Explainable AI : Shapley Values and Concept Activation Vectors

1. A Unified Approach to Interpreting Model Predictions

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

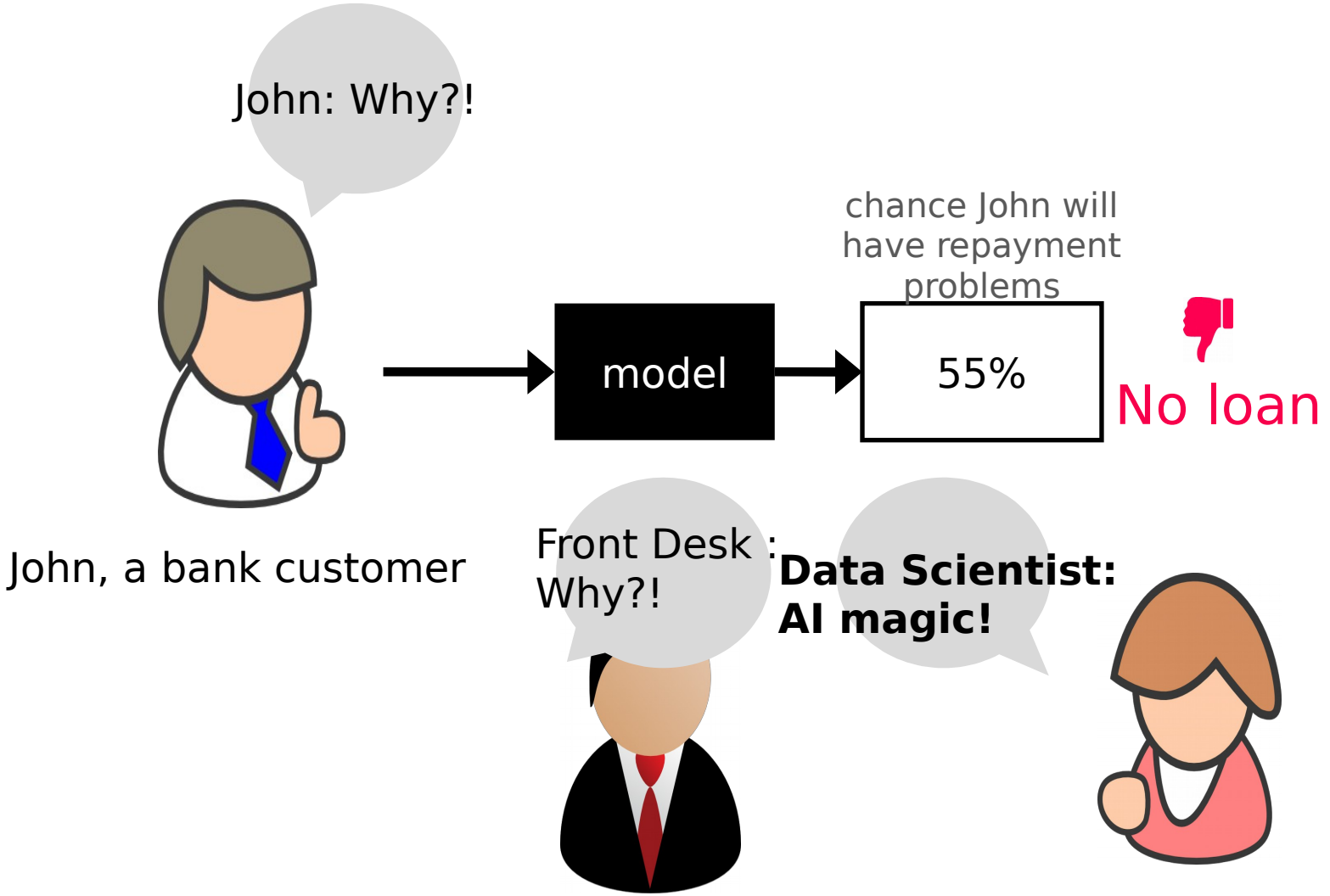
Github:leexa90/
Explainable_AI_image_classification

