Using optimal transport to mitigate unfair predictions

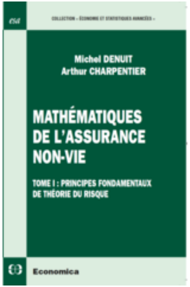
Arthur Charpentier, François Hu, Agathe Fernandes-Machado & Philipp Ratz

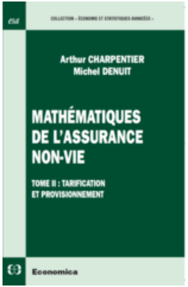
Københavns Universitet, June 2024

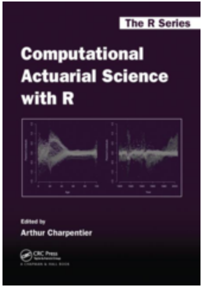
Bio (short)

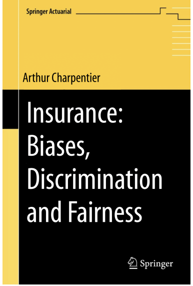
Arthur Charpentier Professor at Université du Québec à Montréal,

- Denuit and Charpentier (2004, 2005) Mathématiques de l'Assurance Non-Vie,
- Charpentier (2014) Computational Actuarial Science with R,
- Bénéplanc et al. (2022) Manuel d'Assurance,
- Charpentier (2024) Insurance: Biases, Discrimination and Fairness.









Bio (short)

François Hu Postdoctoral fellow, Université de Montréal

Philipp Ratz PhD Student, Université du Québec à Montréal

Agathe Fernandes-Machado PhD Student, Université du Québec à Montréal

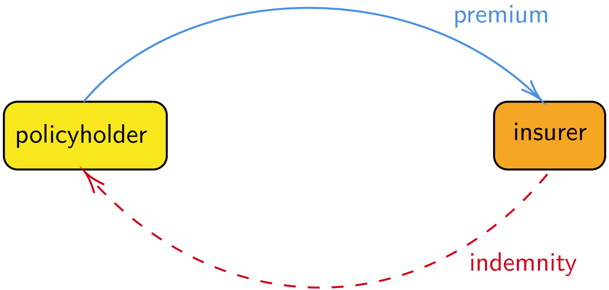
Olivier Côté PhD Student, Université Laval, Québec

Also interns **Ana Patrón Piñerez** and **Suzie Grondin**,

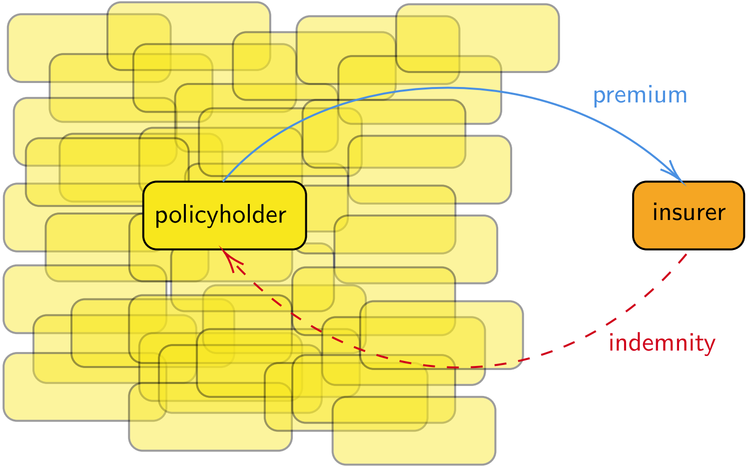
Insurance (and “Actuarial Fairness”)

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



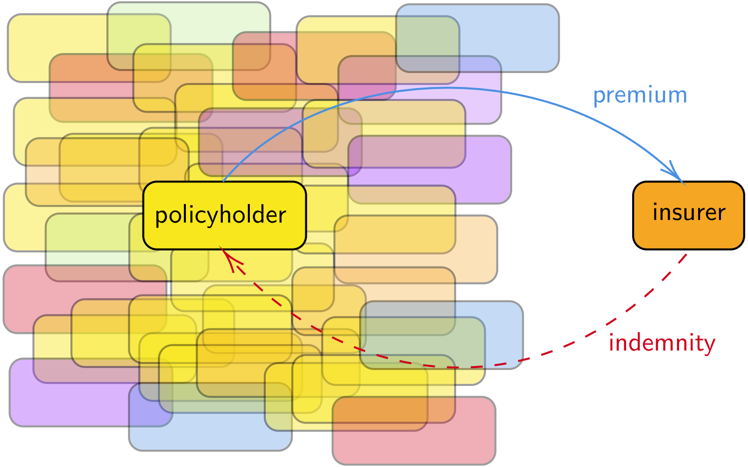
Insurance (and "Actuarial Fairness")

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



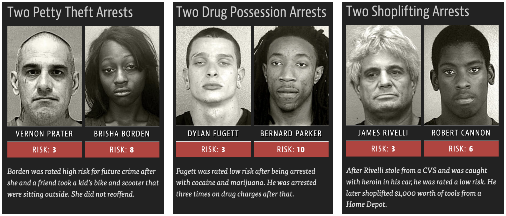
Insurance (and "Actuarial Fairness")

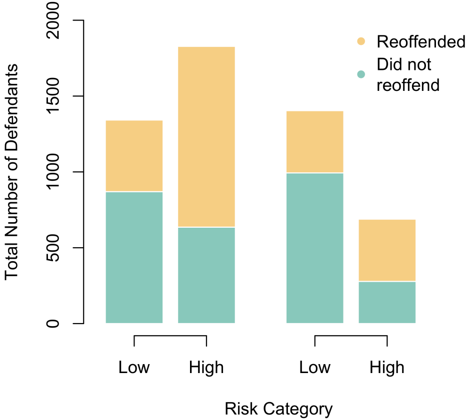
- "*Insurance is the contribution of the many to the misfortune of the few*"



Motivation (1. Propublica, Actuarial Justice)

- Concept of "**actuarial justice**" as coined in Feeley and Simon (1994)
- **C**orrectional **O**ffender **M**anagement **P**rofiling for **A**lternative **S**anctions (**COMPAS**), Perry (2013)

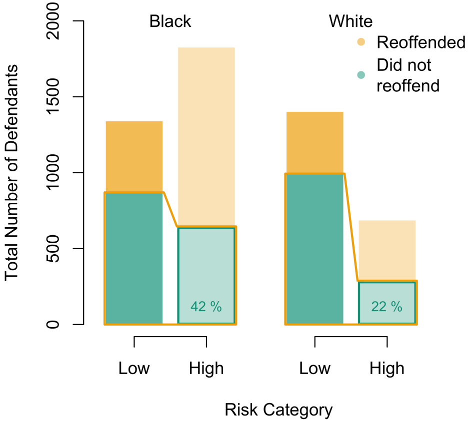




- https://github.com/propublica/compas-analysis
- Angwin et al. (2016) Machine Bias
  Dressel and Farid (2018)

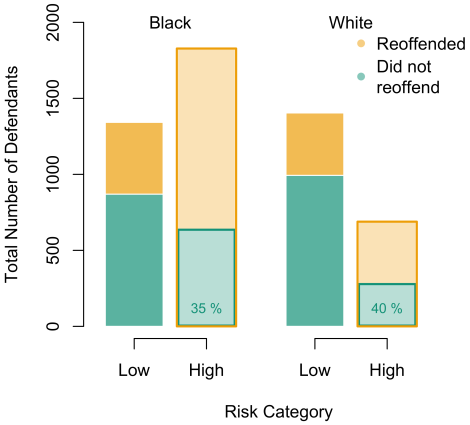
Motivation (1. Propublica, Actuarial Justice)

- From Feller et al. (2016),
  - for White people, among those who did not re-offend, 22% were wrongly classified,
  - for Black people, among those who did not re-offend, 42% were wrongly classified,
  - problem, since 42% $\gg$ 22%



Motivation (1. Propublica, Actuarial Justice)

- From Dieterich et al. (2016),
  - for White people, among those who where classified as high risk, 40% did not re-offend,
  - for Black people, among those who where classified as high risk, 35% did not re-offend,
  - no problem, since 40% $\approx$ 35%



Motivation (2. 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 (2. 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 (2. 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 (2. 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 (2. Legal Aspects)

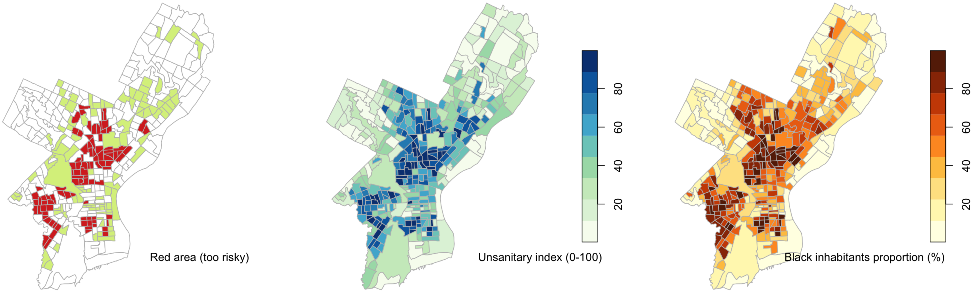
- In France, Loi n° 2008-496 du 27 mai 2008

– 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 (3. 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 (4. Competition)

- Québec Province, life table can be gender based
- Ontario Province, life table must be unisex

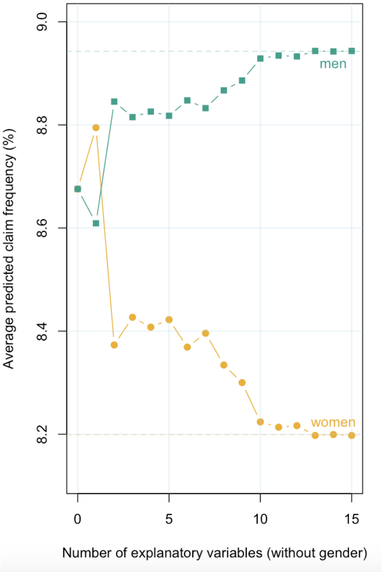
Age	men		unisex		women	
x	L_x	${}_{10}p_x$	L_x	${}_{10}p_x$	L_x	${}_{10}p_x$
0	100,000	1%	100,000	1%	100,000	0%
10	99,373	0%	99,451	0%	99,528	0%
20	99,147	1%	99,269	0%	99,391	0%
30	98,589	1%	98,869	1%	99,148	0%
40	97,727	2%	98,203	1%	98,679	0%
50	96,194	4%	97,242	2%	98,290	1%
60	92,482	10%	95,115	10%	97,748	9%
70	83,189	25%	85,891	21%	88,593	18%
80	62,635	59%	67,649	53%	72,662	47%
90	25,452	94%	31,908	90%	38,364	87%
100	1,473		3,142		4,811	

Motivation (5. 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

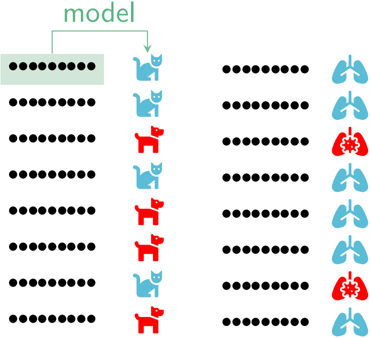
"*Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitly ask for*," Kearns and Roth (2019)

"*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)

Classifiers (or why actuarial science $\neq$ computer science)

Classifiers on pictures,
$\rightarrow$ (cats) – (dogs)
$\rightarrow$ (healthy) – (sick)

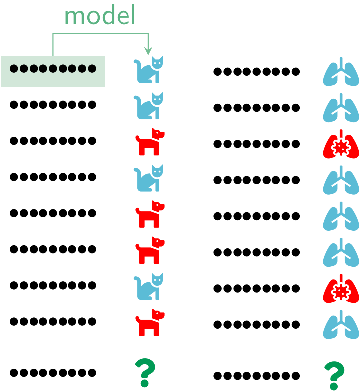


Classifiers (or why actuarial science $\neq$ computer science)

Classifiers on pictures,

$\rightarrow$ (cats) – (dogs)

$\rightarrow$ (healthy) – (sick)



Classifiers (or why actuarial science $\neq$ computer science)

Classifiers on pictures,

$\rightarrow$ (cats) – (dogs)

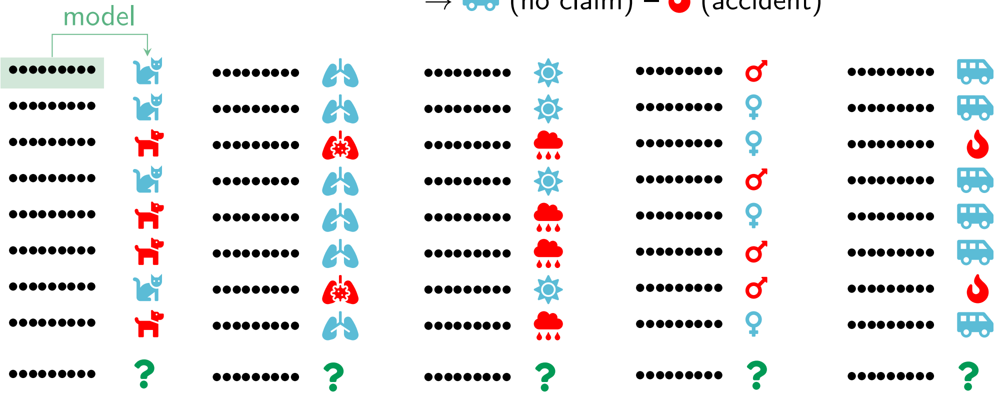
$\rightarrow$ (healthy) – (sick)

Classifiers, we need some "**probabilities**"

$\rightarrow$ (sunny) – (rainy)

$\rightarrow$ (woman) – (man)

$\rightarrow$ (no claim) – (accident)



