Fair Questions

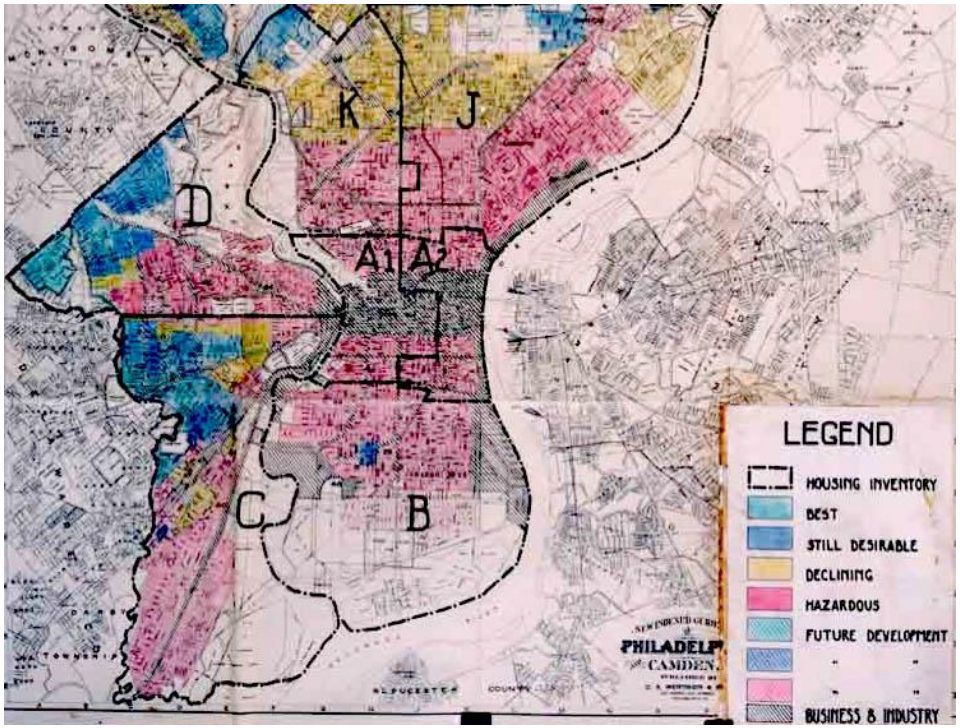
Cynthia Dwork, Harvard University & MSR

Outline

- Fairness in Classification: the one-shot case
  - Metrics
- The Sui Generis Semantics of Composition
  - Situational Awareness
- Beyond Classification
  - Nothing known
- The Data Don’t Tell
  - Recognizing failure
- Final Remarks

Adversary Goals

- "Catalog of Evils"
  - Redlining (exploiting redundant encodings), (reverse) tokenism, deliberately targeting "wrong" subset of S,…



Statistical Parity

- ☑ Demographics of selected group = demographics of population
  - Pr[x in S| outcome = o] = Pr[x in S]
  - Pr[x mapped to o | x in S] = Pr[x mapped to o | x in S^c]
  - Completely neutralizes redundant encodings

- ✘ Permits several evils in the catalog
  - E.g., intentionally targeting the subset of S unable to buy

