Algorithmic Fairness & Loss Minimization

Omer Reingold, Stanford University

- Based on a joint works with
  Cynthia Dwork, , Shafi Goldwasser, Parikshit Gopalan, Úrsula Hébert-Johnson, Adam Kalai, Christoph Kern, Michael P. Kim, Frauke Kreuter, Guy N. Rothblum, Vatsal Sharan, Udi Wieder, Gal Yona

We hold these truths to be self-evident

- That algorithms are making and informing decisions all around.
  - Medical diagnoses.
  - Employment.
  - Bail.
  - Dating
  - Driving Cars.
  - Dating partners.
  - Ads we see.
  - Content we consume.
  - ….

We hold these truths to be self-evident

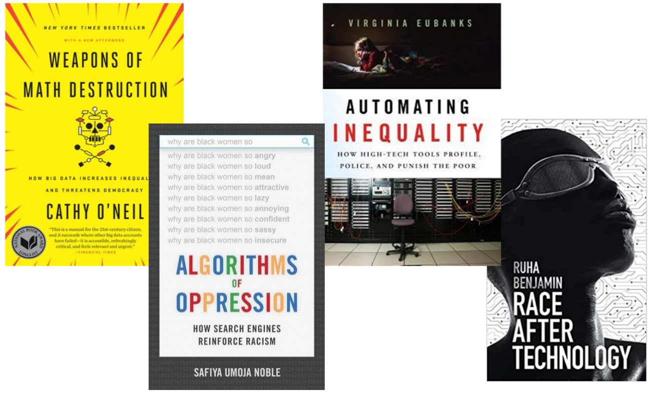
- That algorithms are making and informing decisions all around.
- That there is nothing particularly objective about algorithms.
  - Created by humans and relies on design choices.
  - In the case of learning, also rely on historic data.

We hold these truths to be self-evident

- That algorithms are making and informing decisions all around.
- That there is nothing particularly objective about algorithms.
- That concerns of unfair algorithmic discrimination are real.

LOUISE MATSAKIS BUSINESS 04.06.19 07:00 AM

FACEBOOK'S AD SYSTEM MIGHT BE HARD-CODED FOR DISCRIMINATION



Who's a CEO? Google image results can shift gender biases

UNIVERSITY OF WASHINGTON

SHARE PRINT E-MAIL

IMAGE: PERCENTAGE OF WOMEN IN TOP 100 GOOGLE IMAGE SEARCH RESULTS FOR CEO IS: 11 PERCENT. PERCENTAGE OF US CEOS WHO ARE WOMEN IS: 27 PERCENT. view more ›

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

By James Vincent on October 10, 2018 7:09 am

We hold these truths to be self-evident

- That algorithms are making and informing decisions all around.
- That there is nothing particularly objective about algorithms.
- That concerns of unfair algorithmic discrimination are real.
- That algorithmic fairness is multidisciplinary.
  - Philosophy, Law, Economics, Statistics, Social Science, …
  - Policy, Activism, Industry …

We hold these truths to be self-evident

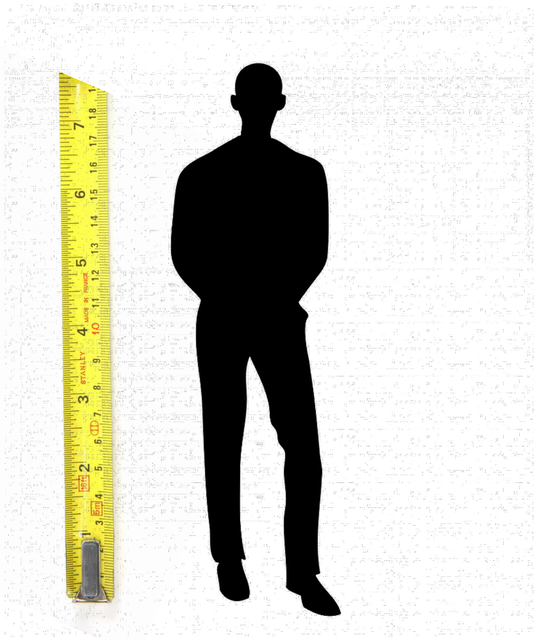
- That algorithms are making and informing decisions all around.
- That there is nothing particularly objective about algorithms.
- That concerns of unfair algorithmic discrimination are real.
- That algorithmic fairness is multidisciplinary.
- That computer scientists are needed in this multidisciplinary effort, and as a field we have a moral obligation to contribute.
  - Part of the problem - part of the solution.

