# CSE/STAT 416

## Bias and Fairness in ML

Tanmay Shah
Paul G. Allen School of Computer Science & Engineering
University of Washington

July 8, 2024

# ML and Society

# ML Systems Gone Wrong

INDEPENDENT

GOOGLE'S ALGORITHM SHOWS PRESTIGIOUS JOB ADS TO MEN, BUT NOT TO WOMEN

REUTERS

Business Markets World Politics TV More

BUSINESS NEWS OCTOBER 9, 2018 / 8:12 PM / 5 MONTHS AGO

**Amazon scraps secret AI recruiting tool that showed bias against women**

Jeffrey Dastin

8 MIN READ

The New York Times

***Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says***

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019

168

MIT Technology Review

**Intelligent Machines**

**How to Fix Silicon Valley's Sexist Algorithms**

Computers are inheriting gender bias implanted in language data sets—and not everyone thinks we should correct it.

PROPUBLICA

**Machine Bias**

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

# COMPAS

An ML model created by NorthPointe used to predict likelihood of inmates to "recidivate". Eventually started use in Florida in judges' decision for parole

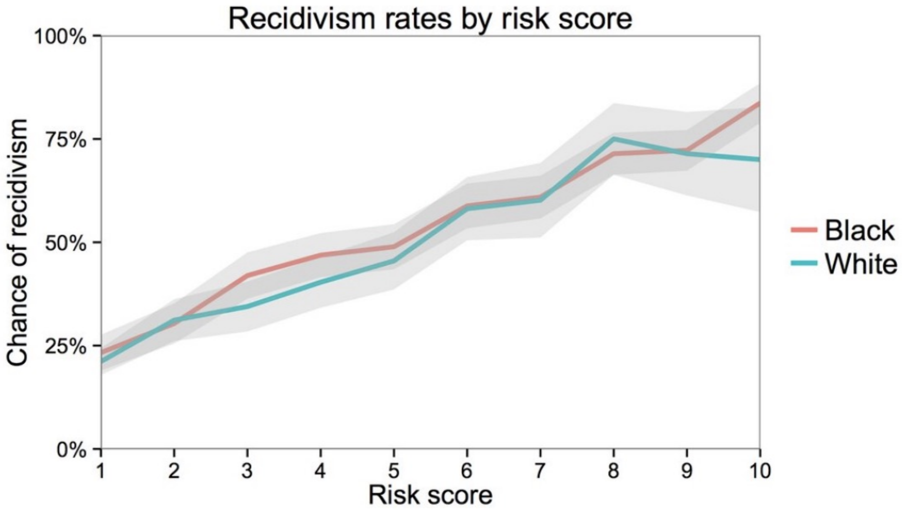
ProPublica (a news org) investigated the model and wrote that the model exhibited biased behavior against people of color. Particularly, they found that the model would predict higher risk scores for black people.

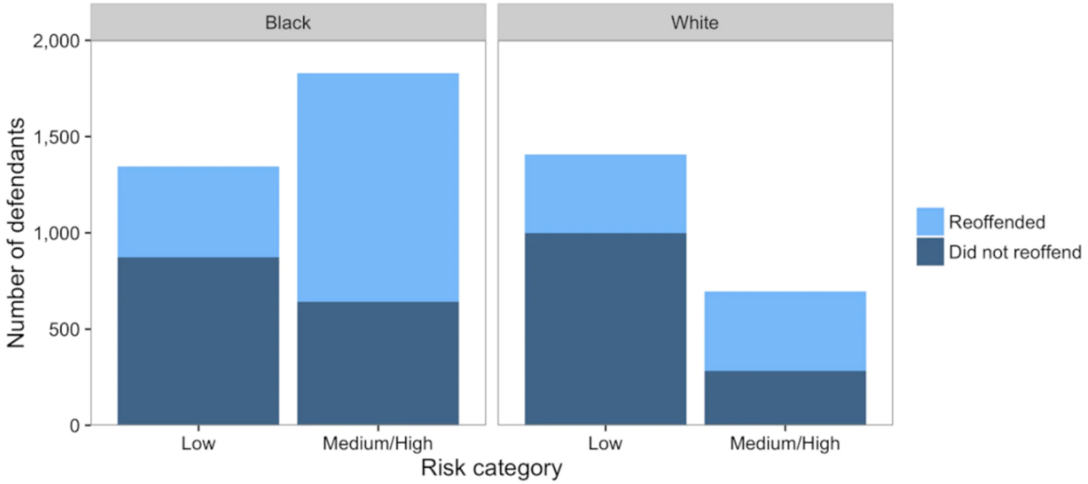
Northpointe countered and claimed that their scores were well **calibrated** (e.g., when the predict score of 9/10 that person recidivates about 90% of the time).

- Interesting follow up from ProPublica

So the question is: Who is right? Is it right to use this model?

# COMPAS


Recidivism rates by risk score
100%
75%
50%
25%
0%
Chance of recidivism
1
2
3
4
5
6
7
8
9
10
Risk score
Black
White


Black
White
2,000
1,500
1,000
500
0
Number of defendants
Low
Medium/High
Low
Medium/High
Risk category
Reoffended
Did not reoffend

# Why Biased Outcomes?

Probably not the case that someone explicitly coded the model to be biased against a particular race. In fact, race was not even a question that was on the survey inmates took!

More often than not, biased outcomes from a model come from **the data it learns from** rather than some explicit choice from the modeler.

> “Garbage in → Garbage out”
>
> “Bias in → Bias out”

# Sources of Bias

# Sources of Bias

Discussion heavily based on Suresh and Guttag (2020)

Six common sources of bias:

