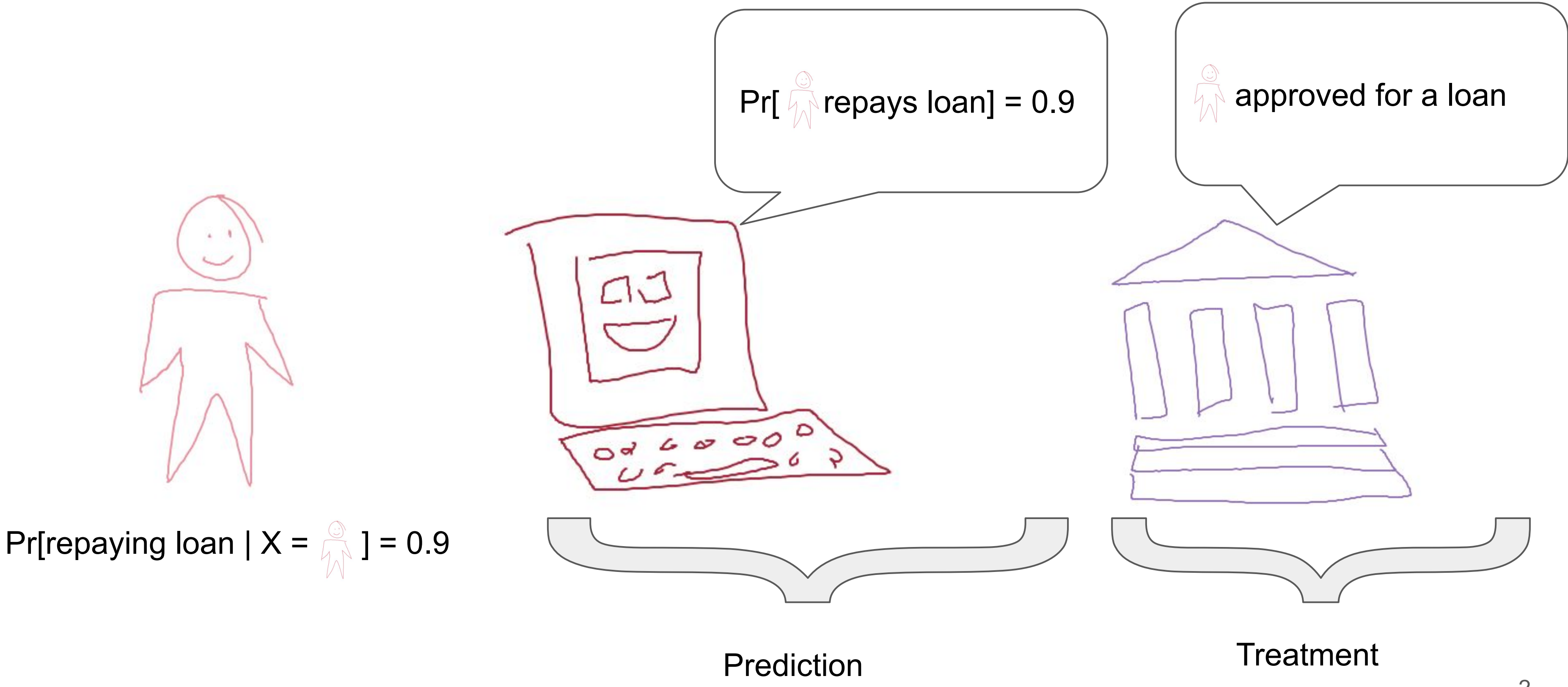
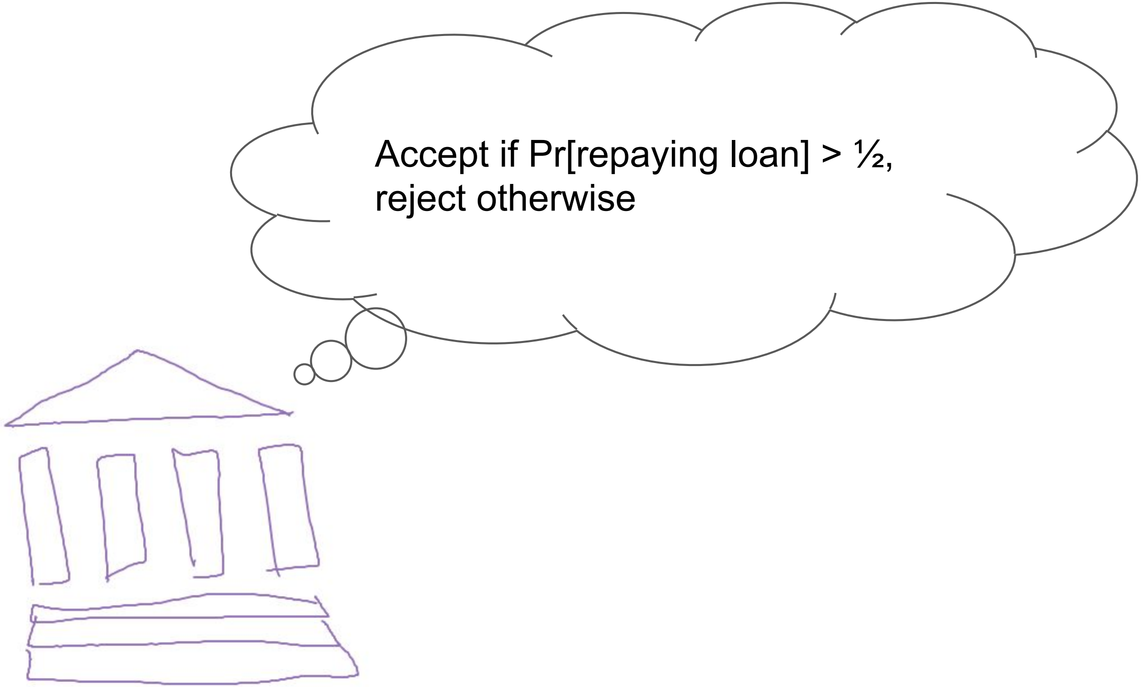
Visualizing our values: using property elicitation to understand the consequences of constraints

