Explainable Machine Learning: Understanding the Limits & Pushing the Boundaries

Hima Lakkaraju

Tutorial Outline

- Motivation
- Interpretability vs. Explainability
- Overview of Explanation Methods
- Limitations of Explanation Methods
- The Road Ahead

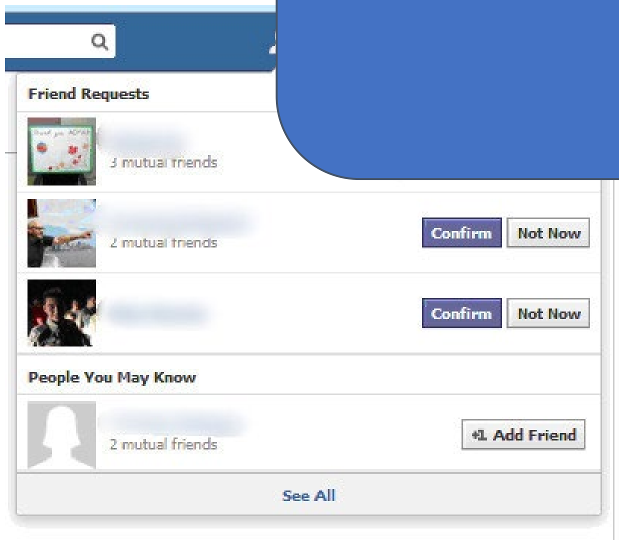
Tutorial Outline

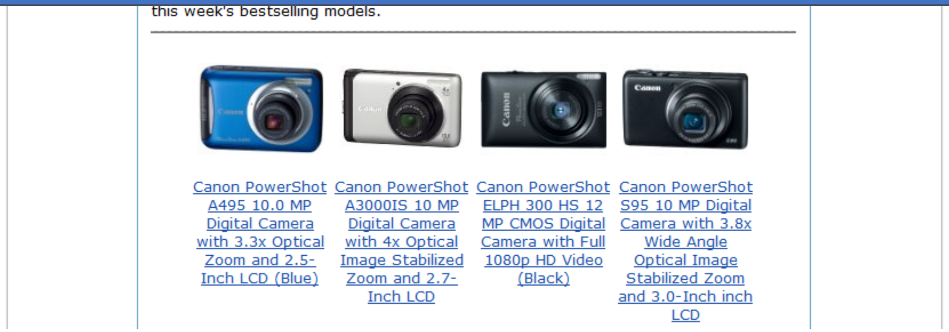
- **Motivation**
- Interpretability vs. Explainability
- Overview of Explanation Methods
- Limitations of Explanation Methods
- The Road Ahead

[Weller 2017]

Motivation

Machine Learning is EVERYWHERE!!



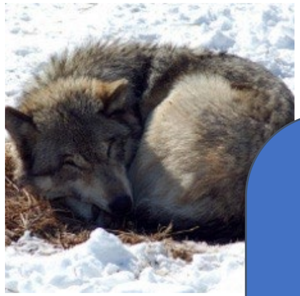


Motivation

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

Motivation: Why Model Understanding?

Input



Predictive Model

Prediction = Siberian Husky

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

Model understanding facilitates debugging.

Motivation: Why Model Understanding?

This prediction is

Defendant Details

Model understanding facilitates bias detection.

Predictive Model

Prediction = Risky to Release

Motivation: Why Model Understanding?

Loan Applicant Details

FILE

I have some means
Let me
my
pay
ne.

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

to get a loan

Predictive Model

Prediction = Denied Loan

Loan Applicant

Motivation: Why Model Understanding?

Patient Data

25, Femal
32, Male,
31, Male,
.
.
.
.

Model Understanding

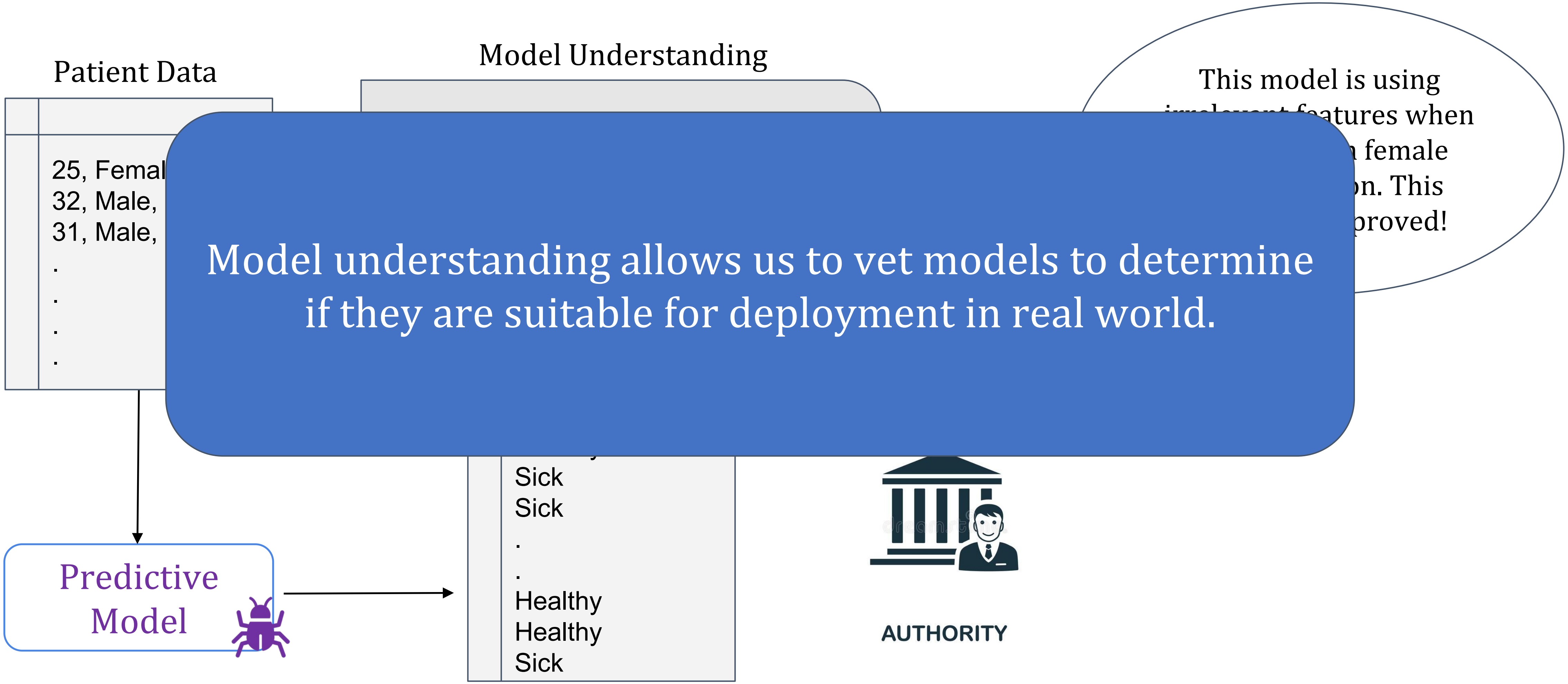
Predictive Model

Sick
Sick
.
.
Healthy
Healthy
Sick

This model is using
irrelevant features when
on female
. I should
edictions
oup.

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

Motivation: Why Model Understanding?



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

Motivation: Why Model Understanding?

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)

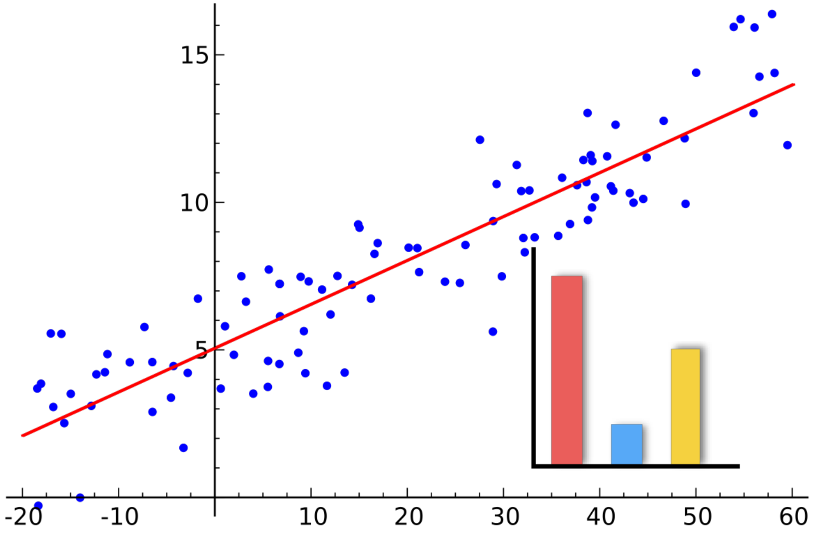
Researchers and engineers

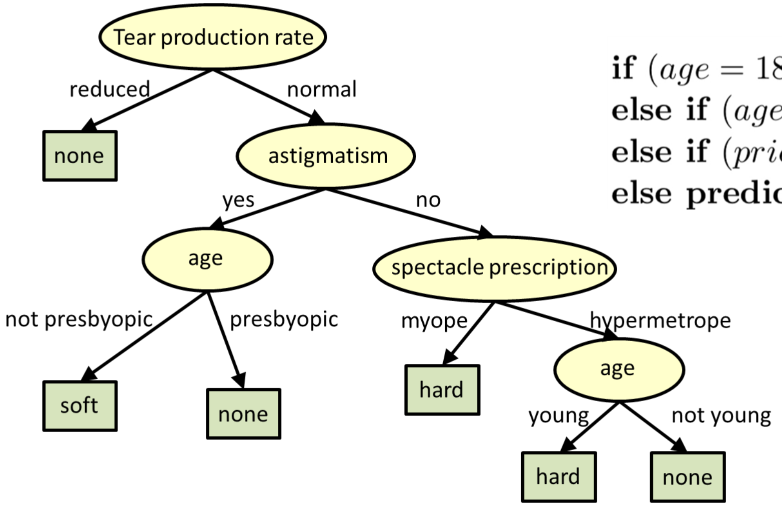
Tutorial Outline

- Motivation
- **Interpretability vs. Explainability**
- Overview of Explanation Methods
- Limitations of Explanation Methods
- The Road Ahead

Achieving Model Understanding

Take 1: Build *inherently interpretable* predictive models



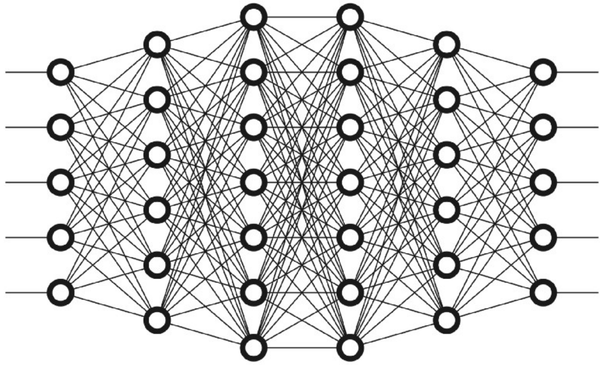


if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

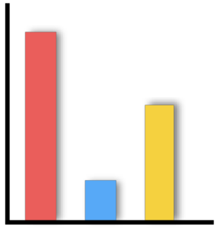
[Ribeiro et. al. 2016, 2018; Lakkaraju et. al. 2019]