**Historical bias**

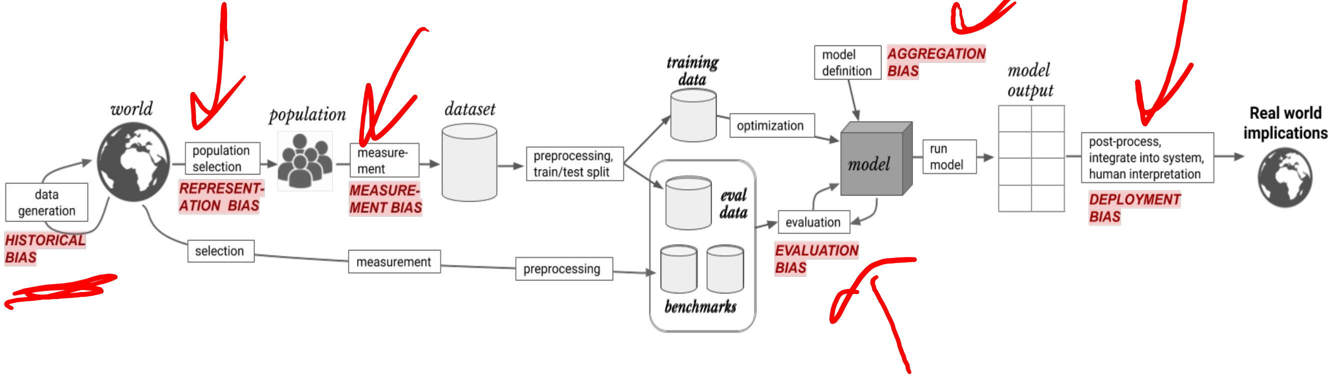
**Representation Bias**

**Measurement Bias**

**Aggregation Bias**

**Evaluation Bias**

**Deployment Bias**




[A FRAMEWORK FOR UNDERSTANDING UNINTENDED CONSEQUENCES OF MACHINE LEARNING](), BY HARINI SURESH AND JOHN V. GUTTAG, 2020

# Historical Bias

The world we lived in is one that contains biases for/against certain demographics. Even ‘accurate’ data could still be harmful.

Historical bias exists even with perfect sampling or feature measurement (other sources of bias are possible)!

Examples:

In 2018, 5% of Fortune 500 CEOs were women. Should search results for “CEO” match this statistic? Could reflecting the world (even if accurately) perpetuate more harm?

# Representation Bias

When the *training data* we collect does not contain representative samples of the true distribution.

Examples:

- If we use data gathered from smart phones, we would likely be underestimating poorer and older populations.
- ImageNet (a very popular image dataset) with 1.2 million images. About 45% of these images were taken in the US and the majority of the rest in North America and Western Europe. Only about 1% and 2.1% of the images come from China and India respectively.

# Measurement Bias

Often we are gathering data that contains (noisy) proxies of characteristics of interest. Some examples:

- Financial responsibility → Credit Score
- Crime Rate → Arrest Rate
- Intelligence → SAT Score

If these measurements are not measured equally across groups or places (or aren't relevant to the task at hand), this can be another source of bias.

# Measurement Bias (cont.)

Examples:

If factory workers are monitored more often, more errors are spotted. This can result in a **feedback loop** to encourage more monitoring in the future.

- Same principles at play with predictive policing. Minoritized communities were more heavily policed in the past, which causes more instances of documented crime, which then leads to more policing in the future.

Women are more likely to be misdiagnosed (or not diagnosed) for conditions where self-reported pain is a symptom. In this case aspect of our data “diagnosed with X” is a biased proxy for “has condition X”.

The feature we measure is a poor representation of the quality of interest (e.g., SAT score doesn’t actually measure intelligence)

# Aggregation Bias

When we use a “one-sized fits all” model that does not accurately serve every group equally.

Examples:

HbA1c levels (used to monitor and diagnose diabetes) differ in very complex ways across ethnicities and sexes. One model for everyone might not be the right choice, even if everyone is represented well in the training data.

# Evaluation Bias

Similar to representation bias, but focused more on the data we evaluate or test ourselves against. If the evaluation dataset or benchmark doesn't represent the world well, we have evaluation bias.

> **Benchmarks** are common datasets used to evaluate models from different researchers.

Examples:

> If it is common to report accuracy on a benchmark, this might hide disparate performance on subgroups.
>
> Drastically worse performance for facial recognition software when used on faces of darker-skinned females. Common evaluation datasets for facial recognition only had 5-7% had faces of darker-skinned women.

# Deployment Bias

When how a model was intended to be used and how it is actually used when deployed in the real-world.

Examples:

- Crime risk prediction models might be evaluated to achieve good calibration, but the model designers might not have evaluated the model's use in the context of determining prison sentence lengths.
- People are complex and when using models to aid their decisions, might make incorrect assumptions about what a model says.

# Sources of Bias

Discussion heavily based on Suresh and Guttag (2020)

Six common sources of bias:

**Historical bias**

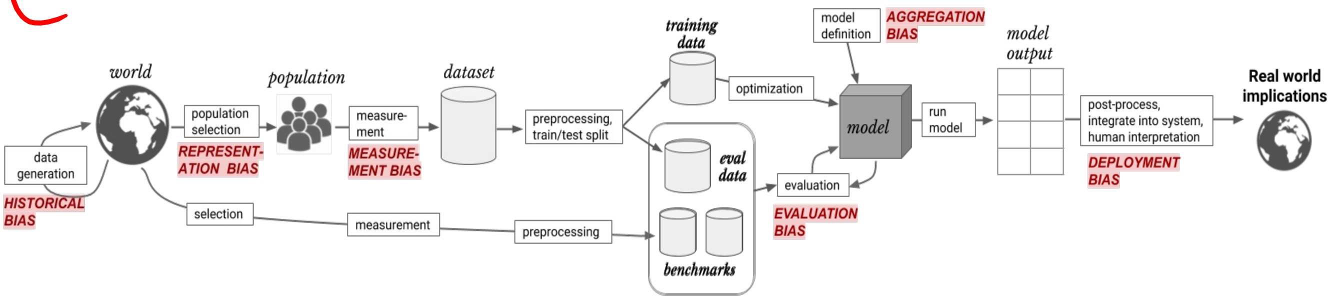
**Representation Bias**

**Measurement Bias**

**Aggregation Bias**

**Evaluation Bias**

**Deployment Bias**




[A FRAMEWORK FOR UNDERSTANDING UNINTENDED CONSEQUENCES OF MACHINE LEARNING](), BY HARINI SURESH AND JOHN V. GUTTAG, 2020

# Fairness in ML

# Fairness

What does it mean for a model to be fair or unfair? Can we come up with a numeric way of measuring fairness?

Lots of work in the field of ML and fairness is looking into mathematical definitions of fairness to help us spot when something might be unfair.

- There is not going to be one central definition of fairness, as each definition is a mathematical statement of which behaviors are/aren't allowed.
- Different definitions of fairness can be contradictory!

Today, we will focus on notions of **group fairness** in an attempt to prevent discriminatory outcomes.

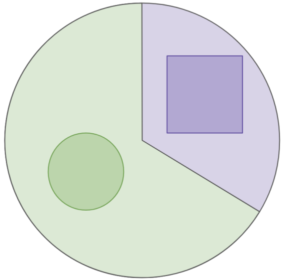
# Example: College Admissions

Will use a very simplified example of college admissions. This is **not** an endorsement of such a system or a statement of how we think the world does/should work.  Will make MANY simplifying assumptions (which are unrealistic).