Jessie Finocchiaro
Harvard CRCS / NSF Math
10 July 2023
EC Gender Inclusion Workshop

Make predictions about people all the time



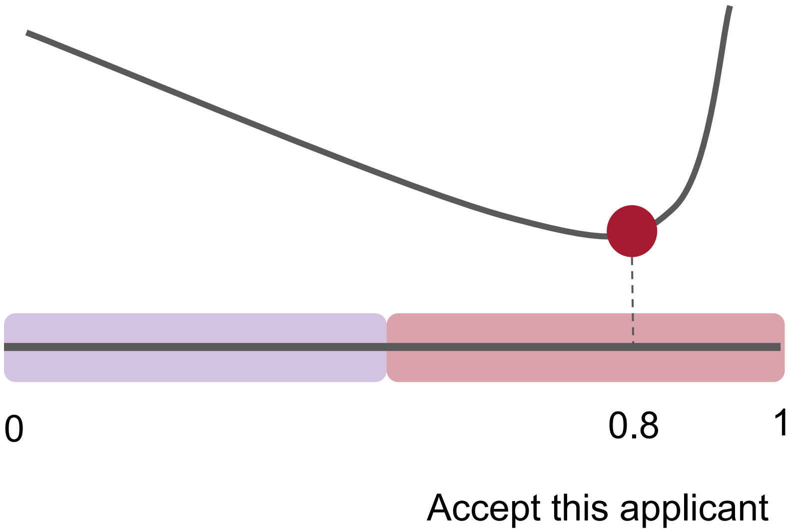
Treatments reflect some summary statistic of belief



Treatments reflect some summary statistic of belief

If Pr[repaying]...
In [0, ½), reject
In [½, ¾), reject with expedited reapplication
Over ¾, accept

Design loss functions to elicit such statistics

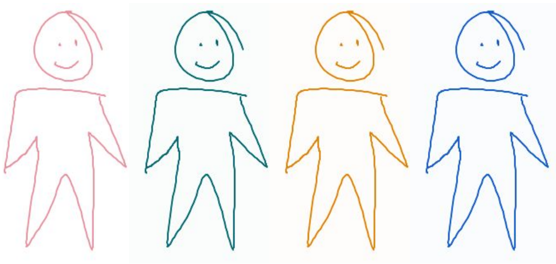


Set of outcomes Y	Y = {repay, default}
True p $\in \Delta_Y$	p = Pr[repay] = 0.8
Set of predictions U	U = [0,1]
Set of treatments T	T = {award loan, reject loan}

