Topic 10: ML Transparency

Interpreting & Explaining Model Predictions

CMSC 491/691 Robust Machine Learning

UMBC

Some slides from Hima Lakkaraju

Reminders / Announcements ...

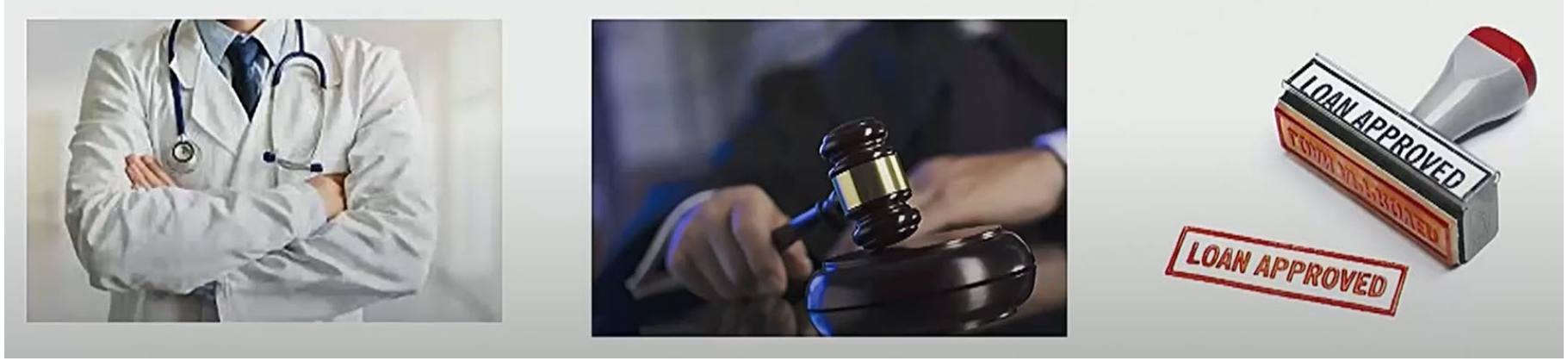
- If you've presented in class, submit your slides!
  - I will only grade this assignment after you submit the slides
- Check your email / Blackboard messages!
  - I'm chasing missing submissions (i.e. I'm being nice) to avoid giving you a **0**
- No one has submitted the extra credit assignment
  - These are extra credits! Don't you want them ??!!
- Project: 1 month left! Keep me in the loop!

Model Transparency

- ML models make mistakes
  - his entire class is about the different types of failures
- Can we explain those failures?
- Can we even interpret a decision made by the model?
  - **Why does the model think this is a cat?**
- Unfortunately, so far in this class, the model is like a "black box"
  - No justification / explanation / ability to interpret predictions

Model Transparency

- Important in high stakes applications of ML



- Why did Uber autonomous car not detect the pedestrian?

ARIZONA

Self-driving Uber car kills Arizona pedestrian, police say

By Travis Fedschun , · **Fox News**

Published March 19, 2018 6:58pm EDT | Updated March 20, 2018 7:34am EDT

Model Transparency

- Important in high stakes applications o

- Why did Uber autonomous car not det

Self-driving Ub
pedestria

By Travis F

Published March 19, 2018 6:58pm

NBC NEWS WATCH LIVE

TECH NEWS

Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk

The automated car lacked "the capability to classify an object as a pedestrian unless that object was near a crosswalk," an NTSB report said.

Why is it challenging?

- Impossible to enumerate all possible scenarios
  - Either through training data or hardcoding

- Key criteria for "safety" are hard to quantify
  - Current attitude is "you will know it when you see it"

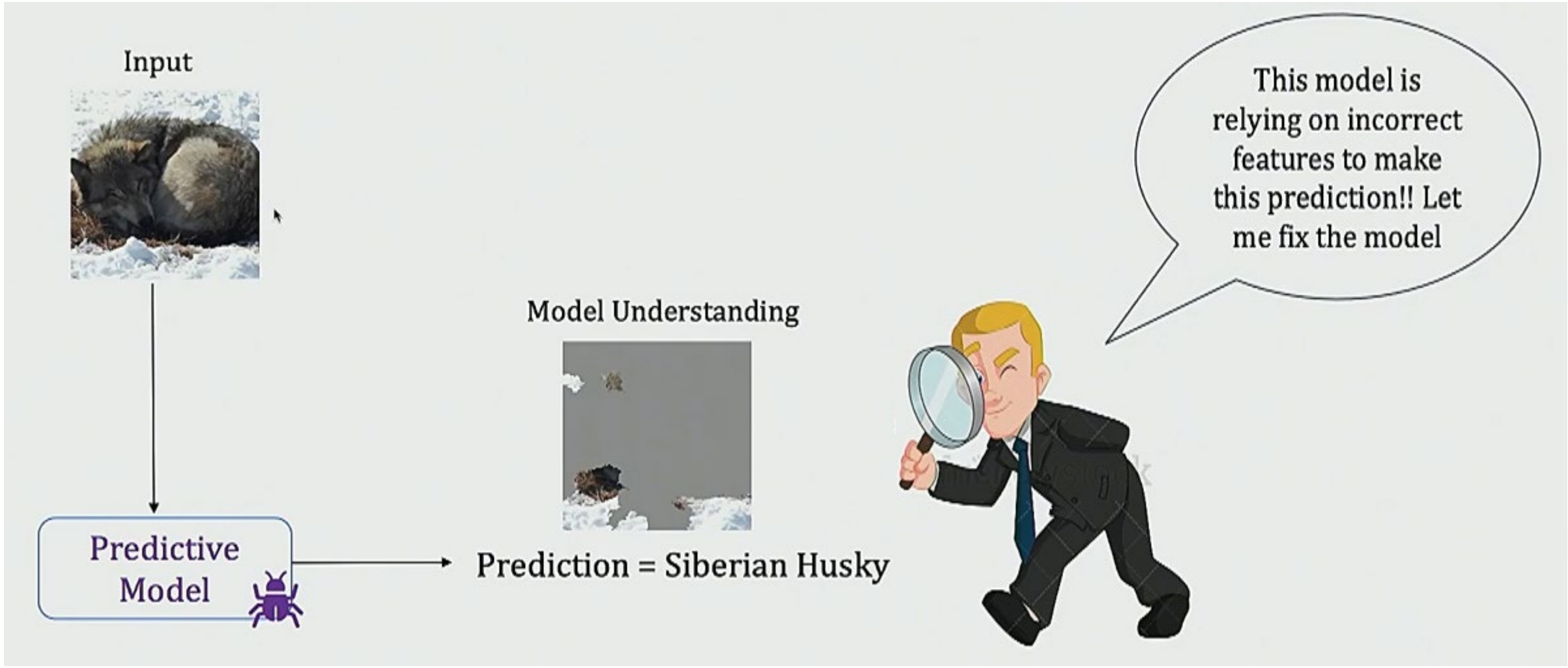
quote attributed to Hima Lakkaraju

Why is Model Transparency Important?

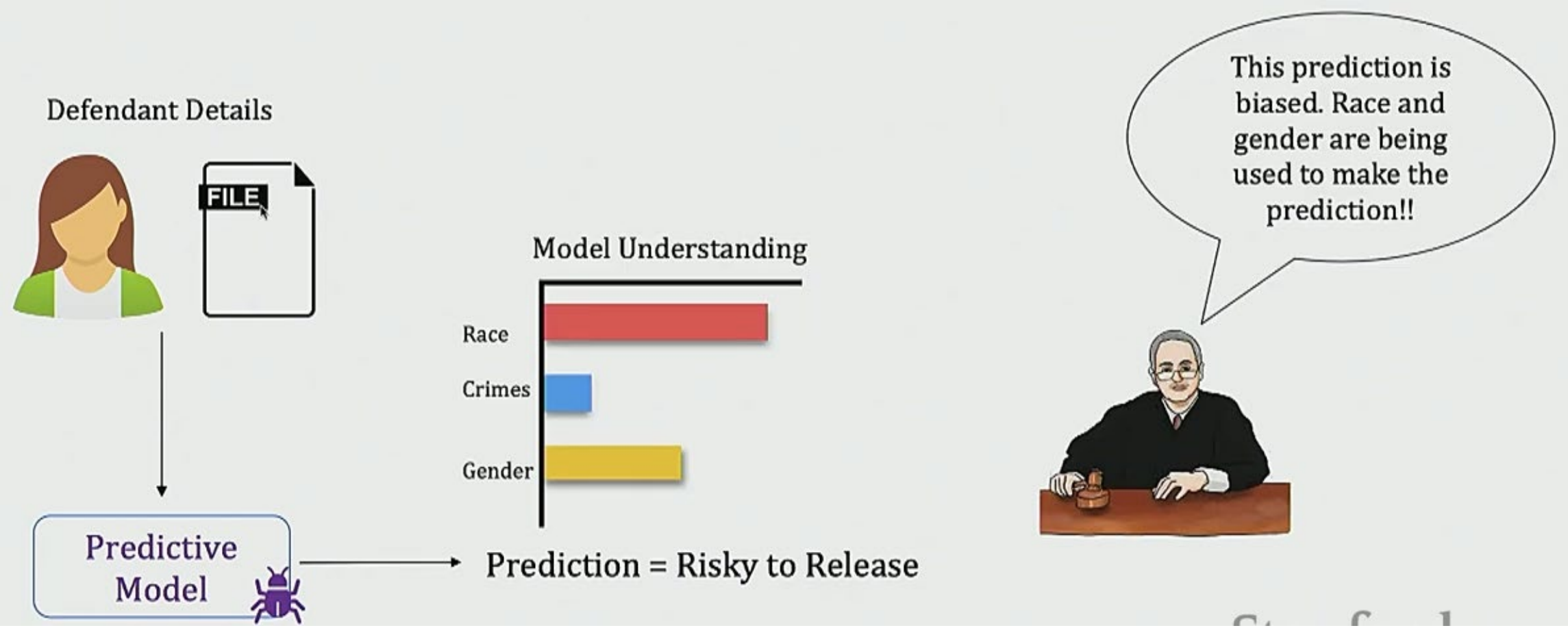
Motivating Examples

(adapted from Hima Lakkaraju's Seminar)

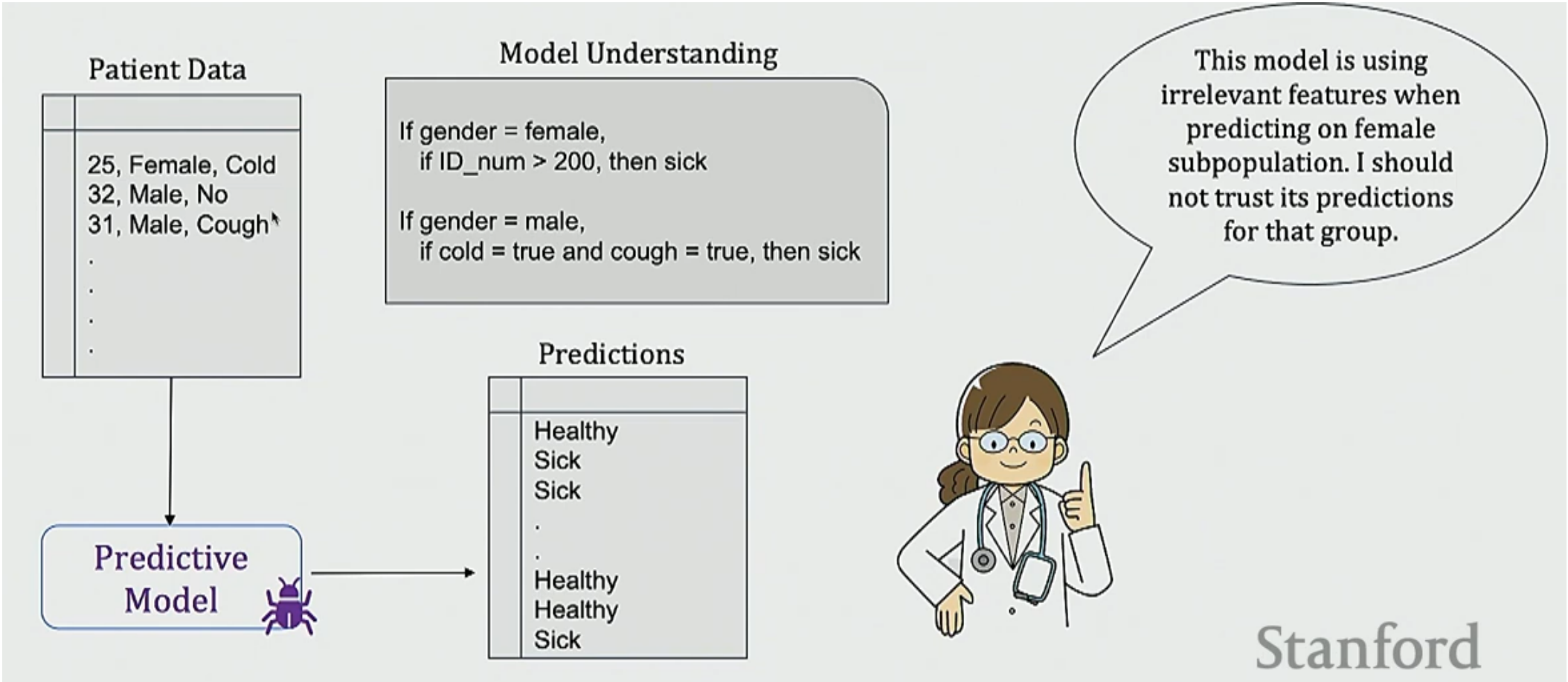
Model Transparency ⇒ Debugging ML Models?



