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

Hima Lakkaraju
Julius Adebayo
Sameer Singh

Hima Lakkaraju
Harvard University

Julius Adebayo
MIT

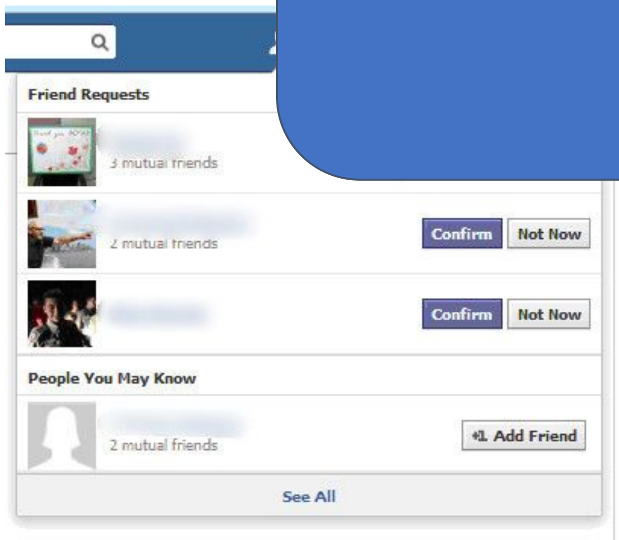
Sameer Singh
UC Irvine

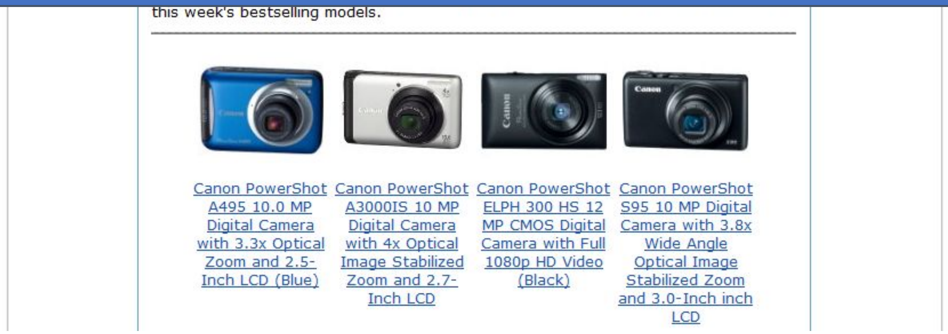
Slides and Video: explainml-tutorial.github.io

[Weller 2017]

Motivation

Machine Learning is EVERYWHERE!!





Motivation

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

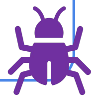
Motivation: Why Model Understanding?

Input

This model is
incorrect
make
on!! Let
model

Model understanding facilitates debugging.

Predictive
Model



Prediction = Siberian Husky

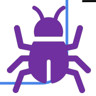
Motivation: Why Model Understanding?

Defendant Detai

This prediction is
ing
the
!

Model understanding facilitates bias detection.

Predictive Model



Prediction = Risky to Release

Motivation: Why Model Understanding?

Loan Applicant Details

FILE

I have some means
Let me
my
pay
ne.

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

to get a loan

Predictive Model

Prediction = Denied Loan

Loan Applicant

Motivation: Why Model Understanding?

Patient Data

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

Model Understanding

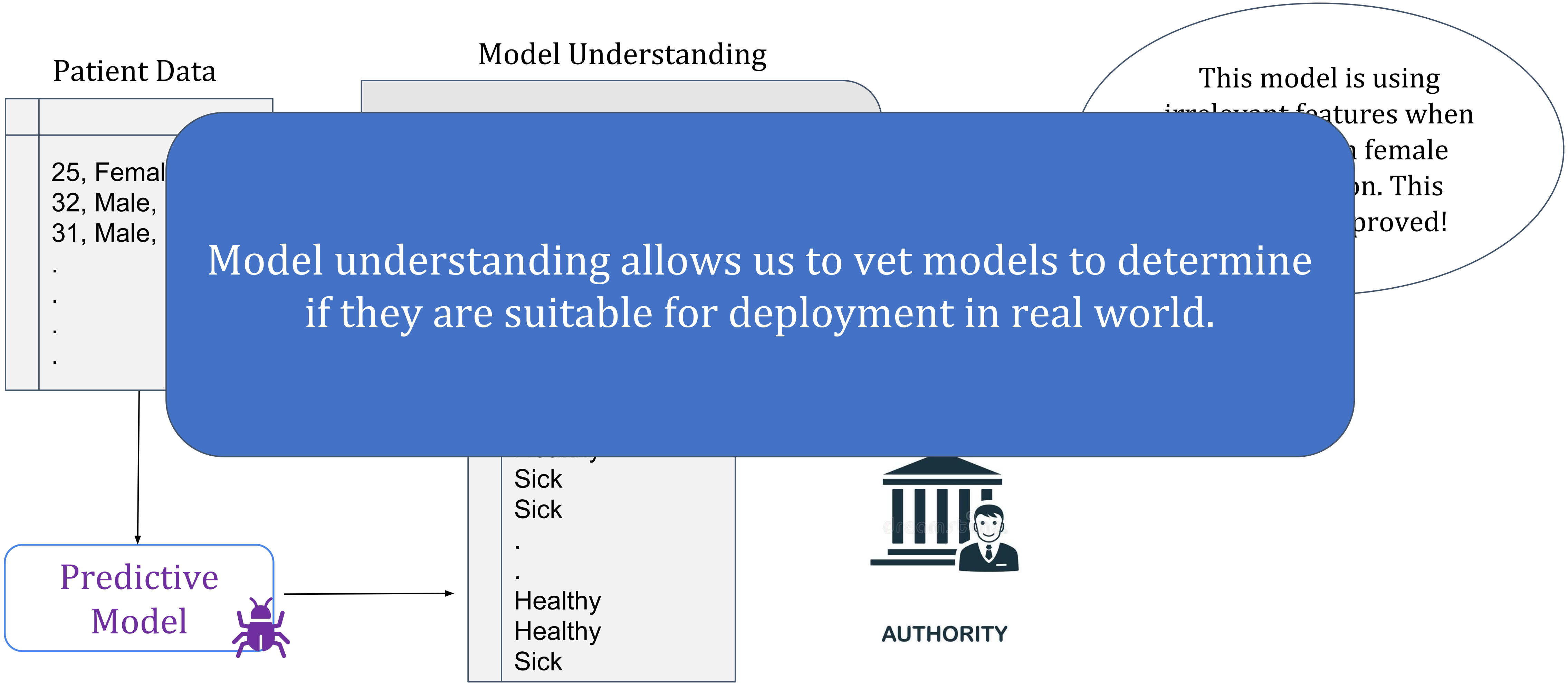
This model is using
irrelevant features when
on female
. I should
edictions
oup.

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

Sick
Sick
.
.
Healthy
Healthy
Sick

Predictive Model

Motivation: Why Model Understanding?



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

Motivation: Why Model Understanding?

Utility

Debugging

Bias Detection

Recourse

If and when to trust model predictions

Vet models to assess suitability for deployment

Stakeholders

End users (e.g., loan applicants)

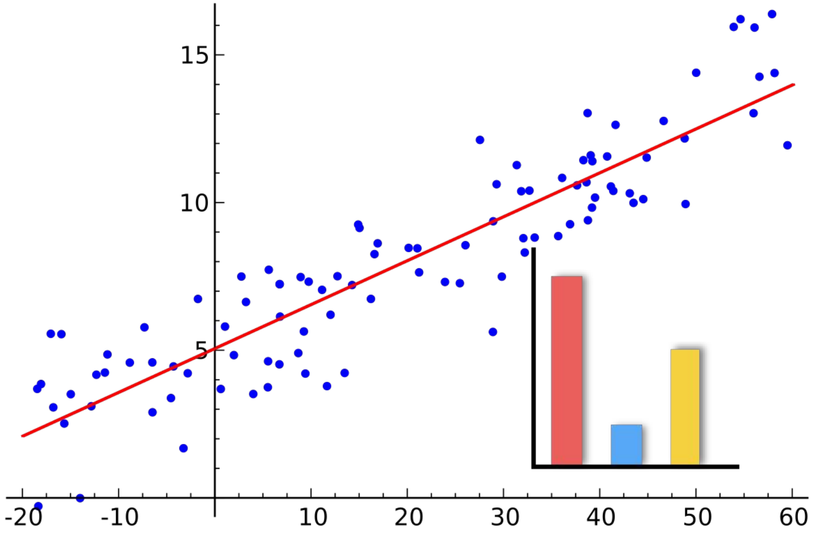
Decision makers (e.g., doctors, judges)

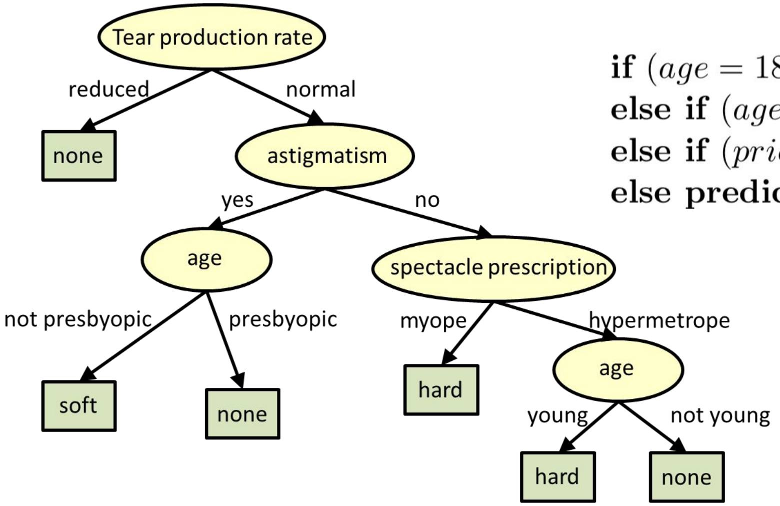
Regulatory agencies (e.g., FDA, European commission)

Researchers and engineers

Achieving Model Understanding

Take 1: Build *inherently interpretable* predictive models



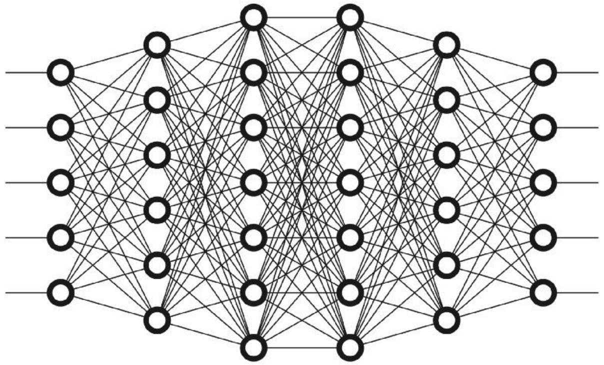


if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

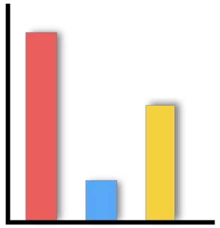
[Ribeiro et. al. 2016, Ribeiro et al. 2018; Lakkaraju et. al. 2019]

Achieving Model Understanding

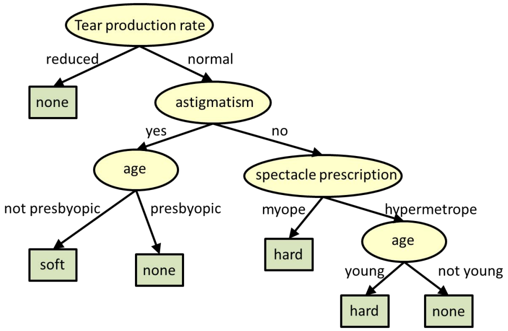
Take 2: *Explain* pre-built models *in a post-hoc manner*



Explainer



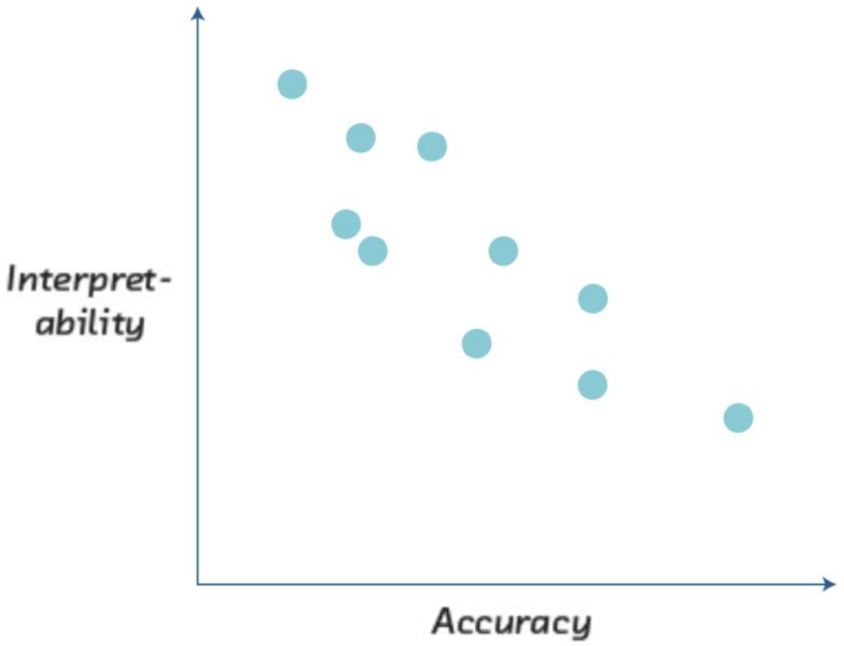
if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

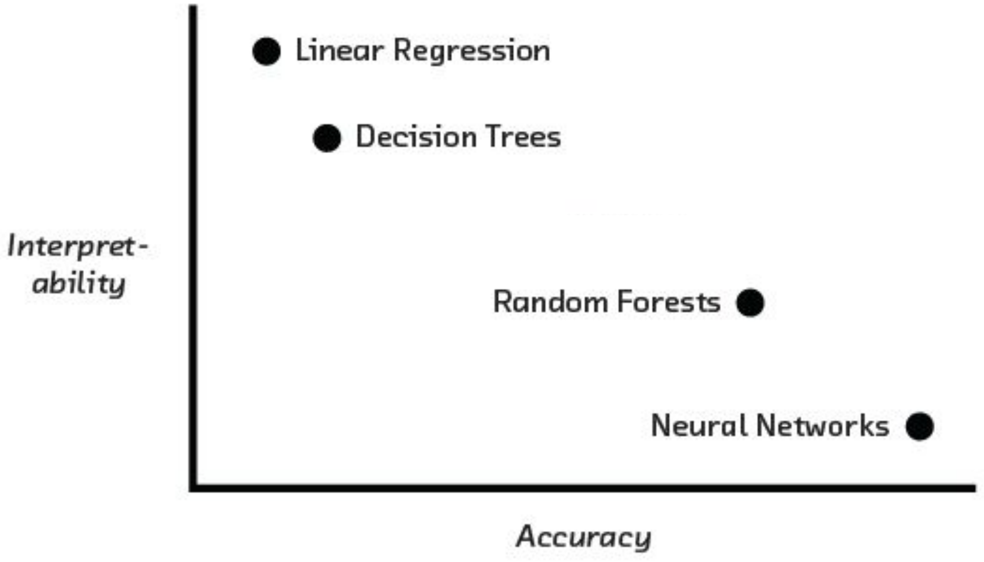


[Cireşan et. al. 2012, Caruana et. al. 2006, Frosst et. al. 2017, Stewart 2020]

Inherently Interpretable Models vs. Post hoc Explanations

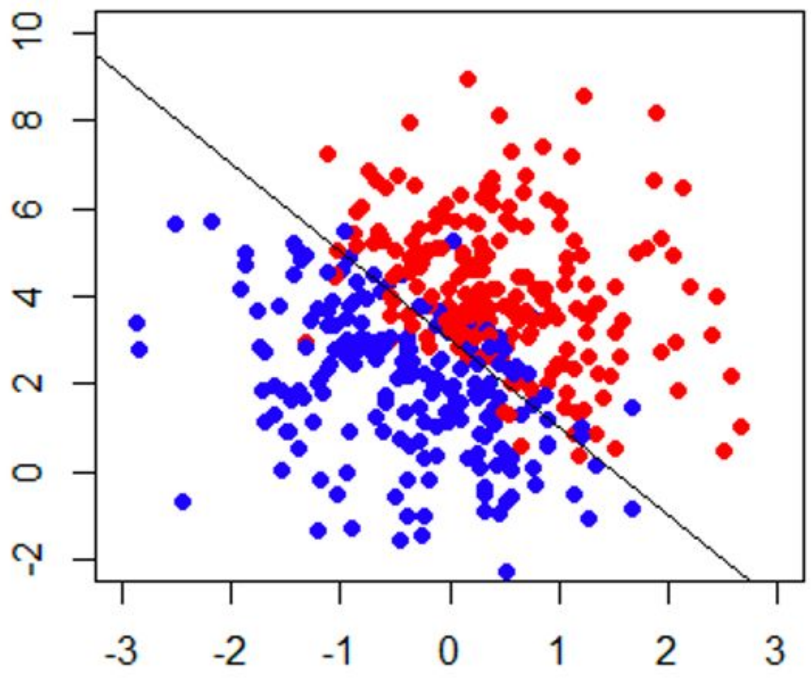
Example



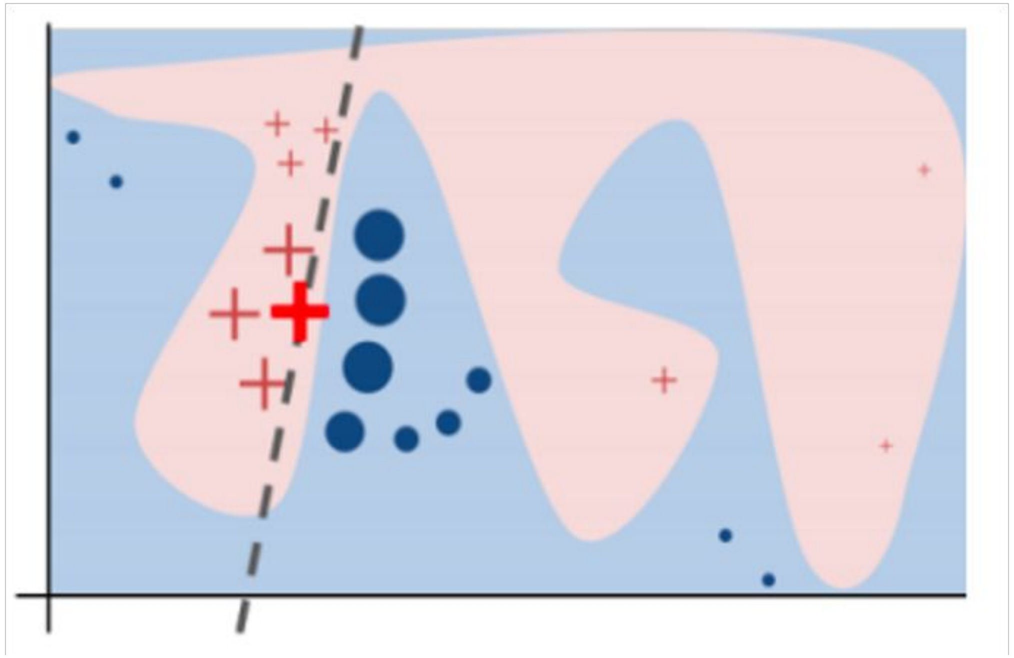


In ***certain*** settings, *accuracy-interpretability trade offs* may exist.

Inherently Interpretable Models vs. Post hoc Explanations



can build interpretable + accurate models



complex models might achieve higher accuracy

[Ribeiro et. al. 2016]

Inherently Interpretable Models vs. Post hoc Explanations

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

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

[Rudin 2019]

Inherently Interpretable Models vs. Post hoc Explanations

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

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

This 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

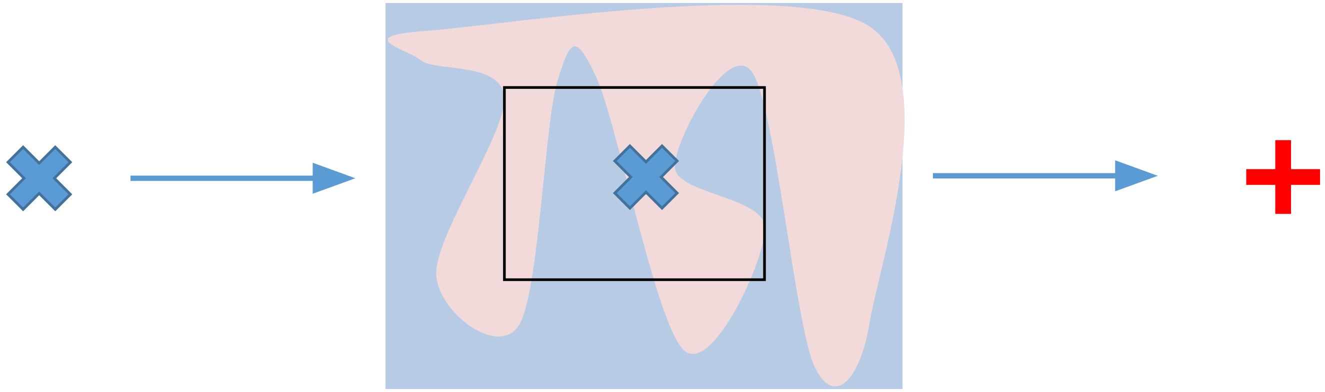
Describe how to *flip* the model prediction

...

User

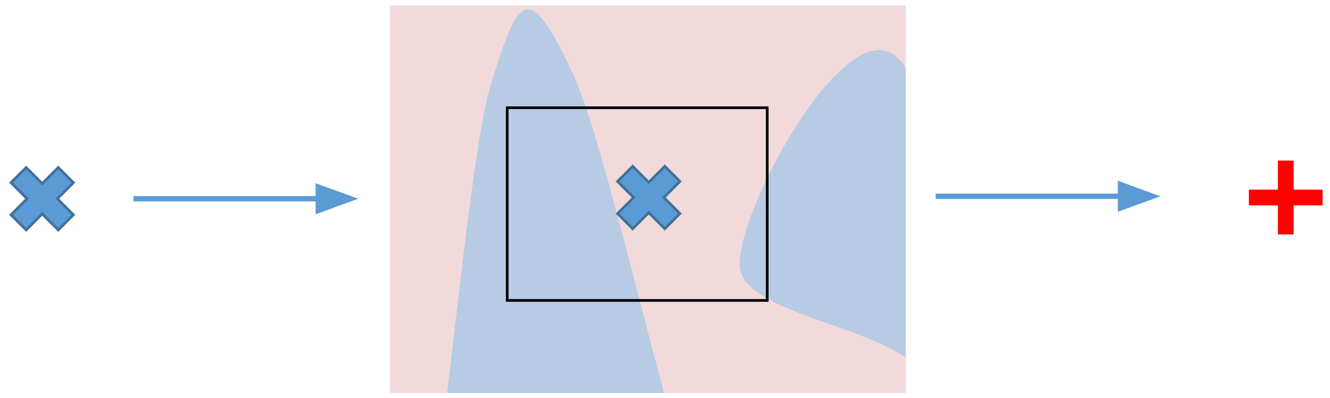
Local versus Global Explanations

Global explanation may be too complicated



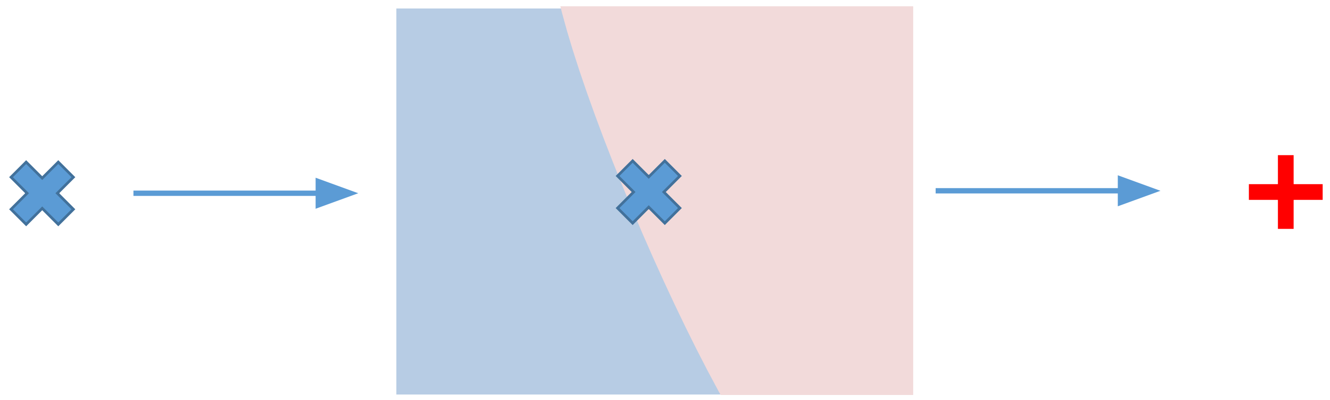
Local versus Global Explanations

Global explanation may be too complicated



Local versus Global Explanations

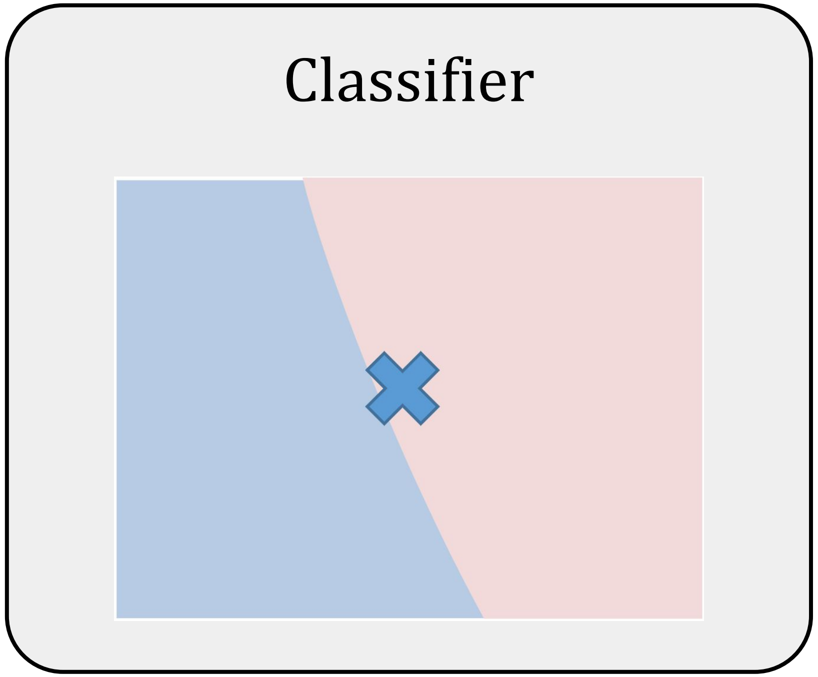
Global explanation may be too complicated



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

Local Explanations

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



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
- Explanations in Different Modalities
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Explanations in Different Modalities
- 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
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

Approaches for Post hoc Explainability

Local Explanations

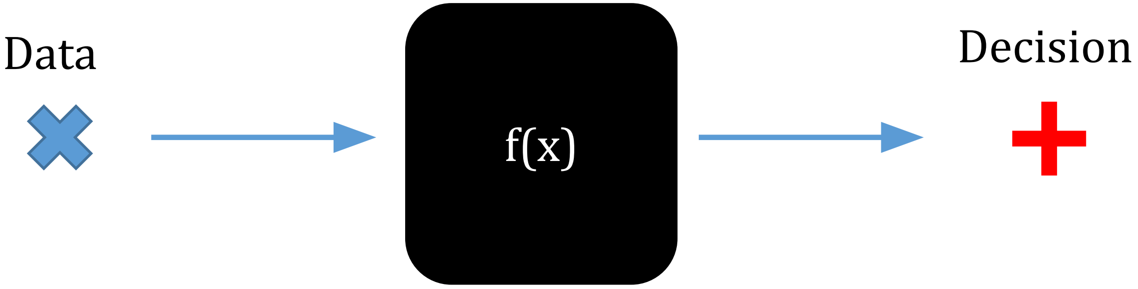
- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

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

Being Model-Agnostic…

No access to the internal structure…



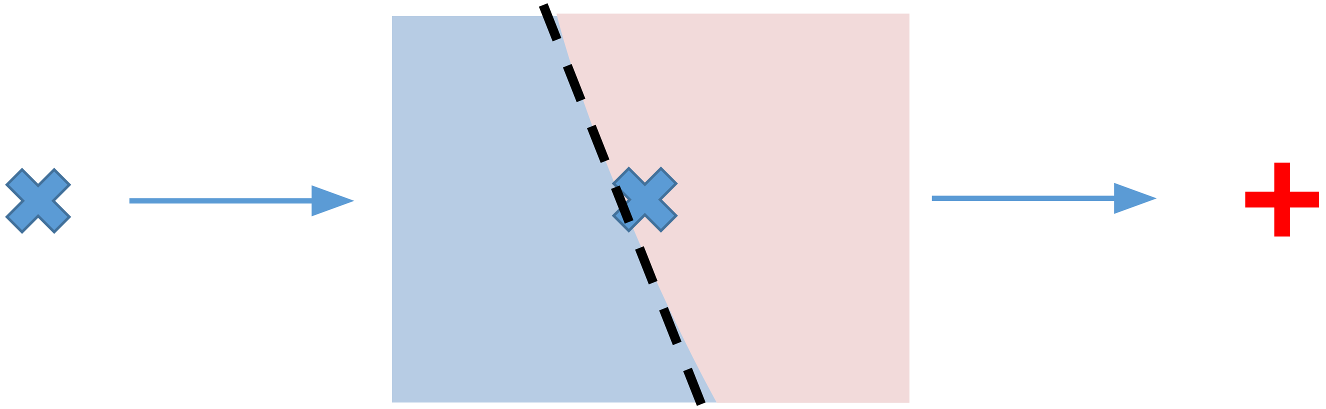
Not restricted to specific models

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

Study models that you don’t have access to!

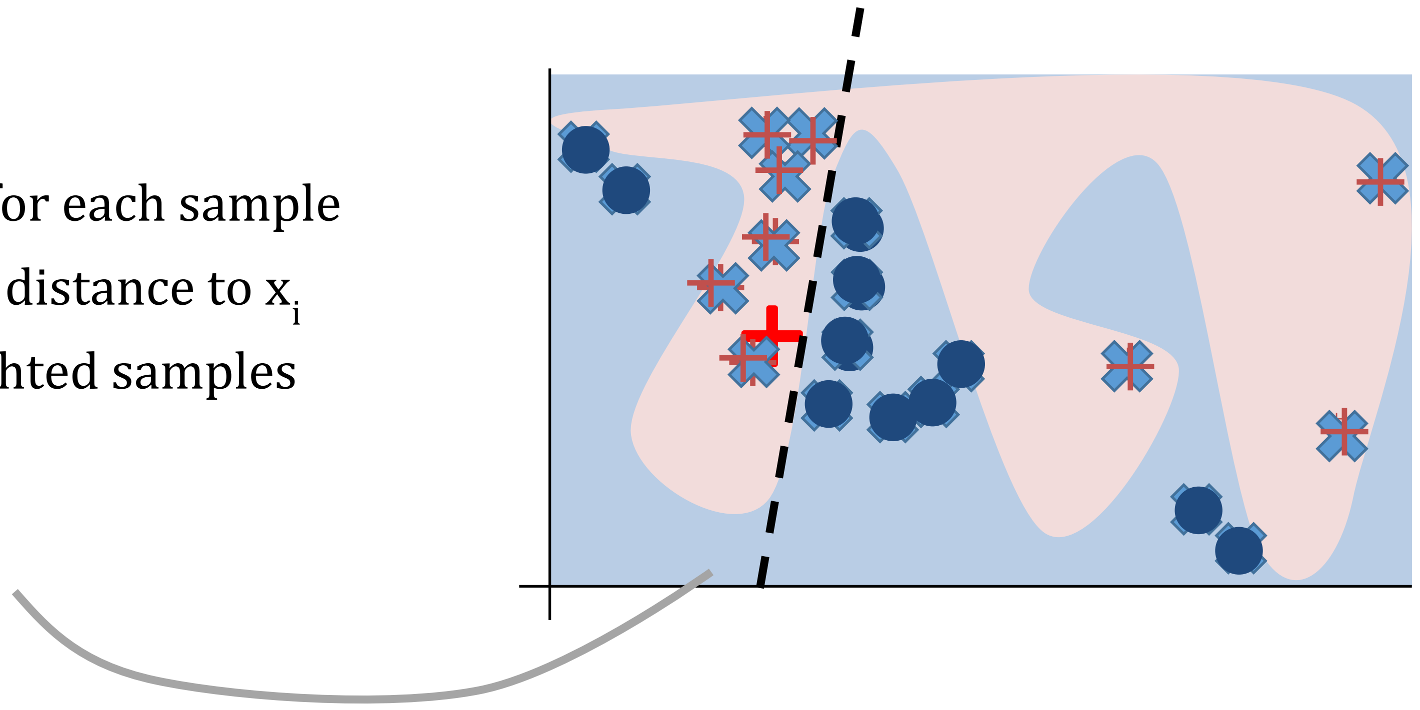
LIME: Sparse, Linear Explanations

Identify the important dimensions,
and present their relative importance



LIME: Sparse Linear Explanations

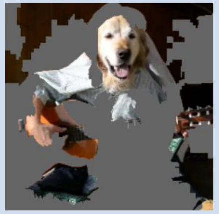
1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn simple model on weighted samples
5. Use simple model to explain



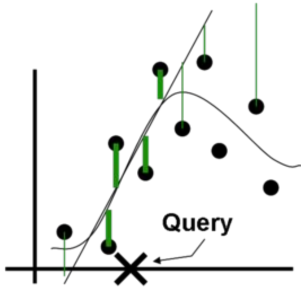
LIME Example - Images

Original Image

P(labrador) = 0.21

Perturbed Instances	P(Labrador)
	0.92
	0.001
	0.34

Maybe to a fault?



Locally weighted regression

Explanation

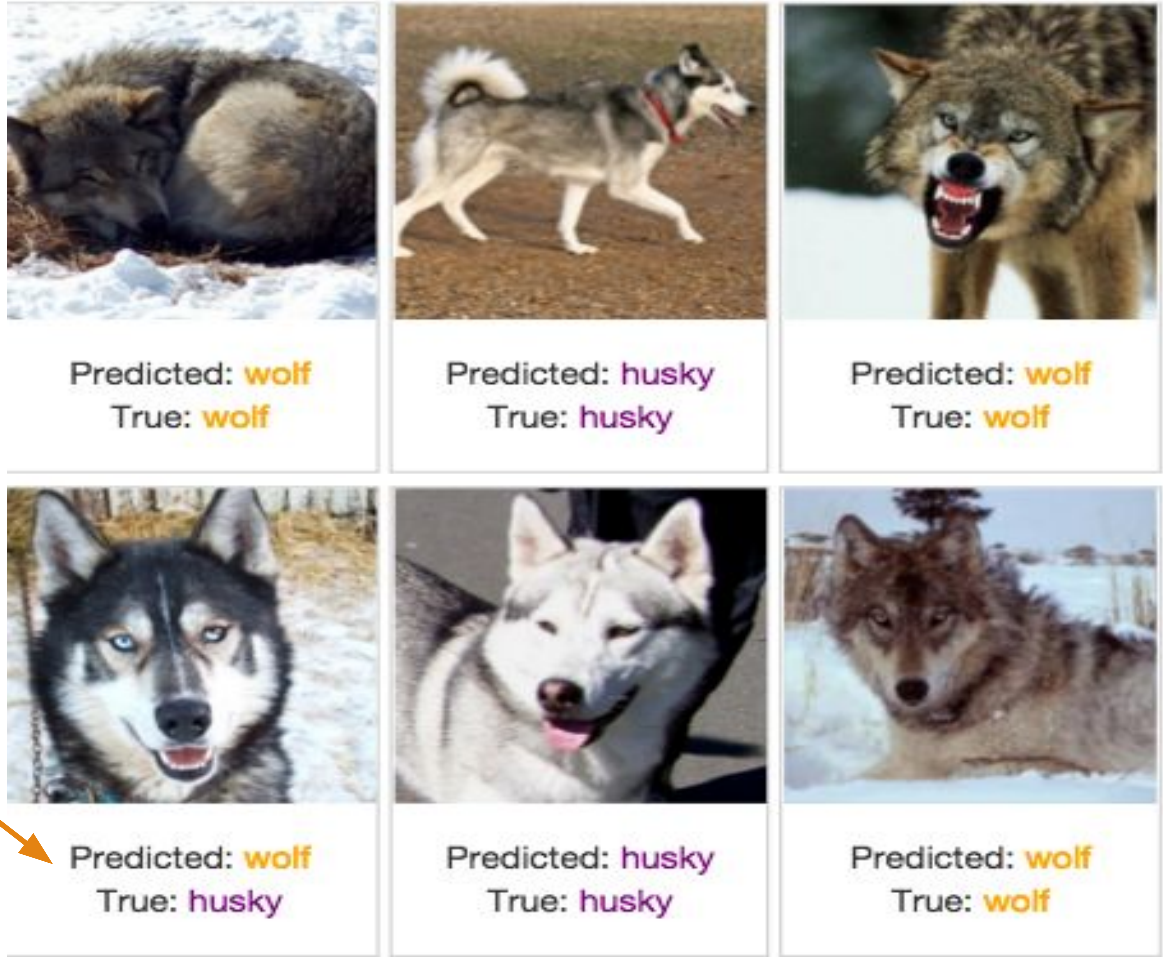
LIME is quite customizable:

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

[Ribeiro et al. 2016]

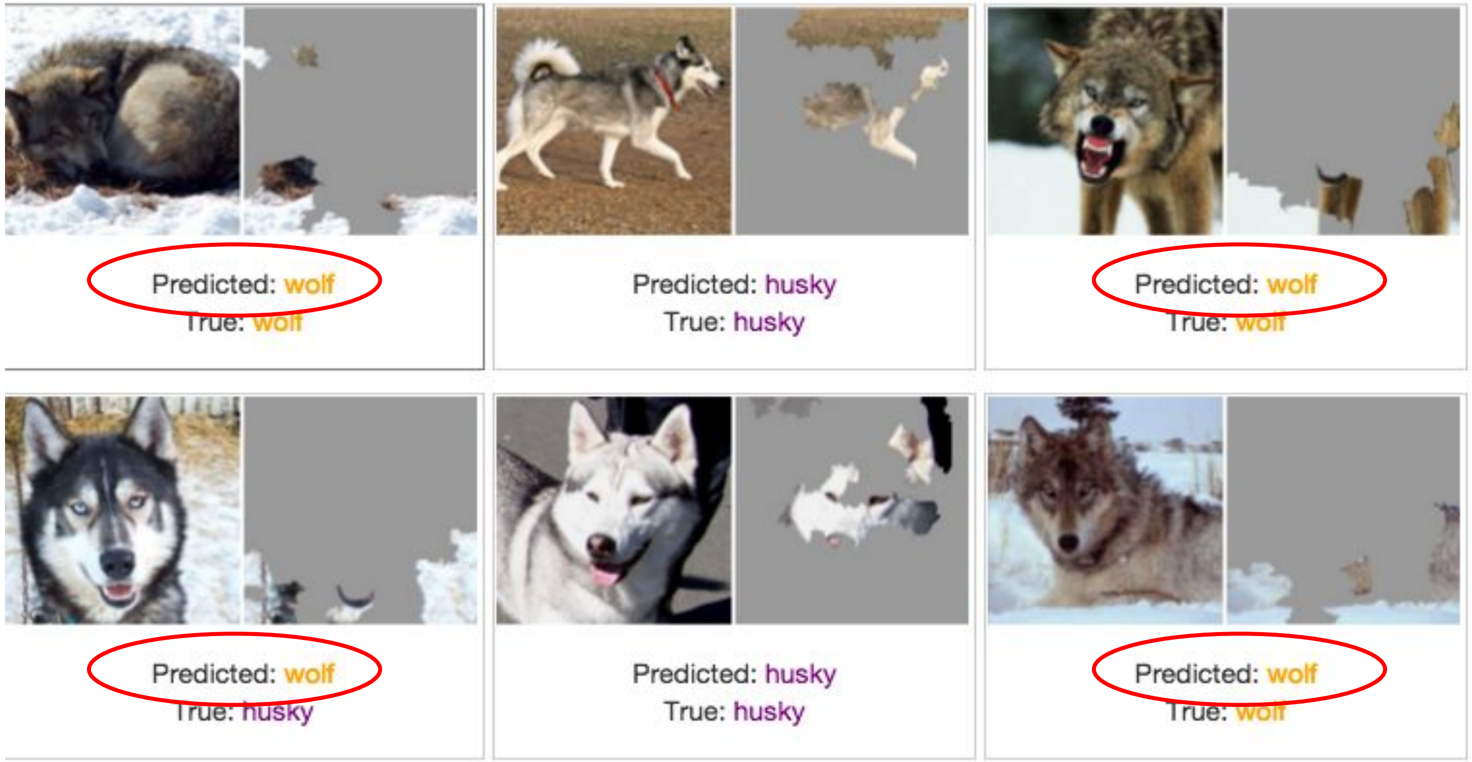
Predict Wolf vs Husky

Only 1 mistake!



