Post Hoc Explanations For ML

Exposing Shortcomings and Improving Reliability

Dylan Slack
UC Irvine

2022

Goals

Introduce Explanations

What are post hoc explanations, and why should you care?

Where do they fall short?

What are the issues with these methods?

How can we do better?

How are we working to fix these issues?

Collaborators

Sophie Hilgard

Hima Lakkaraju

Sameer Singh

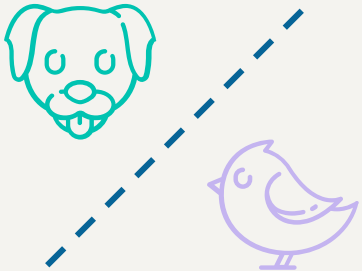
Emily Jia

With Support From

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

My Awesome Classifier Idea

Let's build a model to classify birds and dogs!

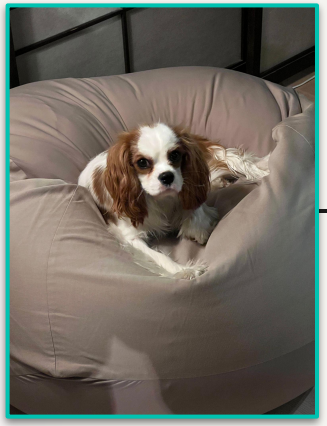


Bird or Dog?

Deep Neural Network

Bird!

Bird or Dog?

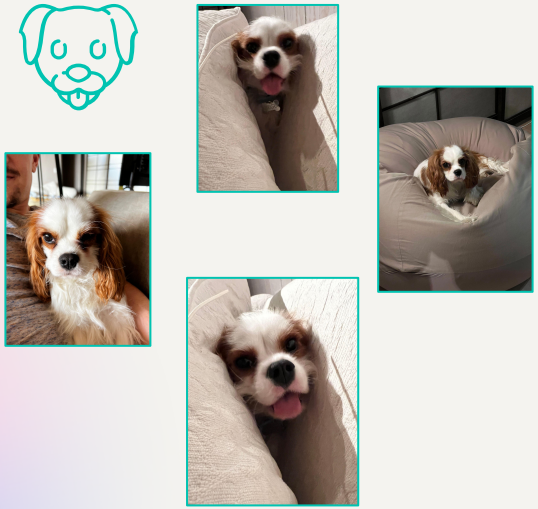


Deep Neural Network

Dog!

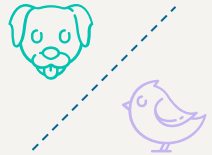
Let's Build a Model

01 *Collect Data*

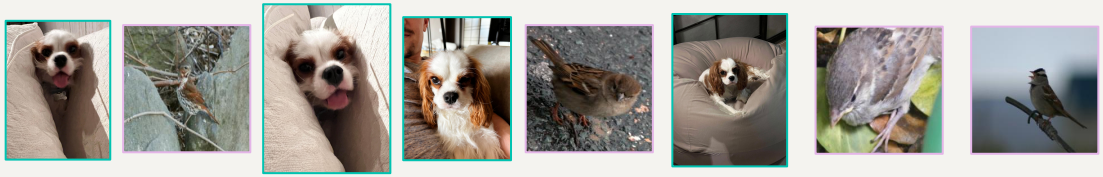


2022

Let's Build a Model



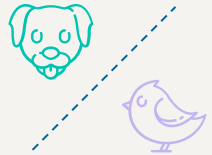
02 *Train Model*



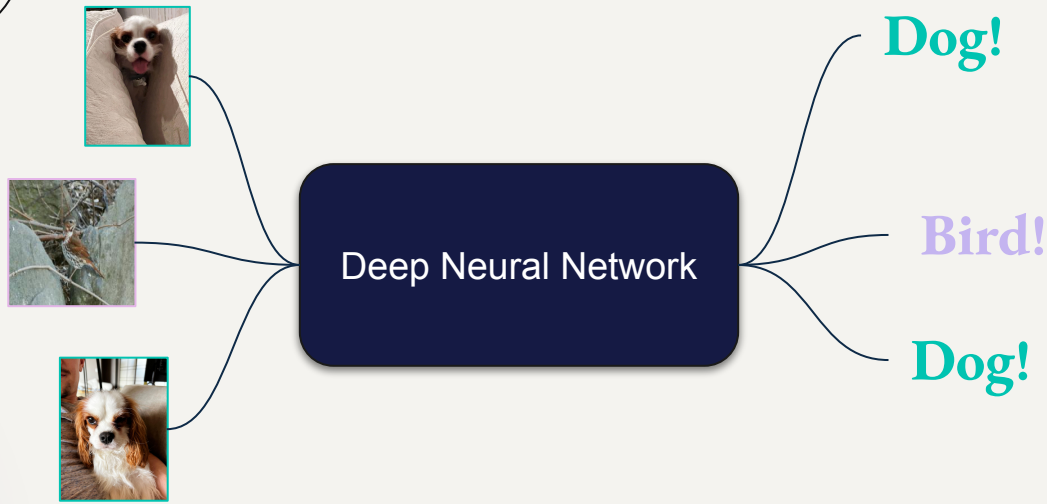
Deep Neural Network

2022

Let's Build a Model



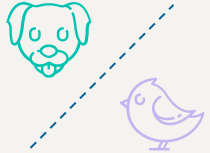
03 *Predict!*



Awesome!

2022

We've built a black-box!



2022

Trust

Does it make decisions for the right reasons?

Understanding

Do I know why this model is making decisions?

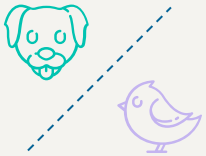
Fixing

If something goes wrong, can I fix it?

Failures abound

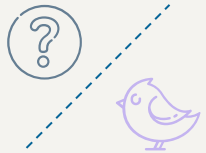
Deep Neural Network

Bird!

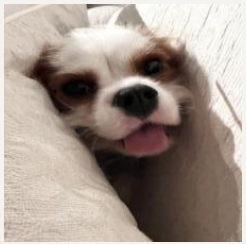
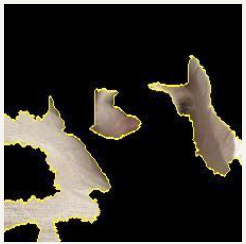
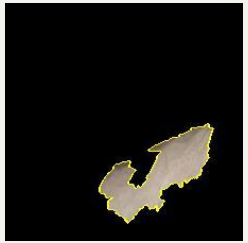
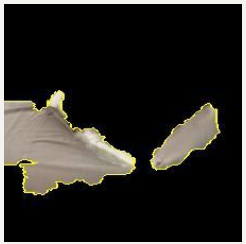


Why is the model bad in-the-wild?

Post hoc Explanations



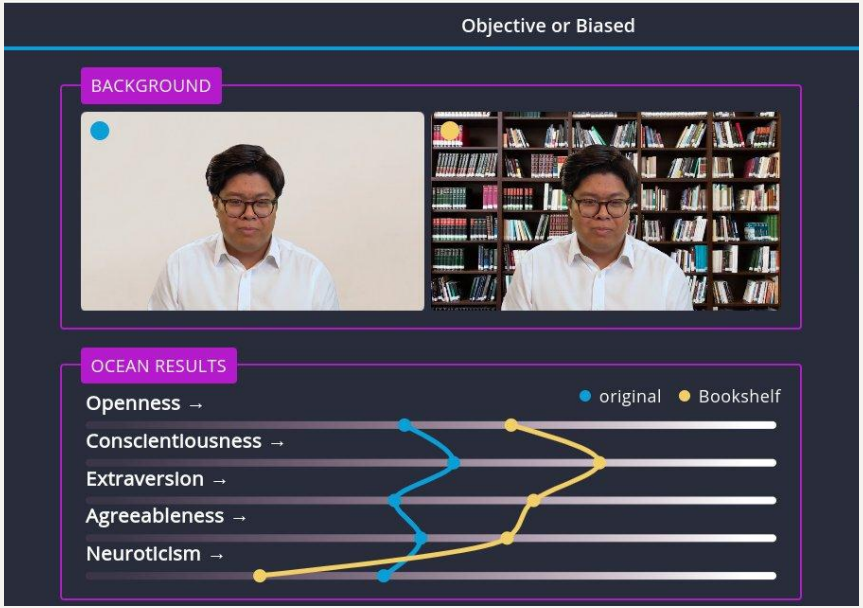
- Show important parts of the image for prediction

Image	Explanation	Image	Explanation
			
			

Uh oh, looks like we've built a couch detector!

In the Real World

- AI interview system uses image background



[Source: Washington Post]

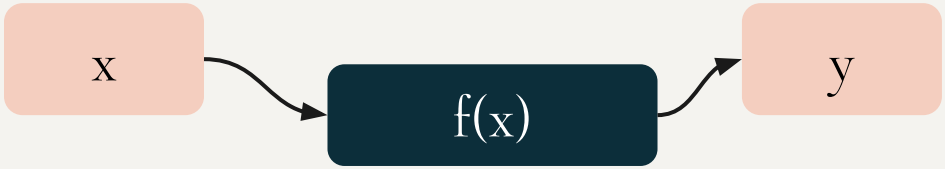
Post hoc Explanations

What are they, and why care?

2022

Explaining Predictions

What parts of the data are most responsible for predictions?



