

# IA et éthique en Assurance: discriminations, biais et équité

**Arthur Charpentier<sup>1</sup> & Laurence Barry<sup>2</sup>**

<sup>1</sup> Université du Québec à Montréal    <sup>2</sup> Chaire Pari

Journée IA et éthique, 2022

# Agenda & Mots Clés

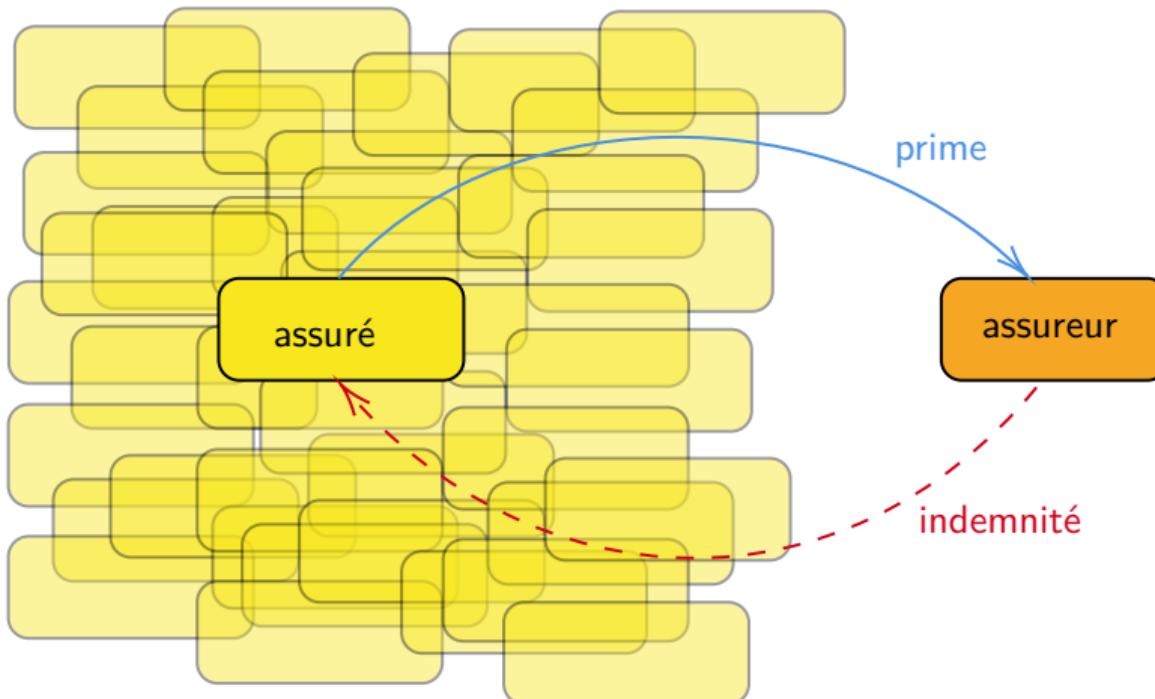
“Technology is neither good nor bad; nor is it neutral” , Kranzberg (1986)

- ▶ Assurance, mutualisation, solidarité vs. individualisation, hétérogénéité
- ▶ Discrimination, *actuarial fairness*, aspects légaux, discrimination par proxy
- ▶ Biais observation vs. expérience, biais de sélection, biais de variable omise
- ▶ Équité,  $\hat{Y} \perp\!\!\!\perp P$ ,  $\hat{Y} \perp\!\!\!\perp P | Y$  ou  $Y \perp\!\!\!\perp P | \hat{Y}$ , et équité individuelle (contrefactuels)
- ▶ Explicabilité et interprétabilité

Cf Charpentier (2022) pour plus de détails

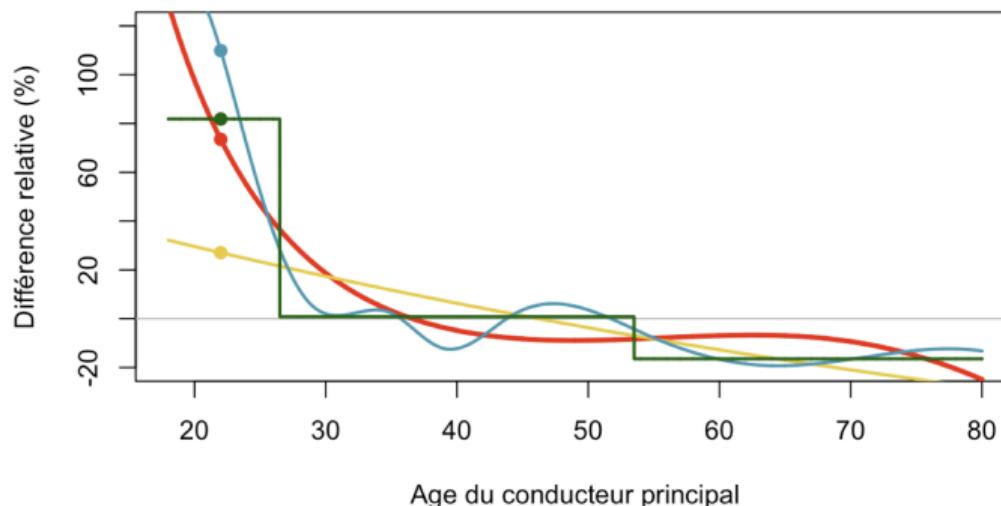
# Assurance, mutualisation des risques & solidarité

- ▶ Insurance is the contribution of the many to the misfortune of the few



# Risques hétérogènes I

- ▶ Fréquence de sinistre, en fonction de l'âge du conducteur (Charpentier (2014))



