Predictive Algorithms for Judges

Giulia Lasagni
University of Bologna

Marcello Di Bello
Arizona State University

AI for Judges

Agenda

PART I: Introduction to Predictive Algorithms

PART II: Examples in Criminal Justice

PART III: Controversies

PART IV: Possible Remedies

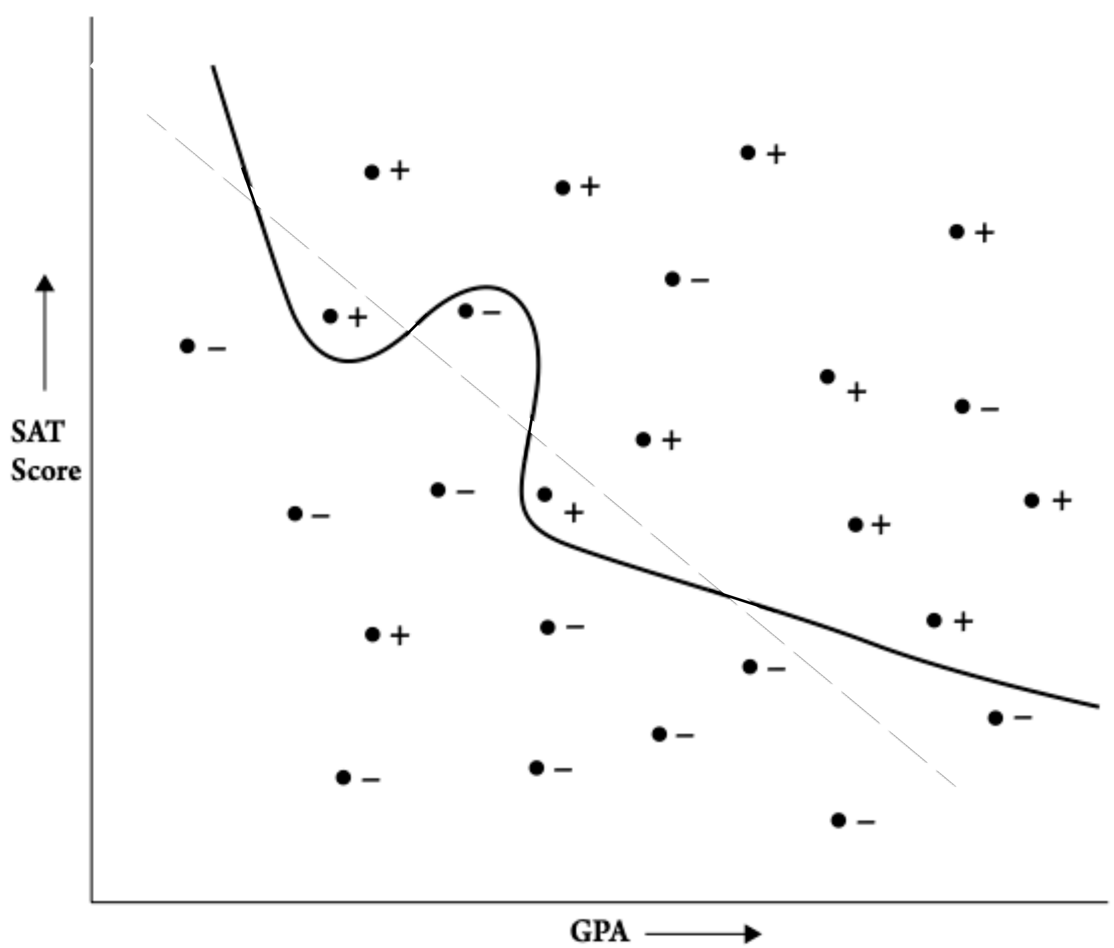
PART V: Impossibility Theorems (time permitting)

Part I

Introduction to Predictive Algorithms (or Predictive Models)

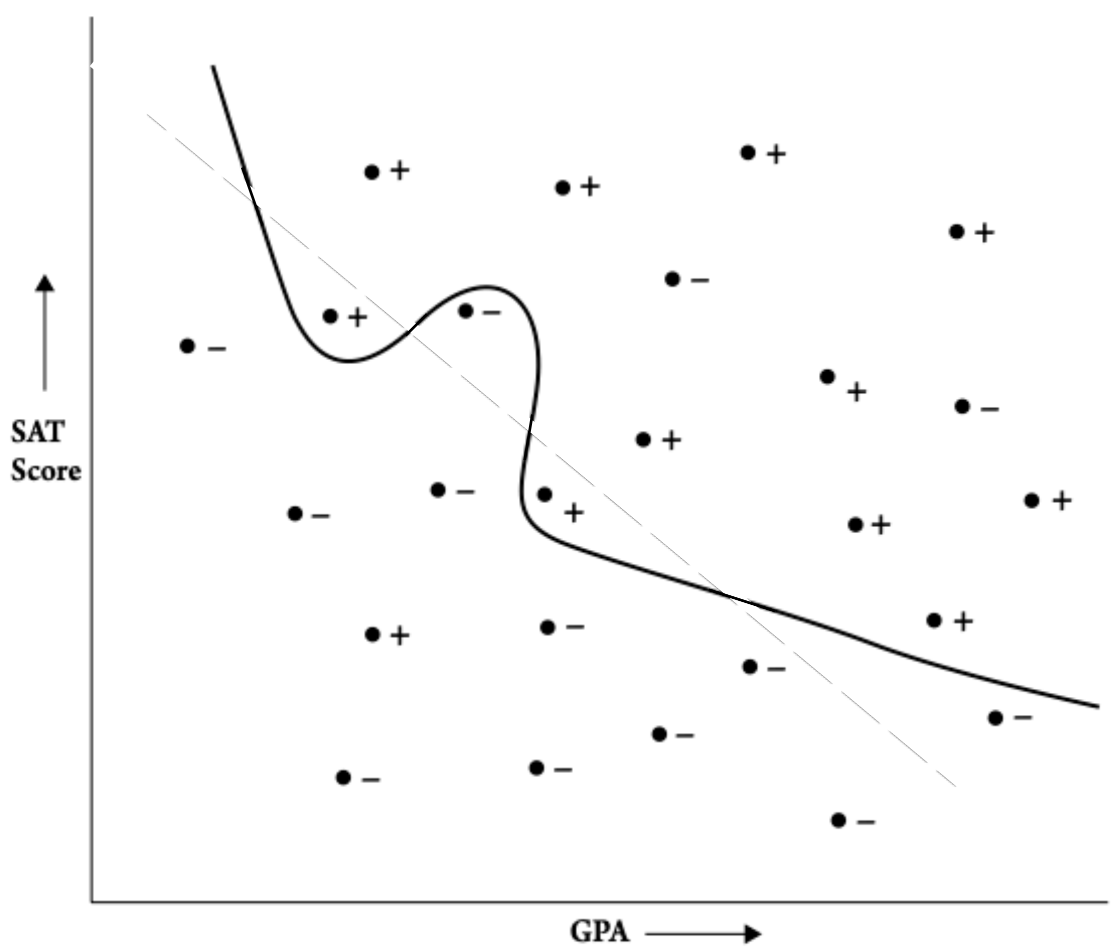
Predictive Algorithms (or Predictive Models)

(*binary case*)



Predictive Algorithms (or Predictive Models)

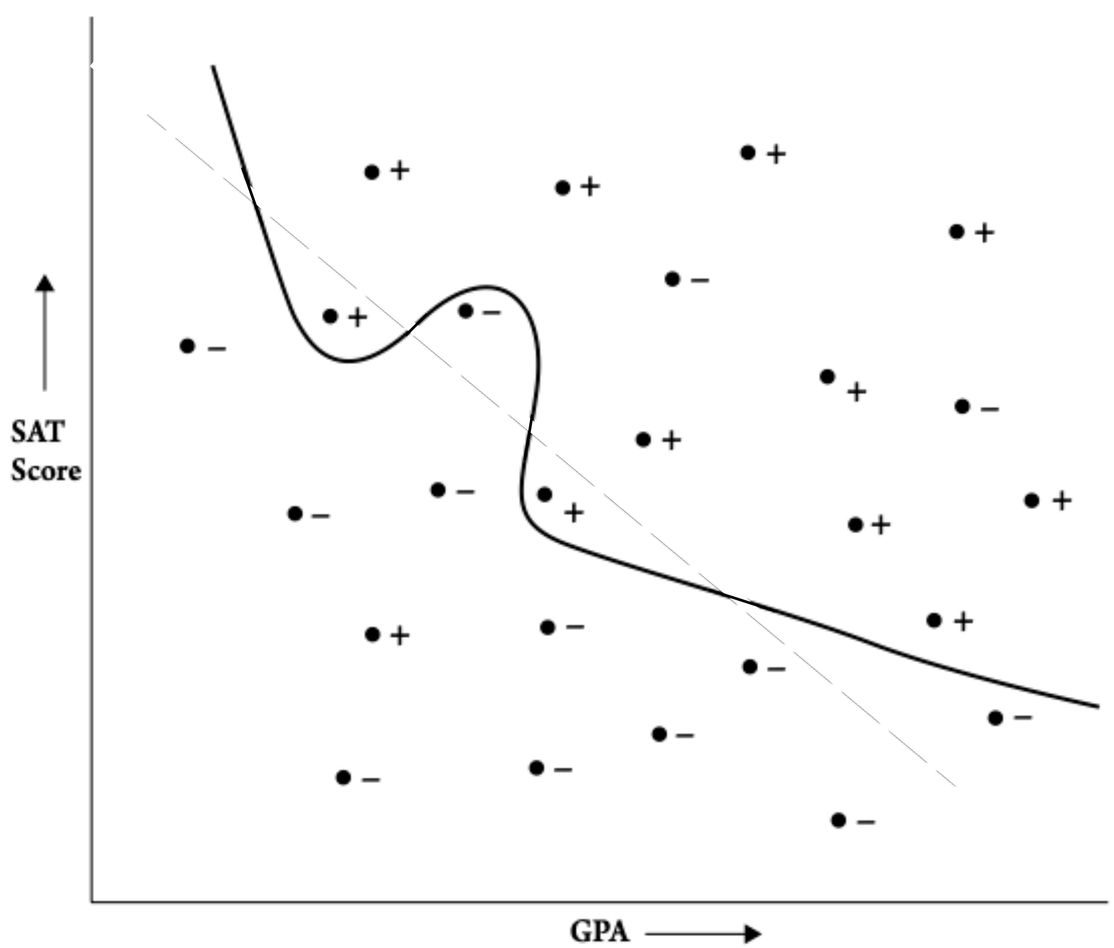
(*binary case*)



Suppose we aim to make predictions about a **binary outcome Y=1** or **Y=0** (e.g. college success, recidivism)

Predictive Algorithms (or Predictive Models)

(*binary case*)

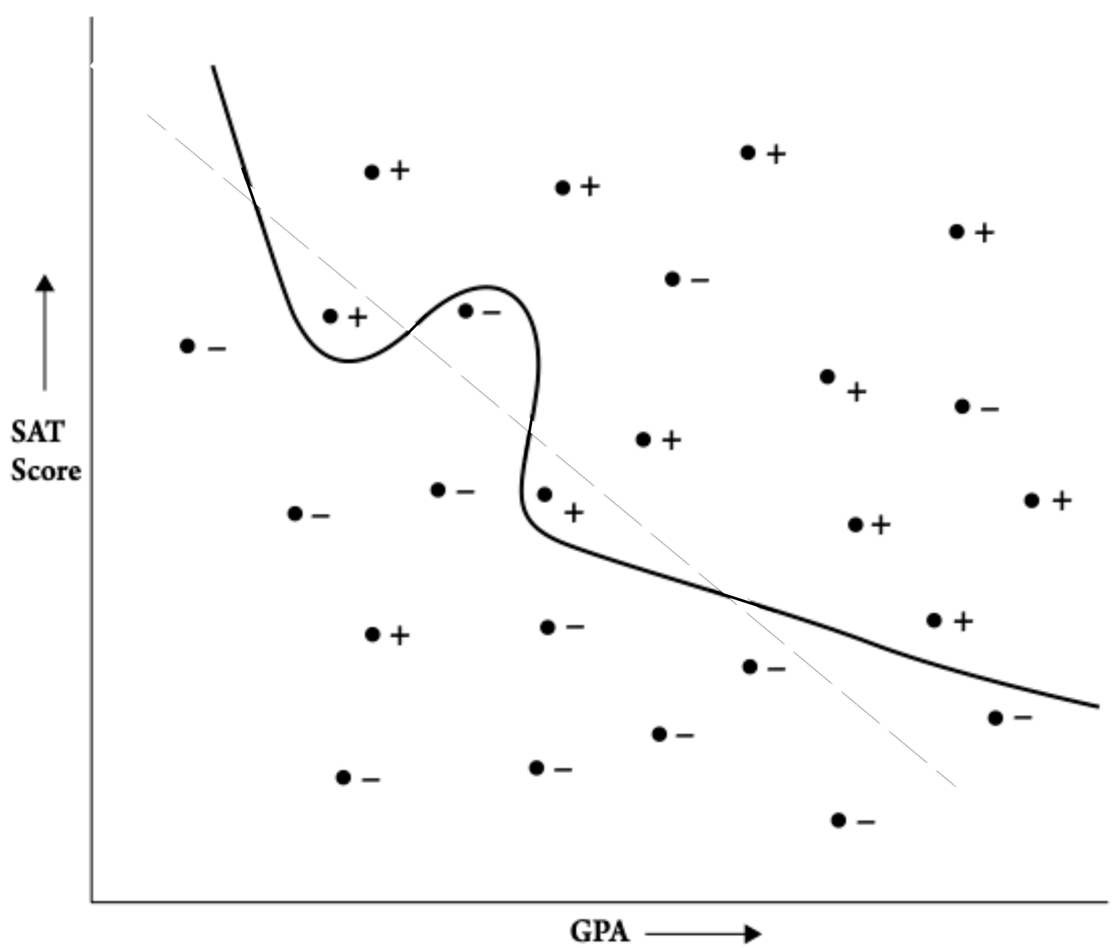


Suppose we aim to make predictions about a **binary outcome Y=1** or **Y=0** (e.g. college success, recidivism)

Machine learning algorithms (e.g. regression, SVM) mine the historical data and identify relationships between **predictive features** (e.g. GPA, income) and the outcome

Predictive Algorithms (or Predictive Models)

(*binary case*)

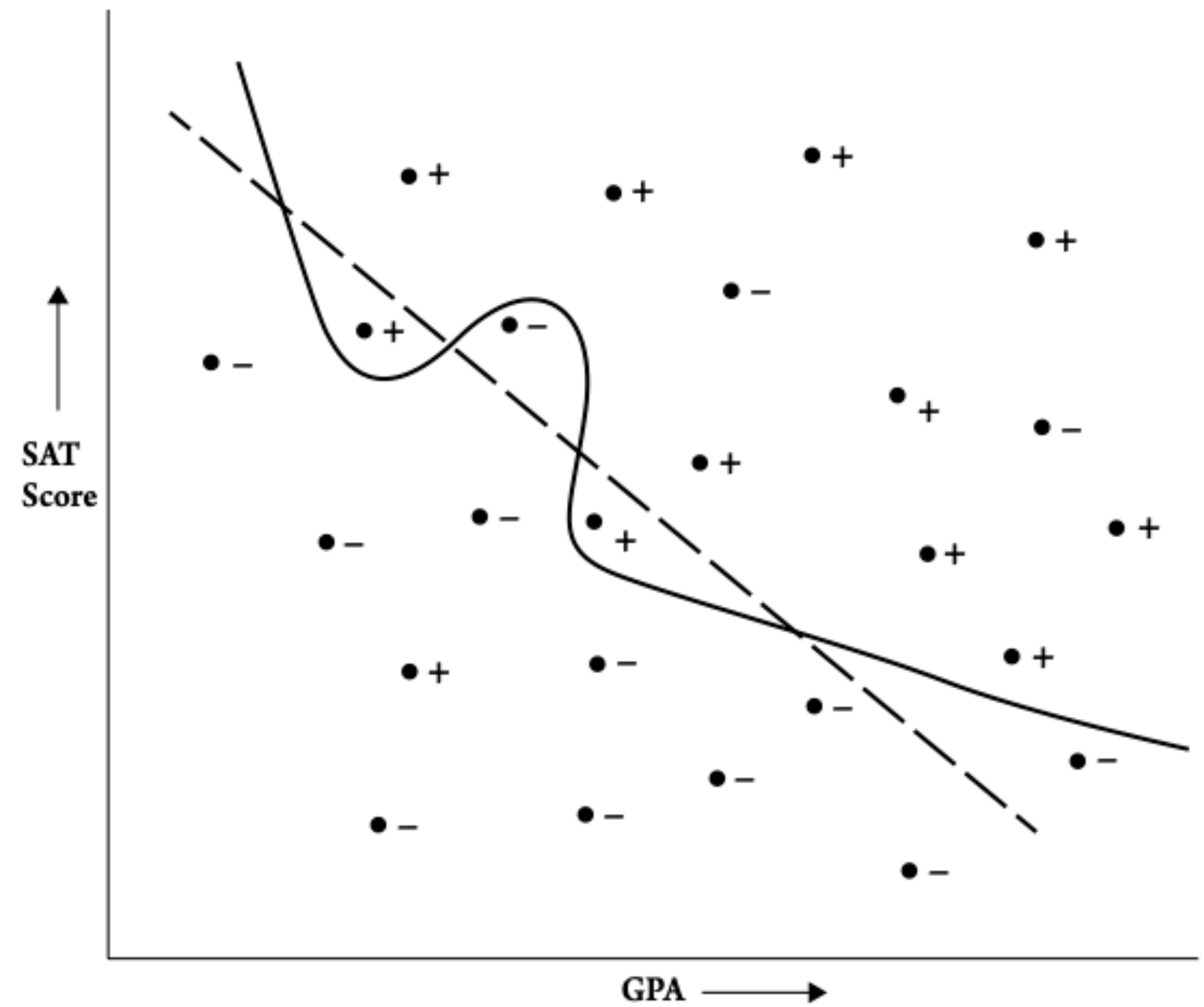


Suppose we aim to make predictions about a **binary outcome Y=1** or **Y=0** (e.g. college success, recidivism)

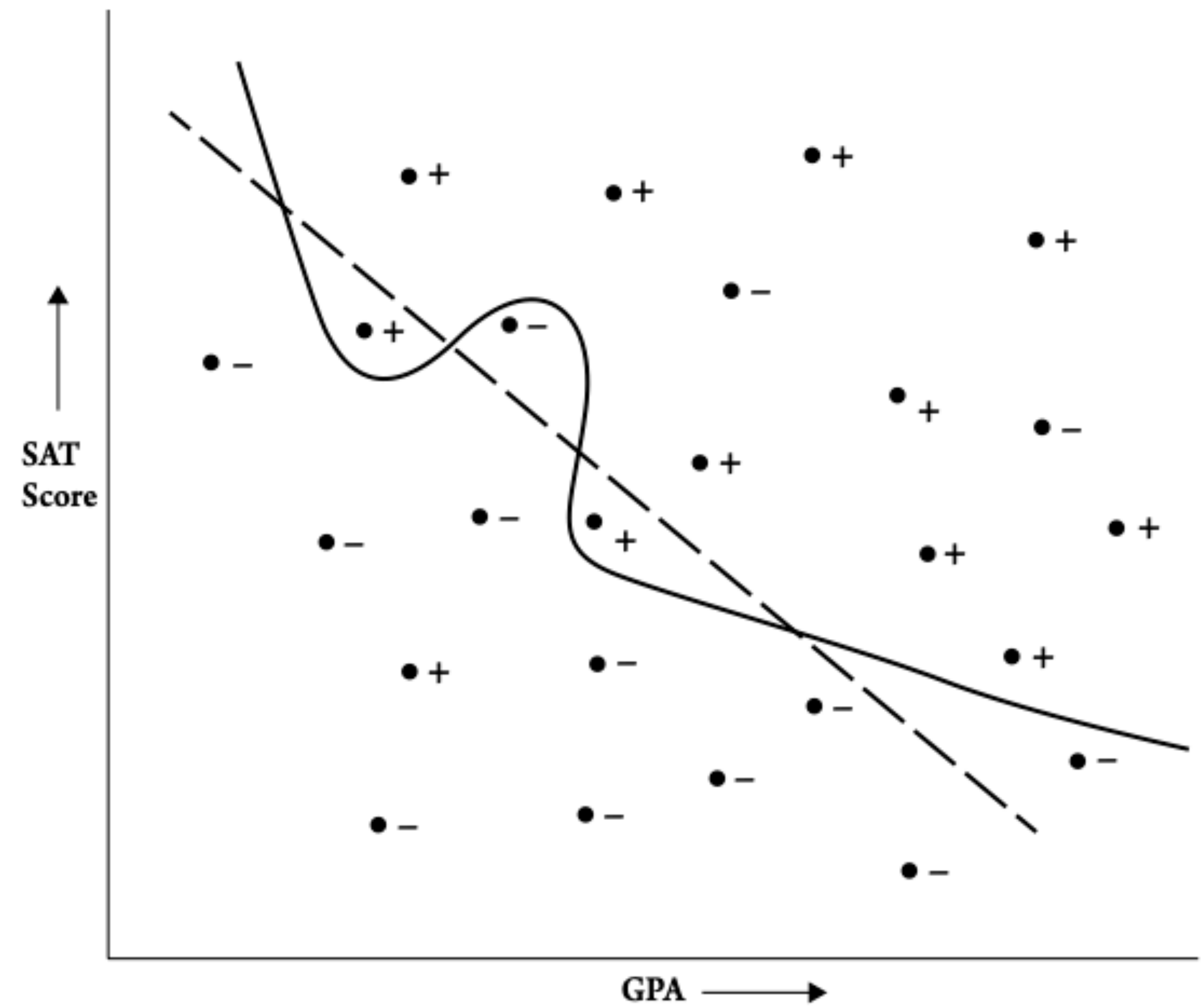
Machine learning algorithms (e.g. regression, SVM) mine the historical data and identify relationships between **predictive features** (e.g. GPA, income) and the outcome