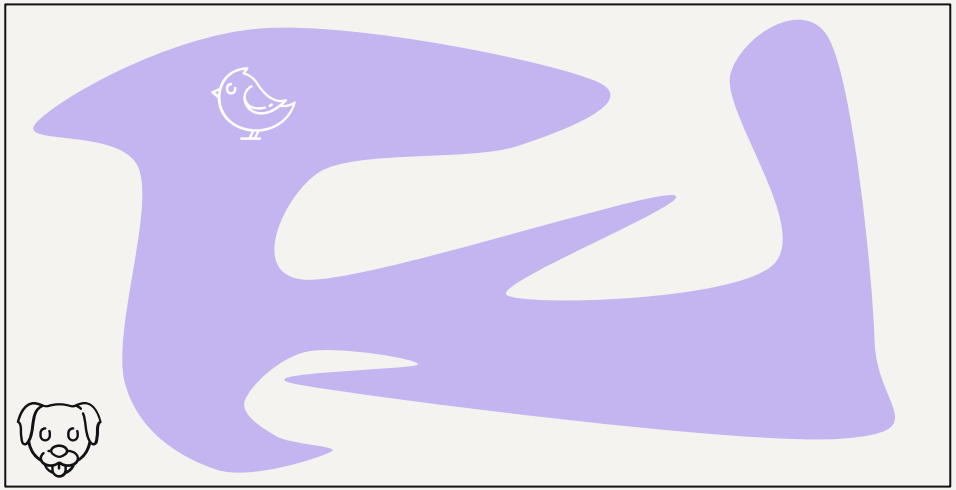
x could be:

"The quick brown fox jumps over the lazy dog"

Income:	75,000
Age:	32
Credit Score:	720

Model Agnostic Local Explanations

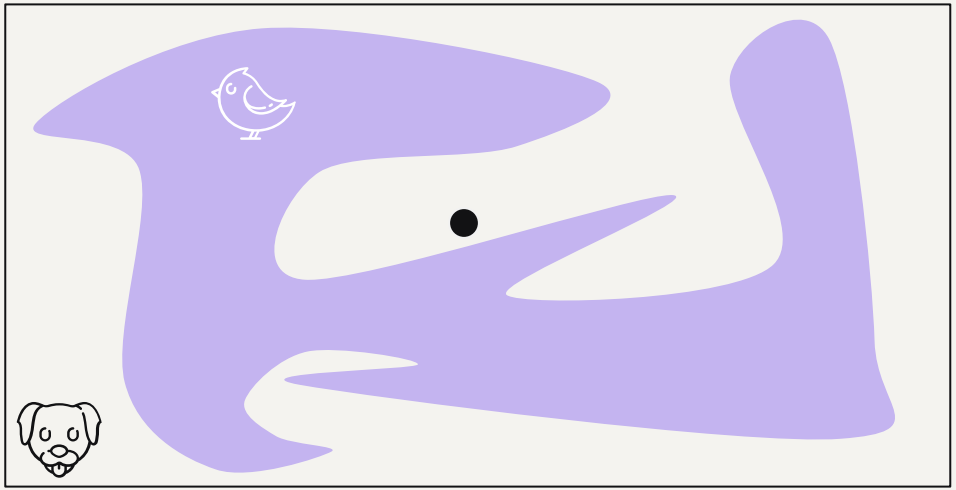
Complex, global decision surface



Difficult to explain entire decision surface

Model Agnostic Local Explanations

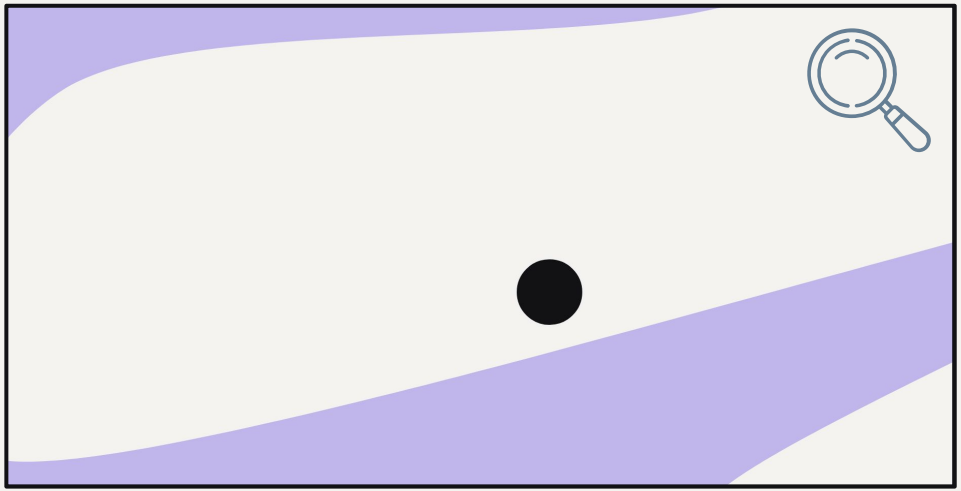
Complex, global decision surface



Instead give locally accurate explanation for a single point

Model Agnostic Local Explanations

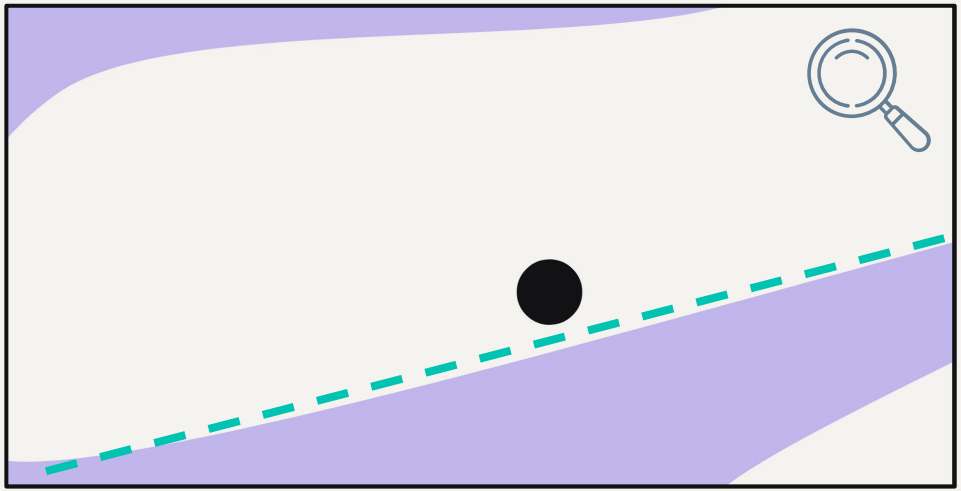
Complex, global decision surface



Instead give locally accurate explanation for a single point

Model Agnostic Local Explanations

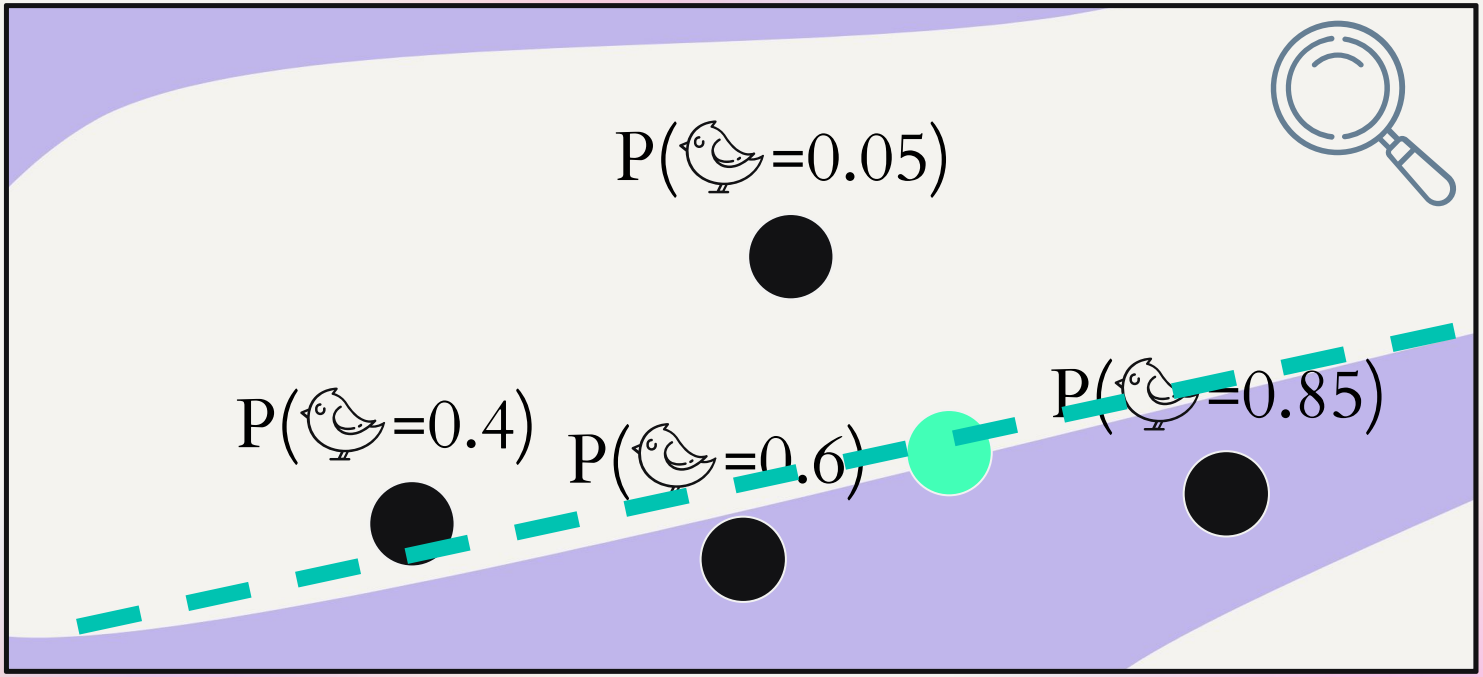
Complex, global decision surface



Local explanation is interpretable, locally accurate model

Model Agnostic = We can run explanations for any model!

