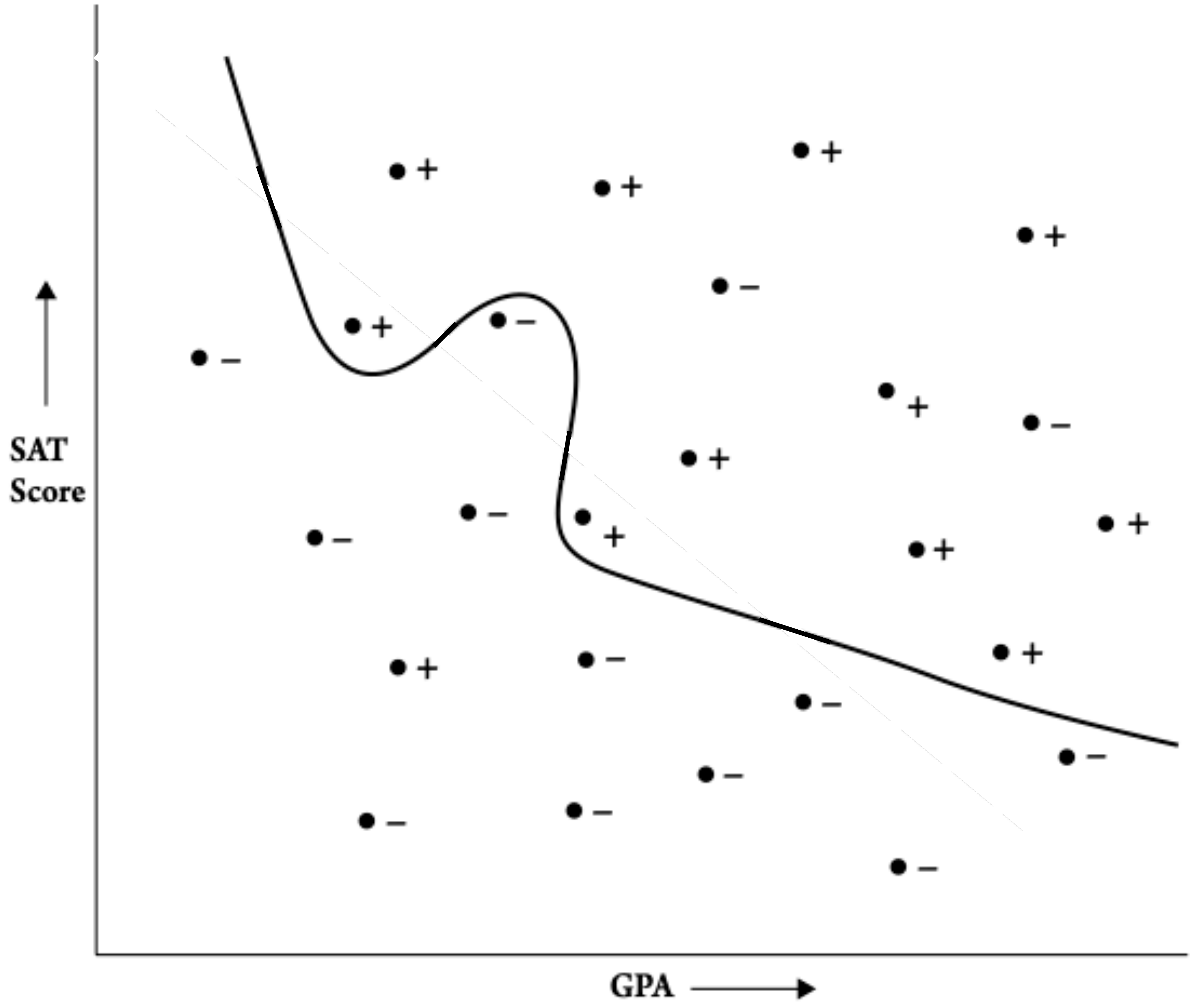
Is Algorithmic Fairness Possible?

Marcello Di Bello
Arizona State University

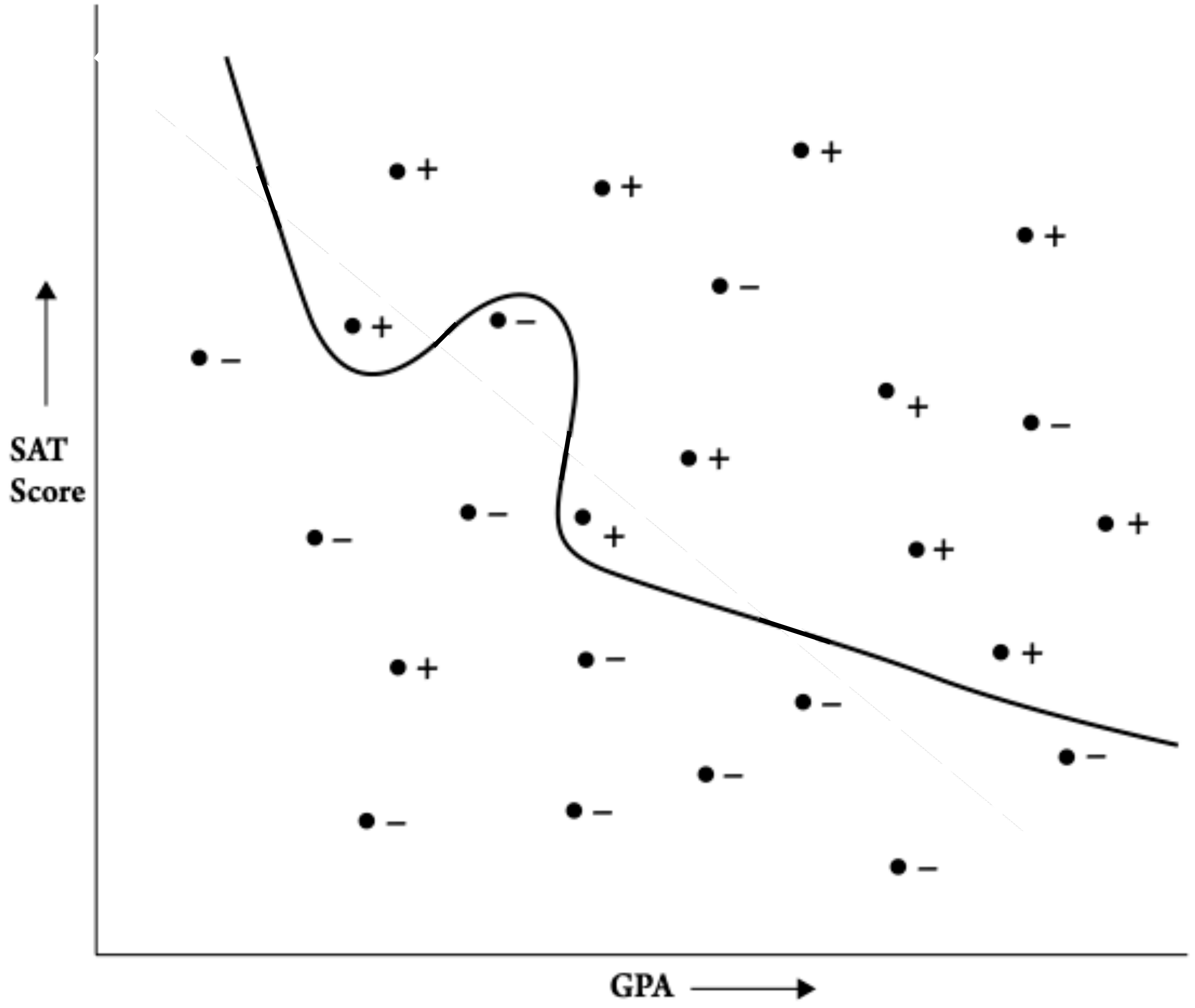
November 13, 2021 - PSA ~~2020~~ 2021 - Baltimore

0. Background

Predictive Models (*binary case*)