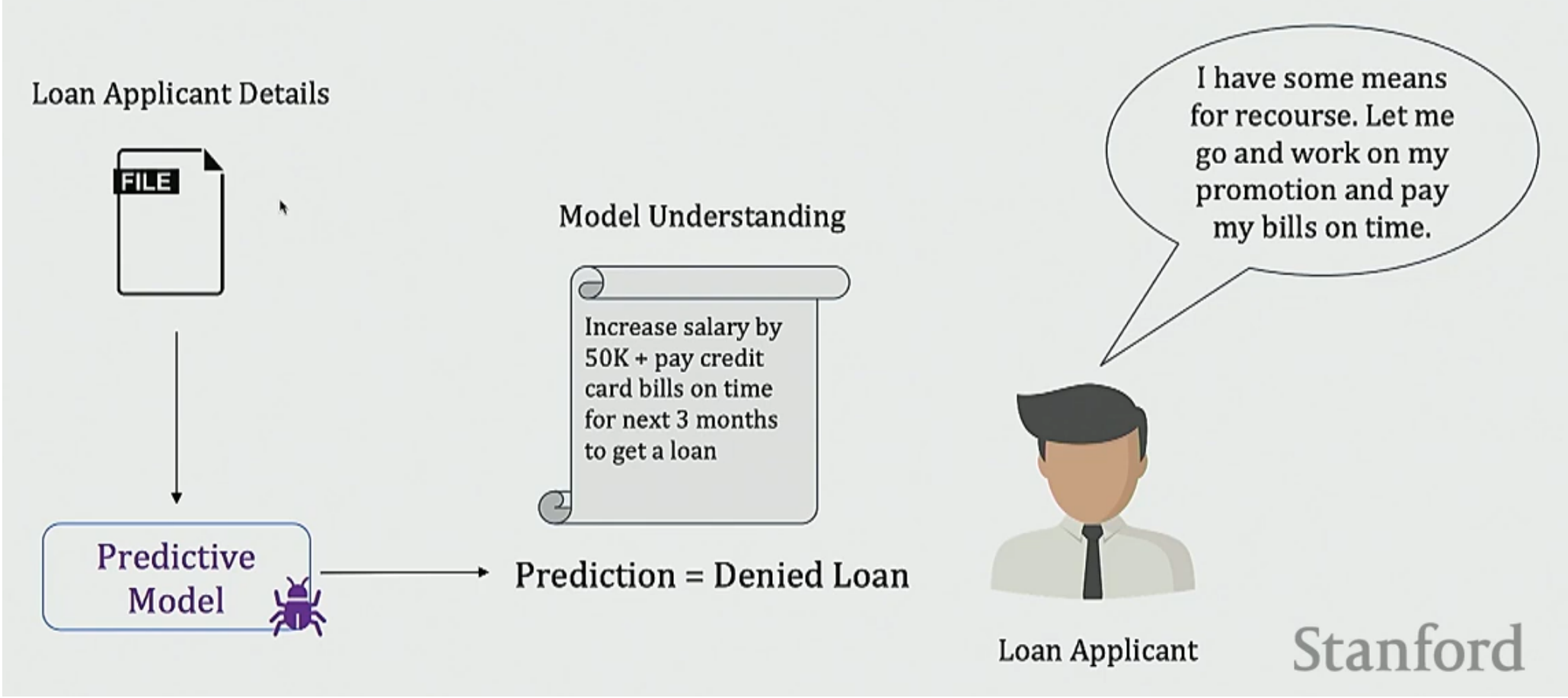
Model Transparency ⇒ Bias Identification



Model Transparency ⇒ Assess Trustworthiness



Model Transparency ⇒ Recourse to Users



Summary: Why Model Transparency?

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)

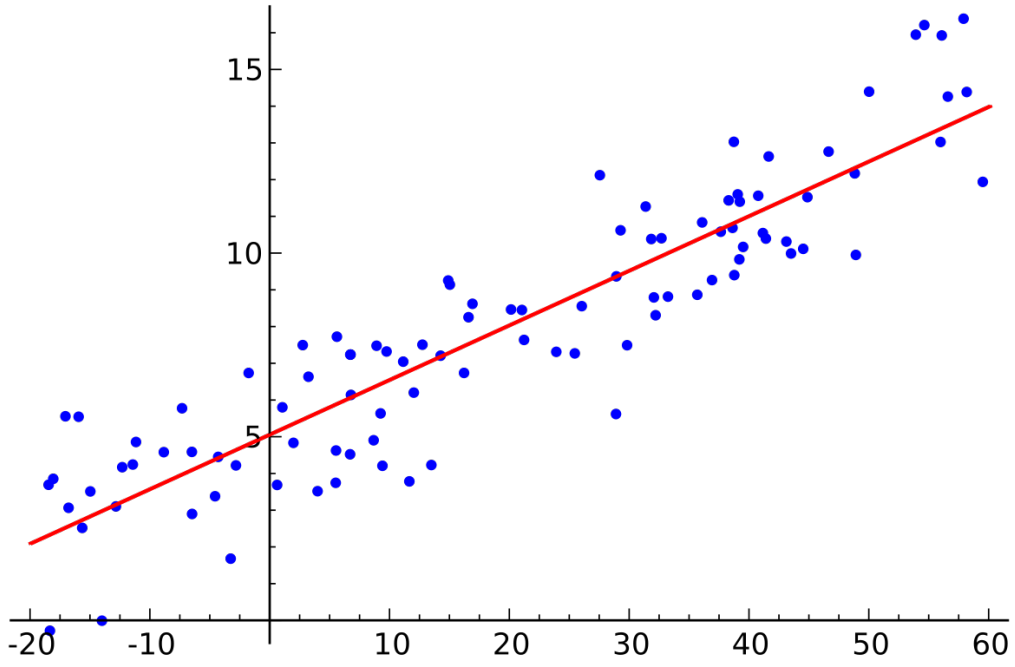
Decision makers (e.g., doctors, judges)

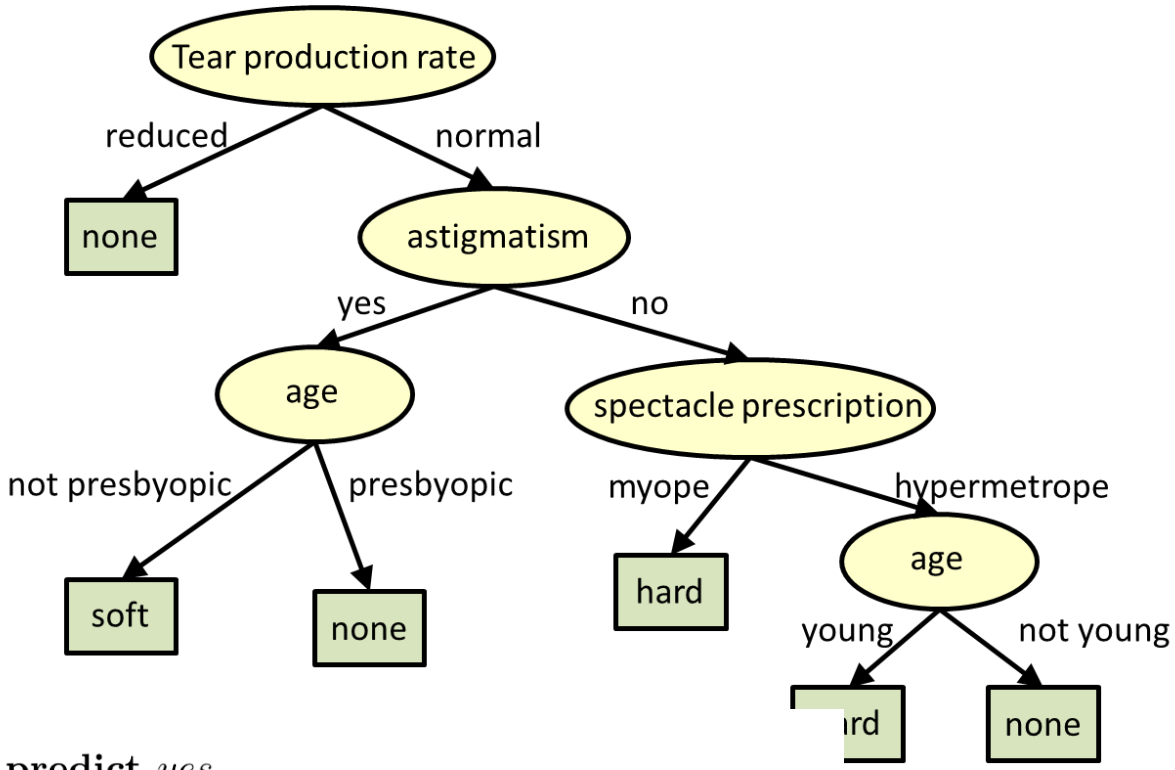
Regulatory agencies (e.g., FDA, European commission)

Researchers and engineers

How can we design ML models that are transparent?

Approach 1: Inherently Interpretable Models

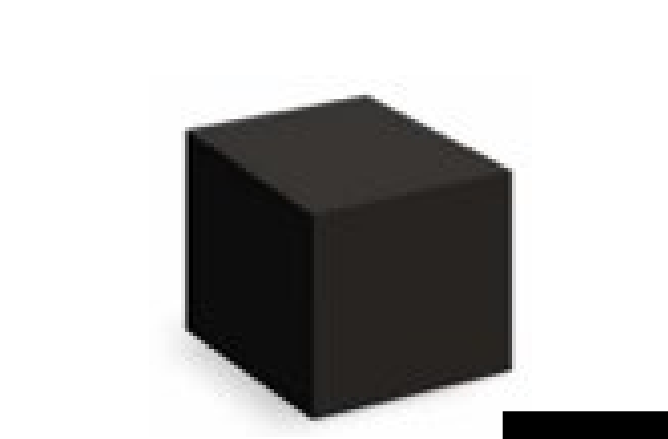


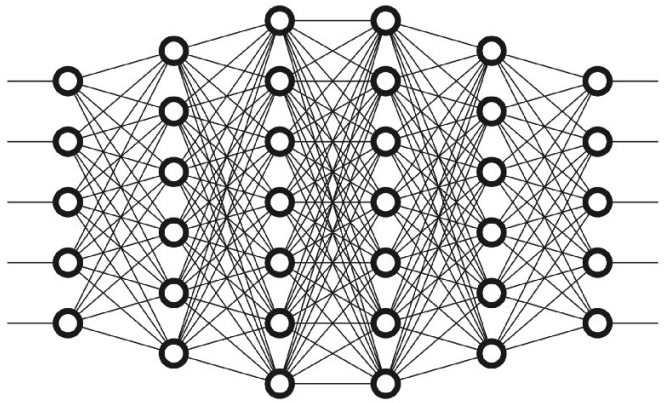


if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** *yes*
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** *yes*
else if $(priors > 3)$ **then predict** *yes*
else predict *no*

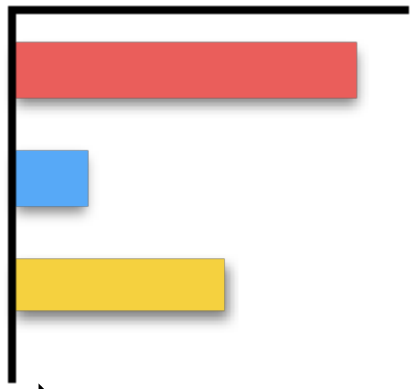
Figure 1: An example rule list that predicts two-year recidivism for the ProPublica data set, found by CORELS.

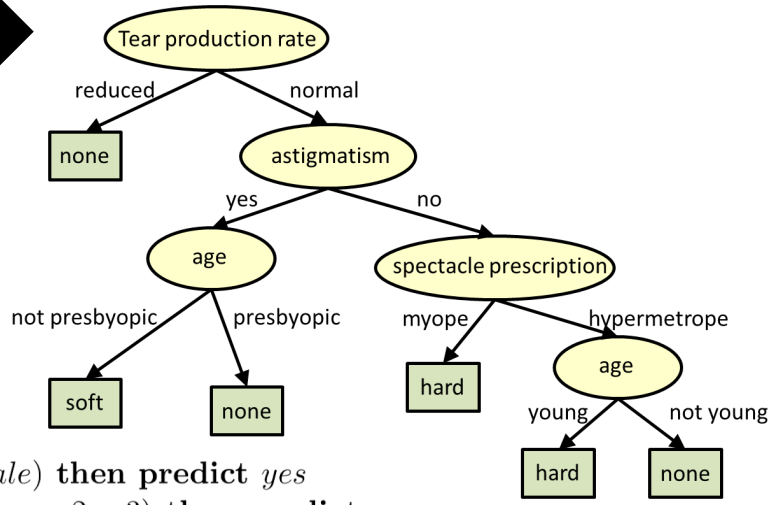
Approach 2: Post-Hoc Explanations





Explainer



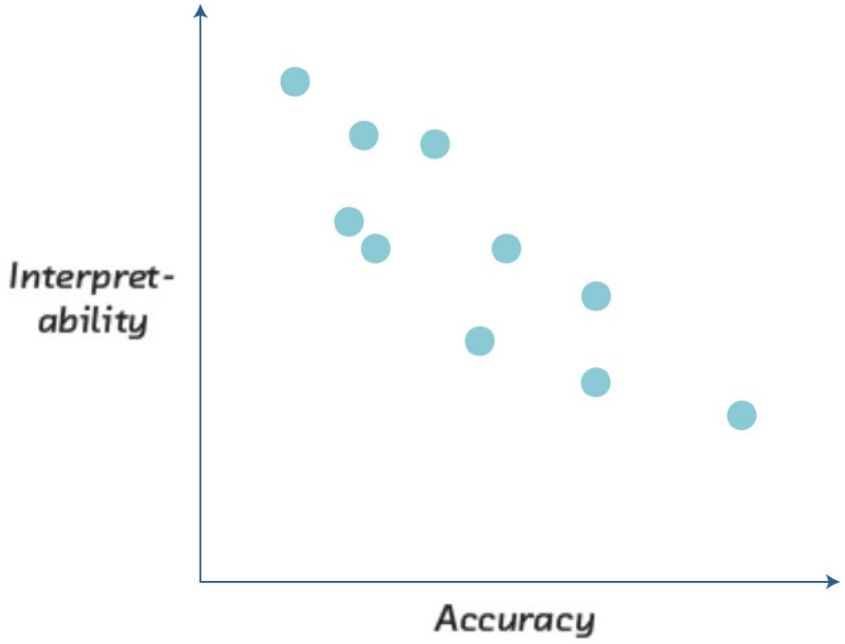


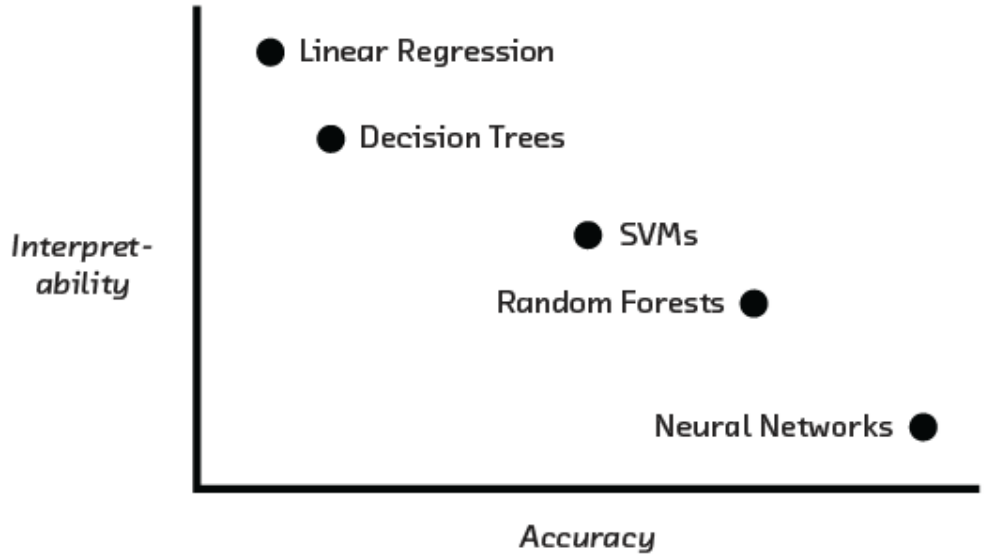
if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** *yes*
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** *yes*
else if $(priors > 3)$ **then predict** *yes*
else predict *no*

Figure 1: An example rule list that predicts two-year recidivism for the ProPublica data set, found by CORELS.

