# Optimal transport and fairness of predictive models

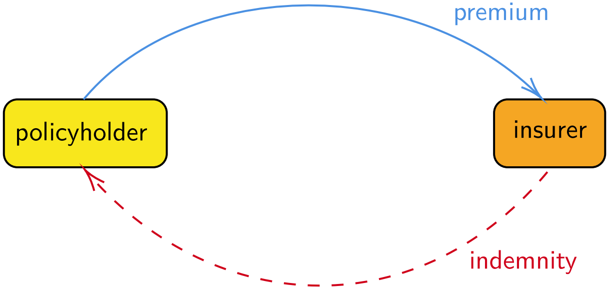
**Arthur Charpentier**, François Hu, Agathe Fernandes-Machado & Philipp Ratz

SCAI (Sorbonne Center for Artificial Intelligence), September 2024

SCOR FOUNDATION FOR SCIENCE

# Insurance (and “Actuarial Fairness”)

- Insurance is a **risk transfer** (from a policyholder to an insurance company)



premium

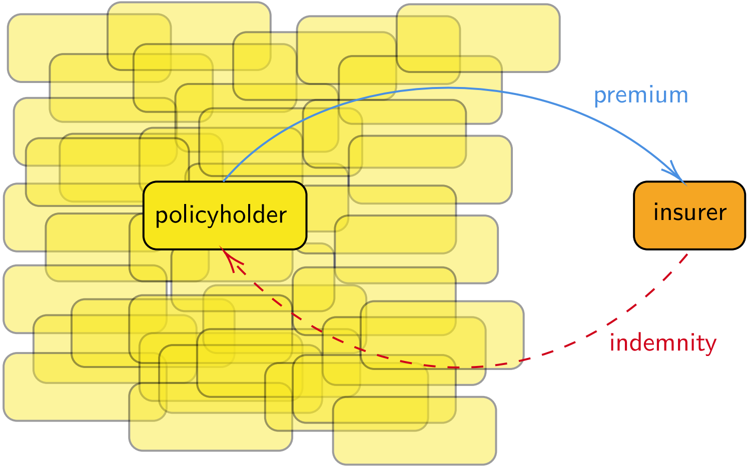
policyholder

insurer

indemnity

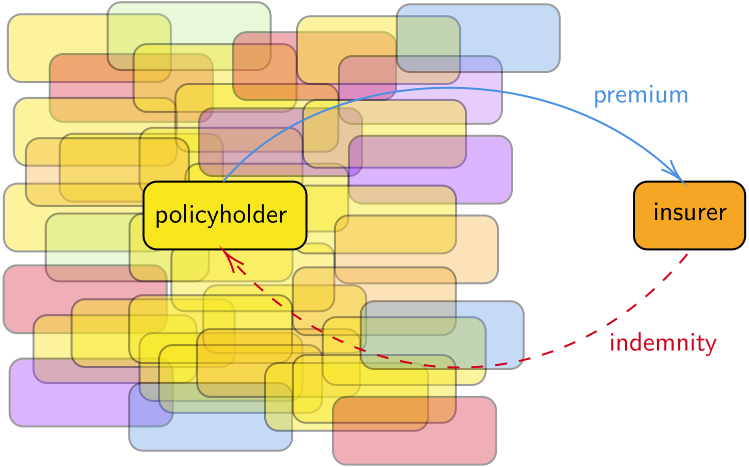
# Insurance (and “Actuarial Fairness”)

- “*Insurance is the contribution of the many to the misfortune of the few*”



# Insurance (and "Actuarial Fairness")

- "*Insurance is the contribution of the many to the misfortune of the few*"



# Motivation (1. Legal Aspects)

- EU Directive (2004/113/EC), 2004 version

– Article 5 (Actuarial factors) –

1. Member States shall ensure that in all new contracts concluded after 21 December 2007 at the latest, the use of sex as a factor in the calculation of premiums and benefits for the purposes of insurance and related financial services shall not result in differences in individuals' premiums and benefits.

2. Notwithstanding paragraph 1, Member States may decide before 21 December 2007 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. The Member States concerned shall inform the Commission and ensure that accurate data relevant to the use of sex as a determining actuarial factor are compiled, published and regularly updated.

# Motivation (1. Legal Aspects)

- Au Québec, Charte des droits et libertés de la personne (C-12)

– Article 20.1 –

In an insurance or pension contract, a social benefits plan, a retirement, pension or insurance plan, or a public pension or public insurance plan, a distinction, exclusion or preference based on age, sex or civil status is deemed non-discriminatory where the use thereof is warranted and the basis therefor is a risk determination factor based on actuarial data

# Motivation (1. Legal Aspects)

- September 27, 2023, the Colorado Division of Insurance exposed a new proposed regulation entitled Concerning Quantitative Testing of External Consumer Data and Information Sources, Algorithms, and Predictive Models Used for Life Insurance Underwriting for Unfairly Discriminatory Outcomes

– Section 5 (Estimating Race and Ethnicity) –

Insurers shall estimate the race or ethnicity of all proposed insureds that have applied for coverage on or after the insurer's initial adoption of the use of ECDIS, or algorithms and predictive models that use ECDIS, including a third party acting on behalf of the insurer that used ECDIS, or algorithms and predictive models that used ECDIS, in the underwriting decision-making process, by utilizing: BIFSG and the insureds' or proposed insureds' name and geolocation (...)

- Bayesian Improved First Name Surname Geocoding, or "BIFSG"
- External Consumer Data and Information Source, or "ECDIS"

# Motivation (1. Legal Aspects)

- EU Directive (2010/41/EU), 2010 version (on the application of the principle of equal treatment between men and women)

– Article 3 (Definition) –

(a) ‘direct discrimination’: where one person is treated less favourably on grounds of sex than another is, has been or would be, treated in a comparable situation;
(b) ‘indirect discrimination’: where an apparently neutral provision, criterion or practice would put persons of one sex at a particular disadvantage compared with persons of the other sex, unless that provision, criterion or practice is objectively justified by a legitimate aim, and the means of achieving that aim are appropriate and necessary;

# Motivation (1. Legal Aspects)

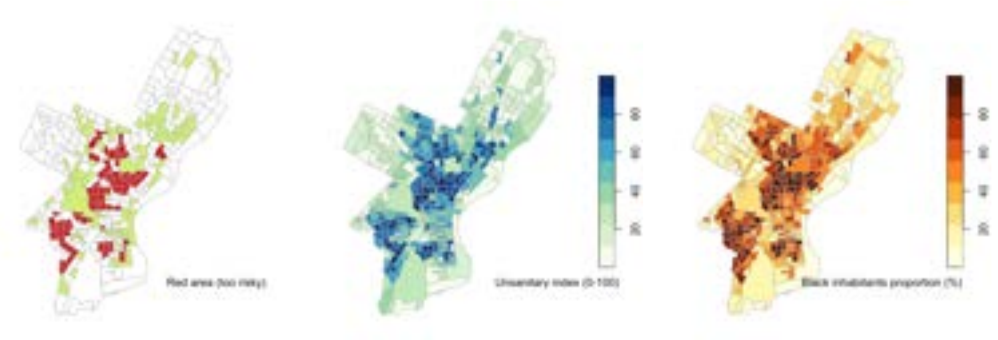
- In France, Loi n° 2008-496 du 27 mai 2008

– Article 1 –

Constitue une discrimination indirecte une disposition, un critère ou une pratique neutre en apparence, mais susceptible d'entraîner, pour l'un des motifs mentionnés au premier alinéa, un désavantage particulier pour des personnes par rapport à d'autres personnes, à moins que cette disposition, ce critère ou cette pratique ne soit objectivement justifié par un but légitime et que les moyens pour réaliser ce but ne soient nécessaires et appropriés.

