

FRONTIERS OF FAIRNESS IN MACHINE LEARNING

MICHAEL KEARNS
UNIVERSITY OF PENNSYLVANIA

ARTIFICIAL INTELLIGENCE IN
CONSUMER FINANCE
FEDERAL RESERVE BANK
NOVEMBER 9, 2021

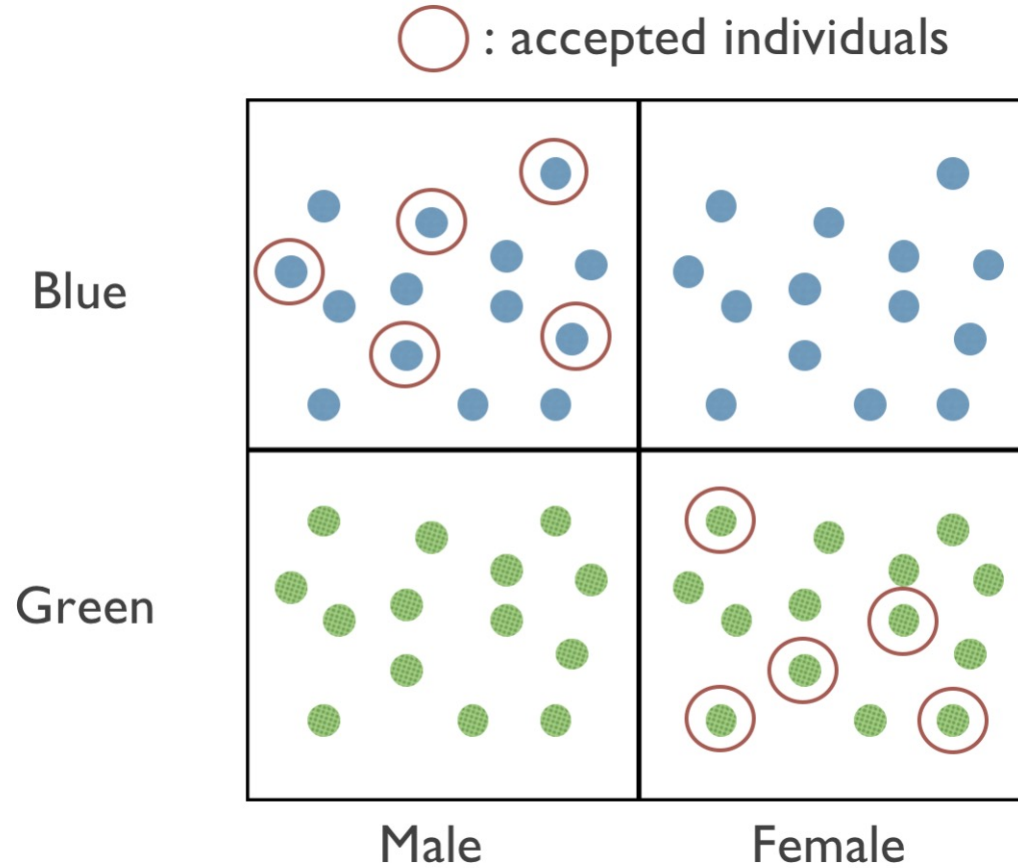
TYPES OF FAIRNESS DEFINITIONS

- Group Fairness
 - E.g. equality of error or false negative rates across gender, racial groups, etc.
 - Strong theory and algorithms, practical implementations
 - But no guarantees to *individuals*
- Individual Fairness
 - E.g. metric fairness (“fairness through awareness”), meritocratic fairness
 - Binds at the individual level
 - But strong (non-statistical) assumptions required have prevented practical *implementations*
- What about *interpolations*?

