The Calculus of Inclusion

Cynthia Dwork
Harvard University
Microsoft Research

Law and the Crystal Ball: Predicting Behavior with Statistical Inference and Individualized Judgment

B. D. Underwood • Published 1979 • Economics • Yale Law Journal

Important benefits and burdens are distributed in American society on the basis of predictions about individual behavior. Release from prison, places in schools, jobs, and retail credit are among the benefits distributed to those applicants who are found most likely to succeed. The effort to predict an applicant's behavior can be made in a variety of ways: by professional experts or ordinary laymen, by use of individualized judgment or formulas that assign fixed weights to predetermined characteristics of the applicant. No matter what method is used, it typically generates controversy. This controversy is expressed in policy debates over the fairness or wisdom of choosing a particular method for selecting applicants. It also appears in litigation challenging a selection system on the ground that it violates some constitutional or statutory requirement. When the decisionmaker is a government agency, such as the parole authority or a public school, then the choice of a selection system is plainly a matter of public concern. As a political matter it involves the allocation of public resources, and as a legal matter it is subject to the requirements of fairness contained in the due process and equal protection clauses of the United States Constitution. But even when the decisionmaker is a private institution, such as a private employer or lender, its practices are often subject to public scrutiny and legal control. Many private decisionmakers are prohibited by law from discrimination on the basis of race, sex, and various other attributes. Enforcement of that prohibition requires the decisionmaker to respond to claims of illegal discrimination, by explaining his selection system, and thereby

Law and the Crystal Ball: Predicting Behavior with Statistical Inference and Individualized Judgment

B. D. Underwood · Published 1979 · Economics · Yale Law Journal

Important benefits and burdens are distributed in American society on the basis of predictions about individual behavior. Release from prison, places in schools, jobs, and retail credit are among the benefits distributed to those applicants who are found most likely to succeed. The effort to predict an applicant's behavior can be made in a variety of ways: by professional experts or ordinary laymen, by use of individualized judgment or formulas that assign fixed weights to predetermined characteristics of the applicant. No matter what method is u olicy debates ov applicants. that it violates so government agency, such as the parole authority or a public school, then the choice of a selection system is plainly a matter of public concern. As a political matter it involves the allocation of public resources, and as a legal matter it is subject to the requirements of fairness contained in the due process and equal protection clauses of the United States Constitution. But even when the decisionmaker is a private institution, such as a private employer or lender, its practices are often subject to public scrutiny and legal control. Many private decisionmakers are prohibited by law from discrimination on the basis of race, sex, and various other attributes. Enforcement of that prohibition requires the decisionmaker to respond to claims of illegal discrimination, by explaining his selection system, and thereby

"two different ways of approaching the task of moving from evidence to facts"

Theory of Algorithmic Fairness

- Definitions: Group vs Individual

  Group notions fail under scrutiny

- Group Fairness Examples
  - *Statistical parity*: demographics of accepted students are same as in population
    - 48.7% female
  - *Balance for positive class*: the average score for a positive member of A is the same as the average score for a positive member of B

DworkHardtPitassiReingoldZemel2012

Theory of Algorithmic Fairness

- Definitions: Group vs Individual
  - *Group notions fail under scrutiny*
  - *– steak ads for vegetarians*
  - *– very different distributions, reward minority that "look like" majority*
  - *– which groups? Intersectionality?*
  - *– surprisingly hard to test*
  - *– natural desiderata are mutually exclusive*

- Group Fairness Examples
  - *Statistical parity*: demographics of accepted students are same as in population
  - *Balance for positive class*: the average score for a positive member of A is the same as the average score for a positive member of B

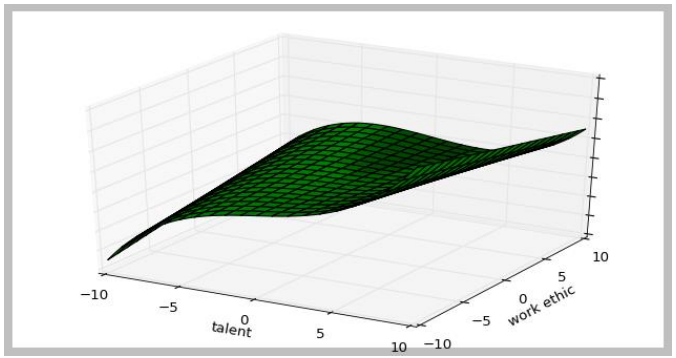
Chouldechova 2016; KleinbergMullainathanRaghavan2016

