FAIR MACHINE LEARNING

The Measure and Mismeasure of Fairness:

A Critical Review of Fair Machine Learning

Sam Corbett-Davies
Stanford University

Sharad Goel
Stanford University

August 14, 2018

STRUCTURE

- MOTIVATION
- FAIRNESS
- ALGORITHMIC FAIRNESS
  - RISK ASSESSMENT
    - BASICS & ASSUMPTIONS
  - FORMAL DEFINITIONS OF FAIRNESS
    - LIMITS
- PROBLEMS WITH DESIGNING FAIR ALGORITHMS
- SUMMARY

MOTIVATION

Quelle: https://www.newyorker.com/magazine/2019/04/15/who-belongs-in-prison

Quelle: https://www.wn.de/Service/Verbrauchertipps/Kredite-Die-Top-10-Verwendungszwecke-fuer-einen-Privatkredit

schufa

Quelle: https://de.wikipedia.org/wiki/Datei:Schufa_Logo.svg

FAIRNESS

- fairness = !(discrimination)
- fair algorithms are algorithms that do not discriminate
- *disparate impact*
  = decision provokes unjustified differences between groups

ALGORITHMIC FAIRNESS: RISK ASSESSMENT

RISK ASSESSMENT: FLOW

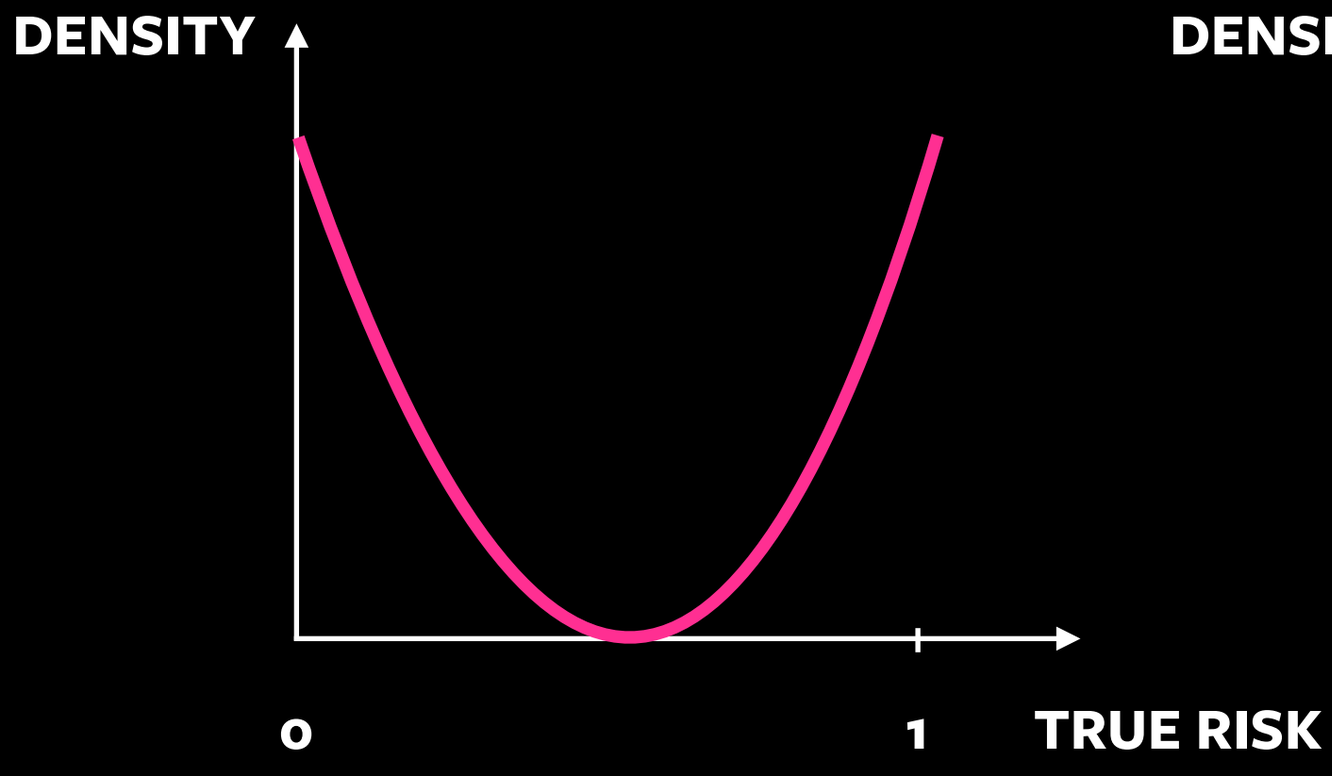
X = (age, gender, race, credit history)