Other Group Fairness Notions

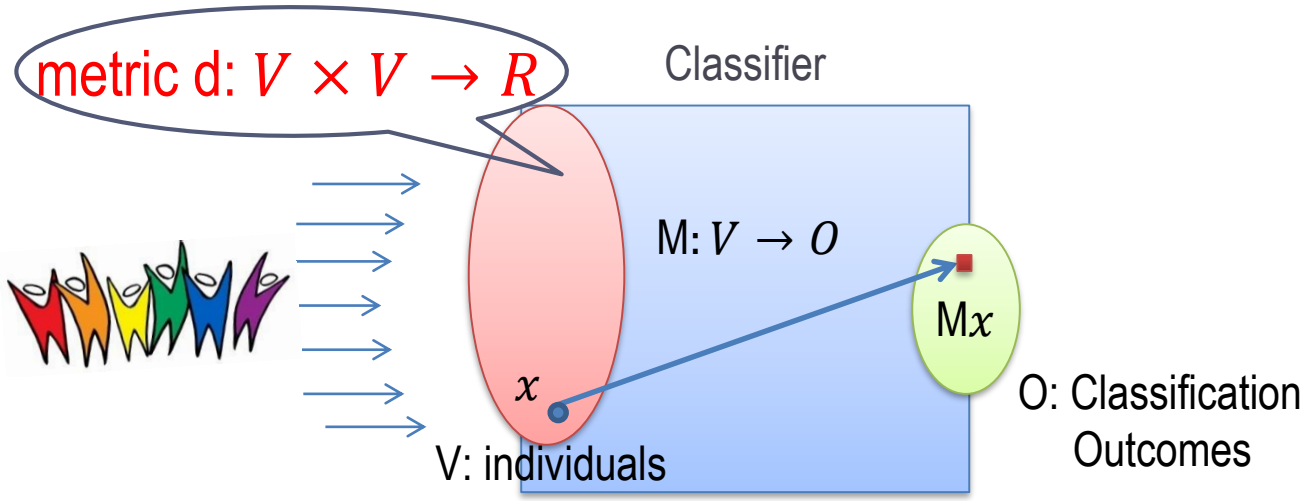
- Equal False Positive Rate (FPR) across groups
- Equal False Negative Rate (FNR) across groups
- Equal Positive Predictive Value (PPV) across groups
- Equal False Discovery Rate (FDR) across groups
- …
- No imperfect classifier can simultaneously ensure equal FPR, FNR, PPV unless the base rates are equal

$$\mathrm{FPR} = \left(\frac{p}{1-p}\right)\left(\frac{1-\mathrm{PPV}}{\mathrm{PPV}}\right)(1-\mathrm{FNR})$$

Chouldechova 2017; Kleinberg, Mullainathan, Raghavan 2017

Individual Fairness

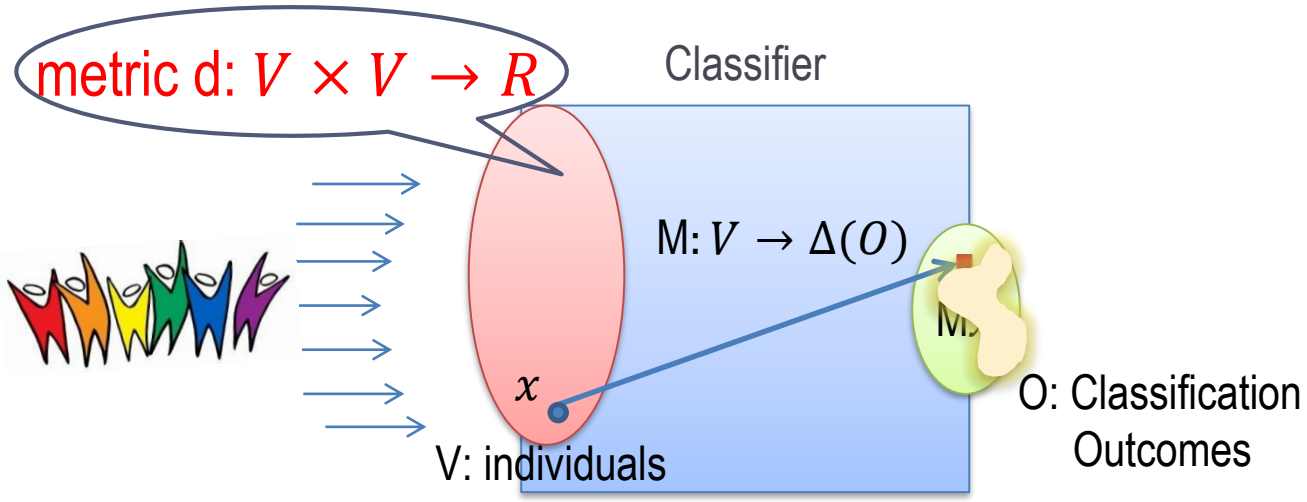
- People who are similar with *respect to a specific classification task* should be treated similarly
  - S + math $\sim$ S^c + finance
  - "Fairness Through Awareness"



