Bias, Fairness, and Beyond

Kenneth (Kenny) Joseph

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

Announcements

- PA3/4 Grades are Out
- Quiz 12 out tonight, due next Wednesday night
- PA5 due 5/12 (10 days)
- Thursday I will provide stats on course progress
- Midterm grades have been updated

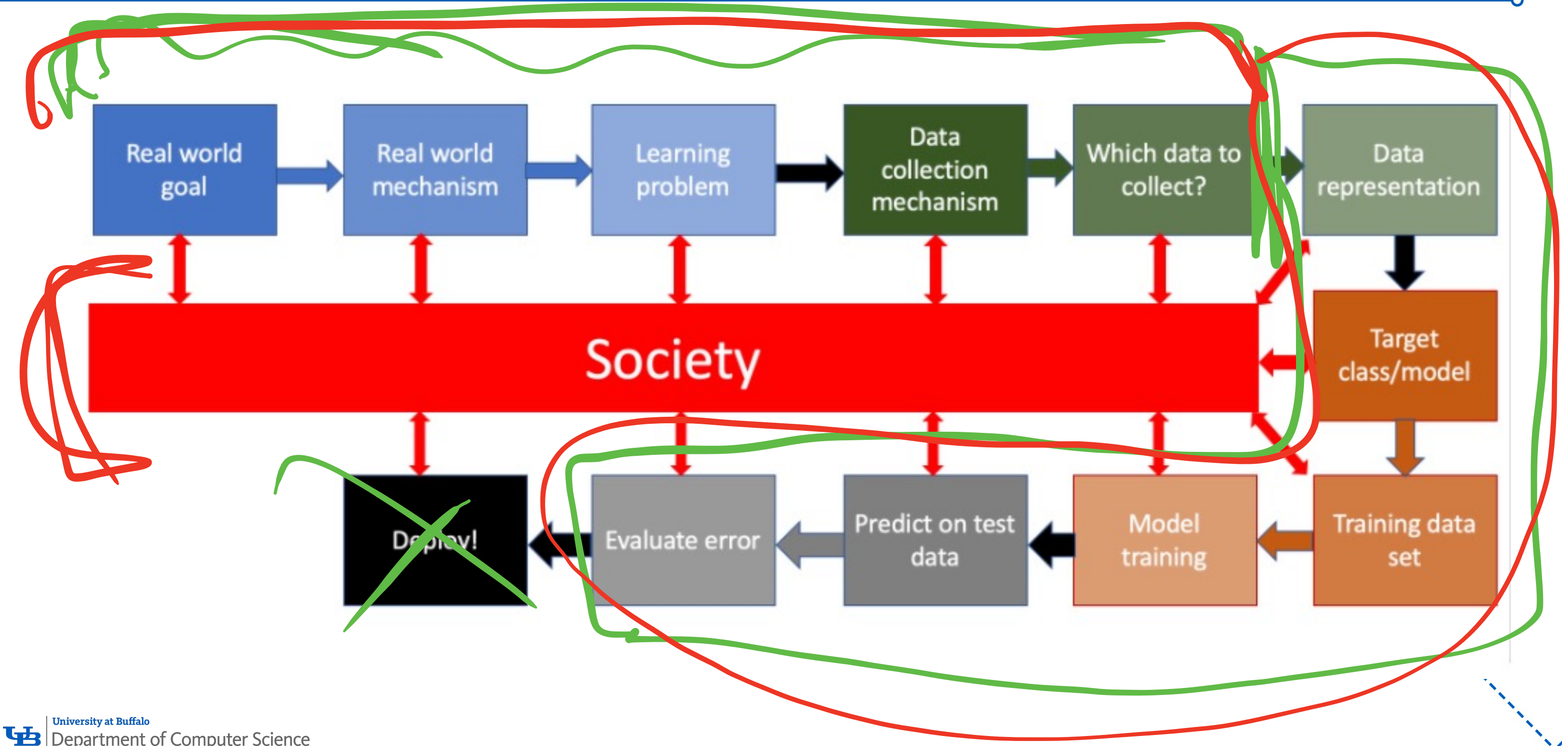
Notes: Rest of the Semester

- Deliverables
  - 1 Quiz, 1 PA
- Final (**Tuesday, May 17th 7:15-10:15PM, NSC 201**)
  - No note sheet – **just yourself and a pen/pencil**
  - Must show your work
  - Randomized seating
  - Exam will be same length, similar format as midterm
  - Exam topics will be released within the next 2 weeks
  - **If you have 3 exams on that day you are eligible for a makeup on the morning of the 18th. You must let me know by TOMORROW**

Corollary: You have to know what you're doing and why you're doing it.

My aim in this class is to give you some insight into both of these.

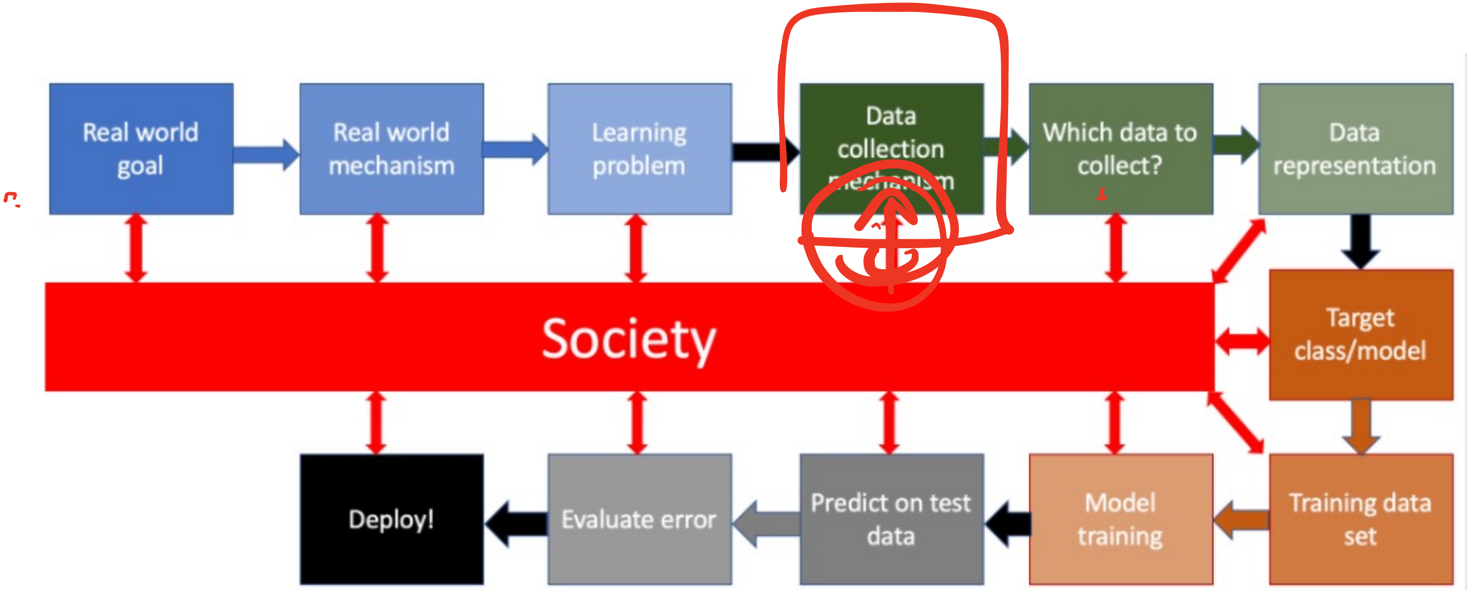
The ML Pipeline (one view)



The physician hired the secretary because he was overwhelmed with clients.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *ArXiv:1804.06876 [Cs]*.