Fairness for Classifiers

$$\begin{cases} \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{`explanatory' variables} \\ s \in \{A, B\} : \text{"sensitive variable"} \\ y \in \{0,1\} : \text{classification problem} \\ \widehat{y} \in \{0,1\} : \text{prediction, classically } \widehat{y} = \mathbf{1}(m(\boldsymbol{x}, s) > t) \end{cases}$$

Fairness for Classifiers

$$\begin{cases} \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{'explanatory' variables} \\ s \in \{A, B\} : \text{"sensitive variable"} \\ y \in \{0,1\} : \text{classification problem} \\ \widehat{y} \in \{0,1\} : \text{prediction, classically } \overbrace{\boxed{\widehat{y}}}^{\text{class} \in \{0,1\}} = \mathbf{1}(\, m(\boldsymbol{x}, s) \, > t) \end{cases}$$

Fairness for Classifiers

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

Fairness for Classifiers

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

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 $(\boldsymbol{X}, S, Y)$ $\leftarrow$ demographic parity

Fairness for Classifiers

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

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 $(\boldsymbol{X}, S, Y)$ $\leftarrow$ 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 $(\boldsymbol{X}, S, Y)$ $\leftarrow$ equalized odds

Fairness for Classifiers

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

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

classical definition of "demographic parity" for a classifier

$$\mathbb{P}[\ \widehat{Y}=1 \mid S=A\ ] \stackrel{?}{=} \mathbb{P}[\ \widehat{Y}=1 \mid S=B\ ]$$

Fairness for Classifiers

classical definition of "demographic parity" for a classifier

$$\mathbb{P}[\widehat{Y}=1\,|\,\overbrace{S=A}^{\text{sensitive}}]\overset{?}{=}\mathbb{P}[\widehat{Y}=1\,|\,\overbrace{S=B}^{\text{sensitive}}]$$

Fairness for Classifiers

classical definition of "demographic parity" for a classifier

$$\mathbb{P}[\underbrace{\widehat{Y}=1}_{\text{class prediction}} \mid \overbrace{S=A}^{\text{sensitive}}] \overset{?}{=} \mathbb{P}[\underbrace{\widehat{Y}=1}_{\text{class prediction}} \mid \overbrace{S=B}^{\text{sensitive}}]$$

Fairness for Classifiers

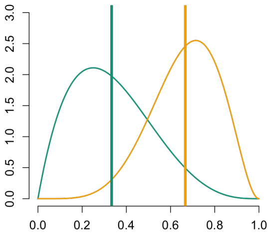
classical definition of "demographic parity" for a classifier

$$\mathbb{E}[\underbrace{\widehat{Y}}_{\text{class prediction}} \mid \overbrace{S = A}^{\text{sensitive}}] \overset{?}{=} \mathbb{E}[\underbrace{\widehat{Y}}_{\text{class prediction}} \mid \overbrace{S = B}^{\text{sensitive}}]$$

Fairness for Classifiers

(**weak**) definition of "demographic parity" for a classifier

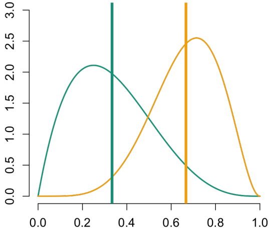
$$\mathbb{E}[\underbrace{m(\boldsymbol{X},S)}_{\text{score}} \mid \overbrace{S=\mathrm{A}}^{\text{sensitive}}] \overset{?}{=} \mathbb{E}[\underbrace{m(\boldsymbol{X},S)}_{\text{score}} \mid \overbrace{S=\mathrm{B}}^{\text{sensitive}}]$$



Fairness for Classifiers

(weak) definition of "demographic parity" for a classifier

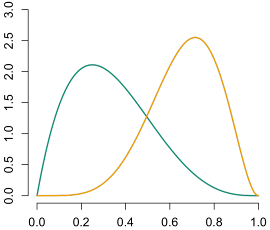
$$\mathbb{E}[\underbrace{m(\boldsymbol{X},S)}_{\text{score}} \mid \overbrace{S=\mathsf{A}}^{\text{sensitive}}] \stackrel{?}{=} \mathbb{E}[\underbrace{m(\boldsymbol{X},S)}_{\text{score}} \mid \overbrace{S=\mathsf{B}}^{\text{sensitive}}]$$



(**strong**) definition of "demographic parity" for a classifier

$$\mathbb{P}[m(\boldsymbol{X},S)\in\mathcal{I} \mid S=\mathsf{A}] \stackrel{?}{=} \mathbb{P}[m(\boldsymbol{X},S)\in\mathcal{I} \mid S=\mathsf{B}]$$

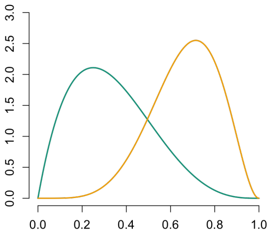
$\forall\mathcal{I}\subset[0,1]$,



Fairness for Classifiers

(weak) definition of "demographic parity" for a classifier

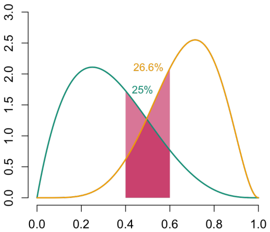
$$\mathbb{E}[\underbrace{m(\boldsymbol{X},S)}_{\text{score}} \mid \underbrace{S=A}_{\text{sensitive}}] \stackrel{?}{=} \mathbb{E}[\underbrace{m(\boldsymbol{X},S)}_{\text{score}} \mid \underbrace{S=B}_{\text{sensitive}}]$$



(**strong**) definition of "demographic parity" for a classifier

$$\mathbb{P}[m(\boldsymbol{X},S)\in\mathcal{I} \mid S=A] \stackrel{?}{=} \mathbb{P}[m(\boldsymbol{X},S)\in\mathcal{I} \mid S=B]$$

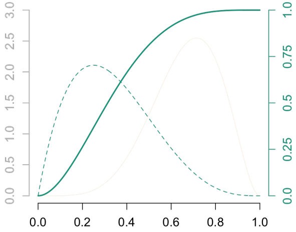
$\forall\mathcal{I}\subset[0,1]$, e.g. $[40\%;60\%]$.



Fairness for Classifiers using Optimal Transport

$$F_A(u) = \mathbb{P}[\underbrace{m(\boldsymbol{X}, S)}_{\text{score}} \leq u | \underbrace{S = A}_{\text{sensitive}}]$$

let F_A denote the cumulative distribution function of scores in group A

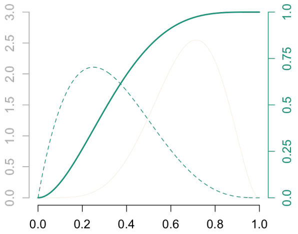


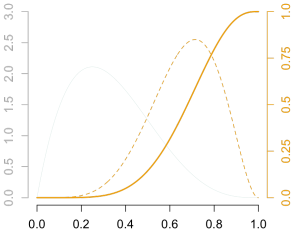
Fairness for Classifiers using Optimal Transport

$$F_A(u) = \mathbb{P}[\overbrace{m(\boldsymbol{X}, S)}^{\text{score}} \leq u | \overbrace{S = A}^{\text{sensitive}}]$$

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

and F_B denote the cumulative distribution function of scores in group B





Fairness for Classifiers using Optimal Transport

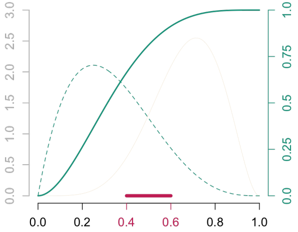
$$F_A(u) = \mathbb{P}[\, m(\boldsymbol{X}, S) \leq u \mid S = A \,]$$

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

and F_B denote the cumulative distribution function of scores in group B

Consider individuals in group A such that

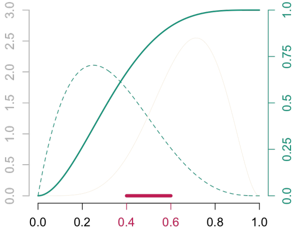
$$\underbrace{m(\boldsymbol{X}, S)}_{\text{score}} \in [0.4; 0.6] \mid \underbrace{S = A}_{\text{sensitive}}$$



Fairness for Classifiers using Optimal Transport

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

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



and F_B denote the cumulative distribution function of scores in group B

Consider individuals in group A such that

$$\underbrace{m(\boldsymbol{X},S)}_{\text{score}} \in \underbrace{[0.4;0.6]}_{\textbf{levels}\text{ of the score}} \mid \underbrace{S = A}_{\text{sensitive}}$$

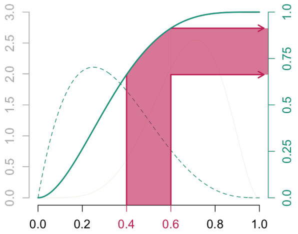
Fairness for Classifiers using Optimal Transport

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

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

and F_B denote the cumulative distribution function of scores in group B

Consider individuals in group A such that



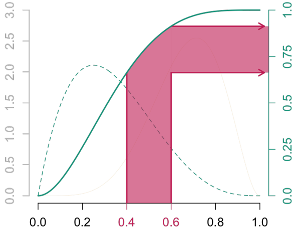
$$m(\boldsymbol{X}, S) \in [0.4; 0.6] \mid S = A \text{ then } \text{ranks}(m(\boldsymbol{X}, S)) \in [66.3\%; 91.3\%] \mid S = A$$

quantile

Fairness for Classifiers using Optimal Transport

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

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



and F_B denote the cumulative distribution function of scores in group B

Consider individuals in group A such that

$$m(\boldsymbol{X}, S) \ \in \ [0.4; 0.6] \ | \ S = \text{A} \ \text{ then } \ \boxed{\text{ranks}(m(\boldsymbol{X}, S))} \in \boxed{[66.3\%; 91.3\%]} \ | \ S = \text{A}$$

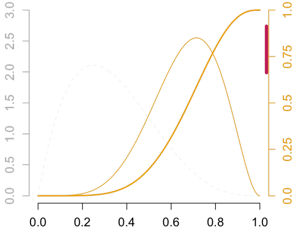
quantile

probabilities associated to the score

Fairness for Classifiers using Optimal Transport

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

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



and F_B denote the cumulative distribution function of scores in group B

Consider individuals in group A such that

$$m(\boldsymbol{X},S)\ \in\ [0.4;0.6]\ |\ S=A\ \text{ then }\ \text{ranks}(m(\boldsymbol{X},S))\ \in\ [66.3\%;91.3\%]\ |\ S=A$$

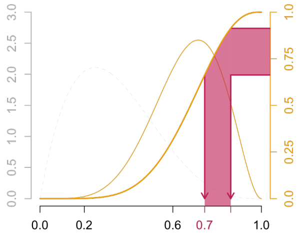
then, in group B

$$\text{if}\ \text{ranks}(m(\boldsymbol{X},S))\ \in\ [66.3\%;91.3\%]\ |\ S=B$$

Fairness for Classifiers using Optimal Transport

$$F_{A}\,(u) = \mathbb{P}[\; m(\boldsymbol{X},S)\;\leq u|\; S = A\;]$$

$$F_{B}\,(u) = \mathbb{P}[\; m(\boldsymbol{X},S)\;\leq u|\; S = B\;]$$



and F_B denote the cumulative distribution function of scores in group B

Consider individuals in group A such that

$$m(\boldsymbol{X},S)\;\in\;[0.4;0.6]\;|\;S=A\;\;\text{then}\;\;\text{ranks}(m(\boldsymbol{X},S))\;\in\;[66.3\%;91.3\%]\;|\,S=A$$

then, in group B

$$\text{if}\;\;\text{ranks}(m(\boldsymbol{X},S))\;\in\;[66.3\%;91.3\%]\;|\,S=B\;\text{then}\;\;m(\boldsymbol{X},S)\;\in\;[0.743;0.861]\;|\;S=B$$

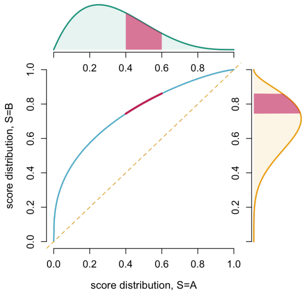
score

sensitive

Fairness for Classifiers using Optimal Transport

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

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



and F_B denote the cumulative distribution function of scores in group B

Consider individuals in group A such that

$$m(\boldsymbol{X}, S)\ \in \boxed{[0.4; 0.6]}\ |\ S = A$$

then, in group B

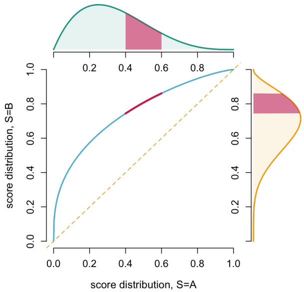
optimal transport mapping $T^\star$

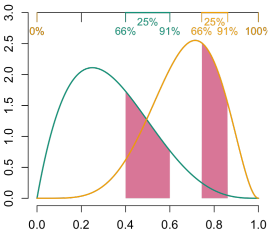
if ranks$(m(\boldsymbol{X}, S))\ \in\ [66.3\%; 91.3\%]\ |\ S = B$ then $m(\boldsymbol{X}, S)\ \in \boxed{[0.743; 0.861]}\ |\ S = B$

