

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)

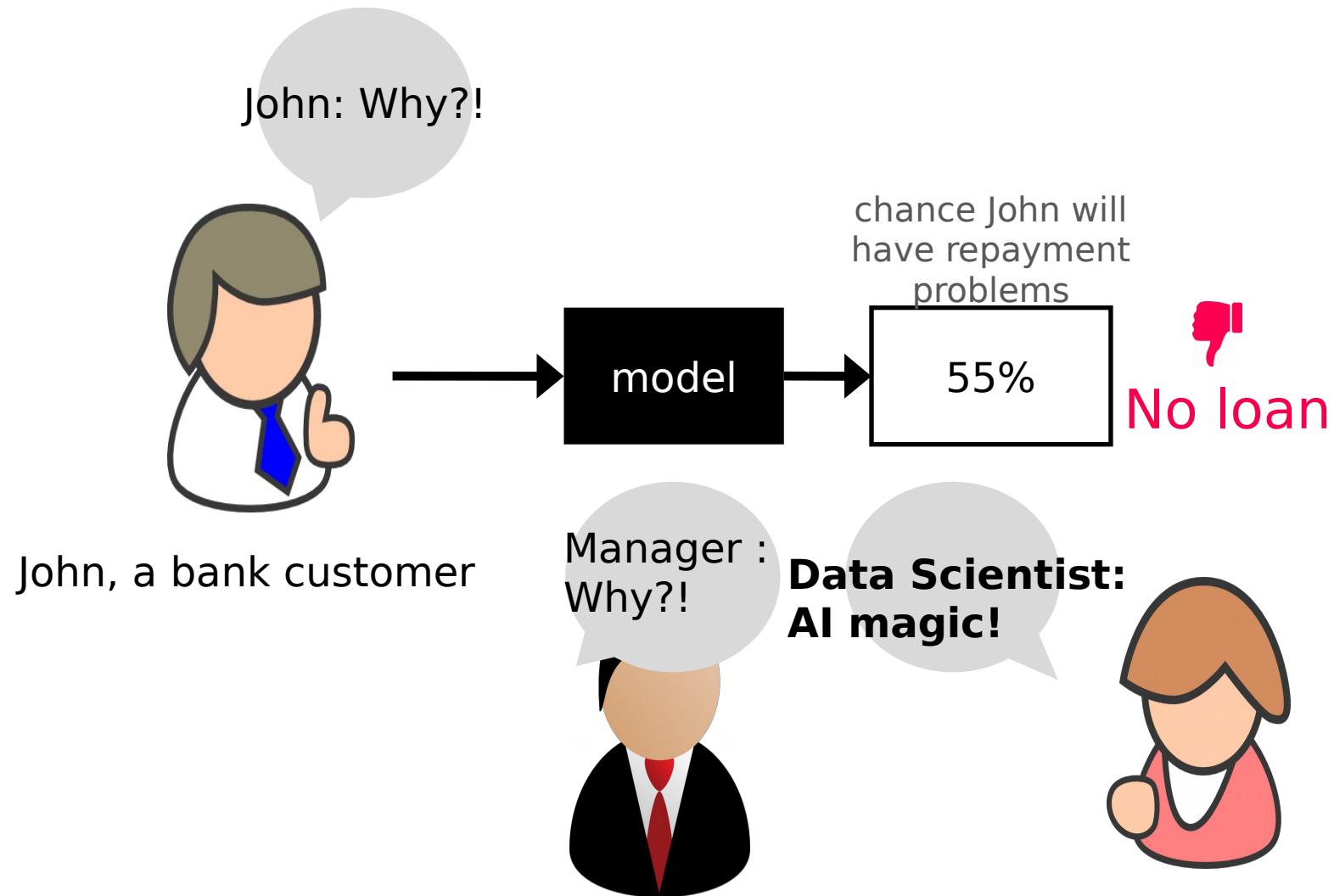
https://colab.research.google.com/github/leexa90/Explainable_AI_image_classification/blob/master/colabs_script.ipynb