We hold these truths to be self-evident

- That algorithms are making and informing decisions all around.
- That there is nothing particularly objective about algorithms.
- That concerns of unfair algorithmic discrimination are real.
- That algorithmic fairness is multidisciplinary.
- That computer scientists are needed in this multidisciplinary effort, and as a field we have a moral obligation to contribute.
- That theory has an important role to play.
  - In models, definitions, algorithms etc. (following the examples of cryptography, privacy, algorithmic game theory, …)
  - A language for discussing fairness

Risk Scores

Probability repay the loan



Probability click on this article

Probability of heart attack in 10 years

Problem Setup

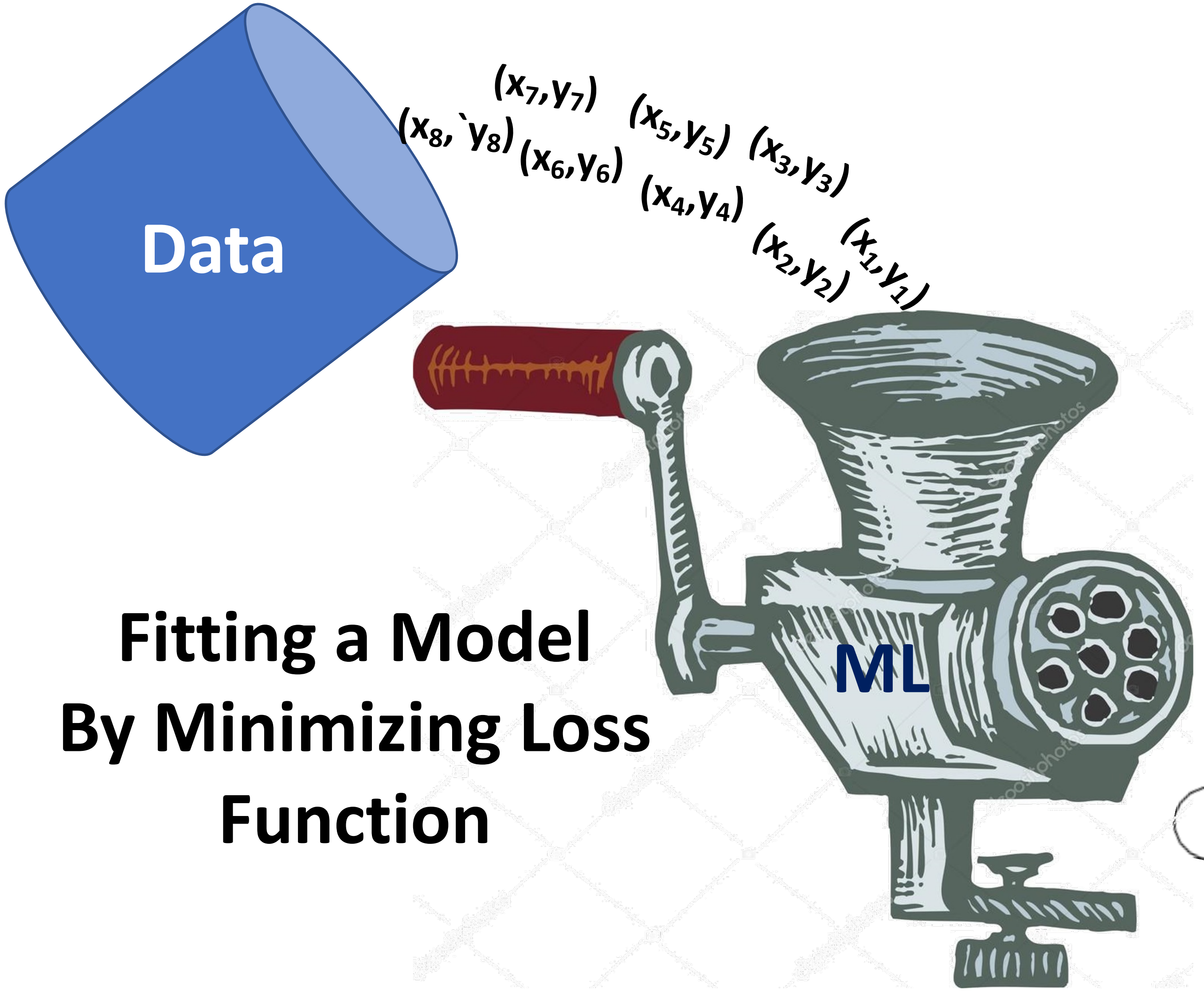
- Population χ
- $x \in \chi$ (arbitrary set of features, often identifies individual)
- y_x^* – outcome (to be predicted, binary for this talk)
- $p^*(x)$ - true Pr[y_x^* |x]

- Learning Algorithm's Input: a sample of (x, y_x^*).
- Learning Algorithm's Output: a predictor $\tilde{p}$
  - $\tilde{p}(x)$ – algorithm's estimate of $p^*(x)$.

