Ethics in Machine Learning: Measuring and Mitigating Algorithmic Bias

COMP90049
Introduction to Machine Learning
Semester 2, 2021

Lida Rashidi, CIS

Acknowledgement: Lea Frermann

Roadmap

So far... ML nuts and bolts

- Supervised learning / classification
- Unsupervised learning
- Features selection
- Evaluation

Roadmap

So far... ML nuts and bolts

- Supervised learning / classification
- Unsupervised learning
- Features selection
- Evaluation

Today... ML in the world

- What is Bias and where does it come from?
- Algorithmic Fairness
- How to make ML algorithms fairer

Introduction

Applications

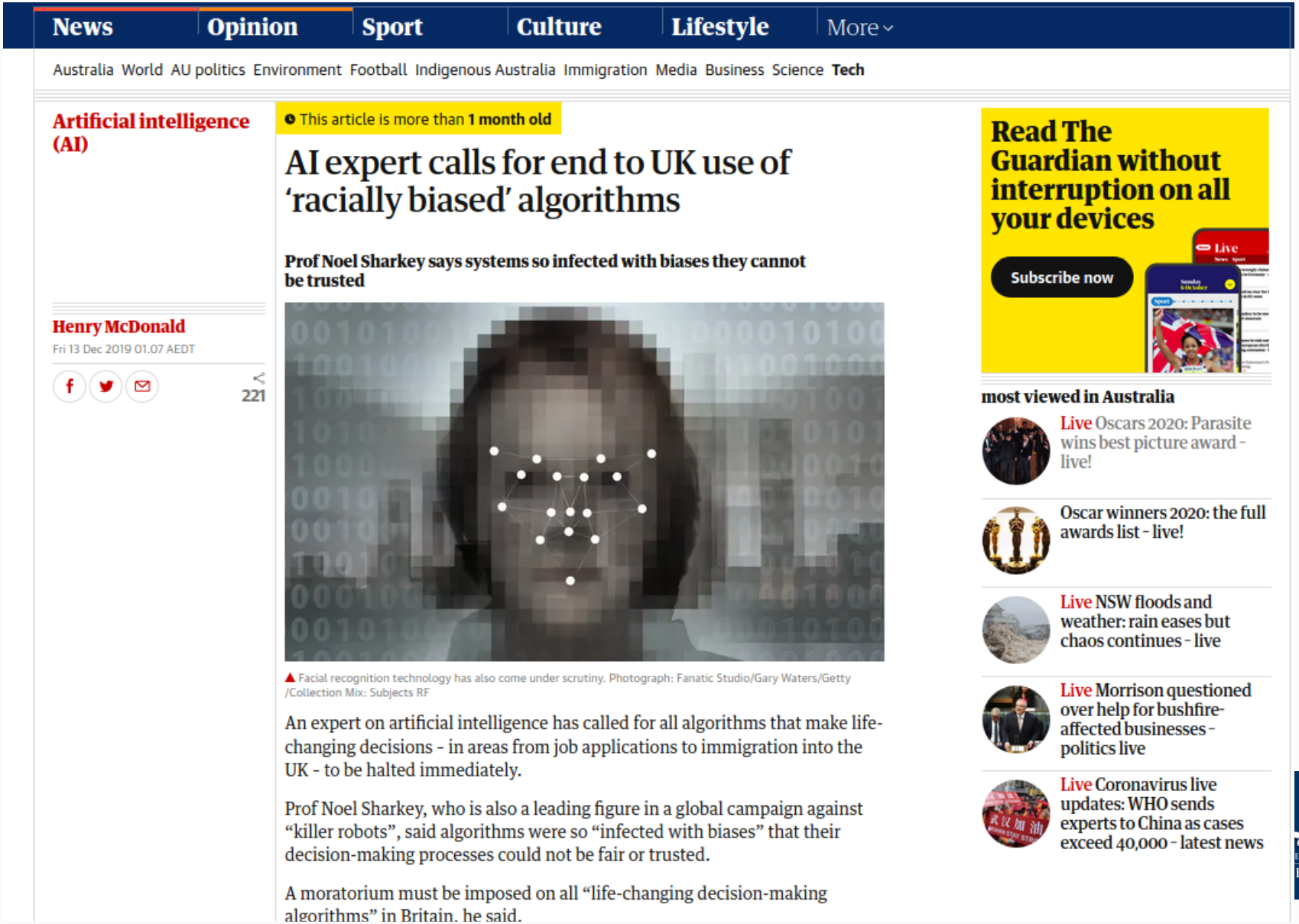
- medical diagnoses
- language generation
- hate speech detection
- stock market prediction
- insurance policy suggestion
- search and recommendation
- spam / malware detection
- predictive policing
- ...

Applications

- medical diagnoses
- language generation
- hate speech detection
- stock market prediction
- insurance policy suggestion
- search and recommendation
- spam / malware detection
- predictive policing
- ...

What can possibly go wrong?

A quick (and biased) press review



News | Opinion | Sport | Culture | Lifestyle | More

Australia World AU politics Environment Football Indigenous Australia Immigration Media Business Science **Tech**

Artificial intelligence (AI)

This article is more than **1 month old**

AI expert calls for end to UK use of 'racially biased' algorithms

Prof Noel Sharkey says systems so infected with biases they cannot be trusted

Henry McDonald

Fri 13 Dec 2019 01.07 AEDT

221

▲ Facial recognition technology has also come under scrutiny. Photograph: Fanatic Studio/Gary Waters/Getty /Collection Mix: Subjects RF

An expert on artificial intelligence has called for all algorithms that make life-changing decisions - in areas from job applications to immigration into the UK - to be halted immediately.

Prof Noel Sharkey, who is also a leading figure in a global campaign against "killer robots", said algorithms were so "infected with biases" that their decision-making processes could not be fair or trusted.

A moratorium must be imposed on all "life-changing decision-making algorithms" in Britain, he said.

most viewed in Australia

- Live Oscars 2020: Parasite wins best picture award - live!
- Oscar winners 2020: the full awards list - live!
- Live NSW floods and weather: rain eases but chaos continues - live
- Live Morrison questioned over help for bushfire-affected businesses - politics live
- Live Coronavirus live updates: WHO sends experts to China as cases exceed 40,000 - latest news

https://www.theguardian.com/technology/2019/dec/12/ai-end-uk-use-racially-biased-algorithms-noel-sharkey

A quick (and biased) press review

News | Opinion | Sport | Culture | Lifestyle | More

World ▸ Europe US Americas Asia Australia Middle East Africa Inequality Cities Global development

Artificial intelligence (AI)

Welfare surveillance system violates human rights, Dutch court rules

Government told to halt use of AI to detect fraud in decision hailed by privacy campaigners

Jon Henley *and* Robert Booth

Thu 6 Feb 2020 00.18 AEDT

1,303

▲ People in Rotterdam, the Netherlands. The Dutch system aimed to predict the likelihood of an individual committing benefit or tax fraud, or violating labour laws. Photograph: Geography Photos/UIG via Getty Images

A Dutch court has ordered the immediate halt of an automated surveillance system for detecting welfare fraud because it violates human rights, in a judgment likely to resonate well beyond the Netherlands.

The case was seen as an important legal challenge to the controversial but growing use by governments around the world of artificial intelligence (AI) and risk modelling in administering welfare benefits and other core services.

Campaigners say such "digital welfare states" - developed often without consultation, and operated secretively and without adequate oversight -

most viewed in Australia

Live Oscars 2020: Parasite wins best picture award - live!

Oscar winners 2020: the full awards list - live!

Live NSW floods and weather: rain eases but chaos continues - live

Live Morrison questioned over help for bushfire-affected businesses - politics live

Live Coronavirus live updates: WHO sends experts to China as cases exceed 40,000 - latest news

https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules

A quick (and biased) press review

Artificial intelligence (AI)

New AI fake text generator may be too dangerous to release, say creators

The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse

Alex Hern

@alexhern

Fri 15 Feb 2019 04.00 AEDT

6,686 572

▲ The AI wrote a new passage of fiction set in China after being fed the opening line of Nineteen Eighty-Four by George Orwell (pictured). Photograph: Mondadori/Getty Images

The creators of a revolutionary AI system that can write news stories and works of fiction – dubbed "deepfakes for text" – have taken the unusual step of not releasing their research publicly, for fear of potential misuse.

OpenAI, an nonprofit research company backed by Elon Musk, Reid Hoffman, Sam Altman, and others, says its new AI model, called GPT2 is so good and the risk of malicious use so high that it is breaking from its normal practice of releasing the full research to the public in order to allow more time to discuss the ramifications of the technological breakthrough.

most viewed in Australia

Live Oscars 2020: Parasite wins best picture award – live!

Oscar winners 2020: the full awards list – live!

Live NSW floods and weather: rain eases but chaos continues – live

Live Morrison questioned over help for bushfire-affected businesses – politics live

Live Coronavirus live updates: WHO sends experts to China as cases exceed 40,000 – latest news

https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction

A quick (and biased) press review

Opinion
Computing

Can the planet really afford the exorbitant power demands of machine learning?

