4th Summer School on Machine Learning
13 July, 2019. IIIT Hyderabad

Fairness in Machine Learning

Chetan Arora
Department of Computer Science and Engineering
Indian Institute of Technology Delhi

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Object Detection

Method	data	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Fast[6]	07++12	68.4	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73.0	55.0	87.5	80.5	80.8	72.0	35.1	68.3	65.7	80.4	64.2
Faster[2]	07++12	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
Faster[2]	07++12+COCO	75.9	87.4	83.6	76.8	62.9	59.6	81.9	82.0	91.3	54.9	82.6	59.0	89.0	85.5	84.7	84.1	52.2	78.9	65.5	85.4	70.2
YOLO[5]	07++12	57.9	77.0	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8
SSD300	07++12	72.4	85.6	80.1	70.5	57.6	46.2	79.4	76.1	89.2	53.0	77.0	60.8	87.0	83.1	82.3	79.4	45.9	75.9	69.5	81.9	67.5
SSD300	07++12+COCO	77.5	90.2	83.3	76.3	63.0	53.6	83.8	82.8	92.0	59.7	82.7	63.5	89.3	87.6	85.9	84.3	52.6	82.5	**74.1**	**88.4**	74.2
SSD512	07++12	74.9	87.4	82.3	75.8	59.0	52.6	81.7	81.5	90.0	55.4	79.0	59.8	88.4	84.3	84.7	83.3	50.2	78.0	66.3	86.3	72.0
SSD512	07++12+COCO	**80.0**	**90.7**	**86.8**	**80.5**	**67.8**	**60.8**	**86.3**	**85.5**	**93.5**	**63.2**	**85.7**	**64.4**	**90.9**	**89.0**	**88.9**	**86.8**	**57.2**	**85.1**	72.8	**88.4**	**75.9**

Table 4: **PASCAL VOC2012 `test` detection results.** Fast and Faster R-CNN use

SSD: Single Shot MultiBox Detector

W Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A. C. Berg


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fairness in ML Systems

- Names that are associated with black people are found to be significantly more associated with unpleasant than with pleasant terms, compared to names associated with whites.

- The models learned on such text data for opinion or sentiment mining have a possibility of inheriting the prejudices reflected in the data.

A. Caliskan-Islam, J. J. Bryson, and A. Narayanan.
Semantics derived automatically from language corpora necessarily contain human biases.

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fairness in ML Systems

- Strong ethnic bias in COMPAS score: a predictive model for the “risk of crime recidivism”

- According to the score:
  - A black who did not re-offend were classified as high risk twice as much as whites who did not re-offend, and
  - White repeat offenders were classified as low risk twice as much as black repeat offenders

http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

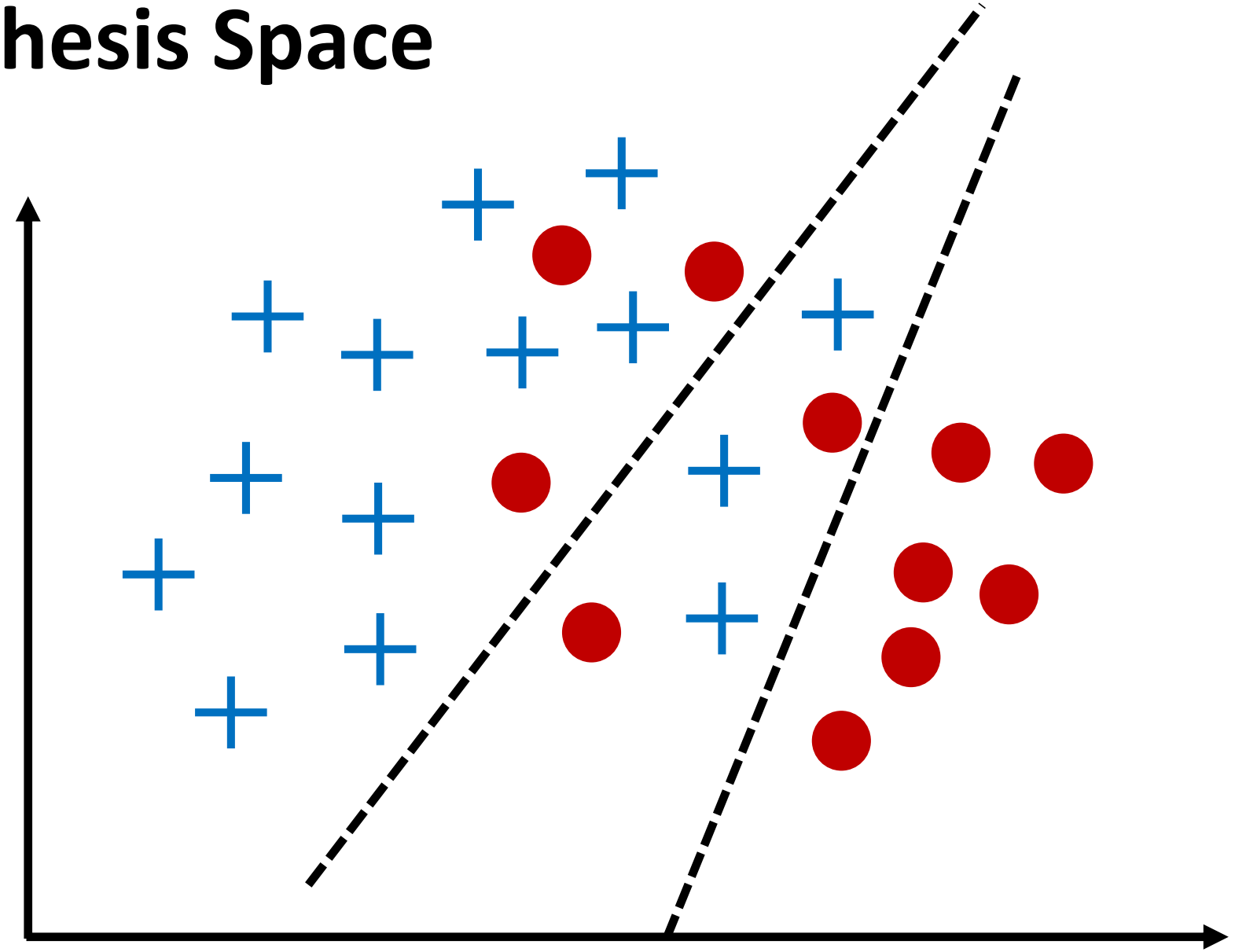
Fairness in ML Systems

- In 2016, the software used to determine the areas of the US to which Amazon would offer free same-day delivery, unintentionally restricted minority neighborhoods from participating in the program (often when every surrounding neighborhood was allowed)

http://www.techinsider.io/how-algorithms-can-be-racist-2016-4

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Hypothesis Space




Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Inductive Bias

Set of assumptions that the learner uses to predict outputs given inputs that it has not encountered

