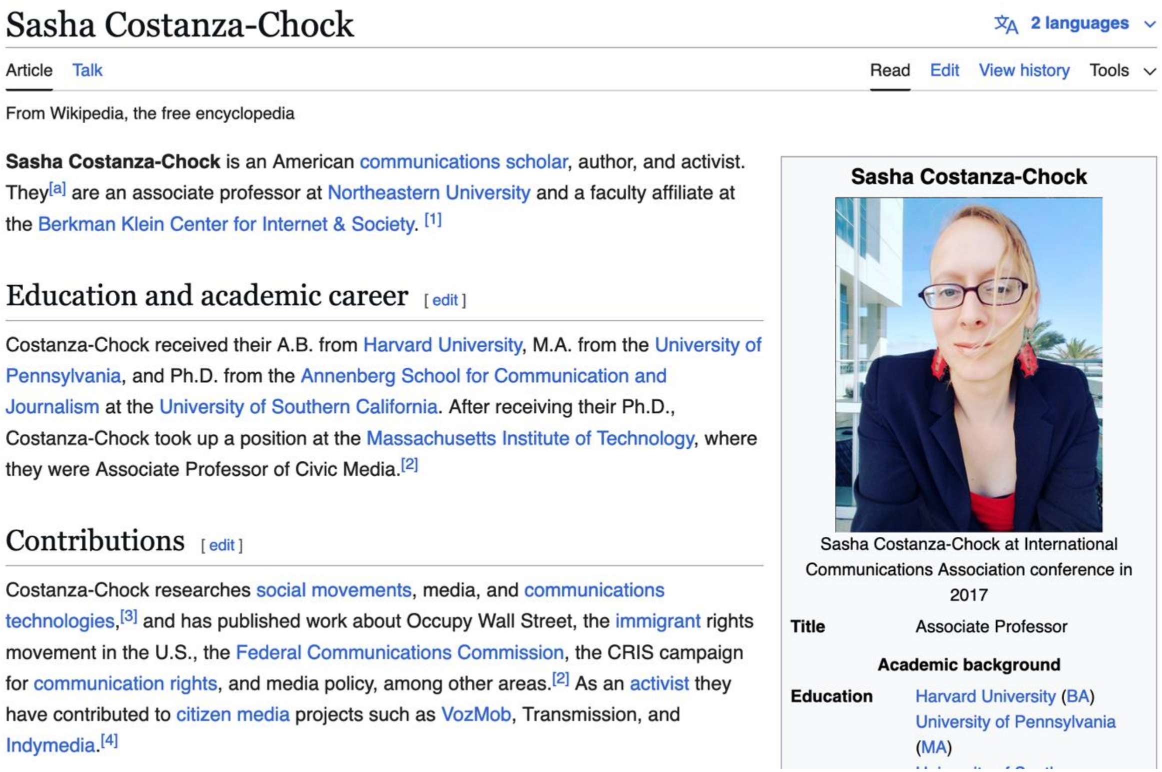
Impossibility of “Ideal” Fairness

Lecture 2

ML and Society (Spring 2025)

Atri Rudra

Pass phrase for today: Sasha Costanza-Chock



Sasha Costanza-Chock

From Wikipedia, the free encyclopedia

Sasha Costanza-Chock is an American communications scholar, author, and activist. They[a] are an associate professor at Northeastern University and a faculty affiliate at the Berkman Klein Center for Internet & Society. [1]

Education and academic career [edit]

Costanza-Chock received their A.B. from Harvard University, M.A. from the University of Pennsylvania, and Ph.D. from the Annenberg School for Communication and Journalism at the University of Southern California. After receiving their Ph.D., Costanza-Chock took up a position at the Massachusetts Institute of Technology, where they were Associate Professor of Civic Media.[2]

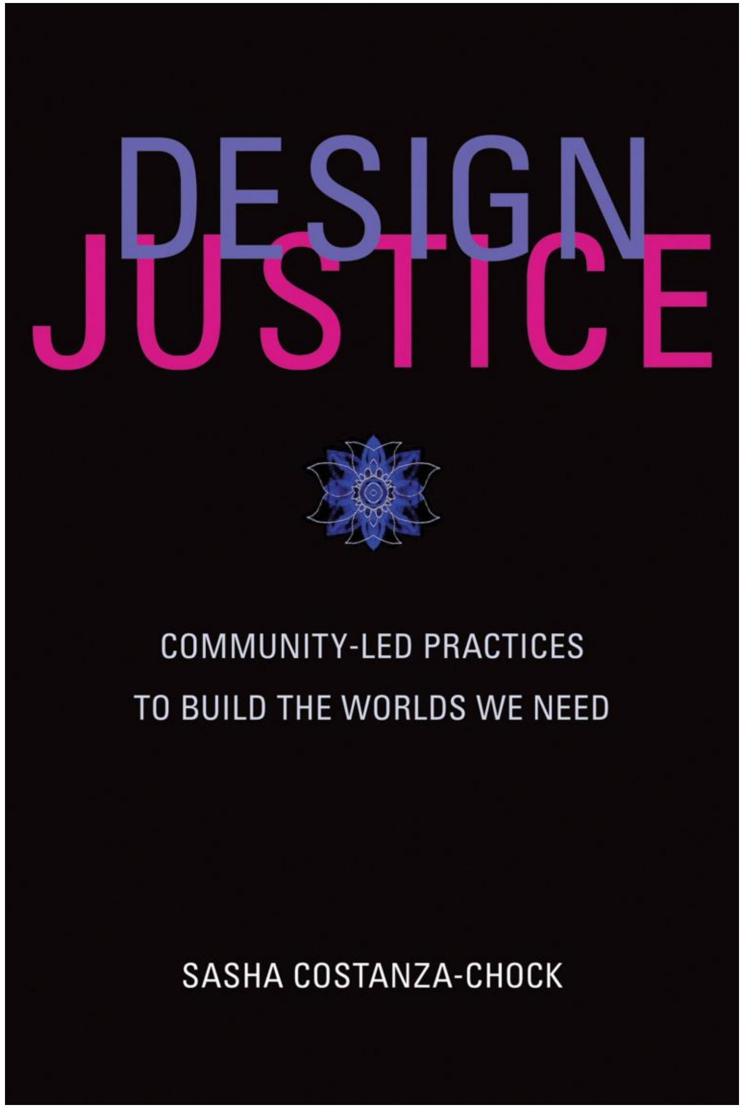
Contributions [edit]

Costanza-Chock researches social movements, media, and communications technologies,[3] and has published work about Occupy Wall Street, the immigrant rights movement in the U.S., the Federal Communications Commission, the CRIS campaign for communication rights, and media policy, among other areas.[2] As an activist they have contributed to citizen media projects such as VozMob, Transmission, and Indymedia.[4]

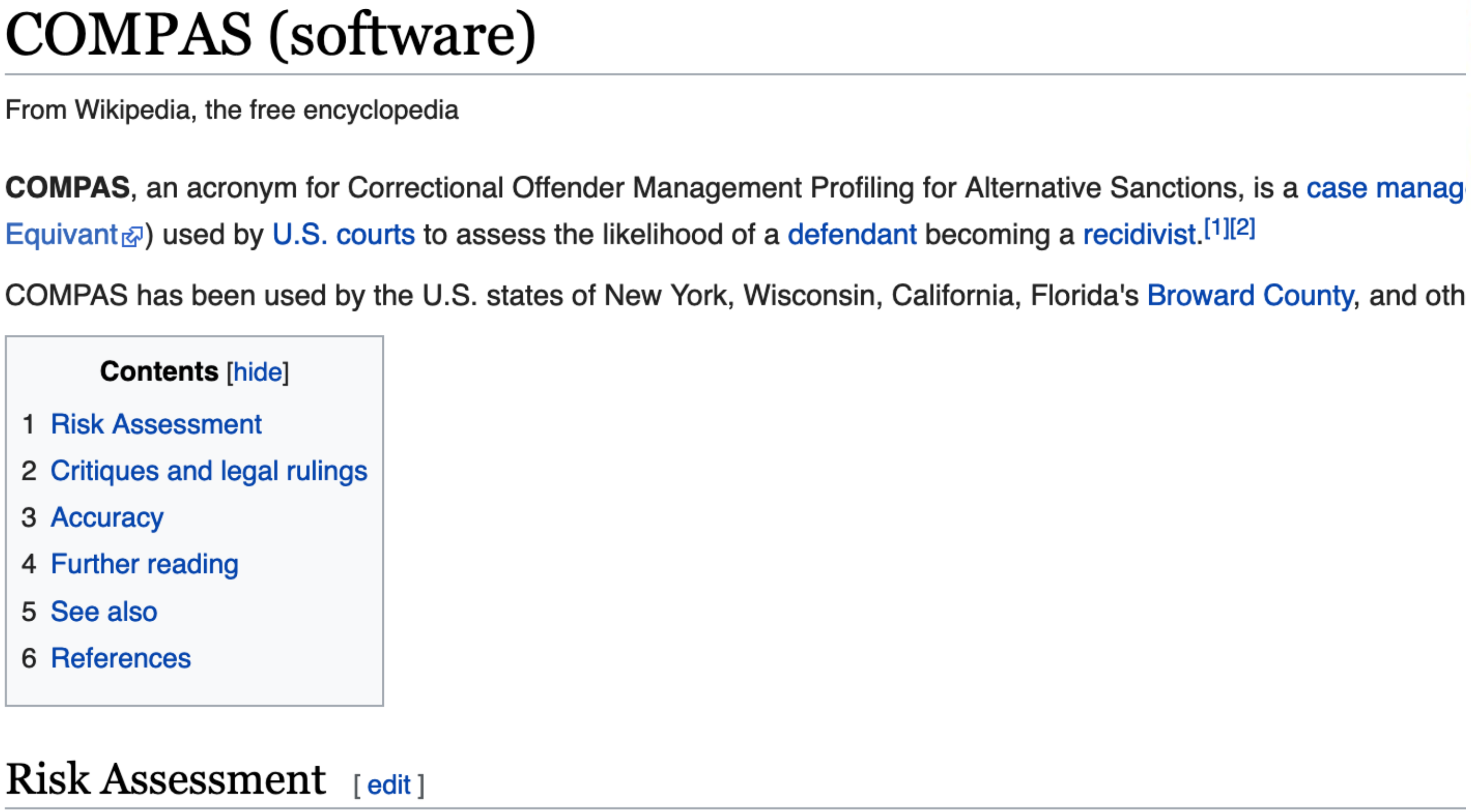
Sasha Costanza-Chock

Sasha Costanza-Chock at International Communications Association conference in 2017

Title	Associate Professor
Academic background	
Education	Harvard University (BA) University of Pennsylvania (MA)



Recall COMPAS



COMPAS (software)

