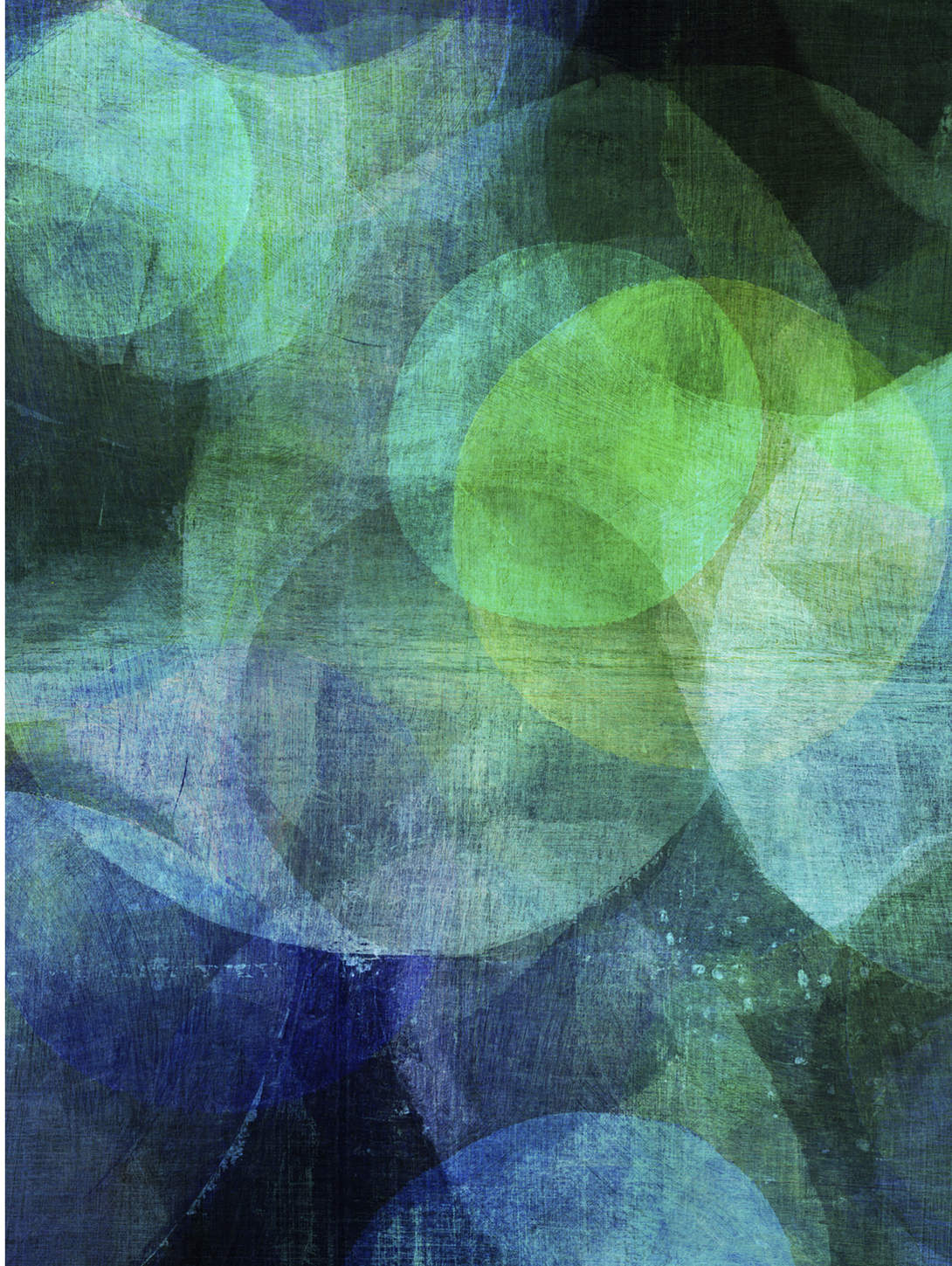
ALGORITHMIC HOOKS FOR PARTICIPATORY MACHINE LEARNING

Ira Globus-Harris
Cornell
globusharris.github.io



HI I'M IRA

- ML theory
- Currently: research prof at Cornell & Cornell Tech
  - hosted by Nikhil Garg (operations research) and danah boyd (communications)
- Previous life: Harvard privacy tools project/differential privacy/secure multiparty computation

ALGORITHMIC "HOOKS"

How can we design algorithmic systems which are *inherently* "human-centric"?

> i.e. with designated "hooks" to incorporate human input and expertise within each part of the design pipeline?

AI which…

- Empowers the expertise and lived experiences of the people the system interacts with?
- Recognizes the network of relationships and societal context it is embedded within?

How can we formalize any of this with mathematical objectives?

WHY WOULD YOU WANT TO DO THIS??

