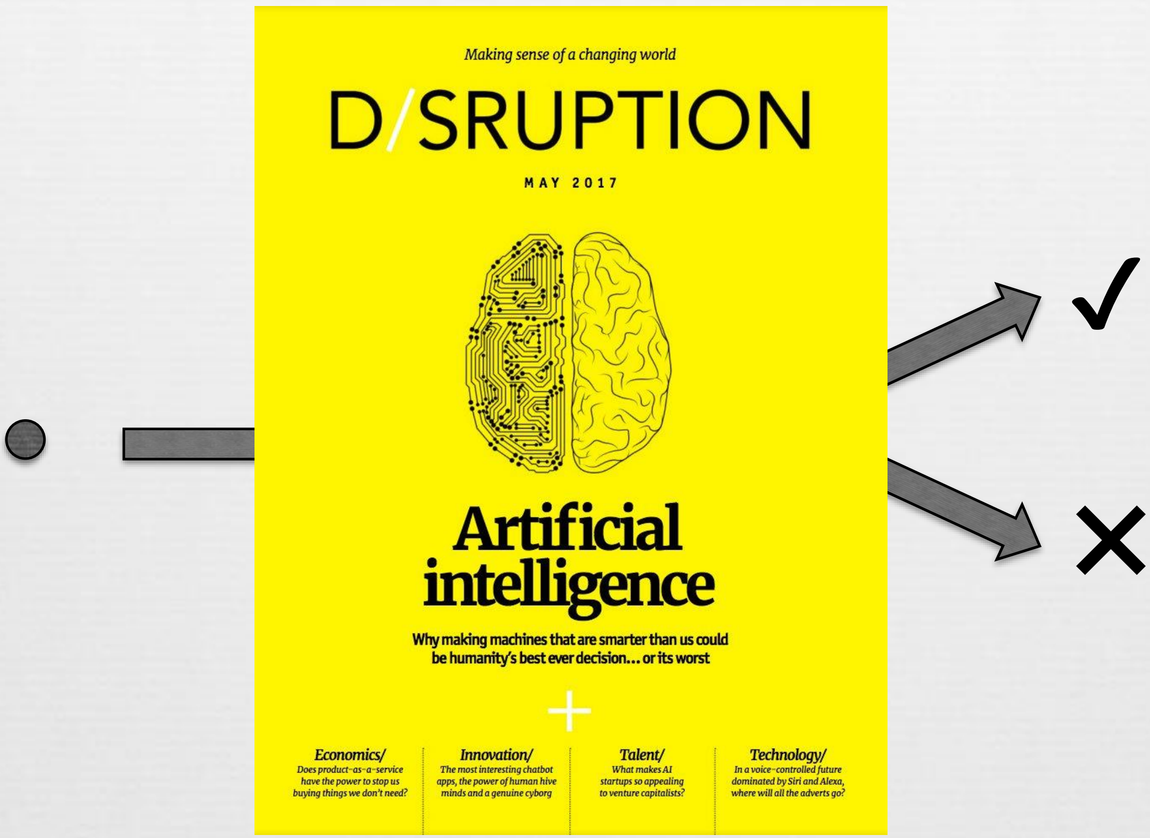
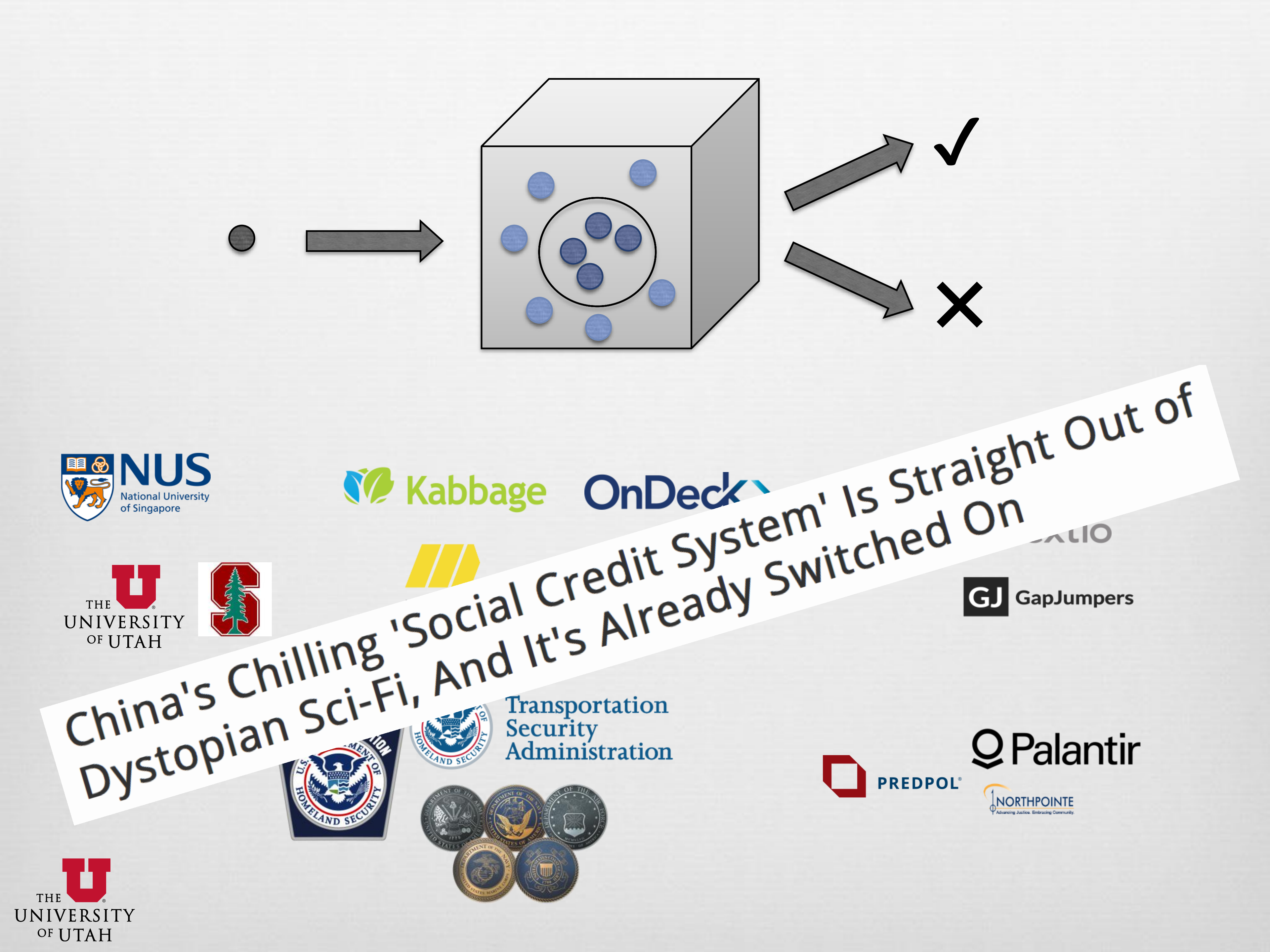
Algorithmic (un)fairness: a research agenda

