AIM 2: Artificial Intelligence in Medicine II

Harvard - BMIF 203 and BMI 702, Spring 2025

Lecture 7: Explainability and interpretability in medical AI, Feature importance and Shapley values, Bias and fairness in biomedical AI, Discussion: Is explainability critical or overrated?

HARVARD MEDICAL SCHOOL

Kempner INSTITUTE For the Study of Natural & Artificial Intelligence at Harvard University

BROAD INSTITUTE

Marinka Zitnik
marinka@hms.harvard.edu

Outline of today’s class

- **What is trustworthy AI?**
- Explaining AI predictions
- Definitions of fairness in AI
- Framework for fair AI
- Algorithmic fairness criteria
    - Individual fairness
    - Group fairness

Trustworthy ML

- ML models are increasingly being deployed in real-world applications
  - It is critical to ensure that these models are behaving responsibly and are trustworthy
- There has been growing interest to develop and deploy ML models and algorithms that are:
  - Not only accurate
  - But also explainable, fair, privacy-preserving, causal, and robust
- This broad area of research is commonly referred to as trustworthy ML

Motivation

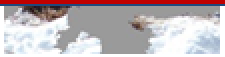
Model understanding is absolutely critical in several domains - particularly those involving **high stakes decisions**

Why model understaning?

Input

Predictive Model

Prediction = Siberian Husky



This model is relying on incorrect features to make
iction!! Let
model

Model understanding facilitates debugging

Why model understanding?

Defendant Detai

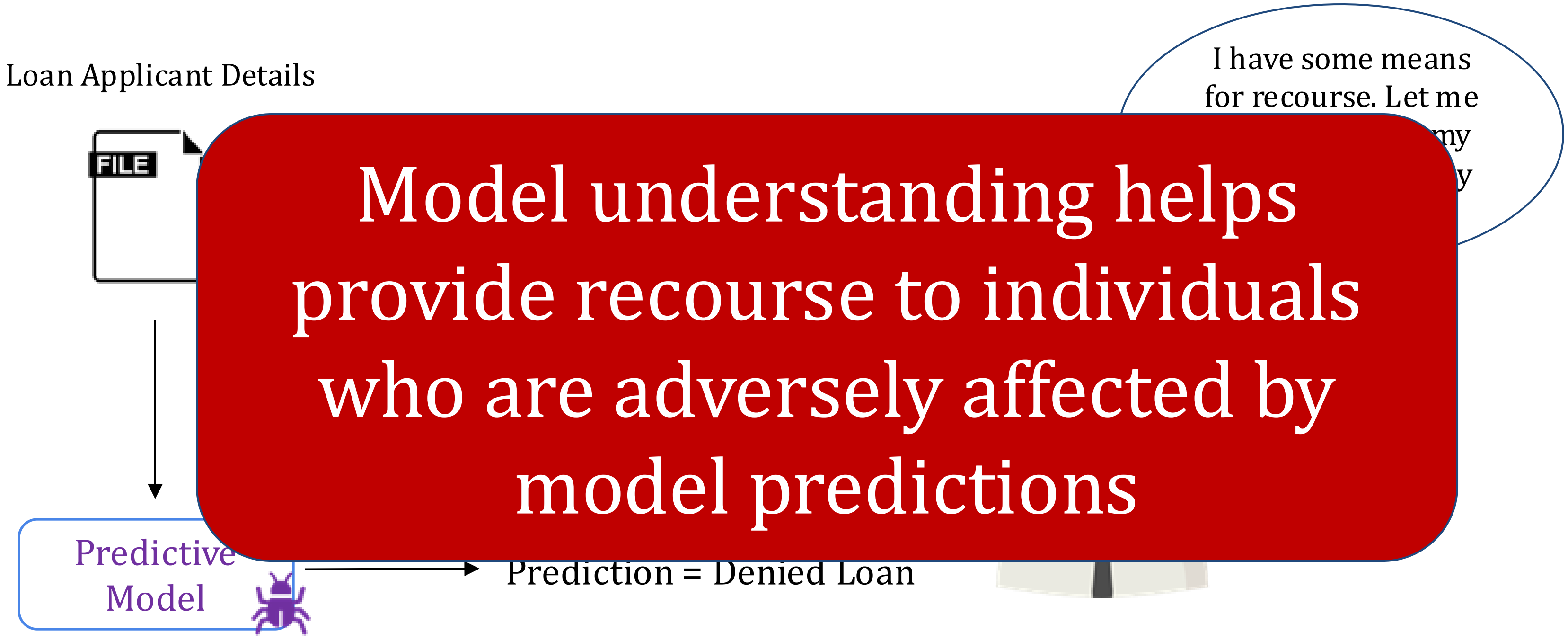
Model understanding facilitates bias detection

This prediction is
biased. Race and
ing
the
!

Predictive Model

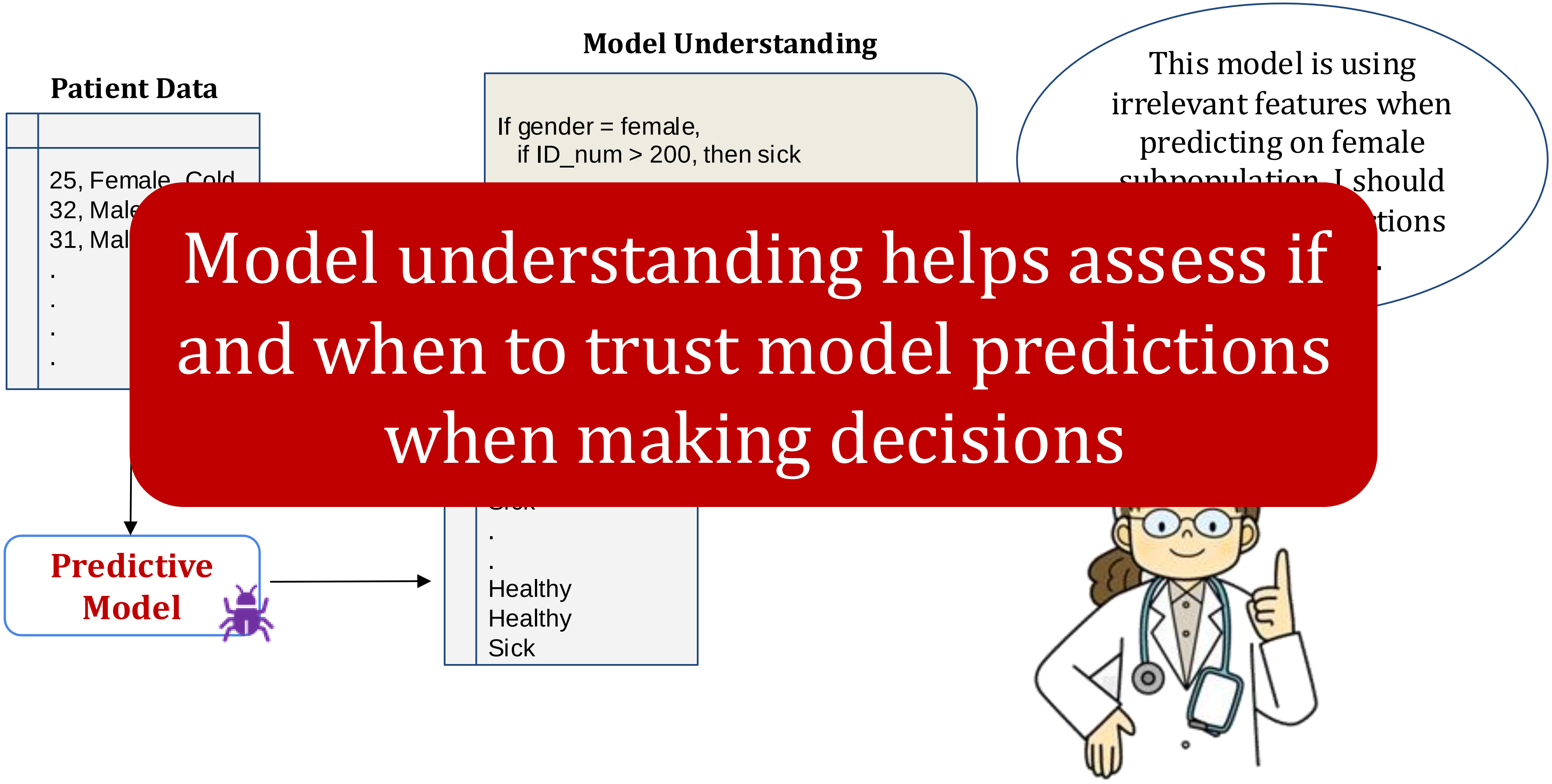
Prediction = Do not release on bail

Why model understanding?



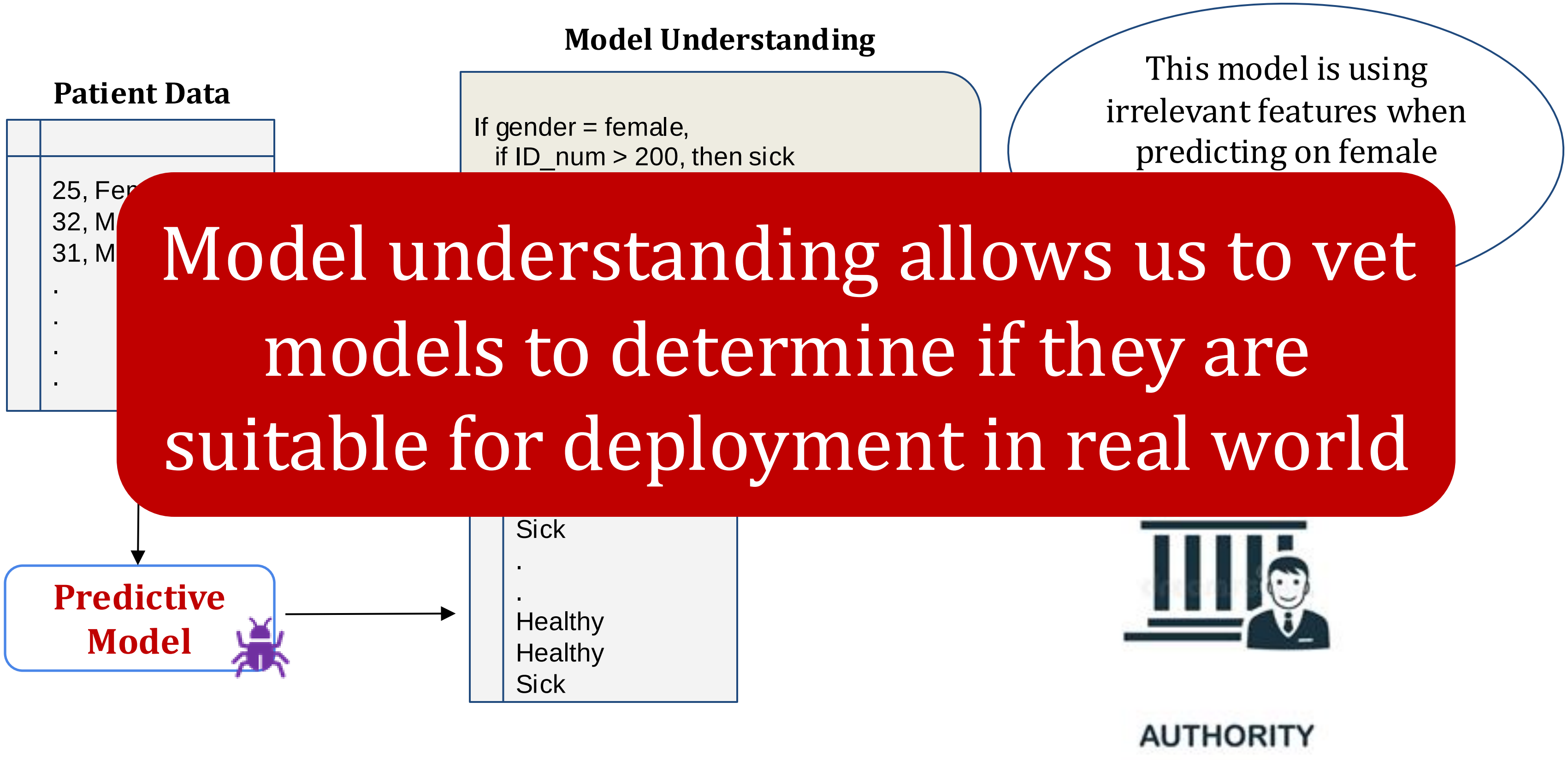
Model understanding helps provide recourse to individuals who are adversely affected by model predictions

Motivation: Why model understanding?



Model understanding helps assess if and when to trust model predictions when making decisions

Motivation: Why model understanding?



Patient Data

25, Fe
32, M
31, M
.
.
.
.

Model Understanding

If gender = female,
if ID_num > 200, then sick

This model is using irrelevant features when predicting on female

Predictive Model

Sick
.
.
Healthy
Healthy
Sick

AUTHORITY

Model understanding allows us to vet models to determine if they are suitable for deployment in real world

Why should I care about understanding ML models?

Utility

Debugging

Bias Detection

Recourse

If and when to trust model predictions

Vet models to assess suitability for deployment

Stakeholders

End users (e.g., loan applicants)

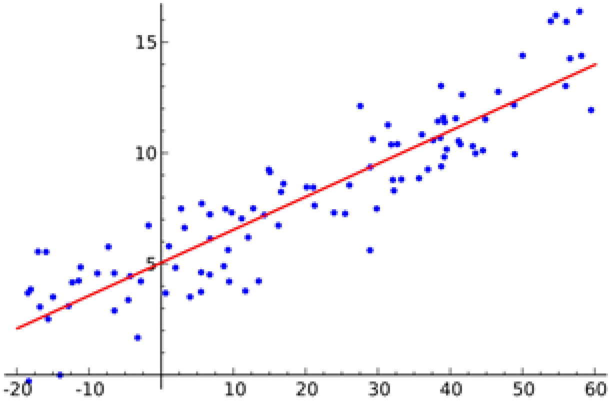
Decision makers (e.g., doctors, judges)

Regulatory agencies (e.g., FDA, European commission)

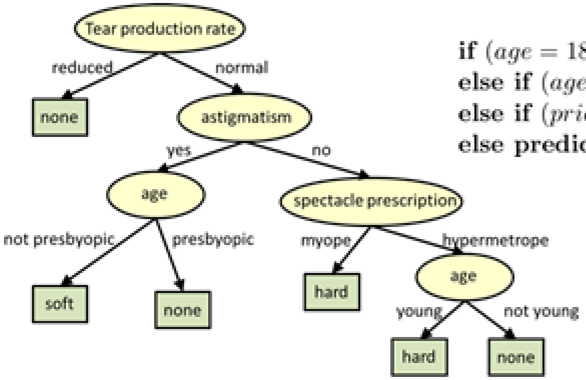
Researchers and engineers

Achieving model understanding

Goal: Build <u>inherently interpretable</u> predictive models



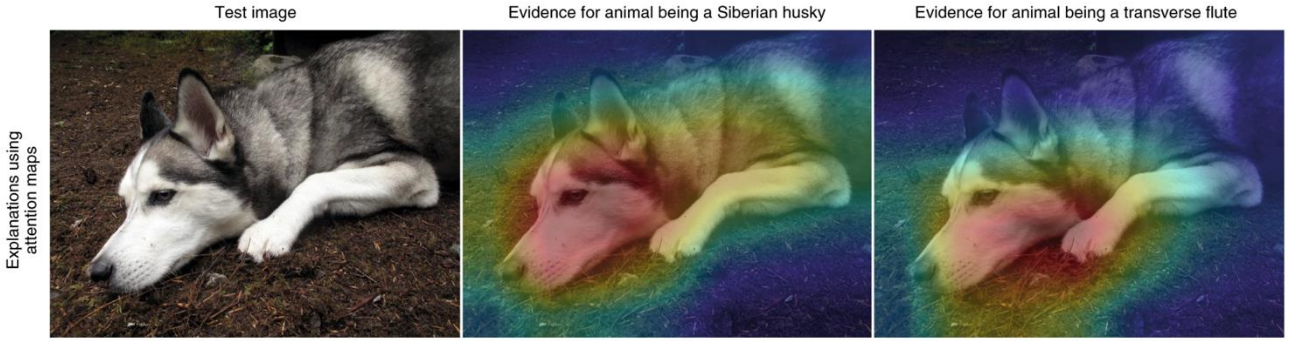
Linear regression



Decision trees

if $(age = 18 - 20)$ and $(sex = male)$ then predict yes
else if $(age = 21 - 23)$ and $(priors = 2 - 3)$ then predict yes
else if $(priors > 3)$ then predict yes
else predict no

Decision rules



Explanations using attention maps
Test image
Evidence for animal being a Siberian husky
Evidence for animal being a transverse flute