John Naughton

Sun 17 Nov 2019 03.00 AEDT

270 348

The environmental impact of such technological advances can be huge

▲ Only huge firms such as Facebook can house the number of processors that machine learning requires. Photograph: Jim Thompson/Zuma Press/eyevine

There is, alas, no such thing as a free lunch. This simple and obvious truth is invariably forgotten whenever irrational exuberance teams up with digital technology in the latest quest to "change the world". A case in point was the bitcoin frenzy, where one could apparently become insanely rich by "mining" for the elusive coins. All you needed was to get a computer to solve a complicated mathematical puzzle and - lo! - you

most viewed in Australia


Live Oscars 2020: Parasite wins best picture award - live!


Oscar winners 2020: the full awards list - live!


Live NSW floods and weather: rain eases but chaos continues - live

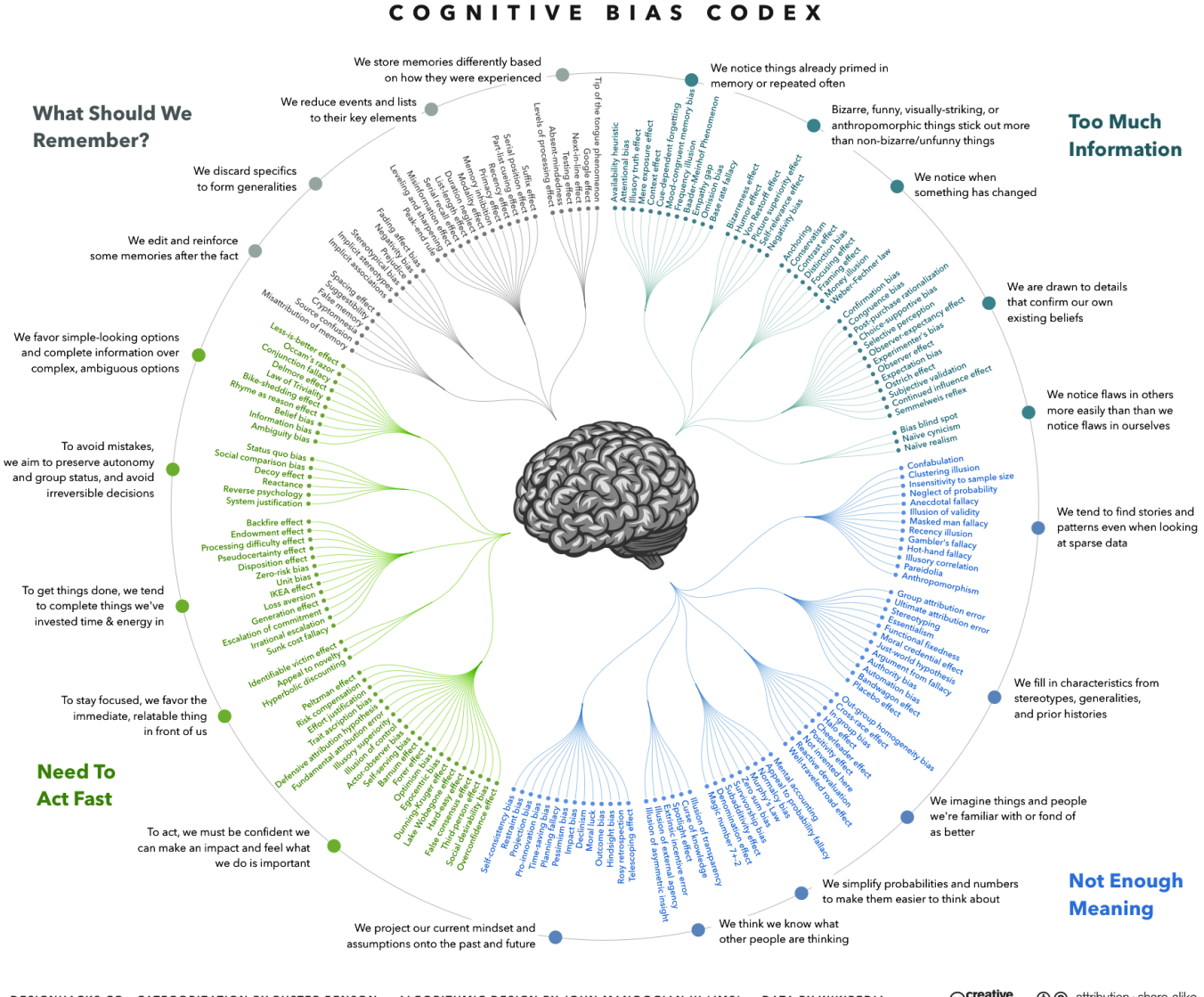

Live Morrison questioned over help for bushfire-affected businesses - politics live


Live Coronavirus live updates: WHO sends experts to China as cases exceed 40,000 - latest news

https://www.theguardian.com/commentisfree/2019/nov/16/can-planet-afford-exorbitant-power-demands-of-machine-learning

Sources and types of bias in ML

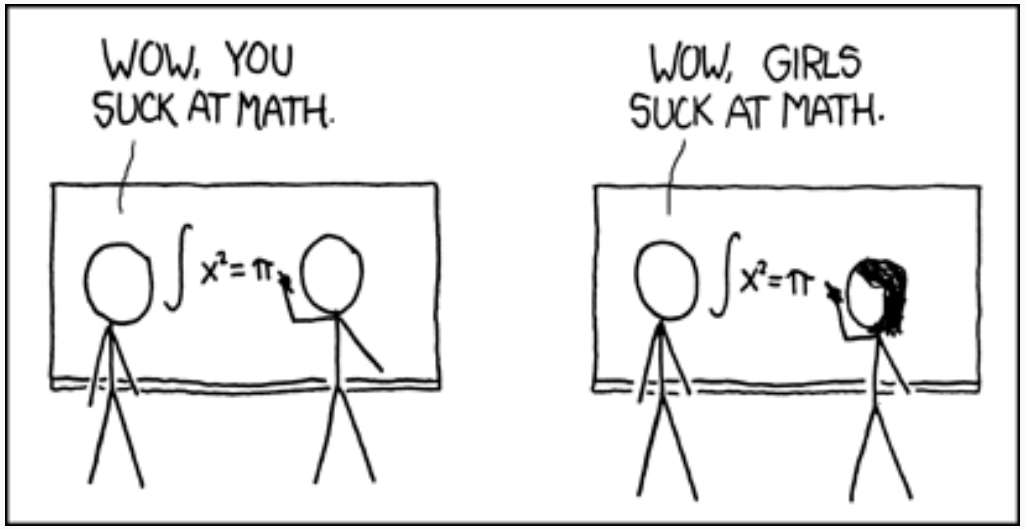
Humans are biased



https://commons.wikimedia.org/wiki/File:Cognitive_bias_codex_en.svg

Humans are biased

Out-group homogeneity bias (Stereotypes/Prejudice)

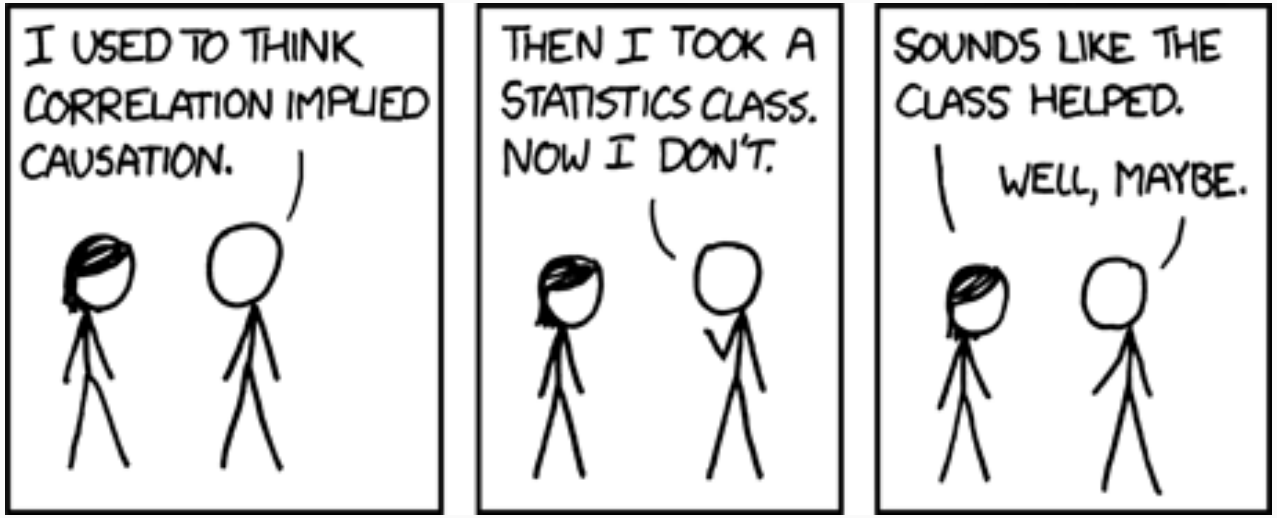


https://xkcd.com/385/

Humans tend to perceive out-group members as less nuanced than in-group members.

