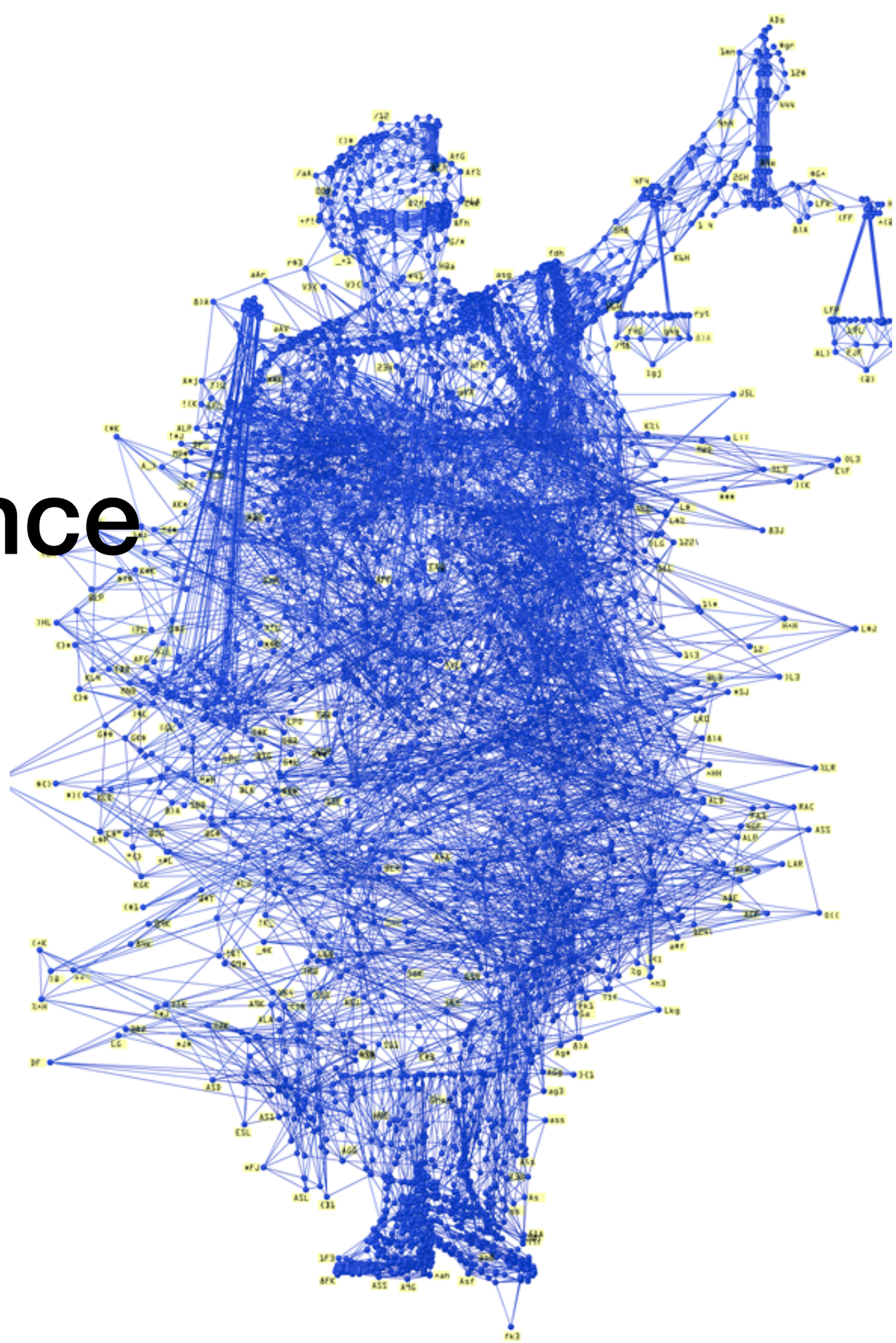
Lecture 10-1

Making Artificial Intelligence Fair



13-Oct-2020

Dr. Atoosa Kasirzadeh

- We want
    - AI predictions and recommendations to be **fair**
    - AI algorithms be **transparent & accountable**
    - AI algorithms respect our **privacy**

- In sum, we want ML algorithms **be ethical**

Let AI systems learn about ethical behaviour

Force AI researchers follow ethical codes

Align AI systems with human values

Code ethical theories in AI systems

What do we mean by Ethics of AI?

