

# Explainable AI : Shapley Values and other stuff

A Unified Approach to Interpreting Model Predictions

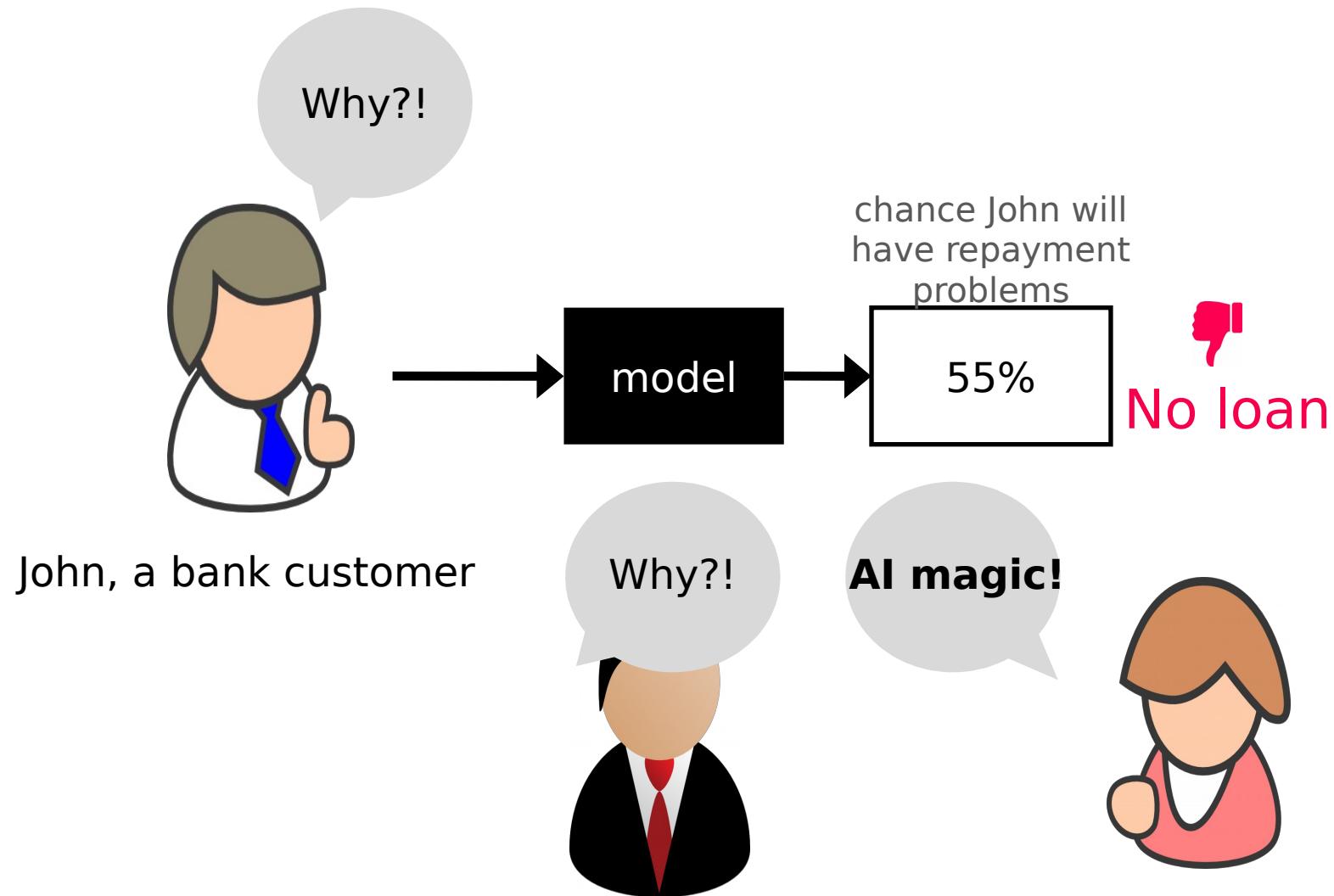
**Scott Lundberg, Su-In Lee**

[https://colab.research.google.com/github/leexxa90/Explainable\\_AI\\_image\\_classification/blob/master/colabs\\_script.ipynb](https://colab.research.google.com/github/leexxa90/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, now in Crop analytics

# Need for Explainable AI



# 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

A screenshot of a Twitter post from user @pmddomingos. The post features a profile picture of Pedro Domingos, the name "Pedro Domingos" in bold, and the handle "@pmddomingos". To the right of the handle is a blue Twitter bird icon. The tweet itself reads: "Starting May 25, the European Union will require algorithms to explain their output, making deep learning illegal." Below the tweet is the timestamp "11:59 AM - Jan 29, 2018". At the bottom left are engagement metrics: "344" with a heart icon and "249" with a speech bubble icon, followed by the text "people are talking about this". On the far right is a small grey info icon.

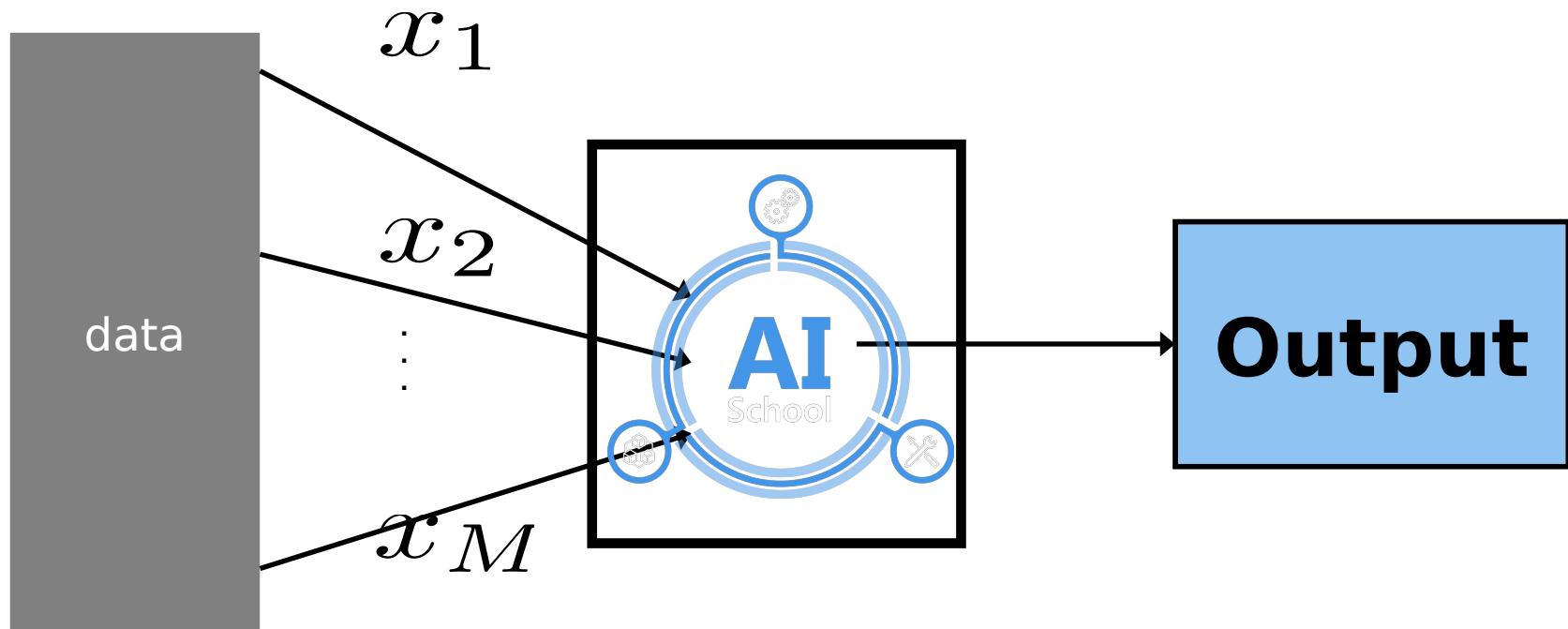
Starting May 25, the European Union will require algorithms to explain their output, making deep learning illegal.