Achieving Model Understanding

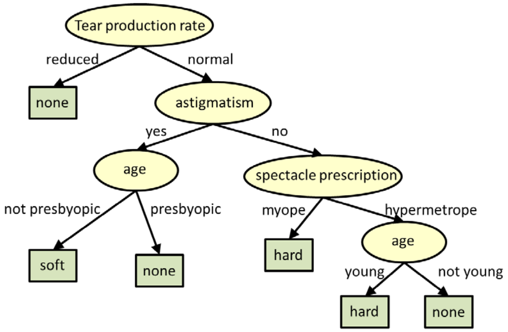
Take 2: *Explain* pre-built models *in a post-hoc manner*



Explainer



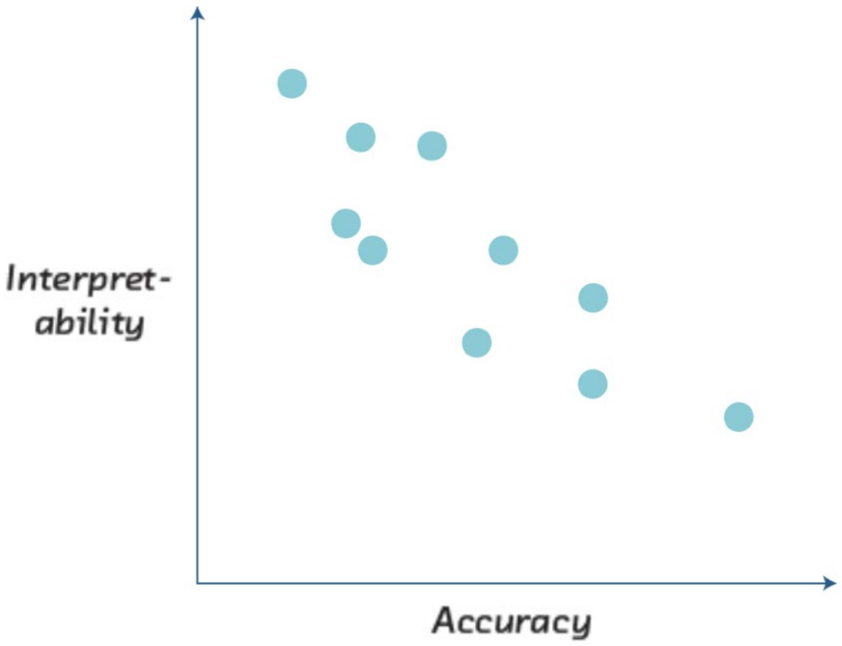
$\textbf{if } (age = 18-20) \textbf{ and } (sex = male) \textbf{ then predict } yes$
$\textbf{else if } (age = 21-23) \textbf{ and } (priors = 2-3) \textbf{ then predict } yes$
$\textbf{else if } (priors > 3) \textbf{ then predict } yes$
$\textbf{else predict } no$

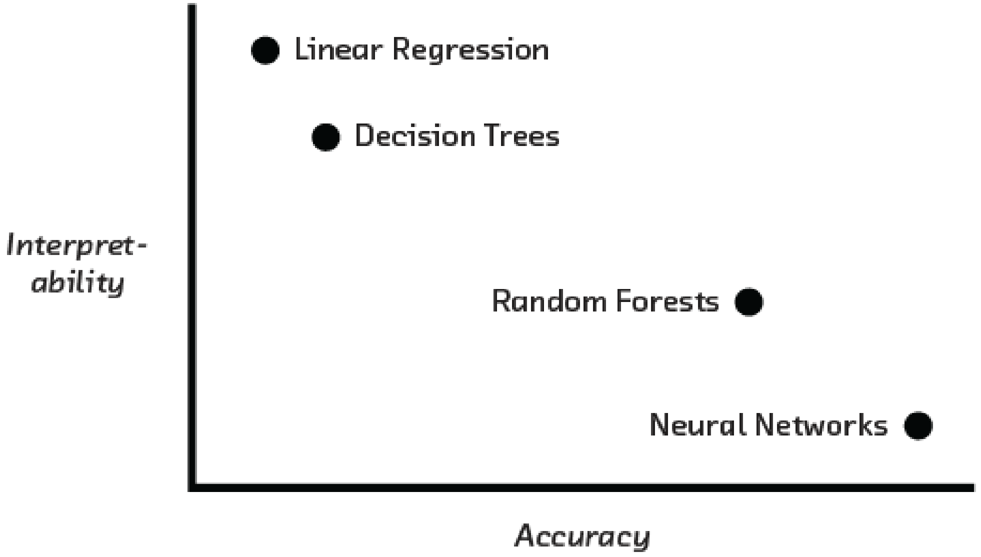


[Cireşan et. al. 2012, Caruana et. al. 2006, Frosst et. al. 2017, Stewart 2020]

Inherently Interpretable Models vs. Post hoc Explanations

Example





In ***certain*** settings, *accuracy-interpretability trade offs* may exist.

[Ribeiro et. al. 2016]

Inherently Interpretable Models vs. Post hoc Explanations

Sometimes, you don't have enough data to build your model from scratch.

And, all you have is a (proprietary) black box!

[Rudin 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!

Otherwise, ***post hoc explanations*** come to the rescue!

This talk will focus on post hoc explanations!

Tutorial Outline

- Motivation
- Interpretability vs. Explainability
- **Overview of Explanation Methods**
- Limitations of Explanation Methods
- The Road Ahead

What is an Explanation?

Definition: Interpretable description of the model behavior

Classifier

Faithful

Explanation

Understandable

User

Overview of Explanation Methods

Local Explanations vs. Global Explanations

Local Explanations	Global Explanations
Explain individual predictions	Explain complete behavior of the model
Help unearth biases in the *local neighborhood* of a given instance	Sheds light on *big picture biases* affecting larger subgroups
Help vet if individual predictions are being made for the right reasons	Help vet if the model, at a high level, is suitable for deployment

Overview of Explanation Methods

Local Explanations

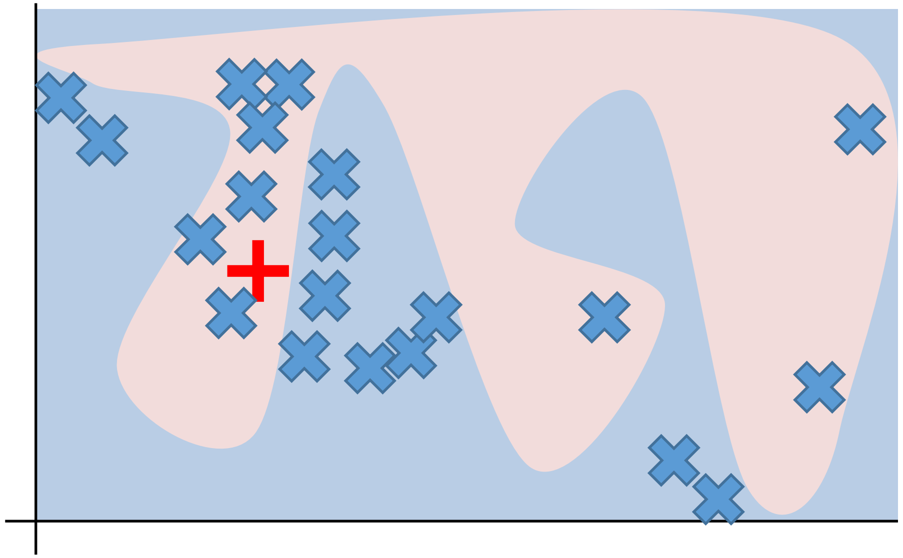
- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

LIME: Local Interpretable Model-Agnostic Explanations

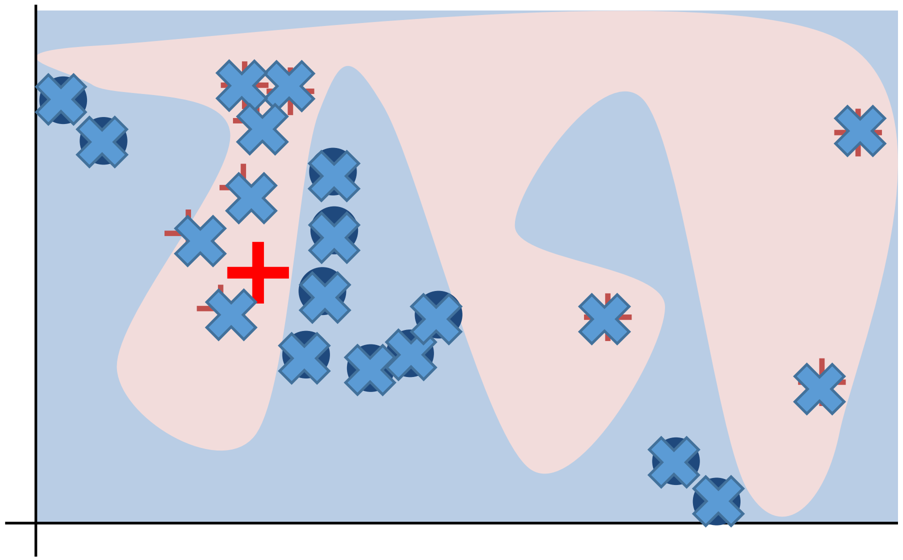
1. Sample points around x_i



[Ribeiro et al. 2016]

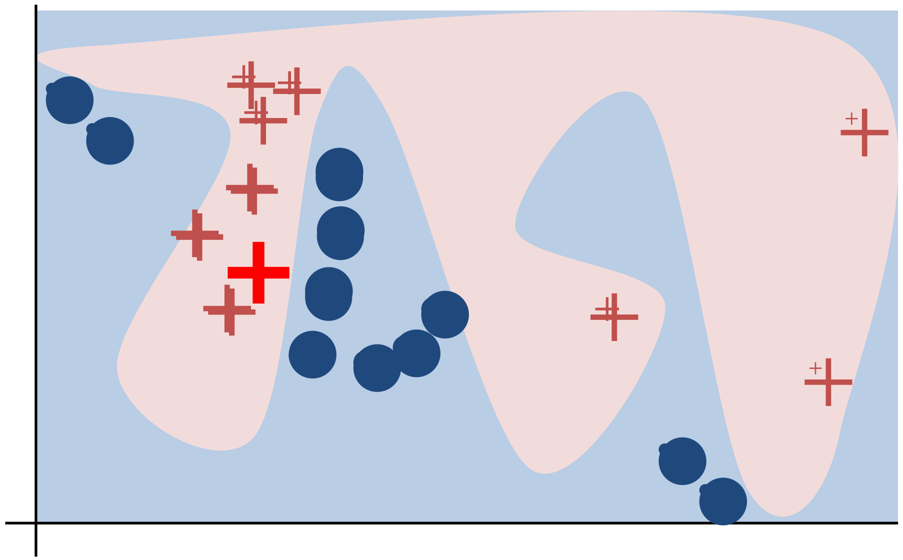
LIME: Local Interpretable Model-Agnostic Explanations

