Algorithmic Statistical Discrimination in Criminal Risk Prediction

FATML
July 11, 2015

Zubin Jelveh
New York University (CRISSP)
& Crime Lab New York

Michael Luca
Harvard Business School

Risk Prediction Tools

- Courts:
  - Pretrial bail and release
  - Sentencing
- Probation and parole:
  - Levels of Supervision
- Prison and Jails:
  - Security Classifications
  - Program targeting
- Law Enforcement:
  - Hotspot prediction
  - High-risk offenders

Current Models

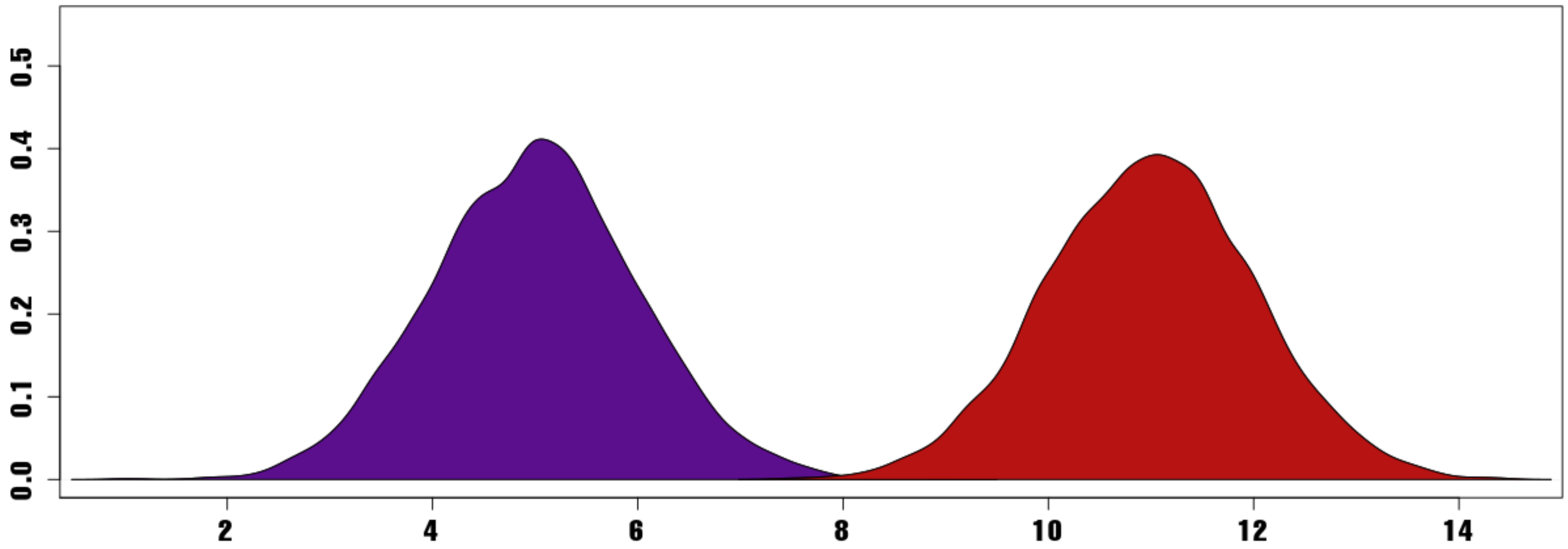
- MODELS: Federal Pretrial Risk Assessment, HCR-20, LSI-R, Federal Post Conviction Risk Assessment, COMPAS
- FACTORS:
  - Age, Alcohol/Drugs, Arrested before age 16, Citizenship, Companions, Criminal acquaintances, Criminal associates, Criminal attitudes, Criminal family, Criminal involvement, Criminal personality, Current drug problems, Current offense class, Current offense type, Current violence, Dissatisfaction with marital situation, Distress, Drug problems, **Educational attainment**, Emotional/Personal, Employment history, Employment problems, Escape history, Exposure to destabilizers, Family criminality, Financial problems, **High crime neighborhood**, History of violence, Impulsivity, Institutional adjustment problems, Interaction with parents, Interactions with authorities, Lack of insight, Lack of personal support, Lack of social support, Major mental illness, Married, Mental health treatment, Motivated to change, Negative attitudes, Participation in organized activity, Participation in school activities, Peer interactions, Plans lack feasibility, Previous violence, Prior adult convictions, **Prior incarceration**, Prior psychological assessment, Prior supervision failure, Psychopathy, Psychosis, Punished for institutional misconduct, Relationship instability, Reliance on social assistance, Risk Management, School suspensions, Social isolation, Substance abuse, Substance use problems, Treatment nonresponse
- **Don’t include race, ethnicity, gender, etc.**

Considerations

- Desire: More robust method to combat discrimination
- Accuracy vs. Fairness trade-off
  - Algorithmic/practitioner concern
- Equity vs. Efficiency trade-off
  - How do we define equity in criminal justice context?
  - Who gains, who loses?
- Data Generating Processes:
  - Impact how we make an algorithm discrimination free?