What happens when we think about the population: adding regularizers

When treatments are individual, simply consider each treatment individually

$$\min_{\vec{u}} L(\vec{u};\vec{p}) := \frac{1}{m}\sum_{i=1}^{m} L(u_i, p_i)$$



Fairness concerns often merit adding regularizers to losses

$$\min_{\vec{u}} L^{\lambda,R}(\vec{u};\vec{s};\vec{p}) := (1-\lambda)\frac{1}{m}\sum_{i=1}^{m} L(u_i, p_i) + \lambda R(\vec{u};\vec{s};\vec{p})$$

Now we need to consider population as a whole, and cannot abstract decisions to the individual level

Property elicitation

A loss L elicits a property Γ if, for all p∈$\Delta^m_{\mathcal{Y}}$,

$$\Gamma(\vec{p}) = \arg\min_{\vec{u}} L(\vec{u}; \vec{p})$$

Since L is additive in u, this decomposes into $\{\Gamma(p_i)\}_i$

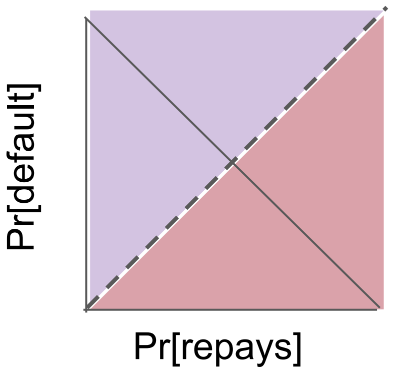
Fix s. A regularized loss elicits a regularized property ϴ if, for all p in $\Delta^m_{\mathcal{Y}}$,

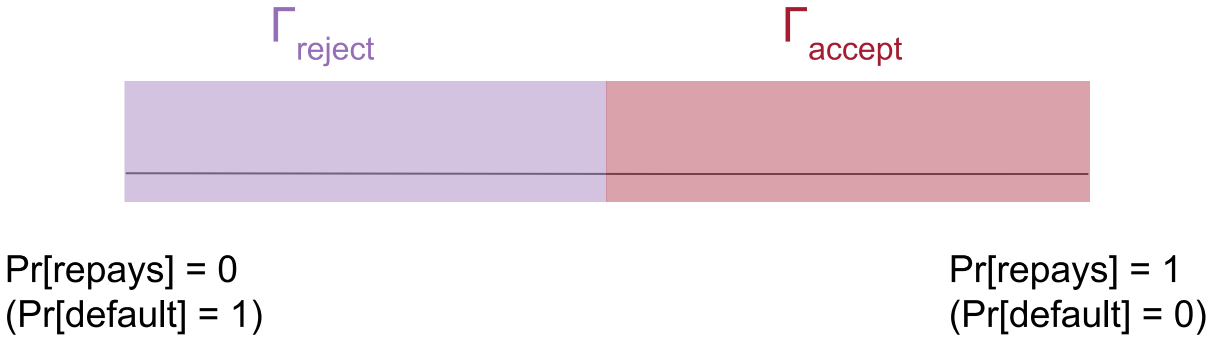
$$\Theta(\vec{p}) = \arg\min_{\vec{u}} L^{\lambda,R}(\vec{u}; \vec{s}; \vec{p})$$

Level sets of properties

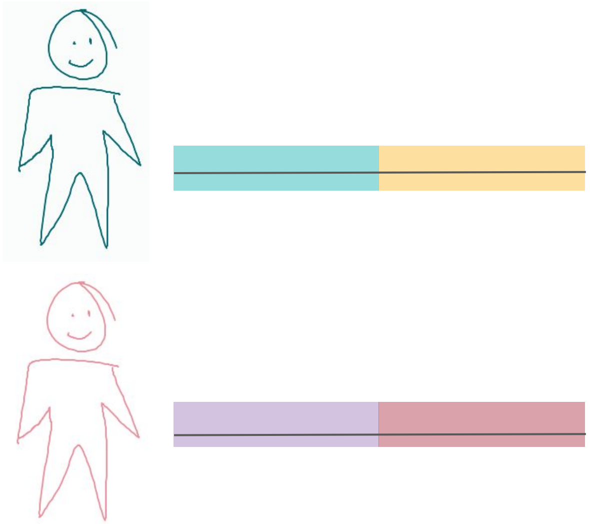
Predictions don’t have to be perfect, so long as treatments are correct

