Explainable AI : Shapley Values and other stuff

A Unified Approach to Interpreting Model Predictions

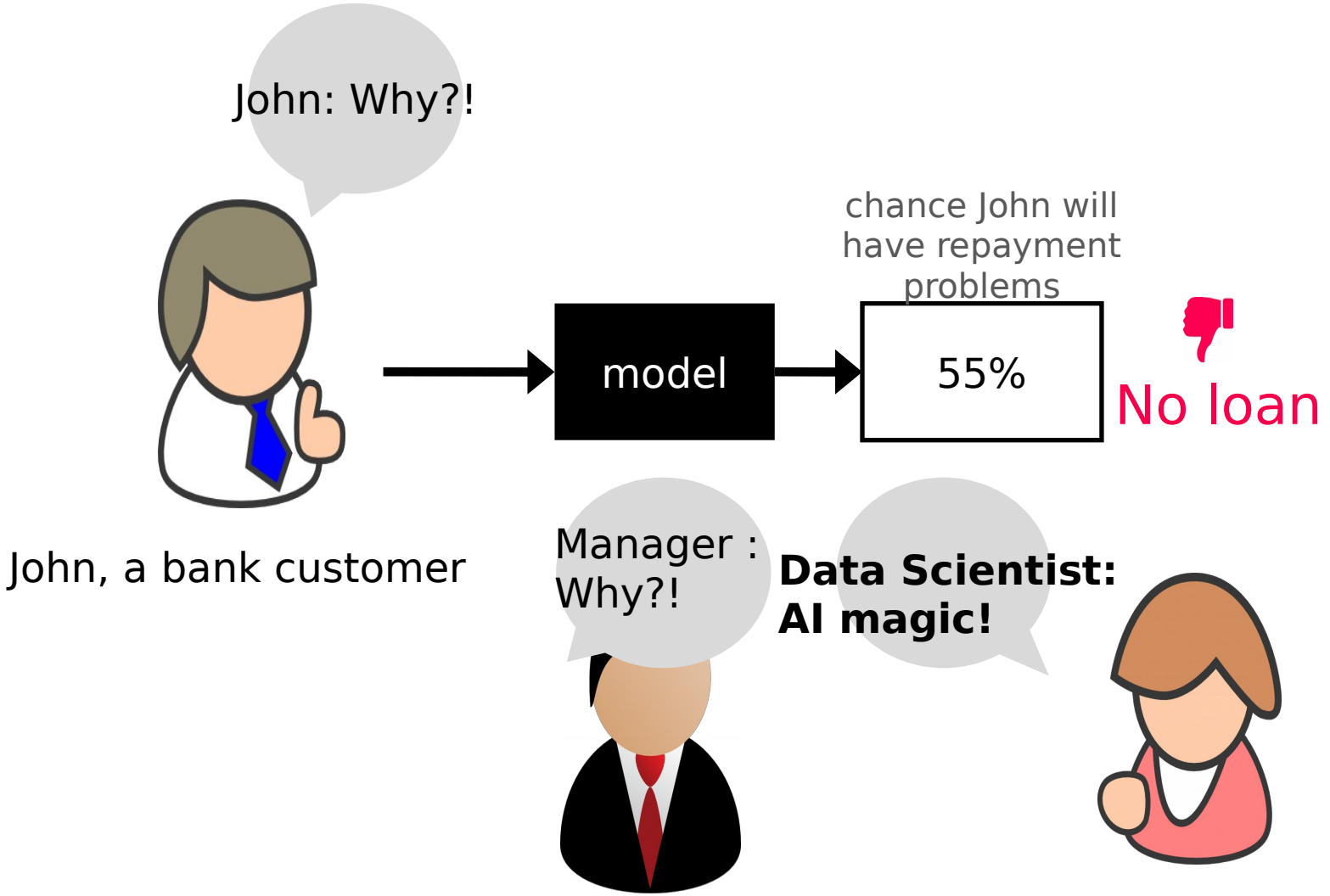
Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)

https://colab.research.google.com/github/**leexa90**/Explainable_AI_image_classification/blob/master/colabs_script.ipynb