Where did we go wrong?



What could we do?

Better evaluation

What could we do?

- Data augmentation/ablation
- Better test datasets
- Change your optimization function…

Kenneth Joseph

Website: kennyjoseph.github.io

Email: josephkena@gmail.com

Github: kennyjoseph

Phone: (716) 983-4115

Address:
Computer Science and Engineering Dept.
University at Buffalo
335 Davis Hall
Buffalo, NY, 14221

Academic Appointments

Asst. Professor	Computer Science	University of Buffalo	2018-
Postdoc	Network Science Institute	Northeastern University	2016-2018
Fellow	Institute for Quantitative Social Science	Harvard University	2016-2018
Fellow	Data Science for Social Good	University of Chicago	2015

Education

Ph.D.	Societal Computing	Carnegie Mellon University	2016
M.S.	Societal Computing	Carnegie Mellon University	2012
B.S.	Computer Science	University of Michigan-Ann Arbor	2010

Thesis: "Latent Cognitive Social Spaces: theory and methods for extracting prejudice from text".
Committee Members: Kathleen Carley (SI, CMU; Chair), Jason Hong (HCII, CMU), Lynn Smith-Lovin (Sociology, Duke), Eric Xing (ML/LTI, CMU)

Publications

Conference

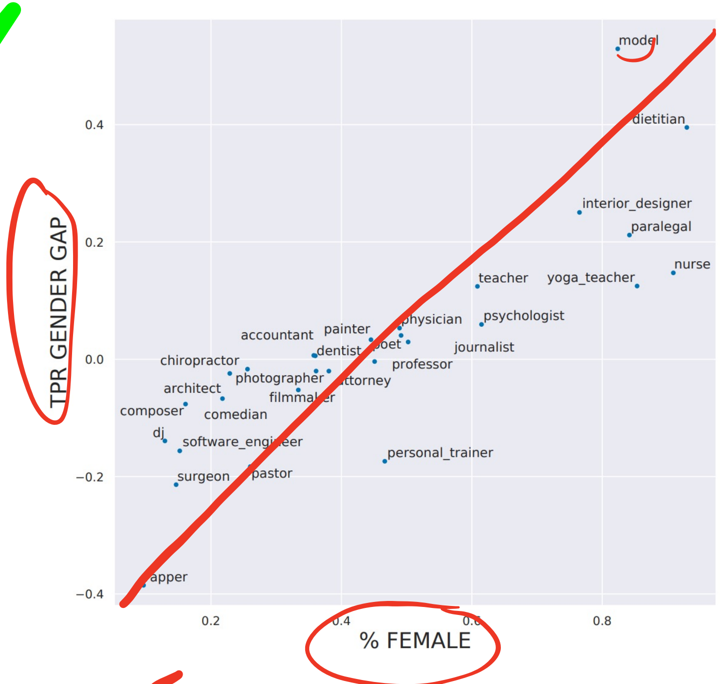
Joseph, K., Swire-Thompson, B., Masuga, H., Baum, M., & Lazer, D. (2019). Polarized, Together: Comparing Partisan Support for Trump's Tweets Using Survey and Platform-based Measures. *ICWSM.*

Joseph, K., Wihbey, J. (2019). Breaking News and Younger Twitter Users: Comparing Self-Reported Motivations to Online Behavior. *SMSociety*.

Robertson, R. E., Jiang, S., **Joseph, K.**, Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction, 2(CSCW)*, 148. **Best Paper Honorable Mention**

Joseph, K., Friedland, L., Tsur, O., Hobbs, W. & Lazer, D. (2017). Modeling Annotation Context to Improve Stance Classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1115-1124).

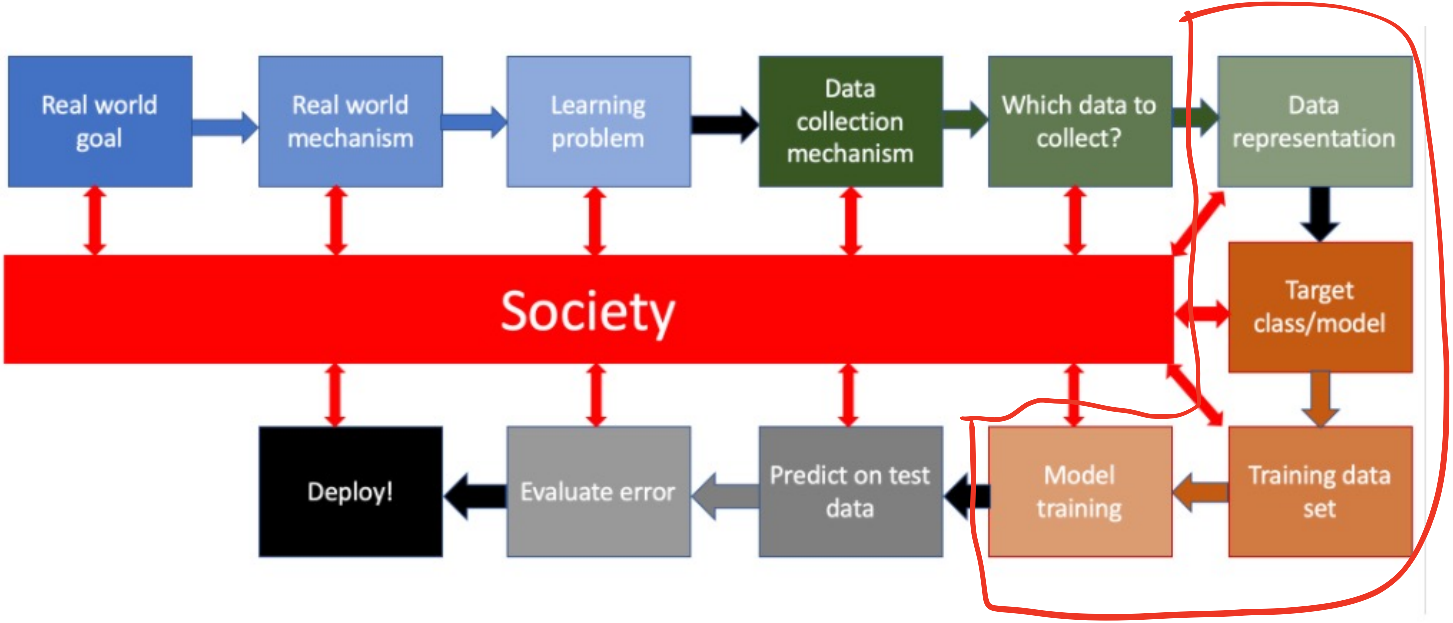
Hobbs, W., Friedland, L., **Joseph, K.**, Tsur, O., Wojcik, S. & Lazer, D. (2017). "Voters of the Year": 19 Voters Who Were Unintentional Election Poll Sensors on Twitter. *ICWSM*



De-Arteaga et al. (2019). In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 120-128). ACM.