Background on myself

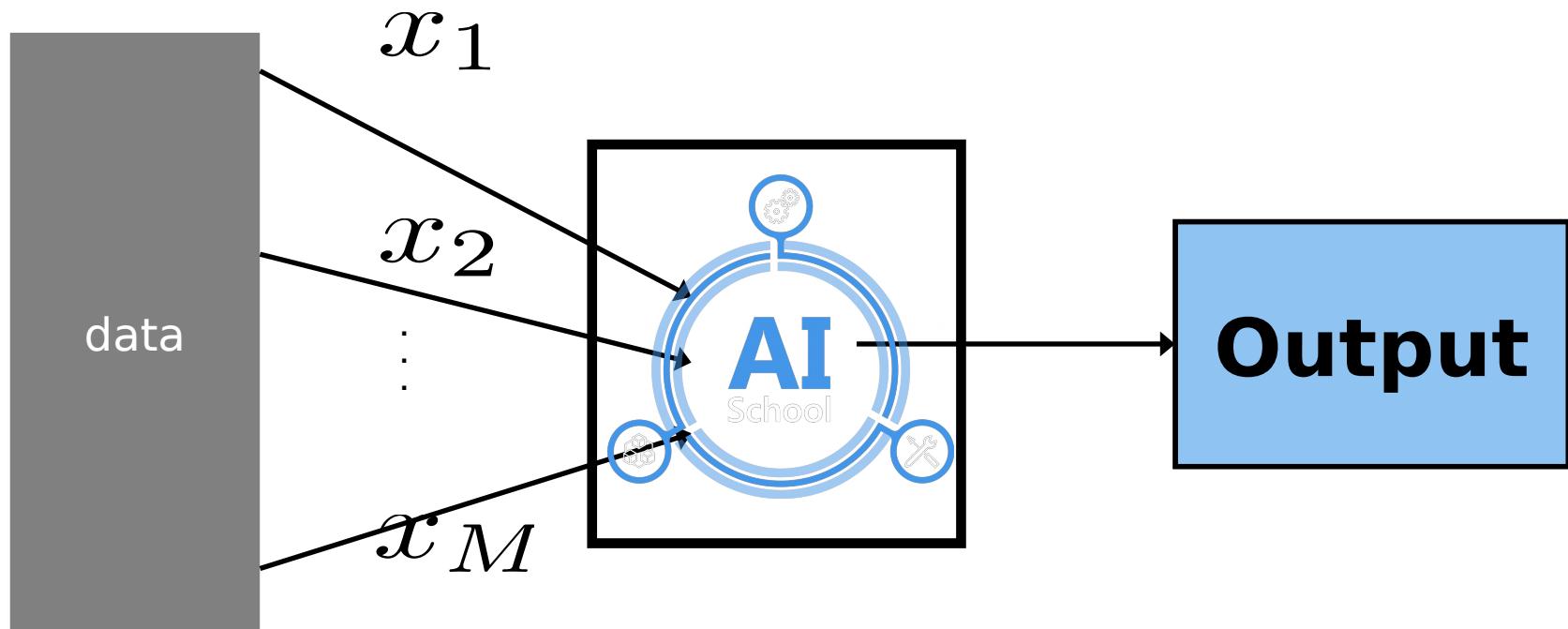
- Graduated from NUS science 2015
- Attended Deep Learning Developer Course 2017
- Working in A*STAR Bioinformatics Institute, using various methods and analysis on crop analytics in smart urban farms



Need for Explainable AI



Complicated AI Model



Explainable model: Shapley Values (1)

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

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

Φ_i is the shapley value of the feature i

Explainable model: Shapley Values (2)