- ▶ “actuaries smoothed because smoothing was a ‘mathematical and ethical’ good”, Bouk (2015)
- ▶ interprétabilité et explicabilité, “les jeunes conducteurs ont plus de chance d'avoir un accident”

## Risques hétérogènes II

- ▶ Assurance vie, tables de mortalité en fonction de l'âge et du genre
- ▶ Tables hommes / femmes, 1720 ([Struyck \(1912\)](#), page 231)

hommes			femmes		
0	1000	29.0%	45	371	16.6%
5	710	5.6%	50	313	19.2%
10	670	4.2%	55	253	22.9%
15	642	5.5%	60	195	27.2%
20	607	6.6%	65	142	31.7%
25	567	7.9%	70	97	37.1%
30	522	9.2%	75	61	45.9%
35	474	10.5%	80	33	51.5%
40	424	12.5%	85	16	

## Risques hétérogènes III

- ▶ Assurance vie, tables de mortalité en fonction de l'âge et du genre
- ▶ Tables plus récentes (TV, TD et INED)

TD 73-77		TV 73-77		TD 88-90		TV 88-90		INED (H)		INED (F)	
0	100000	0	100000	0	100000	0	100000	0	100000	0	100000
10	97961	10	98447	10	98835	10	99129	10	99486	10	99578
20	97105	20	98055	20	98277	20	98869	20	99281	20	99471
30	95559	30	97439	30	96759	30	98371	30	98656	30	99247
40	93516	40	96419	40	94746	40	97534	40	97661	40	98810
50	88380	50	94056	50	90778	50	95752	50	95497	50	97645
60	77772	60	89106	60	81884	60	92050	60	90104	60	94777
70	57981	70	78659	70	65649	70	84440	70	78947	70	89145
80	28364	80	52974	80	39041	80	65043	80	59879	80	77161
90	4986	90	14743	90	9389	90	24739	90	25123	90	44236
100	103	100	531	100	263	100	1479	100	1412	100	4874
110	0	110	0	110	0	110	2				

## Risques hétérogènes IV

- ▶ Assurance vie, espérance de vie résiduelle (en années) en fonction de l'âge, du genre et sur le statut de fumeur (ou pas), (données Benjamin and Michaelson (1988) 1970-1975, US)
- ▶ Hoffman (1931), Johnston (1945) “*it is clear that smoking is an important cause of mortality*”, Miller and Gerstein (1983)

hommes		femmes		
	non-fumeur	fumeur	non-fumeur	fumeur
25	48.4	42.8	25	52.8
35	38.7	33.3	35	43.0
45	29.2	24.2	45	33.5
55	20.3	16.5	55	24.5
65	12.8	10.4	65	16.2

## Risques hétérogènes V

- ▶ Assurance vie, espérance de vie résiduelle (en années) en fonction de l'âge, du genre et sur le poids (BMI) (données Steensma et al. (2013) US)  
regular [ $18.5; 25\text{kg}/\text{m}^2$ ], overweighted [ $25; 30\text{kg}/\text{m}^2$ ], obesity I [ $30; 35\text{kg}/\text{m}^2$ ],  
obesity II [ $35, 100\text{kg}/\text{m}^2$ ])
- ▶ Crossley (2005), Czerniawski (2007) ou Kelly and Markowitz (2009)

		hommes				femmes			
		regular	over.	obesity I	obesity II	regular	over.	obesity I	obesity II
20		57.2	61.0	59.1	53.5	20	62.8	66.5	64.6
30		47.6	51.4	49.4	44.1	30	53.0	56.7	54.8
40		38.1	41.7	39.9	34.7	40	43.3	46.9	45.0
50		28.9	32.4	30.6	25.8	50	33.8	37.3	35.5
60		20.4	23.6	21.9	17.6	60	24.9	28.1	26.4
70		13.2	15.8	14.4	10.9	70	16.8	19.7	18.2

## Risques hétérogènes VI

- ▶ handicap et tests génétiques
- ▶ “*the insurance industry has generally regarded handicapped persons as undesirable risks*” Baker and Karol (1977)
- ▶ “*the denial of insurance coverage to an individual whose (noninherited) cancer had been long cured would not constitute genetic discrimination, while the denial of insurance to that individual's relatives because of the (erroneous) belief that that type of cancer is heritable would be genetic discrimination*” Natowicz et al. (1992)
- ▶ Schatz (1986), Clifford and Iculano (1987) (HIV), Jacobs and Sommers (2015) (inférence à partir de prescriptions de médicaments)

# Assurance et “individualisation” de la prime I

- ▶ “*il convient en effet de distinguer deux choses lorsque l'on parle d'assurance. La première, l'opération d'assurance, relève de la technique et a une dimension collective, la seconde, le contrat d'assurance, relève du droit et a une dimension individuelle*”, Bigot and Cayol (2020) (aussi Thiery and Van Schoubroeck (2006), Lehtonen and Liukko (2015))
- ▶ **Approche individualiste**
  - ▶ L'approche individualiste de l'égalité analyse les droits fondamentaux, tels que le droit à l'égalité de traitement, en termes d'individus.
  - ▶ Un individu ne peut être traité différemment en raison de son appartenance à tel ou tel groupe, en particulier à un groupe auquel il n'a pas choisi d'appartenir.
- ▶ **Approche de groupe**
  - ▶ La tradition de l'assurance, quant à elle, analyse les risques, les primes et les barèmes de prestations en termes de groupes.
  - ▶ Contrairement à l'approche "individualiste", les systèmes de classification des assurances reposent sur l'hypothèse que les individus répondent aux caractéristiques moyennes (stéréotypées) d'un groupe auquel ils appartiennent.

