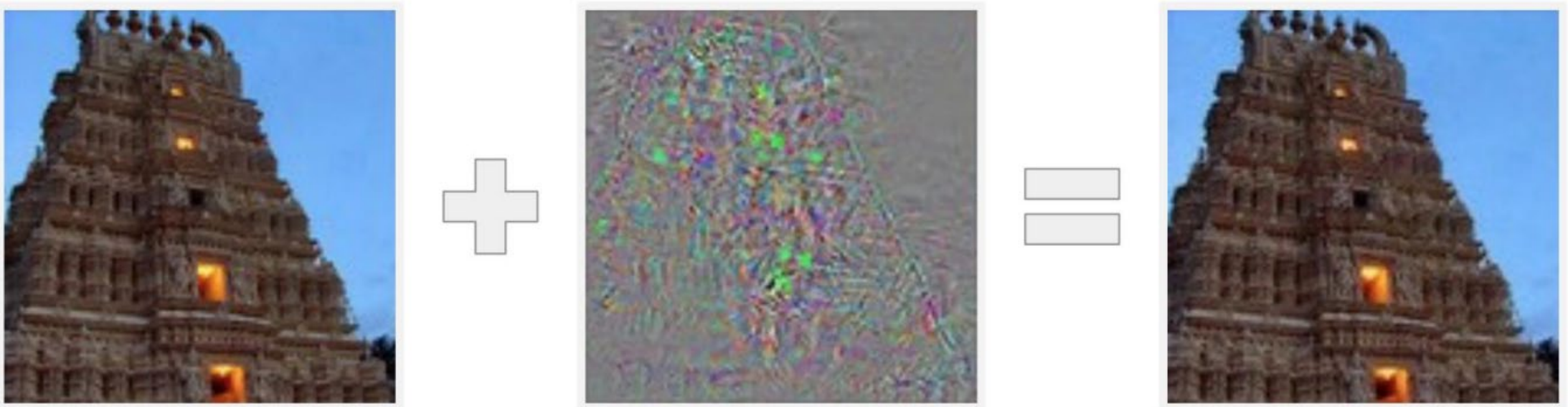
THE UNIVERSITY of TENNESSEE KNOXVILLE

Lecture 33: Explainable ML

Instructor: Sergei V. Kalinin

Adversarial attacks



https://course.ccs.neu.edu/ds4420sp20/slides/18-ds4420-interpretability.pdf

What are the problems with ML models?

- We don’t trust the models
- We don’t know what happens in extreme cases
- Mistakes can be expensive / harmful
- Does the model make similar mistakes as humans ?
- How to change model when things go wrong ?

What do we want to get?

- Interactive feedback - can model learn from human actions in online setting ? (Can you tell a model to not repeat a specific mistake ?)
- Recourse – Can a model tell us what actions we can take to change its output ? (For example, what can you do to improve your credit score?)

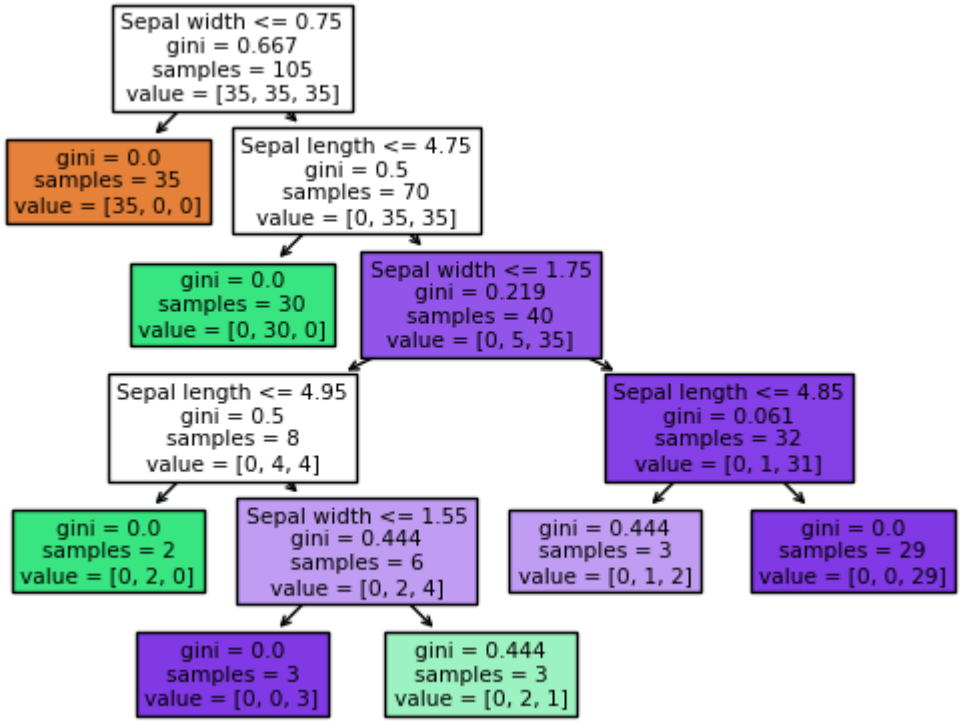
Models: Explainable and Not

Some models are explainable:

1. Linear or physics-defined function
2. Decision trees

But what about:

1. Image segmentation
2. Natural language processing
3. Classification
4. …



What is explainability?

- **Faithfulness:** how to provide explanations that accurately represent the true reasoning behind the model's final decision.
- **Plausibility:** Is the explanation correct or something we can believe is true, given our current knowledge of the problem
- **Understandable:** Can I put it in terms that end user without in-depth knowledge of the system can understand ?
- **Stability:** Does similar instances have similar interpretations ?

What do we expect from explainer?

1. Interpretable: It should provide a qualitative understanding between the input variables and the response. It should be easy to understand.

2. Local Fidelity: It might not be possible for an explanation to be completely faithful unless it is the complete description of the model itself. Having said that it should be at least locally faithful, i.e it must replicate the model's behavior in the vicinity of the instance being predicted.

3. Model Agnostic: The explainer should be able to explain any model and should not make any assumptions about the model while providing explanations.

4. Global perspective:The explainer should explain a representative set to the user so that the user has a global intuition of the model

https://towardsdatascience.com/decrypting-your-machine-learning-model-using-lime-5adc035109b5

Ways to explain ML methods

Global vs local:

- Do we explain individual prediction (Heatmaps, Rationales)?
- Do we explain entire model (Linear Regression, Decision Trees)?

Inherent or post-hoc:

- Is the explainability built into the model (Linear Regression, Decision Trees, Natural Language Explanations)
- Is the model black-box and we use external method to try to understand it (Heatmaps)?

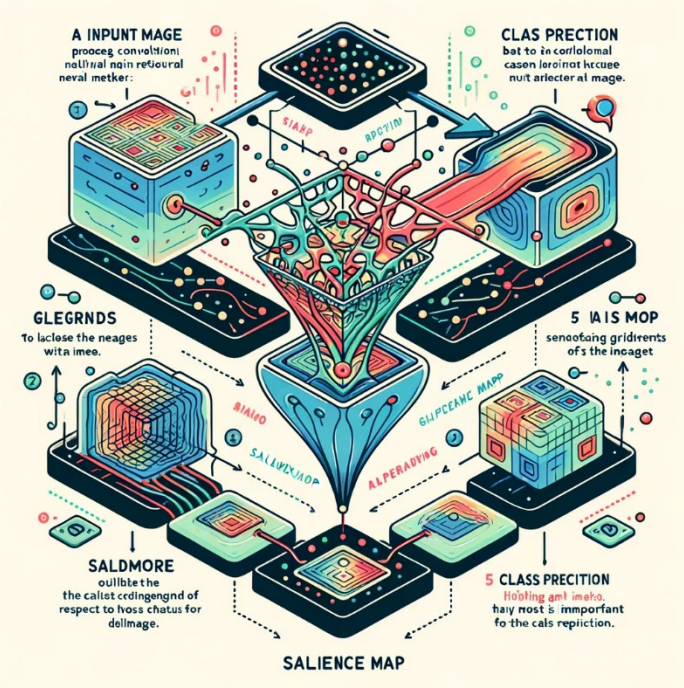
Model based vs Model Agnostic

- Can it explain only few classes of models (attention gradients – differentiable models only)
- Can it explain any model (LIME, SHAP)?