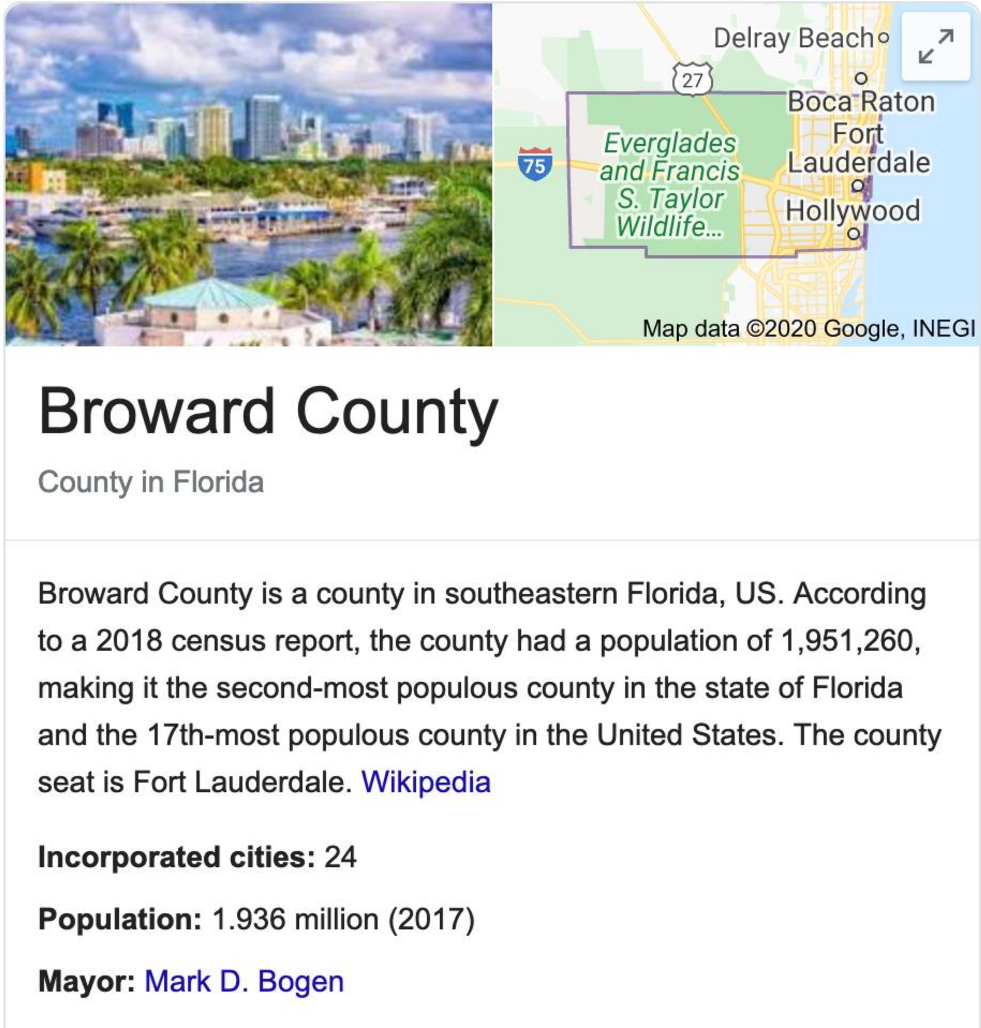
From Wikipedia, the free encyclopedia

COMPAS, an acronym for Correctional Offender Management Profiling for Alternative Sanctions, is a case manag Equivant) used by U.S. courts to assess the likelihood of a defendant becoming a recidivist.[1][2]

COMPAS has been used by the U.S. states of New York, Wisconsin, California, Florida's Broward County, and oth

Contents [hide]

Risk Assessment [edit]



Map data ©2020 Google, INEGI

Broward County

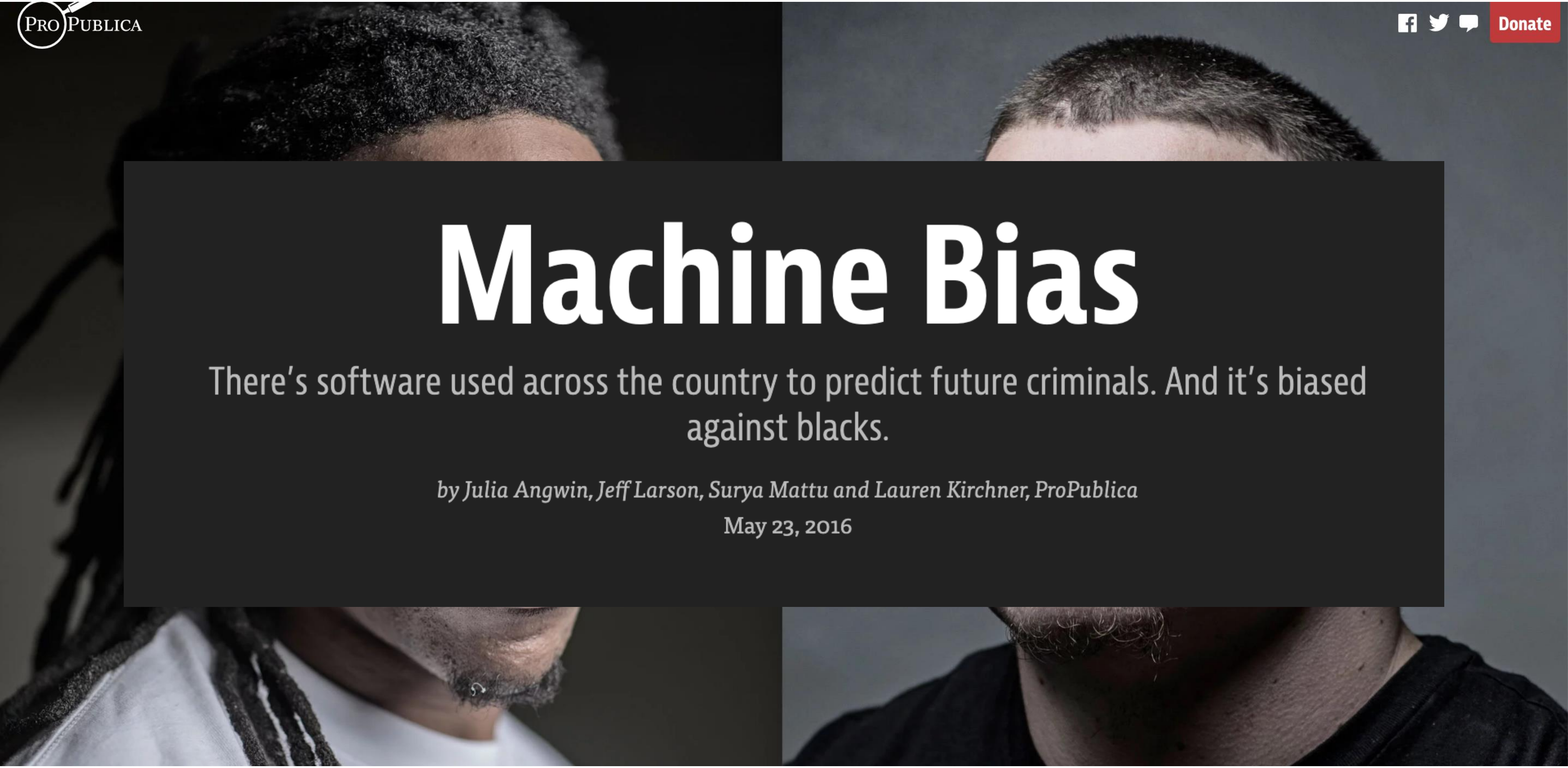
County in Florida

Broward County is a county in southeastern Florida, US. According to a 2018 census report, the county had a population of 1,951,260, making it the second-most populous county in the state of Florida and the 17th-most populous county in the United States. The county seat is Fort Lauderdale. Wikipedia

Incorporated cities: 24

Population: 1.936 million (2017)

Mayor: Mark D. Bogen



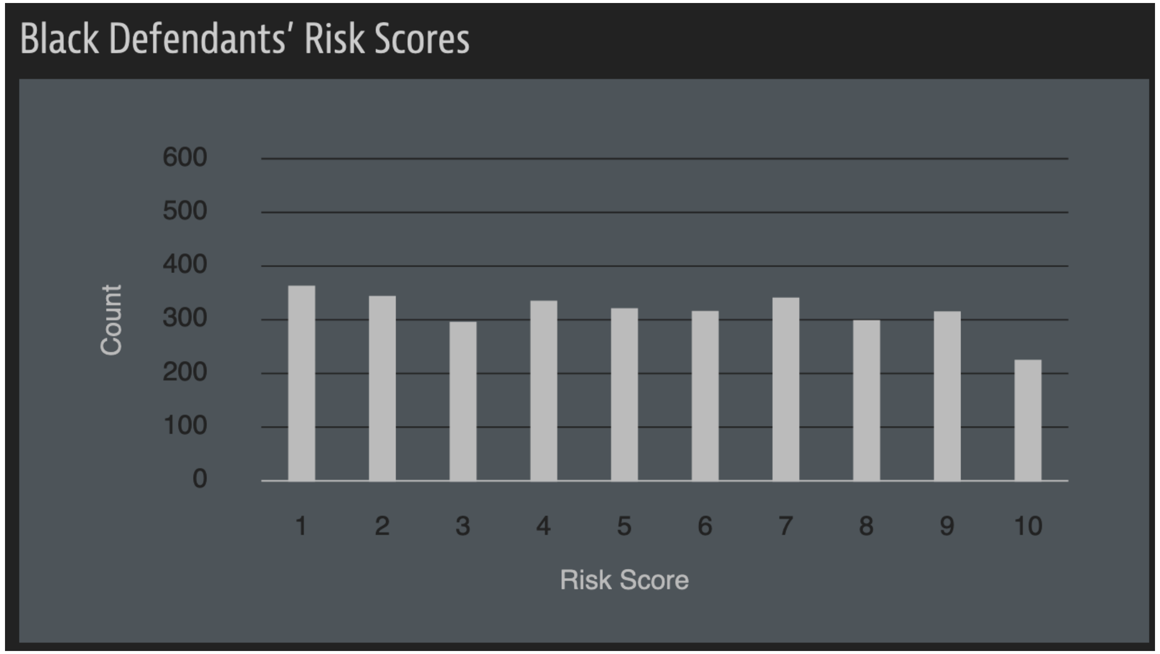
Machine Bias

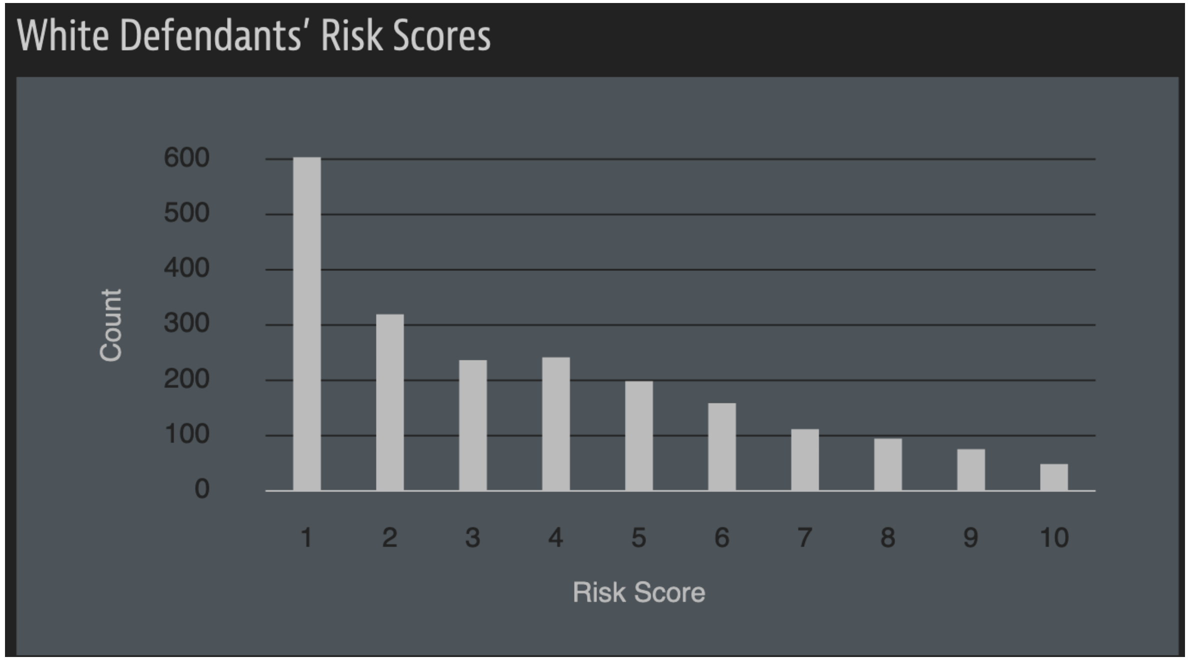
There’s software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

