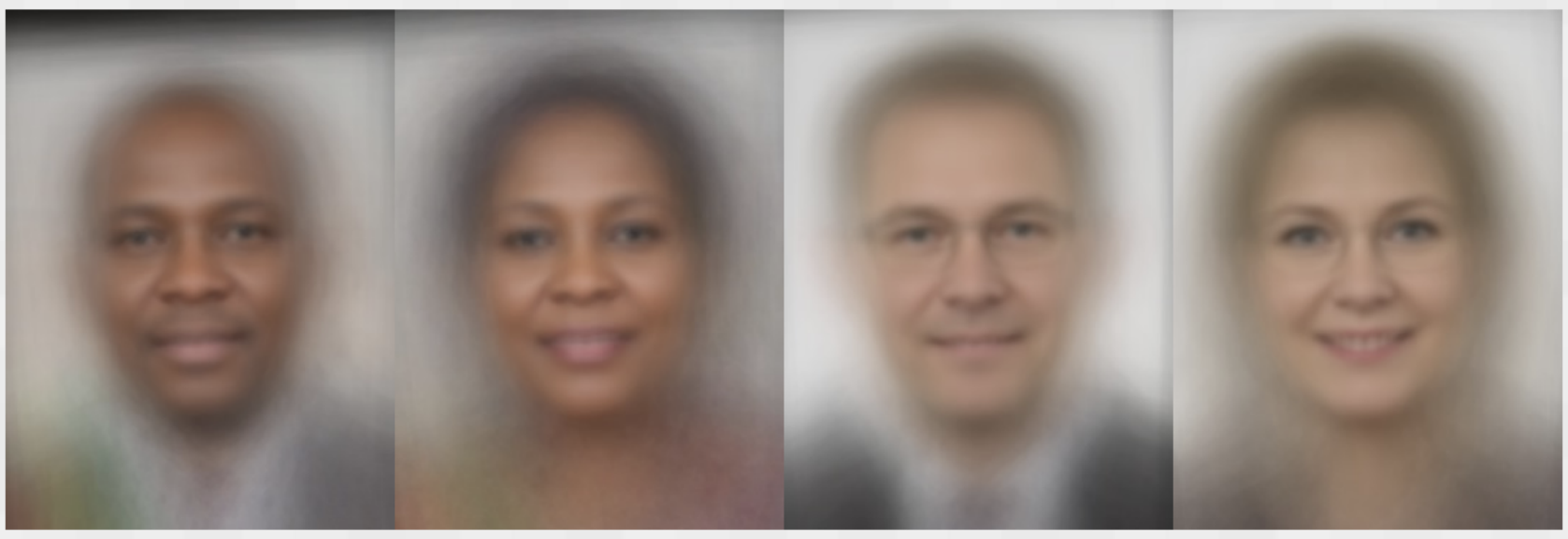
Fairness in Machine Learning

Patrick Loiseau

Inria

Hi! Paris summer school, July 2023

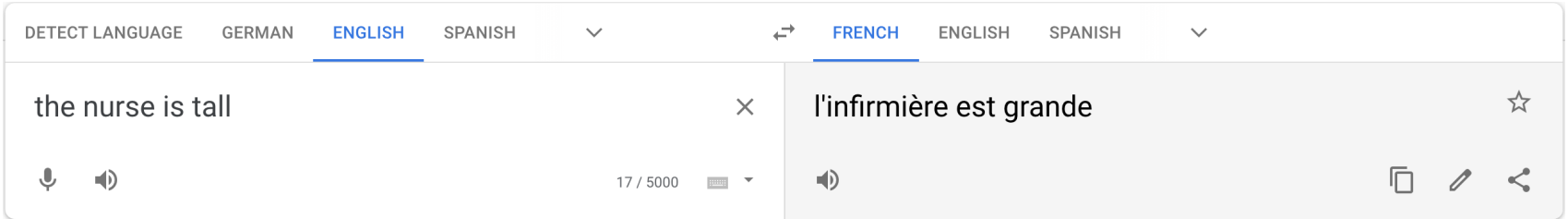
Discrimination in automatic face recognition



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

`http://gendershades.org/`, [Buolamwini & Gebru 2018]

Discrimination in automatic translation systems

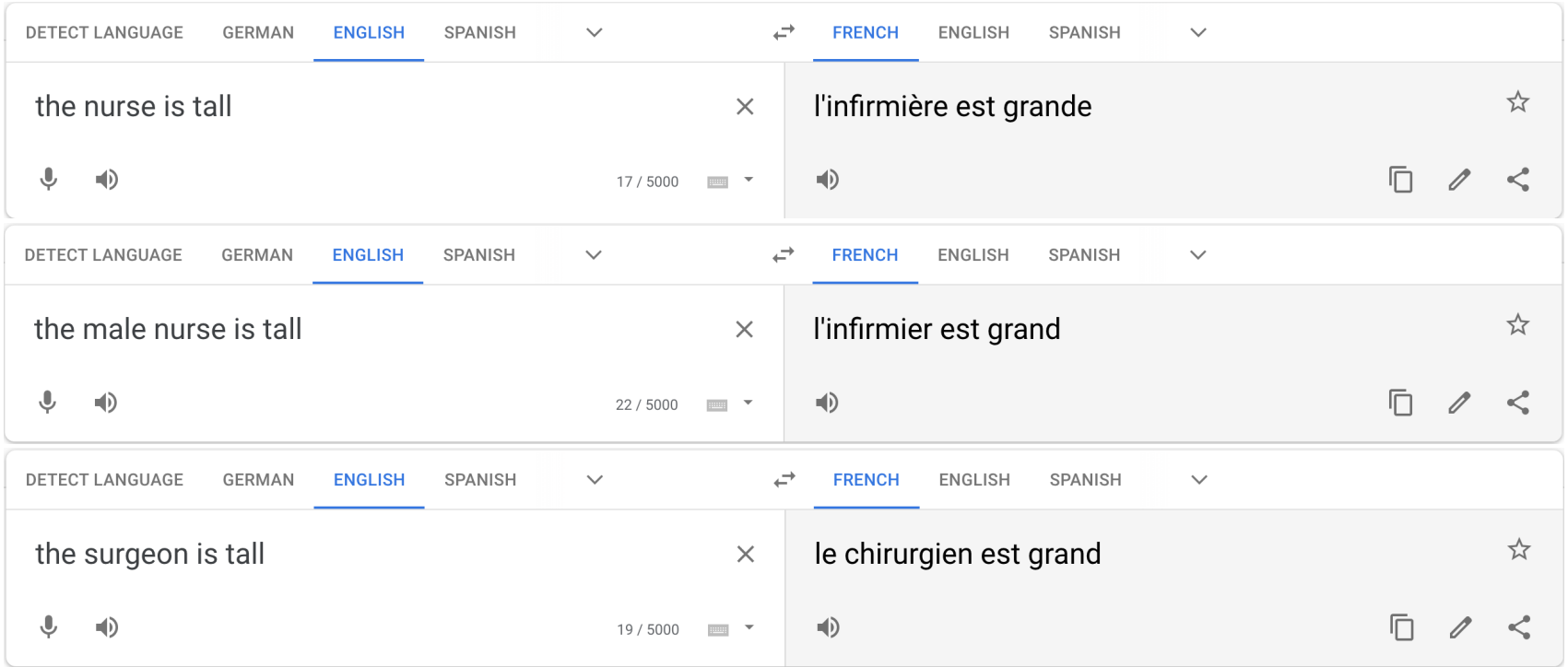


Google translate, Jan 15, 2021

Discrimination in automatic translation systems

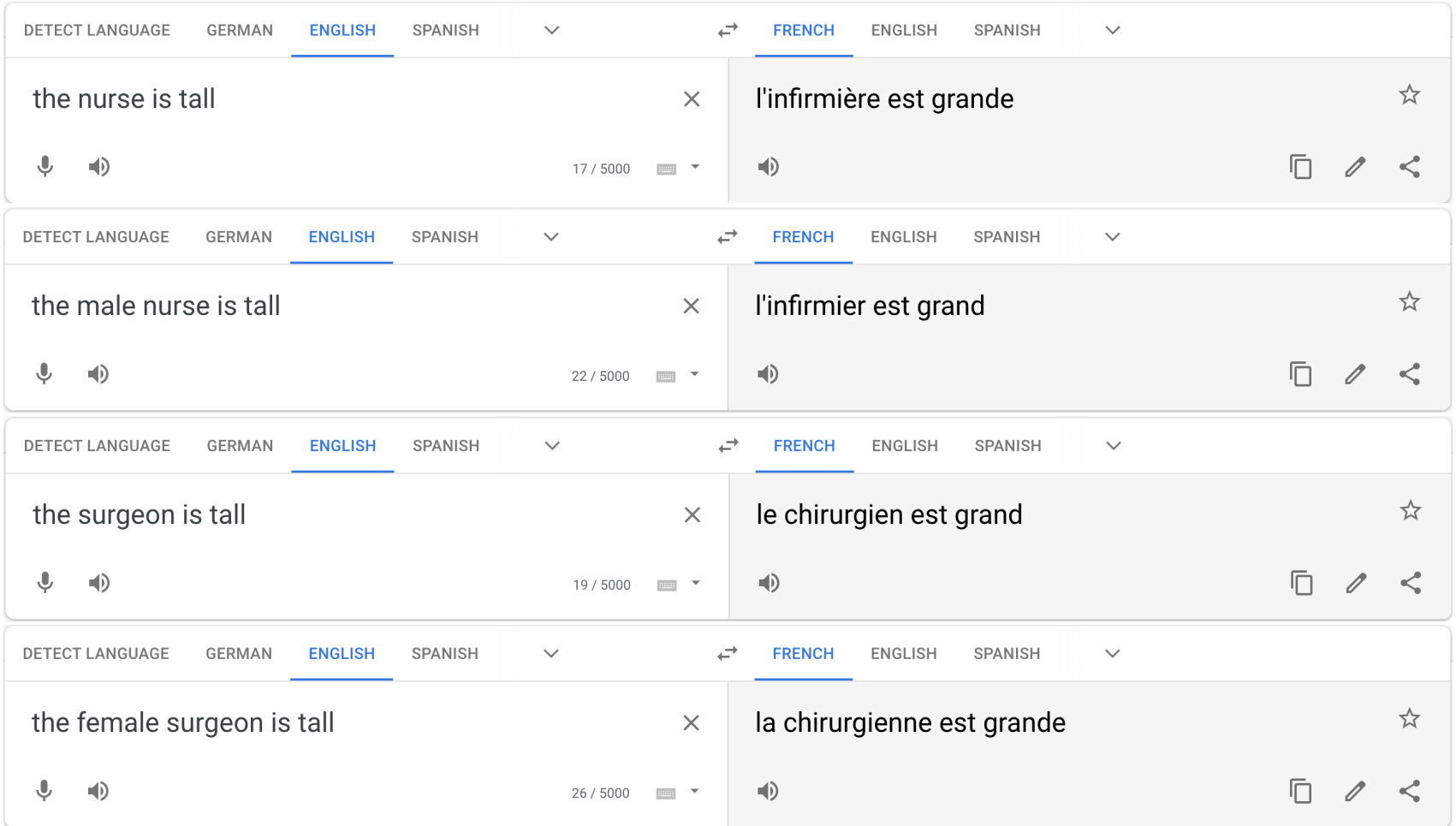
Google translate, Jan 15, 2021

Discrimination in automatic translation systems



Google translate, Jan 15, 2021

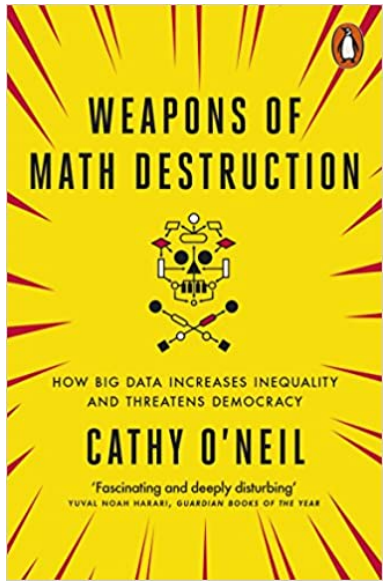
Discrimination in automatic translation systems