We can do this for any model!



LIME

Perturbations

Class Probability

Original Image

Class Probability **0.8**

Classifier

0.7

Classifier

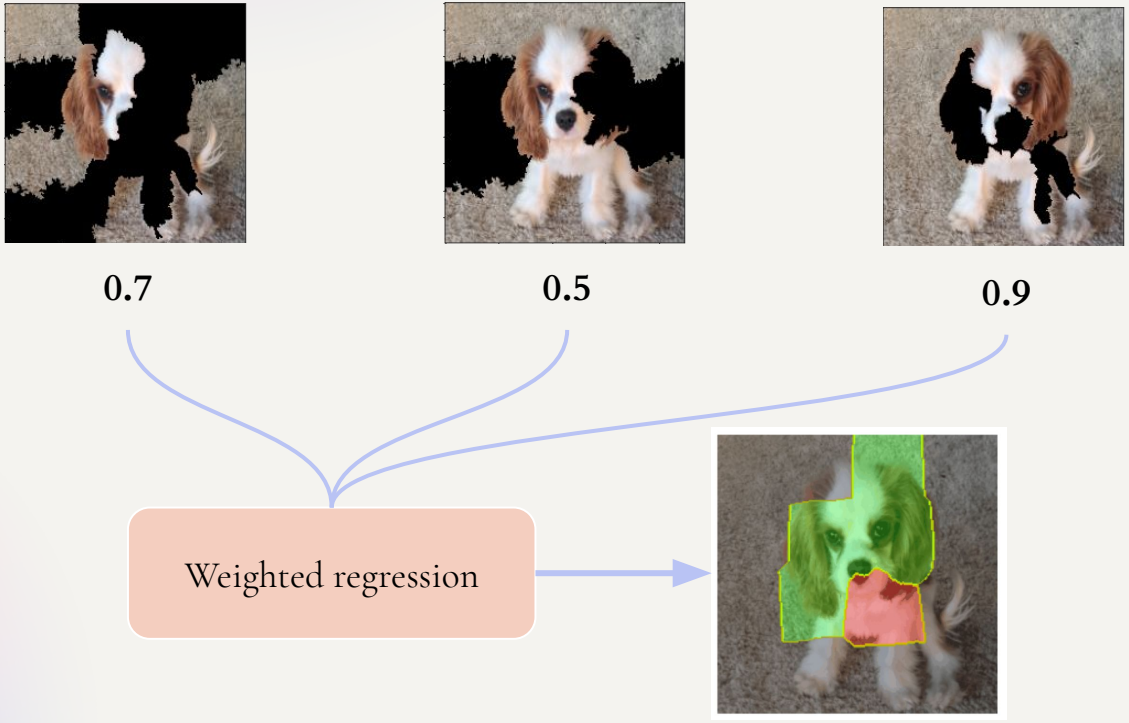
0.5

Classifier

0.9

2022

LIME



2022

Tradeoffs of Local Model Agnostic Methods

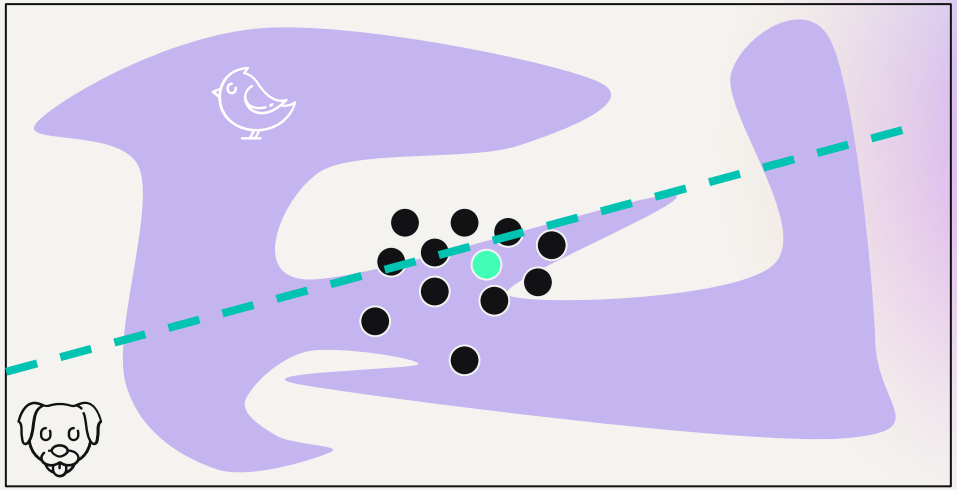
Advantages

- Only needs black-box access!
- Easy to use
- Highly flexible
  - Perturbations, locality, precision can all be customized
- Helpful for understanding model behavior

Tradeoffs of Local Model Agnostic Methods

Disadvantages

Explanations are highly sensitive to the perturbations



Tradeoffs of Local Model Agnostic Methods

Consequences

- Unstable
- Have difficult to set hyperparameters
- Hard to determine when you have a “good” explanation

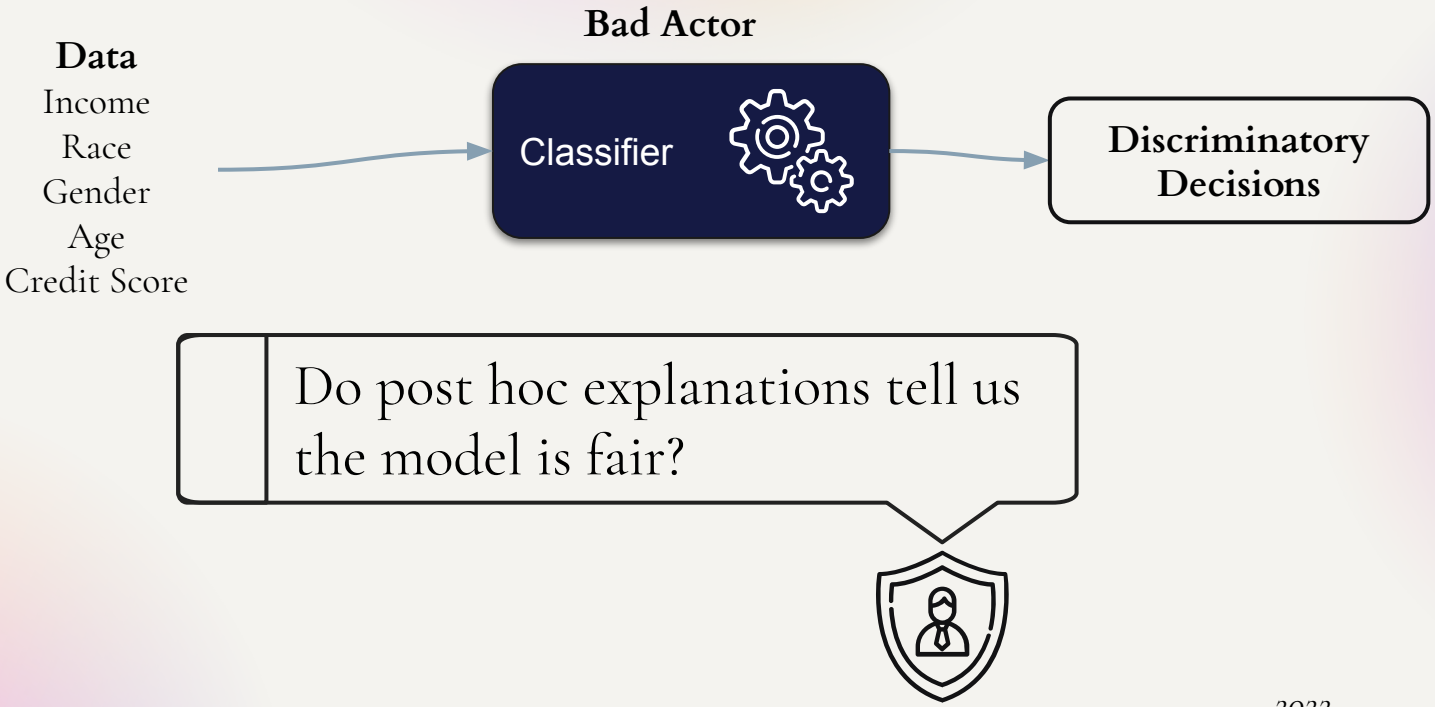
Unclear when to trust explanations

Can We Fool Post Hoc Explanations?