NeilWinship2019

DworkHardtPitassiReingoldZemel2012

Theory of Algorithmic Fairness

- Definitions: Group vs Individual
  - *Group notions fail under scrutiny*
  - *Individual Fairness requires a task-specific metric*



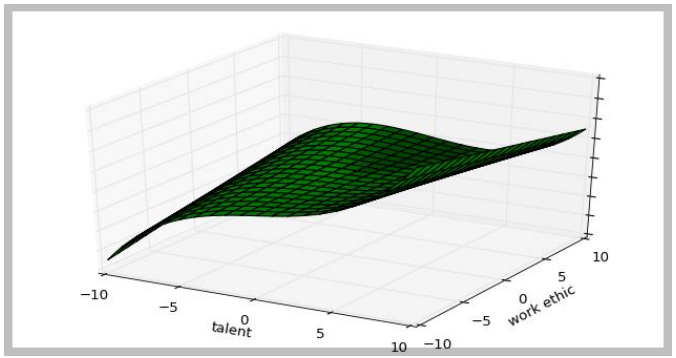
- Individual Fairness
  - People who are similar with respect to a given classification task should be treated similarly
    - $||C(x) - C(z)|| \leq d_T(x, z)$

Theory of Algorithmic Fairness

- Definitions: Group vs Individual

  Group notions fail under scrutiny

  Individual Fairness requires a task-specific metric



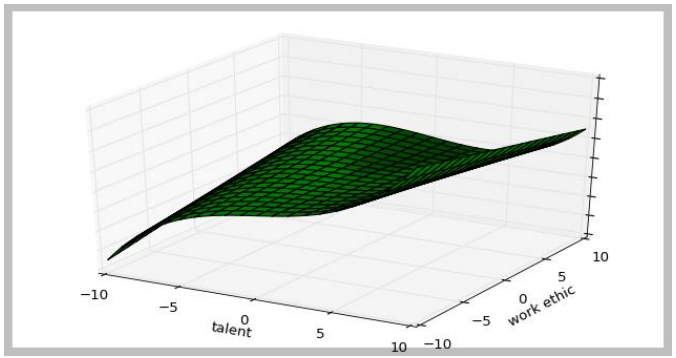
- Individual Fairness
  - People who are similar with respect to a given classification task should be treated similarly
    - $||C(x) - C(z)|| \leq d_T(x, z)$
    - Strong legal foundation
      - $d_T(x, z)$?
    - Ilvento19: O(1) hard queries
    - GillenJungRothKearns18
    - KimReingoldRothblum18
    - RothblumYona18

Theory of Algorithmic Fairness

- Definitions: Group vs Individual

  Group notions fail under scrutiny

  Individual Fairness requires a task-specific metric



- Individual Fairness
  - People who are similar with respect to a given classification task should be treated similarly
    - $\left|\left|C(x) - C(z)\right|\right| \leq d_T(x,z)$
    - Strong legal foundation
      - $d_T(x,z)$?
    - The "Metric Conjecture": a metric can be *extracted* from any "fair" system or "fairness" oracle

DworkIlventoRothblumSur2020

Multi-Group Fairness

- Definitions: Group vs Individual
  - *Group notions fail under scrutiny*
  - *Individual Fairness requires a task-specific metric*
- "Multi-X"
  - Hébert-JohnsonKimReingoldRothblum 2017
  - LearnsNeelRothWu 2017

Requirement applies simultaneously to sets in pre-specified collection C

Specifies a group fairness guarantee

Omer Reingold's Talk: Multi-Calibration

- Definitions: Group vs Individual
  - *Group notions fail under scrutiny*
  - *Individual Fairness requires a task-specific metric*
- "Multi-X"
  - Hébert-JohnsonKimReingoldRothblum 2017
  - KearnsNeelRothWu 2017

Requirement applies simultaneously to sets in pre-specified collection C

Multi-Calibration

Calibration as Fairness [KMR16]

- $\tilde{p}: Z \to [0,1]$
- $\tilde{p}$ is calibrated
- Fairness: calibrated simultaneously on (disjoint) demographic groups
  - v "means the same thing" in each group
- *Not aspirational*