Google translate, Jan 15, 2021

AI systems for high-stakes decisions about individuals

- Multitude of domains
  - Hiring
  - Credit approval
  - Insurance
  - Predictive policing
  - Justice
  - Education
  - Pricing
  - etc



- Discrimination: Certain *demographic groups* receive "less favorable outcomes" (disparate impact)
  - Prohibited by law
  - Sensitive attribute: race, age, gender, sexual orientation, disability, religious belief, etc. (full lists varies per country/domain)
  - Ethical concern

Discrimination in ML-based hiring systems

The New York Times

TheUpshot

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

117

By Claire Cain Miller

June 25, 2015

"hiring could become faster and less expensive, and [...] lead recruiters to more highly skilled people [...]. Another potential result: a more diverse workplace. The software relies on data to surface candidates from a wide variety of places and match their skills to the job requirements, free of human biases."

Discrimination in ML-based hiring systems

The New York Times

TheUpshot

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

117

By Claire Cain Miller

June 25, 2015

"hiring could become faster and less expensive, and [...] lead recruiters to more highly skilled people [...]. Another potential result: a more diverse workplace. The software relies on data to surface candidates from a wide variety of places and match their skills to the job requirements, free of human biases."

The New York Times

TheUpshot

HIDDEN BIAS

When Algorithms Discriminate

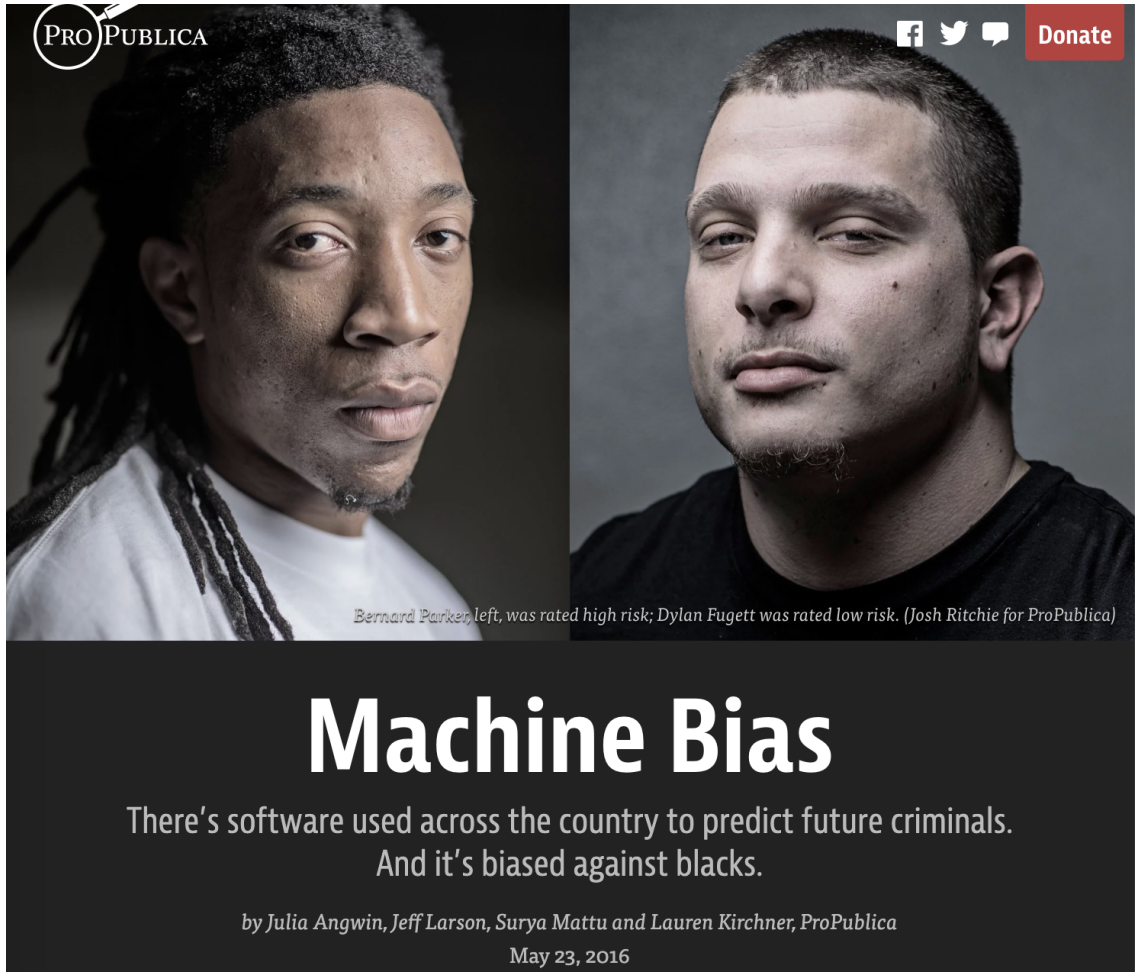
147

By Claire Cain Miller

July 9, 2015

"But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result, [...] algorithms can reinforce human prejudices."

Discrimination in ML-based justice



[Angwin et al., Propublica 2016]

Discrimination in online advertising

- Opportunity ads (jobs, financial services, housing, etc.)
- Law prohibits discrimination at ad stage (i.e., not just final decision)

[1] [Ali et al., 2019]

Discrimination in online advertising

- Opportunity ads (jobs, financial services, housing, etc.)
- Law prohibits discrimination at ad stage (i.e., not just final decision)


DIGITAL

Online Ads for High-Paying Jobs Are Targeting Men More Than Women

New study uncovers gender bias

By Garett Sloane | July 7, 2015 PREMIUM


Facebook, Amazon, and hundreds of companies post targeted job ads that screen out older workers

Facebook users are suing them for age discrimination.

By Alexia Fernández Campbell | @AlexiaCampbell | alexia@vox.com | May 31, 2018, 8:50am EDT

[1] [Ali et al., 2019]

Discrimination in online advertising

- Opportunity ads (jobs, financial services, housing, etc.)
- Law prohibits discrimination at ad stage (i.e., not just final decision)


DIGITAL

Online Ads for High-Paying Jobs Are Targeting Men More Than Women

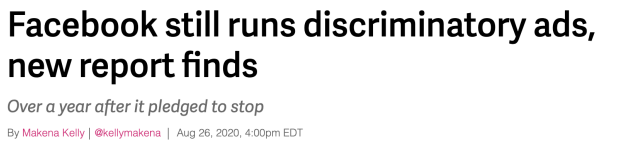
New study uncovers gender bias

By Garett Sloane | July 7, 2015 PREMIUM


Facebook, Amazon, and hundreds of companies post targeted job ads that screen out older workers

Facebook users are suing them for age discrimination.

By Alexia Fernández Campbell | @AlexiaCampbell | alexia@vox.com | May 31, 2018, 8:50am EDT


Facebook still runs discriminatory ads, new report finds

Over a year after it pledged to stop

By Makena Kelly | @kellymakena | Aug 26, 2020, 4:00pm EDT

[1] [Ali et al., 2019]

Discrimination in online advertising

- Opportunity ads (jobs, financial services, housing, etc.)
- Law prohibits discrimination at ad stage (i.e., not just final decision)


DIGITAL

Online Ads for High-Paying Jobs Are Targeting Men More Than Women

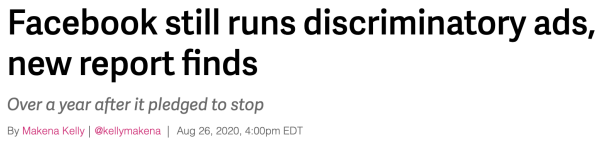
New study uncovers gender bias

By Garett Sloane | July 7, 2015 PREMIUM


Facebook, Amazon, and hundreds of companies post targeted job ads that screen out older workers

Facebook users are suing them for age discrimination.

By Alexia Fernández Campbell | @AlexiaCampbell | alexia@vox.com | May 31, 2018, 8:50am EDT


Facebook still runs discriminatory ads, new report finds

Over a year after it pledged to stop

By Makena Kelly | @kellymakena | Aug 26, 2020, 4:00pm EDT

- Removing sensitive attributes does not work!
  - Proxies, custom audience, lookalike audience
- The AI matching algorithm itself discriminates[1]

[1] [Ali et al., 2019]

Examples of proxies in Facebook ads

- Study based on voter record in the US[2]

Race	Most inclusive	Most exclusive
Asian	US Politics: Liberal (8%, 2.76) Frequent travelers (15%, 2.70) Interest: Vegetarianism (7%, 2.23)	US Politics: Very Conservative (14%, 0.30) African American affinity (17%, 0.41) Interest: Country music (20%, 0.48)
Black	African American affinity (17%, 7.06) US Politics: Very Liberal (12%, 6.44) Interest: Online games (9%, 4.91)	US Politics: Very Conservative (14%, 0.18) US Politics: Conservative (17%, 0.22) Interest: Mountain biking (6%, 0.35)
Indian	Interest: Motorcycles (7%, 2.08) Interest: Online games (9%, 2.04) Interest: Ecotourism (6%, 1.96)	US Politics: Very Conservative (14%, 0.50) Away from hometown (22%, 0.51) Primary OS Mac OS X (7%, 0.56)
White	US Politics: Very Conservative (14%, 5.19) US Politics: Conservative (17%, 3.77) Interest: Hiking (11%, 2.27)	African American affinity (17%, 0.15) US Politics: Very Liberal (12%, 0.16) Interest: Online games (9%, 0.20)

Free-form Attribute	Potential Target (PT)	PT Audience (%)	US Audience (%)
Marie Claire	Female	90%	54%
myGayTrip.com	Man interested in Man	38.6%	0.38%
BlackNews.com	African American affinity	89%	16%
Hoa hoc Tro Magazine	Asian American affinity	95%	3.4%
Nuestro Diario	Hispanic affinity	98%	16%