- Isn’t this falling into all of our worst traps?
  - technosolutionism, “objectivity”, abstraction…

- Yes…
- But ML models are here and being used.
  - articulating what we can demand of these systems in technical language is one avenue to hold them accountable.
  - no(?) ML theorist thinks models are perfect or can predict on individuals perfectly
  - we do still have people involved and these people hold knowledge and expertise that we could acknowledge/leverage/honor.

A COMPUTER SCIENTIST'S DESIDERATA

- Scalable: some degree of automation necessary.
- Generalizable: works outside of the training sample
- Efficient: algorithms don't run forever or require infinite data, communication, etc.
  - Corollary: want to avoid throwing out all the work you previously did on model training.
- Incentive compatible and strategy-proof:
  - Want rational, utility-maximizing players to opt in to truthful participation.

EXAMPLE 1: TWITTER

- End users, not employees, notice the cropping centers white faces.

Image from PetaPixel

EXAMPLE 1: TWITTER

- End users, not employees, notice the cropping centers white faces.

Welcome to Twitter's first algorithmic bias bounty challenge!

Entry period 7/30/21 9:01 am PT through 8/6/21 11:59 pm PT

Winners will be announced at the DEF CON AI Village workshop hosted by Twitter on August 9th, 2021.

Optionally, we invite the winners to present their work during the workshop at DEF CON although conference attendance is not a requirement to compete. The winning teams will receive cash prizes via HackerOne:

$3,500 1st Place
$1,000 2nd Place
$500 3rd Place
$1,000 for Most Innovative
$1,000 for Most Generalizable (i.e., applies to the most types of algorithms).

Images from PetaPixel & HackerOne

EXAMPLE 1: BIAS BOUNTIES

Twitter tried this, but they used a panel of judges.

- How do you scale?
- How do you fix the model so the harm doesn't occur again?
  - Iteratively, without undoing previous fixes
- How do you convince companies to do this?
  - Admitting a harm has occurred or judging the "worthiness" of an example of bias admits legal liability
  - Incentive-compatible and non-gameable payment/reward mechanisms?

EXAMPLE 1: BIAS BOUNTIES

Big Tech Company Inc©™

Model f

Description of unfairness

Concerned Community Members

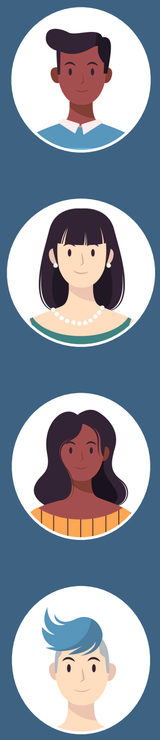
EXAMPLE 1: BIAS BOUNTIES

Big Tech Company Inc©™

Model f

Description of unfairness

Model f'



Concerned Community Members

EXAMPLE 1: BIAS BOUNTIES

Big Tech Company Inc©™

Model f

Description of unfairness

Model f'

Reward to bounty hunter

Concerned Community Members

NORMATIVE DESIDERATA

- Honor the knowledge that people hold
- Catch harms before severe impacts
- Empower the people involved: provides them with recourse, reparation, ??
- Prevent the harm from happening again

WHAT IS A SUFFICIENT DESCRIPTION OF UNFAIRNESS OR BIAS?

Your model can't identify cats!

Bounty Hunter

Big Tech Company Inc©™

WHAT IS A SUFFICIENT DESCRIPTION OF UNFAIRNESS OR BIAS?

Your model can't identify cats!

Bounty Hunter

This is a mission-critical problem!

Big Tech Company Inc©™

WHAT IS A SUFFICIENT DESCRIPTION OF UNFAIRNESS OR BIAS?

Your model can't identify cats!

Bounty Hunter

Big Tech Company Inc©™

WHAT IS A SUFFICIENT DESCRIPTION OF UNFAIRNESS OR BIAS?

Your model can't identify cats!