Suresh Venkatasubramanian
University of Utah

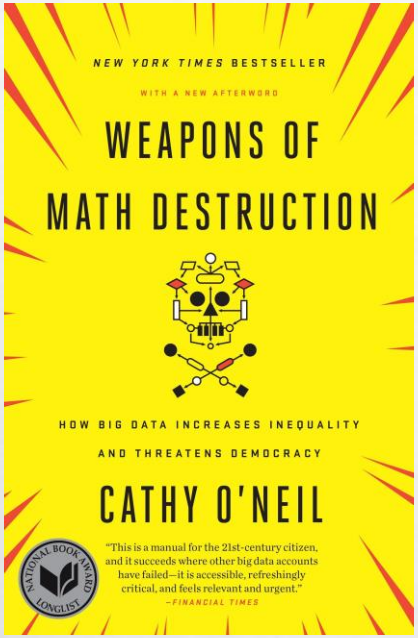
NUS, Singapore
Jan 9, 2019

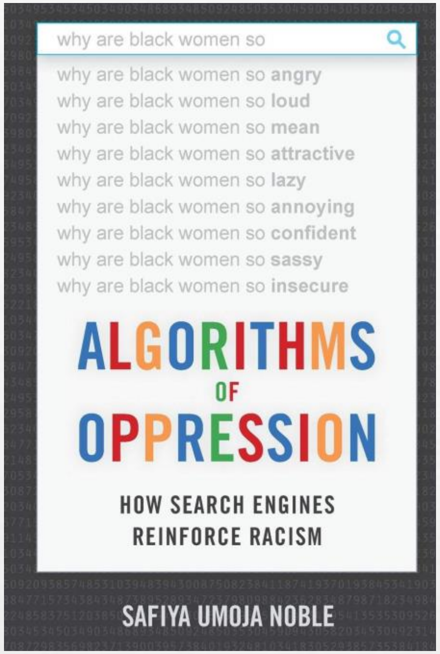


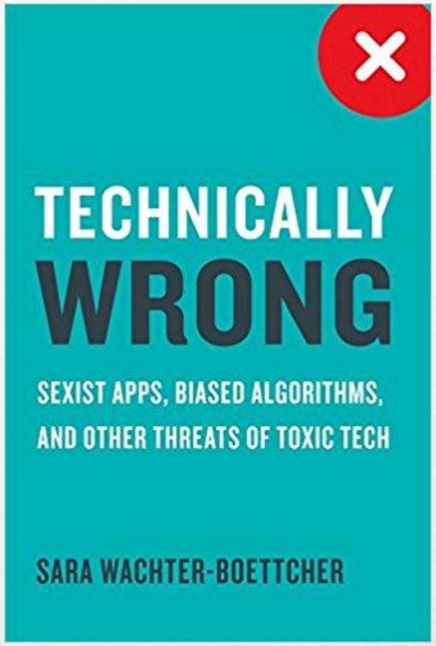
THE UNIVERSITY OF UTAH

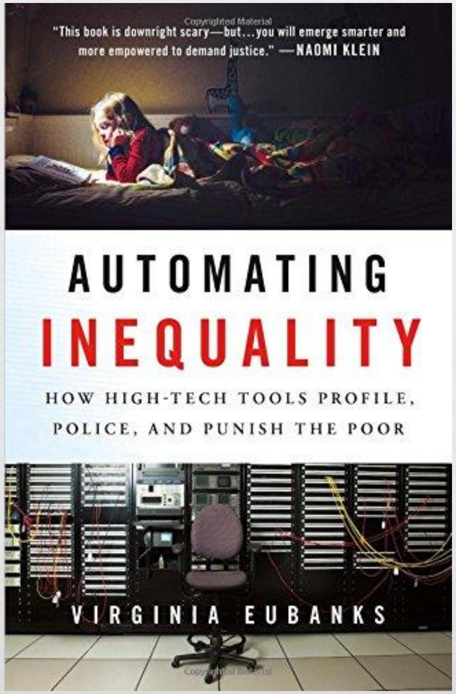


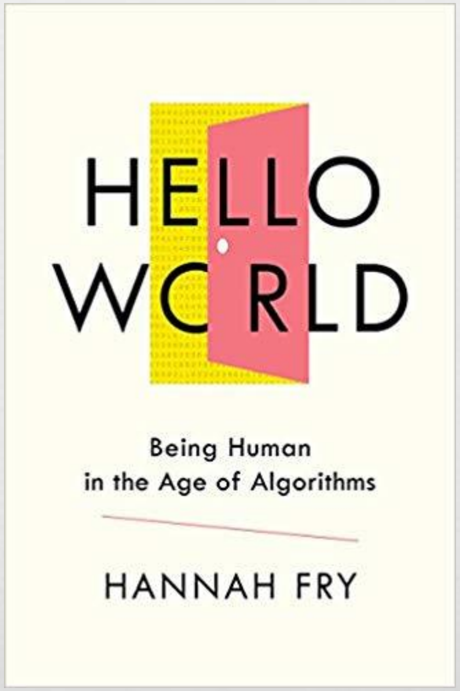
China's Chilling 'Social Credit System' Is Straight Out of
Dystopian Sci-Fi, And It's Already Switched On
NUS
National University of Singapore
Kabbage
OnDeck
GapJumpers
THE UNIVERSITY OF UTAH
Transportation Security Administration
PREDPOL
Palantir
NORTHPOINTE
THE UNIVERSITY OF UTAH