Baseline Method

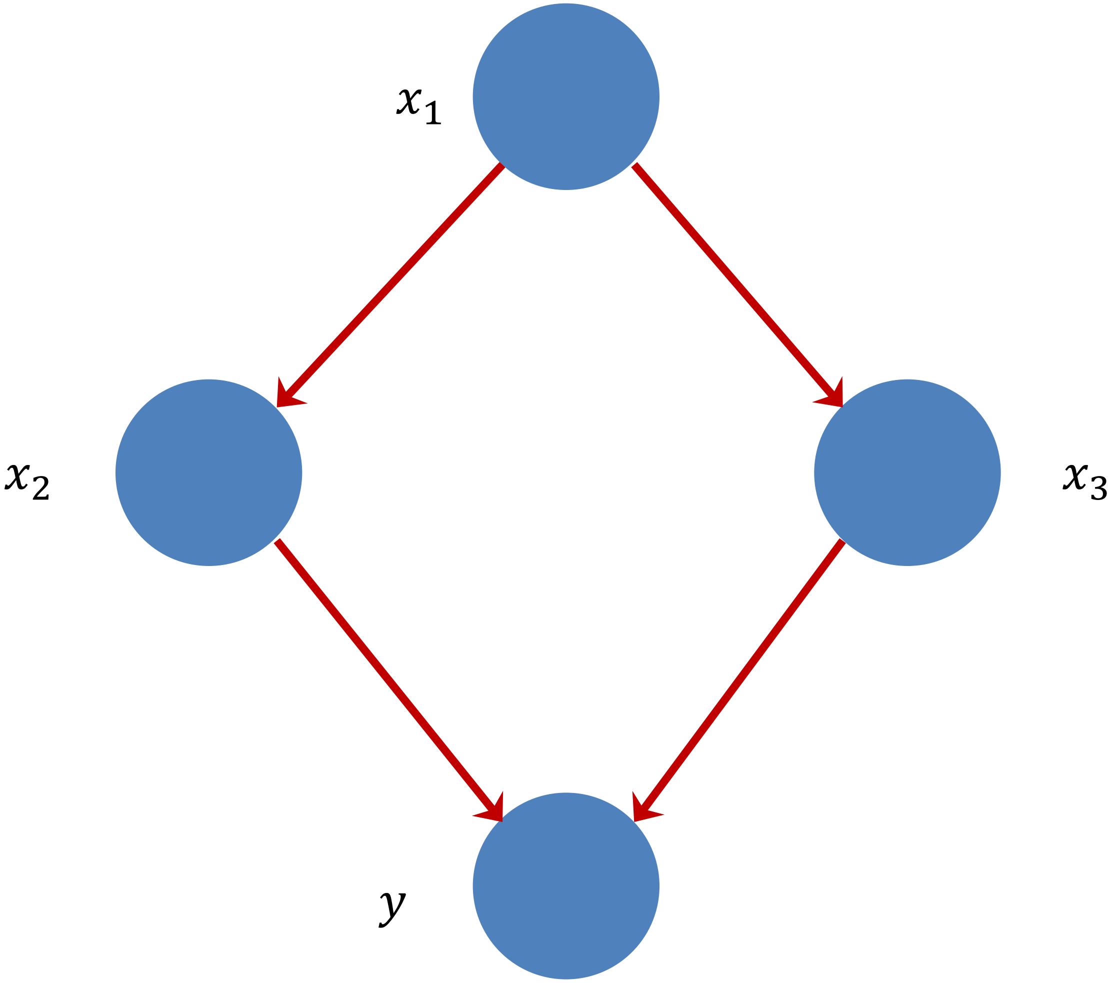
- $\{s_1, \ldots, s_k\}, \{x_1, \ldots, x_n\}, y$
- Assume linear data-generating process
- Massage input data:
  - $\mathrm{x}_{i,j} \ = \alpha_j \ + \beta_{j,1} \, s_{i,1} + \cdots + \beta_{j,k} s_{i,k}$
  - Residual/Empirical error is uncorrelated with sensitive variables
  - $y_i = \kappa \ + \delta_1 \, \tilde{x}_{i,1} + \cdots + \delta_n \, \tilde{x}_{i,n}$
- Equivalent to (Pope & Sydnor 2011):
  - $y_i = \alpha + \delta_1 x_{i,1} + \cdots + \delta_n x_{i,n} + \beta_1 s_{i,1} + \cdots + \beta_k s_{i,k}$
  - $\hat{y}_i = \hat{\delta}_1 x_{i,1} + \cdots + \hat{\delta}_n x_{i,n}$
- **Correlation between $\widehat{y}$ and s will be close to zero**


0.0
0.1
0.2
0.3
0.4
0.5
2
4
6
8
10
12
14

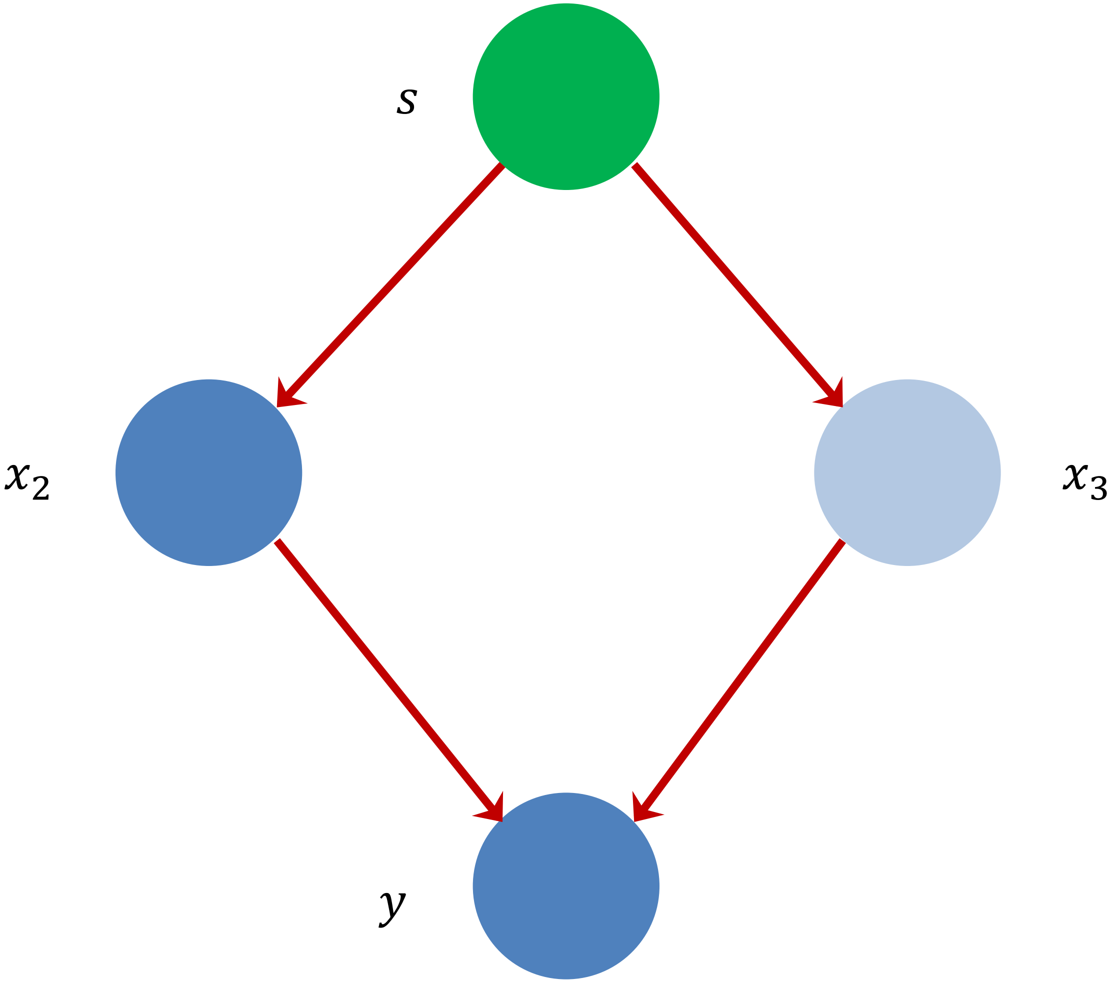
What Accuracy Loss Is “Ideal”?