https://course.ccs.neu.edu/ds4420sp20/slides/18-ds4420-interpretability.pdf

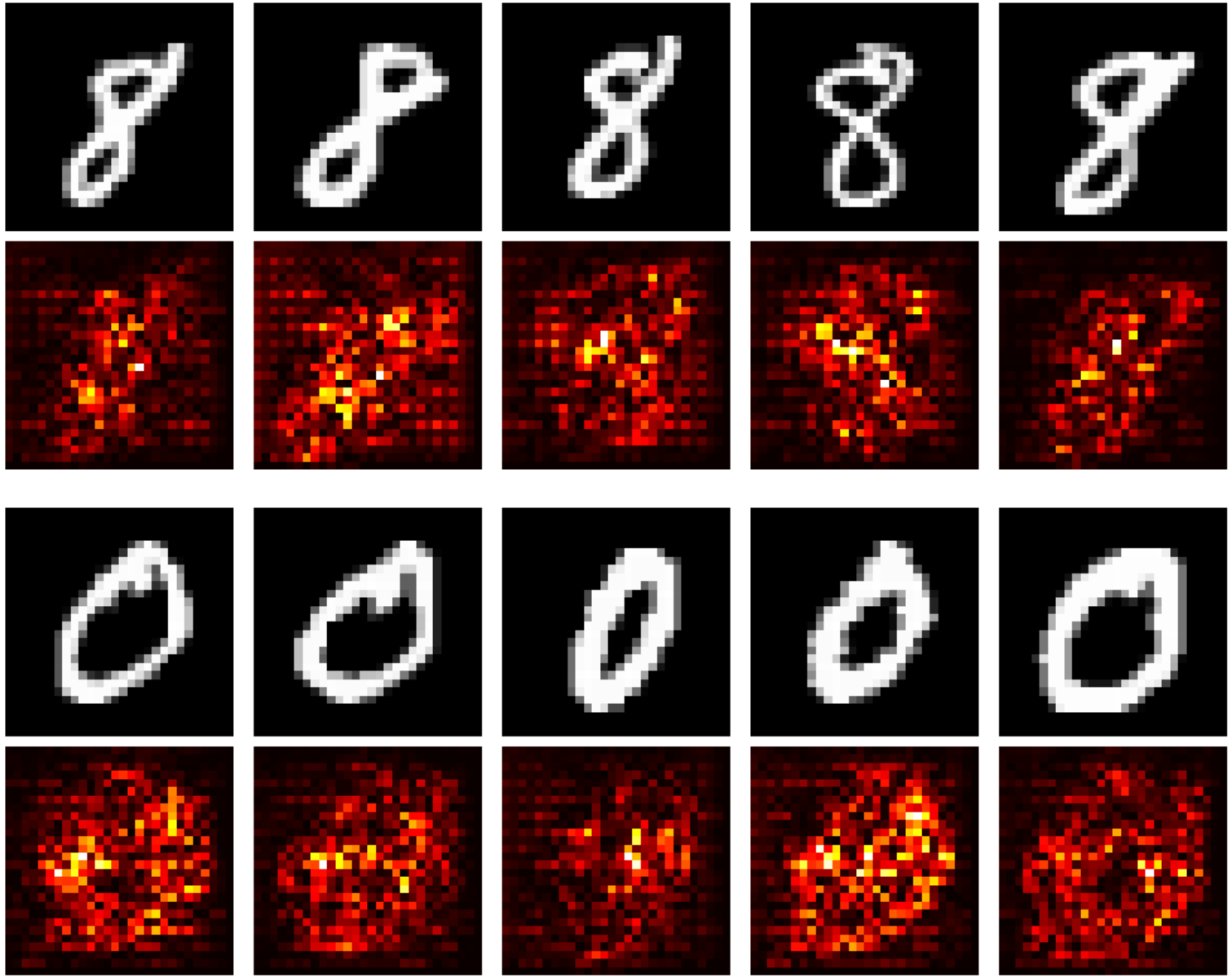
Saliency maps for differentiable models

- Choose the target class for which you want to compute the saliency map. This could be the class predicted by the network or any other class of interest.
- Pass the image through the model to get the output predictions. In the case of classification, this output is typically a probability distribution over classes.
- Extract the model's output (e.g., the probability or the logit) corresponding to the target class.

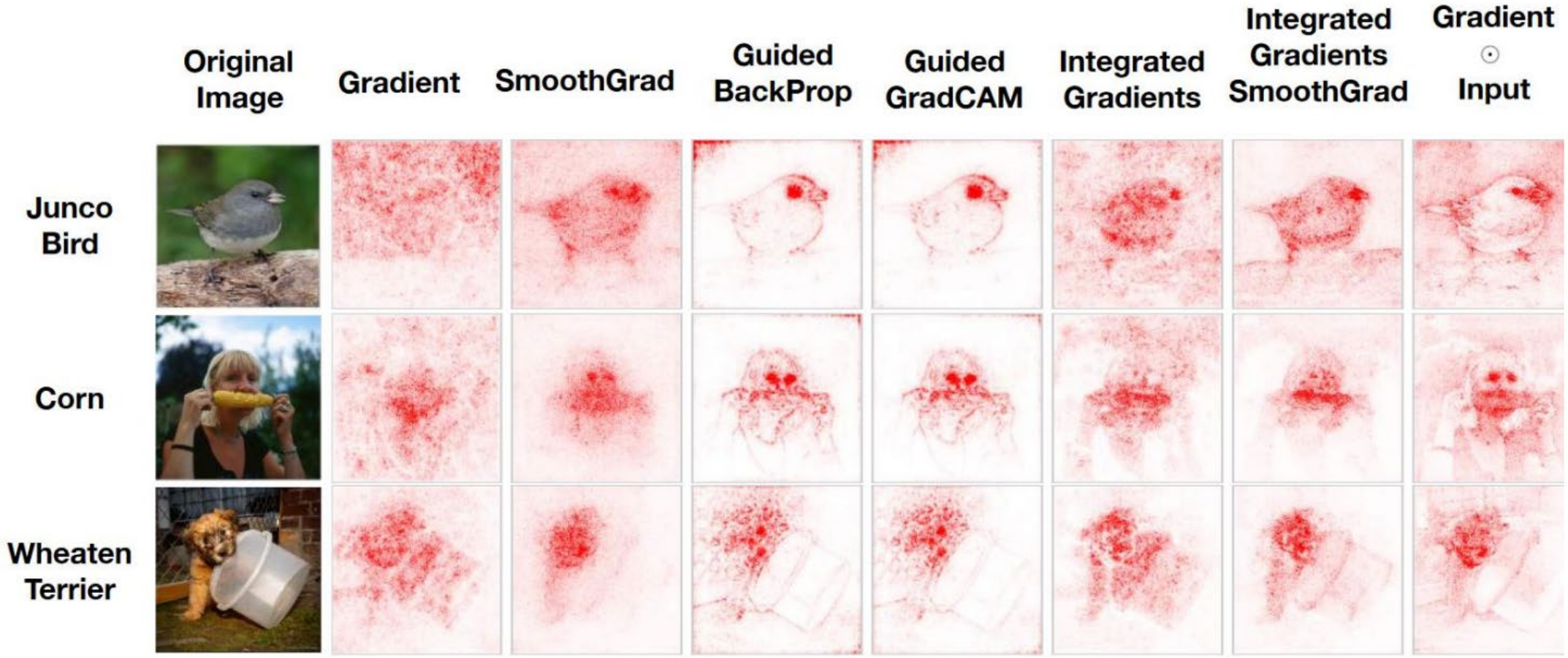


- Calculate the gradient of the output for the target class with respect to the input image. It highlights how much each pixel in the input image contributes to the output value for the chosen class.
- Post-process the gradients to create a saliency map:
  - Taking the absolute value of the gradient.
  - Collapsing the gradient across the color channels, often by taking the maximum or the average across channels.

