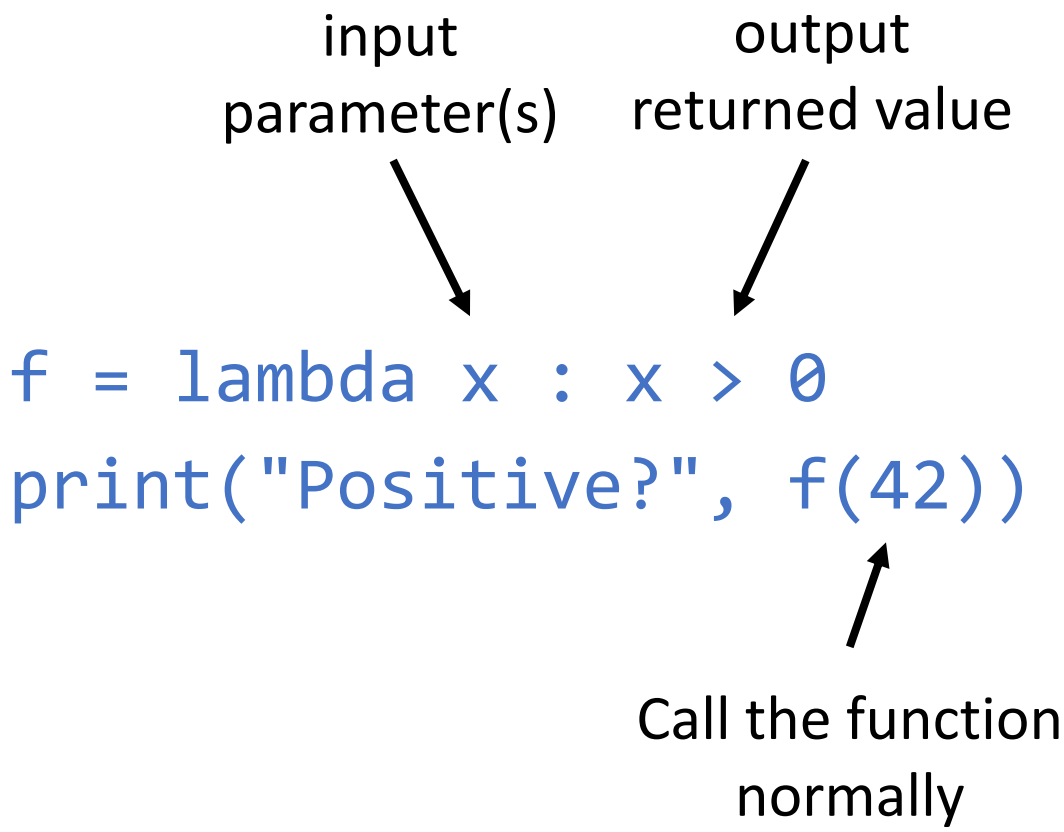
Bonus Slide

Lambda Expressions

Sometimes a function only needs to return a simple expression.

A **lambda expression** lets you define a function in a single line of code!

This is possible because in Python, functions are secretly values, just like variables and data.



Unit 5: CS Ethics

Ethics of Algorithm Design

15-110 – Friday 09/04

Announcements

- **No class next Monday due to Labor Day**
- Hw2 due **Tuesday noon** (adjusted deadline)
- Hw1 revisions also due **Tuesday noon**
- First quizlet next **Wednesday**

Learning Objectives

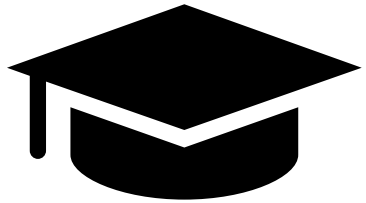
- Discuss the complicating factors behind using algorithms in **kidney-matching, criminal risk assessment,** and **facial recognition**
- Recognize the impact **algorithmic bias** can have in varying scenarios
- **Formulate arguments** for and against using algorithms in varying scenarios

Algorithm Design

Algorithms are Made by People

Algorithms are being used to help make decisions that greatly impact people's lives – decisions about university admission, jobs, healthcare, and more.