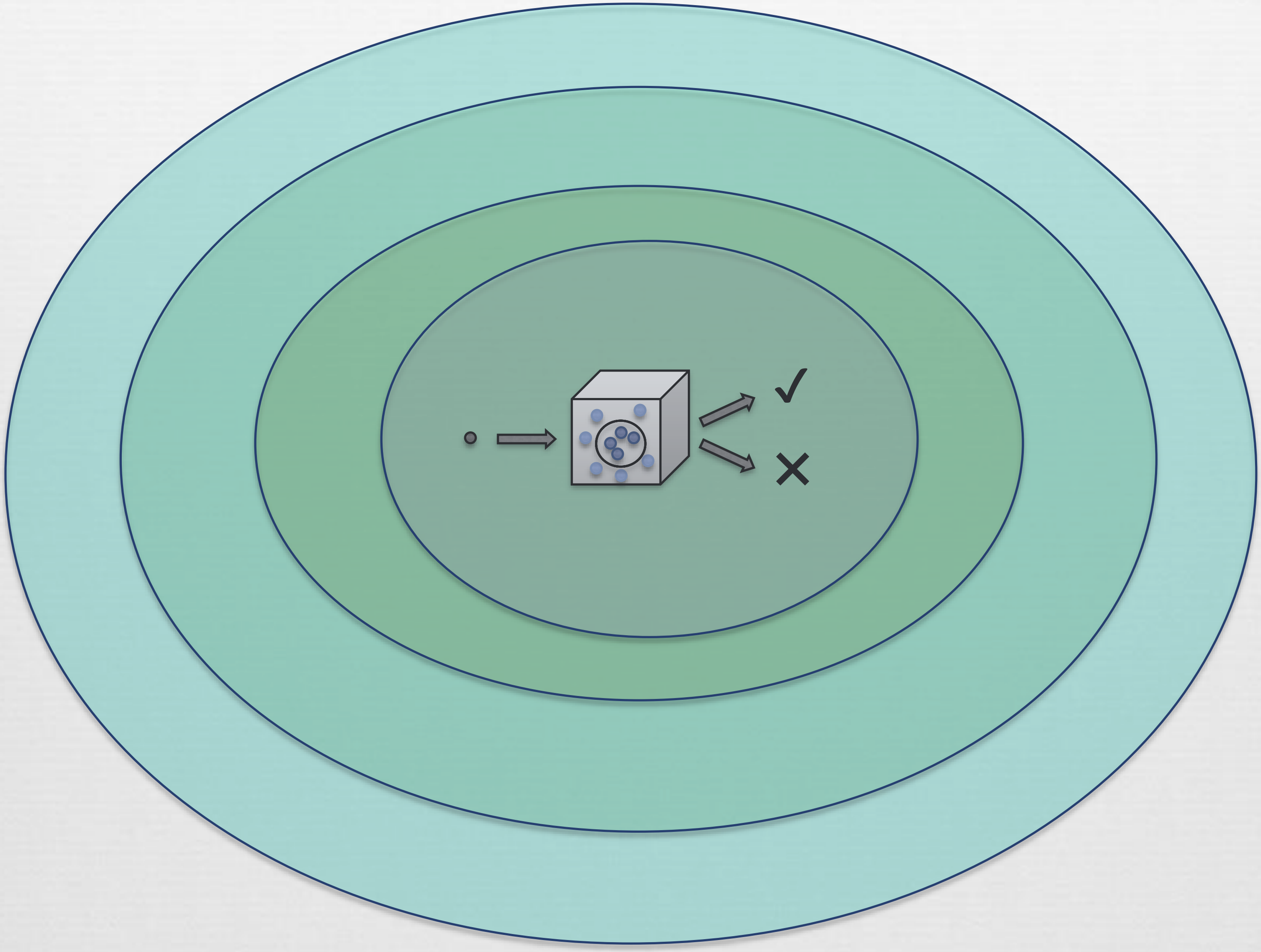








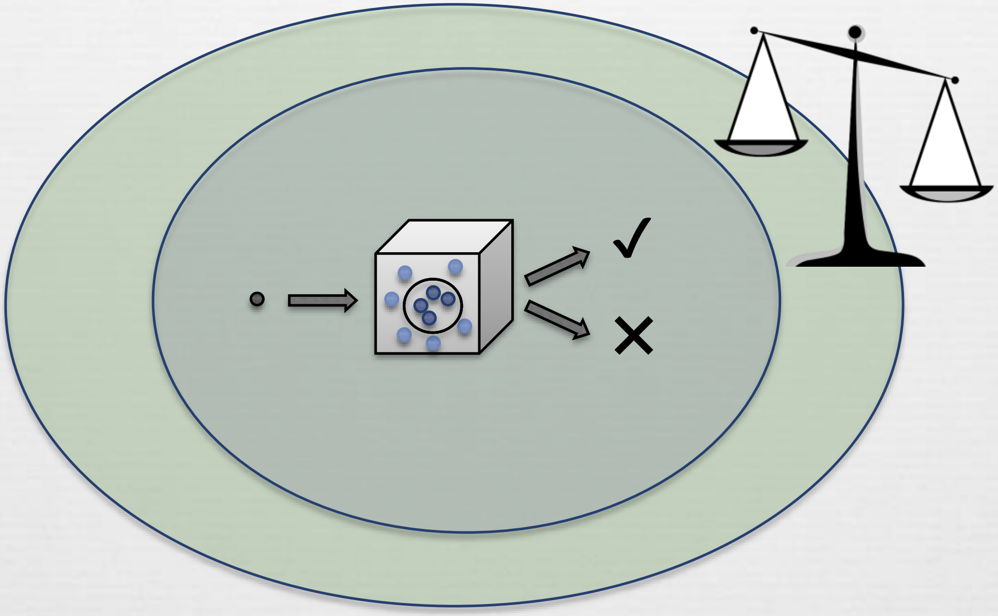
THE UNIVERSITY OF UTAH



Society

THE UNIVERSITY OF UTAH

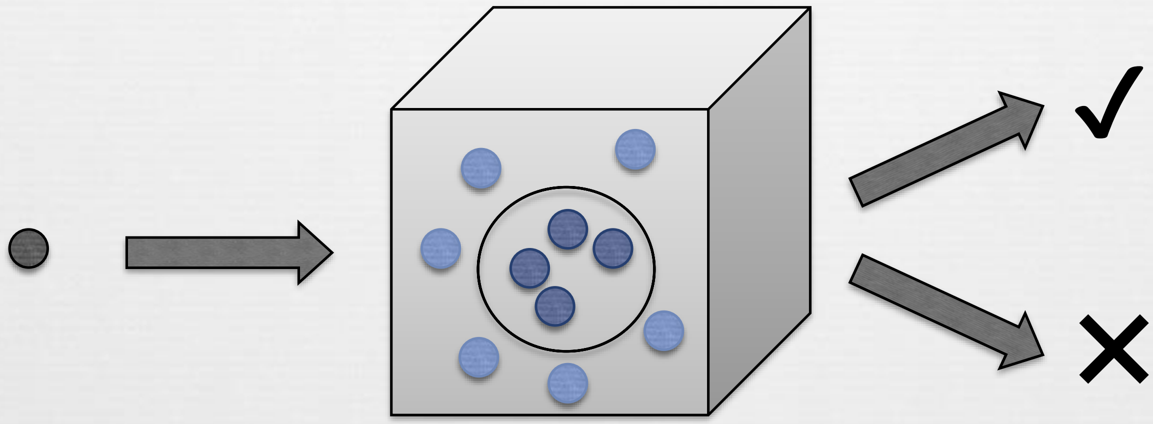
Expanding context: Fairness



THE UNIVERSITY OF UTAH

Definitions of fairness

A simple problem: classification



Hiring

College admission

Loan

Definitions of fairness

I treat you differently because of your race

Individual fairness

Individuals with *similar* abilities should be treated the *same*

Group fairness

Structural bias against groups

Groups should all be treated *similarly*

Definitions of fairness

- Individual fairness

$$d(x,y) \leq \epsilon \Rightarrow D(C(x),C(y)) \leq \delta$$

- Group fairness

$$|\Pr(C=1 \mid g=0) - \Pr(C=1 \mid g=1)| \leq \epsilon$$

$$\frac{\Pr(C=1 \mid g=0)}{\Pr(C=1 \mid g=1)} \geq 1-\epsilon$$

Definitions of fairness

- Individual fairness

$$d(x,y) \leq \epsilon \Rightarrow d(C(x),C(y)) \leq \delta$$

- Group fairness

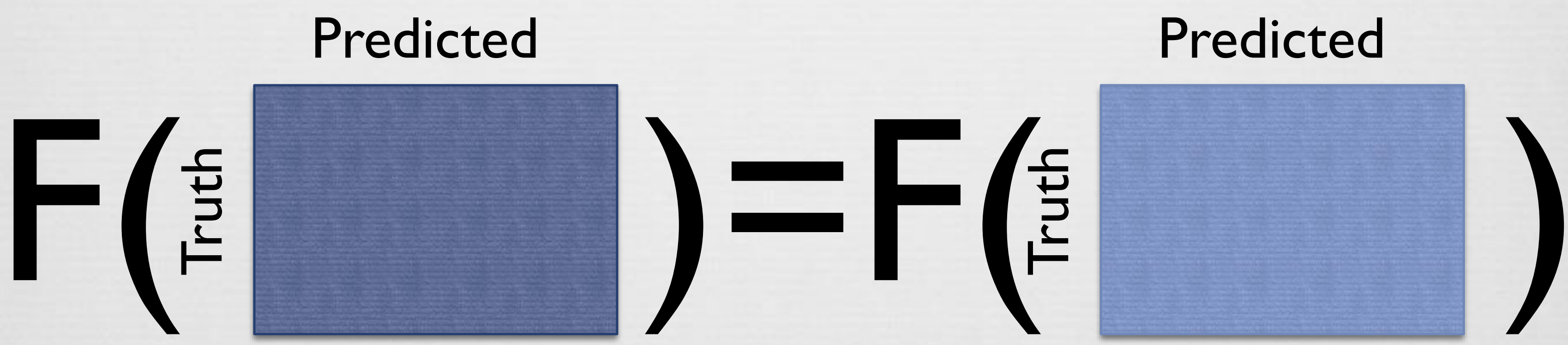
$$(FP,FN)_0 \simeq (FP,FN)_1$$

Definitions of fairness

- Individual fairness

$$d(x, y) \leq \epsilon \Rightarrow d(C(x), C(y)) \leq \delta$$

- Group fairness



Unifying notions of fairness

$$\mathbf{y} \perp \mathbf{a} \mid \{\mathbf{z} = z\}$$

[RSV17]

Outcome independent of group given other factors

$$\left(a + \sum_i b_i \mathbf{y}_i\right) \perp \mathbf{s},$$
$$\mathbf{y}_i = \Pr(\mathbf{y} = 1 | E_i)$$

[CHKV19]

Linear combination of conditional outcomes independent of group

$$\mathbf{y} \perp \mathbf{c} \mid \{\mathbf{e} = e\}$$

[HLGK19]

Outcome independent of circumstances, given efort

THE UNIVERSITY OF UTAH

A computational notion of fairness

Group: $g_M = \{x \mid M(x) = 1\}$

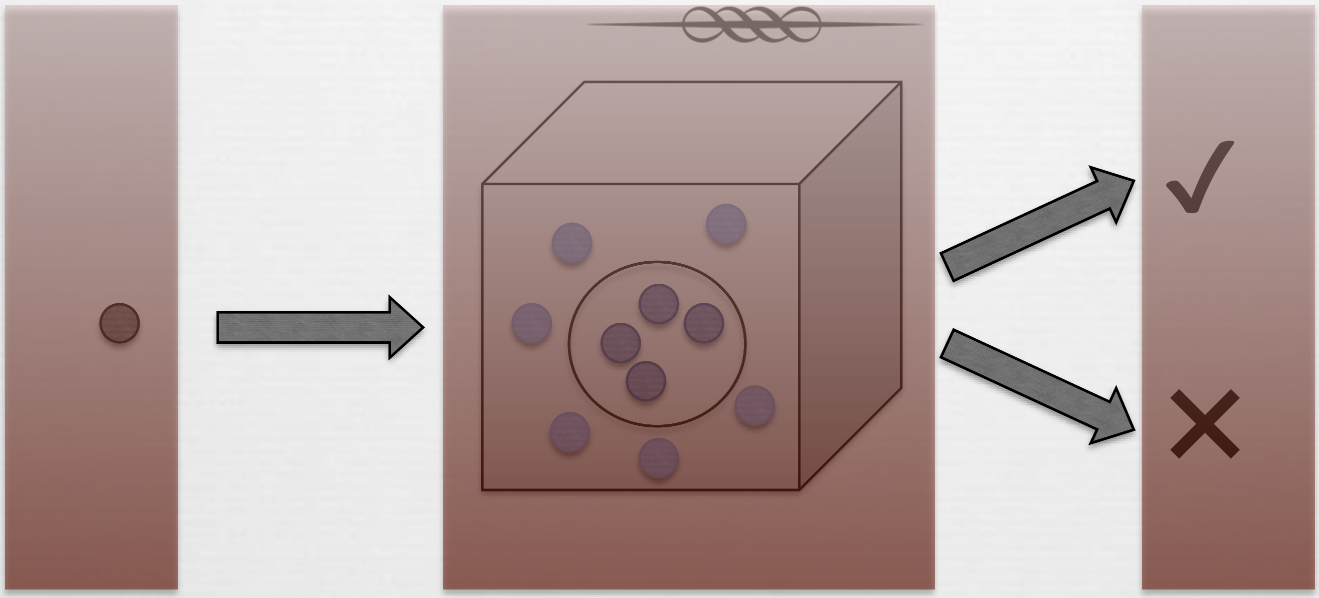
M : circuit of size s

Decision procedure is **fair** if it is fair for any group that can be defined with respect to a size-s circuit M. [**HKRR17, KNRW17**]

Connections to hardness of agnostic learning.

Fairness mechanisms

Make algorithmic decision-making fair

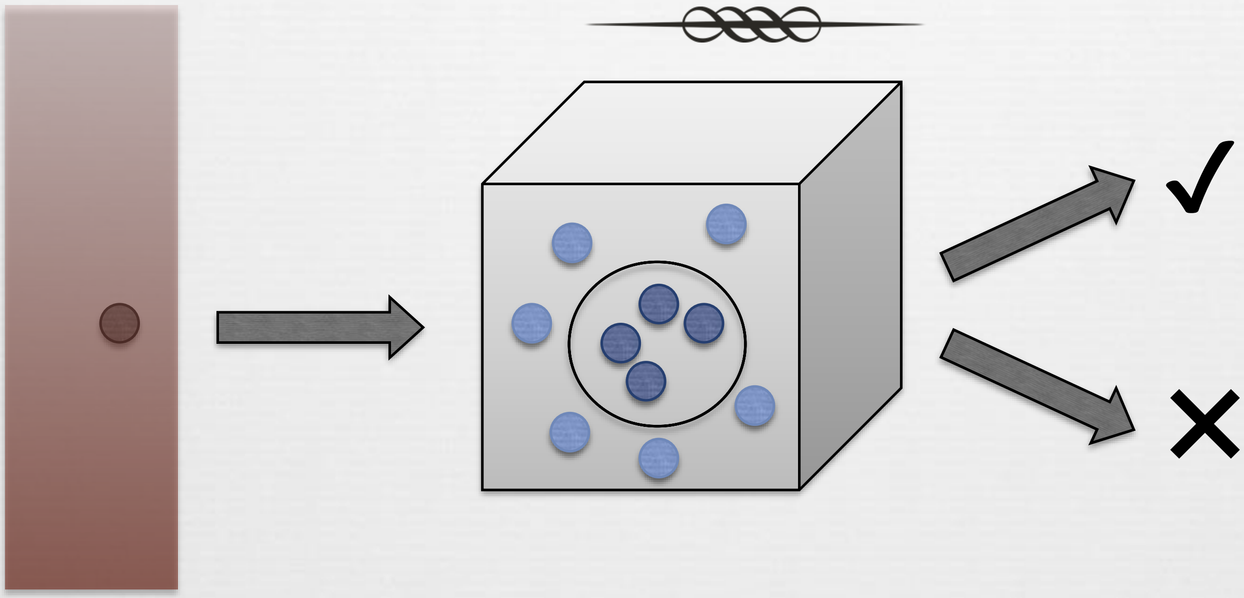


Modify the input

Modify the algorithm

Modify the output

Make algorithmic decision-making fair



Modify the input

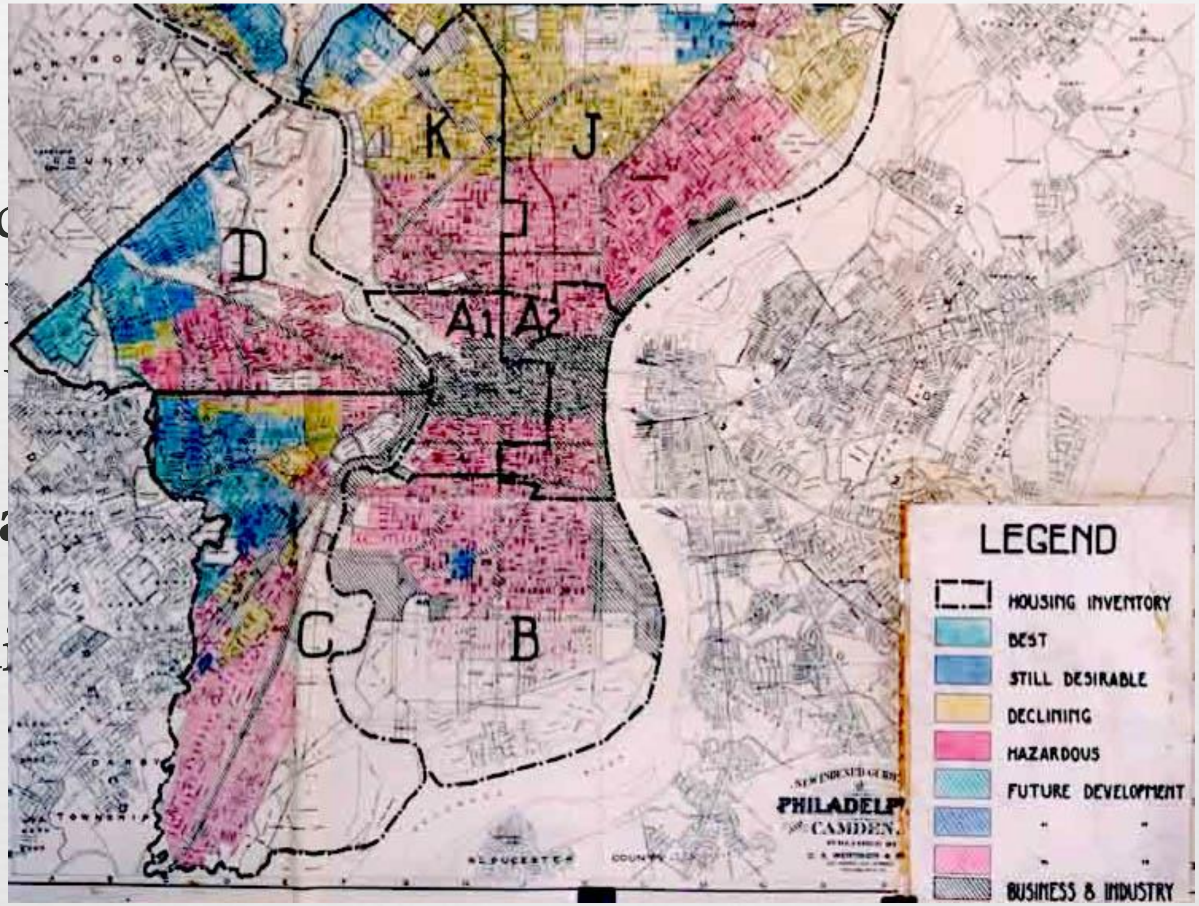
Direct and Indirect Bias

Source: Library of Congress (http://www.loc.gov/exhibits/civil-rights-act/segregation-era.html#obj24)

THE UNIVERSITY OF UTAH

Direct and Indirect Bias

- D: c
  -
  -
- **Goa**
- **Indi** with X)



By http://cml.upenn.edu/redlining/HOLC_1936.html,
Public Domain, https://commons.wikimedia.org/w/index.php?curid=34781276

Information content and indirect influence

the information content of a feature can be estimated by trying to predict it from the remaining features

Given variables X, Y that are correlated, find Y' conditionally **independent** of X such that Y' is as similar to X as possible.