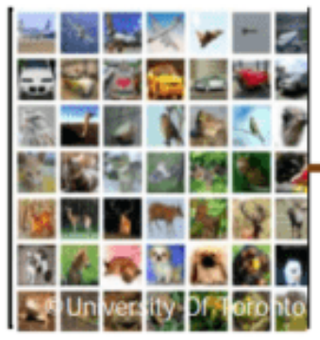
Today: Define some ethical principles (fairness) and operationalise them into AI systems

Resolve the problem of the centralised power (surveillance, privacy, power held by a few mega corporations)

Regulate AI (example: General Data Protection Regulation)

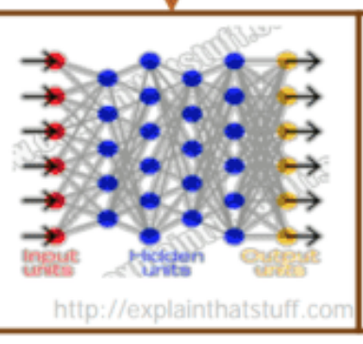
Remainder discussion from last week: trust and the explainable AI

Today



Training Data

Learning Process



Learned Function

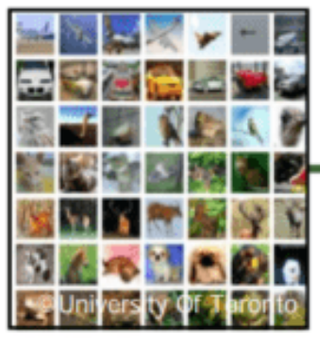
This is a cat
(p = .93)

Output

User with a Task

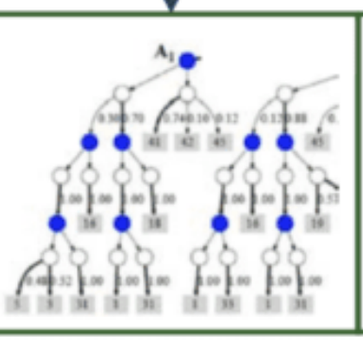
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Tomorrow



Training Data

New Learning Process



Explainable Model

This is a cat:
- It has fur, whiskers, and claws.
- It has this feature:

Explanation Interface

User with a Task

- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

“panda”