When we design an algorithm, we may embed our own values and judgements into that process.



Case Study: Kidney Matching Algorithms

Many people need to get their kidney replaced. There are more people waiting for kidneys than donors willing to provide them. This is complicated by the fact that there must be a **biological match** between a donor and recipient based on factors like matching blood types and antigens.

Computer scientists can design algorithms to help match donors to recipients. But when multiple recipients are eligible, how should we decide which one gets the match?

Discuss: Factors to Consider

You do: what factors should be considered when comparing/ranking possible recipients of a kidney?

An Actual Kidney-Matching Algorithm

The Alliance for Paired Kidney Donation publicly posts the algorithm they use for 'scoring' possible recipients. The factors they assess are:

- Medical factors that increase the likelihood of the transfer working
  - Also distance to the transplant center, which has been shown to directly impact success rates
- Medical factors that make it harder for the candidate to find a matching kidney
- Age
- Whether the candidate previously donated a kidney

Discuss: compare this to the factors we generated. What do you observe?

Bias in Algorithm Design

Case Study: Criminal Risk Assessment

When someone has been charged with a crime, the criminal justice system tries to determine their risk of **recidivism** – how likely they are to commit another crime if released. This may impact whether they will be released while awaiting trial, or whether they are eligible for rehabilitation programs.

This assessment can be done by **humans** who look at the full picture of a person and make a judgement call. Or they can be done by **algorithms** which analyze data about a defendant to predict how risky they are.

Ethical Questions Have a Spectrum of Answers

Our central question is: **should criminal risk assessment be conducted by algorithms or not?** This is an **ethical** question as our decision will be affected by our moral principles.

You do: what is your current position on this issue? You answer can lie on the spectrum of possibilities from 'no algorithms at all' to 'decision made entirely by algorithms'.

https://bit.ly/110-f26-risk1

No algorithms **All algorithms**

Bias in Human Risk Assessment

Studies have shown that judges are subject to human biases when predicting the risk that a defendant will reoffend, classifying black defendants as riskier than white ones. (1)

Why? Some analyses show that judges "rely extensively on intuition, more than deliberative judging, in deciding matters before the bench". (2) And studies have shown that fatigue greatly influences the accuracy of judge's rulings. (3)

1: https://academic.oup.com/qje/article-abstract/133/4/1885/5025665
2: https://libguides.law.uconn.edu/implicit/courts
3: https://www.pnas.org/doi/10.1073/pnas.1018033108

COMPAS Algorithm

The COMPAS algorithm is used by some courts in pre-trial assessments.

Input: a range of data about prior criminal history and the defendant's answers to a series of background and behavioral questions.

Output: a score from 1 to 10 ranking the defendant's risk of recidivism (likelihood of committing another crime)

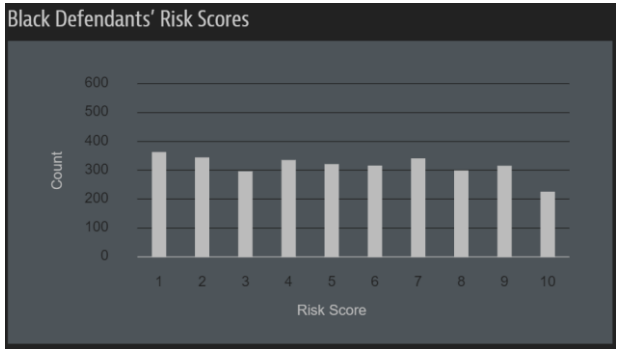
1. Do any current offenses involve family violence?
☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
☐ No ☑ Yes
5. Number of pending charges or holds?
☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
☑ No ☐ Yes