Dwork, Hardt, Pitassi, Reingold, Zemel 2012

Individual Fairness

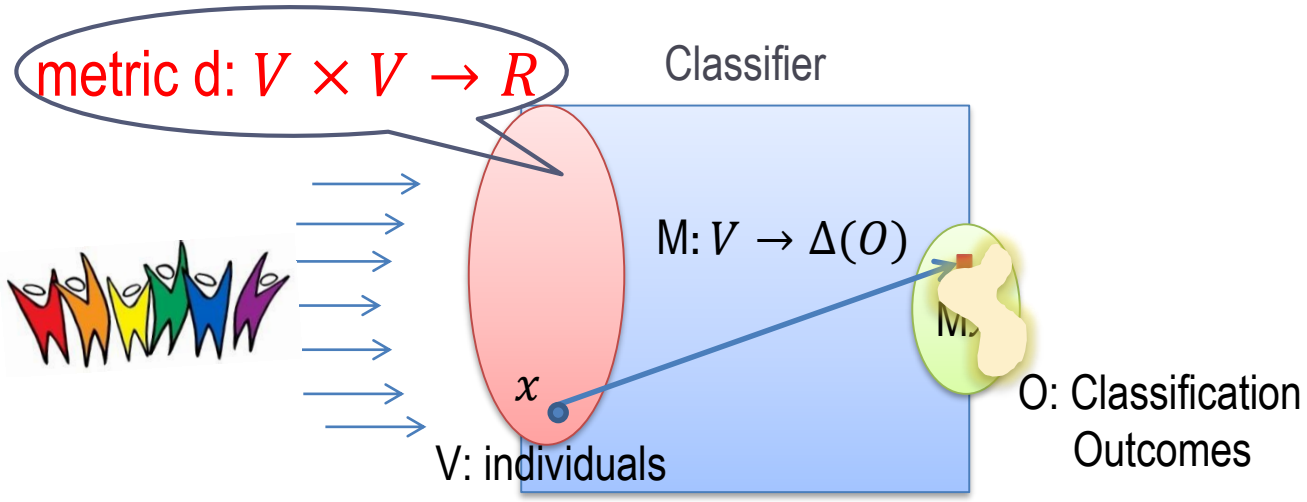
$$M: V \rightarrow \Delta(O)$$
$$\left|\left|M(u) - M(v)\right|\right| \leq d(u, v)$$



Dwork, Hardt, Pitassi, Reingold, Zemel 2012

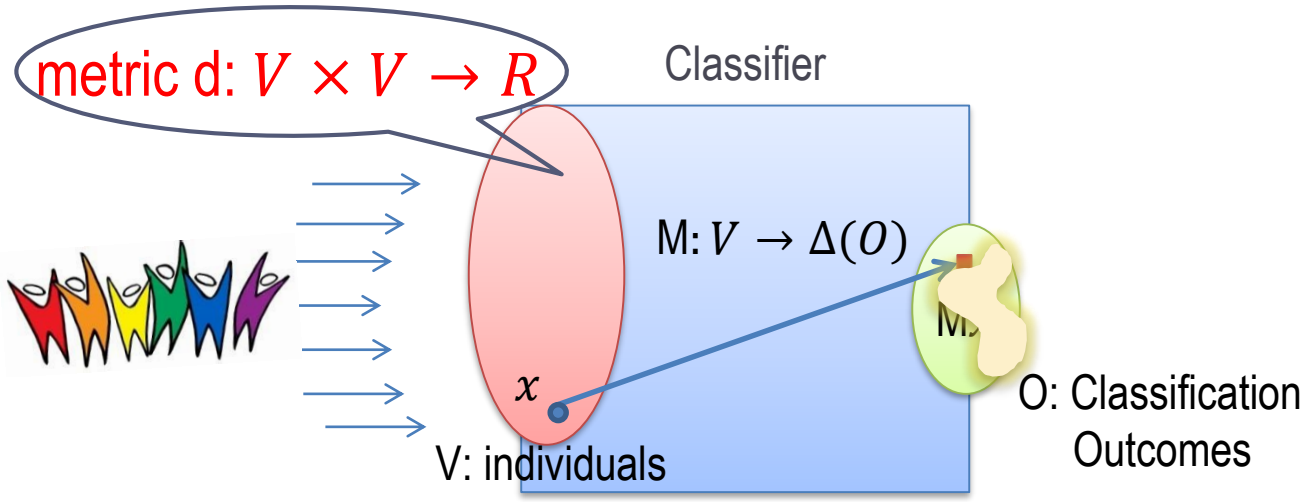
Individual Fairness

- Science Fiction: task-specific similarity metric
  - Ideally, ground truth
  - In reality, no better than society's "best approximation"



