MACHINE LEARNING FOR HEALTHCARE

6.S897, HST.S53

Lecture 4: Fairness and bias

Prof. David Sontag

MIT EECS, CSAIL, IMES

(Thanks to Nati Srebro, Moritz Hardt, and Rich Zemel for some slides)

Massachusetts Institute of Technology

Outline

1. Commercialization of risk scores in healthcare
2. ProPublica article on machine bias
3. Formalizing fairness

Example commercial product

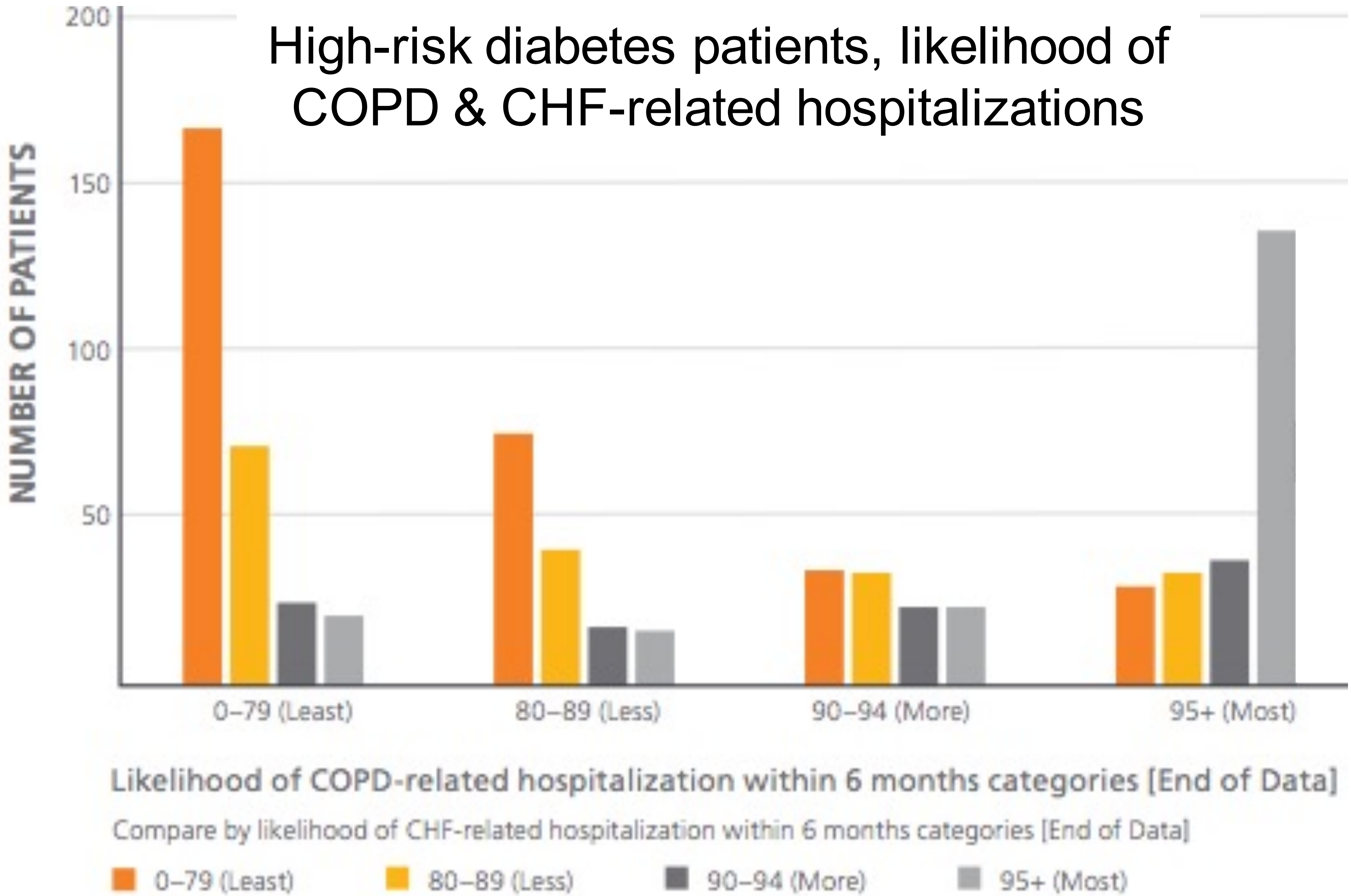
Area Under the Receiver Operating Curve (C-STATS)

HOSPITAL ADMISSIONS MODELS	IDN MODEL	NON-IDN MODEL
CONGESTIVE HEART FAILURE MODEL		
Training sample	0.757	0.742
Avg of testing samples	0.739	0.708
CHRONIC OBSTRUCTIVE PULMONARY DISEASE MODEL		
Training sample	0.833	0.802
Avg of testing samples	0.830	0.799
DIABETES MELLITUS MODEL		
Training sample	0.765	0.754
Avg of testing samples	0.781	0.765
PEDIATRIC ASTHMA MODEL		
Training sample	0.784	0.739
Avg of testing samples	0.761	0.716

NOTE: *Models developed using data from over 30M patients (inclusive of all conditions). All models predict both initial admission and readmission, for both inpatient and emergency department. Pediatric asthma model also predicts observation visits.*

