Fairness and discrimination in actuarial pricing

Arthur Charpentier[1], Laurence Barry[2], Vincent Grari[3], Lamprier Sylvain[3], Detyniecki Marcin[4]

[1] Univeristé du Québec à Montréal [2] Chaire Pari [3] Sorbonne Université [4] AXA

Séminaire de la chaire DIALog (Digital Insurance And Long-term risks) 2022

Agenda

- Barry and Charpentier (2022) The Fairness of Machine Learning in Insurance: New Rags for an Old Man?, *ArXiv:2205.08112*
- Charpentier (2022) Insurance: Discrimination, Biases and Fairness, *Institut Louis Bachelier*
- Grari et al. (2022) A fair pricing model via adversarial learning, *ArXiv:2202.12008*

Ethics, Fairness and Discrimination I

NEWSLETTERS
Sign up to read our regular email newsletters

NewScientist

SUBSCRIBE AND SAVE 69%

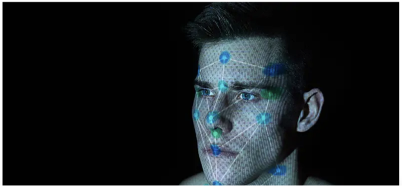
News Podcasts Video Technology Space Physics Health More Shop Courses Events Tours Jobs

Discriminating algorithms: 5 times AI showed prejudice

Artificial intelligence is supposed to make life easier for us all – but it is also prone to amplify sexist and racist biases from the real world

TECHNOLOGY 12 April 2018 , updated 27 April 2018

By Daniel Cossins



TRENDING LATEST VIDEO FREE

1. Dingo genome suggests Australian icon not descended from domestic dogs
2. How *Minecraft* is helping children with autism make new friends
3. A third of people aged over 70 are sexually active, survey reveals
4. Harmful air pollution now affects 99 per cent of everyone on Earth
5. Breaking the News exhibition shows Edward Snowden's smashed drives

PROPUBLICA

Donate

How We Analyzed the COMPAS Recidivism Algorithm

by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin

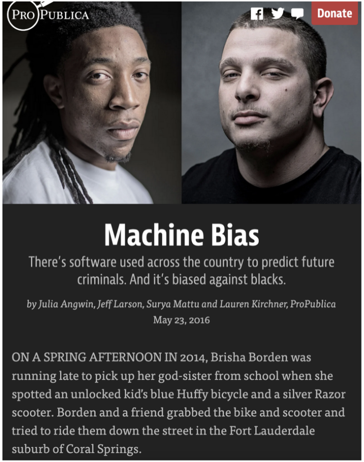
May 23, 2016

← Read the story

Across the nation, judges, probation and parole officers are increasingly using algorithms to assess a criminal defendant's likelihood of becoming a recidivist – a term used to describe criminals who re-offend. There are dozens of these risk assessment algorithms in use. Many states have built their own assessments, and several academics have written tools. There are also two leading nationwide tools offered by commercial vendors.

We set out to assess one of the commercial tools made by Northpointe, Inc. to discover the underlying accuracy of their recidivism algorithm and to test whether the algorithm was biased against certain groups.

Ethics, Fairness and Discrimination II



PROPUBLICA

Donate

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

AKE ACTION

IABA

PUBLIC CRITICISM OF INSURANCE PRICING PRACTICES

A 2015 Study by the Consumer Federation of America (CFA) claimed that "on average, a good driver in a predominantly African American Community will pay considerably more for state-mandated auto insurance coverage than a similarly situated driver in a predominantly White community."

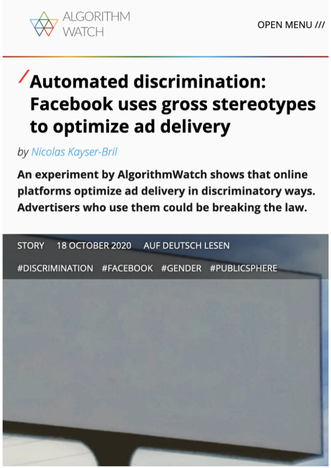
An analysis by ProPublica and Consumer Reports in 2017 drew a similar conclusion, stating that "this disparity may amount to a subtler form of redlining, a term that traditionally refers to denial of services or products to minority areas." While there were methodological flaws in the analysis, the article raised a number of questions about whether insurance rates were biased against minorities.

RACE AND INSURANCE

CONGRESS TARGETS DISCRIMINATION IN AUTO INSURANCE

H.R. 1756: Preventing Credit Score Discrimination in Auto Insurance Act
This bill intended to amend the Fair Credit Reporting Act to prohibit the use of consumer credit information in auto insurance decision-making but did not receive a vote in the 116th Congress.

H.R. 3693: Prohibit Auto Insurance Discrimination Act



ALGORITHM WATCH

OPEN MENU ///

Automated discrimination: Facebook uses gross stereotypes to optimize ad delivery

by Nicolas Kayser-Bril

An experiment by AlgorithmWatch shows that online platforms optimize ad delivery in discriminatory ways. Advertisers who use them could be breaking the law.

STORY 18 OCTOBER 2020 AUF DEUTSCH LESEN

#DISCRIMINATION #FACEBOOK #GENDER #PUBLICSPHERE

Ethics, Fairness and Discrimination III

Fairness Through Awareness

Cynthia Dwork* Moritz Hardt† Toniann Pitassi‡ Omer Reingold§ Richard Zemel¶

November 30, 2011

Algorithmic decision making and the cost of fairness

Sam Corbett-Davies, Stanford University, scorbett@stanford.edu
Emma Pierson, Stanford University, emmap1@stanford.edu
Avi Feller, Univ. of California, Berkeley, afeller@berkeley.edu
Sharad Goel, Stanford University, scgoel@stanford.edu
Aziz Huq, University of Chicago, huq@uchicago.edu

Equality of Opportunity in Supervised Learning

Moritz Hardt Eric Price Nathan Srebro

October 11, 2016

Fairness in Criminal Justice Risk Assessments: The State of the Art

Richard Berk[a,b], Hoda Heidari[c], Shahin Jabbari[c], Michael Kearns[c], Aaron Roth[c]

Fair prediction with disparate impact: A study of bias in recidivism prediction instruments

Alexandra Chouldechova *

The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*

Sam Corbett-Davies, Stanford University
Sharad Goel, Stanford University

August 14, 2018

HUMAN DECISIONS AND MACHINE PREDICTIONS*

Jon Kleinberg
Himabindu Lakkaraju
Jure Leskovec
Jens Ludwig
Sendhil Mullainathan

The Frontiers of Fairness in Machine Learning

Alexandra Chouldechova* Aaron Roth†

October 23, 2018

DOI:10.1145/3376898

A group of industry, academic, and government experts convene in Philadelphia to explore the roots of algorithmic bias.

BY ALEXANDRA CHOULDECHOVA AND AARON ROTH

A Snapshot of the Frontiers of Fairness in Machine Learning

A Survey on Bias and Fairness in Machine Learning

NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, and ARAM GALSTYAN, USC-ISI

FAIRNESS IN MACHINE LEARNING: A SURVEY

A PREPRINT

Simon Caton, University College Dublin, Dublin, Ireland, simon.caton@ucd.ie
Christian Haas, University of Nebraska at Omaha, Omaha, US, christianhaas@unomaha.edu

Ethics, Fairness and Discrimination IV

- "*at the core of insurance business lies discrimination between risky and non-risky insureds*", Avraham (2017)

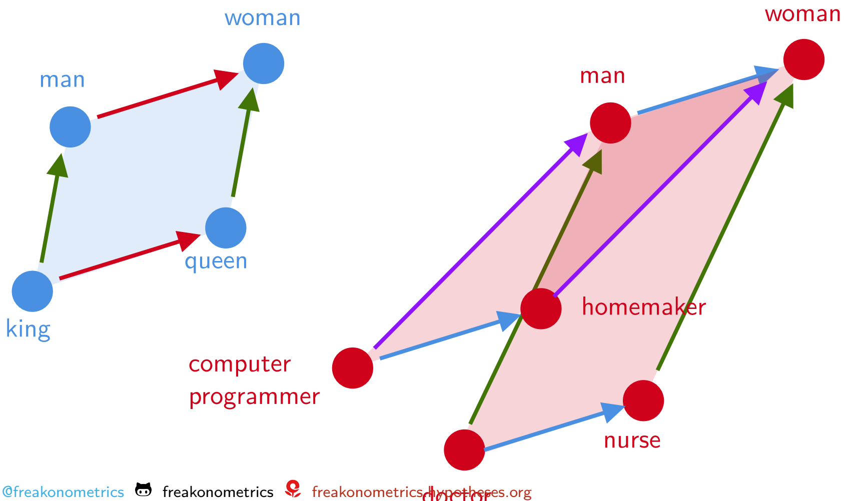
- "*Technology is neither good nor bad; nor is it neutral* " , Kranzberg (1986)
- "*Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitely ask for* ", Kearns and Roth (2019)

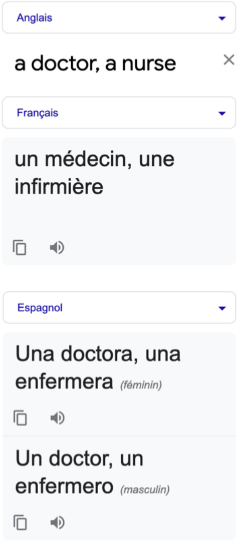
It is a complex problem...

- Accuracy : $\pi(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X}=\boldsymbol{x}]$ ($\mathbb{P}$ historical probability) (is)
- Fairness : $\pi^\star(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}^\star}[Y|\boldsymbol{X}=\boldsymbol{x}]$ ($\mathbb{P}^\star$ targeted probability) (ought, Hume (1739))

Ethics, Fairness and Discrimination V

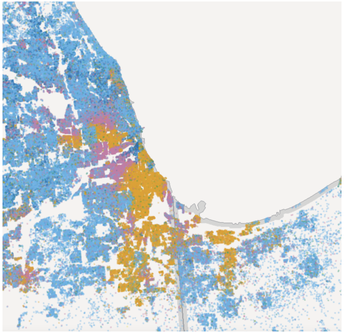
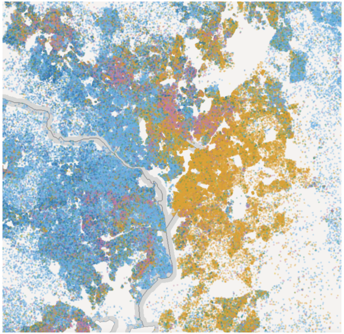
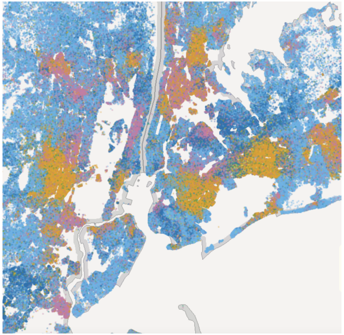
Word embedding exhibiting gender bias, Bolukbasi et al. (2016)





Ethics, Fairness and Discrimination VI

Spatial information and racial bias (redlining)

Protected Attributes ?