Lee Xiong An

Background on myself

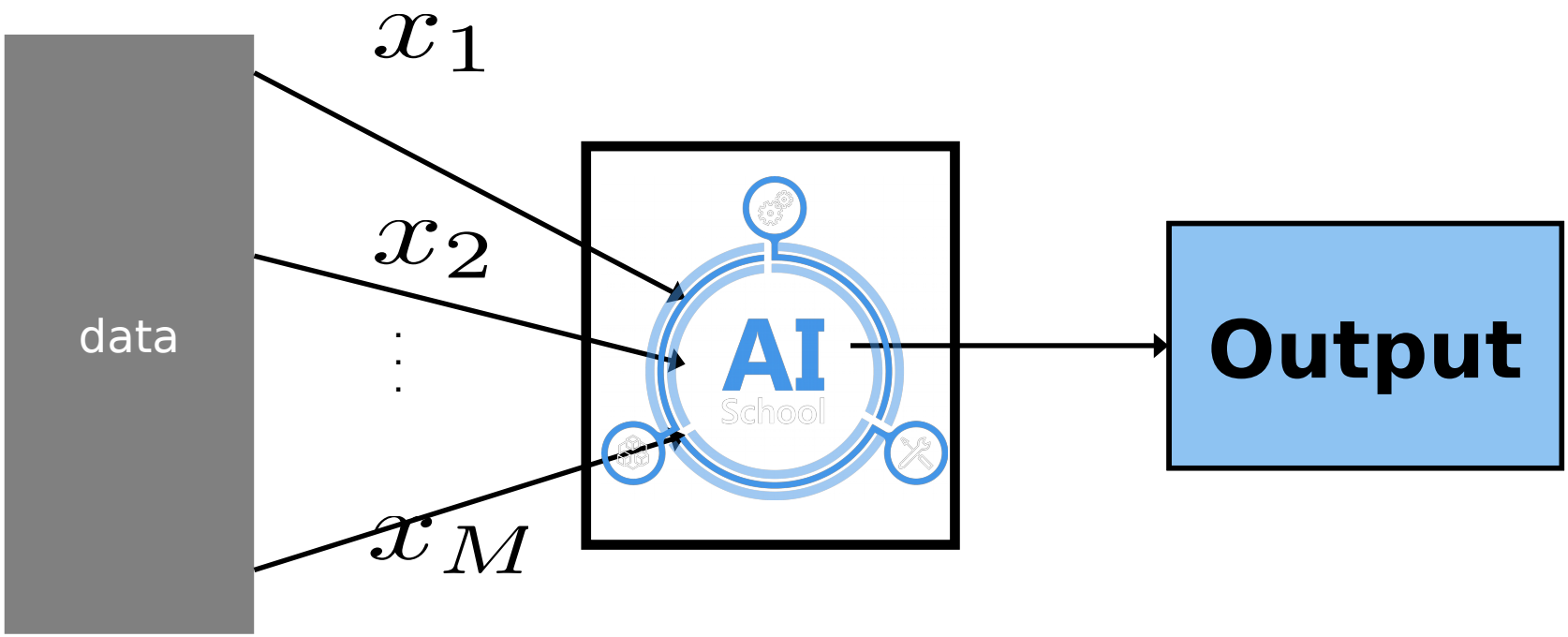
- Graduated from NUS science 2015
- Attended Deep Learning Developer Course 2017
- Working in A*STAR Bioinformatics Institute, using various methods and analysis on crop analytics in smart urban farms

Need for Explainable AI



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

Complicated AI Model



Feature Additive Methods: Shapley Values (1)

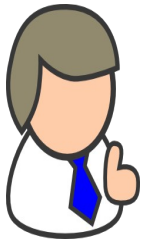
Output = $\sum_{i=1}^{M} \phi_i$

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

$\boldsymbol{\phi}_i$ is the shapley value of the feature i

https://github.com/slundberg/shap

Feature Additive Methods: Shapley Values (2)



Base rate

20%

$E[f(x)]$

Prediction for John

55%

$f(x)$

0

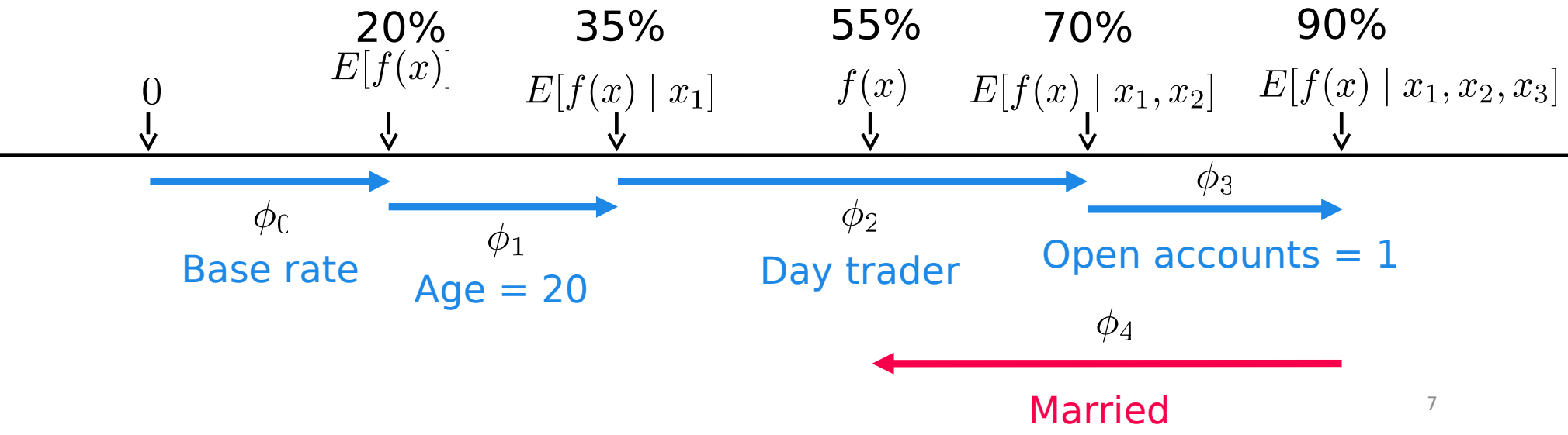
55% = $\sum_{i=1}^{M} \phi_i$

How did we get here?