- **Minimum cross-validation error:** When trying to choose among hypotheses, select the hypothesis with the lowest cross-validation error.

- **Maximum margin:** when drawing a boundary between two classes, attempt to maximize the width of the boundary.


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Inductive Bias

Set of assumptions that the learner uses to predict outputs given inputs that it has not encountered

- **Minimum features:** Delete a feature unless there is good evidence that it is useful. This is the assumption behind feature selection algorithms.

- **Nearest neighbors:** Assume that most of the cases in a small neighborhood in feature space belong to the same class. Used in the k-nearest neighbors and many manifold learning algorithms.


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Undesirable Biases in ML Systems

- **Selection Bias**
  - Bias introduced by the selection of data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population
  - **Sampling Bias:** Bias in which a sample is collected in such a way that some members of the intended population have a lower sampling probability than others

- **Reporting Bias**
  - Selective revealing or suppression of information. Authors under-reporting unexpected or undesirable experimental results


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Differentiation Vs Discrimination

- Differentiation is the basis of prediction in any ML system.

- Unjustified basis for differentiation is Discrimination

  - **Practical irrelevance**
    - Gender, Religion, Caste, Financial Status etc.

  - **Moral irrelevance**
    - Caste, Disability etc.

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Legal Regulations Against Discrimination

- Credit
- Education
- Employment
- Housing
- Public Accommodation

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Doctrines of Discrimination

1. Disparate Treatment

- **Formal**
  - Using protected attributes explicitly
- **Intentional**
  - Using proxies for the protected attributes

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Doctrines of Discrimination

2. Disparate Impact

- **Unjustified**
  - Predicted outcome is significantly different for two groups
- **Avoidable**
  - There is way to achieve the same score with lesser disparate impact

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Targets of Discrimination Doctrines

- **Disparate Treatment**
  - Procedural fairness
  - Equality of opportunity

- **Disparate Impact**
  - Distributive justice
  - Minimized inequality of outcome

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Discrimination Vs Affirmative Action

- **Equality of opportunity**
  - Ensure that decision-making treats similar people similarly on the basis of relevant features, given their current degree of similarity
  - Organize society in such a way that people of equal talents and ambition can achieve equal outcomes over the course of their lives

- **Affirmative Action**
  - Equality of opportunity should be balanced with the need to treat seemingly dissimilar people similarly, on the belief that their current dissimilarity is the result of past injustice

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Tradeoff: Disparate Treatment and Impact

- Bias in the data may imply that there will be inherent contradiction or tradeoff between need to treat each group similarly or need to have similar outcome for each group.

- Visible regularly in the case of affirmative action

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Does ML Prevent Disparate Treatment?

- Automated decision leaves no scope of human discretion.
- Model learns what the data supports
- Protected attributes can be withheld

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

How Machines Learn to Discriminate

- **Skewed sample**
  - School A produces good students
  - Feedback Loop
- **Tainted examples**
  - Unreliable labels
  - Group A are tax evaders.
- **Proxies**
  - Considering features that are correlated with protected attribute

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

How Machines Learn to Discriminate

- **Limited features**
  - Features may be less informative or less reliably collected for certain parts of the population
  - A feature set that supports accurate predictions for the majority group may not for a minority group
  - Different models with the same reported accuracy can have a very different distribution of error across population

- **Sample size disparity**
  - Objective functions are biased against the minority classes

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Handling Discrimination in ML Systems

- **Discovering unobserved differences in performance**
  - Skewed samples
  - Tainted examples

- **Coping with observed differences in performance**
  - Limited features
  - Sample size disparity

- **Understanding the causes of disparities in predicted outcome**
  - Proxies

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Problem Formulation

Example: A company wants to hire a software engineer (SWE) and is going to advertise for the same. An ML system needs to predict which persons to show the advertisement based upon if he/she is currently a SWE.

- X: features of an individual (browsing history etc.)
- A: sensitive or protected attributes (gender etc.)
- $C = c(X, A)$: predicted score/class (show ad or not). Also denoted as $\hat{Y}$
- Y: target variable (whether the person is a SWE)

Notation: $\mathbb{P}_a\{E\} = \mathbb{P}\{E|A = a\}$

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Problem Formulation

- Score function is any random variable $R = r(X, A) \in [0,1]$
- Can be turned into (binary) predictor by thresholding
- Example: Bayes optimal score given by $r(x, a) = \mathbb{E}[Y|X = x, A = a]$

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fundamental Criterion for Fairness

- **Independence:** C independent of A
- **Separation:** C independent of A conditional on Y
- **Sufficiency:** Y independent of A conditional on C

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

First criterion: Independence

- Require C and A to be independent. Denoted as $C \perp A$

- That is, for all groups g_1, g_2 and all predictions c:

$$\mathbb{P}_{g_1}\{C = c\} = \mathbb{P}_{g_2}\{C = c\}$$

- When C is a binary 0/1-variable:

$$\mathbb{P}_{g_1}\{C = 1\} = \mathbb{P}_{g_2}\{C = 1\}, \forall a, b$$

- Also called **demographic parity**, or **statistical parity**

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

First criterion: Independence

Approximate versions

- $\frac{\mathbb{P}_a\{C=1\}}{\mathbb{P}_b\{C=1\}} \geq 1-\epsilon$

- Also known in the legal context as “80% Rule”. Declare unfair if the impact differs by more than 20%