- A simple model:
  - Potential predictors: x_1, x_2, x_3
  - Outcome of Interest: y
  - One sensitive variable
  - One unobserved variable
- “Ideal”:
  - Statistical parity
  - Non-sensitive information not removed (No over-adjustment)
  - Respects DGP: De-correlate if function of causal process

A Causal Process

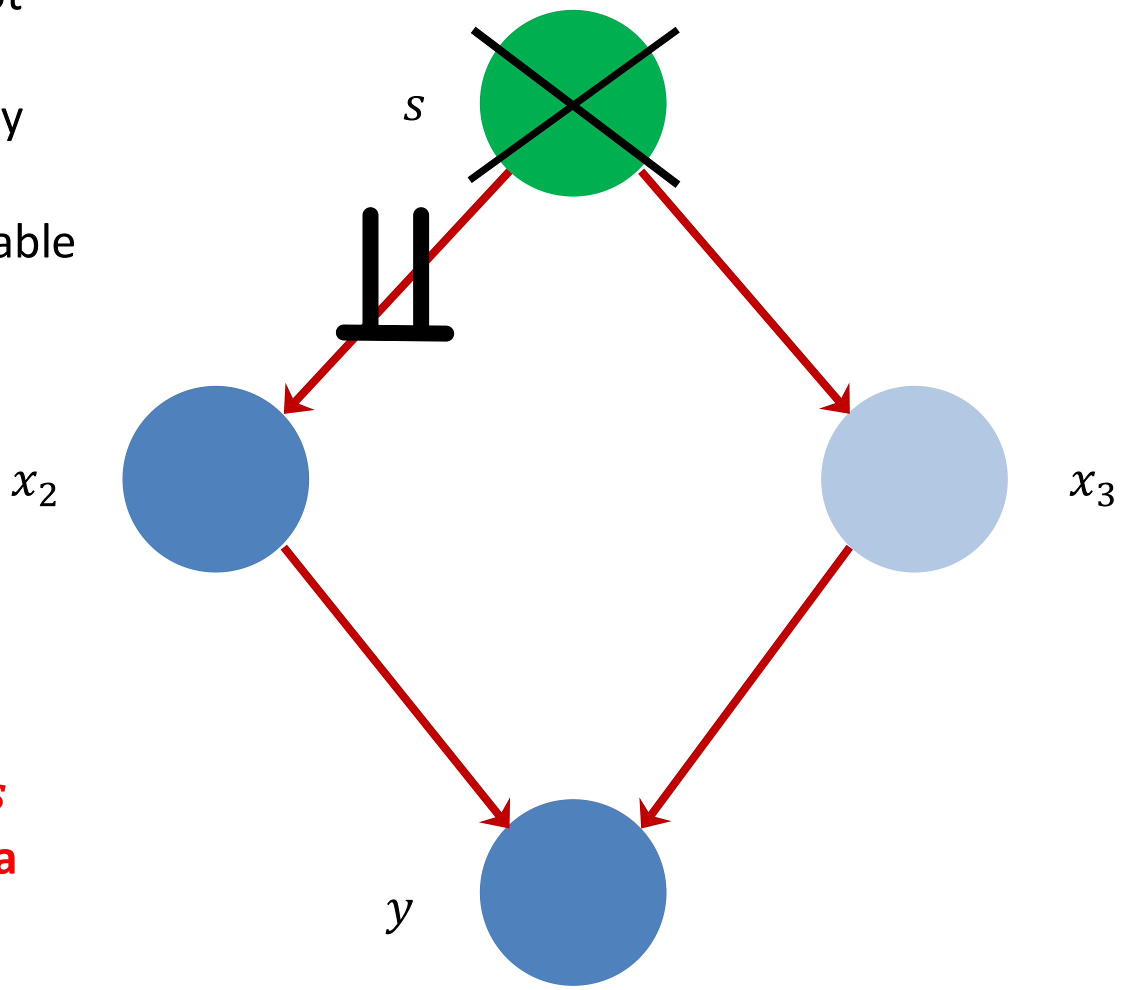