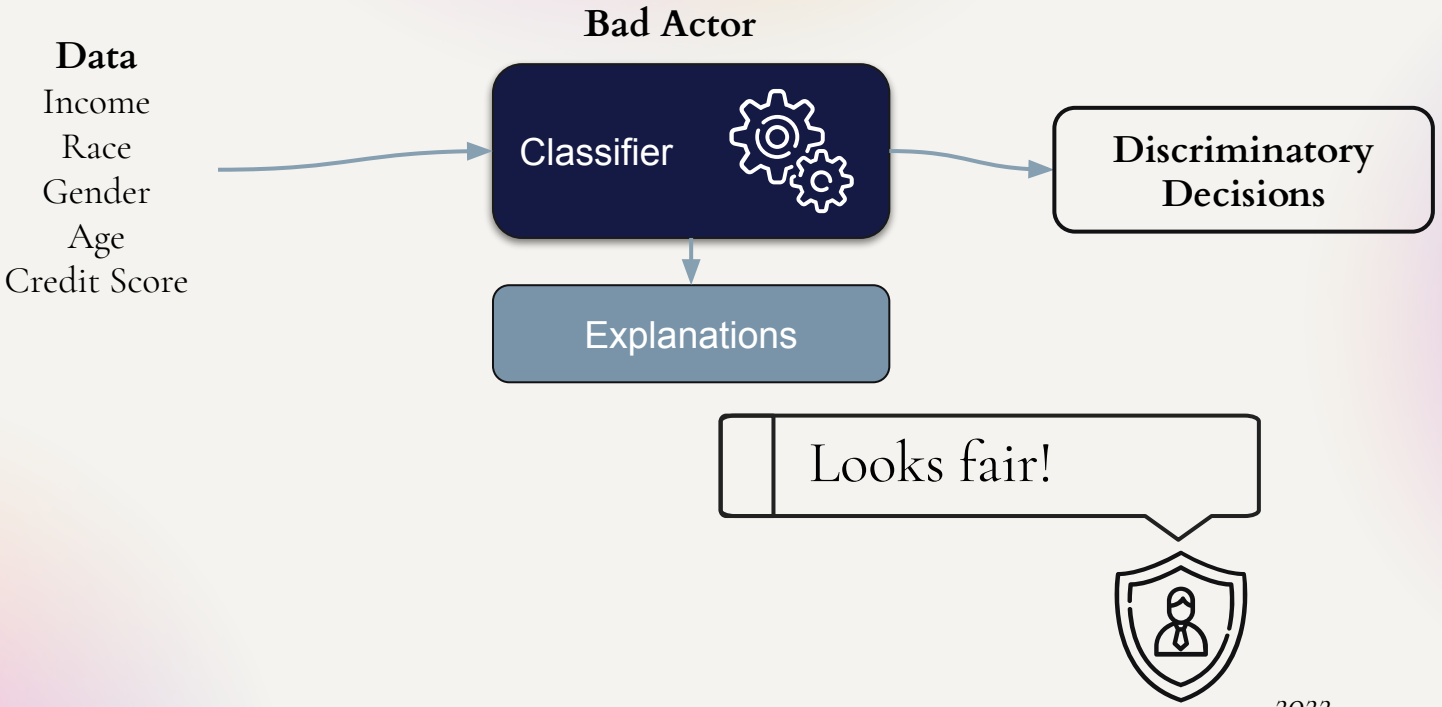
Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods
Dylan Slack*, Sophie Hilgard*, Emily Jia, Sameer Singh, and Hima Lakkaraju
AIES 2020

2022

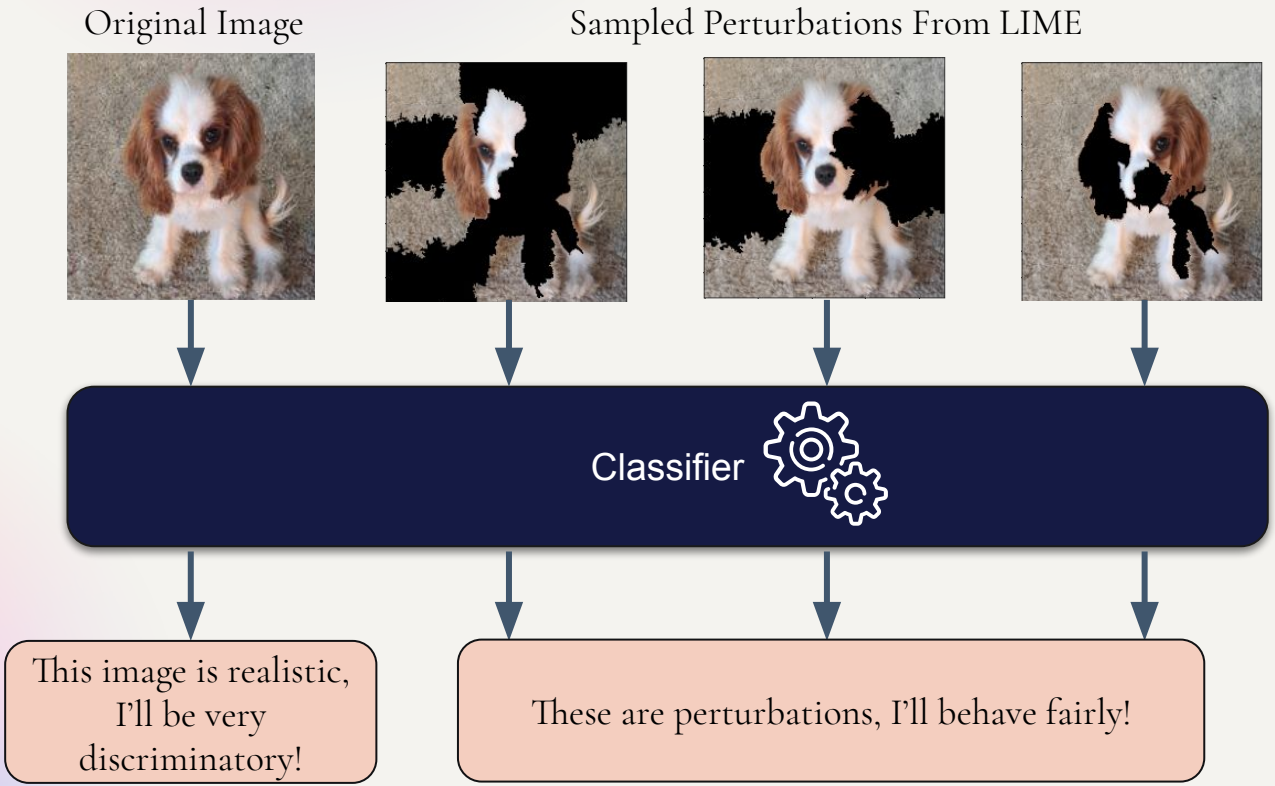
Setup


Bad Actor
Data
Income
Race
Gender
Age
Credit Score
Classifier
Discriminatory Decisions
Do post hoc explanations tell us the model is fair?

Setup


Bad Actor
Data
Income
Race
Gender
Age
Credit Score
Classifier
Discriminatory Decisions
Explanations
Looks fair!

How is this possible?



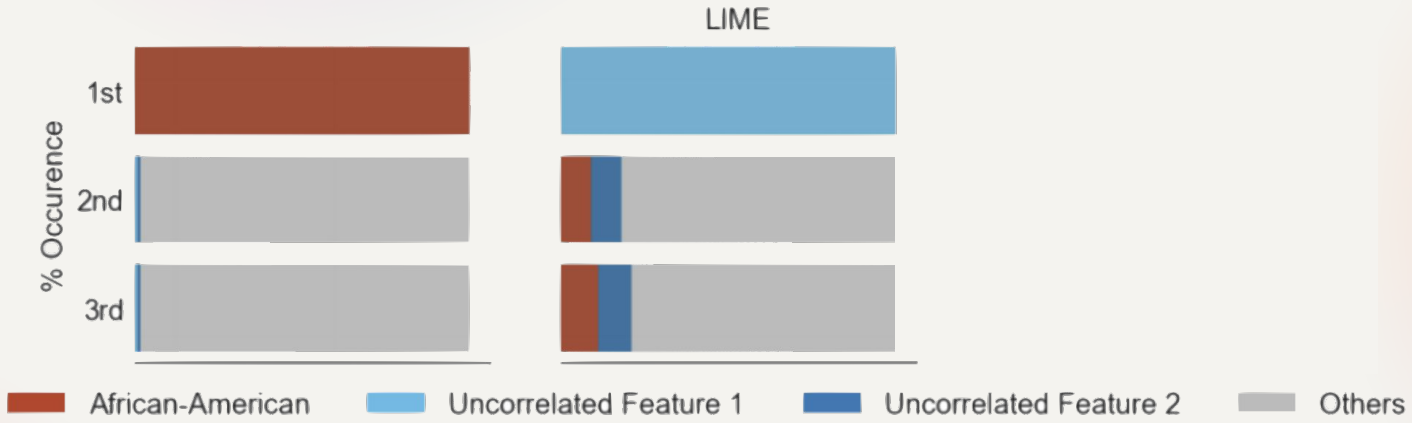
2022

Compas Recidivism Dataset

The model only uses the “race” feature!

