

# Wyjaśnialność systemów uczących się

wykład 15

Jerzy Stefanowski

Instytut Informatyki PP

2021 – update 2024

Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI-TECH)  
projekt finansowany z środków Programu Operacyjnego Polska Cyfrowa  
POPC.03.02.00-00-0001/20



**Fundusze  
Europejskie**  
Polska Cyfrowa



**Rzeczpospolita  
Polska**

**Unia Europejska**  
Europejski Fundusz  
Rozwoju Regionalnego



# Plan wykładu

- Motywacje do wyjaśnialnej sztucznej inteligencji
- Terminologia
- Klasyfikacja rodzajów wyjaśnień
- Klasyfikacja podstawowych metod
- Wybrane metody XAI
  - Ocena ważności atrybutów
  - LIME
  - SHAP
  - Kontrfakty
- Podsumowanie

# Spojrzenie na rozwój AI / ML

## Obserwacje z ostatnich kilkudziesięciu lat

- Gwałtowny rozwój metod ML, w szczególności głębokiego uczenia
- Przejście z środowiska akademickiego do fazy technologicznej produktów/ środowisk wykorzystywanych praktycznie / zainteresowanie przemysłu i biznesu
- Spektakularne zastosowania systemów ML/DANN z wysoką zdolnością predykcyjną (obrazy, wizja komputerowa, teksty, tłumaczenie języka, rozpoznawanie mowy,..)
- Lecz są to b. złożone systemy typu „black box” w zakresie oferowania informacji jak wewnątrz działają, czego się nauczyły i jak doszły do konkretnej decyzji

# Wyjaśnialność inteligentnych systemów

Wzrost zainteresowania tzw. XAI – ang. explainable AI

Próba definicji – cyt ang. [F.Giannotti et al]:

Explainable-AI explores and investigates methods to produce or complement AI models to **make accessible and interpretable** the internal logic and the outcome of the algorithms, making such process **understandable by humans**.

Działania w celu wyjaśnienie zasad działania systemu AI oraz rezultatów ich działania

Zarówno dla ekspertów, jak nie wyspecjalizowanych użytkowników

# Motywacje – ograniczenia automatycznej predykcji

- Predykcja - to to jedna ze możliwych perspektyw ML, w rzeczywistych zastosowaniach inteligentnych systemów oczekujemy także innych kryteriów oceny
- Czy ludzie ufają tzw. black box models oferowanym przez obecne ML?
- Cytaty:
  - If a machine learning model performs well, why do not we just trust the model and ignore why it made a certain decision? “The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks.” (Doshi-Velez and Kim 2017)
  - Do you just want to know what is predicted? For example, the probability that a customer will churn or how effective some drug will be for a patient. Or do you want to know why the prediction was made and possibly pay for the interpretability with a drop in predictive performance? (Molnar book)

# Motywacje – ograniczenia automatycznej predykcji

Ponadto

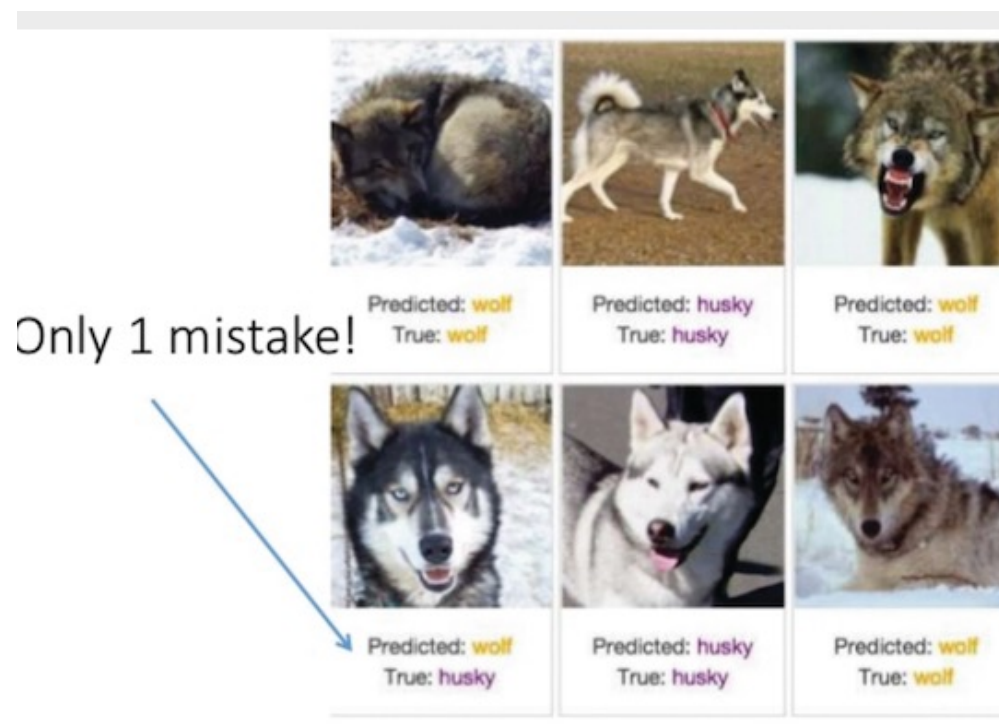
- Inne miary określają inne perspektywy spojrzenia na systemy uczące się oraz ich praktyczne wykorzystanie
- Zamknięty vs. otwarty świat – czy dysponujemy wystarczającymi danymi oraz przygotowaliśmy dogłębne testy?
- Jak radzić sobie z krytycznymi błędami systemu oraz zaskakującymi nietypowymi sytuacjami.
- Podatność na tzw. bias w danych i procesie uczenia się oraz niesprawiedliwość (ang. unfair) decyzji
- Mała odporność na tzw. ataki zewn. „hacking and adversarial attacks”

# Motywujące przykłady

- Cynthia Rudin et al – analiza działania systemu prawnego predykcji możliwości bycia w przyszłości recydywistą i zwolnień warunkowych w USA -> tzw. COMPAS system : błędne predykcje – przetrzymywania w więzieniu

Detale: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

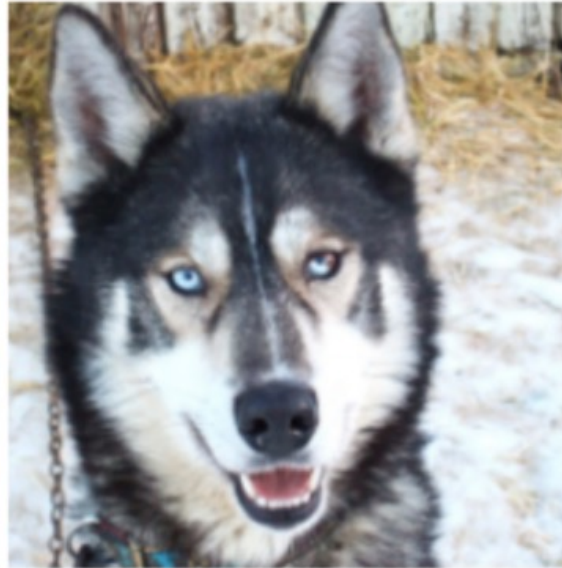
# Błędne działanie sieci CNN w rozpoznawaniu obrazów



Czy Syberian Husky jest podobny do wilka?



# Efekt tła obrazu



(a) Husky classified as wolf



(b) Explanation

Sieć nauczyła się tła – śniegu na którym występowały wilki

Literatura: liczne inne przykłady -DNN nie uczą się właściwych cech

# „Krytyczne” systemy

- Systemy sterujące urządzeniami, pojazdami, oraz o bardzo ważnym znaczeniu dla sprawnego działania ważnych organizacji
  - System musi być wysoce niezawodny i zachowywać tę niezawodność w różnorodnych i częściowo nieprzewidywalnych sytuacjach.

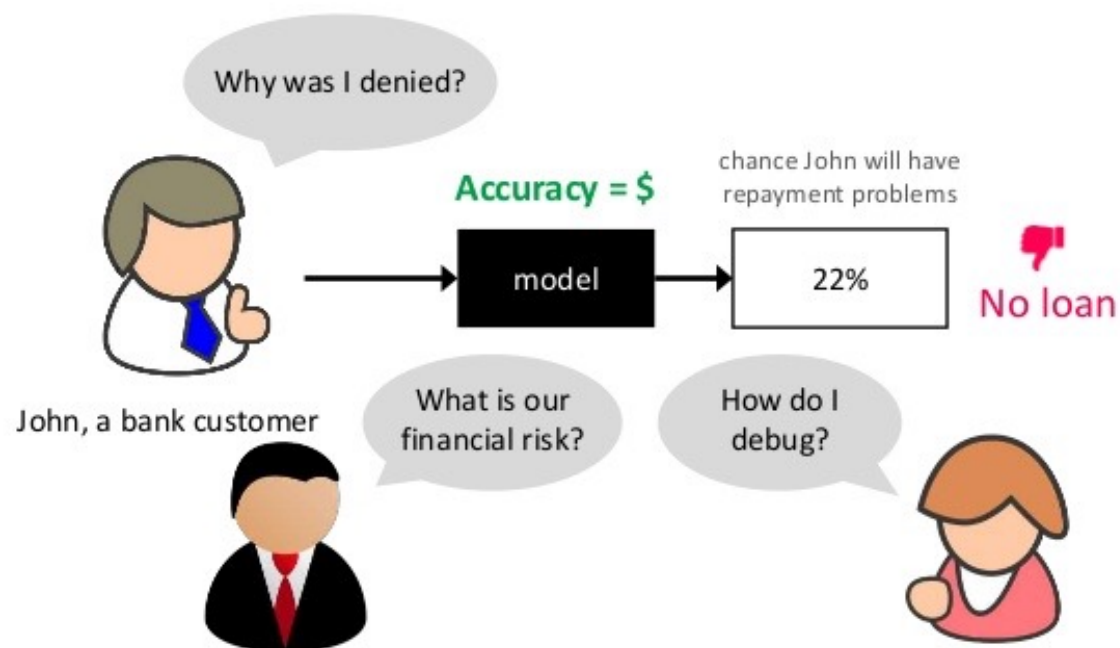


- Rygorystyczne wymagania wobec ich analizy i weryfikacji poprawności działania
  - Wyjaśnialność może pomóc w analizie błędów lub sytuacji odpowiadającym „adversarial attacks”

# Potrzeby udzielania wyjaśnień – podejścia regulacyjne

Ocena ryzyka kredytowego, dostęp do produktów finansowych, ubezpieczenia = często automatyzowanie procesu oceny wniosku klienta

Także decyzje administracyjne



Za wykład Scott Lundberg'a h2o world nyc 2019

# Prawo obywateli EU do wyjaśnień działania automatycznych systemów

(...) Algorytmy których decyzje są związane z predykcją użytkowników (ludzi) i mają znaczący wpływ na nich, powinny dostarczać wyjaśnień”

Dokumenty z 2016 nt. algorytmicznego podejmowania decyzji  
+2018 EU GDPR → a right to explanation

„a user can ask for an explanation of an algorithmic decision that was made about them”

2019 The high level expert group AI – the ethics guidelines for trustworthy AI



# Potrzeby wynikające z zastosowań ML

- Interpretowalność modeli uczenia maszynowego jest niezbędna dla **pozyskania zaufania ludzi** wobec takich systemów, zwłaszcza jeśli automatyczne podjęte decyzje są zaskakujące, nawet dla ekspertów [Stefanowski, Woźniak]
- Samek, Muller zauważają potrzeby wyjaśnialności w stosowaniu systemów predykcyjnych:
  - Weryfikowalność poprawności działania systemu predykcyjnego [przykład wykrycia niedoskonałości danych w problemie diagnozy zapalenia płuc].
  - Zrozumienie działania i wsparcie poprawy modelu
  - Uczenie się od systemów sztucznej inteligencji
  - Zgodność z prawodawstwem i postulatami regulacyjnymi

# Zdolność udzielania wyjaśnień o pracy system ML

System ML powinien móc udzielić wyjaśnień, dlaczego rekomenduje określoną decyzję.

- Rozróżnia się poziomy:
  - Funkcjonalnej interpretacji (jak i dlaczego model działa) i
  - Nisko-poziomowe algorytmiczne zrozumienie detali algorytmu, wpływu parametrów itd.
- Należy „powiązać atrybuty opisujące przykłady z wyjściem systemu w prosty, informatywny oraz posiadający znaczenie sposób”.
- Użytkownicy złożonych systemów muszą mieć możliwość odtworzenia całego procesu podejmowania decyzji

# Terminologia

- Rozróżnienie pomiędzy wyjaśnialnością (ang. Explainability) a interpretowalnością systemów ML – nie jest rygorystycznie precyzyjnie definiowane.
- F.Giannotti et al. artykuł:
  - **Explanation** – możliwości odpowiedzi na pytania “why”
  - **Interpretability** – opisanie wnętrza lub działania systemu w sposób potencjalnie zrozumiały dla człowieka (zależne od przygotowania wstępnego odbiorcy, jego wiedzy oraz kontekstu użycia)

## Elementy terminologiczne ad. Interpretowalność cd.

Pojęcie zrozumienia przez człowieka jest najważniejszych aspektów dobrej interpretowalności, lecz może zarówno dotyczyć zrozumienia procesu zbudowania modelu, jego ewentualnej reprezentacji, jak również zasad jego działania.