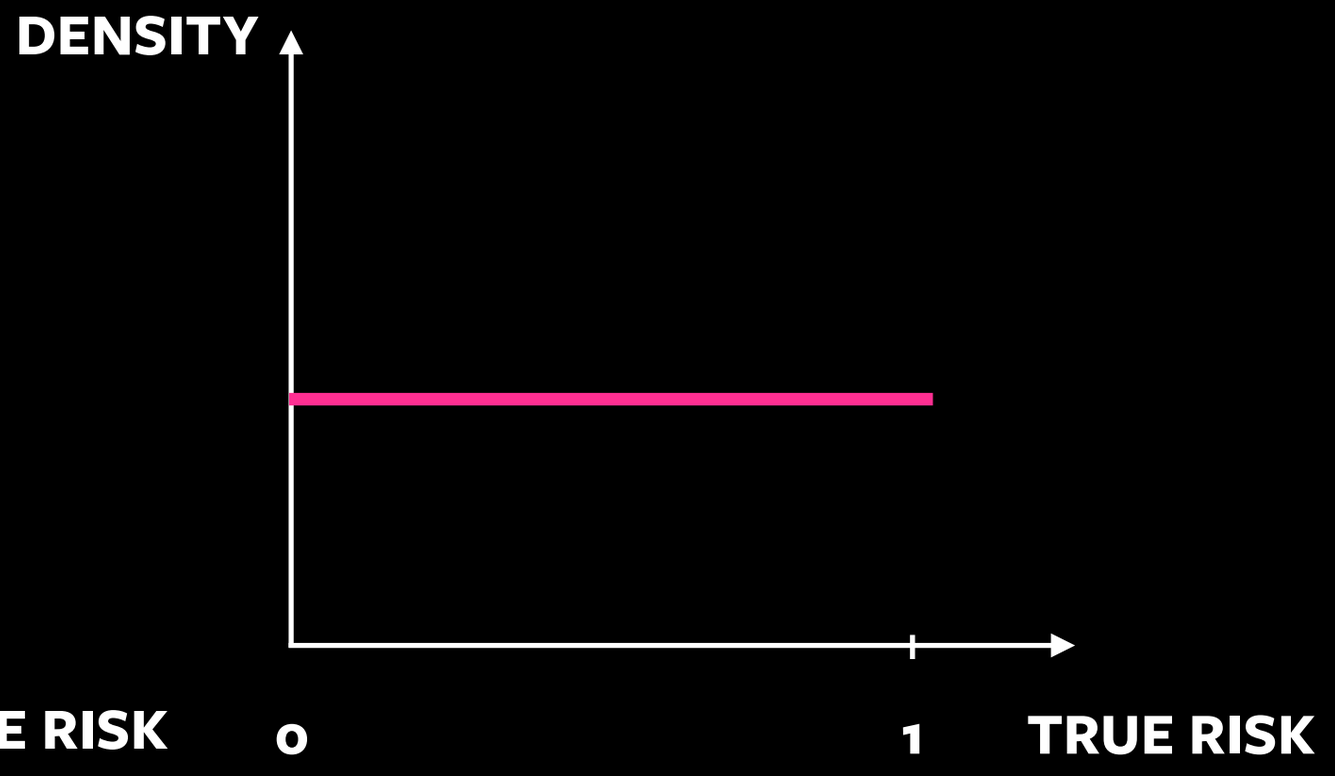
x = (xp, xu)

RISK ASSESSMENT: ASSUMPTION

TRUE RISK: r(x) = Pr(Y=1 | X = x)

