

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

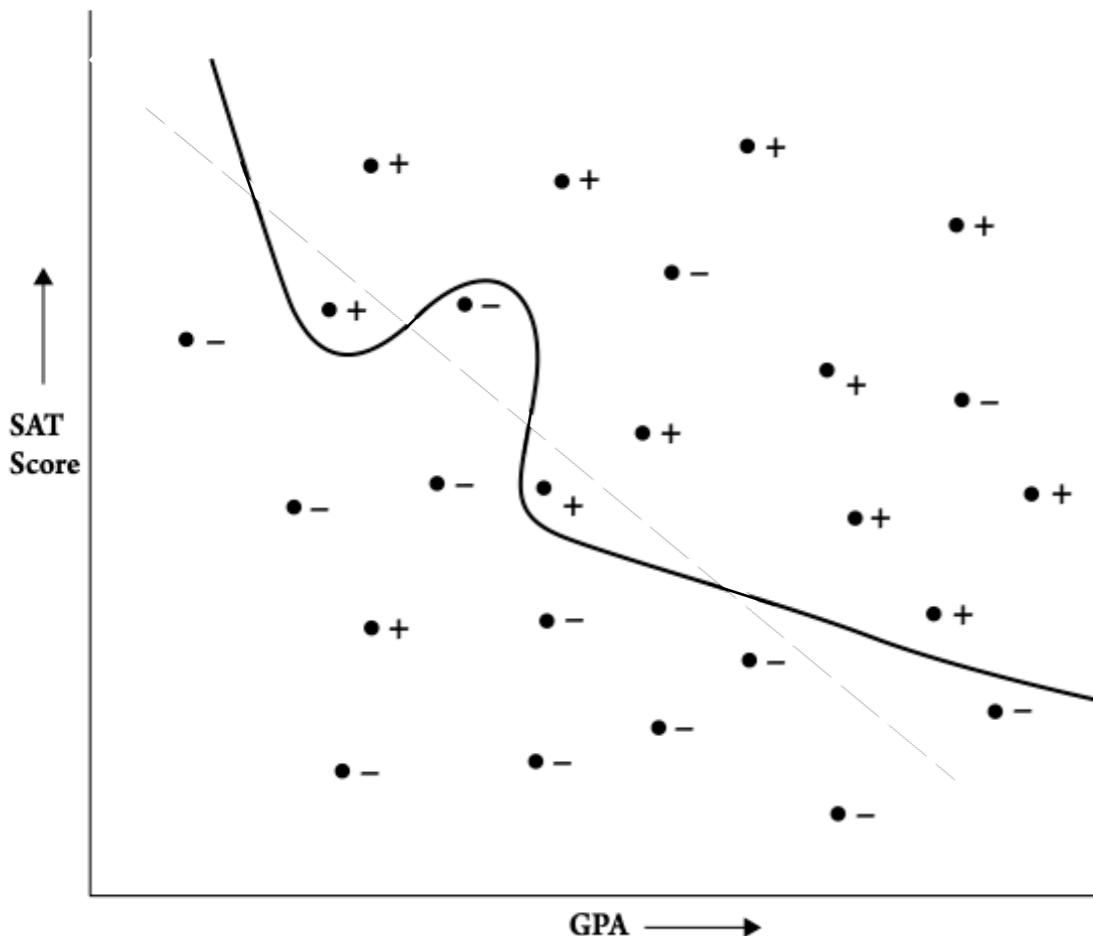
PART V: Impossibility Theorems (time permitting)

Part I

Introduction to Predictive Algorithms (or Predictive Models)

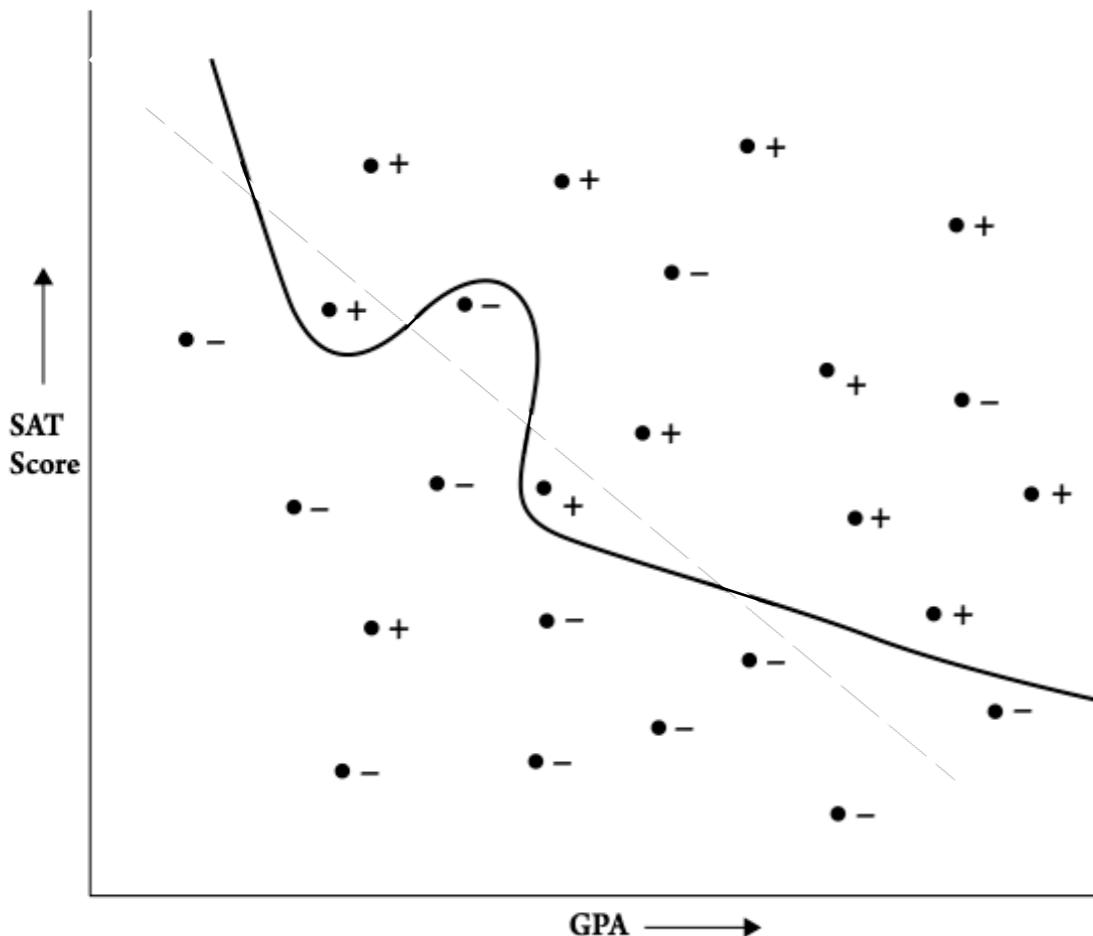
Predictive Algorithms (or Predictive Models)

(binary case)



Predictive Algorithms (or Predictive Models)

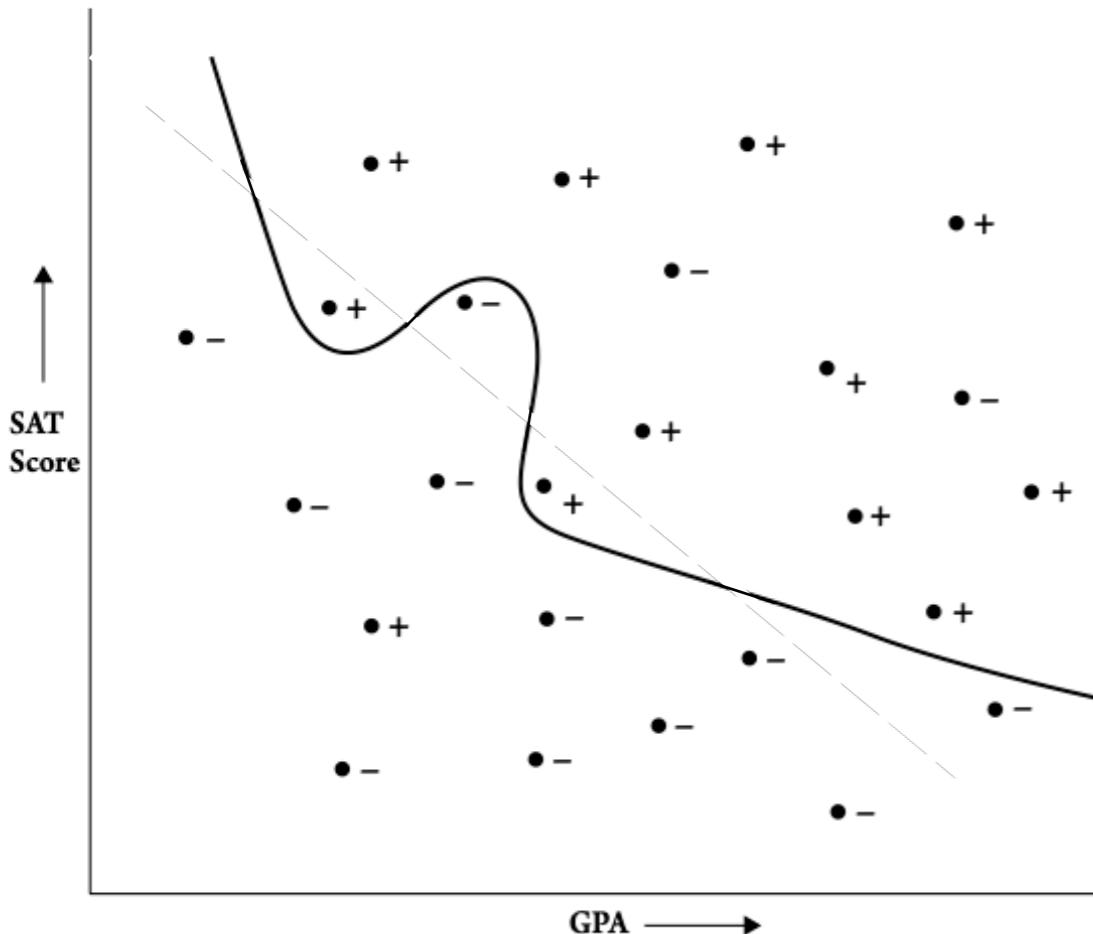
(binary case)



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)

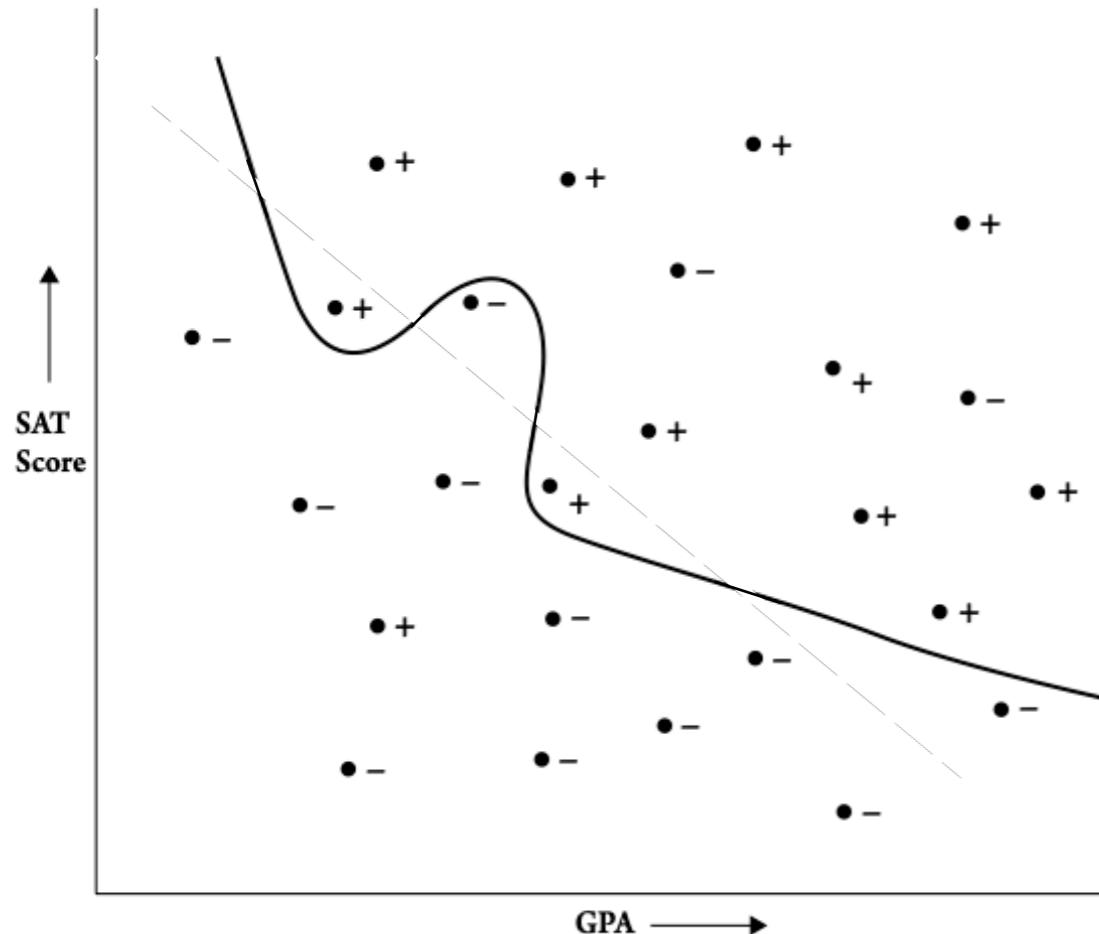


Suppose we aim to make predictions about a **binary outcome** $Y=1$ or $Y=0$ (e.g. college success, recidivism)

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*)

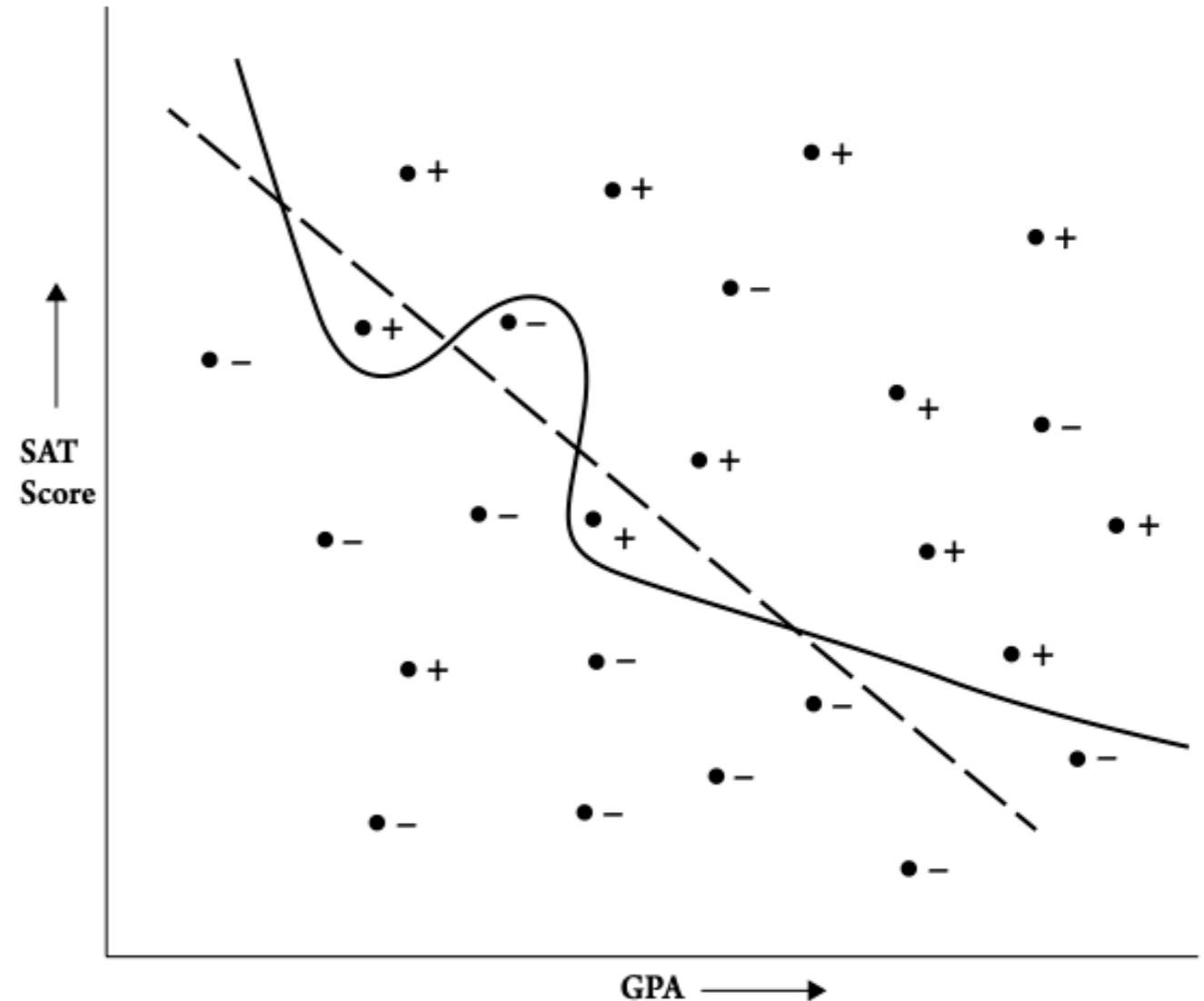


Suppose we aim to make predictions about a **binary outcome $Y=1$ or $Y=0$** (e.g. college success, recidivism)