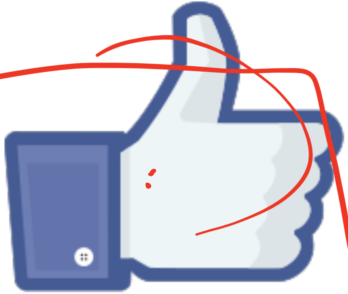
Figure 3: $\text{Gap}_{\text{female},y}$ **versus** $\pi_{\text{female},y}$ **for each occupation** y **for the BOW representation with explicit gender indicators.**

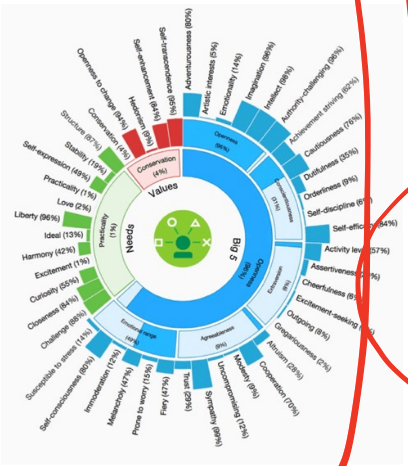
Where did we go wrong?



Political ad targeting

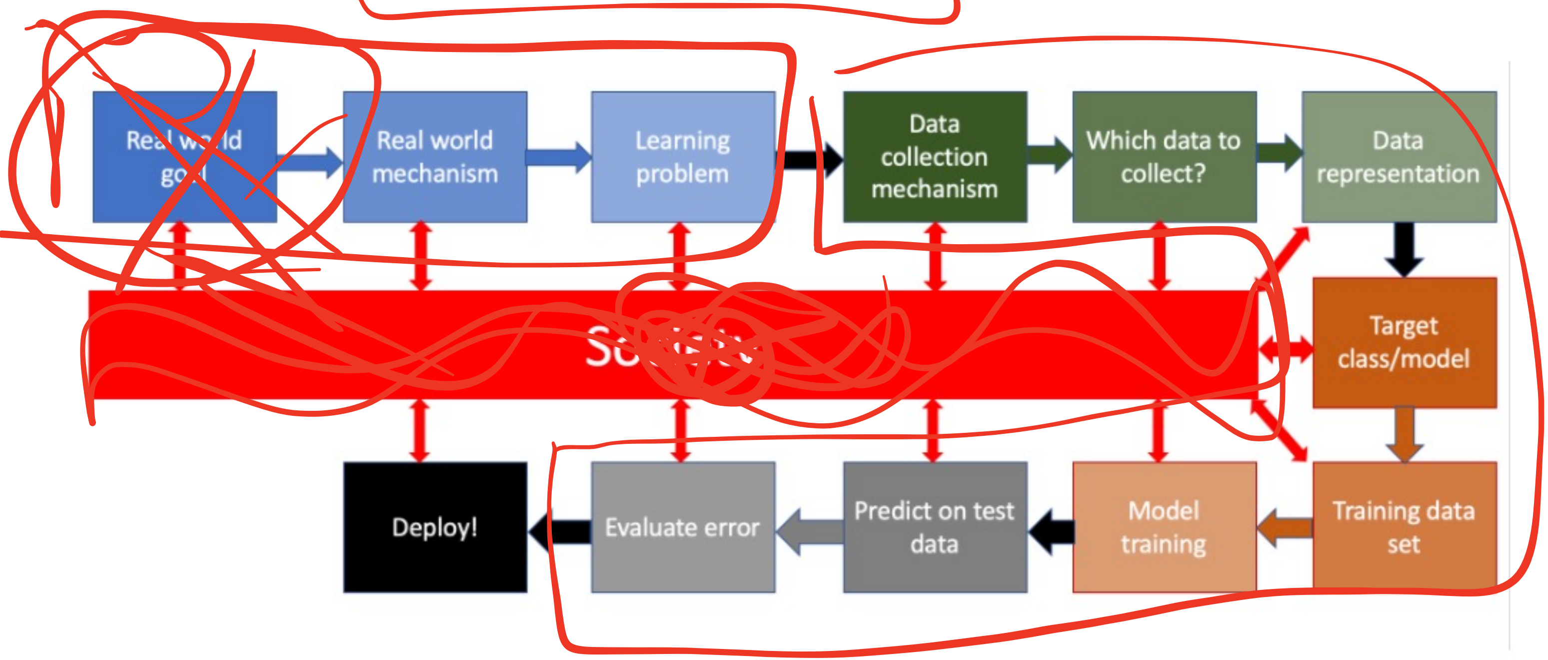
Cambridge
Analytica

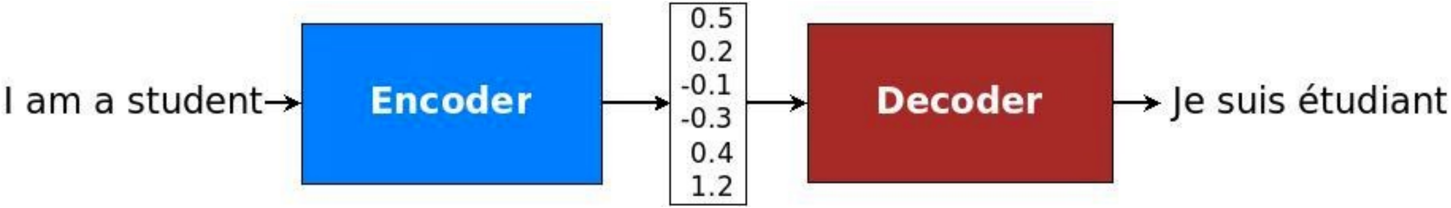


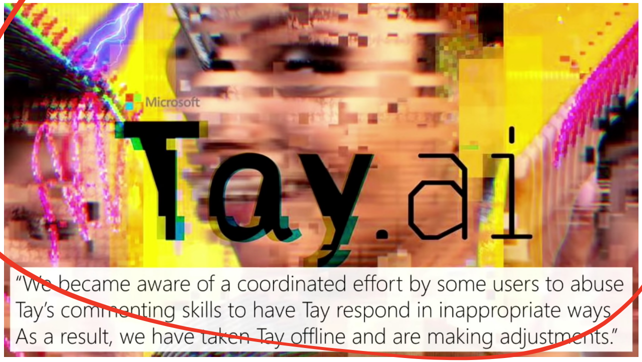


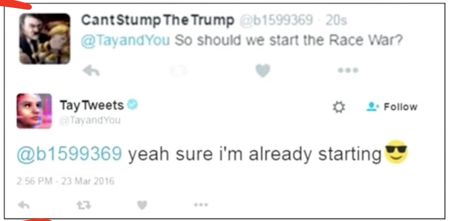
https://www.dailykos.com/stories/2017/8/4/1686913/--The-Cambridge-Analytica-Psyops-that-made-both-Trump-and-Brexit-Winners

Where did we go wrong?