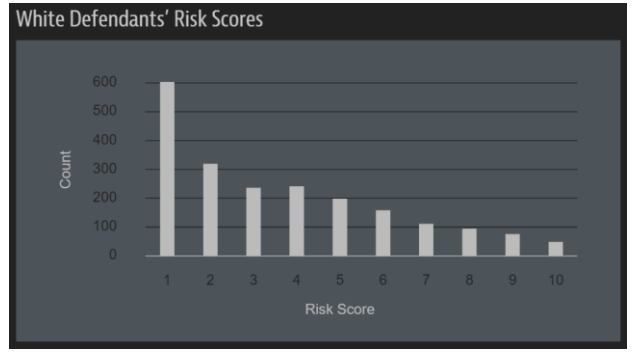
71. Did you complete your high school diploma or GED?
☑ No ☐ Yes
72. What was your final grade completed in school?
9
73. What were your usual grades in high school?
☐ A ☐ B ☑ C ☐ D ☐ E/F ☐ Did Not Attend
74. Were you ever suspended or expelled from school?
☐ No ☑ Yes
75. Did you fail or repeat a grade level?
☑ No ☐ Yes
76. How often did you have conflicts with teachers at school?
☐ Never ☑ Sometimes ☐ Often

Sample questions from the input survey:
https://embed.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE/

Bias in Algorithmic Risk Assessment

ProPublica analyzed actual results produced by the COMPAS algorithm and found bias in the scores assigned to white vs. black defendants.





The overall accuracy of the risk assessment was roughly the same for both groups (61%). However, the ratio of **false positives and false negatives** was drastically different.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

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

True and False Positives and Negatives

We use false positives and false negatives to determine how reliable an algorithm is when it classifies something as part of a category vs. not.

For example, if we have a test that screens people for the flu:

	They have the flu	They don't have the flu
Test says they have the flu	**True Positive**	**False Positive**
Test says they don't have the flu	**False Negative**	**True Negative**

Bias in Algorithmic Risk Assessment

Why does the COMPAS algorithm have different false positive/negative rates across races?

We don't know – the algorithm is proprietary and secret.

Even if we did know, some algorithms (like those created through machine learning) are so complicated that it's hard to **explain** why they make any given decision.

Algorithms Used in New Contexts

Risk assessments are supposed to be used only to determine which defendants are eligible for probation or treatment programs.

However, they have also been used by judges in sentencing, among other factors.

Key takeaway: **algorithms may be used for different purposes than what they were designed for**.

You Do: Arguments For and Against

No Algorithms in Risk Assessments:

All Algorithms in Risk Assessments:

Algorithmic Bias

Algorithmic Bias occurs when an algorithm produces unfair or discriminatory outcomes for certain groups.

This can happen due to bias that exists in the algorithm's design, its training data, or how the algorithm's results are evaluated/applied.

What is training data? What does it mean for an algorithm to be 'unfair'?

https://www.ibm.com/think/topics/algorithmic-bias

Bias in Training Data

Algorithms Derived From Data

These days, many algorithms are created using **machine learning.** We call these algorithms **models**.

We'll talk specifically about **classifiers**, which are models that take a piece of data and assign it to one of several categories.

For example, a classifier can take an email and classify it as Important, Normal, or Spam.

Learning From Examples

If you want to learn a new skill, you can practice that skill by repeating it on many practice problems. Eventually you can apply the skill to real-world problems.

Similarly, classifiers use **many examples** to find patterns across a set of categories. Eventually they can apply those patterns to new data to predict which group it belongs to.