[2] [Speicher et al., 2018]

Biases extend to look-alike audiences

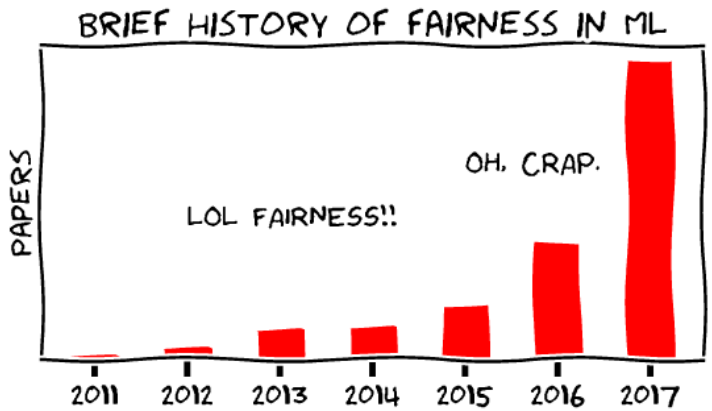
- Study based on voter record in the US[3]

Over-represented Attributes	Under-represented Attributes
Source Audience	
African American affinity (5.52)	Asian American affinity (0.09)
US politics: very liberal (3.21)	Hispanic (Spanish dominant) affinity (0.09)
Liberal content engagement (2.98)	Expats: Mexico (0.11)
Interest: Gospel music (2.64)	Hispanic (all) affinity (0.18)
Interest: Dancehalls (2.51)	Expats: all countries (0.22)
2% Look-Alike Audience	
African American affinity (5.24)	Hispanic (Spanish dominant) affinity (0.10)
Liberal content engagement (4.16)	Expats: Mexico (0.13)
US politics: very liberal (3.29)	Asian American affinity (0.13)
Interest: Gospel music (3.07)	Hispanic (all) affinity (0.19)
Interest: Soul music (2.32)	Expats: all countries (0.24)
2–4% Look-Alike Audience	
African American affinity (5.06)	Asian American affinity (0.17)
Liberal content engagement (3.61)	Hispanic (Spanish dominant) affinity (0.18)
US politics: very liberal (3.37)	Expats: Mexico (0.19)
Interest: Gospel music (2.72)	Hispanic (all) affinity (0.29)
Interest: Dancehalls (2.54)	Expats: all countries (0.37)

[3] [Speicher et al., 2018]

Fairness in ML is a complex topic

- Various origins of discrimination: skewed samples, feedback loops, limited features, proxies, metric optimized, bias/variance, etc.
- Highly interdisciplinary
- Different fairness doctrines (disparate impact vs treatment)
- Domain/task specific definitions



[Illustration by Hardt]

- This tutorial discusses only a small part: outcome group-fairness in classification, extensions to other tasks
- Lab based on the COMPAS dataset (classification)

Outline

Outline

Multiple definitions of fairness

There exist many different definitions of fairness, even just for classification.[4]

- Individual fairness
  - Similar individual should receive similar outcomes
  - Requires a metric of similarity and randomized classifiers

[4] [Narayanan, 2019]

Multiple definitions of fairness

There exist many different definitions of fairness, even just for classification.[4]

- Individual fairness
  - Similar individual should receive similar outcomes
  - Requires a metric of similarity and randomized classifiers
- Utility based fairness
  - Link to social choice theory, Pareto optimility
  - Economic mesures of inequalities such as Gini index

[4] [Narayanan, 2019]

Multiple definitions of fairness

There exist many different definitions of fairness, even just for classification.[4]

- Individual fairness
  - Similar individual should receive similar outcomes
  - Requires a metric of similarity and randomized classifiers
- Utility based fairness
  - Link to social choice theory, Pareto optimility
  - Economic mesures of inequalities such as Gini index
- Group fairness
  - Groups based on sensitive attributes
  - Disparate treatment: groups should be treated similarly (also called process fairness)
  - Disparate impact: groups should have similar outcomes

[4] [Narayanan, 2019]

Multiple definitions of fairness

There exist many different definitions of fairness, even just for classification.[4]

- Individual fairness
  - Similar individual should receive similar outcomes
  - Requires a metric of similarity and randomized classifiers
- Utility based fairness
  - Link to social choice theory, Pareto optimility
  - Economic mesures of inequalities such as Gini index
- Group fairness
  - Groups based on sensitive attributes
  - Disparate treatment: groups should be treated similarly (also called process fairness)
  - Disparate impact: groups should have similar outcomes

We focus on the last, but still many definitions exist depending on what we mean by outcome.

[4] [Narayanan, 2019]

French law on discrimination

Constitue une discrimination directe la situation dans laquelle, sur le fondement de son origine, de son sexe, de sa situation de famille, de sa grossesse, de son apparence physique, de la particulière vulnérabilité résultant de sa situation économique, apparente ou connue de son auteur, de son patronyme, de son lieu de résidence ou de sa domiciliation bancaire, de son état de santé, de sa perte d'autonomie, de son handicap, de ses caractéristiques génétiques, de ses mœurs, de son orientation sexuelle, de son identité de genre, de son âge, de ses opinions politiques, de ses activités syndicales, de sa capacité à s'exprimer dans une langue autre que le français, de son appartenance ou de sa non-appartenance, vraie ou supposée, à une ethnie, une nation, une prétendue race ou une religion déterminée, une personne est traitée de manière moins favorable qu'une autre ne l'est, ne l'a été ou ne l'aura été dans une situation comparable.

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.

[LOI no 2008-496 du 27 mai 2008]

Setup for group-fairness in classification

- $X \in \mathcal{X}$: feature vector (e.g., GPA). Typically $\mathcal{X} = \mathbb{R}^d$
- $A \in \mathcal{A}$: sensitive attribute (e.g., gender). Typically $\mathcal{A} = \{a, b\}$
- $Y \in \{0, 1\}$: (binary) label (e.g., performs well at job)
- $\hat{Y} \in \{0, 1\}$: predicted label (e.g., hired or not)
- $R = r(X, A)$: score (e.g., Bayesian classifier $R = \mathbb{E}[Y|X, A]$)
  - Get $\hat{Y}$ by thresholding on R (i.e., $\hat{Y} = 1$ iif $R \geq \theta$)

- There exist many fairness criteria
  - Demographic/statistical parity, equal opportunity, calibration...
- They belong to three main categories
  - Independence, separation, sufficiency

[Barocas, Hardt, Narayanan, 2020]: Chapter 3

Fairness metric I: Independence

- Definition: $\hat{Y}$ is independent of A (denoted $\hat{Y} \perp\!\!\!\perp A$)
- Requires $\mathbb{P}(\hat{Y} = 1|A = a) = \mathbb{P}(\hat{Y} = 1|A = b)$
  - Equal selection rate across groups
- Also called demographic parity or statistical parity
- Approximate versions (e.g., 4/5-th rule)
- Simple notion of fairness, often used with principled arguments
- Shortcomings
  - Rules out perfect classifier $\hat{Y} = Y$ if Y correlated with A
  - Allows selecting carefully in one group and randomly in other

Fairness metric II: Separation

- Definition: $\hat{Y} \perp\!\!\!\perp A | Y$, that is
  $\mathbb{P}(\hat{Y} = \hat{y} | Y = y, A = a) = \mathbb{P}(\hat{Y} = \hat{y} | Y = y, A = b), \forall y, \hat{y}$

- Uniform errors across groups, allows $\hat{Y} = Y$
- For binary labels, equivalent to
  $\mathbb{P}(\hat{Y} = 1 | Y = 1, A = a) = \mathbb{P}(\hat{Y} = 1 | Y = 1, A = b)$ and
  $\mathbb{P}(\hat{Y} = 1 | Y = 0, A = a) = \mathbb{P}(\hat{Y} = 1 | Y = 0, A = b)$
  - equal TPR and FPR (also called equalized odds)

Proposed independently by [Zafar et al., 2017] and [Hardt et al., 2016]

Fairness metric II: Separation

- Definition: $\hat{Y} \perp\!\!\!\perp A | Y$, that is
  $\mathbb{P}(\hat{Y} = \hat{y} | Y = y, A = a) = \mathbb{P}(\hat{Y} = \hat{y} | Y = y, A = b), \forall y, \hat{y}$
- Uniform errors across groups, allows $\hat{Y} = Y$
- For binary labels, equivalent to
  $\mathbb{P}(\hat{Y} = 1 | Y = 1, A = a) = \mathbb{P}(\hat{Y} = 1 | Y = 1, A = b)$ and
  $\mathbb{P}(\hat{Y} = 1 | Y = 0, A = a) = \mathbb{P}(\hat{Y} = 1 | Y = 0, A = b)$
  - equal TPR and FPR (also called equalized odds)
- Relaxed "equal opportunity": equalize TPR across groups
  - equivalently, equalize FNR
  - also called disparate mistreatement

Proposed independently by [Zafar et al., 2017] and [Hardt et al., 2016]

Fairness metric III: Sufficiency

- Definition: $Y \perp\!\!\!\perp A | R$
  - For the purpose of predicting Y, R subsumes A

- For binary labels:
  $\mathbb{P}(Y = 1 | R = r, A = a) = \mathbb{P}(Y = 1 | R = r, A = b), \quad \forall r$
- If $R \in \{0, 1\}$, this is equalizing PPV/NPV (positive/negative predictive value) across groups

[5] See, e.g., [Barocas, Hardt, Narayanan, 2020]: Chapter 3

Fairness metric III: Sufficiency

- Definition: $Y \perp\!\!\!\perp A | R$
  - For the purpose of predicting Y, R subsumes A

- For binary labels:
  $\mathbb{P}(Y = 1 | R = r, A = a) = \mathbb{P}(Y = 1 | R = r, A = b), \quad \forall r$
- If $R \in \{0, 1\}$, this is equalizing PPV/NPV (positive/negative predictive value) across groups

- Sufficiency is implied by calibration per group:
  $\mathbb{P}(Y = 1 | R = r, A = g) = r, \quad \forall r, g$
  - Score interpreted as a probability
  - Satisfied by Bayes-optimal score $R = \mathbb{E}[Y | X, A]$
  - There exist standard techniques to make a score calibrated (e.g., Platt scaling)[5]

[5] See, e.g., [Barocas, Hardt, Narayanan, 2020]: Chapter 3

Relationship between fairness metrics

- Any two notions are mutually exclusive in most cases

Proposition (Independence *vs* Sufficiency)

If $Y \not\perp\!\!\!\perp A$, then independence and sufficiency cannot both hold.

Proof: If $Y \not\perp\!\!\!\perp A$ and $Y \perp\!\!\!\perp A|R$, then $A \not\perp\!\!\!\perp R$ □

Illustration. Proofs in [Barocas et al., 2020]: Chapter 3. Original versions (and other formulations) of trade-offs in [Chouldechova, 2017] and [Kleinberg et al., 2016].