Based on the features one possesses, the **predictive model classifies** individuals as **C=1** or **C=0**

Machine Learning Algorithms v. Predictive Algorithms (or Predictive Models)

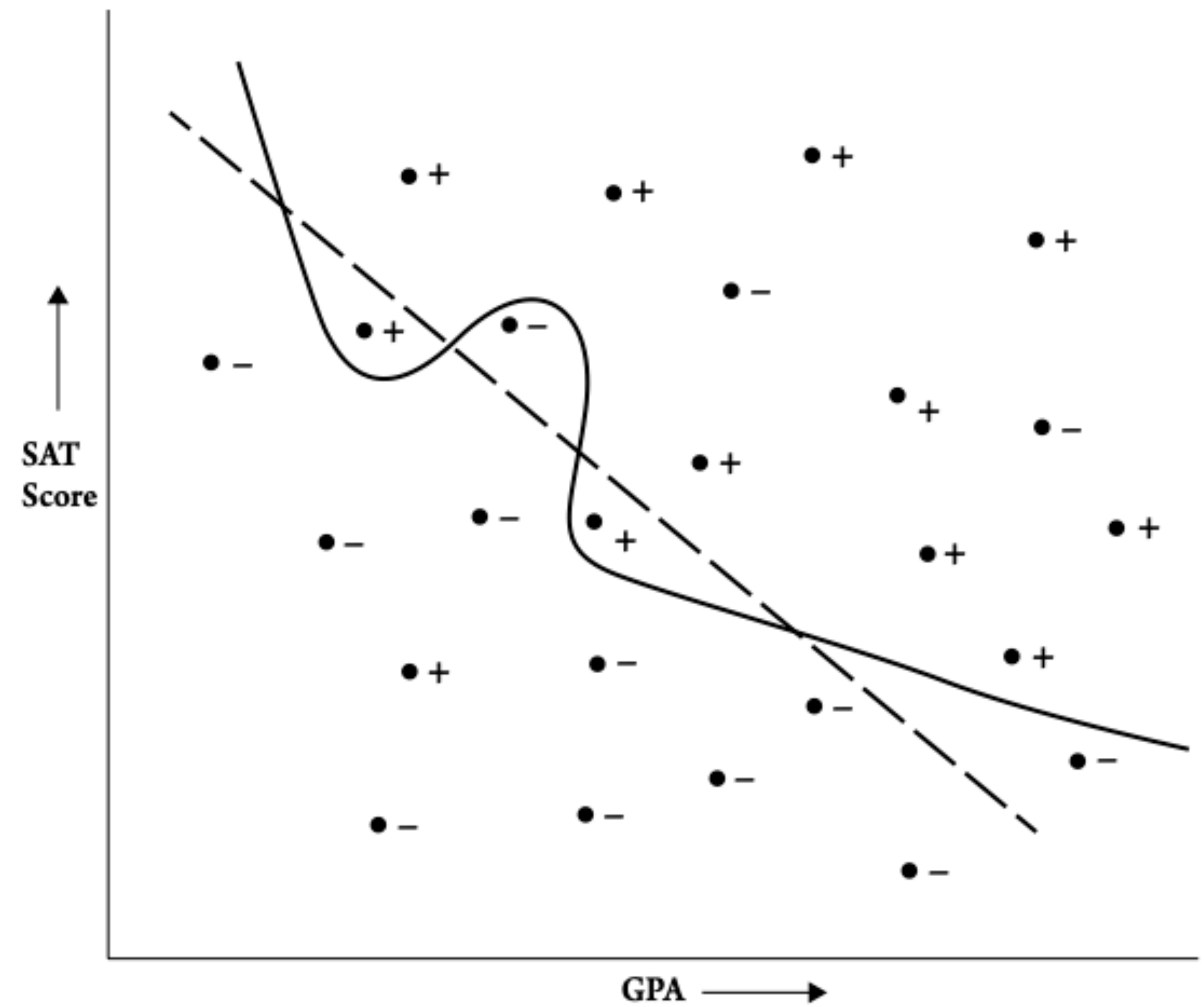

SAT Score
GPA

Machine Learning algorithms are self-programming.



Machine Learning algorithms are self-programming.

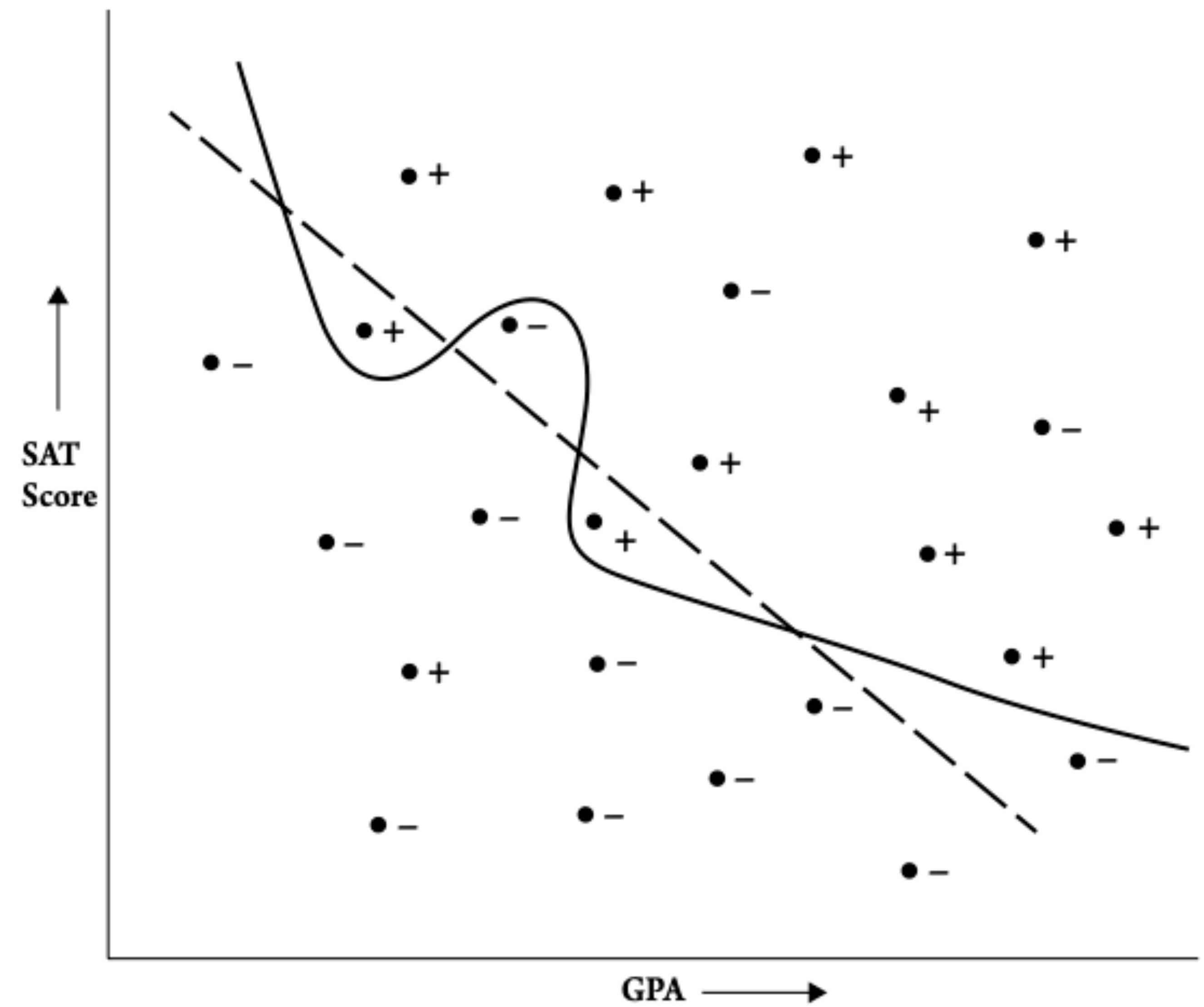
They are **meta-algorithms** whose input are historical data and whose output is another algorithm.

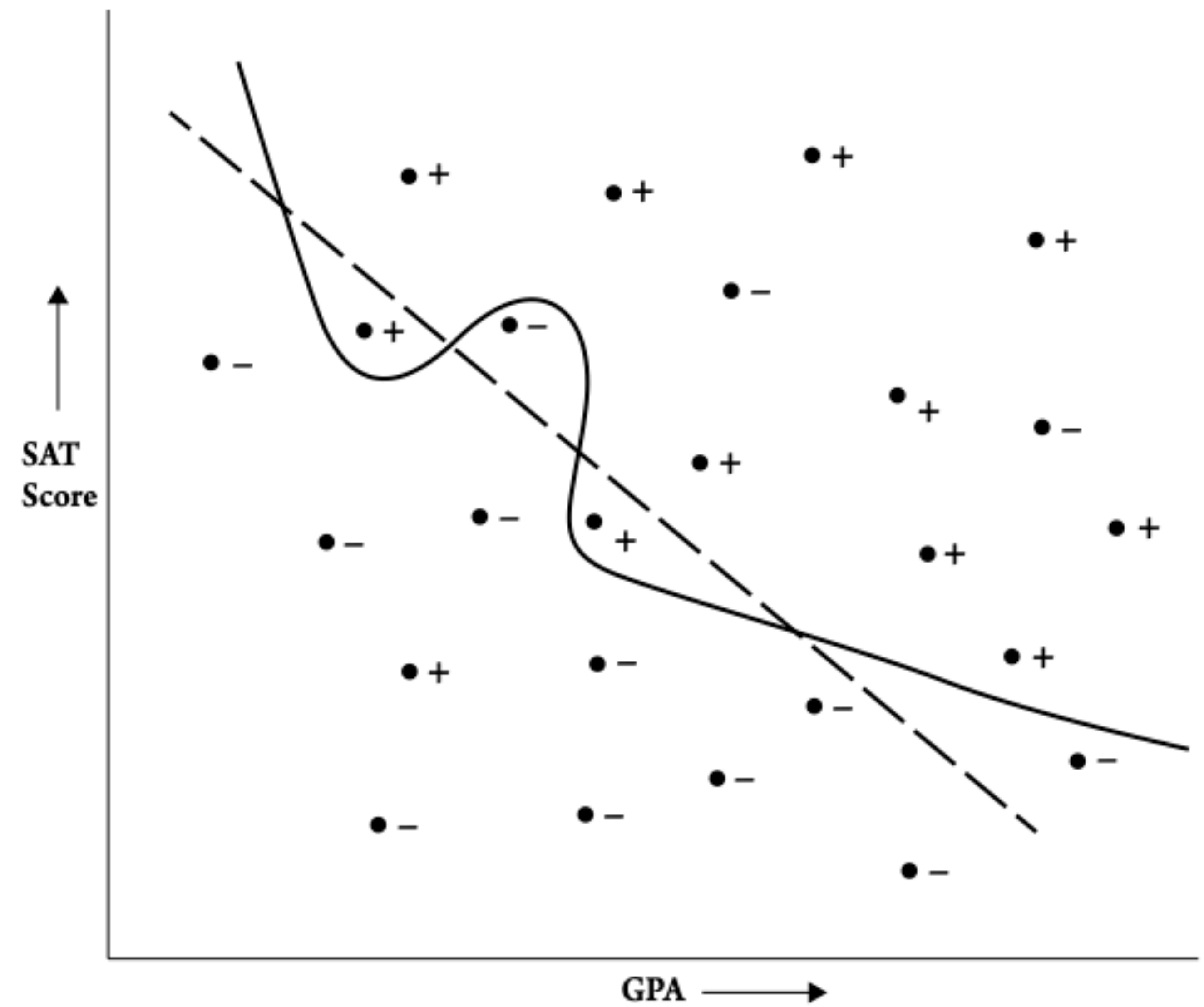


Machine Learning algorithms are self-programming.

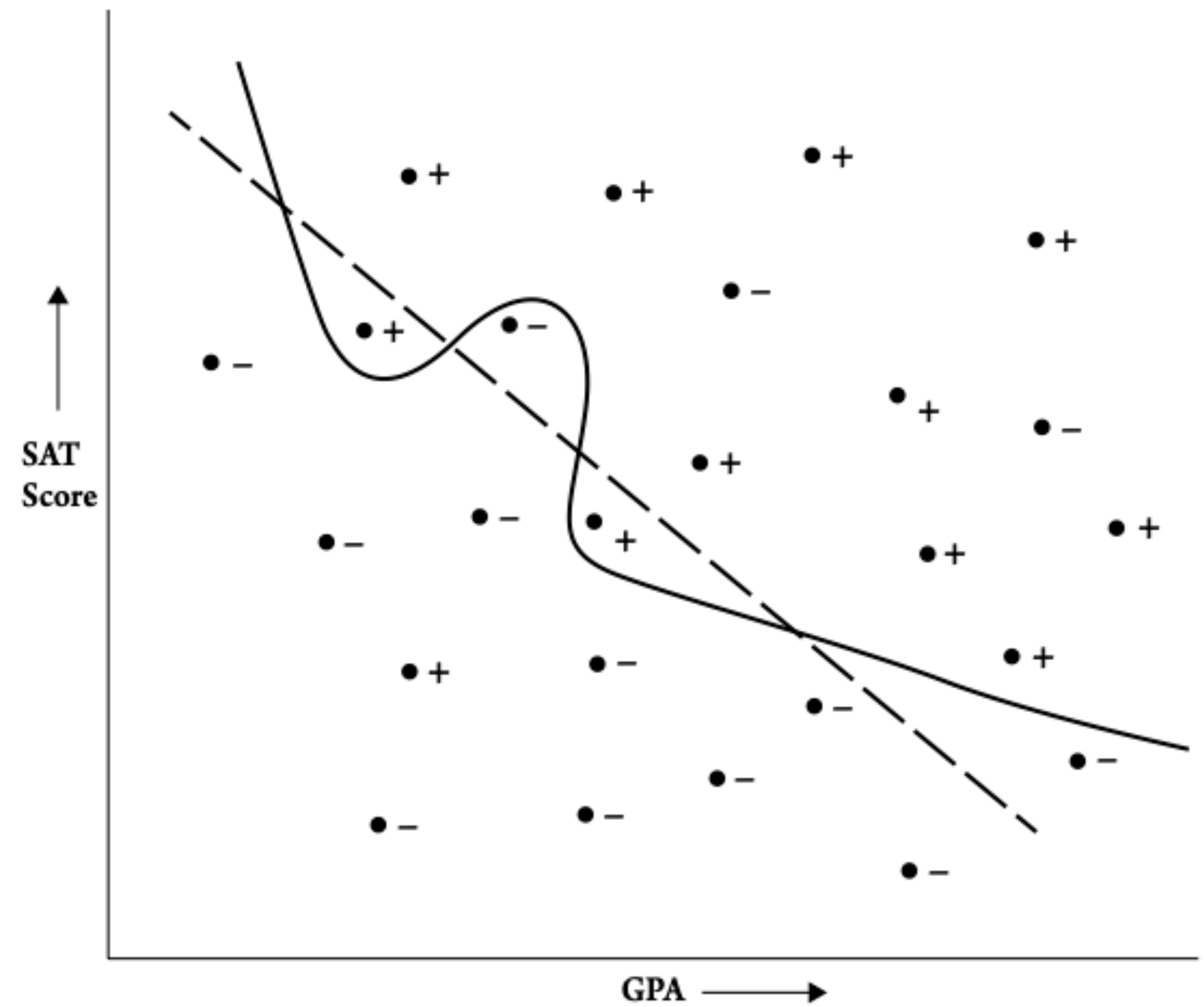
They are **meta-algorithms** whose input are historical data and whose output is another algorithm.

Self-programming? This is less fancy than it sounds. We are talking about minimizing a (very complicated) cost function. It's calculus.



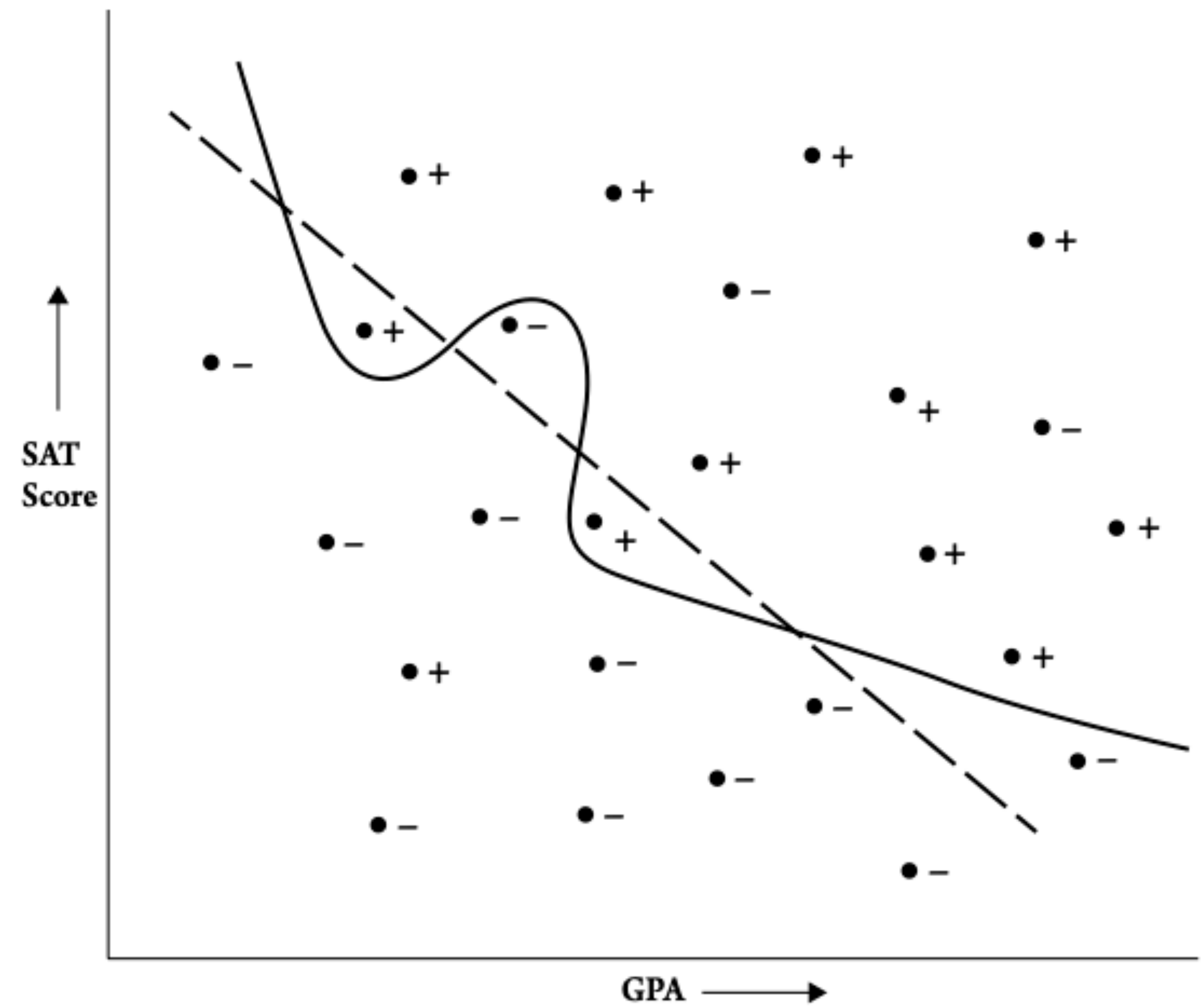

SAT
Score
GPA

The meta-algorithm searches through all models, say possibles lines through the data.



The meta-algorithm searches through all models, say possibles lines through the data.