- "Multi-X"
  - Hébert-JohnsonKimReingoldRothblum 2017
  - KearnsNeelRothWu 2017

Requirement applies simultaneously to sets in pre-specified collection C

Multi-Calibration

Multi-Calibration

- Powerful framework, with far-reaching applications
  - Kim, Kern, Goldwasser, Kreuter, and Reingold: Universal Adaptability
    - Propensity score reweighting *functions* captured by C allows one-time effort to yield statistics on as-yet unseen target distributions
  - Gopalan, Kalai, Reingold, Sharan, and Wieder: Omnipredictors
    - Allows one-time training to be post-processed later to approximate "best-in-class C" optimization with respect to any convex Lipschitz loss function

The Defining Problem of AI

Risk predictors assign numbers in [0,1] to individual instances:

- What is the probability that it will rain *tomorrow*?
- What is the probability that *X* will repay the loan?
- What is the probability that *this* tumor will metastasize?
- What is the probability that *Y* will commit a violent crime?

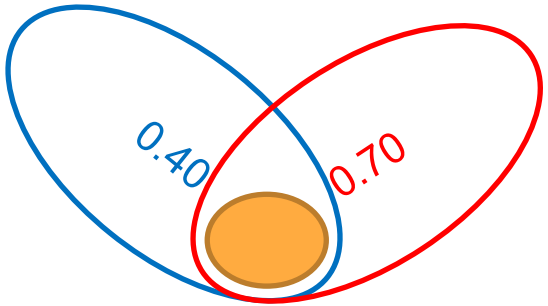
What is the “probability” of a non-repeatable event?

The Tumor Example

- "Probabilities" are learned from binary outcomes data
  – did vs did not metastasize

Locations considered in Study 1

Locations considered in Study 2

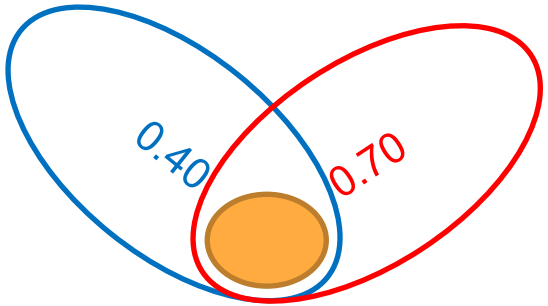


The Tumor Example

- Representation matters!
  – vector for introduction of bias

Locations considered in Study 1

Locations considered in Study 2



A Different Talk: Outcome Indistinguishability

- Definitions: Group vs Individual
  - *Group notions fail under scrutiny*
  - *Individual Fairness requires a task-specific metric*
- "Multi-X"
  - Hébert-JohnsonKimReingoldRothblum 2017
  - KearnsNeelRothWu 2017
  - DworkKumReingoldRothblumYona 2020

Requirement applies simultaneously to sets in pre-specified collection C

Outcome Indistinguishability at level $i \in [4]$

Which Sets?

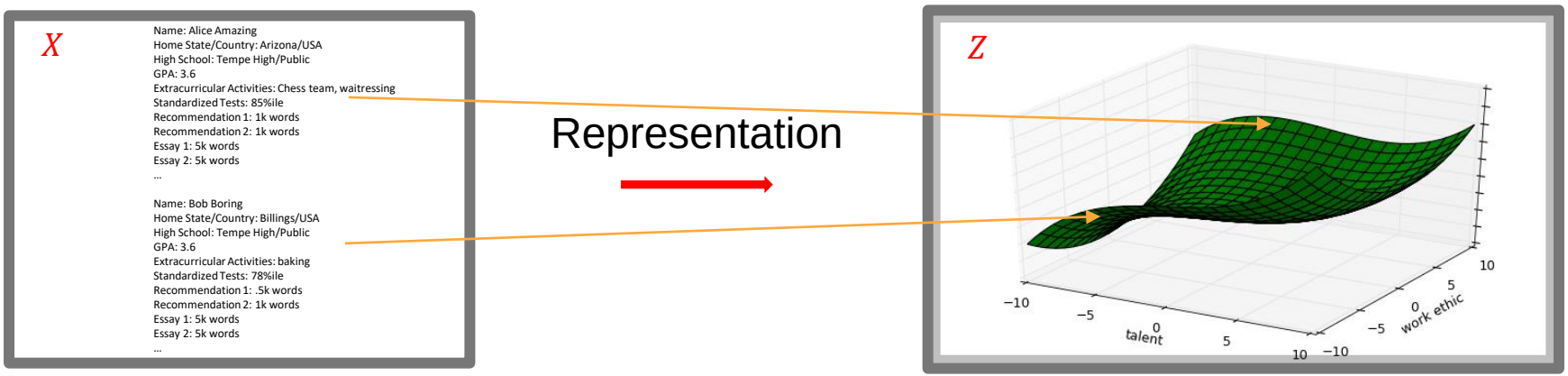
- Definitions: Group vs Individual
  - *Group notions fail under scrutiny*
  - *Individual Fairness requires a task-specific metric*
- "Multi-X"
  - Hébert-JohnsonKimReingoldRothblum 2017
  - KearnsNeelRothWu 2017
  - DworkKumReingoldRothblumYona 2020