Example of saliency maps for MNIST



There are many ways to get salience maps



[Adebayo et al 2018]

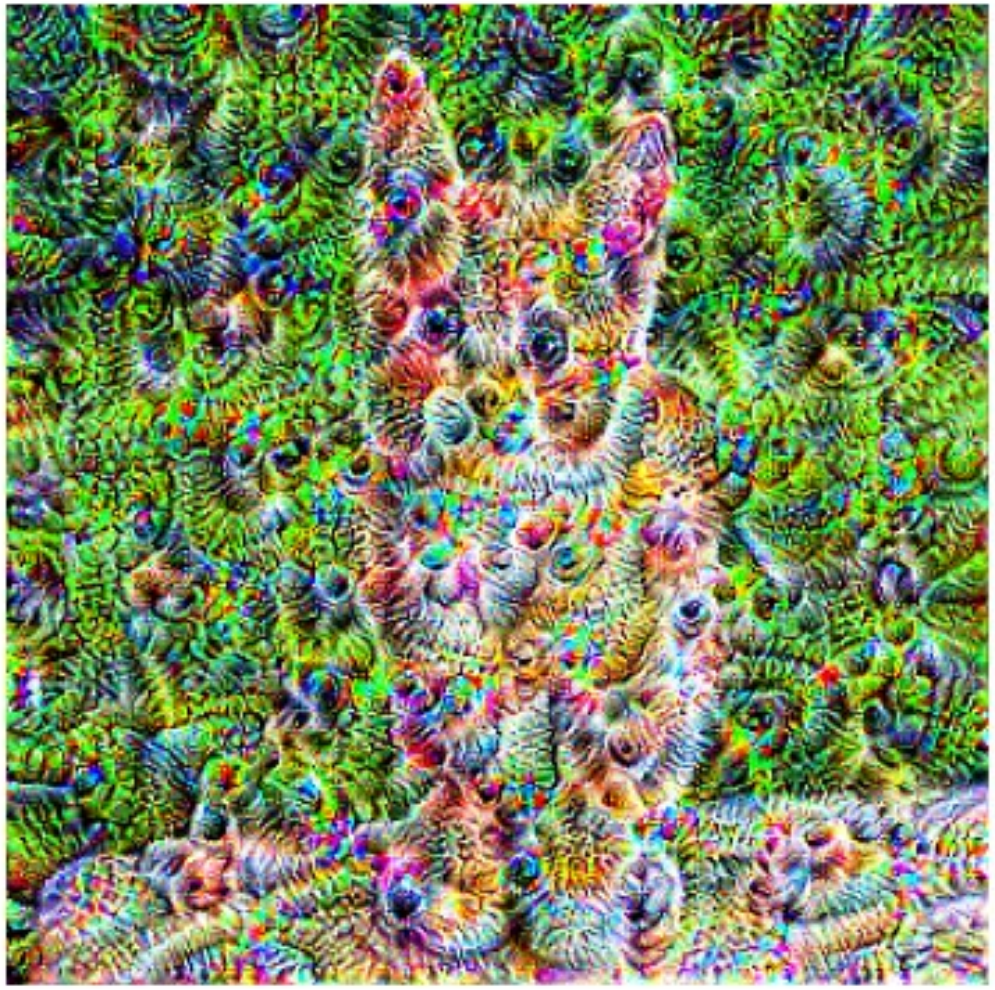
- Only capture first order information
- Not very reliable

https://course.ccs.neu.edu/ds4420sp20/slides/18-ds4420-interpretability.pdf

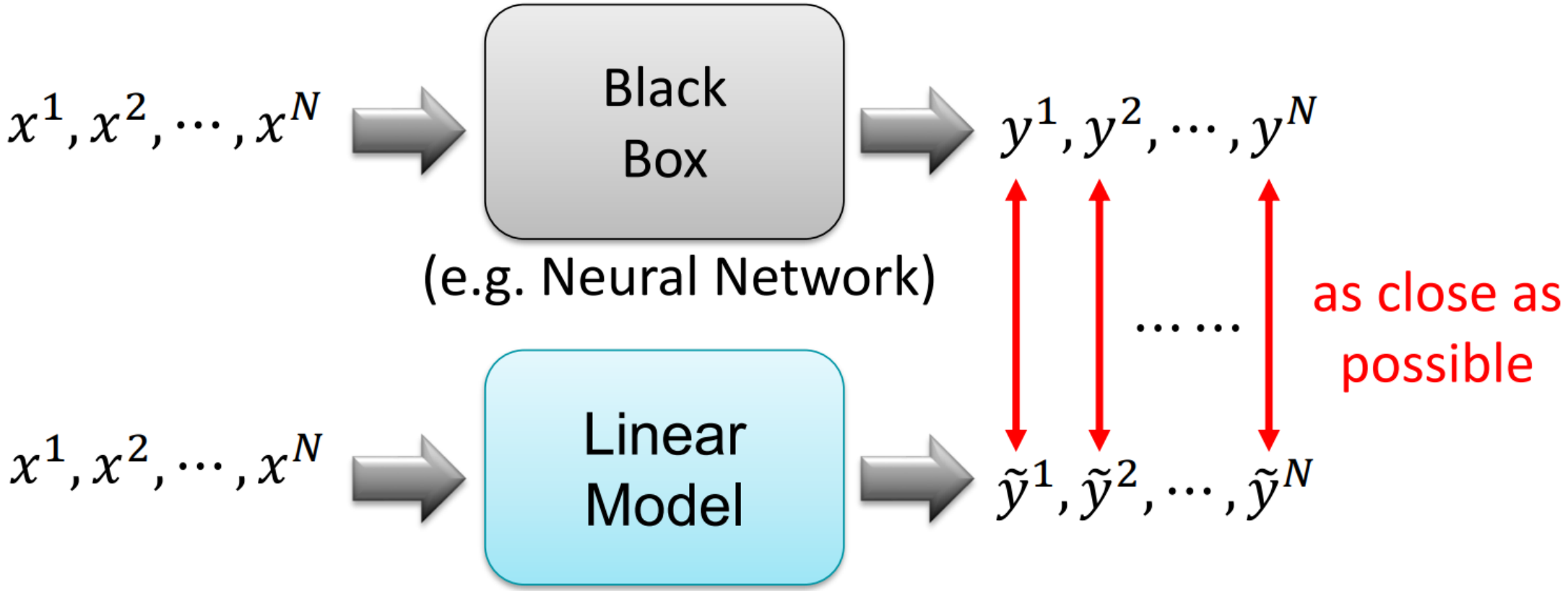
Closely related – Deep Dream

Original Image

Deep Dream Image



Locally Interpretable Manifold Embedding



https://course.ccs.neu.edu/ds4420sp20/slides/18-ds4420-interpretability.pdf

Locally Interpretable Manifold Embedding

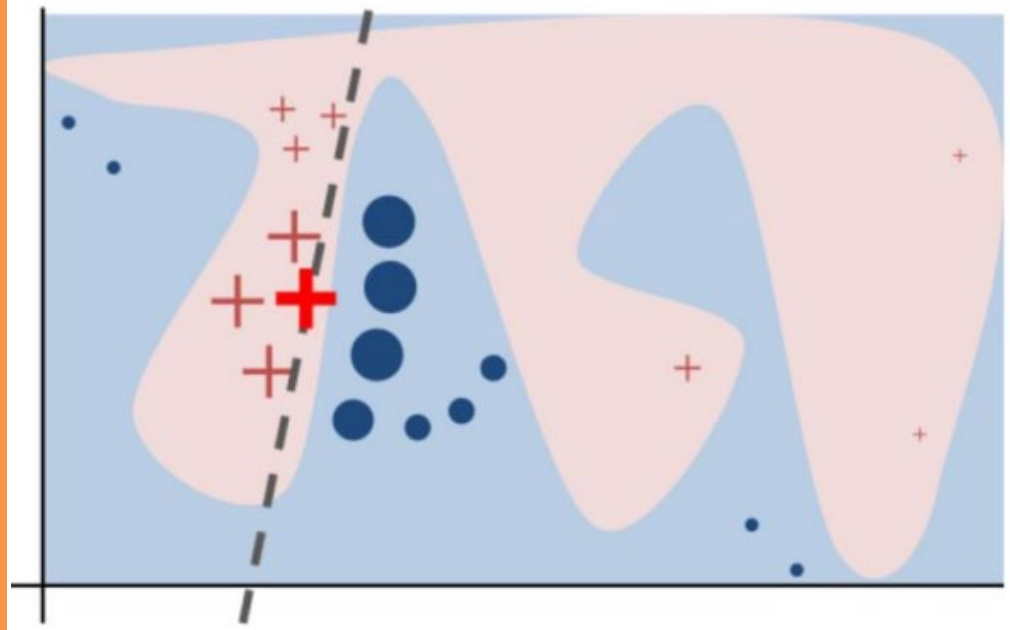


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

Ribeiro et al, 2016

Locally Interpretable Manifold Embedding

Fidelity-Interpretability trade-off: We want an explainer that is faithful (replicate our model's behavior locally) and interpretable. Towards this goal, LIME minimizes the following

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g) \qquad (1)$$

- f: an original predictor
- x: original features
- g: explanation model which could be a linear model, decision tree, or falling rule lists
- Pi: proximity measure between an instance of z to x to define locality around x. It weighs z' (perturbed instances) depending upon their distance from x.

