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

Sameer Singh

UCI nlp

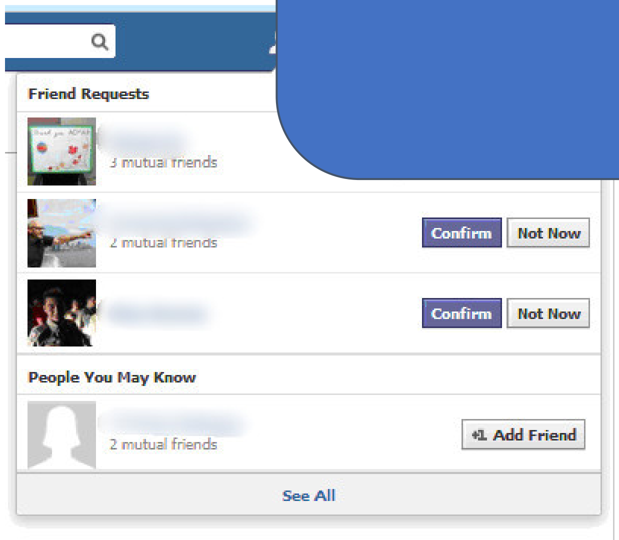
Julius Adebayo
MIT

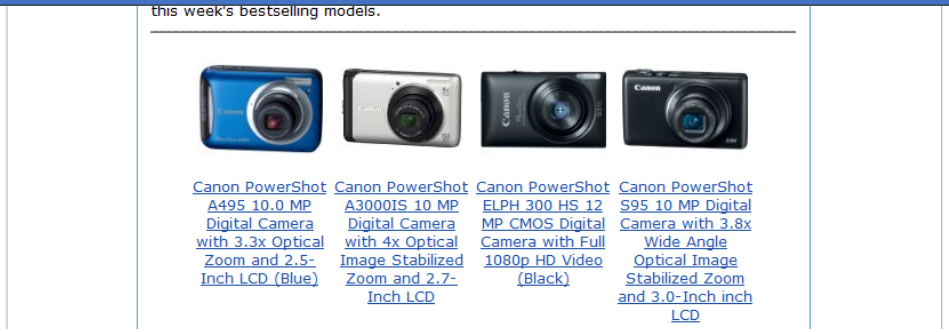
Hima Lakkaraju
Harvard University

Slides and Video: explainml-tutorial.github.io

Motivation

Machine Learning is EVERYWHERE!!





Motivation

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

Motivation: Why Model Understanding?

Input

This model is
incorrect
make
on!! Let
model

Model understanding facilitates debugging.

Predictive Model

Prediction = Siberian Husky

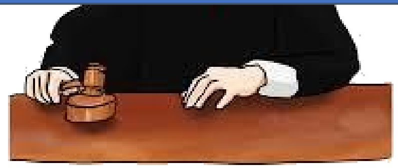
Motivation: Why Model Understanding?

Defendant Detai

Model understanding facilitates bias detection.

This prediction is

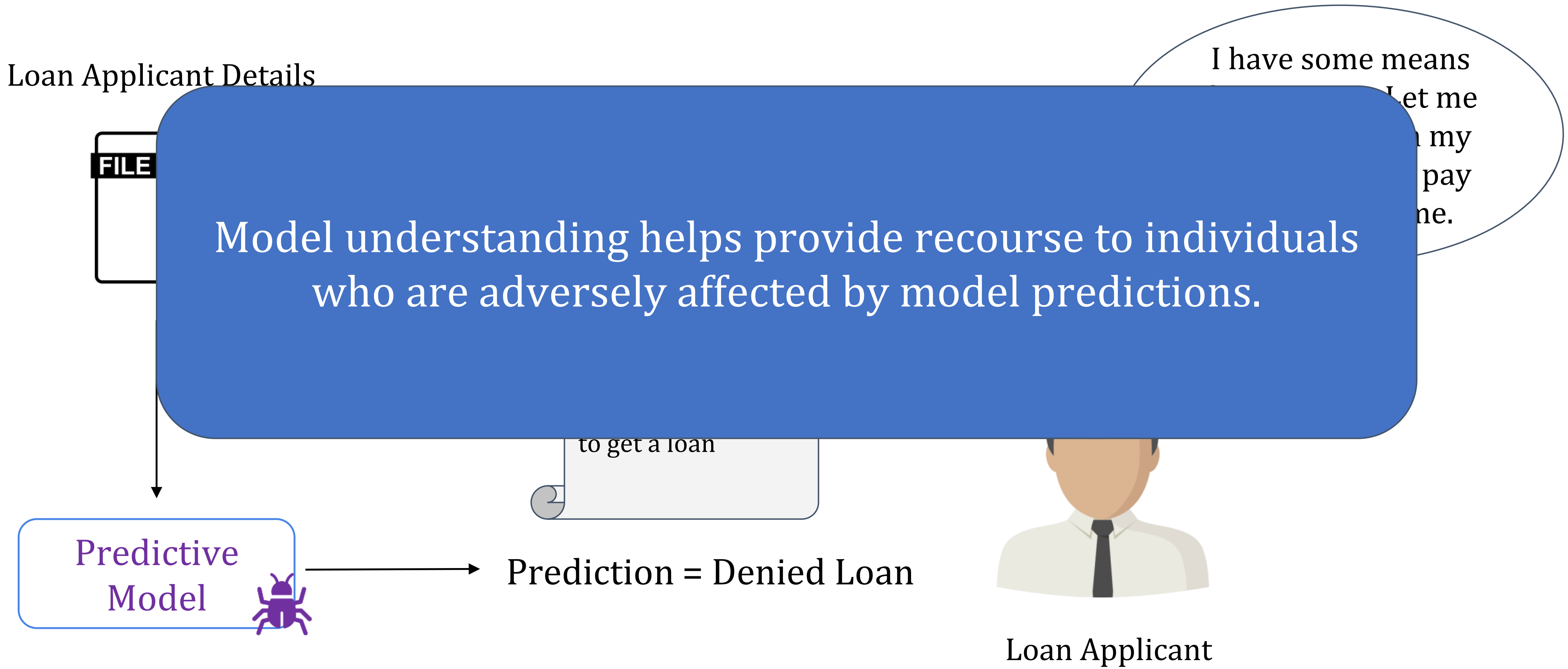
Gender



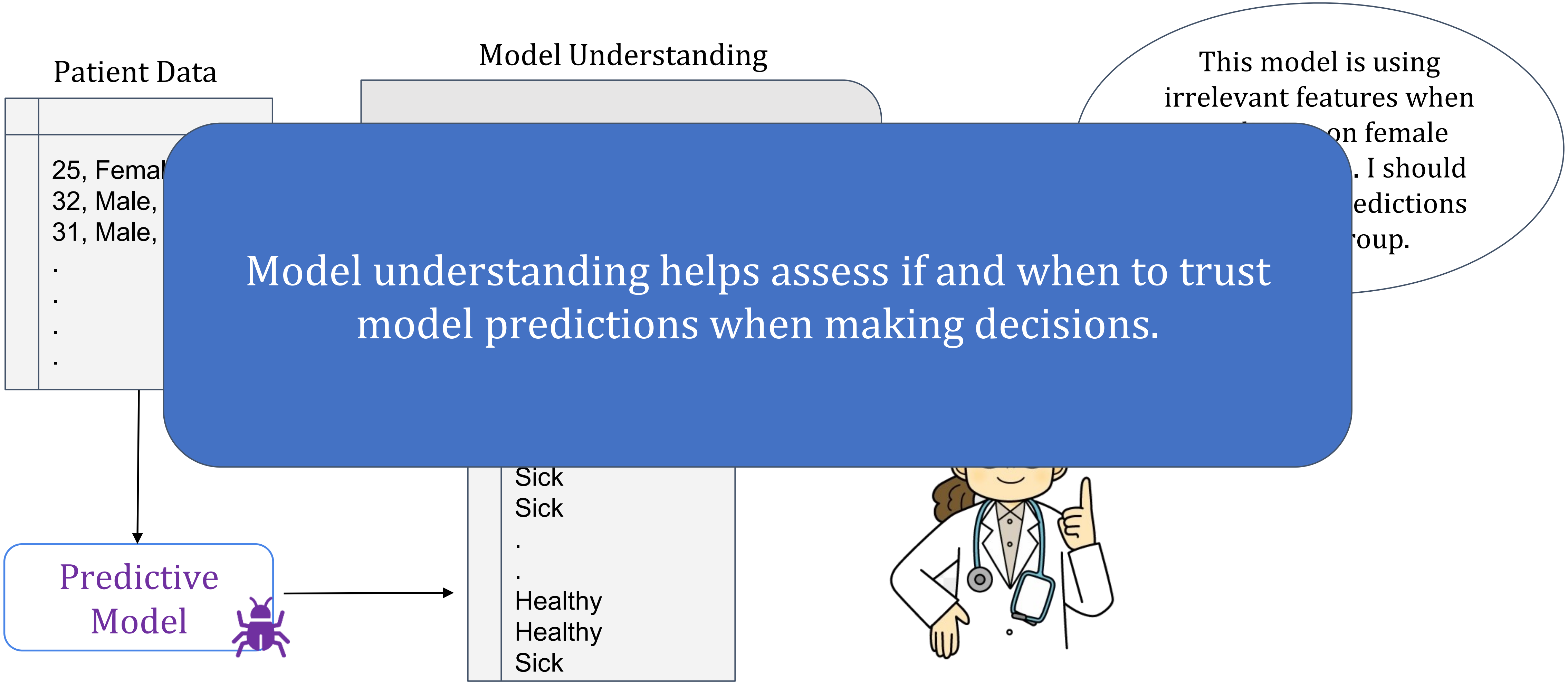
Predictive Model

Prediction = Risky to Release

Motivation: Why Model Understanding?

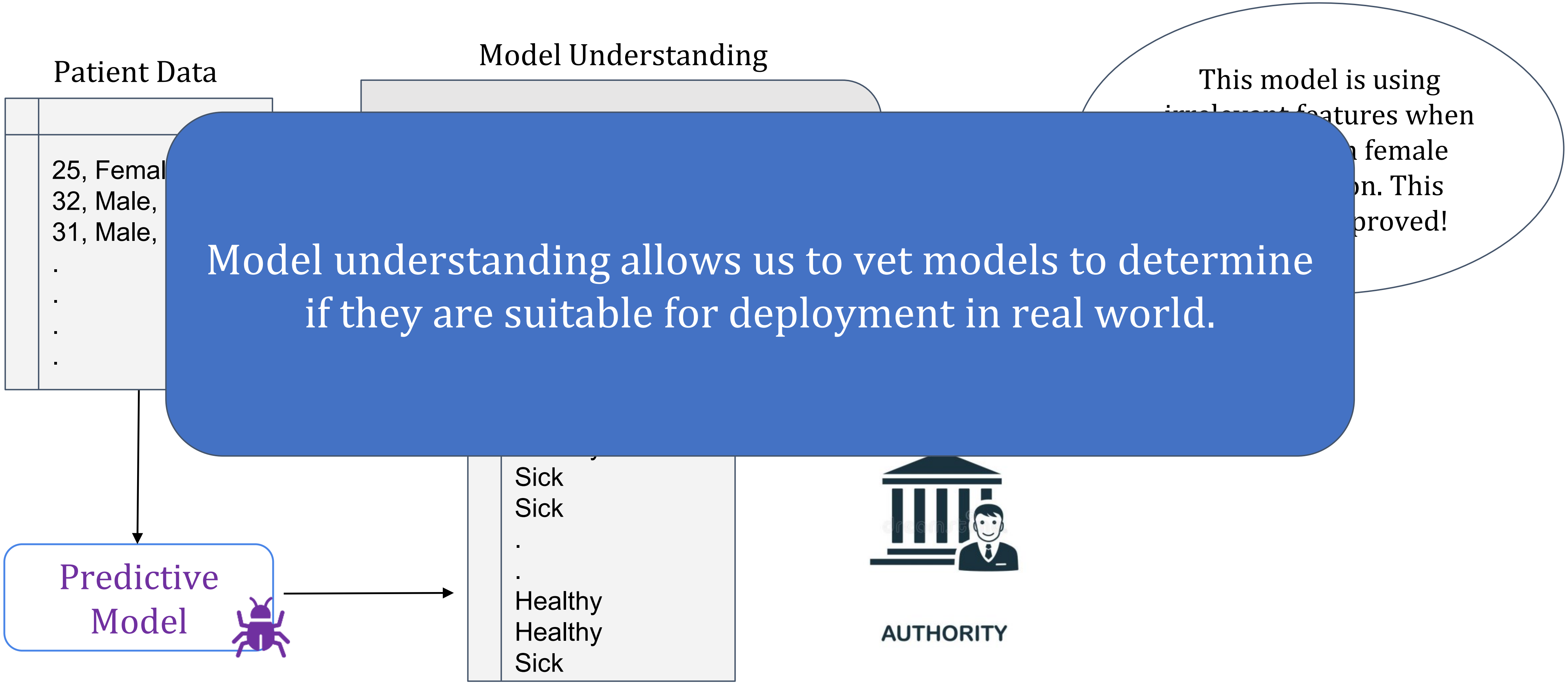


Motivation: Why Model Understanding?



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

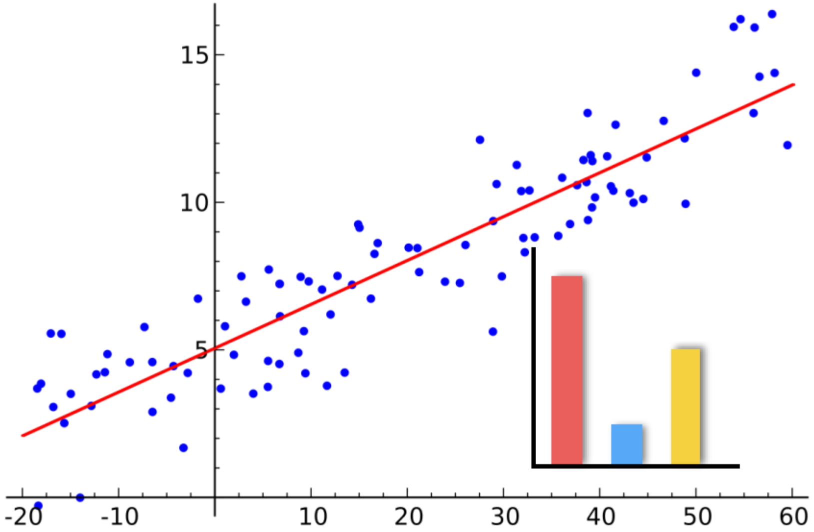
Motivation: Why Model Understanding?

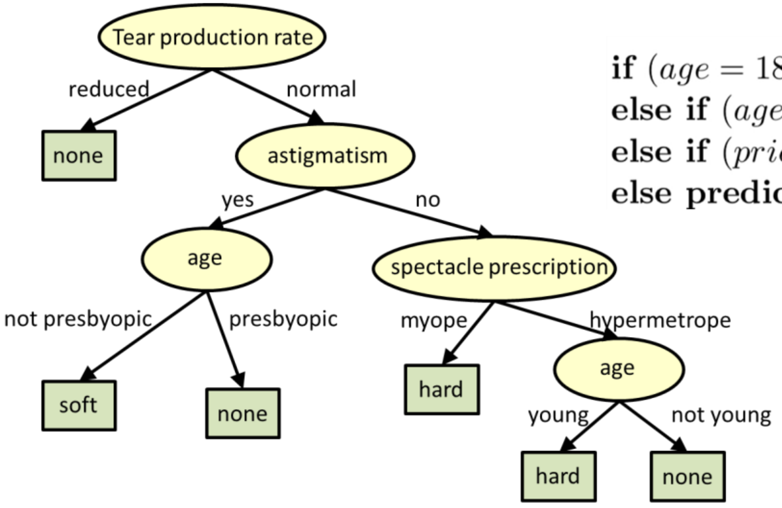


Model understanding allows us to vet models to determine if they are suitable for deployment in real world.

Achieving Model Understanding

Take 1: Build *inherently interpretable* predictive models

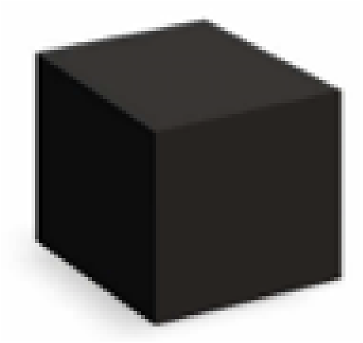


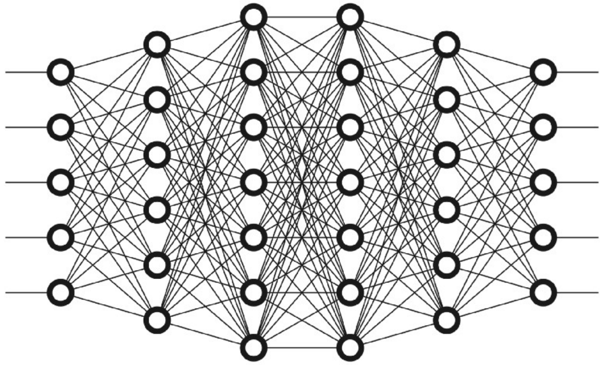


if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

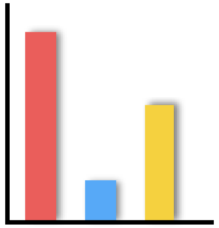
Achieving Model Understanding

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

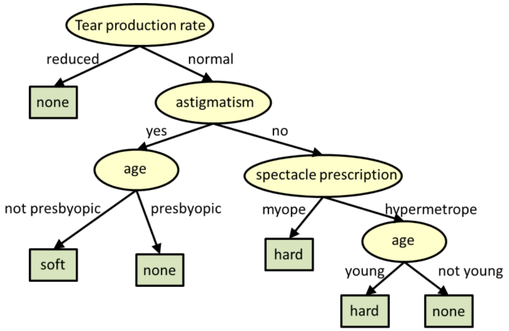




Explainer

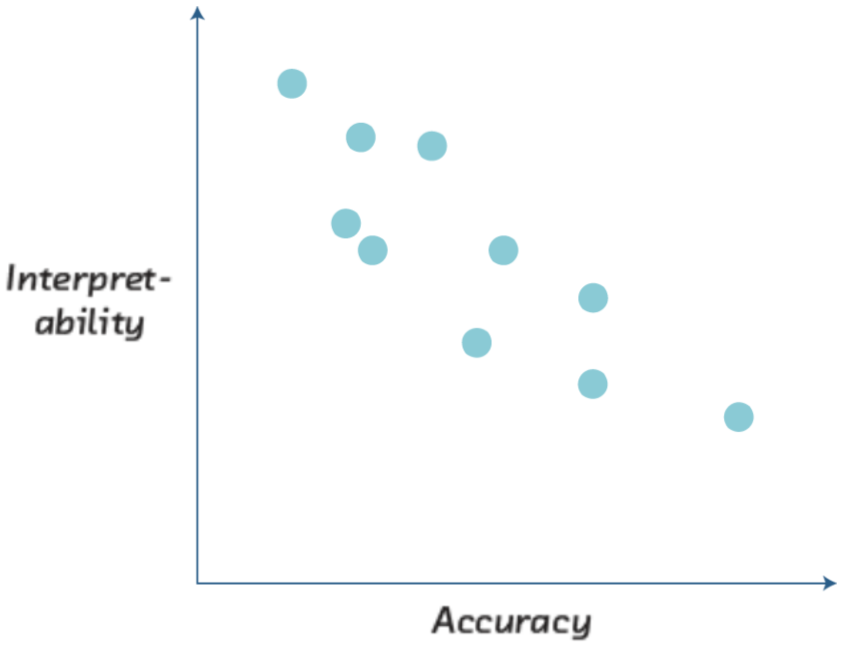


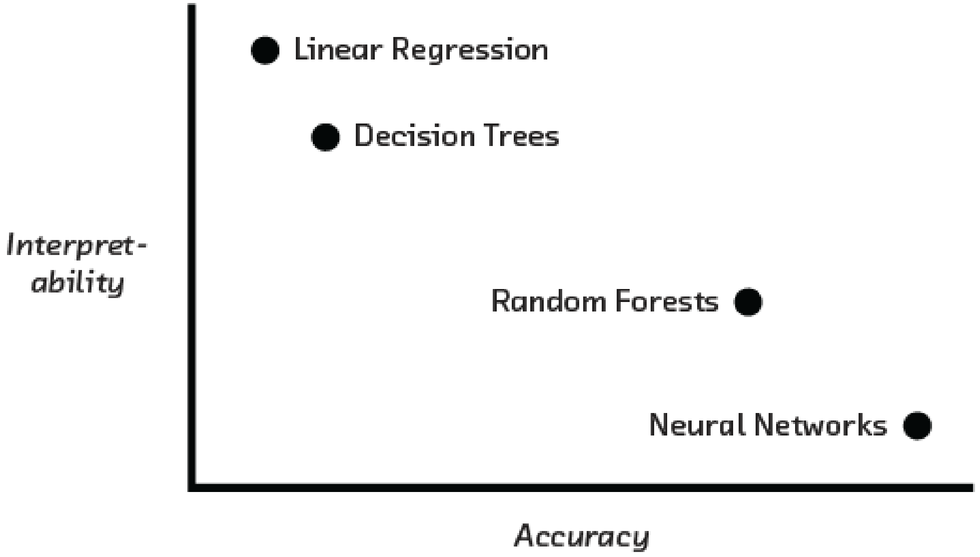
if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no



Inherently Interpretable Models vs. Post hoc Explanations

Example





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

[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

What is an Explanation?

Definition: Interpretable description of the model behavior

Classifier

Send all the model parameters θ?

Send many example predictions?