[Ribeiro et al. 2016]

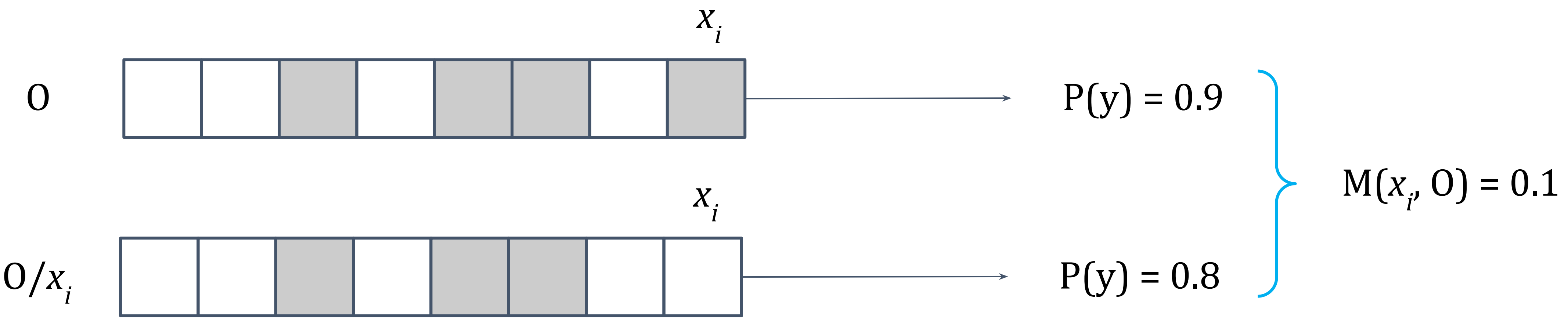
Predict Wolf vs Husky



We've built a great snow detector…

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

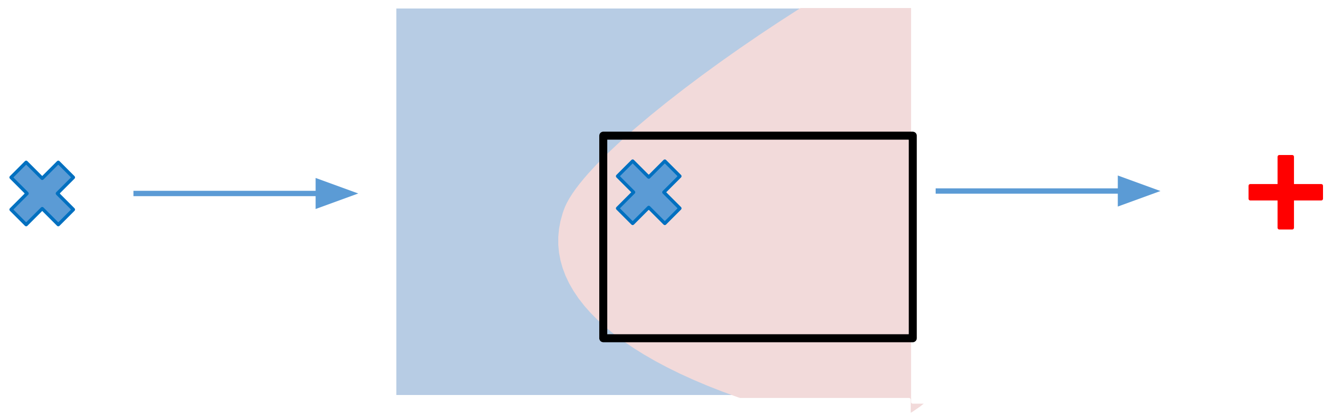
Local Explanations

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

Global Explanations

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

Anchors: Sufficient Conditions

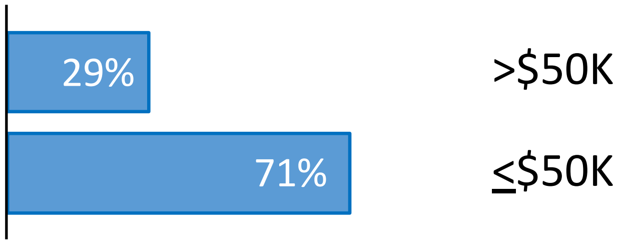


Identify the conditions under which the classifier has the same prediction

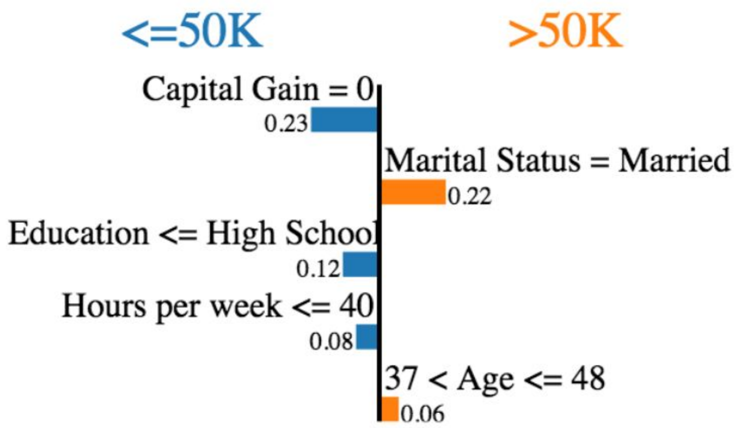
Salary Prediction

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

Salary



LIME



Anchors

IF Education ≤ High School
Then Predict Salary ≤ 50K

Approaches for Post hoc Explainability

Local Explanations

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

Global Explanations

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

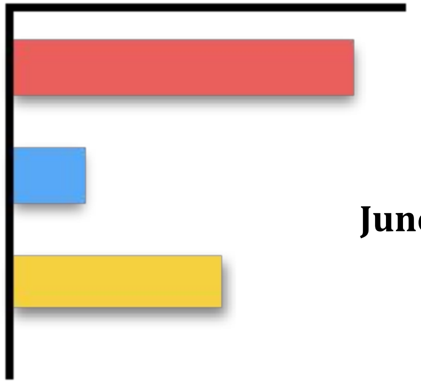
Saliency Map Overview

Input

Model

F_θ

Predictions



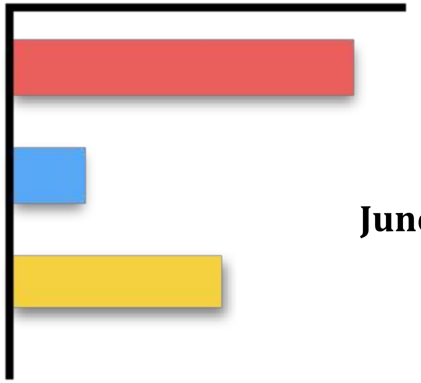
Junco Bird

Saliency Map Overview

Input

Model

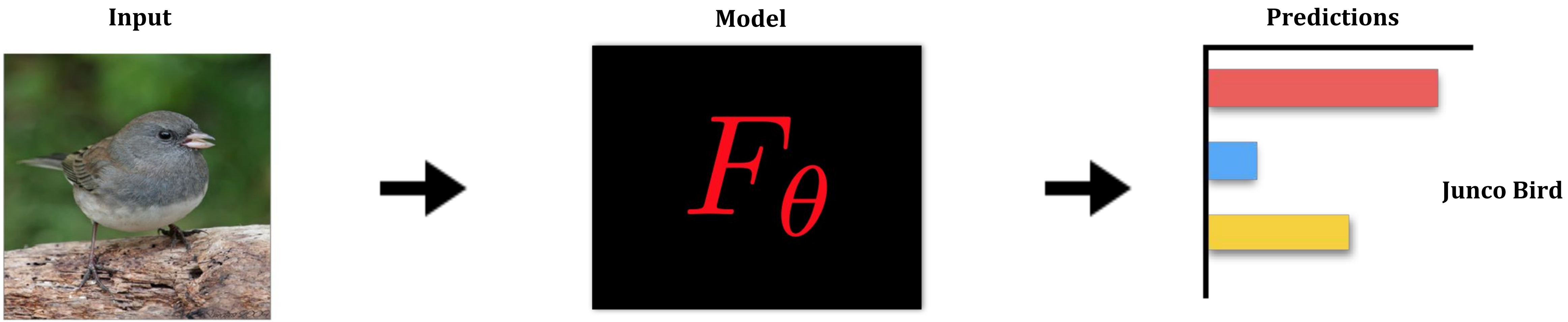
Predictions



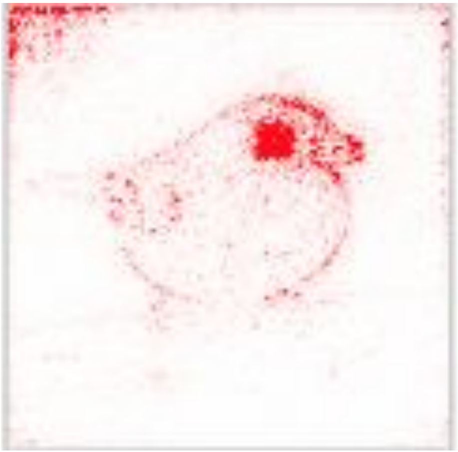
Junco Bird

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’?



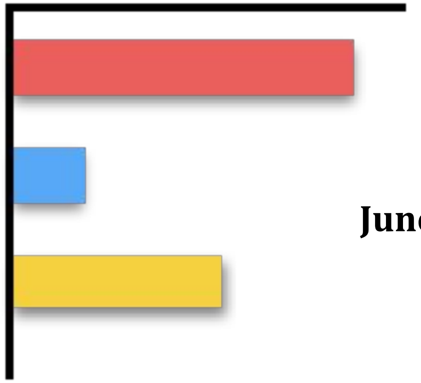
Saliency Map Overview

Input

Model

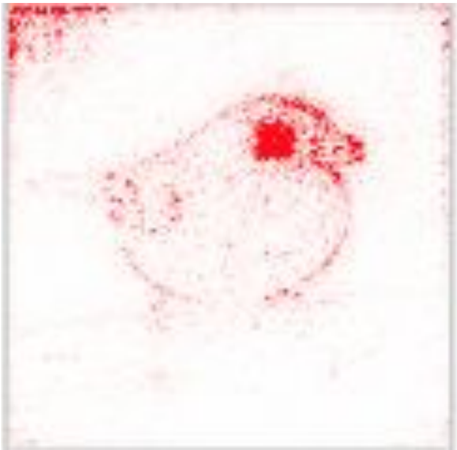
F_θ

Predictions



Junco Bird

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



- **Feature Attribution**
- **'Saliency Map'**
- **Heatmap**

A Linear Model Detour

$$y = w^\top x \quad x \in \mathbb{R}^d$$

$$y = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

A Linear Model Detour: Sensitivity

$$y = w^\top x \qquad x \in \mathbb{R}^d$$

$$y = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

How much does a unit change in an input dimension induce in the output?

A Linear Model Detour: Sensitivity

$$y = w^\top x \qquad x \in \mathbb{R}^d$$

$$y = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

How much does a unit change in an input dimension induce in the output?

$$\nabla_x y = w$$

$$\downarrow$$

$$\text{Sensitivity} \equiv (w_1, w_2, \ldots, w_d)$$

A Linear Model Detour: Attribution

$$y = w^\top x \qquad x \in \mathbb{R}^d$$

$$y = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

how can we apportion the output across all the input dimensions?

Another notion of relevance

$$y = w^\top x \qquad x \in \mathbb{R}^d$$

$$y = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

how can we apportion the output across all the input dimensions?

$$\downarrow$$

$$(w_1 x_1, w_2 x_2, \ldots, w_d x_d)$$

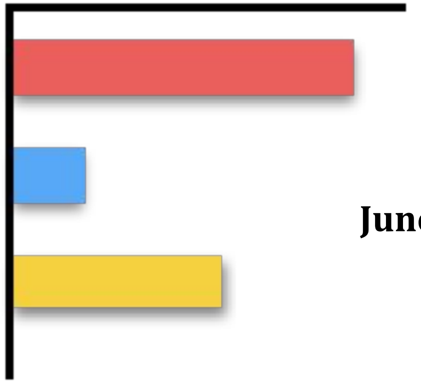
Modern DNN Setting

Input

Model

F_θ

Predictions



Junco Bird

$$F : \mathbb{R}^d \rightarrow \mathbb{R}^c$$

Model

$$F_i : \mathbb{R}^d \rightarrow \mathbb{R}$$

class specific logit

Input-Gradient

Input

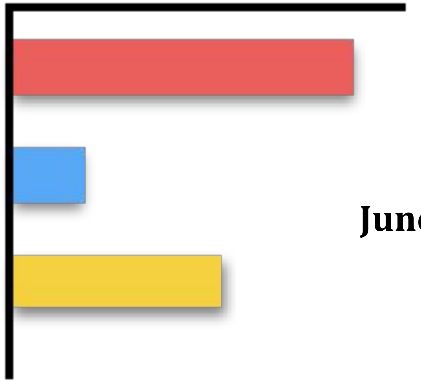
→

Model

$$F_\theta$$

→

Predictions



Junco Bird

↓

Input-Gradient

$$\nabla_x F_i(x) \rightarrow \in \mathbb{R}^d$$

Logit

Input

Same dimension as the input.

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

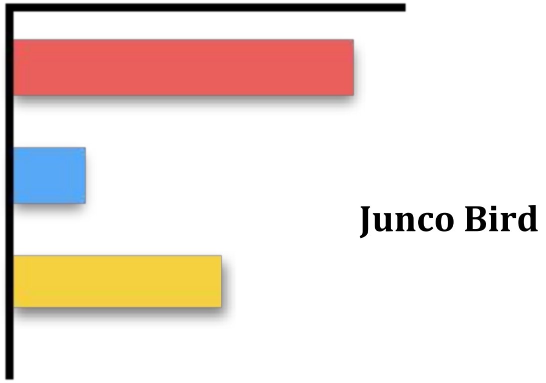
Input-Gradient

Input

Model

$$F_\theta$$

Predictions



Input-Gradient

$$\nabla_x F_i(x)$$

Logit

Input

Visualize as a heatmap

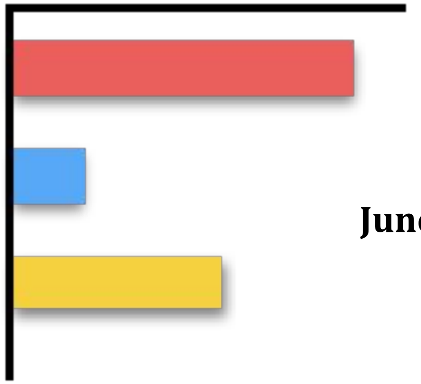
Baehrens et. al. 2010; Simonyan et. al. 2014 .

Input-Gradient

Input

Model

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](#) .

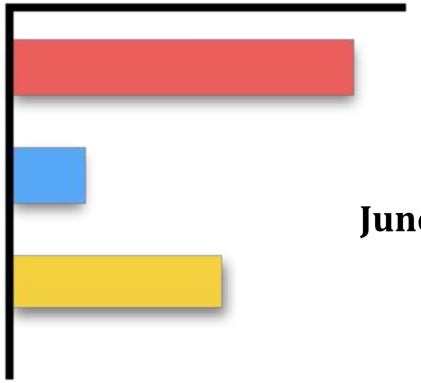
SmoothGrad

Input

Model

$$F_\theta$$

Predictions



SmoothGrad

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

Gaussian noise

Average Input-gradient of 'noisy' inputs.

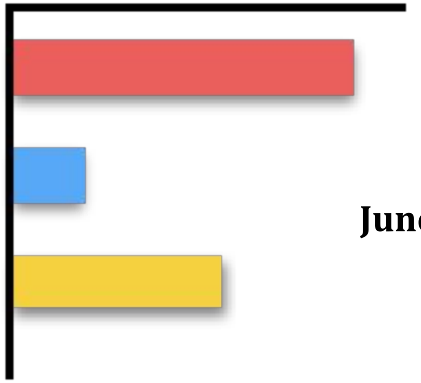
[Smilkov et. al. 2017]

SmoothGrad

Input

Model

Predictions



Junco Bird

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](#)

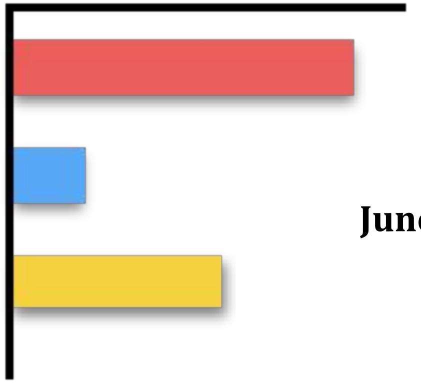
Integrated Gradients

Input

Model

F_θ

Predictions



Junco Bird

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

Baseline input

Path integral: ‘sum’ of interpolated gradients

[Sundararajan et. al. 2017]

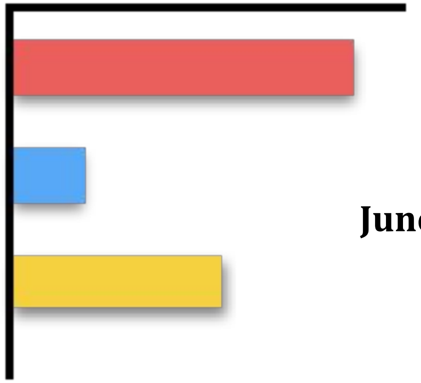
Integrated Gradients

Input

Model

F_θ

Predictions



Junco Bird

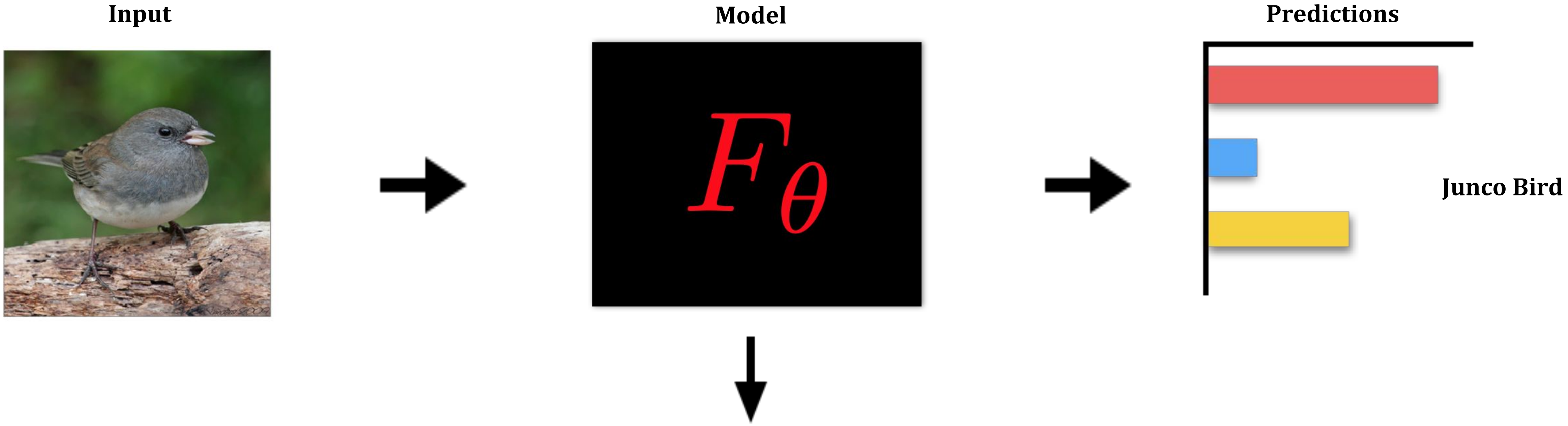
$$(x - \tilde{x}) \times \int_{\alpha=0}^{1} \frac{\partial F(\tilde{x} + \alpha \times (x - \tilde{x}))}{\partial x}$$

Baseline input

Path integral: ‘sum’ of interpolated gradients

[Sundararajan et. al. 2017](#)

Gradient-Input



Gradient-Input

$$\nabla_x F(x) \odot x$$

Input gradient **Input**

Element-wise product of input-gradient and input.

[Shrikumar et. al. 2017](), [Ancona et. al. 2018.]()

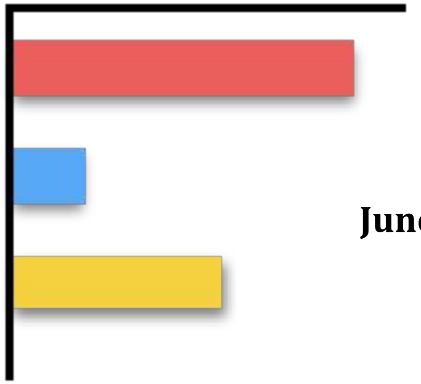
Gradient-Input

Input

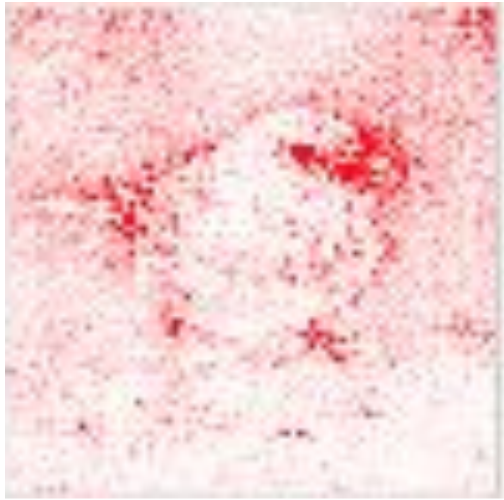
Model

F_θ

Predictions



Junco Bird



Gradient-Input

$$\nabla_x F(x) \odot x$$

logit gradient

Input

Element-wise product of input-gradient and input.

[Shrikumar et. al. 2017](), [Ancona et. al. 2018.]()

‘Modified Backprop’ Approaches

Compute feature relevance by modifying the backpropagation.

‘Modified Backprop’ Approaches

Compute feature relevance by modifying the backpropagation.

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

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

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

‘Modified Backprop’ Approaches

Compute feature relevance by modifying the backpropagation.

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

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

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

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

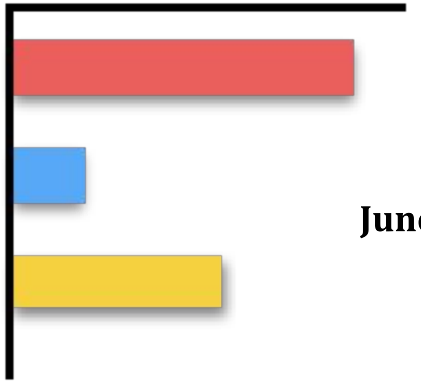
Attribution: Guided BackProp

Input

Model

F_θ

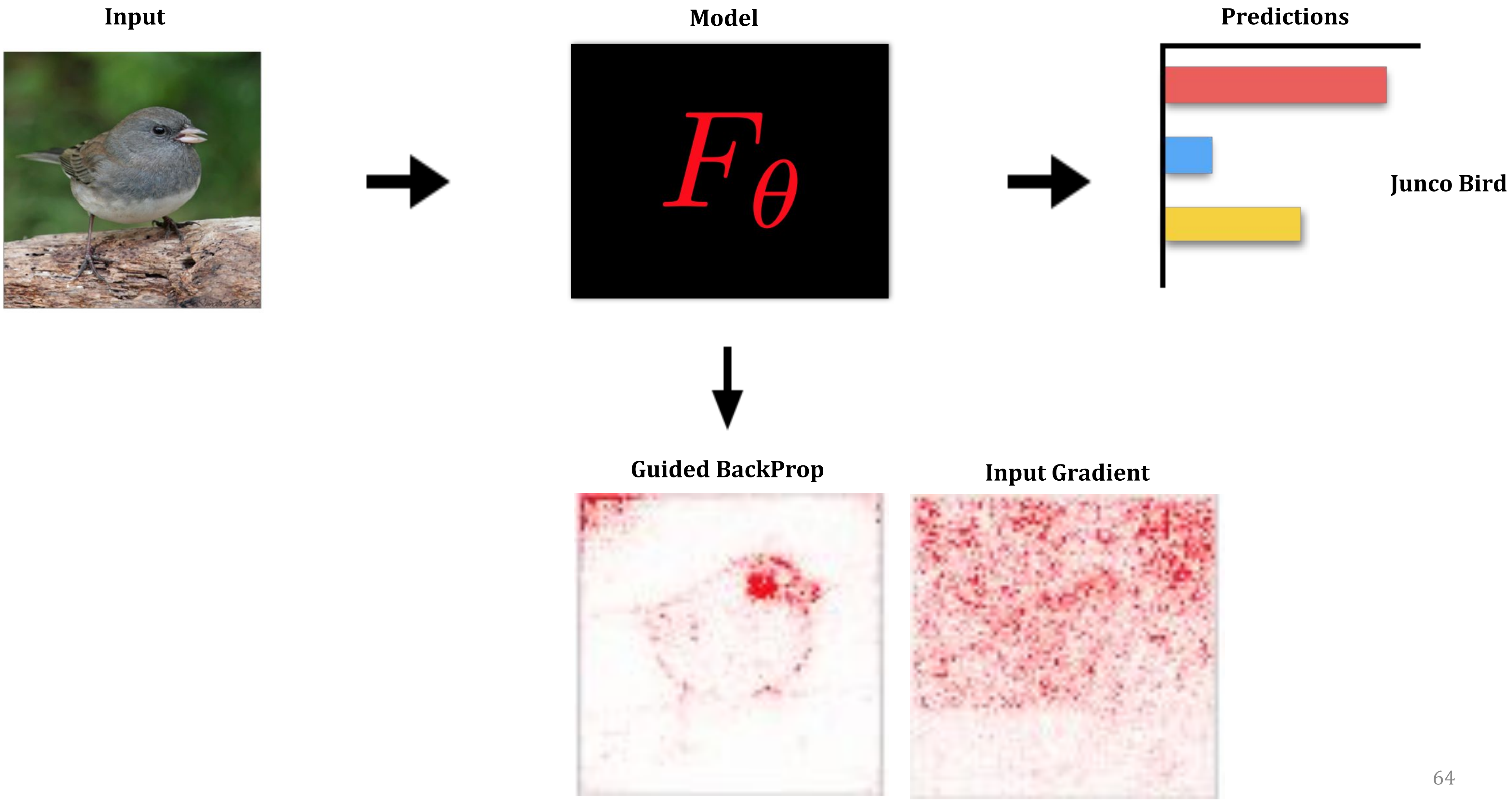
Predictions



Junco Bird

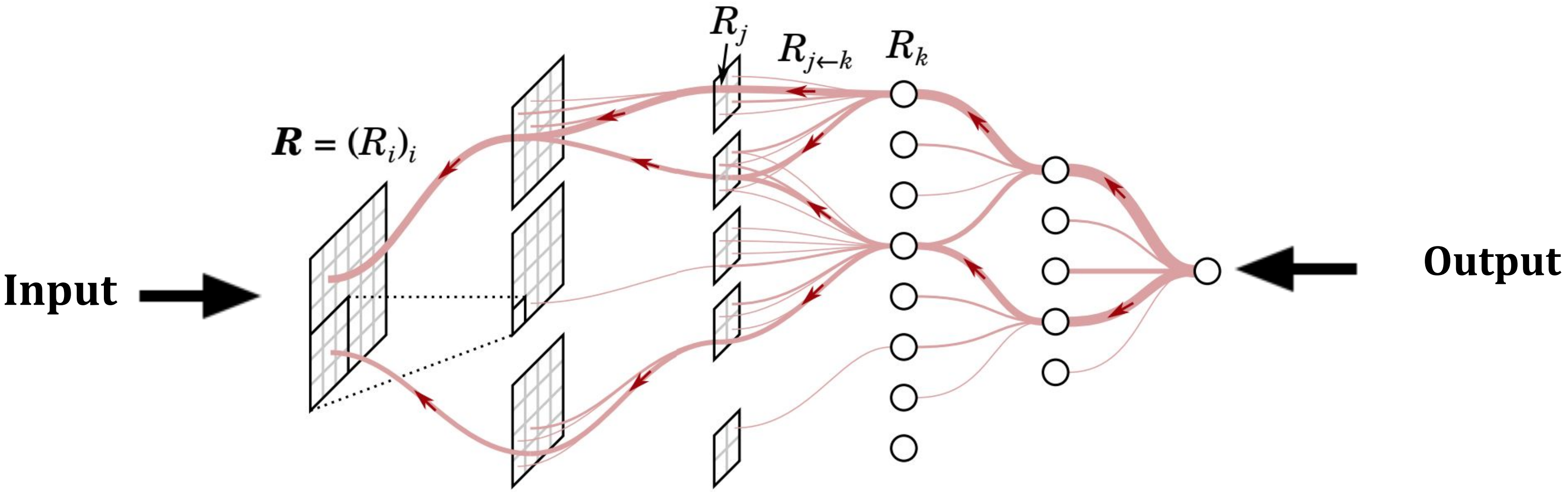
Guided BackProp

Attribution: Guided BackProp



Layer Relevance Propagation (LRP)

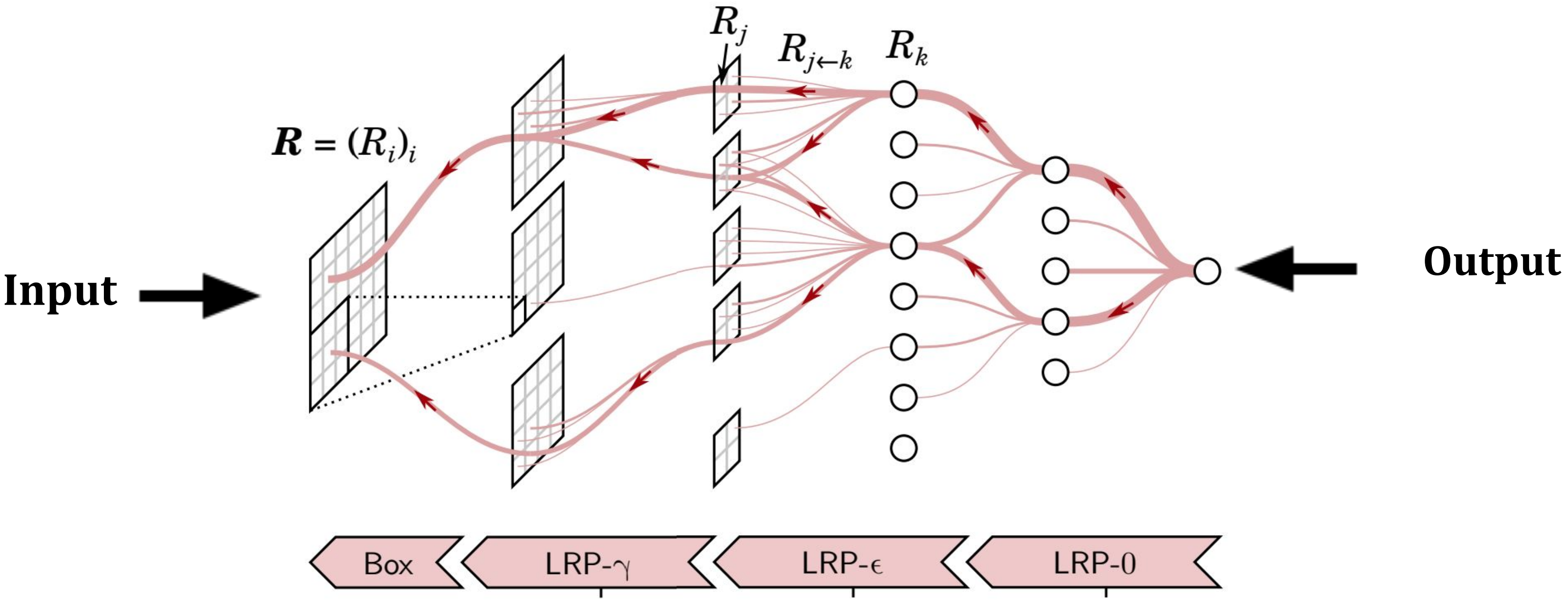
Compute feature relevance iteratively and propagate. Different **propagation rules** can be specified.



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

Layer Relevance Propagation (LRP)

Compute feature relevance iteratively and propagate. Different **propagation rules** can be specified.



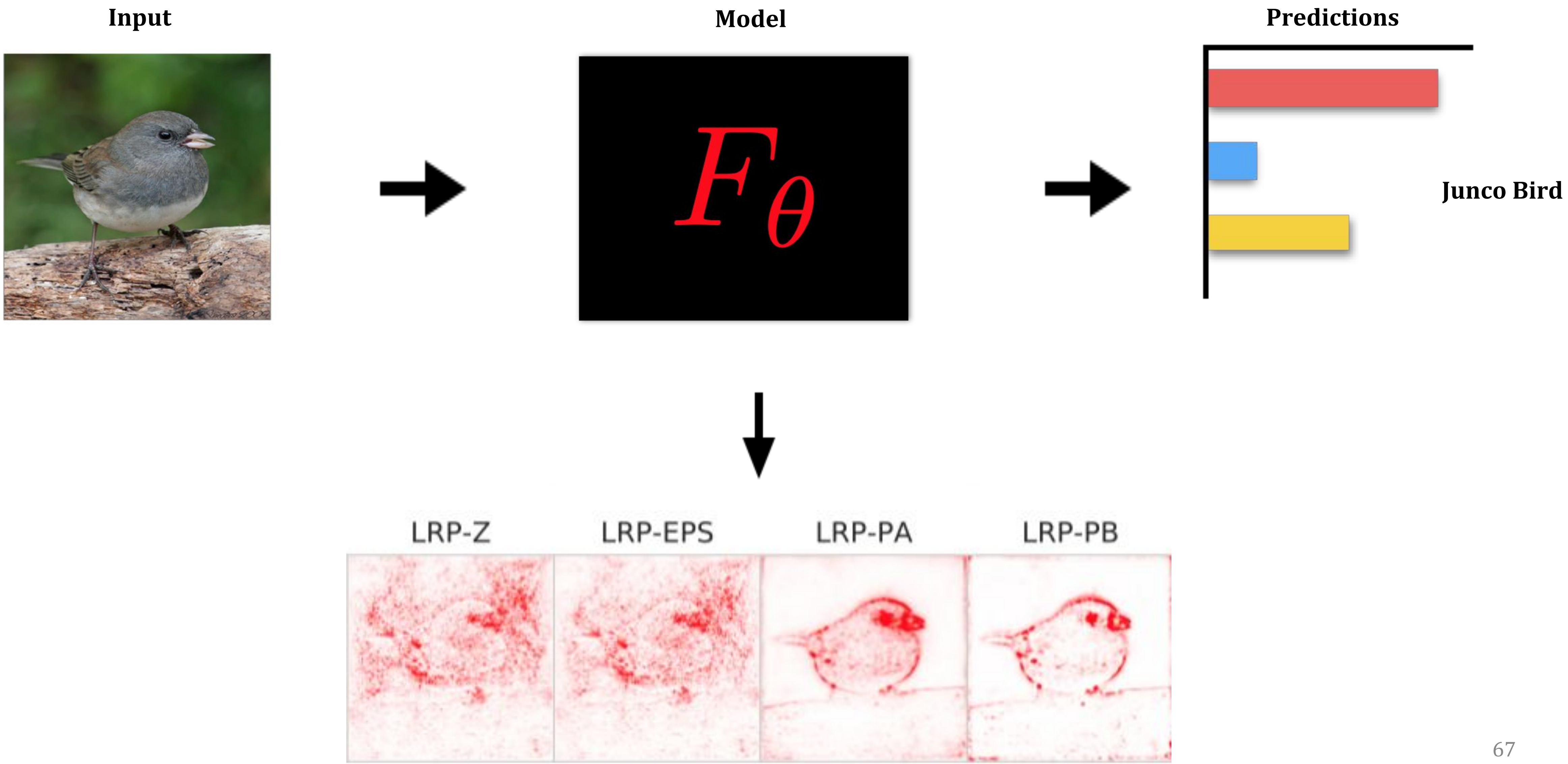
$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

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

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

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

Layer Relevance Propagation (LRP)



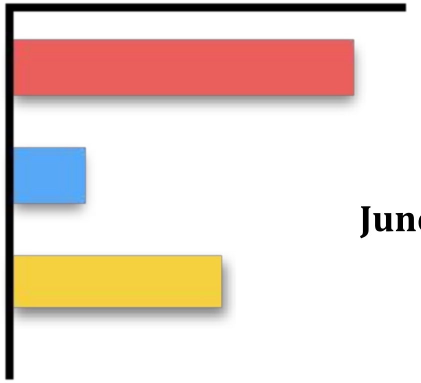
Recap

Input

Model

F_θ

Predictions



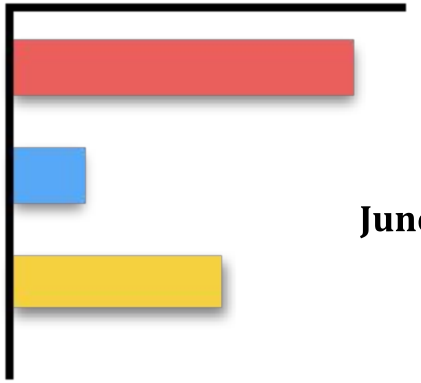
Junco Bird

Recap

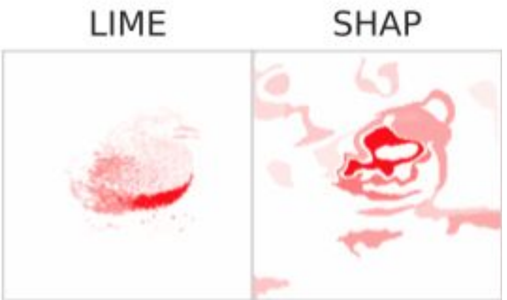
Input

Model

Predictions



Junco Bird

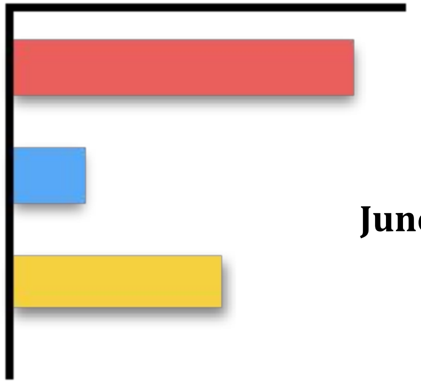


Recap

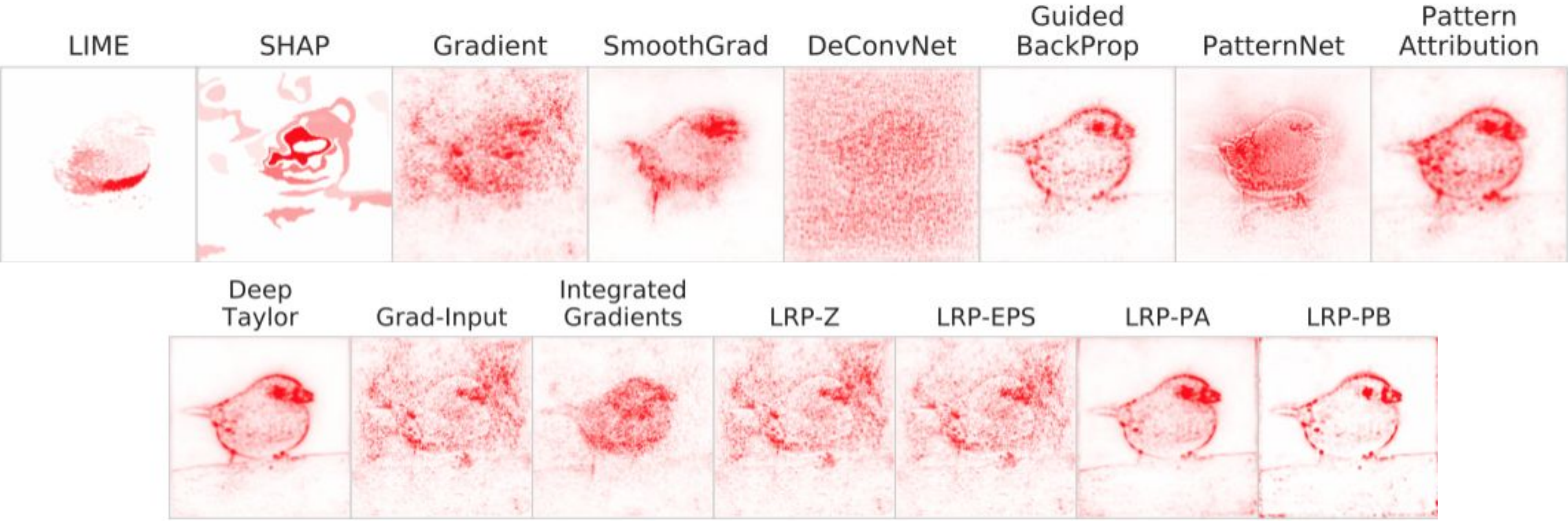
Input

Model

Predictions



Junco Bird



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
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

Prototype Approaches

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

Prototype Approaches

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

Insights

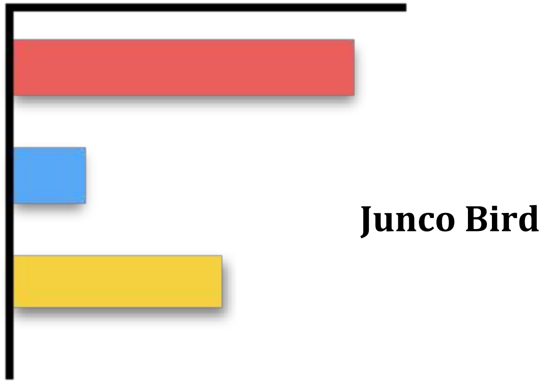
- What kind of input is the model **most likely to misclassify**?
- Which training samples are **mislabelled**?
- Which input **maximally activates** an intermediate neuron?

Training Point Ranking via Influence Functions

Input

Model

Predictions



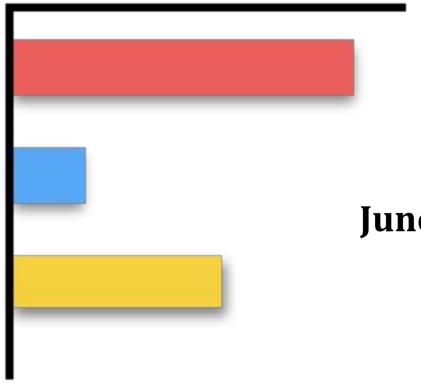
Which training data points have the most ‘*influence*’ on the test loss?

Training Point Ranking via Influence Functions

Input

Model

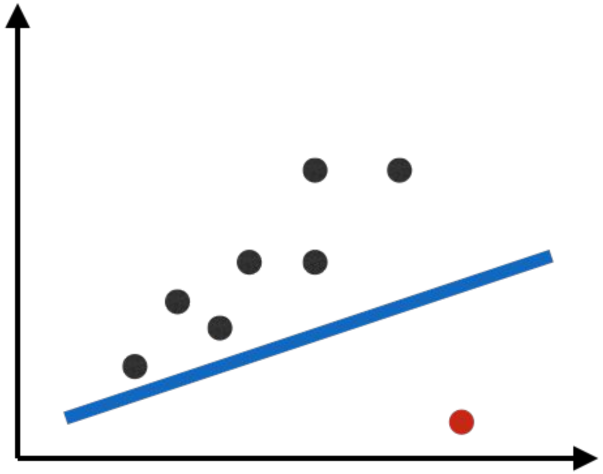
Predictions

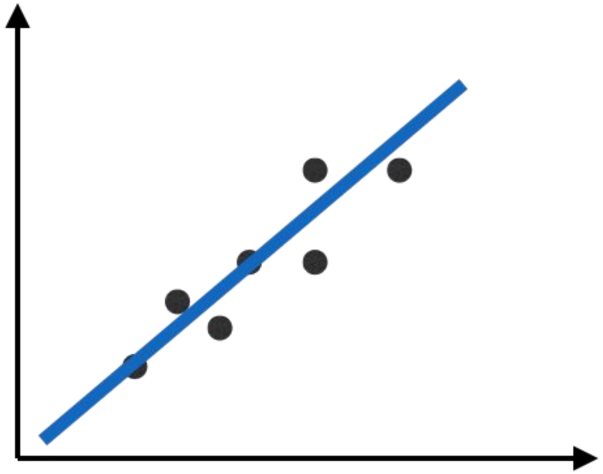


Which training data points have the most ‘*influence*’ on the test loss?

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).





Instead of refitting model for every data point, **Cook's distance** provides analytical alternative.

Training Point Ranking via Influence Functions

Koh & Liang (2017) extend the 'Cook's distance' insight to modern machine learning setting.

$z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ $\quad z_j = (x_j, y_j)$ ⬅ **Training sample point** $\quad z_{\text{test}}$

Training Point Ranking via Influence Functions

Koh & Liang (2017) extend the 'Cook's distance' insight to modern machine learning setting.

$z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ $\quad z_j = (x_j, y_j)$ ← **Training sample point** $\quad z_{\text{test}}$

ERM Solution

$$\hat{\theta} := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i; \theta)$$

UpWeighted ERM Solution

$$\hat{\theta}_{\epsilon, z_j} := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i; \theta) + \epsilon \ell(z_j; \theta) \qquad \epsilon = -\frac{1}{n}$$

Koh & Liang 2017

Training Point Ranking via Influence Functions

Koh & Liang (2017) extend the 'Cook's distance' insight to modern machine learning setting.

$$z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \qquad z_j = (x_j, y_j) \leftarrow$$ **Training sample point** $$z_{\text{test}}$$

ERM Solution

$$\hat{\theta} := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i; \theta)$$

UpWeighted ERM Solution

$$\hat{\theta}_{\epsilon, z_j} := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i; \theta) + \epsilon \ell(z_j; \theta) \qquad \epsilon = -\frac{1}{n}$$

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

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

Koh & Liang 2017

Training Point Ranking via **Influence Functions**

Applications:

- compute self-influence to identify mislabelled examples;
- diagnose possible domain mismatch;
- craft training-time poisoning examples.

