Fairness in Machine Learning: Practicum

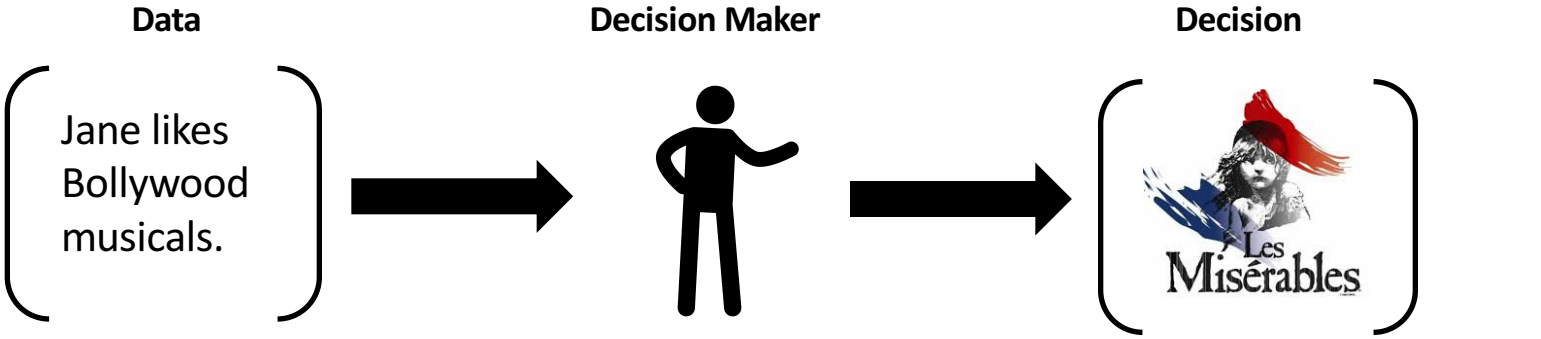
Privacy & Fairness in Data Science

CS848 Fall 2019

UNIVERSITY OF WATERLOO | DSg Data Systems Group

Human Decision Making

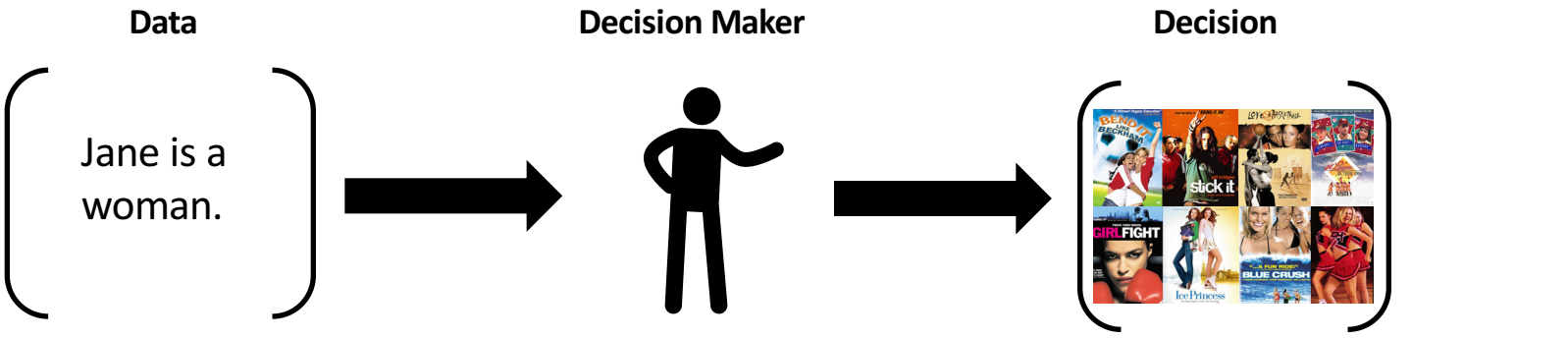
Suppose we want to recommend a movie.



Jane: "Nice try, Bob, but you clearly don't understand how to generalize from your prior experience."