- Additive version: $|\mathbb{P}_a\{C=1\} - \mathbb{P}_b\{C=1\}| \leq \epsilon$

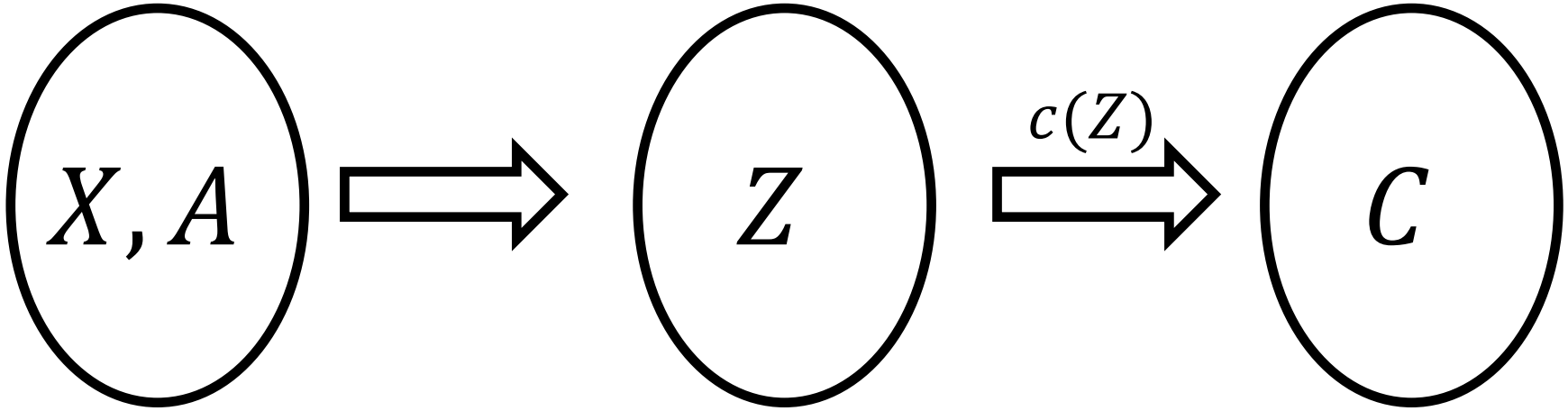
Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Independence

Representation Learning Approach

- Maximize *Mutual Information* between X and latent representation Z while minimizing the same between A and Z



$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} \mathbb{P}(x,y) \log \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)\mathbb{P}(y)}$$

R. Zemel, Y. Wu, K. Swersky T. Pitassi, and C. Dwork
Learning Fair Representations. ICML 2013


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Problems with Independence based Fairness

- Ignores possible correlation between Y and A.
  - There could be less woman software engineers.
- Rules out perfect predictor $C = Y$.
- Permits laziness: Accept the qualified in one group, random people in other
- Allows to trade false negatives for false positives.
- Conflates desirable long-term goal with algorithmic constraint

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

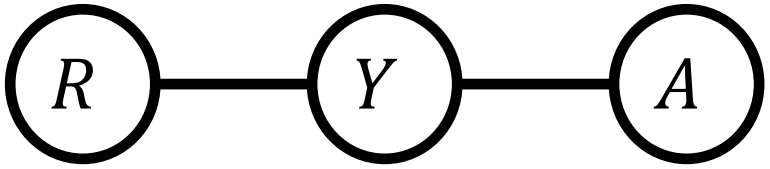

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Second criterion: Separation

- Require R and A to be independent, conditional on target variable Y
- Denoted as $R \perp A \mid Y$
- For all groups a, b and all values r and y:

$$\mathbb{P}_a\{R = r \mid Y = y\} = \mathbb{P}_b\{R = r \mid Y = y\}$$

- PGMs: Random variable R **separated** from A if $R \perp A \mid Y$



Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Desirable Properties of Separation

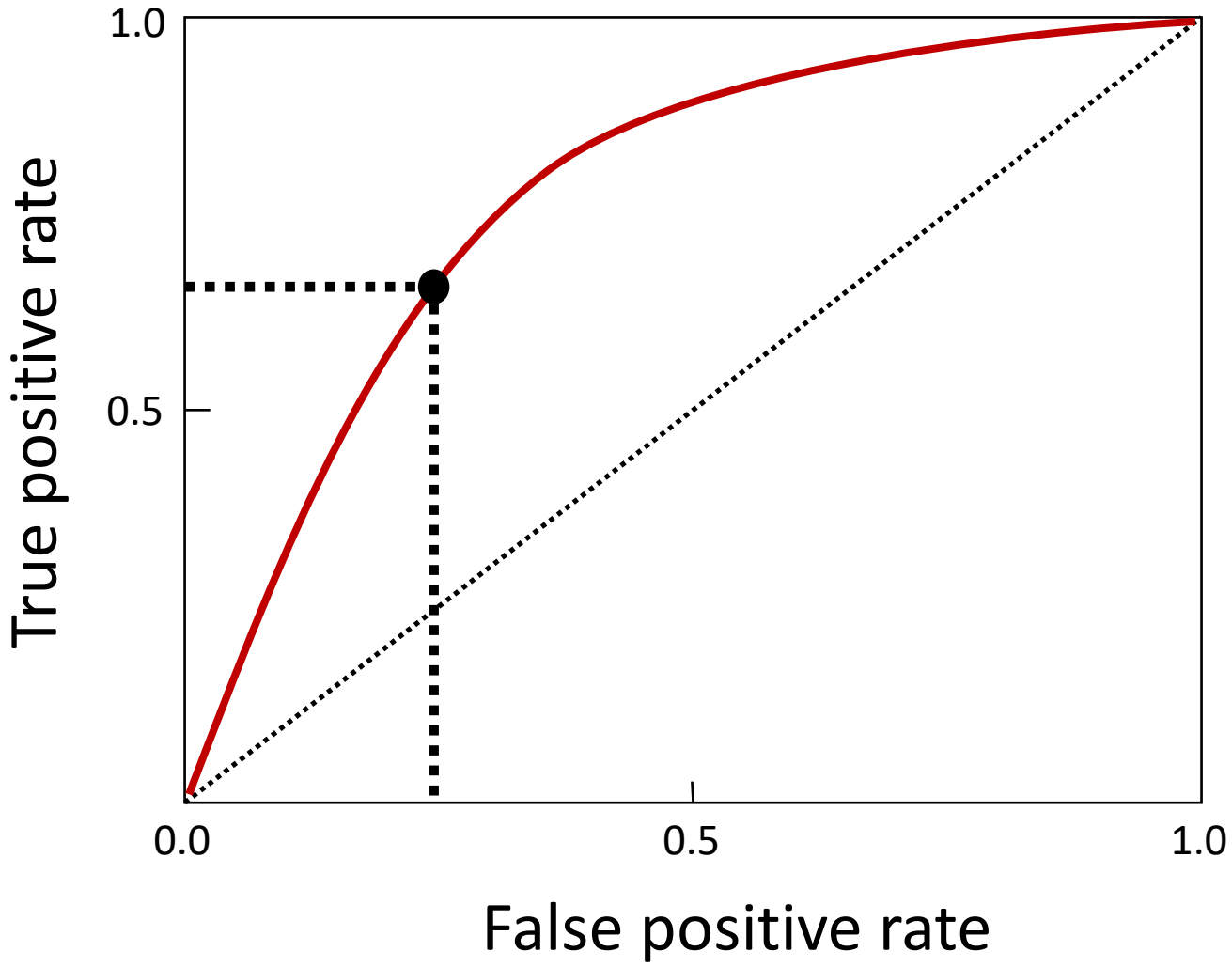
$$\mathbb{P}_a\{R = r \mid Y = y\} = \mathbb{P}_b\{R = r \mid Y = y\}$$