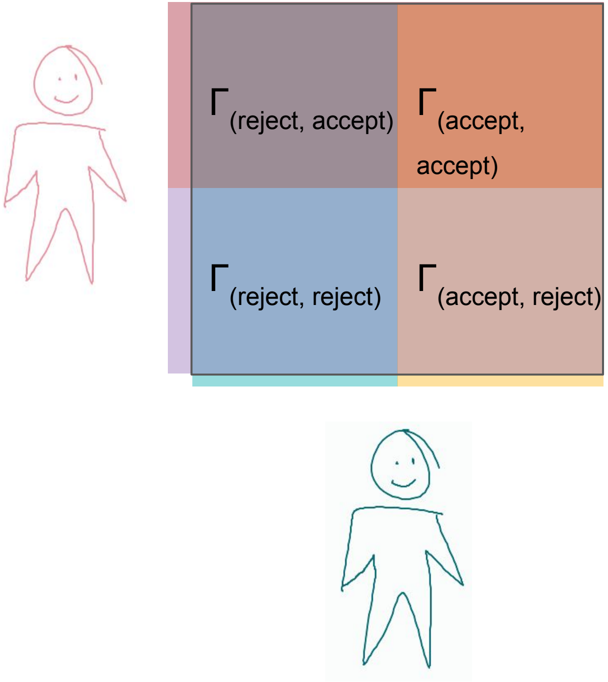
$$\Gamma_t = \{\vec{p} \in \Delta_{\mathcal{Y}}^m : t \in \Gamma(\vec{p})\}$$





Example visualization: 2 agents, binary classification

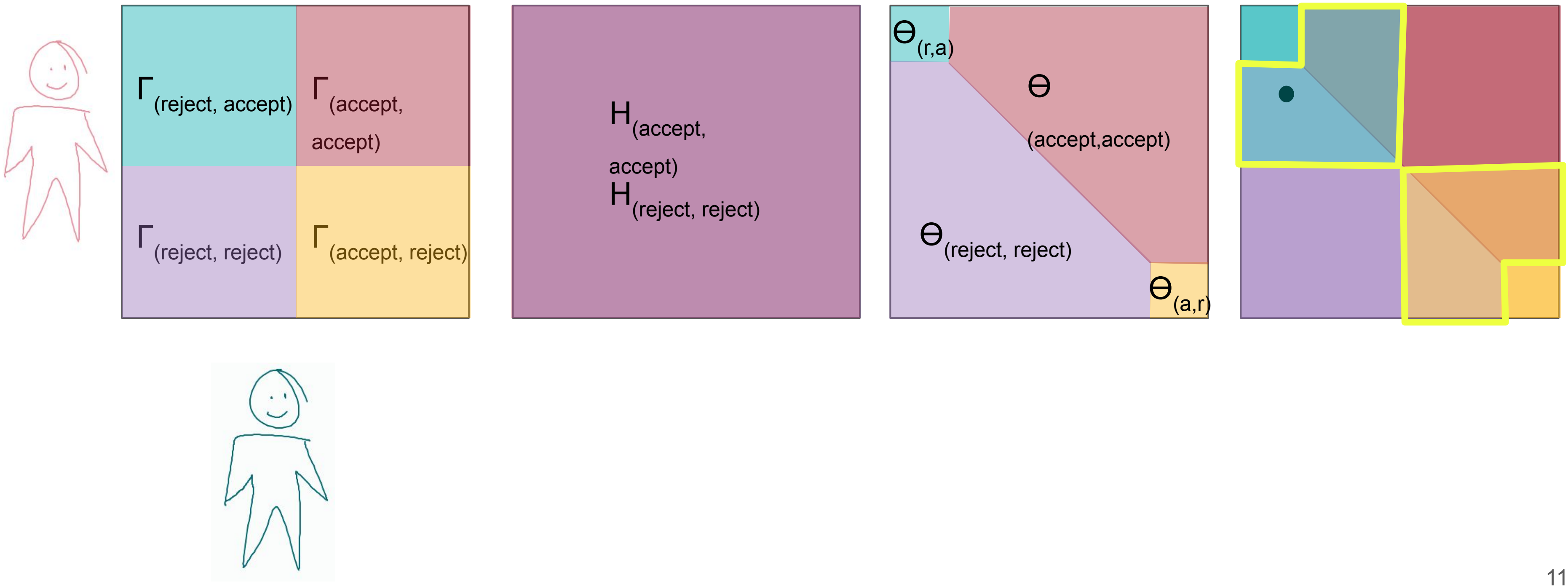




When do regularizers change the original property?