	CA	HI	GA	NC	NY	MA	PA	FL	TX	AL	ON	NB	NL	QC
Gender	×	×	●	×	●	×	×	●	●	●	●	×	×	●
Age	×	×	●	×*	●	×	●	●	●	●*	●	×	×	●
Driving experience	●	×	●	●	●	●	●	●	●	●	●	●	●	●
Credit history	×	×	●	●	●	×	●*	●	●	×*	×	●*	×	●
Education	×	×	×	×	×	×	●	●	●	●	●	●	●	●
Occupation	×	×	×	●	×	×	●	●	●	●	●	●	●	●
Employment status	×	×	×	●	×	×	●	●	●	●	●	●	●	●
Marital status	●	×	●	●	●	×	●	●	●	●	●	●	●	●
Housing situation	×	×	●	●	●	×	●	●	●	×	×	●	●	●
Address/ZIP code	●	●	●	●	●	●	●	●	●	×	×	●	●	●
Insurance history	●	●	●	●	●	●	●	●	●	●	●	●	●	●

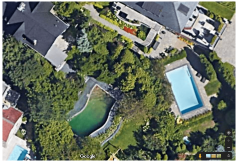
CA: Californie, HI: Hawaii, GA: Georgia, NC: Caroline du nord, NY: New York, MA: Massachusetts, PA: Pennsylvanie, FL: Floride, TX: Texas, AL: Alberta, ON: Ontario, NB: Nouveau-Brunswick, NL: Terre-Neuve-et-Labrador, QC: Québec

Big Data and Proxies I

More and more features, possibly (strongly) correlated with protected variables

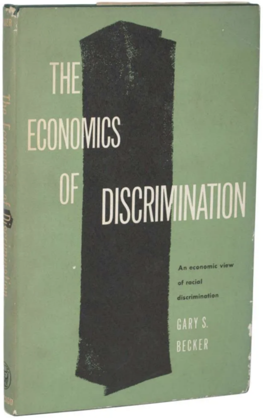
- location (policyholder home adress)

  Jean et al. (2016), Seresinhe et al. (2017), Gebru et al. (2017), Law et al. (2019), Ilic et al. (2019), Kita and Kidziński (2019), see also redlining



Big Data and Proxies II

Bohren et al. (2019) on statistical discrimination, or efficient discrimination, as in Becker (1957) (inspired by Edgeworth (1922) up to Phelps (1972))
Becker (2005) says "*if young Moslem Middle Eastern males were in fact much more likely to commit terrorism against U.S. than were other groups, putting them through tighter security clearance would reduce current airport terrorism*",



"racial profiling" is "effective", even though "*such profiling is 'unfair' to the many young male Moslems who are not terrorists, and to the many minority shoppers who are honest ...*"

Big Data and Proxies III

"... *That could be made up in part by compensating groups who are forced to go through more careful airport screening through putting them in shorter security lines, or in other ways. Similarly, innocent shoppers who are stopped and searched could be compensated for their embarrassment and time*"

See also Boczar et al. (2021) (Insurance Data Science Conference in London)

phrenology (Lombroso (1876) and Bertillon and Chervin (1909)) and ugly laws (TenBroek (1966) and Burgdorf and Burgdorf Jr (1974))

source https://nvlabs-fi-cdn.nvidia.com/stylegan2-ada-pytorch/, cf Karras et al. (2020)

Big/Small Data and Possible Pitfalls I

Missing an important covariate is an important issue

- $y_i = \beta_0 + \boldsymbol{x}_1^\top \boldsymbol{\beta}_1 + \boldsymbol{x}_2^\top \boldsymbol{\beta}_2 + \varepsilon_i$: true model
- $y_i = b_0 + \boldsymbol{x}_1^\top \boldsymbol{b}_1 + \eta_i$: estimated model

Maximum likelihood estimator of $\boldsymbol{b}_1$ is

$$
\begin{aligned}
\widehat{\boldsymbol{b}}_1 &= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{y} \\
&= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top [\boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}] \\
&= (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_2 + (\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta}_1 + \underbrace{(\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{X}_2 \boldsymbol{\beta}_2}_{\boldsymbol{\beta}_{12}} + \underbrace{(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon}}_{\nu_i}
\end{aligned}
$$

so that $\mathbb{E}[\widehat{\boldsymbol{b}}_1] = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{12} \neq \boldsymbol{\beta}_1$.

Big/Small Data and Possible Pitfalls II

From Bickel et al. (1975) (see also Alipourfard et al. (2018))

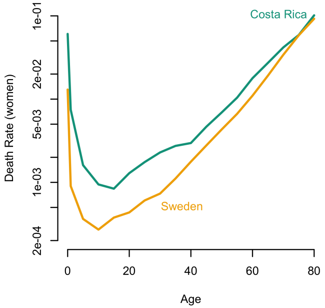
	Total	Men	Women	Proportions
Total	$5233/12763 \sim 41\%$	$3714/8442 \sim \mathbf{44}\%$	$1512/4321 \sim 35\%$	66%-34%
Top 6	$1745/4526 \sim 39\%$	$1198/2691 \sim \mathbf{45}\%$	$557/1835 \sim 30\%$	59%-41%
A	$597/933 \sim 64\%$	$512/825 \sim 62\%$	$89/108 \sim \mathbf{82}\%$	88%-12%
B	$369/585 \sim 63\%$	$353/560 \sim 63\%$	$17/\ \ 25 \sim \mathbf{68}\%$	96%- 4%
C	$321/918 \sim 35\%$	$120/325 \sim \mathbf{37}\%$	$202/593 \sim 34\%$	35%-65%
D	$269/792 \sim 34\%$	$138/417 \sim 33\%$	$131/375 \sim \mathbf{35}\%$	53%-47%
E	$146/584 \sim 25\%$	$53/191 \sim \mathbf{28}\%$	$94/393 \sim 24\%$	33%-67%
F	$43/714 \sim \ \ 6\%$	$22/373 \sim \ \ 6\%$	$24/341 \sim \ \ \mathbf{7}\%$	52%-48%

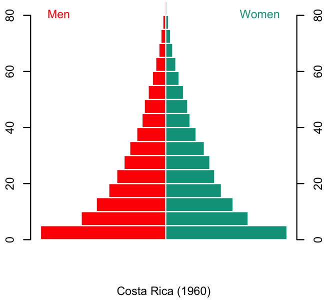
Big/Small Data and Possible Pitfalls III

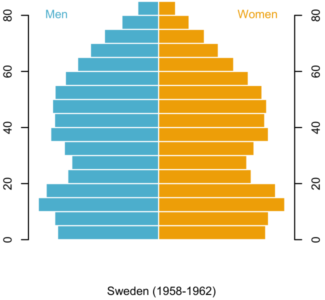
As mentioned in Cohen (1986),

$$\mathbb{P}[T \leq 1 | \boldsymbol{X} = \text{Costa Rica}] < \mathbb{P}[T \leq 1 | \boldsymbol{X} = \text{Sweden}]$$

$$\mathbb{P}[T \leq 1 | \boldsymbol{X} = (\text{Costa Rica}, x)] > \mathbb{P}[T \leq 1 | \boldsymbol{X} = (\text{Sweden}, x)], \ \forall x$$







Defining Group Fairness I

$$\begin{cases} y \in \{0,1\} & \text{variable of interest (classically binary)} \\ p \in \{0,1\} & \text{protected variable (sensitive)} \\ \boldsymbol{x} \in \mathbb{R}^d & \text{`explanatory' variables} \\ s \in [0,1] & \text{score, classically } s = s(\boldsymbol{x}, p) \\ \widehat{y} \in \{0,1\} & \text{classifier, classically } \widehat{y} = \mathbf{1}(s > t) \end{cases}$$

Fairness Through Unawareness, Kusner et al. (2017)
Protected attribute p is not explicitly used in decision function $\widehat{y}$.

- what is y ?
- how to define a fair pricing ?

Defining Group Fairness II

policyholder

$(\boldsymbol{x}_i, p_i)$

Defining Group Fairness III

policyholder

$(\boldsymbol{x}_i, p_i)$

infered data

$m_d(\boldsymbol{x}, \boldsymbol{p})$

possibles scores

Defining Group Fairness IV

policyholder

$(\boldsymbol{x}_i, p_i)$

underwriting

$m_u(\boldsymbol{x}, \boldsymbol{p})$

infered data

$m_d(\boldsymbol{x}, \boldsymbol{p})$

- commercial discount
- multiple decisions from policyholder (related to various personal features)
- multiple decisions from agent / broker

Defining Group Fairness V

policyholder

$(\boldsymbol{x}_i, p_i)$

underwriting

$m_u(\boldsymbol{x}, \boldsymbol{p})$

claim a loss

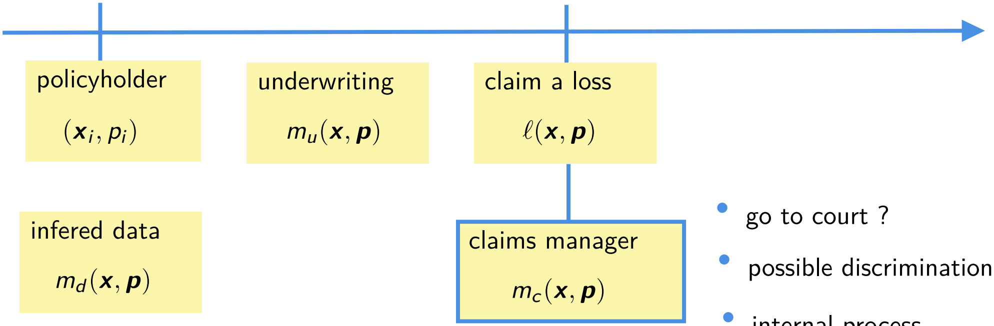
$\ell(\boldsymbol{x}, \boldsymbol{p})$

(true value of the loss)