Background on myself

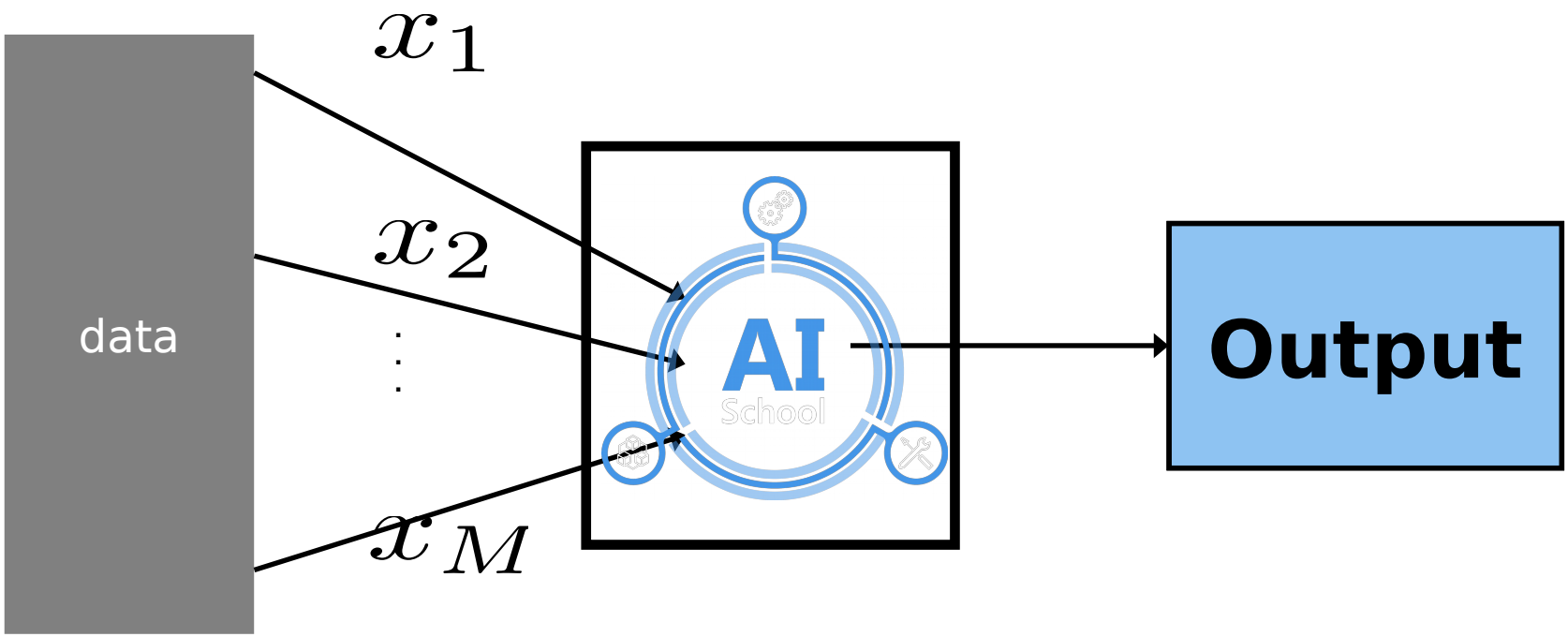
- Graduated from NUS science 2015
- Attended Deep Learning Developer Course 2017
- Working in A*STAR Bioinformatics Institute, using various methods and analysis on crop analytics in smart urban farms

Need for Explainable AI



https://github.com/slundberg/shap/blob/master/docs/presentations/NIPS%202017%20Talk.pptx

Complicated AI Model

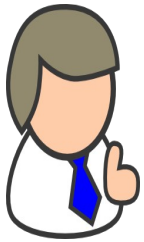


Explainable model: Additive feature attribution model

$$\textbf{Output} = \sum_{i=1}^{M} \phi_i$$

M is the number of simplified input features, and $\phi_i \in \mathbb{R}$.

SHapley Additive exPlanation - (SHAP) values (1)



Base rate

20%

$E[f(x)]$

Prediction for John

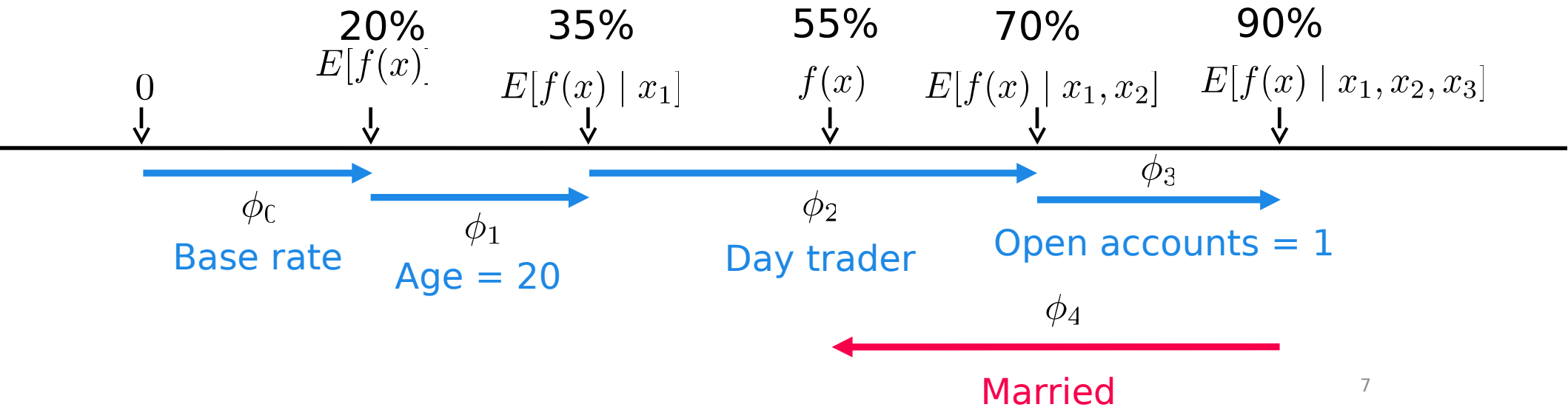
55%

$f(x)$

0

How did we get here?