Scenario 1



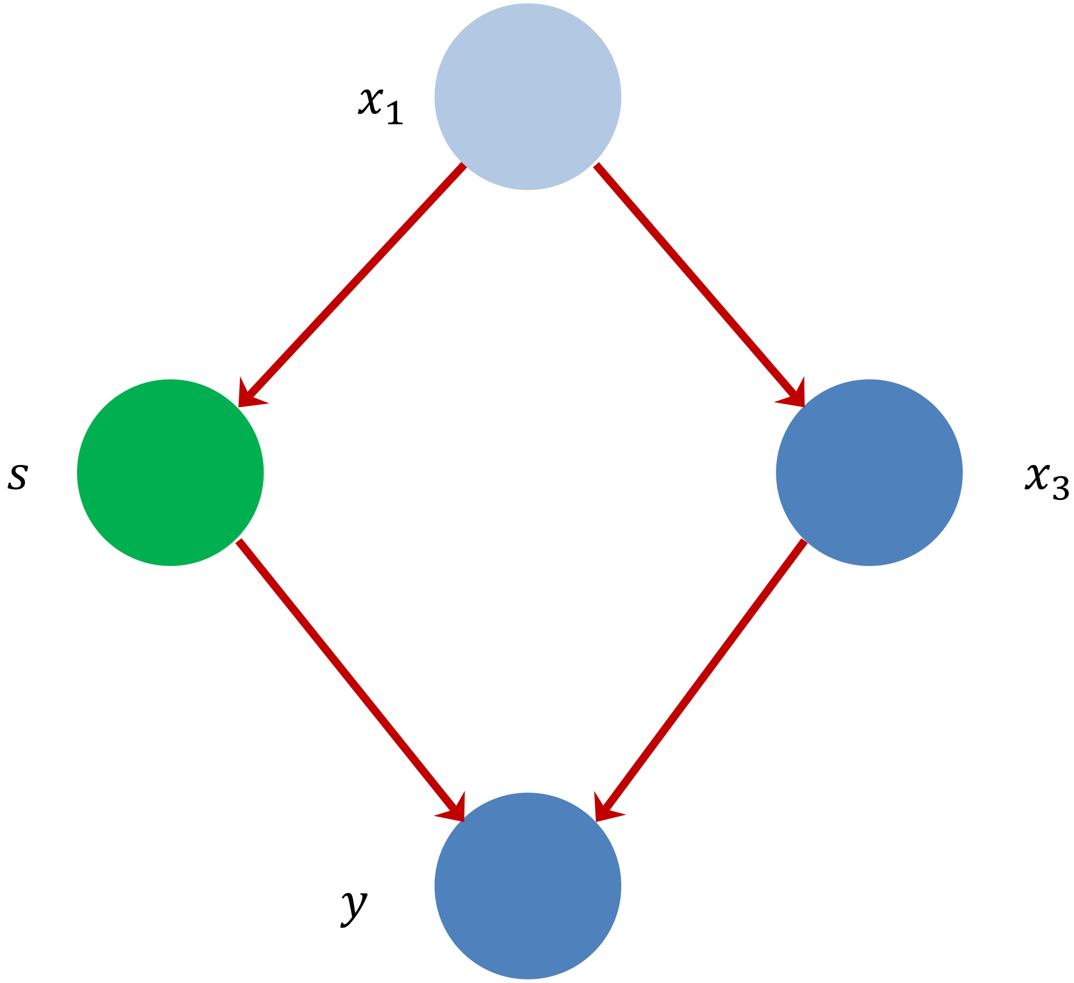
Scenario 1

- Non-sensitive information not removed
- Statistical Parity
- Respects DGP
- s non-manipulable
  - Race
  - Ethnicity
  - Age
  - Gender



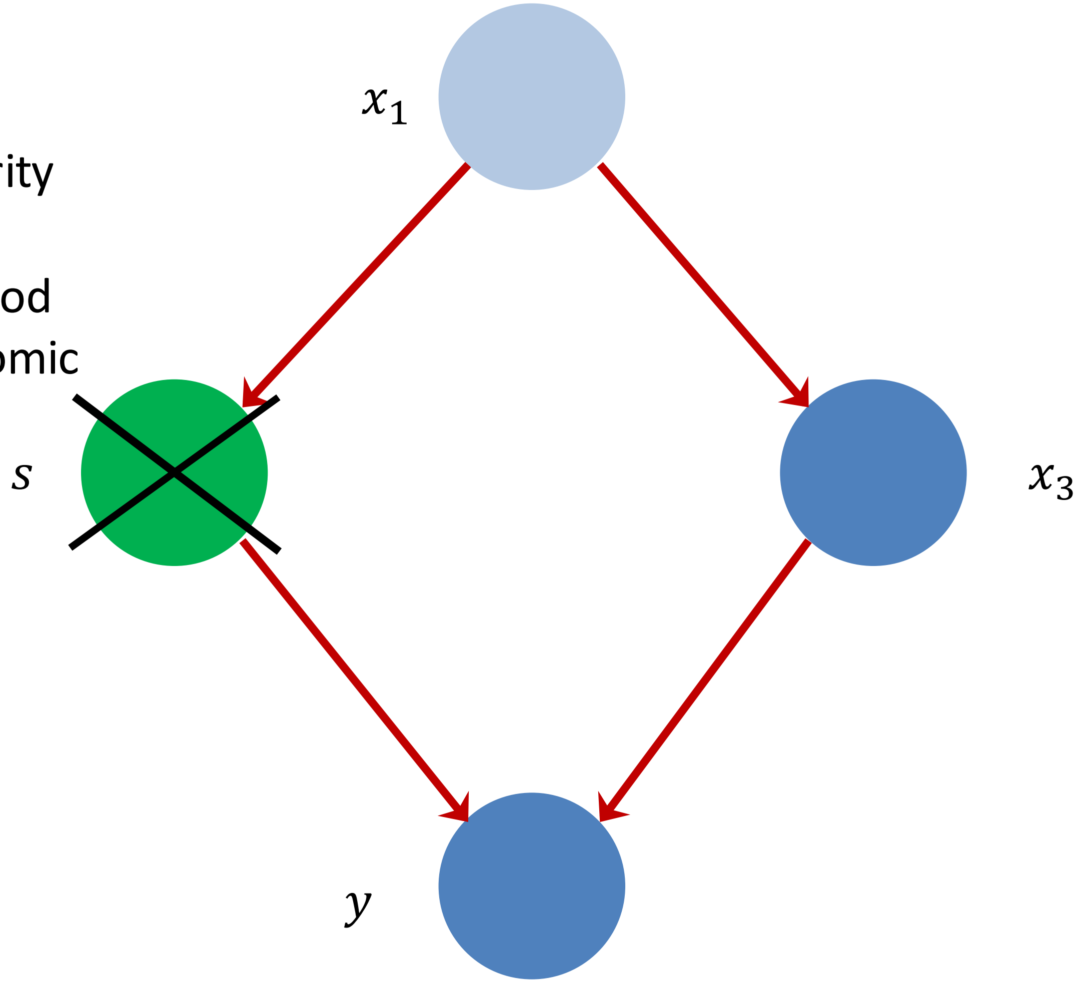
- **DISCLAIMER: s causes $x_{\downarrow}2$ as a result of discrimination**