infered data

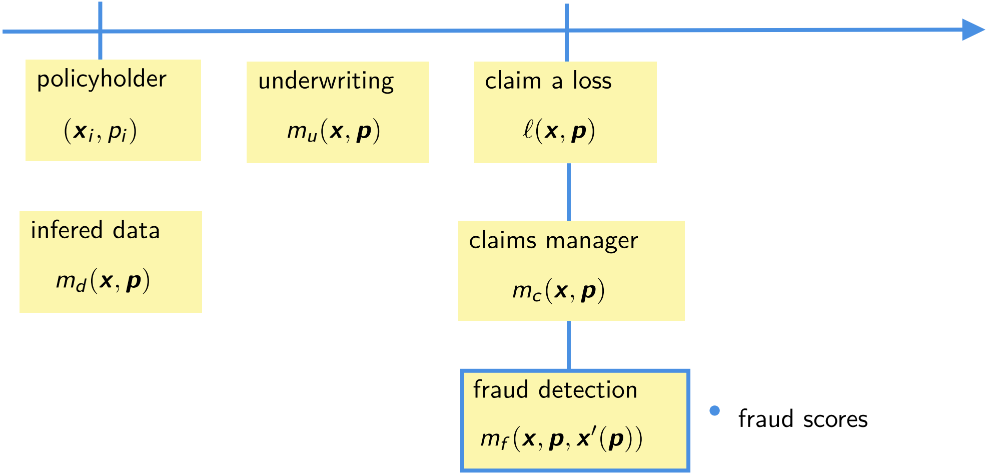
$m_d(\boldsymbol{x}, \boldsymbol{p})$

Defining Group Fairness VI



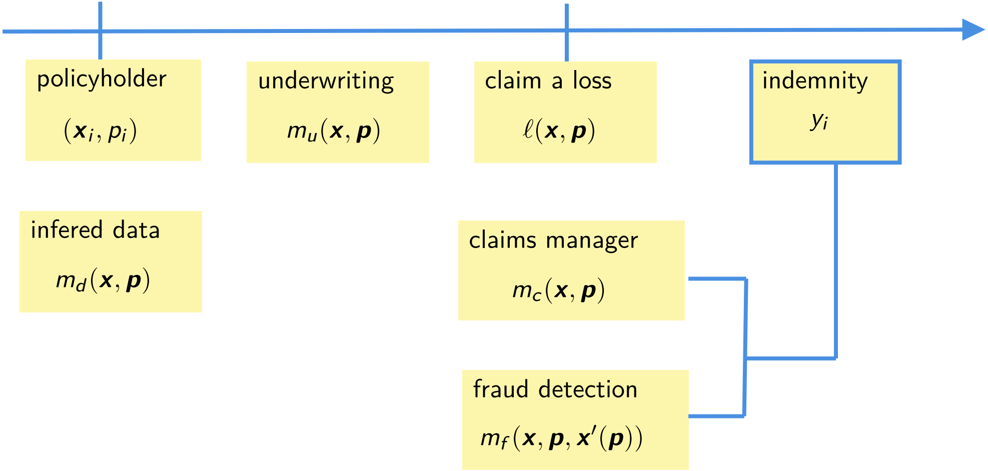
- go to court ?
- possible discrimination
- internal process

Defining Group Fairness VII



- fraud scores

Defining Group Fairness VIII



Defining Group Fairness IX

policyholder

$(\boldsymbol{x}_i, p_i)$

underwriting

$m_u(\boldsymbol{x}, \boldsymbol{p})$

claim a loss

$\ell(\boldsymbol{x}, \boldsymbol{p})$

indemnity

y_i

infered data

$m_d(\boldsymbol{x}, \boldsymbol{p})$

claims manager

$m_c(\boldsymbol{x}, \boldsymbol{p})$

dataset

$(y_i, \boldsymbol{x}_i, \boldsymbol{p}_i)$

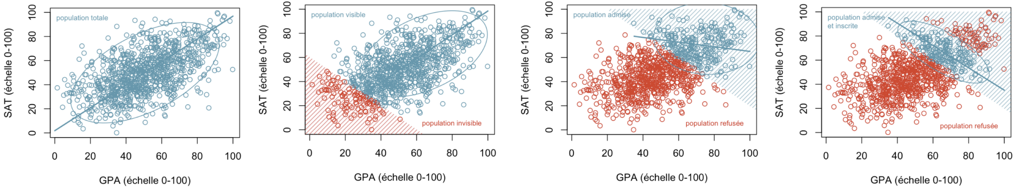
fraud detection

$m_f(\boldsymbol{x}, \boldsymbol{p}, \boldsymbol{x}'(\boldsymbol{p}))$

(training data)

Defining Group Fairness X

Not to mention multiple biases, from feedback bias (telematics and gamification) to selection bias



(see Hand (2020) on dark data)

Defining Group Fairness XI

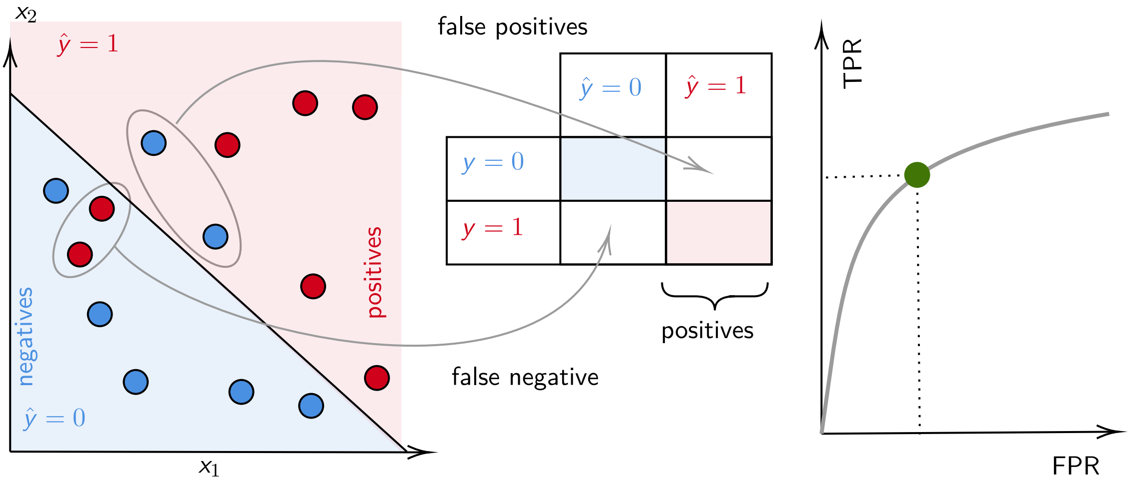
$$\begin{cases} y \in \{0,1\} & \text{variable of interest (classically binary)} \\ p \in \{0,1\} & \text{protected variable (sensitive)} \\ \boldsymbol{x} \in \mathbb{R}^d & \text{`explanatory' variables} \\ s \in [0,1] & \text{score, classically } s = s(\boldsymbol{x}, p) \\ \widehat{y} \in \{0,1\} & \text{classifier, classically } \widehat{y} = \mathbf{1}(s > t) \end{cases}$$

Demographic Parity, (Corbett-Davies et al. (2017), Agarwal (2021))
Decision function $\widehat{y}$ satisfies demographic parity if $\widehat{Y} \perp\!\!\!\perp P$, i.e.

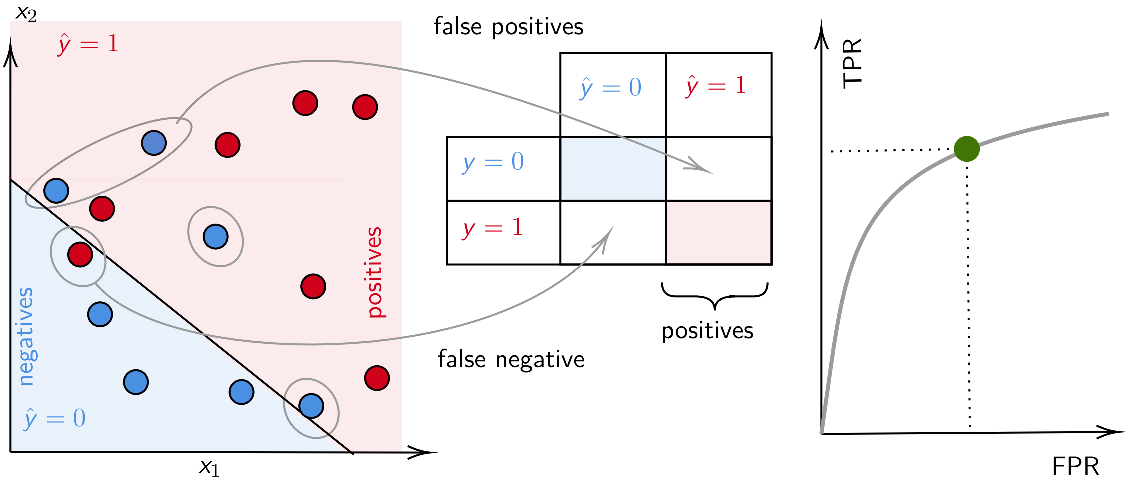
$$\mathbb{P}[\widehat{Y} = y | P = 0] = \mathbb{P}[\widehat{Y} = y | P = 1],\ \forall y \text{ or } \mathbb{E}[\widehat{Y} | P = 0] = \mathbb{E}[\widehat{Y} | P = 1]$$

Since y is binary, a classical intermediary component is the score $s(\boldsymbol{x})$
and a classical tool is the ROC curve (obtained by changing threshold t)