Check information flow via computation

- Take data set D containing X
- Strip out X in some way, to get Y
- See if we can predict X' = X from Y with the best possible method.
- If error is high, then X and Y *have very little shared information.* **[FFMSV15]**

THE UNIVERSITY OF UTAH

Disparate Impact

$$\frac{\Pr(C = 1 \mid g = 0)}{\Pr(C = 1 \mid g = 1)} \geq 1 - \epsilon$$

$\epsilon < 0.2$

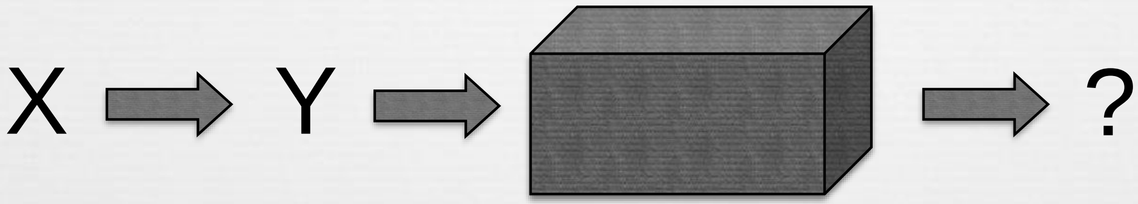
"4/5 rule":

There is a potential for disparate impact if the ratio of class-conditioned success probabilities is at most 4/5

Focus on **outcome,** rather than **intent.**

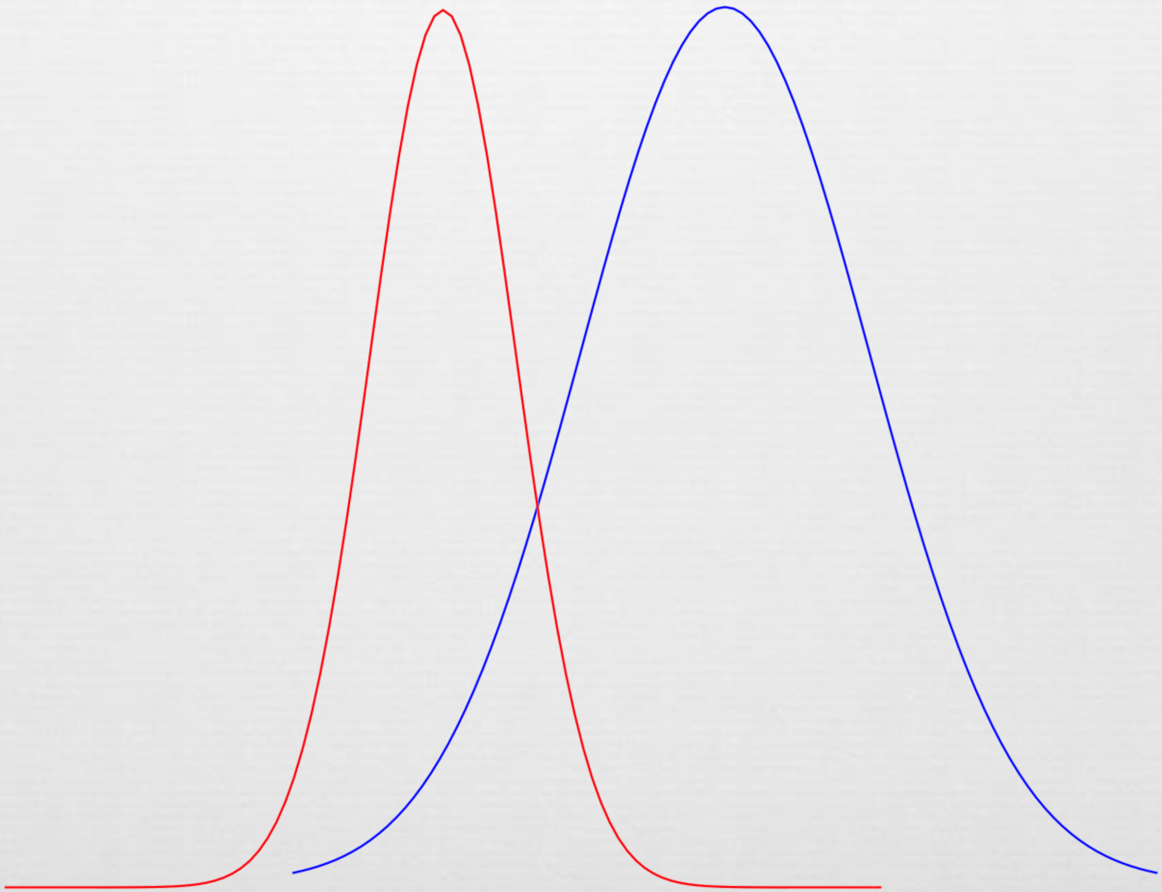
THE UNIVERSITY OF UTAH

Certification via prediction



Theorem: If we can predict X from Y with probability $\boldsymbol{\varepsilon}$**,** then our classifier has **potential disparate impact** with level g($\boldsymbol{\varepsilon}$).

Fixing data bias



Using the earthmover distance

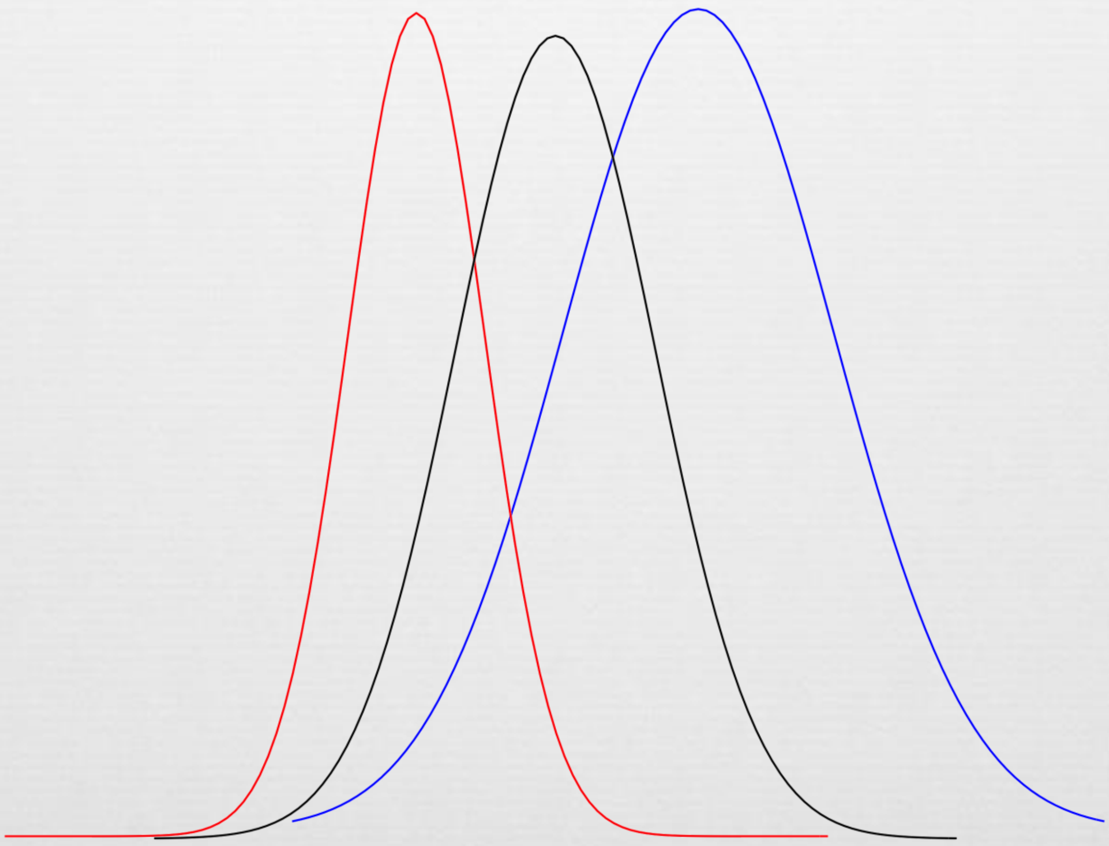
Let
$$P_i = \Pr(Y = y | X = i)$$
$$F_i = \text{cdf of } P_i$$