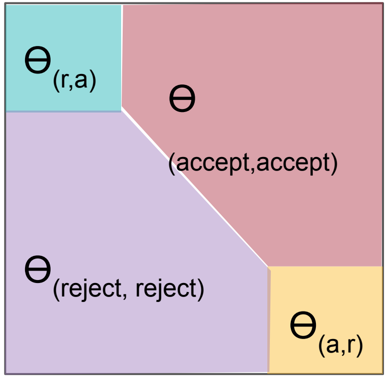
Theorem (informal): Fix $\lambda \in (0,1)$. Let L elicit Γ, $L^{R,\lambda}$ elicit Θ, and R (which is nonconstant) elicit H. Then $\Gamma = \Theta$ if and only if $H = \Gamma$.

Proof by picture: Counterexample with Demographic Parity

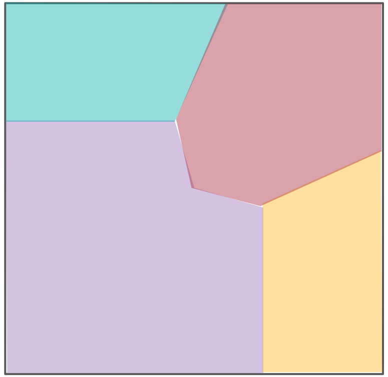


Corollary: common group fairness metrics change it up

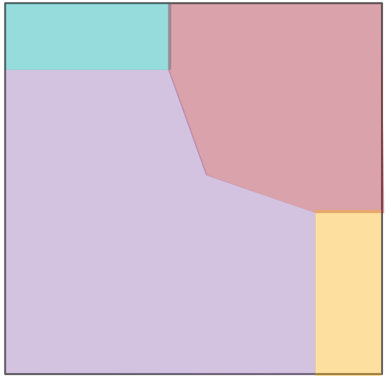
- Most group fairness regularizers change the property
  - They are not additive, so regardless of Γ
- Notable exception: calibration
  - Implies changes imposed by calibration constraints are a result of expressiveness of the model