Inherently Interpretable Models vs Post-Hoc Explanations

Example

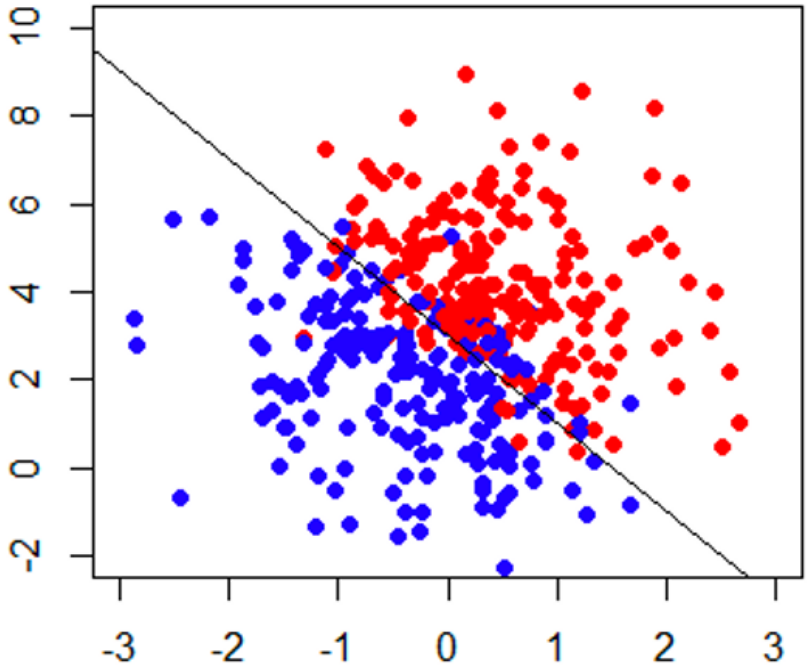




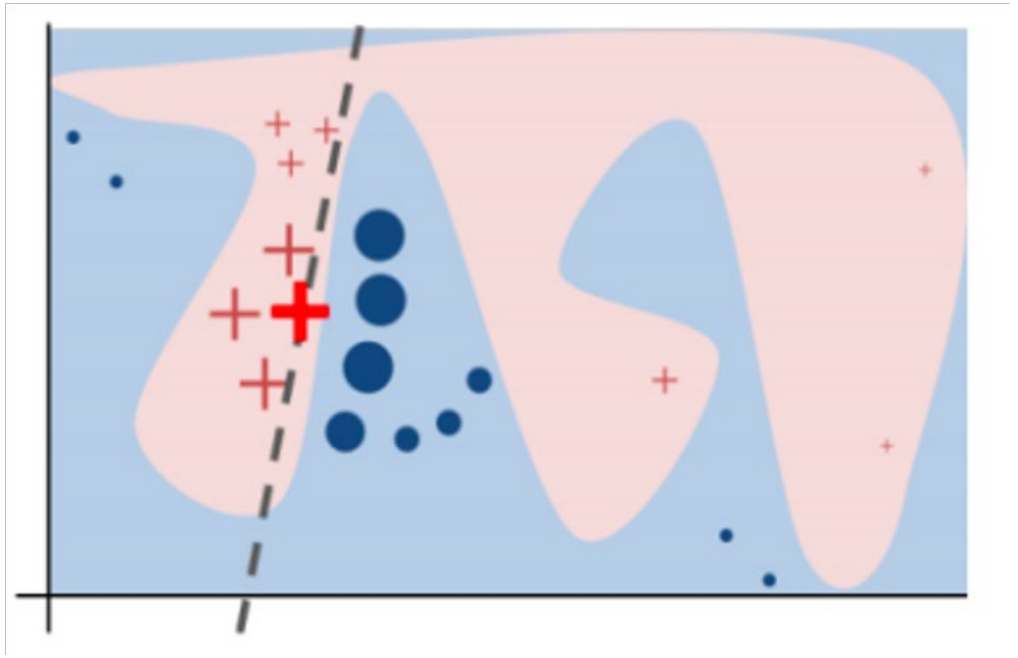
In certain settings, accuracy—interpretability tradeoff might exist ...

Cireșan et. al. 2012, Caruana et. al. 2006, Frosst et. al. 2017, Stewart 2020

Inherently Interpretable Models vs Post-Hoc Explanations



can build interpretable + accurate models



complex models might achieve higher accuracy

In certain settings, accuracy—interpretability tradeoff might exist ...

Cireșan et. al. 2012, Caruana et. al. 2006, Frosst et. al. 2017, Stewart 2020

Inherently Interpretable Models vs Post-Hoc Explanations

Sometimes ...

- You don't have enough data to build an inherently interpretable model from scratch
- You only have proprietary (black-box) models
  - No access to parameters / training data etc. ⇒ no ability to retrain
  - E.g. API-based models like ChatGPT, etc.

- If you can build an inherently interpretable model AND have high accuracy, DO IT !
  - Otherwise develop post-hoc explanation techniques

Cireșan et. al. 2012, Caruana et. al. 2006, Frosst et. al. 2017, Stewart 2020

Examples of Inherently Interpretable Models

- Rule-based Models

if hemiplegia **and** age $>$ 60 **then** *stroke risk* 58.9% (53.8%–63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8%–50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5%–28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2%–19.6%)
else if altered state of consciousness **and** age $>$ 60 **then** *stroke risk* 16.0% (12.2%–20.2%)
else if age $\leq$ 70 **then** *stroke risk* 4.6% (3.9%–5.4%)
else *stroke risk* 8.7% (7.9%–9.6%)

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

Examples of Inherently Interpretable Models

- Risk Scores
  - widely used in medicine and criminal justice
    - assess risk of mortality in ICU, assess the risk of recidivism
    - decision makers find them easy to understand

1.	*Prior Arrests ≥ 2*	1 point		···
2.	*Prior Arrests ≥ 5*	1 point	+	···
3.	*Prior Arrests for Local Ordinance*	1 point	+	···
4.	*Age at Release between 18 to 24*	1 point	+	···
5.	*Age at Release ≥ 40*	-1 point	+	···
	ADD POINTS FROM ROWS 1–5	**SCORE**	=	···

SCORE	-1	0	1	2	3	4
RISK	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

Recidivism

1.	*Call between January and March*	1 point		···
2.	*Called Previously*	1 point	+	···
3.	*Previous Call was Successful*	1 point	+	···
4.	*Employment Indicator < 5100*	1 point	+	···
5.	*3 Month Euribor Rate ≥ 100*	-1 point	+	···
	ADD POINTS FROM ROWS 1–5	**SCORE**	=	···

SCORE	-1	0	1	2	3	4
RISK	4.7%	11.9%	26.9%	50.0%	73.1%	88.1%

Loan Default

Examples of Inherently Interpretable Models

- Risk Scores
  - widely used in medicine and criminal justice
  - Until very recently, risk scores were constructed manually by domain experts.
    - Can we learn these in a data-driven fashion?

Definition 1 (Risk Score Problem, RiskSlimMINLP)
The risk score problem is a discrete optimization problem with the form:

$$\begin{aligned} \min_{\boldsymbol{\lambda}} \quad & l(\boldsymbol{\lambda}) + C_0 \|\boldsymbol{\lambda}\|_0 \\ \text{s.t.} \quad & \boldsymbol{\lambda} \in \mathcal{L}, \end{aligned} \tag{1}$$

where:

- $l(\boldsymbol{\lambda}) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-\langle \boldsymbol{\lambda}, y_i \boldsymbol{x}_i \rangle))$ *is the normalized logistic loss function;*
- $\|\boldsymbol{\lambda}\|_0 = \sum_{j=1}^{d} \mathbb{1}[\lambda_j \neq 0]$ *is the ℓ_0-seminorm;*
- $\mathcal{L} \subset \mathbb{Z}^{d+1}$ *is a set of feasible coefficient vectors (user-provided);*
- $C_0 > 0$ *is a trade-off parameter to balance fit and sparsity (user-provided).*

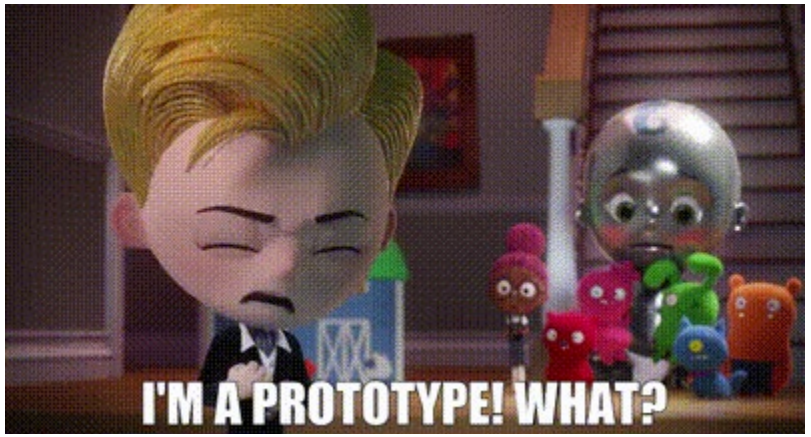
Examples of Inherently Interpretable Models

- Generalized Additive Models (GAMs)

Model	Form	Intelligibility	Accuracy
Linear Model	$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Generalized Linear Model	$g(y) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$	+++	+
Additive Model	$y = f_1(x_1) + ... + f_n(x_n)$	++	++
Generalized Additive Model	$g(y) = f_1(x_1) + ... + f_n(x_n)$	++	++
Full Complexity Model	$y = f(x_1, ..., x_n)$	+	+++

Examples of Inherently Interpretable Models

- Prototype-based Models



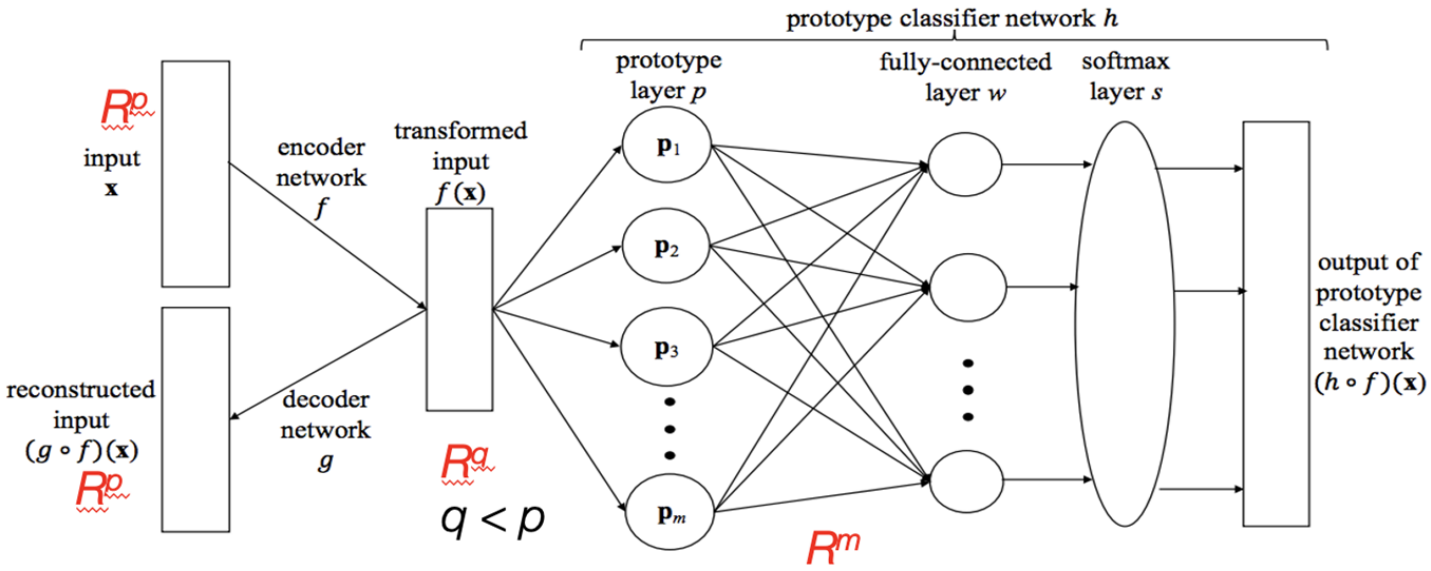
  - Goal: identify K prototypes (instances) from the data s.t. a new instance which will be assigned the same label as the closest prototype will be correctly classified (with a high probability)

  - Let each instance "cover" the ϵ - neighborhood around it.