$$P_* = \arg\min \sum_i d_{EM}(P, P_i)$$

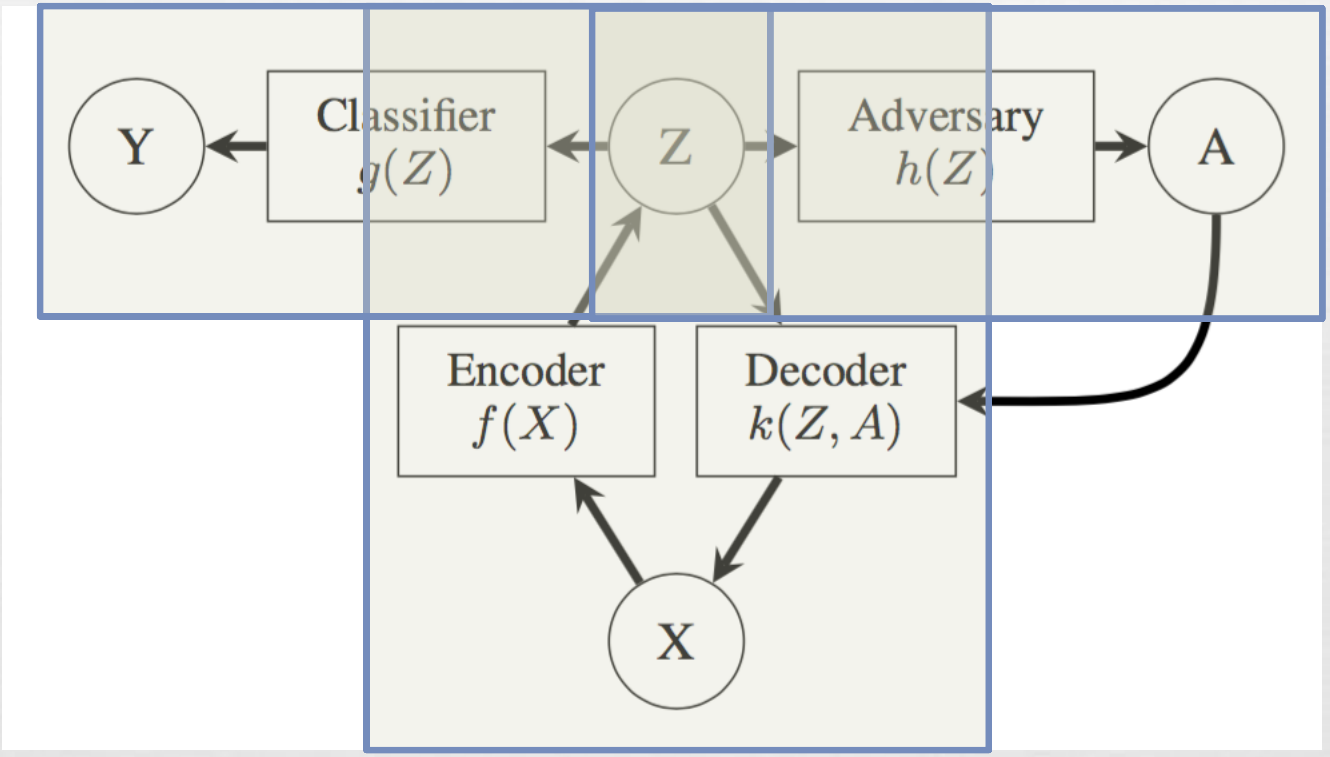
$$F_*^{-1}(\lambda) = \text{median}\, F_i^{-1}(\lambda)$$

We find a new distribution that is “close” to all conditional distributions.

Moving them together



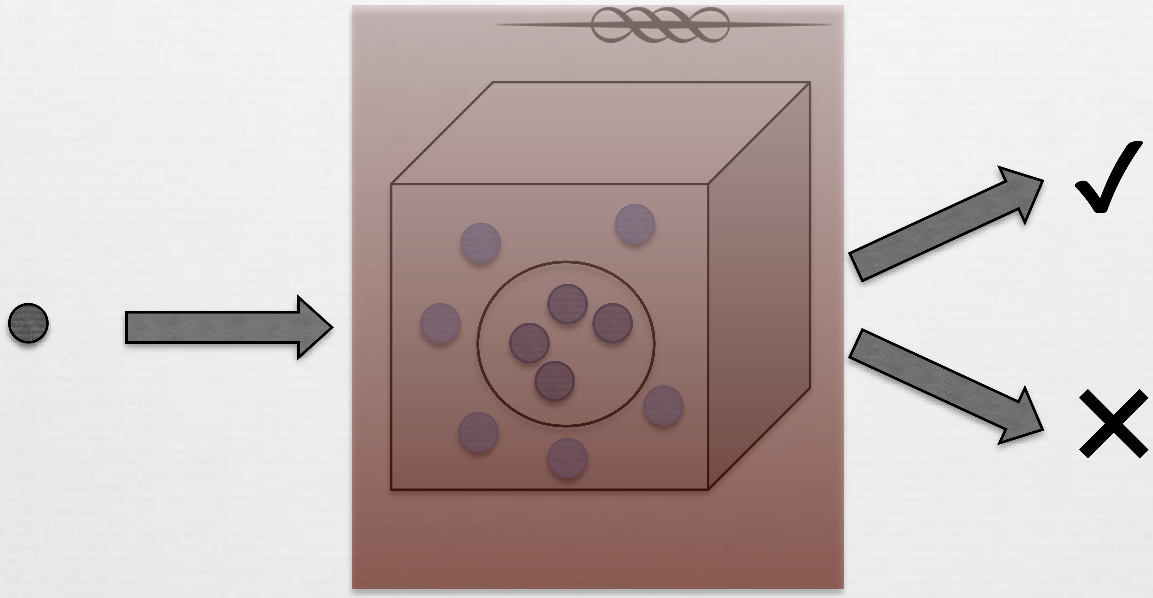
Learning fair representations



[ZWSPD13, ES16, MCPZ18]

THE UNIVERSITY OF UTAH

Make algorithmic decision-making fair



Modify the algorithm

Defining proxies for fairness

Classifier $\hat{y} = f(x;\theta)$

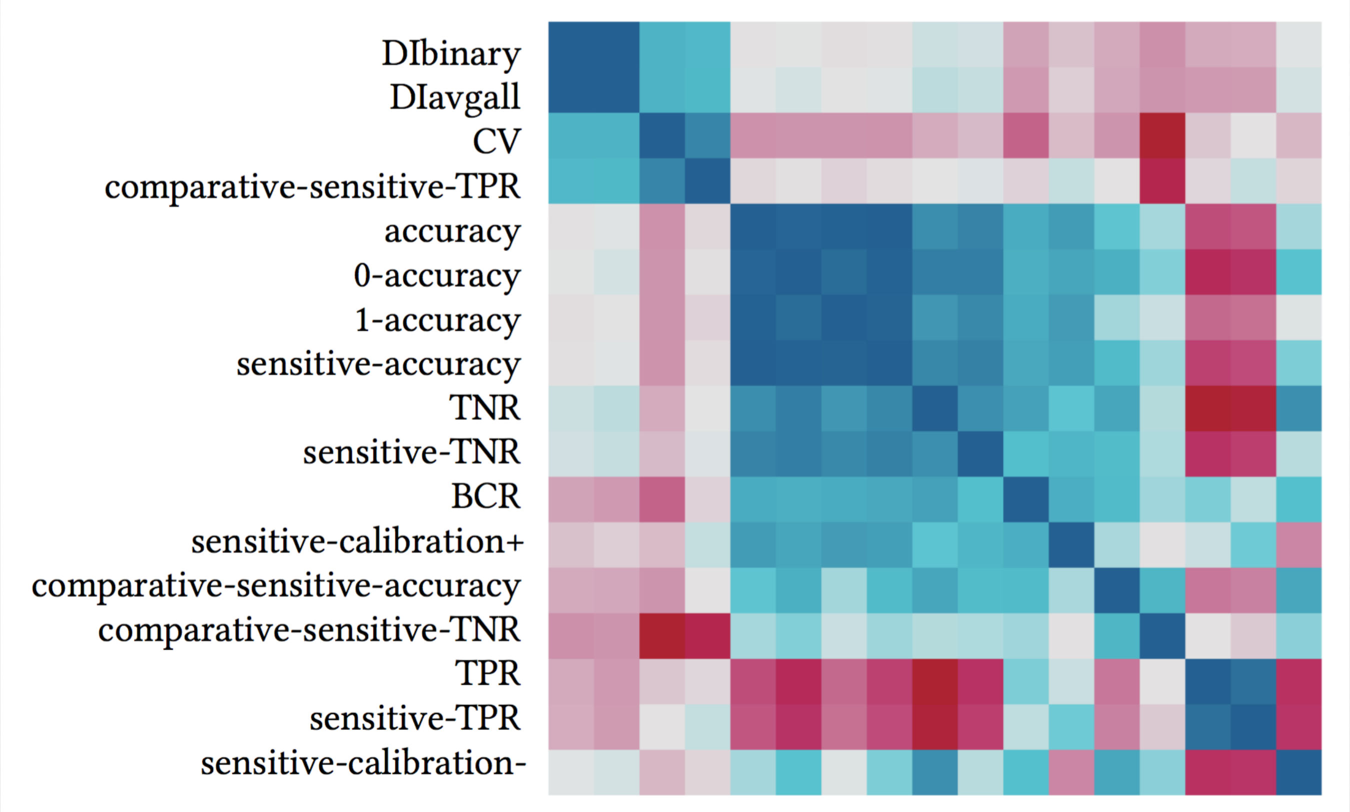
$$\begin{array}{ll} \min & L(\theta) \\ & |\Pr(\hat{y} \neq y \mid z = 1) - \Pr(\hat{y} \neq y \mid z = 0)| \leq \epsilon \end{array}$$

Goal [**ZVRG16**] : Eliminate correlation between sensitive attribute and (signed) *distance to decision boundary:*

$$\begin{aligned} \mathrm{Cov}(z, g_\theta(y,x)) &= \mathbb{E}[(z - \bar{z})(g_\theta(y,x) - \bar{g}_\theta(y,x))] \\ &\simeq \frac{1}{n}\sum(z - \bar{z})g_\theta(y,x) \\ \text{where} \quad g_\theta(y,x) &= \min(0, y d_\theta(x)) \end{aligned}$$