Wymienia się także inne dodatkowe pojęcia

- przejrzystość i czytelność [transparency],
- wierność odwzorowania zależności [fidelity],
- dopasowanie do zdolności poznawczych odbiorcy,
- zaufanie do modelu,

Pytanie o miary oceny spełnienia postulatów



# Wyjaśnienia dla kogo (Who)?

Nie ma uniwersalnych wyjaśnień!

Różni odbiorcy:

- Specjaliści ML
- Eksperci od zastosowania
- Użytkownicy (end user)
- Audytorzy/ regulatorzy

Odbiór wyjaśnienia w zależności od przygotowania odbiorcy i jego wiedzy

Patrz A.Weller Challenges for Transparency ICML 2017

A.Arrieta et al Explainable AI ... 2020 oraz F. Rossi, AI Ethics for Enterprise AI (2019)

# Kategoryzacja metod wyjaśnialności

- Explanation by design (wnętrze systemu)
- Black box explanation (post-hoc)
  - Model (próba wyjaśnienia podstaw całego systemu)
  - Outcome (przyczyny podjęcia specyficznej decyzji)
  - Inspection (na ogół wizualizacja graficzna fragmentu działania)
- Global vs. local explanations
- Model-specific vs. model agnostic

# Czytelne reprezentacje modeli ML

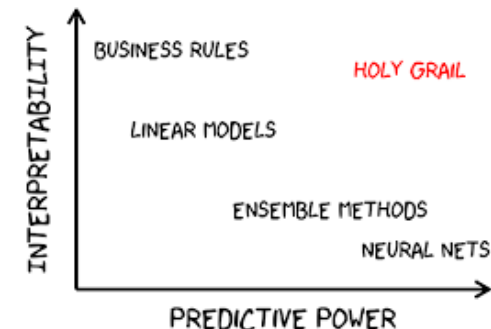
## Łatwiejsze do interpretacji

- Modele liniowe
- Reguły
- Drzewa
- K-NN
- Uogólnione model liniowe
- Modele Bayesowskie

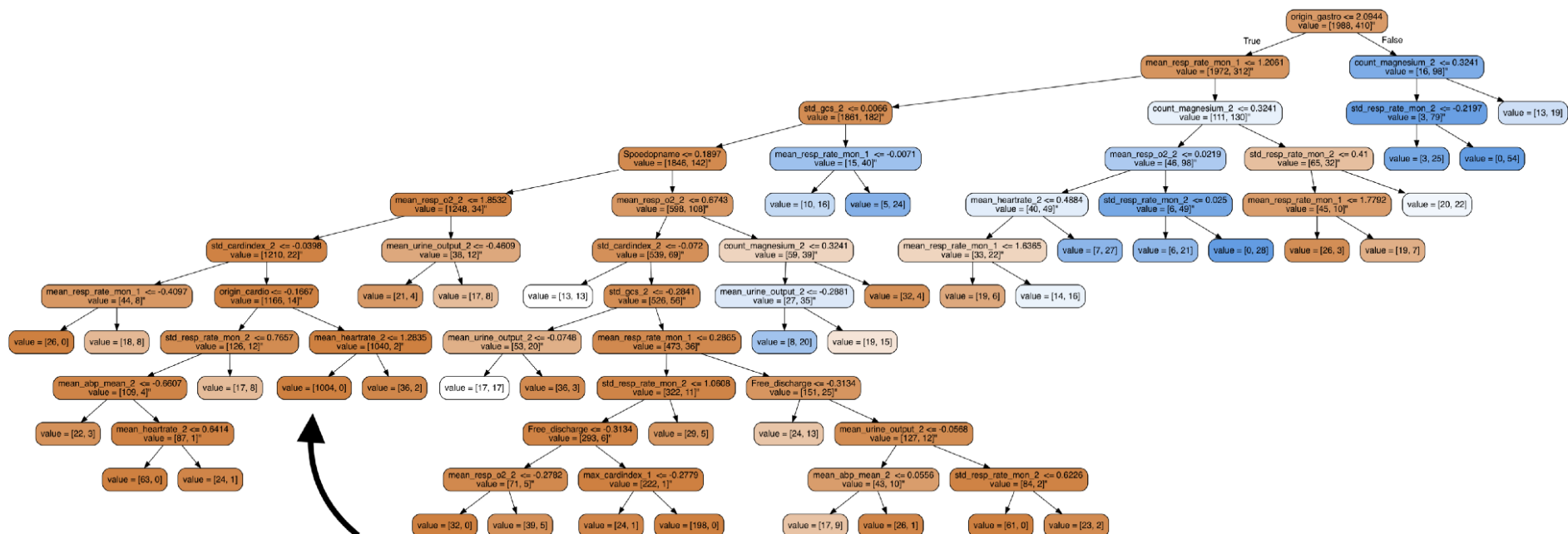
## Trudne do interpretacji

- SVM
- Zespoły modeli predykcyjnych (ensembles)
- ANN
- DNN
- Hybrydowe modele złożone

Przetarg między trafnością a złożonością  
/ potencjalną czytelnością modelu



# Przykład oceny stanu chorych IOM –VUMc Amsterdam (Pacmed system)



do not come from gastroenterology

had low respiratory rate on the first day of recovery

have a stable score in the Glasgow Coma Scale (measuring alertness and being a proxy for neurological disorders)

do not come from an emergency setting



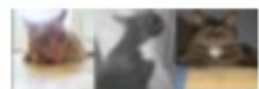
have respiratory measurements that are not worrisome at the moment

come from cardiology

have heart-rate measurements that are not worrisome at the moment

....

# Przykłady metod

TABULAR	IMAGE	TEXT																																																									
<p><b>Rule-Based (RB)</b></p> <p>A set of premises that the record must satisfy in order to meet the rule's consequence.</p> <p><math>r = Education \leq College</math> <math>\rightarrow \leq 50k</math></p>	<p><b>Saliency Maps (SM)</b></p> <p>A map which highlight the contribution of each pixel at the prediction.</p> 	<p><b>Sentence Highlighting (SH)</b></p> <p>A map which highlight the contribution of each word at the prediction.</p> <p>the movie is not that bad</p>																																																									
<p><b>Feature Importance (FI)</b></p> <p>A vector containing a value for each feature. Each value indicates the importance of the feature for the classification.</p> <table><tr><td>capitalgain</td><td>0.00</td></tr><tr><td>education-num</td><td>14.00</td></tr><tr><td>relationship</td><td>1.00</td></tr><tr><td>hourspersweek</td><td>3.00</td></tr></table>	capitalgain	0.00	education-num	14.00	relationship	1.00	hourspersweek	3.00	<p><b>Concept Attribution (CA)</b></p> <p>Compute attribution to a target “concept” given by the user. For example, how sensitive is the output (a prediction of zebra) to a concept (the presence of stripes)?</p> 	<p><b>Attention Based (AB)</b></p> <p>This type of explanation gives a matrix of scores which reveal how the word in the sentence are related to each other.</p> <table><tr><td></td><td>the</td><td>movie</td><td>is</td><td>not</td><td>that</td><td>bad</td></tr><tr><td>the</td><td>0.8</td><td>0.1</td><td>0.2</td><td>0.3</td><td>0.2</td><td>0.1</td></tr><tr><td>movie</td><td>0.1</td><td>0.9</td><td>0.1</td><td>0.2</td><td>0.3</td><td>0.2</td></tr><tr><td>is</td><td>0.2</td><td>0.1</td><td>0.8</td><td>0.1</td><td>0.2</td><td>0.1</td></tr><tr><td>not</td><td>0.3</td><td>0.2</td><td>0.1</td><td>0.9</td><td>0.2</td><td>0.1</td></tr><tr><td>that</td><td>0.2</td><td>0.3</td><td>0.2</td><td>0.2</td><td>0.8</td><td>0.1</td></tr><tr><td>bad</td><td>0.1</td><td>0.2</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.9</td></tr></table>		the	movie	is	not	that	bad	the	0.8	0.1	0.2	0.3	0.2	0.1	movie	0.1	0.9	0.1	0.2	0.3	0.2	is	0.2	0.1	0.8	0.1	0.2	0.1	not	0.3	0.2	0.1	0.9	0.2	0.1	that	0.2	0.3	0.2	0.2	0.8	0.1	bad	0.1	0.2	0.1	0.1	0.1	0.9
capitalgain	0.00																																																										
education-num	14.00																																																										
relationship	1.00																																																										
hourspersweek	3.00																																																										
	the	movie	is	not	that	bad																																																					
the	0.8	0.1	0.2	0.3	0.2	0.1																																																					
movie	0.1	0.9	0.1	0.2	0.3	0.2																																																					
is	0.2	0.1	0.8	0.1	0.2	0.1																																																					
not	0.3	0.2	0.1	0.9	0.2	0.1																																																					
that	0.2	0.3	0.2	0.2	0.8	0.1																																																					
bad	0.1	0.2	0.1	0.1	0.1	0.9																																																					
<p><b>Prototypes (PR)</b></p> <p>The user is provided with a series of examples that characterize a class of the black box</p> <p><math>p = Age \in [35, 60], Education \in [College, Master] \rightarrow “\geq 50k”</math></p> <p><math>p = </math>  <math>\rightarrow “cat”</math></p> <p><math>p = “... not bad ...” \rightarrow “positive”</math></p>																																																											

# Najpopularniejsze podejścia

- Feature summary statistics, evaluating their role, impact
- Tzw. PDP (partial dependency plots)
- Visualization (np. heat maps for ANN, ...)
- Model internal parameters (easier for typical interpretable models)
- Focus on some data elements (interpreting some data points / example based explanations) – tzw. Prototypy
- Kontrfakty
- Sensitivity / Perturbation analysis
- Use (or transform into) Intrinsically interpretable model: surrogate = zbliżone do post-hoc
- Specialized approaches to DeepANN and images (e.g.LRP Layer-Wise Relevance Propagation, decomposition)

Inne podziały:

Patrz literatura

# Ocena ważności atrybutów

- Przypomnij sobie np. feature importance w drzewach lub lasach (Breiman)

Najprostsza heurystyka analizy struktury drzewa – im wyżej i częściej występuje warunek z atrybutem

Oceń dla warunku redukcję miary (impurity) oraz wagę – liczbę przykładów w węźle

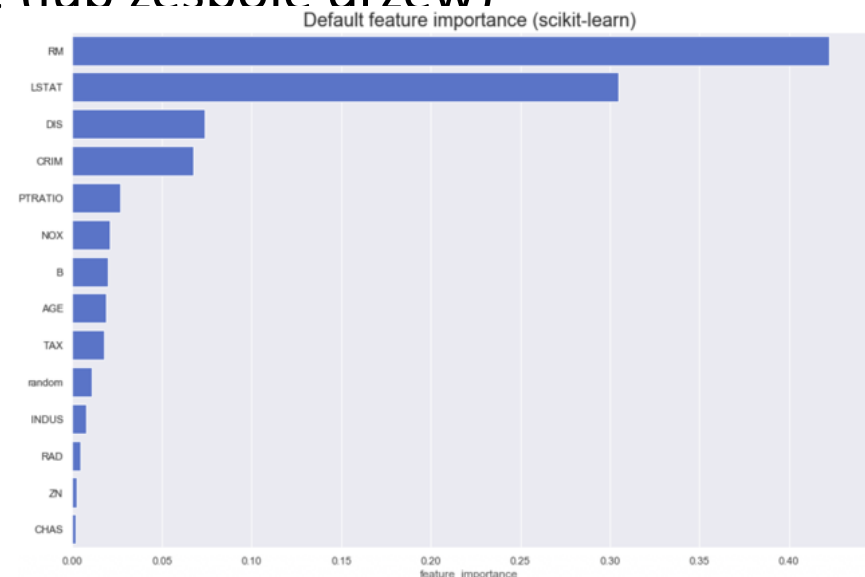
np. w węźle  $t$  dla atrybutu  $A$  spadek entropi  $\Delta(t)=0.1$  oraz  $N = 60$  przykładów, to ocena  $\text{atr } A = 0.1 * 60 = 6$

Sumuj takie wystąpienia w całym drzewie (lub zespole drzew)

Pro: – szybkie obliczenia

Cons: - przybliżone obliczenia i może nadmiernie faworyzować wielo\_ wartościowe oraz liczbowe atrybuty;