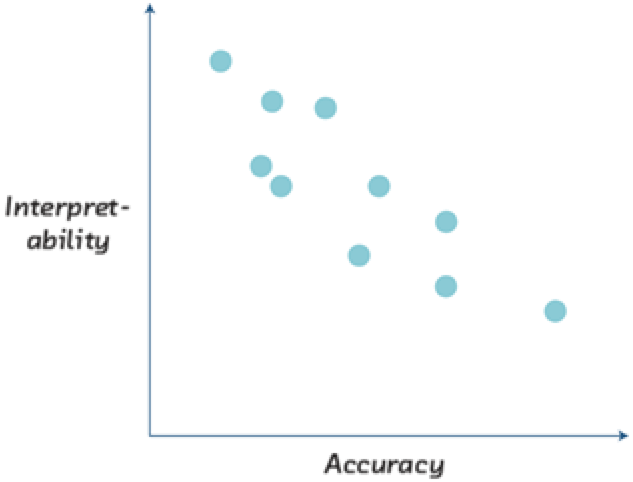
Saliency map of a black box (deep learning) model does not explain anything except where the model is looking: We have no idea why this image is labeled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes

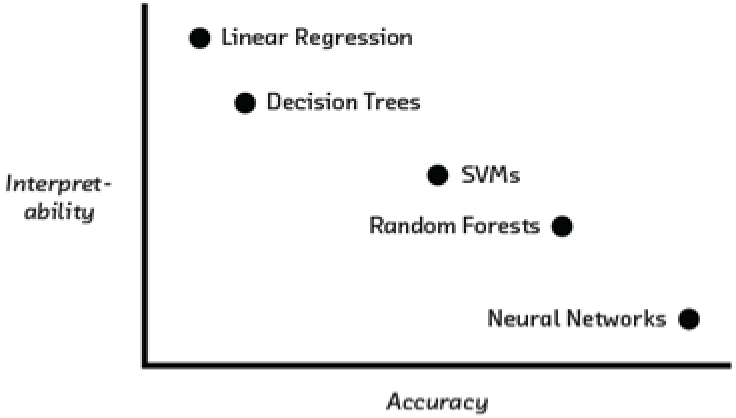
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence* 2019

Inherently interpretable models vs. post hoc explanations

Accuracy-interpretability trade offs may exist in certain settings

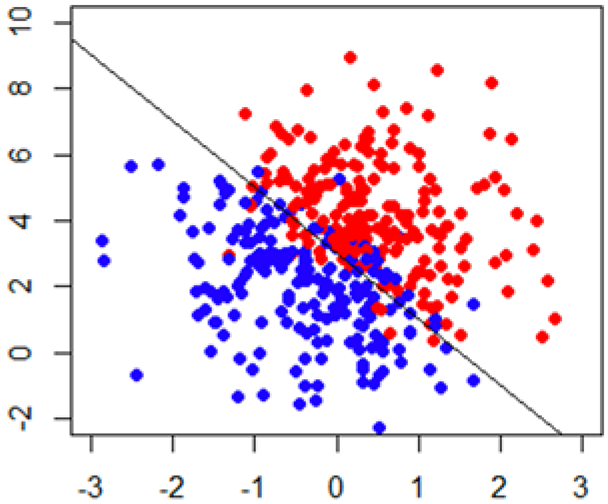
Example



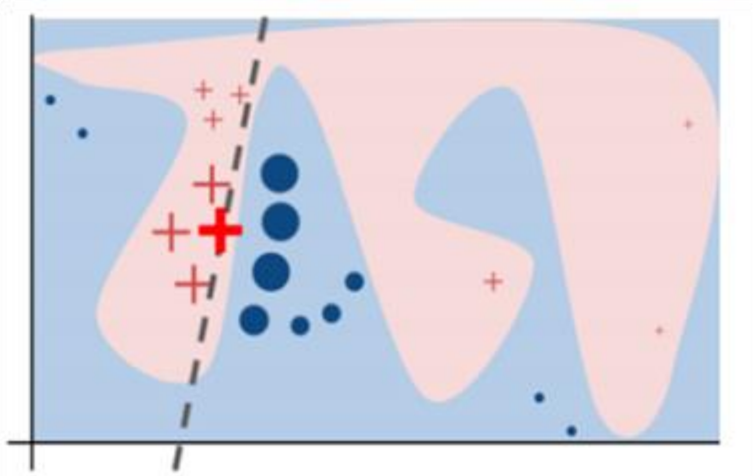


[Cireșan et. al. 2012, Caruana et. al. 2006, Frosst et. al. 2017, Stewart 2020]

Inherently interpretable models vs. post hoc explanations



Build interpretable and accurate models

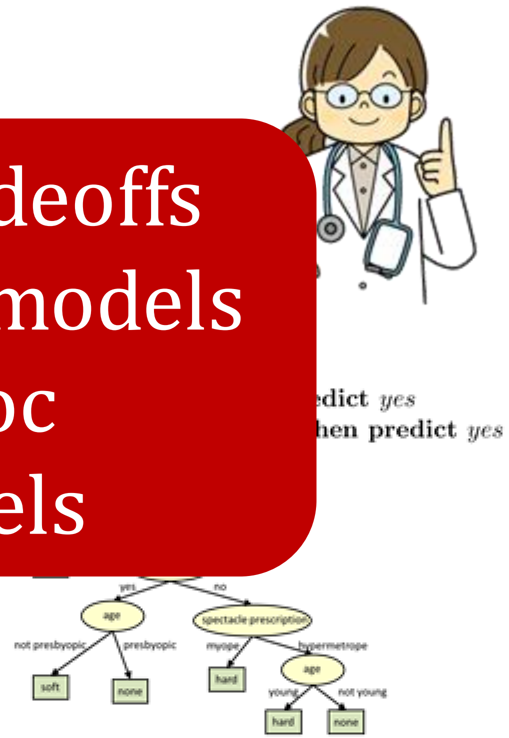


Complex models might achieve higher accuracy

Achieving model understanding

Explain pre-built models *in a post-hoc manner*

Interpretability/accuracy tradeoffs and proliferation of black box models force us to rely on post hoc “explanations” of ML models



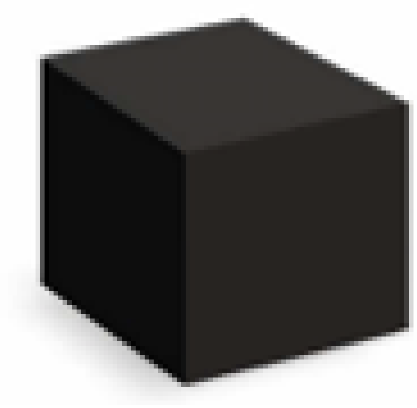
[Ribeiro et. al. 2016, 2018; Lakkaraju et. al. 2019]

Inherently interpretable models vs. post hoc explanations

- If you can build an interpretable model which is also adequately accurate for your setting, DO IT!

- Sometimes, you don’t have enough data to build your model from scratch
- And, all you have is a (proprietary) black box!
- Post hoc explanations come to the rescue!

Next: Overview of post hoc explanations methods

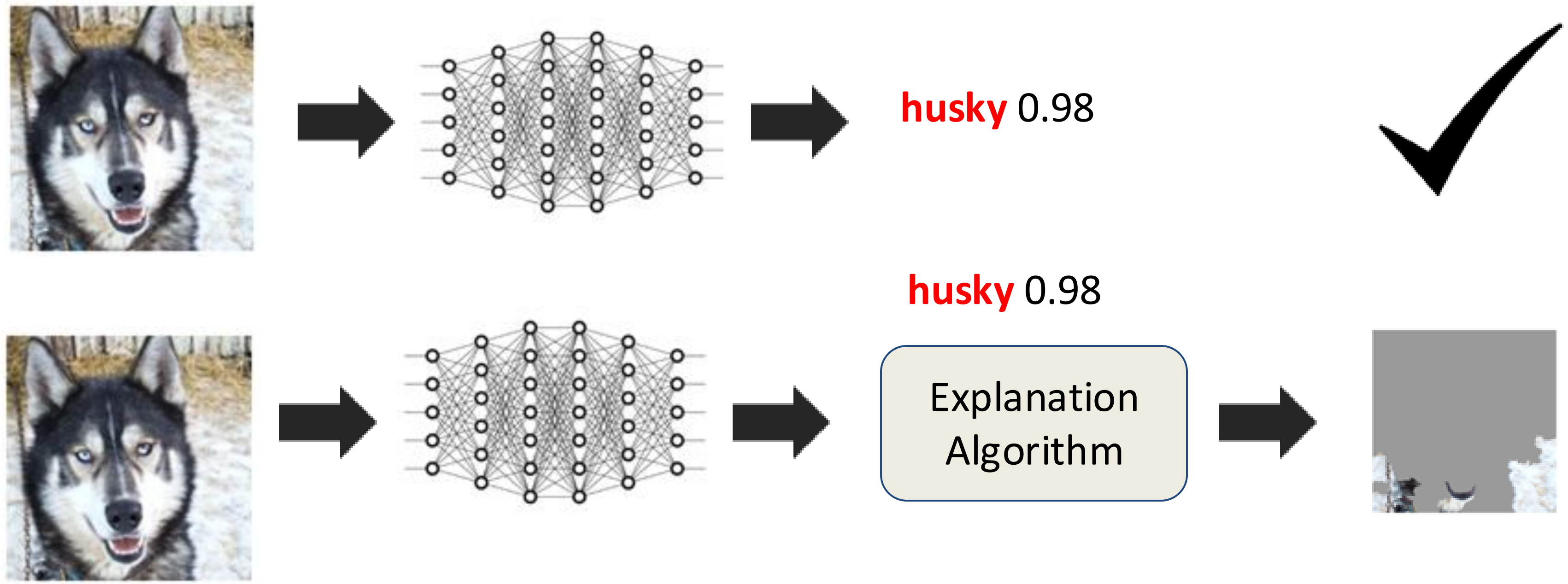


Outline of today’s class

- **What is trustworthy AI?**
- **Explaining AI predictions**
- Definitions of fairness in AI
- Framework for fair AI
- Algorithmic fairness criteria
    - Individual fairness
    - Group fairness

Explainable AI

“Explainable AI refers to the set of approaches that provide an interpretable description of the behavior of a given (complex) model to end users.”



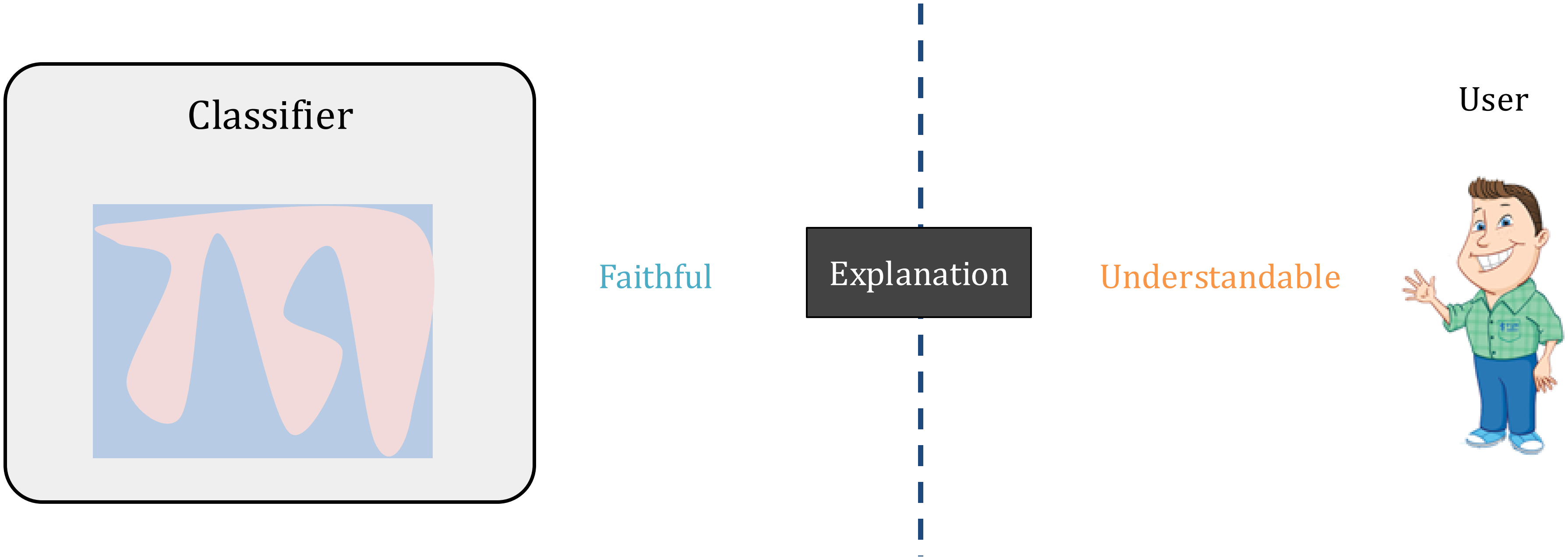
husky 0.98

husky 0.98

Explanation Algorithm

What is an explanation?

- **Definition:** Interpretable description of the model behavior



Classifier

Faithful

Explanation

Understandable

User

Overview of explanation methods

Local explanations

- Explain individual predictions
- Help unearth biases in the *local neighborhood* of a given instance
- Help vet if individual predictions are being made for the right reasons

Global explanations

- Explain complete behavior of the model
- Sheds light on *big picture biases* affecting larger subgroups
- Help vet if the model, at a high level, is suitable for deployment

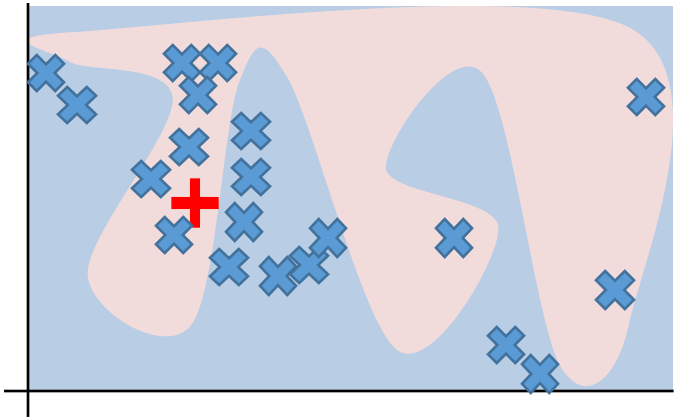
Overview of explanation methods

- Local explanation methods:
  - Feature importance scoring
  - Integrated gradients
  - Prototype explanations
  - Counterfactuals

- Global explanation methods:
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

LIME: Local interpretable model-agnostic explanations

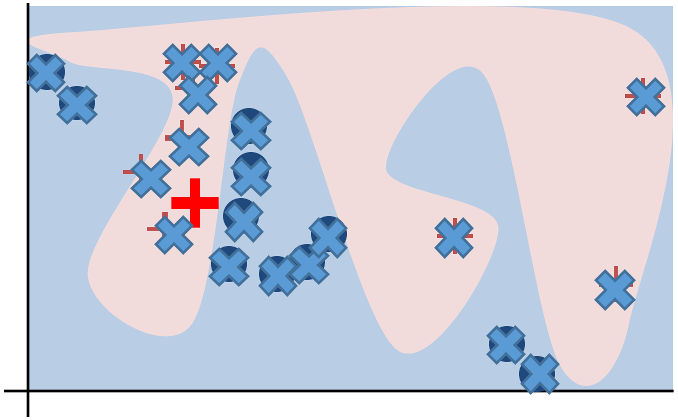
1. Sample points around x_i



[Ribeiro et al. 2016]

LIME: Local interpretable model-agnostic explanations

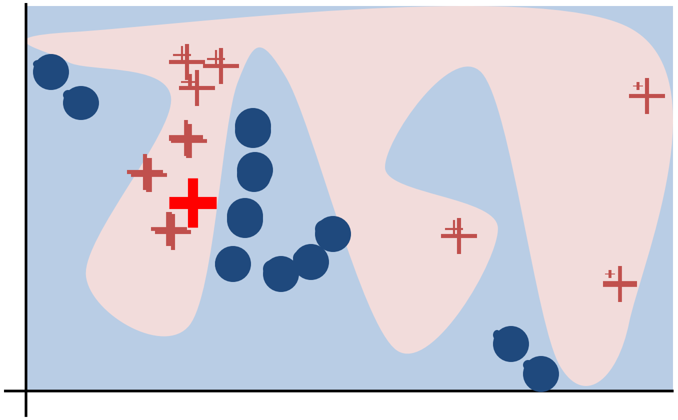
1. Sample points around x_i
2. Use model to predict labels for each sample



[Ribeiro et al. 2016]

LIME: Local interpretable model-agnostic explanations

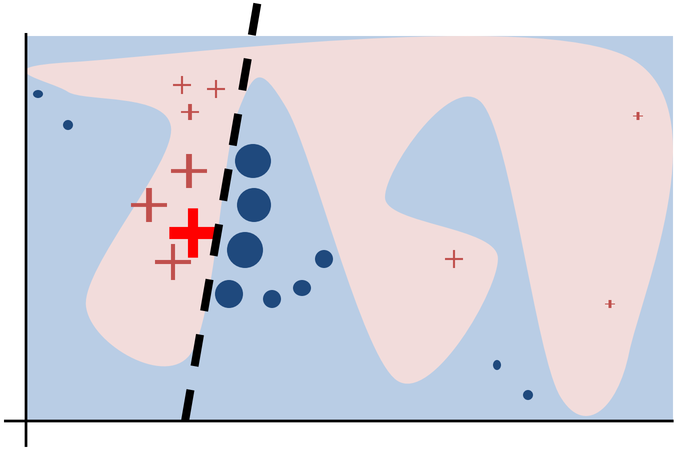
1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i



