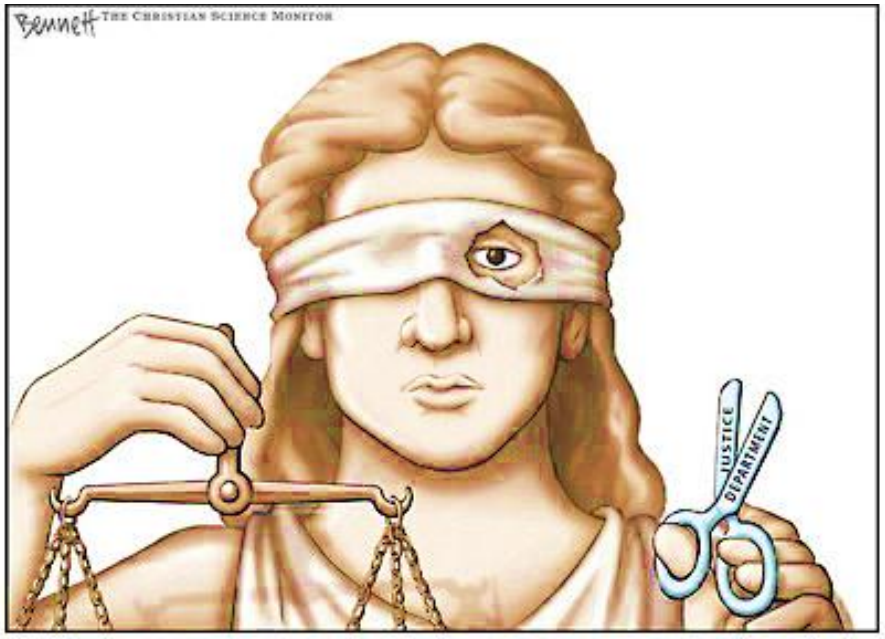
ENSURING FAIR DECISIONS



RICHARD ZEMEL

MAY 10, 2016

CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

WHY WAS I NOT SHOWN THIS AD?

FAIRNESS IN AUTOMATED DECISIONS

Health Care

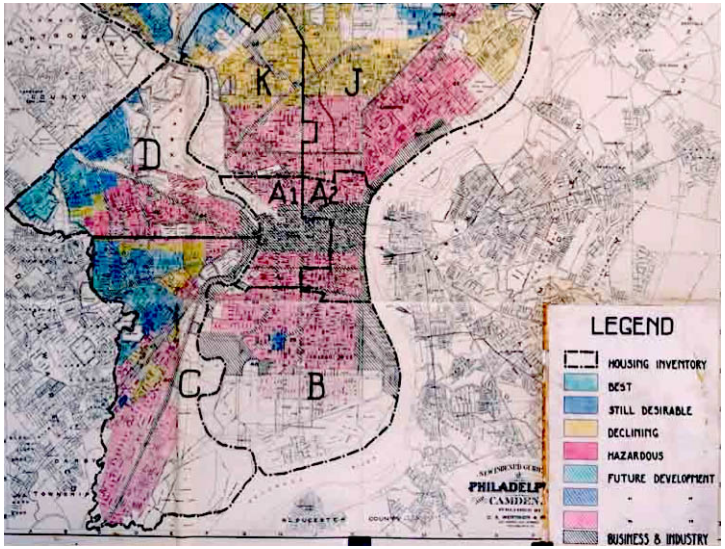
CONCERN: DISCRIMINATION

- Population includes minorities
  - Ethnic, religious, medical, geographic
- Protected by law, policy, ethics
- (If) we cannot completely control our data, can we regulate how it is used, how decisions are made based on it?

FORMS OF DISCRIMINATION

- *Steering* minorities into higher rates (advertising)

- *Redlining:* deny service, change rates based on area



DISCRIMINATION IN HIRING DECISIONS

Legal, public policy issues once decisions automated – responsibility of ML algorithm generally ignored

Learning algorithm finds optimal employee: lives near job, has reliable transportation, uses 1-4 social networks

"Practices that even unintentionally filter out older or minority applicants can be illegal under federal equal opportunity laws. If a hiring practice is challenged in court as discriminatory, a company must show the criteria it is using are proven to predict success in the job"

- http://science.slashdot.org/story/12/09/21/1437253/when-the-hiring-boss-is-an-algorithm

