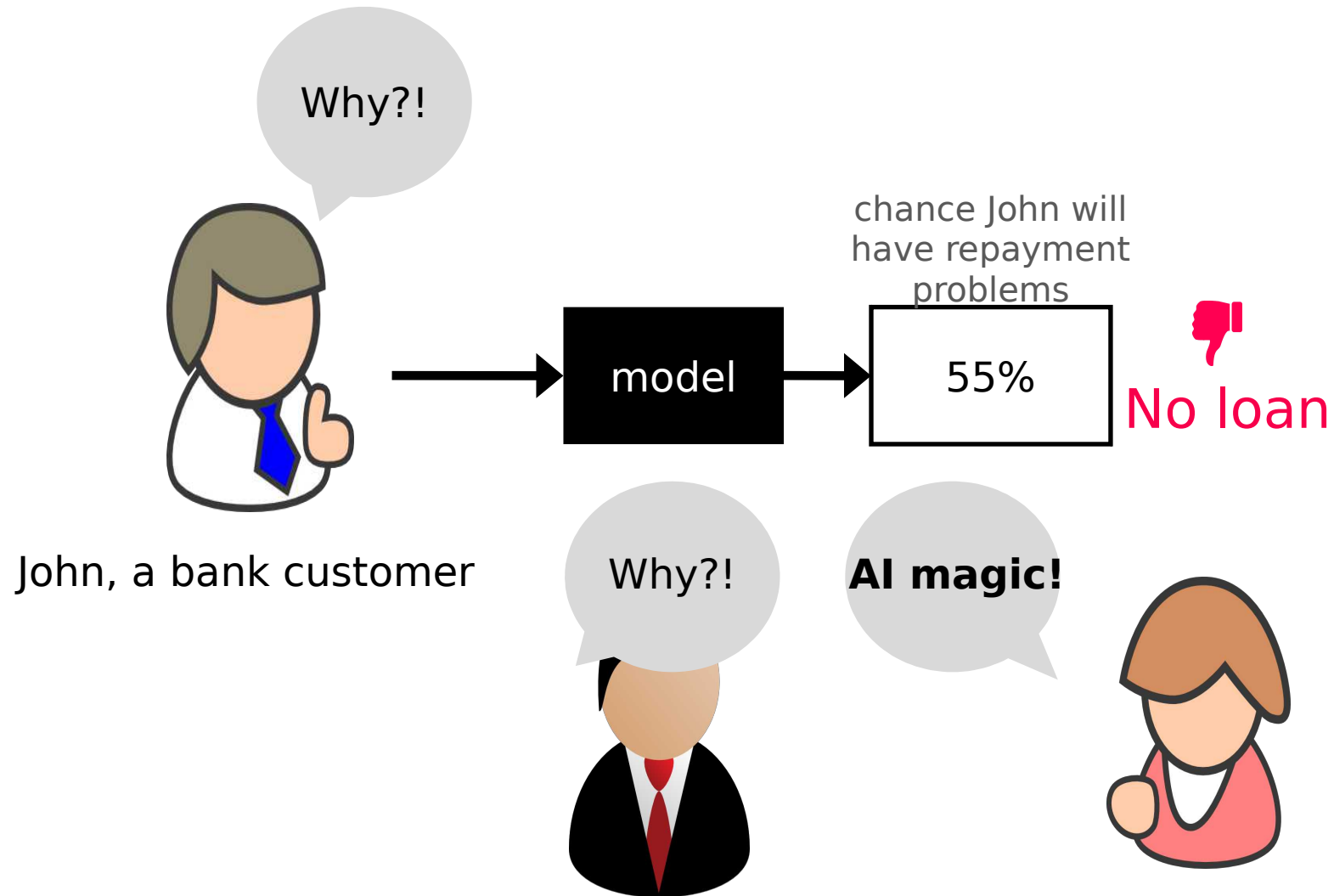


# Explainable AI : Shapley Values

A Unified Approach to Interpreting Model  
Predictions

**Scott Lundberg**, Su-In Lee

# Need for Explainable AI



# Need for Explainable AI

Some of the articles of GDPR can be interpreted as requiring explanation of the decision made by a machine learning algorithm, when it is applied to a human subject.

UW Prof. Pedro Domingos, a leading AI researcher, started a firestorm with his tweet



**Pedro Domingos**

@pmddomingos



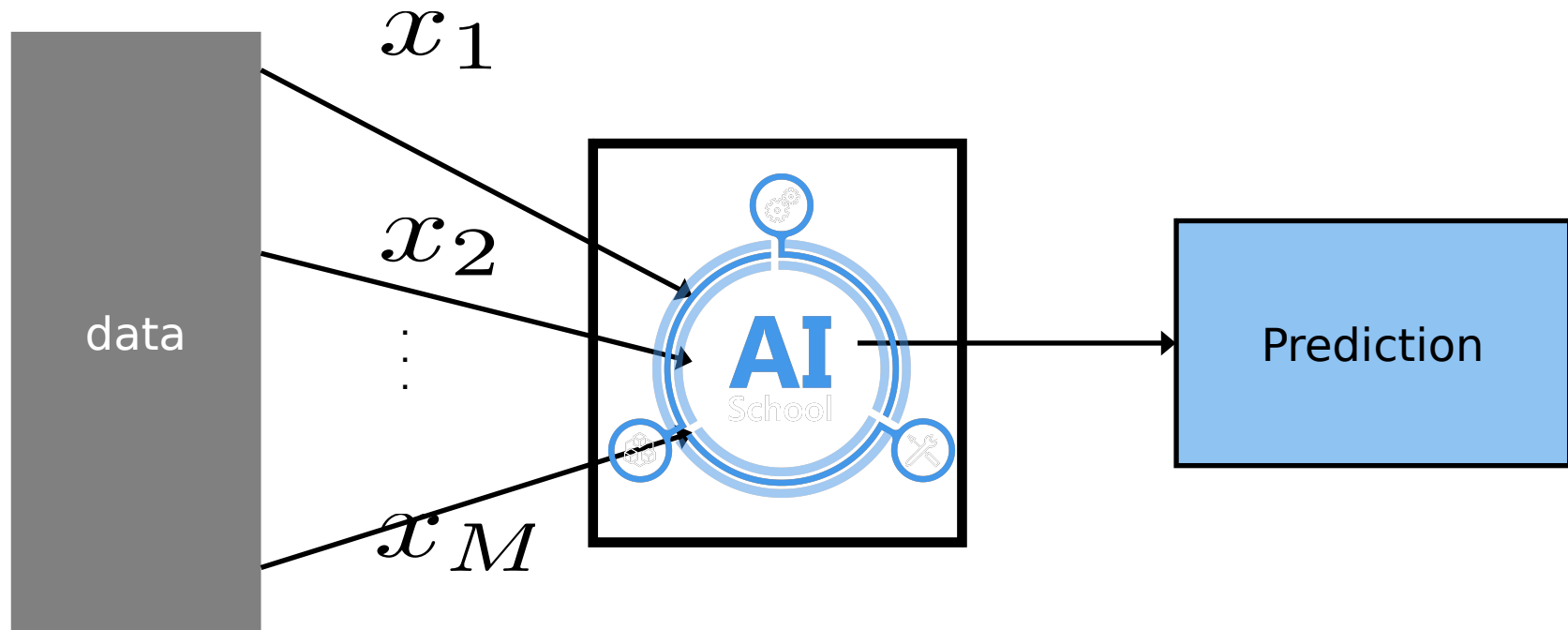
Starting May 25, the European Union will require algorithms to explain their output, making deep learning illegal.

