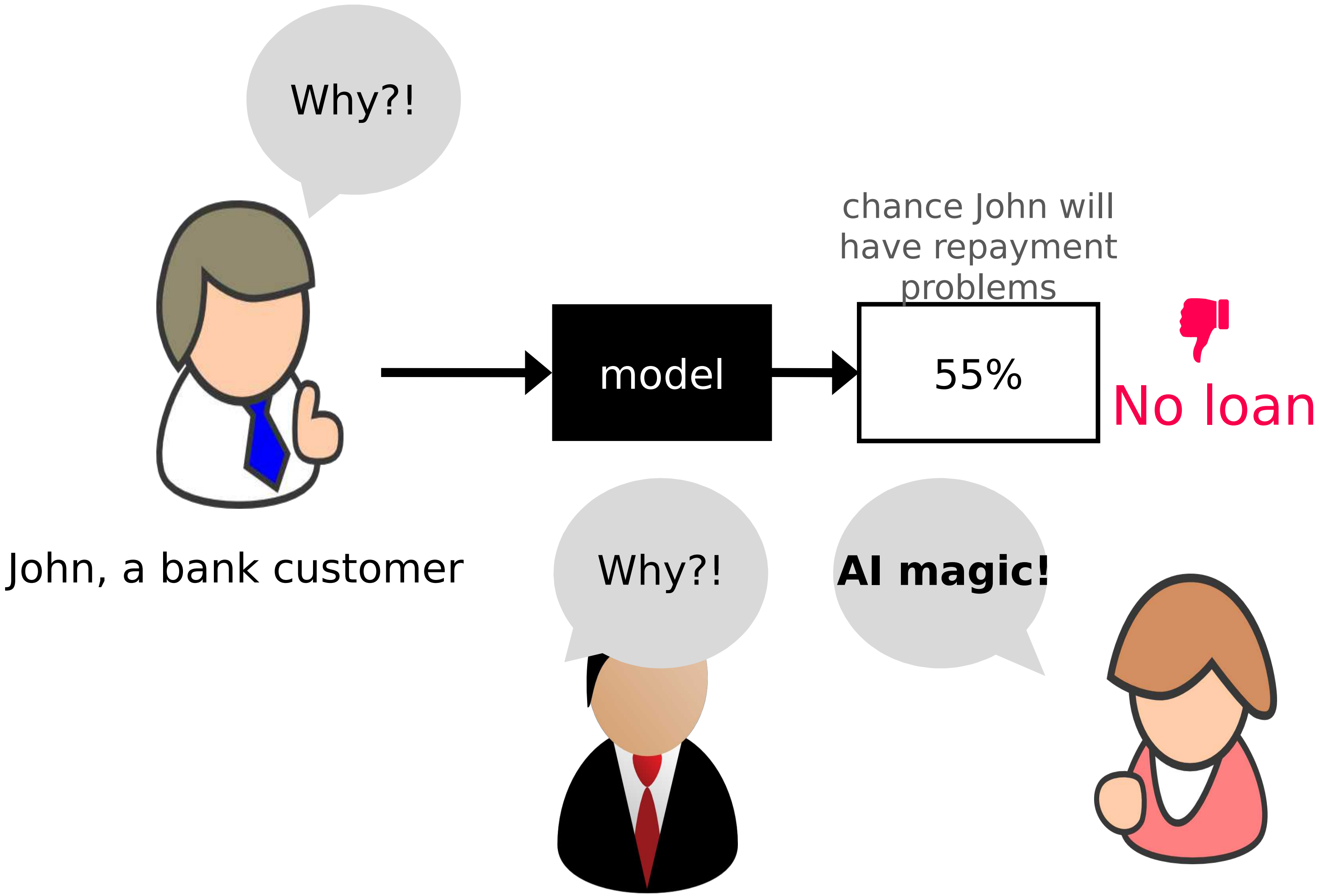
Explainable AI : Shapley Values

A Unified Approach to Interpreting Model Predictions

Scott Lundberg, Su-In Lee

Need for Explainable AI



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

Need for Explainable AI

Some of the articles of GDPR can 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