Human Decision Making

Or even worse:



Jane: "Actually Bob, that's a sexist recommendation that doesn't reflect well on you as a person or your understanding of cinema."

What if we use machine learning algorithms instead?
They will generalize well and be less biased, right?

Algorithmic Decision Making

"A blackbox collaborative filtering algorithm suggests you would like this movie."

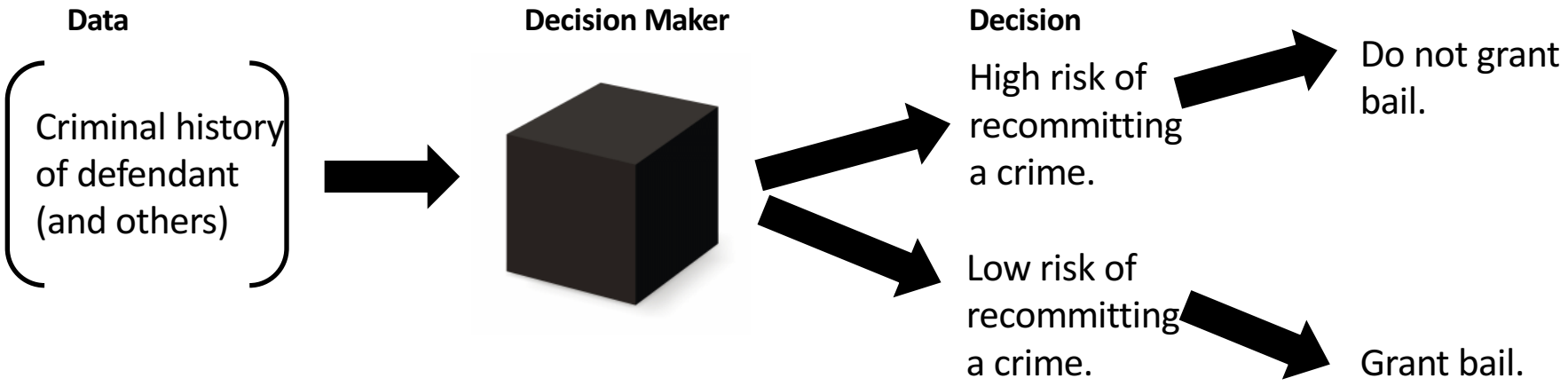
Jane: "Wow Netflix, that was a great recommendation, and you didn't negatively stereotype me in order to generalize from your data!"

Problem solved! Right?

Recidivism Prediction

- In many parts of the U.S., when someone is arrested and accused of a crime, a judge decides whether to grant *bail*.
- In practice, this decides whether a defendant gets to wait for their trial at home or in jail.
- Judges are allowed or even encouraged to make this decision based on how likely a defendant is to re-commit crimes, i.e., recidivate.

Recidivism Prediction



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

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Practicum Activity.

- The ProPublica team studied a proprietary algorithm (COMPAS) and found that it discriminated against African Americans.

- In this activity, **you** will take on the role of the reporters and data analysts looking for discrimination of more standard machine learning algorithms (SVM and Logistic Regression).