[Koh & Liang 2017]

Training Point Ranking: NLP Application

Han et. al. (2020) use influence-based training point ranking to study spurious training artifacts in NLP setting.

Test input	
P: The manager was encouraged by the secretary. *H:* The secretary encouraged the manager.	{entail}
Most supporting training examples	
P: Because you're having fun. *H:* Because you're having fun.	[entail]
P: I don't know if I was in heaven or hell, said Lillian Carter, the president's mother, after a visit. *H:* The president's mother visited.	[entail]
P: Inverse price caps. *H:* Inward caps on price.	[entail]
P: Do it now, think 'bout it later. *H:* Don't think about it now, just do it.	[entail]
Most opposing training examples	
P: H'm, yes, that might be, said John. *H:* Yes, that might be the case, said John.	[non-entail]
P: This coalition of public and private entities undertakes initiatives aimed at raising public awareness about personal finance and retirement planning. *H:* Personal finance and retirement planning are initiatives aimed at raising public awareness.	[non-entail]

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

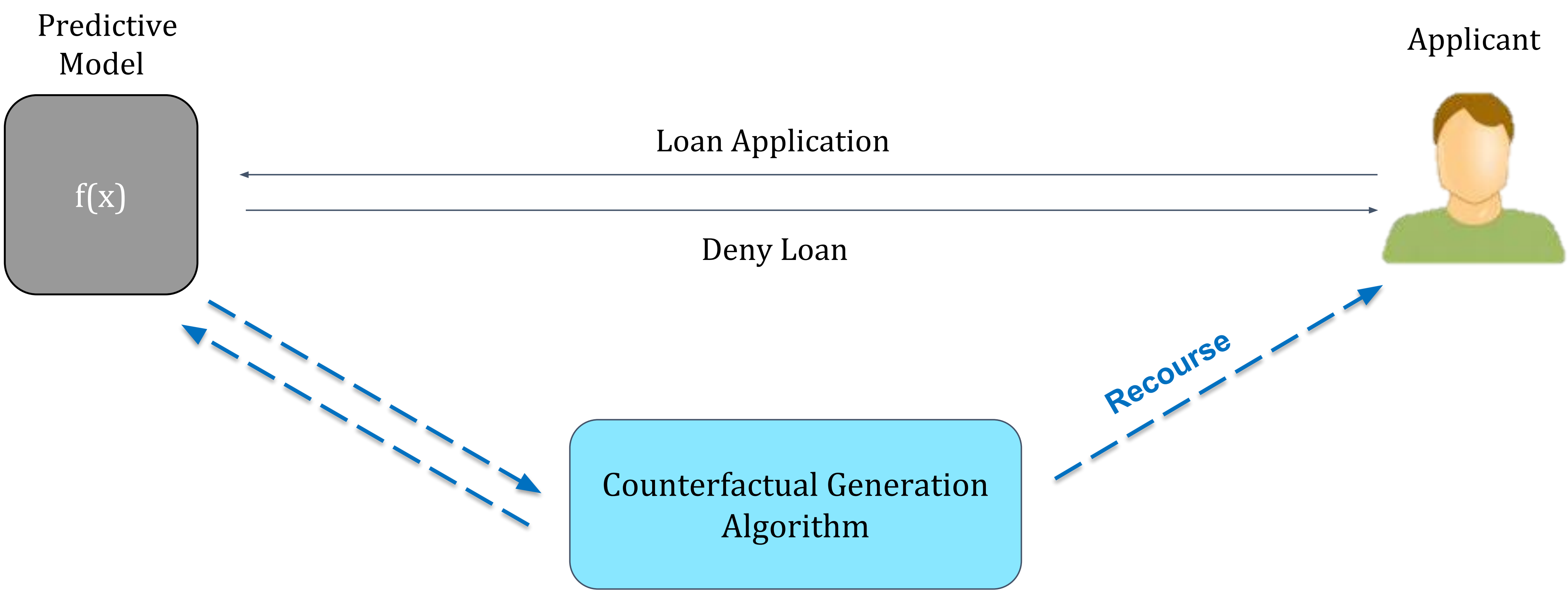
- Collection of Local Explanations
- Representation Based
- Model Distillation
- Summaries of Counterfactuals

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



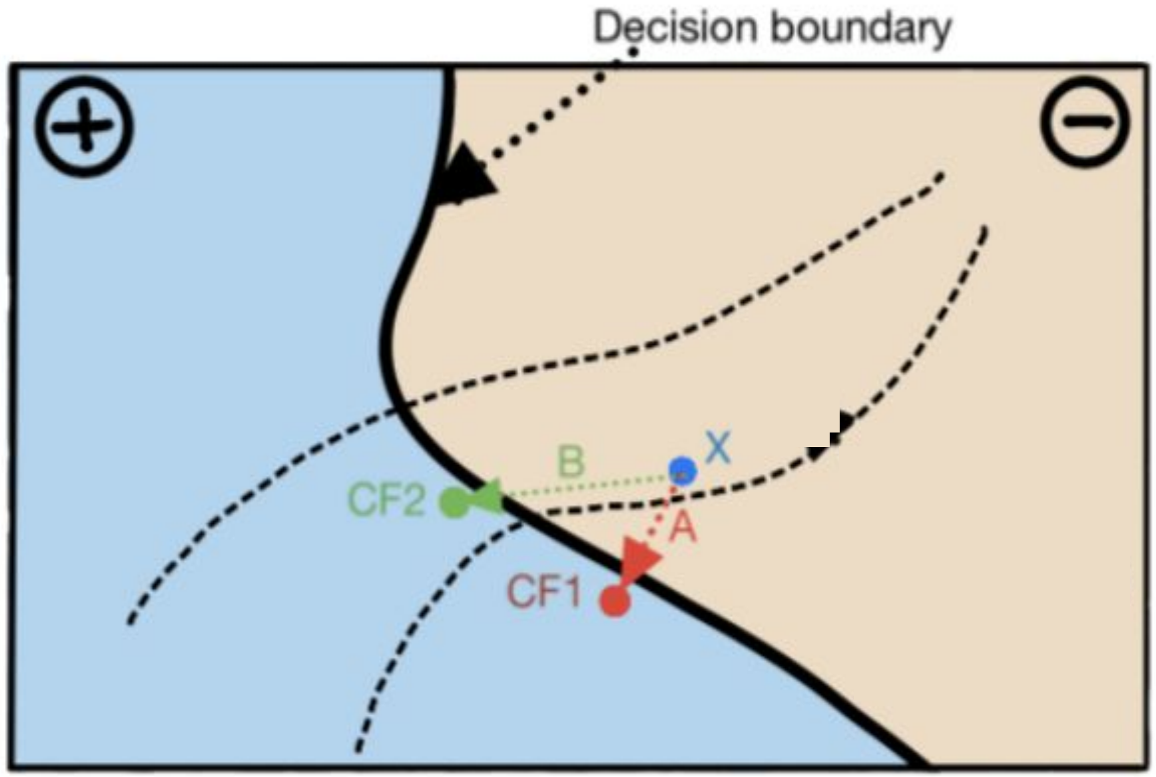
Recourse: Increase your salary by 50K & pay your credit card bills on time for next 3 months

Counterfactual Explanations

- Important to provide “recourse” to affected individuals (GDPR)

- 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).*

Generating Counterfactual Explanations: Intuition



Proposed solutions differ on:

1. How to choose among candidate counterfactuals?
2. How much access is needed to the underlying predictive model?

[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

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

$$\arg\min_{x'} \lambda \, (f(x') - y')^2 + d(x, x')$$

Wachter et. al. solve a differentiable, unconstrained version of the objective using ADAM optimization algorithm with random restarts.

This method ***requires access to gradients*** of the underlying predictive model.

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* is modeled as total log-percentile shift
  - Changes become harder when starting off from a higher percentile value

Take 2: Feasible and Least Cost Counterfactuals

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

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

- Ustun et. al. only consider the case where the model is a linear classifier
  - Objective formulated as an IP and optimized using CPLEX
- Requires ***complete access*** to the linear classifier i.e., weight vector

[Ustun et. al., 2019]

Take 2: Feasible and Least Cost Counterfactuals

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

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

Question: What if we have a black box or a non-linear classifier?

Answer: generate a local linear model approximation (e.g., using LIME) and then apply Ustun et. al.'s framework

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

Changing one feature without affecting another might not be possible!

Take 3: Causally Feasible Counterfactuals

Loan Applicant

Recourse:
Reduce current debt
from 3250$ to 1000$

After 1 year

Loan Applicant

My current debt has
reduced to 1000$.
Please give me loan.

Predictive Model

f(x)

Your age increased by 1
year and the recourse is
no longer valid! Sorry!

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

But how?!

Take 3: Causally Feasible Counterfactuals

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

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

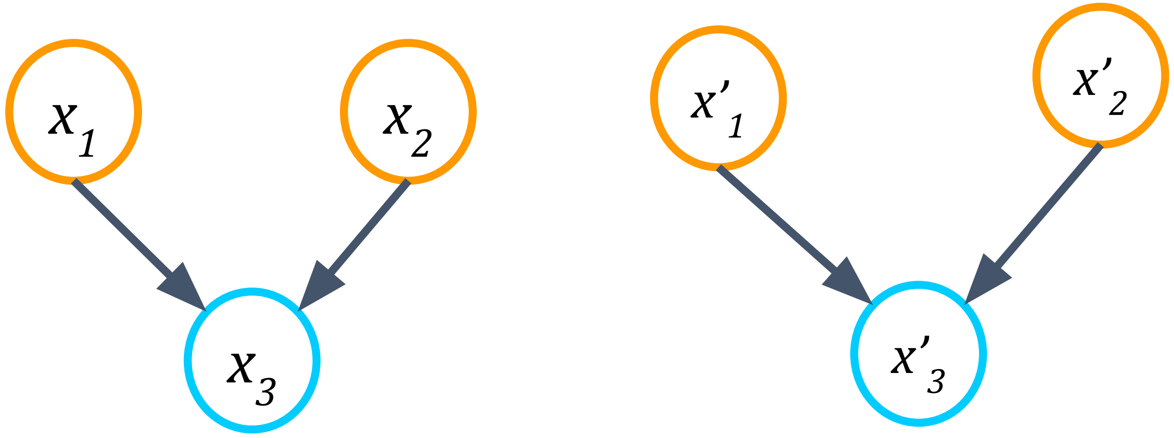
Leverage Structural Causal Model (SCM) to define this new distance metric

Take 3: Causally Feasible Counterfactuals

$$d\_causal(x, x') =$$

$$\sum_{u \in U} \underbrace{d(x_u, x'_u)}_{\text{Standard L1/L2 distance for each variable } u \text{ with no parents}} +$$

$$\sum_{v \in V} \underbrace{d(x_v, \mathbb{E}[x'_v | x'_{v_{p1}}, x'_{v_{p2}}, \cdots x'_{v_{pM}}])}_{\text{For variables } v \text{ with parents, compute L1/L2 distance between value of } v \text{ for original instance and expected value of } v \text{ given its parents for counterfactual}}$$



Original Instance

Counterfactual

U is set of nodes without parents in the graph;
V is set of nodes with parents in the graph

[Mahajan et. al., 2019]

Take 3: Causally Feasible Counterfactuals

- Requires knowledge of full causal graph
  - *Empirically*, partial knowledge also seems to work fine
  - Learn about feasibility constraints/partial causal graph from user inputs

- Solving the objective: Leverage a Variational Autoencoder (VAE)
  - ***requires access to gradients*** of the underlying predictive model.

[Verma et. al., 2020, Poyiadzi et. al. 2020]

Other Takes on Feasible Counterfactuals

- Data Manifold Closeness: Generated counterfactual should be "close to" the original data distribution.

- Sparsity: Ideal to change small number of features in the counterfactual

[Verma et. al., Poyiadzi et. al. 2020]

Other Takes on Feasible Counterfactuals

- Data Manifold Closeness: Generated counterfactual should be “close to” the original data distribution.
    - Include term to minimize the distance (e.g., averaged Euclidean distance) between counterfactual and all original data instances

- Sparsity: Ideal to change small number of features in the counterfactual
    - Include term to minimize the total number of features being changed to obtain desired outcome (e.g., L0/L1 norm)

Approaches for Post hoc Explainability

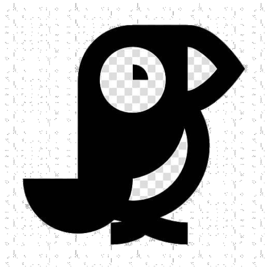
Local Explanations

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

Global Explanations

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

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

Local vs. Global Explanations

Explain individual predictions

Help unearth biases in the *local neighborhood* of a given instance

Help vet if individual predictions are being made for the right reasons

Explain complete behavior of the model

Help shed light on *big picture biases* affecting larger subgroups of the population

Help vet if the model, at a high level, is suitable for deployment

Local vs. 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

Approaches for Post hoc Explainability

Local Explanations

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

Global Explanations

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

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

What local explanation technique to use?
How to choose the subset of k local explanations?

[Ribeiro et al. 2016]

Global Explanations from Local Feature Importances: SP-LIME

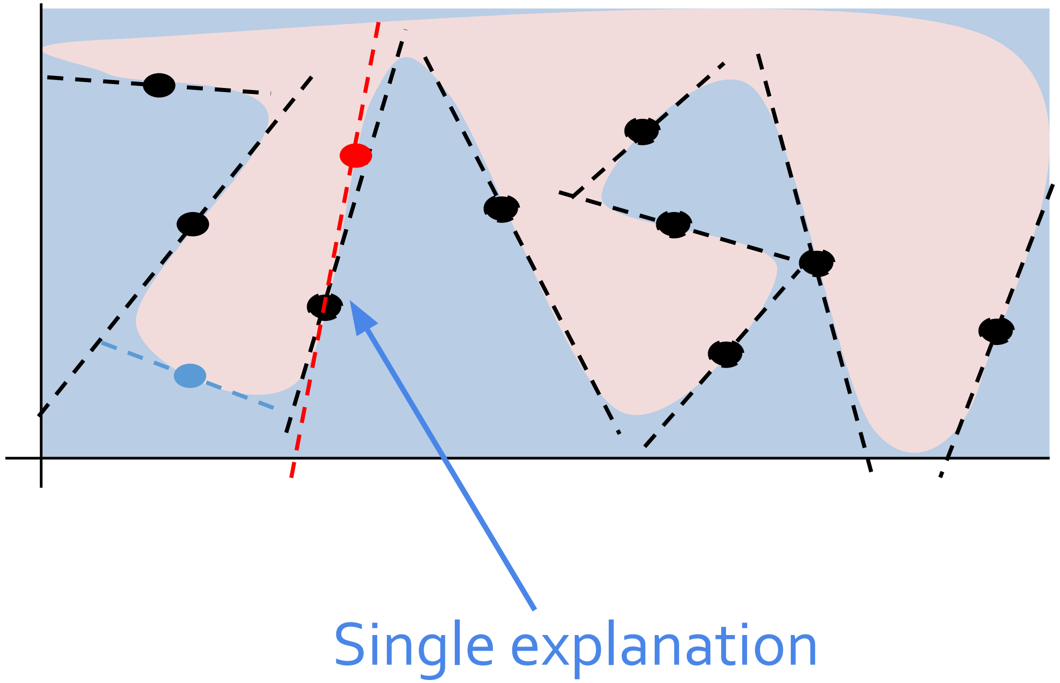
LIME explains a single prediction
local behavior for a single instance

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



Model Agnostic

Picking *k* Explanations: Intuition

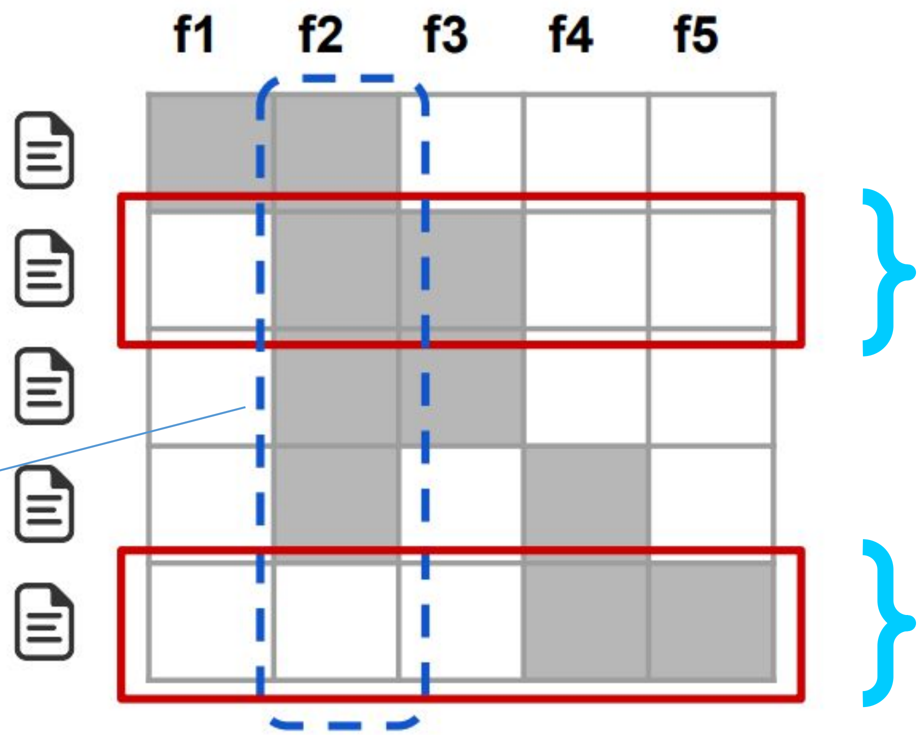
Aggregate
Feature Importances
across all instances

+

"Coverage" of
Features

[Ribeiro et al. 2016]

Global Explanations from Local Feature Importances: SP-LIME



f2 is an important feature across several instances

Explanations of these instances will be selected by submodular optimization algorithm

Rows represent instances

Columns represent features

[Ribeiro et. al., 2018]

Global Explanations from Local Rule Sets: SP-Anchor

Model Agnostic

Use the same approach as above with Anchors algorithm (instead of LIME) which produces local rule sets as explanations.

Global Explanations from Local Rule Sets: SP-Anchor

- Use Anchors algorithm discussed earlier to obtain local rule sets for every instance in the data

- Use the same procedure to *greedily select a subset of k local rule sets* to correspond to the global explanation

Model Agnostic

Approaches for Post hoc Explainability

Local Explanations

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

Global Explanations

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

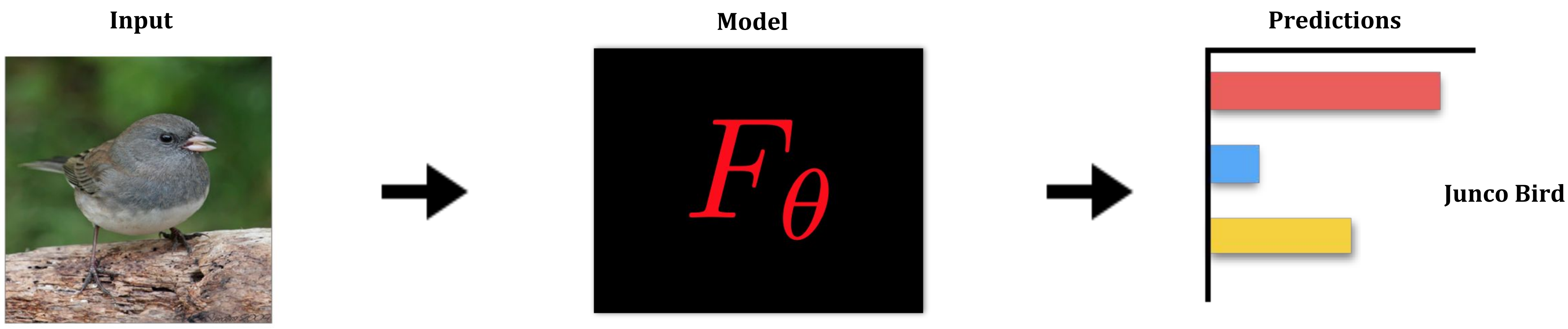
Representation Based Approaches

- Derive model understanding by analyzing intermediate representations of a DNN.

- Determine model’s reliance on ‘concepts’ that are semantically meaningful to humans.

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.



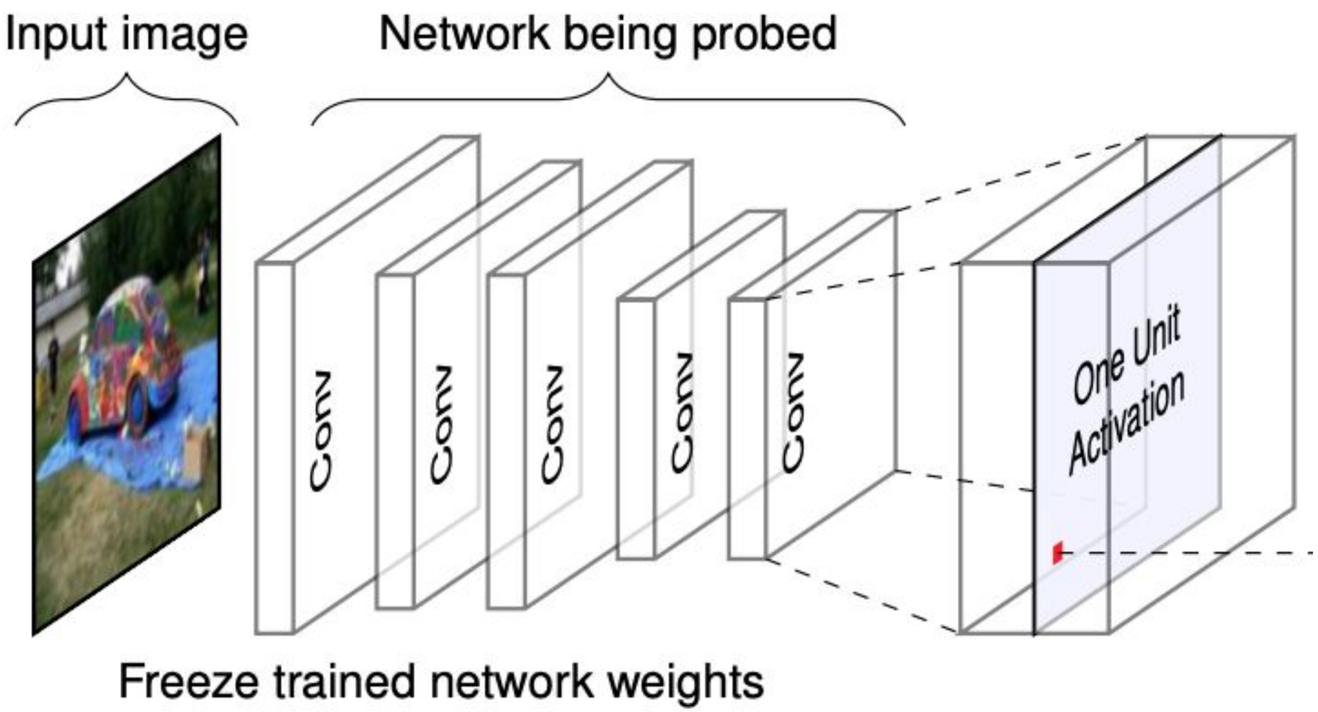
Does the model rely on the ‘**green background**’?

Network Dissection

1. Identify a broad set of human-labeled visual concepts.

Bau & Zhou et. al. 2017

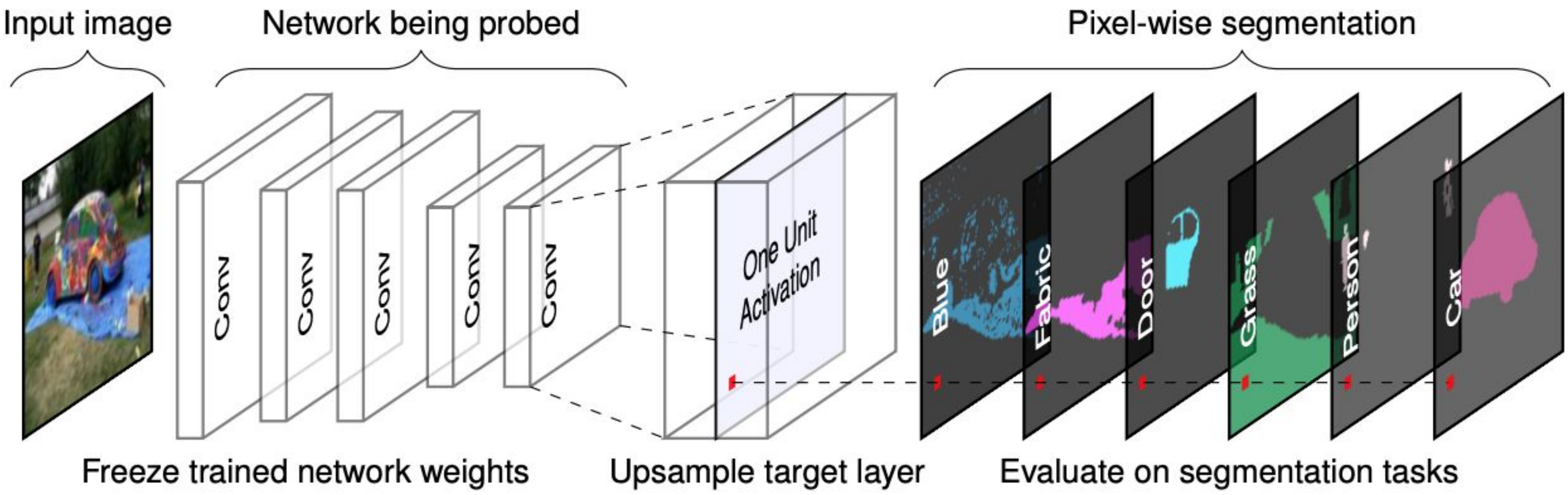
Network Dissection



1. Identify a broad set of human-labeled visual concepts.
2. Gather the response of hidden variables (convolutional filters) to known concepts.

[Bau & Zhou et. al. 2017]

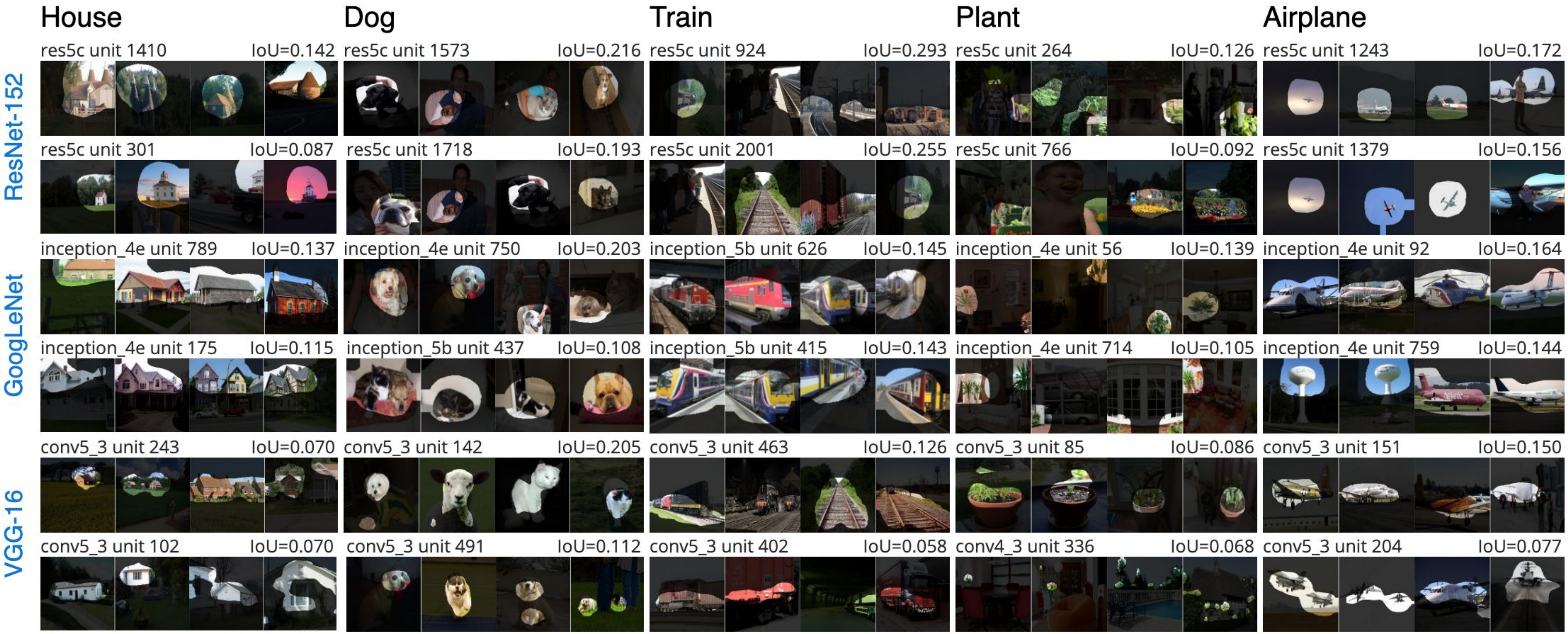
Network Dissection



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

[Bau & Zhou et. al. 2017]

Network Dissection



[Bau & Zhou et. al. 2017]

Compositional Extension

Natural Language Inference

Unit 870 (gender-sensitive)

`((((NOT hyp:man) AND pre:man) OR hyp:eating) AND (NOT pre:woman)) OR hyp:dancing`

IoU **0.123** w_{entail} -0.046 $w_{neutral}$ -0.021 w_{contra} 0.040

Pre A guy pointing at a giant blackberry.
Hyp A woman tearing down a giant display.
Act **29.31** True **contra** Pred **contra**

Pre A man in a hat is working with...flowers.
Hyp Women are working with flowers.
Act **27.64** True **contra** Pred **contra**

Unit 99 (high overlap)

`((NOT hyp:JJ) AND overlap-75% AND (NOT pre:people)) OR pre:basket OR pre:tv`

IoU **0.118** w_{entail} 0.043 $w_{neutral}$ -0.029 w_{contra} -0.021

Pre A woman in a light blue jacket is riding a bike.
Hyp A women in a jacket riding a bike.
Act **19.13** True **entail** Pred **entail**

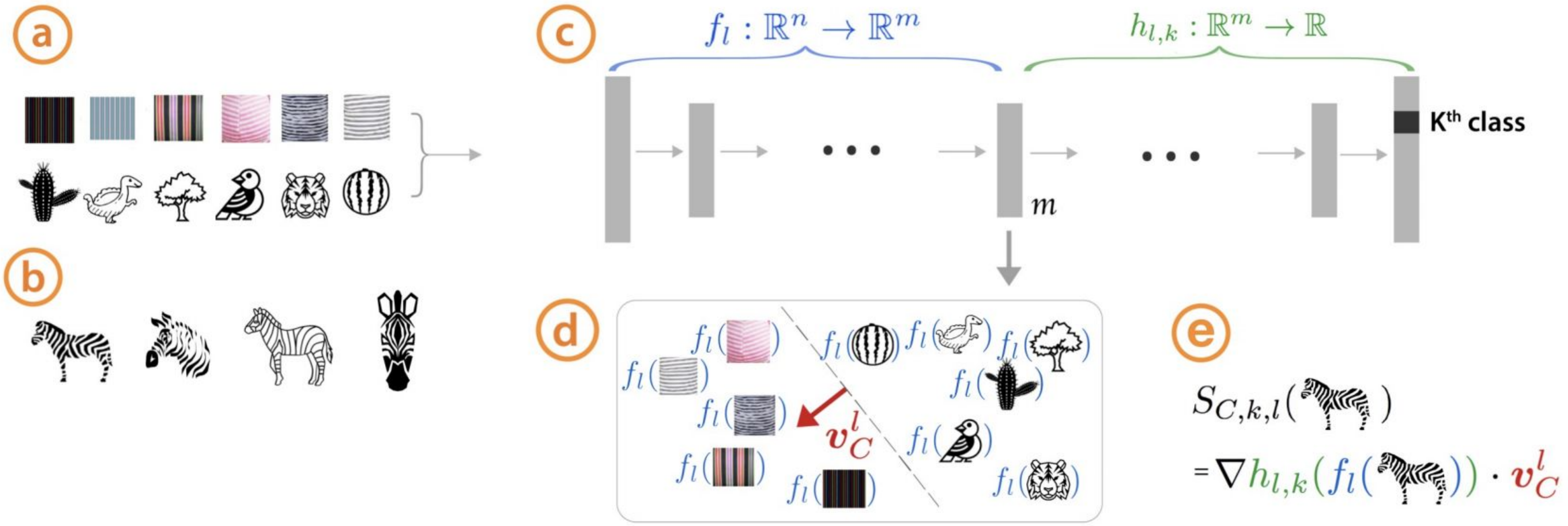
Pre A girl in a pumpkin dress sitting at a table.
Hyp There is a girl in a pumpkin dress sitting at a table.
Act **17.84** True **entail** Pred **entail**

Vision

Mu & Andreas at Neurips 2020

Quantitative Testing with Concept Activation Vectors (TCAV)

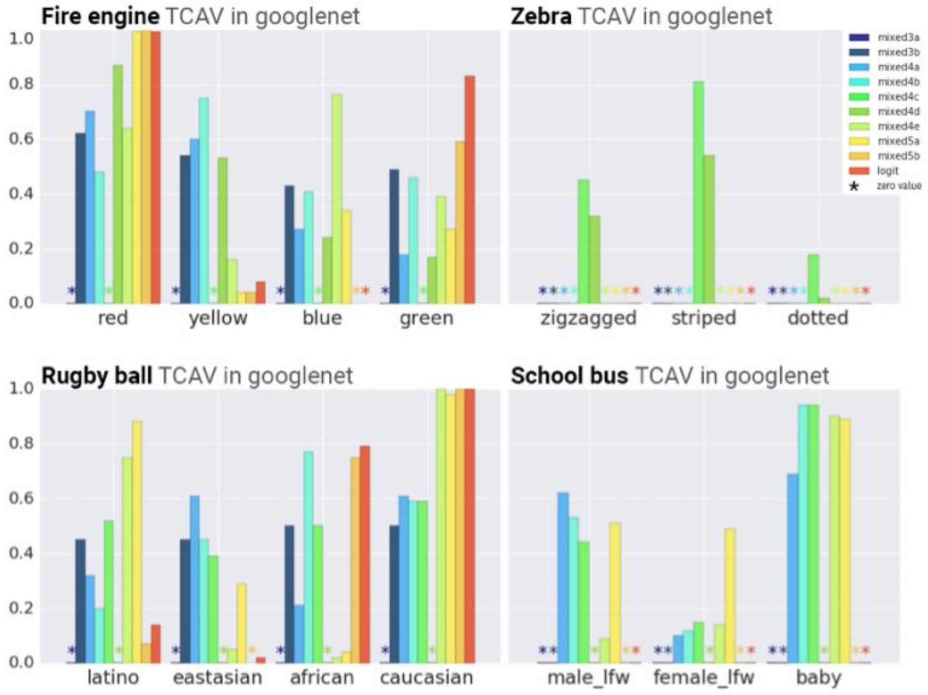
TCAV measures the sensitivity of a model's prediction to **user provided concept** using the model **internal representations**.

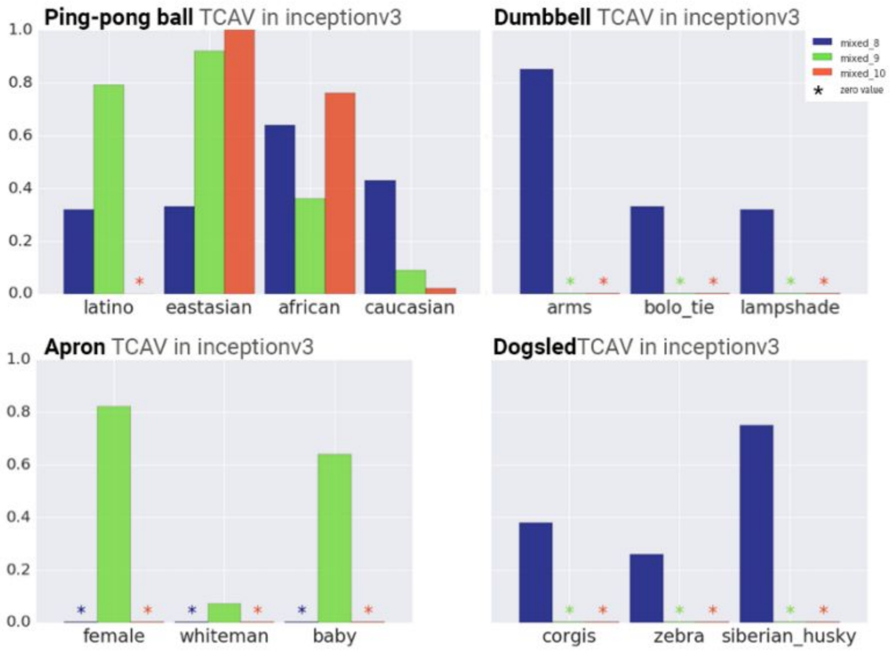


[Kim et. al. 2018](#)

Quantitative Testing with Concept Activation Vectors (TCAV)

Insights from Googlenet and Inceptionv3

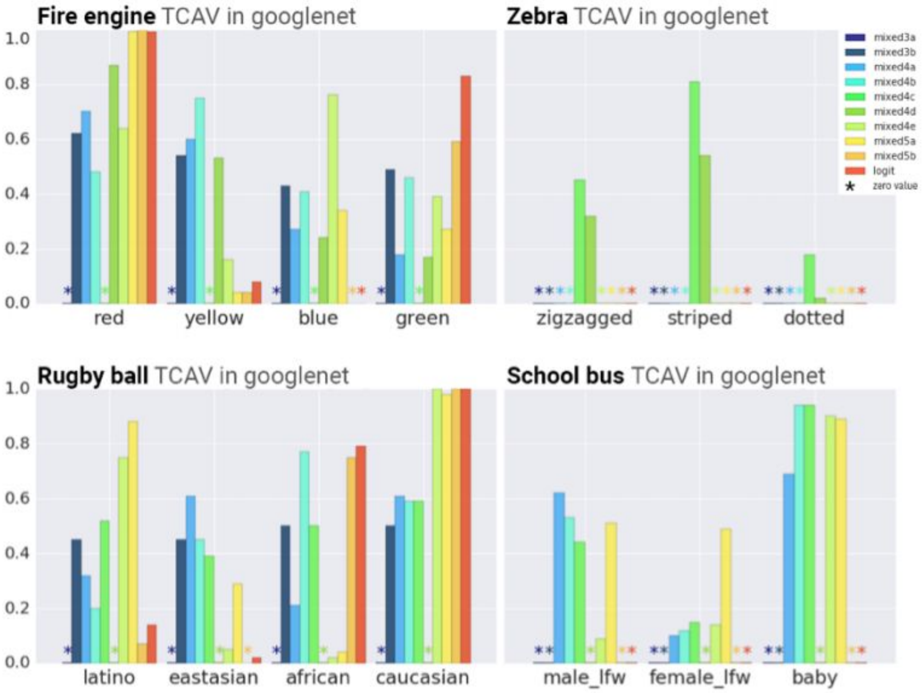


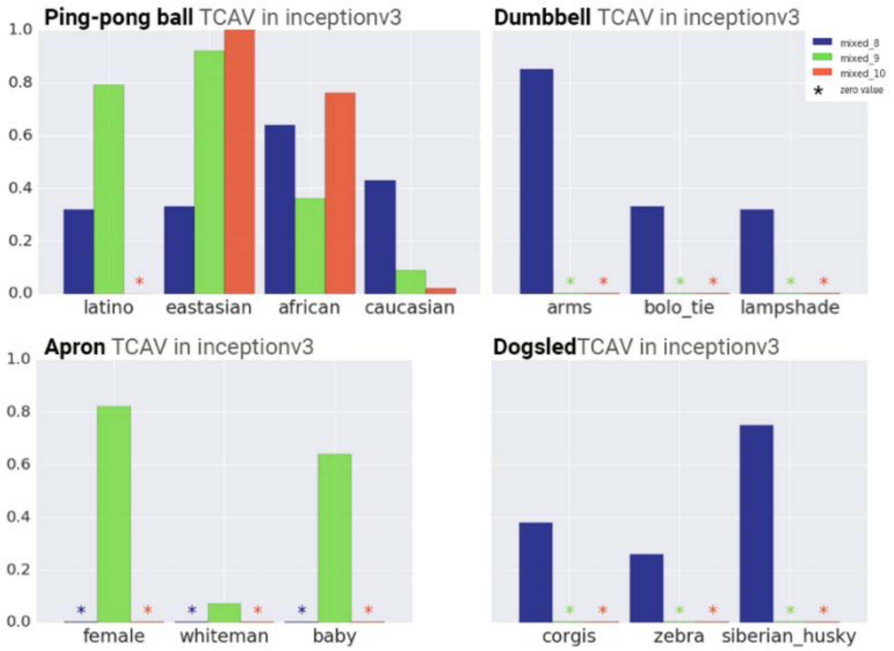


Images from [Kim et. al. 2018]

Quantitative Testing with Concept Activation Vectors (TCAV)

Insights from Googlenet and Inceptionv3





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**

Connections to Probing and Representational Similarity

- The line of work presented has connections to the literature on **probing in NLP**.

- See recent tutorial by Belinkov, Gehrmann, & Pavlick at ACL 2020 for additional discussion

Connections to Probing and Representational Similarity

- The line of work presented has connections to the literature on **probing in NLP**.

- See recent tutorial by Belinkov, Gehrmann, & Pavlick at ACL 2020 for additional discussion

Representational Similarity

1. How similar are the representations at the lower layers of a model compared to its higher layers.

2. How similar are the representations of one model to another?

See: Raghu et. al. 2017 & Kornblith et. al. 2019 for techniques that can provide insights on the questions above.