[Ribeiro et al. 2016]

LIME: Local interpretable model-agnostic explanations

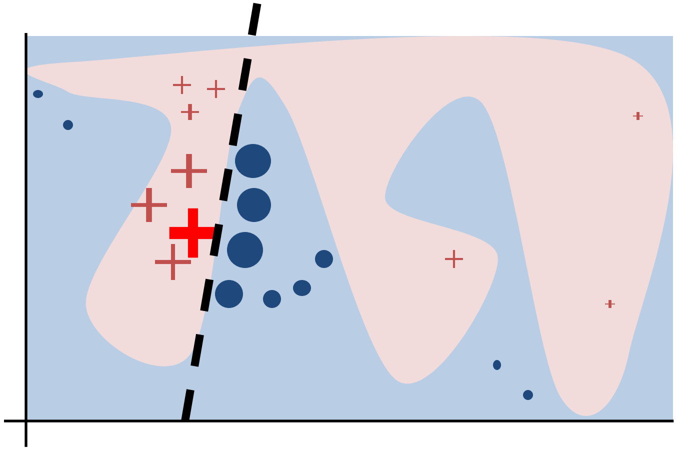
1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn simple linear model on weighted samples



[Ribeiro et al. 2016]

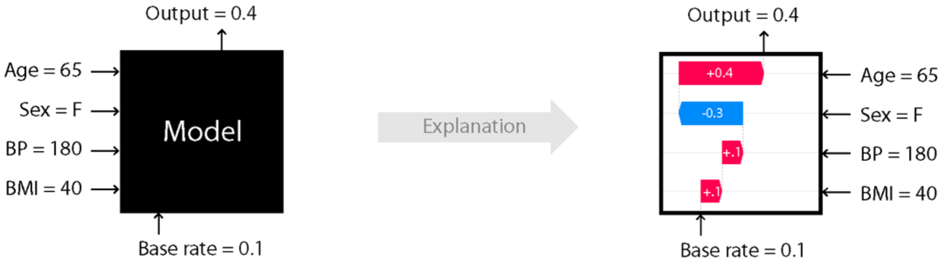
LIME: Local interpretable model-agnostic explanations

1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn simple linear model on weighted samples
5. Use simple linear model to explain x_i



Another popular method which outputs feature importance scores: SHAP

SHAP values are based on game theory and assign an importance value to each feature in a model. Features with positive SHAP values positively impact the prediction, while those with negative values have a negative impact. The magnitude is a measure of how strong the effect is



[Ribeiro et al. 2016; Lundberg & Su-In Lee 2017]

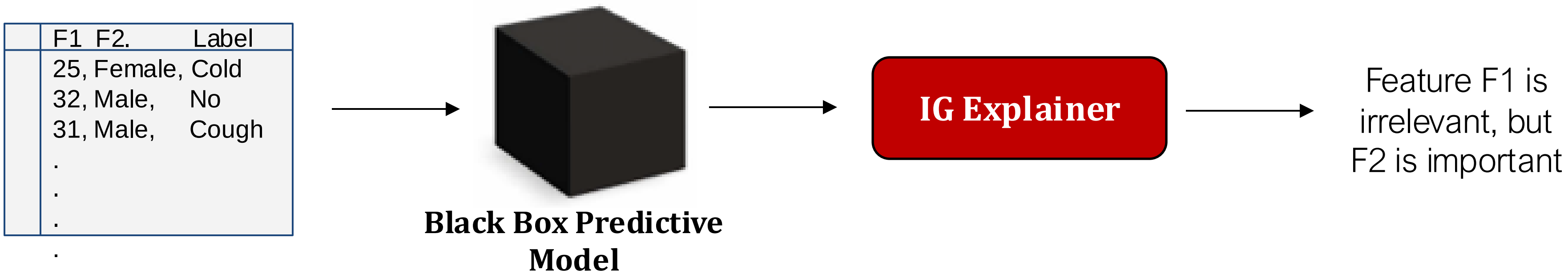
Overview of explanation methods

- Local explanation methods:
  - Feature importance scoring
  - Integrated gradients
  - Prototype explanations
  - Counterfactuals

- Global explanation methods:
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

Integrated Gradients (IG)

- Integrated Gradients (IG) is an explanation method for deep neural networks
- It identifies important features that contribute most to the model's prediction



F1	F2.	Label
25,	Female,	Cold
32,	Male,	No
31,	Male,	Cough
.		
.		
.		
.		

Black Box Predictive Model
IG Explainer
Feature F1 is irrelevant, but F2 is important

- Appealing properties of integrated gradients:
  - It can be applied to any differentiable model like models for images, text, or structured data
  - It requires no modification to the original ML model

Axiomatic Attribution for Deep Networks, *ICML 2017*

https://distill.pub/2020/attribution-baselines

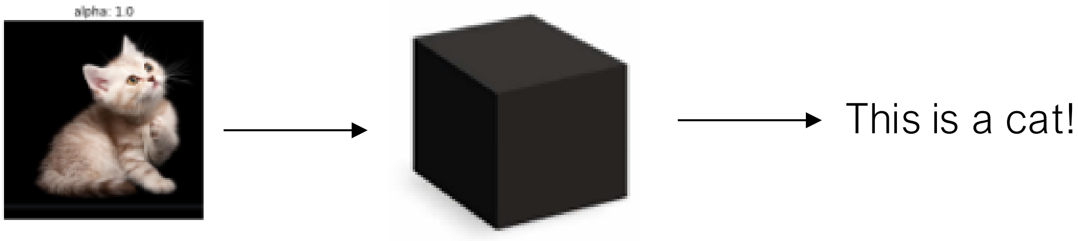
How does IG work?

- IG computes gradients of the model’s prediction w.r.t. input features
- IG is built on two axioms which need to be satisfied:
  - Sensitivity and
  - Implementation invariance
- Sensitivity:
  - We establish a baseline instance as a starting point
  - We then build a sequence of instances which we interpolate from a baseline instance to the actual instance to calculate
- Implementation invariance:
  - Implementation invariance is satisfied when two functionally equivalent models have identical attributions for the same input image and the baseline image.
  - Two models are functionally equivalent when their outputs are equal for all inputs despite having very different implementations

Axiomatic Attribution for Deep Networks, *ICML 2017*

Calculating and visualizing IG

- **Setup:**
  - Let’s consider an ML model for image classification
  - We aim to use IG to explain the predicted image label



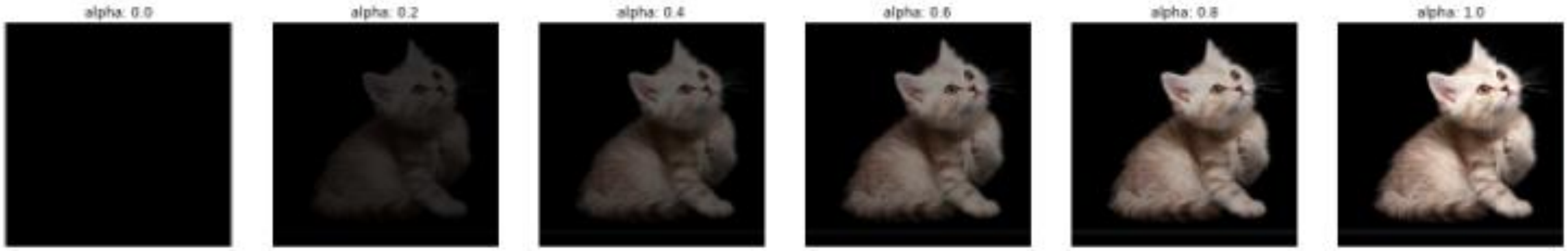
This is a cat!

- **Step 1:**
  - Start from a baseline where the baseline can be a black image whose pixel values are all zero or an all-white image, or a random image
  - Baseline input is one where the prediction is neutral and is central to any explanation method and visualizing pixel feature importance scores

Axiomatic Attribution for Deep Networks, *ICML 2017*

Calculating and visualizing IG

- Step 2:
  - Generate a linear interpolation between the baseline and the original image
  - Interpolated images are small steps(α) in the feature space between your baseline and input image and consistently increase with each interpolated image's intensity



Axiomatic Attribution for Deep Networks, *ICML 2017*

Calculating and visualizing IG

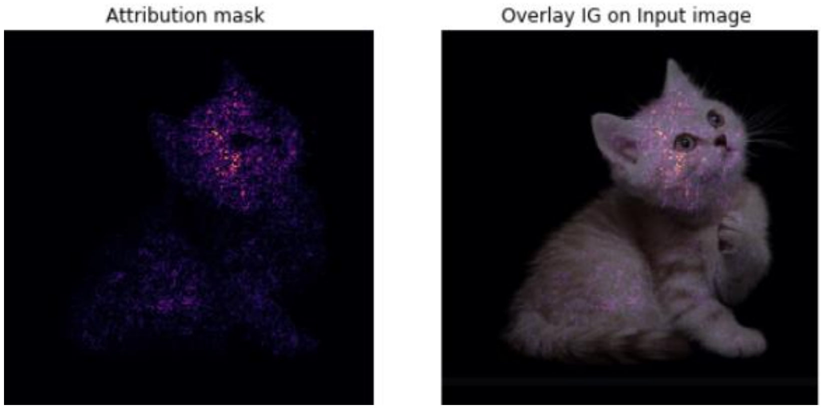
- Step 3: Calculate gradients to measure the relationship between changes to a feature and changes in the model's predictions
- The gradient informs which pixel has the strongest effect on the models predicted class probabilities
  - Varying variable changes the output, and the variable will receive some attribution to help calculate the feature importances for the input image
  - Variable that does not affect the output gets no attribution
- Step 4: Compute the numerical approximation through averaging gradients (that's why the method's name is integrated gradients)

Axiomatic Attribution for Deep Networks, *ICML 2017*

Calculating and visualizing IG

- Step 5:
  - Scale IG to the input image to ensure that the attribution values are accumulated across multiple interpolated images are all in the same units
  - Represent the IG on the input image with the pixel importances

IG helps us explain what an ML model looks at to make a prediction by highlighting the feature importances.
It does this by computing the gradient of the model's prediction output to its input features.



Axiomatic Attribution for Deep Networks, *ICML 2017*

Overview of explanation methods

- **Local explanation methods:**
  - Feature importance scoring
  - Integrated gradients
  - Prototype explanations
  - Counterfactuals

- **Global explanation methods:**
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

Prototype-based explanations

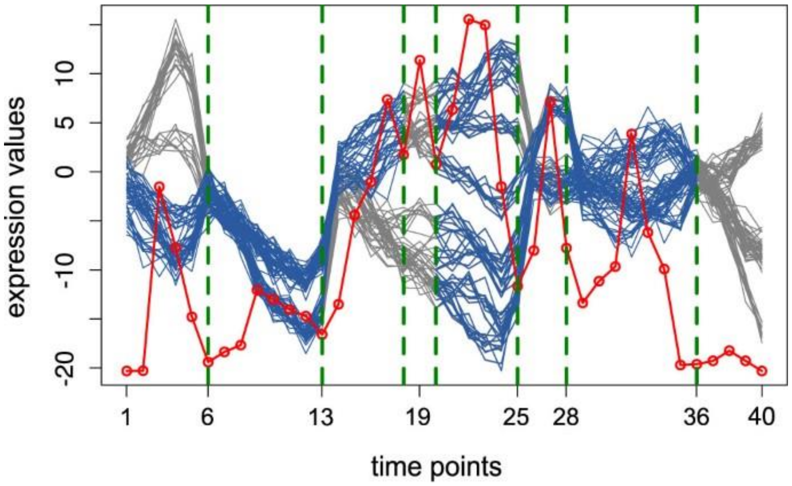
- Use examples (synthetic or natural) to explain individual predictions

- Influence Functions (Koh & Liang 2017)
  - Identify instances in the training set that are responsible for the prediction of a given test instance

- Activation Maximization (Erhan et al. 2009)
  - Identify examples (synthetic or natural) that strongly activate a function (neuron) of interest

Prototypes for explaining time series models

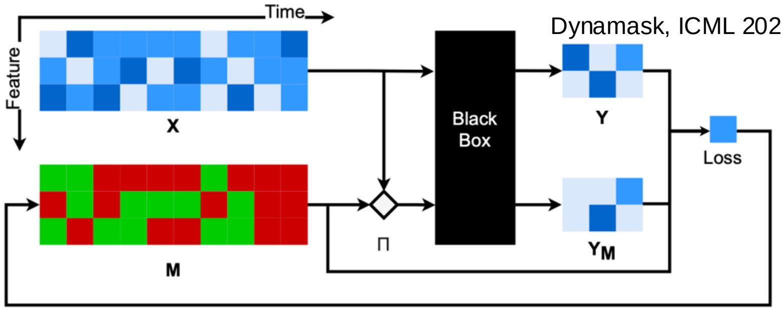
- Time series are not easily visually interpretable
  - Noisy samples
  - Dense informative features, unlike imaging and text modalities
- **Temporal patterns**
  - Only show up when looking at time segments and long-term behaviors
- Perturbations matter
  - Setting a value to zero does not ignore that time point
  - Temporal dependencies cannot be ignored



Omranian et al., 2015

Existing time series explainers are inadequate

1. Perturbations are continuous
   - Can deform shape of samples
2. Give only instance-based explanations
   - Cannot relate patterns across samples
3. Fail to match performance of generic explainers
   - Post-hoc methods suffer from a lack of faithfulness and stability



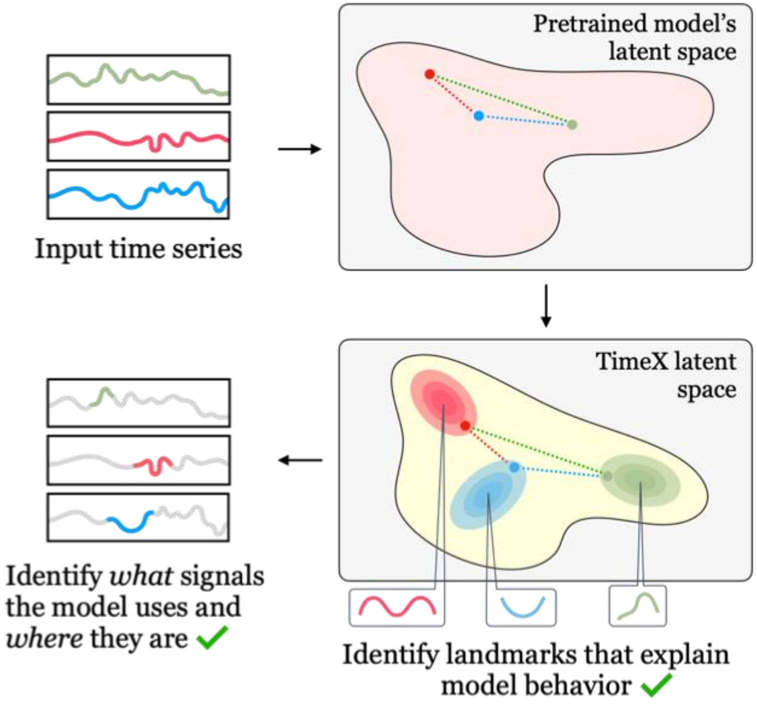
Dynamask, ICML 2021

Desiderata for time series explanations

- Temporally connected and visually digestible
- Identify the location of predictive time series signals and underlying interpretable patterns
- Connect explanations across samples