## Assurance et “individualisation” de la prime II

- ▶ “at the core of insurance business lies discrimination between risky and non-risky insureds”, Avraham (2017)
- ▶ segmentation parfaite avec un facteur de risque latent observable  $\Theta$

	assuré	assureur
perte	$\mathbb{E}[Y \Theta]$	$Y - \mathbb{E}[Y \Theta]$
perte moyenne	$\mathbb{E}[Y]$	0
variance	$\text{Var}[\mathbb{E}[Y \Theta]]$	$\text{Var}[Y - \mathbb{E}[Y \Theta]]$

$$\text{Var}[Y] = \underbrace{\mathbb{E}[\text{Var}[Y|\Theta]]}_{\rightarrow \text{assureur}} + \underbrace{\text{Var}[\mathbb{E}[Y|\Theta]]}_{\rightarrow \text{assuré}}.$$

## Assurance et “individualisation” de la prime III

- segmentation (statistique) avec des caractéristiques observables  $\mathbf{X} = (X_1, \dots, X_k)$   
“categorization based on immutable characteristics”, Crocker and Snow (2013)

	assuré	assureur
perte	$\mathbb{E}[Y \mathbf{X}]$	$Y - \mathbb{E}[Y \mathbf{X}]$
perte moyenne	$\mathbb{E}[Y]$	0
variance	$\text{Var}[\mathbb{E}[Y \mathbf{X}]]$	$\mathbb{E}[\text{Var}[Y \mathbf{X}]]$

$$\begin{aligned}\mathbb{E}[\text{Var}[Y|\mathbf{X}]] &= \mathbb{E}\left[\mathbb{E}\left[\text{Var}[Y|\Theta]|\mathbf{X}\right]\right] + \mathbb{E}\left[\text{Var}\left[\mathbb{E}[Y|\Theta]|\mathbf{X}\right]\right] \\ &= \underbrace{\mathbb{E}\left[\text{Var}[Y|\Theta]\right]}_{\text{segmentation parfaite}} + \underbrace{\mathbb{E}\left\{\text{Var}\left[\mathbb{E}[Y|\Theta]|\mathbf{X}\right]\right\}}_{\text{misfit}}.\end{aligned}$$

- “kanssolidariteit” vs “subsidierende solidariteit”, De Pril and Dhaene (1996)

# Assurance(s) & solidarité

- ▶ assurance santé
- ▶ assurance collective
- ▶ catastrophes naturelles

*“La Nation proclame la solidarité et l'égalité de tous les Français devant les charges qui résultent des calamités nationales”, Constitution du 27 octobre 1946*

*“la solidarité en assurance, c'est décider de ne pas segmenter le marché du risque correspondant sur une base des caractéristiques observables des risques des individus”, Gollier (2002).*

- ▶ assurance non-vie

*“Tout ce qui n'est pas défendu par la Loi ne peut être empêché”, Déclaration des Droits de l'Homme et du Citoyen, 1789, art. 5*

## Réglementation et aspects légaux I

- ▶ “accéder à l’assurance s’entend non seulement de la possibilité même de souscrire un contrat en vue d’une couverture, mais peut-être, également, à un coût économique raisonnable, non prohibitif partant non dissuasif” ([Noguéro \(2010\)](#))
- ▶ le **sexé** ou le **genre** (art. A. 111-6 du Code des assurances, Commission européenne (Arr. 18 déc. 2012, NOR : EFIT1238658A, relatif à l’égalité entre les hommes et les femmes en assurance, JO 20 déc., mod. par Arr. 3 févr. 2014, NOR : EFIT1400411A, JO 11 févr.)
- ▶ distinction fondée sur l’**âge** (C. pén., art. 225-1 et 225-2), (“*belonging to a particular race or sex is akin to joining one specific ‘club at the moment of conception, whereas age...*”, [Macnicol \(2006\)](#))
- ▶ la **situation de famille** ou sur l’ **orientation sexuelle** (C. pén., art. 225-1 et 225-2)
- ▶ en raison du **lieu de résidence** d’une personne constitue une discrimination au sens pénal (C. pén., art. 225-1)
- ▶ “*Nul ne peut faire l’objet de discriminations en raison de ses caractéristiques génétiques*” (C. C., art. 16-13)