There is a single definition of "success" for college applicants, and the goal of an admissions decision is to predict "success"

The only thing we will use as part of our decision is SAT Score

To talk about group fairness, will assume everyone belongs to exactly one of two races: Circles (66%) or Squares (33%).



# Notation

Example: College admission only using SAT Score

$X$ input about a person for prediction

Example: $X$ = SAT Score

$A$ variable indicating which group $X$ belongs in

Example: $A =$ ■ or $A = \bigcirc$

$Y$ the "true label"

Example: $Y = +$ if truly successful in college, $Y = -$ if not

$\hat{Y} = \hat{f}(X)$ is our prediction for $Y$ using a learned model $\hat{f}$

Example: $\hat{Y} = +$ if predicted successful, $\hat{Y} = -$ otherwise

# Fairness Definition 1: "Shape Blind"

To avoid unfair decisions, prevent the model from every looking at protected attribute (e.g., if the applicant is Circle/Square).

Often called **"Fairness through unawareness"**

**Doesn't work in practice.** This does not prevent historical or measurement bias. Protected attributes can be unintentionally inferred from other, related attributes (e.g., in some cities, zip code can be deeply correlated with race).

# Confusion Matrix

For binary classification, there are only two types of mistakes

$$\hat{y} = +1, \;\; y = -1$$

$$\hat{y} = -1, \;\; y = +1$$

Generally we make a **confusion matrix** to understand mistakes.

| | | **Predicted Label** | |
|---|---|---|---|
| | | + | − |
| **True Label** | + | *True Positive (TP)* | *False Negative (FN)* |
| | − | *False Positive (FP)* | *True Negative (TN)* |

# Binary Classification Measures

Notation

$$C_{TP} = \#\text{TP}, \quad \text{C}_{\text{FP}} = \#\text{FP}, \quad \text{C}_{\text{TN}} = \#\text{TN}, \quad \text{C}_{\text{FN}} = \#\text{FN}$$

$$N = C_{TP} + C_{FP} + C_{TN} + C_{FN}$$

$$N_P = C_{TP} + C_{FN}, \quad N_N = C_{FP} + C_{TN}$$

**Error Rate**

$$\frac{C_{FP} + C_{FN}}{N}$$

**Accuracy Rate**

$$\frac{C_{TP} + C_{TN}}{N}$$

**False Positive rate (FPR)**

$$\frac{C_{FP}}{N_N}$$

**False Negative Rate (FNR)**

$$\frac{C_{FN}}{N_P}$$

**True Positive Rate or Recall**

$$\frac{C_{TP}}{N_P}$$

**Precision**

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

**F1-Score**

$$2\,\frac{Precision \cdot Recall}{Precison + Recall}$$

[See more!]

# Fairness Definition 2: Statistical Parity

Idea: “Admit decisions are equivalent across groups”

66%

33%

$$\boxed{\Pr(\hat{Y} = +|\, A = \blacksquare)} = \boxed{\Pr(\hat{Y} = +|\, A = \bigcirc)}$$

Also phrased as matching demographic statistics (e.g., if 33% of population are Squares, 33% of those admitted should be Square).

**Pros:**

Aligns with certain legal definitions of equity.

**Cons:**

A rather weak in fairness requirements. Allows for strategies that might not be desirable (e.g., random selection, self-fulfilling prophecy)

# Fairness Definition 3: Equal Opportunity

Idea: True positive rate should be equivalent across groups

$$\Pr(\hat{Y} = +|\, A = \blacksquare, Y = +) = \Pr(\hat{Y} = +|\, A = \bigcirc, Y = +)$$

**Pros:**

Better controls for true outcome

**Cons:**

More complex to explain to non-experts

Only protects for the positive outcome

Note: Equality of true positives is the same as equality of false negatives

# Fairness Definition 4: Predictive equality

Idea: True negative rate should be equivalent across groups

$$\Pr(\hat{Y} = - \mid A = \blacksquare, Y = -) = \Pr(\hat{Y} = - \mid A = \bigcirc, Y = -)$$

Same idea as equal opportunity, but controlling for different statistic. Might be favorable in situations you care more about false positives than a false negative.

Note: Equality of true negatives is the same as equality of false positives

# And many, many more

List of demographic fairness criteria

| Name | Closest relative | Note | Reference |
|---|---|---|---|
| Statistical parity | Independence | Equivalent | Dwork et al. (2011) |
| Group fairness | Independence | Equivalent | |
| Demographic parity | Independence | Equivalent | |
| Conditional statistical parity | Independence | Relaxation | Corbett-Davies et al. (2017) |
| Darlington criterion (4) | Independence | Equivalent | Darlington (1971) |
| Equal opportunity | Separation | Relaxation | Hardt, Price, Srebro (2016) |
| Equalized odds | Separation | Equivalent | Hardt, Price, Srebro (2016) |
| Conditional procedure accuracy | Separation | Equivalent | Berk et al. (2017) |
| Avoiding disparate mistreatment | Separation | Equivalent | Zafar et al. (2017) |
| Balance for the negative class | Separation | Relaxation | Kleinberg, Mullainathan, Raghavan (2016) |
| Balance for the positive class | Separation | Relaxation | Kleinberg, Mullainathan, Raghavan (2016) |
| Predictive equality | Separation | Relaxation | Chouldechova (2016) |
| Equalized correlations | Separation | Relaxation | Woodworth (2017) |
| Darlington criterion (3) | Separation | Relaxation | Darlington (1971) |
| Cleary model | Sufficiency | Equivalent | Cleary (1966) |
| Conditional use accuracy | Sufficiency | Equivalent | Berk et al. (2017) |
| Predictive parity | Sufficiency | Relaxation | Chouldechova (2016) |
| Calibration within groups | Sufficiency | Equivalent | Chouldechova (2016) |
| Darlington criterion (1), (2) | Sufficiency | Relaxation | Darlington (1971) |

Table from [Fairness and machine learning](Fairness and machine learning) by Barocas, Hardt, Narayanan

# Which one to use?

We can't tell you! Each definition makes its own statement on what fairness means. Choosing a fairness measure is an explicit statement of what values we hold when thinking about fairness.

**Takeaway**: Discrimination in ML models is a crucial problem we need to work on. It's not a problem that will only be solved algorithmically. We need people (e.g., policymakers, regulators, philosophers, developers) to be in the loop to determine the values we want to encode into our systems.