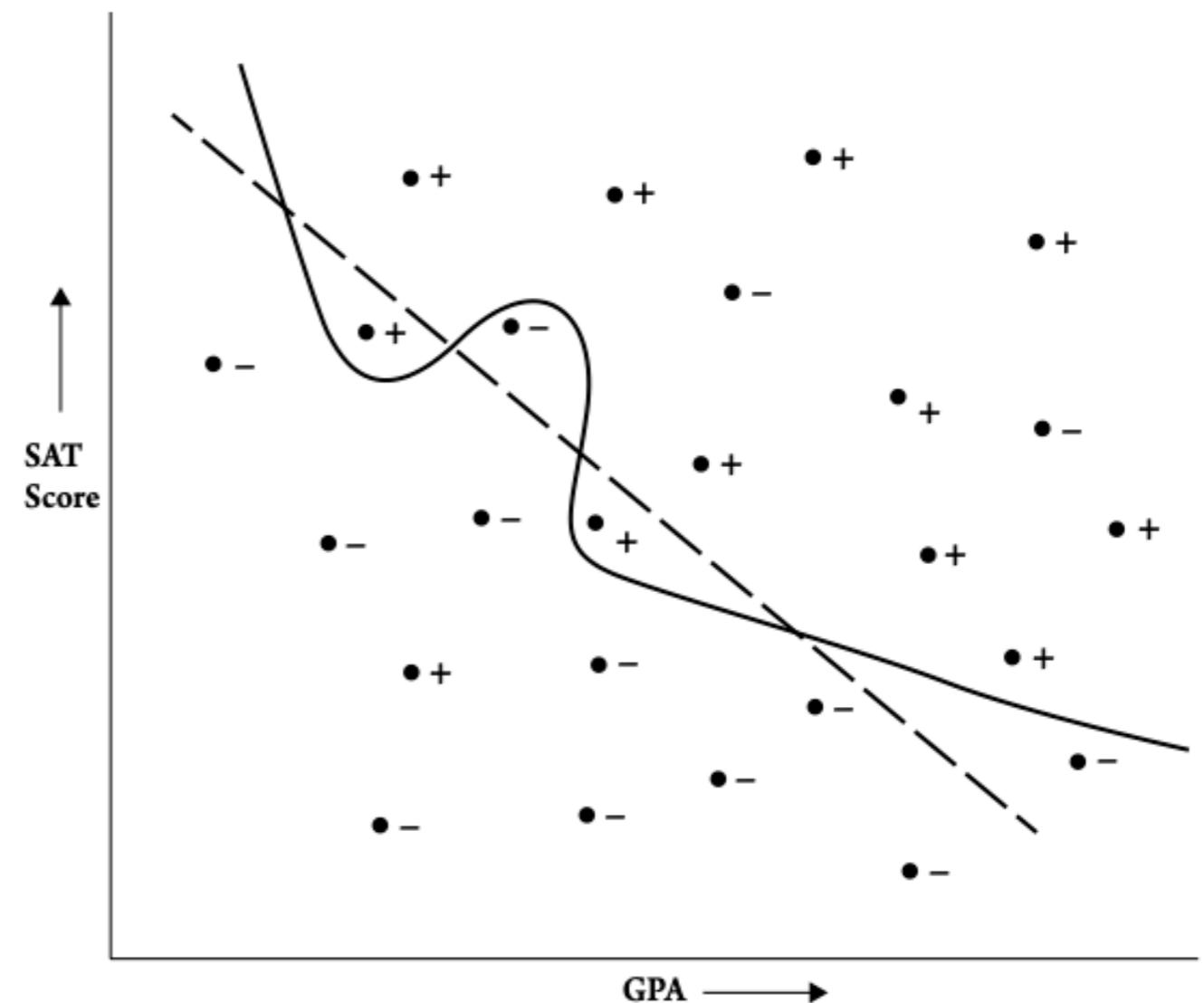
Machine learning algorithms (e.g. regression, SVM) mine the historical data and identify relationships between **predictive features** (e.g. GPA, income) and the outcome

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)

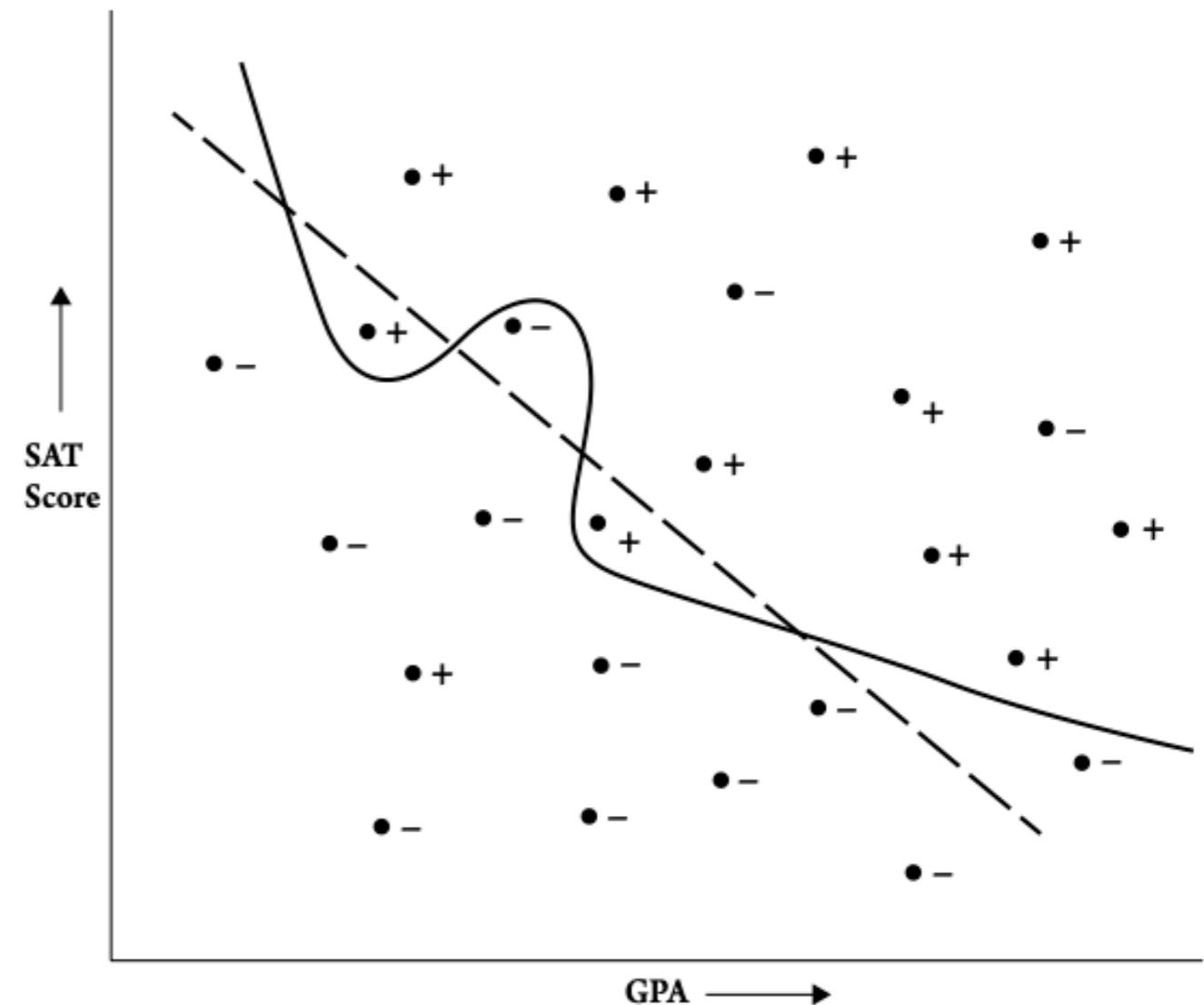


Machine Learning
algorithms are self-
programming.



Machine Learning
algorithms are self-
programming.

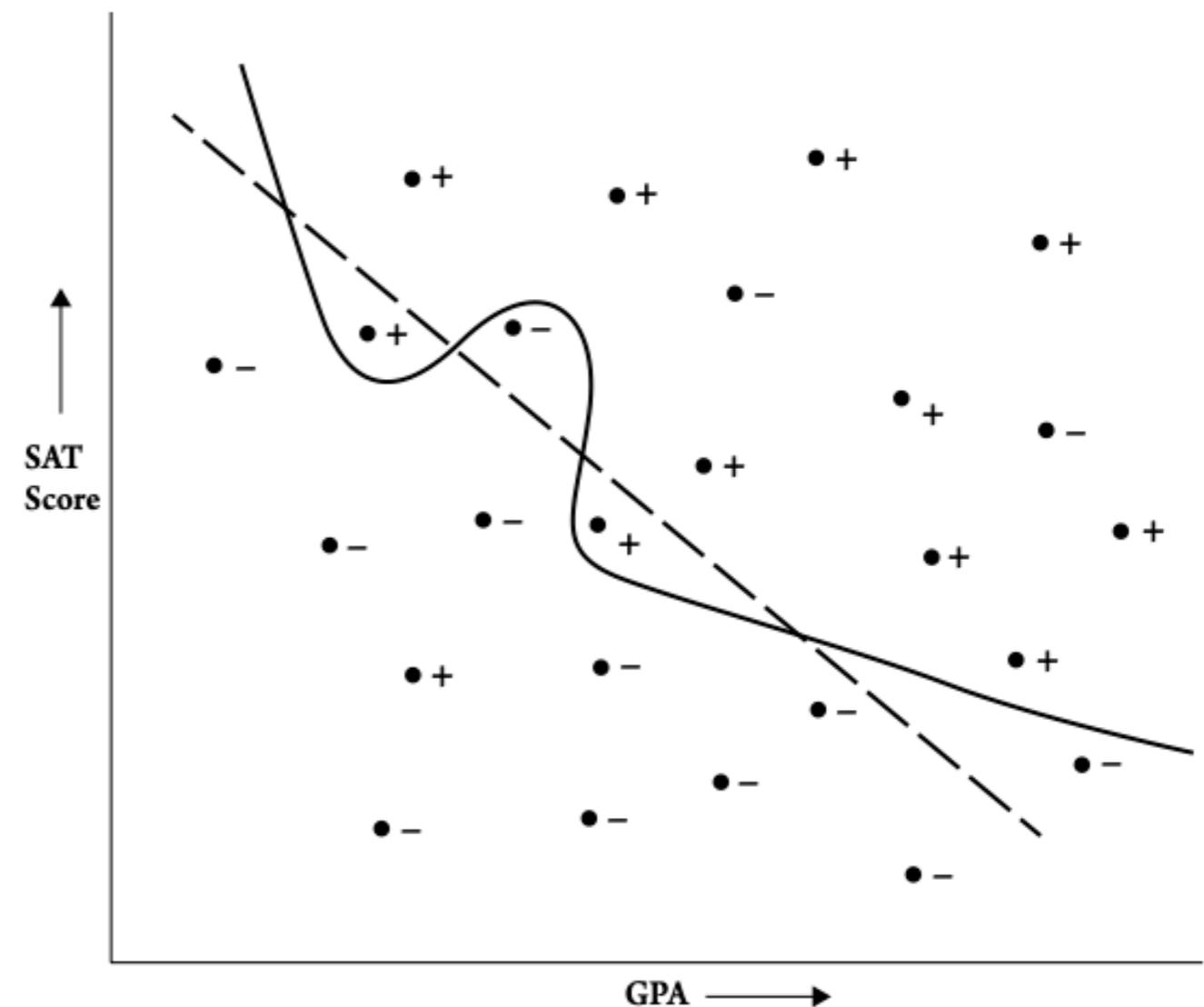
They are **meta-algorithms**
whose input are historical
data and whose output is
another algorithm.

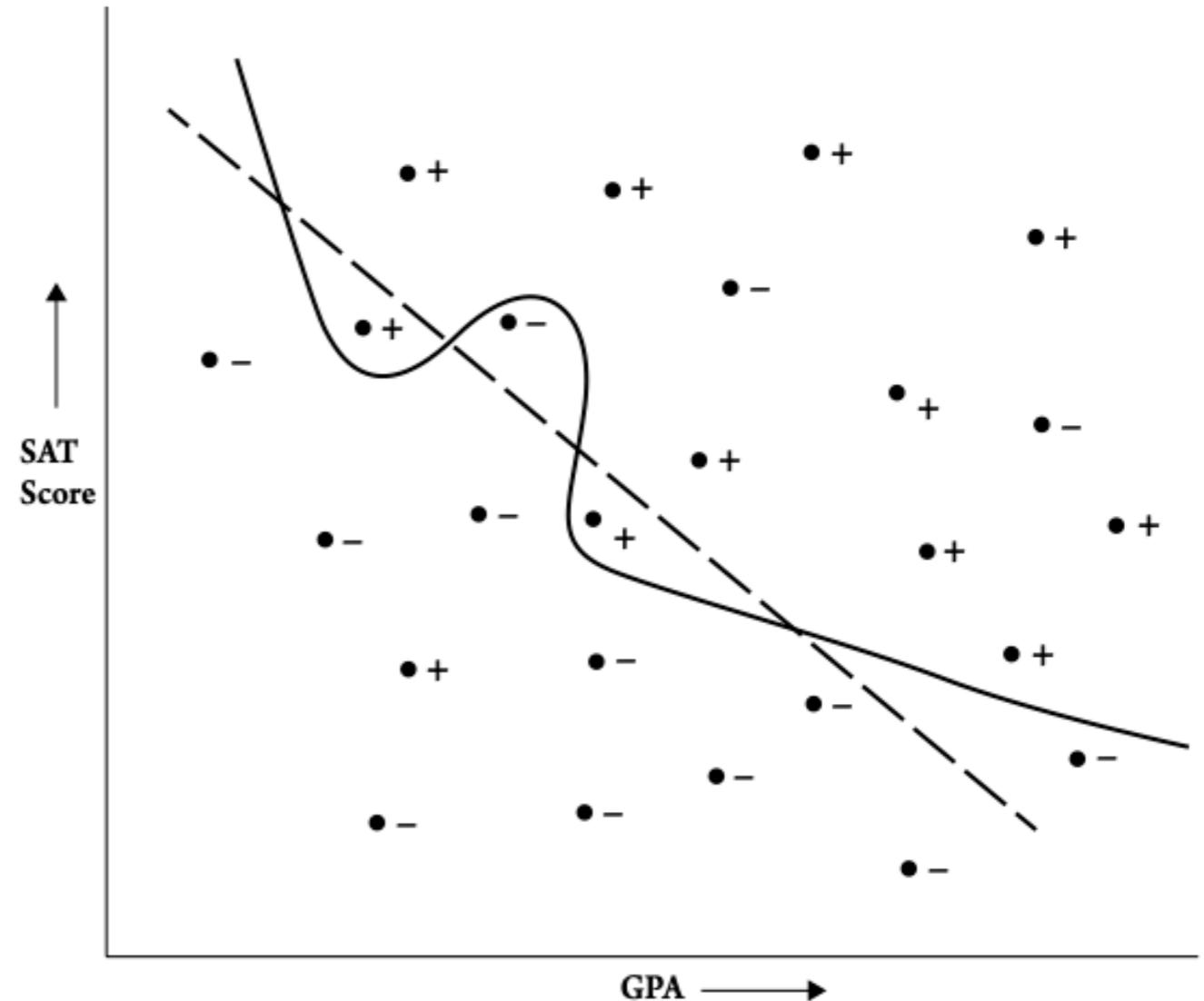


Machine Learning
algorithms are self-
programming.