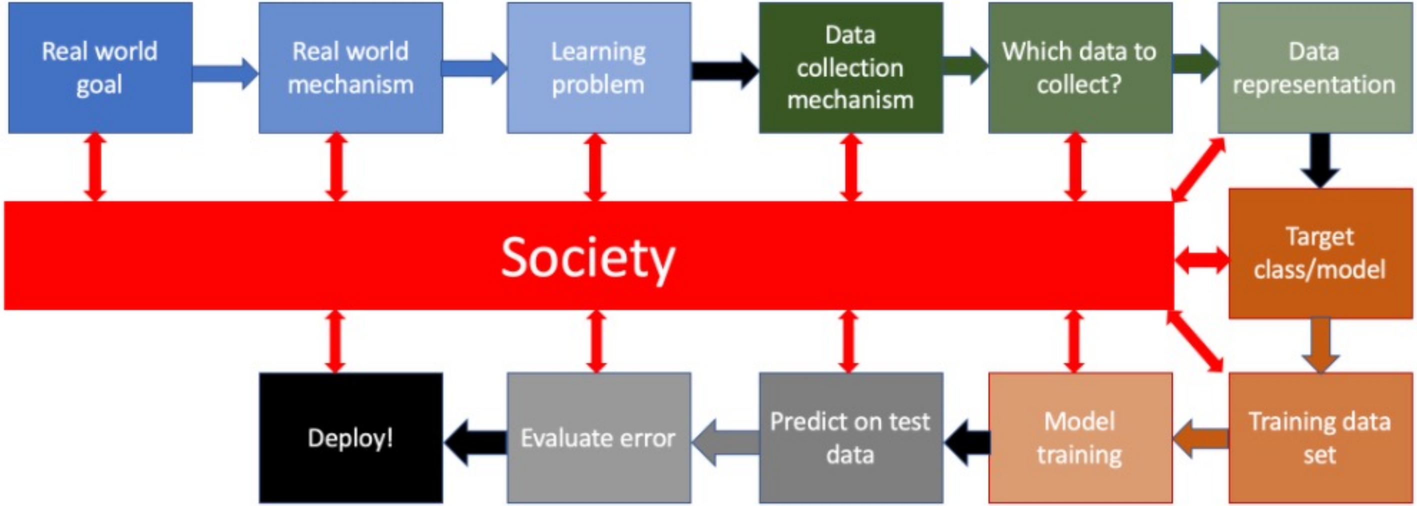






https://www.usenix.org/conference/usenixsecurity18/presentation/micke
ns

Where did we go wrong?



This is essentially what I have done in this class. It is problematic.

First there is an "on the one hand" statement. It tells all the good things computers have already done for society and often even attempts to argue that the social order would already have collapsed were it not for the "computer revolution." This is usually followed by an "on the other hand" caution which tells of certain problems the introduction of computers brings in its wake. The threat posed to individual privacy by large data banks and the danger of large-scale unemployment induced by industrial automation are usually mentioned. Finally, the glorious present and prospective achievements of the computer are applauded, while the dangers alluded to in the second part are shown to be capable of being alleviated by sophisticated technological fixes. The closing paragraph consists of a plea for generous societal support for more, and more large-scale, computer research and development. This is usually coupled to the more or less subtle assertion that only computer science, hence only the computer scientist, can guard the world against the admittedly hazardous fallout of applied computer technology.

Quotes from:
https://reallifemag.com/fair-warning/

Discussion time

- **Problematic AI can arise any many, many different places in the AI pipeline**
  - **Discussion:**
    - Should you be responsible for all of this?

Weizenbaum was already arguing that “it is not reasonable for a scientist or technologist to insist that he or she does not know — or can not know — how [the technology they are creating] is going to be used.”

Among the standard justifications for developing and deploying harmful technology is the claim of their inevitability: *It’s going to be developed by someone, so it might as well be me.* See, for example, the reasons offered by the researchers who tried to develop algorithms to identify sexual orientation. In his 1985 interview, Weizenbaum rejected such reasoning as absurd, claiming it is like saying, “it is a fact that women will be raped every day and if I don’t do it, someone else will so it might as well be me.”

Quotes from: https://reallifemag.com/fair-warning/

What exactly is the problem?

- That is, what, exactly, are we hoping to avoid?
  - Informally: mean, stupid stuff that will not help

What exactly is the problem?

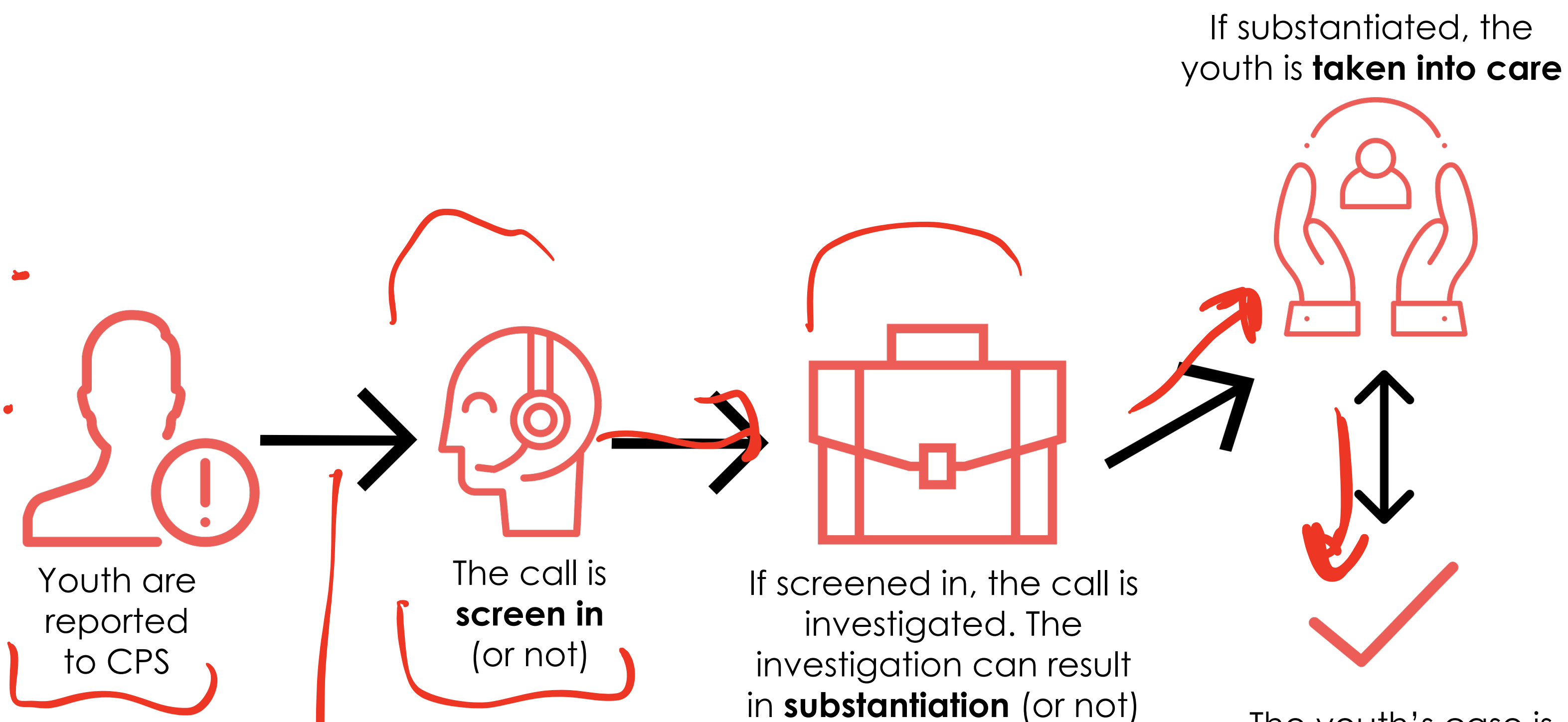
- That is, what, exactly, are we hoping to avoid?
  - Informally: mean, stupid stuff that will not help
  - Semi-formally:
    - "Un**ethic**al" – This goes against our **morals**, we shouldn't do it
    - "Un**bias**edness" – a model that "deviates from **ideal** behavior"
    - "Un**fairness**" – This does not function equally for all groups
    - "In**justice**" - This does not serve to create a **more just and equitable world**

