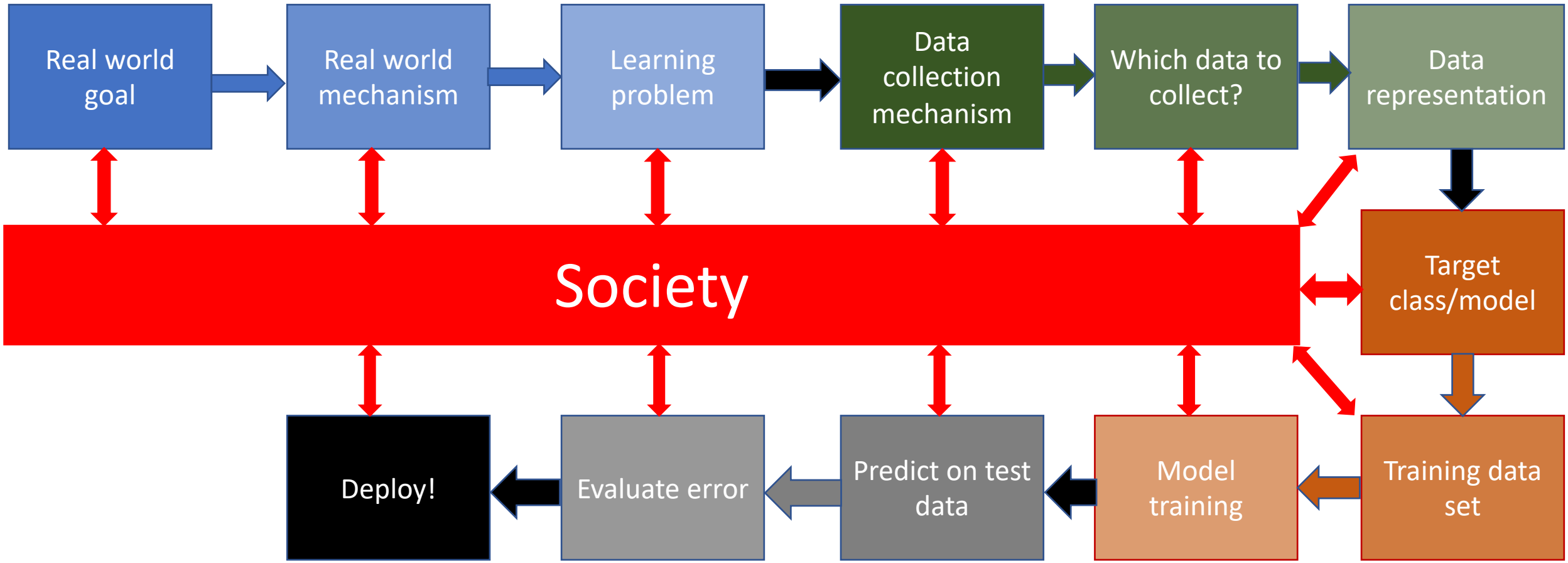
ML and Society

Mar 31, 2022

Done with ML pipeline



Do you remember 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]

1 Risk Assessment
2 Critiques and legal rulings
3 Accuracy
4 Further reading
5 See also
6 References

Risk Assessment [edit]

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

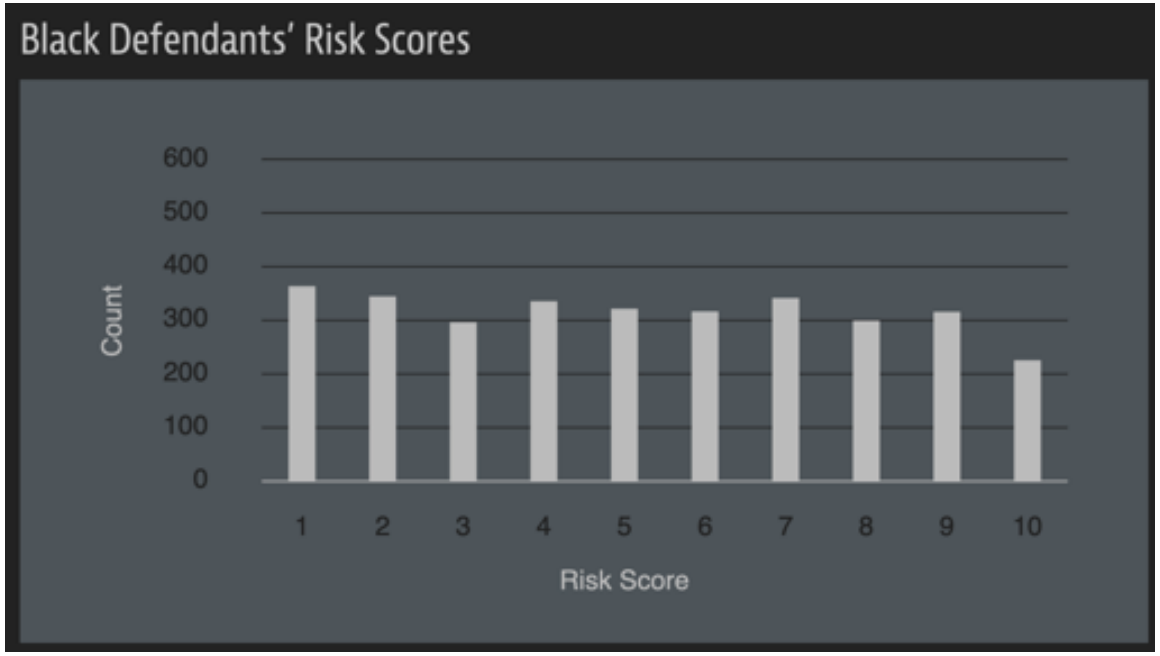
Machine Bias

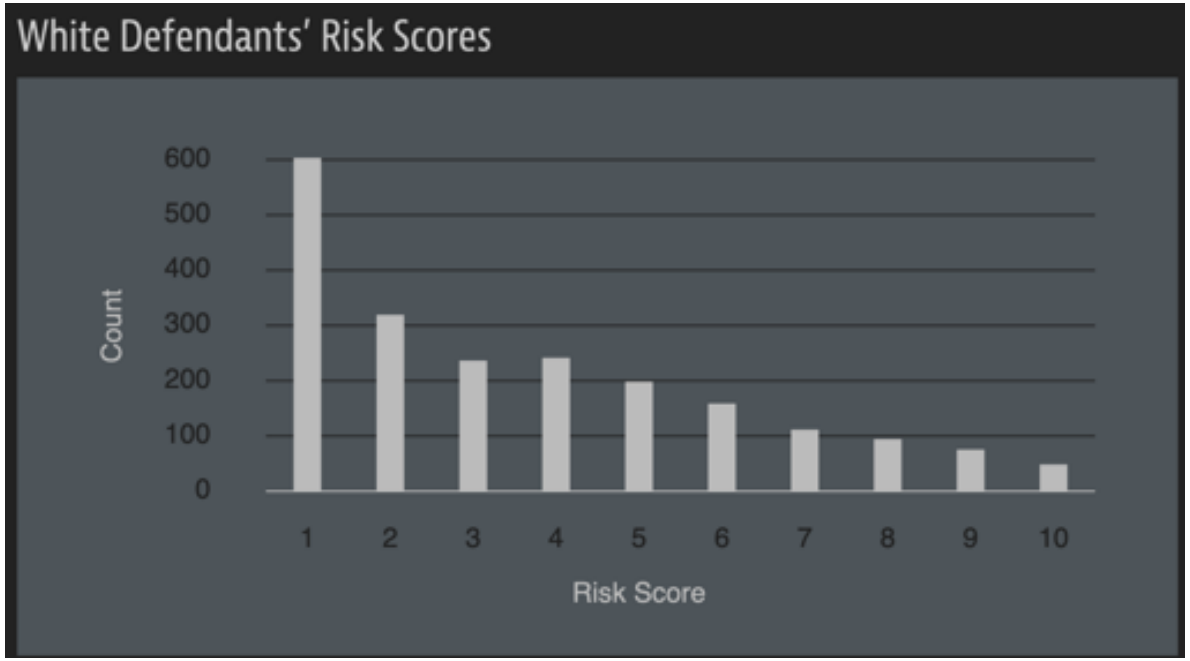
There's software used across the country to predict future criminals. And it's biased against blacks.

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

What is fair? Take-1

Individual Fairness

What is fair? Take-2

Equality/
Group
Fairness

Very limited coverage

Book (in progress) on fairness and ML

If you are interested in more details, we refer you to the book by Barocas, Hardt and Narayanan on this topic. As of the time of writing of these notes, this book is still not complete but it does contain a lot of the known material that we will not cover in this course.

Will we get **the** definition of fairness?

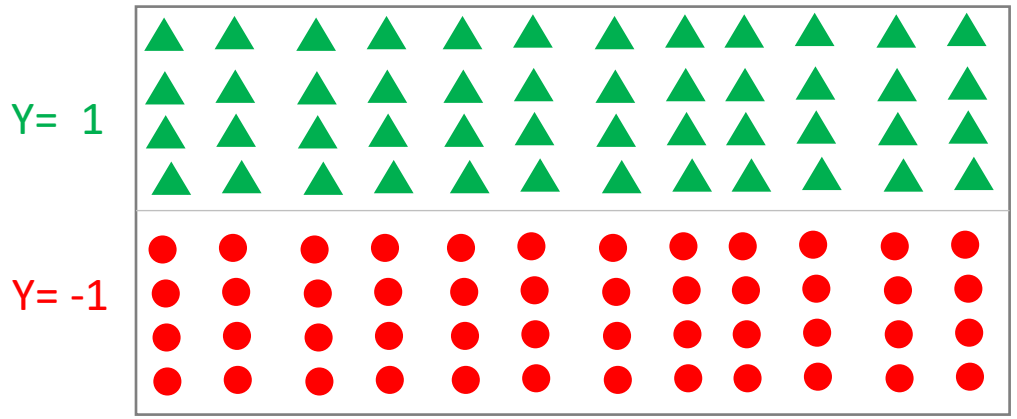
What about bias?

What is bias?

