Preventing Fairness Gerrymandering in Machine Learning

Michael Kearns*, Seth Neel*, Aaron Roth*, Steven Wu^

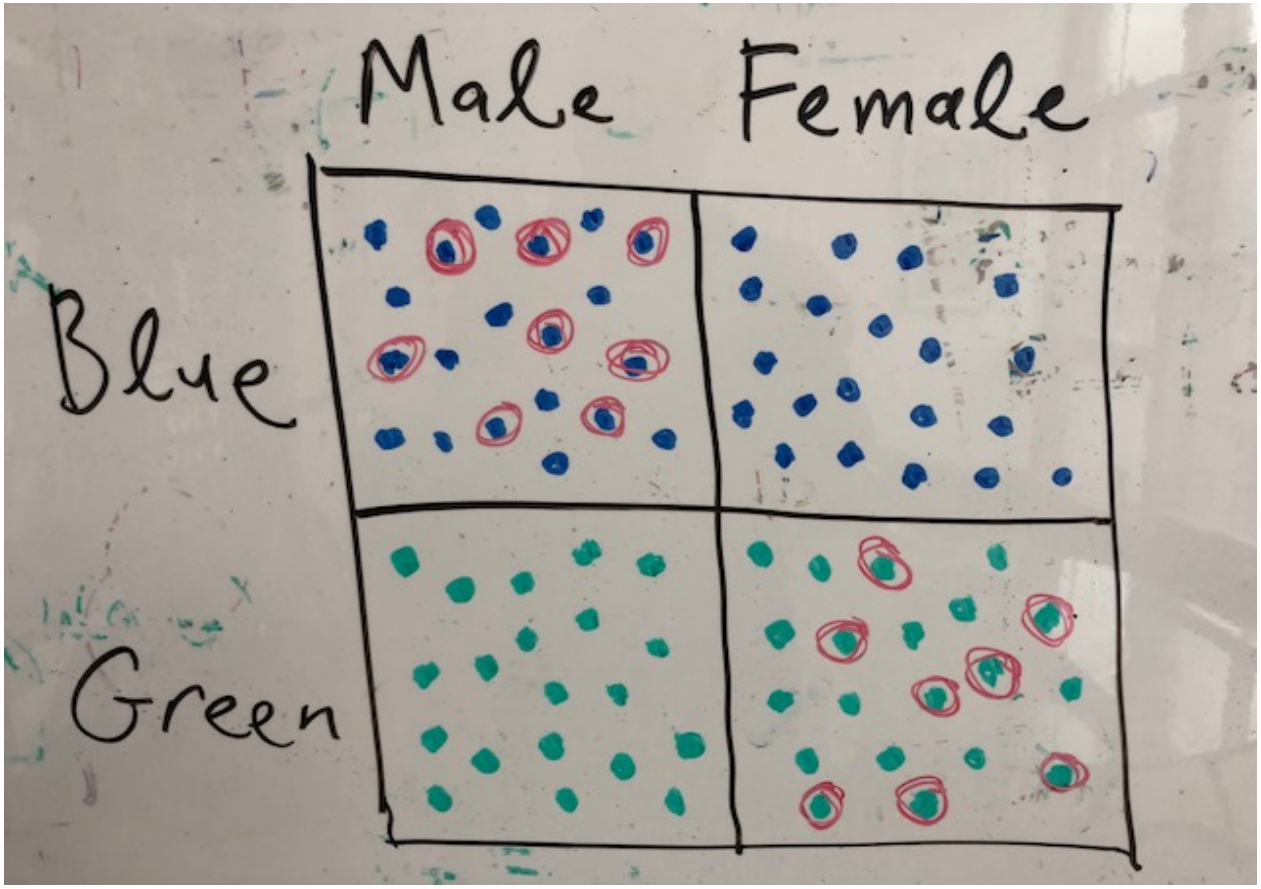
* University of Pennsylvania
^ Microsoft Research NYC

Warren Center Meet and Greet
February 20, 2018

Fairness in ML/Stats

- Identify the groups/attributes you want to protect: e.g. race, gender, age…
- Choose a (statistical) measure of equality or fairness
  - statistical parity (approx. equal loan rates across groups, ignoring creditworthiness)
  - equality of false positive/negative rates (accounts for merit)
  - calibration (estimate 30% repayment → actual 30% repayment)
- Try to design models minimizing error subject to (approximate) fairness
- *No promises made to individuals or finer-grained groups*

"Fairness Gerrymandering": The Cartoon



"Fairness Gerrymandering": The Reality

- Communities and Crime dataset:
  - census and other data on 2K U.S. communities
  - target prediction: high vs. low violent crime rate
  - 122 features total; 18 protected (racial group pct, incomes, police)
  - fairness notion is false positive
- Ran standard ML algo constrained by fairness wrt 18 features separately
- Quickly finds accurate classifier with less than 0.03 FP disparity
- But "Auditor" finds *subgroup* of weight 0.67 with FP disparity 0.26
- May be exponentially/infinitely many potentially discriminated subgroups

"Fairness Gerrymandering": The Solution

- Designing subgroup-fair learning algorithms:
  - formulate as a 2-player, repeated, zero-sum game
  - Learner has pure strategy space H of hypotheses
  - Auditor has pure strategy space G of subgroups
  - Learner objective: minimize error subject to (approximate) fairness wrt G
- Theorem (Informal): Provably (and rapidly) convergent learning algorithm for H that is fair wrt all subgroups in G.

"Fairness Gerrymandering": The Experiments

- Communities and Crime dataset
- Both H (d = 122) and G (d = 18) are linear threshold functions
- One free parameter C, bound on Auditor's (dual player) variables