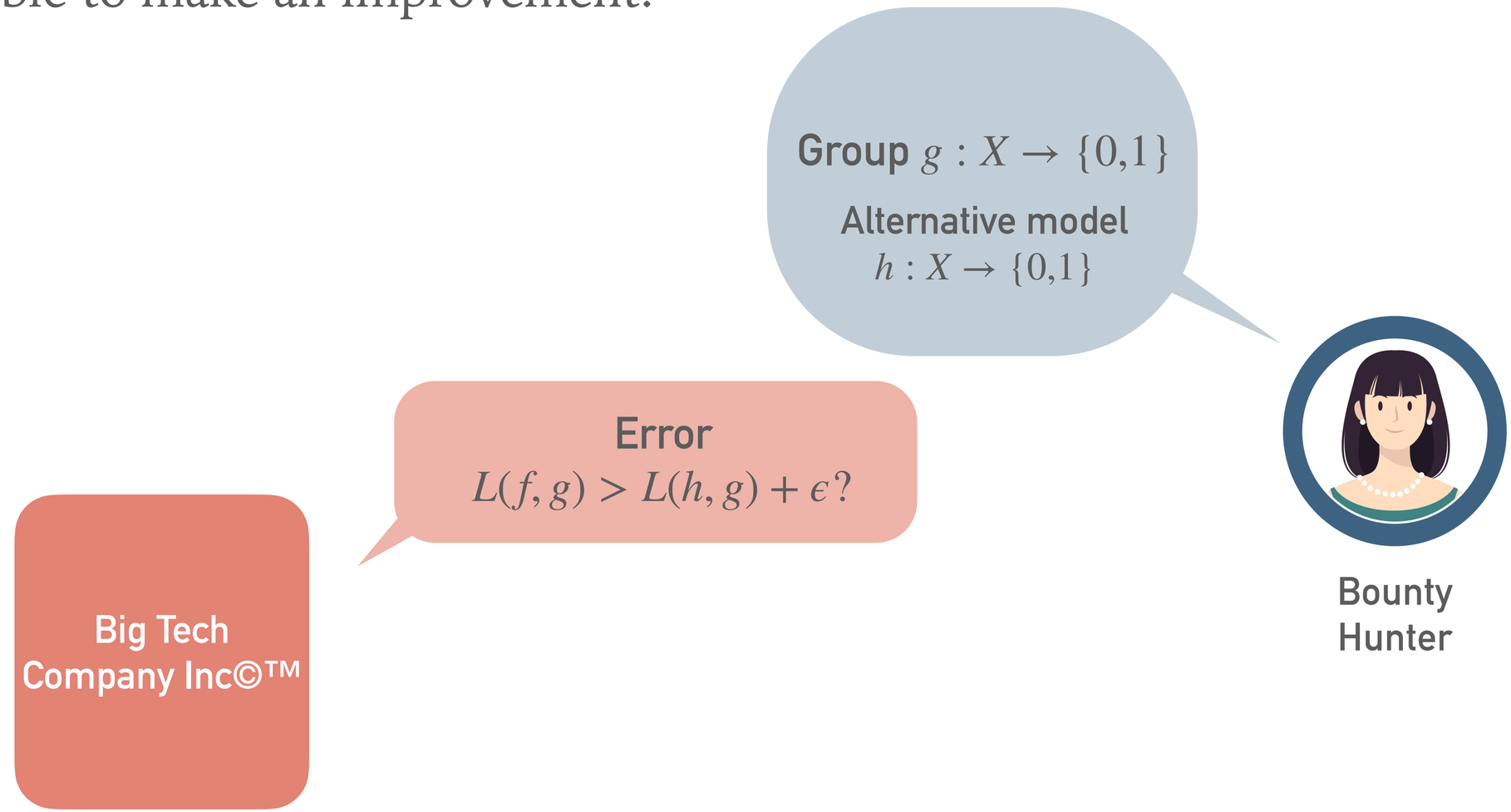
Bounty Hunter

Your cat is very blurry!

Big Tech Company Inc©™

ONE SOLUTION: A HIGHER BAR FOR THE SUBMISSIONS

- Require submissions not only demonstrate existence of model error, but that it is possible to make an improvement.



ONE SOLUTION: A HIGHER BAR FOR THE SUBMISSIONS

- Require submissions not only demonstrate existence of model error, but that it is possible to make an improvement.
- Then, ensemble over the models in a way that guarantees monotonic improvement in overall & group-wise error rates.
- Prevents gaming via restrictions on amount you can infer from interactions with the system.

CHALLENGES

- Requires submissions to be from the same distribution as the training was.
  - In reality, harms occur due to edge events that are not well-supported by training data.
- High technical barrier to entry for bounty hunters.
  - The real dream: this form of system as a back-end to a crowdsourced incident reporting system.
  - Open questions here:
    - How to reward individual incident reports?
    - How to evaluate/prevent spamming and noise in submissions?

EXAMPLE 2: COLLABORATIVE MACHINE LEARNING

Should I admit this patient?