Requirement applies simultaneously to sets in pre-specified collection C

Multi-Calibration

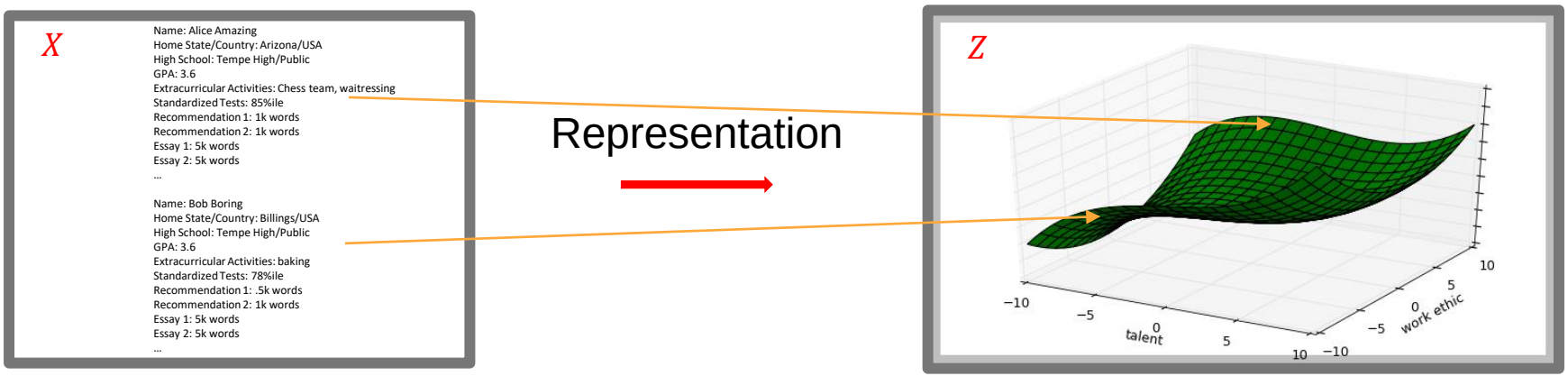
Representations (Informal)

- X: All possible real people
- Algorithm operates only on a representation of the person
  - The algorithm only knows what it is told about you
  - Distinct individuals may be mapped to the same representation



Representations (Informal)

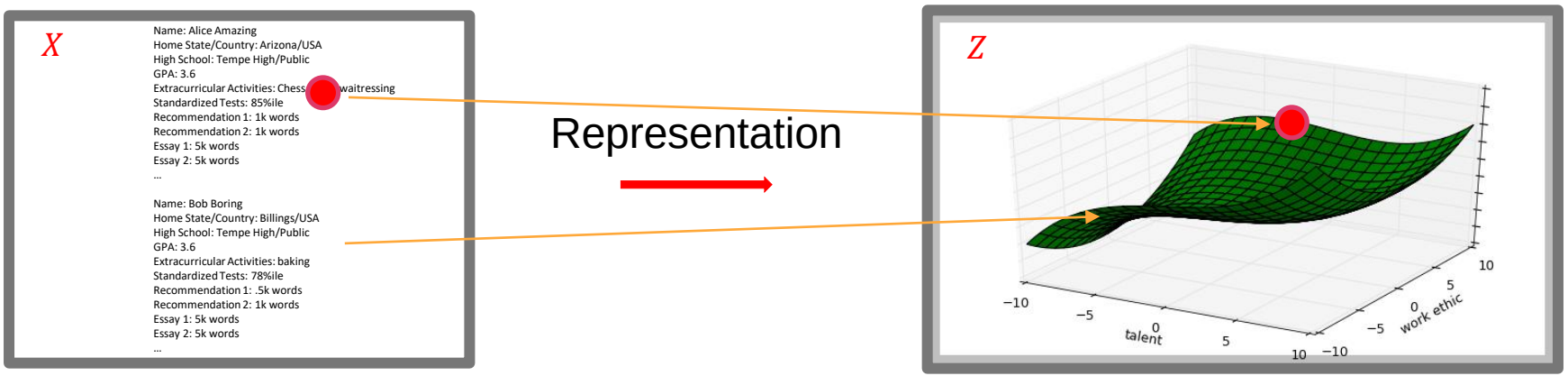
- X: All possible real people
- Algorithm operates only on a representation of the person
  - The algorithm only knows what it is told about you
  - Distinct individuals may be mapped to the same representation
  - We assume representations are rich; no collisions



