Microsoft Research

Each year Microsoft Research hosts hundreds of influential speakers from around the world including leading scientists, renowned experts in technology, book authors, and leading academics, and makes videos of these lectures freely available.

2016 © Microsoft Corporation. All rights reserved.
The Emerging Theory of Algorithmic Fairness

Cynthia Dwork
Harvard University
Radcliffe Institute for Advanced Study
Microsoft Research

All the good slides: Christina Ilvento
Algorithmic Fairness

Population is diverse: ethnic, religious, geographic, medical, gender, class, sexual preference, etc.

- Bank receives detailed user information before serving a page
- Concern: steering minorities to credit card offerings of less desirable terms
Algorithmic Fairness

Population is diverse: ethnic, religious, geographic, medical, gender, class, sexual preference, etc.

- “Hide” the sensitive information?
  - Redlining, birds-of-a-feather
Algorithmic Fairness

- Population is diverse: ethnic, religious, geographic, medical, gender, class, sexual preference, etc.

- “Hide” the sensitive information?
  - Redlining, birds-of-a-feather
  - “Culturally aware” algorithms are more accurate
Algorithmic Fairness

- Population is diverse: ethnic, religious, geographic, medical, gender, class, sexual preference, etc.

- "Hide" the sensitive information?
  - Redlining, birds-of-a-feather
  - "Culturally aware" algorithms are more accurate

Sage

Thyme
Algorithmic Fairness

- Population is diverse: ethnic, religious, geographic, medical, gender, class, sexual preference, etc.

- “Train” on historical data?
  - Imbibe biases in the data

- No general source of ground truth
Classification Algorithms

- Binary
Classification Algorithms

- Ternary
Classification Algorithms

- Dependent
Classification Algorithms

- Risk Assessment Scoring
Classification Algorithms

- Risk Assessment Scoring
  - Translates score to decision
A Rigorous Approach to Algorithmic Fairness

- Basic Paradigm
  - Define fairness by articulating the goals of the adversary
  - Create fair algorithmic building blocks
  - Prove a “composition theorem” showing that fair pieces combine to give (possibly slightly less) fair systems.
Defining Fairness
Adversary Goals

- "Catalog of Evils"

  - Redlining (exploiting redundant encodings), (reverse) tokenism, deliberately targeting "wrong" subset of protected group $S$, ...

Dwork, Hardt, Pitassi, Reingold, Zemel 2012
Defining Fairness for Groups

- Group fairness properties are statistical requirements
  - Statistical Parity: demographics of people with positive (negative) classification are the same as the demographics of the general population
Defining Fairness for Groups

- Group fairness properties are statistical requirements
  - Statistical Parity: demographics of people with positive (negative) classification are the same as the demographics of the general population
  - Meaningful in the breach
Defining Fairness for Groups

- Group fairness properties are statistical requirements
  - Statistical Parity: demographics of people with positive (negative) classification are the same as the demographics of the general population
  - Permits targeting the wrong subset of S
Defining Fairness for Groups

- Group fairness properties are statistical requirements
  - Statistical Parity: demographics of people with positive (negative) classification are the same as the demographics of the general population
  - Fairness gerrymandering (sage eating coffee drinkers)
Defining Fairness for Groups

- Group fairness properties are statistical requirements
  - Statistical Parity: demographics of people with positive (negative) classification are the same as the demographics of the general population
  - Sets of group fairness properties may be incompatible
Unequal Error Rates as Group Unfairness

- Equal False Positive Rate (FPR) across groups
- Equal False Negative Rate (FNR) across groups
- Equal Positive Predictive Value (PPV) across groups

No imperfect classifier can simultaneously ensure equal FPR, FNR, PPV unless the base rates are equal

\[
FPR = \left(\frac{p}{1 - p}\right) \left(\frac{1 - PPV}{PPV}\right) (1 - FNR)
\]

\[ p = 5/13 \]

Chouldechova 2016; Kleinberg, Mullainathan, Raghavan 2016
Unequal Error Rates as Group Unfairness

- Equal False Positive Rate (FPR) across groups
- Equal False Negative Rate (FNR) across groups
- Equal Positive Predictive Value (PPV) across groups

No *imperfect* classifier can simultaneously ensure equal FPR, FNR, PPV unless the base rates are equal

\[
FPR = \left( \frac{p}{1 - p} \right) \left( \frac{1 - PPV}{PPV} \right) (1 - FNR)
\]

\[ p = \frac{5}{13} \]

