Explaining Machine Learning Predictions: State-of-the-art, Challenges, Opportunities

Hima Lakkaraju   Julius Adebayo   Sameer Singh

Julius Adebayo
MIT

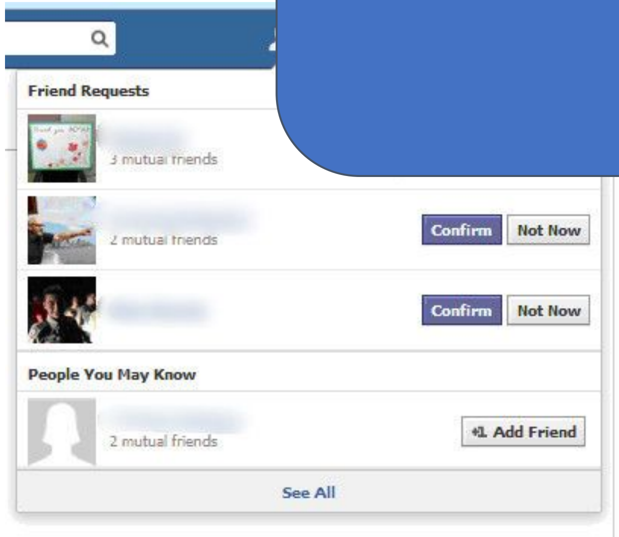
Hima Lakkaraju
Harvard University

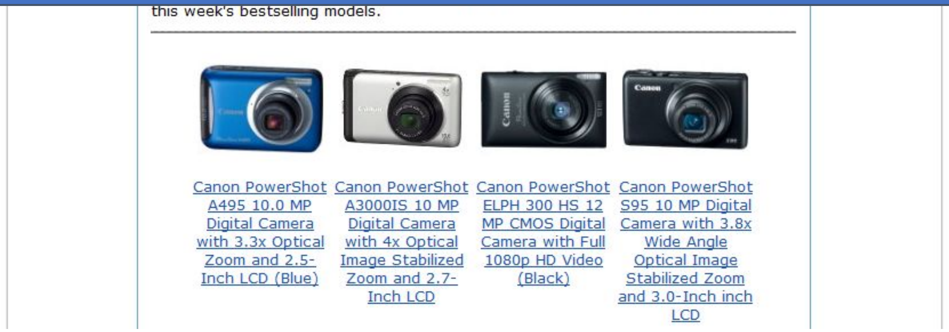
Sameer Singh
UC Irvine

Slides and Video: explainml-tutorial.github.io

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

This model is
incorrect
make
on!! Let
model

Model understanding facilitates debugging.

Predictive Model

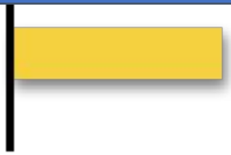
Prediction = Siberian Husky

Motivation: Why Model Understanding?

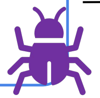
This prediction is

Defendant Detai

Model understanding facilitates bias detection.



Predictive Model



Prediction = Risky to Release

Motivation: Why Model Understanding?

Loan Applicant Details

I have some means

get a loan

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

Predictive Model

Prediction = Denied Loan

Loan Applicant

Motivation: Why Model Understanding?

Patient Data

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

Model Understanding

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.

Predictive Model

Sick
Sick
.
.
Healthy
Healthy
Sick

Motivation: Why Model Understanding?

Patient Data

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

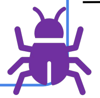
Model Understanding

This model is using
irrelevant features when
n female
n. This
proved!

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

Sick
Sick
.
.
Healthy
Healthy
Sick

Predictive Model



AUTHORITY

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)

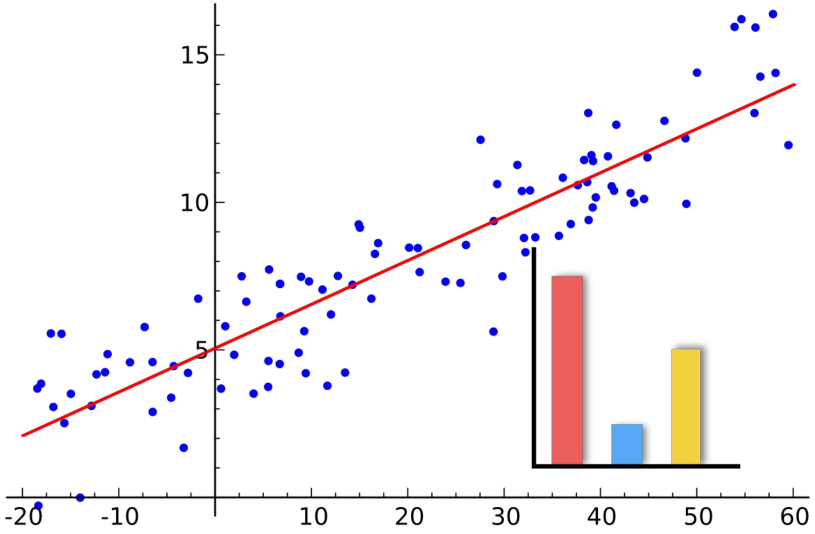
Regulatory agencies (e.g., FDA, European commission)

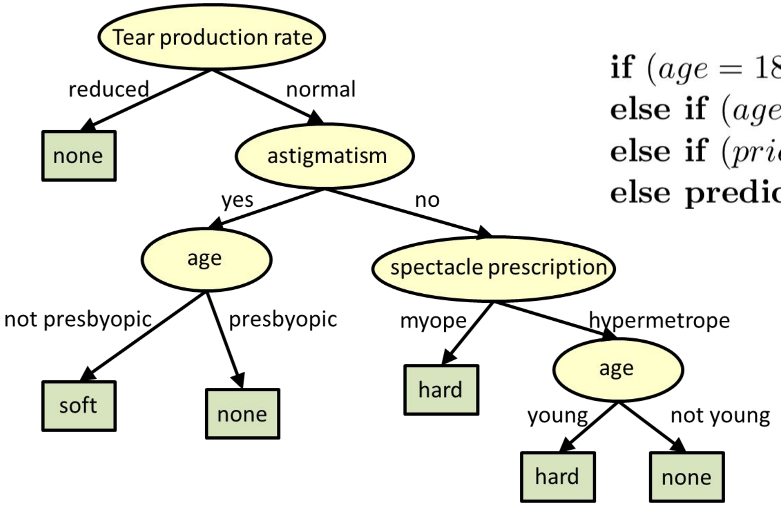
Researchers and engineers

[Letham and Rudin 2015; Lakkaraju et. al. 2016]

Achieving Model Understanding

Take 1: Build *inherently interpretable* predictive models

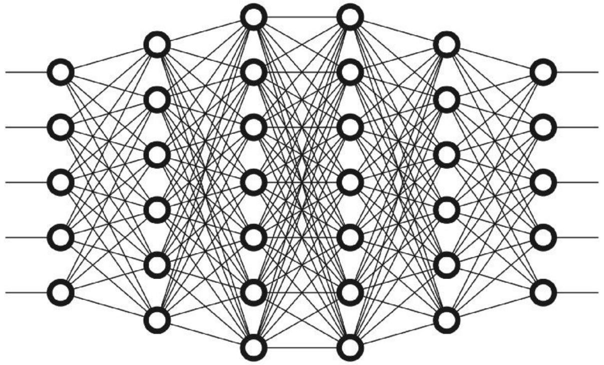




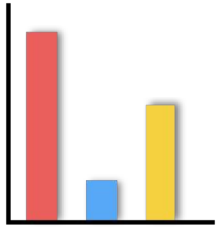
if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

Achieving Model Understanding

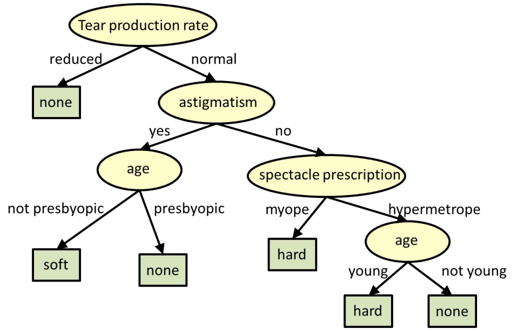
Take 2: *Explain* pre-built models *in a post-hoc manner*



Explainer



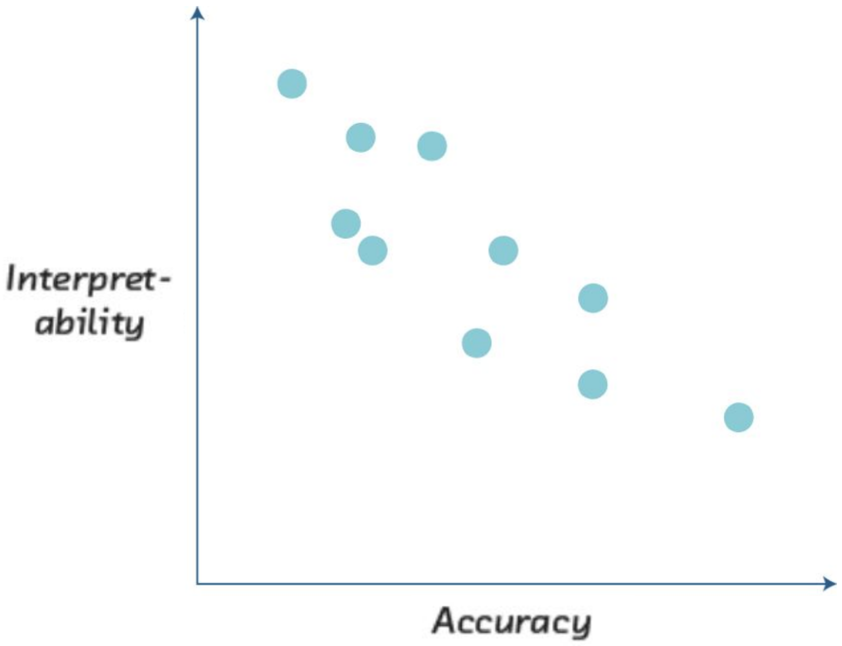
if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

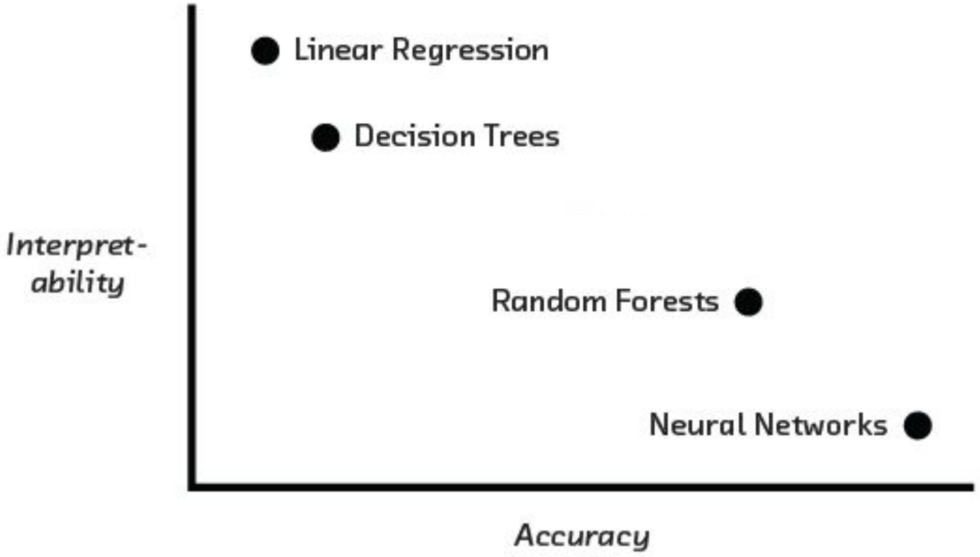


[Cireşan et. al. 2012, Caruana et. al. 2006, Frosst et. al. 2017, Stewart 2020]

Inherently Interpretable Models vs. Post hoc Explanations

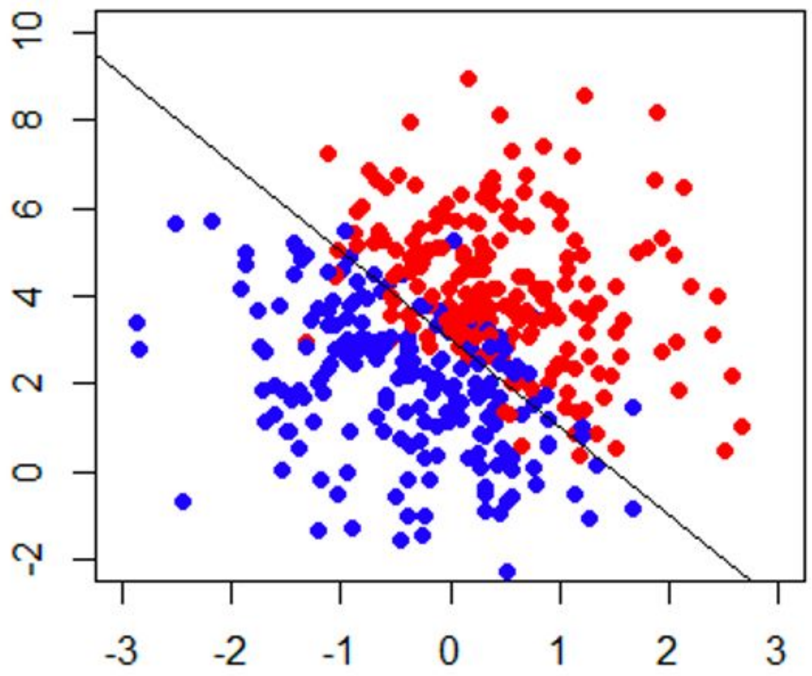
Example



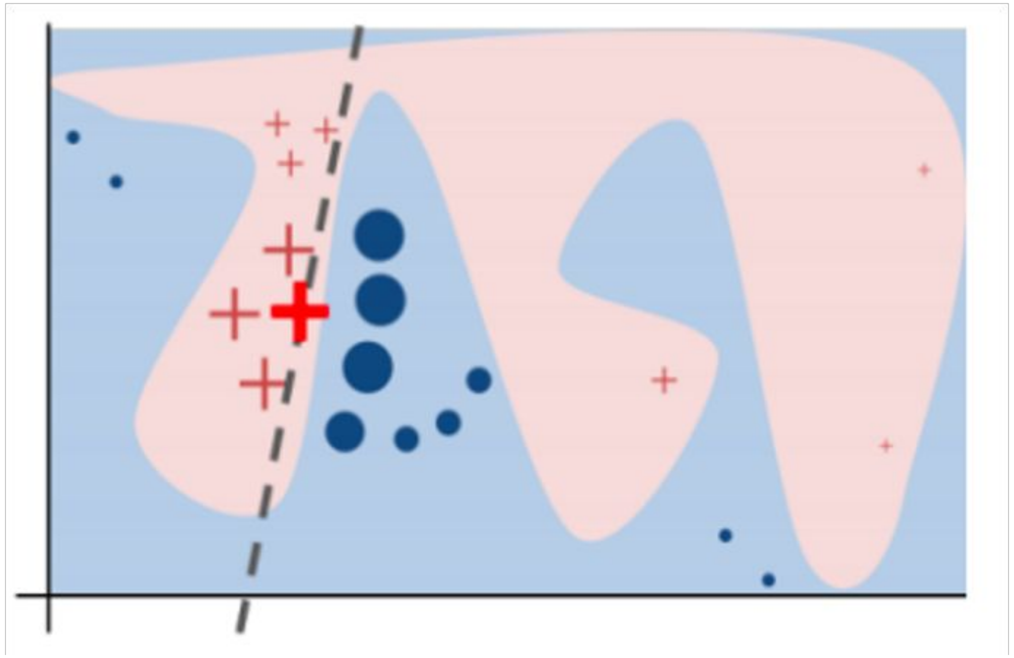


In ***certain*** settings, *accuracy-interpretability trade offs* may exist.

Inherently Interpretable Models vs. Post hoc Explanations



can build interpretable + accurate models



complex models might achieve higher accuracy

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!

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 tutorial will focus on post hoc explanations!

What is an Explanation?

What is an Explanation?

Definition: Interpretable description of the model behavior

Classifier

Faithful

Explanation

Understandable

User

[Lipton 2016]

What is an Explanation?

Definition: Interpretable description of the model behavior

Classifier

Send all the model parameters θ?

Send many example predictions?

Summarize with a program/rule/tree

Select most important features/points

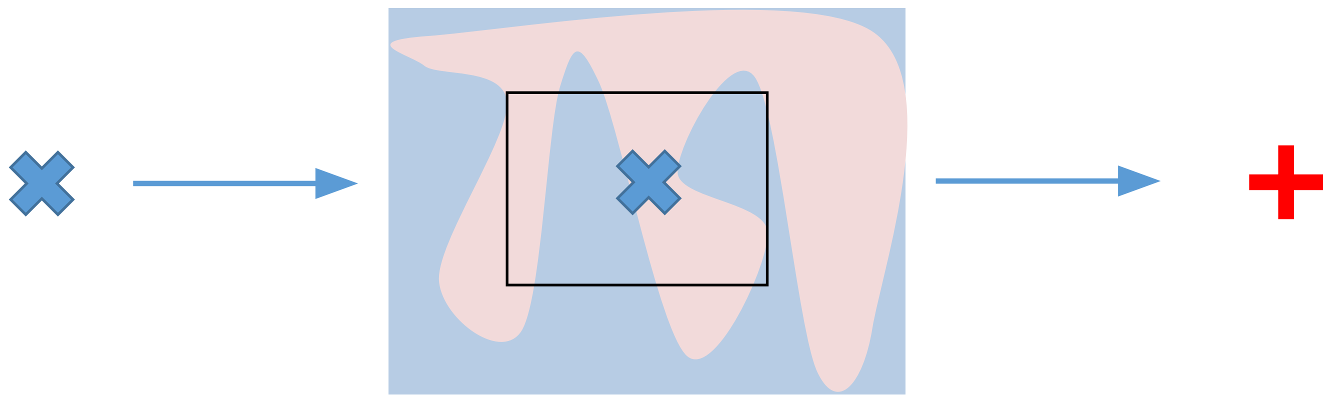
Describe how to *flip* the model prediction

...

User

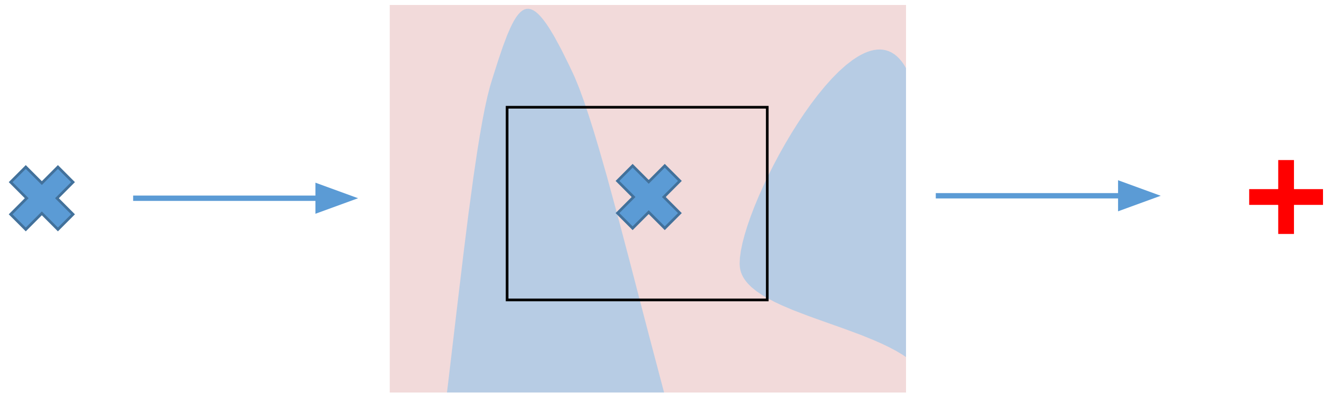
Local versus Global Explanations

Global explanation may be too complicated



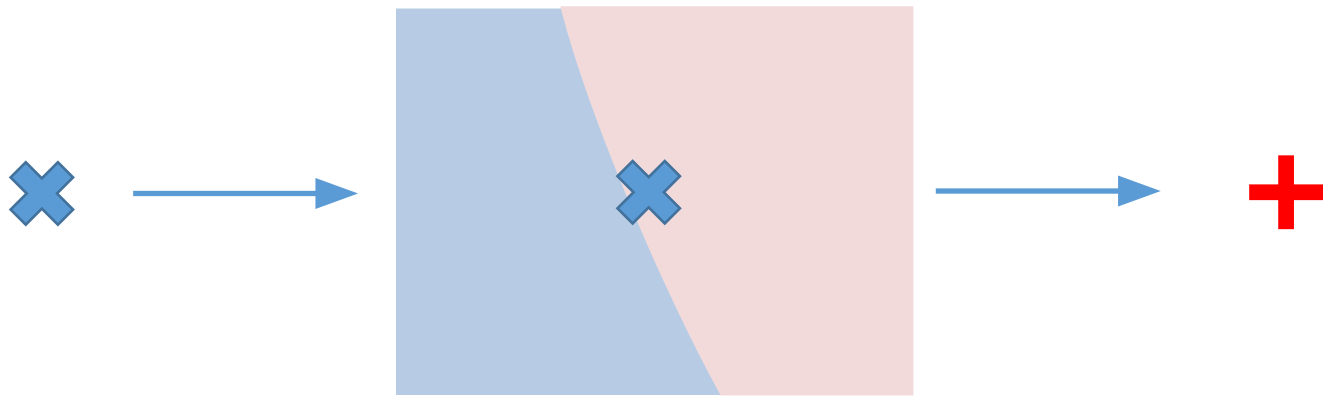
Local versus Global Explanations

Global explanation may be too complicated



Local versus Global Explanations

Global explanation may be too complicated

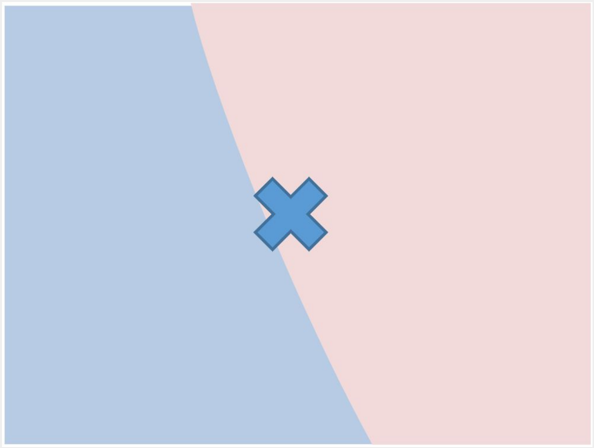


Definition: Interpretable description of the model behavior *in a target neighborhood*.

Local Explanations

Definition: Interpretable description of the model behavior *in a target neighborhood*.

Classifier



Send many example predictions?

Summarize with a program/rule/tree

Select most important features/points

Describe how to *flip* the model prediction

...

User

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	Help shed 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

Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Approaches for Post hoc Explainability

Approaches for Post hoc Explainability

Local Explanations

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

Global Explanations

- Collection of Local Explanations
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

Approaches for Post hoc Explainability

Local Explanations

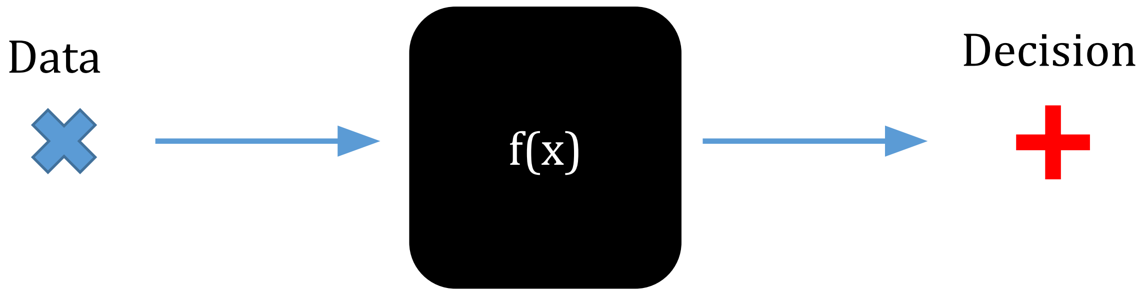
- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

Being Model-Agnostic…

No access to the internal structure…



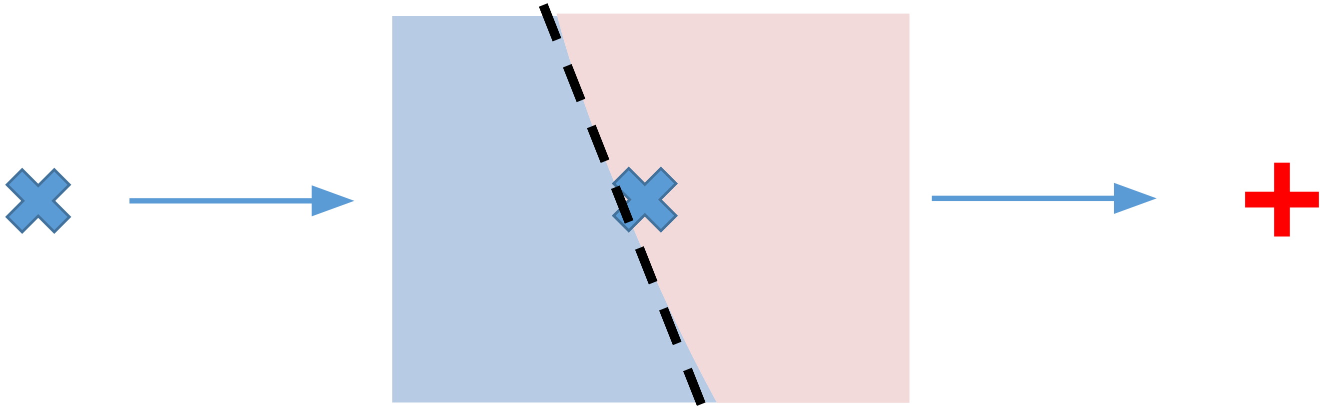
Not restricted to specific models

Practically easy: not tied to PyTorch, Tflow, etc.

Study models that you don't have access to!

LIME: Sparse, Linear Explanations

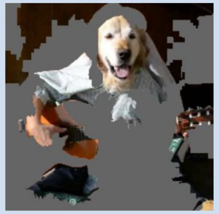
Identify the important dimensions,
and present their relative importance



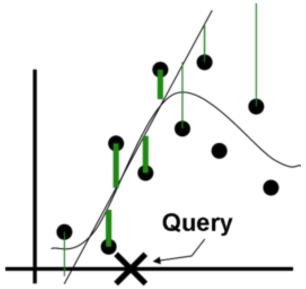
LIME Example - Images

Original Image

P(labrador) = 0.21

Perturbed Instances	P(Labrador)
	0.92
	0.001
	0.34

Maybe to a fault?



Locally weighted regression

Explanation

LIME is quite customizable:

- How to perturb?
- Distance/similarity?
- How *local* you want it to be?
- How to express explanation

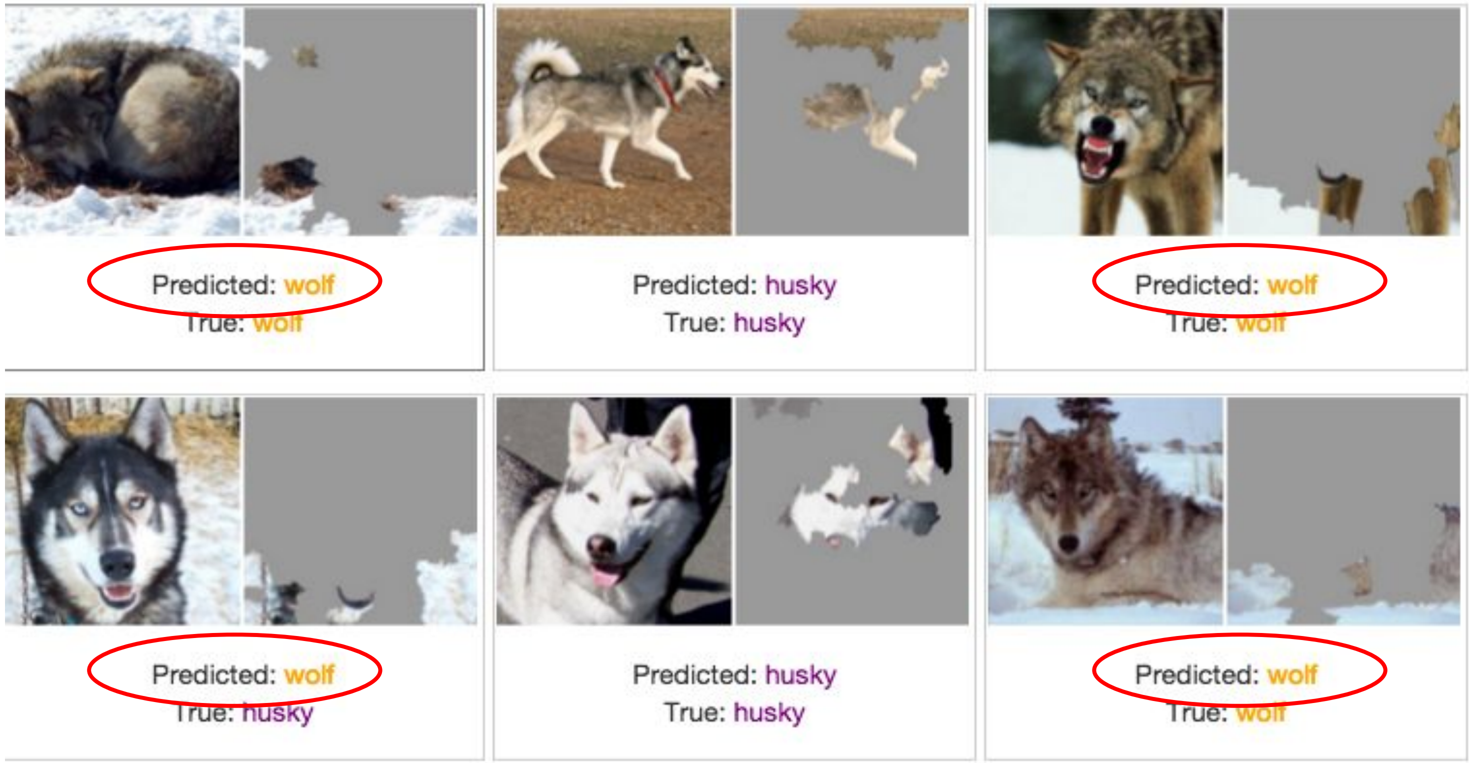
[Ribeiro et al. 2016]

Predict Wolf vs Husky

Only 1 mistake!

[Ribeiro et al. 2016]

Predict Wolf vs Husky

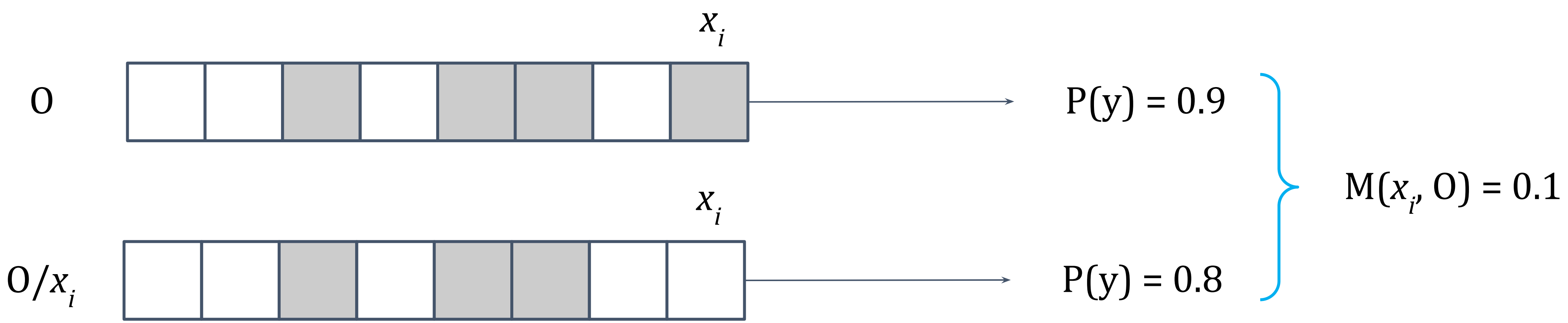


We've built a great snow detector…

[Lundberg & Lee 2017]

SHAP: Shapley Values as Importance

Marginal contribution of each feature towards the prediction, averaged over all possible permutations.



Fairly attributes the prediction to all the features.