  - Once we define the neighborhood covered by each instance
    - this problem becomes similar to the problem of finding rule sets

Examples of Inherently Interpretable Models

- Prototype-based Deep Learning



Prototype layer is responsible for computing the prototypes

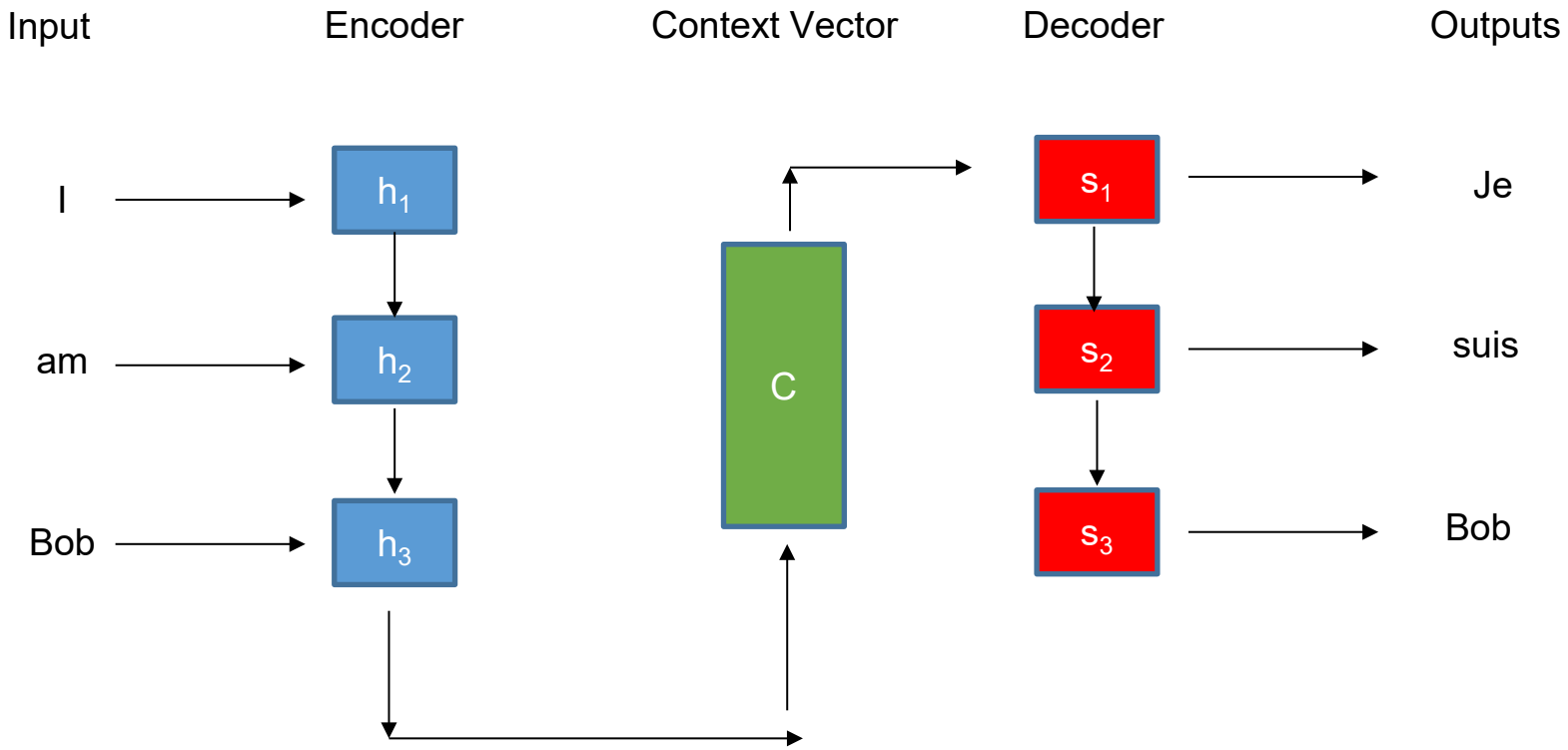
$$\mathbf{z} = f(\mathbf{x}) \quad p(\mathbf{z}) = \left[\|\mathbf{z} - \mathbf{p}_1\|_2^2, \quad \|\mathbf{z} - \mathbf{p}_2\|_2^2, \quad \ldots \quad \|\mathbf{z} - \mathbf{p}_m\|_2^2\right]^\top$$

Each node in layer p computes one of the above elements

[Li et. al. 2017, Chen et. al. 2019]

Examples of Inherently Interpretable Models

- Attention-based Models
  - E.g. machine translation



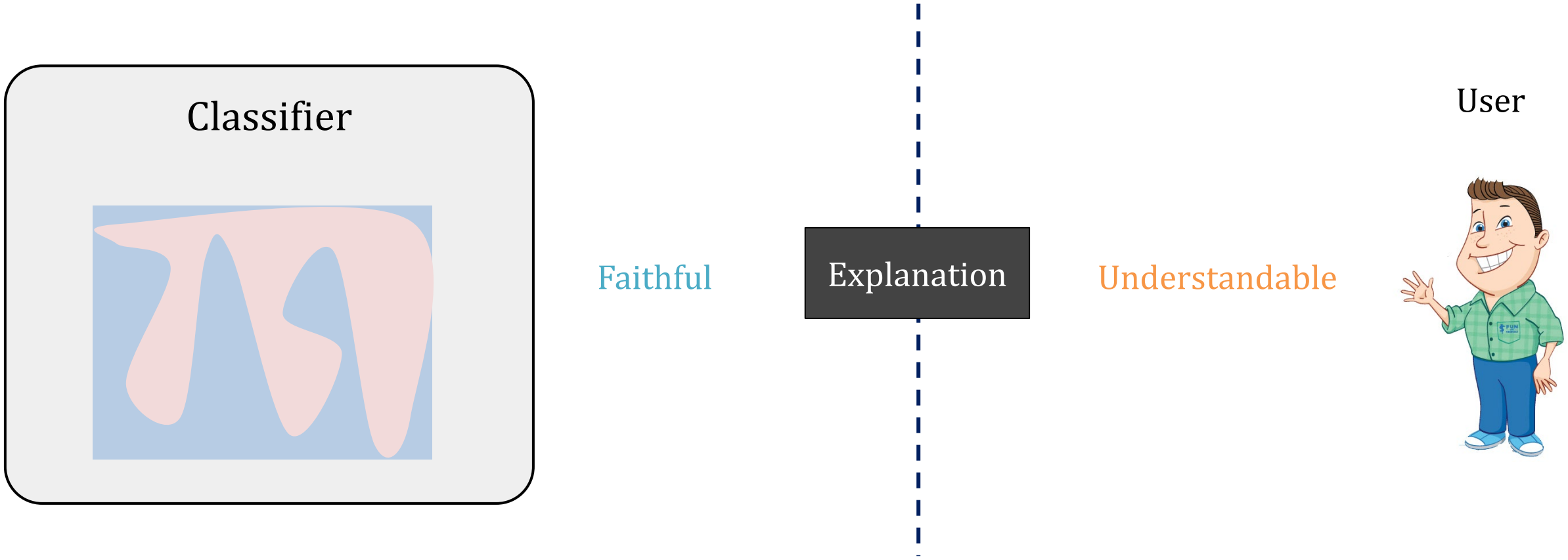
- Context vector

$$c_i = \sum_{j=1}^{3} a_{ij} h_j$$

- a_{ij} captures the attention placed on input token j when determining the decoder hidden state s_i; it can be computed as a softmax of the "match" between s_{i-1} and h_j

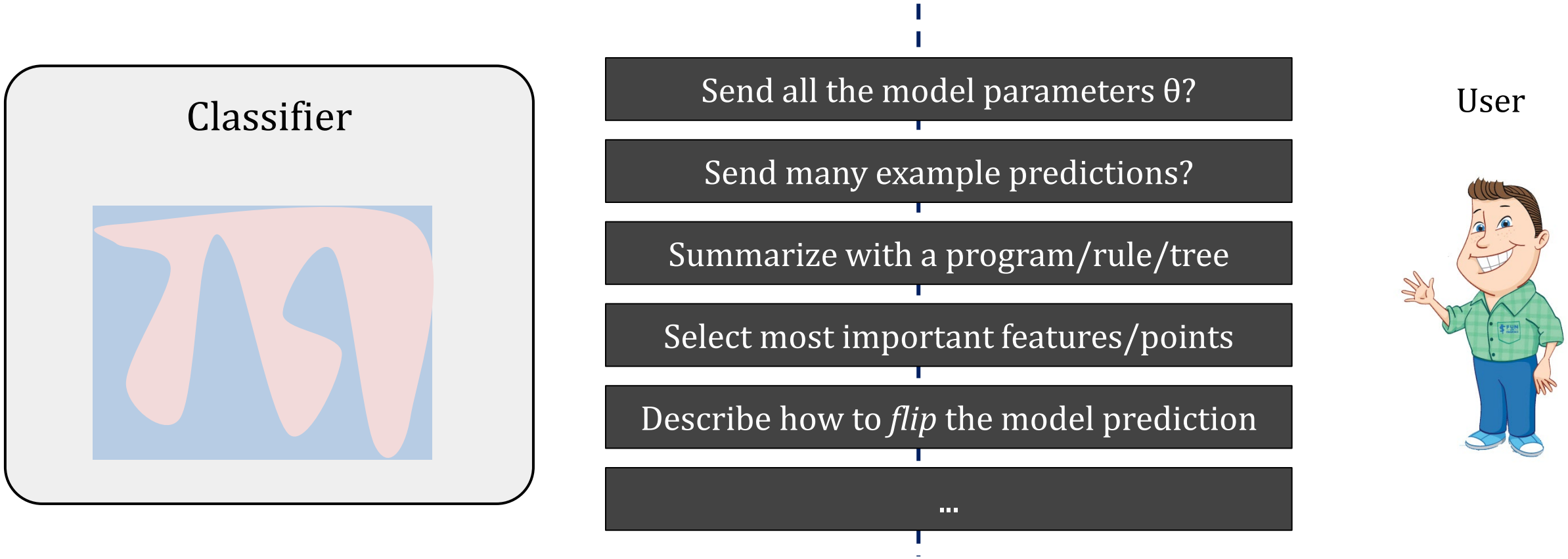
Approach 2: Explanations

- What is an explanation? An "understandable" description of model behavior



Approach 2: Explanations

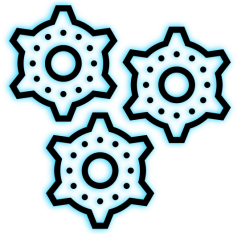
- What is an explanation? An "understandable" description of model behavior



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

Approaches for Post hoc Explainability



Local Explanations

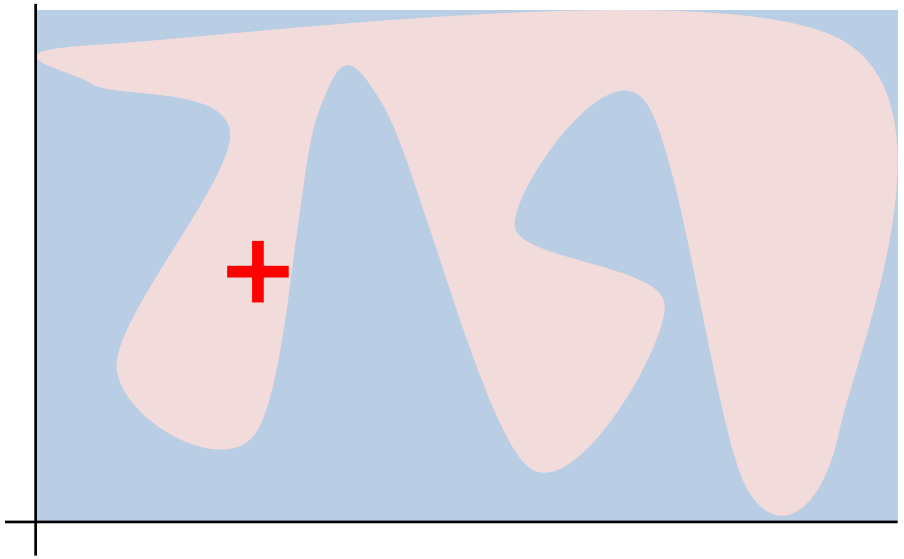
- Feature Importances
- Rule Based
- Saliency Maps
- Prototypes/Example Based
- Counterfactuals

Global Explanations

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

LIME: Local Interpretable Model-Agnostic Explanations

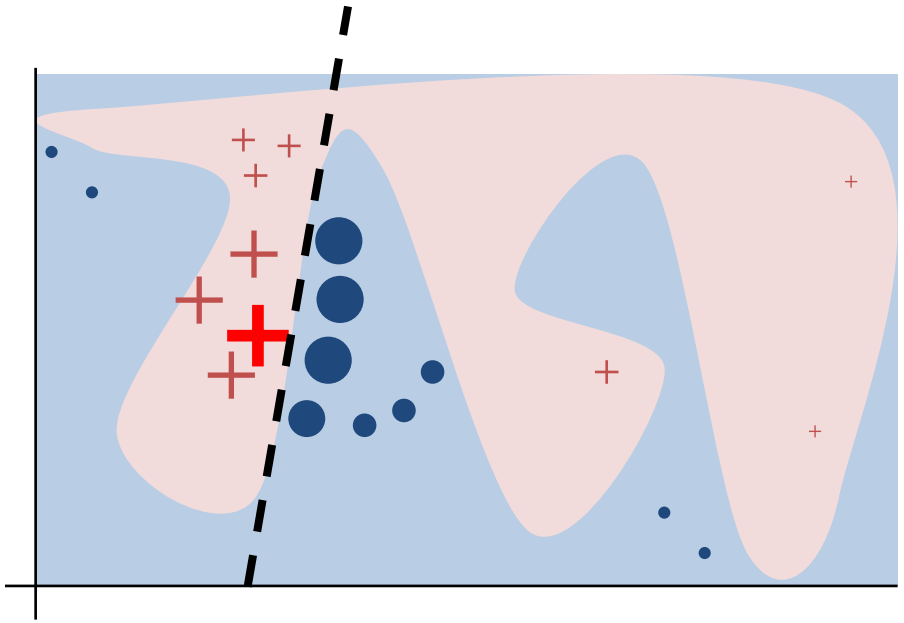
1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn simple linear model on weighted samples
5. Use simple linear model to explain



Riberio, Singh, and Guestrin. KDD 2016

LIME: Local Interpretable Model-Agnostic Explanations

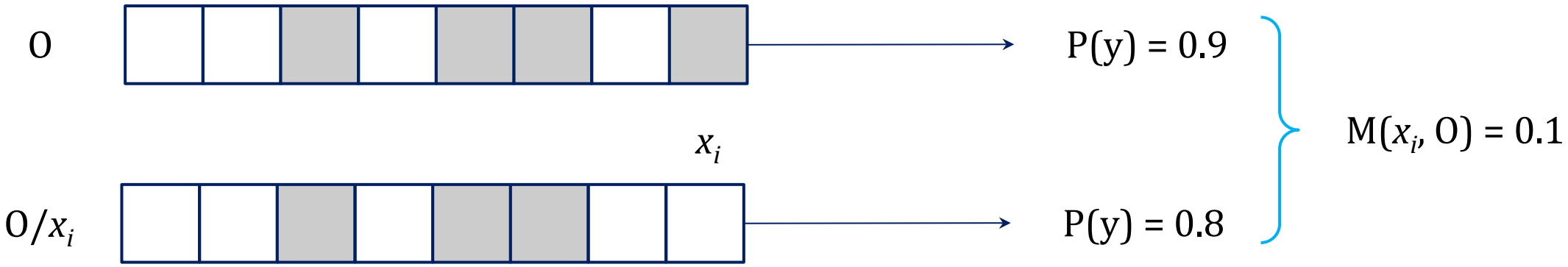
1. Sample points around x_i
2. Use model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn simple linear model on weighted samples
5. Use simple linear model to explain