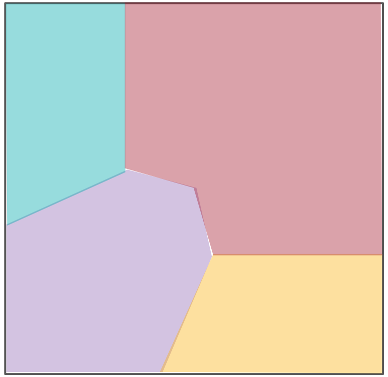
Demographic Parity



False Positive Rates

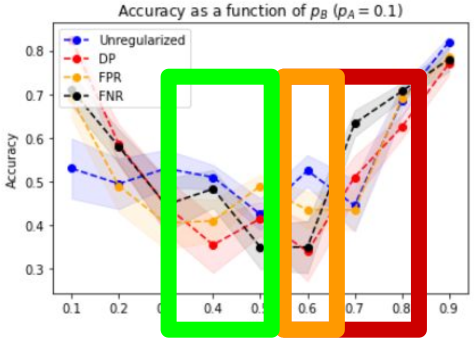


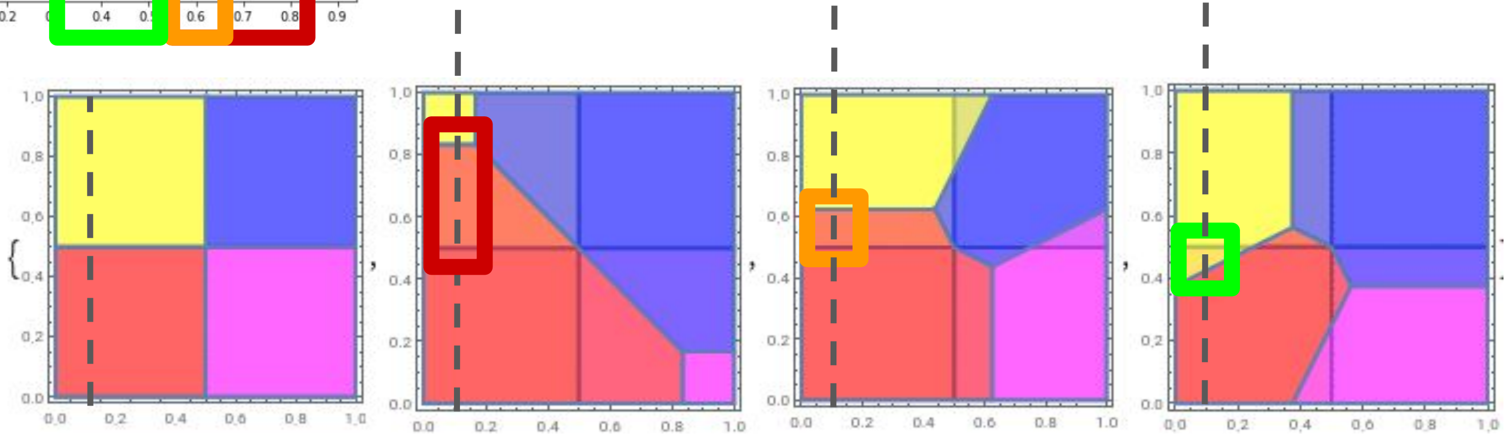
Equalized Odds



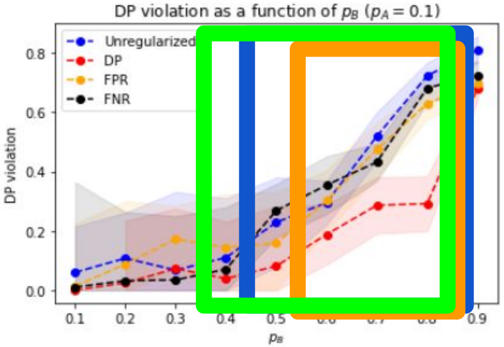
False Negative Rates

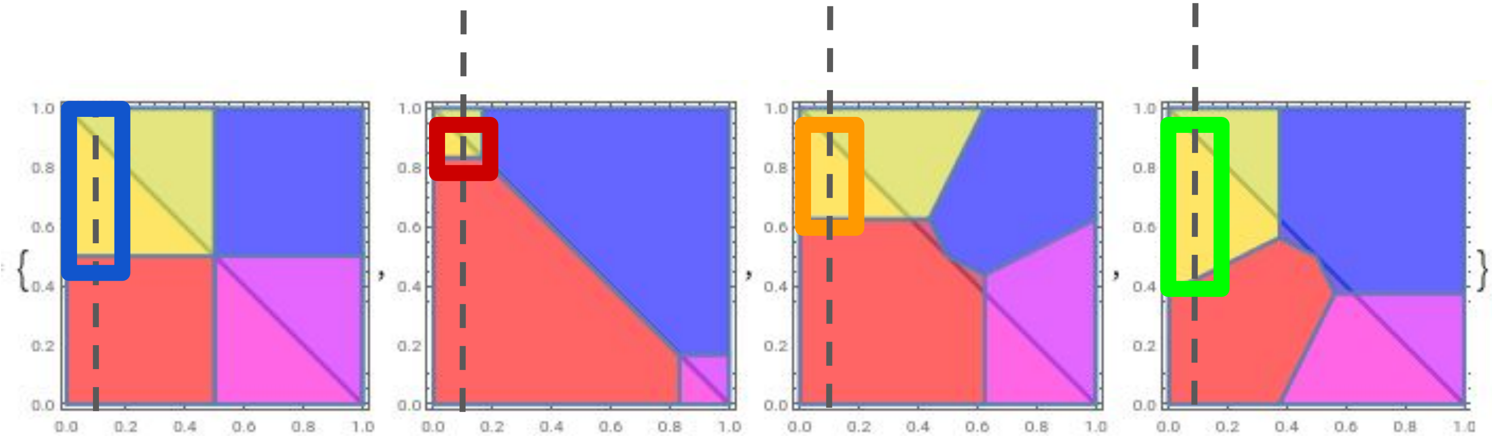
How decisions change as we go through distribution space





Fairness violations when regularized





In summary, come chat!