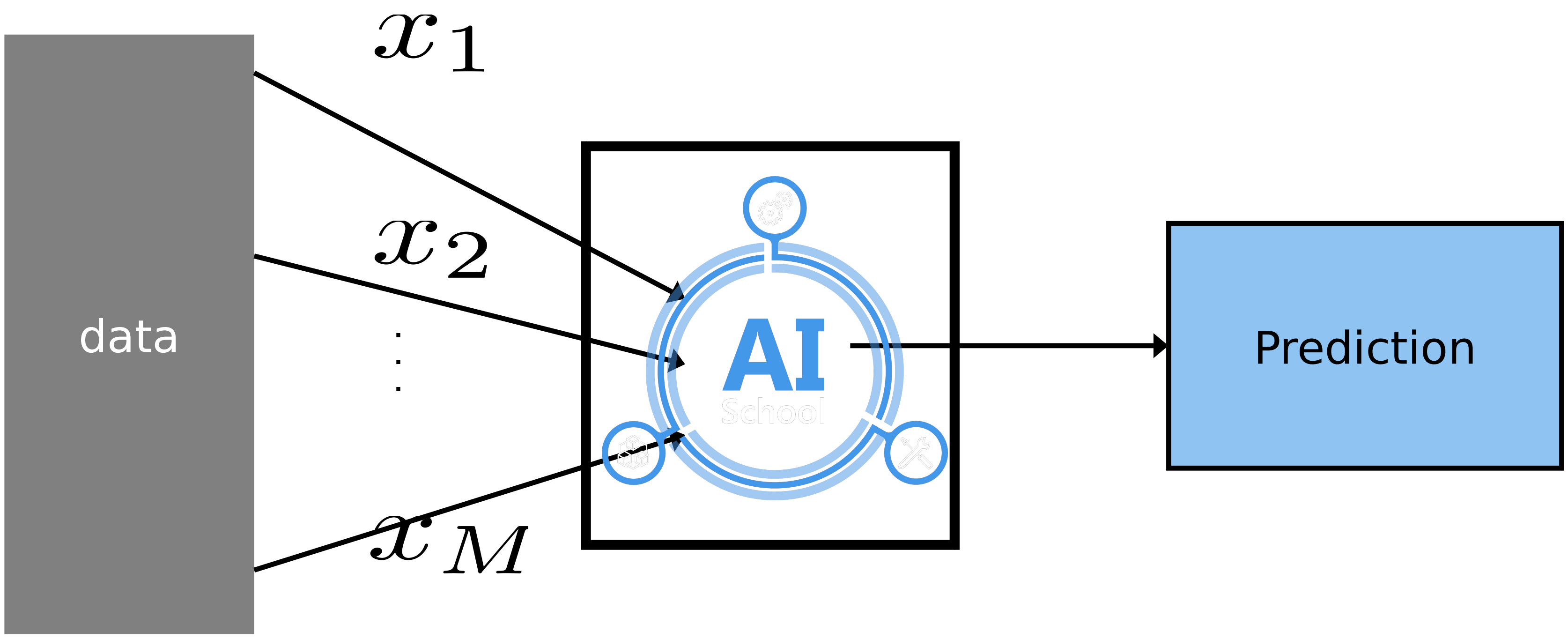
Pedro Domingos
@pmddomingos

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

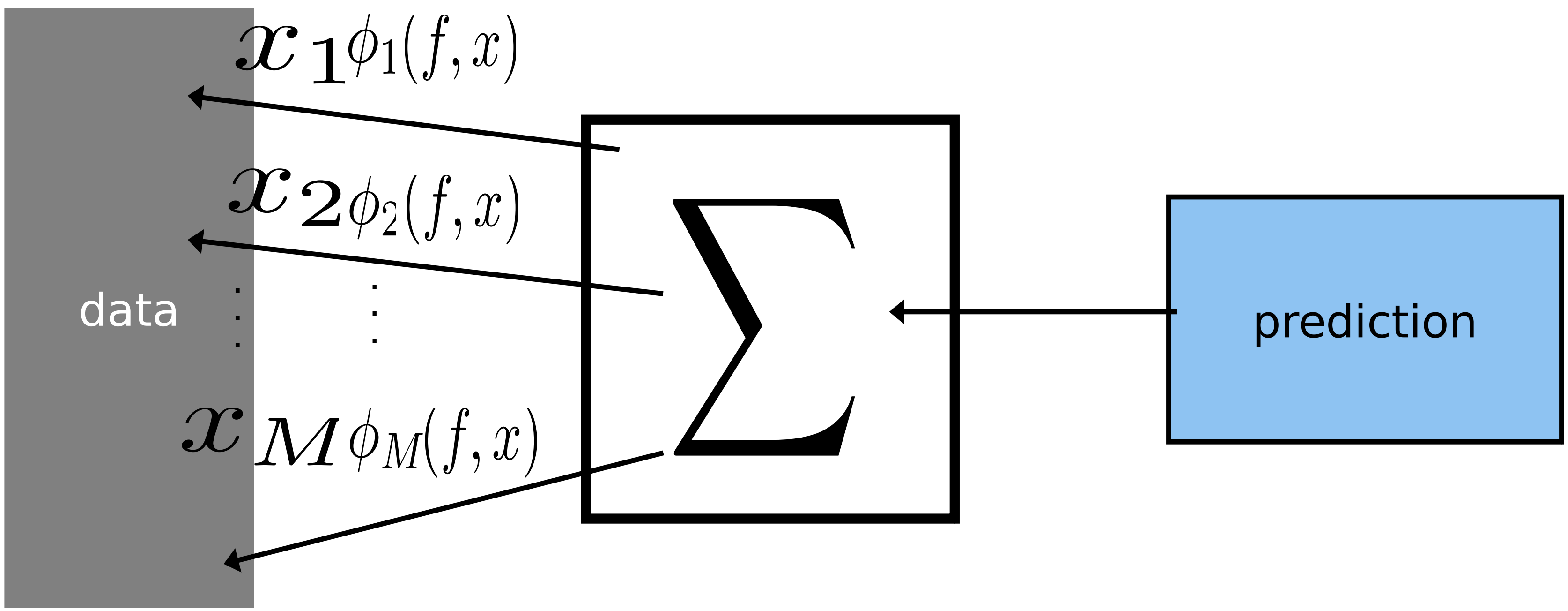
11:59 AM - Jan 29, 2018

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

Σ

Shapley reg. values

Lipovetsky et al. 2001

Relevance prop.