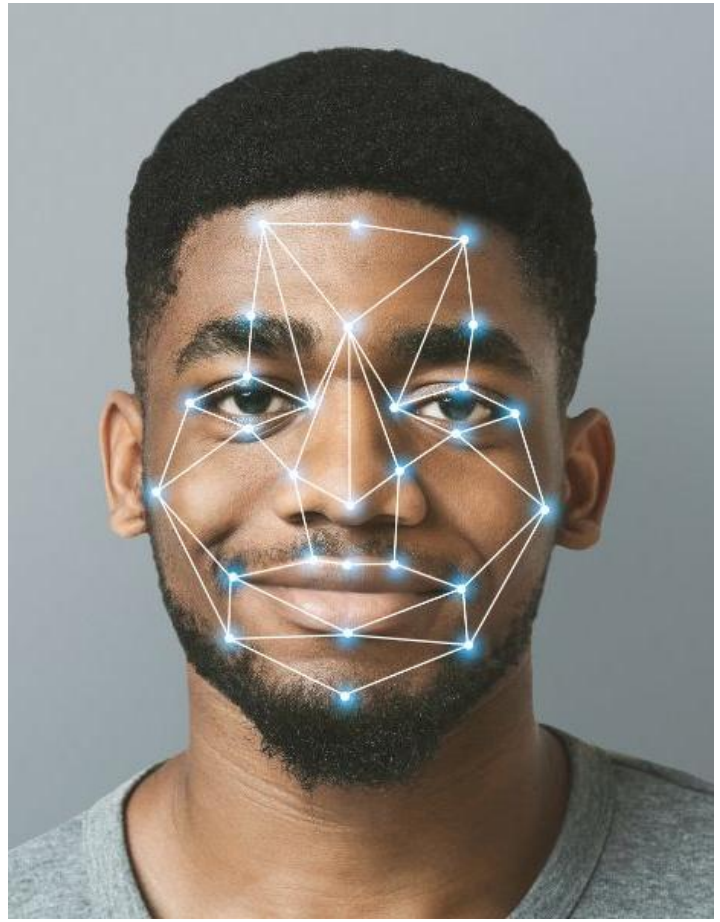
Where do the examples come from? Datasets!

Case Study: Facial Recognition

Since the 1960s computer scientists have worked on facial recognition – using a person's features to automatically identify them.

This can be used for biometric identification, automatic tagging of photos, surveillance, and more.

As part of this effort, some groups developed **benchmark datasets** that contained many pictures of individual people.

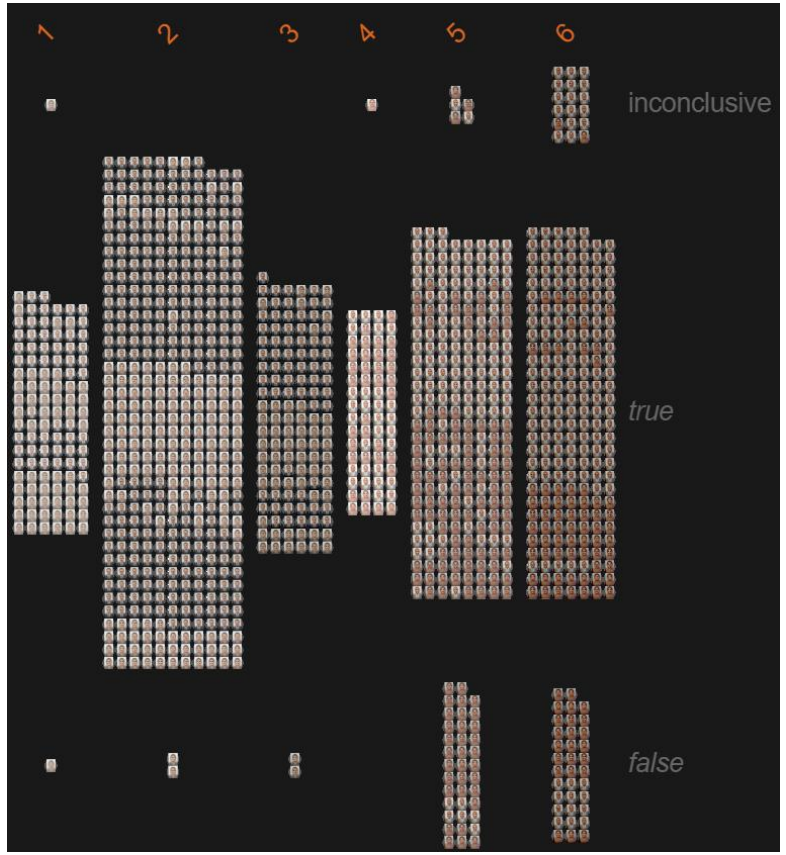


Unequal Representation in Data

The Gender Shades project found that popular datasets available in 2017 were heavily unbalanced in the proportions of subjects with different skin colors, with ~80% light-skinned subjects compared to ~20% dark-skinned.