Model

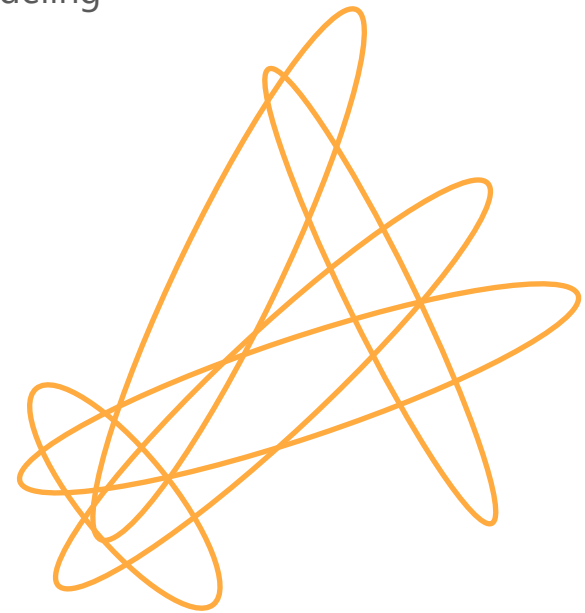
$p_i^* \in [0,1]$ assigned to all $i \in X$ by Nature; $o_i^* \sim \text{Bernoulli}(p_i^*)$

No collisions ⇒ can think of p_i^* as attaching to representation of individual i



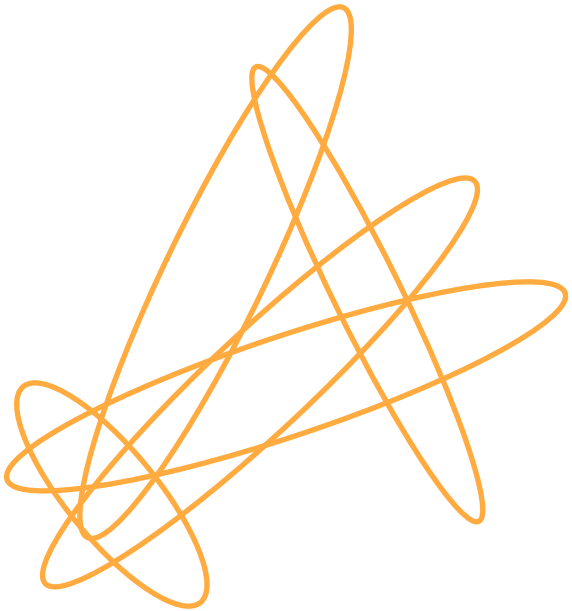
The Set Collection C

- Which sets?
  - A ubiquitous problem, eg, in synthetic data generation and modeling
  - How to think "outside the box"?
    - Non-binary individuals
    - Women without access to safe abortion
  - Inappropriate to place the onus on the members of G
    - Energy, time, knowledge of salience?



Multi-Calibration

- Which sets?
  - Complexity theory rocks!
    - Non-binary individuals
    - Women without access to safe abortion
  - … provided membership is identifiable in the base class $\mathcal{C}$



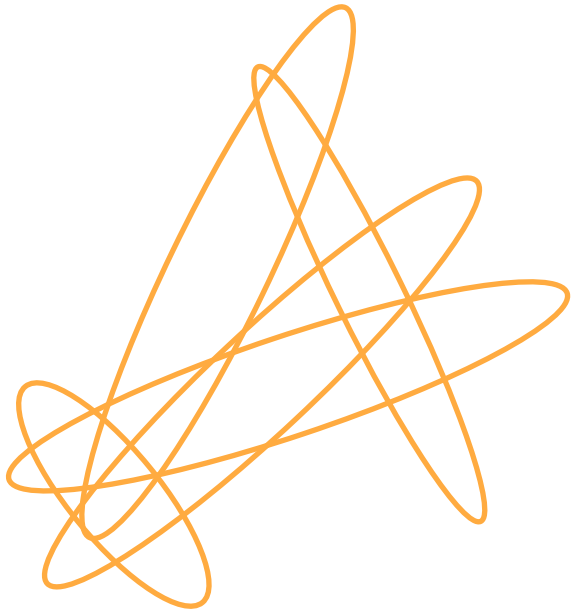
Multi-Calibration

- Which sets?
  - Complexity theory rocks!
    - Non-binary individuals
    - Women without access to safe abortion
  - … provided membership is identifiable in the base class C