Individual Fairness

- Science Fiction: task-specific similarity metric
  - Ideally, ground truth
  - In reality, no better than society's "best approximation"
- How can we use AI to learn the (conjecture: unavoidable) metric?



Individual Fairness: Composition

- Composition subtle, *sui generis* semantics
  - Unlike in differential privacy, cryptography
  - Eg: Fair classifiers for ads "competing" for a slot on a web page

- Troubling Scenario
  - Consider phenomenon observed by Datta, Datta, and Tchantz

  - Maybe:
    - Job-related advertiser: pay same modest amount for M, W
    - Appliance advertiser: pay very little for M, a lot for W
  - What would the ad network do?

Individual Fairness: Composition

- Theorem: For any tasks T, T' with not identical non-trivial metrics d, d' on universe U, ∃ individually fair classifiers C, C' that *when naively composed* violate multiple-task fairness: $\exists u, v \in U$ s.t. at least one of:

$$|\Pr[S(u)_T = 1] - \Pr[S(v)_T = 1]| > d(u, v)$$
$$|\Pr[S(u)_{T'} = 1] - \Pr[S(v)_{T'} = 1]| > d'(u, v)$$

Dwork and Ilvento, 2017

Individual Fairness: Composition

- Theorem: For any tasks T, T' with not identical non-trivial metrics d, d' on universe U, ∃ individually fair classifiers C, C' that *when naively composed* violate multiple-task fairness.

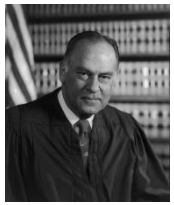
- How can AI develop situational awareness for fair composition?

Dwork and Ilvento, 2017

Beyond Classification

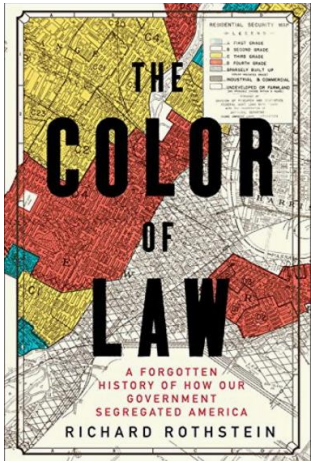
- I am represented by an AI
  - Eg: In my online negotiations
- Source of great inequity
  - Replace “AI” with “lawyer”
  - Exaggerated in online setting?
  - Should agents give each other some slack?
- Completely Open
  - Basic definitions, notions of composition

The Myth of *de facto* Segregation



Justice Potter Stewart, 1974: “The Constitution simply does not allow federal courts to attempt to change that situation unless and until it is shown that the State, or its political subdivisions, have contributed to cause the situation to exist.”

Chief Justice John Roberts, 2007: racially separate neighborhoods might result from “societal discrimination” but remedying discrimination “not traceable to [government’s] own actions” can never justify a constitutionally acceptable, racially conscious, remedy.



Richard Rothstein

Does Your Training Set Know History?

- Very complete data on the status quo may not reveal causality.

- How can AI recognize failure / need for scholarship?

Doaa Abu-Eloyunas, Frances Ding, Christina Ilvento,
Toni Pitassi, Guy Rothblum, Yo Shavit, Pragya Sur,
Saranya Vijayakumar, Greg Yang

NIPS, December 7, 2017

Individual Fairness: Composition

- Composition subtle, *sui generis* semantics
  - Unlike in differential privacy, cryptography
  - Eg: Fair classifiers for ads for job coaching service and appliances "competing" for a slot on a newspaper web page