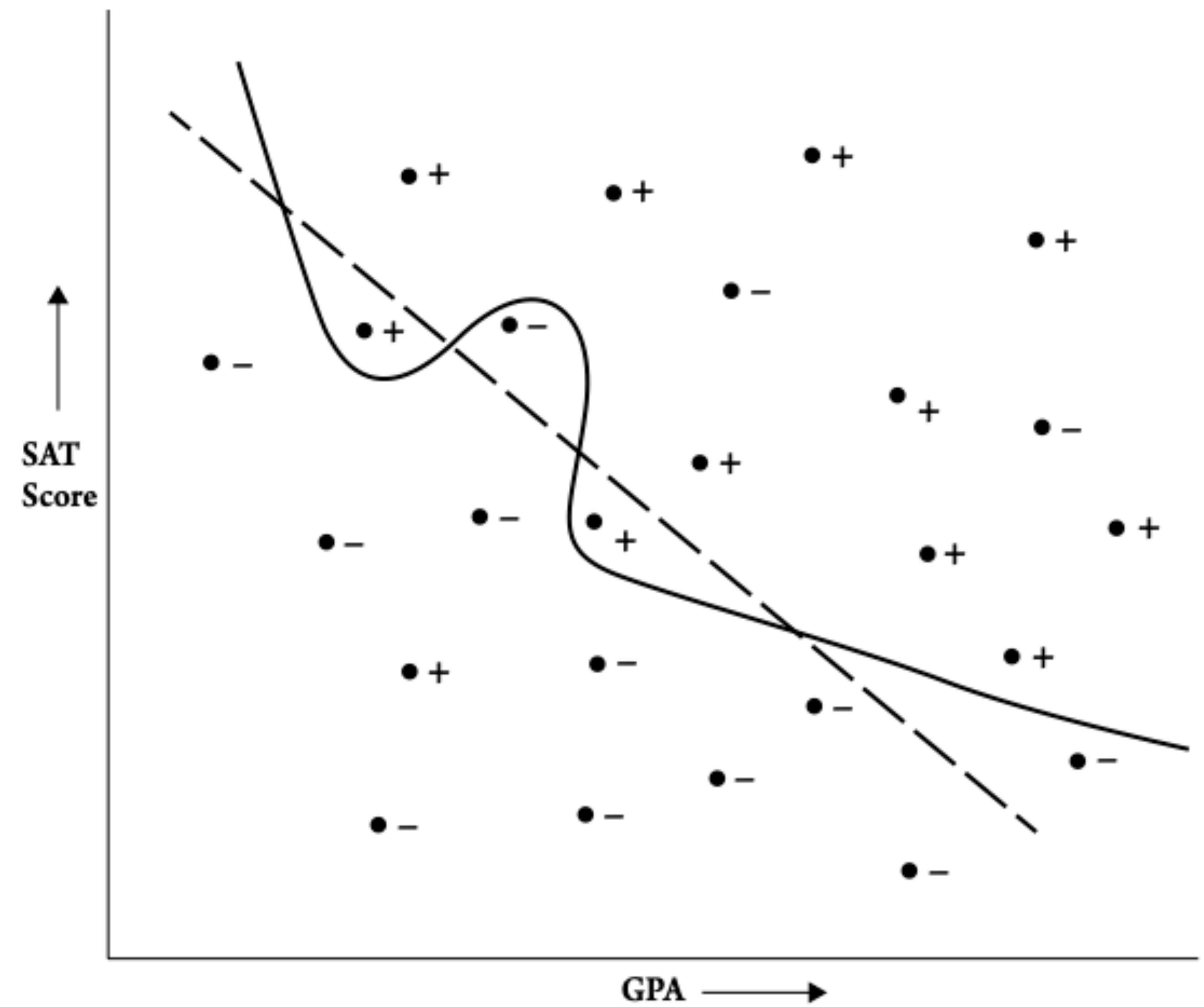
Lines are good for 2-dimensional data (e.g. SAT and GPA) and a binary outcome (graduate/not graduate).



The meta-algorithm searches through all models, say possibles lines through the data.

Lines are good for 2-dimensional data (e.g. SAT and GPA) and a binary outcome (graduate/not graduate).

By a process of optimization, the meta-algorithm selects the predictive model (first-order algorithm) that minimizes errors.

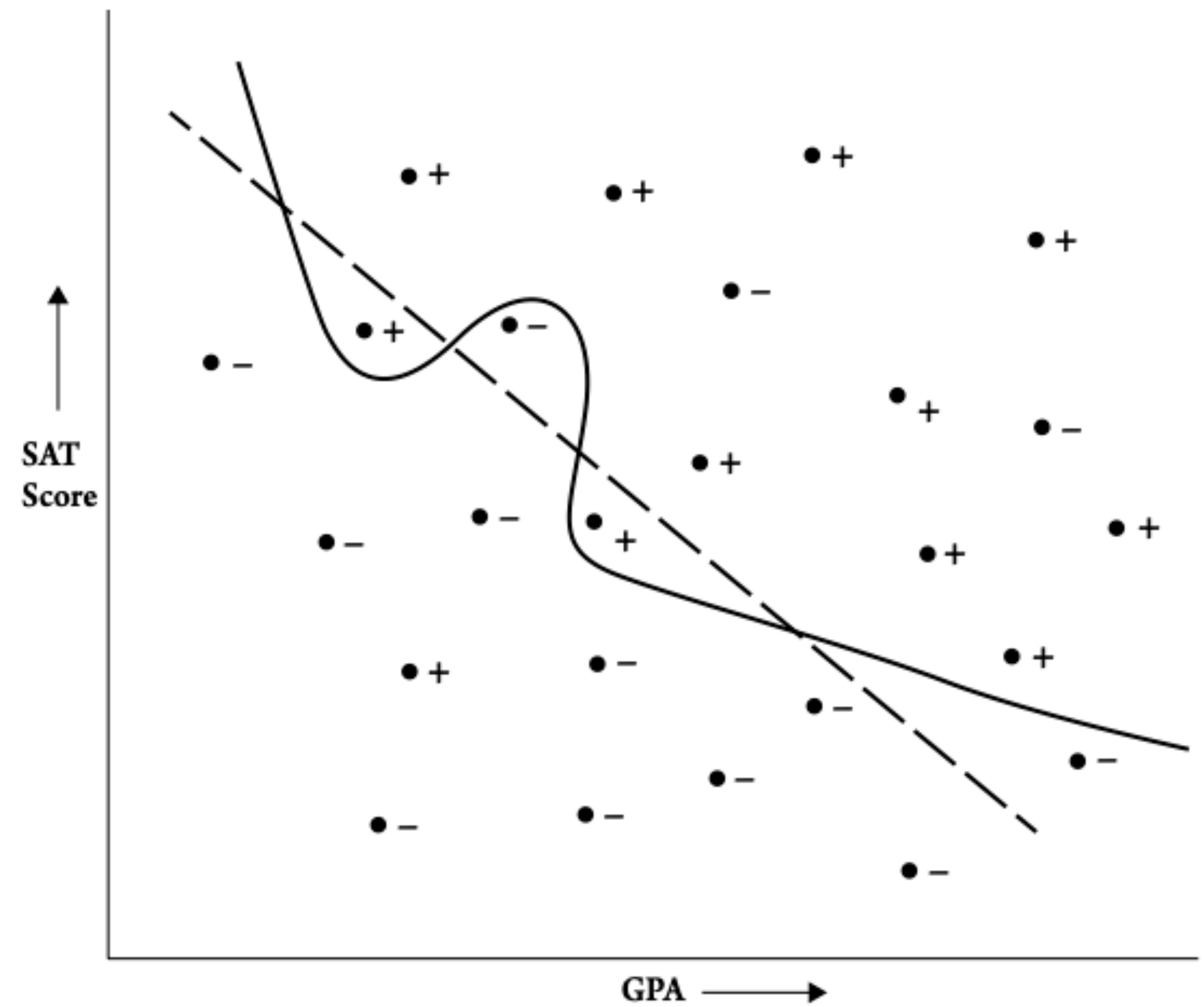


The meta-algorithm searches through all models, say possibles lines through the data.

Lines are good for 2-dimensional data (e.g. SAT and GPA) and a binary outcome (graduate/not graduate).

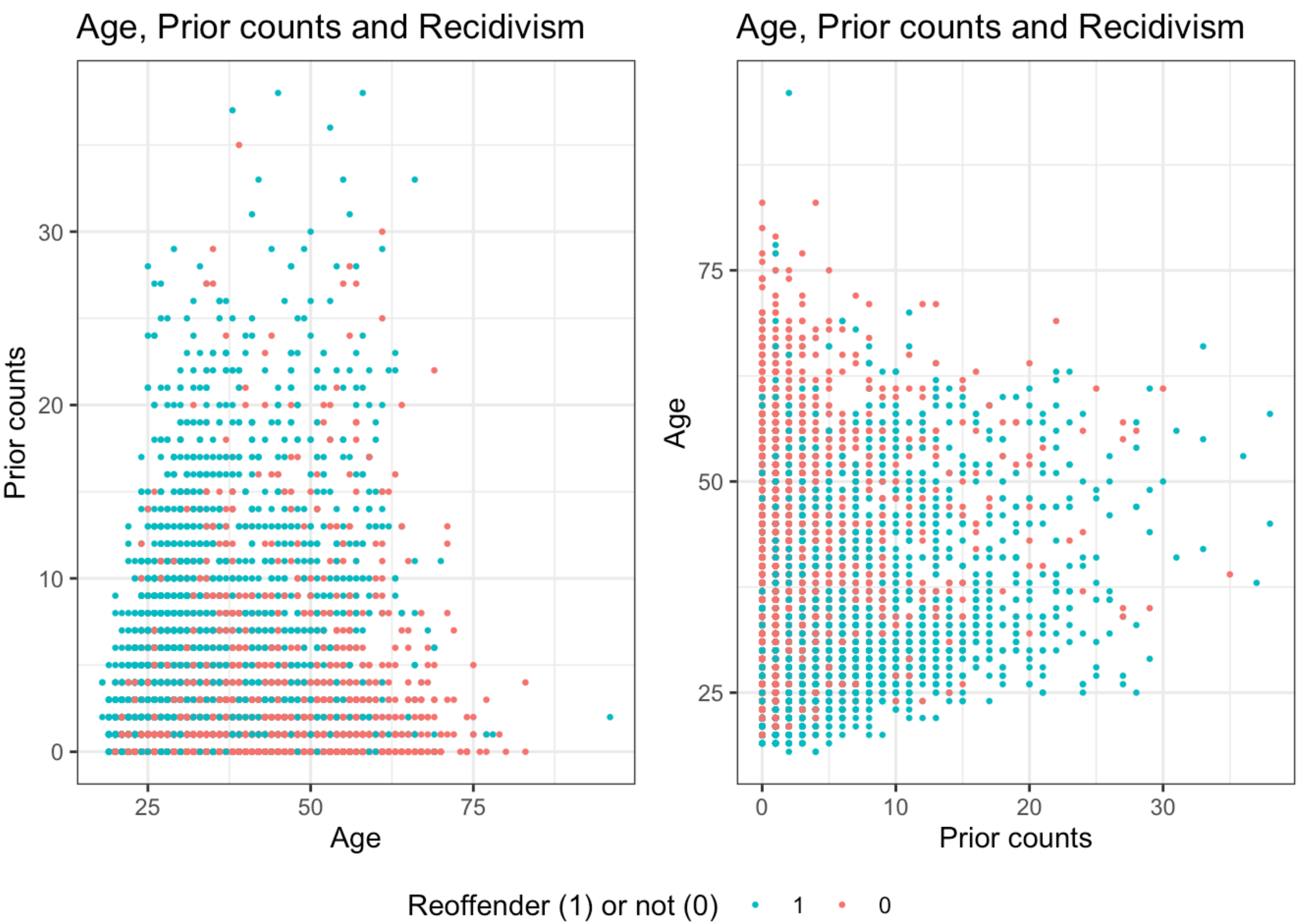
By a process of optimization, the meta-algorithm selects the predictive model (first-order algorithm) that minimizes errors.

This is an example of ***supervised learning***. The model learns by comparing its prediction with the actual outcome in the training data.



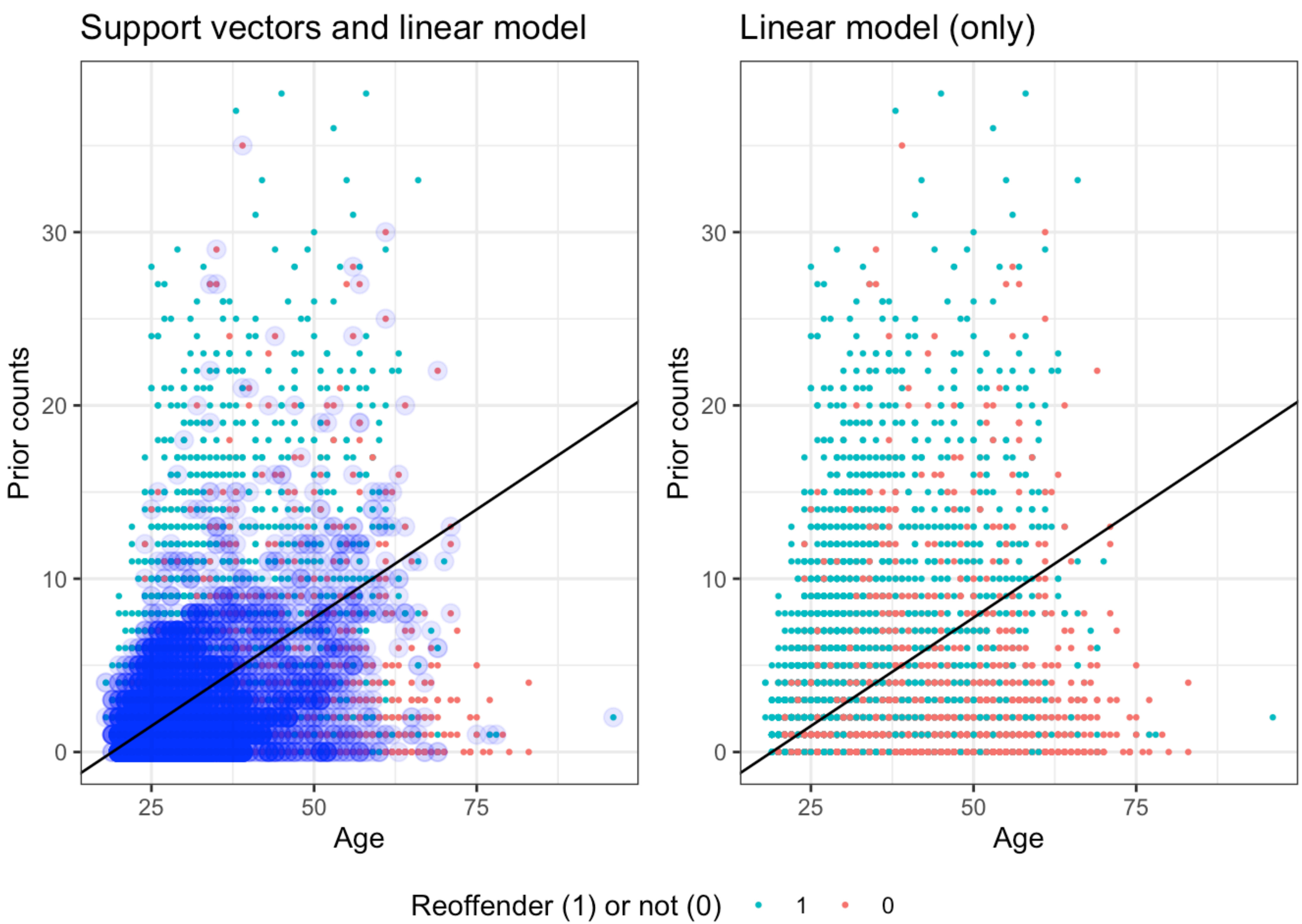
Toy Example
Support Vector Machine (SVM)
Algorithm

Historical Data: Age, Prior Counts, Reoffeding



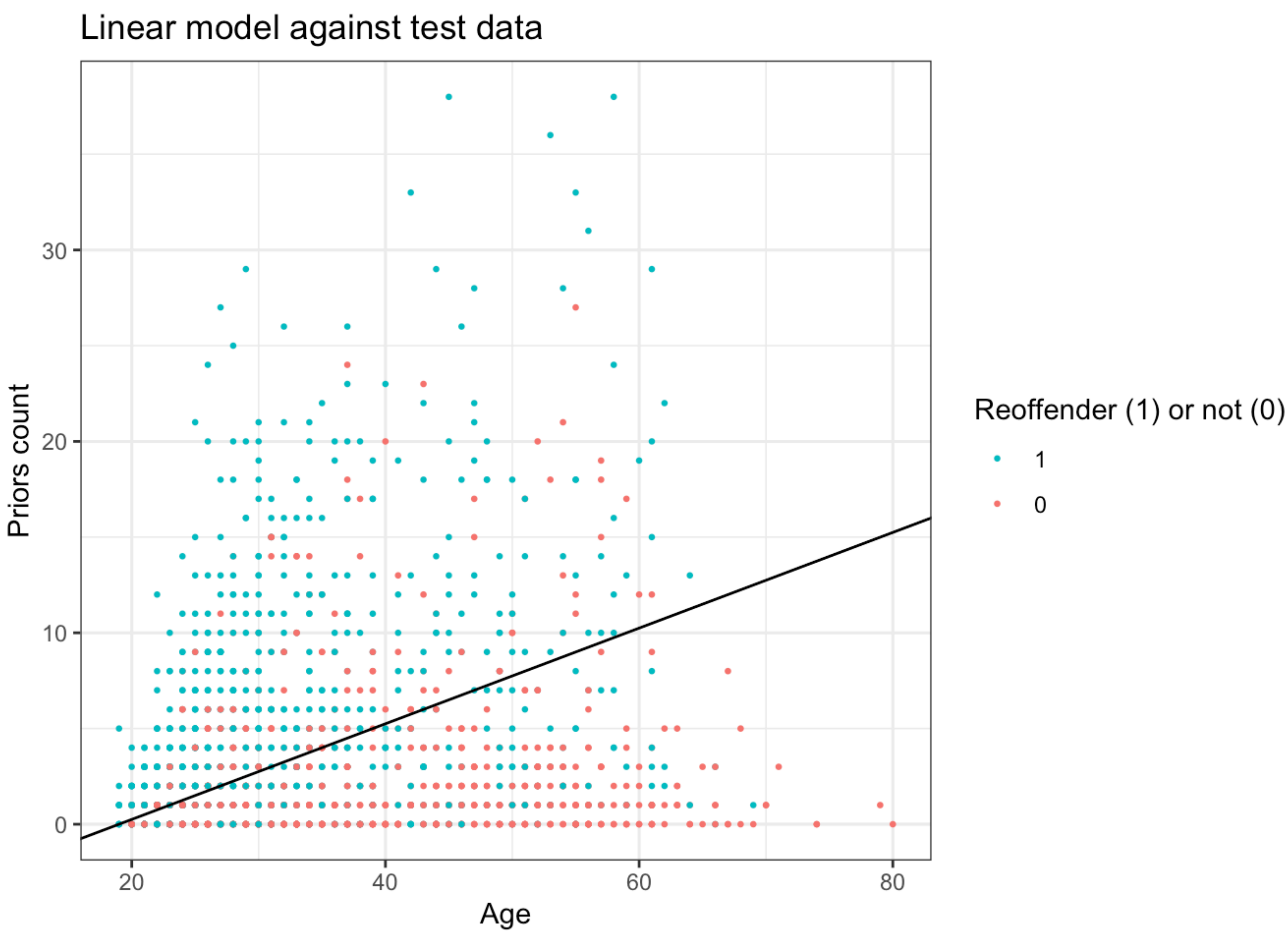
It is not at all clear where the line (predictive model) should be drawn to minimize errors

SVM Risk Model: Support Vectors and Line

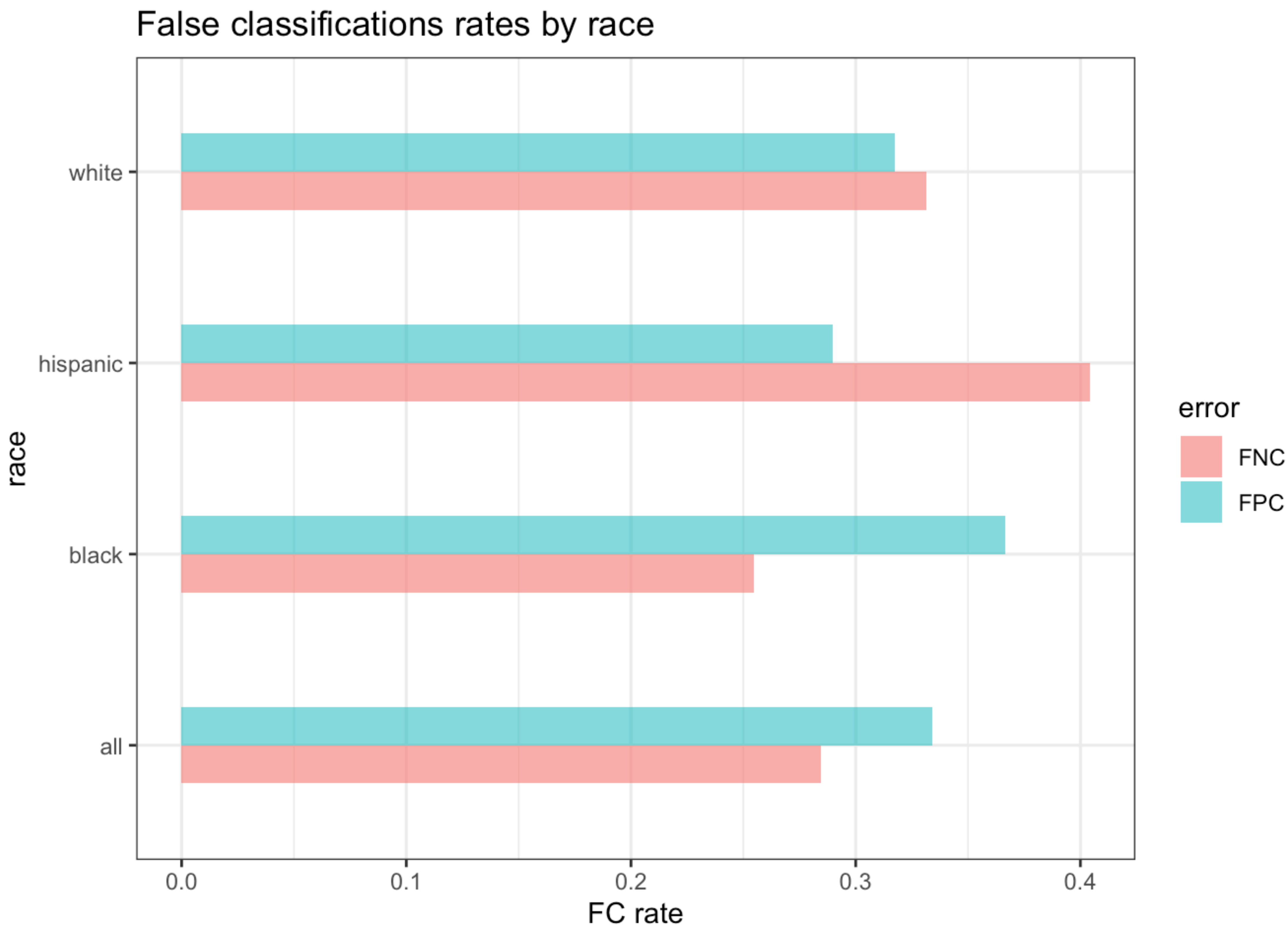


Training Data v. Test Data

Validating Model Against Test Data



Examples of Error Rates



PART II

Examples of Predictive Algorithms in Criminal Justice

Predictive Policing

Functions

In criminal investigations, algorithmic systems are reported to be used at least for the following purposes (*RAND report, 2013*)