Approaches for Post hoc Explainability

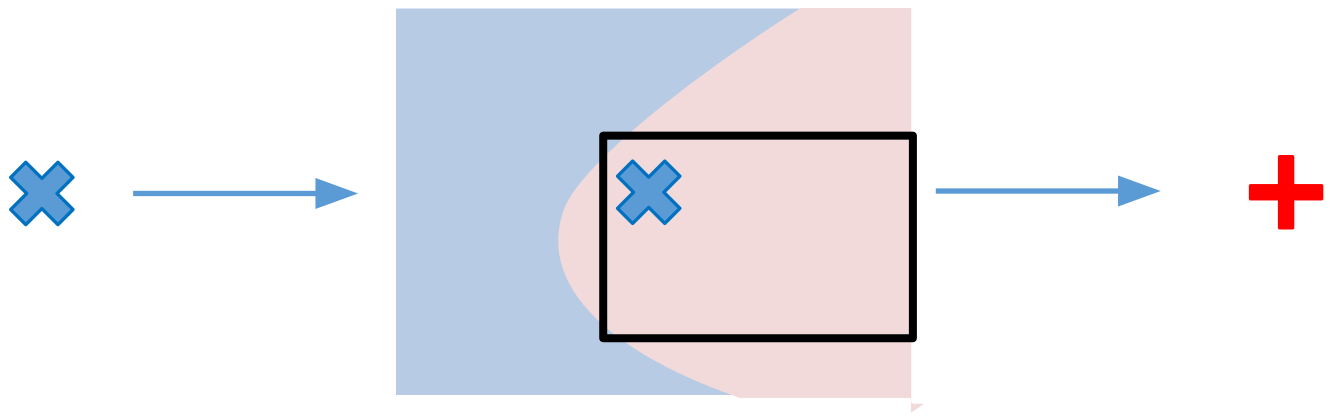
Local Explanations

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

Global Explanations

- Collection of Local Explanations
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

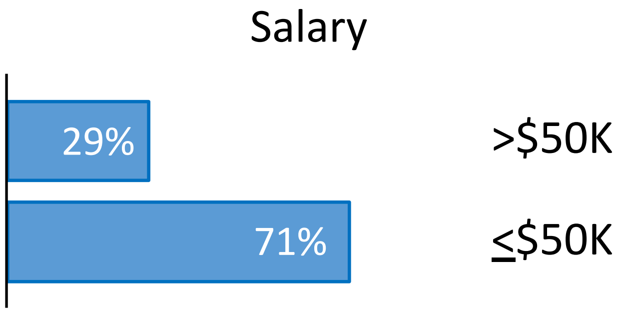
Anchors: Sufficient Conditions



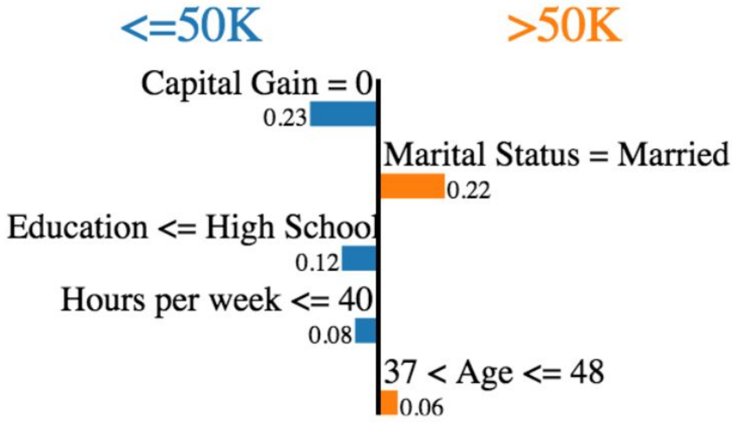
Identify the conditions under which the classifier has the same prediction

Salary Prediction

Feature	Value
Age	$37 < \text{Age} \leq 48$
Workclass	Private
Education	$\leq$ High School
Marital Status	Married
Occupation	Craft-repair
Relationship	Husband
Race	Black
Sex	Male
Capital Gain	0
Capital Loss	0
Hours per week	≤ 40
Country	United States



LIME



Anchors

IF Education $\leq$ High School
Then Predict Salary $\leq$ 50K

Approaches for Post hoc Explainability

Local Explanations

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

Global Explanations

- Collection of Local Explanations
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

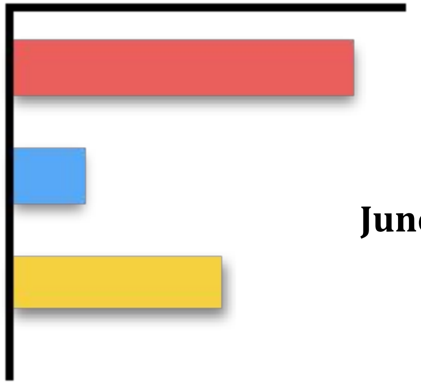
Saliency Map Overview

Input

Model

$$F_\theta$$

Predictions



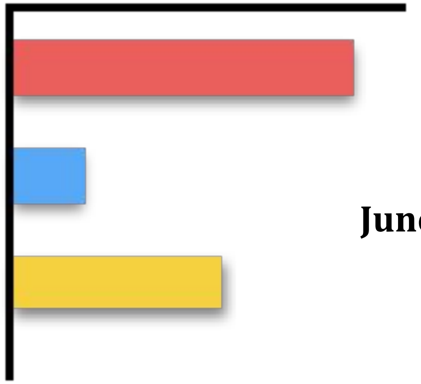
Junco Bird

Saliency Map Overview

Input

Model

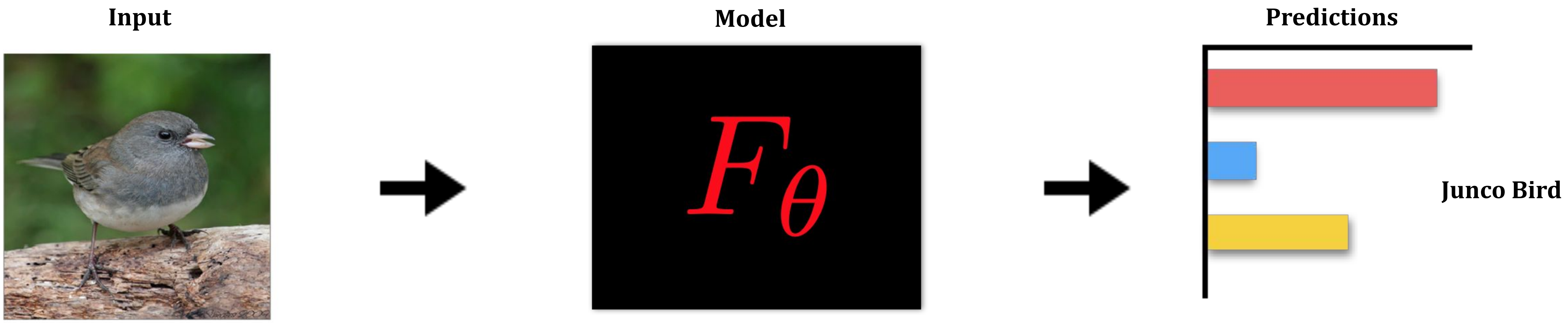
Predictions



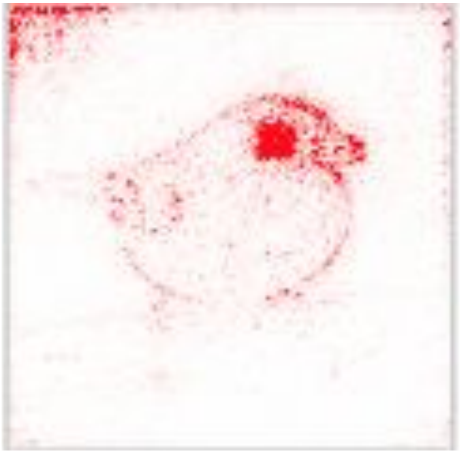
Junco Bird

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

Saliency Map Overview



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



- **Feature Attribution**
- **‘Saliency Map’**
- **Heatmap**

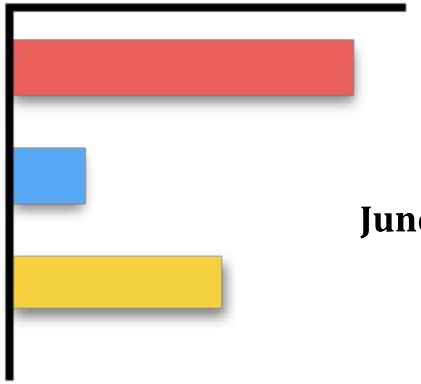
Input-Gradient

Input

Model

$$F_\theta$$

Predictions



Junco Bird

Input-Gradient

$$\nabla_x F_i(x)$$

Logit

Input

Visualize as a heatmap

[Baehrens et. al. 2010](); [Simonyan et. al. 2014]() .

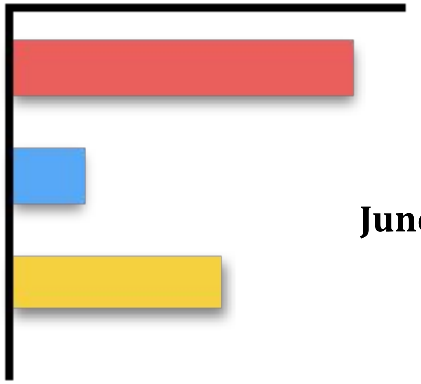
Input-Gradient

Input

Model

$$F_\theta$$

Predictions



Junco Bird

Input-Gradient

$$\nabla_x F_i(x)$$

Logit

Input

Challenges

- **Visually noisy & difficult to interpret.**
- **‘Gradient saturation.’**

[Shrikumar et. al. 2017](#).

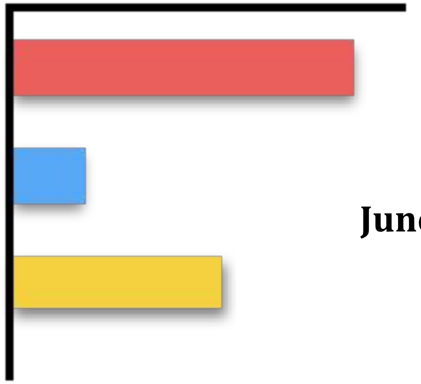
[Baehrens et. al. 2010](#); [Simonyan et. al. 2014](#) .

SmoothGrad

Input

Model

Predictions



SmoothGrad

$$\frac{1}{N}\sum_{i}^{N}\nabla_{(x+\epsilon)}F_i(x+\epsilon)$$

Gaussian noise

Average Input-gradient of 'noisy' inputs.

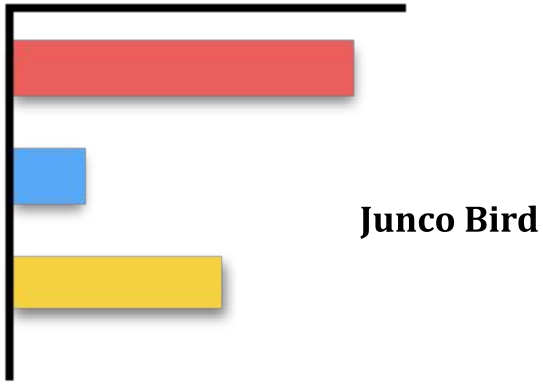
[Smilkov et. al. 2017](#)

Integrated Gradients

Input

Model

Predictions



$$(x - \tilde{x}) \times \int_{\alpha=0}^{1} \frac{\partial F(\tilde{x} + \alpha \times (x - \tilde{x}))}{\partial x}$$

Baseline input

Path integral: 'sum' of interpolated gradients

[Sundararajan et. al. 2017]

‘Modified Backprop’ Approaches

Compute feature relevance by modifying the backpropagation via **positive aggregation**.

‘Modified Backprop’ Approaches: Guided BackProp

Compute feature relevance by modifying the backpropagation via **positive aggregation**.

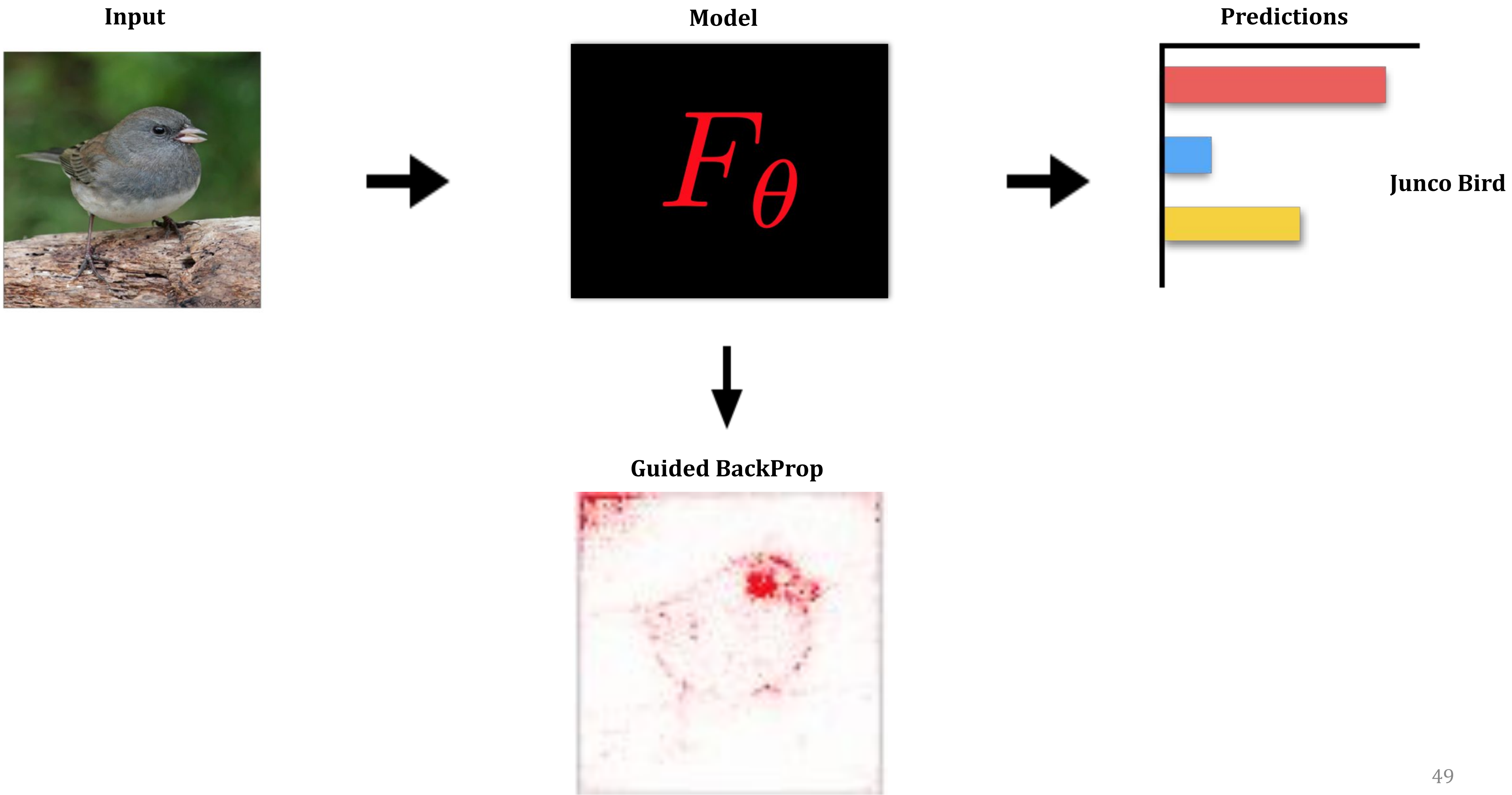
activation: $$f_i^{l+1} = relu(f_i^l) = \max(f_i^l, 0)$$

backpropagation: $$R_i^l = (f_i^l > 0) \cdot R_i^{l+1}, \text{ where } R_i^{l+1} = \frac{\partial f^{out}}{\partial f_i^{l+1}}$$

guided backpropagation: $$R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1}$$

Source: [Springenberg & Dosovitskiy et. al. 2015]

Attribution: Guided BackProp

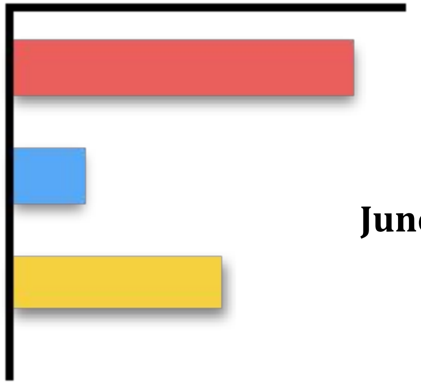


Attribution: Guided BackProp

Input

Model

Predictions



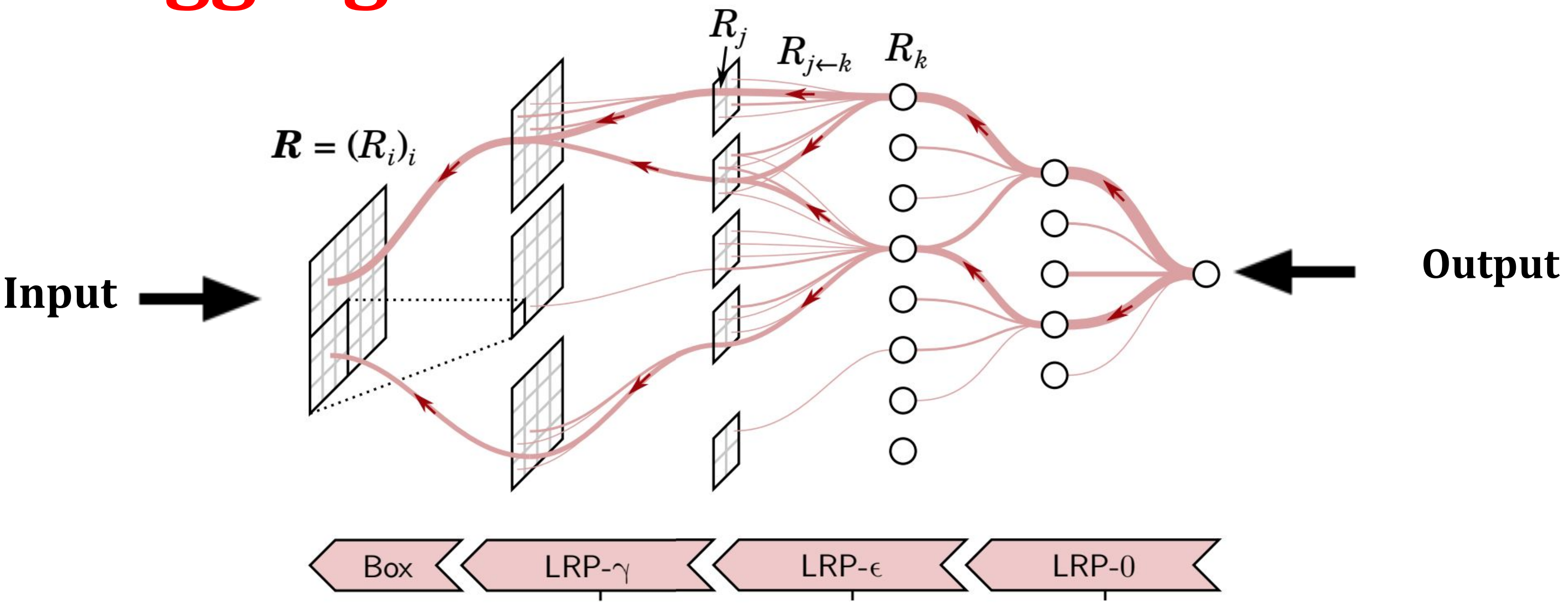
Junco Bird

Guided BackProp

Input Gradient

‘Modified Backprop’ Approaches: LRP

Compute feature relevance by modifying the backpropagation via **positive aggregation**.



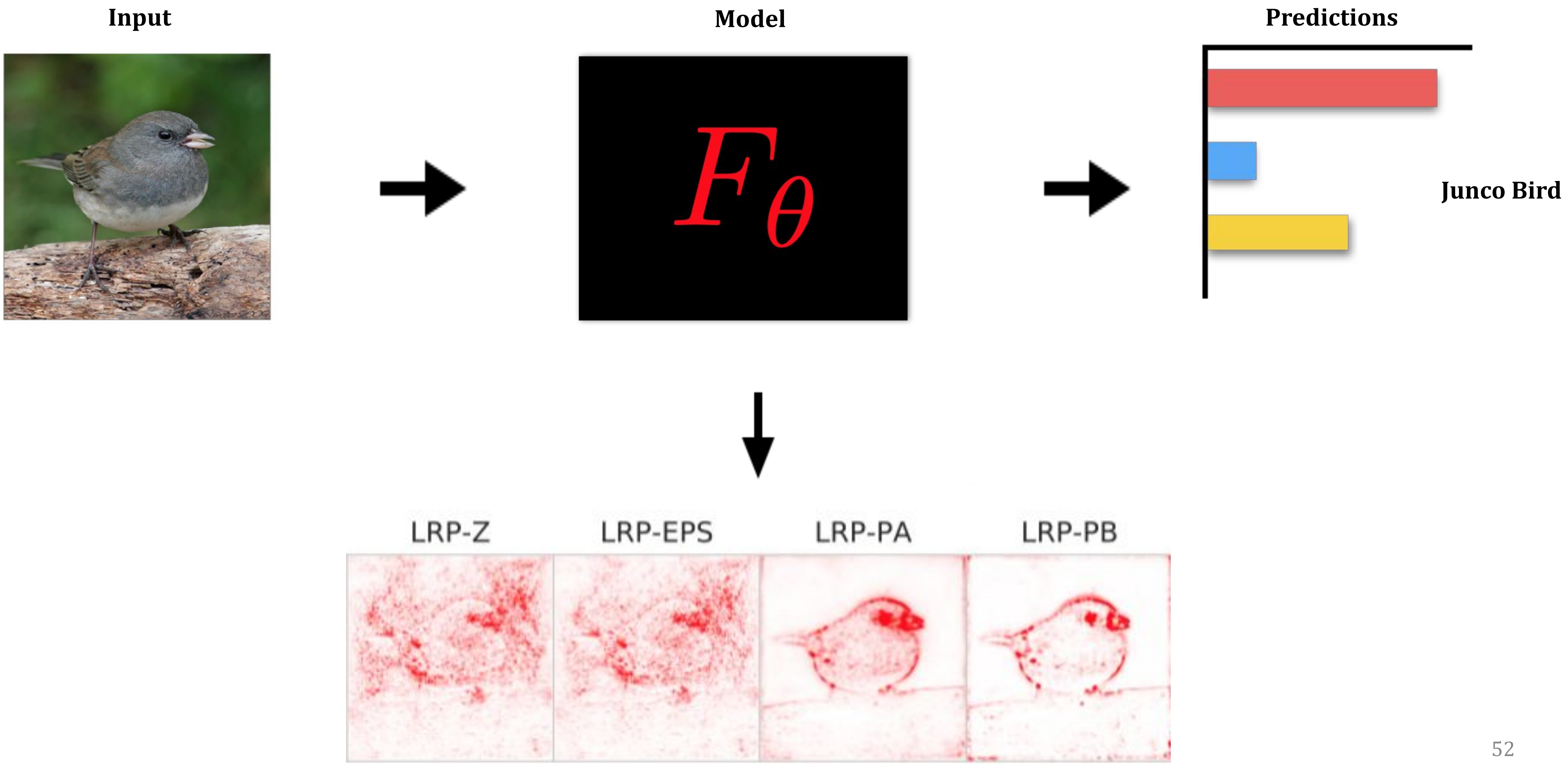
$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

$$R_j = \sum_k \frac{a_j w_{jk}}{\varepsilon_k + \sum_{0,j} a_j w_{jk}} R_k$$

$$R_j = \sum_k \frac{a_j (w_{jk} + \gamma w_{jk}^{+})}{\sum_{0,j} a_j (w_{jk} + \gamma w_{jk}^{+})} R_k$$

Source: heatmapping.org; ECML-PKDD 2020 Tutorial.

Layer Relevance Propagation (LRP)



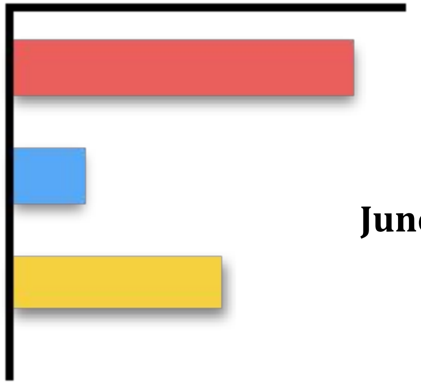
Recap

Input

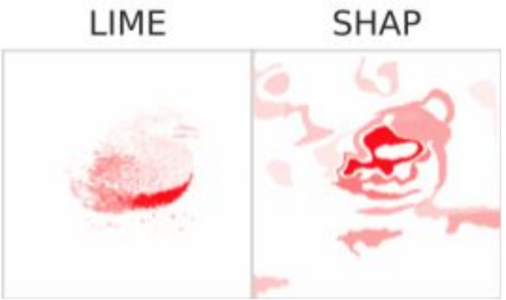
Model

$$F_\theta$$

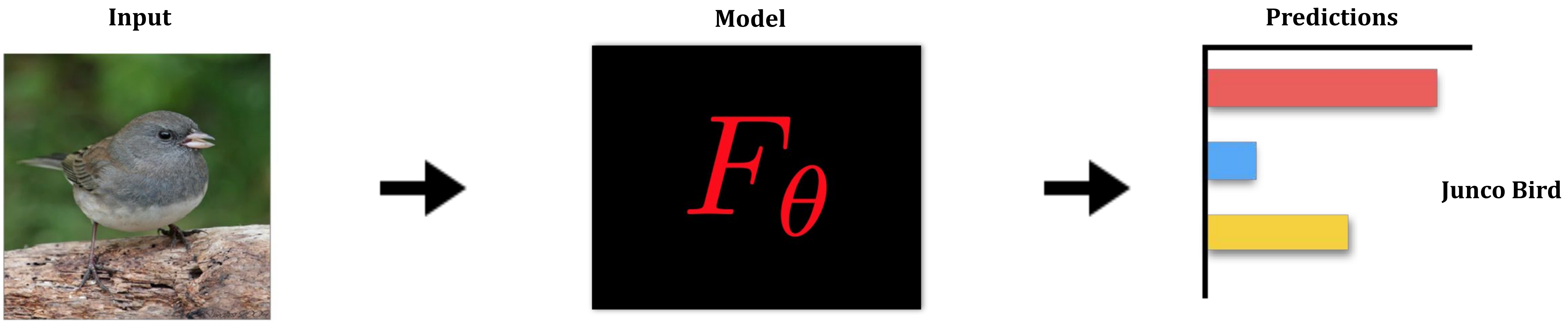
Predictions

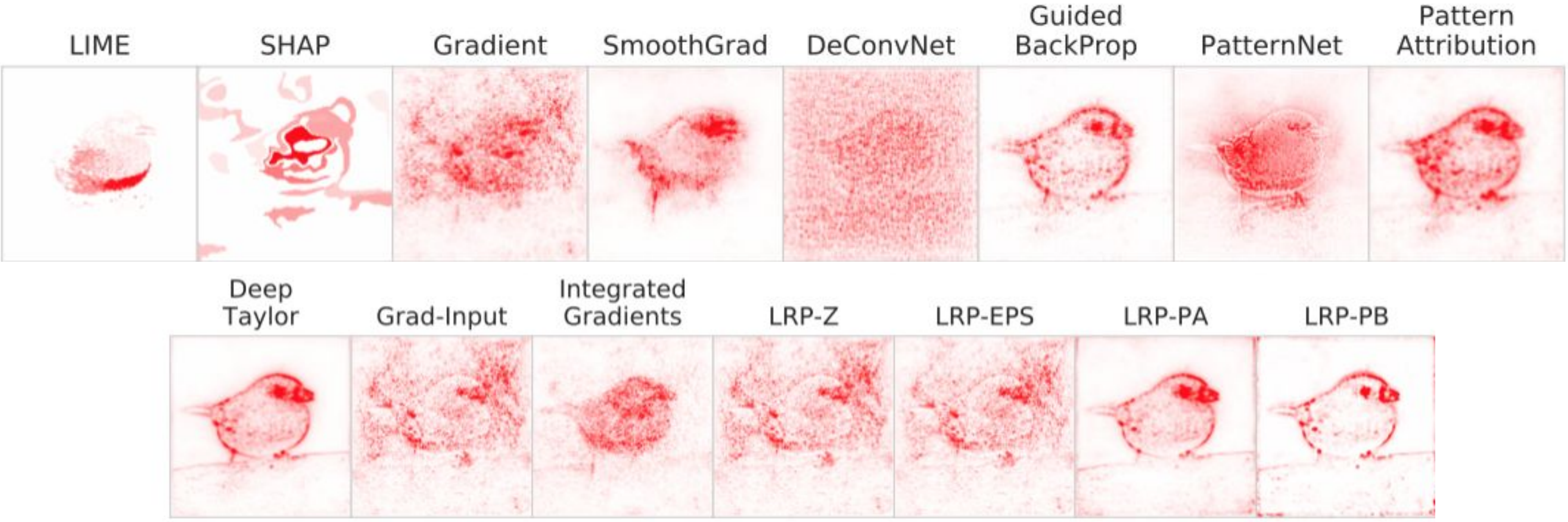


Junco Bird



Recap


Input
Model
F_θ
Predictions
Junco Bird


LIME
SHAP
Gradient
SmoothGrad
DeConvNet
Guided BackProp
PatternNet
Pattern Attribution
Deep Taylor
Grad-Input
Integrated Gradients
LRP-Z
LRP-EPS
LRP-PA
LRP-PB

Additional Methods

- **Class Activation Mapping** (Zhou et. al. 2016).
- **Meaningful Perturbation** (Fong et. al. 2017).
- **RISE** (Petsuik et. al. 2018).
- **Extremal Perturbations** (Fong & Patrick 2019).
- **DeepLift** (Shrikumar et. al. 2018).
- **Expected Gradients** (Erion et. al. 2019)
- **Excitation Backprop** (Zhang et. al. 2016)
- **GradCAM** (Selvaraju et. al. 2016)
- **Guided GradCAM** (Selvaraju et. al. 2016)
- **Occlusion** (Zeiler et. al. 2014).
- **Prediction Difference Analysis** (Gu. et. al. 2019).
- **Internal Influence** (Leino et. al. 2018).

See for additional methods: Samek & Montavon et. al. 2020

Approaches for Post hoc Explainability

Local Explanations

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

Global Explanations

- Collection of Local Explanations
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

Prototype Approaches

Explain a model with synthetic or natural input ‘**examples**’.

Prototype Approaches

Explain a model with synthetic or natural input ‘**examples**’.

Insights

- What kind of input is the model **most likely to misclassify**?
- Which training samples are **mislabelled**?
- Which input **maximally activates** an intermediate neuron?

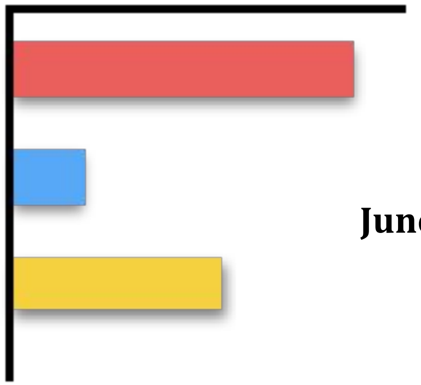
Training Point Ranking via Influence Functions

Input

Model

F_θ

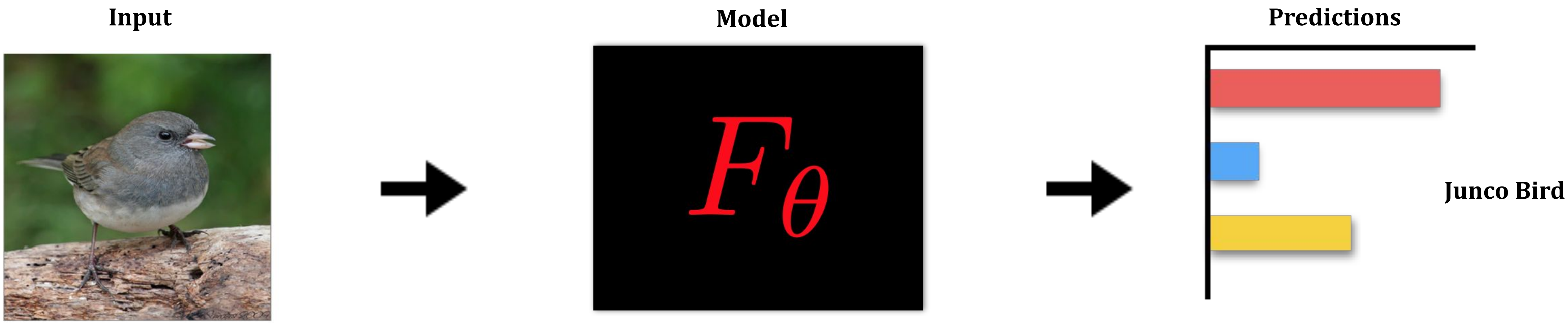
Predictions



Junco Bird

Which training points have the most ‘influence’ on test input’s loss?

Training Point Ranking via Influence Functions



Which training points have the most 'influence' on test input's loss?

Training Point Ranking via Influence Functions

Influence Function: classic tool used in robust statistics for assessing the effect of a sample on regression parameters (Cook & Weisberg, 1980).