Optum Whitepaper, “Predictive analytics: Poised to drive population health"

Example commercial product



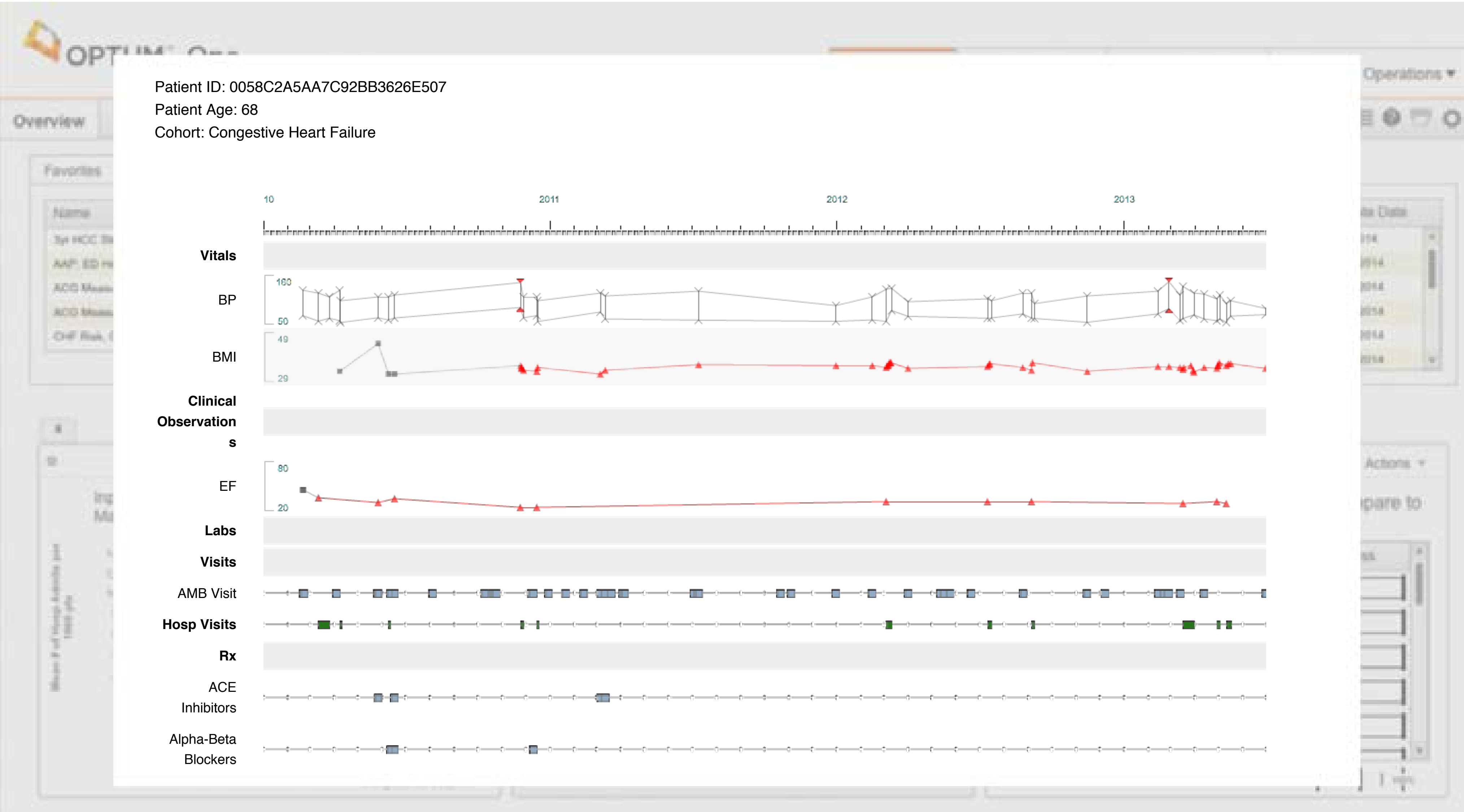
Optum Whitepaper, “Predictive analytics: Poised to drive population health"

Example commercial product

High-risk diabetes patients missing tests	# of A1c tests	# of LDL tests	Last A1c	Date of last A1c	Last LDL	Date of last LDL
Patient 1	2	0	9.2	5/3/13	N/A	N/A
Patient 2	2	0	8	1/30/13	N/A	N/A
Patient 3	0	0	N/A	N/A	N/A	N/A
Patient 4	0	2	N/A	N/A	133	8/9/13
Patient 5	0	0	N/A	N/A	N/A	N/A
Patient 6	0	1	N/A	N/A	115	7/16/13
Patient 7	1	0	10.8	9/18/13	N/A	N/A
Patient 8	0	0	N/A	N/A	N/A	N/A
Patient 9	0	0	N/A	N/A	N/A	N/A
Patient 10	0	0	N/A	N/A	N/A	N/A

Optum Whitepaper, “Predictive analytics: Poised to drive population health"

Example commercial product



Optum Whitepaper, “Predictive analytics: Poised to drive population health"