A sample of their result

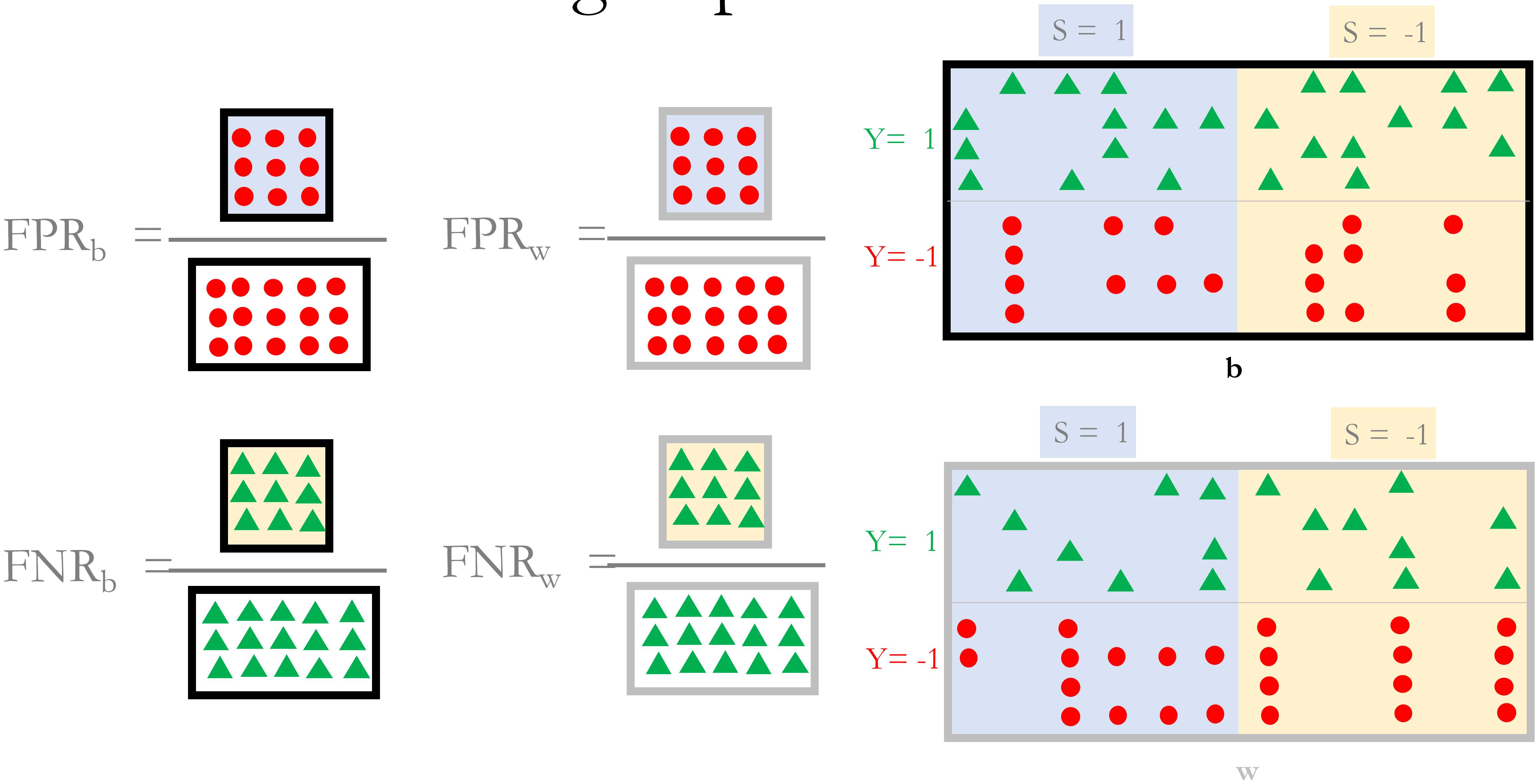




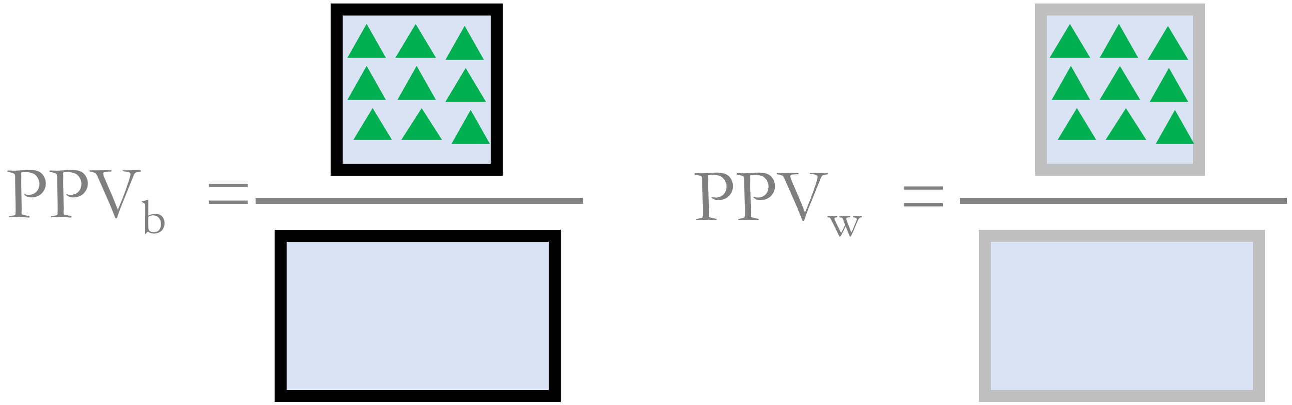
False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks."

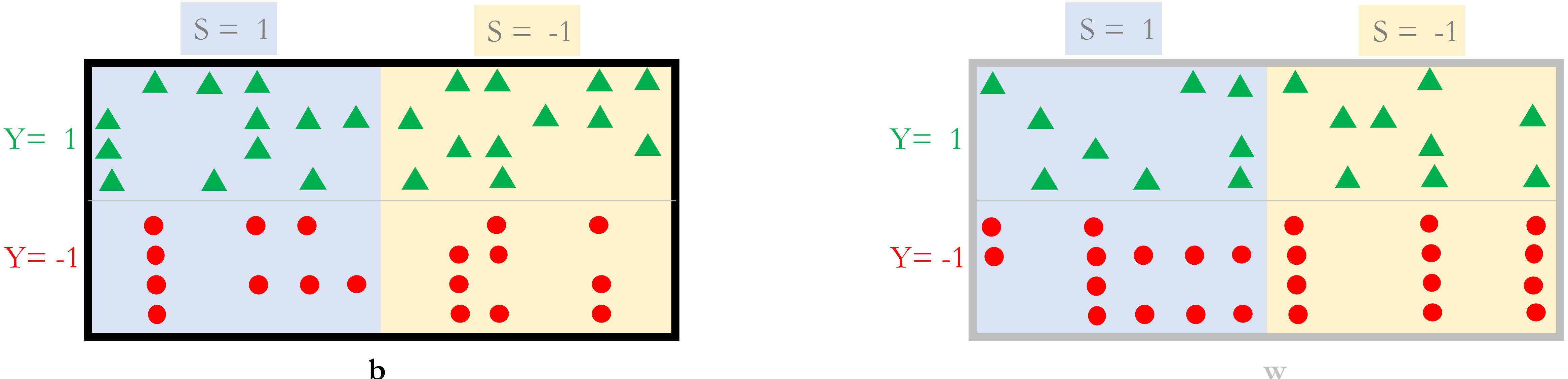
Anthony W. Flores
California State University, Bakersfield
Kristin Bechtel
Crime and Justice Institute at CRJ
Christopher T. Lowenkamp
Administrative Office of the United States Courts
Probation and Pretrial Services Office

FPR and FNR for groups

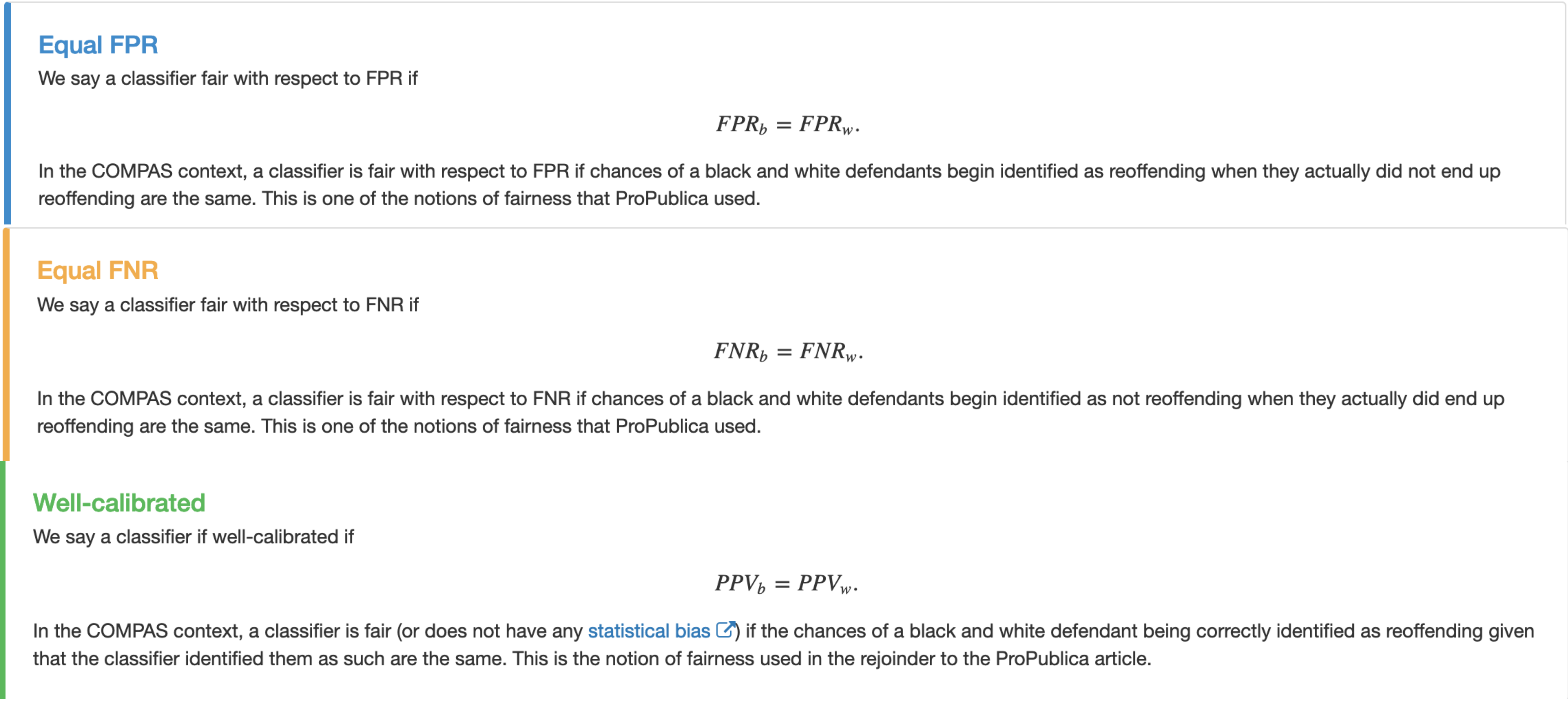


PPV for groups





Fairness definitions



Equal FPR

We say a classifier fair with respect to FPR if

$$FPR_b = FPR_w.$$

In the COMPAS context, a classifier is fair with respect to FPR if chances of a black and white defendants begin identified as reoffending when they actually did not end up reoffending are the same. This is one of the notions of fairness that ProPublica used.