RISK ASSESSMENT:
RISK DISTRIBUTIONS

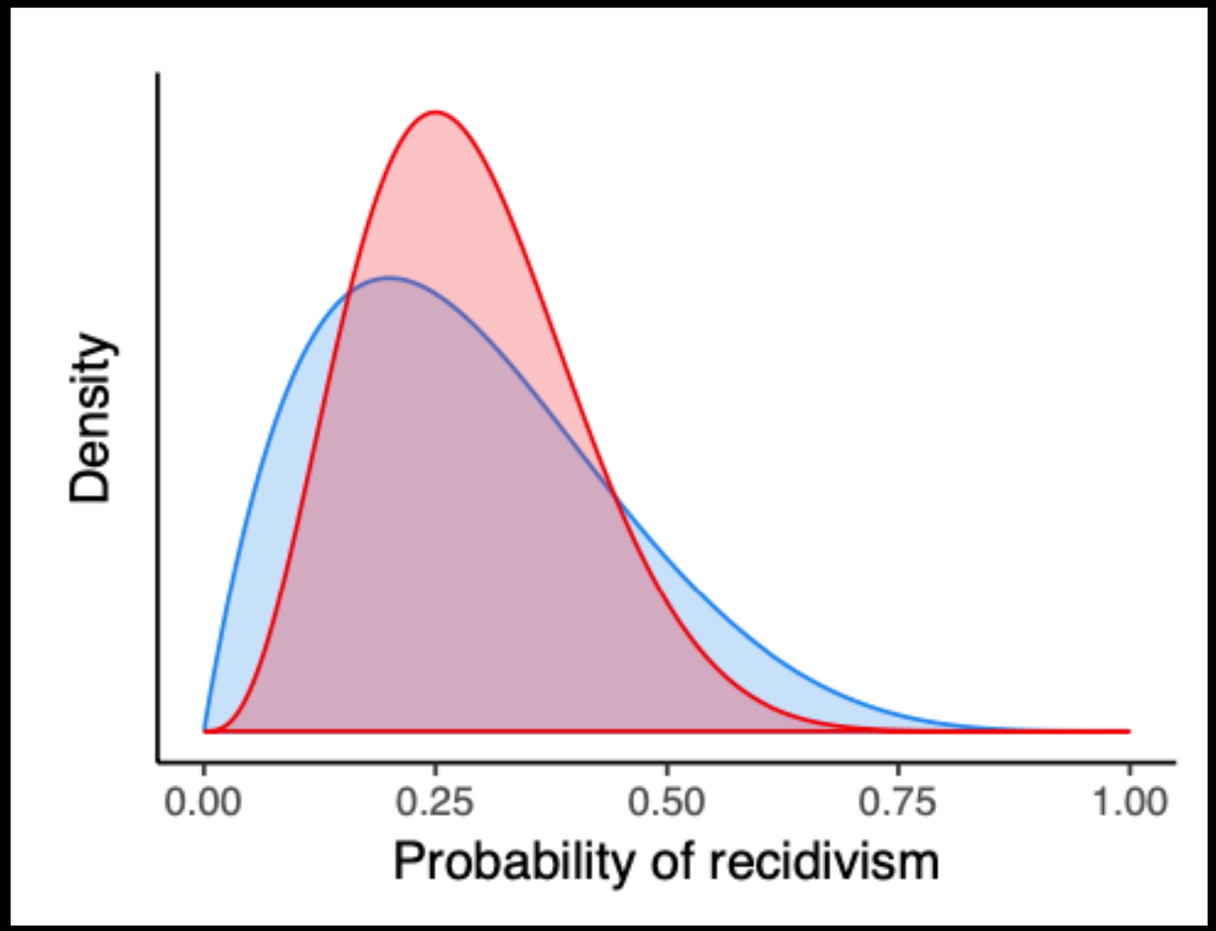




RISK ASSESSMENT: RISK DISTRIBUTIONS

- **ASSUMPTIONS**

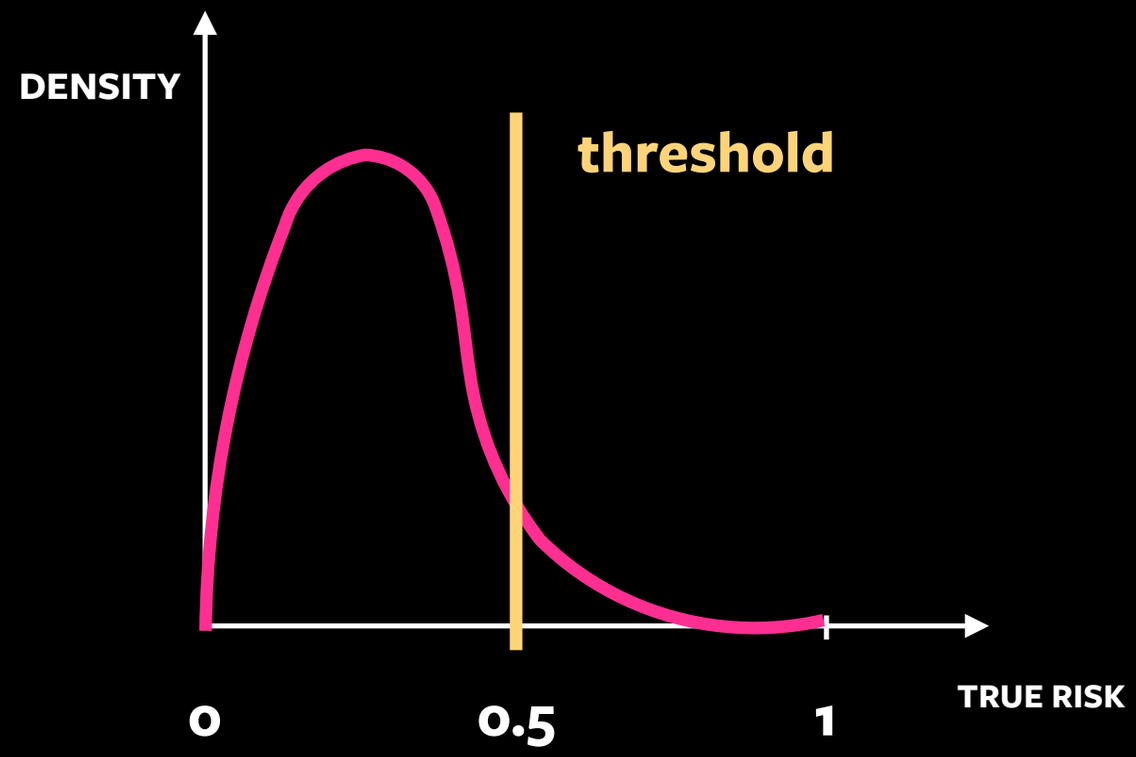
- 2 groups
- same mean (fixed for all groups)
- different distribution
- distribution depends on x



Quelle: S. Corbett-Davies, S. Goel, 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *Computer Research Repository*.

RISK ASSESSMENT: FORMING DECISIONS

- **DISTRIBUTION -> DECISION**
  - apply threshold on risk
  - trade-off costs and benefits
  - maximum utility of decision = sweet spot between costs and benefits



RISK ASSESSMENT: UTILITY FUNCTIONS

- **FIND OPTIMAL DECISION**
  - maximize utility

- $u(o) = b_{oo} * (1-r(x)) - c_{o1} * r(x)$

decision

benefit of correct positive / negative decision

costs of incorrect positive / negative decision

- $u(1) = b_{11} * r(x) - c_{10} * (1 - r(x))$

RISK ASSESSMENT: THRESHOLD RULES

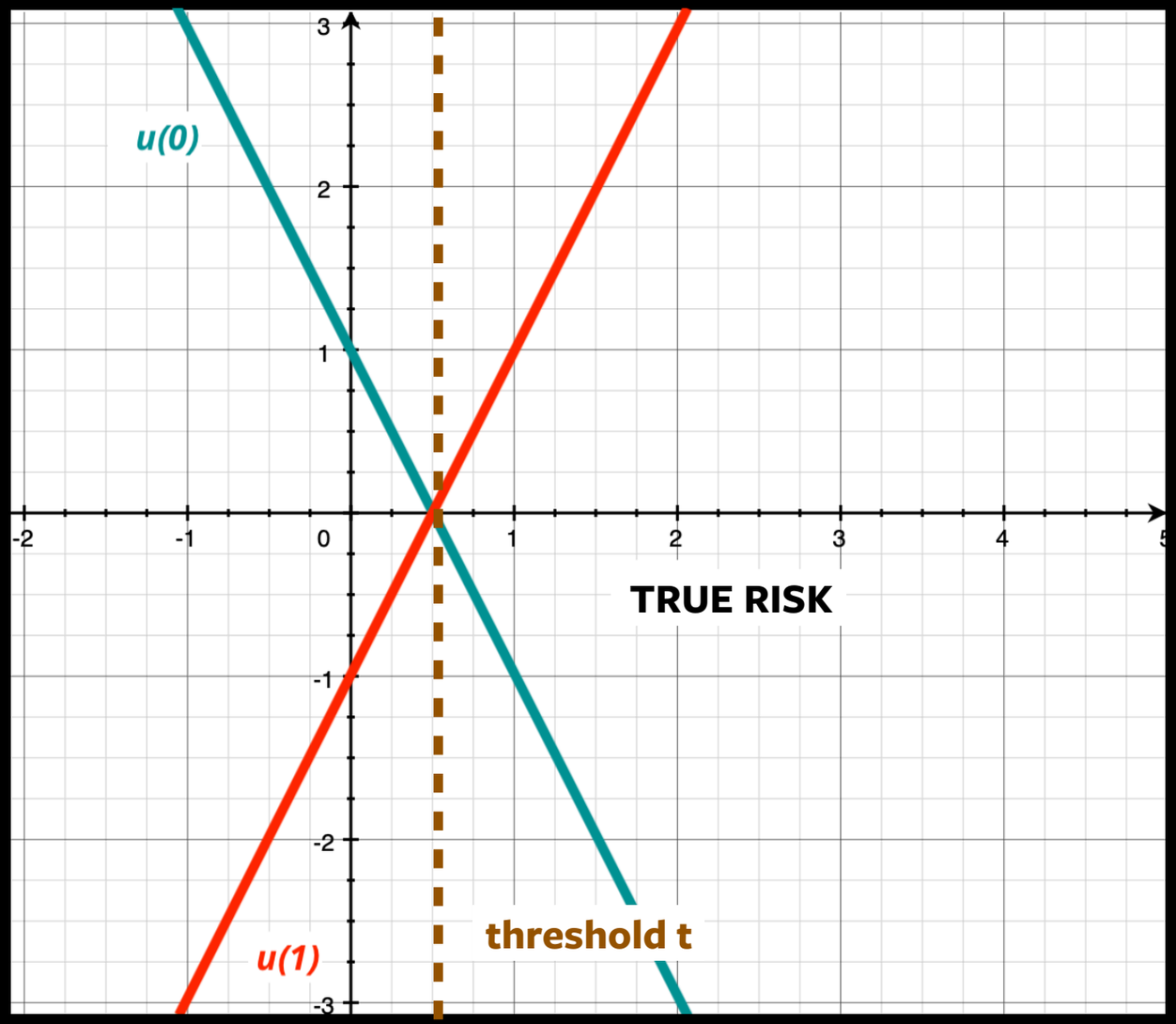
- **THRESHOLD RULES**
- from utility functions:

$$u(1) \geq u(0) \leftrightarrow r(x) \geq (b_{00} + c_{10}) / (b_{00} + b_{11} + c_{01} + c_{10})$$

threshold that produces optimal decisions

RISK ASSESSMENT: THRESHOLD RULES

- **UTILITY FUNCTIONS & THRESHOLD RULES**



PROBLEM?

with
$b_{00} = b_{11} = c_{10} = c_{01} = 1$

ALGORITHMIC FAIRNESS:
FORMAL DEFINITIONS

FORMAL DEFINITIONS: ANTI-CLASSIFICATION

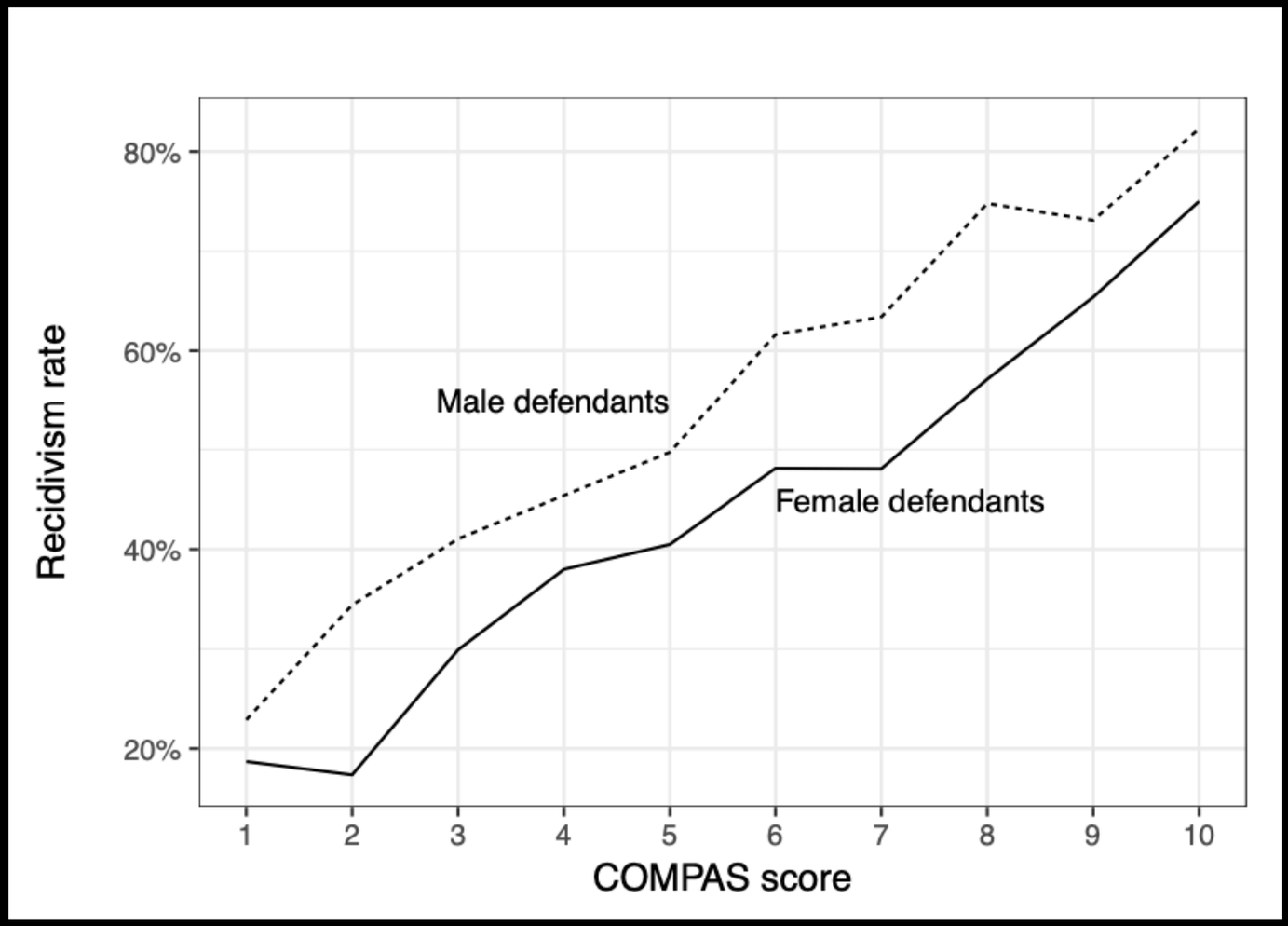
- **IDEA**
  - decisions should not explicitly depend on protected attributes
- forbids use of protected features in X