Base rate

Prediction for John

20%

55%

0
↓

$E[f(x)]$
↓

$f(x)$
↓

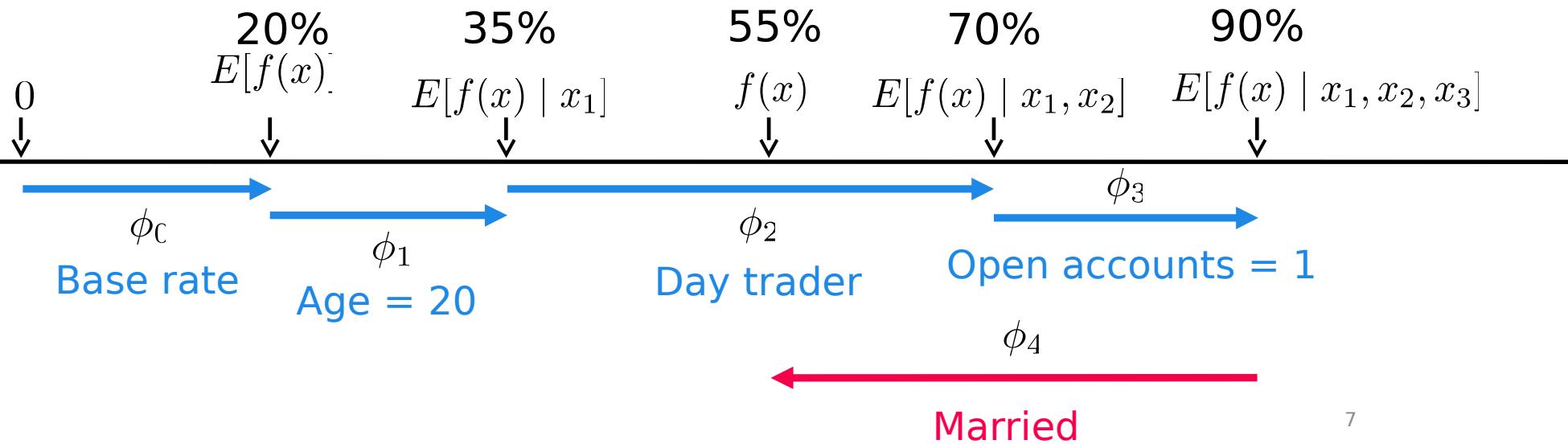
$$55\% = \sum_{i=1}^M \phi_i$$

How did we get here?

Explainable model: Shapley Values (2)



$$55\% = \sum_{i=1}^M \phi_i$$



Explainable model: Shapley Values (3) – Computation

- Train AI model

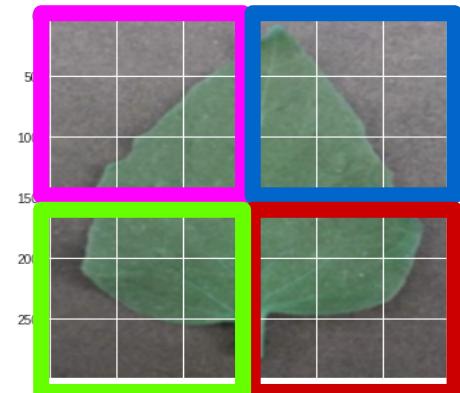
For each picture containing 4 superpixels

{*Explain model :*

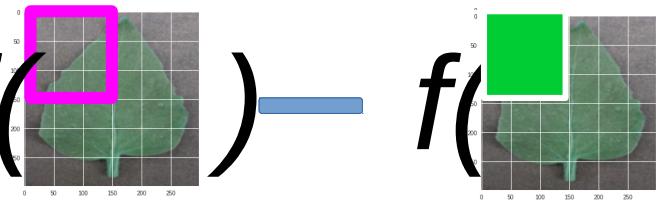
$$output = \phi_{pink} + \phi_{blue} + \phi_{green} + \phi_{red}$$

ϕ is the shapley value of the feature

}



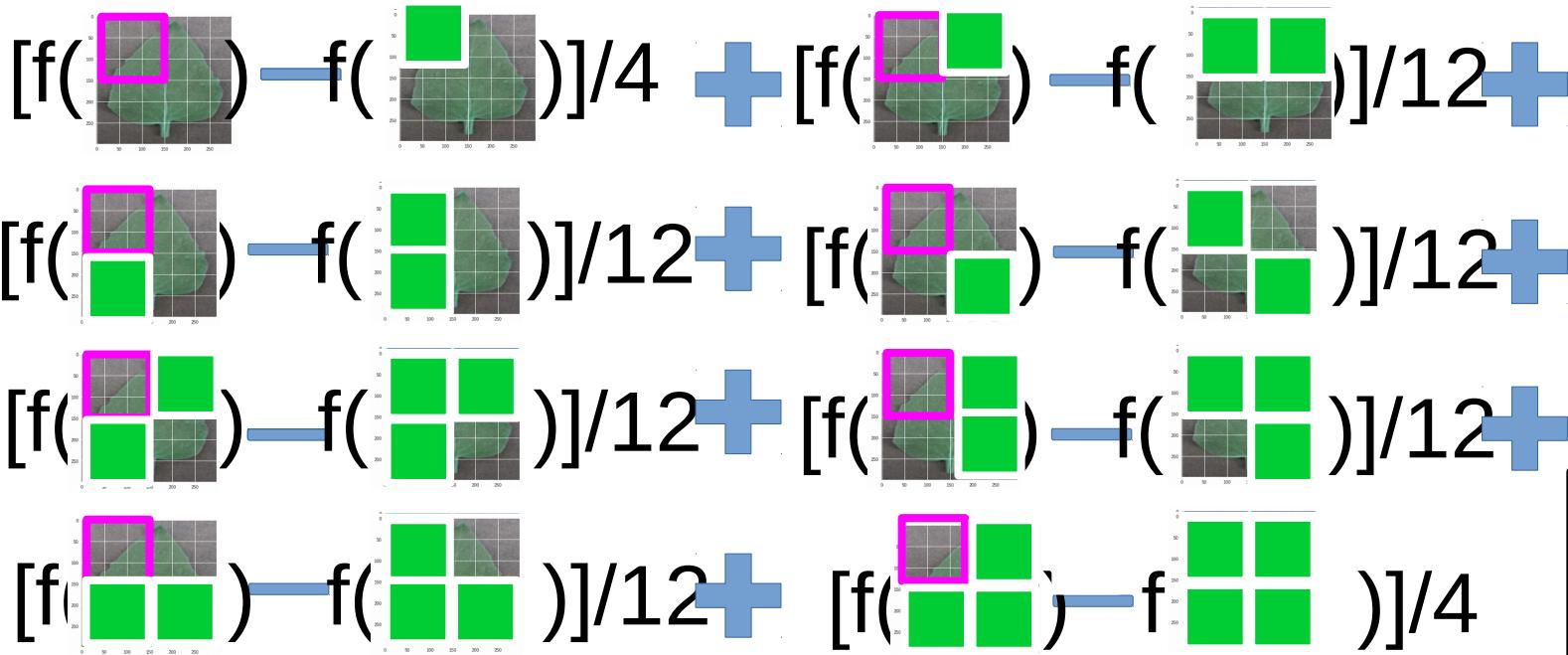
Explainable model: Shapley Values (4) – Computation