https://github.com/slundberg/shap

Feature Additive Methods: Shapley Values (2)

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



https://github.com/slundberg/shap

Feature Additive Methods: Shapley Values (3) – Toy Example

- Train AI model

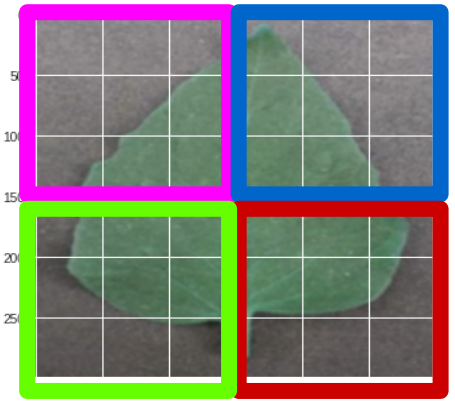
For each picture containing 4 superpixels

{*Explain model :*

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

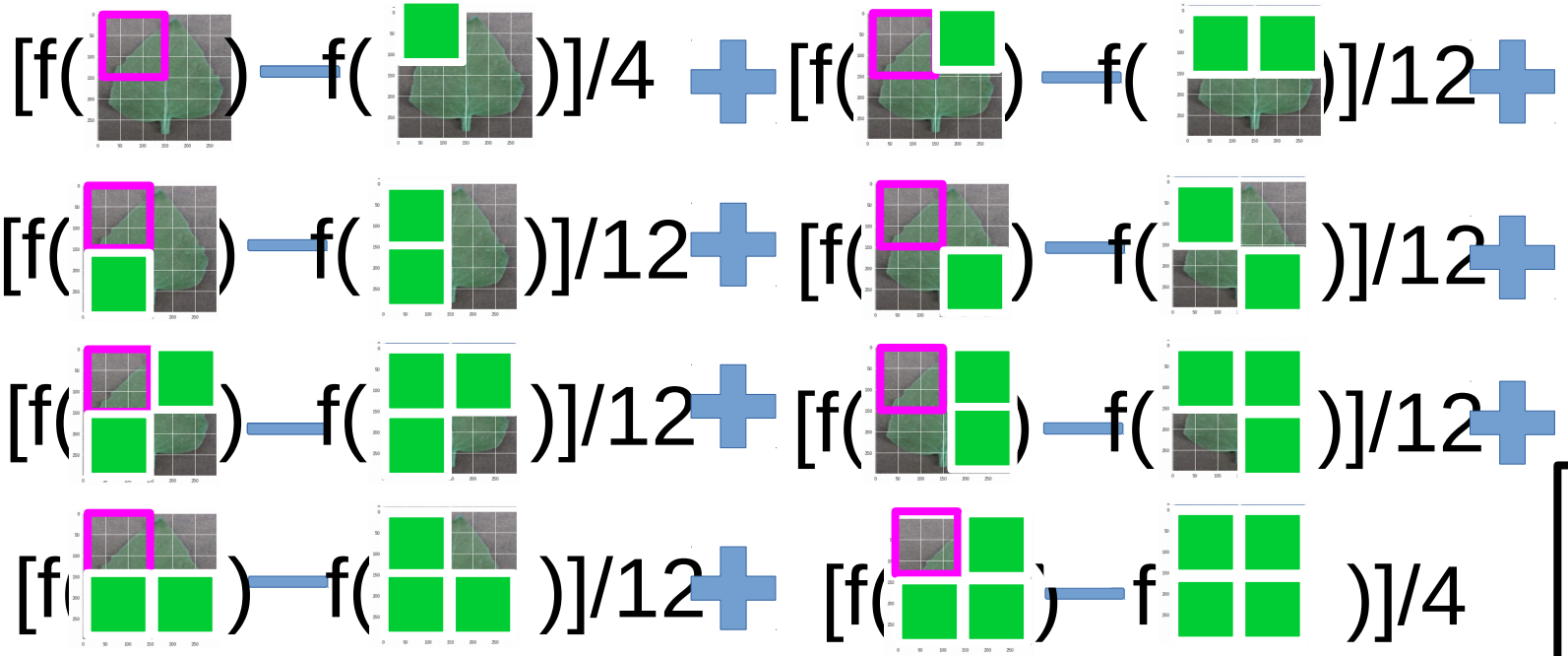
ϕ is the shapley value of the feature

}



Feature Additive Methods: Shapley Values (4) – Toy Example

