# Shapley Flow:

## A Graph-based Approach to Interpreting Model Predictions

J. Wang, J. Wiens, S. Lundberg

AISTATS 2021

**WKD, M. Krzyziński, 29.07.2021**

# Motywacja

*Explaining a model's predictions by assigning importance to its inputs (i.e., feature attribution) is* ***critical*** *to many applications in which a user interacts with a model to either make decisions or gain a better understanding of a system.*

*However,* ***correlation among input features presents a challenge*** *when estimating feature importance.*

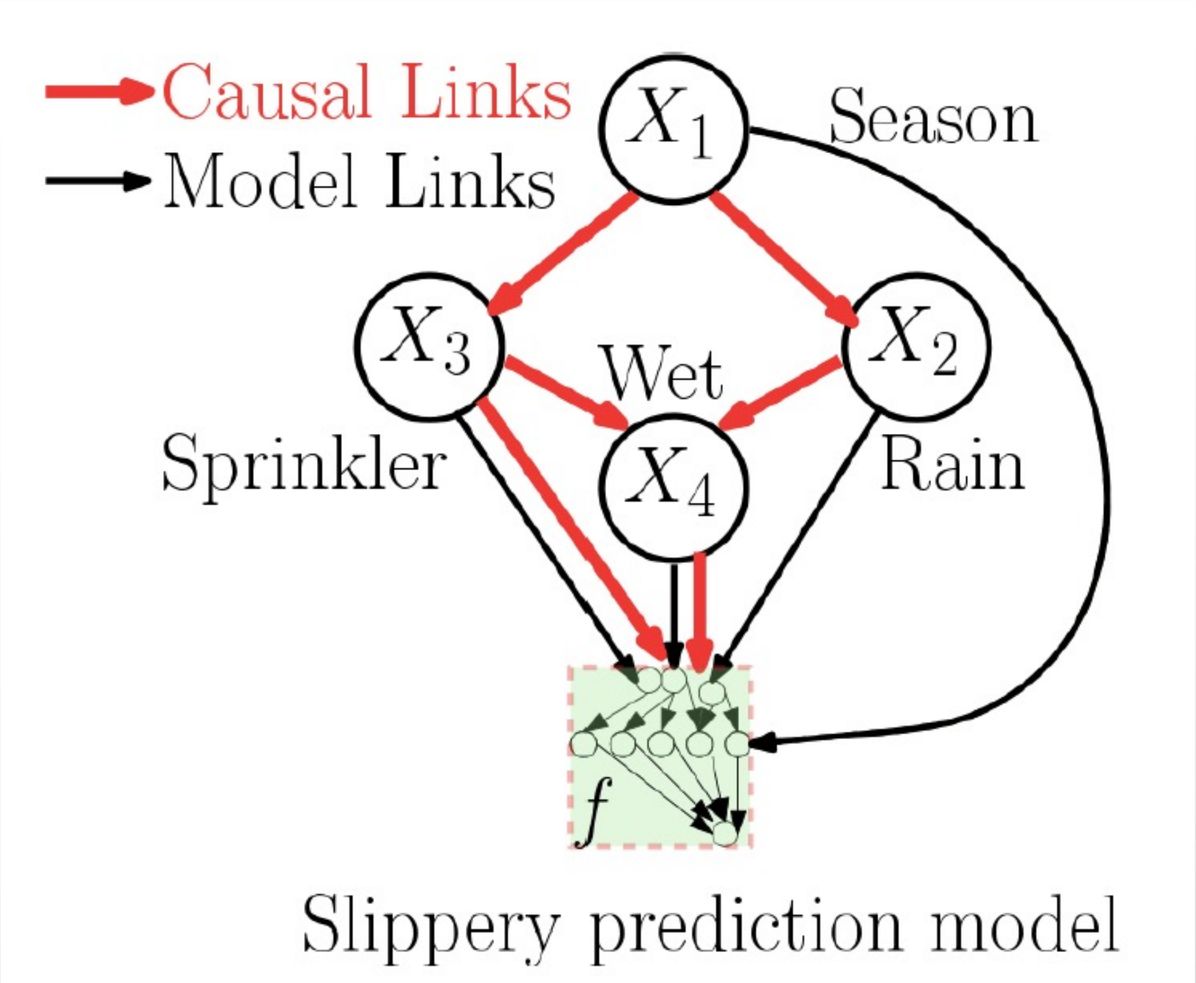
*Our key contributions are as follows.*

- *We propose **the first** (to the best of our knowledge) **generalization of Shapley value feature attribution to graphs**, providing a complete system-level view of a model.*