Extension of "Loi n° 72-546 du 1 juillet 1972", which removed the requirement for specific intent.

# Motivation (2. Redlining)



(Fictitious maps, inspired by a Home Owners' Loan Corporation map from 1937)

- Federal Home Loan Bank Board (FHLBB) "*residential security maps*" (for real-estate investments), Crossney (2016) and Rhynhart (2020)
- Unsanitary index and proportion of Black inhabitants
- Discrimination as an "**ill-posed problem**"?

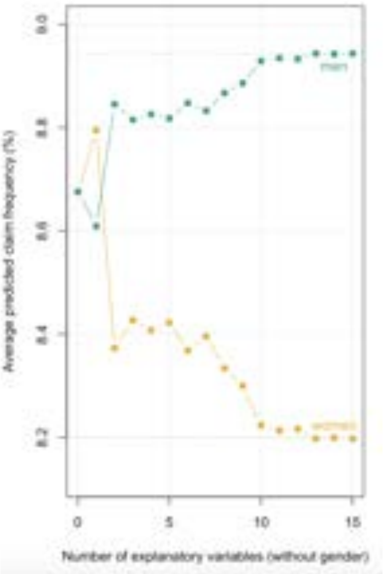
# Motivation (3. Proxies)

- On a French motor dataset, average claim frequencies are 8.94% (men) 8.20% (women).

- Consider some logistic regression to estimate annual claim frequency, on $k$ explanatory variables excluding gender.

| | men | women |
|---|---|---|
| $k = 0$ | 8.68% | 8.68% |
| $k = 2$ | 8.85% | 8.37% |
| $k = 8$ | 8.87% | 8.33% |
| $k = 15$ | 8.94% | 8.20% |
| empirical | 8.94% | 8.20% |

- Models simply tend to reproduce what was observed in the data (see "**is-ought**" problem, in Hume (1739)).



# Discrimination and Insurance

"*Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitly ask for*," Kearns and Roth (2019)

"*What is unique about insurance is that even statistical discrimination which by definition is absent of any malicious intentions, poses significant moral and legal challenges. Why? Because on the one hand, policy makers would like insurers to treat their insureds equally, without discriminating based on race, gender, age, or other characteristics, even if it makes statistical sense to discriminate* (...) *On the other hand, at the core of insurance business lies discrimination between risky and non-risky insureds. But riskiness often statistically correlates with the same characteristics policy makers would like to prohibit insurers from taking into account.* " Avraham (2017)

"*Technology is neither good nor bad; nor is it neutral*," Kranzberg (1986)

# Fairness for Classifiers

$$\begin{cases} \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{‘explanatory’ variables} \\ s \in \{A, B\} : \text{"sensitive variable"} \\ y \in \{0,1\} : \text{classification problem} \\ \widehat{y} \in \{0,1\} : \text{prediction, classically } \overset{\text{class } \in \{0,1\}}{\widehat{y}} = \boldsymbol{1}(\underset{\text{score } \in [0,1] \subset \mathbb{R}}{m(\boldsymbol{x}, s)} > t) \end{cases}$$

Following Barocas et al. (2017), standard definitions are

A model $m$ satisfies the **independence property** if $m(\boldsymbol{X}, S) \perp\!\!\!\perp S$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$ ← demographic parity

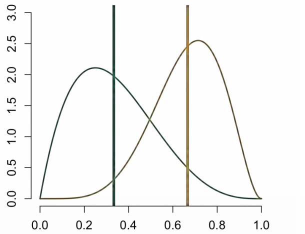
A model satisfies the **separation property** if $m(\boldsymbol{X}, S) \perp\!\!\!\perp S \mid Y$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$ ← equalized odds

A model satisfies the **sufficiency property** if $Y \perp\!\!\!\perp S \mid m(\boldsymbol{X}, S)$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$ ← calibration

# Fairness for Classifiers

(weak) definition of “demographic parity” for a classifier

$$\mathbb{E}[\underset{\text{score}}{m(\boldsymbol{X},S)} \mid \overset{\text{sensitive}}{S=A}] \overset{?}{=} \mathbb{E}[\underset{\text{score}}{m(\boldsymbol{X},S)} \mid \overset{\text{sensitive}}{S=B}]$$



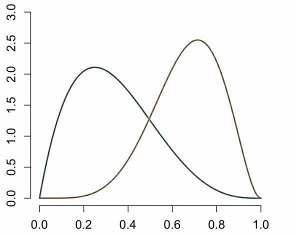
(**strong**) definition of “demographic parity” for a classifier

$$\mathbb{P}[m(\boldsymbol{X},S)\leq u \mid S=A] \overset{?}{=} \mathbb{P}[m(\boldsymbol{X},S)\leq u \mid S=B]$$

$\forall u\in[0,1]$, or $F_A \overset{?}{=} F_B$ where

$$F_A(u)=\mathbb{P}[m(\boldsymbol{X},S)\leq u \mid S=A]$$

$$F_B(u)=\mathbb{P}[m(\boldsymbol{X},S)\leq u \mid S=B]$$



# Formalizing Optimal Transport

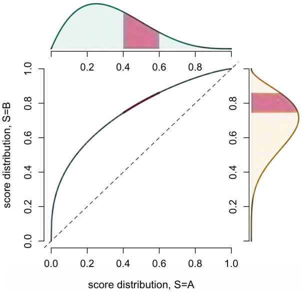
Consider the following $[0,1] \to [0,1]$ mapping

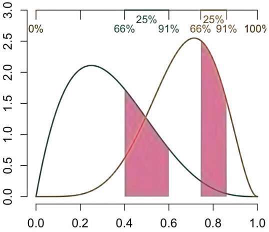
$$T^\star(x) = F_B^{-1} \circ F_A(x)$$

$$T^\star = \underset{T:[0,1]\to[0,1]}{\text{argmin}} \int_0^1 \big( T(x) - x \big)^2 dF_A(x)$$

i.e. $\underset{T:[0,1]\to[0,1]}{\text{argmin}} \; \mathbb{E}\big[\big( \underbrace{T(X)}_{Y \text{ with } Y \sim F_B} - X \big)^2\big]$ where $X \sim F_A$,

corresponding to Monge (1781) problem,
revisited by Kantorovich (1942).
(the minimum value is called **Wasserstein distance**)