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



Legend
Green solid squared are mean-filled super-pixels

Feature Additive Methods: 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

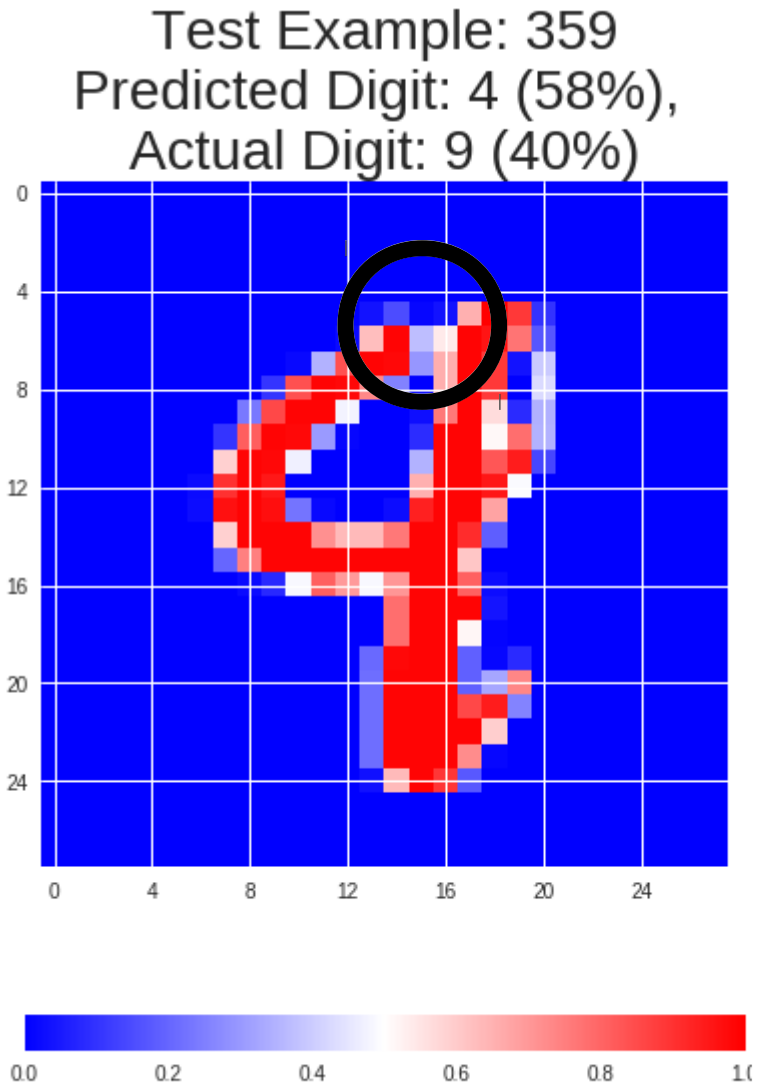
Applying Shapley to Mnist (1)

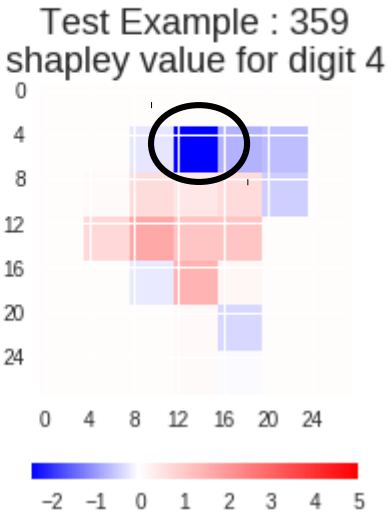
- Trained Mnist model
- for each test image {
  - Use 7*7 = 49 super-pixels instead of 784 pixels
  - Sampled 7367 permutations of mean-filled super-pixels
  - Calculate shapley values for each super-pixel using weighted regression