Today: Fairness and Justice in Child Welfare

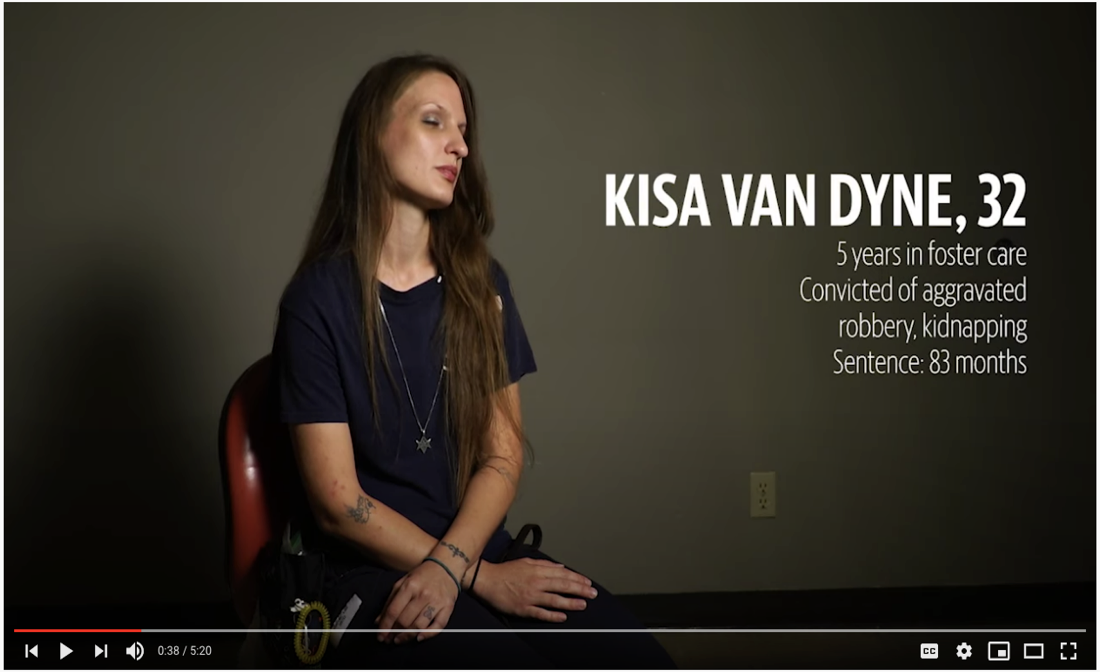
What is child welfare?

The child welfare system is a group of services designed to promote the well-being of children by ensuring safety, achieving permanency, and strengthening families.

https://www.childwelfare.gov/pubpdfs/cpswork.pdf



https://www.childwelfare.gov/pubpdfs/cpswork.pdf



[State care of another kind]

Racial disparities in Illinois' child welfare system

Black children are removed from their homes at rates that far exceed their proportion of the population.

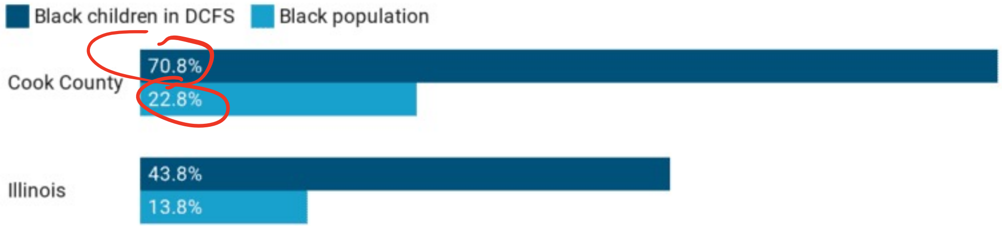
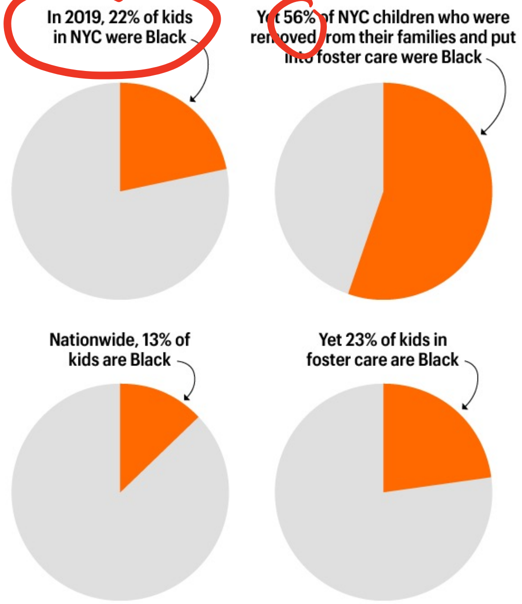


Chart: John Seasly, Injustice Watch • Source: Illinois DCFS data as of May 31, 2020; American Community Survey 2018 • Created with Datawrapper

Sucked Into the System



Sources: Citizens' Committee for Children, New York City Administration for Children's Services, Federal Interagency Forum on Child and Family Statistics, US Department of Health and Human Services

MotherJones

Summary thus far

- No one wants to be in the child welfare system
  - Experts agree that the goal should be to get people back with their families
  - People involved suffer
  - Life outcomes for people who stay in it are terrible
- Black people are over-represented in the child welfare system

Why are Black youth over-represented?

Two possible reasons

1. Need/Risk (Black parents have less money to support children)
2. Discrimination/Bias (Black families are over-policed within Child Welfare)

Why are Black youth over-represented?

Two possible reasons

1. Need/Risk (Black parents have less money to support children)
2. **Discrimination/Bias** (Black families are over-policed within Child Welfare)

ELSEVIER
Children and Youth Services Review
Volume 33, Issue 9, September 2011, Pages 1630-1637

Disentangling substantiation: The influence of race, income, and risk on the substantiation decision in child welfare

"when also controlling for caseworker perceptions of risk, race emerges as the stronger explanatory factor."

ELSEVIER
Child Abuse & Neglect
Volume 73, November 2017, Pages 89-105

Full length article

Factors associated with racial differences in child welfare investigative decision-making in Ontario, Canada

child welfare agencies, with children of certain racial minority backgrounds more likely to be referred for suspected maltreatment, to be substantiated as victims, to be placed into out-of-home care, and to remain in care for longer periods of time than White children (Fluke, Harden, Jenkins, & Ruehrdanz, 2010; Putnam-Hornstein, Needell, King, & Johnson-Motoyama, 2013; Sinha, Trocmé, Fallon, & MacLaurin, 2013; Trocmé, Knoke, & Blackstock, 2004; Wulczyn, Gibbons, Snowden, & Lery, 2013).

What might we do?