11:59 AM - Jan 29, 2018

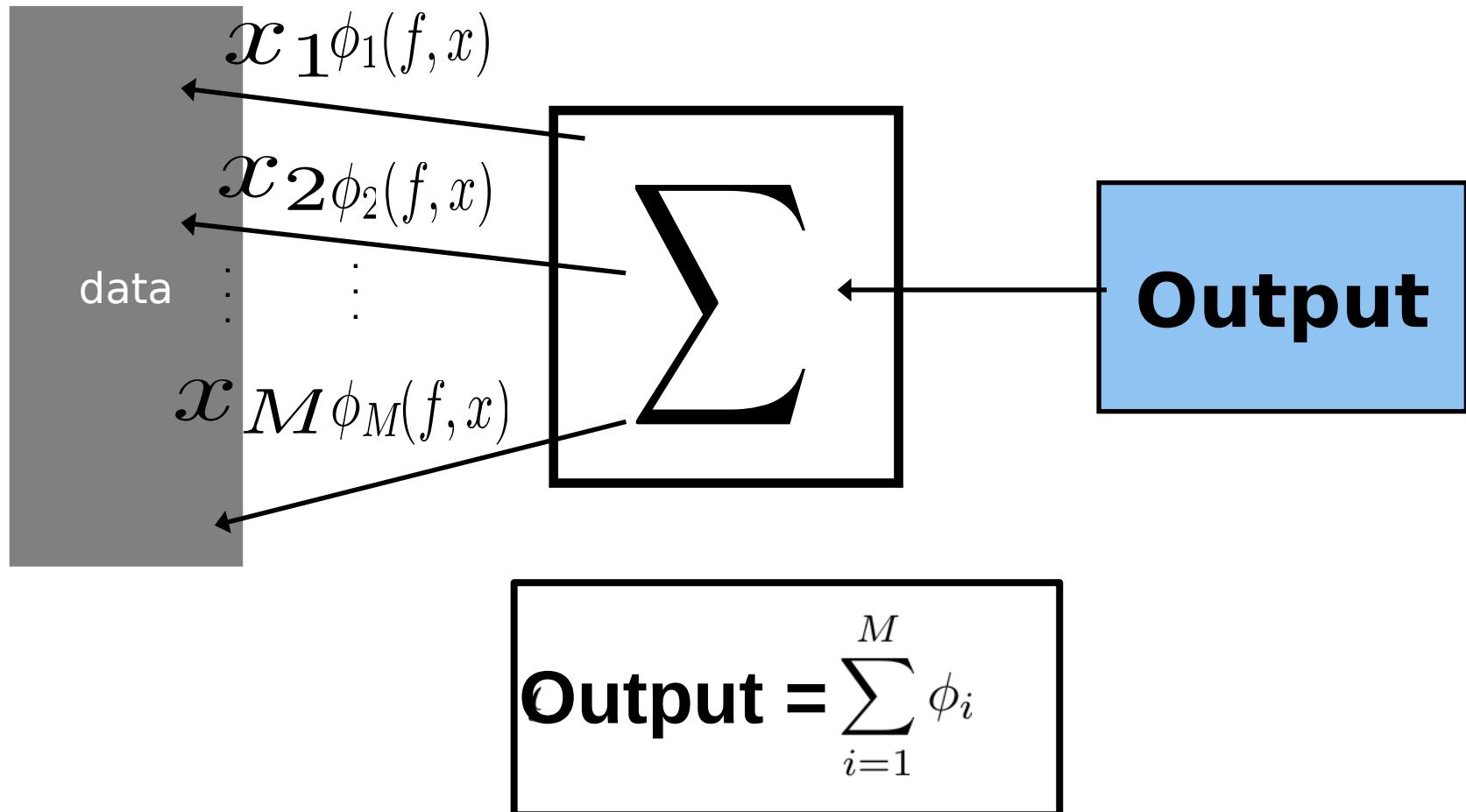
344 people are talking about this

## Cognitive, Human-Like, Empathetic & Explainable Machine-Learning (CHEEM)

# Complicated AI Model



# Explainable model: Additive feature attribution model



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

# Additive feature attribution methods

LIME

Ribeiro et al. 2016

Shapley reg.  
values

Lipovetsky et al. 2001

QII

Datta et al. 2016

$$\sum$$

DeepLIFT

Shrikumar et al. 2016

Relevance prop.