Another loaded term that we will use is the term `bias`. In particular, there are roughly three kinds of notions of bias that is relevant to these notes:

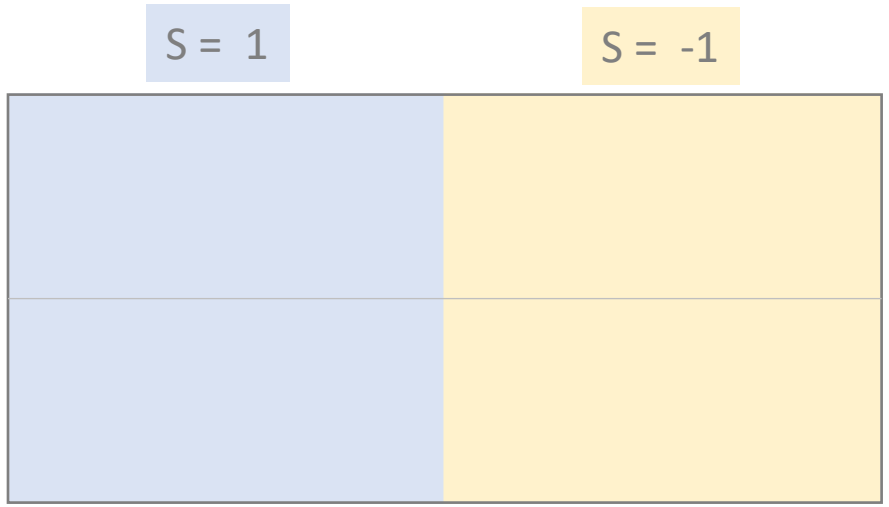
1. The first notion (which might be the least known) occurs in a dataset where there are certain specific collection of input variable values occur more than others. This essentially measure how far away from a truly random dataset the given dataset is. Note that this notion is bias is **necessary** for ML to work. If all the datapoints are completely random (i.e. both their input and target variable values are completely random), then there is no bias for a classifier to "exploit"-- in other words, one might as well just output a random label for prediction.
2. The second notion of bias is that of statistical bias, where in our setting this would mean that the binary classifier outcome does not reflect the distribution of the underlying target variable. Such a classifier would be well calibrated, if this does not happen. One could consider a well-calibrated binary classifier to be fair in some sense. This will be one notion of fairness that will come up in the COMPAS story. (This is the notion of fairness used in the rejoinder to the ProPublica article).
3. The finally notion of bias is the colloquial use of the term that is mean to denote an outcome that is **not fair**. Most of the definitions of fairness in the literature deal with this notion of bias. And a couple of definition of this kind of fairness will also play a part in the COMPAS story (this is the notion of fairness used in the ProPublica article).

Going beyond # correctly classified points

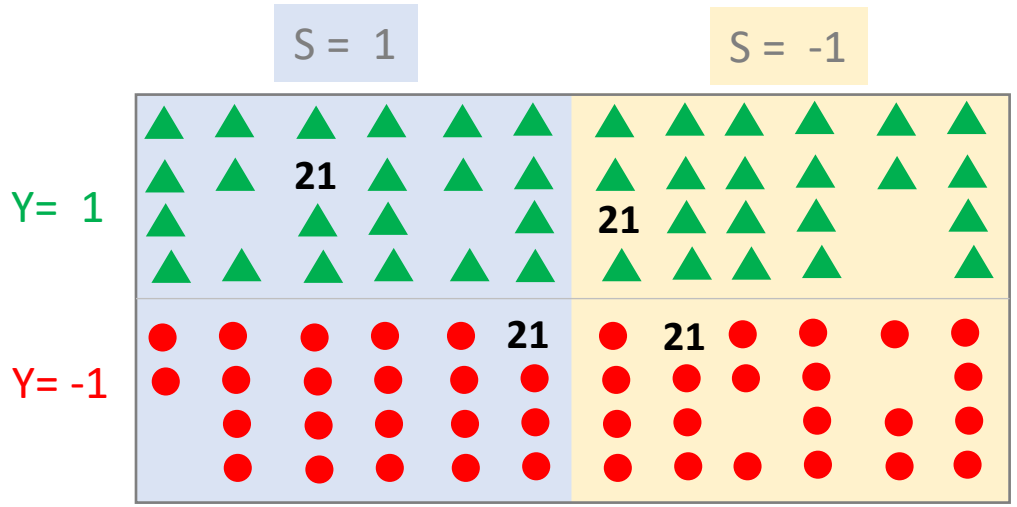


Points are NOT 2D!

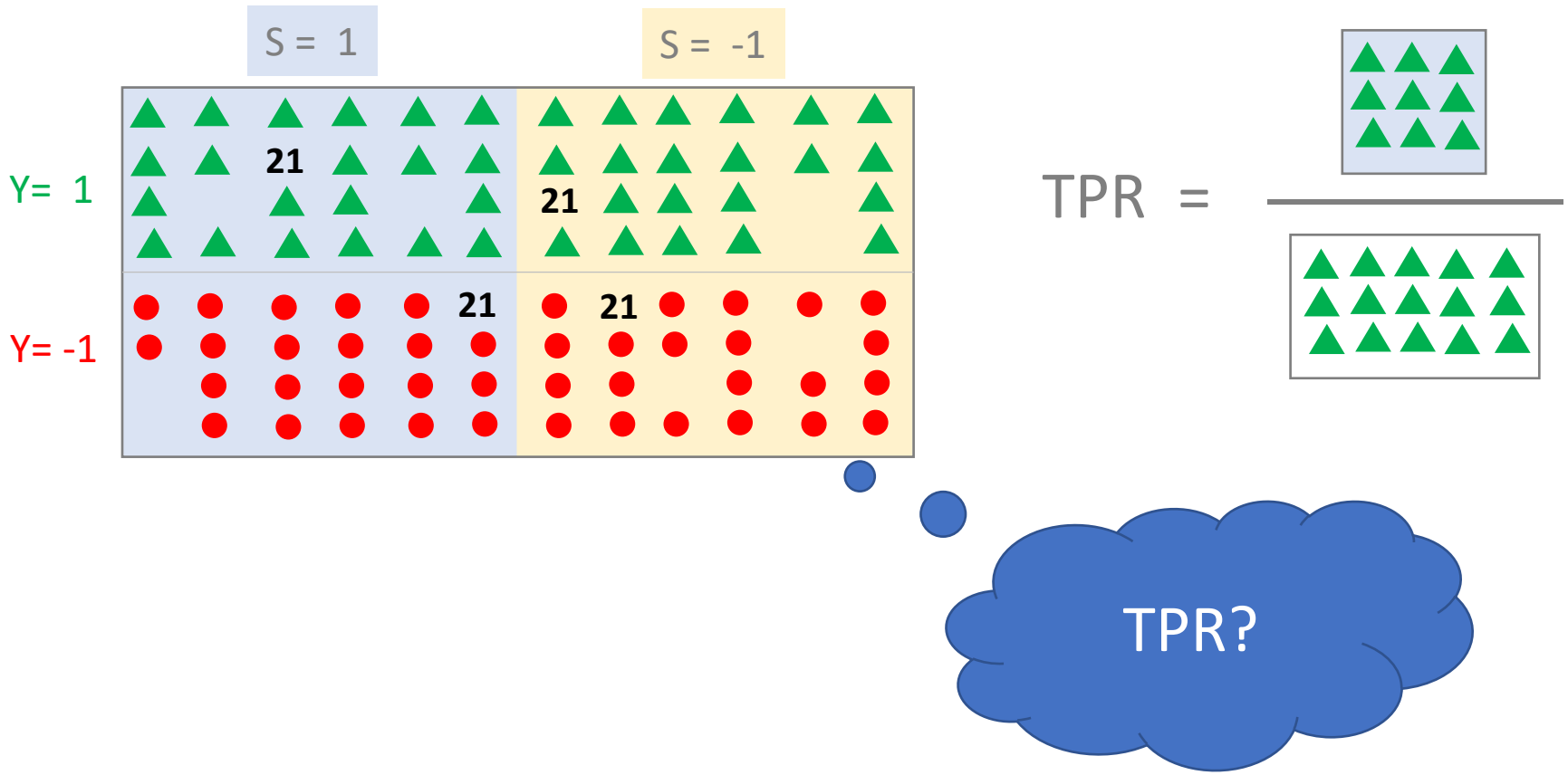
Binary classifier output



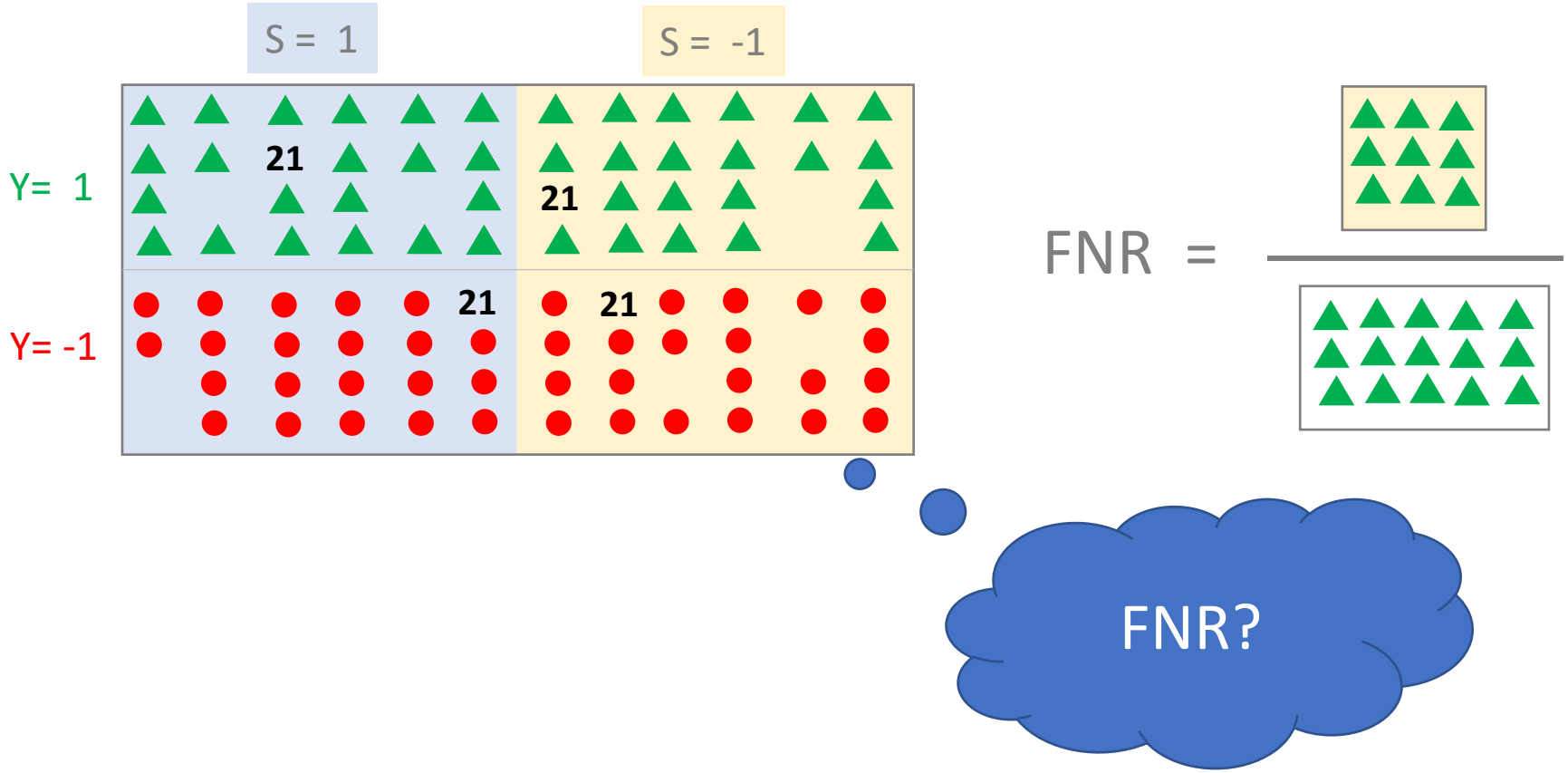
“Confusion matrix”