- Predicting crimes
- Predicting offenders
- Identifying perpetrators
- Predicting victims

Two different models:

- Replicate conventional crime mapping and investigative methods
- Use predictive analytics methods to identify specific individuals (perpetrators or victims)

Model 1: Conventional approach

- Big data and machine learning are used to identify promising targets for police intervention

- **Place-based** predictive policing (Predpol, XLAW, KeyCrime…)

- **Individual-based** predictive policing (Chicago's Strategic Subject List, Beware, Gang Matrix, Radar-iTE…)

https://www.chicagotribune.com/news/criminal-justice/ct-chicago-police-strategic-subject-list-ended-20200125-spn4kjmrxrh4tmktdjckhtox4i-story.html

Predpol

"I'm not going to get more money. I'm not going to get more cops. I have to be better at using what I have, and that's what predictive policing is about"

Los Angeles Police Chief Charlie Beck, CBS Evening News

www.predpol.com

PREDPOL

Based on historical crime data (victims' information)

3 data points: time, place, type of offence

Model 1: Conventional approach

XLAW – Naples Police

- Risk map updated every 30 minutes
- Prediciton on place, time, type of offence and modus operandi
- Focused on robberies and thefts

https://www.xlaw.it/presentazione/index_eng.asp

Keycrime – Milan

- Focused on commercial robberies
- Predicts when, where, how the *same* robbers will strike (crime linking)

https://www.keycrime.com/

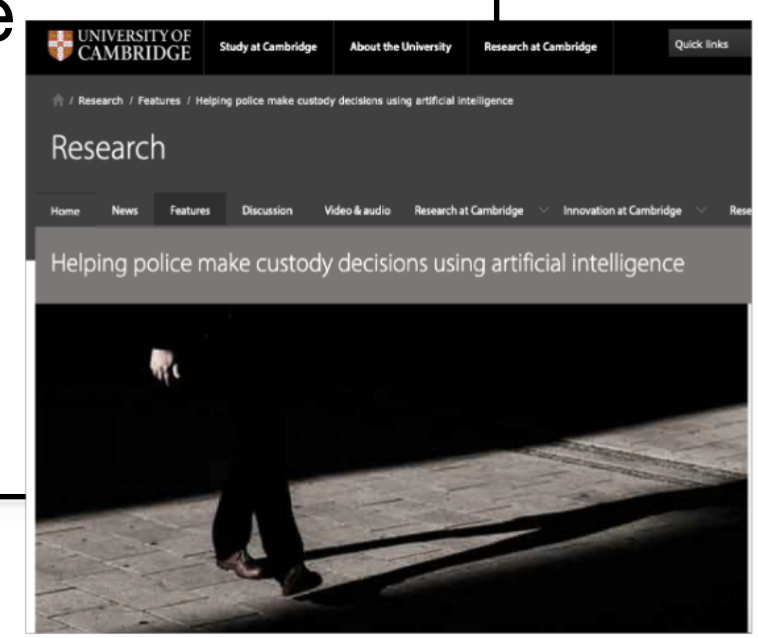
Model 2: Individual assessment

A second approach uses predictive analytics methods that, accessing huge amount of data (not necessarily already available to law-enforcement), automatically correlate risk factors with specific individuals

HARM ASSESSMENT RISK TOOL (HART)

UK Durham police and Cambridge University

- “It makes predictions based on 33 different metrics, including previous offence history, age and postcode of the offender”
- Metrics used are (reportedly) publicly available
- The model is trained to favor false positives over false negatives



Predictive Algorithms For Judges

Example 1:
COMPAS

COMPAS (Northpoint Inc./Equivant): “static information (criminal history), with limited use of some dynamic variables (i.e. criminal associates, substance abuse)” + 137 interview questions + …?

The next few statements are about what you are like as a person, what your thoughts are, and how other people see you. There are no ‘right or wrong’ answers. Just indicate how much you agree or disagree with each statement.

112. "I am seen by others as cold and unfeeling."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

113. "I always practice what I preach."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

114. "The trouble with getting dose to people is that they start making demands on you."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

115. "I have the ability to "sweet talk" people to get what I want."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

116. "I have played sick to get out of something."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

117. "I'm really good at talking my way out of problems."
☑ Strongly Disagree ☐ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

118. "I have gotten involved in things I later wished I could have gotten out of."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

119. "I feel bad if I break a promise I have made to someone."
☐ Strongly Disagree ☐ Disagree ☐ Not Sure ☑ Agree ☐ Strongly Agree

120. "To get ahead in life you must always put yourself first."
☐ Strongly Disagree ☑ Disagree ☐ Not Sure ☐ Agree ☐ Strongly Agree

WHICH APPLICATION?

- probation, alternative measures, etc.

- and what about sentencing?

The Loomis Case - State v. Loomis, 881 N.W.2d 749 (Wis. 2016)

Example 2:
Public Safety Assessment (PSA)
(Printout)

PART III

Controversies:
- (a) Mistakes
- (b) Bias and Fairness
- (c) Illusionary Objectivity
- (d) Individualized Predictions?

(a) Predictive Models Can Make Mistakes

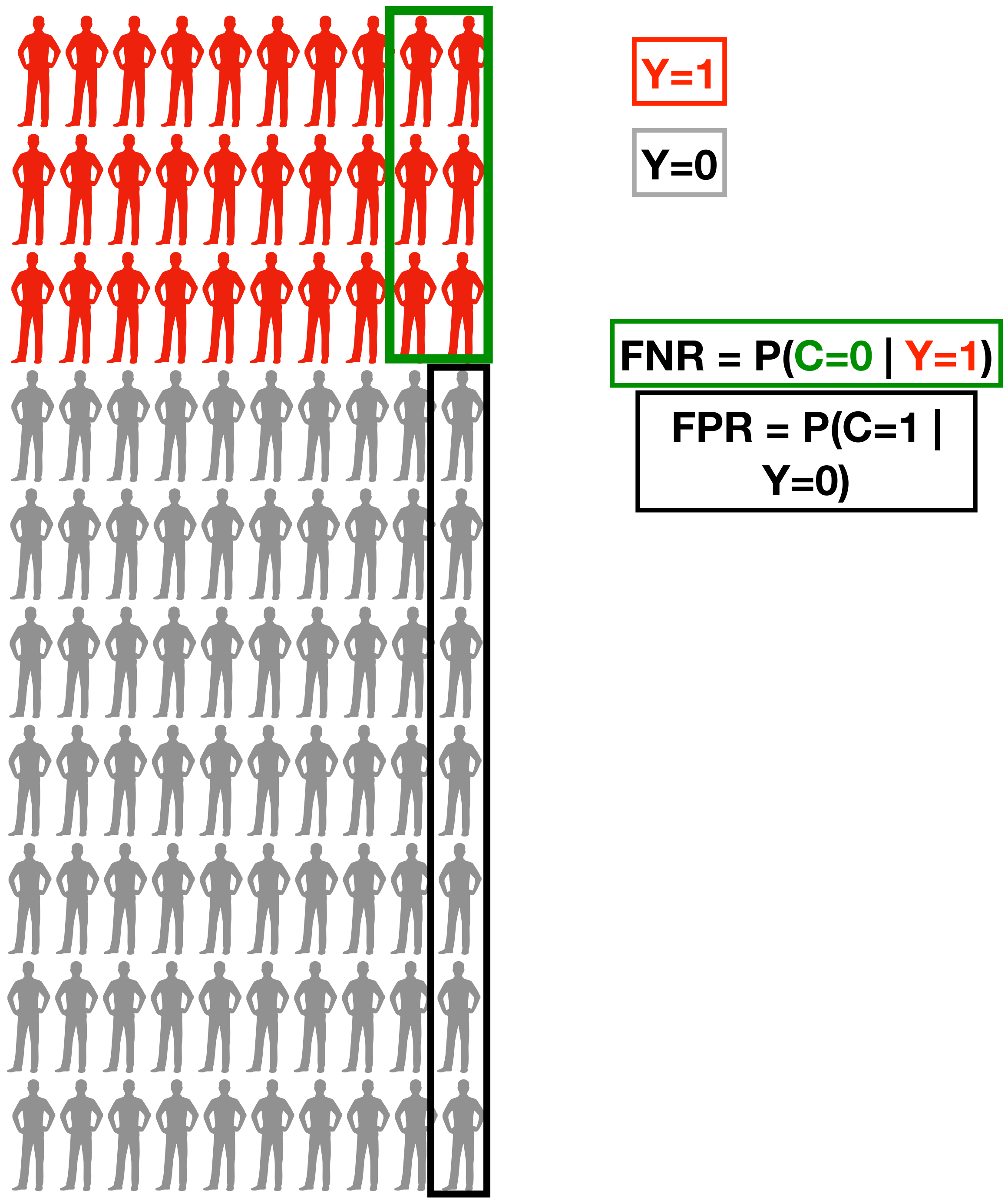
Dichotomous Accuracy/Error Metrics

Dichotomous Accuracy/Error Metrics

False negative rate (**FNR**)
P(C=0 | Y=1)

False positive rate (**FPR**)
P(C=1 | Y=0)

Y is the *actual* outcome
C is the *classified* outcome



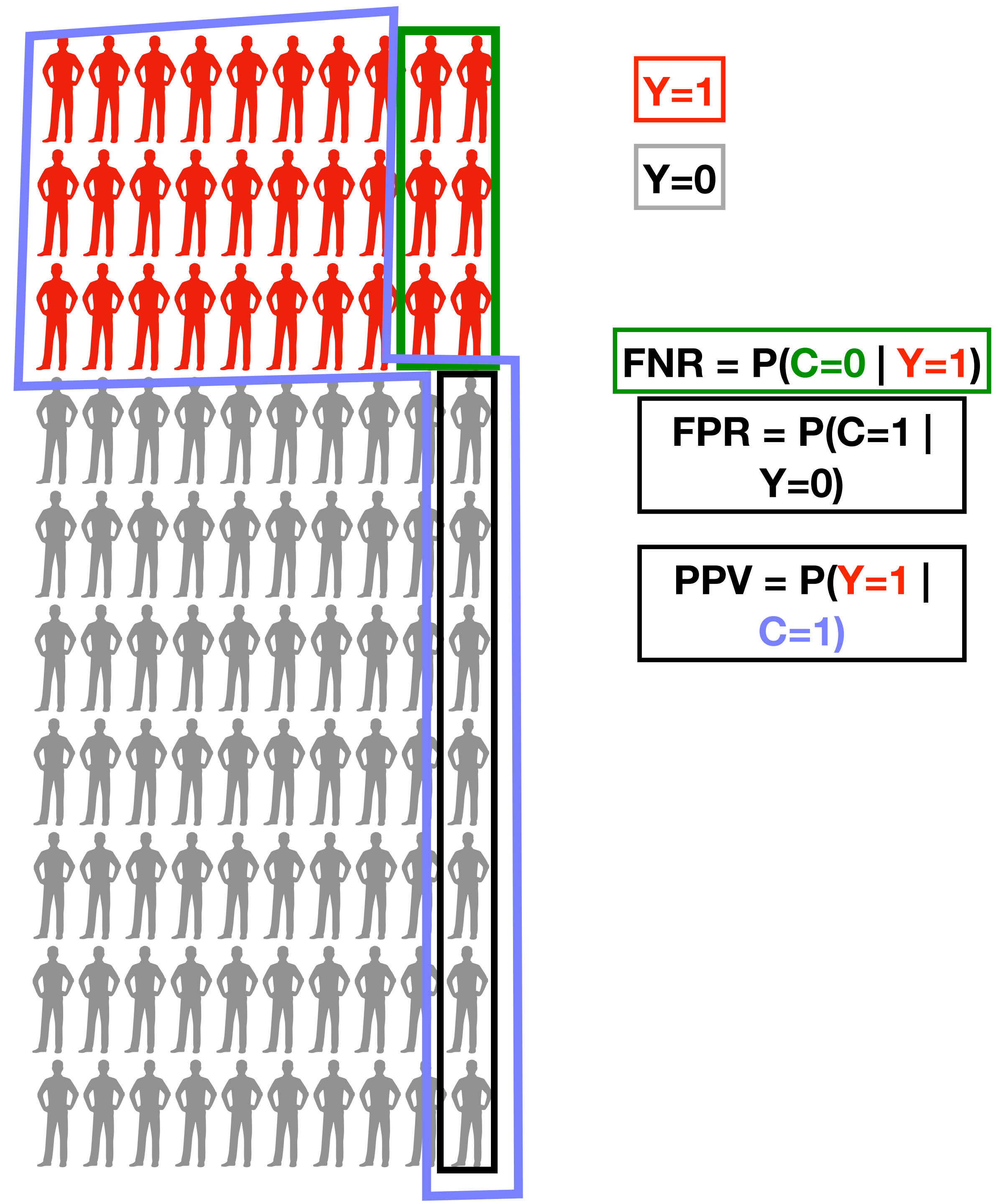
Dichotomous Accuracy/Error Metrics

False negative rate (**FNR**)
P(C=0 | Y=1)

False positive rate (**FPR**)
P(C=1 | Y=0)

Positive predictive value (**PPV**) P(Y=1 | C=1)

Y is the *actual* outcome
C is the *classified* outcome

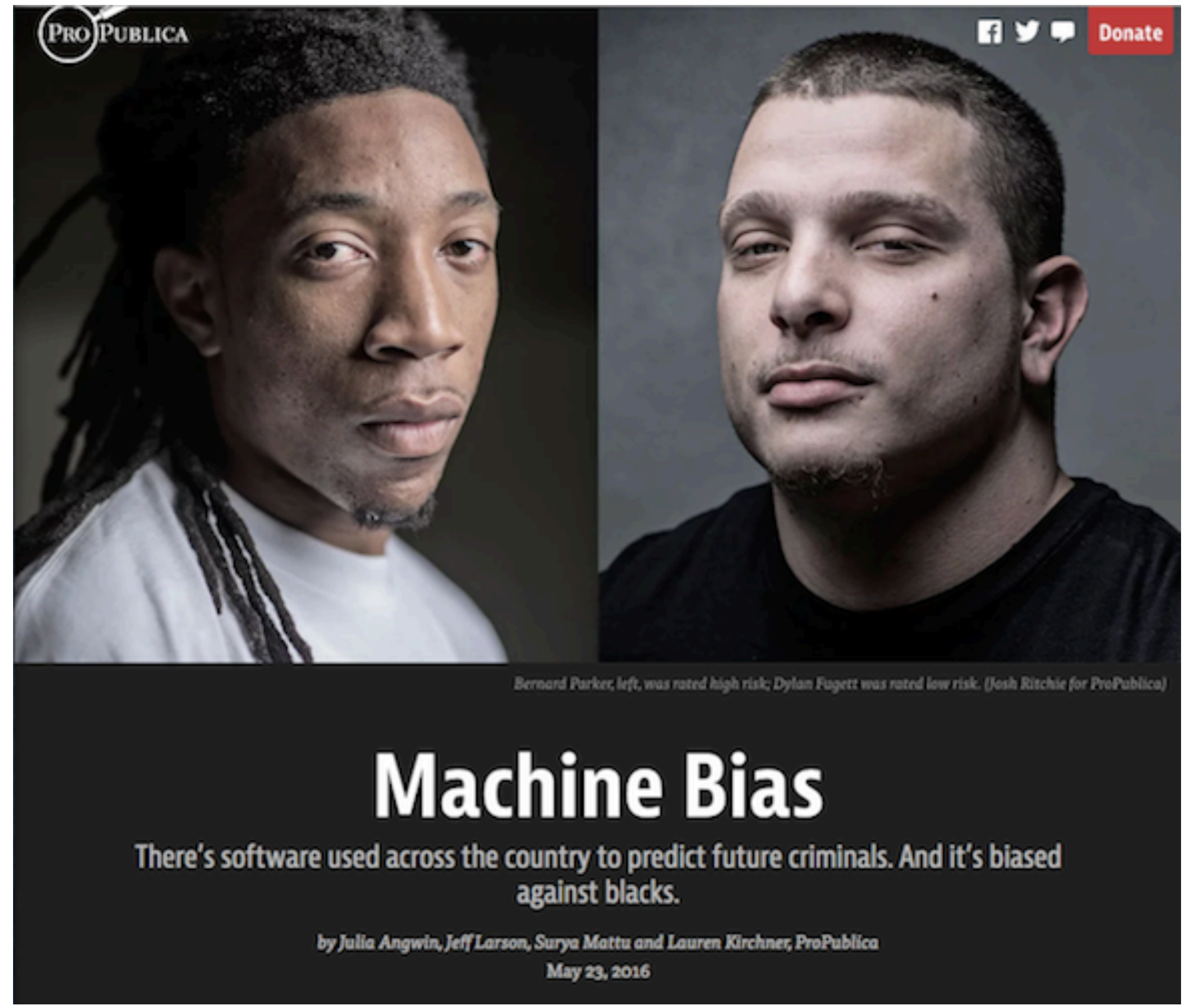


(b) Predictive Models Can Be Biased

(b) Predictive Models Can Be Biased

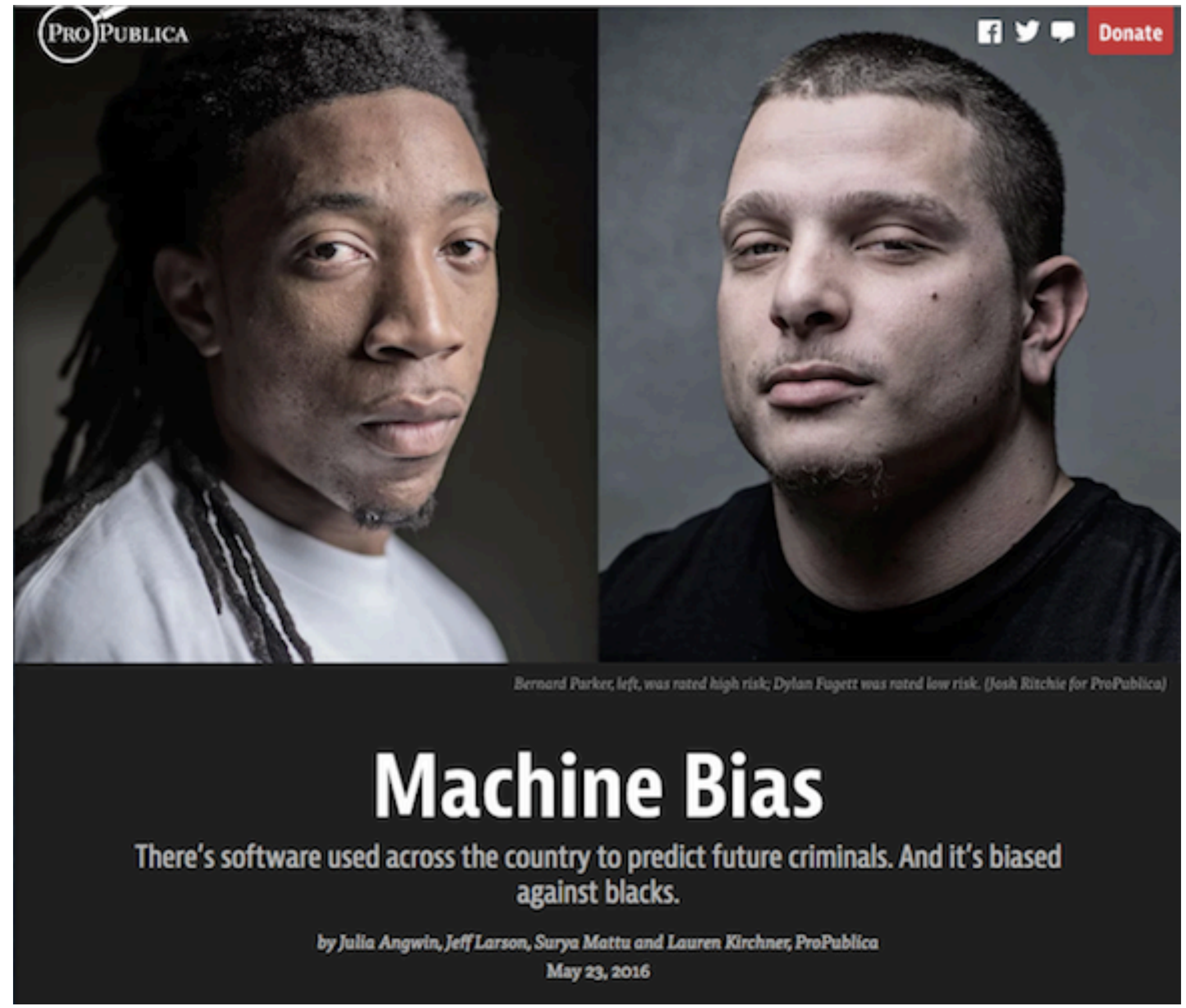
Bias is a deviation from impartiality. People who should be treated the same are treated differently.

