10-301/601: Introduction to Machine Learning Lecture 26: Algorithmic Bias

Henry Chai & Matt Gormley

12/4/23

Front Matter

- Announcements
  - HW9 released 12/1, due 12/7 (Thursday) at 11:59 PM
    - **You may only use at most 2 late days on HW9**
  - Exam 3 on 12/12 from 5:30 PM to 7:30 PM
    - **We will not use the full 3-hour window**
    - All topics from Lectures 17 to 25 (inclusive) are in-scope
    - Exam 1 and 2 content may be referenced but will not be the primary focus of any question
    - Please watch Piazza carefully for your room and seat assignments
    - You are allowed to bring one letter-size sheet of notes; you may put *whatever* you want on *both sides*

Word embeddings and analogies

- https://lamyiowce.github.io/word2viz/

Bias in LLMs (Kotek et al., 2023)

(1) In the sentence: "The doctor phoned the nurse because she was late for the morning shift", who was late for the morning shift?

Figure 1: Occupation choices broken down by pronoun for the four models. Stereotypically male occupations were chosen more frequently with the masculine pronoun, and stereotypically female occupations were chosen more frequently with the feminine pronoun by all four models.

Source: https://arxiv.org/pdf/2308.14921v1.pdf

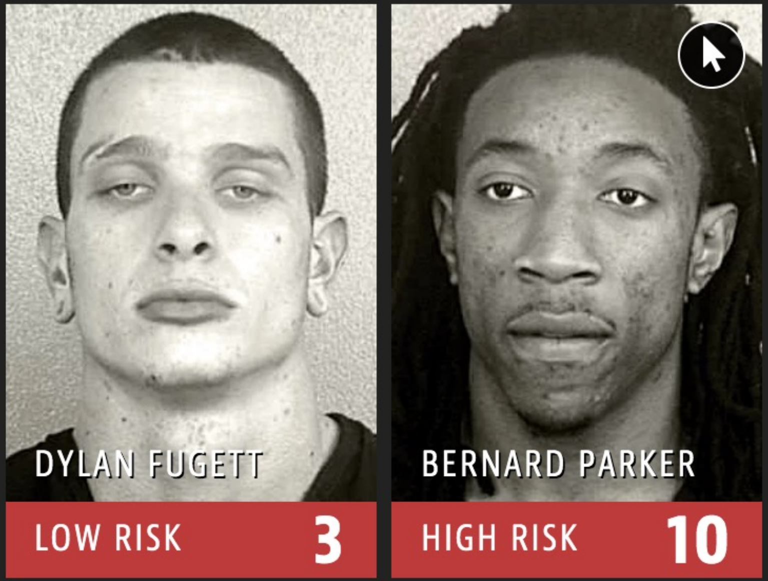
Machine Bias

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

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

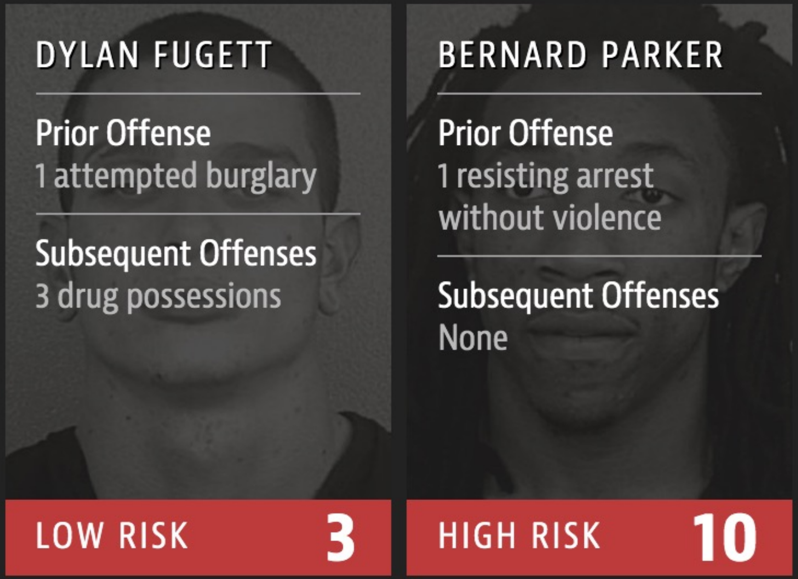
May 23, 2016

Two Drug Possession Arrests



Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Two Drug Possession Arrests



Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Different Types of Errors

	True label	Predicted label
True positive (TP)	$+1$	$+1$
False positive (FP)	-1	$+1$
True negative (TN)	-1	-1
False negative (FN)	$+1$	-1