- *Our approach **unifies three previous game theoretic approaches** to estimating feature importance.*

- *Through **examples** on real data, we demonstrate how our approach facilitates understanding feature importance.*

# Przykład




Judea Pearl, *Causality*

# Jak poradziłyby sobie inne metody?

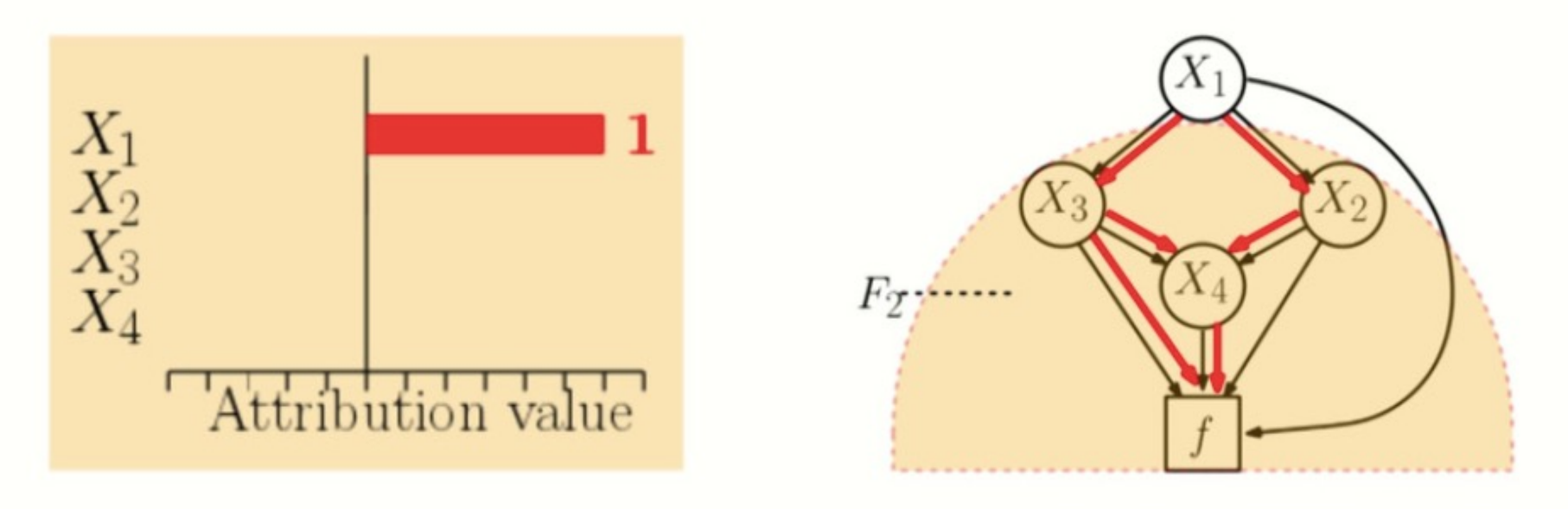
## Independent SHAP


Nie pokazuje **pośredniego** wpływu zmiennych.