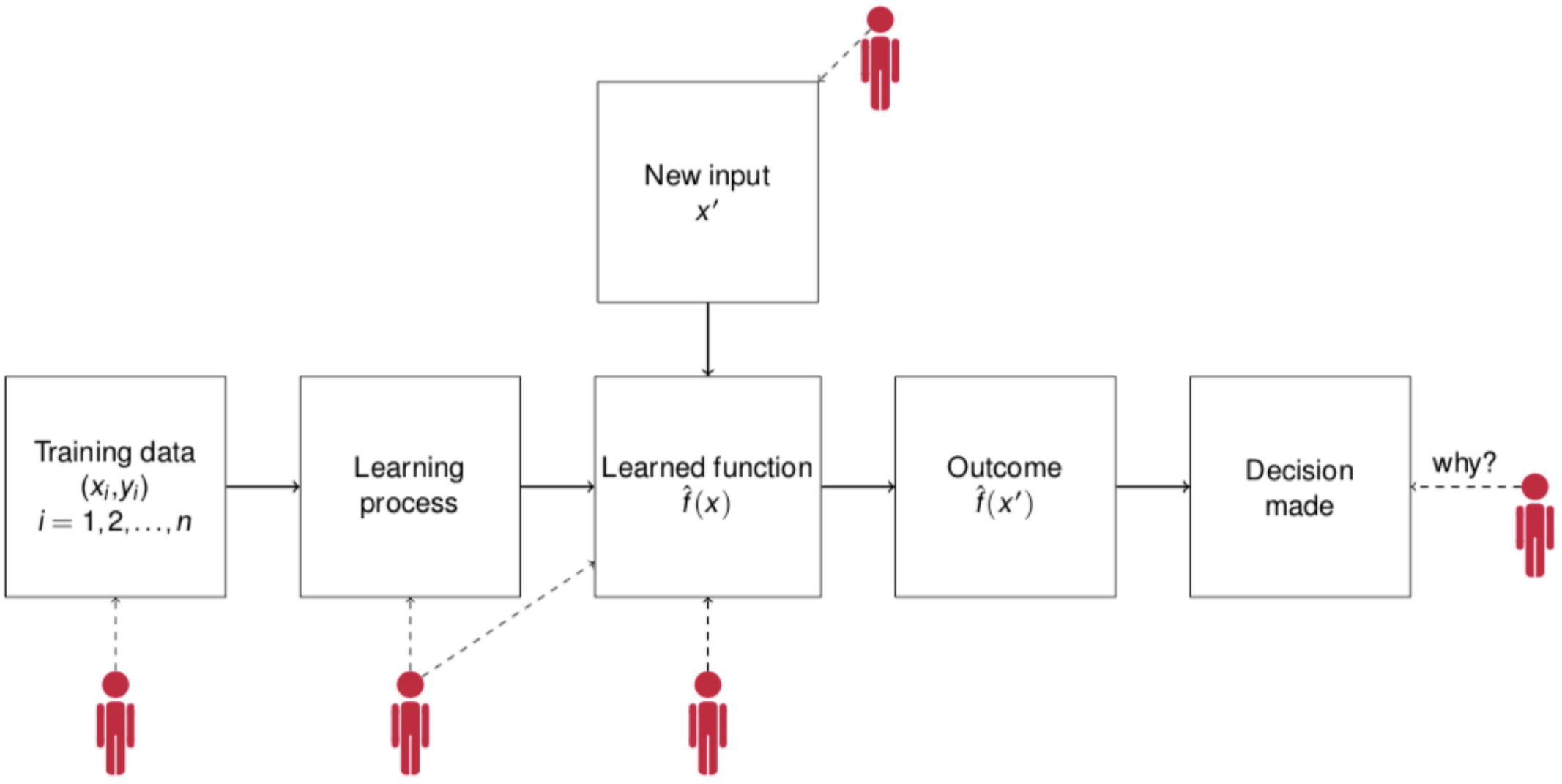
81.97% confidence

$+ 0.01 \times$

$=$

“papillon dog”