Individual Probabilities?

- But what do individual probabilities mean?
  - What is $p^*(x)$? Non-repeatable experiment …
- Debated for decades within Statistics.
- Randomness in the environment (Nature)?
  - Limited Information.
  - Bounded computational resources.
- Scale of algorithmic decision-making calls for revisiting the question from a computational perspective.
- Cannot talk about ML fairness without providing an answer.

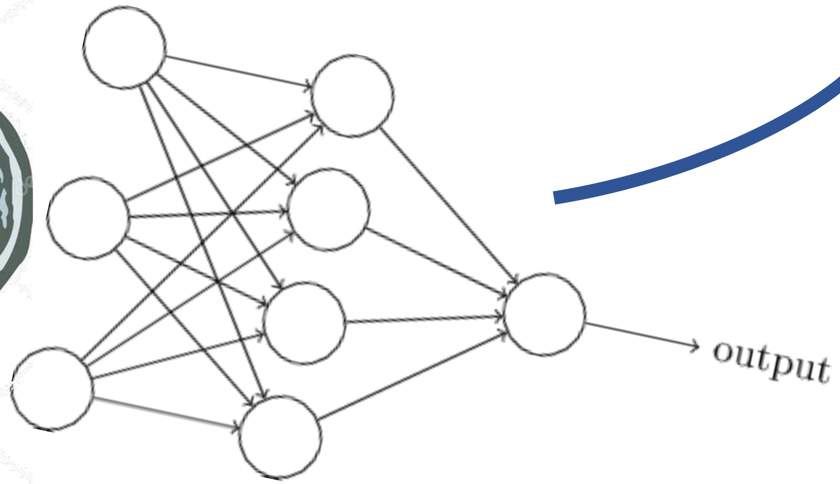
How Do Risk-Score Predictors Come to The World?



Fitting a Model By Minimizing Loss Function

Predict (and classify)

Probability of heart attack in 10 years



Learned Predictor

What's The Promise?

- Find $c \in C$ minimizing $E\left[\ell\left(y, c(x)\right)\right]$ for some loss function ℓ.
- What is the implication for individual probabilities?
- What about subgroups?

- Which loss function?

Plan

An alternative paradigm:

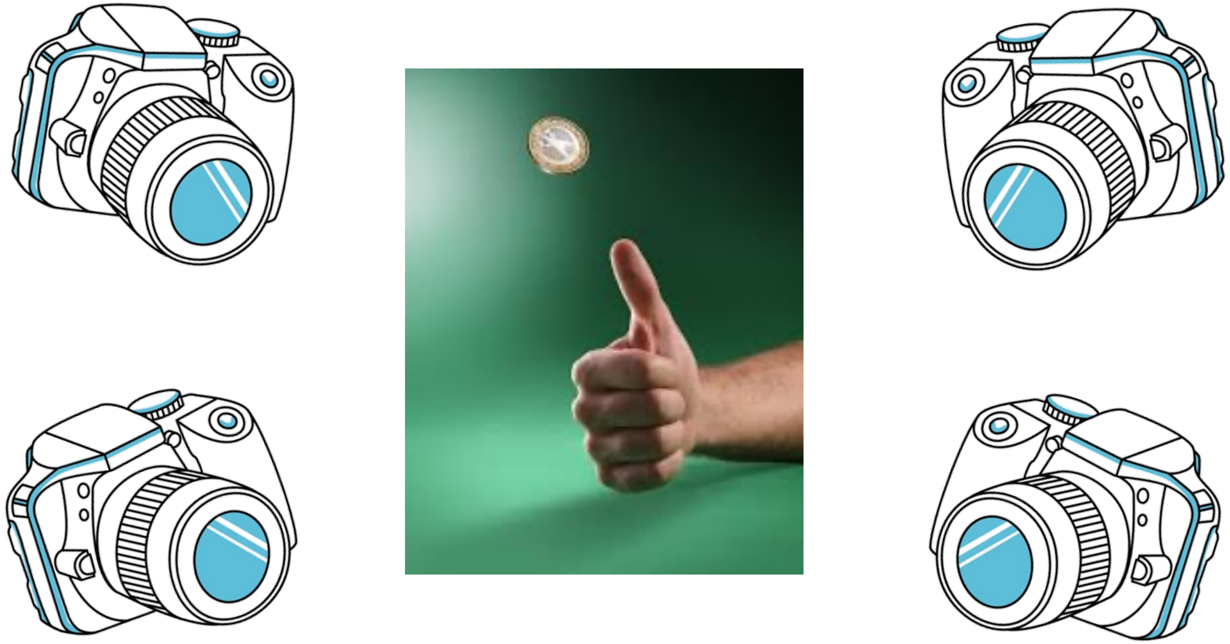
- Outcome Indistinguishability: computational perspective on the meaning of individual probabilities.
- Multicalibration: multi-group fairness – “equivalent to” OI