2022

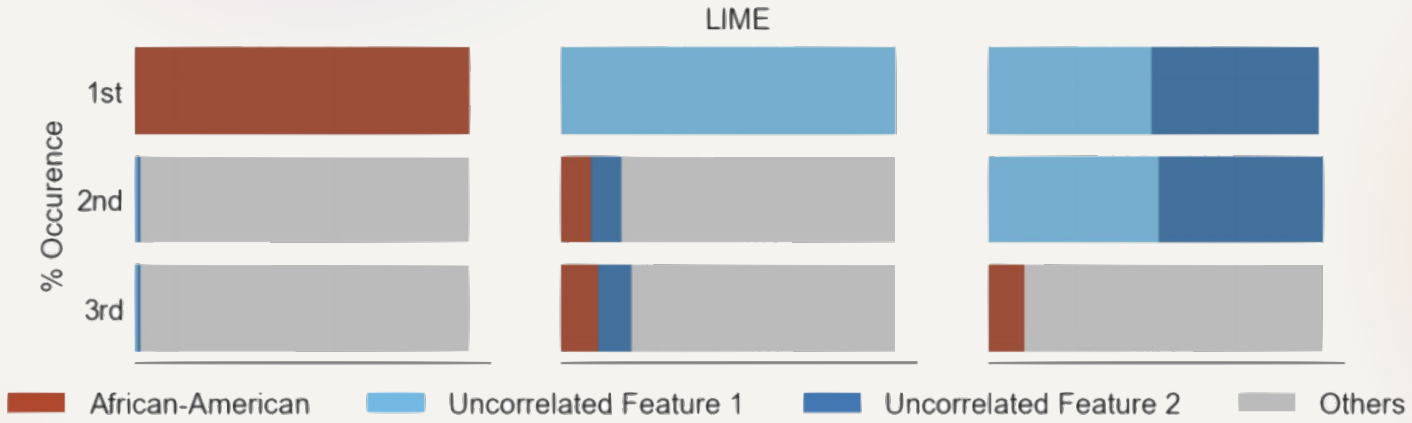
Compas Recidivism Dataset



The explanations don't show race is important!

2022

Compas Recidivism Dataset

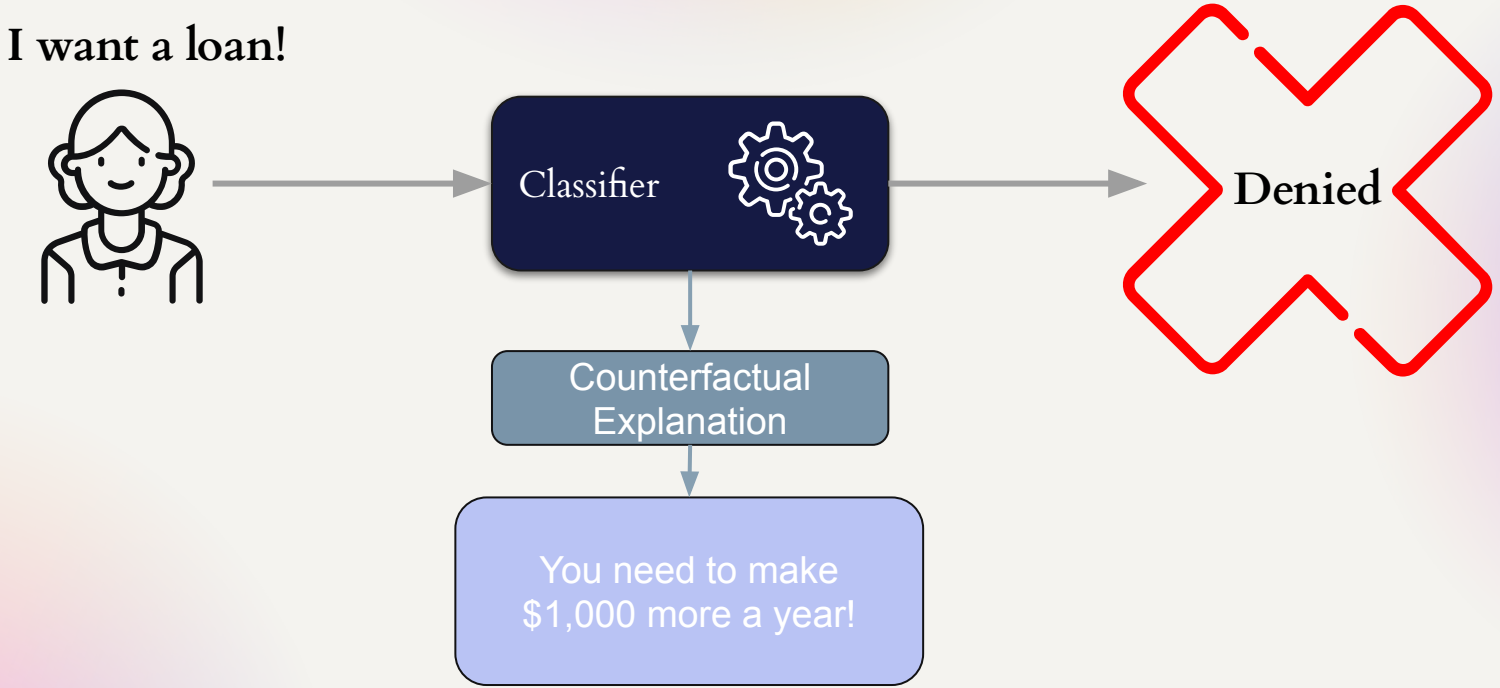


The explanations don't show race is important!

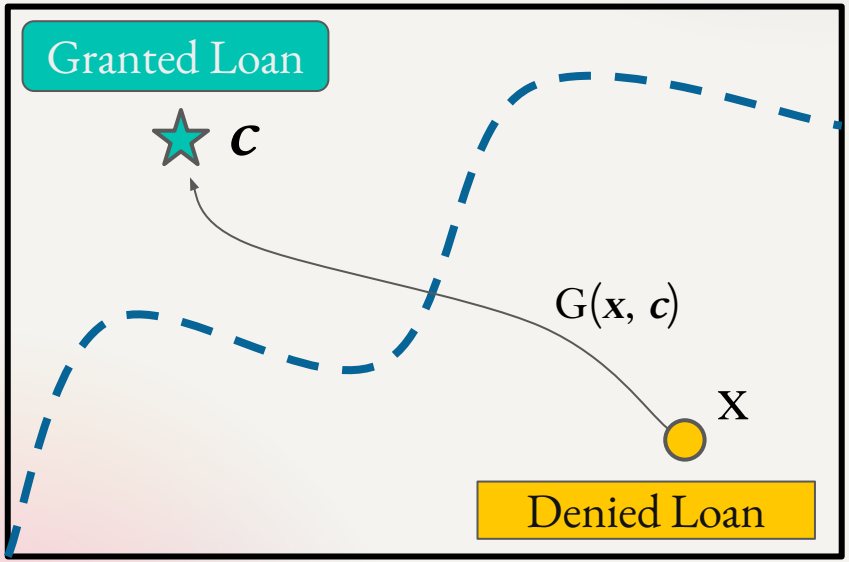
2022

Other types of explanations can be *attacked*

Counterfactual Explanations



How CFE's Work!

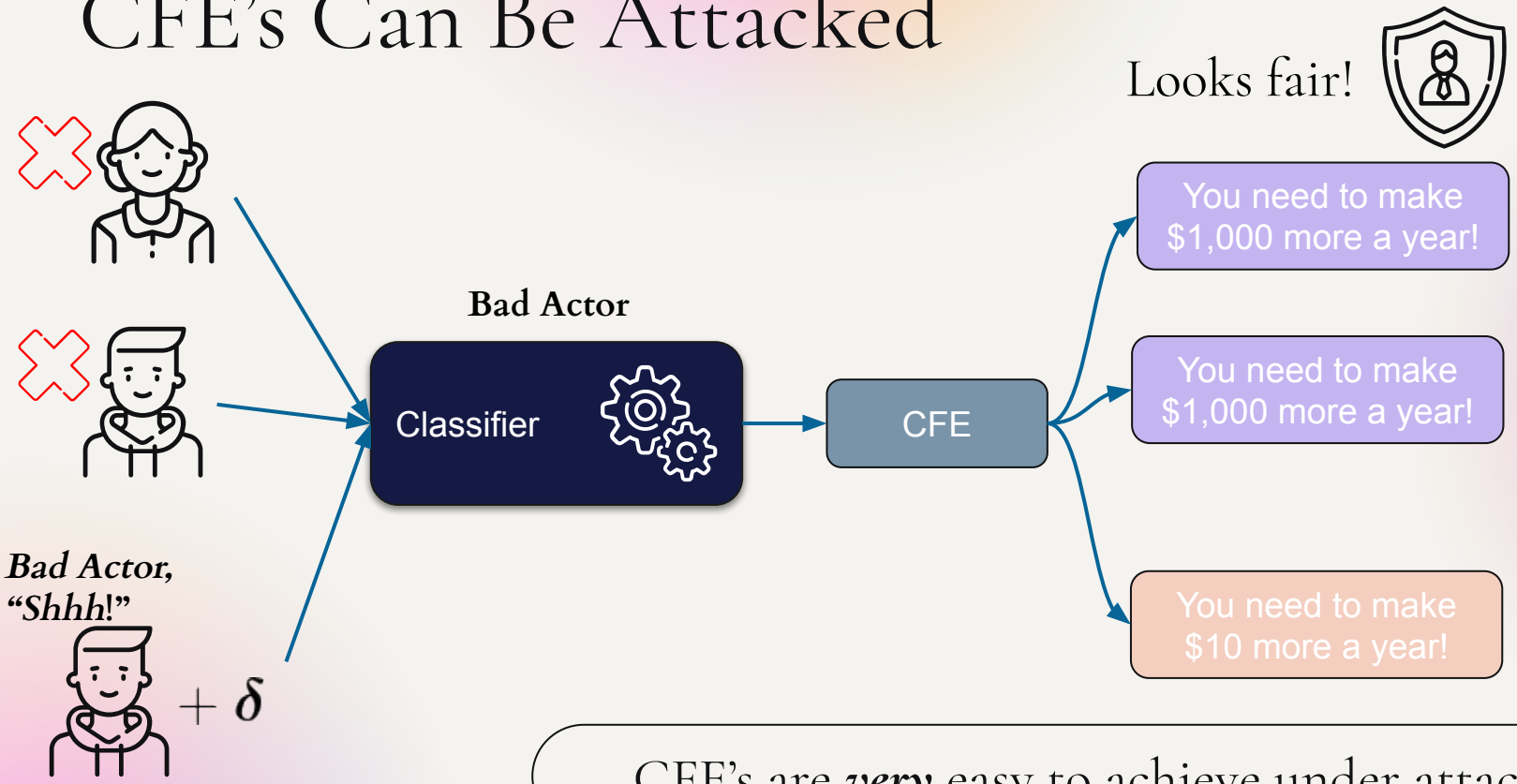


Objective: G(***x***, ***c***)