- Use high-dimensional property elicitation to study the impacts of different regularizers
    - Examples: group fairness constraints
- Can be used to explain performance gaps and translation across different fairness regularizers

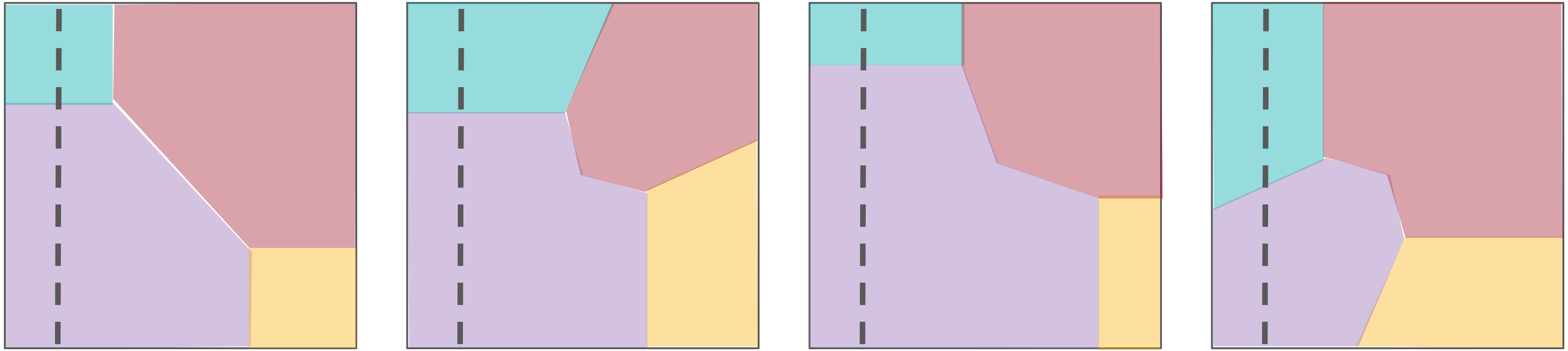
Interested in collaborating, questions?

Email: jessie@seas.harvard.edu

Online: www.jessiefin.com

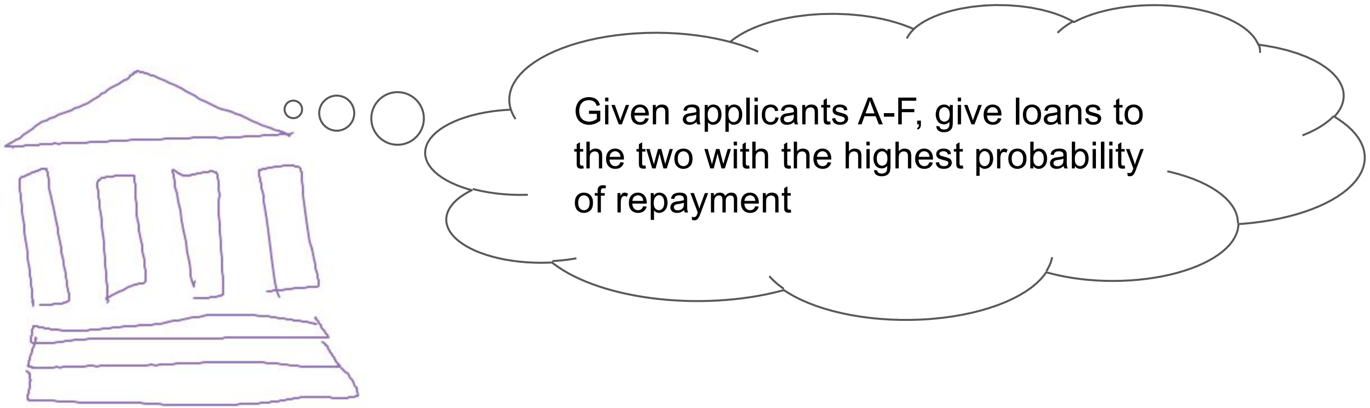
CRCS Center for Research on Computation and Society at Harvard John A. Paulson School of Engineering and Applied Sciences

Experimental results



Demographic Parity | False Positive Rates | Equalized Odds | False Negative Rates

Treatments reflect some summary statistic of belief



Treatments reflect some summary statistic of belief