Good Karma:

- Omnipredictors: OI/Multicalibration implies loss minimization on steroids (answering “which loss function?”)
- Universal Adaptability: OI/Multicalibration implies an alternative to learning propensity scores.
- Multicalibration in the wild.

Randomness is in the Eye of the Beholder

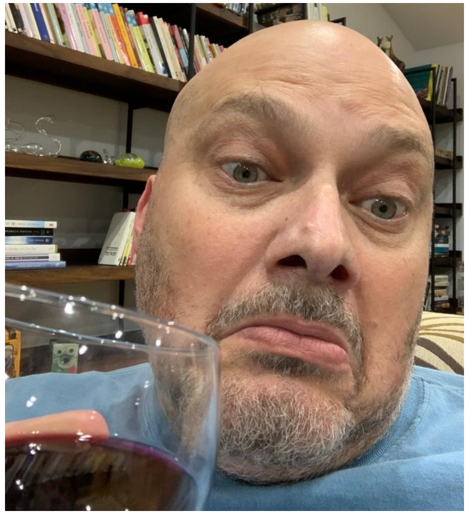
$p^* = ½$



$p^* = 0$ or1

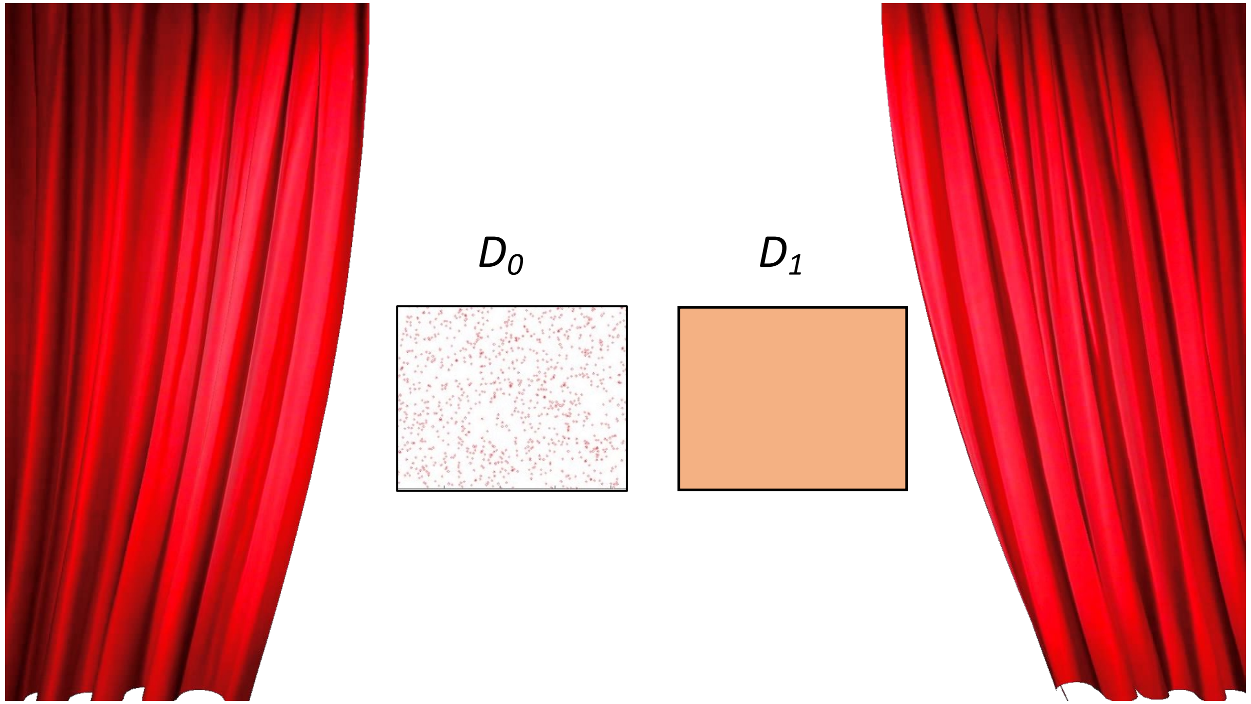
Computational Indistinguishability

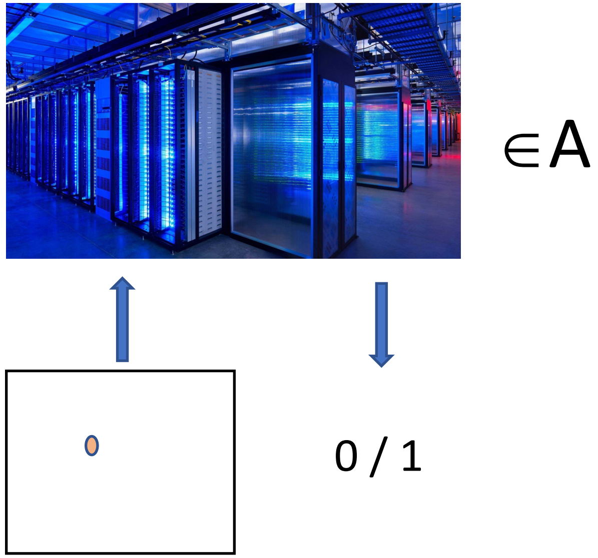
- The area of pseudorandomness (cryptography, complexity theory, …) deals with distributions that “look uniform” (indistinguishable from uniform) although they are not.



Computational Indistinguishability

- The area of pseudorandomness (cryptography, complexity theory, …) deals with distributions that "look uniform" (indistinguishable from uniform) although they are not.

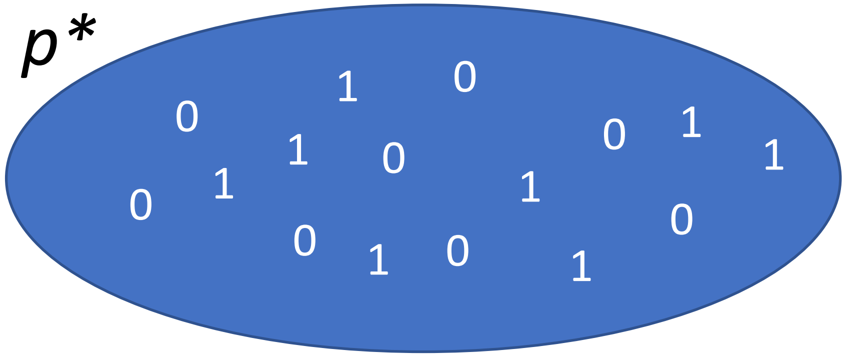




The Role of Computation

- Goal: "approximate" $p*$ from a sample of outcomes $\{(x, y_x^*)\}_x$
- Individual accuracy impossible (what we don't see we don't know), unless we make (unreasonably?) strong assumptions.
- *Any* "accuracy" depends on computational resources:

$p*$