Maximize: ***c*** is in desired class

Minimize: Difficulty of ***c***

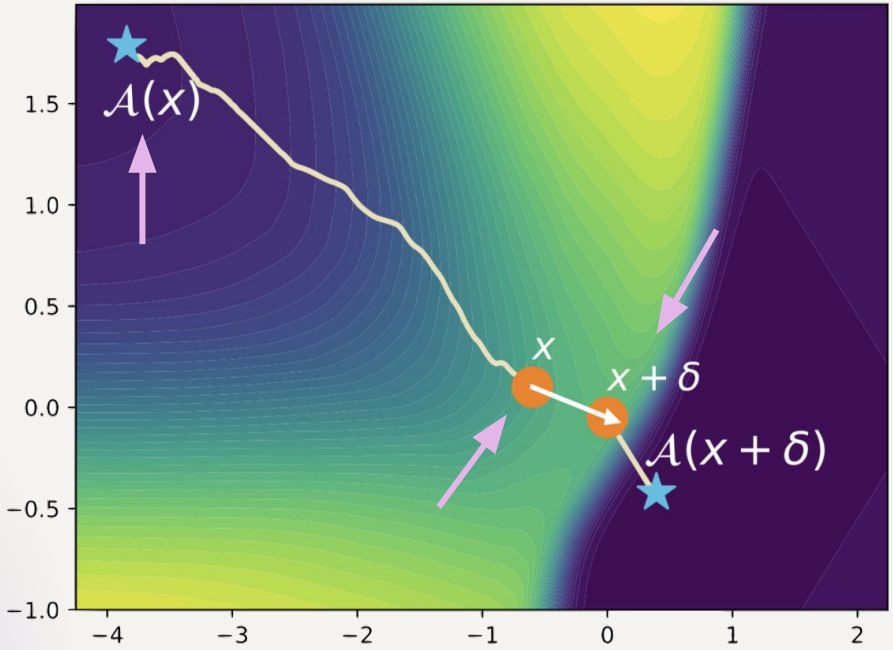
CFE's Can Be Attacked



CFE's are ***very*** easy to achieve under attack!

Why Does This Attack Work?

Key Idea 🔑 : CFE search converges to different local minimums



Counterfactual Explanations Can Be Manipulated

Dylan Slack, Sophie Hilgard, Hima Lakkaraju, and Sameer Singh

NeurIPS 2021

Building More Reliable Explanations

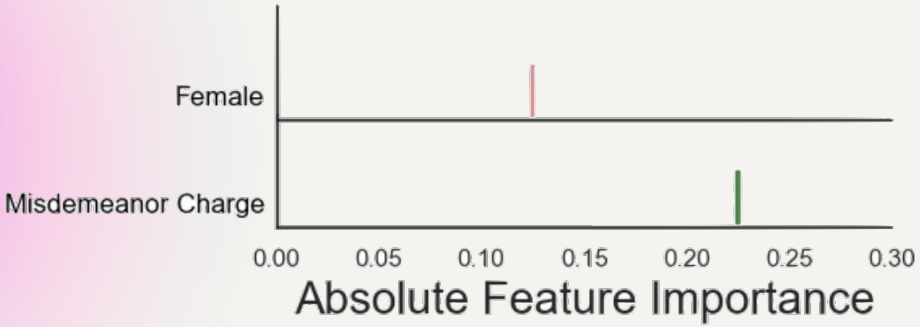
Reliable Post hoc Explanations: Modeling Uncertainty in Explainability
Dylan Slack*, Sophie Hilgard, Sameer Singh, and Hima Lakkaraju
NeurIPS 2021

2022

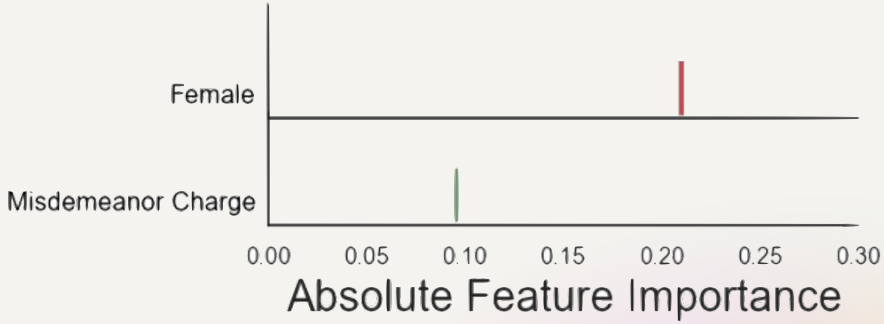
Highlighting Issues With LIME

Two explanations on the same instance, different hyperparameters

Red: Negative Contribution
Green: Positive Contribution



200 Perturbations

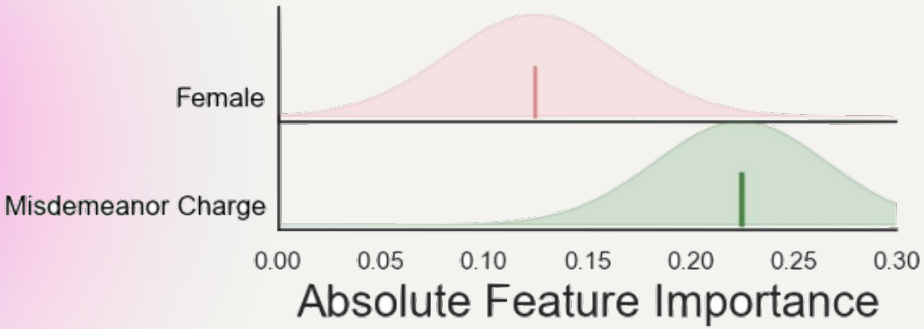


2,000 Perturbations

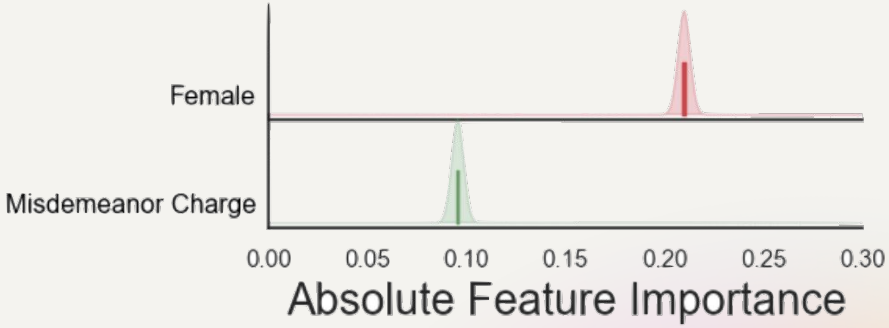
Uncertainty is the Solution!

Define feature importances as distributions.