$$\phi_{pink} = weight_avg(f(\text{[pink square]}) - f(\text{[green square]}))$$


$$[f(\text{[pink square]}) - f(\text{[green square]})]/4 + [f(\text{[pink square]}) - f(\text{[green square]})]/12 +$$

$$[f(\text{[pink square]}) - f(\text{[green square]})]/12 + [f(\text{[pink square]}) - f(\text{[green square]})]/12 +$$

$$[f(\text{[pink square]}) - f(\text{[green square]})]/12 + [f(\text{[pink square]}) - f(\text{[green square]})]/12 +$$

$$[f(\text{[pink square]}) - f(\text{[green square]})]/12 + [f(\text{[pink square]}) - f(\text{[green square]})]/4$$


Legend
 Green solid squared are mean-filled super-pixels

Explainable model: Shapley 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 Shapley to Mnist (1)

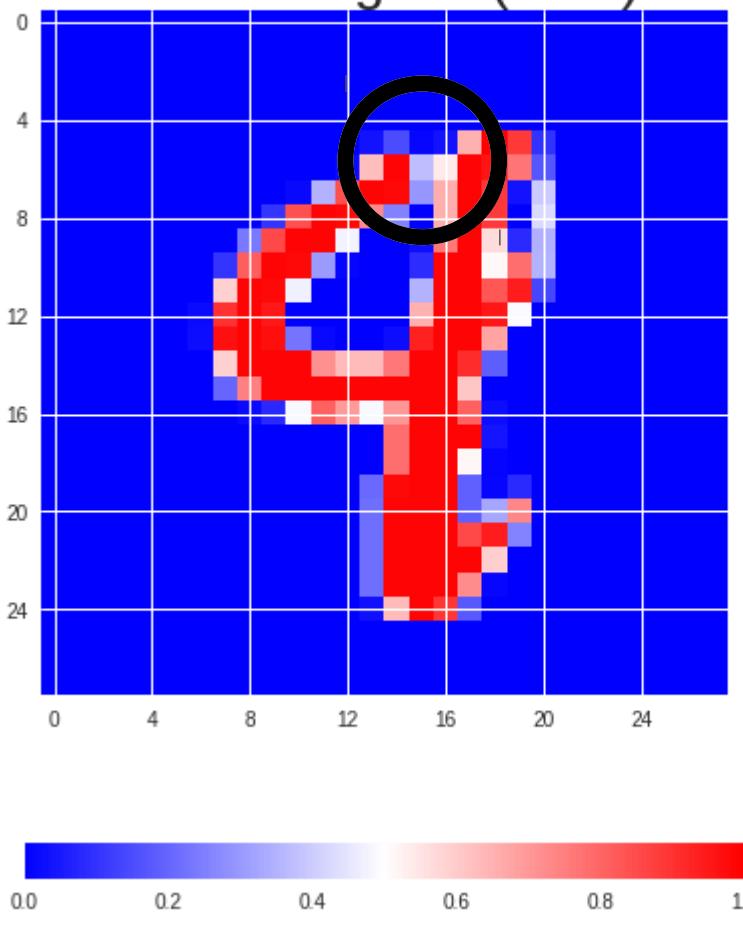
- Trained Mnist model
- for each test image {
 - Split image to $7 \times 7 = 49$ super-pixels
 - Sampled 7367 permutations of mean-filled super-pixels
 - Calculate shapley values for each super-pixel using weighted regression}