If distribution of 0's and 1's is computationally indistinguishable then $\tilde{p} \equiv ½$ is irrefutable based on outcomes:

- Cannot distinguish real "Nature" from "simulated Nature" operating based on $\tilde{p} \equiv ½$ (even that $\tilde{p}$ very different from $p*$).

Outcome Indistinguishability [Dwork,Kim,Reingold,Rothblum,Yona 2021]

- A predictor $\tilde{p}$ gives a generative model for outcomes, where the probability x sees a positive outcome is $\tilde{p}_x$.
  - Let $\tilde{y}_x$ be outcomes sampled this way.
- Outcome indistinguishability (one version):

(x,$\tilde{p}_x$, y_x^*)

(x, $\tilde{p}_x$,$\tilde{y}_x$)

$\in \mathfrak{A}$

(x, $\tilde{p}_x$,y) 0 / 1

OI – Some Comments

- Comparing (x,$\tilde{p}_x$, y_x^*) with (x, $\tilde{p}_x$, $\tilde{y}_x$). No reference to "real, individual probabilities" p_x^* just to outcomes y_x^*.
- Definition parametrized by family $\mathfrak{A}$ of distinguishers (computational resources) and representation of individuals (information).
- We cannot empirically refute $\tilde{p}$ (given the information and computational resources):
  - Cannot distinguish true outcomes y_x^* from simulated/generated outcomes $\tilde{y}_x$.