Relationship between fairness metrics

- Any two notions are mutually exclusive in most cases

Proposition (Independence *vs* Sufficiency)

If $Y \not\perp\!\!\!\perp A$, then independence and sufficiency cannot both hold.

Proof: If $Y \not\perp\!\!\!\perp A$ and $Y \perp\!\!\!\perp A|R$, then $A \not\perp\!\!\!\perp R$ □

Proposition (Independence *vs* Separation)

If $Y \not\perp\!\!\!\perp A$ and $Y \not\perp\!\!\!\perp R$, then independence and separation cannot both hold.

Proof: If $R \perp\!\!\!\perp A$ and $R \perp\!\!\!\perp A|Y$, then either $Y \perp\!\!\!\perp A$ or $Y \perp\!\!\!\perp R$ □

Proposition (Separation *vs* Sufficiency)

If all events (Y, A, R) have positive (joint) probability and $Y \not\perp\!\!\!\perp A$, then sufficiency and separation cannot both hold.

Illustration. Proofs in [Barocas et al., 2020]: Chapter 3. Original versions (and other formulations) of trade-offs in [Chouldechova, 2017] and [Kleinberg et al., 2016].

Outline

How to achieve the different fairness metrics?

Three main categories of methods:

- Post-processing: take a classifier without changes and massage the output to satisfy fairness metrics
  - Works for any black-box classifier, potentially large utility loss

How to achieve the different fairness metrics?

Three main categories of methods:

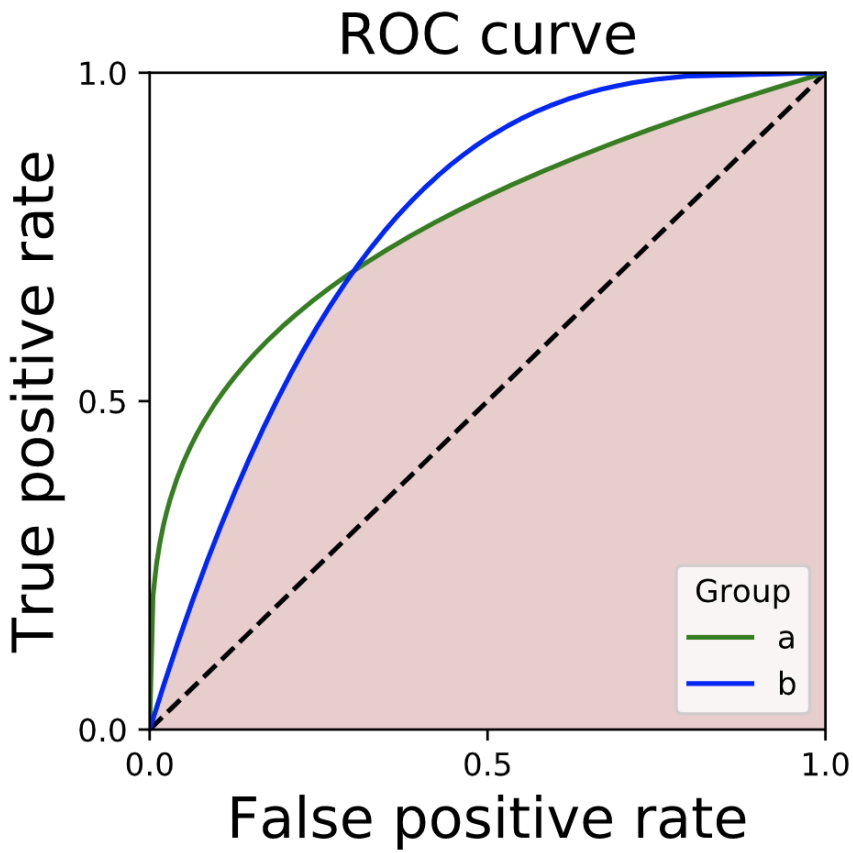
- Post-processing: take a classifier without changes and massage the output to satisfy fairness metrics
  - Works for any black-box classifier, potentially large utility loss
- In-processing: modify the training to include fairness constraints
  - Better utility, classifier-specific, requires access to data, leads to difficult optimization

How to achieve the different fairness metrics?

Three main categories of methods:

- Post-processing: take a classifier without changes and massage the output to satisfy fairness metrics
  - Works for any black-box classifier, potentially large utility loss
- In-processing: modify the training to include fairness constraints
  - Better utility, classifier-specific, requires access to data, leads to difficult optimization
- Pre-processing: learn fair representations and use them for downstream classification task
  - Agnostic to downstream task, specific to a fairness metric

Post-processing for equalized odds or equal opportunity



- Equal opportunity: take appropriate A-dependent thresholds
- Equalized odds: randomize between classifiers to achieve any point in the pink region

Method from [Hardt et al., 2016]. Illustration from [Barocas et al., 2020].

Optimal Bayesian classifier under demographic parity

The optimal Bayesian classifier is a group-dependent threshold on η.

Theorem

Let $p_g = P(A = g)$ *for* $g \in \{a, b\}$ *and* $\eta(X, A) = \mathbb{E}[Y|X, A]$. *Let*

$$\lambda^* \in \arg\max_{\lambda \in \mathbb{R}} \sum_{g \in \mathcal{A}} \mathbb{E}\left[\max\{p_g(2\eta(X, A) - 1) - \lambda(\mathbb{1}_{g=a} - \mathbb{1}_{g=b}), 0\}|A = g\right].$$

Then the following classifier minimizes risk amongst all demographically fair classifiers:

$$h^*(x, g) = \mathbb{1}\left(\eta(x, g) \geq \frac{1}{2} + \frac{(\mathbb{1}_{g=a} - \mathbb{1}_{g=b})\lambda^*}{2p_g}\right).$$

That is, h^* *is a solution of*

$$\min_{h:\mathcal{X}\times\mathcal{A}\to\{0,1\}} \{\mathbb{P}(h(X, A) \neq Y) : \mathbb{P}(h(X, A) = 1|A = a) = \mathbb{P}(h(X, A) = 1|A = b)\}$$

Various versions in a series of papers [Chzhen & Schreuder, 2022], [Gaucher et al., 2023] initiated by [Corbett-Davies et al., 2017]

Remarks on the optimal Bayesian classifier

- If (X, Y) is independent from A, then $\lambda^* = 0$
  $\Rightarrow$ we recover the unconstrained case
- Requires to known η, but also the distribution of $X|A$
- It is possible to construct a double plug-in estimator and analyze its properties[6]

[6] See e.g., [Chzhen et al., 2020]

In-processing

Standard classification: find $h \in \mathcal{H}$ that minimizes empirical risk

$$L(h) = 1/N \sum_{i=1}^{N} l(h(X, A), Y)$$

for some $\mathcal{H}$ and l that make the problem convex

Fair classification:

- Add a constraint such as $h(X, A) \perp\!\!\!\perp A | Y$
  - Highly intractable

- Key: replace constraint by a proxy that makes the problem tractable
- Typically based on covariances (i.e., second order approximation)
  - [Woodworth et al., 2017]
  - [Zafar et al., 2017]
  - [Zafar et al., 2017b] (for disparate impact)
  - [Zafar et al., 2019] (general framework)

Code available here for the Zafar et al. papers.

Example: in-processing for equal opportunity

Assume binary sensitive attributes $\mathcal{A} = 0, 1$, and recall the EO constraint: $\mathbb{P}(\hat{Y} = 1 | Y = 1, A = 0) = \mathbb{P}(\hat{Y} = 1 | Y = 1, A = 1)$.

Method from [Zafar et al., 2017]

Example: in-processing for equal opportunity

Assume binary sensitive attributes $\mathcal{A} = 0, 1$, and recall the EO constraint: $\mathbb{P}(\hat{Y} = 1 | Y = 1, A = 0) = \mathbb{P}(\hat{Y} = 1 | Y = 1, A = 1)$.

Take a class of classifiers parameterized by θ such that $L(\theta)$ is convex

- Example of logistic regression:
  $L(\theta) = 1/N \sum_{i=1}^{N} \log p(y_i | x_i, \theta)$ where $p(y_i | x_i, \theta) = \frac{1}{1+\exp(-\theta^T x_i)}$

Method from [Zafar et al., 2017]

Example: in-processing for equal opportunity

Assume binary sensitive attributes $\mathcal{A} = 0, 1$, and recall the EO constraint: $\mathbb{P}(\hat{Y} = 1 | Y = 1, A = 0) = \mathbb{P}(\hat{Y} = 1 | Y = 1, A = 1)$.

Take a class of classifiers parameterized by θ such that $L(\theta)$ is convex

- Example of logistic regression:
  $L(\theta) = 1/N \sum_{i=1}^{N} \log p(y_i | x_i, \theta)$ where $p(y_i | x_i, \theta) = \frac{1}{1+\exp(-\theta^T x_i)}$

Denote by $d_\theta(x))$ the distance from x to the decision boundary and $d_\theta^-(x)) = \min\{0, d_\theta(x))\}$; and take the proxy for EO:

$$Cov(A, Y d_\theta^-(X)) \simeq 1/N \sum_{i=1}^{N} (a_i - \bar{a}) y_i d_\theta^-(x_i)$$
$$= -\frac{N_1}{N} \sum_{D_0} y_i d_\theta^-(x_i) + \frac{N_0}{N} \sum_{D_1} y_i d_\theta^-(x_i),$$

where D_0 is the subdataset for $A = 0$, and N_0 its size (same for D_1, N_1).

