{DE} = 0$

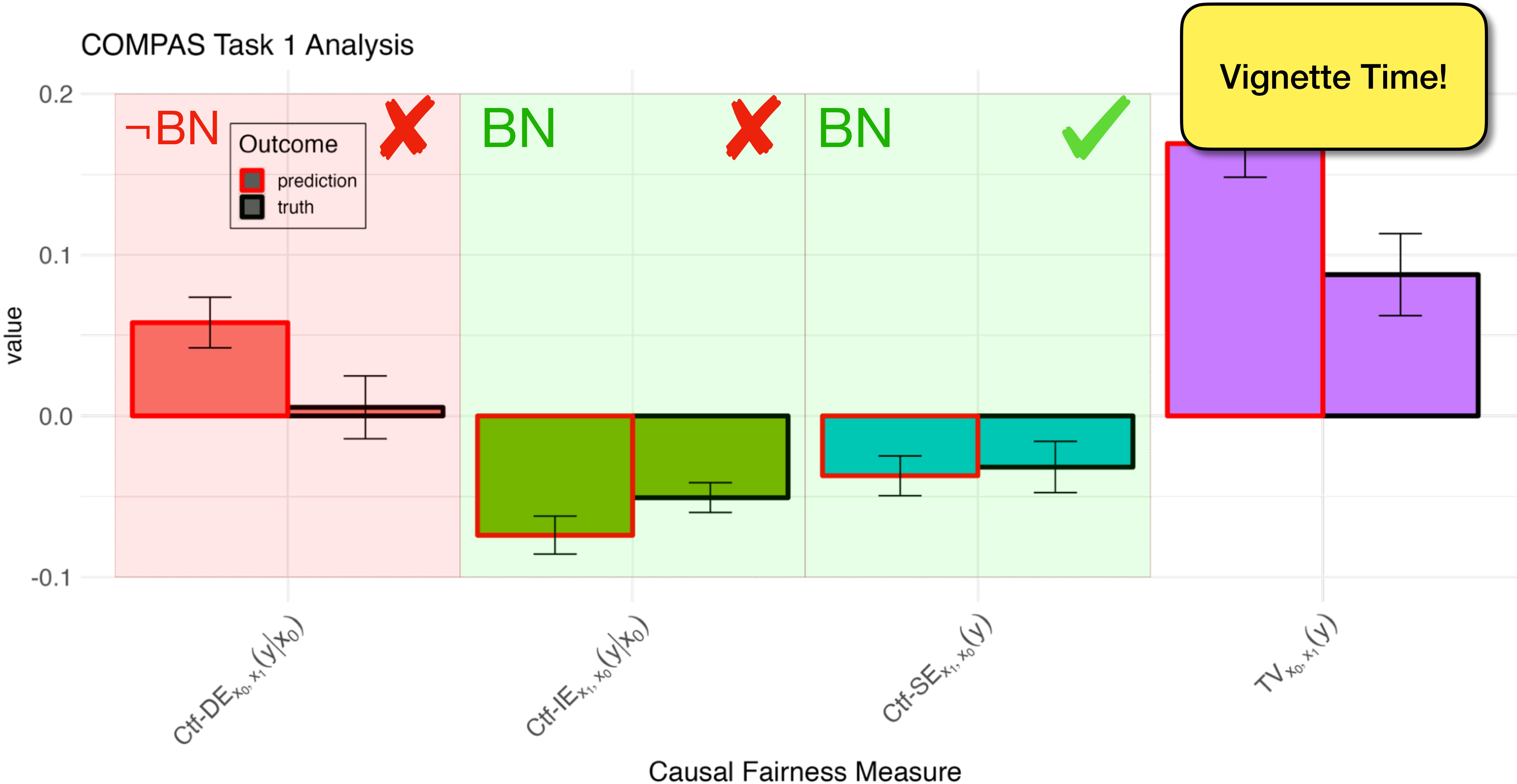
$\text{SE} = 0$

~~IE = arbitrary?~~

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

