Fairness and discrimination in actuarial pricing

Arthur Charpentier\textsuperscript{1}, Laurence Barry\textsuperscript{2}, Vincent Grari\textsuperscript{3}, Lamprier Sylvain\textsuperscript{3}, Detyniecki Marcin\textsuperscript{4}

\textsuperscript{1} Univeristé du Québec à Montréal \textsuperscript{2} Chaire Pari \textsuperscript{3} Sorbonne Université \textsuperscript{4} AXA

CIRM - MLISTRAL - September 2022
Agenda

▶ Charpentier (2022) Insurance: Discrimination, Biases and Fairness, *Institut Louis Bachelier*


▶ “Technology is neither good nor bad; nor is it neutral”, Kranzberg (1986)

▶ “Machine learning won’t give you anything like gender neutrality ‘for free’ that you didn’t explicitely ask for”, Kearns and Roth (2019)
Motivation

- Accuracy: \( \pi(x) = \mathbb{E}_P[Y|X = x] \) (historical probability)
- Fairness: \( \pi^*(x) = \mathbb{E}_{P^*}[Y|X = x] \) (targeted probability)
“at the core of insurance business lies discrimination between risky and non-risky insureds”, Avraham (2017)

\[
\begin{align*}
  x \in \mathcal{X} \subset \mathbb{R}^d &: \text{‘explanatory’ variables} \\
  x_c \in \mathcal{X}_c \subset \mathbb{R}^{d_c} &: \text{car-vehicle ‘explanatory’ variables} \\
  x_c' = m_c(x_c; x) \in \mathbb{R} &: \text{car-vehicle scoring} \\
  x_g \in \mathcal{X}_g \subset \mathbb{R}^{d_g} &: \text{geographic ‘explanatory’ variables} \\
  x_g' = m_g(x_g; x) \in \mathbb{R} &: \text{geographic scoring} \\
  y \in \{0, 1\}, \mathbb{N} \text{ or } \mathbb{R}_+ &: \text{variable of interest} \\
  \hat{y} = m(x, x_c', x_g') &: \text{prediction (or score)}
\end{align*}
\]

Classically, a two-stage approach is considered to create a geographic score

- Fit a model $\hat{m}$ to predict $y$ based on $x$, compute residuals $\hat{\varepsilon}$
- Fit a model $\hat{m}_g(\cdot; x)$ to predict $\hat{\varepsilon}$ based on $x_g$
- Define a score $x_g' = \hat{m}_g(x_g; x)$ (that is function of $x$, too)
## Protected Attributes for Motor Insurance in North America

<table>
<thead>
<tr>
<th>Attribute</th>
<th>CA</th>
<th>HI</th>
<th>GA</th>
<th>NC</th>
<th>NY</th>
<th>MA</th>
<th>PA</th>
<th>FL</th>
<th>TX</th>
<th>AL</th>
<th>ON</th>
<th>NB</th>
<th>NL</th>
<th>QC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>X</td>
<td>X</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Age</td>
<td>X</td>
<td>X</td>
<td>•</td>
<td>X*</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Driving experience</td>
<td>•</td>
<td>X</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Credit history</td>
<td>X</td>
<td>X</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>X</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>X*</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Education</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Occupation</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Employment status</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Marital status</td>
<td>•</td>
<td>X</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Housing situation</td>
<td>X</td>
<td>X</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Address/ZIP code</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Insurance history</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
</tbody>
</table>

Protected Attributes ? Motor insurance in North America

Spatial information and racial bias (redlining)
Defining Group Fairness when $y$ is binary I

$$\begin{align*}
\mathbf{x} &\in \mathcal{X} \subset \mathbb{R}^d : \text{‘explanatory’ variables} \\
p &\in \{0, 1\} : \text{protected variable or sensitive attribute} \\
\mathbf{x}_c &\in \mathcal{X}_c \subset \mathbb{R}^{d_c} : \text{car-vehicle ‘explanatory’ variables} \\
x'_c &= m_c(x_c; x, p) \in \mathbb{R} : \text{car-vehicle scoring} \\
\mathbf{x}_g &\in \mathcal{X}_g \subset \mathbb{R}^{d_g} : \text{geographic ‘explanatory’ variables} \\
x'_g &= m_g(x_g; x, p) \in \mathbb{R} : \text{geographic scoring} \\
y &\in \{0, 1\} : \text{variable of interest (binary)} \\
s &= m(x, p, x'_c, x'_g) : \text{score} \\
\hat{y} &= 1(s > \text{threshold}) : \text{prediction}
\end{align*}$$