Riberio, Singh, and Guestrin. KDD 2016

SHAP: Shapley Values as Importance

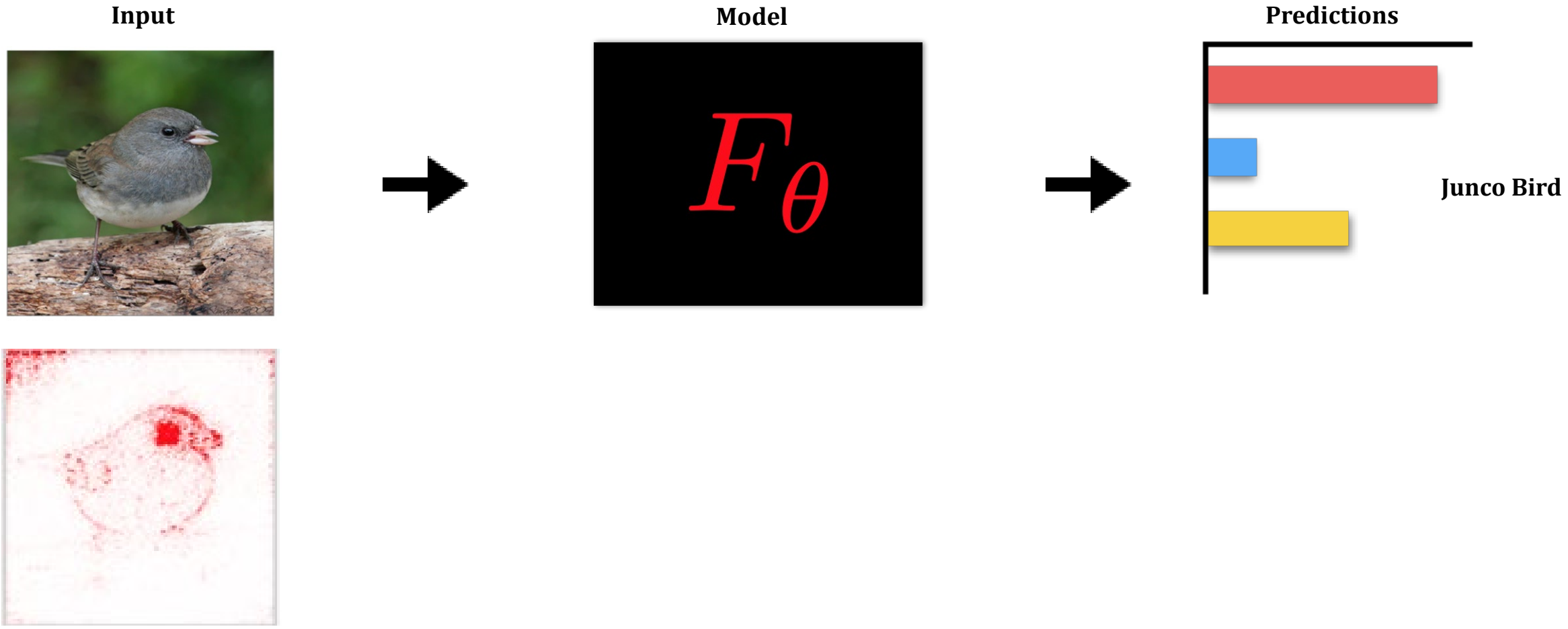
- Estimate Marginal contribution of each feature towards the prediction, averaged over all possible permutations.
- Example: what is the marginal contribution of feature x_i ?



- The prediction is "attributed to" each feature.
  - But .... Deep learning has way too many features and way too many permutations ...

Saliency Maps

- What parts of the input are most relevant to the model's prediction?



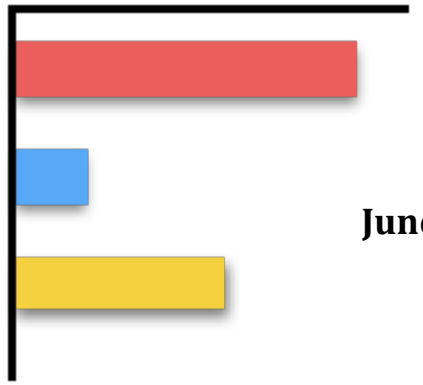
Input-Gradient

Input

Model

F_θ

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 .

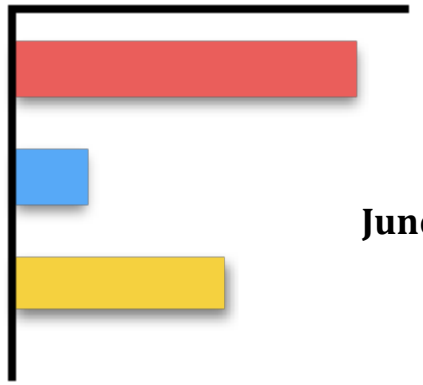
Input-Gradient

Input

Model

F_θ

Predictions



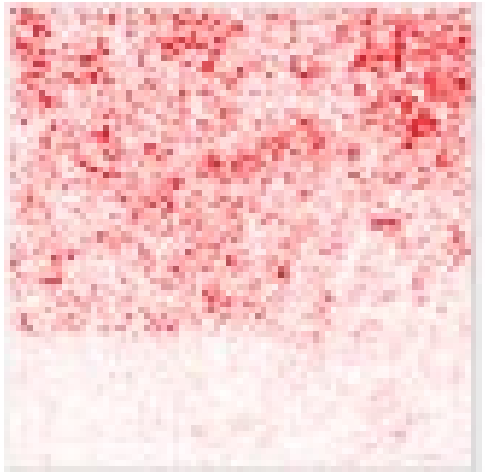
Junco Bird

Input-Gradient

$$\nabla_x F_i(x)$$

Logit

Input



Visualize as a heatmap

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

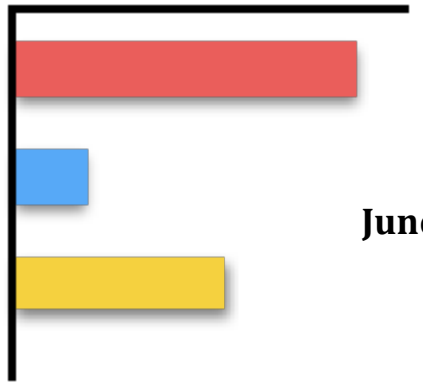
Input-Gradient

Input

Model

F_θ

Predictions



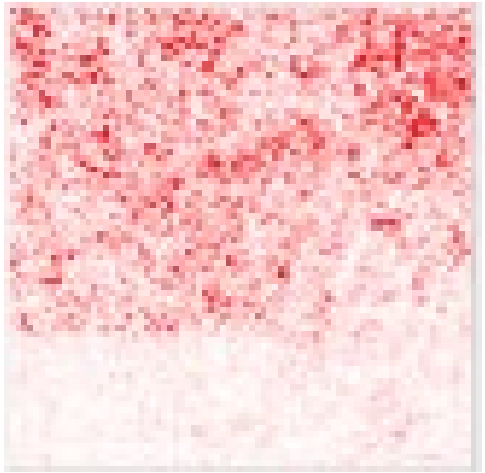
Junco Bird

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 .

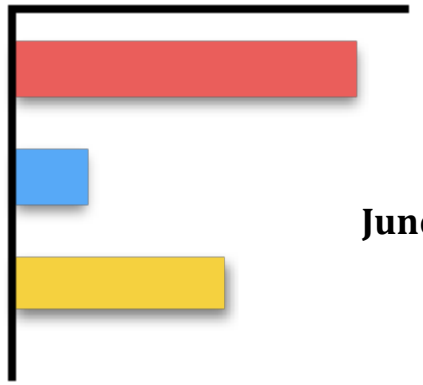
SmoothGrad

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]

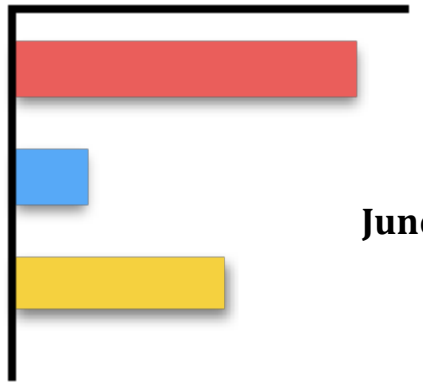
SmoothGrad

Input

Model

F_θ

Predictions

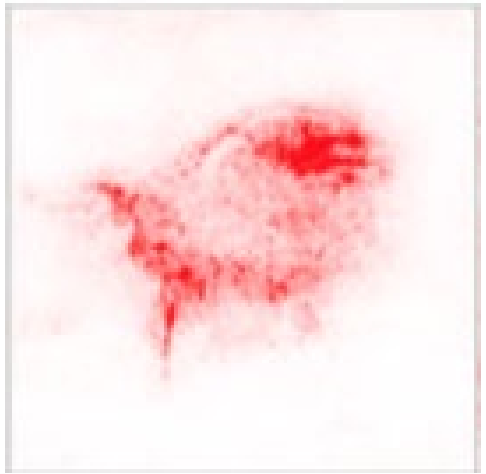


Junco Bird

SmoothGrad

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

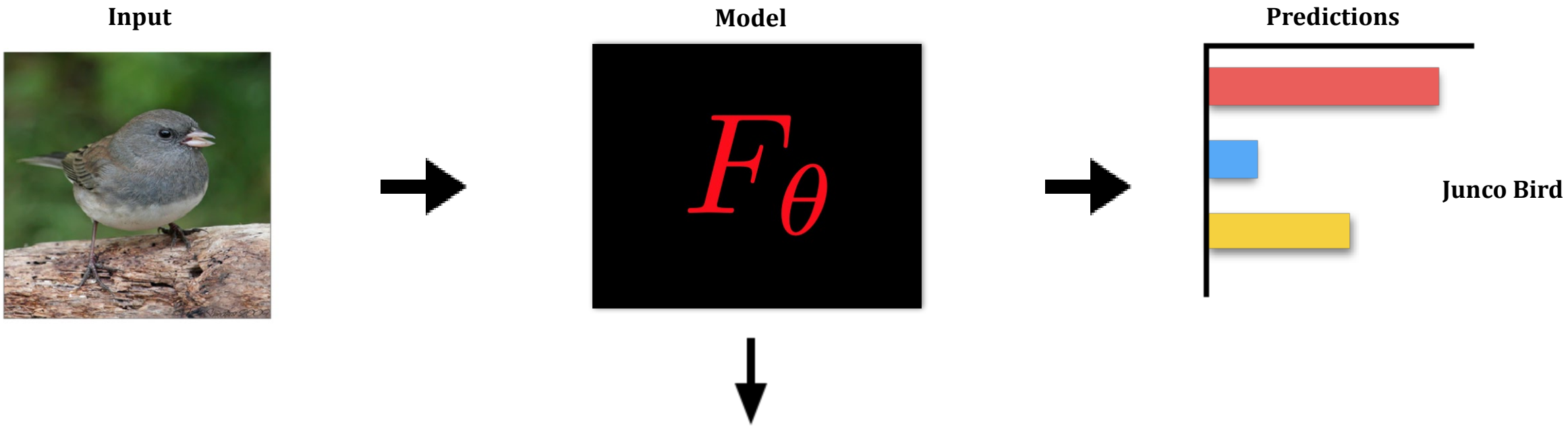
Gaussian noise



Average Input-gradient of 'noisy' inputs.

[Smilkov et. al. 2017]

Integrated Gradients



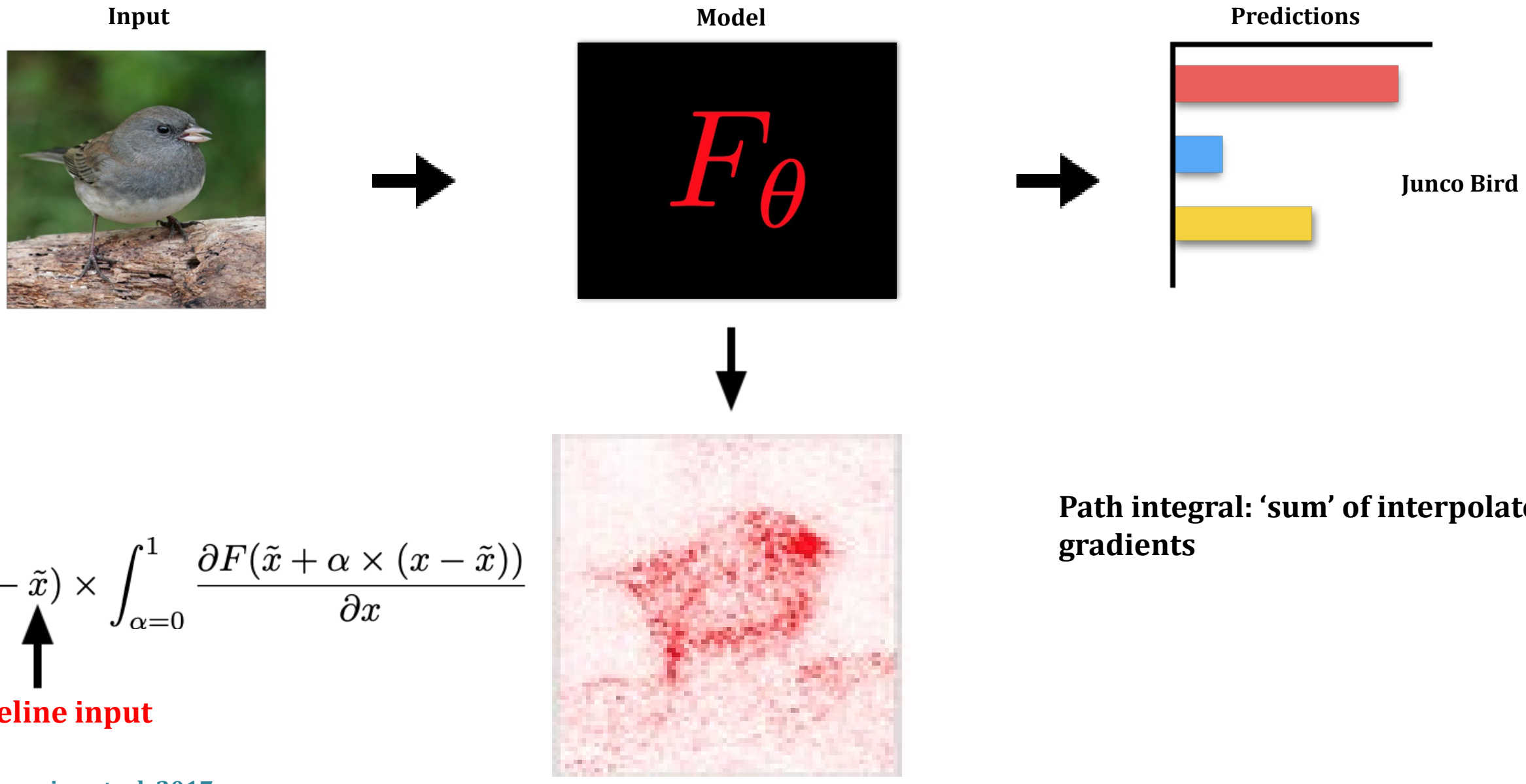
$$(x - \tilde{x}) \times \int_{\alpha=0}^{1} \frac{\partial F(\tilde{x} + \alpha \times (x - \tilde{x}))}{\partial x}$$