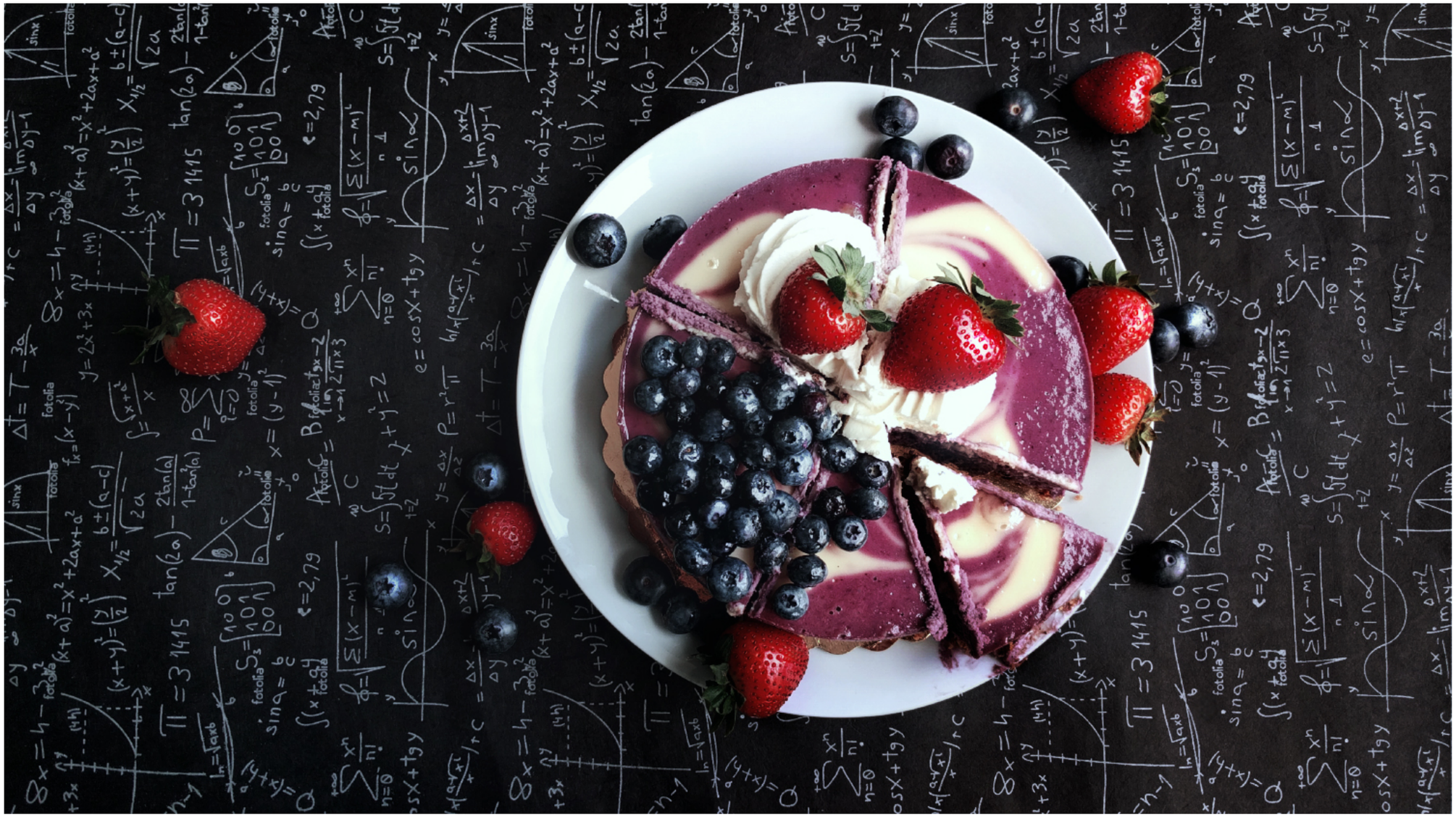
99.56% confidence


New input
x'
Training data
(x_i, y_i)
$i = 1, 2, \ldots, n$
Learning
process
Learned function
$\hat{f}(x)$
Outcome
$\hat{f}(x')$
Decision
made
why?

Today's lecture:

Fairness as a way to make Machine Learning algorithms ethical

What is a fair decision? Let's start with a non-algorithmic example



- Imagine you have the task of dividing this cake between a group of 10 people, including yourself. How would you divide it such that the **distribution** of the cake pieces is **fair**?

- Is there a unique answer?

Who gets care? Given that it is practically impossible for everyone to get care, what is a fair decision for distribution of medical services?

Distribution of goods among a set of individuals and fairness criteria

- Lessons from economics and social choice theory (utility the individuals obtain; allocation function)

- Design different set of conditions

    - Pareto efficiency: when the circumstances of one individual cannot be made better without making the situation worse for another individual.

    - Envy-freeness: every agent feels that their share is at least as good as the share of any other agent, and thus no agent feels envy.

    - Proportionality: a division of a resource among n partners such that each partner receives a part worth for her at least 1/n of the whole.

    - Equitability: equal life chances regardless of identity, to provide all citizens with a basic and equal minimum of income, goods, and services

- What is fairness?
    - Equality: everyone should have an equal opportunity or outcome or …

But.. people do not start from the same starting point

    - Maximin principle: maximize the welfare of those at the minimum level of society.

- Procedural fairness (looks at the procedure by which we've arrived at the decision)
- Substantive fairness (looks at the outcome)
- Many other characterizations given by philosophers, social scientists, psychologists, economists, etc
- Fairness is morally and legally motivated

- Population is diverse: ethnic, religious, geographic, medical, class, sexual preferences, etc