Scenario 2

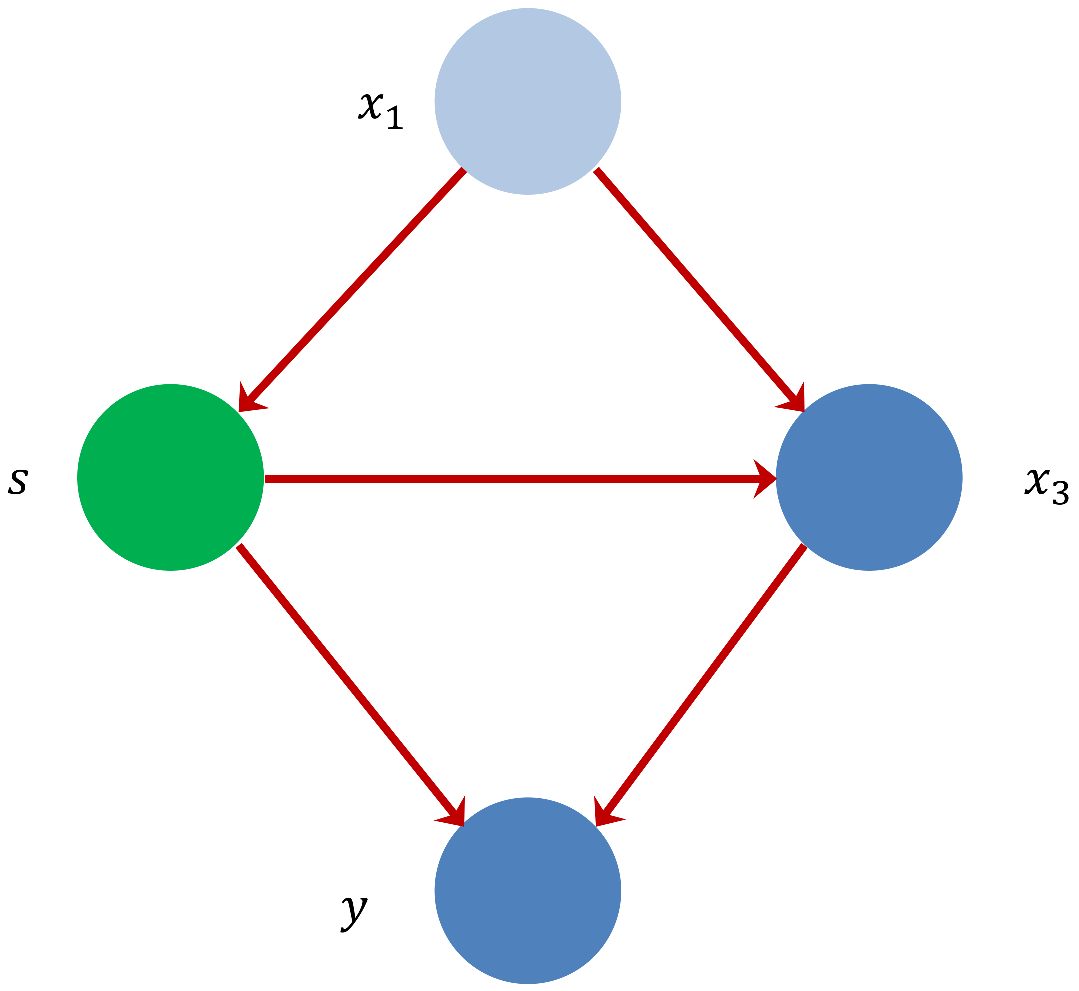


Scenario 2

- Non-sensitive information not removed
- Respects DGP
- No Statistical Parity
- s manipulable
  - Neighborhood
  - Socio-economic status