Equal FNR

We say a classifier fair with respect to FNR if

$$FNR_b = FNR_w.$$

In the COMPAS context, a classifier is fair with respect to FNR if chances of a black and white defendants begin identified as not reoffending when they actually did end up reoffending are the same. This is one of the notions of fairness that ProPublica used.

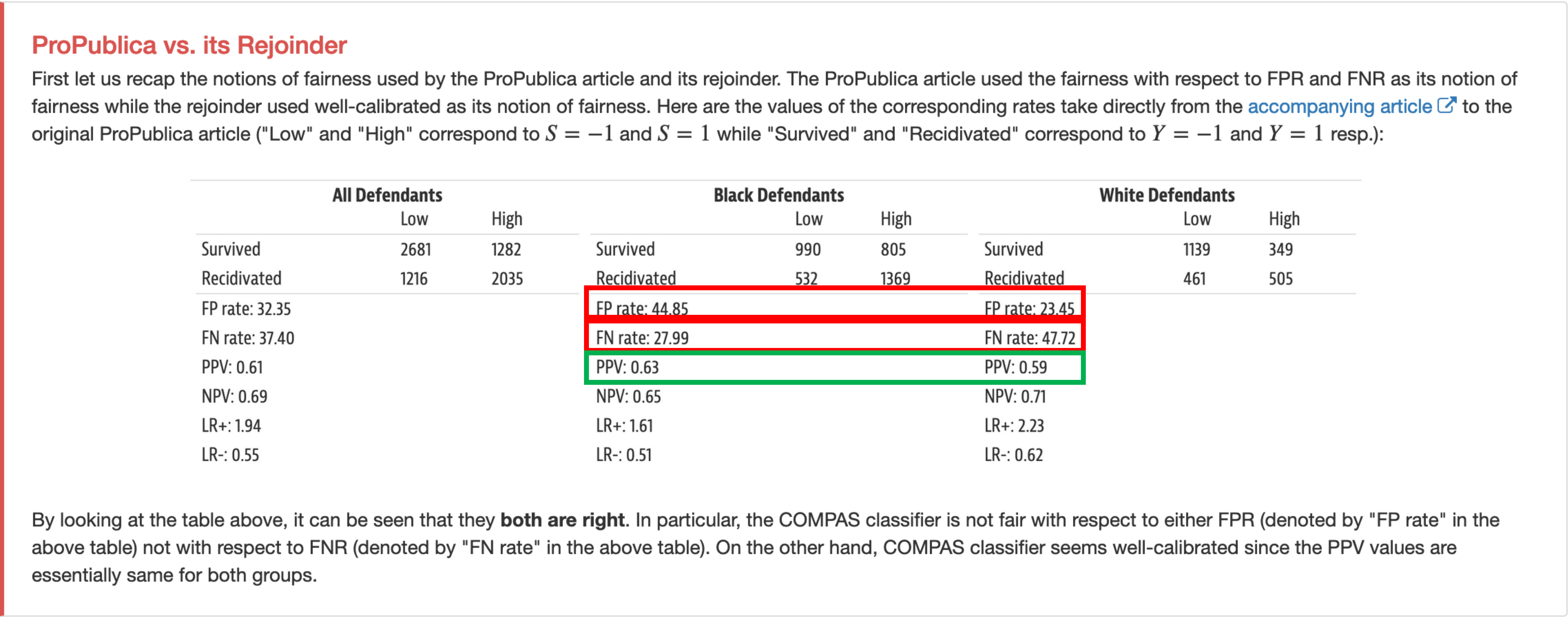
Well-calibrated

We say a classifier if well-calibrated if

$$PPV_b = PPV_w.$$

In the COMPAS context, a classifier is fair (or does not have any statistical bias) if the chances of a black and white defendant being correctly identified as reoffending given that the classifier identified them as such are the same. This is the notion of fairness used in the rejoinder to the ProPublica article.

Connecting back to COMPAS story



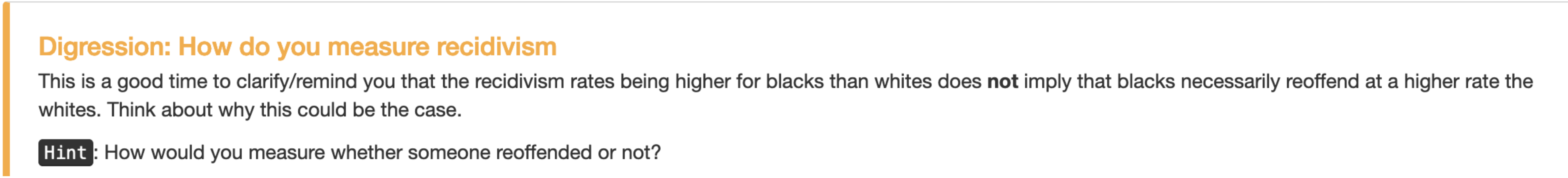
ProPublica vs. its Rejoinder

First let us recap the notions of fairness used by the ProPublica article and its rejoinder. The ProPublica article used the fairness with respect to FPR and FNR as its notion of fairness while the rejoinder used well-calibrated as its notion of fairness. Here are the values of the corresponding rates take directly from the accompanying article to the original ProPublica article ("Low" and "High" correspond to $S = -1$ and $S = 1$ while "Survived" and "Recidivated" correspond to $Y = -1$ and $Y = 1$ resp.):

All Defendants			Black Defendants			White Defendants		
	Low	High		Low	High		Low	High
Survived	2681	1282	Survived	990	805	Survived	1139	349
Recidivated	1216	2035	Recidivated	532	1369	Recidivated	461	505
FP rate: 32.35			FP rate: 44.85			FP rate: 23.45		
FN rate: 37.40			FN rate: 27.99			FN rate: 47.72		
PPV: 0.61			PPV: 0.63			PPV: 0.59		
NPV: 0.69			NPV: 0.65			NPV: 0.71		
LR+: 1.94			LR+: 1.61			LR+: 2.23		
LR-: 0.55			LR-: 0.51			LR-: 0.62		

By looking at the table above, it can be seen that they **both are right**. In particular, the COMPAS classifier is not fair with respect to either FPR (denoted by "FP rate" in the above table) not with respect to FNR (denoted by "FN rate" in the above table). On the other hand, COMPAS classifier seems well-calibrated since the PPV values are essentially same for both groups.

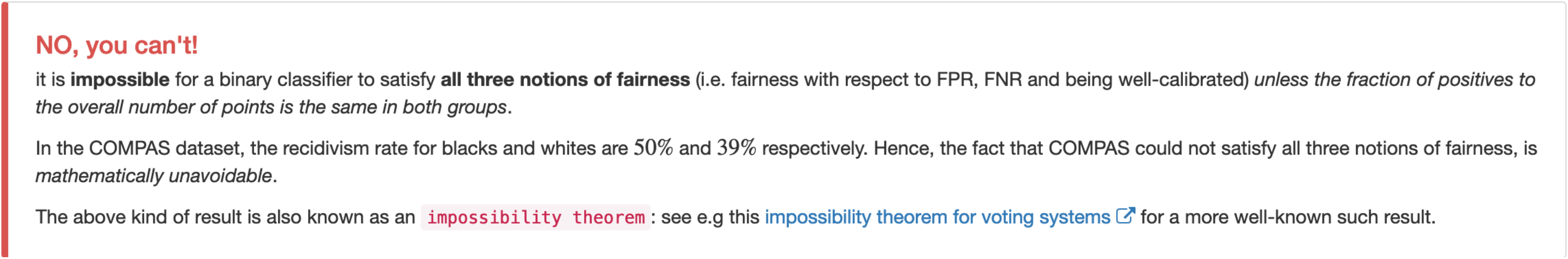
Perhaps COMPAS can be improved?



Digression: How do you measure recidivism

This is a good time to clarify/remind you that the recidivism rates being higher for blacks than whites does **not** imply that blacks necessarily reoffend at a higher rate the whites. Think about why this could be the case.

`Hint`: How would you measure whether someone reoffended or not?



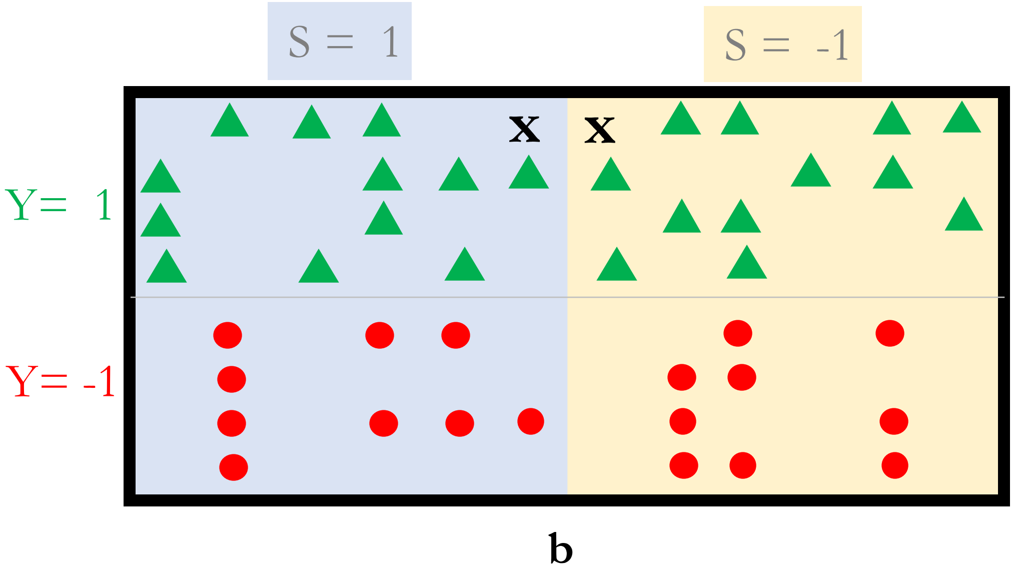
NO, you can't!

