Making Decisions that Reduce Discriminatory Impact

Matt J. Kusner[1,2] Chris Russell[1,3] **Joshua R. Loftus**[4]
Ricardo Silva[1,5]

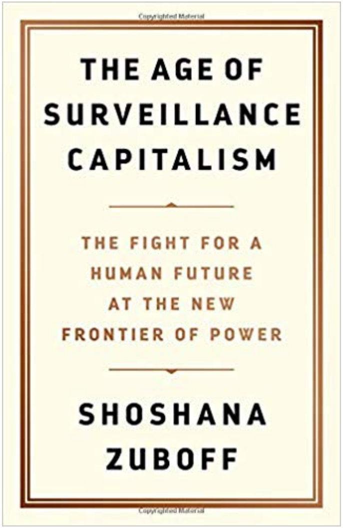
[1]Alan Turing Institute, [2]Oxford, [3]Surrey, [4]NYU, [5]UCL

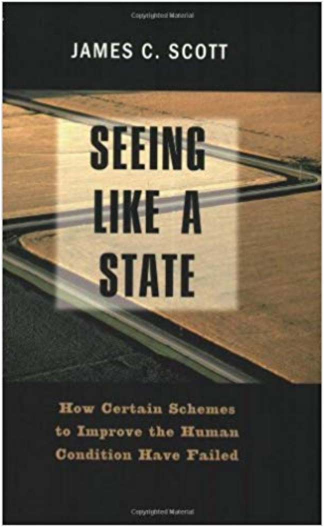
6/13/2019

Data-driven processes: not necessarily fair by default

Source: flir.com "SkyWatch"

Maybe closer to the opposite of fair by default...





This.paper()

- Propose to formalize the **impact problem**
- Design fair(er) interventions under **causal interference**

Defining impact

An **impact** is an event caused jointly by the decisions under our control and other real world factors. *Decisions about one individual can impact another individual.*
See also Liu et al. (ICML 2018), Green & Chen (FAT* 2019)

Fair predictions/decisions do not imply fair impacts, since other downstream factors can make the impact unfair (possibly to different individuals than the subjects of the original prediction/decision)

This.paper()

- Propose to formalize the **impact problem**
- Design fair(er) interventions under **causal interference**

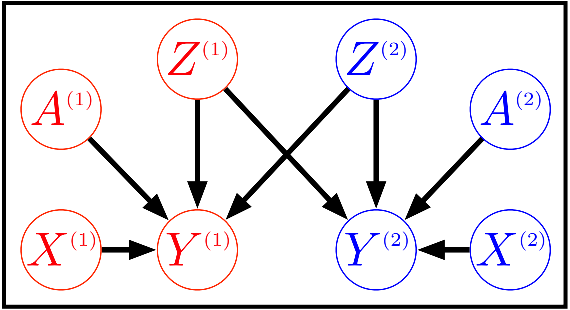
Defining impact

An **impact** is an event caused jointly by the decisions under our control and other real world factors. *Decisions about one individual can impact another individual*.

See also Liu et al. (ICML 2018), Green & Chen (FAT* 2019)

Fair predictions/decisions do not imply fair impacts, since other downstream factors can make the impact unfair (possibly to different individuals than the subjects of the original prediction/decision)

We use the structural causal model (SCM) framework



Z is the intervention or policy we want to optimize, **A** the protected attribute, **X** other predictors, and **Y** the outcome (higher values are desirable), superscript for observation index

School example

- Budget to pay for calculus classes in highschools (that do not already have them)
- Intervention: $\mathbf{Z}^{(i)} = 1$ if school i receives funding for a class and 0 otherwise
- Outcome: $\mathbf{Y}^{(i)}$ percent of students at school i taking the SAT (planning to go to college)
- Protected attribute: $\mathbf{A}^{(i)}$ encodes whether school i is majority black, Hispanic, or white
- Interference: students at school i *may be able to take a calculus class at nearby schools*

Given causal model and data, design the best **fair** intervention $\mathbf{Z}$

Fair? What does that mean?