SHapley Additive exPlanation (SHAP) values (2)



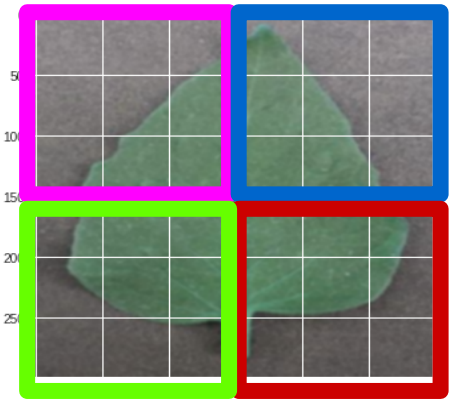
SHapley Additive exPlanation (SHAP) values (3) – Computation

- Train AI model

For each picture containing 4 superpixels

{ *Explain model :*

$$output = \varphi_{pink} + \varphi_{blue} + \varphi_{green} + \varphi_{red}$$

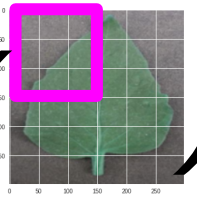
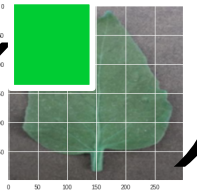


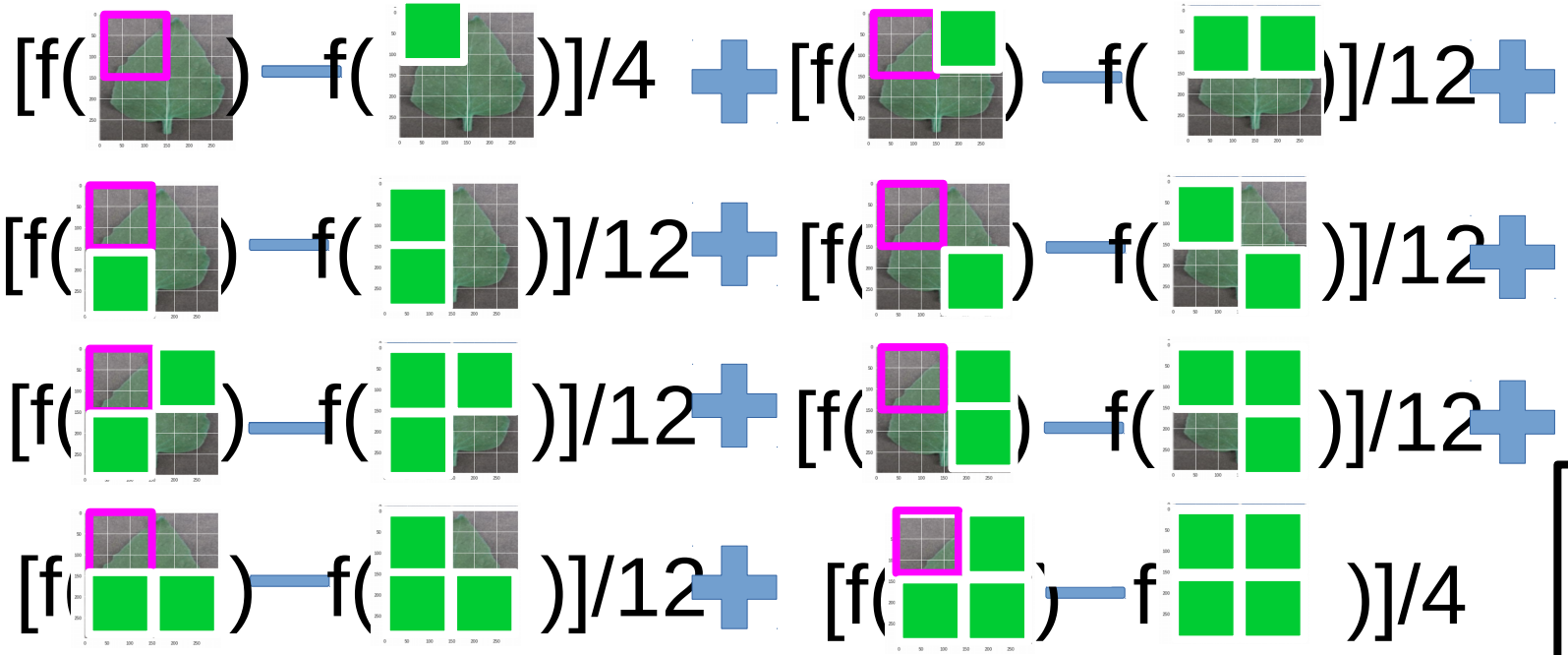
φ is the shapley value of the feature

}

$$\varphi_{pink} = weight\_avg(f(\quad) - f(\quad))$$

SHapley Additive exPlanation (SHAP) values (4) – computation

$$\varphi_{pink} = weight\_avg(f(\text{  }) - f(\text{  }))$$



Legend
Green solid squared are mean-filled super-pixels

SHapley Additive exPlanation (SHAP) values (5) – solved using weighted linear regression