THE UNIVERSITY OF UTAH

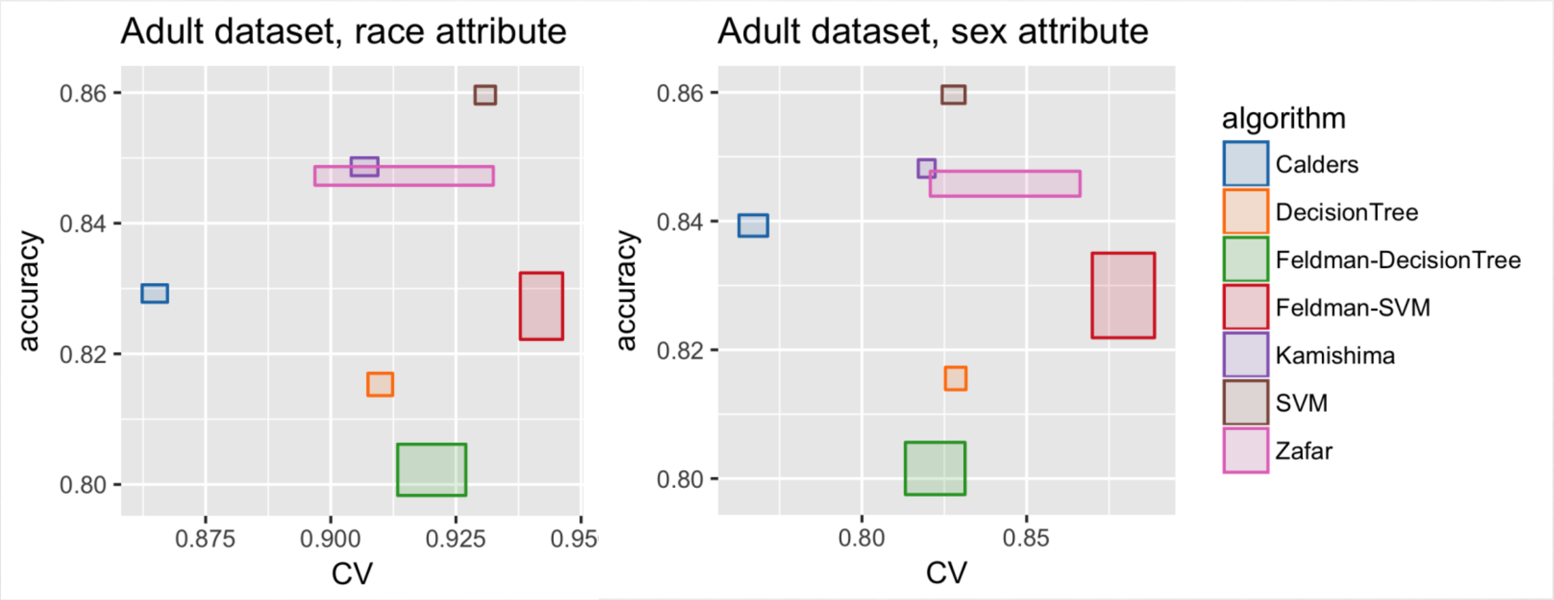
Comparing measures of fairness



[FSVCHR19]

THE UNIVERSITY OF UTAH

Comparing mechanisms for fairness



[FSVCHR19]

THE UNIVERSITY OF UTAH

But wait… there's more

- **Recourse [USL19]**
  - Measure the amount of *effort* it would take to move a point from a negative to positive classification
- **Counterfactual fairness [KLRS17, KRPHJS17]**
  - How would the algorithm have changed decisions if the sensitive attribute was flipped?

THE UNIVERSITY OF UTAH

Expanding Context: Audits

Research Question

- Given a black box function

$$Y = f(x_1, \ldots, x_n)$$

- Determine the *influence* each variable has on the outcome
  - How do we quantify influence
  - How do we model it (random perturbations?)
  - How do we handle *indirect* and *joint* influence

Landscape of work

$$y = f(\mathbf{x} = (x_1, \ldots, x_d))$$

- To what extent does a feature influence the model?
  - Determine whether model is using impermissible or odd features
- To what extent did the feature influence the outcome for **x? [RSG16, SSZ18]**
  - Generate an explanation for a decision, or a method of recourse (GDPR)

THE UNIVERSITY OF UTAH

Influence via perturbation [B01]

Key is the design of the intervention distributon

$$Y = f(x_1, \ldots, x_n$$

$$x_1' \sim B_\epsilon(x_1)$$

$$Y' = f(x_1', \ldots, x_n)$$

$$\mathrm{inf}_\epsilon(x_1) = |Y - Y'| = \Delta(Y)$$

[HPBAP14, DSZ16, LL18,…]

Information content and indirect influence

the information content of a feature can be estimated

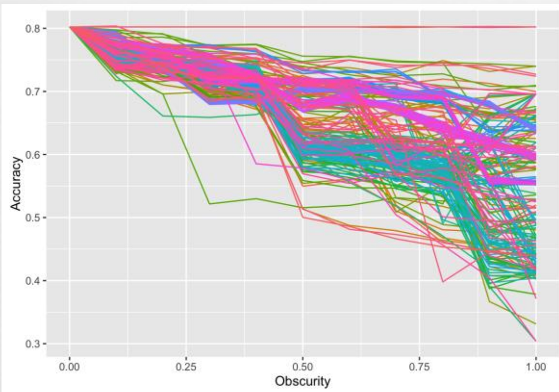
by trying to predict it from the remaining features
[AFFNRSSV16,17]

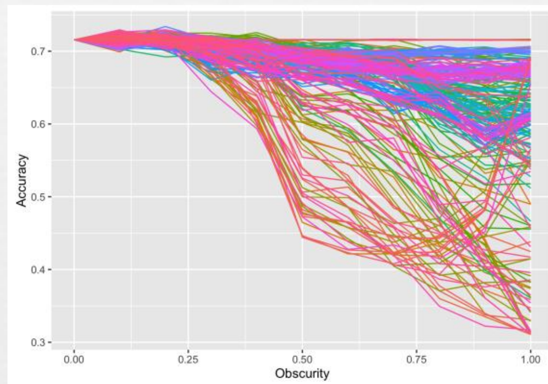
Given variables X, W that are correlated, find W' conditionally **independent** of X such that W' is as similar to W as possible.

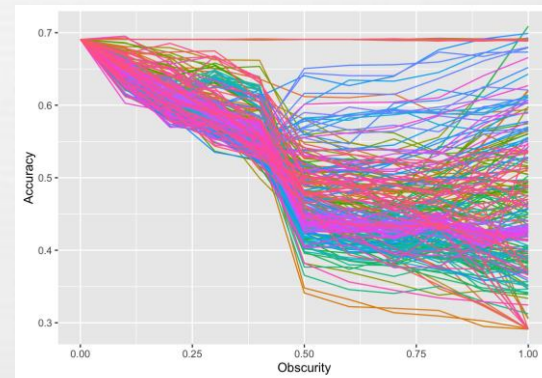
Influence(W) (without X) = $\Delta(Y)$

THE UNIVERSITY OF UTAH

Can we understand a model?



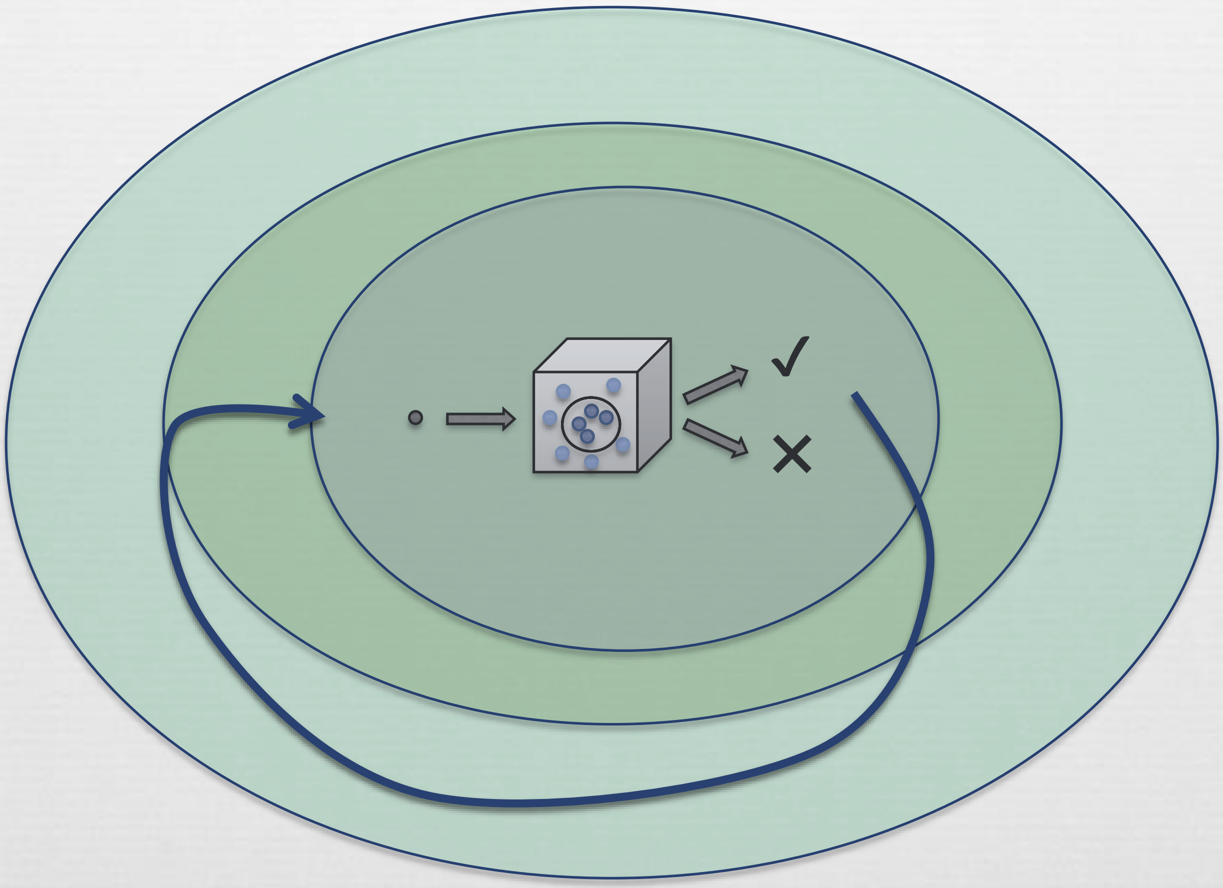




- Dark reactions project: predict presence/absence of a certain compound in a complex reaction.
- 273 distinct features.
- Approach identified key variables for further study that appear to influence the models.

THE UNIVERSITY OF UTAH

Feedback loops



THE UNIVERSITY OF UTAH

Predictive Policing

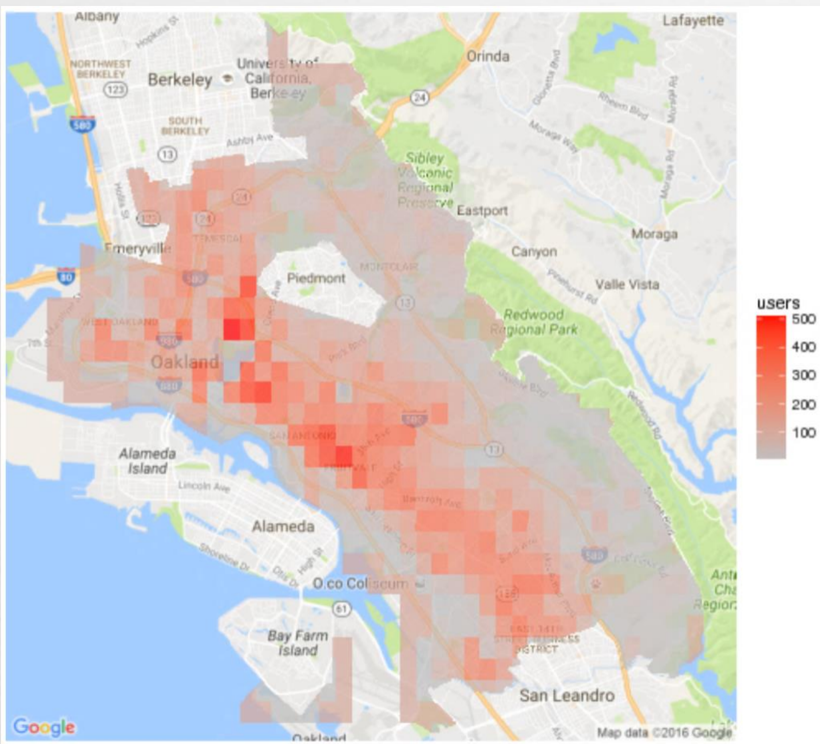
Given historical data about crime in different neighborho...
build a model to predict crime and use thi...