They demonstrated how this had a direct impact on the accuracy of facial recognition algorithms on dark-skinned individuals. **The algorithms performed worse because they had less training on that group.**

Let's explore: https://gs.ajl.org/



https://proceedings.mlr.press/v81/buolamwini18a.html

Unequal Representation in Data

NIST found similar results when they ran an analysis in 2019. Most facial recognition algorithms they tested had different false positive rates across different demographics when verifying identities.

Interestingly, **algorithms developed in Asia had lower false positive rates on Asian faces than algorithms developed in the US.** One possible explanation is that the Asian developers used datasets with more Asian faces.

Key takeaway: your algorithm is only as good as the data you train it on

https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software

Fairness in Algorithm Design

What Does Fairness Mean?

Discuss: what would it mean for an algorithm to be entirely fair?

Metrics of Fairness

Computer scientists have designed metrics to measure how fair an algorithm is.

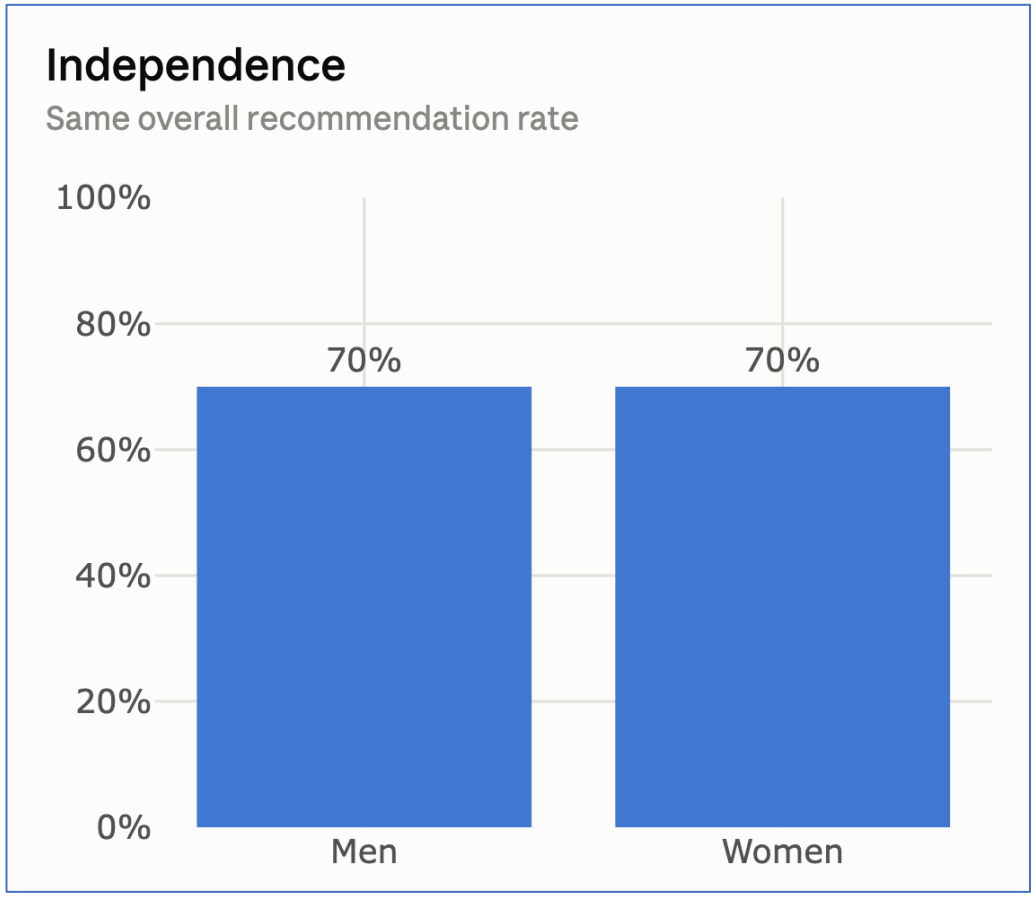
Let's consider them in the context of an algorithm used to recommend college admissions. We'll discuss **independence** and **separation**. (**Sufficiency** is another common metric, but we won't discuss it here.)