Bach et al. 2015

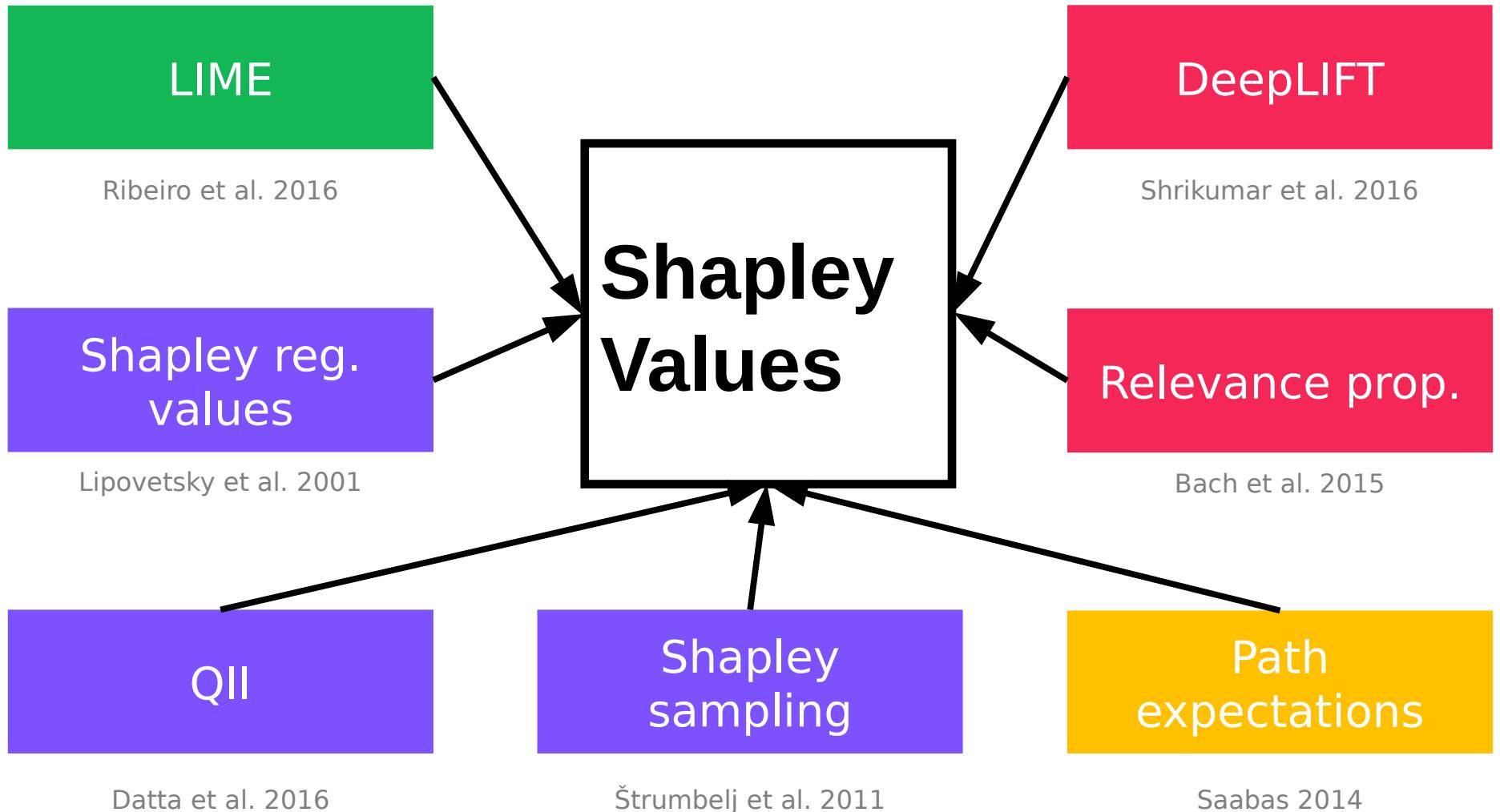
Shapley  
sampling

Štrumbelj et al. 2011

Path  
expectations

Saabas 2014

# Additive feature attribution methods



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



Base rate

Prediction for John

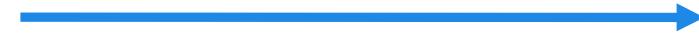
20%

55%

0  
↓

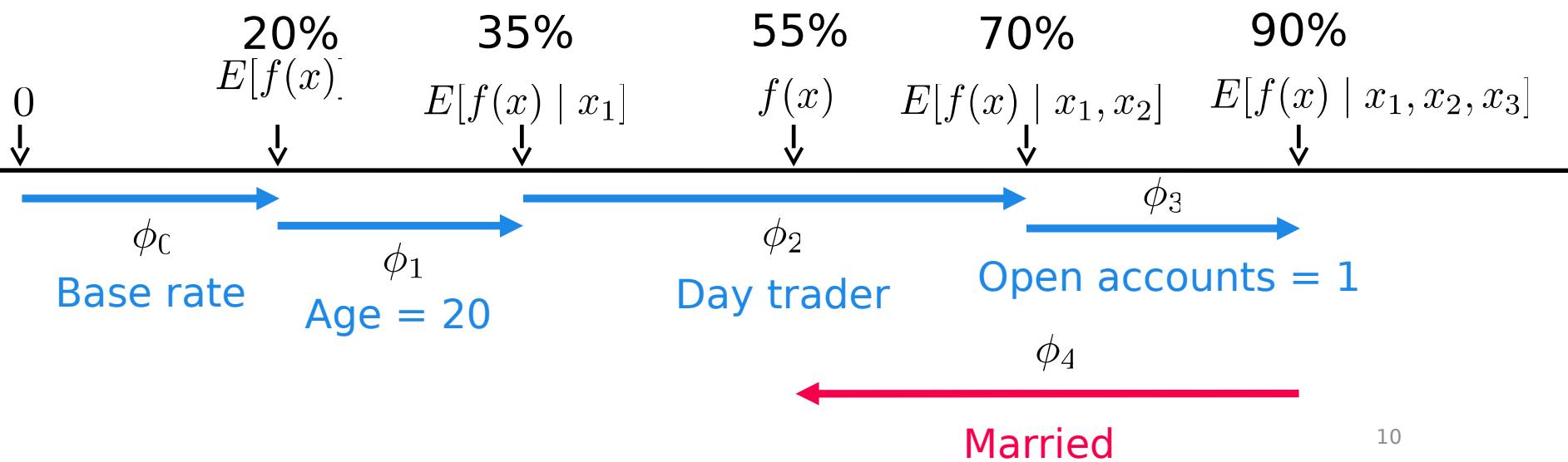
$E[f(x)]$   
↓

$f(x)$   
↓

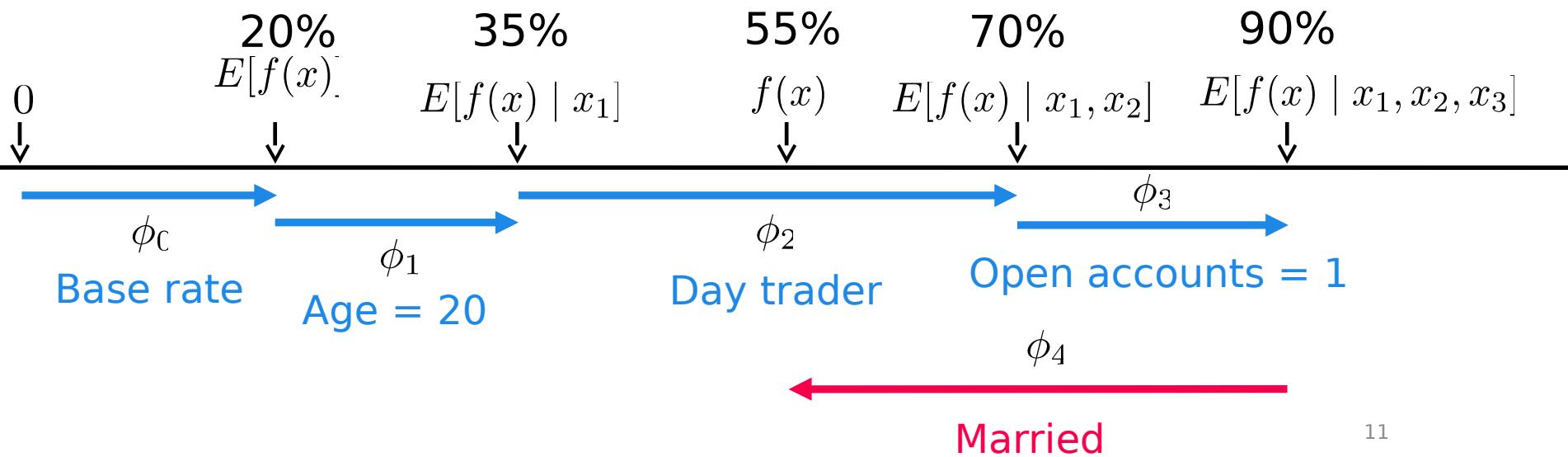


How did we get here?