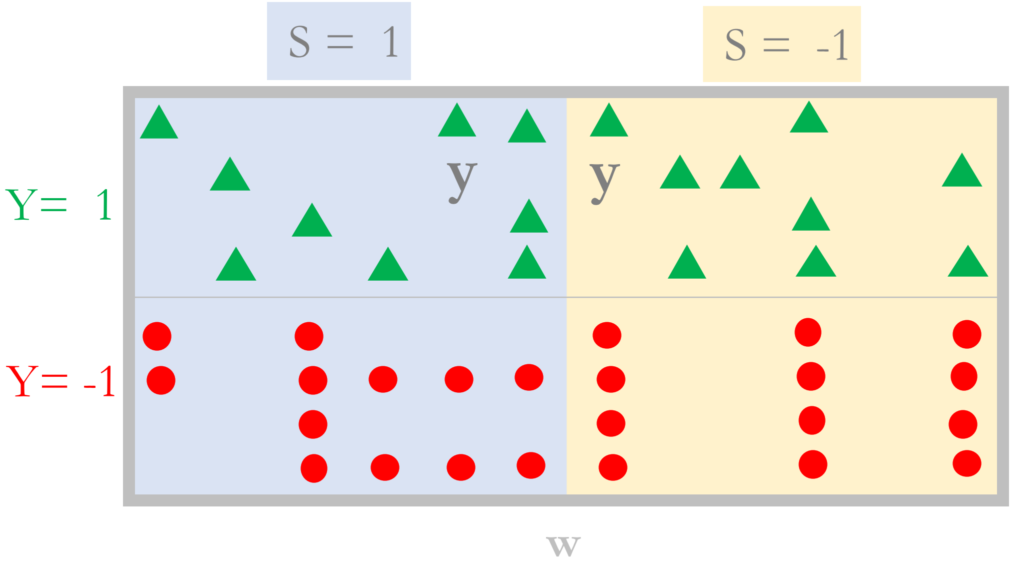
it is **impossible** for a binary classifier to satisfy **all three notions of fairness** (i.e. fairness with respect to FPR, FNR and being well-calibrated) *unless the fraction of positives to the overall number of points is the same in both groups*.

In the COMPAS dataset, the recidivism rate for blacks and whites are 50% and 39% respectively. Hence, the fact that COMPAS could not satisfy all three notions of fairness, is *mathematically unavoidable*.

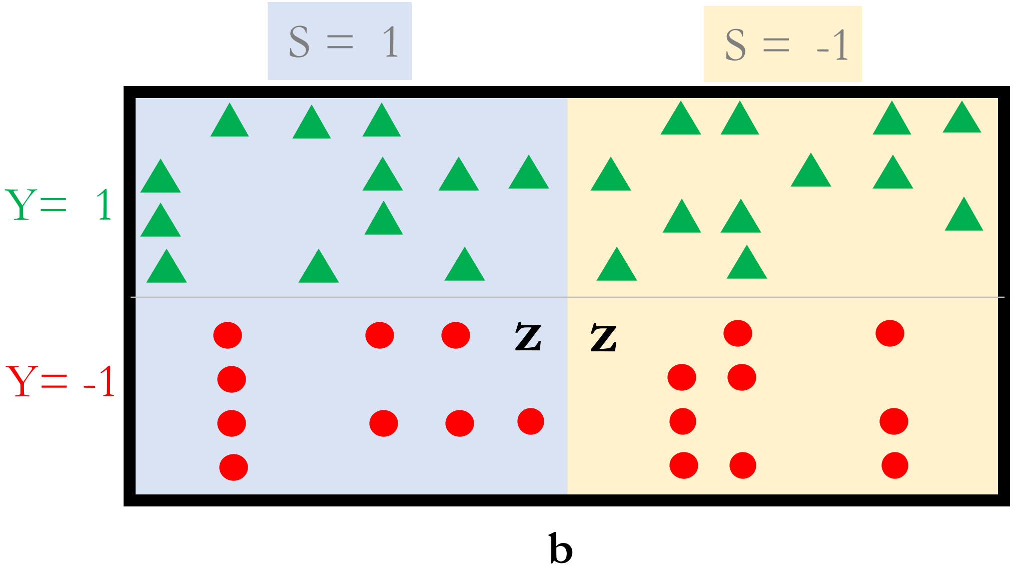
The above kind of result is also known as an `impossibility theorem`: see e.g this impossibility theorem for voting systems for a more well-known such result.

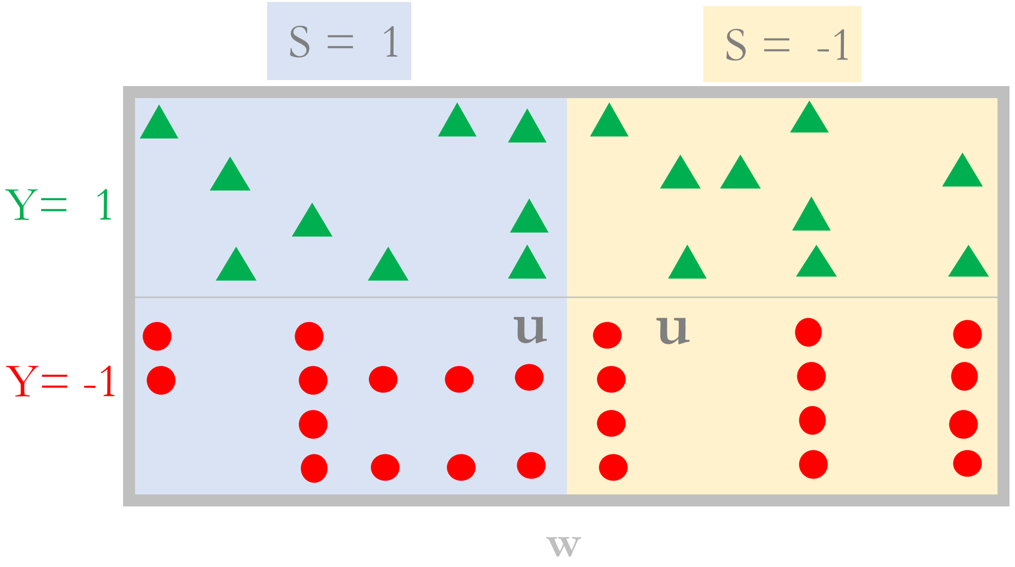
$$\mathrm{FNR}_b = \mathrm{FNR}_w = 1/2$$





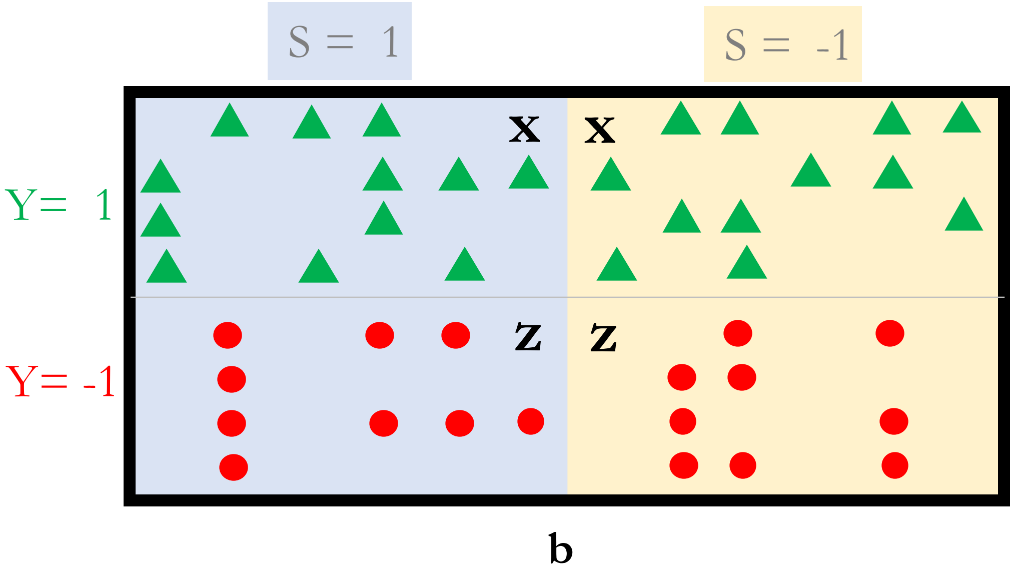
$$\mathrm{FPR}_b = \mathrm{FPR}_w = 1/2$$

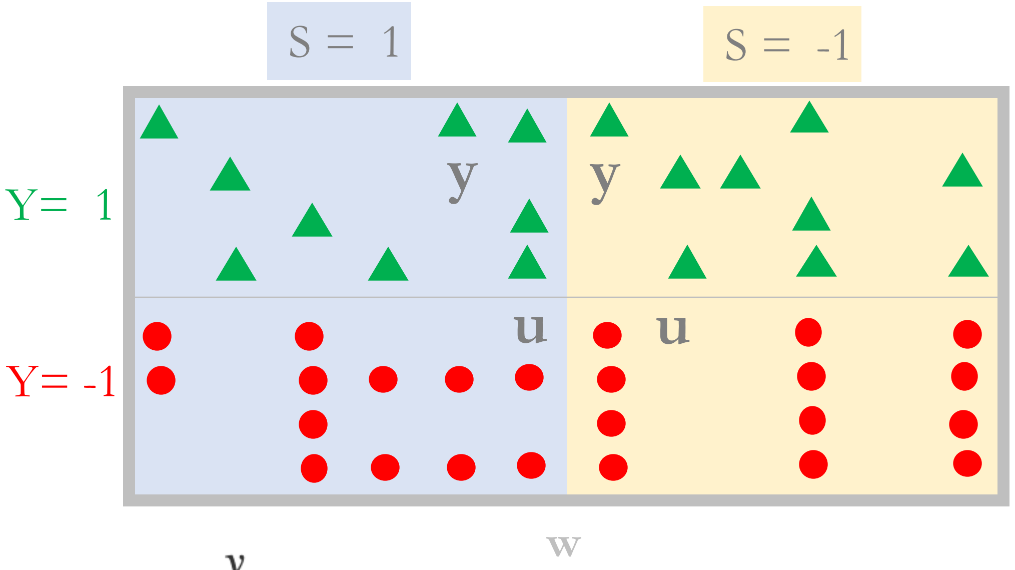




Overall situation

What are PPV_b and PPV_w?