Maintain Proportions Across Subgroups

Independence: all subgroups in a population should have **equal proportions** across all categories.

	All	Men	Women
Recommended	70%	70%	70%
Not Recommended	30%	30%	30%

Mathematically speaking, the subgroup should be **independent** of the predicted group.

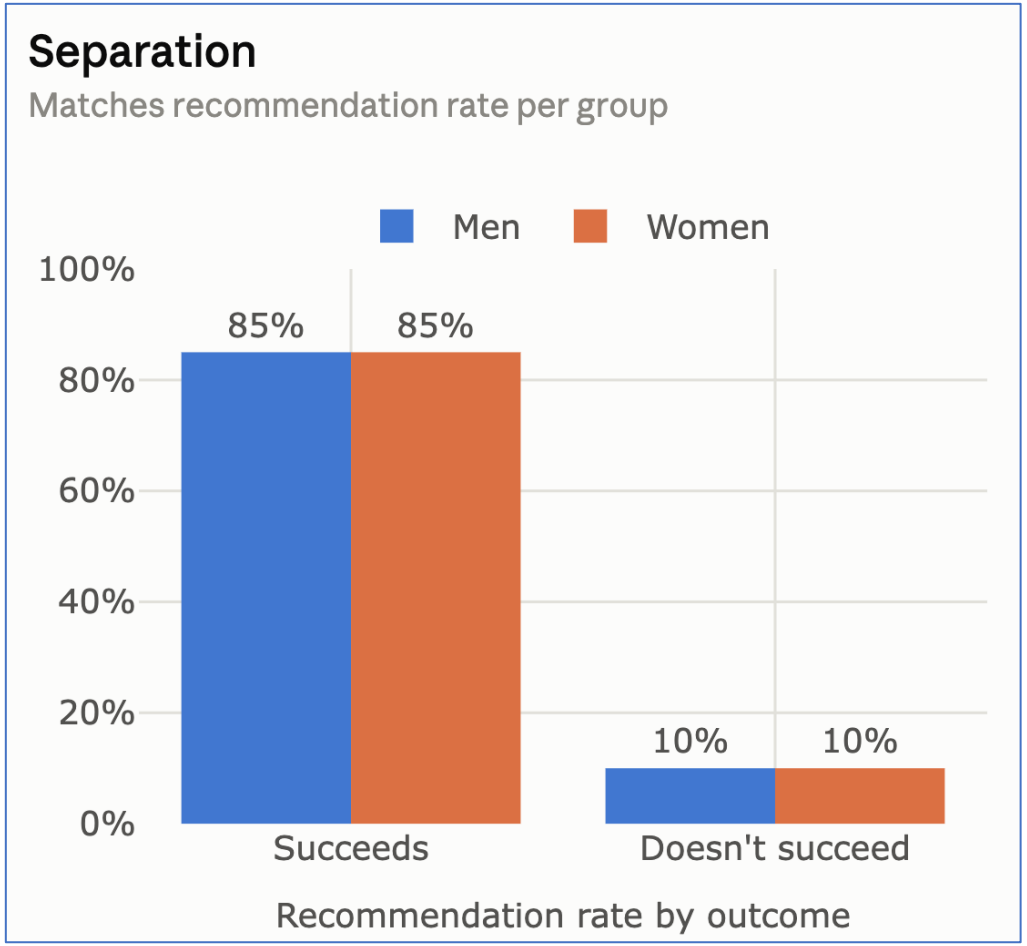


Equalize False Positives Across Subgroups

Separation: all subgroups should have **equal true positive / false positive rates** across all categories

	All	Men	Women
Succeeds, Recommended	85%	85%	85%
Doesn't Succeed, Recommended	10%	10%	10%

Mathematically speaking, the subgroup should be independent of the predicted group **given the true classification**.