First Term: the measure of the unfaithfulness of g in approximating f in the locality defined by Pi. This is termed as locality-aware loss in the original paper

Last term: a measure of model complexity of explanation g. For example, if your explanation model is a decision tree it can be the depth of the tree or in the case of linear explanation models it can be the number of non-zero weights

https://towardsdatascience.com/decrypting-your-machine-learning-model-using-lime-5adc035109b5

LIME: Sparse Linear Explanation:

1. g(z') = w . z' (Making the explanation model linear)
2. Locally-aware loss = square loss
3. Pi(z) : exp(-D(x,z)(2)/sigma(2)) (proximity weighing for the samples)
4. D(x,z) : Distance function

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2 \qquad (2)$$

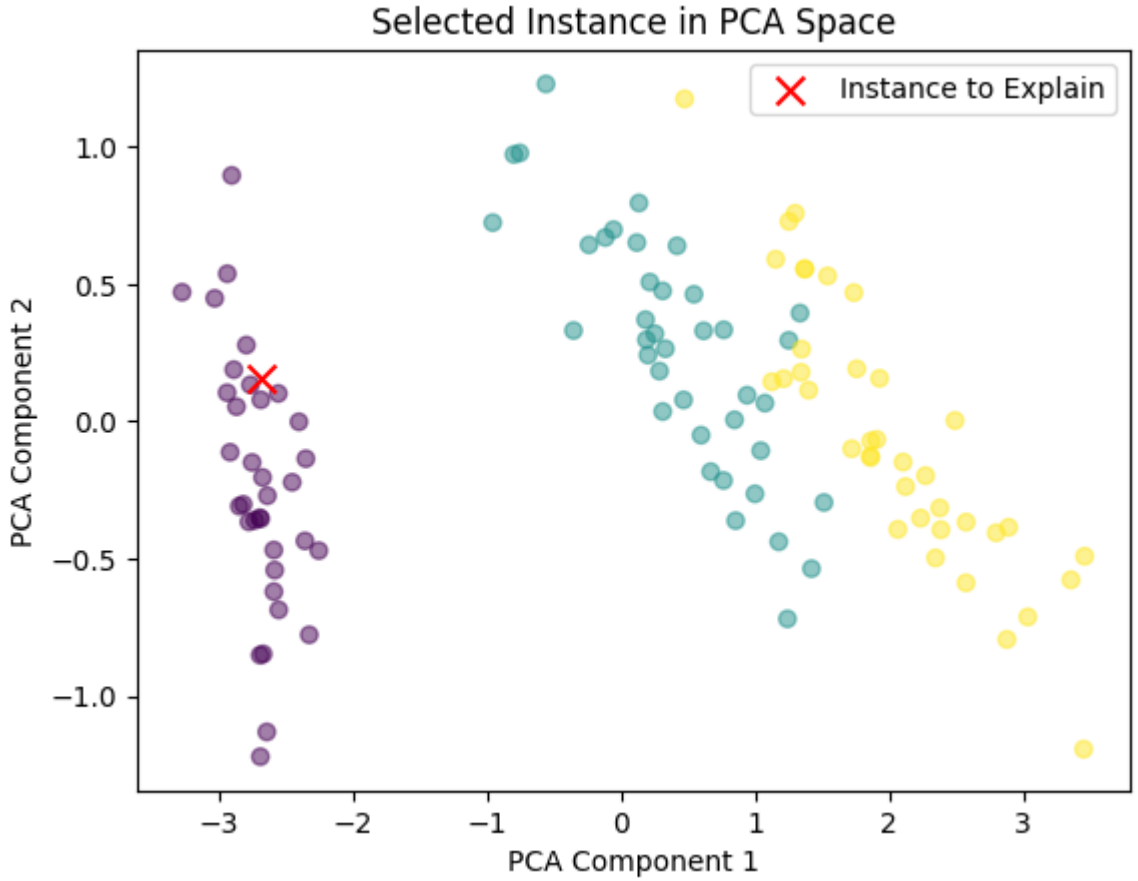
Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

```
Z ← {}
for i ∈ {1, 2, 3, ..., N} do
    z'_i ← sample_around(x')
    Z ← Z ∪ ⟨z'_i, f(z_i), π_x(z_i)⟩
end for
w ← K-Lasso(Z, K)  ▷ with z'_i as features, f(z) as target
return w
```

https://towardsdatascience.com/decrypting-your-machine-learning-model-using-lime-5adc035109b5

LIME on Iris



Prediction probabilities

setosa	1.00
versicolor	0.00
virginica	0.00

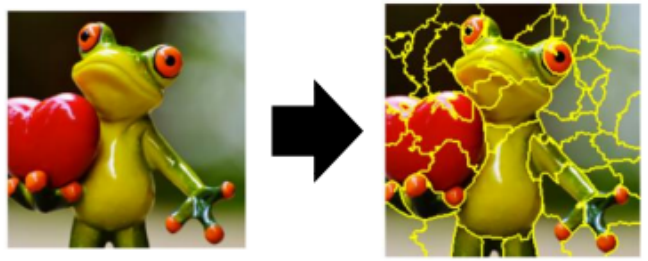
NOT setosa setosa

petal width (cm) <= 0.30
0.44
petal length (cm) <= 1.60
0.43
sepal length (cm) <=...
0.04
3.00 < sepal width (cm...
0.00

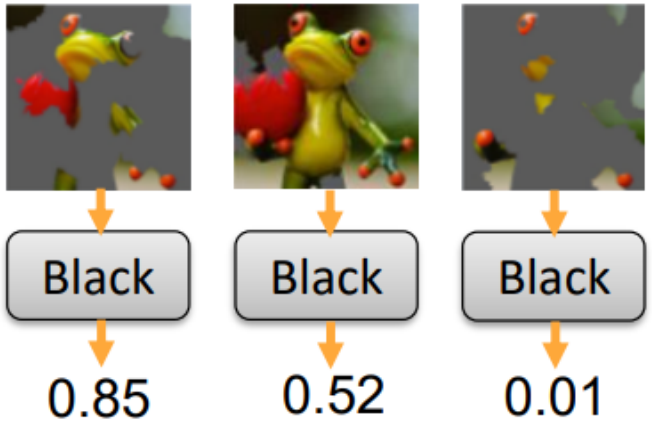
Feature	Value
petal width (cm)	0.20
petal length (cm)	1.60
sepal length (cm)	4.70
sepal width (cm)	3.20

LIME on Images

LIME – Image



- 1. Given a data point you want to explain
- 2. Sample at the nearby - Each image is represented as a set of superpixels (segments).



Randomly delete some segments.