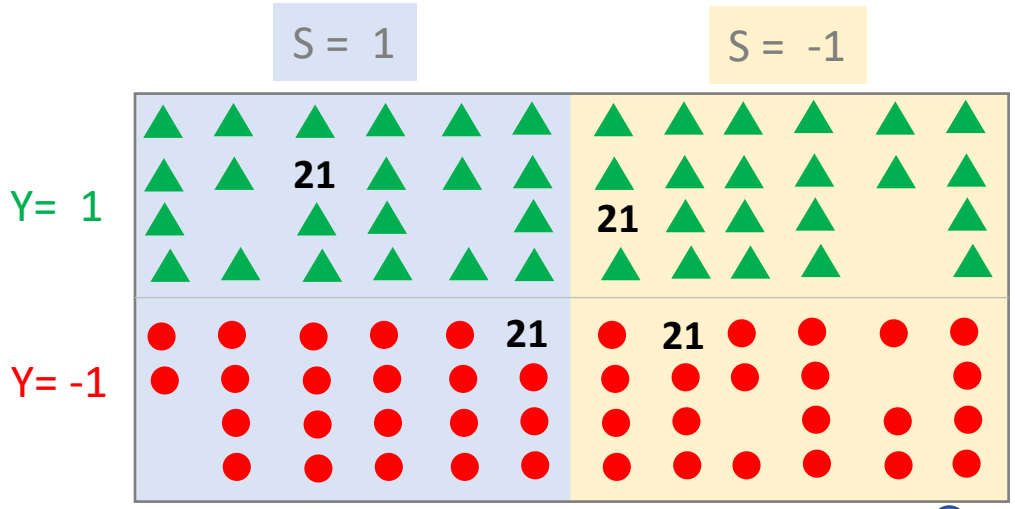
True Positive rate



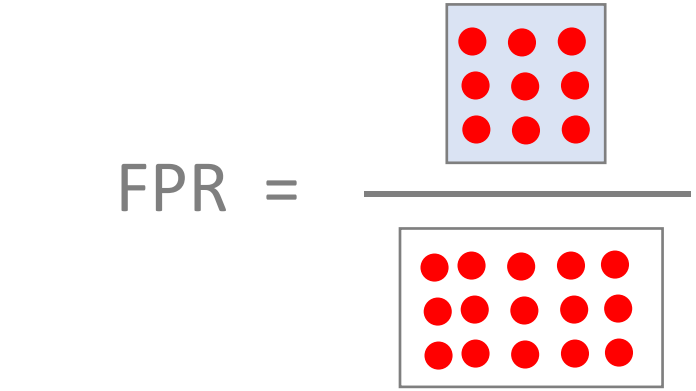
False Negative Rate (FNR)



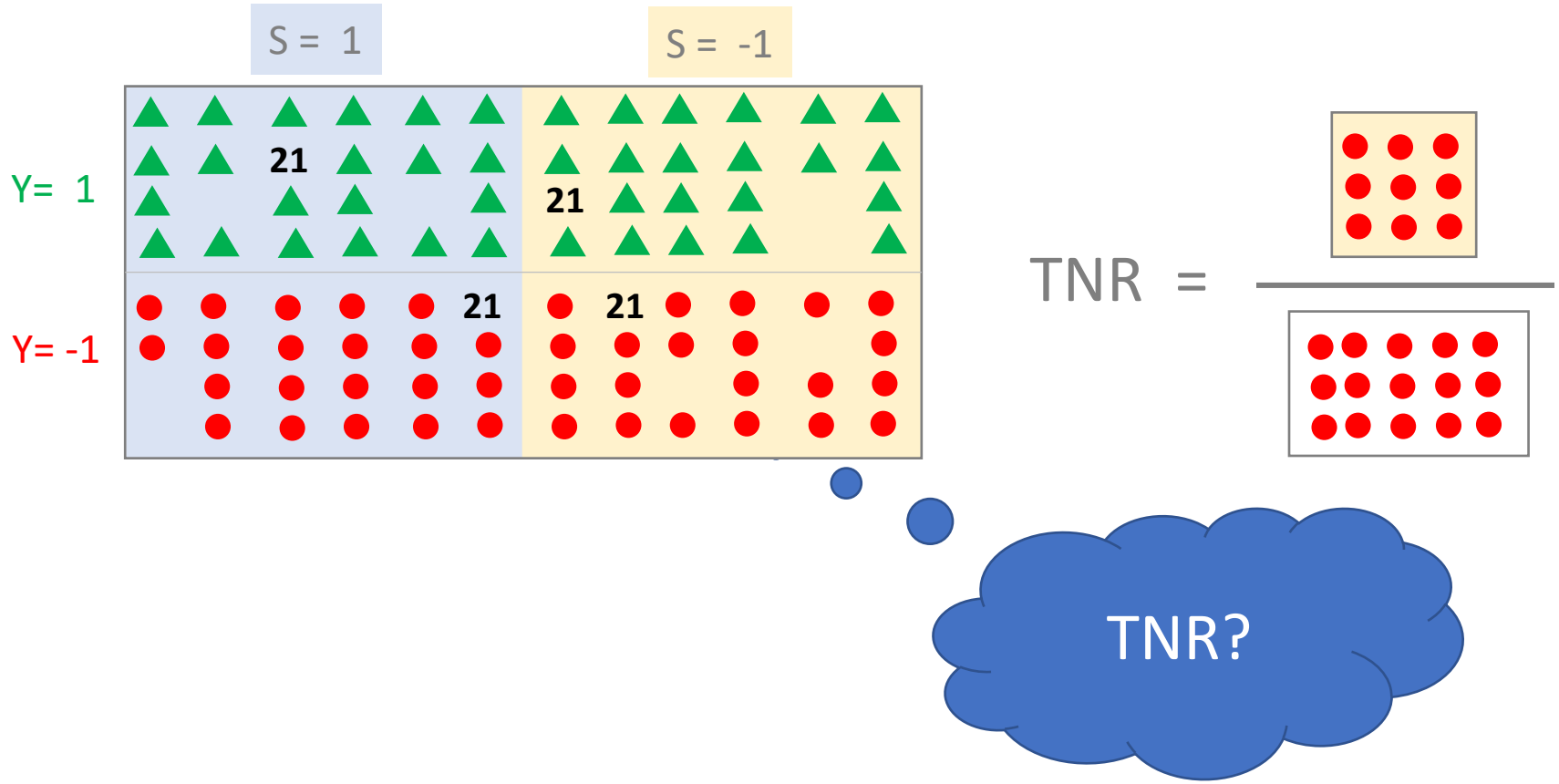
False Positive Rate (FPR)



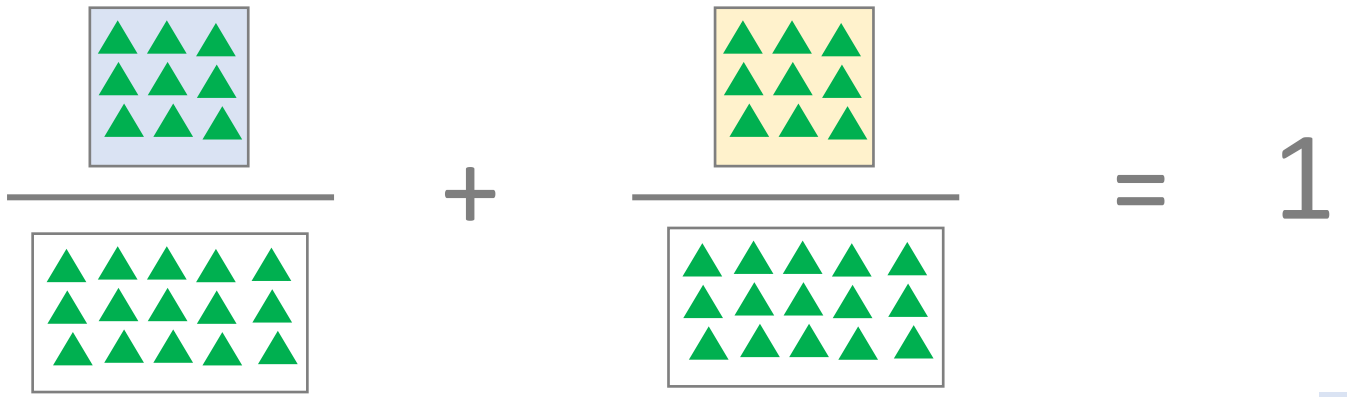
FPR =

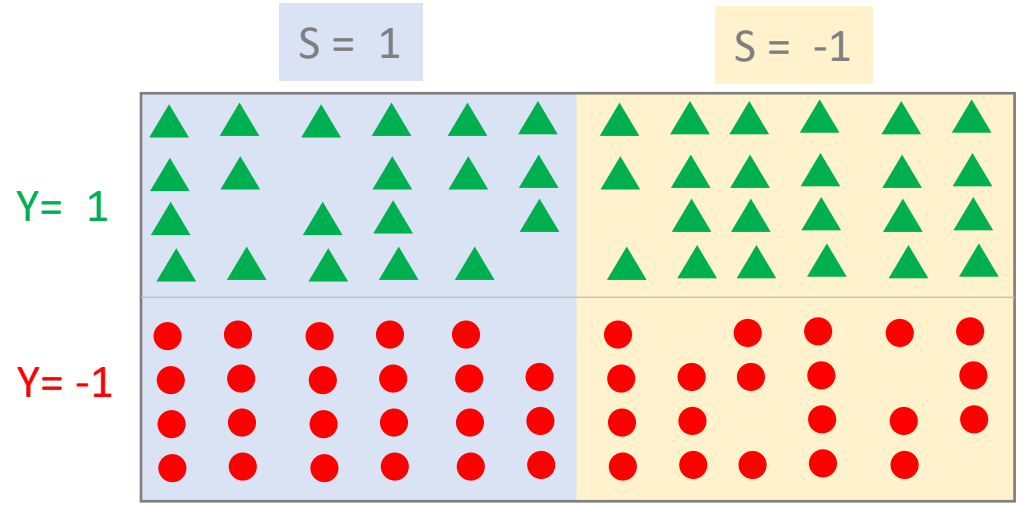


True Negative Rate (TNR)