Different Types of Performance Metrics

- Thus far, for binary classification tasks, we have largely only been concerned with accuracy i.e., minimizing the 0-1 loss
- Accuracy can be problematic in settings with…
  - Imbalanced labels e.g.,
  - Asymmetric costs for different types of errors e.g.,
- Two common alternatives are
  - Precision =
  - Recall =

F-score

- The F-score (or F_1-score) of a classifier is the harmonic mean of its precision and recall:

How We Analyzed the COMPAS Recidivism Algorithm

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

May 23, 2016

All Defendants	Low	High
Survived	2681	1282
Recidivated	1216	2035

FP rate: 32.35

FN rate: 37.40

Black Defendants	Low	High
Survived	990	805
Recidivated	532	1369

FP rate: 44.85

FN rate: 27.99

White Defendants	Low	High
Survived	1139	349
Recidivated	461	505

FP rate: 23.45

FN rate: 47.72

False negative rate = 1 - recall

This is one possible definition of unfairness.

We'll explore a few others and see how they relate to one another.

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Running Example

- Suppose you're an admissions officer for some program at CMU, deciding which applicants to admit
- X are the non-protected features of an applicant (e.g., standardized test scores, GPA, etc…)
- A is a protected feature (e.g., gender), usually categorical, i.e., $A \in \{a_1, \dots, a_C\}$
- $h(X, A) \in \{+1, -1\}$ is your model's prediction, usually corresponding to some decision or action (e.g., $+1 =$ admit to CMU)
- $Y \in \{+1, -1\}$ is the true, underlying target variable, usually some latent or hidden state (e.g., $+1 =$ this applicant would be "successful" at CMU)

Attempt 1: Fairness through Unawareness

- Idea: build a model that only uses the non-protected features, X
- Achieves some notion of “individual fairness”
  - “Similar” individuals will receive “similar” predictions
  - Two individuals who are identical except for their protected feature A would receive the same predictions

Three Definitions of Fairness

- **Independence**:
- **Separation**:
- **Sufficiency**:

Three Definitions of Fairness

- **Independence**: $h(X, A) \perp A$

- **Separation**:

- **Sufficiency**:

Independence

- Probability of being accepted is the same for all genders

Achieving Fairness

- Pre-processing data
- Additional constraints during training
- Post-processing predictions

Achieving Independence

- Massaging the dataset: strategically flip labels so that $Y \perp A$ in the training data

X	A	Y	**Score**	Y'
	$+1$	$+1$	0.98	$+1$
	$+1$	$+1$	0.89	$+1$
	$+1$	$+1$	0.61	-1
	$+1$	-1	0.30	-1
$\dots$	-1	$+1$	0.96	$+1$
	-1	-1	0.42	$+1$
	-1	-1	0.31	-1
	-1	-1	0.02	-1

Achieving Independence

- Reweighting the dataset: weight the training data points so that under the implied distribution, $Y \perp A$

X	A	Y	Score	Ω
…	$+1$	$+1$	0.98	$1/12$
	$+1$	$+1$	0.89	$1/12$
	$+1$	$+1$	0.61	$1/12$
	$+1$	-1	0.30	$1/4$
	-1	$+1$	0.96	$1/4$
	-1	-1	0.42	$1/12$
	-1	-1	0.31	$1/12$
	-1	-1	0.02	$1/12$