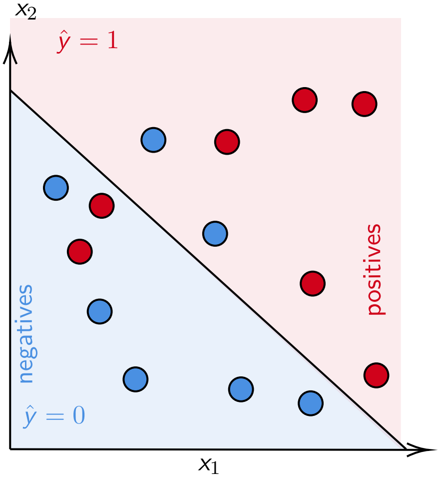
Defining Group Fairness XII

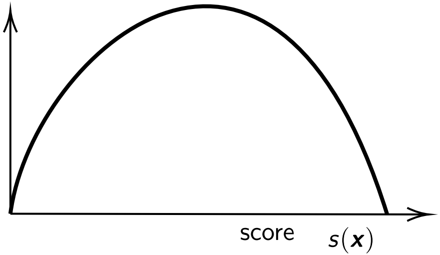


Defining Group Fairness XIII

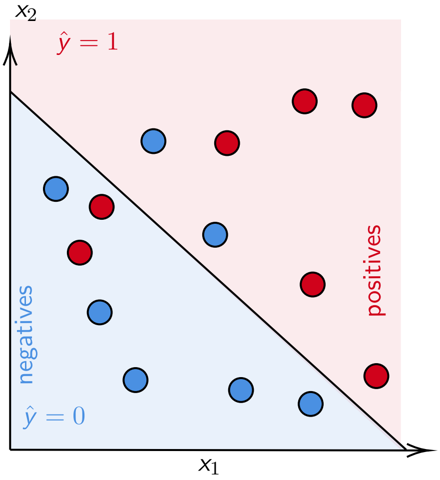


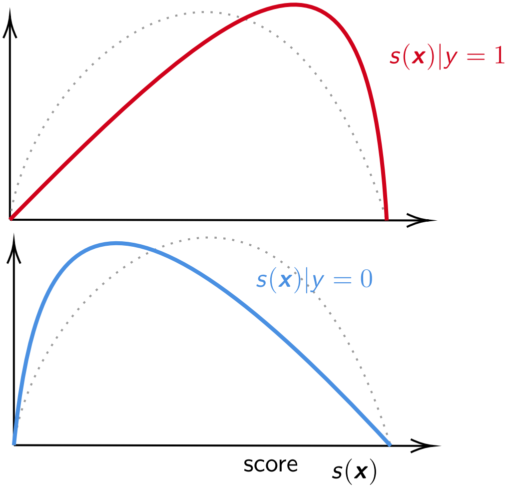
Defining Group Fairness XIV



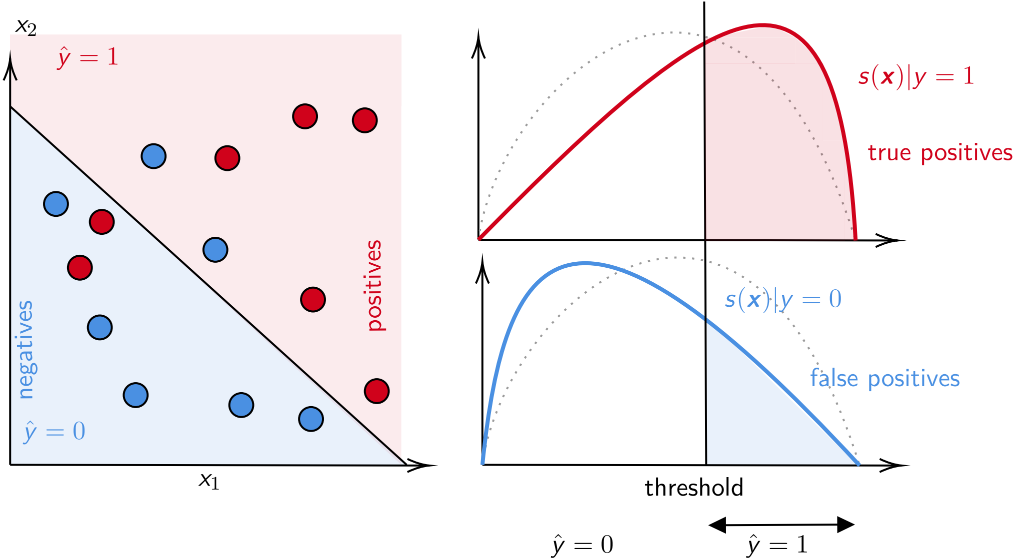


Defining Group Fairness XV



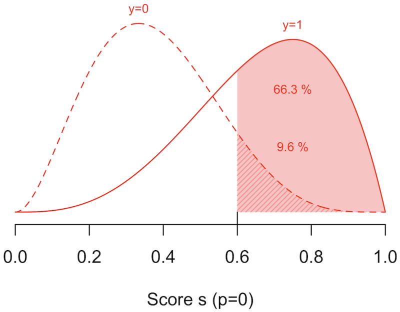


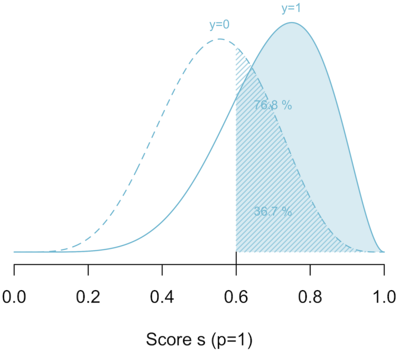
Defining Group Fairness XVI



Defining Group Fairness XVII

We can compare $s(\boldsymbol{X})$ conditional on Y, but also on P





Defining Group Fairness XVIII

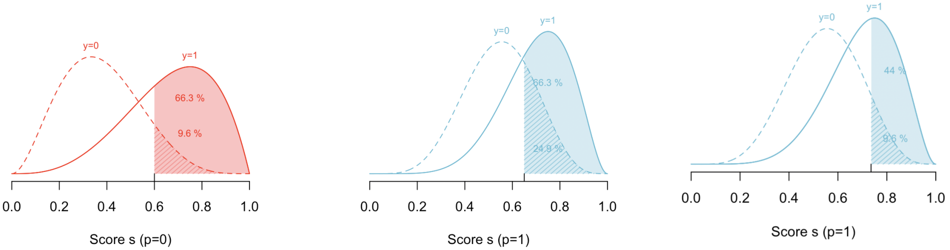
Equal Opportunity, Hardt et al. (2016)

True positive parity

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

or false positive parity

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



Defining Group Fairness XIX

statistical parity	Dwork et al. (2012)	$\mathbb{P}[\widehat{Y}=1\|P=p]=\text{cst},\ \forall p$	independence
conditional statistical parity	Corbett-Davies et al. (2017)	$\mathbb{P}[\widehat{Y}=1\|P=p,X=x]=\text{cst}_x,\ \forall p,y$	$\widehat{Y}\perp\!\!\!\perp P$
equalized odds	Hardt et al. (2016)	$\mathbb{P}[\widehat{Y}=1\|P=p,Y=y]=\text{cst}_y,\ \forall p,y$	separation
equalized opportunity	Hardt et al. (2016)	$\mathbb{P}[\widehat{Y}=1\|P=p,Y=1]=\text{cst},\ \forall p$	
predictive equality	Corbett-Davies et al. (2017)	$\mathbb{P}[\widehat{Y}=1\|P=p,Y=0]=\text{cst},\ \forall p$	$\widehat{Y}\perp\!\!\!\perp P\mid Y$
balance (positive)	Kleinberg et al. (2017)	$\mathbb{E}[S\|P=p,Y=1]=\text{cst},\ \forall p$	$S\perp\!\!\!\perp P\mid Y$
balance (negative)	Kleinberg et al. (2017)	$\mathbb{E}[S\|P=p,Y=0]=\text{cst},\ \forall p$	
conditional accuracy equality	Berk et al. (2017)	$\mathbb{P}[Y=y\|P=p,\widehat{Y}=y]=\text{cst}_y,\ \forall p,y$	sufficiency
predictive parity	Chouldechova (2017)	$\mathbb{P}[Y=1\|P=p,\widehat{Y}=1]=\text{cst},\ \forall p$	
calibration	Chouldechova (2017)	$\mathbb{P}[Y=1\|P=p,S=s]=\text{cst}_s,\ \forall p,s$	$Y\perp\!\!\!\perp P\mid\widehat{Y}$
well-calibration	Chouldechova (2017)	$\mathbb{P}[Y=1\|P=p,S=s]=s,\ \forall p,s$	
accuracy equality	Berk et al. (2017)	$\mathbb{P}[\widehat{Y}=Y\|P=p]=\text{cst},\ \forall p$	
treatment equality	Berk et al. (2017)	$\dfrac{\text{FN}_p}{\text{FP}_p}=\text{cst}_p,\ \forall p$	

From correlation to causality I

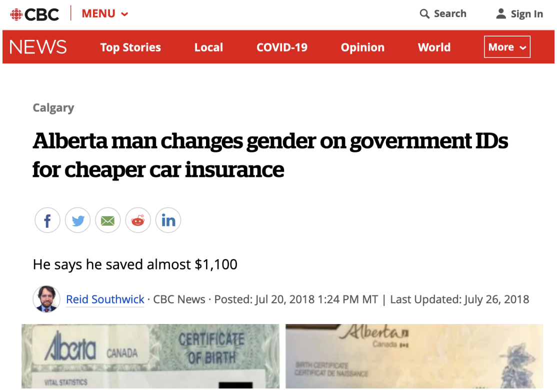
- "*classifying projection methods as using demographic/actuarial models or non-demographic/causal models*"

  Keilman (2003) and Hudson (2007)
- "*Article 5(2) allowed Member States to Permit proportionate differences in individuals premiums and benefits where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data.*"

  Thiery and Van Schoubroeck (2006) and Schmeiser et al. (2014)
- "*Two judges on the Supreme Court dissented in the Zurich case. In their view, an insurer must not only prove a statistical correlation between a particular group and higher risk, but a causal connection*"

  Gomery et al. (2011)

From correlation to causality II

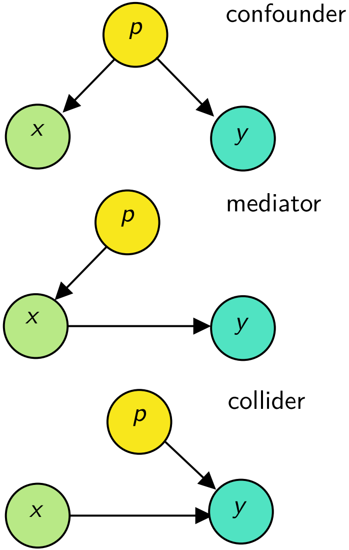

CBC | MENU Search Sign In
NEWS Top Stories Local COVID-19 Opinion World More

Calgary

Alberta man changes gender on government IDs for cheaper car insurance

He says he saved almost $1,100

Reid Southwick · CBC News · Posted: Jul 20, 2018 1:24 PM MT | Last Updated: July 26, 2018



- DAGs are important
- Looking for a counterfactual

From correlation to causality III

Consider some distances D on $\{0,1\}\times\{0,1\}$ or $[0,1]\times[0,1]$, and d on $\mathbb{R}^p\times\mathbb{R}^p$,

Lipschitz property, Duivesteijn and Feelders (2008)

$$D(\widehat{y}_i,\widehat{y}_j) \text{ or } D(s_i,s_j)\leq d(\boldsymbol{x}_i,\boldsymbol{x}_j),\ \forall i,j=1,\cdots,n.$$