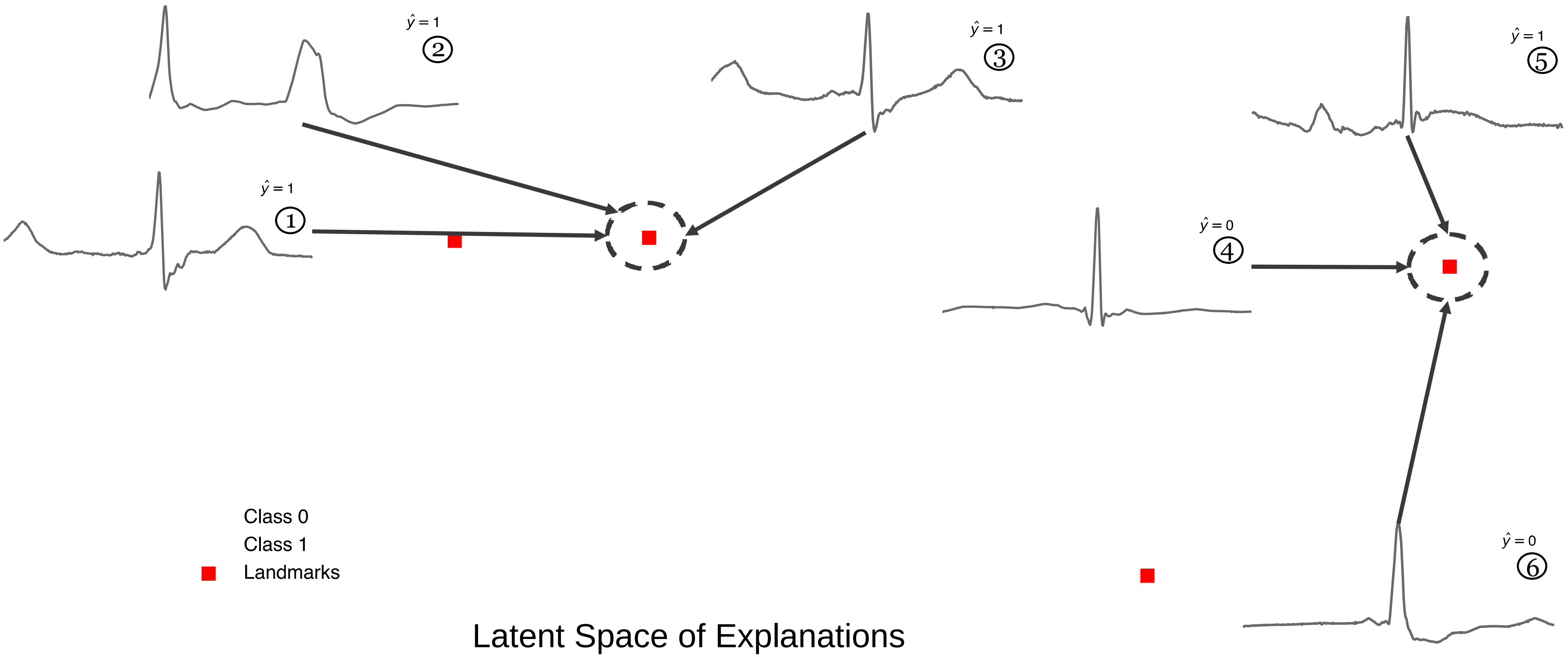
TimeX is a time-series consistency explainer

- Surrogate model to **mimic the behavior of a pretrained time series model**
- TimeX makes inferences on masked samples
- **Model behavior consistency**
  - Enforces faithfulness at the level of the latent space
  - Learns a flexible latent space of explanations


Input time series
Pretrained model's latent space
TimeX latent space
Identify *what* signals the model uses and *where* they are ✓
Identify landmarks that explain model behavior ✓

Encoding Time-Series Explanations through Self-Supervised Model Behavior Consistency, NeurIPS 2023

Learned landmarks represent important patterns in physiological time series



Landmarks partition the latent space of explanations into interpretable temporal patterns

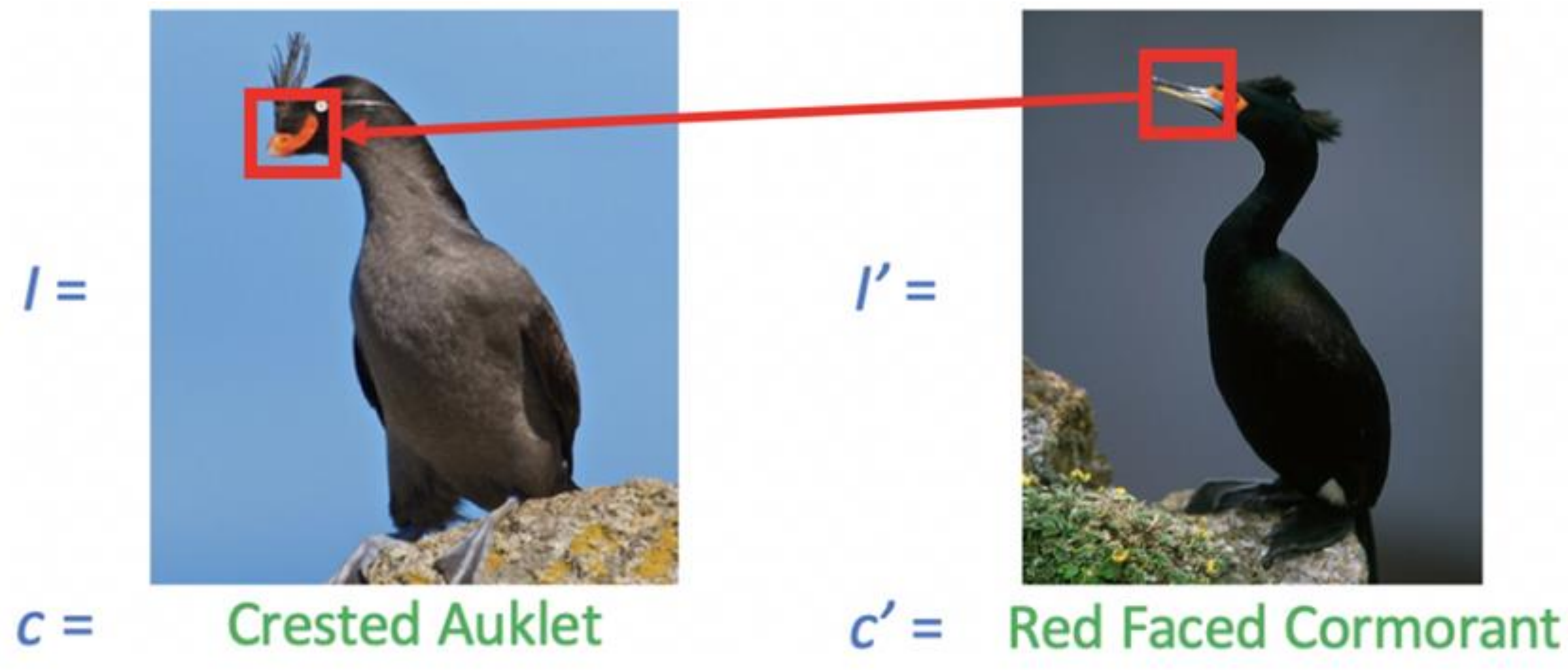
Overview of explanation methods

- Local explanation methods:
  - Feature importance scoring
  - Integrated gradients
  - Prototype explanations
  - Counterfactuals

- Global explanation methods:
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

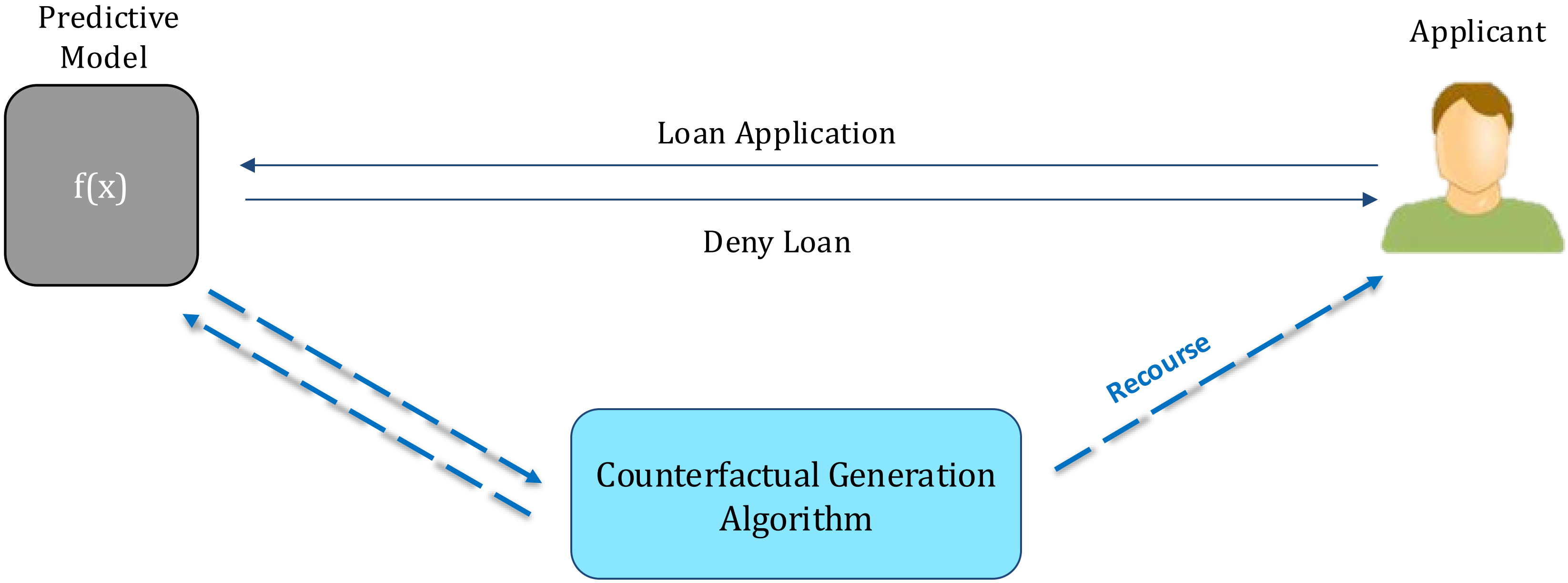
Counterfactual explanations

What features need to be changed and by how much to flip a model's prediction?



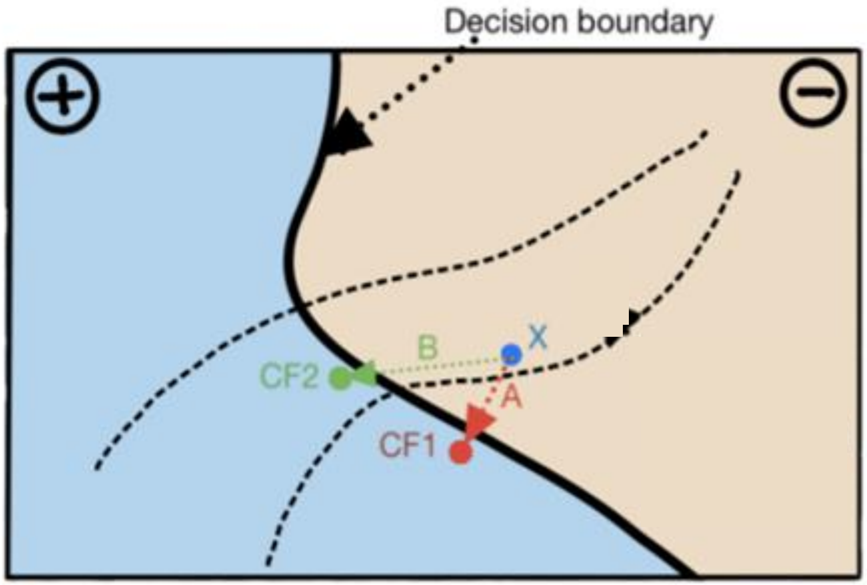
[Goyal et. al., 2019]

Counterfactual explanations



Recourse: Increase your salary by 50K & pay your credit card bills on time for next 3 months

Generating counterfactual explanations: Intuition



Proposed solutions differ on:

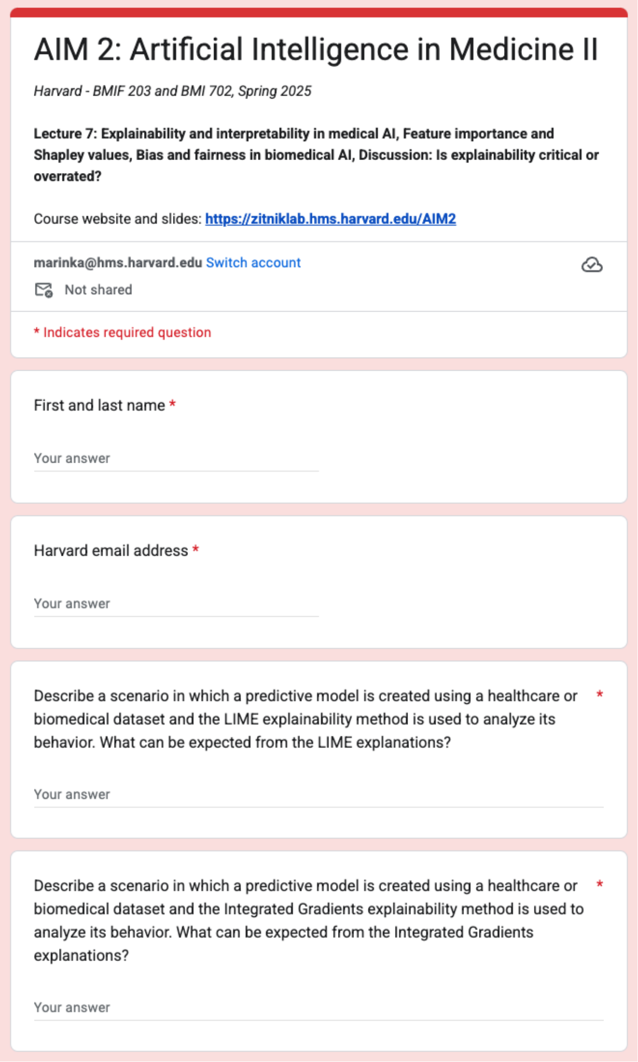
1. How to choose among candidate counterfactuals?

1. How much access is needed to the underlying predictive model?

[Verma et. al., 2020]

Quick Check

https://forms.gle/r7B2PKemuzpG1uSC7



AIM 2: Artificial Intelligence in Medicine II

Harvard - BMIF 203 and BMI 702, Spring 2025

Lecture 7: Explainability and interpretability in medical AI, Feature importance and Shapley values, Bias and fairness in biomedical AI, Discussion: Is explainability critical or overrated?

Course website and slides: https://zitniklab.hms.harvard.edu/AIM2

marinka@hms.harvard.edu Switch account

Not shared

* Indicates required question

First and last name *

Your answer

Harvard email address *

Your answer

Describe a scenario in which a predictive model is created using a healthcare or biomedical dataset and the LIME explainability method is used to analyze its behavior. What can be expected from the LIME explanations? *

Your answer

Describe a scenario in which a predictive model is created using a healthcare or biomedical dataset and the Integrated Gradients explainability method is used to analyze its behavior. What can be expected from the Integrated Gradients explanations? *

Your answer

Overview of explanation methods

- Local explanation methods:
  - Feature importance scoring
  - Integrated gradients
  - Prototype explanations
  - Counterfactuals

- Global explanation methods:
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

Global explanations from local feature importances: SP-LIME

LIME explains a single prediction

local behavior for a single instance

Can’t examine all explanations

Instead pick *k* explanations to show to the user

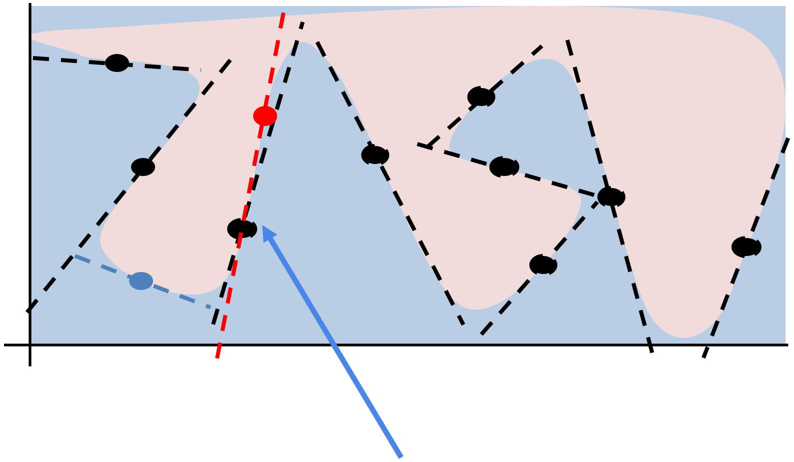
Representative

Should summarize the model’s global behavior

Diverse

Should not be redundant in their descriptions

SP-LIME uses submodular optimization and *greedily* picks k explanations



Single explanation

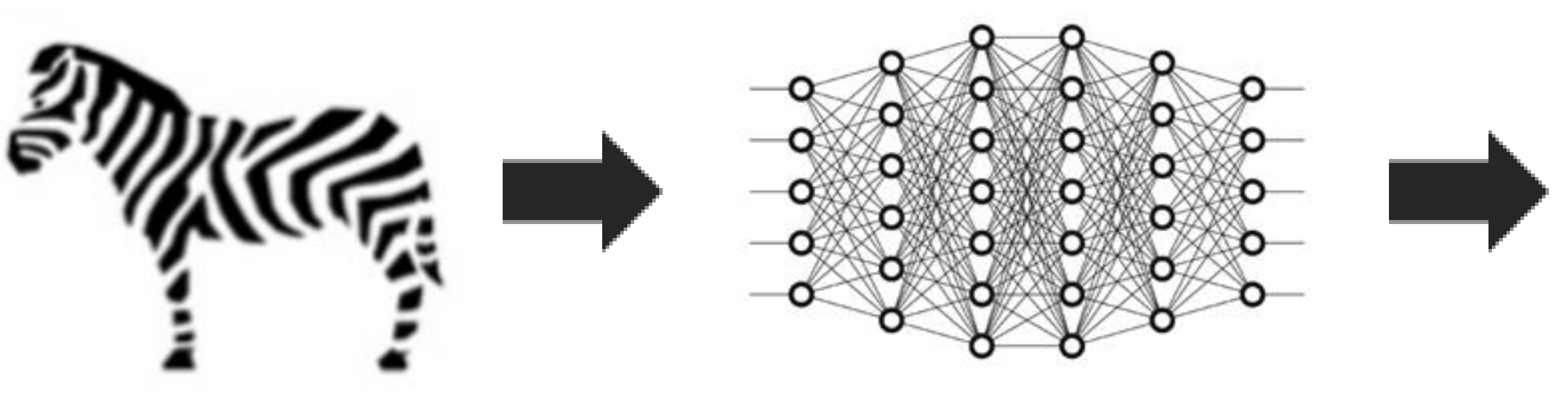
[Ribeiro et. al., 2016]

Overview of explanation methods

- **Local explanation methods:**
  - Feature importance scoring
  - Integrated gradients
  - Prototype explanations
  - Counterfactuals

- **Global explanation methods:**
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

Representation-based explanations



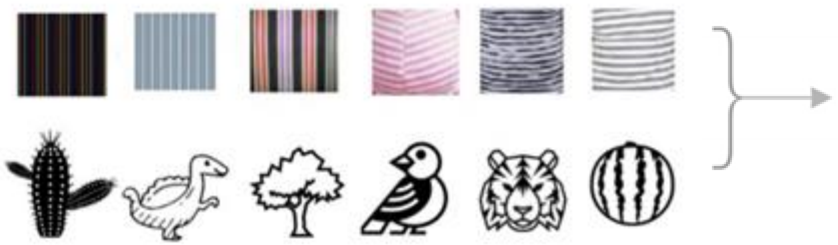
Zebra
(0.97)

How important is the notion of “stripes” for this prediction?