Brak interakcji pomiędzy cechami



# Permutations – “noised-up” method

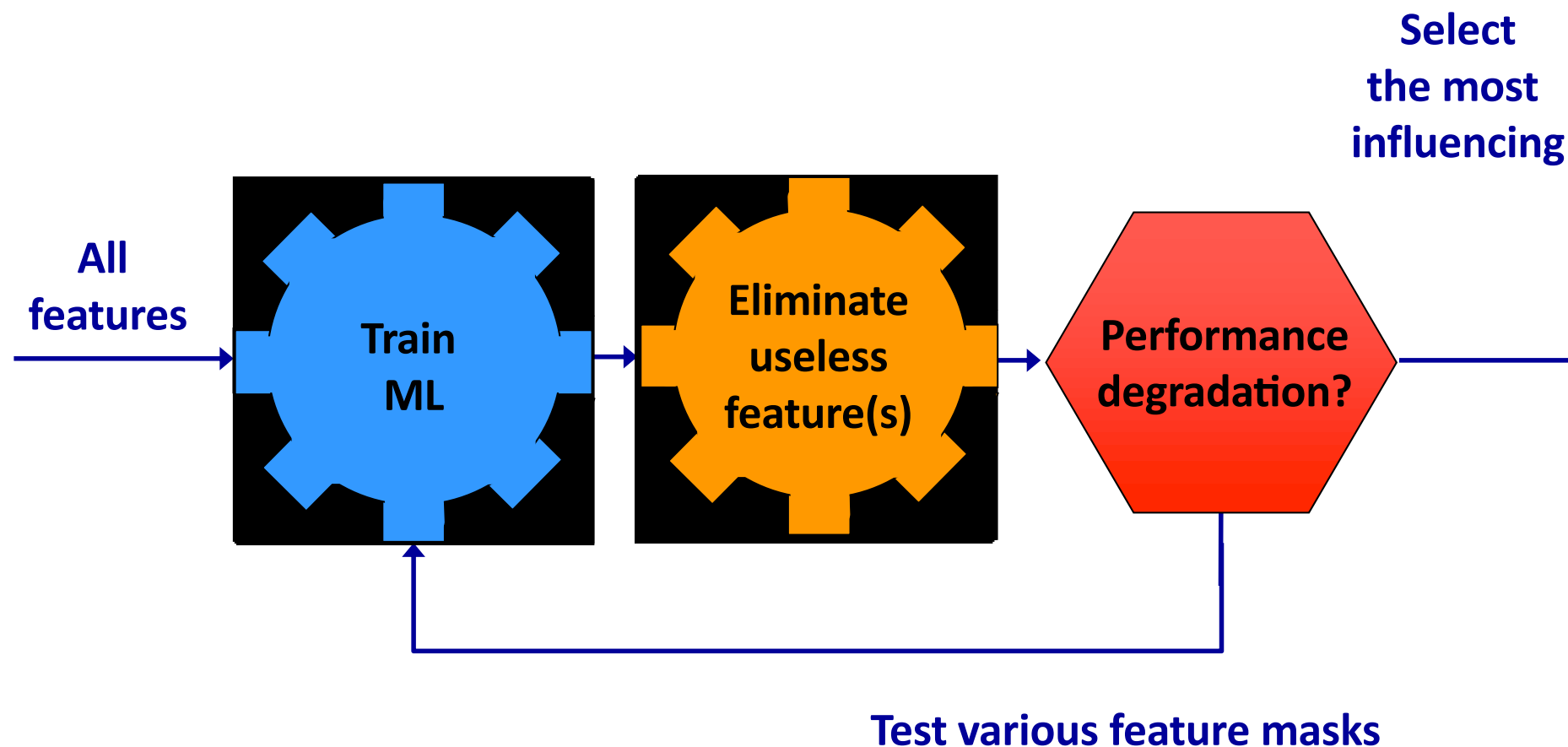
- L. Breiman wprowadził dla innego rankingu ważności cech poprzez analizę klasyfikowania w random forests / bagging
- Hipoteza – cecha jest tym ważniejsza, jeśli losowa zmiana jej wartości (permutacja) w zbiorze testowych mocno pogarsza trafność predykcji
- W bagging / RF – wykorzystuje się zbiór out of bag (OOB) dla każdego z drzew
  - najpierw ocenia się predykcję z oryginalnymi cechami  $O$ ;
  - następnie dla każdej z cech tworzy zastępczy zbiór  $O'$ :  
losowo permutuje się wartości cechy i ponownie testuje klasyfikatorzachowuje się informacje o predykcjach w każdej z tych sytuacji
- Po zbudowaniu całego zespołu:
  - Dla każdego z przykładów uczących  $z_i$  ( $1 \dots N$ ) – określ predykcje klasyfikatora używając pamięci testowania OOB jeśli zawierają  $z_i$  – i oblicz dec. zespołu; następnie łączny błąd klasyfikowania  $e$
  - Analogicznie dla każdej cechy  $x_j$  oblicz z pamięci OOB –  $O'$  błąd zespołu klasyfikatora
- Ważność cechy  $F_j = e_j - e$       Im większa, tym ważniejsza  $x_j$

Pro: b. wiarygodna; stosowalna także dla innych klasyfikatorów niż drzewa

Cons: przeszacowuje skorelowane atrybuty, kosztown., nie w podst. bibliotekach

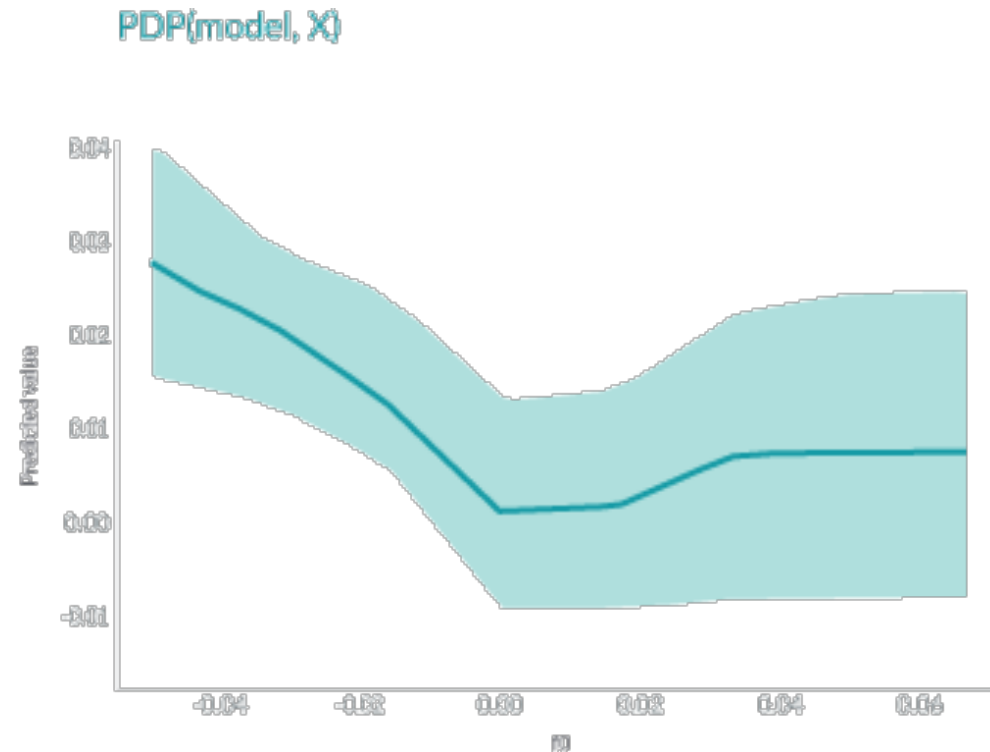


# Wrażliwość na wybrane cechy



Recursive Feature Elimination (RFE) SVM. *Guyon-Weston*, - also for ranking them

# Partial Dependency Plots

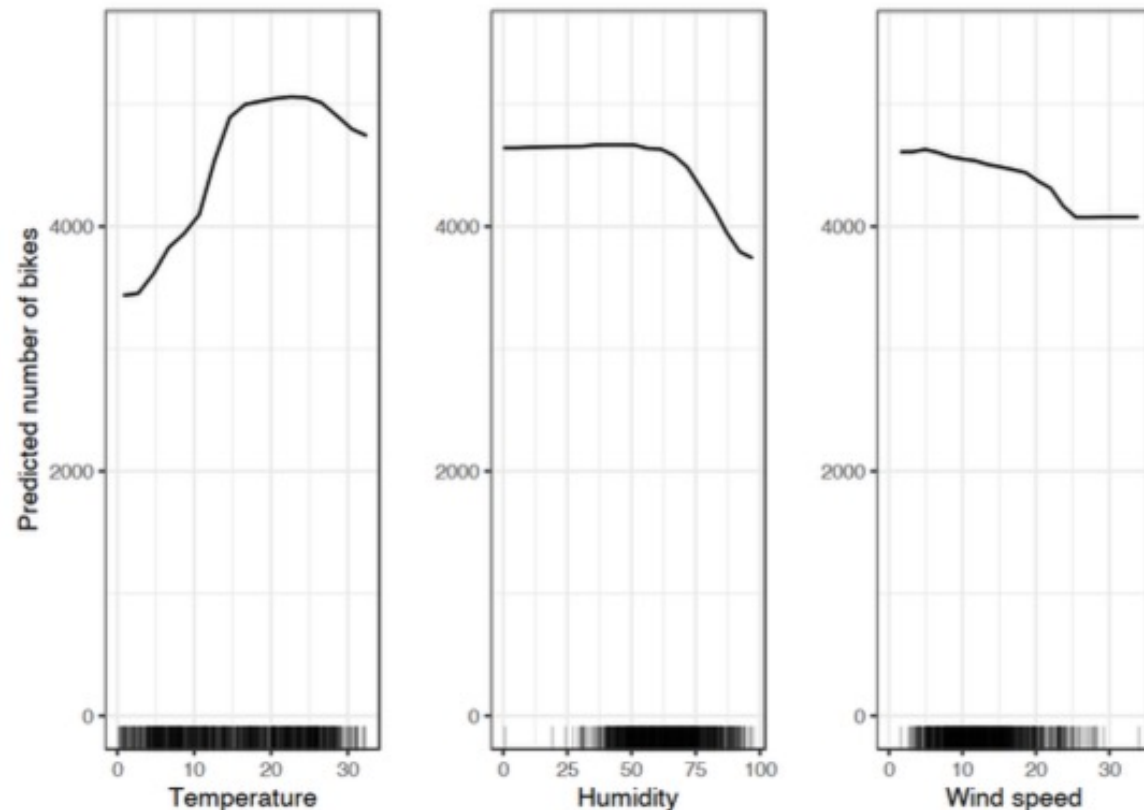
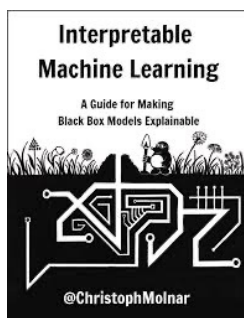


Wpływ zmian wartości jednej cechy na wyjście modelu / dla pojedynczej obserwacji “zamrażamy” inne cechy/ wynik zbiorczy wraz z zaznaczoną wartością średnią

Prosta analiza wrażliwości – lecz bez dogłębniejszej analizy cech i ich interakcji

# Przykład użycia PDP

The task of prediction regression of bicycles in the city depending on external conditions[see Molnar]

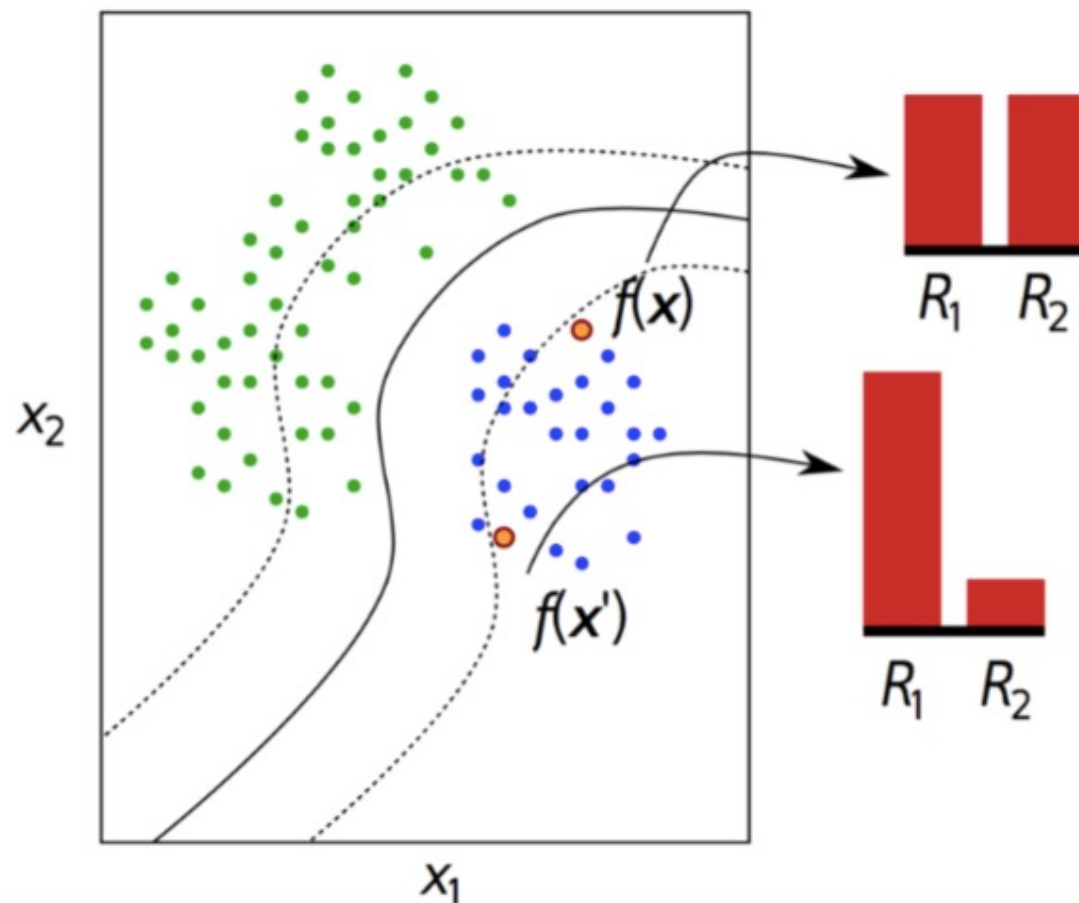


**FIGURE 7.1** PDPs for the bicycle count prediction model and temperature, humidity and wind speed. The largest differences can be seen in the temperature. The hotter, the more bikes are rented. This trend goes up to 20 degrees Celsius, then flattens and drops slightly at 30. Marks on the x-axis indicate the data distribution.