Summarize with a program/rule/tree

Select most important features/points

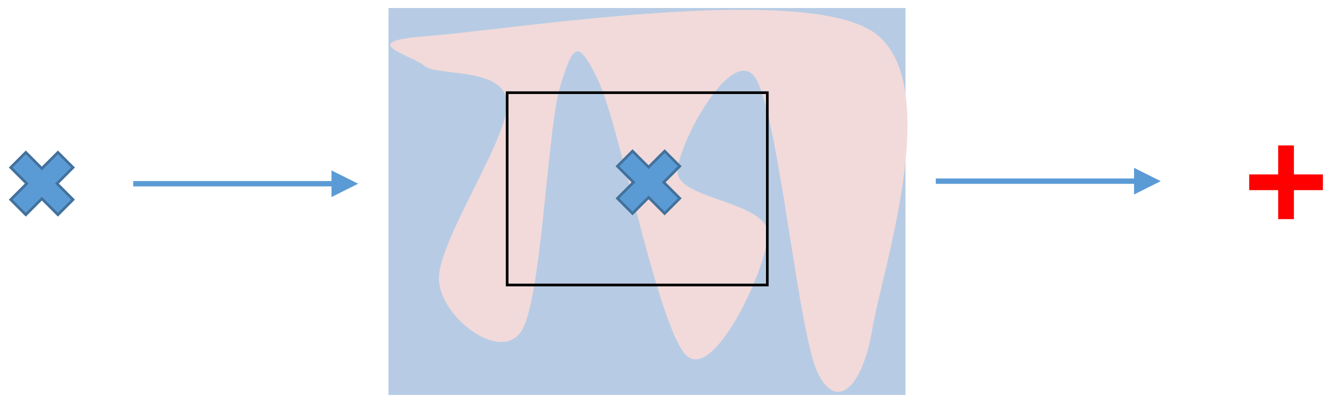
Describe how to *flip* the model prediction

...

User

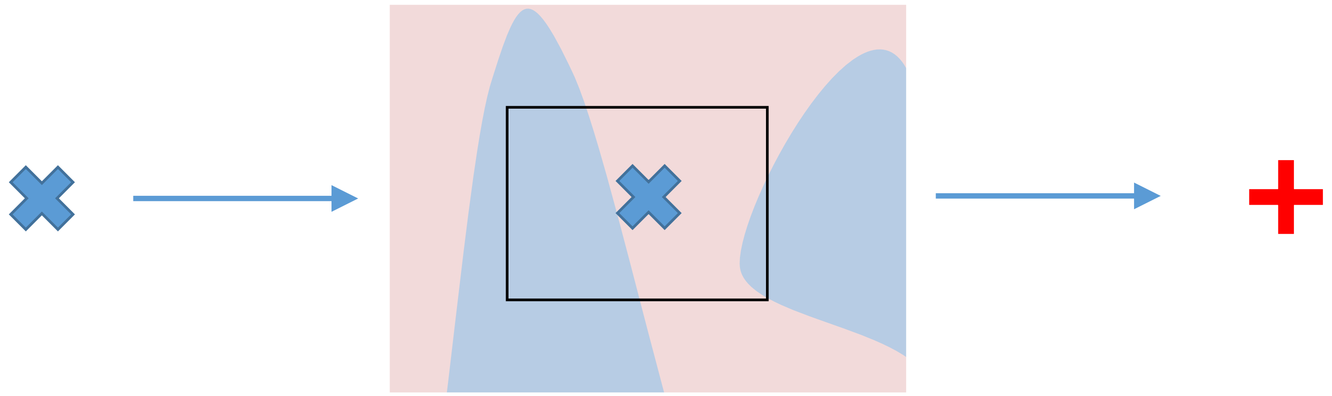
Local versus Global Explanations

Global explanation may be too complicated



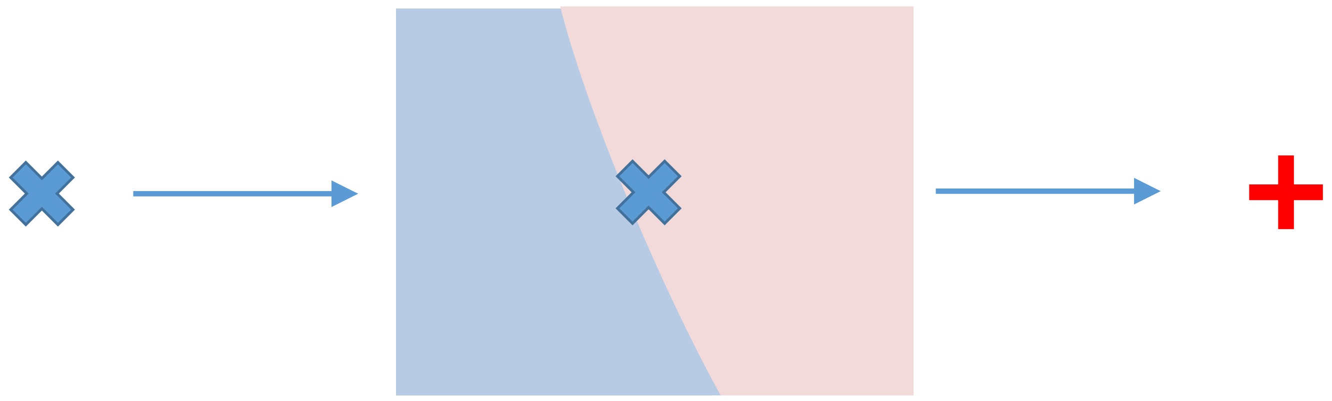
Local versus Global Explanations

Global explanation may be too complicated



Local versus Global Explanations

Global explanation may be too complicated

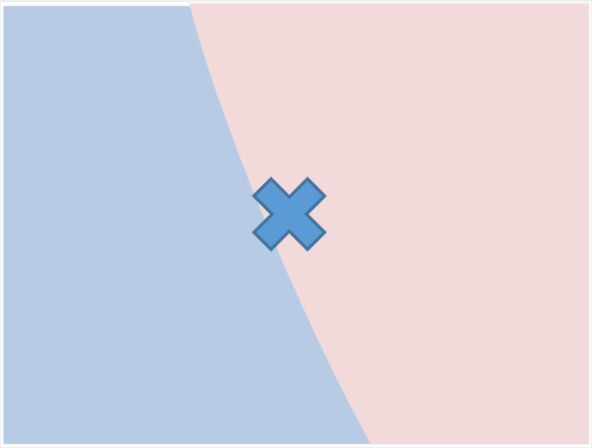


Definition: Interpretable description of the model behavior *in a target neighborhood*.

Local Explanations

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

Classifier



Send many example predictions?

Summarize with a program/rule/tree

Select most important features/points

Describe how to *flip* the model prediction

...

User

Local Explanations vs. Global Explanations

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

Tutorial on Post hoc Explanations

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

Tutorial on Post hoc Explanations

Approaches for Post hoc Explainability

Evaluation of Explanations

Limits of Post hoc Explainability

Future of Post hoc Explainability

Approaches for Post hoc Explainability

Approaches for Post hoc Explainability

Local Explanations

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

Global Explanations

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

Approaches for Post hoc Explainability

Local Explanations

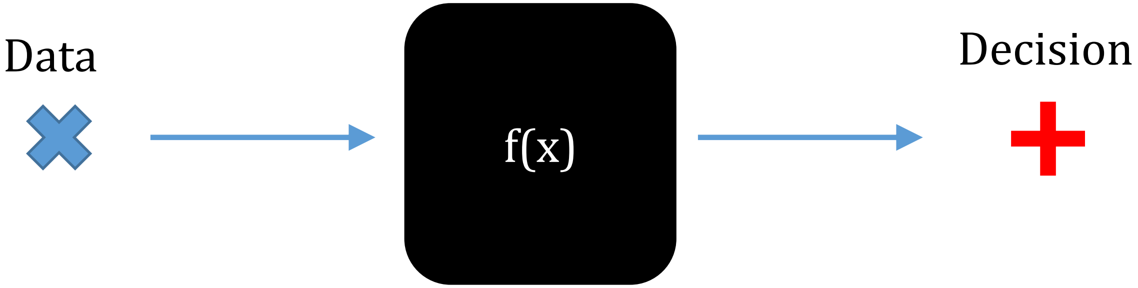
- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

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

Being Model-Agnostic…

No access to the internal structure…



Not restricted to specific models

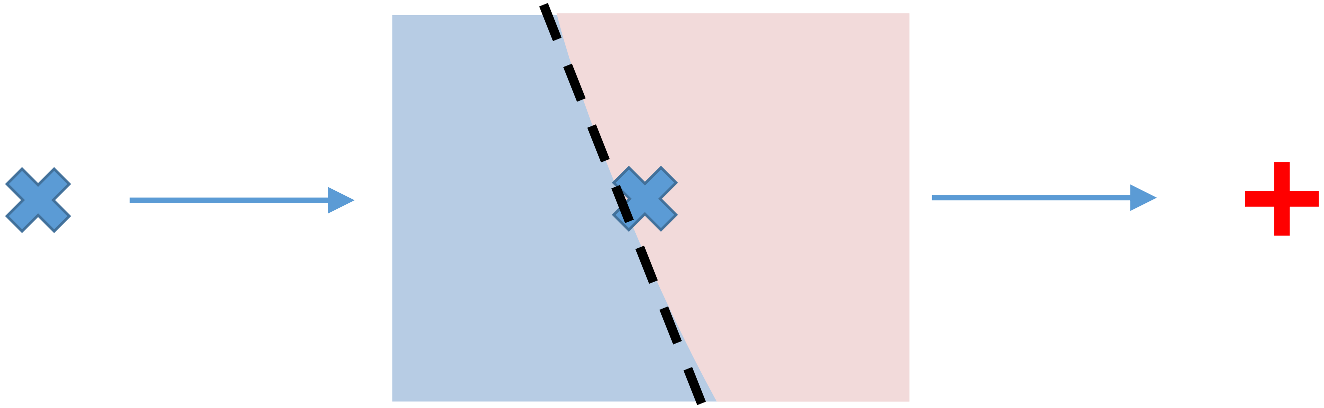
Practically easy: not tied to PyTorch, Tflow, etc.

Study models that you don't have access to!

[Ribeiro et al. 2016]

LIME: Sparse, Linear Explanations

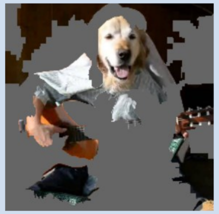
Identify the important dimensions,
and present their relative importance



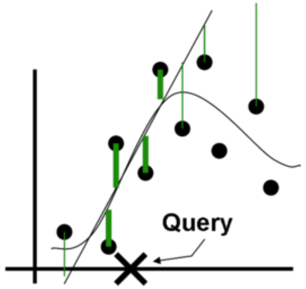
LIME Example - Images

Original Image

P(labrador) = 0.21

Perturbed Instances	P(Labrador)
	0.92
	0.001
	0.34

Maybe to a fault?



Locally weighted regression

Explanation

LIME is quite customizable:

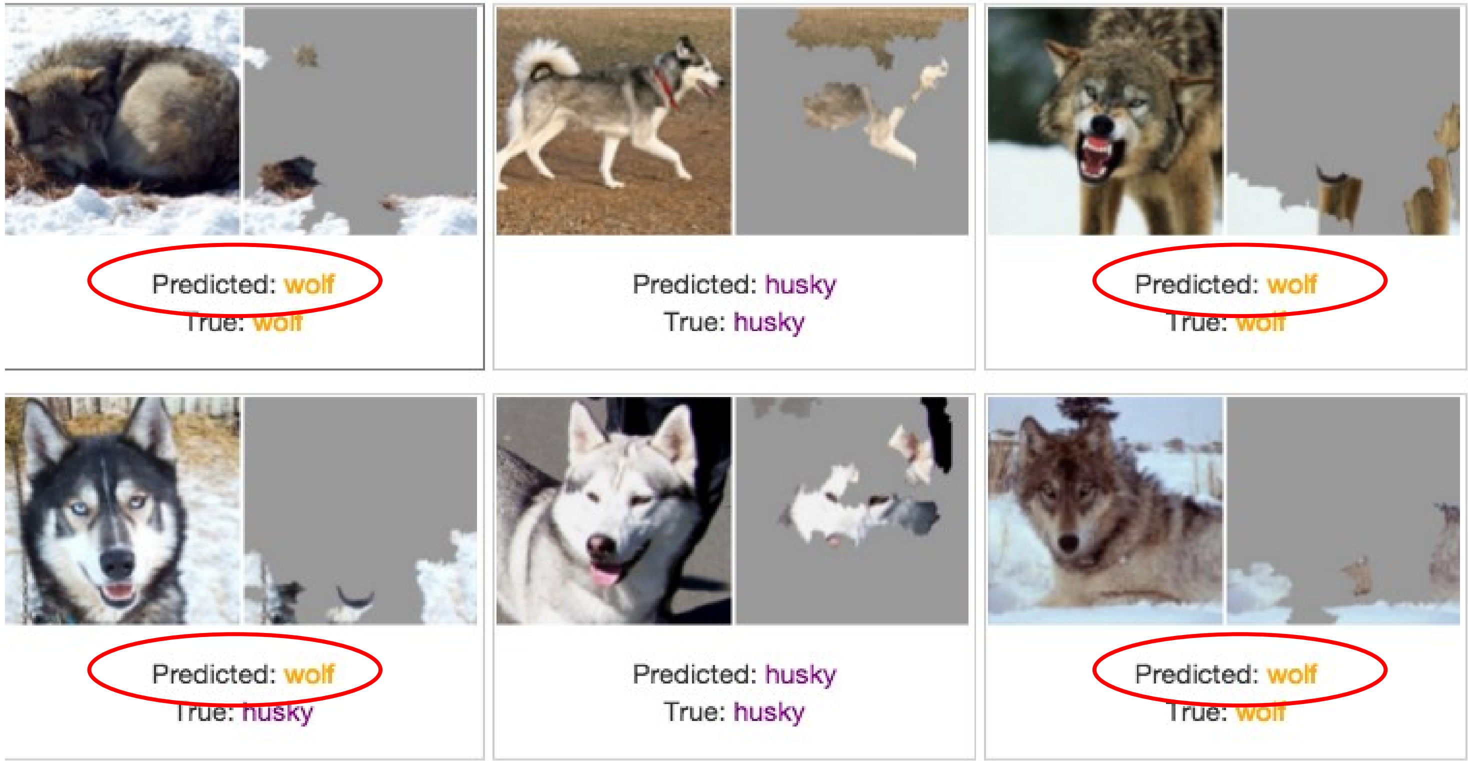
- How to perturb?
- Distance/similarity?
- How *local* you want it to be?
- How to express explanation

[Ribeiro et al. 2016]

Predict Wolf vs Husky

Only 1 mistake!

Predict Wolf vs Husky

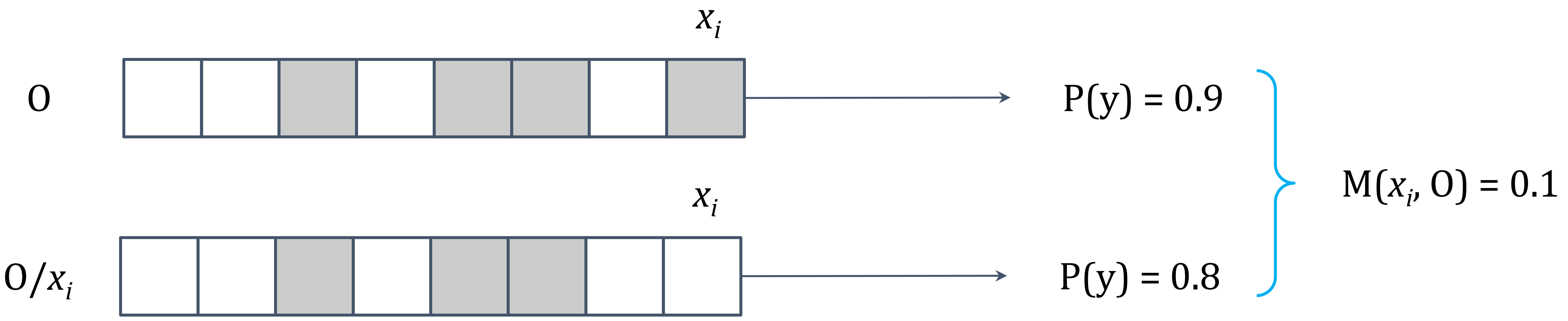


We've built a great snow detector…

[Lundberg & Lee 2017]

SHAP: Shapley Values as Importance

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



Fairly attributes the prediction to all the features.

Approaches for Post hoc Explainability

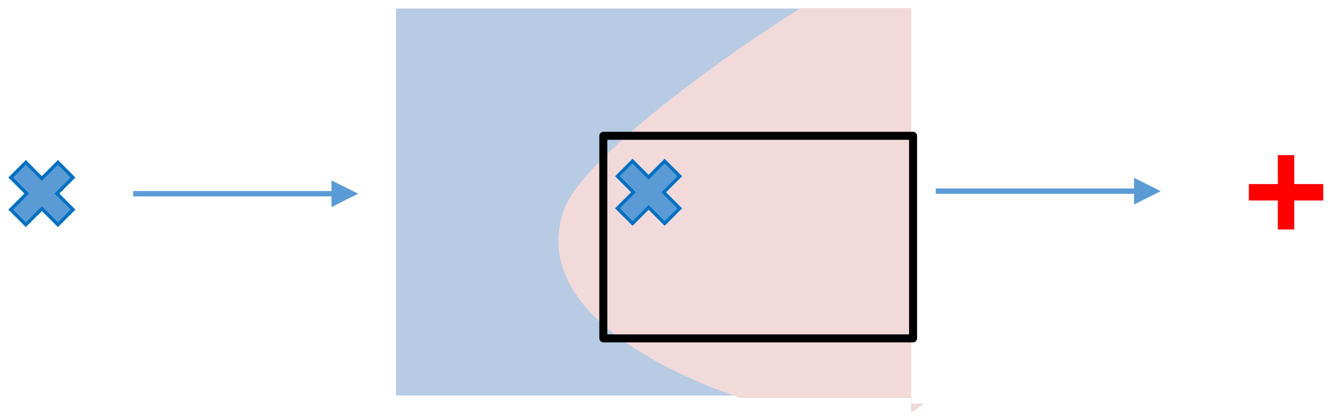
Local Explanations

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

Global Explanations

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

Anchors: Sufficient Conditions



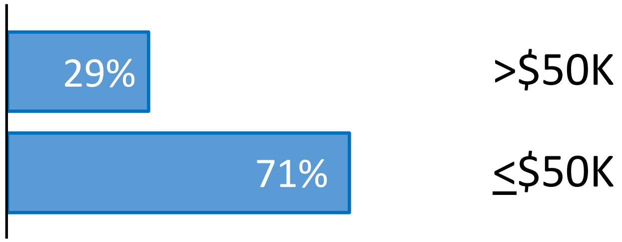
Identify the conditions under which the classifier has the same prediction

[Ribeiro et al. 2018]

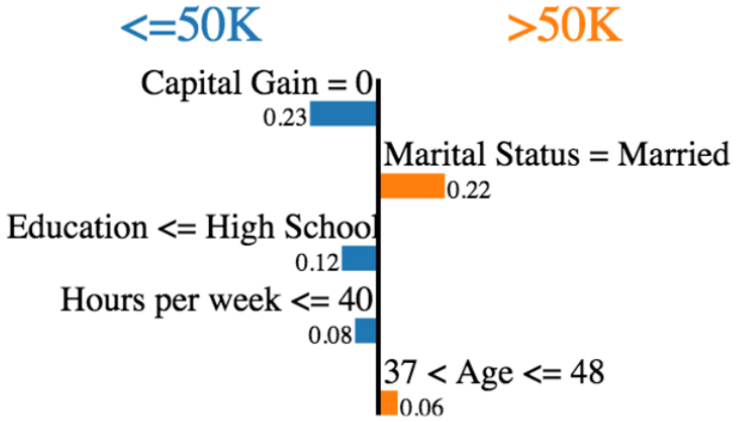
Salary Prediction

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
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

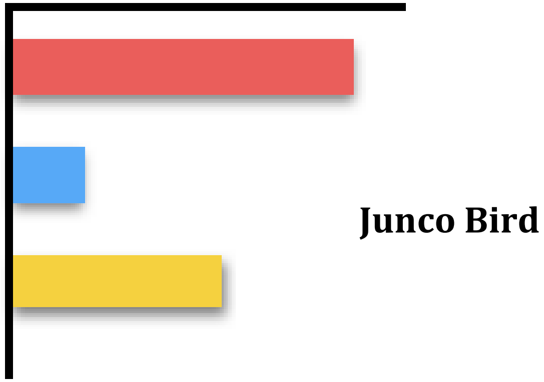
Saliency Map Overview

Input

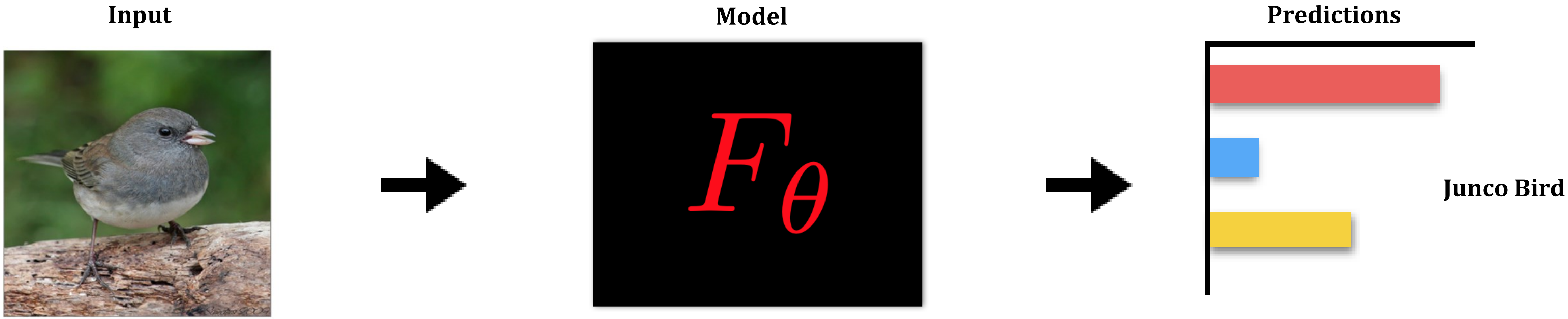
Model

F_θ

Predictions



Saliency Map Overview



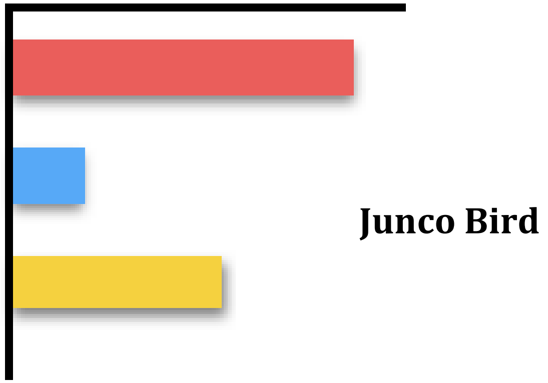
What parts of the input are most relevant for the model's prediction: ‘Junco Bird’?