Let's discuss some limitations in these definitions (particularly how they contradict) and how we can think about fairness as a philosophy (or worldview).

# Fairness

What does it mean for a model to be fair or unfair? Can we come up with a numeric way of measuring fairness?

Lots of work in the field of ML and fairness is looking into mathematical definitions of fairness to help us spot when something might be unfair.

- There is not going to be one central definition of fairness, as each definition is a mathematical statement of which behaviors are/aren't allowed.
- Different definitions of fairness can be contradictory!

# Group Fairness

Fairness through Unawareness

Statistical Parity

- Require admissions match demographics in data

Equal Opportunity

- Require false-negative rate to be equal across groups

Predictive Equality

- Require false-positive rate to be equal across groups

# (Im)possibility of Fairness

Four reasonable conditions we want in a real world ML Model:

1. Statistical Parity
2. Equal Opportunity (Equality across false negative rates)
3. Predictive Equality (Equality across false positive rates)
4. Good accuracy of the model across subgroups

In general, can't satisfy all 4 simultaneously *unless* groups have the exact same underlying distribution.

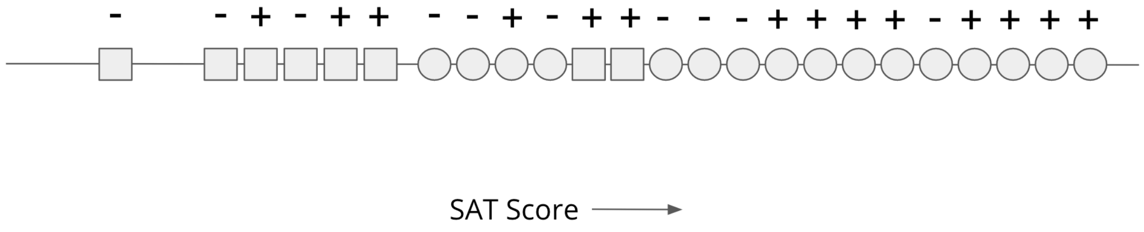
This condition is rarely met in practice as we mentioned earlier when there are so many places for bias to enter our data collection.

# Brain Break

# College Admissions - Continued

Continuing overly simplistic college admissions example, with a fake dataset.

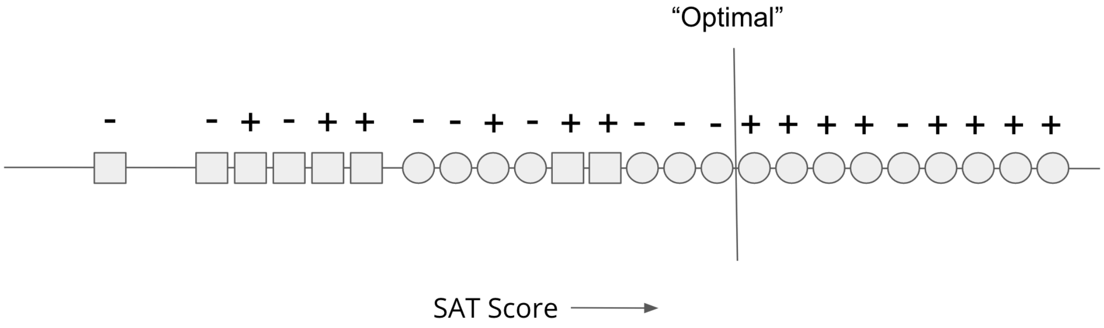
- Majority (2/3) are Circle, the remaining 1/3 are Square
- SAT score for Circles tends to be inflated when compared to Squares. Possibility: Systematic barriers and access to SAT Prep
- Even though we see statistical differences between groups in our data, the rate in which they are actually successful is the same.



# Accuracy and Fairness

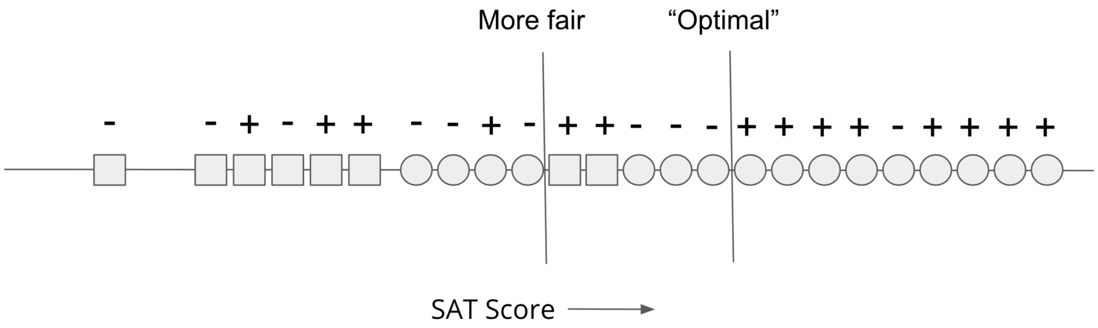
With only one feature, we will consider a simple threshold classifier (a linear classifier with 1 input!).

The most accurate model is not necessarily the most fair.



# Fairness-Accuracy Tradeoff

In general, we find there is a tradeoff between accurate models and fair models. Making a model more fair tends to decrease accuracy by some amount.



# Notes on Tradeoff

Might argue that my example is overly simplistic (it is!), but I'll claim this is a proof of concept. We saw lots of examples of "accurate" models that were unfair.

This is not a statement that a tradeoff necessarily must exist, it just generally happens in real-world datasets.

Originally just cared about finding the most accurate model, saw unfairness as a byproduct. Controlling for fairness will yield a different model than you found before.

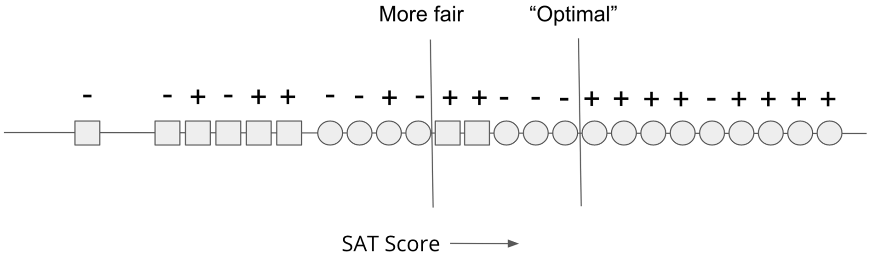
If we recognize data can encode biases and accuracy is determined in terms of that data, trying to achieve fairness will likely hurt accuracy.