# Explaining Individual Decisions

---

**Goal:** Determine the relevance of each input variable for a given decision  $f(x_1, x_2, \dots, x_d)$ , by assigning to these variables *relevance scores*  $R_1, R_2, \dots, R_d$ .



## Model Explainability (degree of explainability)

1) linearity

2) monotonicity

make predictions easier & reliable

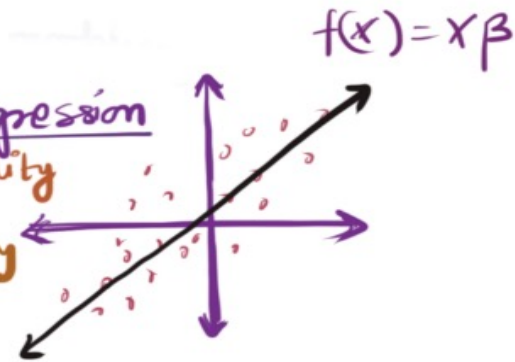
$f: X \rightarrow Y$  (mapping from  $x$  to  $y$  by  $f$ )

3) classification problem

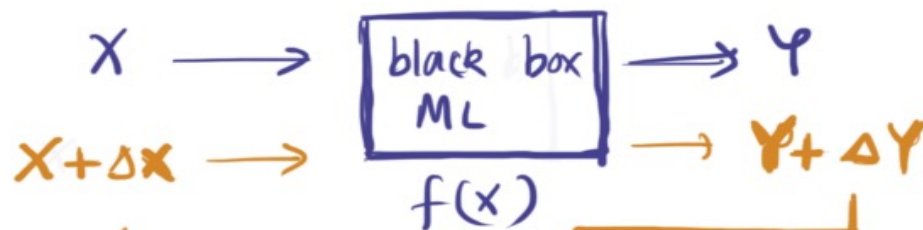


globally : complex boundary  
locally : linear boundaries

1) linear regression  
global fidelity  
= local fidelity



2) piecewise linear  
local fidelity  
≠ global fidelity



$\Delta X$  : small change in  $X$   
 $\Delta Y$  : small change in  $Y$

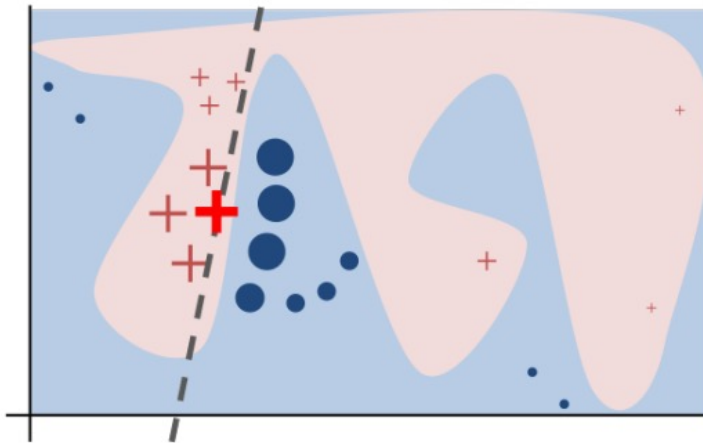


# LIME - Lokalne wyjaśnienie wpływu atrybutów na predykcje

- LIME – **L**ocal **i**nterpretable **m**odel agnostic **e**xplanations [Ribeiro et al KDD Conf. 2016]
- Globalne agnostyczne przybliżenie działania złożonego modelu  $f$  może być trudne
- Autorzy skupiają uwagę na przybliżeniu lokalnym

# LIME – ogólny schemat

- Podejście agnostyczne – lokalne – dla konkretnego przykładu  $x$ 
  - Skupienie zainteresowania na sąsiedztwie  $N(x)$  punktu  $x$



LIME (Ribeiro et. al.)

- Naucz model zastępczy  $g$  na podstawie danych z sąsiedztwa  $N(x)$  rozszerzone o losowe zaburzone punkty
- Liniowy model / regresji
- Interpretacja zmiennych w modelu linowym

# LIME – do czego dążymy?

Wyjaśnienie ważności cech / atrybutów i ich wartości dla predykcji modelu black box

- Lecz rozpatrywanych lokalnie



# LIME funkcja celu

Poszukuj modelu liniowego  $g$  o dużej lokalnie zgodności predykcji z “black box model”  $f$

$$\xi(x) = \operatorname{argmax}_{(g \in G)} L(f, g, \pi_x) + \Omega(g)$$

gdzie  $\pi_x$  ważona odległość punktów z sąsiedztwa do  $x$

$\Omega$  czynnik regularyzacyjny – złożoność liniowego modelu  $f$  w odniesieniu do zmiennych

Autorzy LIME stosowali LASSO i regresję liniową

# Ogólny schemat algorytmu

To find an explanation for a single data point and a given classifier

- Sample the locality around the selected single data point uniformly and at random and generate a dataset of perturbed data points with its corresponding prediction from the model  $f$  we want to be explained
- Use the specified feature selection methodology to select the number of features that is required for explanation
- Calculate the sample weights using a kernel function and a distance function. (this captures how close or how far the sampled points are from the original point)
- Fit an interpretable linear model on the perturbed dataset using the sample weights to weigh the objective function.
- Provide local explanations using the newly trained interpretable model

Więcej w artykule autorów oraz na blogu

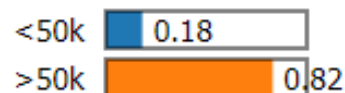
<https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html>

# Przykład ilustracyjny – dane o dochodach mieszkańców USA

```
marital_status      Married
education_num       Bachelors
hours_per_week      40
fnlwgt              167065
sex                 Male
age                 50
random              0.0412146
workclass_ Private  1
occupation_ Exec-managerial  1
race_ White         1
Name: 23706, dtype: object

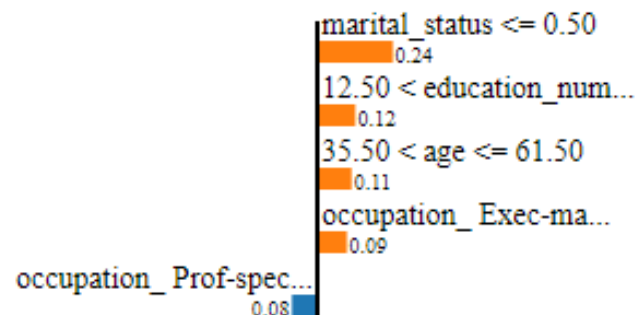
Label: >50K
Prediction: >50K
```

Prediction probabilities



<50k

>50k

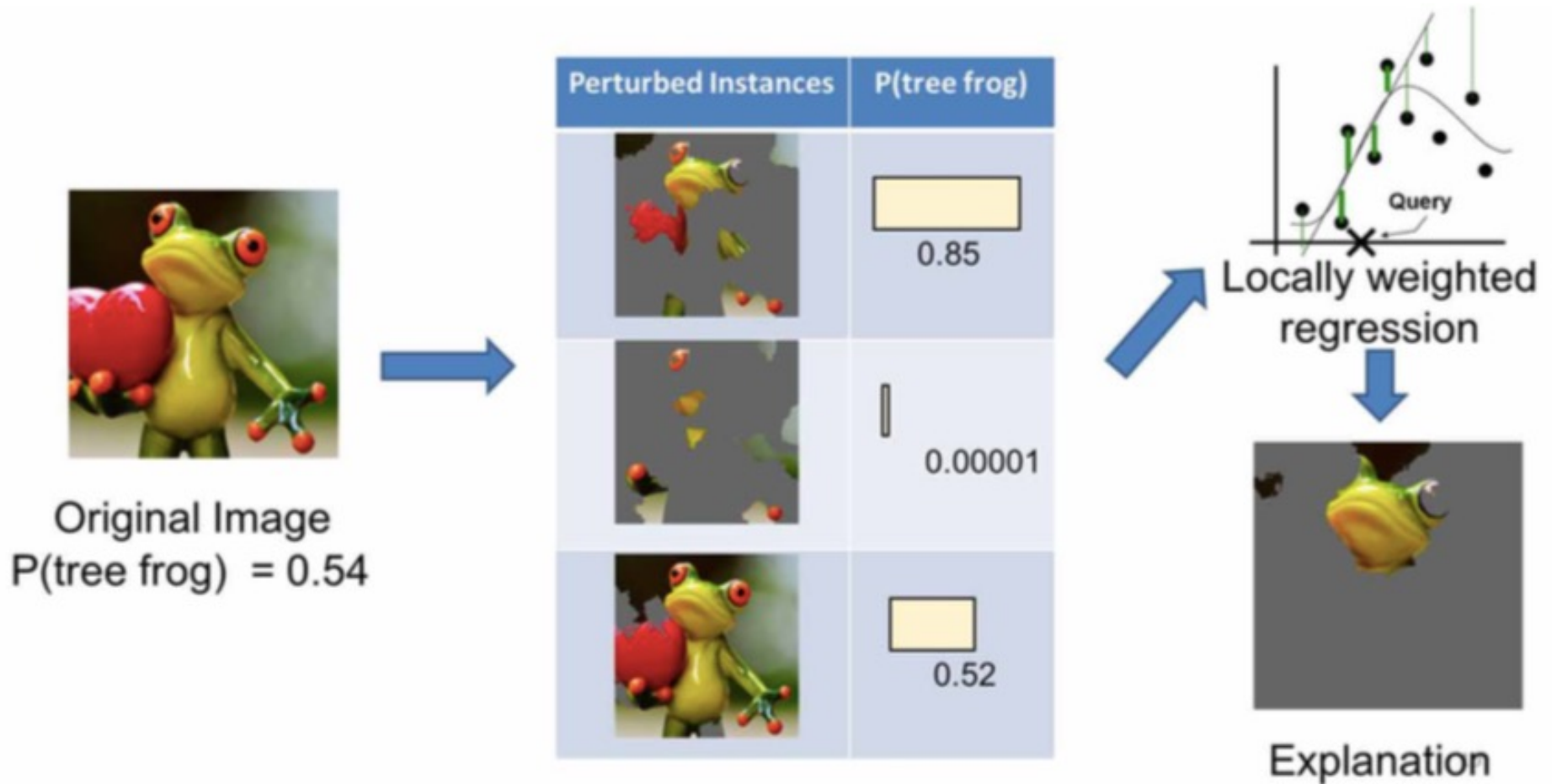


Feature Value

marital_status	0.00
education_num	13.00
age	50.00
occupation_ Exec-managerial	1.00
occupation_ Prof-specialty	0.00

Klasyfikacja : dwie kategorie dochodu; atrybuty opisujące prace, wykształcenie, działalność, rasę, itd.

# LIME – image recognition



# 20 newgroups text classification

Prediction probabilities

atheism	0.58
christian	0.42

atheism

christian

Posting: 0.15  
Host: 0.14  
NNTP: 0.11  
edu: 0.04  
com: 0.04  
have: 0.04  
There: 0.04

## Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)  
Subject: Another request for Darwin Fish  
Organization: University of New Mexico, Albuquerque  
Lines: 11  
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.