11:59 AM - Jan 29, 2018

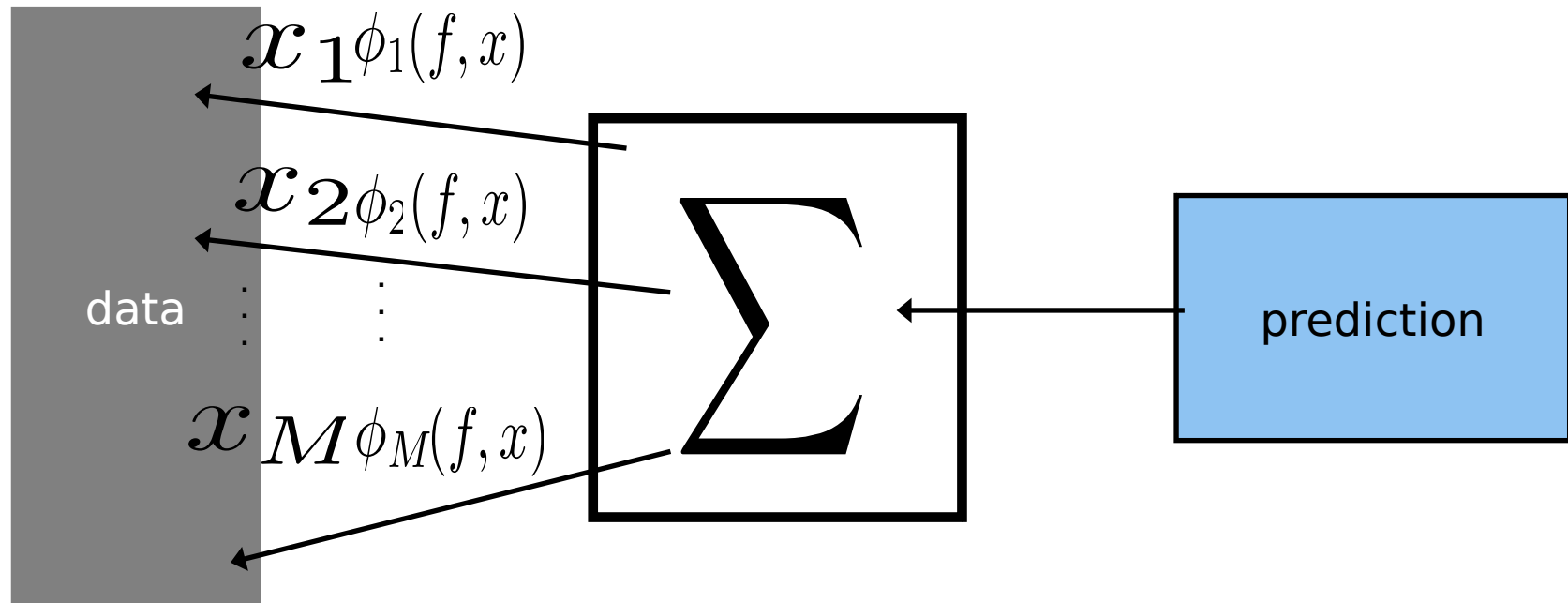
♡ 344 💬 249 people are talking about this



# Complicated AI Model



# Explainable model: Additive feature attribution model



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

where  $z' \in \{0, 1\}^M$ ,  $M$  is the number of simplified input features, and  $\phi_i \in \mathbb{R}$ .

# Additive feature attribution methods

LIME

Ribeiro et al. 2016

DeepLIFT

Shrikumar et al. 2016

Shapley reg.  
values

Lipovetsky et al. 2001

$\Sigma$

Relevance prop.

Bach et al. 2015

QII

Datta et al. 2016

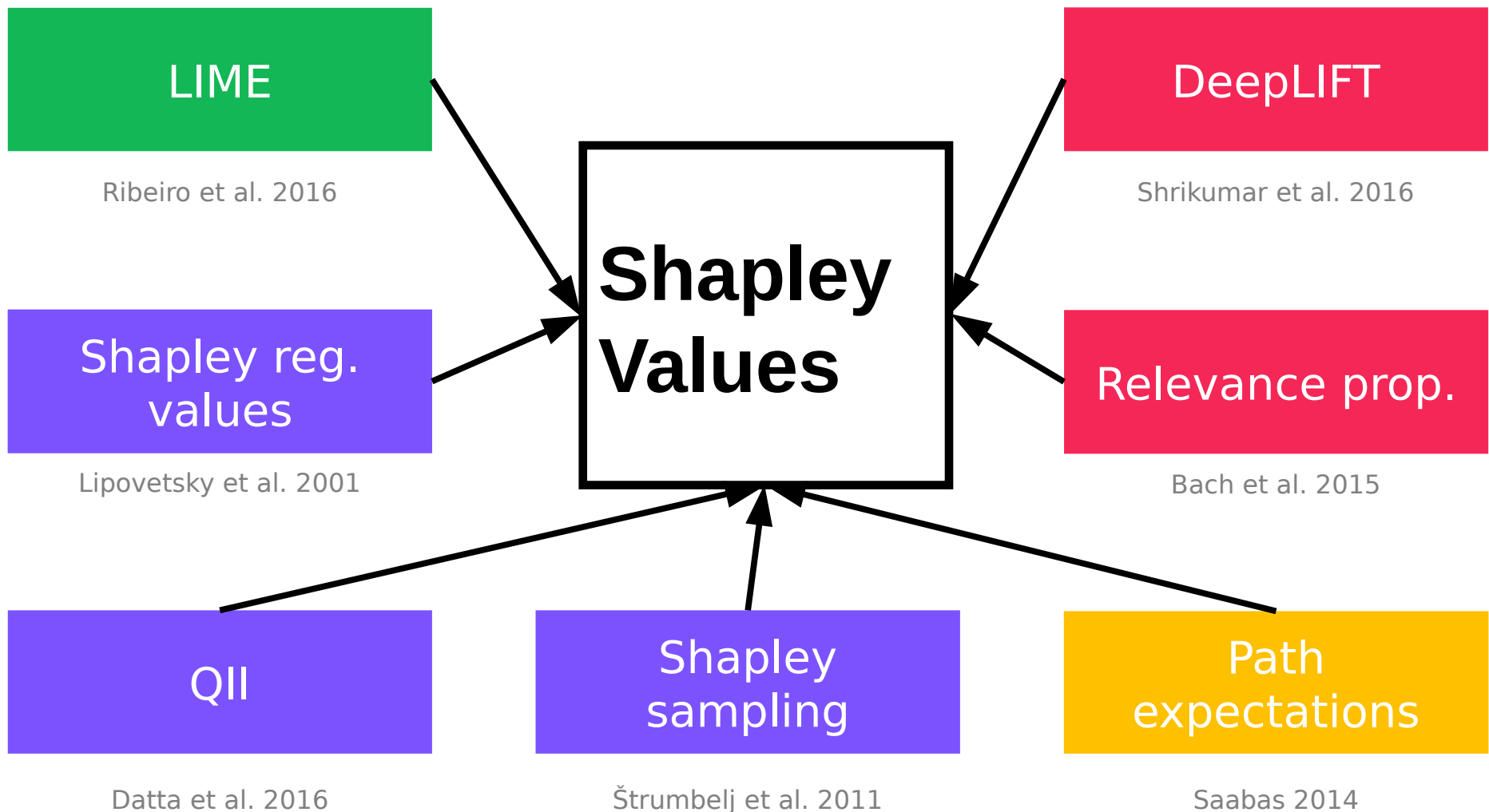
Shapley  
sampling

Štrumbelj et al. 2011

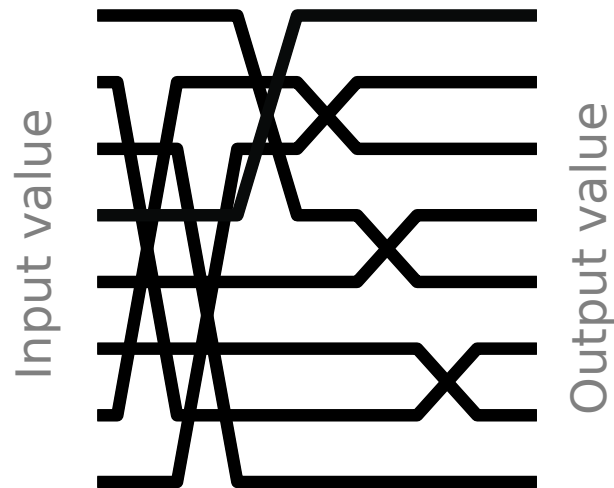
Path  
expectations

Saabas 2014

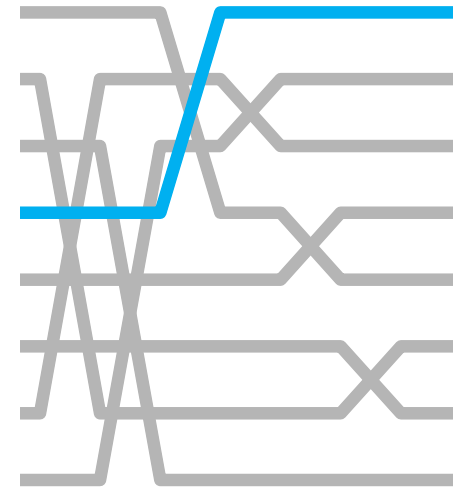
# Additive feature attribution methods



# Why additive feature attribution methods may work



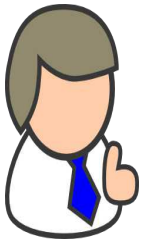
Complex models are inherently complex!