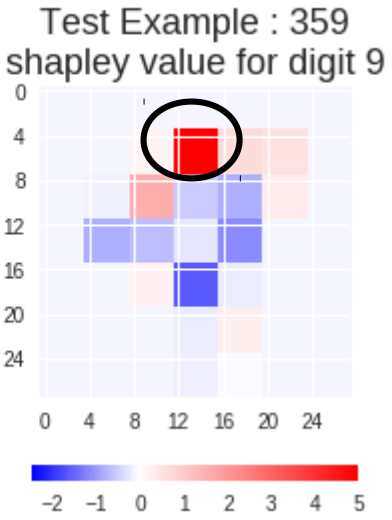
  }

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

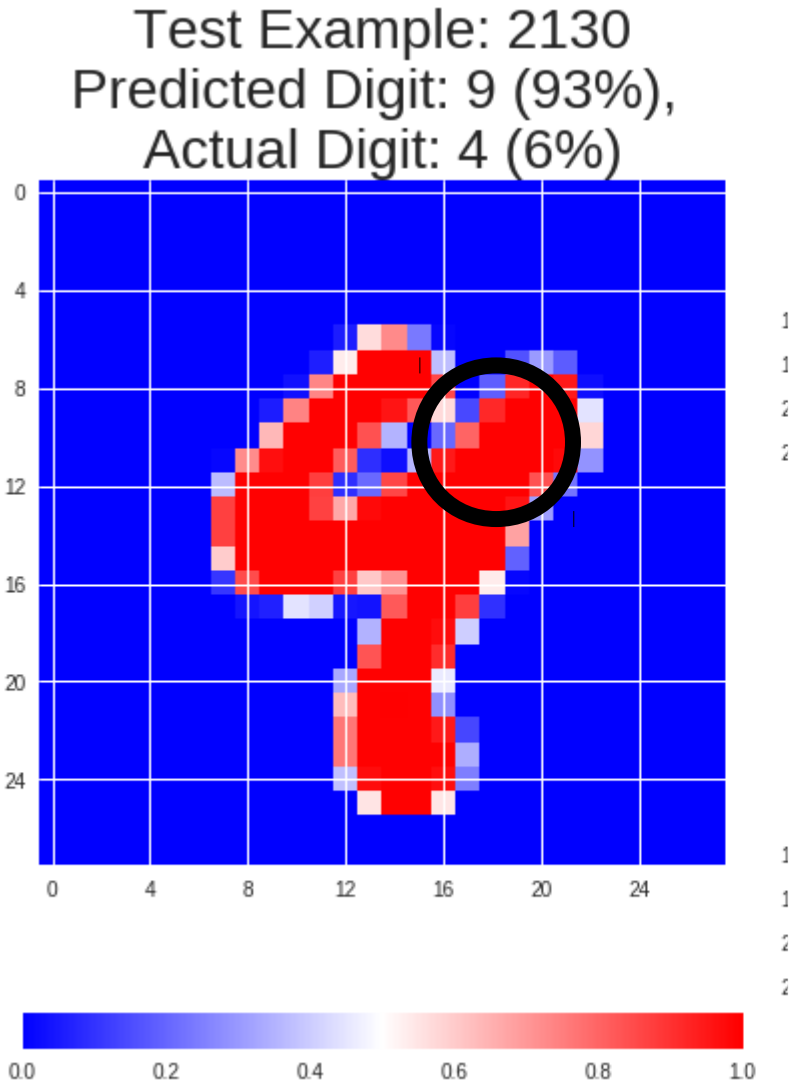
Applying Shapley to Mnist (2) – example (a)

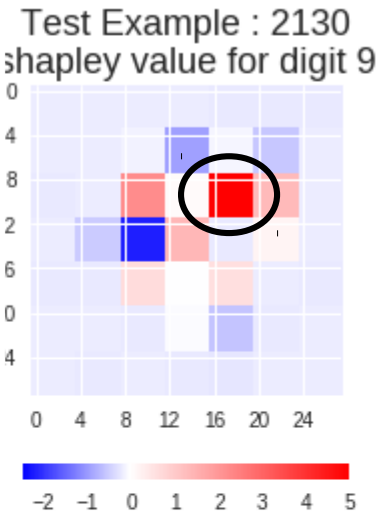


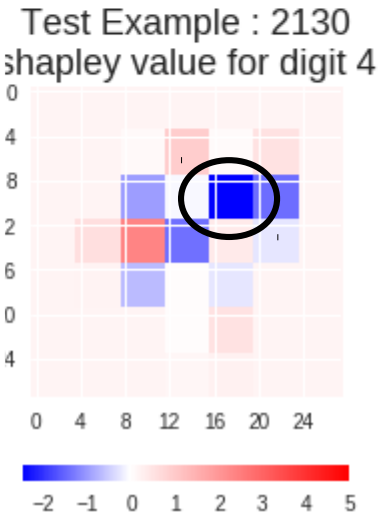




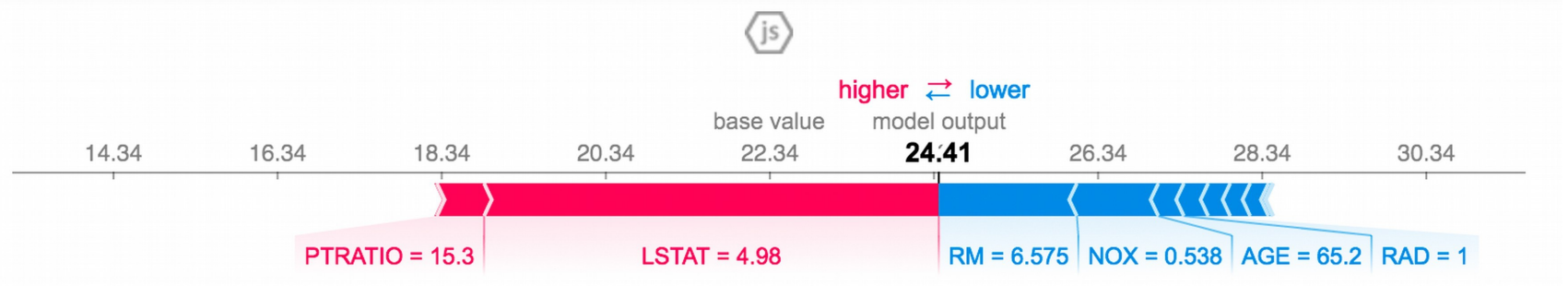
Applying Shapley to Mnist (3) – example (b)