# SHapley Additive exPlanation (SHAP) values (2)



# SHapley Additive exPlanation (SHAP) values (2)



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

- Train AI model

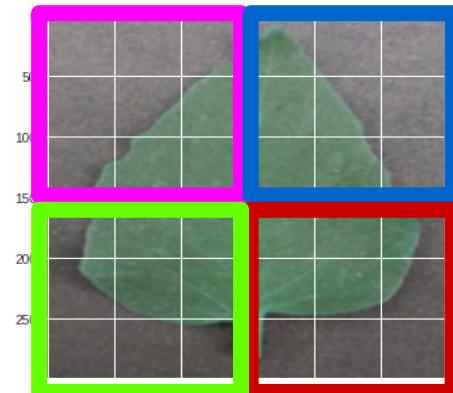
For each picture containing 4 superpixels

{ *Explain model :*

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

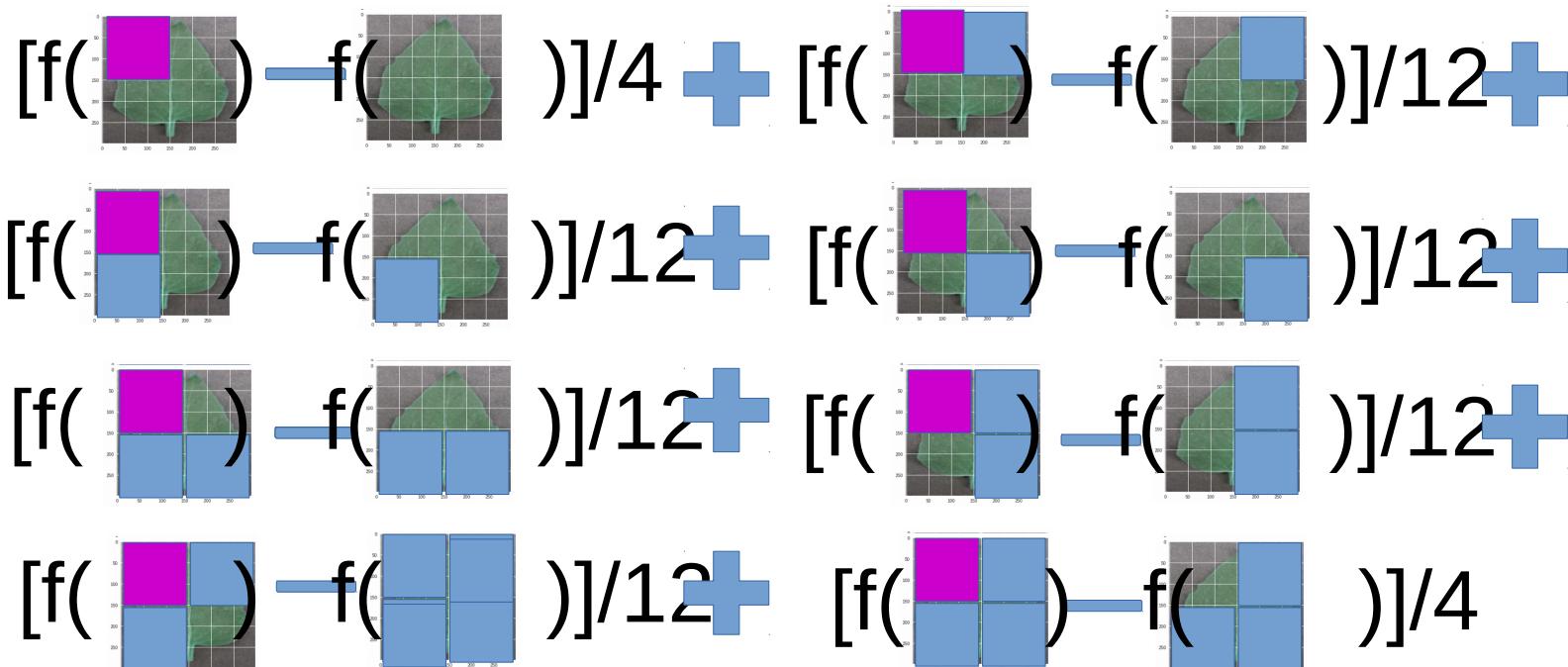
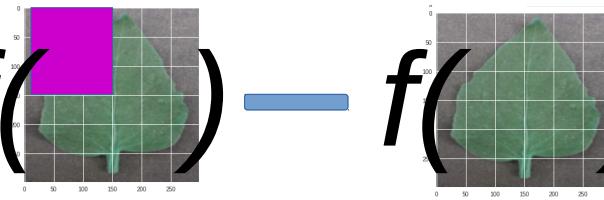
$\varphi_i$  is the shapley value of the feature

