광운대학교
KwangWoon University

설명 가능한 인공지능 II

(Shapley value and SHAP)

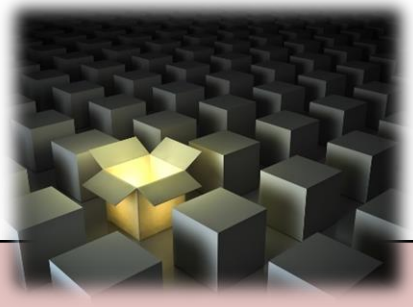
이상민

정보융합학부, 소프트웨어융합대학

광운대학교

Agenda

1. Background
2. Motivation, Why Important?
3. Applying XAI
4. Model-Agnostic methods
    1. Partial Dependence Plot (PDP)
    2. Individual Conditional Expectation (ICE)
    3. Local Surrogate Models (LIME)
    4. Shapley Additive Explanations (SHAP)



Local Interpretable Model-agnostic Expectations (LIME)

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

1 Introduction

The ability to correctly interpret a prediction model's output is extremely important. It engenders appropriate user trust, provides insight into how a model may be improved, and supports understanding of the process being modeled. In some applications, simple models (e.g., linear models) are often preferred for their ease of interpretation, even if they may be less accurate than complex ones. However, the growing availability of big data has increased the benefits of using complex models, so bringing to the forefront the trade-off between accuracy and interpretability of a model's output. A wide variety of different methods have been recently proposed to address this issue [5, 8, 9, 3, 4, 1]. But an understanding of how these methods relate and when one method is preferable to another is still lacking.

Here, we present a novel unified approach to interpreting model predictions.[1] Our approach leads to three potentially surprising results that bring clarity to the growing space of methods:

1. We introduce the perspective of viewing *any* explanation of a model's prediction as a model itself, which we term the *explanation model*. This lets us define the class of *additive feature attribution methods* (Section 2), which unifies six current methods.