FORMAL DEFINITIONS: ANTI-CLASSIFICATION

- **LIMITS OF ANTI-CLASSIFICATION**
  - implicit dependence on included features
  - sometimes explicit use of group membership needed for fair decision

ALGORITHMIC FAIRNESS: ANTI-CLASSIFICATION

- **LIMITS OF ANTI-CLASSIFICATION**
  - excluding could lead to unjustified disparate impact
    - example gender-neutral vs. gender-specific recidivism rate



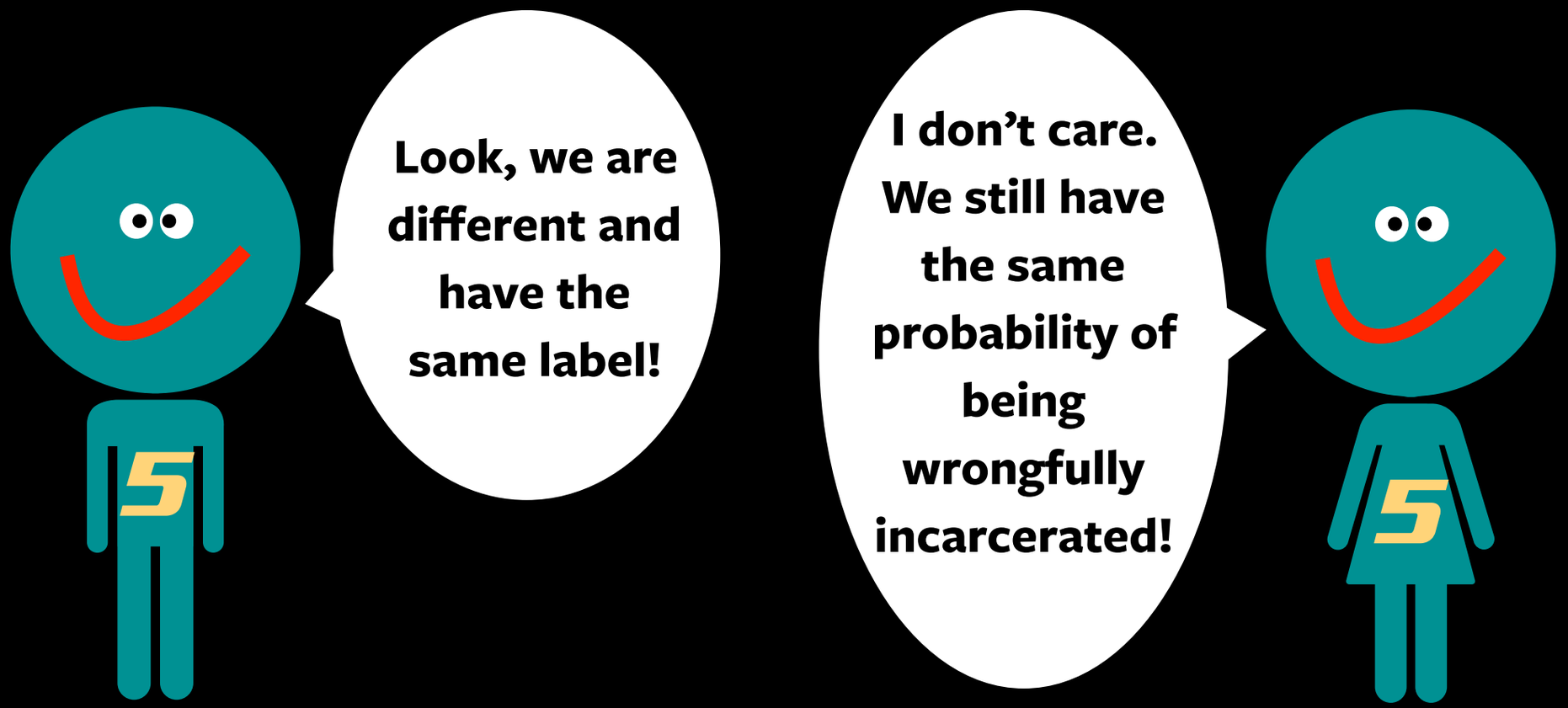
Quelle: S. Corbett-Davies, S. Goel, 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *Computer Research Repository.*

FORMAL DEFINITIONS: CLASSIFICATION PARITY

- **IDEA**
  - all groups have the same classification errors
    - classification errors: false positive / negative rates, precision, recall, proportion of positive decisions