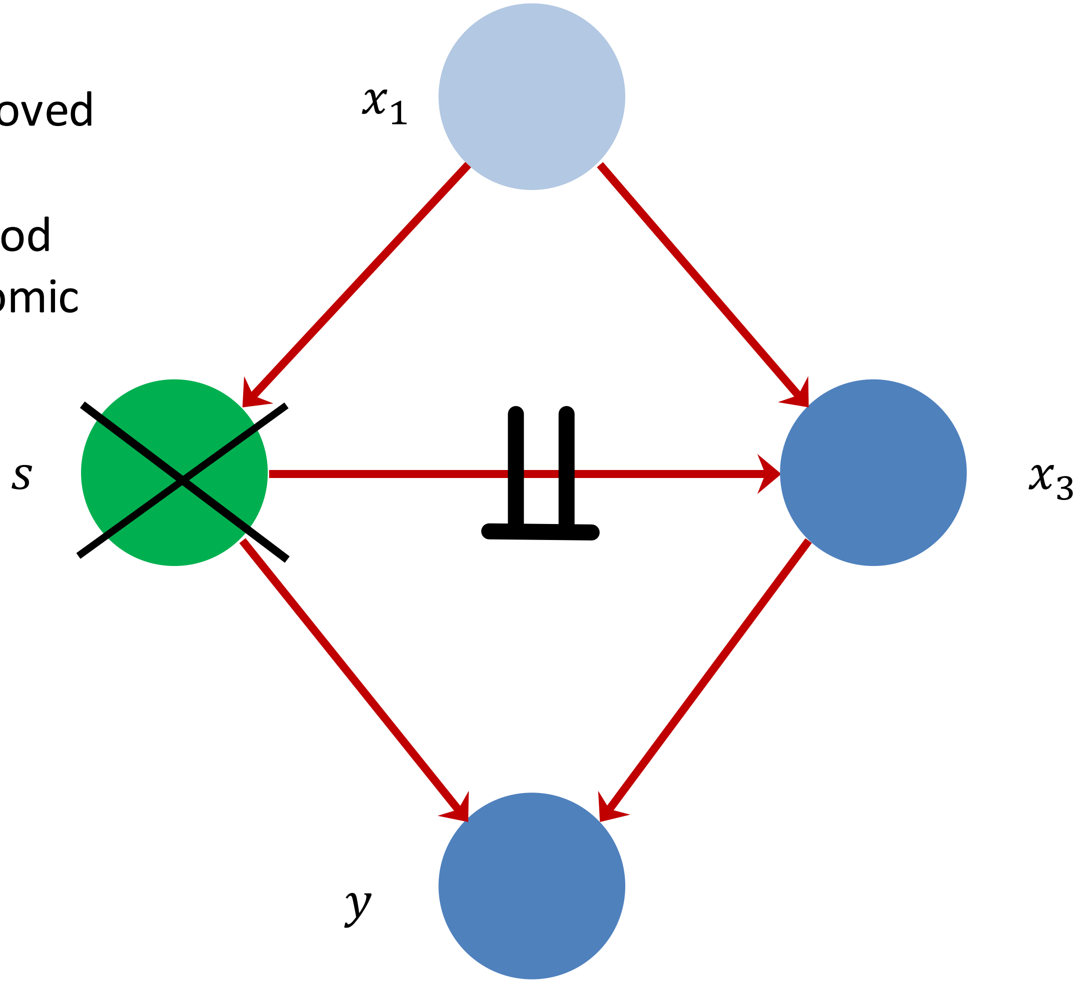


Scenario 3

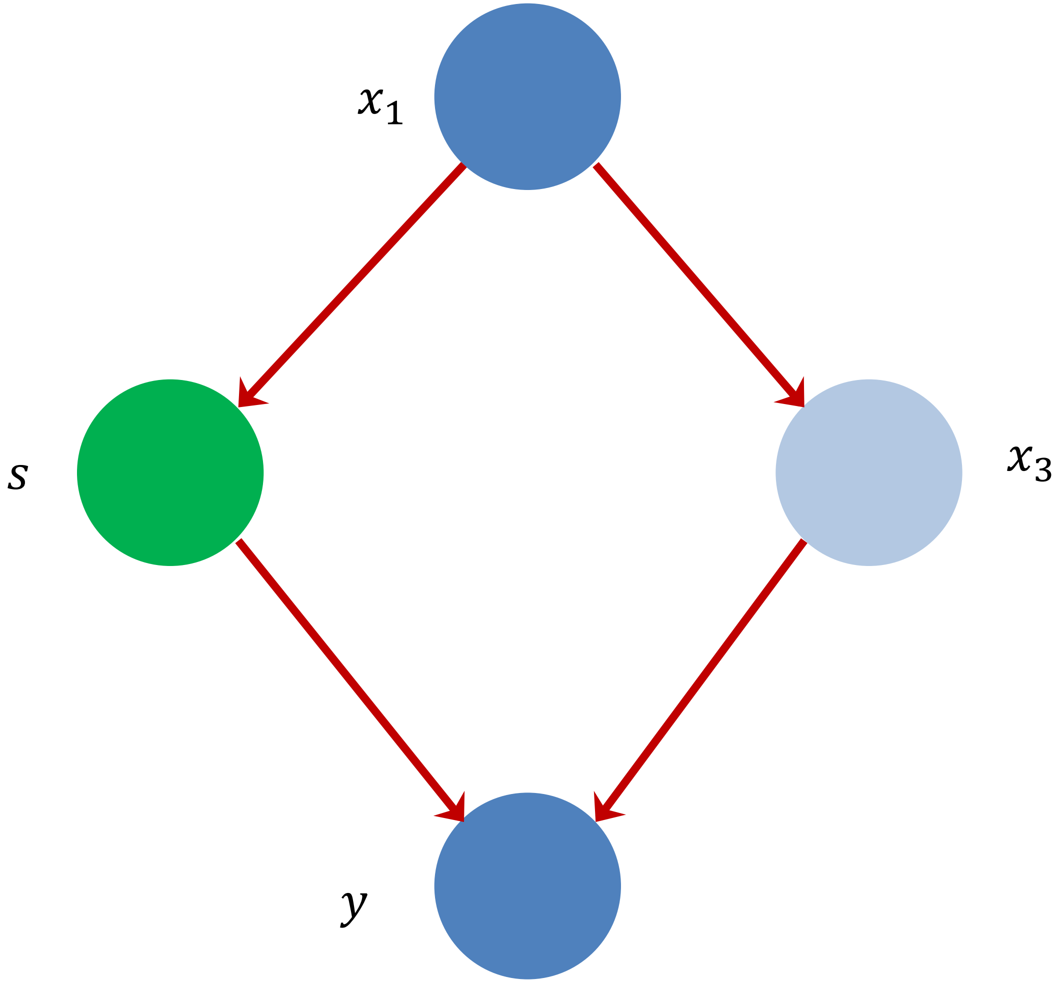


Scenario 3

- Respects DGP
- Statistical Parity
- Non-sensitive information removed
- s manipulable
  - Neighborhood
  - Socio-economic status

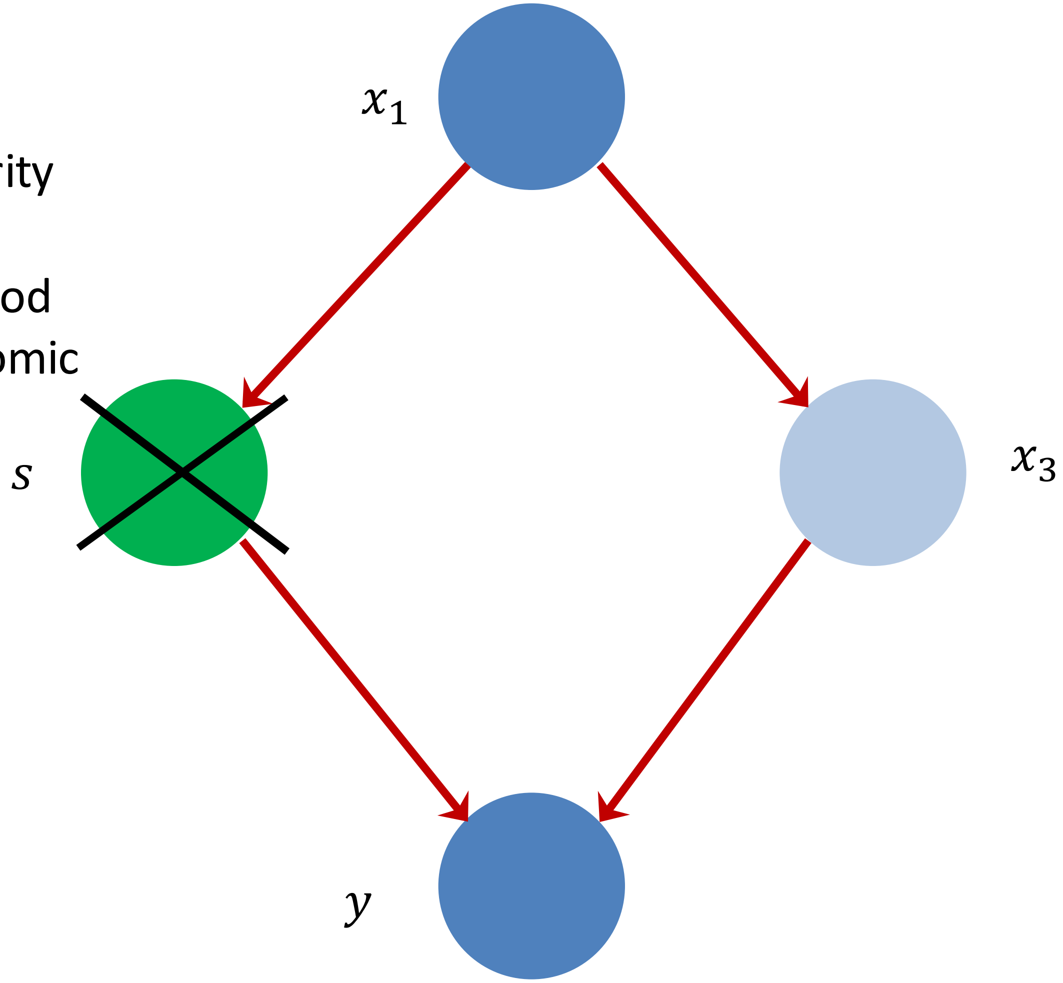


Scenario 4



Scenario 4

- Non-sensitive information not removed
- Respects DGP
- No Statistical Parity
- s manipulable
  - Neighborhood
  - Socio-economic status