**Fairness Through Unawareness**, Kusner et al. (2017)

Protected attribute $p$ is not explicitly used in decision function $\hat{y}$. 
Defining Group Fairness when $y$ is binary II

**Demographic Parity**, (Corbett-Davies et al. (2017), Agarwal (2021))

Decision function $\hat{y}$ satisfies demographic parity if $Y \perp \perp P$, i.e.

$$P[\hat{Y} = y | P = 0] = P[\hat{Y} = y | P = 1], \forall y \text{ or } E[\hat{Y} | P = 0] = E[\hat{Y} | P = 1]$$

We can compare $s(X)$ conditional on $Y$, but also on $P$
Defining Group Fairness when $y$ is binary III

**Equal Opportunity**, Hardt et al. (2016)

True positive parity

$$P[\hat{Y} = 1|P = 0, Y = 1] = P[\hat{Y} = 1|P = 1, Y = 1]$$

or false positive parity

$$P[\hat{Y} = 1|P = 0, Y = 0] = P[\hat{Y} = 1|P = 1, Y = 0]$$
### Defining Group Fairness when $y$ is binary IV

<table>
<thead>
<tr>
<th>Fairness Measure</th>
<th>Authors</th>
<th>Condition</th>
<th>Independence/Equivalence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>statistical parity</strong></td>
<td>Dwork et al. (2012)</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p] = \text{cst}, \forall p$</td>
</tr>
<tr>
<td><strong>conditional statistical parity</strong></td>
<td>Corbett-Davies et al. (2017)</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p, X = x] = \text{cst}_x, \forall p, y$ $\hat{Y} \perp P$</td>
</tr>
<tr>
<td><strong>equalized odds</strong></td>
<td>Hardt et al. (2016)</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p, Y = y] = \text{cst}_y, \forall p, y$</td>
</tr>
<tr>
<td><strong>equalized opportunity</strong></td>
<td>Hardt et al. (2016)</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p, Y = 1] = \text{cst}, \forall p$</td>
</tr>
<tr>
<td><strong>predictive equality</strong></td>
<td>Corbett-Davies et al. (2017)</td>
<td>$P[\hat{Y} = 1</td>
<td>P = p, Y = 0] = \text{cst}, \forall p$</td>
</tr>
<tr>
<td><strong>balance (positive)</strong></td>
<td>Kleinberg et al. (2017)</td>
<td>$E[S</td>
<td>P = p, Y = 1] = \text{cst}, \forall p$</td>
</tr>
<tr>
<td><strong>balance (negative)</strong></td>
<td>Kleinberg et al. (2017)</td>
<td>$E[S</td>
<td>P = p, Y = 0] = \text{cst}, \forall p$</td>
</tr>
<tr>
<td><strong>conditional accuracy equality</strong></td>
<td>Berk et al. (2017)</td>
<td>$P[Y = y</td>
<td>P = p, \hat{Y} = y] = \text{cst}_y, \forall p, y$</td>
</tr>
<tr>
<td><strong>predictive parity</strong></td>
<td>Chouldechova (2017)</td>
<td>$P[Y = 1</td>
<td>P = p, \hat{Y} = 1] = \text{cst}, \forall p$</td>
</tr>
<tr>
<td><strong>calibration</strong></td>
<td>Chouldechova (2017)</td>
<td>$P[Y = 1</td>
<td>P = p, S = s] = \text{cst}_s, \forall p, s$</td>
</tr>
<tr>
<td><strong>well-calibration</strong></td>
<td>Chouldechova (2017)</td>
<td>$P[Y = 1</td>
<td>P = p, S = s] = s, \forall p, s$</td>
</tr>
<tr>
<td><strong>accuracy equality</strong></td>
<td>Berk et al. (2017)</td>
<td>$P[\hat{Y} = Y</td>
<td>P = p] = \text{cst}, \forall p$</td>
</tr>
<tr>
<td><strong>treatment equality</strong></td>
<td>Berk et al. (2017)</td>
<td>$\frac{FN_p}{FP_p} = \text{cst}_p, \forall p$</td>
<td></td>
</tr>
</tbody>
</table>
Dependence measures I