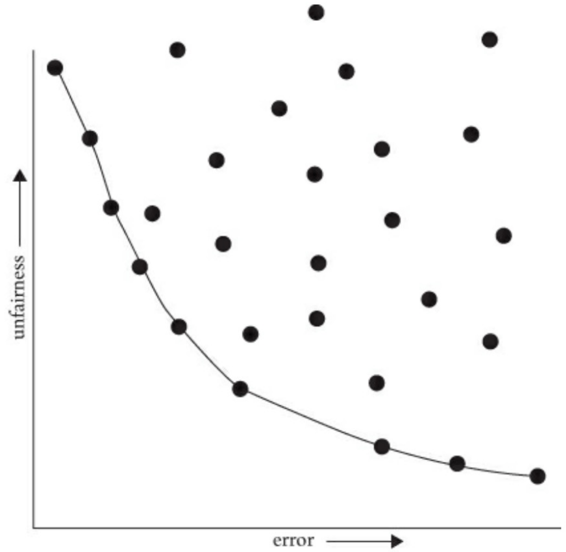
- In the example before, the artificial difference in SAT scores caused the problem.

# Pareto Frontier

Visualizing the tradeoff between fairness and accuracy

**Does not** tell you which tradeoff is appropriate!



# Thoughts on Pareto Frontier

This feels a bit cold-hearted, it's okay to like this is weird. Michael Kearns and Aaron Roth write in *The Ethical Algorithm*

While the idea of considering cold, quantitative trade-offs between accuracy and fairness might make you uncomfortable, the point is that there is simply no escaping the Pareto frontier. Machine learning engineers and policymakers alike can be ignorant of it or refuse to look at it. But once we pick a decision-making model (which might in fact be a human decision-maker), there are only two possibilities. Either that model is not on the Pareto frontier, in which case it's a "bad" model (since it could be improved in at least one measure without harm in the other), or it is on the frontier, in which case it implicitly commits to a numerical weighting of the relative importance of error and unfairness. Thinking about fairness in less quantitative ways does nothing to change these realities—it only obscures them.

Making the trade-off between accuracy and fairness quantitative does ***not*** remove the importance of human judgment, policy, and ethics—it simply focuses them where they are most crucial and useful, which is in deciding exactly which model on the Pareto frontier is best (in addition to choosing the notion of fairness in the first place, and which group or groups merit protection under it, [...]). Such decisions should be informed by many factors that cannot be made quantitative, including what the societal goal of protecting a particular group is and what is at stake. Most of us would agree that while both racial bias in the ads users are shown online and racial bias in lending decisions are undesirable, the potential harms to individuals in the latter far exceed those in the former. So in choosing a point on the Pareto frontier for a lending algorithm, we might prefer to err strongly on the side of fairness—for example, insisting that the false rejection rate across different racial groups be very nearly equal, even at the cost of reducing bank profits. We'll make more mistakes this way—both false rejections of creditworthy applicants and loans granted to parties who will default—but those mistakes will not be disproportionately concentrated in any one racial group.

## Brain Break

# Fairness as Worldview

# Context

So far have discussed notions of **group fairness**, but other notions of fairness exist. Provide a framework for how to approach learning tasks and what assumptions we make. Based on Friedler et al. (2016).

High level ideas:

- Data gathering and modeling
- Individual fairness vs. group fairness
- Common world-views that dictate which fairness is appropriate
- How these worldviews can contradict each other

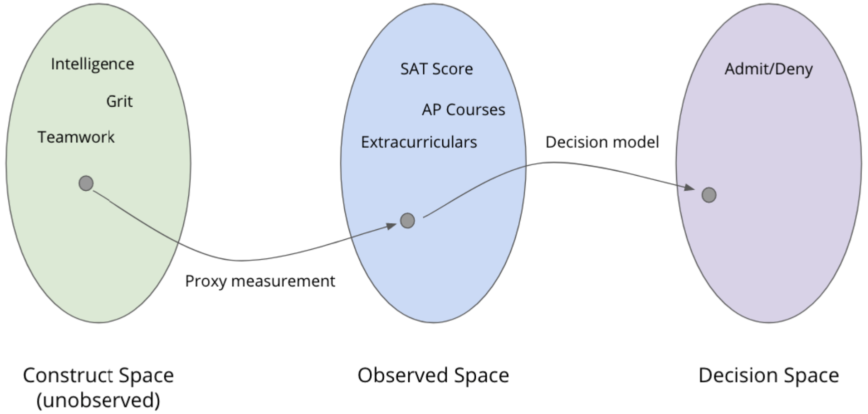
# ML and Spaces

Defined modeling as transformation through three spaces

**Construct space:** True quantities of interest (unobserved)

**Observed space:** Data gathered to (hopefully) represent constructs. Achieved through measurement of proxies.

**Decision space:** The decisions of the model. Models take observed data and make decisions.

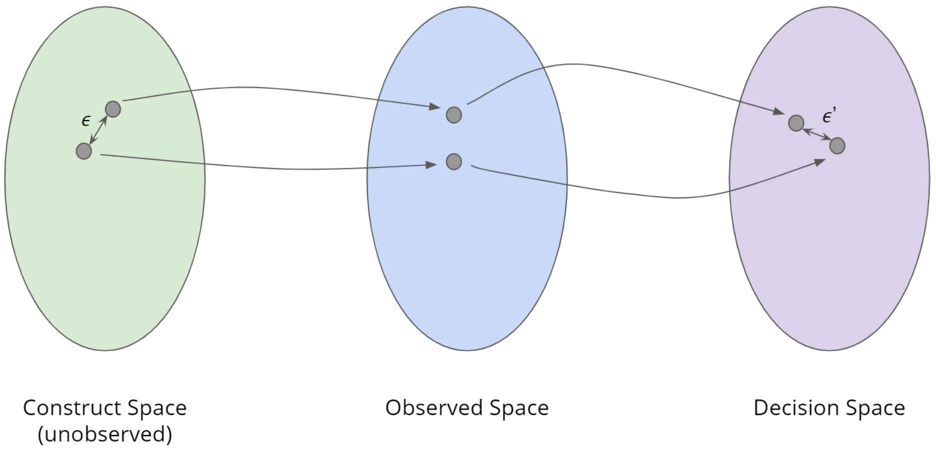


# Individual Fairness

Idea: If two people are close in construct space, they should receive similar decisions.