Approaches for Post hoc Explainability

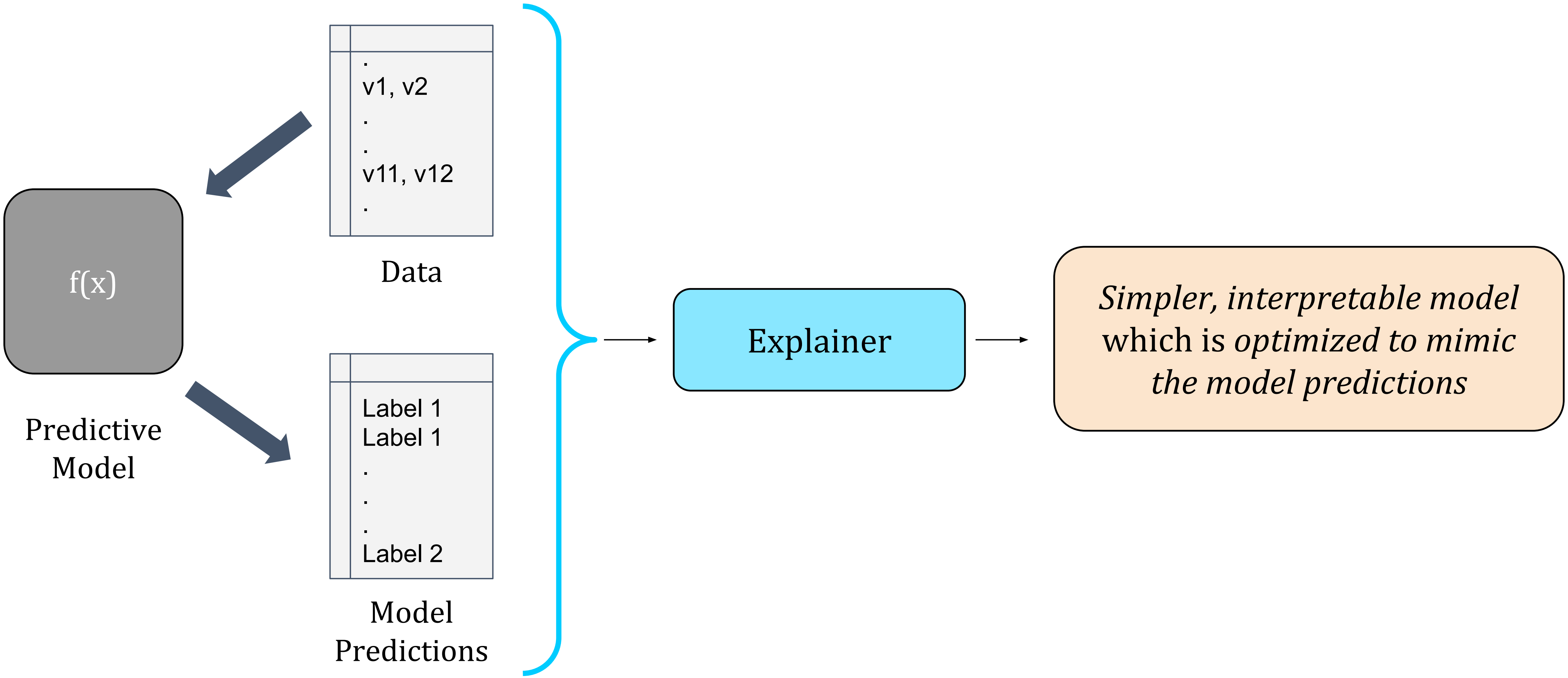
Local Explanations

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

Global Explanations

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

Model Distillation for Generating Global Explanations



[Tan et. al., 2019]

Generalized Additive Models as Global Explanations

Black Box Model

.
v1, v2
.
.
v11, v12
.

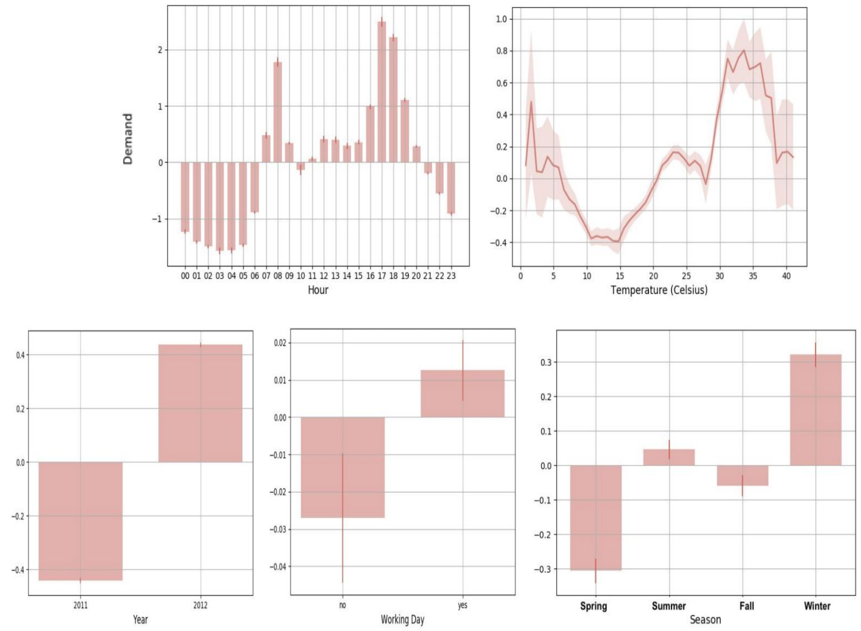
Data

Label 1
Label 1
.
.
.
Label 2

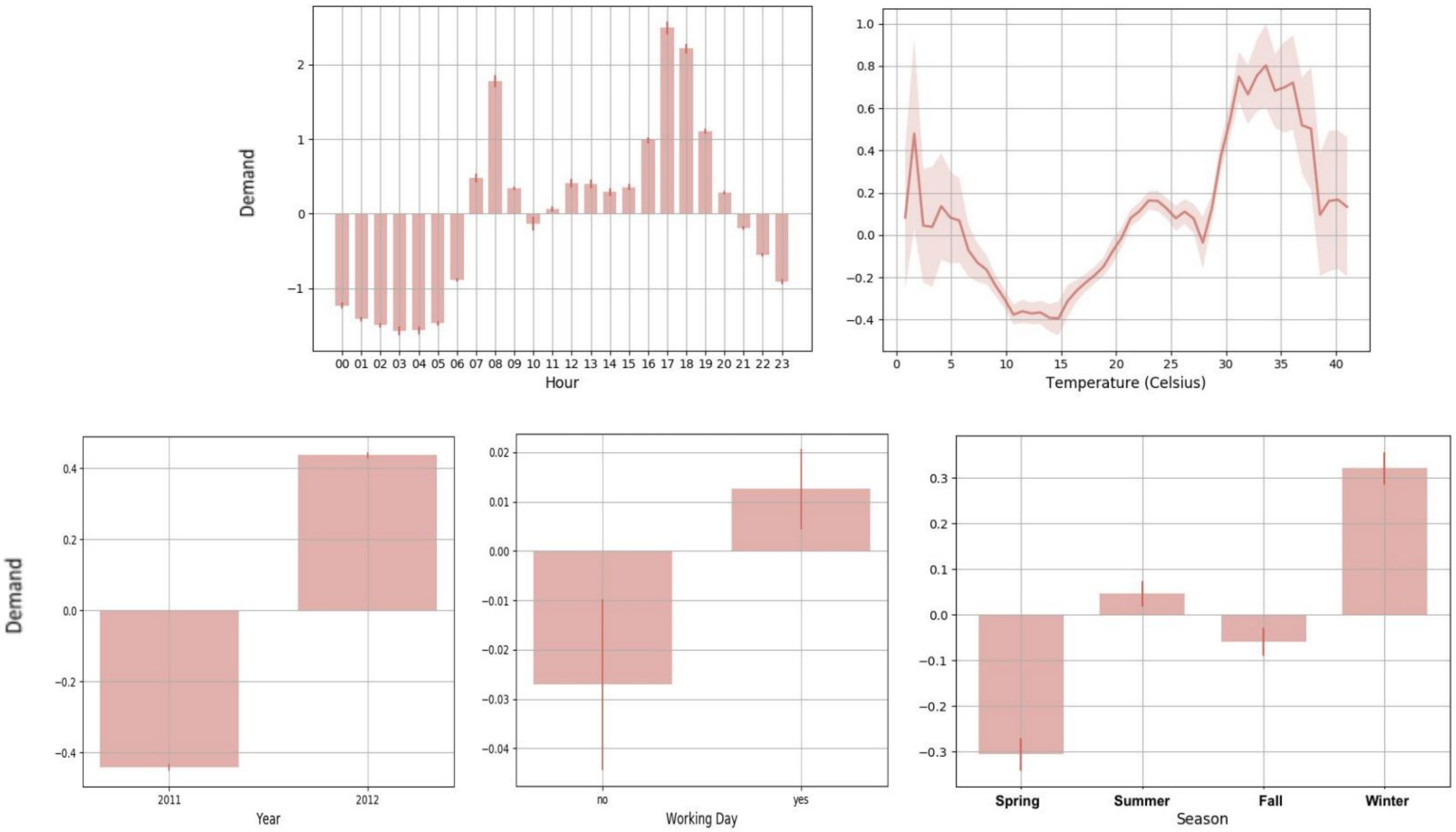
Model Predictions

Model Agnostic

Explainer



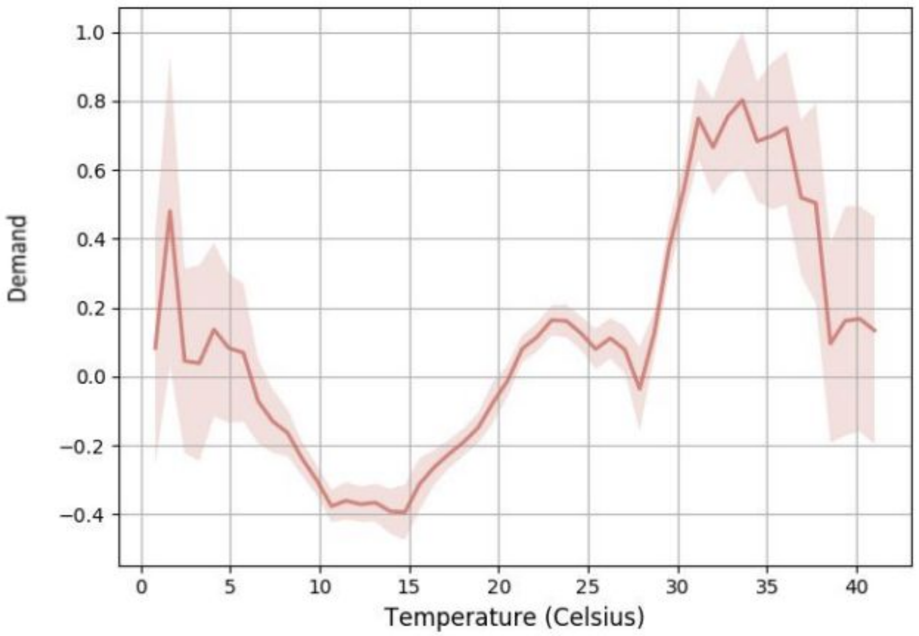
Generalized Additive Models as Global Explanations: *Shape Functions* for Predicting Bike Demand



[Tan et. al., 2019]

Generalized Additive Models as Global Explanations: *Shape Functions* for Predicting Bike Demand

How does bike demand vary as a function of temperature?



Generalized Additive Models as Global Explanations

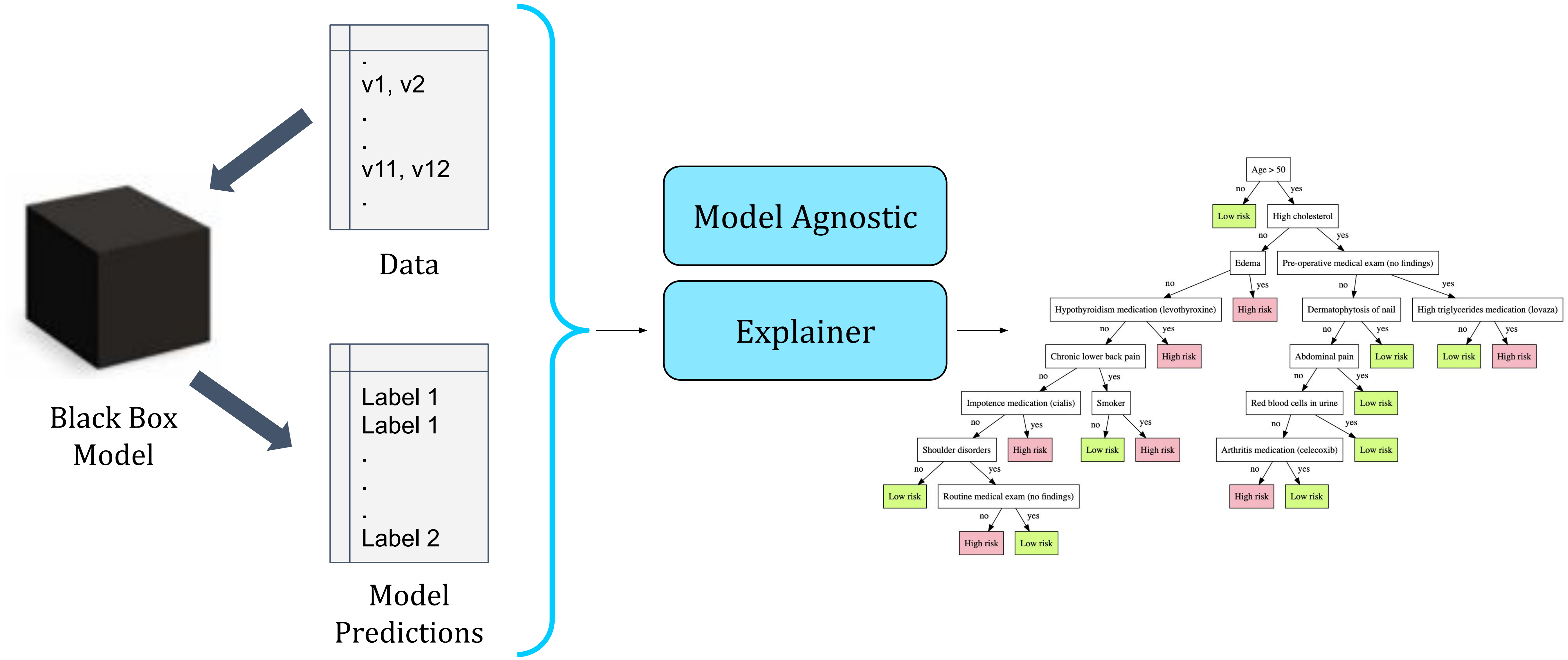
Generalized Additive Model (GAM) :

$$\hat{y} = h_0 + \underbrace{\sum_i h_i(x_i)}_{\text{Shape functions of individual features}} + \underbrace{\sum_{i \neq j} h_{ij}(x_i, x_j) + \sum_{i \neq j} \sum_{j \neq k} h_{ijk}(x_i, x_j, x_k) + \cdots}_{\text{Higher order feature interaction terms}}$$

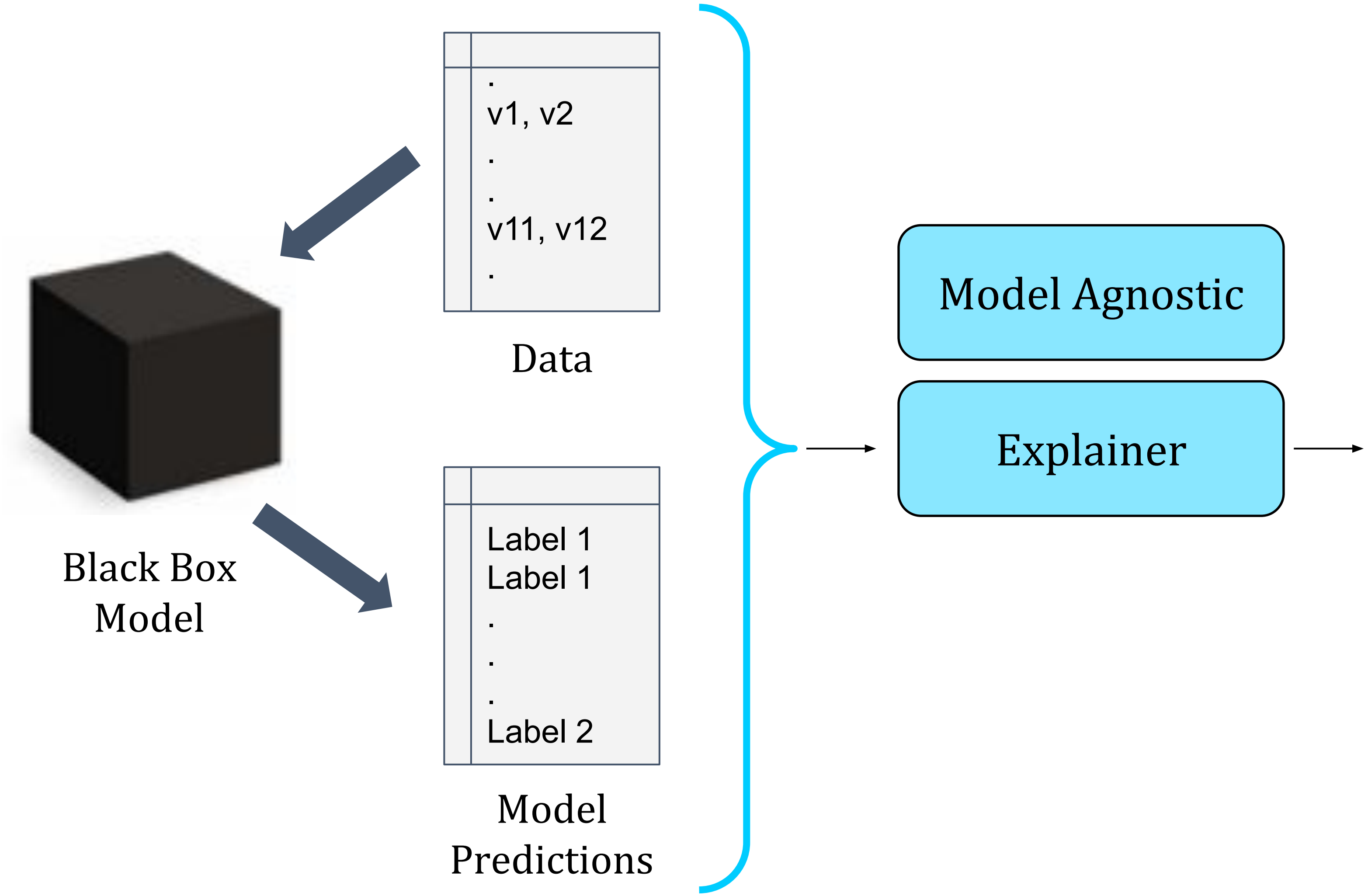
Fit this model to the predictions of the black box to obtain the shape functions.

[Bastani et. al., 2019]

Decision Trees as Global Explanations



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 $\geq$ 50 **and** Male =No:

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

Default:

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

[Lakkaraju et. al., 2019]

Customizable Decision Sets as Global Explanations

Subgroup Descriptor

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

Decision Logic

If Age $\geq$ 50 **and** Male $=$No:

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

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

Default:

If Past-Depression $=$Yes **and** Tiredness $=$No **and** Exercise $=$No **and** Insomnia $=$Yes, **then** Depression

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

[Lakkaraju et. al., 2019]

Customizable Decision Sets as Global Explanations

If Exercise $=$Yes **and** Smoking $=$No:

If Rapid-Weight-Gain $=$Yes **and** Tiredness $=$Yes **and** Melancholy $=$Yes **and** Insomnia $=$Yes **and** Age $<$50, **then** Depression

If Tiredness $=$Yes **and** Melancholy $=$Yes **and** Age $\geq$ 50, **then** Depression

If Tiredness $=$No **and** Melancholy $=$No, **then** Healthy

If Smoking $=$Yes:

If Rapid-Weight-Gain $=$Yes **and** Melancholy $=$Yes, **then** Depression

If Tiredness $=$No **and** Insomnia $=$No **and** Melancholy $=$No **and** Rapid-Weight-Gain $=$No, **then** Healthy

If Insomnia $=$Yes **and** Past-Depression $=$Yes **and** Tiredness $=$Yes, **then** Depression

Default:

If Past-Depression $=$Yes **and** Tiredness $=$Yes **and** Melancholy $=$Yes, **then** Depression

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

If Family-Depression $=$Yes **and** Age $\geq$ 50 **and** Male $=$No **and** Tiredness $=$Yes, **then** Depression

[Lakkaraju et. al., 2019]

Customizable Decision Sets as Global Explanations: Desiderata & Optimization Problem

Desiderata	Optimization Problem
Fidelity Describe model behavior accurately	**Fidelity** Minimize number of instances for which explanation's label $\neq$ model prediction
Unambiguity No contradicting explanations	**Unambiguity** Minimize the number of duplicate rules applicable to each instance
Simplicity Users should be able to look at the explanation and reason about model behavior	**Simplicity** Minimize the number of conditions in rules; Constraints on number of rules & subgroups;
Customizability Users should be able to understand model behavior across various subgroups of interest	**Customizability** Outer rules should only comprise of features of user interest (candidate set restricted)

Customizable Decision Sets as Global Explanations

- The complete optimization problem 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

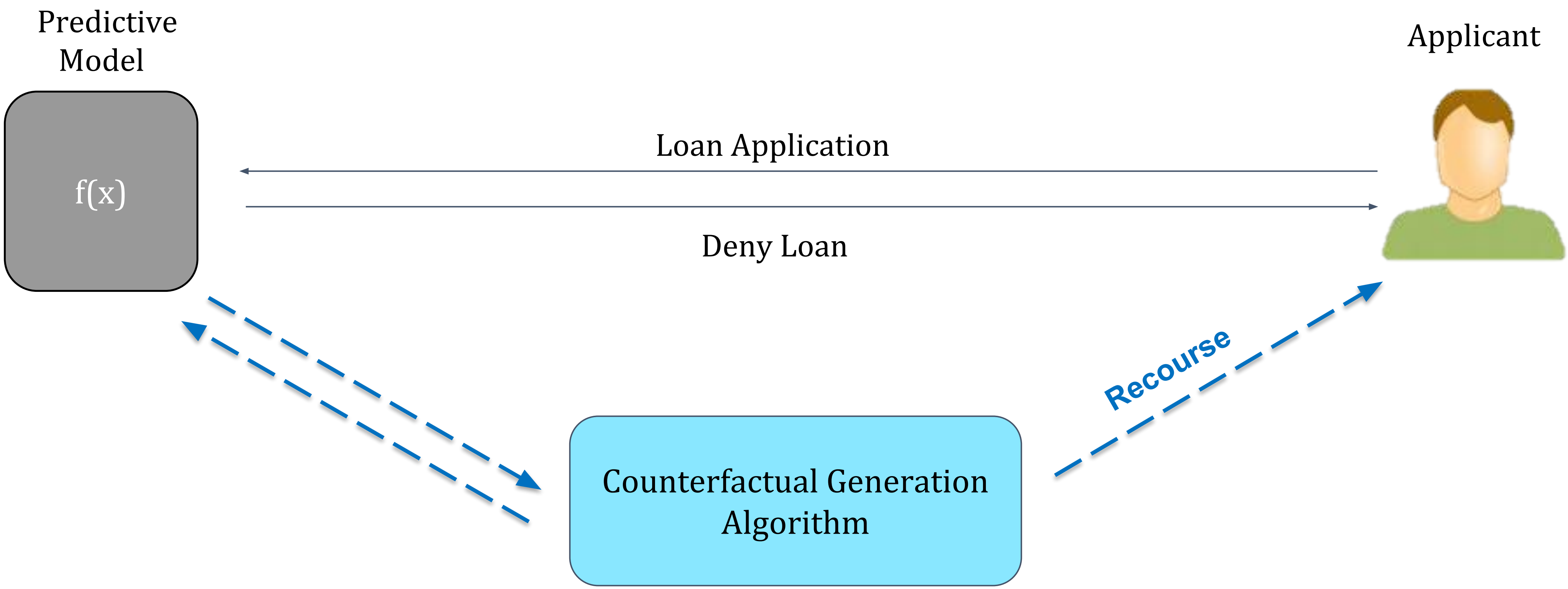
Local Explanations

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

Global Explanations

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

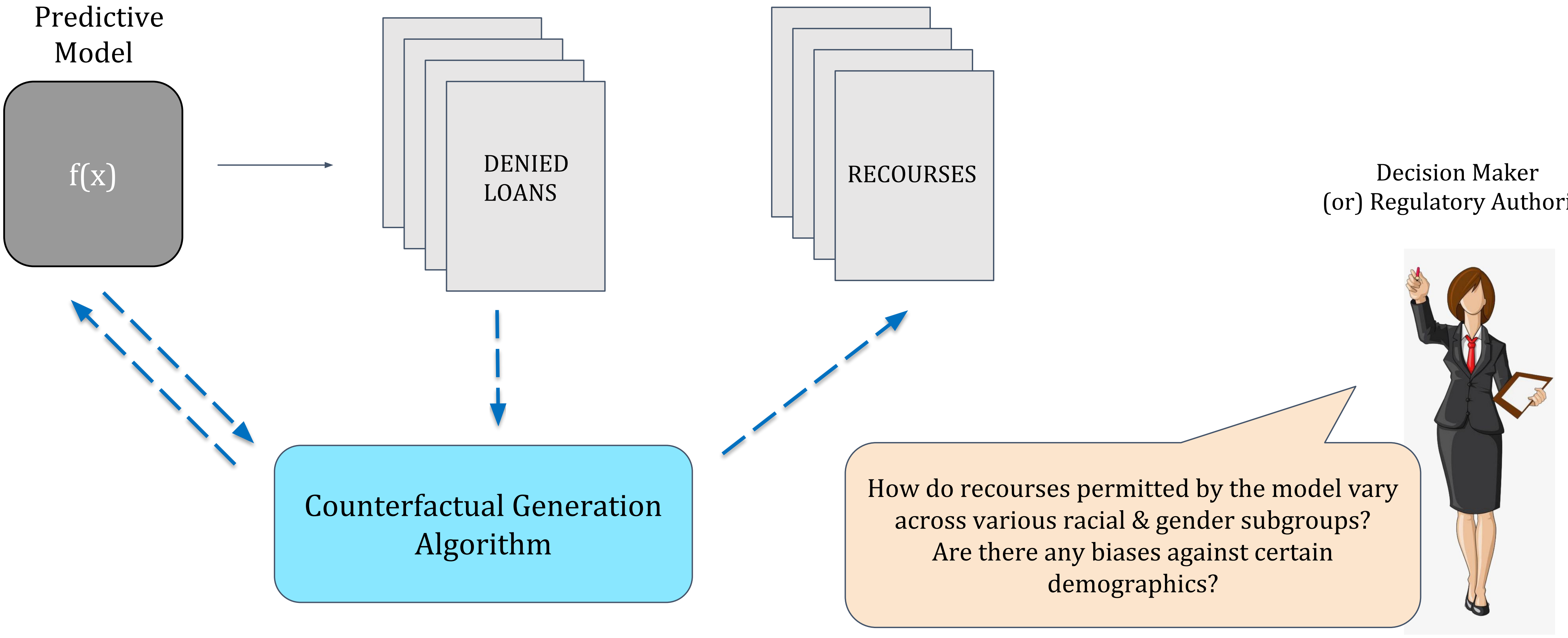
Counterfactual Explanations



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

[Rawal and Lakkaraju, 2020]

Counterfactual Explanations



[Rawal and Lakkaraju, 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

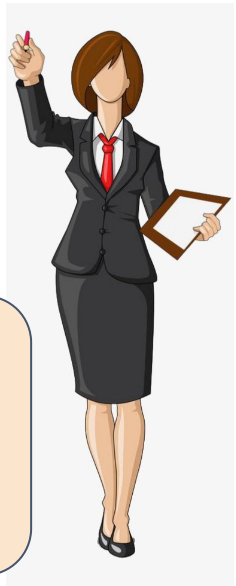
If Race = Caucasian **and** Gender = Female:

- **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 “act upon” 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

If Race ≠ 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

Recourse Rules

[Rawal and Lakkaraju, 2020]

Customizable Global Summaries of Counterfactuals: Desiderata & Optimization Problem

Desiderata	Optimization Problem
Recourse Correctness Prescribed recourses should obtain desirable outcomes	**Recourse Correctness** Minimize number of applicants for whom prescribed recourse does not lead to desired outcome
Recourse Coverage (Almost all) applicants should be provided with recourses	**Recourse Coverage** Minimize number of applicants for whom recourse does not exist (i.e., satisfy no rule).
Minimal Recourse Costs Acting upon a prescribed recourse should not be impractical or terribly expensive	**Minimal Recourse Costs** Minimize total ***feature costs*** as well as ***magnitude of changes*** in feature values
Interpretability of Summaries Summaries should be readily understandable to stakeholders (e.g., decision makers/regulatory authorities).	**Interpretability of Summaries** Constraints on # of rules, # of conditions in rules & # of subgroups
Customizability Stakeholders should be able to understand model behavior across various subgroups of interest	**Customizability** Outer rules should only comprise of features of stakeholder interest (candidate set restricted)

Customizable Global Summaries of Counterfactuals: Feature Costs & Magnitude of Changes

- Feature Costs: *Each feature is associated with a cost which indicates how hard it is change that feature*.

- How to obtain feature costs?
    - Obtain pairwise feature comparison inputs from domain experts
    - Apply Bradley Terry model which connects pairwise feature comparisons to individual feature costs and estimate these costs.

- Magnitude of Changes: are penalized via total log percentile shift

[Rawal and Lakkaraju, 2020]

Customizable Global Summaries of Counterfactuals: Feature Costs & Magnitude of Changes

- Feature Costs: *Each feature is associated with a cost which indicates how hard it is change that feature*.

- How to obtain feature costs?
  - Obtain pairwise feature comparison inputs from domain experts
  - Apply Bradley Terry model which connects pairwise feature comparisons to individual feature costs and estimate these costs.

$$p_{ij} = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}}$$

- Magnitude of Changes: are penalized via total log percentile shift

Customizable Global Summaries of Counterfactuals

- The complete optimization problem 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.

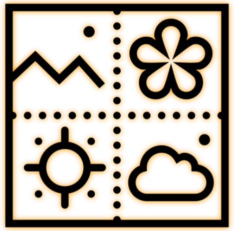
Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Explanations in Different Modalities
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Explanations in Different Modalities
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Post hoc Explanations in Different Modalities



Different Data Modalities

Structured Data

Computer Vision

Natural Language

Different Data Modalities

Structured Data

Computer Vision

Natural Language

Structured Data

NoDefaultN	Married	Single	Age_lt_25	Age_in_25_t	Age_in_40_t	Age_geq_6	EducationLev	MaxBillAmo	MaxPayment	MonthsWith	MonthsWith	MonthsWith	MostRecentB	MostRecentP	TotalOverdu	TotalMonths	HistoryOfOverduePayments
0	1	0	1	0	0	0	2	120	20	0	6	0	120	0	1	4	1
0	0	1	0	1	0	0	2	110	60	0	6	0	80	0	2	4	1
1	0	1	0	1	0	0	2	890	155.32	0	5	0	890	50	0	0	0
1	1	0	0	1	0	0	2	1510	60	0	0	3	1430	60	0	0	0
1	1	0	0	0	1	0	2	1090	1120	1	2	0	260	60	0	0	0
1	0	1	0	1	0	0	3	1970	80	0	0	3	1970	80	0	0	0
1	0	1	0	1	0	0	3	16570	1680	0	0	5	11240	1680	0	0	0
1	0	1	1	0	0	0	2	360	50	4	6	0	360	10	0	0	0
1	1	0	0	1	0	0	1	430	100	0	6	0	340	100	1	2	1
1	0	1	0	1	0	0	1	420	400	1	4	0	0	0	0	0	0
1	0	1	0	1	0	0	1	340	110	1	6	0	340	70	1	2	1
1	0	1	0	0	1	0	3	680	680	2	6	0	370	670	1	2	1
1	0	1	0	0	1	0	2	370	200	0	6	0	370	30	0	0	0
0	0	1	0	1	0	0	2	2060	100	0	0	4	2010	100	2	7	1
1	0	1	0	1	0	0	3	2160	90	0	0	0	2160	90	0	0	0
1	0	0	1	0	0	0	1	1550	50	0	0	1	1550	0	1	3	1
0	0	1	1	0	0	0	3	580	100	0	0	5	470	100	1	8	1
1	1	0	0	0	1	0	3	7730	5970	1	1	0	7730	320	0	0	0
1	1	0	0	0	1	0	3	0	0	0	6	0	0	0	1	1	1
1	0	1	0	1	0	0	3	0	0	0	6	0	0	0	1	1	1
1	0	1	0	1	0	0	1	1170	1030	1	4	0	1170	90	0	0	0

Categorical Data

Ordinal Data

Numerical Data
(Discrete & Continuous)

Structured Data: Why care about explainability?

- Lot of information in various real world settings available as structured data

- Lots of applications deal with structured data
  - Disease diagnosis and treatment (e.g., weight, age, symptoms, glucose level)
  - Risk prediction in education/lending/criminal justice (e.g., credit scores, previous crimes, student GPAs, education level)
  - Recommender systems for movies/products (e.g., list of movies liked in the past)

Challenges for Structured Data

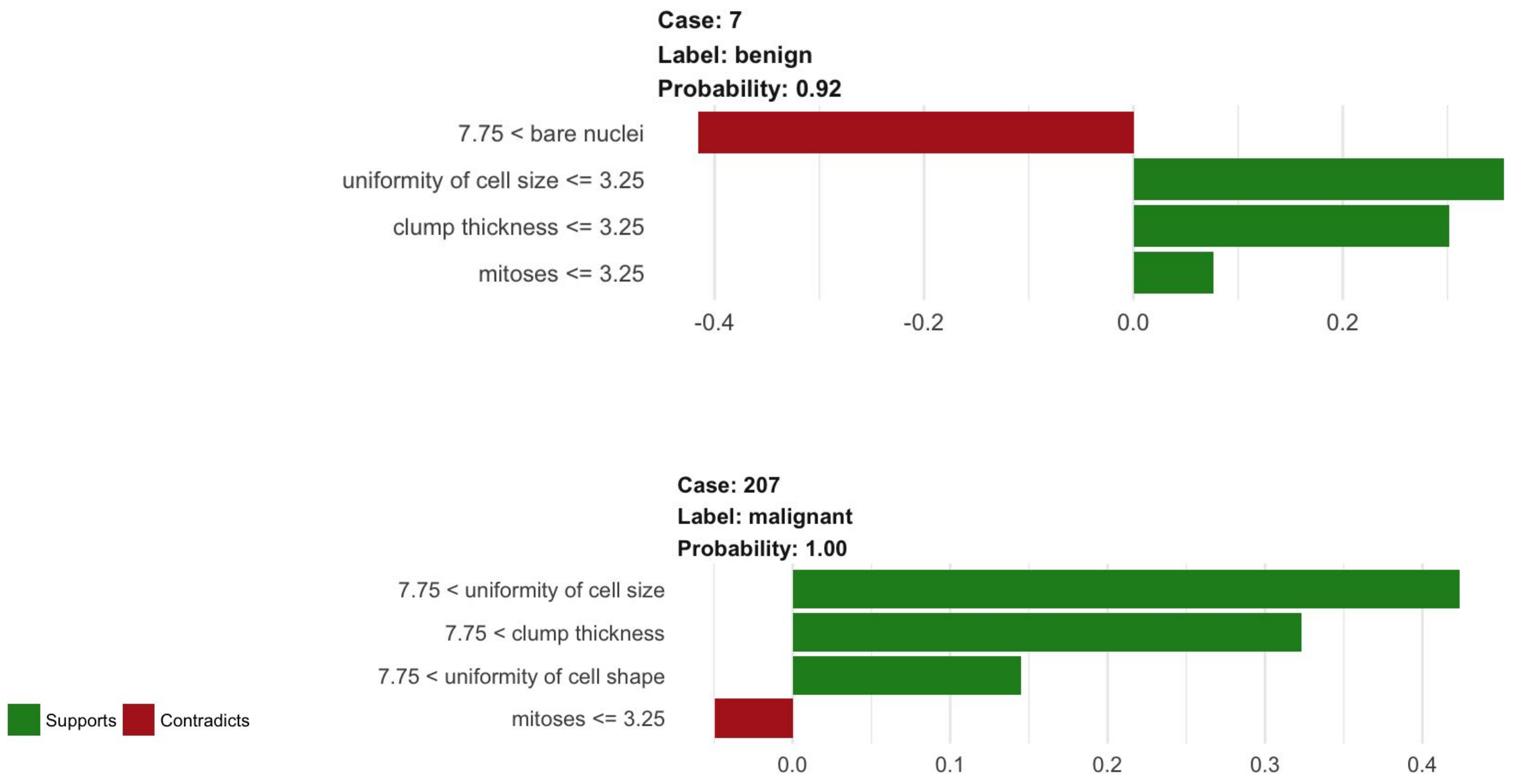
- Different types of variables in the data
  - Different types of variables call for different similarity/perturbation functions
  - gradients may not always be meaningful

- Depending on the task/domain, data could be either low or high dimensional
  - E.g., movie recommendations -- user x movie matrix (high dimensional)

Structured Data: Explainability Techniques

- Feature importance based explanations
  - Perturbation methods e.g., LIME/SHAP
  - Saliency maps and other gradient based methods not very meaningful

Feature Importance Based Explanations



Structured Data: Explainability Techniques

- Feature importance based explanations
  - Perturbation methods e.g., LIME/SHAP
  - Saliency maps and other gradient based methods not very meaningful

- Prototype/example based explanations
  - might not always be interpretable
  - e.g., an instance with 100 feature values as prototype

[Koh and Liang, 2017]

Prototype Based Explanations

Prediction: Not Diabetic

Influential instances driving the prediction:

Instance #	Age	Weight	Smoking	Exercise	Prediction
1	32	153	No	Yes	Not Diabetic
2	27	172	Yes	Yes	Not Diabetic
3	55	163	No	Yes	Not Diabetic
4	18	147	No	No	Not Diabetic

Structured Data: Explainability Techniques

- Feature importance based explanations
  - Perturbation methods e.g., LIME/SHAP
  - Saliency maps and other gradient based methods not very meaningful
- Prototype/example based explanations
  - might not always be interpretable
  - e.g., an instance with 100 feature values as prototype
- Rule based explanations

[Letham et. al., 2015; Lakkaraju et. al. 2016]

Rule Based Explanations

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age$\geq$ 50 **then** Lung Cancer

If Risk-LungCancer=Yes **and** Blood-Pressure$\geq$ 0.3 **then** Lung Cancer

If Risk-Depression=Yes **and** Past-Depression=Yes **then** Depression

If BMI$\geq$ 0.3 **and** Insurance=None **and** Blood-Pressure$\geq$ 0.2 **then** Depression

If Smoker=Yes **and** BMI$\geq$ 0.2 **and** Age$\geq$ 60 **then** Diabetes

If Risk-Diabetes=Yes **and** BMI$\geq$ 0.4 **and** Prob-Infections$\geq$ 0.2 **then** Diabetes

If Doctor-Visits $\geq$ 0.4 **and** Childhood-Obesity=Yes **then** Diabetes

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age$\geq$ 50 **then** Lung Cancer

Else if Risk-Depression=Yes **then** Depression

Else if BMI $\geq$ 0.2 **and** Age$\geq$ 60 **then** Diabetes

Else if Headaches=Yes **and** Dizziness=Yes, **then** Depression

Else if Doctor-Visits$\geq$ 0.3 **then** Diabetes

Else if Disposition-Tiredness=Yes **then** Depression

Else Diabetes

Structured Data: Explainability Techniques

- Feature importance based explanations
  - Perturbation methods e.g., LIME/SHAP
  - Saliency maps and other gradient based methods not very meaningful
- Prototype/example based explanations
  - might not always be interpretable
  - e.g., an instance with 100 feature values as prototype
- Rule based explanations
- Counterfactual explanations

[Ustun et. al., 2019; Rawal and Lakkaraju, 2020]

Counterfactual Explanations

Features to Change	Current Values		Required Values
MostRecentPaymentAmount	\$0	⟶	\$500
MonthsWithLowSpendingOverLast6Months	6	⟶	5
MonthsWithZeroBalanceOverLast6Months	1	⟶	2

If Female =No **and** Foreign Worker =No:

If Missed Payments =Yes **and** Critical Loans =Yes, **then** Missed Payments =Yes **and** Critical Loans =No

If Unemployed =Yes **and** Critical Loans =Yes **and** Has Guarantor =No,

then Unemployed =Yes **and** Critical Loans =No **and** Has Guarantor =Yes

If Female =No **and** Foreign Worker =Yes:

If Skilled Job =No **and** Years at Job ≤ 1, **then** Skilled Job =Yes **and** Years at Job ≥ 4

If Unemployed =Yes **and** Has Guarantor =No **and** Has CoAppplicant =No,

then Unemployed =No **and** Has Guarantor =Yes **and** Has CoAppplicant =Yes

If Female =Yes:

If Married =No **and** Owns House =No, **then** Married =Yes **and** Owns House =Yes