Can't Have Them All

Here's where things get tricky: **mathematically, it's impossible to optimize for both independence and separation**.

Consider our admission example. We can't get both of these tables to be balanced across genders – one will show different rates.

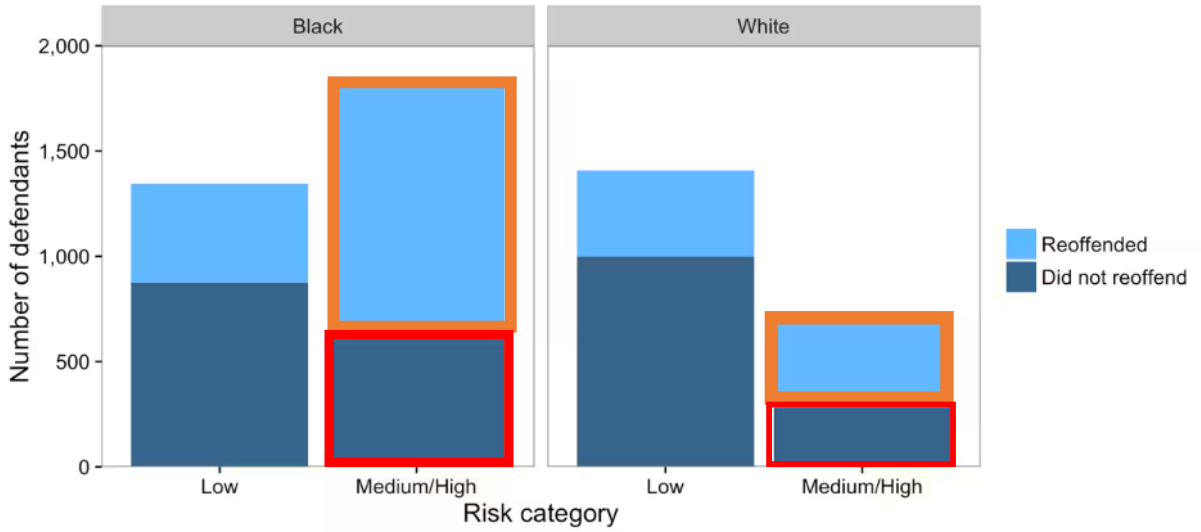
	All	Men	Women
Recommended	70%	70%	70%
Not Recommended	30%	30%	30%

	All	Men	Women
Recommended, Succeeds	85%	85%	85%
Recommended, Doesn't Succeed	10%	10%	10%

COMPAS – One Measure of Fairness

COMPAS defined fairness as **equal rates of recidivism across races in each risk category**.

Because black defendants were more likely to recidivate (52% vs 39%), there were more in the medium/high category, and thus more false positives.



Distribution of defendants across risk categories by race. Black defendants reoffended at a higher rate than whites, and accordingly, a higher proportion of black defendants are deemed medium or high risk. As a result, blacks who do not reoffend are also more likely to be classified higher risk than whites who do not reoffend.

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

Revisit the Question

https://bit.ly/110-f26-risk2

You do: think back on everything we've discussed. Where do you stand on the spectrum of algorithms in criminal risk assessment now?

No algorithms **All algorithms**

Learning Objectives

- Discuss the complicating factors behind using algorithms in **kidney-matching algorithms, criminal risk assessment,** and **facial recognition**
- Recognize the impact **algorithmic bias** can have in varying scenarios
- **Formulate arguments** for and against using algorithms in varying scenarios