Example commercial product

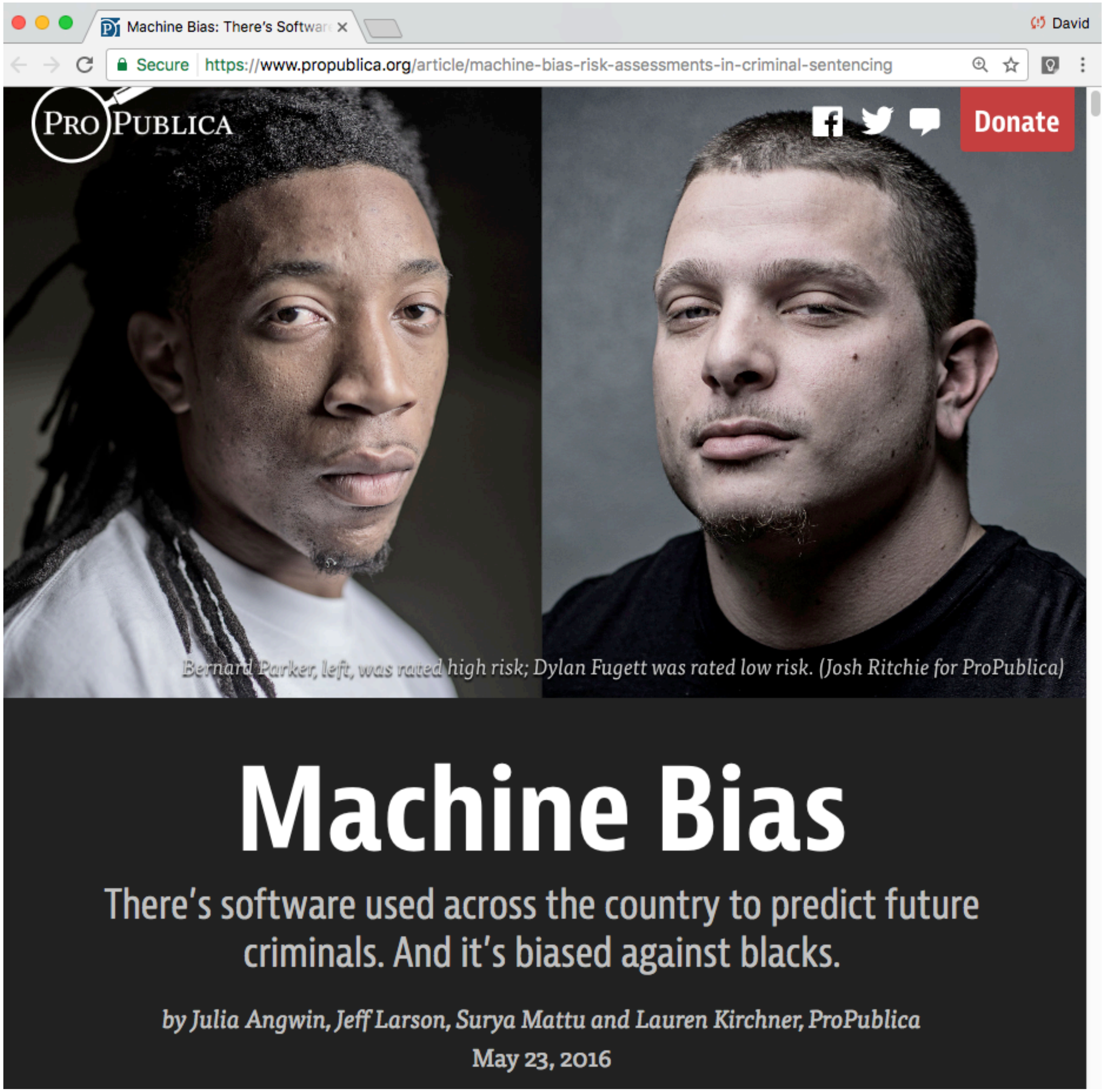
Score Calculation

Description	12m
Lower cost infectious disease	0.1725
CAD, heart failure, cardiomyopathy, II	0.3932
Endocrinology Specialty	0.1715
Cardiology Specialty	0.2840
If 2 A&E Attendances in last 3 month period	0.7340
If sum of Length of Stay less than 5 days in period	0.3645
Male aged between 45-54	0.9491
If greater than 3 first or follow-up Outpatient Attendances in last 3 month period	0.2930

Intercept	**-5.4605**
TOTAL (-Intercept)	**-2.0987**
Exp (TOTAL)	**0.1092**

Optum Whitepaper, “HealthNumerics-RISC Predictive Models: A Successful Approach to Risk Stratification"

ProPublica article



Discussion points

- What are other areas of healthcare where we might be concerned with machine bias?
- What are the relevant protected groups?

- How do we *measure* bias if we don't observe the counterfactual?