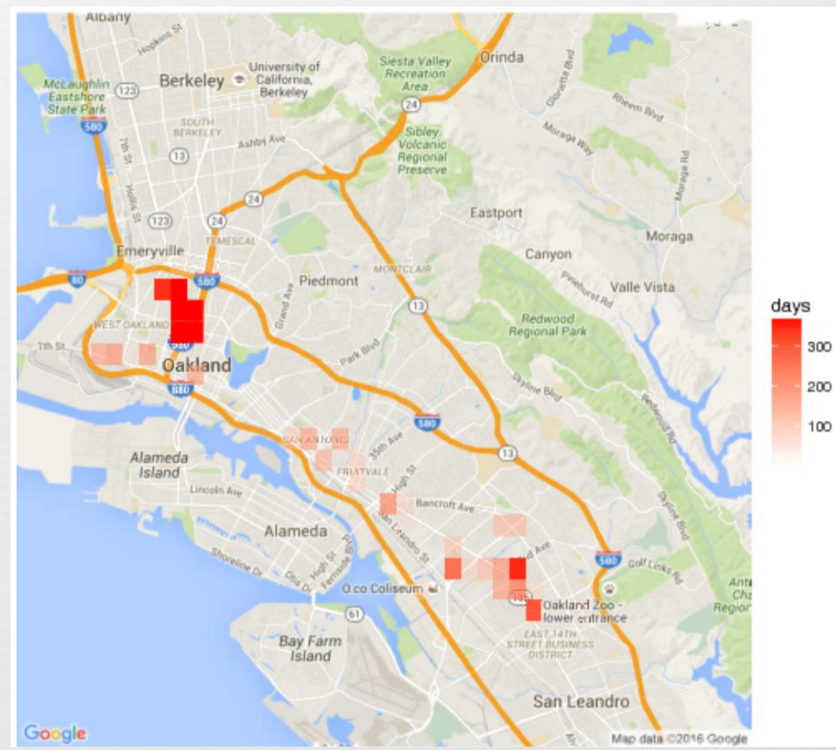
The LAPD Has a New Surveillance Formula, Powered by Palantir

HunchLab

Feedback Loops

To Predict and Serve, Lum and Isaac (2016)





THE UNIVERSITY OF UTAH

Building a model

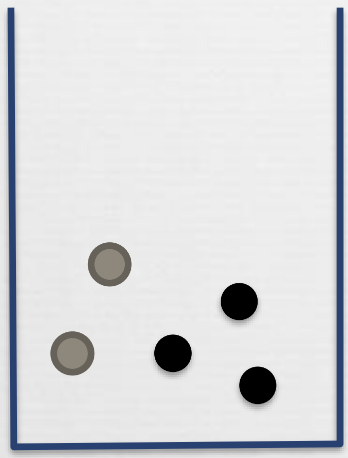
Assumptions.

1. Officer tosses coin based on current model to decide where to go next
2. Only information retained about crime is the count
3. If officers goes to area with baseline crime rate r, they will see crime with probability r.

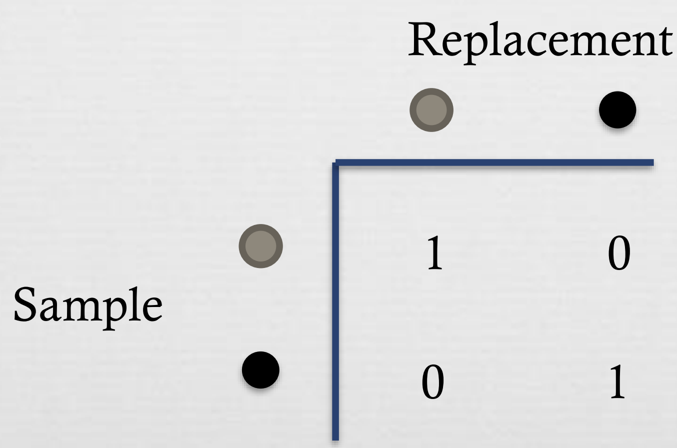
Goal:

A region with X% of crime should receive X% of policing.

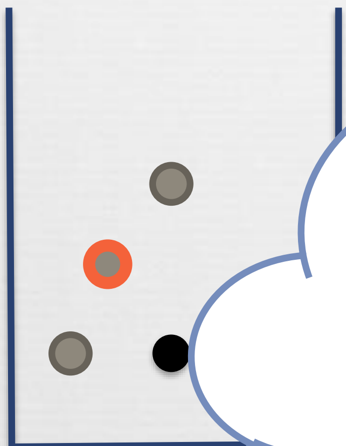
Urn Models



1. Sample a ball at random from the urn
2. Replace the ball and add/remove more balls depending on the color (replacement matrix)
3. Repeat



Urn Models



from the

ve

lor

What is the limiting fraction of ● in the urn?

Sample

0

0 1

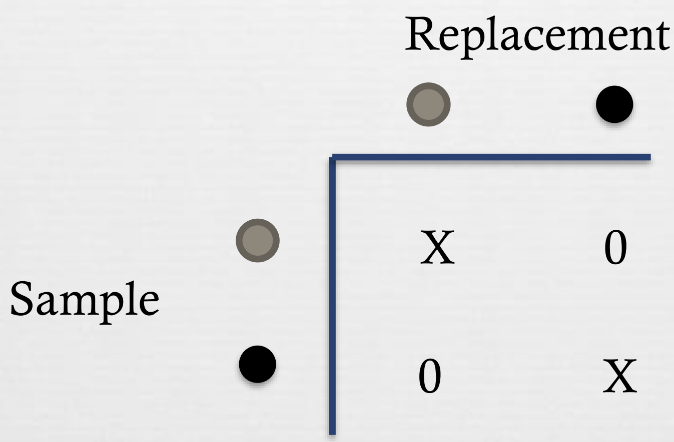
From policing to urns

- Assume we have two neighborhoods, and that each is one **color.**
- Visiting neighborhood = sampling ball of that color (Assumption 1)
- Observing crime = adding a new ball of that color.

THE UNIVERSITY OF UTAH

Urn 1: Uniform crime rates

- Assume both regions have the same crime rate r.



$$E[X] = r$$

*This is an urn **conditioned** on the events where a ball is inserted.*

Urn 1: Uniform crime rates

Theorem (folklore)

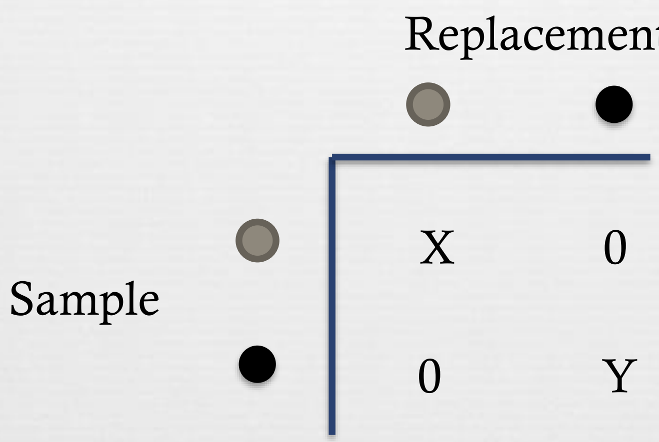
If the urn starts with A● and B●, then the limiting probability of ● is a random draw from the distribution Beta(A, B)

Implication

This is independent of the actual crime rate, and is only governed by initial conditions (i.e initial *belief*).

Urn 2: Different crime rates

- Regions have crime rates r_A and r_B



$$E[X] = r_A$$
$$E[Y] = r_B$$

*This is an urn **conditioned** on the events where a ball is inserted (proof in our paper).*

Urn 2: Different crime rates

- **Theorem (Renlund2010)**

Limiting probability of
is root of quadratic equation

		Replacement	
Sample		a	b
		c	d

$$(c+d-a-b)x^2+(a-2c-d)x+c=0$$

THE UNIVERSITY OF UTAH

Urn 2: Different crime rates

- **Theorem (Renlund2010)**

$$(c+d-a-b)x^2+(a-2c-d)x+c=0$$

- b = c = 0, a = r_A, d = r_B

Implication

If $r_A > r_B$, estimated probability of crime in A = 1.

THE UNIVERSITY OF UTAH

Blackbox Solution [EFNSV18]

- Using prior estimates to sample from urn creates biased estimator.
- Intuition: only update the model if the sample is "surprising".
  - If probability of ● is p, then only update model when seeing ● with probability 1-p = p(●).
  - Guarantees that model estimates are proportional to true probabilities
  - "rejection-sampling" variant of Horvitz-Thompson estimator.

Whitebox solution [EFNSV18b]

- Model problem as a reinforcement learning question
  - Specifically as a *partial monitoring problem*
- Yields no-regret algorithms for predictive policing
- Improvements and further strengthening by **[EJJKNRS19]**

Game Theoretic Feedback

- Can we design a decision process that cannot be gamed by users seeking an advantage **[HMPW16]**?
  - **[MMDH18]:** any attempt to be strategy-proof can cause an extra burden tp disadvantaged groups.
  - **[HIV18]:** if groups have different costs for improving themselves, strategic classification can hurt weaker groups and subsidies can hurt both groups.

THE UNIVERSITY OF UTAH

But wait… there's more

- Suppose the decision-making process is a sequence of decisions
  - Admission to college ➔Getting a job ➔Getting promoted
- Do fairness interventions "compose"?
  - NO! **[BKNSVV17, ID18]**
  - Can we make intermediate interventions so as to achieve end-to-end fairness? **[HC17, KRZ18]**

THE UNIVERSITY OF UTAH

Expanding context: Society

History of (un)fairness [HM19]

- Notions of fairness first studied in context of standardized testing and race-based discrimination (early 60s)
- Virtually **all** modern discussions of fairness and unfairness mirrors this earlier literature.
- Recommendations: focus more on *unfairness* rather than fairness, and how to reduce it.