Bach et al. 2015

QII

Datta et al. 2016

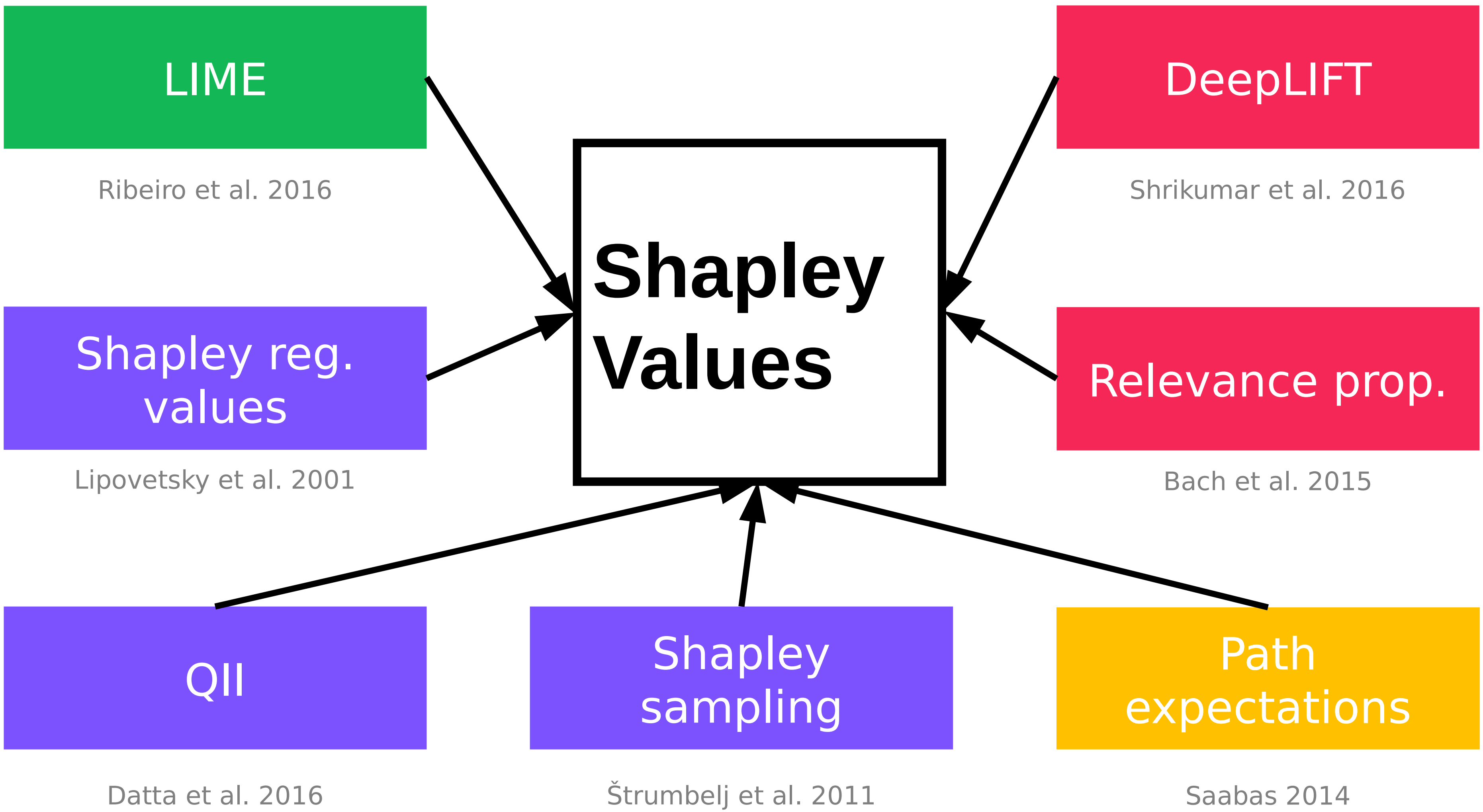
Shapley sampling

Štrumbelj et al. 2011

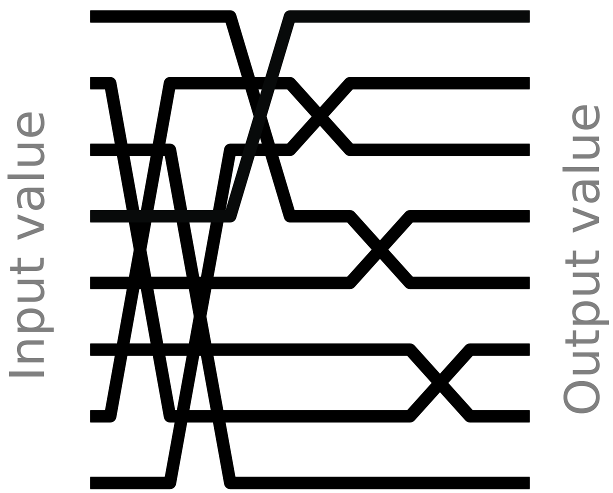
Path expectations

Saabas 2014

Additive feature attribution methods



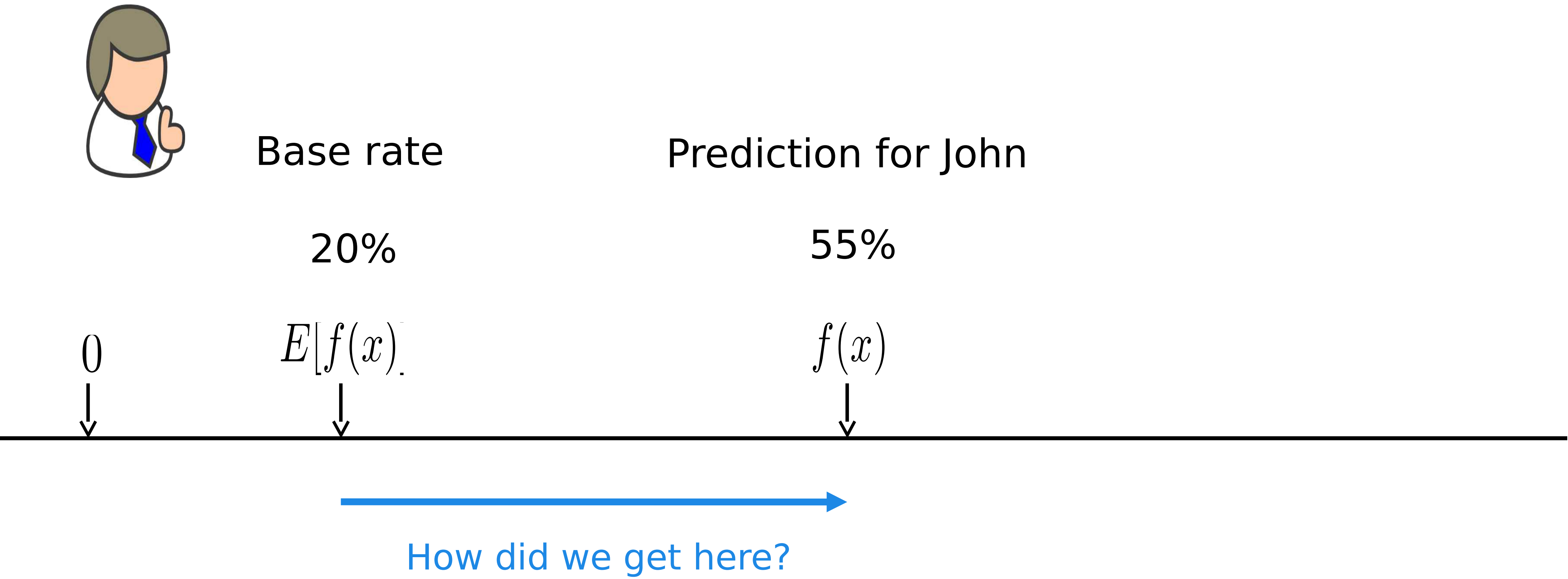
Why additive feature attribution methods may work



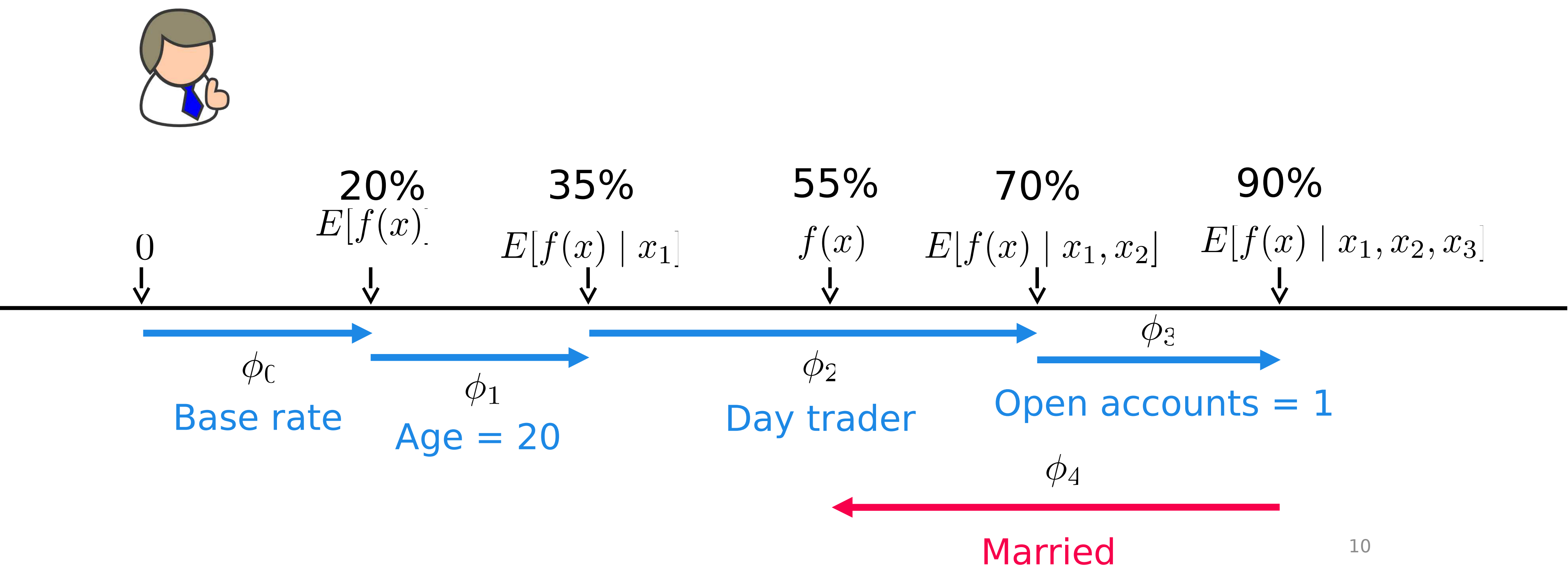
Complex models are
inherently complex!