DISCRIMINATION IN LENDING DECISIONS

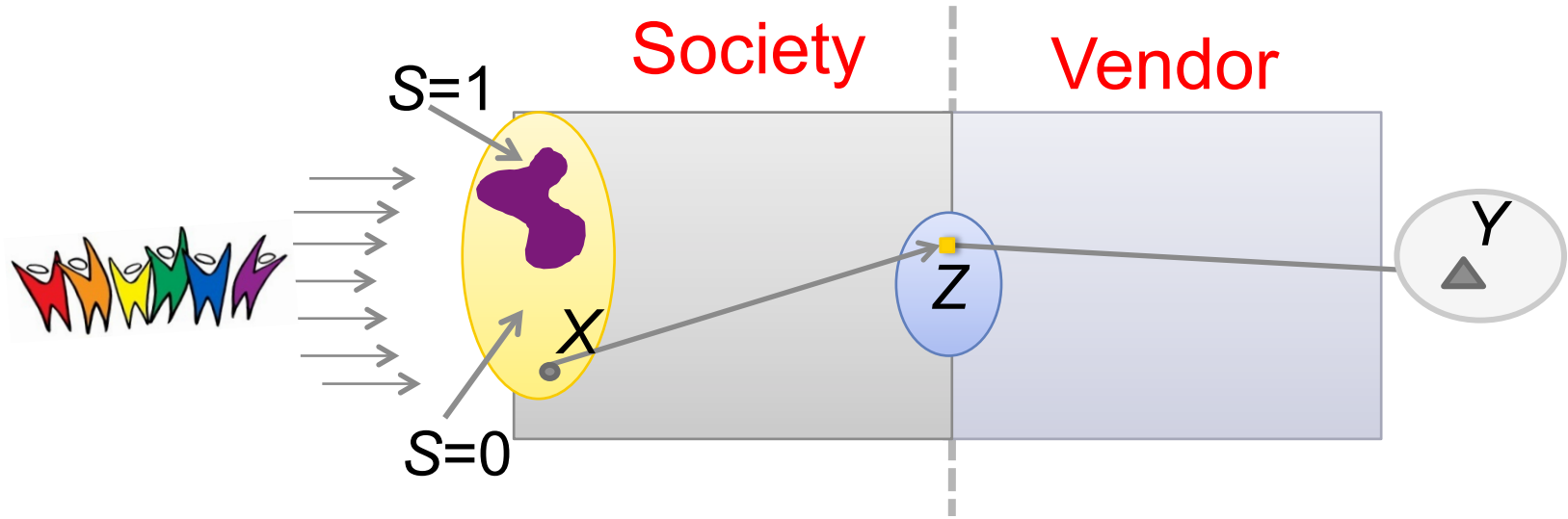
"Applying the modern techniques of data science to consumer lending raises questions, especially for regulators who enforce anti-discrimination laws….

By law, lenders cannot discriminate against loan applicants on the basis of race, religion, national origin, sex, marital status, age or the receipt of public assistance. Big-data lending, though, relies on software algorithms largely working on their own and learning as they go.

The danger is that with so much data and so much complexity, an automated system is in control. The software could end up discriminating against certain racial or ethnic groups without being programmed to do so."

- *Banking Start-Ups Adopt New Tools for Lending* (Jan 18, 2015)