A FRAMEWORK FOR FAIR ML

- Begin by expressing training as a constrained optimization problem
 - E.g. minimize error subject to various fairness constraints
- Recast as two-player, zero-sum game
 - Learner: wants to minimize overall error
 - Regulator: enforces constraints, allowing violations *less than γ*
 - Nash equilibrium is solution to constrained optimization problem
- If we can:
 - Formulate best responses as instances of *standard classification*
 - Implement at least one player as a *no-regret* algorithm w.r.t. their strategy space... then algorithm *provably converges*
- Directly implement on top of your favorite *non-fair* learning heuristic
- Applications:
 - Preventing “fairness gerrymandering”
 - Average individual fairness
 - Subjective individual fairness
 - Minimax and lexicographic fairness
 - Downstream proxies

PREVENTING “FAIRNESS GERRYMANDERING”

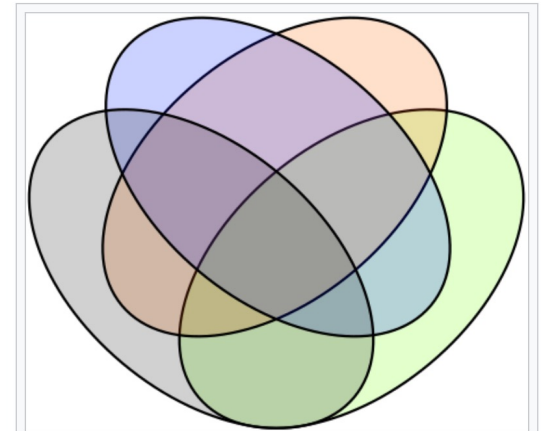



Intersectionality

From Wikipedia, the free encyclopedia

Intersectionality is an [analytical framework](#) for understanding how aspects of a person's [social and political identities](#) combine to create different modes of [discrimination](#) and [privilege](#). Examples of these aspects include [gender](#), [caste](#), [sex](#), [race](#), [class](#), [sexuality](#), [religion](#), [disability](#), [physical appearance](#),^{[1][2]} and [height](#).^[3] Intersectionality identifies multiple factors of advantage and disadvantage.^[4] These intersecting and overlapping social identities may be both empowering and oppressing.^{[5][6]} For example, a black woman might face discrimination from a business that is not distinctly due to her [race](#) (because the business does not discriminate against black men) nor distinctly due to her [gender](#) (because the business does not discriminate against white women), but due to a combination of the two factors.

Intersectionality broadens the lens of the [first](#) and [second waves of feminism](#), which largely focused on the experiences of women who were both [white](#)



An intersectional analysis considers  all the factors that apply to an individual in combination, rather than considering each factor in isolation.