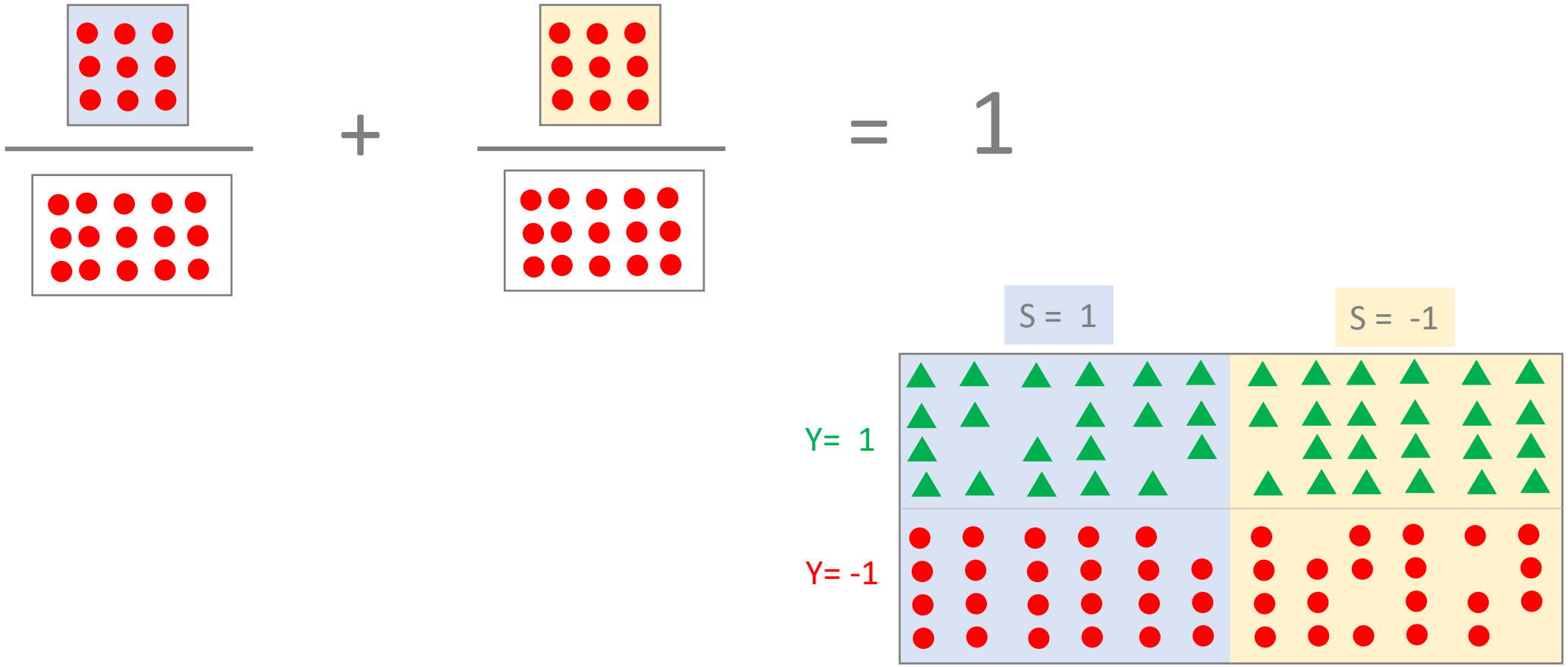


TPR + FNR = 1

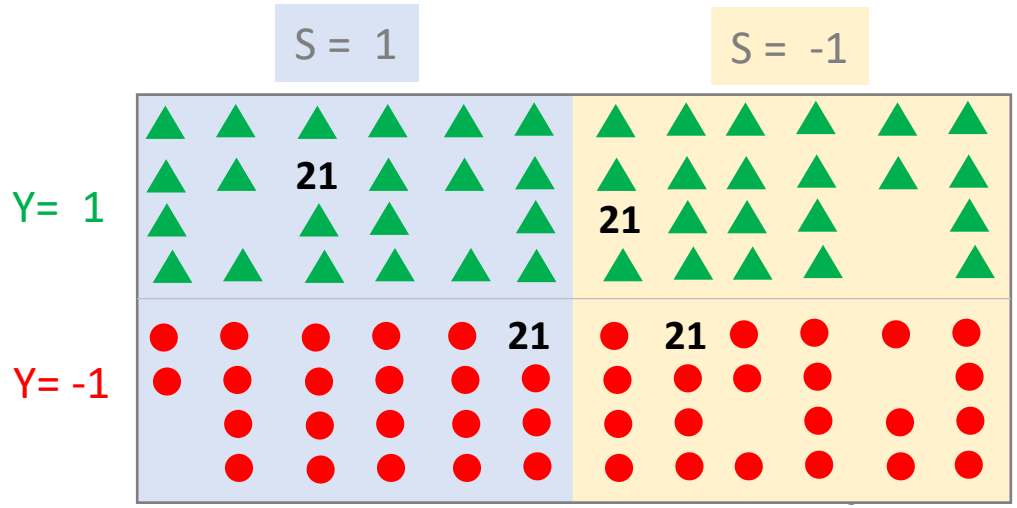




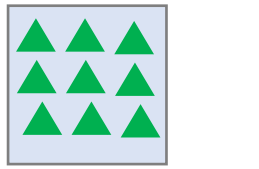
FPR + TNR = 1

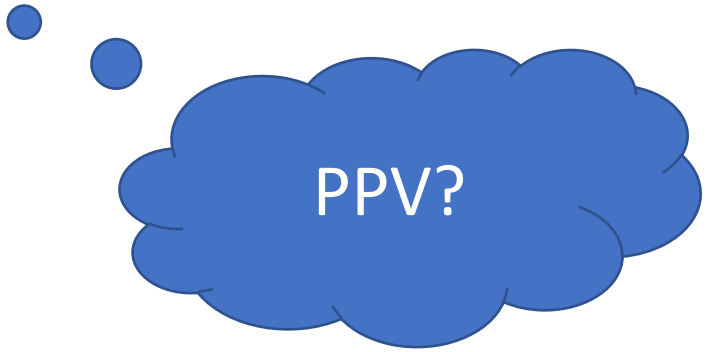


Positive Predictive Value (PPV)



PPV =





Back to fairness

Protected/Sensitive attribute

To define **group** fairness, we have to well, define a *group* first. Towards this, we will use the notion of a `protected attribute` or `sensitive attribute` (we will use both terminology interchangeably): this will be a special attribute R (which takes few pre-defined values i.e. is a categorical variable)-- and each choice of the value of R defines a separate group. There is precedence in US law: grouping this way is used in the concept of protected class in US anti-discrimination law-- i.e. one cannot discriminate on the basis of any protected class.

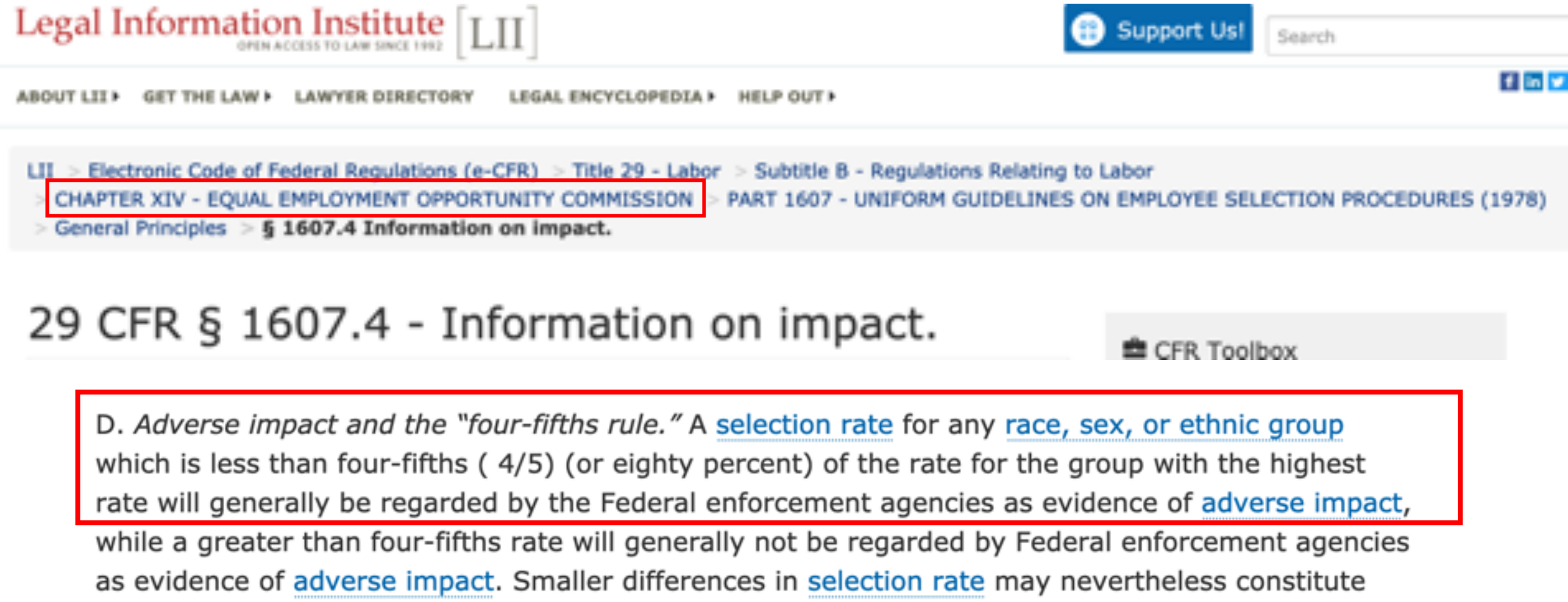
Coming back to the COMPAS example, we will use R to denote the race and for simplicity we will assume the two values R can take are b (for *black*) and w (for *white*). While clearly these are not the only racial classification, the results of ProPublica mentioned earlier focus on these two value of race and hence we concentrate on these two possibilities.