General Framework



X: Original Representation of Person

Z: New Representation

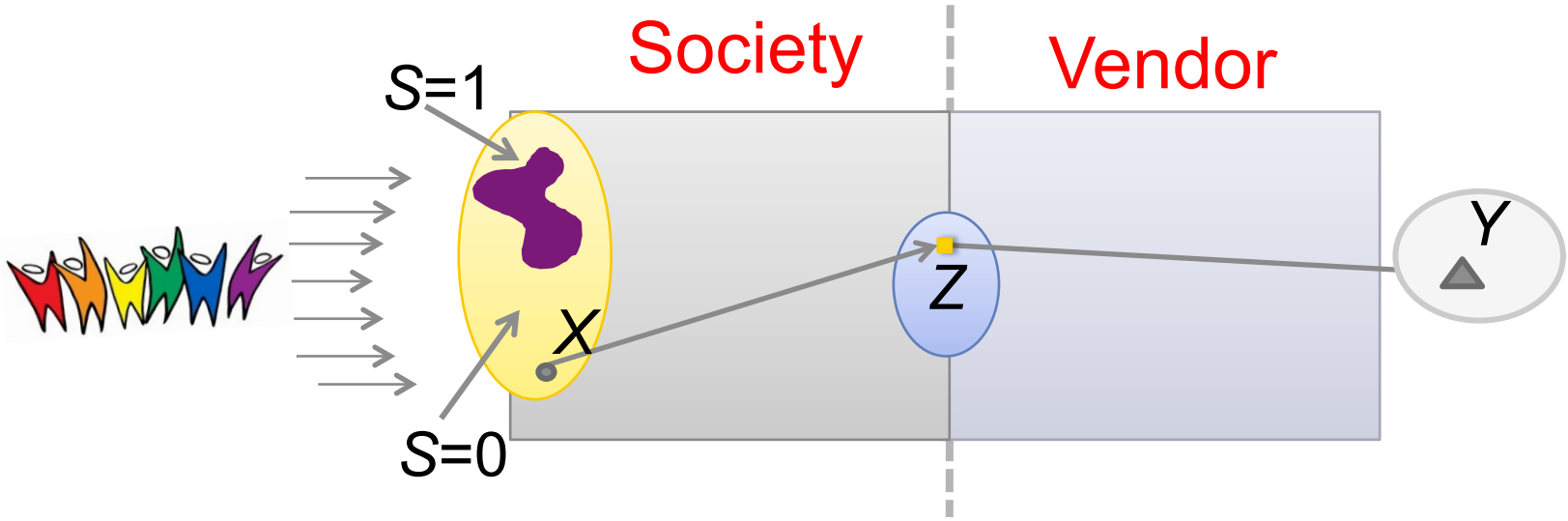
Y: Vendor Action

FAIRNESS VIA S-BLINDNESS?

Remove or ignore the "membership in S" bit

- Fails: Membership in S may be encoded in other attributes

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*)

EXPERIMENTS

1. German Credit

   Task: classify individual as good or bad credit risk

   Sensitive feature: Age

2. Adult Income

   Size: 45,222 instances, 14 attributes

   Task: predict whether or not annual income > 50K

   Sensitive feature: Gender

3. Heritage Health

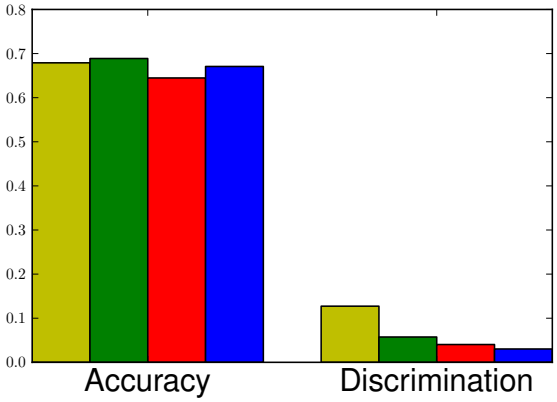
   Size: 147,473 instances, 139 attributes