**Individual Fairness:** A model $f: CS \rightarrow DS$ is said to be fair if objects close in CS are close in DS. Specifically, it is $(\varepsilon, \varepsilon')$-fair if for any $x, y \in CS$

$$d_{CS}(x, y) \leq \varepsilon \quad \Rightarrow \quad d_{DS}\big(f(x), f(y)\big) \leq \varepsilon'$$



# Worldview 1: WYSIWYG

Problem: We can’t tell if two objects are close in CS. So if we want to use individual fairness, we must make an assumption about how the world workds

**What You See is What You Get (WYSIWYG):** The Observed Space is a good representation of the Construct Space.

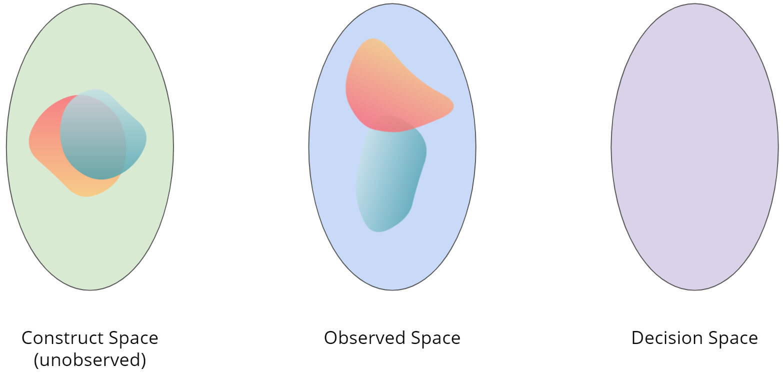
> Example: For college admissions, things like SAT correlate well with intelligence.

With WYSIWYG, you can ensure fairness by comparing objects in the Observed Space as a good proxy for the Construct Space

# Worldview 2: Structural Bias + WAE

What if we don't believe the Observed Space represents the Construct Space well? What if there is some **structural bias** that make people close in the construct space look different in the observed space?

Example: SAT doesn't just measure intelligence, but also measures ability to afford SAT prep. People who are just as intelligent as someone else, can end up with different observations.



# Worldview 2: Structural Bias + WAE

When considering Structural Bias, commonly will also assume **We're All Equal (WAE).**

**We're All Equal (WAE):** Membership in some protected group (e.g., race) *should not* be the cause of a meaningful difference for the task at hand (e.g., academic preparation). Not saying every group is exactly equal in all ways, but for the task at hand we are equal enough that it shouldn't be the cause of difference.

> Differences seen in groups in Observed Space are the result of structural bias!

Notions of group fairness make sense with Structural Bias + WAE

# Which One?

So which is right? WYSIWYG or Structural Bias + WAE?

- No way to know! They are statements of belief!
- Which worldview you use determines what you think is fair

**If you assume WYSIWYG**

- Individual fairness is right and easy to achieve
- Non-discrimination may violate individual fairness

**If you assume Structural Bias + WAE**

- Non-discrimination is right and is possible (saw group fairness mechanisms)
- Attempts to achieve individual fairness may result in discrimination.

# Takeaways

Models can have a huge impact on society, both positive and negative.

- If we are not careful, our models will at best, perpetuate and at worst, amplify injustice in our society.

Historically, people thought defining things like accuracy was easy but defining what is/isn't fair was not. Only recently (~10 years) have ML researchers tried to define what fairness might mean and how to enforce it in our models.

It's clear that defining and enforcing fairness, but what fairness and how is a crucial problem we need humans (and not just ML engineers) in the loop to determine. These are questions of values, and we need humans to make informed decisions of what is right.

# Recap I

**Theme**: It's important to give terms to abstract notions like bias and fairness so we can have concrete things to look out for. There is not one right perspective though!

**Ideas:**

Calibration

Impacts of ML Systems on society

Sources of bias

- Historical bias
- Representation Bias
- Measurement Bias
- Aggregation Bias
- Evaluation Bias
- Deployment Bias

Definitions of fairness

- Fairness through unawareness
- Statistical parity
- Equal opportunity
- Predictive equality

# Recap II

Theme: Thinking about fairness and the limitations of learning as a worldview.

**Concepts**:

Impossibility to achieve all fairness and accuracy

Fairness-accuracy tradeoff

Pareto Frontier

Modeling Spaces

- Construct space
- Observed space
- Decision space

Individual fairness

What You See is What You Get (WYSIWYG)

Structural Bias + We're All Equal (WAE)