If Unemployed =No **and** Has Guarantor =Yes **and** Has CoAppplicant =No,

then Unemployed =No **and** Has Guarantor =Yes **and** Has CoAppplicant =Yes

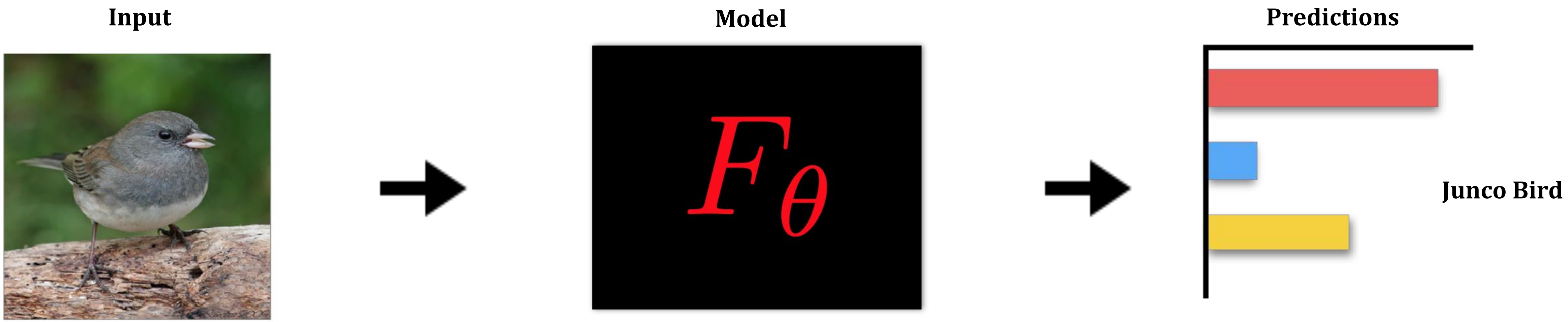
Different Data Modalities

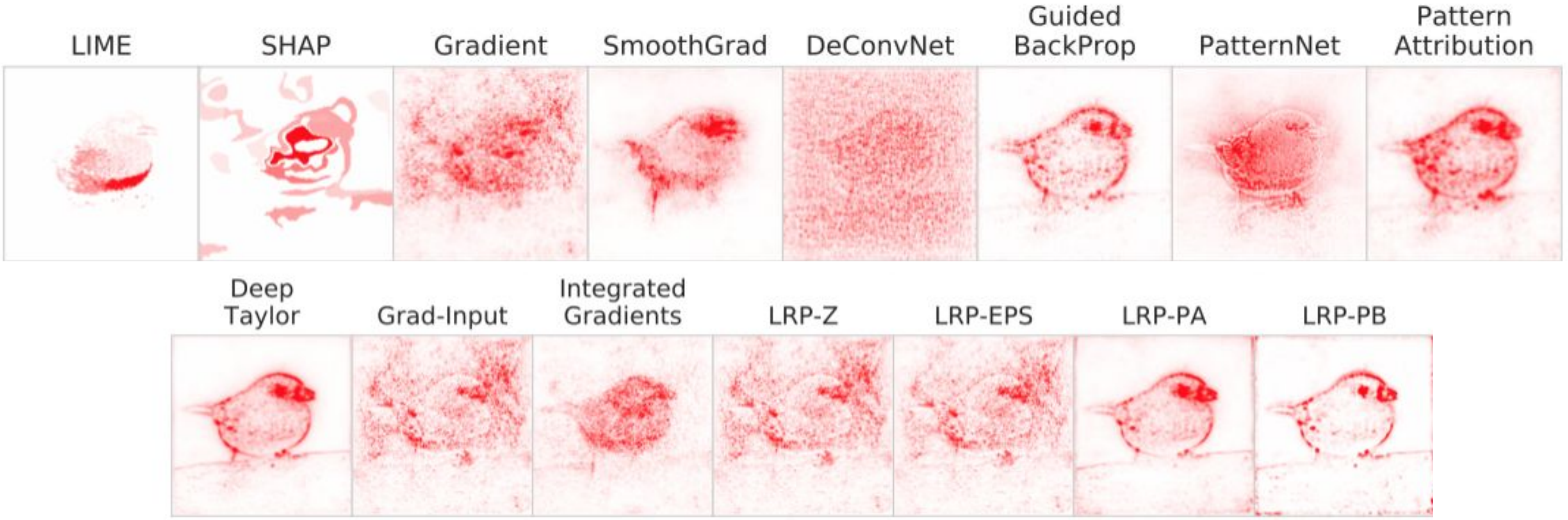
Structured Data

Computer Vision

Natural Language

Feature Importance Approaches on VGG-16





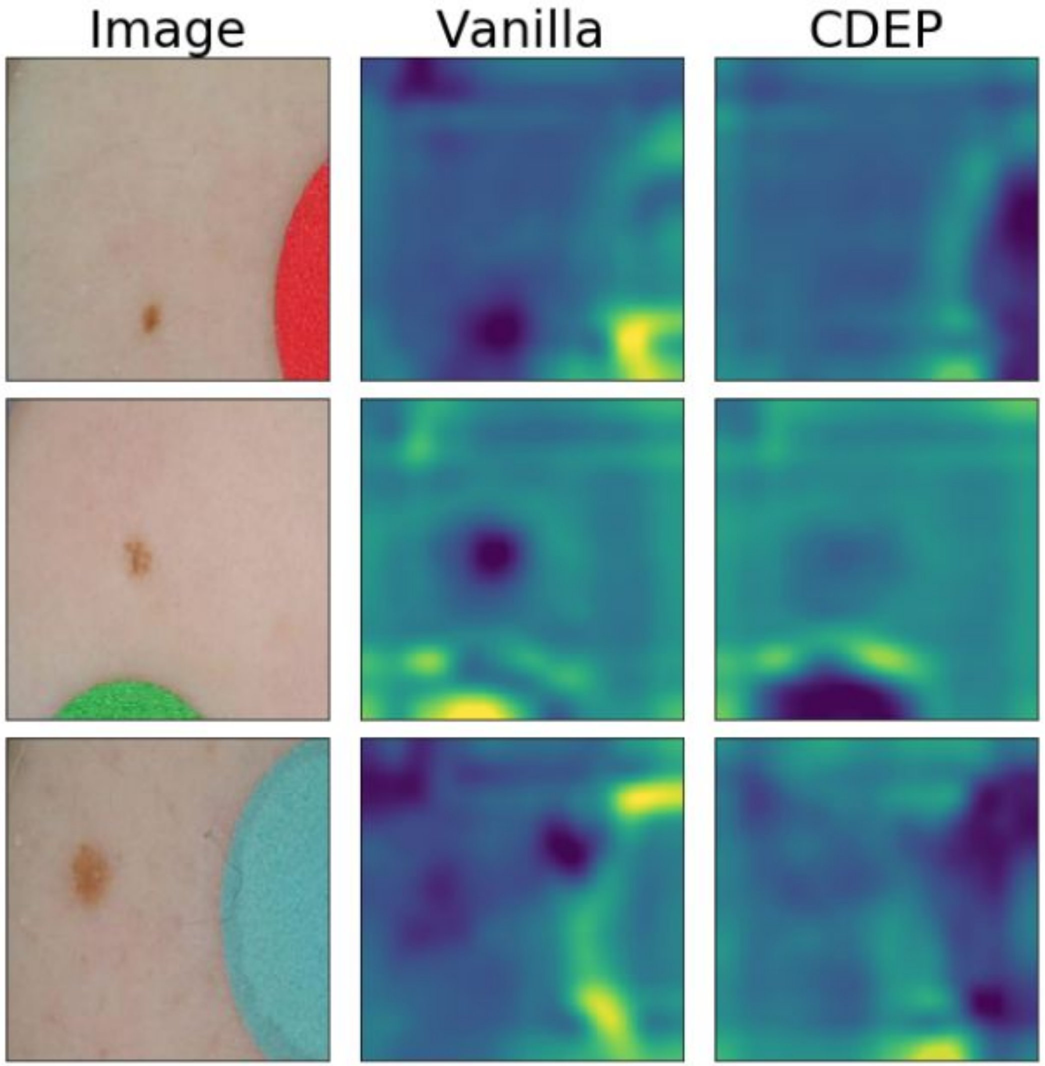
Neuron Shapley Importance for Inception-V3 Trained on ImageNet

Images from [Ghorbani et. al. at Neurips 2020]

Saliency Map for Bone Age Model

Input		Model		Predictions
	→		→	Bone Age

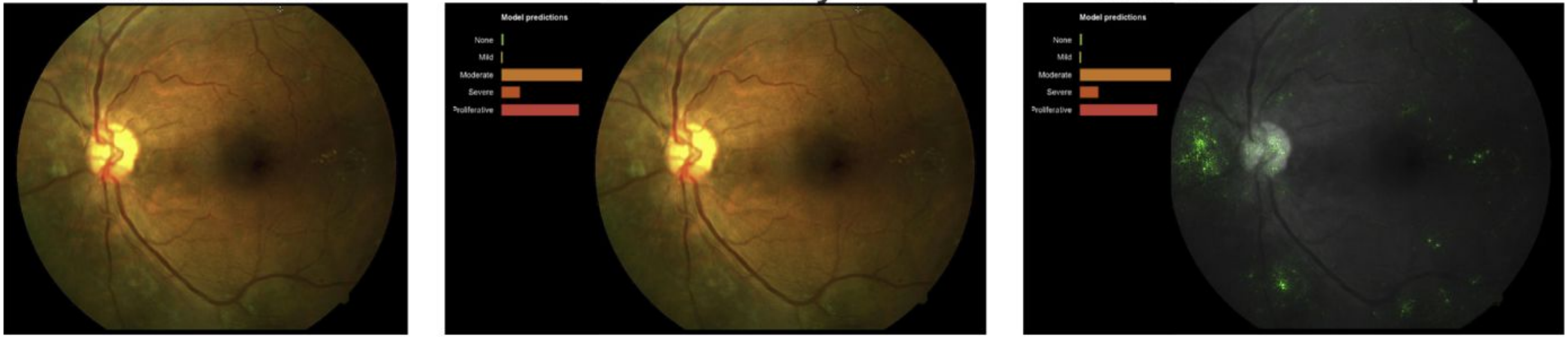
Contextual Decomposition for a Skin Cancer Prediction Model



Images from [Reiger et. al. 2020]

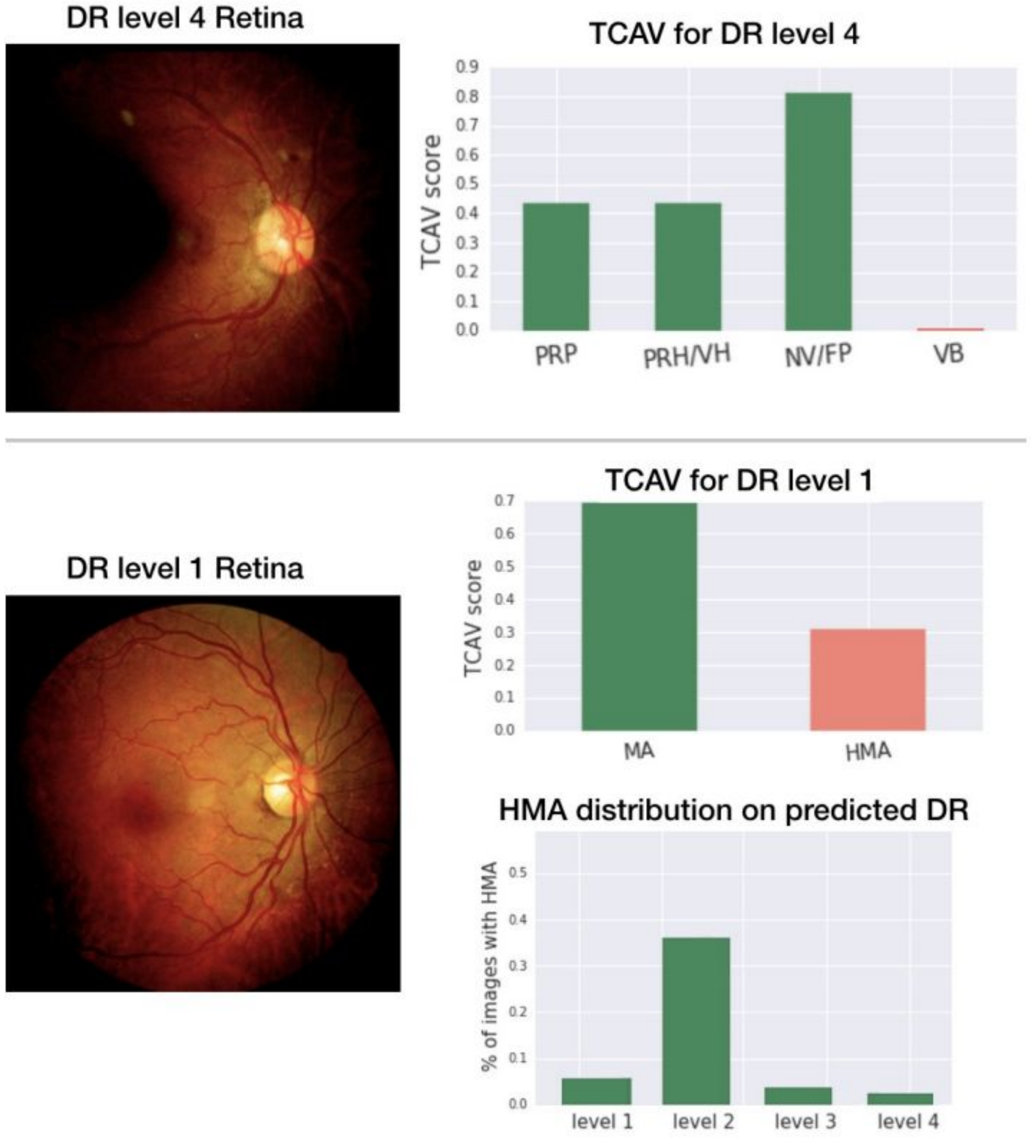
Integrated Gradients for Diabetic Retinopathy Model

Unassisted | Grades Only | Grades + Heatmap



Images from [Sayres et. al. 2019]

TCAV for Diabetic Retinopathy Model

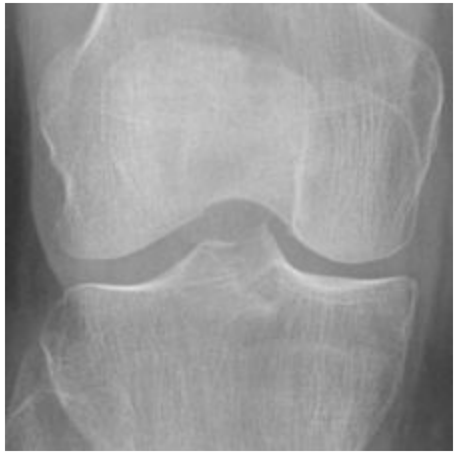


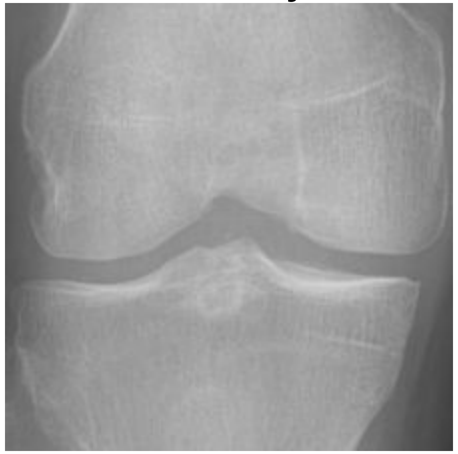
Images from [Kim et. al. 2018]

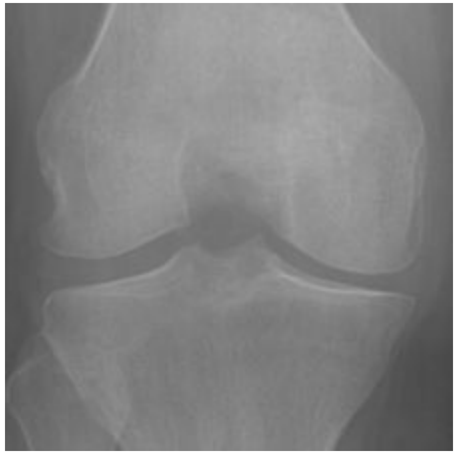
Challenges Transfering Approaches to Medical Setting

Adapting explanation methods developed for benchmark tasks like ImageNet and CIFAR to medical imaging setting is challenging in practice due to input **homogeneity**.

Knee Xray







Different Data Modalities

Structured Data

Computer Vision

Natural Language

Natural Language Processing

- Why should we care about interpretability for NLP?
  - Lots of NLP applications everywhere
    - Translation, Social Media Analysis, Hate Speech Filtering, Digital Assistants, …
  - Quickly evolving, in major ways, last few years
    - Word Embeddings, ELMo, BERT, GPT-2/3, T5, ...
  - Gap between what the benchmarks show and how good they are is **vast**
    - Lots of question answering, classification, textual entailment, etc. are "solved"
  - Brings up unique and additional challenges (that are more general)
    - Domains with discrete/structured/combinatorial inputs...

Challenges for NLP

- Discrete space of inputs
  - E.g. gradients are not directly applicable (or as meaningful)
- Not all combinations are well defined
  - They need not to be nonsense, ungrammatical
- Difficult to write a similarity/perturbation functions
- Format is not fixed: not everything is classification
  - structured prediction, text generation, span selection, ...
- Language does not lend itself to “simple explanations”

Word Attribution for NLP

Sentiment — an intelligent fiction about learning through cultural clash.

QA — What company won free advertisement due to QuickBooks contest ?

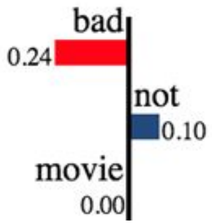
MLM — [CLS] The [MASK] ran to the emergency room to see her patient . [SEP]

Perturbation-based Explanations for NLP

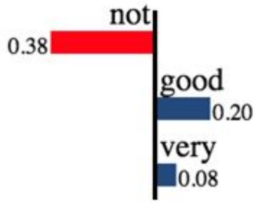
LIME

Anchors

\+ This movie is not bad.



− This movie is not very good.



This movie is **not bad**

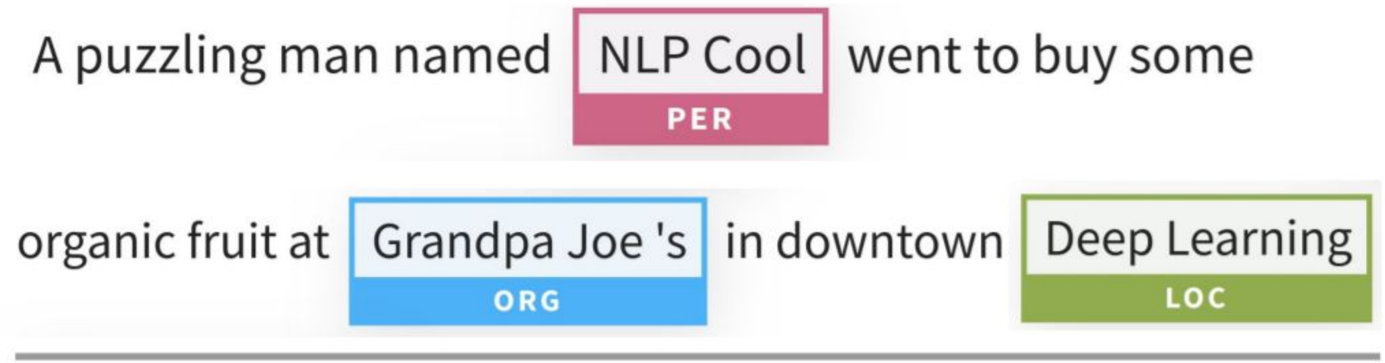
This audio is not bad

This novel is not bad

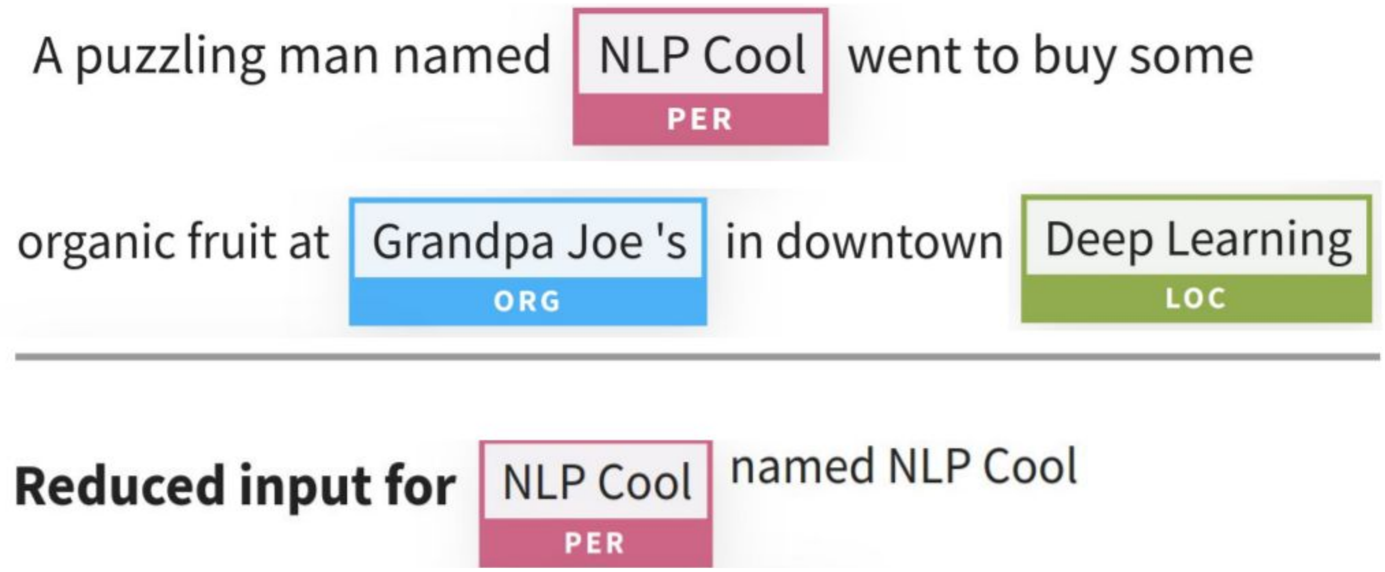
This footage is not bad

What is the mustache made of?	banana

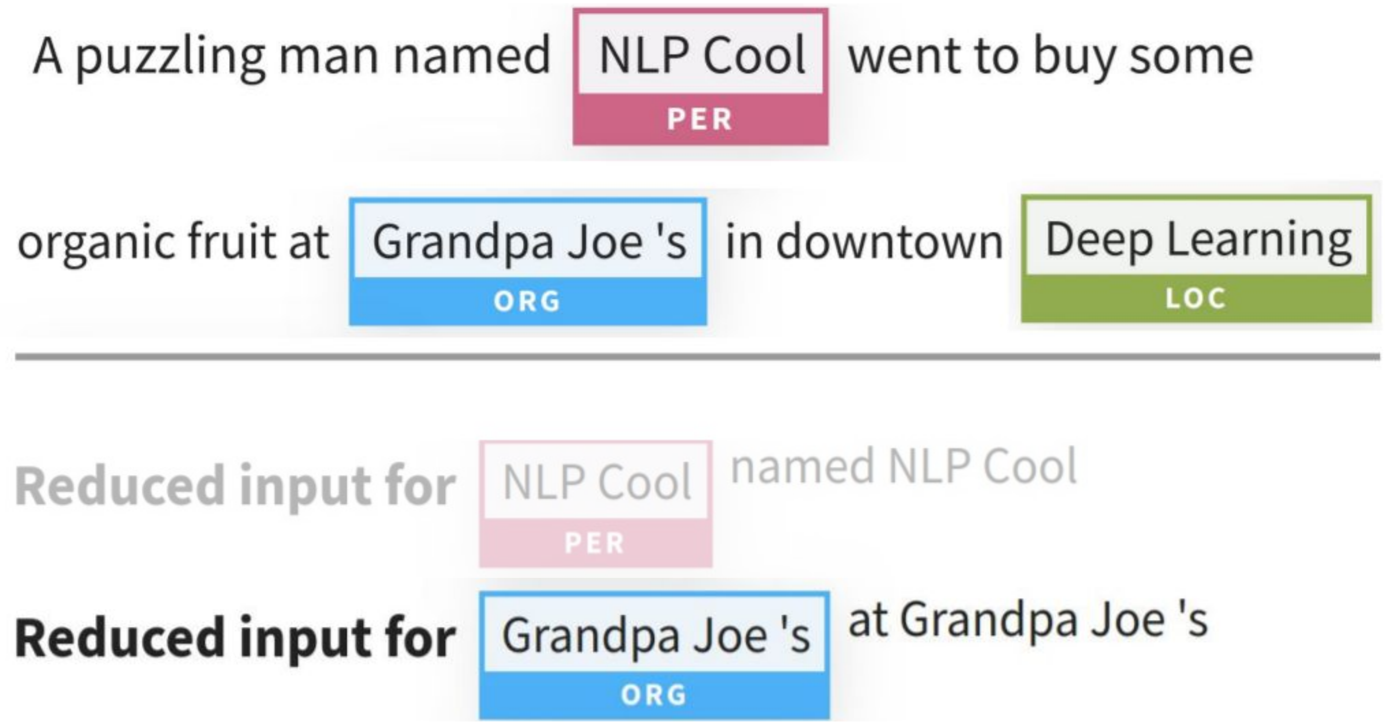
Input Reduction



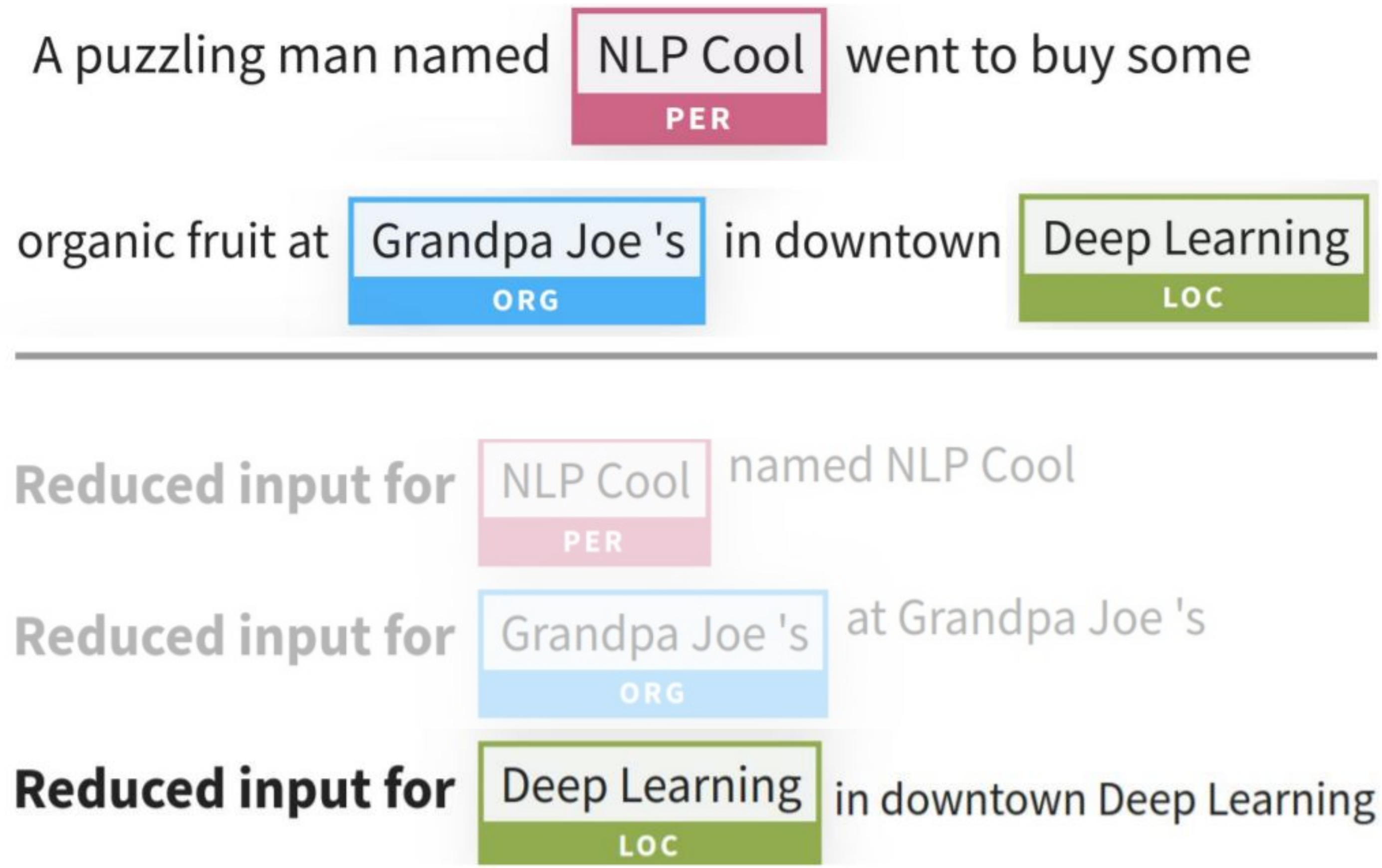
Input Reduction



Input Reduction



Input Reduction



[Han et al. 2020]

Prototypes for NLP

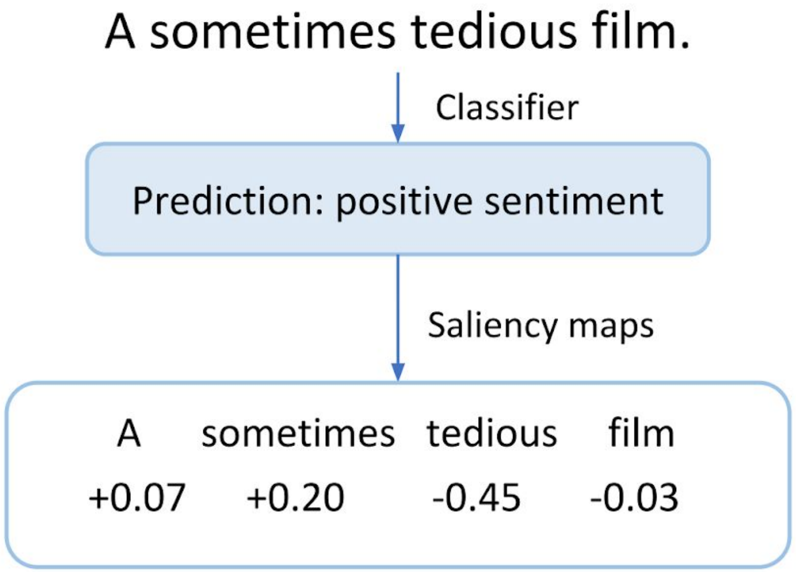
A sometimes tedious film.

↓ Classifier

Prediction: positive sentiment

[Han et al. 2020]

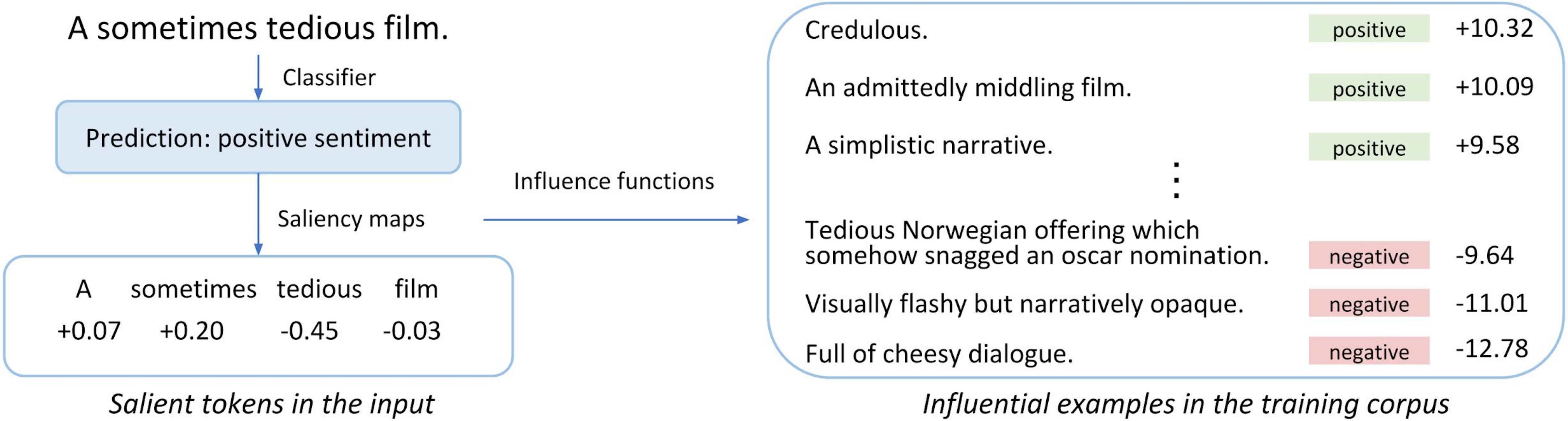
Prototypes for NLP



Salient tokens in the input

[Han et al. 2020]

Prototypes for NLP



Useful Implementations

Lots of code available (in no particular order):

- https://captum.ai/tutorials/Bert_SQUAD_Interpret
- https://github.com/PAIR-code/lit
- https://allennlp.org/interpret
- https://github.com/QData/TextAttack
- https://github.com/interpretml/interpret-text
- Influence functions for text
- Triggers Code
- Anchors Code
- LIME Code

Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Explanations in Different Modalities
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Explanations in Different Modalities
- 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?

[Lage et al, 2019]

Two Different Factors

What are you evaluating?

How we evaluate it?

	Understand the Behavior	Useful for Debugging	Help make decisions
Application-grounded			
Human-grounded			
Functionally-grounded			

Evaluating Post hoc Explanations

Understand the Behavior

Help make decisions

Useful for Debugging

Evaluating Post hoc Explanations

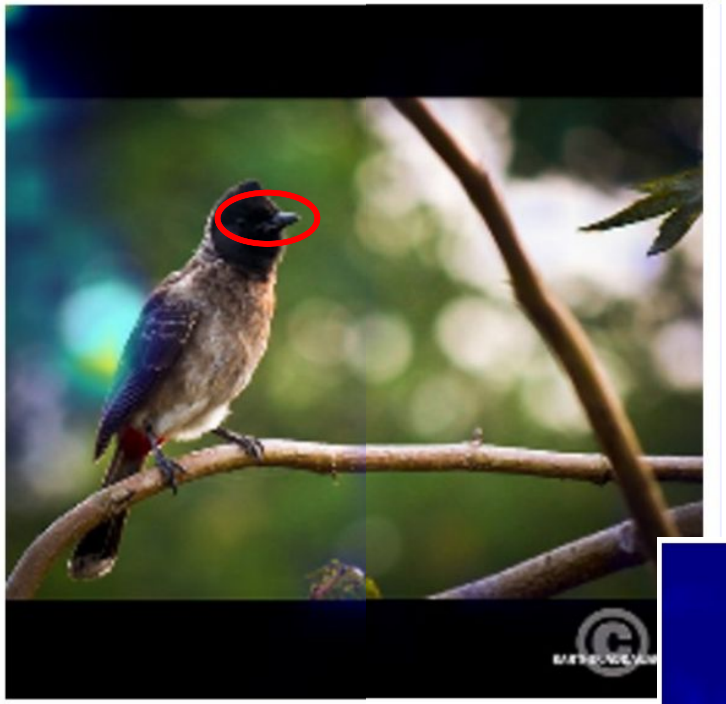
Understand the Behavior

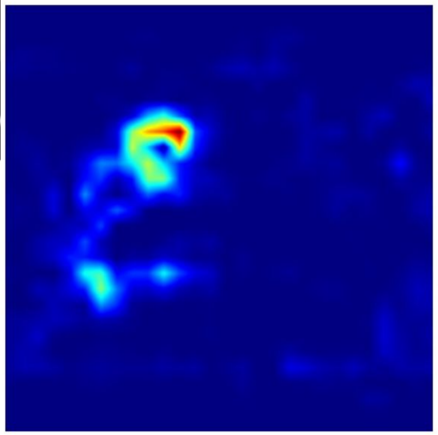
Help make decisions

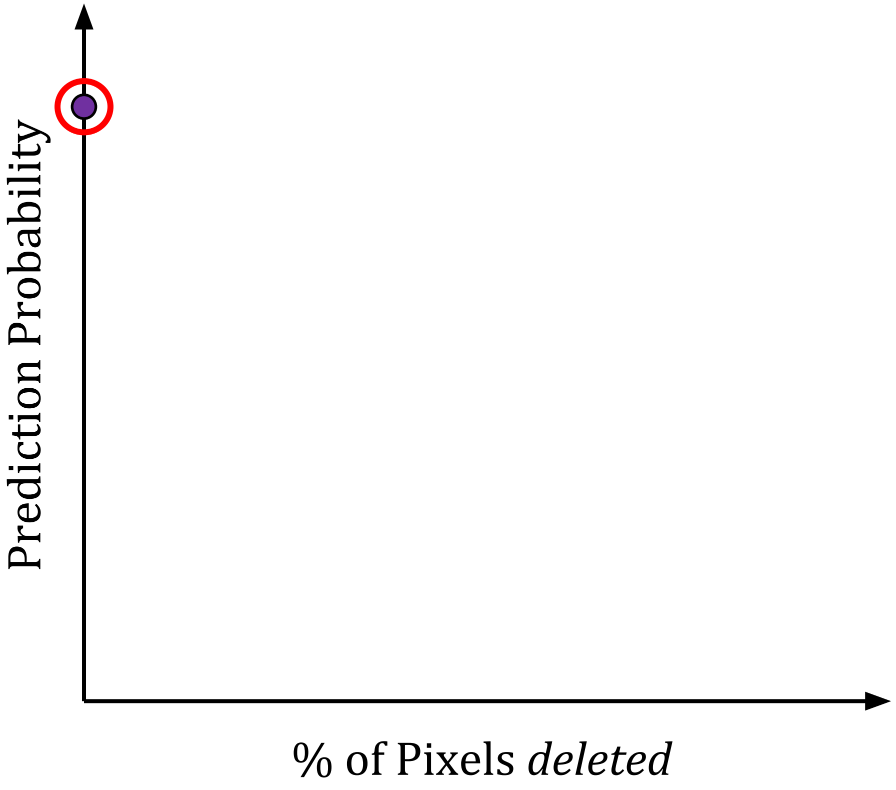
Useful for Debugging

How important are selected features?

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

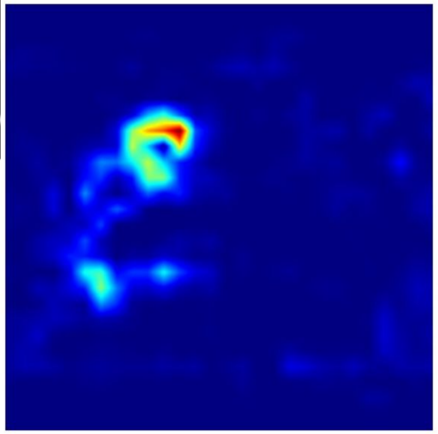


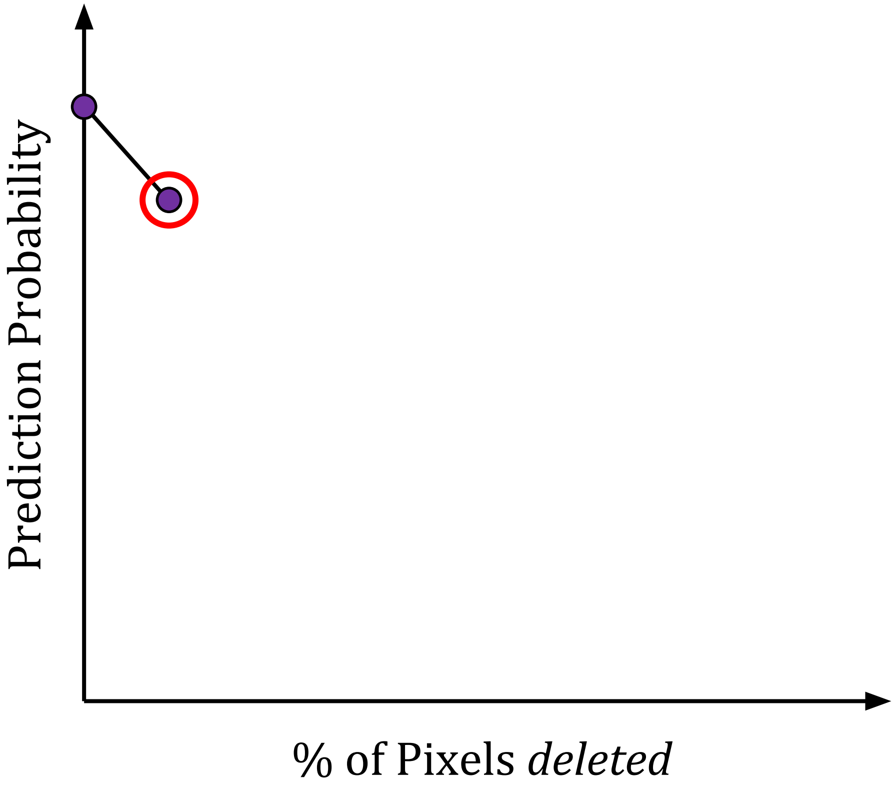




How important are selected features?

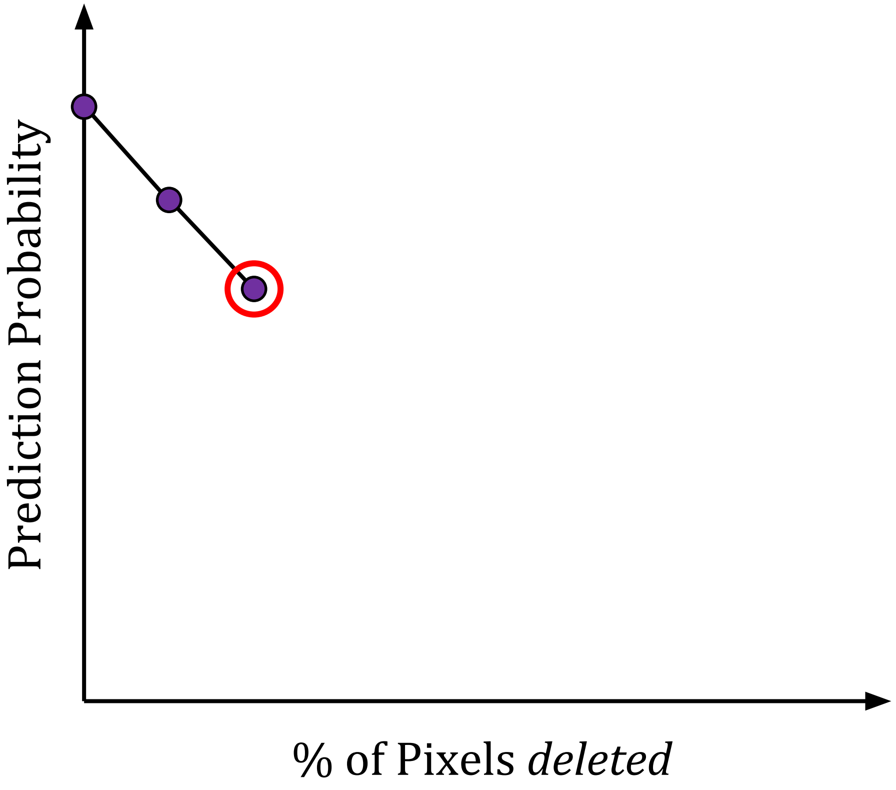
- Deletion: remove important features and see what happens..





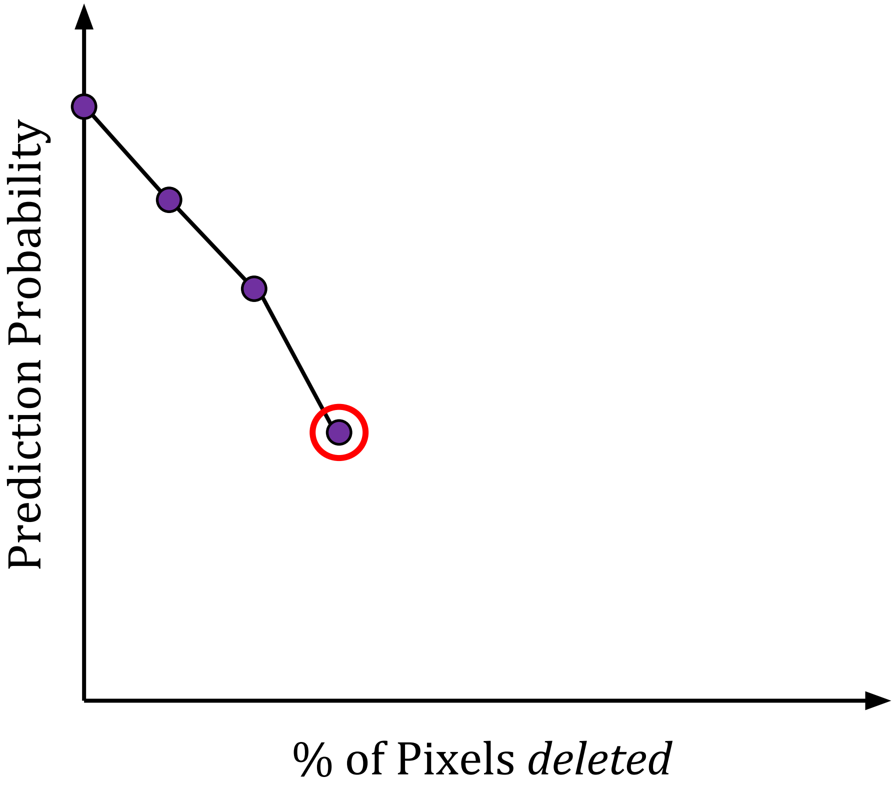
How important are selected features?