Formalizing Optimal Transport

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

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



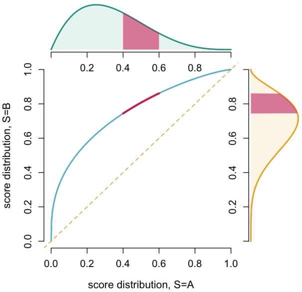


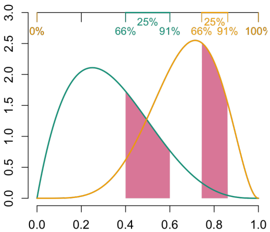
Formalizing Optimal Transport

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

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

probability p associated with score x in group A





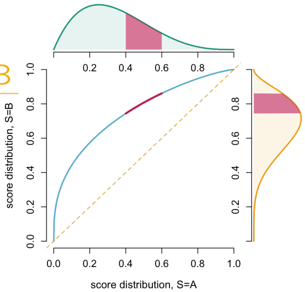
Formalizing Optimal Transport

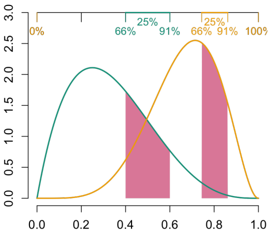
Consider the following $[0,1] \to [0,1]$ mapping

quantile of level p in group B

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

probability p associated with x in group A





Fairness for Classifiers

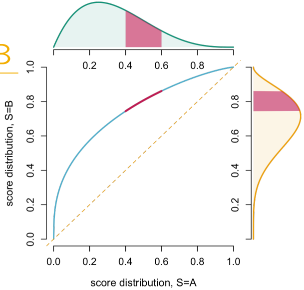
Consider the following $[0,1] \to [0,1]$ mapping

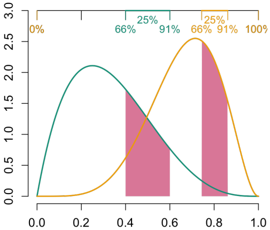
optimal transport mapping

quantile of level p in group B

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

probability p associated with x in group A





Formalizing Optimal Transport

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

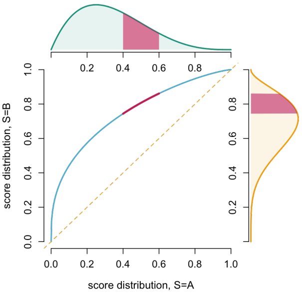
$$T^\star(x) = F_B^{-1} \circ F_A(x)$$

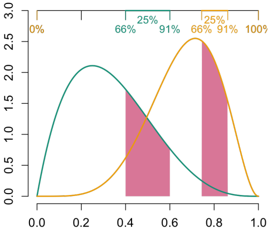
$$T^\star = \underset{T:[0,1]\to[0,1]}{\text{argmin}} \int_0^1 \big(T(x) - x\big)^2 dF_A(x)$$

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

Y with $Y \sim F_B$

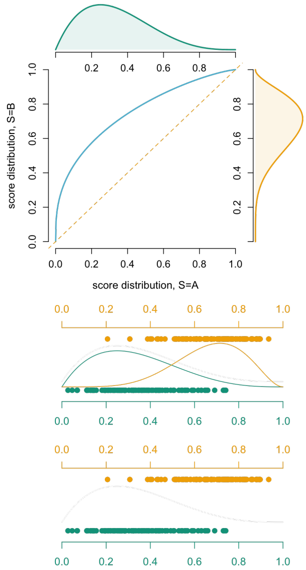
corresponding to Monge (1781) problem,
revisited by Kantorovich (1942).
(the minimum value is called **Wasserstein distance**)





Optimal Transport with a Finite Sample (another interpretation)

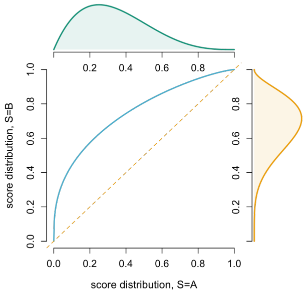
Consider two samples, $\big(m(\boldsymbol{x}_i, s_i = A)\big)$ and $\big(m(\boldsymbol{x}_i, s_i = B)\big)$

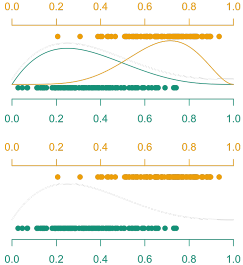


Optimal Transport with a Finite Sample (another interpretation)

$$\underbrace{m_1^A \leq m_2^A \leq \cdots \leq m_n^A}$$

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



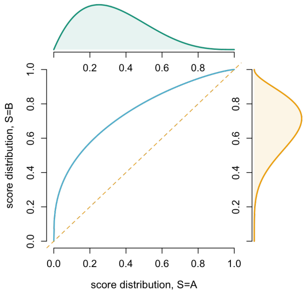


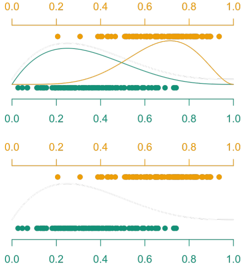
Optimal Transport with a Finite Sample (another interpretation)

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

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

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



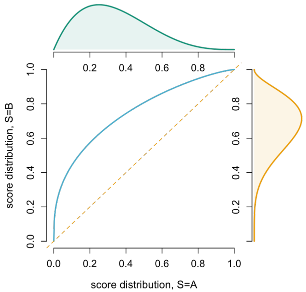


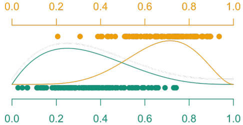
Optimal Transport with a Finite Sample (another interpretation)

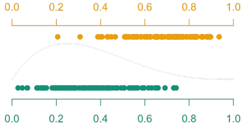
$m_1^B \leq m_2^B \leq \cdots \leq m_n^B$

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

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







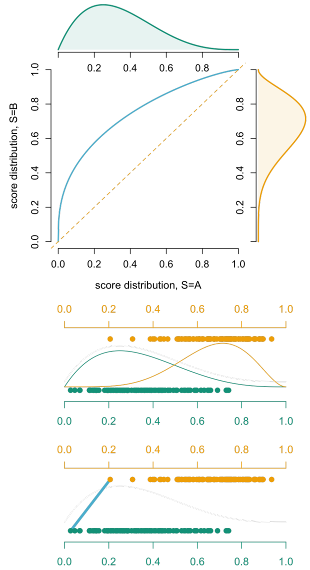
Optimal Transport with a Finite Sample (another interpretation)

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

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

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

$\boxed{m_1^A}$ $\longleftrightarrow$ $\boxed{m_1^B}$



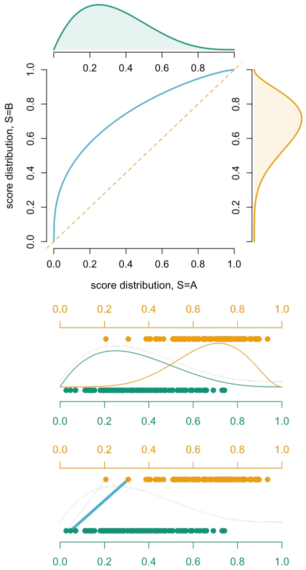
Optimal Transport with a Finite Sample (another interpretation)

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

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

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

$m_1^A \leq \boxed{m_2^A}$ $\qquad$ $m_1^B \leq \boxed{m_2^B}$



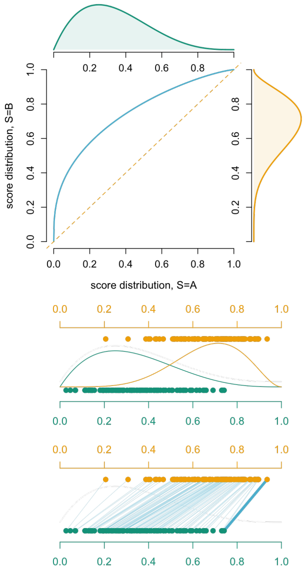
Optimal Transport with a Finite Sample (another interpretation)

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

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

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

$m_1^A \leq m_2^A \leq \cdots \leq \boxed{m_n^A} \qquad m_1^B \leq m_2^B \leq \cdots \leq \boxed{m_n^B}$



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

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

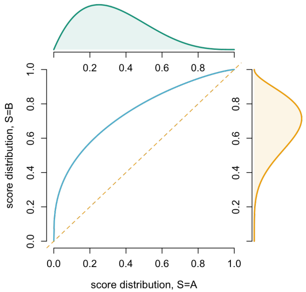
m is not fair with respect to s if $T^\star(x) \neq x$, or $m_i^A \neq m_i^B$

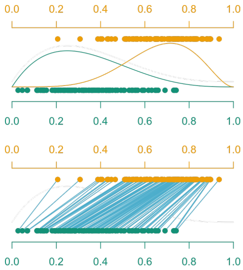
optimal transport mapping

quantile of level p in group B

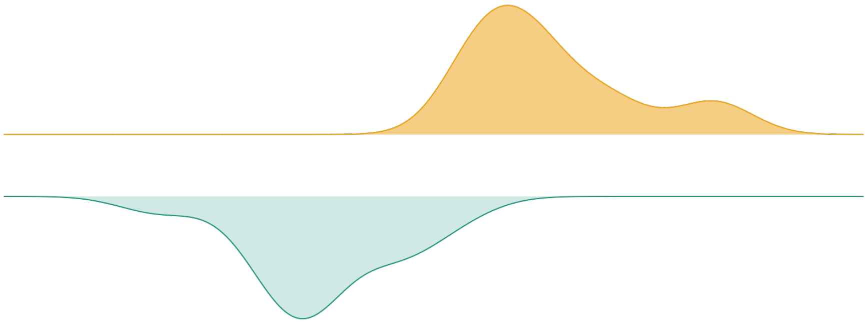
$$T^\star(x) = F_B^{-1} \circ F_A(x) \neq x$$

probability p associated with u in group A



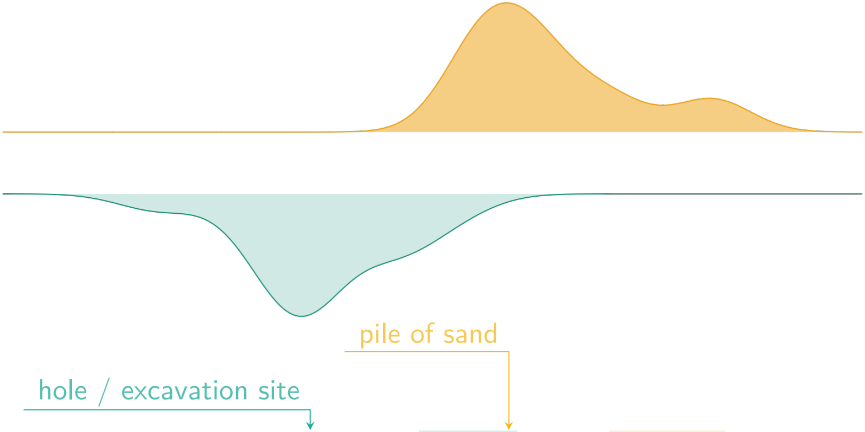


“Optimal Transport” (a side note / a cultural interlude)



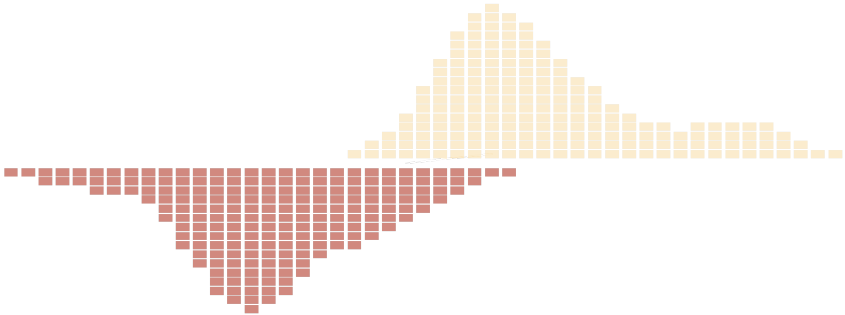
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

"Optimal Transport" (a side note / a cultural interlude)



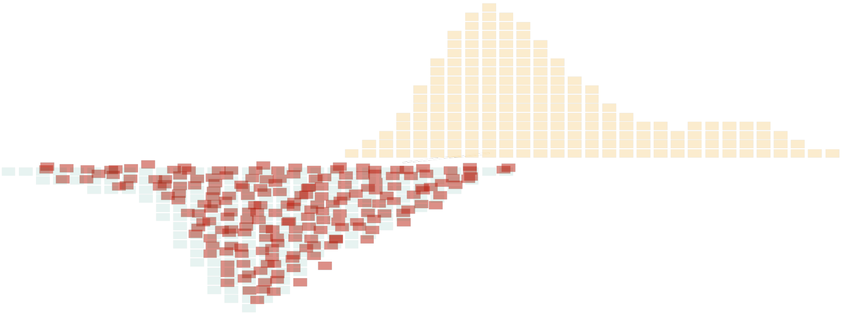
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

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



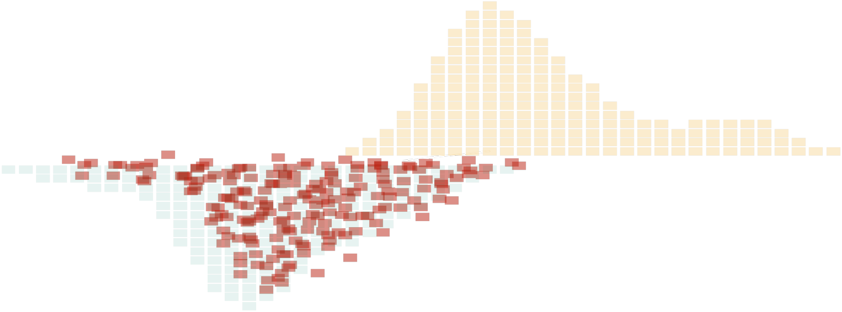
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

"Optimal Transport" (a side note / a cultural interlude)