- Optimality compatibility: $R = Y$ is allowed
- Penalizes laziness: Incentive to reduce errors uniformly in all groups
- Recall, none of the above is achieved by Independence.
- Assume Y is unbiased: Problematic with tainted and skewed samples

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

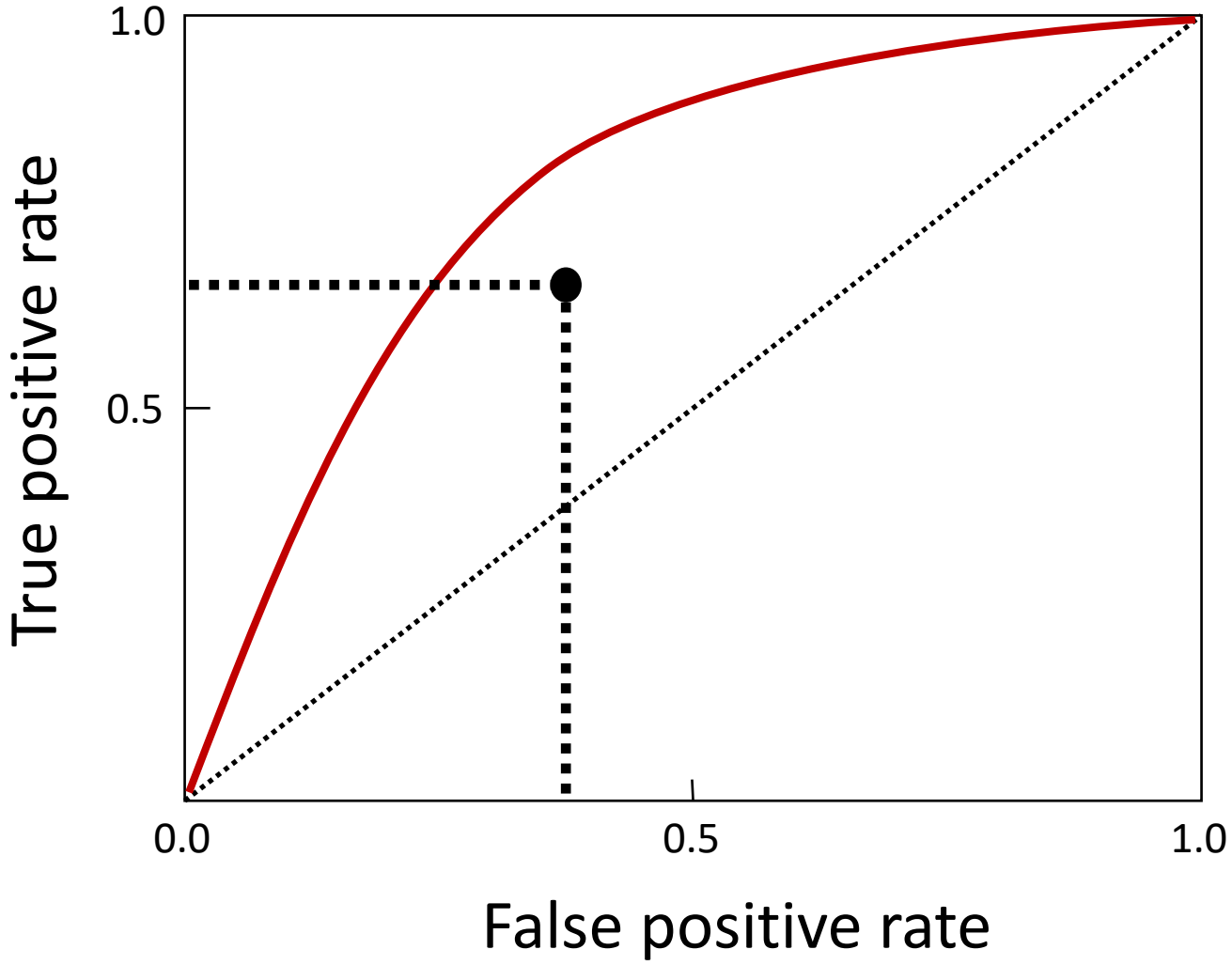
Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Separation



Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

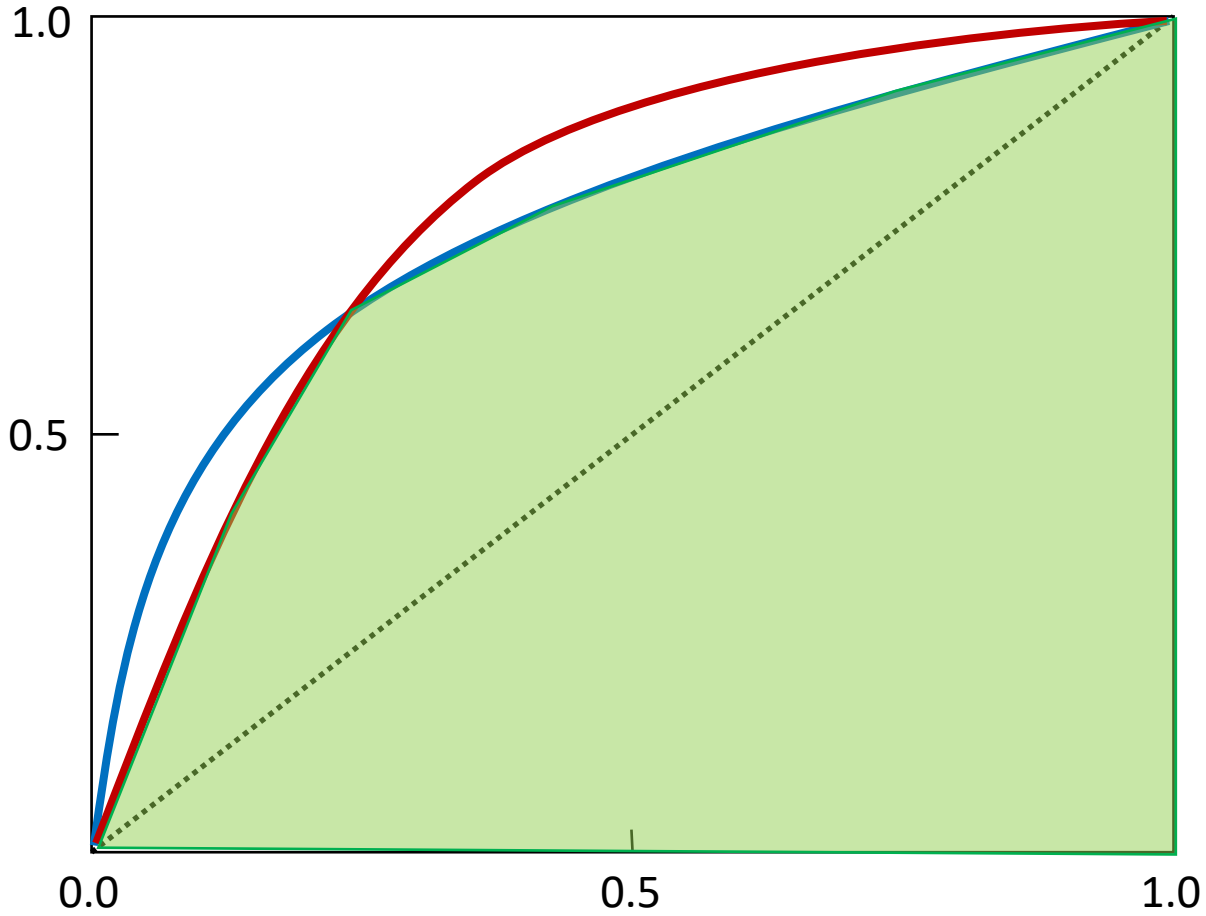
Achieving Fairness by Separation



Create Random False Positives

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Separation

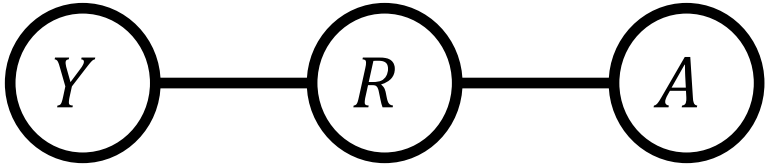


Moritz Hardt, Eric Price, Nathan Srebro
Equality of Opportunity in Supervised Learning. NIPS 2016

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Third criterion: Sufficiency

- Random variable R is sufficient for A if $Y \perp A \mid R$



- For the purpose of predicting Y, we don't need to see A when we have R

- For the purpose of showing the SWE advertisement, I don't need to know gender, if I already know the programming score.

- For the purpose of approving loan, I don't need to the know the race, when I know the credit rating

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Sufficiency

$$Y \perp A \mid R$$

- Sufficiency satisfied by Bayes optimal score:

$$r(X, A) = \mathbb{E}[Y \mid X = x, A = a]$$

- Can be achieved using **calibration by group**: $\mathbb{P}\{Y = 1 \mid R = r, A = a\} = r$

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Achieving Fairness by Sufficiency

$$\mathbb{P}\{Y = 1 \mid R = r, A = a\} = r$$

Plat Scaling:

- Given uncalibrated score R, fit a sigmoid function: $S = \frac{1}{1+\exp(\alpha R+\beta)}$

- Minimize log loss: $-\mathbb{E}[Y \log S + (1 - Y)\log(1 - S)]$
  - Objective function nudges S to take similar value as Y
  - $\frac{Y}{S} - \frac{1-Y}{1-S} = 0. \quad \Rightarrow \frac{Y}{S} = \frac{1-Y}{1-S} \quad \Rightarrow Y(1 - S) = S(1 - Y). \quad \Rightarrow Y = S$

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Fairness Tradeoffs

Any two of the three criteria we saw are mutually exclusive except in degenerate cases.

For example:

- If $A \not\perp Y$ and $R \not\perp Y$, then either independence or separation holds, not both.

- Recall that Separation $\Rightarrow R \perp A \mid Y$, and Independence $\Rightarrow R \perp A$

- $\frac{\mathbb{P}\{R,A\}}{\mathbb{P}\{Y\}} = \frac{\mathbb{P}\{R\}}{\mathbb{P}\{Y\}} \times \frac{\mathbb{P}\{A\}}{\mathbb{P}\{Y\}}$. $A \not\perp Y$ and $R \not\perp Y \Rightarrow \mathbb{P}(Y)$ can not be removed in R.H.S.