But a single prediction involves only a small piece of that complexity.



# SHapley Additive exPlanation - (SHAP) values (1)



Base rate

Prediction for John

20%

55%

0

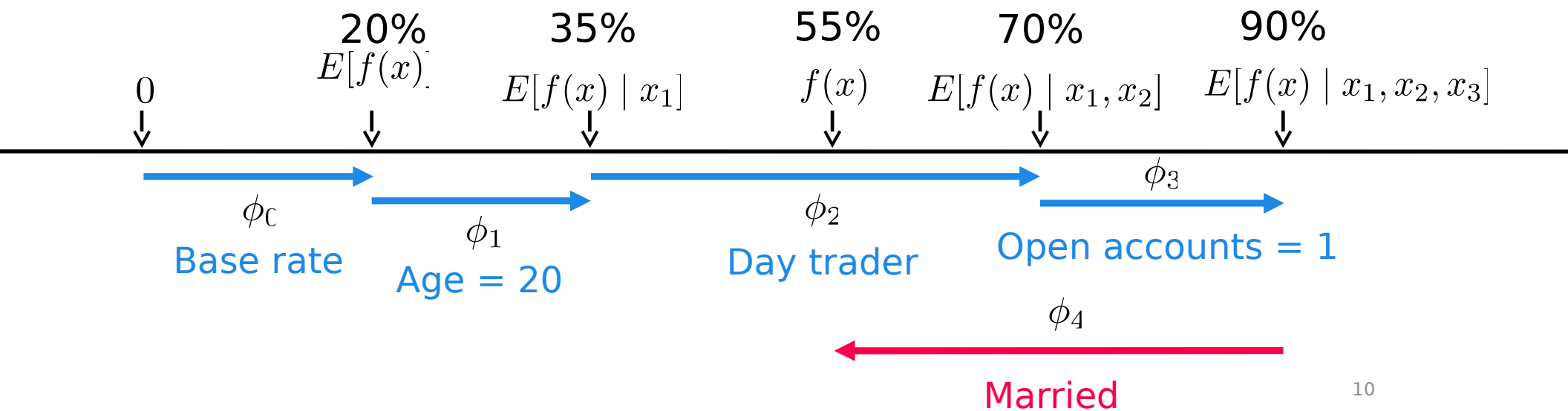
$E[f(x)]$

$f(x)$

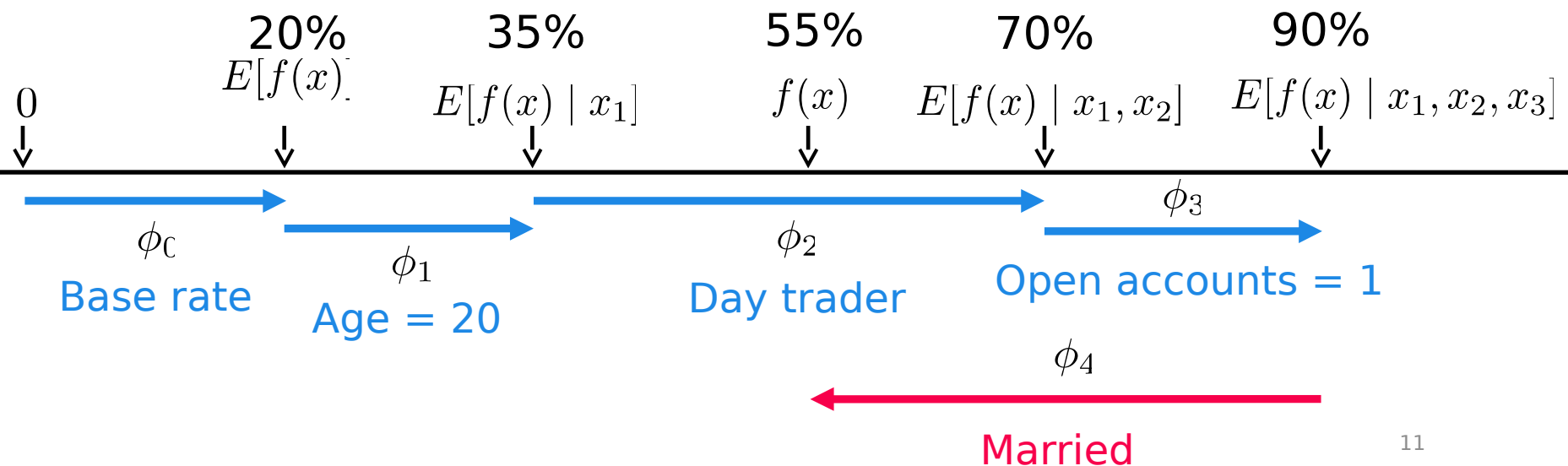


How did we get here?

# SHapley Additive exPlanation (SHAP) values (2)



# SHapley Additive exPlanation (SHAP) values (2)



# SHapley Additive exPlanation (SHAP) values (3) – phi values

$$\text{Explain model} = \sum_{i=1}^{m \text{ features}} \varphi_i X_i$$

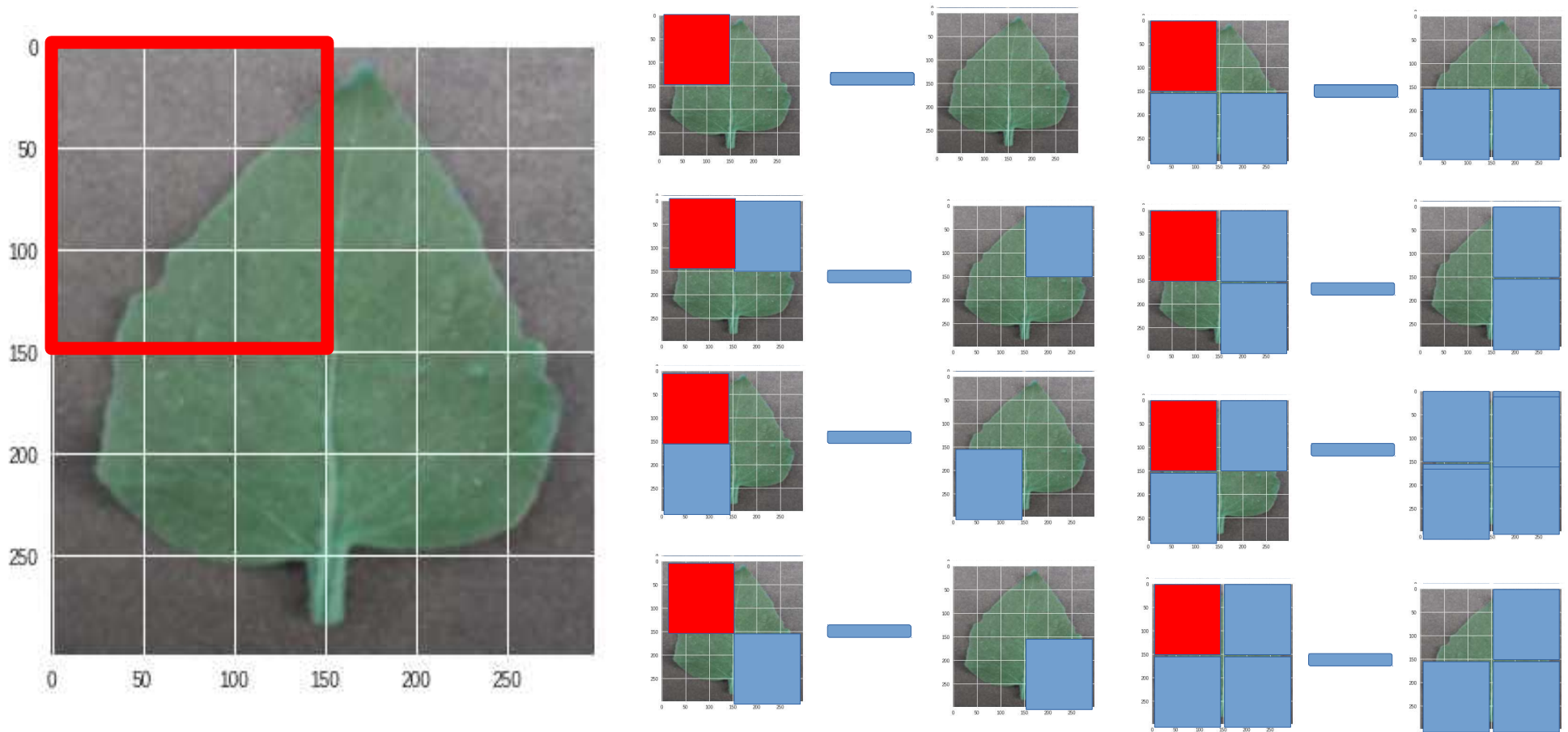
where  $X_i$  an input and  $\varphi_i$  is the effect of  $X_i$  on the model.