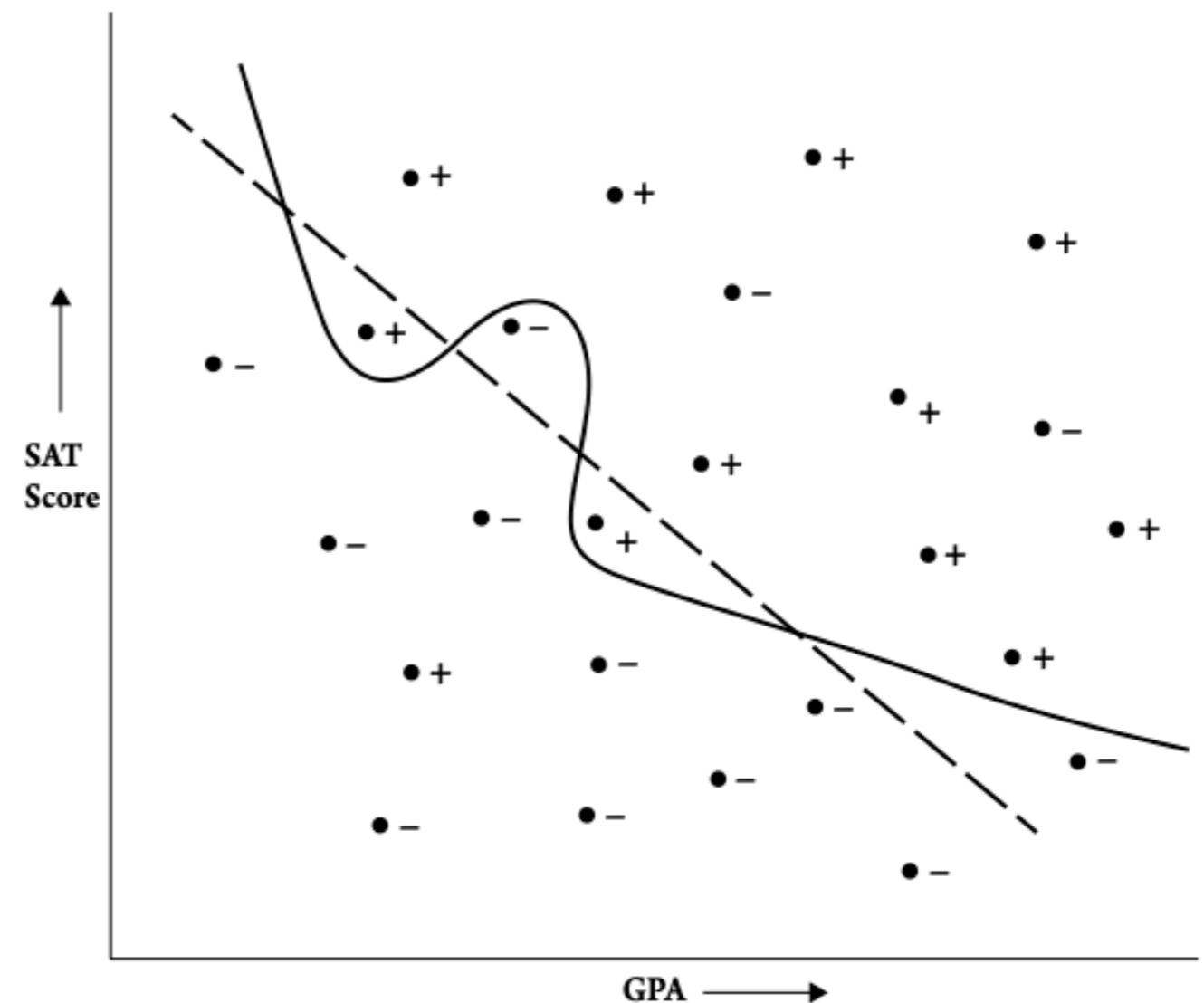
They are **meta-algorithms**
whose input are historical
data and whose output is
another algorithm.

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



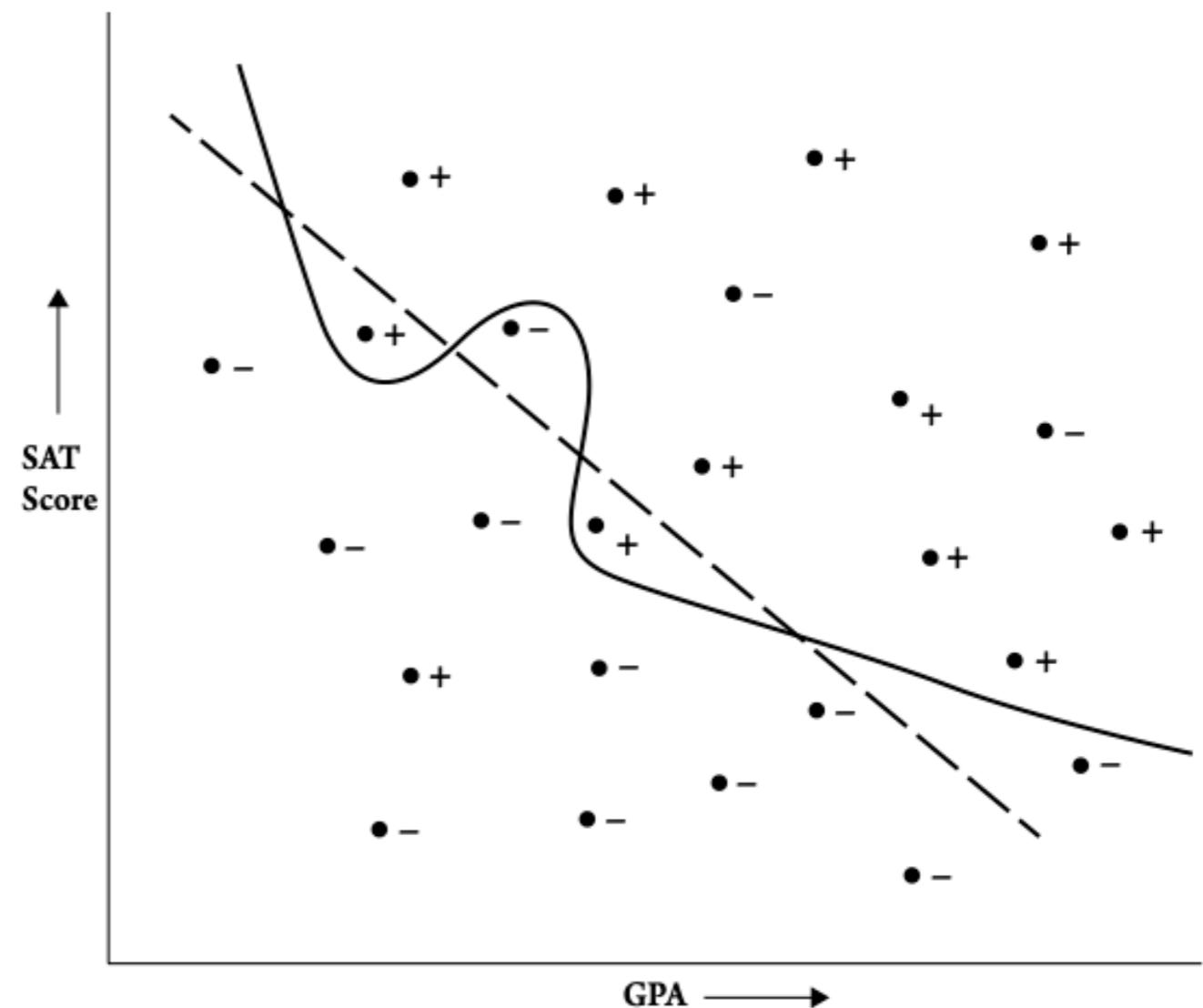


The meta-algorithm searches through all models, say possible lines through the data.



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

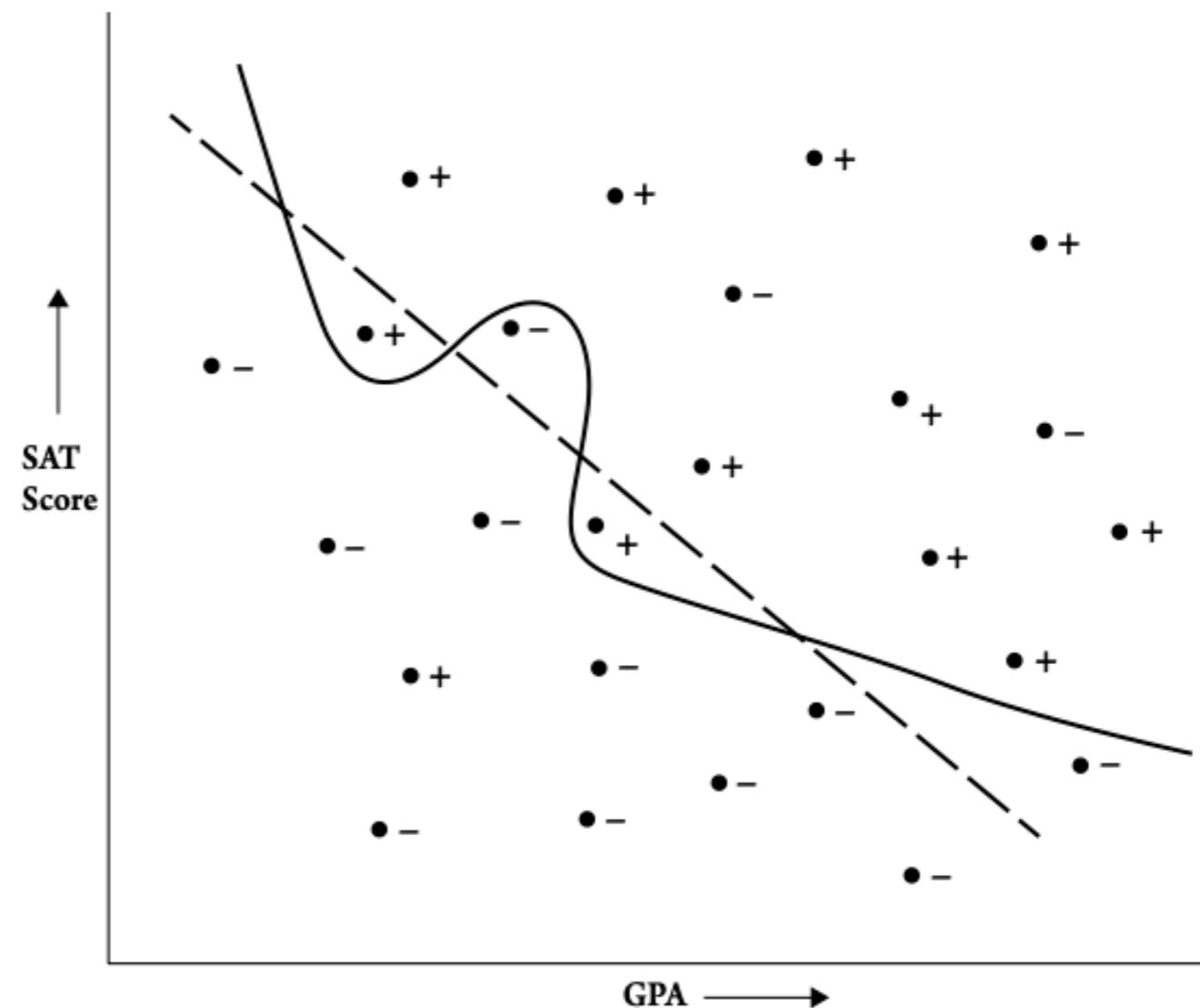
Lines are good for 2-dimensional data (e.g. SAT and GPA) and a binary outcome (graduate/not graduate).



The meta-algorithm searches through all models, say possible 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.

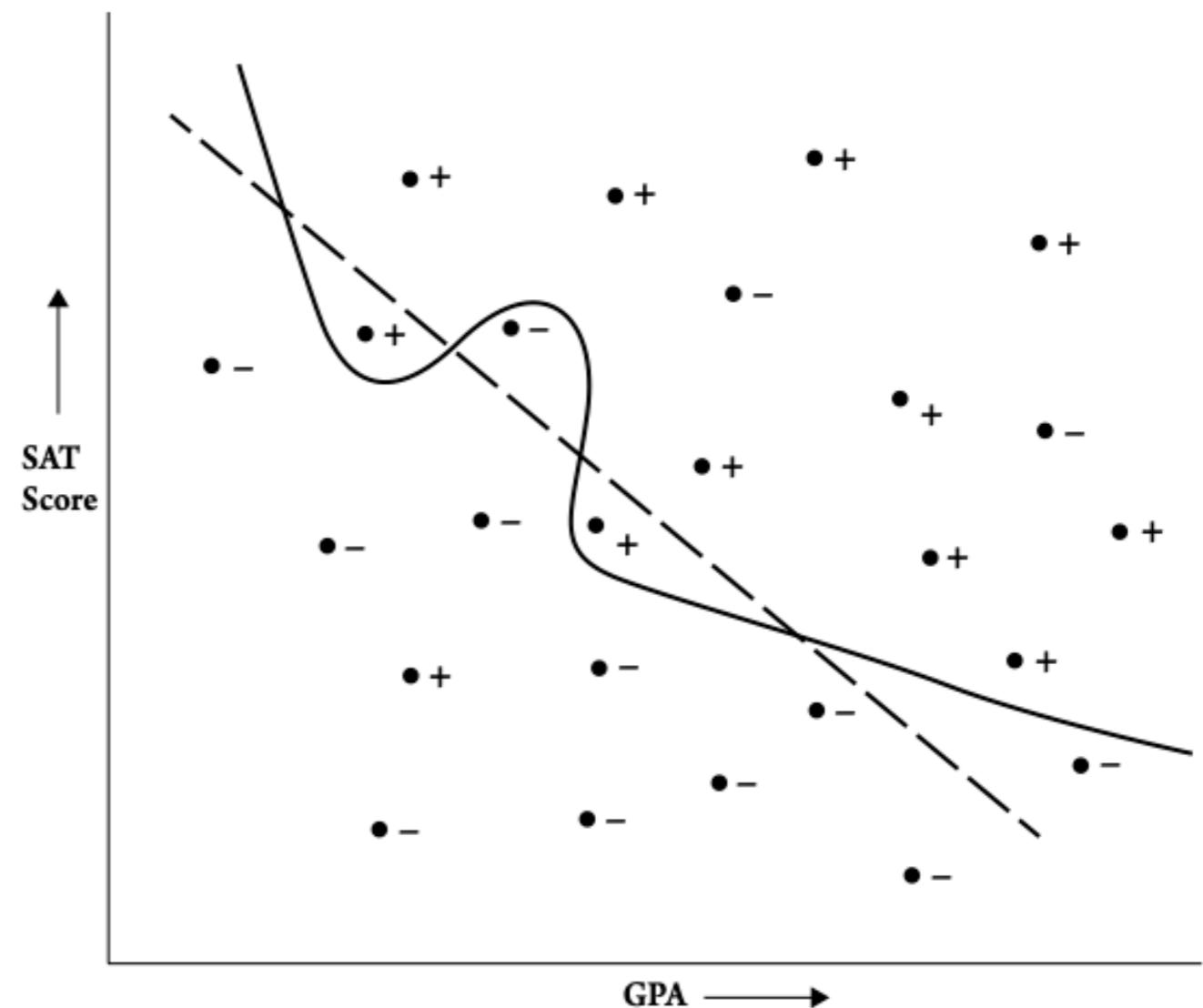


The meta-algorithm searches through all models, say possible lines through the data.

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.

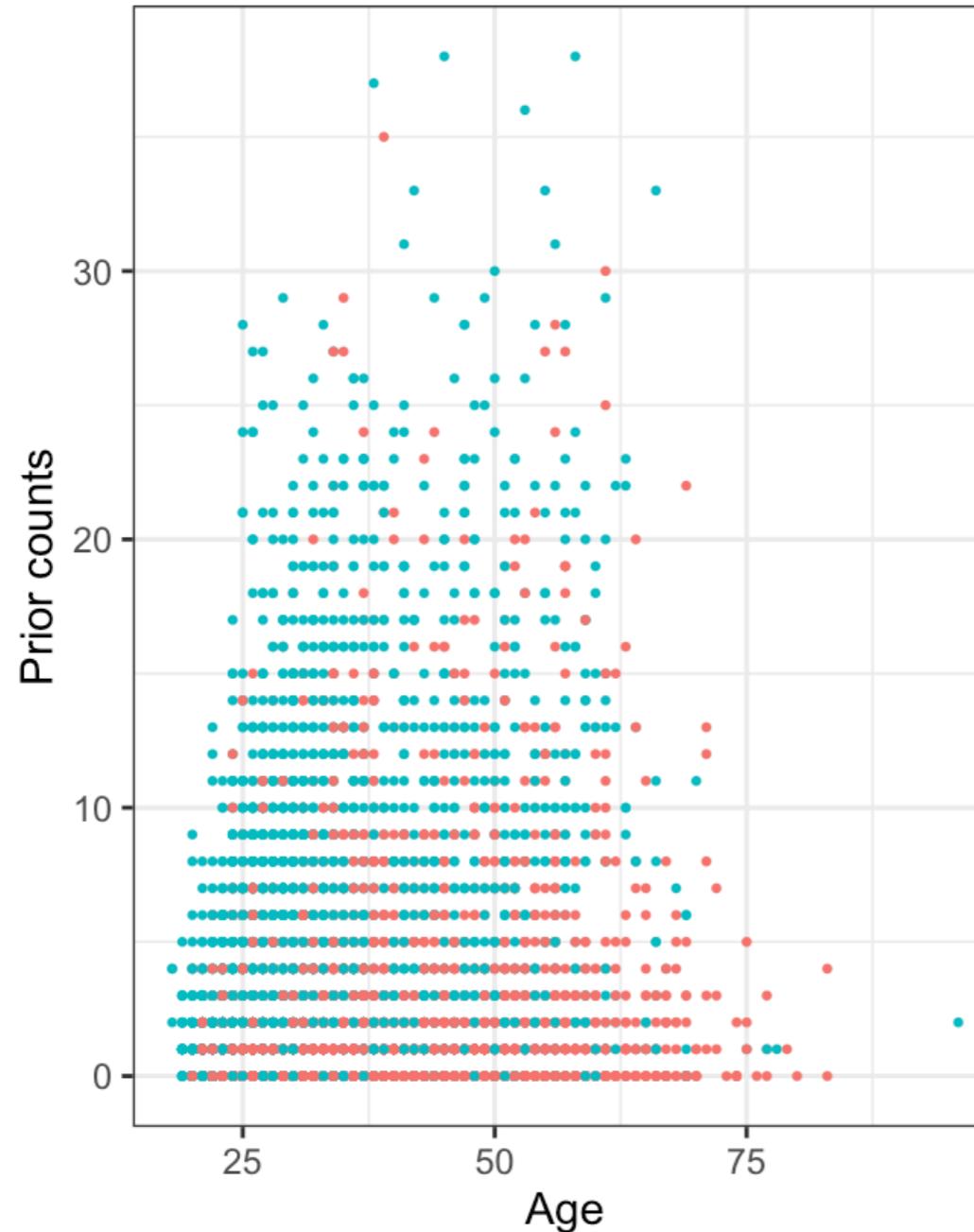
This is an example of ***supervised learning***. The model learns by comparing its prediction with the actual outcome in the training data.