Moritz Hardt, Eric Price, Nathan Srebro
Equality of Opportunity in Supervised Learning. NIPS 2016

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

The COMPAS Debate

Bernard Parker, left, was rated high risk; Dylan Fugett, right, was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

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

Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica May 23, 2016


Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Essence of COMPAS debate

ProPublica's main charge: Black defendants face higher false positive rate.

Northpointe's main defence: Scores are calibrated by group.

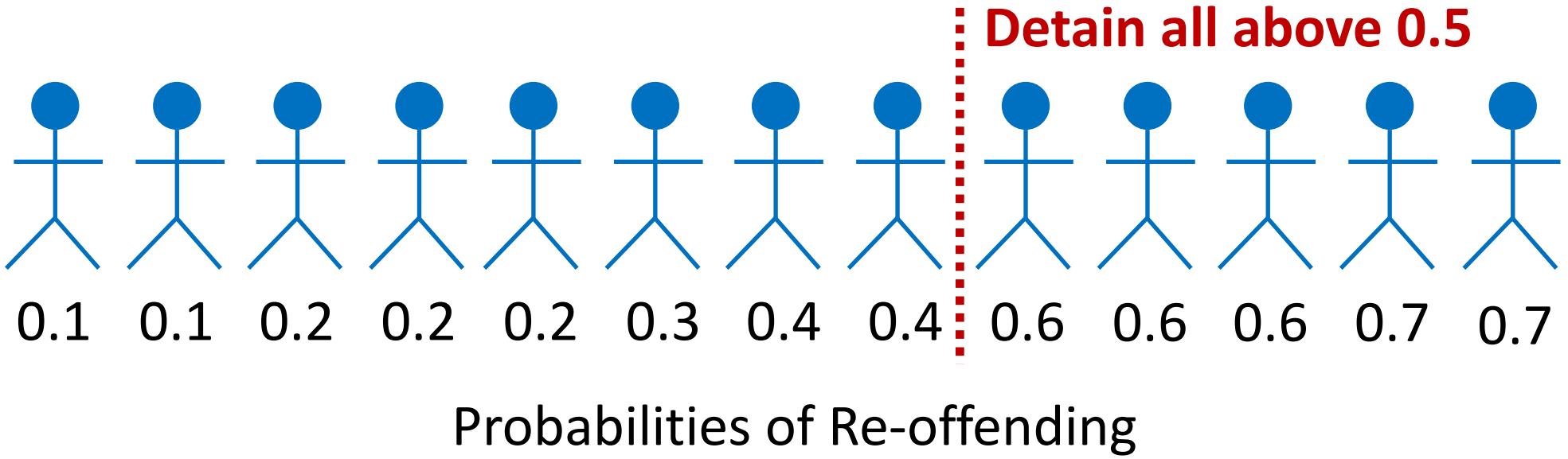
- Recall that sufficiency $\Rightarrow Y \perp A \mid R$, and can be achieved using **calibration by group**: $\mathbb{P}\{Y = 1 \mid R = r, A = a\} = r$ implemented by Platt Scaling

Neither calibration nor equality of false positive rates rule out blatantly unfair practices