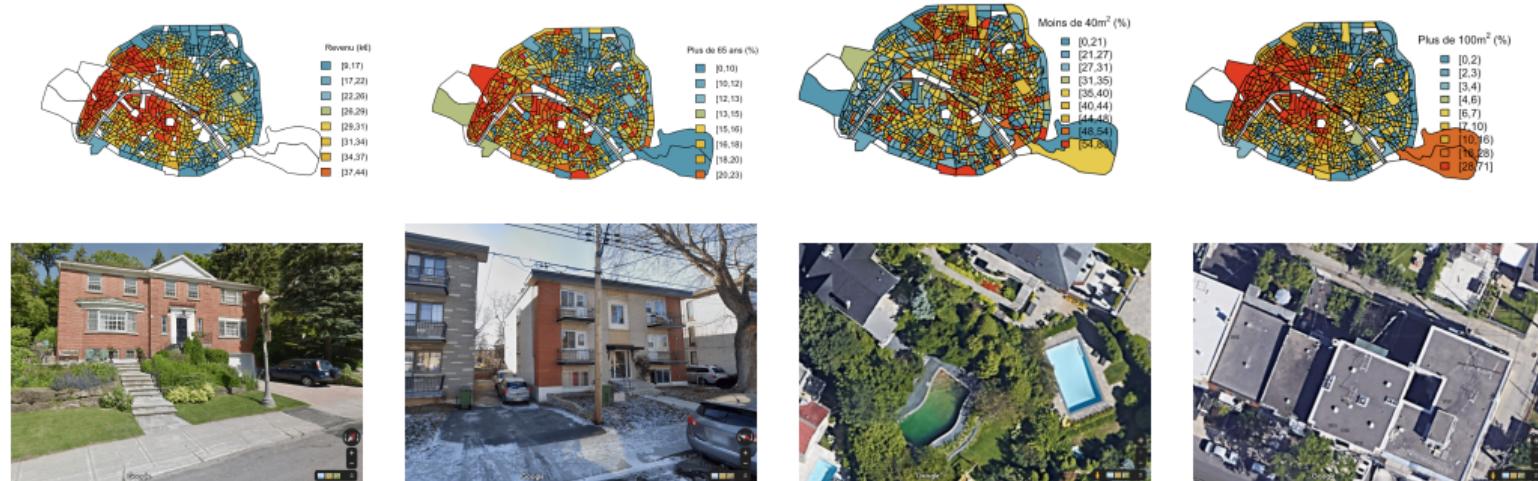
## Réglementation et aspects légaux II

	CA	HI	GA	NC	NY	MA	PA	FL	TX
Genre	X	X	•	X	•	X	X	•	•
Âge	X	X	•	X*	•	X	•	•	•
Expérience de conduite	•	X	•	•	•	•	•	•	•
Antécédents de crédit	X	X	•	•	•	X	•*	•	•
Éducation	X	X	X	X	X	X	•	•	•
Profession	X	X	X	•	X	X	•	•	•
Situation d'emploi	X	X	X	•	X	X	•	•	•
Situation de famille	•	X	•	•	•	X	•	•	•
Situation résidentielle	X	X	•	•	•	X	•	•	•
Adresse/code postal	•	•	•	•	•	•	•	•	•
Antécédents d'assurance	•	•	•	•	•	•	•	•	•

# Discrimination par proxies (?) I

## ► localisation (adresse de l'assuré)

Jean et al. (2016), Seresinhe et al. (2017), Gebru et al. (2017), Law et al. (2019), Illic et al. (2019), Kita and Kidziński (2019)



## Discrimination par proxies (?) II

- ▶ visages, récemment, certains assureurs ont envisagé l'idée d'utiliser la reconnaissance faciale pour prédire certaines maladies, Shikhare (2021)



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

- ▶ cf "phrénologie" Lombroso (1876) et Bertillon and Chervin (1909)
- ▶ cf "ugly laws" TenBroek (1966) et Burgdorf and Burgdorf Jr (1974)

# Discrimination par proxies (?) III

- ▶ credit scoring, très important en Amérique du Nord

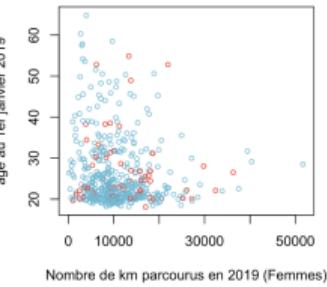
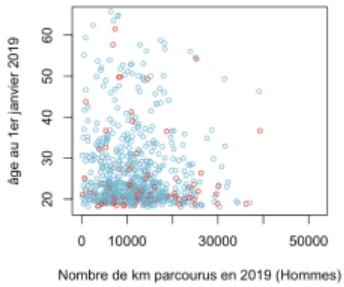
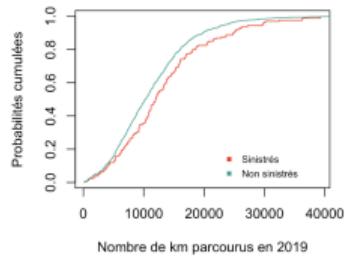
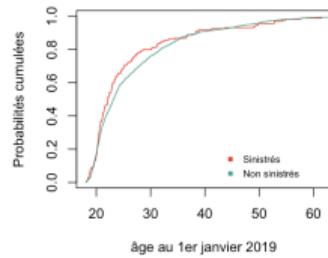
Kabler (2004), Arya et al. (2013), Miller et al. (2003) Bartik and Nelson (2016), O'neil (2016), Lauer (2017), Morris et al. (2017), Kiviat (2019)



source <https://www.incharge.org/debt-relief/credit-counseling/>

# Discrimination par proxies (?) IV

► **télématique**, données "comportementales"



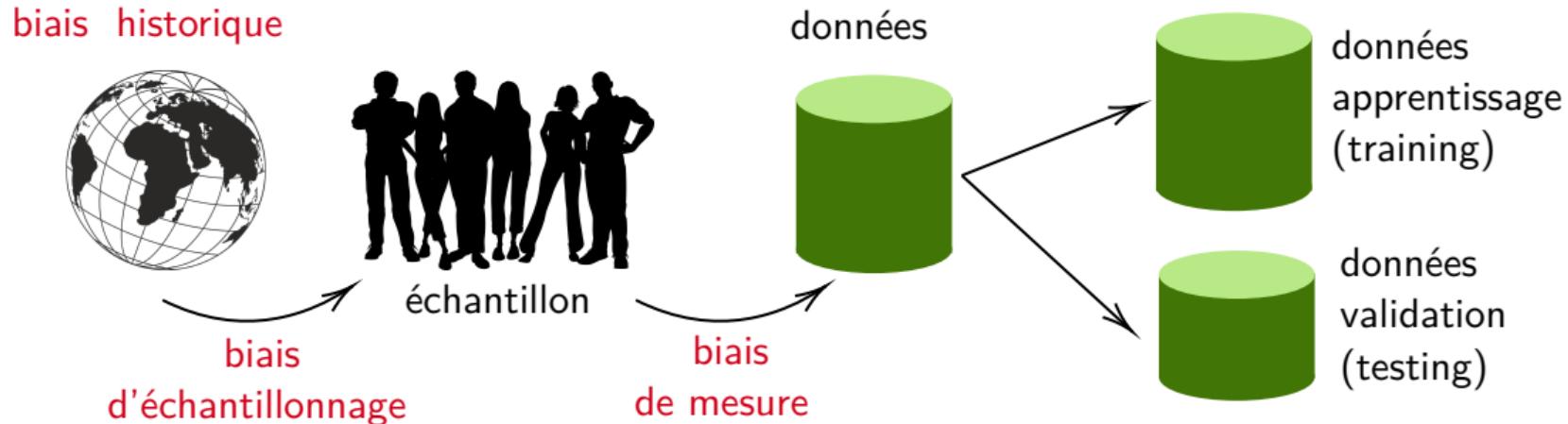
E.g. distance parcourue et genre de l'assuré **Verbelen et al. (2018)**