Compute the probability of "frog" by black box

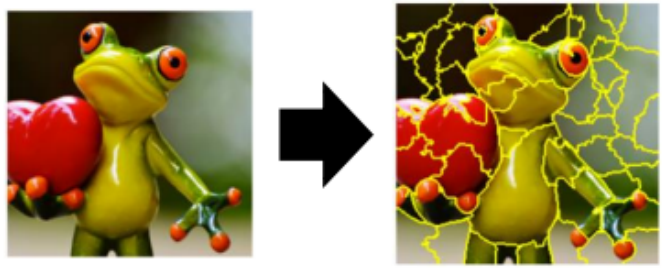
Ref: https://medium.com/@kstseng/lime-local-interpretable-model-agnostic-explanation%E6%8A%80%E8%A1%93%E4%BB%8B%E7%B4%B9-a67b6c34c3f8

(Slide Credit – Hung-yi Lee)

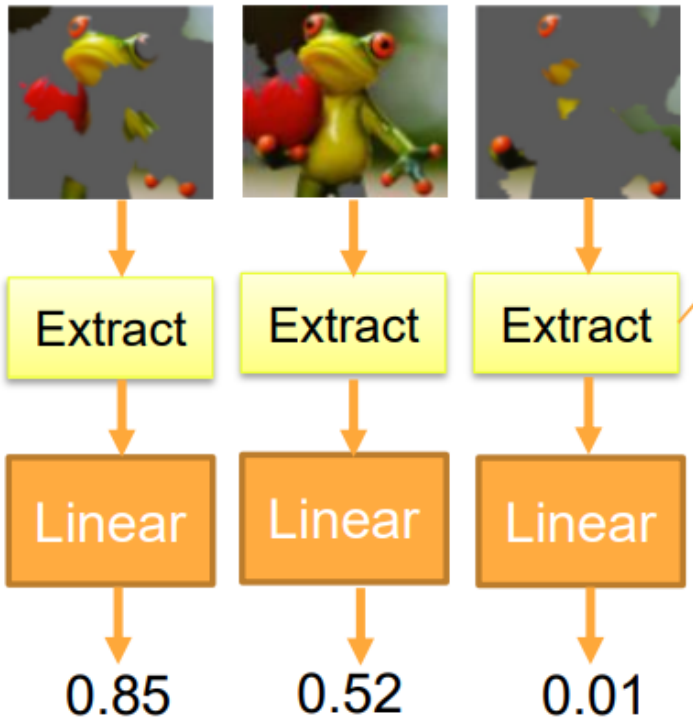
https://course.ccs.neu.edu/ds4420sp20/slides/18-ds4420-interpretability.pdf

LIME on Images

LIME – Image



- 3. Fit with linear (or interpretable) model



$x_1 \quad \cdots\cdots \quad x_m \quad \cdots\cdots \quad x_M$

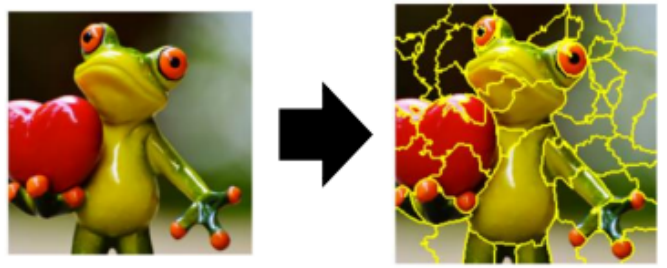
$$x_m = \begin{cases} 0 & \text{Segment m is deleted.} \\ 1 & \text{Segment m exists.} \end{cases}$$

M is the number of segments.

(Slide Credit – Hung-yi Lee)

LIME on Images

LIME – Image



- 4. Interpret the model you learned

$$y = w_1 x_1 + \cdots + w_m x_m + \cdots + w_M x_M$$

$$x_m = \begin{cases} 0 & \text{Segment m is deleted.} \\ 1 & \text{Segment m exists.} \end{cases}$$

M is the number of segments.

If $w_m \approx 0$ ➡ segment m is not related to “frog”

If w_m is positive ➡ segment m indicates the image is “frog”

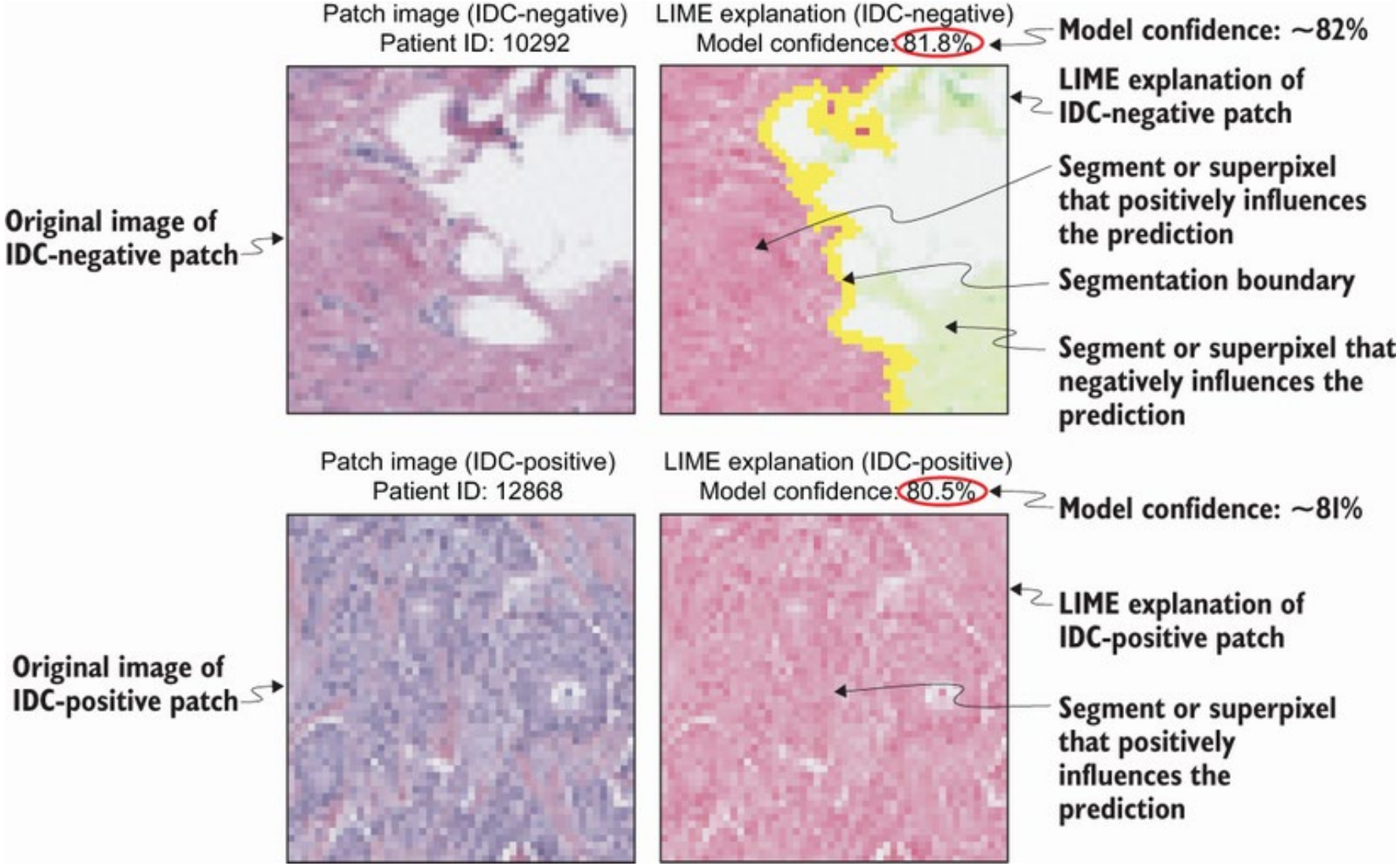
If w_m is negative ➡ segment m indicates the image is not “frog”

(Slide Credit – Hung-yi Lee)