This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

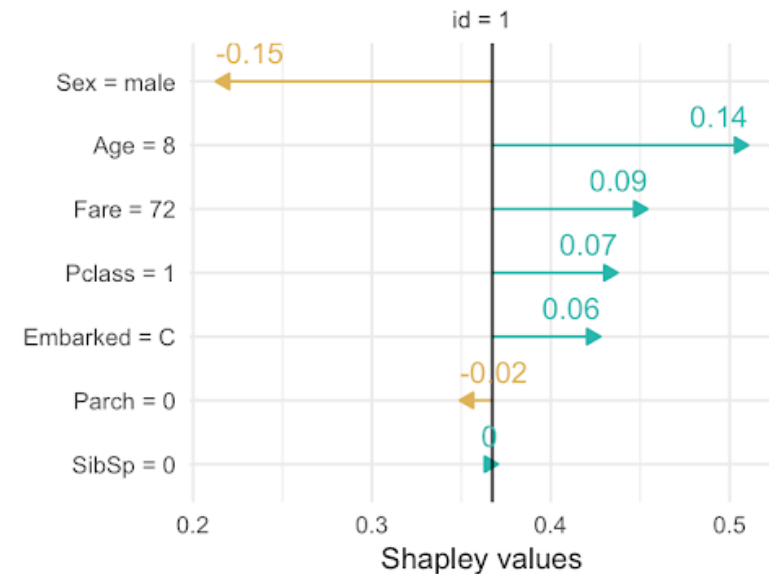
# SHAP

- Podejście agnostyczne do wyjaśniania modeli klasyfikacyjnych i regresyjnych
- Ocena globalna lub lokalna ważności / wkładu każdego atrybutu do predykcji
  - Wykorzystanie przybliżonych oszacowań wskaźników Shapley'a
  - Nacisk na ich efektywne obliczenie

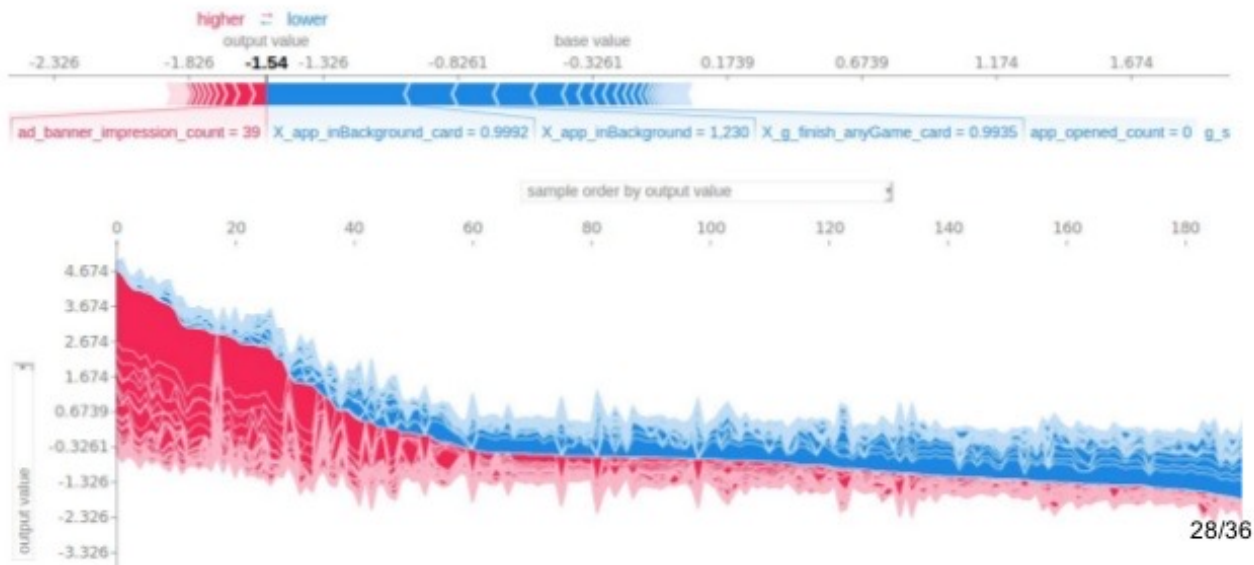
# Możliwe wyniki działania SHAP

Ocena wpływu atrybutów

Shapley values



SHAP: force plot



# Lloyd Stowell Shapley



- Amerykański matematyki i laureat Nagrody Nobla (1923-2016).
- Wkład naukowy w zastosowania ekonomiczne matematyki, w szczególności teorię gier
- Nowe propozycja w tzw. cooperative game 1953.
- Wskaźniki (Shapley values) – oszacowanie wkładu **każdego uczestnika do tzw. coalition game.**
  - Assume there are  $N$  players and  $S$  is a subset of the  $N$  players. Let  $v(S)$  be the total value of the  $S$  players. When player  $\{i\}$  join the  $S$  players, Player  $i$ 's marginal contribution is  $v(S \cup \{i\}) - v(S)$ .



# Podstawy tzw. funkcji na zbiorach

- $X = \{1, 2, \dots, n\}$  zbiór elementów (gracze, uczestnicy);  $P(X)$  – zbiór potęgowy  $X$  = zbiór wszystkich możliwych podzbiorów  $X$   
Podstawowa funkcja oceny - set function  $\mu : P(X) \rightarrow [0, 1]$
- Funkcja  $\mu$  - spełnia minimum założeń:
  - $\mu(\emptyset) = 0$  i  $\mu(X) = 1$
  - $A \subseteq B$  implikuje  $\mu(A) \leq \mu(B)$
  - „1” uznaje się za wartość max
- Praktyczna interpretacja  $\mu$  zależy od zastosowania
  - Zysk otrzymany przez graczy / agentów
  - Ważność kryteriów / atrybutów w MCDA lub analizy danych
- Transformacje funkcji  $\mu$ 
  - Wartości Shapley’a i Banzhaf’a odnosi się do pojedynczych elementów  $i \in X$ , lecz także interakcji w parach, podzbiorów  $A \subseteq X$
  - Möbius representation  $m: P(X) \rightarrow R$

# Przykład ilustracyjny – Shapley value

- Shapley value – średni udział elementu w koalicji – zbiorze
- Niech  $X=\{1,2,3\}$  gdzie zysk z akcji udziału agenta w koalicjach  $\mu(\{1\})=5$ ,  $\mu(\{2\})=7$ ,  $\mu(\{3\})=4$ ,  $\mu(\{1,2\})=15$ ,  $\mu(\{1,3\})=12$ ,  $\mu(\{2,3\})=14$  and  $\mu(\{1,2,3\})=30$
- Jak podzielić zysk 30 jednostek agentów uwzględniając ich udział/wkład do różnych koalicji?
- Rozważając wspólny udział w  $A \subseteq X$ , podziel równo  $m(A)$  pomiędzy agentów  $m(A)/|A|$
- Każdy agent powinien otrzymać udział będący wartością Shapley'a (Shapley value)

$$\varphi_i(\mu) = \sum_{A \subseteq X: i \in A} \frac{m(A)}{|A|}$$

## Przykład ilustracyjny – Shapley value

- $X=\{1,2,3\}$  i zyski z ich udziału  $\mu(\{1\})=5$ ,  $\mu(\{2\})=7$ ,  $\mu(\{3\})=4$ ,  $\mu(\{1,2\})=15$ ,  $\mu(\{1,3\})=12$ ,  $\mu(\{2,3\})=14$  and  $\mu(\{1,2,3\})=30$

Shapley values dla każdego agenta

- $\phi_1(\mu)=m(\{1\})/1+m(\{1,2\})/2+m(\{1,3\})/2+m(\{1,2,3\})/3 = 5+3/2+3/2+5/3=9.67$
- $\phi_2(\mu)=m(\{2\})/1+m(\{1,2\})/2+m(\{2,3\})/2+m(\{1,2,3\})/3 = 7+3/2+3/2+5/3=11.67$
- $\phi_3(\mu)=m(\{3\})/1+m(\{1,3\})/2+m(\{2,3\})/2+m(\{1,2,3\})/3 = 5+3/2+3/2+5/3=9.67$

# Inne sformułowanie wzoru i rozszerzenia

Shapley value:

$$\Phi_i(\mu) = \sum_{A \subseteq X - \{i\}} \frac{(|X - A| - 1)! |A|!}{|X|!} \cdot [\mu(A \cup \{i\}) - \mu(A)]$$

Banzhaf value:

$$\Phi_{Bi}(\mu) = \frac{1}{2^{|X|-2}} \sum_{A \subseteq X - \{i\}} [\mu(A \cup \{i\}) - \mu(A)]$$

Interpreacja jako średni wkład elementu  $i$  we wszystkich możliwych koalicjach  $A$

Interaction indices  $(i,j) \rightarrow$  Morofushi i Soneda; Roubens

$$I_{MS}(i,j) = \sum_{A \subseteq X - \{i,j\}} \frac{(|X - A| - 2)! |A|!}{(|X| - 1)!} \cdot [\mu(A \cup \{i,j\}) - \mu(A \cup \{i\}) - \mu(A \cup \{j\}) + \mu(A)]$$

$$I_R(i,j) = \frac{1}{2^{n-2}} \sum_{A \subseteq X - \{i,j\}} [\mu(A \cup \{i,j\}) - \mu(A \cup \{i\}) - \mu(A \cup \{j\}) + \mu(A)]$$

# Inny przykład

Trzech znajomych chce partycypować w kosztach obiadu

$$v(c) = \begin{cases} 80, & \text{if } c = \{A\} \\ 56, & \text{if } c = \{B\} \\ 70, & \text{if } c = \{C\} \\ 80, & \text{if } c = \{A, B\} \\ 85, & \text{if } c = \{A, C\} \\ 72, & \text{if } c = \{B, C\} \\ 90, & \text{if } c = \{A, B, C\} \end{cases}$$

Wkład osoby A => 51.17

Zapis wszystkich obliczeń dostępny na KDDBlog

<https://www.kdnuggets.com/2019/12/interpretability-part-3-lime-shap.html>

# Wartości Shapley'a

- Ciekawa interpretacja – dostarcza więcej informacji o wpływie (zmiennej) niż prostsze metody statystyczne

Lecz,

- Obliczenia – wymaga rozważenia wszystkich permutacji elementów (w ML atrybutów).
- Koszty obliczeniowe + badanie wszystkich możliwych koalicji coraz trudniejsze dla rosnącej liczby atrybutów = zbyt kosztowne dla ML

2013 E. Štrumbelj I. Kononnenko zaproponował przybliżone oszacowanie poprzez losowanie permutacji cech metodą Monte-Carlo

Nadal potrzebne efektywniejsze obliczenie przybliżenia wartości Shapley'a!

# Użycie Shapley value w metodzie SHAP

- SHAP – **SH**apley **A**dditive ex**P**lanations nowe podejścia dla wyjaśnień modeli predykcyjnych
- Zaproponowane przez Lundberg i Lee (NIPS 2016) jako ogólne podejście do oceny klasyfikatorów oraz modeli regresji
- Popularne dzięki efektywnej obliczeniowo implementacji
- Spójrz na authors' repository  
<https://github.com/slundberg/shap>

Obliczają przybliżenie wartości Shapley lecz w innym środowisku modeli predykcyjnych

# Przeformułowanie w metodzie SHAP

Shapley regression value

$$\Phi_i = \sum_{A \subseteq F - \{i\}} \frac{(|F - A| - 1)! |A|!}{|F|!} \cdot [f(A \cup \{i\}) - f(A)]$$

Oszacowanie ważności atrybutu w modelu liniowych / nawet przy ich skorelowaniu

Predykcja -> model ze zbiorem cech A (bez elementu i) oraz analogicznego modelu ze zbiorem rozszerzonym o {i}

A- wszystkie możliwe podzbiory z F

Matematycznie ich wartości sumują się do 1

Ponadto – można oszacować wartości Shapleya dla predykcji przykładu x za pomocą addytywnego modelu liniowego

$$\hat{\phi}_j = \frac{1}{M} \sum_{m=1}^M \left( \hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m) \right)$$

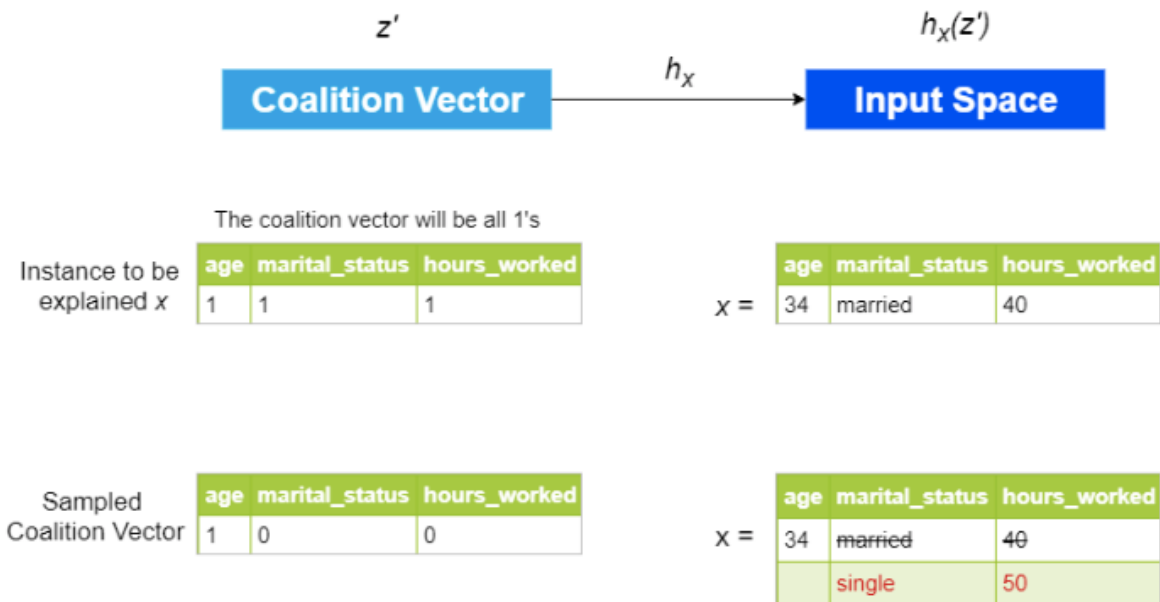


# SHAP – specjalne przybliżenia

- Kernel (partly extend linear model inspired like LIME) – do not require the evaluation of all  $2^M$  sets
- Instead an additive attribute model – a weighted linear regression with simplified inputs  $z$  and estimation Shapley values by making calculation over a sample of instance predictions

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i,$$

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



Representation of how the coalition vector is converted to original input space.