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



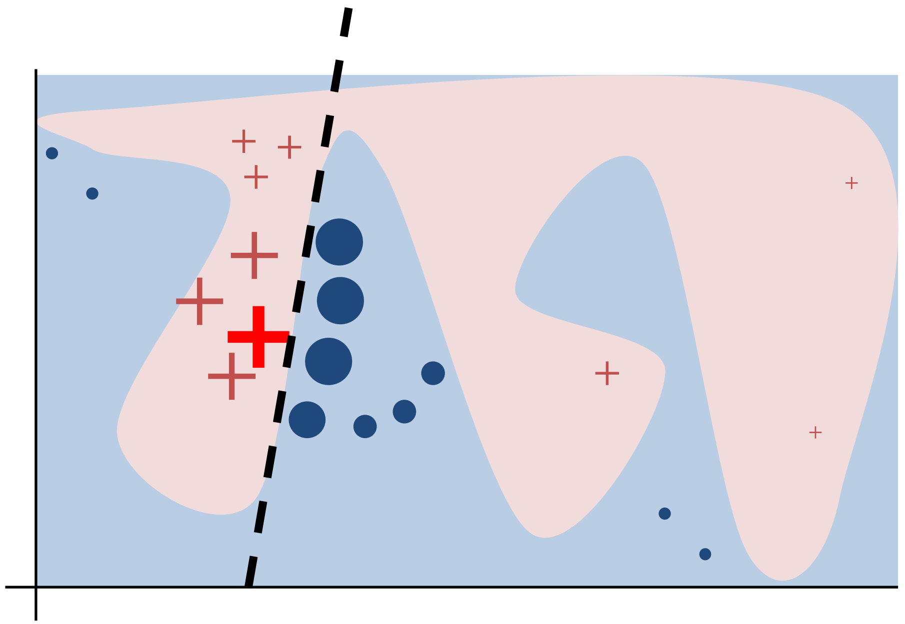
LIME: Local Interpretable Model-Agnostic Explanations

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



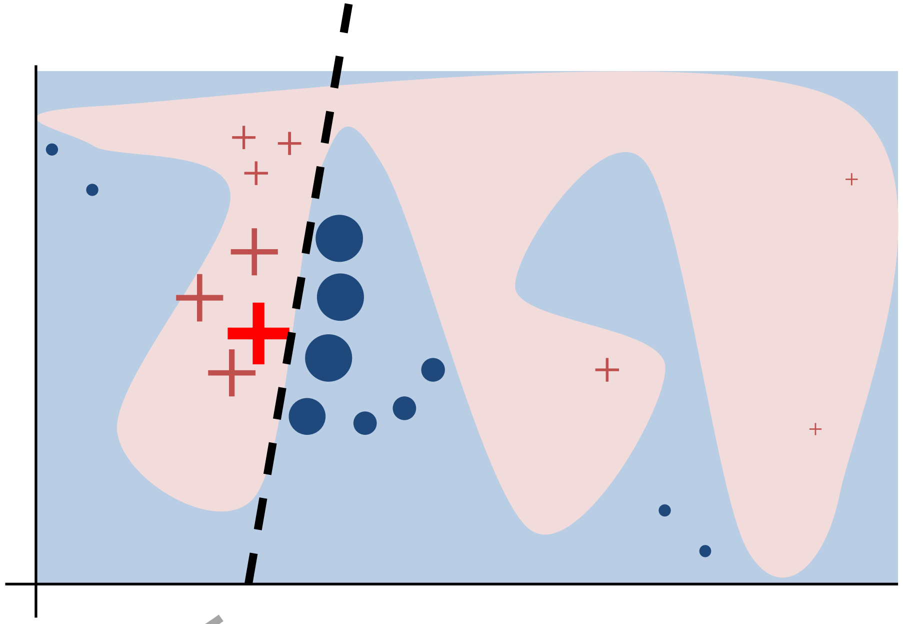
LIME: Local Interpretable Model-Agnostic Explanations

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



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



Another popular method which outputs feature importances: SHAP

Overview of Explanation Methods

Local Explanations

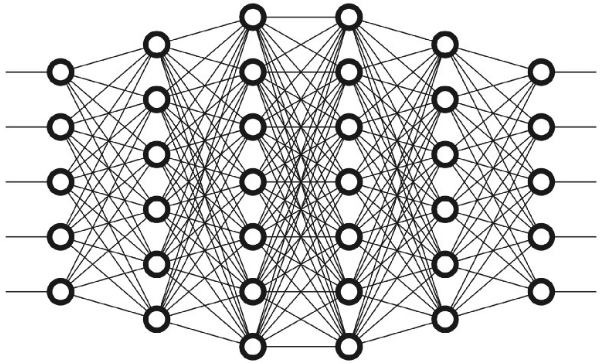
- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

Saliency Maps

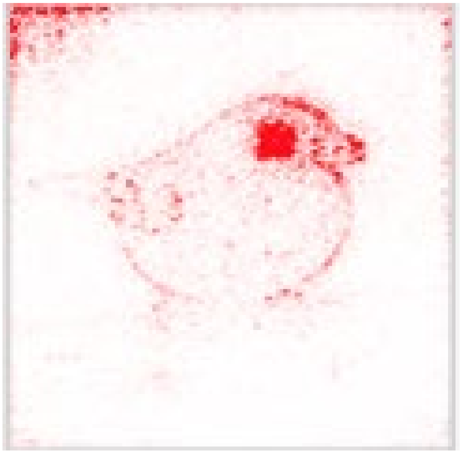
Input



Prediction

Junco Bird

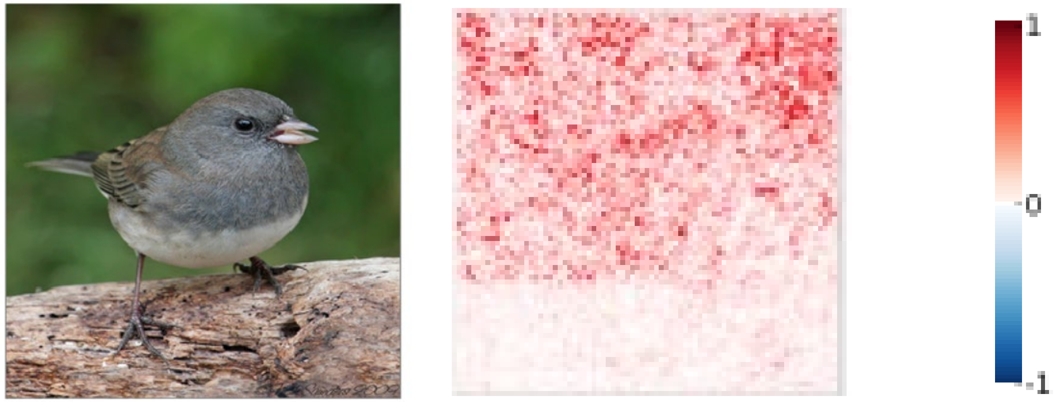
What parts of the input are most relevant for the model's prediction: 'Junco Bird'?



Saliency Map

Saliency Maps

Gradient



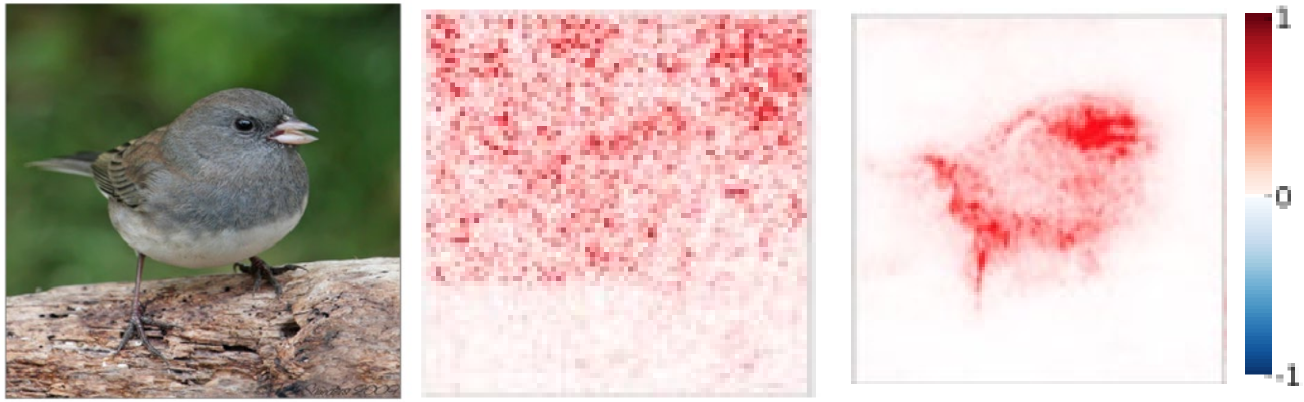
$$\nabla_{x} f$$

Problems:

- noisy and uninterpretable

Saliency Maps

Gradient **SmoothGrad**



$$\frac{1}{n}\sum_{1}^{n}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}+\mathcal{N}(0,\sigma^2))$$

Problems:

- ~~noisy and uninterpretable~~

Overview of Explanation Methods

Local Explanations

- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

Prototypes/Example

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

Overview of Explanation Methods

Local Explanations

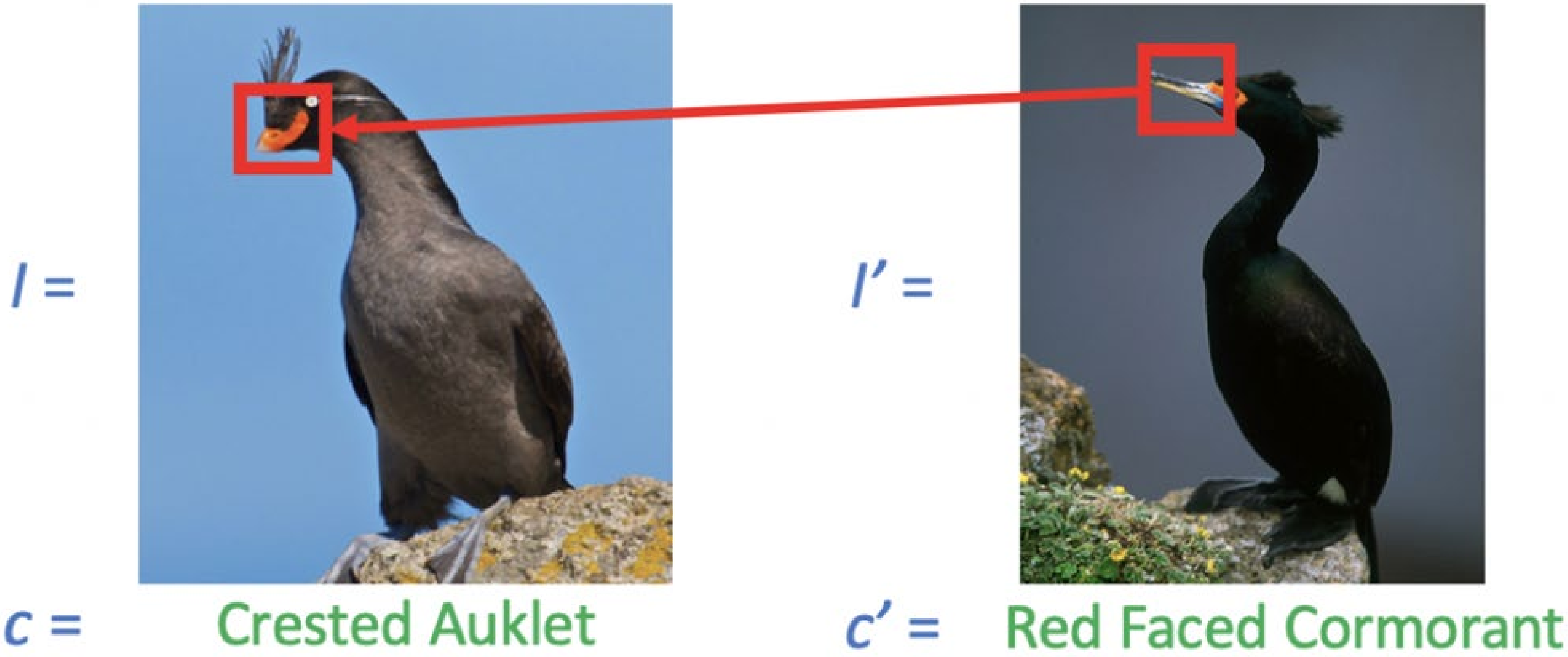
- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