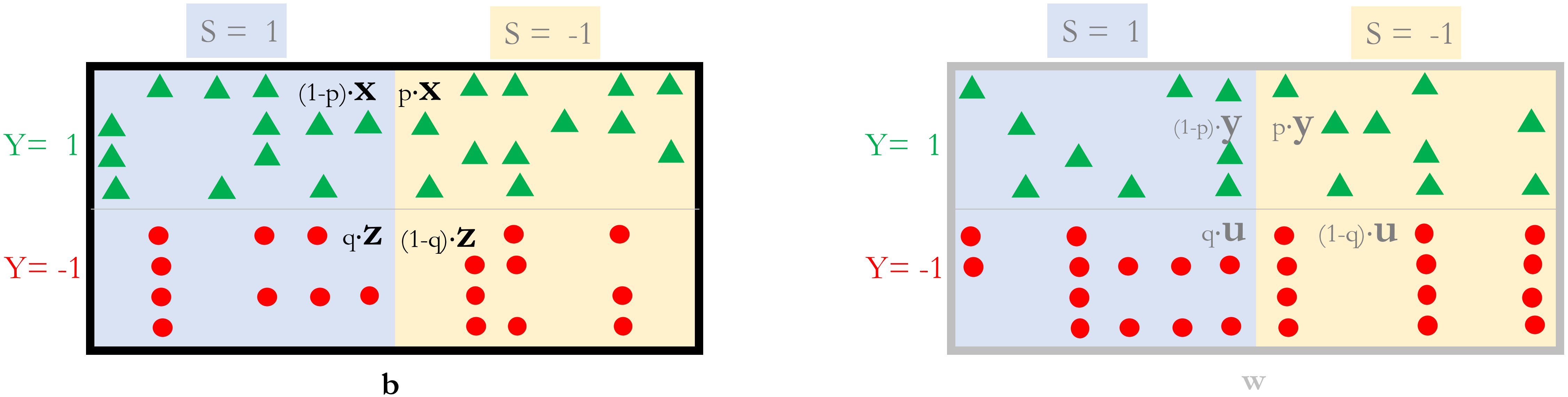
$$PPV_b = \frac{x}{x+z} \text{ and } PPV_w = \frac{y}{y+u}$$

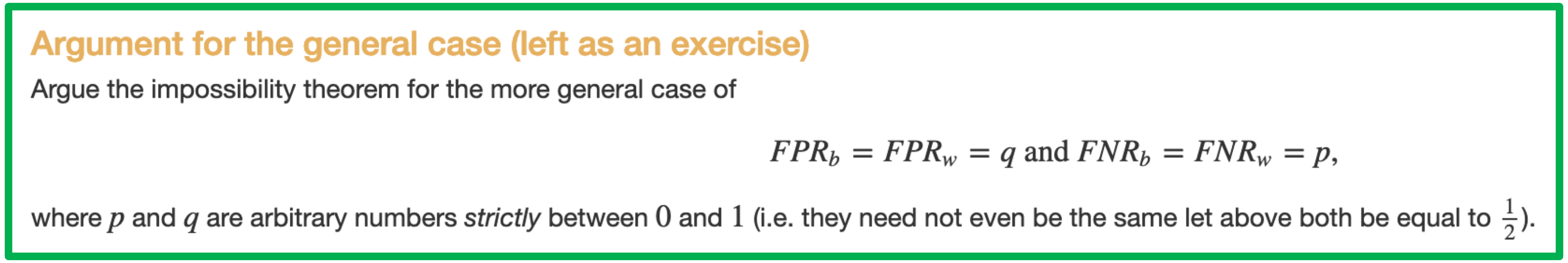
$$p_b = \frac{x+x}{z+z+x+x} = \frac{x}{x+z} = PPV_b.$$

When is PPV_b = PPV_w?

$$p_w = \frac{y+y}{y+y+u+u} = \frac{y}{y+u} = PPV_w.$$

Argue the general case!





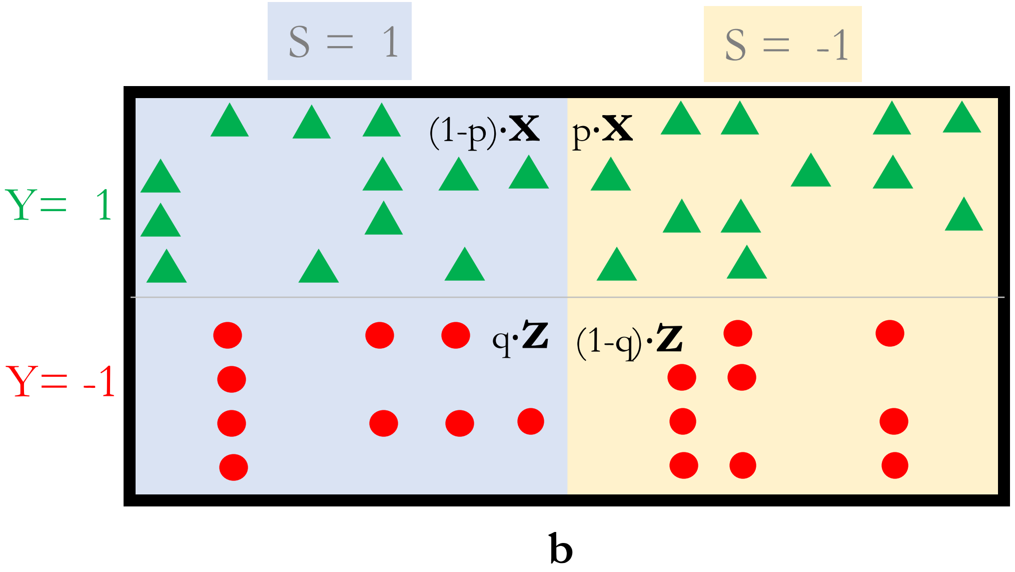
Argument for the general case (left as an exercise)

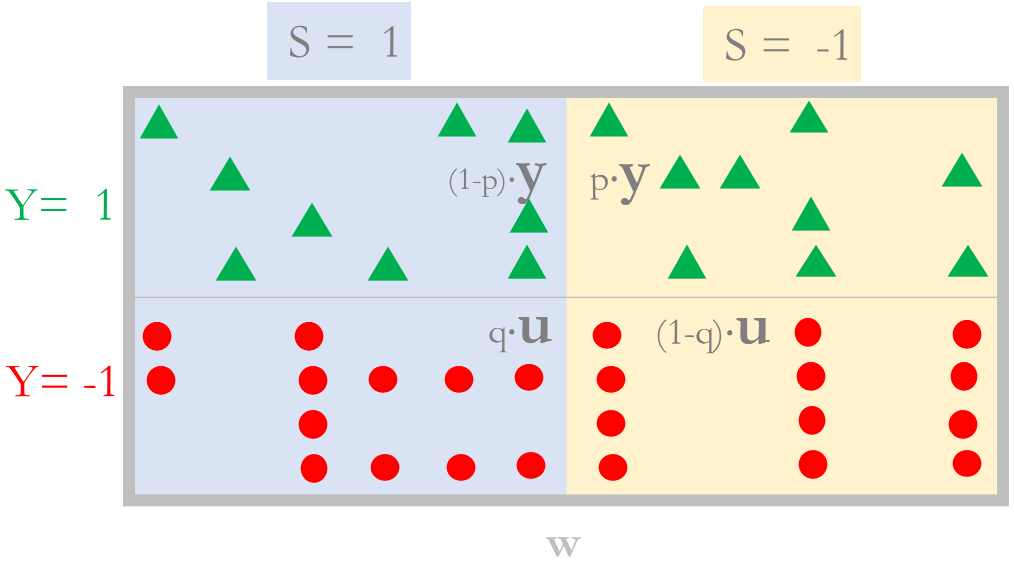
Argue the impossibility theorem for the more general case of

$$FPR_b = FPR_w = q \text{ and } FNR_b = FNR_w = p,$$

where p and q are arbitrary numbers *strictly* between 0 and 1 (i.e. they need not even be the same let above both be equal to $\frac{1}{2}$).

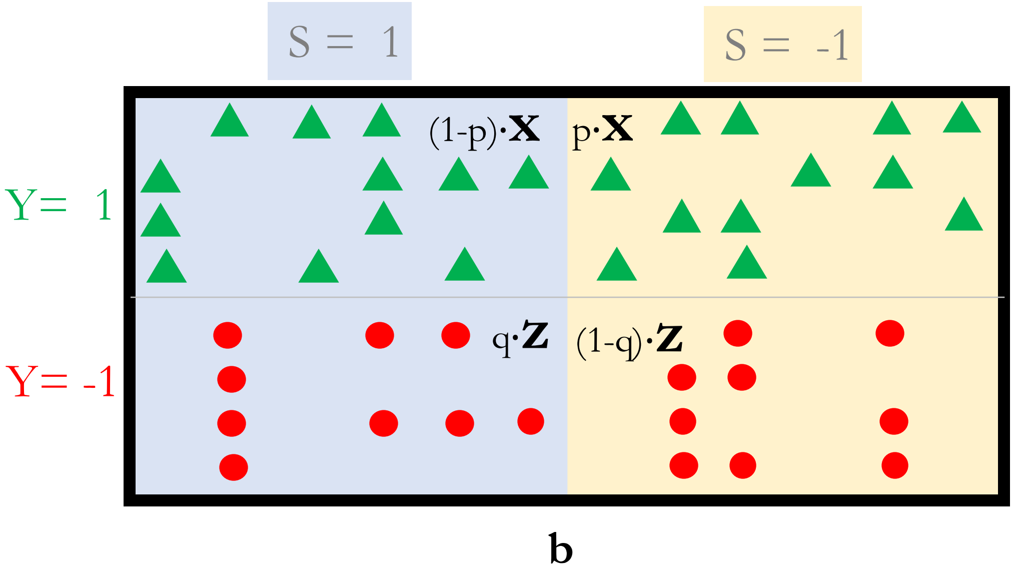
Exercise: What are PPV_b and PPV_w?

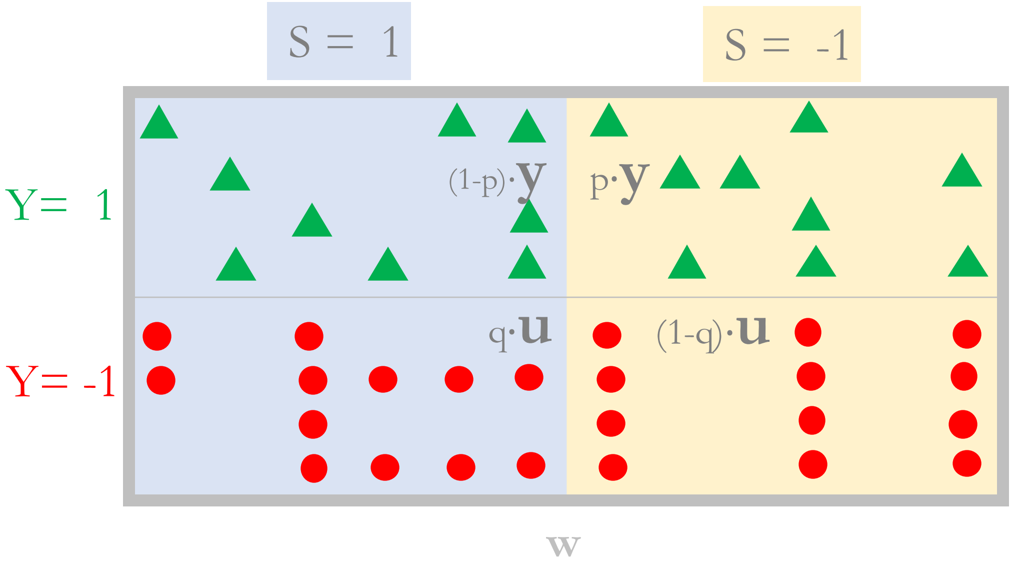




$$PPV_b = \frac{(1-p)\cdot x}{(1-p)\cdot x + q\cdot z} \quad \text{and} \quad PPV_w = \frac{(1-p)\cdot y}{(1-p)\cdot y + q\cdot u}.$$

Exercise: What are p_b and p_w?