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

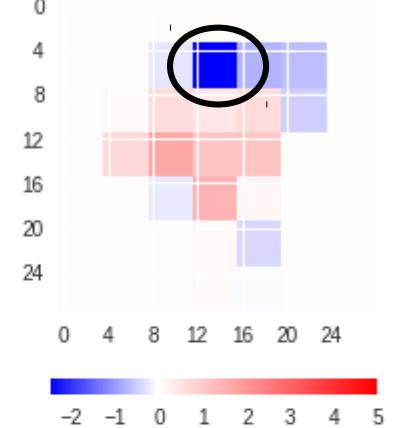
Applying Shapley to Mnist (2) – Individual analysis (a)

Test Example: 359

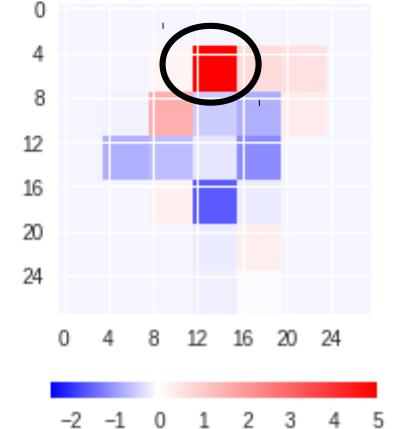
Predicted Digit: 4 (58%),
Actual Digit: 9 (40%)



Test Example : 359
shapley value for digit 4



Test Example : 359
shapley value for digit 9

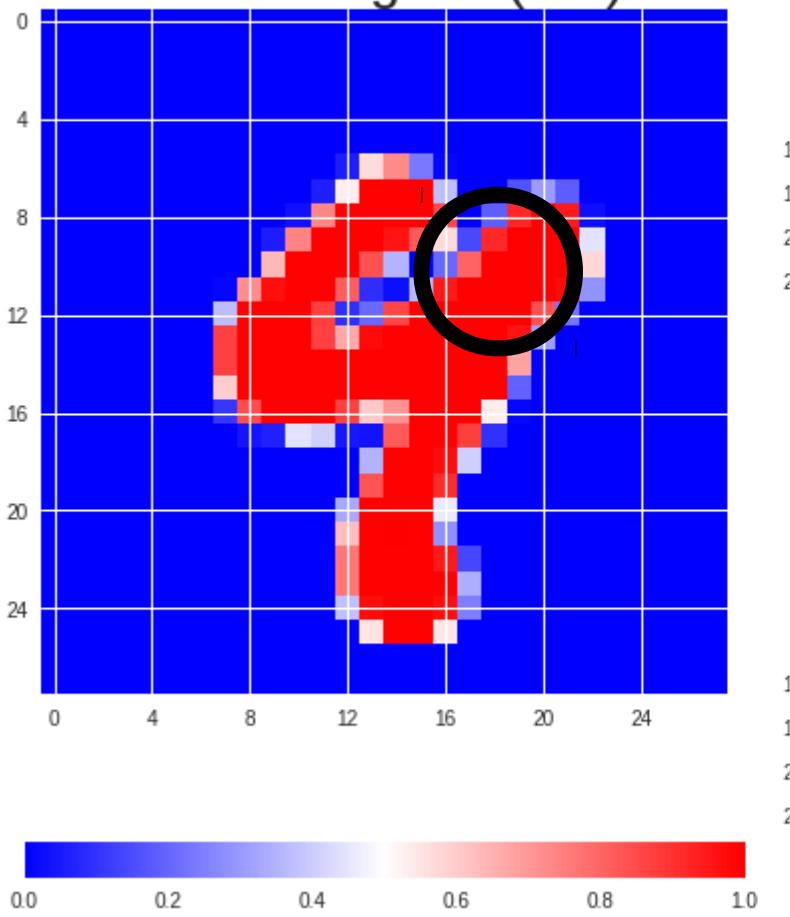


Applying Shapley to Mnist (3) – Individual analysis (b)