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

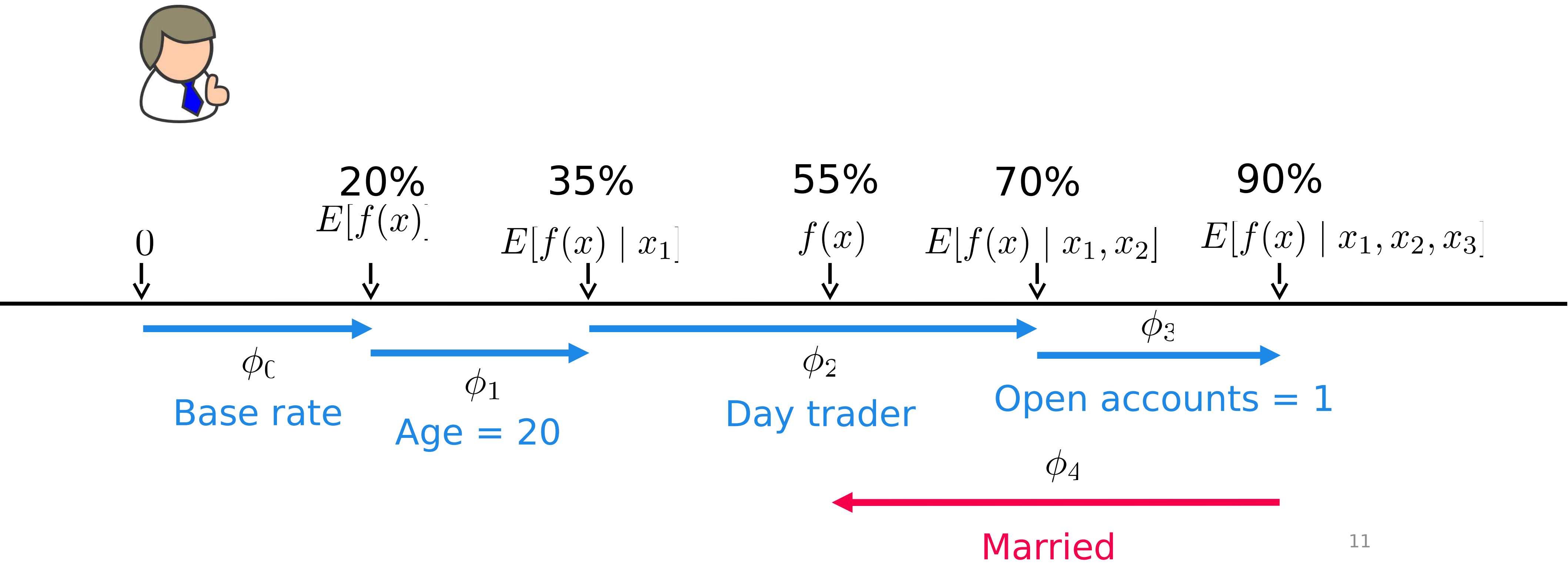
SHapley Additive exPlanation - (SHAP) values (1)



SHapley Additive exPlanation (SHAP) values (2)



SHapley Additive exPlanation (SHAP) values (2)



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

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

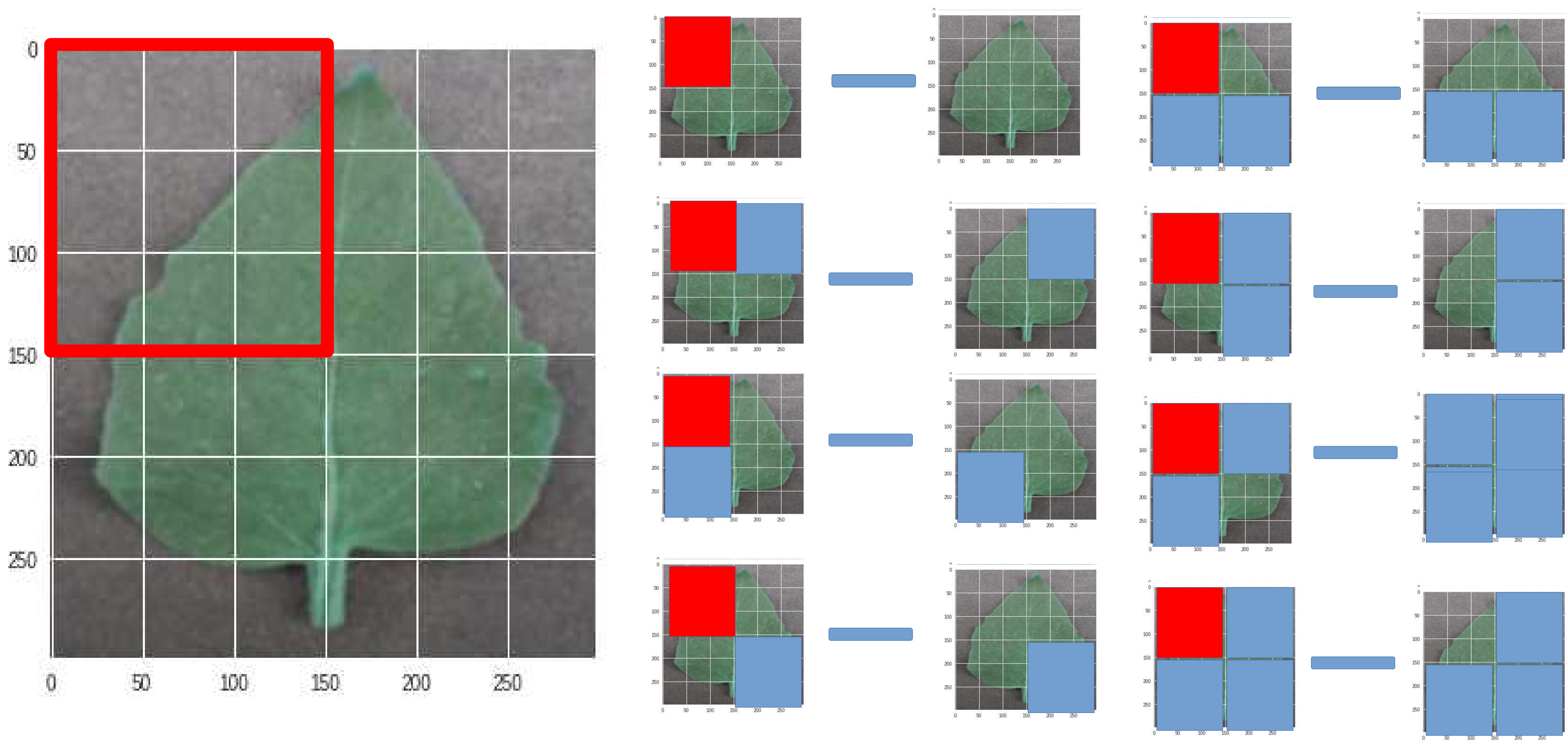
where X_i an input and φ_i is the effect of X_i on the model.