S. Lundberg, S.-I. Lee, *A unified approach to interpreting model predictions*

# Jak poradziłyby sobie inne metody?

## ASV (Asymmetric Shapley Values)

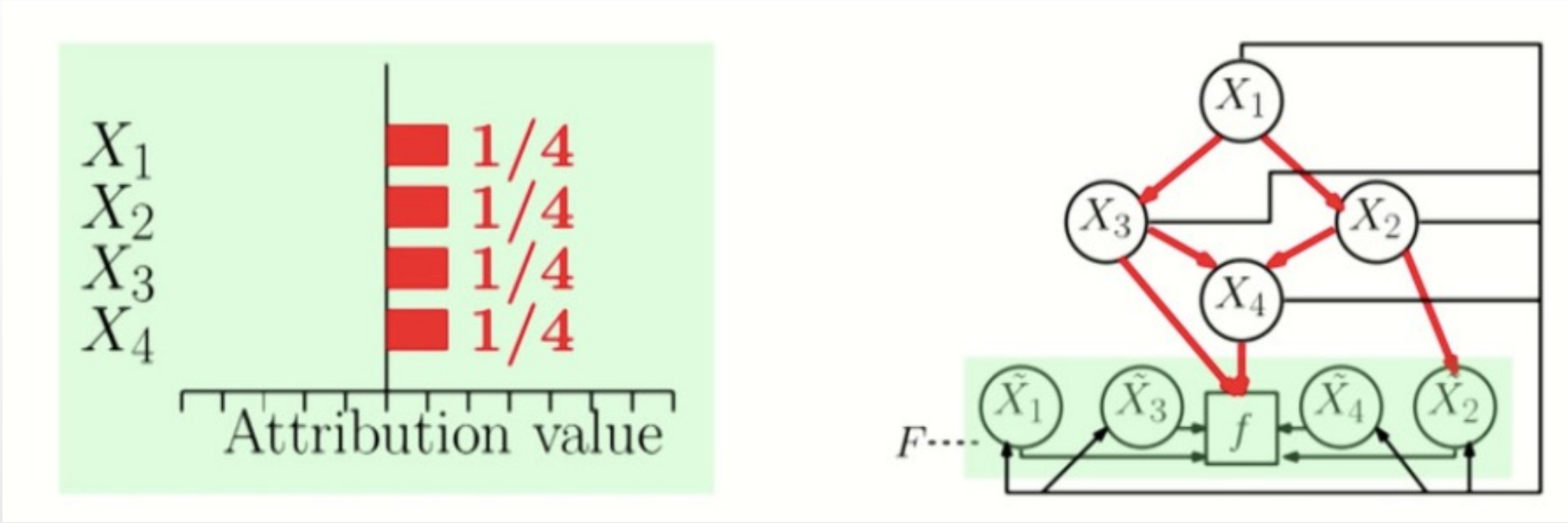



Nie pokazuje **bezpośredniego** wpływu zmiennych.

C. Frye et al., *Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability*

# Jak poradziłyby sobie inne metody?

## On-manifold SHAP




Ważność nie opisuje wprost wpływu żadnej ze zmiennych.

K. Aas, *Explaining individual predictions when features are depen*dent: More accurate approximations to shapley values

# Jak zrobić to *lepiej*?

## Shapley Flow

- metoda, które bierze pod uwagę związki przyczynowo-skutkowe między zmiennymi, uwzględniając jednocześnie zarówno bezpośrednie i pośrednie wpływy

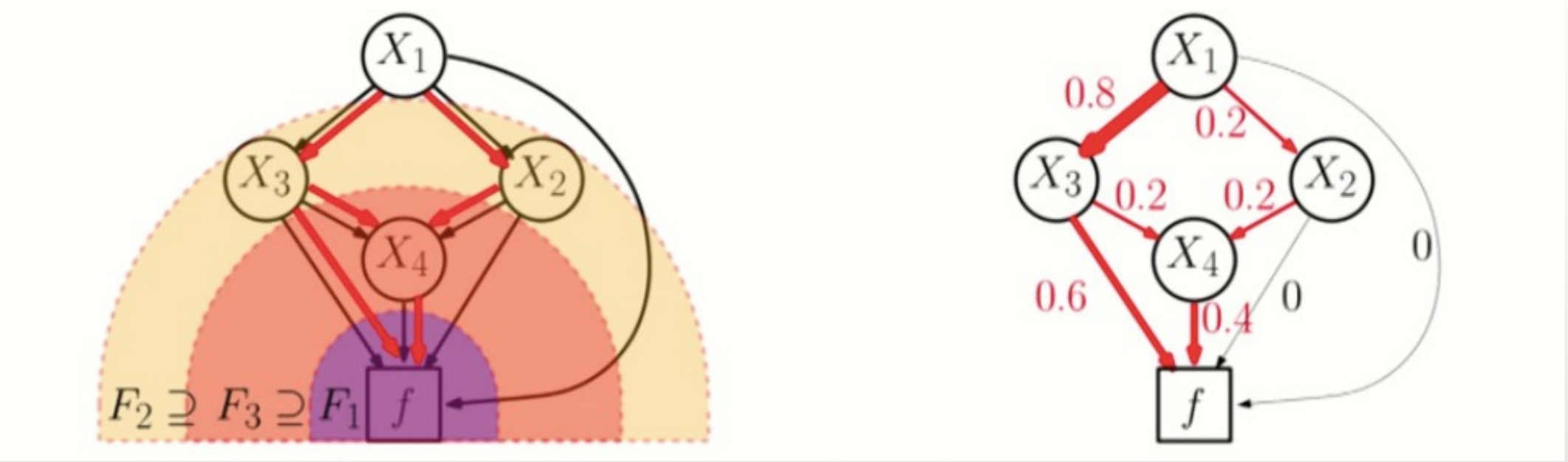
- przeformułowanie problemu:

przypisanie ważności wierzchołkom causal grafu

przypisanie creditsów krawędziom causal grafu

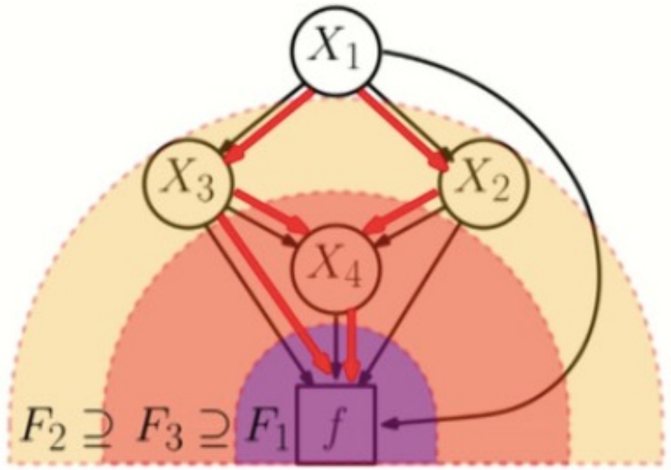
# Jak zrobić to *lepiej*?

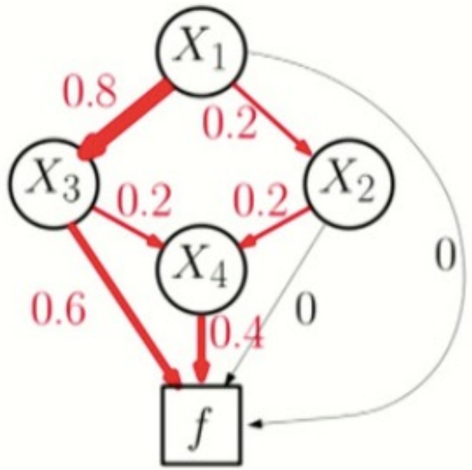
## Shapley Flow



# Shapley Flow

## Granice wyjaśnienia

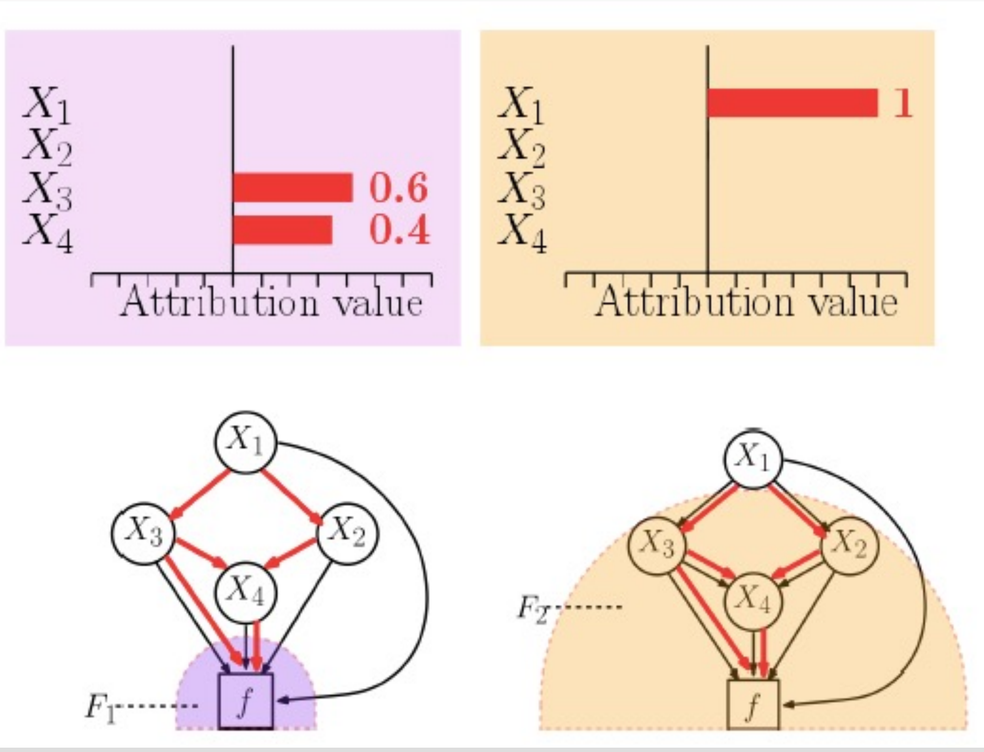



Przyjęcie innej granicy wyjaśnienia (*boundary of explanation)* daje inne rezultaty.