# Optimal Transport with a Finite Sample (another interpretation)

$m_1^B \leq m_2^B \leq \cdots \leq m_n^B$

$m_1^A \leq m_2^A \leq \cdots \leq m_n^A$

Consider two samples, $(m(\boldsymbol{x}_i, s_i = A))$ and $(m(\boldsymbol{x}_i, s_i = B))$

$m_1^A \leq m_2^A \leq \cdots \leq m_n^A$ and $m_1^B \leq m_2^B \leq \cdots \leq m_n^B$

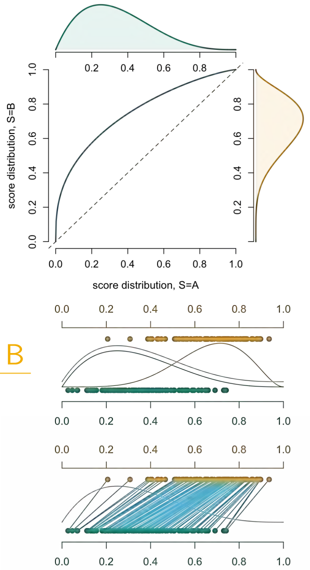
$m$ is not fair with respect to $s$ if $T^\star(x) \neq x$, or $m_i^A \neq m_i^B$

optimal transport mapping

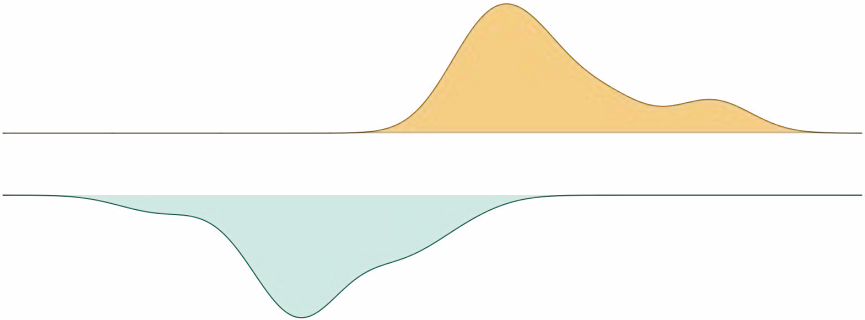
quantile of level $p$ in group B

$$T^\star(x) = F_B{}^{-1} \circ F_A(x) \neq x$$

probability $p$ associated with $u$ in group A



# “Optimal Transport” (a side note / a cultural interlude)

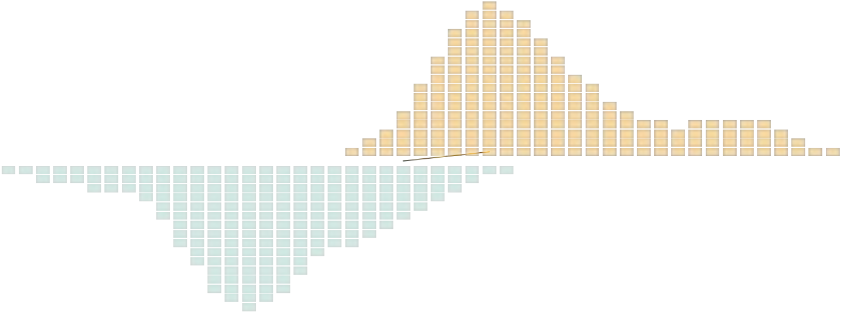


pile of sand

hole / excavation site

Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

# “Optimal Transport” (a side note / a cultural interlude)



This “monotone” (increasing) mapping is optimal $x_i^A \xrightarrow{T^\star} y_i^B$

# Mitigating Discrimination with Wasserstein Barycenters

Mitigation is about finding some $m^\star$ "in-between" (Demographic Parity)

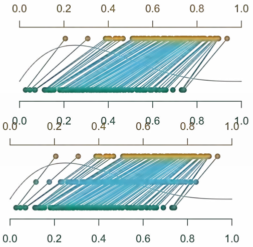
For individual $i$, why not

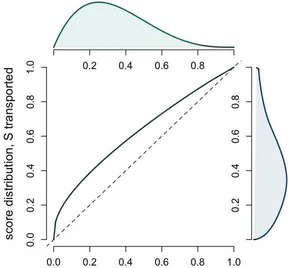
$$m_i^\star = \frac{1}{2} m_i^A + \frac{1}{2} m_i^B$$

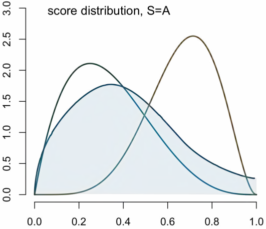
corresponding to

$$m^\star(x, A) = \frac{1}{2} m^A(x) + \frac{1}{2} T^\star(m^A(x))$$

$\mathbb{P}[S = A]$ $\mathbb{P}[S = B]$ associated score in group B





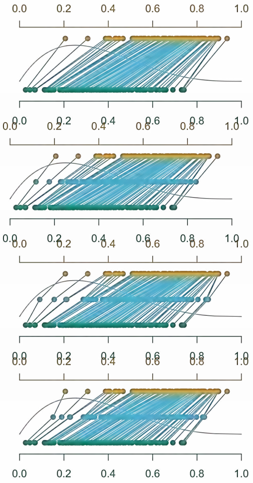


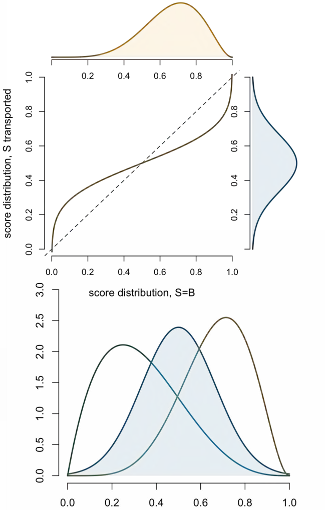
# Mitigating Discrimination with Wasserstein Barycenters

Mitigation is about finding some $m^\star$ “in-between” (Demographic Parity)

other “averages” could be considered that one (“**Wasserstein barycenter**”) is actually optimal in terms of (empirical) risk

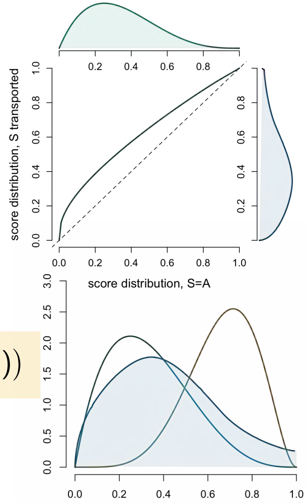
Given a model $m$ (regression, boosting, random forest, neural nets, etc) we can easily derive a “**fair model**”