Causal PP

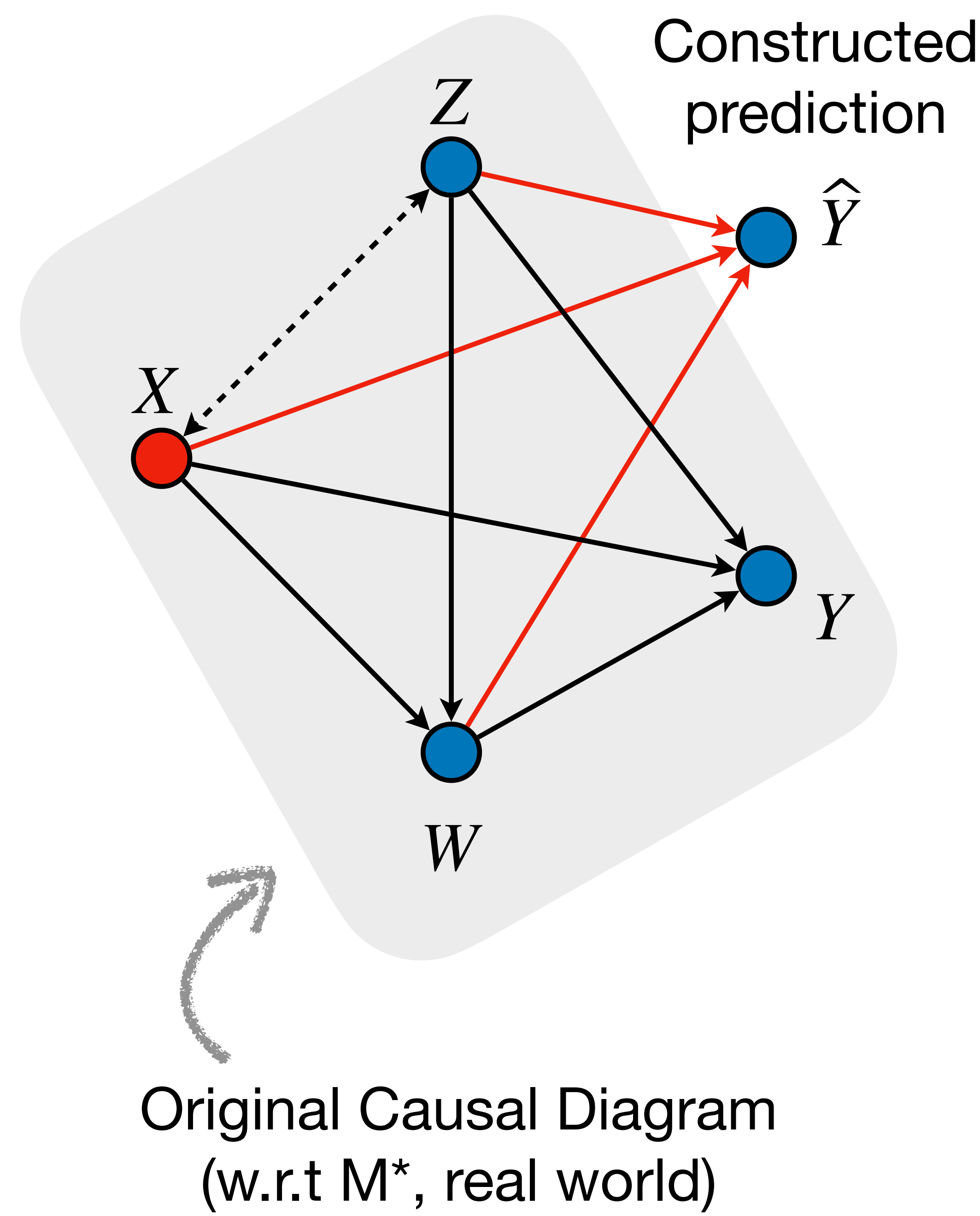
Task 1C: Bias Detection for Y and $\widehat{Y}$ (COMPAS)