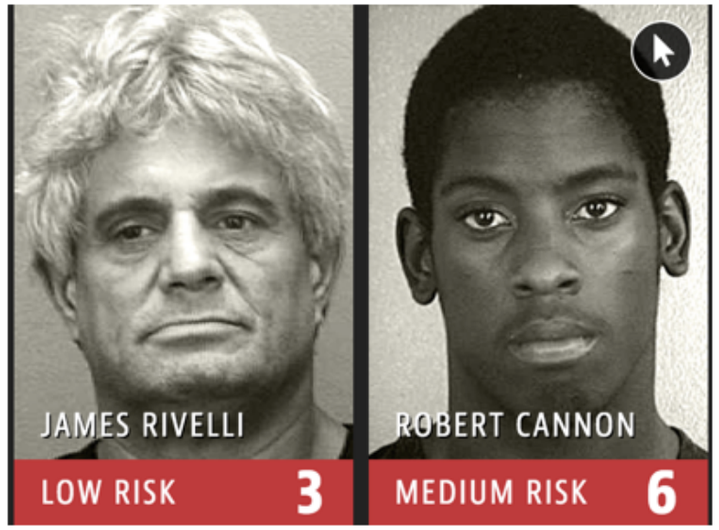
- In many classification tasks, some features implicitly or explicitly encode certain characteristics of an individual

Why algorithmic fairness?

- A motivating example is membership in a racial minority in the context of banking
- An article in The Wall Street Journal (2010) describes the practices of a credit card company and its use of a tracking network to learn detailed demographic information about each visitor to the site, such as approximate income, where she shops, the fact that she rents children's videos, and so on
- According to the article, this information is used to "decide which credit cards to show first-time visitors" to the web site, raising the concern of steering, namely the (illegal) practice of guiding members of minority groups into less advantageous credit offerings." (Dwork et al., 2012)

COMPAS & Criminal justice

- COMPAS: an algorithm used in US criminal justice system to predict whether criminals will re-offend
- Basic operation: assign a level of risk to each defendant
- Predicting if a defendant should receive bail based on a set of individual's features
- Propublica analysis of COMPAS algorithm (2016)
- Unbalanced false positive rates
- African-American defendants who didn't subsequently re-offend had higher average scores than white defendants who subsequently didn't reoffend
- White defendants who subsequently re-offended had lower average scores than African-American defendants who subsequently re-offend



	African-American	White
Wrongly labeled high risk	**44.9%**	**23.5%**
Wrongly labeled low risk	**28%**	**47.7%**

Fairness through blindness/unawareness: Ignore the sensitive attributes

Fairness through blindness

- $P(X, A) = P(X)$
- Easy to use and legal support
- Ineffective and harmful in many settings: highly correlated features as proxies to *sensitive attributes*
- Correlation between sensitive attributes (gender) and visiting a website (artofmanliness.com)
- A typical browsing history includes numerous non-sensitive features that become slightly predictive of sensitive attributes
- In large feature spaces, sensitive attributes become redundant given the non-sensitive features

Predicting income from DNA

??

- A start-up wants to predict incomes based on DNA information
- DNA encodes information about ancestry
- Sometimes harmful (in some cases, medication and race)

Group vs. Individual fairness

Group fairness

- Group fairness measures require statistical analysis
- Group fairness answers to questions such as:
    - Do outcomes systematically differ between demographic groups (or other population groups)?

- Algorithms are trained on data: History of explicit discrimination, implicit attitudes and stereotypes about certain groups or social attributes

- Are observed disparities discrimination? Are they justified? Are they harmful?

- Intervening to minimize disparities

Individual fairness

- Similar people are treated **similarly** (with respect to a classification task)

- Goal: prevent discrimination against individuals based on their group membership while maintaining the highest possible utility for the classifier

- [?] Examples: admission to university, bank loan, advertisement, etc

- [?] Dwork et al. (2012): A fair classification requires:

1. a similarity metric between individuals wrt the classification task at hand, and
2. an algorithm for maximizing utility subject to a fairness constraint

- Similarity can be cashed out in terms of distance

Let’s make our intuitions about **fairness** mathematically precise: measuring inequalities

Classification algorithms

- The goal of classification: determine a plausible value for an unknown variable Y given an observed variable X

- Predict whether a loan applicant will **pay back her loan or not** by looking at her characteristics such as credit history, income, and net worth

- Binary classification (whether an image is of a dog or a cat)

- Ternary classification (whether an image is of a dog or a tiger or a Persian cat)