Chouldechova 2016; Kleinberg, Mullainathan, Raghavan 2016
Cucumbers vs Grapes

Brosnan and de Waal
Individual Fairness

“Similar people” are treated similarly

Need “right” notion of (dis)similarity $d(u, v)$ for the specific classification task
Individual Fairness

“Similar people” are treated similarly

“You have to draw the line somewhere”
Individual Fairness

“Similar people” are treated similarly

“You have to draw the line somewhere”
Individual Fairness

“Similar people” have similar probabilities of “Yes” and “No” outcomes

“You have to draw the line somewhere”
... or do you?
Individual Fairness

“Similar people” have similar *probabilities* of “Yes” and “No” outcomes

\[ C : U \to \Delta(O) \]
\[ ||C(x) - C(y)|| \leq d(x, y) \]

Dwork, Hardt, Pitassi, Reingold, Zemel 2012
Algorithms
Vendor’s Input

\[ L(v, o) = \text{“Loss” incurred by mapping } v \text{ to } o \]
Assemble Ingredients

"Minimize vendor's utility loss, subject to the fairness conditions."

Loss function: soft constraints
Fairness conditions: hard constraints

\[
\min_{M=\{\mu_v\}_{v \in V}} \mathbb{E}_{v \in V} \mathbb{E}_{o \sim \mu_v} L(v, o)
\]

\[
\mu_v \in \Delta(\mathcal{O})
\]

\[
\|\mu_u - \mu_v\| \leq d(u, v) \quad \text{(Lipschitz)}
\]

Yields statistical parity iff \( EM(S, T) = 0 \)

Dwork, Hardt, Pitassi, Reingold, Zemel 2012
Assemble Ingredients

"Minimize vendor’s utility loss, subject to the fairness conditions."

Loss function: soft constraints
Fairness conditions: hard constraints

\[
\min_{M=\{\mu_v\}_{v \in V}} E_{v \in V} E_{o \sim \mu_v} L(v, o)
\]

\[
\mu_v \in \Delta(0)
\]

\[
\|\mu_u - \mu_v\| \leq d(u, v) \quad \text{(Lipschitz)}
\]

Generalizability via relaxing to “Probably Approximately Fair”

Rothblum and Yona 2018
Dwork, Hardt, Pitassi, Reingold, Zemel 2012
2016:

- **HPS:** Take equal FNR (and FPR) as fairness requirement
  - First learn (unfair) predictor; Apply post-processing (with appropriate training data) to learn a modified predictor which is also fair

- **JKMR:** Bandit setting, one arm per demographic group
  - Issue: Paucity of training data for $S$
  - Require worse applicant never favored over a better one (despite uncertainty)
  - Gap: Regret of learning a fair policy and regret when fairness not a concern
2017

- **KNRW: Preventing Fairness Gerrymandering**
  - Group Fairness – learn a classifier ensuring Statistical Parity (or equalizing FPR/FNR) across very large numbers of large, intersecting, groups defined by a class of functions (circuits)
  - *Fairness auditing* is hard, by reduction from agnostic learning; efficient given an appropriate algorithm for agnostic learning of the set of groups

- **HKRR: Multicalibration**
  - Approximate *calibration* for every set defined by a circuit in a predetermined set \( C \). Similar negative results to KNRW; an efficient algorithm given an agnostic learner for \( C \).
  - Use in post-processing to convert predictor to multicalibrated one with no accuracy loss.
  - *Among those rated \( v \in [0,1] \), the fraction who truly have a positive label is close to \( v \); Correct on average for those rated \( v \)
2018

- **KRR: Multifairness**
  - Oracle provides distances for a set of comparisons; algorithm learns to treat similar subpopulations similarly in expectation, for large numbers of intersecting large subsets, while preserving utility

- **GJKR: The Ethical Friend**
  - Oracle provides corrective feedback on a set of probabilities (contextual bandit setting); learns a fair policy while minimizing regret and violations of fairness constraints

---

Gillen, Jung, Kearns, Roth
Kim, Reingold, Rothblum
Evolution of Variables of Interest

- Outcomes affect future distributions
  - Chen and Hu 2017
    - Employment, temporary labor market
  - Liu, Dean, Rolf, Simchowitz, Hardt 2018
    - Loans
Adversarial Learning of Fair Representations