Formalizing fairness

- Fairness through blindness
- Demographic parity / group fairness / statistical parity
- Calibration / predictive parity
- Error rate balance / equalized odds
- Individual fairness

Fairness through Blindness

The case of ProPublica versus Northpointe

- Score S=S(x) satisfies *predictive parity* at threshold s_{HR} if

$$\mathbb{P}(Y = 1 \mid S > s_{\mathrm{HR}}, R = b) = \mathbb{P}(Y = 1 \mid S > s_{\mathrm{HR}}, R = w)$$

  where R is the protected attribute taking two states, *b* or *w*

- I.e., positive predictive value (PPV) same across groups

(Chouldechova, “Fair prediction with disparate impact”,’17)

The case of ProPublica versus Northpointe

- Score S=S(x) satisfies *error rate balance* at threshold s_{HR} if

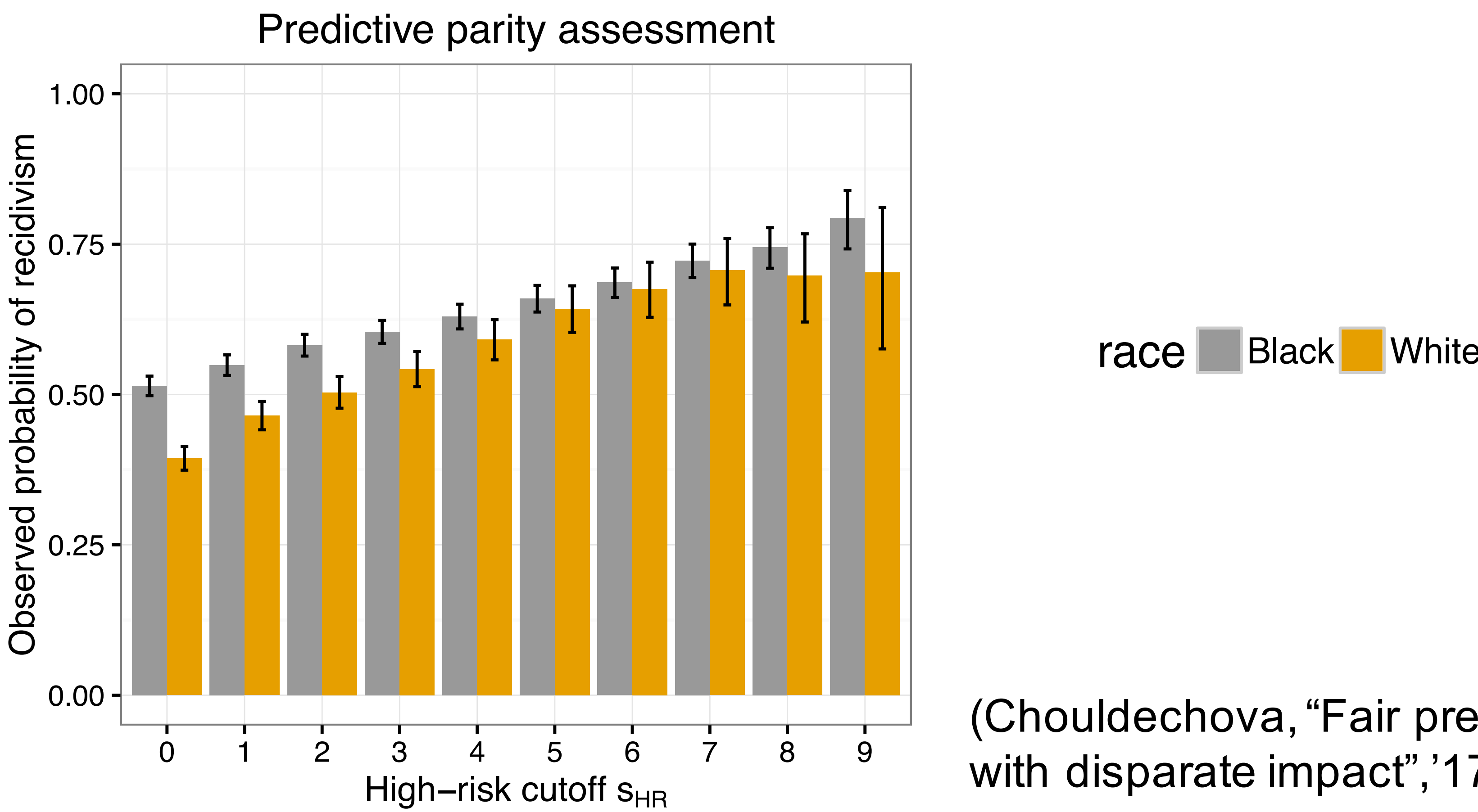
$$\mathbb{P}(S > s_{\mathrm{HR}} \mid Y = 0, R = b) = \mathbb{P}(S > s_{\mathrm{HR}} \mid Y = 0, R = w) \ , \quad \text{and}$$
$$\mathbb{P}(S \leq s_{\mathrm{HR}} \mid Y = 1, R = b) = \mathbb{P}(S \leq s_{\mathrm{HR}} \mid Y = 1, R = w),$$

where R is the protected attribute taking two states, *b* or *w*

(Chouldechova, “Fair prediction with disparate impact”,’17)