}



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

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



**Legend**  
Blue solid  
squared are  
mean-filled  
super-pixels

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

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

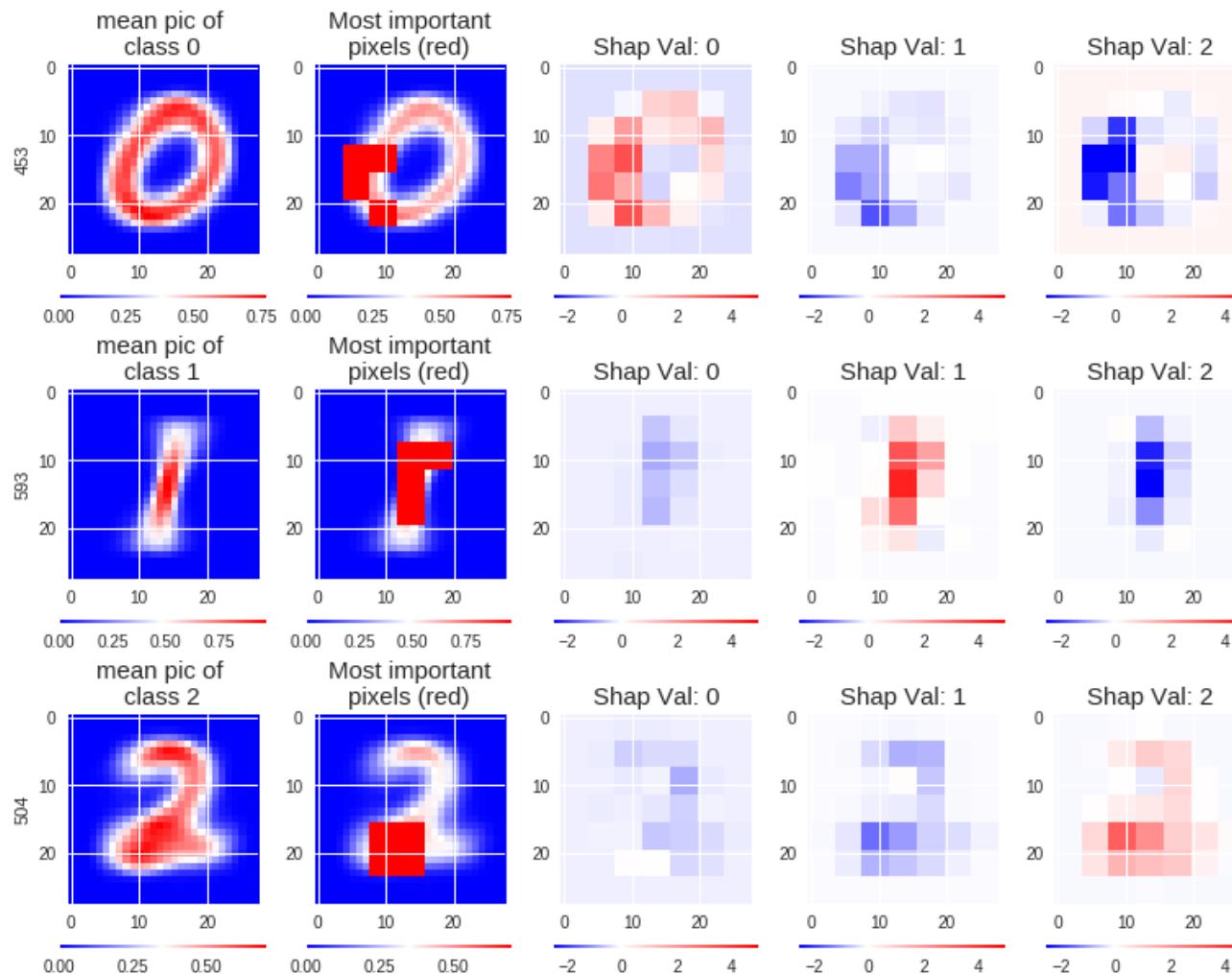
Refer to proof in paper for details, A Unified Approach to Interpreting Model Predictions

# 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  $7 \times 7 = 49$  superpixels for shapley value computation
  - Sample 7367 combinations of pixels
    - ~ all 1 mean-filled pixel images,  ${}^{49}C_1 = 49$