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

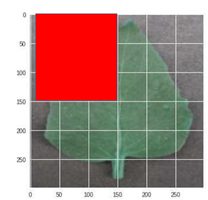
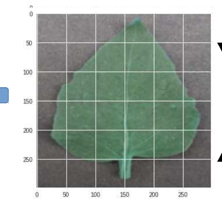
f is your model output, eg accuracy, squared error
$features_{some}$ is the set containing subset of features

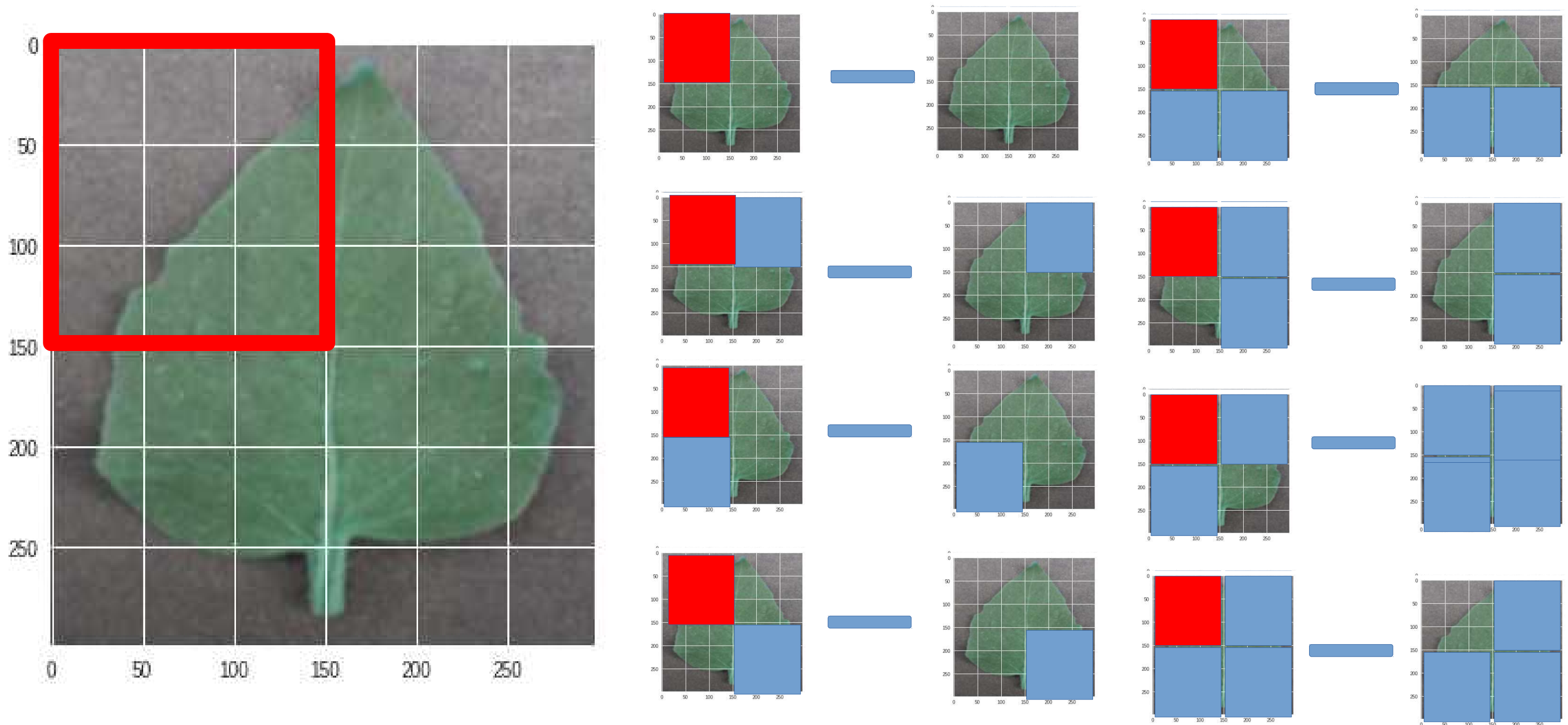
SHapley Additive exPlanation (SHAP) values (4) – phi values

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

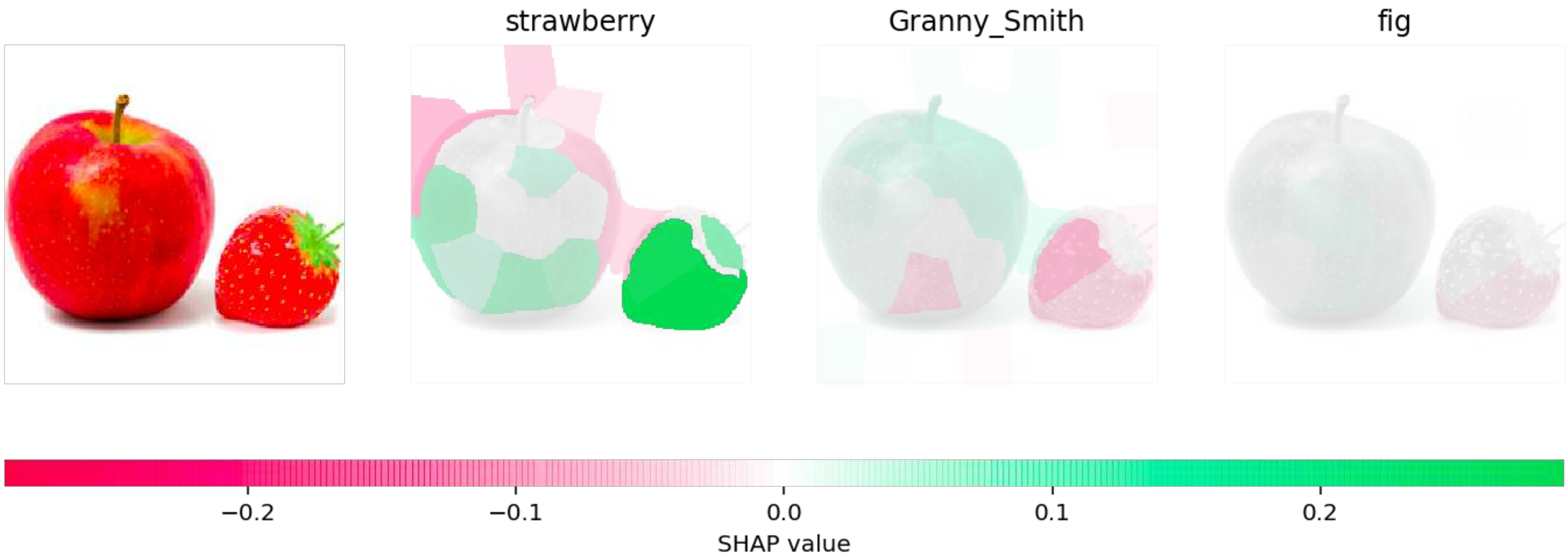


SHapley Additive exPlanation (SHAP) values (4) – phi values φ

φ_{pink} = weight_avg( )