The case of ProPublica versus Northpointe

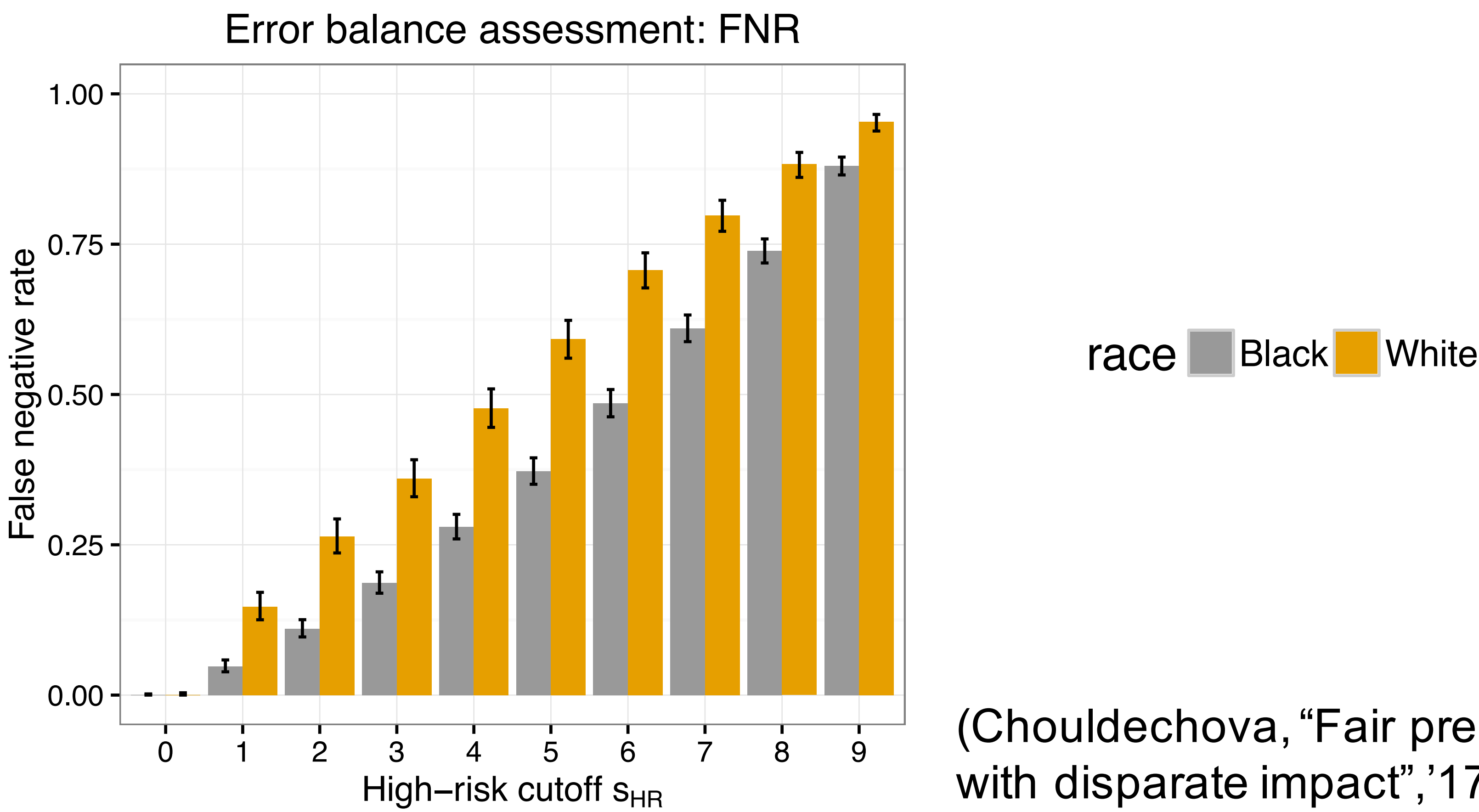
- Northpointe score approximately satisfies *predictive parity*: $\mathbb{P}(Y = 1 \mid S > s_{\mathrm{HR}}, R = b)$



(Chouldechova, “Fair prediction with disparate impact”,’17)