THE UNIVERSITY OF UTAH

How do people accept algorithmic decision-making?

- What did judges do when risk assessment tools for pretrial hearings were rolled out? [**Stevenson18**]
  - Changes in bail
  - Little to no change in pretrial release
  - Reversion to pre-RAT behavior over time.
- How are people likely to behave when given algorithmic "guidance"? [**Green-Chen 19**]
  - Exhibit biased behavior even with guidance
  - Underperform algorithm.

Two computer scientists, two sociologists and a lawyer walk into a bar…

Two computer scientists, two sociologists and a lawyer walk into a ~~bar~~…

conference room

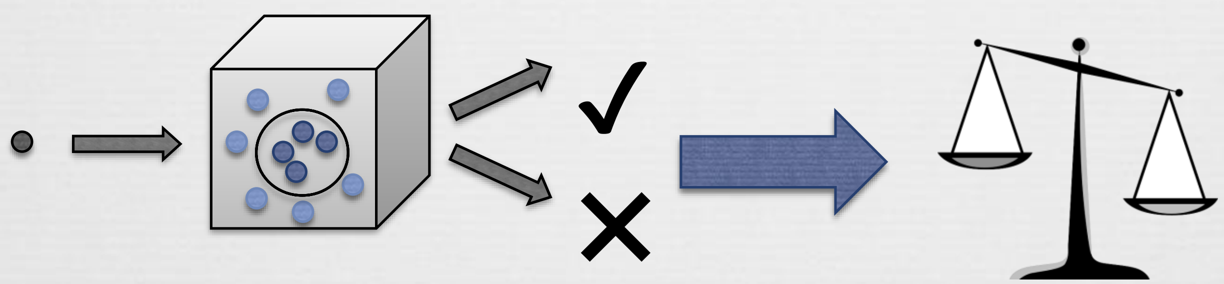
Fairness And Abstraction in Sociotechnical Systems, FAT 2019.*
Selbst, boyd, Friedler, V. and Vertesi.

The problem with abstraction

- CS modeling falls into *traps* when modeling sociotechnical systems
- Traps are rooted in the desire for abstraction.
- Proposed solutions are ineffective at best, and exacerbate the problems if worse.
- We need to identify these traps to avoid constantly falling into them.

1. The Framing Trap

Failure to model the entire system over which a social criterion, like fairness, will be enforced.

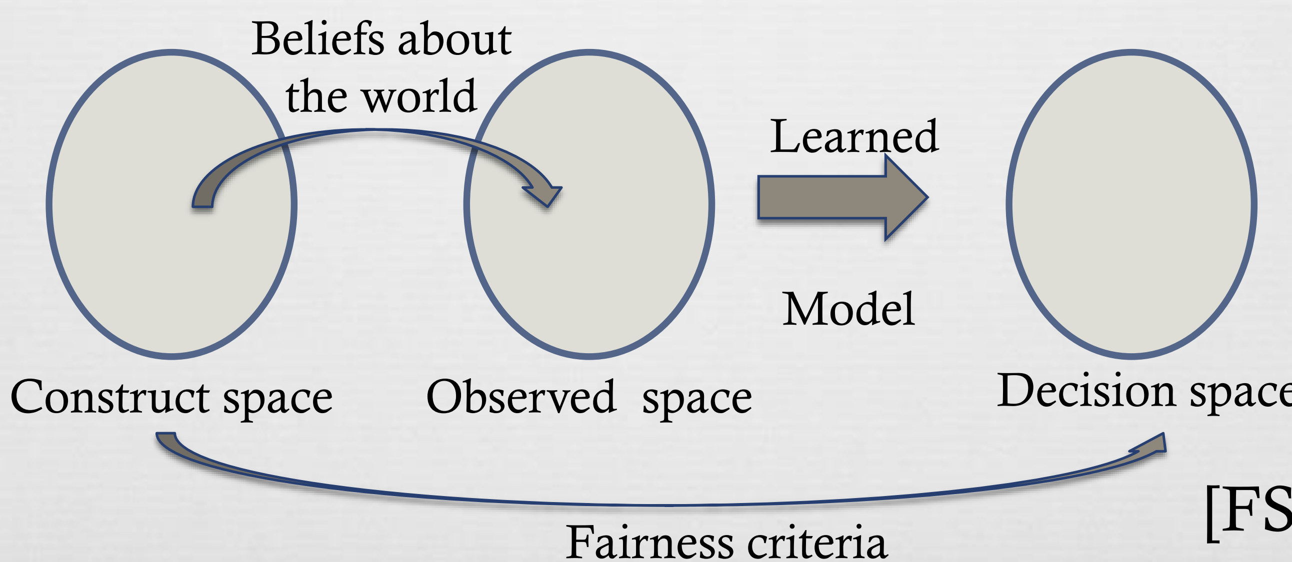


Fair risk assessment provides guarantees on disparate impact

Judge disregards recommendation when it doesn't align with "gut instinct"

2. The Modularity Trap

Failure to understand how repurposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context



[FSV16]

THE UNIVERSITY OF UTAH

3. The Formalism Trap

Failure to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms

Definitions of fairness are:

- Process-based rather than outcome-based
- Depend on the context in which they are being used.
- Contested depending on the stakeholders involved.

THE UNIVERSITY OF UTAH

4. The Ripple Effect Trap

Failure to understand how the insertion of technology into an existing social system changes the behaviors and embedded values of the pre-existing system

Why Amazon's Automated Hiring Tool Discriminated Against Women

PREDPOL®

THE UNIVERSITY OF UTAH

5. The Solutionism Trap

Failure to recognize the possibility that the best solution to a problem may not involve technology

Elon Musk's Chicago Tunnel: A Breakthrough or a Pipe Dream?

Musk proposes rescuing boys trapped in Thai cave with a 'submarine' made from SpaceX rocket part

Science and Technology Studies

- Recognize that we are dealing with **socio**technical systems
- Understand the social actors that interact with technology and shape it.
- Use studies of past adoption of technology to understand how new adoption might play out.

THE UNIVERSITY OF UTAH

Avoiding the traps

Framing Trap

Heterogeneous engineering or "human in the loop" design [GC19]

Modularity Trap

Model cards [MW+18]
Data sheets [GMV+18,BF19]
Nutrition labels [YSA+18, MIT Media Lab]

Formalism Trap

Interpretive flexibility.
Avoid rhetorical closure.

Ripple effect Trap

Model feedback loops [EFNS**V**+18a,EFNS**V**+18b,EJJ+18]
Strategic classification [HIV18,MM+18]

THE UNIVERSITY OF UTAH

The research

Defining **(un)fairness** and fairness-enhancing procedures

Understanding **interaction** between system and agents.

Understanding **influence** of inputs to black/gray-box procedures

Evaluating interventions in larger social context

Things I didn't touch on

Articulating **harms of representation** (GIGO)

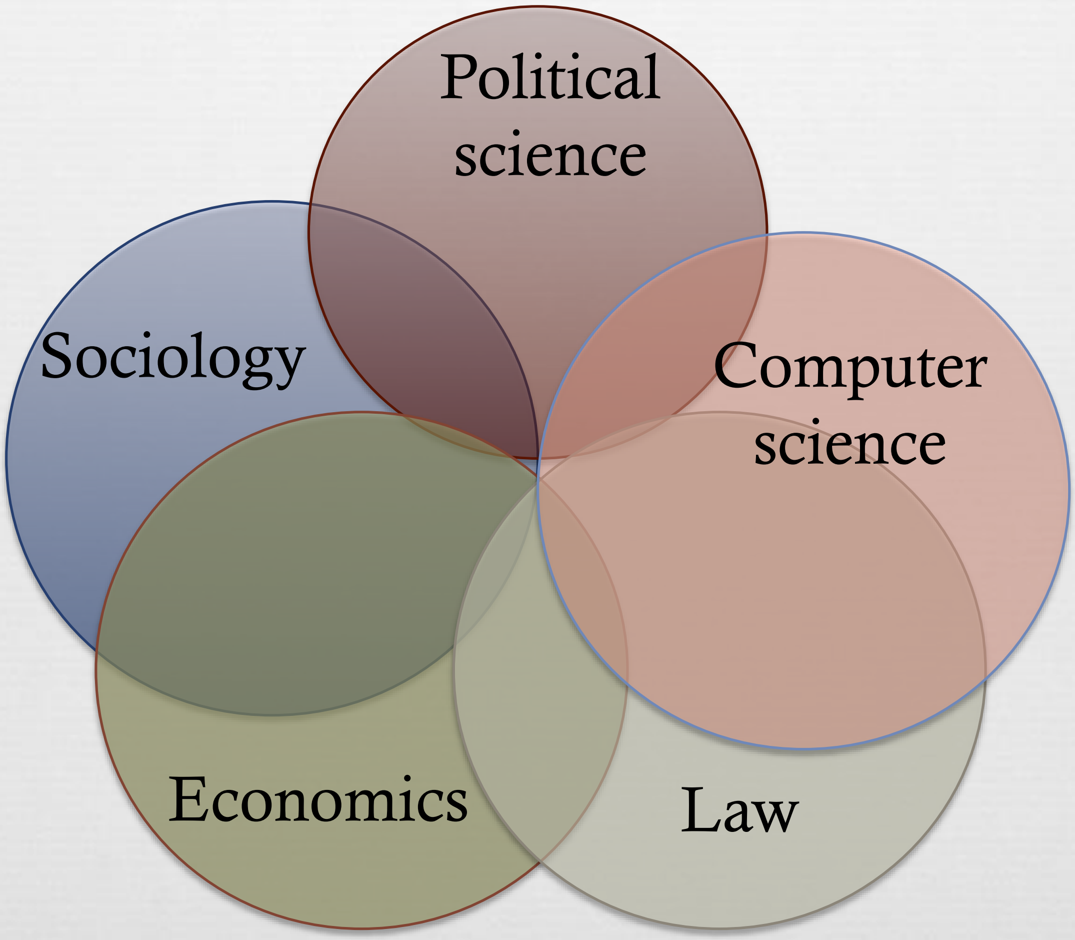
Tools to **interpret** and **explain** decisions (GDPR)

Interaction between **policy**, **technology** and the **law**.

Tensions between **privacy** and the desire for fairness**.**

THE UNIVERSITY OF UTAH

The questions



THE UNIVERSITY OF UTAH

danah boyd
Sorelle Friedler
Carlos Scheidegger
Andrew Selbst
Janet Vertesi

Thank you!

suresh@cs.utah.edu

Mohsen Abbasi
Sonam Choudhary
Danielle Ensign
Scott Neville