- etc

- Let us call the classifier: $\hat{Y}$
- What makes $\hat{Y}$ a good classifier?
    - Accurate prediction of the target variable $P(Y = \hat{Y})$
- For a binary classifier, we can consider the conditional probability *P*(event | condition)

Event	Condition	*P*(event | condition)
$\hat{Y} = 1$	$Y = 1$	True positive rate
$\hat{Y} = 0$	$Y = 1$	False negative rate
$\hat{Y} = 1$	$Y = 0$	False positive rate
$\hat{Y} = 0$	$Y = 0$	True negative rate

Formal properties for group non-discrimination/ group fairness

1- Independence

- R : the classifier; Y: the target variable; A*:* sensitive attributes
- Independence: $R \perp A$
- Achieve an equal acceptance rate in all groups

$$P(R = r|A = a) = P(R = r|A = b)$$

  - Proxy for a belief about human nature: Certain intrinsic human traits such as intelligence are independent of human race or gender.
  - A long-term societal goal: desire to live in a society where the sensitive attribute is statistically independent of outcomes such as financial well-being

- Also called demographic parity/statistical parity

- Demographic of people with positive/negative classification are the same as the demographic of general population

- The same proportions are classified positively/negatively

- Disparate impact is the reverse of demographic parity

- Some problems with the “independence” condition:

  - Sometimes discrimination is explainable in terms of legitimate grounds.

  - Works only if classifier is extremely accurate

2- Separation

- Allow correlations between the classifier and the sensitive attribute to the extent that it is justified by the target variable: $R \perp A \mid Y$

- For a binary classifier:

$$P(R = 1 \mid Y = 1, A = a) = P(R = 1 \mid Y = 1, A = b)$$ Equality of opportunity

$$P(R = 1 \mid Y = 0, A = a) = P(R = 1 \mid Y = 0, A = b)$$

- All subgroups experience the same error rate (false negative or false positive).

- Also called Equalized Odds, or Positive Rate Parity

3- Sufficiency

- The probability of actually being in each of the groups is equal for two individuals with different sensitive characteristics given that they were predicted to belong to the same group: $Y \perp A \,|\, R$

$$P(Y = 1 \,|\, R = 1, A = a) = P(Y = 1 \,|\, R = 1, A = b)$$

Let’s satisfy all these fairness measures

- Can we be so cool and satisfy all fairness measures?
- To answer, we need to know:
    - What are the relationships between these measures?

Independence, separation, sufficiency

- If A and Y are not statistically independent, then sufficiency and independence cannot both hold
- Assuming Y is binary, if A and Y are not statistically independent, and R and Y are not statistically independent either, then independence and separation cannot both hold

Equal error rates

- Equal false positive rates (future hasn't happened yet)
- Equal false negative rates (future hasn't happened yet)
- Equal positive predictive value $(PPV = \frac{TP}{TP + FP})$

	Labeled High risk	Labeled Low risk
Did not recidivate	**FP**	**TN**
Recidivated	**TP**	**FN**

- **Which fairness measure?**
  - Decision maker looking at only true positive and true negatives (Predictive positive value)
  - Defendant: what's the probability I will be incorrectly classified high risk? (False positive)
  - Society: is the selected set demographically balanced? (demographic parity)
- Different definitions matter to different stakeholders.
- **There is no absolutely "right answer".**

An Impossibility result for algorithmic fairness

- Equal false positive rates (future hasn't happened yet)
- Equal false negative rates (future hasn't happened yet)
- Equal positive predictive value (PPV across groups)

$$FPR = (\frac{P}{1-P})(\frac{1-PPV}{PPV})(1-FNR)$$

- Valid only when the base rate is the same (the groups have the same fraction of positive instances) or we have a perfect predictor
- No imperfect classifier can simultaneously ensure equal *FPR*, *PPV*, *FNR* unless the base rates are equal (mathematically impossible).