Group fairness is characterized by independence or conditional independence properties. Given two random variables $U$ and $V$, 

$$C(U, V) = \begin{cases} 
\text{corr}[U, V] \text{ Pearson's correlation} \\
\text{corr}[F_U(U), F_V(V)] \text{ Spearman's rank correlation} \\
\tau[U, V] \text{ Kendall's tau}
\end{cases}$$

that can be extended to conditional measures, as Lawrance (1976), since

$$\text{corr}[U, V] = \mathbb{E}[UV] \text{ when } \begin{cases} 
\mathbb{E}[U] = \mathbb{E}[V] = 0 \\
\mathbb{E}[U^2] = \mathbb{E}[V^2] = 1
\end{cases}$$

- **Demographic Parity**: $\hat{Y} \perp P \implies C(\hat{Y}, P) = 0$
- **Equalized Odds**: $\hat{Y} \perp P|Y \implies C(\hat{Y}, P|Y = y) = 0, \forall y$
Dependence measures II

Hirschfeld (1935), Gebelein (1941) and Rényi (1959)

\[ HGR(U, V) = \max \, \{ \text{corr}[f(U), g(V)] \} = \max_{f \in S_U, g \in S_V} \{ \mathbb{E}[f(U)g(V)] \} \]

where \( S_U = \{ f : U \to \mathbb{R} : \mathbb{E}[f(U)] = 0 \text{ and } \mathbb{E}[f(U)^2] = 1 \} \) and similarly \( S_V \).

One can also consider a conditional version,

\[ HGR(U, V|Z) = \max_{f \in S_U|Z, g \in S_V|Z} \left\{ \mathbb{E}[f(U)g(V)|Z] \right\} \]

where \( S_U|Z = \{ f : U \to \mathbb{R} : \mathbb{E}[f(U)|Z] = 0 \text{ and } \mathbb{E}[f(U)^2|Z] = 1 \} \).

This measure can be used to characterize independence,

\[ U \perp \perp V \iff HGR(U, V) = 0, \]

and if \((U, V)\) is a Gaussian vector, \( HGR(U, V) = |\text{corr}(U, V)| \).
Thus, this measure can be used to characterize fairness,

\[
\begin{align*}
\text{Demographic Parity} & : \hat{Y} \perp P \iff HGR(\hat{Y}, P) = 0 \\
\text{Equalized Odds} & : \hat{Y} \perp P | Y \iff HGR(\hat{Y}, P | Y = y) = 0, \quad \forall y
\end{align*}
\]

\(HGR\) can be difficult to estimate, but one can use some Neural Networks

\[
HGR_{NN}(U, V) = \max_{\omega_f, \omega_g} \{ \mathbb{E}[f_\omega(U)g_\omega(V)] \}
\]

See also Breiman and Friedman (1985) on the estimation of this maximal correlation, in the context of regression.
Dependence measures IV

More generally, \( \mathbf{V} \) can be a vector on \( \mathcal{V} \subset \mathbb{R}^k \), then

\[
HGR(U, \mathbf{V}) = \max_{h: \mathcal{V} \to \mathbb{R}} \{ HGR[U, h(\mathbf{V})] \} = \max_{f \in \mathcal{S}_U, g \in \mathcal{S}_V} \{ \mathbb{E}[f(U)g(\mathbf{V})] \}
\]

where \( \mathcal{S}_V = \{ g : \mathcal{V} \to \mathbb{R} : \mathbb{E}[g(\mathbf{V})] = 0 \text{ and } \mathbb{E}[g(\mathbf{V})^2] = 1 \} \). A conditional version exists, and one can estimate that measure using a neural network,

\[
HGR_{NN}(U, \mathbf{V}) = \max_{\omega_f, \omega_g} \{ \mathbb{E}[f_{\omega_f}(U)g_{\omega_g}(\mathbf{V})] \}
\]