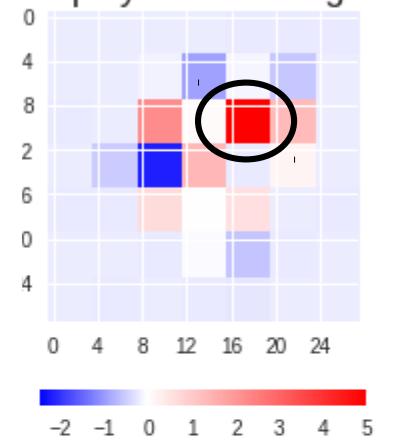
Test Example: 2130

Predicted Digit: 9 (93%),

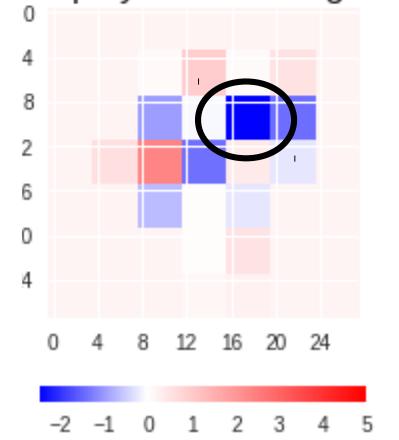
Actual Digit: 4 (6%)



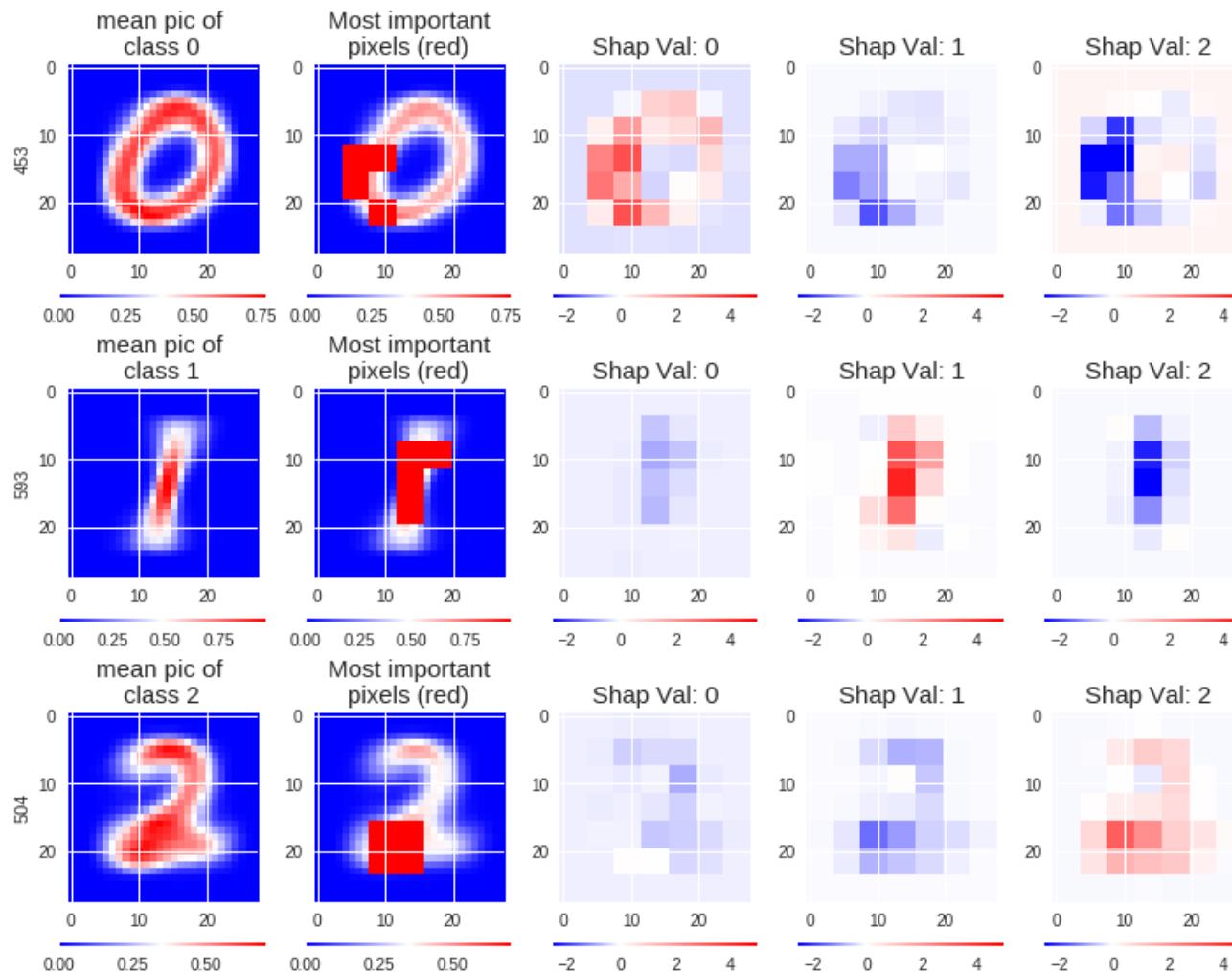
Test Example : 2130
shapley value for digit 9