# Mitigating Discrimination with Wasserstein Barycenters

$$\begin{cases} m^\star(\boldsymbol{x}, s=A) = \mathbb{P}[S=A]\cdot m(\boldsymbol{x}, s=A) \\ \qquad\qquad\qquad + \mathbb{P}[S=B]\cdot F_B^{-1}\circ F_A(m(\boldsymbol{x}, s=A)) \\ m^\star(\boldsymbol{x}, s=B) = \mathbb{P}[S=A]\cdot F_A^{-1}\circ F_B(m(\boldsymbol{x}, s=B)) \\ \qquad\qquad\qquad + \mathbb{P}[S=B]\cdot m(\boldsymbol{x}, s=B). \end{cases}$$



score in group A

$p = F_A(m(\boldsymbol{x}, s=A))$

$$\mathbb{P}[S=A]\cdot m(\boldsymbol{x}, s=A) + \mathbb{P}[S=B]\cdot F_B^{-1}\circ F_A(m(\boldsymbol{x}, s=A))$$

weights

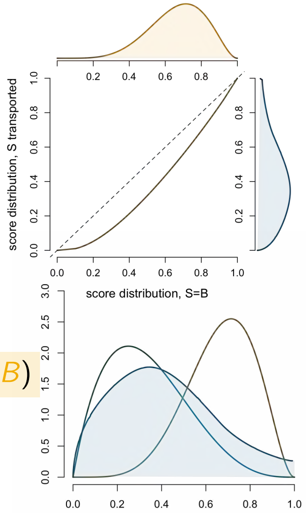
# Mitigating Discrimination with Wasserstein Barycenters

$$\begin{cases} m^\star(\boldsymbol{x}, s=A) = \mathbb{P}[S=A]\cdot m(\boldsymbol{x}, s=A) \\ \qquad\qquad\qquad\quad + \mathbb{P}[S=B]\cdot F_B^{-1}\circ F_A(m(\boldsymbol{x}, s=A)) \\ m^\star(\boldsymbol{x}, s=B) = \mathbb{P}[S=A]\cdot F_A^{-1}\circ F_B(m(\boldsymbol{x}, s=B)) \\ \qquad\qquad\qquad\quad + \mathbb{P}[S=B]\cdot m(\boldsymbol{x}, s=B). \end{cases}$$

$p = F_B(m(\boldsymbol{x}, s=B))$

score in group B

$$\boxed{\mathbb{P}[S=A]}\cdot\boxed{F_A^{-1}\circ\ F_B(m(\boldsymbol{x}, s=A))}+\boxed{\mathbb{P}[S=B]}\cdot\boxed{m(\boldsymbol{x}, s=B)}$$



# Mitigation with Wasserstein Barycenter

We have defined the risk of a model $m \in \mathcal{M}$ as $\mathcal{R}(m) = \mathbb{E}\big[\ell(Y, m(\boldsymbol{X}))\big]$.
Define the classes of fair models,

$$\begin{cases}\mathcal{M}_{\mathrm{DP}} = \{m \in \mathcal{M} \text{ s.t. } m(\boldsymbol{X}) \perp\!\!\!\perp S\} \\ \mathcal{M}_{\mathrm{EO}} = \{m \in \mathcal{M} \text{ s.t. } m(\boldsymbol{X}) \perp\!\!\!\perp S \mid Y\}\end{cases}$$

Fairness is achieved by projection onto a fair subspace

$$\widehat{m}_{\mathrm{fair}} \in \underset{m \in \mathcal{M}_{\mathrm{fair}}}{\operatorname{argmin}} \{\widehat{\mathcal{R}}_n(m)\}$$

Given a risk $\mathcal{R}$, a class $\mathcal{M}$ and the fair-subclass $\mathcal{M}_{\mathrm{fair}}$, the **price of fairness**

$$\mathcal{E}_{\mathrm{fair}}(\mathcal{M}) = \min_{m \in \mathcal{M}_{\mathrm{fair}}} \{\mathcal{R}(m)\} - \min_{m \in \mathcal{M}} \{\mathcal{R}(m)\}.$$

# Mitigation with Wasserstein Barycenter

Recall that Bayes estimator is the best model, for the $\ell_2$ loss,

$$\mu(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}] \text{ and set } \begin{cases} \mu_{\mathtt{A}}(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, S = \mathtt{A}] \\ \mu_{\mathtt{B}}(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, S = \mathtt{B}] \end{cases}$$

From the definition of Wasserstein distance,

$$W_2(p,q) = \left( \inf_{\pi \in \Pi(p,q)} \int |x-y|^2 d\pi(x,y) \right)^{1/2}$$

Thus,

$$\mathbb{E}\big[|m(\boldsymbol{X},S) - \mu_S(\boldsymbol{X})|^2 \big| S = s\big] \geq W_2(\mathbb{P}_m, \mathbb{P}_s)^2$$

# Mitigation with Wasserstein Barycenter

## Price of fairness and Wasserstein Barycenter

$$\mathcal{E}_{\text{fair}}(\mathcal{M}) = \min_{m\in\mathcal{M}_{\text{fair}}}\{\mathcal{R}(m)\} - \min_{m\in\mathcal{M}}\{\mathcal{R}(m)\} \geq \min_{g\in\mathcal{M}}\left\{\mathbb{E}\left(W_2(\mathbb{P}_S,\mathbb{P}_{S,g})^2\right)\right\}$$

where $\mathbb{P}_S$ is the condition distribution of $\mu(\boldsymbol{X},S)$, given $S$, and $\mathbb{P}_{S,g}$ is the condition distribution of $g(\boldsymbol{X},S)$, given $S$. Moreover, if $\mathcal{M}_{\text{fair}} = \mathcal{M}_{\text{DP}}$, and if $\mathbb{P}_s$ is absolutely continuous (w.r.t. Lebesgue measure),

$$\mathcal{E}_{\text{DP}}(\mathcal{M}) = \min_{g\in\mathcal{M}}\left\{\mathbb{E}\left(W_2(\mathbb{P}_S,\mathbb{P}_{S,g})^2\right)\right\} = \min_{g\in\mathcal{M}}\left\{\sum_s \mathbb{P}[S=s]\cdot W_2(\mathbb{P}_s,\mathbb{P}_{s,g})^2\right\}$$

See Gouic et al. (2020) for a complete proof.

We recognize on the right the barycenter, with weights $\mathbb{P}[S=s]$ and distance $W_2$.

# Group Fairness Definitions

weak demographic parity →
$$\mathbb{E}[\,m(\boldsymbol{X},S)\mid \overbrace{S=A}^{\text{sensitive}}\,] \overset{?}{=} \mathbb{E}[\,m(\boldsymbol{X},S)\mid \overbrace{S=B}^{\text{sensitive}}\,]$$
score

strong demographic parity →
$$\mathbb{P}[\,m(\boldsymbol{X},S)\leq u\mid \overbrace{S=A}^{\text{sensitive}}\,] \overset{?}{=} \mathbb{P}[\,m(\boldsymbol{X},S)\leq u\mid \overbrace{S=B}^{\text{sensitive}}\,],\forall u$$

equalized odds →
$$\mathbb{E}[\,m(\boldsymbol{X},S)\mid Y=y,\ \overbrace{S=A}^{\text{sensitive}}\,] \overset{?}{=} \mathbb{E}[\,m(\boldsymbol{X},S)\mid Y=y,\ \overbrace{S=B}^{\text{sensitive}}\,],\forall y$$
score

calibration →
$$\mathbb{E}[\,Y\mid m(\boldsymbol{X},S)=u,\ \overbrace{S=A}^{\text{sensitive}}\,] \overset{?}{=} \mathbb{E}[\,Y\mid m(\boldsymbol{X},S)=u,\ \overbrace{S=B}^{\text{sensitive}}\,],\forall u$$
score

# From Discrimination to Calibration (an Epistemological Detour)

$$\text{calibration} \rightarrow \mathbb{E}[\,Y \mid m(\boldsymbol{X},S)=u\,,\ \overset{\text{sensitive}}{S=\mathrm{A}}\,] \underset{\text{score}}{\overset{?}{=}} \mathbb{E}[\,Y \mid m(\boldsymbol{X},S)=u\,,\ \overset{\text{sensitive}}{S=\mathrm{B}}\,], \forall u$$

Property $\mathbb{E}[\,Y \mid m(\boldsymbol{X},S)=u\,]=u$, $\forall u\in[0,1]$ corresponds to "**calibration**".