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



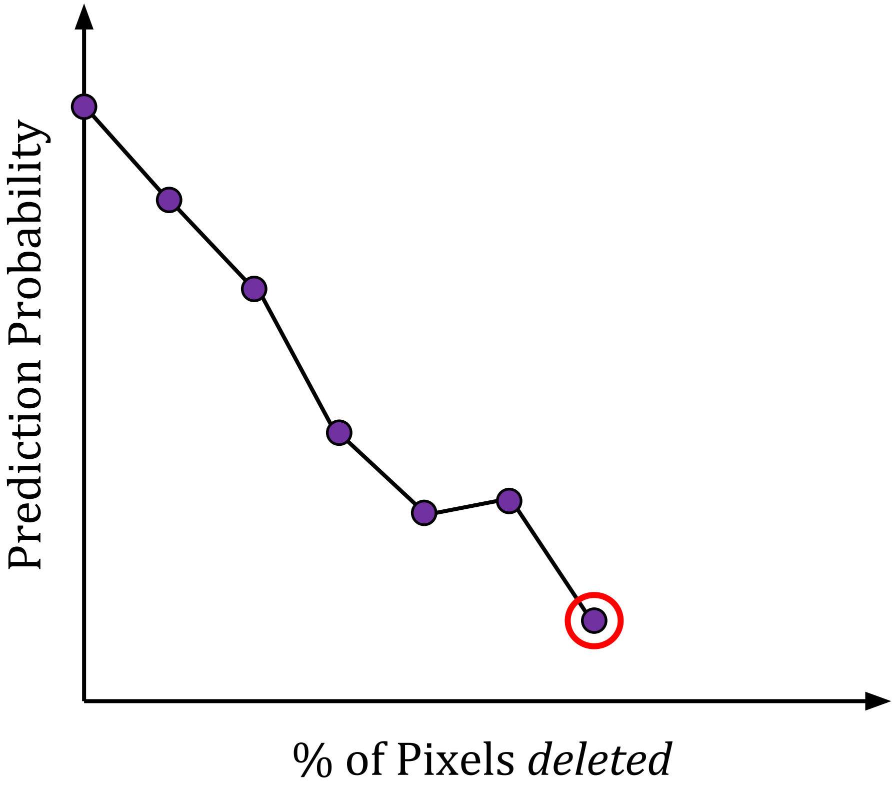
How important are selected features?

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



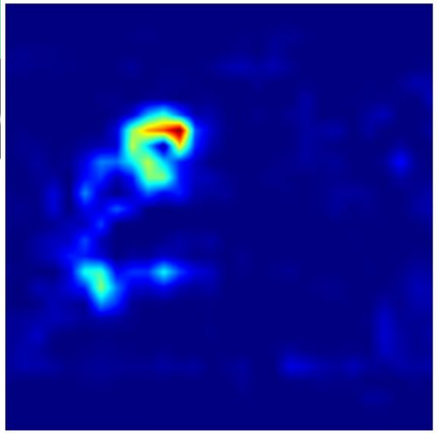
How important are selected features?

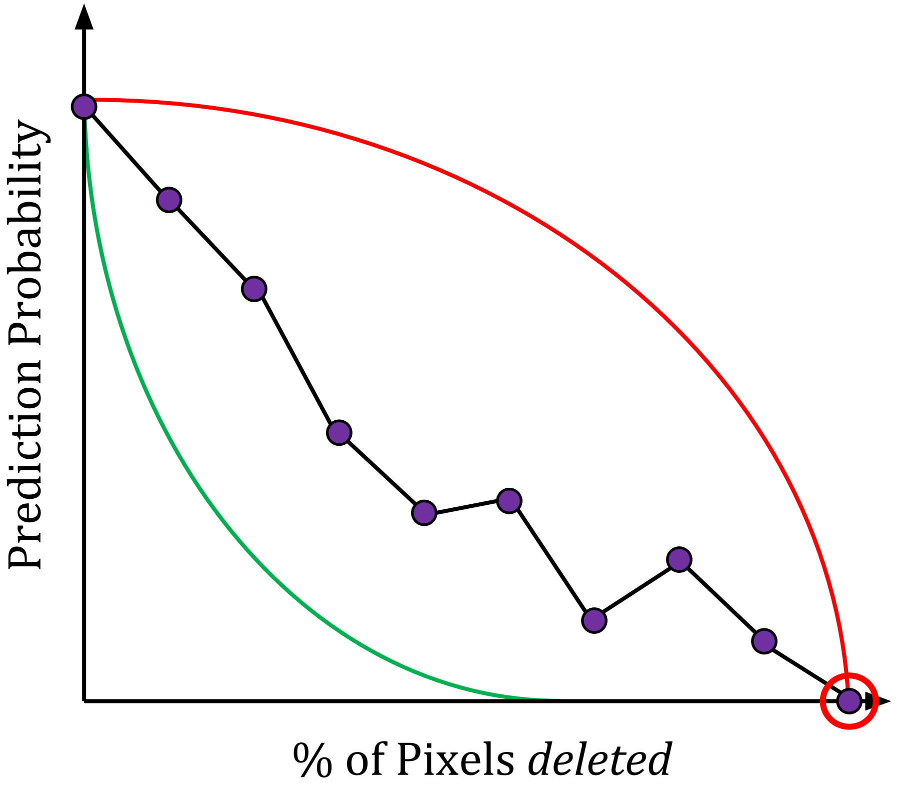
- Deletion: remove important features and see what happens..



How important are selected features?

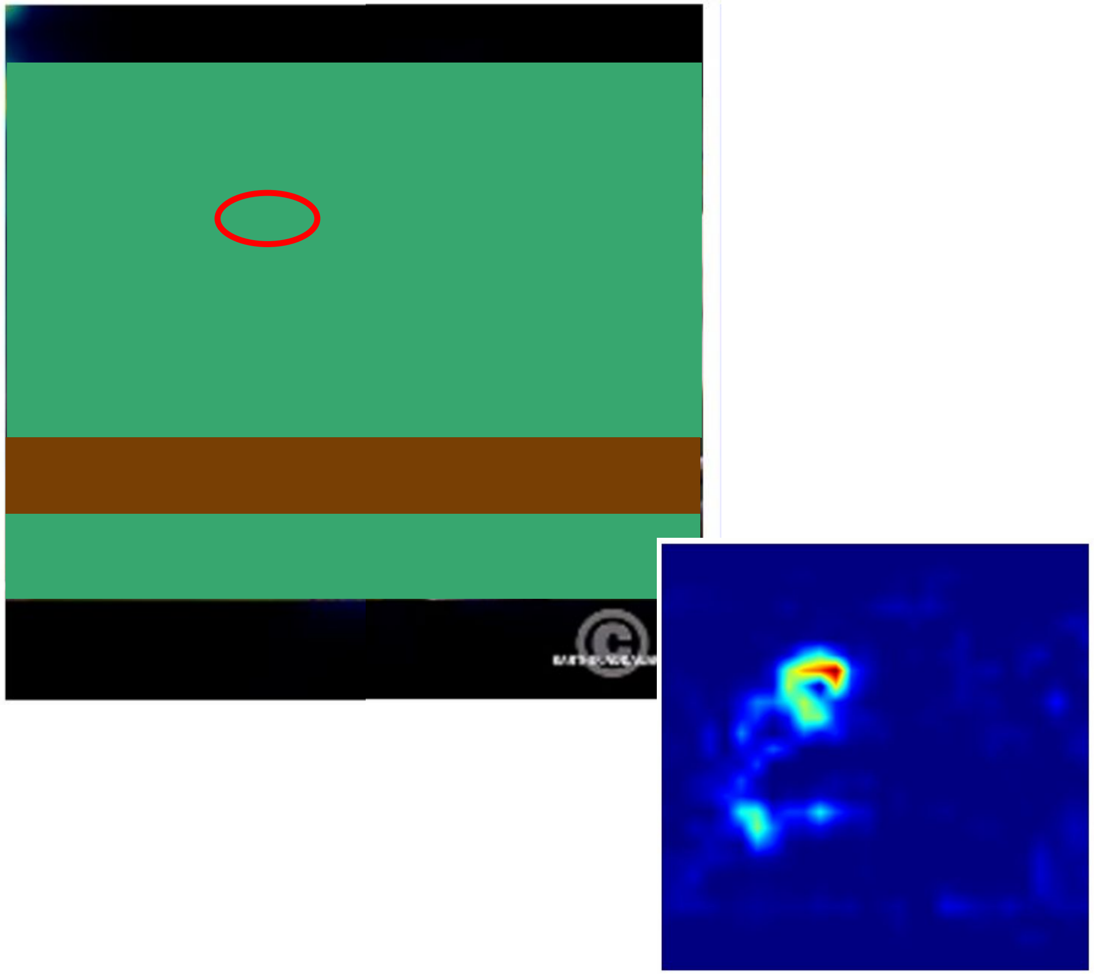
- Deletion: remove important features and see what happens..

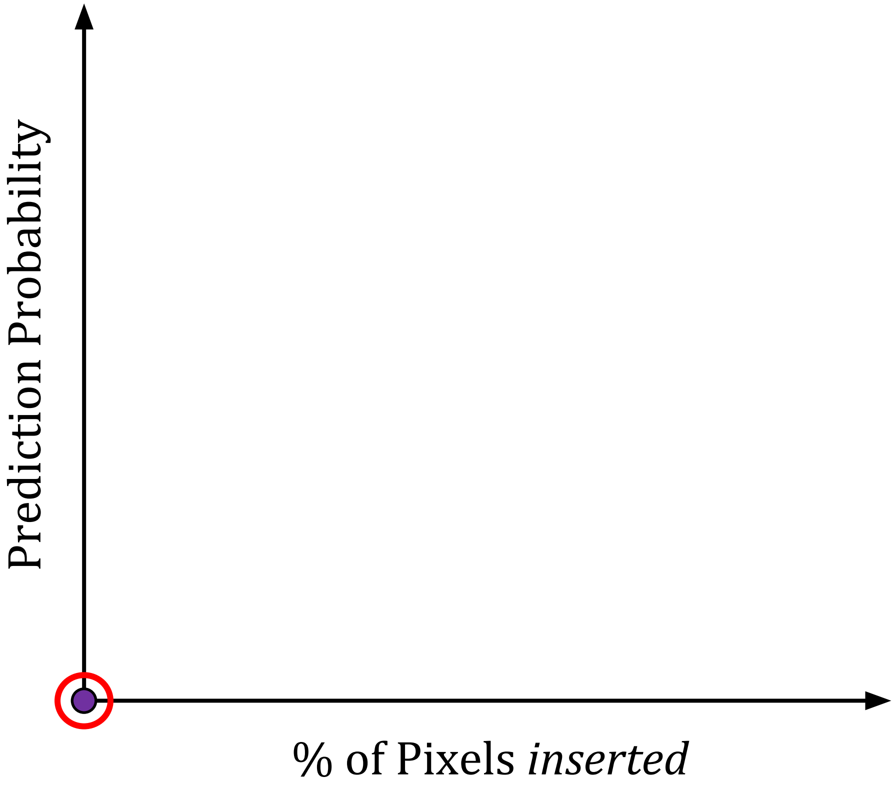




How important are selected features?

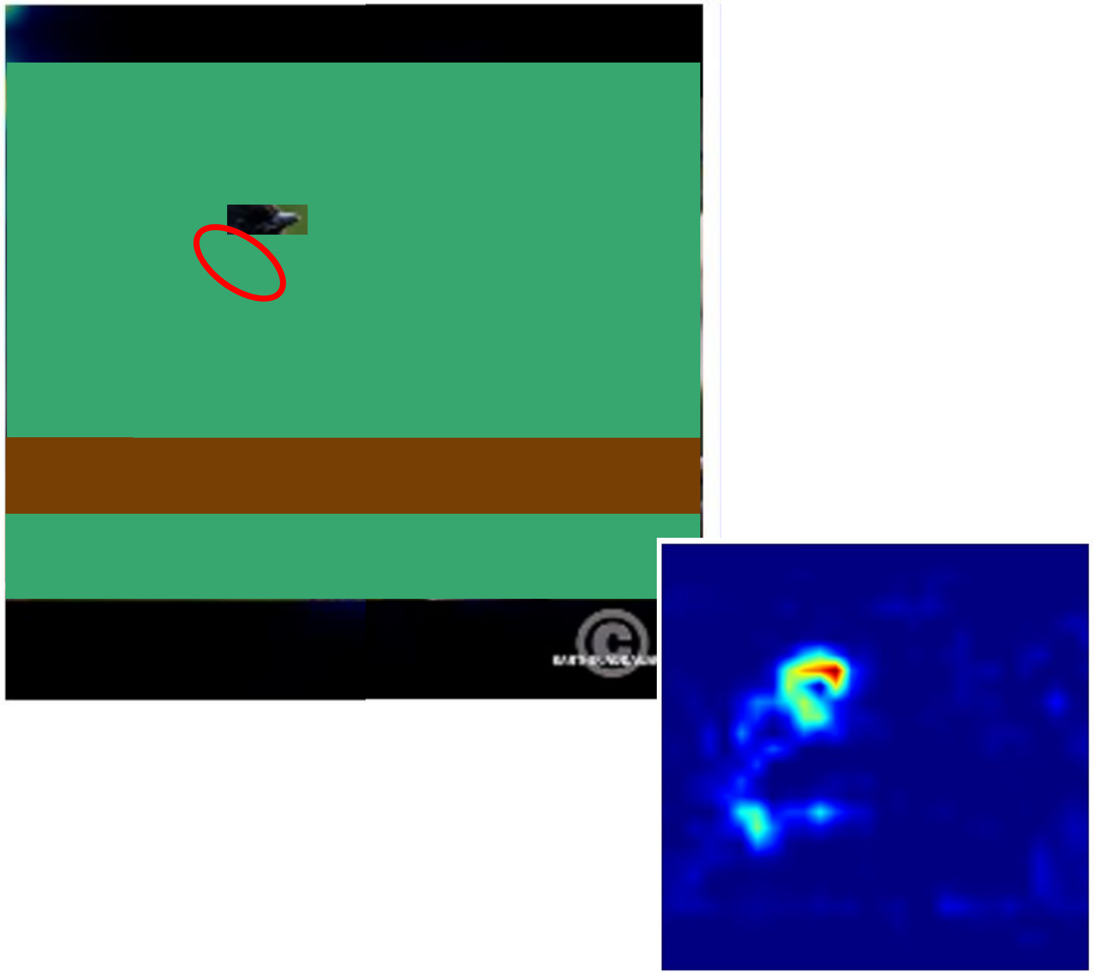
- Insertion: add important features and see what happens..

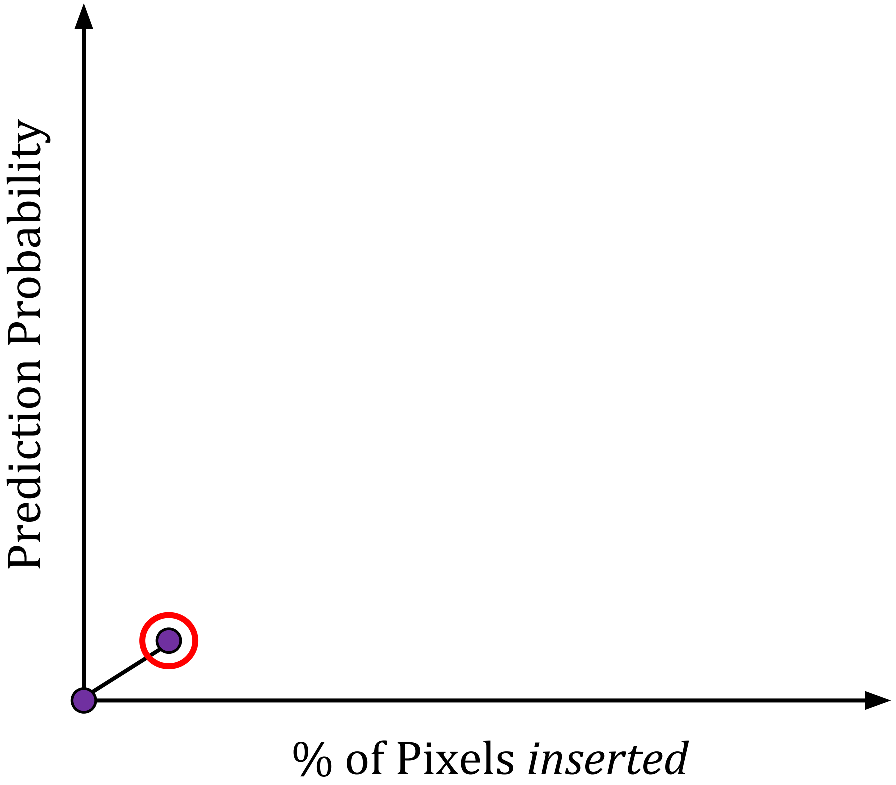




How important are selected features?

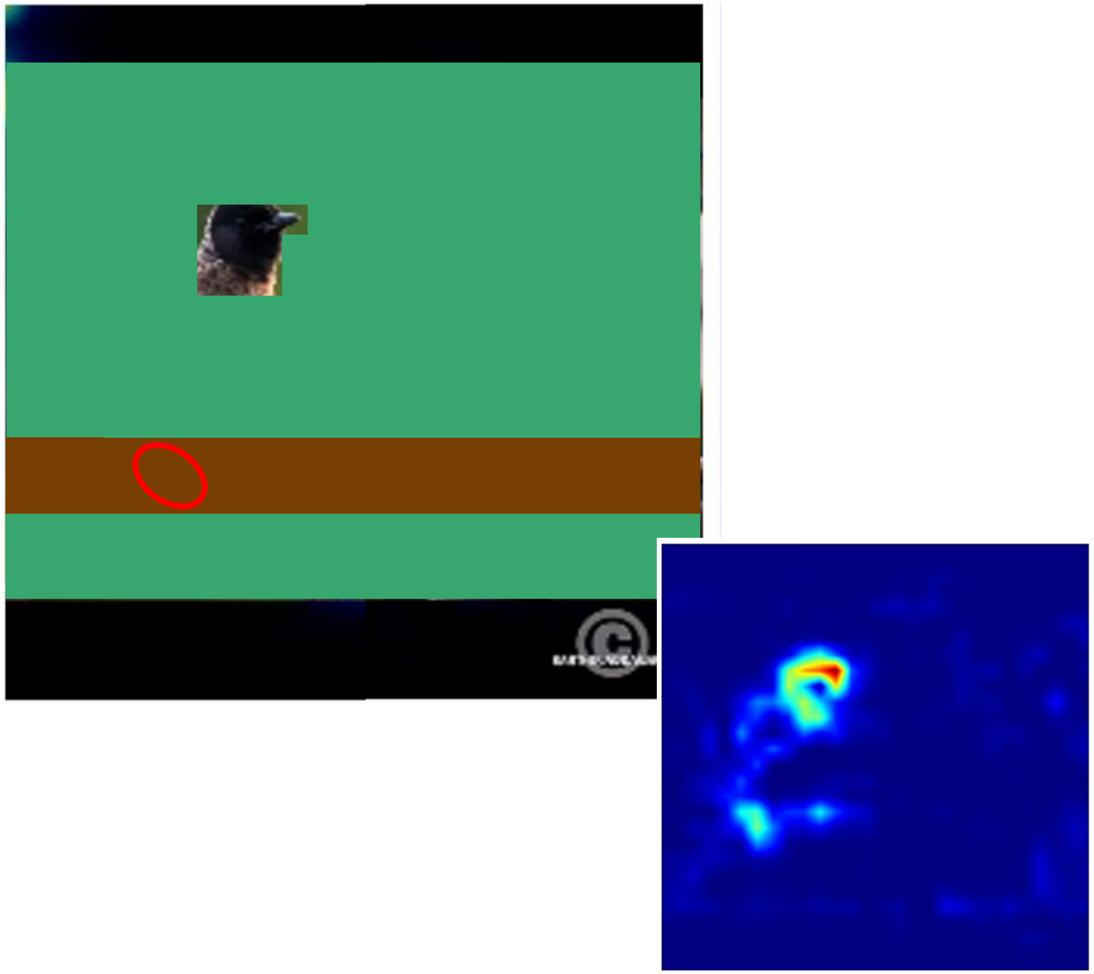
- Insertion: add important features and see what happens..

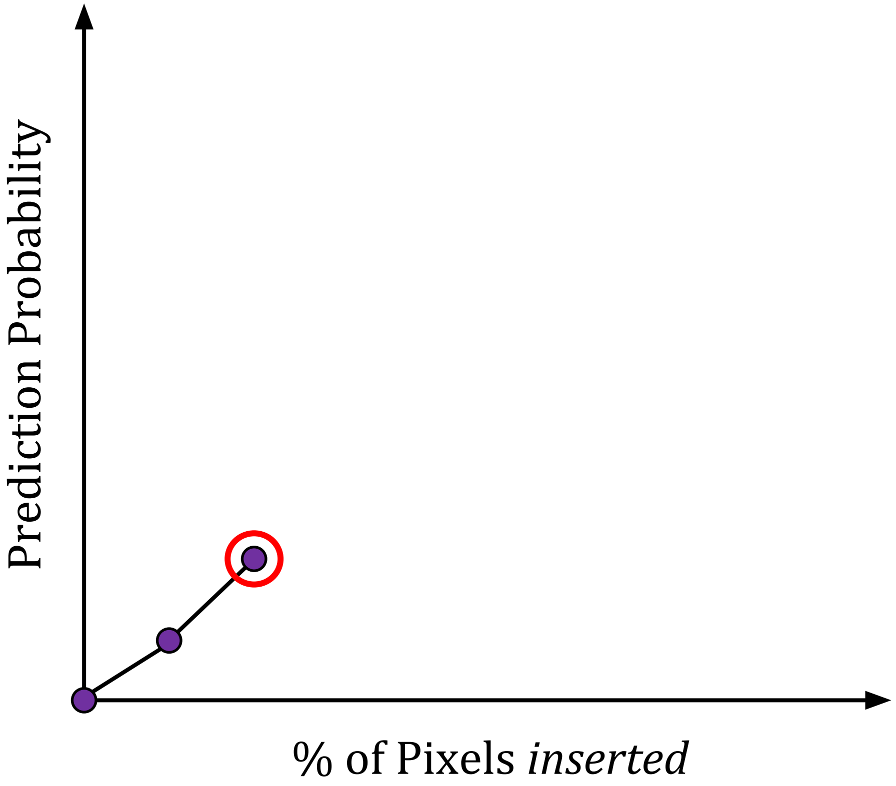




How important are selected features?

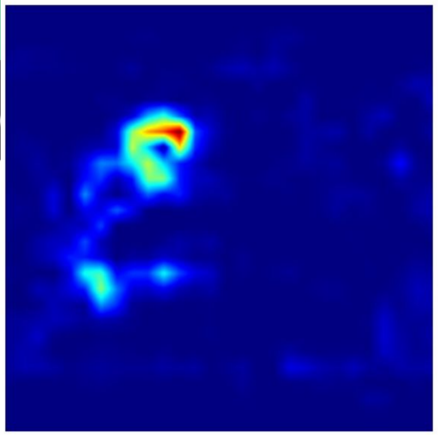
- Insertion: add important features and see what happens..

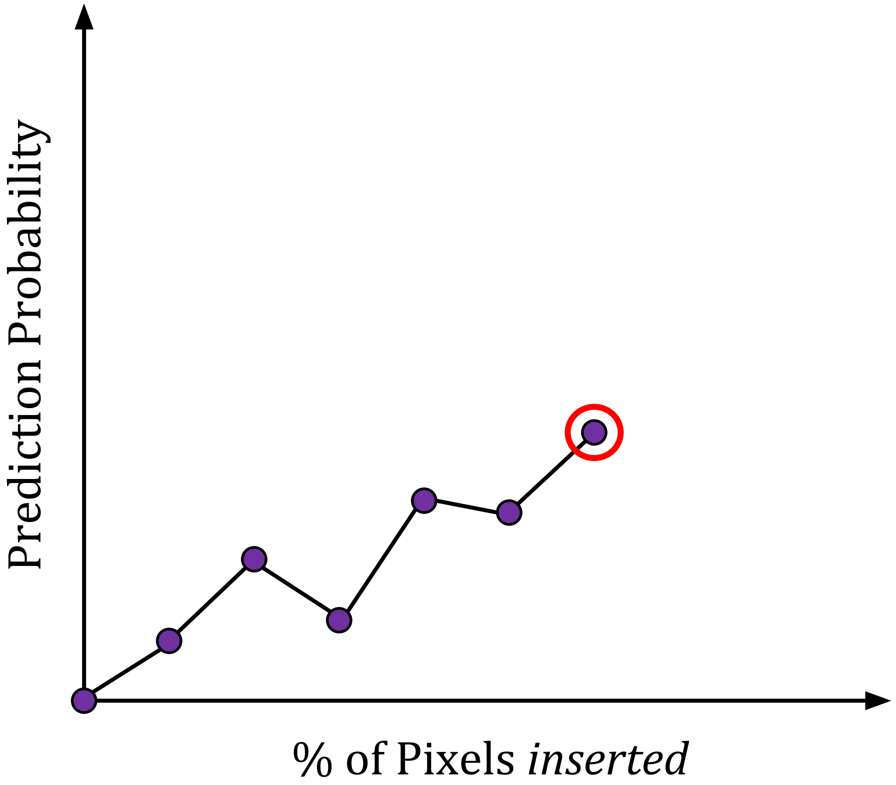




How important are selected features?

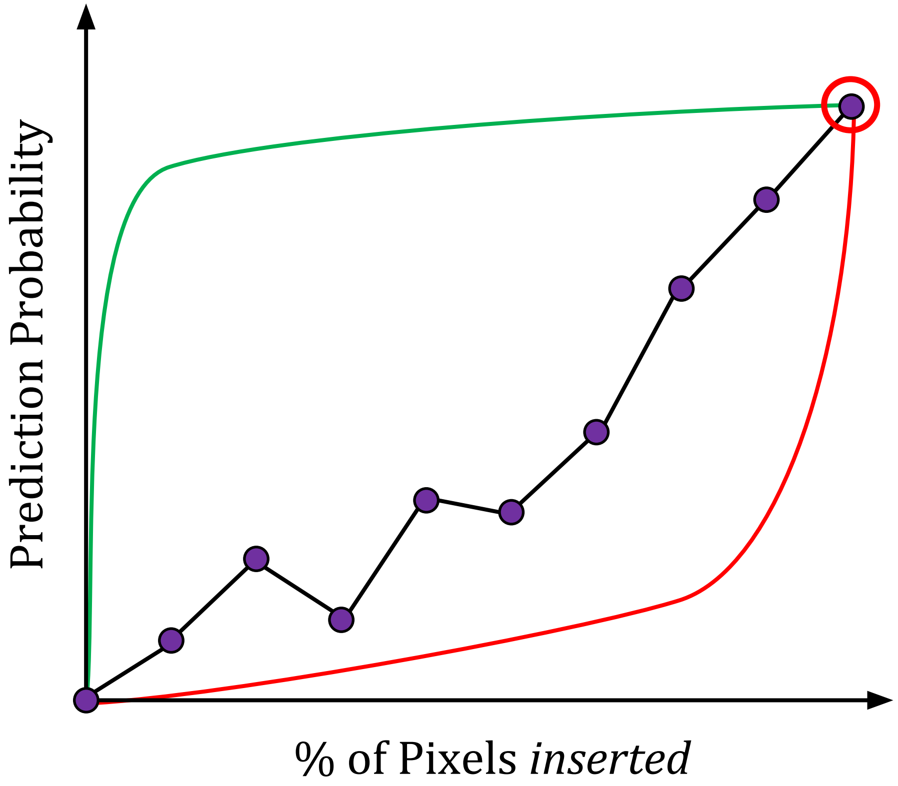
- Insertion: add important features and see what happens..





How important are selected features?

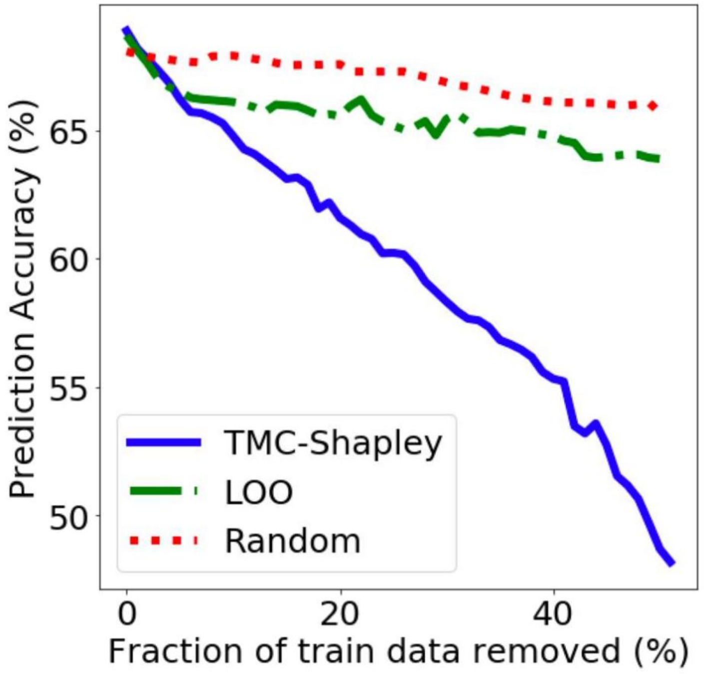
- Insertion: add important features and see what happens..



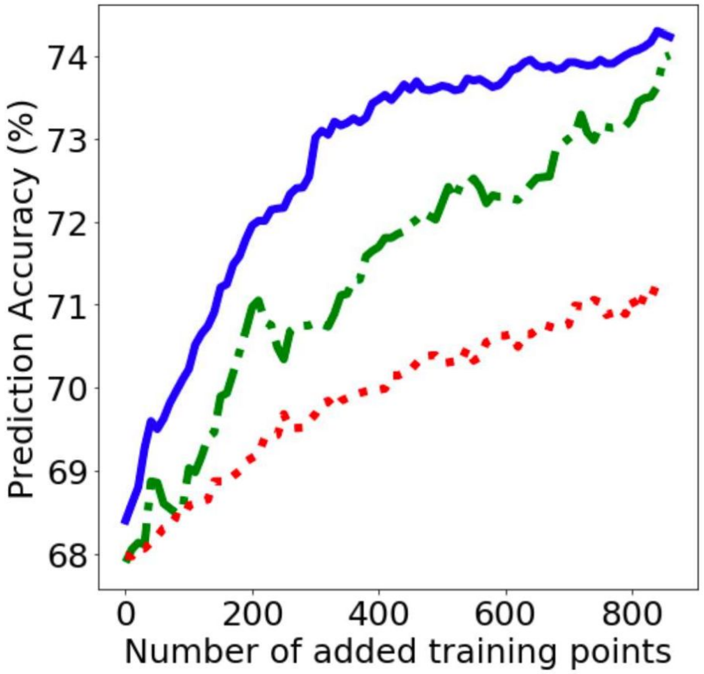
Same Idea: For *Training* Data

Add/remove influential training data, see what happens

Removing high value data

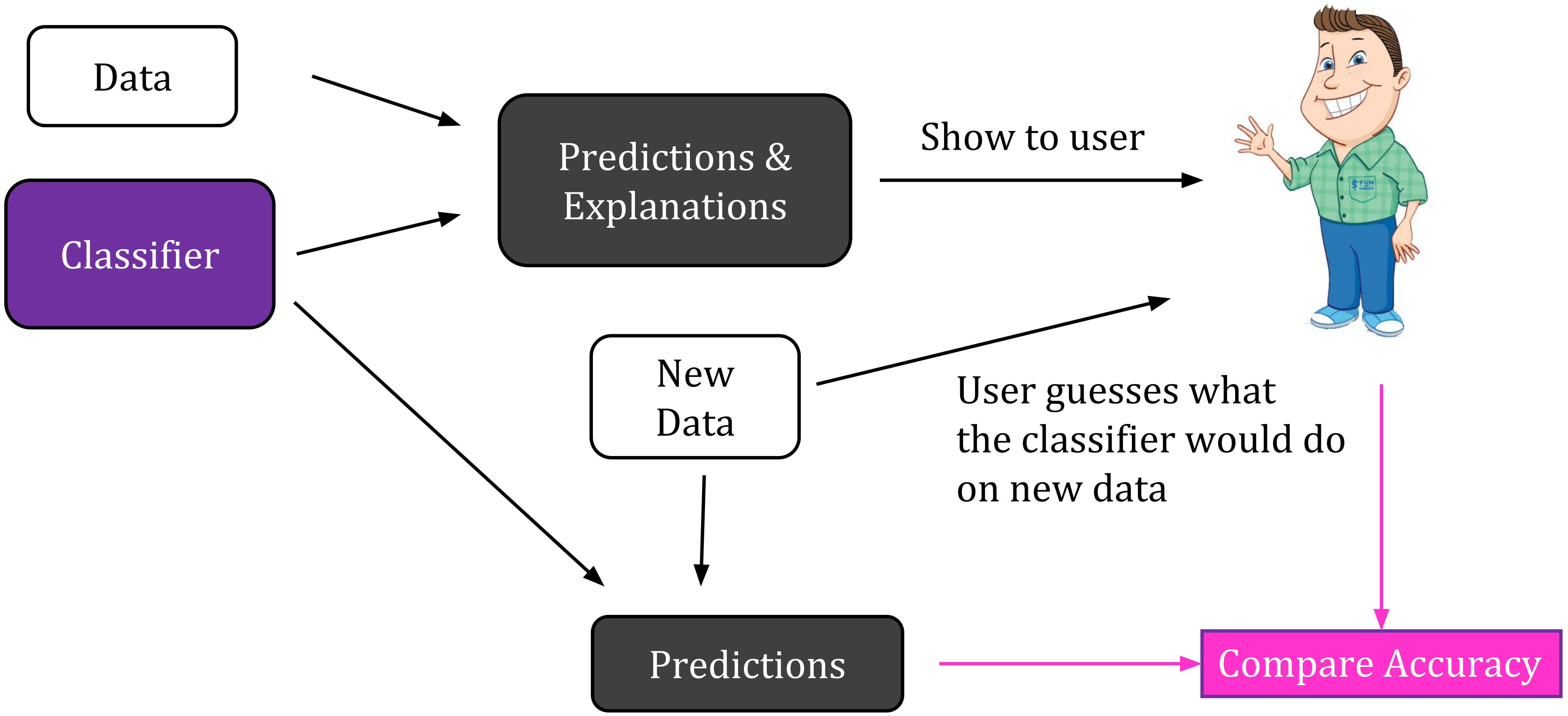


Adding high value data

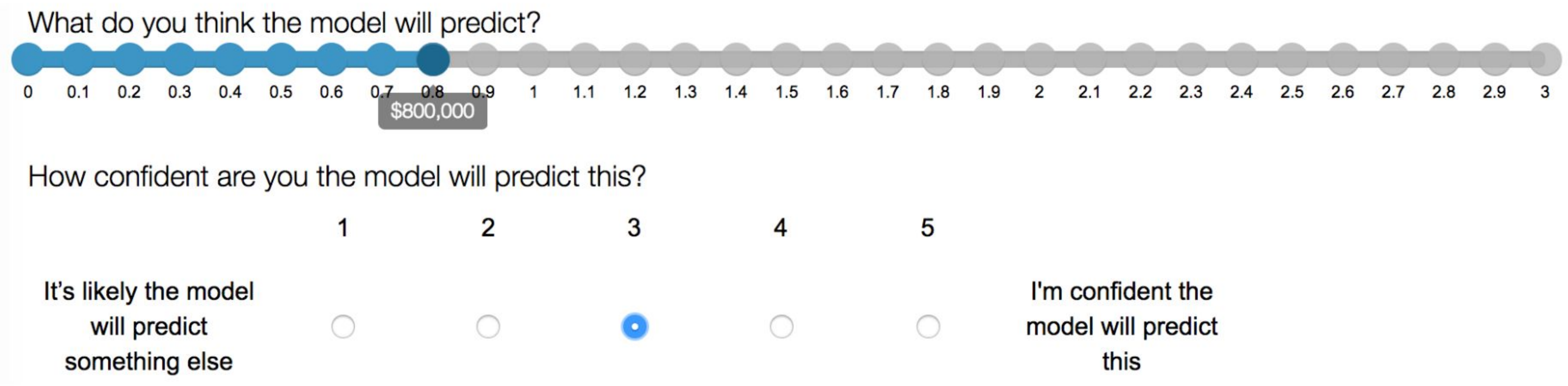


[Ribeiro et al. 2018, Hase and Bansal 2020]

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

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?

[Ribeiro et al. 2016]

2. Comparing Classifiers

Classifier → Explainer

Classifier → Explainer

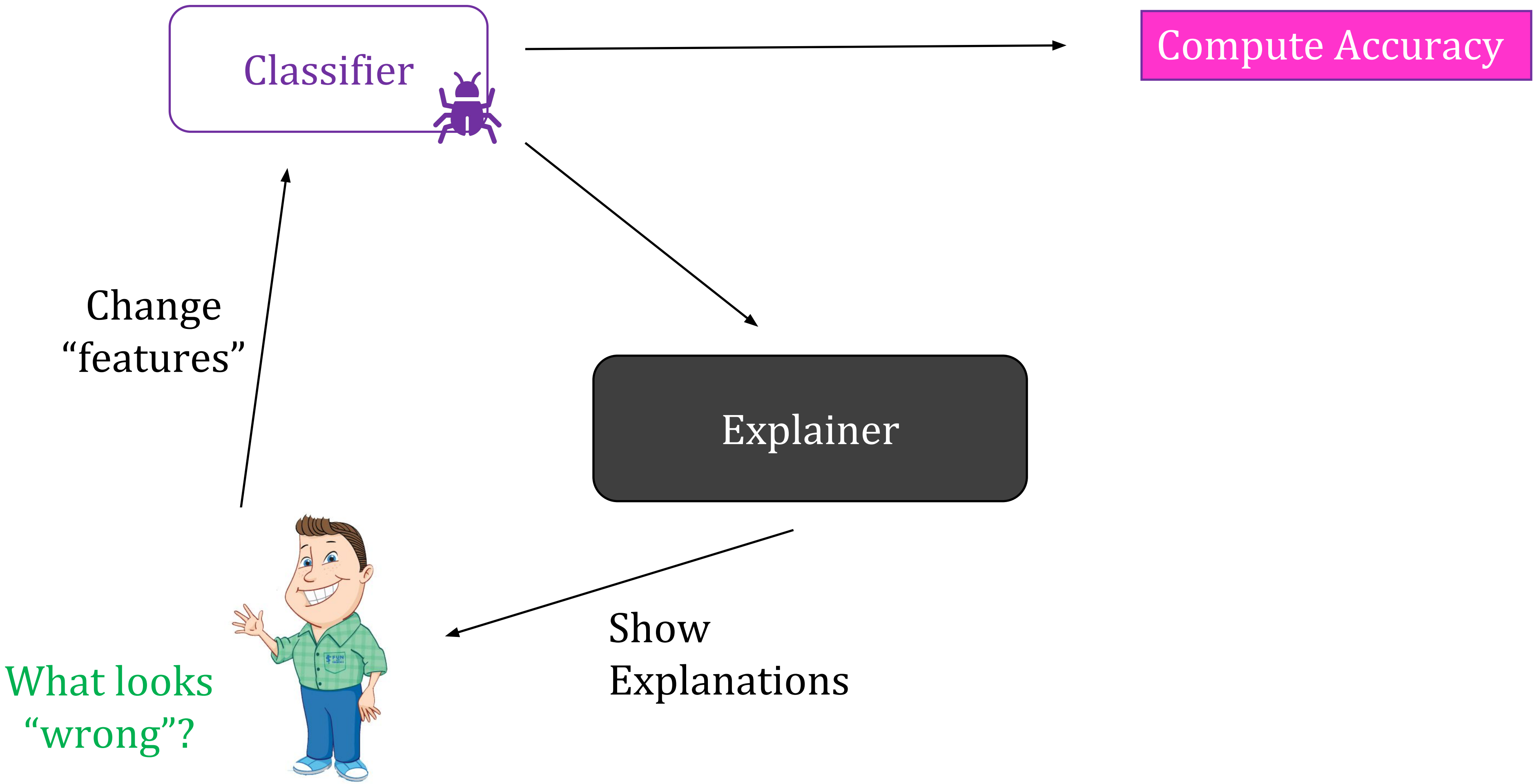
Show
Explanations

Question
Which algorithm is better?

Did they pick the right one?