   Task: predict whether patient spends any nights in hospital

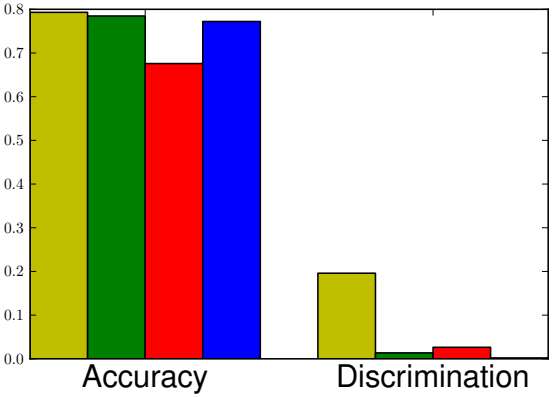
   Sensitive feature: Age

EXPERIMENTAL RESULTS

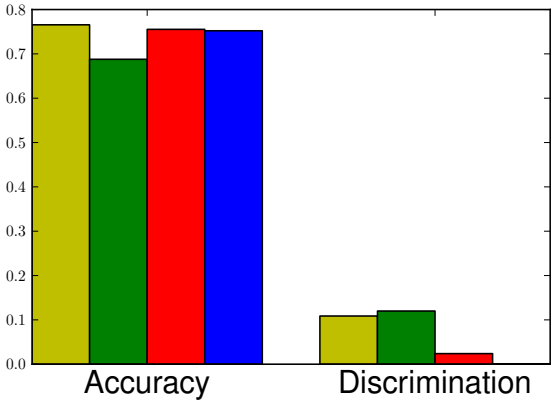
German



Adult