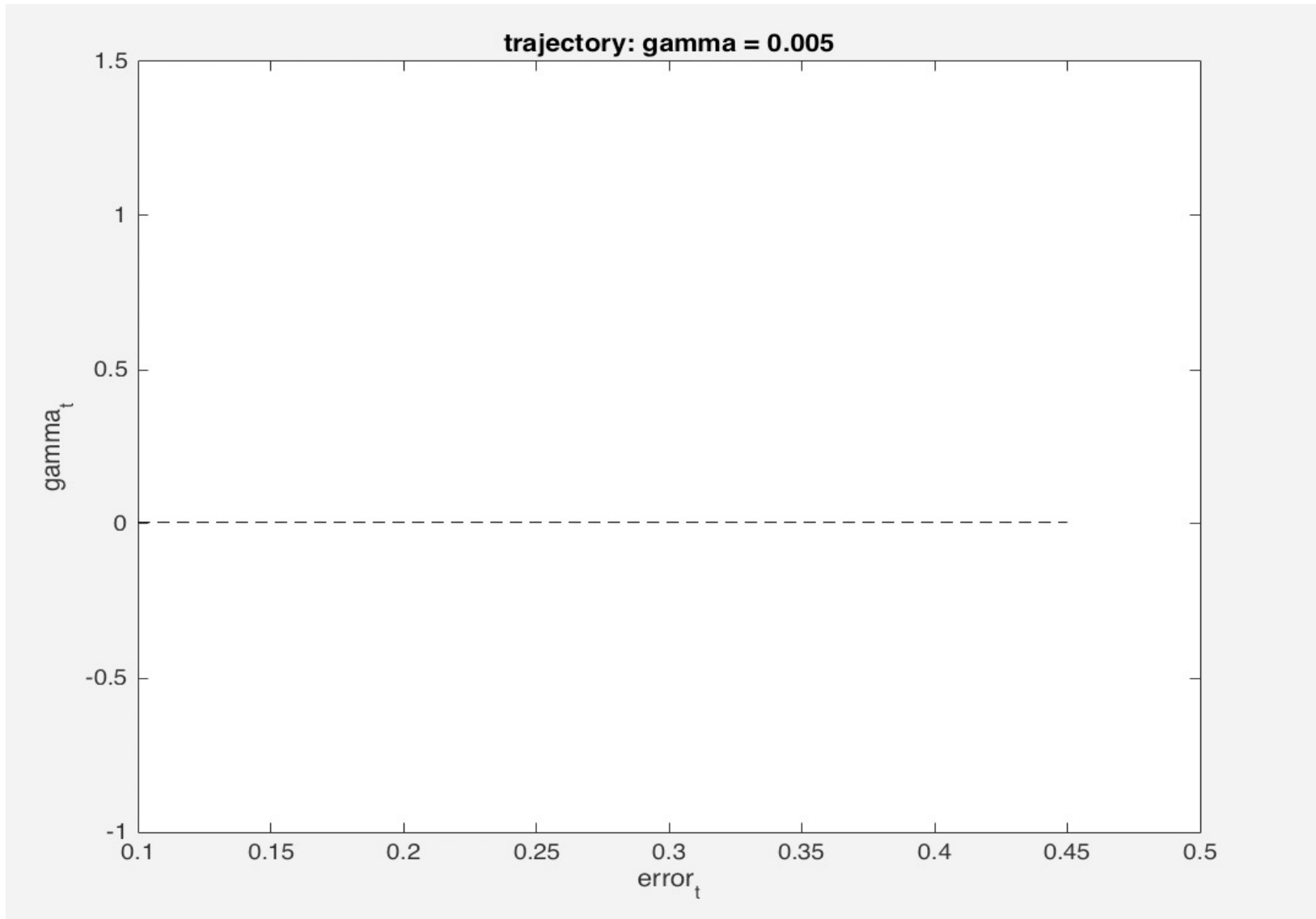
INTERPOLATING BETWEEN GROUPS AND INDIVIDUALS

- Problem: achieving group fairness by subgroup discrimination
 - E.g. disabled Hispanic women over age 55 earning less than \$25K
 - N.B. Facebook hate speech policy
 - No reason to expect it won't happen under standard fairness notions
- But cannot generally protect arbitrarily refined subgroups (e.g. individuals)
- Constrained optimization problem:

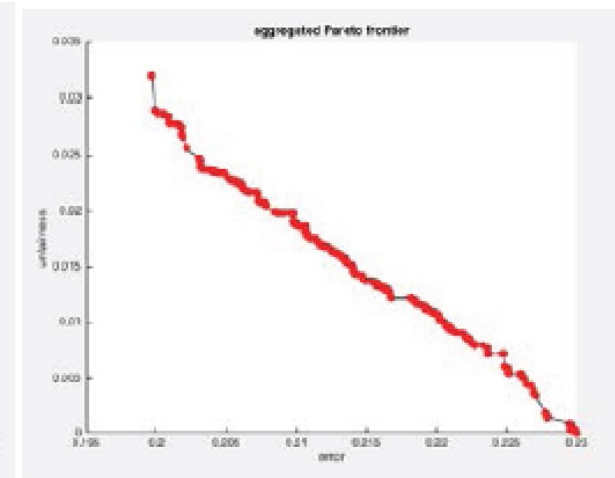
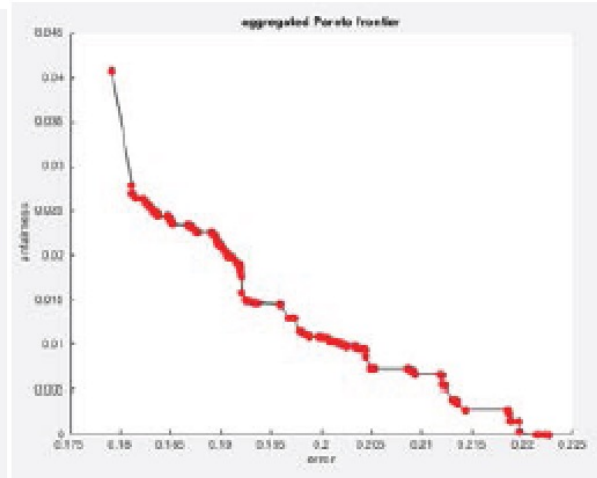
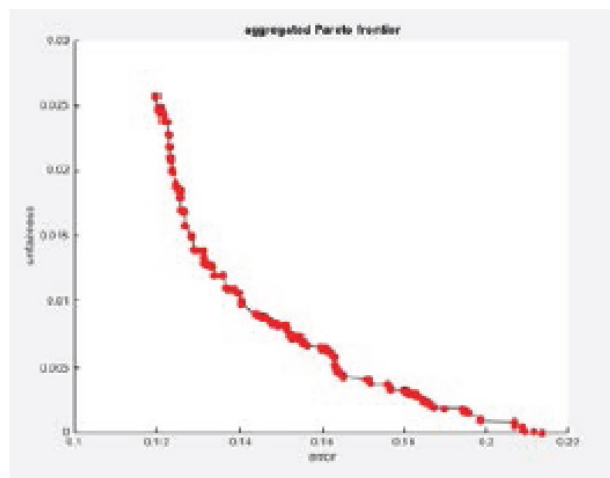
$$\min_{D \in \Delta_{\mathcal{H}}} \mathbb{E}_{h \sim D} [\text{err}(h, \mathcal{P})]$$

such that $\forall g \in \mathcal{G} \quad \alpha_{FP}(g, \mathcal{P}) \beta_{FP}(g, D, \mathcal{P}) \leq \gamma.$