The case of ProPublica versus Northpointe

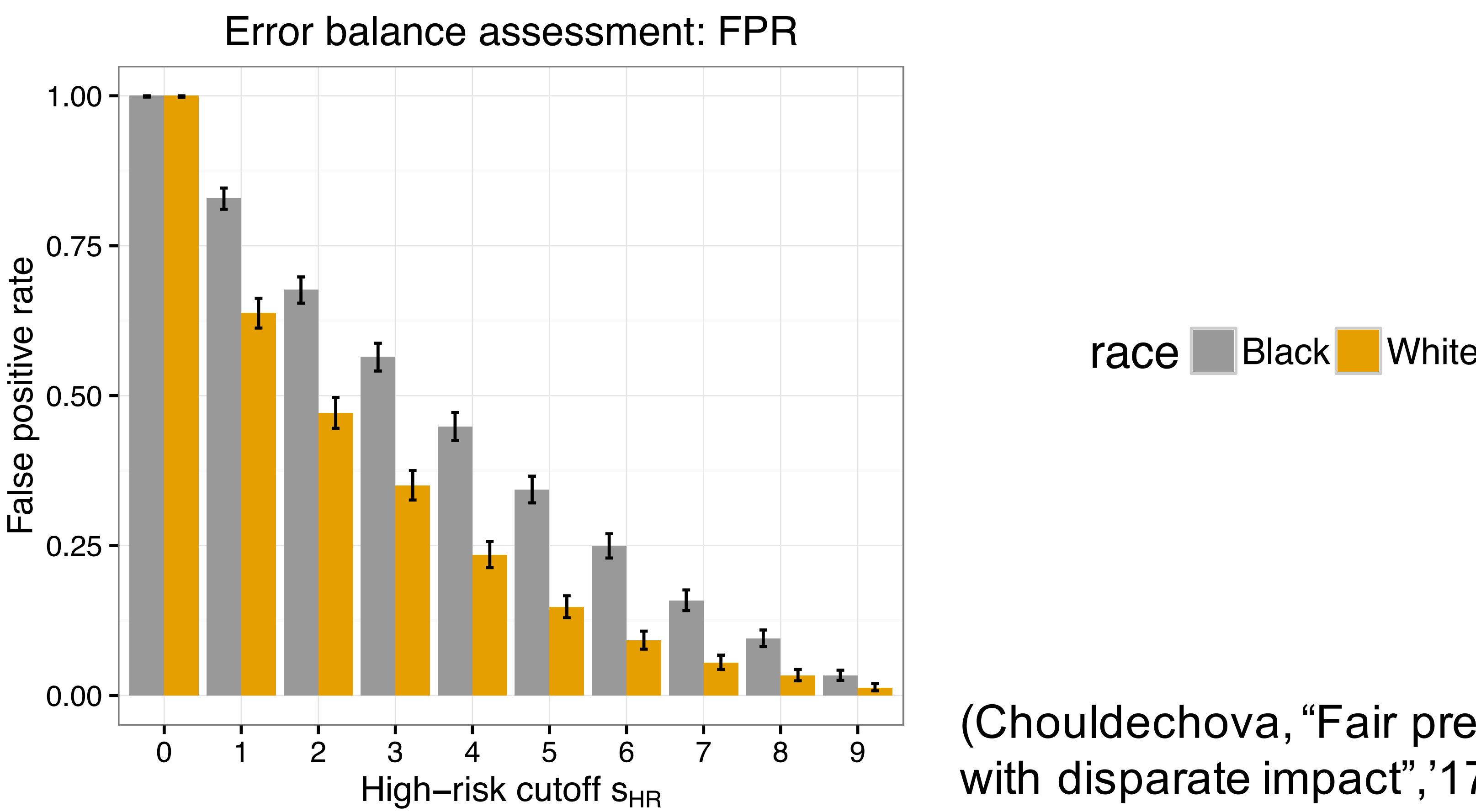
- Northpointe score does *not* satisfy *error rate balance*: $\mathbb{P}(S \le s_{\mathrm{HR}} \mid Y = 1, R = w)$



(Chouldechova, "Fair prediction with disparate impact",'17)

The case of ProPublica versus Northpointe

- Northpointe score does *not* satisfy *error rate balance*: $\mathbb{P}(S > s_{\mathrm{HR}} \mid Y = 0, R = w)$



(Chouldechova, “Fair prediction with disparate impact”,’17)

Impossibility of satisfying all 3 criteria

- Consider the following confusion matrix:

	Low-Risk	High-Risk
$Y=0$	TN	FP
$Y=1$	FN	TP

- Let *p* be the prevalence within a group. Then,

$$\mathrm{FPR} = \frac{p}{1-p}\frac{1-\mathrm{PPV}}{\mathrm{PPV}}(1-\mathrm{FNR})$$

- If PPV is the *same* across groups but *p* is *different* across groups, FPR/(1-FNR) must also be different across groups

(Chouldechova, “Fair prediction with disparate impact”,’17)

Non-Discrimination in Supervised Learning

- Formal setup:
  - Available features X (e.g. credit history, payment history, rent and house purchase history, number of dependents, driving record, employment record, education, etc)
  - Protected attribute A (e.g. race)
  - Prediction target Y (e.g. not defaulting on loan)
  - Learn predictor $\hat{Y}(X)$ or $\hat{Y}(X, A)$ for Y

- Learn based on training set $\{(x_i, a_i, y_i)\}_{i=1..m}$

  …but for now assume population distribution (X, A, Y) is known

- **What does it mean for $\widehat{Y}$ to be non-discriminatory?**

Demographic Parity

- Require the same fraction of $\hat{Y} = 1$ decisions in each population
  - If 70% of whites get loans, then also 70% of blacks should

- Can be stated as: $\hat{Y} \perp A$

Problems:

- What if true Y correlates with A?
- Even $\hat{Y} = Y$ (if we could somehow predict it perfectly) doesn't satisfy requirement
  - e.g. giving loans exactly to those that won't default

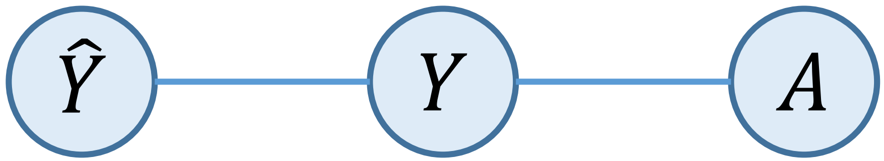
- Also too weak: doesn't control different error rate
  - e.g. allows giving loans to qualified $A = 0$ people and random $A = 1$ people

- Typical relaxation (with some legal standing), "The 80% Rule":

$$P\left(\hat{Y} = 1 \middle| A = 1\right) \leq 0.80 \cdot P(\hat{Y} = 1 | A = 0)$$

Suggested Notion: Equalized Odds

$$\hat{Y} \perp A | Y$$



- Prediction does not provide any additional information about A beyond what the truth Y already tells us on A