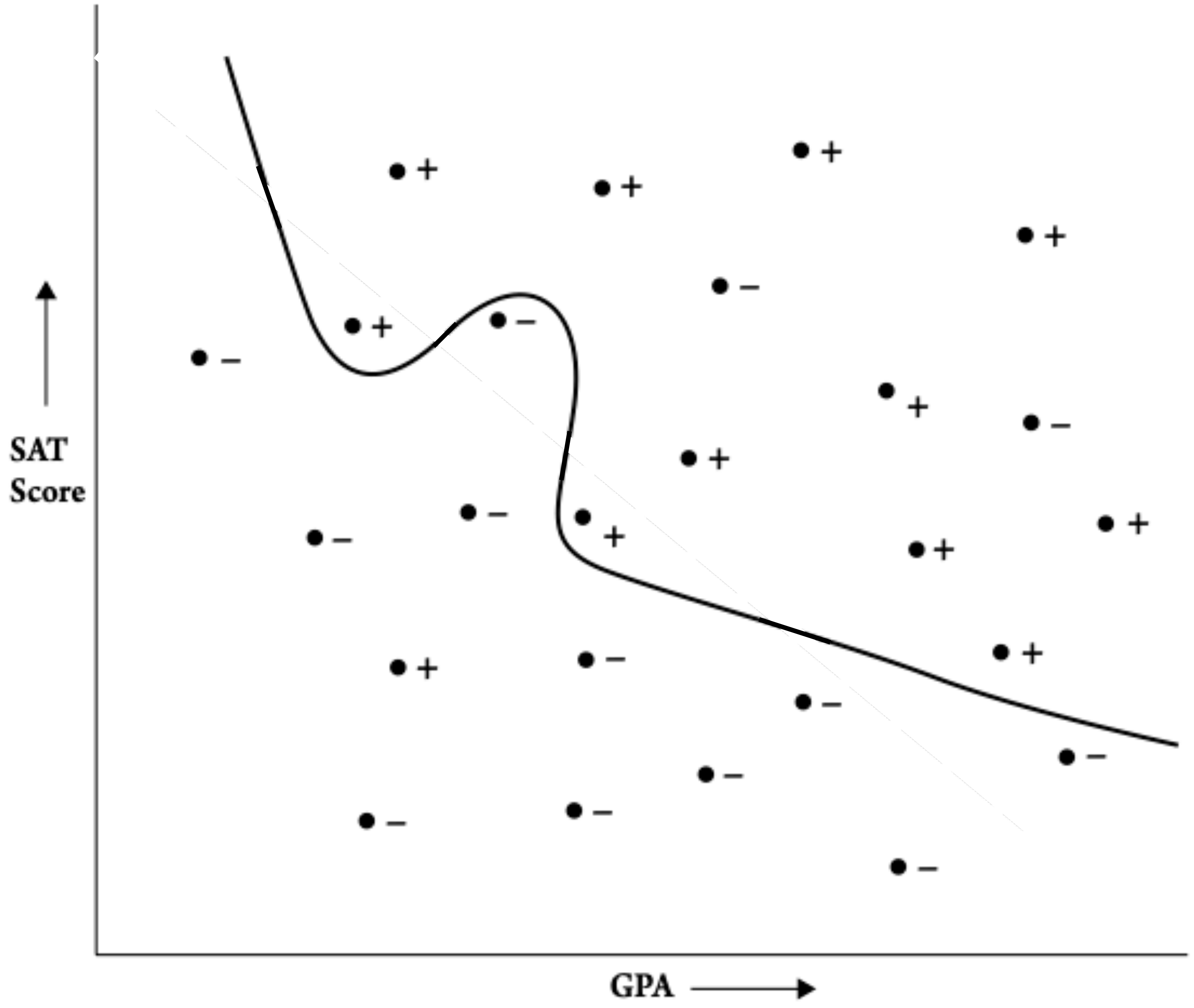


Predictive Models (*binary case*)



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

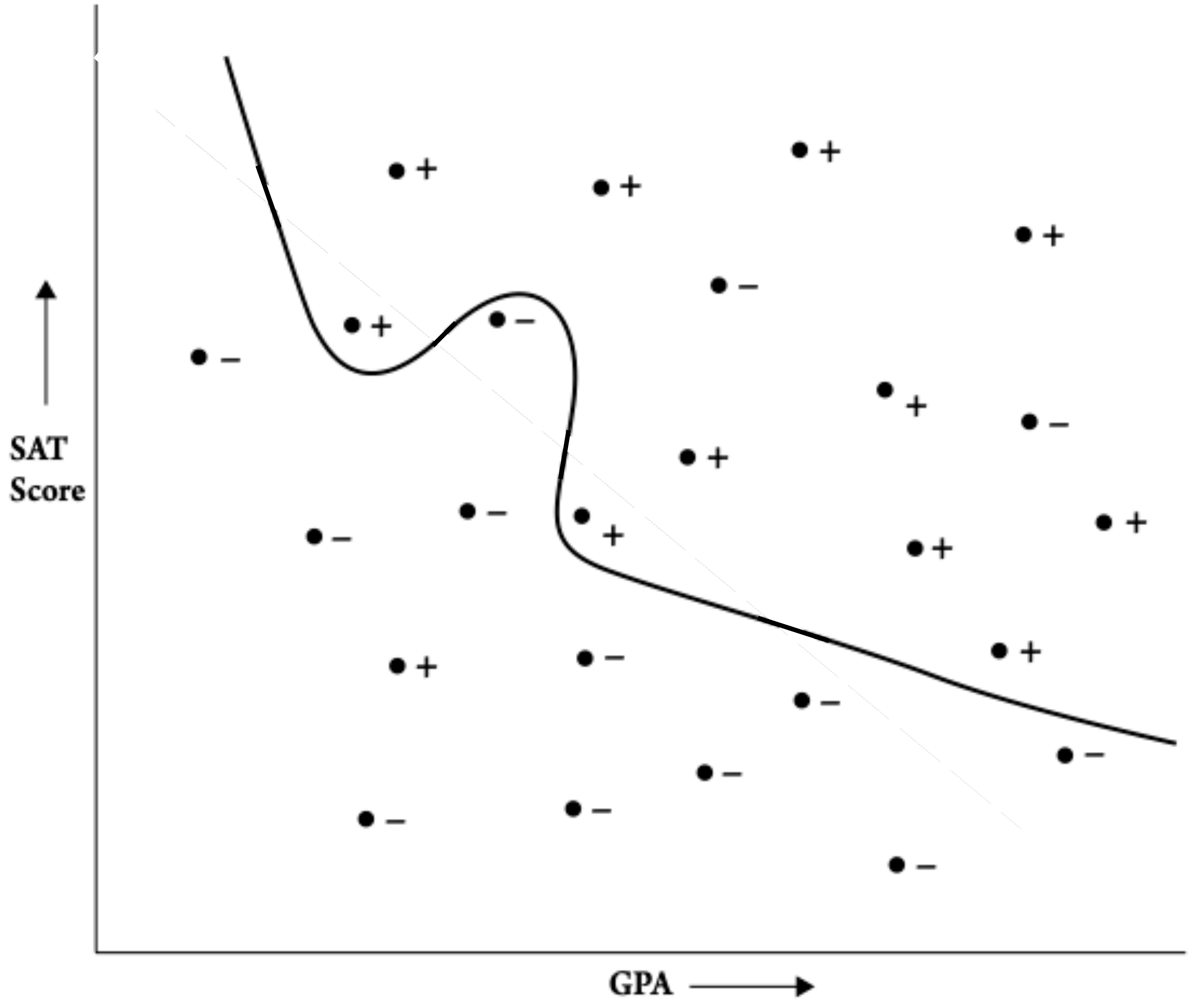
Predictive Models (*binary case*)



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 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**