Humans are biased

Correlation Fallacy

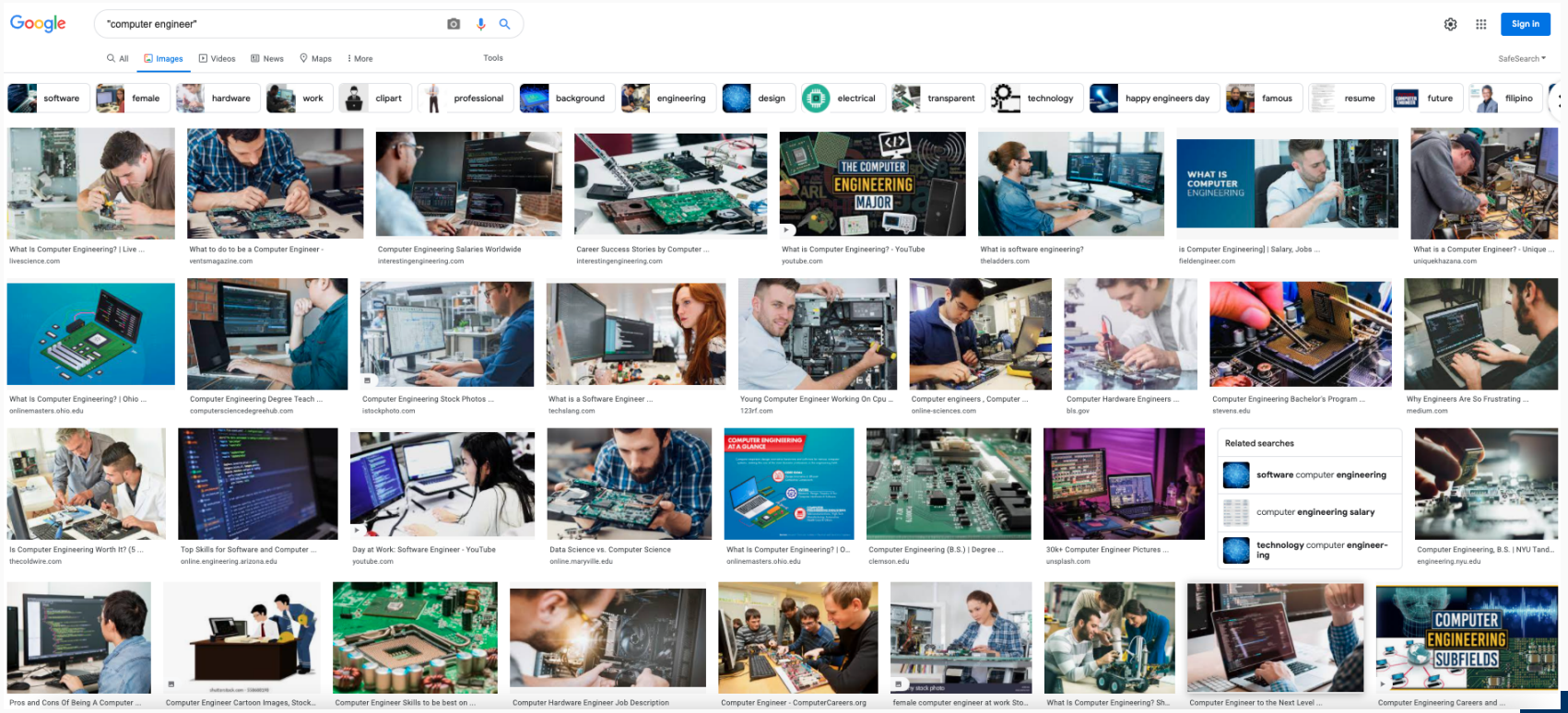


https://xkcd.com/552/

Humans have a tendency to mistake correlation (two co-incidentally co-occurring events) with causation.

Data is biased

Historical Bias: A randomly sampled data set, reflects the world *as it was* including existing biases which should not be carried forward



Data is biased

Representation bias / Reporting bias

- The data sets do not faithfully represent the whole population
- Minority groups are underrepresented
- Obvious facts are underrepresented. Anomalies are overemphasized.

Word	*Teraword*	*Word*	*Teraword*
spoke	11,577,917	hugged	610,040
laughed	3,904,519	blinked	390,692
murdered	2,843,529	was late	368,922
inhaled	984,613	exhaled	168,985
breathed	725,034	was punctual	5,045

From: Gordon and van Durme (2013)

Data is biased

Measurement bias

1. Noisy measurement $\rightarrow$ errors or missing data points which **are not randomly distributed**
    - e.g., records of police arrests differ in level of detail across postcode areas

2. Mistaking a **(noisy) proxy** for a label of interest
    - e.g., 'hiring decision' as a proxy for 'applicant quality'. **(why noisy?)**

3. **Oversimplification** of the quantity of interest
    - e.g., classifying political leaning into: 'Democrat' vs. 'Republican' (USA); binarizing gender into: 'Male' vs. 'Female'

Data is biased

Measurement bias

1. Noisy measurement $\rightarrow$ errors or missing data points which **are not randomly distributed**
    - e.g., records of police arrests differ in level of detail across postcode areas

2. Mistaking a **(noisy) proxy** for a label of interest
    - e.g., 'hiring decision' as a proxy for 'applicant quality'. **(why noisy?)**

3. **Oversimplification** of the quantity of interest
    - e.g., classifying political leaning into: 'Democrat' vs. 'Republican' (USA); binarizing gender into: 'Male' vs. 'Female'

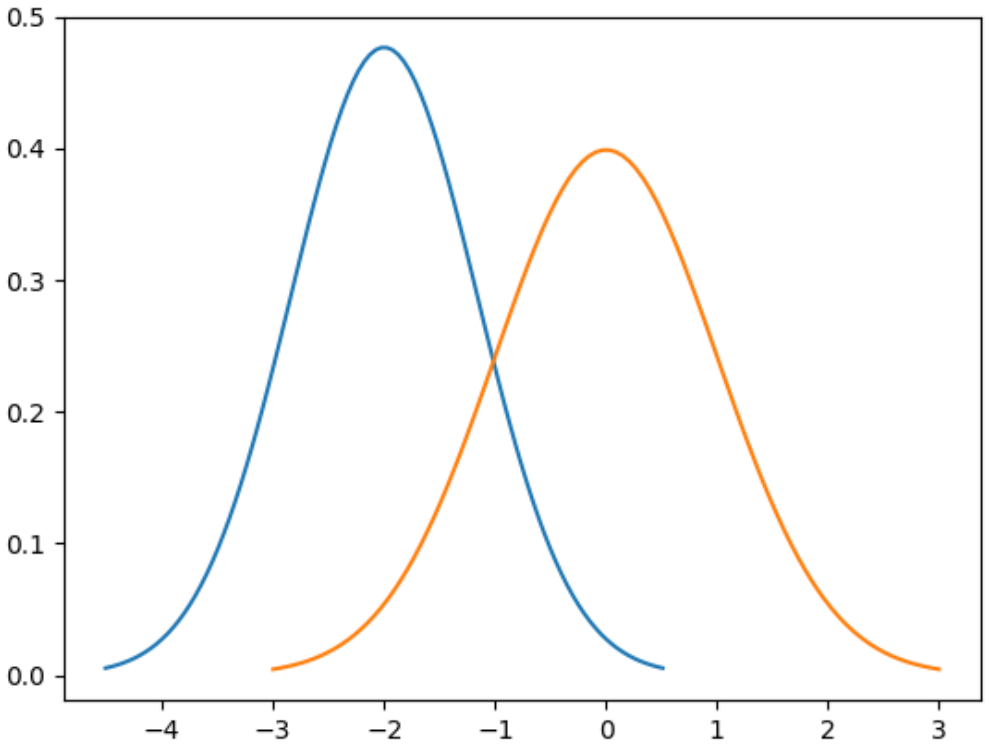
1. Know your domain 2. Know your task 3. Know your data

THE UNIVERSITY OF MELBOURNE

Models are Biased

Model Fit

- Weak models: high bias – low variance
- Unjustified model assumptions



Models are Biased

Biased Loss Function

- Blind to certain types of errors
- E.g., 0/1 loss will tend to tolerate errors in the minority class for highly imbalanced data

Models are Biased

Biased Loss Function

- Blind to certain types of errors
- E.g., 0/1 loss will tend to tolerate errors in the minority class for highly imbalanced data

1. Carefully consider model assumptions
2. Carefully choose loss functions
3. Model groups separately (e.g., multi-task learning)
4. Represent groups fairly in the data

THE UNIVERSITY OF MELBOURNE

Bias in Evaluation or Deployment

Evaluation bias

- Test set not representative of target population
- Overfit to a *test* set. Widely used benchmark data sets can reinforce the problem.
- Evaluation metrics may not capture all quantities of interest (disregard minority groups or average effects). E.g.,
  - Accuracy?
  - Face recognition models largely trained/evaluated on images of ethnically white people.

Deployment bias

- Use of systems in ways they weren't intended to use. Lack of education of end-users.