Risk Predictor

Risk Score

Yes/No

Doctor uses combination of predictor and their intuition to make decision

HERE, THE DOCTOR AND RISK PREDICTOR KNOW DIFFERENT THINGS ABOUT THE PATIENT

Medical history

Family history

Lab reports

Training data

Distributed patient data

Pre-trained info

Risk Predictor

Risk Score

Patient disposition

Patient appearance

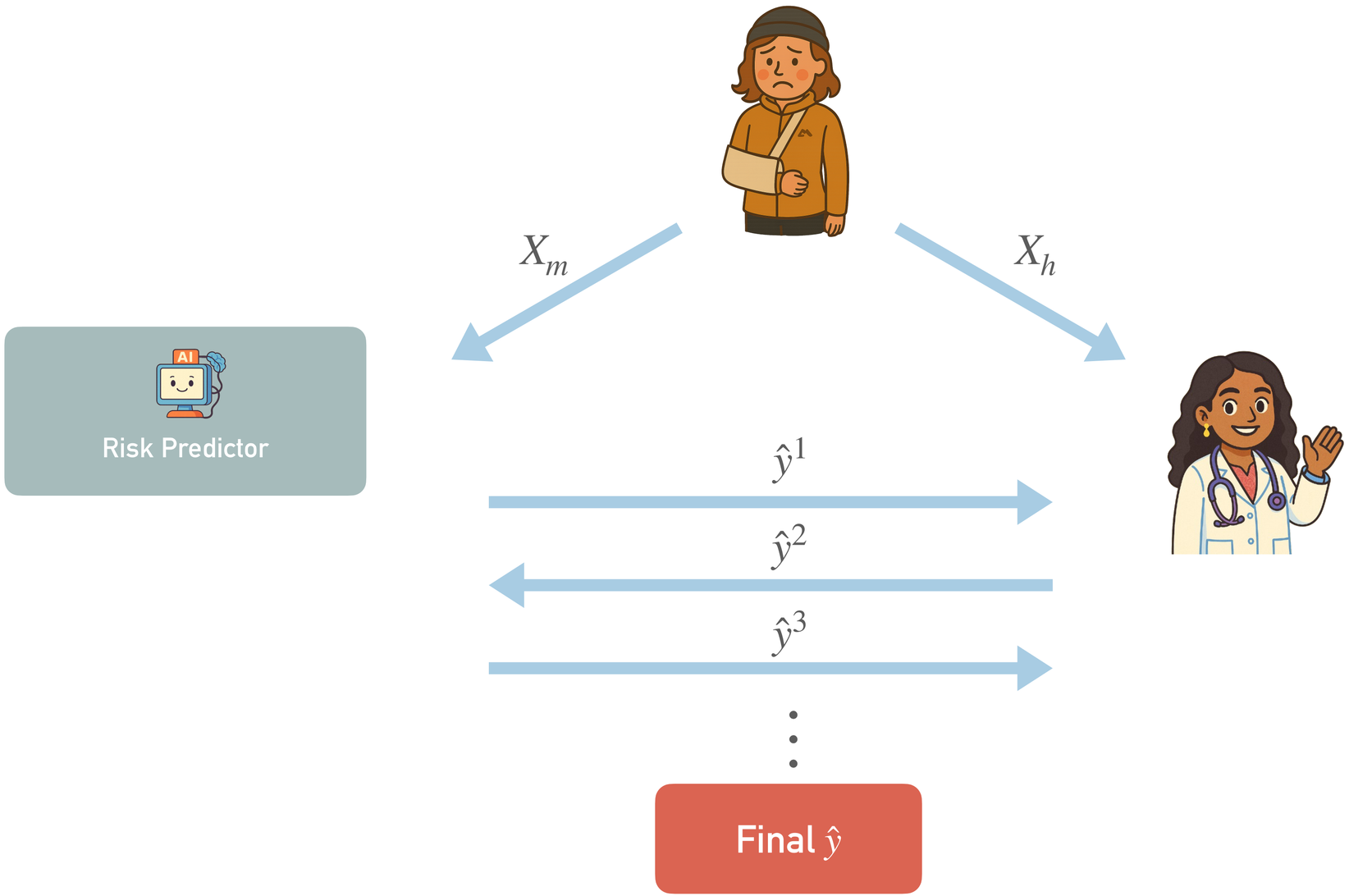
Relevant context

How do we build ML systems that are optimized for this?

CHALLENGES OF THE COLLABORATIVE SETTING

- The AI model and human doctor can't easily communicate the "features" about the patient.
- Could just have them communicate their predictions and then aggregate them.
  - Problem: there will always be a "hard case" where your aggregation rule makes you worse off than without collaboration.
- Alternative approach: iterative communication until agreement.

COLLABORATIVE LEARNING FRAMEWORK



GUARANTEES OF COLLABORATIVE FRAMEWORK

- Efficient conversations end in agreement
- The thing they agree on is as good on average as the better of the two agents
- Assuming that for any collection of patients, one of the two agents can do marginally better than a constant predictor, the thing they agree on is on average as good as what they could have predicted in the world where they told each other all the features.

THE CHALLENGES…

- “Efficient” means “polynomial”…
- “Weak assumptions on human decision makers” means “they have access to a squared error oracle in the sky instead of assuming they are a rational Bayesian”…
- If the agents have misaligned utilities (e.g. their best response action is different for the same context), no guarantees.