DYI LIME

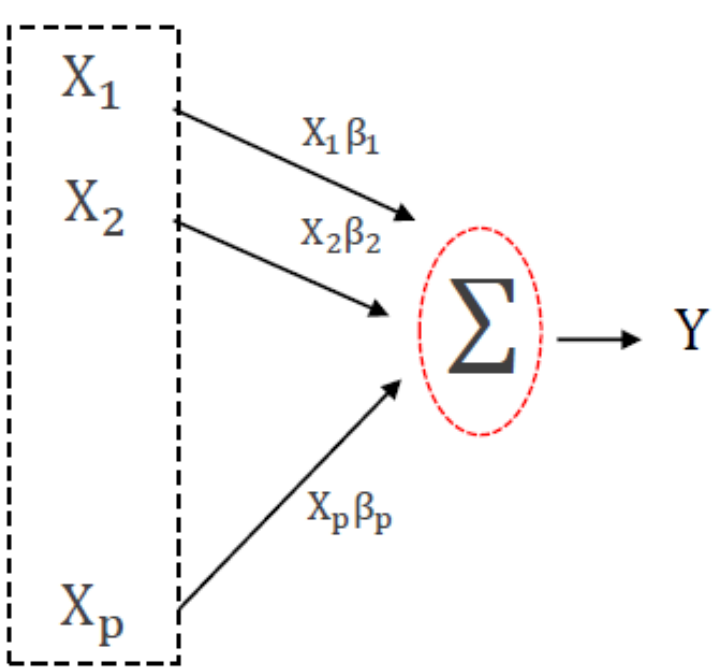

Num features: 2
Num features: 3
Num features: 4
Num features: 5
Num features: 6
Num features: 7
Num features: 8
Num features: 9
Num features: 10
Num features: 11

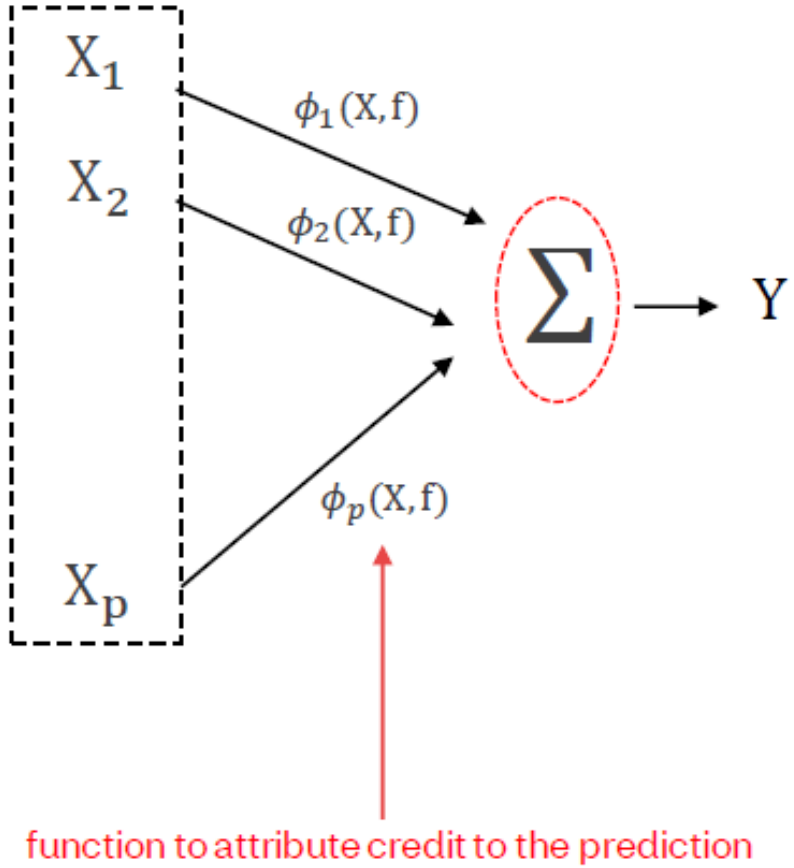
LIME on Images



https://livebook.manning.com/book/interpretable-ai/chapter-5/94

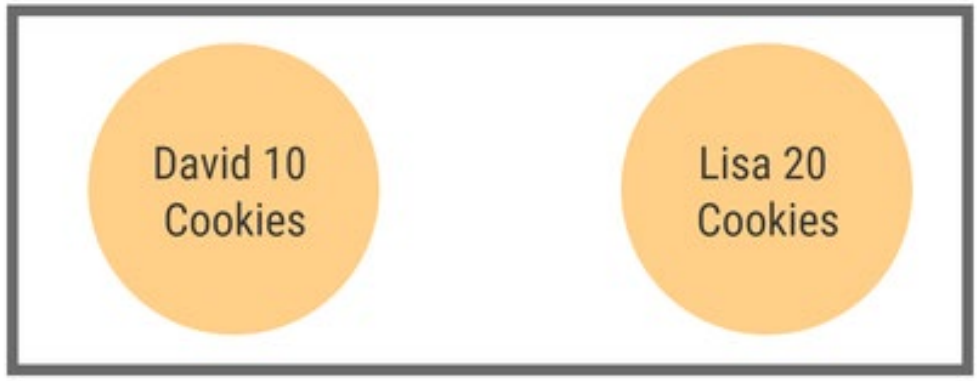
SHAP: Shapley Additive exPlanations

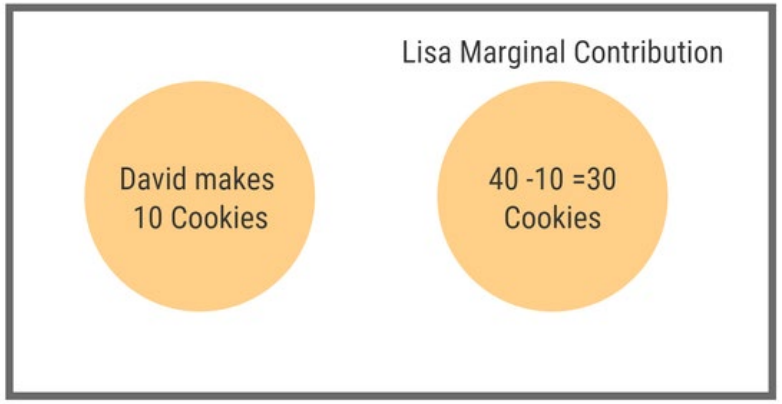


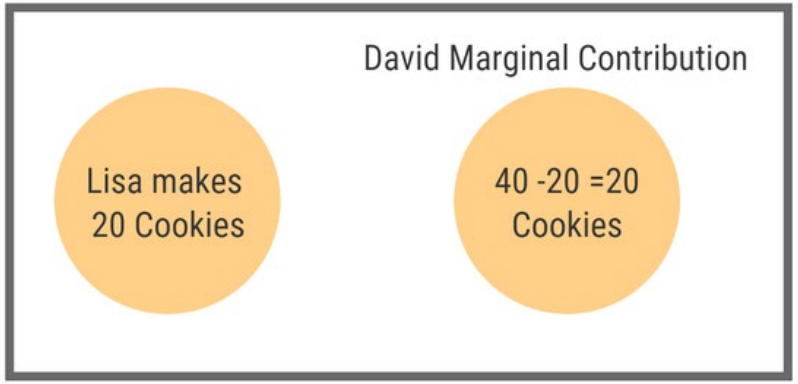


https://towardsdatascience.com/interpretable-machine-learning-using-shap-theory-and-applications-26c12f7a7f1a

SHAP: Shapley Additive exPlanations







- For Lisa, the contribution to the coalition is 30 cookies in the first case and her contribution to the coalition in the second case is 20 cookies. The Shapley value will be (20+30)/2 = 25
- To find the Shapley value of David, we need to average them: (10+20)/2 = 15.

https://abhishek-maheshwarappa.medium.com/shap-values-for-explainable-ai-58652645d881

SHAP: Shapley Additive exPlanations

A method of dividing up the gains or costs among players according to the value of their individual contributions.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S))$$

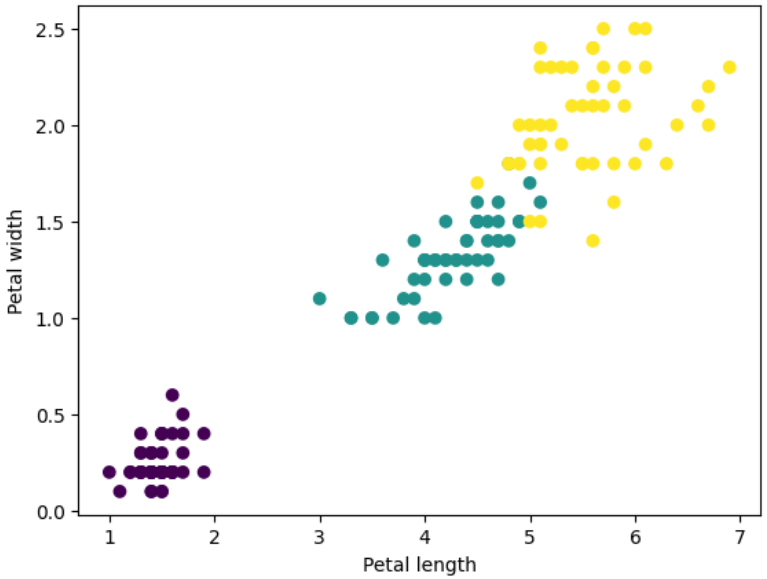
1. Marginal contribution
The contribution of each player is determined by what is gained or lost by removing them from the game. This is called their marginal contributions.