Influence of Training Point on Parameters

$$\mathcal{I}_{z_j} = \left.\frac{d\hat{\theta}_{\epsilon,z_j}}{d\epsilon}\right|_{\epsilon=0} = -H_{\hat{\theta}}^{-1}\nabla_\theta \ell(z_j,\hat{\theta})$$

Influence of Training Point on Test-Input's loss

$$\mathcal{I}_{z_j,z_{\text{test}},\text{loss}} = -\nabla_\theta \ell(z_{\text{test}},\hat{\theta})^\top H_\theta^{-1}\nabla_\theta \ell(z_j,\hat{\theta})$$

Koh & Liang 2017

Challenges and Other Approaches

Influence function Challenges:

1. **scalability**: computing hessian-vector products can be tedious in practice.

2. **non-convexity:** possibly loose approximation for deeper networks (Basu et. al. 2020).

Challenges and Other Approaches

Influence function Challenges:

1. **scalability**: computing hessian-vector products can be tedious in practice.

2. **non-convexity:** possibly loose approximation for deeper networks (Basu et. al. 2020).

Alternatives:

- **Representer Points** (Yeh et. al. 2018).

- **TracIn** (Pruthi et. al. appearing at NeuRIPs 2020).

‘Activation Maximization’

These approaches identify examples, synthetic or natural, that **strongly activate a function (neuron) of interest**.

‘Activation Maximization’

These approaches identify examples, synthetic or natural, that **strongly activate a function (neuron) of interest**.

Implementation Flavors:

- Search for **natural examples within a specified set** (training or validation corpus) that strongly activate a neuron of interest;

- **Synthesize examples**, typically via gradient descent, that strongly activate a neuron of interest.

Feature Visualization

Dataset Examples show us what neurons respond to in practice

Optimization isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.

Baseball—or stripes? *mixed4a, Unit 6*

Animal faces—or snouts? *mixed4a, Unit 240*

Clouds—or fluffiness? *mixed4a, Unit 453*

Buildings—or sky? *mixed4a, Unit 492*

[Olah et. al. 2017]()

Approaches for Post hoc Explainability

Local Explanations

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

Global Explanations

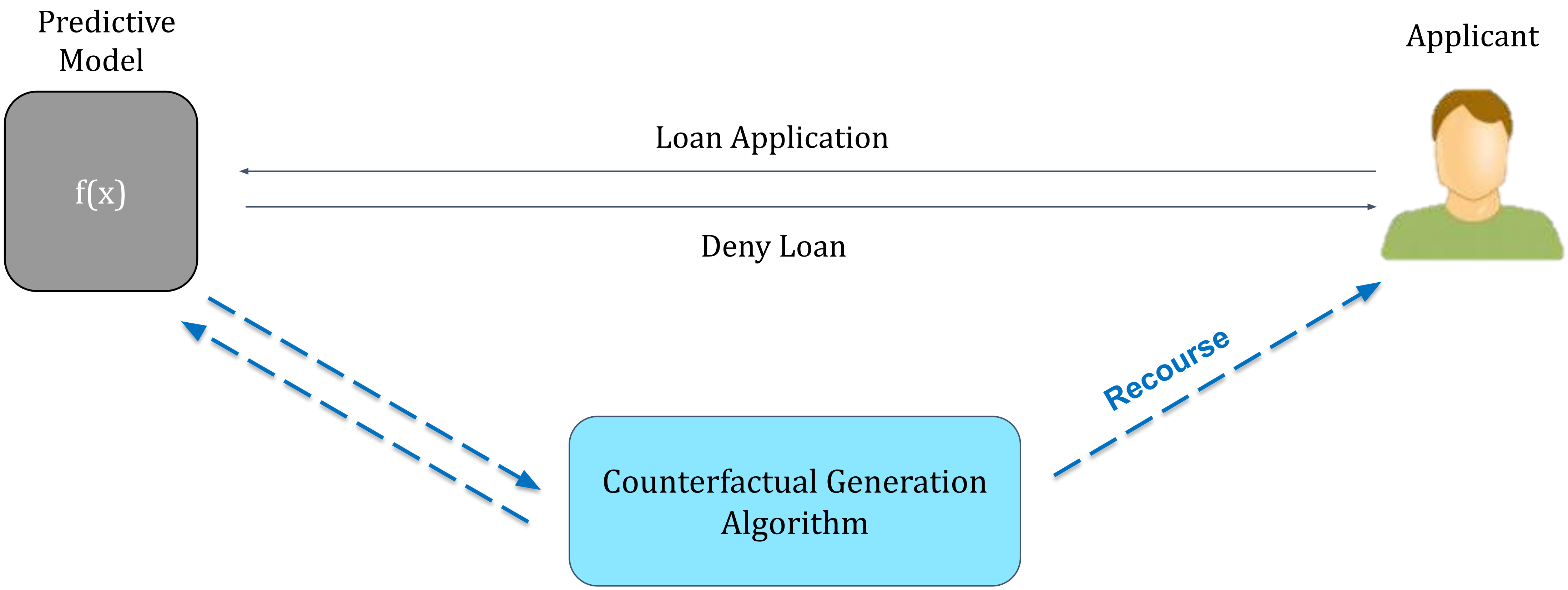
- Collection of Local Explanations
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

Counterfactual Explanations

As ML models increasingly deployed to make high-stakes decisions (e.g., loan applications), it becomes important to provide recourse to affected individuals.

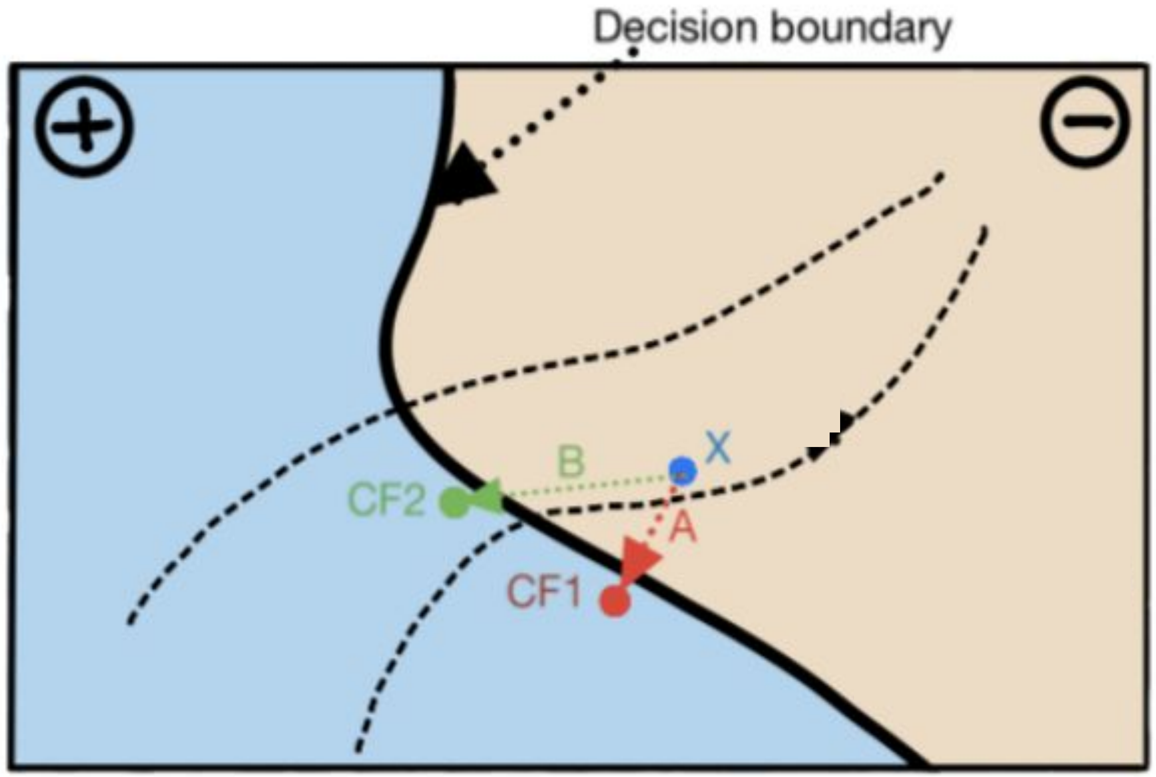
Counterfactual Explanations
What features need to be changed and by how much to flip a model's prediction ?
(i.e., to reverse an unfavorable outcome).

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:

How to choose among candidate counterfactuals?

[Wachter et. al., 2018]

Take 1: Minimum Distance Counterfactuals

Distance Metric

Counterfactual

$$\arg\min_{x'} d(x, x')$$

Original Instance

$$s.t. \; f(x') = y'$$

Predictive Model

Desired Outcome

Choice of distance metric dictates what kinds of counterfactuals are chosen.

Wachter et. al. use normalized Manhattan distance.

Take 1: Minimum Distance Counterfactuals

Person 1: If your LSAT was 34.0, you would have an average predicted score (0).

Person 2: If your LSAT was 32.4, you would have an average predicted score (0).

Person 3: If your LSAT was 33.5, and you were ‘white’, you would have an average predicted score (0).

Person 4: If your LSAT was 35.8, and you were ‘white’, you would have an average predicted score (0).

Person 5: If your LSAT was 34.9, you would have an average predicted score (0).

Not feasible to act upon these features!

[Ustun et. al., 2019]

Take 2: Feasible and Least Cost Counterfactuals

$$\arg\min_{x'} d(x, x') \quad s.t.\ f(x') = y'$$

$$\arg\min_{x' \in \boxed{\mathcal{A}}} \boxed{cost}(x, x') \quad s.t.\ f(x') = y'$$

- $\mathcal{A}$ is the set of feasible counterfactuals (input by end user)
  - E.g., changes to race, gender are not feasible
- *Cost* to capture how hard it is to go from *x* to *x*'

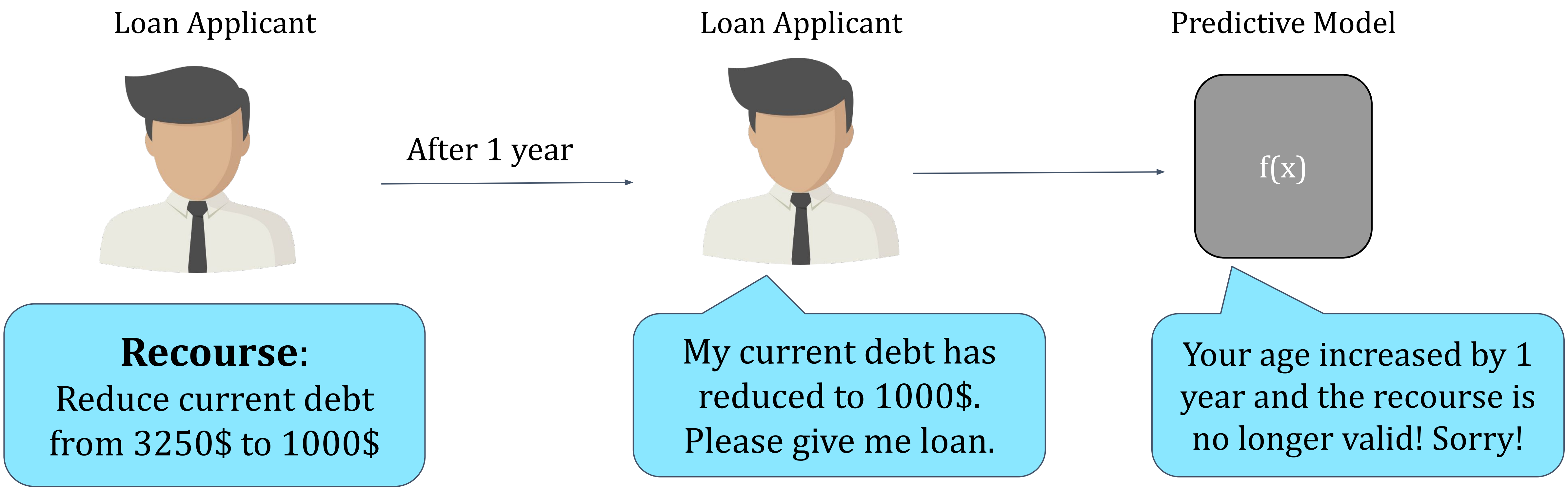
[Ustun et. al., 2019]

Take 2: Feasible and Least Cost Counterfactuals

Features to Change	Current Values		Required Values
n_credit_cards	5	⟶	3
current_debt	$3,250	⟶	$1,000
has_savings_account	FALSE	⟶	TRUE
has_retirement_account	FALSE	⟶	TRUE

[Mahajan et. al., 2019, Karimi et. al. 2020]

Take 3: Causally Feasible Counterfactuals



Important to account for *feature interactions* when generating counterfactuals!

But how?!

Take 3: Causally Feasible Counterfactuals

$$\arg\min_{x'} d(x, x') \quad s.t.\ f(x') = y'$$

$$\arg\min_{x'} d\_causal(x, x') \quad s.t.\ f(x') = y'$$

Leverage Structural Causal Model (SCM) to define this new distance metric

Underlying causal models capture the feature interactions

Approaches for Post hoc Explainability

Local Explanations

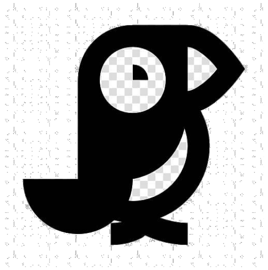
- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

Global Explanations

- Explain the complete behavior of a given (black box) model
  - Provide a *bird's eye view* of model behavior



- Help detect *big picture* model biases persistent across larger subgroups of the population
  - Impractical to manually inspect local explanations of several instances to ascertain big picture biases!

- Global explanations are complementary to local explanations

Approaches for Post hoc Explainability

Local Explanations

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

Global Explanations

- Collection of Local Explanations
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

Global Explanation as a Collection of Local Explanations

How to generate a global explanation of a (black box) model?

- Generate a local explanation for every instance in the data using one of the approaches discussed earlier

- Pick a subset of k local explanations to constitute the global explanation

[Ribeiro et. al., 2016]

Global Explanations from Local Feature Importances: SP-LIME

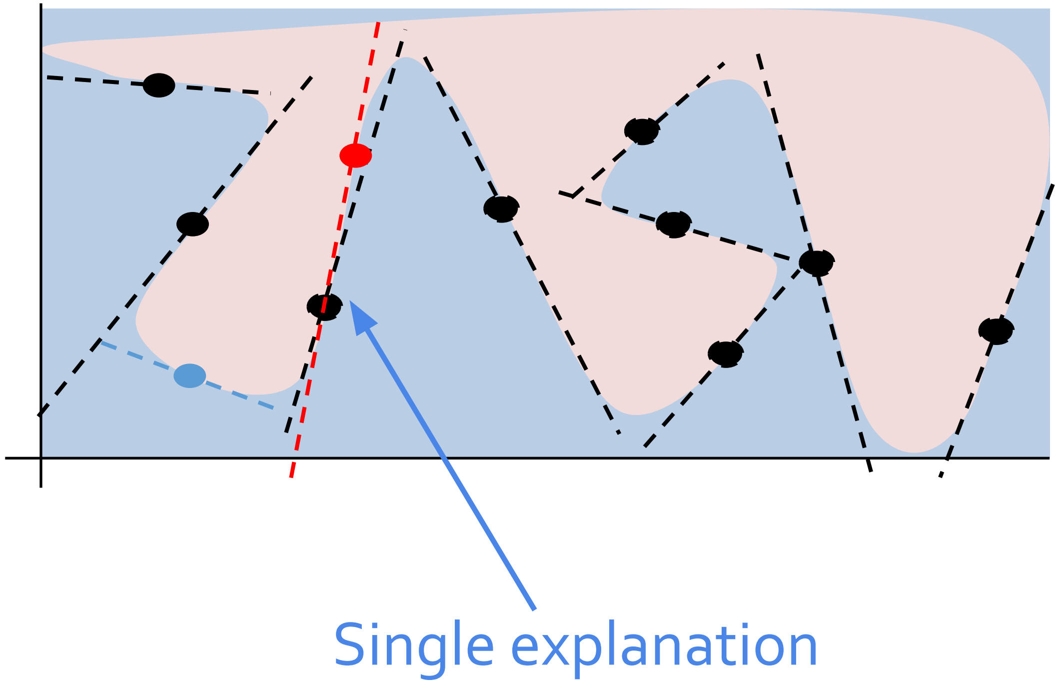
LIME explains a single prediction
local behavior for a single instance

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



Approaches for Post hoc Explainability

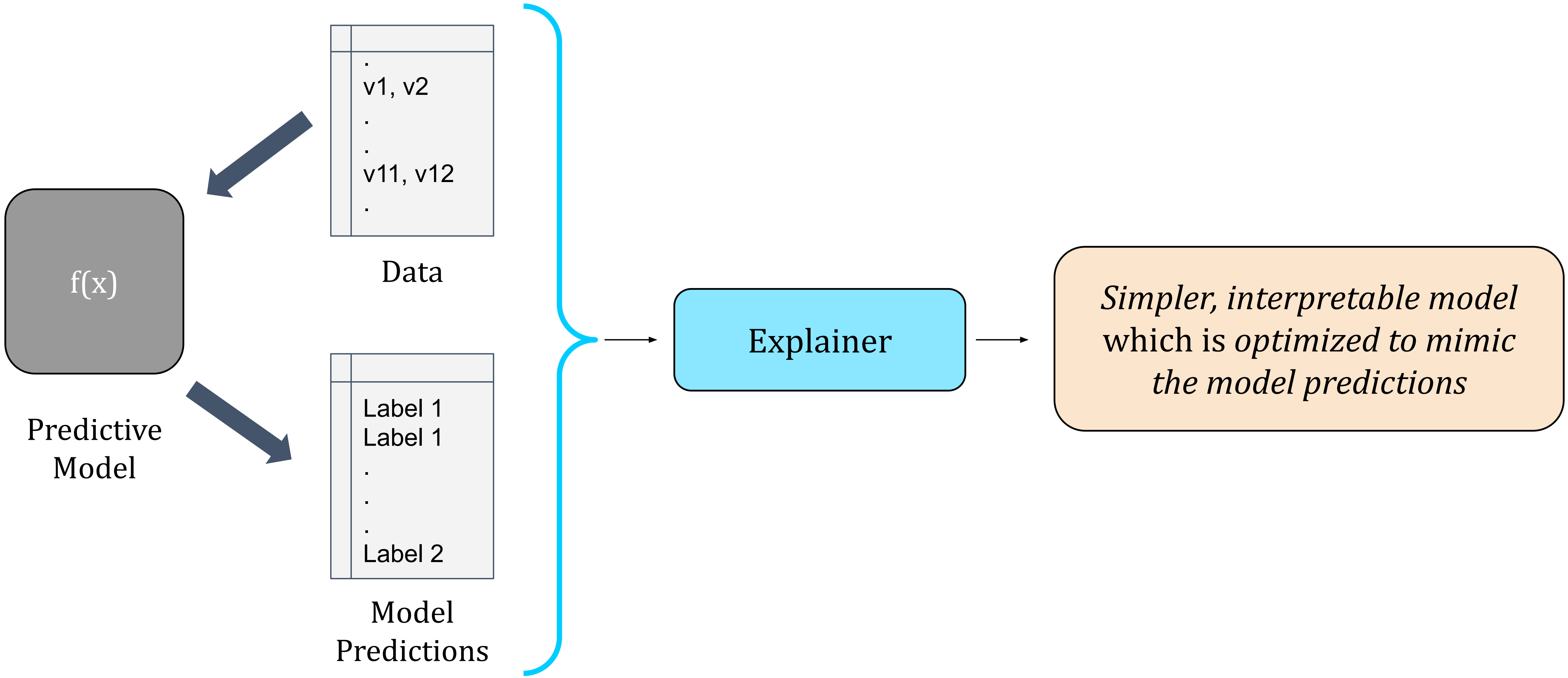
Local Explanations

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

Global Explanations

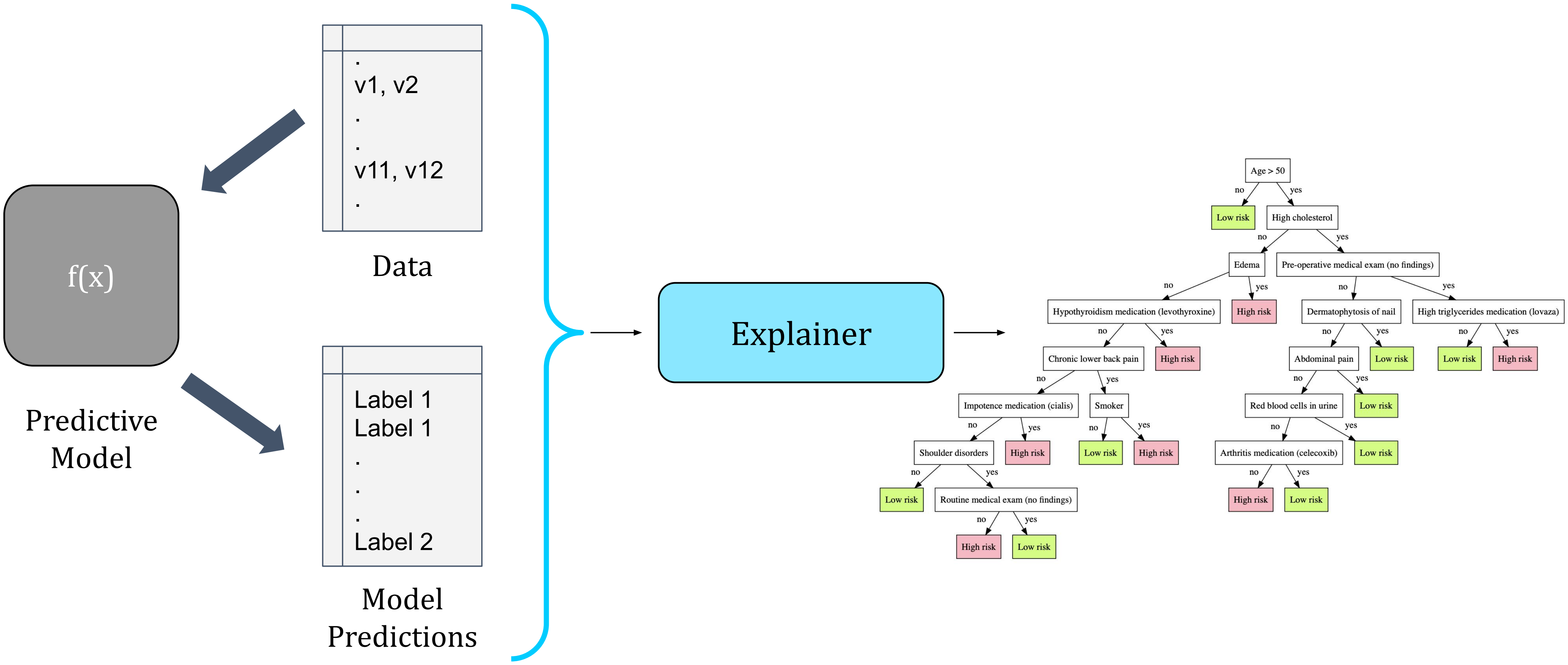
- Collection of Local Explanations
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

Model Distillation for Generating Global Explanations



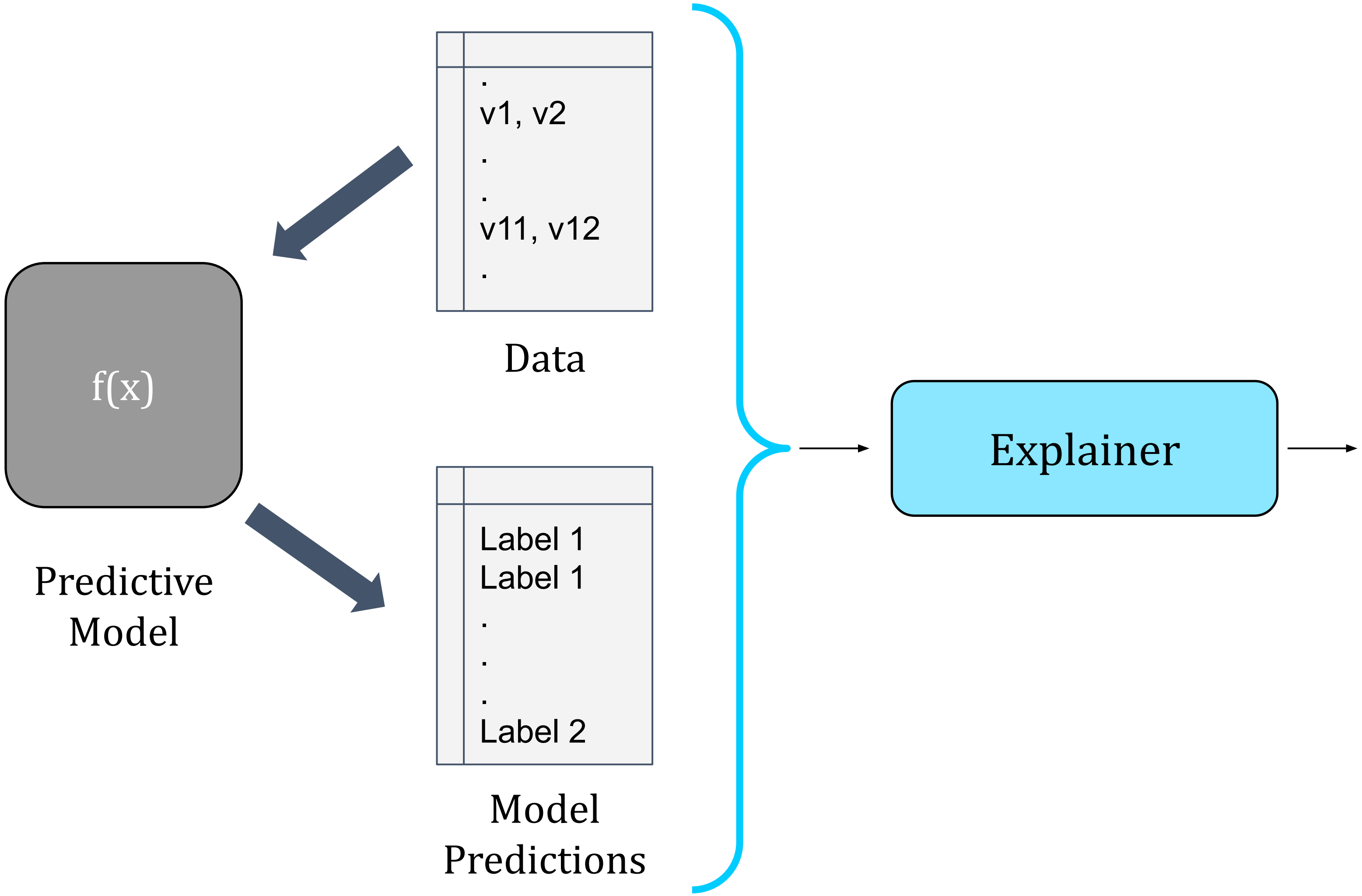
[Bastani et. al., 2019]

Decision Trees as Global Explanations



[Lakkaraju et. al., 2019]

Customizable Decision Sets as Global Explanations



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

[Tan et. al., 2019]

Generalized Additive Models as Global Explanations

f(x)

Predictive Model

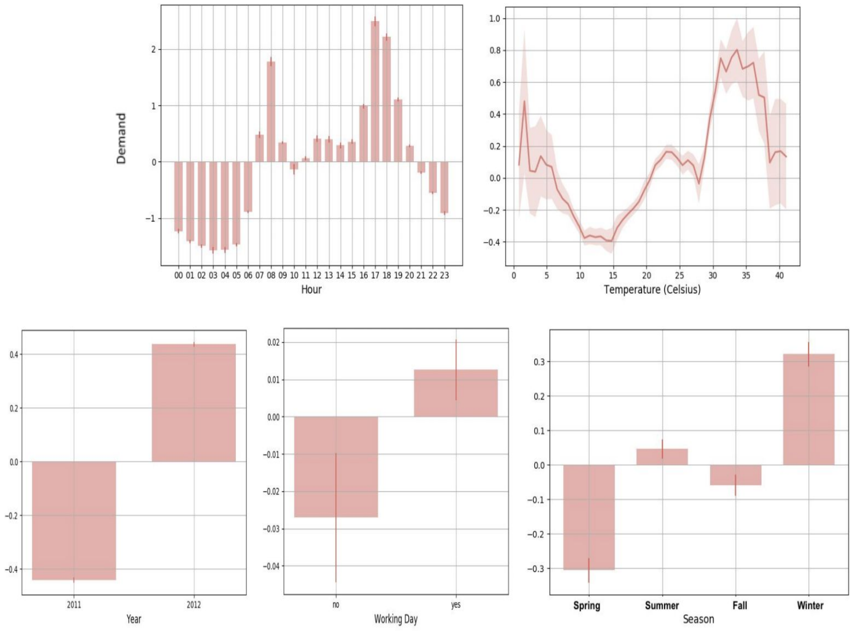
.
v1, v2
.
.
v11, v12
.

Data

Label 1
Label 1
.
.
.
Label 2

Model Predictions

Explainer



Approaches for Post hoc Explainability

Local Explanations

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

Global Explanations

- Collection of Local Explanations
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

[Rawal et. al., 2020]

Customizable Global Summaries of Counterfactuals

Predictive Model

f(x)

DENIED LOANS

Algorithm for generating global summaries of counterfactuals

If Race = Caucasian **and** Gender = Male:

If Married =No **and** Property =No and Has Job =No, **then** Married =No **and** Property =No and Has Job =Yes

If Drugs =Yes **and** School =No **and** Pays Rent =No , **then** Drugs =No **and** School =No **and** Pays Rent =No

If Race = Caucasian **and** Gender = Female:

If Married =No **and** Property =No and Has Job =No, **then** Married =No **and** Property =No and Has Job =Yes

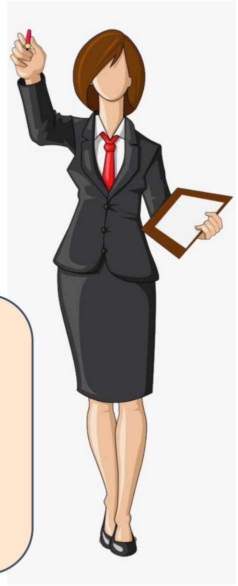
If Drugs =Yes **and** School =No **and** Pays Rent =No , **then** Drugs =No **and** School =No **and** Pays Rent =Yes

If Race $\neq$ Caucasian:

If Married =No **and** Property =No and Has Job =No, **then** Married =No **and** Property =Yes and Has Job =Yes

If Drugs =Yes **and** School =No **and** Pays Rent =No , **then** Drugs =No **and** School =Yes **and** Pays Rent =Yes

How do recourses permitted by the model vary across various racial & gender subgroups? Are there any biases against certain demographics?



Customizable Global Summaries of Counterfactuals

Subgroup Descriptor

Omg! this model is biased. It requires certain demographics to “change” lot more features than others.

If Race = Caucasian **and** Gender = Male:

If Married =No **and** Property =No and Has Job =No, **then** Married =No **and** Property =No and Has Job =Yes

If Drugs =Yes **and** School =No **and** Pays Rent =No , **then** Drugs =No **and** School =No **and** Pays Rent =No

If Race = Caucasian **and** Gender = Female:

If Married =No **and** Property =No and Has Job =No, **then** Married =No **and** Property =No and Has Job =Yes

If Drugs =Yes **and** School =No **and** Pays Rent =No , **then** Drugs =No **and** School =No **and** Pays Rent =Yes

Recourse Rules

If Race $\neq$ Caucasian:

If Married =No **and** Property =No and Has Job =No, **then** Married =No **and** Property =Yes and Has Job =Yes

If Drugs =Yes **and** School =No **and** Pays Rent =No , **then** Drugs =No **and** School =Yes **and** Pays Rent =Yes

Customizable Global Summaries of Counterfactuals

- An optimization problem which is *non-negative*, *non-normal*, *non-monotone*, and *submodular* with *matroid constraints*

- Solved using the well-known *smooth local search* algorithm (Feige et. al., 2007) with best known optimality guarantees.

Approaches for Post hoc Explainability

Local Explanations

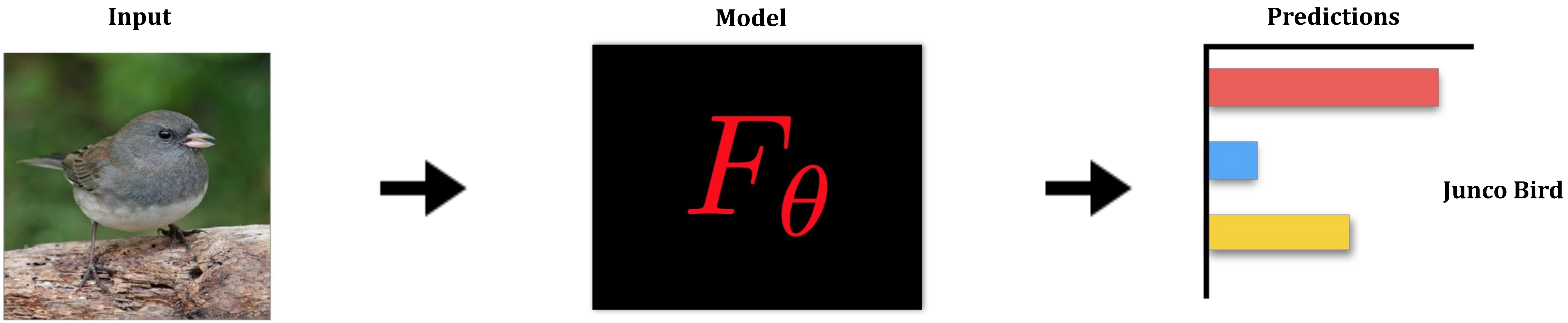
- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

- Collection of Local Explanations
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

Representation Based Approaches

- Derive model understanding by analyzing intermediate representations of a DNN.
- Determine model’s reliance on ‘concepts’ that are semantically meaningful to humans.