Bias in Evaluation or Deployment

Evaluation bias

- Test set not representative of target population
- Overfit to a *test* set. Widely used benchmark data sets can reinforce the problem.
- Evaluation metrics may not capture all quantities of interest (disregard

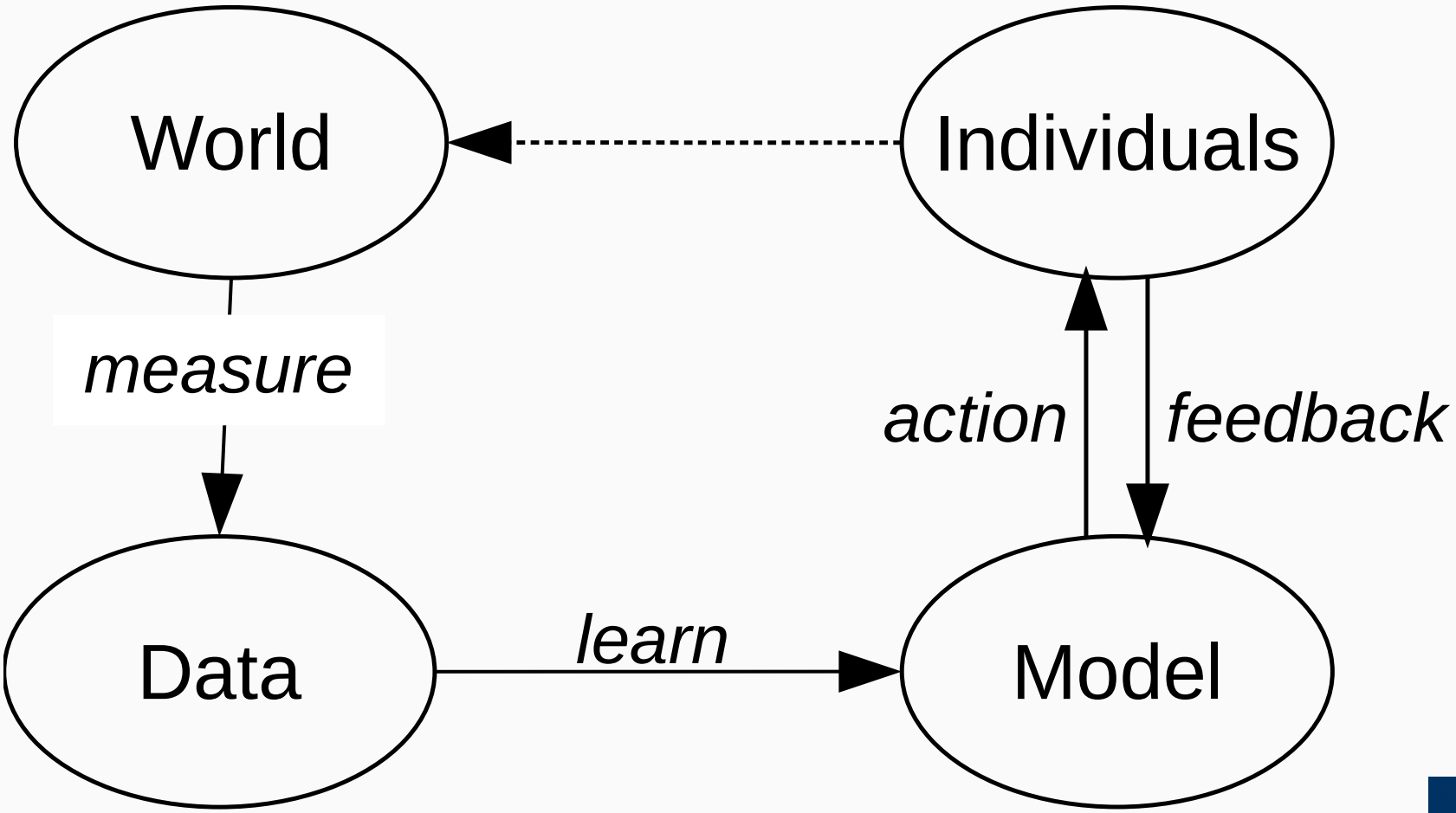
1. Carefully select your evaluation metrics
2. Use multiple evaluation metrics
3. Carefully select your test sets and benchmarks
4. Document your models to ensure they are used correctly

s of ethnically

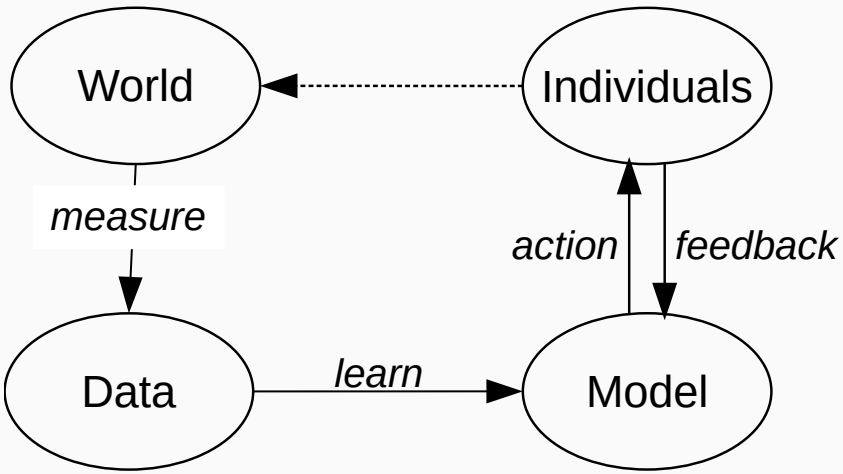
D

- Use of systems in ways they weren't intended to use. Lack of education of end-users.

The Machine Learning Pipeline



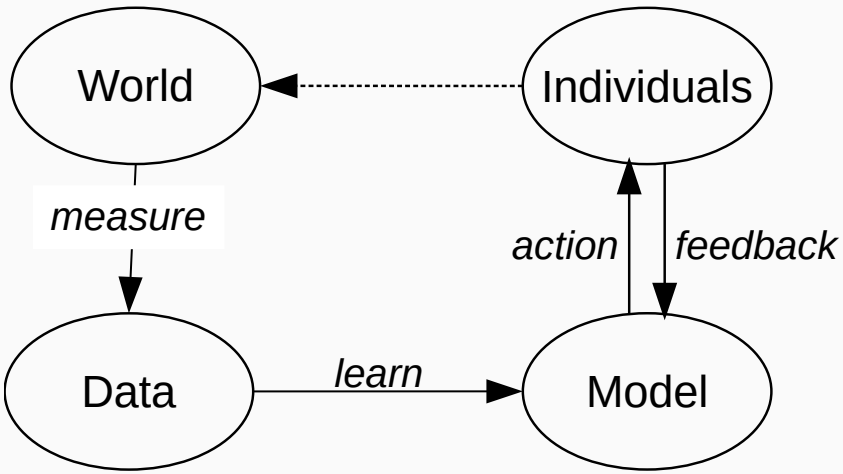
The Machine Learning Pipeline



Measurement

- Define your variables of interests
- Define your target variable
- Especially critical, if target variable is not measured explicitly.
  E.g., *hiring decision→applicant quality* or *income→creditworthiness*

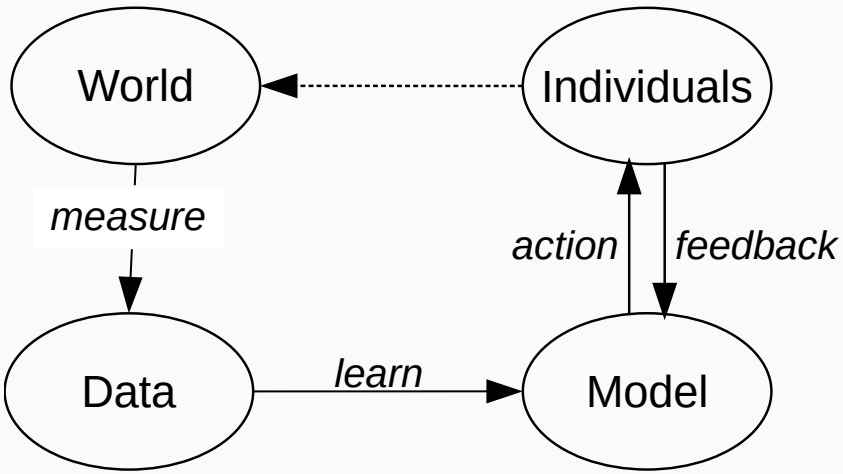
The Machine Learning Pipeline



Learning

- Models are faithful to the data (by design!)
- Data contains “knowledge” (*smoking causes cancer*)
- Data contains “Stereotypes” (*boys like blue, girls like pink*)
- What’s the difference? Based on social norms, no clear line!