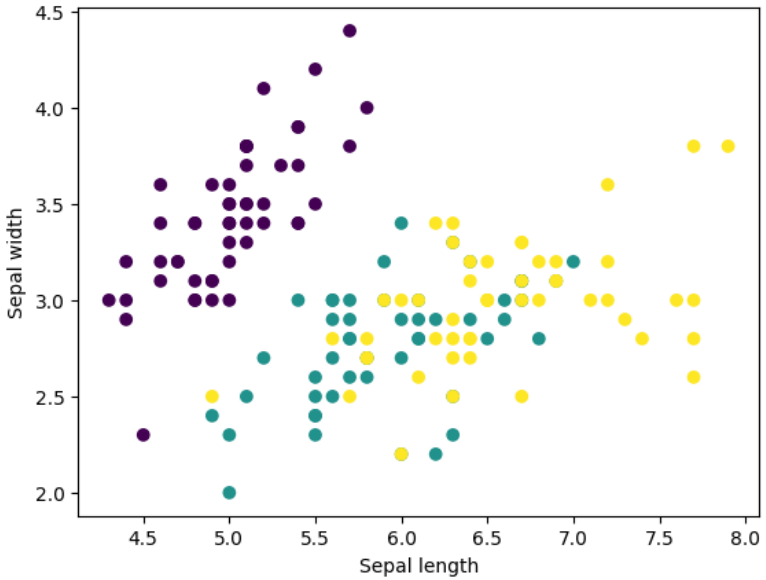
2. Interchangeable players have an equal value
If two parties bring the same things to the coalition, they should have to contribute the same amount and should be rewarded for their contributions.

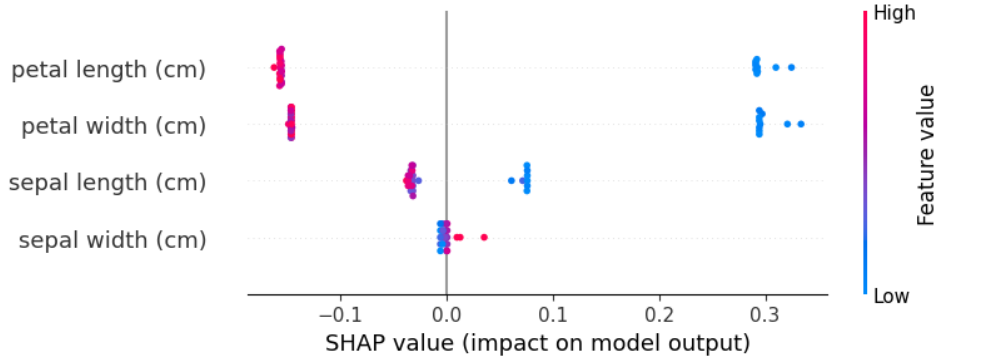
3. Dummy player has zero values
If a member of the coalition contributes nothing, then they should receive nothing. But it might not be fair in all cases, let us take an example of this thing more clear:

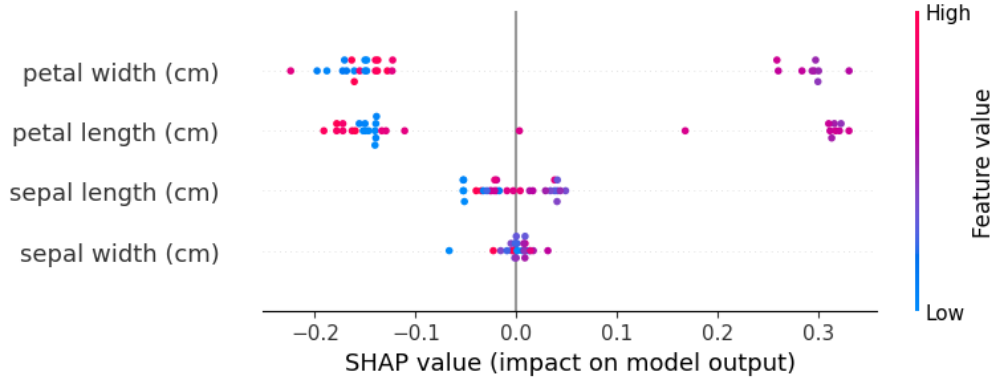
https://abhishek-maheshwarappa.medium.com/shap-values-for-explainable-ai-58652645d881