OI – Through Multicalibration

- Outcome Indistinguishability is closely related to an earlier notion of multicalibration [Hebert Johnson-Kim-Reingold-Rothblum 18].
- Multicalibration introduced in the context of algorithmic fairness.
- An alternative to loss minimization with surprising implications.
  - In fact, it’s the same alternative.
- Let’s retell this story …

Group Notions of "Fairness"

- For a few protected groups *S*, make sure that your predictor "behaves similarly" on *S* and on the general population *U* (statistical parity, calibration, balance, …).
- Easy to work with - prevailing notions (unfortunate!).
- Very weak (easy to abuse, may cause more harm) [DHPRZ'12, …]
- Are at odds with each other and often at odds with utility [KMR16,C16].
- Alternative: Individual Fairness ("fairness through awareness" [DHPRZ'12]).
    - Treat similarly situated individuals similarly.

Which Groups? A Computational Perspective

- Often the weakness of group notions of fairness is that they do not protect important subgroups
  - Advertise burger-joint to vegetarians in the group S you want to exclude [DHPRZ'12]
- Fairness relies on identifying subgroups that are relevant to the task at hand (carnivores, qualified job applicants, …)
- Multi-Group Fairness [Hebert Johnson-Kim-Reingold-Rothblum 18, Kearns-Neel-Roth-Wu 18] offer "fairness protection" to every (large) set that can be identified given the data and given computational limitations
  - In an exact sense: the best possible

Calibration (Group Notion)

- Let S be a protected set. One fear: $\tilde{p}$ downplays fitness of S.
- $\tilde{p}$ is α-calibrated on S if
  $\forall v \in [0,1]$ and $S_v = \{x \in S : \tilde{p}(x) = v\}$
  - $\left|v - E_{x \in S_v}[o_x^*]\right| \leq \alpha$
  - (also let α-fraction of predictions be arbitrary)
- A prediction v on average means what it says.
- Extremely weak. For example, $\tilde{p}$ can be fixed on S (to the expectation) = algorithmic stereotyping.

Multicalibration (Multi-Group)

- Calibration too weak – may discriminate against qualified members of S.
- Multicalibration: calibration on every (large) set that can be identified given the data and given computational resources
- For a family of subsets C:
  $\tilde{p}$ is α-multicalibrated on C if $\forall T \in C$
  - $\tilde{p}$ is α-calibrated on T
- Think of C as computational bounds (decision trees of depth 5)
- Comes with algorithms (post-processing for multicalibration).
  - Efficient if weak agnostic learning of C is efficient.

Accuracy as Fairness?

- Multicalibration aims to address additional discrimination by ML that is not substantiated in the training data.
- It can serve as a more refined basis for affirmative action (to address other kinds of unfairness) and as a criteria to rejecting the data.
- Multi-group notions have been suggested in a variety of other settings, including to facilitate social engineering.
- Sometimes fairness is rooted in accuracy. Example, Ageism in Health Care:

  Certain diseases in elderly patients are underdiagnosed.
  - Masked as age-related symptoms.
  - Risk: ML algorithm may choose to optimize on younger patients.

  Don't want to overcorrect
  - Sometimes those *are* age-related symptoms.

OI ≅ Multicalibration

- Calibration tests ≅ general distinguishers
- Multicalibration more relatable to statisticians and ML whereas OI more relatable to complexity theoreticians and cryptographers.
- Multicalibration more natural for designing algorithms (can be viewed as a solution concept to agnostic boosting).
- OI more amendable to variants – giving the distinguisher more or less information/power.
- Several works on the relation of Multicalibration and loss minimization. Multicalibration is unlikely to be obtainable by loss minimization. Weaker notions are. (Topic for a separate talk.)

Applications of OI/Multi-Calibration

- Omnipredictors: loss minimization that simultaneously works for a huge family of loss functions.
- Universal Adaptability: adapting statistical findings to a large family of target distributions (and alternative to learning propensity scores).
- Practical basis for learning in a heterogeneous population.
- Much more
  - Real-valued labels [Jung-Lee-Pai-Roth-Vohra’20,DKRRY’22]
  - Online learning [Gupta-Jung-Noarov-Pai-Roth’21]
  - Semi-supervised learning/importance weights [Gopalan-Reingold-Sharan-Wieder’21]
  - …

Omnipredictors
[Gopalan, Tauman-Kalai, Reingold, Sharan, Wieder 2021]

- Given samples from $\mathcal{D} \sim (\mathcal{X}, \mathcal{Y})$
- Compute a hypothesis : $t \colon \mathcal{X} \to \mathbb{R}$

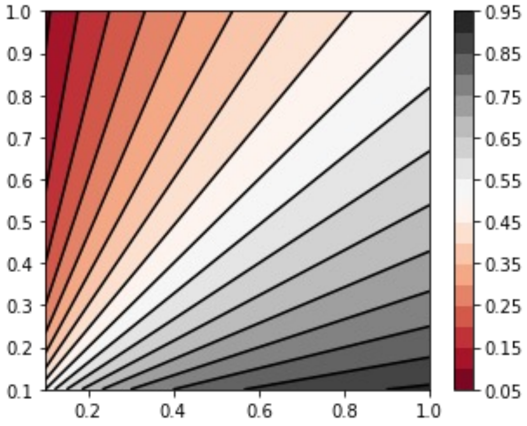
- Should this person be tested?
- What dosage of … n to give?