The impossibility result

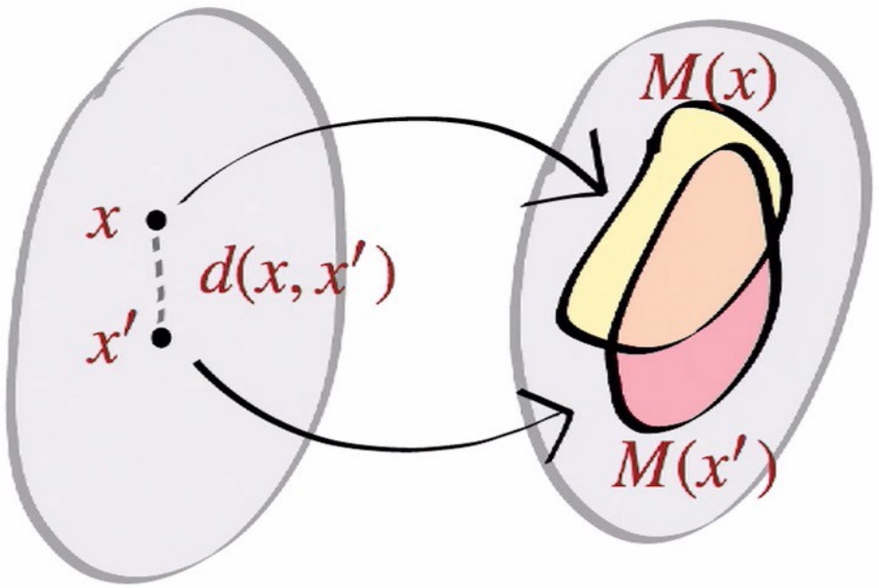
- In any instance of risk score assignment where all three properties can be achieved, we must have either perfect prediction or equal base rate

- Implication: bias in criminal risk scores is **mathematically inevitable**

Individual fairness

- Similar people are treated similarly (have similar probabilities of yes/No outcomes)
- A set V of individuals, A set A of outcomes, A randomized mapping M : V→ Δ(A).
- To classify x ∈ V, choose an outcome a ∈ A according to the probability distribution M(x) Lipschitz condition on the classifier.
- M : V→ Δ(A) satisfies the (D, d)–Lipschitz property if, for any x, y ∈ V,