True to the model or true to the data? → Inne podejście przekłada się na inną granicę wyjaśnień.

Taki rezultat wyjaśnienia (na grafie) eliminuje potrzebę wielu wyjaśnień (barplotów), jednocześnie dając szeroki, holistyczny obraz sytuacji.



# Shapley Flow

## Intuicja

***Shapley Flow is the unique assignment of credit to edges***
***such that a relaxation of the classic Shapley value axioms are satisfied***
***for all possible boundaries of explanation.***

# Shapley Flow

## Prawdziwa Intuicja

~~*Shapley Flow is the unique assignment of credit to edges such that a relaxation of the classic Shapley value axioms are satised for all possible boundaries of explanation.*~~

**Krawędź jest istotna, jeśli jej usunięcie spowoduje dużą zmianę w predykcji modelu.**

**Co to znaczy „usunąć krawędź”?**
Usunięcie krawędzi polega na tym, że z wierzchołka źródłowego do wierzchołka docelowego „nie dochodzi informacja” o jego wartości zmiennej.

Przy czym operacje na krawędziach nie są od siebie niezależne i nie wykonujemy tylko jednej pojedynczej operacji. Bierzemy pod uwagę wszystkie możliwe scenariusze (historie) opisujące sposób zmiany wartości zmiennych.

**Wartość Shapley Flow dla krawędzi jest różnicą w predykcji modelu po jej usunięciu, uśrednioną dla wszystkich historii, które są *boundary consistent*.**