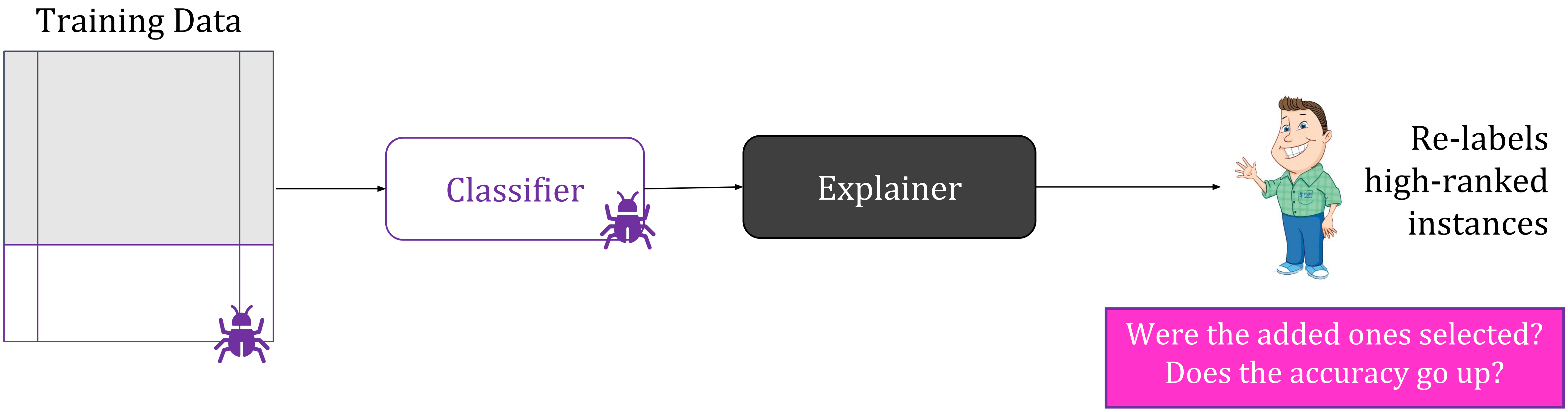
[Ribeiro et al. 2016, Lertvittayakumjorn et al. 2020]

3. “Fixing” Features of Classifiers



4. Finding Errors in Training Data

- Prototypical Explanations: important instances from training data



Evaluating Posthoc Explanations

Understand the Behavior

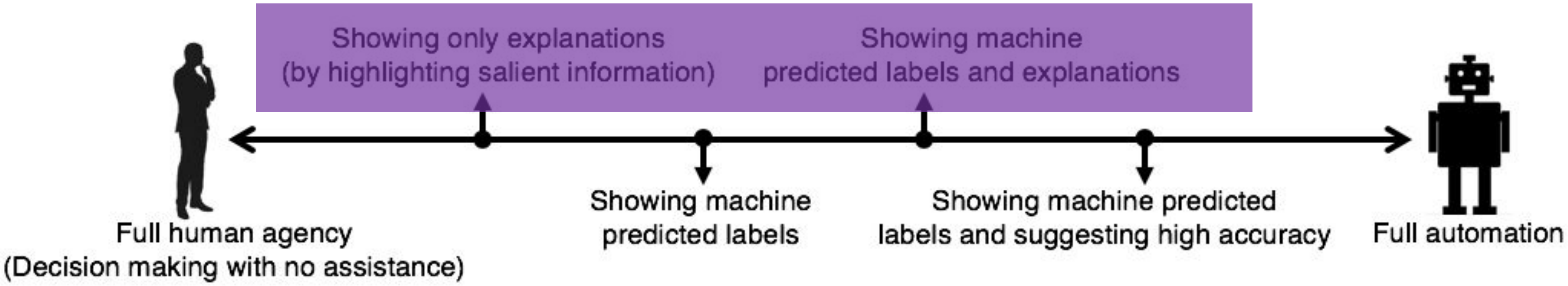
Help make decisions

Useful for Debugging

[Lai and Tan, 2019]

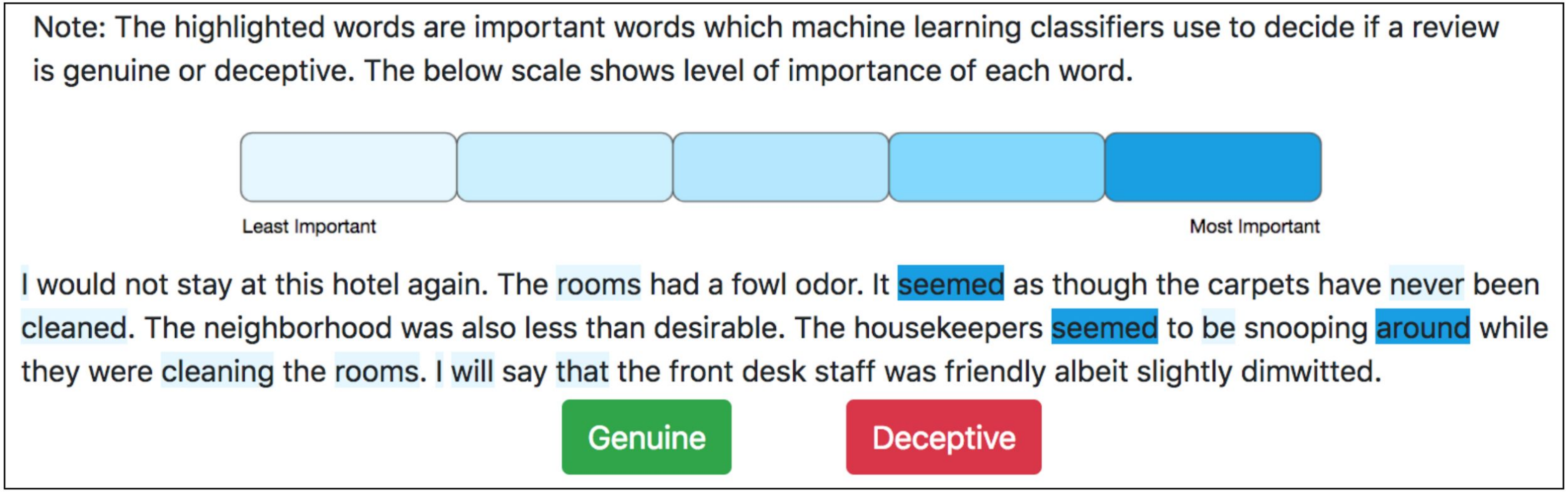
Human-AI Collaboration

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



Human-AI Collaboration

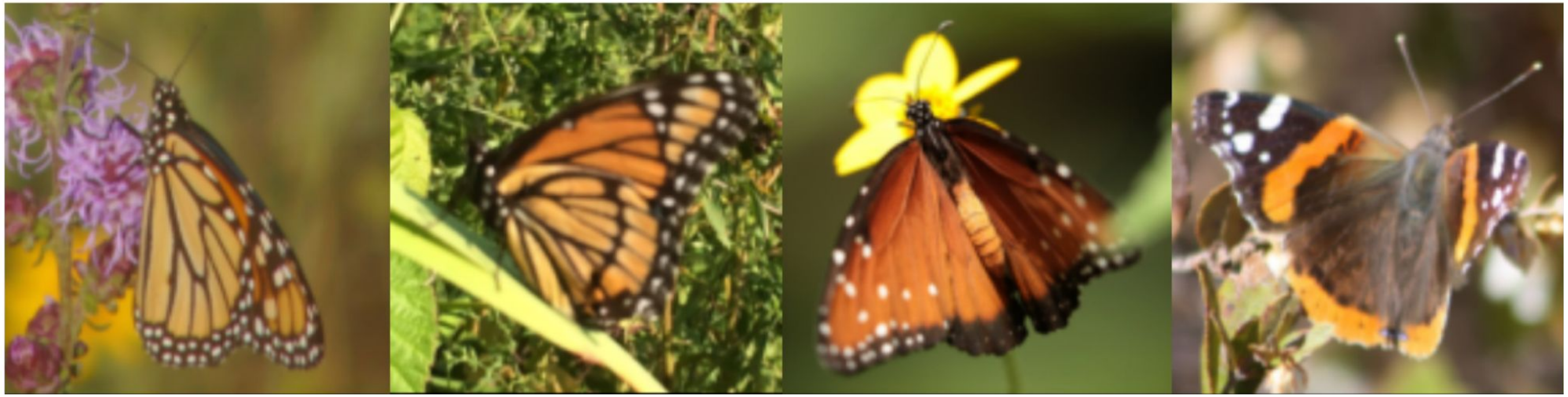
- Deception Detection: Identify fake reviews online
  - Are Humans better detectors with explanations?



https://machineintheloop.com/deception/

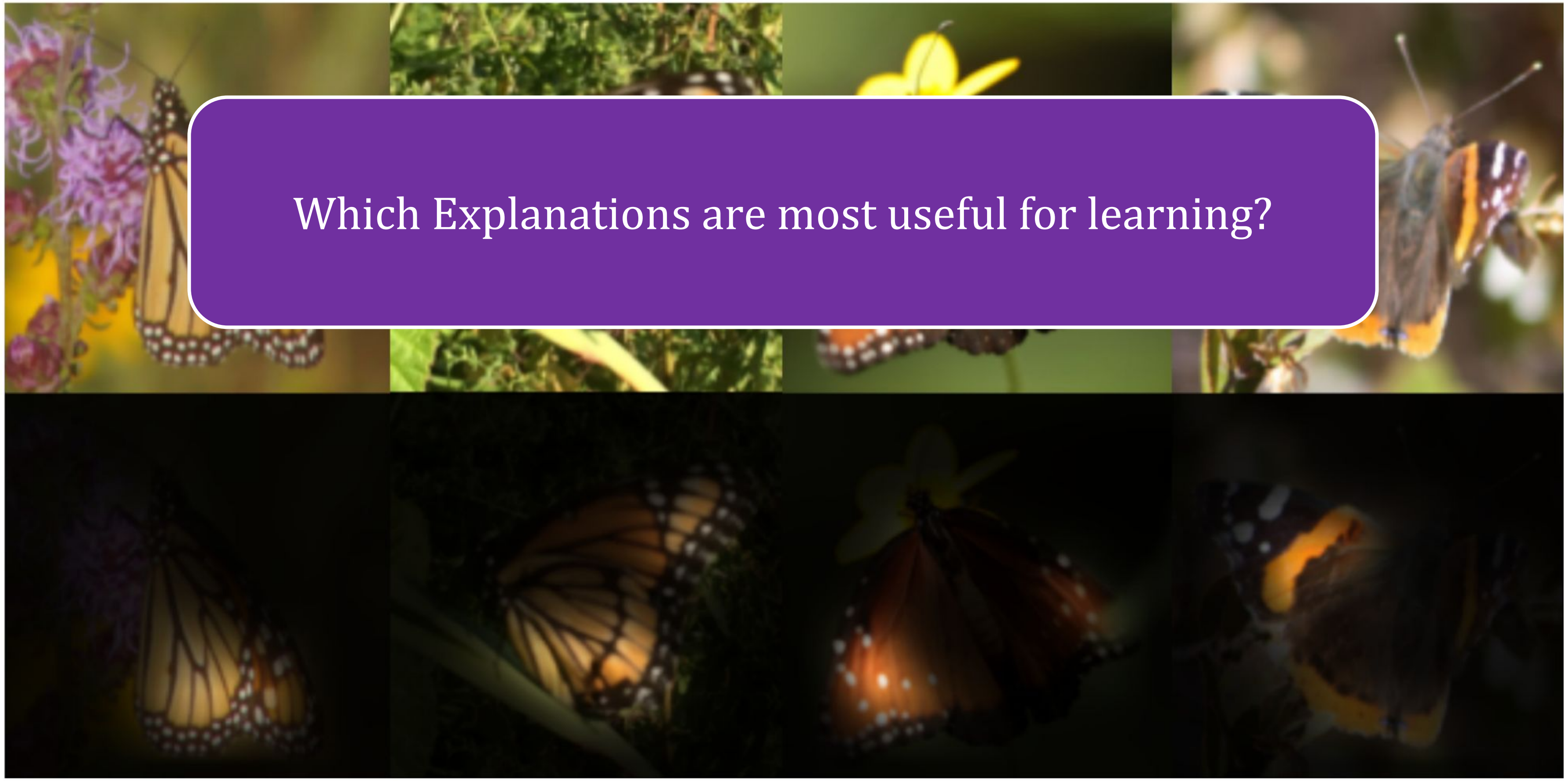
Machine Teaching

Monarch Viceroy Queen Red Admiral



Machine Teaching

Monarch Viceroy Queen Red Admiral



Which Explanations are most useful for learning?

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
- Explanations in Different Modalities
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Explanations in Different Modalities
- 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.

Limitations

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

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

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.

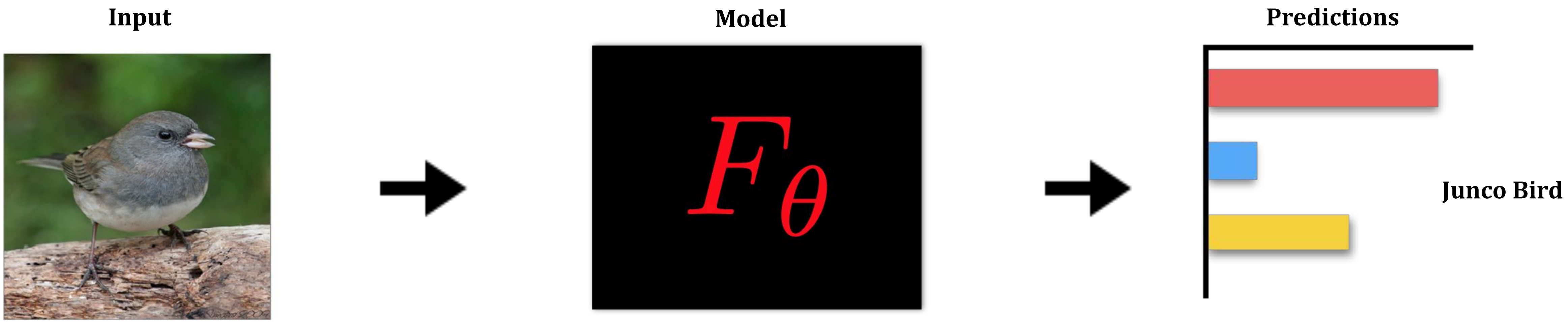
Limitations

- **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.

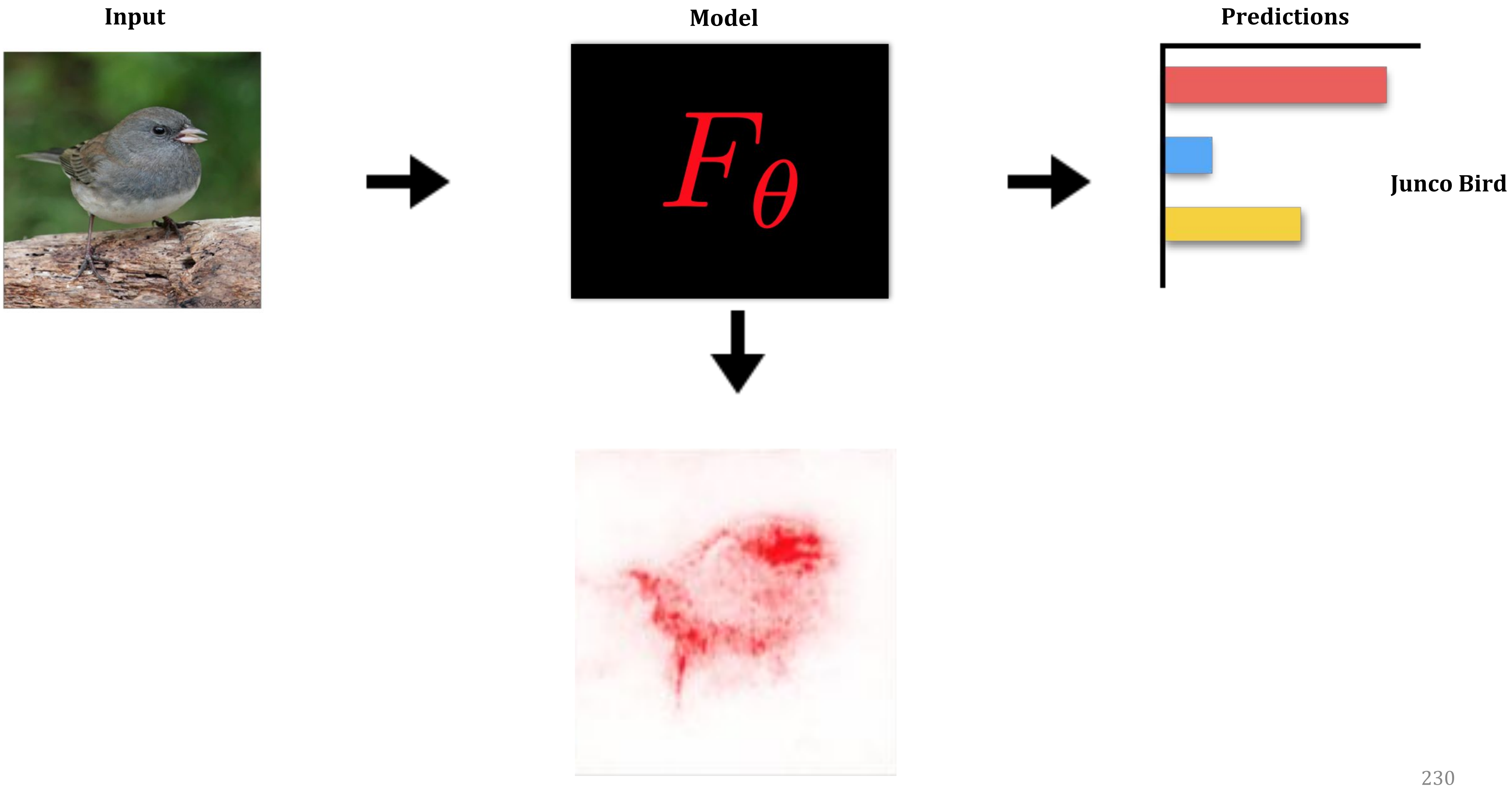
Limitations

- **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?



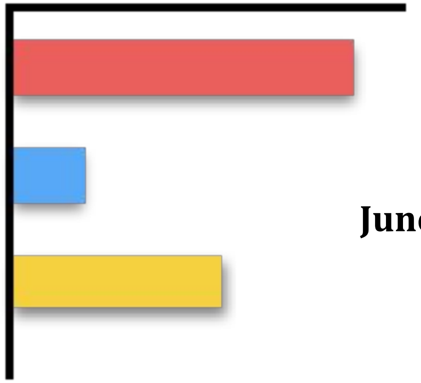
Do Explanations Capture Model-based Discriminative Signals?

Input

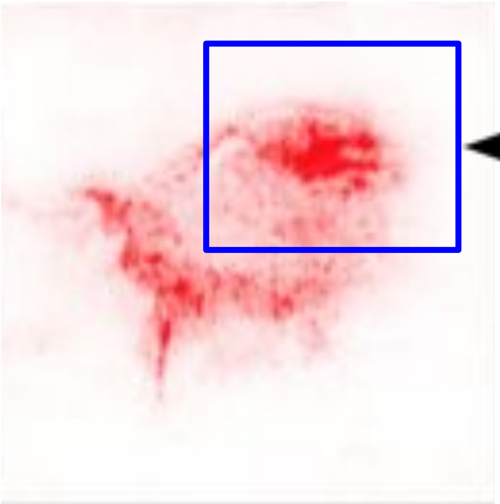
Model

F_θ

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?

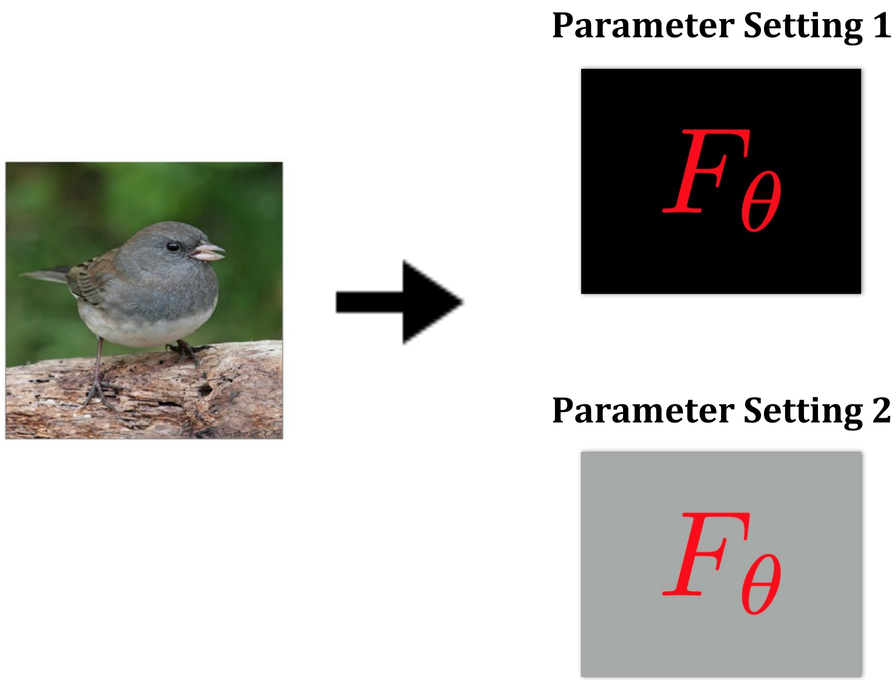
Sanity Check for Faithfulness/Fidelity

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

Adebayo et. al. 2018

Sanity Check for Faithfulness/Fidelity

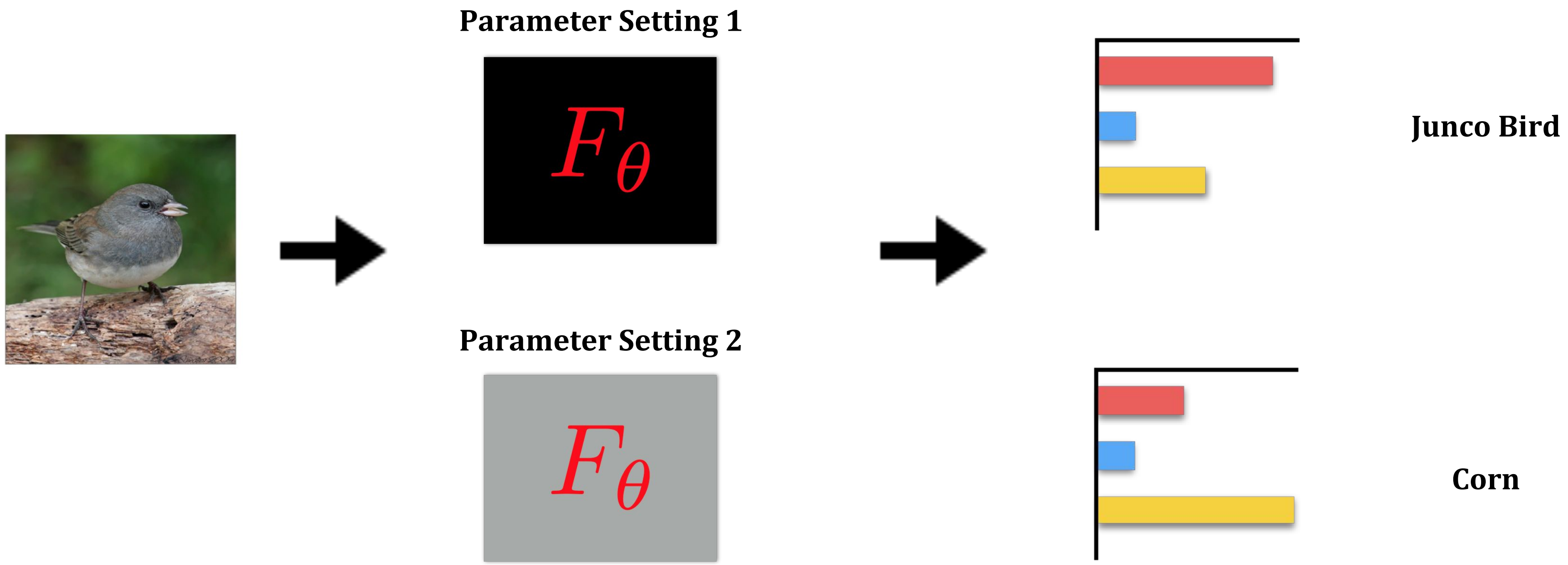
- **Sensitivity to Model Parameters:** if the parameter settings change, the explanations should change.



[Adebayo et. al. 2018]()

Sanity Check for Faithfulness/Fidelity

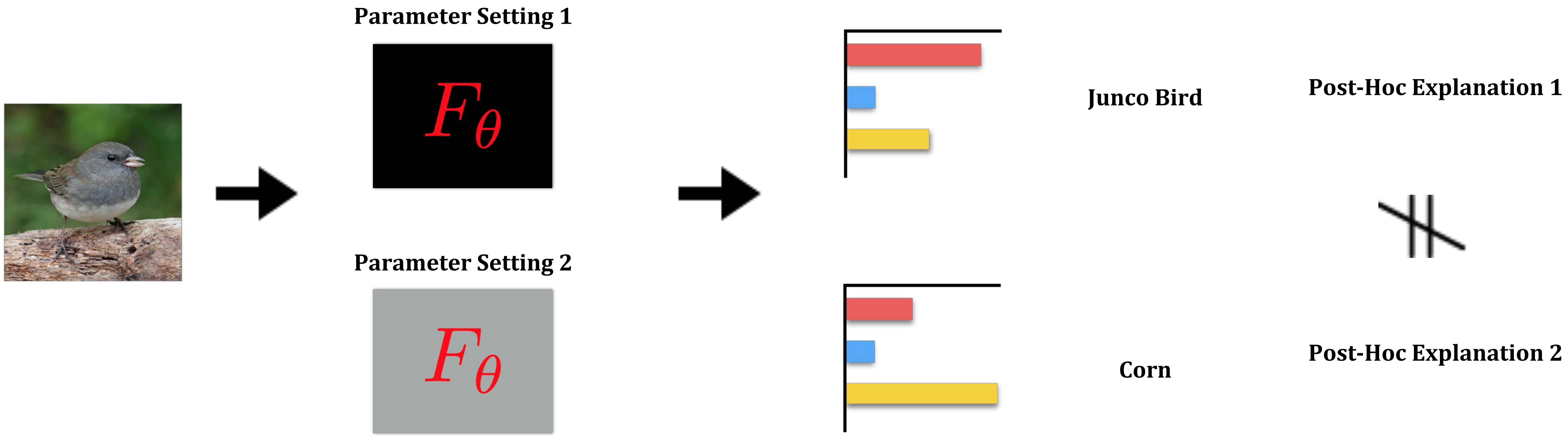
- **Sensitivity to Model Parameters:** if the parameter settings change, the explanations should change.



[Adebayo et. al. 2018]()

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.

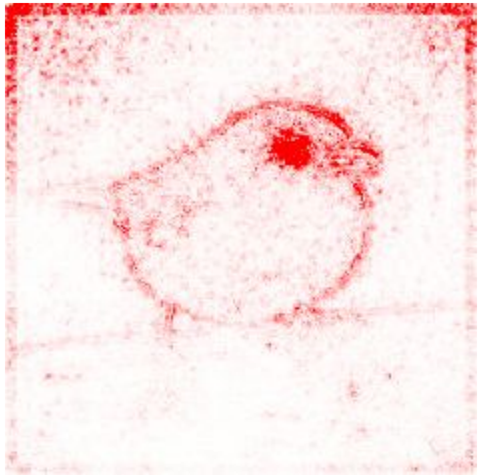
Guided BackProp Explanation Inception-V3 ImageNet

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](#)

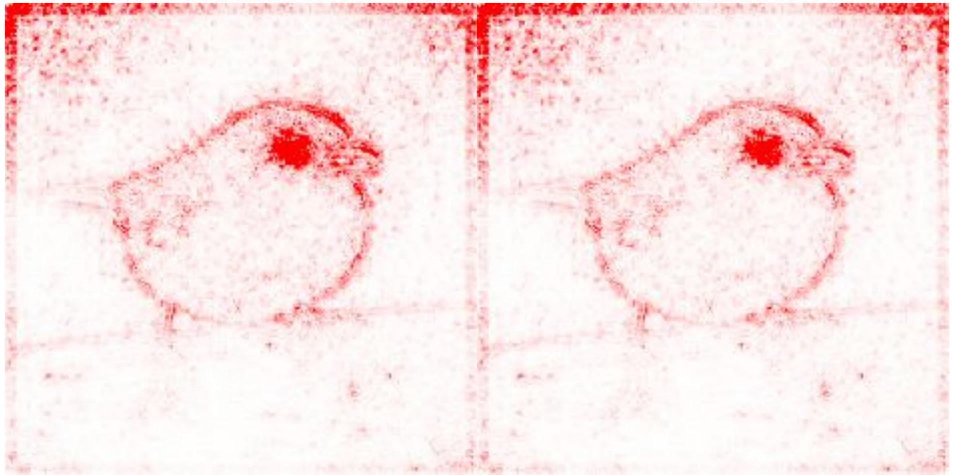
Cascading Randomization Inception-V3

- **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



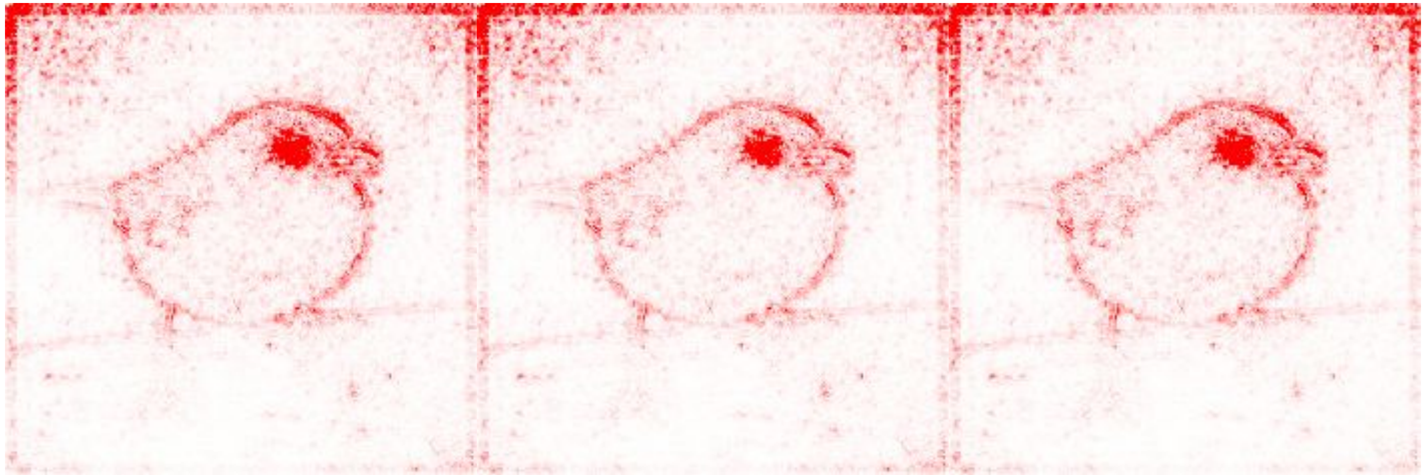
[Adebayo et. al. 2018](#)

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**

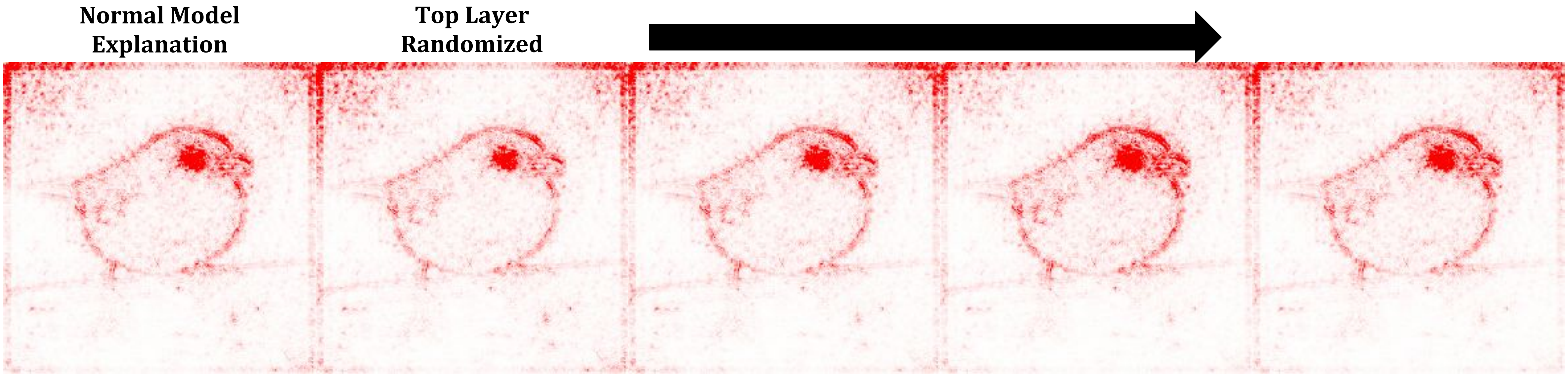


[Adebayo et. al. 2018]

Cascading Randomization Inception-V3

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

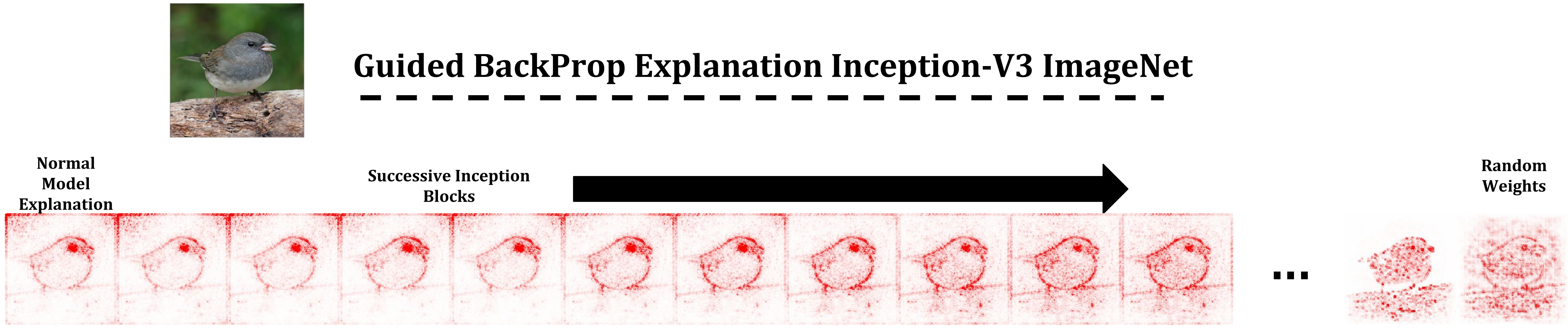
Guided BackProp Explanation Inception-V3 ImageNet



[Adebayo et. al. 2018](#)

Cascading Randomization Inception-V3

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

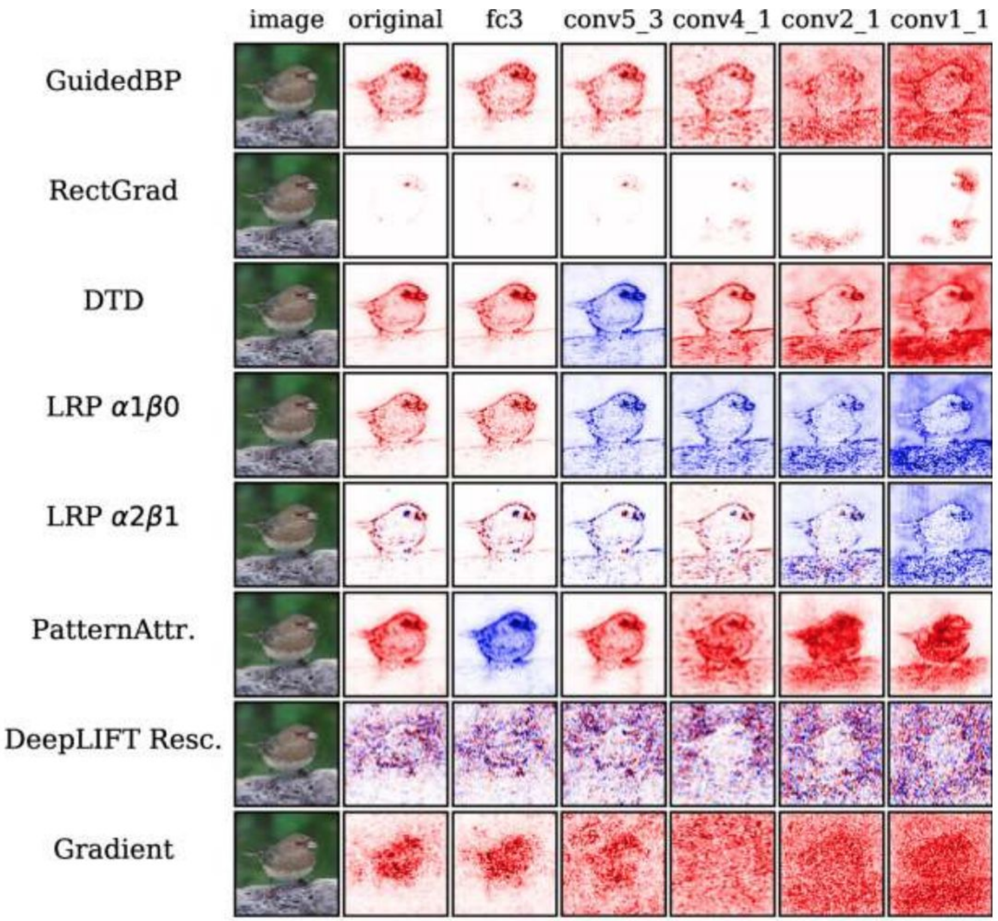


Guided BackProp is invariant to the higher level weights.

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).

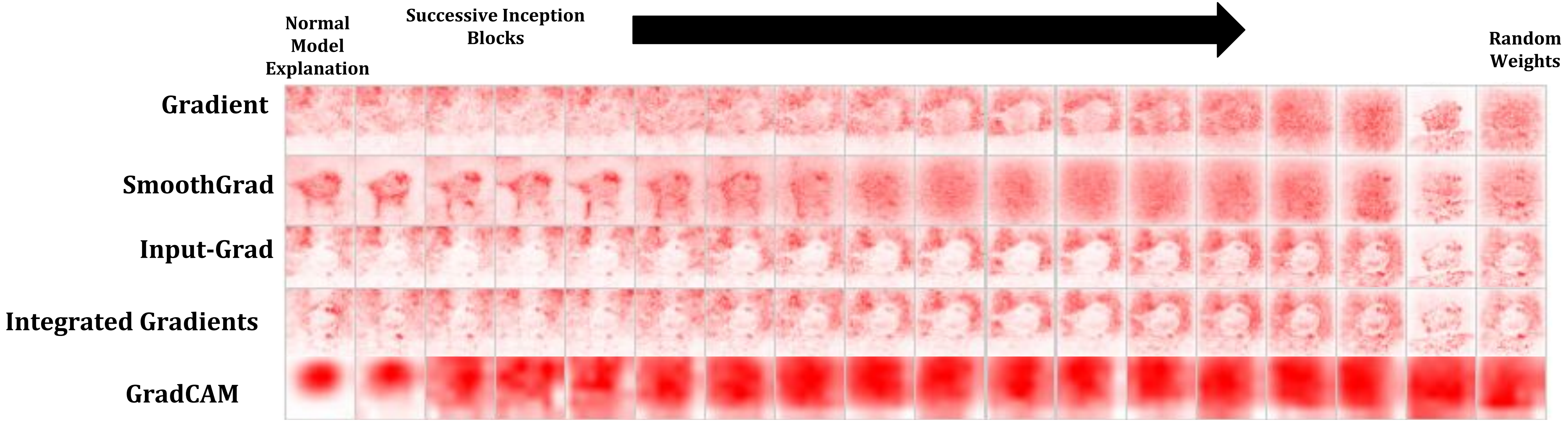
- **These modified backprop methods converge to a rank-1 matrix**! This is because the product of a sequence of non-negative matrices (non-orthogonal columns, along with other assumptions) converges to a rank-1 matrix (*Theorem 1 in Sixt et. al. 2020*).

- -

- **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

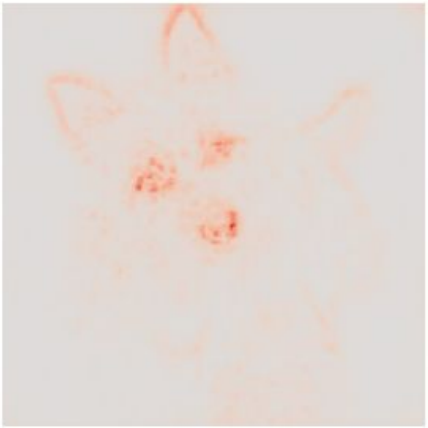
Post-hoc explanations can be easily manipulated.

Original Image

[Dombrowski et. al. 2019]

Post-hoc Explanations are Fragile

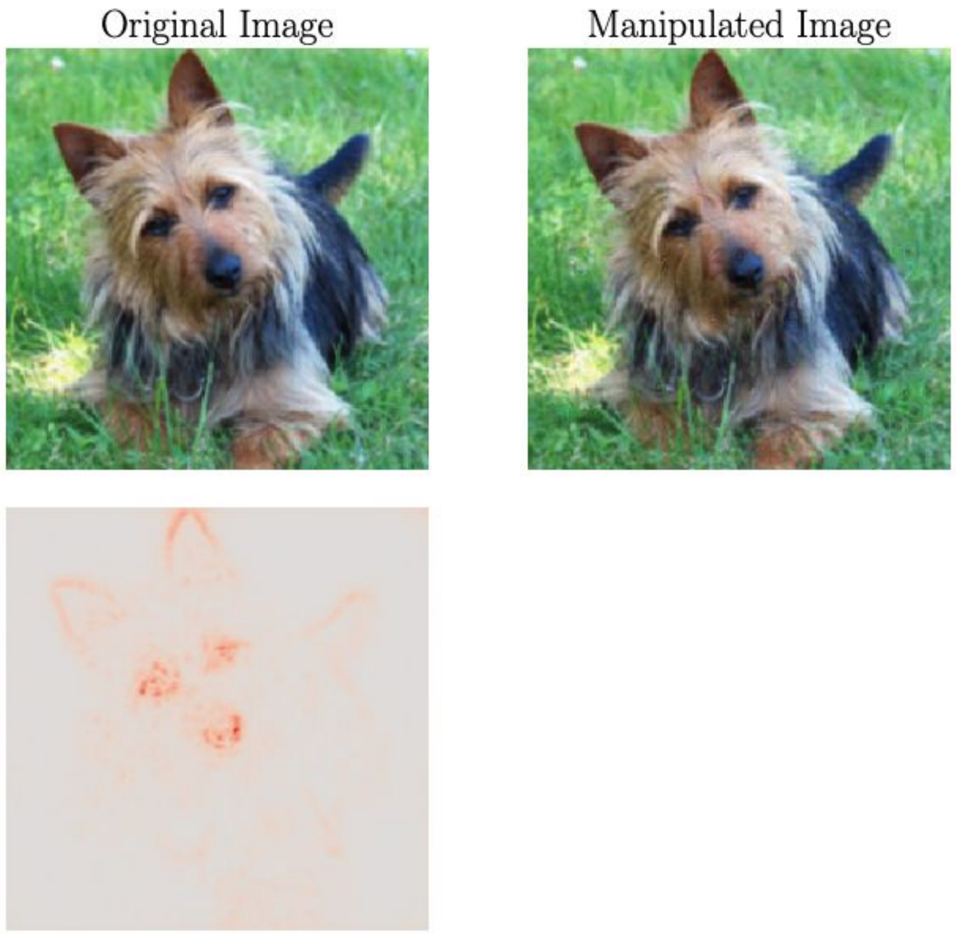
Post-hoc explanations can be easily manipulated.



[Dombrowski et. al. 2019]

Post-hoc Explanations are Fragile

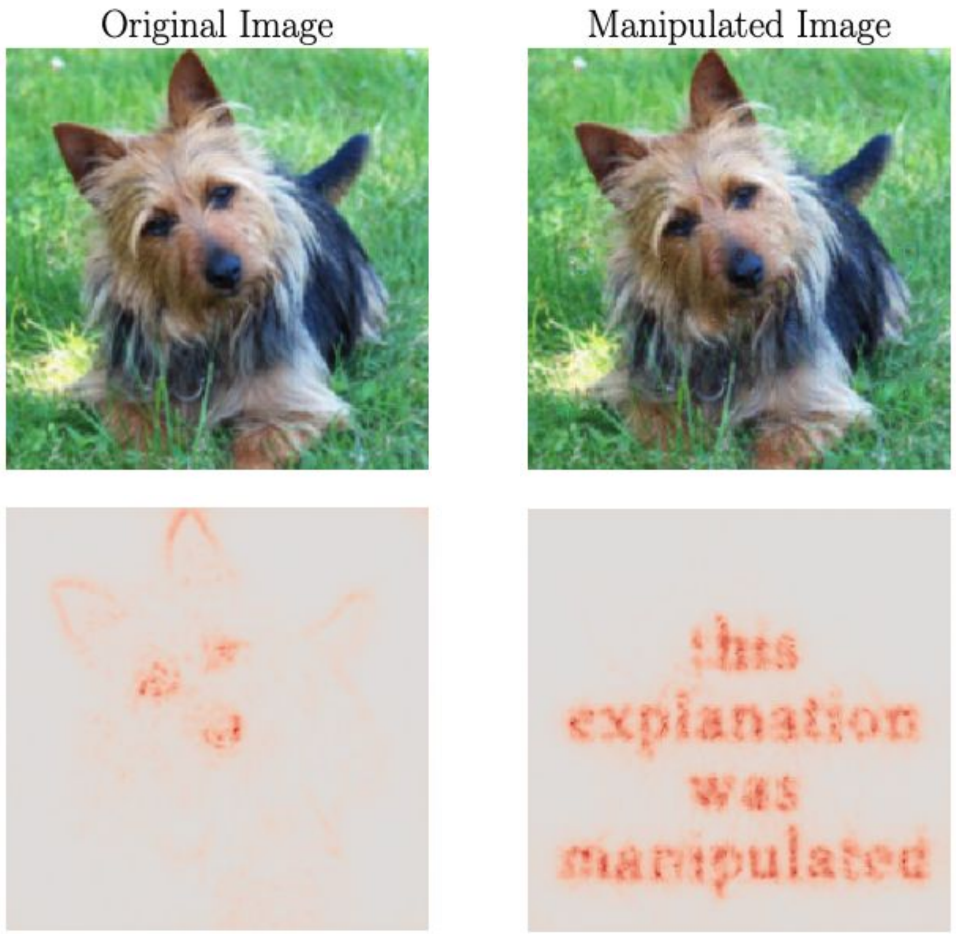
Post-hoc explanations can be easily manipulated.