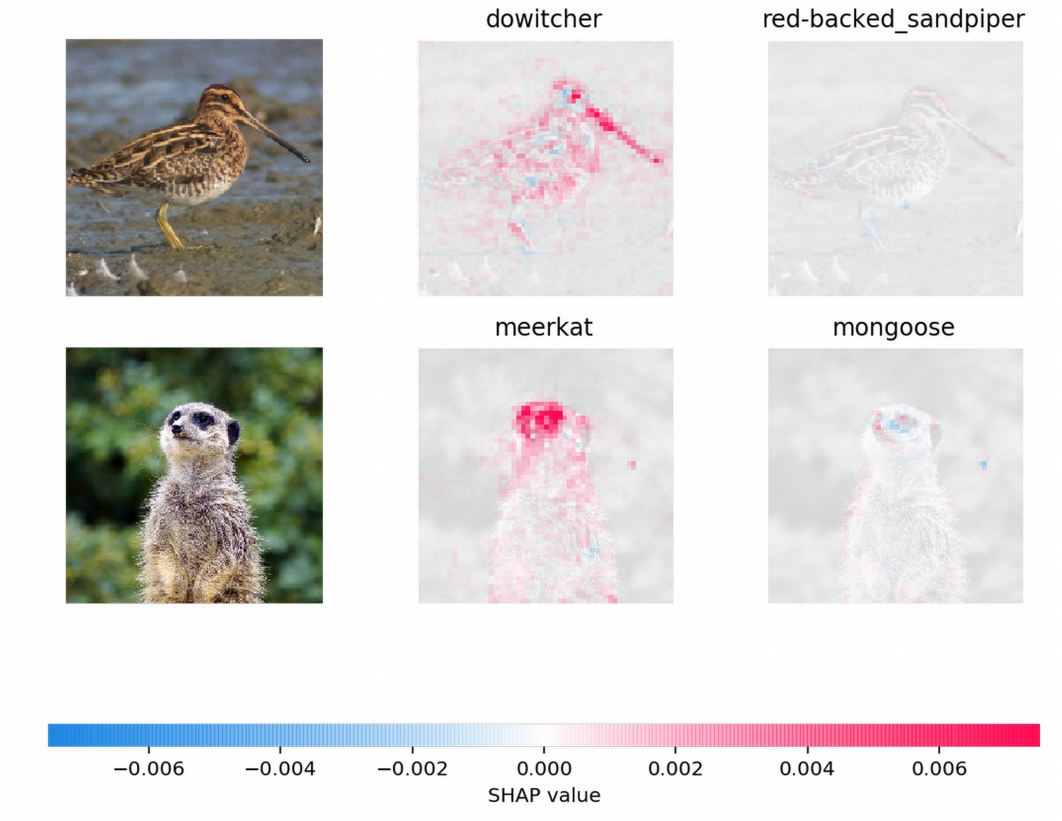






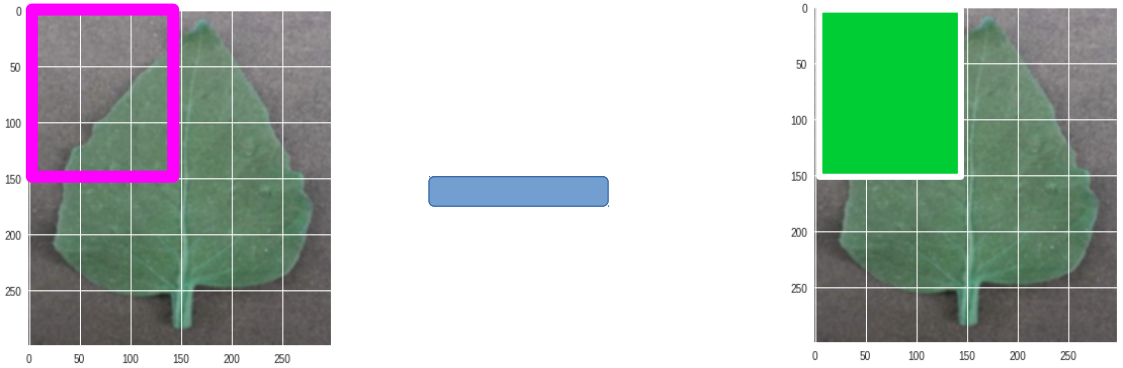
https://github.com/slundberg/shap





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 != local explanation

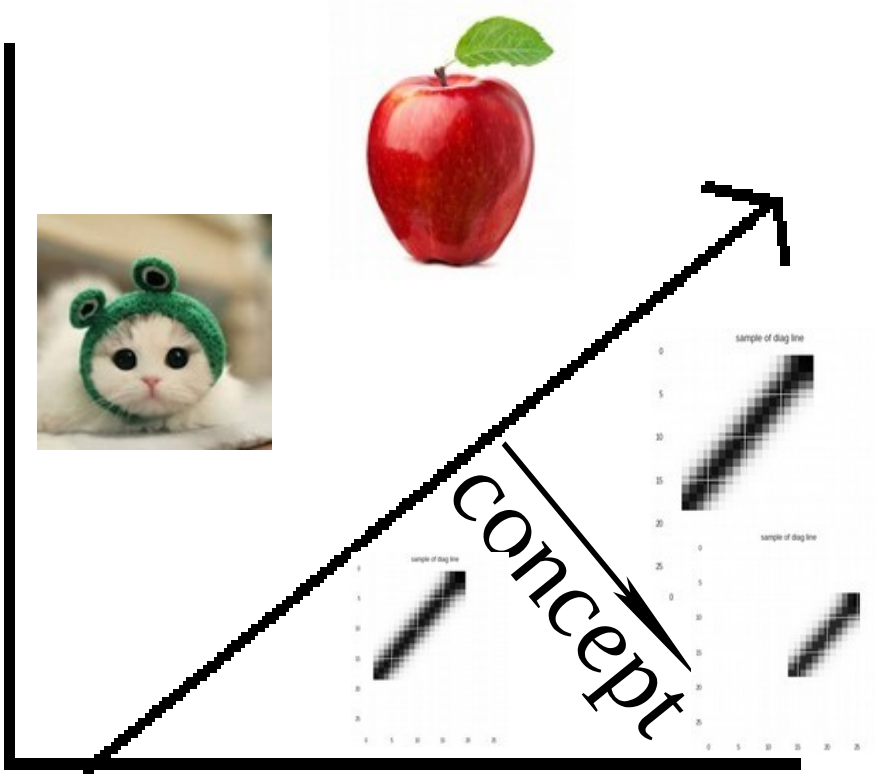
Part 2

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

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, Rory Sayres

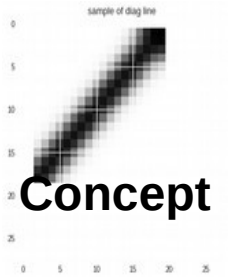
Gradient based methods : concept directional derivative