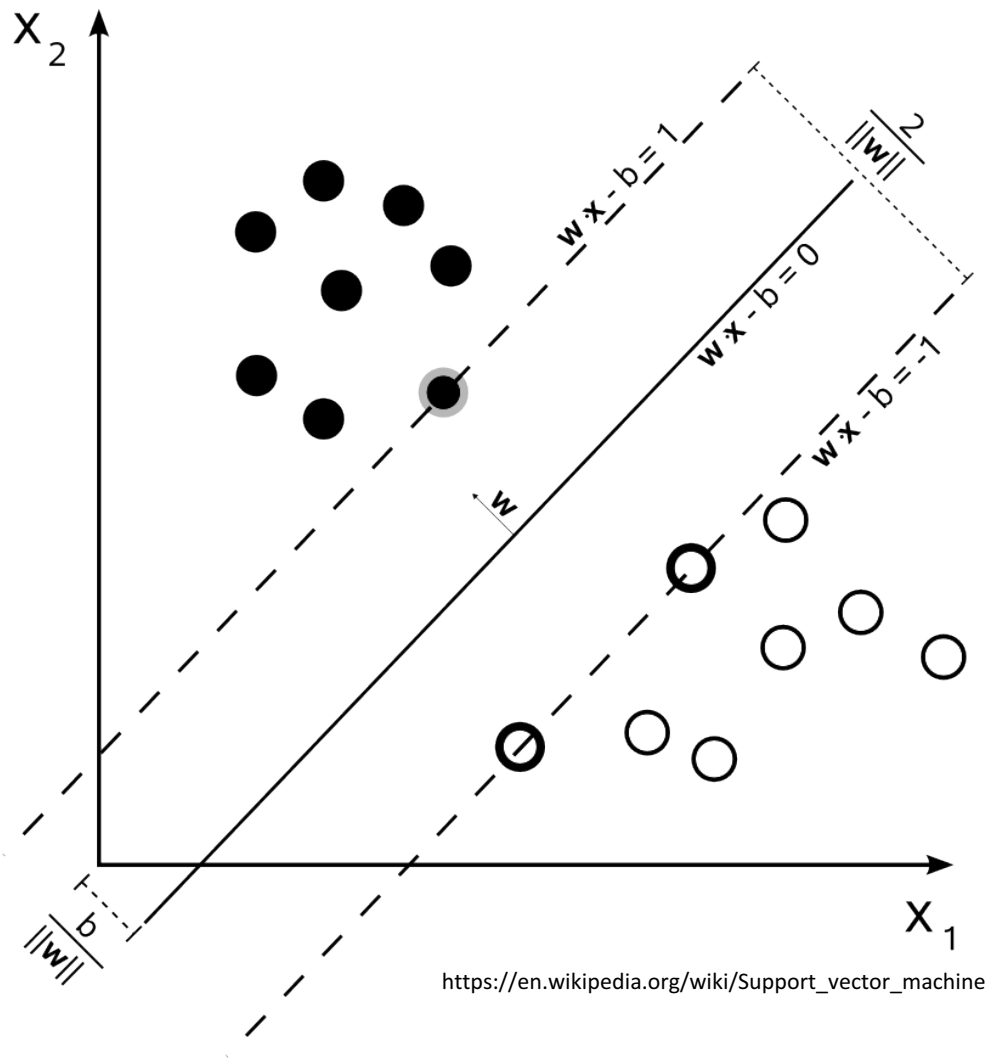
Supervised Learning – Brief Aside

- In supervised learning, we want to make predictions of some target value.
- We are given **training data**, a matrix where every row represents a data point and every column is a **feature**, along with the true target value for every data point.
- What we "learn" is a function from the feature space to the prediction target. E.g., if there are m features, the feature space might be $\mathbb{R}^m$, in which case a **binary classifier** is a function

$$f: \mathbb{R}^m \to \{0, 1\}.$$

Supervised Learning – Brief Aside

- Support vector machines and logistic regression are different algorithms for generating such classifiers, given training data.
- A **support vector machine** (with a linear kernel) just learns a linear function of the feature variables.
- In other words, it defines a **hyperplane** in the feature space, mapping points on one side to 0 and the other side to 1. It chooses the hyperplane that minimizes the **hinge loss**: max(0, distance to hyperplane).
- Visually:



https://en.wikipedia.org/wiki/Support_vector_machine

Supervised Learning – Brief Aside

- **Logistic regression** is used to predict the probabilities of binary outcomes. We can convert it to a classifier by choosing the more likely outcome, for example.
- Let $\vec{x}$ be the independent variables for an individual for whom the target value is 1 with probability p($\vec{x}$).
- Logistic regression assumes $\log \frac{p(\vec{x})}{1-p(\vec{x})}$ is a linear function of $\vec{x}$, and then computes the best linear function using maximum likelihood estimation.

Practicum Activity.

Break into groups of 3. Download the activity from the website (it's a Jupyter notebook). Think creatively and have fun!