Demographic Parity: \( \hat{Y} \perp \perp P \iff HGR(\hat{Y}, P) = 0 \)

Equalized Odds: \( \hat{Y} \perp \perp P|Y \iff HGR(\hat{Y}, P|Y = y) = 0, \forall y \)
In a classical ML or econometric pricing model, solve

$$\arg\min_{\theta} \{ \mathcal{L}(\hat{y}, y) \}, \text{ where } \mathcal{L}(\hat{y}, y) = \sum_{i=1}^{n} \ell(\hat{y}_i, y_i) \text{ and } \hat{y} = h_{\theta}(x)$$

either related to some loss, or some log-likelihood,

To avoid over-fit: penalize complexity (penalty $\mathcal{P}$)

$$\arg\min_{\theta} \{ \mathcal{L}(h_{\theta}(x), y) + \lambda \mathcal{P}(h_{\theta}) \}$$
Adversarial Approach II

Inspired by Goodfellow et al. (2018) (but also Bechavod and Ligett (2017) or Cho et al. (2020)), to avoid un-fairness: penalize according to $\text{HGR}(\hat{y}, p)$ (for demographic parity),

$$ \argmin_{\theta, \omega_f, \omega_g} \{ \mathcal{L}(h_\theta(x), y) + \lambda \text{HGR}_{\omega_f, \omega_g}(h_\theta(x), p) \} $$

i.e.

$$ \argmin_{\theta} \left\{ \max_{\omega_f, \omega_g} \{ \mathcal{L}(h_\theta(X), Y) + \lambda \mathbb{E}_{(X, S) \sim D}(\hat{f}_{\omega_f}(h_\theta(X))\hat{g}_{\omega_g}(P)) \} \right\} $$

or $\text{HGR}(\hat{y}, p|y)$ (for equalized odds), i.e. when $y \in \{0, 1\}$

$$ \argmin_{\theta} \left\{ \max_{\omega_{f0}, \omega_{g0}, \omega_{f1}, \omega_{g1}} \{ \mathcal{L}(h_\theta(X), Y) + \lambda_0 \mathbb{E}_{(X, P) \sim D_0}(\hat{f}_{\omega_{f0}}(h_\theta(X))\hat{g}_{\omega_{g0}}(P)) ight\} \right. $$

$$ + \left. \lambda_1 \mathbb{E}_{(X, P) \sim D_1}(\hat{f}_{\omega_{f1}}(h_\theta(X))\hat{g}_{\omega_{g1}}(P)) \} \right\} $$

or, more generally when $y \in \Omega_Y$ (e.g. $\{0, 1, 2, 3+\}$), if $k = \#\Omega_Y$
Adversarial Approach III

\[
\arg\min_{\theta} \left\{ \max_{\omega_0,\omega_g,\ldots,\omega_{f_k},\omega_{g_k}} \left\{ \mathcal{L}(h_{\theta}(X), Y) + \sum_{y \in \Omega_y} \lambda_y E(X,P)\sim_D \mathbb{E}(h_{\theta}(X)g_{\omega_y}(P)) \right\} \right\}
\]

\[
\frac{\partial \mathcal{L}(h_{w_h}(X), Y)}{\partial w_h} + \lambda \frac{\partial E(\hat{f}_{w_f}(h_{w_h}(X))\hat{g}_{w_g}(P))}{\partial w_h} - \frac{\partial E(\hat{f}_{w_f}(h_{w_h}(X))\hat{g}_{w_g}(P))}{\partial w_f} = E(\hat{f}_{w_f}(h_{w_h}(X))\hat{g}_{w_g}(P))
\]
Dealing with high dimension I