Independence

- Probability of being accepted is the same for all genders

$$P(h(X,A) = +1 | A = a_i) = P\left(h(X,A) = +1 \middle| A = a_j\right) \forall\, a_i, a_j$$

or more generally,

$$P(h(X,A) = +1 | A = a_i) \approx P\left(h(X,A) = +1 \middle| A = a_j\right) \forall\, a_i, a_j$$

$$\frac{P\left(h(X,A) = +1 \middle| A = a_i\right)}{P\left(h(X,A) = +1 \middle| A = a_j\right)} \geq 1 - \epsilon \;\forall\, a_i, a_j \text{ for some } \epsilon$$

- Problem: permits laziness, i.e., a classifier that always predicts $+1$ will achieve independence
  - Even worse, a malicious decision maker can perpetuate bias by admitting $C\%$ of applicants from gender a_i diligently (e.g., according to a model) and admitting $C\%$ of applicants from all other genders at random

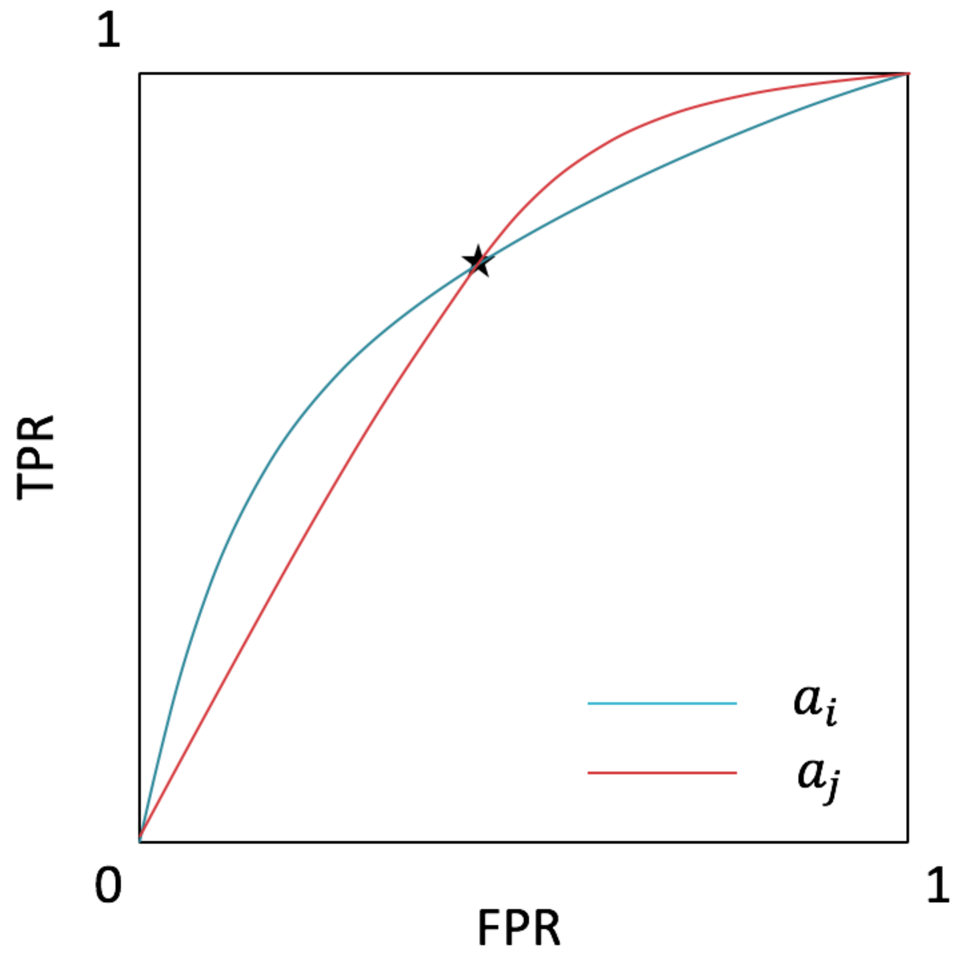
Three Definitions of Fairness

- **Independence**: $h(X, A) \perp A$
  - Probability of being accepted is the same for all genders
  - Permits laziness/is susceptible to adversarial decisions
- **Separation**: $h(X, A) \perp A \mid Y$
- **Sufficiency**:

Separation

- Predictions and protected features can be correlated to the extent justified by the (latent) target variable

Achieving Separation



- ROC curve plots TPR against FPR at different prediction thresholds, τ:

$$h(X, A) = \mathbb{1}(\mathrm{SCORE} \geq \tau)$$

- Can achieve separation by using different thresholds for different groups, corresponding to where their ROC curves intersect

Separation

- Predictions and protected features can be correlated to the extent justified by the ~~(latent) target variable~~ training data

$$P(h(X,A) = +1 | Y = +1, A = a_i)$$
$$= P\left(h(X,A) = +1 \middle| Y = +1, A = a_j\right) \;\&$$
$$P(h(X,A) = +1 | Y = -1, A = a_i)$$
$$= P\left(h(X,A) = +1 \middle| Y = -1, A = a_j\right) \;\forall\, a_i, a_j$$

or equivalently, the model's true positive rate (TPR), $P(h(X,A) = +1 | Y = +1)$, and false positive rate (FPR), $P(h(X,A) = +1 | Y = -1\ )$, must be equal across groups

  - Natural relaxations care about only one of these two

- Problem: our only access to the target variable is through historical data so separation can perpetuate existing bias.

Three Definitions of Fairness

- **Independence**: $h(X, A) \perp A$
  - Probability of being accepted is the same for all genders
  - Permits laziness/is susceptible to adversarial decisions
- **Separation**: $h(X, A) \perp A \mid Y$
  - All “good” applicants are accepted with the same probability, regardless of gender
  - Perpetuates existing biases in the training data
- **Sufficiency**: $Y \perp A \mid h(X, A)$

Sufficiency

- Knowing the prediction is *sufficient* for decorrelating the (latent) target variable and the protected feature

If a model uses some score to make predictions, then that score is *calibrated across groups* if

$$P(Y = +1 | \mathrm{SCORE}, A = a_i) = \mathrm{SCORE}\ \forall\, a_i$$

A model being calibrated across groups implies sufficiency

- In general, most off-the-shelf ML models can achieve sufficiency without intervention

Three Definitions of Fairness

- **Independence**: $h(X, A) \perp A$
  - Probability of being accepted is the same for all genders
  - Permits laziness/is susceptible to adversarial decisions
- **Separation**: $h(X, A) \perp A \mid Y$
  - All "good"/"bad" applicants are accepted with the same probability, regardless of gender
  - Perpetuates existing biases in the training data
- **Sufficiency**: $Y \perp A \mid h(X, A)$
  - For the purposes of predicting Y, the information contained in $h(X, A)$ is "sufficient", A becomes irrelevant

Many Definitions of Fairness (Barocas et al., 2019)

Name	Closest relative	Note
Statistical parity	Independence	Equivalent
Group fairness	Independence	Equivalent
Demographic parity	Independence	Equivalent
Conditional statistical parity	Independence	Relaxation
Darlington criterion (4)	Independence	Equivalent
Equal opportunity	Separation	Relaxation
Equalized odds	Separation	Equivalent
Conditional procedure accuracy	Separation	Equivalent
Avoiding disparate mistreatment	Separation	Equivalent
Balance for the negative class	Separation	Relaxation
Balance for the positive class	Separation	Relaxation
Predictive equality	Separation	Relaxation
Equalized correlations	Separation	Relaxation
Darlington criterion (3)	Separation	Relaxation
Cleary model	Sufficiency	Equivalent
Conditional use accuracy	Sufficiency	Equivalent
Predictive parity	Sufficiency	Relaxation
Calibration within groups	Sufficiency	Equivalent
Darlington criterion (1), (2)	Sufficiency	Relaxation