$$\phi = (X^T W X)^{-1} X^T W y$$

Refer to proof in paper for details, A Unified Approach to Interpreting Model Predictions

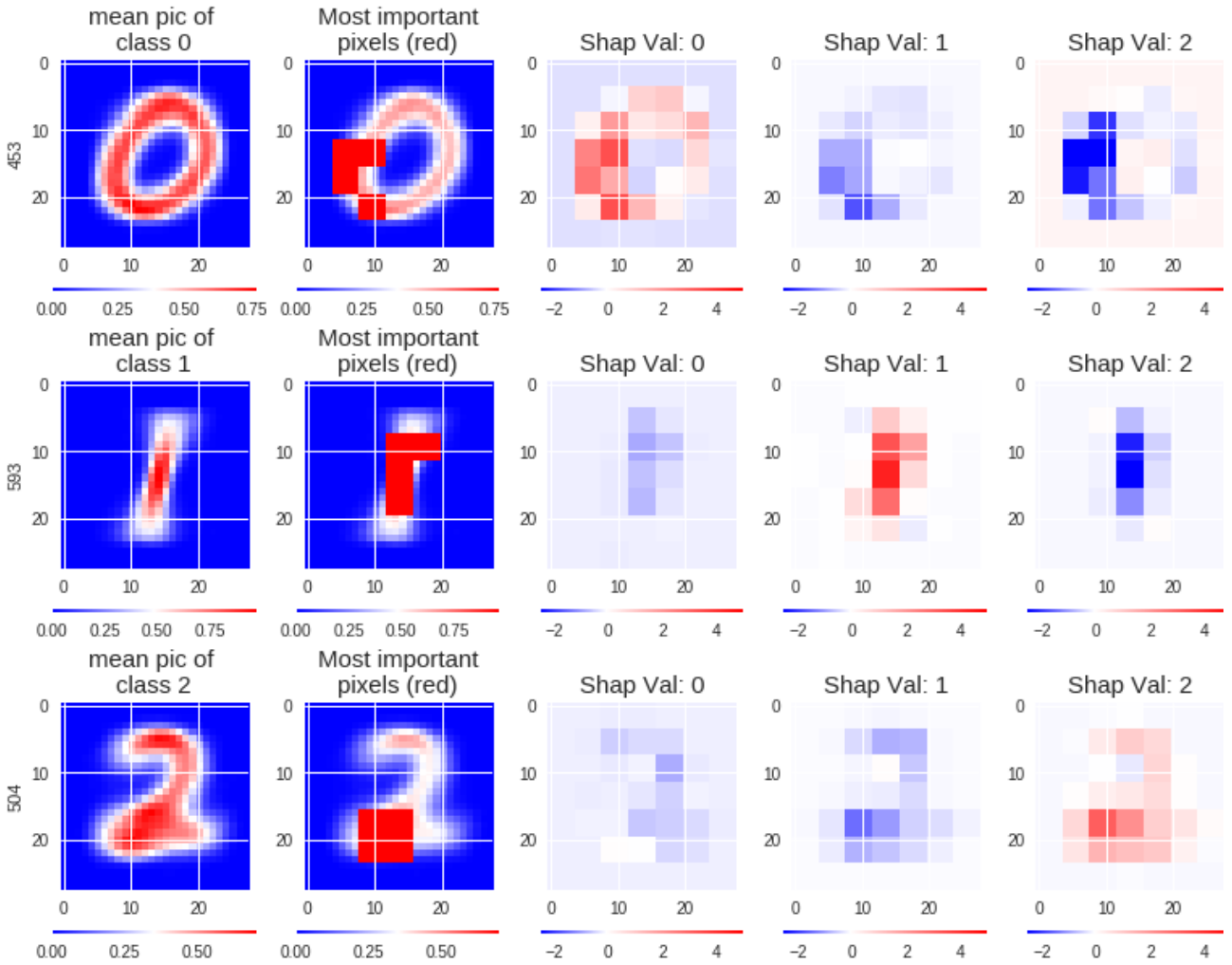
Applying to Mnist (1)