ERROR-UNFAIRNESS TRAJECTORY



EFFICIENT FRONTIERS

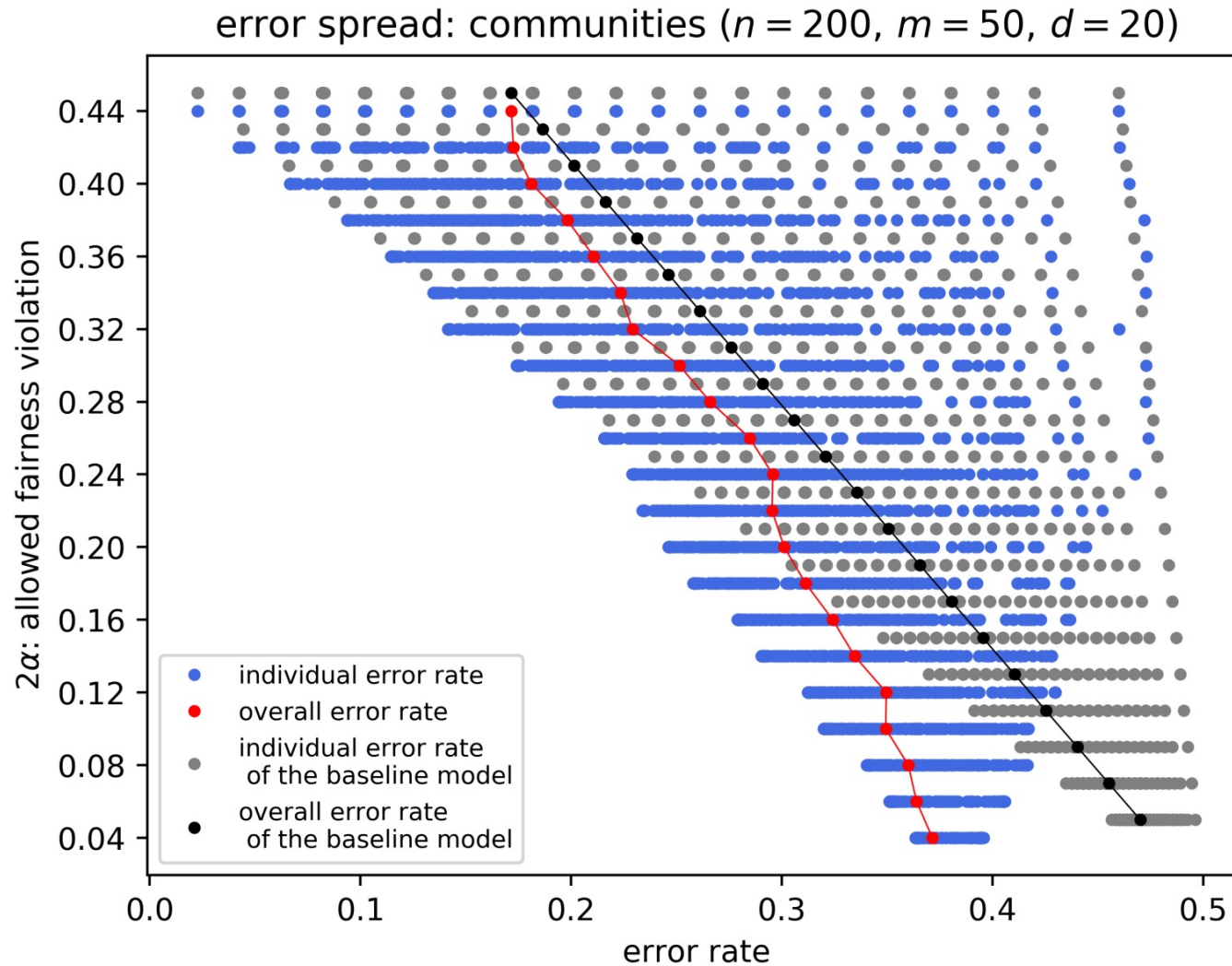


Examples of Pareto frontiers of error (x axis) and an unfairness measure (y axis) for three different real data sets. The curves differ in their shapes and the actual numeric values on the error and fairness axes, thus presenting different trade-offs.

AVERAGE INDIVIDUAL FAIRNESS

- Imagine we make *many* predictions/decisions about each individual
- E.g. product recommendations, ads, image labels
- Now sensible to talk about error *rate* for an individual *across predictions*
- Fairness constraint: individual error rates (approximately) equalized
- Binds at individual level

ERROR-AIF TRADEOFFS



SUBJECTIVE INDIVIDUAL FAIRNESS

- What if fairness is *subjective* and *complex*?
- *Elicit* pairwise fairness constraints from subjects/stakeholders/committee
- “A and B should receive same outcome”
- “A should receive at least as good an outcome as B”
- A *distributional* form of individual fairness

FAIRNESS ELICITATION

sex	age	race	juv. felony count	juv. misdemeanor count	juv. other count	priors count	severity of charge
Male	22	Caucasian	0	0	0	2	Felony

vs.