- Learn a censored mapping to a space $Z$
  - Encoder tries to hide membership bit, learn classifier on $Z$
  - Decoder tries to reconstruct $x$ from $z = \text{Enc}(x)$
  - Adversary tries to distinguish $\text{Enc}(x \in S)$ from $\text{Enc}(x \in T)$
Adversarial Learning of Fair Representations

- Learn a censored mapping to a space $Z$
  - Encoder tries to hide membership bit, learn classifier on $Z$
  - Decoder tries to reconstruct $x$ from $z = \text{Enc}(x)$
  - Adversary tries to distinguish $\text{Enc}(x \in S)$ from $\text{Enc}(x \in T)$

Intuition
Predictor not more powerful than $A$ \implies predictions relatively unbiased
Composition
Intuition

If all of the parts are fair, then the whole should be fair.
Reality

It’s complicated.
Task-Competitive Composition

Tasks ‘compete’ for individuals.

- Example: Advertisers compete for a single ad slot
- Goal: Individual Fairness for tech jobs advertising and groceries advertising simultaneously
Naïve Task-Competitive Composition

1. Grocery service decides whether to bid ($1)
Naïve Task-Competitive Composition

1. Grocery service decides whether to bid ($1)

2. Tech company bids among those not claimed by groceries ($0.50)

Analogous problems arise for group fairness
Tie-Breaking Function Irrelevant

**Theorem.** For any two tasks T and T’ and for any tie-breaking function, not necessarily the same for each individual, there exist classifiers C and C’ which are individually fair in isolation, but when naively combined violate multiple task fairness.

**Proof kernel (for case ties go to T)**
The difference in probability of positive classification for T’ is 

$$(1-p_u)p'_u - (1-p_v)p'_v$$

When C’ maxes out its allowance this is

$$d'(u,v) + p_vp'_v - p_up'_u$$
Subgroups & Composition

Intersection of subgroups
Fairness Gerrymandering [KNRW 17] and Multicalibration [HKRR 17] both look at how to ensure fairness on (sufficiently large) well defined subgroups, without composition.
A Fair Algorithm: Randomize Then Classify

Procedure:
- Fix a probability distribution $X$ over the tasks.
- Choose a task $T \sim X$
- Classify using a fair classifier for $T$.

Fairness:
- For every task $T$ and for every $u, v \in U$
  - $\Pr[\text{System classifies } u \text{ positively for } T] = \Pr[T \text{ chosen}] \Pr[C_T(u) = 1]$
  - $\Pr[\text{System classifies } v \text{ positively for } T] = \Pr[T \text{ chosen}] \Pr[C_T(v) = 1]$
Functional Composition
OR Fairness: Applying to multiple colleges

**Relevant outcome:** get in to at least one college.

**Definition.** A set of classifiers $C$ for task $T$ with metric $D$ satisfy OR-Fairness if for all $u, v$ in $U$ $|p_u - p_v| \leq D(u,v)$, where $p_w$ is the probability that $w$ is accepted by at least one classifier in $C$.

**Theorem.** For any nontrivial task, there exists a set of classifiers which are fair in isolation, but violate OR-Fairness.

**Theorem.** Any set of individually fair classifiers for a task which have average (over choice of classifier) probability of positive classification $> \frac{1}{2}$ for all $u \in U$ also satisfy OR-Fairness.
Dependent Classifications

In many cases outcomes are not independent:

- Can only accept $n$ students
- Must accept at least $n/2$ students who can pay tuition

Two main settings:

- *(Constrained)* Cohort Selection
- Universe Subset Problems
Interpretability and Causality

This slide intentionally left blank
On the Metric

- Locus of Unfairness
  - $\pi_x$ is the probability, over randomness in $x$ and Environment, that $x$ will succeed at Nice News
  - $|\pi_x - \pi_y|$ small $\Rightarrow x$ and $y$ should have similar probabilities of being classified as "hire"
On the Metric

- Locus of Unfairness
  - $\pi_x$ is the probability, over randomness in $x$ and Environment, that $x$ will succeed at Nasty News
  - $|\pi_x - \pi_y|$ small $\Rightarrow x$ and $y$ should have similar probabilities of being classified as “hire”

- What should metric capture? Success probability? Talent?
Final Remarks

- Truth is elusive
  - Not remedied by computer
  - Sunshine for the metrics

- Computers are not necessarily worse than humans
  - More accurate
  - More easily tested