Red: Negative Contribution
Green: Positive Contribution



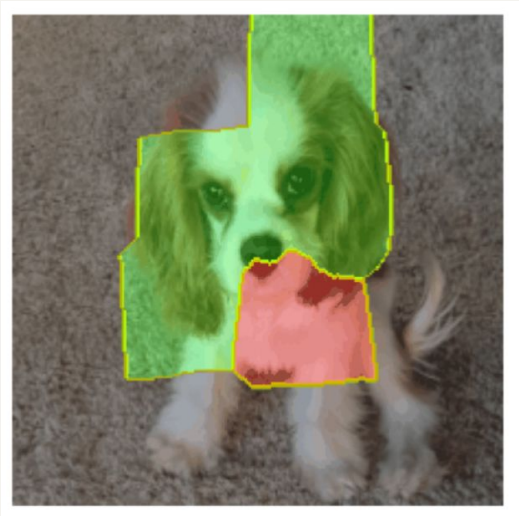
200 Perturbations



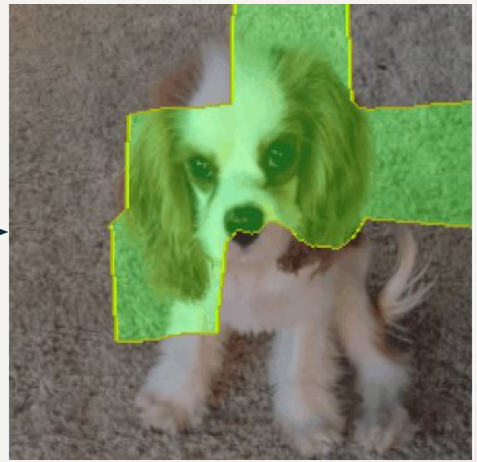
2,000 Perturbations

Visualizing Uncertainty In Explanations

LIME



BayesLIME

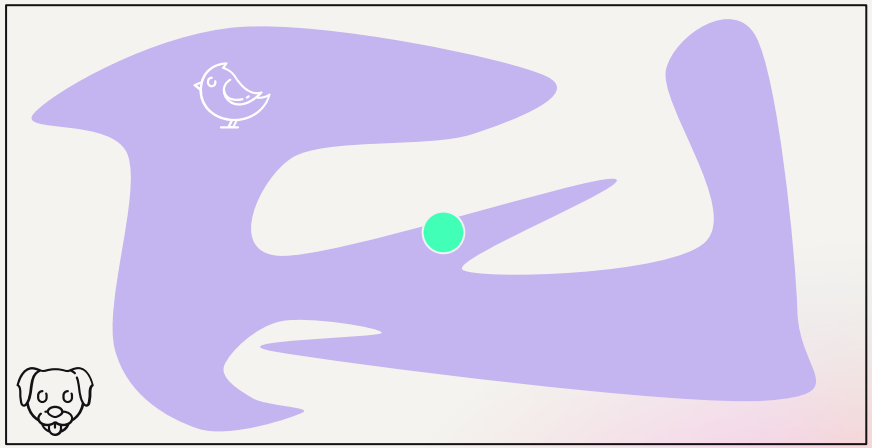


Formulation

$$y|z, \phi, \epsilon \quad \sim \quad \phi^T z$$

ϕ : Feature Importance

z : Perturbations

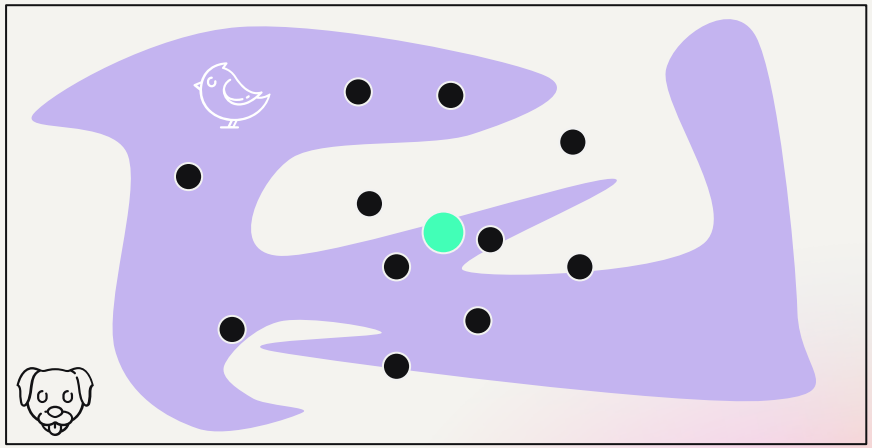


Formulation

$$y|z, \phi, \epsilon \quad \sim \quad \phi^T z$$

ϕ : Feature Importance

z : Perturbations



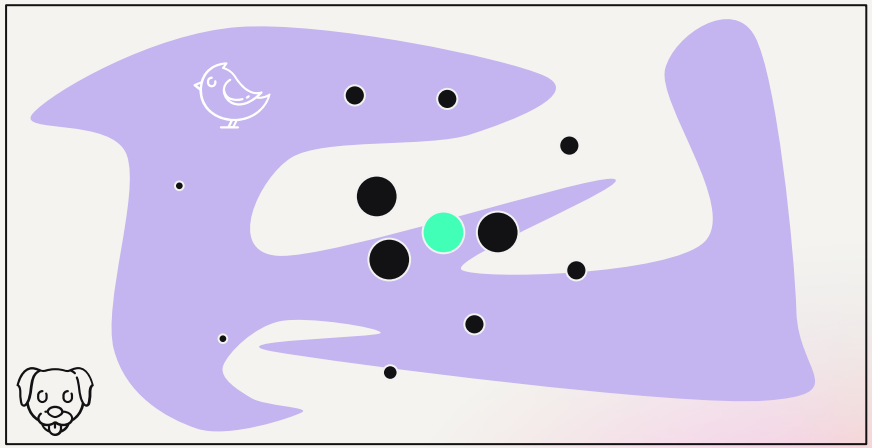
Formulation

$$y|z, \phi, \epsilon \sim \phi^T z + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \frac{\sigma^2}{\pi_x(z)})$$

ϕ: Feature Importance

z: Perturbations

$\pi_x(z)$: Weighting Function



Formulation

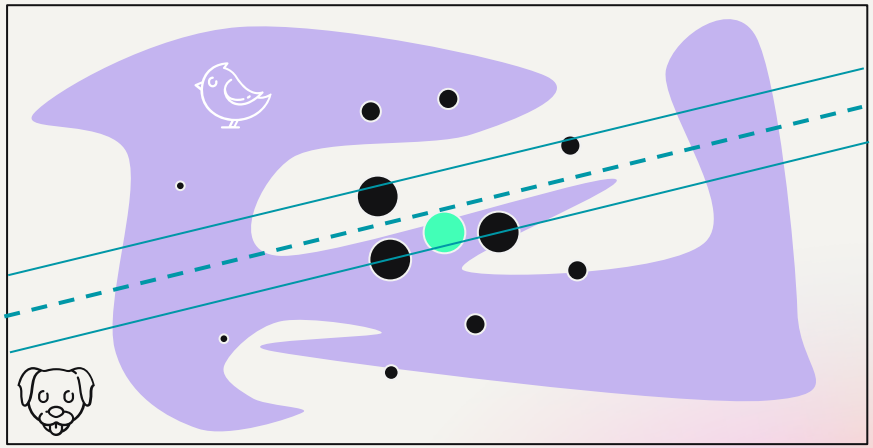
$$y|z, \phi, \epsilon \sim \phi^T z + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \frac{\sigma^2}{\pi_x(z)})$$

$$\phi|\sigma^2 \sim \mathcal{N}(0, \sigma^2 \mathbb{I}) \qquad \sigma^2 \sim \text{Inv-}\chi^2(n_0, \sigma_0^2).$$

ϕ: Feature Importance

z: Perturbations

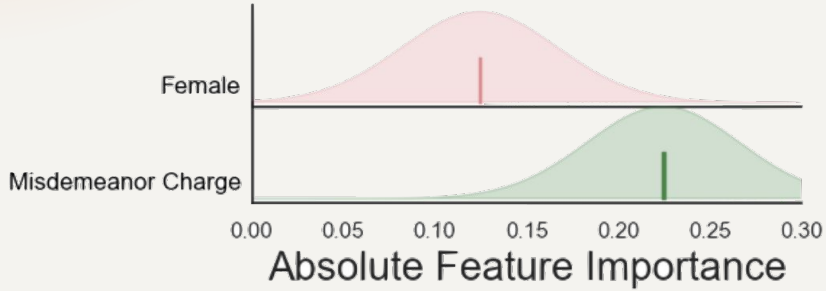
$\pi_x(z)$: Weighting Function



Notions of Uncertainty

Feature Importance Uncertainty

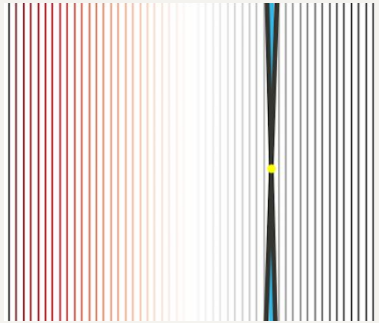
Goes to zero with sufficient perturbations



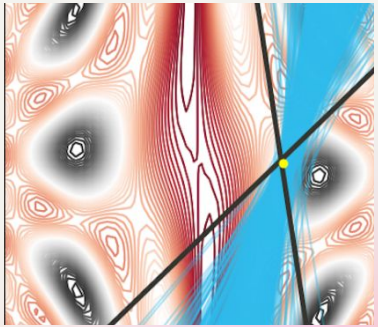
Error Uncertainty

Converges to the error of the explanation

Linear, low error uncertainty



Nonlinear, higher error uncertainty



Uncertainty is Calibrated

Data set	Calibration *Closer to 95.0 is better*	
	BayesLIME	*BayesSHAP*
ImageNet	94.8	89.9
MNIST	97.2	90.1
COMPAS	95.5	87.9
German Credit	96.9	89.6

Estimating Required Number of Perturbations (PTG)

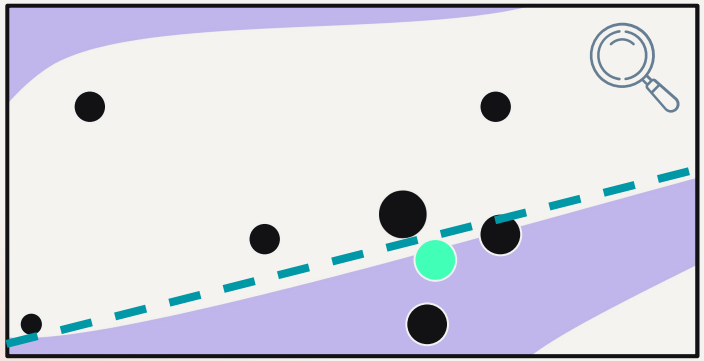
"Get me an explanation with 95% credible interval width of 0.01 for this image!"

Perturbations-to-go (PTG)

You need to use 7,634 perturbations!

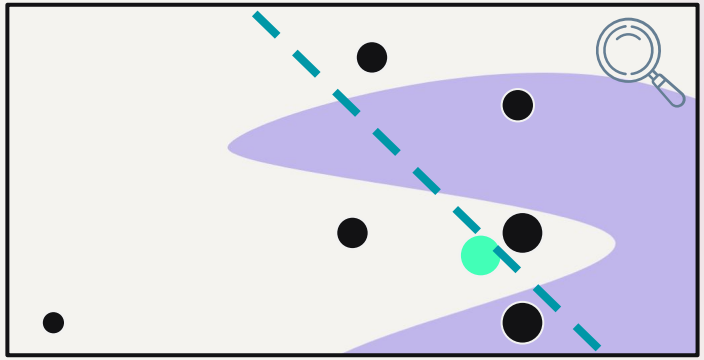
Estimating Required Number of Perturbations (PTG)

Linear Local Decision Surface



You probably don't need to sample too much more!