For the rest of the section, we will **only consider groups corresponding to $R(x) = b$ and $R(x) = w$** (i.e. groups based on whether race of x is black or white).

Statistical parity

At a high level we would like the accuracy of binary classifier to be the same across groups. Since in real life false positive positives and false negatives have different costs, various instantiation of statistical parity definitions follows by asking that different notions of accuracy be the same across groups.

Why statistical parity across groups?



Notes on ML and law

Discrimination, Law and ML

This page will do a quick overview of anti-discrimination law and how it could/would interact with the ML pipeline.

Under Construction

This page is still under construction. In particular, nothing here is final while this sign still remains here.

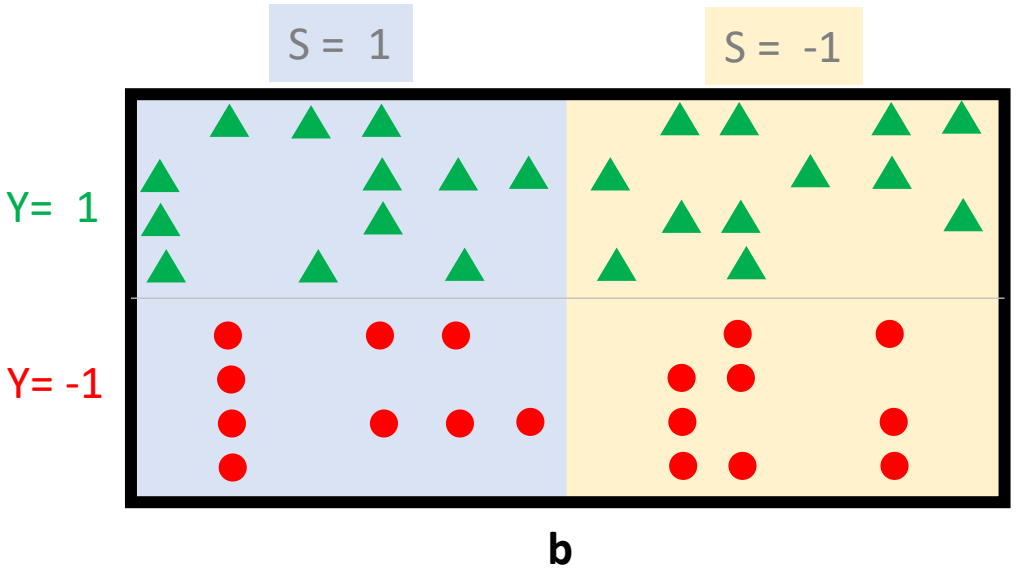
A Request

I know I am biased in favor of references that appear in the computer science literature. If you think I am missing a relevant reference (outside or even within CS), please email it to me.

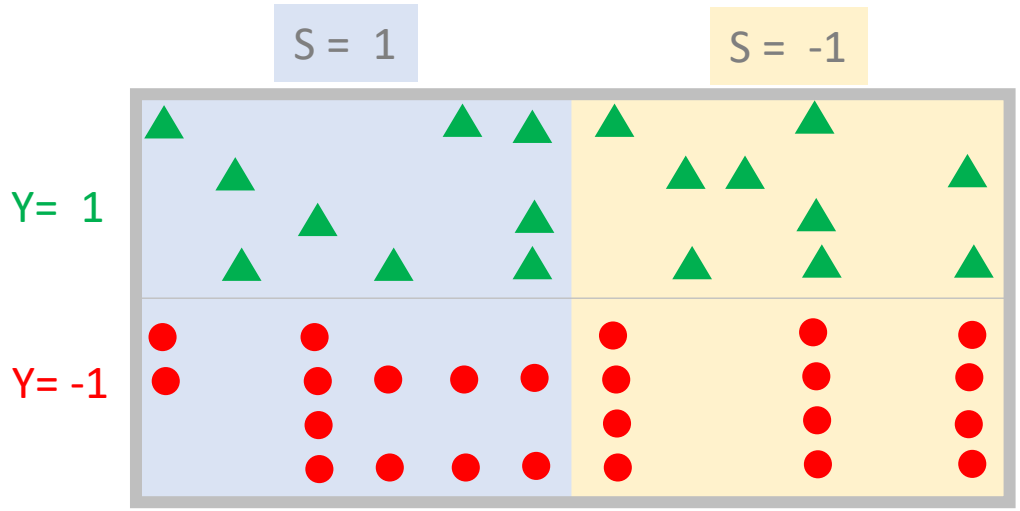
Anti-discrimination law

In this section, we will review anti-discrimination law as part of Title VII of the Civil rights act of 1964.