Baseline input

Path integral: 'sum' of interpolated gradients

Sundararajan et. al. 2017

Integrated Gradients



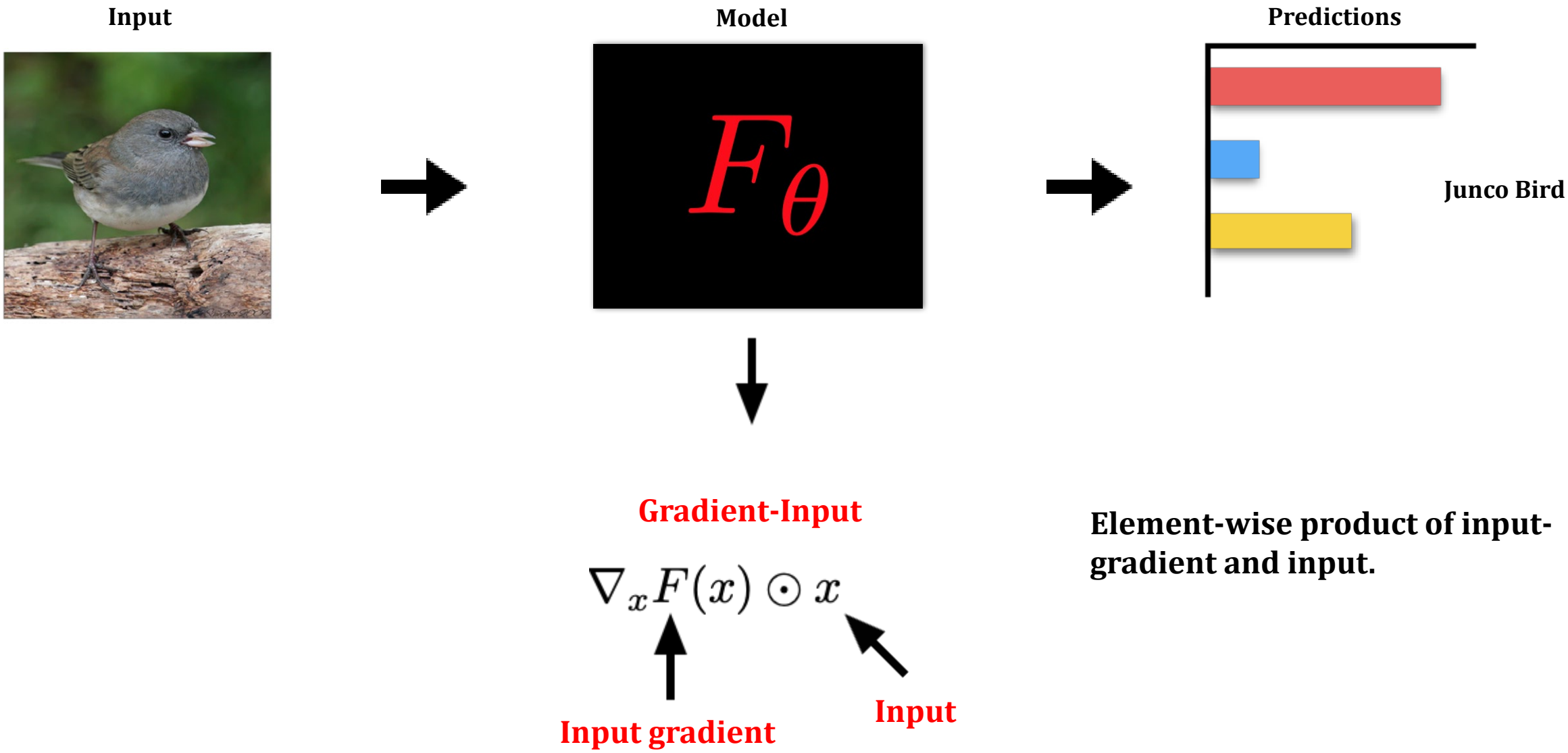
$$(x - \tilde{x}) \times \int_{\alpha=0}^{1} \frac{\partial F(\tilde{x} + \alpha \times (x - \tilde{x}))}{\partial x}$$

Baseline input

Path integral: 'sum' of interpolated gradients

[Sundararajan et. al. 2017]

Gradient-Input



Element-wise product of input-gradient and input.

Shrikumar et. al. 2017, Ancona et. al. 2018.

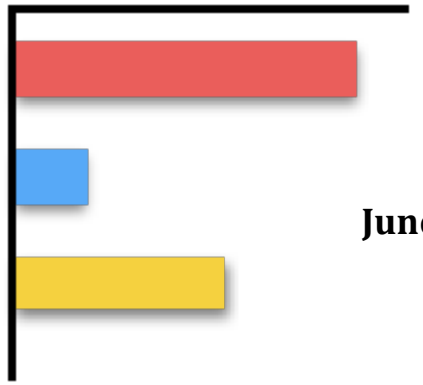
Gradient-Input

Input

Model

F_θ

Predictions



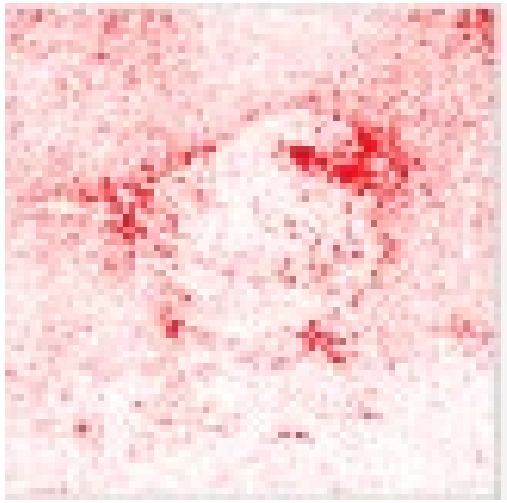
Junco Bird

Gradient-Input

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

logit gradient

Input

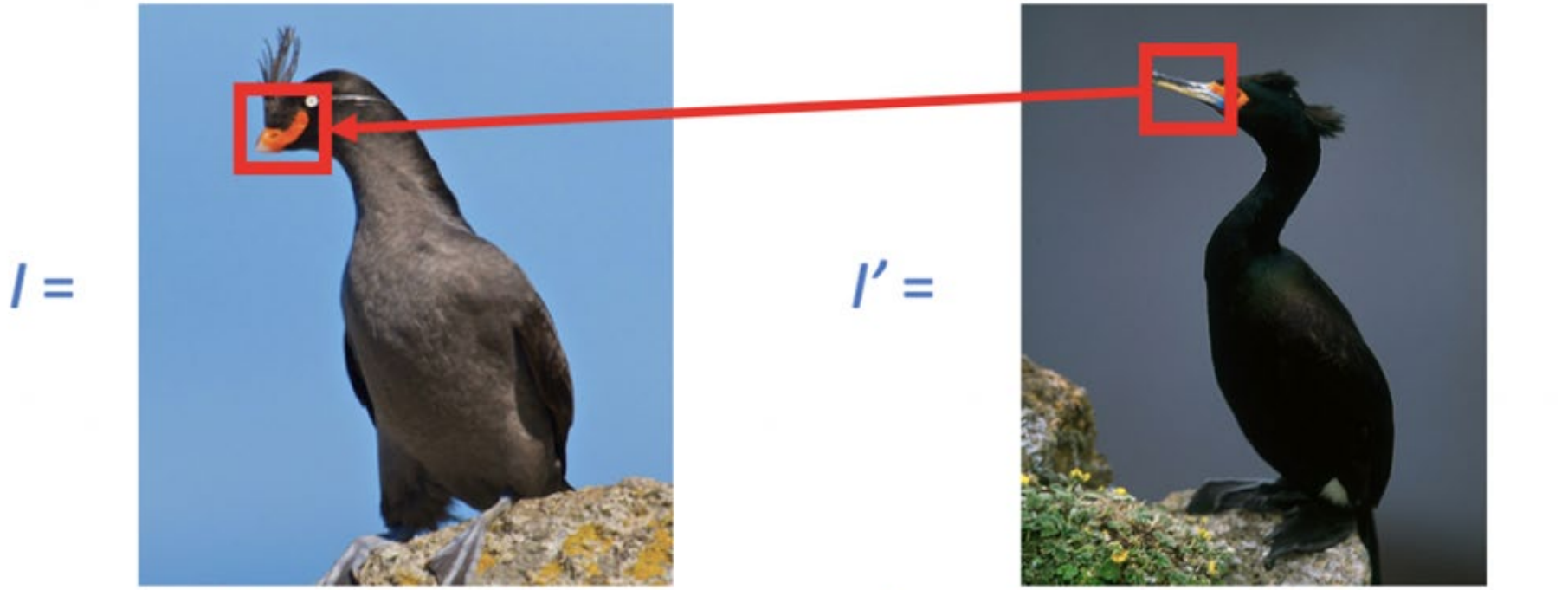


Element-wise product of input-gradient and input.

Shrikumar et. al. 2017, Ancona et. al. 2018.

Counterfactual Explanations

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



[Goyal et. al., 2019]

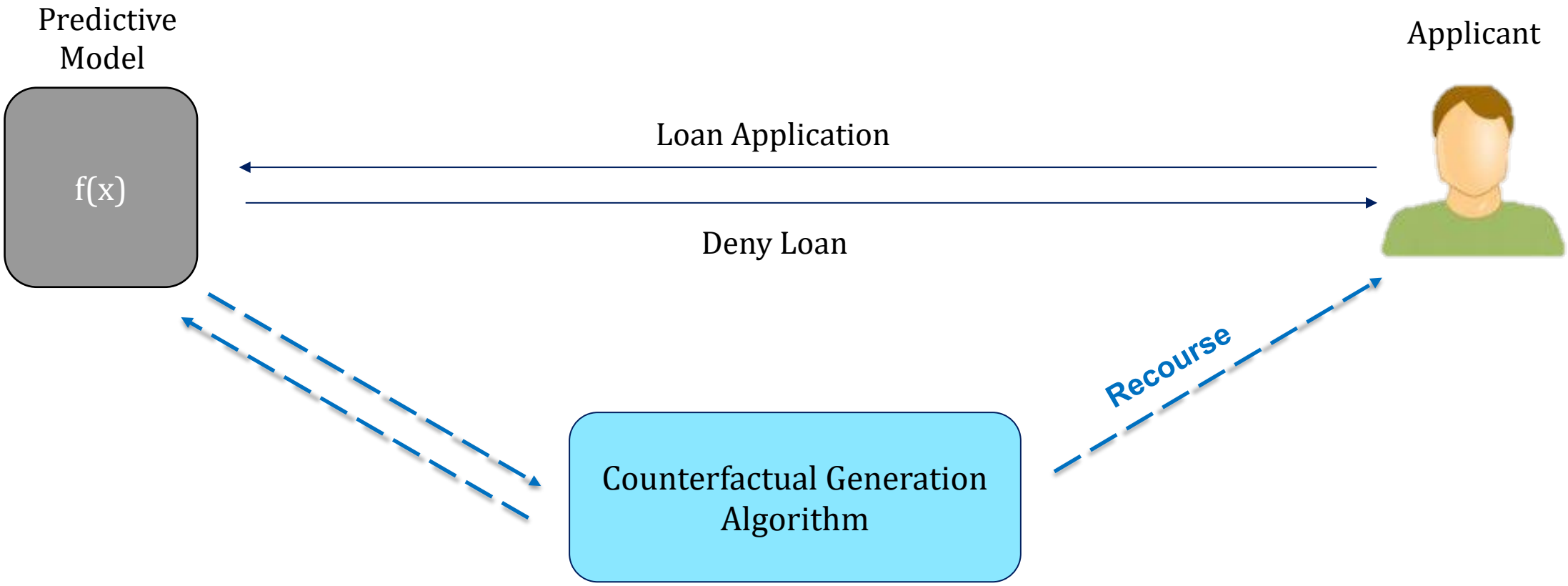
Counterfactual Explanations

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

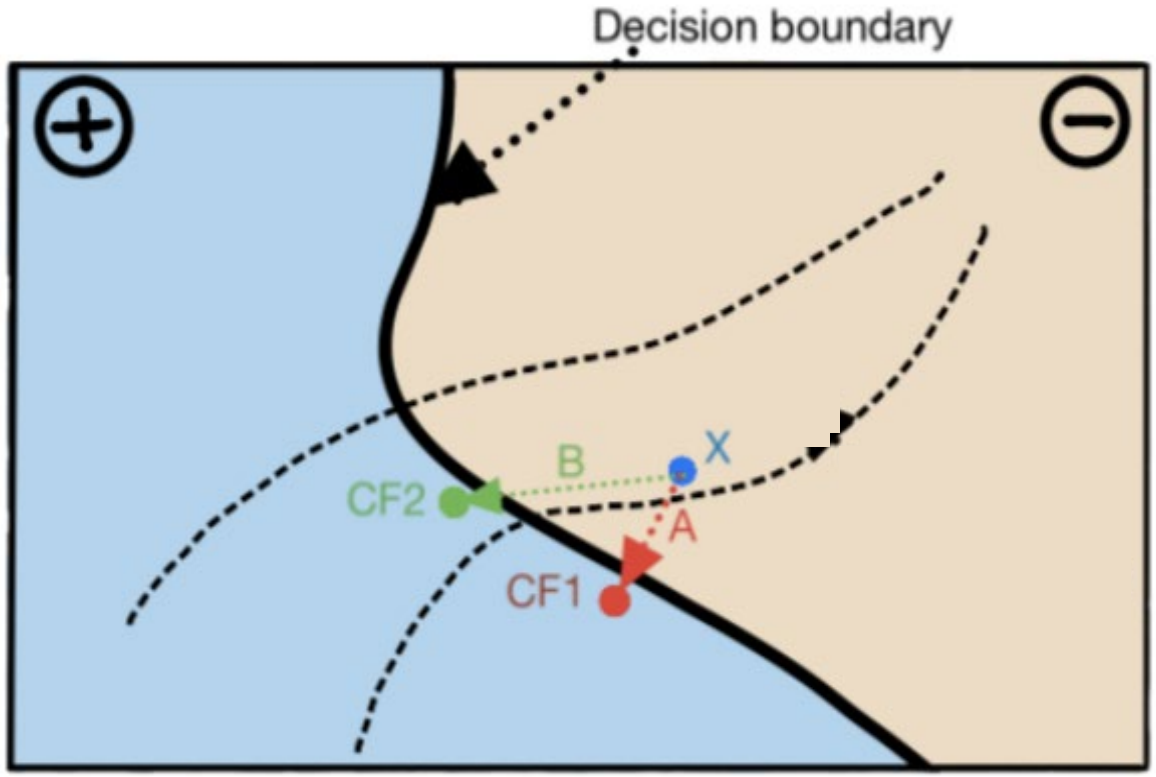
What features need to be changed and by how much to flip a model's prediction ?
(i.e., to reverse an unfavorable outcome).