COMPAS Algorithm



COMPAS Algorithm

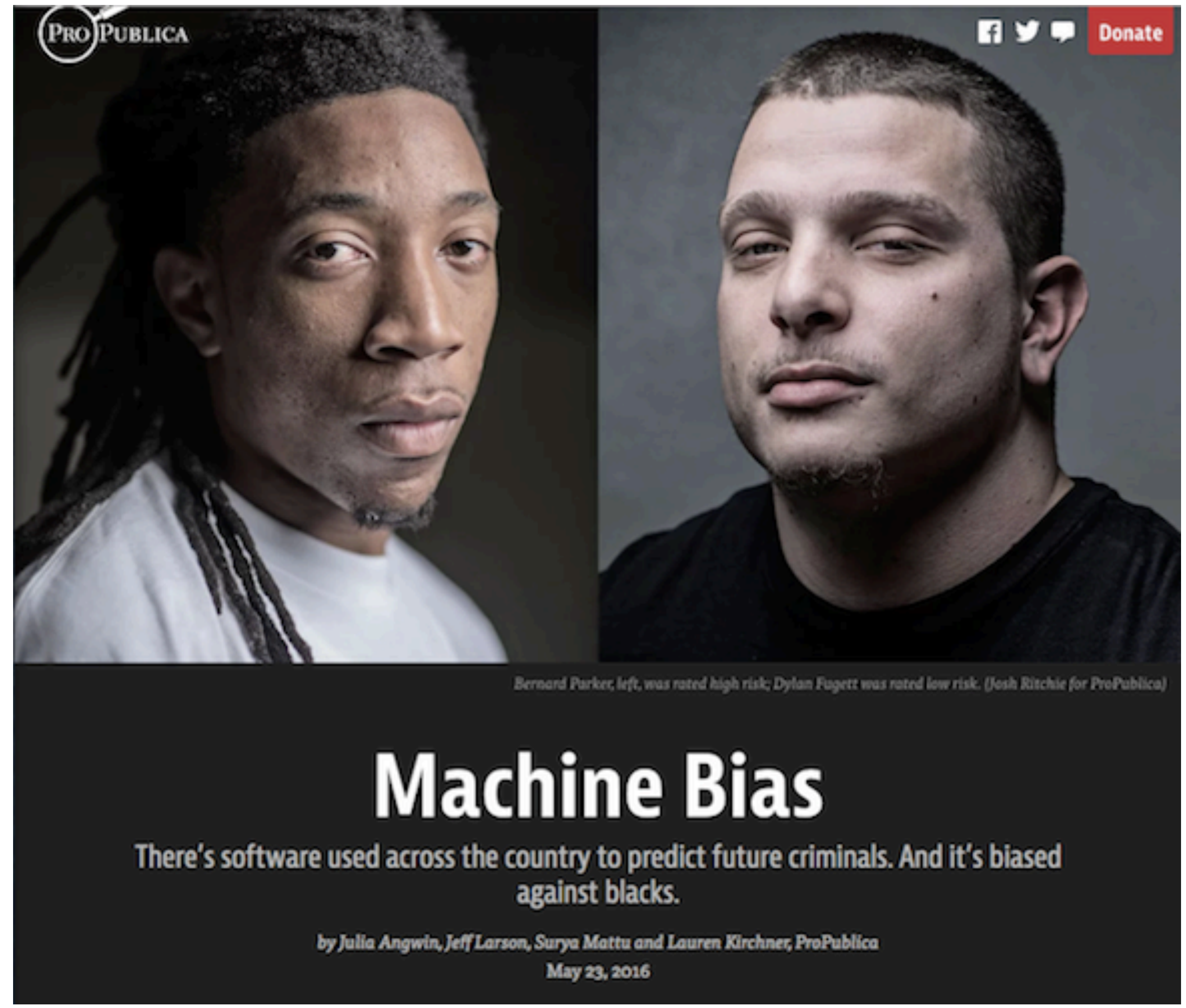
Even if ‘race’ is not among the predictive features used:



COMPAS Algorithm

Even if 'race' is not among the predictive features used:

- False positive rate (**FPR**) was **higher for blacks** than whites

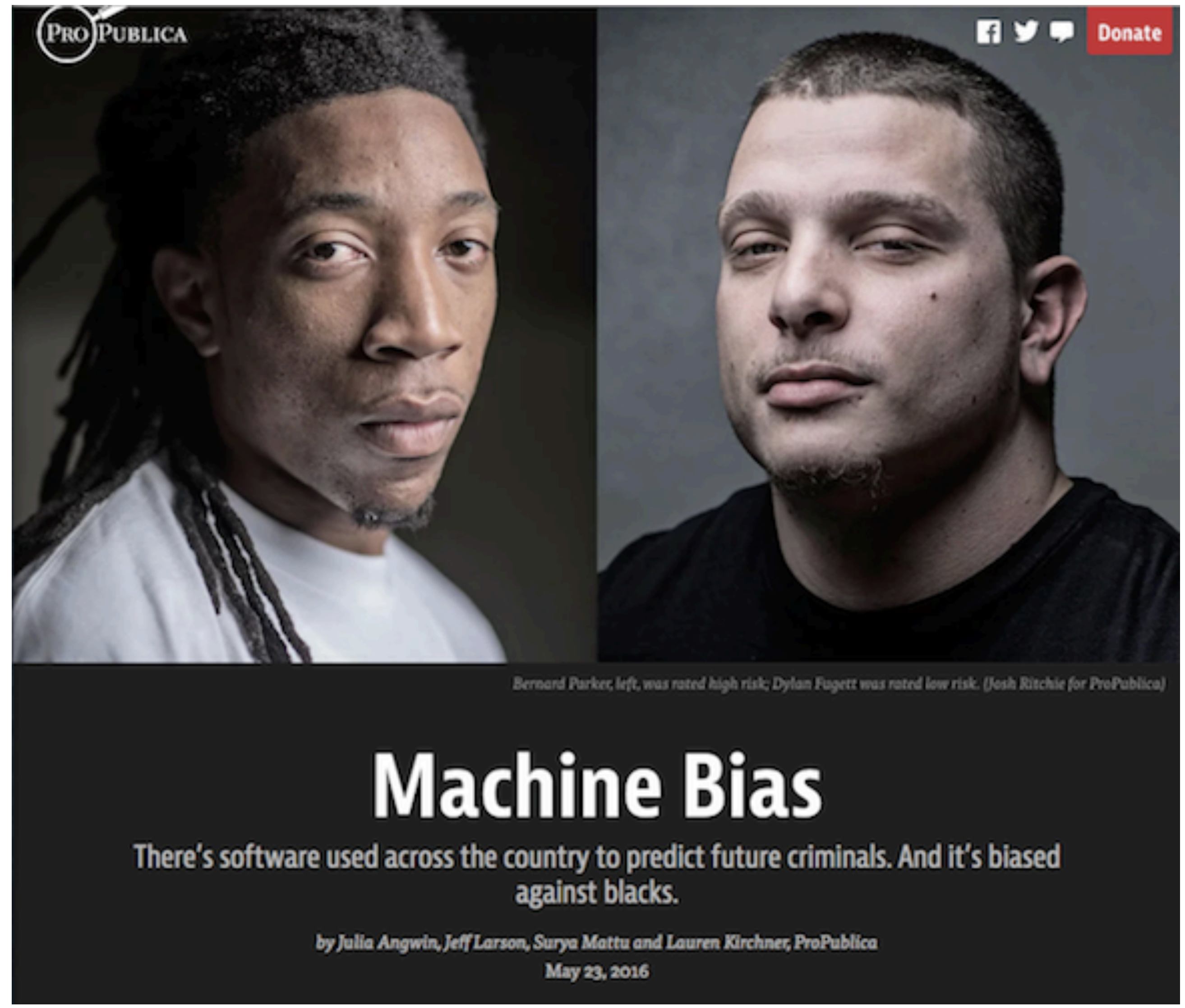


COMPAS Algorithm

Even if ‘race’ is not among the predictive features used:

- False positive rate (**FPR**) was **higher for blacks** than whites

- False negative rate (**FNR**) was **higher for whites** than blacks



COMPAS Algorithm

Even if ‘race’ is not among the predictive features used:

- False positive rate (**FPR**) was **higher for blacks** than whites

- False negative rate (**FNR**) was **higher for whites** than blacks

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

Machine Bias

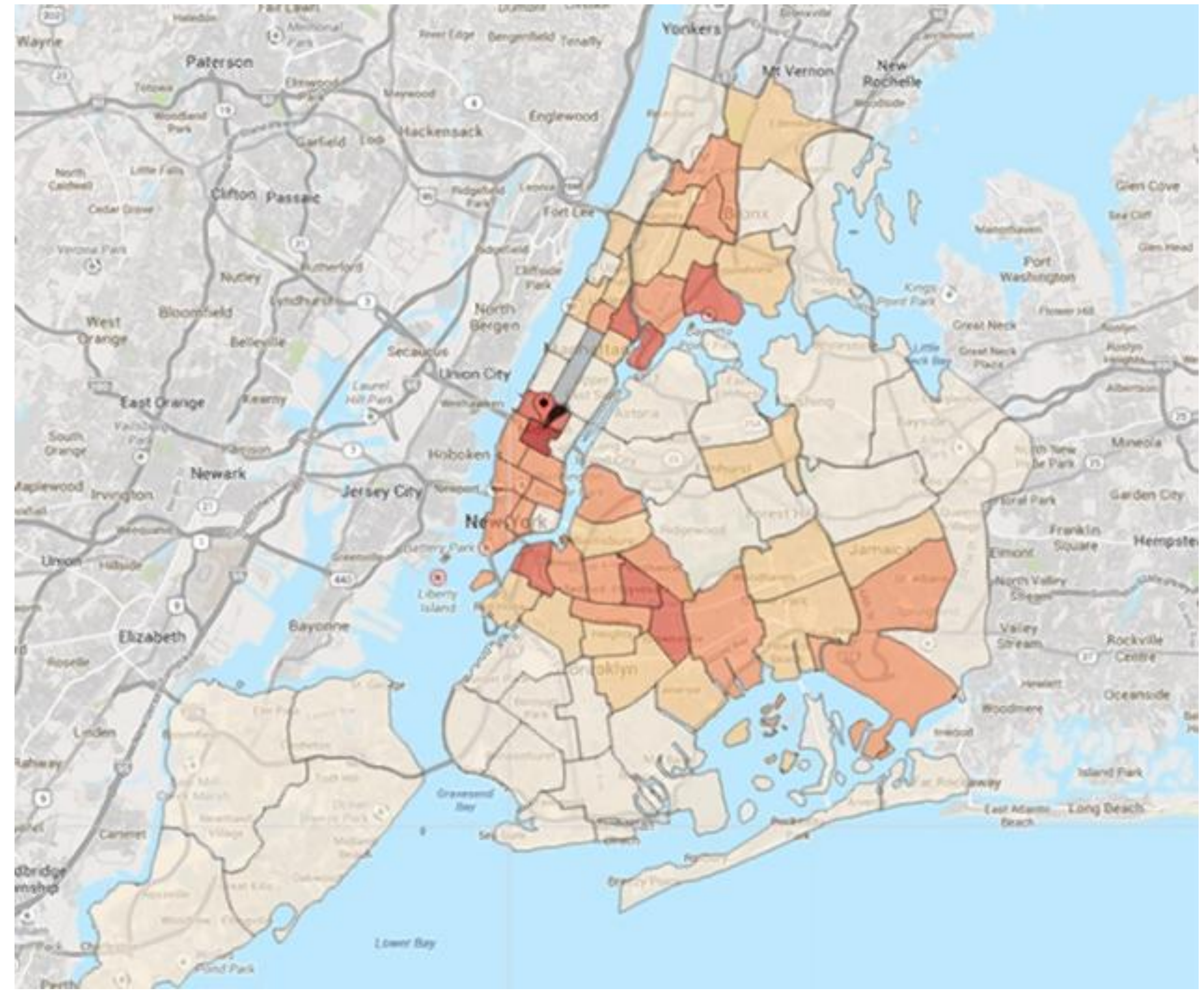
Prediction Fails Differently for Black Defendants

	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%

Sources of Algorithmic Bias

Sources of Algorithmic Bias

1. **Biased data:** one group is oversampled, data about one group contain more noise, etc.

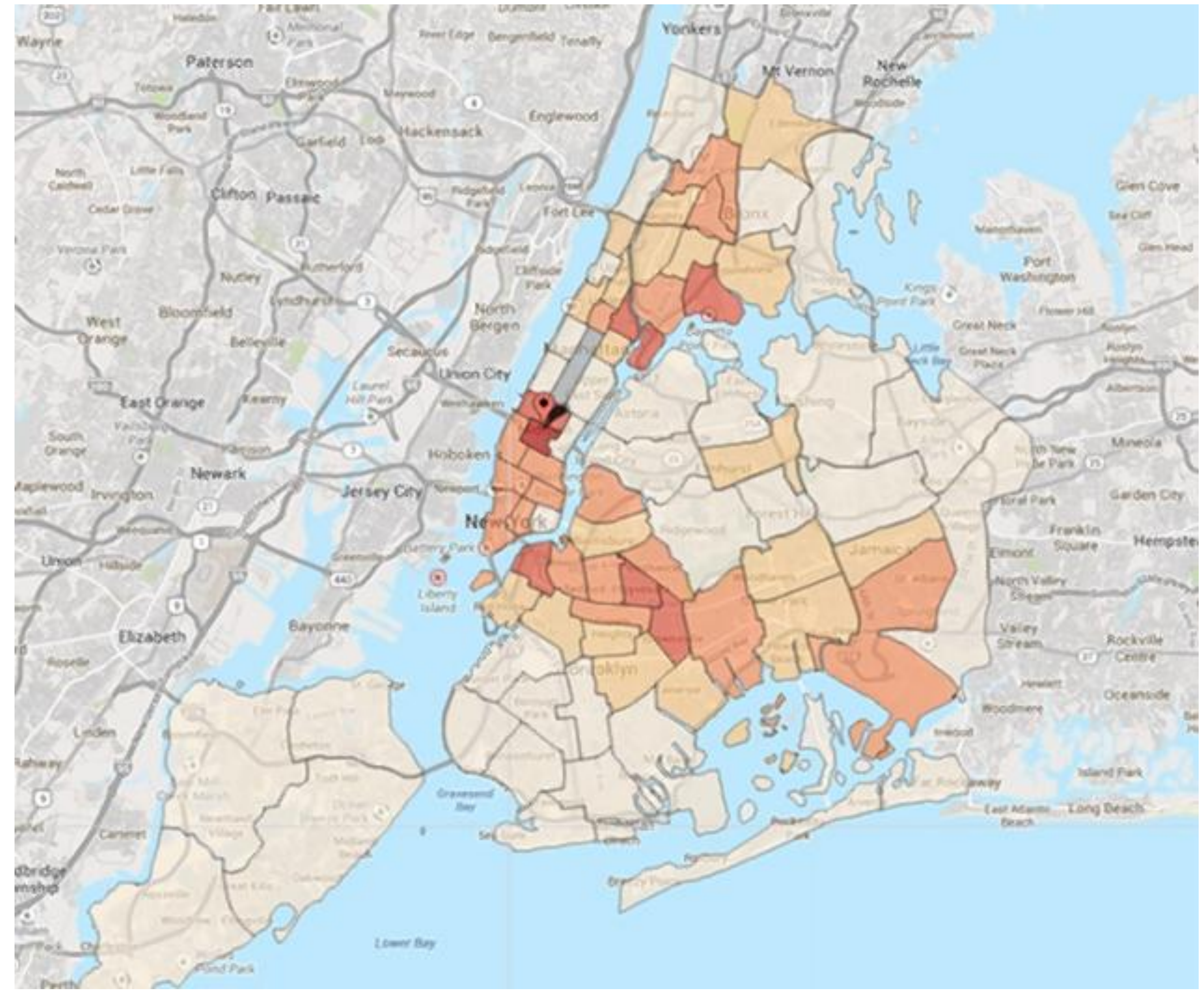


Call this the **biased data argument** about algorithmic bias

Sources of Algorithmic Bias

1. **Biased data:** one group is oversampled, data about one group contain more noise, etc.

   - Use of proxies variables can be pernicious, e.g., when ‘arrest’ is used as a proxy for recidivism or ‘healthcare cost’ as a proxy for ‘care need’

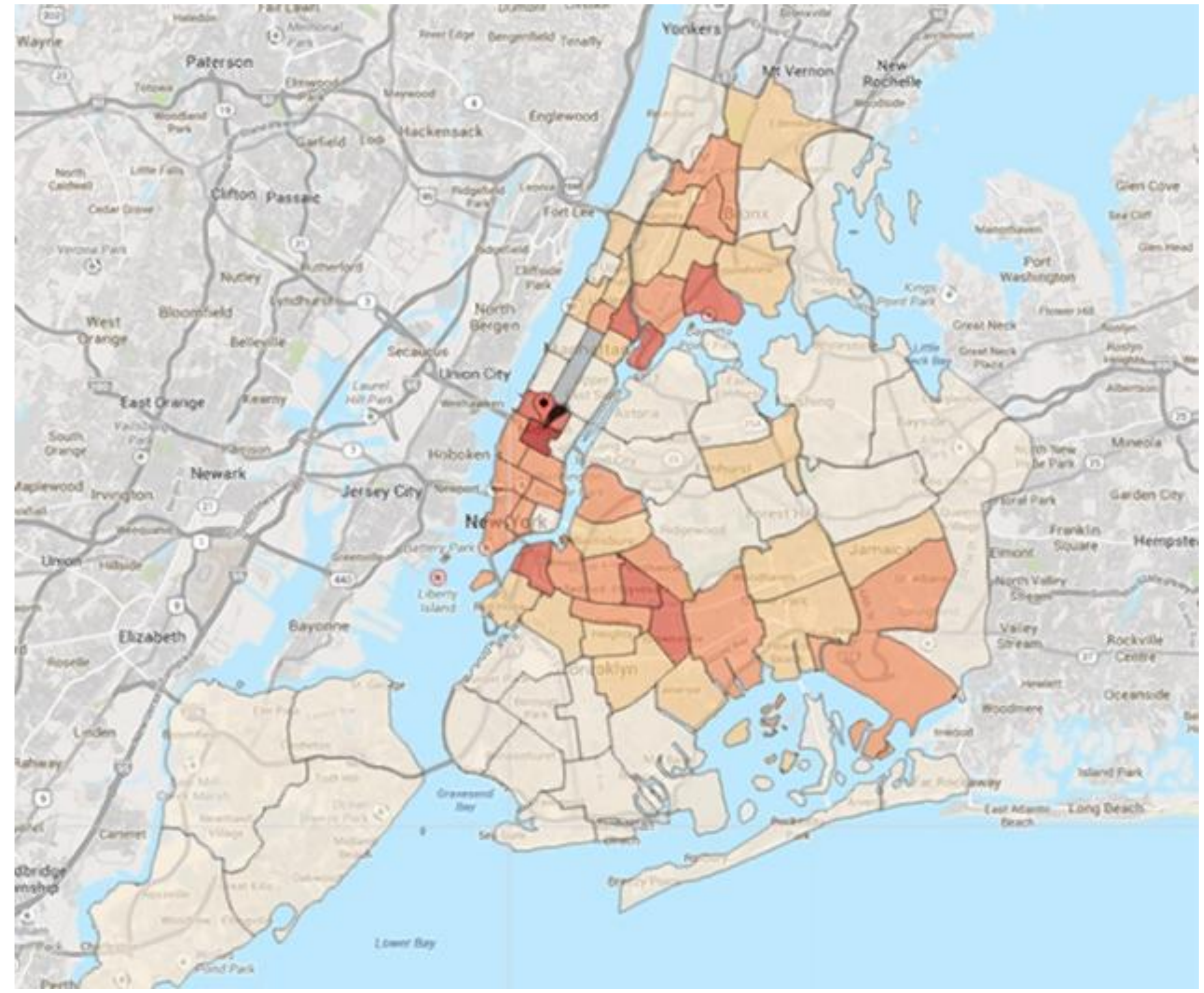


- Call this the **biased data argument** about algorithmic bias

Sources of Algorithmic Bias

1. **Biased data:** one group is oversampled, data about one group contain more noise, etc.