Does the model rely on the **‘green background’**?

Representation Based Approaches

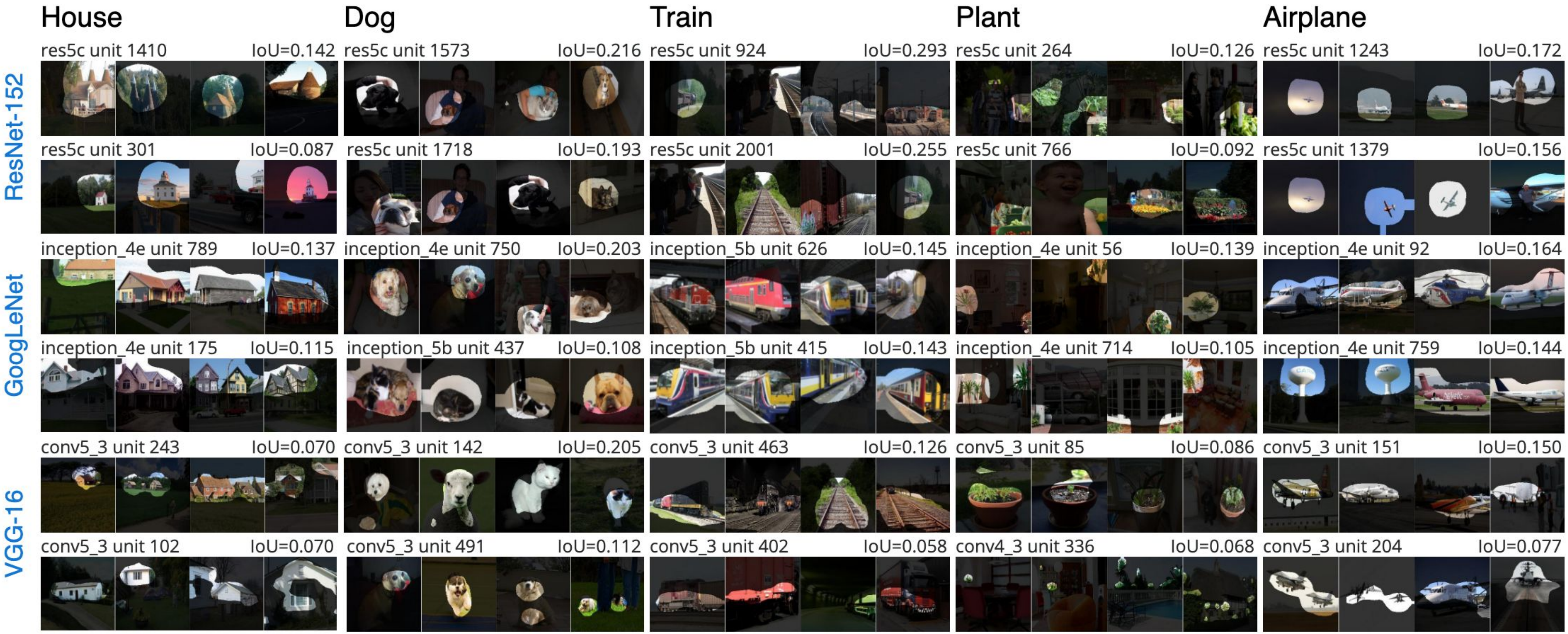
- Network Dissection (**Bau & Zhou et. al. 2017**).

- TCAV (**Kim et. al. 2018**).

Process

1. Identify human-labeled concepts.
2. Gather the response of hidden variables (convolutional filters) to known concepts.
3. Quantify alignment of hidden variable-concept pairs

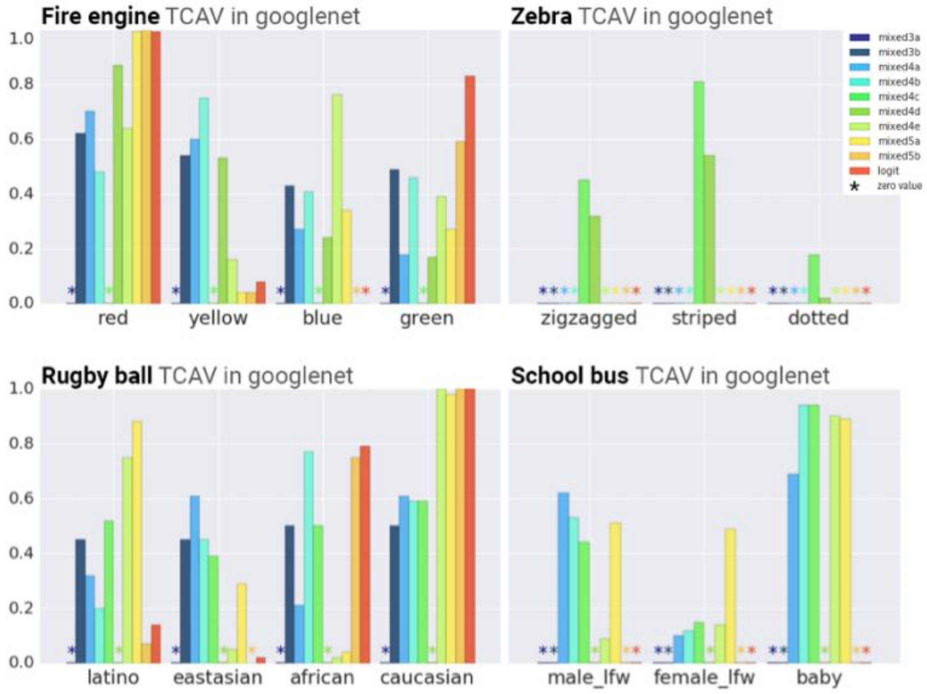
Network Dissection

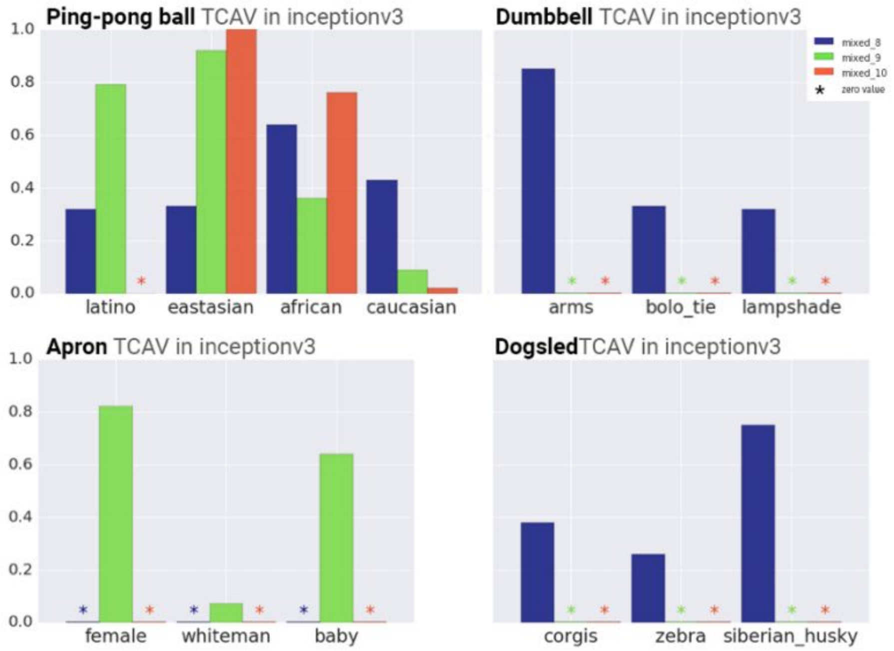


[Bau & Zhou et. al. 2017]

Quantitative Testing with Concept Activation Vectors (TCAV)

Insights from Googlenet and Inception-v3





Additional Variants:

- Regression problems in medical domain ([Graziani et. al. 2019](#)).
- Automatic extraction of visual concepts ([Ghorbani et. al. 2019](#)).

Images from **Kim et. al. 2018**

Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Evaluation of Post hoc Explanations

[Doshi-Velez and Kim, 2017]

How we evaluate explanations?

Evaluating Post hoc Explanations

Understand the Behavior

Help make decisions

Useful for Debugging

Evaluating Post hoc Explanations

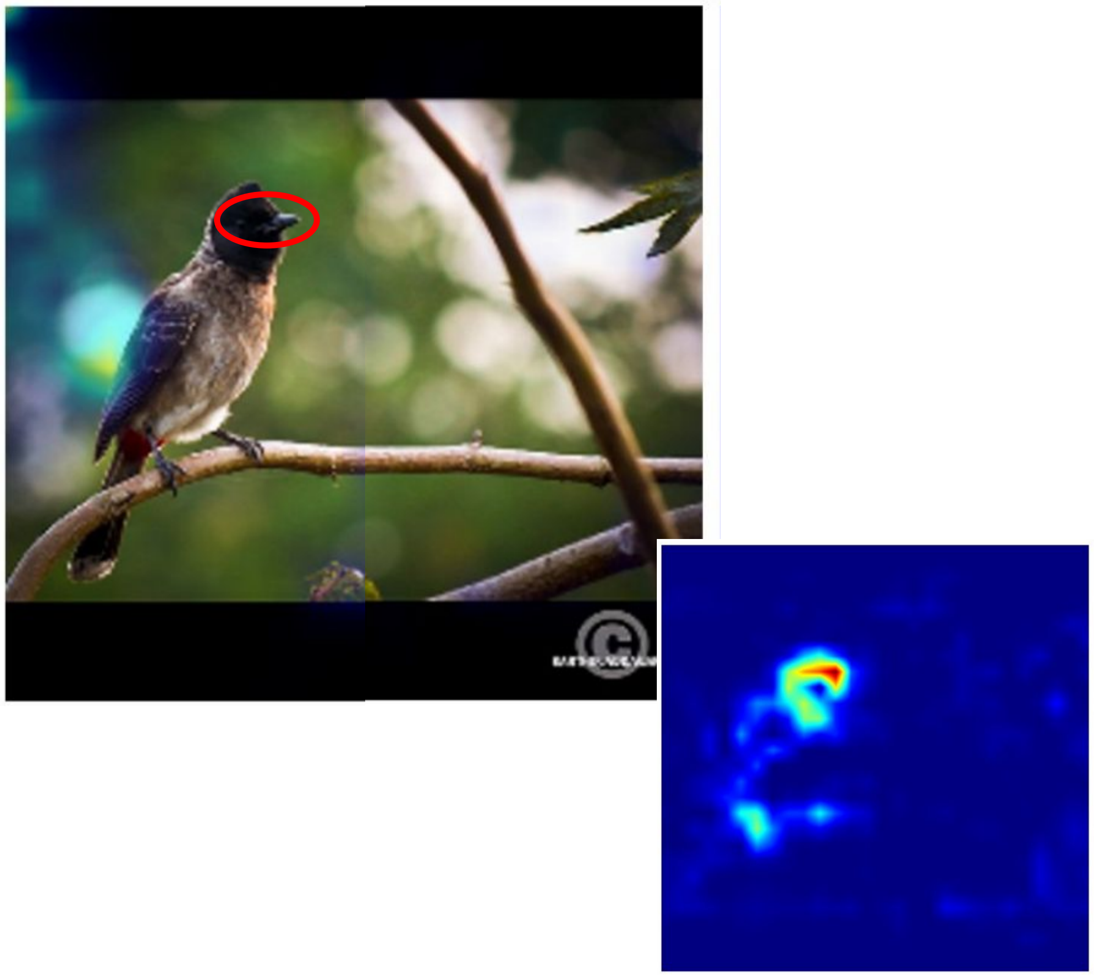
Understand the Behavior

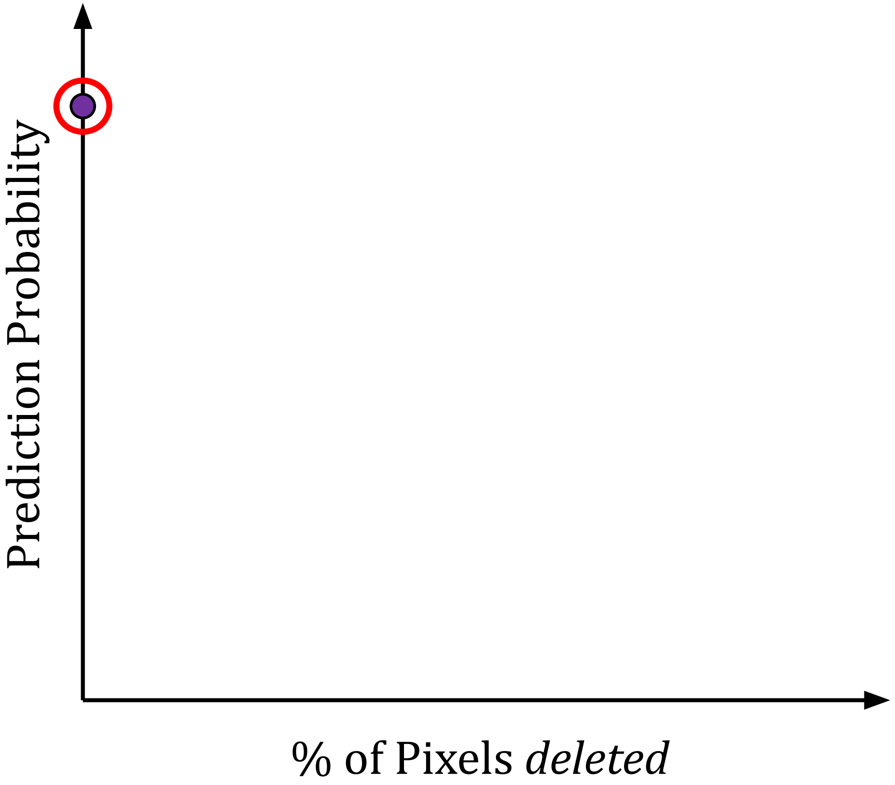
Help make decisions

Useful for Debugging

How important are selected features?

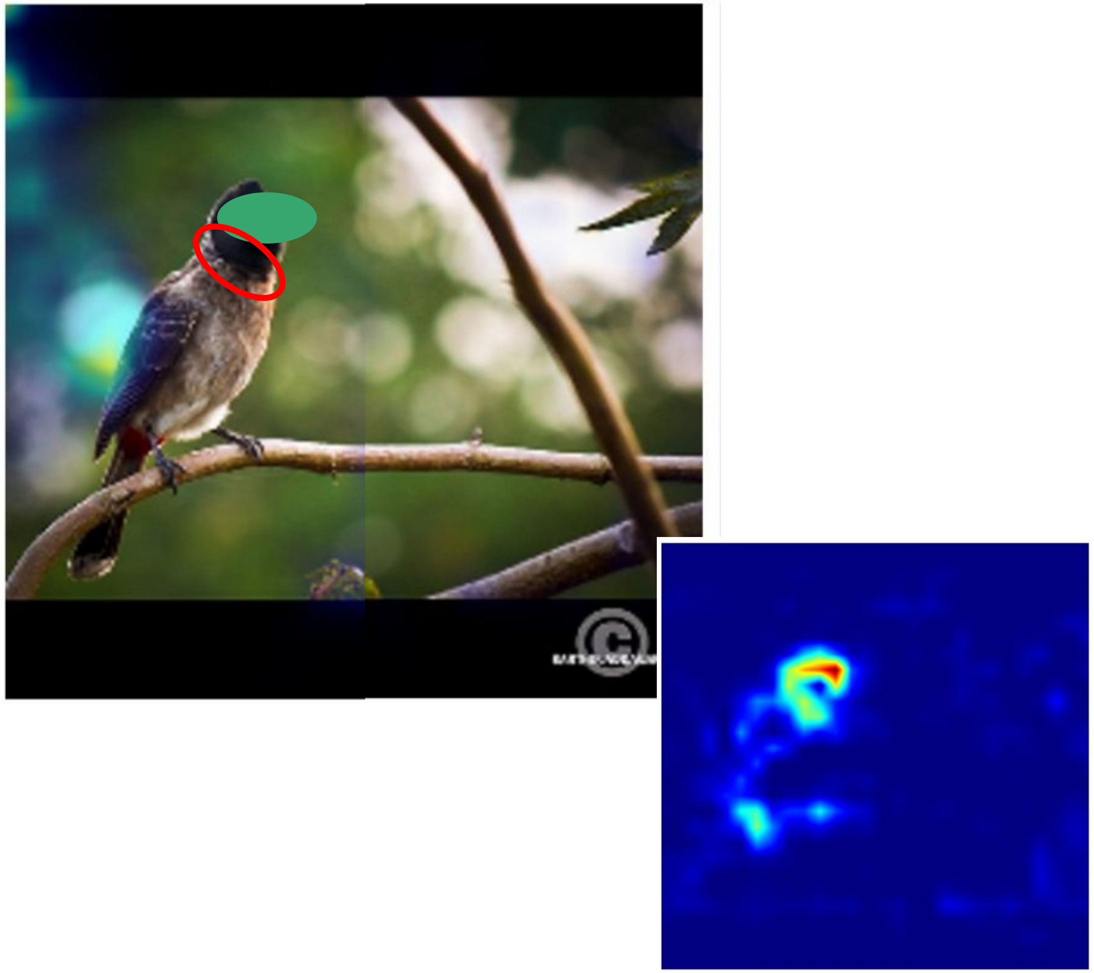
- Deletion: remove important features and see what happens..

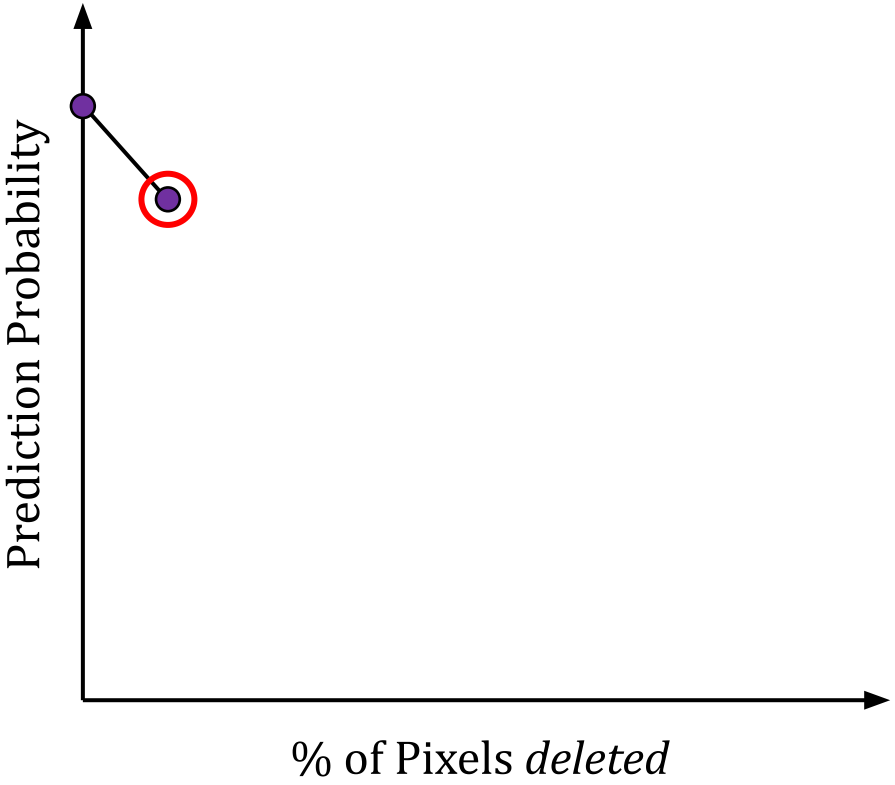




How important are selected features?

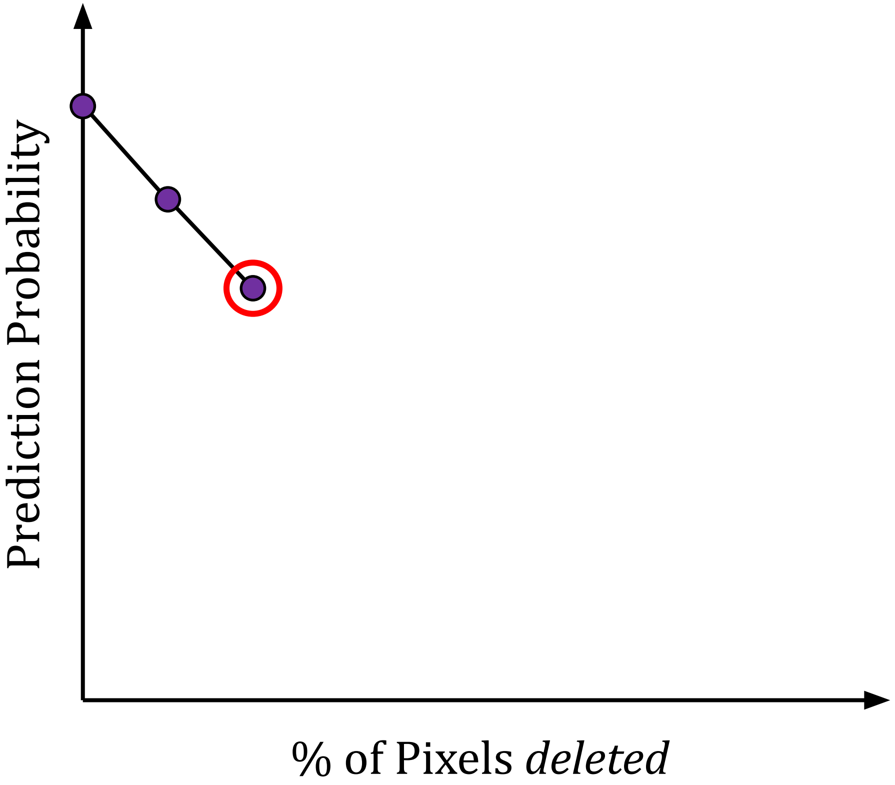
- Deletion: remove important features and see what happens..





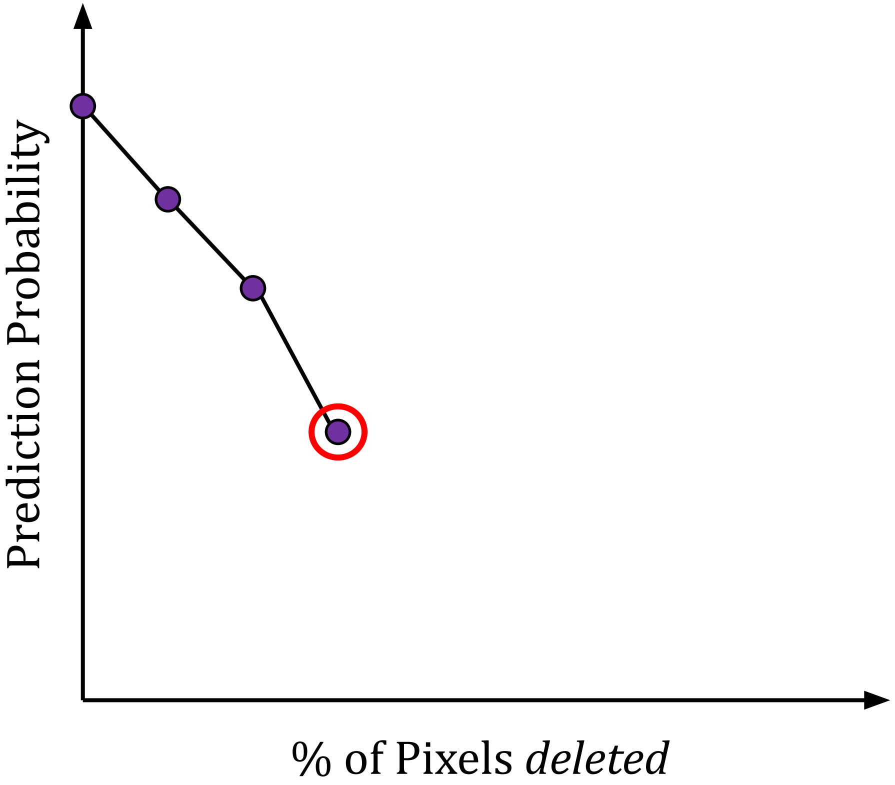
How important are selected features?

- Deletion: remove important features and see what happens..



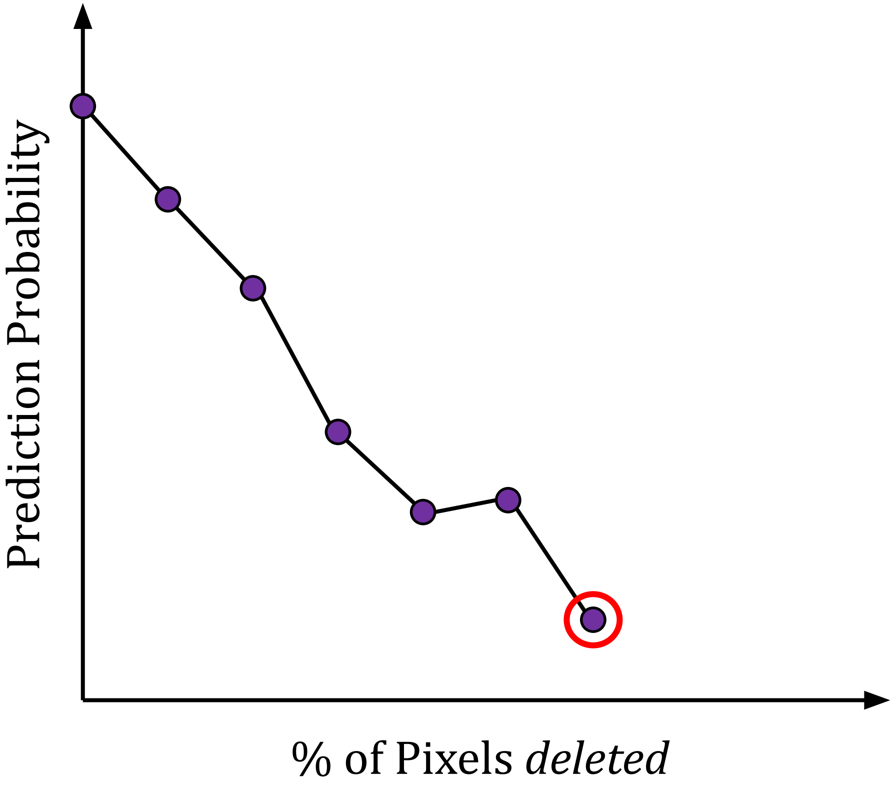
How important are selected features?

- Deletion: remove important features and see what happens..



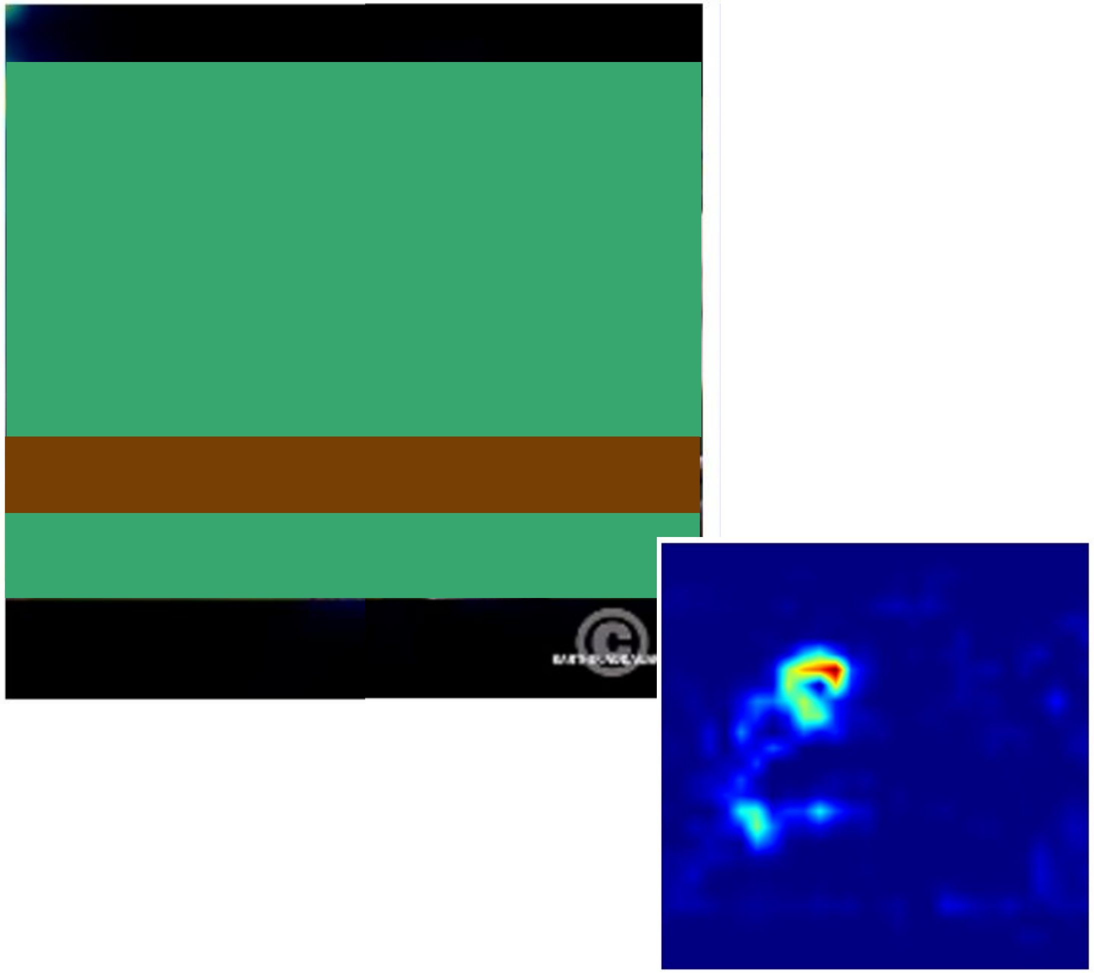
How important are selected features?

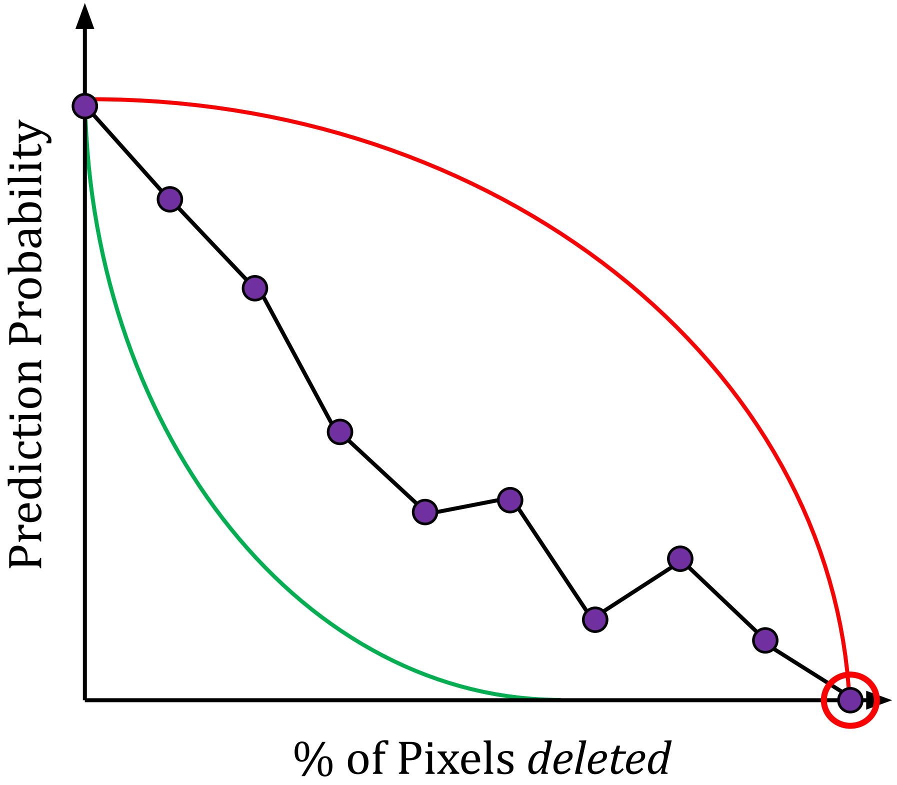
- Deletion: remove important features and see what happens..



How important are selected features?