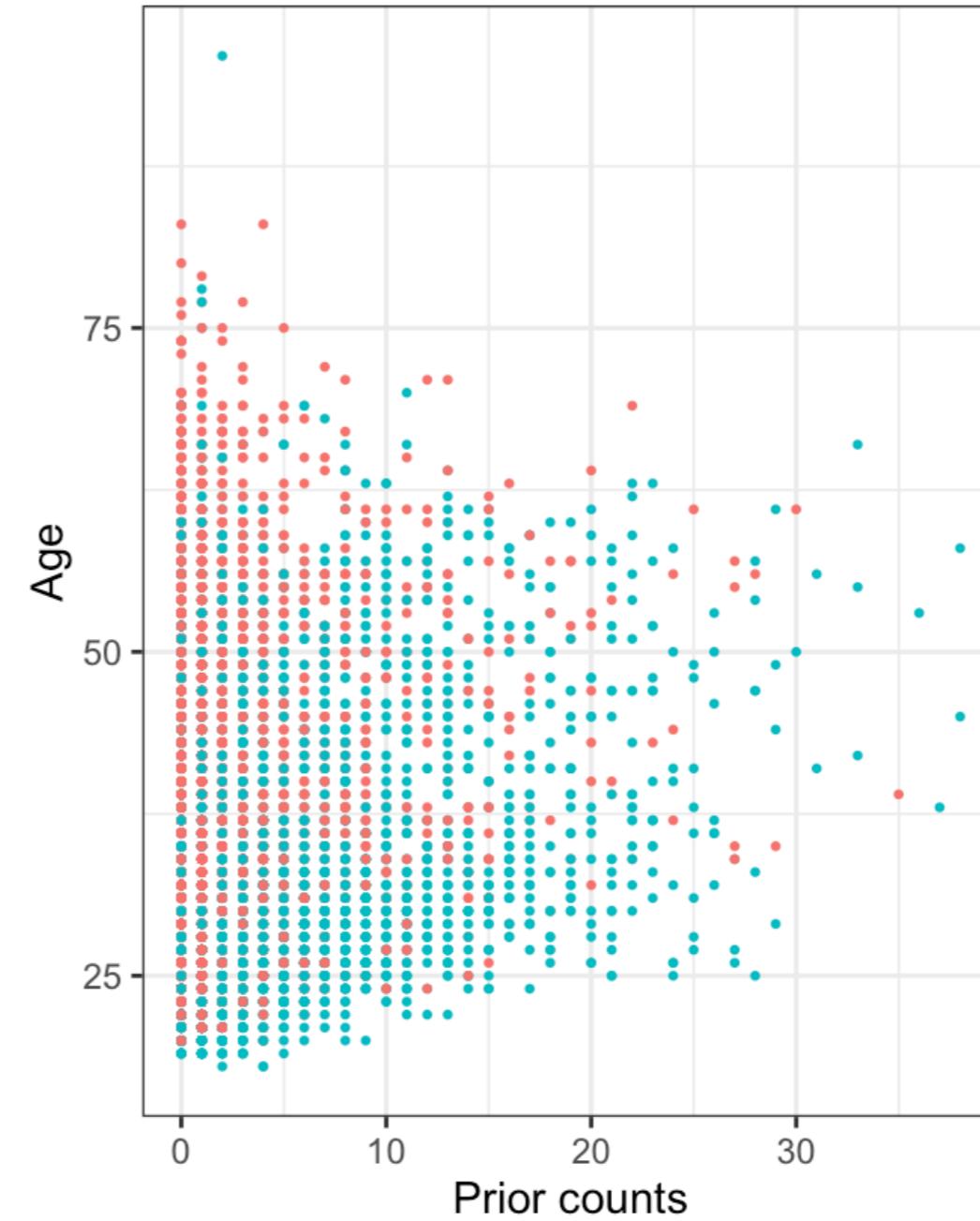
Toy Example Support Vector Machine (SVM) Algorithm

Historical Data: Age, Prior Counts, Reoffending

Age, Prior counts and Recidivism



Age, Prior counts and Recidivism

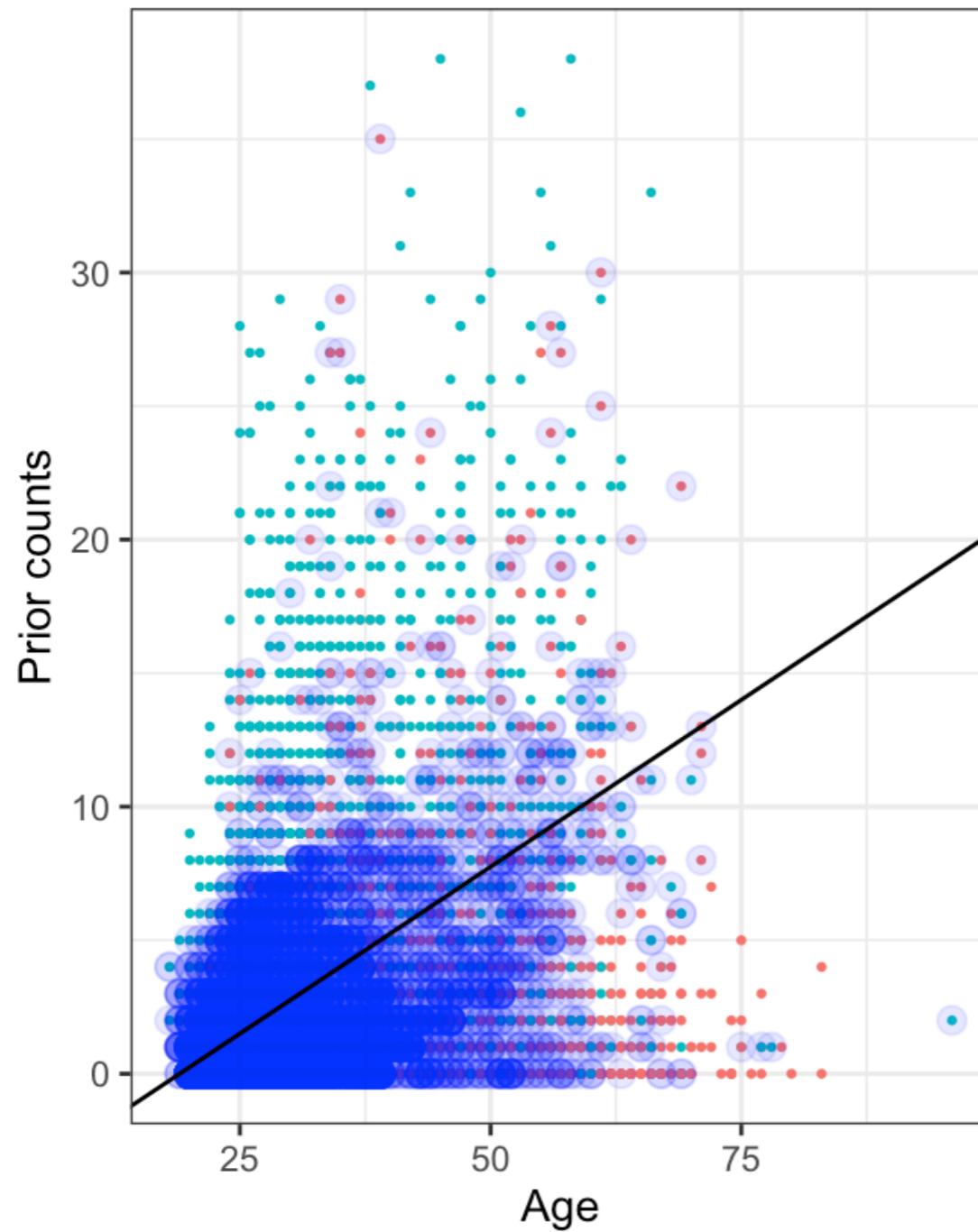


Reoffender (1) or not (0) • 1 • 0

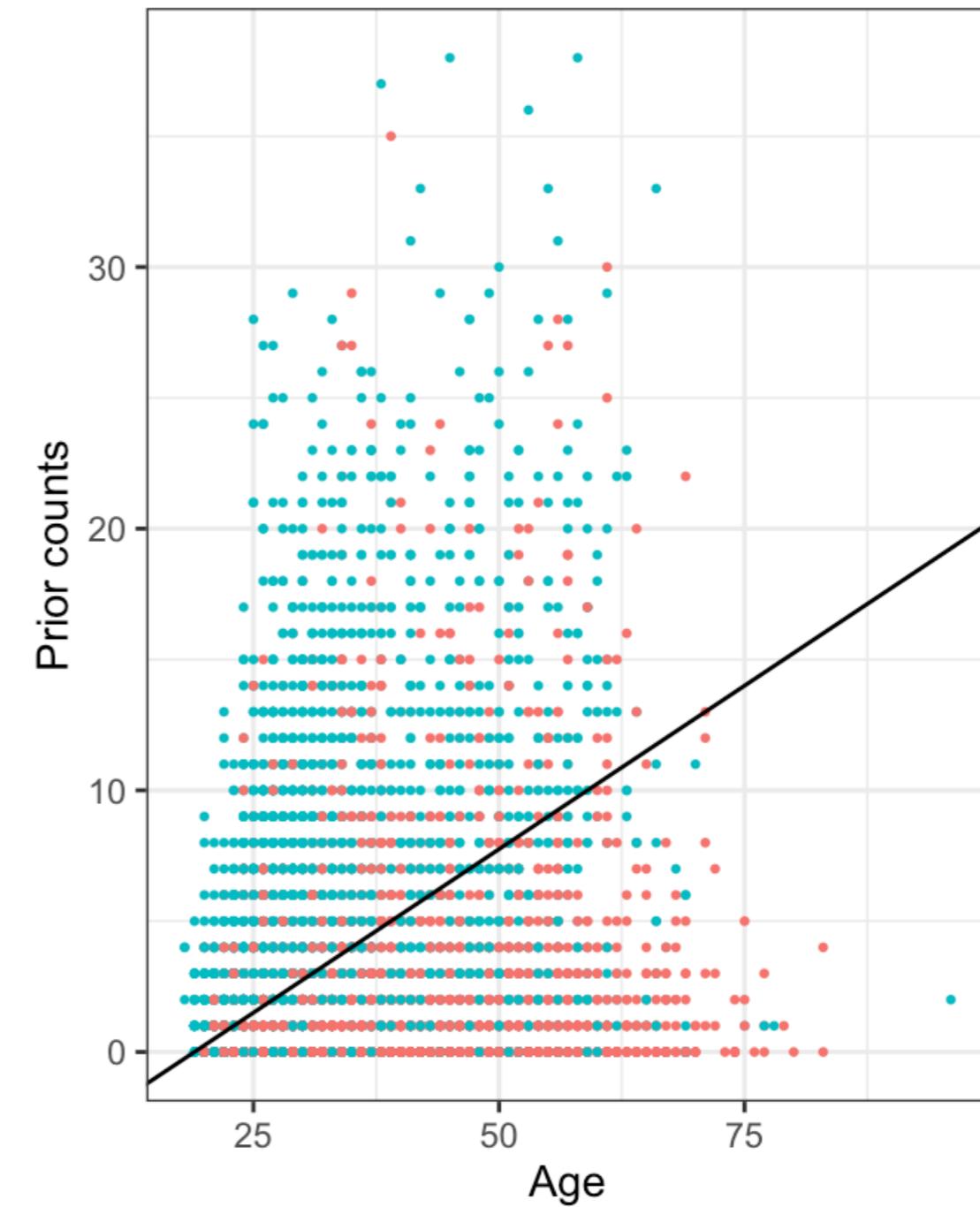
**It is not at all clear where the line
(predictive model) should be
drawn to minimize errors**

SVM Risk Model: Support Vectors and Line

Support vectors and linear model



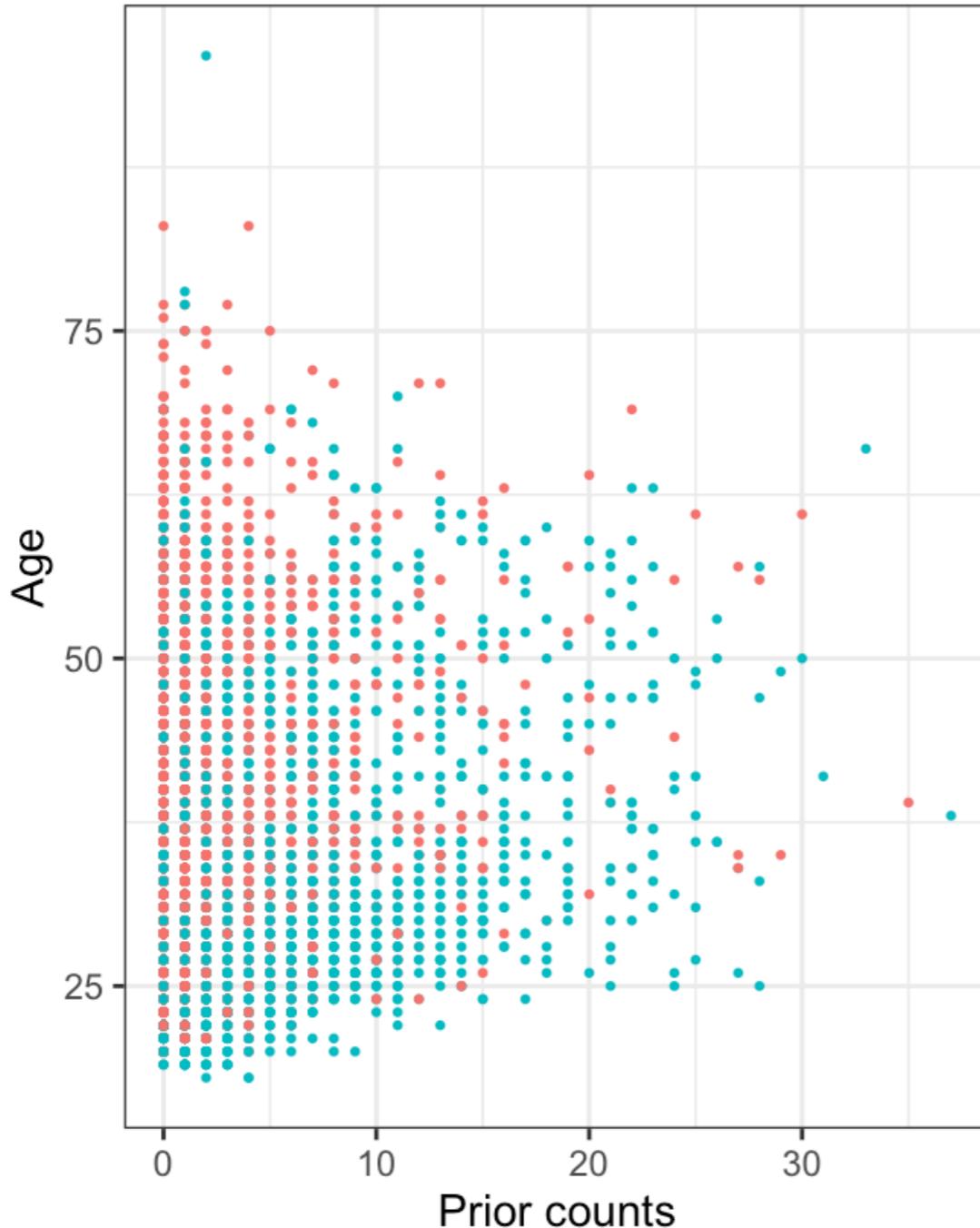
Linear model (only)