- geographic / spatial information, $X_g$
- car type / make / model, $X_c$
- other classical ratemaking variables, $X$ (non protected)

Some features can be in high dimension, natural solution would be PCA or autoencoders (see Shi and Shi (2021) about feature embedding in high dimension).
Dealing with high dimension II

Principal Component Analysis (PCA)

\[
\min_{P \in \Pi} \left\{ ||X - P^T PX||_F^2 \right\} \text{ s.t. rank}(P) = k
\]

where \( \Pi \) is the set of projection matrices.

\[
\begin{align*}
\min ||X - X'||^2 &= \min ||X - P^T PX||^2 \\
&= \min \sum_{i=1}^{n} \left( P^T P x_i - x_i \right)^T \left( P^T P x_i - x_i \right)
\end{align*}
\]
Dealing with high dimension III

Autoencoder

\[
\min_{\psi} \left\{ \| X - \psi \circ \varphi X \|_F^2 \right\}
\]

\[
X \in \mathcal{X} \quad \quad Z \in \mathcal{Z} \quad \quad X' \in \mathcal{X}
\]

\[
\min \| X - X' \|^2 = \min \| X - \psi \circ \varphi (X) \|^2
\]

\[
\min \sum_{i=1}^{n} (\psi \circ \varphi (x_i) - x_i)^\top (\psi \circ \varphi (x_i) - x_i)
\]

"bottleneck" hidden layer

encoder

decoder
Dealing with high dimension IV

\[ \mathcal{L}(h_w(X_p, \gamma_w(X_g), c_w(X_c)), Y) + \lambda \text{HGR}(\bar{Y}, P) \]

\[ d\mathcal{L}(h_w(X_p, \gamma_w(X_g), c_w(X_c)), Y) + \lambda \text{HGR}(\bar{Y}, P) \]

\[ \text{Adversary } \hat{f} \quad \text{Weight: } w_f \]

\[ \text{Adversary } \hat{g} \quad \text{Weight: } w_g \]

\[ \text{HGR}(\bar{Y}, P) = E(\hat{f}_{w_f}(h_w(X_p, C, G))\hat{g}_{w_g}(P)) \]

\[ \partial E(\hat{f}_{w_f}(h_w(X_p, \gamma_w(X_g), c_w(X_c)))\hat{g}_{w_g}(P)) \]

\[ \partial E(\hat{f}_{w_f}(h_w(X_p, \gamma_w(X_g), c_w(X_c))\hat{g}_{w_g}(P)) \]

\[ \partial E(\hat{f}_{w_f}(h_w(X_p, \gamma_w(X_g), c_w(X_c))\hat{g}_{w_g}(P)) \]
Application on synthetic data I

- S: sensitive / protected (gender)
- Col: color of the car
- Sp: maximum speed of the car
- Sal: average salary of the policyholders area
- Age: age of the driver
- Ina: inattention
- Agg: aggressivity
- Y: total cost
Application on synthetic data II

Performance Accuracy vs Fairness

Dependence between Car risk and $S$ by $\lambda$

Dependence between Predictions and Color by $\lambda$

- $\lambda$: fairness penalty
- $p$-rule: $\min \left\{ \frac{\mathbb{P}(\hat{Y} = 1|P = 1)}{\mathbb{P}^*(\hat{Y} = 1|P = 0)}, \frac{\mathbb{P}(\hat{Y} = 1|P = 0)}{\mathbb{P}^*(\hat{Y} = 1|P = 1)} \right\}$
Application on real data (pricing game 2015) I

\[ y \in \{0, 1\} \text{ (claim occurrence), and } p \text{ is the (binary) gender} \]
Application on real data (pricing game 2015) II

\[ y \in \{0, 1, 2+\} \text{ (claim frequency), and } p \text{ is the (binary) gender} \]

\[ y \in \{0, 1, 2+\} \text{ (claim frequency), and } p \text{ is the (binary) gender} \]

see Grari et al. (2022) for more examples (including the case where \( y \in \mathbb{R}^+ \))
$y \in \{0, 1, 2+\}$ (claim frequency), and $p$ is spatial information (redlining)
From correlation to causality I