# Szybsze obliczeniowo przybliżanie wartości Shapleya

- Podstawowe podejście - Kernel SHAP metoda przybliżenia wartości Shapleya wykorzystująca model addytywny liniowy.
- Możliwe jest szybsze przybliżanie wartości Shapleya, kiedy dodamy informację o rodzaju zastosowanego modelu uczenia maszynowego. Obecnie SHAP jest wyspecjalizowany dla:
  - Tree SHAP - XGBoost, LightGBM, CatBoost, scikit-learn, pyspark
  - Deep SHAP - TensorFlow, Keras, PyTorch
  - Linear SHAP

# SHAP oferuje różne wyjaśnienia

- **The global interpretability** - SHAP values wskazują wkład każdego atrybutu : pozytywny lub negatywny do wartości wyjściowej (ang. target variable)
- **The local interpretability** - dla każdego przykładu (i atrybutów z nim skojarzonych) tzw. local set of SHAP values oraz wpływu na konkretną wartość predykcji.
- Różne formy graficznej wizualizacji

# Ilustracja wykorzystania SHAP



# SHAP motywacyjne rozważanie

Rozważmy predykcje ceny mieszkań w pewnym mieście

- Wybieramy trzy najważniejsze atrybuty: flat size, year of building and a localization (region/district of the city)
- Dla oferty “40m<sup>2</sup>, building from 1920 and inside Old City this” model przewiduje cenę 450.000
- Wiedząc że **średnie ceny w tym rejonie miasta są approx. 400.000**, stawiamy pytania: jaka jest przyczyna wyższej ceny, jaki jest wkład wartość każdego atrybutu do ceny?
- SHAP values wskazały pozytywny wkład lokalizacji mieszkania (podniesienie ceny o około 70.000); negatywny związek z wiekiem (obniżenie o około 20.000), wielkość mieszkania nie ma znaczącego wpływu na cenę.

# Boston housing data

Dane o 506 nieruchomościach opisanych 13 atrybutami i jednym wyjściem (MEDV – the price of the house)

Atrybuty:

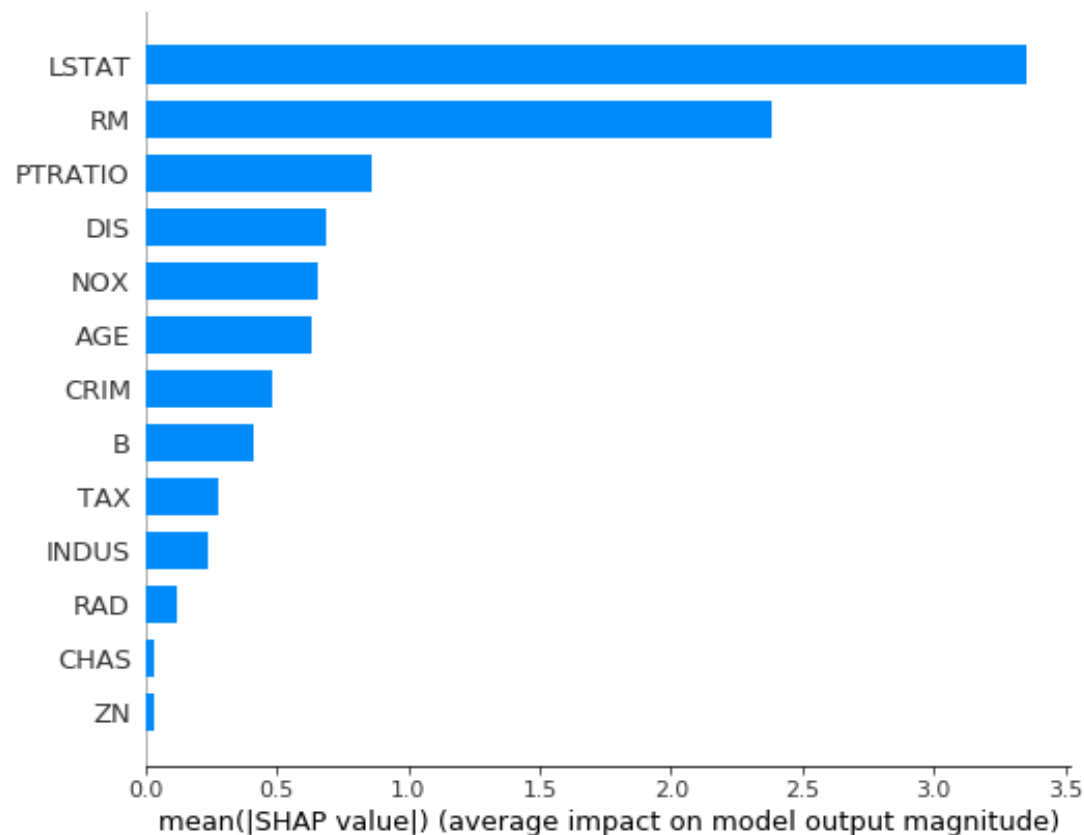
- CRIM – wskaźnik przestępczości na mieszkańca według miasta
- ZN – część działki pod zabudowę mieszkaniową pod działki o powierzchni ponad 25 000 **stóp kwadratowych**
- INDUS – odsetek niedetalicznych akrów biznesowych na miasto.
- CHAS – zmienna zmienna Charles River (1, jeśli trasa ogranicza rzekę; 0 w przeciwnym razie)
- NOX – stężenie tlenków azotu (części na 10 milionów)
- RM – średnia liczba pokoi na mieszkanie
- AGE – odsetek jednostek zajmowanych przez właścicieli wybudowanych przed 1940 r
- DIS – ważone odległości do pięciu centrów zatrudnienia w Bostonie
- RAD – indeks dostępności do radialnych autostrad
- TAX- pełna stawka podatku od nieruchomości od 10 000 USD
- PTRATIO – stosunek liczby uczniów do nauczycieli według miasta
- B –  $1000 (B_k - 0,63)^2$ , gdzie  $B_k$  to odsetek czarnych według miasta
- LSTAT -% niższy status populacji
- MEDV – Mediana wartości domów zajmowanych przez właścicieli w tysiącach dolarów

Modele – wybrano XGBoost regressor vs. linear regression)

# Globalna interpretacja

Typowa wizualizacja rankingu atrybutów wg. Wartości Shapley'a / im wyżej, tym bardziej wpływowe

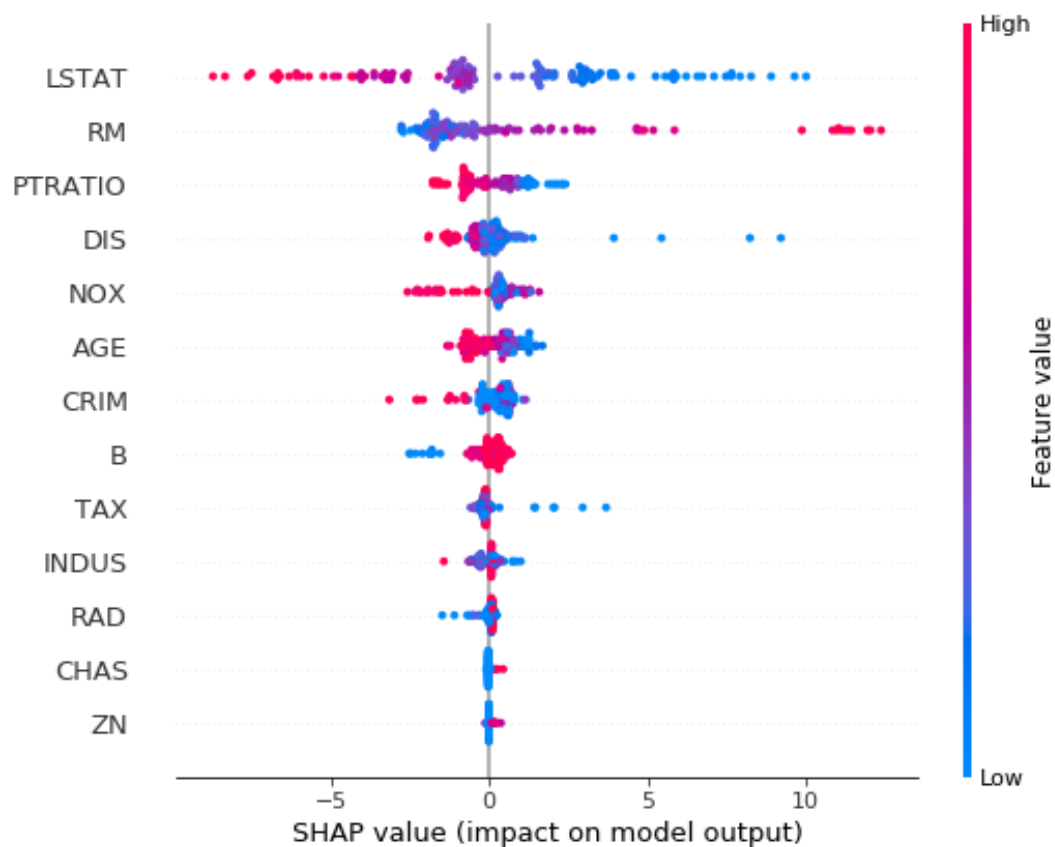
LSTAT, RM, ... najbardziej wpływowe, a CHAS i ZN najmniej



# Global kierunek wpływu

Można pokazać pozytywny lub negatywny wpływ każdego atrybutu na wartości zmiennej wyjściowej

Czerwony kolor oznacza wpływ na podwyższenie wartości, a niebieski na zmniejszenie wartości wyjścia (tutaj ceny nieruchomości)

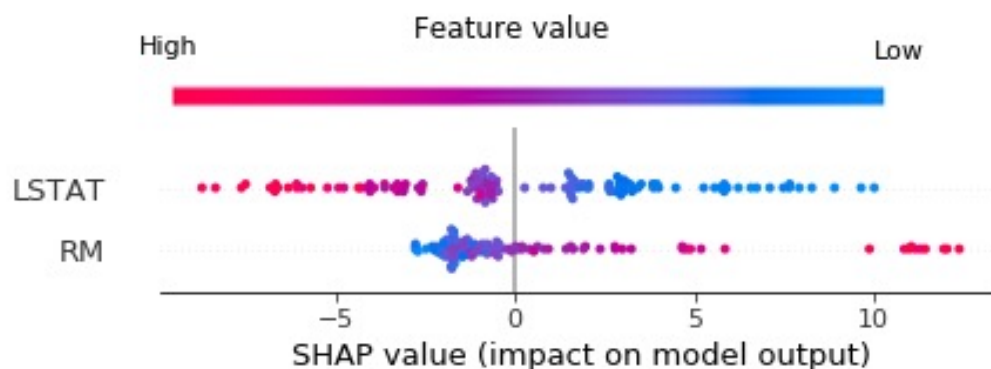




# Wykres kierunku wpływu

Wykres obejmuje przykłady uczące – każda kropka to pojedynczy przykład:

- Ranking ważności atrybutów wg. wyższych wartości Shapleya
- Kierunek wpływ: linia pozioma – czy oryginalna wartość atrybutu dla obiektu wpływa na wyższe (czerwone) czy zmniejszenie (niebieskie) wartości predykcji  $y$

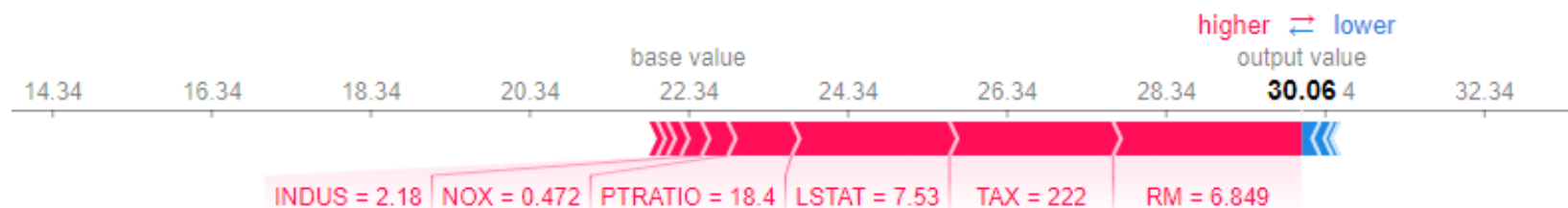


LSTAT : niższe wartości – większy wpływ na wyższą cenę

RM : inny rodzaj wpływu– wyższe wartości powiązane z wyższą ceną

## Lokalne wyjaśnienie dla pojedynczej predykcji – jak wartości konkretnego przykładu wpływają na wartość wyjścia

Dla tego przykładu:



Wyjście – predykcja modelu ( $y = 30,06$ ) – wyższa niż średnia wartość predykcji dla wszystkich przykładów (base value 22.34)

Kolory – czerwone – atrybuty, które swoimi wartościami wpływają na podwyższenie wartości  $y$  (ceny) i w jakim stopniu, niebieskie wkład do obniżenia

Wskazane atrybuty częściowo zgodne z rankingiem ważności / największy wkład do wysokiej ceny mają RM, TAX później LSTAT i PTRATIO

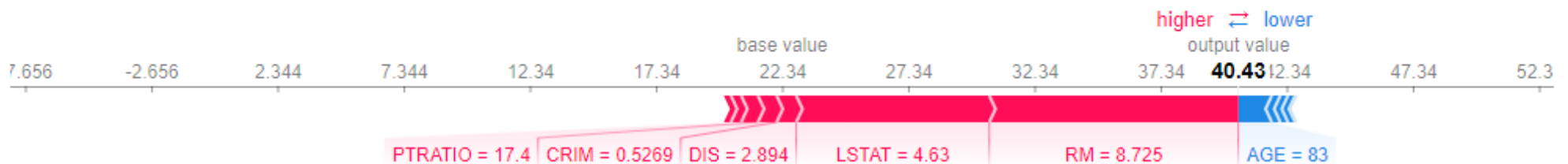
# Porównywanie kilku predykcji

Porównaj analizę predykcji dla ofert nr 17, 23 i 54

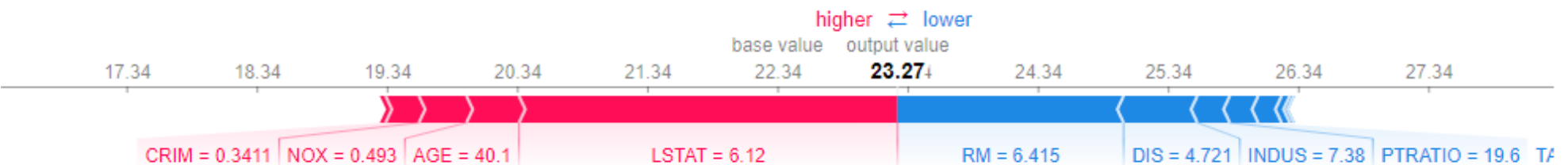
Zauważ inne wkłady atrybutów dla wyższych i mniejszych cen

Możliwość dostawienia ofert dla klientów

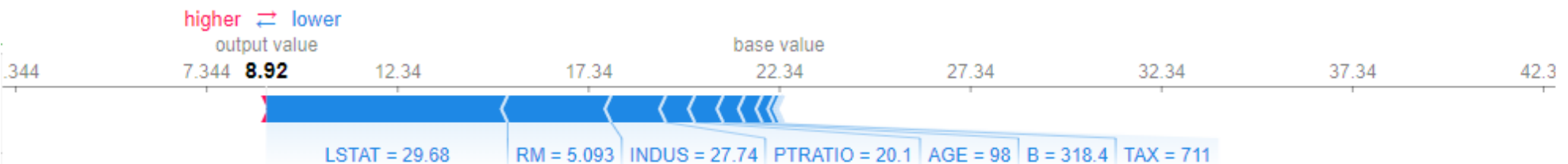
Obserwacja nr: 17



Obserwacja nr: 23



Obserwacja nr: 54



# Niektóre biblioteki

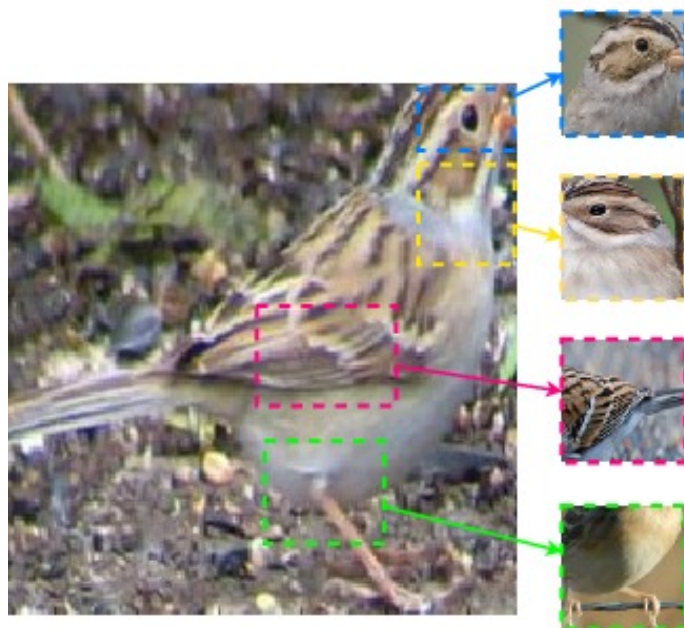
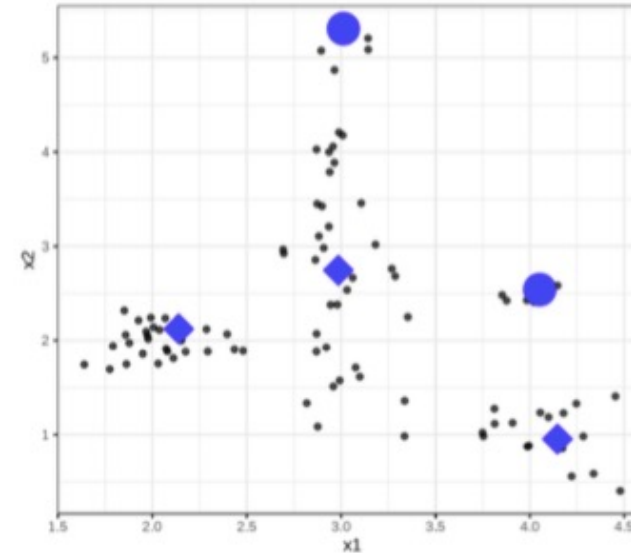
- DeepExplain: perturbation and gradient-based attribution methods for Deep Neural Networks interpretability.  
[github.com/marcoancona/DeepExplain](https://github.com/marcoancona/DeepExplain)
- iNNvestigate: A toolbox to iNNvestigate neural networks' predictions.  
[github.com/albermax/innvestigate](https://github.com/albermax/innvestigate)
- SHAP: SHapley Additive exPlanations. [github.com/slundberg/shap](https://github.com/slundberg/shap)
- ELI5: A library for debugging/inspecting machine learning classifiers and explaining their predictions. [github.com/TeamHG-Memex/eli5](https://github.com/TeamHG-Memex/eli5)
- Skater: Python Library for Model Interpretation/Explanations.  
[github.com/datascienceinc/Skater](https://github.com/datascienceinc/Skater)
- Yellowbrick: Visual analysis and diagnostic tools to facilitate machine learning model selection. [github.com/DistrictDataLabs/yellowbrick](https://github.com/DistrictDataLabs/yellowbrick)
- Lucid: A collection of infrastructure and tools for research in neural network interpretability. [github.com/tensorflow/lucid](https://github.com/tensorflow/lucid)

# Wyjaśnienia z prototypami

Prototyp – przykład, który jest reprezentatywny dla podzbioru danych

- Rzeczywisty przykład uczący
- Centroid skupiska
- Sztuczny przykład o specjalnych właściwościach

Klasyfikacja  $x$  – może być wyjaśniona poprzez podobieństwo do prototypów



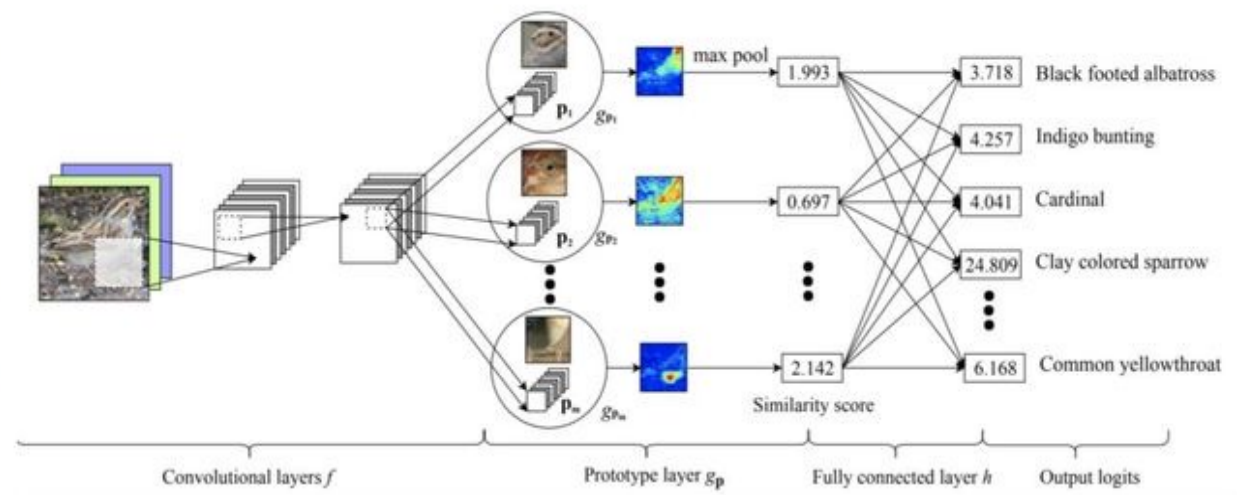
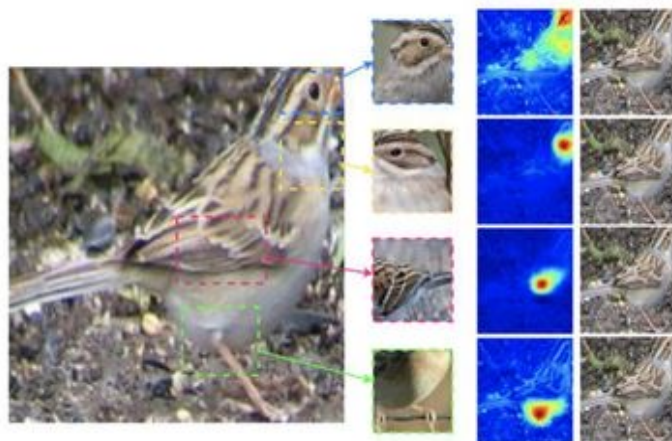
Obrazy – np. PROTOPNET [Chen et al 2019] This looks like this – specjalna odmiana CNN

Prototyp – na podstawie ukrytej reprezentacji w warstwach sieci – powiązany do części obrazu

Klasyfikacja = określa ich ważne podobieństwo w obrazie testowych do wyuczonych prototypów – proste do interpretacji

# ProtoNet – porównywalna trafności do innych CNN

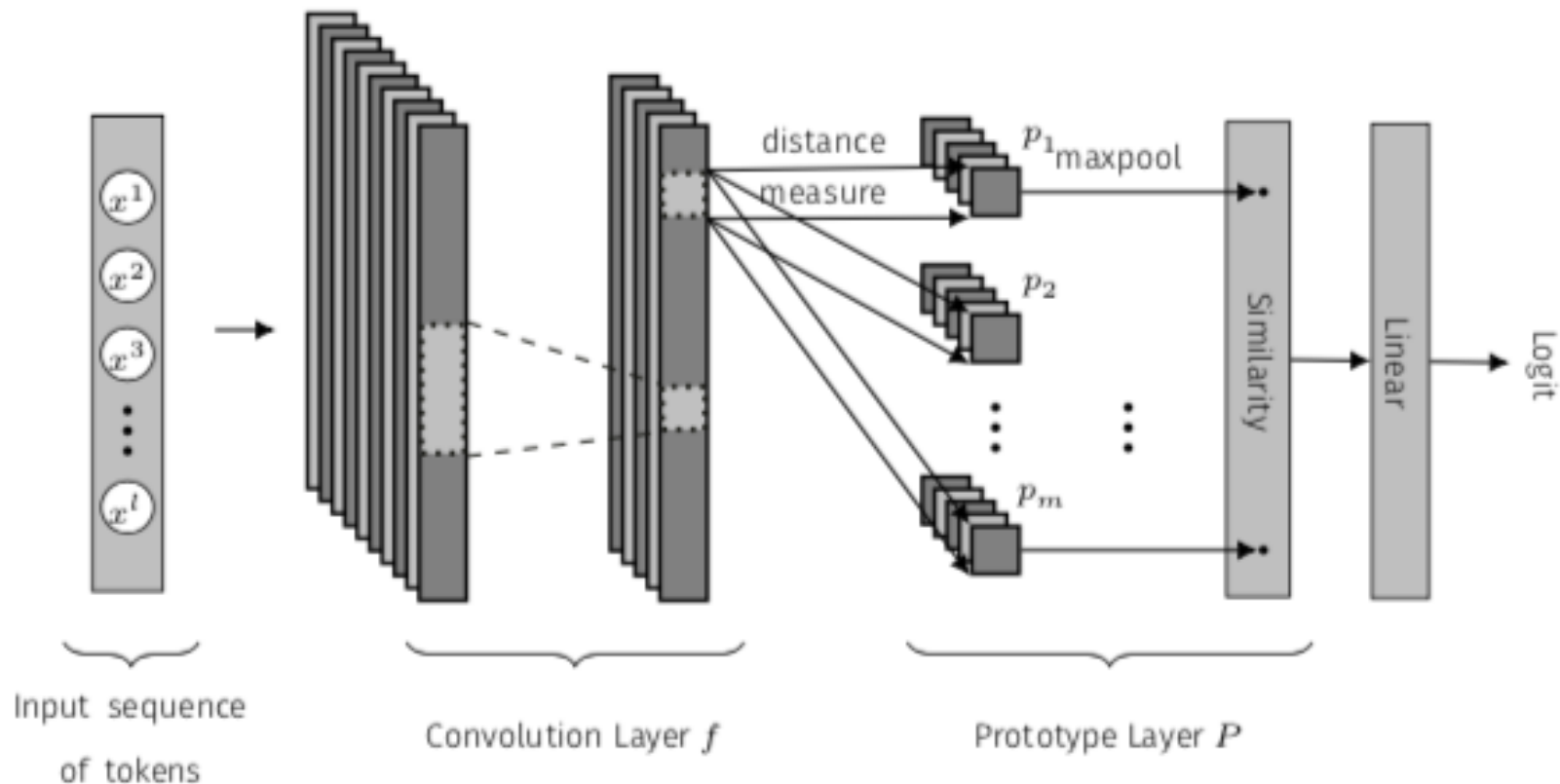
## Interpretable Deep Learning



This Looks Like That: Deep Learning for Interpretable Image Recognition. Chen et al 2018.