Method from [Zafar et al., 2017]

Example: in-processing for equal opportunity (cont'd)

Now, we have to solve:

$$\min_{\theta} L(\theta)$$

$$s.t. -\frac{N_1}{N}\sum_{D_0} y_i d_\theta^-(x_i) + \frac{N_0}{N}\sum_{D_1} y_i d_\theta^-(x_i) \leq c$$

$$-\frac{N_1}{N}\sum_{D_0} y_i d_\theta^-(x_i) + \frac{N_0}{N}\sum_{D_1} y_i d_\theta^-(x_i) \geq -c.$$

Method from [Zafar et al., 2017]

Example: in-processing for equal opportunity (cont'd)

Now, we have to solve:

$$\min_{\theta} L(\theta)$$

$$s.t. -\frac{N_1}{N}\sum_{D_0} y_i d_\theta^-(x_i) + \frac{N_0}{N}\sum_{D_1} y_i d_\theta^-(x_i) \leq c$$

$$-\frac{N_1}{N}\sum_{D_0} y_i d_\theta^-(x_i) + \frac{N_0}{N}\sum_{D_1} y_i d_\theta^-(x_i) \geq -c.$$

- This is a "Disciplined Concave-Convex Program"
  - This objective is convex
  - The constraints are differences of convex functions
- There are known heuristics to solve these [Shen et al., 2016]

Method from [Zafar et al., 2017]

Example: in-processing for equal opportunity (cont'd)

Now, we have to solve:

$$\min_{\theta} L(\theta)$$

$$s.t. -\frac{N_1}{N}\sum_{D_0} y_i d_\theta^-(x_i) + \frac{N_0}{N}\sum_{D_1} y_i d_\theta^-(x_i) \leq c$$

$$-\frac{N_1}{N}\sum_{D_0} y_i d_\theta^-(x_i) + \frac{N_0}{N}\sum_{D_1} y_i d_\theta^-(x_i) \geq -c.$$

- This is a "Disciplined Concave-Convex Program"
  - This objective is convex
  - The constraints are differences of convex functions
- There are known heuristics to solve these [Shen et al., 2016]

- Experimentally: performs relatively well
  - Similar to post-processing [Hardt et al., 2016], a bit less accuracy drop
- No need for the sensitive attribute at test time (decision time)

Method from [Zafar et al., 2017]

Some other in-processing methods

- Reduction to cost-sensitive classification[7]
  - Take randomized classifiers μ over a class of classifiers
  - Do empirical risk minimization over the set of random classifiers, s.t. a fairness constraint
  - Introduce a Lagrangian with multiplier λ:
    $L(\mu, \lambda) = \mathit{risk}(\mu) + \lambda \cdot \mathit{constraint}(\mu)$
  - Now the problem is equivalent to finding a saddle point
    $\min_\mu \max_\lambda L(\mu, \lambda)$
  - Solve by classical exponentiated gradient descent
    - ★ the minimization of μ step is reduced to cost-sensitive classification
    - ★ can be solved efficiently in some cases

[7] [Agarwal et al., 2018]
[8] [Zhang et al., 2018]

Some other in-processing methods

- Reduction to cost-sensitive classification[7]
  - Take randomized classifiers μ over a class of classifiers
  - Do empirical risk minimization over the set of random classifiers, s.t. a fairness constraint
  - Introduce a Lagrangian with multiplier λ:
    $L(\mu, \lambda) = \mathit{risk}(\mu) + \lambda \cdot \mathit{constraint}(\mu)$
  - Now the problem is equivalent to finding a saddle point
    $\min_\mu \max_\lambda L(\mu, \lambda)$
  - Solve by classical exponentiated gradient descent
    - ★ the minimization of μ step is reduced to cost-sensitive classification
    - ★ can be solved efficiently in some cases
- Adversarial training[8]
  - Maximize predictor's ability to predict label Y
  - Minimize adversary's ability to predict sensitive attribute A

[7] [Agarwal et al., 2018]
[8] [Zhang et al., 2018]

Pre-processing

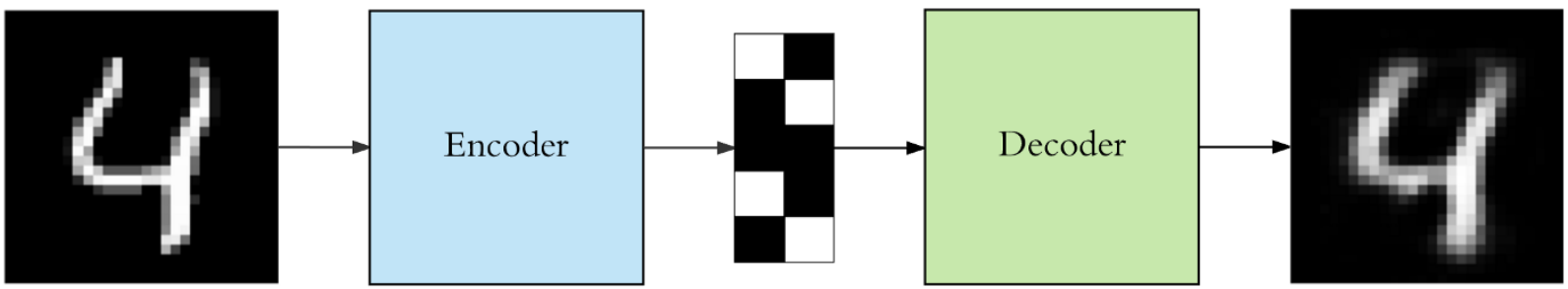
- Learn a *representation* $g : X, A \rightarrow Z$ such that
  - Z retains as much information as possible from X
  - Z is independent from A

- What is a representation and how to learn a good representation?
  - Learn a mapping from raw data to low-dimensional representation
  - Useful for many things...
    - ★ Multi-task, transfer learning, domain adaptation, generation, etc.
  - "Good" representation
    - ★ Makes subsequent task easy
    - ★ Disentangle factors of variation
    - ★ Allows introducing fairness constraints

Outline

Autoencoders: basic idea

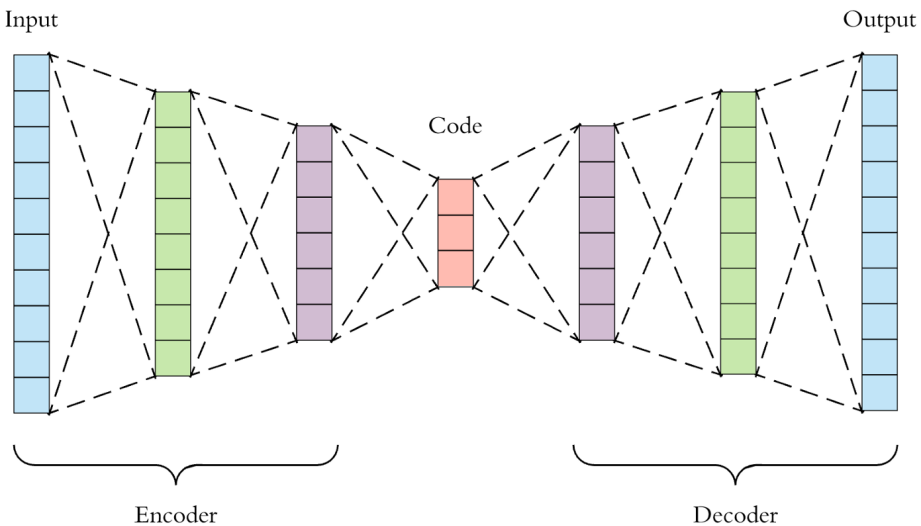
- Learn automatically a code that allows to reproduce the input
- Minimize reconstruction error $\mathbb{E}||x - \hat{x}||_2^2$
  - input x
  - code $h = encoder(x)$
  - output $\hat{x} = decoder(h) = decoder(encoder(x))$



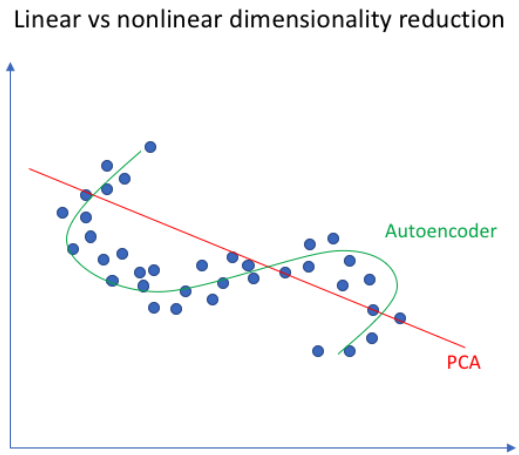
[Picture by A. Dertat]

Nonlinear autoencoder

- Use neural networks with non-linear activation to encode non-linear function



[Picture by A. Dertat]

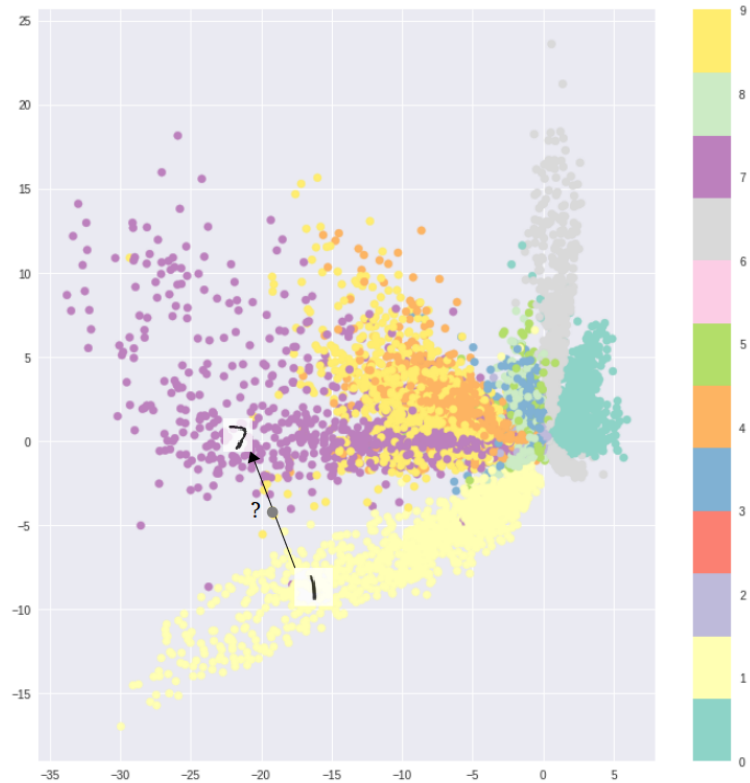


[Picture by J. Jordan]

- Training through backpropagation (same as regular neural nets)

Limitation of autoencoders

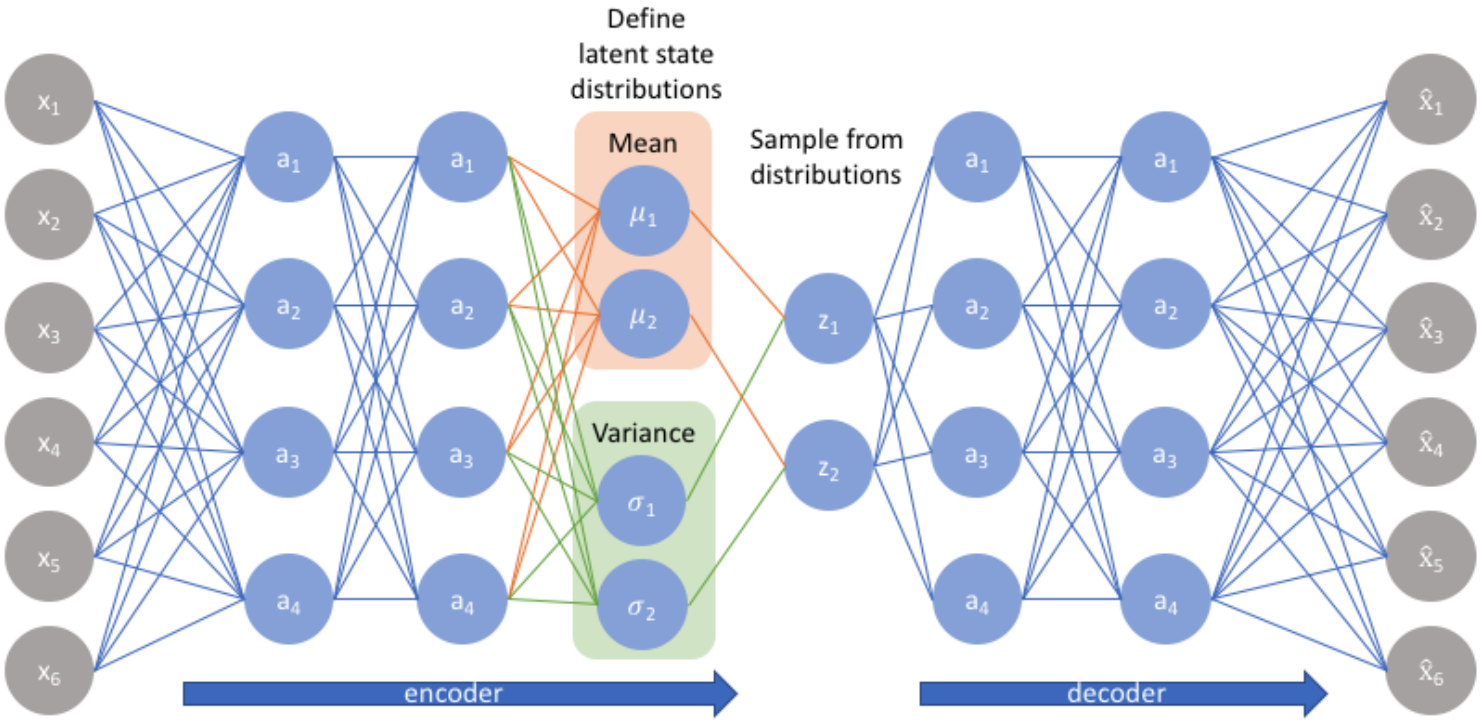
- No regularity of latent space $\Rightarrow$ cannot be used for generation