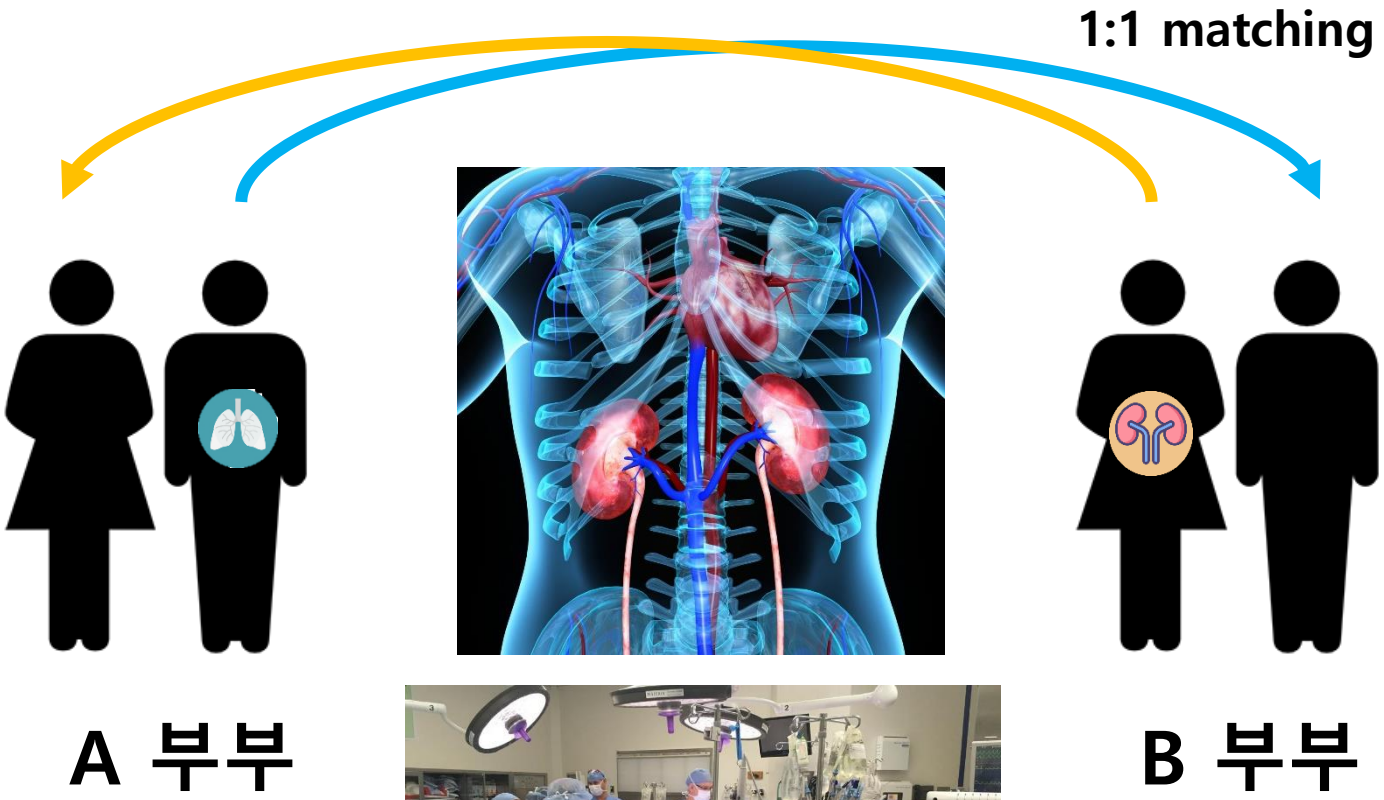
[1] https://github.com/slundberg/shap

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Lloyd Shapley in 2012

Shapley Value Application

- **Transplant Program, New England**



Cooperative game theory

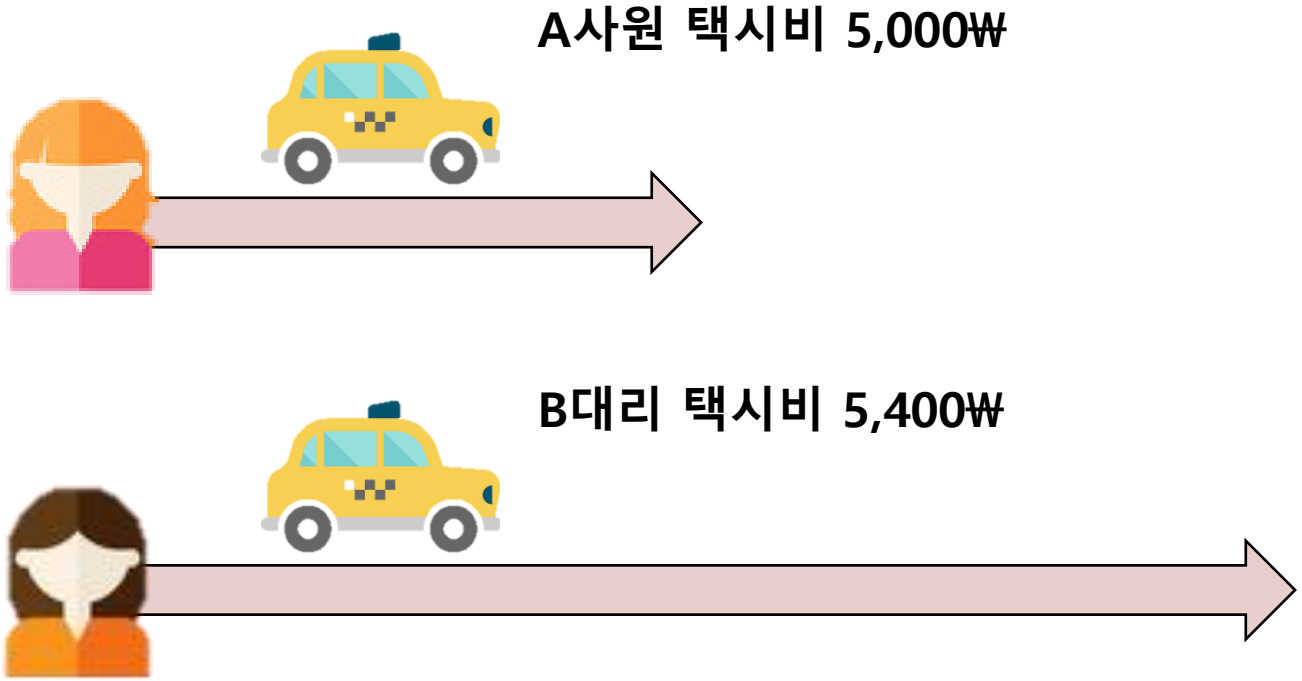
- **Conventional Game Theory**
  - ✓ **Branch of micro-economics dealing with interactions between decision-making agents**
  - ✓ **Limitations: information is not shared between agents**
- **Cooperative Game Theory**
  - ✓ **Subfield of game theory where players are "working together" to achieve a common goal**
  - ✓ **Regarding AI, we can view the features of the model as the "players", and the prediction of the model as the "game's result"**

SHAP (Shapley Additive exPlanations)

- **Definition**
  - ✓ **The goal of SHAP is to explain the prediction of a data point x by computing the contribution of each feature to the prediction**
  - ✓ **The SHAP explanation method computes Shapley values from coalitional game theory**
  - ✓ **A player can be an individual feature value**
  - ✓ **the Shapley value explanation is represented as an additive feature attribution method, a linear model**
  - ✓ **That view connects LIME and Shapley Values**

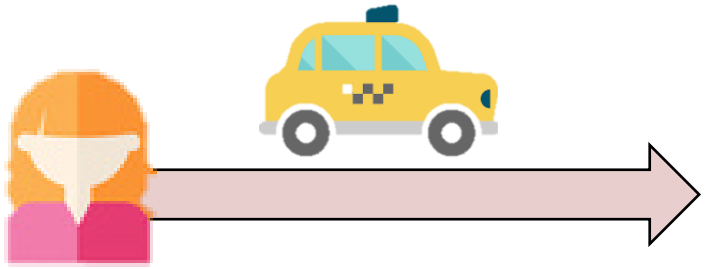
Shapley Value

- **Shapley value 예시**

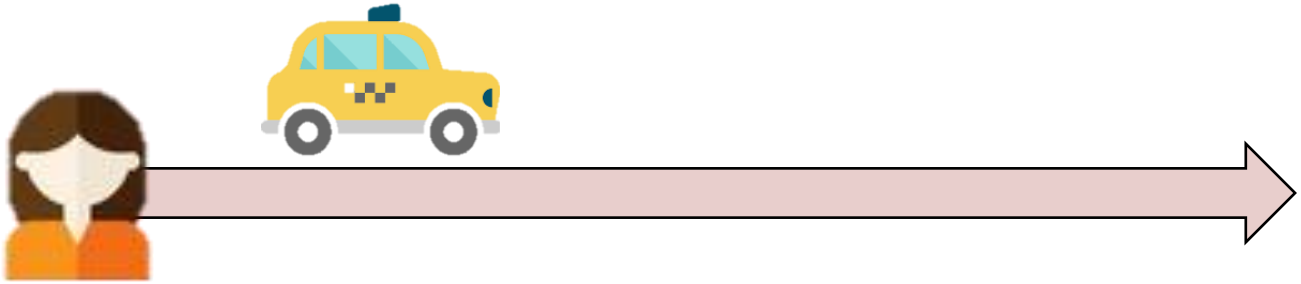


Shapley Value

- Shapley value 예시 (택시를 동승하면 요금을 아낄 수 있다.)



A사원, B대리 합승 택시비8,000₩



Shapley Value

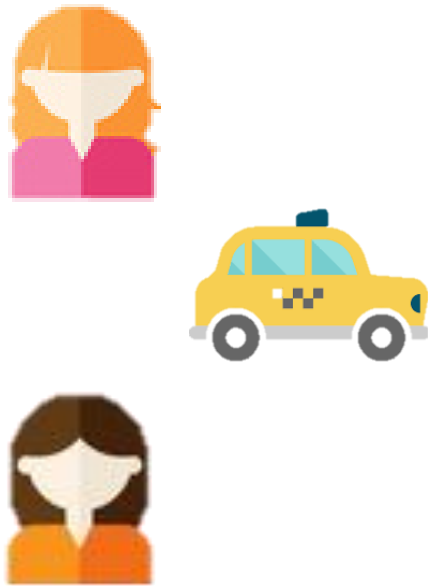
- **Shapley value 예시 (어떻게 분담할지 이견이 있음)**

"... 그럴 바에는 혼자 타고 갈래요, 대리님"

"A사원 택시비 절반씩 나눕시다. 4,000원씩"