$$\varphi_{age} = \langle f(\text{age} \cup \text{features}_{\text{some}}) - f(\text{features}_{\text{some}}) \rangle_{\text{shapley values}}$$

$f$  is your model output, eg accuracy, squared error  
 $\text{features}_{\text{some}}$  is the set containing subset of features

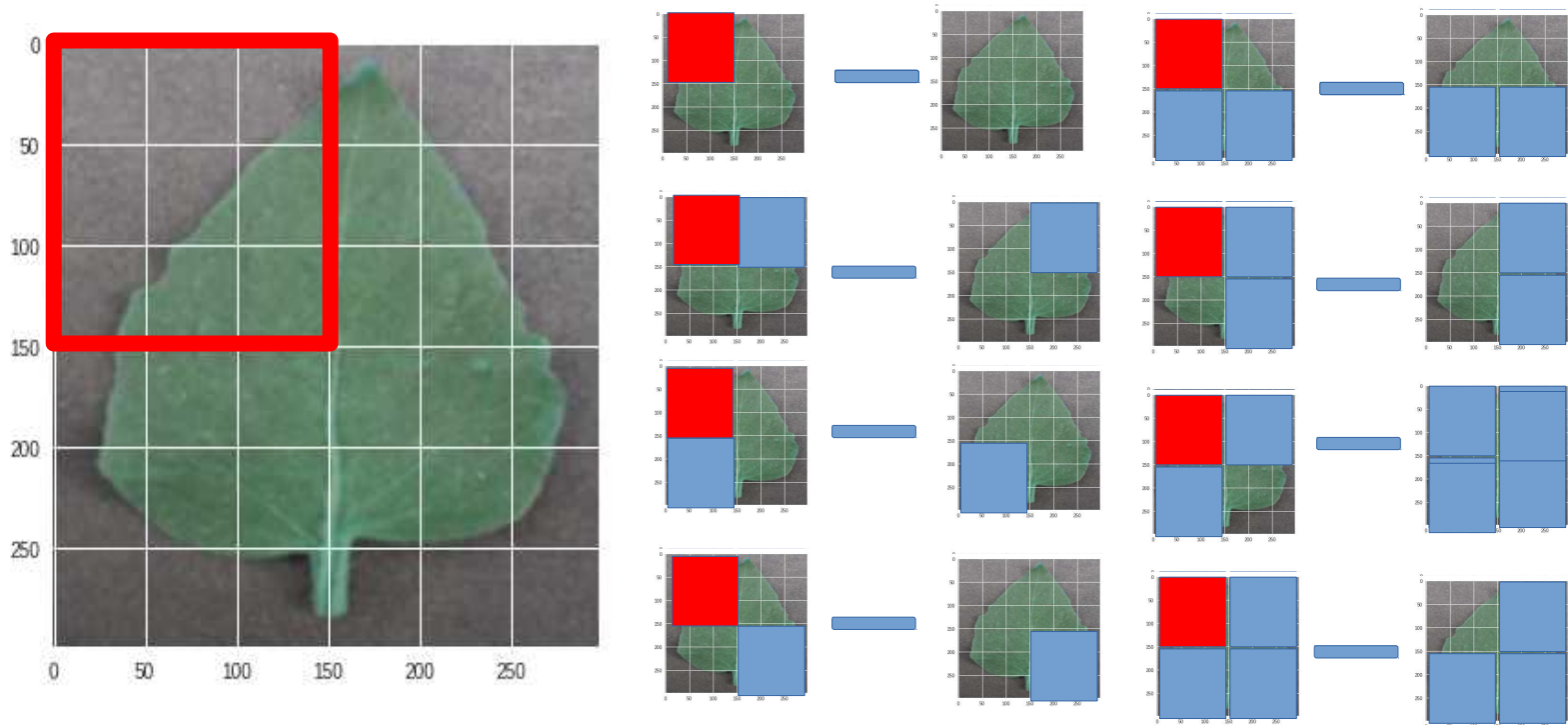
# SHapley Additive exPlanation (SHAP) values (4) – phi values

$$\varphi_{pink} = \langle f(pink \cup features_{some}) - f(features_{some}) \rangle_{shapley\ values}$$