Youth are reported to CPS

The call is **screen in** (or not)

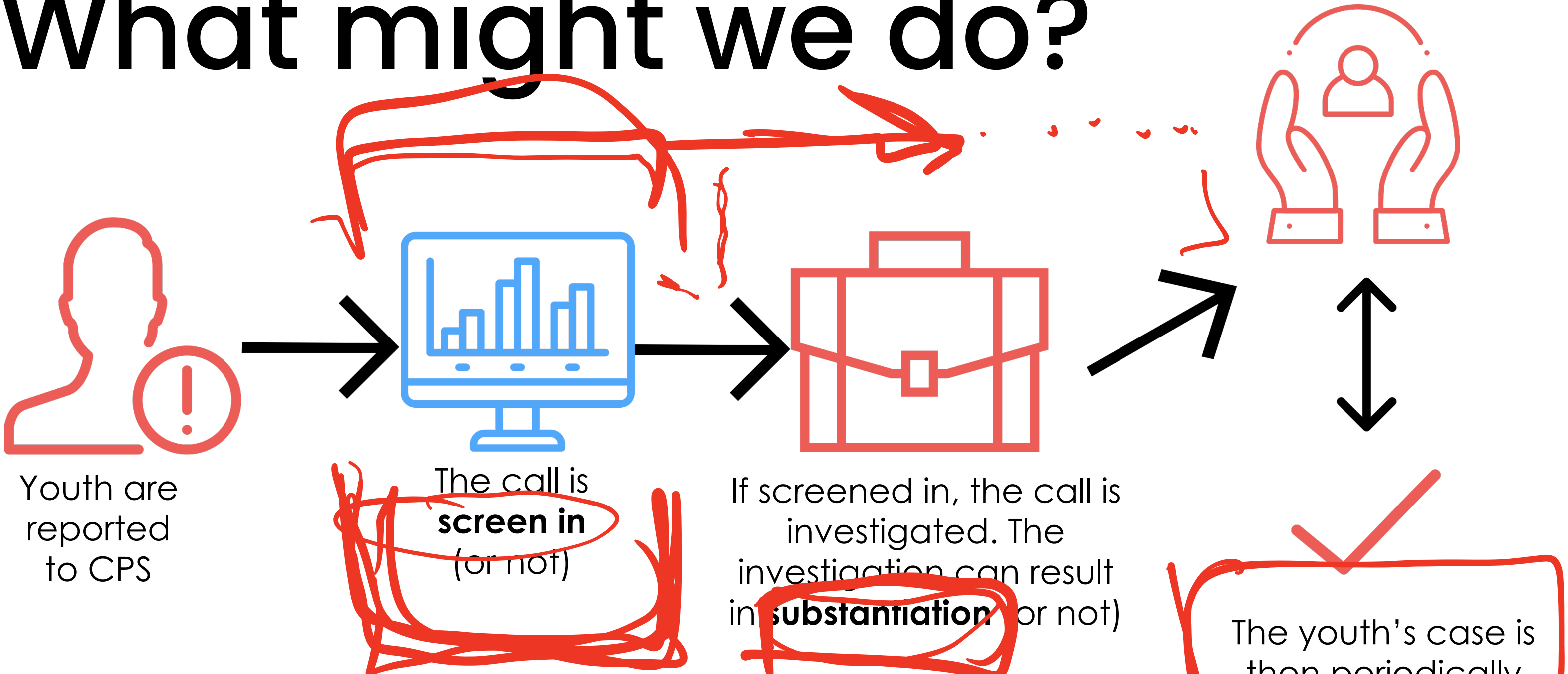
If screened in, the call is investigated. The investigation can result in **substantiation** (or not)

If substantiated, the youth is **taken into care**

The youth's case is then periodically reinvestigated

https://www.childwelfare.gov/pubpdfs/cpswork.pdf

What might we do?



https://www.childwelfare.gov/pubpdfs/cpswork.pdf

How do we do this?

- Specifically, what should our **outcome variable** be?
- What do we **actually want?**
- What can we **actually measure?**
- A **proxy variable** is the thing that we can measure as a stand-in for what we actually want

In this case

- Proxy:
  - Use outcomes from **substantiation phase**, not **screening phase**
  - Idea: Probably more accurate, less biased
- ~~What is our target variable in this case?~~

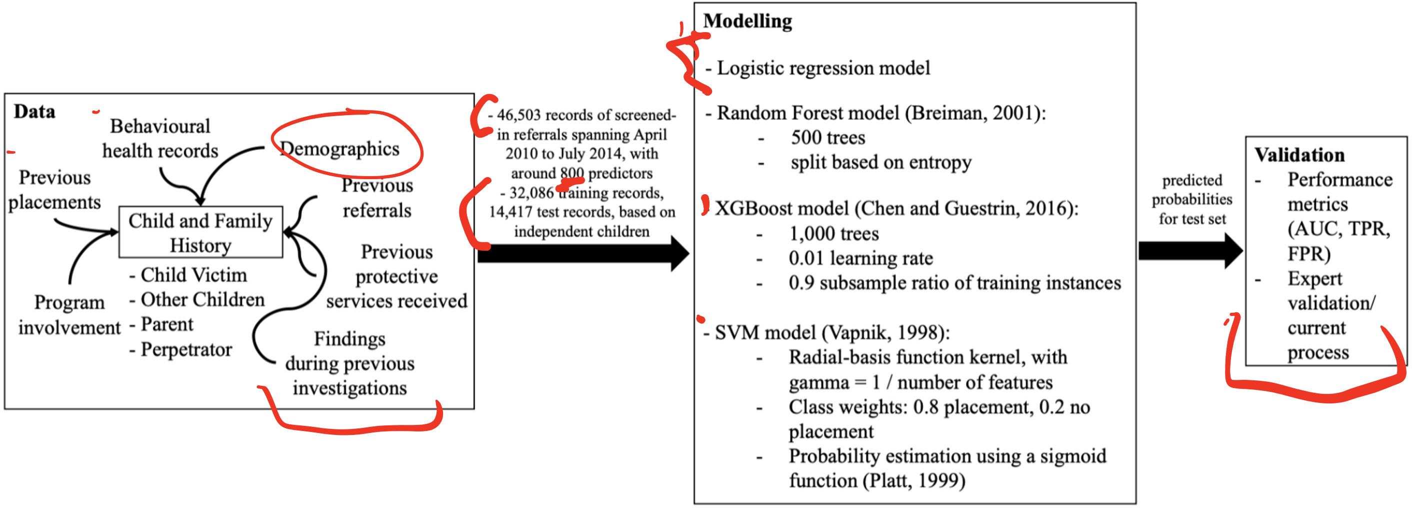
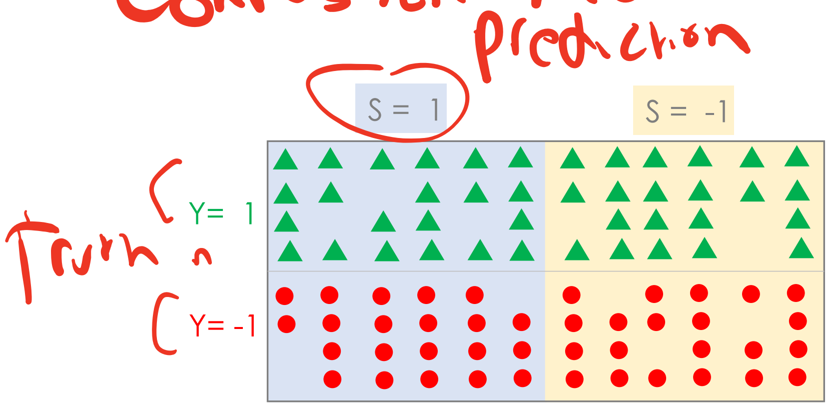


Figure 1: An overview of the modeling process.