Shapley Value

- **Shapley value은 이득의 공평한 배분으로 이 문제를 해결**



✓ **합승의 이득**

$$(5{,}000 + 5{,}400) - (8{,}000) = 2{,}400$$

✓ **이득을 공평히 나누자**

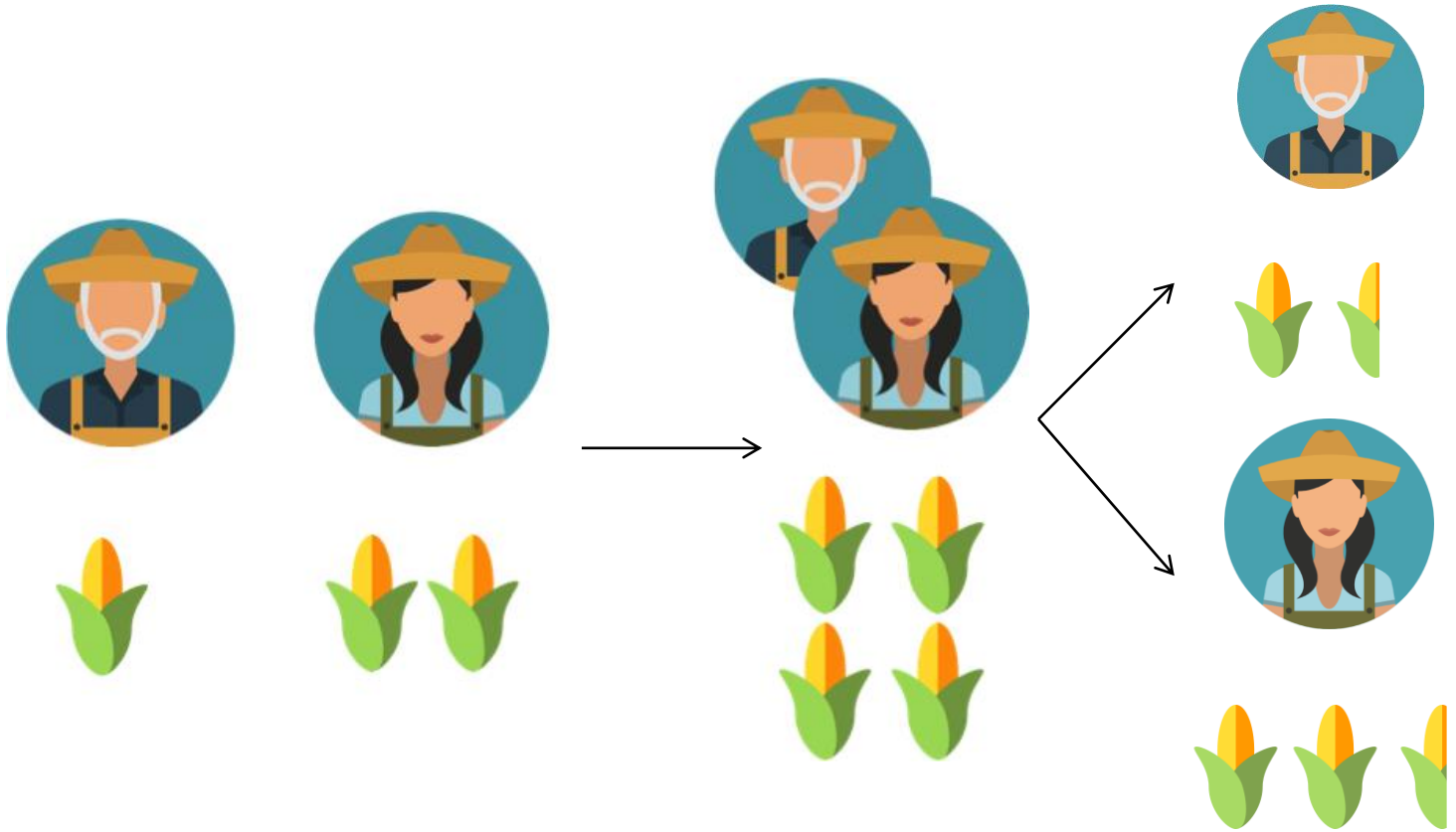
$$\frac{2{,}400}{2} = 1{,}200$$

✓ **각자의 택시비는 얼마?**

A사원: 5,000 – 1,200 = 3,800

B대리: 5,400 – 1,200 = 4,200

Coalitional game theory identifies key players

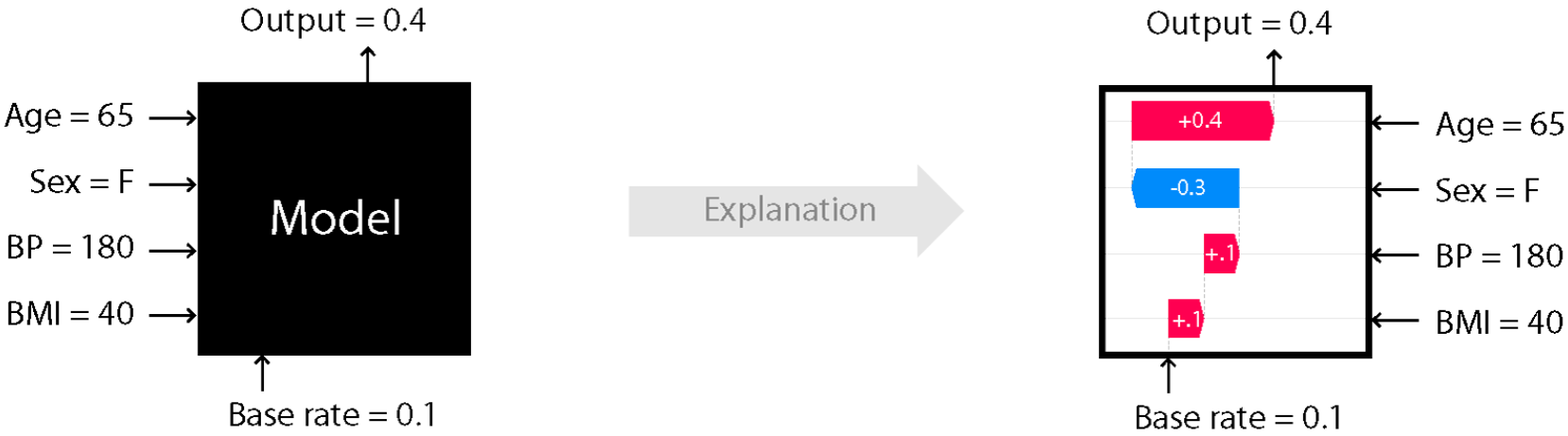


$$\phi_i(N,v) = \frac{1}{N!} \sum_{S \subseteq N, i \subseteq S} |S|! * (|N| - |S| - 1)! * [v(S) - v(S \backslash \{i\})]$$

Reference: Moretti et al. 2008

SHAP (Shapley Additive exPlanations)

SHAP



Reference: https://github.com/slundberg/shap

SHAP (Shapley Additive exPlanations)

- **Formula**

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

$z' \in \{0,1\}^M$: coalition vector

M: maximum coalition size

$\phi_j \in \mathbb{R}$: feature attribute for a feature j

$$g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j$$

SHAP Characteristics

- **Local accuracy**
- **Missingness**
- **Consistency**

SHAP Characteristics

- **Local accuracy**

$$f(x) = g(x') = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j$$

If you define $\phi_0 = E_X\left(\hat{f}(x)\right)$ and set all x'_j to 1, this is the Shapley efficiency property

$$f(x) = \phi_0 + \sum_{j=1}^{M} \phi_j x'_j = E_X(\hat{f}\,(X)) + \sum_{j=1}^{M} \phi_j$$

SHAP Characteristics

- **Missingness**
  - ✓ Missingness says that a missing feature gets an attribution of zero

$$x'_j = 0 \Rightarrow \phi_j = 0$$

  - ✓ The presence of a 0 would mean that the feature value is missing for the instance of interest.
  - ✓ minor book-keeping property