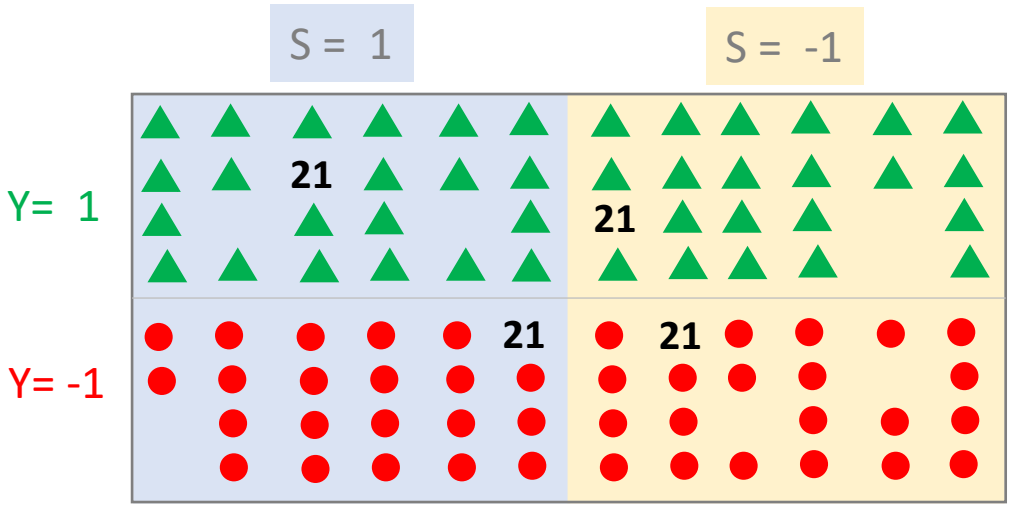
Rates for groups



b

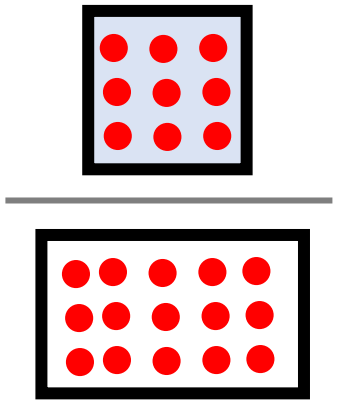


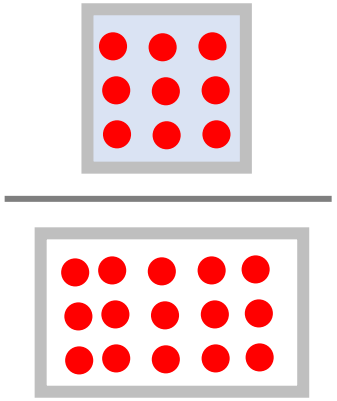
w

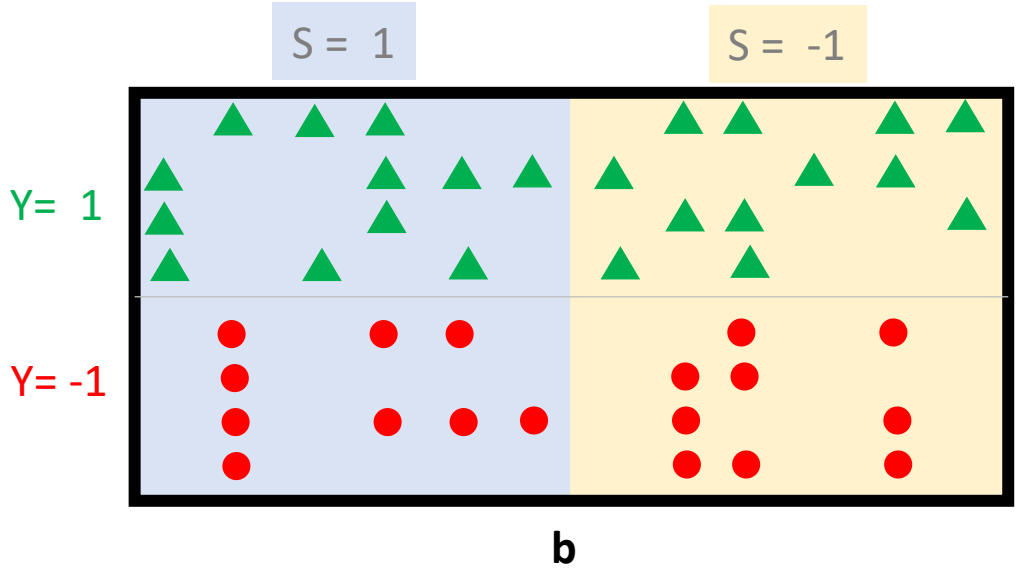


FPR and FNR for groups

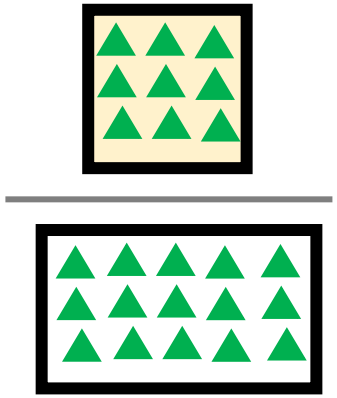
Calculate the rates

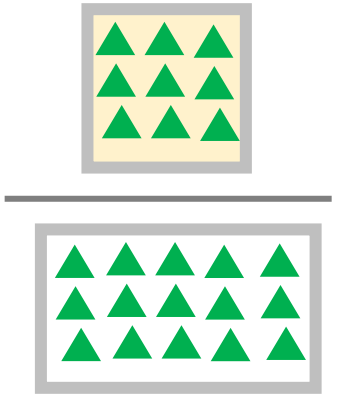
$FPR_b =$ 

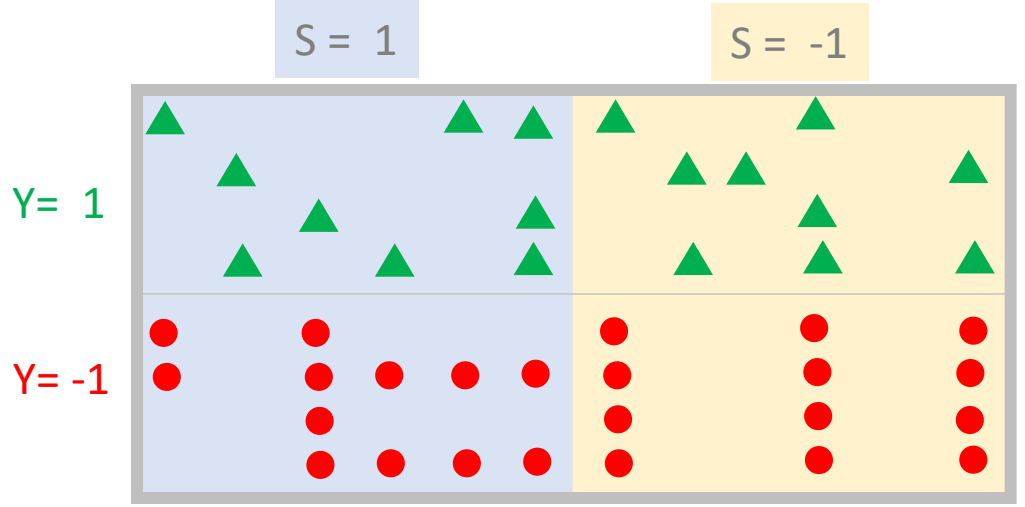
$FPR_w =$ 



b

$FNR_b =$ 

$FNR_w =$ 



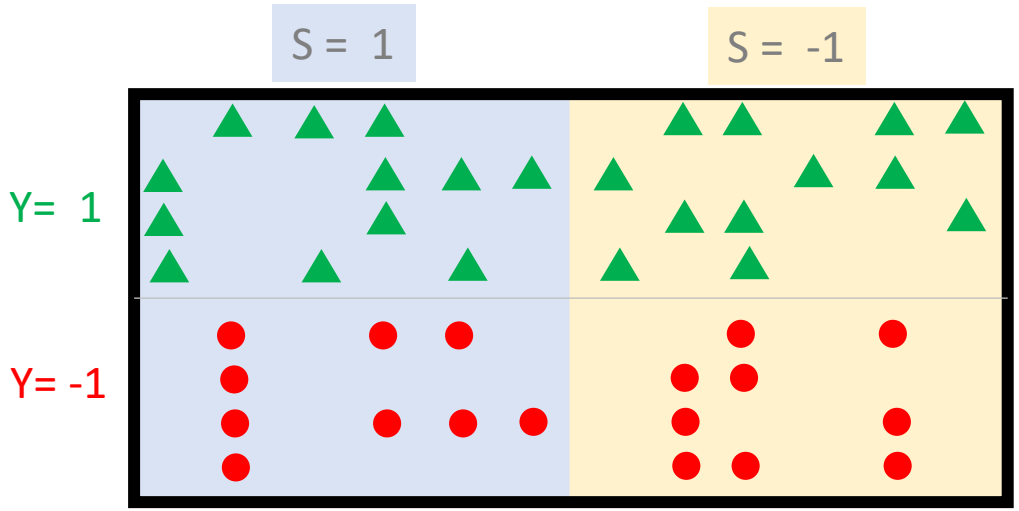
w

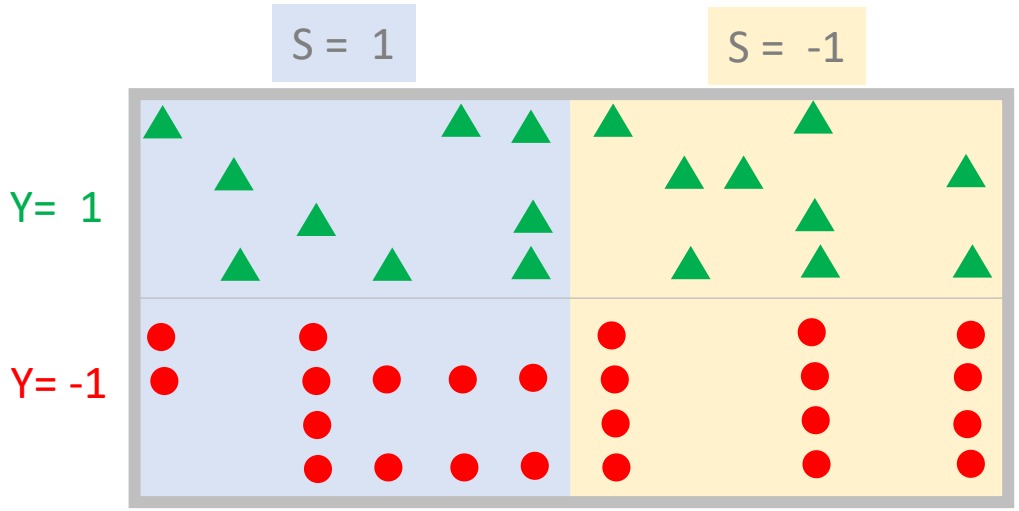
PPV for groups

Calculate the values

$PPV_b = \frac{\text{▲▲▲ ▲▲▲ ▲▲▲}}{\text{▭}}$

$PPV_w = \frac{\text{▲▲▲ ▲▲▲ ▲▲▲}}{\text{▭}}$





Finally, the formal 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.

Why no Equal TPR and TNR?

You might be wondering why we did not define notions of fairness with respect to TPR and TNR? The answer is that because we already have! Do you see why?

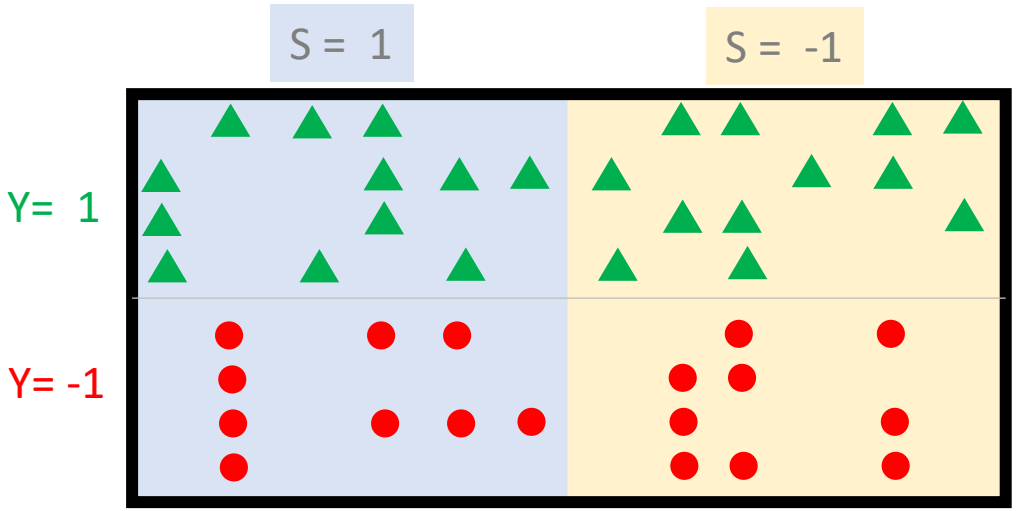
`Hint`: We claim those follow from fairness with respect to FNR and FPR respectively.

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.

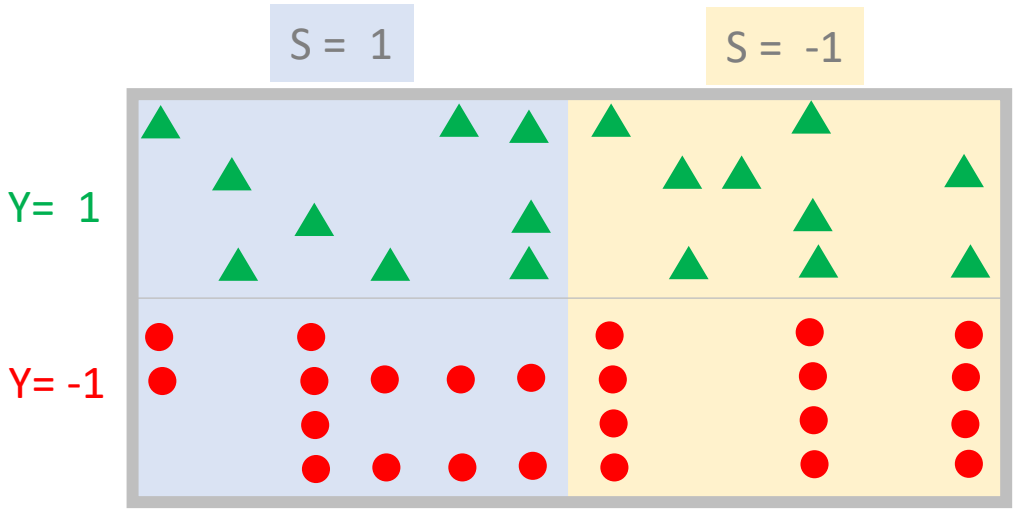
Exercise!