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

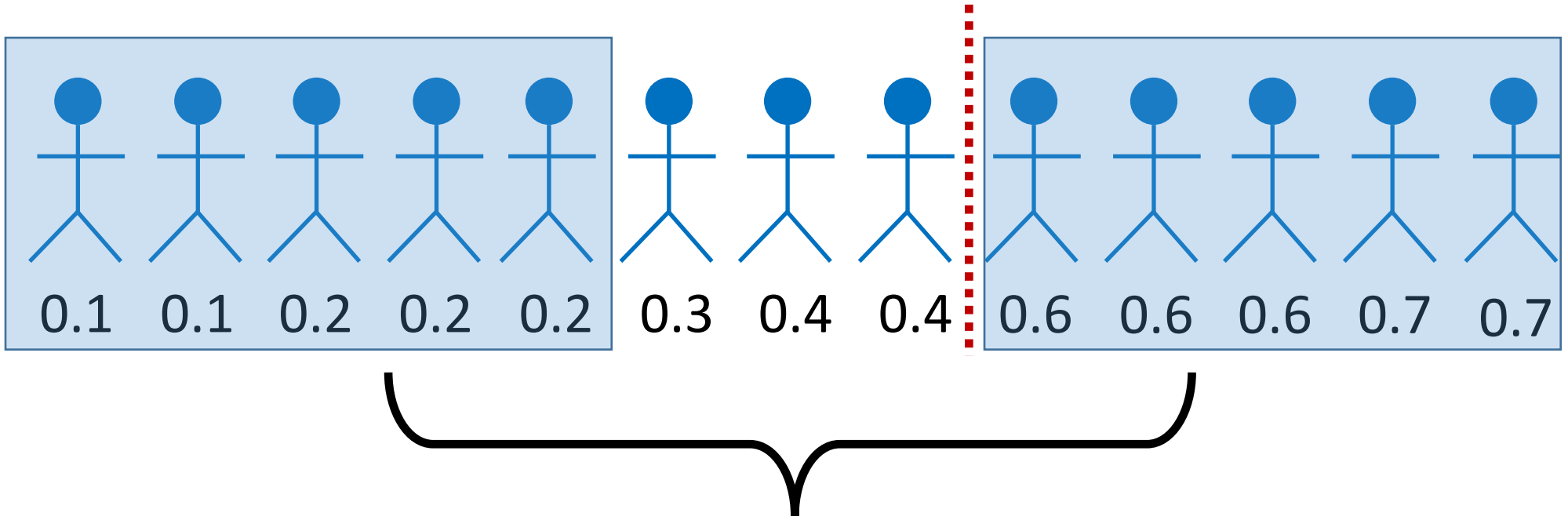
Calibration in Insufficient



Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

Calibration is Insufficient

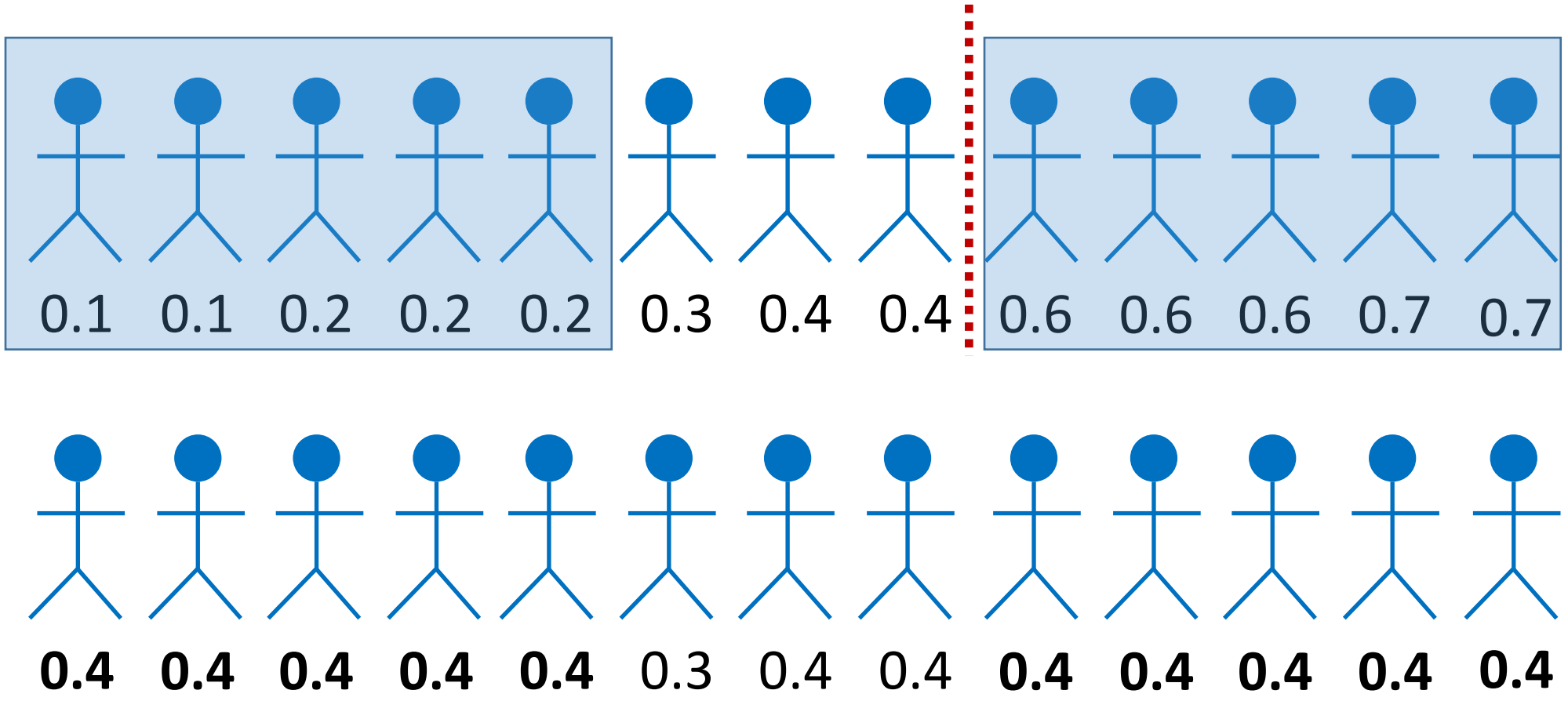


Average Re-offending Rate = 0.4

Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

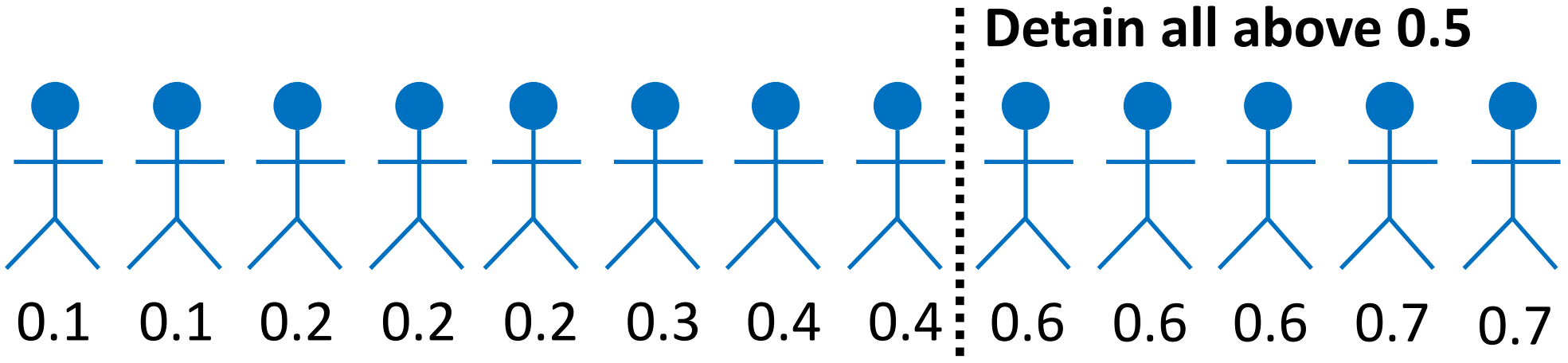
Calibration is Insufficient



Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