[Dombrowski et. al. 2019](#)

Post-hoc Explanations are Fragile

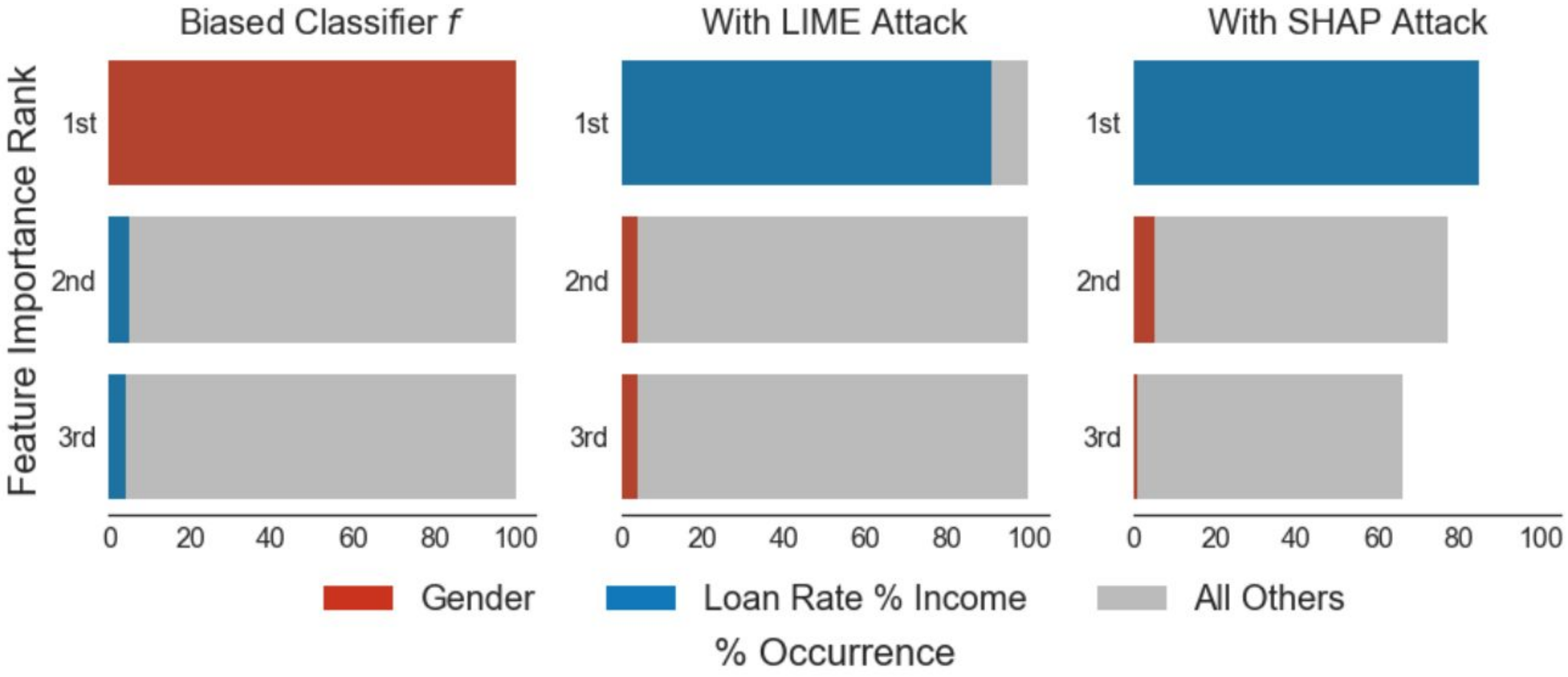
Post-hoc explanations can be easily manipulated.



[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**.

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

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

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

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).

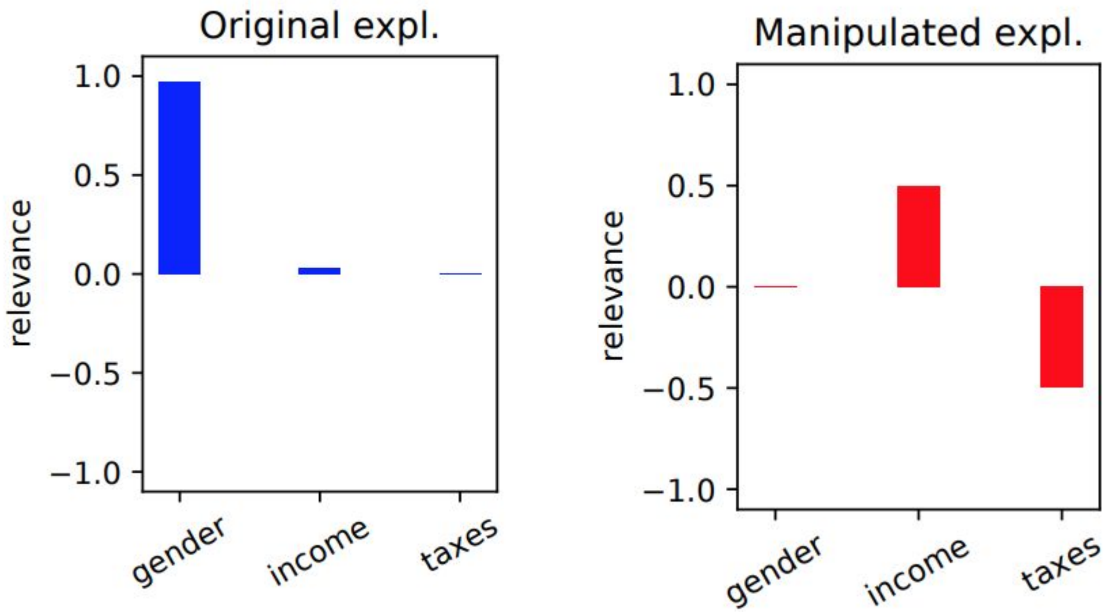
Methods Affected

- **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

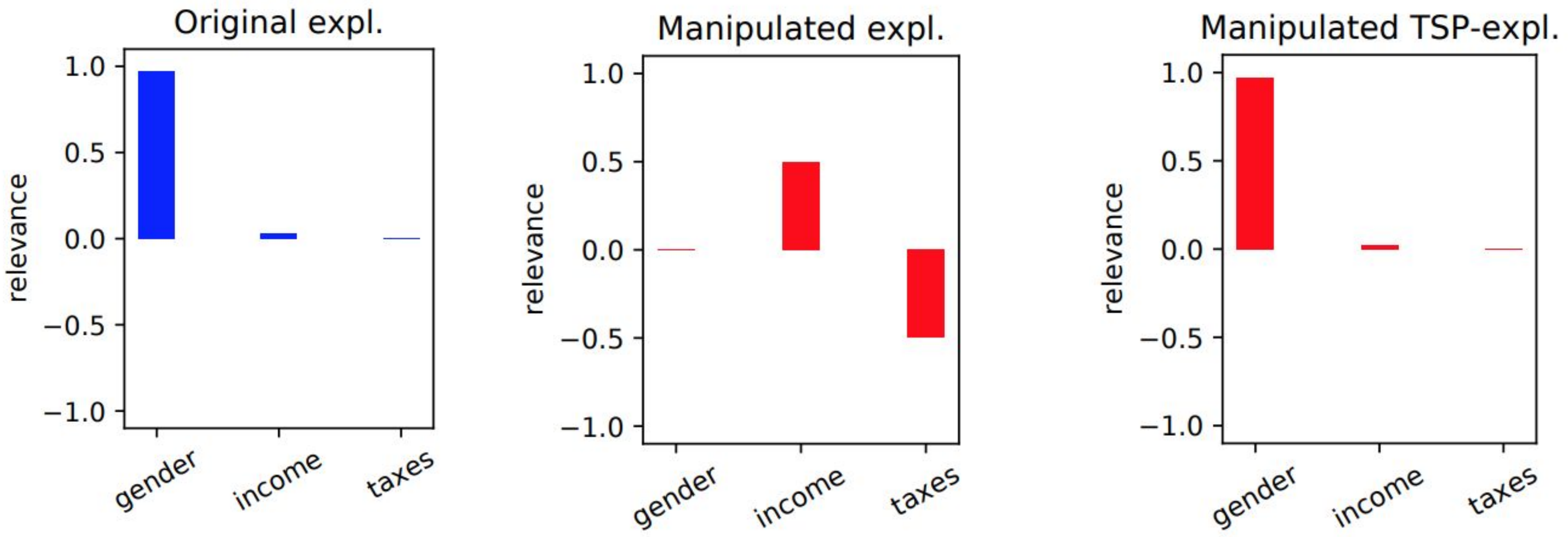


[Anders et. al., 2020]()

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.

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

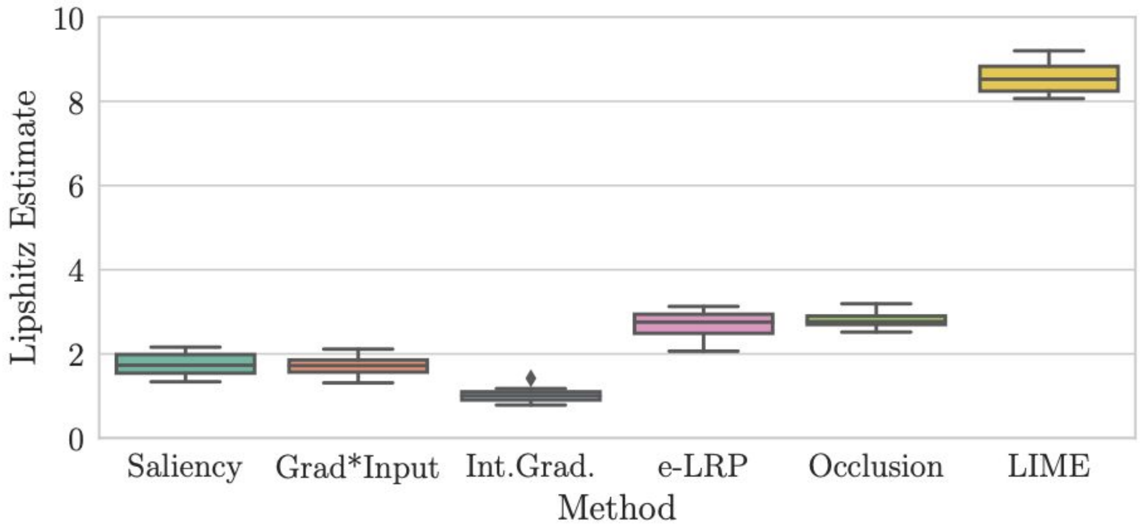
Input

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.

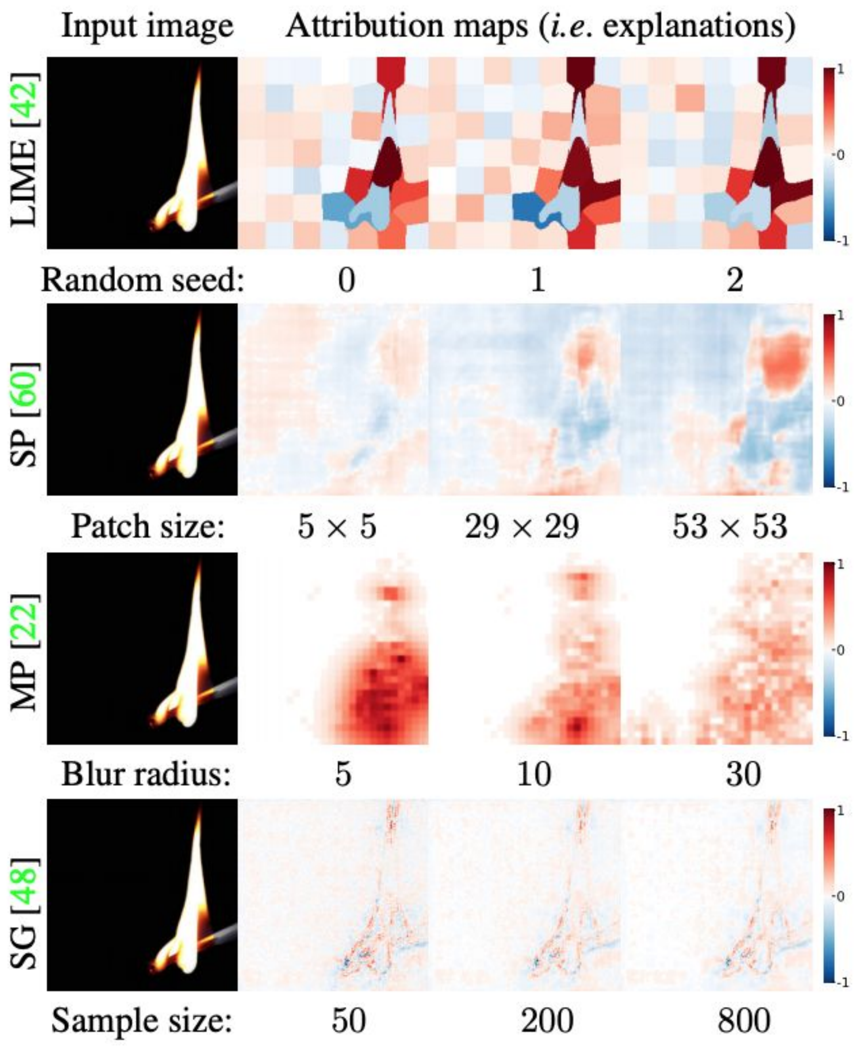


Estimate for 100 tests for an MNIST Model.

[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

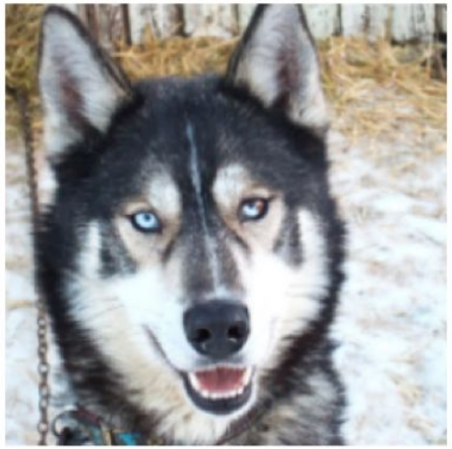
- ~~**Faithfulness/Fidelity**~~
  - ~~Some explanations do not reflect the underlying model.~~

- ~~**Stability**~~

- **Useful in practice?**
  - Unclear if a data scientist (ML engineer)/lay person use explanations to isolate errors, improve 'trust', and 'simulatability' in practice?

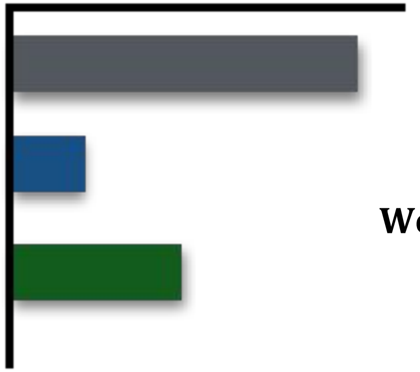
Model Debugging: Spurious Signals

True Label: Siberian Husky



Model

Predictions



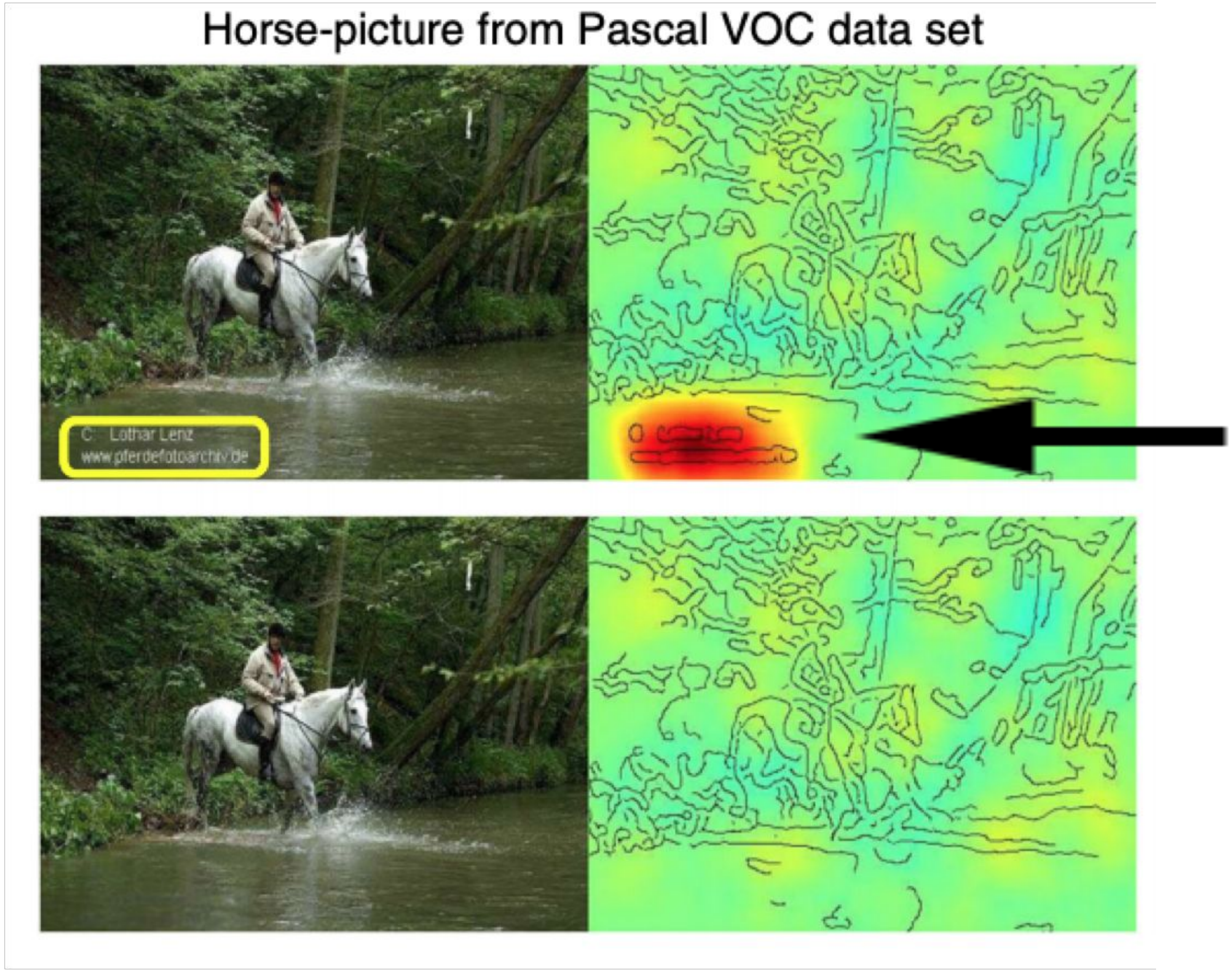
Wolf

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

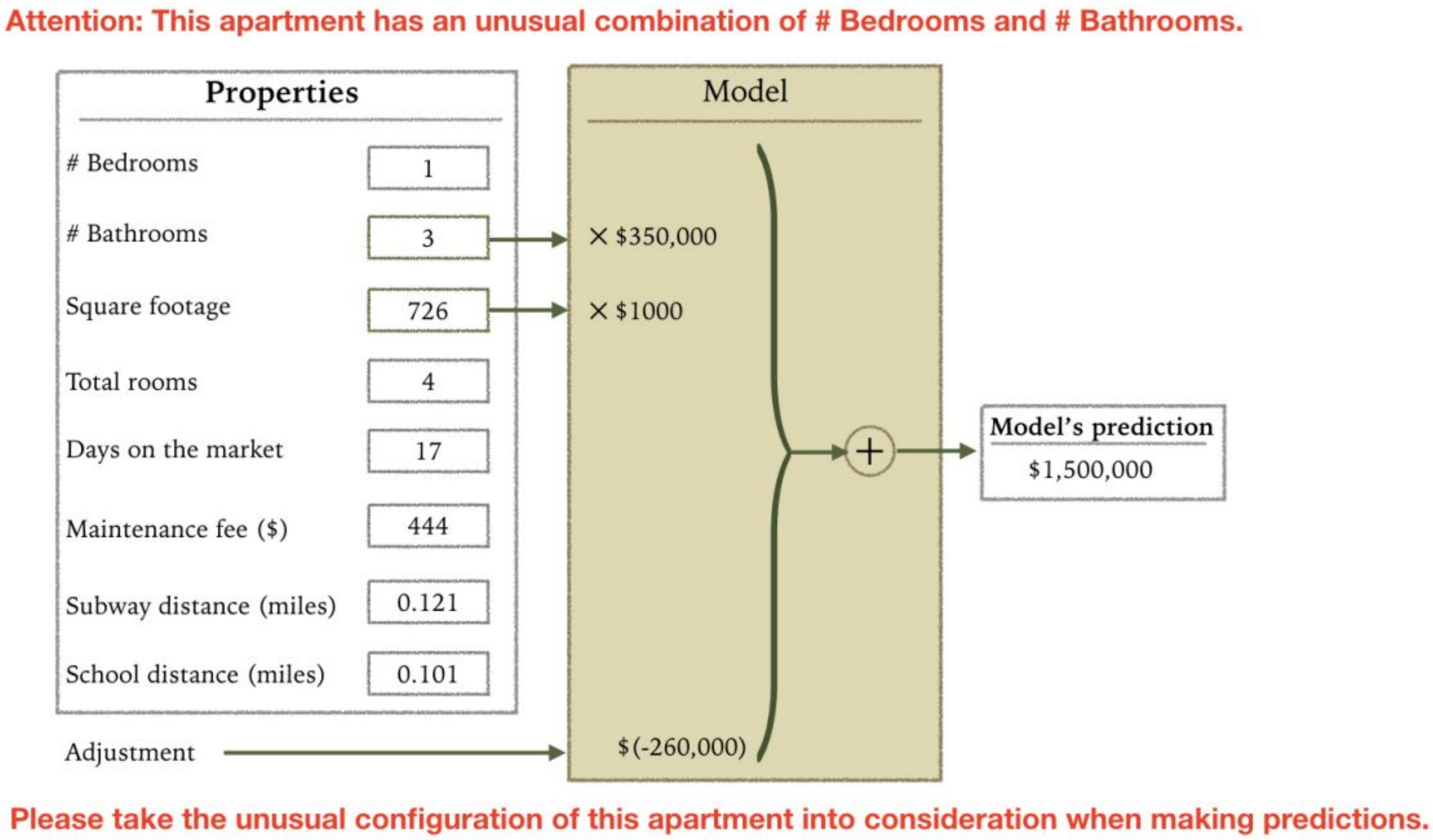
If Current-Offense = Misdemeanor and Prior-Arrests ≤ 1:
- If Has-Kids = No and Owns-House = No and Prior-Jail-Incarcerations = Yes, then Risky
- If Age ≥ 50 and Has-Kids = Yes and Prior-FTA = No, then Not Risky

Default: Not Risky

Lakkaraju & Bastani 2019.

Difficulty using explanations for debugging

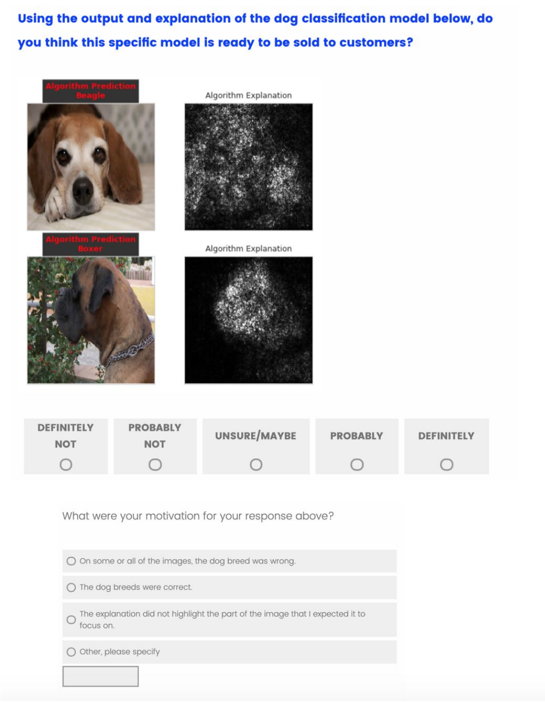
In a housing price prediction task, Amazon mechanical turkers are unable to use linear model coefficients to diagnose model mistakes.



[Poursabzi-Sangdeh et. al. 2019]

Difficulty using explanations for debugging

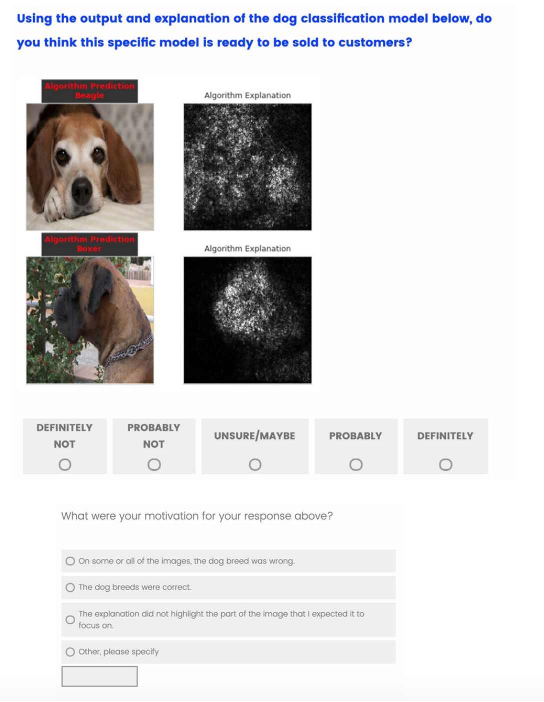
In a dog breeds classification task, users familiar with machine learning **rely on labels, instead of saliency maps,** for diagnosing model errors.

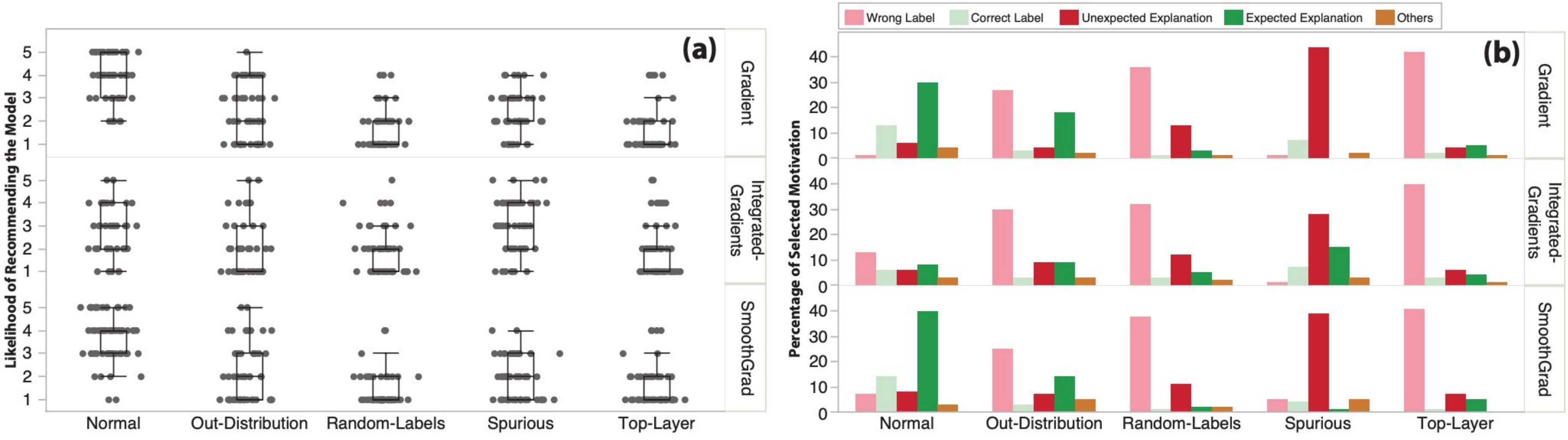


[Adebayo et. al., 2020.]

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.

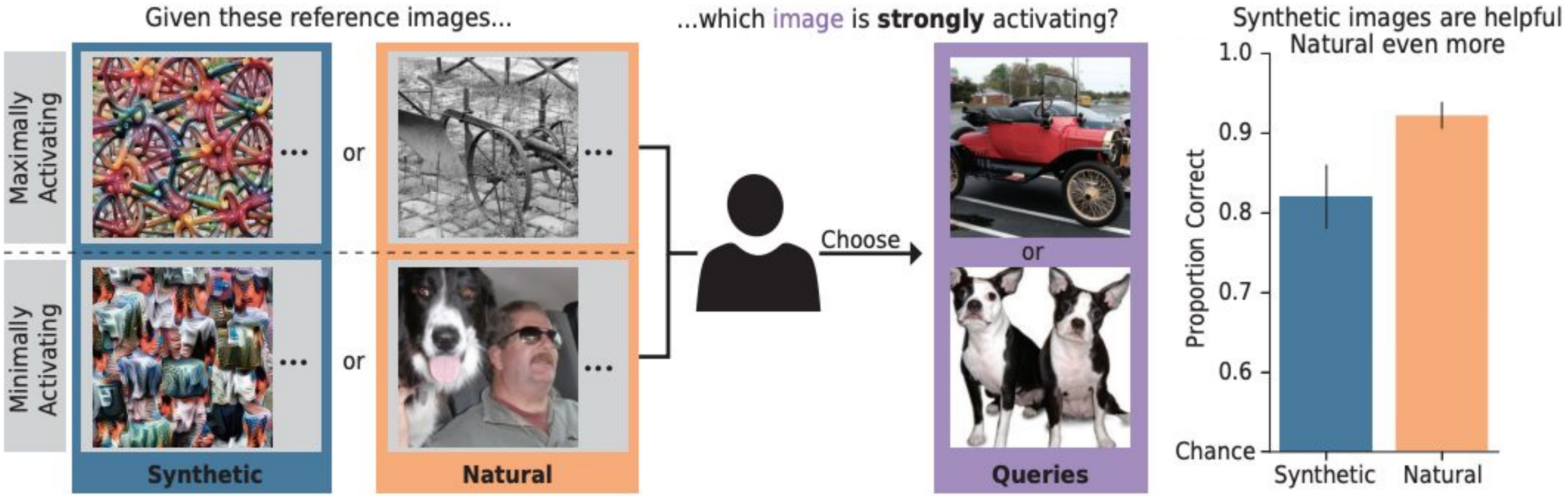




[Adebayo et. al., 2020.](#)

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

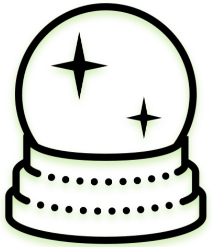
- Approaches for Post hoc Explainability
- Explanations in Different Modalities
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

Tutorial on Post hoc Explanations

- Approaches for Post hoc Explainability
- Explanations in Different Modalities
- 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]

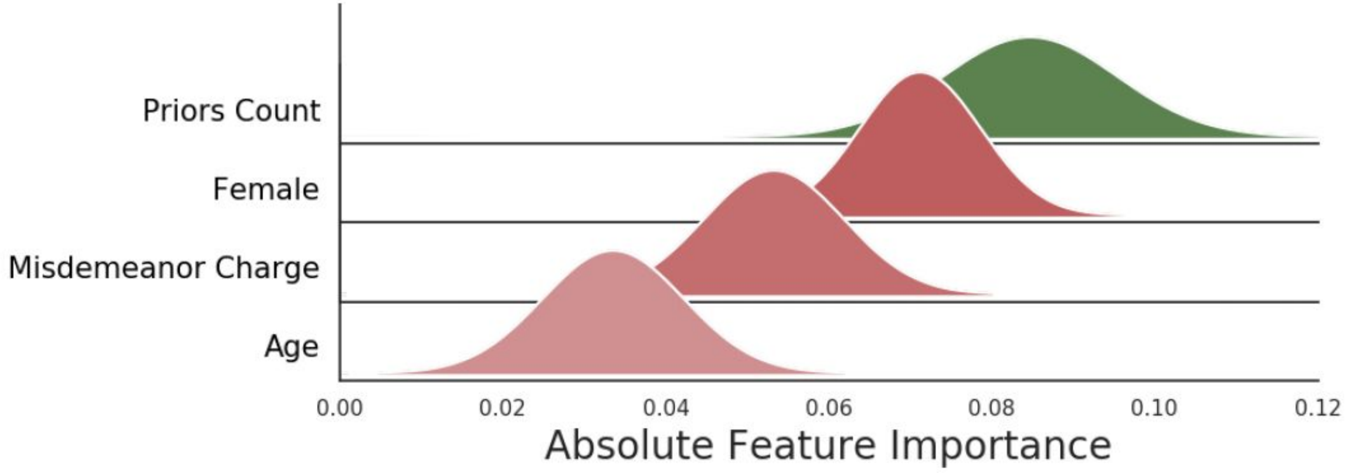
- Generating post hoc explanations that are stable as well as robust to distribution shifts [Chalasani et. al., 2020, Lakkaraju et. al. 2020]

- Generating causal explanations that are faithful to the underlying model [Goyal et. al., 2020]

[Guo et. al., 2018, Slack et. al., 2020]

Modeling Uncertainty in Post hoc Explanations

Model Agnostic



Bayesian versions of LIME/SHAP with closed form solutions

Generate post hoc explanations with user specified confidence levels

I need an explanation where true feature importance lies within ±0.5 of estimated values with 95% confidence

[Lakkaraju et. al., 2020]

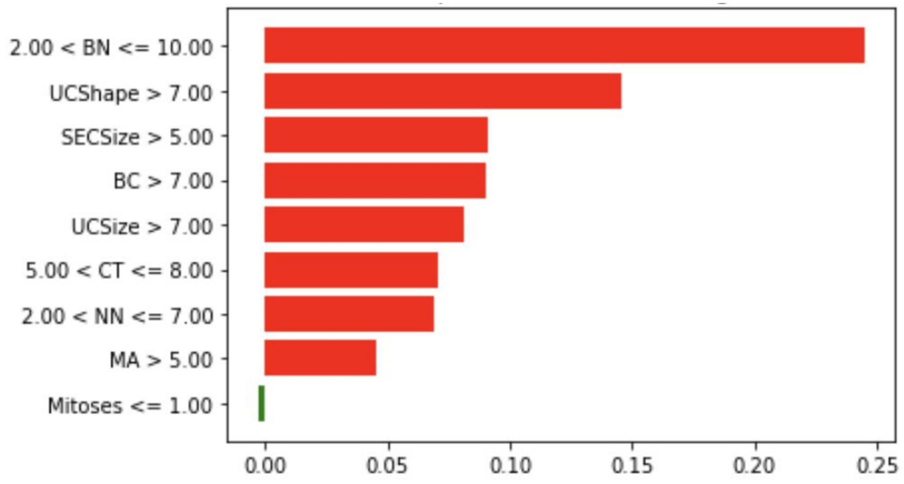
Stable and Robust Post hoc Explanations

- Leverages minimax objective and adversarial training to generate explanations that are stable and robust to distribution shifts

$$\hat{E} = \arg\min_{E \in \mathcal{E}} \max_{\delta \in \Delta} \underbrace{\mathbb{E}_{p_\delta(x)}[\ell(E(x), B^*(x))]}_{\text{mismatch between explanation and black box predictions}}$$

worst-case over distribution shifts

- Generic framework -- can be instantiated to generate *model agnostic local/global explanations* of various types (e.g., feature importances, rules)



If $X_1 < 7.0$ *and* $X_2 < 2.0$, *then Benign*

If $X_2 \geq 5.0$, *then Malignant*

If $X_6 \geq 9.0$, *then Malignant*

If $X_1 \geq 7.0$, *then Malignant*

If $X_4 \geq 4.0$, *then Malignant*

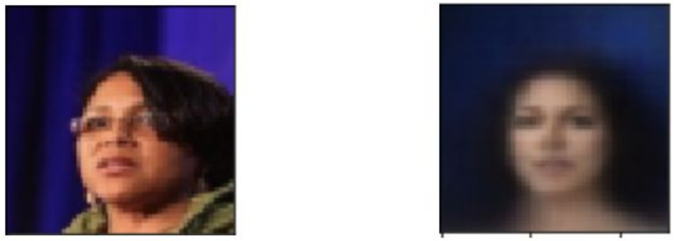
Default Rule (Benign)

Faithful Causal Explanations

ce

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

- causal effects



Future of Post hoc Explainability

Towards Better Post hoc Explanations

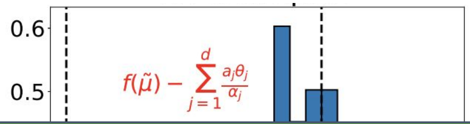
- Methods for More Reliable Post hoc Explanations
- Theoretical Analysis of Post hoc Explanation Methods
- Rigorous Evaluation of the Utility of Post hoc Explanations

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*”

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)

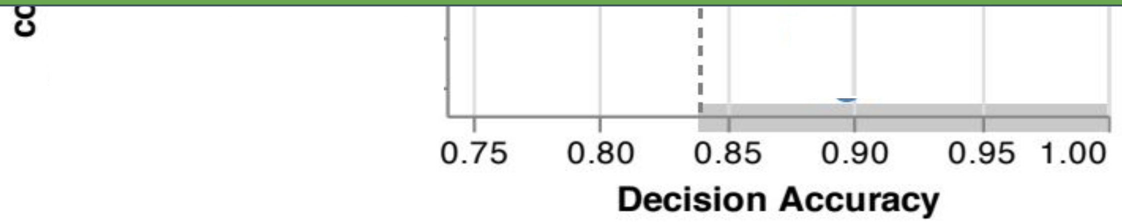
"[The tool] assigns a value that is important to know, but it's showing that in a way that makes you misinterpret that value. Now I want to go back and check all my answers"... [later] "Okay, so, it's not showing me a whole lot more than what I can infer on my own. Now I'm thinking... is this an 'interpretability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

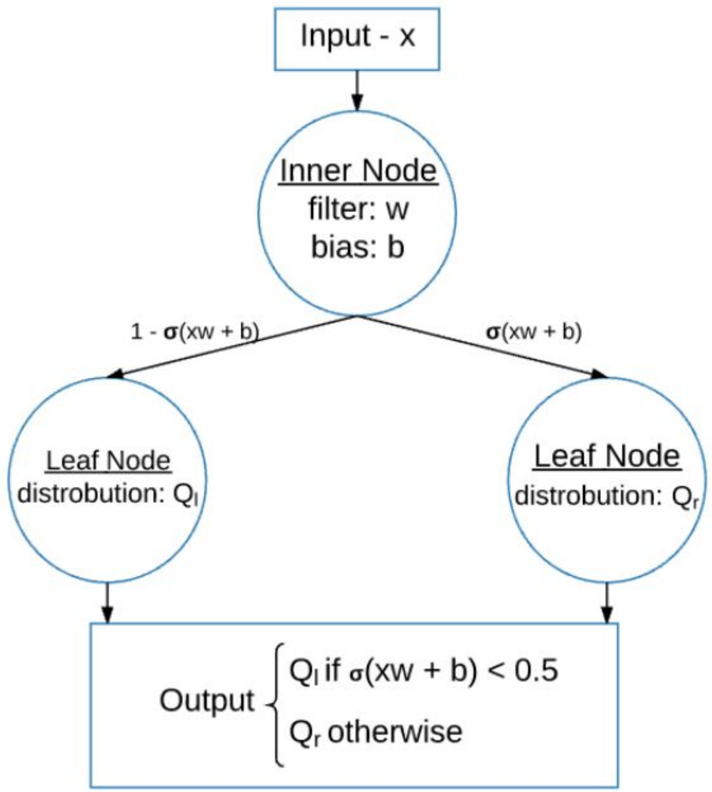
Intersections with Differential Privacy

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

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
in
th

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

[Harder et. al., 2020; Patel et. al. 2020]

Intersections with Differential Privacy

•

•

learning them

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

Future of Post hoc Explainability

Towards Better Post hoc Explanations

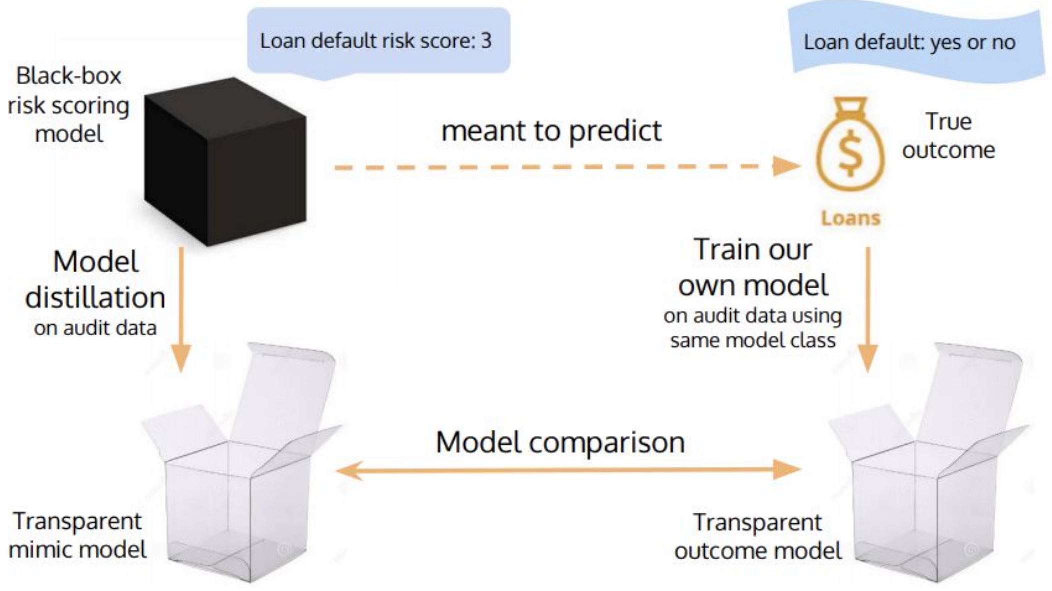
- Methods for More Reliable Post hoc Explanations
- Theoretical Analysis of Post hoc Explanation Methods
- Rigorous Evaluation of the Utility of Post hoc Explanations

Other Emerging Directions

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

Intersections with 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

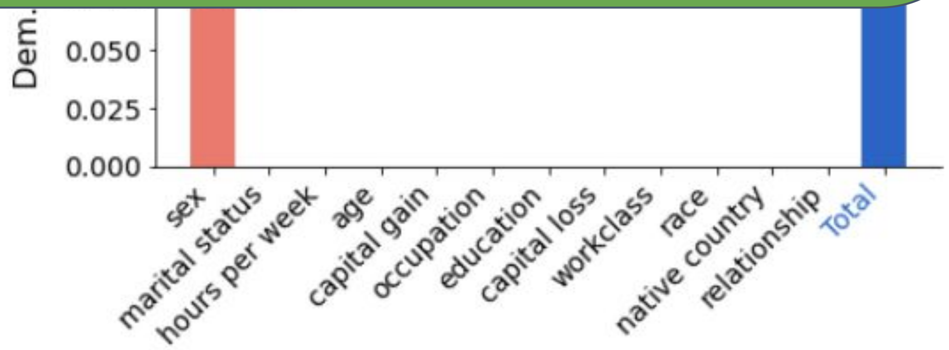
t

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
- Explanations in Different Modalities
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

In Conclusion

Summary of Tutorial

- Motivation for Explainability
- Approaches for Post hoc Explainability
- Explanations in Different Modalities
- Evaluation of Explanations
- Limits of Post hoc Explainability
- Future of Post hoc Explainability

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

Parting Thoughts…

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!

Hima Lakkaraju
Harvard University

Julius Adebayo
MIT

Sameer Singh
UC Irvine

Slides and Video: explainml-tutorial.github.io