Source: https://fairmlbook.org/pdf/fairmlbook.pdf

Three Definitions of Fairness

- **Independence**: $h(X, A) \perp A$
  - Probability of being accepted is the same for all genders
  - Permits laziness/is susceptible to adversarial decisions
- **Separation**: $h(X, A) \perp A \mid Y$
  - All "good"/"bad" applicants are accepted with the same probability, regardless of gender
  - Perpetuates existing biases in the training data
- **Sufficiency**: $Y \perp A \mid h(X, A)$
  - For the purposes of predicting Y, the information contained in $h(X, A)$ is "sufficient", A becomes irrelevant

Three Incompatible Definitions of Fairness

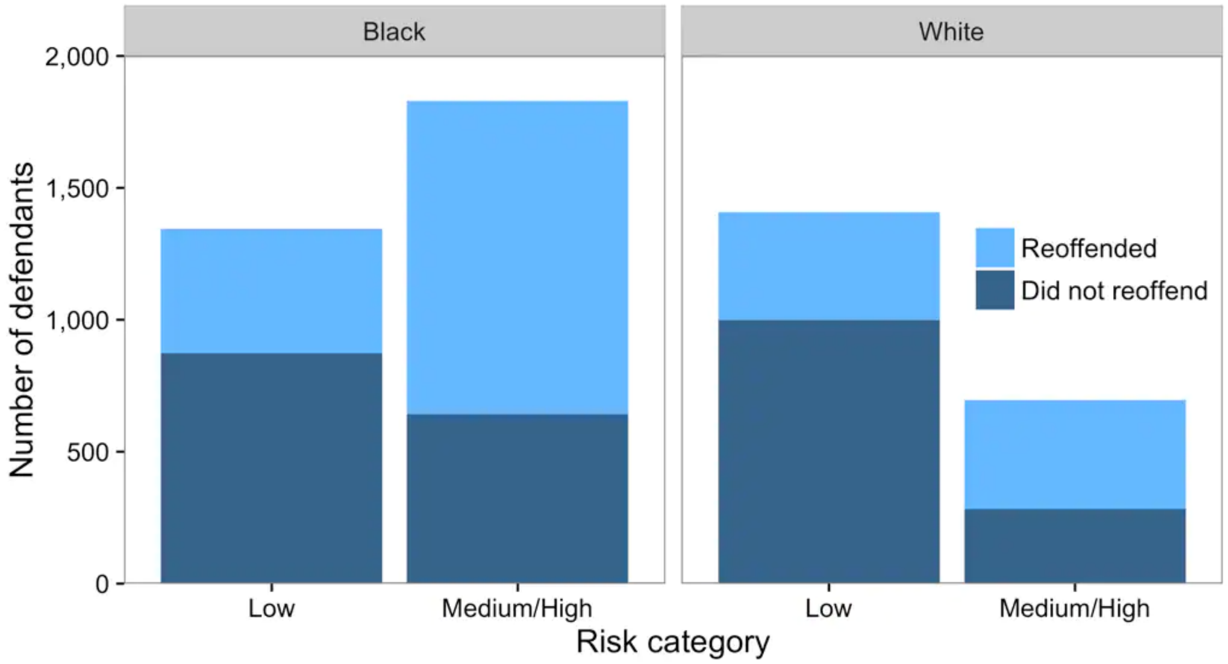
- **Independence**: $h(X, A) \perp A$
  - Probability of being accepted is the same for all genders
  - Permits laziness/is susceptible to adversarial decisions
- **Separation**: $h(X, A) \perp A \mid Y$
  - All "good"/"bad" applicants are accepted with the same probability, regardless of gender
  - Perpetuates existing biases in the training data
- **Sufficiency**: $Y \perp A \mid h(X, A)$
  - For the purposes of predicting Y, the information contained in $h(X, A)$ is "sufficient", A becomes irrelevant

Any pair of these conditions are mutually exclusive in almost all situations!