# SHapley Additive exPlanation (SHAP) values (4) – phi values $\phi$

$$\varphi_{\text{pink}} = \text{weight\_avg}(\text{img1}, \text{img2})$$



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

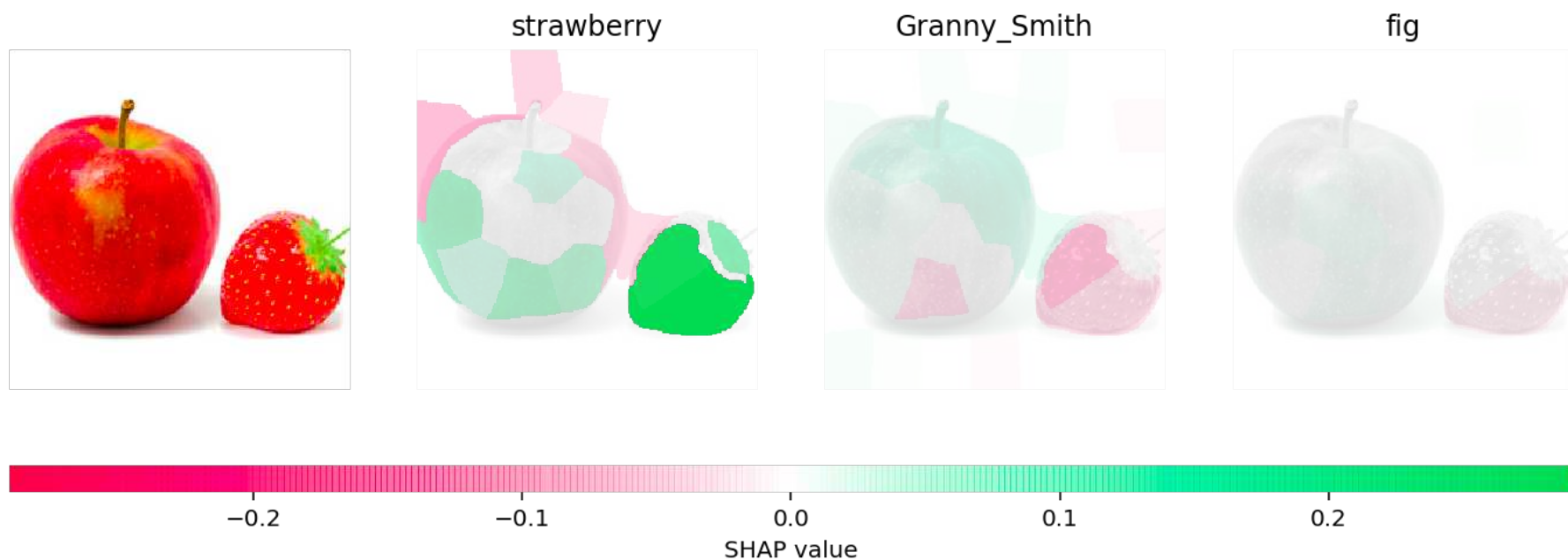
$$\phi = (X^T W X)^{-1} X^T W y$$

$X$  is the feature binary vector of all combinations of  
 $X$

$W$  is weights for each example

$y$  is model output for  $X$

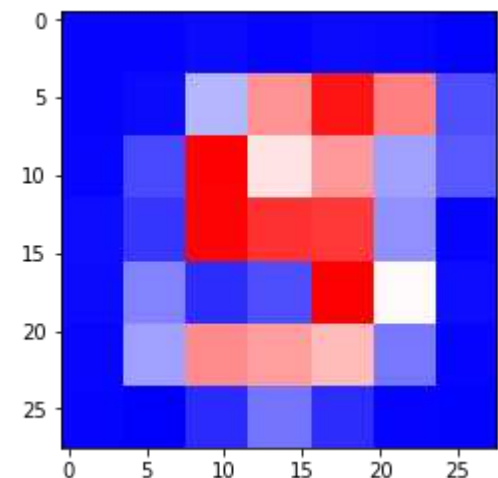
# Another Example : VGG16



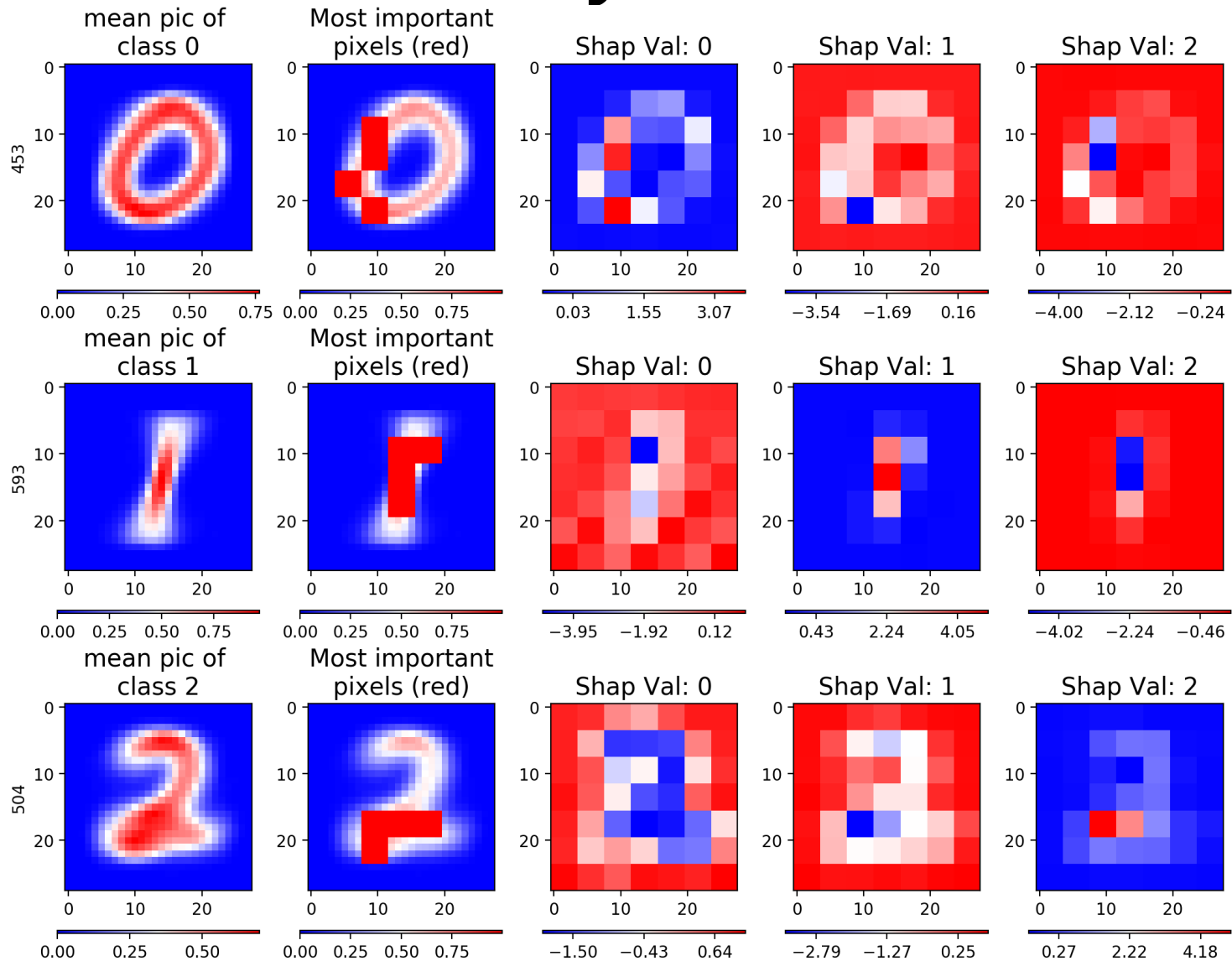


# Applying to Mnist (1)

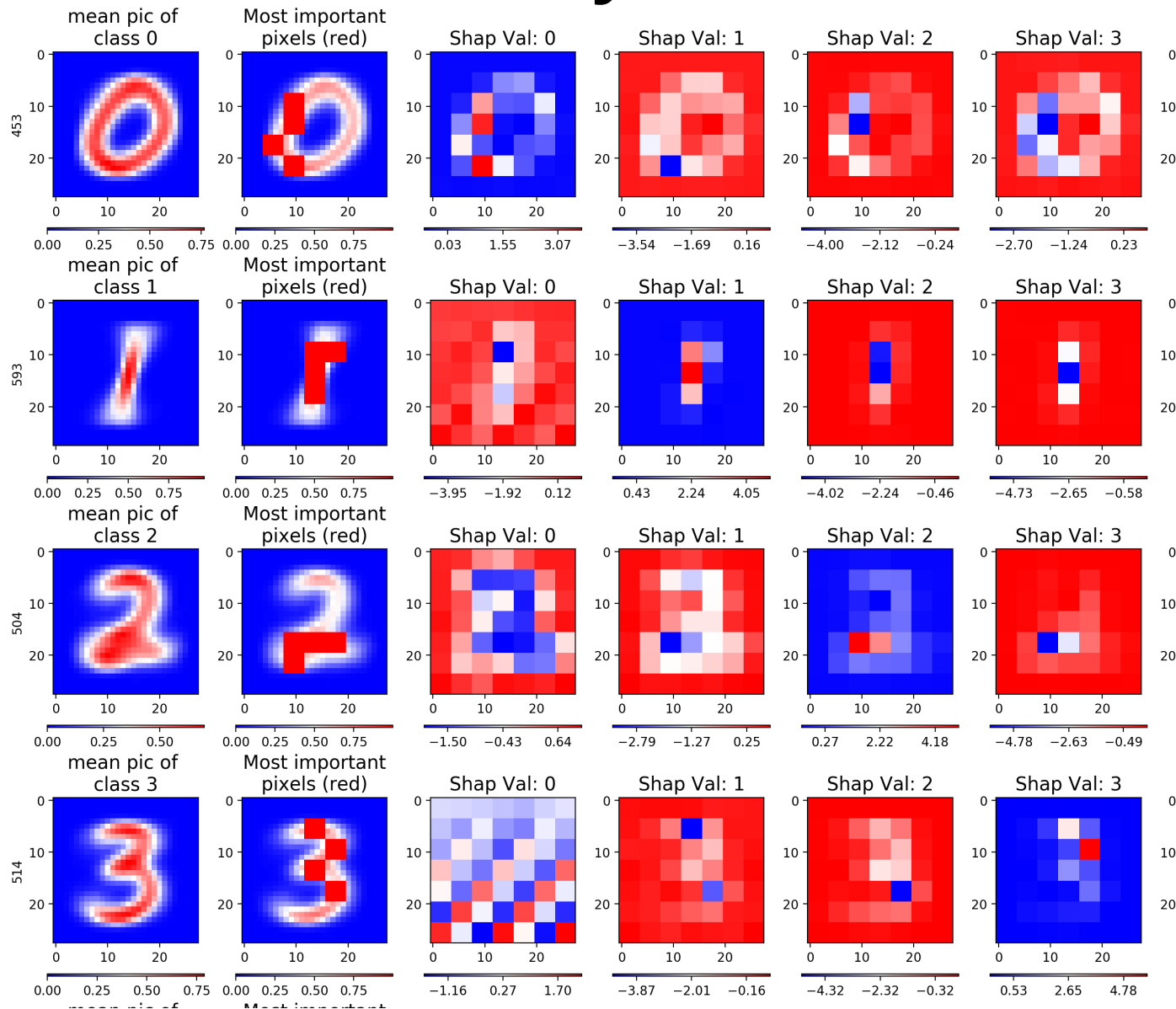
- Mnist model with 4 convolutional layers and 2 dense layers.
- Accuracy is 99.6%
- for each test image {
  - Split image to  $7 \times 7 = 49$  pixels for shapley value computation
  - Sample 7367 combinations of pixels
    - ~ all -1 pixel images,  ${}^{49}C_1 = 49$
    - ~ all -2 pixel images,  ${}^{49}C_2 = 1176$
    - ~ all -3 pixel images,  ${}^{49}C_3 = 6142$
  - Calculate shapley values for each pixel using weighted regression}