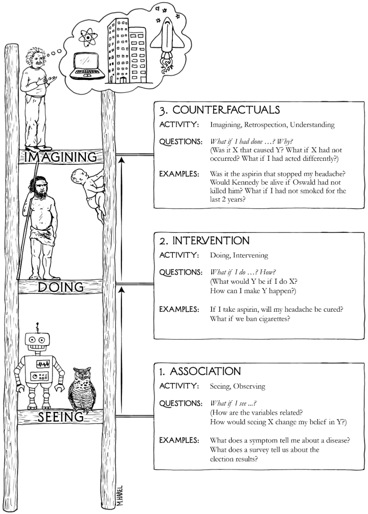
Conterfactual fairness, Kusner et al. (2017) If the prediction in the real world is the same as the prediction in the counterfactual world where the individual would have belonged to a different demographic group, we have counterfactual equity, i.e.

$$\mathbb{P}\big[Y^\star_{P\leftarrow p}=y|\boldsymbol{X}=\boldsymbol{x}\big]=\mathbb{P}\big[Y^\star_{P\leftarrow p'}=y|\boldsymbol{X}=\boldsymbol{x}\big],\ \forall p',\boldsymbol{x},y.$$

From correlation to causality IV

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

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



(picture Pearl & Mackenzie (2018))

From correlation to causality V

Causal inference litterature,

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

	treatement	outcome			age	gender	height	weight
	t_i	y_i	$y^\star_{i,T\leftarrow 1}$	$y^\star_{i,T\leftarrow o}$	$x_{1,i}$	$x_{2,i}$	$x_{3,i}$	$x_{4,i}$
1	1	121	121	?	37	F	160	56
2	0	109	?	109	28	F	156	54
3	1	162	162	?	53	M	190	87

There will be a significant impact of treatment t on y if $y^\star_{T\leftarrow 0} \neq y^\star_{T\leftarrow 1}$ (see Rubin (1974), Hernán and Robins (2010) or Imai (2018)).
The causal effect for individual i is $\tau_i = y^\star_{i,T\leftarrow 1} - y^\star_{i,T\leftarrow 0}$

From correlation to causality VI

One can define the sample average treatment effect (SATE)

$$\text{SATE} = \frac{1}{n}\sum_{i=1}^{n} y^\star_{i,T\leftarrow 1} - y^\star_{i,T\leftarrow 0}$$

the average treatment effect (ATE)

$$\tau = \text{ATE} = \mathbb{E}\big[Y^\star_{i,T\leftarrow 1} - Y^\star_{i,T\leftarrow 0}\big]$$

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

$$\tau(\boldsymbol{x}) = \text{CATE}(\boldsymbol{x}) = \mathbb{E}\big[Y^\star_{i,T\leftarrow 1} - Y^\star_{i,T\leftarrow 0}\big|\boldsymbol{X} = \boldsymbol{x}\big]$$

From correlation to causality VII

Naive difference-in-means estimators, but Splawa-Neyman et al. (1990)

$$\widehat{\tau}=\frac{1}{n_1}\sum_{i=1}^n t_iy_i-\frac{1}{n_0}\sum_{i=1}^n(1-t_i)y_i,\text{ where }n_j=\sum_{i=1}^n t_i$$

or the regression-based estimators (also called single learner)

$$\widehat{\tau}(\boldsymbol{x})=\widehat{\mu}_1(\boldsymbol{x})-\widehat{\mu}_0(\boldsymbol{x})$$

where $\widehat{\mu}_t(\boldsymbol{x})=\mathbb{E}[Y|T=t,\boldsymbol{X}=\boldsymbol{x}]$
One can also use the propensity score $\pi_i=\mathbb{P}[T=1|\boldsymbol{X}=\boldsymbol{x}_i]$

$$\widehat{\tau}=\frac{1}{n}\sum_{i=1}^n\frac{t_iy_i}{\pi_i}-\frac{(1-t_i)y_i}{1-\pi_i}$$

to derive a inverse-propensity score weighted estimator.
This can be use to derive a doubly-robust estimator

Counterfactual and optimal transport I

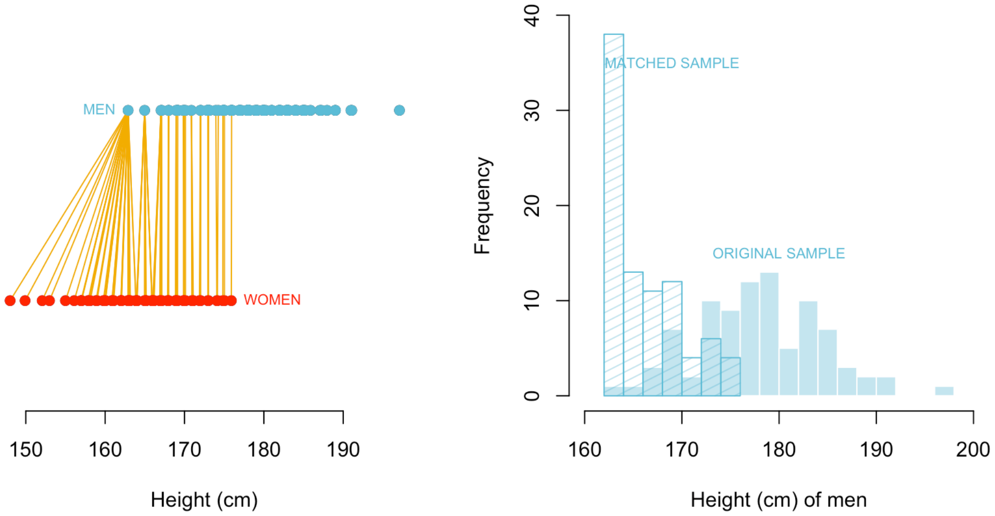
One can also use matching technique,

- for treated individual i ($t_i = 1$), match that individual to someone in the non-treated group ($t_j = 0$), $j_i^\star = \underset{j:t_j=0}{\text{argmin}}\{d(\boldsymbol{x}_i, \boldsymbol{x}_j)\}$,

E.g. x_i is the height of a person, t_i is the gender of that person, then

$$\widehat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n_1} t_i \cdot \big(y_i - y_{j_i^\star}\big)$$

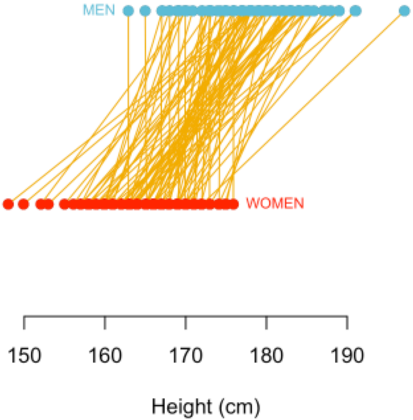
Counterfactual and optimal transport II

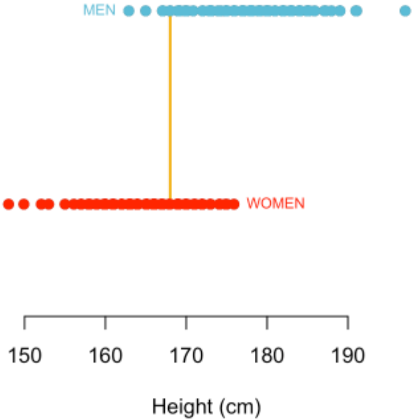


Counterfactual and optimal transport III

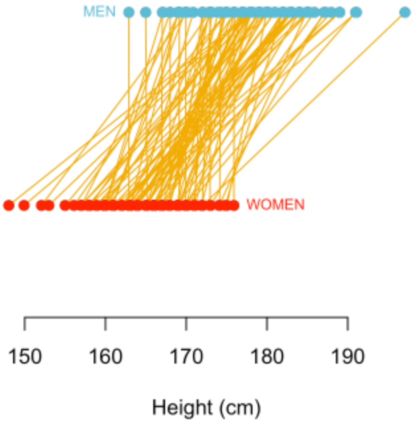
- consider a permutation of all treated individuals ($t_i = 1$),
- for treated individual i ($t_i = 1$), match that individual to someone in the non-treated group ($t_j = 0$), $j_i^\star = \underset{j:t_j=0}{\text{argmin}}\{d(\boldsymbol{x}_i, \boldsymbol{x}_j)\}$,
- then we remove the untreated observation from the database, and iterate (so as to match all treated individuals with an untreated person