Saliency Map Overview

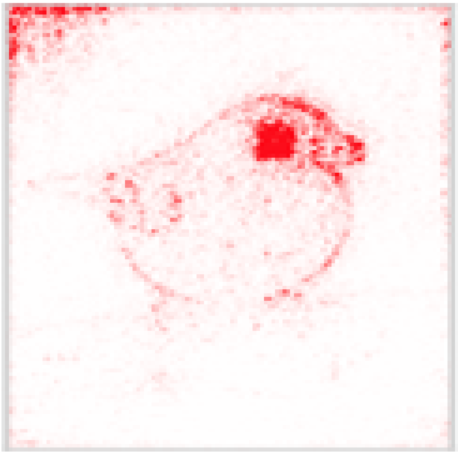
Input

Model

Predictions



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



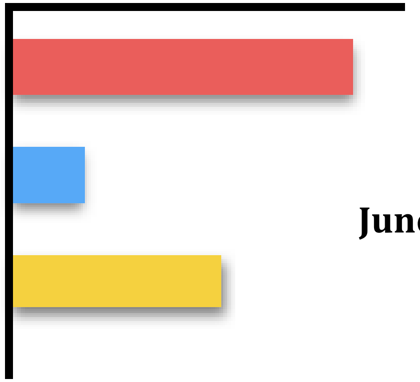
- **Feature Attribution**
- **'Saliency Map'**
- **Heatmap**

Input-Gradient

Input

Model

Predictions



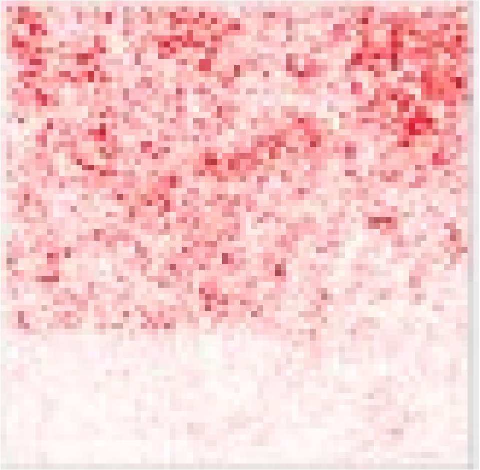
Junco Bird

Input-Gradient

$$\nabla_x F_i(x)$$

Logit

Input



Visualize as a heatmap

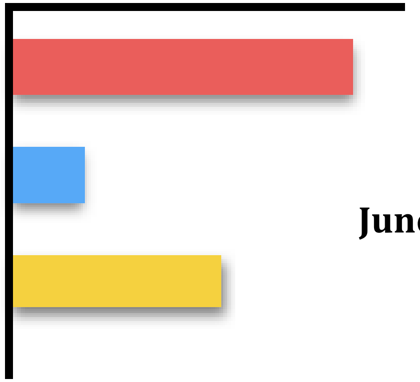
[Baehrens et. al. 2010]; [Simonyan et. al. 2014] .

Input-Gradient

Input

Model

Predictions



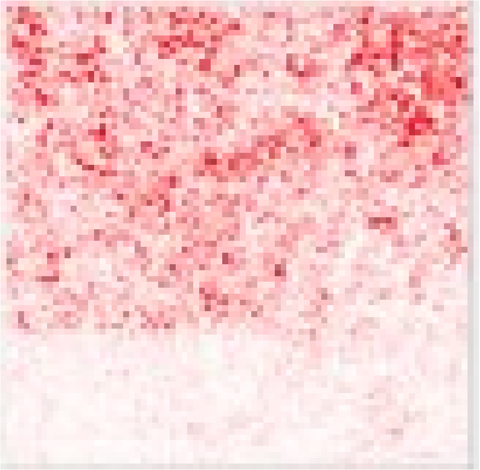
Junco Bird

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 .

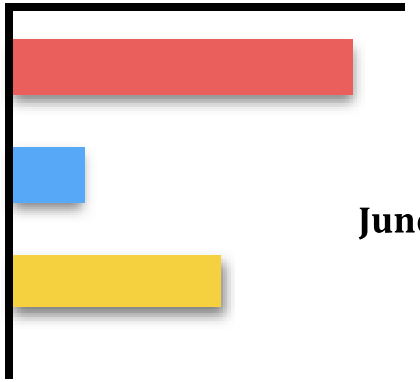
SmoothGrad

Input

Model

F_θ

Predictions

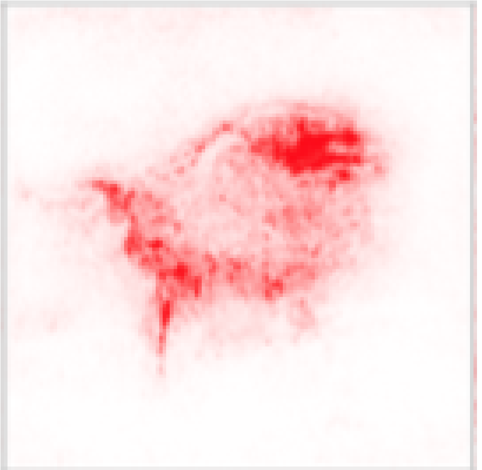


Junco Bird

SmoothGrad

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

Gaussian noise



Average Input-gradient of ‘noisy’ inputs.

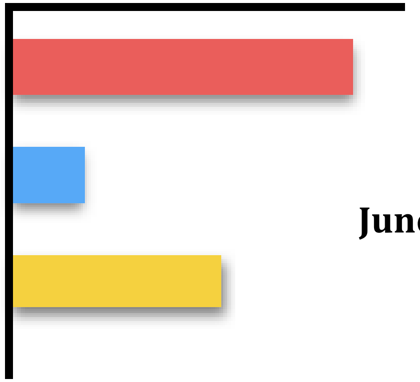
[Smilkov et. al. 2017](#)

Integrated Gradients

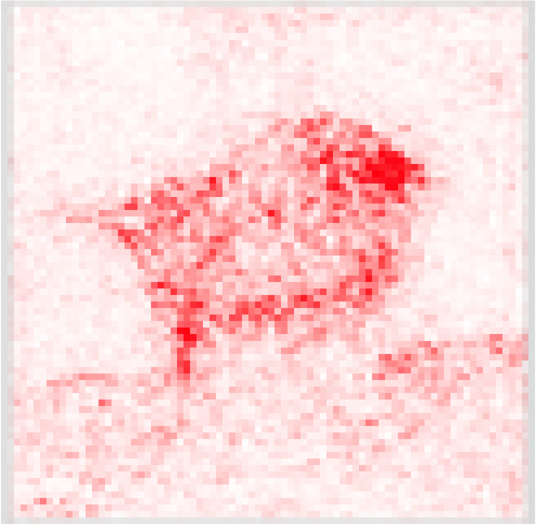
Input

Model

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

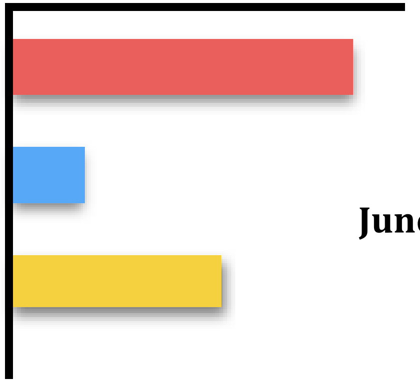
[Sundararajan et. al. 2017]

Recap

Input

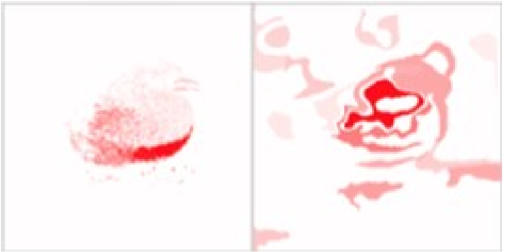
Model

Predictions



Junco Bird

LIME SHAP

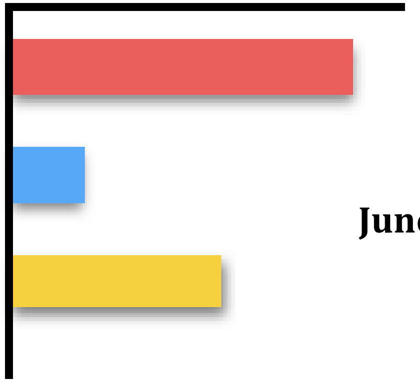


Recap

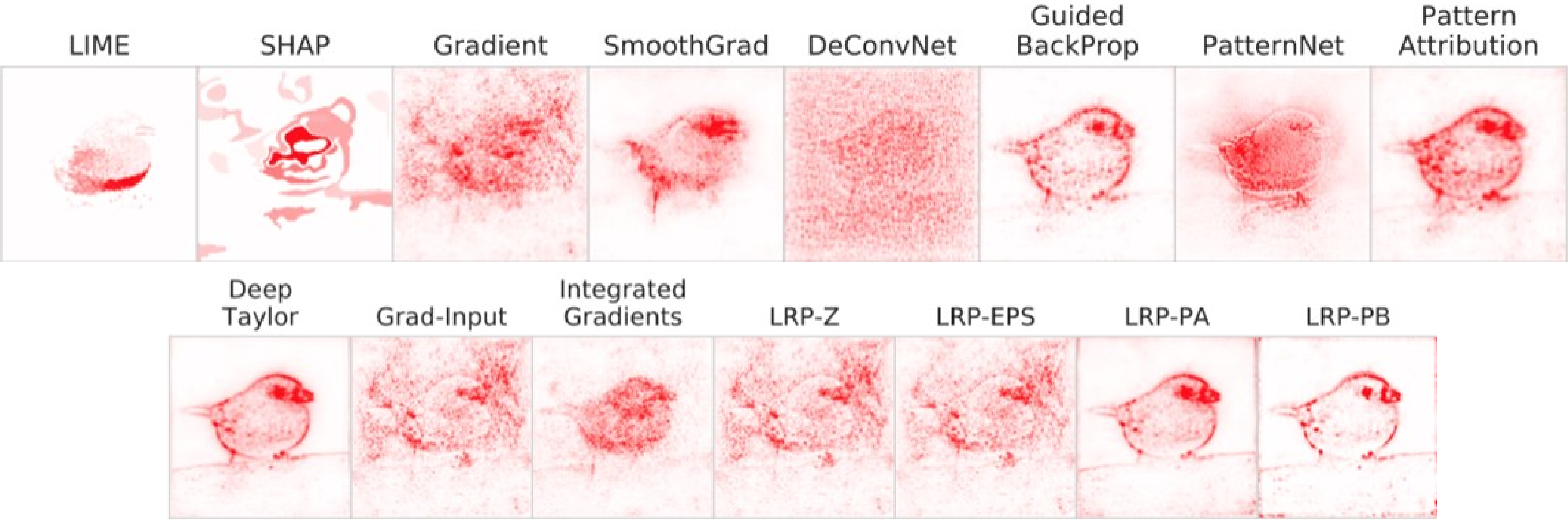
Input

Model

Predictions



Junco Bird



Approaches for Post hoc Explainability

Local Explanations

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

Global Explanations

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

Prototype Approaches

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

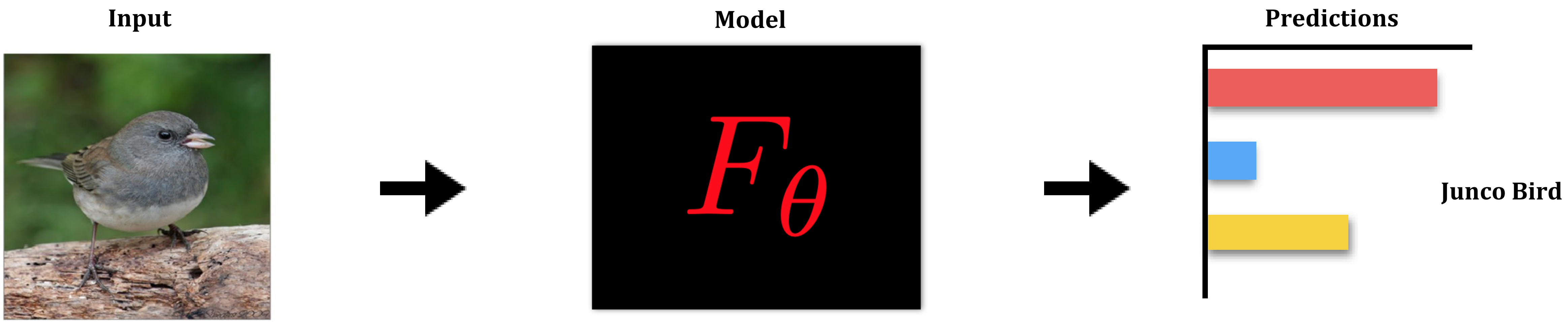
Prototype Approaches

Explain a model with synthetic or natural input '**examples**'.

Insights

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

Training Point Ranking via Influence Functions



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

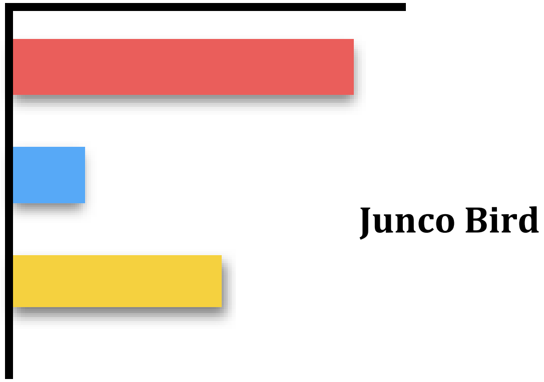
[Koh & Liang 2017](#) ; [Yeh et. al. 2018](#) ; [Pruthi et. al. 2020](#)

Training Point Ranking via Influence Functions

Input

Model

Predictions



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

[Koh & Liang 2017](#) ; [Yeh et. al. 2018](#) ; [Pruthi et. al. 2020](#)

Approaches for Post hoc Explainability

Local Explanations

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

Global Explanations

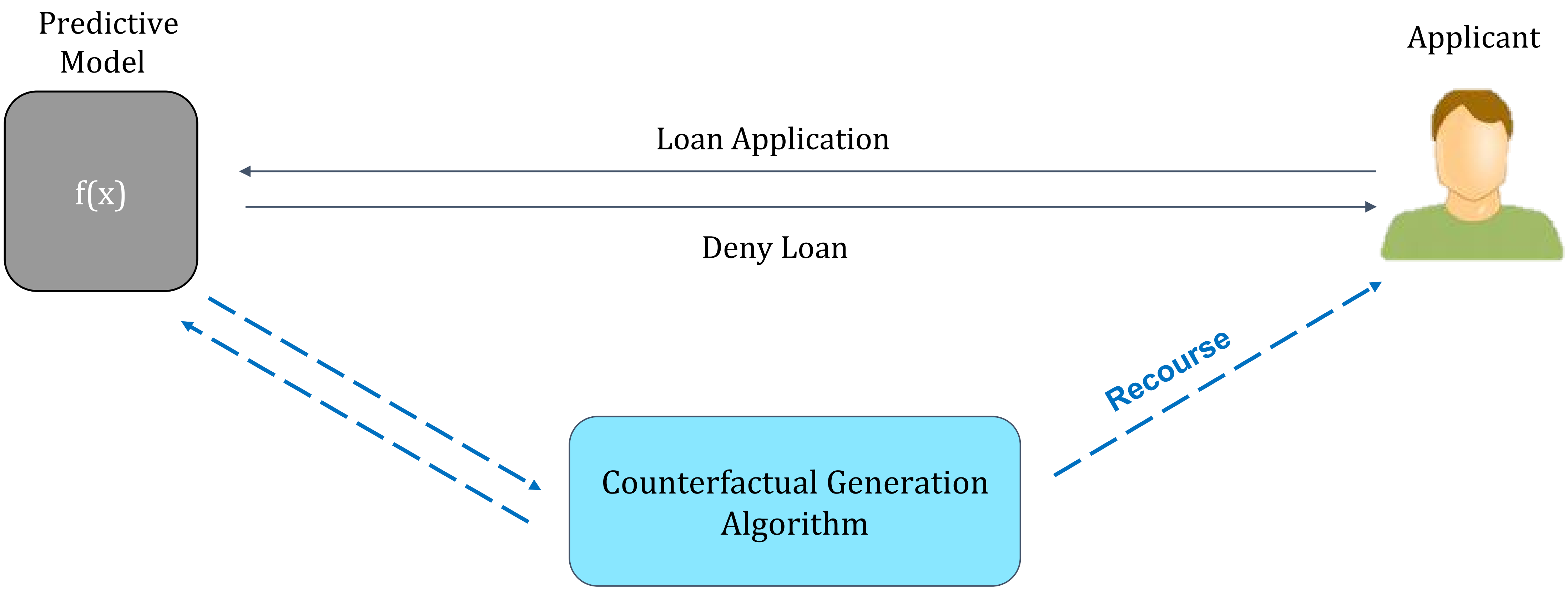
- Collection of Local Explanations
- Model Distillation
- Summaries of Counterfactuals
- Representation Based

Counterfactual Explanations

It's important to provide recourse to affected individuals.

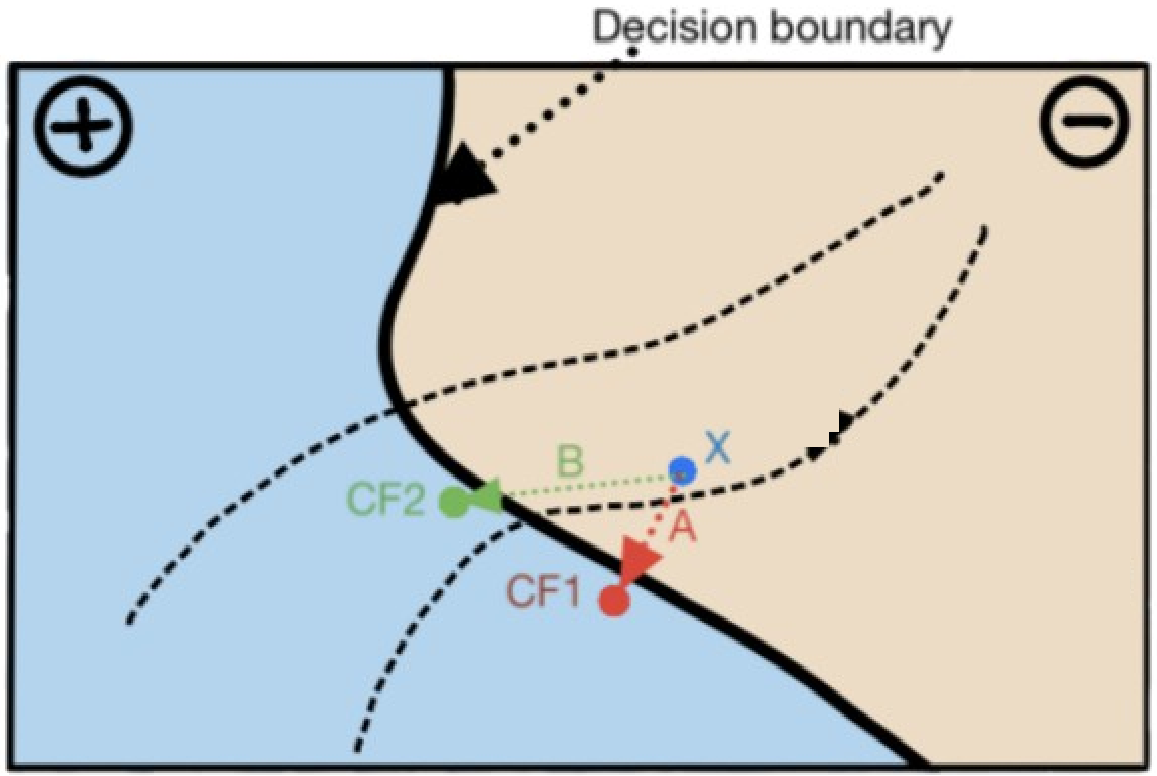
Counterfactual Explanations
What features need to be changed and by how much to flip a model's prediction ?
(i.e., to reverse an unfavorable outcome).

Counterfactual Explanations



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

Counterfactual Explanations: Intuition



Proposed solutions differ on:

How to choose among candidate counterfactuals?