If not identifiable, the learned predictor may still discriminate

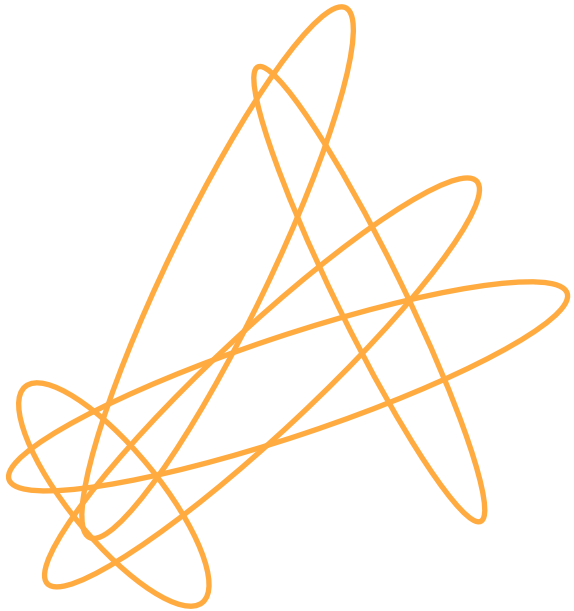
Multi-Calibration

- Which sets?
  - Complexity theory rocks!
    - Non-binary individuals
    - Women without access to safe abortion
  - … provided membership is identifiable in the base class $\mathcal{C}$
- Computational Cost?
  - Weak agnostic learning to audit for “unhappiness”
  - Use heuristics, if not learnable



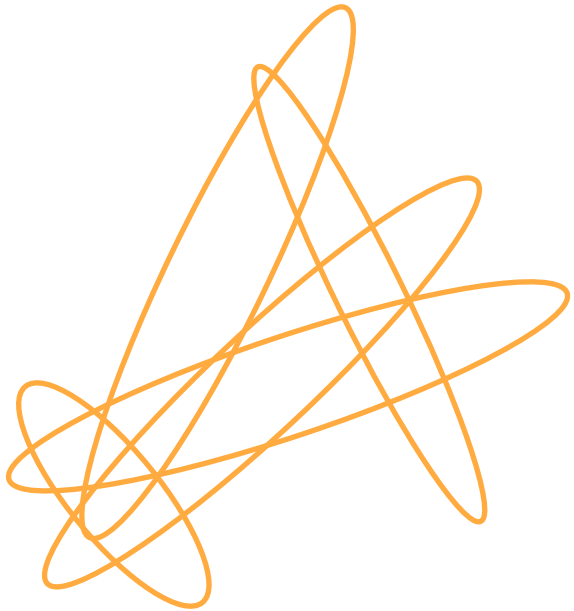
Accuracy?

If sets $S \in C$ are random, or $\perp$ to p^*, then $\hat{p}(x) = E_{(X,Y)\sim D}[Y]$ is MC wrt C



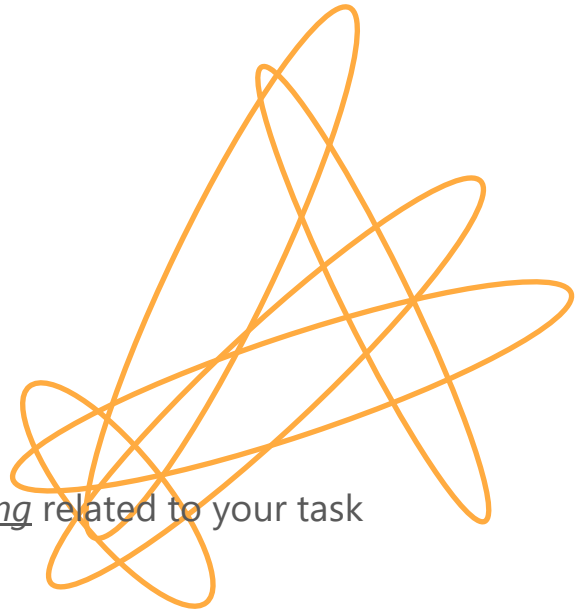
Multi-Calibration

- Which sets?
  - Complexity theory rocks!
    - Non-binary individuals
    - Women without access to safe abortion
  - … provided membership is identifiable in the base class $\mathcal{C}$
- Sets play two roles
  - Demographic
  - Differentiation