"*Out of all the times you said there was a 40 percent chance of rain, how often did rain actually occur? If, over the long run, it really did rain about 40 percent of the time, that means your forecasts were well calibrated*," Silver (2012)

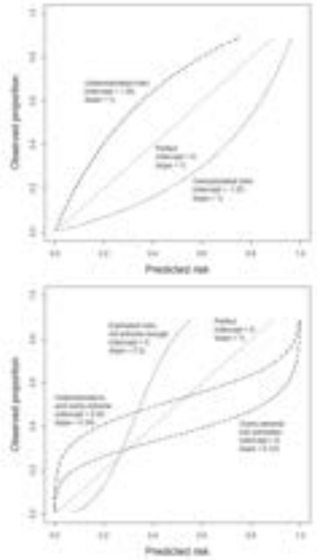
# From Discrimination to Calibration (an Epistemological Detour)

As explained in Van Calster et al. (2019), ”*among patients with an estimated risk of 20%, we expect 20 in 100 to have or to develop the event*,”

- If 40 out of 100 in this group are found to have the disease, the risk is underestimated
- If we observe that in this group, 10 out of 100 have the disease, we have overestimated the risk.

Most machine learning models can be poorly calibrated, Denuit et al. (2021), Machado et al. (2024).

(picture source: Van Calster et al. (2019))

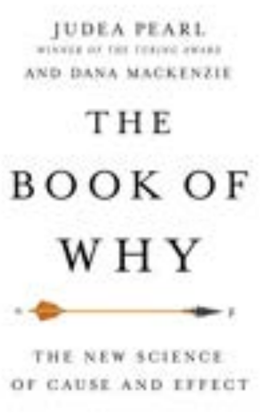


# Individual Fairness

We have **counterfactual fairness** if “*had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same*,” Kusner et al. (2017)

“Ladder of causation” from Pearl et al. (2009), Pearl and Mackenzie (2018)

- 3. **Counterfactuals** (Imagining, “*what if I had done...*”)
- 2. **Intervention** (Doing, “*what if I do...*”)
- 1. **Association** (Seeing, “*what if I see...*”)



JUDEA PEARL
WINNER OF THE TURING AWARD
AND DANA MACKENZIE

THE BOOK OF WHY

THE NEW SCIENCE OF CAUSE AND EFFECT

# Counterfactual Fairness

If the protected variable is considered as the treatment, individual fairness is close a measuring a **treatment effect**.

What does "*other things being equal*" really mean ?

It is possible to suppose that the protected attribute $s$ could affect some explanatory variables $\boldsymbol{x}$ in a non-discriminatory way, Kilbertus et al. (2017) (concept of "revolving variable").

See **ceteris paribus** and **mutatis mutandis CATE**, in Charpentier et al. (2023)

$$\begin{cases}\text{"ceteris paribus CATE"} : \mathbb{E}[Y^\star(B)|\boldsymbol{X}=\boldsymbol{x}] - \mathbb{E}[Y^\star(A)|\boldsymbol{X}=\boldsymbol{x}] \\ \text{"mutatis mutandis CATE"} : \mathbb{E}[Y^\star(B)|\boldsymbol{X}=\boldsymbol{x}^\star(B)] - \mathbb{E}[Y^\star(A)|\boldsymbol{X}=\boldsymbol{x}]\end{cases}$$

suggested also in Plečko and Meinshausen (2020),Plečko et al. (2021) and De Lara et al. (2024). We need to transport $\boldsymbol{X}|S=A$ to $\boldsymbol{X}|S=B$ (multivariate transport).

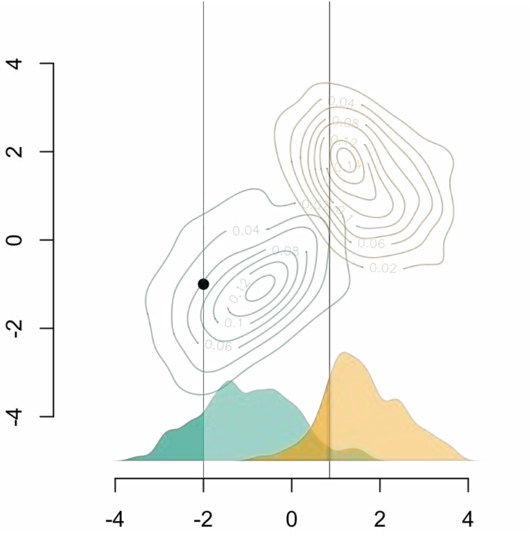
# Counterfactual Fairness

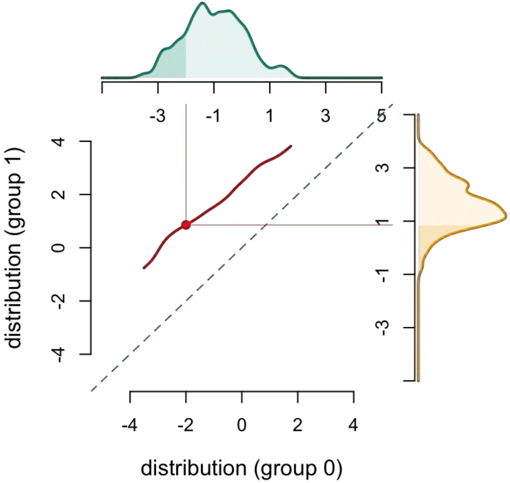
As explained in Villani (2003); Carlier et al. (2010); Bonnotte (2013), the Knothe-Rosenblatt rearrangement is directly inspired by the Rosenblatt chain rule, from Rosenblatt (1952), and some extensions obtained on general measures by Knothe (1957). The **Knothe-Rosenblatt rearrangement** is

$$T_{\overline{kr}}(x_1,\cdots,x_d)=\begin{pmatrix}T^\star_{\overline{1}}(x_1|x_2,\cdots,x_d)\\T^\star_{\overline{2}}(x_2|x_3,\cdots,x_d)\\\vdots\\T^\star_{\overline{d-1}}(x_{d-1}|x_d)\\T^\star_{\overline{d}}(x_d)\end{pmatrix}\text{ or }T_{\underline{kr}}(x_1,\cdots,x_d)=\begin{pmatrix}T^\star_{\underline{1}}(x_1)\\T^\star_{\underline{2}}(x_2|x_1)\\\vdots\\T^\star_{\underline{d-1}}(x_{d-1}|x_1,\cdots,x_{d-2})\\T^\star_{\underline{d}}(x_d|x_1,\cdots,x_{d-1})\end{pmatrix}.$$