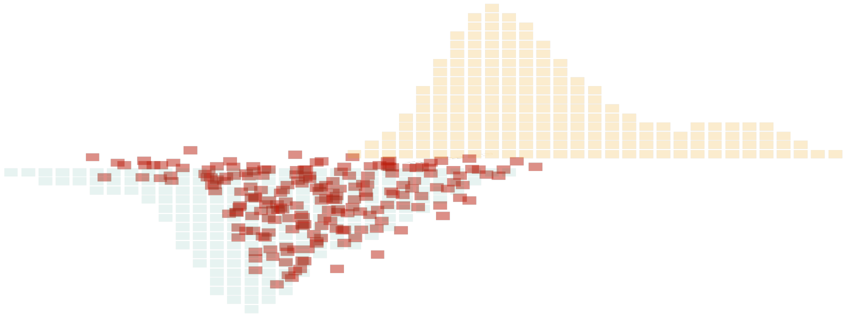
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

"Optimal Transport" (a side note / a cultural interlude)



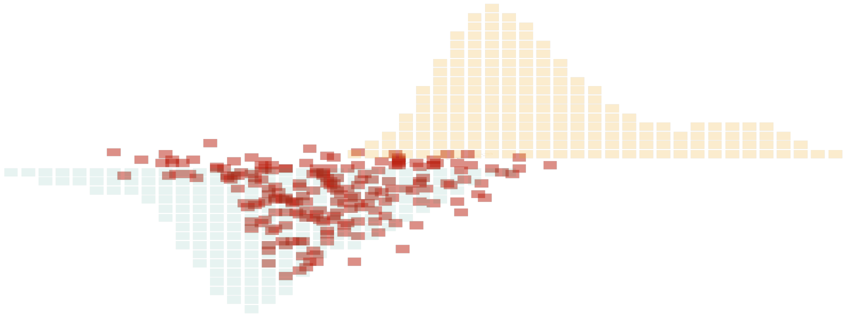
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

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



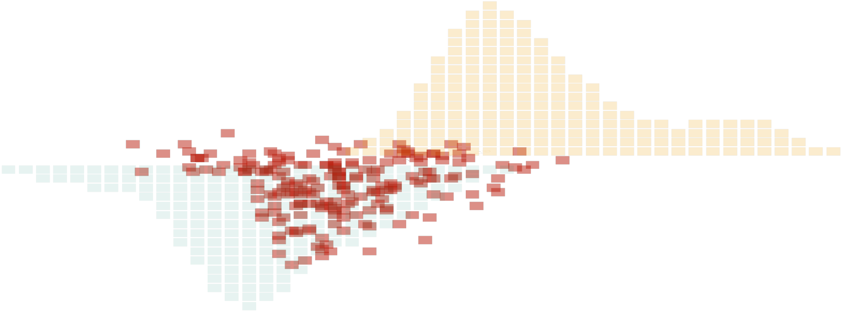
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

"Optimal Transport" (a side note / a cultural interlude)



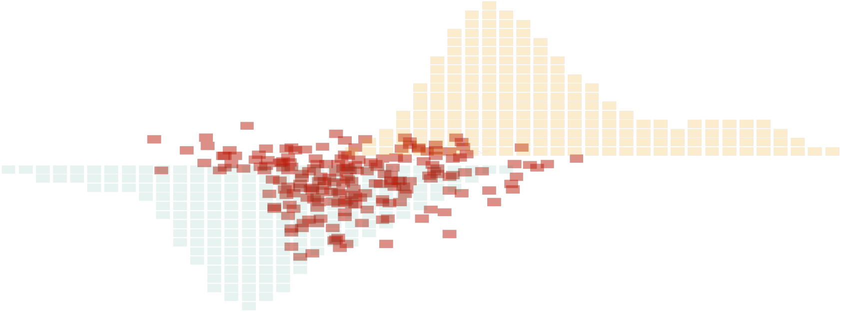
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

"Optimal Transport" (a side note / a cultural interlude)



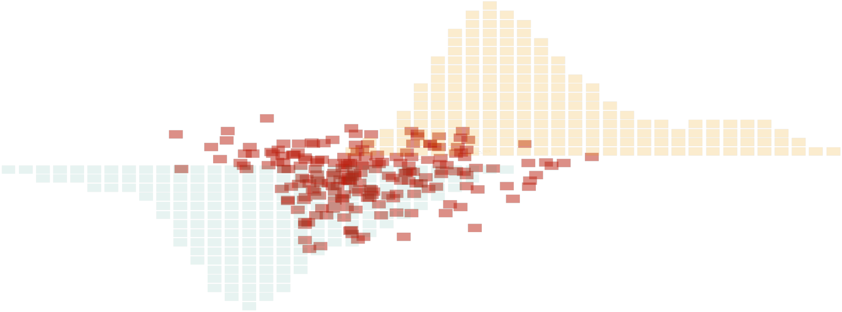
Monge (1781), "Mémoire sur la théorie des déblais et des remblais "

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



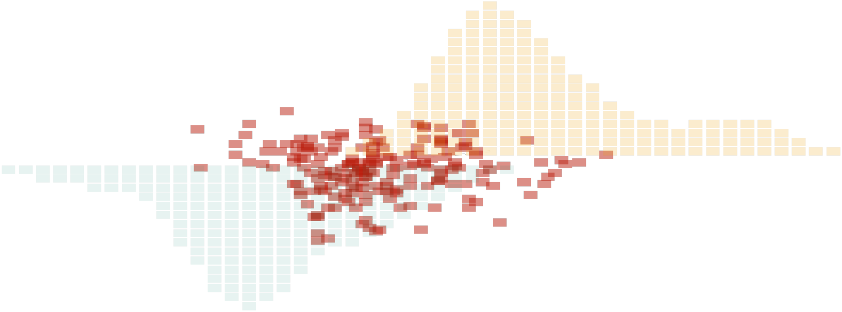
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



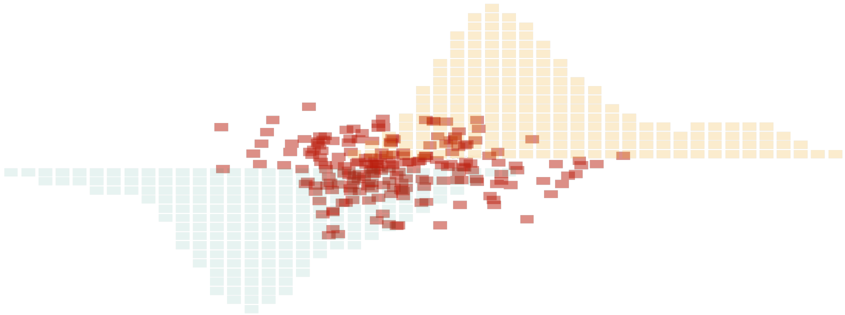
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



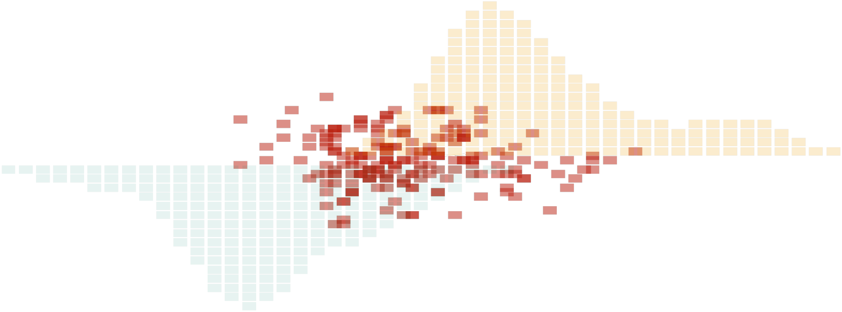
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



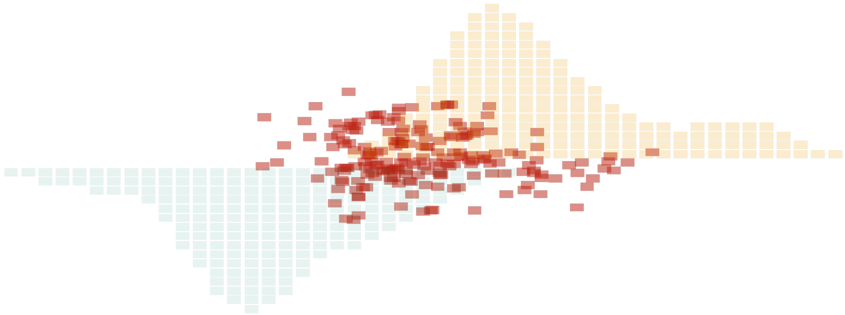
Monge (1781), “Mémoire sur la théorie des déblais et des remblais”

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



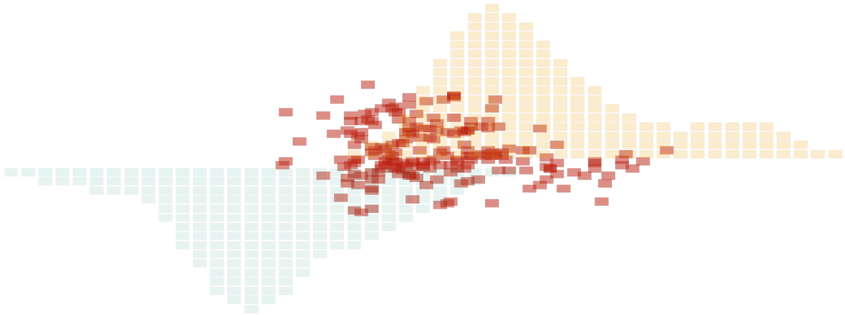
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



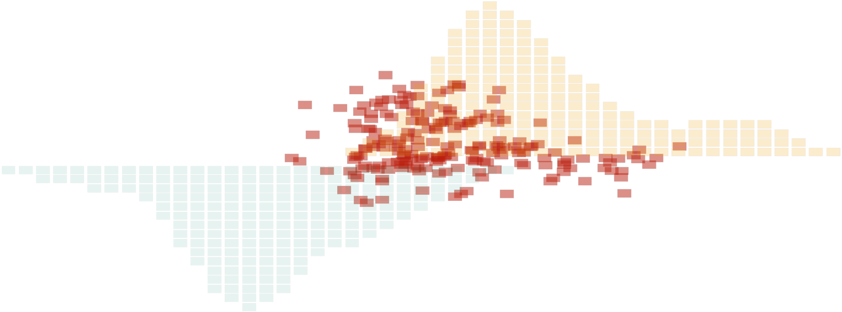
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



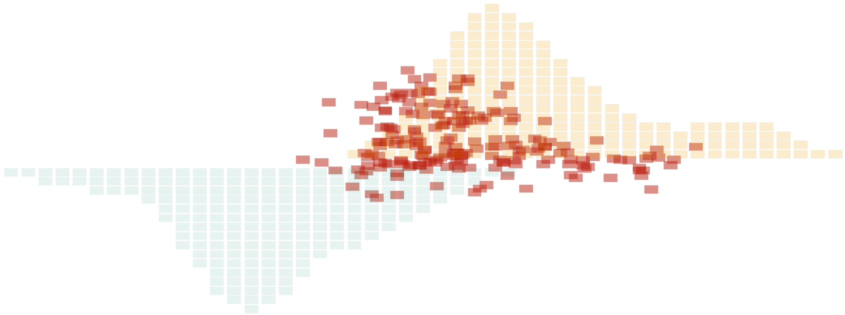
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



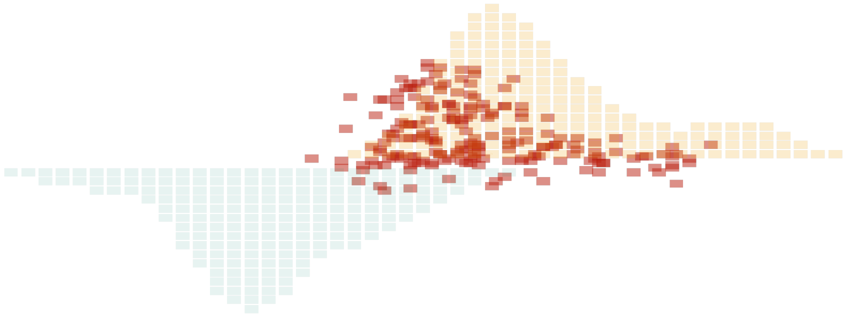
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



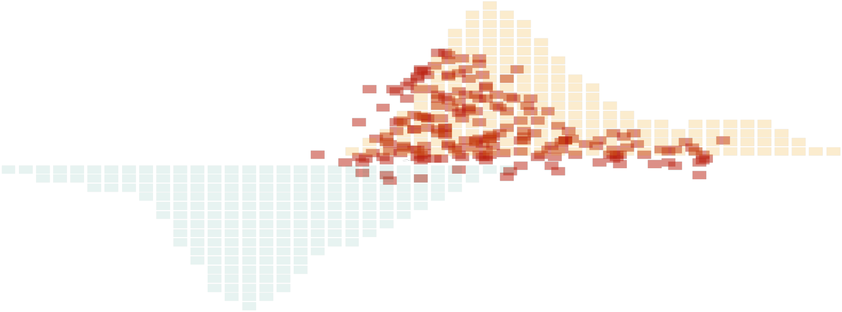
Monge (1781), “Mémoire sur la théorie des déblais et des remblais”

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



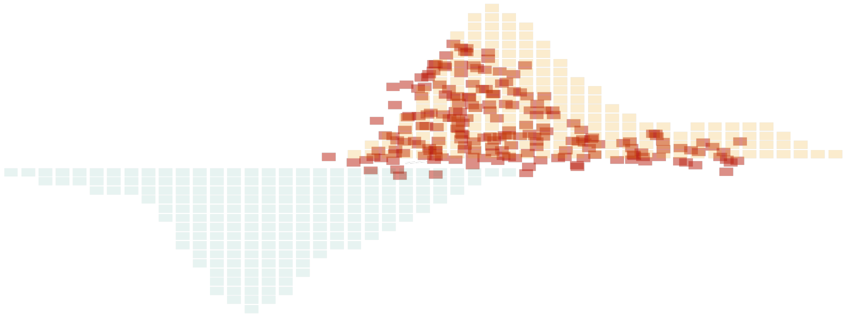
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