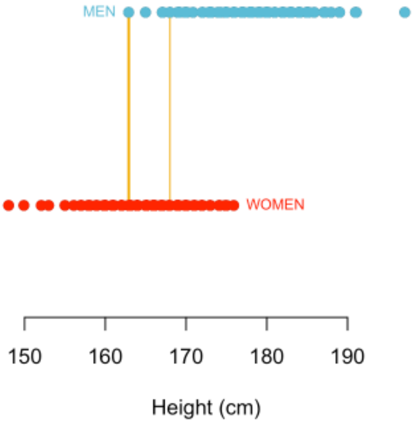
Counterfactual and optimal transport IV



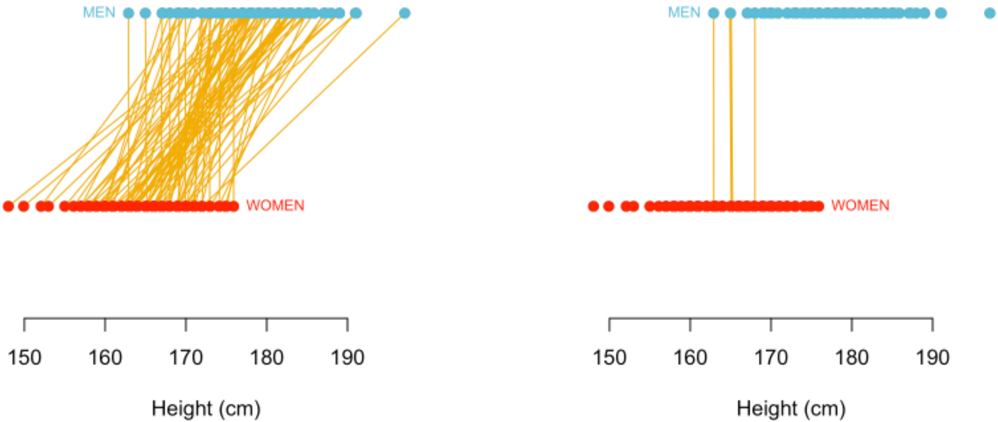


Counterfactual and optimal transport V

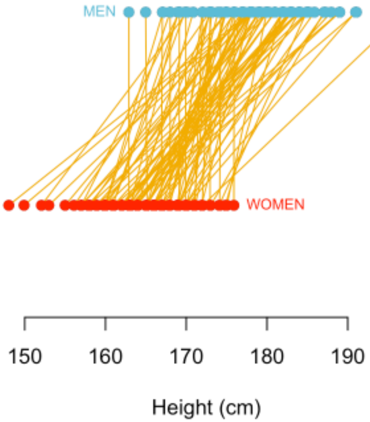


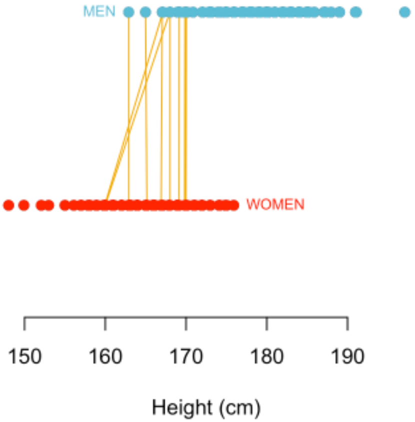


Counterfactual and optimal transport VI

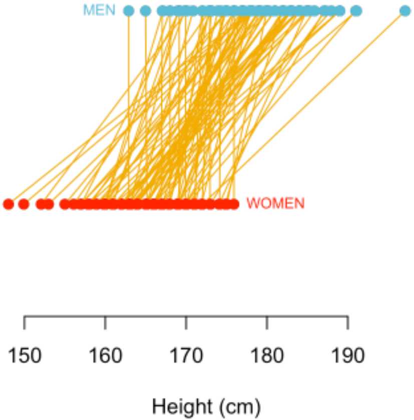


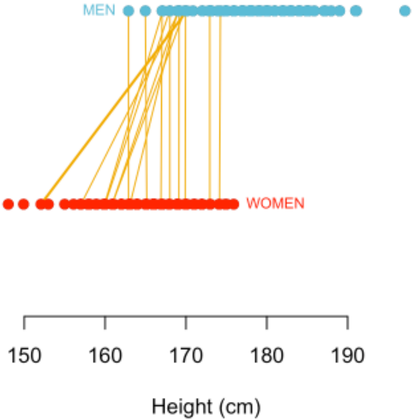
Counterfactual and optimal transport VII



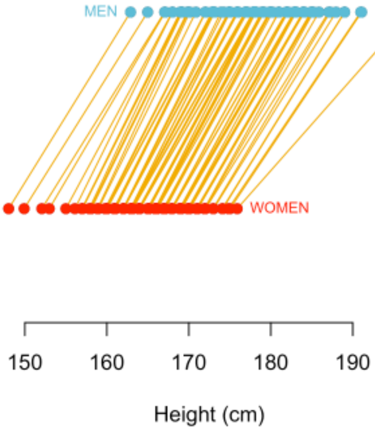


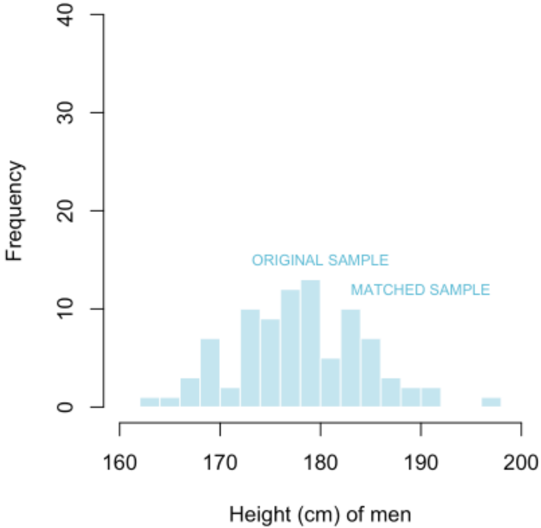
Counterfactual and optimal transport VIII



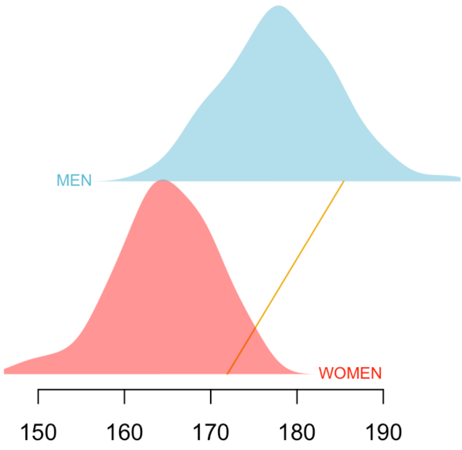


Counterfactual and optimal transport IX





Counterfactual and optimal transport X



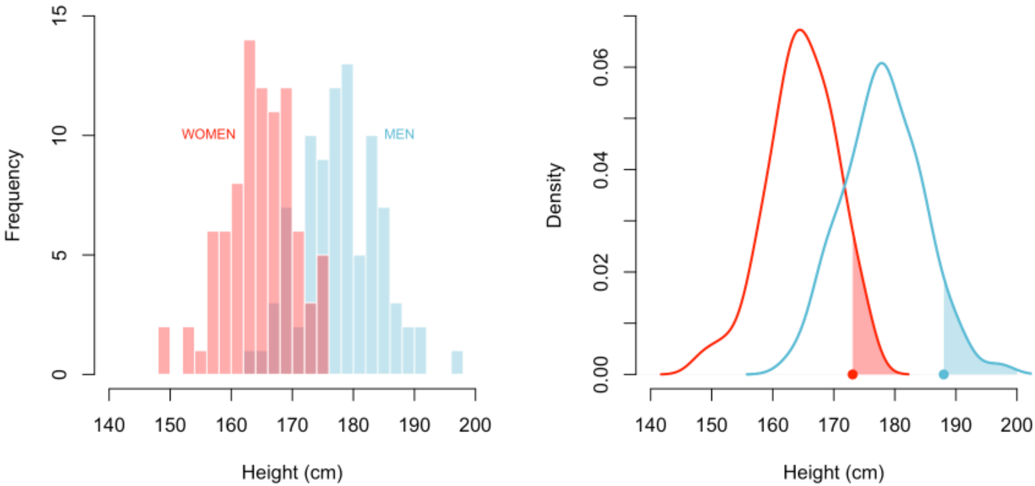
See Monge (1781)'s problem (déblais/remblais - excavation/infill)
We want to go from $\mathbb{P}_0$ to $\mathbb{P}_1$ (distributions on $\mathbb{R}^k$). Given $T : \mathbb{R}^k \to \mathbb{R}^k$, define the "*push-forward*" measure

$$\mathbb{P}_1(A) = T_{\#}\mathbb{P}_{\nvdash}(A) = \mathbb{P}\big(T^{-1}(A)\big),\ \forall A \subset \mathbb{R}^k.$$

An optimal transport $T^\star$ (in Brenier's sense) from $\mathbb{P}_0$ towards $\mathbb{P}_1$ will be solution of

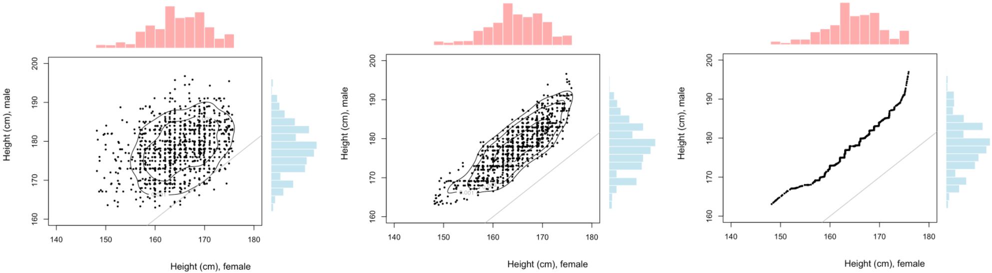
$$T^\star \in \underset{T:T_{\#}\mathbb{P}=\mathbb{Q}}{\text{arginf}} \int_{\mathbb{R}^k} \|\boldsymbol{x} - T(\boldsymbol{x})\|^2 d\mathbb{P}(\boldsymbol{x}).$$

Counterfactual and optimal transport XI



Counterfactual and optimal transport XII

One can prove (see) that $T^\star = \nabla\psi$ where ψ is a convex function. If $k=1$, T is an increasing function, i.e. if $F_0(x)=\mathbb{P}_0[X\leq x]$ and $F_1(x)=\mathbb{P}_1[X\leq x]$, then $T^\star(x)=F_1^{-1}\circ F_0(x)$ satisfies $T^\star_{\#}\mathbb{P}_0=\mathbb{P}_1$ (since $F_1(x)=F_0(T^{\star-1}(x))$ and $T^\star$ is optimal.



Counterfactual and optimal transport XIII

That is an optimal coupling problem, that we can write

$$\min_{\text{copula } C} \int_{\mathbb{R}^2} c(x,y)dF(x,y), \text{ where } F(x,y) = C(F_0(x), F_1(y))$$

As Gordaliza et al. (2019) and Black et al. (2020), a classifier m obtained from score s ($m(\boldsymbol{x},p) = \mathbf{1}(s(\boldsymbol{x},p) > \text{seuil})$) is unfair if $m(\boldsymbol{x},0) \neq m(T^\star(\boldsymbol{x}),1)$. Black et al. (2020) define the FlipSet as

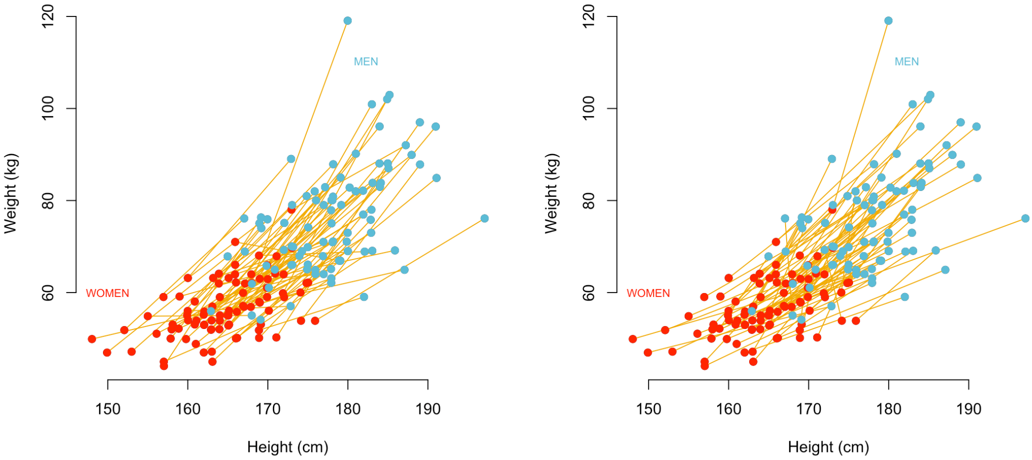
$$\mathcal{X}_F(m,T^\star) = \big\{\boldsymbol{x} \in \mathcal{X} : m(\boldsymbol{x},0) \neq m(T^\star(\boldsymbol{x}),1)\big\}.$$

$$\begin{cases} \mathcal{X}_F^+(m,T^\star) = \big\{\boldsymbol{x} \in \mathcal{X} : m(\boldsymbol{x},0) > m(T^\star(\boldsymbol{x}),1)\big\}. \\ \mathcal{X}_F^-(m,T^\star) = \big\{\boldsymbol{x} \in \mathcal{X} : m(\boldsymbol{x},0) < m(T^\star(\boldsymbol{x}),1)\big\}. \end{cases}$$

and if $T^\star_\# \mathbb{P}_0 = \mathbb{P}_1$, the average causal effect is

$$ACE = \mathbb{E}[Y^\star_{T\leftarrow 1}] - \mathbb{E}[Y^\star_{T\leftarrow 0}] = \mathbb{P}_0[\mathcal{X}_F^-(m,T^\star)] - \mathbb{P}_0[\mathcal{X}_F^+(m,T^\star)]$$