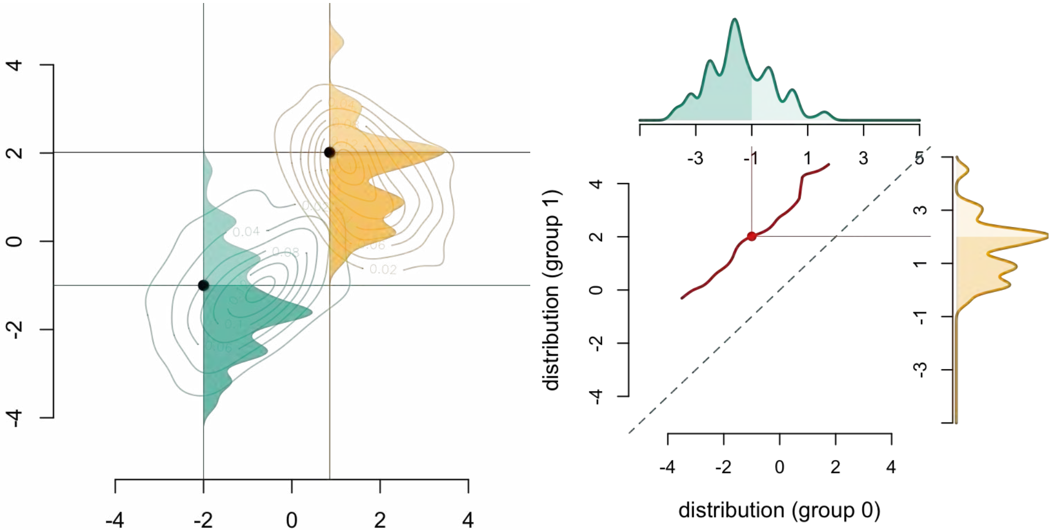
the "**monotone lower triangular map**," defined in Bogachev et al. (2005).

# Counterfactual Fairness

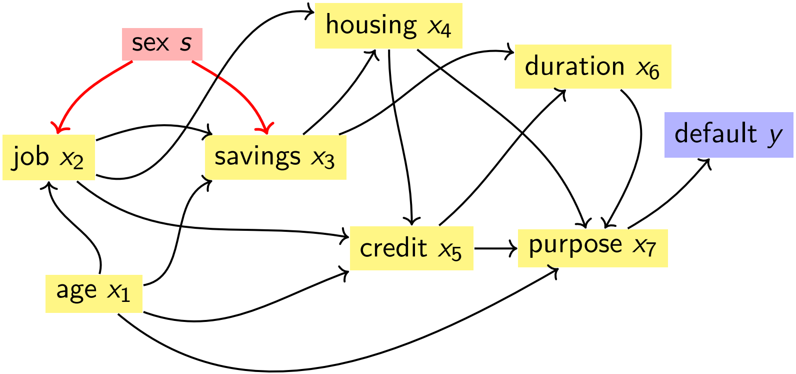




# Counterfactual Fairness



# Counterfactual Fairness



Causal graph in the German Credit dataset from Watson et al. (2021), or DAG.
(acyclical probablistic graphical models)

# Counterfactual Fairness

The joint distribution of $\boldsymbol{X}$ satisfies the (global) **Markov property** w.r.t. $\mathcal{G}$:

$$\mathbb{P}[x_1,\cdots,x_d]=\prod_{j=1}^{d}\mathbb{P}[x_j|\text{parents}(x_j)],$$

where parents$(x_i)$ are nodes with edges directed towards $x_i$, in $\mathcal{G}$.

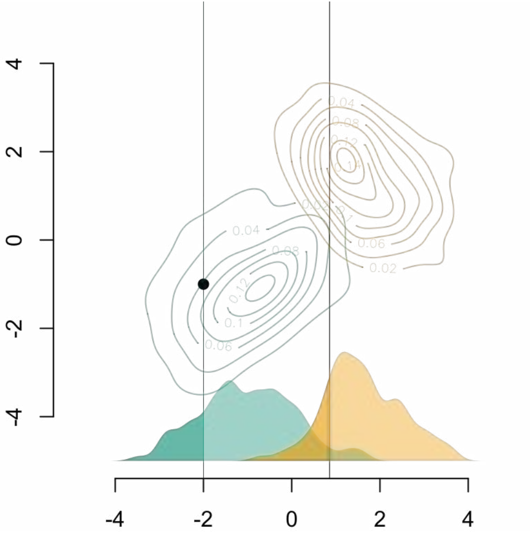
# Counterfactual Fairness

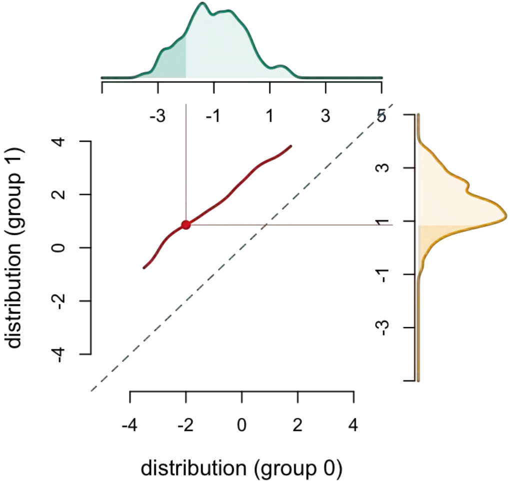
Consider some acyclical causal graph $\mathcal{G}$ on $(s, \boldsymbol{x})$ where variables are topologically sorted, where $s \in \{A, B\}$ is a binary variable , defining two measures $\mu_A$ and $\mu_B$ on $\mathbb{R}^d$, by conditioning on $s = A$ and $s = B$, respectively, factorized according to $\mathcal{G}$. Define

$$T^\star_{\mathcal{G}}(x_1, \cdots, x_d) = \begin{pmatrix} T^\star_1(x_1) \\ T^\star_2(x_2 | \text{ parents}(x_2)) \\ \vdots \\ T^\star_{d-1}(x_{d-1} | \text{ parents}(x_{d-1})) \\ T^\star_d(x_d | \text{ parents}(x_d)) \end{pmatrix}.$$

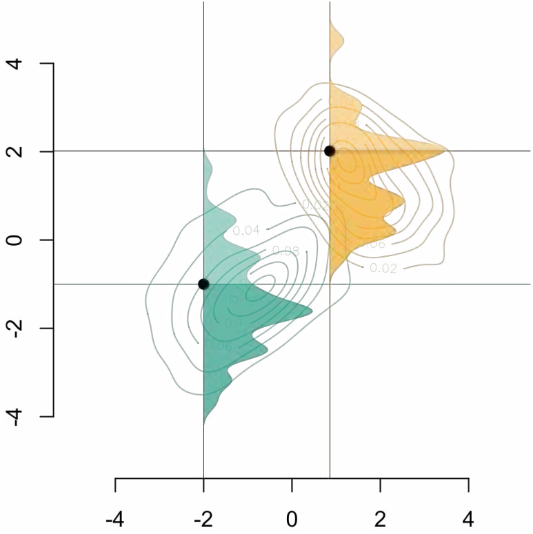
This mapping will be called “sequential conditional transport on the graph $\mathcal{G}$.”
The counterfactual value will be obtained by propagating “downstream” the causal graph (following the topological order), when $s$ changes from $A$ to $B$.