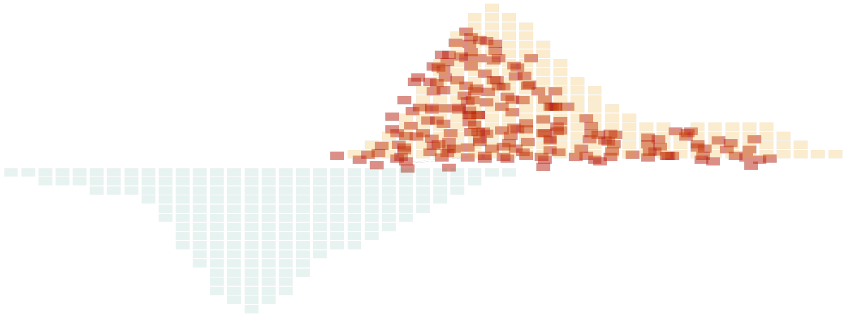
Monge (1781), “Mémoire sur la théorie des déblais et des remblais”

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



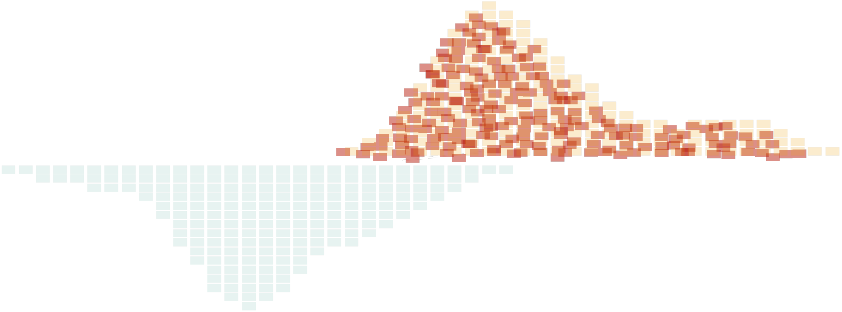
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

"Optimal Transport" (a side note / a cultural interlude)



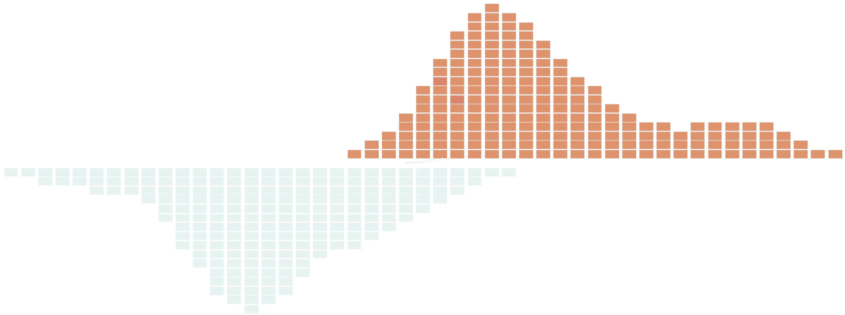
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

"Optimal Transport" (a side note / a cultural interlude)



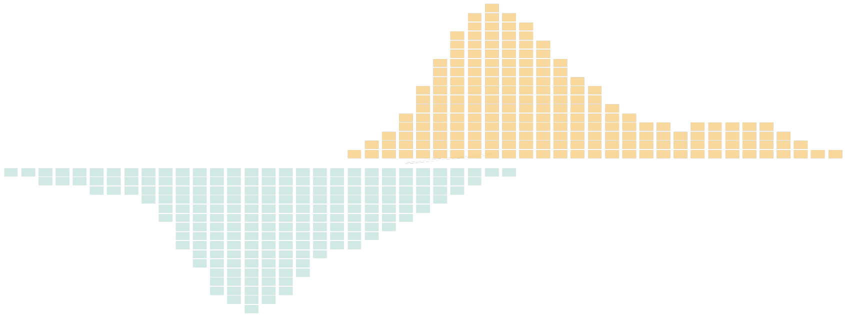
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

"Optimal Transport" (a side note / a cultural interlude)



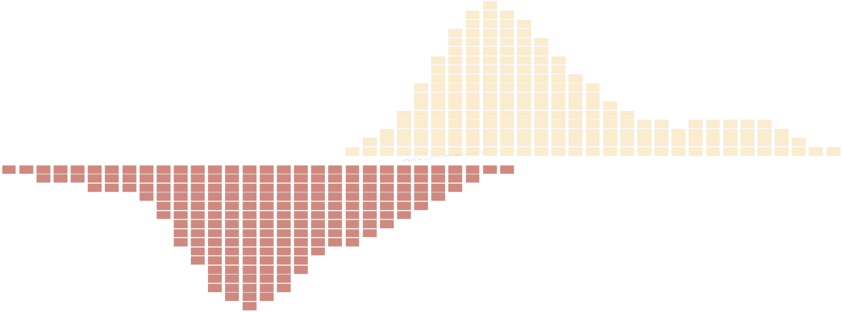
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

"Optimal Transport" (a side note / a cultural interlude)



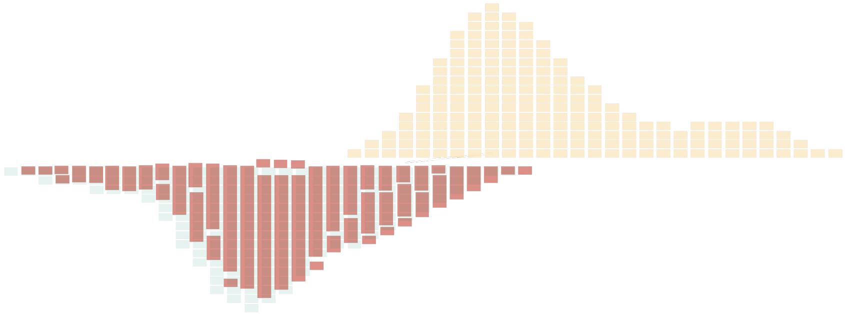
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

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



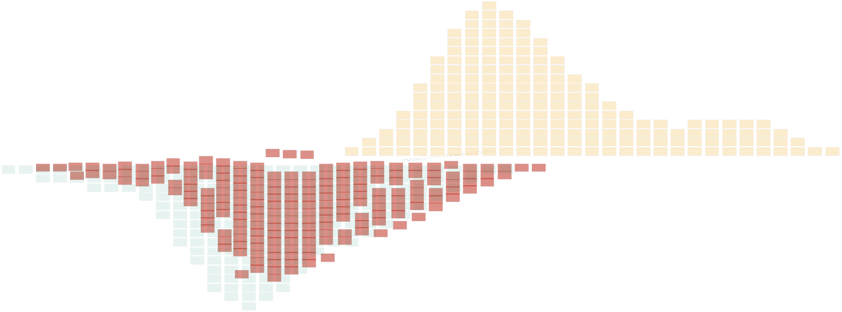
Monge (1781), “Mémoire sur la théorie des déblais et des remblais”

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



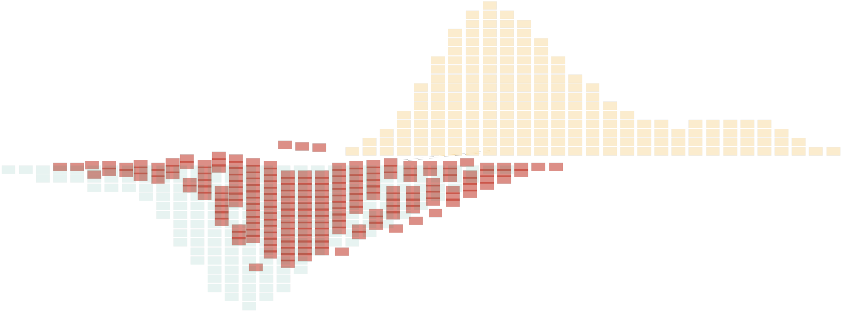
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



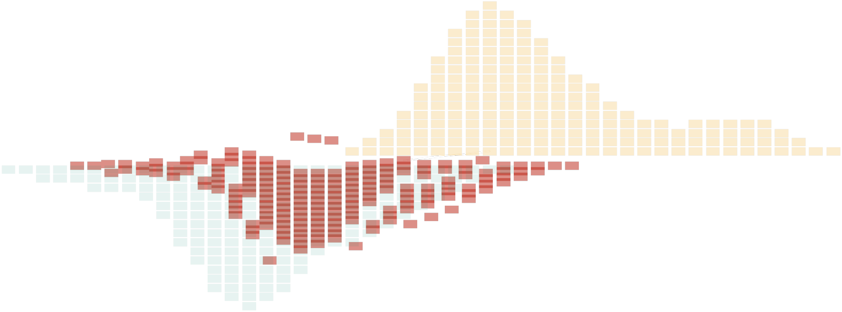
Monge (1781), “Mémoire sur la théorie des déblais et des remblais”

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



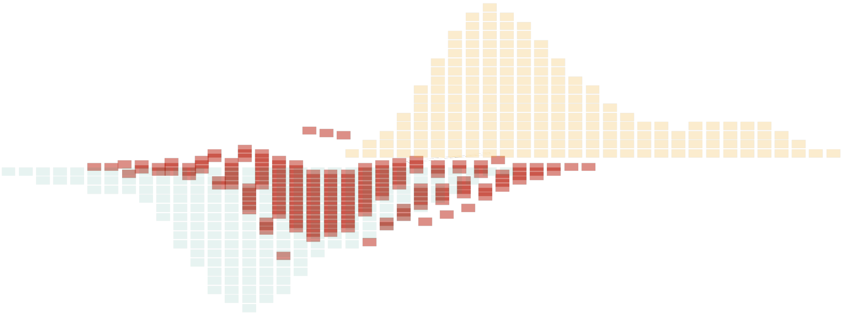
Monge (1781), “Mémoire sur la théorie des déblais et des remblais”

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



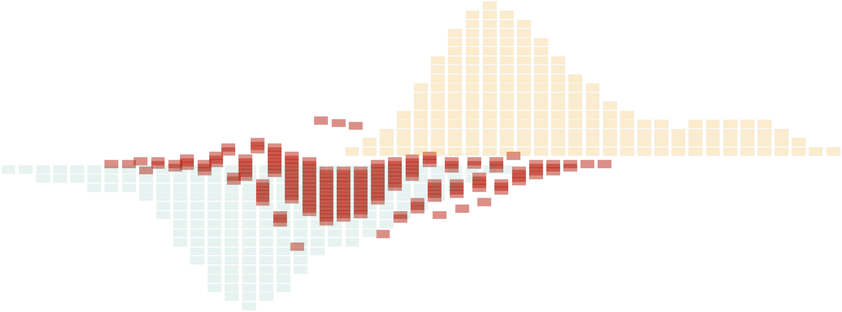
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

"Optimal Transport" (a side note / a cultural interlude)



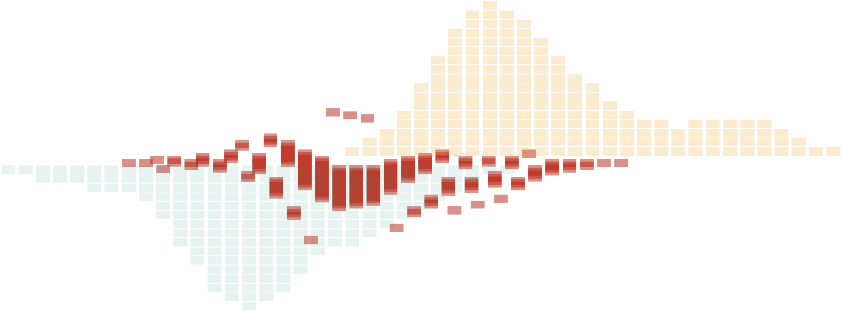
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

"Optimal Transport" (a side note / a cultural interlude)



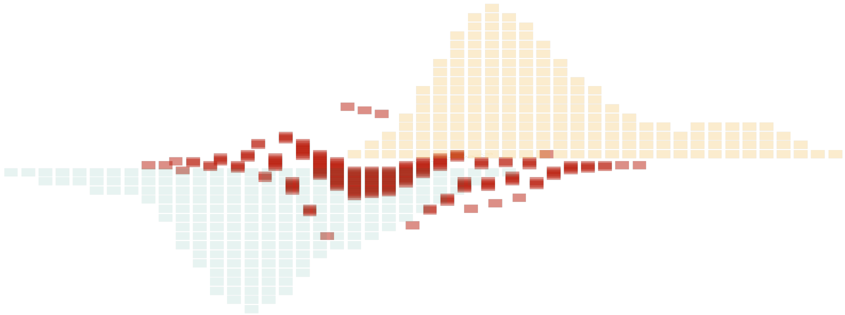
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

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



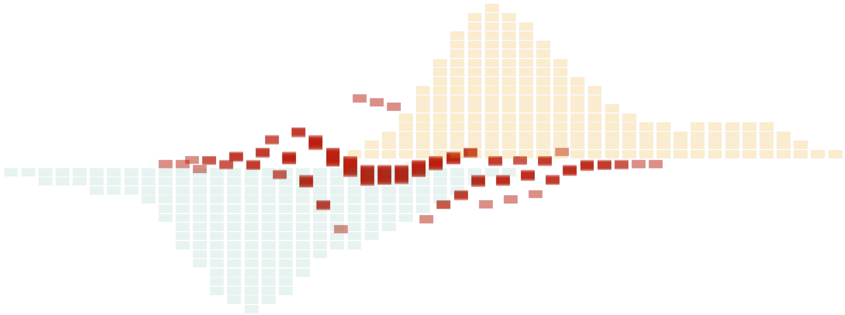
Monge (1781), “Mémoire sur la théorie des déblais et des remblais”