SHAP Characteristics

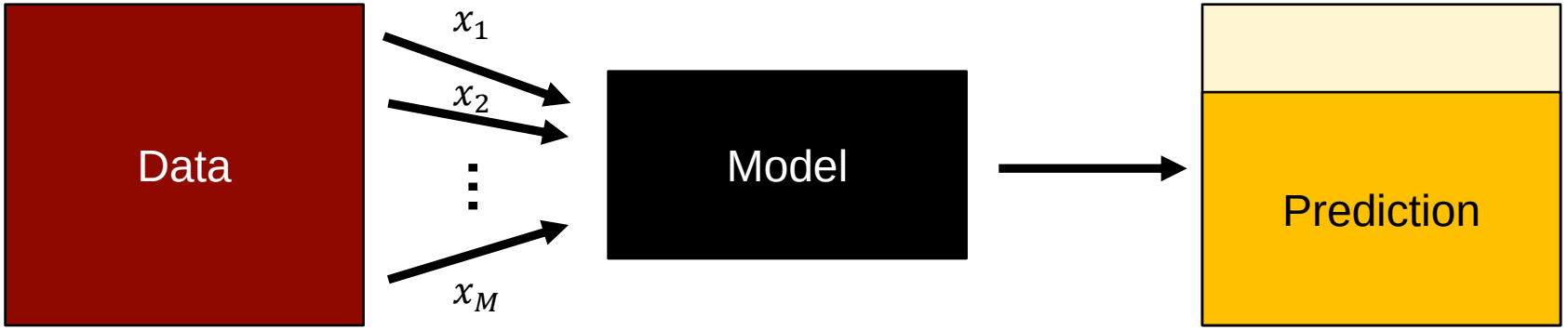
- **Consistency**
  - ✓ If a model changes so that the marginal contribution of a feature value increases or stays the same, the Shapley value also increases or stays the same
  - ✓ We set $f_x(z') = f(h_x(z'))$ and $z_{\setminus j'}$ denotes that $z'_j = 0$