Conflicting Worldviews

# Brain Break

# One Slide

Regression

Overfitting

Training, test, and generalization error

Bias-Variance tradeoff

Ridge, LASSO

Cross validation

Gradient descent

Classification

Logistic regression

Bias and Fairness

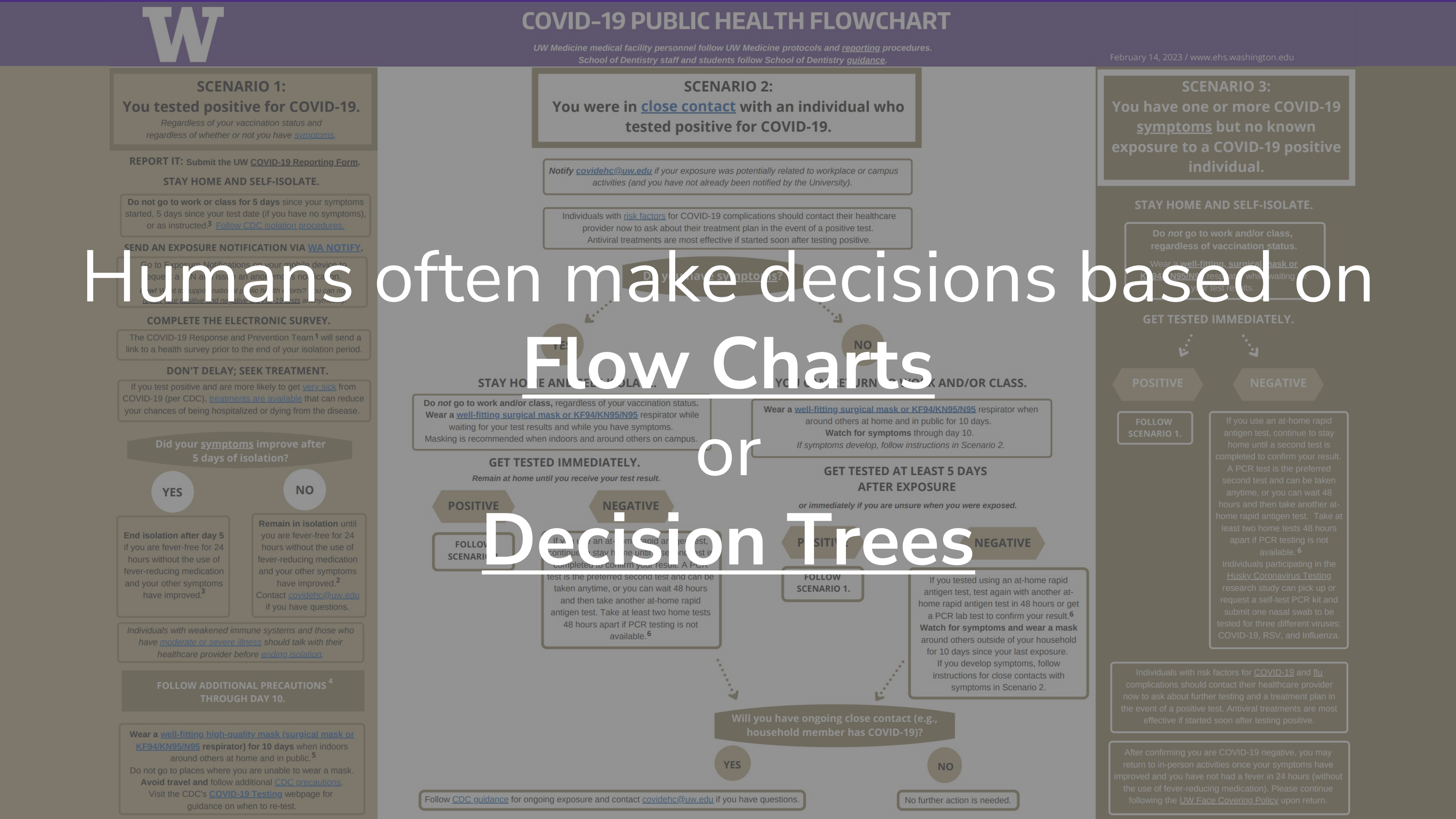
# Decision Trees

# Humans often make decisions based on

# Flow Charts

# or

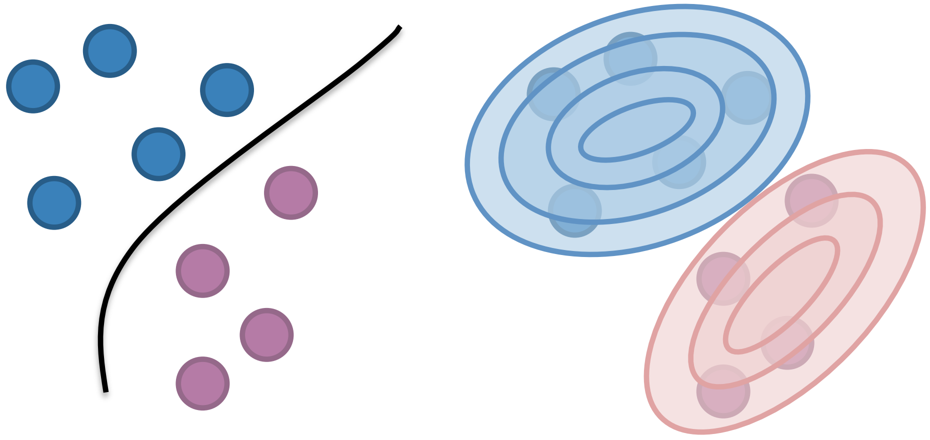
# Decision Trees



# Compare Models

**Generative:** defines a model for generating x (e.g. Naïve Bayes)

**Discriminative:** only cares about defining and optimizing a decision boundary (e.g. Logistic Regression)



# Parametric vs. Non-Parametric Methods

**Parametric Methods**: make assumptions about the data distribution

- Linear Regression ⇒ assume the data is linear
- Logistic Regression ⇒ assume probability has the shape of of a logistic curve and linear decision boundary
- Those assumptions result in a *parameterized* function family. Our learning task is to learn the parameters.

**Non-Parametric Methods**: (mostly) don't make assumptions about the data distribution

- Decision Trees, k-NN (soon)
- We're still learning something, but not the parameters to a function family that we're assuming describes the data.
- Useful when you don't want to (or can't) make assumptions about the data distribution.

# XOR

A line might not always support our decisions.

# What makes a loan risky?

I want to buy a new house!

Loan Application


Credit History
★★★★

Income
★★★

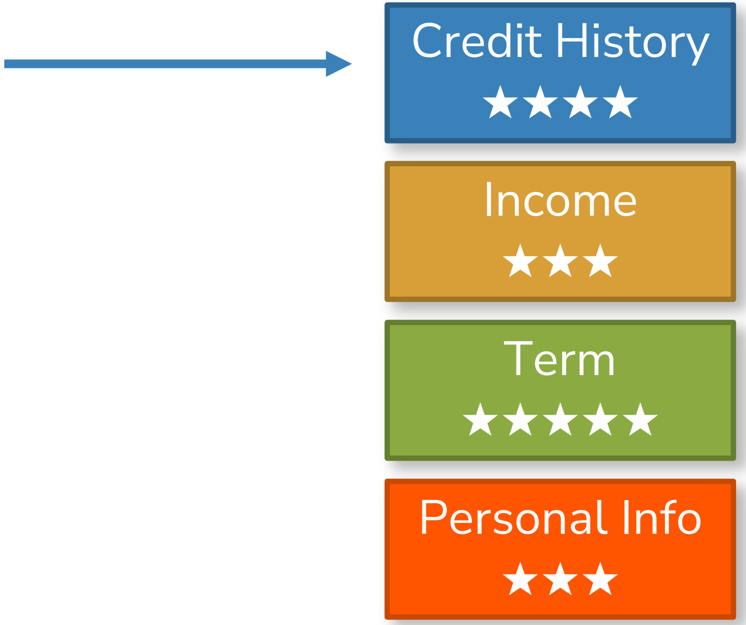
Term
★★★★★

Personal Info
★★★

# Credit history explained

Did I pay previous loans on time?

Example:
excellent, good, or fair



# Income

What's my income?

Example:
$80K per year

Credit History
★★★★

Income
★★★

Term
★★★★★

Personal Info
★★★

# Loan terms

How soon do I need to pay the loan?

Example: 3 years, 5 years,...