Non-linear Local Decision Surface



Eh, you need to sample more!

Estimating Required Number of Perturbations (PTG)

How many perturbations you need $=$

$$G(W, \alpha, x) =$$

Estimating Required Number of Perturbations (PTG)

$$\text{How many perturbations you need} = \frac{\text{Local Error}}{} - \text{Number of Sampled Perturbations}$$

$$G(W, \alpha, x) = \frac{4s_S^2}{} - S$$

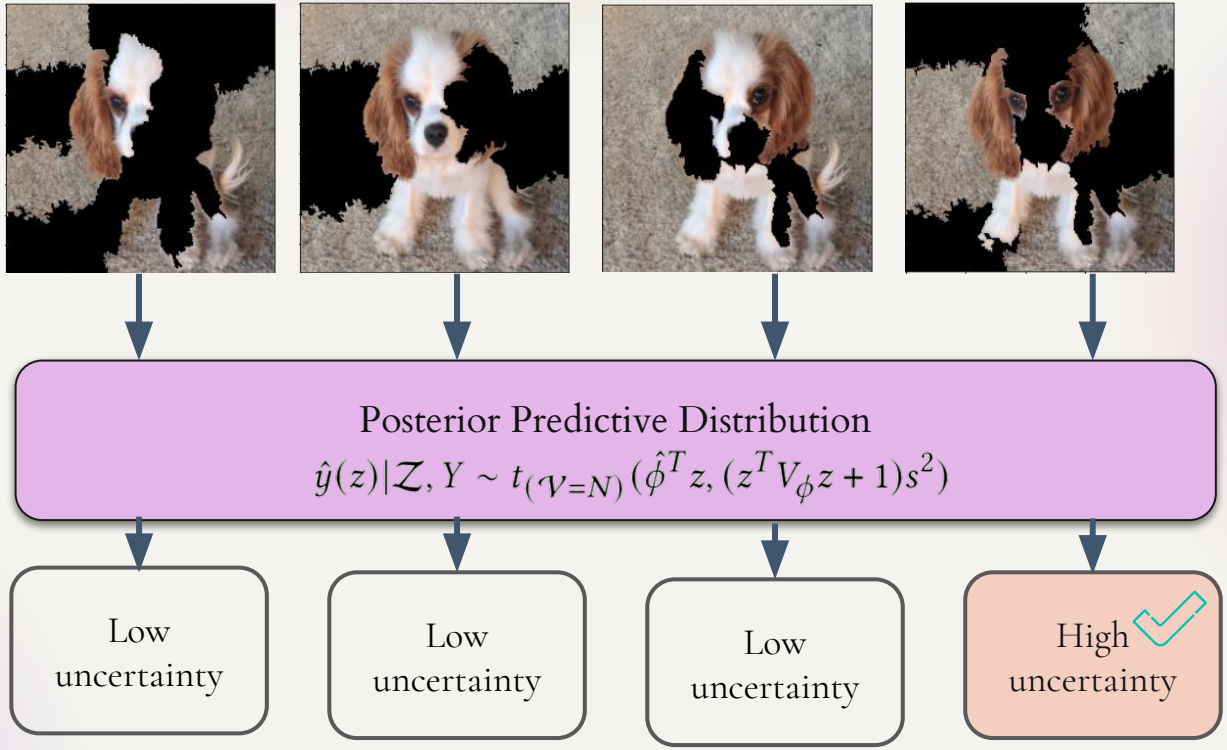
Estimating Required Number of Perturbations (PTG)

$$\text{How many perturbations you need} = \frac{\text{Local Error}}{\text{Perturbation Proximity} \times \text{Desired Uncertainty}} - \text{Number of Sampled Perturbations}$$

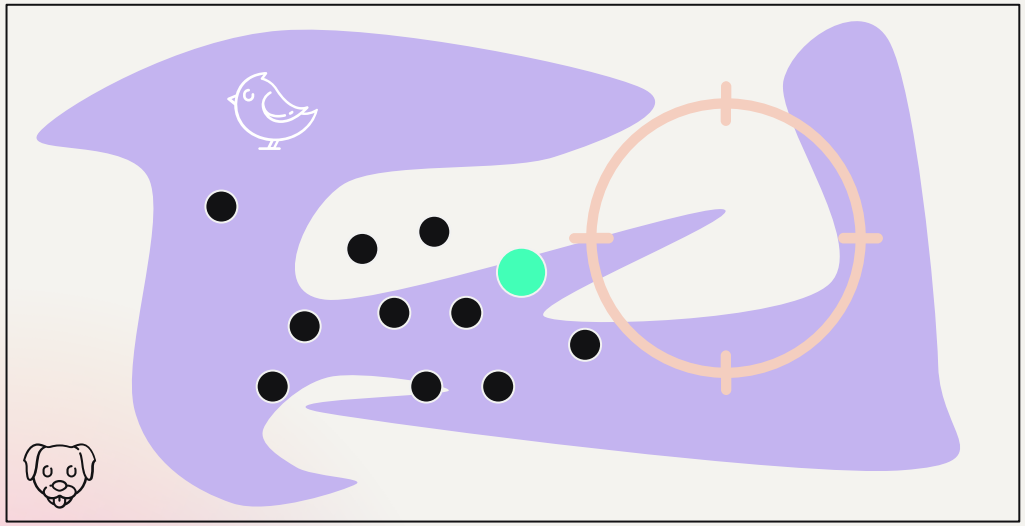
$$G(W, \alpha, x) = \frac{4s_S^2}{\bar{\pi}_S \times \left[\frac{W}{\Phi^{-1}(\alpha)}\right]^2} - S$$

Focused Sampling of Perturbations

Perturbations



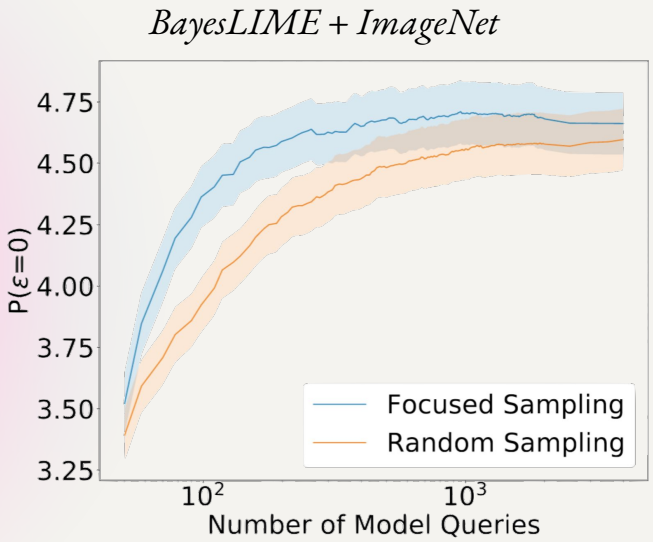
Focused Sampling of Perturbations



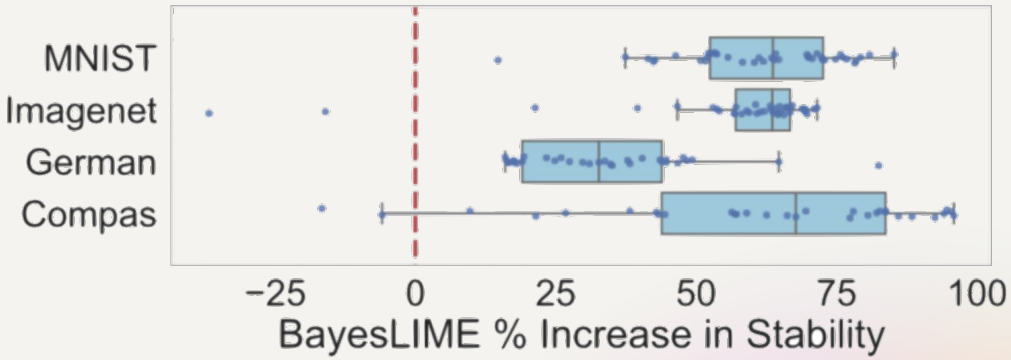
I should probably sample more over here...

Benefits of Focused Sampling

Converges more quickly



Improves Stability



Conclusion

- We need methods to figure out whether to trust ML models
- Post hoc explanations can help us figure this out
  - Revealing “why” models make decisions
  - Very flexible
- However, there are shortcomings to post hoc explanations
- Modeling uncertainty helps us overcome some challenges

Bibliography

Reliable Post hoc Explanations: Modeling Uncertainty in Explainability
Dylan Slack*, Sophie Hilgard, Sameer Singh, and Hima Lakkaraju
NeurIPS 2021

Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods
Dylan Slack*, Sophie Hilgard*, Emily Jia, Sameer Singh, and Hima Lakkaraju
AIES 2020

Counterfactual Explanations Can Be Manipulated
Dylan Slack, Sophie Hilgard, Hima Lakkaraju, and Sameer Singh
NeurIPS 2021

A face-scanning algorithm increasingly decides whether you deserve the job
Drew Harwell
Washington Post 2019
Accessed January 2022

Slides Template Downloaded from Slidesgo