Predictive Models Can Make Mistakes

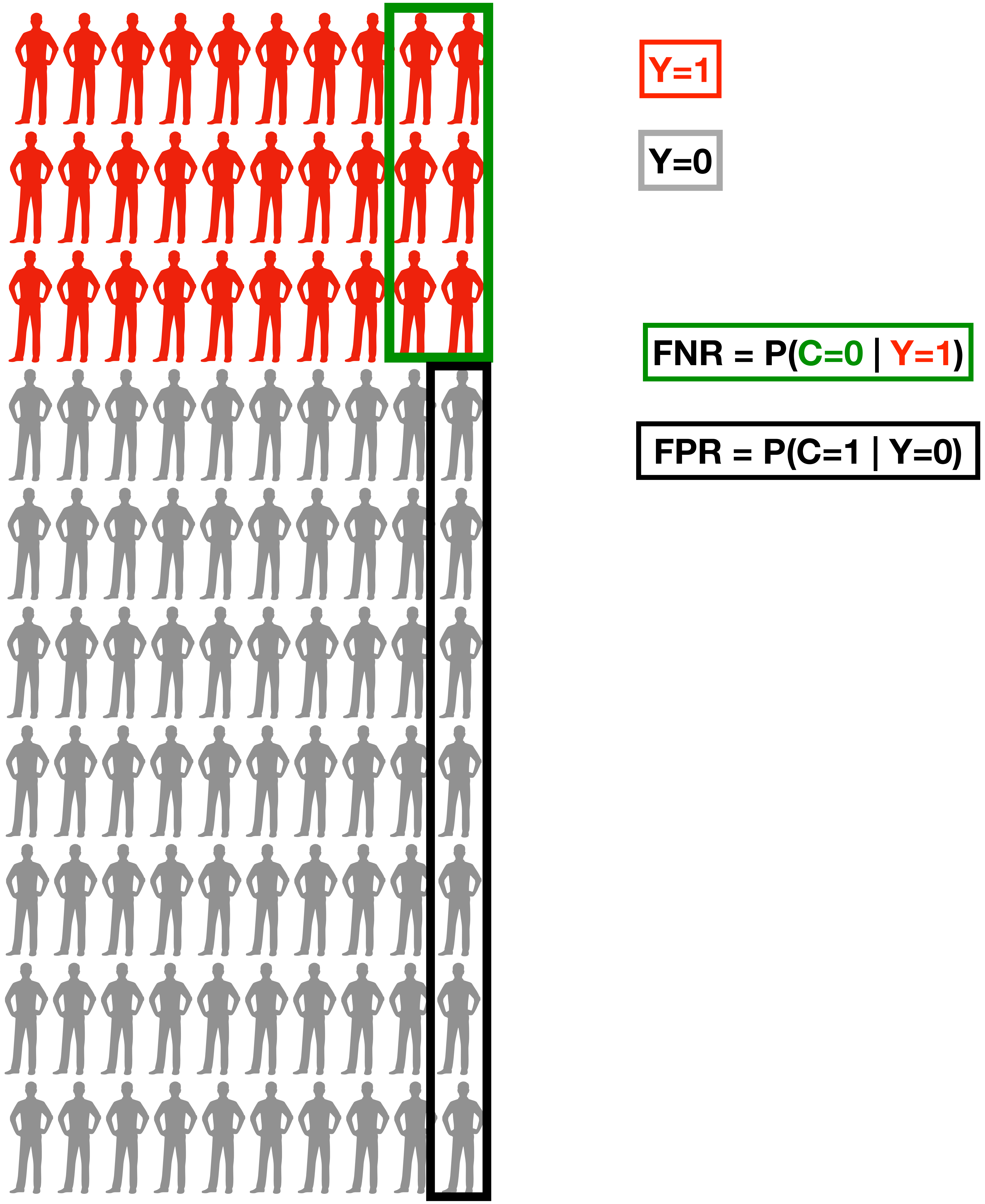
Dichotomous Accuracy/Error Metrics

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



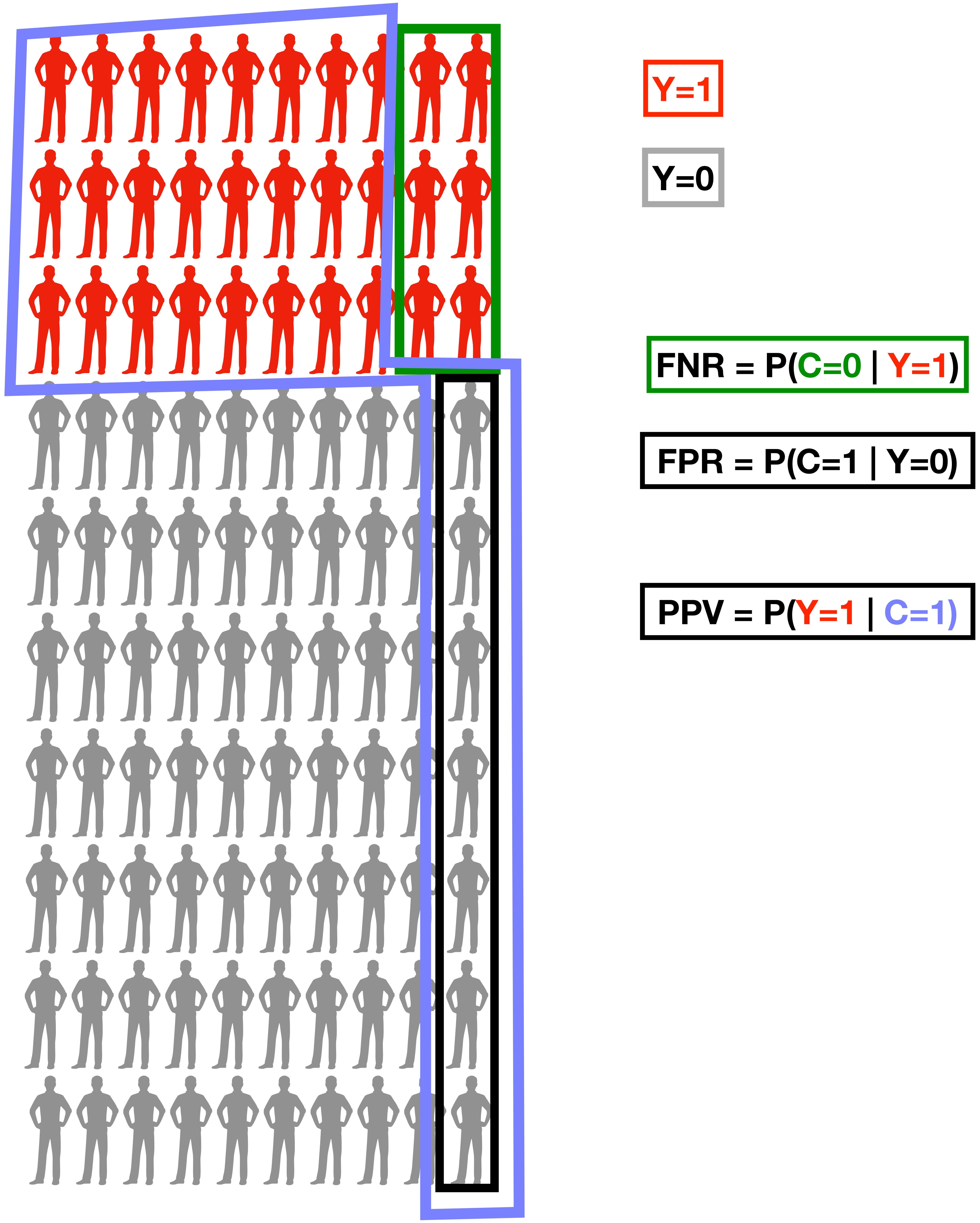
Dichotomous Accuracy/Error Metrics

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

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

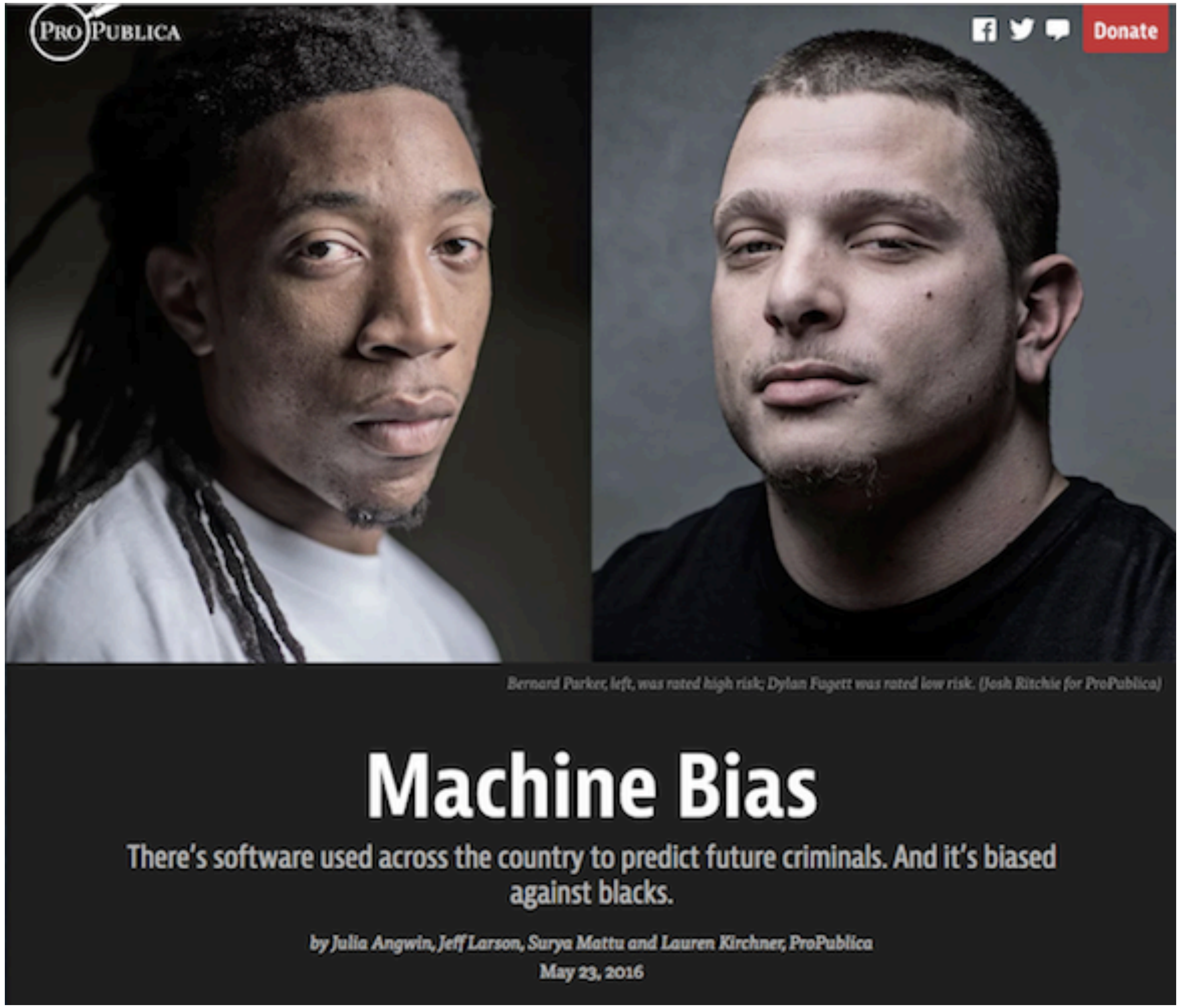
Positive predictive value (**PPV**)
P(Y=1 | C=1)