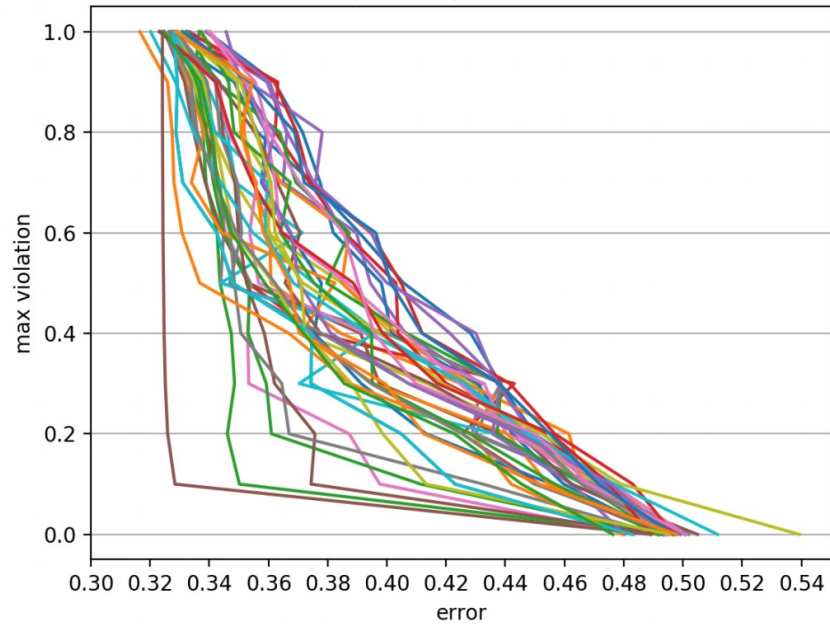
sex	age	race	juv. felony count	juv. misdemeanor count	juv. other count	priors count	severity of charge
Male	35	African-American	0	0	0	1	Felony

Should be treated equally

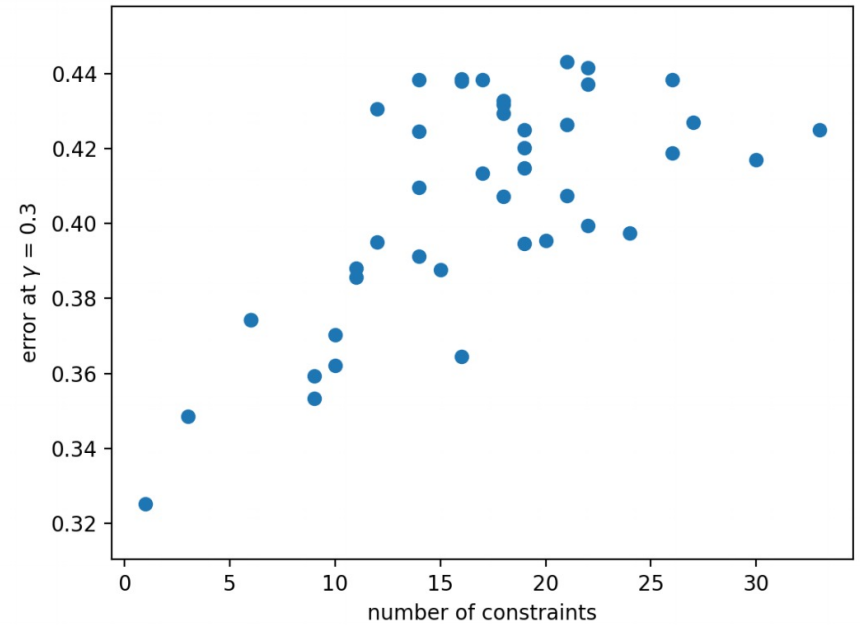
Ok to treat differently, or no opinion

INTERPERSONAL VARIABILITY

Variability of Subject Pareto Curves



correlation = 0.681134



OTHER APPLICATIONS

- *Minimax* group fairness
 - Prevent artificial inflation of lower group errors
 - Pareto dominates equalization notions
- *Lexicographic* group fairness
 - Minimax to its logical extreme
- Proxies for *downstream* fairness
 - When sensitive attribute not available
 - “Non-disclosive” proxies?

THANKS!