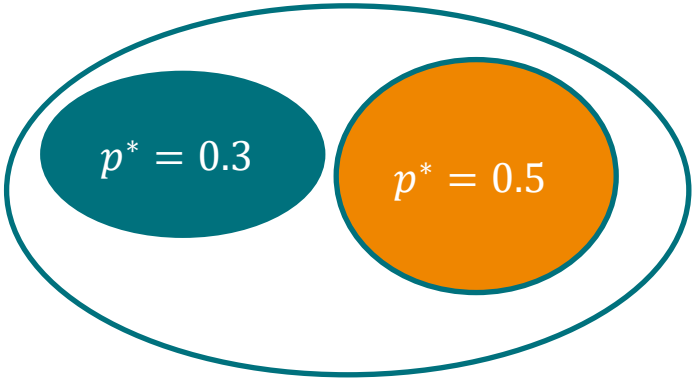
Multi-Calibration

- Which sets?
  - Complexity theory rocks!
    - Non-binary individuals
    - Women without access to safe abortion
  - … provided membership is identifiable in the base class C
- Sets play two roles
  - Demographic
  - Differentiation
    - Assume your base computational objects can do *something* related to your task



Taking Stock

- Fairness & Accuracy
  1. *De*scriptive vs *Pre*scriptive
  2. Both Fairness and Accuracy appear to depend on richness of the collection C of sets
     - Construction costs can incur factor of $|C|$
- Can we efficiently find a "small but mighty" collection C?
  - "Scaffolding sets": multi-calibration wrt to C yields a good approximation to p^*
  - Gedanken: level sets of p^*
    - Calibration on level sets ⇒ accuracy
    - Accuracy everywhere ⇒ calibration everywhere



Scaffolding Set Problem

- Efficiently find a modest-sized collection $\mathcal{S}$ of sets such that multi-calibration with respect to $\mathcal{S}$ yields a good approximation of p^*

- Proof of concept: Yes, (sometimes) we can!

Burhanpurkar,Deng,Dwork,Zhang 2021

Philosophy

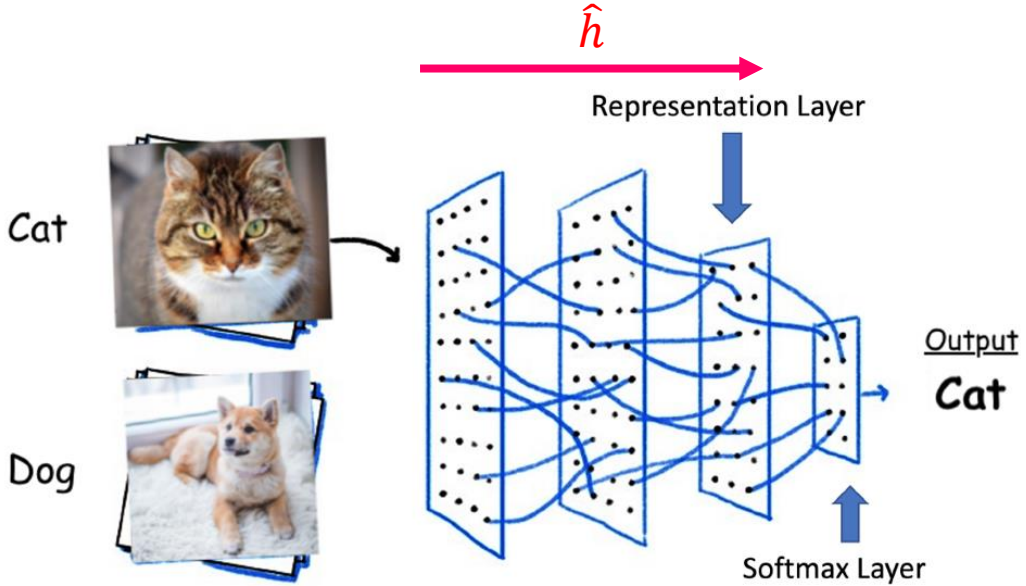
1. Use NNs to find a potential Scaffolding Set Collection $\mathcal{S}$
   - Impossible to know whether or not we have succeeded!
2. Multi-calibrate with respect to $C \cup \mathcal{S}$
   - First multi-calibrate with respect to $\mathcal{S}$ (this is easy)
   - Then post-process any way you can to also multi-calibrate wrt C