Predictions or decisions should be the same in the actual world and in a counterfactual world where the value of the protected attribute had been different

- Changing a to a' also changes descendents of **A** in the SCM graph (*model-based counterfactuals*)
- **Counterfactual fairness** (Kusner et al, NeuRIPs 2017) is the property of invariance to those specific changes
- In this paper we instead bound **counterfactual privilege**

$$\mathbb{E}[\widehat{\mathbf{Y}}(a, \mathbf{Z})] - \mathbb{E}[\widehat{\mathbf{Y}}(a', \mathbf{Z})] < \tau$$

- In practice these *asymmetric constraints* will only be active for privileged values of a (actual, left term), and inactive otherwise

Fair? What does that mean?

Predictions or decisions should be the same in the actual world and in a counterfactual world where the value of the protected attribute had been different

- Changing a to a' also changes descendents of **A** in the SCM graph (*model-based counterfactuals*)
- **Counterfactual fairness** (Kusner et al, NeuRIPs 2017) is the property of invariance to those specific changes
- In this paper we instead bound **counterfactual privilege**

$$\mathbb{E}[\widehat{\mathbf{Y}}(a, \mathbf{Z})] - \mathbb{E}[\widehat{\mathbf{Y}}(a', \mathbf{Z})] < \tau$$

- In practice these *asymmetric constraints* will only be active for privileged values of a (actual, left term), and inactive otherwise

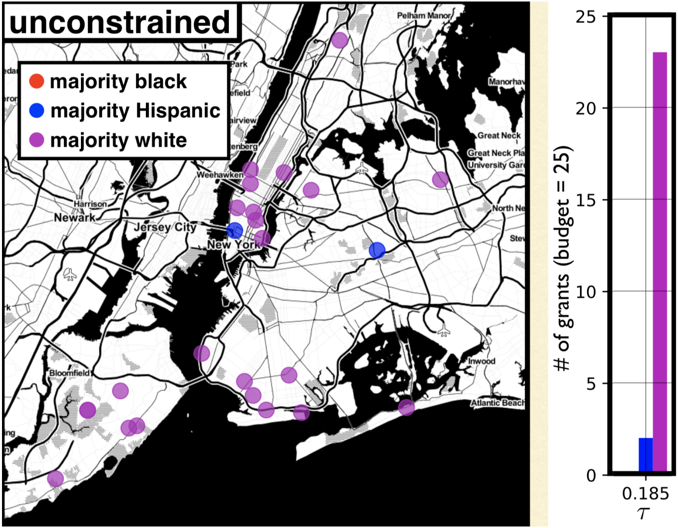
Optimal intervention under interference

- Our goal is to design optimal interventions or policies $\mathbf{Z}$ subject to a budget constraint, e.g.

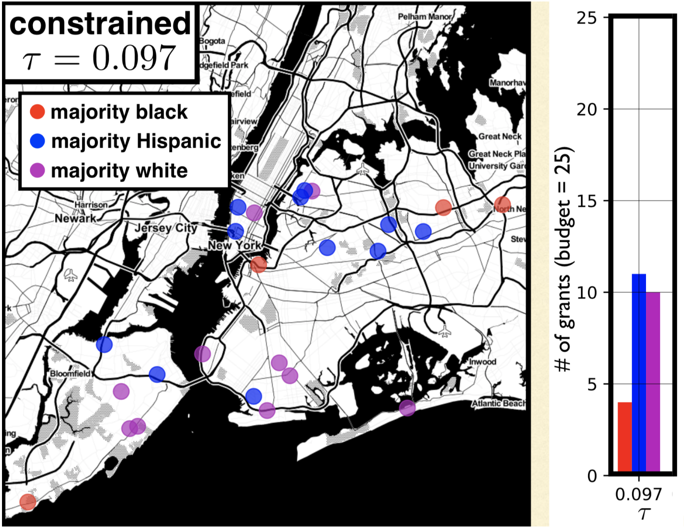
$$\mathbf{Z} = \arg\max \sum_i \mathbb{E}\left[\widehat{\mathbf{Y}}^{(i)}(a^{(i)}, \mathbf{Z}) | \mathbf{A}^{(i)}, \mathbf{X}^{(i)}\right] \quad s.t. \quad \sum_i \mathbf{Z}^{(i)} \leq b$$

- Interference means $\mathbf{Y}^{(i)}$ is potentially a function of all of $\mathbf{Z}$ and not just $\mathbf{Z}^{(i)}$
- Next two slides: optimal interventions with and without counterfactual privilege constraint

School resource allocation without fairness constraint



School resource allocation bounded counterfactual privilege



Thanks for your attention! See paper/poster(138) for more details

Matt Kusner

Chris Russell

Ricardo Silva