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



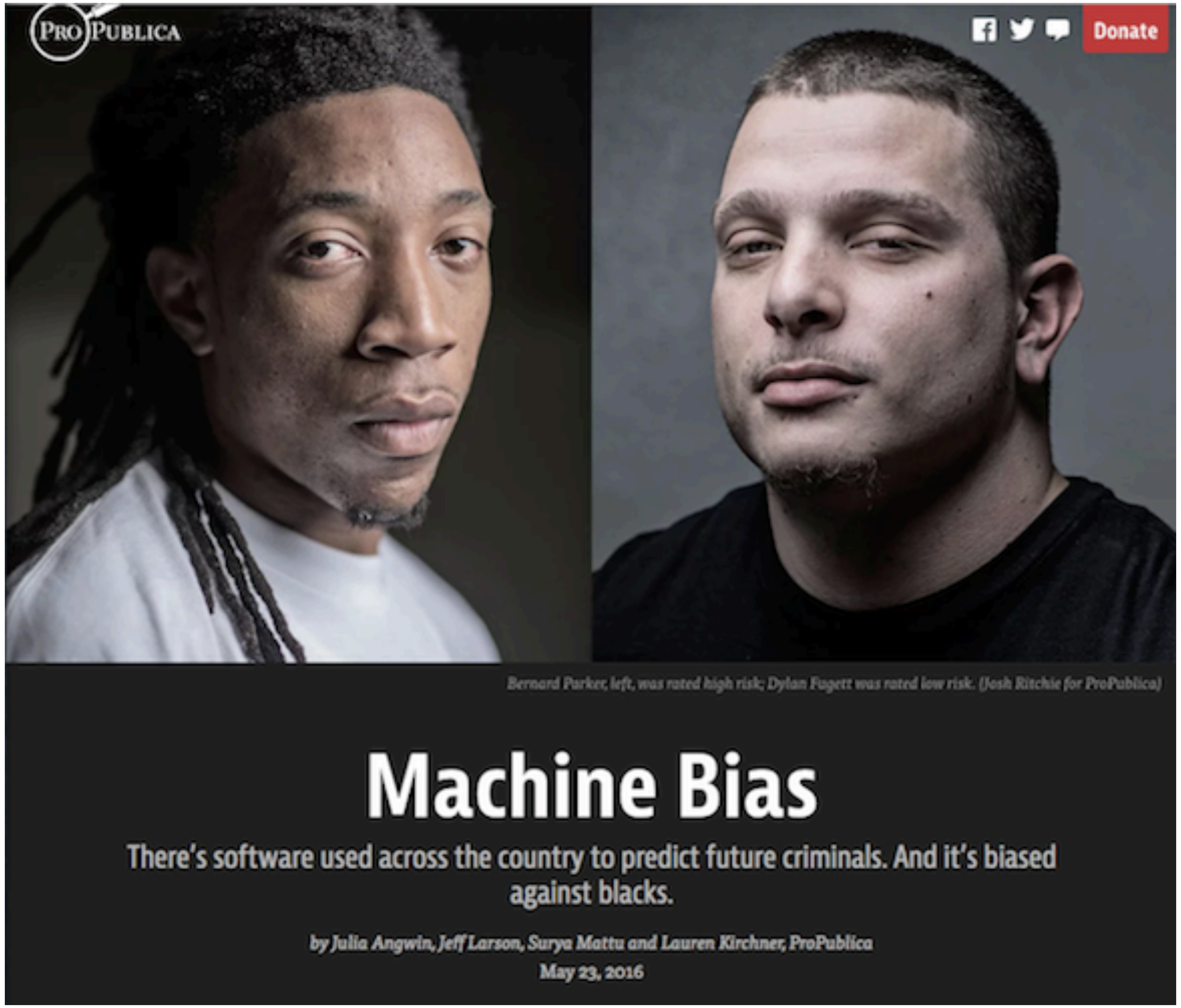
Predictive Models Can Be Unfair

COMPAS Algorithm



COMPAS Algorithm

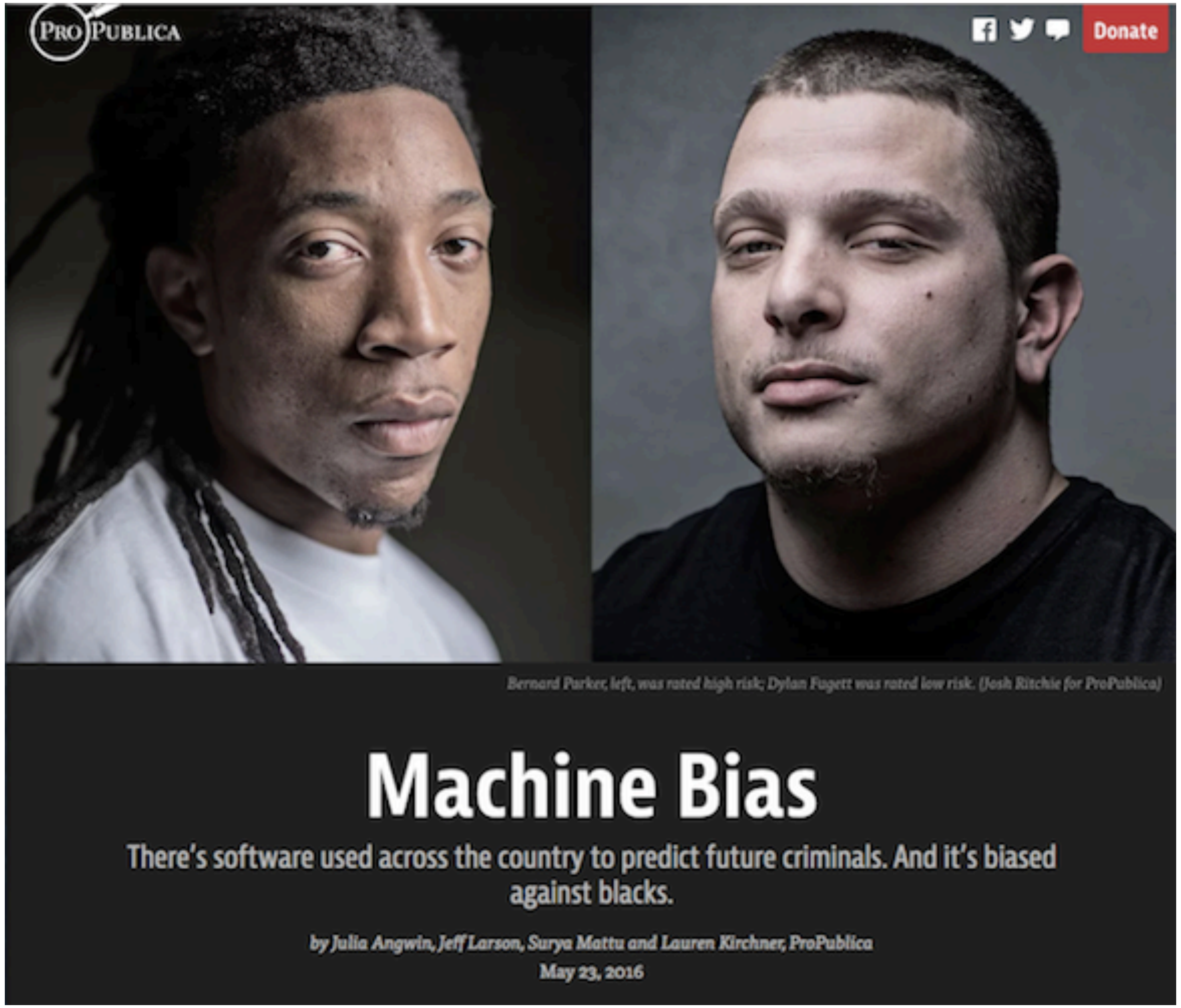
Even if 'race' was not among the predictive features used:



COMPAS Algorithm

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

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

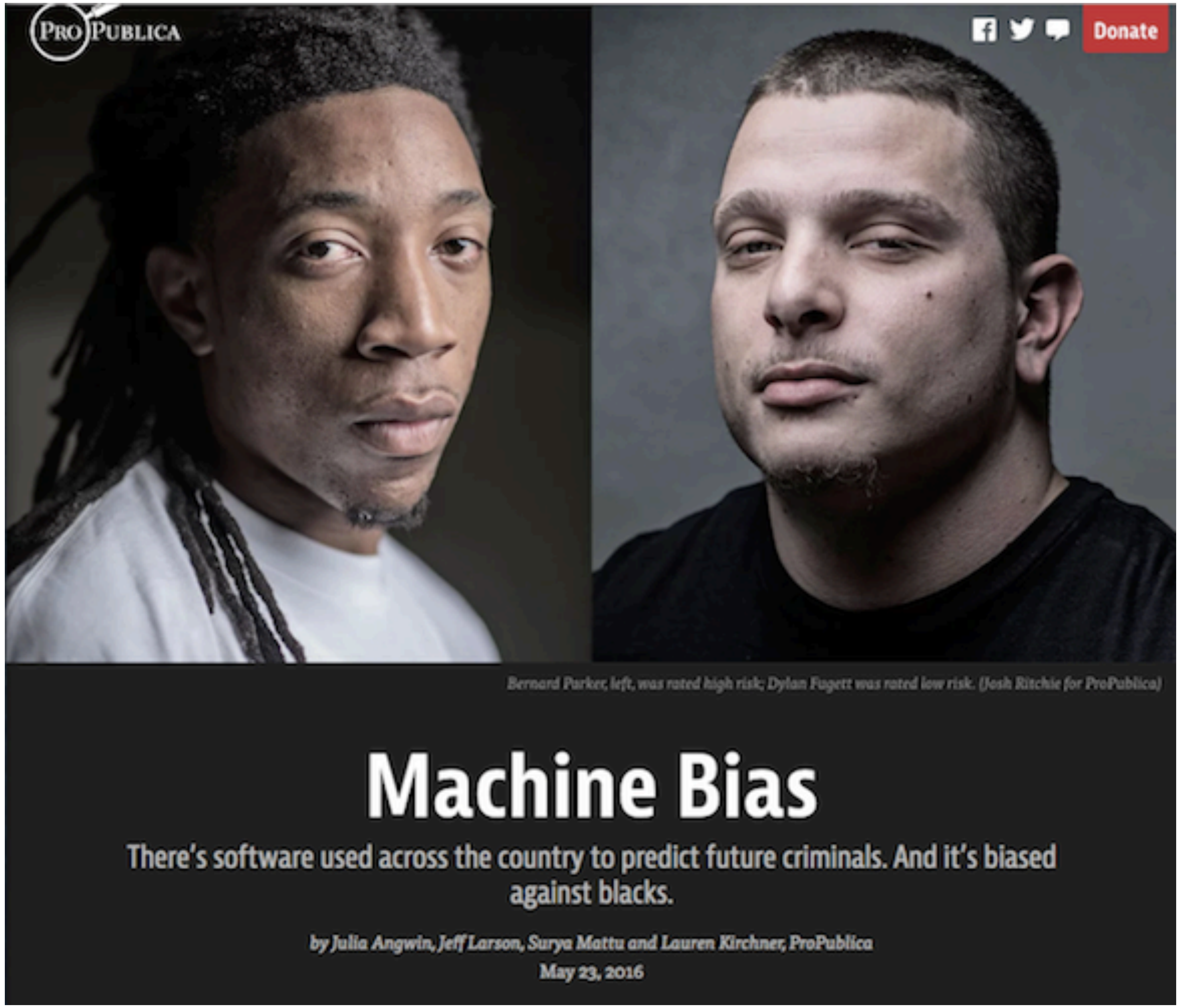


COMPAS Algorithm

Even if 'race' was 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



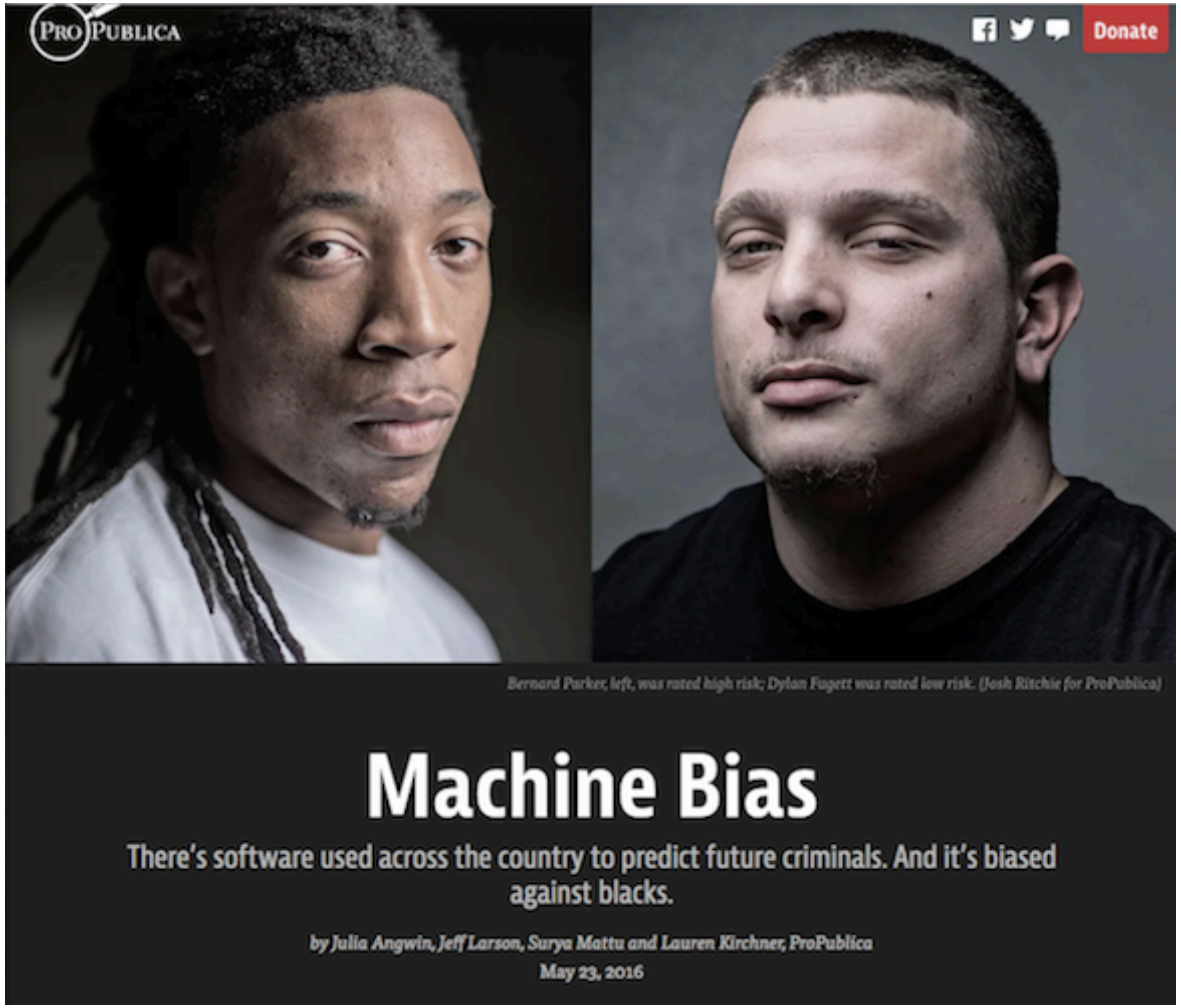
COMPAS Algorithm

Even if ‘race’ was 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

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

Dichotomous Accuracy Metrics

Dichotomous (Group) Fairness Metrics

Dichotomous Accuracy Metrics	Dichotomous (Group) Fairness Metrics
False negative rate (**FNR**) $P(C=0 \mid Y=1)$	**Same FNR** across groups: $P(C=1 \mid Y=0 \& G=1) = P(C=1 \mid Y=0 \& G=0)$
False positive rate (**FPR**) $P(C=1 \mid Y=0)$	**Same FPR** across groups: $P(C=0 \mid Y=1 \& G=1) = P(C=0 \mid Y=1 \& G=0)$
Positive predictive value (**PPV**) $P(Y=1 \mid C=1)$	**Same PPV** across groups: $P(Y=1 \mid C=1 \& G=1) = P(Y=1 \mid C=1 \& G=0)$

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

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**)
$P(Y=1 \mid C=1)$

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

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

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

1. Impossibility Theorems

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

How Broadly Do Impossibility Theorems Apply?

Theorems such as Chouldechova's are usually formulated in the literature on predictive algorithms (Kleinberg, Mullainathan, Raghavan 2016) and in the literature about the fairness of test scores (Borsboom, Romeijn, Wicherts, 2008)

How Broadly Do Impossibility Theorems Apply?

Theorems such as Chouldechova's are usually formulated in the literature on predictive algorithms (Kleinberg, Mullainathan, Raghavan 2016) and in the literature about the fairness of test scores (Borsboom, Romeijn, Wicherts, 2008)

Shall we assume these impossibility theorems *only* apply to:

1. predictive evidence
2. quantitative evidence

They Apply Broadly

Impossibility theorems about algorithmic fairness **generalize to any dichotomous evidence-based decisions** (in fact, any dichotomous decision), whether or not the decision is aided by a predictive model.

They Apply Broadly

Impossibility theorems about algorithmic fairness **generalize to any dichotomous evidence-based decisions** (in fact, any dichotomous decision), whether or not the decision is aided by a predictive model.

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

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

They Apply Broadly

Impossibility theorems about algorithmic fairness **generalize to any dichotomous evidence-based decisions** (in fact, any dichotomous decision), whether or not the decision is aided by a predictive model.

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

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

Thus, any dichotomous decision would seem unfair one way or another.

2. Are All Decisions Unfair?

Predictive / Diagnostic Evidence

Predictive / Diagnostic Evidence

Decisions based on ***predictive evidence*** — e.g. a bank refuses to grant a mortgage to an applicant because the bank predicts the applicant will not repay it

Predictive / Diagnostic Evidence

Decisions based on ***predictive evidence*** — e.g. a bank refuses to grant a mortgage to an applicant because the bank predicts the applicant will not repay it

Decisions based on ***diagnostic evidence*** — e.g. a doctor diagnoses the presence of appendicitis based on a CT scan showing inflammation of the patient’s appendix

Predictive / Diagnostic Evidence

Decisions based on ***predictive evidence*** — e.g. a bank refuses to grant a mortgage to an applicant because the bank predicts the applicant will not repay it

Decisions based on ***diagnostic evidence*** — e.g. a doctor diagnoses the presence of appendicitis based on a CT scan showing inflammation of the patient's appendix

It is tricky to draw the distinction. In the medical context, the distinction is formulated in terms of **screening** v. **diagnostic tests**

Example (1): Medical Diagnosis

Example (1): Medical Diagnosis

Group **H**: **higher prevalence** of outcome Y=1

Group **L**: **lower prevalence** of outcome Y=1

Example (1): Medical Diagnosis

Group **H**: **higher prevalence** of outcome $Y=1$

Group **L**: **lower prevalence** of outcome $Y=1$

Everybody undergoes a test (or series of tests) with **sensitivity** $P(T=1 \mid Y=1)$ and **specificity** $P(T=0 \mid Y=0)$.

The diagnostic properties of the test do not vary across groups.

Example (1): Medical Diagnosis

Group **H**: **higher prevalence** of outcome Y=1

Group **L**: **lower prevalence** of outcome Y=1

Everybody undergoes a test (or series of tests) with **sensitivity** P(T=1 | Y=1) and **specificity** P(T=0 | Y=0).

The diagnostic properties of the test do not vary across groups.

Using Bayes' theorem, the doctor combines

(a) **prevalence data**

(b) the outcome of **diagnostic test**

This yields a posterior probability ***p*** that **Y=1**

The decision (here a diagnosis) is dichotomous:

C=1 iff *p*>*t* and C=0 otherwise

Example (1): Medical Diagnosis

Group **H**: **higher prevalence** of outcome Y=1

Group **L**: **lower prevalence** of outcome Y=1

Everybody undergoes a test (or series of tests) with **sensitivity** P(T=1 | Y=1) and **specificity** P(T=0 | Y=0).

The diagnostic properties of the test do not vary across groups.

Using Bayes' theorem, the doctor combines

(a) **prevalence data**

(b) the outcome of **diagnostic test**

This yields a posterior probability ***p*** that **Y=1**

The decision (here a diagnosis) is dichotomous:

C=1 iff *p>t* and C=0 otherwise

By Chouldechova's impossibility, ***either*** *predictive parity* ***or*** *classification parity are violated* across groups H and L

What If We Only Use Diagnostic Evidence?