# Applying to Mnist (2) – Global analysis



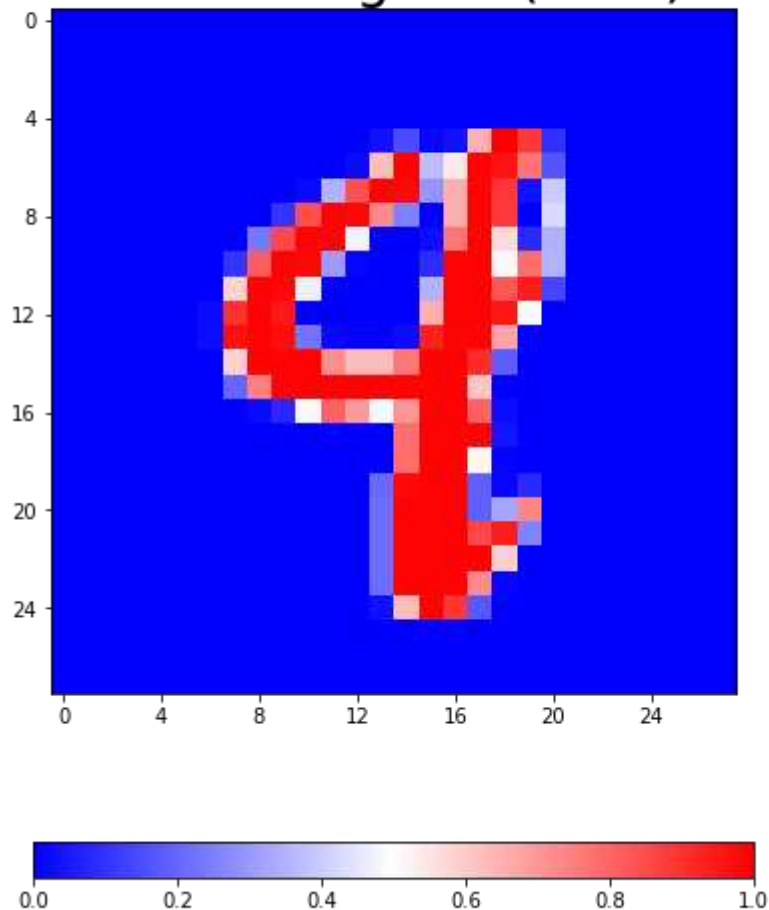
# Applying to Mnist (2) – Global analysis



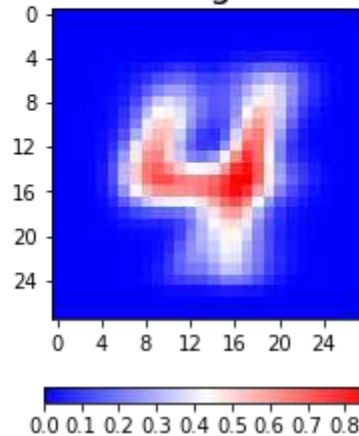
# Applying to Mnist (3) – Individual analysis (a)

Test Example: 359

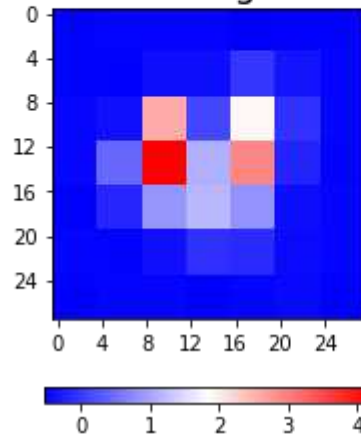
Predicted Digit: 4 (58%),  
Actual Digit: 9 (40%)



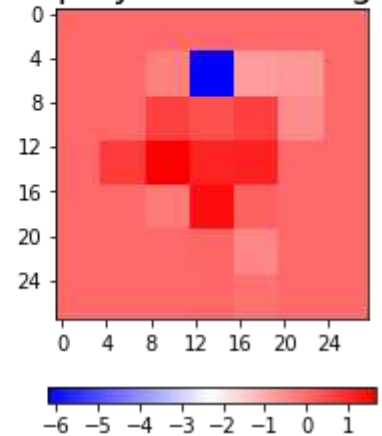
Median pic of  
all digit 4



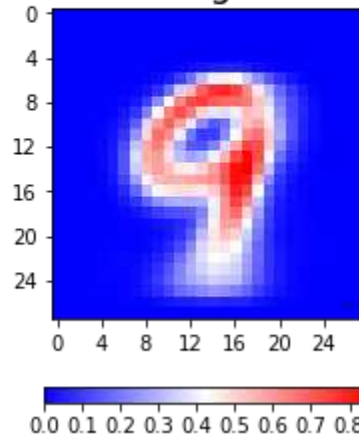
Median shapley value  
of all digit 4



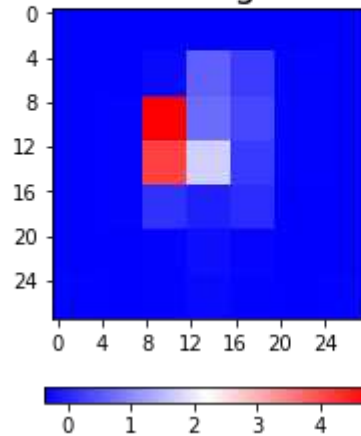
Test Example : 359  
shapley value for digit 4