$[\ell(y, t(x))]$

Measures the penalty of $t(x)$ given x, y
ex: $\|t(x) - y\|_p$

- Di… different optimal *t.*

- $C =$ constant functions
- $\ell(y,t) = \|y - t\|_2$ learns the mean
- $\ell(y,t) = \|y - t\|_1$ learns the median



Which Loss?

- May not know the correct loss function at time of learning.
- May want to learn for very different loss functions (daily aspirin vs. surgery).
- May want to work for future loss functions (a future medical intervention).
- If we learned true probabilities $p^*(x)$ then, for arbitrary (somewhat nice) loss function, easy to compute optimal action.
- Omnipredictors obtain the same for a wide set of loss functions!
- Multicalibration -> omnipredictors (compared with the class *C*).
- Related work on multi-group loss minimization [Rothblum-Yona'21]

Universal Adaptability
[Kim, Kern, Goldwasser, Kreuter, Reingold 2021]

- Stanford Hospital conducts an experiment (e.g., success rate of a health policy).
  - Can Princeton Hospital rely on this study?
- Assume no unobserved confounders: Pr $(Y=1 | X=x)$, is the same in Stanford and in Princeton.
- The distribution of patients in Stanford (source) is different than in Princeton (target).
  - Subpopulations may be over/under-represented.
  - If subpopulations somewhat represented, there is a chance

Propensity Score Weighting

- Propensity score = ratio in probability that individual x appears in source (Stanford) and target (Princeton).
- Obtain unlabeled samples from source and target.
- Learn propensity score *g* from a class *C*.
- Reweight samples by *g*.
- Estimate *Y* on reweighted samples.
- Need unlabeled samples from target when training. Realistic?
  - May want to apply the Stanford study to numerous other hospitals around the world.
  - May want to apply the Stanford study to Stanford in 5 years.

Universal Adaptation?

- Intuition: if estimator learned in Stanford is multicalibrated it will directly apply for a target distributions that weigh those subpopulations differently.
- Provable: if g comes from C then C-multicalibration works as well as propensity scoring.
  - without a need for samples from target (needed only in inference time)
  - without a need to learn the propensity scores.
- Experiments: competitive and at times better performance (even when the propensity scores not in class).

Subpopulation miscalibration – an empirical evaluation of the problem and possible solution