[Picture by I. Shafkat]

$\Rightarrow$ Solution: variational autoencoder (VAE)

Variational autoencoder (VAE)

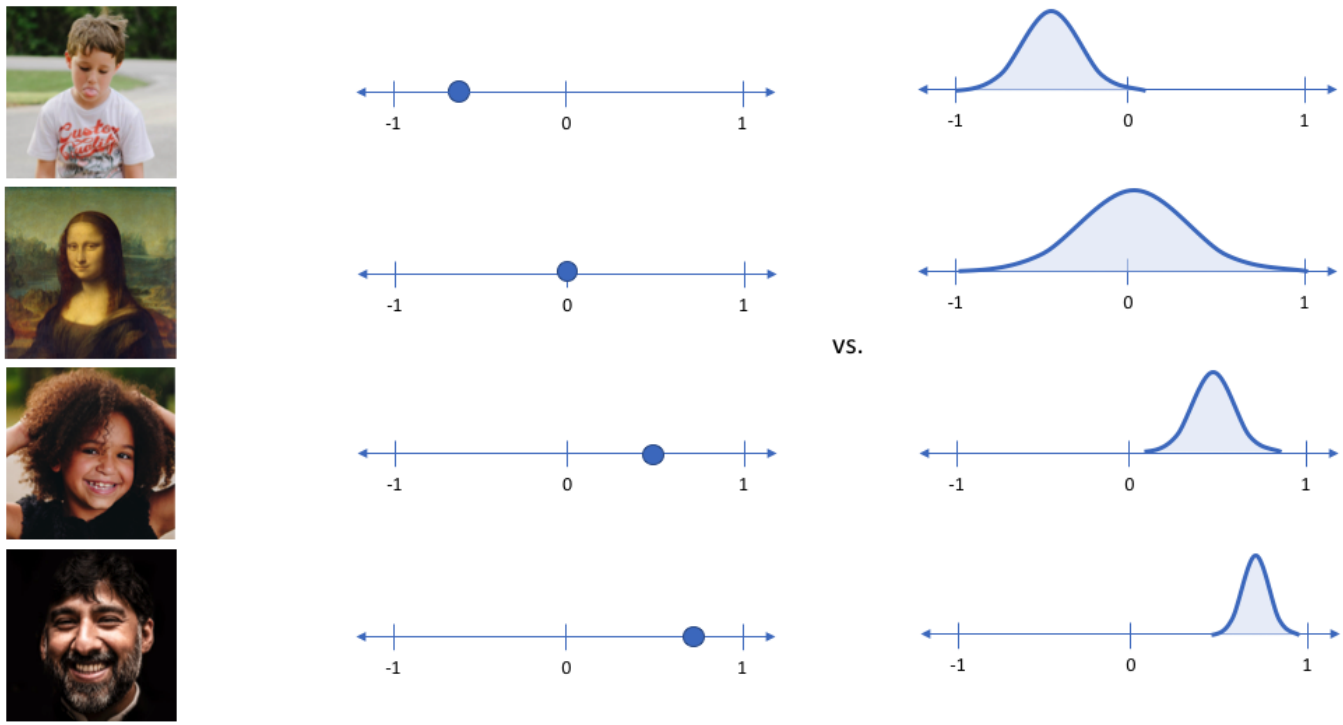


[Picture by J. Jordan]

- Similar to an AE encoding data as distributions

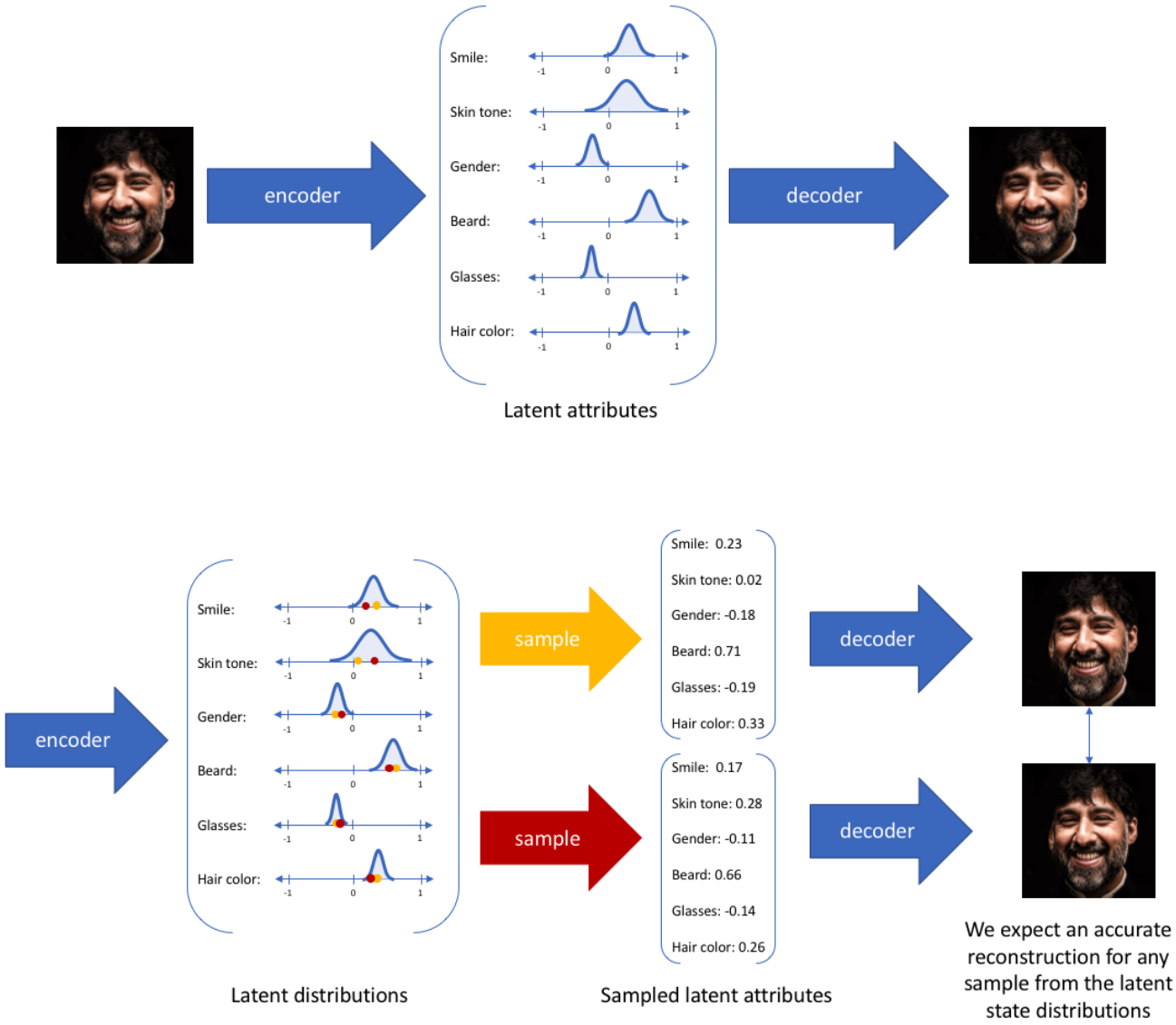
Proposed by [Kingma, Welling, 2014]. Tutorial by [Doersch, 2016].

VAE illustration



[Picture by J. Jordan]

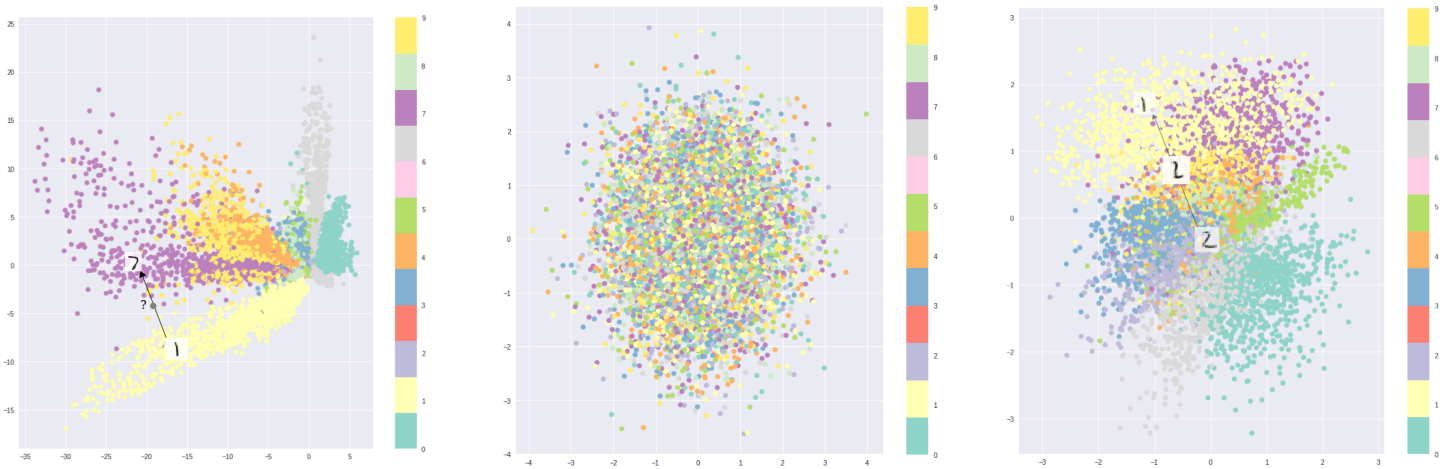
VAE illustration II



[Picture by J. Jordan]