## Pour aller plus loin sur la discrimination

- ▶ Notion de **variable sensible** dans le RGPD
- ▶ Forte composante culturelle
- ▶ En grande dimension (beaucoup de variables explicatives  $x$ ), il y a de fortes chances d'avoir des variables (très) corrélées à une variable sensible

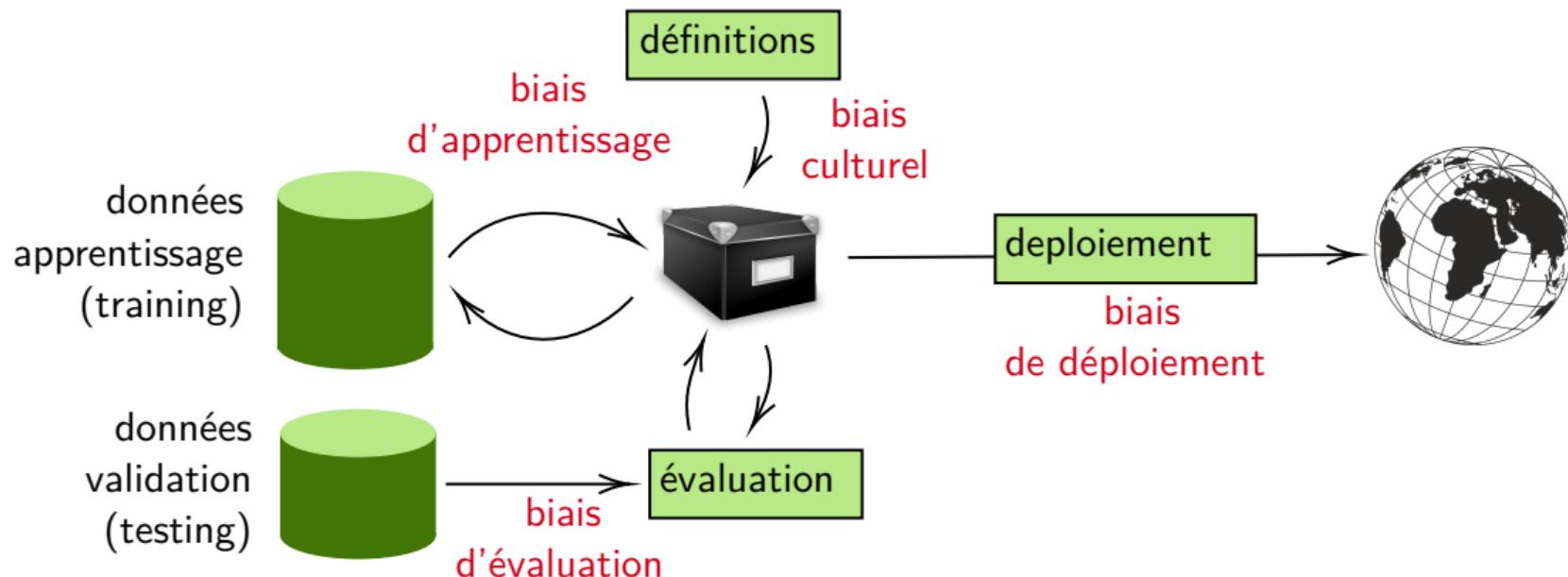
	Total	Men	Women	Proportions
Total	5233/12763 ~ 41%	3714/8442 ~ <b>44%</b>	1512/4321 ~ 35%	66%-34%
Top 6	1745/4526 ~ 39%	1198/2691 ~ <b>45%</b>	557/1835 ~ 30%	59%-41%
A	597/933 ~ 64%	512/825 ~ 62%	89/108 ~ <b>82%</b>	88%-12%
B	369/585 ~ 63%	353/560 ~ 63%	17/ 25 ~ <b>68%</b>	96%- 4%
C	321/918 ~ 35%	120/325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

# Biais dans la génération des données



(inspiré de Suresh and Guttag (2019)).

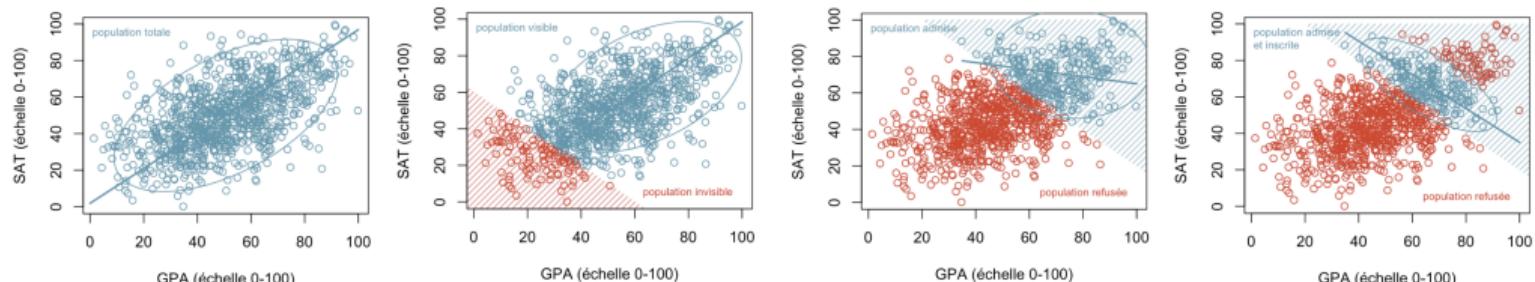
# Biais dans la construction d'un modèle



(inspiré de Suresh and Guttag (2019)).