- Theorem: For any tasks T, T' with not identical non-trivial metrics D, D' on universe U, ∃ individually fair classifiers C, C' that *when naively composed* violate multiple-task fairness: $\exists u, v \in U$ *s.t.*

$$|\Pr[S(u)_T = 1] - \Pr[S(v)_T = 1| \leq D(u, v)$$
$$|\Pr[S(u)_{T'} = 1] - \Pr[S(v)_{T'} = 1]| > D'(u, v)$$

Dwork and Ilvento, 2017

Individual Fairness: Composition

- Special Case: $\forall w \in U$: T is preferred to T'.
  - *$\forall w$: if w is positively classified by both C and C', it gets the ad T*
- Proof: Fix some u, v such that $D(u, v) \neq 0$

  $\Pr[S(u)_{T'} = 1] = (1 - p_u)p'_u$; $\Pr[S(v)_{T'} = 1] = (1 - p_v)p'_v$

  Difference = $[p'_u - p'_v] + p_v p'_v - p_u p'_u$

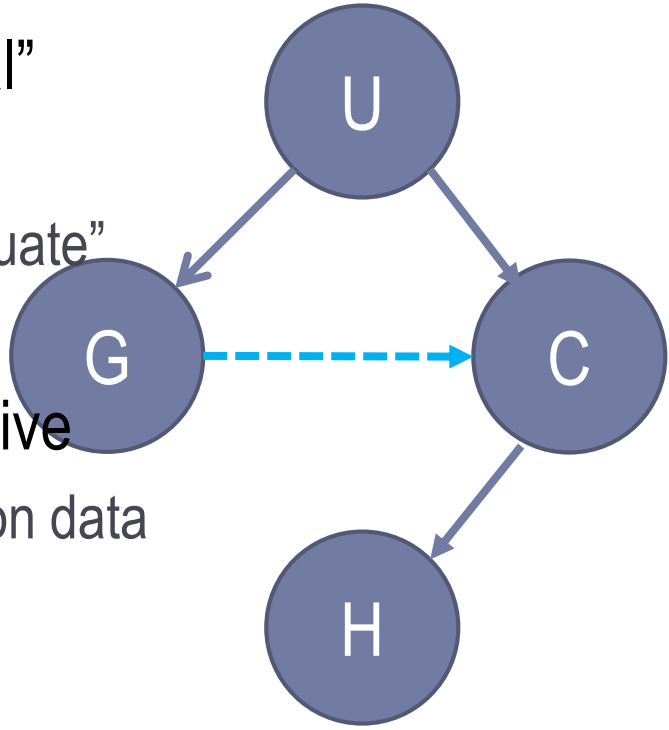
  If $D'(u, v) = 0$ then by Lipschitz $p'_u = p'_v$.
  - C' : $p'_u \neq 0$; C: $p_u - p_v \neq 0$

  If $D'(u, v) \neq 0$
  - C' : $p'_u - p'_v = D'(u, v)$; C : $p_u < p_v$
  - Constrained only by $p_v - p_u \leq D(u, v)$, can easily force $p_v/p_u > p'_u/p'_v$
  - $\Rightarrow p_v p'_v > p_u p'_u$

Dwork and Ilvento, 2017

Causal Inference

- Counterfactuals and Path-Specific Effects
  - Pearl, 2001; Avin, Shpitser, Pearl, 2005, Rubin, 1974, Nabi and Shpitser, 2017; Kusner et al., 2017; Kilbertus et al, 2017
- Aim to capture "everything else being equal"
  - Realizing that this may make no sense
  - No man has qualification "Smith College graduate"
- Unlike (often) prediction, very model-sensitive
  - Different models may yield same distribution on data
  - Fairness definition depends on model. *Brittle.*



Dwork, Ilvento, Rothblum, Sur 2017

Future Directions

- Machine learning of the metric
- Modify the various ML solutions to incorporate individual fairness
  - When does it happen automatically? Eg, points close in latent space decode to similar instances
- Explore the roles for partial solutions
  - Don’t need to solve the trolley problem; can simulate humans in extreme situations, dominating human driving

Doaa Abu-Eloyunas, Frances Ding, Christina Ilvento, Toni Pitassi, Guy Rothblum, Yo Shavit, Pragya Sur, Saranya Vijayakumar, Greg Yang

CAEC, December 1, 2017