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



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 predictions with respect to concepts at a model layer(s)

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 has a positive gradient with respect to the ‘*woman*’ concept
Higher woman concept == higher prediction of apron

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

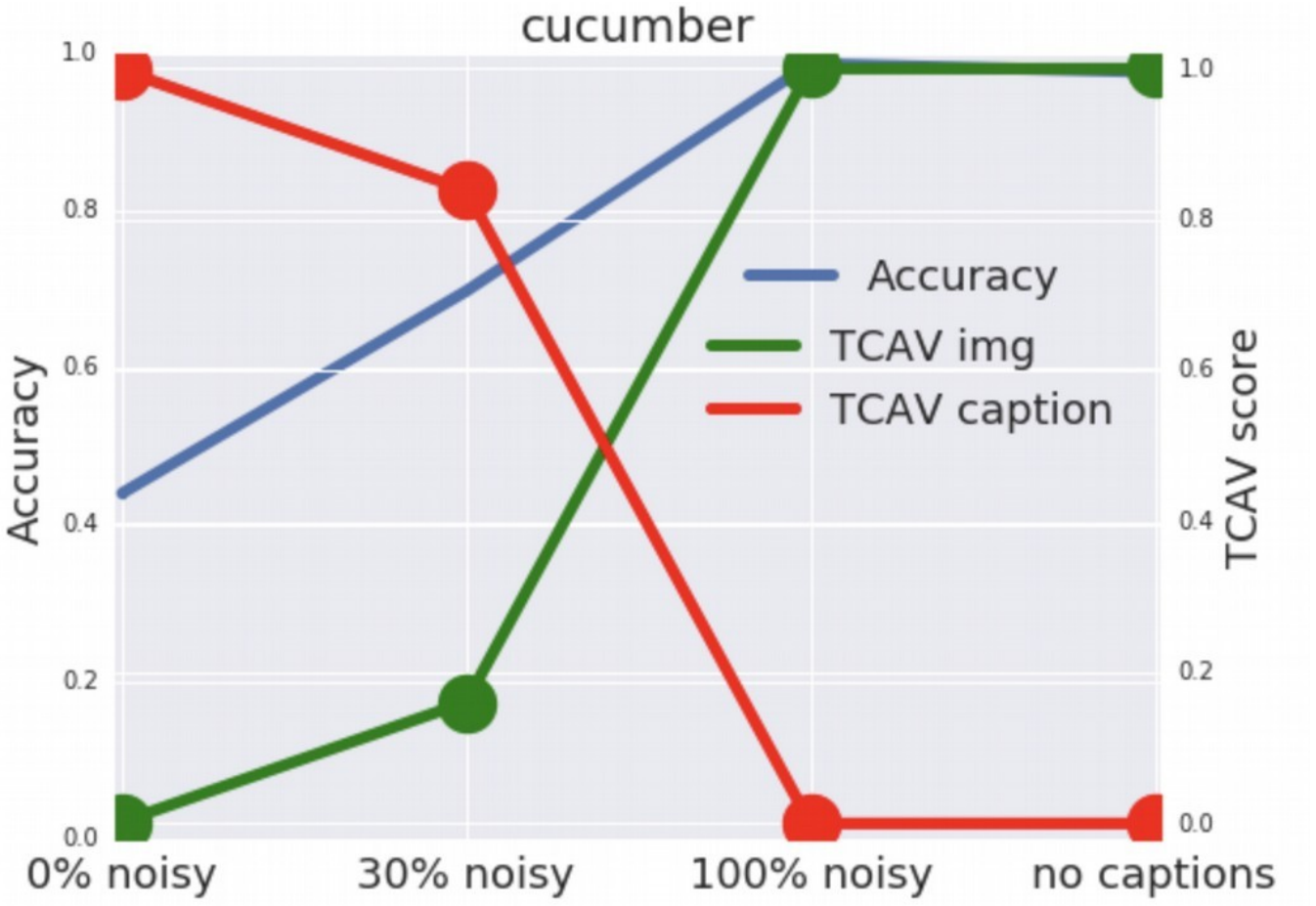
Concept activation of mnist model hidden CNN layer : Inquire about model learning