Counterfactual Explanations



Recourse: Increase your salary by 5K & pay your credit card bills on time for next 3 months

Generating Counterfactual Explanations: Intuition



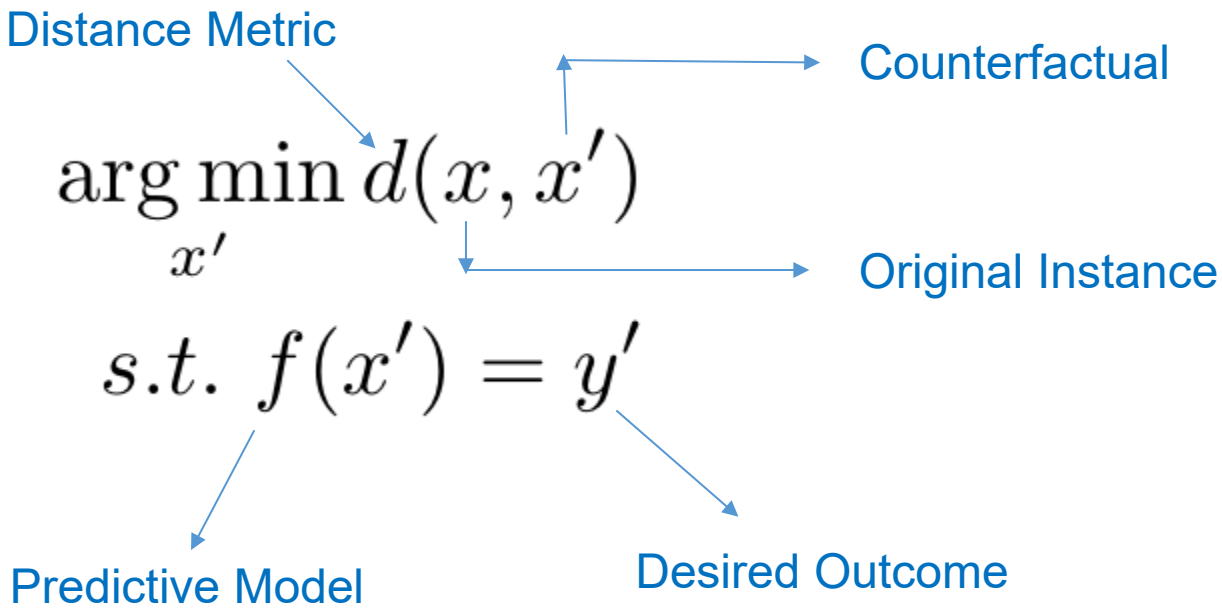
Proposed solutions differ on:

1. How to choose among candidate counterfactuals?

1. How much access is needed to the underlying predictive model?

[Wachter et. al., 2018]

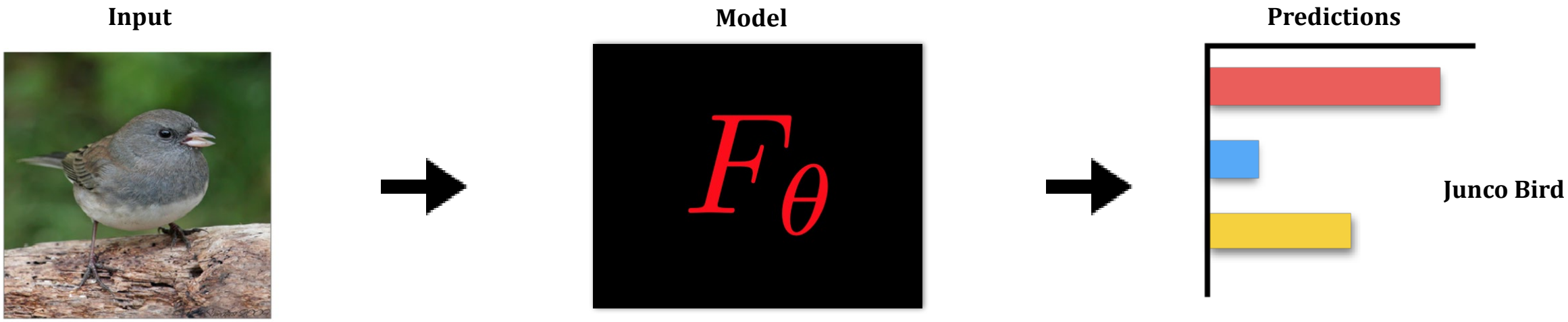
Take 1: Minimum Distance Counterfactuals



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

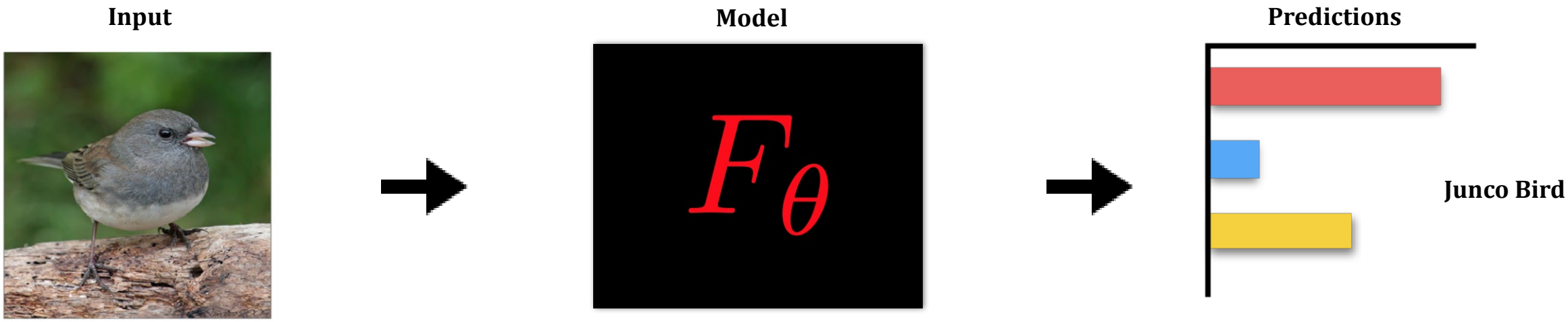
Wachter et. al. use normalized Manhattan distance.

Training Point Ranking via Influence Functions



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

Training Point Ranking via Influence Functions



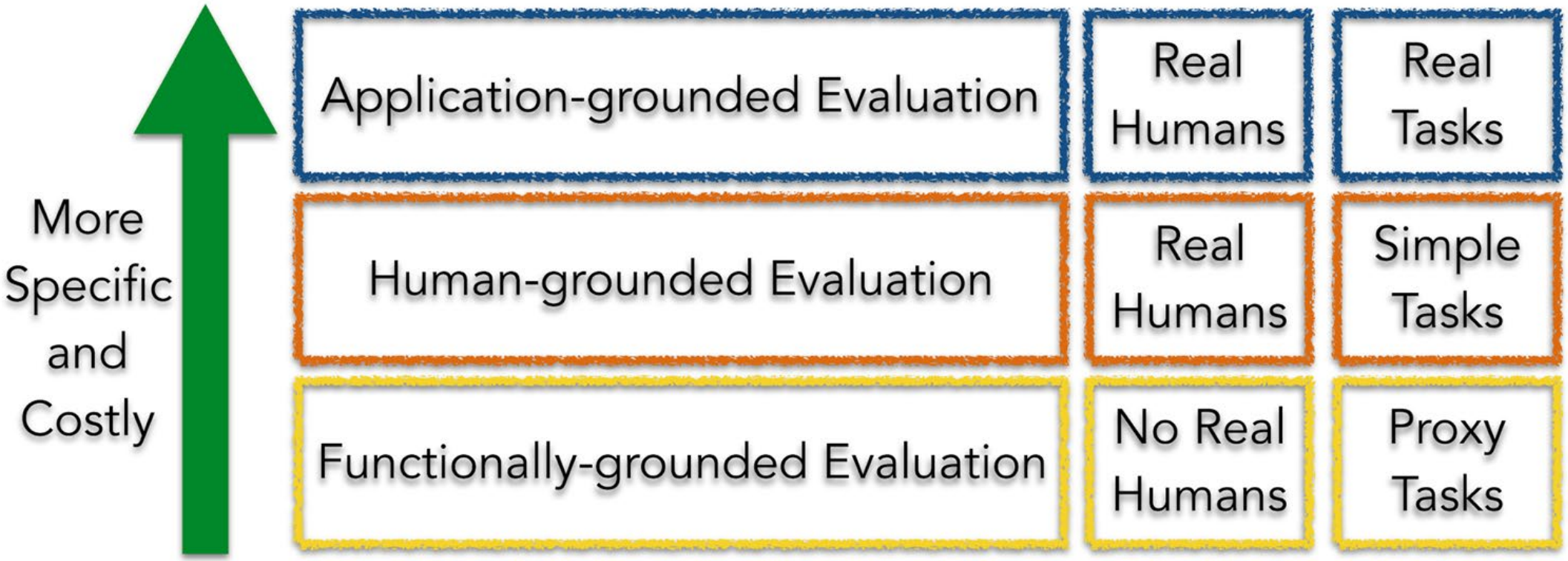
Which training data points have the most '*influence*' on the test loss?

Evaluating Interpretations/Explanations

- Evaluating the meaningfulness or correctness of explanations
  - Diverse ways of doing this depending on the type of model interpretation/explanation
- Evaluating the interpretability of explanations

[Doshi-Velez and Kim, 2017]

Evaluating Interpretability



[Agarwal et. al., 2022]

Evaluating Post hoc Explanations

- Evaluating the faithfulness (or correctness) of post hoc explanations
- Evaluating the stability of post hoc explanations
- Evaluating the fairness of post hoc explanations
- Evaluating the interpretability of post hoc explanations

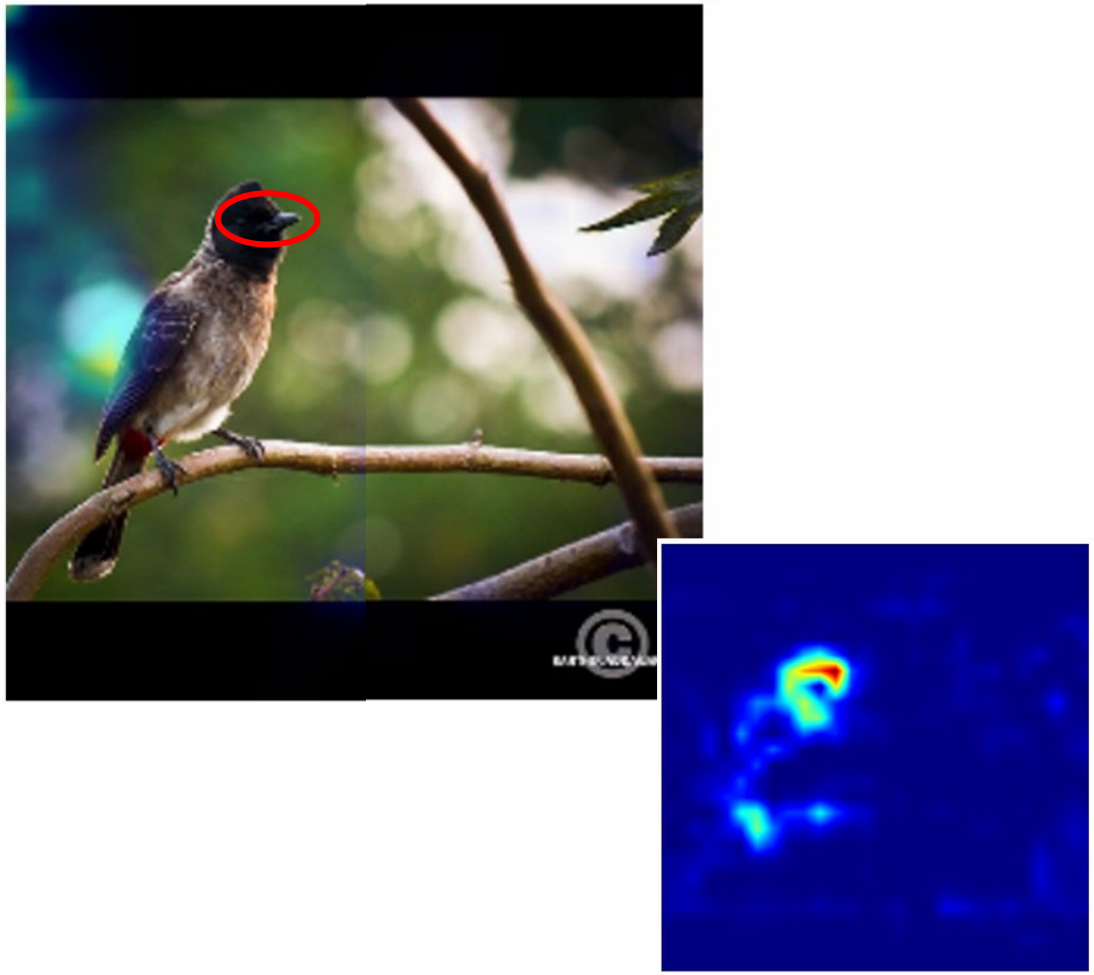
Evaluating Faithfulness of Post hoc Explanations – Explanations as Models

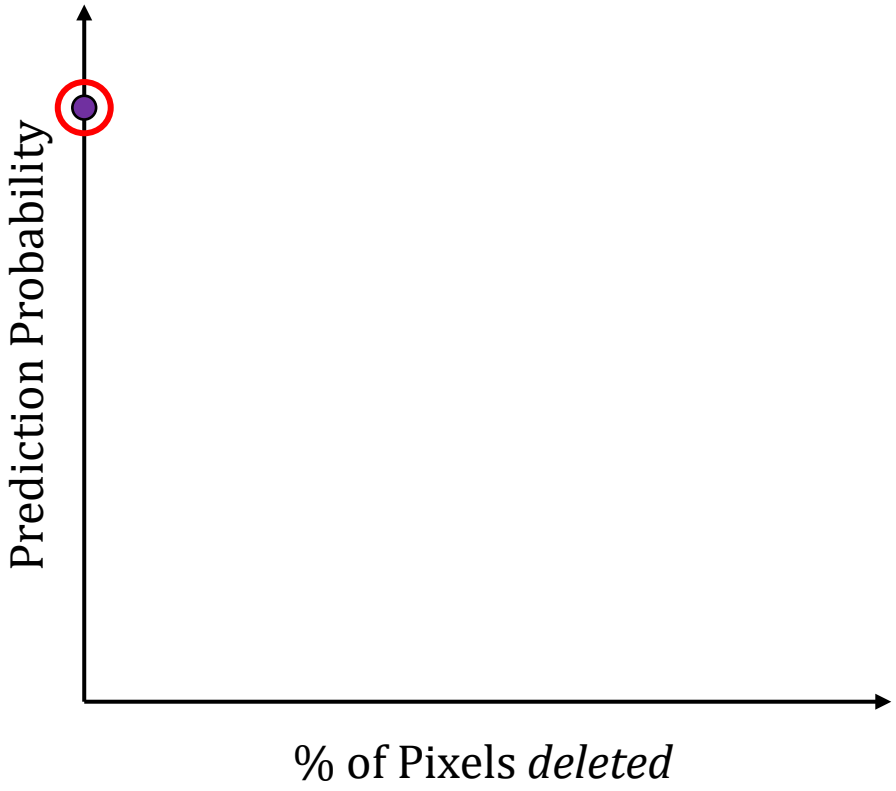
- If the explanation is itself a model (e.g., linear model fit by LIME), we can compute the fraction of instances for which the labels assigned by explanation model match those assigned by the underlying model

[Qi, Khorram, Fuxin, 2020]

How important are selected features?