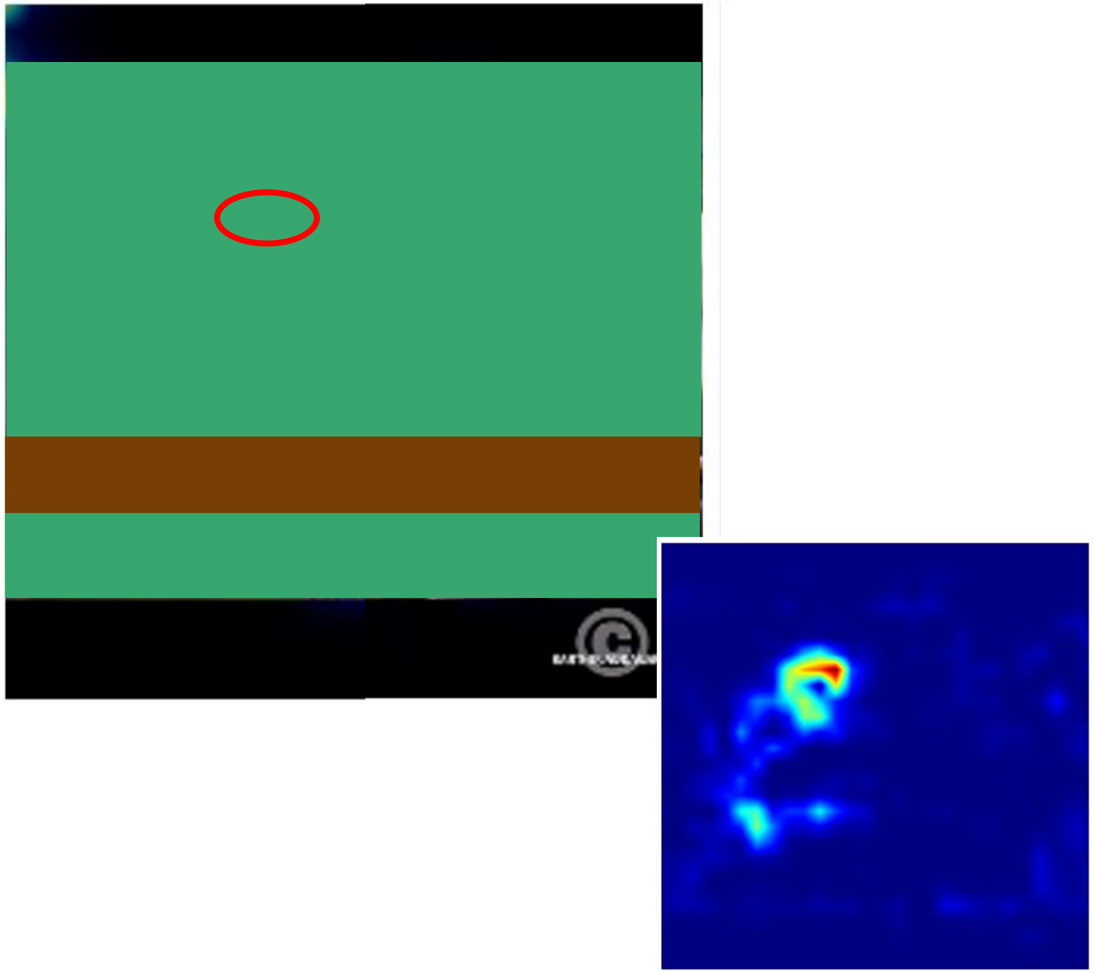
- Deletion: remove important features and see what happens..

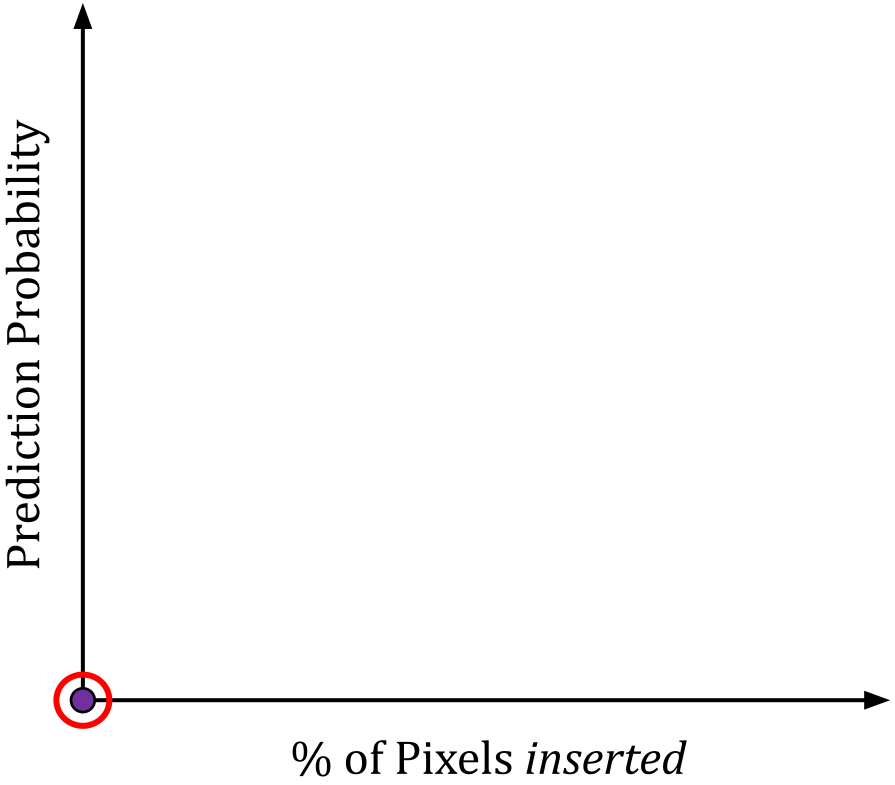




How important are selected features?

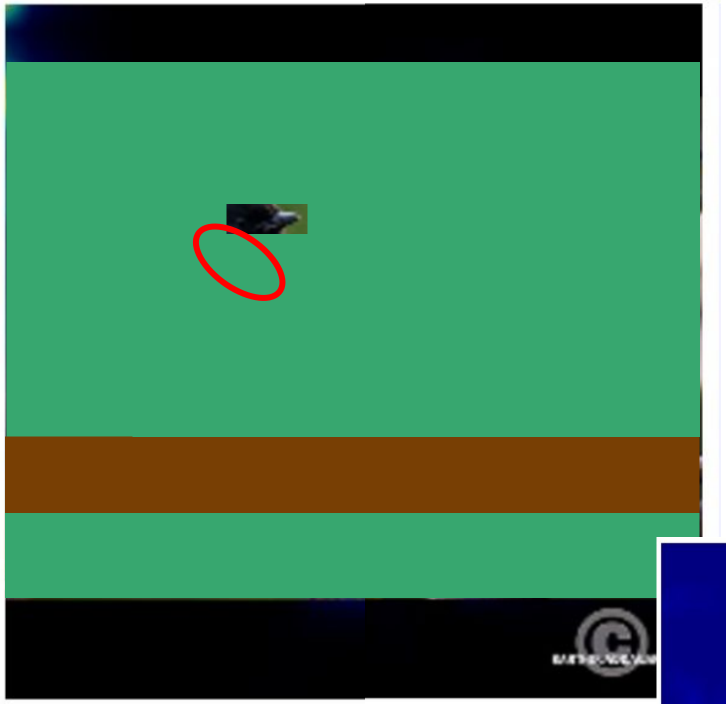
- Insertion: add important features and see what happens..

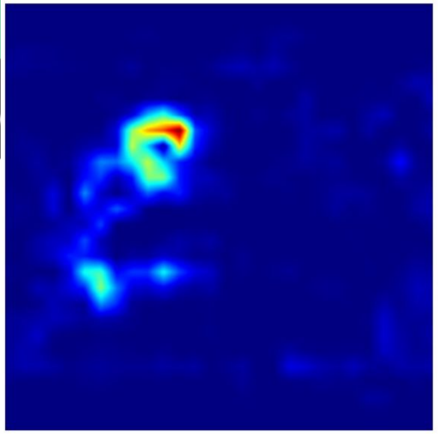


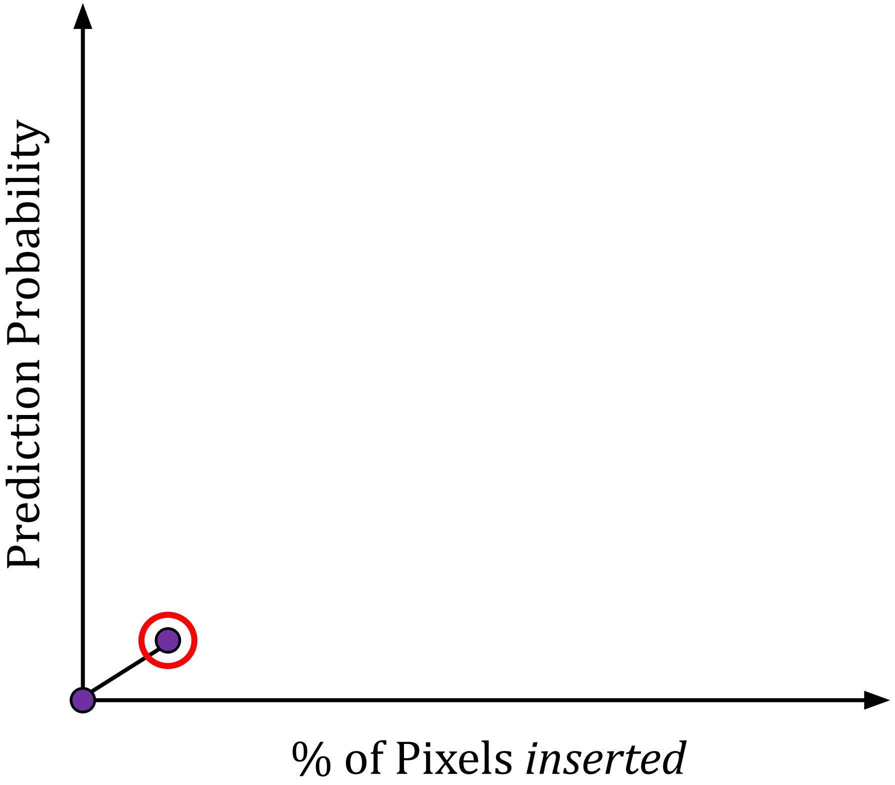


How important are selected features?

- Insertion: add important features and see what happens..

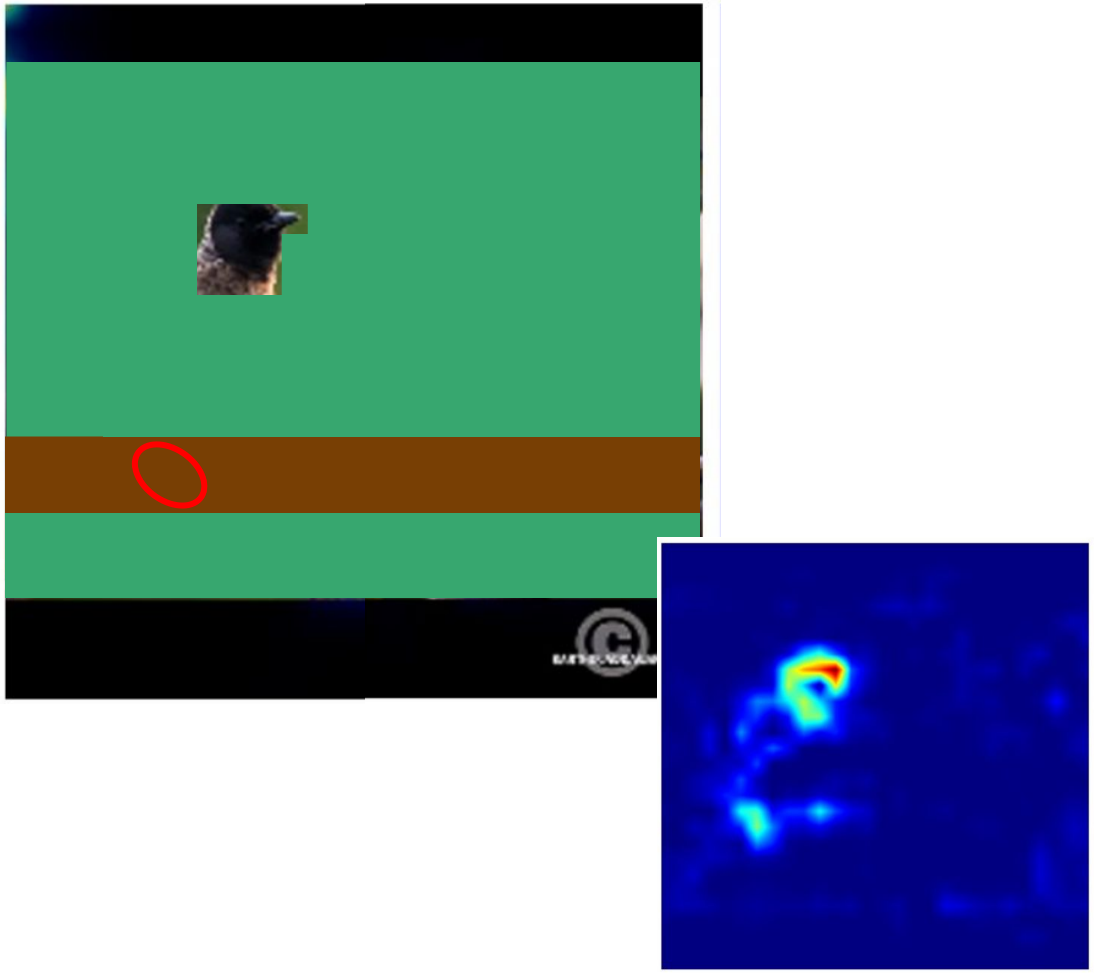


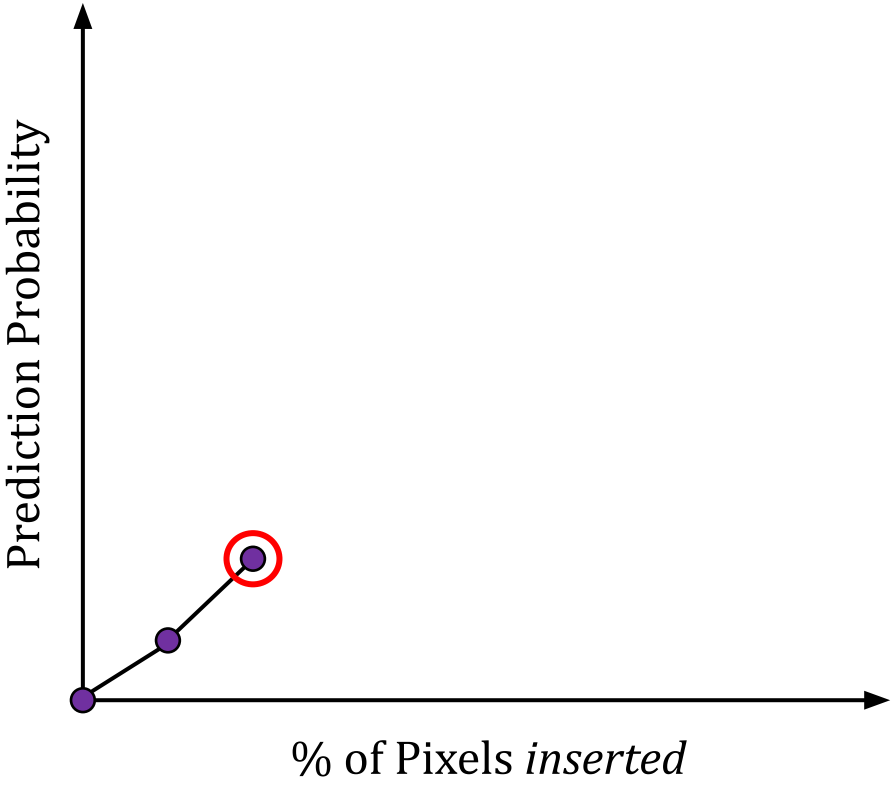




How important are selected features?

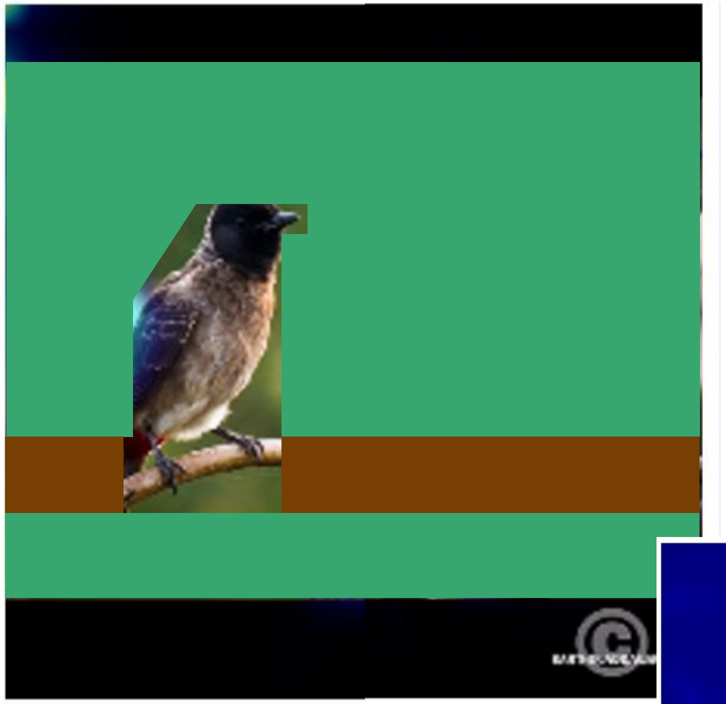
- Insertion: add important features and see what happens..

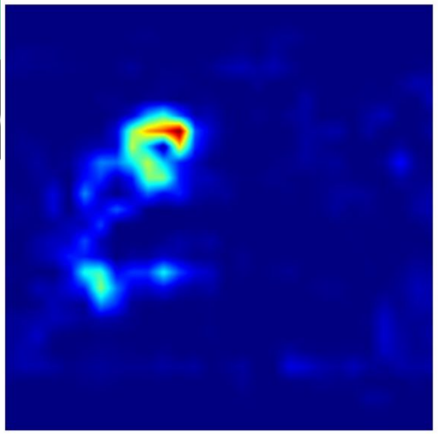


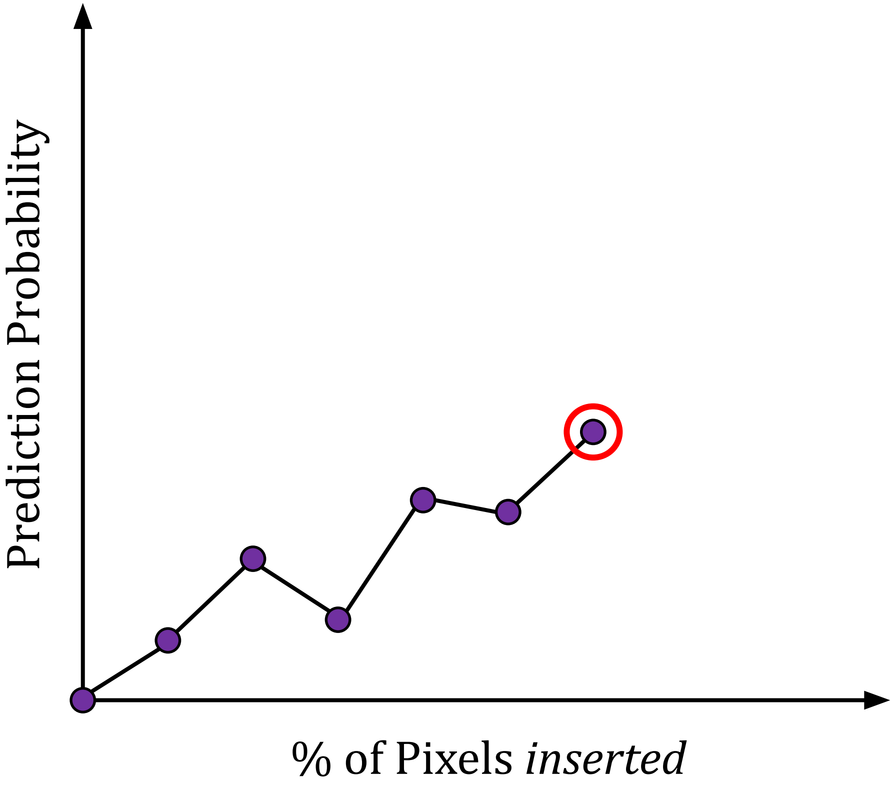


How important are selected features?

- Insertion: add important features and see what happens..

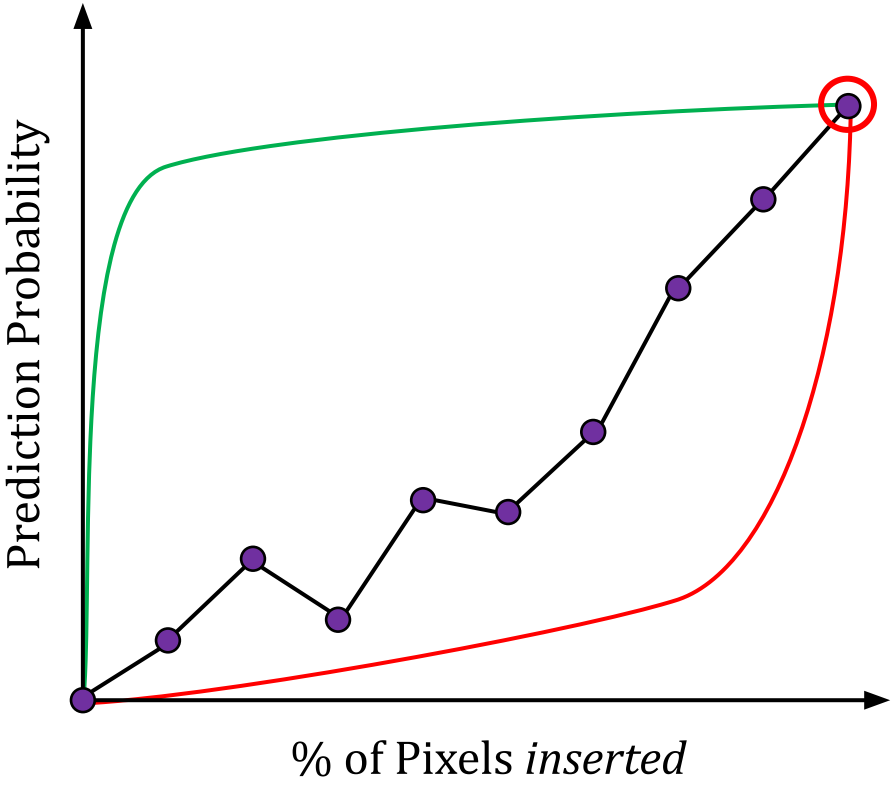






How important are selected features?

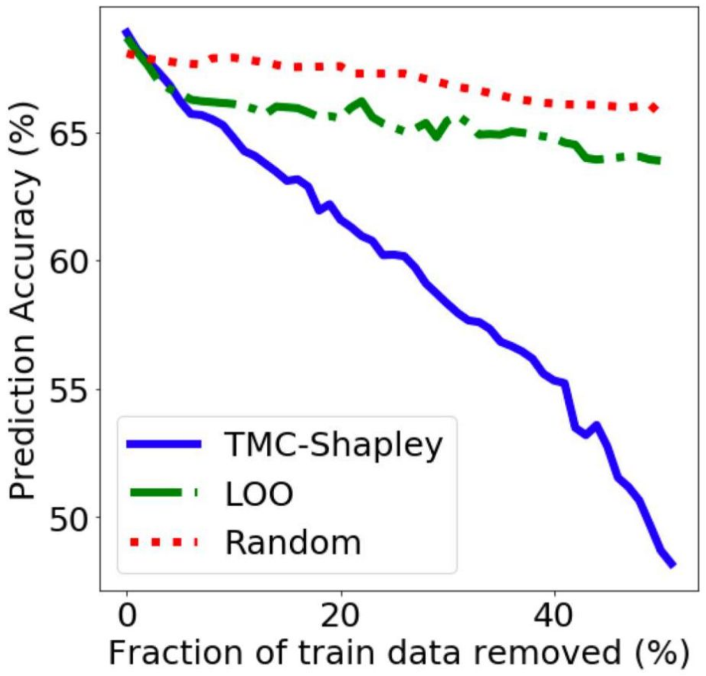
- Insertion: add important features and see what happens..



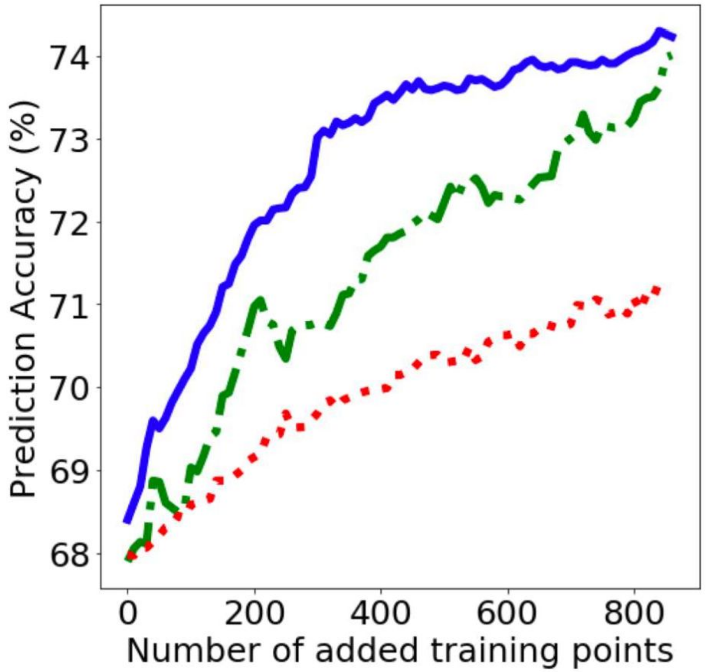
Same Idea: For *Training* Data

Add/remove influential training data, see what happens

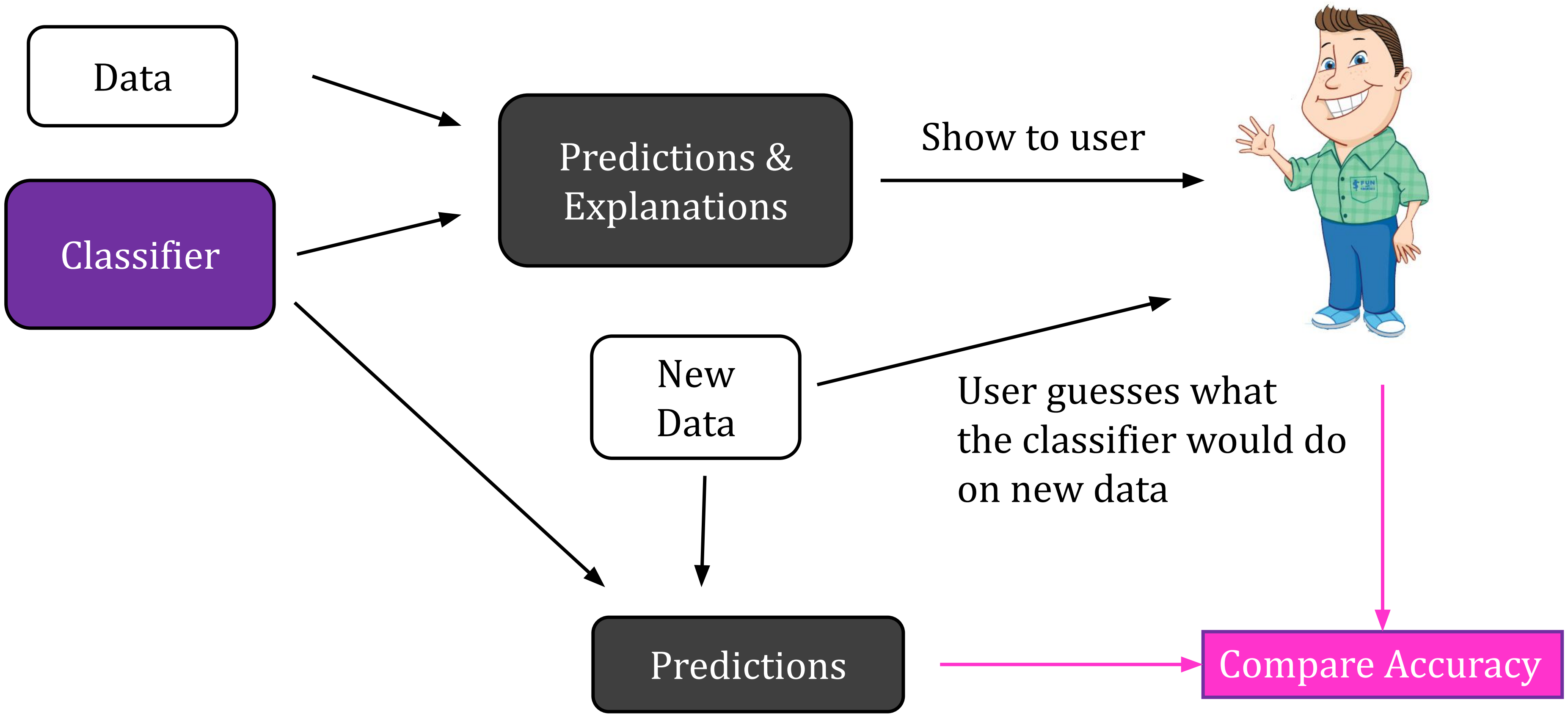
Removing high value data



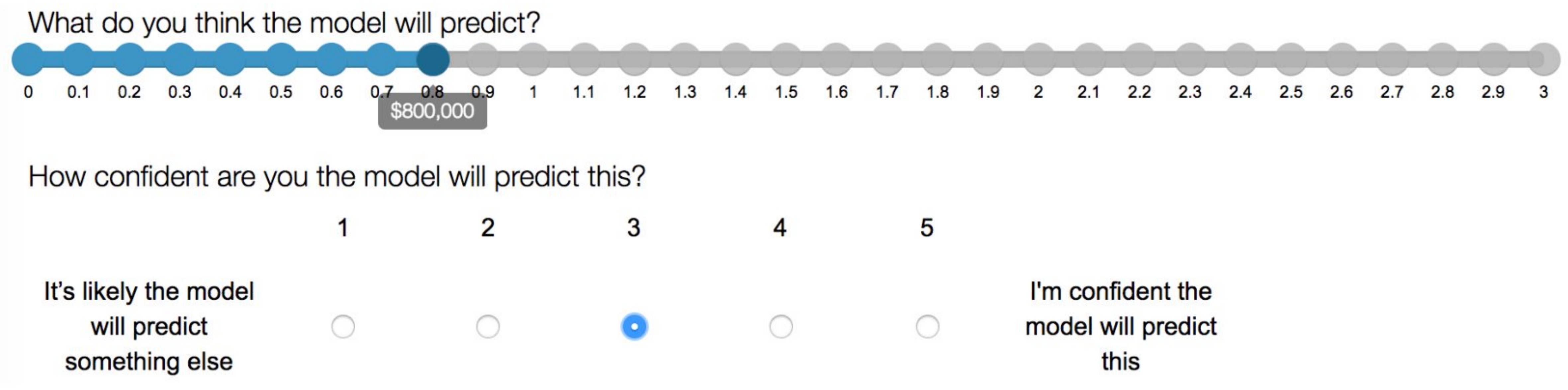
Adding high value data



Predicting Behavior (“Simulation”)



Predicting Behavior ("Simulation")



(a) Step 1: Participants were asked to guess the model's prediction and state their confidence.

Evaluating Post hoc Explanations

Understand the Behavior

Help make decisions

Useful for Debugging

[Ribeiro et al. 2016]

1. Detecting Problems in Classifiers

Classifier → Explainer → Show Explanations

Question 1

Would you trust this model?

Did they say no?

Question 2

What is the classifier doing?

Did they get it right?