Debrief Practicum Activity.

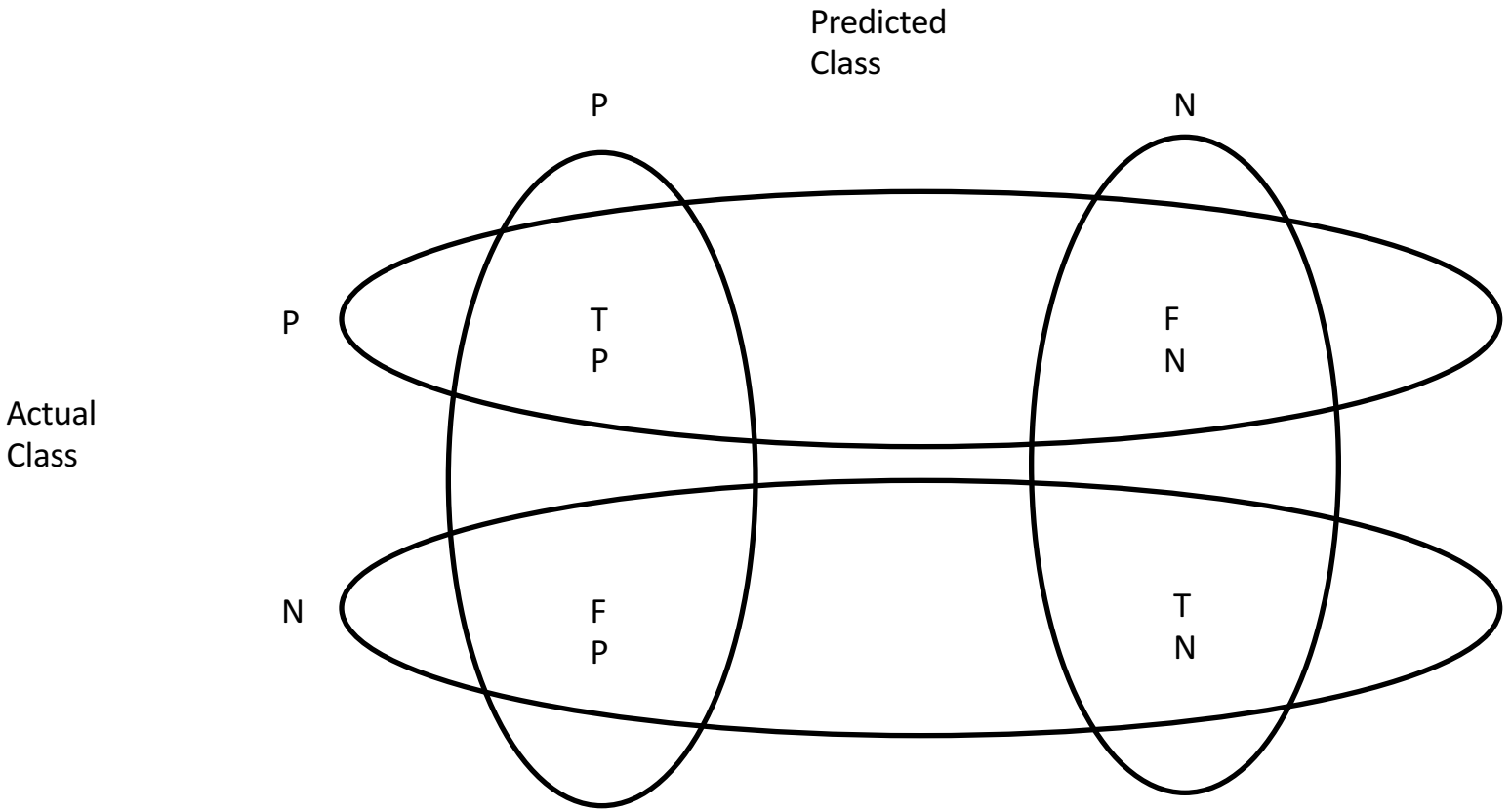
- What arguments did you find for the algorithm(s) being racially biased / unfair?
- What arguments did you find for the algorithm(s) **not** being racially biased / unfair?
- Is one of the algorithms more unfair than the other? Why? How would you summarize the difference between the algorithms?
- Can an algorithm simultaneously achieve high accuracy and be fair and unbiased on this dataset? Why or why not, and with what measures of bias or fairness?