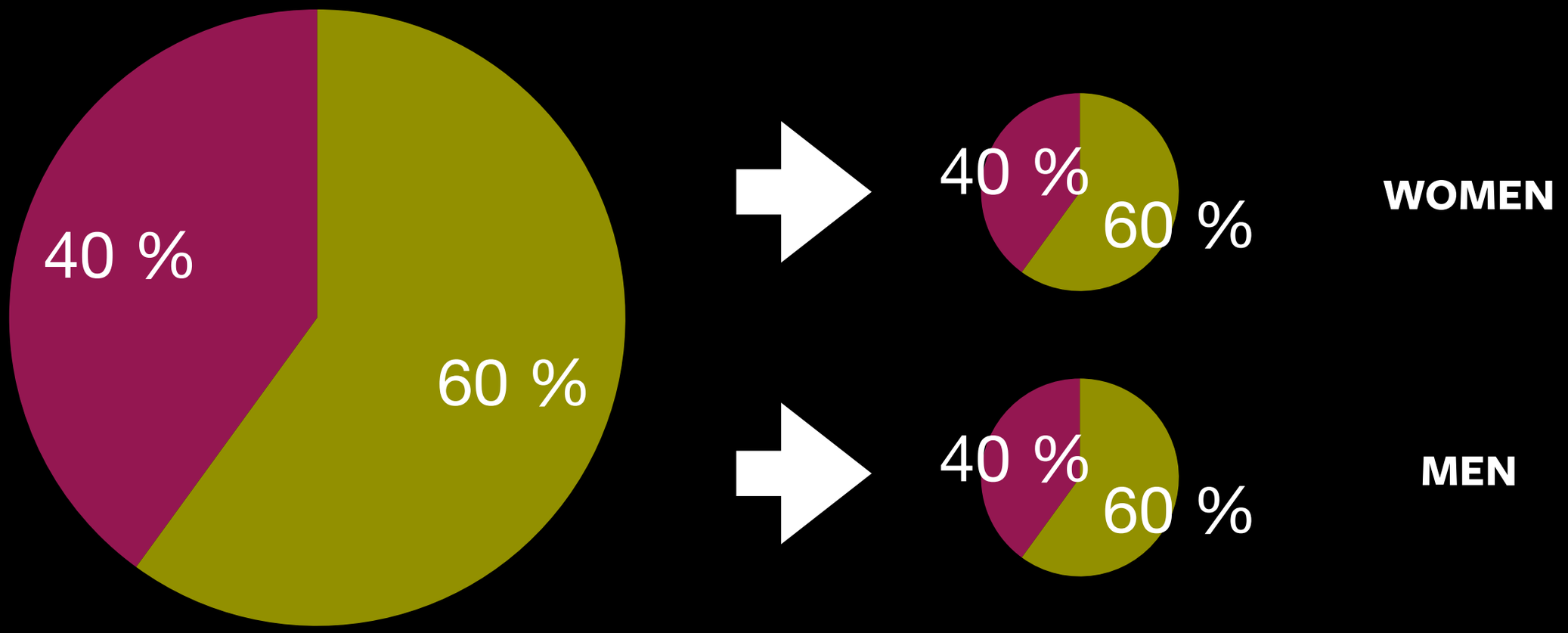
FORMAL DEFINITIONS: CLASSIFICATION PARITY

- **CLASSIFICATION PARITY OF FALSE POSITIVE RATE**



FORMAL DEFINITIONS: CLASSIFICATION PARITY

- **CLASSIFICATION PARITY**



DISTRIBUTION OF ALL DECISIONS

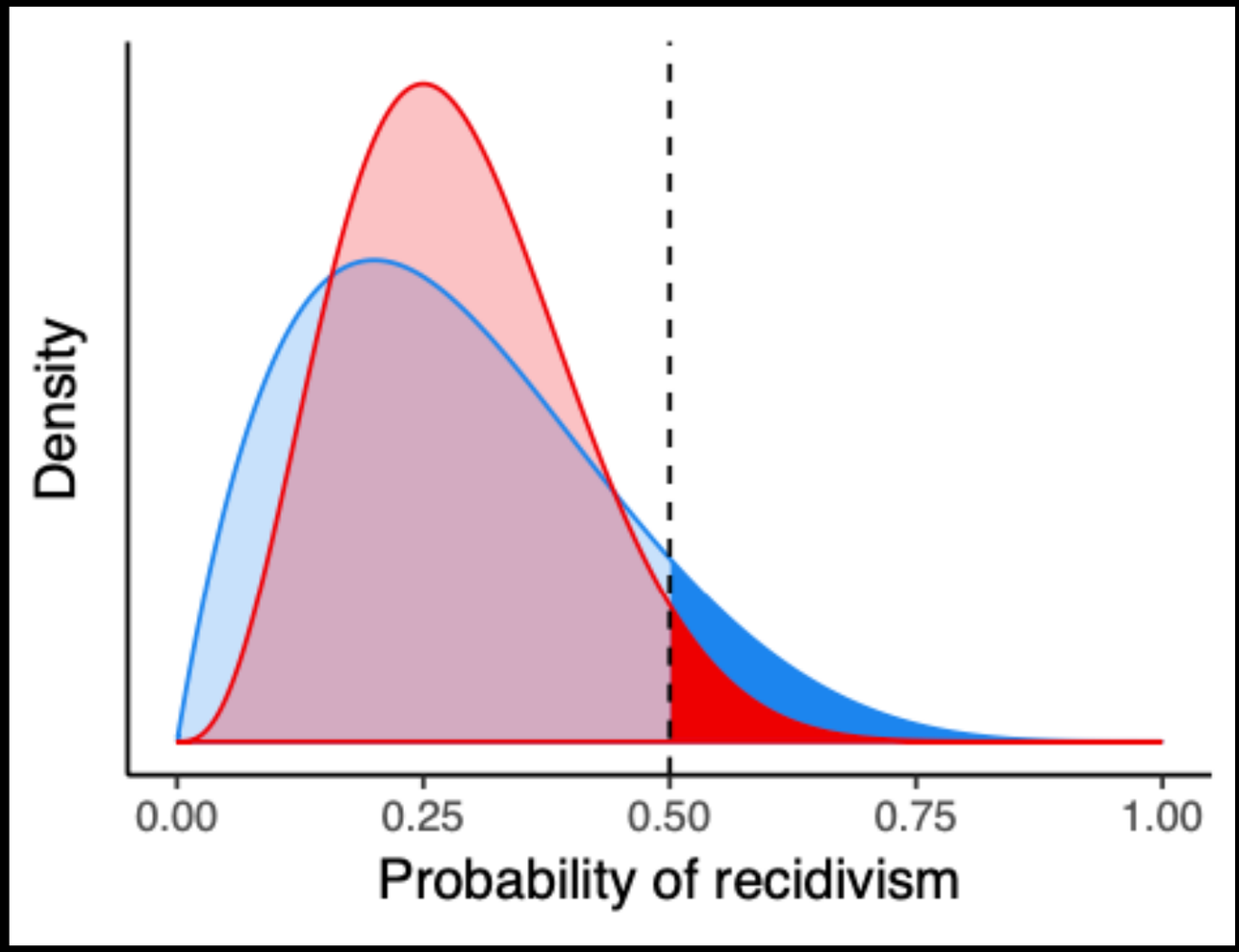
DISTRIBUTION OF SUBSETS

FORMAL DEFINITIONS: CLASSIFICATION PARITY

- **LIMITS OF CLASSIFICATION PARITY**
  - risk distributions differ among groups
    - depend entirely on X and how well X describes the group
  - thresholds lead to unequal classification errors among groups