Another Example : VGG16



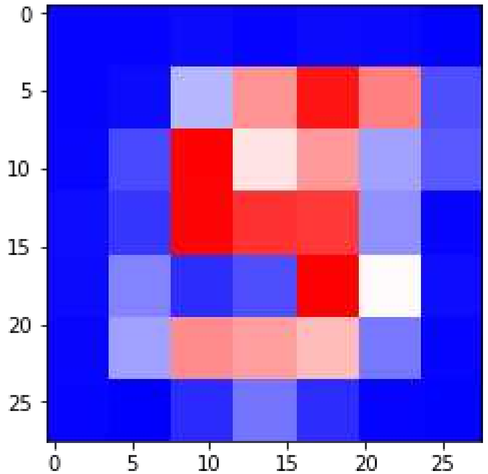
SHapley Additive exPlanation (SHAP) values (5) – solved using weighted linear regression

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

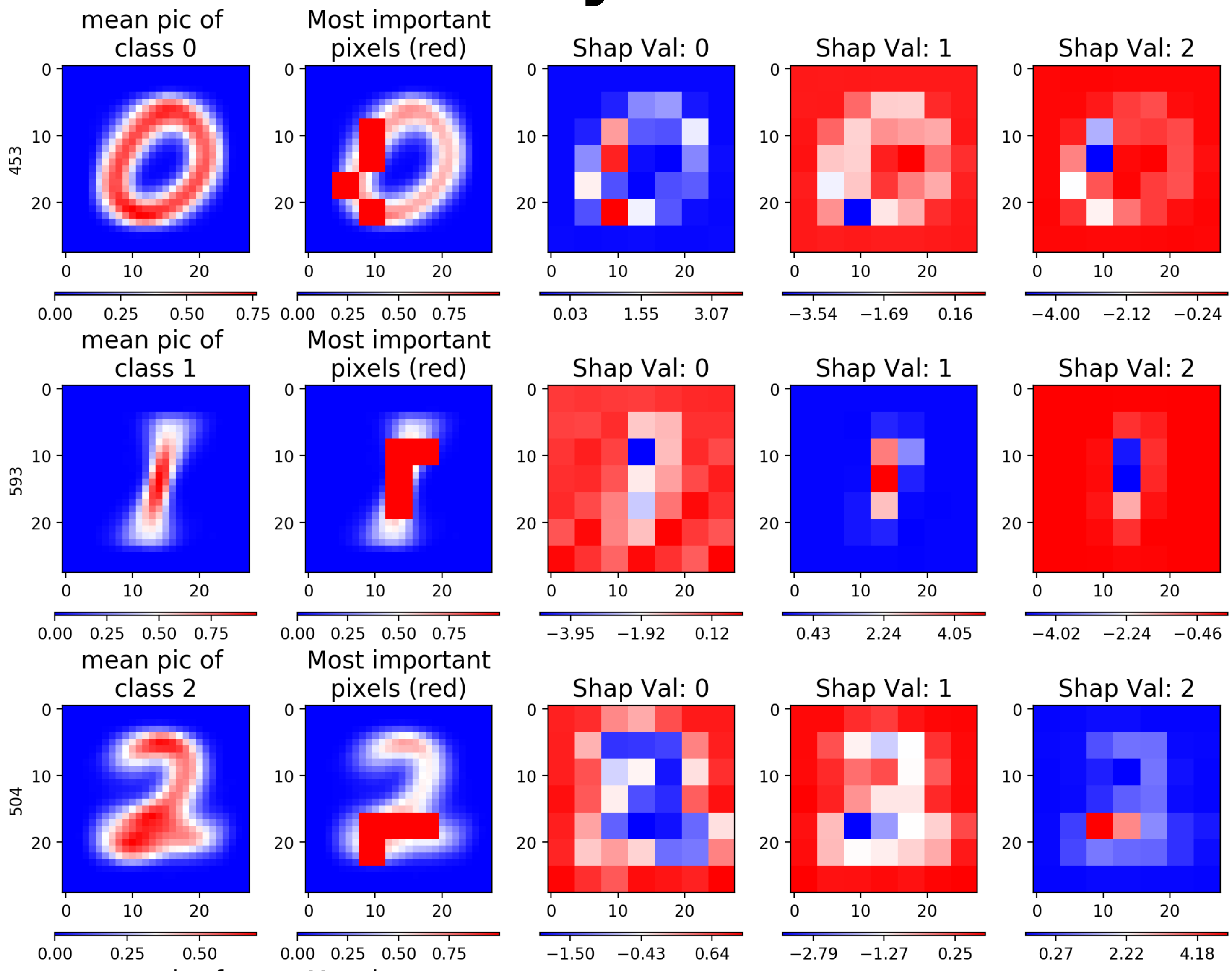
X is the feature binary vector of all combinations of X
W is weights for each example
y is model output for X

Applying to Mnist (1)

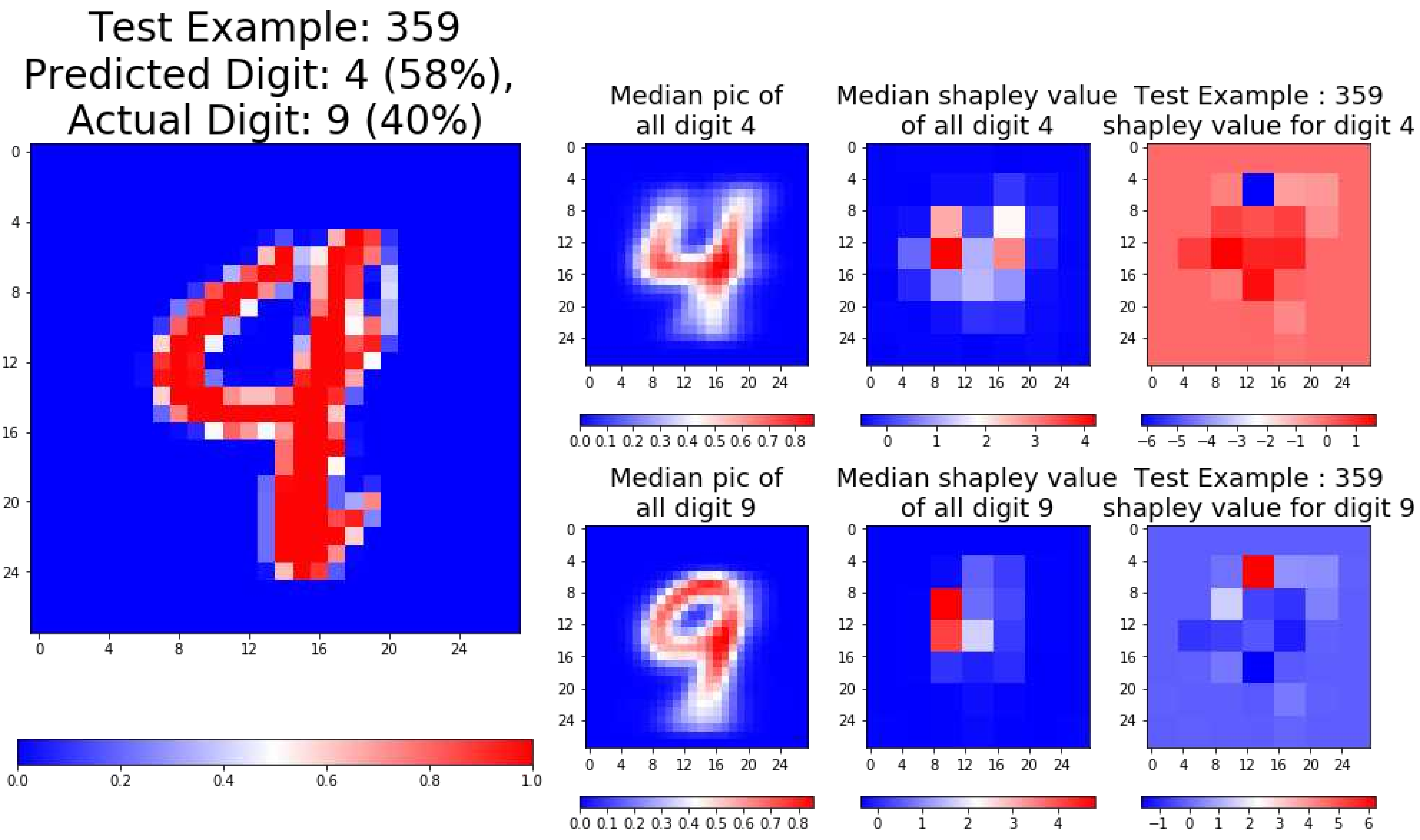
- Mnist model with 4 convolutional layers and 2 dense layers.
- Accuracy is 99.6%
- Split image to 7*7 pixels for shapley computation
- Sampled 7367 combinations for each test image
  ~ all -1 pixel images, ${}^{49}C_1 = 49$
  ~ all -2 pixel images, ${}^{49}C_2 = 1176$
  ~ all -3 pixel images, ${}^{49}C_3 = 6142$