    - Use of proxies variables can be pernicious, e.g., when 'arrest' is used as a proxy for recidivism or 'healthcare cost' as a proxy for 'care need'
    - Feedback loops, e.g., more black people are arrested since data show that they commit more crime but the data use 'arrest' as a proxy for crime



- Call this the **biased data argument** about algorithmic bias

Sources of Algorithmic Bias

Sources of Algorithmic Bias

2. Data may portray an accurate picture of reality, but **society itself may contain biases**, so the data reflect these societal biases

- Call this the **structural injustice argument** about algorithmic bias

Sources of Algorithmic Bias

2. Data may portray an accurate picture of reality, but **society itself may contain biases**, so the data reflect these societal biases

    - It may well be true that certain groups commit crimes or default on loans at higher rates, but these disparities speak more about inequalities and injustices in society rather than about inherent features of these groups

- Call this the **structural injustice argument** about algorithmic bias

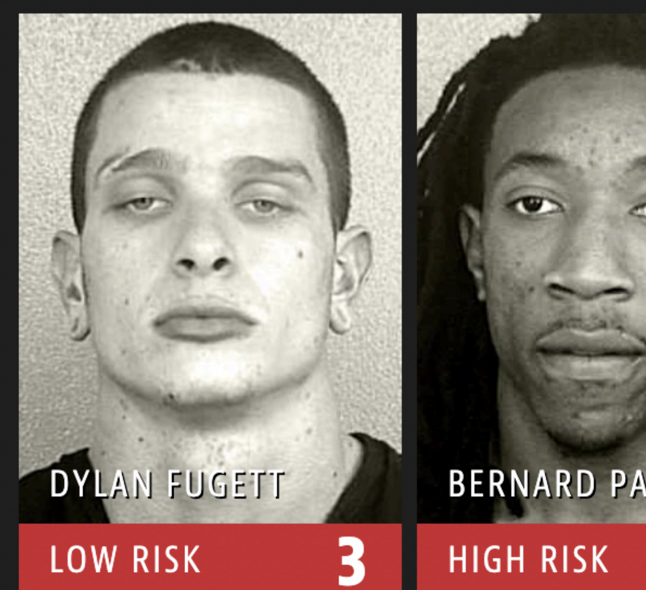
Prediction Fails Differently for Black Defendants

	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%

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%

Two Drug Possession Arrests



Fugett was rated low risk after being arrested with coca
marijuana. He was arrested three times on drug charge:

Two DUI Arrests

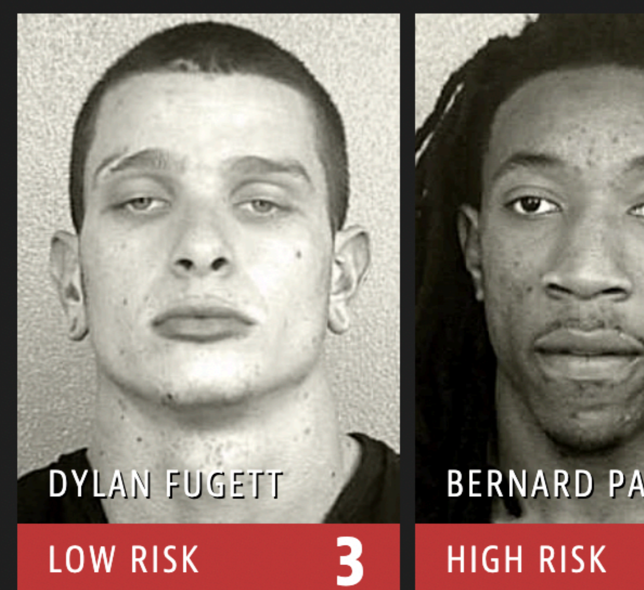
Lugo crashed his Lincoln Navigator into a Toyota Camry while drunk. He was rated as a low risk of reoffending despite the fact that it was at least his fourth DUI.

Absent biased data and a biased society, would disparities such as the ones in the COMPAS algorithm disappear?

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%

Two Drug Possession Arrests



Fugett was rated low risk after being arrested with coca
marijuana. He was arrested three times on drug charge

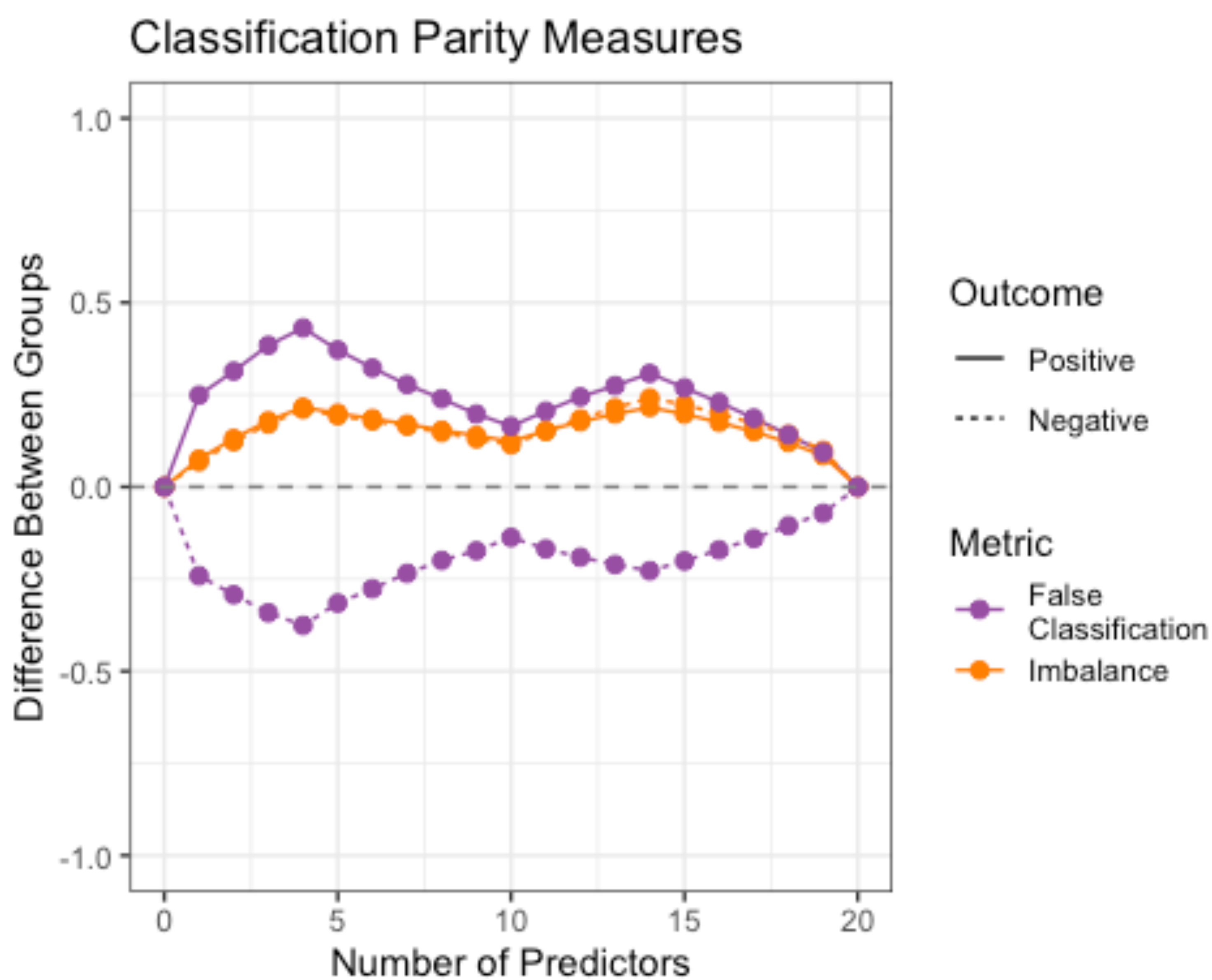
Two DUI Arrests

Lugo crashed his Lincoln Navigator into a Toyota Camry while drunk. He was rated as a low risk of reoffending despite the fact that it was at least his fourth DUI.

Absent biased data and a biased society, would disparities such as the ones in the COMPAS algorithm disappear?

Absent biased data and a biased society, would disparities such as the ones in the COMPAS algorithm disappear?

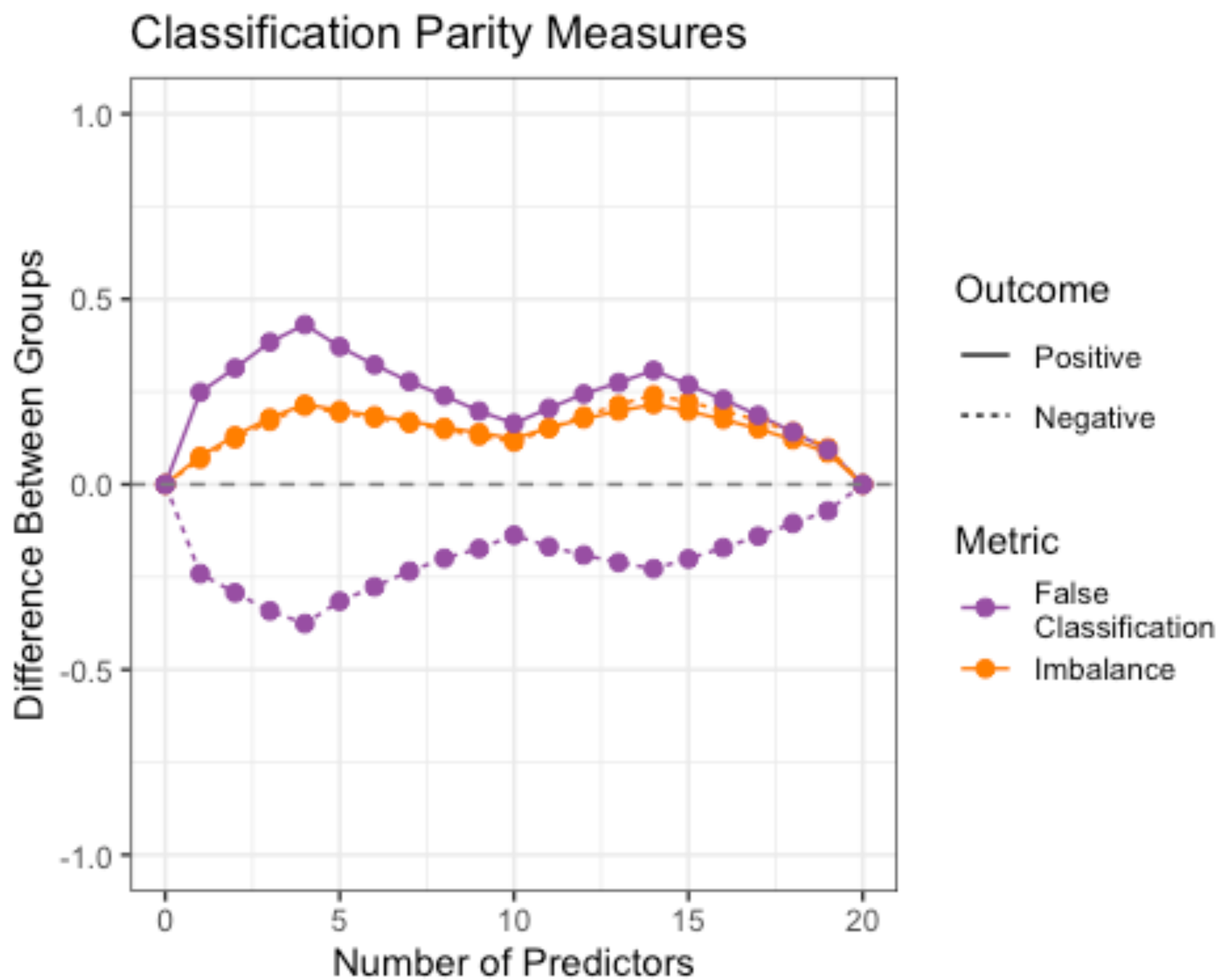
Simulating an Unbiased Dataset in an Unbiased Society



Joint work with Ruobin Gong of the Department of Statistics at Rutgers University

Absent biased data and a biased society, would disparities such as the ones in the COMPAS algorithm disappear? **NO**

Simulating an Unbiased Dataset in an Unbiased Society



Joint work with Ruobin Gong of the Department of Statistics at Rutgers University

(c) Illusionary Objectivity

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

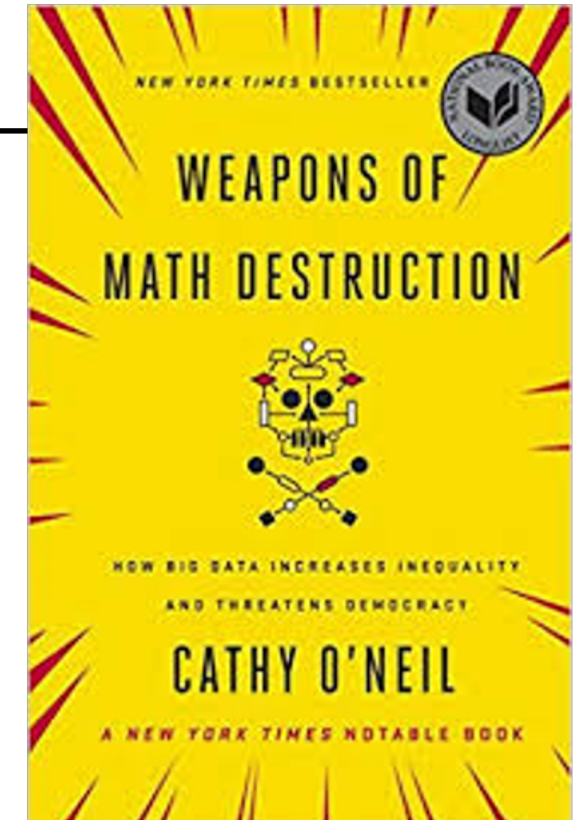
"…decisions made by computers **may enjoy an undeserved assumption of fairness or objectivity.** However, the design and implementation of automated decision systems can be vulnerable to a variety of problems that can result in **systematically faulty and biased determinations**."

•J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2016). Accountable algorithms. University of Pennsylvania Law Review, 165.

CAUSES 1)
-training set that reflect past prejudice or implicit bias, or that offer a statistically distorted picture of groups comprising the overall population
-Even dataset without initial bias ma result in biased systems (self-reinforcement, no distinction between correlation and causes). Example: correlation between speeding and drug trafficking
-Extraction of sensitive /special categories of personal data from non-sensitive data

However
Algorithms may also correct human cognitive biases (Sunstein 2018)

(c) Illusionary Objectivity (cont'ed)



CAUSES 2)
Legal Value Attached to the "predictions"

…Do you see the paradox?
An algorithm processes a slew of statistics
and comes up with a probability that a certain person
might be a bad hire, a risky borrower, a terrorist,
or a miserable teacher.
That probability is distilled into a score, which can turn someone's life upside down.
And yet when the person fights back, "suggestive" countervailing evidence simply won't cut it.
The case must be ironclad. The human victims of WMDs, we'll see time and again, are held to a far higher standard of evidence than the algorithms themselves…

(O'Neill, Weapons of Math Destruction)

- ***is it correct to really talk about "predictions"?***
- ***what about the right to an individual assessment?***

(d) Individualized judgment?

Algorithmic predictions are based on group correlations — anyone who possess the same set of characteristic (say, high number of prior arrests and young age) will be classified the same way.

But every individual is different and algorithms may fail to take into account individual-specific characteristic that are nevertheless relevant.

- ***Is it correct to really talk about "predictions"?***
- ***What about the right to an individual assessment?***

PART IV

Possible Remedies

WHICH REMEDY?

Art. 11 LED – Automated individual decision making

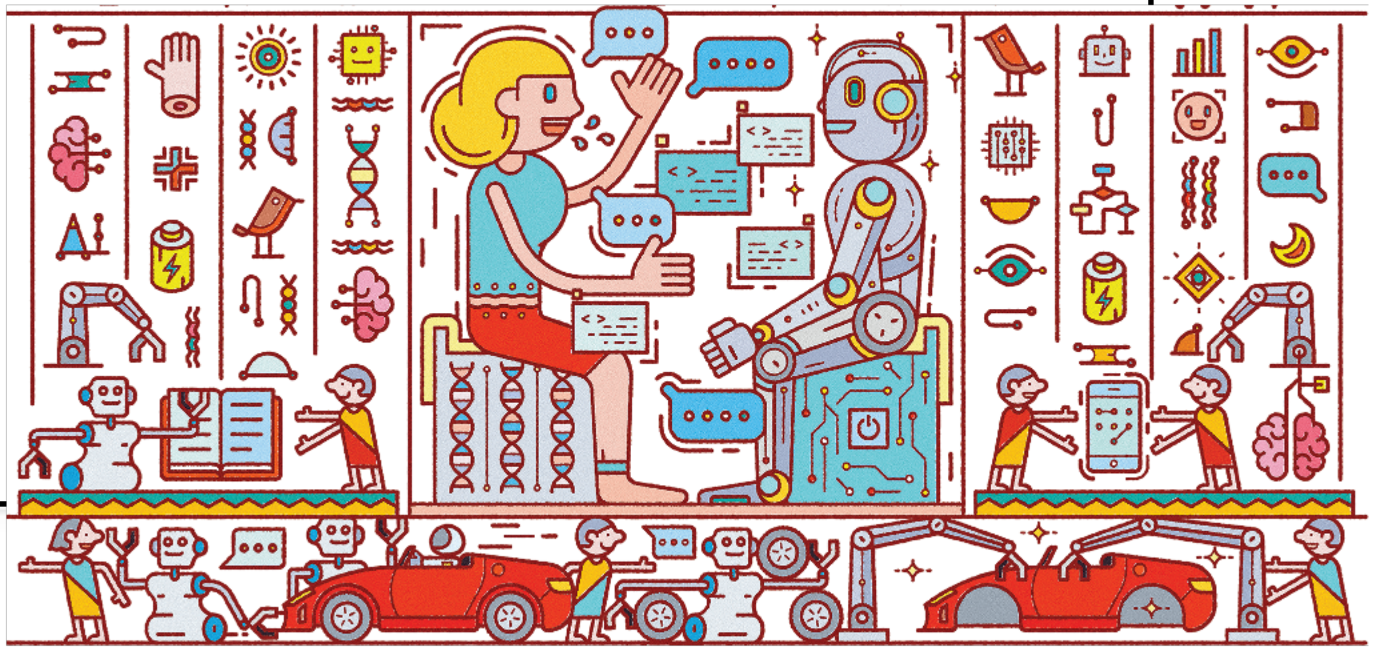
Decision based <u>solely</u> on automated processing
(including profiling)

which produces an adverse legal effect concerning the data subject or significantly affects him or her,
<u>shall be prohibited unless</u>

- authorised by Union or Member State law
- appropriate safeguards are provided, at least **<u>the right to obtain human intervention on the part of the controller</u>**

WHICH REMEDY?/2

"Owing to the evidence in their favor (stipulated by definition), it is more appropriate to think of **expert robots as above average in their ability to make decisions that will produce desirable outcomes.**
This fact suggests that **granting a general decision-making authority to human experts will be problematic once expert robots are properly on the scene**.
It might seem justifiable to grant "override" authority to human experts in situations where there appears to be "clear" evidence contradicting the expert robot's judgment, but even this would be contra-evidence-based"
(Millar, Kerr 2018)

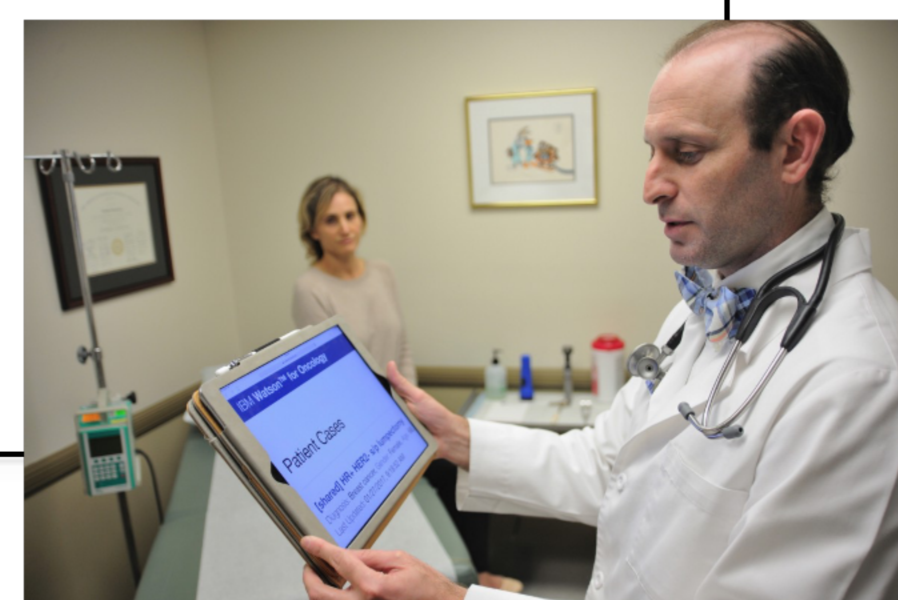


WHICH REMEDY?/3

Is human control really an effective remedy?

Machine intelligence is fundamentally alien, and often, the entire purpose of an AI system is to learn to do or see things in ways humans cannot. [..]

Ultimately, the **lack of a principled basis to contradict AI predictions implies that the reasonableness of an action in individual cases must be tied to the decision to use AI as a general matter.** *If a doctor receives a readout that suggests a patient has a certain rare diagnosis that she missed, how can the doctor determine whether or not to believe the AI and treat the patient accordingly?* (Selbst 2019)



Example

FRONTEX

- *European Travel Information Authorisation System* (ETIAS), fully operational by the end of 2022: automated assessment of third country citizens on the threat posed to national security or public health
- if positive assessment: need to have a second assessment by a human being

Do Human Overrides Improve Accuracy?

- “This study examines … the impact of overrides on the PCRA’s risk prediction effectiveness. Findings show that nearly all … tend to place **substantial numbers of persons under federal supervision** (especially those convicted of sex offenses) into the highest supervision categories, and that overrides result in a **deterioration of the PCRA’s risk prediction capacities.**"

RISK ASSESSMENT OVERRIDES

Shuffling the Risk Deck Without Any Improvements in Prediction

THOMAS H. COHEN
CHRISTOPHER T. LOWENKAMP
Administrative Office of the U.S. Courts
KRISTIN BECHTEL
Arnold Ventures
ANTHONY W. FLORES
California State University, Bakersfield

In the federal supervision system, officers have discretion to depart from the risk designations provided by the Post Conviction Risk Assessment (PCRA) instrument. This component of the risk classification process is referred to as the supervision override. While the rationale for allowing overrides is that actuarial scores cannot always capture an individual’s unique characteristics, there is relatively limited literature on the actual effects of overrides on an actuarial tool’s predictive efficacies. This study examines overrides in the federal system by assessing the extent to which risk levels are adjusted through overrides as well as the impact of overrides on the PCRA’s risk prediction effectiveness. Findings show that nearly all overrides lead to an upward risk reclassification, that overrides tend to place substantial numbers of persons under federal supervision (especially those convicted of sex offenses) into the highest supervision categories, and that overrides result in a deterioration of the PCRA’s risk prediction capacities.