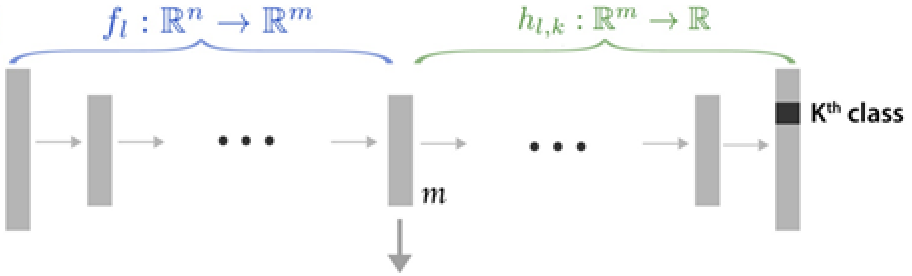
[Kim et. al., 2018]

Representation-based explanations: TCAV approach

Examples of the concept “stripes”



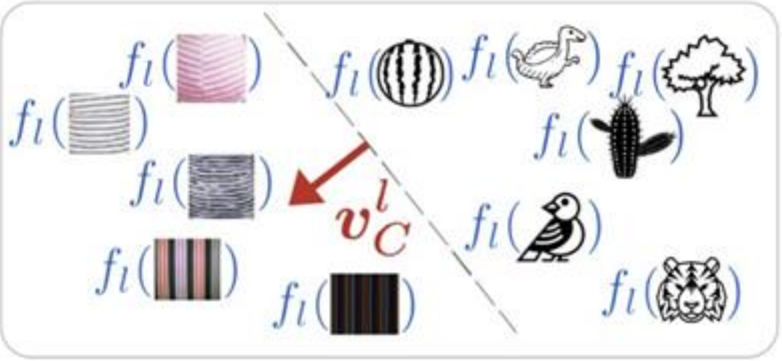
Random examples


$f_l : \mathbb{R}^n \rightarrow \mathbb{R}^m$
$h_{l,k} : \mathbb{R}^m \rightarrow \mathbb{R}$
Kth class
m

Train a linear classifier to separate activations

The vector orthogonal to the decision boundary denotes the concept “stripes”

Compute gradient w.r.t. this vector to determine how important is the notion of stripes for a prediction



[Kim et. al., 2018]

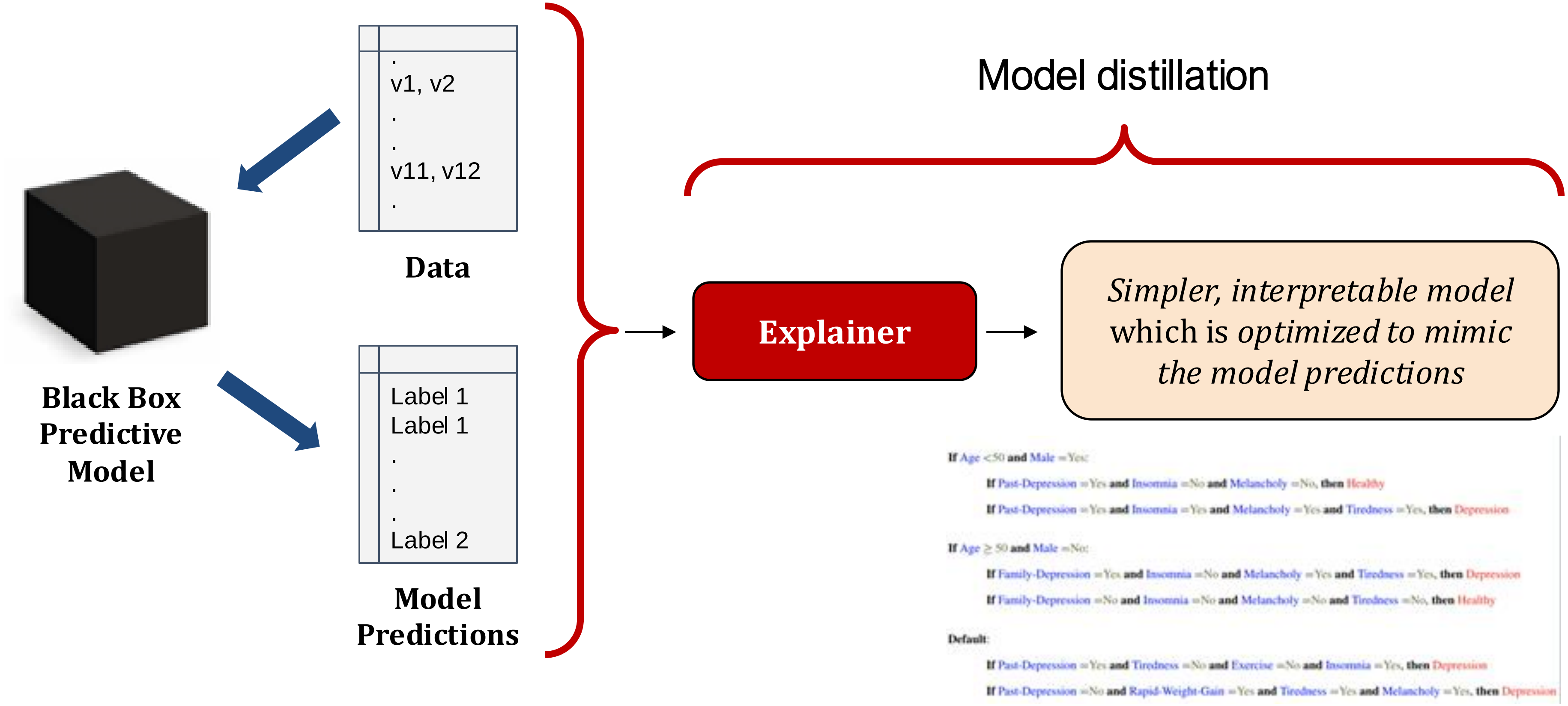
TCAV = testing with concept activation vectors

Overview of explanation methods

- Local explanation methods:
  - Feature importance scoring
  - Integrated gradients
  - Prototypes/Example-based explanations
  - Counterfactuals

- Global explanation methods:
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

Model distillation


.
v1, v2
.
.
v11, v12
.
Data
Black Box Predictive Model
Label 1
Label 1
.
.
.
Label 2
Model Predictions
Model distillation
Explainer
Simpler, interpretable model which is *optimized to mimic the model predictions*

If Age <50 **and** Male =Yes:

- **If** Past-Depression =Yes **and** Insomnia =No **and** Melancholy =No, **then** Healthy
- **If** Past-Depression =Yes **and** Insomnia =Yes **and** Melancholy =Yes **and** Tiredness =Yes, **then** Depression

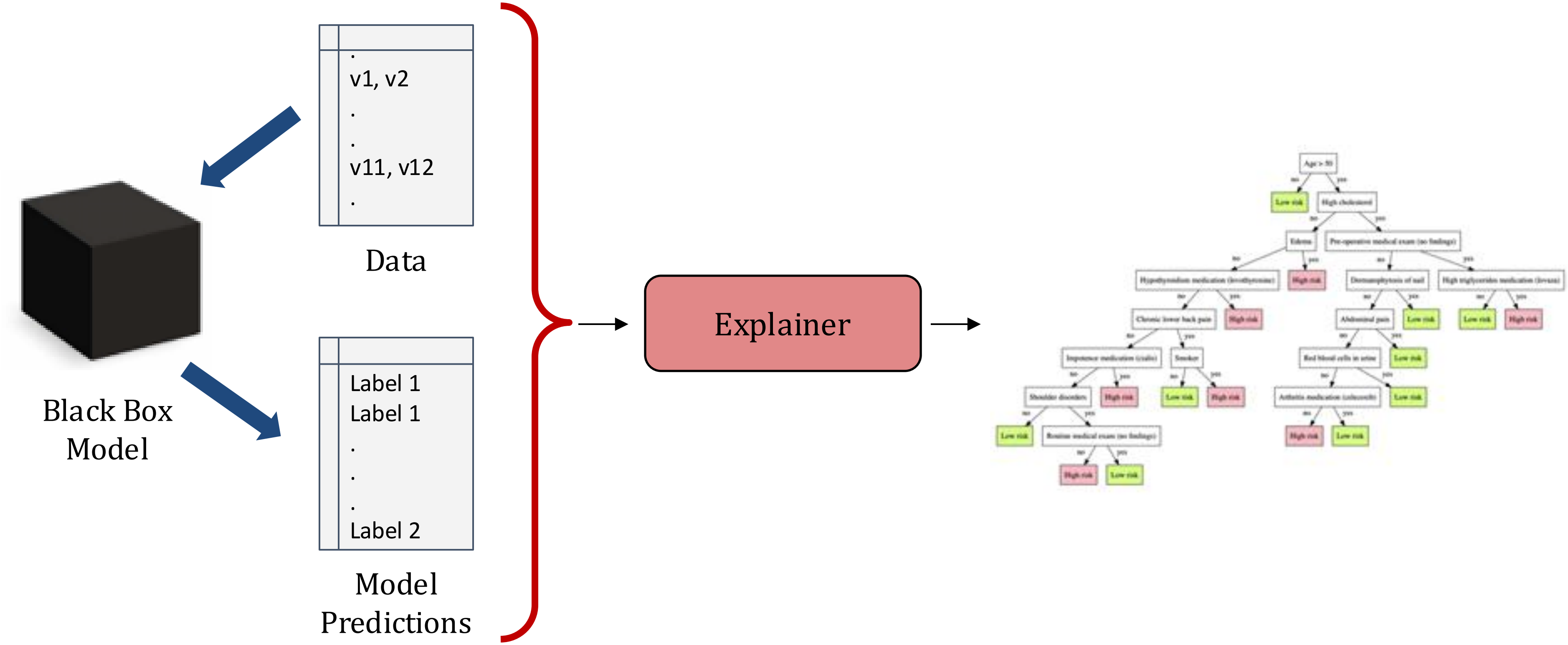
If Age ≥ 50 **and** Male =No:

- **If** Family-Depression =Yes **and** Insomnia =No **and** Melancholy =Yes **and** Tiredness =Yes, **then** Depression
- **If** Family-Depression =No **and** Insomnia =No **and** Melancholy =No **and** Tiredness =No, **then** Healthy

Default:

- **If** Past-Depression =Yes **and** Tiredness =No **and** Exercise =No **and** Insomnia =Yes, **then** Depression
- **If** Past-Depression =No **and** Rapid-Weight-Gain =Yes **and** Tiredness =Yes **and** Melancholy =Yes, **then** Depression

Model distillation using decision trees



Black Box Model

.

v1, v2

.

.

v11, v12

.

Data

Label 1

Label 1

.

.

.

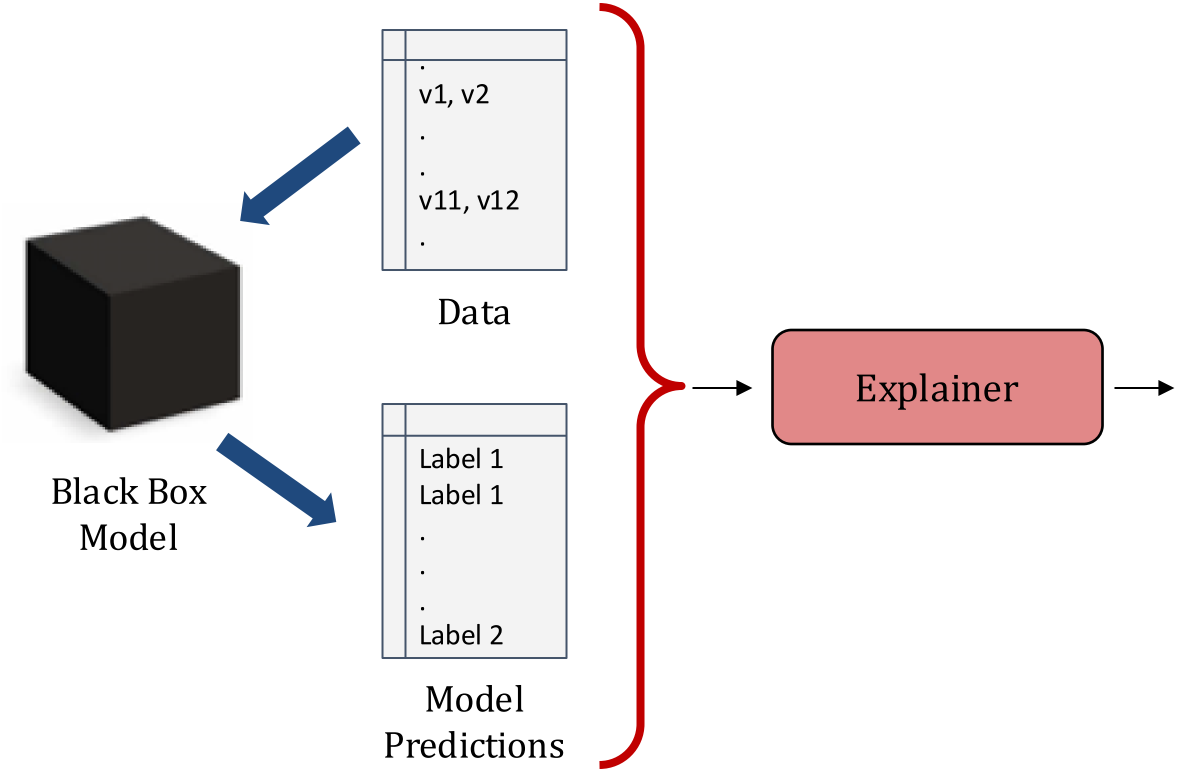
Label 2

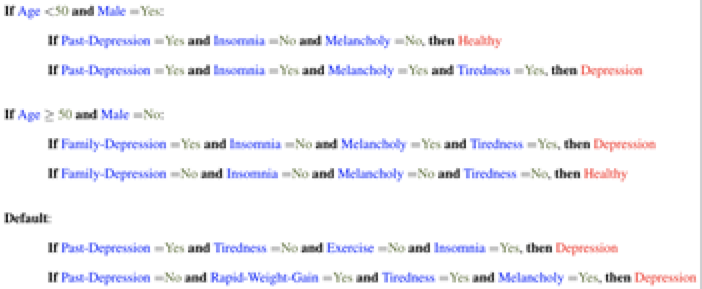
Model Predictions

Explainer

[Bastani et. al., 2019]

Model distillation using decision sets





If Age <50 **and** Male =Yes:

- **If** Past-Depression =Yes **and** Insomnia =No **and** Melancholy =No, **then** Healthy
- **If** Past-Depression =Yes **and** Insomnia =Yes **and** Melancholy =Yes **and** Tiredness =Yes, **then** Depression

If Age ≥ 50 **and** Male =No:

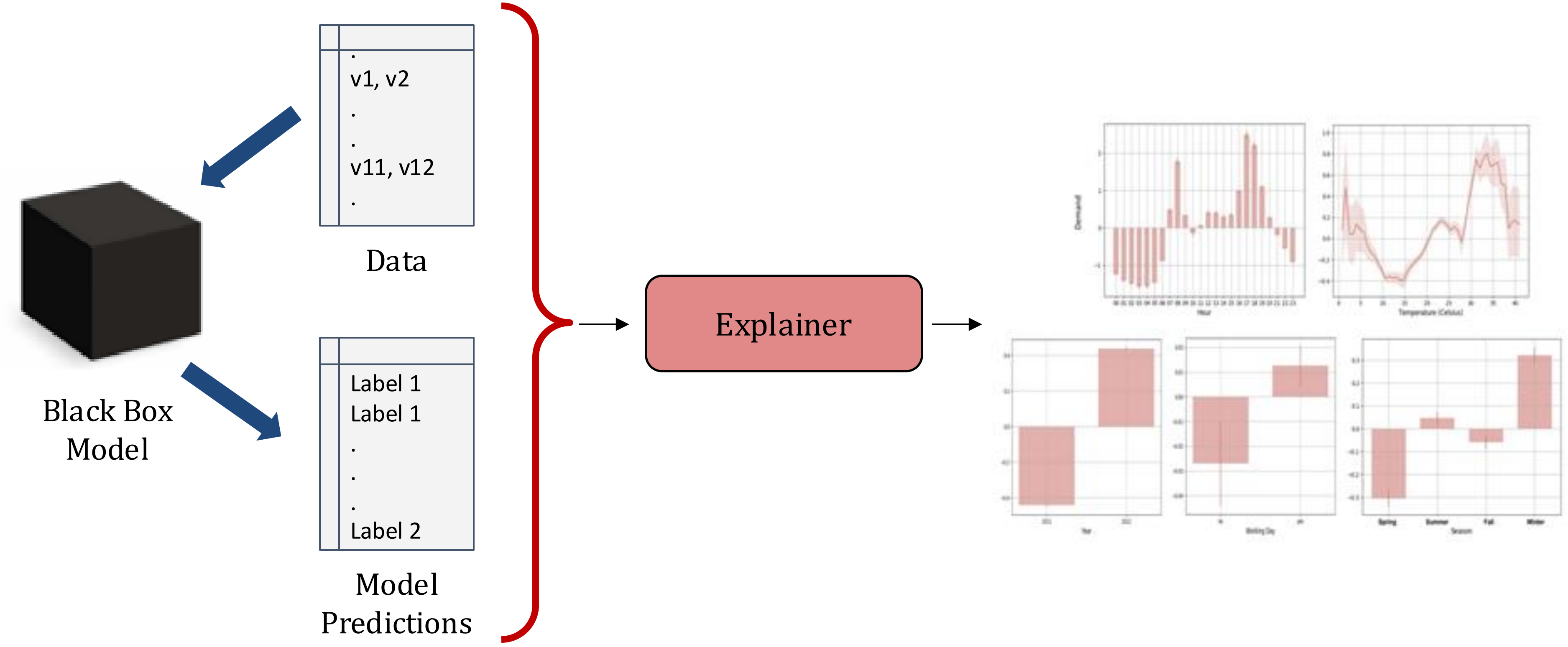
- **If** Family-Depression =Yes **and** Insomnia =No **and** Melancholy =Yes **and** Tiredness =Yes, **then** Depression
- **If** Family-Depression =No **and** Insomnia =No **and** Melancholy =No **and** Tiredness =No, **then** Healthy

Default:

- **If** Past-Depression =Yes **and** Tiredness =No **and** Exercise =No **and** Insomnia =Yes, **then** Depression
- **If** Past-Depression =No **and** Rapid-Weight-Gain =Yes **and** Tiredness =Yes **and** Melancholy =Yes, **then** Depression

[Lakkaraju et. al., 2019]

Model distillation using generalized additive models



.
v1, v2
.
.
v11, v12
.

Data

Black Box Model

Label 1
Label 1
.
.
.
Label 2

Model Predictions

Explainer

[Tan et. al., 2019]

Overview of explanation methods

- Local explanation methods:
  - Feature importance scoring
  - Integrated gradients
  - Prototype explanations
  - Counterfactuals

- Global explanation methods:
  - Collection of local explanations
  - Representation-based explanations
  - Model distillation

Outline of today’s class

- **What is trustworthy AI?**
- **Explaining AI predictions**
- **Definitions of fairness in AI**
- Framework for fair AI
- Algorithmic fairness criteria
  - Individual fairness
  - Group fairness

Adopting AI in high-stakes areas

- Healthcare
- Genomic medicine
- Public health policy
- Child welfare

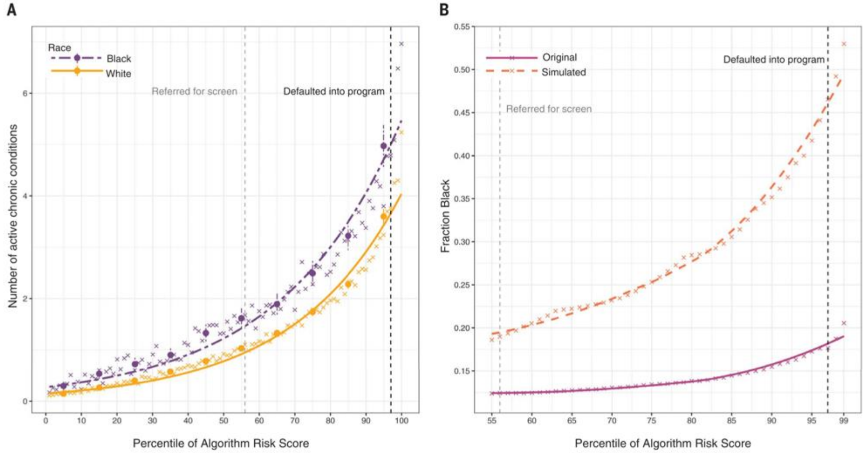


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

Obermeyer et al. *Science* 2019

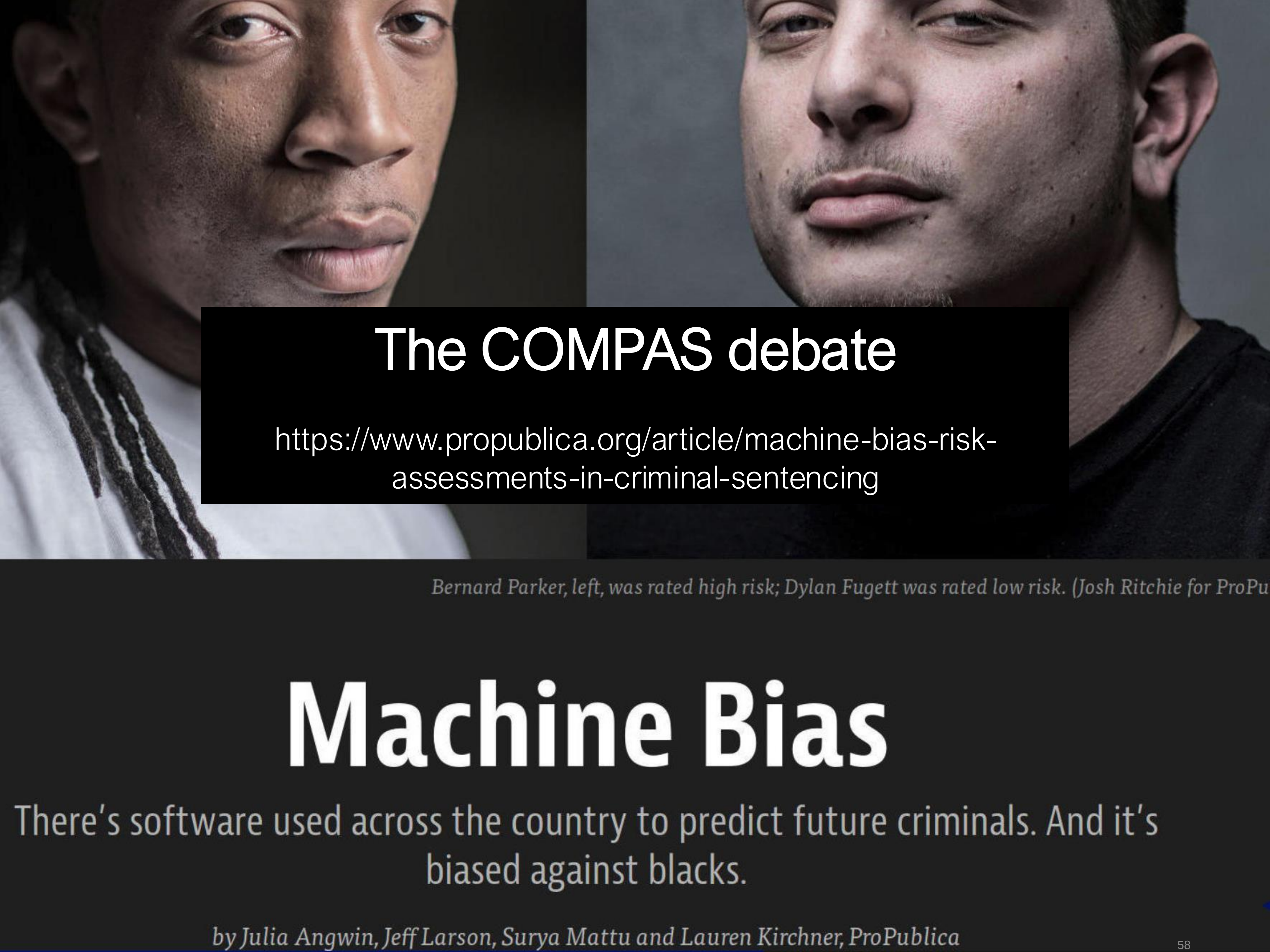
- Criminal risk assessment
- Surveillance
- Financial lending
- Hiring

Three problematic examples

1. High-risk Healthcare Management
   - Commercial prediction models are used by large health systems to identify and help patients with complex health needs.
   - These models can exhibit significant bias: At a given risk score, black patients are considerably sicker than white patients
   - The bias arises because the algorithm predicts health care costs rather than illness
2. Criminal Risk Assessment Tools
   - Defendants are assigned scores that predict the risk of re-committing crimes
   - These scores inform decisions about bail, sentencing, and parole.
   - Some tools have been biased against black defendants
3. Face Recognition Systems
   - Surveillance and self-driving cars
   - Systems can perform poorly for populations that are not well represented in training dataset

The COMPAS debate

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPu

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

COMPAS

- Correctional Offender Management Profiling for Alternative Sanctions
- Used in prisons across country: AZ, CO, DL, KY, LA, OK, VA, WA, WI
- “Evaluation of a defendant’s rehabilitation needs”
- Recidivism = likelihood of criminal to reoffend

COMPAS (continued)

“Our analysis of Northpointe’s tool, called COMPAS, found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk.”

What are protected classes?

- Protected classes in the US:
  - Race
  - Sex
  - Religion
  - National origin
  - Citizenship
  - Pregnancy
  - Disability status
  - Genetic information
- Regulated domains in the US:
  - Credit (Equal Credit Opportunity Act)
  - Education (Civil Rights Act of 1964; Education Amend. of 1972)
  - Employment (Civil Rights Act of 1964)
  - Housing (Fair Housing Act)

Fairness in ML

- It does not necessarily mean being malicious: Bias can occur even when everyone, from data generators to engineers to clinical staff, has the best intentions
- It is not one and done: Just because an algorithm has no bias now does not mean it has no potential bias later
- It is not new: Researchers have raised concerns about it over the last 50 years

- It is defined in many ways, for example, disparate treatment or impact of algorithm
- It can be a culmination of a flawed system
  - Biases in data collection processes
  - Biases in algorithmic design
  - Bias in model implementation/deployment
- It is the vigilance of how technology can amplify/create bias

How to define fairness in ML?

- Fairness through unawareness
- Group fairness
- Calibration
- Error rate balance
- Representational fairness
- Counterfactual fairness
- Individual fairness

Fairness through unawareness

- Idea: Don't record protected attributes, and don't use them in your algorithm
  - Predict risk Y from features X and group S using $P(\hat{Y} = Y|X)$ instead of $P(\hat{Y} = Y|X, S)$
- Pros: Guaranteed to not be making a judgement on protected attribute
- Cons: Other proxies may still be included in a "race-blind" setting, e.g. zip code or conditions

Group fairness

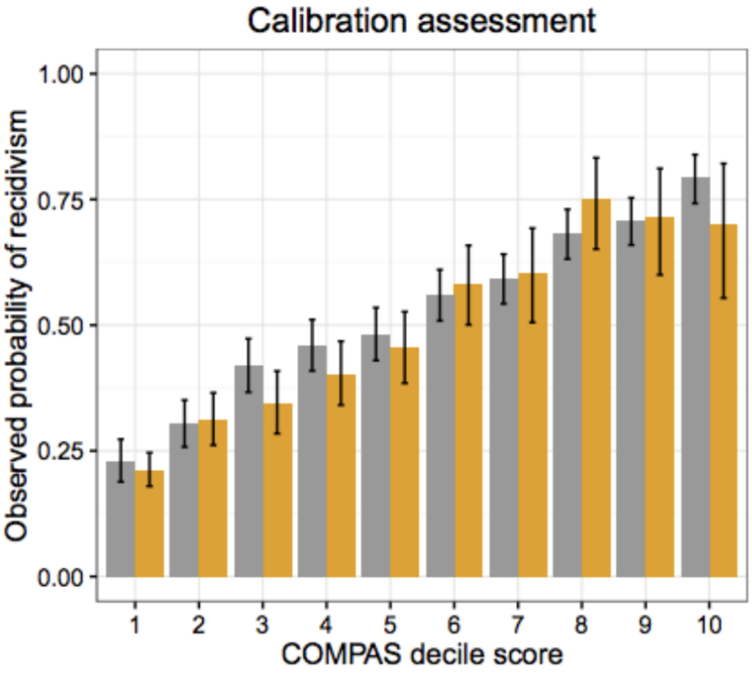
- **Idea:** Require prediction rate be the same across protected groups
  - E.g. "20% of the resources should go to the group that has 20% of population"
- Predict risk Y from features X and group S such that $P(\hat{Y} = 1 | S = 1) = P(\hat{Y} = 1 | S = 0)$
- **Pros:** Literally treats each race equally
- **Cons:**
  - Too strong: Groups might have different base rates. Then, even a perfect classifier wouldn't qualify as "fair"
  - Too weak: Doesn't control error rate. Could be perfectly biased (correct for $S = 0$ and wrong for $S = 1$) and still satisfy

Calibration

- **Idea:** Same positive predictive value across groups
- Predict Y from features X and group S with score R:

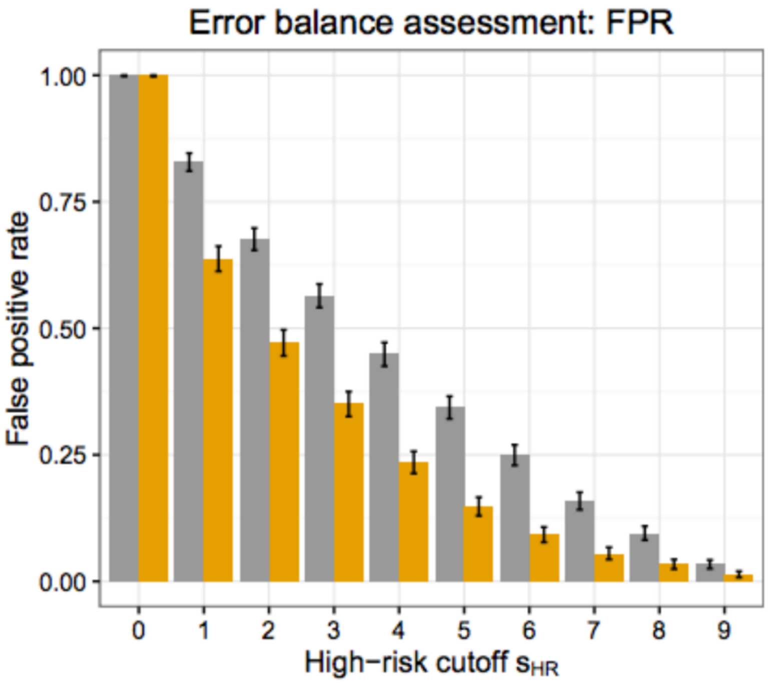
$$P(Y = 1|R = r, A = 1) = P(Y = 1\,|R\ = r, A\ =\ 0)$$