What If We Only Use Diagnostic Evidence?

Suppose the doctor makes decisions based *solely* on diagnostic tests (that is, **T=1 iff C=1** and **T=0 iff C=0**).

Classification parity is satisfied:

$$P(C=1 \mid Y=1 \;\&\; \mathbf{G=H}) = P(C=1 \mid Y=1 \;\&\; \mathbf{G=L})$$

What If We Only Use Diagnostic Evidence?

Suppose the doctor makes decisions based *solely* on diagnostic tests (that is, **T=1 iff C=1** and **T=0 iff C=0**).

C*lassification parity is satisfied:*

$$P(C{=}1 \mid Y{=}1 \;\&\; \mathbf{G{=}H}) = P(C{=}1 \mid Y{=}1 \;\&\; \mathbf{G{=}L})$$

By hypothesis, the tests are assumed to have the same sensitivity and specificity across groups

What If We Only Use Diagnostic Evidence?

Suppose the doctor makes decisions based *solely* on diagnostic tests (that is, **T=1 iff C=1** and **T=0 iff C=0**).

Classification parity is satisfied:

P(C=1 | Y=1 & **G=H**) = P(C=1 | Y=1 & **G=L**)

Predictive parity is violated:

P(Y=1 | C=1 & **G=H**) > P(Y=1 | C=1 & **G=L**)

By hypothesis, the tests are assumed to have the same sensitivity and specificity across groups

What If We Only Use Diagnostic Evidence?

Suppose the doctor makes decisions based *solely* on diagnostic tests (that is, **T=1 iff C=1** and **T=0 iff C=0**).

Classification parity is satisfied:

P(C=1 | Y=1 & **G=H**) = P(C=1 | Y=1 & **G=L**)

Predictive parity is violated:

P(Y=1 | C=1 & **G=H**) > P(Y=1 | C=1 & **G=L**)

By hypothesis, the tests are assumed to have the same sensitivity and specificity across groups

Given the same test outcome T=1, people in group H are more likely to have Y=1 than people in group L since prevalence of Y=1 is higher in H than in L

Example (2): Trial Decisions

Example (2): Trial Decisions

Using Bayes' theorem, jurors combine

(a) "profiling" or predictive evidence (prevalence data)

(b) "individualized" evidence (e.g. eyewitness testimony, trace evidence, fingerprint or DNA matches)

This yields a probability $\boldsymbol{p}$ that **Y=1**

The decision (here a conviction) is dichotomous:

C=1 iff $p>t$ and C=0 otherwise

Example (2): Trial Decisions

Using Bayes' theorem, jurors combine

(a) "profiling" or predictive evidence (prevalence data)

(b) "individualized" evidence (e.g. eyewitness testimony, trace evidence, fingerprint or DNA matches)

This yields a probability ***p*** that **Y=1**

The decision (here a conviction) is dichotomous:

C=1 iff *p>t* and C=0 otherwise

By Chouldechova's impossibility, ***either*** *predictive parity* ***or*** *classification parity are violated* across groups H and L

Example (2): Trial Decisions

Using Bayes' theorem, jurors combine

(a) "profiling" or predictive evidence (prevalence data)

(b) "individualized" evidence (e.g. eyewitness testimony, trace evidence, fingerprint or DNA matches)

This yields a probability ***p*** that **Y=1**

The decision (here a conviction) is dichotomous:

C=1 iff *p>t* and C=0 otherwise

We can assume that the quality of individualized evidence is evenly distributed across across groups and jurors assess evidence quality correctly (by applying some measure of evidence quality)

By Chouldechova's impossibility, ***either*** *predictive parity* ***or*** *classification parity are violated* across groups H and L

Example (2): Trial Decisions

Using Bayes' theorem, jurors combine

(a) "profiling" or predictive evidence (prevalence data)

(b) "individualized" evidence (e.g. eyewitness testimony, trace evidence, fingerprint or DNA matches)

This yields a probability ***p*** that **Y=1**

The decision (here a conviction) is dichotomous:

C=1 iff *p>t* and C=0 otherwise

By Chouldechova's impossibility, ***either*** *predictive parity* ***or*** *classification parity are violated* across groups H and L

We can assume that the quality of individualized evidence is evenly distributed across across groups and jurors assess evidence quality correctly (by applying some measure of evidence quality)

However, if crime prevalence is different across groups, profiling evidence will tend to incriminate more members of higher prevalence groups than lower prevalence groups.

What If We Only Use Individualized Evidence?

What If We Only Use Individualized Evidence?

Suppose jurors make decisions based *solely* on "individualized" evidence (such as eyewitness testimonies, crime traces, fingerprints and DNA matches) and exclude accurate prevalence data ("profiling" or predictive evidence)

Classification parity is satisfied:

P(C=1 | Y=1 & **G=H**) = P(C=1 | Y=1 & **G=L**)

What If We Only Use Individualized Evidence?

Suppose jurors make decisions based *solely* on “individualized” evidence (such as eyewitness testimonies, crime traces, fingerprints and DNA matches) and exclude accurate prevalence data (“profiling” or predictive evidence)

Classification parity is satisfied:

$$P(C=1 \mid Y=1 \ \& \ \mathbf{G=H}) = P(C=1 \mid Y=1 \ \& \ \mathbf{G=L})$$

Assume evidence quality is evenly distributed across across racial groups and jurors convict (C=1) only if the evidence is strongly incriminating (by some measure of the quality of the evidence)

What If We Only Use Individualized Evidence?

Suppose jurors make decisions based *solely* on “individualized” evidence (such as eyewitness testimonies, crime traces, fingerprints and DNA matches) and exclude accurate prevalence data (“profiling” or predictive evidence)

Classification parity is satisfied:

$$P(C=1 \mid Y=1 \ \& \ \mathbf{G=H}) = P(C=1 \mid Y=1 \ \& \ \mathbf{G=L})$$

Predictive parity is violated:

$$P(Y=1 \mid C=1 \ \& \ \mathbf{G=H}) > P(Y=1 \mid C=1 \ \& \ \mathbf{G=L})$$

Assume evidence quality is evenly distributed across across racial groups and jurors convict (C=1) only if the evidence is strongly incriminating (by some measure of the quality of the evidence)

What If We Only Use Individualized Evidence?

Suppose jurors make decisions based *solely* on "individualized" evidence (such as eyewitness testimonies, crime traces, fingerprints and DNA matches) and exclude accurate prevalence data ("profiling" or predictive evidence)

Classification parity is satisfied:

$$P(C=1 \mid Y=1 \,\&\, \mathbf{G=H}) = P(C=1 \mid Y=1 \,\&\, \mathbf{G=L})$$

Predictive parity is violated:

$$P(Y=1 \mid C=1 \,\&\, \mathbf{G=H}) > P(Y=1 \mid C=1 \,\&\, \mathbf{G=L})$$

Assume evidence quality is evenly distributed across across racial groups and jurors convict (C=1) only if the evidence is strongly incriminating (by some measure of the quality of the evidence)

Given the same body of evidence, people in group H are more likely to have Y=1 than people in group L since prevalence of Y=1 is higher in H than in L

3. A Way Out

Question: If the impossibility theorems generalize to any evidence-based decision, *might* the underlying conception of fairness — i.e. classification parity *plus* predictive parity — be too demanding?

Classification Parity *Without* Predictive Parity

Classification Parity *Without* Predictive Parity

Suppose classification parity is satisfied, but predictive parity is violated. *Who would complain?*

Classification Parity *Without* Predictive Parity

Suppose classification parity is satisfied, but predictive parity is violated. *Who would complain?*

- No identifiable group can make a complaint. The group of those convicted or left untreated does not exist *ex ante* of the decision. Anyone could be in it. This group can voice no complaint as it does not exist in a socially stable way.

Classification Parity *Without* Predictive Parity

Suppose classification parity is satisfied, but predictive parity is violated. *Who would complain?*

- No identifiable group can make a complaint. The group of those convicted or left untreated does not exist *ex ante* of the decision. Anyone could be in it. This group can voice no complaint as it does not exist in a socially stable way.
- Instead, the innocent or sick form well-defined groups that exist *ex ante* of the decision being made.