$$D(M(x), M(y)) \leq d(x, y)$$



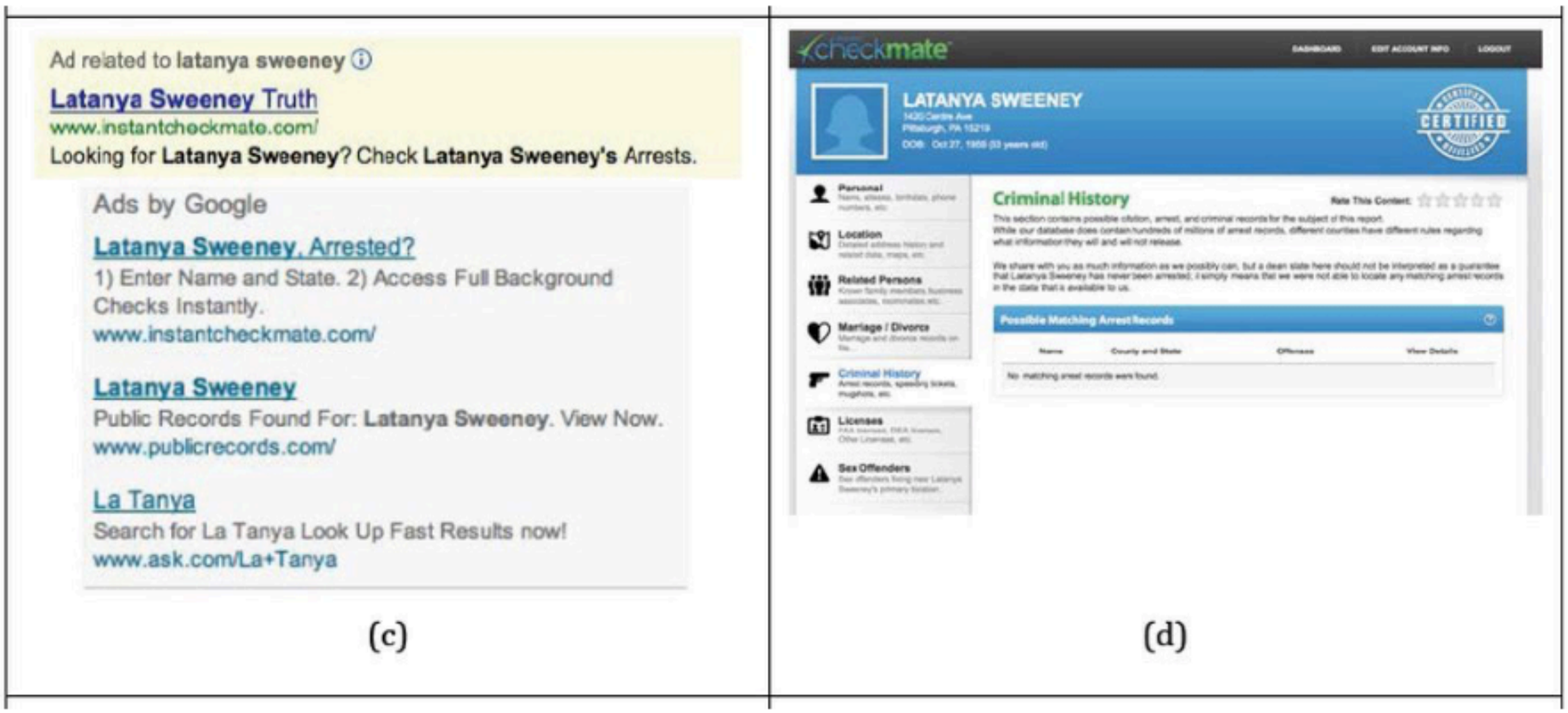
Counterfactual fairness

Racial bias alleged in Google's ad results

Names associated with blacks prompt link to arrest search

By **Hiawatha Bray** Globe Staff, February 6, 2013, 12:00 a.m.

19



Web page results of ads that appeared on-screen when Harvard professor Latanya Sweeney typed her name in a google search. Ads featured services for arrest records. Sweeney conducted a study that concluded searches with "black sounding" names are more likely to get results with ads for arrests records and other negative information. LATANYA SWEENEY

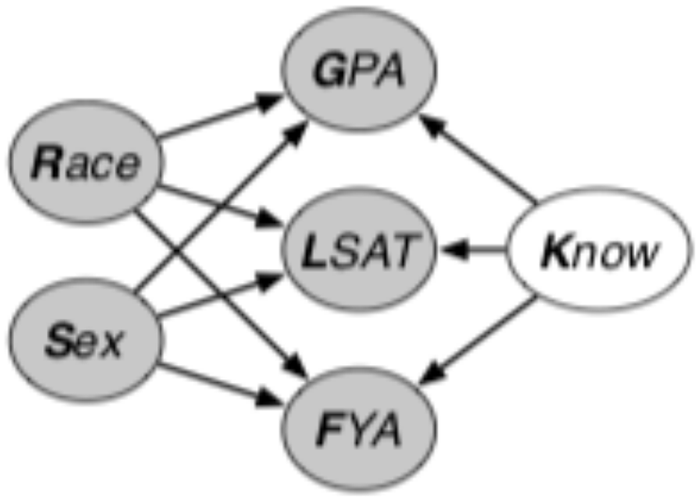
Counterfactual fairness

- Imagine counterfactual scenarios wherein members of protected groups are instead members of the non-protected group

- A predictor $\hat{Y}$ is counterfactually fair if given X (the set of observed non-sensitive attributes) and A,

$$P(\hat{Y}_{A<--a} = y \,|\, X = x, A = a) = P(\hat{Y}_{A<--b} = y \,|\, X = x, A = a)$$

Law school acceptance



Algorithmatization of fairness

Fairness can be applied to machine learning algorithms in three different ways:

- Preprocessing the data used in the algorithm
- Optimization during the training
- Post-processing the answers of the algorithm

References

- Dwork, Cynthia, "Fairness Through Awareness", ITCS, 2012.
- Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." *Advances in neural information processing systems,* 2016.
- Angwin, Julia, "Machine Bias", Propublica, 2016.
- Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." *arXiv preprint arXiv:1609.05807,* 2016.
- Kusner, Matt J., et al. "Counterfactual fairness." *Advances in Neural Information Processing Systems,* 2017.