FORMAL DEFINITIONS: CLASSIFICATION PARITY

- **LIMITS OF CLASSIFICATION PARITY**
  - hypothetical risk distributions
  - infra-marginal statistics differ



Quelle: S. Corbett-Davies, S. Goel, 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *Computer Research Repository*.

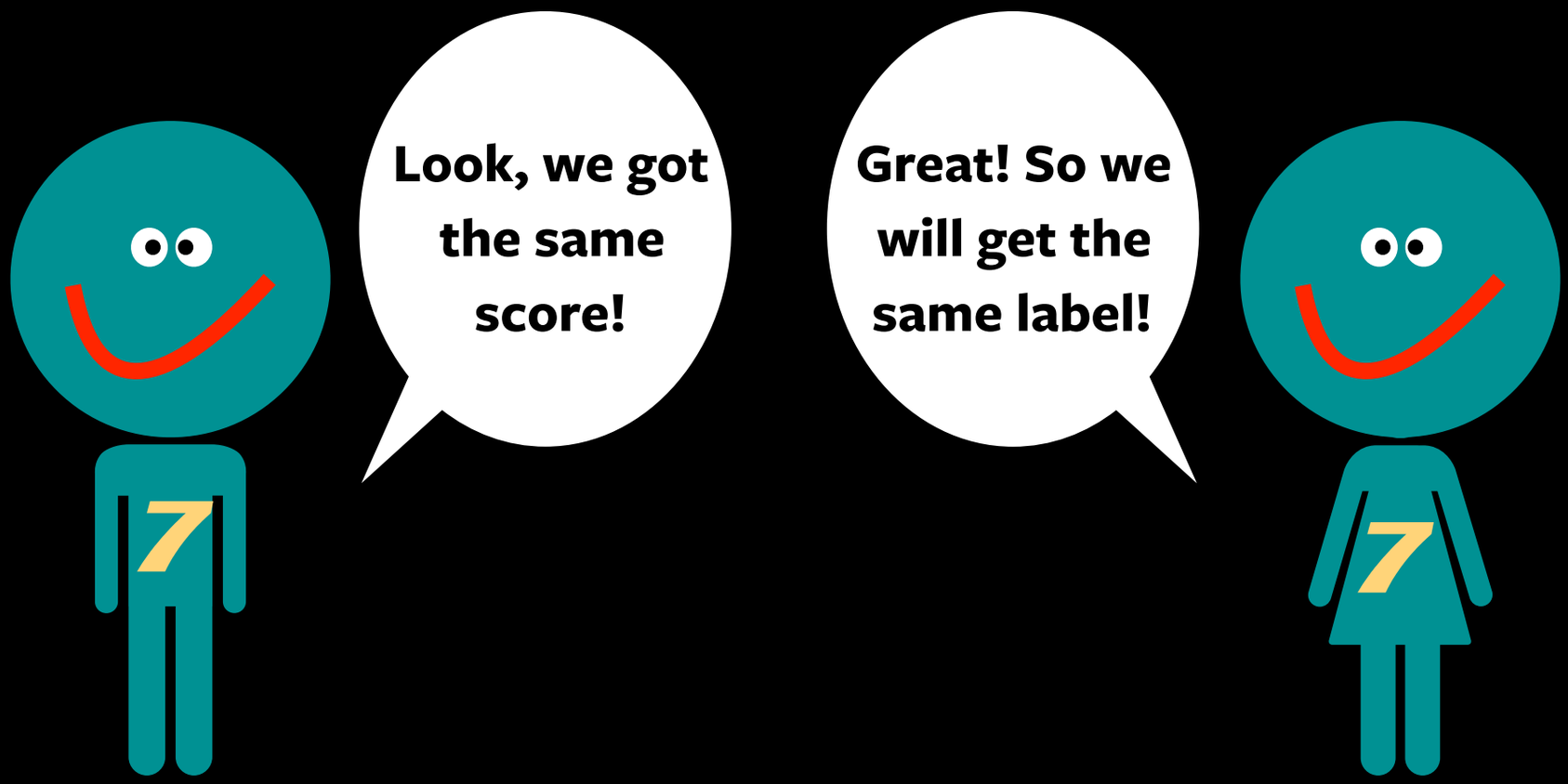
FORMAL DEFINITIONS: CALIBRATION

- **IDEA**
  - given any risk score, decisions must be independent of protected attributes
- ensures equal meaning of risk scores among all groups