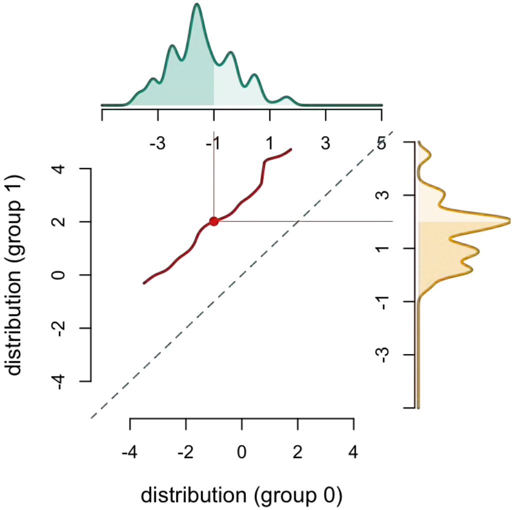
# Counterfactual Fairness



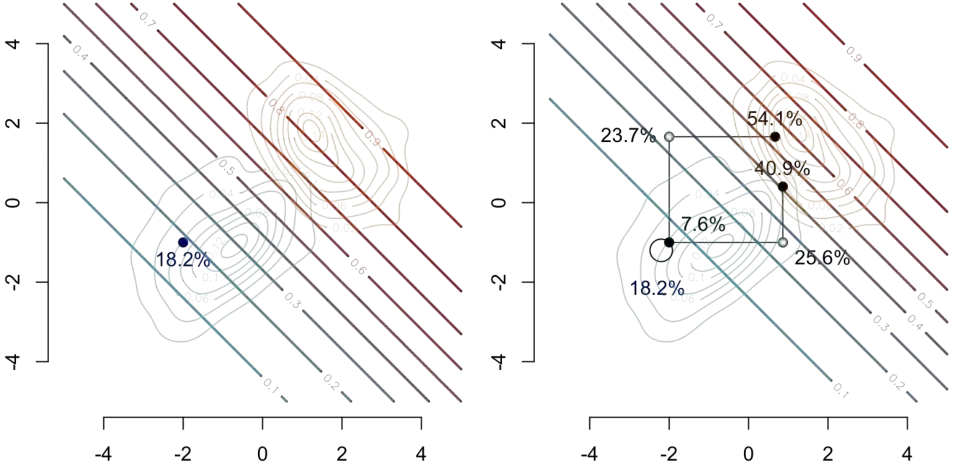


# Counterfactual Fairness





# Counterfactual Fairness



# Counterfactual Fairness

The mutatis mutandis difference $m(s=1,x_1^\star,x_2^\star)-m(s=0,x_1,x_2)$, i.e., $+22.70\%$, is:

$$\begin{array}{rll} & m(s=1,x_1,x_2)-m(s=0,x_1,x_2) & :-10.65\% \\ + & m(s=1,x_1^\star,x_2)-m(s=1,x_1,x_2) & :+17.99\% \\ + & m(s=1,x_1^\star,x_2^\star)-m(s=1,x_1^\star,x_2) & :+15.37\%. \end{array}$$

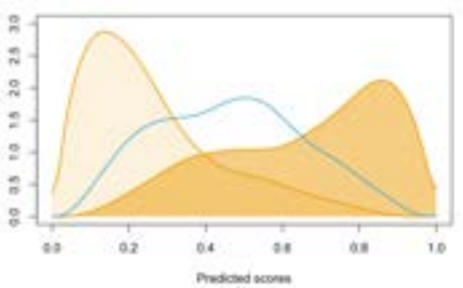
or $m(s=1,x_1^\star,x_2^\star)-m(s=0,x_1,x_2)$, i.e., $+35.82\%$, is:

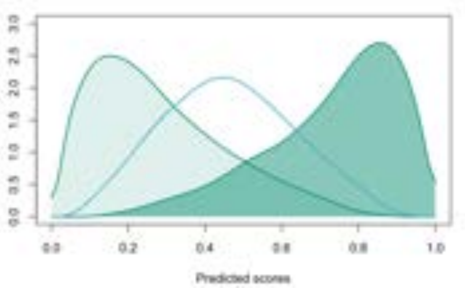
$$\begin{array}{rll} & m(s=1,x_1,x_2)-m(s=0,x_1,x_2) & :-10.66\% \\ + & m(s=1,x_1,x_2^\star)-m(s=1,x_1,x_2) & :+16.07\% \\ + & m(s=1,x_1^\star,x_2^\star)-m(s=1,x_1,x_2^\star) & :+30.41\%. \end{array}$$

The "treatment effect" depends on the causal structure.

# The Case of Multiple Attributes

- Consider a machine Learning model $m$, score predictions and two sensitive attributes, ethnic origin $A_1$ (White/Black) and gender $A_2$ (Male/Female).
- Consider densities of $\nu_{m|A_1=0}$, $\nu_{m|A_1=1}$ (left) and $\nu_{m|A_2=0}$, $\nu_{m|A_2=1}$ (right)
- Plot densities of barycenters, $\nu_{m_{B_1}}$ and $\nu_{m_{B_2}}$





# The Case of Multiple Attributes

- Intersectional Fairness, MSA → Single sensitive attribute (SSA), by intersection,

$$\boldsymbol{a} \in \mathcal{A} = \overset{\text{ethnic origin } A_1}{\mathcal{A}_1} \times \overset{\text{gender } A_2}{\mathcal{A}_2} = \{\text{white}, \text{black}\} \times \{\text{male}, \text{female}\}$$

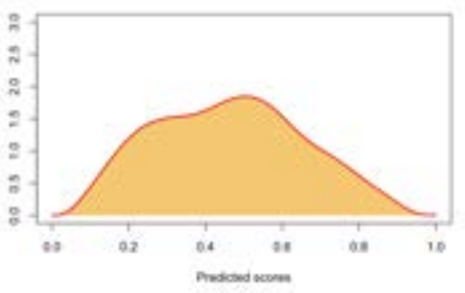
Here $\mathcal{A}$ corresponds to $4 = 2 \times 2$ states,

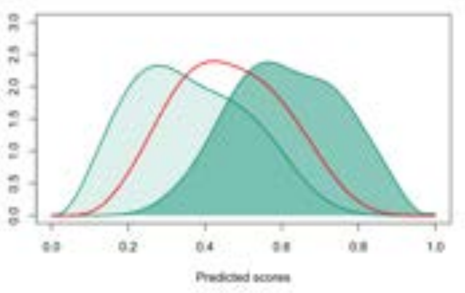
$$\mathcal{A} = \Big\{(\text{white}, \text{male}), (\text{white}, \text{female}), (\text{black}, \text{male}), (\text{black}, \text{female})\Big\}$$