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

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

Approaches for Post hoc Explainability

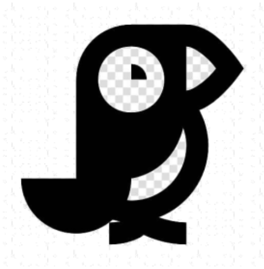
Local Explanations

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

Global Explanations

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

Global Explanations

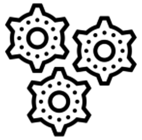


- Explain the complete behavior of a given (black box) model
  - Provide a *bird's eye view* of model behavior

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

- Global explanations are complementary to local explanations

Tutorial on Post hoc Explanations



Approaches for Post hoc Explainability

Evaluation of Explanations

Limits of Post hoc Explainability

Future of Post hoc Explainability

Tutorial on Post hoc Explanations

Approaches for Post hoc Explainability

Evaluation of Explanations

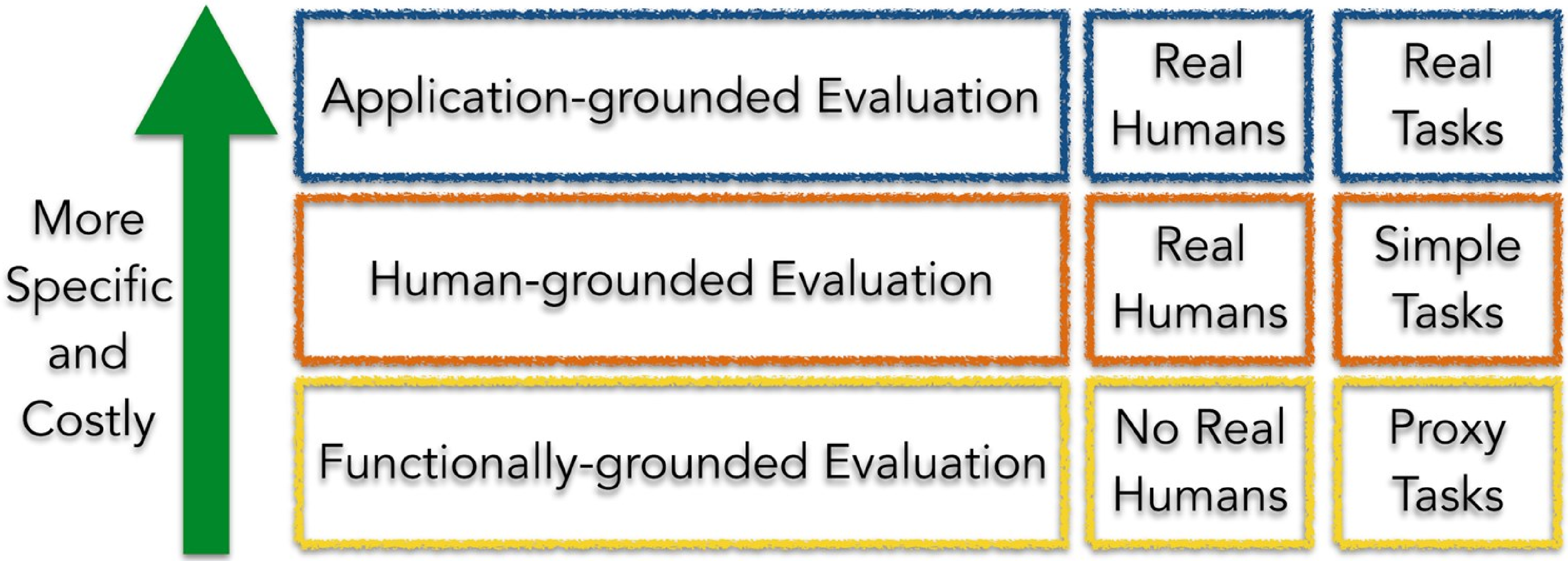
Limits of Post hoc Explainability

Future of Post hoc Explainability

Evaluation of Post hoc Explanations

[Doshi-Velez and Kim, 2017]

How we evaluate explanations?



Evaluating Post hoc Explanations

Understand the Behavior

Help make decisions

Useful for Debugging

Evaluating Post hoc Explanations

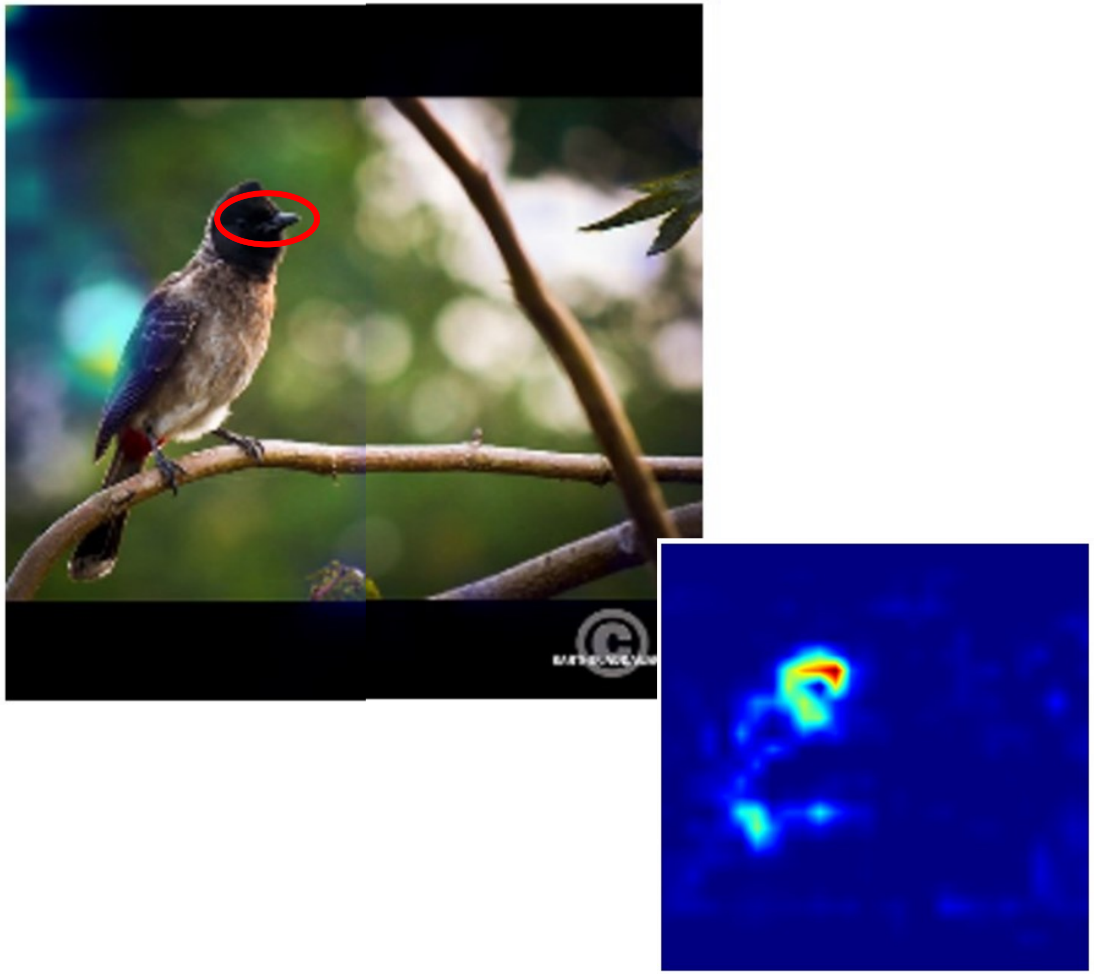
Understand the Behavior

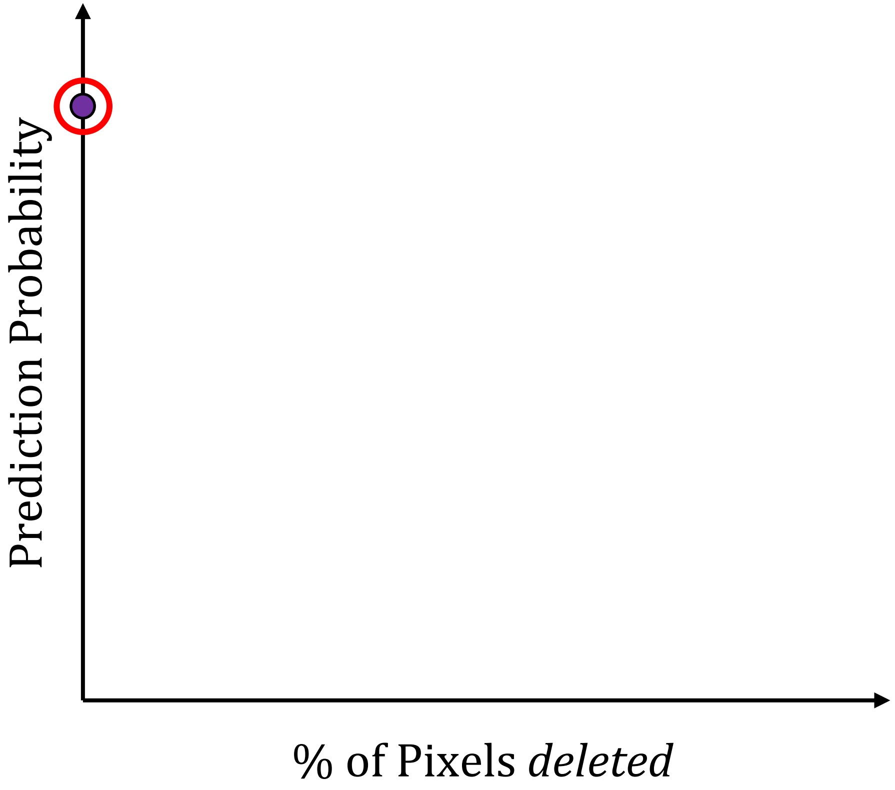
Help make decisions

Useful for Debugging

How important are selected features?

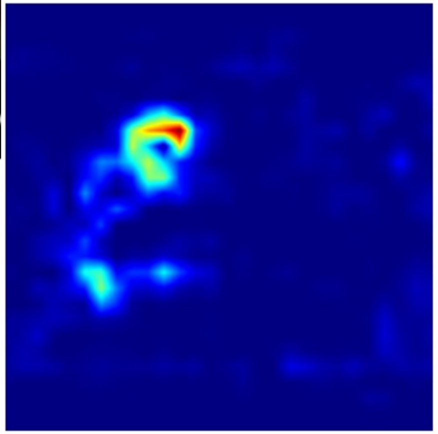
- Deletion: remove important features and see what happens..

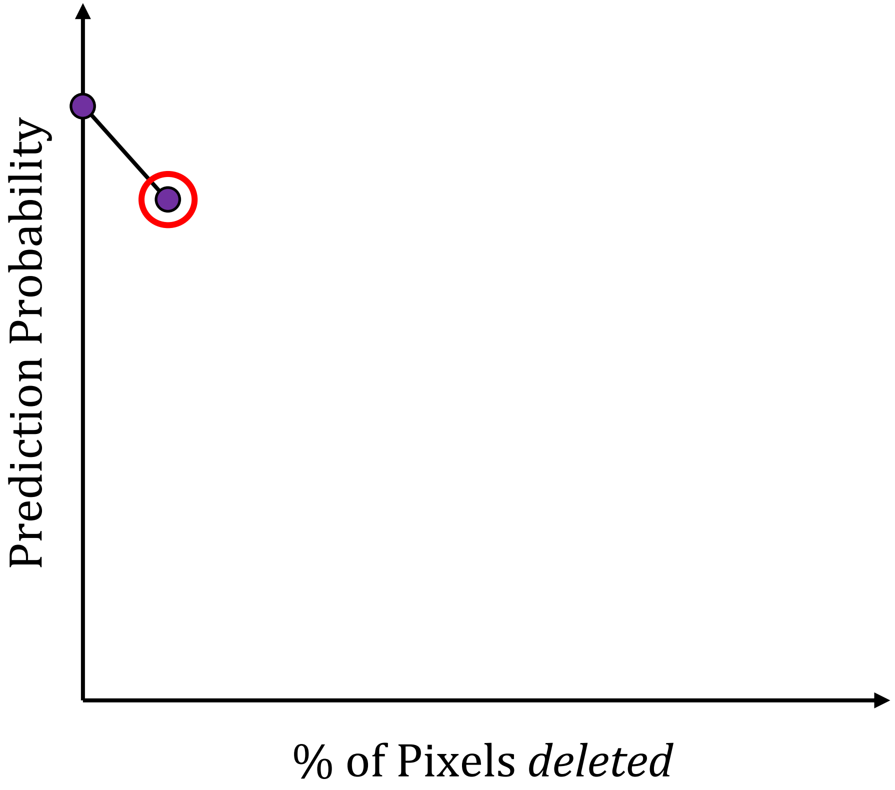




How important are selected features?

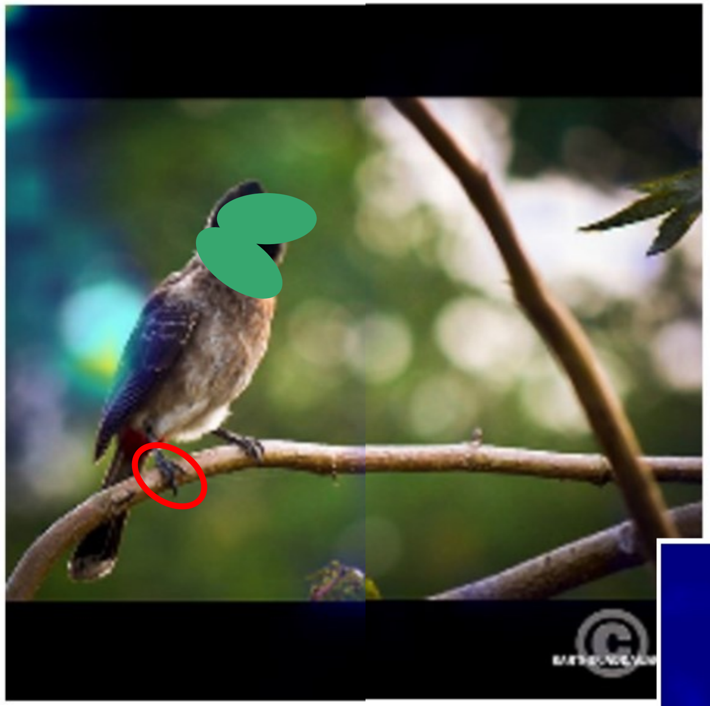
- Deletion: remove important features and see what happens..

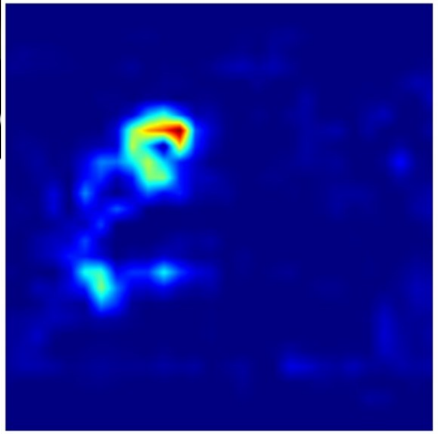


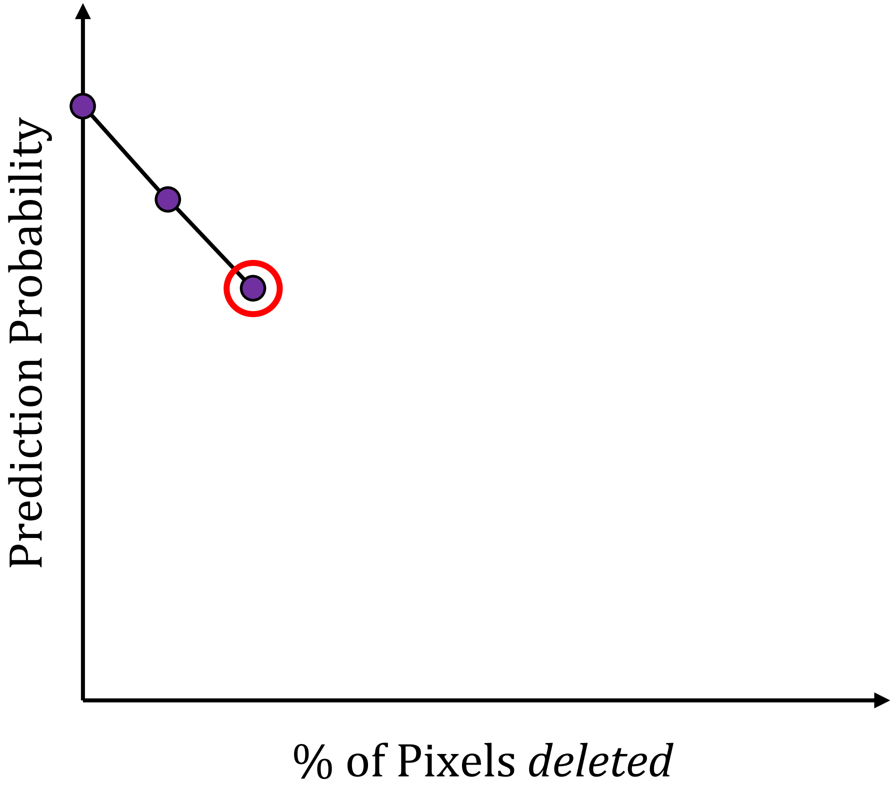


How important are selected features?

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

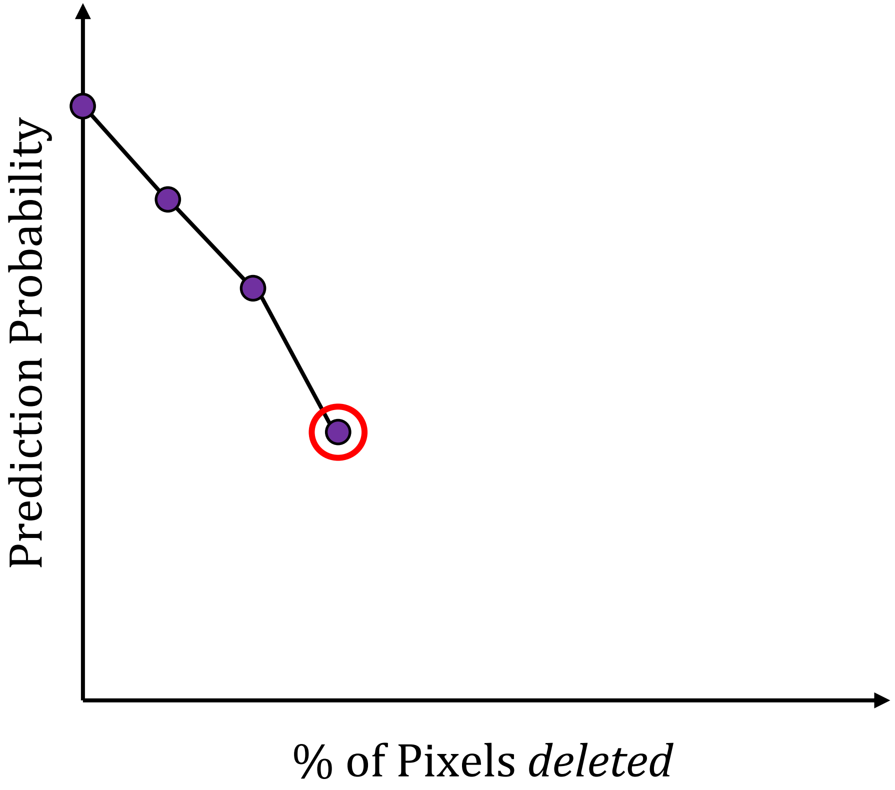






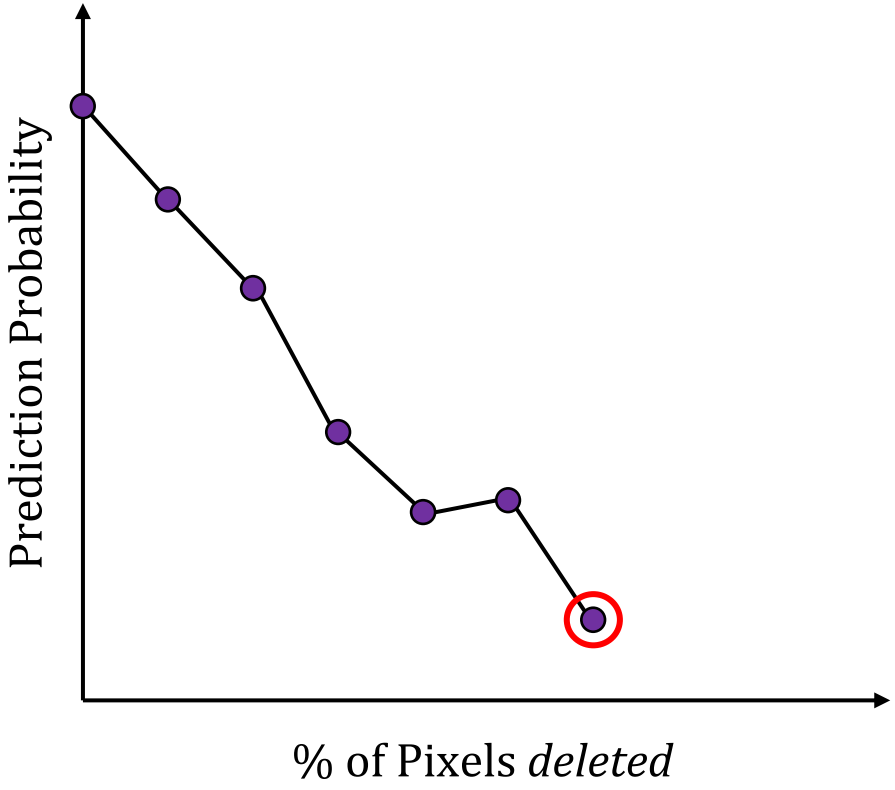
How important are selected features?

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



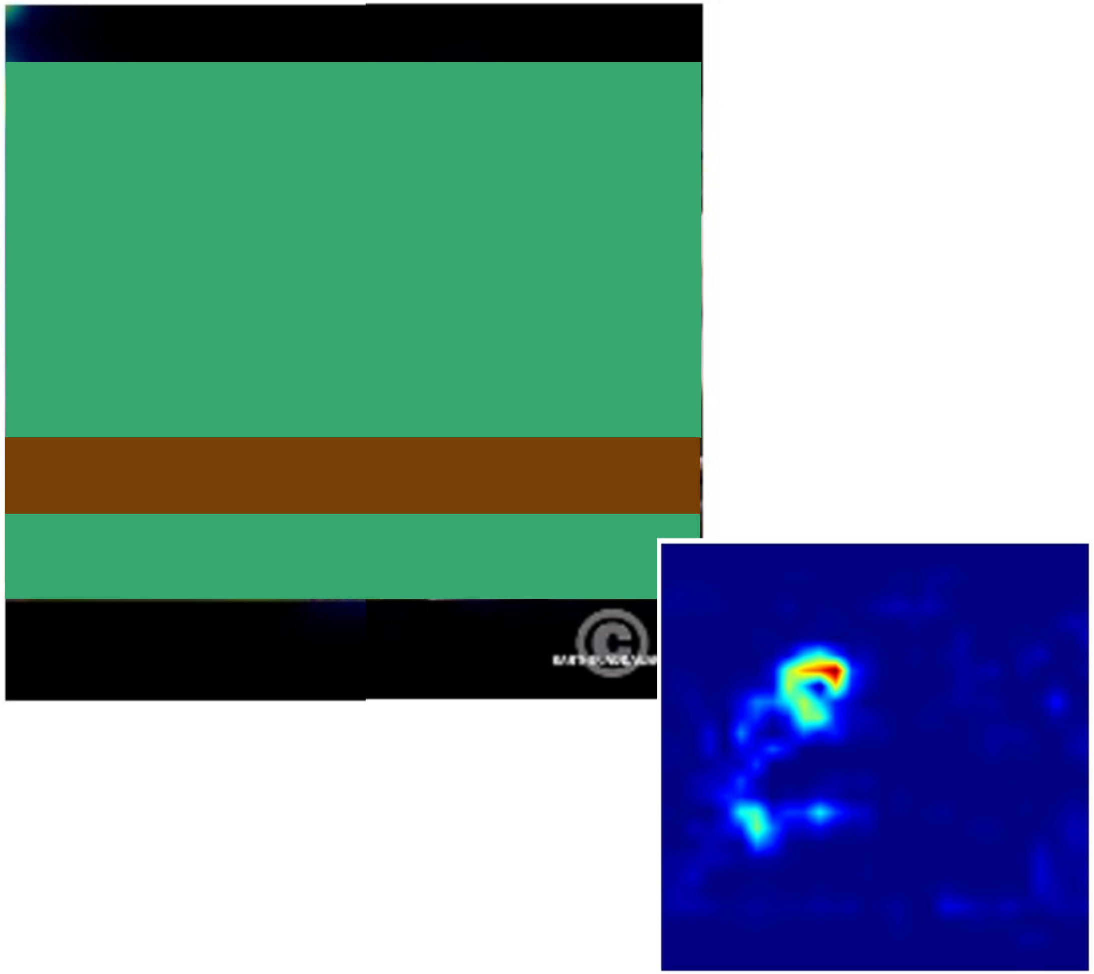
How important are selected features?