- Mnist model with 4 convolutional layers and 2 dense layers, with test accuracy 99.4%
- for each test image {
  - Split image to 7*7 = 49 superpixels for shapley value computation
  - Sample 7367 combinations of pixels
    ~ all images with 1 mean-filled pixel, ${}^{49}C_1 = 49$
    ~ all images with 2-filled pixels, ${}^{49}C_2 = 1176$
    ~ 33% of images with 3 mean-filled pixel, ${}^{49}C_3/3 = 6142$
  - Calculate shapley values for each pixel using weighted regression

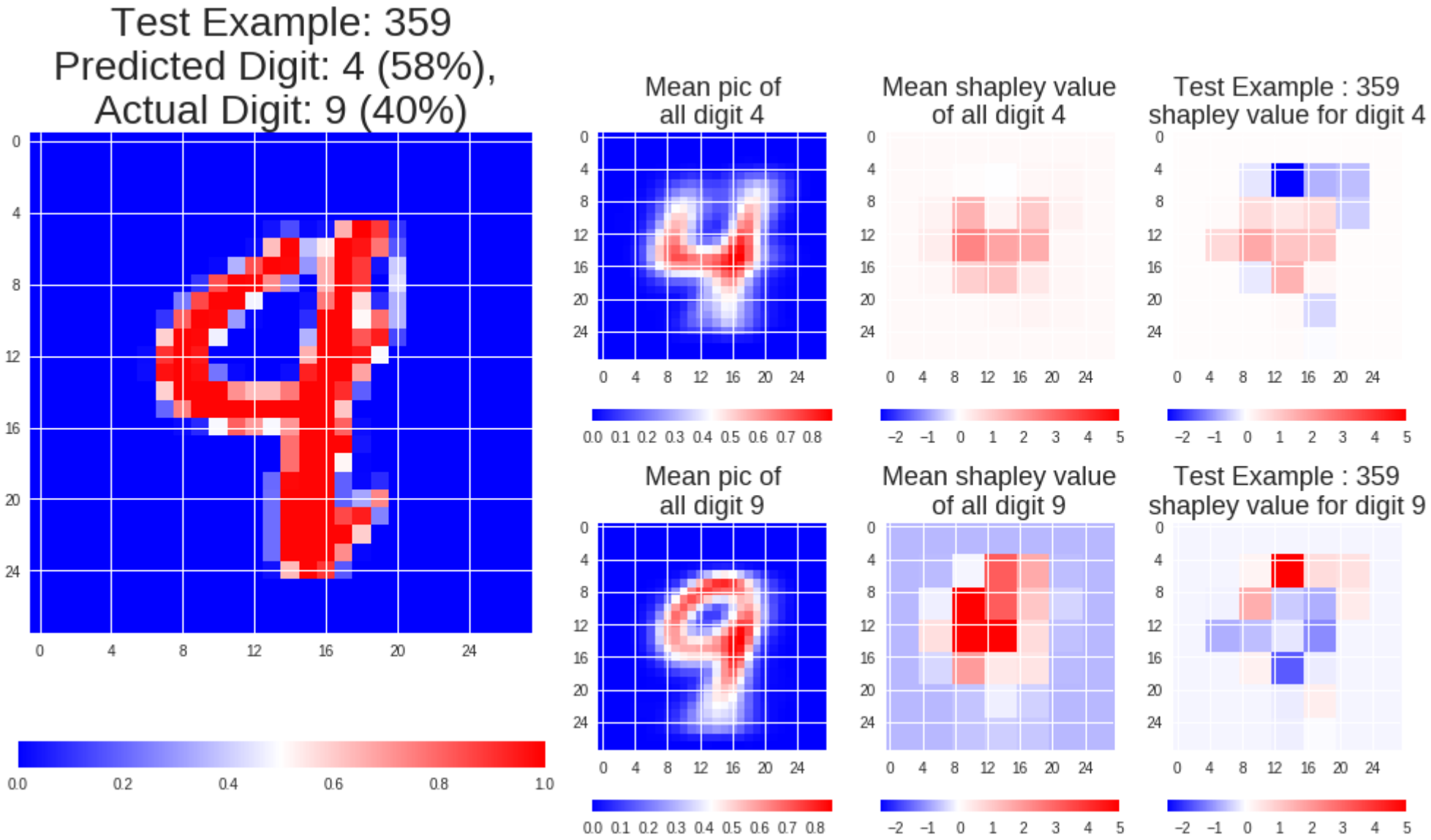
    }

Output = $\sum_{i=1}^{M} \phi_i$

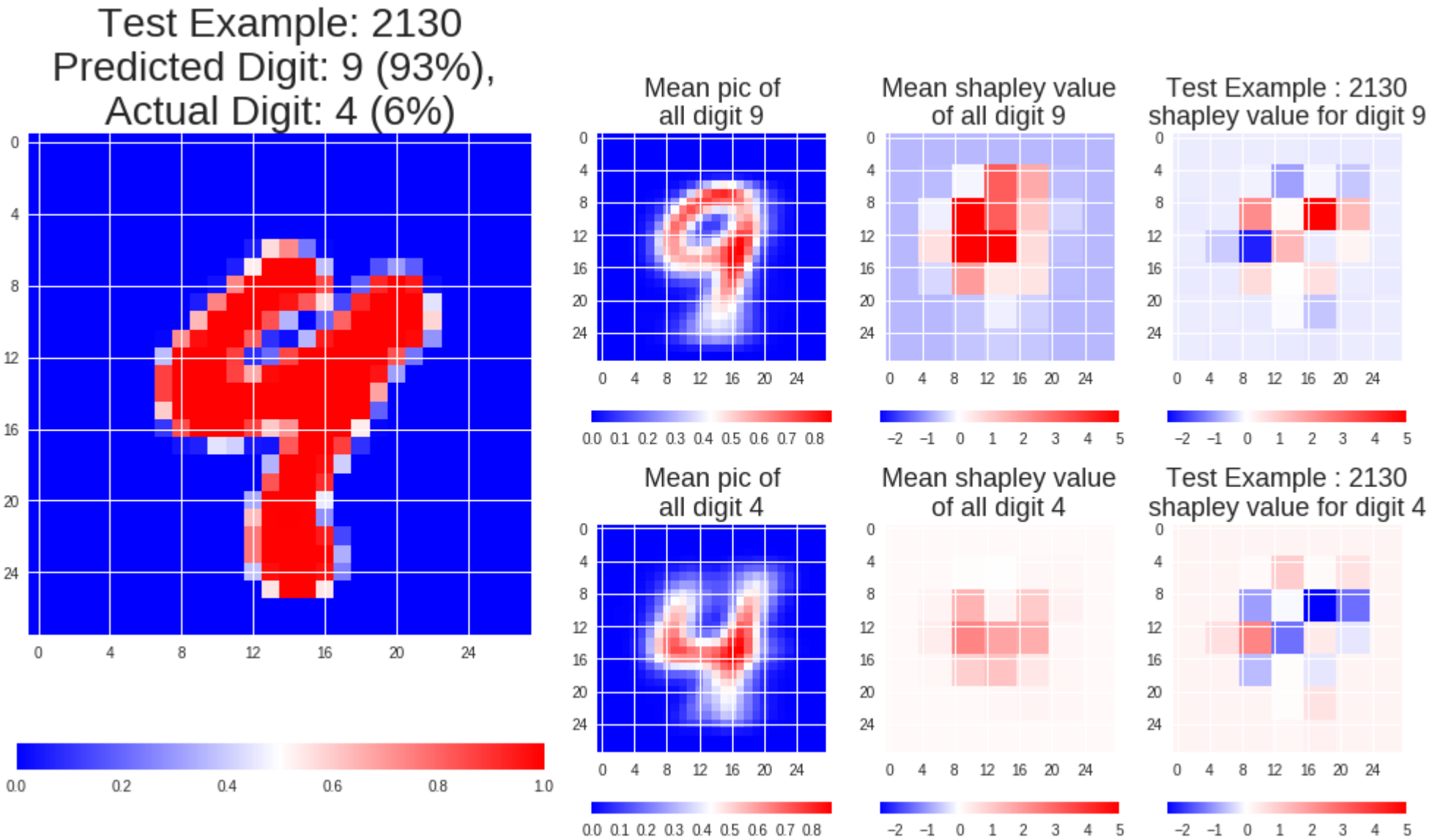
Applying to Mnist (2) – Global analysis



Applying to Mnist (3) – Individual analysis (a)

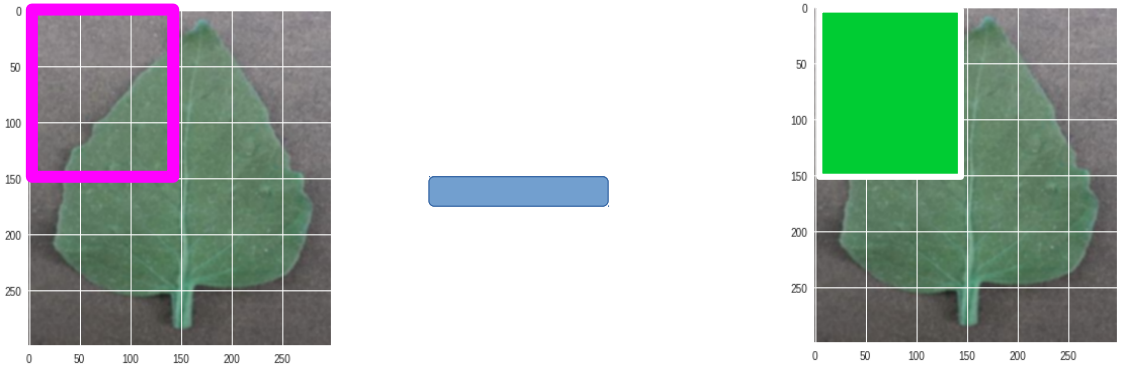


Applying to Mnist (3) – Individual analysis (b)