Counterfactual and optimal transport XIV



Adversarial Approach I

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

$$HGR(U,V) = \max\left\{\text{corr}[f(U),g(V)]\right\} = \max_{f\in\mathcal{S}_U, g\in\mathcal{S}_V}\left\{\mathbb{E}[f(U)g(V)]\right\}$$

where $\mathcal{S}_U = \{f:\mathcal{U}\to\mathbb{R} : \mathbb{E}[f(U)] = 0 \text{ and } \mathbb{E}[f(U)^2] = 1\}$ and similarly $\mathcal{S}_V$.
One can also consider a conditional version,

$$HGR(U,V|Z) = \max_{f\in\mathcal{S}_{U|Z}, g\in\mathcal{S}_{V|Z}}\left\{\mathbb{E}[f(U)g(V)|Z]\right\}$$

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

$$\begin{cases} \textbf{Demographic Parity}: & \widehat{Y} \perp\!\!\!\perp P \quad \text{i.e. } HGR(\widehat{Y},P) = 0 \\ \textbf{Equalized Odds}: & \widehat{Y} \perp\!\!\!\perp P|Y \quad \text{i.e. } HGR(\widehat{Y},P|Y) = 0 \end{cases}$$

Adversarial Approach II

HGR can be difficult to estimate, but one can use some Neural Net,

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

In a classical ML or econometric pricing model, solve

$$\underset{\boldsymbol{\theta}}{\text{argmin}}\left\{\mathcal{L}\big(\widehat{y},y\big)\right\},\text{ where }\mathcal{L}\big(\widehat{y},y\big)=\sum_{i=1}^{n}\ell\big(\widehat{y}_i,y_i\big)\text{ and }\widehat{y}=h_{\boldsymbol{\theta}}(x)$$

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

$$\underset{\boldsymbol{\theta}}{\text{argmin}}\left\{\mathcal{L}\big(h_{\boldsymbol{\theta}}(\boldsymbol{x}),y\big)+\lambda\mathcal{P}(h_{\boldsymbol{\theta}})\right\}$$

Adversarial Approach III

Inspired by Goodfellow et al. (2018), to avoid un-fairness: penalize according to $HGR(\widehat{y}, p)$ (for demographic parity)

$$\underset{\boldsymbol{\theta},\boldsymbol{\omega}_f,\boldsymbol{\omega}_g}{\text{argmin}} \left\{\mathcal{L}\big(h_{\boldsymbol{\theta}}(x), y\big) + \lambda \text{HGR}_{\boldsymbol{\omega}_f,\boldsymbol{\omega}_g}(\widehat{y}, p)\right\}$$

i.e.

$$\underset{\boldsymbol{\theta}}{\text{argmin}}\left\{\max_{\boldsymbol{\omega}_f,\boldsymbol{\omega}_g}\left\{\mathcal{L}(h_{\boldsymbol{\theta}}(\boldsymbol{X}), Y) + \lambda \mathbb{E}_{(\boldsymbol{X},S)\sim\mathcal{D}}(\widehat{f}_{\boldsymbol{\omega}_f}(h_{\boldsymbol{\theta}}(\boldsymbol{X}))\widehat{g}_{\boldsymbol{\omega}_g}(P)\right\}\right\}$$

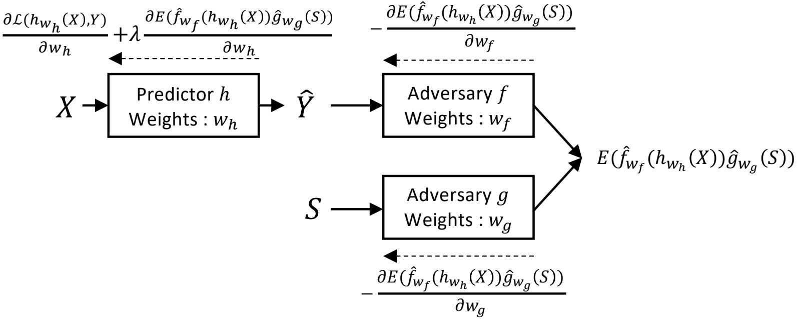
or $HGR(\widehat{y}, p|y)$ (for equalized odds), i.e. when $y \in \{0, 1\}$

$$\begin{aligned}\underset{\boldsymbol{\theta}}{\text{argmin}}\Big\{\max_{\boldsymbol{\omega}_{f0},\boldsymbol{\omega}_{g0},\boldsymbol{\omega}_{f1},\boldsymbol{\omega}_{g1}}\Big\{\mathcal{L}(h_{\boldsymbol{\theta}}(\boldsymbol{X}), Y) \quad &+ \quad \lambda_0 \mathbb{E}_{(\boldsymbol{X},P)\sim\mathcal{D}_0}(\widehat{f}_{\boldsymbol{\omega}_{f_0}}(h_{\boldsymbol{\theta}}(\boldsymbol{X}))\widehat{g}_{\boldsymbol{\omega}_{g_0}}(P)) \\ &+ \quad \lambda_1 \mathbb{E}_{(\boldsymbol{X},P)\sim\mathcal{D}_1}(\widehat{f}_{\boldsymbol{\omega}_{f_1}}(h_{\boldsymbol{\theta}}(\boldsymbol{X}))\widehat{g}_{\boldsymbol{\omega}_{g_1}}(P))\Big\}\Big\}\end{aligned}$$

Adversarial Approach IV

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

$$\underset{\boldsymbol{\theta}}{\text{argmin}}\left\{\max_{\boldsymbol{\omega}_{f0},\boldsymbol{\omega}_{g0},\boldsymbol{\omega}_{fk},\boldsymbol{\omega}_{gk}}\left\{\mathcal{L}(h_{\boldsymbol{\theta}}(\boldsymbol{X}),Y)+\sum_{y\in\Omega_y}\lambda_y\mathbb{E}_{(\boldsymbol{X},P)\sim\mathcal{D}_y}\left(\widehat{f}_{\boldsymbol{\omega}_{fy}}(h_{\boldsymbol{\theta}}(\boldsymbol{X}))\widehat{g}_{\boldsymbol{\omega}_{gy}}(P)\right)\right\}\right\}$$



Dealing with high dimension I

- geographic / spatial information, $\boldsymbol{X}_g$
- car type / make / model, $\boldsymbol{X}_g$
- other classical ratemaking variables, $\boldsymbol{X}_p$ (non protected)

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

Dealing with high dimension II

Principal Component Analysis (PCA)

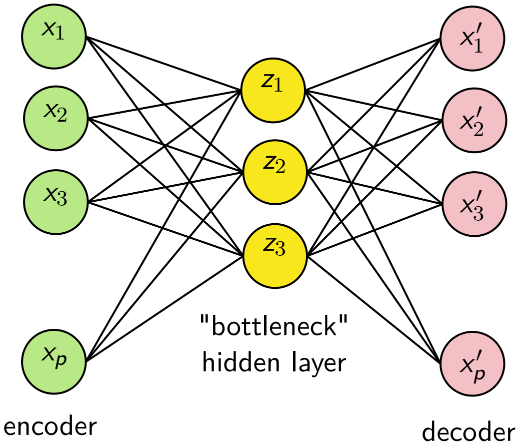
$$\min_{\boldsymbol{P}\in\Pi}\left\{\|\boldsymbol{X}-\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{X}\|_F^2\right\} \text{ s.t. } \text{rank}(\boldsymbol{P})=k$$

where Π is the set of projection matrices.

$$\boxed{\boldsymbol{X}\in\mathcal{X}} \xrightarrow{\boldsymbol{P}} \boxed{\boldsymbol{Z}\in\mathcal{Z}} \xrightarrow{\boldsymbol{P}^\top} \boxed{\boldsymbol{X}'\in\mathcal{X}}$$

$$\min\|\boldsymbol{X}-\boldsymbol{X}'\|^2=\min\|\boldsymbol{X}-\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{X}\|^2$$

$$=\min\sum_{i=1}^n\left(\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{x}_i-\boldsymbol{x}_i\right)^\top\left(\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{x}_i-\boldsymbol{x}_i\right)$$



Dealing with high dimension III

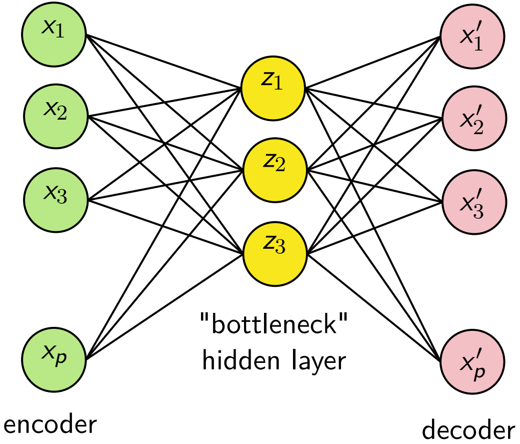
Autoencoder

$$\min_{\psi}\left\{\|\boldsymbol{X}-\psi\circ\varphi\boldsymbol{X}\|_F^2\right\}$$