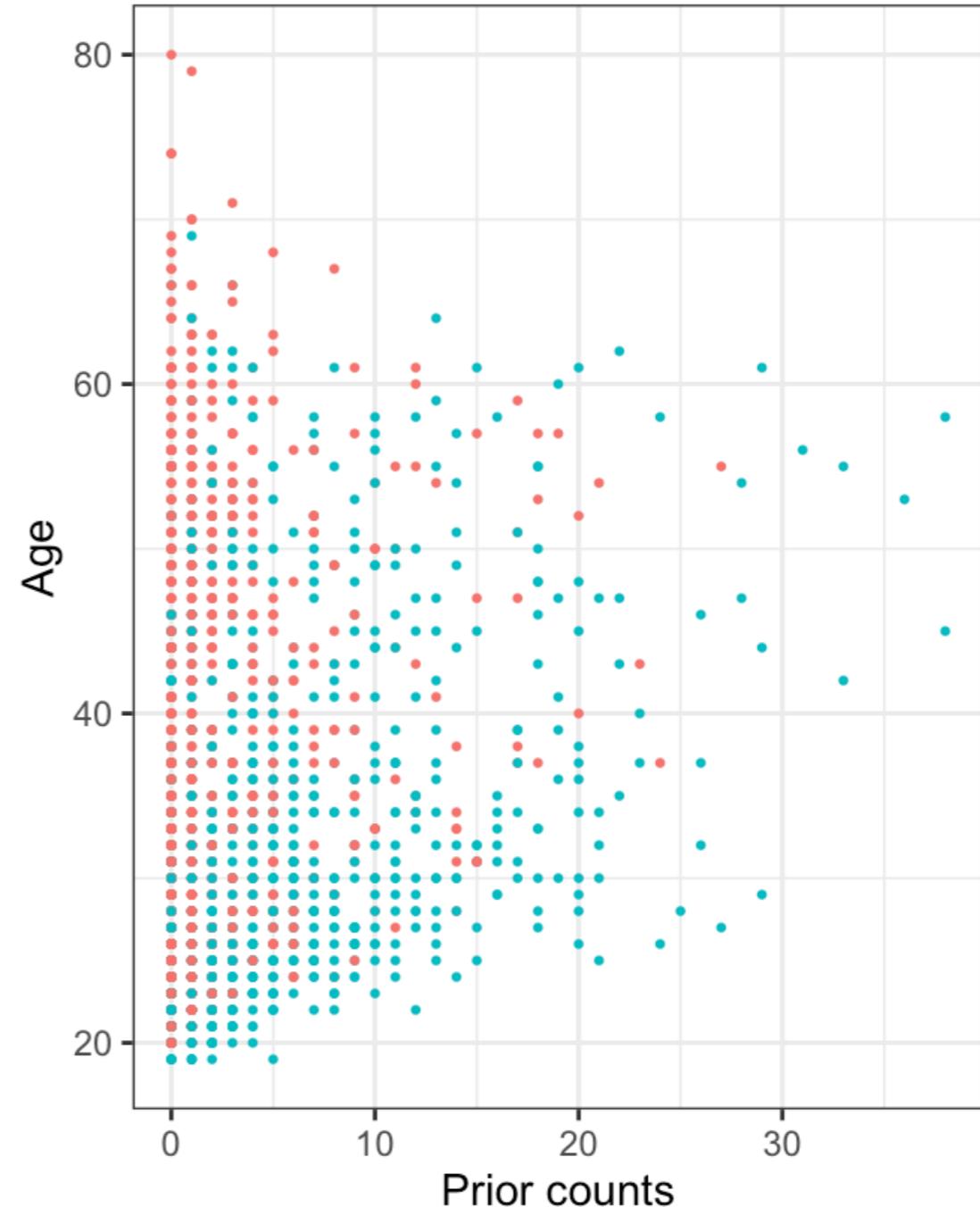
Reoffender (1) or not (0) • 1 • 0

Training Data v. Test Data

Training data



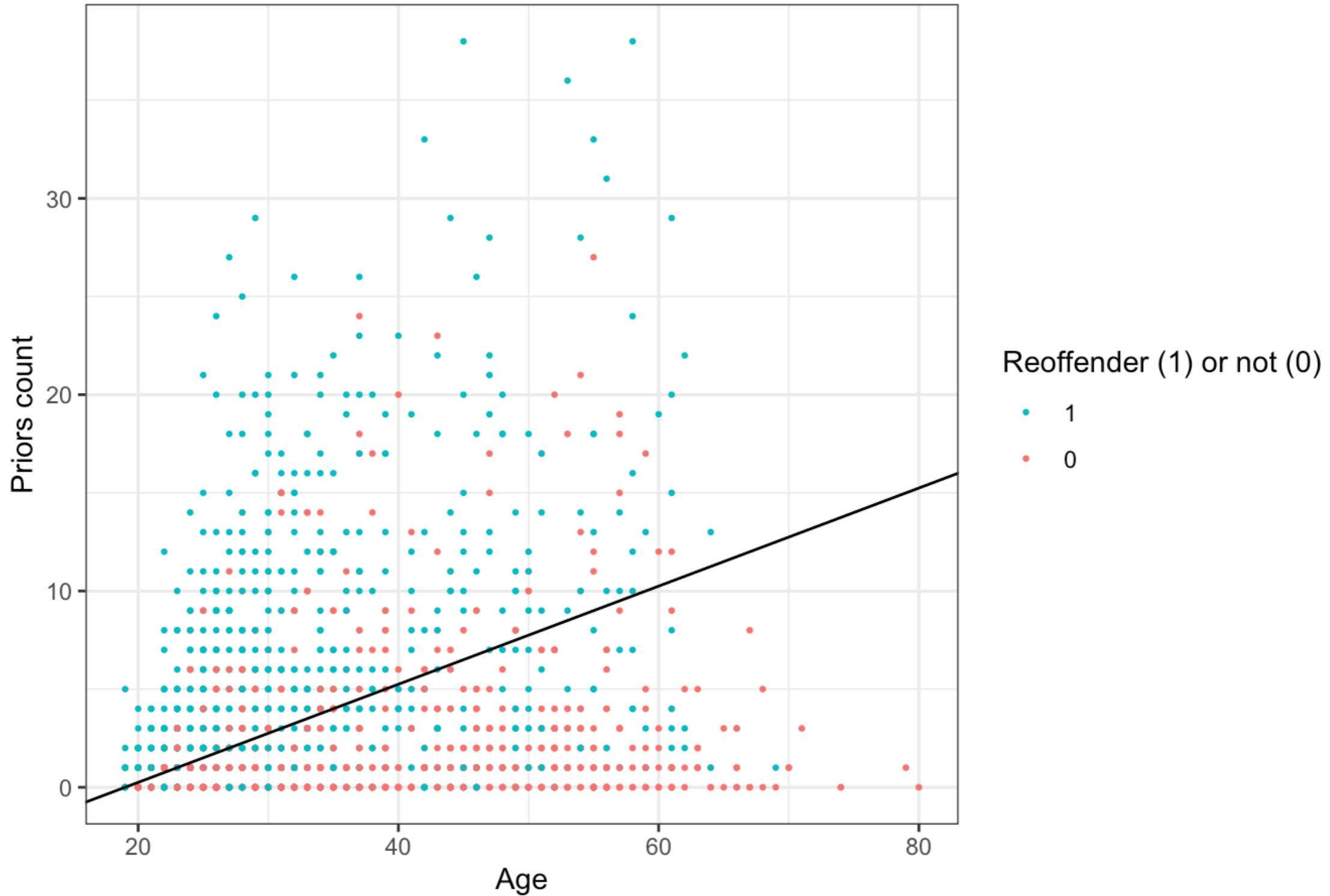
Test data



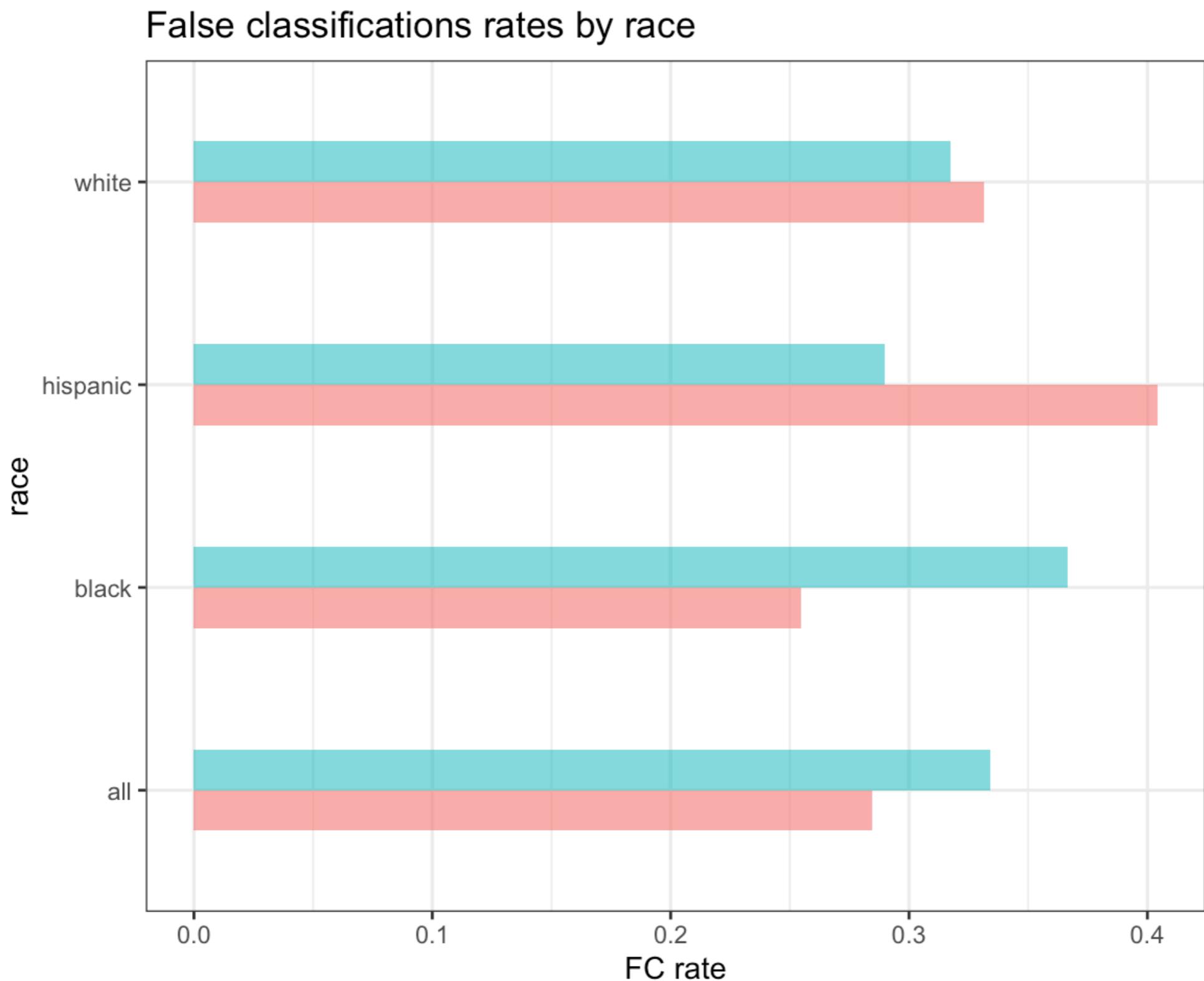
Reoffender (1) or not (0) • 1 • 0

Validating Model Against Test Data

Linear 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



Based on historical crime data
(victims' information)
3 data points: time, place, type of offence

"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

Model 1: Conventional approach

XLAW – Naples Police

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

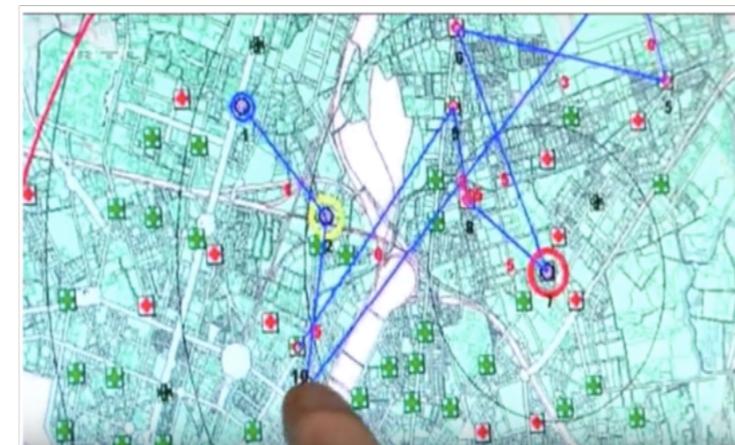
https://www.xlaw.it/presentazione/index_eng.asp



Keycrime – Milan

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

<https://www.keycrime.com/>



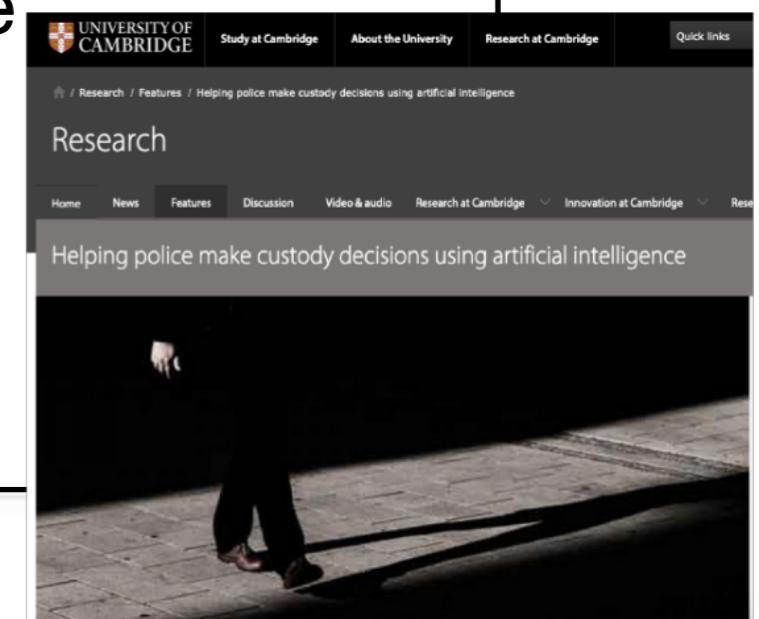
Model 2: Individual assessment

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 close 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

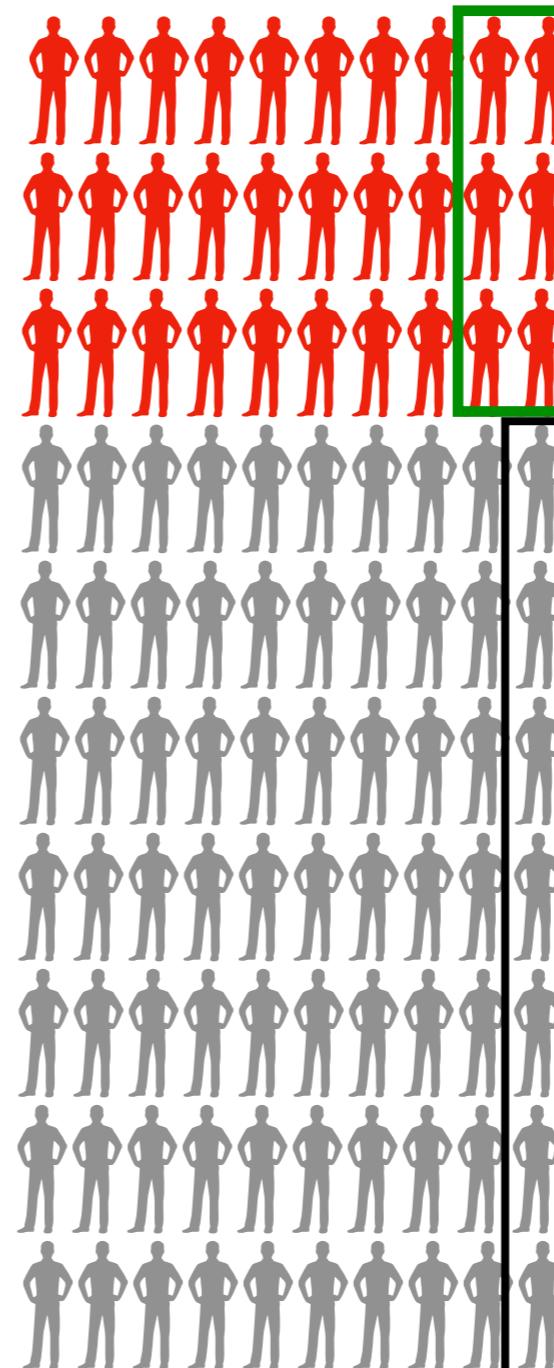


Dichotomous Accuracy/Error Metrics

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

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

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



Y=1

Y=0

FNR = $P(C=0 \mid Y=1)$

FPR = $P(C=1 \mid Y=0)$

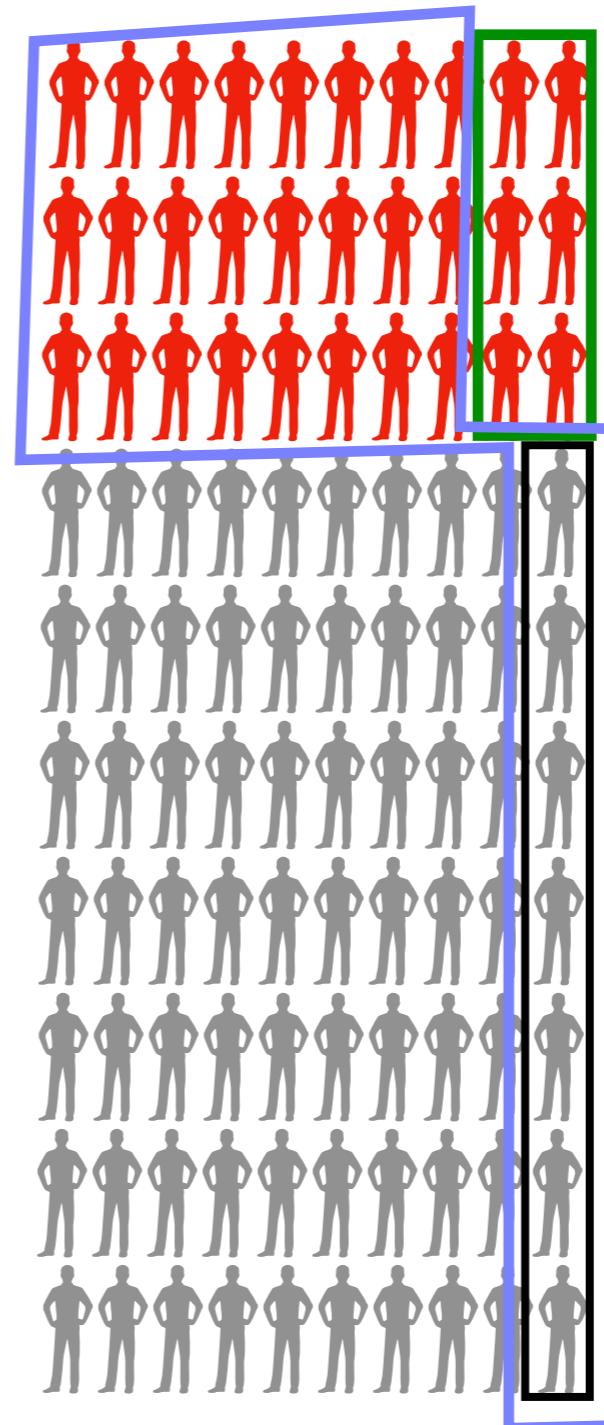
Dichotomous Accuracy/Error Metrics

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



Y=1

Y=0

FNR = $P(C=0 | Y=1)$

FPR = $P(C=1 | Y=0)$

PPV = $P(Y=1 | C=1)$

**(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

PROPUBLICA

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

COMPAS Algorithm

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



PRO PUBLICA

Bernard Parker, left, was rated high risk; Dylan Fugate 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