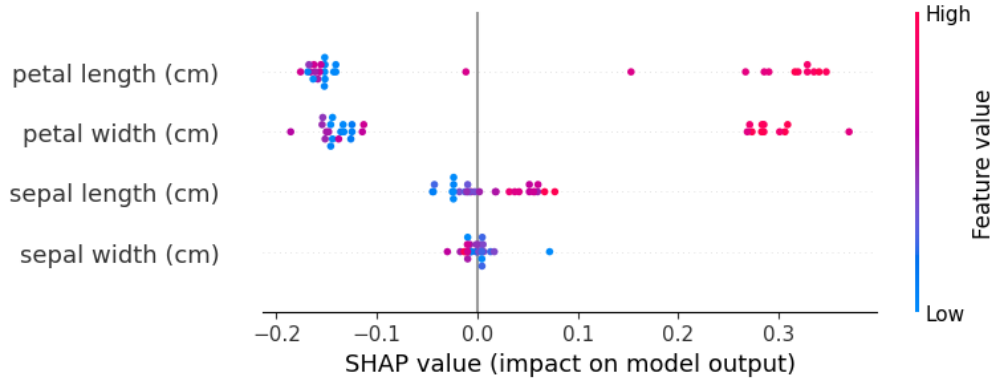
SHAP: Shapley Additive exPlanations



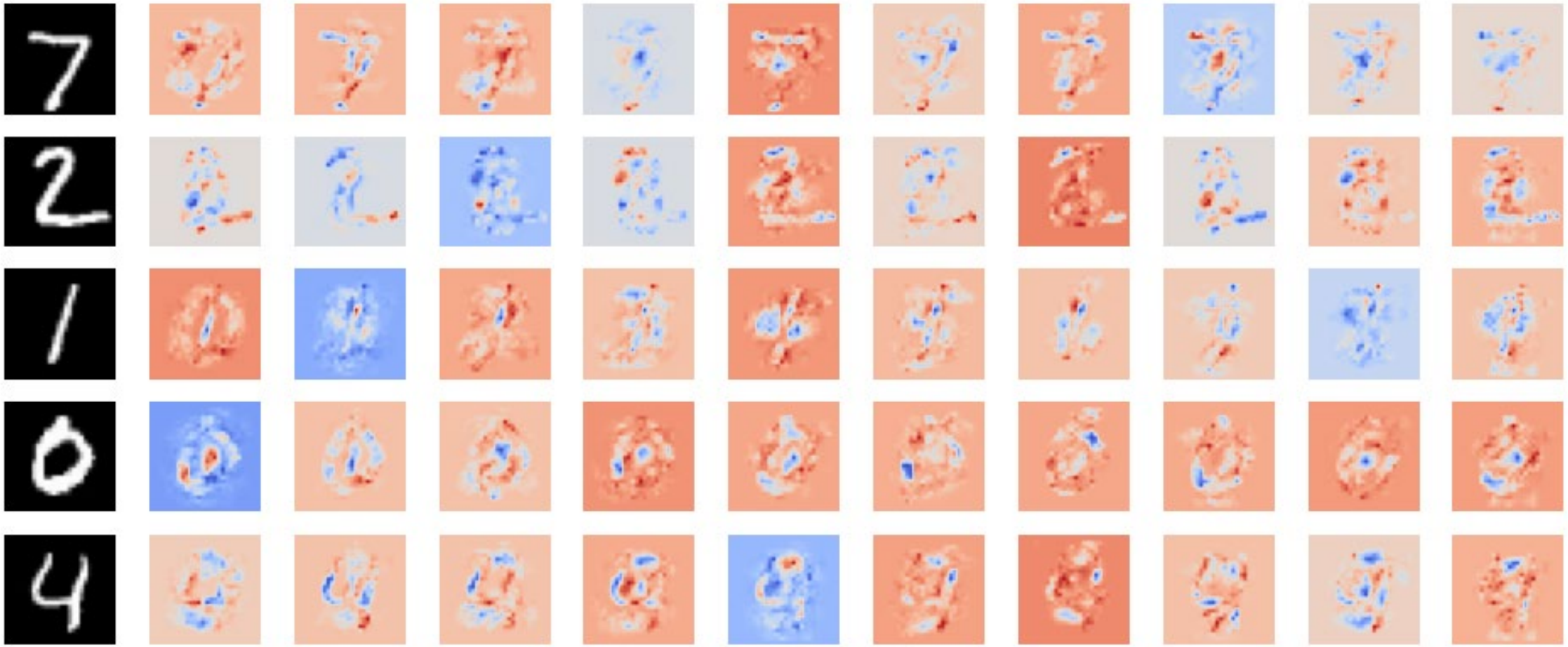


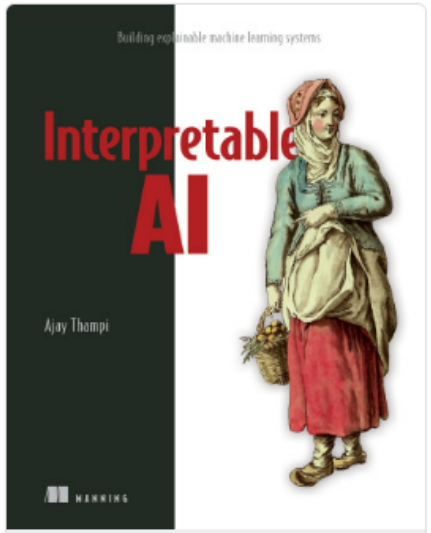






SHAP for Images





look inside

Transparent and understandable AI systems

read this article now in Manning's Free Content Center

resources

Source code

Book Forum

more

Interpretable AI you own this product

Building explainable machine learning systems

6 reviews

Ajay Thampi

May 2022 · ISBN 9781617297649 · 328 pages · printed in black & white

Data

eBook	print	subscription
pdf, ePub, online	includes eBook	from $19.99

AI doesn't have to be a black box. These practical techniques help shine a light on your model's mysterious inner workings. Make your AI more transparent, and you'll improve trust in your results, combat data leakage and bias, and ensure compliance with legal requirements.

In *Interpretable AI*, you will learn:

- Why AI models are hard to interpret
- Interpreting white box models such as linear regression, decision trees, and generalized additive models
- Partial dependence plots, LIME, SHAP and Anchors, and other techniques such as saliency mapping, network dissection, and representational learning
- What fairness is and how to mitigate bias in AI systems
- Implement robust AI systems that are GDPR-compliant

eBook

~~$47.99~~ **$31.19**

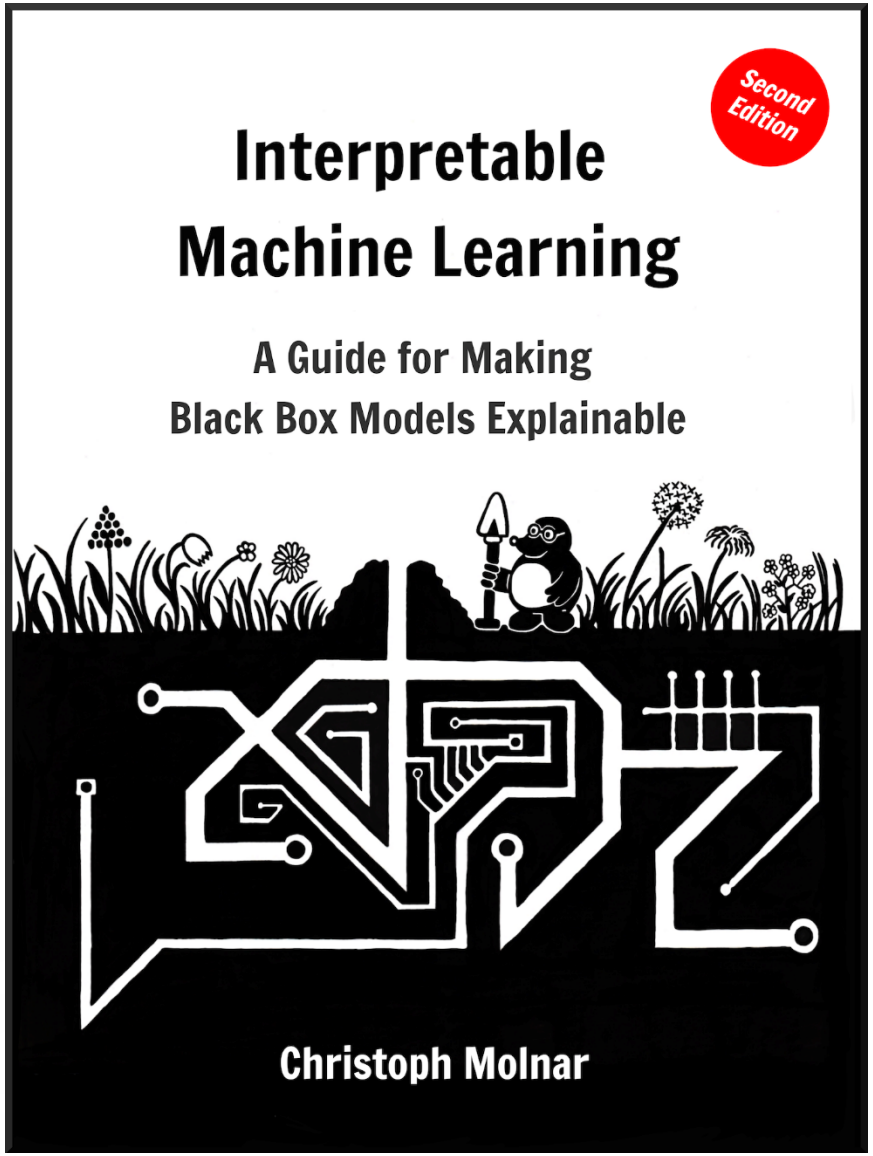
you save $16.80 (35%)

add to cart

buy now

free with subscription

> **A sound introduction for practitioners to the exciting field of interpretable AI.**
>
> Pablo Roccatagliata, Torcuato Di Tella University



https://christophm.github.io/interpretable-ml-book/

Machine learning has great potential for improving products, processes and research. But **computers usually do not explain their predictions** which is a barrier to the adoption of machine learning. This book is about making machine learning models and their decisions interpretable.

After exploring the concepts of interpretability, you will learn about simple, **interpretable models** such as decision trees, decision rules and linear regression. The focus of the book is on model-agnostic methods for **interpreting black box models** such as feature importance and accumulated local effects, and explaining individual predictions with Shapley values and LIME. In addition, the book presents methods specific to deep neural networks.

All interpretation methods are explained in depth and discussed critically. How do they work under the hood? What are their strengths and weaknesses? How can their outputs be interpreted? This book will enable you to select and correctly apply the interpretation method that is most suitable for your machine learning project. Reading the book is recommended for machine learning practitioners, data scientists, statisticians, and anyone else interested in making machine learning models interpretable.