2. Comparing Classifiers

Classifier → Explainer

Classifier → Explainer

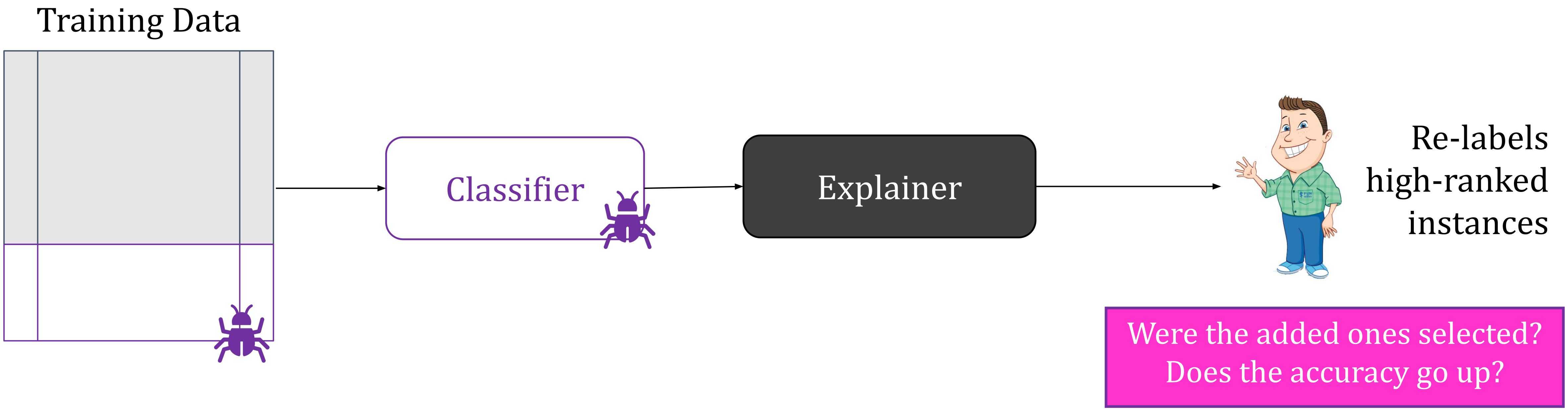
Show
Explanations

Question
Which algorithm is better?

Did they pick the right one?

3. Finding Errors in Training Data

- Prototypical Explanations: important instances from training data



Evaluating Posthoc Explanations

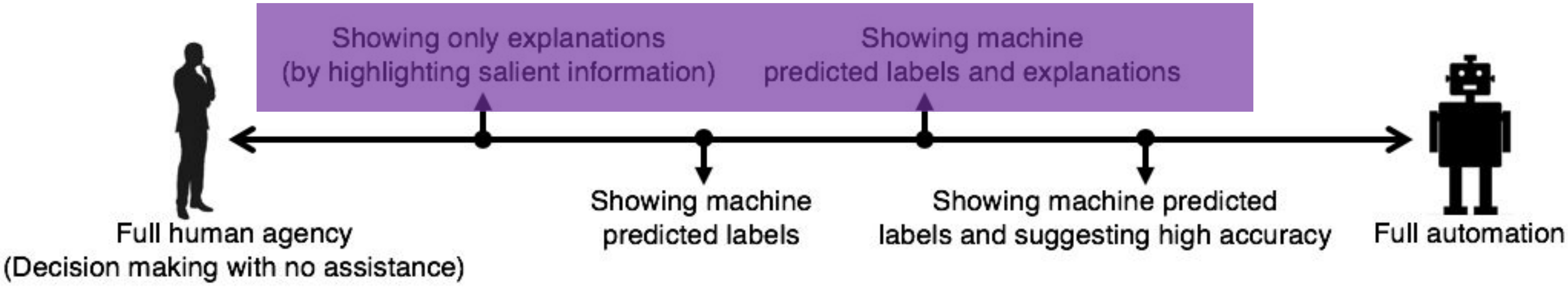
Understand the Behavior

Help make decisions

Useful for Debugging

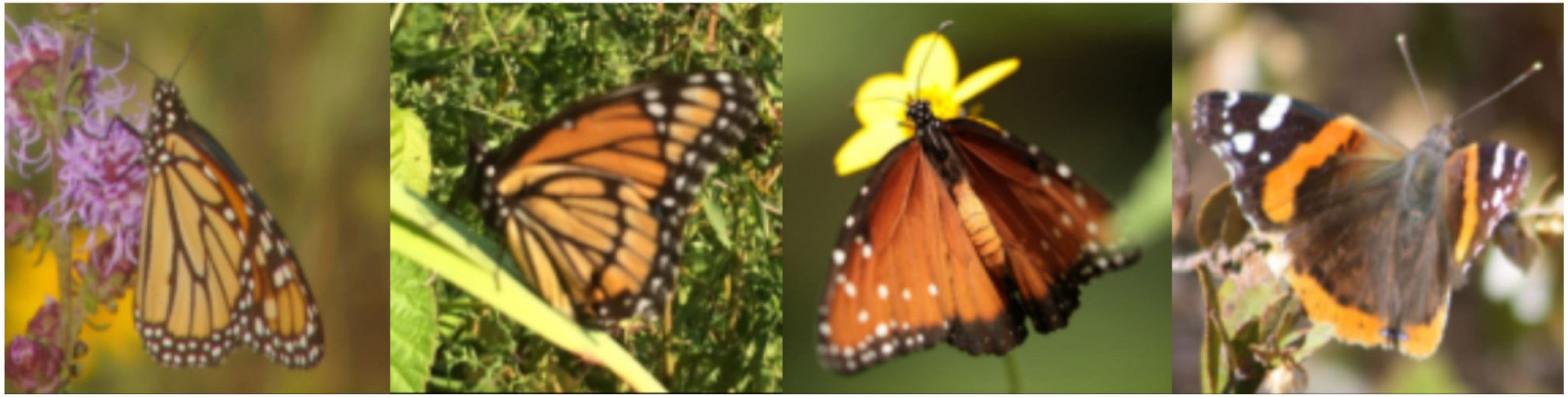
Human-AI Collaboration

- Are Explanations Useful for Making Decisions?
  - For tasks where the algorithms are not reliable by themselves

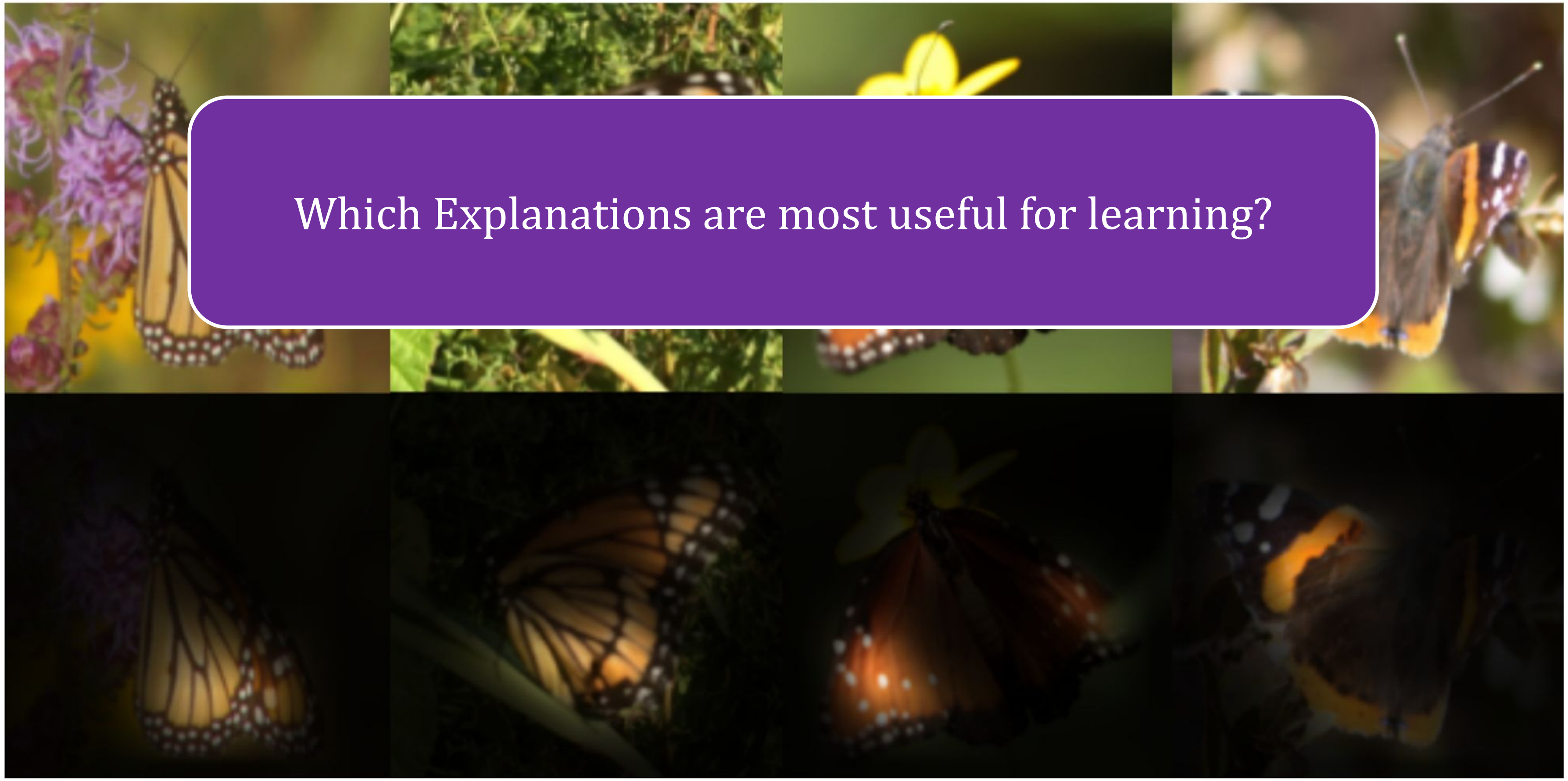


Machine Teaching

Monarch Viceroy Queen Red Admiral



Machine Teaching



Evaluating Posthoc Explanations

Understand the Behavior

Help make decisions

Useful for Debugging

Limitations of Evaluating Explanations

- Evaluation setup is often very easy/simple (or unrealistic)
  - E.g. "bugs" are obvious artifacts, classifiers are different from each other
  - Instances/perturbations create out-of-domain points
- Sometimes flawed
  - E.g. is model explanation same as human explanation?
- Automated metrics can be *optimized*
- User studies are not consistent
  - Affected by choice of: UI, phrasing, visualization, population, incentives, …
  - ML researchers are not trained for this ☹️
- Conclusions are difficult to generalize

Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Tutorial on Post hoc Explanations

Approaches for Post hoc Explainability

Evaluation of Explanations

Limits of Post hoc Explainability

Future of Post hoc Explainability

Limits of Post hoc Explanations

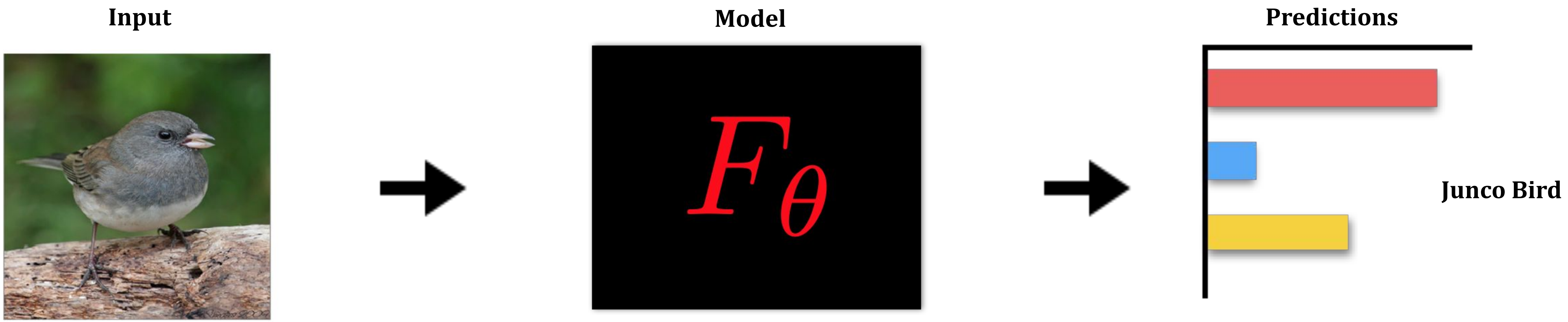
Limitations

- **Faithfulness/Fidelity**
  - Some explanation methods do not ‘*reflect*’ the underlying model.
- **Fragility**
  - Post-hoc explanations can be easily manipulated.
- **Stability**
  - Slight changes to inputs can cause large changes in explanations.
- **Useful in practice?**
  - Unclear if a data scientist (ML engineer)/end-user can use explanations to isolate errors, improve ‘trust’ or simulate the model.

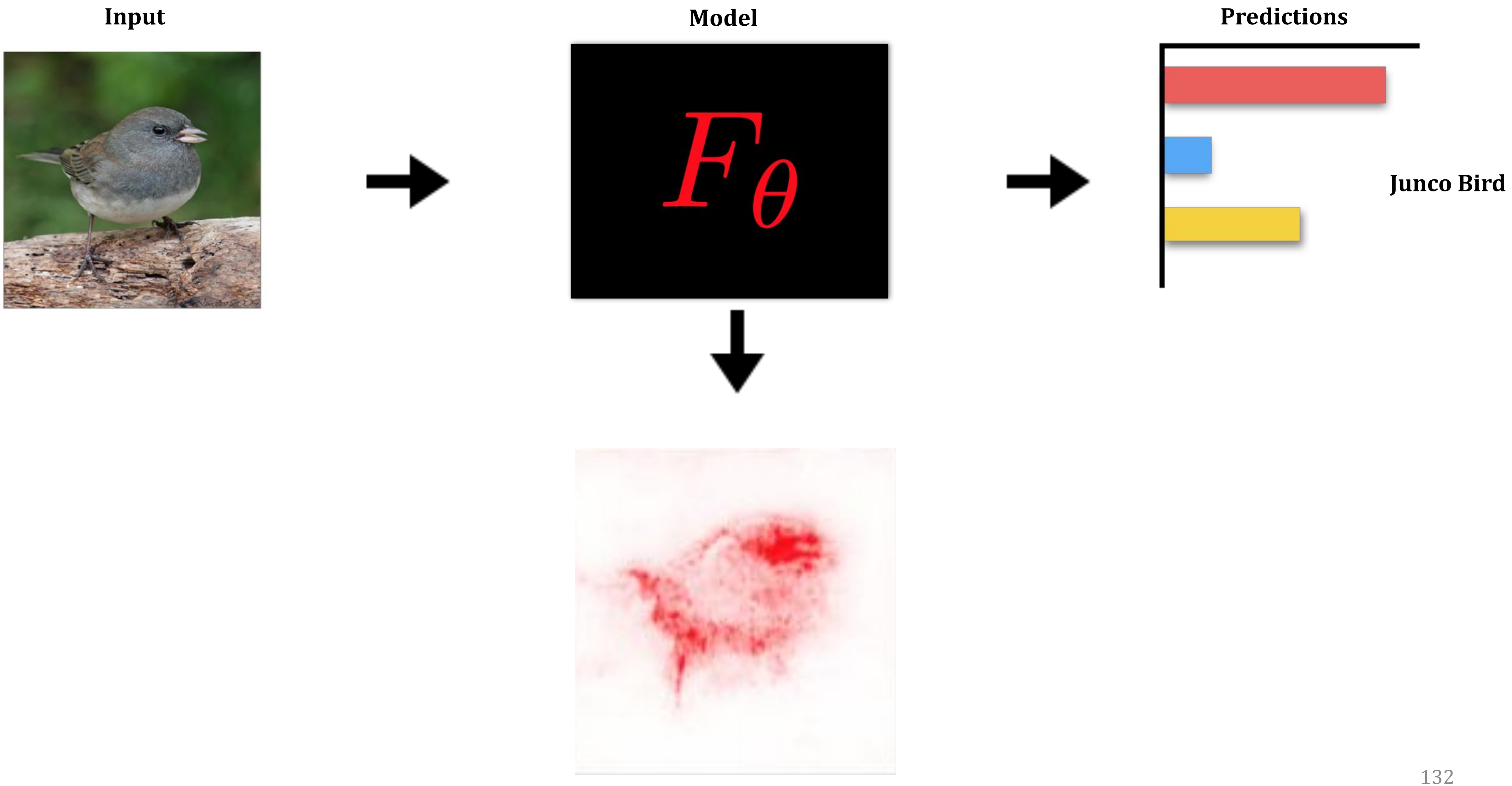
Limitations

- **Faithfulness/Fidelity**
  - Some explanation methods do not ‘*reflect*’ the underlying model.

Do Explanations Capture Model-based Discriminative Signals?



Do Explanations Capture Model-based Discriminative Signals?

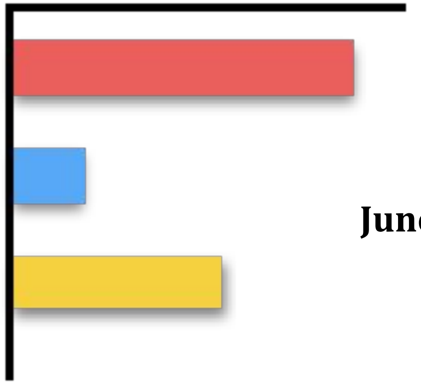


Do Explanations Capture Model-based Discriminative Signals?

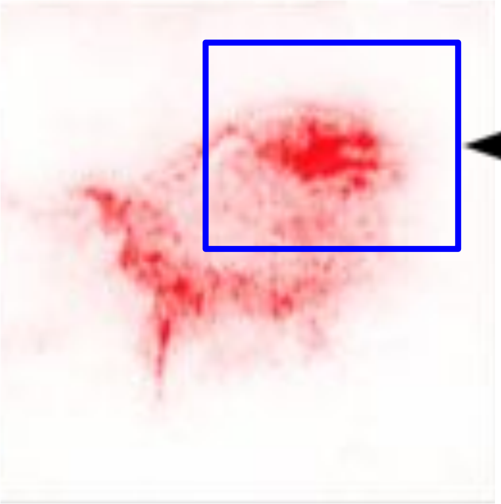
Input

Model

Predictions



Junco Bird



does the model, indeed, rely on these input dimensions to determine the output?

Faithfulness/Fidelity

Does the output of an explanation method reflect the underlying ***‘computation or behavior’*** of the black-box model?

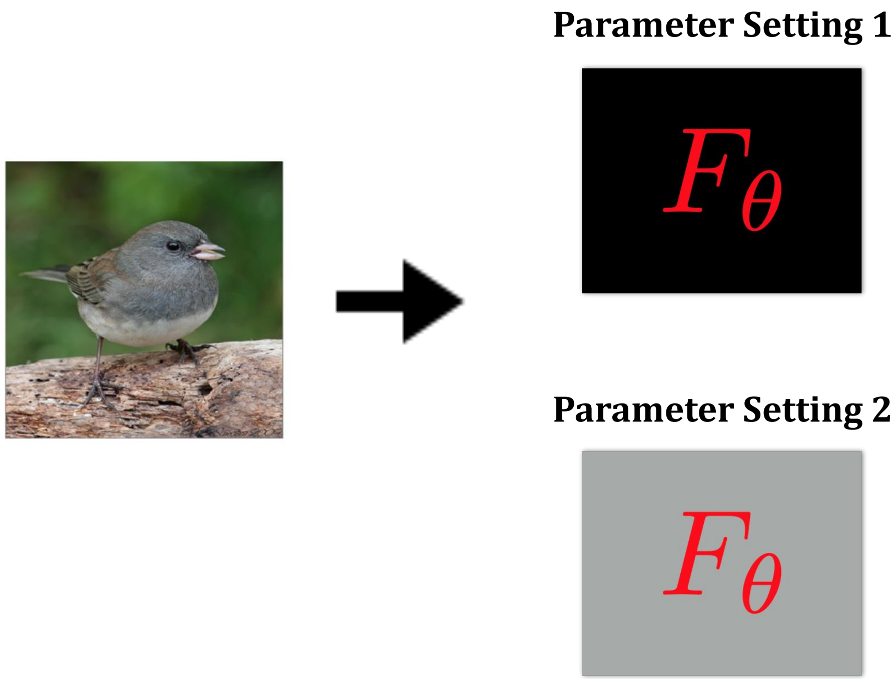
Sanity Check for Faithfulness/Fidelity

- **Sensitivity to Model Parameters:** if the parameter settings change, the explanations should change.

Adebayo et. al. 2018

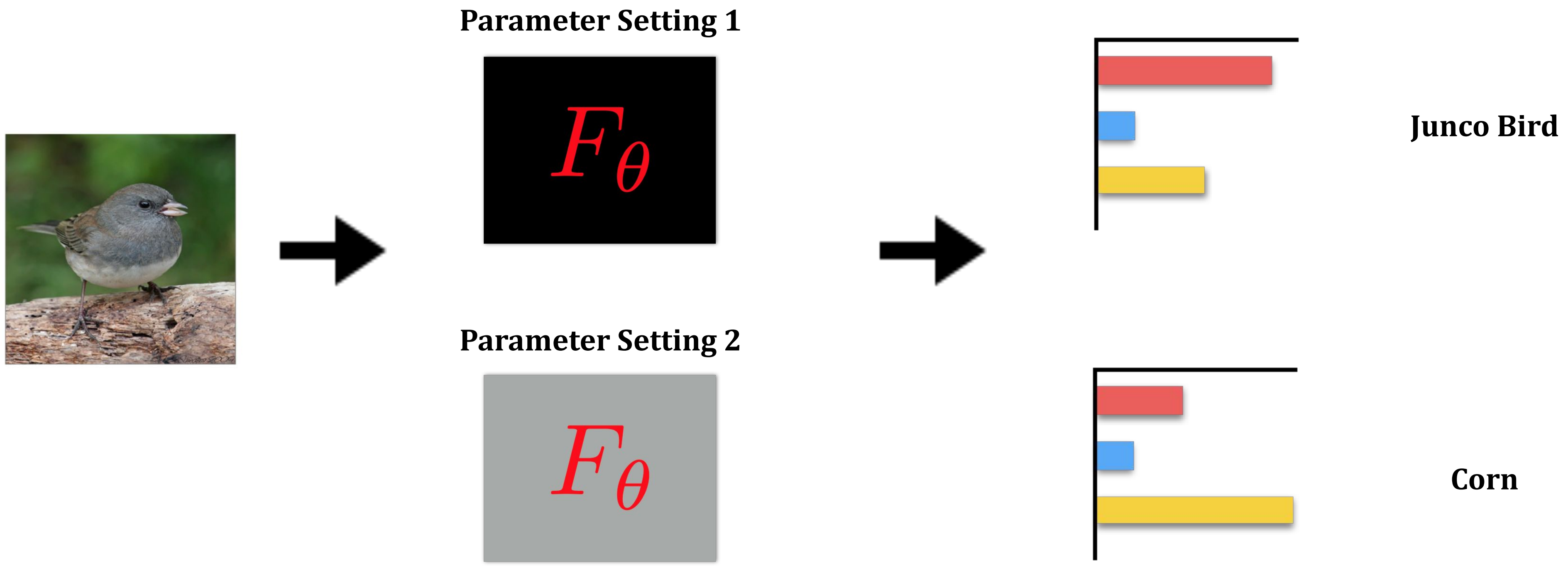
Sanity Check for Faithfulness/Fidelity

- **Sensitivity to Model Parameters:** if the parameter settings change, the explanations should change.



Sanity Check for Faithfulness/Fidelity

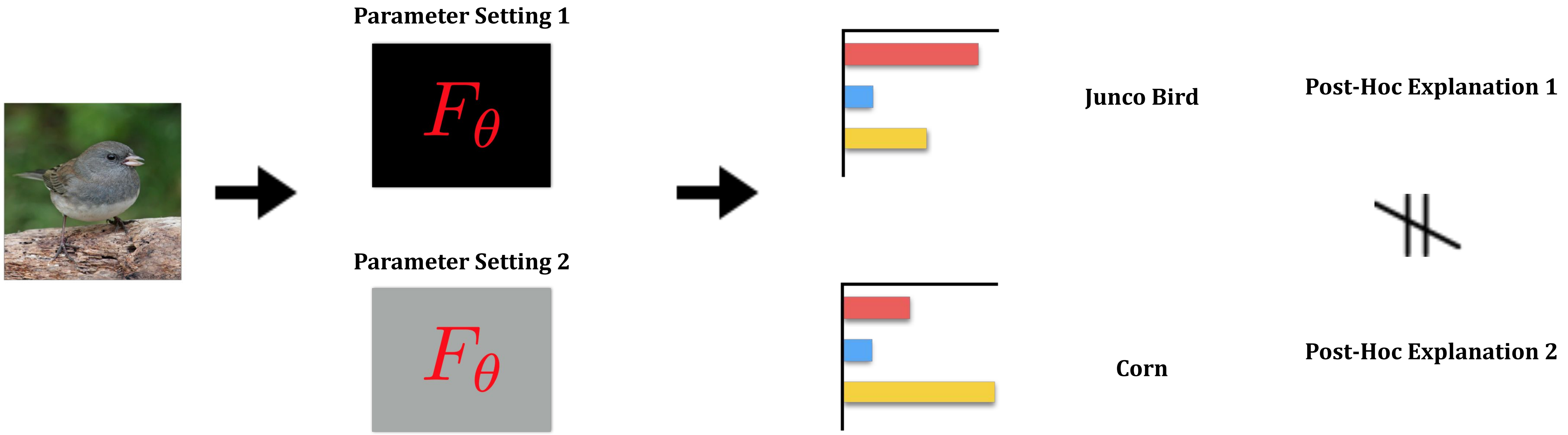
- **Sensitivity to Model Parameters:** if the parameter settings change, the explanations should change.



Adebayo et. al. 2018

Sanity Check for Faithfulness/Fidelity

- **Sensitivity to Model Parameters:** if the parameter settings change, the explanations should change.



[Adebayo et. al. 2018]

Cascading Randomization Inception-V3

- **Randomize (re-initialize)** model parameters starting from top layer all the way to the input.

Guided BackProp Explanation Inception-V3 ImageNet

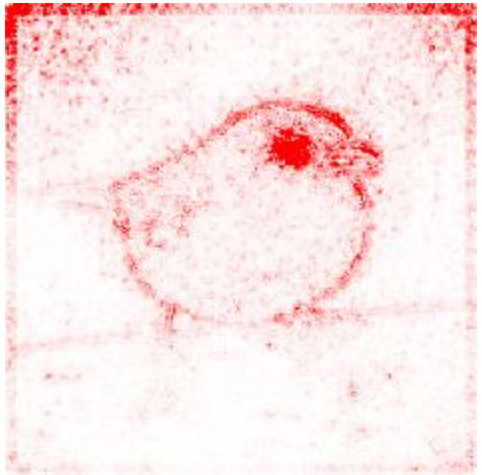
Adebayo et. al. 2018

Cascading Randomization Inception-V3

- **Randomize (re-initialize)** model parameters starting from top layer all the way to the input.

Guided BackProp Explanation Inception-V3 ImageNet

Normal Model Explanation



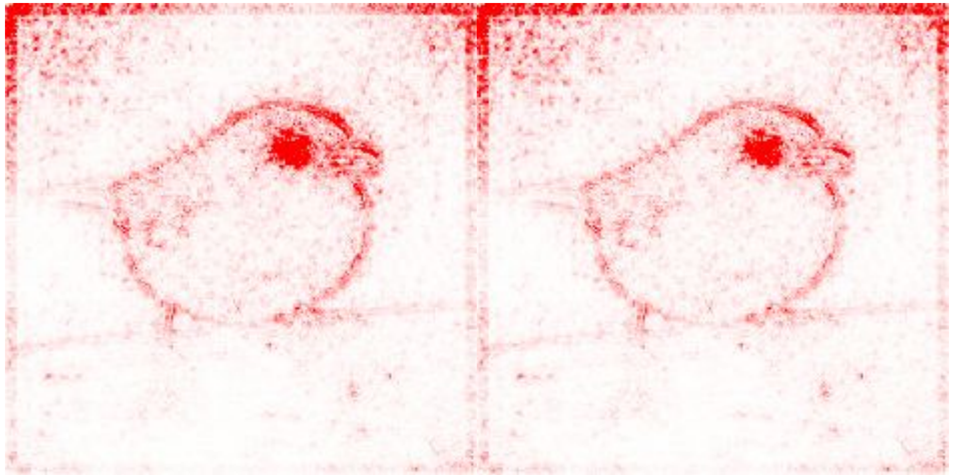
[Adebayo et. al. 2018]()

Cascading Randomization Inception-V3

- **Randomize (re-initialize)** model parameters starting from top layer all the way to the input.

Guided BackProp Explanation Inception-V3 ImageNet

Normal Model Explanation | **Top Layer Randomized**



[Adebayo et. al. 2018](#)

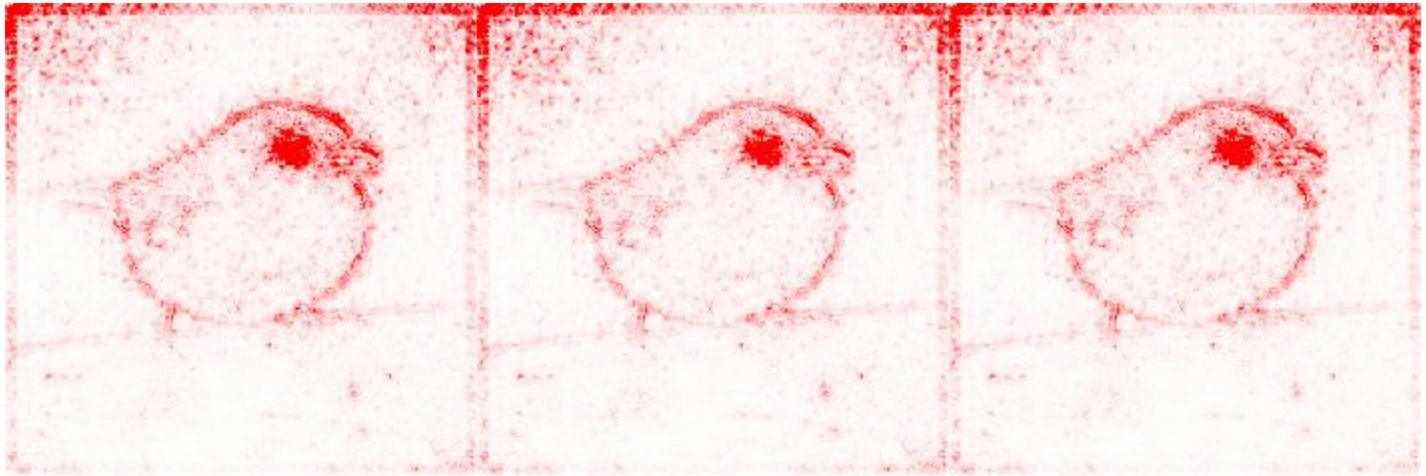
Cascading Randomization Inception-V3

- **Randomize (re-initialize)** model parameters starting from top layer all the way to the input.

Guided BackProp Explanation Inception-V3 ImageNet

Normal Model Explanation

Top Layer Randomized



[Adebayo et. al. 2018](#)

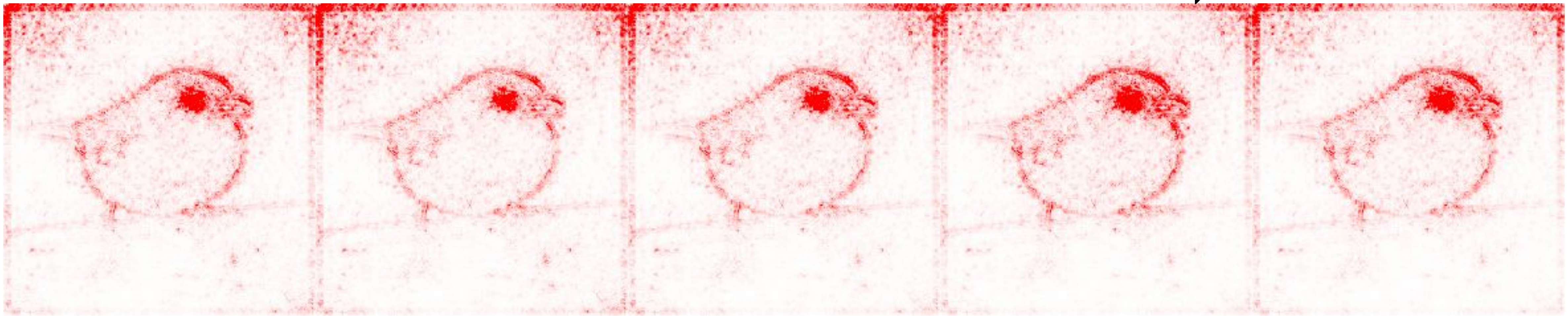
Cascading Randomization Inception-V3

- **Randomize (re-initialize)** model parameters starting from top layer all the way to the input.

Guided BackProp Explanation Inception-V3 ImageNet

Normal Model Explanation

Top Layer Randomized

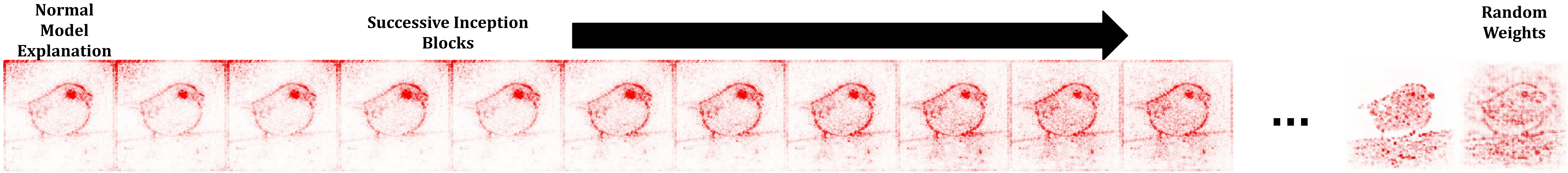


[Adebayo et. al. 2018](#)

Cascading Randomization Inception-V3

- **Randomize (re-initialize)** model parameters starting from top layer all the way to the input.