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



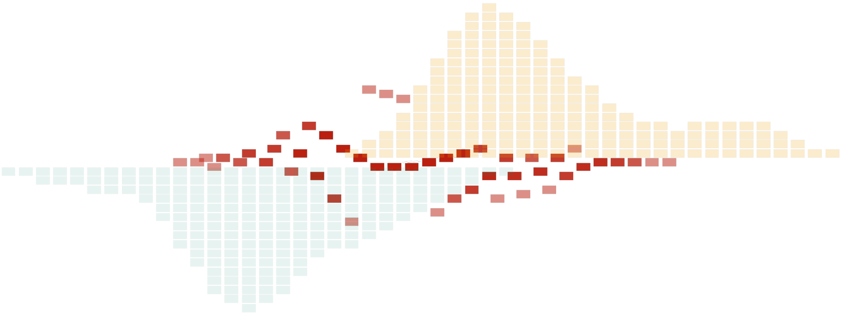
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



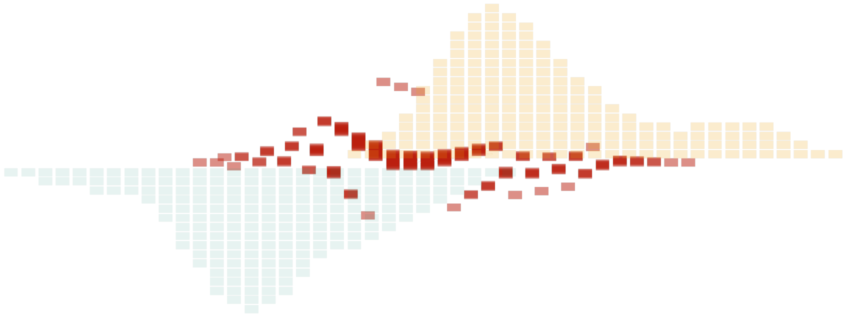
Monge (1781), “Mémoire sur la théorie des déblais et des remblais”

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



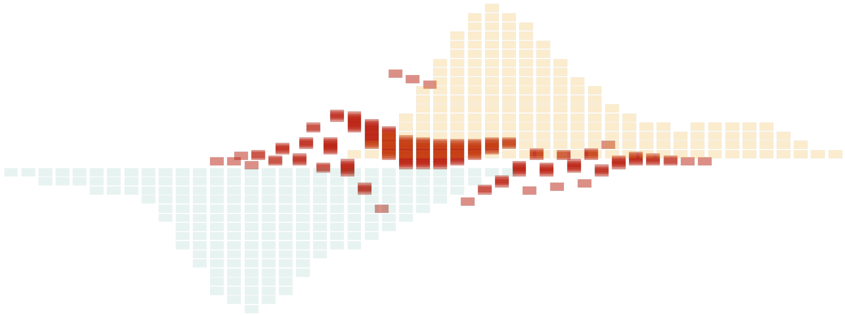
Monge (1781), “Mémoire sur la théorie des déblais et des remblais”

"Optimal Transport" (a side note / a cultural interlude)



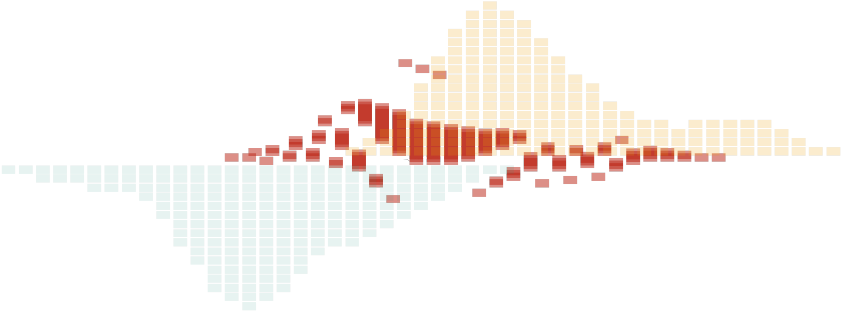
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

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



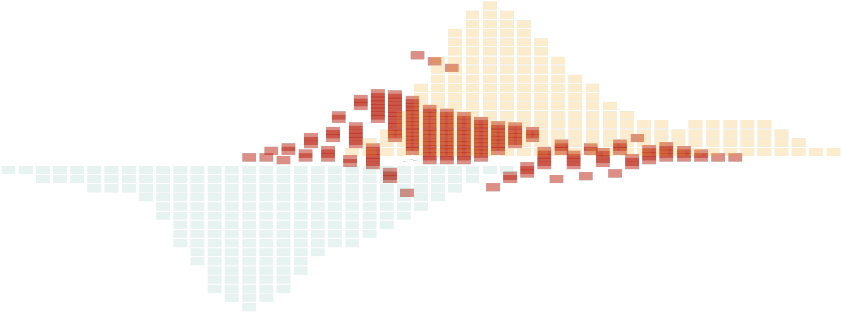
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



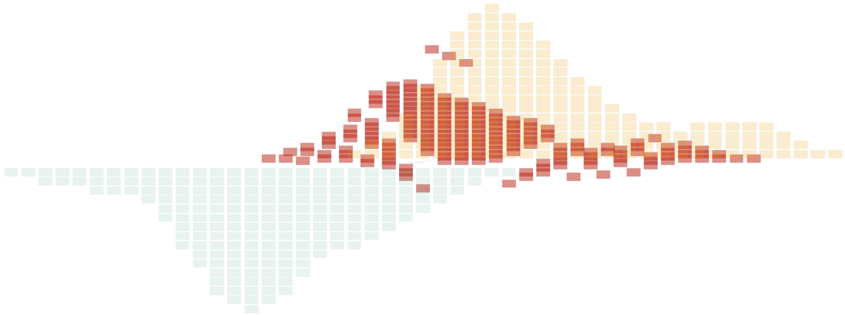
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



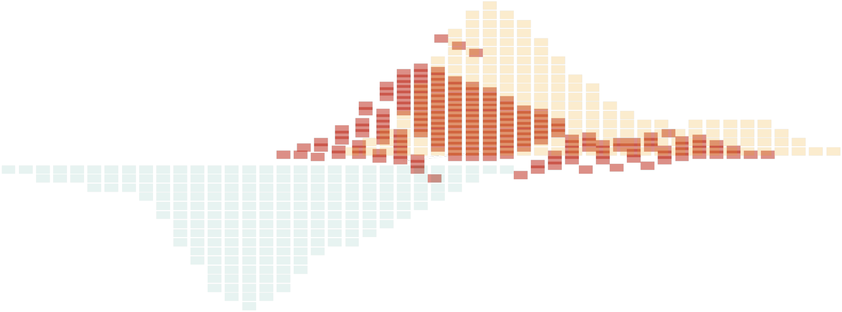
Monge (1781), “Mémoire sur la théorie des déblais et des remblais”

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



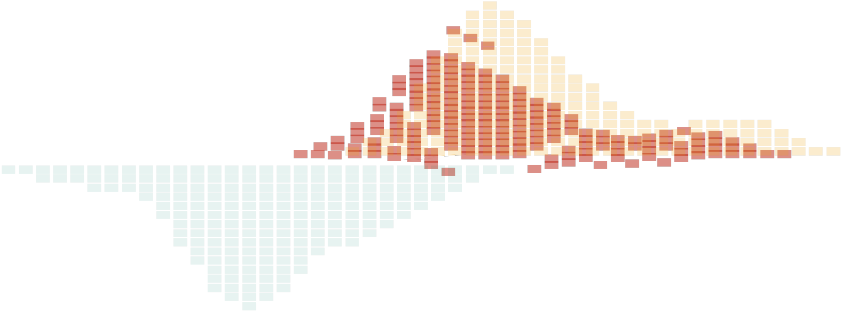
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

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



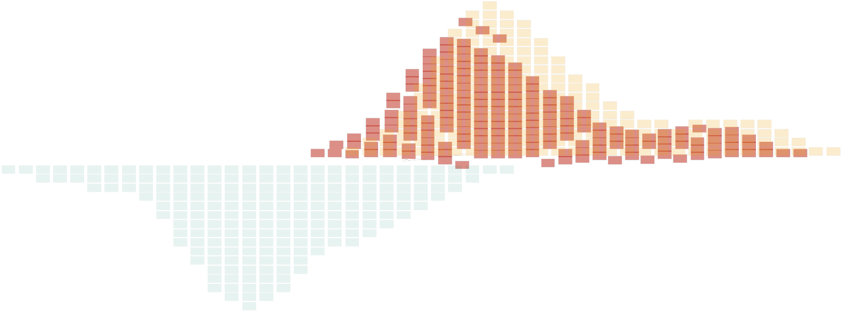
Monge (1781), “Mémoire sur la théorie des déblais et des remblais ”

"Optimal Transport" (a side note / a cultural interlude)



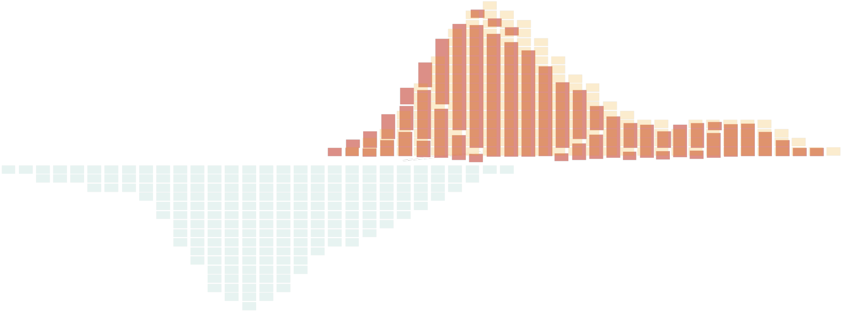
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

"Optimal Transport" (a side note / a cultural interlude)



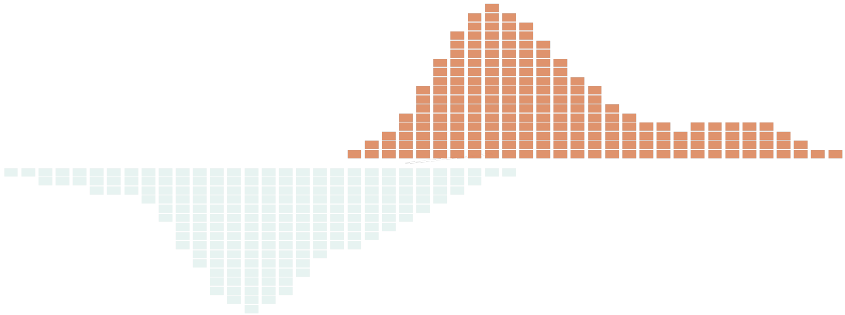
Monge (1781), "Mémoire sur la théorie des déblais et des remblais"

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



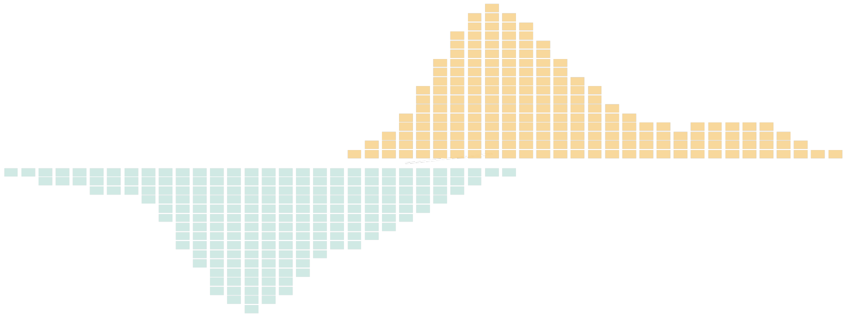
Monge (1781), “Mémoire sur la théorie des déblais et des remblais”

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



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

"Optimal Transport" (a side note / a cultural interlude)



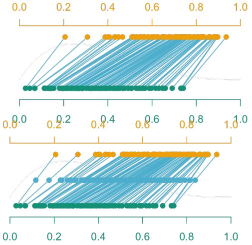
This "monotone" (increasing) mapping is optimal $x_i^A \xrightarrow{T^\star} y_i^B$

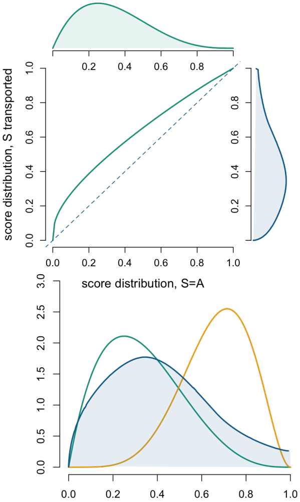
Mitigating Discrimination with Wasserstein Barycenters

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

For individual i, why not

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





Mitigating Discrimination with Wasserstein Barycenters

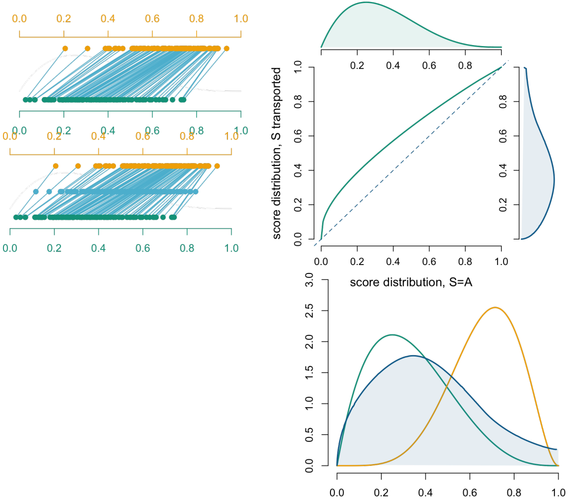
Mitigation is about finding some $m^\star$ "in-between" (Demographic Parity)

For individual i, why not

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

corresponding to

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



Mitigating Discrimination with Wasserstein Barycenters

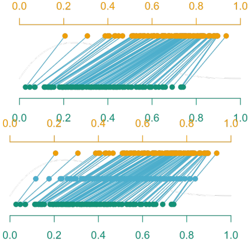
Mitigation is about finding some $m^\star$ "in-between" (Demographic Parity)