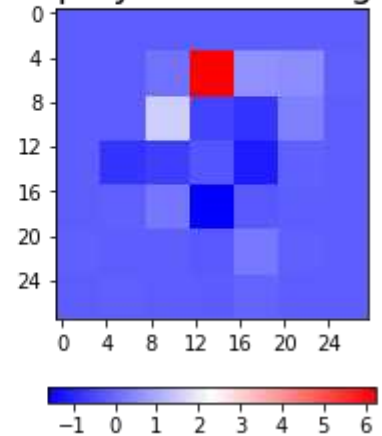
Median pic of  
all digit 9



Median shapley value  
of all digit 9

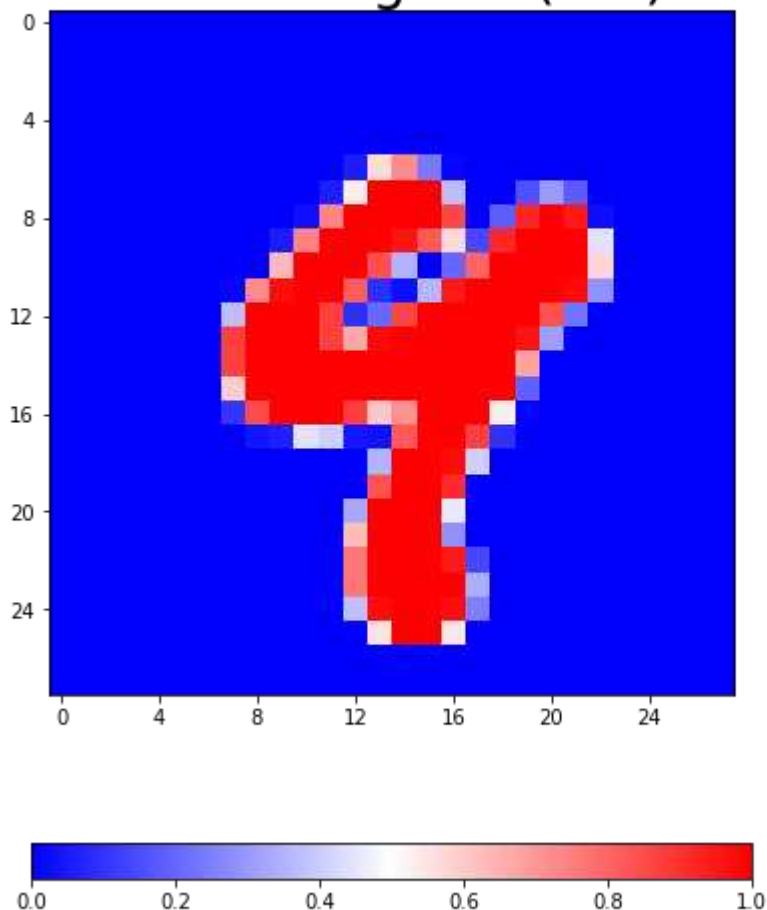


Test Example : 359  
shapley value for digit 9

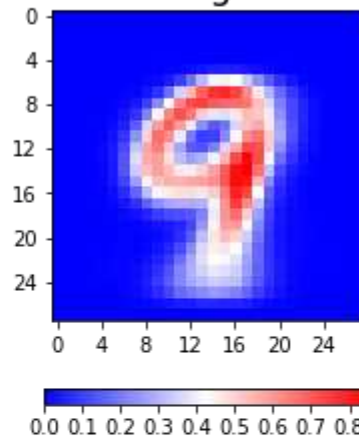


# Applying to Mnist (3) – Individual analysis (b)

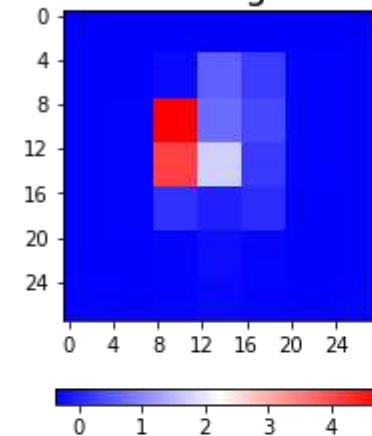
Test Example: 2130  
Predicted Digit: 9 (93%),  
Actual Digit: 4 (6%)



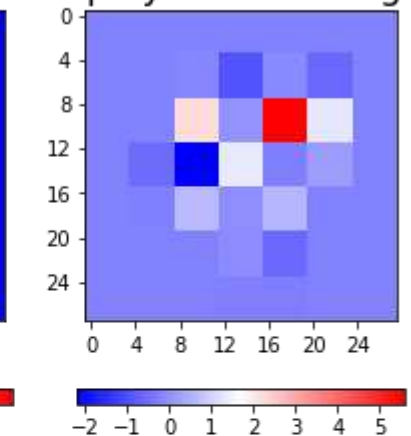
Median pic of  
all digit 9



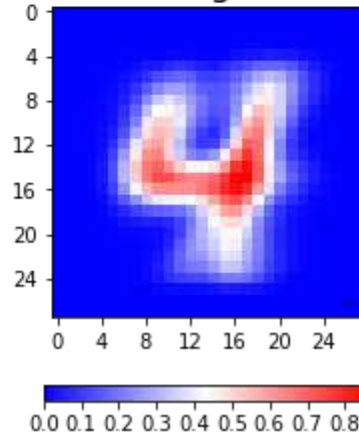
Median shapley value  
of all digit 9



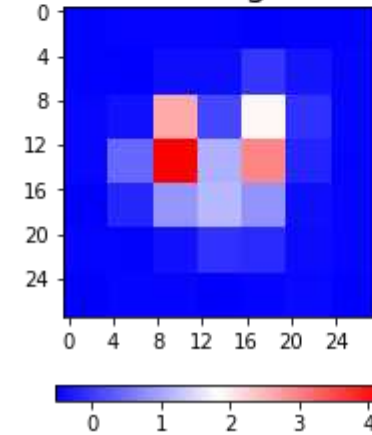
Test Example : 2130  
shapley value for digit 9



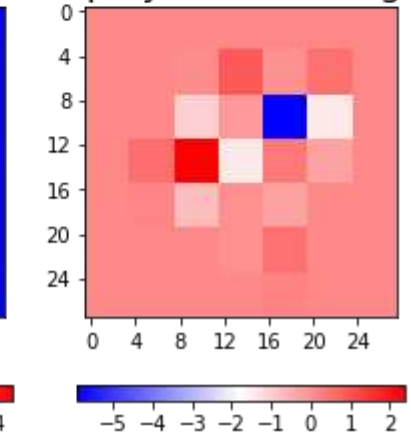
Median pic of  
all digit 4



Median shapley value  
of all digit 4



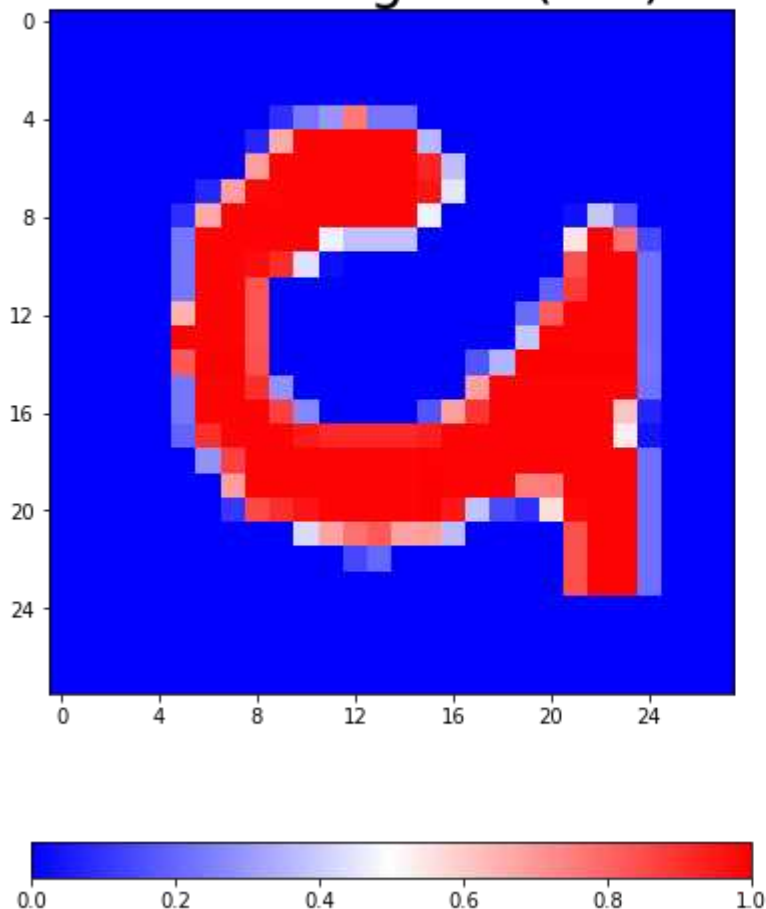
Test Example : 2130  
shapley value for digit 4



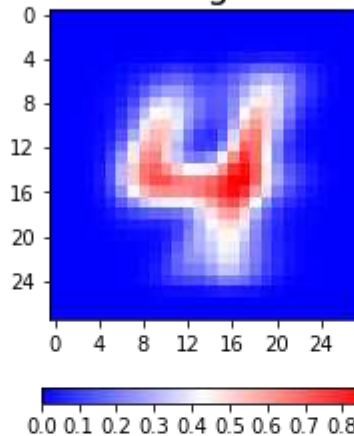
# Applying to Mnist (3) – Individual analysis (c)

Test Example: 2293

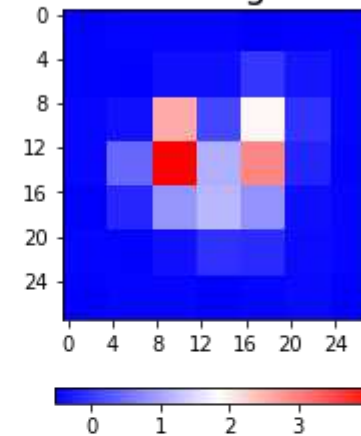
Predicted Digit: 4 (98%),  
Actual Digit: 9 (0%)