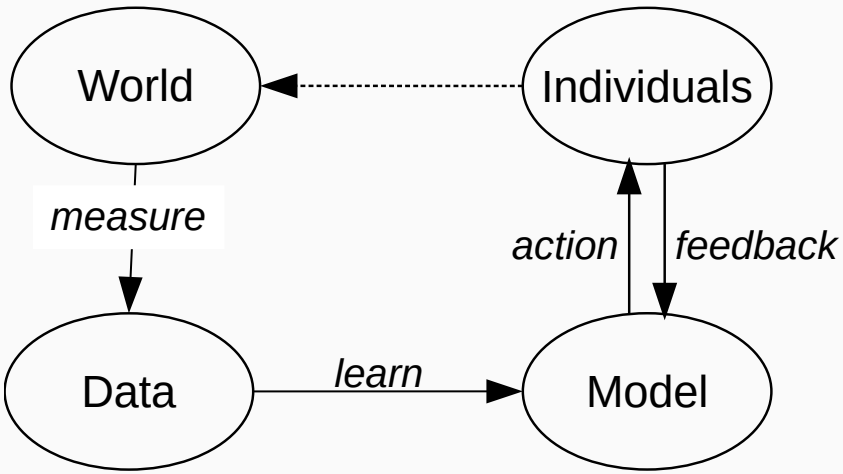
The Machine Learning Pipeline



Action

- ML concept: regression, classification, information retrieval, ...
- Resulting action: Class prediction (Spam, credit granted), search results, hotel recommendation, ...

The Machine Learning Pipeline



Feedback

- Approximated from user behavior
- Ex: click-rates

Demographic Disparity / Sample Size Disparity

Demographic groups will be differently represented in our samples

- Historical bias
- Minority groups
- ...

What does this mean for model fit?

- Models will work better for majorities (ex: dialects, speech recognition)
- Models will **generalize** based on majorities

Effects on society

- Minorities may adopt technology more slowly (increase the gap)
- The danger of feedback loops: (ex: predictive policing $\rightarrow$ more arrests $\rightarrow$ re-enforce model signal ...)

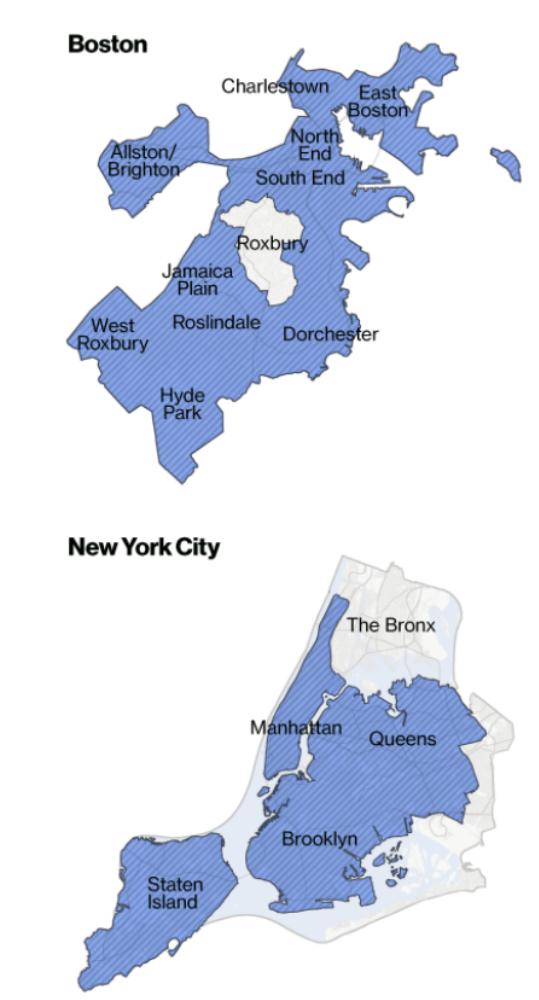
Demographic Disparity / Sample Size Disparity

Two questions

- Is any disparity **justified**?
- Is any disparity **harmful**?

Case study

- Amazon same-day delivery by zip-code (USA)
- Areas with largely black population are left out
- Amazon's objective: min cost / max efficiency
- Is the system biased?
- Is discrimination happening?
- If so: is the discrimination justified? Is it harmful?



https://www.bloomberg.com/graphics/2016-amazon-same-day/

Sensitive Attributes

Notation:

X	non-sensitive features
A	sensitive attributes with discrete labels (male/female, old/young, ...)
Y	true labels
$\hat{Y}$	classifier score (predicted label, for our purposes)

Very often instances have a mix of useful, uncontroversial attributes, and *sensitive attributes* based on which we do not want to make classification decisions.

Different attributes lead to different demographic groups of the population

It is rarely clear which attributes are sensitive and which are not. Choice can have profound impacts.

Fairness through Unawareness

Approach

- Hide all sensitive features from the classifier. Only train on X and remove A.

$$P(\hat{Y}_n|X_n, A_n) \approx P(\hat{Y}_n|X_n)$$

Another case study

- A bank which serves both **humans** and **martians** wants a classifier predicting whether an applicant should get a credit or not. Assume access to features (credit history, education, ...) for all applicants.
  - What are X, A, Y and $\hat{Y}$?
  - Apply "fairness through unawareness". Would the model be fair?

Problem

- General features may be strongly correlated with sensitive features. Example..?

Consequently, this approach does **not generally result in a fair model**

Formal Fairness Criteria

Quick recap of metrics

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	true positive (TP)	false negative (FN)
$y = 0$	false positive (FP)	true negative (TN)

Positive Predictive Value (PPV) (also: precision)

$$\frac{TP}{TP + FP}$$

True Positive Rate (TPR) (also: Recall)

$$\frac{TP}{TP + FN}$$

False Negative Rate (FNR):

$$\frac{FN}{TP + FN} = 1 - TPR$$

Accuracy

$$\frac{TP + TN}{TP + TN + FN + FP}$$

Example Problem: Credit scoring

- We trained a classifier to predict a binary credit score: should an applicant be granted a credit or not?

- Assume a version of the Adult data set as our training data (UCI Machine Learning Repository), which covers *humans* and *martians*. It includes actual credit scoring information

age	workclass	marital-status	race	*class*
39	State-gov	Never-married	White	*<=50K*
49	Self-emp-inc	Married-civ-spouse	White	*>50K*
28	Private	Married-civ-spouse	Other	*<=50K*
35	Private	Divorced	White	*>50K*
38	Private	Divorced	White	*<=50K*
53	Local-gov	Never-married	White	*<=50K*
28	Private	Married-civ-spouse	Black	*<=50K*
37	Private	Married-civ-spouse	Black	*>50K*

- We consider *species* to be the protected attribute: our classifier should make *fair* decisions for both human and martian applicants.

How can we measure fairness?

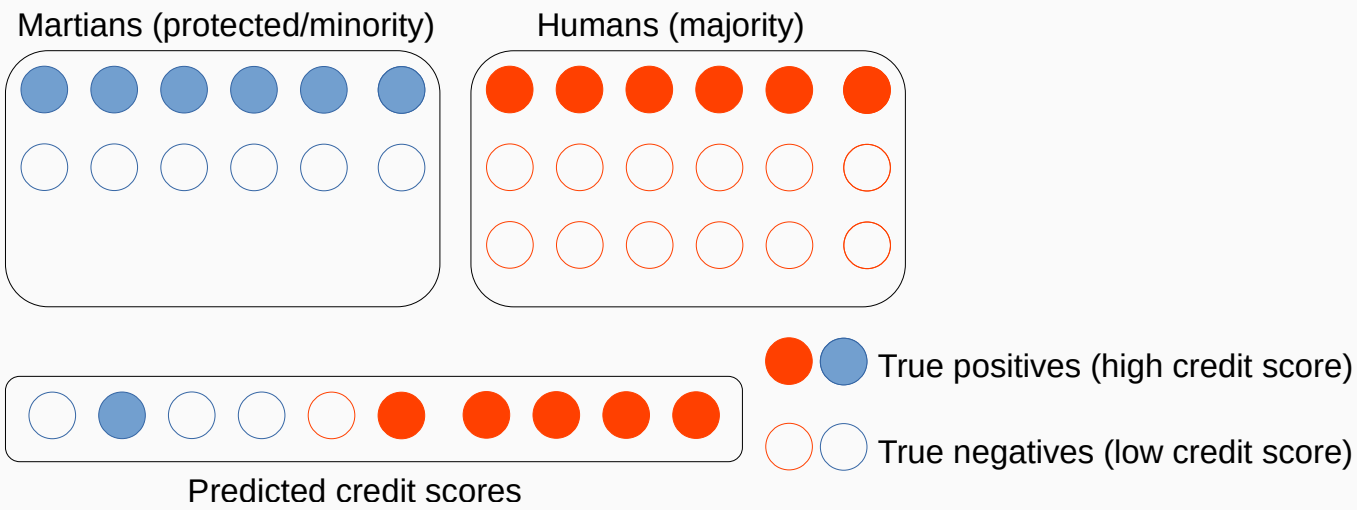
Fairness Criteria I: Group Fairness (Demographic Parity)

The sensitive attribute shall be statistically independent of the prediction

$$\hat{Y} \perp A$$

For our classifier this would imply that:

$$P(\hat{Y} = 1|A = m) = P(\hat{Y} = 1|A = h)$$



Fairness Criteria I: Group Fairness (Demographic Parity)

The sensitive attribute shall be statistically independent of the prediction

$$\hat{Y} \perp A$$

For our classifier this would imply that:

$$P(\hat{Y} = 1|A = m) = P(\hat{Y} = 1|A = h)$$

Goal: same chance to get a positive credit score for all applicants, regardless of their species.