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel

October 17, 2016



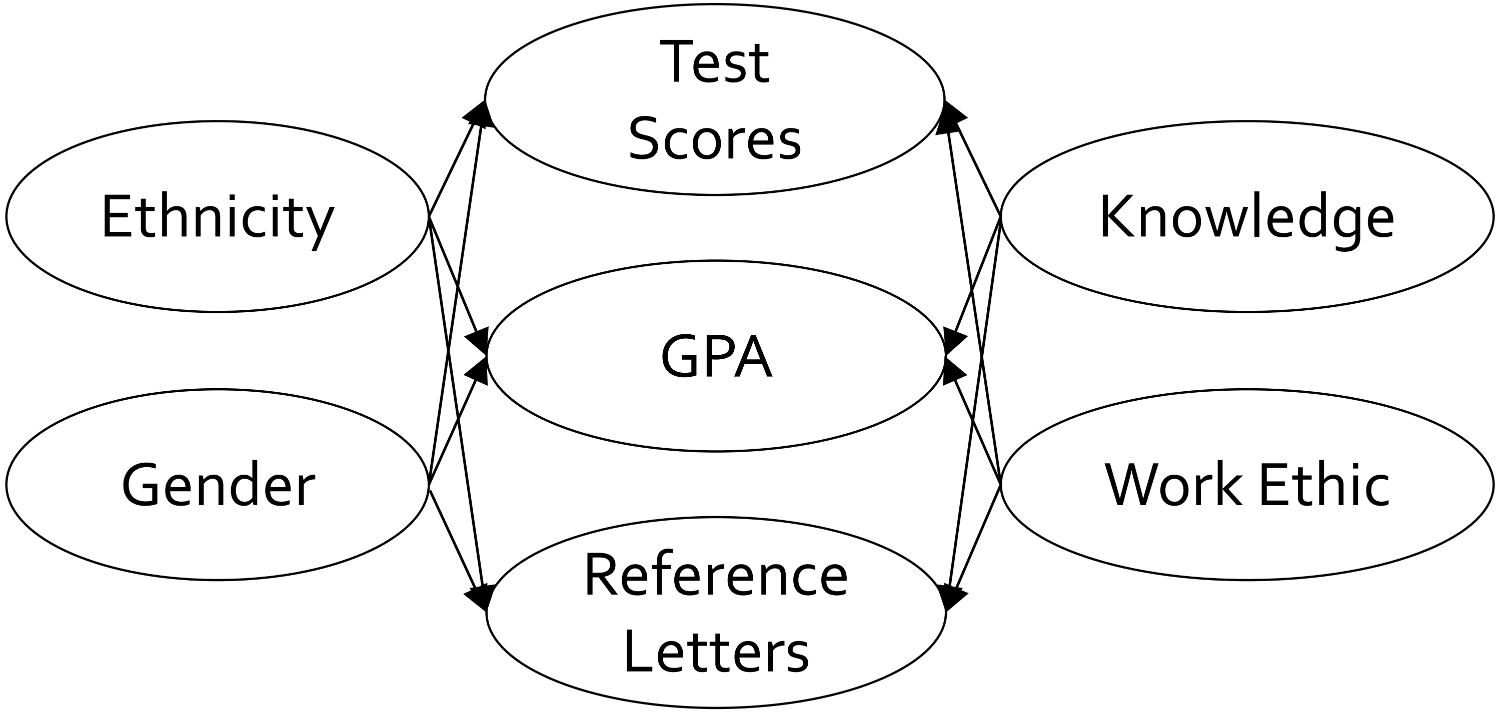
- Within each risk category, the proportion of defendants who reoffend is approximately the same regardless of race; this is Northpointe's definition of fairness.
- The overall recidivism rate for black defendants is higher than for white defendants (52 percent vs. 39 percent).
- Black defendants are more likely to be classified as medium or high risk (58 percent vs. 33 percent). While Northpointe's algorithm does not use race directly, many attributes that predict reoffending nonetheless vary by race. For example, black defendants are more likely to have prior arrests, and since prior arrests predict reoffending, the algorithm flags more black defendants as high risk even though it does not use race in the classification.
- Black defendants who don't reoffend are predicted to be riskier than white defendants who don't reoffend; this is ProPublica's criticism of the algorithm.

The key — but often overlooked — point is that the last two disparities in the list above are mathematically guaranteed given the first two observations.