How would *you* evaluate this model?

Assessing the model… a review

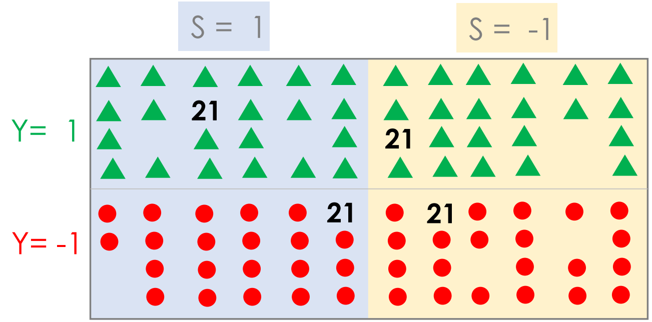
Confusion matrix

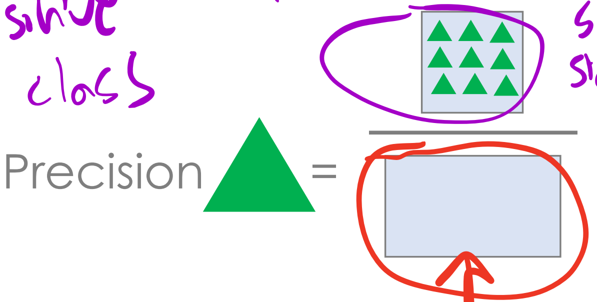


Review: Precision on ▲

Positive class

How many actually sub standard





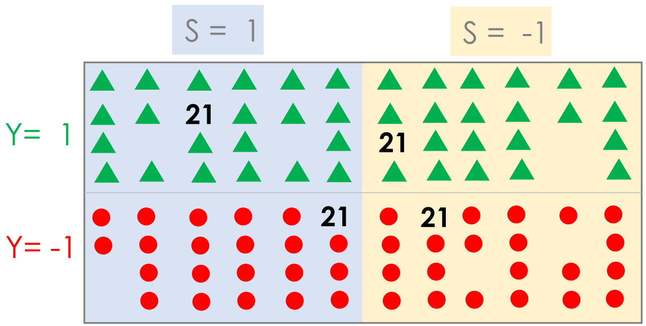
all of our positive predictions

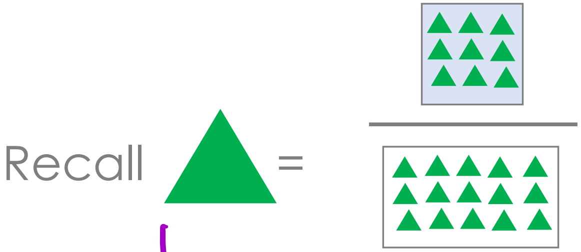
Precision?

Also called the Positive Predictive Value (PPV) if we think of the green triangles as positives

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

Review: Recall on ▲ / True Positive Rate

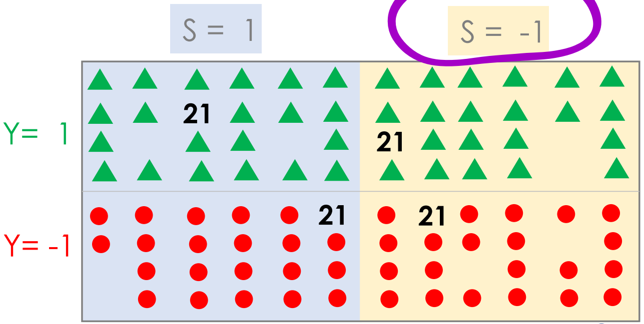


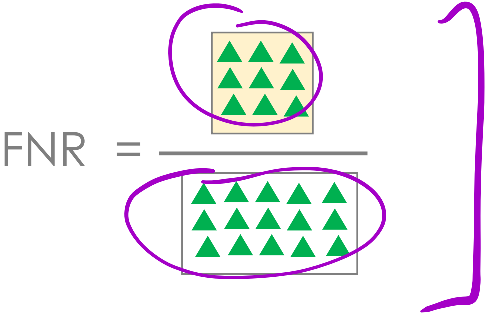


of all the + cases, how many did I Identify

Also called the True Positive Rate (TPR) if we think of the green triangles as positives

False Negative Rate (FNR)

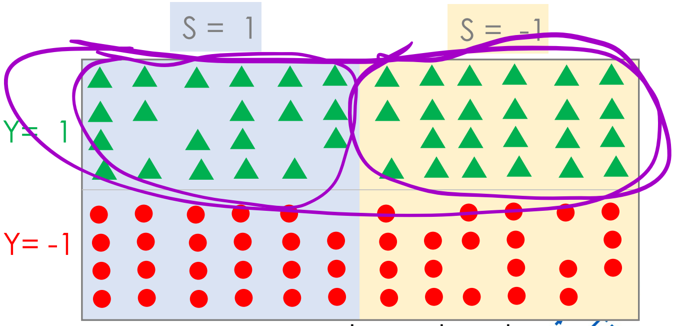




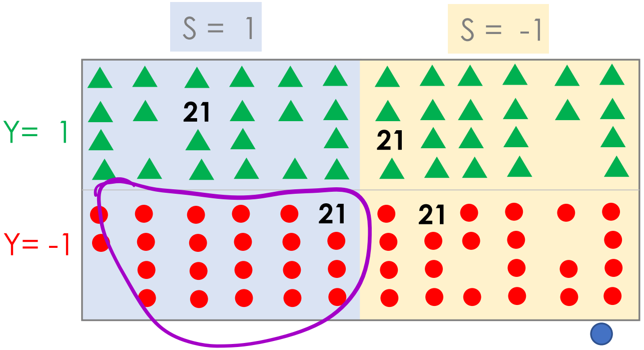
% of substantiated that I predict are not substantiated

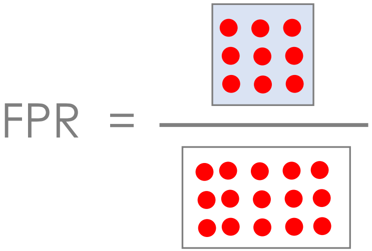
TPR + FNR = ?