    - ~ all 2 mean-filled pixel images,  ${}^{49}C_2 = 1176$
    - ~ 33% of 3 mean-filled pixel images,  ${}^{49}C_3 / 3 = 6142$
  - Calculate shapley values for each pixel using weighted regression, with g
- }

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

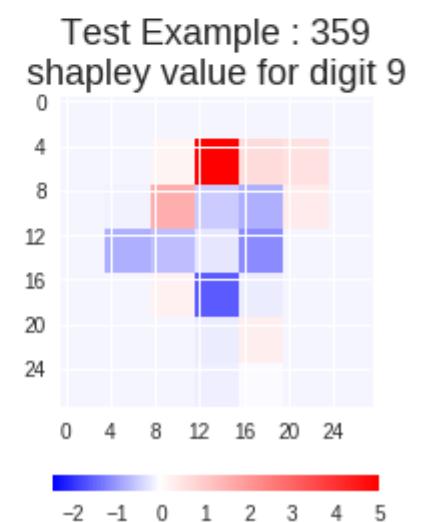
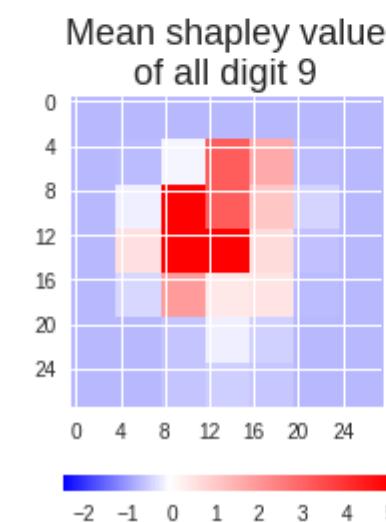
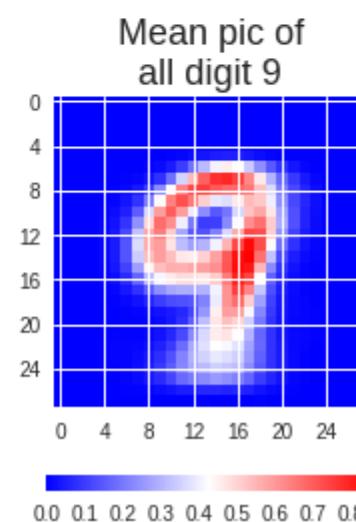
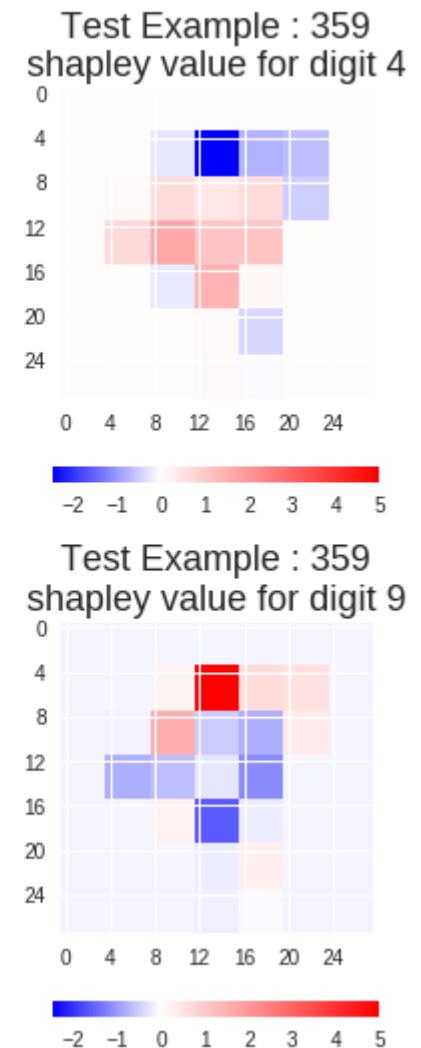
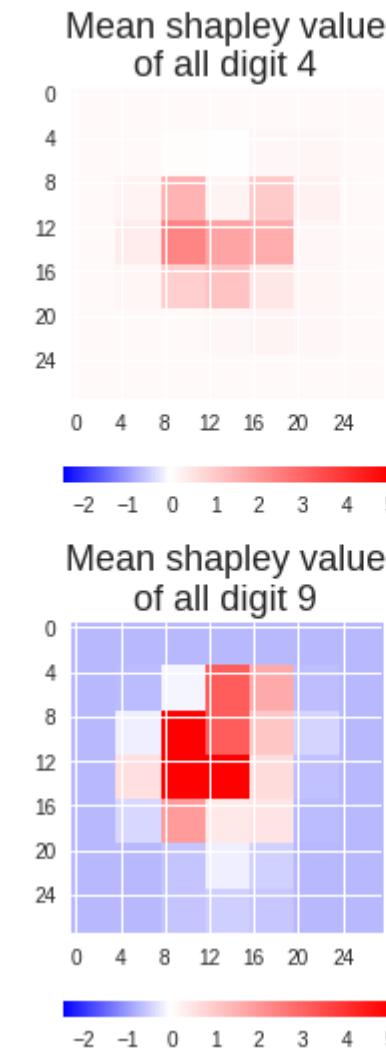
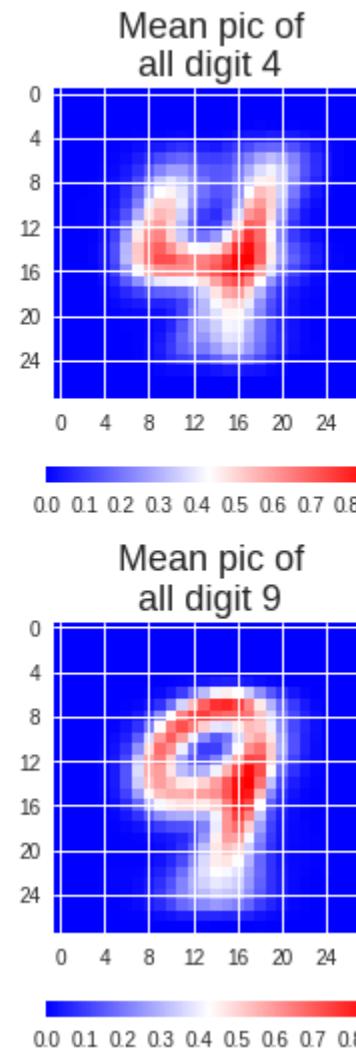
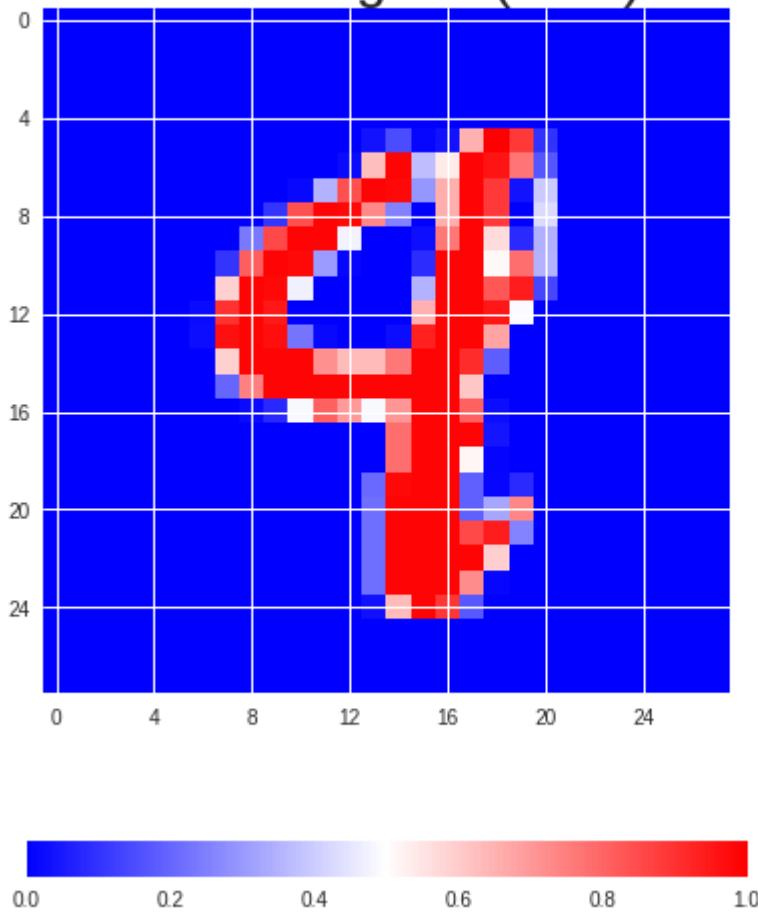
# Applying to Mnist (2) – Global analysis