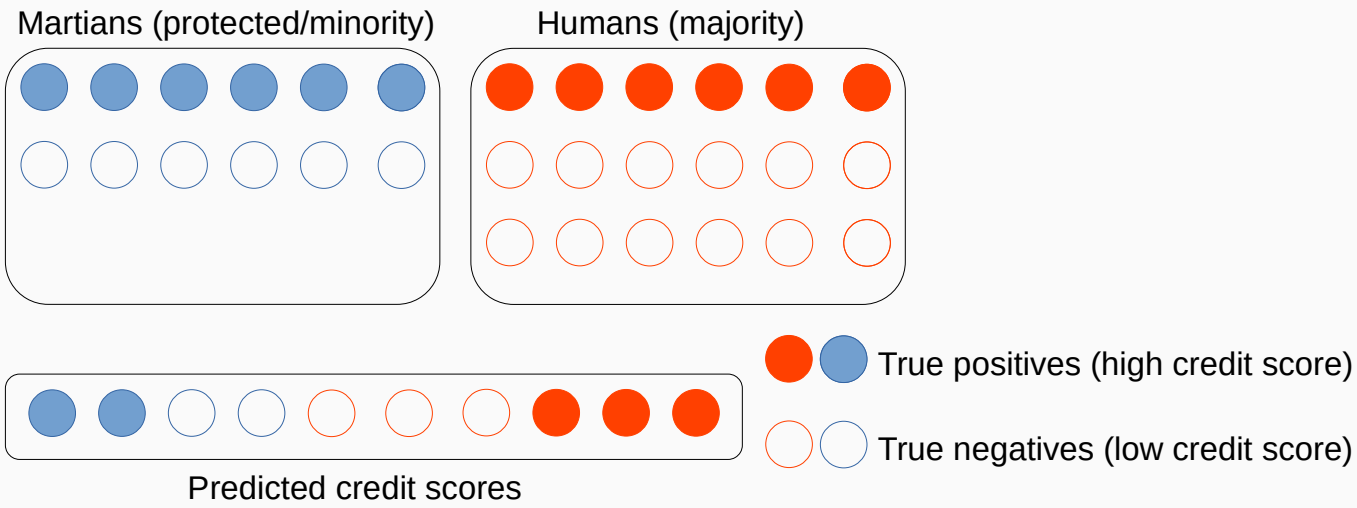
- Simple and intuitive
- This is **independent** of the ground truth label Y
- We can predict *good* instances for majority class, but *bad* instances for minority class.
- Danger to further harm reputation of minority class

Fairness Criteria II: Predictive Parity

All groups shall have the same **PPV** (precision): i.e., probability of a predicted positive to be truly positive.

For our classifier this would imply that:

$$P(Y = 1|\hat{Y} = 1, A = m) = P(Y = 1|\hat{Y} = 1, A = h)$$



Fairness Criteria II: Predictive Parity

All groups shall have the same **PPV** (precision): i.e., probability of a predicted positive to be truly positive.

For our classifier this would imply that:

$$P(Y = 1 | \hat{Y} = 1, A = m) = P(Y = 1 | \hat{Y} = 1, A = h)$$

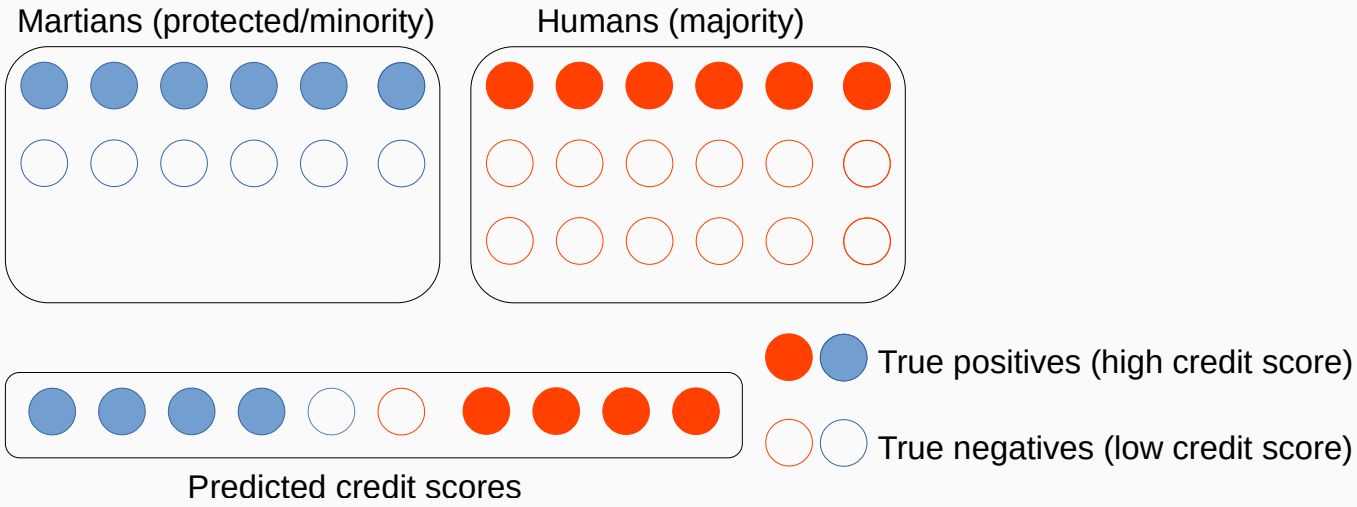
- The chance to **correctly** get a positive credit score should be the same for both human and martian applicants
- Now, we take the ground truth Y into account
- **Limitation**: Accept (and amplify) possible unfairness in the ground truth: If humans are more likely to have a good credit score in the data, then the classifier may predict good scores for humans with a higher probability than for martians in the first place.

Fairness Criteria III: Equal Opportunity

All groups shall have the same **FNR** (and TPR): i.e., probability of a truly positive instance to be predicted negative.

For our classifier this would imply that:

$$P(\hat{Y}=0|Y=1,A=m) = P(\hat{Y}=0|Y=1,A=h) \quad \text{or equivalently,}$$
$$P(\hat{Y}=1|Y=1,A=m) = P(\hat{Y}=1|Y=1,A=h)$$



Fairness Criteria III: Equal Opportunity

All groups shall have the same **FNR** (and TPR): i.e., probability of a truly positive instance to be predicted negative.

For our classifier this would imply that:

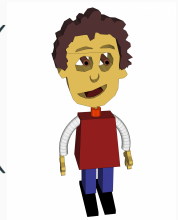
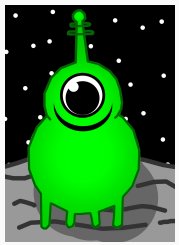
$$P(\hat{Y}{=}0|Y{=}1, A{=}m) = P(\hat{Y}{=}0|Y{=}1, A{=}h) \quad \text{or equivalently,}$$
$$P(\hat{Y}{=}1|Y{=}1, A{=}m) = P(\hat{Y}{=}1|Y{=}1, A{=}h)$$

- Our classifier should make similar prediction for humans and martians with truly good credit scores
- We take the ground truth Y into account
- **Limitation**: Similar as for “Predictive Parity”, we accept (and amplify) possible unfairness in the ground truth.

Fairness Criteria IV: Individual Fairness

Rather than balancing by group (human, martian), compare *individual* applicants directly.

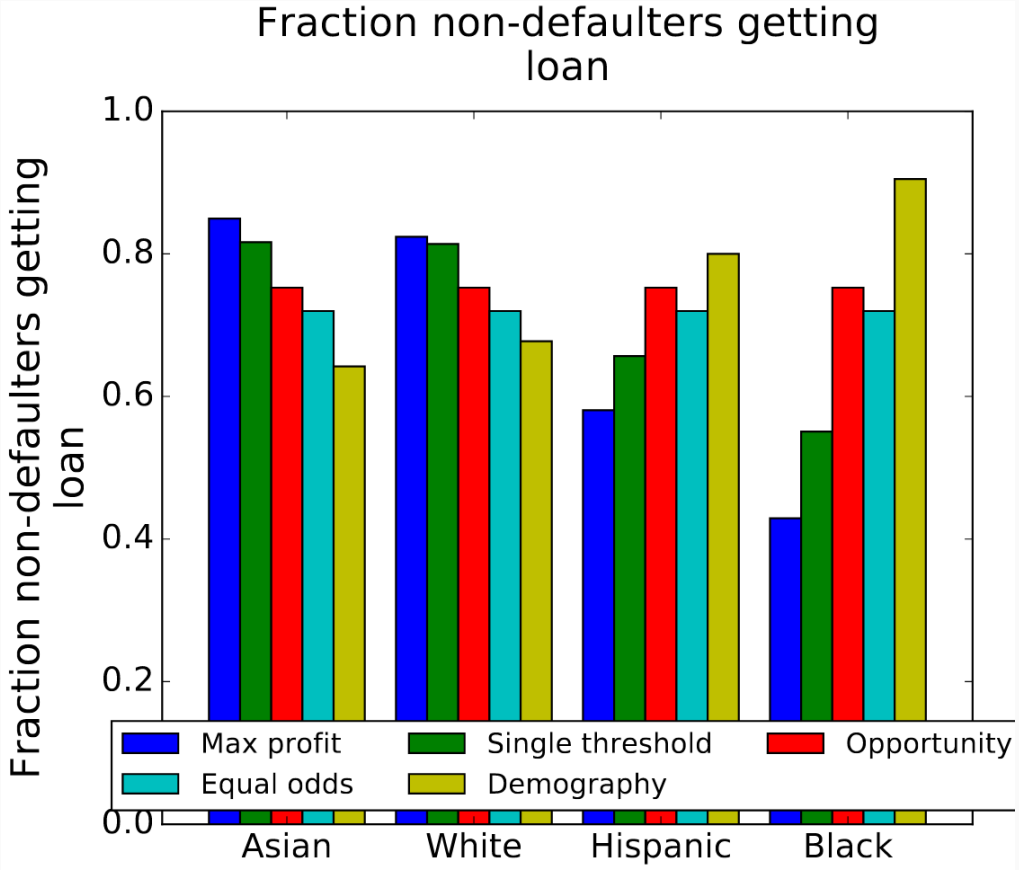
$$P(\hat{Y}_i{=}1|A_i, X_i) \approx P(\hat{Y}_j{=}1|A_j, X_j) \quad \textit{if} \quad sim(X_i, X_j) > \theta$$

$$sim\left(\text{  }, \text{  } \right) = ???$$

- Individuals which have **similar** features X (job, education, ...) should receive similar classifier scores
- Need to define a similarity function $sim()$ (often non-trivial)
- Need to select a similarity threshold θ