Upshot

- Ideal achieved when sensitive variable is root cause
- Otherwise either statistical parity not reached or non-sensitive information removed
- Manipulable sensitive variables typically proxies for non-manipulable
- Related notion in economics (Fryer & Loury, 2005):
  - Color-blind affirmative action less efficient

Economic Models of Bias-Generating Processes

- Taste-based discrimination:
  - Law enforcement derives utility from targeting a sensitive group
  - Two identical individuals except for sensitive variable will be treated differently
  - $y_i = \alpha + \beta_1 x_i + \beta_2 s_i$
  - Economists believe competition should drive out taste-based discrimination

Other Bias-Generating Processes

- Statistical discrimination:
  - Law enforcement can't observe true likelihood of committing crimes
  - No animus against sensitive group
  - To make most efficient decision use historical correlation between group membership and criminal activity

Statistical Discrimination (Autor 2003)

- $y_x \sim N(\bar{y}_x, \sigma_y^2)$
- $\bar{y}_a \leq \bar{y}_b$, equal variances
- $y_i = \bar{y}_x + \epsilon_i$
- Signal error ridden: $\tilde{y}_i = y_i + v_i$, $v \sim N(0, \sigma_v^2)$
- $\tilde{y}_i = \bar{y}_x + \epsilon_i + v_i$
- $E[\tilde{y}_i | y_i] = y_i$, i.e. unbiased

Statistical Discrimination (Autor 2003)

- But what about reverse:
- $E[y_i|\tilde{y}_i, x] = \bar{y}_x + (\tilde{y}_i - \bar{y}_x)\gamma$
- $\gamma = \frac{\sigma_y^2}{\sigma_y^2+\sigma_v^2}$ (regressing y on $\tilde{y}$)
- Unless the signal is not error ridden, identical people, except for race, will have different predictions
- **Not due to animus**

Algorithmic Statistical Discrimination

- Statistical discrimination is efficient in an economic sense
- Perhaps not so in certain legal and ethical concerns
- Algorithms are statistical discriminators
- Devil’s advocate: Data is being used efficiently to serve public policy
- Is this actually illegal?

U.S. Law

- Two notions of discrimination:
  - Disparate Impact
    - Discriminatory effects from policies which may on their face be non-discriminatory
  - Disparate Treatment
    - Discrimination a result of identifiable animus
    - Statistical evidence not enough
  - Disparate Impact ≈ Statistical Discrimination
  - Disparate Treatment ≈ Taste-Based Discrimination

Legal Precedent in Criminal Justice

- *U.S. v Brignoni-Ponce*: Allowed for Mexican heritage to be used as a factor in border police making stops.
- *Anderson v Cornejo*: African-American women more likely to be screened at O’Hare does not imply disparate treatment
- *Johnson v U.S.*: Allowed use of race and ethnicity for prison cell assignment
- *McCleskey v Kemp*: Racial disparities in death penalty in Georgia. “Exceptionally clear proof needed.”

Laws and Attitudes Not Immutable

- **NYPD Racial Profiling Policy:** "prohibits the use of race, color, ethnicity, or national origin as a determinative factor in taking law enforcement action"
- **US DOJ:** race, ethnicity, national origin, **religion**, **gender**, **sexual orientation**
- **NYPD "Stop-and-Frisk"** found to violate Equal Protection (use of both statistical and qualitative evidence)

The Hit-Rate Test (Knowles, Persico, Todd 2001)

- Consider Stop-and-Frisk scenario
- How do we make the process fair?
- ML has tended to focus on search rates:
  - $\min P(\hat{y} = 1|s = 1) - P(\hat{y} = 1|s = 0)$
- Discrepancy could be result of efficient allocation of police resources

The Hit-Rate Test

- Propose a "bright-line" test for disparate treatment/taste-based discrimination
- Rational choice model: Police and targets are utility maximizers
- Law enforcement changes their search behavior until:
  - $P(y = 1 | \hat{y} = 1, s = 1) = P(y = 1 | \hat{y} = 1, s = 0)$
- Two problems:
  - Assume rationality and no systematic error in judgement
  - Use of macro data to infer individual-level intent