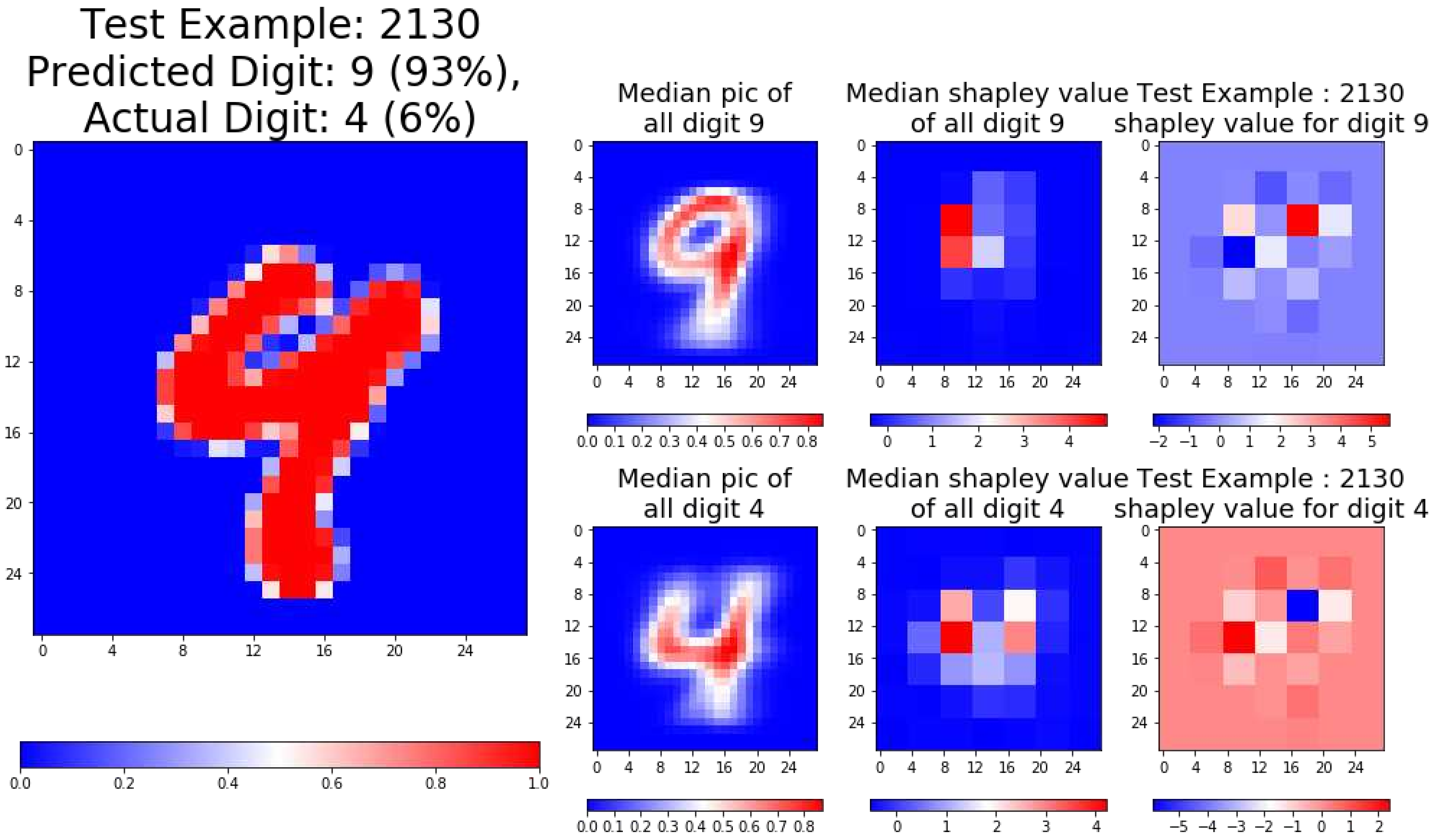
Applying to Mnist (2) – Global analysis



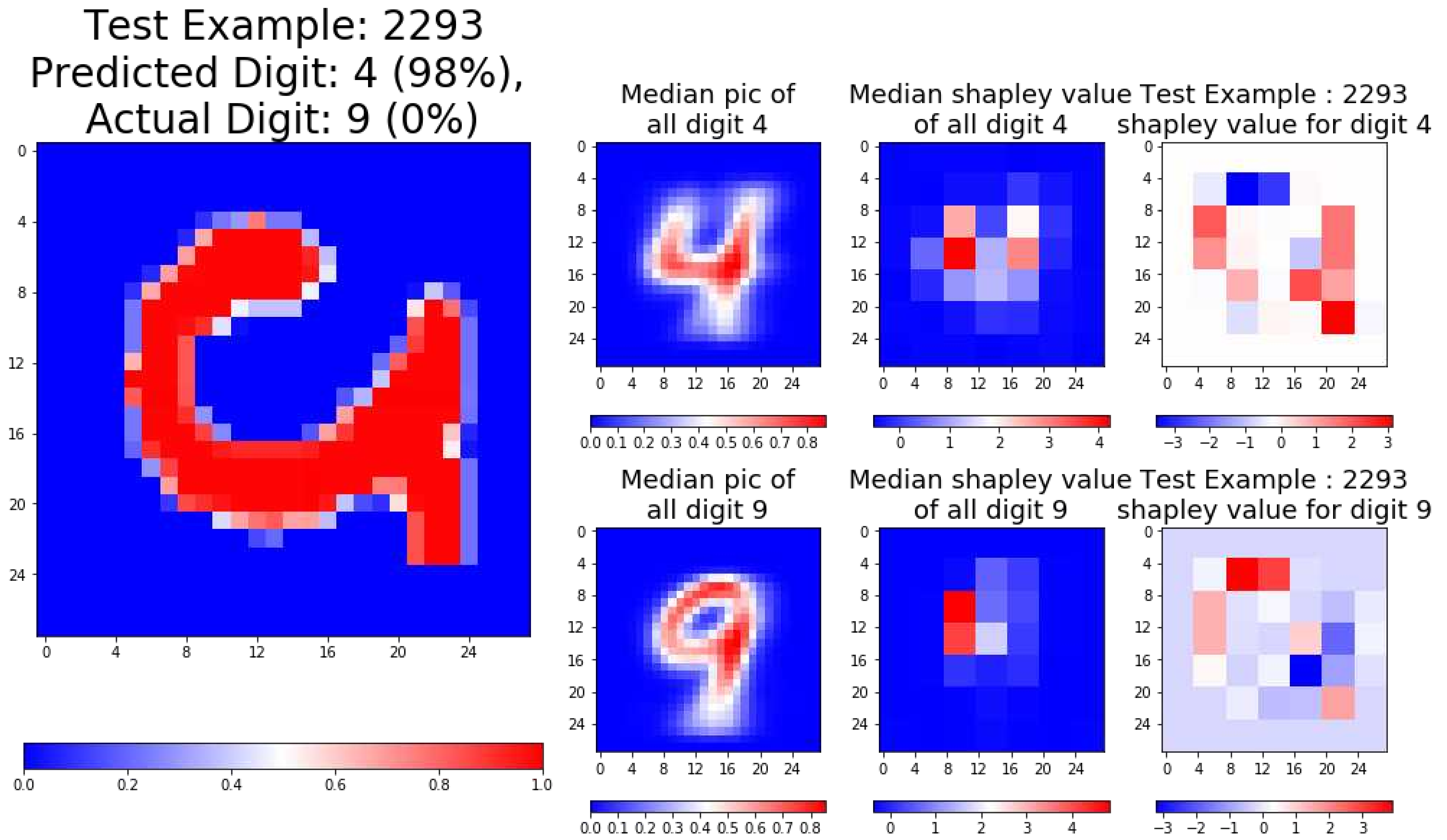
Applying to Mnist (3) – Individual analysis (a)



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

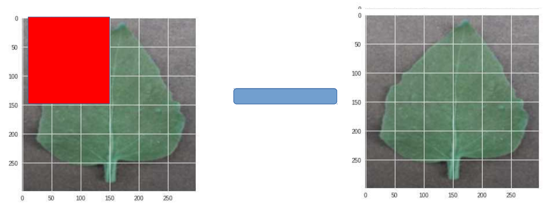


Applying to Mnist (3) – Individual analysis (c)



Drawbacks

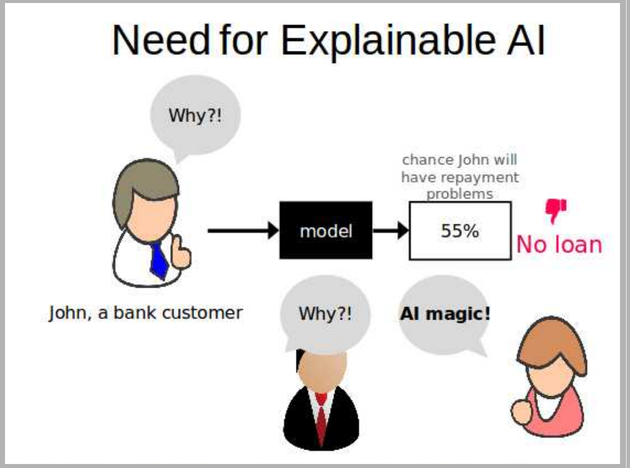
- Computationally intensive, requires to compute 2^m examples for m features followed by inverse of m*m matrix.
  ~ I only sampled 10^3 out of 10^{14} combinations
- How do you appropriately remove a feature ?