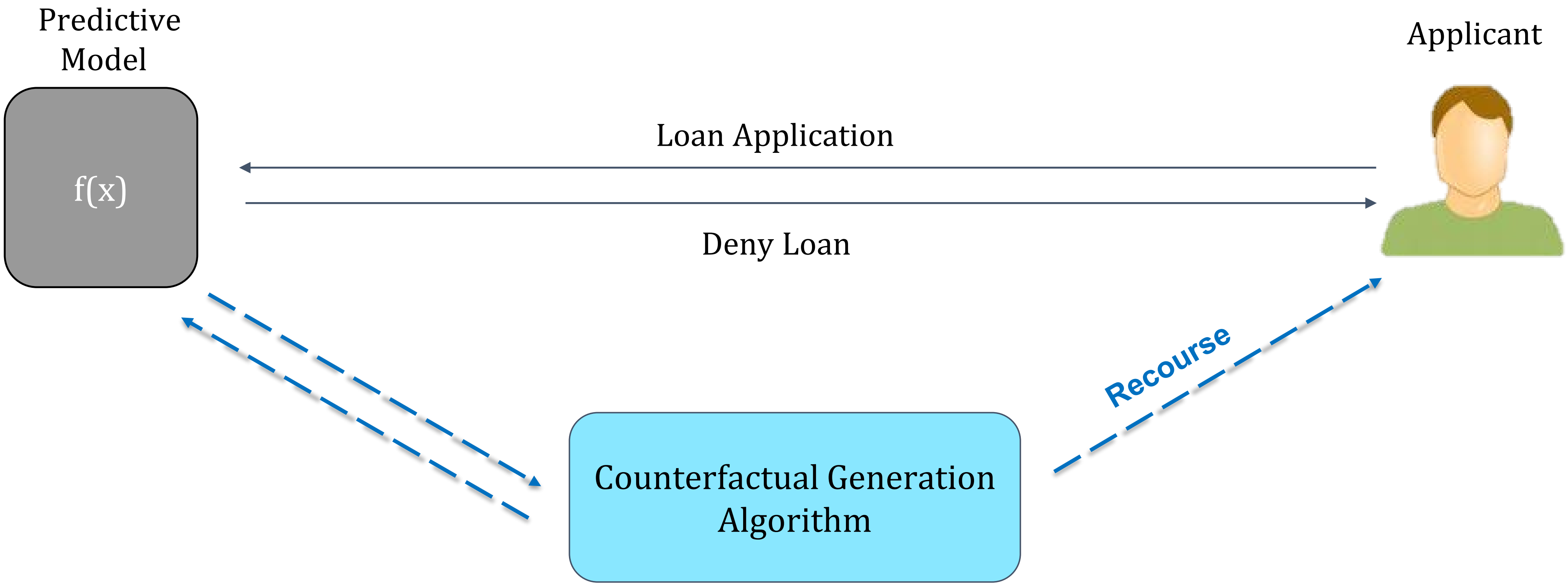
Counterfactual Explanations

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



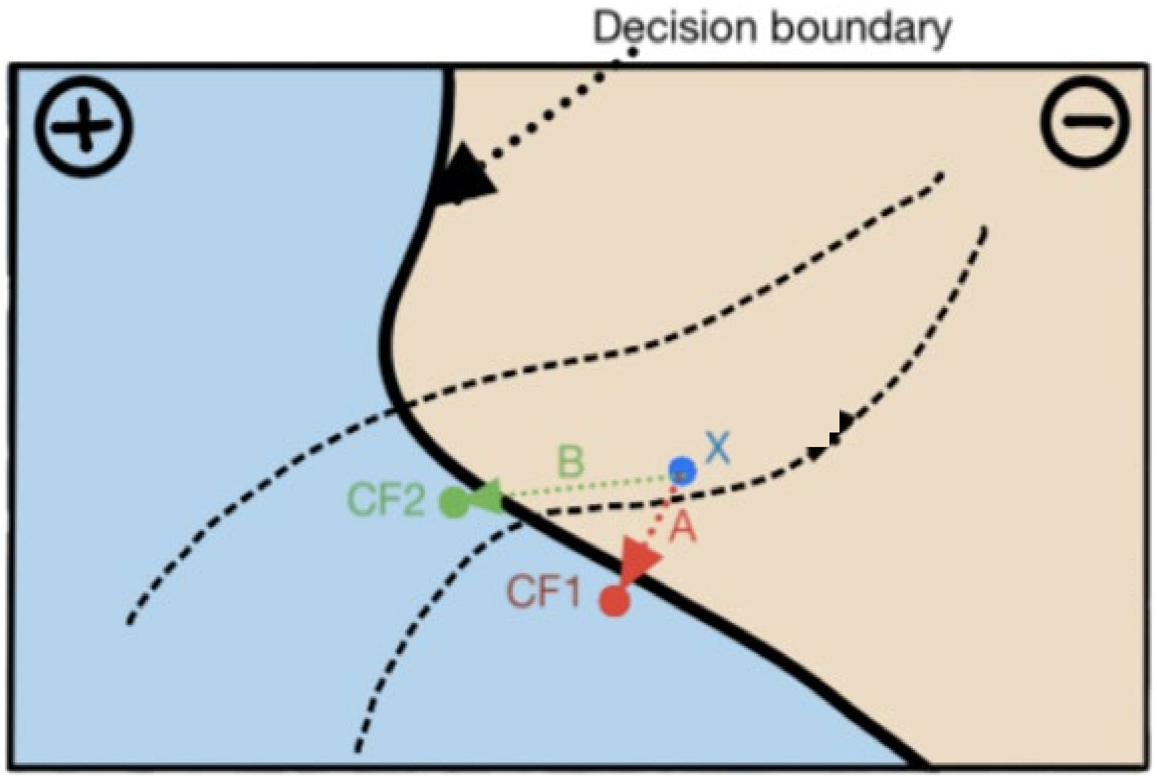
[Goyal et. al., 2019]

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]

Overview of Explanation Methods

Local Explanations

- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

Global Explanations from Local Feature Importances: SP-LIME

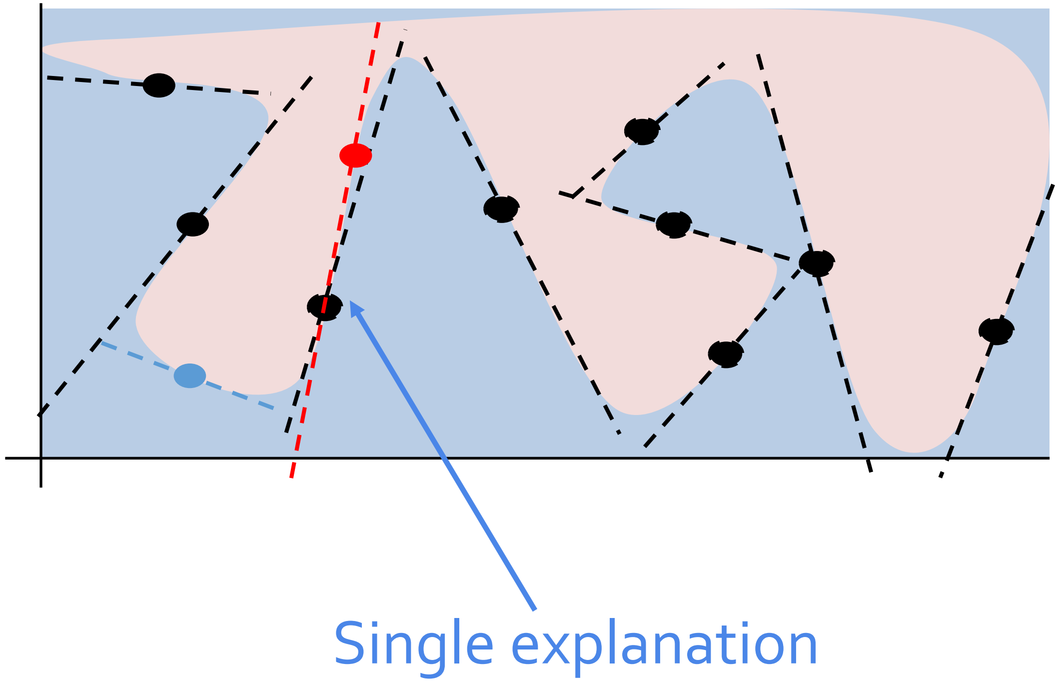
LIME explains a single prediction
local behavior for a single instance

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



[Ribeiro et. al., 2016]

Overview of Explanation Methods

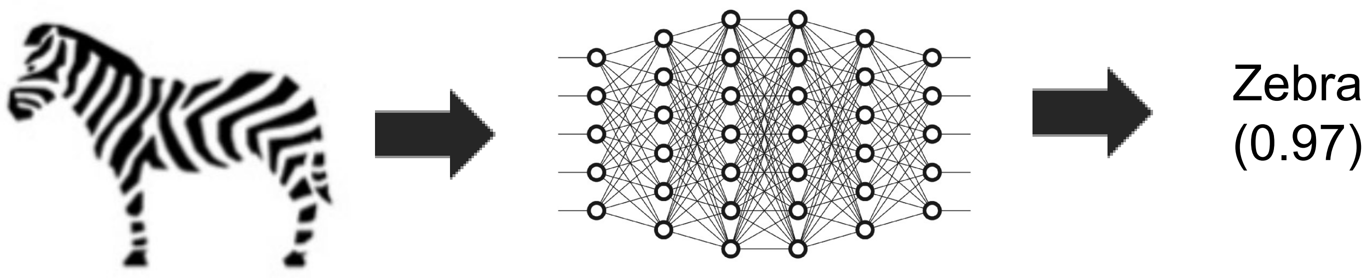
Local Explanations

- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

Representation Based Explanations



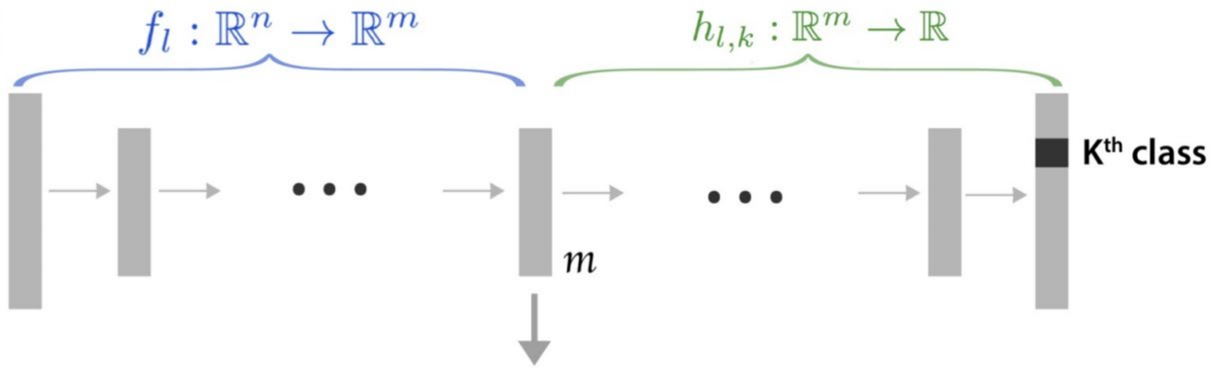
How important is the notion of “stripes” for this prediction?