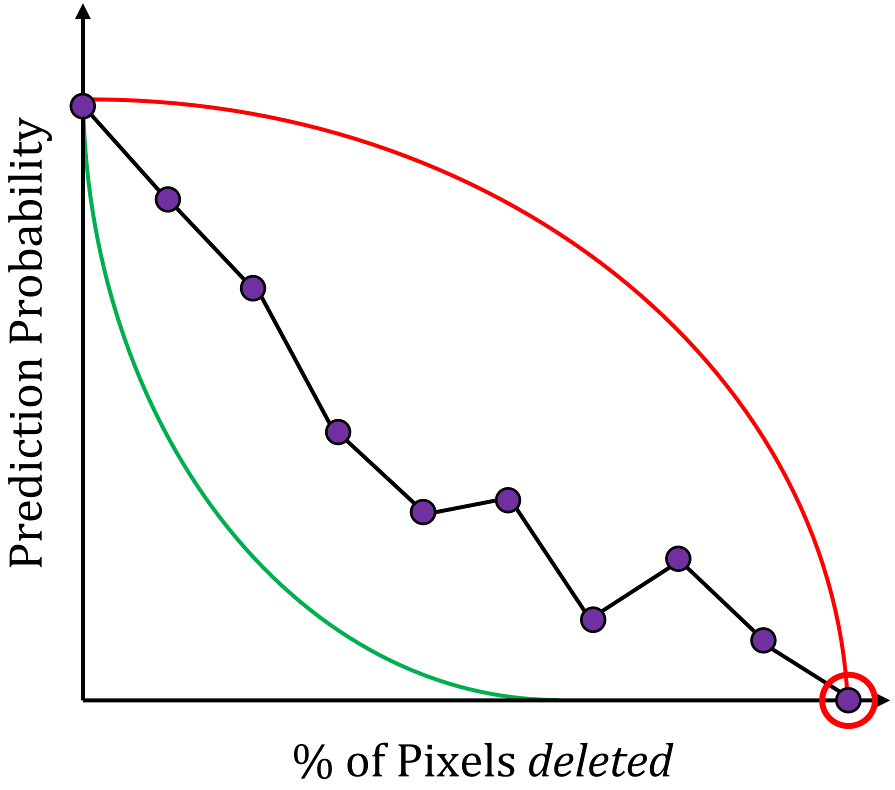
- Deletion: remove important features and see what happens..



How important are selected features?

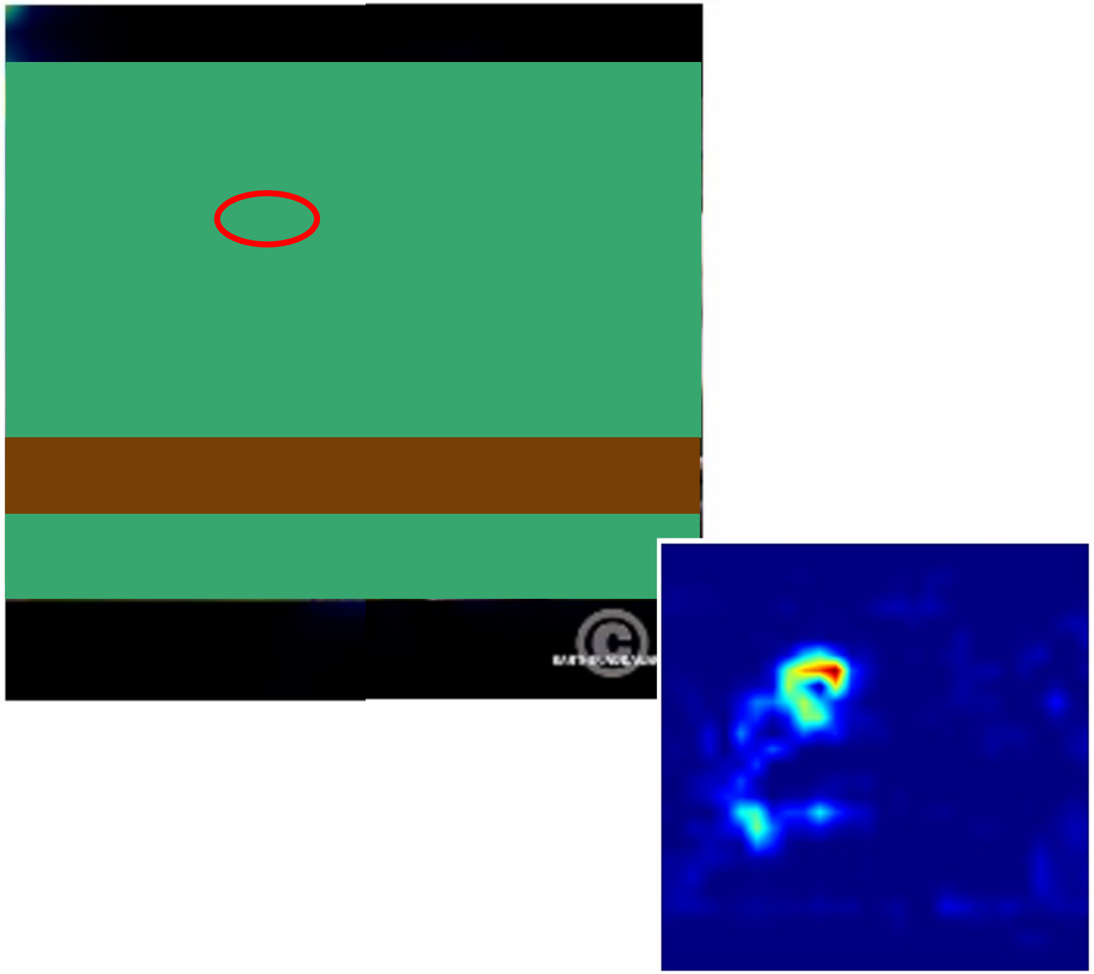
- Deletion: remove important features and see what happens..

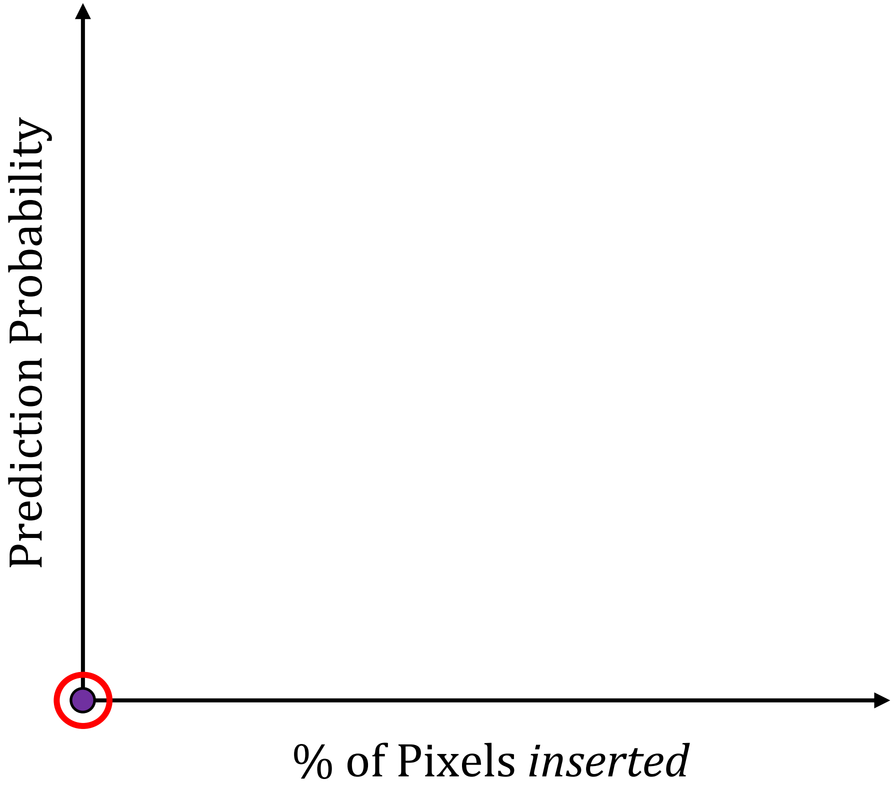




How important are selected features?

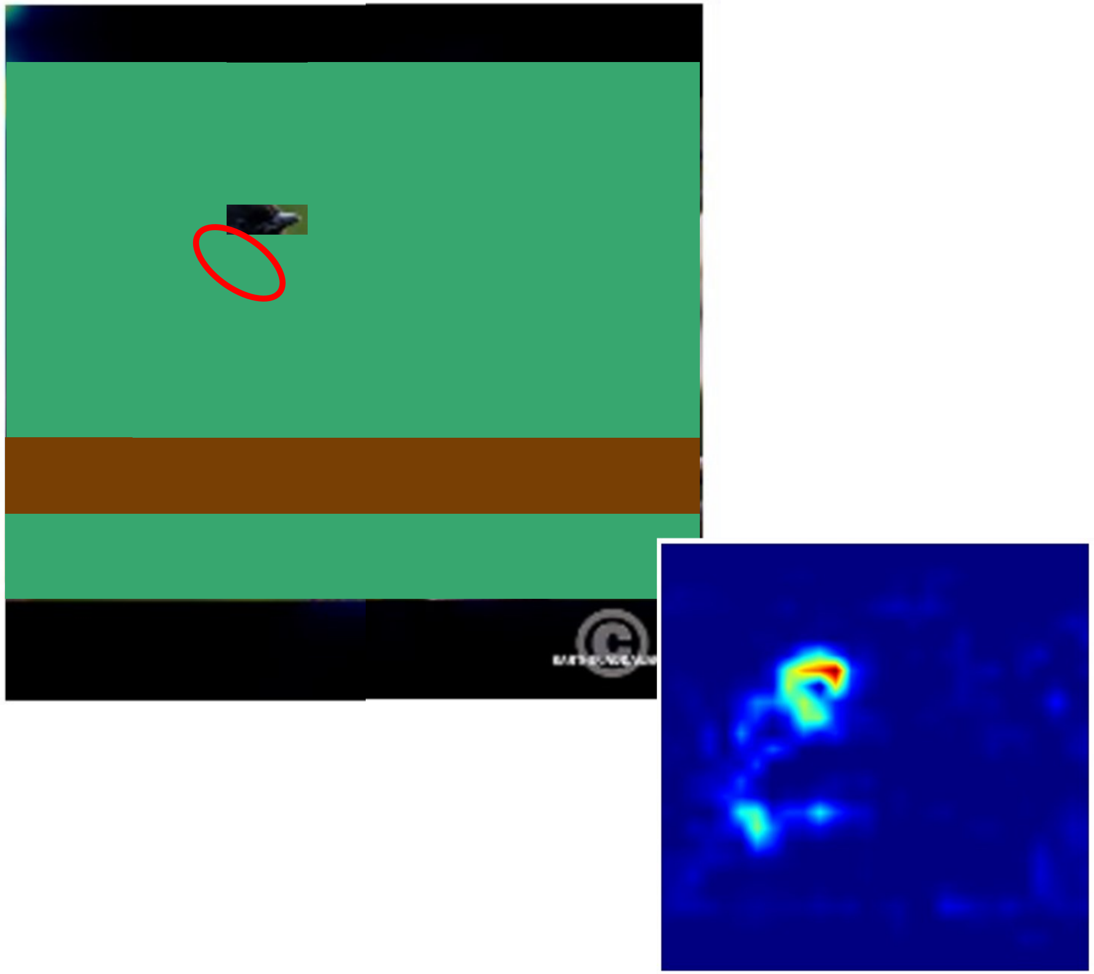
- Insertion: add important features and see what happens..

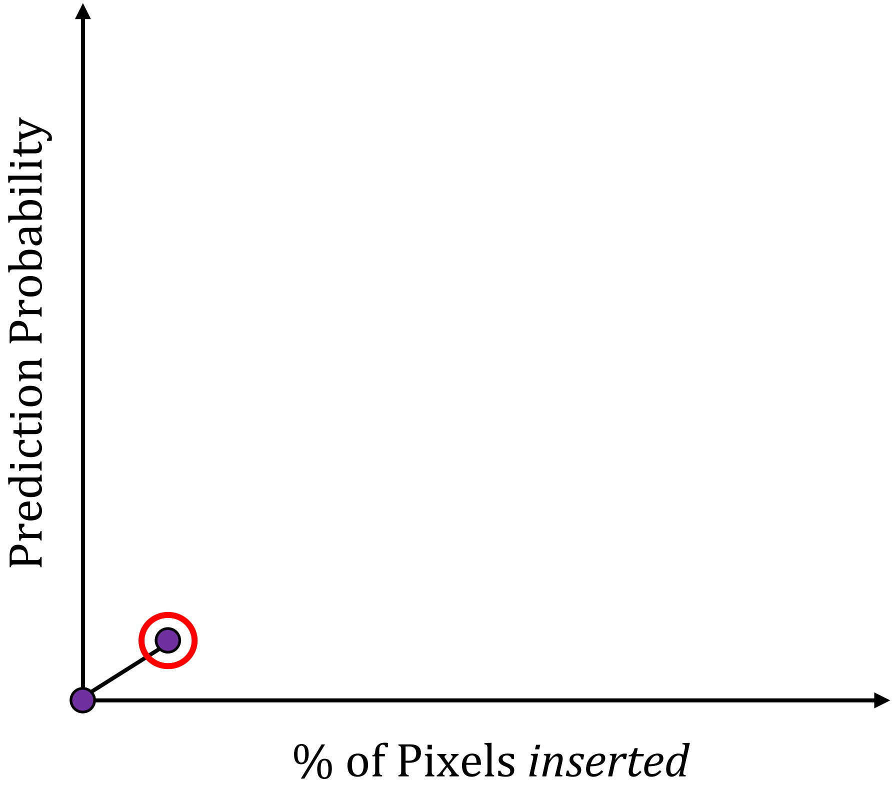




How important are selected features?

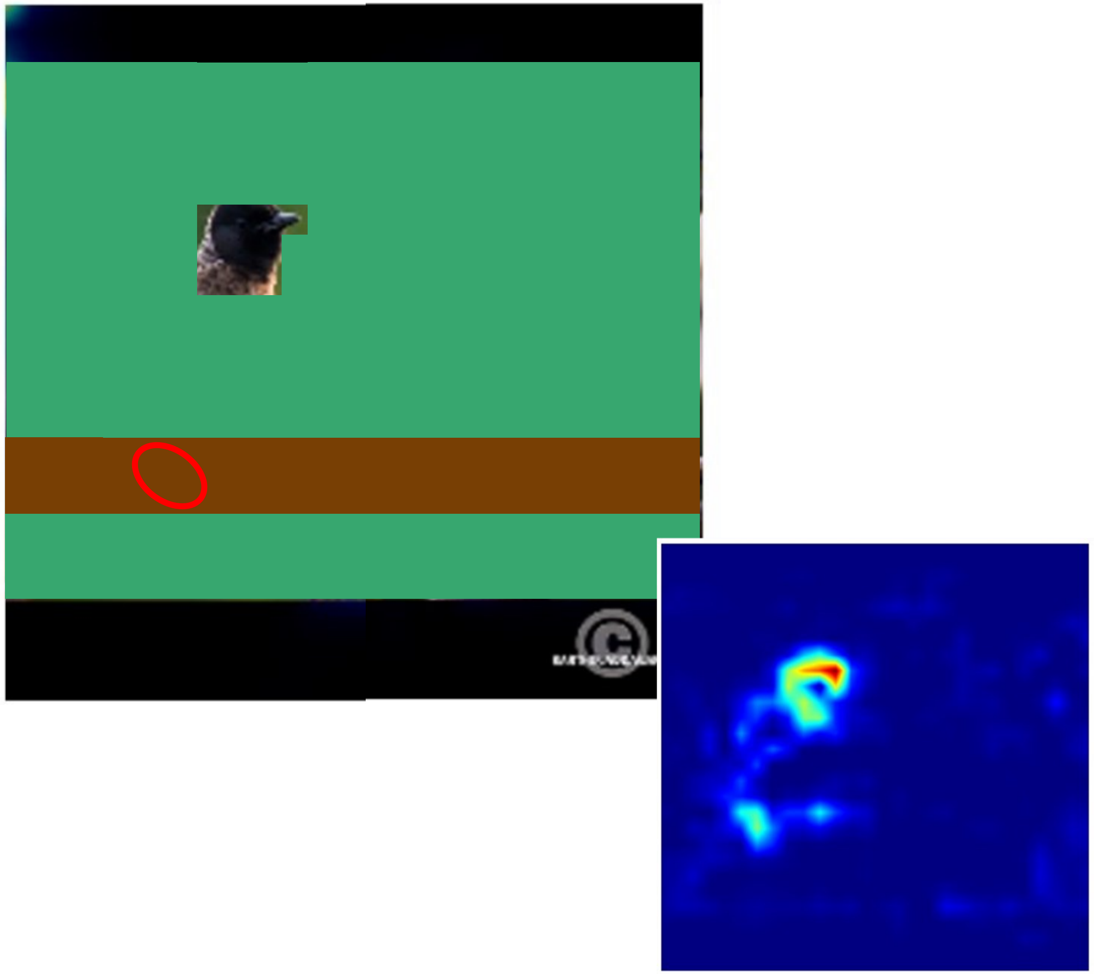
- Insertion: add important features and see what happens..

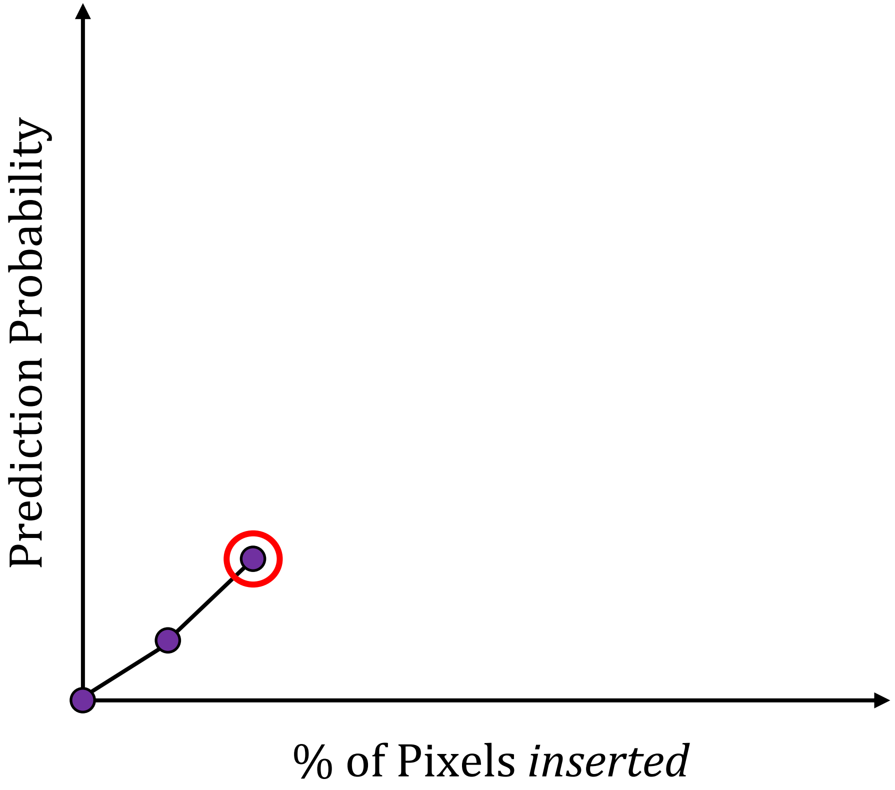




How important are selected features?

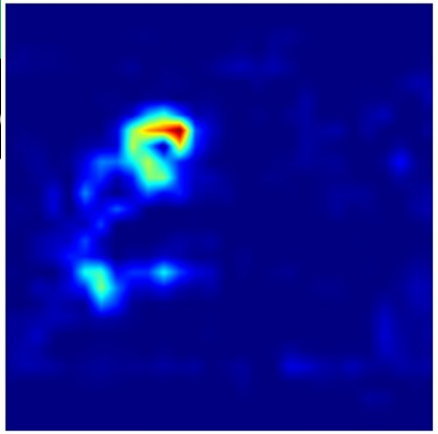
- Insertion: add important features and see what happens..