COMPAS Algorithm

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

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

The image shows a screenshot of a ProPublica article. At the top, there are two side-by-side portraits of men. The man on the left, Bernard Parker, has dark skin, dreadlocks, and is wearing a white t-shirt. The man on the right, Dylan Fuggett, has light skin, short hair, and is wearing a dark t-shirt. Above the portraits is the ProPublica logo and social media sharing icons. Below the portraits is a caption: "Bernard Parker, left, was rated high risk; Dylan Fuggett was rated low risk. (Josh Ritchie for ProPublica)". The main title of the article is "Machine Bias". Below the title is a subtitle: "There's software used across the country to predict future criminals. And it's biased against blacks." At the bottom, the byline reads "by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica" and the date "May 23, 2016".

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 Fugate 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

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



Machine Bias

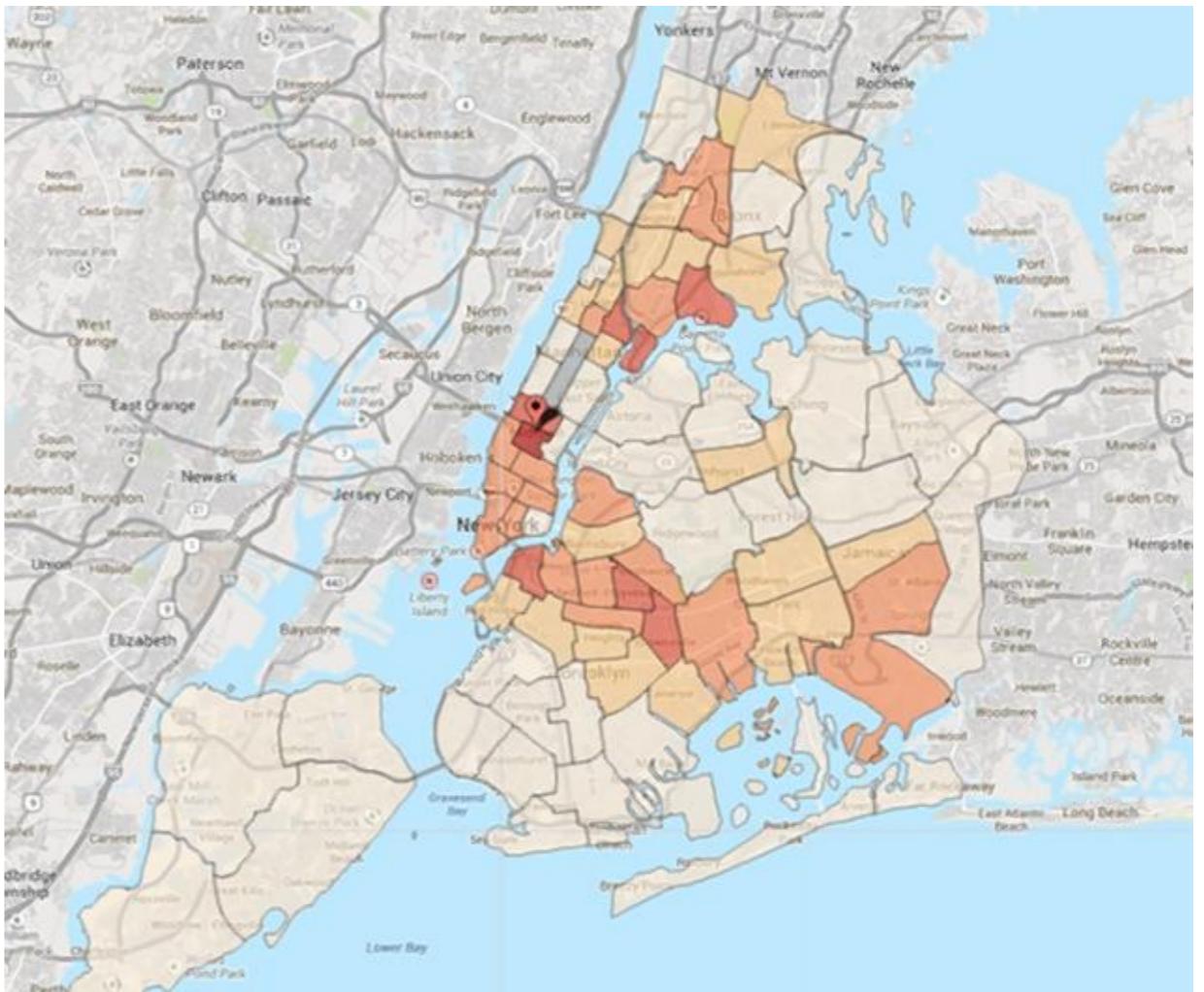
Prediction Fails Differently for Black Defendants

	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

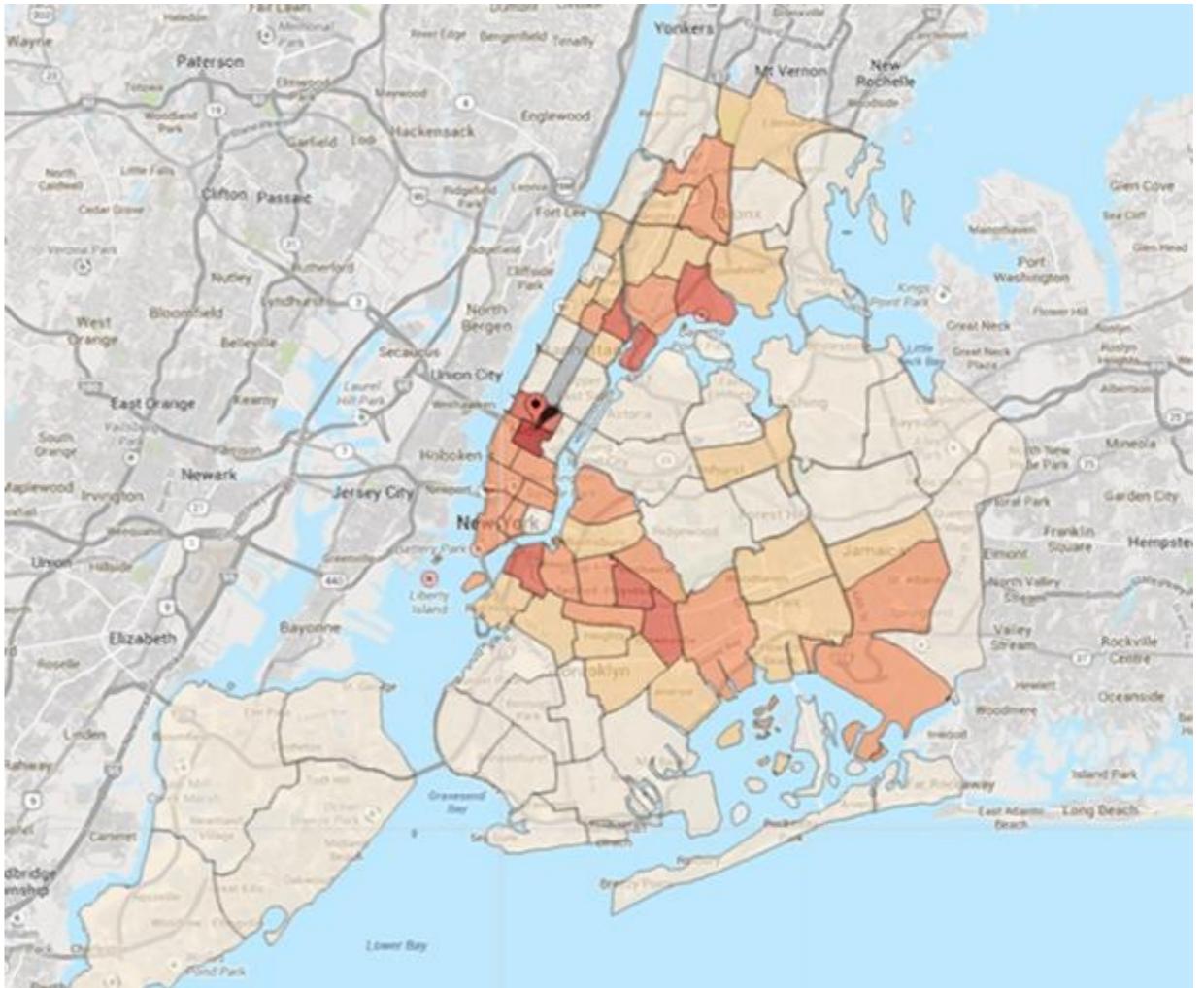
1. **Biased data:** one group is oversampled, data about one group contain more noise, etc.



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

Sources of Algorithmic Bias

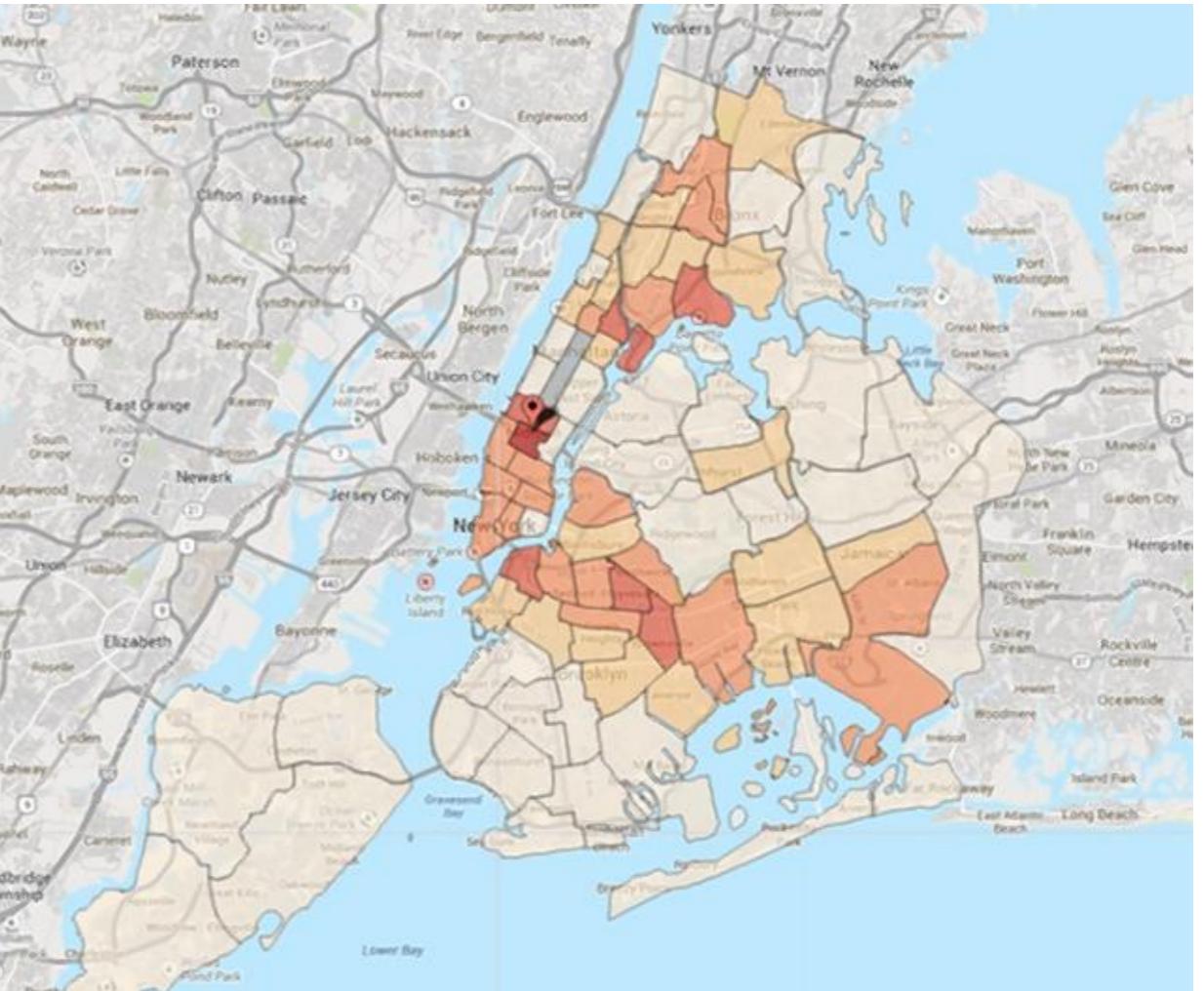
1. **Biased data:** one group is oversampled, data about one group contain more noise, etc.
 - ▶ Use of proxies variables can be pernicious, e.g., when ‘arrest’ is used as a proxy for recidivism or ‘healthcare cost’ as a proxy for ‘care need’



- ▶ 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

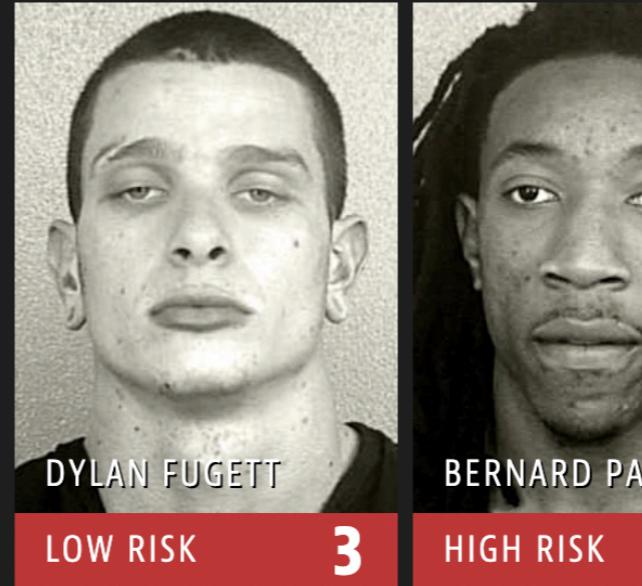
Prediction Fails Differently for Black Defendants

	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 charges.

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 charges.