[Kim et. al., 2018]

Representation Based Explanations: TCAV

Examples of the concept “stripes”

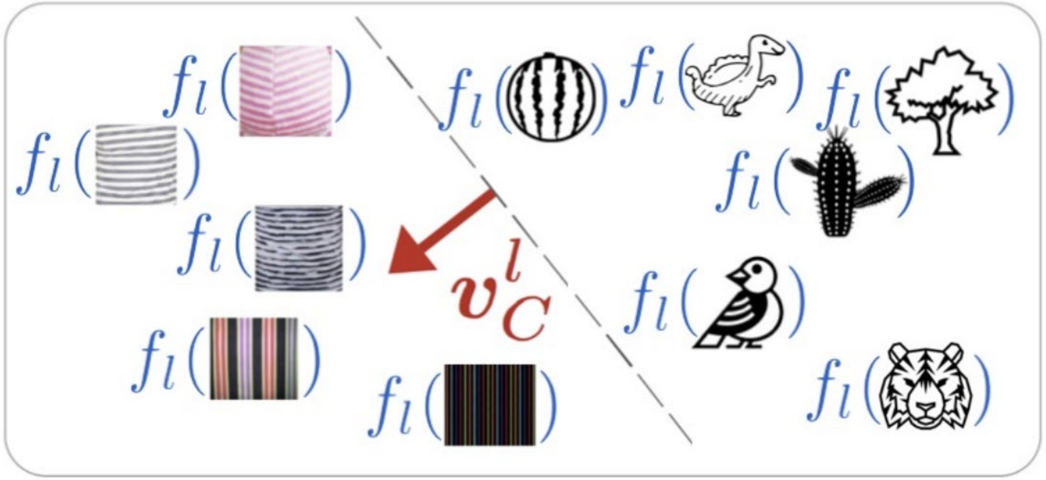
Random examples



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



Overview of Explanation Methods

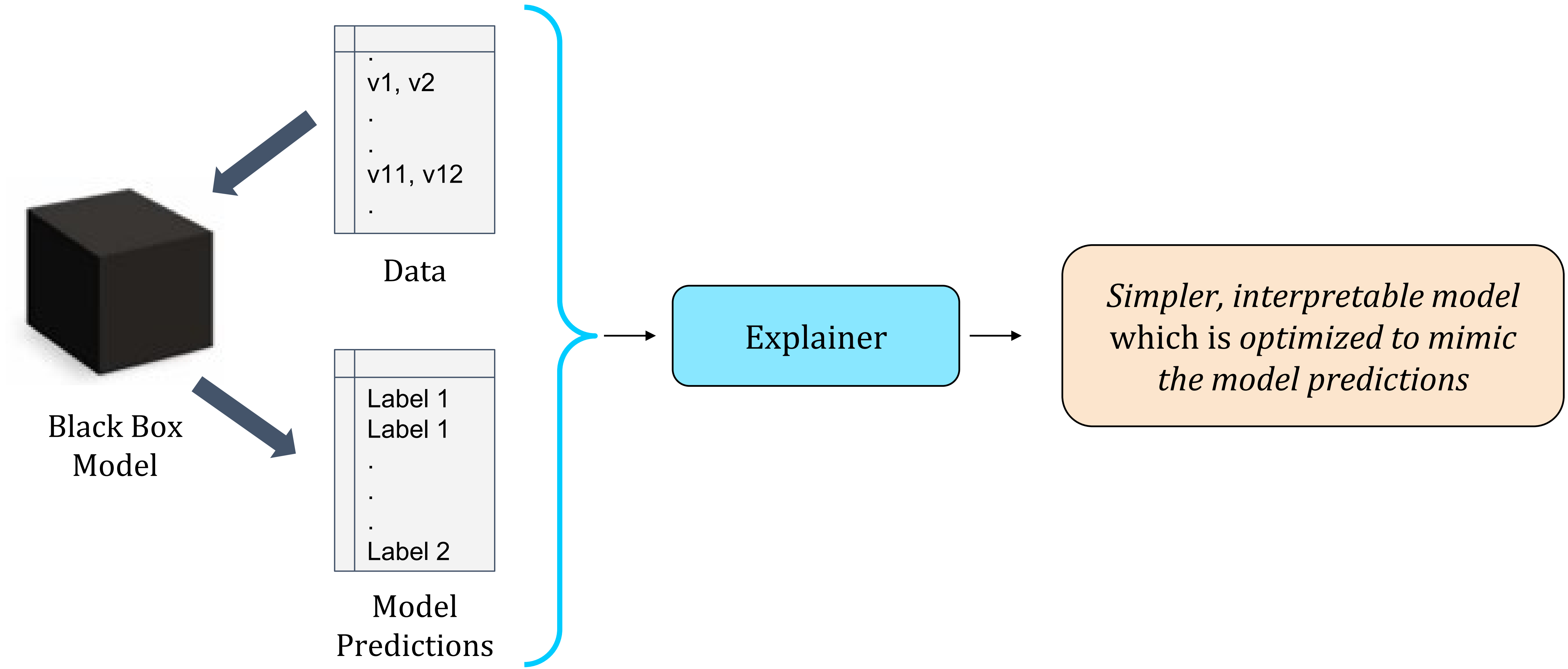
Local Explanations

- Feature Importances
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

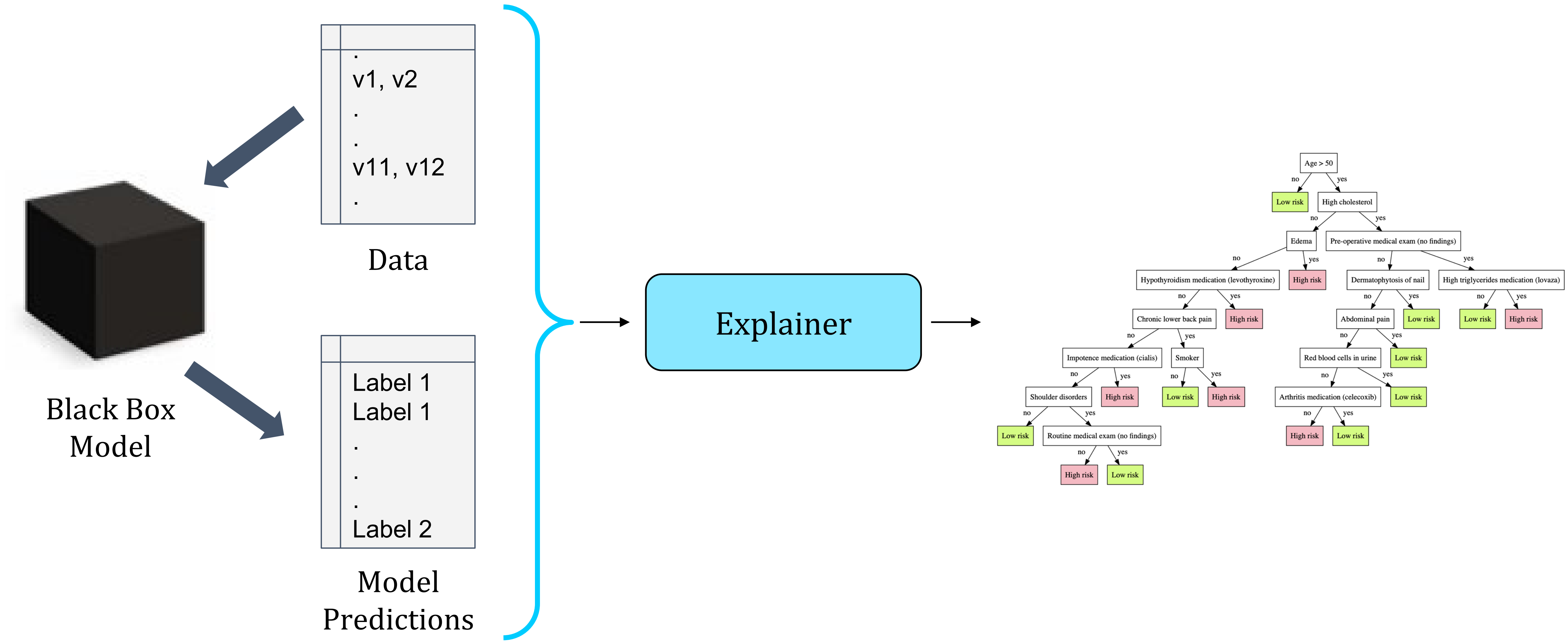
Global Explanations

- Collection of Local Explanations
- Representation Based
- Model Distillation

Model Distillation

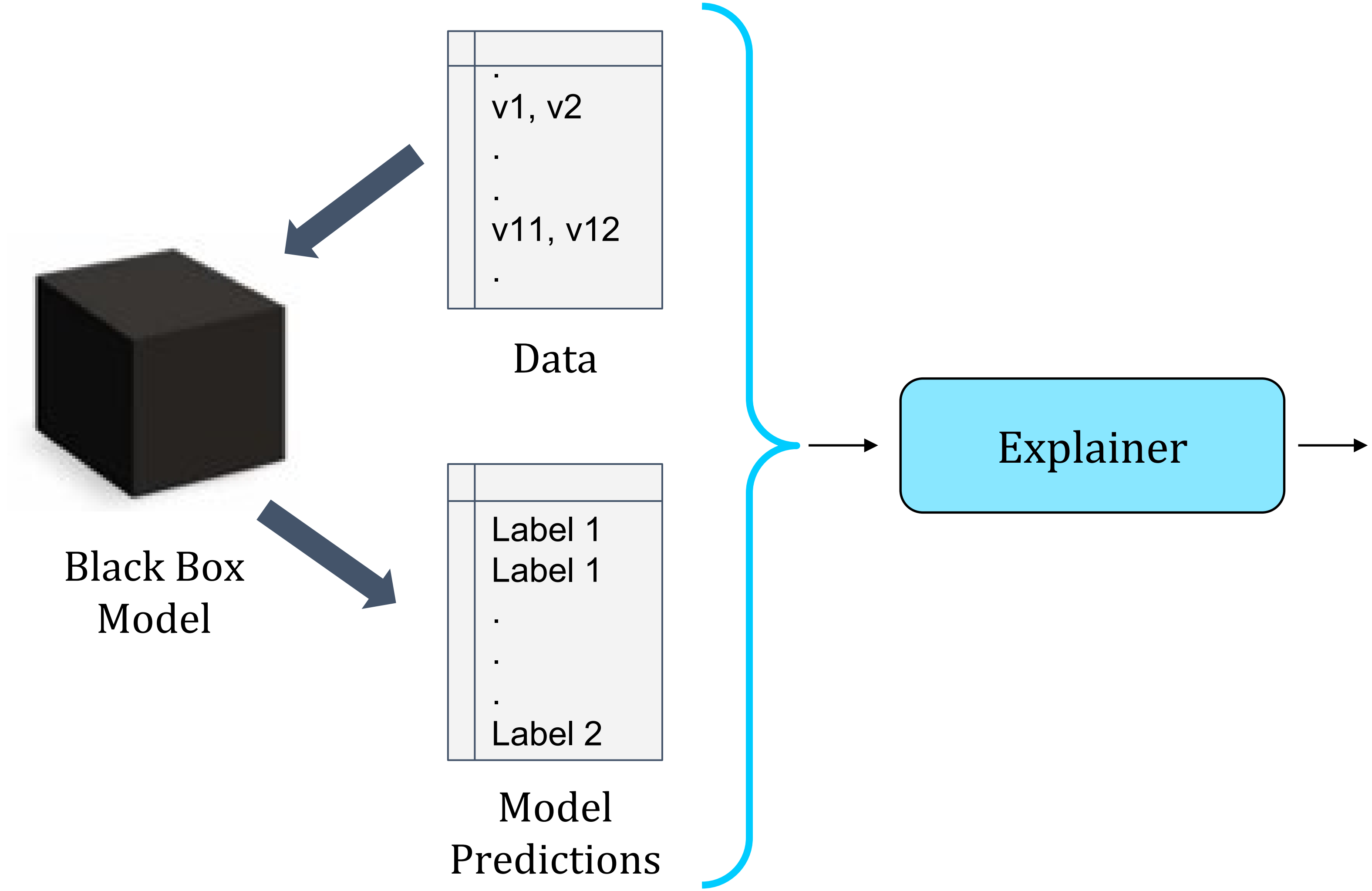


Model Distillation Using Decision Trees



[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 $\geq$ 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