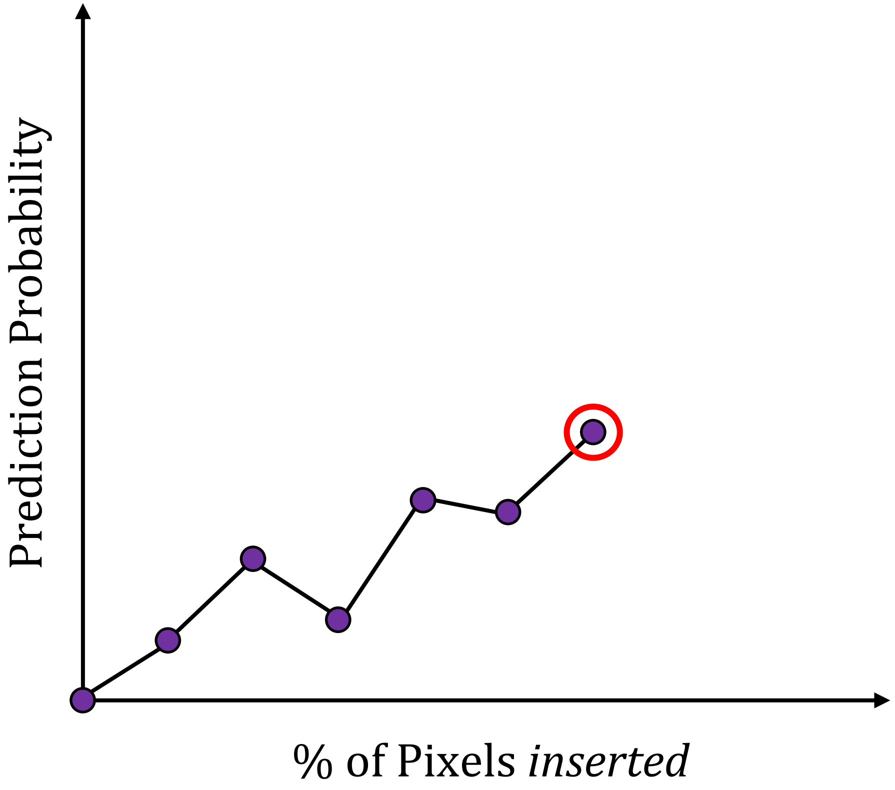




How important are selected features?

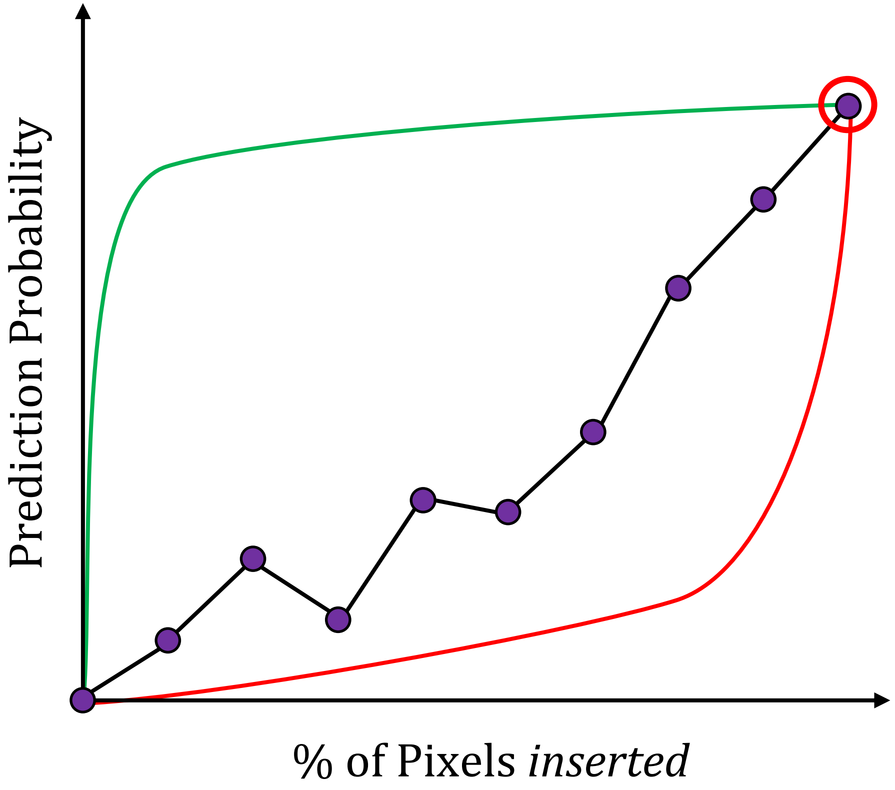
- Insertion: add important features and see what happens..





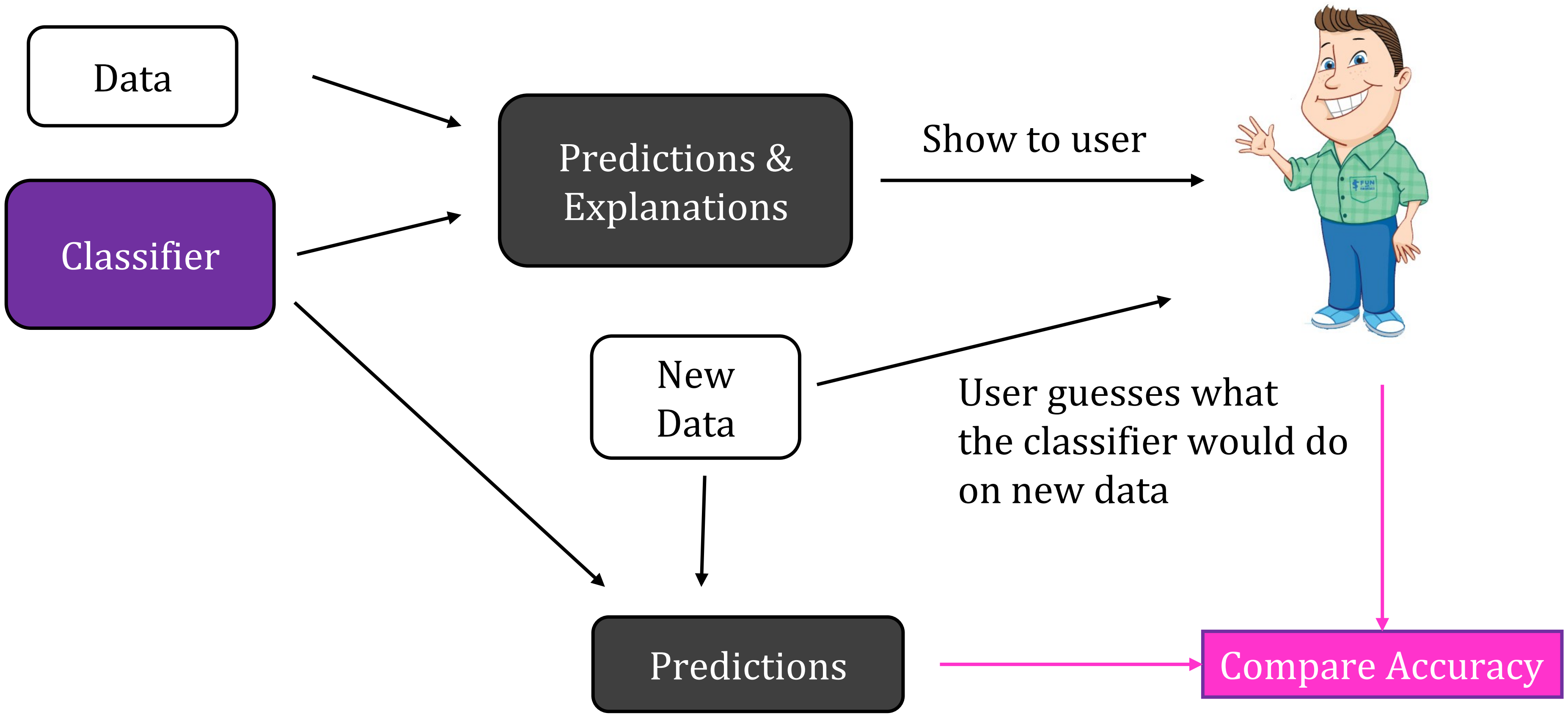
How important are selected features?

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

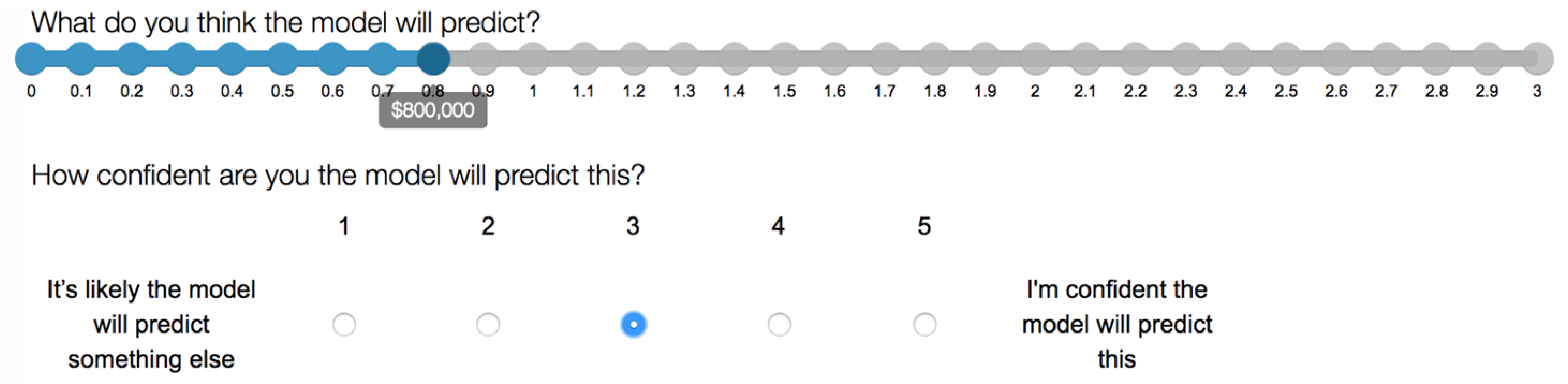


[Ribeiro et al. 2018, Hase and Bansal 2020]

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

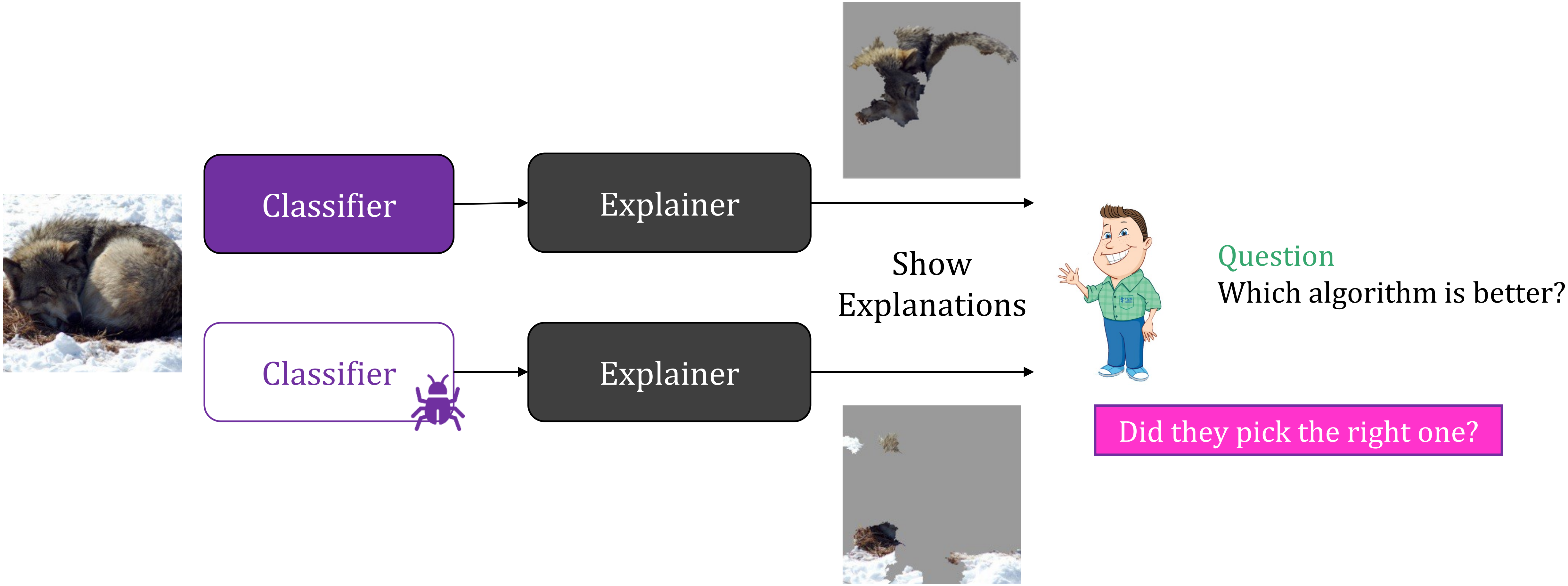
Question 1
Would you trust this model?

Did they say no?

Question 2
What is the classifier doing?

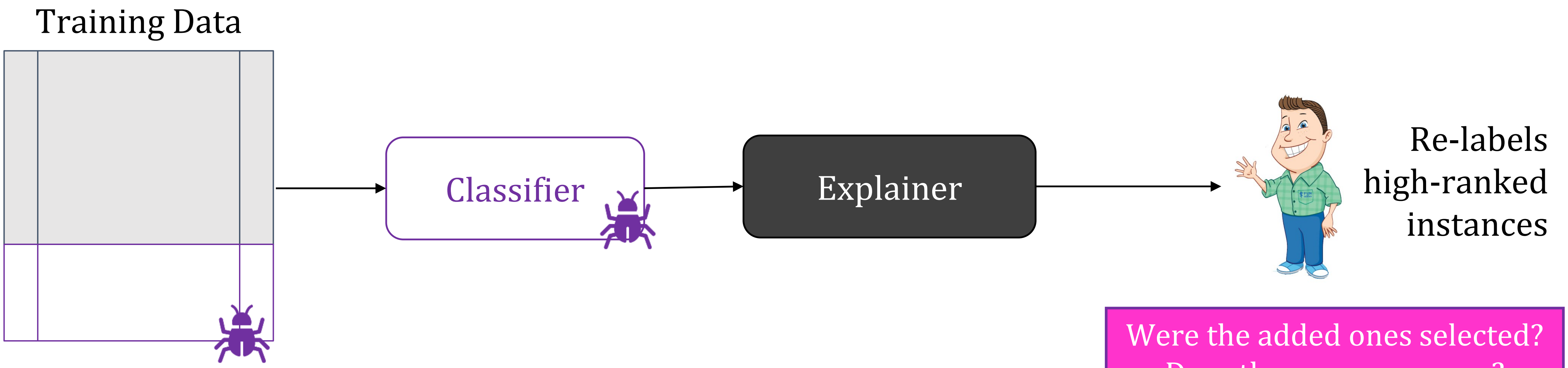
Did they get it right?

2. Comparing Classifiers



3. Finding Errors in Training Data

- Prototypical Explanations: important instances from training data



Were the added ones selected?
Does the accuracy go up?

Evaluating Posthoc Explanations

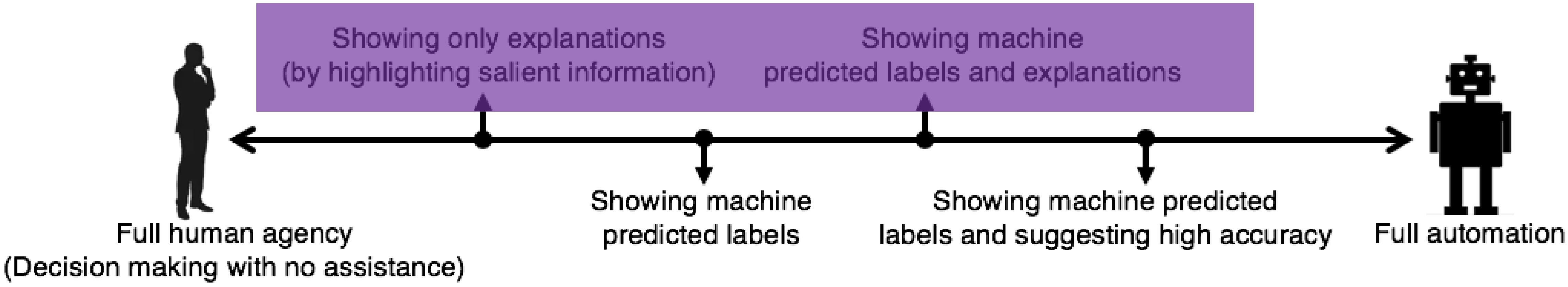
Understand the Behavior

Help make decisions

Useful for Debugging

Human-AI Collaboration

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



Evaluating Posthoc Explanations

Understand the Behavior

Help make decisions

Useful for Debugging

Limitations of Evaluating Explanations

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

Tutorial on Post hoc Explanations

Approaches for Post hoc Explainability

Evaluation of Explanations

Limits of Post hoc Explainability

Future of Post hoc Explainability

Tutorial on Post hoc Explanations

Approaches for Post hoc Explainability

Evaluation of Explanations

Limits of Post hoc Explainability

Future of Post hoc Explainability

Limits of Post hoc Explanations

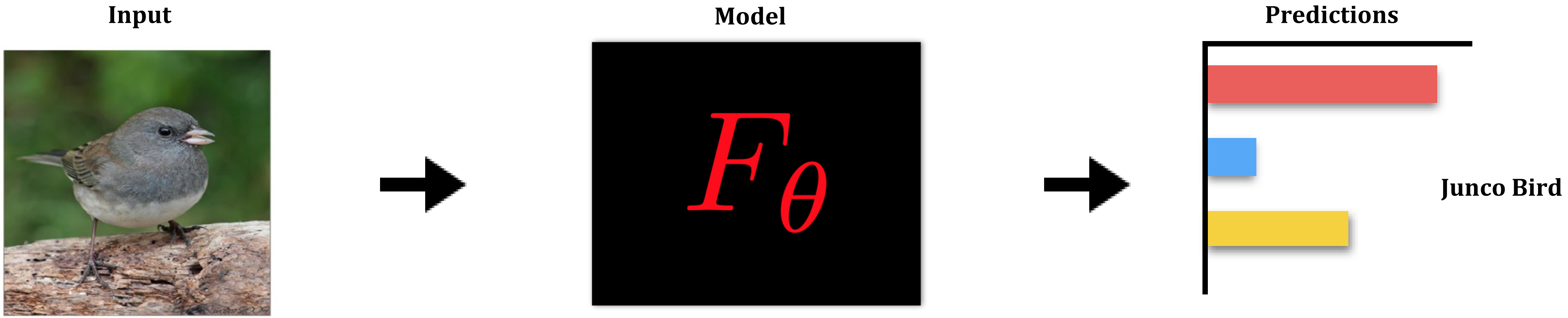
Limitations

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

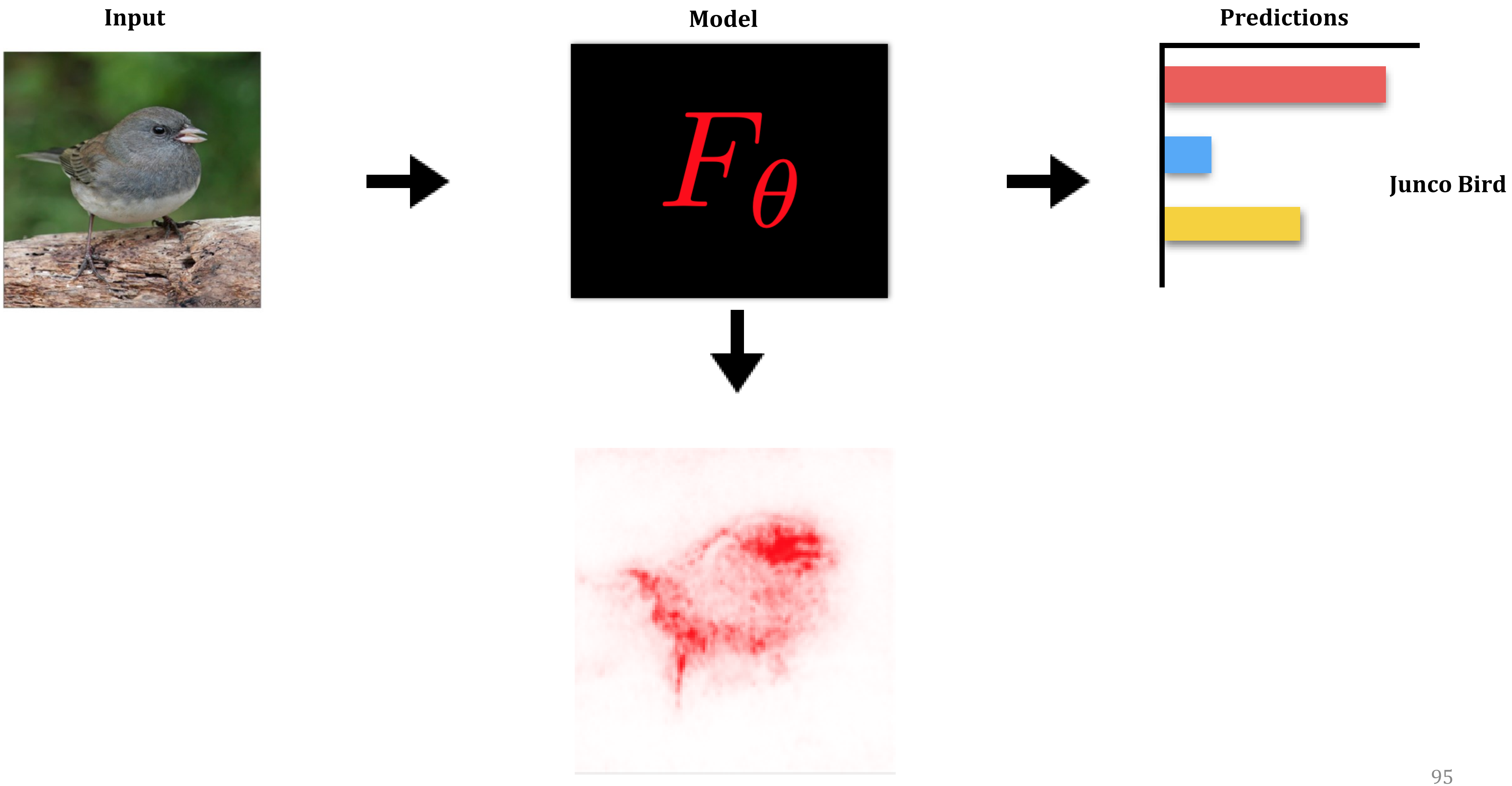
Limitations

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

Do Explanations Capture Model-based Discriminative Signals?