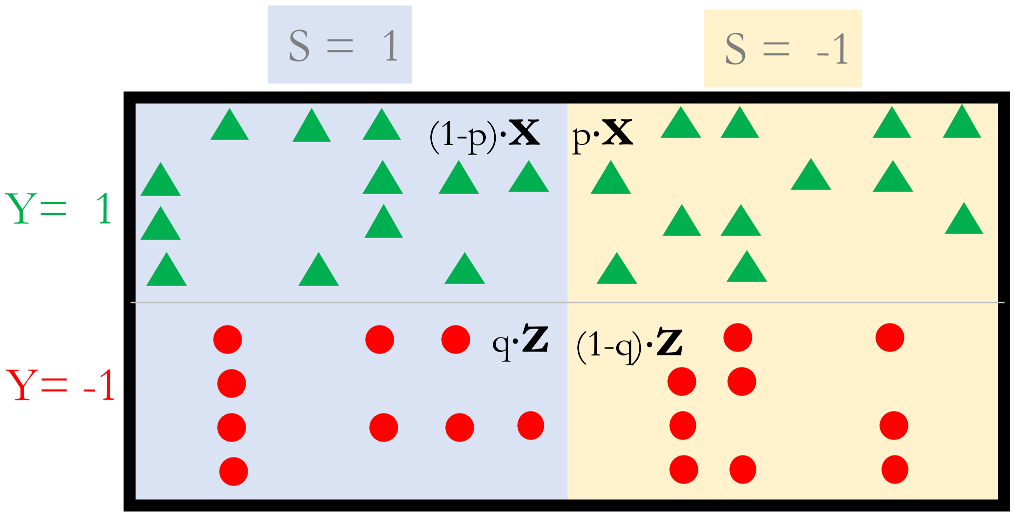




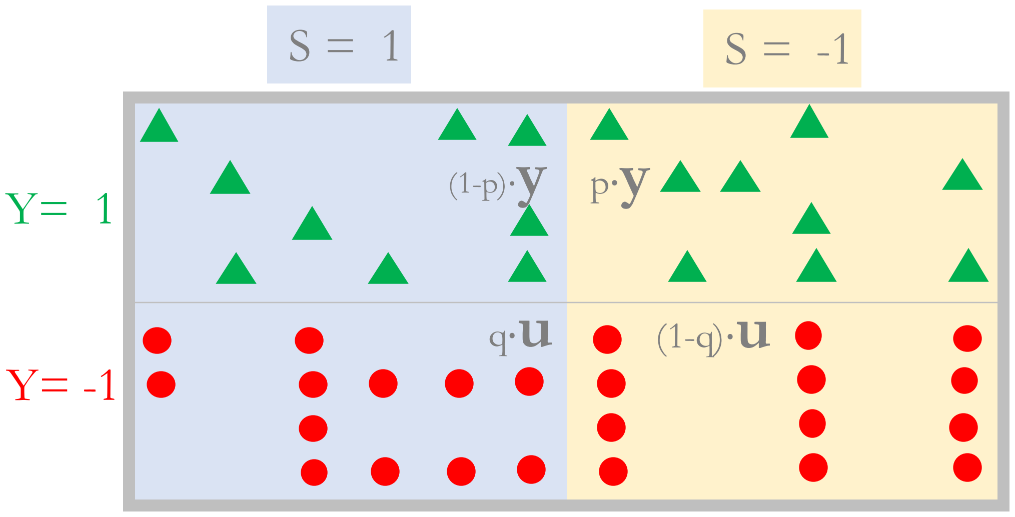
$$p_b = \frac{x}{x+z}$$

$$p_w = \frac{y}{y+u}$$

Exercise: When is $PPV_b = PPV_w$?



b



w

$$p_b = \frac{x}{x+z}$$

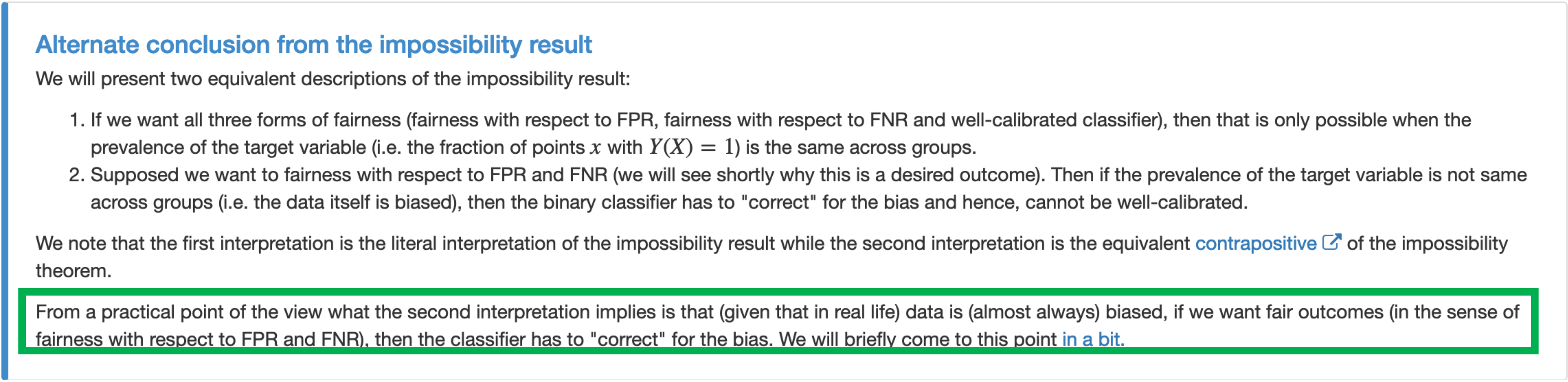
$$PPV_b = \frac{(1-p)\cdot x}{(1-p)\cdot x + q\cdot z} \quad \text{and} \quad PPV_w = \frac{(1-p)\cdot y}{(1-p)\cdot y + q\cdot u}.$$

$$p_w = \frac{y}{y+u}$$

A useful result

Let a, b, c and d be numbers such that $0 < a < b$ and $0 < c < d$. Then the following is true

$$\frac{a}{b} = \frac{c}{d} \text{ if and only if } \frac{a}{b-a} = \frac{c}{d-c} \text{ and } \frac{a}{a+b} = \frac{c}{c+d}.$$



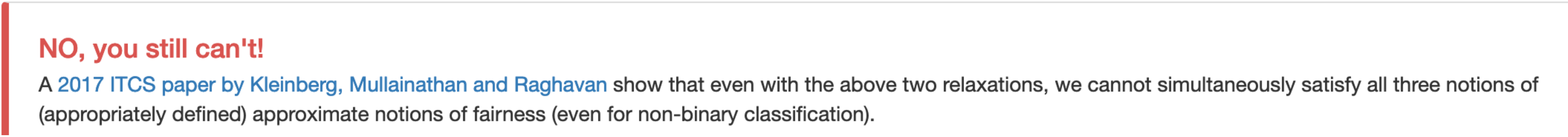
Alternate conclusion from the impossibility result

We will present two equivalent descriptions of the impossibility result:

1. If we want all three forms of fairness (fairness with respect to FPR, fairness with respect to FNR and well-calibrated classifier), then that is only possible when the prevalence of the target variable (i.e. the fraction of points x with $Y(X) = 1$) is the same across groups.
2. Supposed we want to fairness with respect to FPR and FNR (we will see shortly why this is a desired outcome). Then if the prevalence of the target variable is not same across groups (i.e. the data itself is biased), then the binary classifier has to "correct" for the bias and hence, cannot be well-calibrated.

We note that the first interpretation is the literal interpretation of the impossibility result while the second interpretation is the equivalent contrapositive of the impossibility theorem.

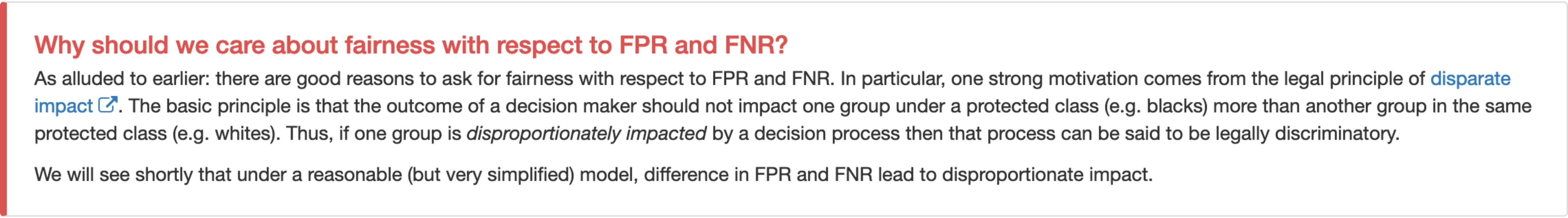
From a practical point of the view what the second interpretation implies is that (given that in real life) data is (almost always) biased, if we want fair outcomes (in the sense of fairness with respect to FPR and FNR), then the classifier has to "correct" for the bias. We will briefly come to this point in a bit.



NO, you still can't!

A 2017 ITCS paper by Kleinberg, Mullainathan and Raghavan show that even with the above two relaxations, we cannot simultaneously satisfy all three notions of (appropriately defined) approximate notions of fairness (even for non-binary classification).

Why should we care fairness wrt FNR and FPR

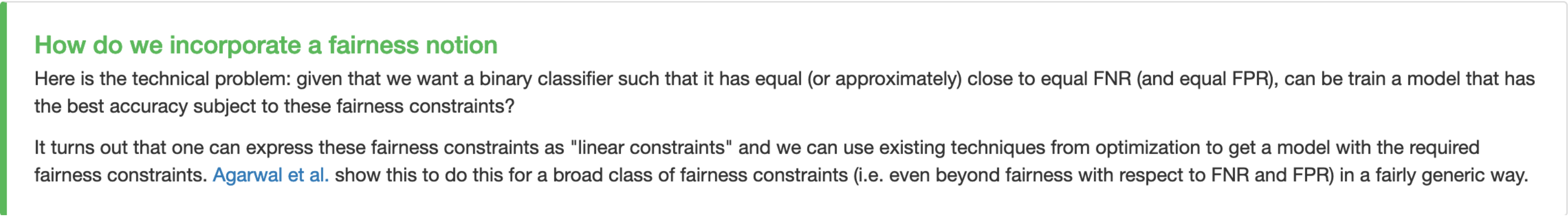


Why should we care about fairness with respect to FPR and FNR?

As alluded to earlier: there are good reasons to ask for fairness with respect to FPR and FNR. In particular, one strong motivation comes from the legal principle of disparate impact. The basic principle is that the outcome of a decision maker should not impact one group under a protected class (e.g. blacks) more than another group in the same protected class (e.g. whites). Thus, if one group is *disproportionately impacted* by a decision process then that process can be said to be legally discriminatory.

We will see shortly that under a reasonable (but very simplified) model, difference in FPR and FNR lead to disproportionate impact.

What if we only care about one fairness def?