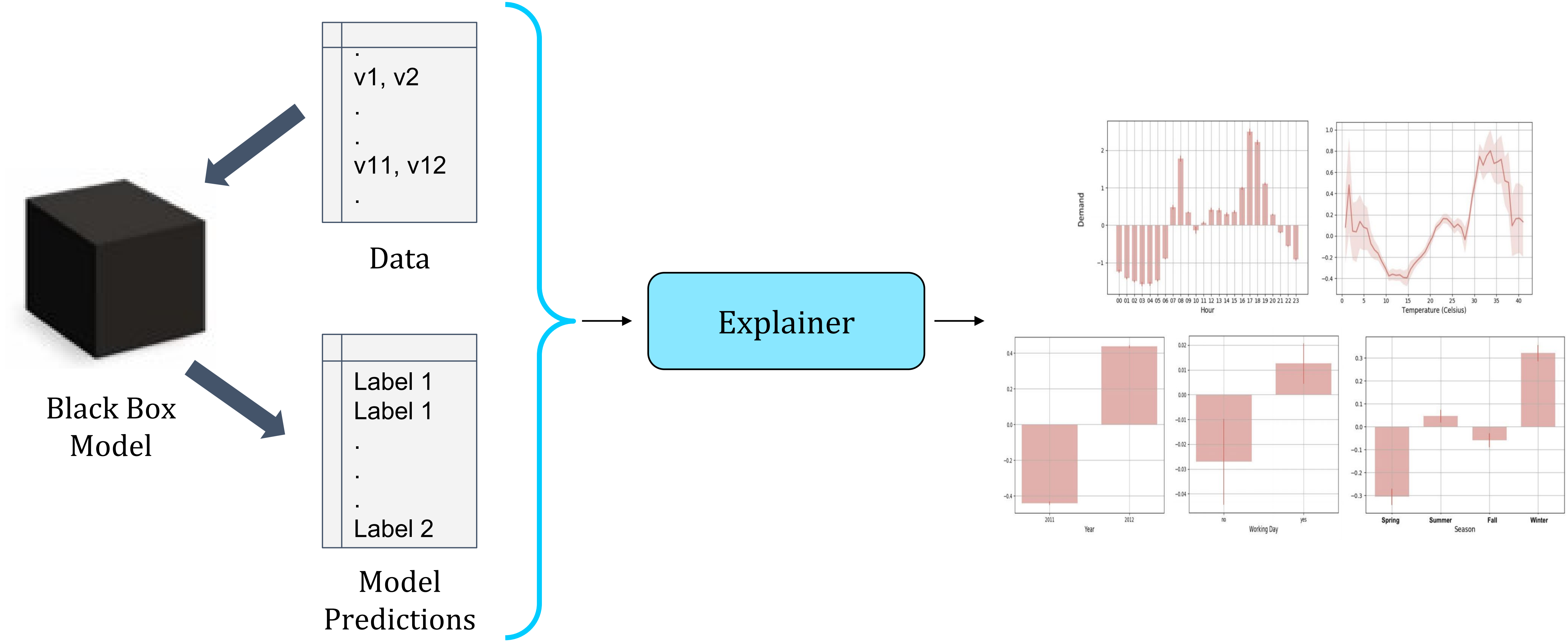
Default:

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



[Tan et. al., 2019]

Tutorial Outline

- Motivation
- Interpretability vs. Explainability
- Overview of Explanation Methods
- **Limitations of Explanation Methods**
- The Road Ahead

Limitations of Explanation Methods

Faithfulness
Some explanation methods do not ‘reflect’ the underlying model.

Stability
Slight changes to inputs can cause large changes in explanations.

Fragility
Post-hoc explanations can be easily manipulated.

Limitations of Explanation Methods

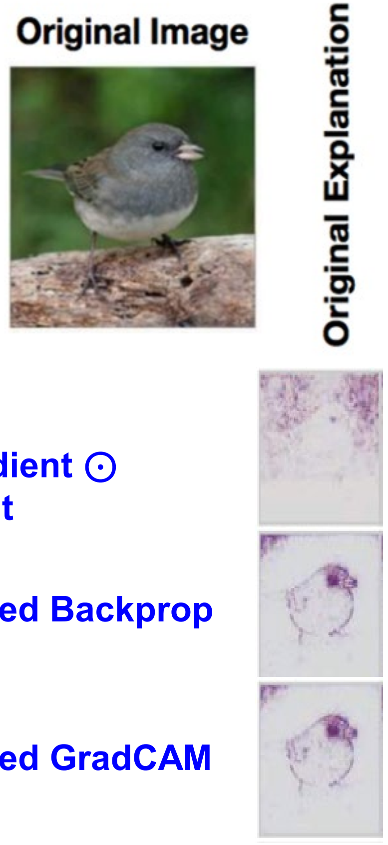
Faithfulness
Some explanation methods do not ‘reflect’ the underlying model.

Stability
Slight changes to inputs can cause large changes in explanations.

Fragility
Post-hoc explanations can be easily manipulated.

Limitations: **Faithfulness**

Model parameter randomization test



Gradient ⊙ Input

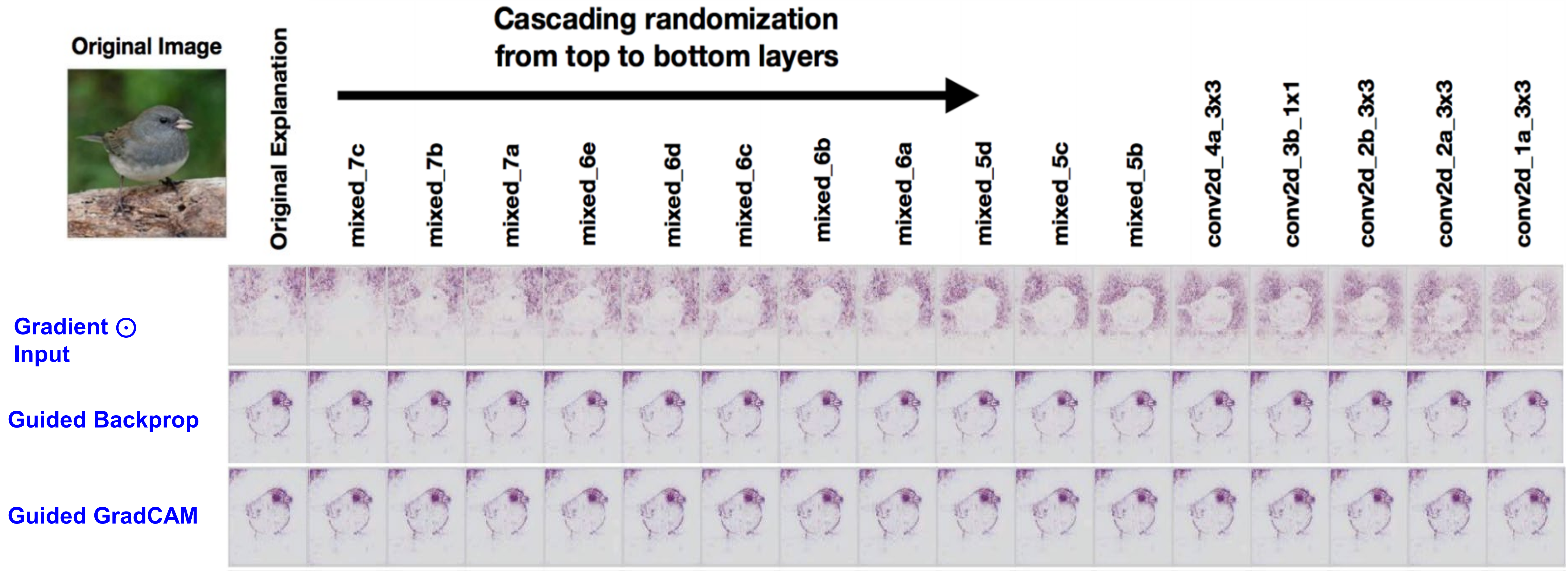
Guided Backprop

Guided GradCAM

[Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.](#)

Limitations: **Faithfulness**

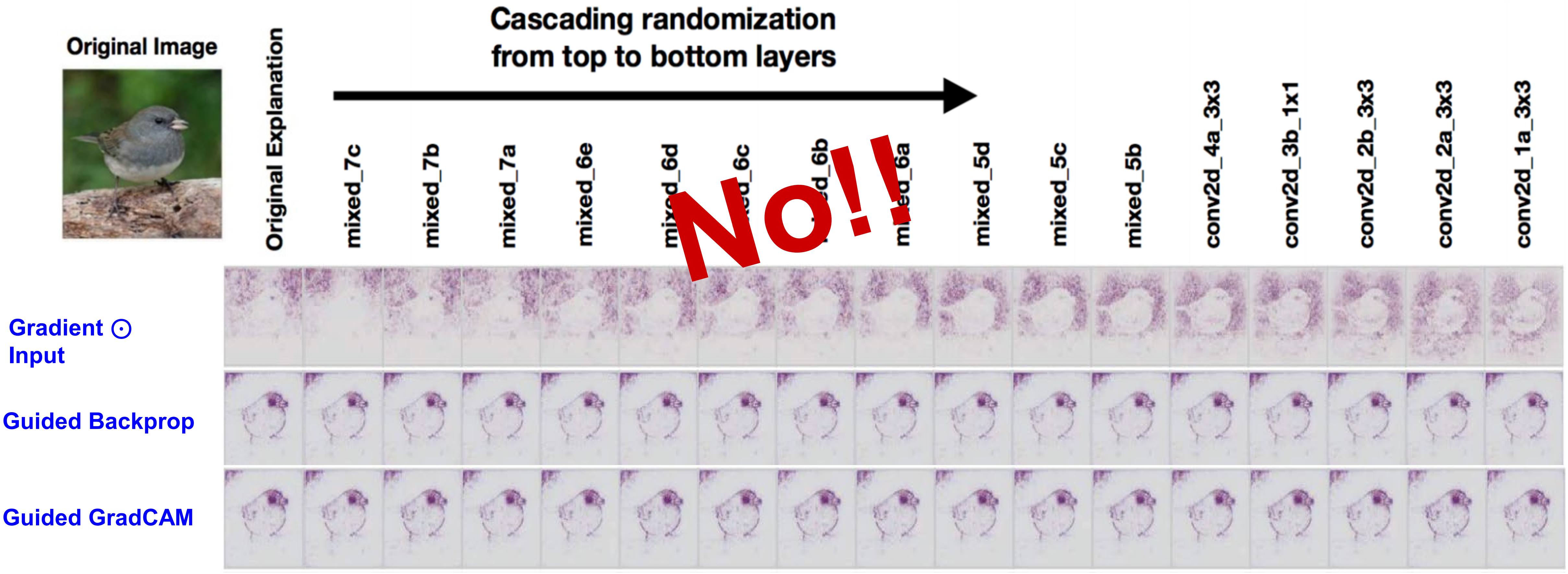
Model parameter randomization test



Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.