Do Explanations Capture Model-based Discriminative Signals?

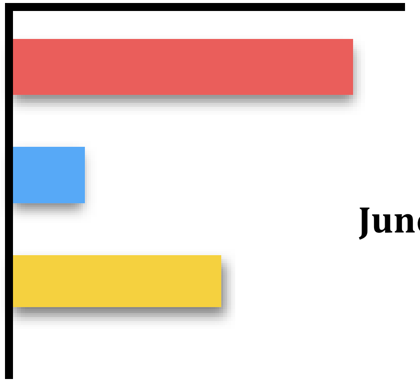


Do Explanations Capture Model-based Discriminative Signals?

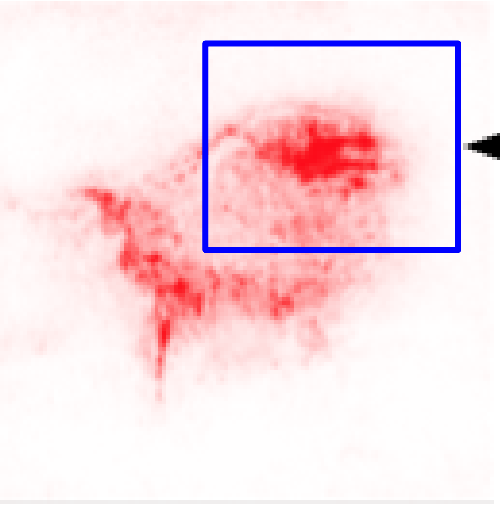
Input

Model

Predictions



Junco Bird



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

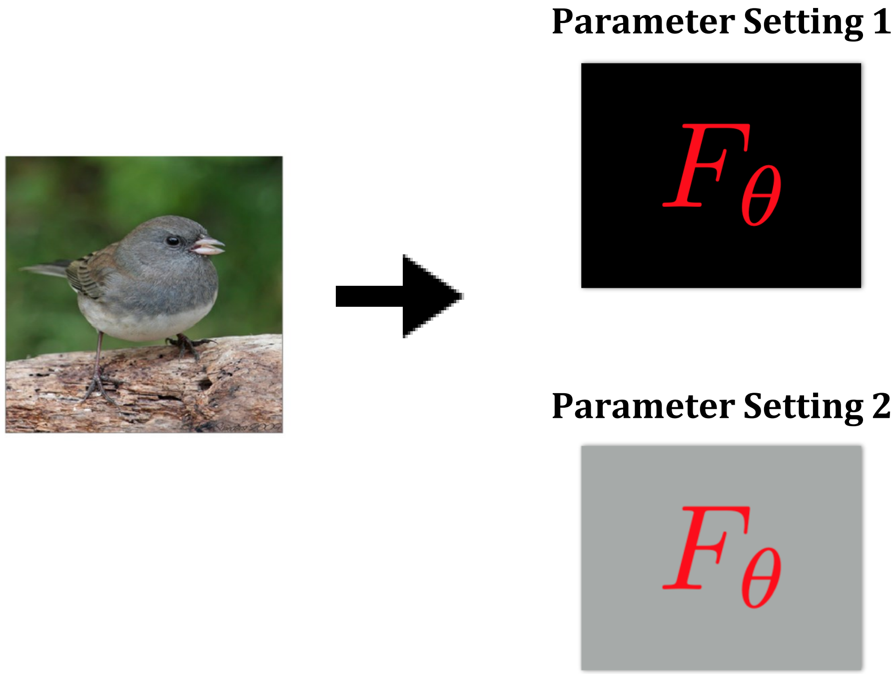
Sanity Check for Faithfulness/Fidelity

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

Adebayo et. al. 2018

Sanity Check for Faithfulness/Fidelity

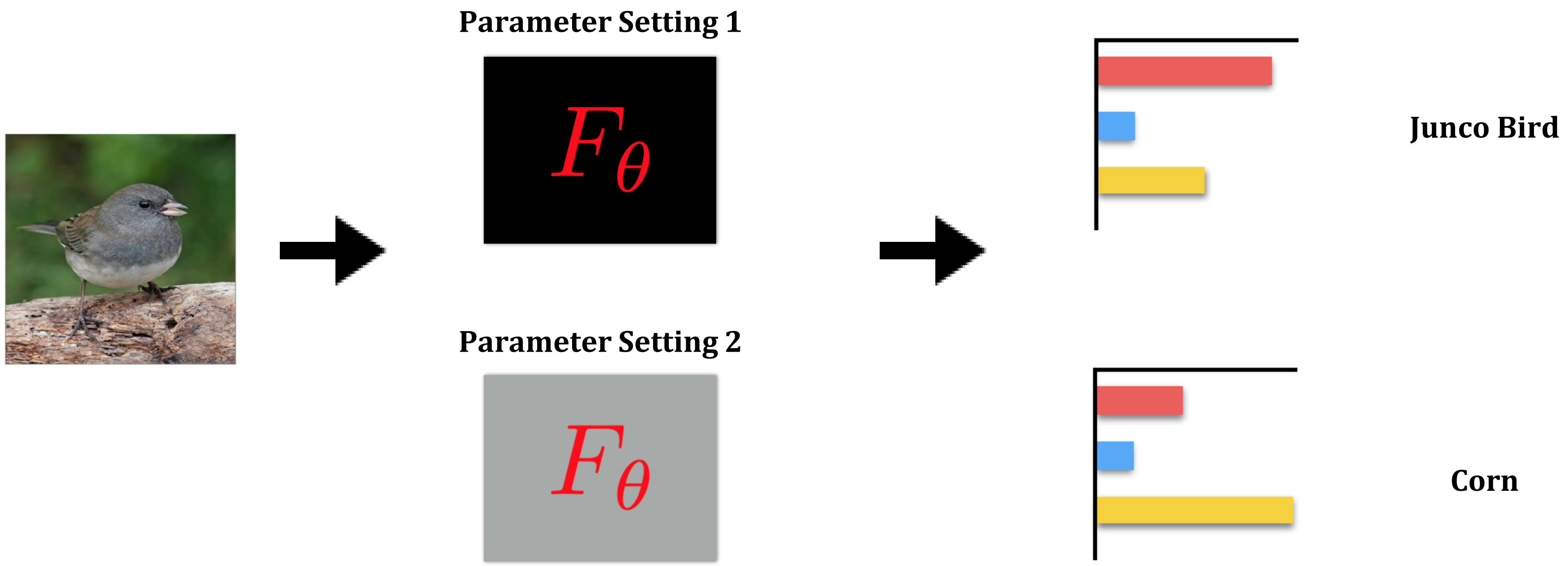
- **Sensitivity to Model Parameters:** if the parameter settings change, the explanations should change.



Adebayo et. al. 2018

Sanity Check for Faithfulness/Fidelity

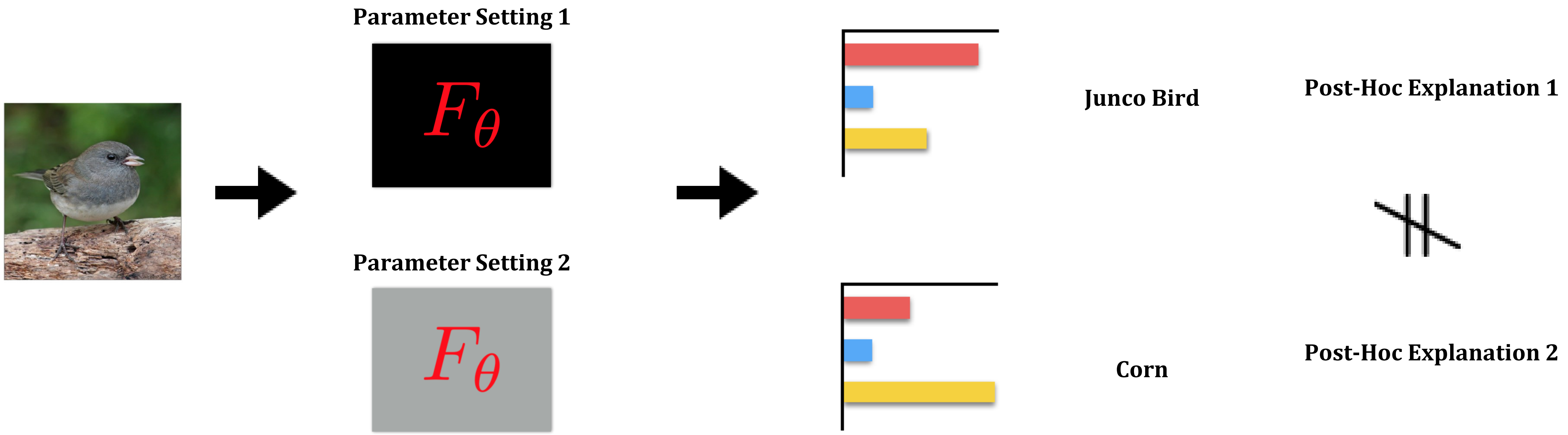
- **Sensitivity to Model Parameters:** if the parameter settings change, the explanations should change.



[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

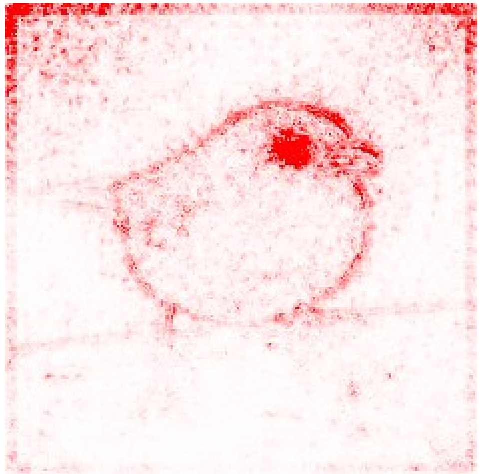
[Adebayo et. al. 2018](#)

Cascading Randomization Inception-V3

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

Guided BackProp Explanation Inception-V3 ImageNet

Normal Model Explanation



[Adebayo et. al. 2018](#)

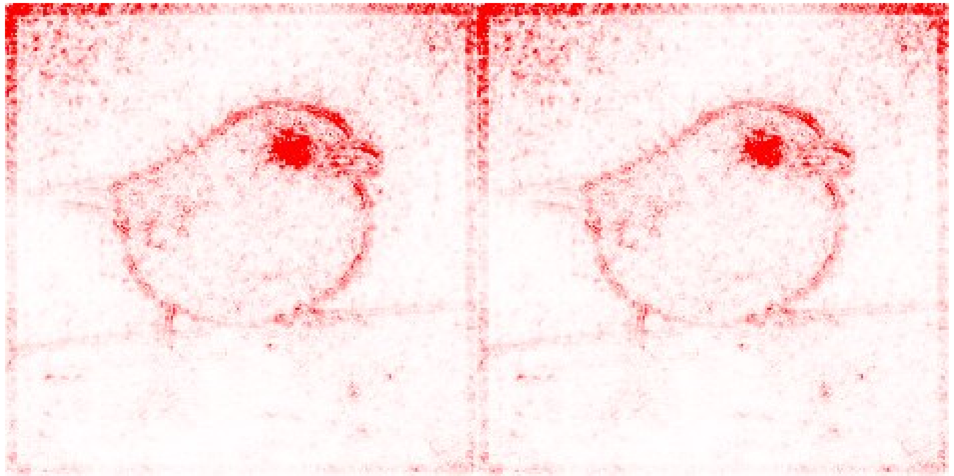
Cascading Randomization Inception-V3

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

Guided BackProp Explanation Inception-V3 ImageNet

Normal Model Explanation

Top Layer Randomized



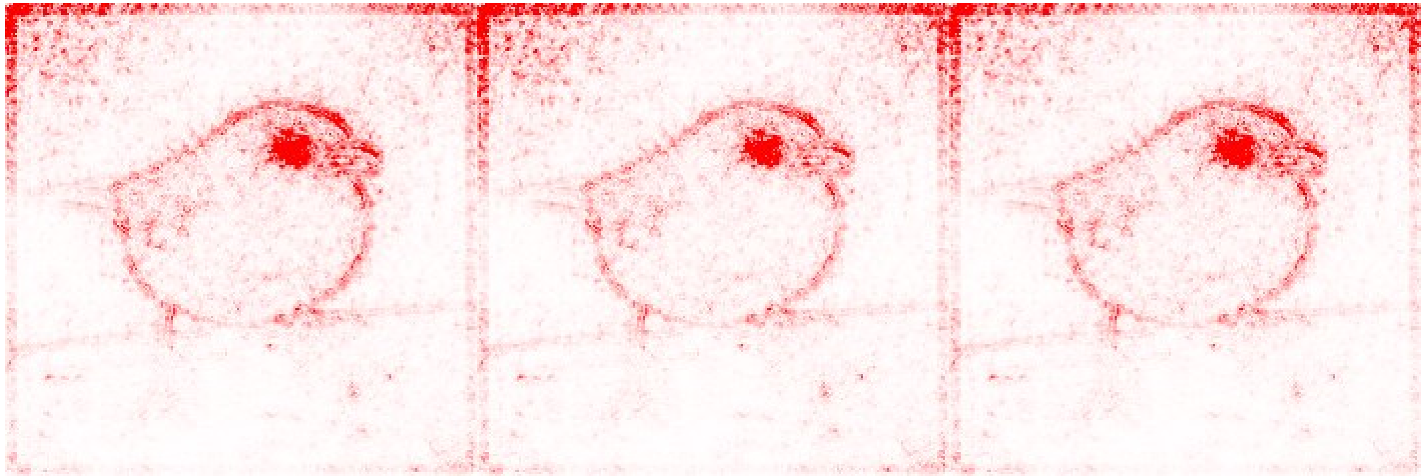
[Adebayo et. al. 2018](#)

Cascading Randomization Inception-V3

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

Guided BackProp Explanation Inception-V3 ImageNet

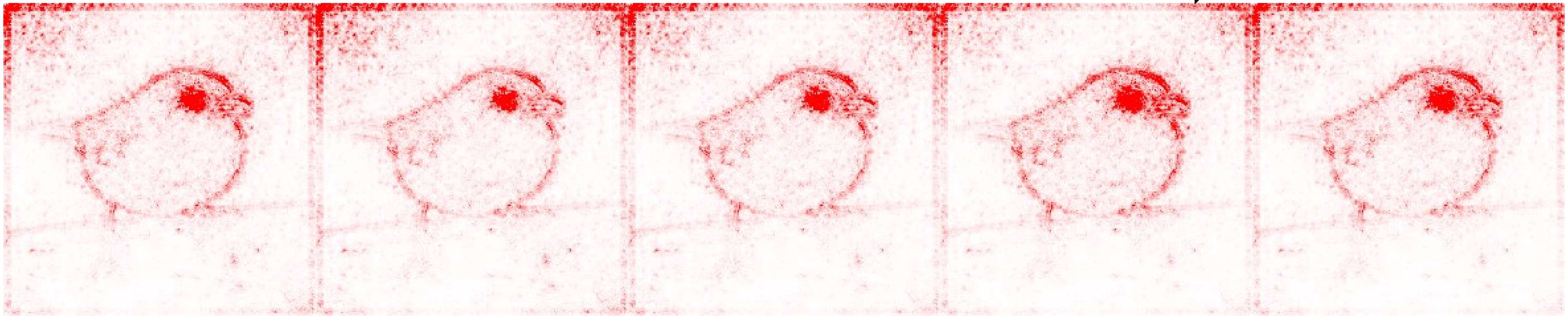
Normal Model Explanation | **Top Layer Randomized**



[Adebayo et. al. 2018](#)

Cascading Randomization Inception-V3

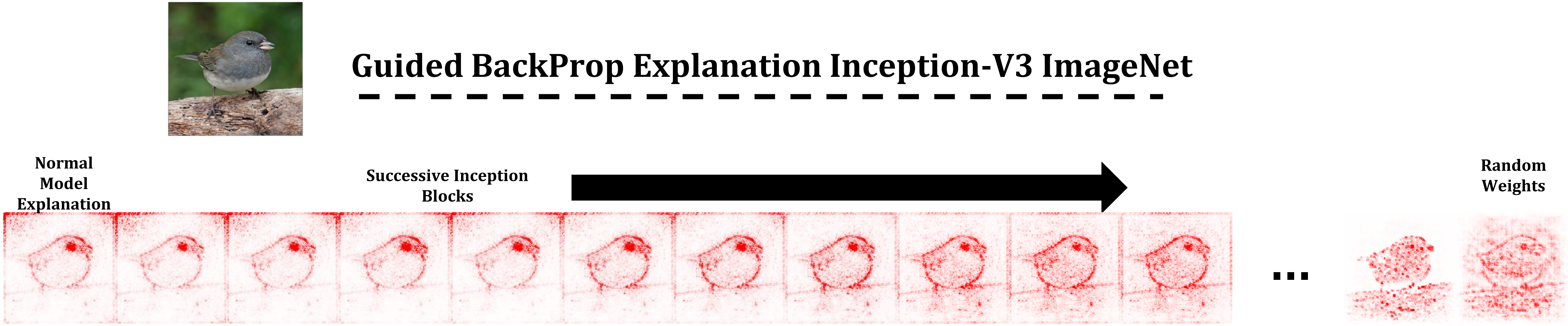
- **Randomize (re-initialize)** model parameters starting from top layer all the way to the input.



Adebayo et. al. 2018

Cascading Randomization Inception-V3

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

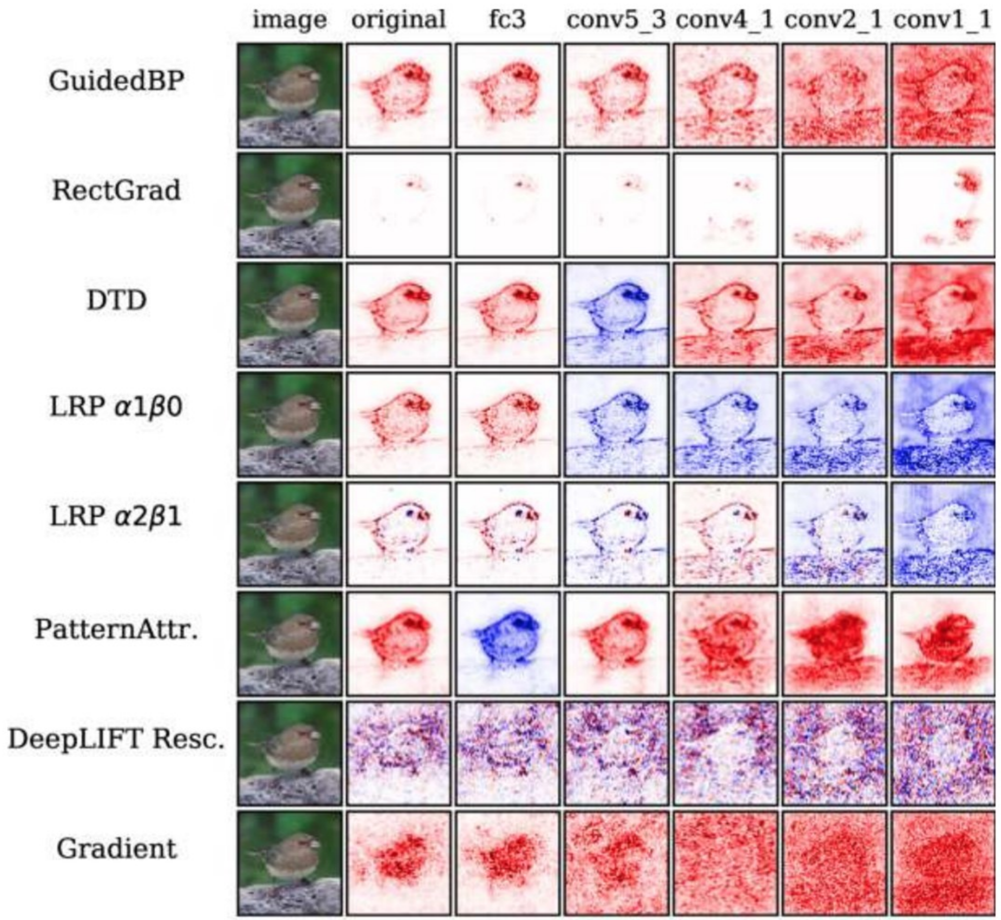


Guided BackProp is invariant to the higher level weights.