- Sequential Fairness, MSA, in Hu et al. (2024)

# The Case of Multiple Attributes

- Given $\nu_{m_{B_1}}$, consider
  - the barycenter $\nu_{m_{B_1}}$ conditional on $A_1$ (no impact, already fair)
  - the barycenter $\nu_{m_{B_2}}$ conditional on $A_2$

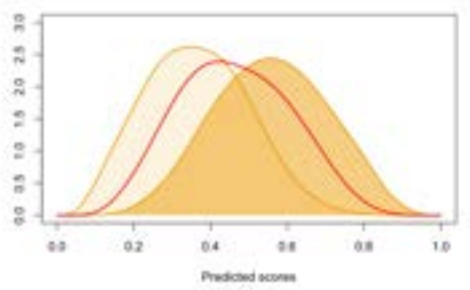


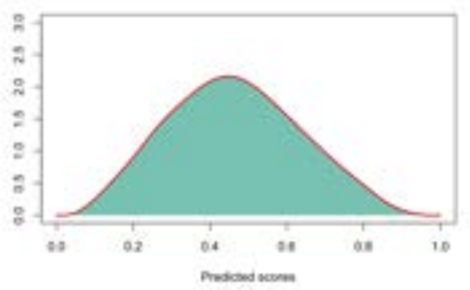


- On the right, distribution of $\nu_{m_{B_2} \circ m_{B_1}}$

# The Case of Multiple Attributes

- Given $\nu_{m_{B_2}}$, consider
  - the barycenter $\nu_{m_{B_1}}$ conditional on $A_1$
  - the barycenter $\nu_{m_{B_2}}$ conditional on $A_2$ (no impact, already fair)

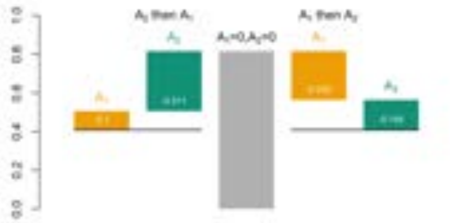


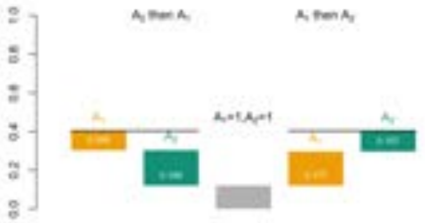


- On the left, distribution of $\nu_{m_{B_1} \circ m_{B_2}}$

## The Case of Multiple Attributes

- The order of this sequential approach leads different interpretations,
  - left hand part, $A_2$ then $A_1$
  - right hand part, $A_1$ then $A_2$





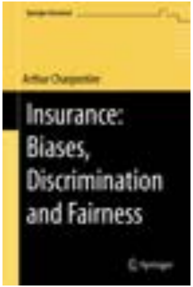
# Mitigating Discrimination ? (brief conclusion)

If it is mandatory to mitigate, there are robust techniques that can guarantee fairness

Supreme Court Justice Harry Blackmun stated, in 1978,
"*In order to get beyond racism, we must first take account of race. There is no other way. And in order to treat some persons equally, we must treat them differently*," Knowlton (1978), cited in Lippert-Rasmussen (2020)

In 2007, John G. Roberts of the U.S. Supreme Court submits
"*The way to stop discrimination on the basis of race is to stop discriminating on the basis of race*," Sabbagh (2007) and Turner (2015)

To go further,
Charpentier (2024) Insurance: Biases, Discrimination and Fairness.



Arthur Charpentier

Insurance: Biases, Discrimination and Fairness

# References

Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.

Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2017.

Bogachev, V. I., Kolesnikov, A. V., and Medvedev, K. V. (2005). Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309.

Bonnotte, N. (2013). From Knothe's rearrangement to Brenier's optimal transport map. *SIAM Journal on Mathematical Analysis*, 45(1):64–87.

Carlier, G., Galichon, A., and Santambrogio, F. (2010). From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576.

Charpentier, A. (2024). *Insurance: biases, discrimination and fairness*. Springer Verlag.

Charpentier, A., Flachaire, E., and Gallic, E. (2023). Causal inference with optimal transport. In Thach, N. N., Kreinovich, V., Ha, D. T., and Trung, N. D., editors, *Optimal Transport Statistics for Economics and Related Topics*. Springer Verlag.

Crossney, K. B. (2016). Redlining. *https://philadelphiaencyclopedia.org/essays/redlining/*.

# References

De Lara, L., González-Sanz, A., Asher, N., Risser, L., and Loubes, J.-M. (2024). Transport-based counterfactual models. *Journal of Machine Learning Research*, 25(136):1–59.

Denuit, M., Charpentier, A., and Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics & Economics*.

Gouic, T. L., Loubes, J.-M., and Rigollet, P. (2020). Projection to fairness in statistical learning. *arXiv*, 2005.11720.

Hu, F., Ratz, P., and Charpentier, A. (2023). Fairness in multi-task learning via wasserstein barycenters. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases – ECML PKDD*.

Hu, F., Ratz, P., and Charpentier, A. (2024). A sequentially fair mechanism for multiple sensitive attributes. *Annual AAAI Conference on Artificial Intelligence*.

Hume, D. (1739). *A Treatise of Human Nature*. Cambridge University Press Archive.

Kantorovich, L. V. (1942). On the translocation of masses. In *Doklady Akademii Nauk USSR*, volume 37, pages 199–201.

Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

# References

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

Knothe, H. (1957). Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39–52.

Knowlton, R. E. (1978). Regents of the university of california v. bakke. *Arkansas Law Review*, 32:499.

Kranzberg, M. (1986). Technology and history:" kranzberg's laws". *Technology and culture*, 27(3):544–560.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.

Lippert-Rasmussen, K. (2020). *Making sense of affirmative action*. Oxford University Press.

Machado, A. F., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024). From uncertainty to precision: Enhancing binary classifier performance through calibration. *arXiv preprint arXiv:2402.07790*.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

# References

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Plečko, D., Bennett, N., and Meinshausen, N. (2021). fairadapt: Causal reasoning for fair data pre-processing.

Plečko, D. and Meinshausen, N. (2020). Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44.

Rhynhart, R. (2020). Mapping the legacy of structural racism in philadelphia. *Philadelphia, Office pf the Controller*.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472.

Sabbagh, D. (2007). *Equality and transparency: A strategic perspective on affirmative action in American law*. Springer.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.

Turner, R. (2015). The way to stop discrimination on the basis of race. *Stanford Journal of Civil Rights & Civil Liberties*, 11:45.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):1–7.

# References

Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Society.

Watson, D. S., Gultchin, L., Taly, A., and Floridi, L. (2021). Local explanations via necessity and sufficiency: Unifying theory and practice. *Uncertainty in Artificial Intelligence*, pages 1382–1392.