Two DUI Arrests

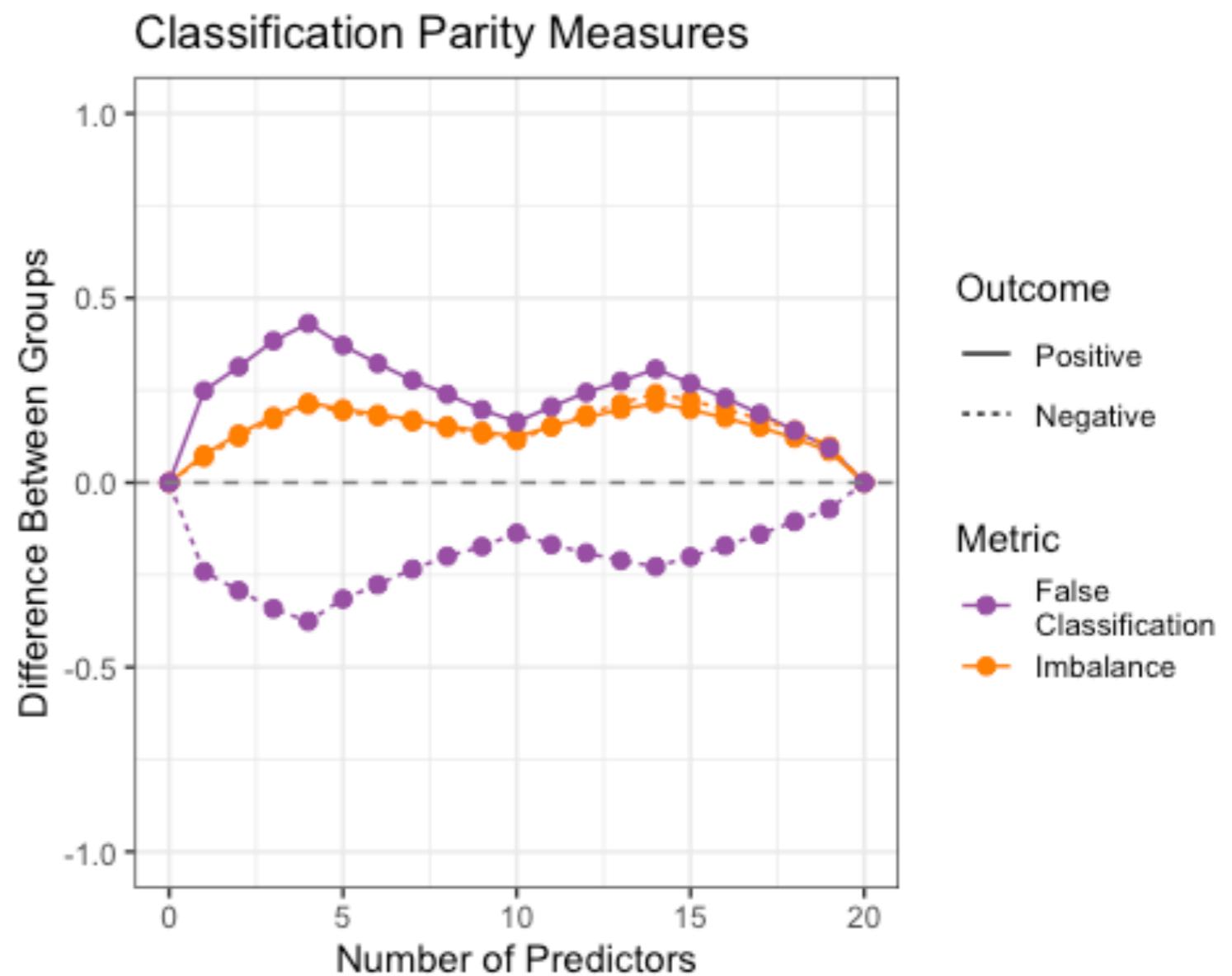


Lugo crashed his Lincoln Navigator into a Toyota Camry while drunk. He was rated as a low risk of reoffending despite the fact that it was at least his fourth DUI.

*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?*

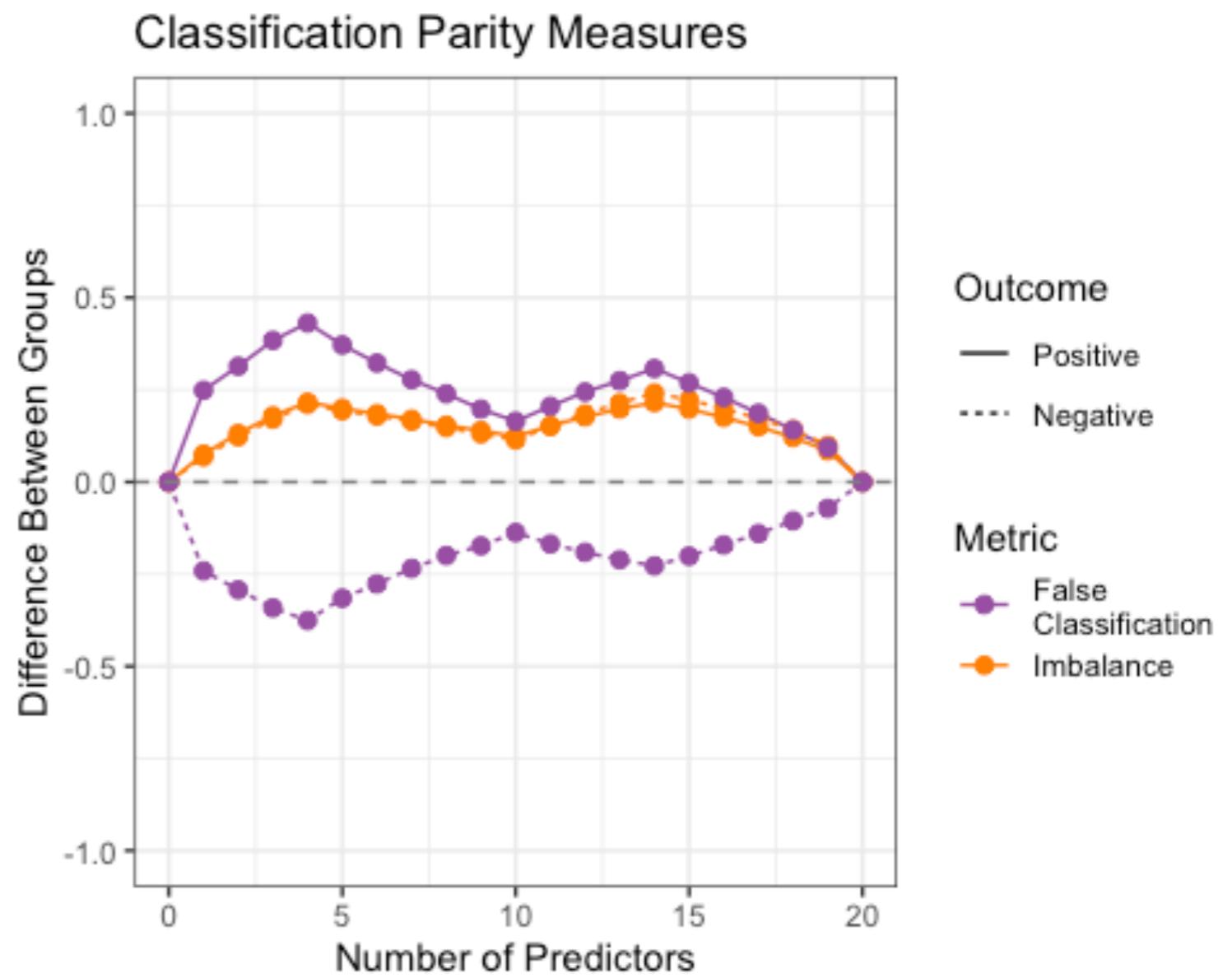
Simulating an Unbiased Dataset in an Unbiased Society



*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

"...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.



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

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 may 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'd)

CAUSES 2)

Legal Value Attached to the “predictions”

...Do you see the paradox?

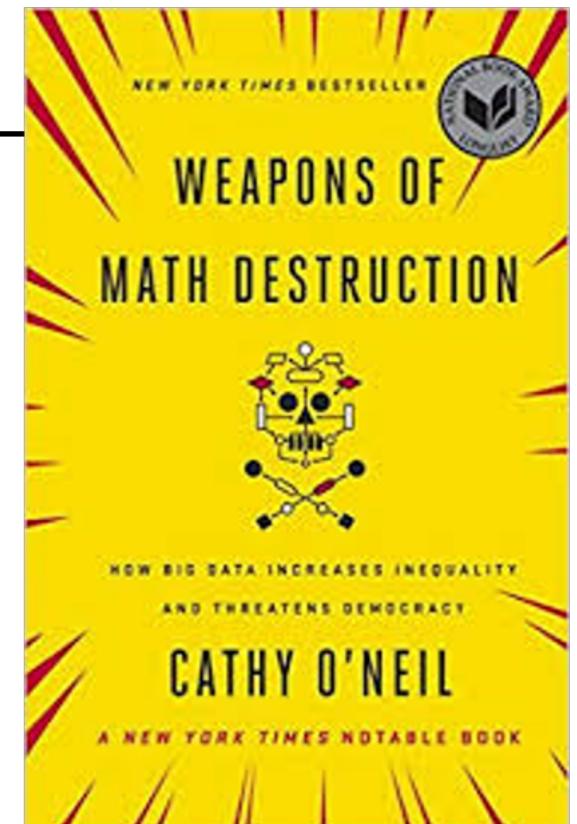
An algorithm processes a slew of statistics and comes up with a probability that a certain person *might* be a bad hire, a risky borrower, a terrorist, or a miserable teacher.

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 solely on automated processing
(including profiling)

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

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

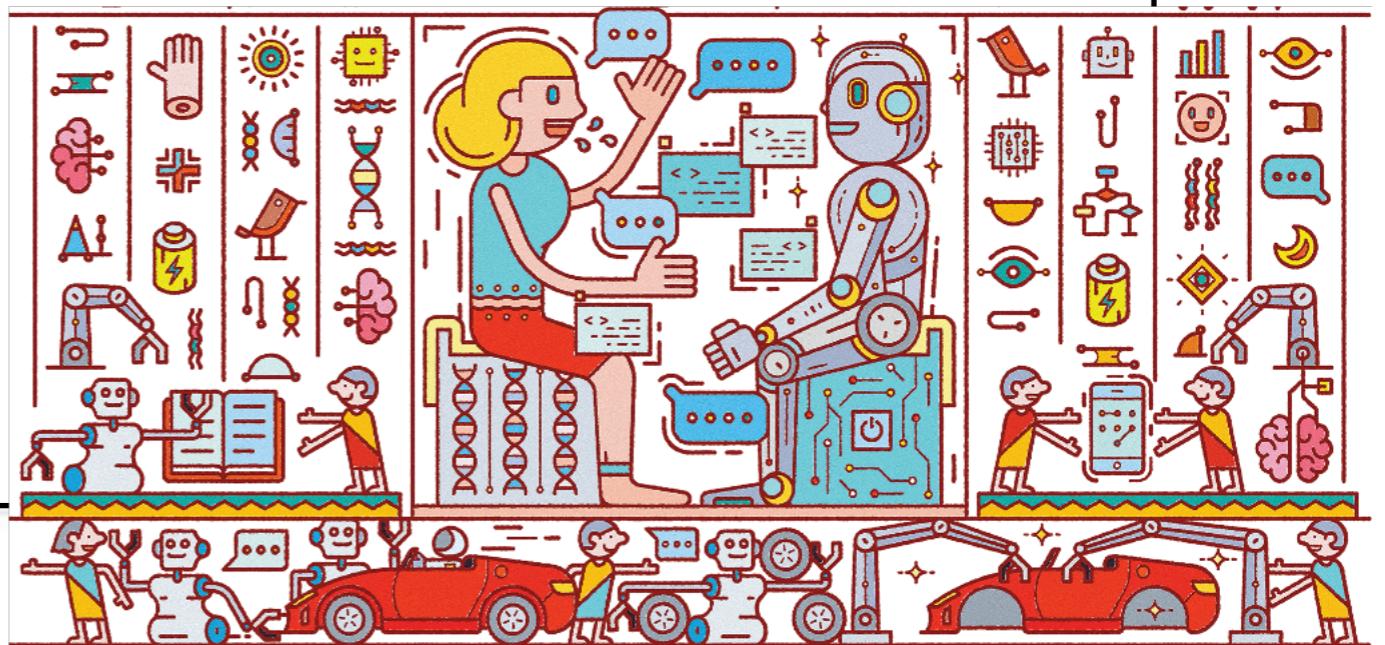
WHICH REMEDY?/2

“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)



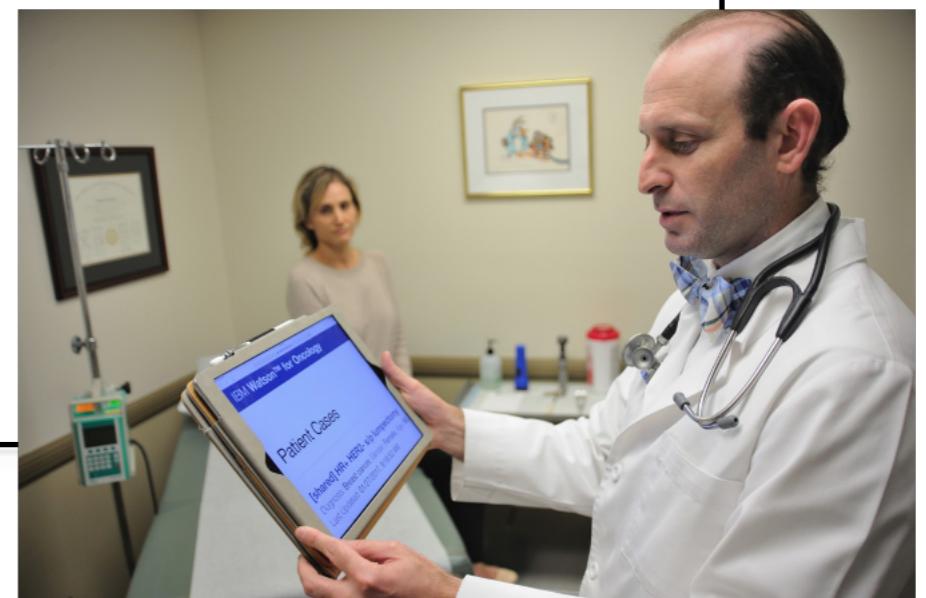
WHICH REMEDY?/3

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

PART V

Impossibility Theorems

Dichotomous Accuracy/Error Metrics

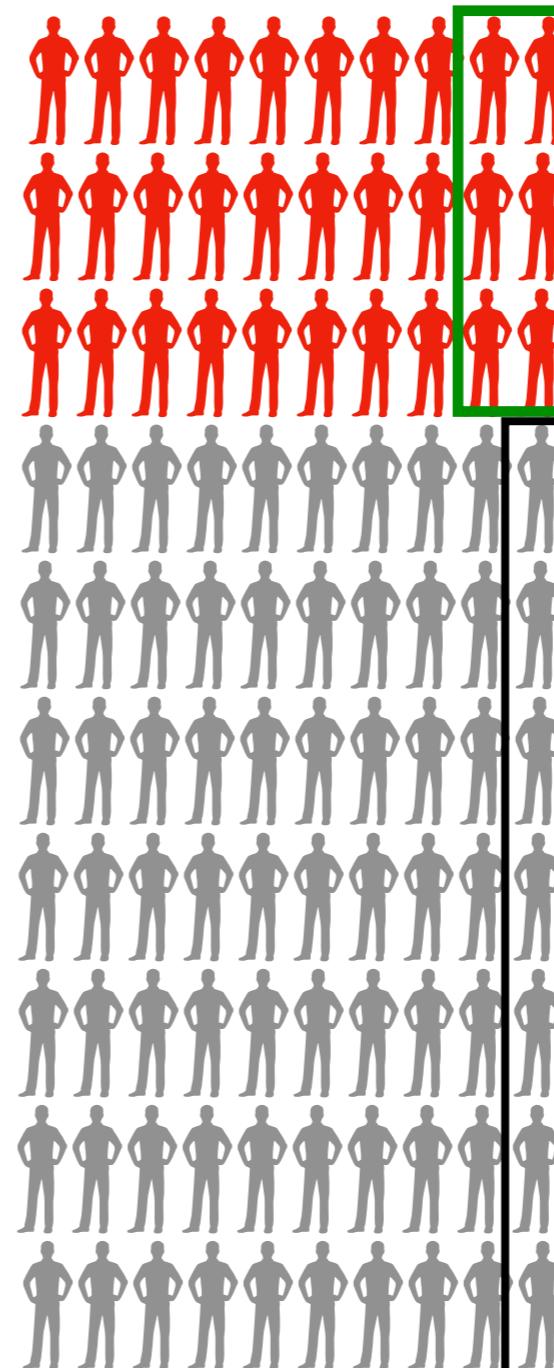


Dichotomous Accuracy/Error Metrics

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

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

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



Y=1

Y=0

FNR = $P(C=0 \mid Y=1)$

FPR = $P(C=1 \mid Y=0)$

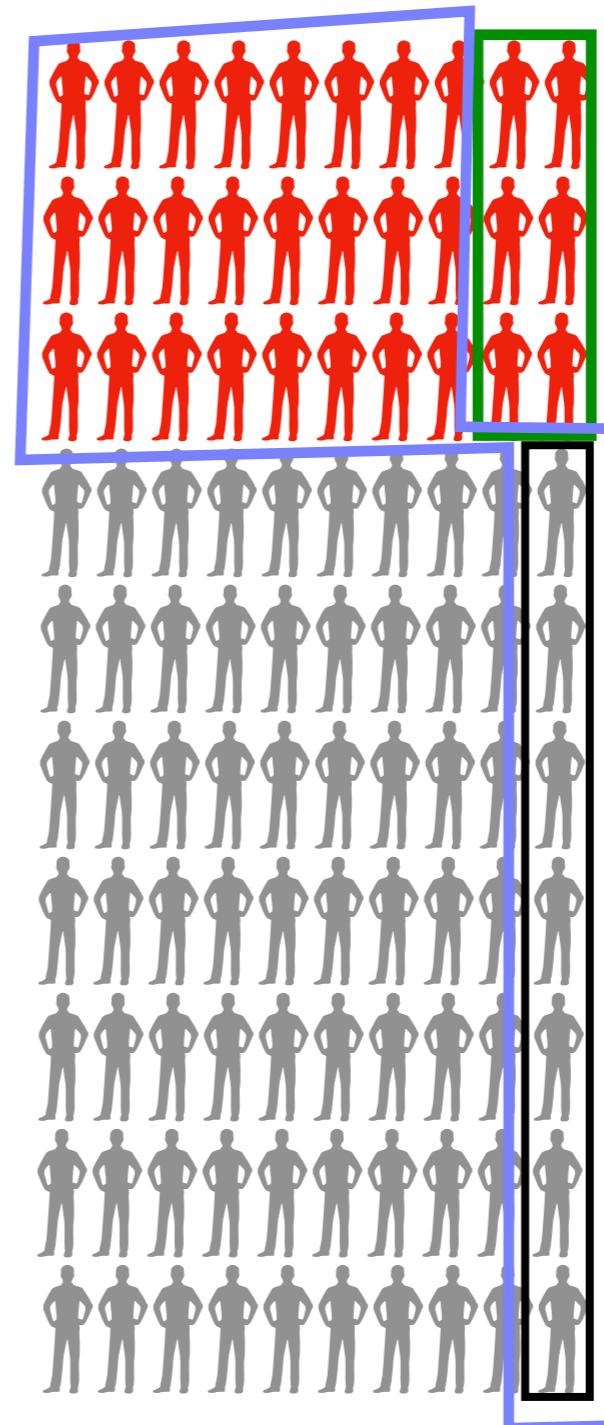
Dichotomous Accuracy/Error Metrics

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



Y=1

Y=0

FNR = $P(C=0 | Y=1)$

FPR = $P(C=1 | Y=0)$

PPV = $P(Y=1 | C=1)$

Back to COMPAS



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

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



The image shows two men from the chest up. On the left, Bernard Parker, a Black man with long dreadlocks, is looking directly at the camera. On the right, Dylan Fugett, a white man with short hair, is also looking directly at the camera. They are both wearing dark-colored shirts. The background is a plain, light color.