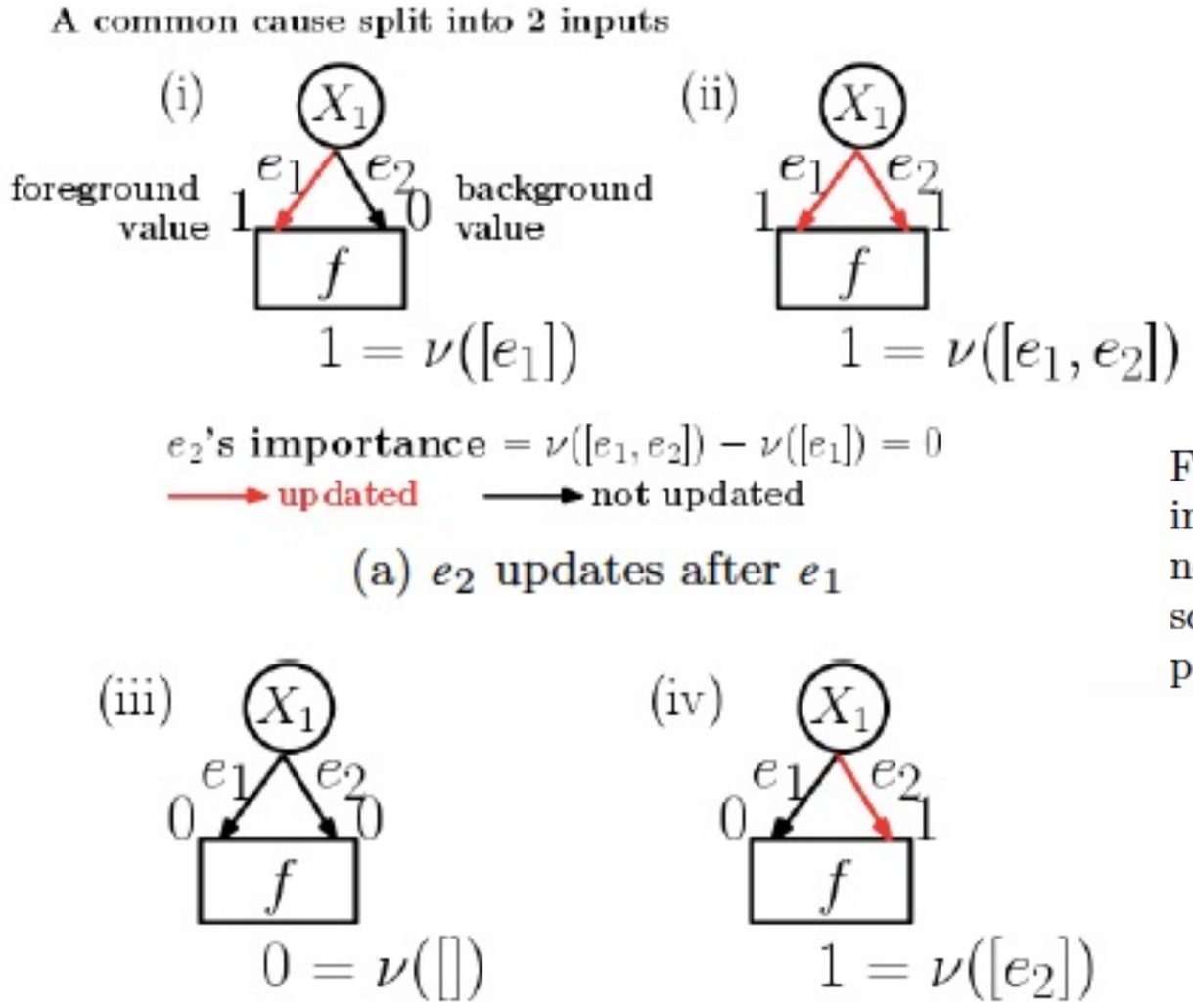



Figure 3: Edge importance is measured by the change in output when an edge is added. When a model is non-linear, say $f = OR$, we need to average over all scenarios in which $e_2$ can be added to gauge its importance. **Section 3.1** has a detailed discussion.

# Shapley Flow

## Algorytm

- Bazuje na przeszukiwaniu w głąb (DFS) – bierzemy pod uwagę wszystkie (lub wybrane losowo) konfiguracje kolejności odwiedzania wierzchołków (ścieżki poszukiwania) począwszy od źródła.

- Dla każdej konfiguracji przetwarzamy krawędzie w odpowiadającej kolejności, tzn. aktualizujemy wartość wierzchołka docelowego poprzez przekazanie wartości wierzchołka źródłowego.

- Jeśli wierzchołek docelowy jest ujściem, różnica w wartości odpowiedzi jest przypisywana każdej krawędzi na ścieżce poszukiwania od źródła do ujścia.

- Końcowy rezultat jest średnią z przypisanych krawędziom wartości we wszystkich przeanalizowanych konfiguracjach.

**Algorithm 1** Shapley Flow pseudo code

**Input:** A computational graph $\mathcal{G}$ (each node $i$ has a function $f_i$), foreground sample $\mathbf{x}$, background sample $\mathbf{x}'$
**Output:** Edge attribution $\phi : E \to \mathbb{R}$
**Initialization:**
$\mathcal{G}$: add an new source node pointing to original source nodes.

**function** SHAPLEYFLOW($\mathcal{G}$, $\mathbf{x}'$, $\mathbf{x}$)
    INITIALIZE($\mathcal{G}$, $\mathbf{x}'$, $\mathbf{x}$) ▷ Set up game $\nu$ for any boundary in $\mathcal{G}$
    $s \leftarrow$ SOURCE($\mathcal{G}$) ▷ Obtain the source node
    **return** DFS($s$, {}, [])
**end function**

**function** DFS($s$, $D$, $S$)
    ▷ $s$ is a node, $D$ is the data side of the current boundary, $S$ is coalition
    ▷ Using Python list slice notation
    Initialize $\phi$ to output 0 for all edges
    **if** IsSINKNODE(s) **then**
        ▷ Here we overload $D$ to refer to its boundary
        $\phi(S[-1]) \leftarrow \nu_D(S) - \nu_D(S[:-1])$ ▷ Difference in output is attributed to the edge
        **return** $\phi$
    **end if**

    **for** $p \leftarrow$ AllOrderings(Children($s$)) **do** ▷ Try all orderings/permutations of the node's children
        **for** $c \leftarrow p$ **do** ▷ Follow the permutation to get the node one by one
            edgeCredit $\leftarrow$ DFS($c$, $D \cup \{s\}$, $S + [(s, c)]$) ▷ Recurse downward

            $\phi \leftarrow \phi + \frac{\text{edgeCredit}}{\text{NumChildren}(s)!}$ ▷ Average attribution over number of runs
            $\phi(S[-1]) \leftarrow \phi(S[-1]) + \frac{\text{edgeCredit}(s,c)}{\text{NumChildren}(s)!}$ ▷ Propagate upward
        **end for**
    **end for**
    **return** $\phi$
**end function**

# Shapley Flow

## Generalizacja innych podejść

- **SHAP:** jeśli causal graf nie zawiera krawędzi pomiędzy wierzchołkami odpowiadającymi zmiennym, ale wszystkie z nich są połączone do wierzchołka odpowiadającego modelowi
- **ASV:** ważności wierzchołków źródłowych, jeśli wszystkie zależności między zmiennymi są modelowane za pomocą causal grafu
- **Owen value** (inne podejście z teorii gier)**:** wartości krawędzi prowadzących do liści, jeśli causal graf jest drzewem