Exercise

For each notion of being fair with respect to FPR, FNR and well-calibrated, decide if it holds for the following instance (that we have seen before):



b



w

Passphrase for today: danah boyd

danah boyd

output:

- danah's blog
- twitter feed
- papers, articles, talks, etc.
- Data & Society
- newsletter: substack

a few key papers:

- "Democracy's Data Infrastructure: The Technopolitics of the US Census" (with Dan Bouk)
- "Data Voids" (with Michael Golebiewski)
- "Situating Methods in the Magic of Big Data and AI" (with M.C. Elish)
- "The Drama! Teen Conflict in Networked Publics" (with Alice Marwick)
- "Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies" (with Alice Marwick)

My name is danah boyd and I'm a Partner Researcher at **Microsoft Research** and the founder of **Data & Society**. I'm also a Distinguished Visiting Professor at Georgetown University and a Visiting Professor at **New York University's Interactive Telecommunications Program**. I am an academic and a scholar and my research examines the intersection between technology and society.

For over a decade, my research focused on how young people use social media as part of their everyday practices. I wrote ***It's Complicated: The Social Lives of Networked Teens*** (2014) to document my findings.

I also co-authored two books - ***Hanging Out, Messing Around, and Geeking Out: Kids Living and Learning with New Media*** (2009) and ***Participatory Culture in a Networked Era*** (2015) to highlight different aspects of that work.

More recently, I have turned to focus on understanding how contemporary social inequities relate to technology and society more generally. My current work centers on what makes data legitimate, based on fieldwork I'm doing around the 2020 US census. At **Data & Society**, a research institute that I founded, I'm collaborating with an amazing network of researchers working on topics like media manipulation, the future of work, fairness and accountability in machine learning, combating bias in data, and the cultural dynamics surrounding artificial intelligence.

Over the years, I have written many **papers** on topics related to media manipulation, algorithmic fairness, social media, privacy, teen drama/bullying, digital backchannels, and social visualization design. I also blog and **tweet** frequently on a wide variety of topics.

In 2008, I completed my PhD at the **School of Information (iSchool)** at the University of California-Berkeley. My **dissertation research** was funded as a part of the **MacArthur Foundation's** initiative on New Media and Learning. My research was supervised by a most astonishing committee: **Mimi Ito**, **Annalee Saxenian**, **Cori Hayden**, and **Jenna Burrell**. My beloved PhD advisor - **Peter Lyman** - lost his battle with brain cancer in July 2007. I miss him dreadfully.

I did my Master's Degree at the MIT Media Lab's **Sociable Media Group** with **Judith Donath** (supervised also by **Henry Jenkins** and Geneviève Bell). My **master's thesis** focused on how people manage their presentation of self in relation to social contextual information in online environments. As an undergraduate, I studied computer science at Brown University, advised by **Andy van Dam**. My **undergrad thesis** focused on how prioritization of depth cues is dependent on levels of sex hormones in the body and how this affects engagement with virtual reality.

Outside of academia, I have worked at various non-profits and corporations. I'm on the boards the **Social Science Research Council** and **Crisis Text Line**, which uses technology to serve people in the midst of a mental health crisis. I'm a former Trustee of the **National Museum of the American Indian**. For five years, I worked at **V-Day**, an organization working to end violence against women and girls worldwide. I helped build an online community to support activists. **For a complete bio, click here.**

On the web, I'm known for two things: maintaining an **Anil Dil'manca lyrics site** and **blogging prolifically**. Personally, I love music, dancing, hiking, reading, and all things fuzzy. At my core, I'm both an advocate and a scholar. I'm also a parent to three kids.

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+: 1.54			LR+: 1.61			LR+: 2.23		
LR-: 0.55			LR-: 0.57			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!

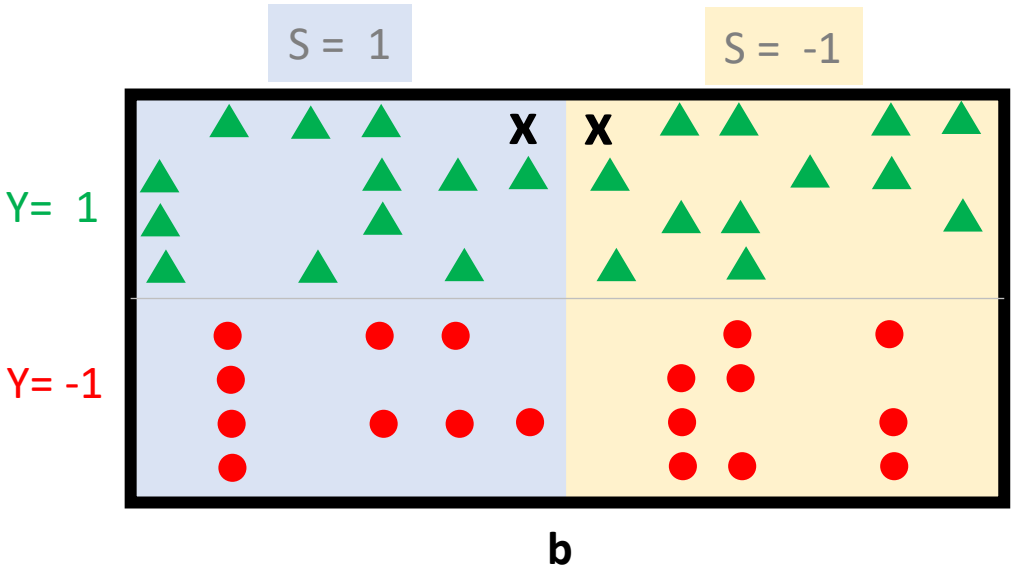
it is **impossible** for a binary classifier to satisfy **all three notions of fairness** (i.e. fairness with respect to FPR, FNR and being well-calibrated) *unless the fraction of positives to the overall number of points is the same in both groups*.

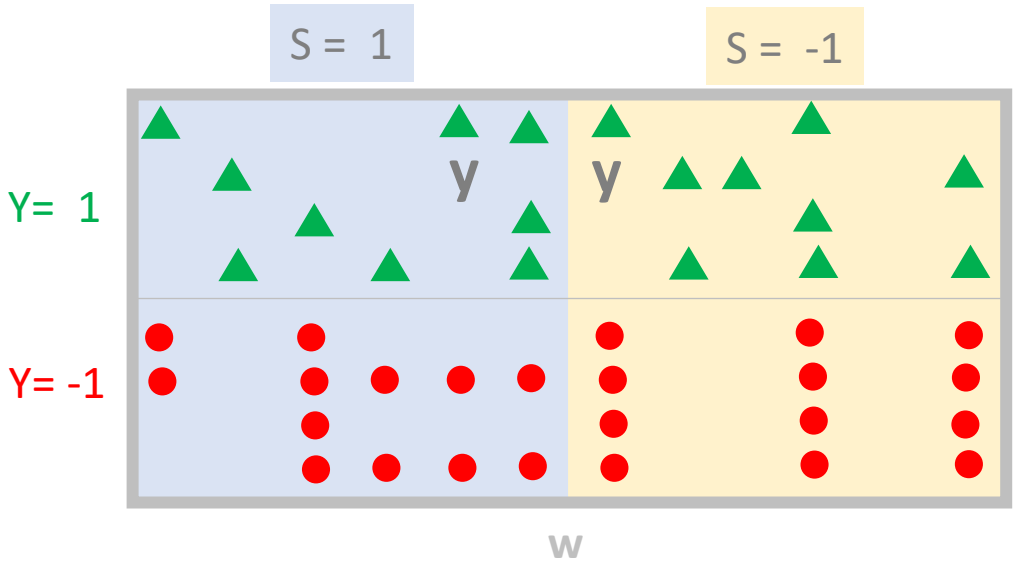
In the COMPAS dataset, the recidivism rate for blacks and whites are 50% and 39% respectively. Hence, the fact that COMPAS could not satisfy all three notions of fairness, is *mathematically unavoidable*.

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.