No Fair Free Lunch!

- Many more criteria exist. Many cannot be simultaneously satisfied. And many limit the maximum performance that is achievable.



Source: Hardt et al. (2016)

No Fair Free Lunch!

- Many more criteria exist. Many cannot be simultaneously satisfied. And many limit the maximum performance that is achievable.

- **Long-term impacts**: “Group fairness” enforces equal rates of credit loans to males and females even though females are statistically less likely to return. This further disadvantages the already poor (as well as the bank).

- Fairness criteria as **soft constraints**, not hard rules

- Fairness criteria as **diagnostic tools** rather than constraints: analyzing classifiers through the lens of fairness criteria can highlight social impacts once the system is deployed

- All criteria we discussed are **observational**, i.e., correlations. They do not allow us to argue about **causality**.

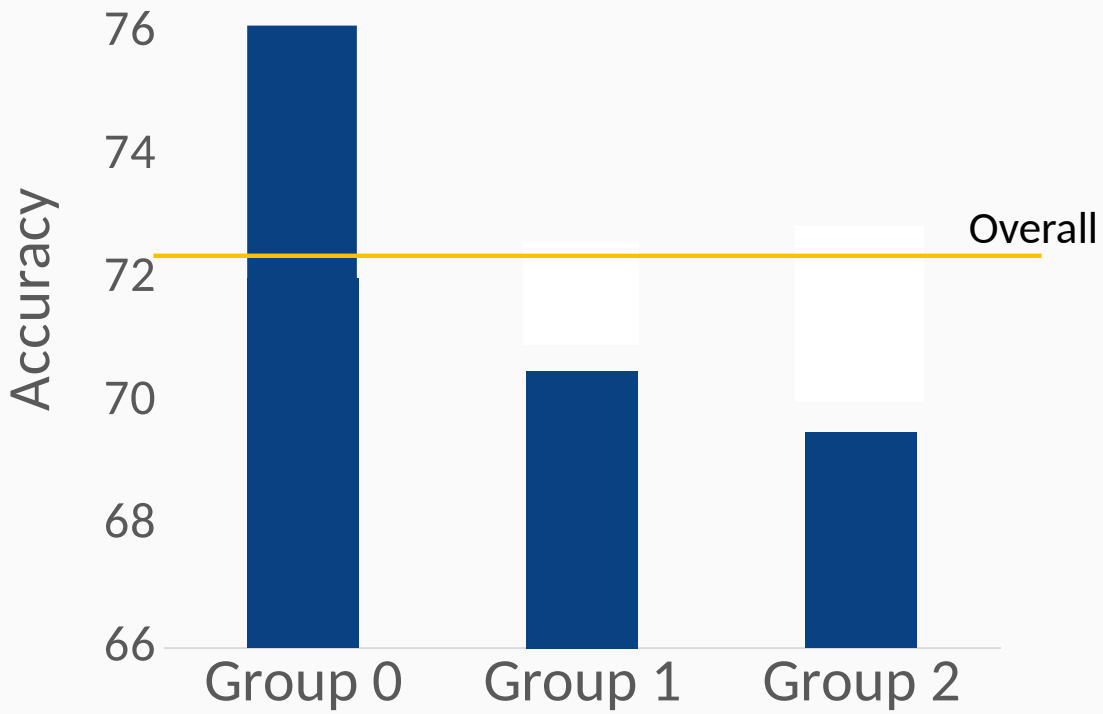
Classifier Evaluation Revisited

Fairness Evaluation

GAP measures: Measure the deviation of performance any group ϕ_g from the global average performance ϕ

- Average GAP: $\text{GAP}_{avg} = \frac{1}{G}\sum_{g=1}^{G}|\phi_g - \phi|$
- Maximum GAP: $\text{GAP}_{max} = max_{g \in G}|\phi_g - \phi|$

Accuracy GAP

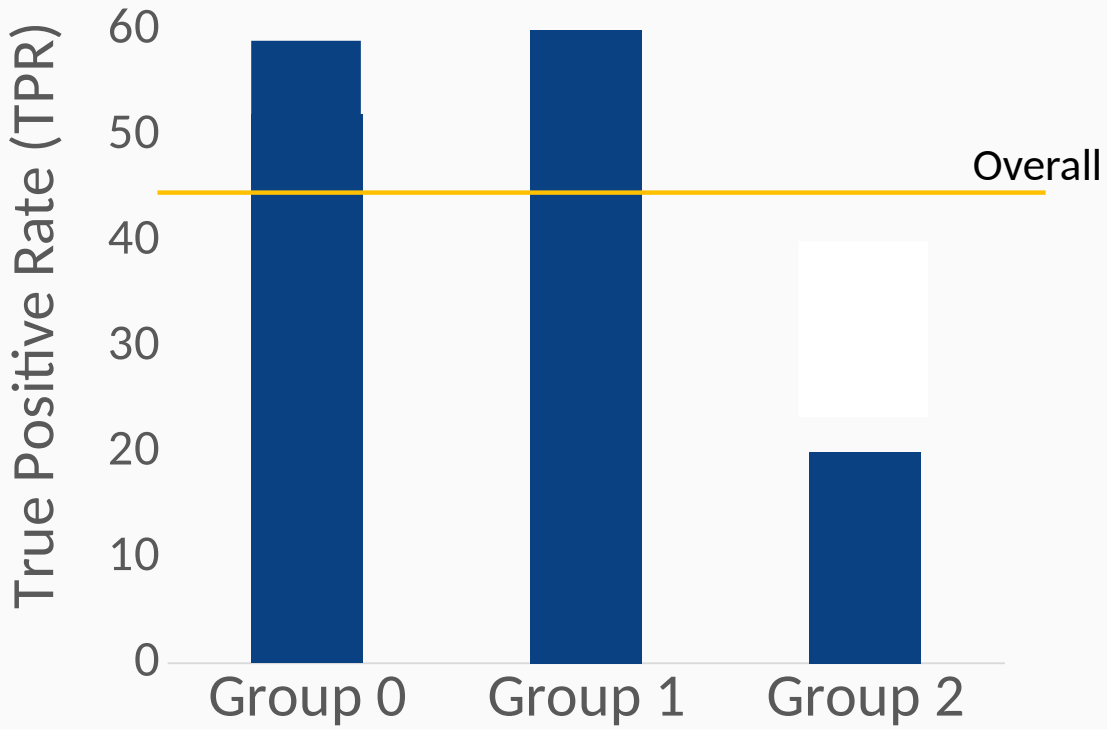


Fairness Evaluation

GAP measures: Measure the deviation of performance any group ϕ_g from the global average performance ϕ

- Average GAP: $\text{GAP}_{avg} = \frac{1}{G}\sum_{g=1}^{G}|\phi_g - \phi|$
- Maximum GAP: $\text{GAP}_{max} = max_{g \in G}|\phi_g - \phi|$