# Shapley Flow

## Przykłady

- Sztucznie wygenerowany causal graph i oparty na nim zbiór danych:
  - graf o 10 wierzchołkach z liniowymi funkcjami o współczynnikach z N(0,1)
  - krawędzie wygenerowane losowo

- Zbiór danych National Health and Nutrition Examination Survey (18 zmiennych)
  - causal graph wygenerowany na podstawie ograniczonej wiedzy domenowej

- Modele: regresja liniowa i XGBoost (100 drzew, maks. głębokość 3)


(a) Causal graph for the nutrition dataset

# Przykłady – poprawność wyjaśnień dla liniowych modeli

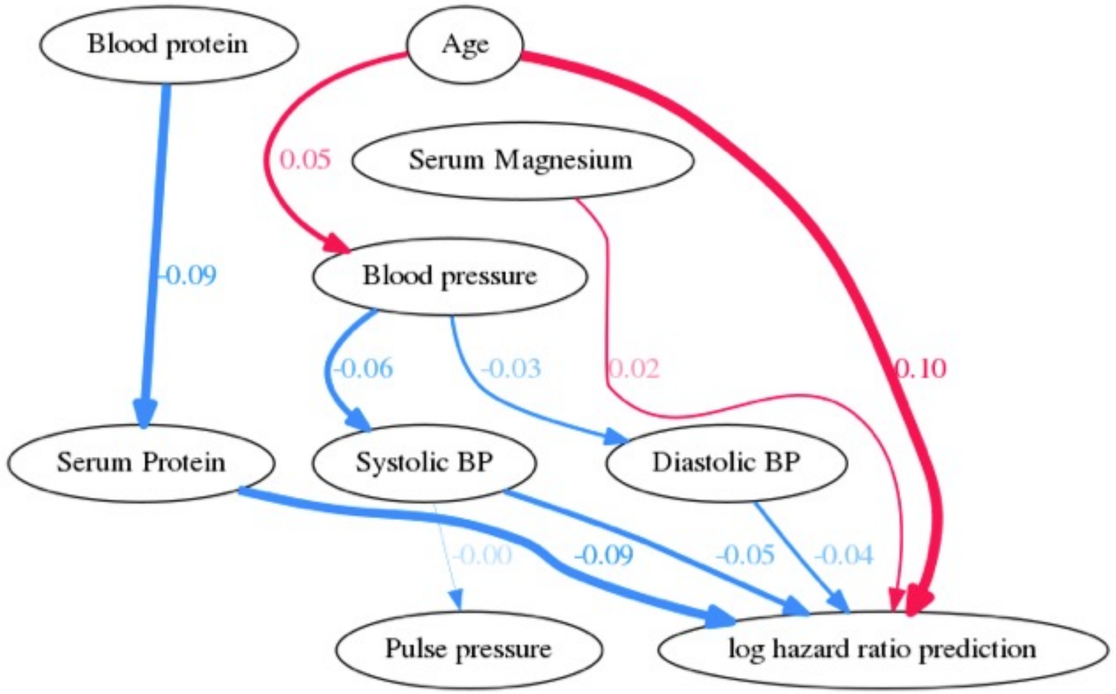
| Methods | Nutrition (**D**) | Synthetic (**D**) | Nutrition (**I**) | Synthetic (**I**) |
|---|---|---|---|---|
| Independent | **0.0** (± 0.0) | **0.0** (± 0.0) | 0.8 (± 2.7) | 1.1 (± 1.4) |
| On-manifold | 1.3 (± 2.5) | 0.8 (± 0.7) | 0.9 (± 1.6) | 1.5 (± 1.5) |
| ASV | 1.5 (± 3.3) | 1.2 (± 1.4) | 0.6 (± 1.9) | 1.1 (± 1.5) |
| Shapley Flow | **0.0** (± 0.0) | **0.0** (± 0.0) | **0.0** (± 0.0) | **0.0** (± 0.0) |

Table 1: Mean absolute error (std) for all methods on direct (**D**) and indirect (**I**) effect for linear models. Shapley Flow makes no mistake across the board.