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

cucumber with caption cab image with caption

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

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



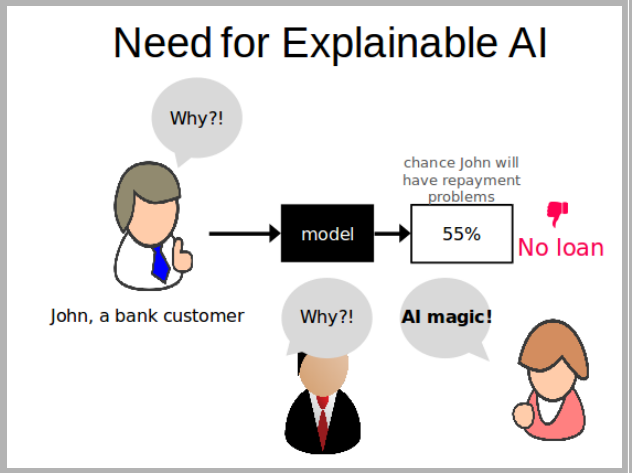
cucumber with caption

Draw backs of methods

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

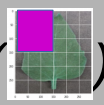
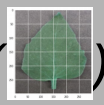
Summary

1.

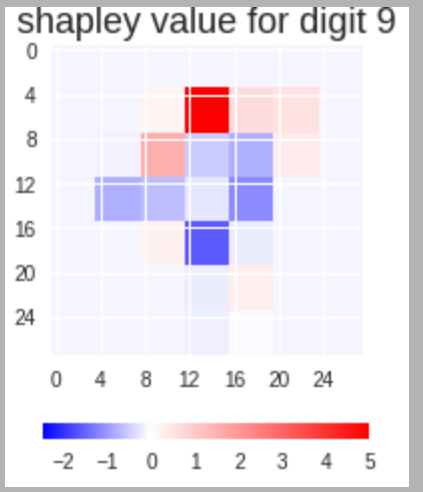


2. Feature Attribution methods

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

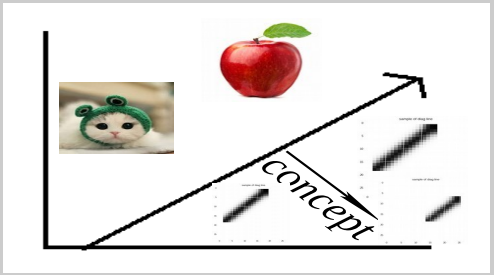
φ_{pink} = *weight_avg(f(*  *) − f(*  *))*

3. Analysis mnist



4. Concept directional derivative

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



5. Applications

Remove model bias
Inquire model learning

6. Drawbacks

References and Questions ?

- Scott's slides https://github.com/slundberg/shap/blob/master/docs/presentations/NIPS%202017%20Talk.pptx
- A Unified Approach to Interpreting Model Predictions(2017)
- Analysis of regression in game theory approach (2001),Stan Lipovetsky, Michael Conklin
- Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), 2017
- **leexa90**/Explainable_AI_image_classification