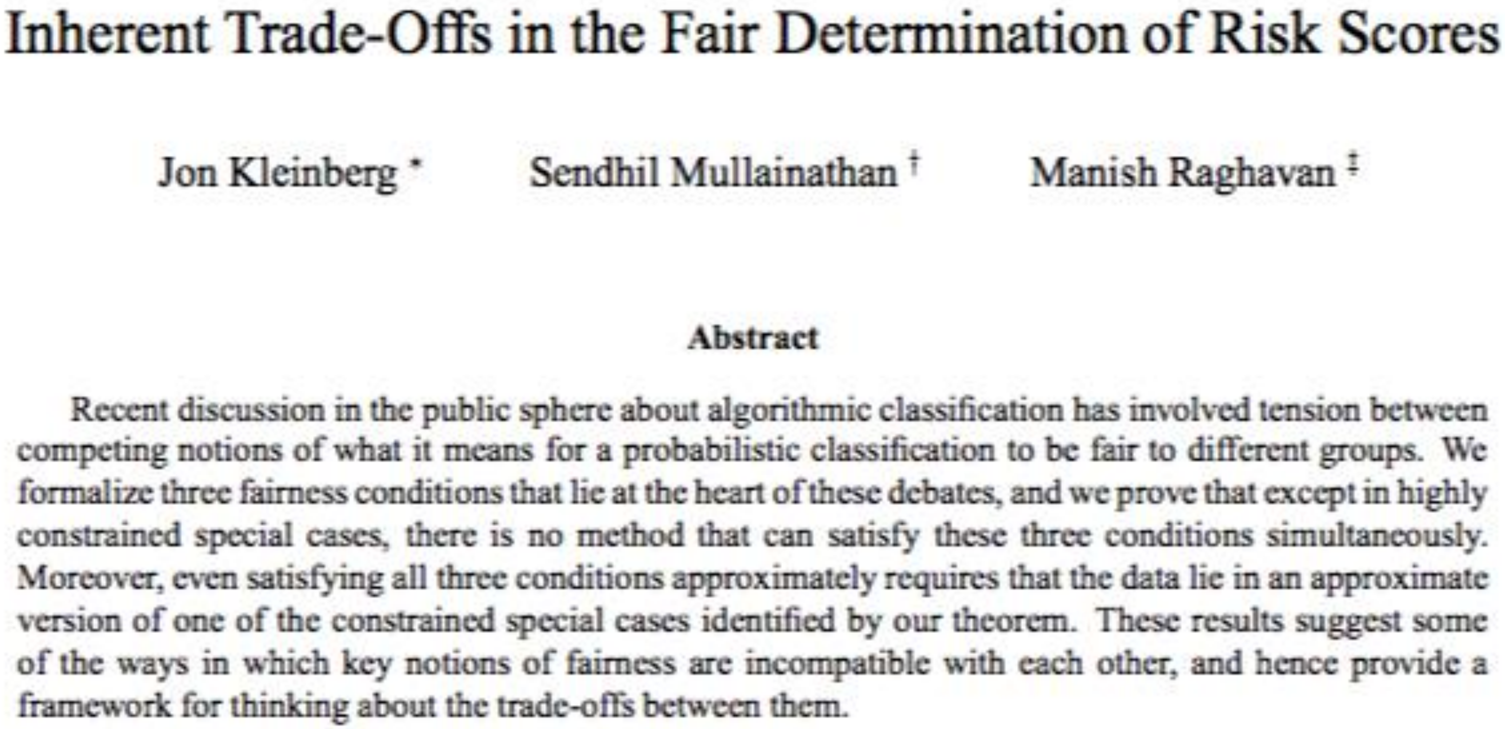
- **Pros:** “Equally right across groups”
- **Cons:** Not compatible with error rate balance (next slide)



Error rate balance

- **Idea:** Equal false positive rates (FPR) across groups
  $$P(\hat{Y} = 1 | Y = 0, S = 1) = P(\hat{Y} = 1 \; Y = 0, S = 0)$$
- **Pros:** “Equally wrong across groups”
- **Cons:** Incompatible with calibration and false negative rates (FNR), could dilute with easy cases



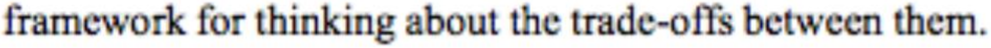


Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg [*] Sendhil Mullainathan [†] Manish Raghavan [‡]

Abstract

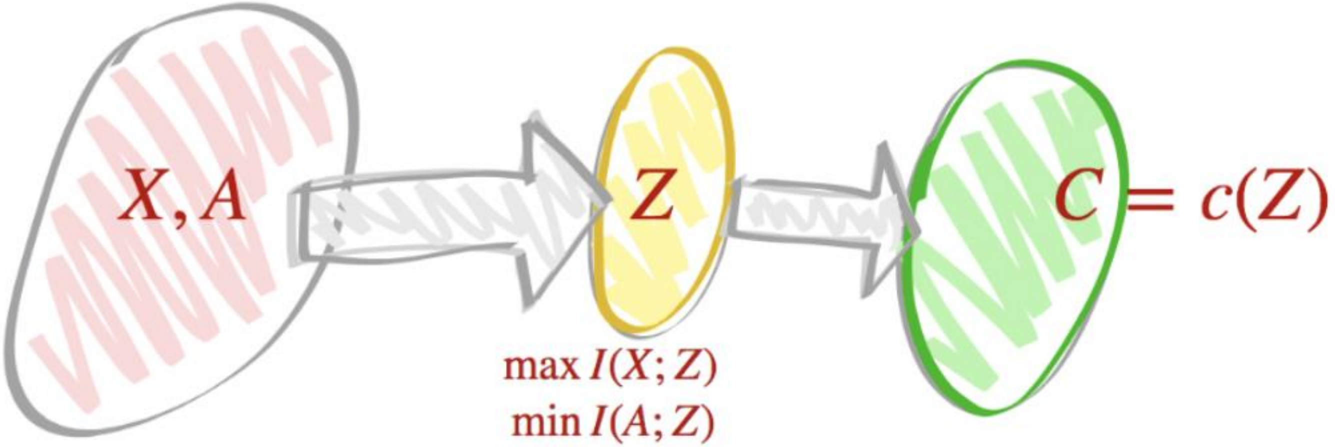
Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.



framework for thinking about the trade-offs between them.

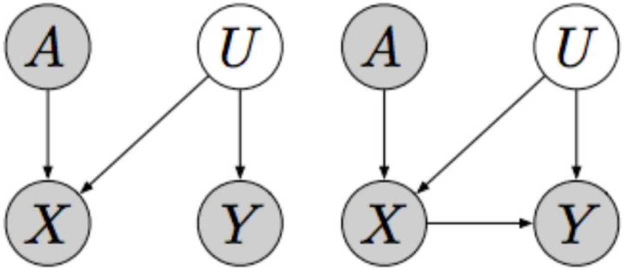
Representational fairness

- Idea: Transform input feature vectors in “fair representations Z to minimize group information
- Pros: Reduce information given to model while still keeping important information
- Cons: Trade-off between accuracy and fairness



Counterfactual fairness

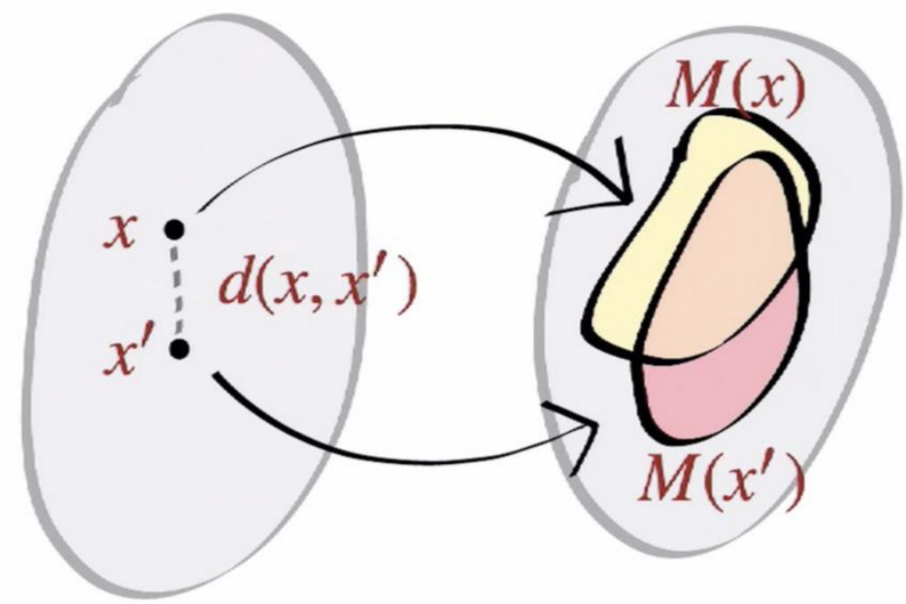
- **Idea:** Group A should not cause prediction $\hat{Y}$
- **Pros:** Can model explicit dependencies between features
- **Cons:**
  - Dependency graphs may not represent real world
  - Inference assumes observed confounders



$$P(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a)$$
$$= P(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a)$$

Individual fairness

- Idea: Similar individuals should be treated similarly
- Pros: Can model heterogeneity within each group
- Cons: Notion of “similar” is hard to define mathematically, especially in high dimensions



How to define "fairness" in ML?

- ~~Fairness through unawareness~~

Not useful

- Group fairness
- Calibration
- Error rate balance

Established strategies

- Representational fairness
- Counterfactual fairness
- Individual fairness

Ongoing and cutting-edge research

One fairness definition or one framework

21 Fairness Definitions and Their Politics. Arvind Narayanan.

ACM Conference on Fairness, Accountability, and Transparency Tutorial (2018)

S. Mitchell, E. Potash, and S. Barocas (2018)
P. Gajane and M. Pechenizkiy (2018)
S. Verma and J. Rubin (2018)

Differences/connections between fairness definitions are difficult to grasp.

We lack common language/framework.

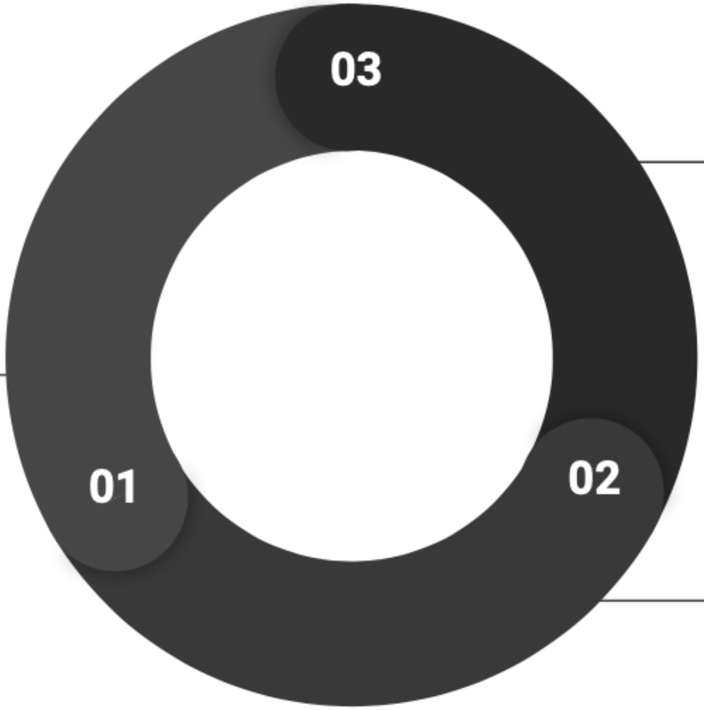
“Nobody has found a definition which is widely agreed as a good definition of fairness in the same way we have for, say, the security of a random number generator.”

“There are a number of definitions and research groups are not on the same page when it comes to the definition of fairness.”

“The search for one true definition is not a fruitful direction, as technical considerations cannot adjudicate moral debates.”

Outline of today’s class

- **What is trustworthy AI?**
- **Explaining AI predictions**
- **Definitions of fairness in AI**
- **Framework for fair AI**
- Algorithmic fairness criteria
  - Individual fairness
  - Group fairness



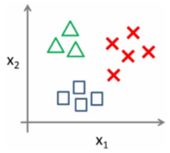
Data Regulator

Determines fairness criteria, determines data source(s), audits results

AUTHORITY

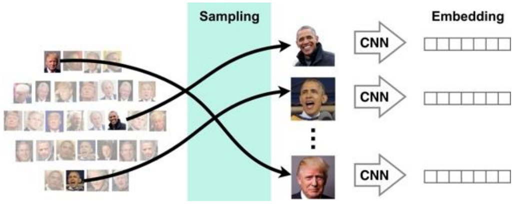
Data User

Computes ML model given sanitized data



Data Producer

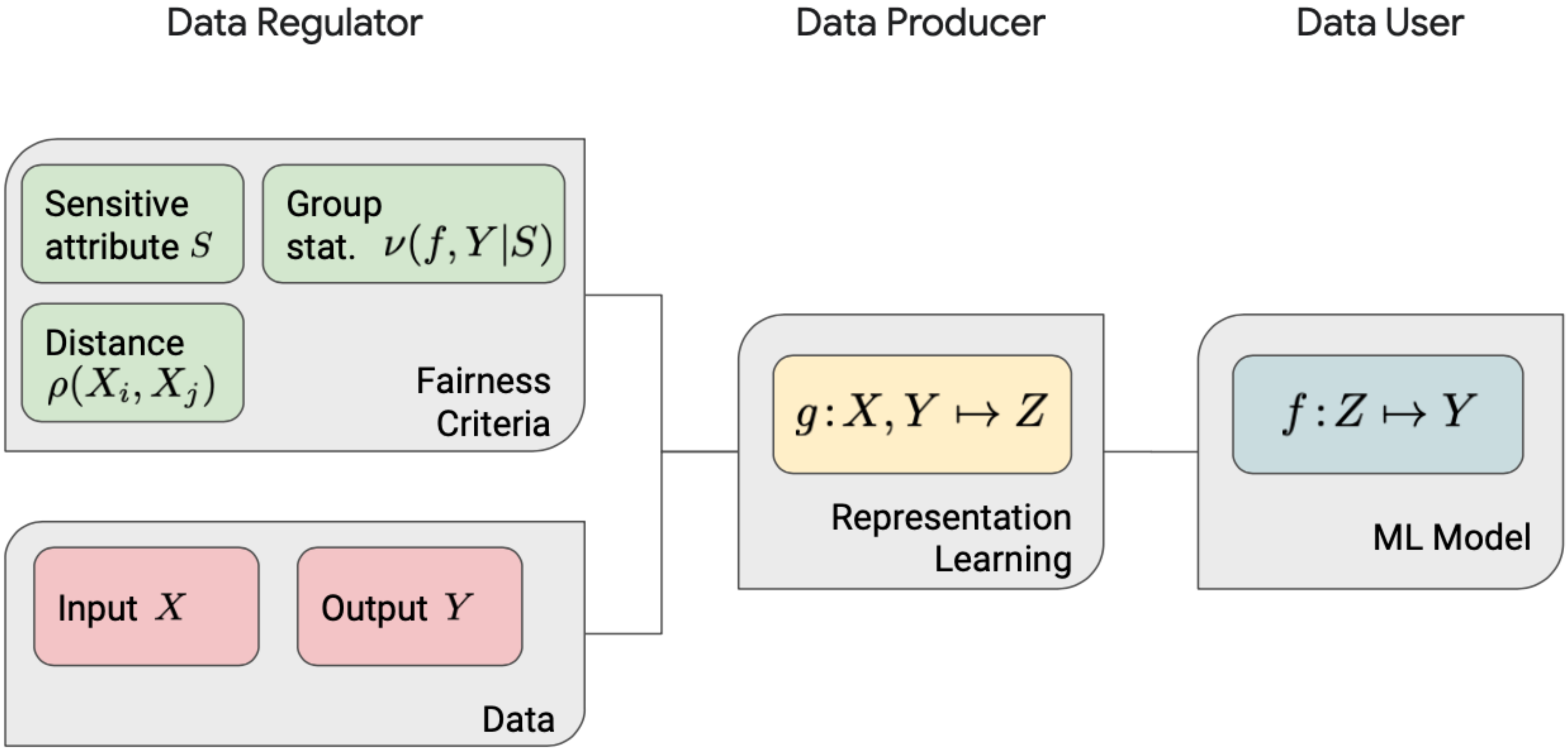
Computes the fair representation given data regulator criteria



McNamara, Ong and Williamson, AIES ‘19

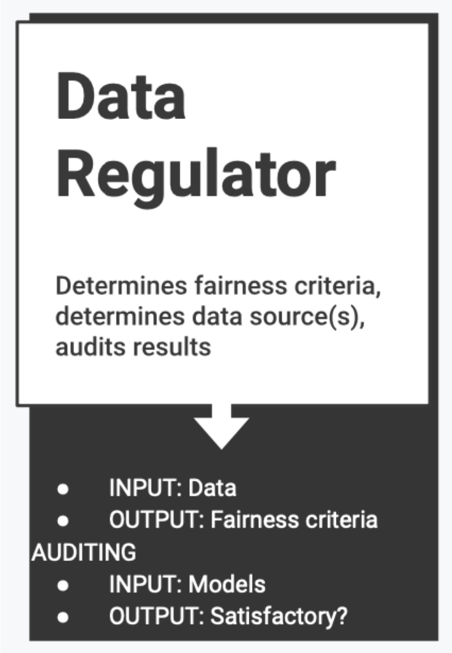
Framework for fair AI/ML

- **Data regulator:** determines fairness measures, audits results
- **Data producer:** creates “fair” feature vectors (i.e., “fair” representations)
- **Data user:** agnostically trains an ML model using “fair” feature vectors



Roles of different parties

- Data regulator determines which fairness criteria to use, and (optionally) audits the results
- When training:
  - Input: interaction with users/experts/judges/policy to determine fairness criteria
  - Output: fairness criteria
- When auditing the ML model:
  - Input (for auditing the data producer):
    - “Fair” representations
  - Input (for auditing the data user):
    - Data and model predictions
  - Output:
    - Are fairness criteria satisfied?



Data Regulator

Determines fairness criteria, determines data source(s), audits results

- INPUT: Data
- OUTPUT: Fairness criteria

AUDITING

- INPUT: Models
- OUTPUT: Satisfactory?