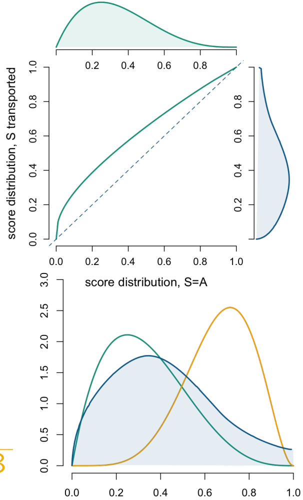
For individual i, why not

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

corresponding to

$$m^\star(x, A) = \underbrace{\frac{1}{2}}_{\mathbb{P}[S=A]} m^A(x) + \underbrace{\frac{1}{2}}_{\mathbb{P}[S=B]} \underbrace{T^\star(m^A(x))}_{\text{associated score in group B}}$$

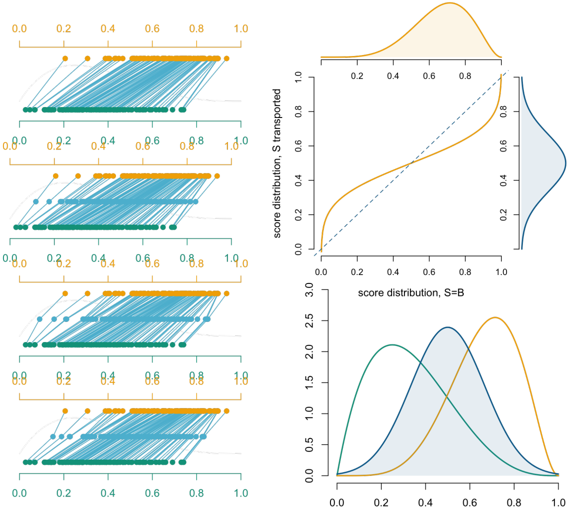




Mitigating Discrimination with Wasserstein Barycenters

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

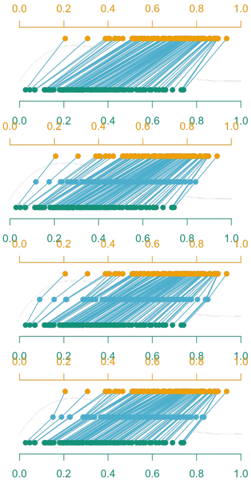
other "averages" could be considered

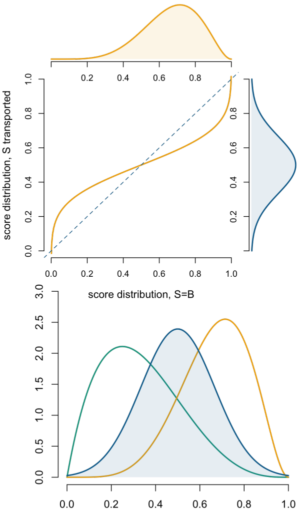


Mitigating Discrimination with Wasserstein Barycenters

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



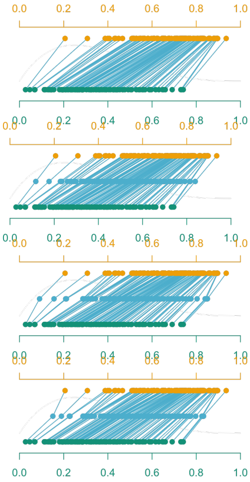


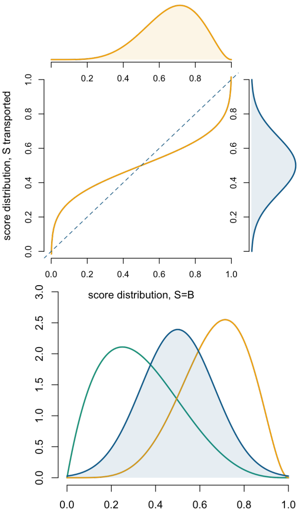
Mitigating Discrimination with Wasserstein Barycenters

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

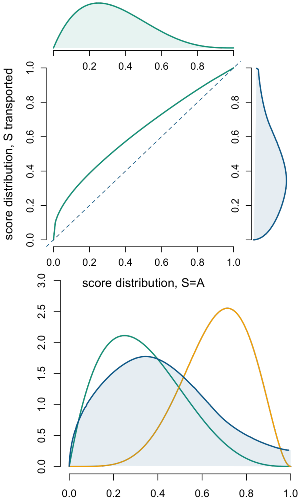
Given a model m (regression, boosting, random forest, neural nets, etc) we can easily derive a "**fair model**"





Mitigating Discrimination with Wasserstein Barycenters

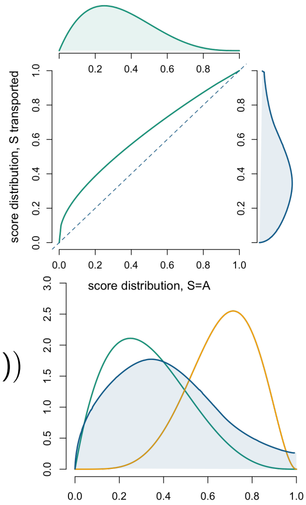
$$\begin{cases} m^\star(\boldsymbol{x}, s = A) = \mathbb{P}[S = A] \cdot m(\boldsymbol{x}, s = A) \\ \qquad\qquad\qquad + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\boldsymbol{x}, s = A)) \\ m^\star(\boldsymbol{x}, s = B) = \mathbb{P}[S = A] \cdot F_A^{-1} \circ F_B(m(\boldsymbol{x}, s = B)) \\ \qquad\qquad\qquad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



Mitigating Discrimination with Wasserstein Barycenters

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

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

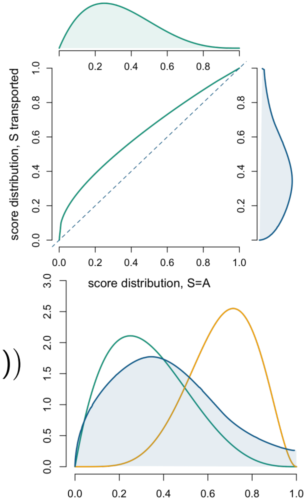


Mitigating Discrimination with Wasserstein Barycenters

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

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

weights



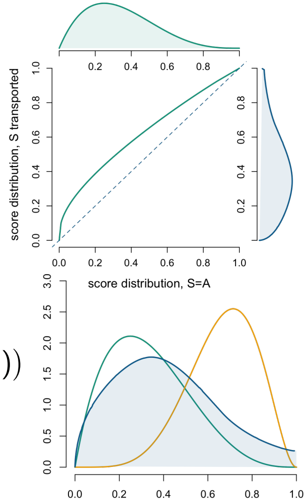
Mitigating Discrimination with Wasserstein Barycenters

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

score in group A

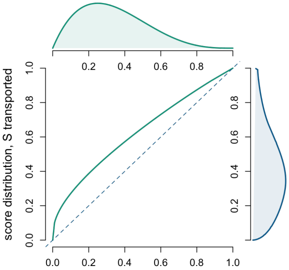
$$
\mathbb{P}[S=A]\cdot m(\boldsymbol{x}, s=A) + \mathbb{P}[S=B]\cdot F_B^{-1}\circ F_A(m(\boldsymbol{x}, s=A))
$$

weights



Mitigating Discrimination with Wasserstein Barycenters

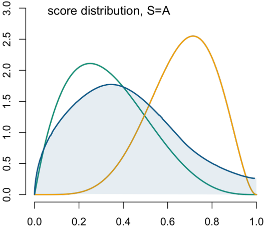
$$
\begin{cases}
m^\star(\boldsymbol{x}, s = A) = \mathbb{P}[S = A] \cdot m(\boldsymbol{x}, s = A) \\
\qquad\qquad\qquad + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\boldsymbol{x}, s = A)) \\
m^\star(\boldsymbol{x}, s = B) = \mathbb{P}[S = A] \cdot F_A^{-1} \circ F_B(m(\boldsymbol{x}, s = B)) \\
\qquad\qquad\qquad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



score in group A

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

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



Mitigating Discrimination with Wasserstein Barycenters

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

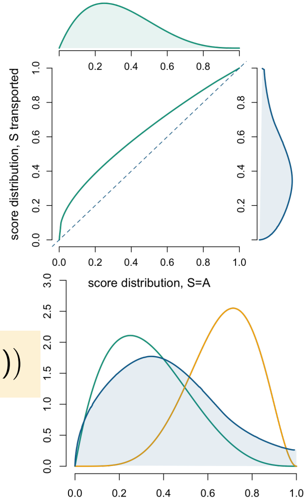
score in group A

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

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

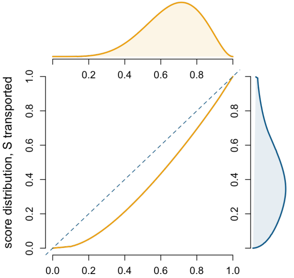
weights

quantile score in group B associated with probability p



Mitigating Discrimination with Wasserstein Barycenters

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

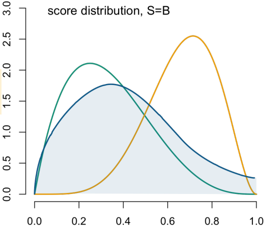


$p = F_B(m(\boldsymbol{x}, s=B))$

score in group B

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

quantile score in group A associated with probability p



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,

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

Fairness is achieved by projection onto a fair subspace

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

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

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

Mitigation with Wasserstein Barycenter

Recall that Bayes estimator is the best model, for the ℓ_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.

Back to the COMPAS Example

$$\begin{cases} S: & \text{race (binary), black \& white} \\ Y: & \text{re-offense (binary), no \& yes} \\ \widehat{Y}: & \text{classifier (risk category), low \& high} \end{cases}$$

(standard) demographic parity would be translated as

$$\mathbb{P}[\underbrace{\widehat{Y} = \text{high}}_{\text{predictions}} \mid \overbrace{S = \text{black}}^{\text{sensitive}}] = 58\% \overset{?}{=} \mathbb{P}[\underbrace{\widehat{Y} = \text{high}}_{\text{predictions}} \mid \overbrace{S = \text{white}}^{\text{sensitive}}] = 33\%,$$

Back to the COMPAS Example, from Discrimination to Calibration

$$\begin{cases} S: & \text{race (binary), black \& white} \\ Y: & \text{re-offense (binary), no \& yes} \\ \widehat{Y}: & \text{classifier (risk category), low \& high} \end{cases}$$

$$\mathbb{P}[\underbrace{\widehat{Y}=\text{high}|Y=\text{no}}_{\text{false positive rate}}, \overbrace{S=\text{black}}^{\text{sensitive}}] = 42\% \overset{?}{=} \mathbb{P}[\underbrace{\widehat{Y}=\text{high}|Y=\text{no}}_{\text{false positive rate}}, \overbrace{S=\text{white}}^{\text{sensitive}}] = 22\%,$$

$$\mathbb{P}[\underbrace{Y=\text{no}|\widehat{Y}=\text{high}}_{\text{false discovery rate}}, S=\text{black}] = 35\% \overset{?}{=} \mathbb{P}[\underbrace{Y=\text{no}|\widehat{Y}=\text{high}}_{\text{false discovery rate}}, S=\text{white}] = 40\%.$$

From Discrimination to Calibration

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

From Discrimination to Calibration

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

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

From Discrimination to Calibration

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

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

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

From Discrimination to Calibration (an Epistemological Detour)

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)

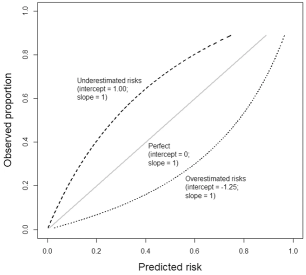
From Discrimination to Calibration (an Epistemological Detour)

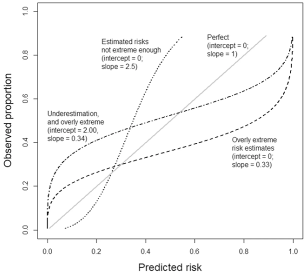
As explained in Van Calster et al. (2019), ”*among patients with an estimated risk of 20%, we expect 20 in 100 to have or to develop the event*,”

- 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))





From Discrimination to Calibration

Designed to Deceive: Do These People Look Real to You?
by Kashmir Hill and Jeremy White Nov. 21, 2020,

The New York Times

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)

From Discrimination to Calibration

Designed to Deceive: Do These People Look Real to You?
by Kashmir Hill and Jeremy White Nov. 21, 2020,

The New York Times

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)

From Discrimination to Calibration

Designed to Deceive: Do These People Look Real to You?
by Kashmir Hill and Jeremy White Nov. 21, 2020,

The New York Times

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)

From Discrimination to Calibration

Designed to Deceive: Do These People Look Real to You?
by Kashmir Hill and Jeremy White Nov. 21, 2020,

The New York Times

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)

From Discrimination to Calibration

Designed to Deceive: Do These People Look Real to You?
by Kashmir Hill and Jeremy White Nov. 21, 2020,

The New York Times

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)

From Discrimination to Calibration

Designed to Deceive: Do These People Look Real to You?
by Kashmir Hill and Jeremy White Nov. 21, 2020,

The New York Times

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)

From Discrimination to Calibration

Designed to Deceive: Do These People Look Real to You?
by Kashmir Hill and Jeremy White Nov. 21, 2020,

The New York Times

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)

From Discrimination to Calibration

Designed to Deceive: Do These People Look Real to You?
by Kashmir Hill and Jeremy White Nov. 21, 2020,

The New York Times

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)

From Discrimination to Calibration

Designed to Deceive: Do These People Look Real to You?
by Kashmir Hill and Jeremy White Nov. 21, 2020,

The New York Times

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)