# K.Pluciński, M.Lango, J.Stefanowski: Prototypical Convolutional Neural Network; 2021

- Prototypy związane z najbardziej podobnymi frazami w dokumencie
- Architektura konwolucyjna z filtrami n-gram + “white” wyjście - liniowa ważona kombinacja podobieństwa przykładu testowego do wyuczonych prototypów – dynamiczny dobór ich liczby



# Przykład dialogu z użytkownikiem

Table 2: Explanation for a correct prediction.

**Input example:** about twenty minutes into this movie i was already bored quite simply these characters were fairly dull occasionally something enjoyable would happen but then things would slow down again fortunately my patience was eventually rewarded and the ending to this movie was n't bad at all however it was by no means good enough to justify sitting through the first ninety minutes so i would say that the movie was mediocre overall and considering all of the talent in the cast i'd call this a disappointment.

Prediction: **Negative**, Gold standard: **Negative**

## Evidence for negative sentiment:

Prototype	Most similar phrase	Similarity * Weight
unbelievable rambling nonsense that should a waste of film was the worst film i	was mediocre overall and considering	$2.17 * 0.96 = 2.08$
	characters were fairly dull occasionally	$2.08 * 0.86 = 1.79$
	the movie was mediocre overall	$2.89 * 0.56 = 1.62$

Sum of evidence: 5.49

## Evidence for positive sentiment:

Prototype	Most similar phrase	Similarity * Weight
the best filmgoing experiences i is a wonderful film full lots of great comedy from	occasionally something enjoyable would happen	$0.60 * 0.97 = 0.58$
	the first ninety minutes so	$0.61 * 0.78 = 0.48$
	my patience was eventually rewarded	$0.82 * 0.45 = 0.36$

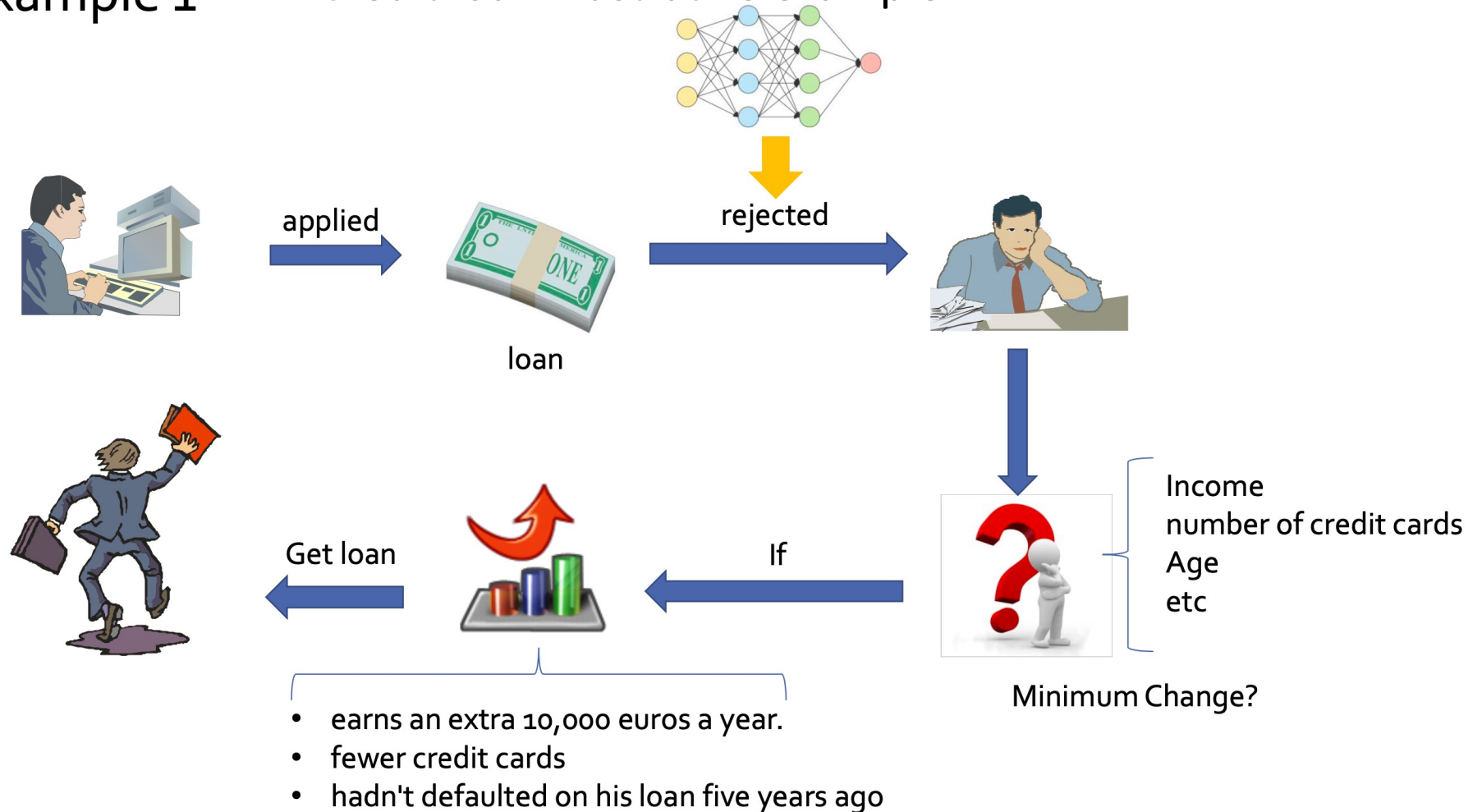
Sum of evidence: 2.01



# Kontrfakty

## Example 1

### Credit loan illustrative example



Molnar, C. (2020). Interpretable machine learning. Lulu. com.

## Example 2 : apartment sale offers

# Wyjaśnienia z kontrfaktami

- **Counterfactual explanations** – opisują zależności przyczynowo skutkowe pomiędzy zmianą przesłanki, a zmianą decyzji,  
“if A had not occurred, then B would not have occurred” [Dandl et al. ]
- W ML dla przykładu (x,y) najmniejsze dopuszczalne zmiany x, które doprowadzą do (pożądaney zmiany predykcji  $b(x') \neq y$

Przykład decyzji kredytowej

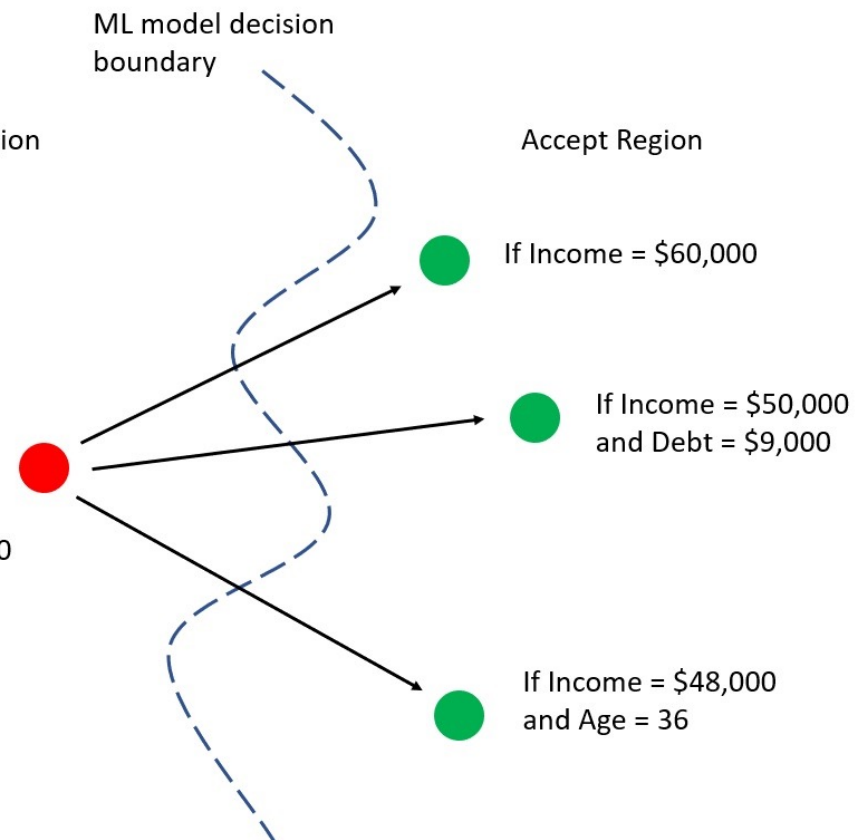
- Dopuszczalne atrybuty
- Optym. funkcji straty [Wachler]

$$C(x) = \arg \min L(b(x'), y) + \lambda d(x, x')$$

Bardziej złożone [DICE] – wiele C

$$C(x) = \arg \min_{x'_1 \dots x'_k} \frac{1}{k} \sum_{i=1}^k \mathcal{L}(f(x'_i), y) + \frac{\lambda_1}{k} d(x'_i, x) - \lambda_2 \cdot dpp_{diversity}(x' \dots x'_k)$$

Income = \$45,000  
Debt = \$11,000  
Age = 29  
Savings = \$6,000



# Uwagi podsumowujące

- XAI i interpretowalne ML – motywowane praktycznymi problemami i nowymi zastosowaniami
- Część rozwiązań wykorzystuje inspiracje z dawnych propozycji w ML i innych dziedzinach nauki
- Otwarte pytania badawcze : czy rozwijać metody post-hoc czy explainable by design?
- Zrozumienie wyniku wielu metod – wymaga jednak przygotowania i wiedzy dziedzinowej
- Miary i techniki oceny – ciągle poszukujemy

# Poczytaj więcej

## Książki:

- Christoph Molnar, “Interpretable Machine Learning: A Guide for making black box models explainable”
- Przemysław Biecek, Explanatory Model Analysis (book under preparation)

## Artykuły:

- Bibal A., Frenay B.: Interpretability of machine learning models and representations: an introduction. W: Proceedings of ESANN 2016
- Bratko I.: Machine learning: between accuracy and interpretability.
- Freitas A.: Comprehensible classification models: a position paper. ACM SIGKDD Exploration Newsletter, Vol. 15, Nr 1, 2014
- Guidotti R, Monreale A., Ruggieri S, Turini F., Giannotti F, Pedreschi D.: A Survey of Methods for Explaining Black Box Models, ACM Comput. Surv., 2018
- Ribeiro M. T., Singh S., Guestrin C.: Why should i trust you? Explaining the predictions of any classifier, W: Proc. of the 22nd ACM SIGKDD 2015
- Samek, W., Wiegand, T., Müller, K. R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296, 2017.
- Stefanowski J. Woźniak M.: Interpretacja modeli uczonych ze złożonych danych medycznych. W Informatyka w medycynie 2019.

# Pytanie i komentarze?

Dalszy kontakt:

[jerzy.stefanowski@cs.put.poznan.pl](mailto:jerzy.stefanowski@cs.put.poznan.pl)

<http://www.cs.put.poznan.pl/jstefanowski/>



**Fundusze  
Europejskie**  
Polska Cyfrowa



**Rzeczpospolita  
Polska**

**Unia Europejska**  
Europejski Fundusz  
Rozwoju Regionalnego