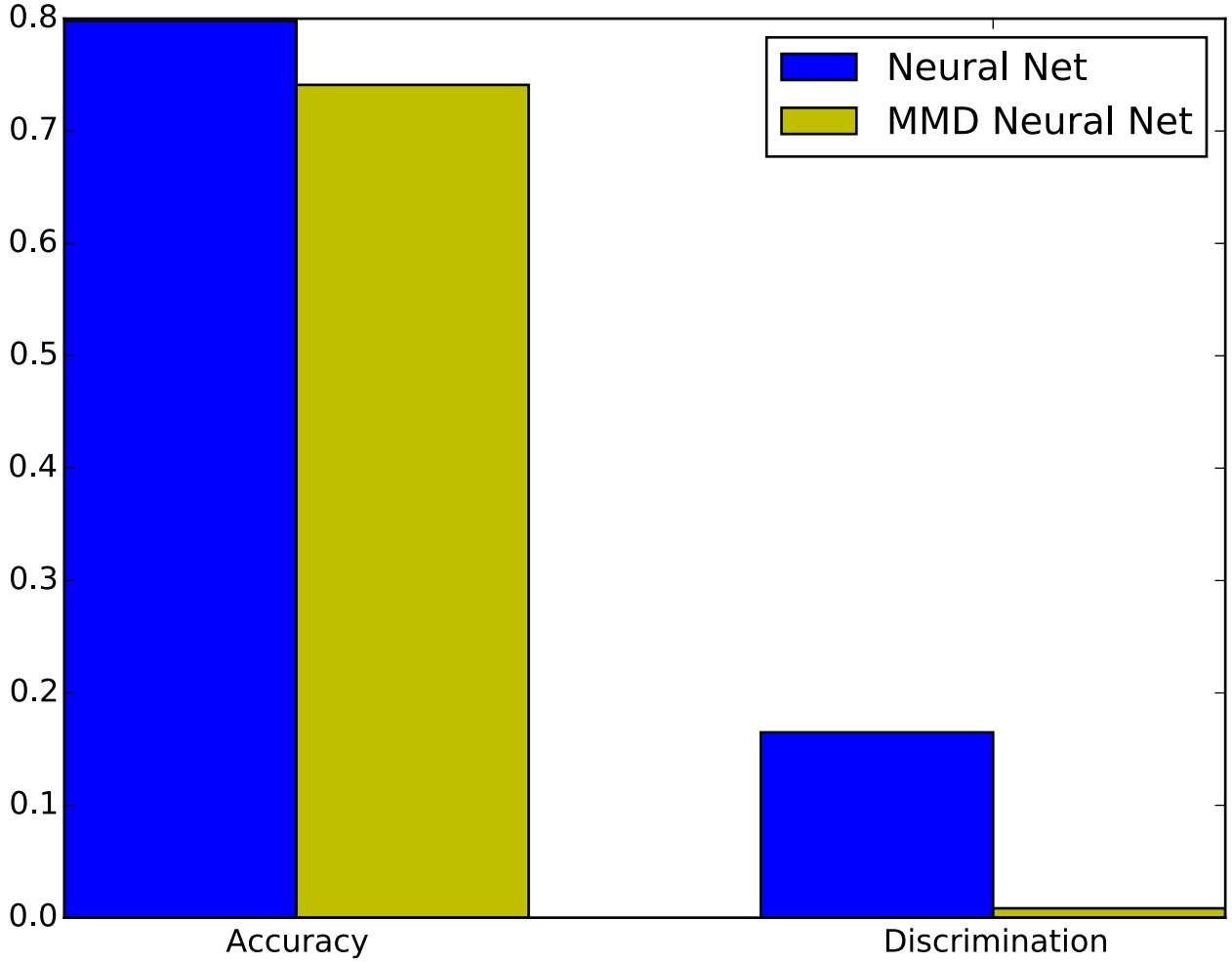


Health



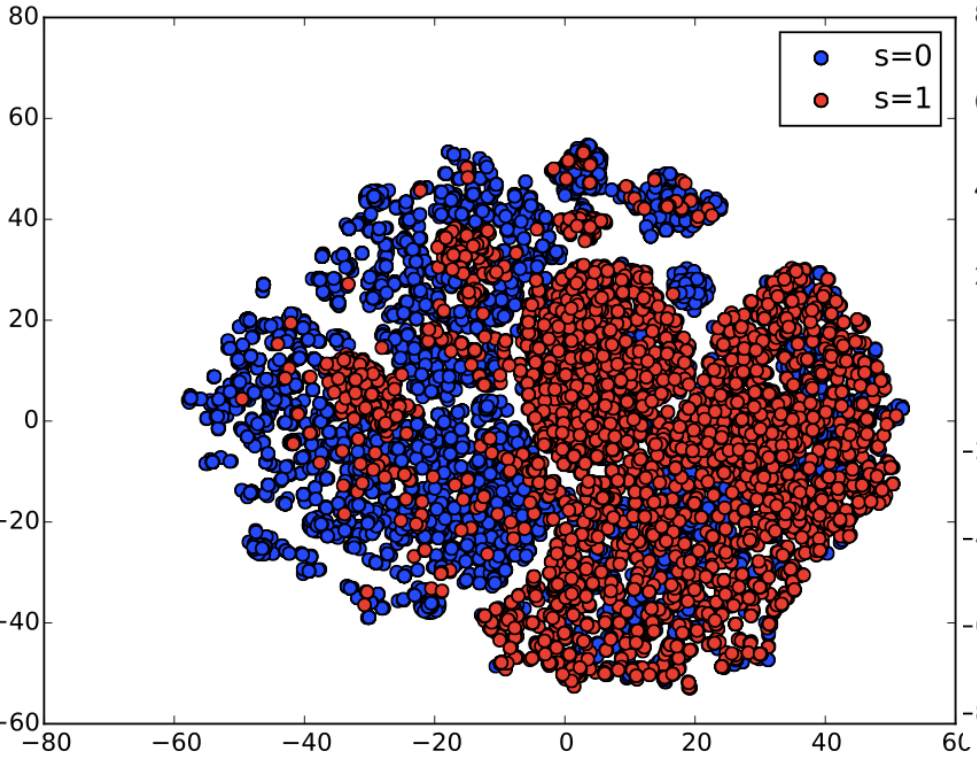
RESULTS: FAIR CLASSIFICATION

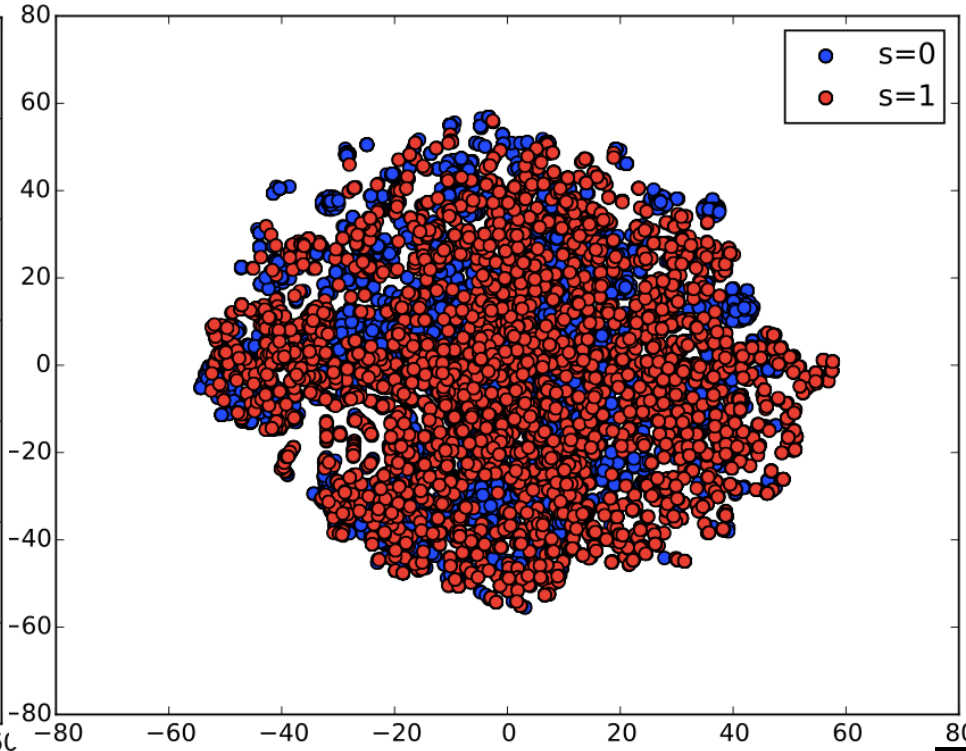
Compare deep network with/out fairness criteria (MMD)