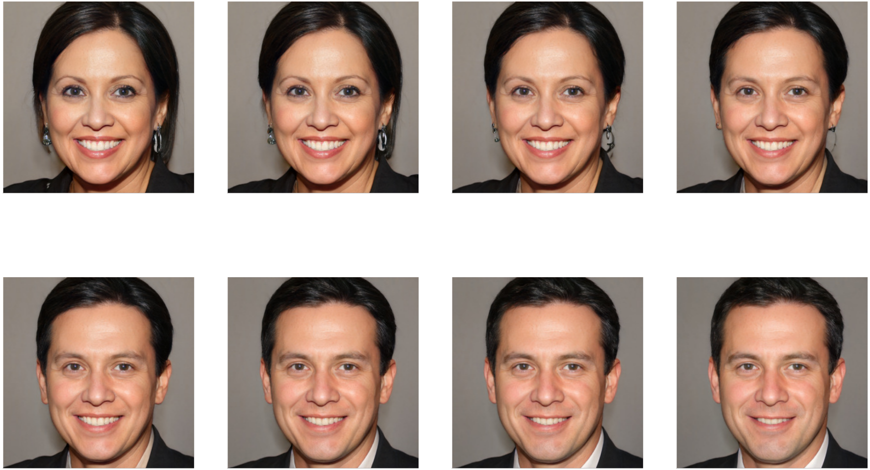
From Discrimination to Calibration

Designed to Deceive: Do These People Look Real to You?
by Kashmir Hill and Jeremy White Nov. 21, 2020,

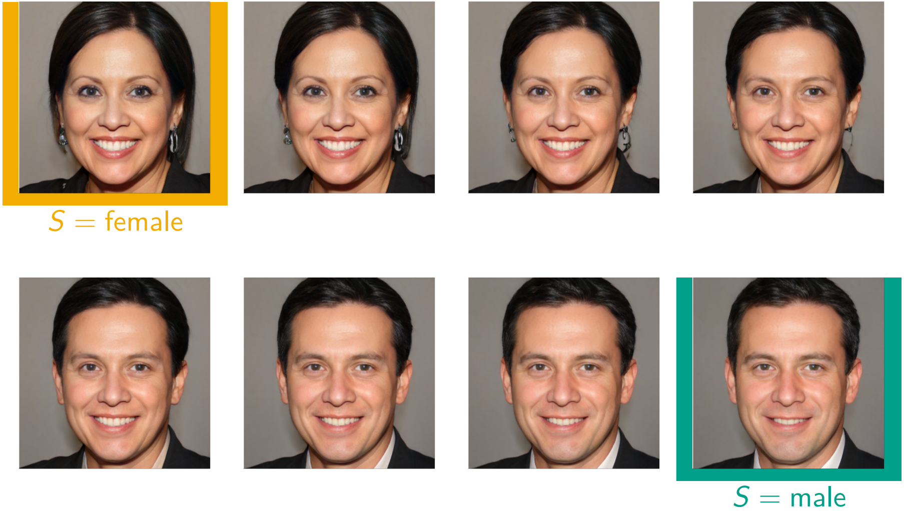
The New York Times

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)

From Discrimination to Calibration



From Discrimination to Calibration



From Discrimination to Calibration

female (0.984)
male (0.016)

female (0.983)
male (0.017)

female (0.982)
male (0.018)

female (0.960)
male (0.040)

female (0.009)
male (0.991)

female (0.013)
male (0.987)

female (0.014)
male (0.986)

female (0.015)
male (0.985)

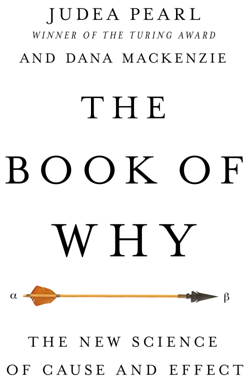
Individual Fairness

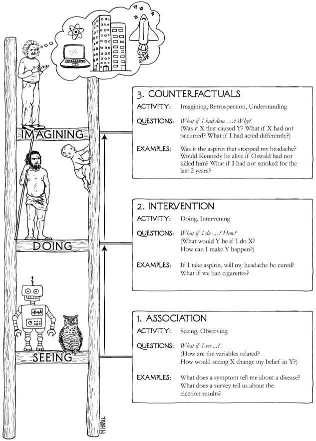
We have **counterfactual fairness** if "*had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same*," Kusner et al. (2017)

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)

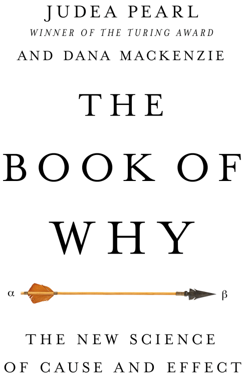


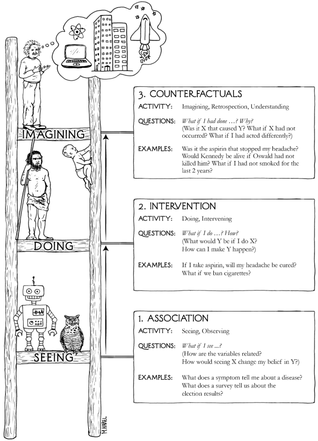


Individual Fairness

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)





- 1. **Association**
(Seeing, "*what if I see...*")

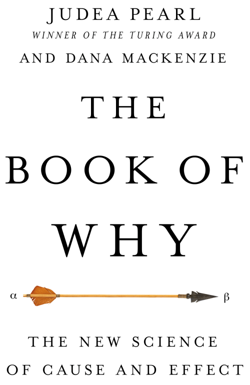
Individual Fairness

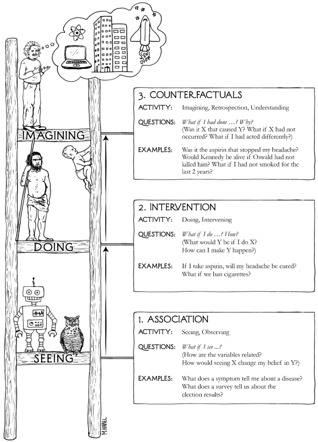
We have **counterfactual fairness** if "*had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same*," Kusner et al. (2017)

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

- 2. **Intervention**
(Doing, "*what if I do...*")

- 1. **Association**
(Seeing, "*what if I see...*")



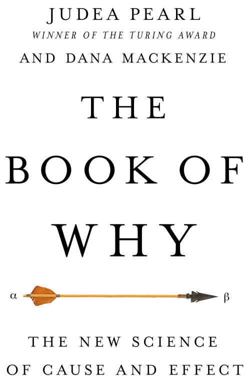


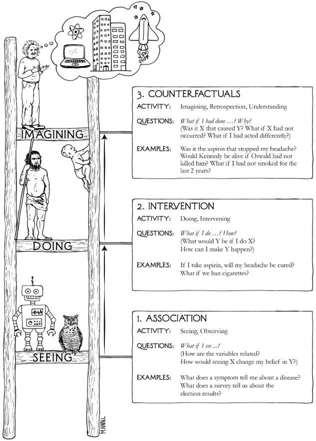
Individual Fairness

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...*")





What About Interpretation ?

"*Humans think in stories rather than facts, numbers or equations - and the simpler the story, the better*," Harari (2018)

For Glenn (2000), insurer's risk selection process has two sides:

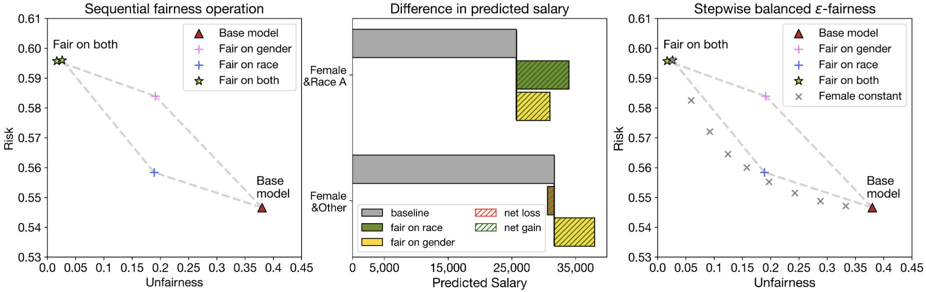
- the one presented to regulators and policyholders (numbers, statistics and objectivity),
- the other presented to underwriters (stories, character and subjective judgment).

The rhetoric of insurance exclusion – numbers, objectivity and statistics – forms what Brian Glenn calls "*the myth of the actuary*," "*a powerful rhetorical situation in which decisions appear to be based on objectively determined criteria when they are also largely based on subjective ones*" or "*the subjective nature of a seemingly objective process*".

What About Interpretation ?

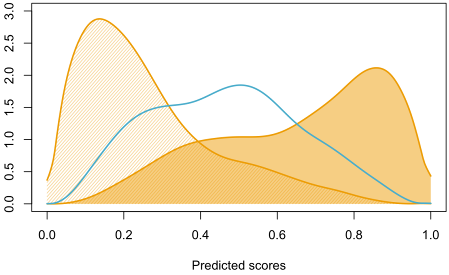
"*The fact that the selection of risk factors is subjective and contingent upon narratives of risk and responsibility has in the past played a far larger role than whether or not someone with a wood stove is charged higher premiums.*"

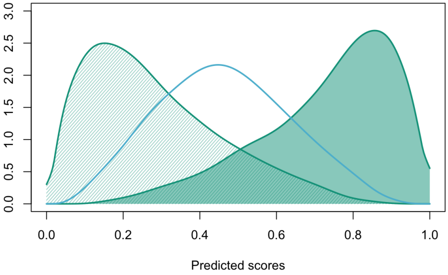
Going further, "*virtually every aspect of the insurance industry is predicated on stories first and then numbers*," Glenn (2003)



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 $\rightarrow$ Single sensitive attribute (SSA), by intersection,

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

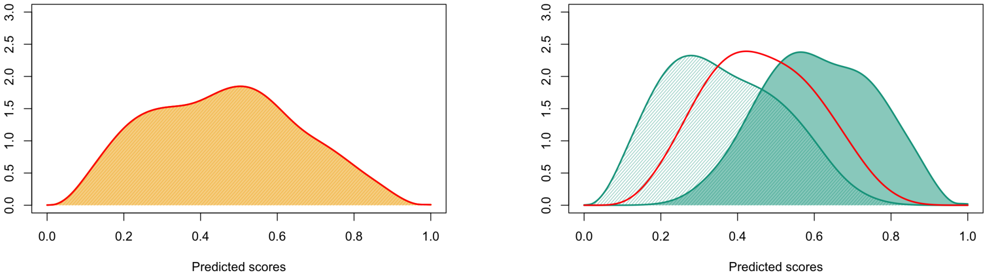
Here $\mathcal{A}$ corresponds to $4 = 2 \times 2$ states,

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

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

The Case of Multiple Attributes

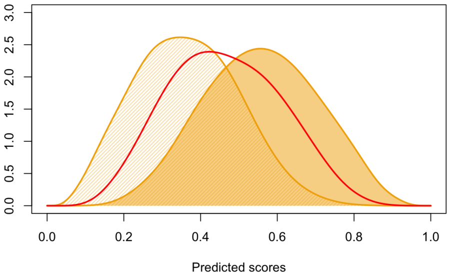
- Given $\nu_{m_{B_1}}$, consider
  - the barycenter $\nu_{m_{B_1}}$ conditional on A_1 (no impact, already fair)
  - the barycenter $\nu_{m_{B_2}}$ conditional on A_2

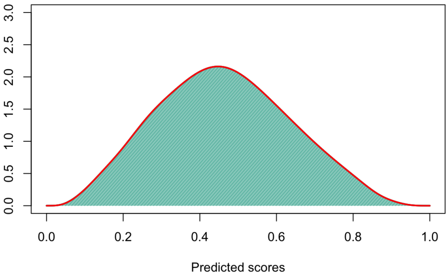


- On the right, distribution of $\nu_{m_{B_2}\circ m_{B_1}}$

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)

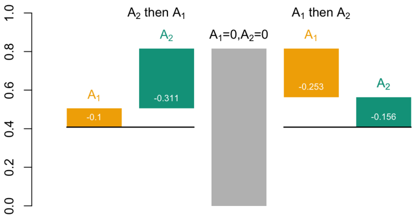


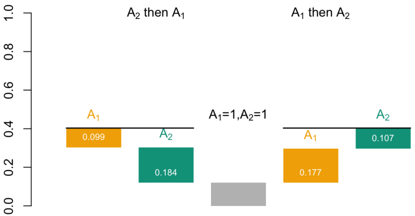


- On the left, distribution of $\nu_{m_{B_1} \circ m_{B_2}}$

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





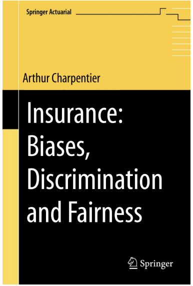
Mitigating Discrimination ? (brief conclusion)

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.



References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*, May 23.

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.

Bénéplanc, G., Charpentier, A., and Thourot, P. (2022). *Manuel d'assurance*. Presses Universitaires de France.

Charpentier, A. (2014). *Computational actuarial science with R*. CRC press.

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

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

Denuit, M. and Charpentier, A. (2004). *Mathématiques de l'assurance non-vie: Tome I Principes fondamentaux de théorie du risque*. Economica.

Denuit, M. and Charpentier, A. (2005). *Mathématiques de l'assurance non-vie: Tome II Tarification et provisionnement*. Economica.

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

References

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4):1.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

Feeley, M. and Simon, J. (1994). Actuarial justice: The emerging new criminal law. *The futures of criminology*, 173:174.

Feller, A., Pierson, E., Corbett-Davies, S., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, October 17.

Glenn, B. J. (2000). The shifting rhetoric of insurance denial. *Law and Society Review*, pages 779–808.

Glenn, B. J. (2003). Postmodernism: the basis of insurance. *Risk Management and Insurance Review*, 6(2):131–143.

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

Harari, Y. N. (2018). *21 Lessons for the 21st Century*. Random House.

References

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.

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.

References

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.

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

Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

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

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.

References

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.