A Compromise

A Compromise

Consider a predictive model that:

1. satisfies classification parity
2. violates predictive parity
3. satisfies a close cousin of predictive parity, namely calibration

A Compromise

Consider a predictive model that:

1. satisfies classification parity
2. violates predictive parity
3. satisfies a close cousin of predictive parity, namely calibration

Same positive predictive value across groups:
P(Y=1 | **C=1** & G=1) = P(Y=1 | **C=1** & G=0)

Predictive parity

A Compromise

Consider a predictive model that:

1. satisfies classification parity
2. violates predictive parity
3. satisfies a close cousin of predictive parity, namely calibration

`Calibration`

Same score predictive value across groups:
P(Y=1 | **S=s** & G=1) = P(Y=1 | **S=s** & G=0)

Same positive predictive value across groups:
P(Y=1 | **C=1** & G=1) = P(Y=1 | **C=1** & G=0)

`Predictive parity`

Classification Parity and *Predictive Parity*

$$G \leftarrow Y \rightarrow C$$
$$G \rightarrow Y \rightarrow C$$
$$G \leftarrow Y \leftarrow C$$

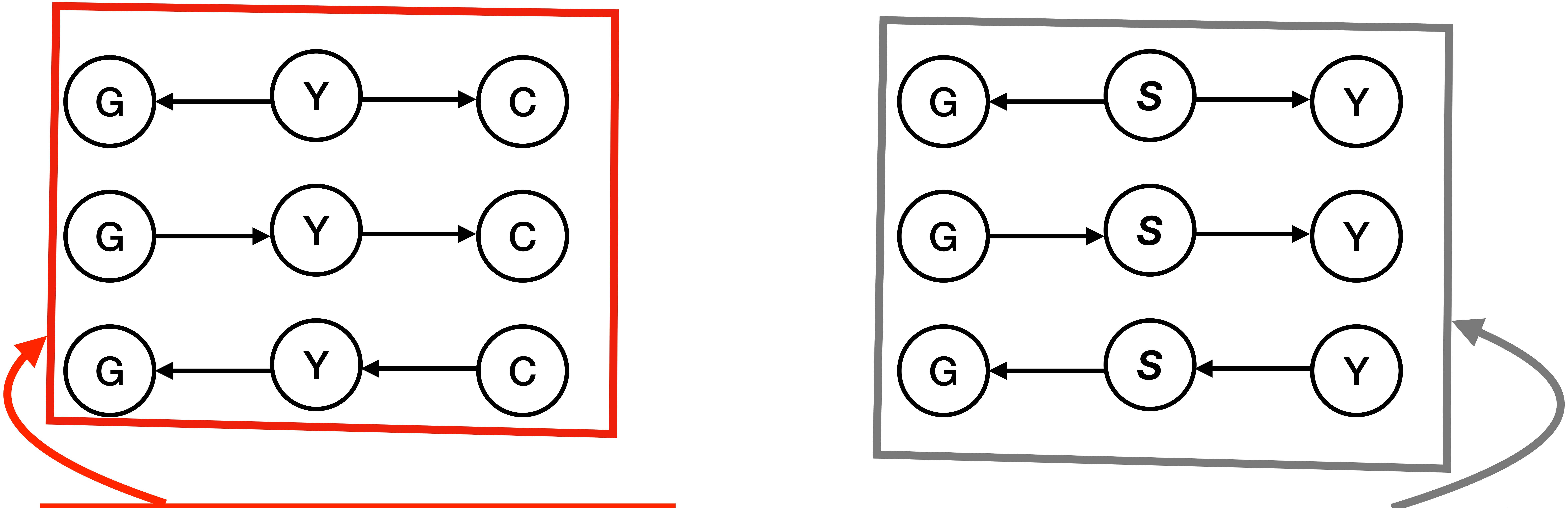
Classification parity:
G and C are independent given Y

$$G \leftarrow C \rightarrow Y$$
$$G \rightarrow C \rightarrow Y$$
$$G \leftarrow C \leftarrow Y$$

Predictive parity:
G and Y are independent given C

No causal diagram can guarantee both no matter the kind of evidence used

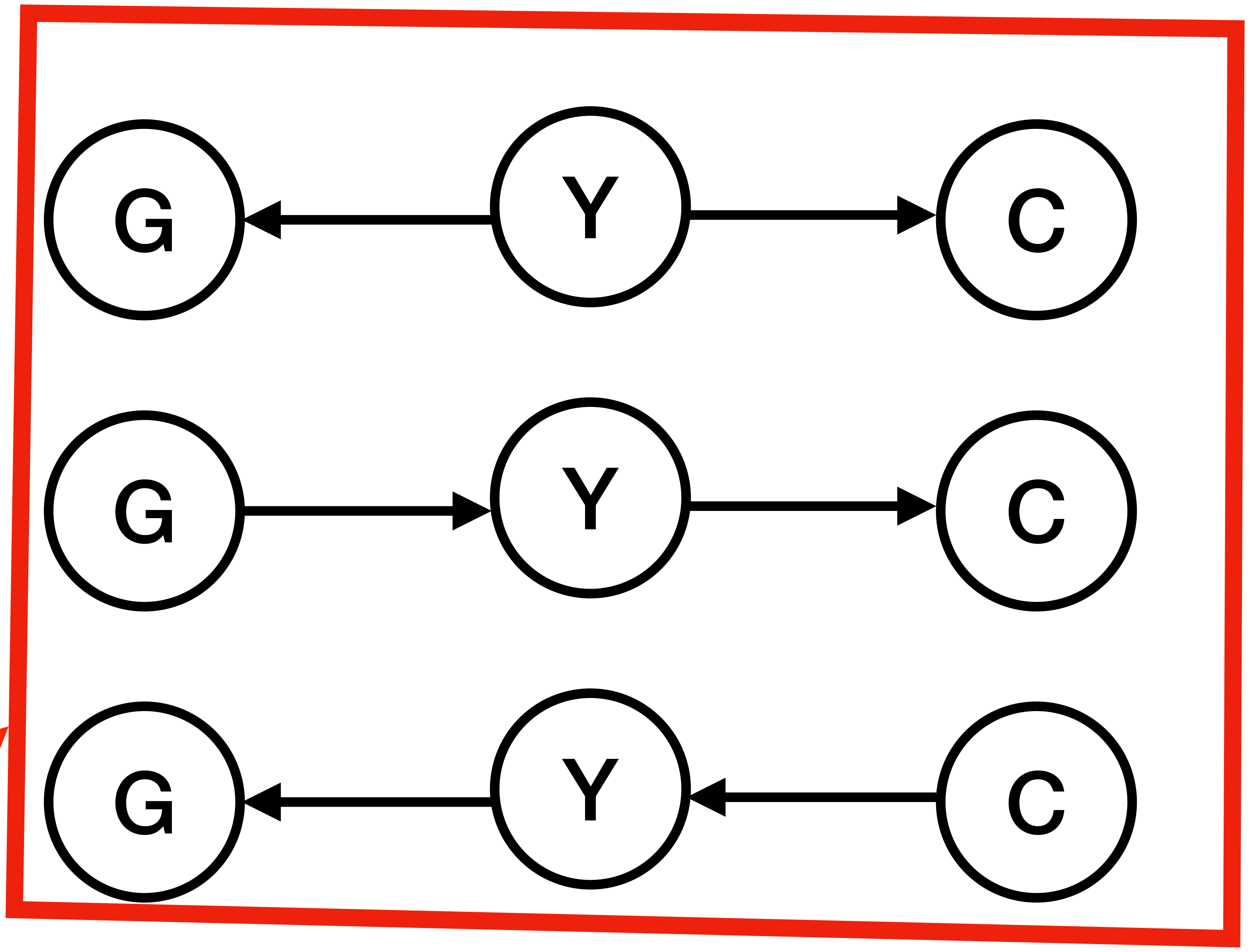
Classification Parity and Calibration



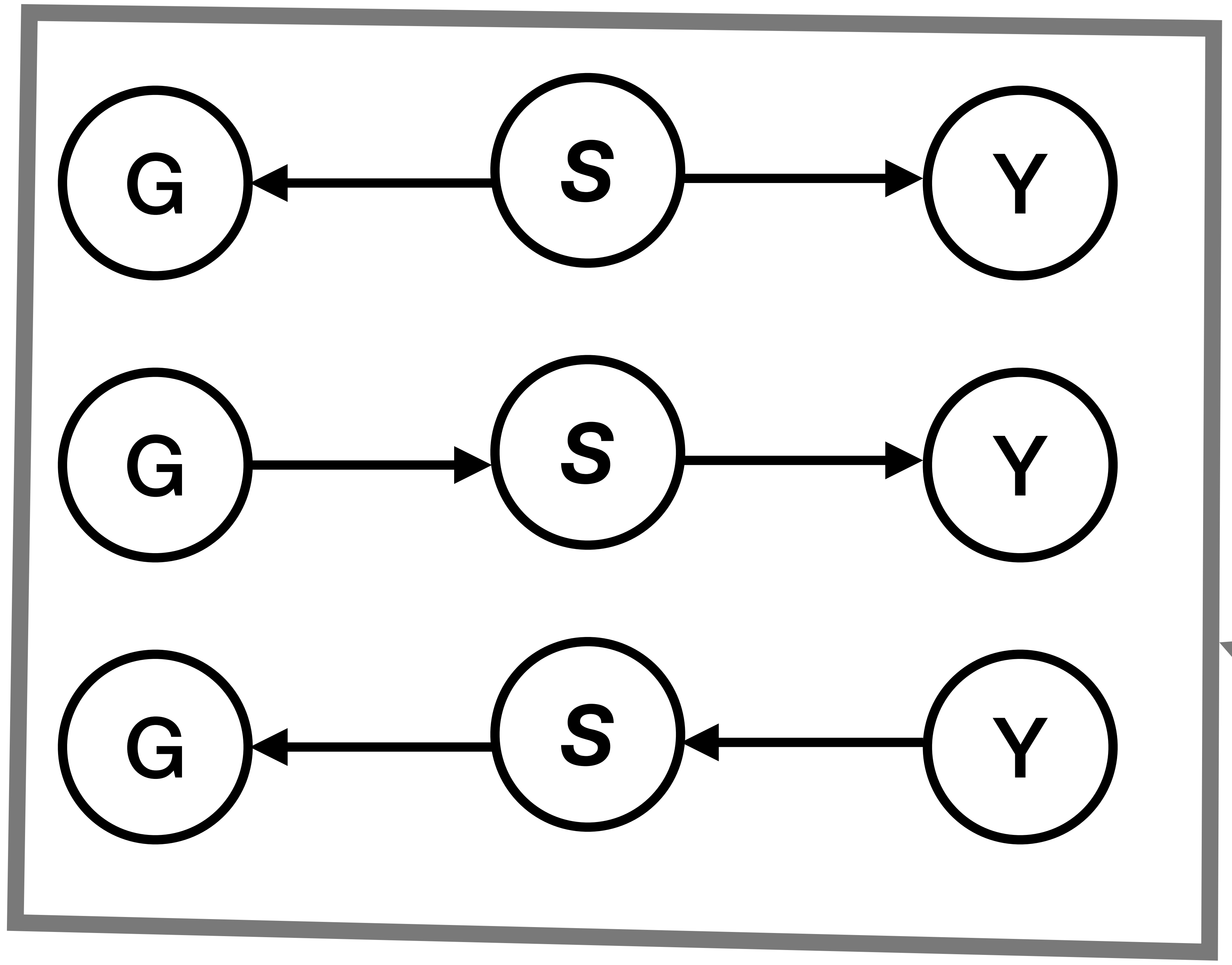
Classification parity:
G and C are independent given Y