# Pour aller plus loin sur les biais

- ▶ notions de “*dark data*” de Hand (2020)
- ▶ **paradoxe** de Simpson, **paradoxe\*** écologique (omission de variables importantes)
  - ex: nombre d'accidents piéton-véhicule et vitesse moyenne, Davis (2004)
  - ex: comparaisons de taux de mortalité (local vs global) Cohen (1986)
- ▶ loi de Goodhart (et biais de rétroaction)
  - “*lorsqu'une mesure devient un objectif, elle cesse d'être une bonne mesure*”
  - ex: covid-related data, Giles (2020)



\* les paradoxes mettent à mal l'interprétabilité et l'explicabilité...

# Mesurer et quantifier l'équité I

Notations:

$$\begin{cases} y \in \{0, 1\} & \text{variable d'intérêt} \\ p \in \{0, 1\} & \text{variable protégée (sensible)} \\ \mathbf{x} \in \mathbb{R}^d & \text{variables 'explicatives'} \\ s \in [0, 1] & \text{score, classiquement } s = s(\mathbf{x}, p) \\ \hat{y} \in \{0, 1\} & \text{prédicteur (classifieur), classiquement } \hat{y} = \mathbf{1}(s > t) \end{cases}$$

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

L'attribut protégé  $p$  n'est pas explicitement utilisé dans la fonction de décision  $\hat{y}$ .

## Mesurer et quantifier l'équité II

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

Une fonction de décision  $\hat{y}$  satisfait à la parité démographique si  $\hat{Y} \perp\!\!\!\perp P$ , soit

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

En pratique, on compare  $DI(\hat{y}, p)$  (**disparate impact**) à 80%

$$DI(\hat{Y}, P) = \frac{\mathbb{P}[\hat{Y} = 1 | P = 0]}{\mathbb{P}[\hat{Y} = 1 | P = 1]} \stackrel{?}{\leq} 80\%$$

cf Feldman et al. (2015), Mercat-Bruns (2016) ou Biddle (2017) utilisé par le State of California Fair Employment Practice Commission (FEPC) depuis 1971  
(voir aussi Besse et al. (2021))

# Mesurer et quantifier l'équité III

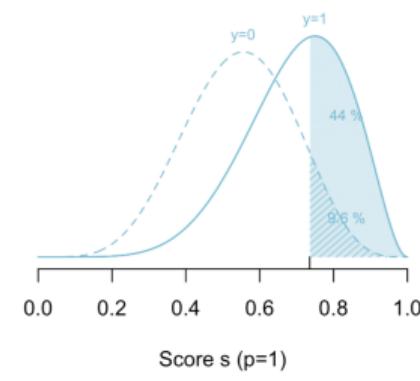
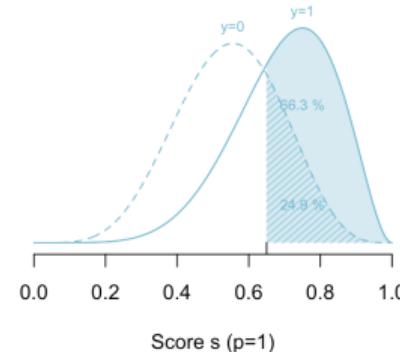
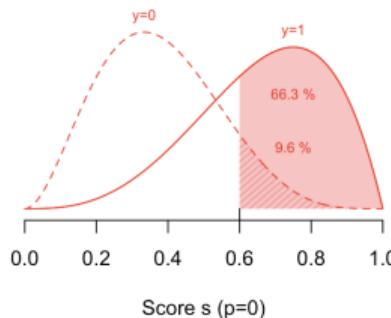
**Equal Opportunity**, Hardt et al. (2016)

Parité des vrais positifs

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

ou parité des faux positifs

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



## Mesurer et quantifier l'équité IV

### Equalized Odds, Hardt et al. (2016)

La parité des faux positifs et des vrais positifs est appelé égalité des chances,

$$\begin{cases} \mathbb{P}[\hat{Y} = 1 | P = 0, Y = 1] = \mathbb{P}[\hat{Y} = 1 | P = 1, Y = 1] \\ \mathbb{P}[\hat{Y} = 1 | P = 0, Y = 0] = \mathbb{P}[\hat{Y} = 1 | P = 1, Y = 0] \end{cases}$$

ou

$$\mathbb{P}[\hat{Y} = 1 | P = 0, Y = y] = \mathbb{P}[\hat{Y} = 1 | P = 1, Y = y], \forall y \in \{0, 1\}$$

autrement dit,  $\hat{Y} \perp\!\!\!\perp P$  conditionnellement à  $Y$ .

Etc... il existe de nombreux concepts, souvent incompatibles entre eux.