Debrief Practicum Activity – Confusion Matrix.

- A common tool for analyzing binary prediction is the **confusion matrix.**

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Debrief Practicum Activity – Confusion Matrix.



Debrief Practicum Activity – False Positive Rate

- The **false positive rate** is measured as

$$FPR = \frac{FP}{FP+TN}$$

In other words: What % of people did we predict would recommit a crime, although in actuality they won't? (perfect classifier gets 0)

	Race 0	Race 1
SVM	0.137	0.094
LR	0.214	0.136

Debrief Practicum Activity – False Positive Rate

- The false positive rate for race 0 is roughly 1.45 times higher than for race 1 using SVM, and 1.57 using LR!
- But they had the same accuracy; how is this possible?
  - Our classifiers tend to make more false positive mistakes for race 0, and more false negative mistakes for race 1.
  - The “accuracy” of the mechanism is indifferent to this difference, but the defendants surely are not!
- **Given** that you will **not** recidivate and are of the protected race, the algorithm looks unfair. Logistic regression (which was slightly more accurate overall) seems slightly worse.

Debrief Practicum Activity – Positive Predictive Value

- The **positive predictive value** is measured as

$$\boldsymbol{PPV} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}}$$

In other words: What % of the people we predicted would recidivate really do recidivate? (perfect classifier gets 1)

	Race 0	Race 1
SVM	0.753	0.686
LR	0.725	0.658

Debrief Practicum Activity – Positive Predictive Value

- By this measure, the algorithms differ on the two racial groups by no more than a factor of 1.1, and seem roughly fair. If anything, the rate is better for the protected group!
- Why doesn't this contradict the false positive finding?
  - Suppose for race 0 we have 100 individuals, 50 of whom recidivate.
  - Suppose for race 1 we have 100 individuals, 20 of whom recidivate.
  - Suppose we make exactly 5 false positives for each racial group, and get everything else correct. Then:
  - $FPR_0 = 5/50 = 0.1$ whereas $FPR_1 = 5/80 = 0.0625$
  - But $PPV_0 = 50/55 = 0.909$ whereas $PPR_1 = 20/25 = 0.8$
- **Given** that the algorithm predicts you will recidivate, it looks roughly fair, logistic regression maybe more so.

Debrief Practicum Activity – Disparate Impact.

- Let $Prob_0$ be the fraction of racial group 0 that we predicted would recidivate (similarly for $Prob_1$).
- The **disparate impact** is measured as

$$\boldsymbol{DI = \frac{Prob_0}{Prob_1}}$$

In other words: How much more (or less) likely were we to predict that an individual of racial group 0 would recidivate vs. racial group 0? (Note that the perfect classifier may not get 1!)

Debrief Practicum Activity – Disparate Impact.

	Race 0	Race 1
SVM	0.284	0.182
LR	0.400	0.242

- So the disparate impact of SVM is 1.56, and for LR is 1.65.
- **Given** that you are a member of racial group 0, the algorithm looks unfair, more so for LR.
- Note that you **can't** get a small disparate impact **and** a high accuracy, in general. This measure is particularly useful if you think the **data itself are biased**.

Debrief Practicum Activity – Conclusion.

- Machine learning algorithms are often black box optimizations without obvious interpretations ex post.
- There is no single perspective on fairness. What looks fair conditioned on some things may look different conditioned on other things.
- Next time, we will dive into these topics in more depth, focusing especially on disparate impact.

Research Project!!!

- Choosing project (before next class, Oct 1 noon)
  - Form team of 1-3
  - Choose a topic
  - Upload a pdf (1 short paragraph of topic and team members)