Guided BackProp Explanation Inception-V3 ImageNet

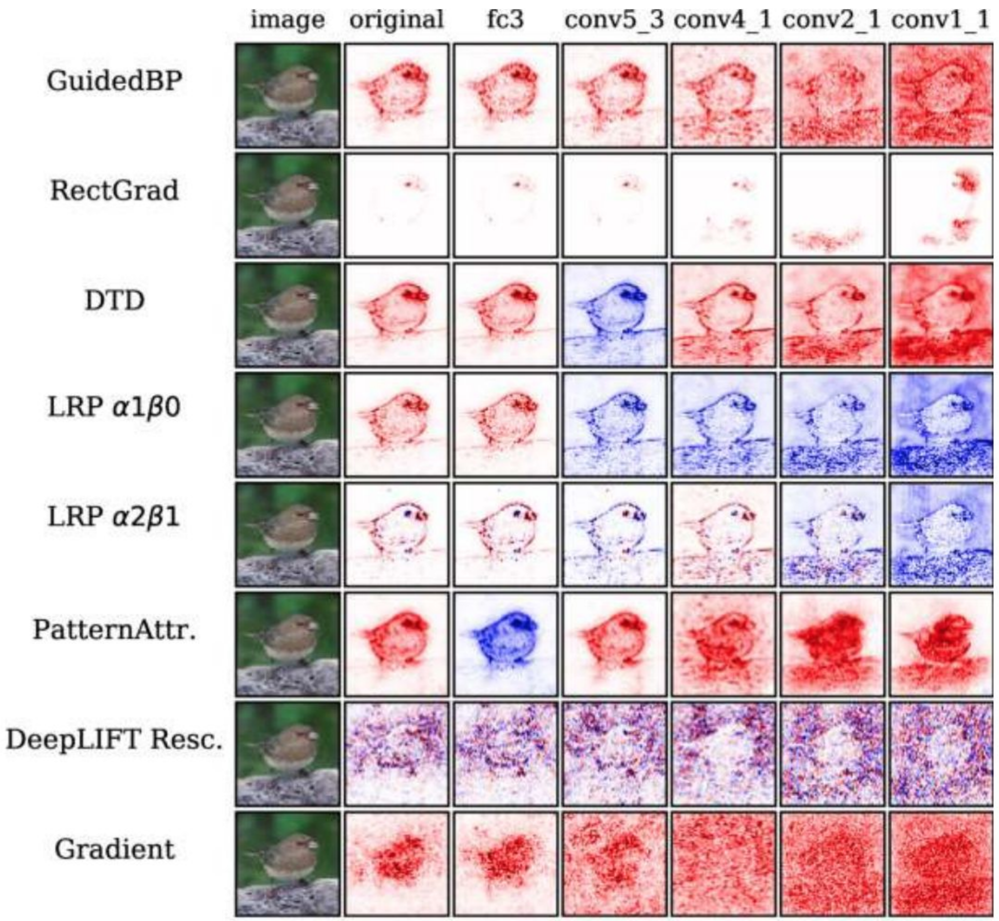


Guided BackProp is invariant to the higher level weights.

[Adebayo et. al. 2018](#)

‘Modified backprop approaches’ are invariant

Method that compute relevance via modified backpropagation and performance positive aggregation along the way are invariant to higher layers.



[Sixt et. al. 2020]

Source of Invariance

- **Guided BackProp and DeConvNet seek to approximately reconstruct the input** (Nie et. al. 2018).

- **These modified backprop methods converge to a rank-1 matrix**! This is because the product of a sequence of non-negative matrices (non-orthogonal columns, along with other assumptions) converges to a rank-1 matrix (*Theorem 1 in Sixt et. al. 2020*).

Source of Invariance

- **Guided BackProp and DeConvNet seek to approximately reconstruct the input** ([Nie et. al. 2018](#)).

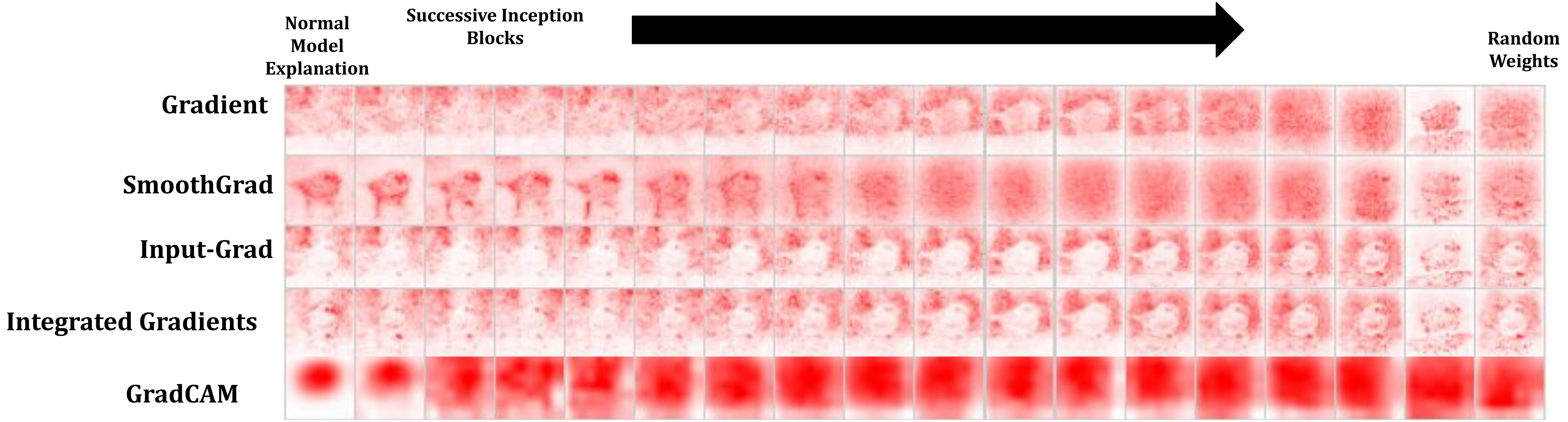
- **These modified backprop methods converge to a rank-1 matrix**! This is because the product of a sequence of non-negative matrices (non-orthogonal columns, along with other assumptions) converges to a rank-1 matrix (*[Theorem 1 in Sixt et. al. 2020](#)*).

\- -

- **DeConvNet**
- **Guided BackProp**
- **Guided GradCAM**

- **Deep Taylor Decomposition**
- **Pattern Net and Pattern Attribution (empirically)**
- **RectGrad**

Cascading Randomization Inception-V3



[Adebayo et. al. 2018](#)

Limitations

- ~~**Faithfulness/Fidelity**~~
  - ~~Some explanation methods do not '*reflect*' the underlying model.~~

- **Fragility**
  - Post-hoc explanations can be easily manipulated.

Post-hoc Explanations are Fragile

Post-hoc explanations can be easily manipulated.

Dombrowski et. al. 2019

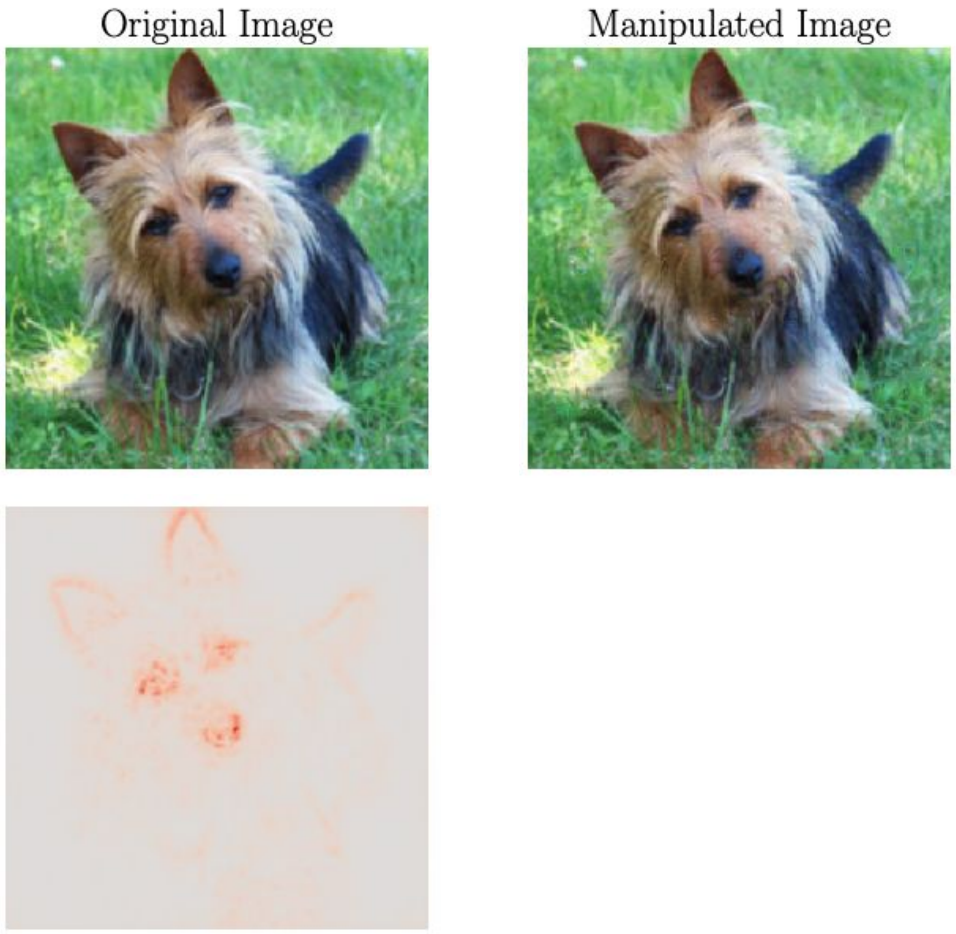
Post-hoc Explanations are Fragile

Post-hoc explanations can be easily manipulated.

Dombrowski et. al. 2019

Post-hoc Explanations are Fragile

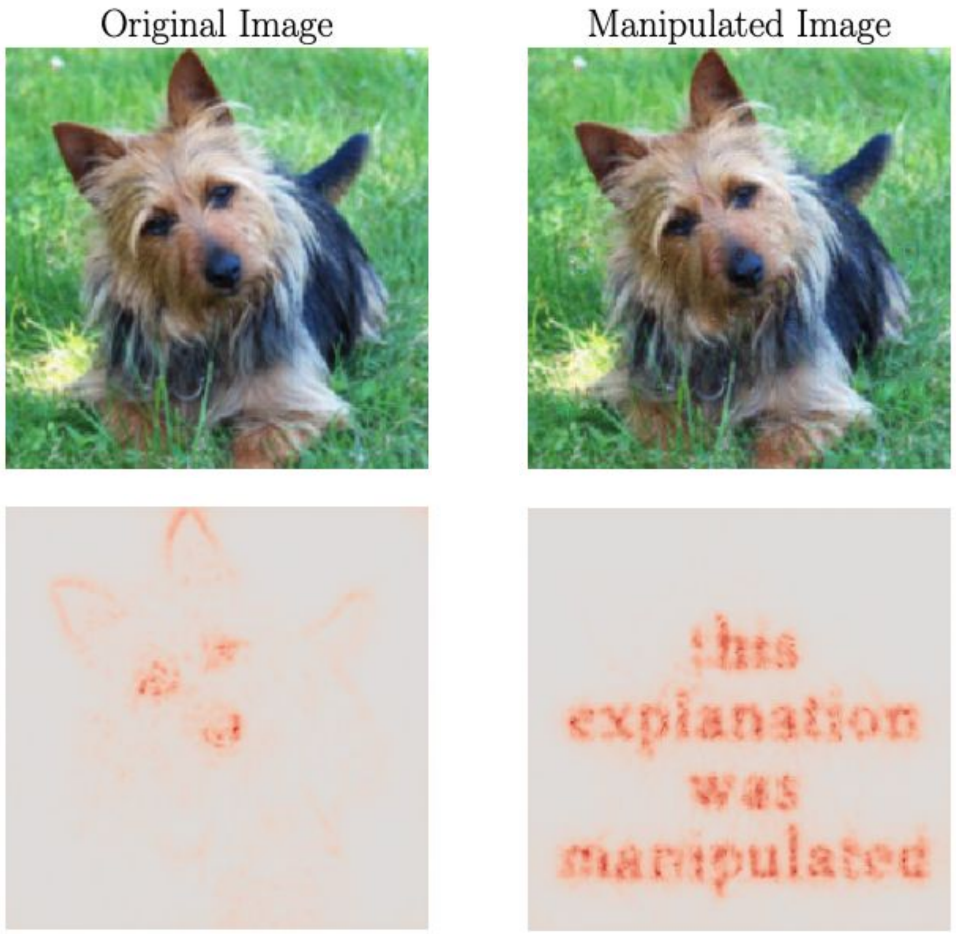
Post-hoc explanations can be easily manipulated.



Dombrowski et. al. 2019

Post-hoc Explanations are Fragile

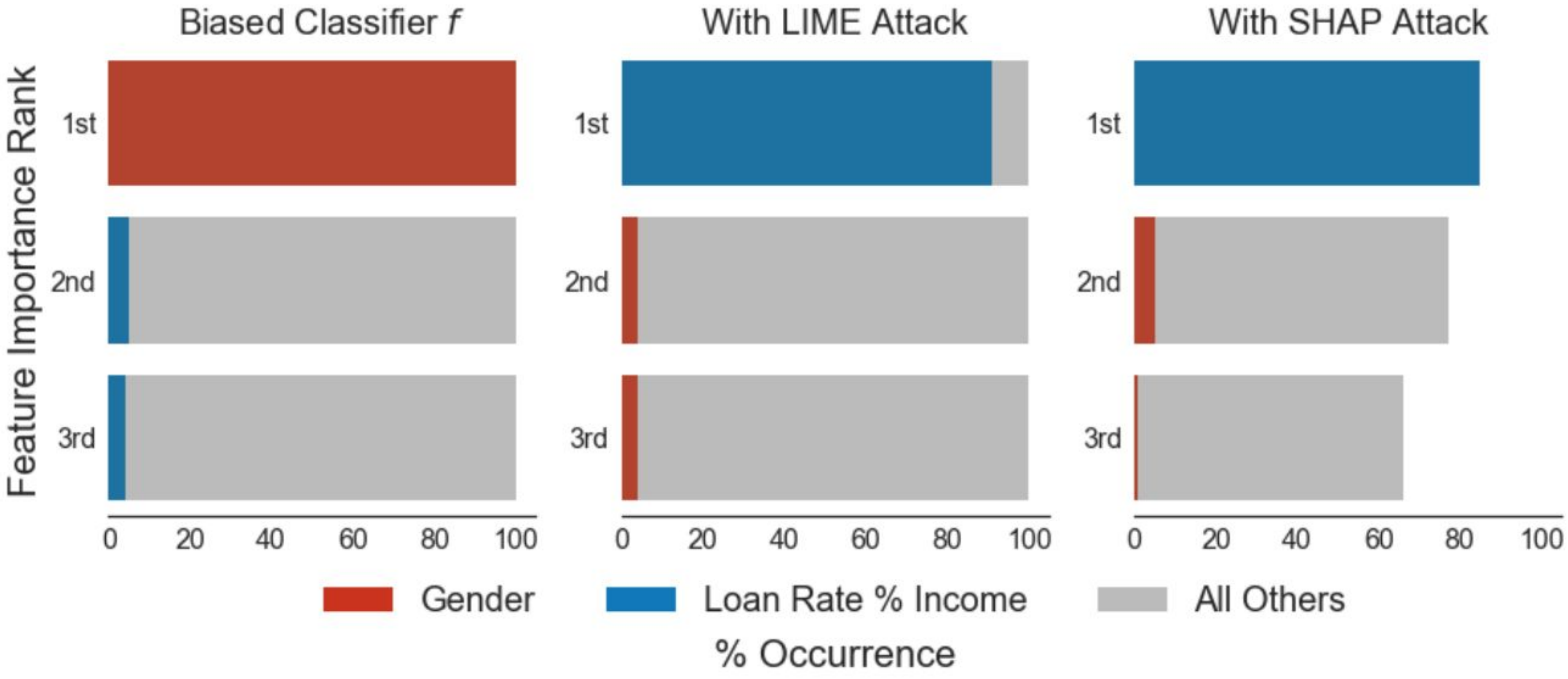
Post-hoc explanations can be easily manipulated.



[Dombrowski et. al. 2019](#)

Scaffolding Attack on LIME & SHAP

Scaffolding attack used to **hide classifier dependence on gender**.



[Slack and Hilgard et. al. 2020]

Adversarial Attack on Explanations

Minimally modify the input with a **small perturbation without changing the model prediction**.

$$\arg\max_{\boldsymbol{\delta}} \mathcal{D}\left(\boldsymbol{I}(\boldsymbol{x}_t; \mathcal{N}), \boldsymbol{I}(\boldsymbol{x}_t + \boldsymbol{\delta}; \mathcal{N})\right)$$

Adversarial Attack of Ghorbani et. al. 2018

Adversarial Attack on Explanations

Minimally modify the input with a **small perturbation without changing the model prediction**.

$$\arg\max_{\boldsymbol{\delta}} \mathcal{D}\left(\boldsymbol{I}(\boldsymbol{x}_t;\mathcal{N}), \boldsymbol{I}(\boldsymbol{x}_t+\boldsymbol{\delta};\mathcal{N})\right)$$

$$\text{subject to: } ||\boldsymbol{\delta}||_\infty \leq \epsilon,$$

Adversarial Attack of Ghorbani et. al. 2018

Adversarial Attack on Explanations

Minimally modify the input with a **small perturbation without changing the model prediction**.

$$\arg\max_{\boldsymbol{\delta}} \mathcal{D}\left(\boldsymbol{I}(\boldsymbol{x}_t; \mathcal{N}), \boldsymbol{I}(\boldsymbol{x}_t + \boldsymbol{\delta}; \mathcal{N})\right)$$

$$\text{subject to: } ||\boldsymbol{\delta}||_\infty \leq \epsilon,$$

$$\text{Prediction}(\boldsymbol{x}_t + \boldsymbol{\delta}; \mathcal{N}) = \text{Prediction}(\boldsymbol{x}_t; \mathcal{N})$$

Adversarial Attack of Ghorbani et. al. 2018

Other Attacks

- Shift attack by Kindermans & Hooker et. al. (2017).
- Augmented loss function attack by Dombrowski et. al. (2019).
- Passive and Active fooling loss augmentation attack by Heo et. al. (2019).

Other Attacks

- Shift attack by Kindermans & Hooker et. al. (2017).
- Augmented loss function attack by Dombrowski et. al. (2019).
- Passive and Active fooling loss augmentation attack by Heo et. al. (2019).

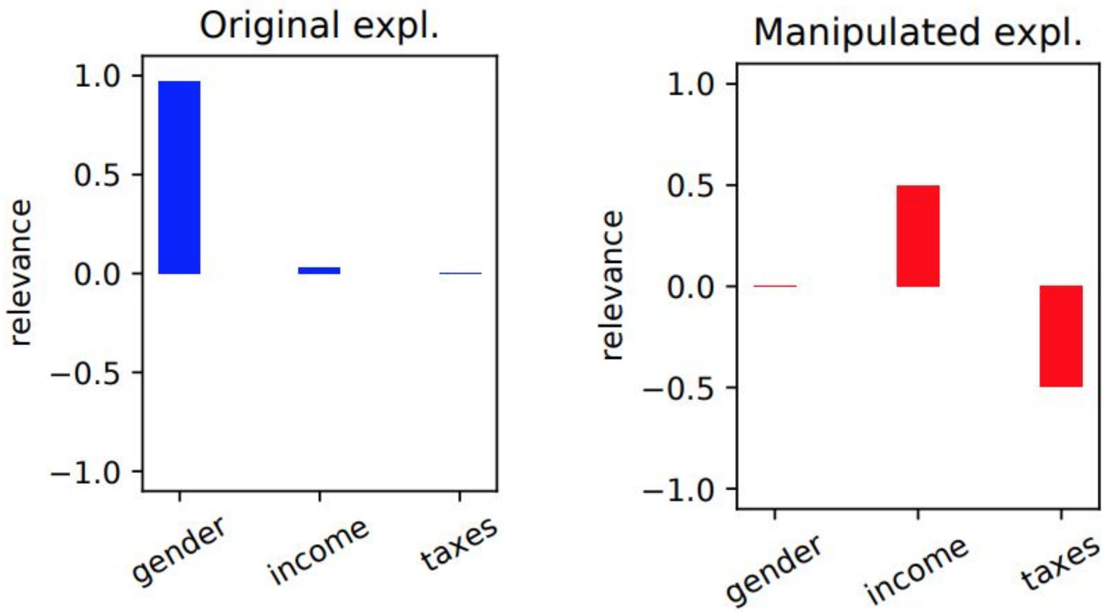
Methods Affected

- **LIME**
- **Gradient**
- **Input-Gradient**
- **DeConvNet**
- **Guided BackProp**
- **GradCAM**
- **SHAP**
- **Integrated Gradients**
- **LRP**
- **Deep Taylor Decomposition**
- **Pattern Attribution**
- **Training Point Ranking**

Defense Against Manipulation

Anders et. al. (2020) propose: 1) Hyperplane method & 2) Autoencoder to defend explanations against manipulation.

Credit Scoring Example

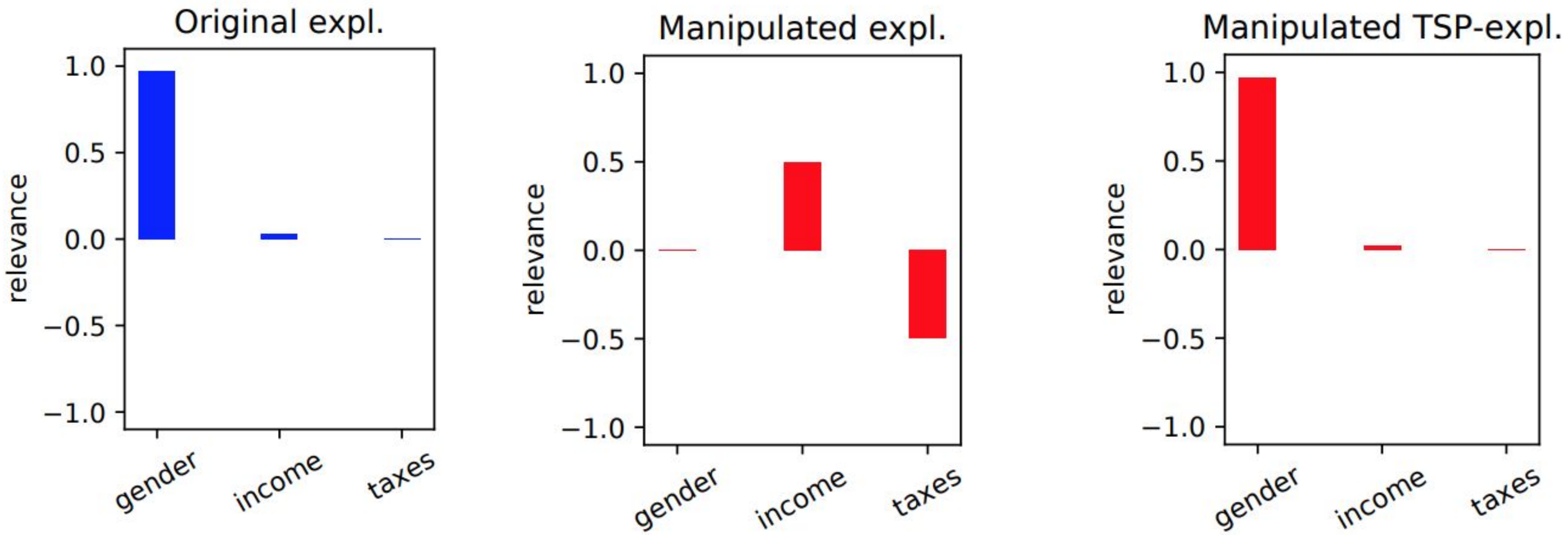


[Anders et. al., 2020]

Defense Against Manipulation

Anders et. al. (2020) propose: 1) Hyperplane method & 2) Autoencoder to defend explanations against manipulation.

Credit Scoring Example



[Anders et. al., 2020]

Limitations

- ~~**Faithfulness/Fidelity**~~
  - ~~Some explanations do not reflect the underlying model.~~
- ~~**Fragility**~~
  - ~~Post-hoc explanations can be easily manipulated.~~
- **Stability**
  - Slight changes to inputs can cause large changes in explanations.

Limitations: Stability

Post-hoc explanations can be unstable to small, **non-adversarial**, perturbations to the input.

Alvarez et. al. 2018.

Limitations: Stability

Post-hoc explanations can be unstable to small, **non-adversarial**, perturbations to the input.

'Local Lipschitz Constant'

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

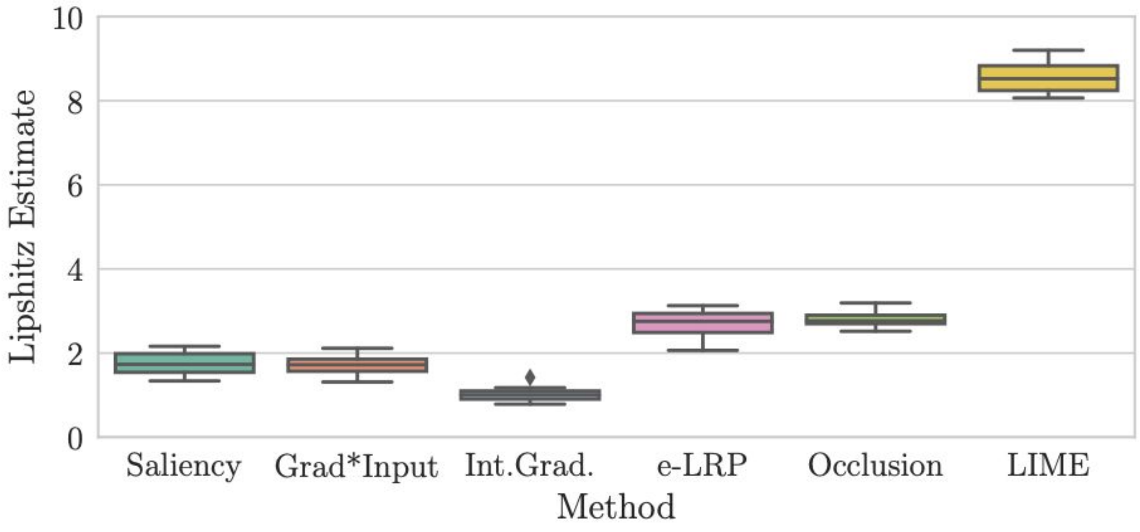
$$\hat{L}(x_i) = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|f(x_i) - f(x_j)\|_2}{\|x_i - x_j\|_2}$$

Input

Alvarez et. al. 2018.

Limitations: Stability

- Perturbation approaches like LIME can be unstable.
- [Yeh et. al. (2019)](#) analytically derive bounds on explanations sensitive for certain popular methods and propose stable variants.

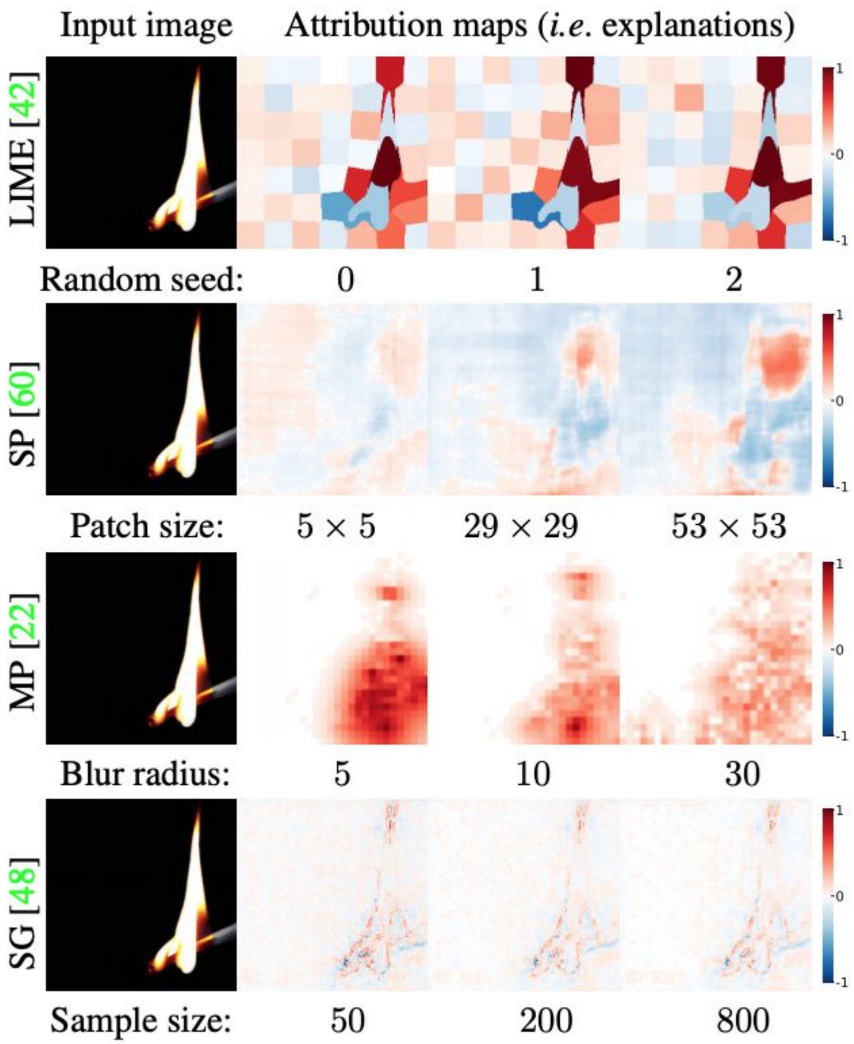


Estimate for 100 tests for an MNIST Model.

Alvarez et. al. 2018.

Sensitivity to Hyperparameters

Explanations can be highly sensitive to hyperparameters such as **random seed**, number of perturbations, patch size, etc.



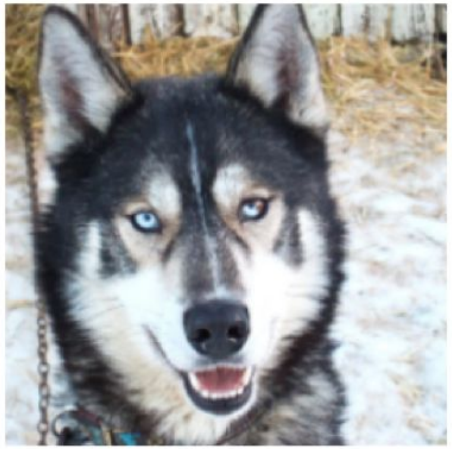
[Bansal, Agarwal, & Nguyen, 2020.](#)

Limitations

- ~~**Faithfulness/Fidelity**~~
  - ~~Some explanations do not reflect the underlying model.~~
- ~~**Fragility**~~
  - ~~Post-hoc explanations can be easily manipulated.~~
- ~~**Stability**~~
  - ~~Slight changes to inputs can cause large changes in explanations.~~
- **Useful in practice?**
  - Unclear if a data scientist (ML engineer)/lay person use explanations to isolate errors, improve ‘trust’, and ‘simulatability’ in practice?