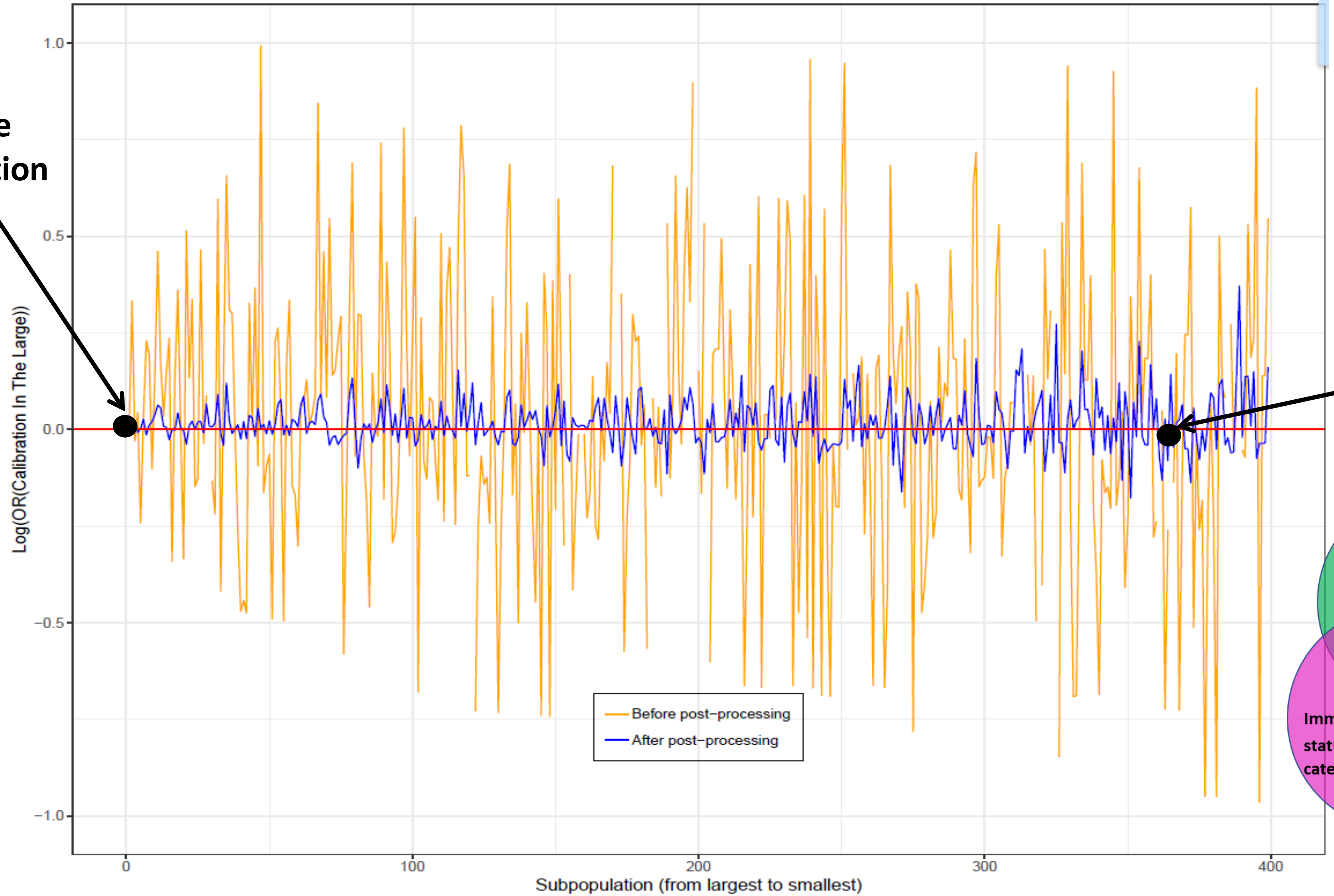
Noam Barda*[1,2], Noa Dagan*[1,3], Guy N. Rothblum[4], Gal Yona[4], Eitan Bachmat[3], Philip Greenland[5], Morton Leibowitz[1], Ran Balicer[1,2]

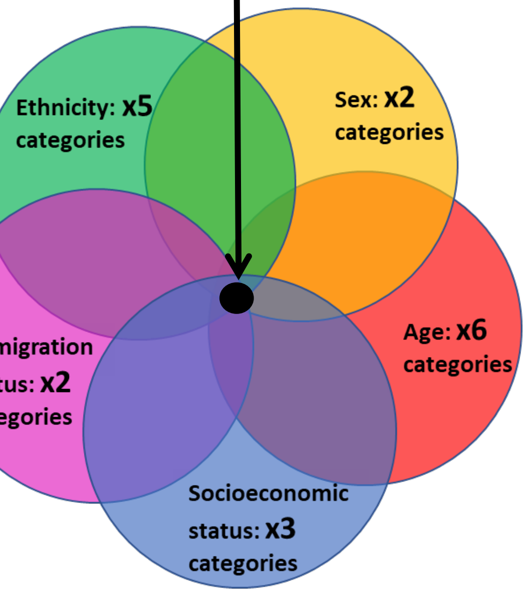
[1]Clalit Research Institute, Clalit Health Services
[2] Faculty of Health Sciences, Ben-Gurion University
[3] Department of Computer Science, Ben-Gurion University
[4] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science
[5] Feinberg School of Medicine, Northwestern University, Chicago, IL, USA
* Equal contribution

Clalit Research Institute

Ben-Gurion University of the Negev

Calibration within subpopulations before and after applying the fairness algorithm





COVID-19 Complication Predictions

- Multicalibration was used by Clalit Israel to post-process a raspatory illness complication predictor into COVID-19 complication predictor based on group statistics in China (when there were too few cases to train a new predictor).
  - In retrospect – quite successful.
- In heterogeneous populations, sometimes, fairness can promote accuracy/utility as it helps identify untapped potential/ unaccounted for risks.
- This pace of transfer from theory to practice is exciting and scary!

Parting Thoughts

- Algorithmic Fairness is both important and scientifically exciting.
- Multi-group fairness and particularly multicalibration gives meaningful fairness guarantees, and practical benefits.
- Outcome indistinguishability – computational perspective on the meaning of individual probabilities a la scientific method.
- Scientific and also practical implications. In particular – alternative to central paradigms on loss minimization and propensity scoring.