$$f'_x(z') - f'_x(z'_{\setminus j}) \geq f_x(z') - f_x(z'_{\setminus j})$$

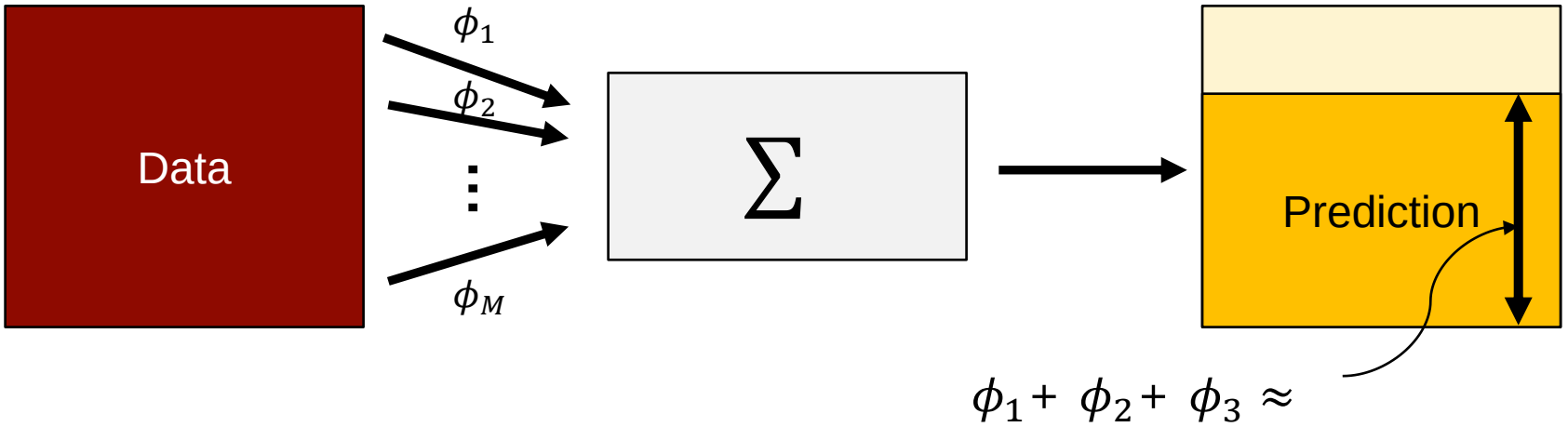
$$\phi_j(f', x) \geq \phi_j(f, x)$$

SHAP Calculation

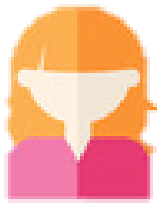
$$x \rightarrow f(x) \rightarrow y$$

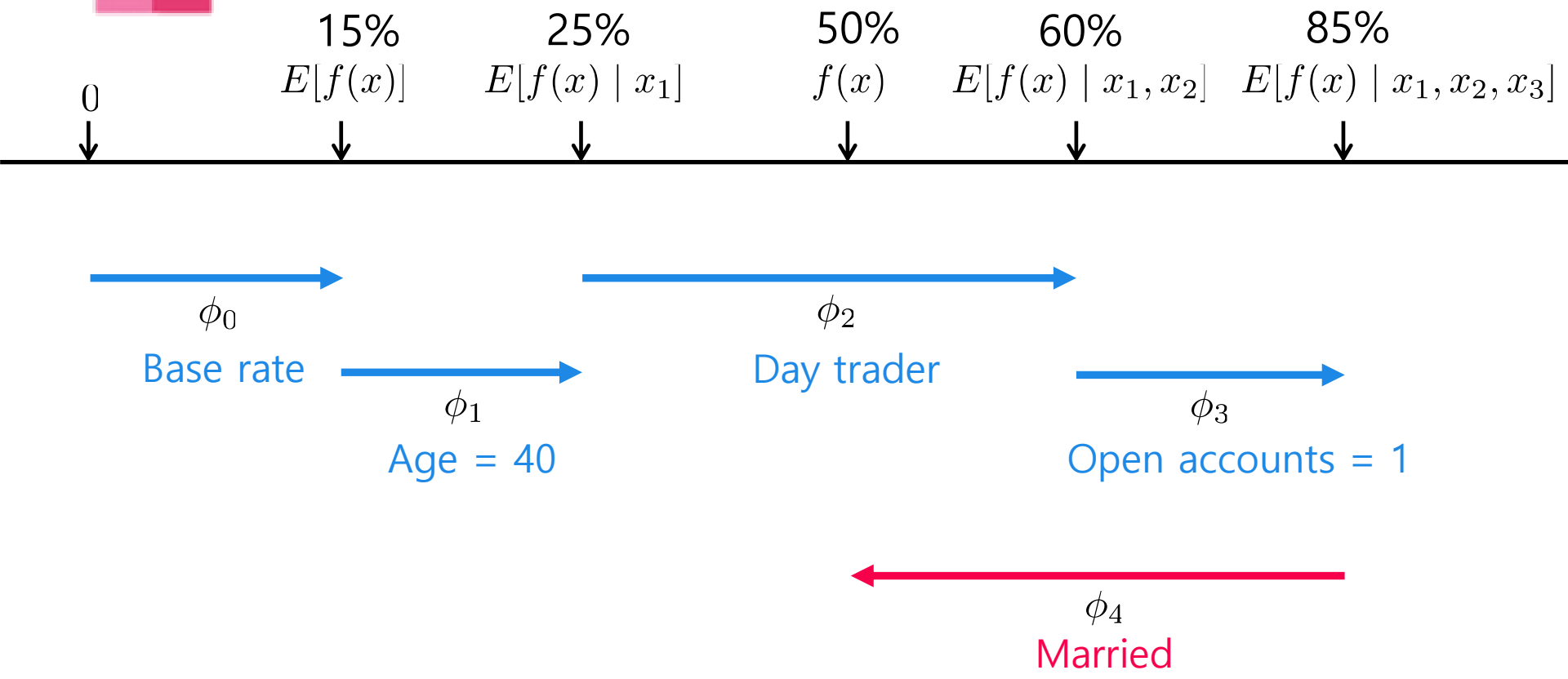


Feature attributions approxmiation

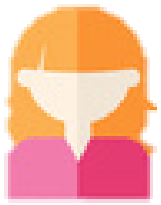


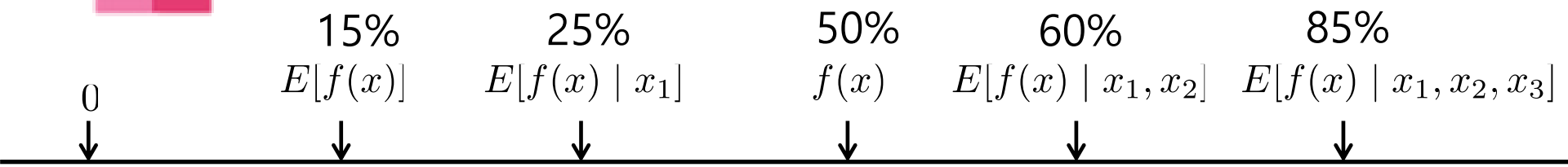
SHAP Calculation





SHAP Calculation



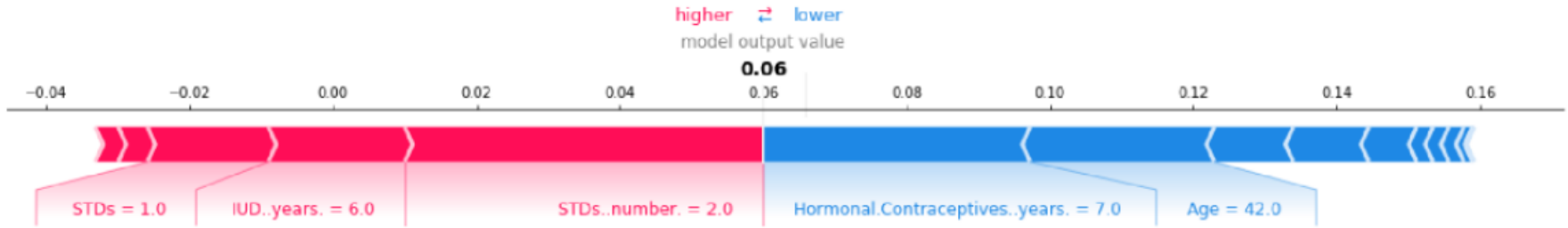


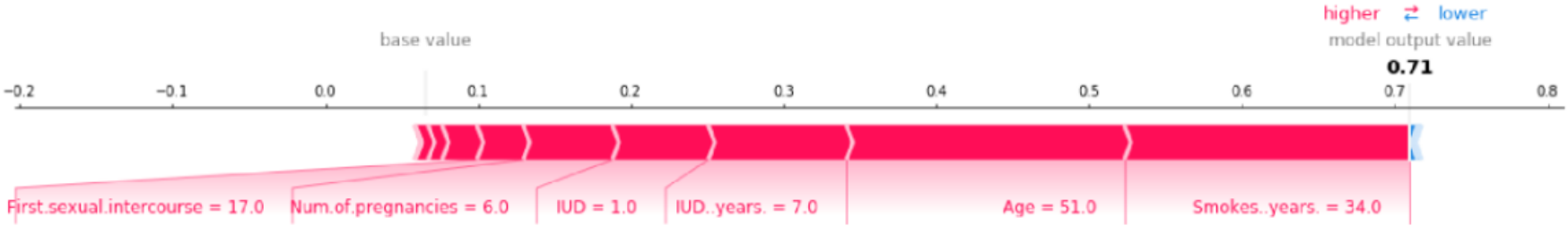
$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)]$$

- **Time complexity for too many combinations**
  - ✓ Impractical approach
  - ✓ What else? KernelSHAP, treeSHAP, DeepSAHP