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

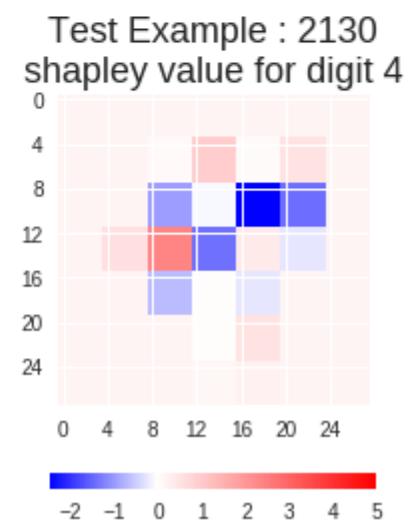
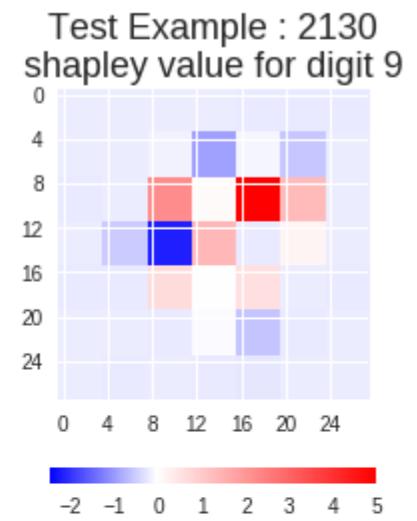
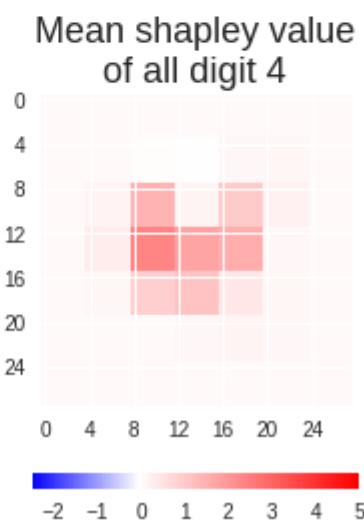
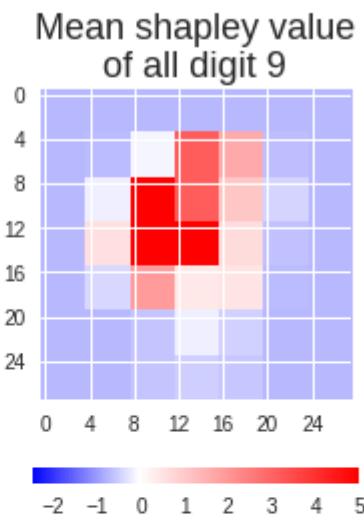
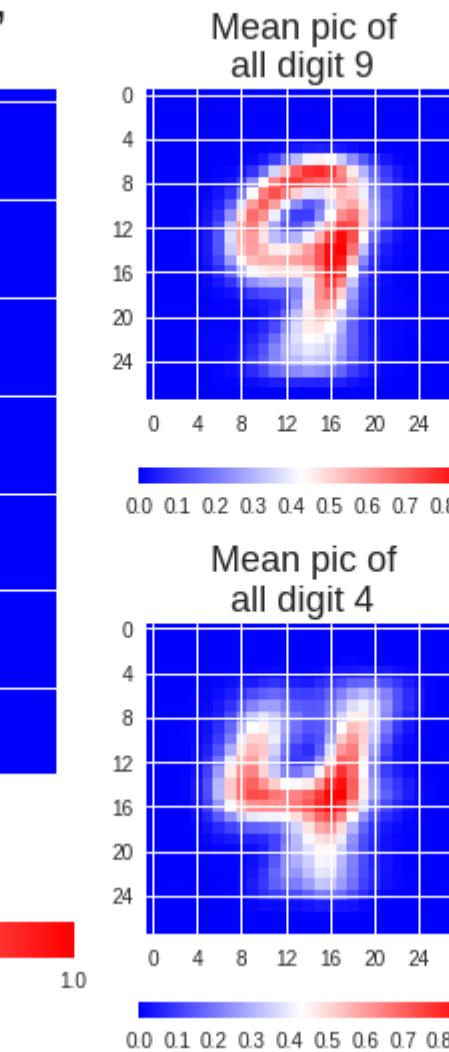
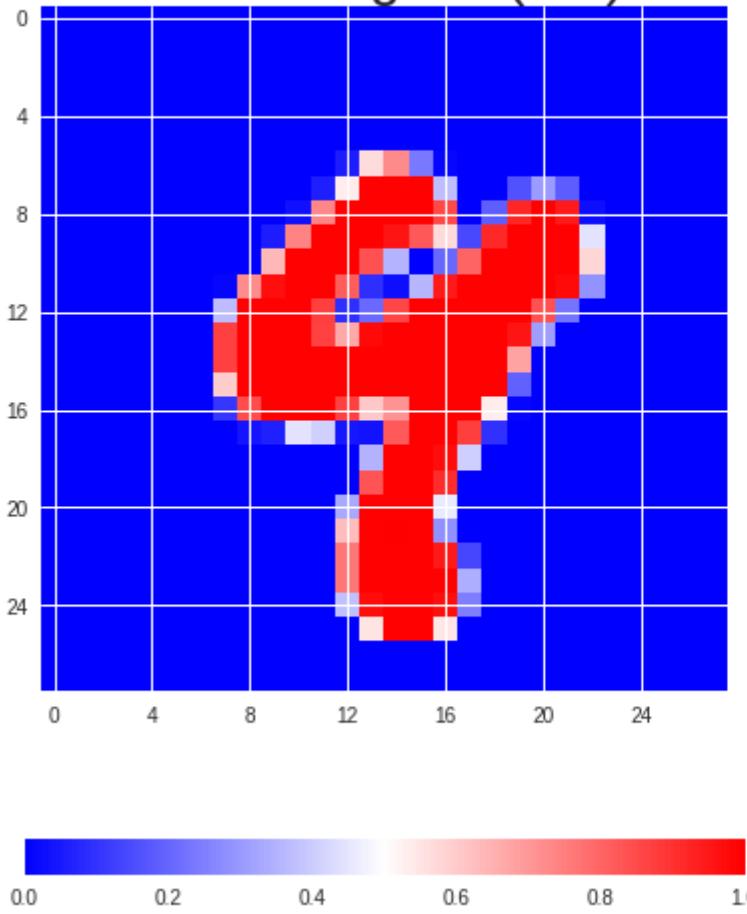
Test Example: 359

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



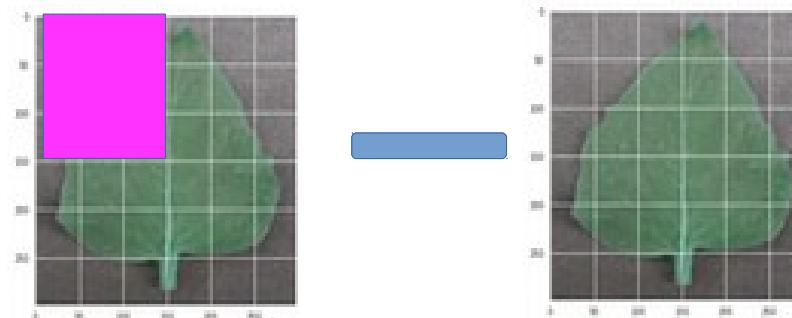
# Applying to Mnist (3) – Individual analysis (b)

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

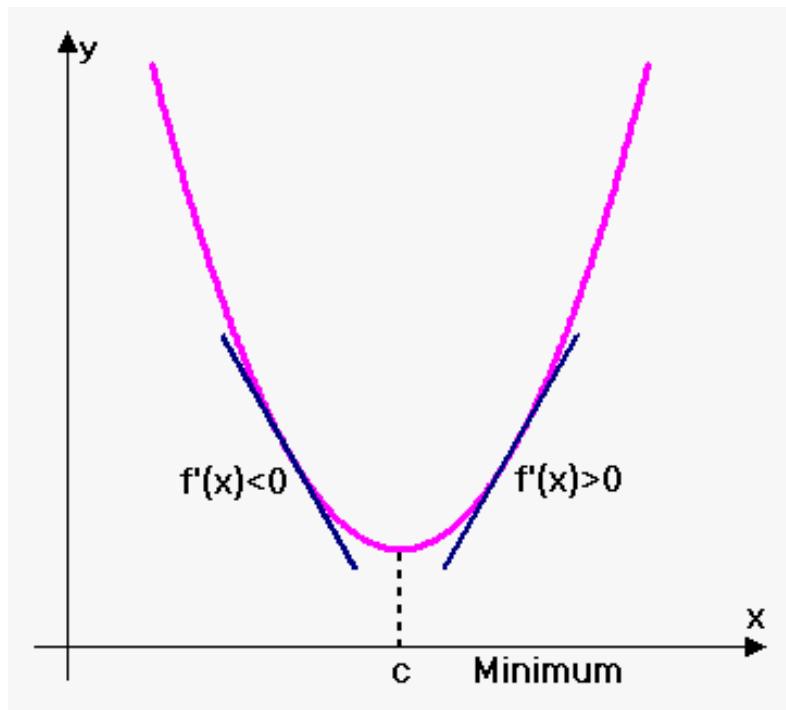


# Drawbacks

- 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 ?