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



Bernard Parker, left, was rated high risk; Dylan Fugate 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



Bernard Parker, left, was rated high risk; Dylan Fuggett 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

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

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

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

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

Dichotomous (Group) Fairness Metrics

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:
 $P(Y=1 | C=1 \& G=1) = P(Y=1 | C=1 \& G=0)$

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**)
 $P(C=0 | Y=1)$

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:
 $P(Y=1 | C=1 \& G=1) = P(Y=1 | C=1 \& G=0)$

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

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

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

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

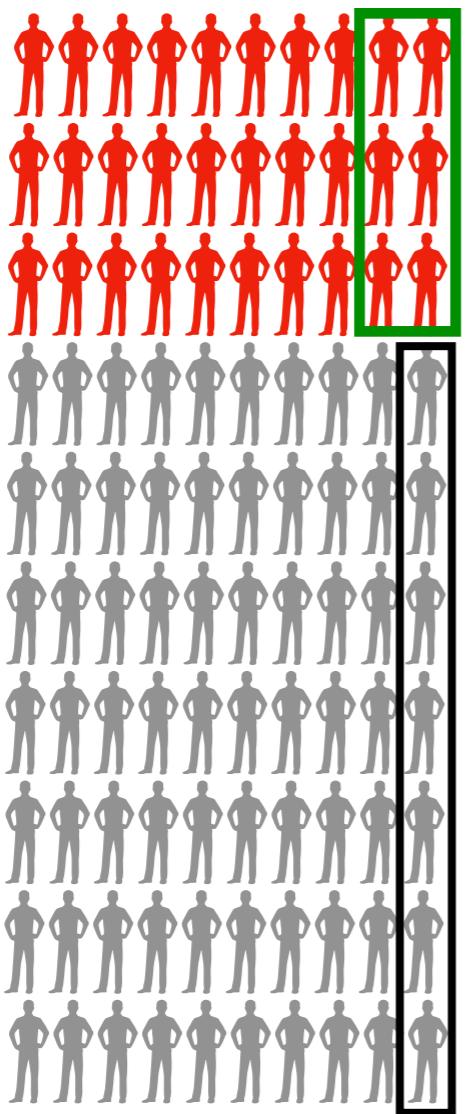
$$\frac{P(Y = 1 | C = 1)}{P(Y = 0 | C = 1)} = \frac{P(C = 1 | Y = 1)}{P(C = 1 | Y = 0)} \times \frac{P(Y = 1)}{P(Y = 0)}$$

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

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

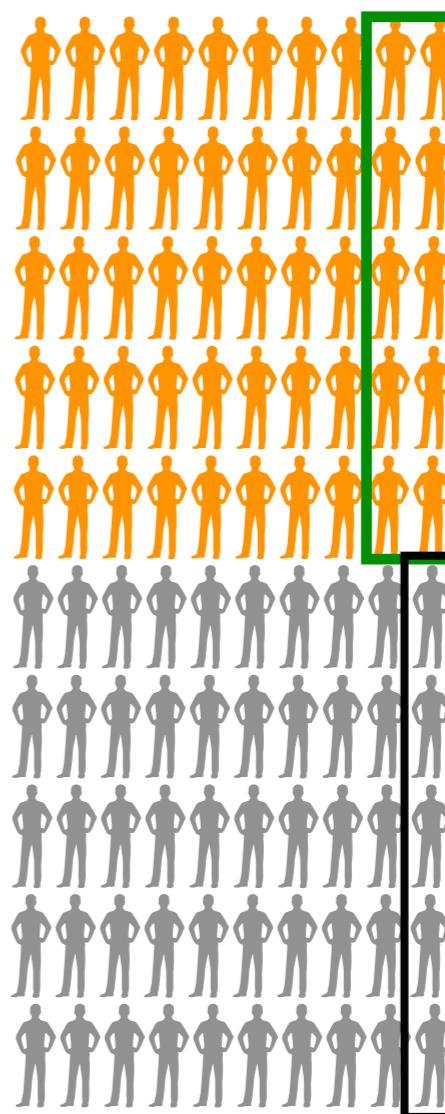
If **FN** and **FP** rates are the same across groups, then **PPV** must be different unless the prevalence rates are the same

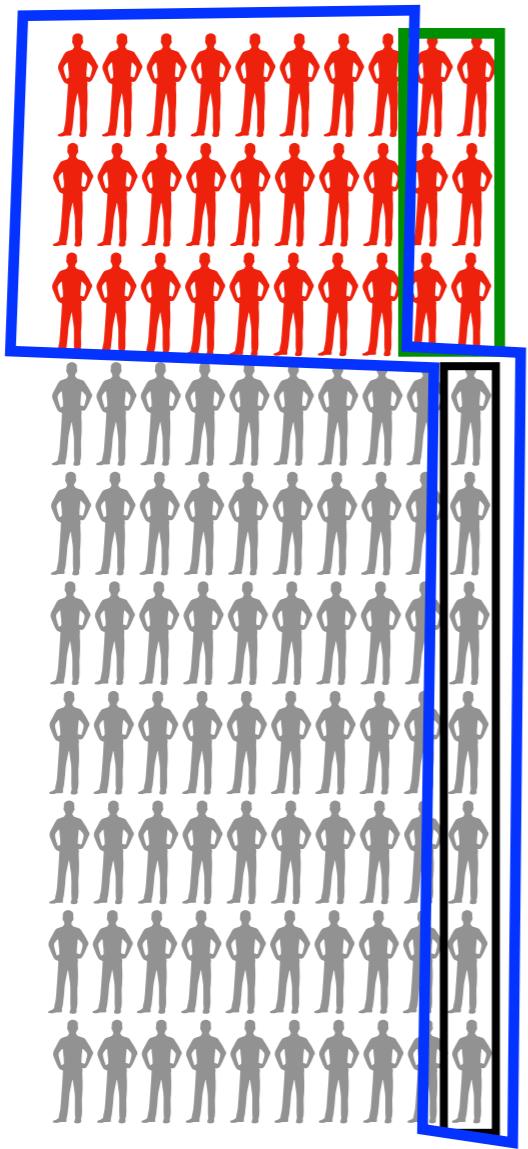
**Suppose FPR and FNR Are
the Same Across Two Groups**



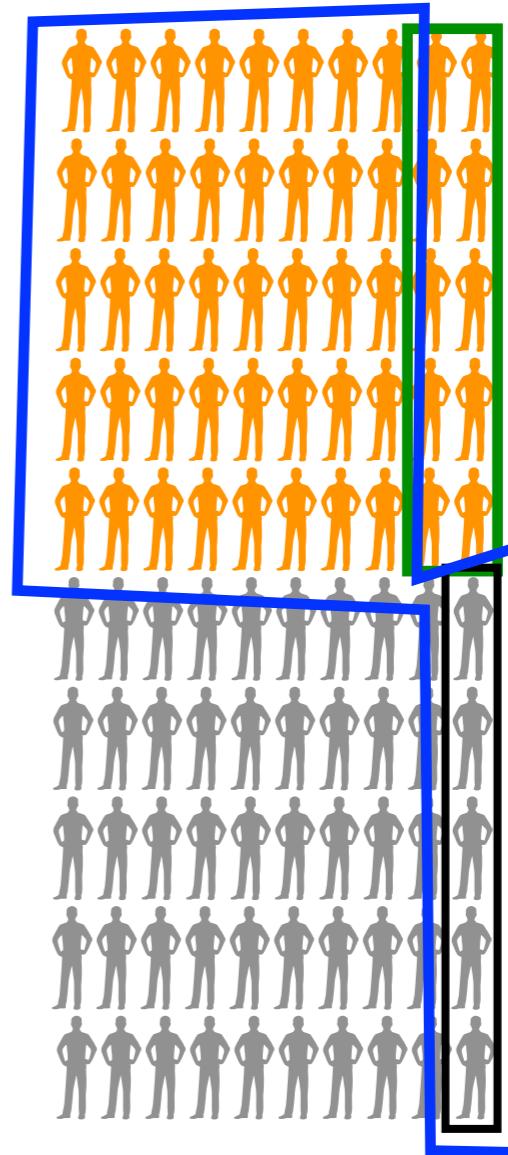
FNR = 20%

FPR = 10%





FNR = 20%



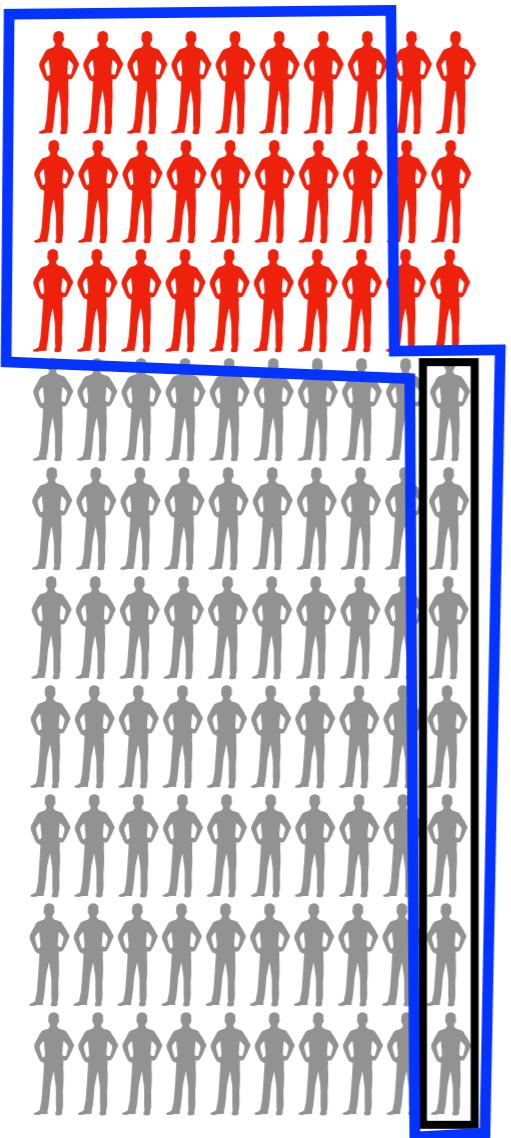
FPR = 10%

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

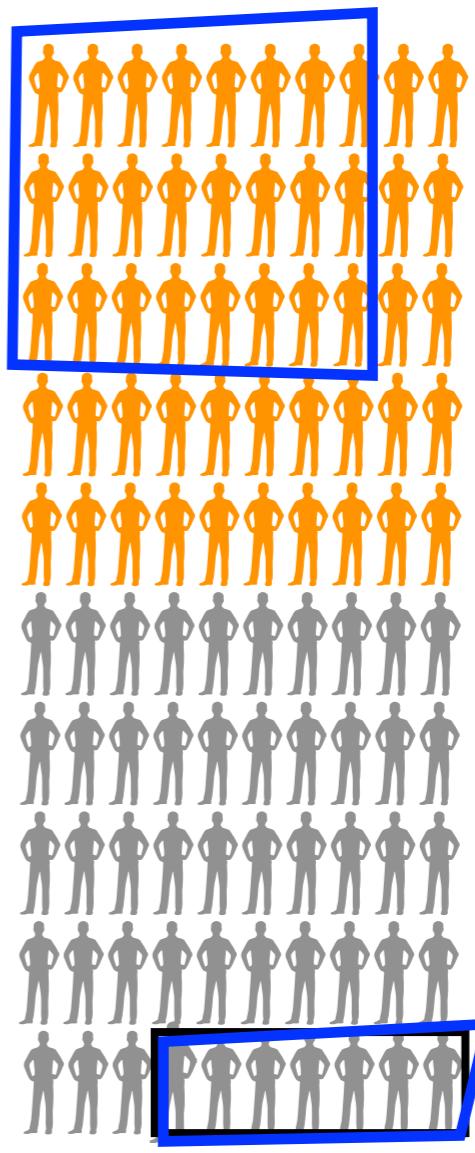
$$PPV_1 = \frac{24 \%}{24 \% + 7 \%} \approx 77 \%$$

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

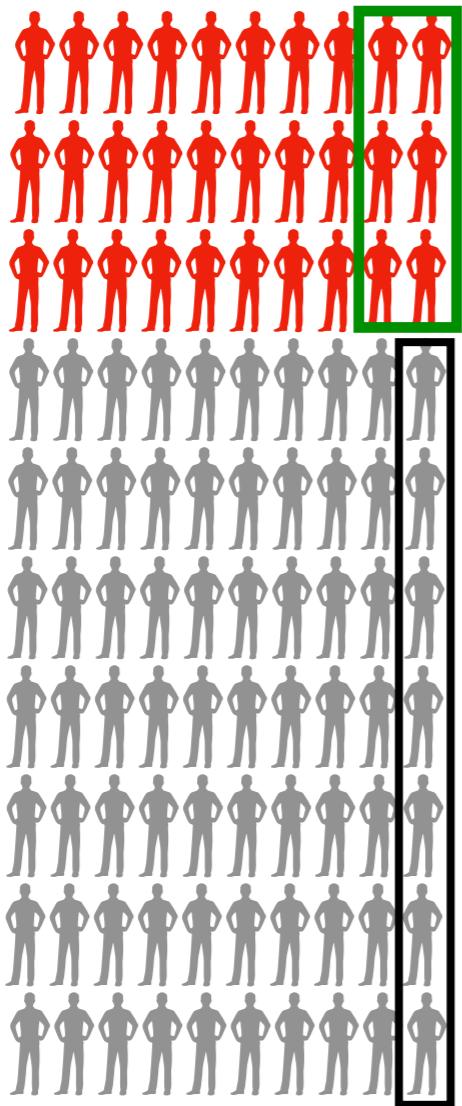
**What If PPV Is the
Same Across Groups?**



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



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

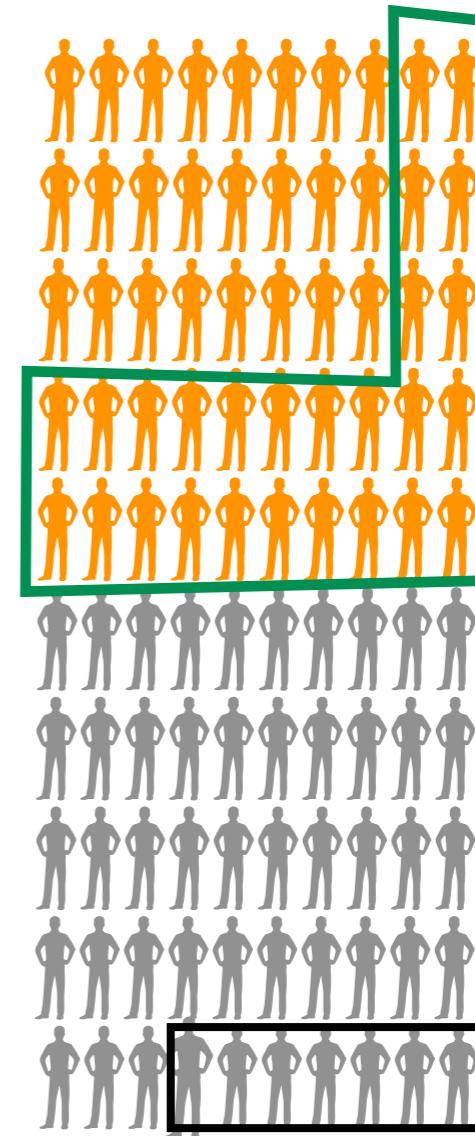


$$FPR_1 = 20\%$$

$$TPR_2 = 26/50=52\%$$

$$FPR_1 = 10\%$$

$$FPR_2 = 7/50 =14\%$$



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

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

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**