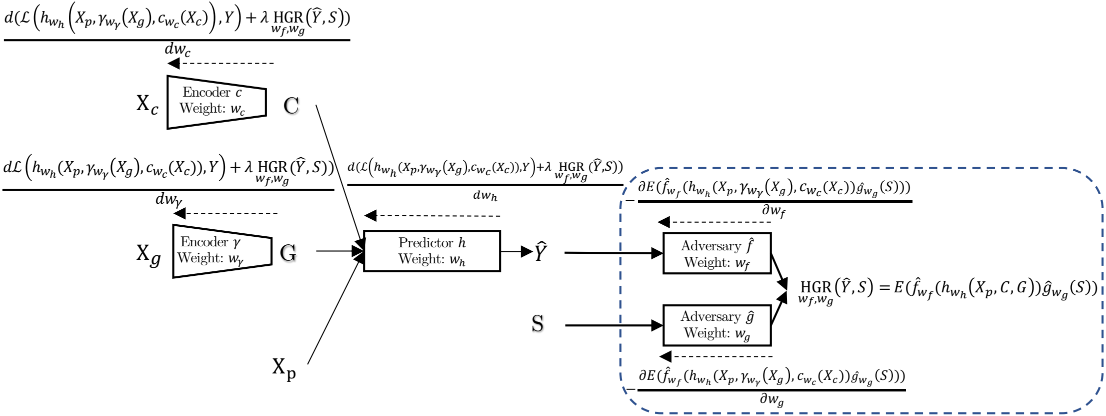
$$\boxed{\boldsymbol{X}\in\mathcal{X}} \xrightarrow{\varphi} \boxed{\boldsymbol{Z}\in\mathcal{Z}} \xrightarrow{\psi} \boxed{\boldsymbol{X}'\in\mathcal{X}}$$

$$\min\|\boldsymbol{X}-\boldsymbol{X}'\|^2=\min\|\boldsymbol{X}-\psi\circ\varphi(\boldsymbol{X})\|^2$$

$$\min\sum_{i=1}^{n}\big(\psi\circ\varphi(\boldsymbol{x}_i)-\boldsymbol{x}_i\big)^\top\big(\psi\circ\varphi(\boldsymbol{x}_i)-\boldsymbol{x}_i\big)$$

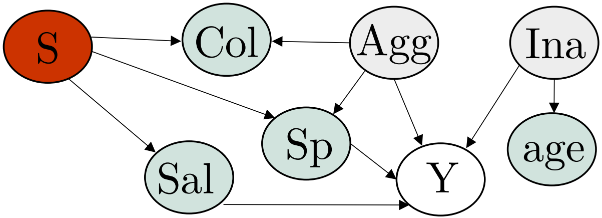


Dealing with high dimension IV

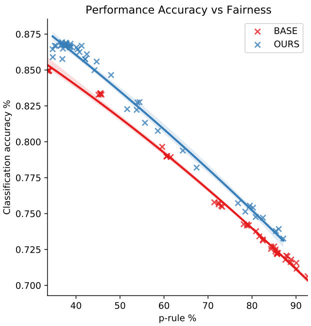


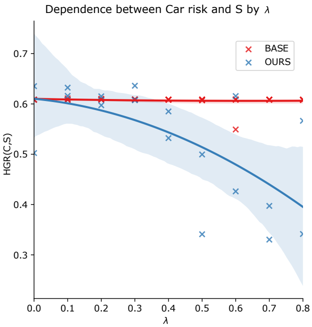
Application on synthetic data I

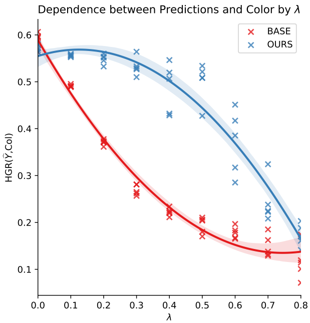
- S: sensitive / protected (gender)
- Col: color of the car
- Sp: maximum speed of the car
- Sal: average salary of the policyholders area
- Age: age of the driver
- Ina: inattention
- Agg: aggressivity
- Y: total cost



Application on synthetic data II



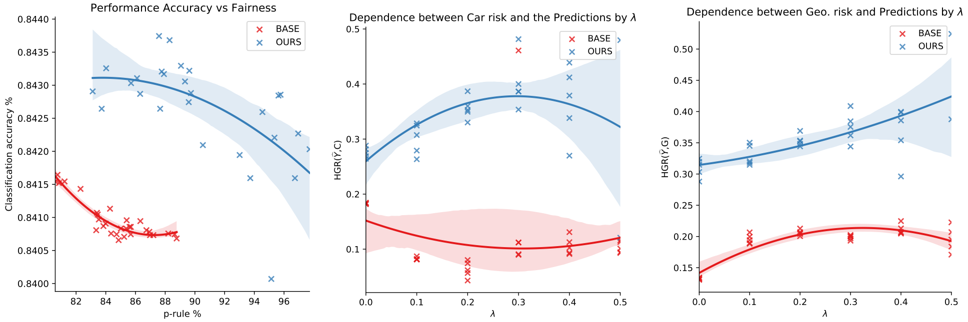




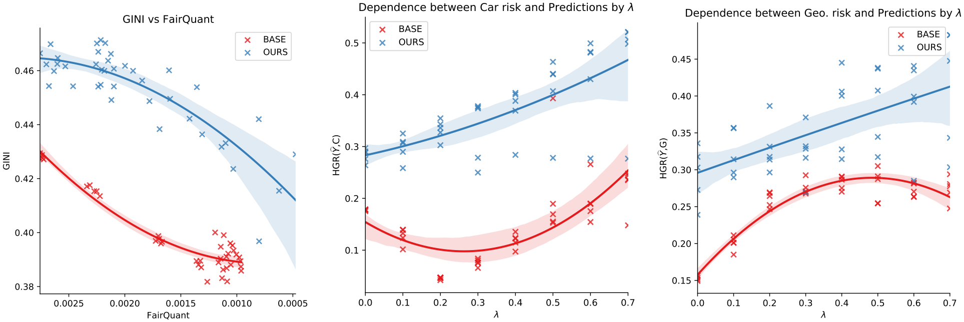
- λ: fairness penalty
- p-rule: $\min\left\{\dfrac{\mathbb{P}(\widehat{Y}=1|P=1)}{\mathbb{P}(\widehat{Y}=1|P=0)}, \dfrac{\mathbb{P}(\widehat{Y}=1|P=0)}{\mathbb{P}(\widehat{Y}=1|P=1)}\right\}$

Application on real data (pricing game 2015) I

$y \in \{0,1\}$ (claim occurrence)



$y \in \{0, 1, 2+\}$ (claim frequency)



see Grari et al. (2022) for more examples (including the case where $y \in \mathbb{R}^+$)

References I

Agarwal, S. (2021). Trade-offs between fairness and interpretability in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*.

Alipourfard, N., Fennell, P. G., and Lerman, K. (2018). Can you trust the trend? discovering simpson's paradoxes in social data. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 19–27.

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

Barry, L. and Charpentier, A. (2020). Personalization as a promise: Can big data change the practice of insurance? *Big Data & Society*, 7(1):2053951720935143.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.

Becker, G. S. (2005). Is ethnic and other profiling discrimination? *The Becker-Posner Blog*, 01-23-2005.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. *arXiv*, 1706.02409.

Bertillon, A. and Chervin, A. (1909). *Anthropologie métrique: conseils pratiques aux missionnaires scientifiques sur la manière de mesurer, de photographier et de décrire des sujets vivants et des pièces anatomiques*. Imprimerie nationale.

References II

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404.

Bigot, R. and Charpentier, A. (2019). Repenser la responsabilité, et la causalité. *Risques*, 120:123–128.

Bigot, R. and Charpentier, A. (2020). Quelle responsabilité pour les algorithmes? *Risques*, 121.

Black, E., Yeom, S., and Fredrikson, M. (2020). Fliptest: Fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 111121.

Boczar, D., Avila, F. R., Carter, R. E., Moore, P. A., Giardi, D., Guliyeva, G., Bruce, C. J., McLeod, C. J., and Forte, A. J. (2021). Using facial recognition tools for health assessment. *Plastic Surgical Nursing*, 41(2):112–116.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2019). Inaccurate statistical discrimination: An identification problem. Technical report, National Bureau of Economic Research.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

References III

Burgdorf, M. P. and Burgdorf Jr, R. (1974). A history of unequal treatment: The qualifications of handicapped persons as a suspect class under the equal protection clause. *Santa Clara Lawyer*, 15:855.

Charpentier, A. (2022a). *Assurance: biais, discrimination et équité*. Institut Louis Bachelier.

Charpentier, A. (2022b). *Insurance: biases, discrimination and fairness*. Institut Louis Bachelier.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Cohen, J. E. (1986). An uncertainty principle in demography and the unisex issue. *The American Statistician*, 40(1):32–39.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *arXiv*, 1701.08230.

Duivesteijn, W. and Feelders, A. (2008). Nearest neighbour classification with monotonicity constraints. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 301–316. Springer.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

References IV

Edgeworth, F. Y. (1922). Equal pay to men and women for equal work. *The Economic Journal*, 32(128):431–457.

Gebelein, H. (1941). Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., and Fei-Fei, L. (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113.

Gomery, S., Renault, O., and John, N. (2011). Gender neutral. *Canadian Underwriter*.

Goodfellow, I., McDaniel, P., and Papernot, N. (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66.

Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365. PMLR.

Grari, V., Charpentier, A., Lamprier, S., and Detyniecki, M. (2022). A fair pricing model via adversarial learning. *ArXiv*, 2202.12008.

Hand, D. J. (2020). *Dark Data: Why What You Dont Know Matters*. Princeton University Press.

References V

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Hernán, M. A. and Robins, J. M. (2010). Causal inference.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520524.

Hudson, R. (2007). Mortality projections and unisex pricing of annuities in the uk. *Journal of Financial Regulation and Compliance*.

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

Ilic, L., Sawada, M., and Zarzelli, A. (2019). Deep mapping gentrification in a large canadian city using deep learning and google street view. *PloS one*, 14(3):e0212814.

Imai, K. (2018). *Quantitative social science: an introduction*. Princeton University Press.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119.

References VI

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

Keilman, N. (2003). Types of models for projecting mortality. *Perspectives on mortality forecasting*, 1:19–27.

Kita, K. and Kidziński, Ł. (2019). Google street view image of a house predicts car accident risk of its resident. *arXiv*, 1904.05270.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

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 Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. NIPS.

Law, S., Paige, B., and Russell, C. (2019). Take a look around: Using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology*, 10(5).

Lombroso, C. (1876). *L'uomo delinquente*. Hoepli.

References VII

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

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review*, 62(4):659–661.

Rényi, A. (1959). On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Schmeiser, H., Störmer, T., and Wagner, J. (2014). Unisex insurance pricing: consumers perception and market implications. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 39(2):322–350.

Seresinhe, C. I., Preis, T., and Moat, H. S. (2017). Using deep learning to quantify the beauty of outdoor places. *Royal Society open science*, 4(7):170170.

Shi, P. and Shi, K. (2021). Nonlife insurance risk classification using categorical embedding. *SSRN*, 3777526.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472.

References VIII

TenBroek, J. (1966). The right to live in the world: The disabled in the law of torts. *Calif. L. Rev.*, 54:841.

Thiery, Y. and Van Schoubroeck, C. (2006). Fairness and equality in insurance classification. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 31(2):190–211.