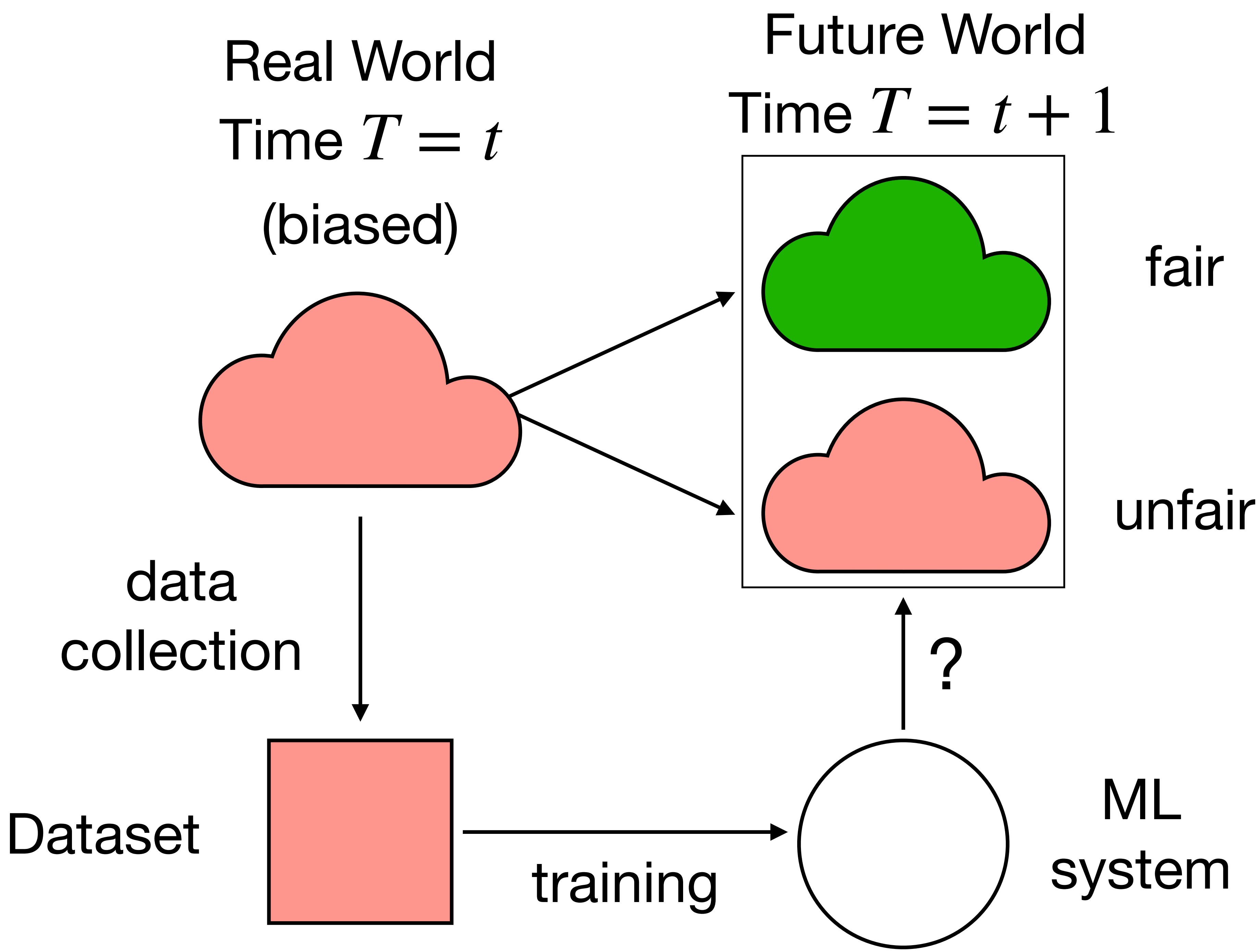
Task 2. Fair Predictions

Prediction Task

- The first talk focused on bias detection, where we just analyze the "observed reality", i.e., nature defines f_Y
- When doing prediction, causally speaking, we are constructing a new mechanism $\widehat{Y} \leftarrow f_{\widehat{Y}}(x, z, w)$ that is under our control (i.e., we are selecting it)
- Typically, in ML, we are simply interested in learning $P(y \mid x, z, w)$
- Does that carry over bias from f_Y?