Calibration:
G and Y are independent given S

Classification Parity and Calibration



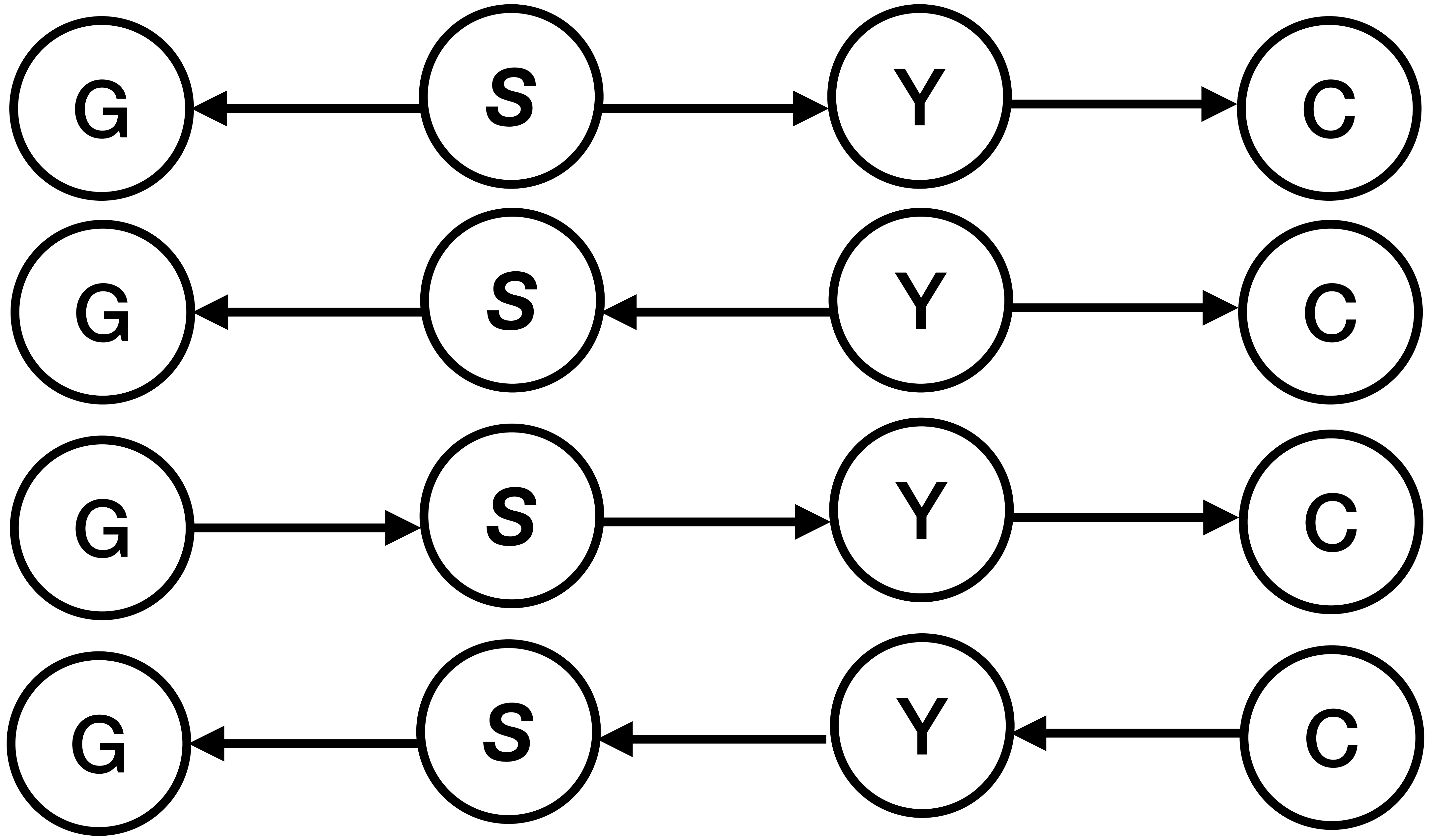
Classification parity:
G and C are independent given Y



Calibration:
G and Y are independent given S

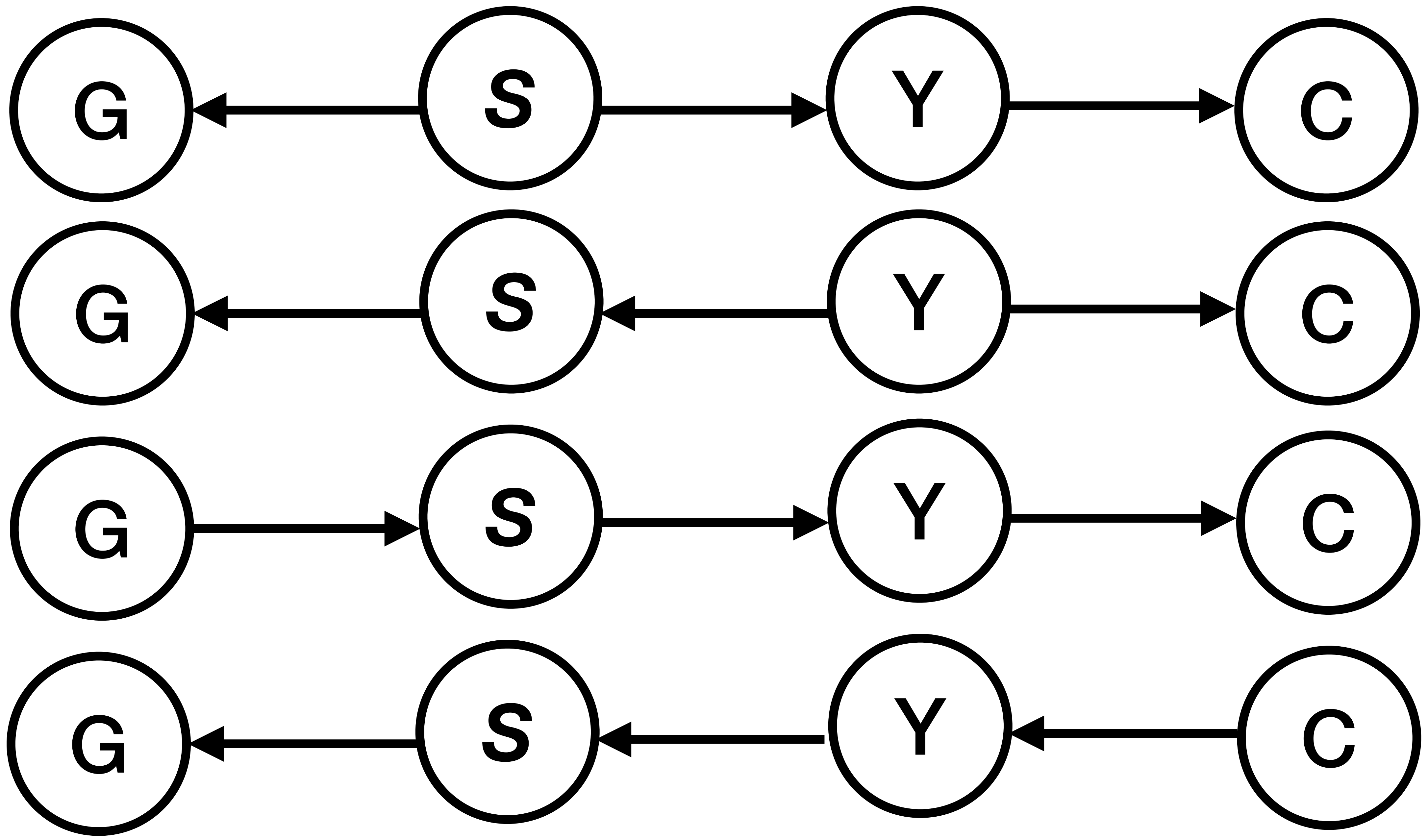
There is a diagram that can guarantee both

Classification Parity and *Calibration*



These diagrams guarantee both

Classification Parity and Calibration



Classification parity:
G and C are independent given Y

These diagrams guarantee both

Classification Parity and Calibration

Calibration:
G and Y are independent given S

Classification parity:
G and C are independent given Y

These diagrams guarantee both

4. Conclusion

Key Points

Key Points

(1) Impossibility theorems about algorithmic fairness reach beyond machine learning and predictive models. They also apply to **diagnostic** and **individualized** evidence in medicine and law.

These impossibility theorems underscore the unfairness of **any** evidence-based decisions.

Key Points

(1) Impossibility theorems about algorithmic fairness reach beyond machine learning and predictive models. They also apply to **diagnostic** and **individualized** evidence in medicine and law.

These impossibility theorems underscore the unfairness of **any** evidence-based decisions.

(2) But *perhaps* we can be content with a compromise:

- satisfy **classification parity** together with **calibration** (a close cousin of predictive parity)

Thank you!