- “classifying projection methods as using demographic/actuarial models or non-demographic/causal models”
  Keilman (2003) and Hudson (2007)

- “Article 5(2) allowed Member States to Permit proportionate differences in individuals premiums and benefits where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data.”
  Thiery and Van Schoubroeck (2006) and Schmeiser et al. (2014)

- “Two judges on the Supreme Court dissented in the Zurich case. In their view, an insurer must not only prove a statistical correlation between a particular group and higher risk, but a causal connection”
  Gomery et al. (2011)
From correlation to causality II

DAGs are important

Looking for a counterfactual

Alberta man changes gender on government IDs for cheaper car insurance

He says he saved almost $1,100

@freakonometrics  ❧  freakonometrics  ❧  freakonometrics.hypotheses.org
Consider some distances $D$ on $\{0,1\} \times \{0,1\}$ or $[0,1] \times [0,1]$, and $d$ on $\mathbb{R}^p \times \mathbb{R}^p$.

**Lipschitz property**, Duivesteijn and Feelders (2008)

$$D(\hat{y}_i, \hat{y}_j) \text{ or } D(s_i, s_j) \leq d(x_i, x_j), \ \forall i, j = 1, \ldots, n.$$ 

**Counterfactual fairness**, Kusner et al. (2017) If the prediction in the real world is the same as the prediction in the counterfactual world where the individual would have belonged to a different demographic group, we have counterfactual equity, i.e.

$$\mathbb{P}[Y^*_{P \leftarrow p} = y \mid X = x] = \mathbb{P}[Y^*_{P \leftarrow p'} = y \mid X = x], \ \forall p', x, y.$$
From correlation to causality IV

- counterfactuals
  *(what if I had done...?)*
- intervention
- association
  *(what if I see...?)*

what would have happened if this person had had treatment 1 instead of treatment 0?

(picture Pearl & Mackenzie (2018))
From correlation to causality V

Causal inference literature,

- \( t \) some binary treatment (\( t \in \{0, 1\} \))
- \( x \) some covariates
- \( y \) denote the observed outcome, \( y_{i, T \leftarrow 1}^* \) and \( y_{i, T \leftarrow 0}^* \) the potential outcomes

<table>
<thead>
<tr>
<th>treatment</th>
<th>outcome</th>
<th>age</th>
<th>gender</th>
<th>height</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_i )</td>
<td>( y_i )</td>
<td>( y_{i, T \leftarrow 1}^* )</td>
<td>( y_{i, T \leftarrow 0}^* )</td>
<td>( x_{1,i} )</td>
<td>( x_{2,i} )</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>121</td>
<td>?</td>
<td>37</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>109</td>
<td>?</td>
<td>28</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>162</td>
<td>?</td>
<td>53</td>
<td>M</td>
</tr>
</tbody>
</table>

There will be a significant impact of treatment \( t \) on \( y \) if \( y_{T \leftarrow 0}^* \neq y_{T \leftarrow 1}^* \) (see Rubin (1974), Hernán and Robins (2010) or Imai (2018)).

The causal effect for individual \( i \) is \( \tau_i = y_{i, T \leftarrow 1}^* - y_{i, T \leftarrow 0}^* \)
From correlation to causality VI

One can define the sample average treatment effect (SATE)

\[
\text{SATE} = \frac{1}{n} \sum_{i=1}^{n} y_{i,T\leftarrow 1} - y_{i,T\leftarrow 0}
\]

the average treatment effect (ATE)

\[
\tau = \text{ATE} = E[Y_{i,T\leftarrow 1} - Y_{i,T\leftarrow 0}]
\]

and, for possibly heterogeneous effects, conditional average treatment effect (CATE)

\[
\tau(x) = \text{CATE}(x) = E[Y_{i,T\leftarrow 1} - Y_{i,T\leftarrow 0}|X = x]
\]