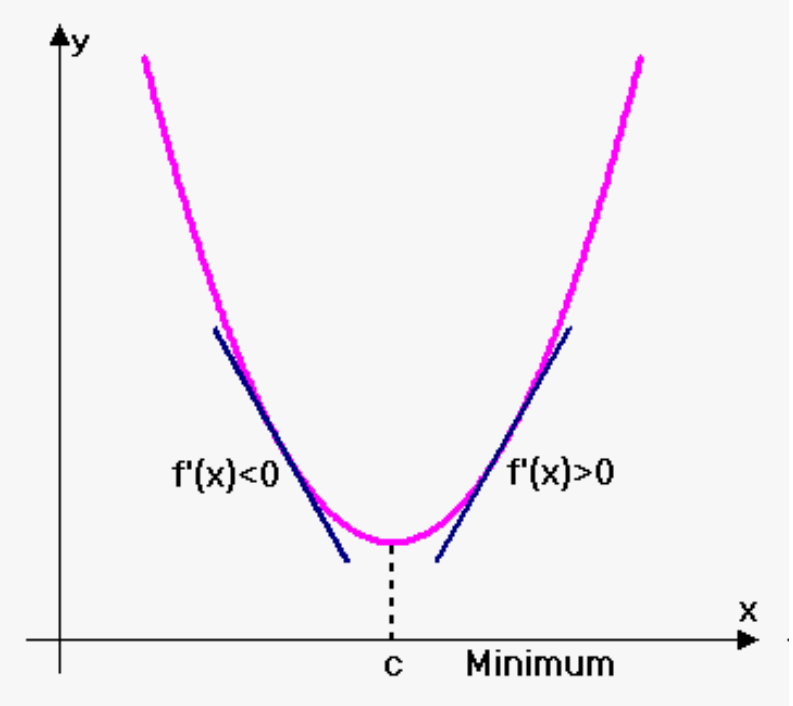
Drawbacks

- Computationally intensive, requires to compute 2^m examples for m features
  ~ I only sampled 10^3 out of 10^{14} combinations
- How do you appropriately remove a feature ?



Global explanation != local explanation

Gradient based methods

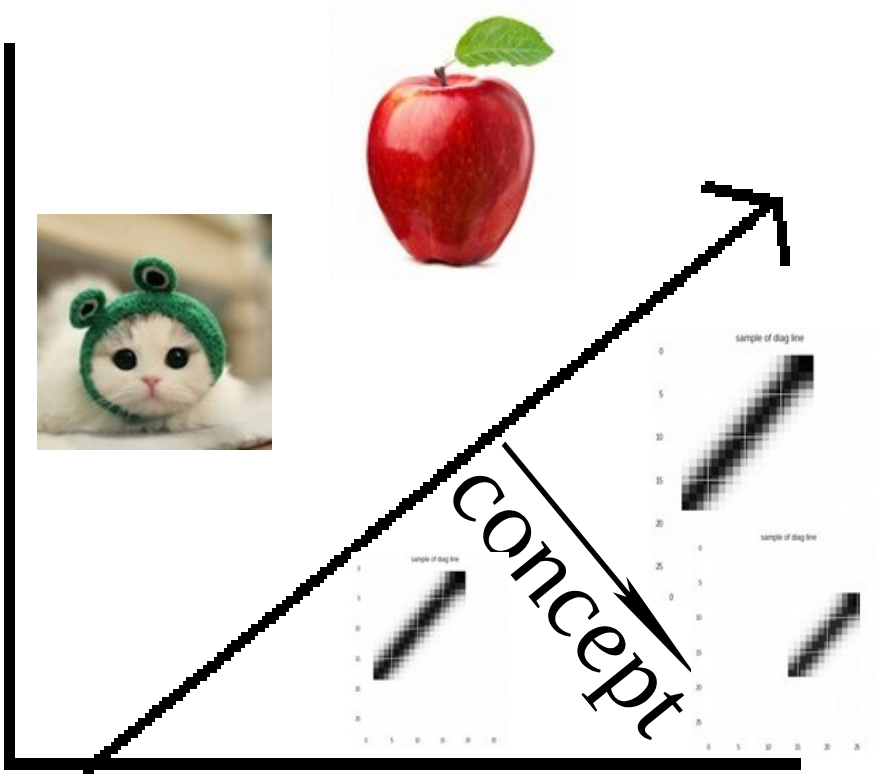


$$\frac{\partial\, output_i}{\partial\, inputs}$$

The gradient shows sensitivity of outputs to inputs

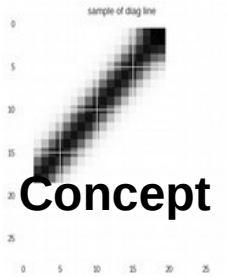
Gradient based methods : concept directional derivative

$$directional\,derivative = \frac{\partial\, output_i}{\partial\, X_{AnyLayer}} \cdot \overrightarrow{concept}$$



Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), 2017

Concept activation of mnist model hidden CNN layer : Toy Example



Concept

$$directional\ derivative = \frac{\partial\, output_i}{\partial\, X_{AnyLayer}} \cdot \overrightarrow{concept}$$

Digit	0	1	2	3	4	5	6	7	8	9
Directional derivative	--	++	--	--	--	--	--	+	---	--

The concept directional derivative measures sensitivity of model predictions with respect to concepts at any model layer

Concept activation of mnist model hidden CNN layer : Remove model bias

$$directional\,derivative = \frac{\partial\, output_{apron}}{\partial\, X_{AnyLayer}} \cdot \overrightarrow{woman}$$

The ‘apron’ predictions was positively correlated with respect to the ‘woman’ concept directional derivative

Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), 2017

Concept activation of mnist model hidden CNN layer : Inquire about model learning

- Train image classifier with captioned images (right)
- Concept directional derivative shows sensitivity of logit output to
  i). Image or
  ii). Captions

cab image with caption

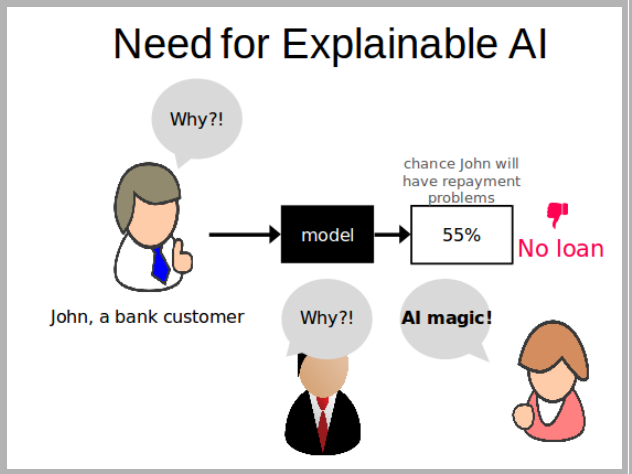
Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), 2017

Draw backs of local methods

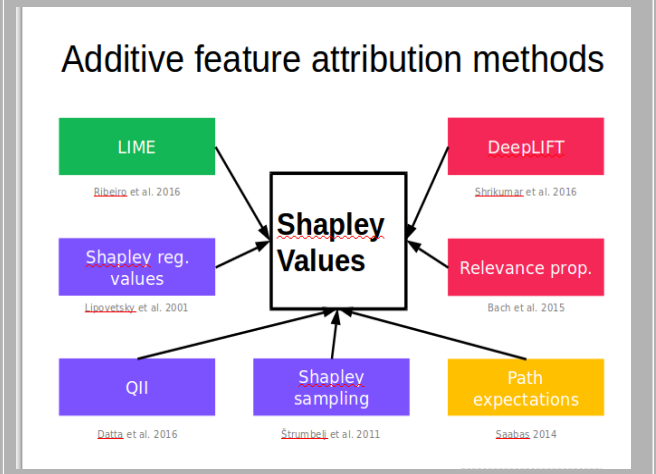
- Indirect – post-processing of model to yields insights

Summary

1.

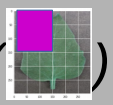
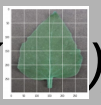


2.

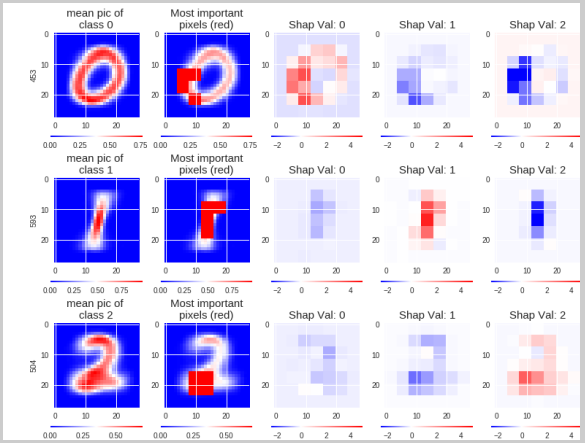


3. Intuition

$$g(z') = \sum_{i=1}^{M} \phi_i$$

φ_{pink} = *weight_avg(f(*  *) − f(*  *))*

4. Analysis mnist



5. Drawbacks

6. Gradient based methods

$$\frac{\partial J}{\partial X} \cdot \overrightarrow{concept}$$

References

- Scotts slides https://github.com/slundberg/shap/blob/master/docs/presentations/NIPS%202017%20Talk.pptx
- A Unified Approach to Interpreting Model Predictions(2017), Scott Lundberg, Su-In Lee
- Analysis of regression in game theory approach (2001),Stan Lipovetsky, Michael Conklin
- Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), 2017