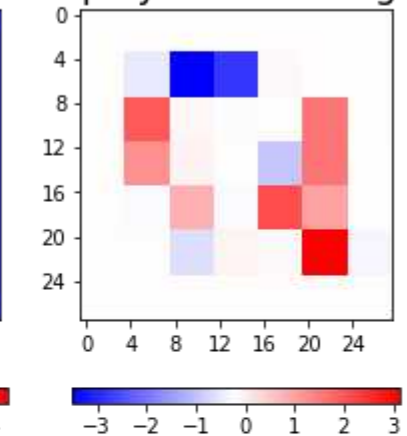
Median pic of  
all digit 4



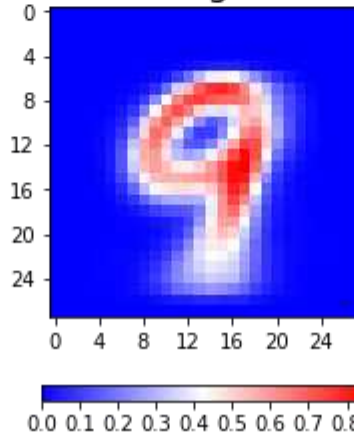
Median shapley value  
of all digit 4



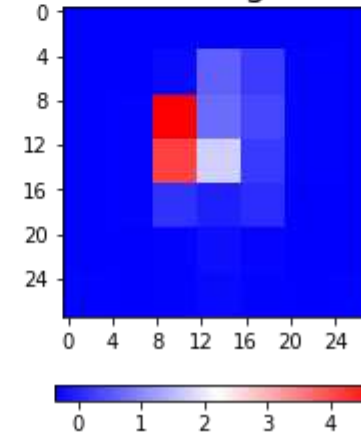
Test Example : 2293  
shapley value for digit 4



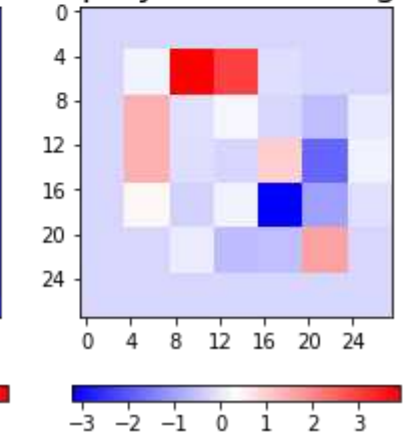
Median pic of  
all digit 9



Median shapley value  
of all digit 9

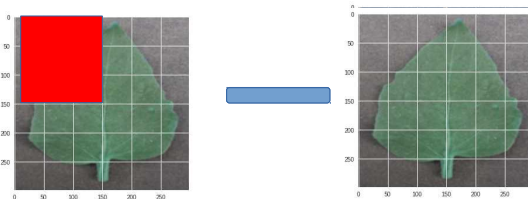


Test Example : 2293  
shapley value for digit 9



# Drawbacks

- Computationally intensive, requires to compute  $2^m$  examples for  $m$  features followed by inverse of  $m \times m$  matrix.  
 $\phi_{pink} = \langle f(pink \cup features_{some}) - f(features_{some}) \rangle_{shapley\ values}$   
 ~ I only sampled  $10^3$  out of  $10^{14}$  combinations
- How do you appropriately remove a feature ?



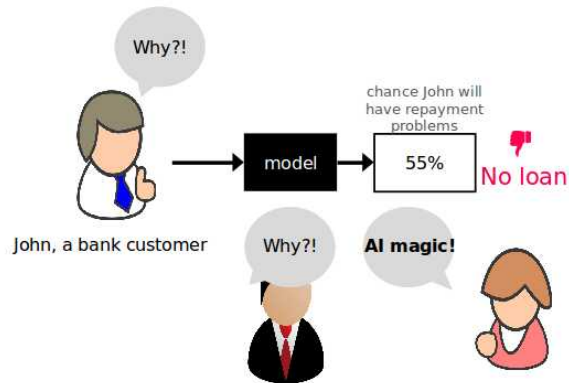
- Method does not explain inner workings, rather it is a model upon a model to explain the final output.



# Summary

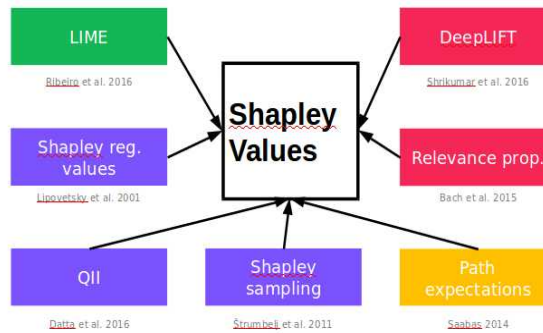
1.

## Need for Explainable AI



2.

## Additive feature attribution methods

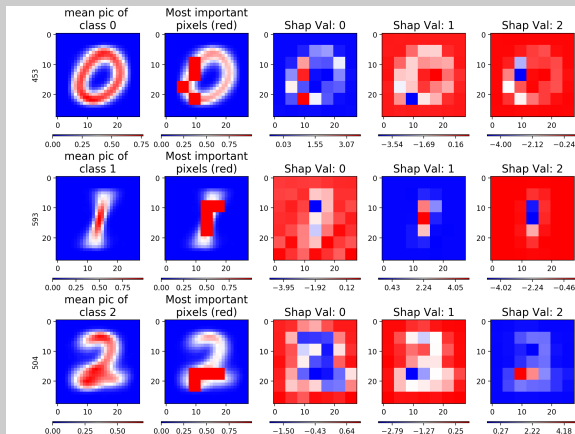


3. Intuition

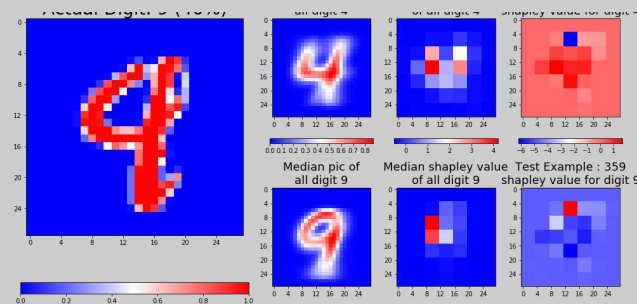
$$g(z') = \sum_{i=1}^M \phi_i z'_i,$$

$$\varphi_{\text{pink}} = \text{weight\_avg}(\text{img}_1 - \text{img}_2)$$

## 4. Analysis of global predictions



## 5. Analysis of each prediction

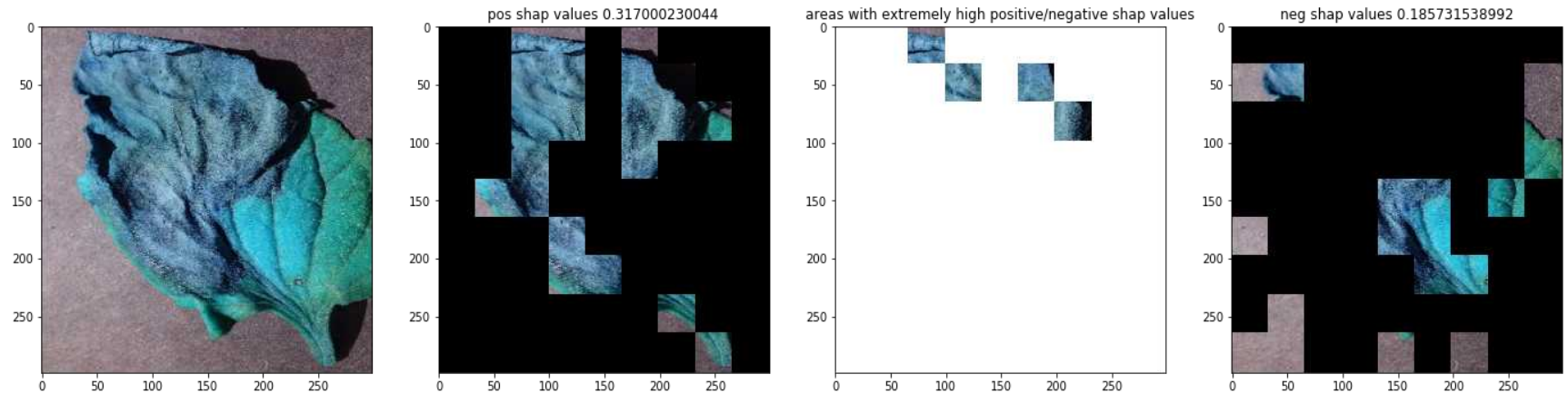


6. Drawbacks





# Another application : Transfer-learned Inception3 model



# 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