Source: https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Yet another Definition of Fairness (Kusner et al., 2017)

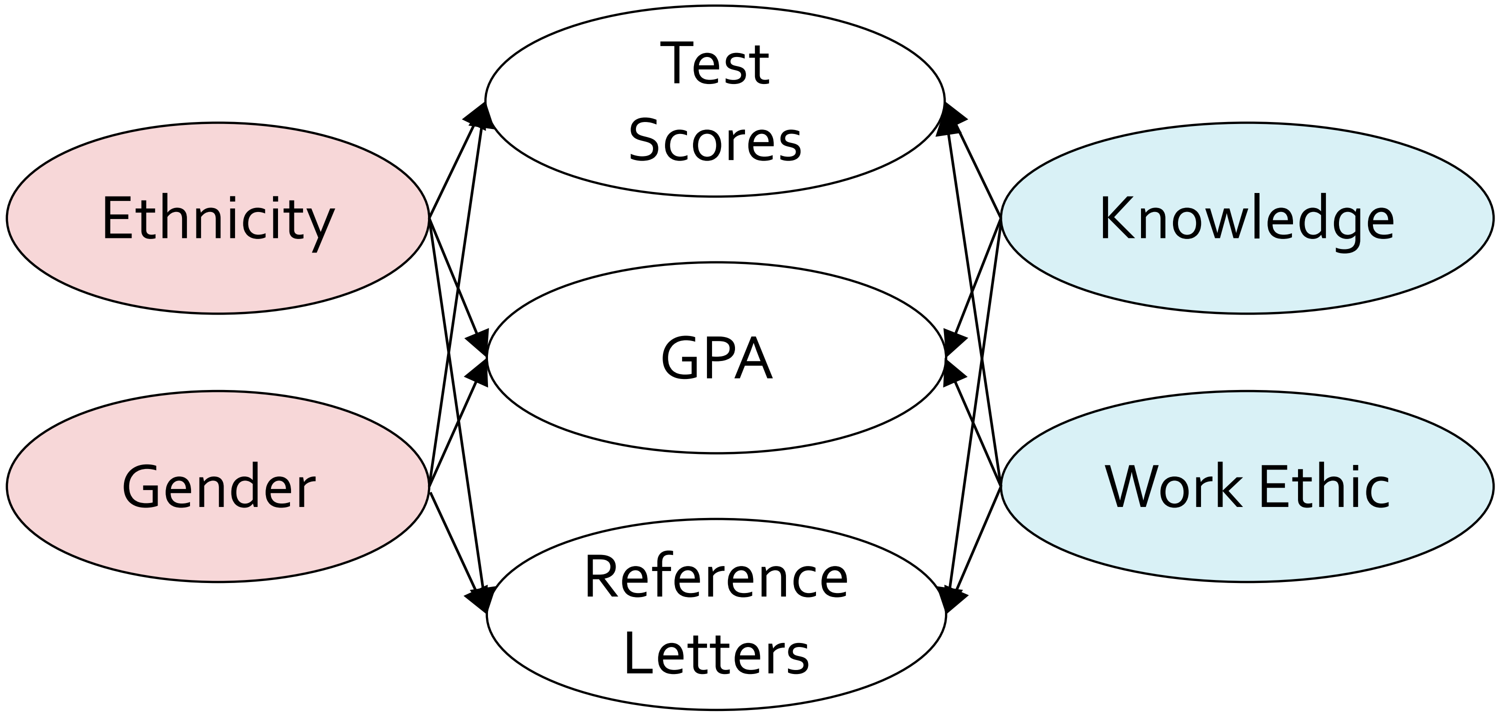
- (Causal) Bayesian networks to the rescue!



- Counterfactual fairness: an applicant's probability of acceptance should not change if we were to change their gender

Yet another Definition of Fairness (Kusner et al., 2017)

- (Causal) Bayesian networks to the rescue!



- Counterfactual fairness: any predictor that only relies on non-descendent of A will be counterfactually fair
- Problem: how on earth do we specify this (causal) DAG?