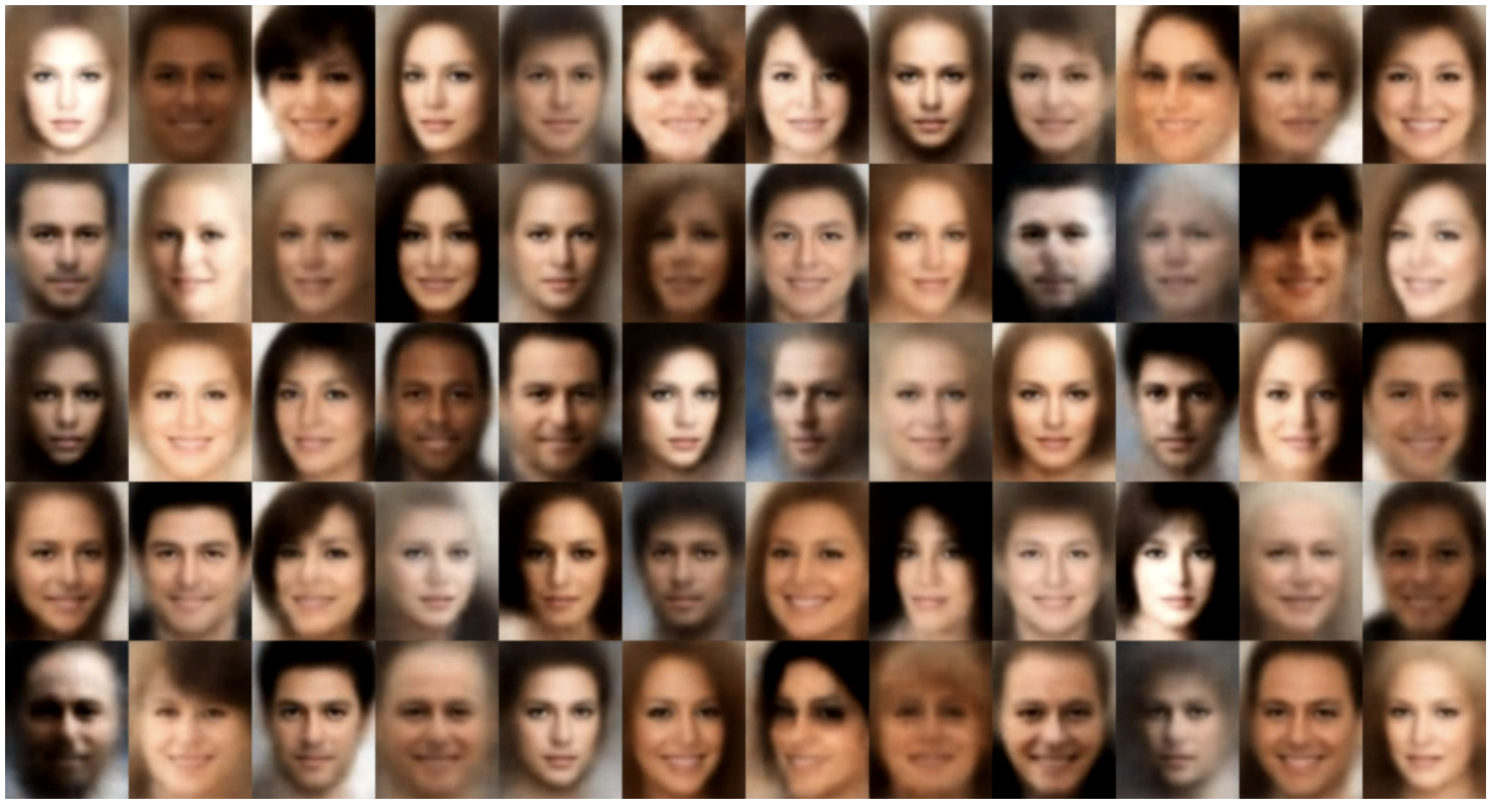
VAE objective function interpretation

- We optimize $\mathcal{L}(\phi, \theta) = \underbrace{\mathbb{E}_{z \sim q_\phi}[\log p_\theta(x|z)]}_{\text{reconstruction error}} - \underbrace{KL(q_\phi(z|x)||p(z))}_{\text{regularization}}$
  - Reconstruction error $\sim ||x - \hat{x}||^2$
    - ★ Estimated by sampling just one value of z (because $q_\phi(z|x)$ concentrated around values of z likely to have generated x)
  - Regularization makes distributions close to $\mathcal{N}(0, 1)$
    - ★ There is a closed-form expression for it
  - Training by back-propagation with the re-parameterization trick



[Picture by I. Shafkat]

VAE example: faces generation

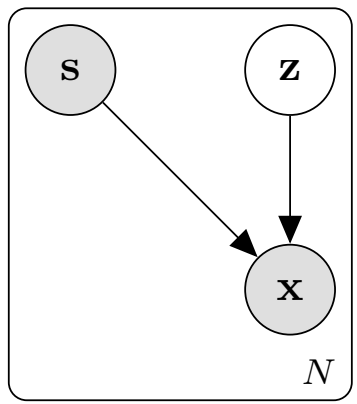


[Picture from https://github.com/wojciechmo/vae]

Outline

The variational fair autoencoder (VFAE)

- Goal: learn a representation $g : X, S, (Y) \rightarrow Z$ such that Z is independent from S (S = sensitive attribute)
  - Statistical parity



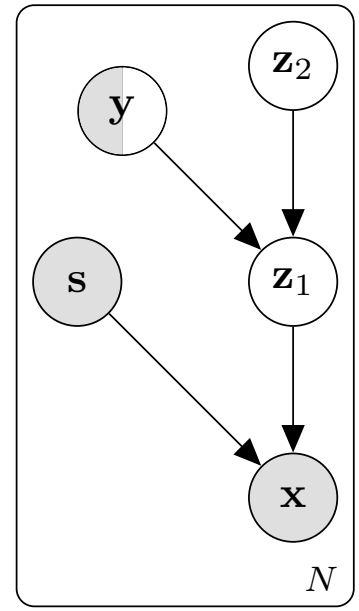
[Figure from [Louizos et al., 2016]]

- VAE architecture with $p_\theta(x|z,s)$ and $q_\phi(z|x,s)$ conditioned on s
- Objective function (trained with both x_i and s_i for all data i):

$$\mathcal{L}(\phi,\theta) = \mathbb{E}_{z\sim q_\phi(z|x,s)}[\log p_\theta(x|z,s)] - KL(q_\phi(z|x,s)||p(z))$$

Proposed by [Louizos et al., 2016].

The variational fair autoencoder II: semi-supervised



[Figure from [Louizos et al., 2016]]

- But we want z to keep information about y:

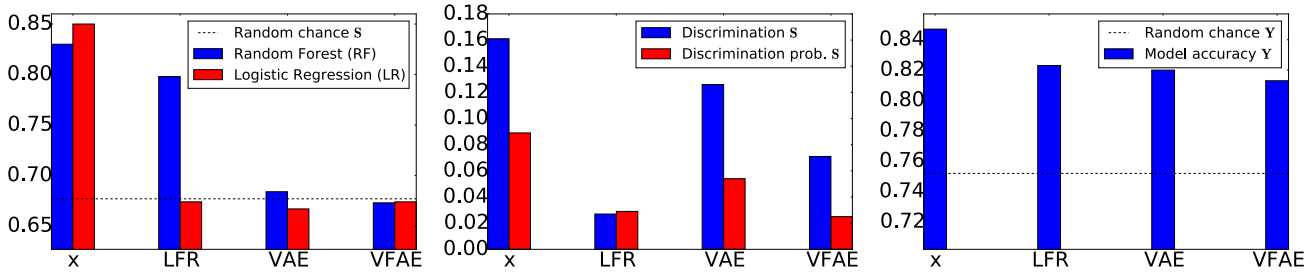
- Objective

$$\mathcal{L}_{VFAE}(\phi, \theta) = \underbrace{\mathcal{L}_{VAE}(\phi, \theta)}_{\text{VAE objective}} - \beta \underbrace{I_{MMD}(q_\phi(z_1|s=a), q_\phi(z_1|s=b))}_{\text{Maximum Mean Discrepency (MMD) penalty}}$$

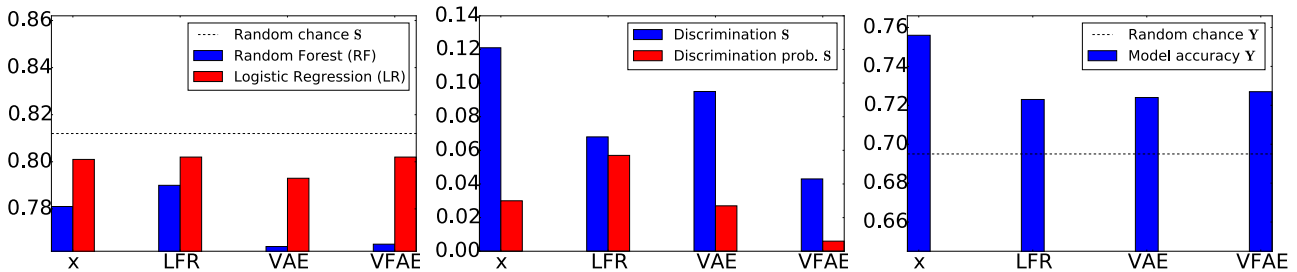
  - MMD penalty forces $q_\phi(z_1|s=a), q_\phi(z_1|s=b)$ to be close

Proposed by [Louizos et al., 2016]. Extensions by [Song et al., 2019].