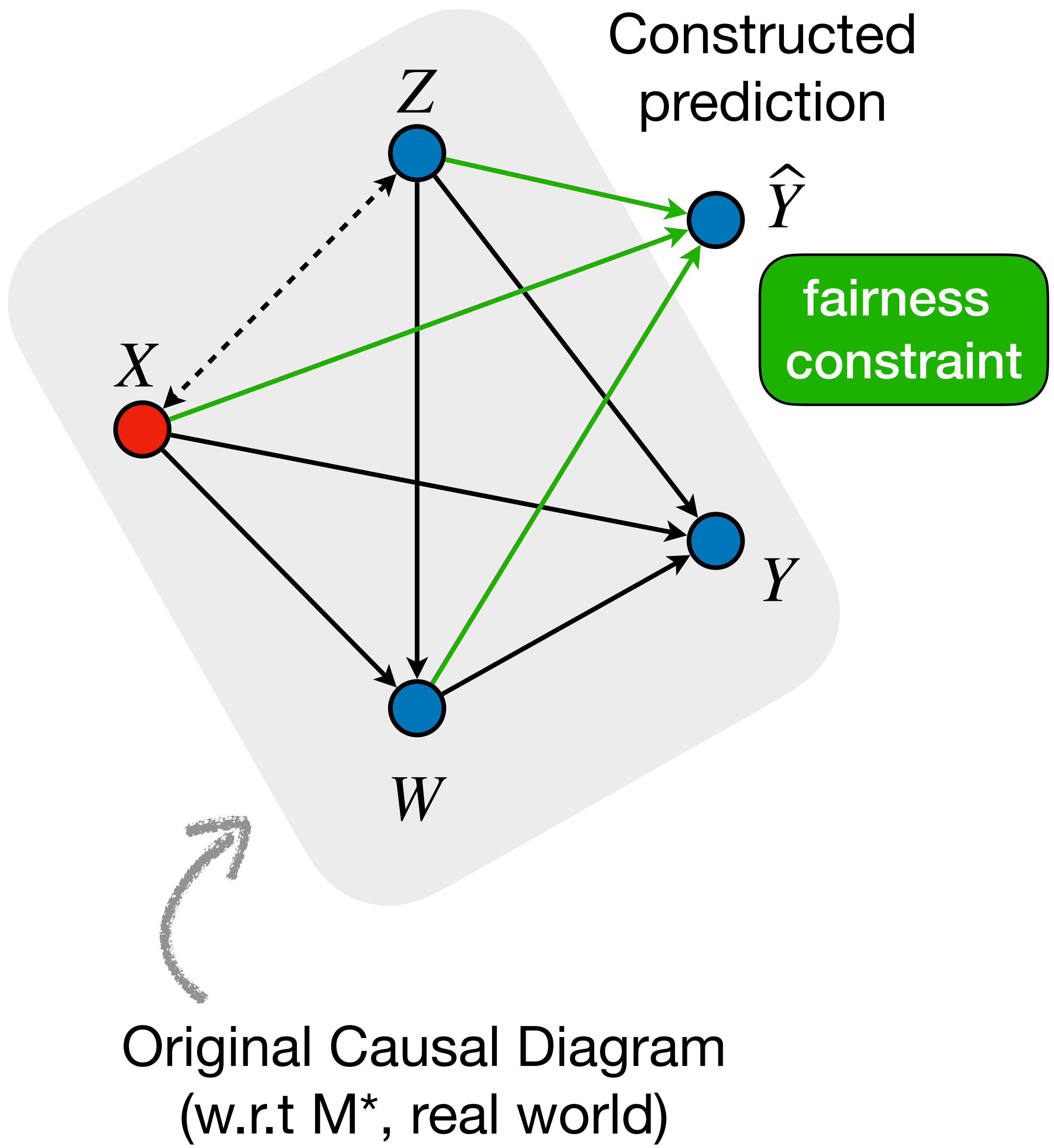


From a biased reality towards a more fair one?



Fair Prediction

- General answer: simply learning $P(y \mid x, z, w)$ will give biased predictions.
- To remove the bias, one might wish for $\widehat{Y}$ to satisfy a pre-specified fairness constraint.
- A commonly considered constraint is to make $\text{TV}_{x_0,x_1}(\widehat{Y}) = 0$.
- In practice, there are different ways to satisfying such a constraint: in particular, we distinguish post-processing, in-processing, and pre-processing methods.

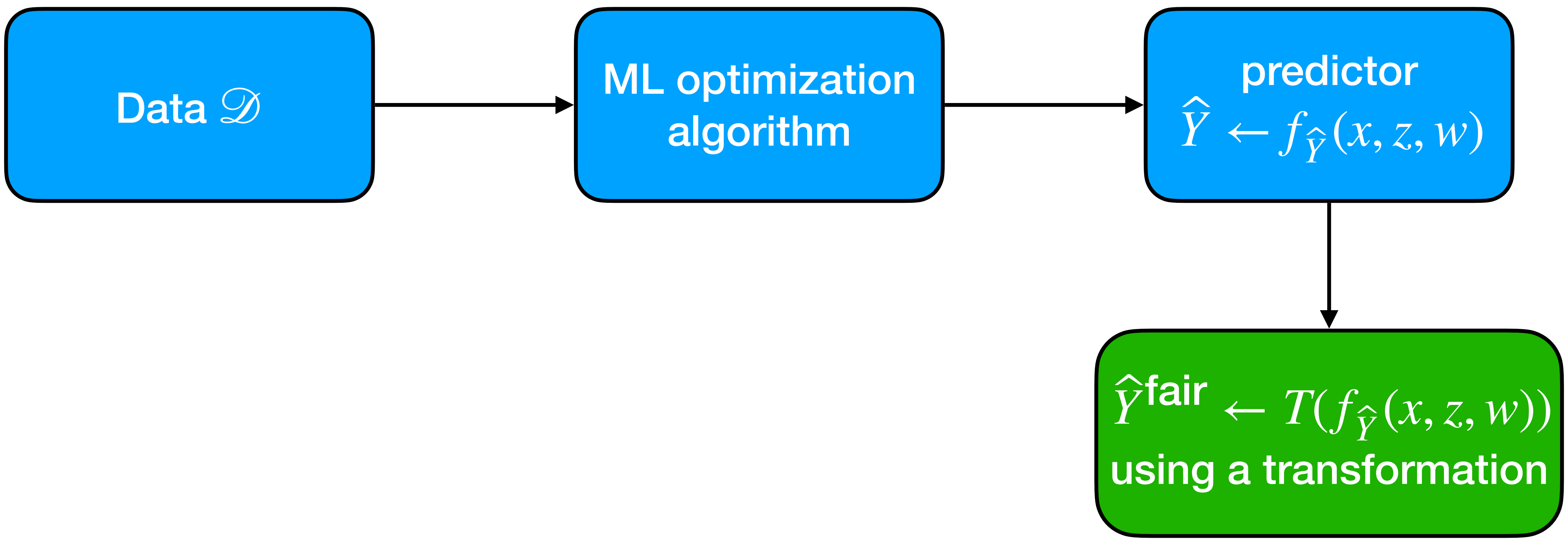


The Typical Approach

Typical ML framework:

Post-processing Methods

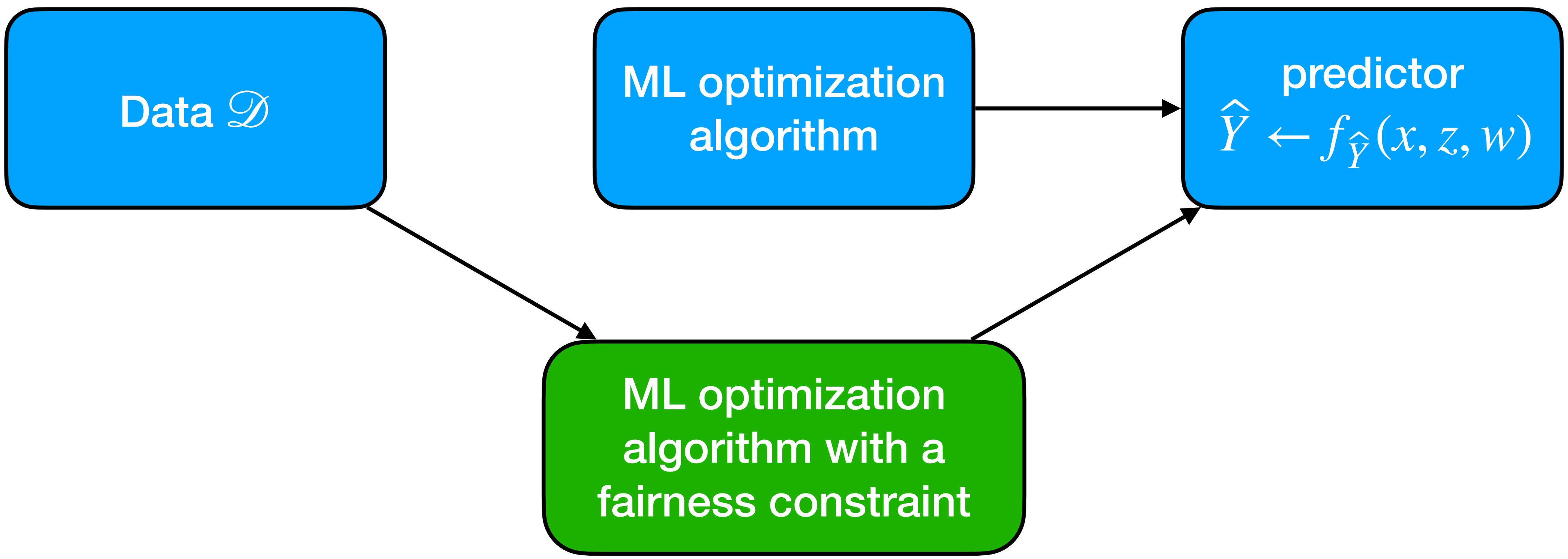
Typical ML framework:



Post-processing:
massage the predictions
to satisfy a constraint

In-processing Methods

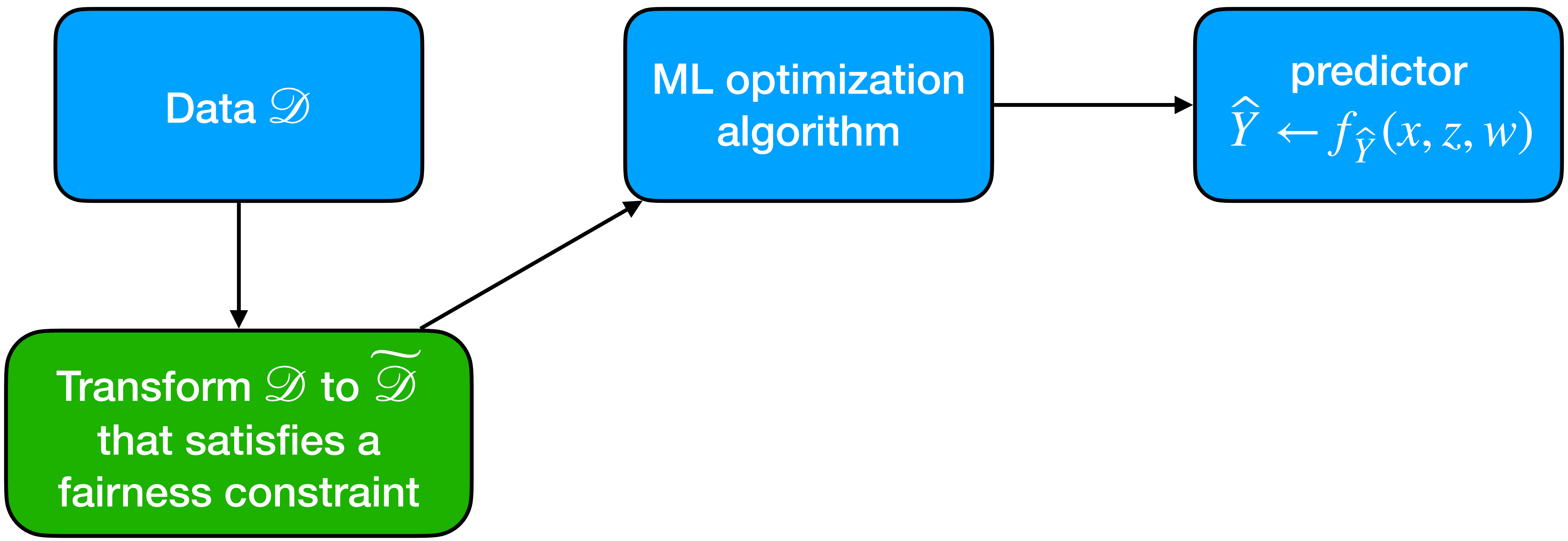
Typical ML framework:



In-processing:
include a fairness constraint in the learning step

Pre-processing Methods

Typical ML framework:



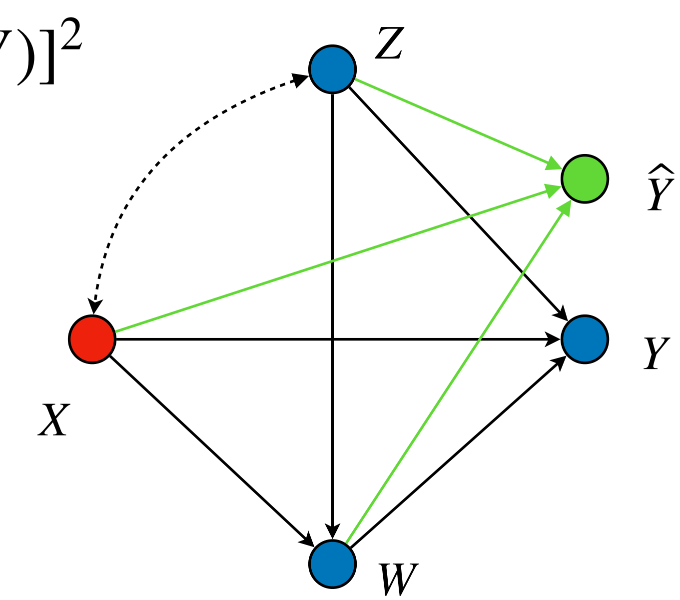
Pre-processing: change the data to satisfy a constraint apriori

Fair Prediction Theorem (FPT)

Theorem. Let SFM(n_Z, n_W) be the SFM with $|Z| = n_Z$ and $|W| = n_W$. Let E denote the set of edges of SFM (n_Z, n_W). Further, let $\mathcal{S}^{\text{linear}}_{n_Z,n_W}$ be the space of linear SCMs (but for the variable X, which is a Bernoulli) compatible with the SFM(n_Z, n_W) and whose structural coefficients are drawn uniformly from $[-1,1]^{|E|}$.

An SCM $M \in \mathcal{S}^{\text{linear}}_{n_Z,n_W}$ is said to be ϵ-TV-compliant if

$$\hat{f}_{\text{fair}} = \text{argmin}_{f \text{ linear}} \quad E[Y - f(X, Z, W)]^2$$
$$\text{subject to} \quad TV_{x_0,x_1}(f) = 0$$



also satifies

$$|\text{Ctf-DE}_{x_0,x_1}(\hat{f}_{\text{fair}} \mid x_0)| \le \epsilon,$$
$$|\text{Ctf-IE}_{x_0,x_1}(\hat{f}_{\text{fair}} \mid x_0)| \le \epsilon,$$
$$|\text{Ctf-SE}_{x_0,x_1}(\hat{f}_{\text{fair}})| \le \epsilon\,.$$

Under the Lebesgue measure ... SF ... e 0.

Furthermore, for any n_Z, n_W t ... TV-

non-vanishing probability of things "going wrong"

Section 5.2
Theorem 5.1

FPT proof sketch

Objective:

$$Y = \sum_{V_i \in X,Z,W} a_{V_iY} V_i + \epsilon_Y, \quad f(X,Z,W) = \sum_{V_i \in X,Z,W} \tilde{a}_{V_iY} V_i .$$

$$E[Y - f(X,Z,W)]^2 = E\Big[\sum_{V_i \in X,Z,W} (a_{V_iY} - \widetilde{a}_{V_iY}) V_i + \epsilon_Y \Big]^2$$

$$= E[\epsilon_Y^2] + E\Big[\sum_{V_i, V_j \in X,Z,W} (a_{V_iY} - \widetilde{a}_{V_iY})(a_{V_jY} - \widetilde{a}_{V_jY}) V_i V_j \Big]$$

$$= 1 + (a_{VY} - \widetilde{a}_{VY})^T E[VV^T](a_{VY} - \widetilde{a}_{VY}),$$

optimizing over $\widetilde{a}_{VY}$

ellipsoid

Linear SCM:

$U \leftarrow N(0,1)$

$X \leftarrow$ **Bernoulli**$($**expit**$(U))$

$Z \leftarrow a_{UZ}U + a_{ZZ}Z\epsilon_Z$

$W \leftarrow a_{XW}X + a_{ZW}Z + a_{WW}W + \epsilon_W$

$Y \leftarrow a_{XY}X + a_{ZY}Z + a_{WY}W + \epsilon_Y$

$$TV_{x_0,x_1}(f) = (E[V \mid x_1] - E[V \mid x_0])^T \widetilde{a}_{VY} = 0.$$

what the constraint is

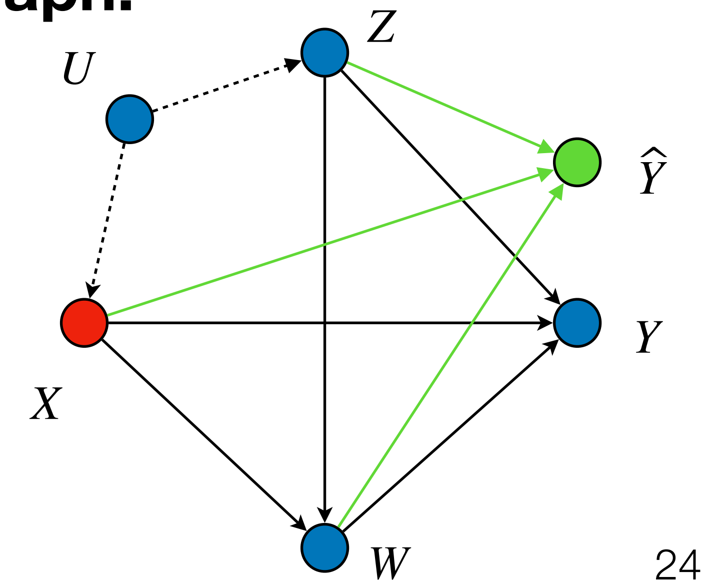
$$Ctf\text{-}DE = \widetilde{a}_{XY}(x_1 - x_0) = 0,$$

$$Ctf\text{-}IE = \sum_{W_i} \widetilde{a}_{W_iY}(E[W_i \mid x_1] - E[W_{i_{x_0}} \mid x_1]) = 0,$$

$$Ctf\text{-}SE = \sum_{W_i} \widetilde{a}_{W_iY}(E[W_{i_{x_0}} \mid x_1] - E[W_i \mid x_0]) + \sum_{Z_i} \widetilde{a}_{Z_iY}(E[Z_i \mid x_1] - E[Z_i \mid x_0]) = 0.$$

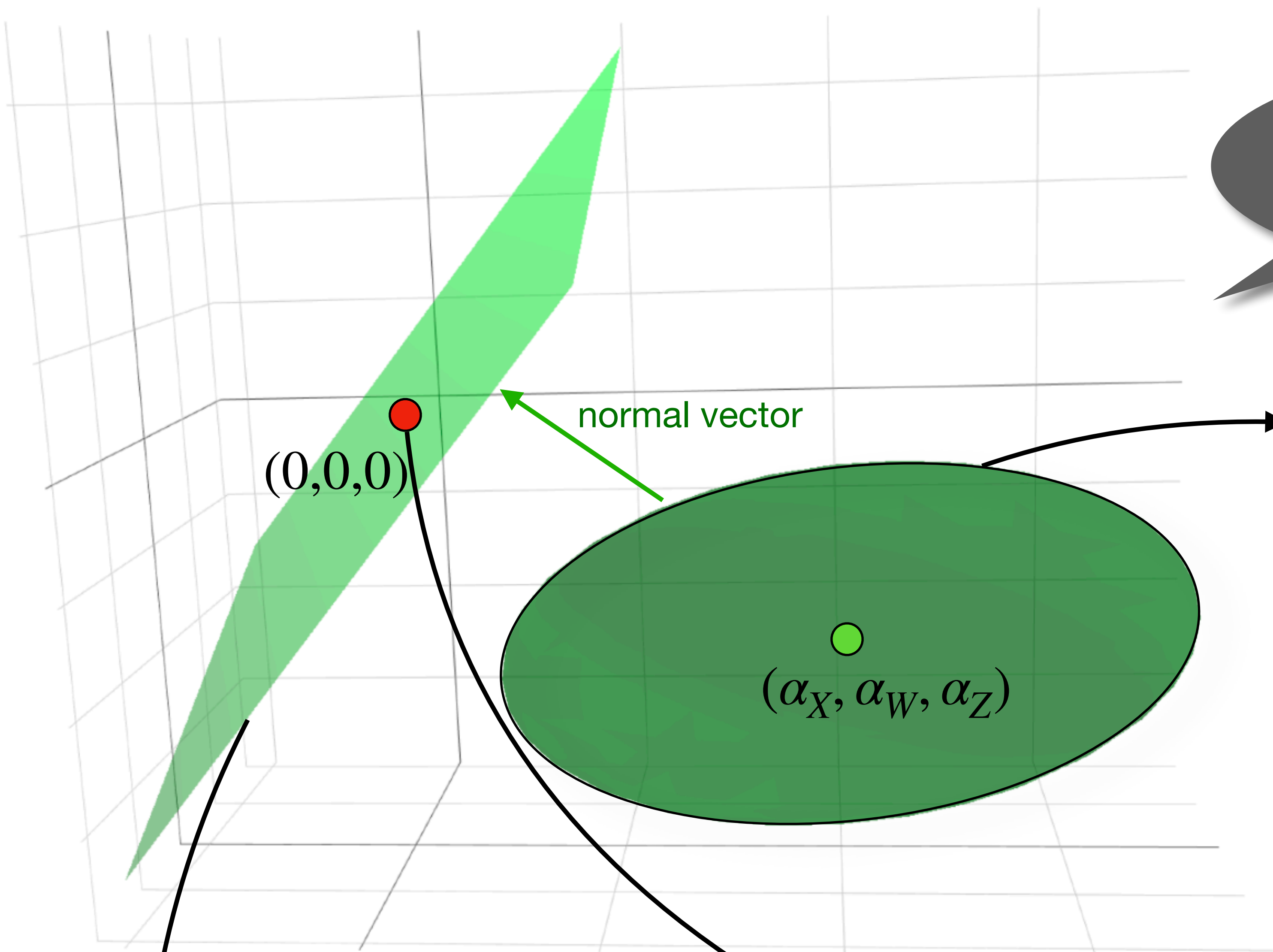
what we actually want

Graph:



FPT visualization

measuring the probability of ellipsoid hitting a point

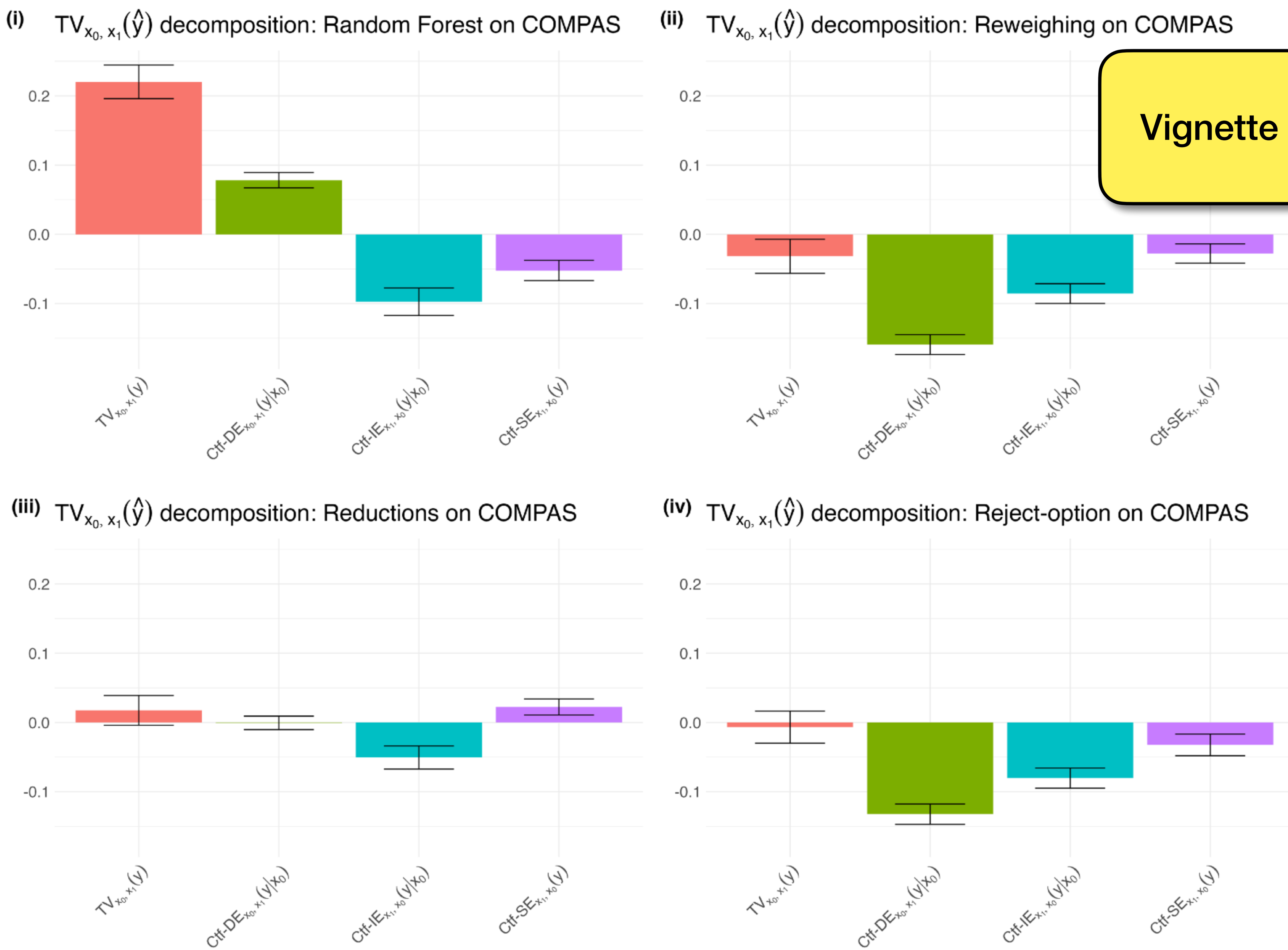


ellipsoid with characteristic matrix $\Sigma = E[VV^T]$

What the constraint is:
$\text{TV}_{x_0,x_1}(\widehat{y}) = 0$
represents a hyperplane through origin.

What we want:
3 linear constraints
Ctf-DE $= 0$, Ctf-IE $= 0$, Ctf-SE $= 0$.
represents a single point

Fair Prediction Theorem in Practice (COMPAS dataset)



Vignette Time!