How to achieve fairness?

- **Post-processing:** Post-process the model outputs

  Doherty et al. (2012), Feldman (2015), Hardt et al. (2016), Kusner et al. (2018), Jiang et al. (2019)

- **Pre-processing:** Pre-process the data to remove bias, or extract representations that do not contain sensitive information during training

  Kamiran and Calder (2012), Zemel et al. (2013), Feldman et al. (2015), Fish et al. (2015), Louizos et al. (2016), Lum and Johndrow (2016), Adler et al. (2016), Edwards and Storkey (2016)

- **In-processing:** Enforce fairness notions by imposing constraints into the optimization, or by using an adversary

  Goh et al. (2016), Corbett-Davies et al. (2017), Agarwal et al. (2018), Cotter et al. (2018), Komiyama et al. (2018), Narasimhan (2018), Wu et al. (2018), Zhang et al. (2018), Jiang et al. (2019)

Outline of today’s class

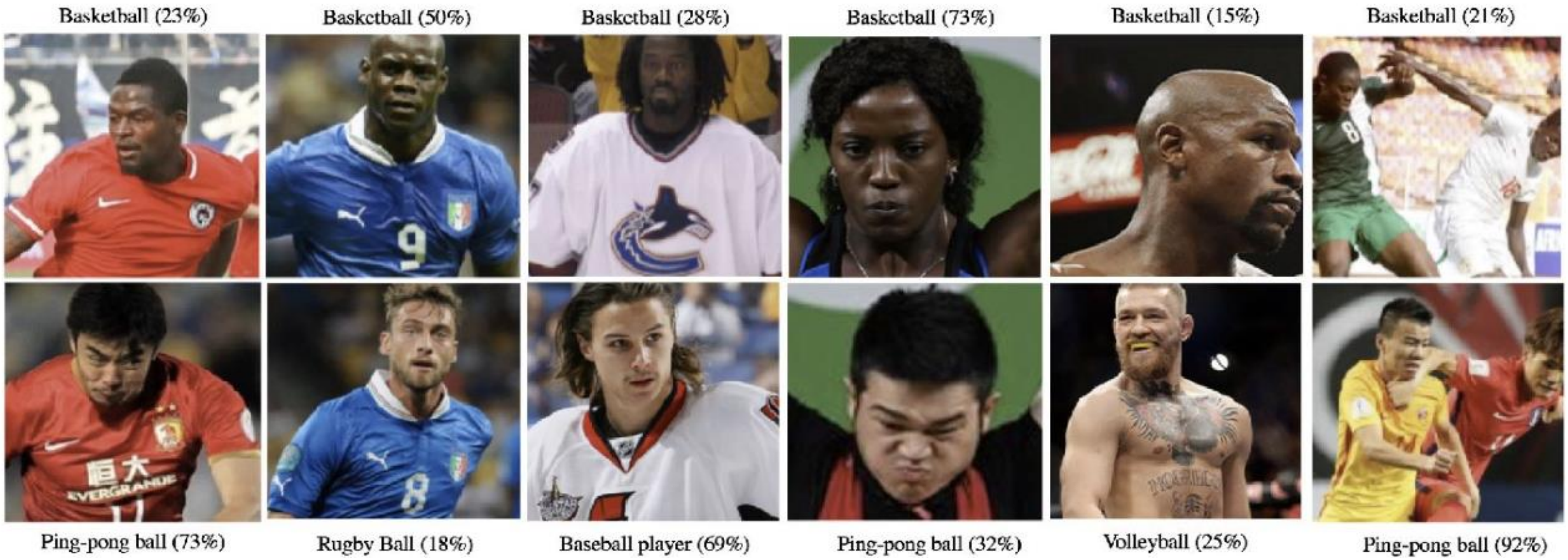
- **What is trustworthy AI?**
- **Explaining AI predictions**
- **Definitions of fairness in AI**
- **Framework for fair AI**
- **Algorithmic fairness criteria**
    - Individual fairness
    - Group fairness

Algorithmic fairness criteria

1) Individual Fairness

2) Group Fairness

Individual fairness: Similar individuals should be treated similarly



Problem: Pairs of similar individuals playing the same sport classified differently. The model is biased against individuals with certain characteristics

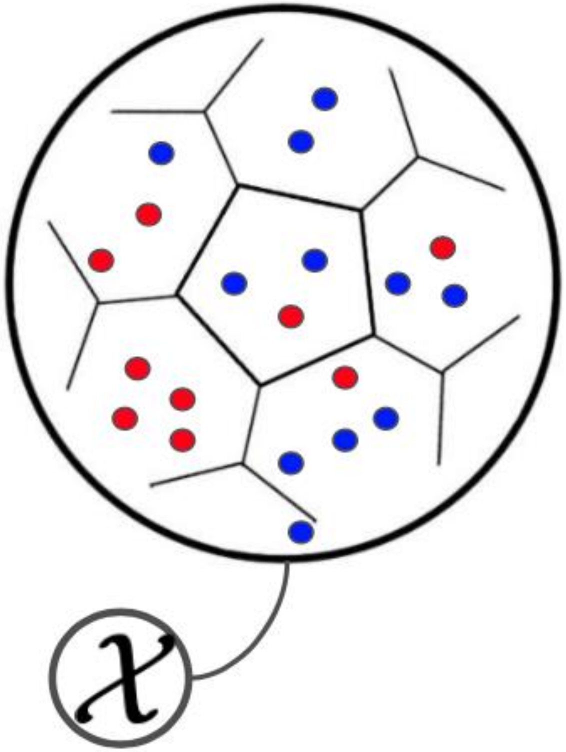
Shown are pairs of pictures (columns) sampled over the Internet along with their prediction by a ResNet-10.

Explore biases of a neural net by analyzing the distance of a sample to the decision boundary using adversarial samples.
The distance to the decision boundary is closely related to the magnitude of the perturbation necessary to make a sample cross it.

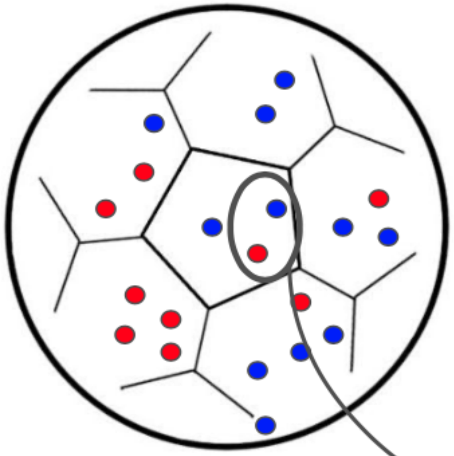
Stock and Cisse, ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases, ‘18

Individual fairness: Similar individuals should be treated similarly

- Data Regulator: Which individuals are similar? equiv., which individuals should be treated similarly?
- One approach:
  - Define a partition of the space into disjoint cells such that similar individuals are in the same cell
  - Individuals in the same cell should be treated similarly even if they are apparently different (e.g., dots with different colored attributes)



Individual fairness: Similar individuals should be treated similarly



Data Regulator: Which individuals are similar?
quiv., which individuals should be treated similarly?

An algorithm $\mathcal{A}_{\mathcal{D}}$ is $(B, \epsilon(\mathcal{D}))$-individually fair if $\mathcal{X}$ can be partitioned into B disjoint subsets denoted $\{C_i\}_{i=1}^{B}$ such that $\forall x_1 \in \mathcal{X}$:

$$x_1, x_2 \in C_i \Rightarrow |l(\mathcal{A}_{\mathcal{D}}, x_1) - l(\mathcal{A}_{\mathcal{D}}, x_2)| \leq \epsilon(\mathcal{D})$$

Remark: Individual fairness implies **algorithmic robustness** (c.f. Xu & Mannor ‘11)

Dwork et al., ‘12; Cisse and Koyejo, ‘20

Individual fairness: Pros and Cons

- Advantages:
  - Intuitive and easy to explain to data producers (and non-experts)
  - Individual fairness implies generalization (c.f. Xu & Mannor, ‘12)
  - Individual fairness implies statistical parity given regularity conditions (Dwork et al., ‘12)

- Challenges:
  - Regulator must provide a metric or a set of examples to be treated similarly
  - Constructing a metric requires significant domain expertise and human insight
  - Fairness of the representation heavily depends on the quality of the metric chosen by the regulator
  - Optimizing and measuring individual fairness is generally more computationally expensive than other measures

Algorithmic fairness criteria

✓ 1) Individual Fairness

☞ 2) Group Fairness

Group fairness: Similar classifier statistics across groups

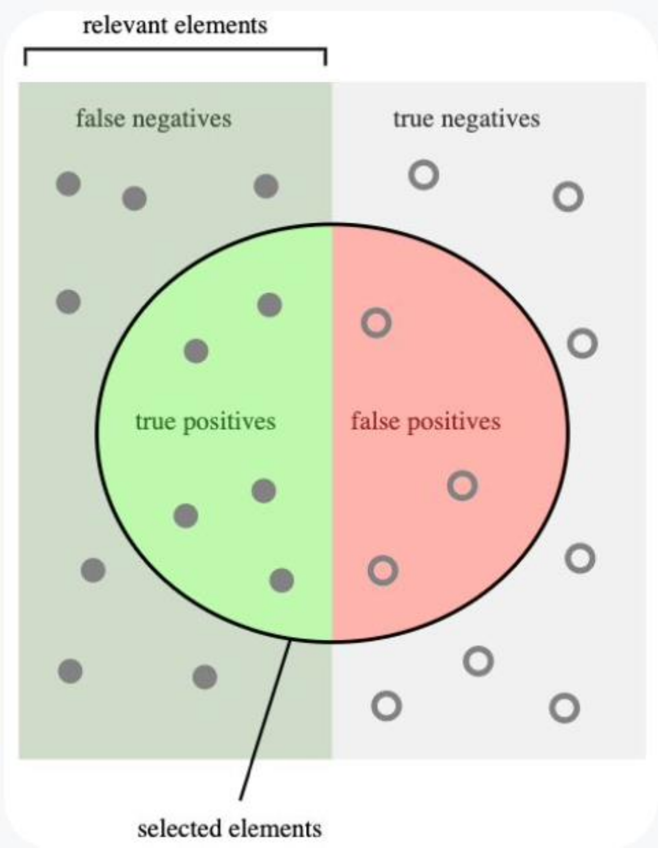
- **Regulator:** Which statistic $\nu(f, Y|S)$ should be equalized across groups S?

- Typical **fairness measure** is a of the ML model performance:
  - **Eq. of opportunity** (Hardt et al., '16)

    $$TP_S = P(Y = 1, f = 1|S)$$
  - **Equalized odds** (Hardt et al., '16)

    $$\{TP_S; FP_S\}$$
  - **Statistical parity** (Dwork et al., '12)

    $$TP_S + FP_S = P(f(Z) = 1|S)$$



Details #1: Statistical Parity

- Statistical parity is a popular measure of group fairness
- Setup:
  - Population is a set X
  - Subset $S \subset X$ that is a "protected" subset of the population

- Example:
  - X is people
  - S is people who dye their hair blue
  - We are afraid that banks give fewer loans to the blues because of hair-colorism, despite blue-haired people being just as creditworthy as the general population on average

Details #2: Statistical parity

- **Assumption:** There is some distribution D over X which represents the probability that any individual will be drawn for evaluation

- Example:
  - Some people will have no reason to apply for a loan (maybe they're filthy rich, or don't like homes, cars, or expensive colleges)
  - D takes that into account
  - Generally, we impose no restrictions on D, and the definition of fairness will work no matter what D is

Details #3: Statistical parity

- Classifier $f: X \rightarrow \{0,1\}$ gives labels to X
  - When given a person x as input $f(x) = 1$ if x gets a loan and 0 otherwise
- Statistical imparity of f on S with respect to X, D:

$$\mathrm{imparity}_f(X,S,D) = \underbrace{P(f(x) = 1|x \in S^C)}_{\text{Probability that a random individual from the complement } S^C \text{ is labeled 1}} - \underbrace{P(f(x) = 1|x \in S)}_{\text{Probability that a random individual drawn from } S \text{ is labeled 1}}$$

- This is the statistical equivalent of adverse impact
  - It measures the difference that the majority and protected classes get a particular outcome

Details #4: Statistical parity

- Statistical imparity measures the difference that the majority and protected classes get a certain outcome
- When the difference is small, the classifier has **statistical parity**, it conforms to this notion of fairness
- **Definition:** ML model $f: X \to \{0,1\}$ achieves statistical parity on D with respect to S up to bias ϵ if $|\mathrm{imparity}_f(X,S,D)| < \epsilon$
- If f achieves statistical parity, it treats the general population statistically similarly as the protected class
  - If 30% of normal-hair-colored people get loans, statistical parity requires roughly 30% of blue also get loans

Group fairness: Pros and Cons

- Advantages:
  - Efficient to compute, measure and enforce for data producer and regulator
  - Often easier to explain to policy-makers (as in terms of population behavior)

- Challenges:
  - Data regulator must determine which classifier statistic(s) to equalize
  - Fairness of the representation depends on the quality of the fairness metric chosen by the regulator
  - Group fairness can lead to (more) violated individual fairness, e.g., intersectionality
  - It can lead to fairness gerrymandering (Kearns et. al., ‘18), and other issues (McNamara et. al., ‘19)

Algorithmic fairness criteria

1) ✓ Individual Fairness

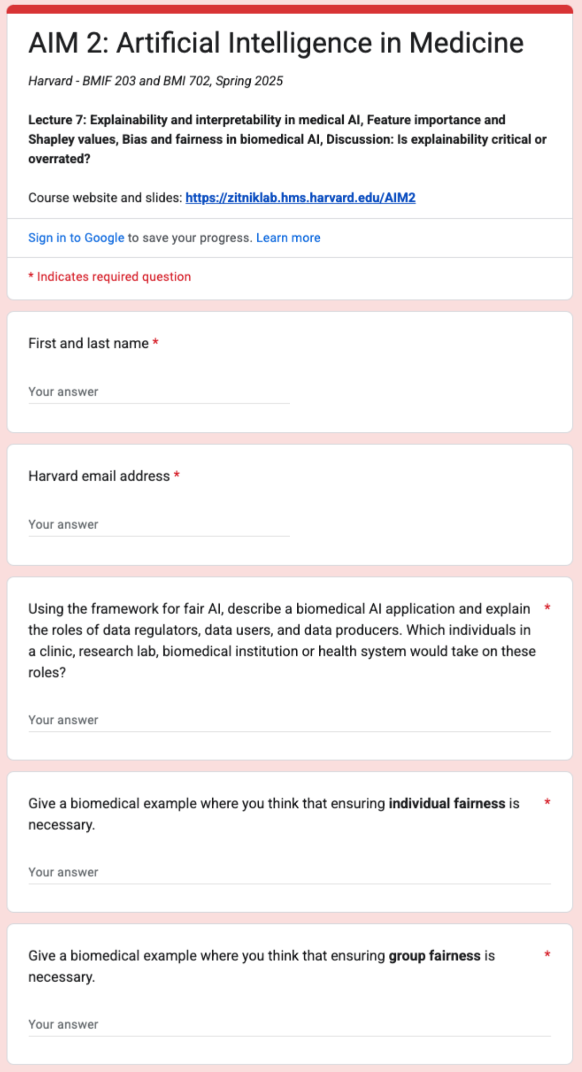
2) ✓ Group Fairness

Data regulator: Measures (un-)fairness

- Regulator must choose how to measure (un-)fairness:
  - **For individual fairness:** must choose the distance metric
  - **For group fairness:** must choose the classifier statistic to equalize
- However, remember that there are no magic metrics:
  - Measurement 101: all measures have **blind spots**
  - *"When a measure becomes a target, it ceases to be a good measure"*
- For ML, we generally specify all measures apriori and optimize them
  - However, **all** metrics will have failure cases, i.e., unusual situations with non-ideal behavior
- One productive approach is to select measures that best capture tradeoffs relevant to the context

Quick Check

https://forms.gle/PwhV3CEN74aywbE68



AIM 2: Artificial Intelligence in Medicine

Harvard - BMIF 203 and BMI 702, Spring 2025

Lecture 7: Explainability and interpretability in medical AI, Feature importance and Shapley values, Bias and fairness in biomedical AI, Discussion: Is explainability critical or overrated?

Course website and slides: https://zitniklab.hms.harvard.edu/AIM2

Sign in to Google to save your progress. Learn more

* Indicates required question

First and last name *

Your answer

Harvard email address *

Your answer

Using the framework for fair AI, describe a biomedical AI application and explain the roles of data regulators, data users, and data producers. Which individuals in a clinic, research lab, biomedical institution or health system would take on these roles? *

Your answer

Give a biomedical example where you think that ensuring **individual fairness** is necessary. *

Your answer

Give a biomedical example where you think that ensuring **group fairness** is necessary. *

Your answer

Outline of today’s class

- What is trustworthy AI?
- Explaining AI predictions
- Definitions of fairness in AI
- Framework for fair AI
- Algorithmic fairness criteria
    - Individual fairness
    - Group fairness