Test Example : 2130
shapley value for digit 4

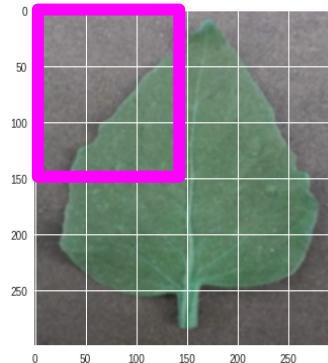


Applying Shapley to Mnist (4) – Global analysis

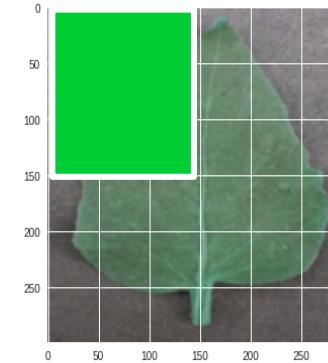


Shapley : Drawbacks

- 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 ?

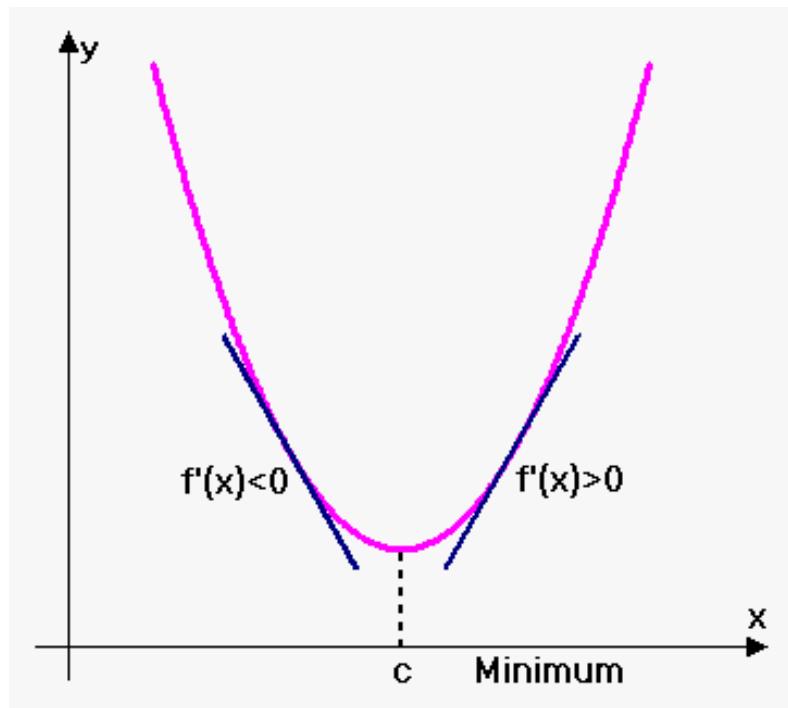


↓



- Global explanation \neq local explanation

Gradient based methods

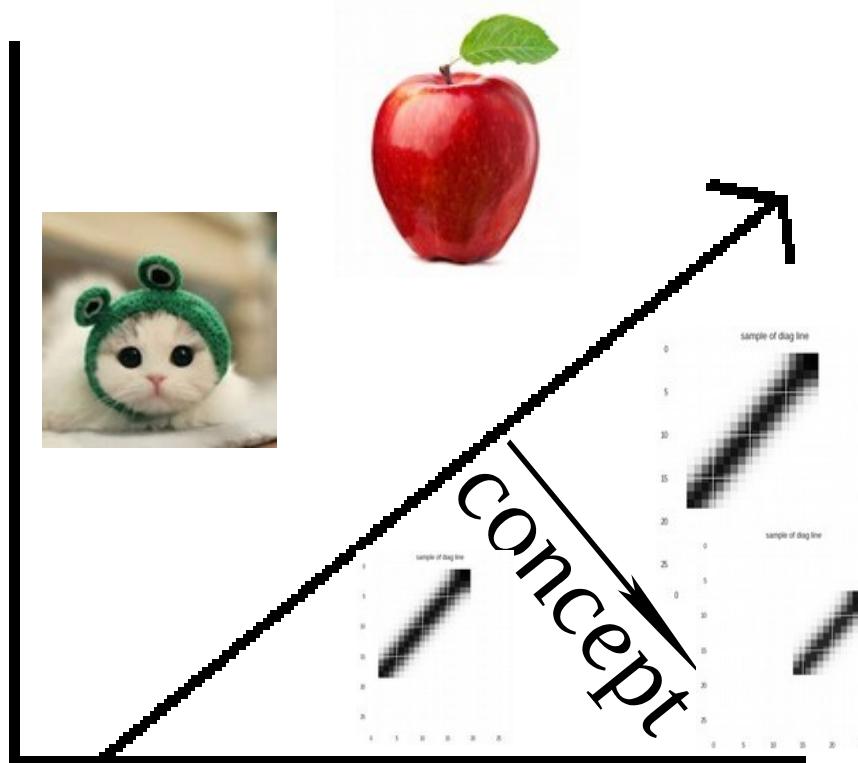


$$\frac{\partial \text{output}_i}{\partial \text{inputs}}$$

The gradient shows sensitivity of outputs to inputs

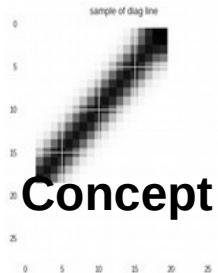
Gradient based methods : concept directional derivative

$$\text{directional derivative} = \frac{\partial \text{output}_i}{\partial X_{\text{AnyLayer}}} \cdot \overrightarrow{\text{concept}}$$



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

Concept activation of mnist model hidden CNN layer : Toy Example



$$\text{directional derivative} = \frac{\partial \text{output}_i}{\partial X_{\text{AnyLayer}}} \cdot \overrightarrow{\text{concept}}$$