# Przykłady – wyjaśnienia dla nieliniowych modeli

| Top features | Age | Serum Magnesium | Serum Protein |
|---|---|---|---|
| Background sample | 35 | 1.37 | 7.6 |
| Foreground sample | 40 | 1.19 | 6.5 |

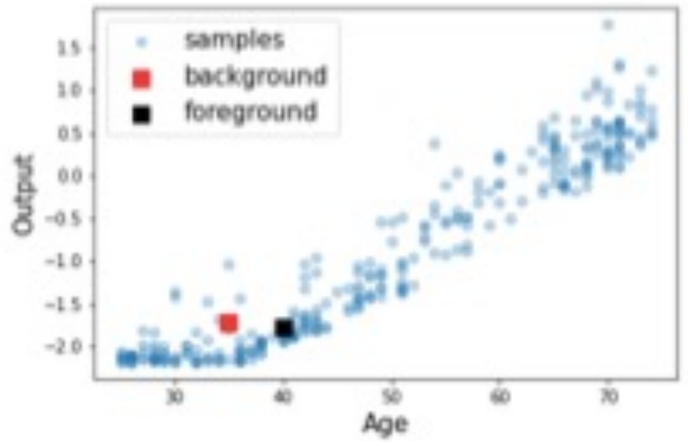
| Attributions | Independent | On-manifold | ASV |
|---|---|---|---|
| Age | 0.1 | -0.26 | 0.16 |
| Serum Magnesium | 0.02 | 0.2 | 0.02 |
| Serum Protein | -0.09 | 0.07 | 0.0 |
| Blood pressure | 0.0 | 0.0 | -0.14 |
| Systolic BP | -0.05 | -0.05 | 0.0 |
| Diastolic BP | -0.04 | -0.07 | 0.0 |
| Serum Cholesterol | 0.0 | -0.15 | 0.0 |
| Serum Albumin | 0.0 | -0.14 | 0.0 |
| Blood protein | 0.0 | 0.0 | -0.08 |
| White blood cells | 0.0 | 0.11 | 0.0 |
| Race | 0.0 | 0.09 | 0.0 |
| BMI | -0.0 | 0.08 | -0.0 |
| TIBC | 0.0 | 0.06 | 0.0 |
| Sex | 0.0 | -0.05 | 0.0 |
| TS | 0.0 | 0.05 | 0.0 |
| Pulse pressure | 0.0 | -0.05 | 0.0 |
| Poverty index | 0.0 | 0.04 | 0.0 |
| Red blood cells | 0.0 | 0.03 | 0.0 |
| Serum Iron | 0.0 | -0.02 | 0.0 |
| Sedimentation rate | 0.0 | 0.0 | 0.0 |
| Iron | 0.0 | 0.0 | -0.0 |
| Inflamation | 0.0 | 0.0 | 0.0 |



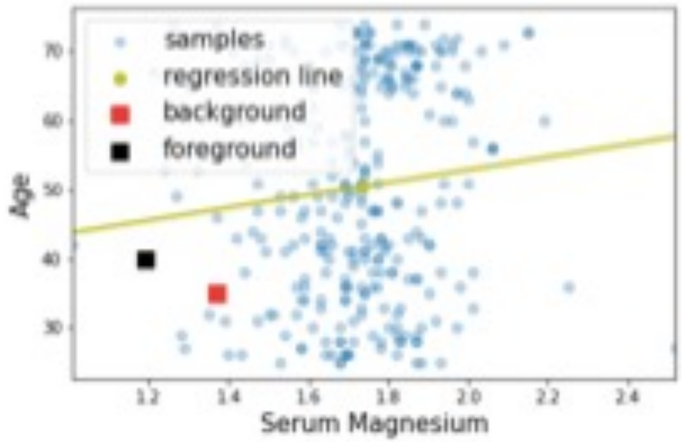

(a) Shapley Flow

Figure 6: Comparison among baselines on a sample (top table) from the nutrition dataset, showing top 10 features/edges.

# Przykłady – wyjaśnienia dla nieliniowych modeli (on-manifold SHAP)




(a) Age vs. output




(b) Age vs. magnesium

Figure 7: Age appears to be protective in on-manifold SHAP because it steals credit from other variables.

# Shapley Flow

## Co z causal graphami?

***While our approach relies on access to a complete causal graph, Shapley Flow is still valuable because:***

a) *there are well-established causal relationships in domains such as healthcare, and ignoring such relationships can produce confusing explanations;*

b) *recent advancements in causal estimation are complementary to our work and make defining these graphs easier;*

c) *finally and most importantly, existing methods already implicitly make causal assumptions, Shapley Flow just makes these assumptions explicit.*

# Shapley Flow

## Co dalej?

- *Can Shapley Flow work with partially defined causal graphs?*
- *How to explore Shapley Flow attribution when the causal graph is complex?*
- *Can Shapley Flow be useful for feature selection?*

# Konkluzja

triplot lepszy 😎