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

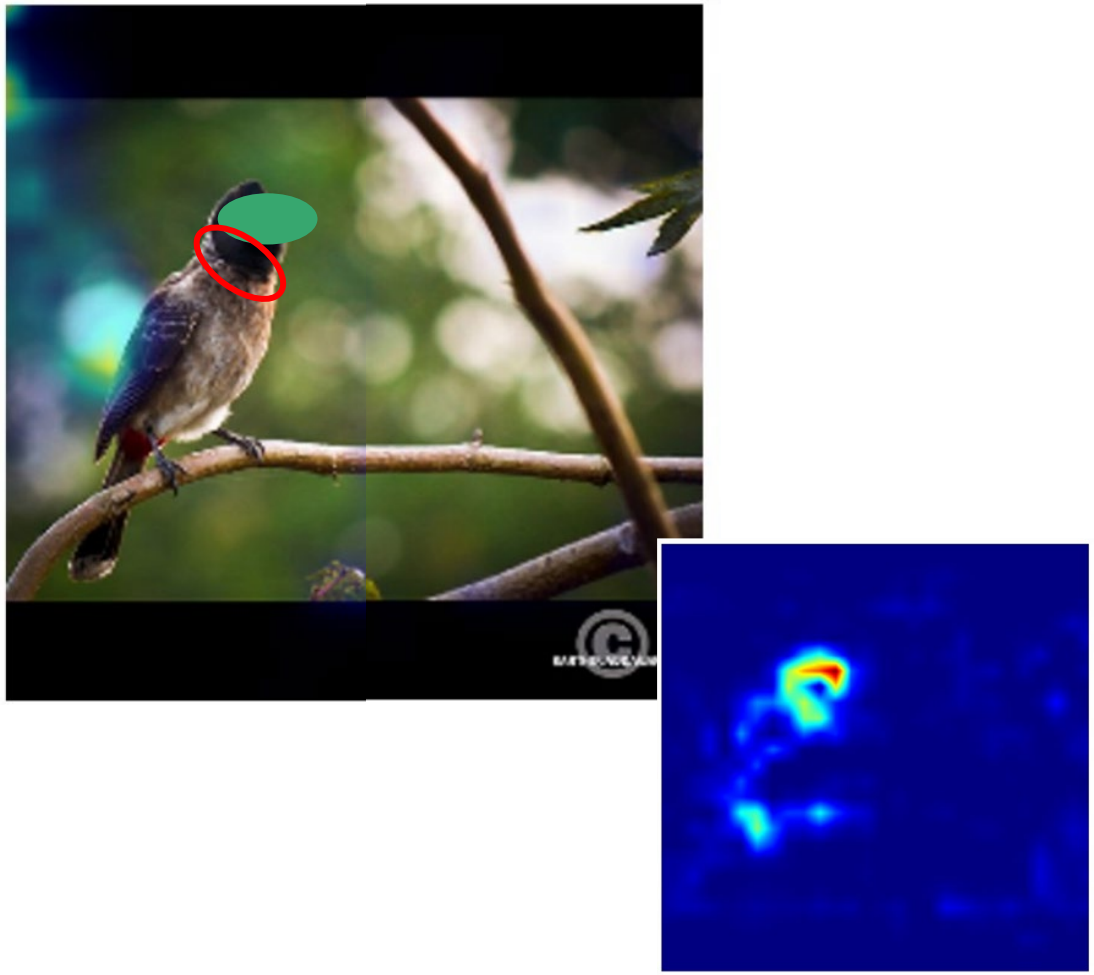


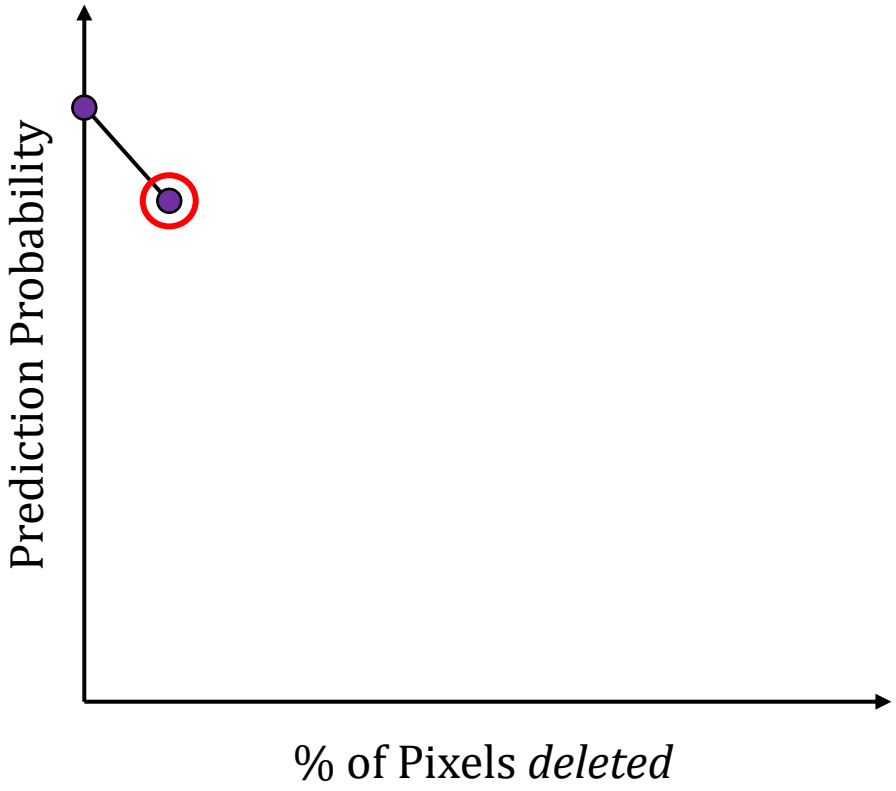


[Qi, Khorram, Fuxin, 2020]

How important are selected features?

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

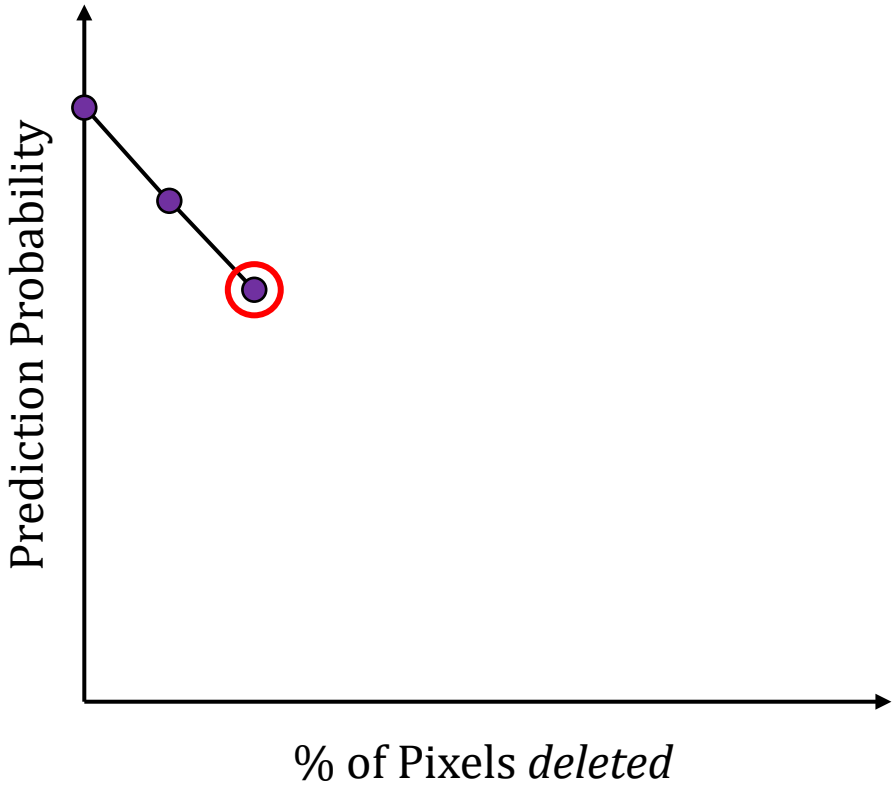




[Qi, Khorram, Fuxin, 2020]

How important are selected features?

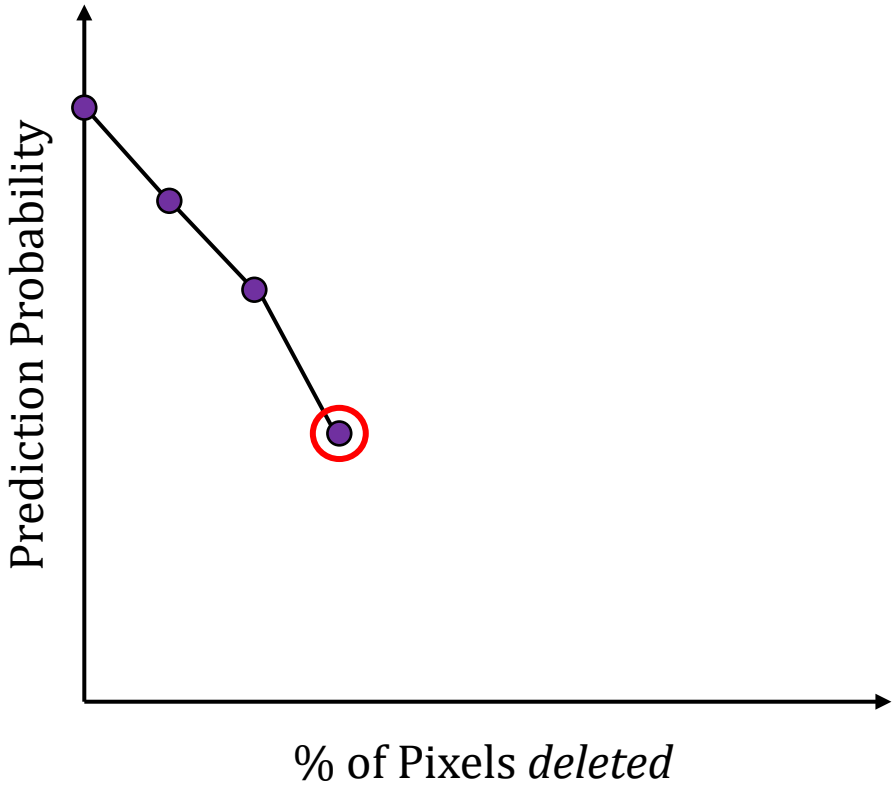
- Deletion: remove important features and see what happens..



[Qi, Khorram, Fuxin, 2020]

How important are selected features?

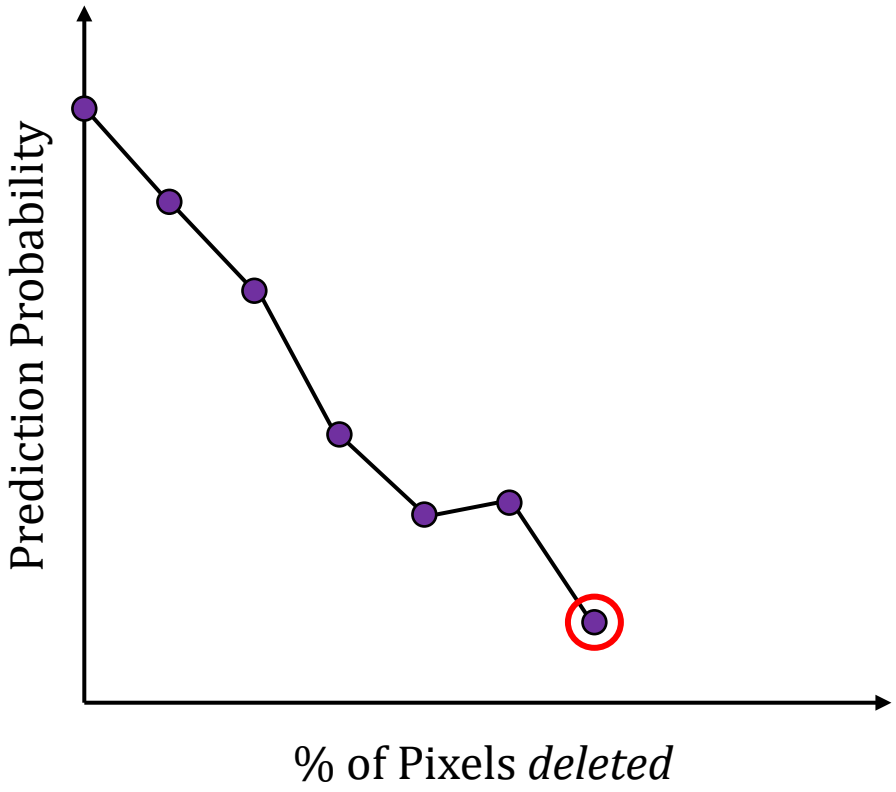
- Deletion: remove important features and see what happens..



[Qi, Khorram, Fuxin, 2020]

How important are selected features?

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



[Alvarez-Melis, 2018; Agarwal et. al., 2022]

Evaluating Stability of Post hoc Explanations

- Are post hoc explanations unstable w.r.t. small input perturbations?

Local Lipschitz Constant

Post hoc Explanation

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

Input

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

Predicting Behavior (“**Simulation**”)