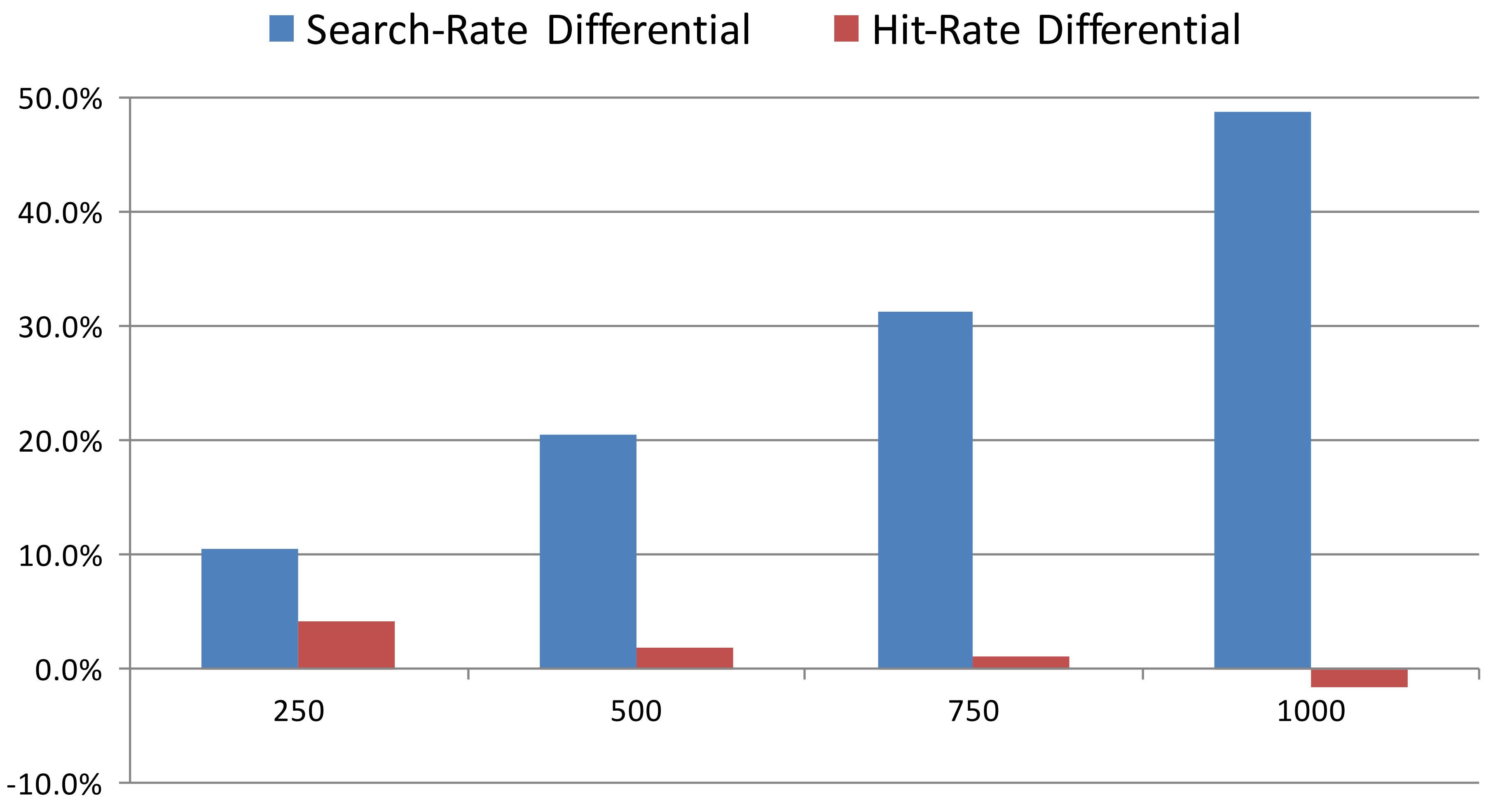
Hit Rates vs. Search Rates

- Which one to optimize for?
- Optimizing for both may unacceptably reduce accuracy
- Given ambiguity in law, optimizing for one may still leave door open to discrimination charges
- Is an algorithm that is discriminatory w/r/t hit-rates a taste-based discriminator?

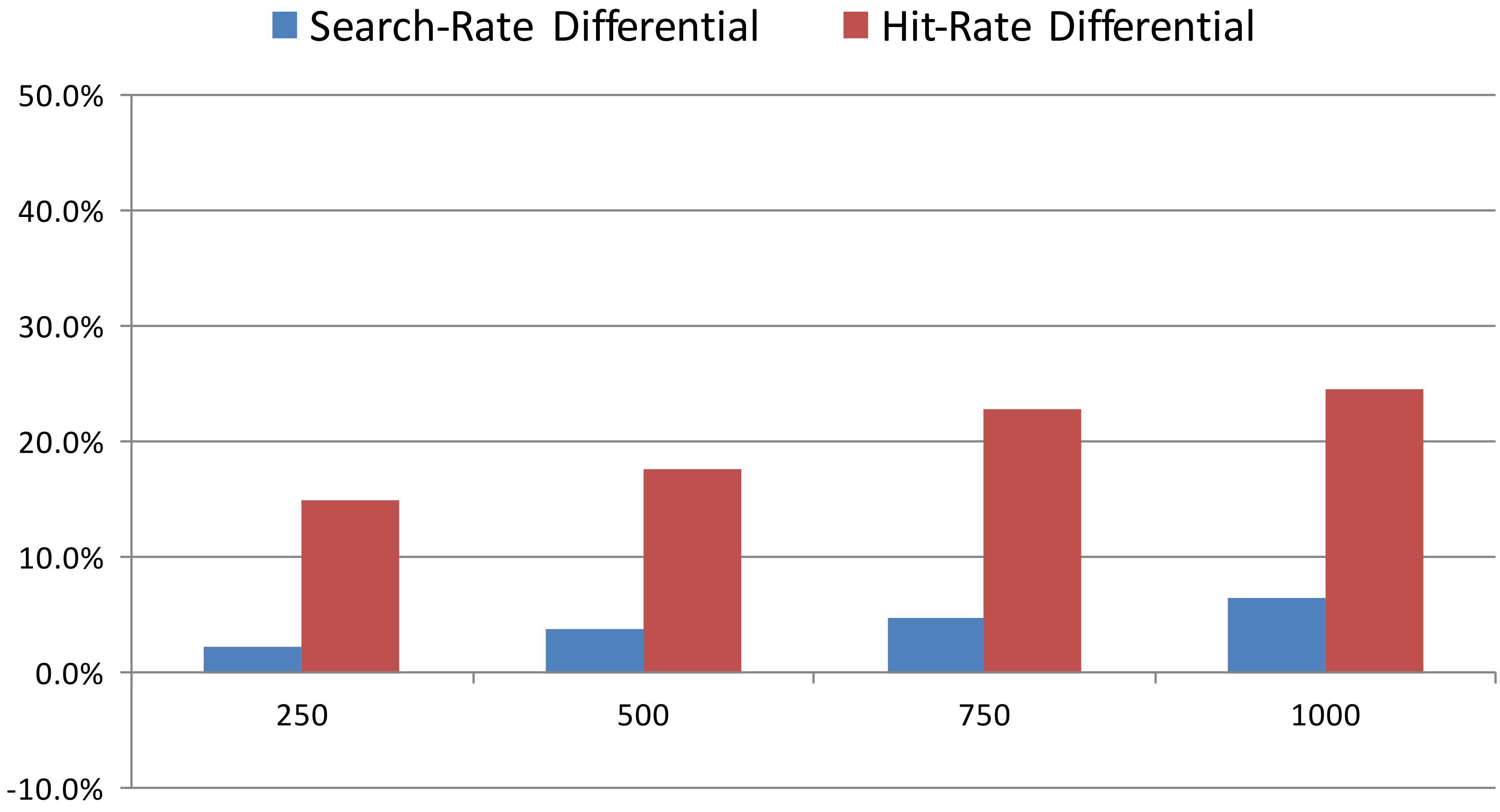
Closing Example

- National Longitudinal Survey of Youth
- Roughly 9,000 individuals born bet. 1980-84
- Variables: Race, gender, height, weight, prior arrests and incarcerations
- Split into pre-period (1980-2003) and post (2004-2011)
- Predict arrest in post-period

Unadjusted Model



Adjusted Model



Final Thoughts

- Investigating different DGPs informs algorithm design
- Still to come:
  - Equity vs. Efficiency:
    - Equity: Do we need to compensate false positives?
    - Efficiency:
      - Standard accuracy loss
      - Will offender behavior change in response to discrimination-aware alg?
  - Algorithm performance when optimizing for both Search and Hit Rates
  - Mix of statistical and taste-based discriminators