SHAP Example

- **SHAP explanations**

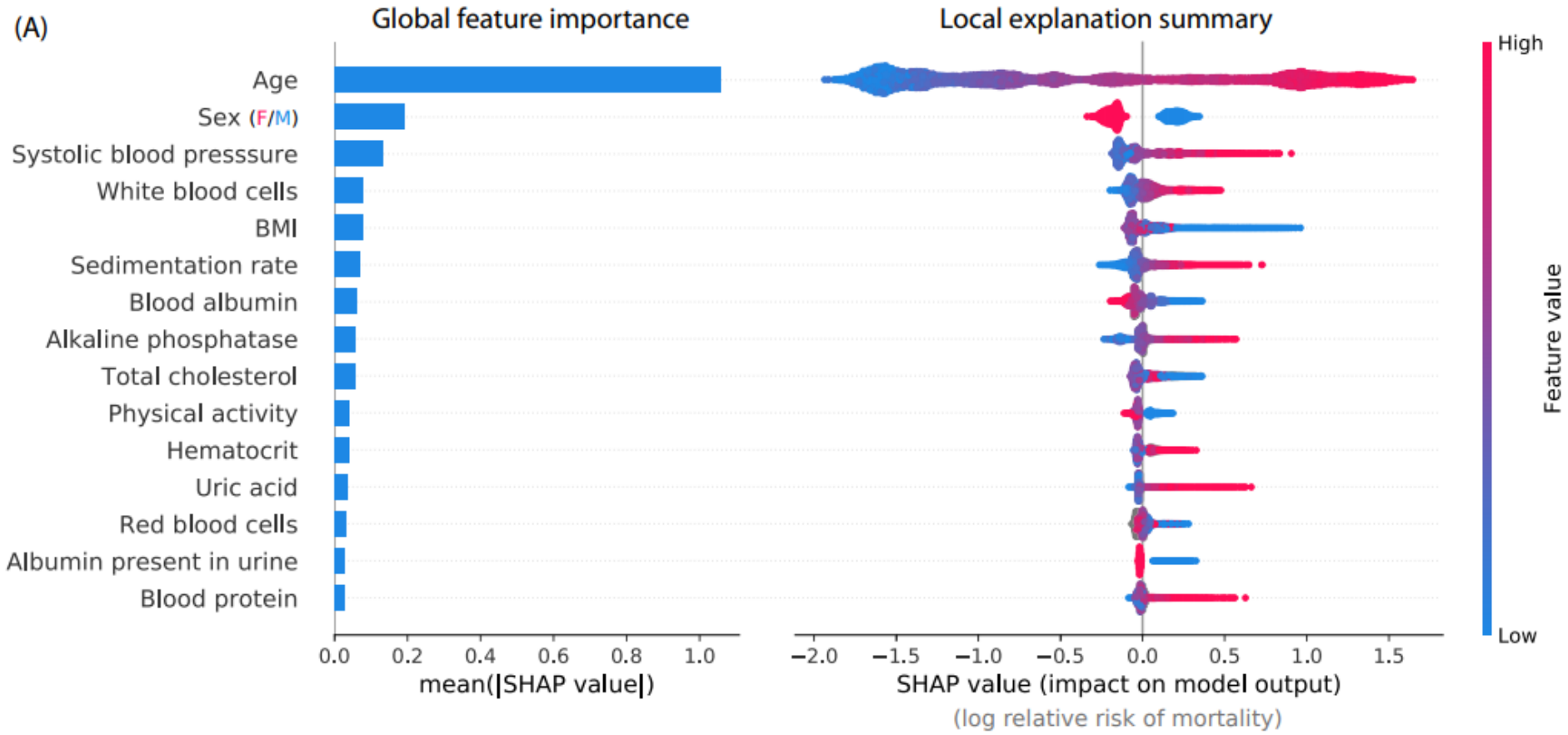




Reference: https://github.com/slundberg/shap

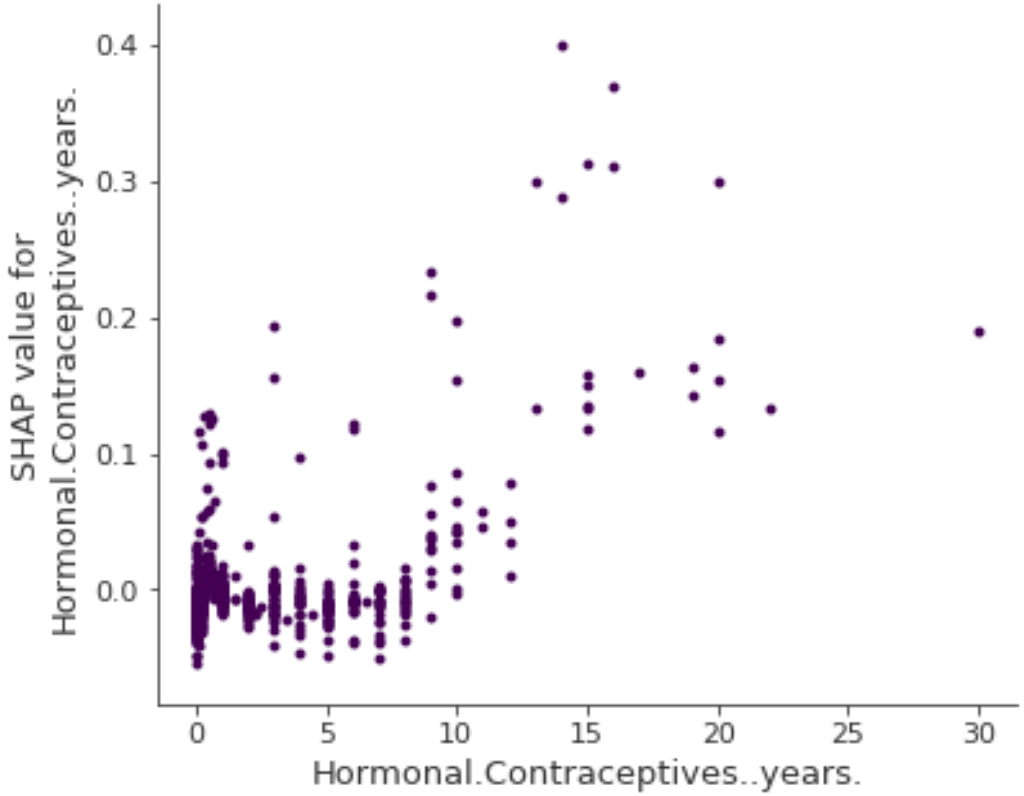
SHAP Summary Plot

- **SHAP Feature Importance**



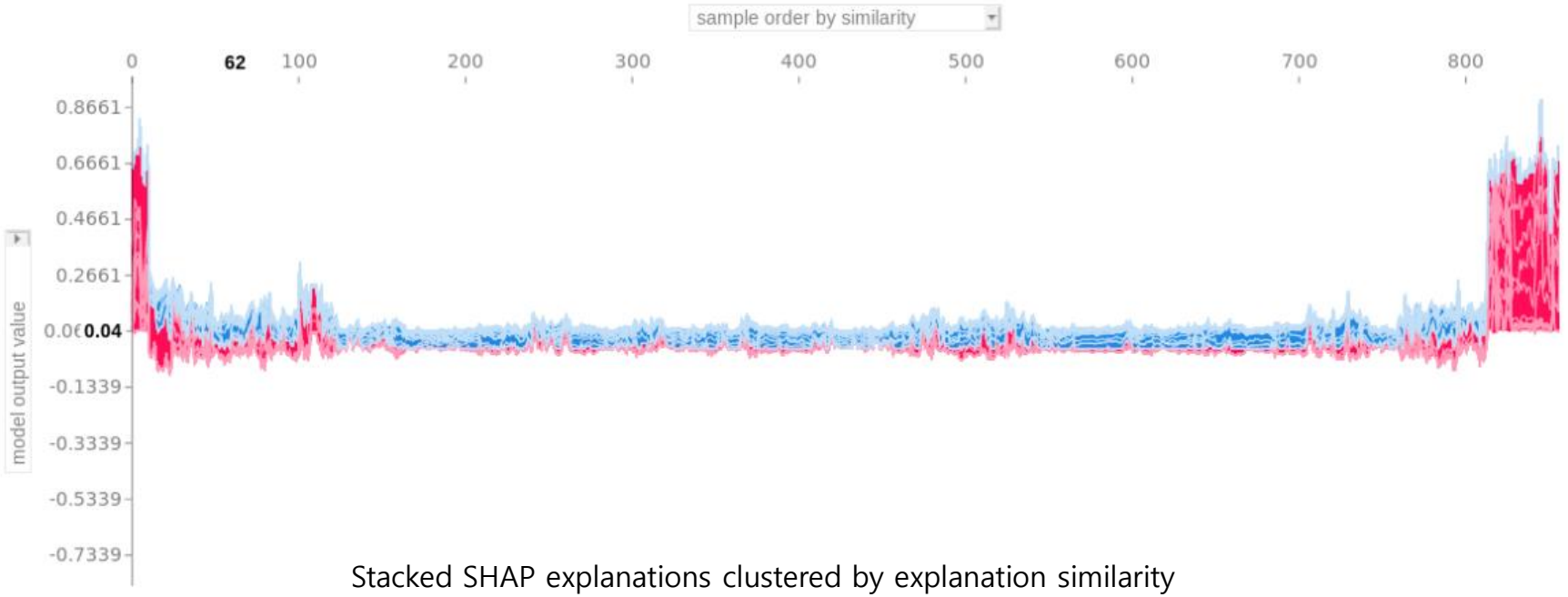
SHAP Dependence Plot (SDP)