# Mesurer et quantifier l'équité V

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

## Mesurer et quantifier l'équité VI

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

$$D(\hat{y}_i, \hat{y}_j) \text{ ou } D(s_i, s_j) \leq d(\mathbf{x}_i, \mathbf{x}_j), \forall i, j = 1, \dots, n.$$

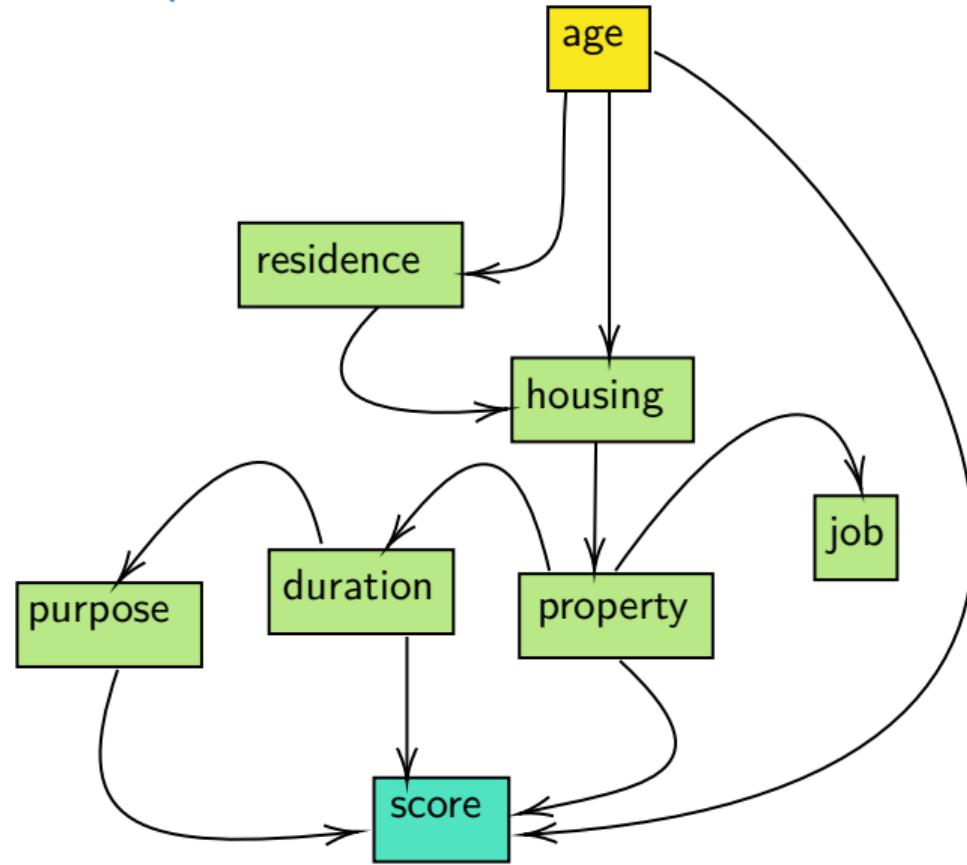
Cf intervention formelle " $\mathbf{X}$  est fixé à  $\mathbf{x}$ ", ce que Pearl (1998) note " $do(\mathbf{X} = \mathbf{x})$ " (ou simplement  $do(\mathbf{x})$ ), (historiquement, de Wright (1921), Neyman et al. (1923) ou Rubin (1974) Holland (1986))

**Counterfactual fairness**, Kusner et al. (2017) Si la prédiction dans le monde réel est la même que celle dans le monde contrefactuel où l'individu aurait appartenu à un groupe démographique différent, on a une équité contrefactuelle, autrement dit

$$\mathbb{P}[Y_{P \leftarrow p}^* = y | \mathbf{X} = \mathbf{x}, P = p] = \mathbb{P}[Y_{P \leftarrow p'}^* = y | \mathbf{X} = \mathbf{x}, P = p], \forall p', \mathbf{x}, y.$$

## Pour aller plus loin sur la mesure de l'équité

- ▶  $P$  doit être observée
- ▶ Recherche de contrefactuel
- ▶ Importance des graphs causaux



## Wrap-up

- ▶ “les États membres peuvent décider (...) d'autoriser des différences proportionnelles dans les primes et les prestations des particuliers lorsque l'utilisation du sexe est un facteur déterminant dans l'évaluation du risque, sur la base de données actuarielles et statistiques pertinentes et précises” modèles **causaux** ?
- ▶ “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”, **Glenn (2000)**  
“virtually every aspect of the insurance industry is predicated on stories first and then numbers”, **Glenn (2003)**  
importance sw l'aspect **narratif** (“*all models are wrong but some models are useful*”, **Box et al. (2011)**).

## References I

- Agarwal, S. (2021). Trade-offs between fairness and interpretability in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*.
- Arya, S., Eckel, C., and Wichman, C. (2013). Anatomy of the credit score. *Journal of Economic Behavior & Organization*, 95:175–185.
- Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.
- Baker, L. D. and Karol, C. (1977). Employee insurance benefit plans and discrimination on the basis of handicap. *DePaul L. Rev.*, 27:1013.
- Barry, L. and Charpentier, A. (2020). Personalization as a promise: Can big data change the practice of insurance? *Big Data & Society*, 7(1):2053951720935143.
- Bartik, A. and Nelson, S. (2016). Deleting a signal: Evidence from pre-employment credit checks. *SSRN*, 2759560.
- Benjamin, B. and Michaelson, R. (1988). Mortality differences between smokers and non-smokers. *Journal of the Institute of Actuaries*, 115(3):519525.
- 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.

## References II

- 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.
- Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.-M., and Risser, L. (2021). A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, 0(0):1–11.
- Biddle, D. (2017). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Routledge.
- Bigot, R. and Cayol, A. (2020). *Le droit des assurances en tableaux*. Ellipses.
- 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.
- Bouk, D. (2015). *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual*. The University of Chicago Press.
- Box, G. E., Luceño, A., and del Carmen Paniagua-Quinones, M. (2011). *Statistical control by monitoring and adjustment*, volume 700. John Wiley & Sons.
- 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.