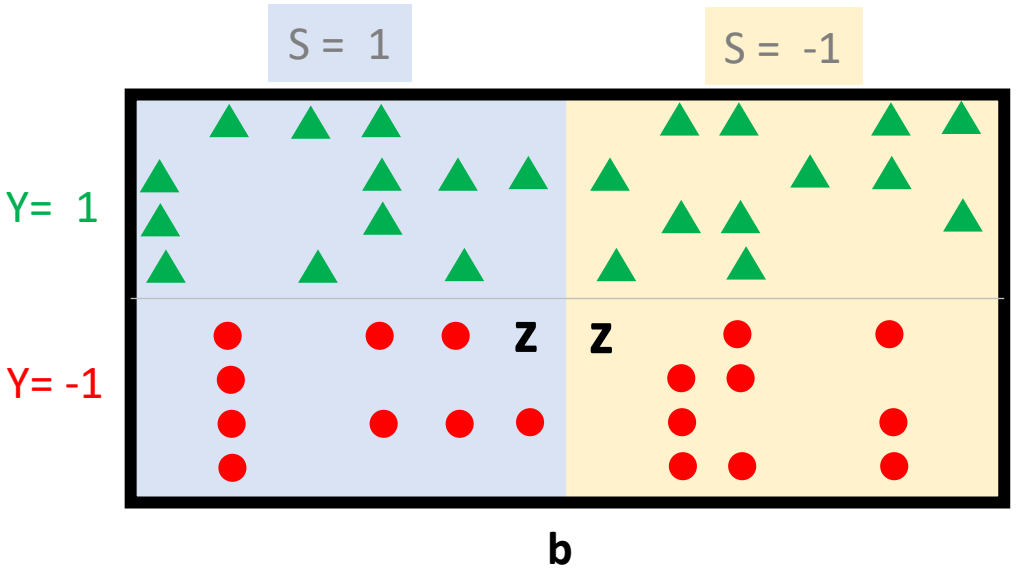


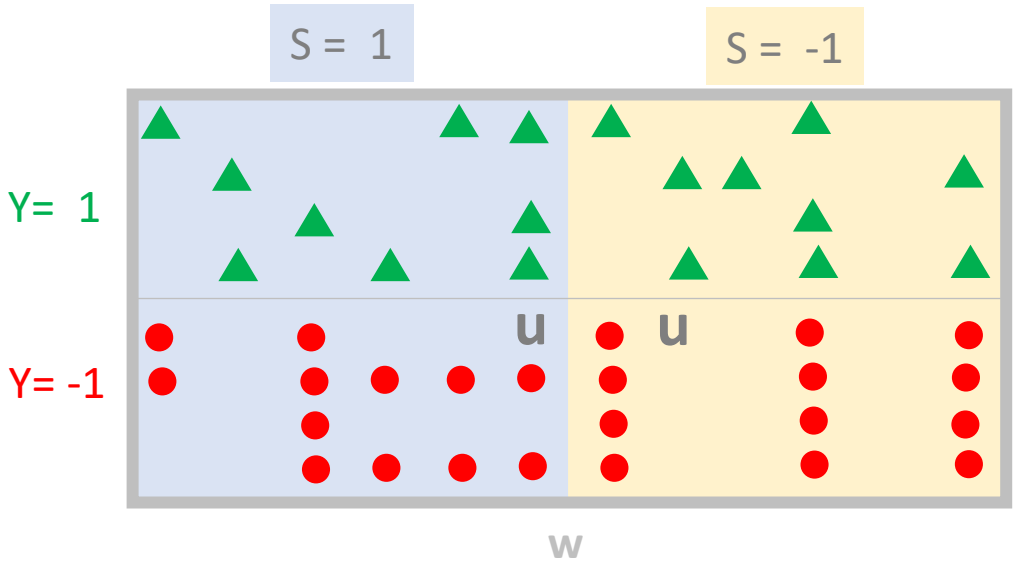

$FNR_b = FNR_w = 1/2$





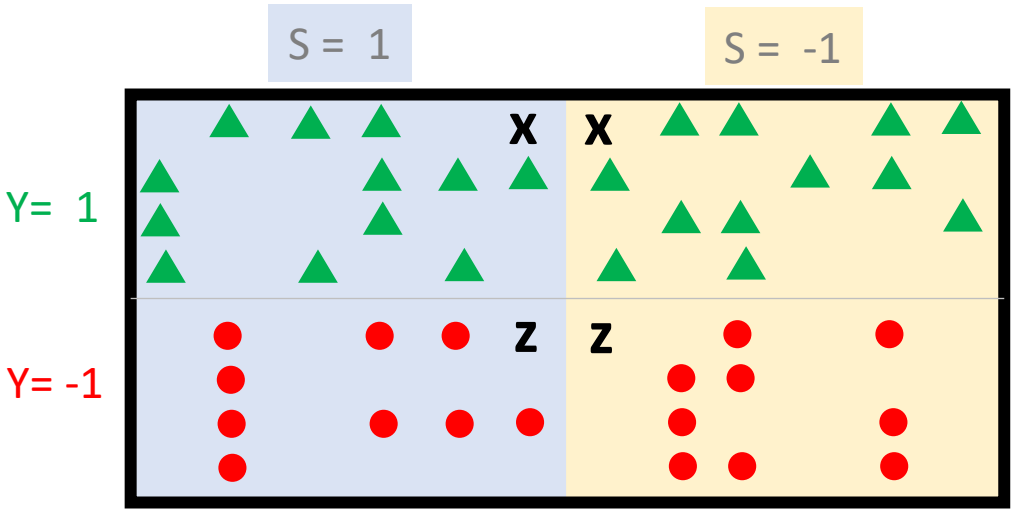
$FPR_b = FPR_w = 1/2$





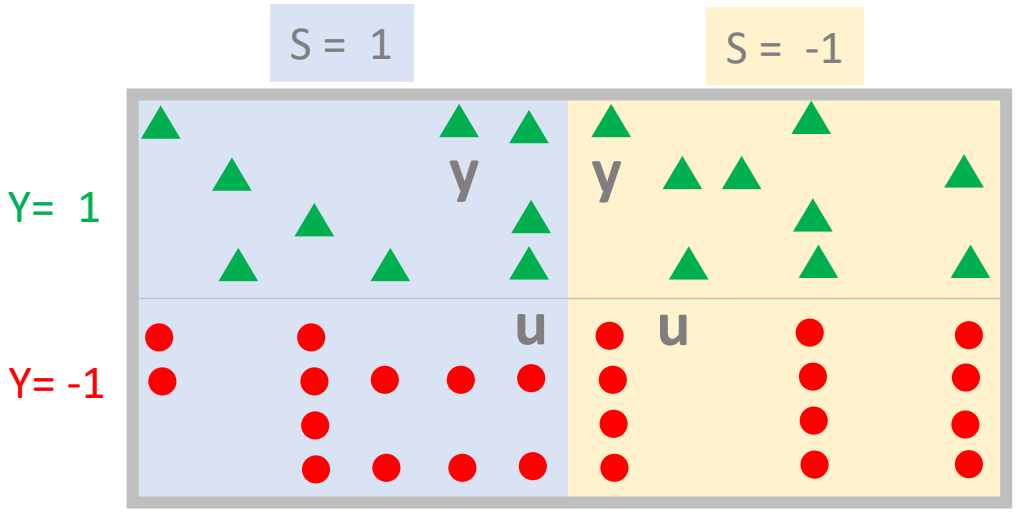
Overall situation

What are PPV_b and PPV_w?



b

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



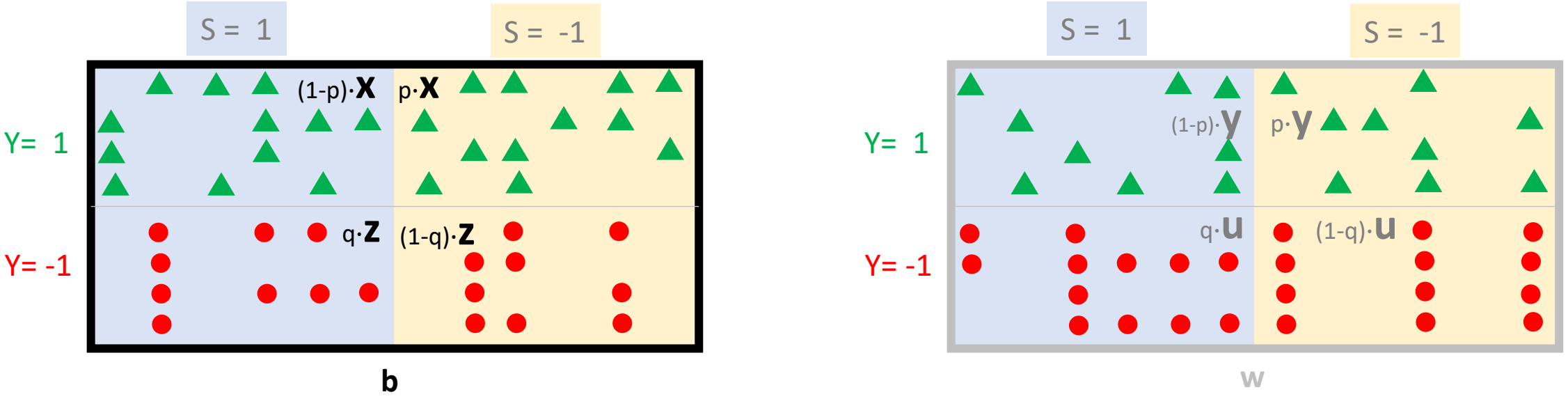
w

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

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

When is PPV_b = 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

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

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

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:

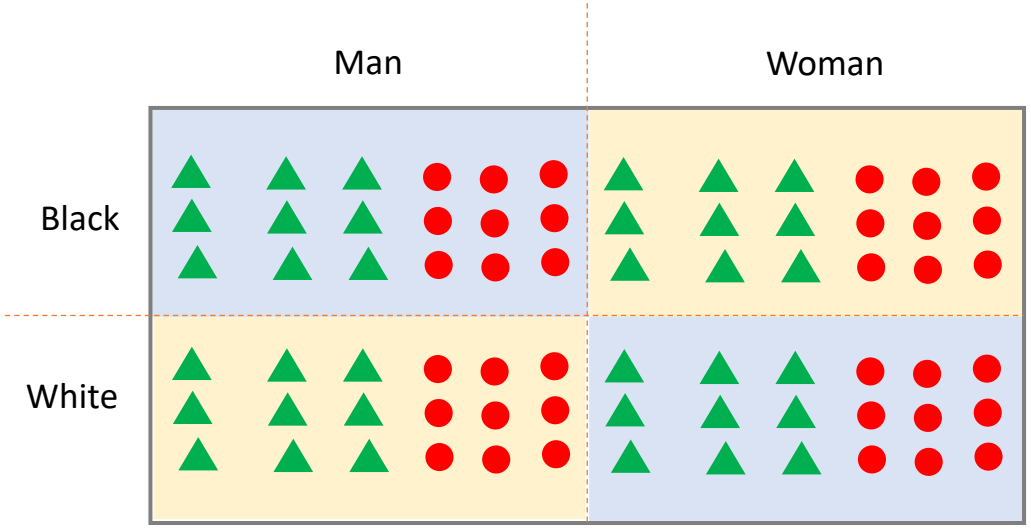
Fairness gerrymandering

Fairness gerrymandering

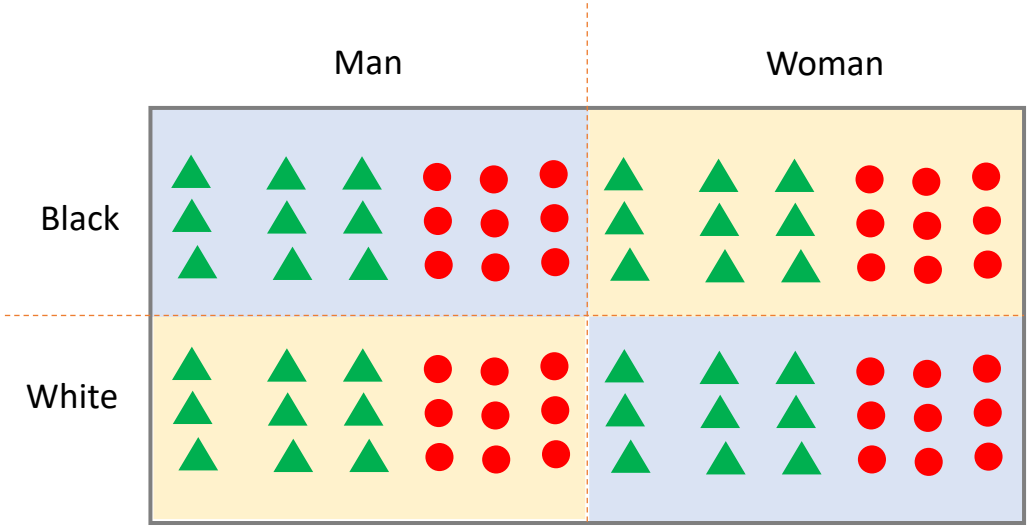
The general idea of this example is from Dwork et al. though the specific version (and indeed the term `fairness gerrymandering`) is from Kearns et al.

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.