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

$$\text{directional derivative} = \frac{\partial \text{output}_{\text{apron}}}{\partial X_{\text{AnyLayer}}} \cdot \overrightarrow{\text{woman}}$$

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

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



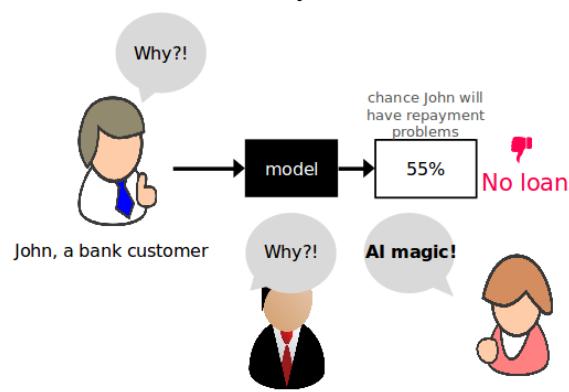
Draw backs of methods

- Indirect – post-processing of model to yields insights
- Explain the concept vector– difficult in high D space

Summary

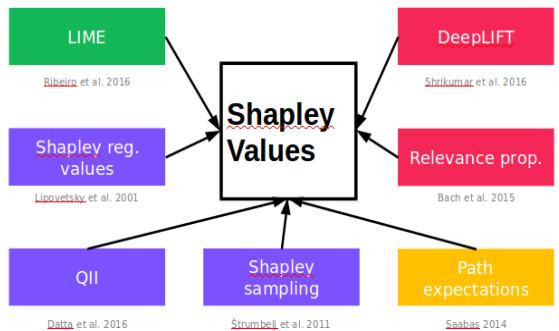
1.

Need for Explainable AI



2.

Additive feature attribution methods

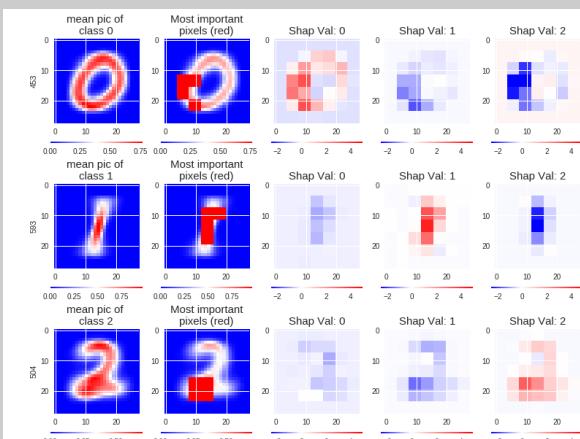


3. Intuition

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

$$\varphi_{pink} = weight_avg(f(\text{pink}) - f(\text{green}))$$

4. Analysis mnist



5. Drawbacks



6. Gradient based methods

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

References and Questions ?

- 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