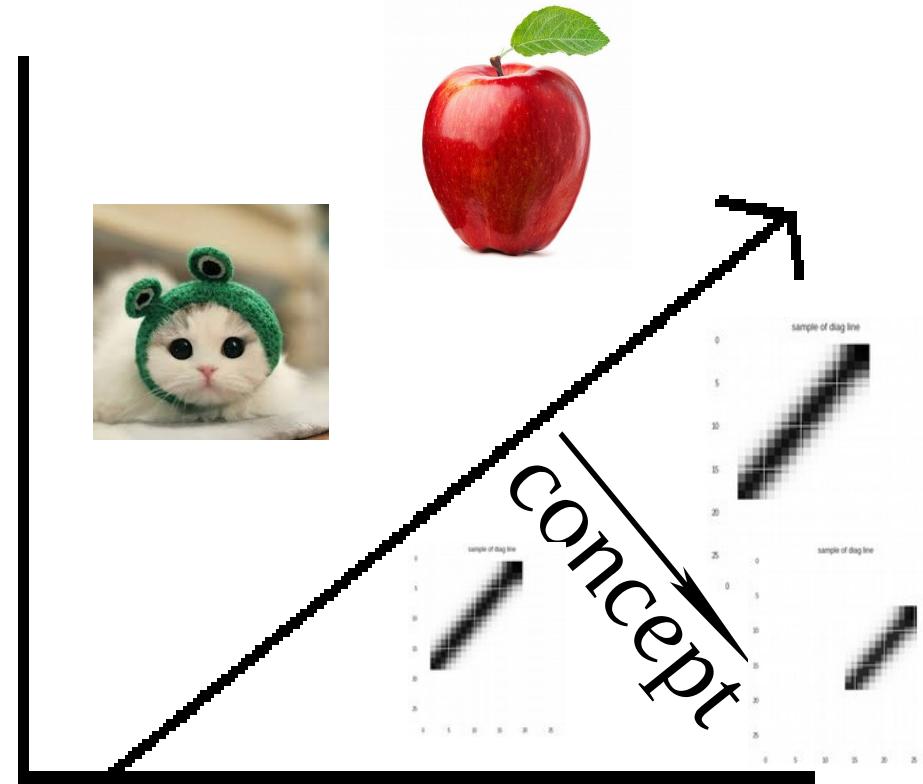
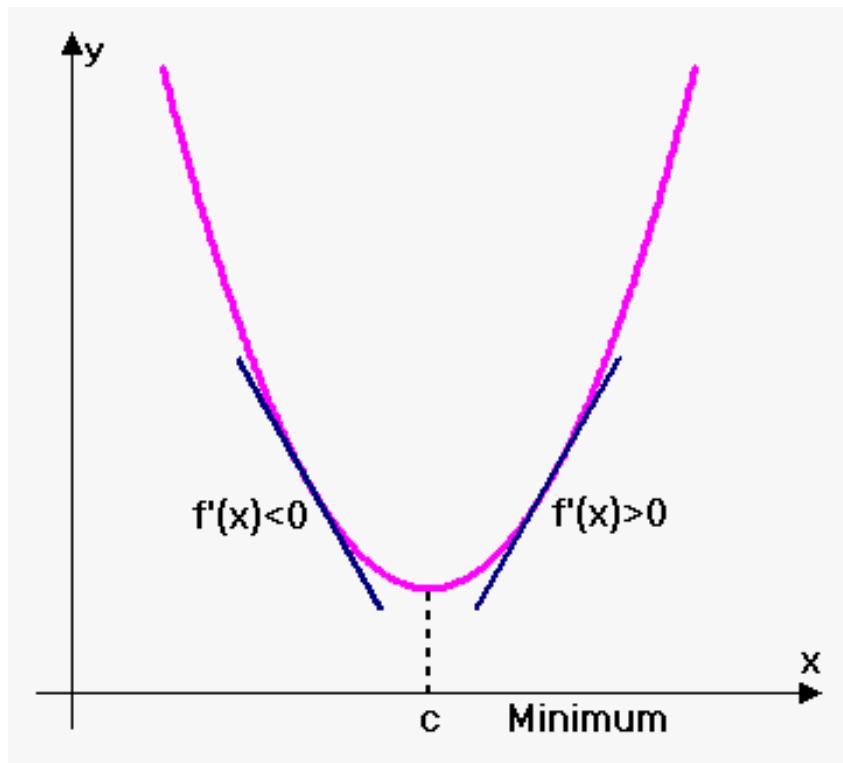
# Gradient based methods



$$\frac{\partial J}{\partial X}.$$

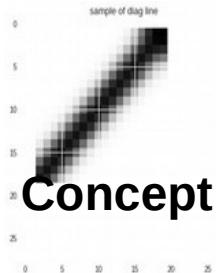
# Gradient based methods

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



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

# Concept activation of mnist model hidden CNN layer



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

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

# Draw backs of local methods

- Global explanation != local explanation
- Example : Young blacks have higher death rates, while old blacks have lower death rates

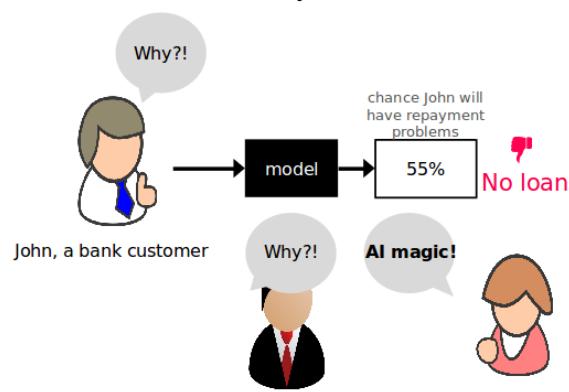
Explainable AI on insurance premium :

- *Because you are black, your price is higher, all things equal relative to your demographic (young, male)*
- *Because you are black, your price is lower, all things equal relative to your demographic (elderly, male)*
- Leads to distrust of the system

# Summary

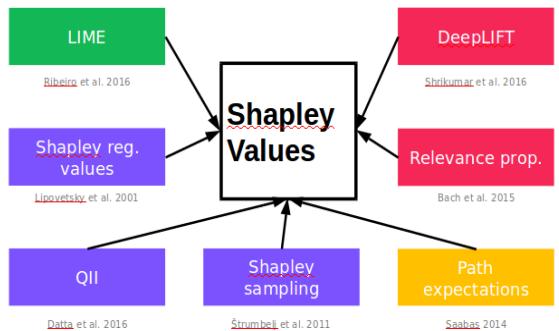
1.

## Need for Explainable AI



2.

## Additive feature attribution methods

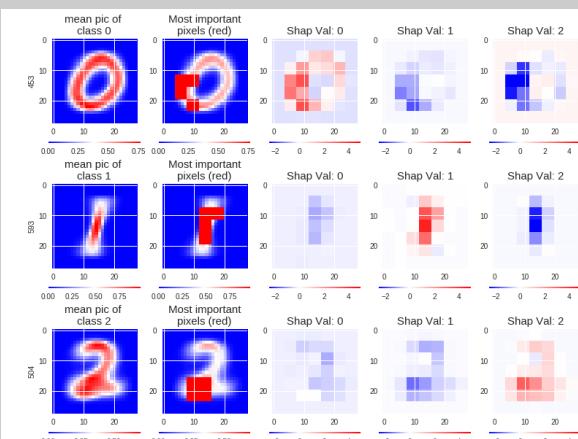


3. Intuition

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

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

4. Analysis mnist



5. Drawbacks



6. Gradient based methods

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

# References

- 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