- **SHAP dependence plot for years on hormonal contraceptives**



Clustering SHAP values

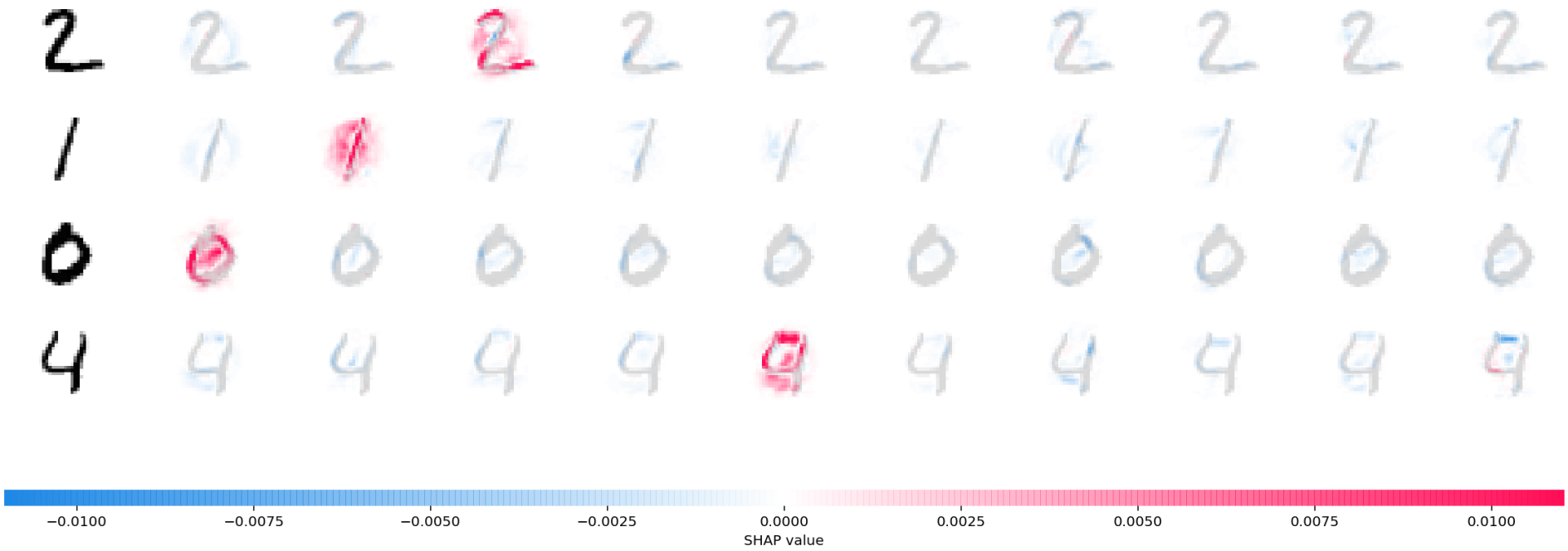
- **SHAP can be used to identify groups of similar data points**
- **SHAP clustering works by clustering on Shapley values of each instance**
  - ✓ **Clustering works with instance explanation similarity**