- The perfect predictor, $\hat{Y} = Y$, always satisfies equalized odds

- Compared to demographic parity:

$$P\left(\hat{Y} \middle| Y = y, A = a\right) = P(\hat{Y} | Y = y, A = a')$$

- Having $\hat{Y} \perp A$ is *not* sufficient for equalized odds

Ensuring Equalized Odds

- Given (possibly unfair) predictor $\hat{Y}(X)$ or $\hat{Y}(X, A)$, and knowledge of $\mathcal{D}\left(Y, X, A, \hat{Y}(X, A)\right)$ create (possibly randomized) $\tilde{Y}(\hat{Y}, A)$ satisfying equalized odds

Focusing on binary $Y, \hat{Y}, A \in \{0,1\}$:

- Can set four parameters:

$$P\left(\tilde{Y} = 1 \middle| \hat{Y} = 0, A = 0\right), P\left(\tilde{Y} = 1 \middle| \hat{Y} = 1, A = 0\right),$$
$$P\left(\tilde{Y} = 1 \middle| \hat{Y} = 0, A = 1\right), P\left(\tilde{Y} = 1 \middle| \hat{Y} = 1, A = 1\right)$$

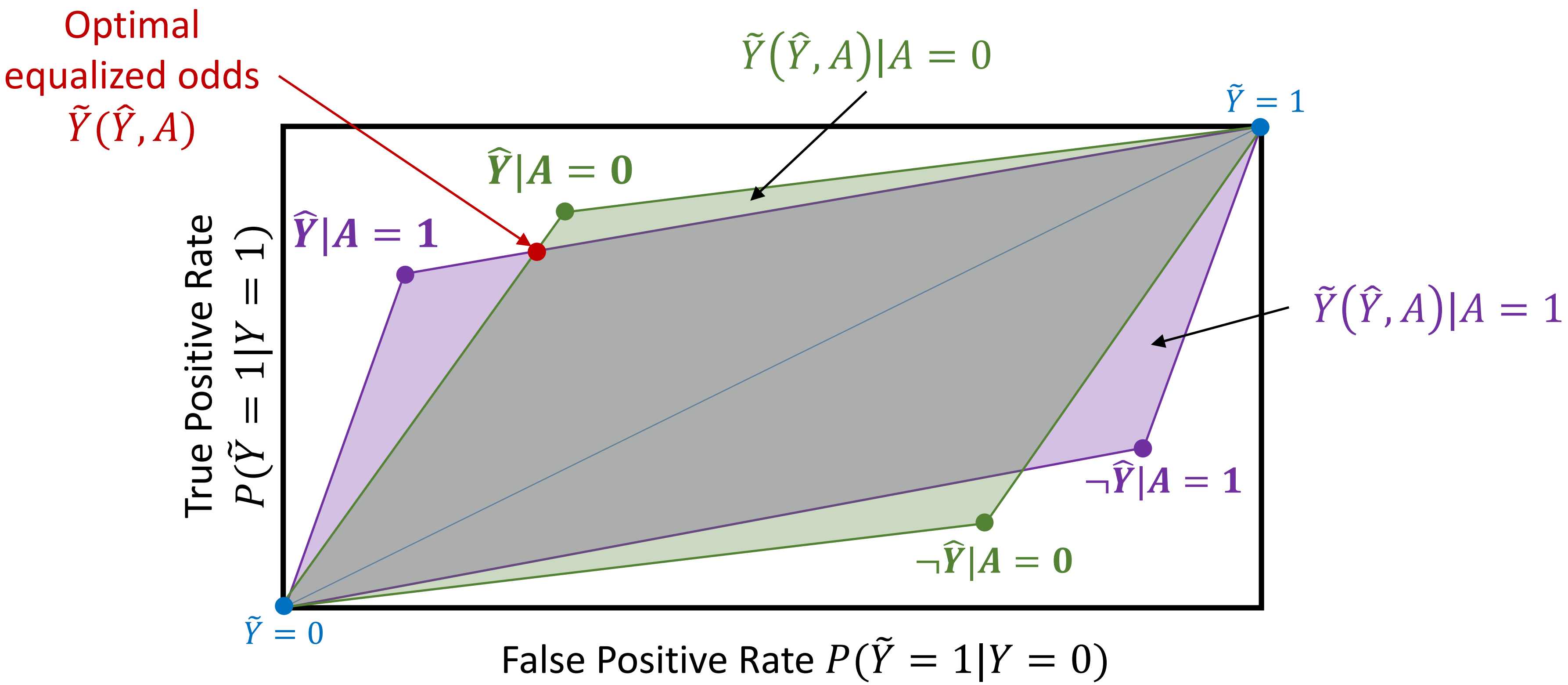
- Need to satisfy two linear constraints:

$$P\left(\tilde{Y} = 1 \middle| Y = 1, A = 0\right) = P\left(\tilde{Y} = 1 \middle| Y = 1, A = 1\right)$$ **True Pos. Rate**

$$P\left(\tilde{Y} = 1 \middle| Y = 0, A = 0\right) = P\left(\tilde{Y} = 1 \middle| Y = 0, A = 1\right)$$ **False Pos. Rate**