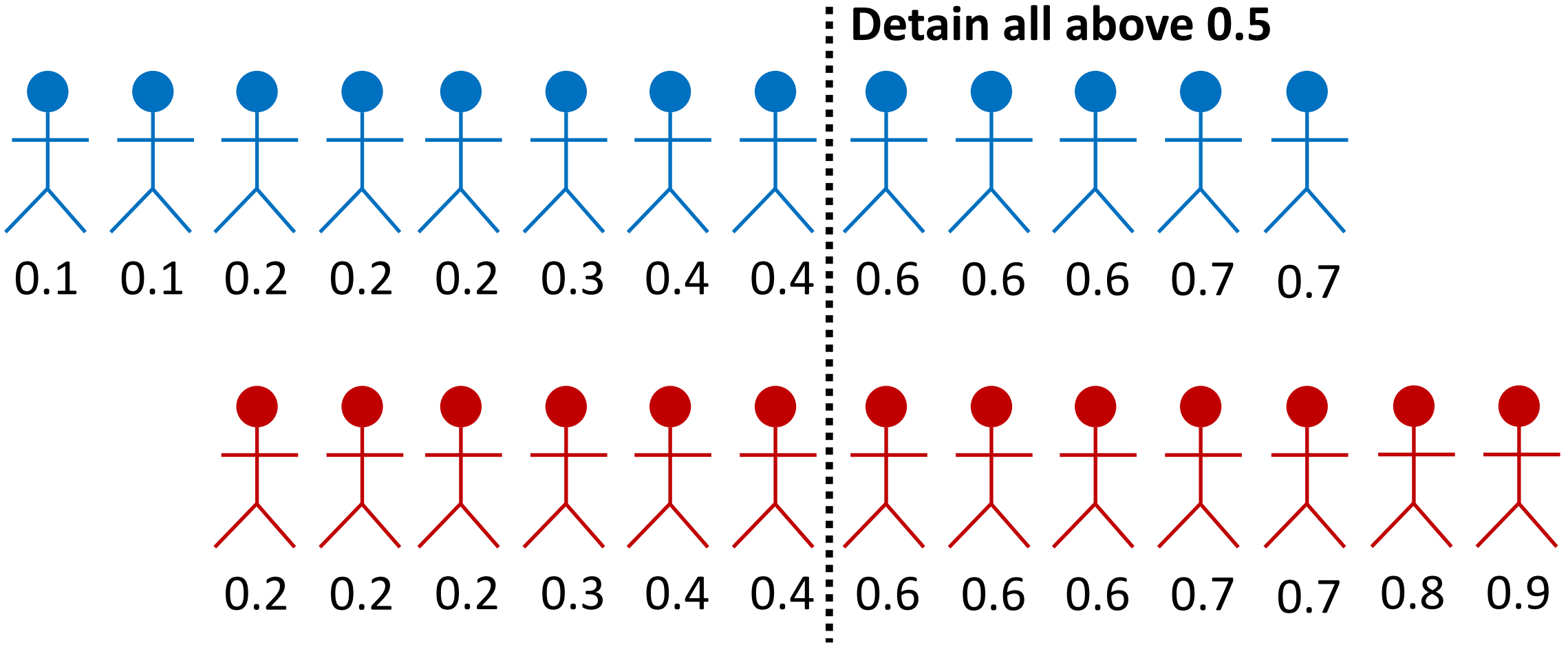
False Positive Rate



Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

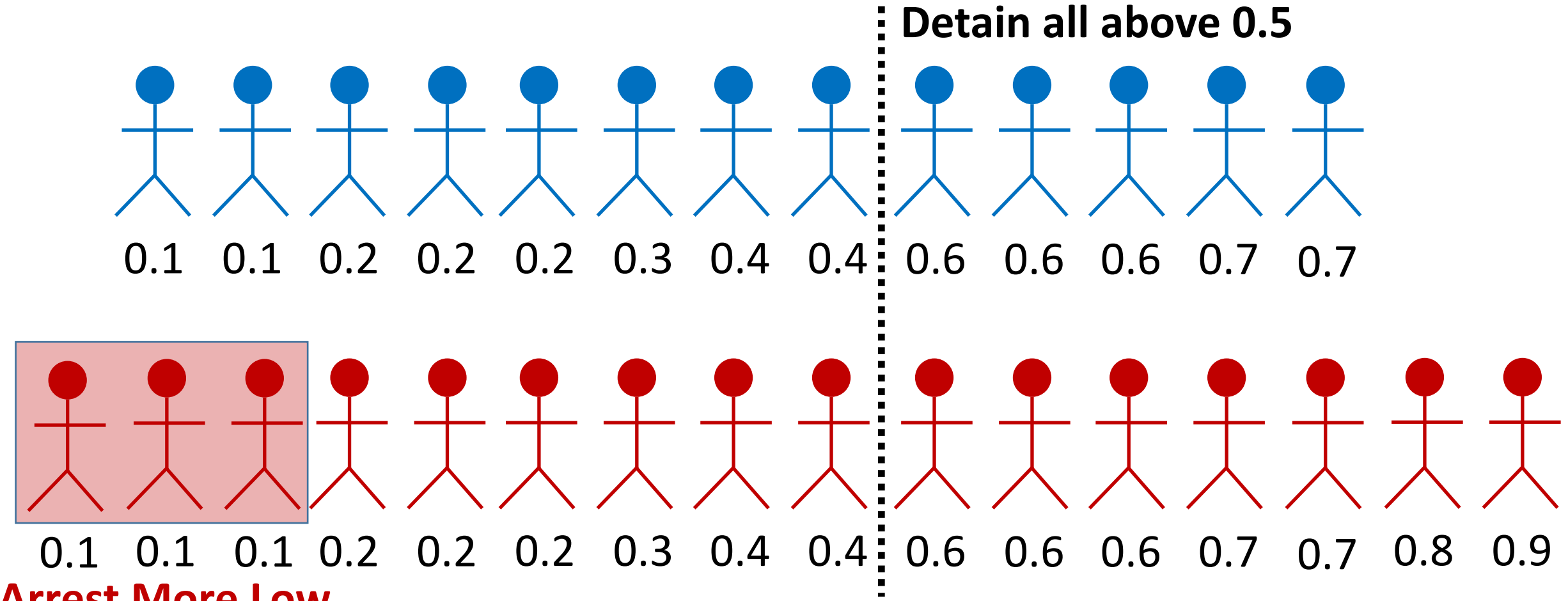
False Positive Rate



Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial

Chetan Arora
Department of Computer Science and Engineering, IIT Delhi

False Positive Rate



Solon Barocas and Moritz Hardt
Fairness in Machine Learning. NIPS 2017 Tutorial