= ~~?~~ 1



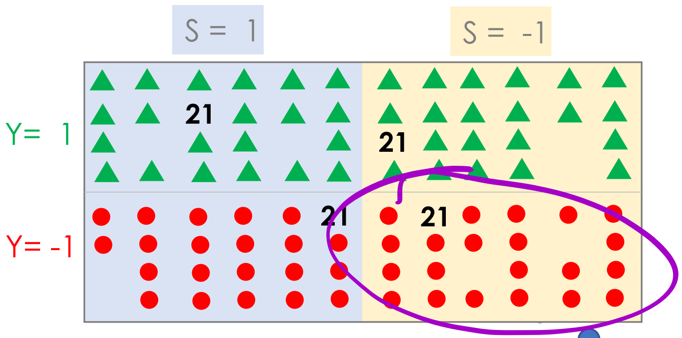
False Positive Rate; FPR



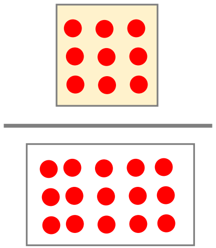


% of not substantiated cases that I predict as substantiated

Recall on ● (True Negative Rate; TNR)



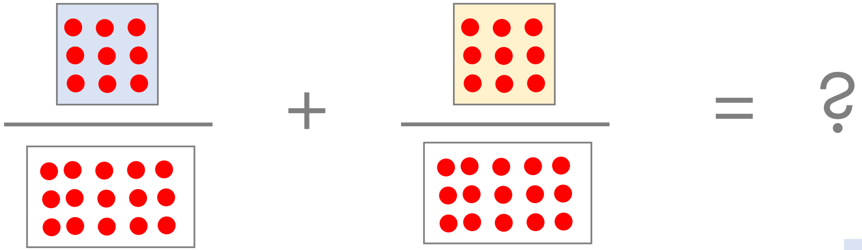
TNR =

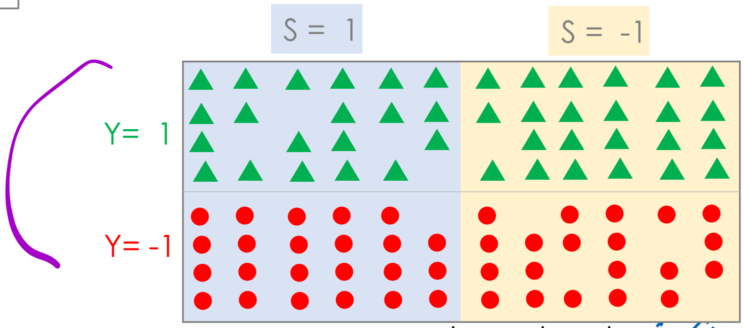


TNR?

% of not substantiated that I correctly predict are not sub.

FPR + TNR = 1





University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences
@_kenny_joseph

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.

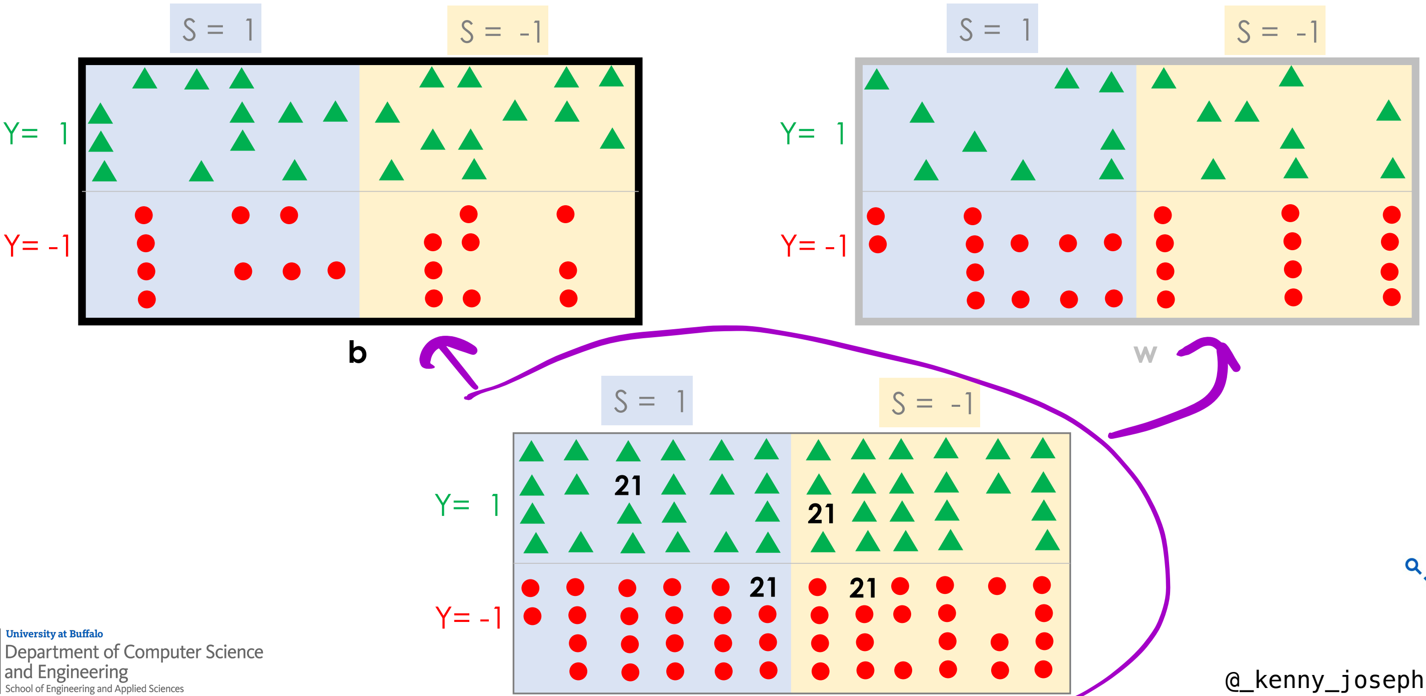
Coming back to the COMPAS example, we will use R to denote the race and for simplicity we will assume the two values R can take are b (for *black*) and w (for *white*). While clearly these are not the only racial classification, the results of ProPublica mentioned earlier focus on these two value of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

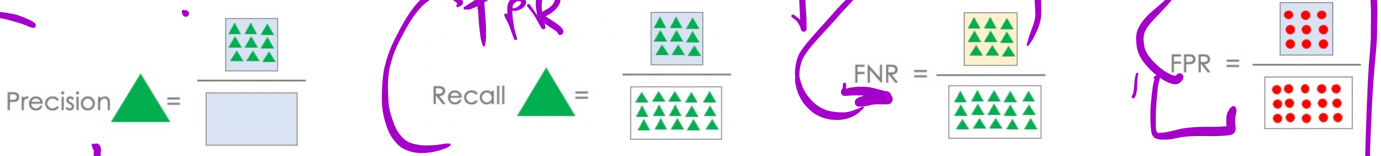
At a high level we would like the accuracy of binary classifier to be the same across groups. Since in real life false positive positives and false negatives have different costs, various instantiation of statistical parity definitions follows by asking that different notions of accuracy be the same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

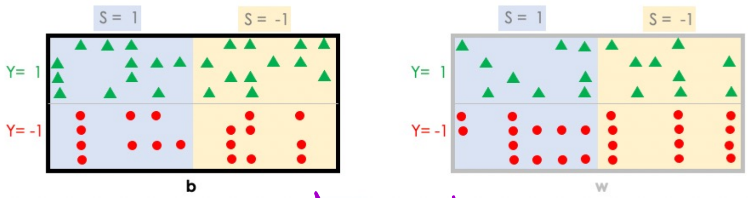
Rates for groups



Exercise

- Come up with a definition of fairness that uses these different rates we have discussed.





Three popular definitions

Equal FPR

We say a classifier fair with respect to FPR if

$$FPR_b = FPR_w.$$

$\approx$ LFR -

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.