RESULTS: OBFUSCATING *S*

Compare user representations without/with fairness:





NEW DATASETS

VoteCompass:

- Website surveys people on their political beliefs
- Tells them which political party their views align most closely with

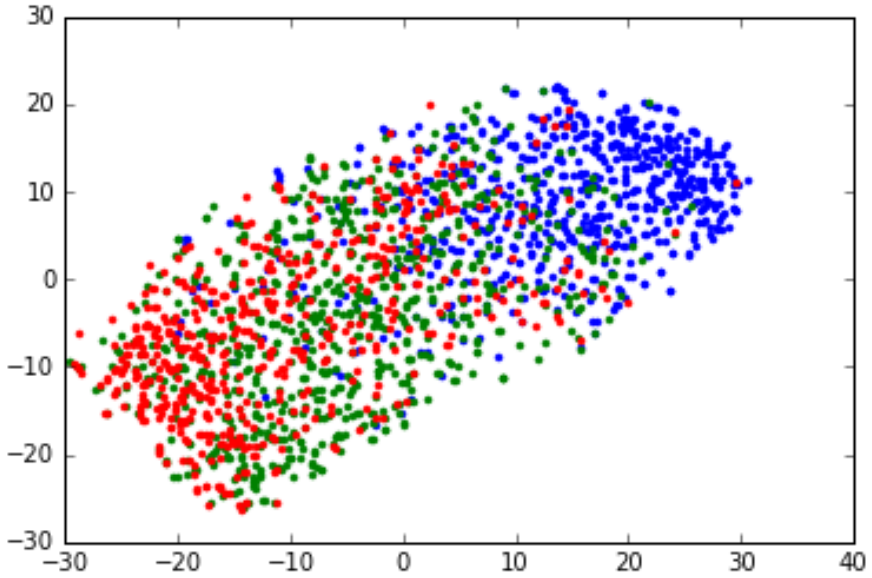
1. **The national budget deficit should be reduced, even if it means fewer public services.**
2. **Students in government and non-government schools should receive the same amount of federal funding.**
3. **How much should the federal government do to tackle climate change?**

FAIR TARGETED ADS

Advertise to respondents, without bias

Example:

- Favorable to party (Y)
- Fair with respect to religion (S)
- Before applying fairness criteria, strong clusters: Liberal-National, Conservative, Greens



CONCLUSION & DISCUSSION

1. **Cannot leave it all up to the algorithm:** **Need to specify aims, criteria**

2. **Inherent trade-off:** **Society's aims of avoiding bias (public utility) vs. decision "accuracy" (private utility)**

3. **What to do about it on an individual basis?**

4. **Refining definition, objectives of fairness: work with legal scholars, public policy experts**
   - Is statistical parity, or quotas, the right goal?
   - Can we help define, formulate the objective (do we know the sensitive variables?)

BIAS IN NEWS: UNAVOIDABLE?

"Facebook's algorithm…prioritizes the stories that should be shown to Facebook users in the trending section. The curators write headlines and summaries of each topic, and include links to news sites. The section… constitutes some of the most powerful real estate on the internet and helps dictate what news Facebook's users—167 million in the US alone—are reading at any given moment.

…workers prevented stories about the right-wing CPAC gathering, Mitt Romney, Rand Paul, and other conservative topics from appearing in the highly-influential section,…instructed to artificially "inject" selected stories…

In other words, Facebook's news section operates like a traditional newsroom, reflecting the biases of its workers and the institutional imperatives of the corporation. Imposing human editorial values onto the lists of topics an algorithm spits out is by no means a bad thing—but it is in stark contrast to the company's claim that the trending module simply lists 'topics that have recently become popular on Facebook.' "

- **Former Facebook Workers: We Routinely Suppressed Conservative News** (May 9, 2016)

THANKS!