[Adebayo et. al. 2018](#)

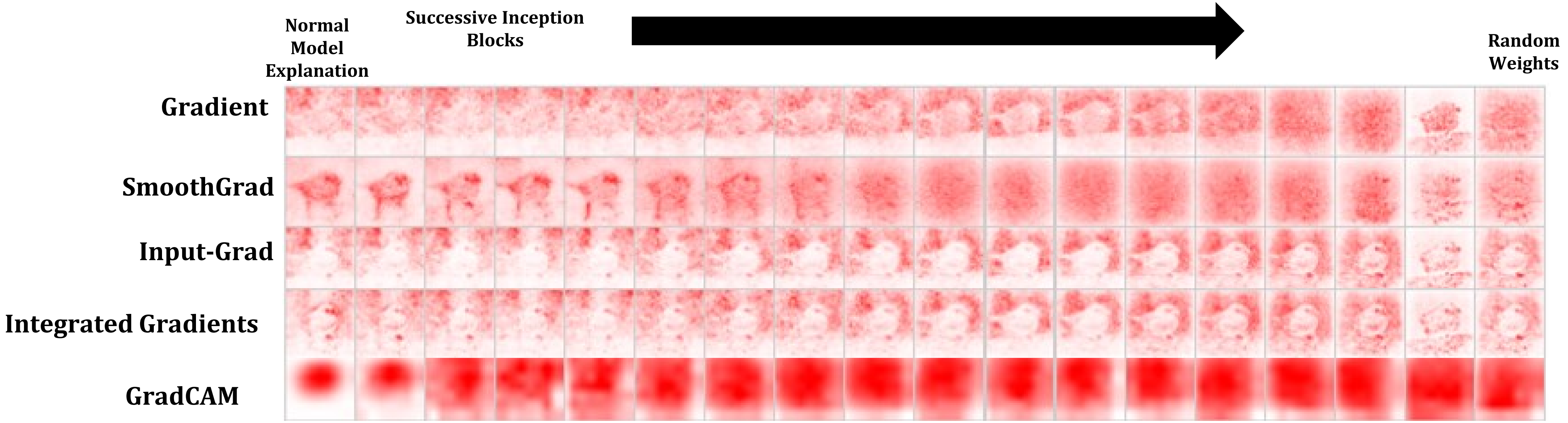
‘Modified backprop approaches’ are invariant

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



[Sixt et. al. 2020]

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

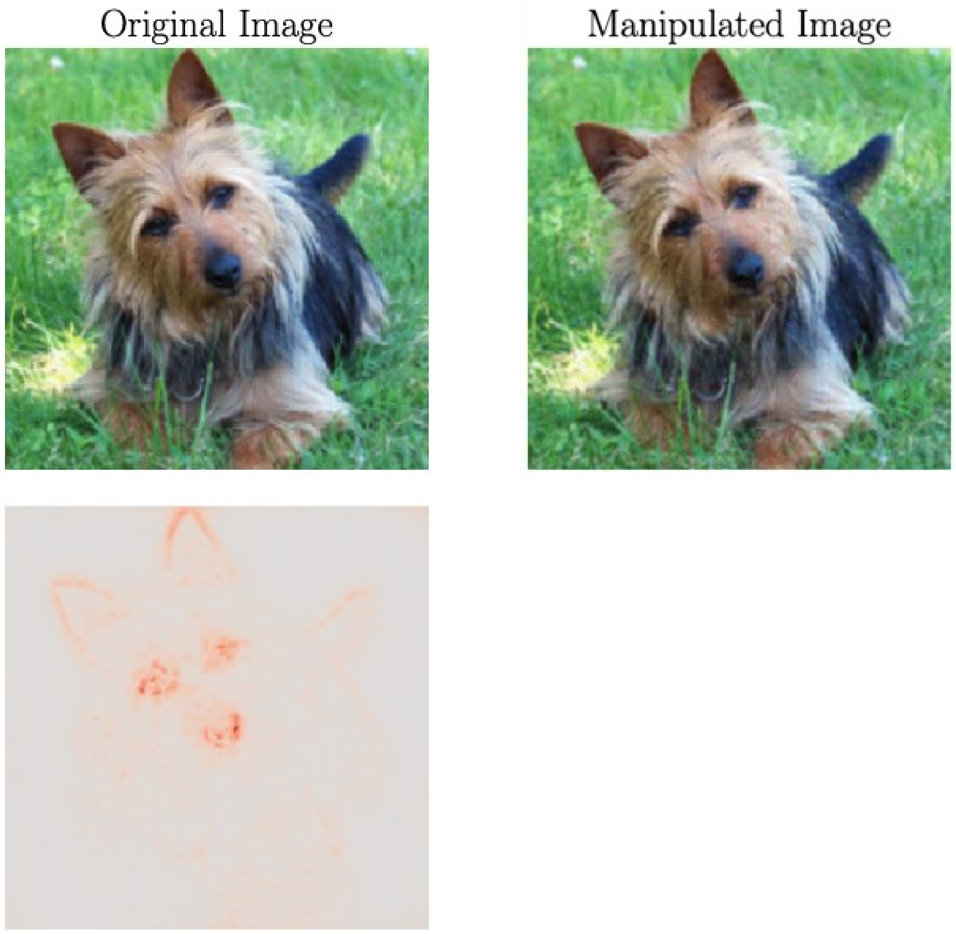
Post-hoc explanations can be easily manipulated.

Original Image

[Dombrowski et. al. 2019]

Post-hoc Explanations are Fragile

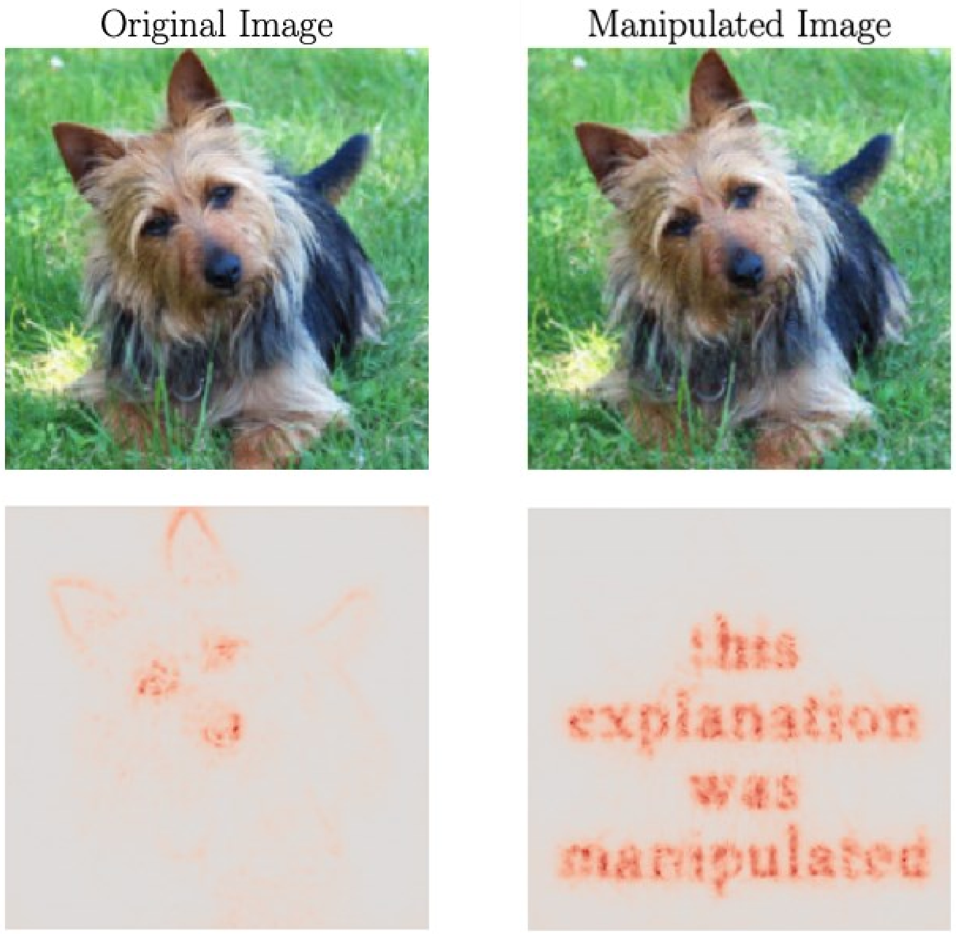
Post-hoc explanations can be easily manipulated.



[Dombrowski et. al. 2019](#)

Post-hoc Explanations are Fragile

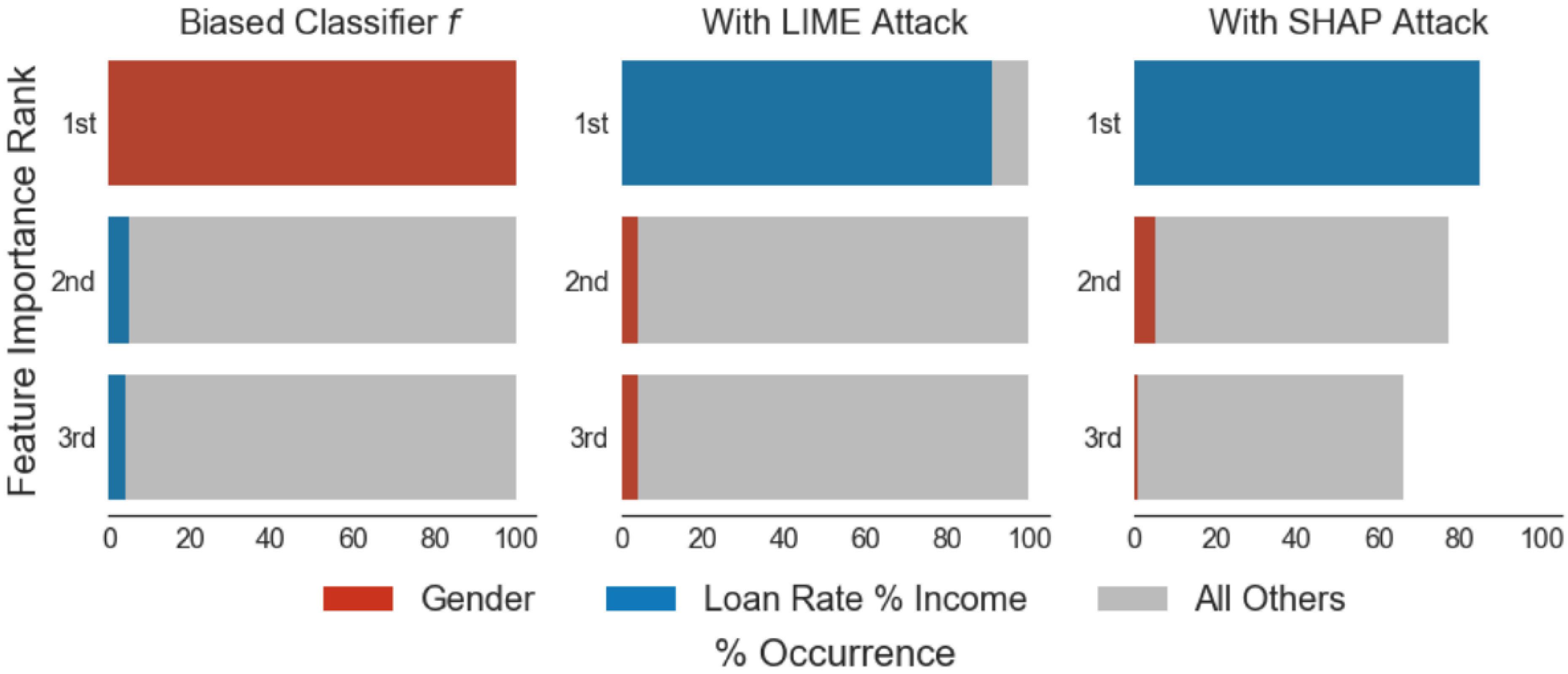
Post-hoc explanations can be easily manipulated.



[Dombrowski et. al. 2019]

Scaffolding Attack on LIME & SHAP

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



Slack and Hilgard et. al. 2020

Limitations

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

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

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

Limitations: Stability

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

Alvarez et. al. 2018.

Limitations: Stability

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

'Local Lipschitz Constant'

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

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

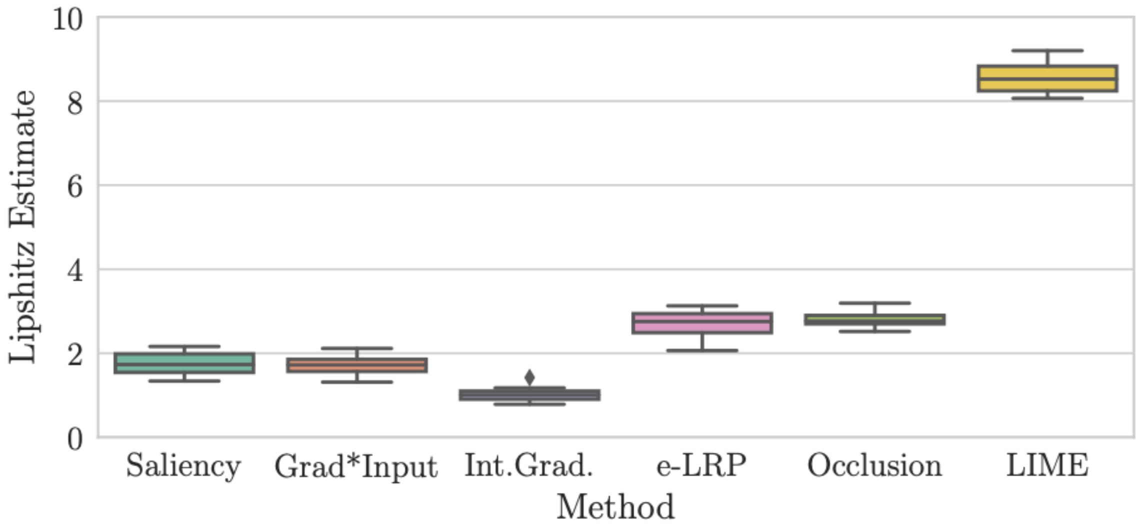
Input

Alvarez et. al. 2018.

Limitations: Stability

- Perturbation approaches like LIME can be unstable.

- [Yeh et. al. (2019)](#) analytically derive bounds on explanations sensitive for certain popular methods and propose stable variants.

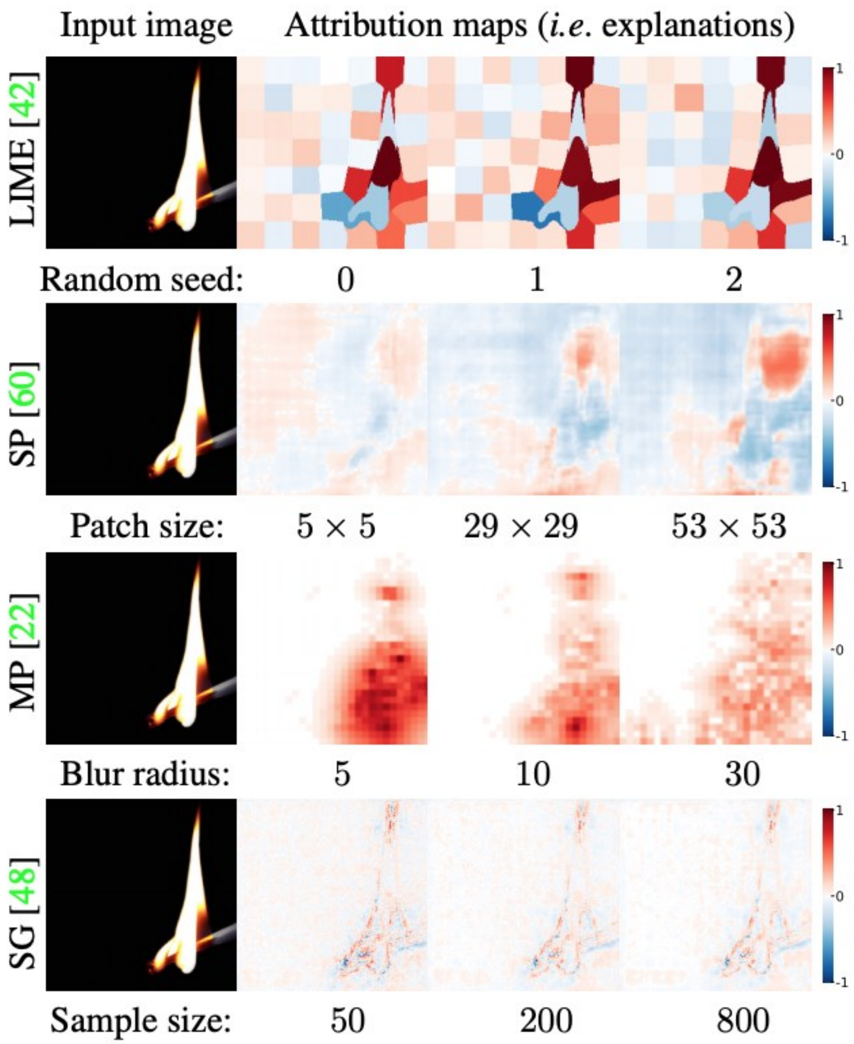


Estimate for 100 tests for an MNIST Model.

Alvarez et. al. 2018.

Sensitivity to Hyperparameters

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



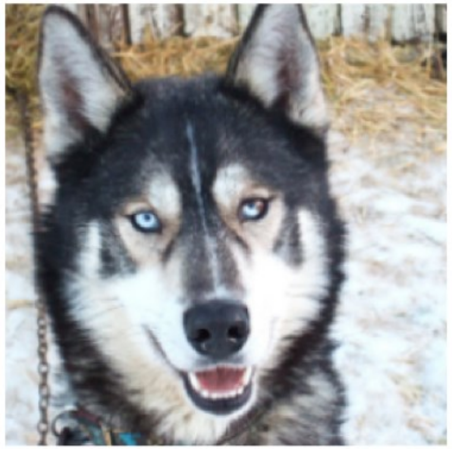
[Bansal, Agarwal, & Nguyen, 2020.](#)

Limitations

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

Model Debugging: Spurious Signals

True Label: Siberian Husky



Model

Predictions

LIME

Relying on snow background

[Riberio et. al. 2017.]

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

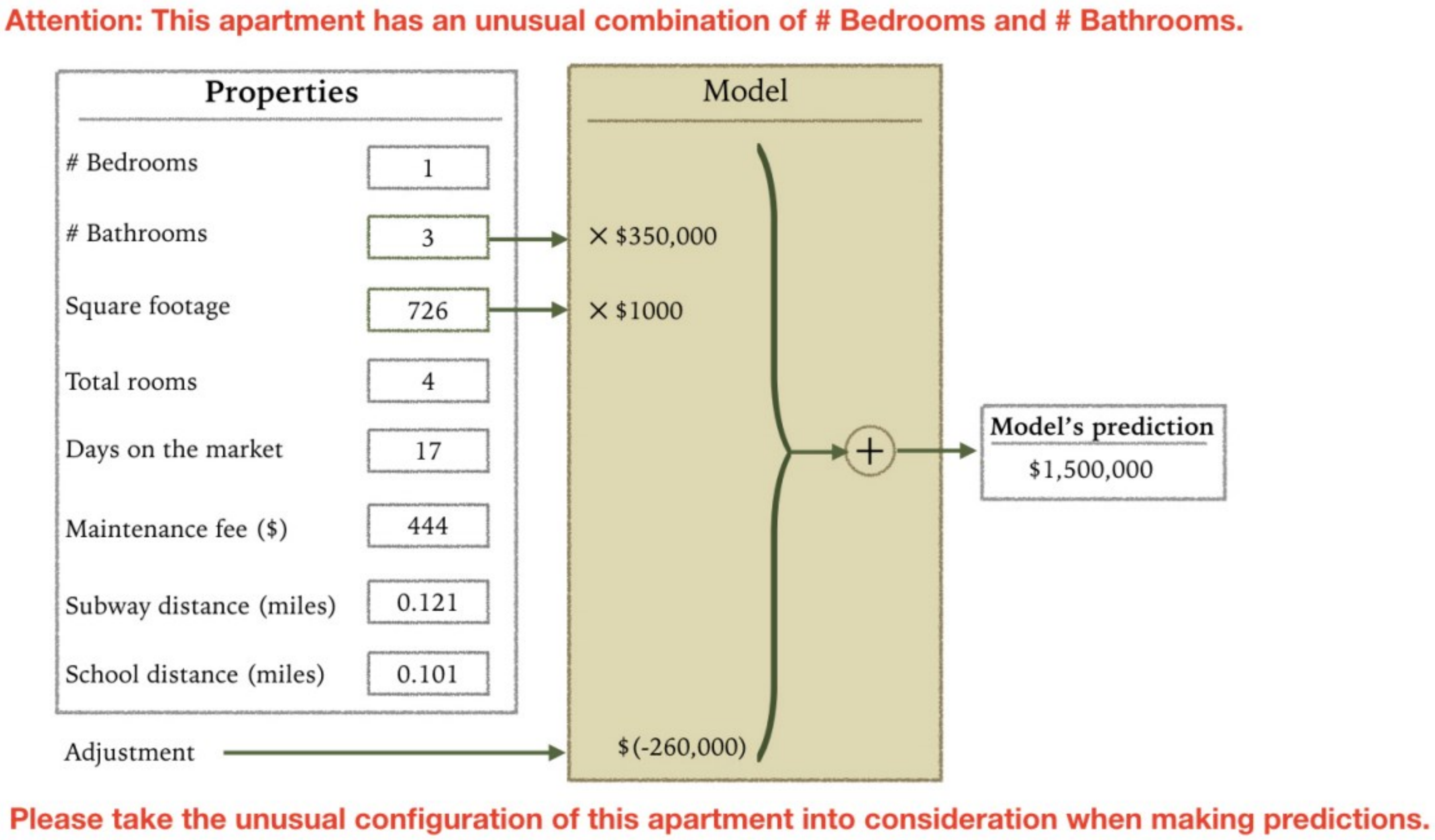
If Current-Offense = Misdemeanor and Prior-Arrests ≤ 1:
- If Has-Kids = No and Owns-House = No and Prior-Jail-Incarcerations = Yes, then Risky
- If Age ≥ 50 and Has-Kids = Yes and Prior-FTA = No, then Not Risky

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

Limitations

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

Tutorial on Post hoc Explanations

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

Tutorial on Post hoc Explanations

Approaches for Post hoc Explainability

Evaluation of Explanations

Limits of Post hoc Explainability

Future of Post hoc Explainability

Future of Post hoc Explainability

Emerging Topics in Explainability Research

Future of Post hoc Explainability

Towards Better Post hoc Explanations

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

Other Emerging Directions

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

Future of Post hoc Explainability

Towards Better Post hoc Explanations

Methods for More Reliable Post hoc Explanations

Theoretical Analysis of Post hoc Explanation Methods

Rigorous Evaluation of the Utility of Post hoc Explanations

Other Emerging Directions

Post hoc Explainability Beyond Classification

Intersections with Differential Privacy

Intersections with Fairness

Methods for More Reliable Post hoc Explanations

Post hoc explanations have several limitations:

not faithful to the underlying model, unstable, fragile

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

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!

- ” model

-

- Local error 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

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!

them -- "*Participants trusted the tools because of their visualizations and their public availability*"

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

Tutorial on Post hoc Explanations

Approaches for Post hoc Explainability

Evaluation of Explanations

Limits of Post hoc Explainability

Future of Post hoc Explainability

Summary of Tutorial

Approaches for Post hoc Explainability

Evaluation of Explanations

Limits of Post hoc Explainability

Future of Post hoc Explainability

Parting Thoughts…

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

When introducing a new explanation method:

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

Thank You!

Sameer Singh
UC Irvine

sameeersingh.org
sameer@uci.edu
@sameer_

Julius Adebayo
MIT

Hima Lakkaraju
Harvard University

Slides and Video: explainml-tutorial.github.io