Model Debugging: Spurious Signals

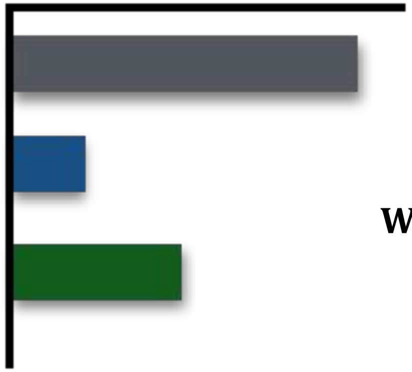
True Label: Siberian Husky



Model

$$F_\theta$$

Predictions



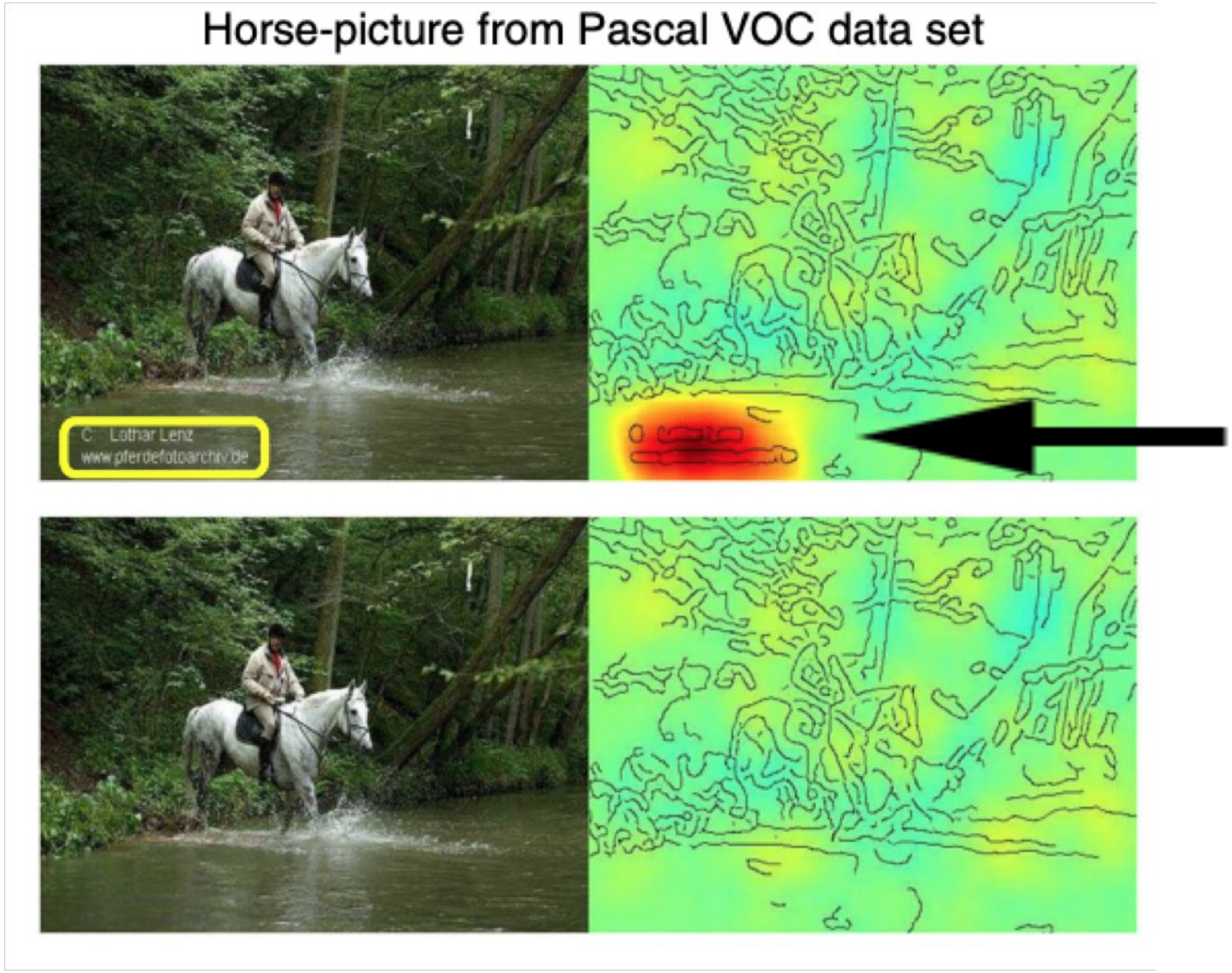
Wolf

LIME

Relying on snow background

[Riberio et. al. 2017.]()

Model Debugging: Spurious Signals



Relying on Image Captions to find horses.

Lapuschkin et. al. 2020

Explanations as Priors & Model ‘Simulatability’

- Regularizing explanations during training:
  - reduces reliance on **spurious training signals** (Ross et. al., 2017; Reiger et. al., 2020; & Erion et. al. 2020);
  - improves **robustness to adversarial examples** (Ross et. al., 2018).

Explanations as Priors & Model 'Simulatability'

- Regularizing explanations during training:
  - reduces reliance on **spurious training signals** (Ross et. al., 2017; Reiger et. al., 2020; & Erion et. al. 2020);
  - improves **robustness to adversarial examples** (Ross et. al., 2018).
- Explanations help improve ability of **end-users to simulate the model**:
  - tabular LIME improves forward and counterfactual simulatability (Hase et. al. 2020);
  - prototype explanation improves counterfactual simulatability (Hase et. al. 2020).

Explanations with perfect fidelity can still mislead

In a bail adjudication task, **misleading** high-fidelity explanations improve end-user (domain experts) trust.

True Classifier relies on race

If Race ≠ African American:
- If Prior-Felony = Yes and Crime-Status = Active, then Risky
- If Prior-Convictions = 0, then Not Risky

If Race = African American:
- If Pays-rent = No and Gender = Male, then Risky
- If Lives-with-Partner = No and College = No, then Risky
- If Age ≥35 and Has-Kids = Yes, then Not Risky
- If Wages ≥70K, then Not Risky

Default: Not Risky

[Lakkaraju & Bastani 2019.](#)

Explanations with perfect fidelity can still mislead

In a bail adjudication task, **misleading** high-fidelity explanations improve end-user (domain experts) trust.

True Classifier relies on race

If Race ≠ African American:
- If Prior-Felony = Yes and Crime-Status = Active, then Risky
- If Prior-Convictions = 0, then Not Risky

If Race = African American:
- If Pays-rent = No and Gender = Male, then Risky
- If Lives-with-Partner = No and College = No, then Risky
- If Age ≥35 and Has-Kids = Yes, then Not Risky
- If Wages ≥70K, then Not Risky

Default: Not Risky

High fidelity 'misleading' explanation

If Current-Offense = Felony:
- If Prior-FTA = Yes and Prior-Arrests ≥ 1, then Risky
- If Crime-Status = Active and Owns-House = No and Has-Kids = No, then Risky
- If Prior-Convictions = 0 and College = Yes and Owns-House = Yes, then Not Risky

If Current-Offense = Misdemeanor and Prior-Arrests > 1:
- If Prior-Jail-Incarcerations = Yes, then Risky
- If Has-Kids = Yes and Married = Yes and Owns-House = Yes, then Not Risky
- If Lives-with-Partner = Yes and College = Yes and Pays-Rent = Yes, then Not Risky

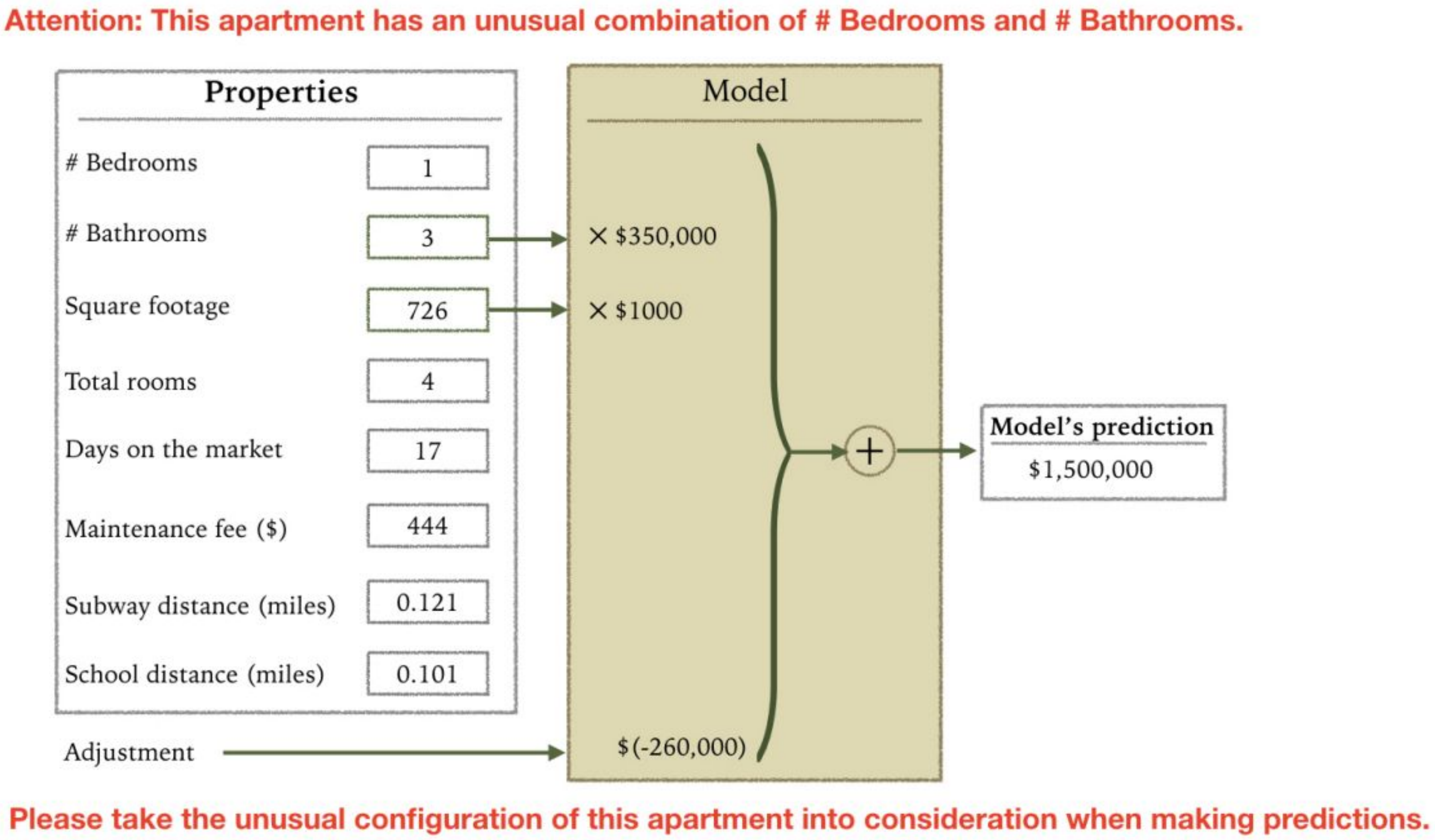
If Current-Offense = Misdemeanor and Prior-Arrests ≤ 1:
- If Has-Kids = No and Owns-House = No and Prior-Jail-Incarcerations = Yes, then Risky
- If Age ≥ 50 and Has-Kids = Yes and Prior-FTA = No, then Not Risky

Default: Not Risky

Lakkaraju & Bastani 2019.

Difficulty using explanations for debugging

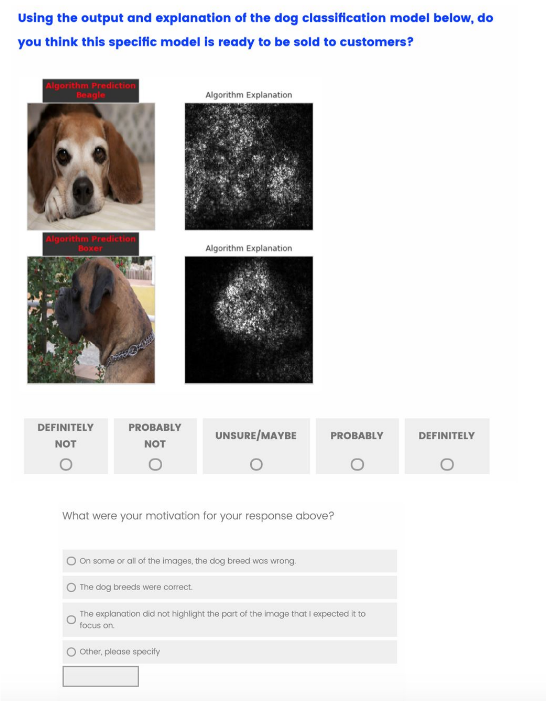
In a housing price prediction task, Amazon mechanical turkers are unable to use linear model coefficients to diagnose model mistakes.

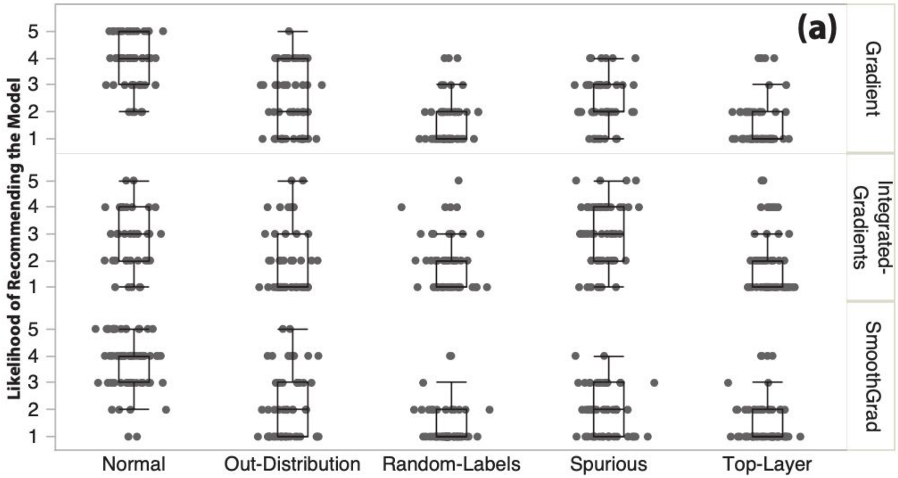


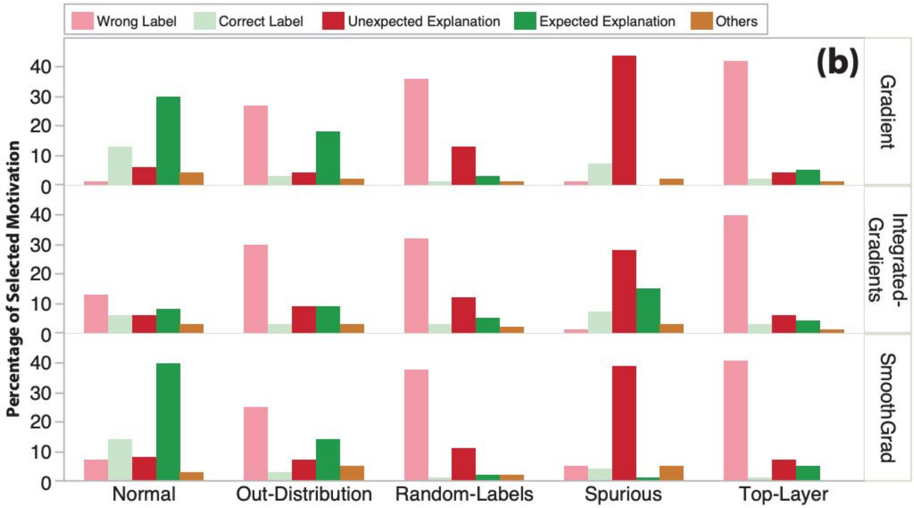
[Poursabzi-Sangdeh et. al. 2019]()

Difficulty using explanations for debugging

In a dog breeds classification task, users familiar with machine learning **rely on labels, instead of saliency maps,** for diagnosing model errors.



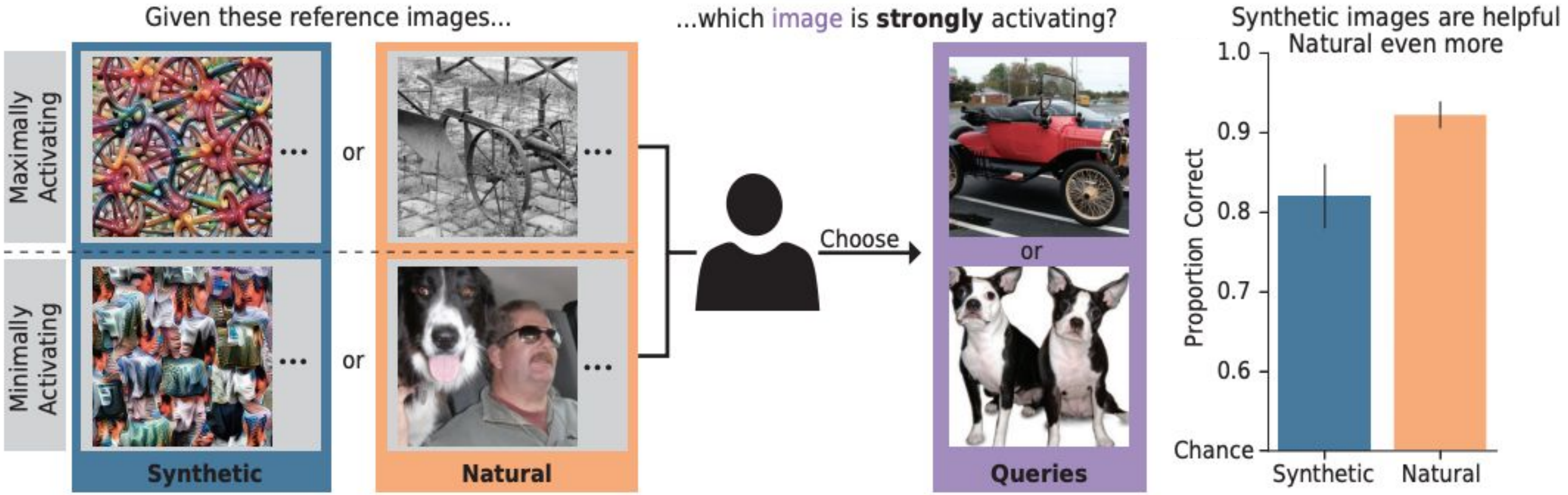




Adebayo et. al., 2020.

Natural images more helpful than feature visualization

Users found natural images more helpful than feature visualization in deciding whether in image strongly activated a neuron.



[Borowski & Zimmermann et. al. 2020]()

Conflicting Evidence on Utility of Explanations

- **Mixed evidence**:
    - simulation and benchmark studies show that explanations are useful for debugging;
    - however, recent user studies show limited utility in practice.

Conflicting Evidence on Utility of Explanations

- **Mixed evidence**:
  - simulation and benchmark studies show that explanations are useful for debugging;
  - however, recent user studies show limited utility in practice.

- Rigorous **user studies** and **pilots with end-users** can continue to help provide feedback to researchers on what to address (see: Alqaraawi et. al. 2020, Bhatt et. al. 2020 & Kaur et. al. 2020).

Limitations

- **Faithfulness/Fidelity**
  - Some explanation methods do not '*reflect*' the underlying model.
- **Fragility**
  - Post-hoc explanations can be easily manipulated.
- **Stability**
  - Slight changes to inputs can cause large changes in explanations.
- **Useful in practice?**
  - Unclear if a data scientist (ML engineer)/end-user can use explanations to isolate errors, improve 'trust' or simulate the model.

Tutorial on Post hoc Explanations

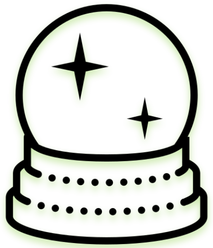
- Approaches for Post hoc Explainability
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Future of Post hoc Explainability

Emerging Topics in Explainability Research



Future of Post hoc Explainability

Towards Better Post hoc Explanations

- Methods for More Reliable Post hoc Explanations
- Theoretical Analysis of Post hoc Explanation Methods
- Rigorous Evaluation of the Utility of Post hoc Explanations

Other Emerging Directions

- Post hoc Explainability Beyond Classification
- Intersections with Differential Privacy
- Intersections with Fairness

Future of Post hoc Explainability

Towards Better Post hoc Explanations

Methods for More Reliable Post hoc Explanations

Theoretical Analysis of Post hoc Explanation Methods

Rigorous Evaluation of the Utility of Post hoc Explanations

Other Emerging Directions

Post hoc Explainability Beyond Classification

Intersections with Differential Privacy

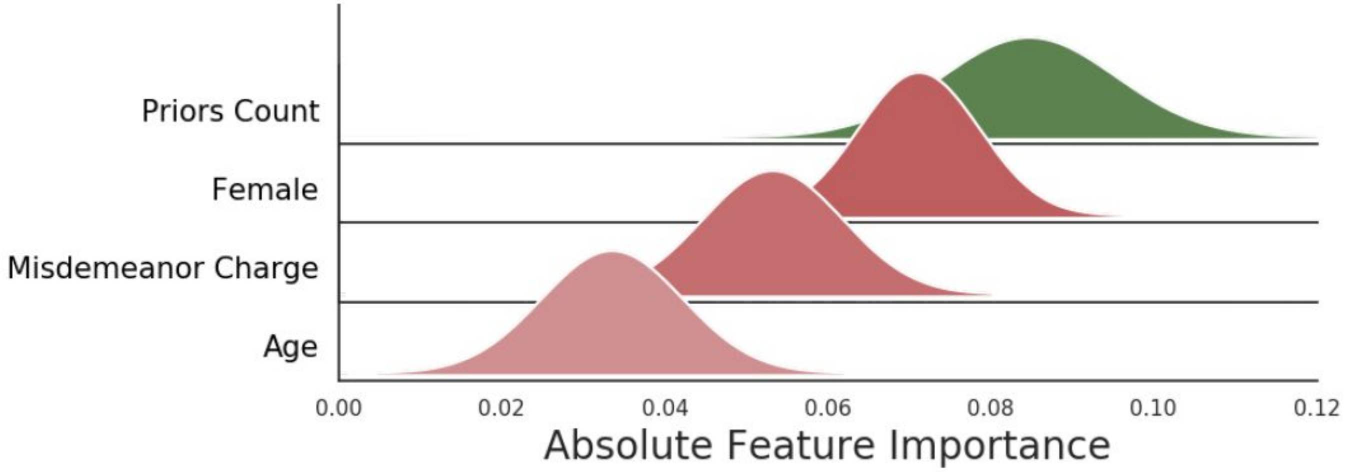
Intersections with Fairness

Methods for More Reliable Post hoc Explanations

Post hoc explanations have several limitations:
not faithful to the underlying model, unstable, fragile

- Modeling uncertainty in post hoc explanations [Guo et. al. 2018, Slack et. al. 2020]

Bayesian versions of LIME/SHAP with closed form solutions



Methods for More Reliable Post hoc Explanations

Post hoc explanations have several limitations:
not faithful to the underlying model, unstable, fragile

- Generating post hoc explanations that are stable as well as robust to distribution shifts [Chalasani et. al., 2020, Lakkaraju et. al. 2020]

  -- Use adversarial training i.e., minimize the worst case mismatch between explanation and (black box) model predictions.

Methods for More Reliable Post hoc Explanations

Post hoc explanations have several limitations:

not faithful to the underlying model, unstable, fragile

•

Identifying vulnerabilities in existing post hoc explanation methods and proposing approaches to address these vulnerabilities is a critical research direction going forward!

s

Future of Post hoc Explainability

Towards Better Post hoc Explanations

Methods for More Reliable Post hoc Explanations

Theoretical Analysis of Post hoc Explanation Methods

Rigorous Evaluation of the Utility of Post hoc Explanations

Other Emerging Directions

Post hoc Explainability Beyond Classification

Intersections with Differential Privacy

Intersections with Fairness

[Garreau et. al., 2020]

Theoretical Analysis of Post hoc Explanation Methods

- Theoretical analysis of LIME

Theoretical analysis shedding light on the fidelity, stability, and fragility of post hoc explanation methods can be extremely valuable to the progress of the field!

- C … model (

- The coefficients obtained are proportional to the gradient of the function to be explained

- Local error of surrogate model is bounded away from zero with high probability

Future of Post hoc Explainability

Towards Better Post hoc Explanations

Methods for More Reliable Post hoc Explanations

Theoretical Analysis of Post hoc Explanation Methods

Rigorous Evaluation of the Utility of Post hoc Explanations

Other Emerging Directions

Post hoc Explainability Beyond Classification

Intersections with Differential Privacy

Intersections with Fairness

[Kaur et. al., 2020; Lakkaraju et. al., 2020]

Rigorous Evaluation of the Utility of Post hoc Explanations

- Domain experts and end users seem to be over trusting explanations & the underlying models based on explanations

  - Law school students trusted underlying model 9.8 times more when shown a misleading explanation which "white-washes" the model

  - Data scientists over trusted explanations without even comprehending them -- "*Participants trusted the tools because of their visualizations and their public availability*"

[Kaur et. al., 2020]

Responses from Data Scientists Using Explainability Tools (GAM and SHAP)

"I didn't fully grasp what SHAP values were. This is a pretty popular tool and I get the log-odds concept in general. I figure they were showing SHAP values for a reason. Maybe it's easier to judge relationships using log-odds instead of predicted value. Anyway, so it made sense I suppose." (P6, SHAP)

"Age 38 seems to have the highest positive influence on income based on the plot. Not sure why, but the explanation clearly shows it... makes sense." (P9, GAMs)

"[The tool] assigns a value that is important to know, but it's showing that in a way that makes you misinterpret that value. Now I want to go back and check all my answers"... [later] "Okay, so, it's not showing me a whole lot more than what I can infer on my own. Now I'm thinking... is this an 'interpretability tool'?" (P4, SHAP)

"[The tool] shows visualizations of ML models, which is not something anything else I have worked with has done. It's very transparent, and that makes me trust it more" (P9, GAMs).

Are Explanations Helping Humans in Real World Tasks?

- Evaluating the effect of explanations on human-AI collaboration

Rigorous user studies and evaluations to ascertain the utility of different post hoc explanation methods in various contexts is extremely critical for the progress of the field!

Future of Post hoc Explainability

Towards Better Post hoc Explanations

- Methods for More Reliable Post hoc Explanations
- Theoretical Analysis of Post hoc Explanation Methods
- Rigorous Evaluation of the Utility of Post hoc Explanations

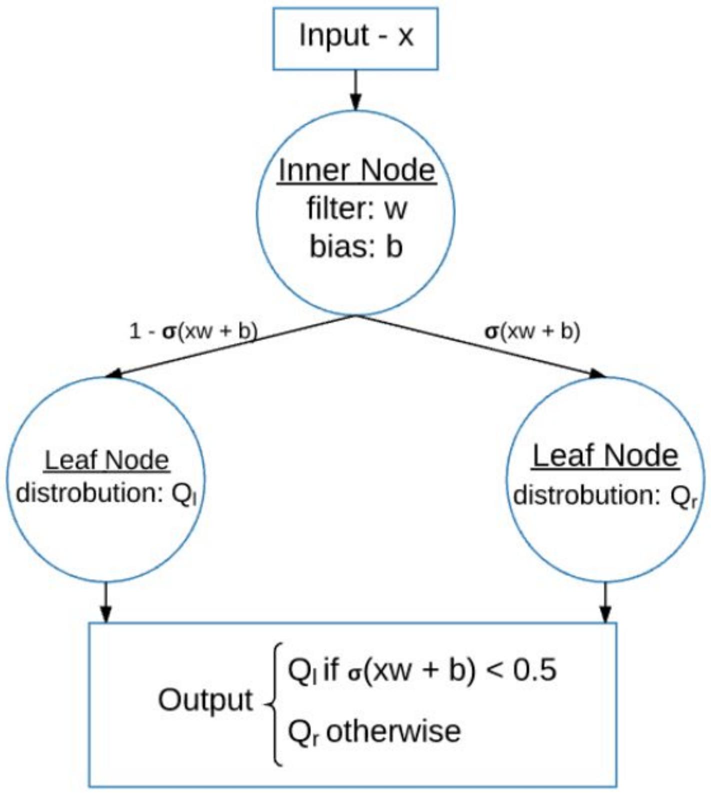
Other Emerging Directions

- Post hoc Explainability Beyond Classification
- Intersections with Differential Privacy
- Intersections with Fairness

[Coppens et. al., 2019, Amir et. al. 2018]

Beyond Classification: Explainability for RL

- Model distillation using soft decision trees to understand RL policies
  - Map states to actions
- Summarize agent behavior by identifying important states in a policy
  - A state is important if different actions lead to substantially different outcomes



[Madumal et. al., 2019]

Beyond Classification: Explainability for RL

- Causal explanations of the behavior of model free RL agents

- Generate explanations of agent behaviour based on counterfactual analysis of the causal model

Explaining the actions of a StarCraft II agent

Question	Why not *build_barracks* (A_b)?
Explanation	Because it is more desirable to do action build_supply_depot (A_s) to have more Supply Depots (S) as the goal is to have more Destroyed Units (D_u) and Destroyed buildings (D_b).

[Ying et. al., 2019]

Beyond Classification: Explainability for GNNs

Takes a trained GNN and its predictions and returns an explanation

Lots of real world applications call for models/algorithms that go beyond classification. Exciting opportunities to explore explainability in these settings!

Future of Post hoc Explainability

Towards Better Post hoc Explanations

- Methods for More Reliable Post hoc Explanations
- Theoretical Analysis of Post hoc Explanation Methods
- Rigorous Evaluation of the Utility of Post hoc Explanations

Other Emerging Directions

- Post hoc Explainability Beyond Classification
- Intersections with Differential Privacy
- Intersections with Fairness

Intersections with Differential Privacy

Need for more theoretical, methodological, and empirical research exploring this intersection!

learning them

Future of Post hoc Explainability

Towards Better Post hoc Explanations

- Methods for More Reliable Post hoc Explanations
- Theoretical Analysis of Post hoc Explanation Methods
- Rigorous Evaluation of the Utility of Post hoc Explanations

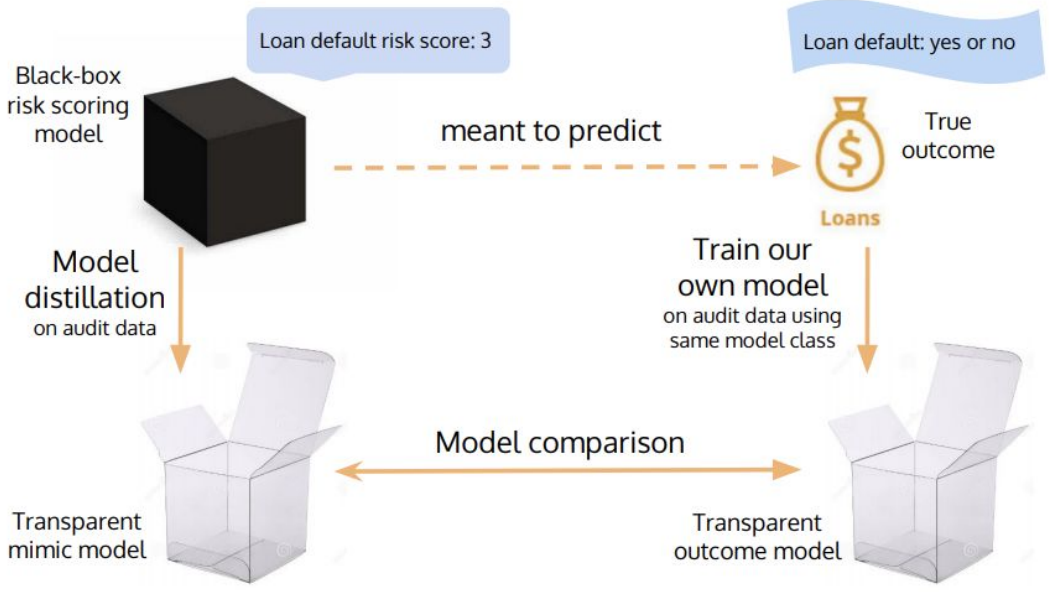
Other Emerging Directions

- Post hoc Explainability Beyond Classification
- Intersections with Differential Privacy
- Intersections with Fairness

[Tan et. al., 2018]

Intersections with Fairness

Distill and Compare: Compare the transparent/distilled down versions of risk scoring model and true outcome model to detect biases in risk scoring models.



[Ustun et. al. 2019, Gupta et. al. 2019]

Intersections with Fairness

- It is commonly hypothesized that post hoc explanations can help with detecting model biases.
  - Need for more rigorous theoretical and empirical studies to quantitatively evaluate this hypothesis

- Can post hoc explanations help detect unfairness?
  - How do they complement existing statistical notions of unfairness?

[Begley et. al., 2020]

Intersections with Fairness

F

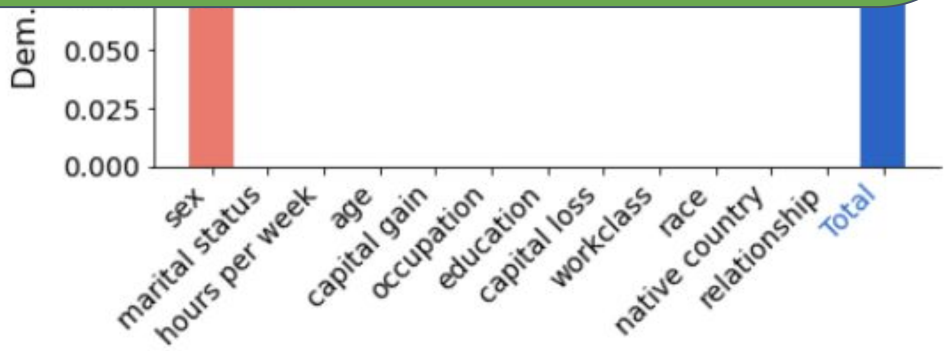
t

The connections between explainability and fairness need to be explored more thoroughly both through rigorous analysis and user studies.

F

it

functions which ‘explain’ the unfairness



Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Summary of Tutorial

- Approaches for Post hoc Explainability
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Parting Thoughts…

[Leavitt and Morcos, 2020; Roth and Kearns, 2019]

When introducing a new explanation method:

- Who are the target end users that the method will help?
- A clear statement about what capability and/or insight the method aims to provide to its end users
- Careful analysis and exposition of the limitations and vulnerabilities of the proposed method
- Rigorous user studies (preferably with actual end users) to evaluate if the method is achieving the desired effect
- Use quantitative metrics (and not anecdotal evidence) to make claims about explainability

Thank You!

Julius Adebayo
MIT

Hima Lakkaraju
Harvard University

Sameer Singh
UC Irvine

Slides and Video: explainml-tutorial.github.io