## References III

- Charpentier, A. (2014). *Computational Actuarial Science*. The R series. CRC Press.
- Charpentier, A. (2019a). Les classes de risques vont-elles plus loin que les stéréotypes? *L'Actuairel*, 32.
- Charpentier, A. (2019b). Les modèles prédictifs peuvent-ils être loyaux et justes. *Risques*, 113.
- Charpentier, A. (2021a). Assurance et discrimination, quel rôle pour les actuaires ? *Risques*, 127.
- Charpentier, A. (2021b). Le mythe de l'interprétabilité et de l'explicabilité des modèles. *Risques*, 128.
- Charpentier, A. (2022). *Assurance: biais, discrimination et équité*. Institut Louis Bachelier.
- Charpentier, A. and Barry, L. (2019). Concilier risques collectifs et décisions individuelles. *Risques*, 123.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.
- Clifford, K. A. and Iculano, R. P. (1987). AIDS and insurance: the rationale for AIDS-related testing. *Harvard law review*, 100(7):1806–1825.
- 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.

## References IV

- Crocker, K. J. and Snow, A. (2013). The theory of risk classification. In Loubergé, H. and Dionne, G., editors, *Handbook of insurance*, pages 281–313. Springer.
- Crossley, M. (2005). Discrimination against the unhealthy in health insurance. *University of Kansas Law Review*, 54:73.
- Czerniawski, A. M. (2007). From average to ideal: The evolution of the height and weight table in the united states, 1836-1943. *Social Science History*, 31(2):273296.
- Davis, G. A. (2004). Possible aggregation biases in road safety research and a mechanism approach to accident modeling. *Accident Analysis & Prevention*, 36(6):1119–1127.
- De Pril, N. and Dhaene, J. (1996). Segmentering in verzekeringen. *DTEW Research Report 9648*, pages 1–56.
- 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.
- Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. 1412.3756(arXiv).

## References V

- 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.
- Giles, C. (2020). Goodharts law comes back to haunt the uks covid strategy. *Financial Times*, 14-5.
- 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.
- Gollier, C. (2002). La solidarite sous langle economique. *Revue Générale du Droit des Assurances*, pages 824–830.
- Hand, D. J. (2020). *Dark Data: Why What You Dont Know Matters*. Princeton University Press.
- Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.
- Hoffman, F. L. (1931). Cancer and smoking habits. *Annals of surgery*, 93(1):50.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

## References VI

- 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.
- Jacobs, D. B. and Sommers, B. D. (2015). Using drugs to discriminateadverse selection in the insurance marketplace. *New England Journal of Medicine*.
- 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.
- Johnston, L. (1945). Effects of tobacco smoking on health. *British Medical Journal*, 2(4411):98.
- Kabler, B. (2004). Insurance-based credit scores: Impact on minority and low income populations in missouri. *State of Missouri Departement of Insurance*.
- 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.
- Kelly, I. R. and Markowitz, S. (2009). Incentives in obesity and health insurance. *Inquiry*, 46(4):418–432.
- Kita, K. and Kidziński, Ł. (2019). Google street view image of a house predicts car accident risk of its resident. *arXiv*, 1904.05270.

## References VII

- Kiviat, B. (2019). The moral limits of predictive practices: The case of credit-based insurance scores. *American Sociological Review*, 84(6):1134–1158.
- 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.
- Lauer, J. (2017). *Creditworthy: A History of Consumer Surveillance and Financial Identity in America*. Columbia University Press.
- 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).
- Lehtonen, T.-K. and Liukko, J. (2015). Producing solidarity, inequality and exclusion through insurance. *Res Publica*, 21(2):155–169.
- Lombroso, C. (1876). *L'uomo delinquente*. Hoepli.

## References VIII

- Macnicol, J. (2006). *Age discrimination: An historical and contemporary analysis*. Cambridge University Press.
- Mercat-Bruns, M. (2016). *Discrimination at Work*. University of California Press.
- Miller, G. and Gerstein, D. R. (1983). The life expectancy of nonsmoking men and women. *Public Health Reports*, 98(4):343.
- Miller, M. J., Smith, R. A., and Southwood, K. N. (2003). The relationship of credit-based insurance scores to private passenger automobile insurance loss propensity. *Actuarial Study, Epic Actuaries*.
- Morris, D. S., Schwarcz, D., and Teitelbaum, J. C. (2017). Do credit-based insurance scores proxy for income in predicting auto claim risk? *Journal of Empirical Legal Studies*, 14(2):397–423.
- Natowicz, M. R., Alper, J. K., and Alper, J. S. (1992). Genetic discrimination and the law. *American Journal of Human Genetics*, 50(3):465.
- Neyman, J., Dabrowska, D. M., and Speed, T. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472.
- Noguéro, D. (2010). Sélection des risques. discrimination, assurance et protection des personnes vulnérables. *Revue générale du droit des assurances*, 3:633–663.
- O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

## References IX

- Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.
- Schatz, B. (1986). The aids insurance crisis: Underwriting or overreaching. *Harvard Law Review*, 100:1782.
- 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.
- Shikhare, S. (2021). Next generation ltc - life insurance underwriting using facial score model. In *Insurance Data Science conference*.
- Steensma, C., Loukine, L., Orpana, H., Lo, E., Choi, B., Waters, C., and Martel, S. (2013). Comparing life expectancy and health-adjusted life expectancy by body mass index category in adult canadians: a descriptive study. *Population health metrics*, 11(1):1–12.
- Struyck, N. (1912). *Les oeuvres de Nicolas Struyck (1687-1769): qui se rapportent au calcul des chances, à la statistique général, z la statistique des décès et aux rentes viagères*. Société générale néerlandaise d'assurances sur la vie et de rentes viagères.

## References X

- Suresh, H. and Guttag, J. V. (2019). A framework for understanding sources of harm throughout the machine learning life cycle. *arXiv*, 1901.10002.
- 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.
- Verbelen, R., Antonio, K., and Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(5):1275–1304.
- Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20.