Success in Step 1 ⇒ pan-calibration; Failure in Step 1 ⇒ no harm

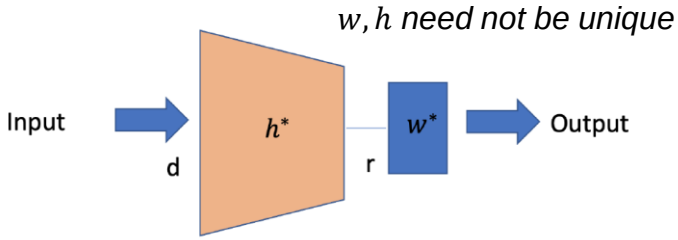
Burhanpurkar,Deng,Dwork,Zhang 2021

Folklore

Intermediate layers in a NN provide high-quality representation of the input

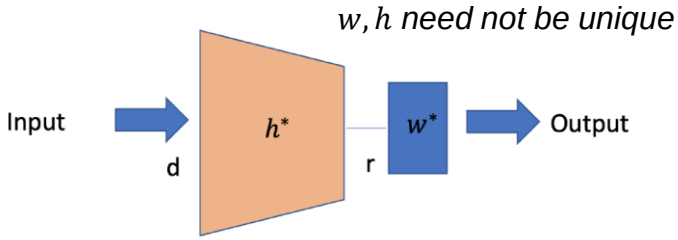


Theorem (informal)



If p^* can be well-approximated by a low-dimensional mapping composed with a low-Lipschitz suffix, then given an approximation $\hat{h}$ of the mapping, we can solve the Scaffolding Set problem for p^*.

Theorem (informal)



If p^* can be well-approximated by a low-dimensional mapping composed with a low-Lipschitz suffix, then given an approximation $\hat{h}$ of the mapping, we can solve the Scaffolding Set problem for p^*.

Examples: general linear models, single index models

Key idea: use the quantiles of $\hat{h}$ to partition range into cells of equal weight

Finding $\hat{h}$

- Learning $\hat{h}$ can be *a lot* easier than learning p^*!
  - Example: k-layer neural nets of fom $p^*(x) = W_k(\sigma(W_{k-1}\sigma(\dots\sigma(W_1 x)))$, $W_1 \in R^d$

  Here, $\hat{h}$ can be found by Ordinary Least Squares minimization with only a linear model(!)

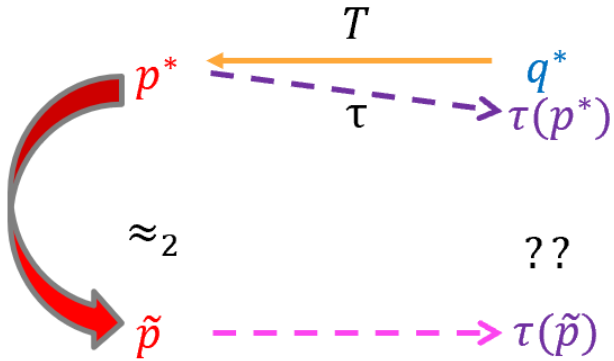
$$\hat{W}_1 = \arg\min_{W_1} \frac{1}{n} \sum_{i=1}^{n} (Y_i - W_1 X_i)^2$$

  when x's drawn from a symmetric, subgaussian, and Σ has bounded eigenvalues

- Training data for $\hat{h}$ may be abundant
  - Example: transfer learning settings (eg, same prefix, different final layer): for both covariate shift and concept shift: maybe not too many samples from any given distribution but lots of samples for training $\hat{h}$ when aggregated over different distributions

Summary and Final Remarks

- Data: The representation mapping
  - Dream: learning algorithms that "Just Say NO" to inadequate representation mappings
- Dilemma: the choice of C
  - Image of groups that are recognizable by humans IRL?
- Computational Complexity: auditing for sets in need of adjustment
  - Weak agnostically learnable
- Machine Learning
  - Heuristics for auditing
  - Scaffolding set construction
- Anti-Subordination
  - Towards the ideal world: q^*



Thank You

FAI, Bocconi University and Cyberspace, June 27, 2022