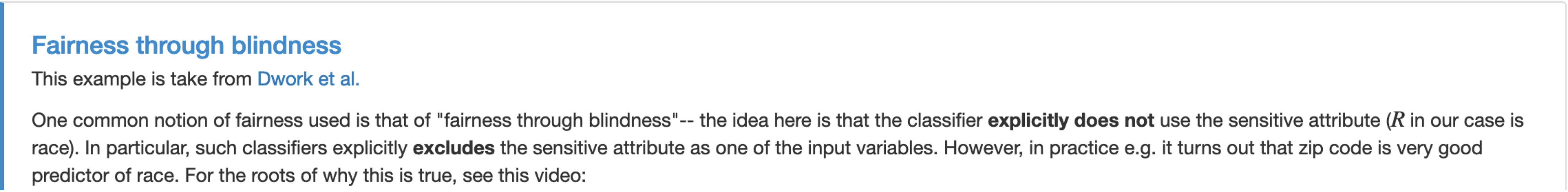


How do we incorporate a fairness notion

Here is the technical problem: given that we want a binary classifier such that it has equal (or approximately) close to equal FNR (and equal FPR), can be train a model that has the best accuracy subject to these fairness constraints?

It turns out that one can express these fairness constraints as "linear constraints" and we can use existing techniques from optimization to get a model with the required fairness constraints. Agarwal et al. show this to do this for a broad class of fairness constraints (i.e. even beyond fairness with respect to FNR and FPR) in a fairly generic way.

Shortcomings of group fairness

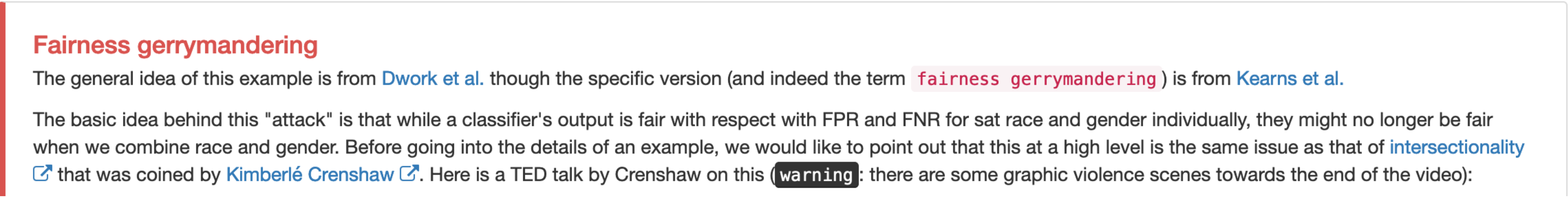


Fairness through blindness

This example is take from Dwork et al.

One common notion of fairness used is that of "fairness through blindness"-- the idea here is that the classifier **explicitly does not** use the sensitive attribute (R in our case is race). In particular, such classifiers explicitly **excludes** the sensitive attribute as one of the input variables. However, in practice e.g. it turns out that zip code is very good predictor of race. For the roots of why this is true, see this video:

Fairness gerrymandering

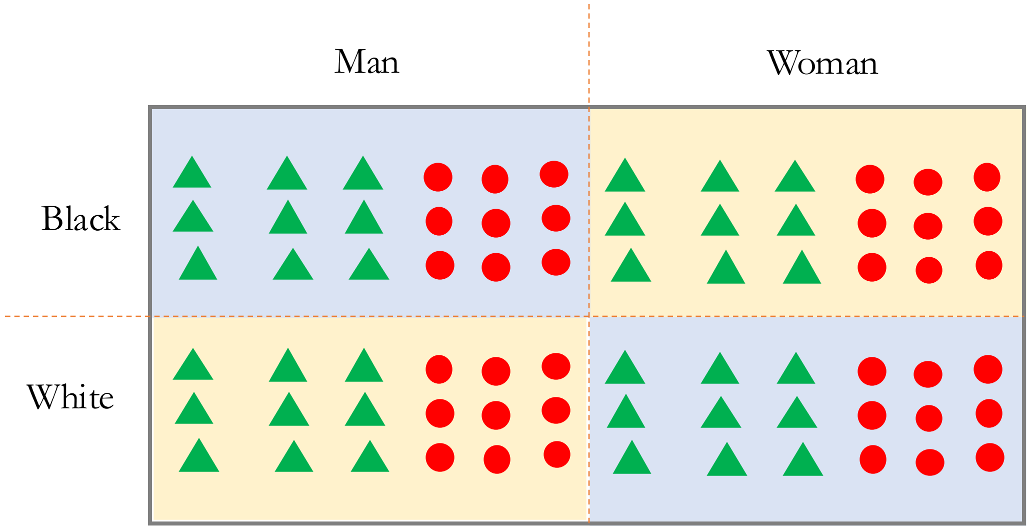


Fairness gerrymandering

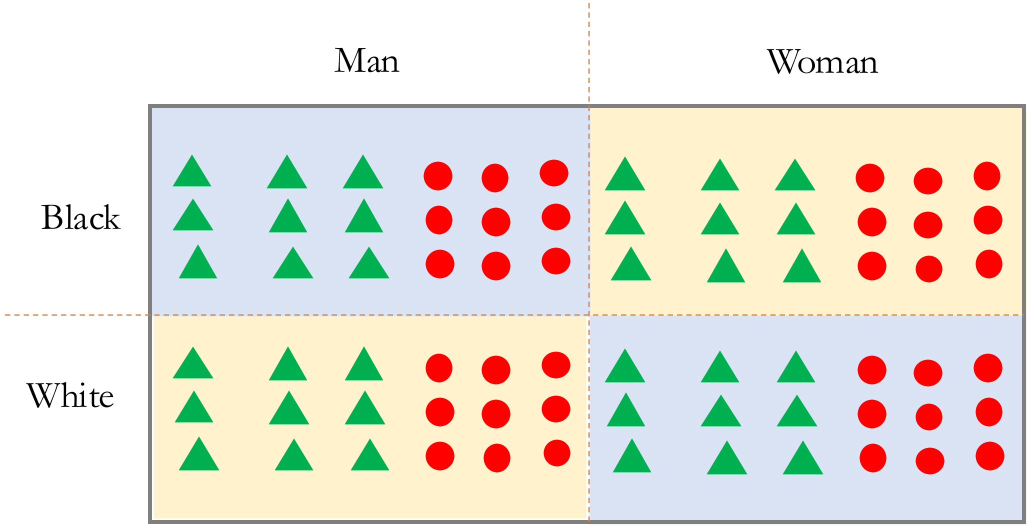
The general idea of this example is from Dwork et al. though the specific version (and indeed the term `fairness gerrymandering`) is from Kearns et al.

The basic idea behind this "attack" is that while a classifier's output is fair with respect with FPR and FNR for sat race and gender individually, they might no longer be fair when we combine race and gender. Before going into the details of an example, we would like to point out that this at a high level is the same issue as that of intersectionality that was coined by Kimberlé Crenshaw. Here is a TED talk by Crenshaw on this (`warning`: there are some graphic violence scenes towards the end of the video):

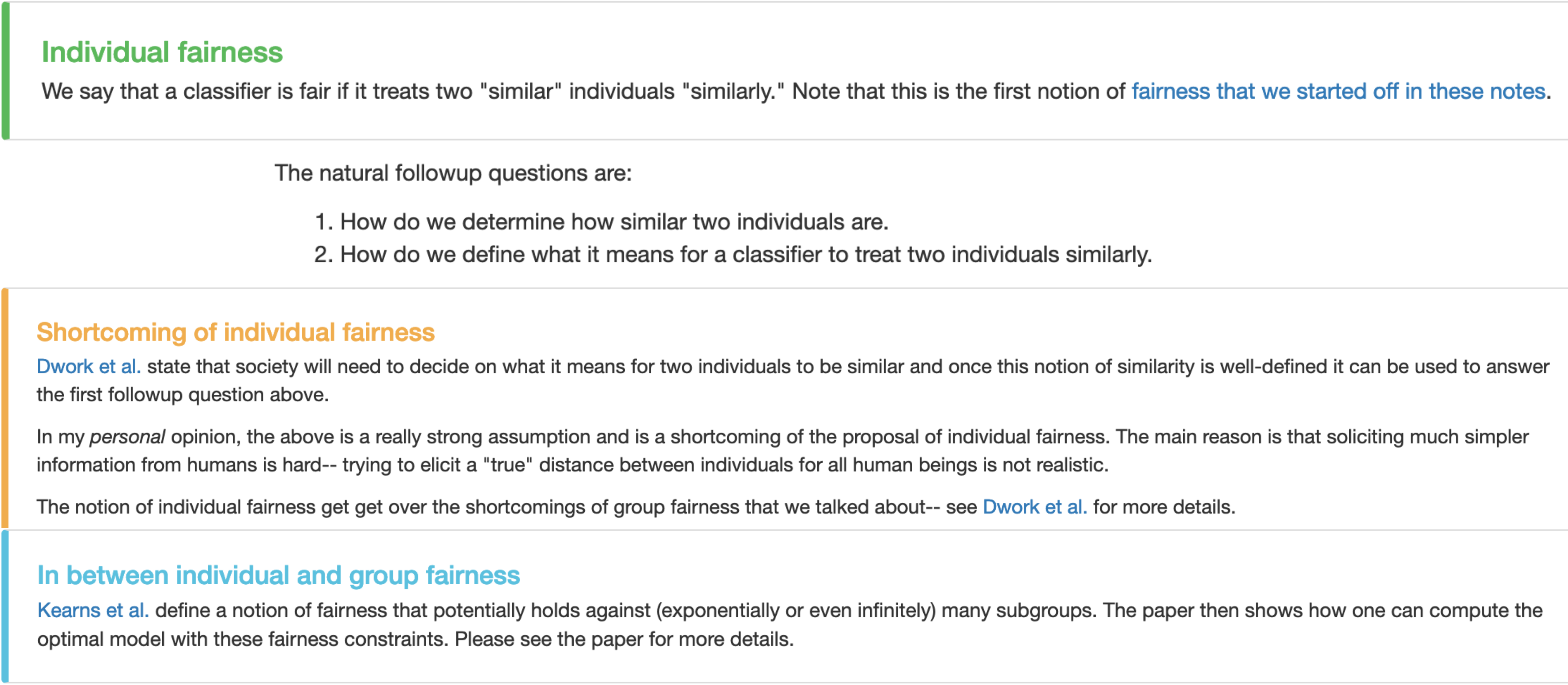
Consider this situation



FNR and FPR of various groups?



Individual Fairness



Individual fairness

We say that a classifier is fair if it treats two "similar" individuals "similarly." Note that this is the first notion of fairness that we started off in these notes.

The natural followup questions are:

1. How do we determine how similar two individuals are.
2. How do we define what it means for a classifier to treat two individuals similarly.

Shortcoming of individual fairness

Dwork et al. state that society will need to decide on what it means for two individuals to be similar and once this notion of similarity is well-defined it can be used to answer the first followup question above.

In my *personal* opinion, the above is a really strong assumption and is a shortcoming of the proposal of individual fairness. The main reason is that soliciting much simpler information from humans is hard-- trying to elicit a "true" distance between individuals for all human beings is not realistic.

The notion of individual fairness get get over the shortcomings of group fairness that we talked about-- see Dwork et al. for more details.

In between individual and group fairness

Kearns et al. define a notion of fairness that potentially holds against (exponentially or even infinitely) many subgroups. The paper then shows how one can compute the optimal model with these fairness constraints. Please see the paper for more details.