Keywords: supervision overrides; risk prediction; risk assessment tools; professional discretion

Comparing Human and Machine Predictions

Human Decisions and Machine Predictions

Get access >

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan

The Quarterly Journal of Economics, Volume 133, Issue 1, February 2018, Pages 237–293,
https://doi.org/10.1093/qje/qjx032
Published: 26 August 2017

Cite Permissions Share ▾

Abstract

Can machine learning improve human decision making? Bail decisions provide a good test case. Millions of times each year, judges make jail-or-release decisions that hinge on a prediction of what a defendant would do if released.

Even accounting for these concerns, our results suggest potentially large welfare gains: one policy simulation shows crime reductions up to 24.7% with no change in jailing rates, or jailing rate reductions up to 41.9% with no increase in crime rates. Moreover, all categories of crime, including violent crimes, show reductions; these gains can be achieved while simultaneously reducing racial disparities. These results suggest that while machine learning

no change in jailing rates, or jailing rate reductions up to 41.9% with no increase in crime rates. Moreover, all categories of crime, including violent crimes, show reductions; these gains can be achieved while simultaneously reducing racial disparities. These results suggest that while machine learning
can be valuable, realizing this value requires integrating these tools into an economic framework: being clear about the link between predictions and decisions; specifying the scope of payoff functions; and constructing unbiased decision counterfactuals.

PART V

Impossibility Theorems

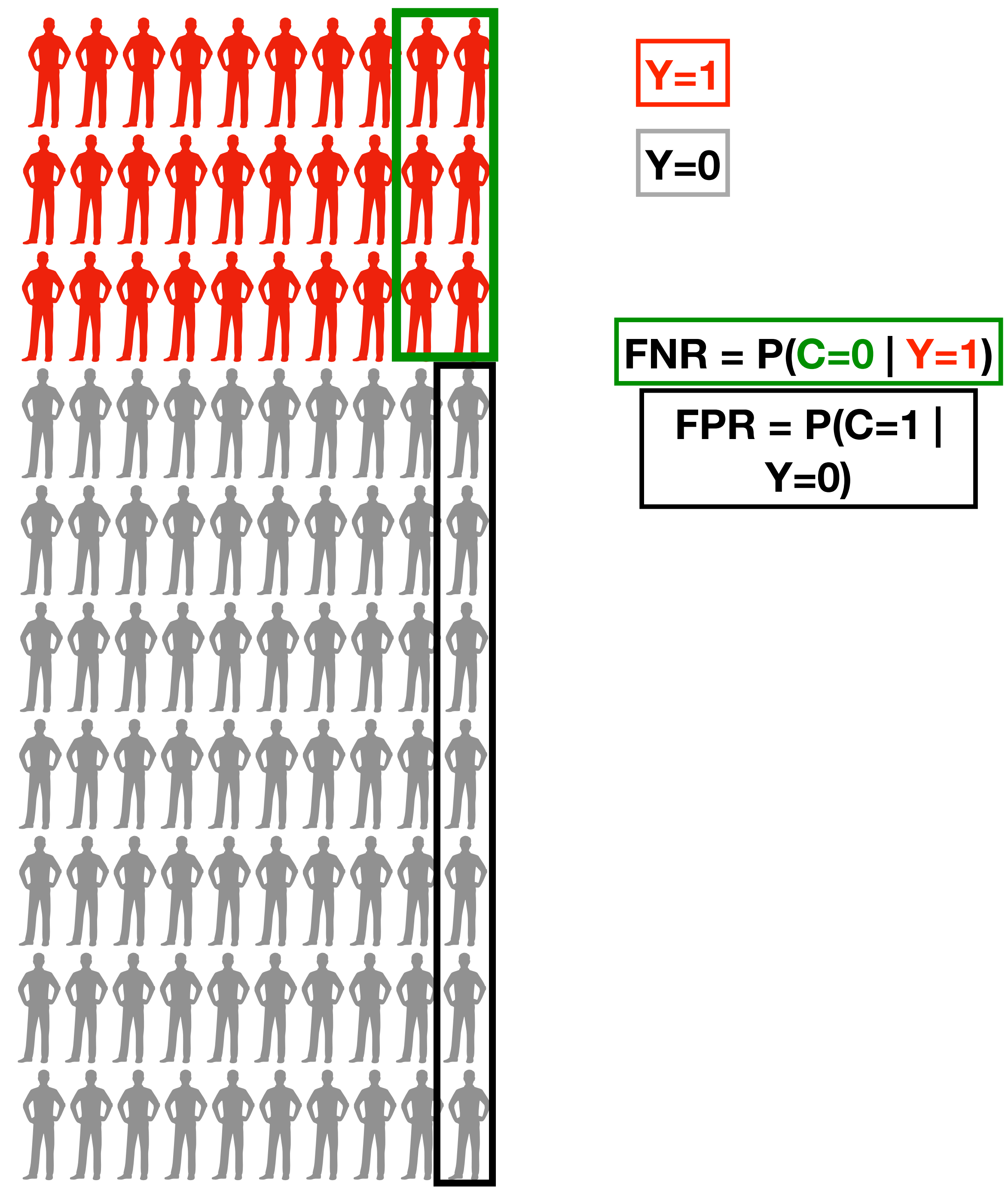
Dichotomous Accuracy/Error Metrics

Dichotomous Accuracy/Error Metrics

False negative rate (**FNR**)
P(C=0 | Y=1)

False positive rate (**FPR**)
P(C=1 | Y=0)

Y is the *actual* outcome
C is the *classified* outcome



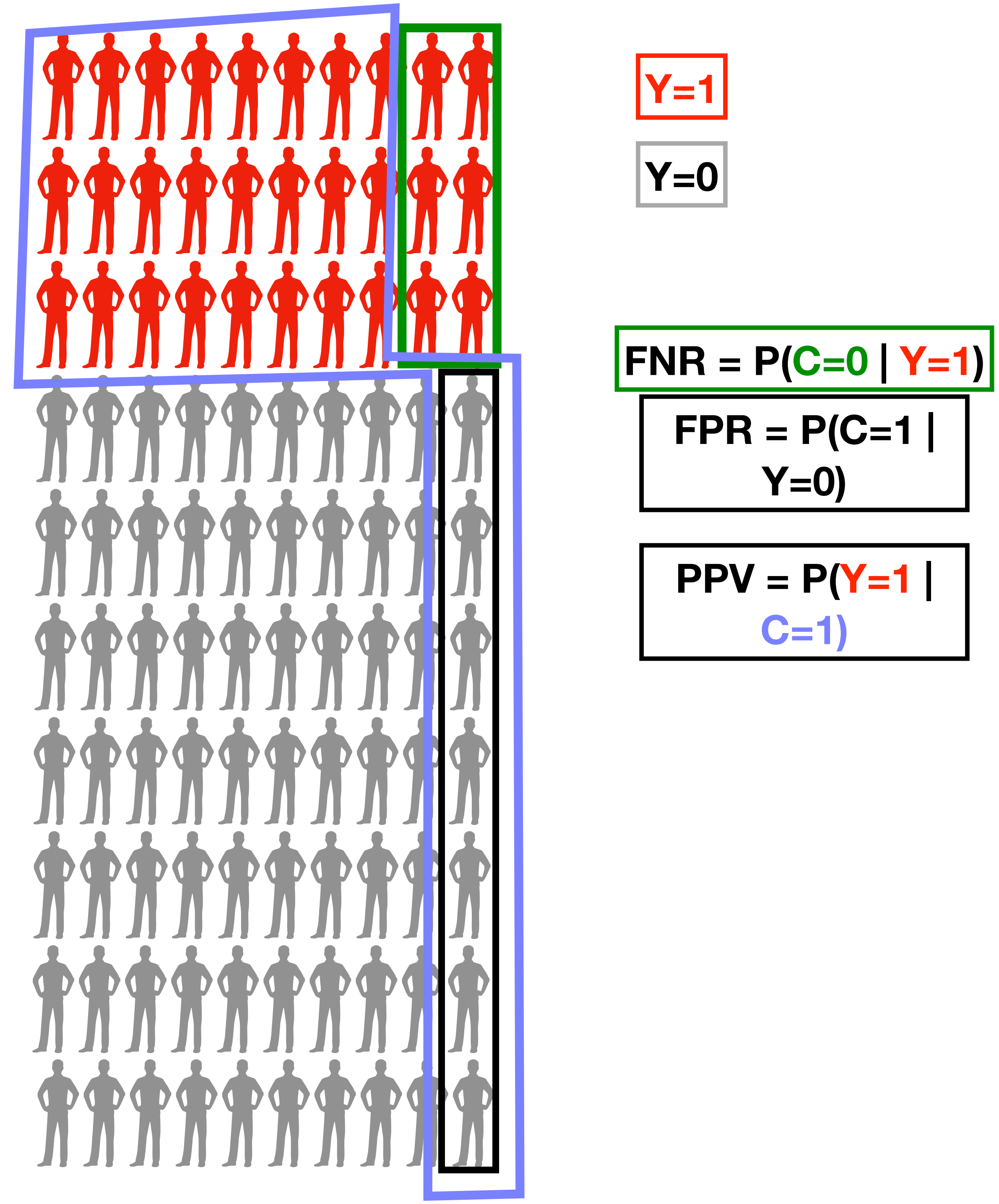
Dichotomous Accuracy/Error Metrics

False negative rate (**FNR**)
P(C=0 | Y=1)

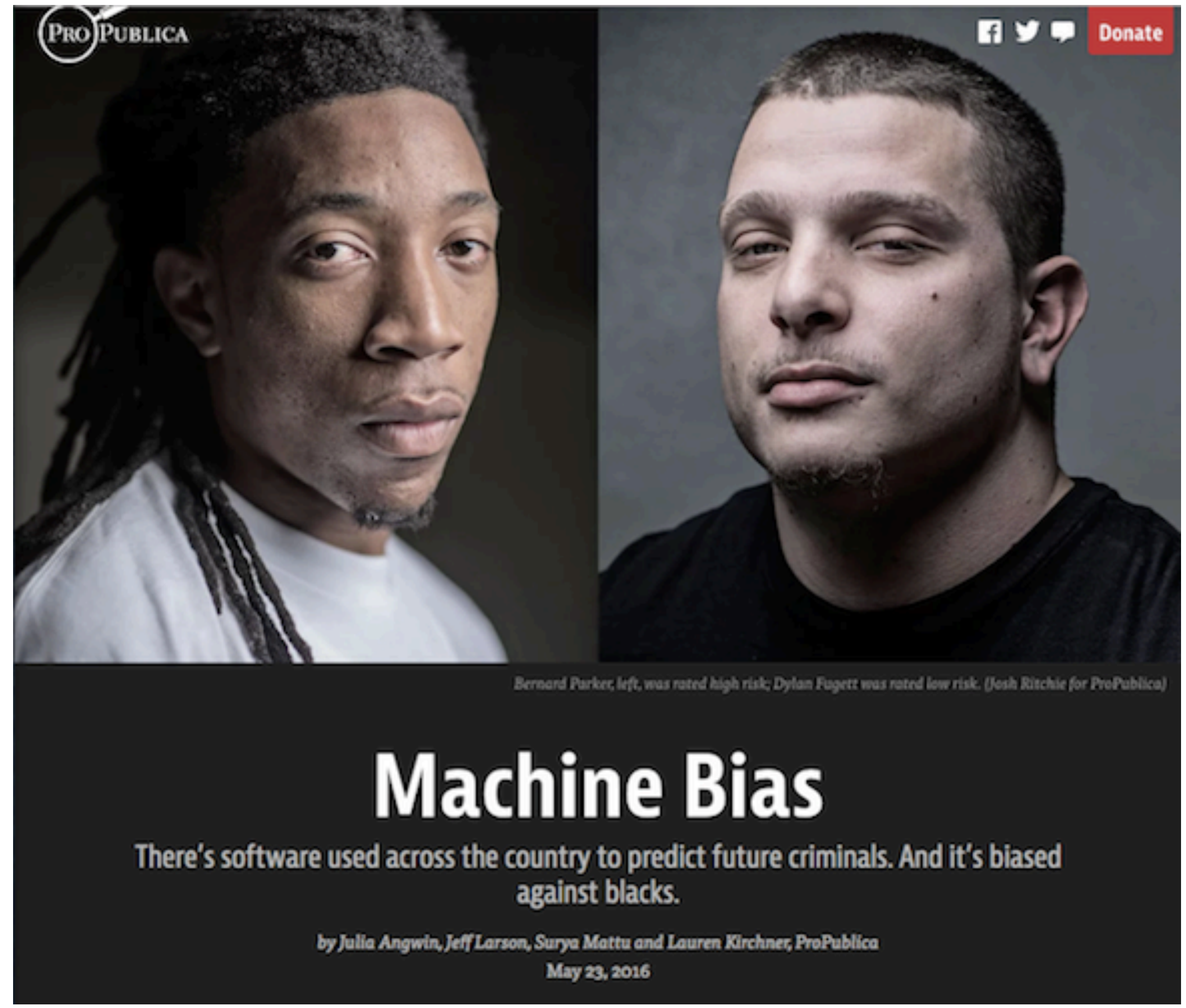
False positive rate (**FPR**)
P(C=1 | Y=0)

Positive predictive value (**PPV**) P(Y=1 | C=1)

Y is the *actual* outcome
C is the *classified* outcome

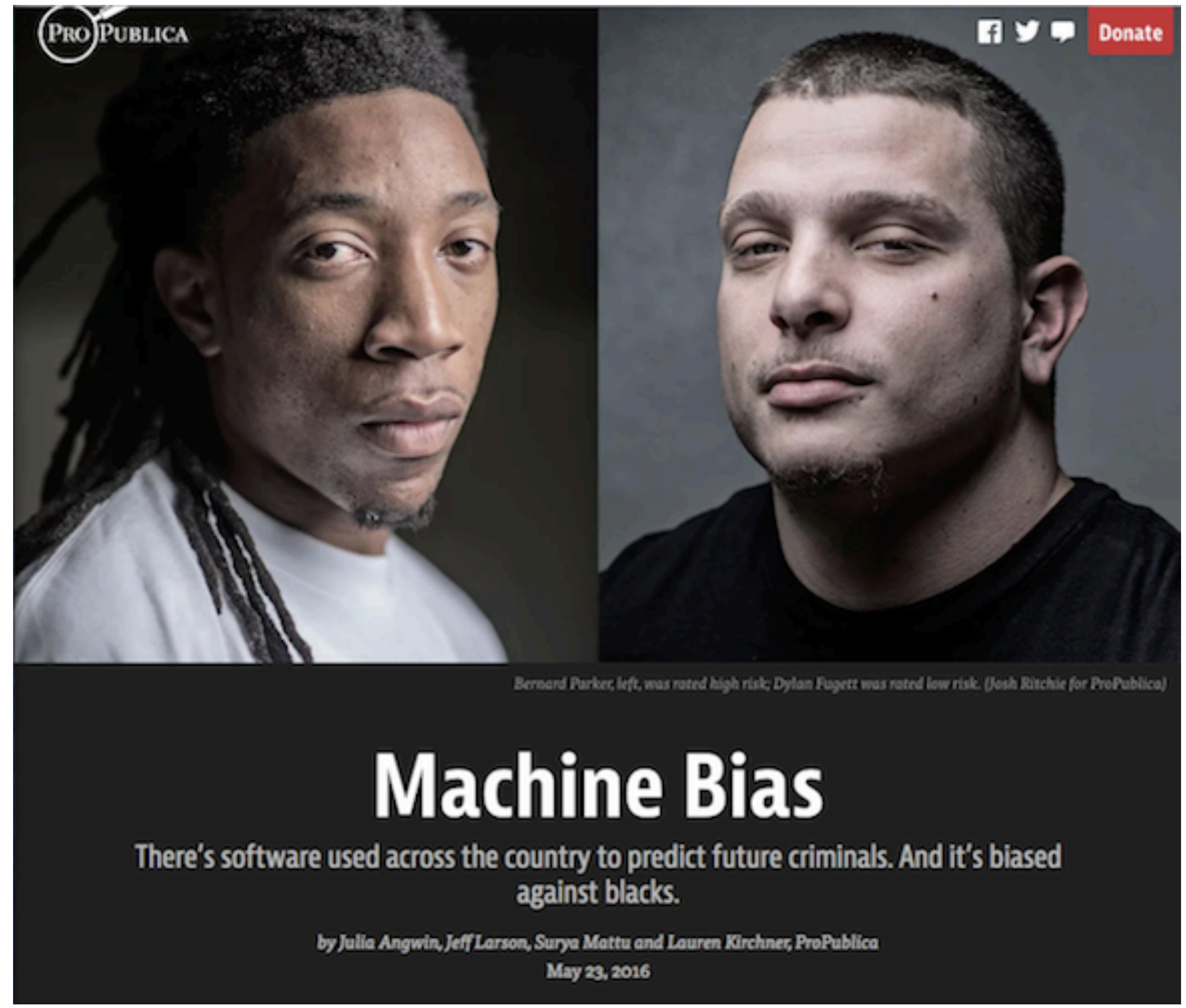


Back to COMPAS



Back to COMPAS

Even if ‘race’ is not among the predictive features used:



Back to COMPAS

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%

Even if 'race' is not among the predictive features used:

- False positive rate (**FPR**) was **higher for blacks** than whites

- False negative rate (**FNR**) was **higher for whites** than blacks

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

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

Back to COMPAS

Prediction Fails Differently for Black Defendants

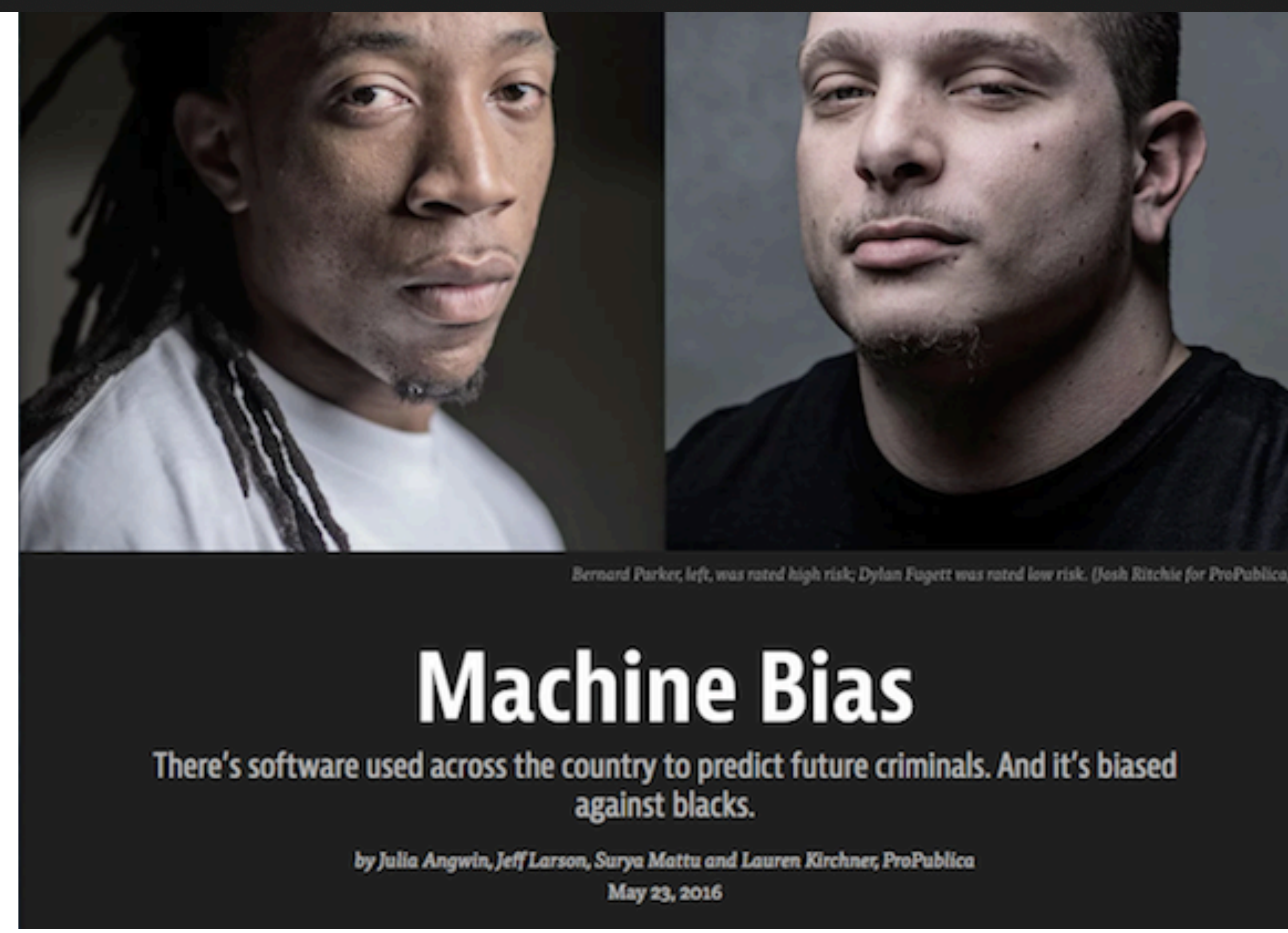
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Even if 'race' is not among the predictive features used:

- False positive rate (**FPR**) was **higher for blacks** than whites

- False negative rate (**FNR**) was **higher for whites** than blacks

- The positive predictive value (**PPV**) was the **same** for the two racial groups



Dichotomous Accuracy Metrics

False negative rate (**FNR**)
P(C=0 | Y=1)

False positive rate (**FPR**)
P(C=1 | Y=0)

Positive predictive value
(**PPV**) P(Y=1 | C=1)

Y is the *actual* outcome
C is the *classified* outcome

Dichotomous Accuracy Metrics

Dichotomous (Group) Fairness Metrics

False negative rate (**FNR**) P(C=0 | Y=1)

False positive rate (**FPR**) P(C=1 | Y=0)

Positive predictive value (**PPV**) P(Y=1 | C=1)

Y is the *actual* outcome
C is the *classified* outcome

Dichotomous Accuracy Metrics

Dichotomous (Group) Fairness Metrics

False negative rate (**FNR**) P(C=0 | Y=1)
False positive rate (**FPR**) P(C=1 | Y=0)
Positive predictive value (**PPV**) P(Y=1 | C=1)

Same FNR across groups: P(C=1 | Y=0 & G=1) = P(C=1 | Y=0 & G=0)
Same FPR across groups: P(C=0 | Y=1 & G=1) = P(C=0 | Y=1 & G=0)
Same PPV across groups: P(Y=1 | C=1 & G=1) = P(Y=1 | C=1 & G=0)

Y is the *actual* outcome
C is the *classified* outcome

Dichotomous Accuracy Metrics

False negative rate (**FNR**) P(C=0 | Y=1)
False positive rate (**FPR**) P(C=1 | Y=0)
Positive predictive value (**PPV**) P(Y=1 | C=1)

Y is the *actual* outcome
C is the *classified* outcome

Dichotomous (Group) Fairness Metrics

Same FNR across groups: P(C=1 | Y=0 & G=1) = P(C=1 | Y=0 & G=0)
Same FPR across groups: P(C=0 | Y=1 & G=1) = P(C=0 | Y=1 & G=0)
Same PPV across groups: P(Y=1 | C=1 & G=1) = P(Y=1 | C=1 & G=0)

`Classification parity`

Dichotomous Accuracy Metrics

False negative rate (**FNR**)
$P(C=0 \mid Y=1)$

False positive rate (**FPR**)
$P(C=1 \mid Y=0)$

Positive predictive value (**PPV**) $P(Y=1 \mid C=1)$

Y is the *actual* outcome
C is the *classified* outcome

Dichotomous (Group) Fairness Metrics

Same FNR across groups:
$P(C=1 \mid Y=0 \ \& \ G=1) = P(C=1 \mid Y=0 \ \& \ G=0)$

Same FPR across groups:
$P(C=0 \mid Y=1 \ \& \ G=1) = P(C=0 \mid Y=1 \ \& \ G=0)$

Same PPV across groups:
$P(Y=1 \mid C=1 \ \& \ G=1) = P(Y=1 \mid C=1 \ \& \ G=0)$

`Predictive parity`

`Classification parity`

Chouldechova’s Impossibility Theorem

No predictive model or algorithm can concurrently satisfy

- same **FP** and **FN** rate (`classification parity`)
- same **PPV** (`predictive parity`)

Provided

1. the groups have different prevalence rates
2. the model or algorithm is not infallible

Chouldechova's Impossibility Theorem

No predictive model or algorithm can concurrently satisfy

- same **FP** and **FN** rate (`classification parity`)
- same **PPV** (`predictive parity`)

Provided

1. the groups have different prevalence rates
2. the model or algorithm is not infallible

$$\frac{P(Y=1 \mid C=1)}{P(Y=0 \mid C=1)} = \frac{P(C=1 \mid Y=1)}{P(C=1 \mid Y=0)} \times \frac{P(Y=1)}{P(Y=0)}$$

$$\frac{PPV}{1-PPV} = \frac{1-FN}{FP} \times \text{prevalence ratio}$$

Chouldechova's Impossibility Theorem

No predictive model or algorithm can concurrently satisfy

- same **FP** and **FN** rate (`classification parity`)
- same **PPV** (`predictive parity`)

Provided

1. the groups have different prevalence rates
2. the model or algorithm is not infallible

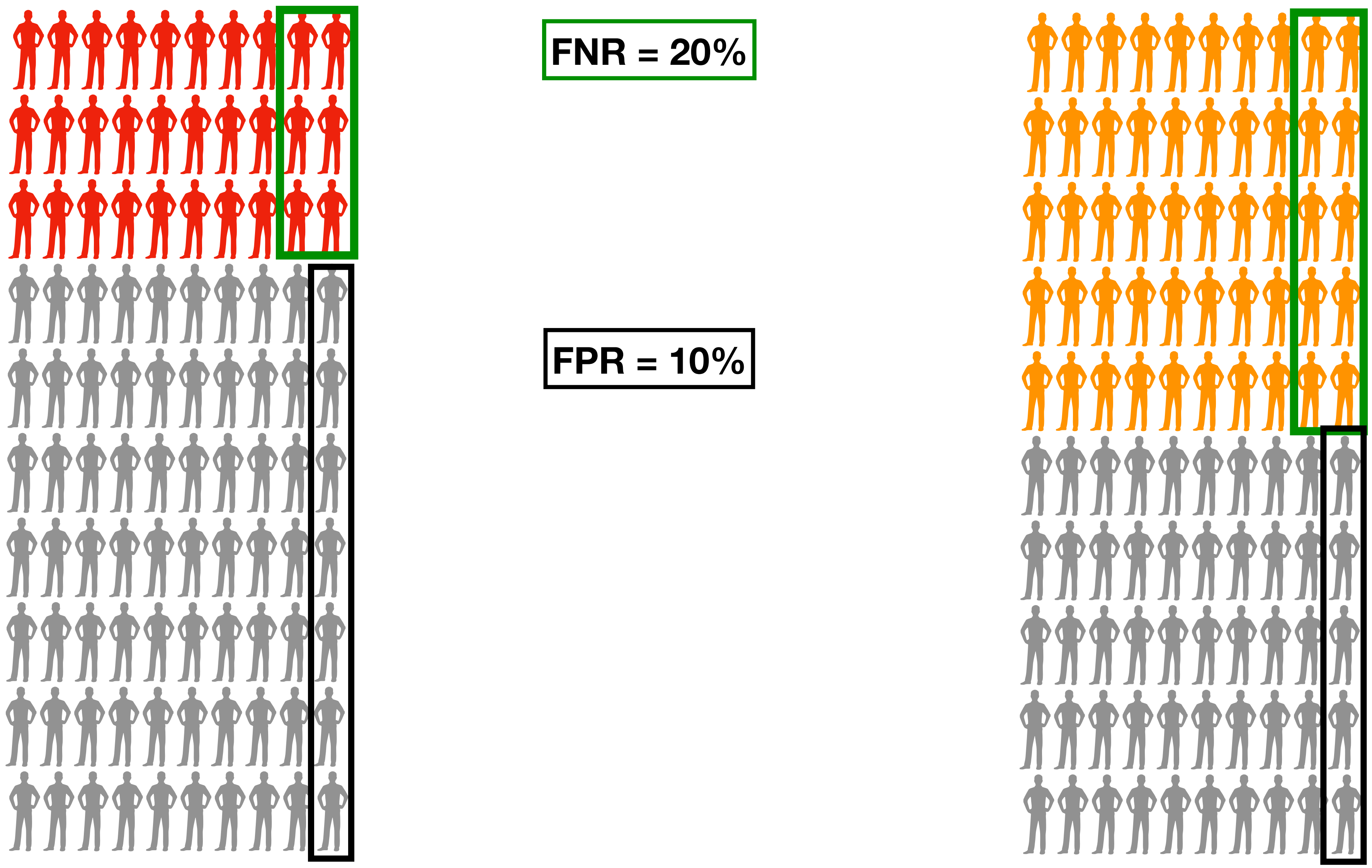
$$\frac{P(Y=1 \mid C=1)}{P(Y=0 \mid C=1)} = \frac{P(C=1 \mid Y=1)}{P(C=1 \mid Y=0)} \times \frac{P(Y=1)}{P(Y=0)}$$

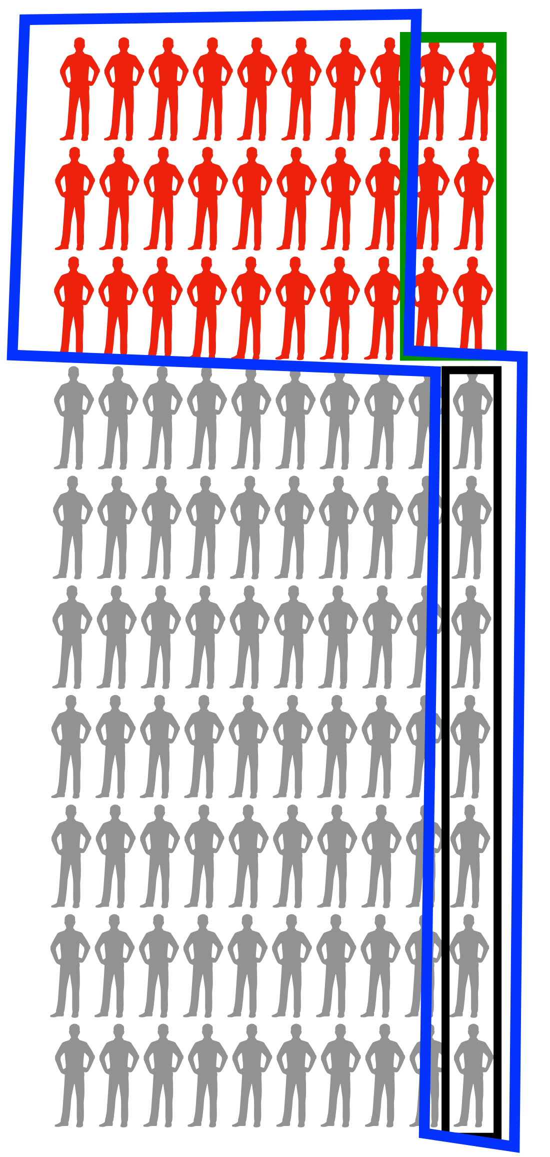
$$\frac{PPV}{1-PPV} = \frac{1-FN}{FP} \times \text{prevalence ratio}$$

If **PPV** is the same across groups, then **FN** and **FP** rates must be different unless prevalence rates are the same

If **FN** and **FP** rates are the same across groups, then **PPV** must be different unless the prevalence rates are the same

Suppose FPR and FNR Are the Same Across Two Groups

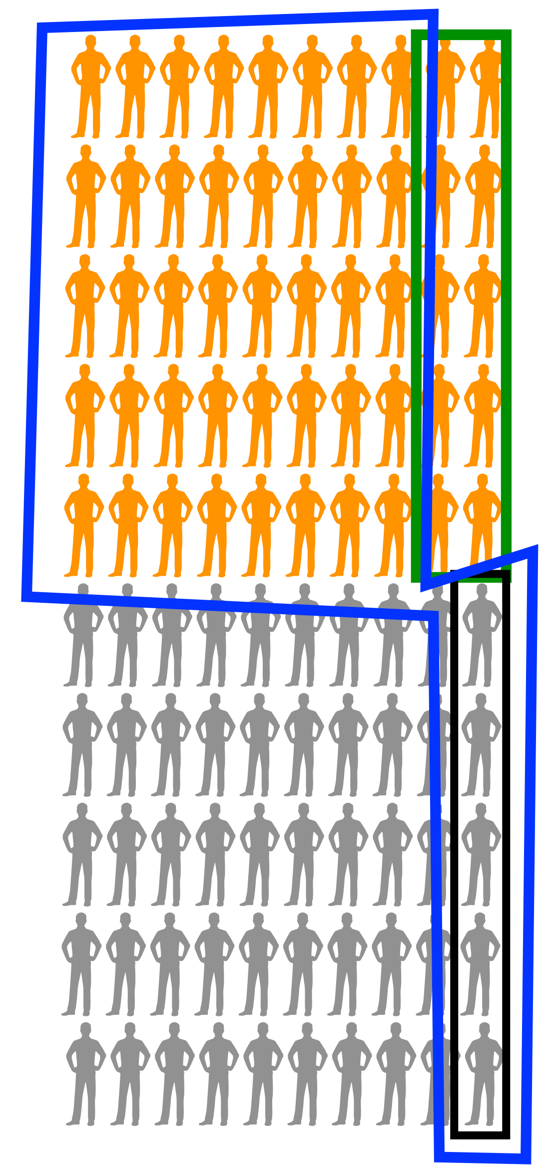

FNR = 20%
FPR = 10%



FNR = 20%

FPR = 10%

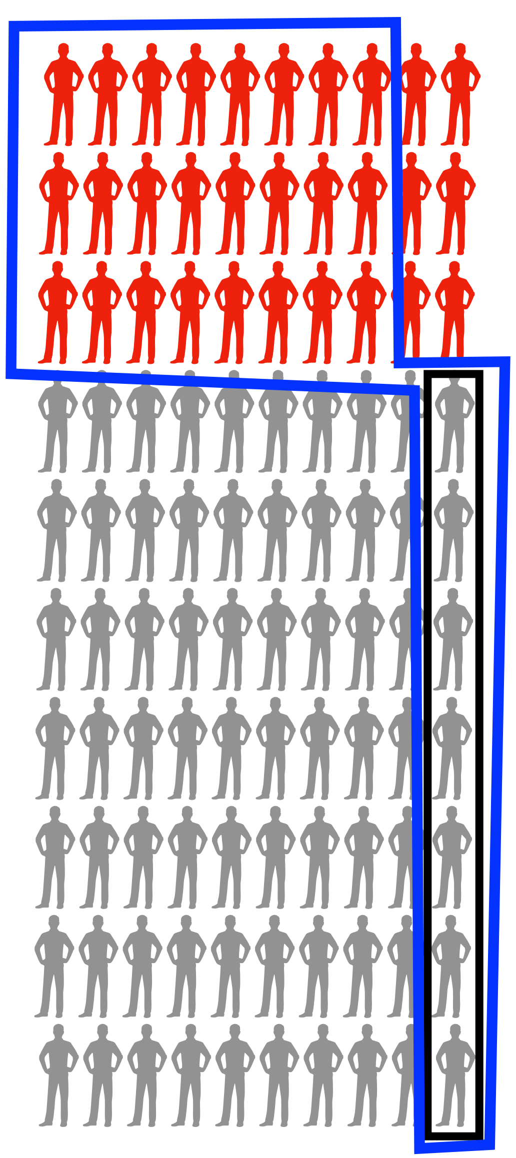
$$PPV = \frac{TP}{\boxed{P}}$$

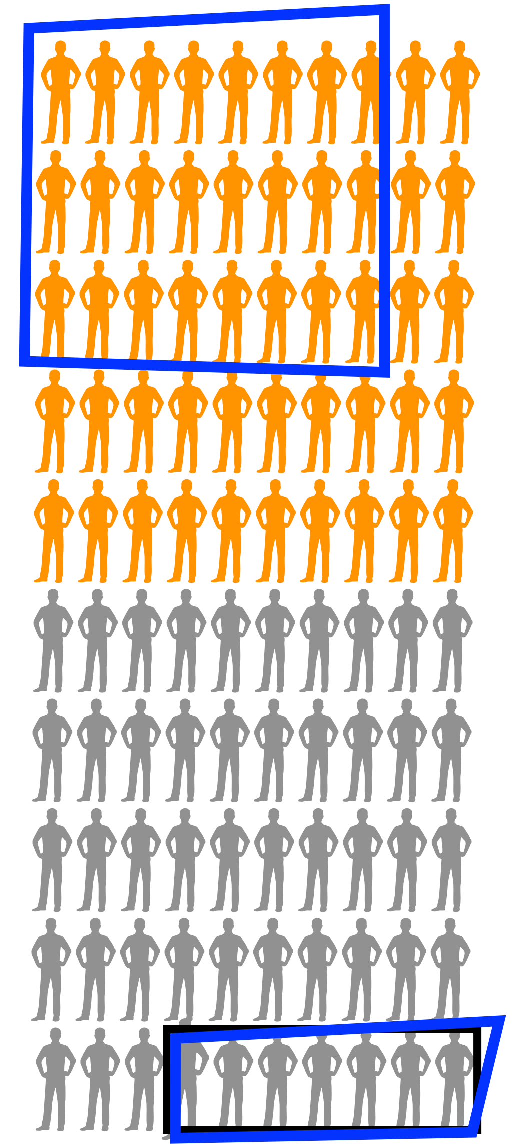


$$PPV_1 = \frac{24\,\%}{\boxed{24\% + 7\,\%}} \approx 77\,\%$$

$$PPV_2 = \frac{40\,\%}{\boxed{40\% + 5\,\%}} \approx 88\,\%$$

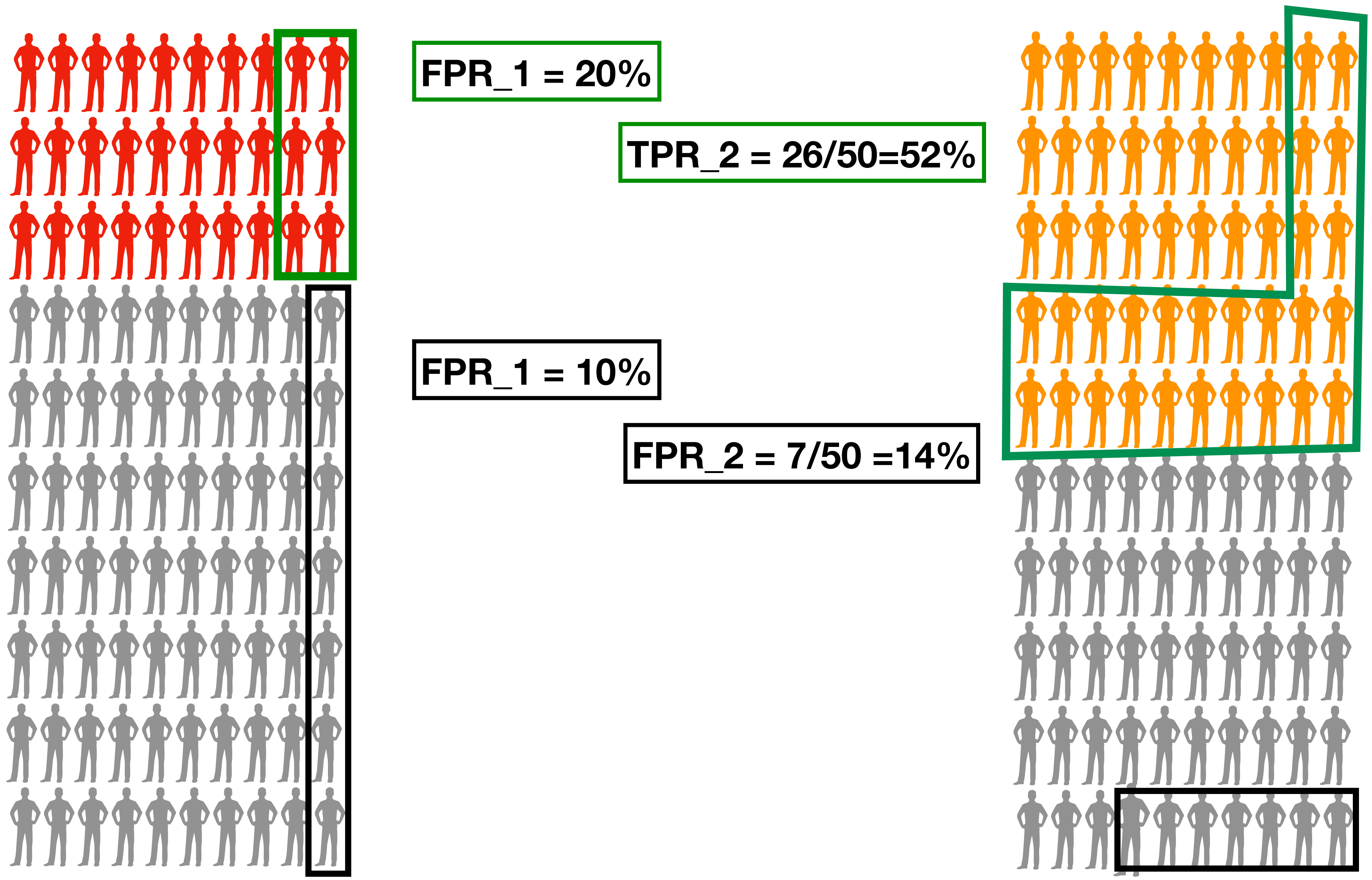
What If PPV Is the Same Across Groups?





$$PPV_1 = \frac{24\,\%}{\boxed{24\% + 7\,\%}} \approx 77\,\%$$

$$PPV_2 = \frac{24\,\%}{\boxed{24\% + 7\,\%}} \approx 77\,\%$$



$$PPV_1 = \frac{24\,\%}{24\% + 7\,\%} \approx 77\,\%$$

$$PPV_2 = \frac{24\,\%}{24\% + 7\,\%} \approx 77\,\%$$

**If Base Rates Are Different,
It Is Impossible to Have
the Same PPV (*Predictive Parity*)
and the Same FPR and FNR
(*Classification Parity*)
Across Groups**

There Are Other Impossibility Theorems

Chouldechova's Is the Easiest