➔ Optimize $\mathbb{E}\left[loss\left(\tilde{Y}; Y\right)\right]$ using Linear Programming

Ensuring Equalized Odds



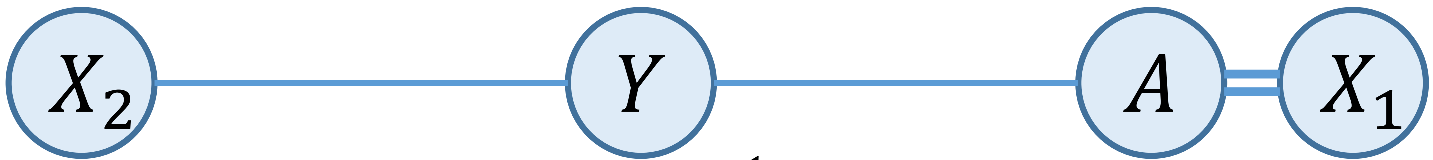
Optimal $\tilde{Y}(\hat{Y}, A)$ is either constant or:

- For $A = 1$ flip from $\hat{Y} = 0$ to $\tilde{Y} = 1$ with prob p
- For $A = 0$ flip from $\hat{Y} = 1$ to $\tilde{Y} = 0$ with prob q

(or the other way around)

Post-Hoc Correction Not Optimal

Example due to Blake Woodworth



$P(X_2 = Y) = 0.9899$

$P(Y = 1) = \frac{1}{2}$

$X_1 = A$

$P(A = Y) = 0.99$

- Optimal unconstrained classifier: $\hat{Y}(X_1, X_2) = X_1$

  ➔ error = $P\left(\hat{Y} \neq Y\right) = 1\%$

- Equalized odds derived from $\hat{Y}, A$ (not learning from features again) must be independent of Y

  ➔ error = $^1/_2$

- Optimal equalized odds predictor : $\hat{Y}(X_1, X_2, A) = X_2$

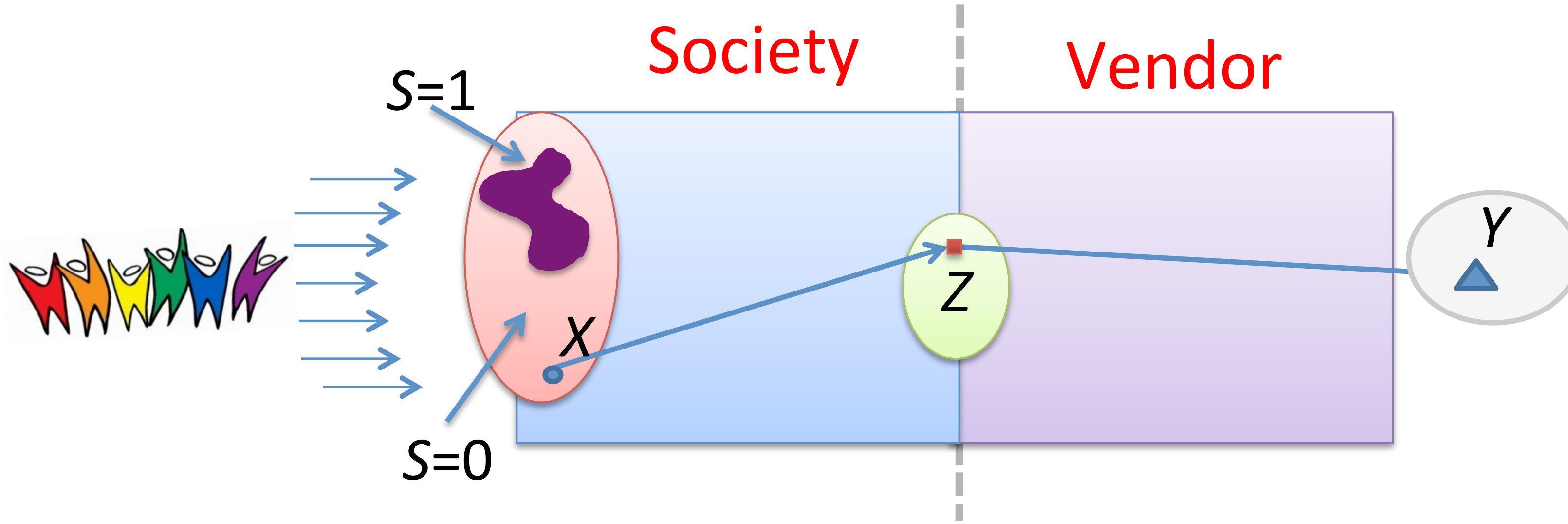
  ➔ error = 1.01%

Learning Fair Representations

Zemel, Yu, Swersky, Pitassi, Dwork
ICML, 2013

- Generalizes to new data: learn general mapping, applies to any individual
- Mapping should satisfy fairness criteria, vendor utility
- Learn prototypes, distances
- Use fair representation for additional classification tasks (transfer learning)
- Working example: dataset of bank loan decisions, protected group (S+) is women

Model Overview



Aims for Z:

1. Lose information about S
   Group Fairness/Statistical Parity: P(Z|S=0) = P(Z|S=1)
2. Preserve information so vendor can max utility

Maximize MI(*Z*, *Y*); Minimize MI(*Z*, *S*)