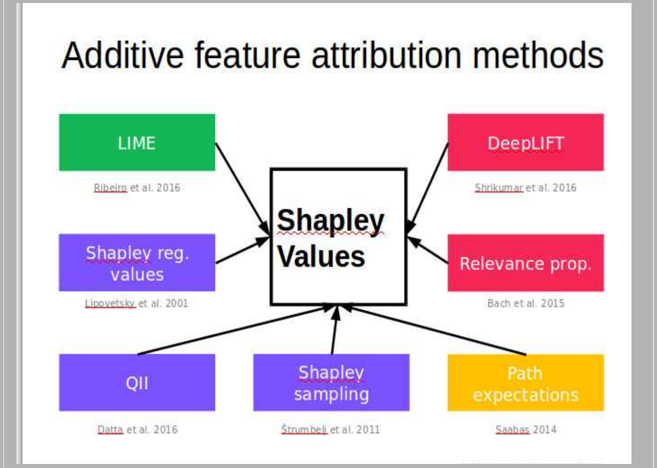
- Method does not explain inner workings, rather it is a model upon a model to explain the final output.

Summary

1.

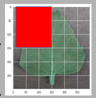
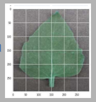


2.

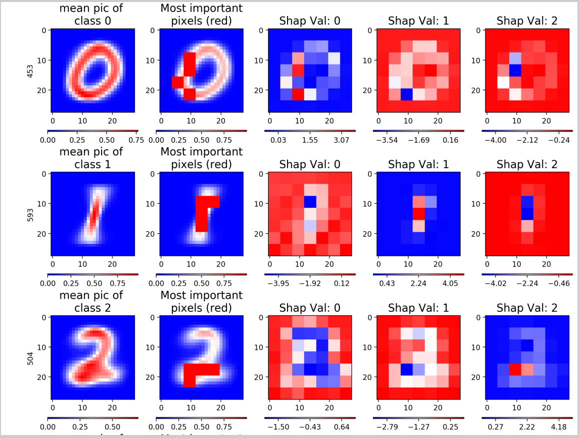


3. Intuition

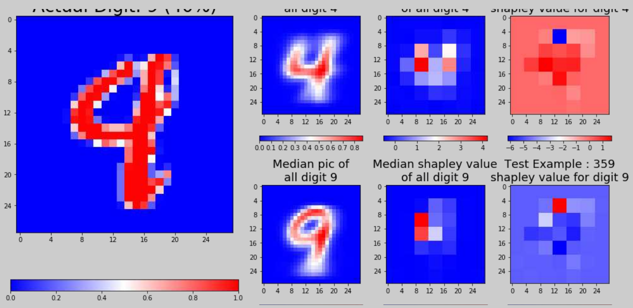
$$g(z') = \sum_{i=1}^{M} \phi_i z'_i,$$

φ_{pink} = weight_avg( − )

4. Analysis of global predictions



5. Analysis of each prediction



6. Drawbacks

Another application : Transfer-learned Inception3 model