Limitations: **Faithfulness**

Model parameter randomization test



[Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.]

Limitations: **Faithfulness**

Randomizing class labels of instances also didn’t impact explanations!

Adebayo, Julius, et al. "Sanity checks for saliency maps." NeurIPS, 2018.

Limitations of Explanation Methods

Faithfulness
Some explanation methods do not ‘reflect’ the underlying model.

Stability
Slight changes to inputs can cause large changes in explanations.

Fragility
Post-hoc explanations can be easily manipulated.

Limitations: **Stability**

Are post-hoc explanations unstable wrt small non-adversarial input perturbation?

Local Lipschitz Constant

Explanation function: LIME, SHAP, Gradient...etc.

$$\hat{L}(x_i) = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|f(x_i) - f(x_j)\|_2}{\|x_i - x_j\|_2}$$

Input

input model

$$A(x, f, H)$$

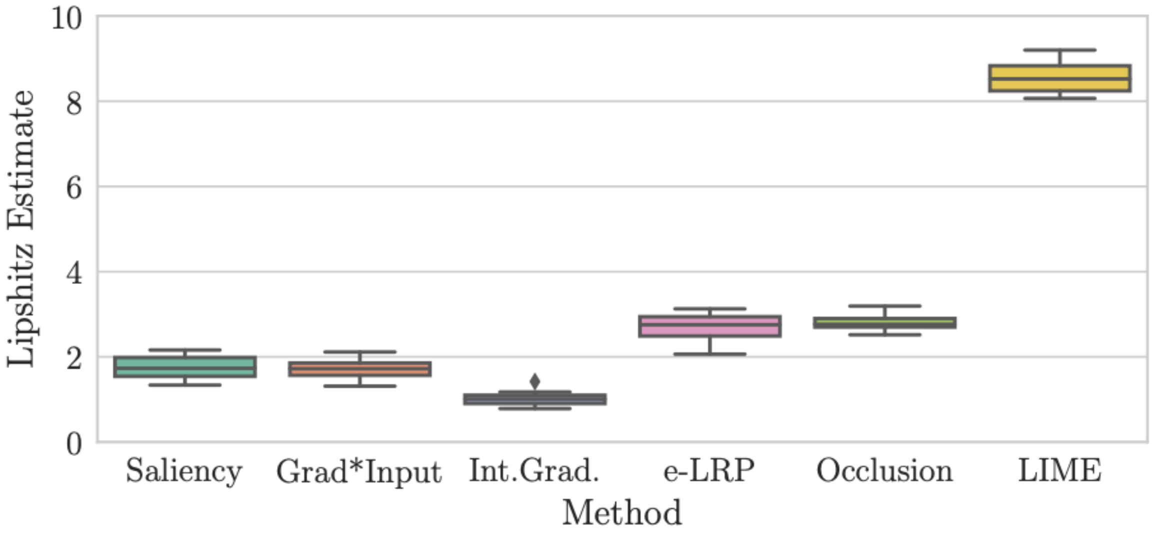
hyperparameters

Alvarez-Melis, David et al. "On the robustness of interpretability methods." WHI, ICML, 2018.

Limitations: **Stability**

Are post-hoc explanations unstable wrt small non-adversarial input perturbation?

- Perturbation approaches like LIME can be unstable.



Estimate for 100 tests for an MNIST Model.

Alvarez-Melis, David et al. "On the robustness of interpretability methods." WHI, ICML, 2018.

Limitations: Stability – Problem is Worse!

(a

Problem with having too few perturbations?
If so, what is the optimal number of perturbations?

When you repeatedly run LIME on the same instance, you get different explanations (blue region)

Limitations of Explanation Methods

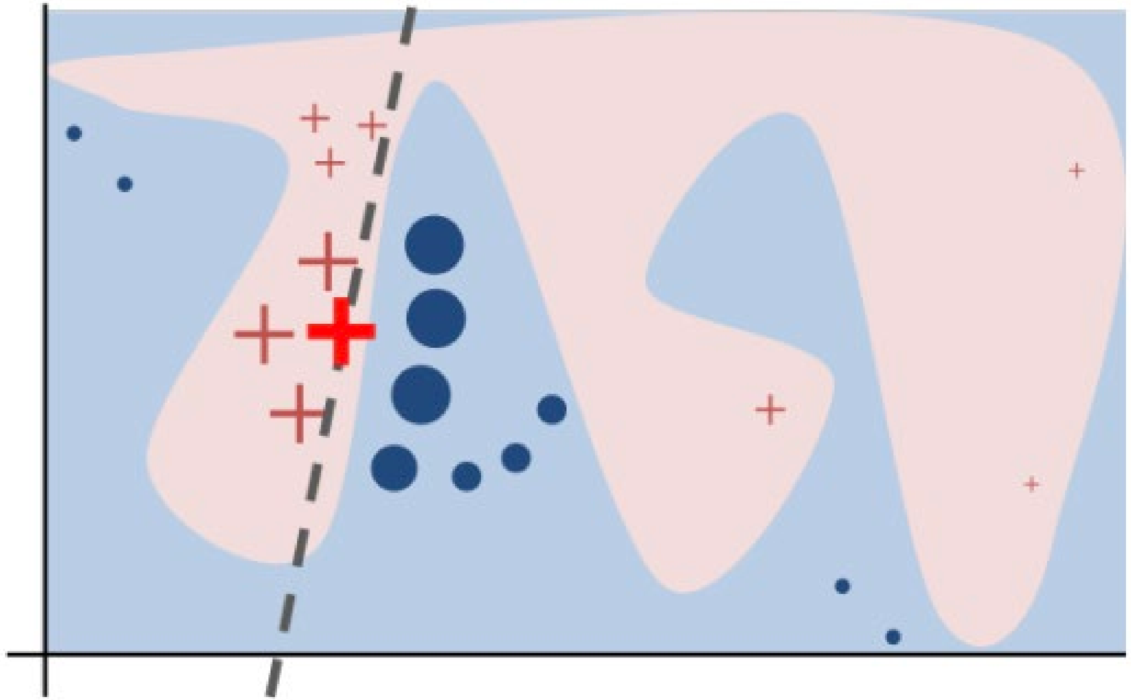
Faithfulness
Some explanation methods do not ‘reflect’ the underlying model.

Stability
Slight changes to inputs can cause large changes in explanations.

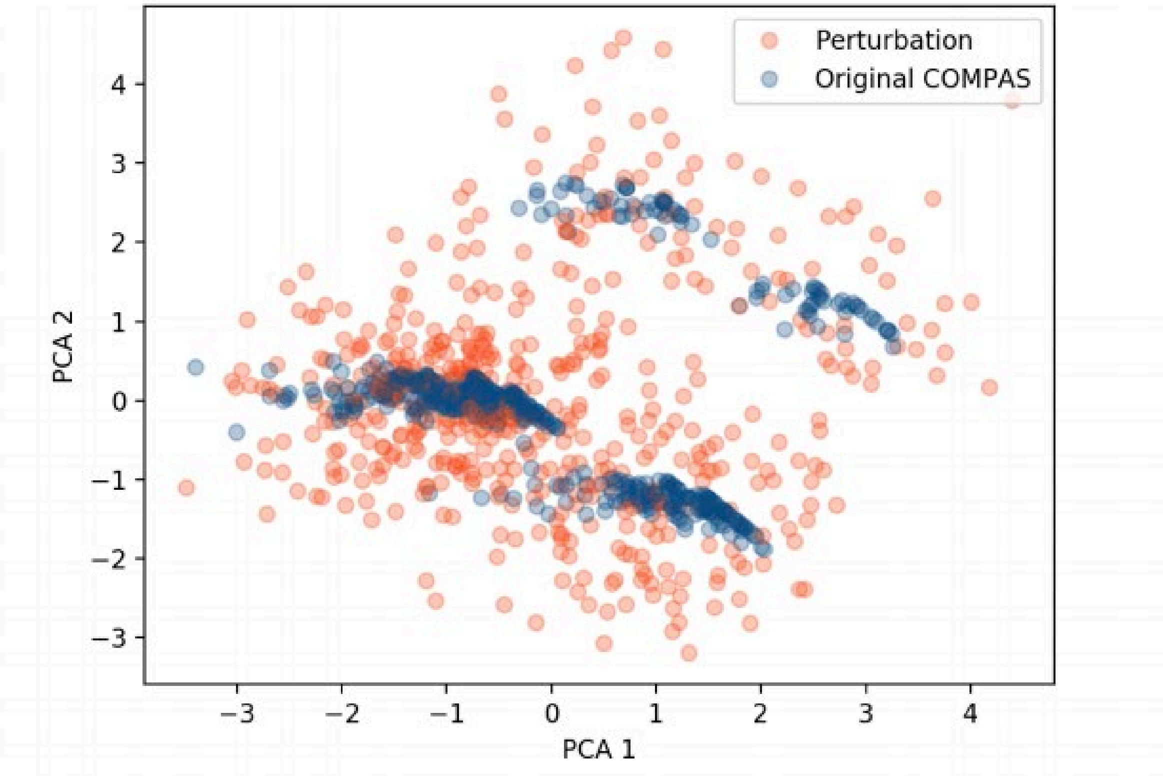
Fragility
Post-hoc explanations can be easily manipulated.

Limitations: Fragility

- LIME

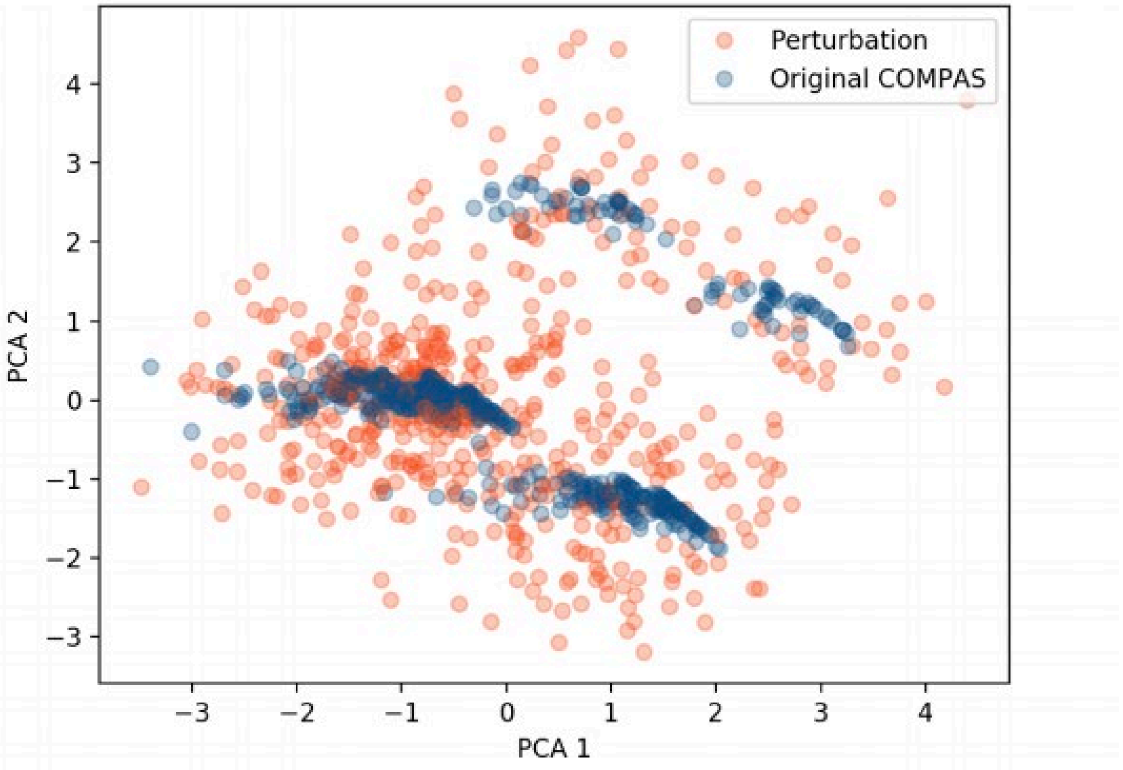


Vulnerabilities of LIME: Intuition



Several perturbed data points are out of distribution (OOD)!