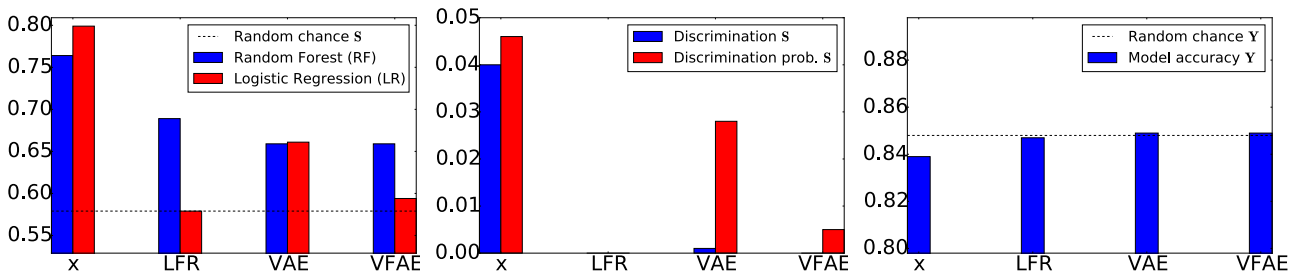
The variational fair autoencoder III: results



(a) Adult dataset



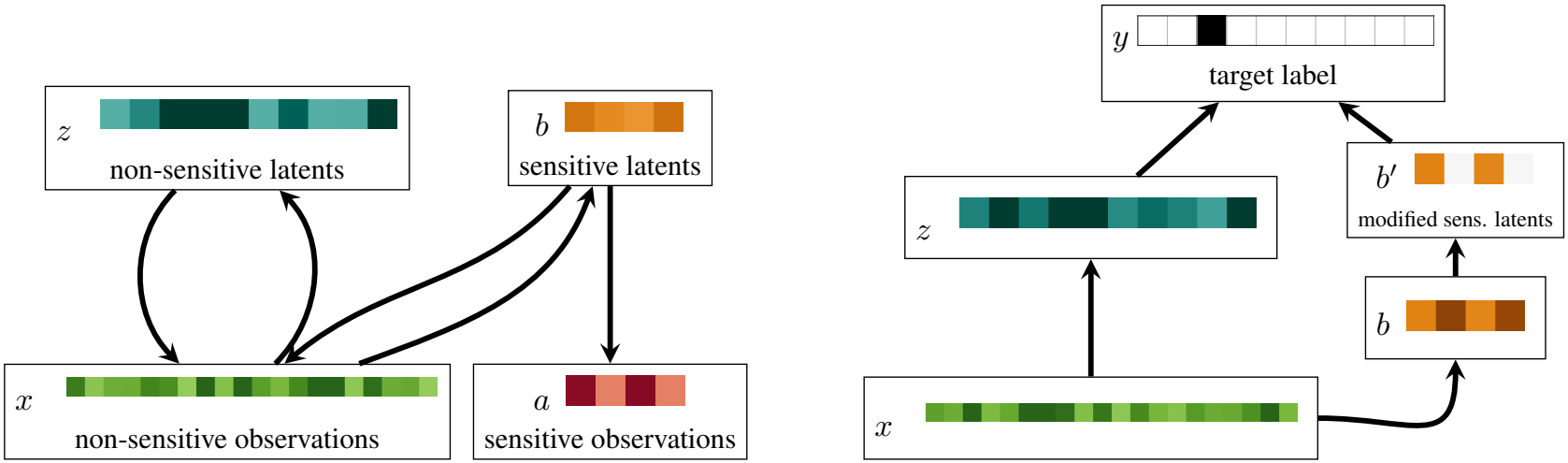
(b) German dataset



(c) Health dataset

[Figure from [Louizos et al., 2016]]

Flexibly Fair Representation Learning by Disentanglement



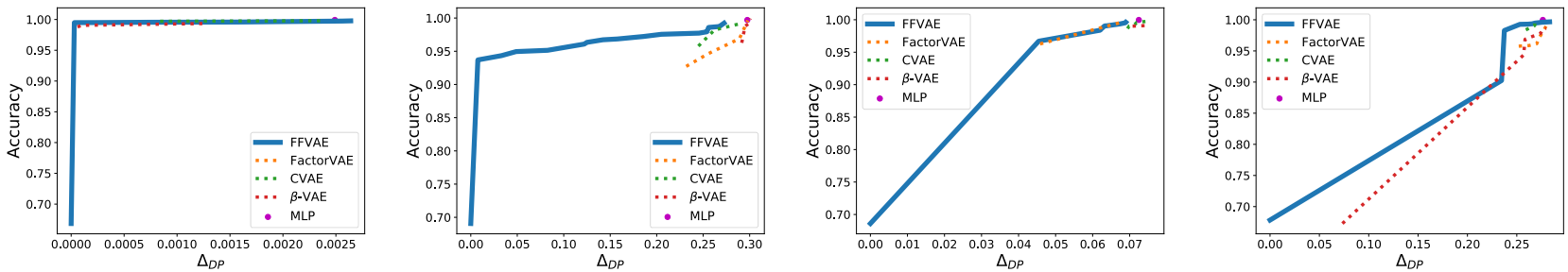
[Figure from [Creager et al., 2019]]

- Objective:

$$\mathcal{L}_{FFVAE}(\phi,\theta) = \overbrace{\mathbb{E}_{q_\phi(z,b|x)}[\log p_\theta(x|z,s)] - KL(q_\phi(z|x,s)||p(z))}^{\text{VAE objective}} + \underbrace{\alpha\mathbb{E}_{q_\phi(z,b|x)} log p_\theta(a|b)}_{\text{predictiveness term}} - \underbrace{\gamma KL[q_\phi(z,b)||q_\phi(z)\cdot\Pi_i q_\phi(b_i)]}_{\text{disentanglement term}}$$

Proposed by [Creager et al., 2019], talk video by R. Zemel here.

FFVAE results



(a) a = Scale (b) a = Shape (c) a = Shape $\wedge$ Scale (d) a = Shape $\vee$ Scale

Figure 2. Fairness-accuracy tradeoff curves, DSpritesUnfair dataset. We sweep a range of hyperparameters for each model and report Pareto fronts. Optimal point is the top left hand corner — this represents perfect accuracy and fairness. MLP is a baseline classifier trained directly on the input data. For each model, encoder outputs are modified to remove information about a. y = XPosition for each plot.

[Figure from [Creager et al., 2019]]

Proposed by [Creager et al., 2019], talk video by R. Zemel here.

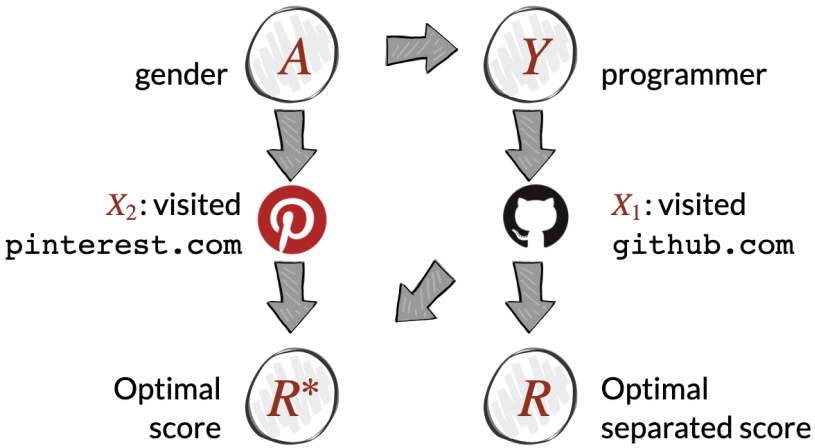
Outline

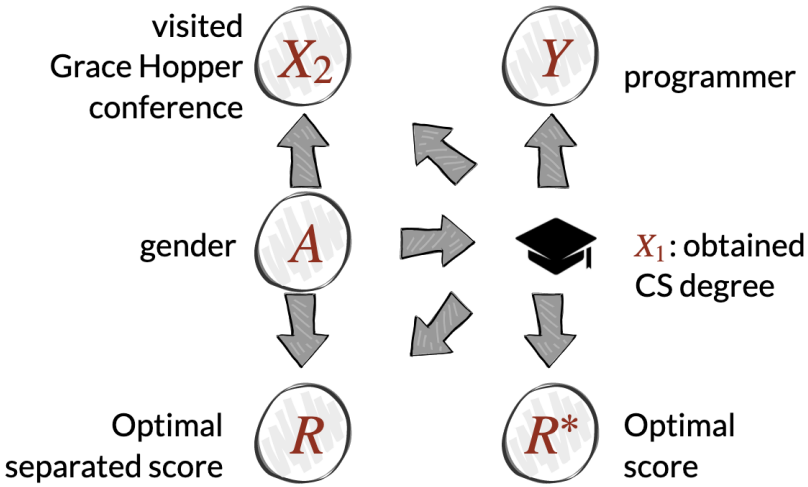
Reminder: observational fairness criteria

- Reminder: attributes X, sensitive A, label Y, score R (outcome $\hat{Y}$)
  - Independence: $R \perp\!\!\!\perp A$
  - Separation: $R \perp\!\!\!\perp A | Y$
  - Sufficiency: $Y \perp\!\!\!\perp A | R$
- All three definitions are observational
  - Depend only on joint distribution over all random variables (X, R, A, Y)
  - Can be estimated from observed data (up to estimation error)

Limitations of observational criteria

- Two scenarios that lead to the same joint distribution
  - cannot be distinguished from data/observations



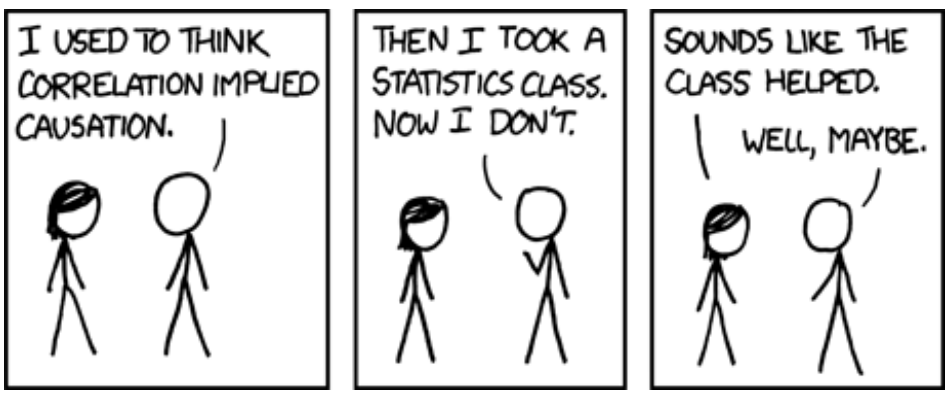


[Figures from [Hardt, 2017]]

$\Rightarrow$ very different interpretation from a fairness standpoint

Causality

- Defines a language and tools to distinguish observation *vs* actions
  - Do-calculus: probability conditioned on doing ($do(X = x)$)
  - Counterfactuals
- Can we recover the causal graph from observations?
  - With extra assumptions (simplicity, form of the noise, etc.)
  - With active experiments (interventions)
- References
  - Section 3.4 of [Barber, 2012]
  - Chapters 3 and 5 of [Barocas et al, 2020]
  - Many books for more details, e.g., [Peters et al., 2017]



[Cartoon from xkcd]

Other/open topics on fairness

- A lot of other problems, many of which are nascent or open
- Tasks other than classification
  - Regression, recommendation, ranking, matching
  - Reinforcement learning, dynamic aspects
- Multi-sided and multi-stakeholders scenarios
- Multi-dimensional sensitive attributes
  - Intersectionality
- Multi-agent systems (e.g., ad auctions)
- Link fairness / privacy
- ...

Main general references

- Book "fairness and ML" [Barocas et al, 2020]
- Tutorials on fairness
  - Fairness-Aware Machine Learning in Practice [Bird et al., 2019]
  - Fairness in ML [Barocas & Hardt, 2017] slides + video
  - 21 fairness definitions and their politics [Narayanan, 2018]
  - Fairness and representation learning tutorial [Cisse, Koyejo, 2019]
- Book "Pattern recognition and ML" [Bishop, 2006]:
- Tutorial on VAE [Doersch, 2016]