Credit History
★★★★

Income
★★★

Term
★★★★★

Personal Info
★★★

# Personal information

Age, reason for the loan, marital status,…

Example: Home loan for a married couple

Credit History
★★★★

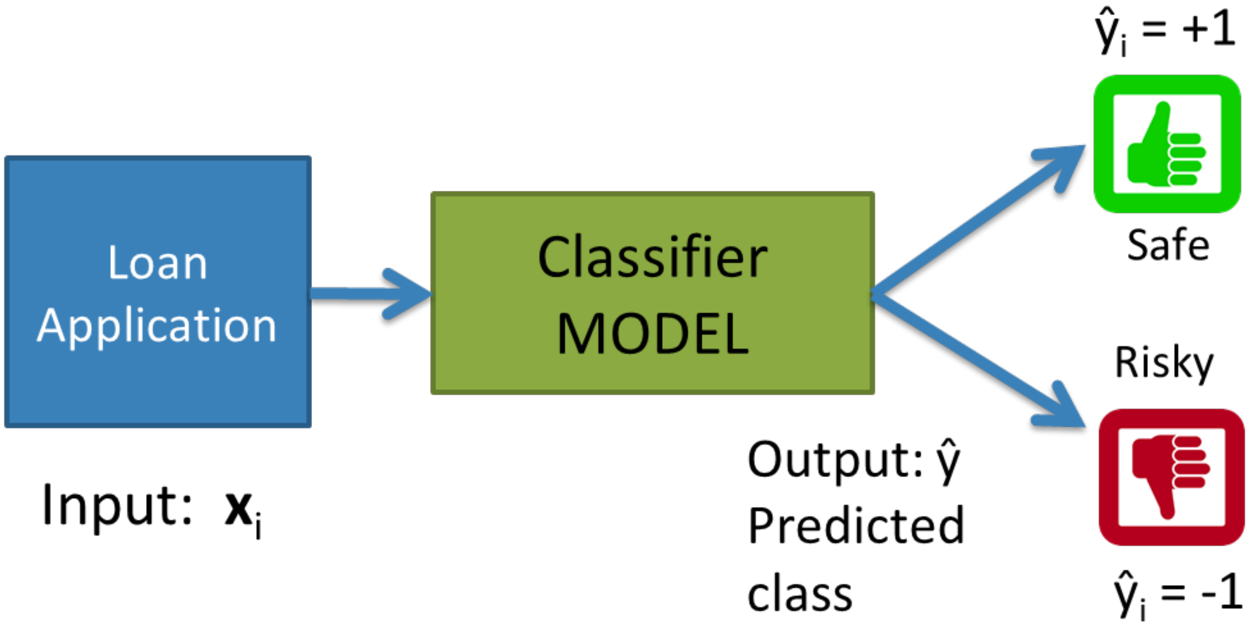
Income
★★★

Term
★★★★★

Personal Info
★★★

# Intelligent application

# Classifier review



# Setup

Data (N observations, 3 features)

| Credit | Term | Income | y |
|---|---|---|---|
| excellent | 3 yrs | high | safe |
| fair | 5 yrs | low | risky |
| fair | 3 yrs | high | safe |
| poor | 5 yrs | high | risky |
| excellent | 3 yrs | low | safe |
| fair | 5 yrs | low | safe |
| poor | 3 yrs | high | risky |
| poor | 5 yrs | low | safe |
| fair | 3 yrs | high | safe |

Evaluation: classification error

Many possible decisions: number of trees grows exponentially!

Poll Everywhere

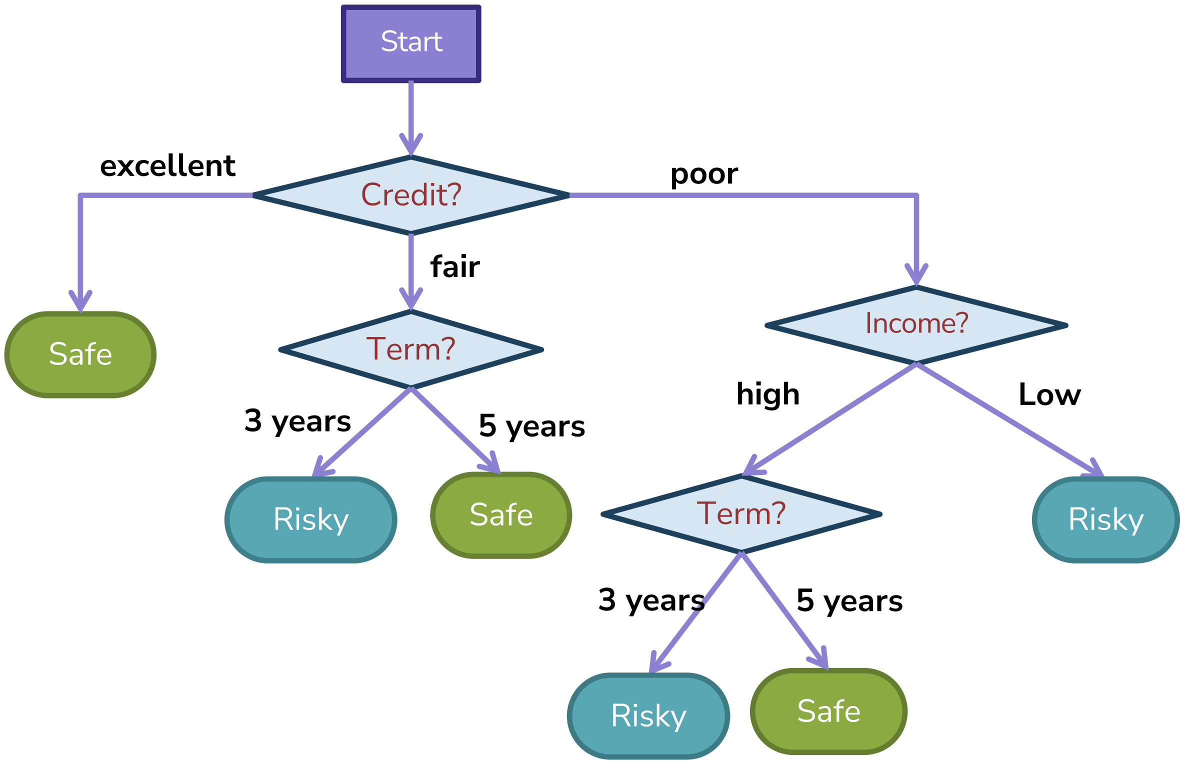
# Think

2 min

With our discussion of bias and fairness from last week, discuss the potential biases and fairness concerns that might be present in our dataset about loan safety.

No poll

# Decision Trees




- **Branch/Internal node:** splits into possible values of a feature
- **Leaf node:** final decision (the class value)

# Next time