$$\Pr(Y = 1 \mid s(X), X_p) = \Pr(Y=1 \mid s(X))$$

FORMAL DEFINITIONS: CALIBRATION

- **IDEA**

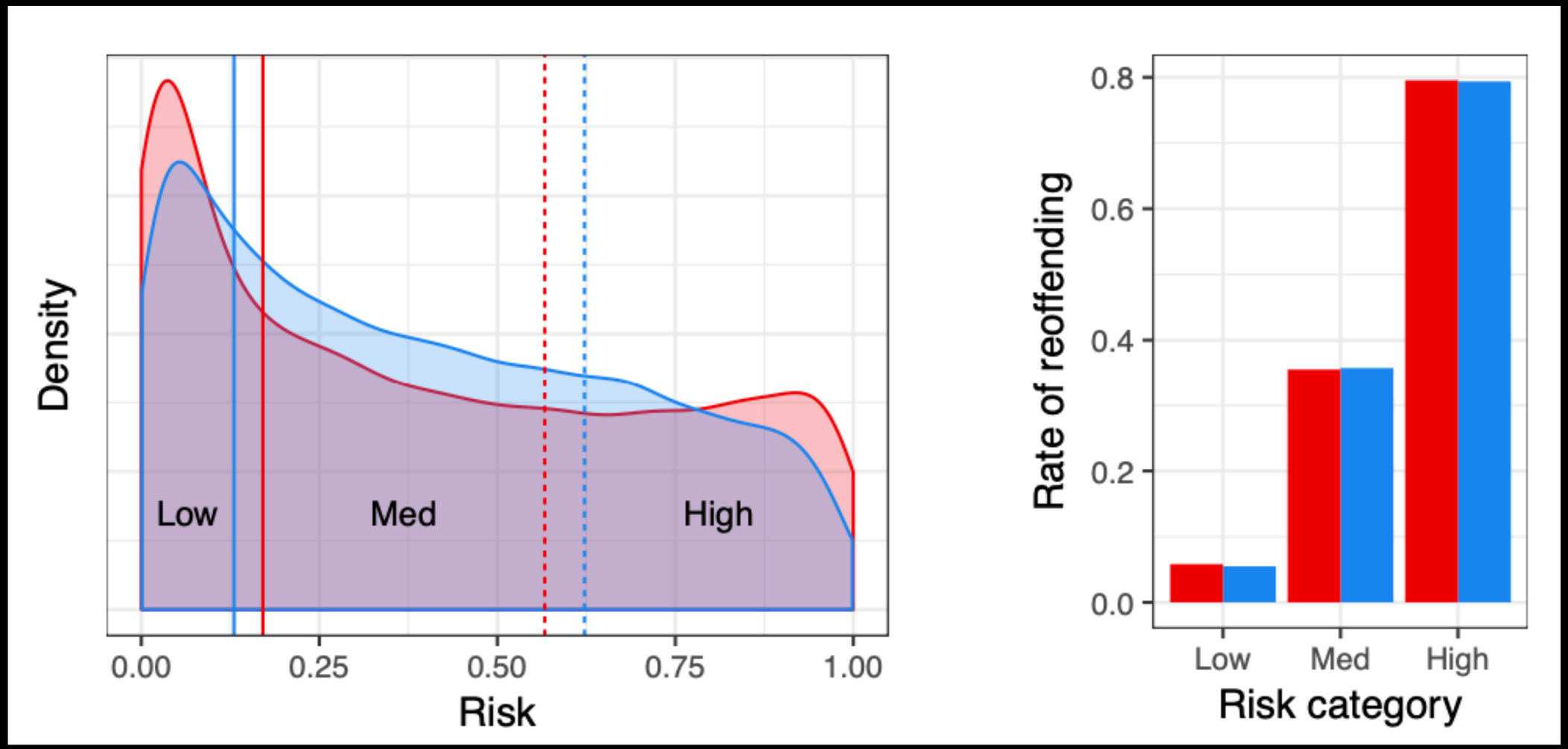


FORMAL DEFINITIONS: CALIBRATION

- **LIMITS OF CALIBRATION**
- insufficient to guarantee:
  - equitable decisions
  - accurate risk scores

FORMAL DEFINITIONS: CALIBRATION

- **LIMITS OF CALIBRATION**



Quelle: S. Corbett-Davies, S. Goel, 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *Computer Research Repository.*

PROBLEMS WITH DESIGNING FAIR ALGORITHMS

PROBLEMS WITH DESIGNING FAIR ALGORITHMS

- **MEASUREMENT ERROR**
  - decisions based on *true* risk
  - not true only approximated through X and Y
    - label bias (errors in Y)
    - feature bias (errors in X)

PROBLEMS WITH DESIGNING FAIR ALGORITHMS

- **MEASUREMENT ERROR: LABEL BIAS**
  - predicted Y in decision != observed Y:
    - Pr(reoffend | released) != Pr(offend)
    - e.g. pretrial: observed Y == crime that we know about
  - no solutions yet
    - but: check estimation strategy

PROBLEMS WITH DESIGNING FAIR ALGORITHMS

- **MEASUREMENT ERROR: FEATURE BIAS**
  - differences in predictive power of features
    - e.g. minorities more likely to be arrested => feature „past criminal behaviour“ could skew data
  - feature vector < real world features
  - solutions:
    - include group membership in predictive model
    - use more data (more features)

PROBLEMS WITH DESIGNING FAIR ALGORITHMS

- **SAMPLE BIAS**
  - sample data should reflect reality
  - problems:
    - reality: true distribution unknown
    - time: model might become outdated
  - no perfect solution:
    - try use representative training data

SUMMARY

- risk assessment tools
- threshold-rules aim to maximize utility conditional on approximated true risk
- imperfect mathematical definitions of fairness
  - anti-classification, classification parity, calibration
- designing fair algorithms bears many other problems
  - e.g. sample bias, feature and label bias

SOURCES

- S. Corbett-Davies, S. Goel, 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *Computer Research Repository.*
- S. Goel, 2019. The Measure and Mismeasure of Fairness. Talk at Berkely University. https://simons.berkeley.edu/talks/measure-and-mismeasure-fairness. [last called Jan. 8th 2020]
- C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (*ITCS '12*).
- M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD '15*).

PROBLEMS WITH DESIGNING FAIR ALGORITHMS

- **MODEL FORM AND INTERPRETABILITY**
  - best case: few features but much training data
    - statistical strategy has little effect on estimates
  - feature space high-dimensional or few training data
    - statistical strategy important
  - limited transparency of decisions
    - field of interpretable machine learning

PROBLEMS WITH DESIGNING FAIR ALGORITHMS

- **EXTERNALITIES AND EQUILIBRIUM EFFECTS**
- risk assessment tools could alter populations to which they are applied
- populations/distributions change
  - model becomes outdated
  - need of new training data