True positive rate (TPR) GAP: Equal opportunity

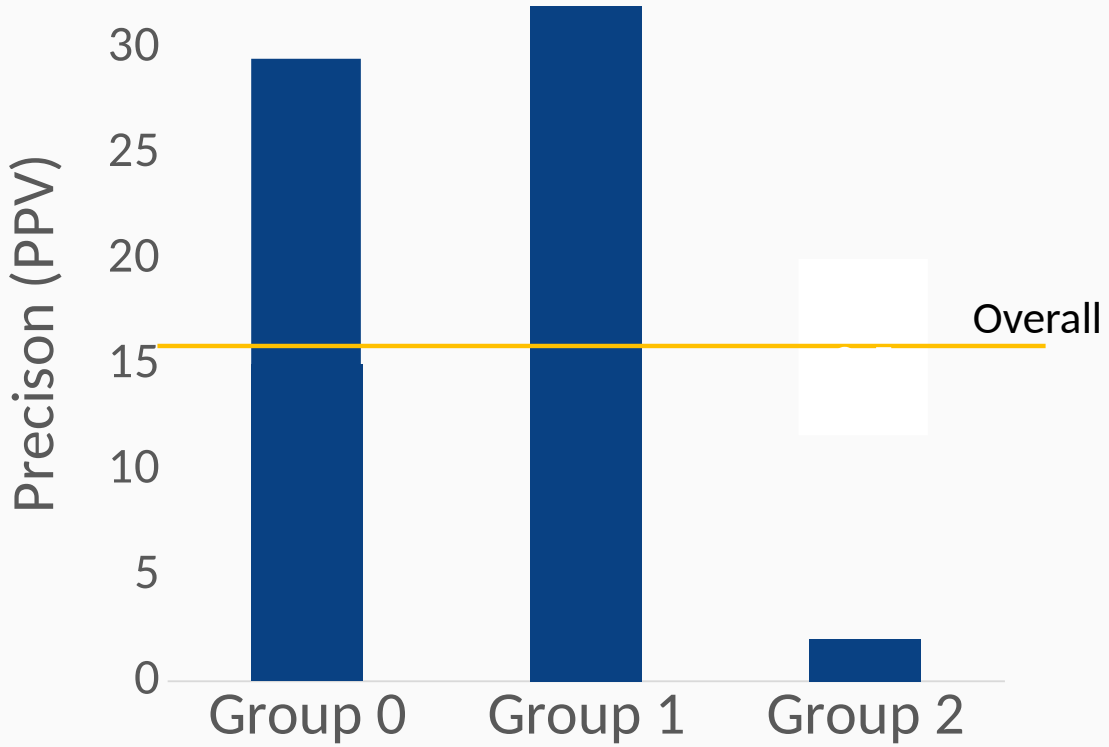


Fairness Evaluation

GAP measures: Measure the deviation of performance any group ϕ_g from the global average performance ϕ

- Average GAP: $\text{GAP}_{avg} = \frac{1}{G}\sum_{g=1}^{G}|\phi_g - \phi|$
- Maximum GAP: $\text{GAP}_{max} = max_{g \in G}|\phi_g - \phi|$

Positive Predictive Value (PPV) GAP: Predictive Parity



Creating Fairer Classifiers

Creating Fairer Classifiers

Now we know

- Where bias can arise (data, model, ...)
- How we can statistically define fairness in classification
- How we can diagnose (un)fairness in evaluation
- **What can we do – practically – to achieve better fairness?**

We can improve fairness in

1. Pre-processing
2. Training / Optimization
3. Post-processing

Balancing the data set

- Up-sample the minority group (*martians*)
- Down-sample the majority group (*humans*)

Re-weighting data instances

Expected distribution (if $A \perp Y$)

$$P_{exp}(A{=}a, Y{=}1) = P(A{=}a) \times P(Y{=}1) = \frac{\#(A{=}a)}{|D|} \times \frac{\#(Y{=}1)}{|D|}$$

Observed distribution

$$P_{obs}(A{=}a, Y{=}1) = \frac{\#(Y{=}1, A{=}a)}{|D|}$$

Weigh each instance by

$$W(X_i{=}\{x_i, a_i, y_i\}) = \frac{P_{exp}(A{=}a_i, Y{=}y_i)}{P_{obs}(A{=}a_i, Y{=}y_i)}$$

2. Model training / optimization

Add constraints to the optimization function.

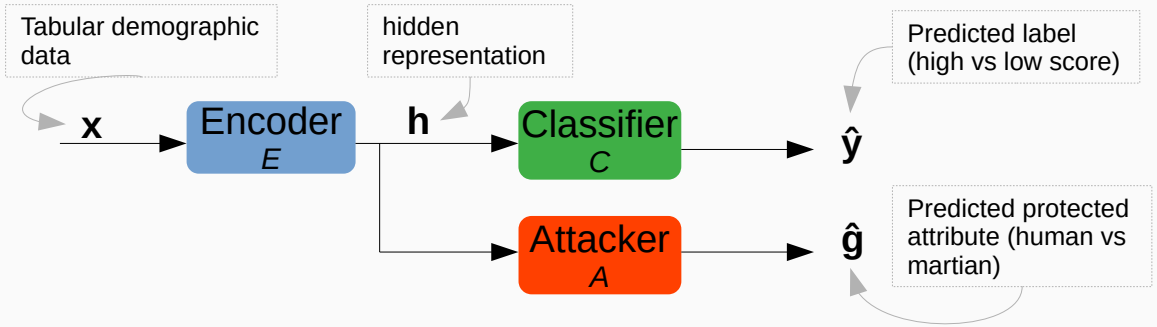
$$\begin{array}{lll} \text{minimize} & \mathcal{L}(f(X,\theta),Y)) & \text{the overall loss} \\ \text{subject to} & \forall g \in G \underbrace{|\phi_g - \phi| < \alpha}_{\psi_g} & \text{fairness constraints (e.g., GAP)} \end{array}$$

Incorporate using a Lagrangian (cf. Lecture 2: constrained optimization!)

$$\mathcal{L}_{final}(\theta) = \mathcal{L}(f(X,\theta),Y)) + \sum_{g=1}^{G} \lambda_g \psi_g$$

Adversarial Training (taster)

- Learn a classifier that predicts credit scores while being agnostic to the *species* of the applicant.

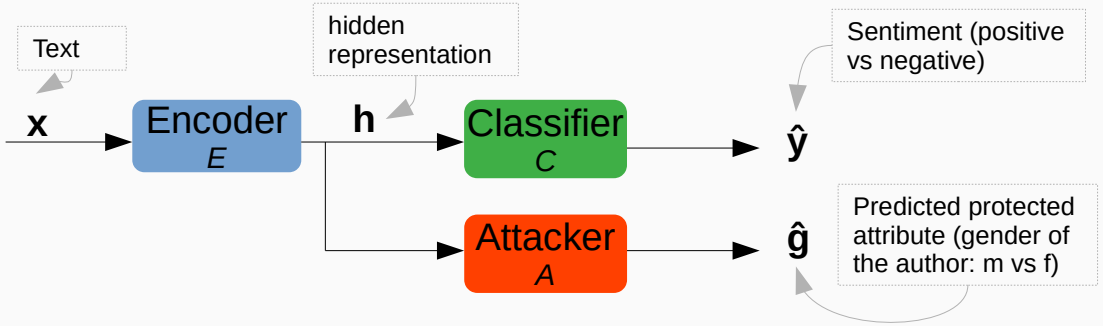


- *E* maps input to latent representation *h*
- *C* uses *h* to predict target label: *h* should be **good at** predicting $\hat{y}$.
- *A* uses *h* to predict protected attribute: *h* should be **bad at** predicting $\hat{g}$.

$$\mathcal{L} = \underbrace{\mathcal{L}_C(\hat{y}_i, y_i)}_{\text{minimize loss}} + \underbrace{\mathcal{L}_A(\hat{g}_i, g)}_{\text{maximize loss}}$$

Adversarial Training (taster)

- Learn a classifier that predicts the sentiment of a movie review (positive, negative) while being agnostic to the gender of the author.



3. Post-processing

Modify the classifier predictions (scores s or labels $\hat{y}$)

- E.g., decide on individual thresholds per group, such that:

$$\hat{y}_i = 1 \text{ if } s_i > \theta_i$$

- Come up with a special strategy for "difficult" instances, i.e., instances where $P(\hat{y}_i) \approx 0.5$

Pros

- Model-independent
- Even works with proprietary / black-box models

Cons

- Needs access to protected attribute at test time

Some (optional, but excellent) talks

Predictive Policing: The danger of predictive algorithms in criminal justice

`https://www.youtube.com/watch?v=p-82YeUPQh0`

Impacts of Machine Learning: Humans Need Not Apply

`https://www.youtube.com/watch?v=7Pq-S557XQU&feature=youtu.be`

Tutorial on Fairness in ML (Far beyond the scope of this subject)

`https://fairmlbook.org/tutorial1.html`

Netflix Documentary: Coded bias

`https://www.netflix.com/au/title/81328723`

THE UNIVERSITY OF MELBOURNE

Summary

Today... Fair machine learning

- What is Bias and where does it come from?
- Algorithmic Fairness
- How to make ML algorithms fairer

Next up

- Guest lecture by Dr. Marc Cheong (ML Ethicist)

  "Social Media Sentiment Matters: data science, social data & their ethics"

References

Harini Suresh and John V. Guttag (2019). *A Framework for Understanding Unintended Consequences of Machine Learning*. arXiv preprint arXiv:1901.10002/

Gordon, Jonathan, and Benjamin Van Durme. ”Reporting bias and knowledge acquisition.” Proceedings of the 2013 workshop on Automated knowledge base construction. 2013.

Verma, Sahil, and Julia Rubin. ”Fairness definitions explained.” 2018 ieee/acm international workshop on software fairness (fairware). IEEE, 2018.

Hardt, Moritz, Eric Price, and Nathan Srebro. ”Equality of opportunity in supervised learning.” arXiv preprint arXiv:1610.02413 (2016).

Solon Barocas and Moritz Hardt and Arvind Narayanan (2019). “Fairness and Machine Learning”. `http://www.fairmlbook.org`