Vulnerabilities of LIME: Intuition



Adversaries can exploit this and build a classifier that is biased on in-sample data points and unbiased on OOD samples!

Building Adversarial Classifiers

- **Setting:**

  - Adversary wants to deploy a biased classifier f in real world.
    - E.g., uses only race to make decisions

  - Adversary must provide black box access to customers and regulators who may use post hoc techniques (GDPR).

  - *Goal of adversary is to fool post hoc explanation techniques and hide underlying biases of f*

Building Adversarial Classifiers

- **Input:** Adversary provides us with the biased classifier *f*, an input dataset $\boldsymbol{X}$ sampled from real world input distribution $\boldsymbol{X}_{\text{dist}}$

- **Output:** Scaffolded classifier *e* which behaves exactly like f when making predictions on instances sampled from $\boldsymbol{X}_{\text{dist}}$ but will not reveal underlying biases of f when probed with perturbation-based post hoc explanation techniques.
  - *e* is the adversarial classifier

Building Adversarial Classifiers

- Adversarial classifier e can be defined as:

$$e(x) = \begin{cases} f(x), & \text{if } x \in \mathcal{X}_{dist} \\ \psi(x), & \text{otherwise} \end{cases}$$

- f is the biased classifier input by adversary.

- ψ is the unbiased classifier (e.g., only uses features uncorrelated to sensitive attributes)

Building Adversarial Classifiers: OOD Detection

- We perturb each data point in the dataset $\boldsymbol{X}$
- Each data point in $\boldsymbol{X}$ is labeled ”not OOD”
- Each data point generated by perturbation is labeled “OOD” unless it is very close to some data point in $\boldsymbol{X}$

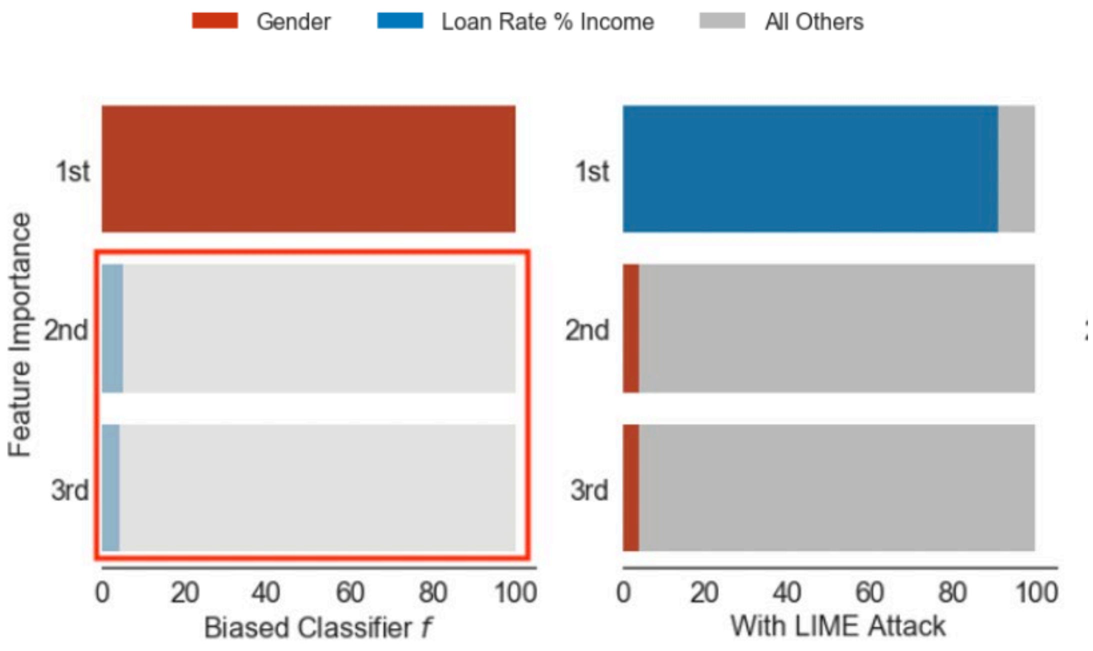
Experimental Evaluation: Data & Setting

Dataset	Size	Features	Positive Class	Sensitive Feature
COMPAS	6172	criminal history, jail and prison time, demographics, COMPAS risk score	High Risk (81.4%)	African-American (51.4%)
Communities and Crime	1994	race, age, education, marriage status, citizenship, police demographics	Violent Crime Rate (50%)	White Population (continuous)
German Credit	1000	account information, credit history, loan purpose, employment, demographics	Good Customer (70%)	Male (69%)

- Standard implementations of LIME/SHAP

- unbiased classifier: we either leverage synthetic feature(s) or existing feature(s) both of which are uncorrelated with sensitive attribute.

Evaluating the Effectiveness of Attacks



German Credit Dataset– LIME

Tutorial Outline

- Motivation
- Interpretability vs. Explainability
- Overview of Explanation Methods
- Limitations of Explanation Methods
- **The Road Ahead**

The Road Ahead

- Explainability as a technology is fragile; Research is in progress
- Improving the Reliability of Explanations
- Developing Evaluation Frameworks for Explanations
- Focusing on the Scalability of Explanation Methods

Thank You!

- Email: hlakkaraju@hbs.edu; hlakkaraju@seas.harvard.edu;
- Course on interpretability and explainability: https://interpretable-ml-class.github.io/
- Trustworthy ML Initiative: https://www.trustworthyml.org/
  - Lots of resources and seminar series on topics related to explainability, fairness, adversarial robustness, differential privacy, causality etc.

RObust & Stable Post hoc Explanations (ROPE)

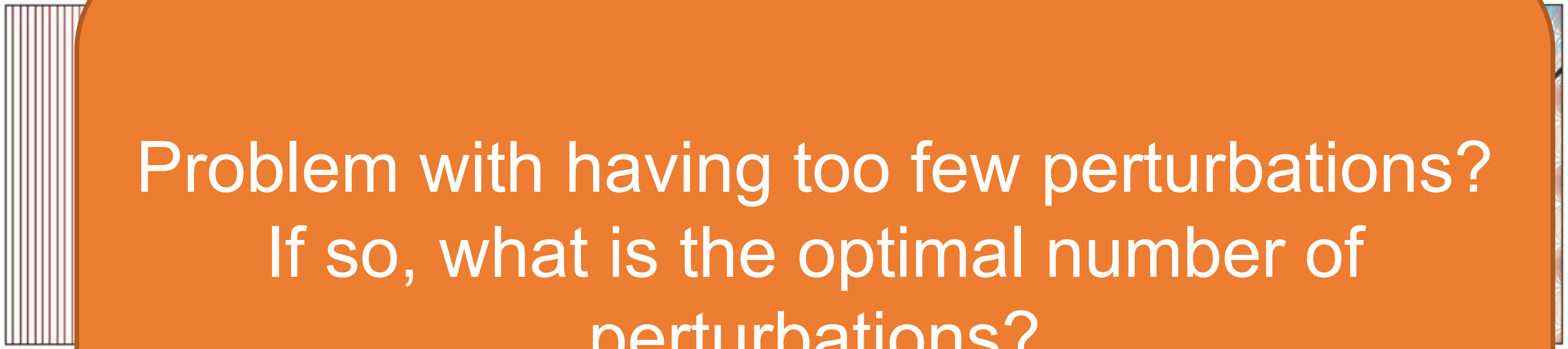
- Framework for generating explanations that are stable and robust to distribution shifts

- It is flexible, e.g., it can be instantiated for linear vs. rule based explanations

$$\hat{E} = \arg\min_{E \in \mathcal{E}} \underbrace{\max_{\delta \in \Delta}}_{\text{worst-case computed over plausible distribution shifts}} \underbrace{\mathbb{E}_{p_\delta(x)}[\ell(E(x), B^*(x))]}_{\text{expected gap between explanation and black box}}.$$

[Slack et. al., 2020]

Limitations: Stability – Problem is Worse!

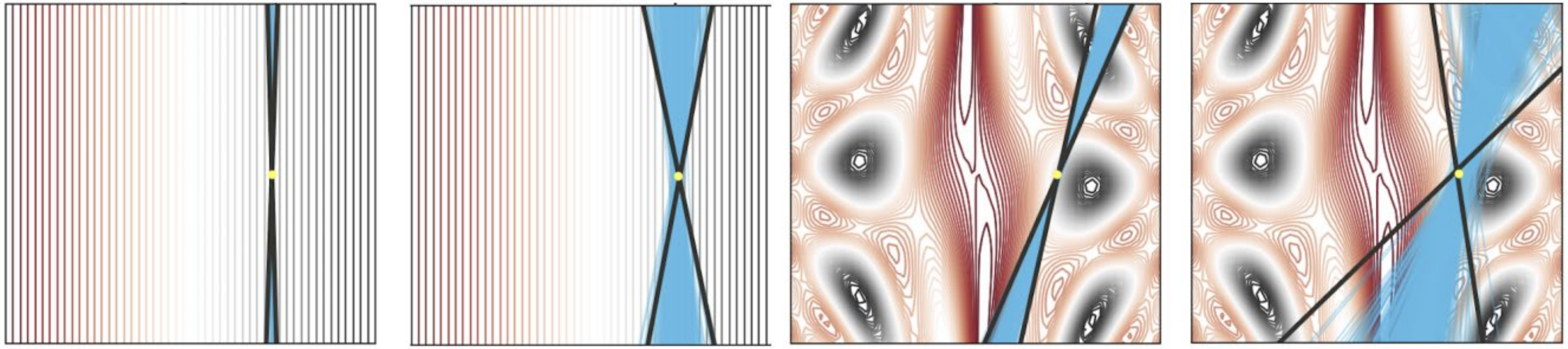


(a

Problem with having too few perturbations?
If so, what is the optimal number of perturbations?

When you repeatedly run LIME on the same instance, you get different explanations (blue region)

Modeling Uncertainty of Black Box Explanations: BayesLIME & BayesSHAP



(a) linear, many samples (b) linear, fewer samples (c) nonlinear, many samples (d) nonlinear, fewer samples

BayesLIME 95% Confidence Interval Shown by Black Lines

Modeling Uncertainty of Black Box Explanations: BayesLIME & BayesSHAP

$$y|z,\phi,\sigma^2 \sim \phi^T z + \underbrace{\mathcal{N}(0, \frac{\sigma^2}{\pi_x(z)})}_{\epsilon}, \quad \forall z \in \mathcal{Z}$$

$$\phi|\sigma^2 \sim \mathcal{N}(\phi_0, \sigma^2\Sigma_0) \qquad \sigma^2 \sim \textbf{Inv-}\chi^2(n_0, \sigma_0^2).$$

Proximity function

Feature importances

- No need to resort to MCMC or VI; Closed form solutions