Stacked SHAP explanations clustered by explanation similarity

SHAP for MNIST dataset

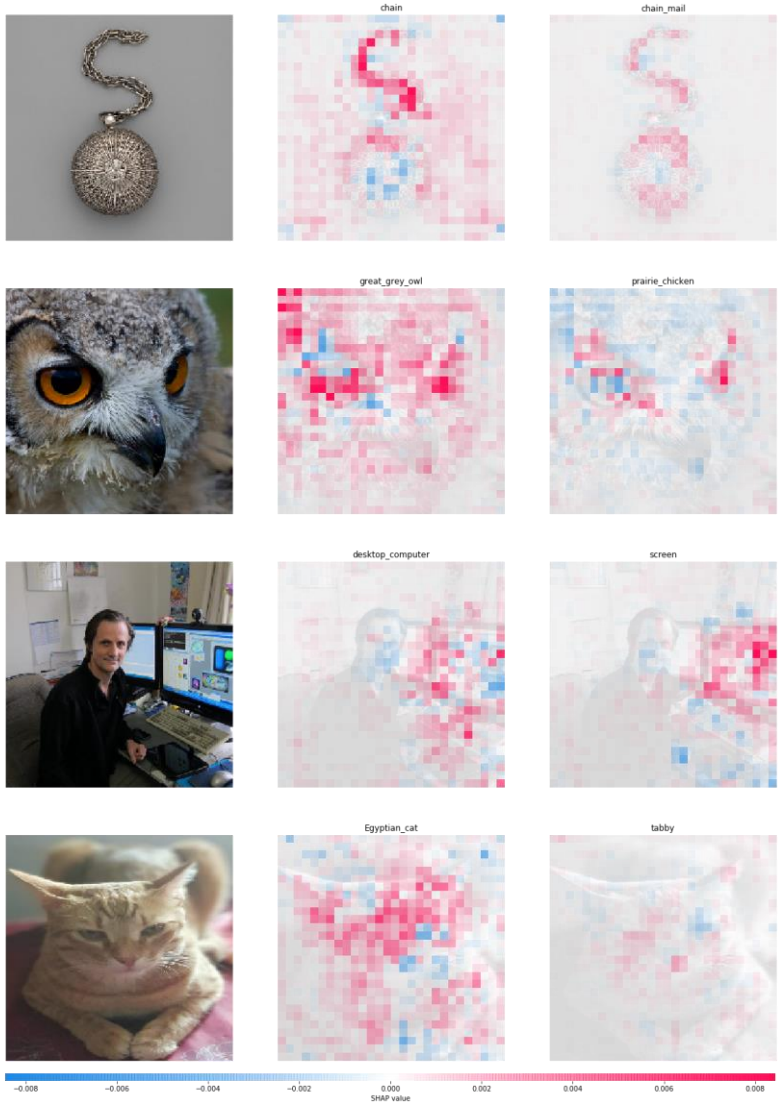
- **DeepSHAP**



Reference: https://github.com/slundberg/shap

SHAP for Image dataset

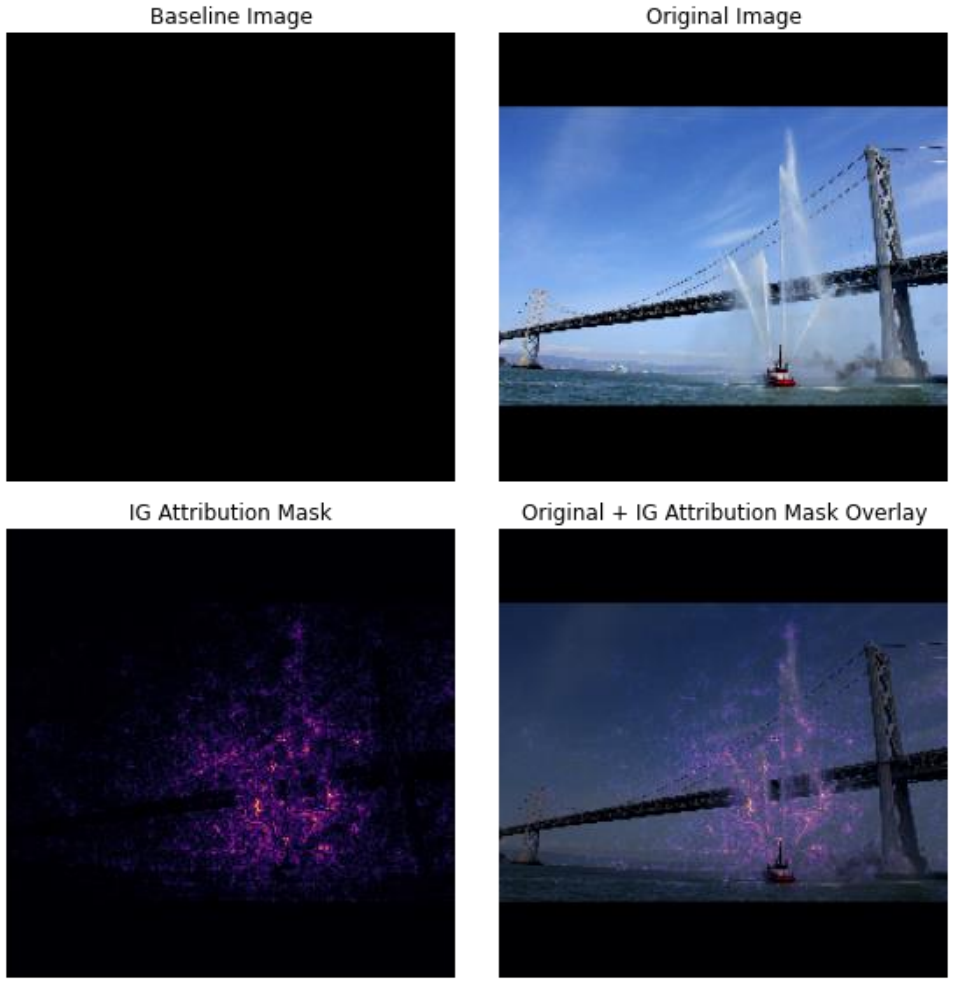
- **DeepSHAP**



Reference: https://medium.com/google-developer-experts/interpreting-deep-learning-models-for-computer-vision-f95683e23c1d

SHAP for Image dataset

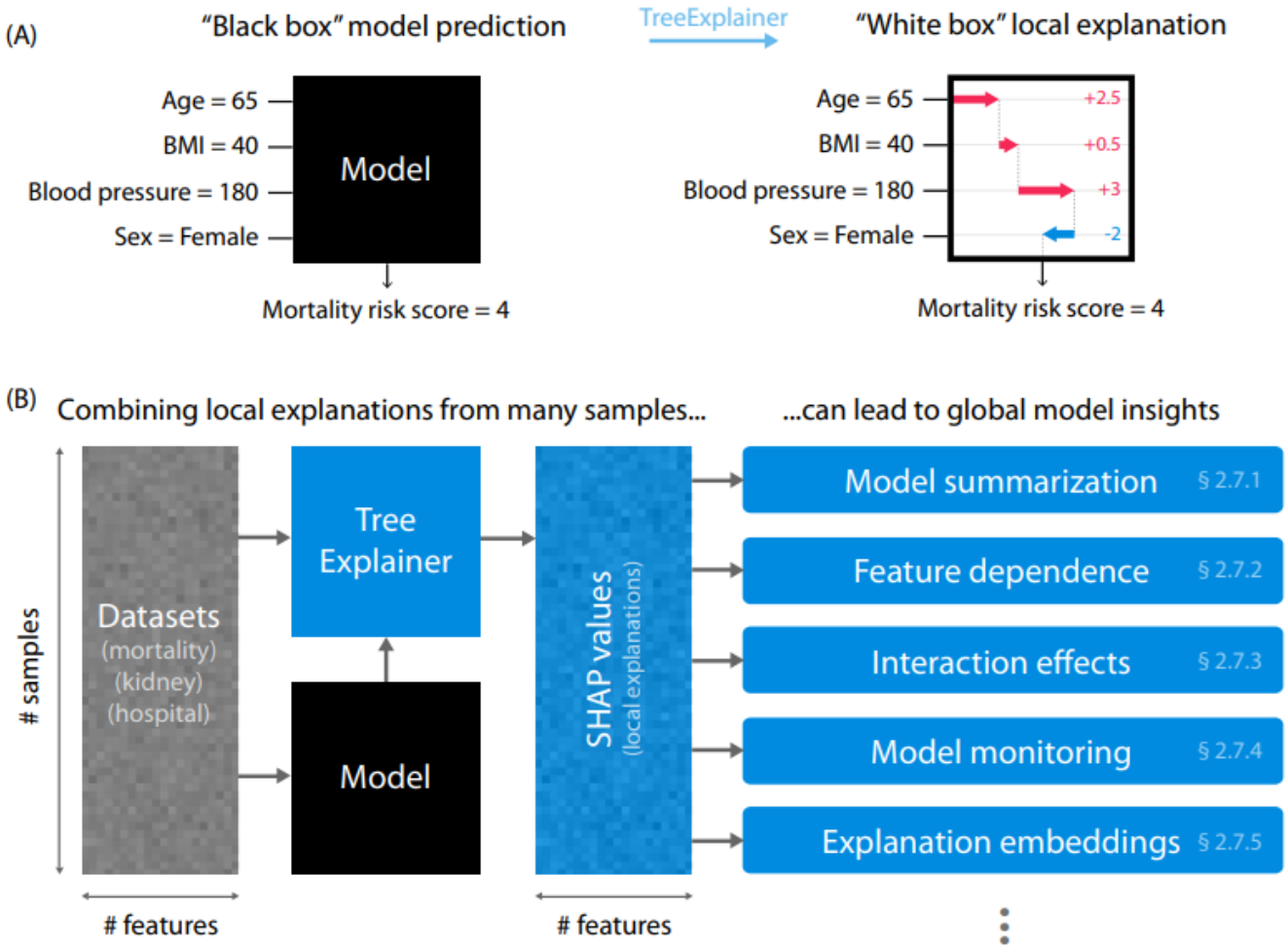
- **Integrated Gradient SHAP**



Reference: https://www.tensorflow.org/tutorials/interpretability/integrated_gradients

TreeExplainer

- **Local explanation based on TreeExplainer**



SHAP Summary

- **Advantages**
  - ✓ SHAP is a surrogate model providing the local and global explanations together
  - ✓ SHAP has a solid theoretical foundation in game theory
  - ✓ SHAP is a contrastive explanations comparing the prediction with average
  - ✓ SHAP connects LIME and Shapley values
  - ✓ SHAP has a fast implementation for tree-based models

- **Disadvantages**
  - ✓ KernelSHAP is slow in terms of the number of data points
  - ✓ KernelSHAP ignores feature dependence regarding permutation
  - ✓ TreeSHAP can produce unintuitive feature attributions