modelStudio & SAFE

Czyli jak robić XAI w 2020 roku (i się nie narobić)

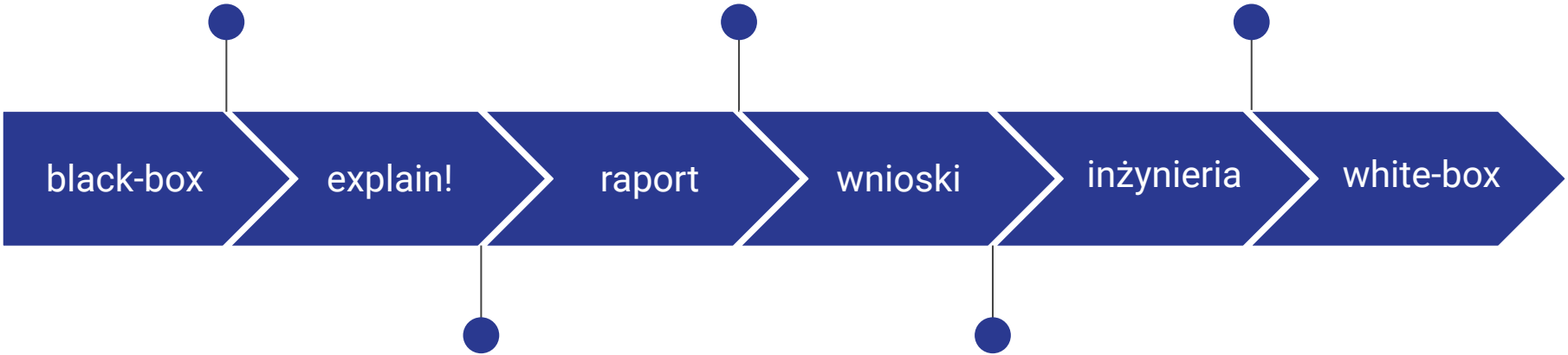
Hubert Baniecki
Mateusz Polakowski

2019

Realny problem rozwiązany czarną skrzynką

Praktyka:
znajomość bibliotek +
tworzenie wyjaśnień +
stworzenie raportu +
reprodukowalność +
zapisanie + ...

Inżynieria cech do wyjaśnialnego modelu



Potrzeba wyjaśnienia modelu: kto? co? jak?

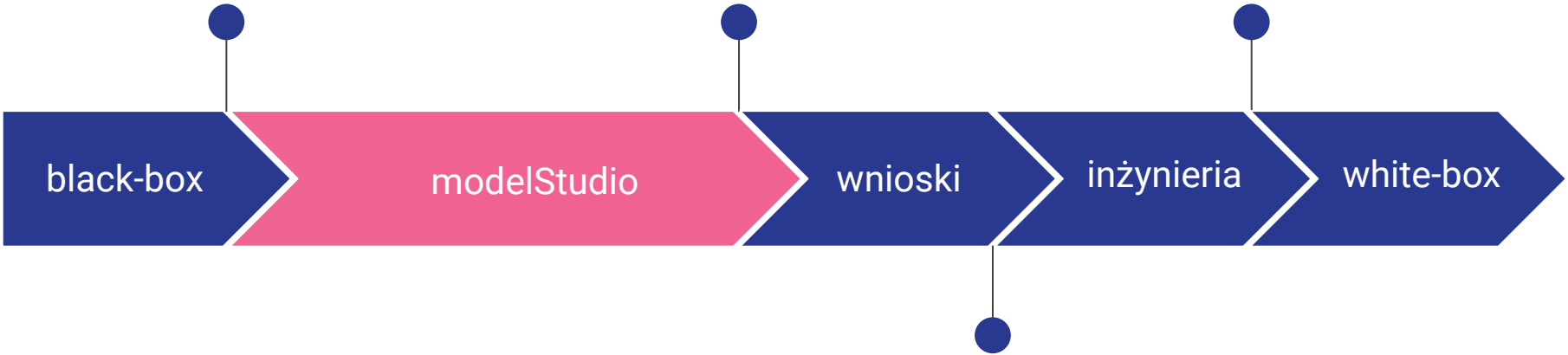
Teoria:
znajomość wyjaśnień +
wybranie poprawnych +
świadome wnioskowanie + ...

2020

Praktyka:
jedna funkcja +
automatyczne tworzenie wyjaśnień
i raportu +
reprodukowalność +
łatwe do zapisania + ….

Realny problem rozwiązany
czarną skrzynką

Inżynieria cech do
wyjaśnialnego modelu



Teoria:
znajomość wyjaśnień +
wybranie poprawnych +
świadome wnioskowanie + ...

modelStudio

Interactive Studio for Explanatory Model Analysis

Overview

The `modelStudio` package **automates the Explanatory Analysis of Machine Learning predictive models.** Generate advanced interactive and animated model explanations in the form of a **serverless HTML site** with only one line of code.

The main `modelStudio()` function computes various (instance and dataset level) model explanations and produces an **interactive, customisable dashboard made with D3.js.** It consists of multiple panels for plots with their short descriptions. Easily **save and share** the dashboard with others. Tools for model exploration unite with tools for EDA (Exploratory Data Analysis) to give a broad overview of the model behavior.

Przykładowy kod:

```
library(modelStudio)

# Create a model:
model <- glm(target ~., data = train)

# Wrap it into an explainer:
explainer <- DALEX::explain(model,
                            data = test,
                            y = test$target,
                            label = "glm")

# Pick some data points:
new_observations <- test[1:4,]

# Make a studio for the model:
modelStudio(explainer, new_observations)
```

Zalety :

- Zestawienie wielu lokalnych i globalnych wyjaśnień obok siebie
- Model agnostic
- Language agnostic*
- Automatyczne tworzenie
- Interaktywny interfejs
- Personalizowany grid
- Możliwość porównania dużej liczby obserwacji
- Wykresy EDA
- Brak serwera
- Łatwo zapisać
- Łatwo udostępnić

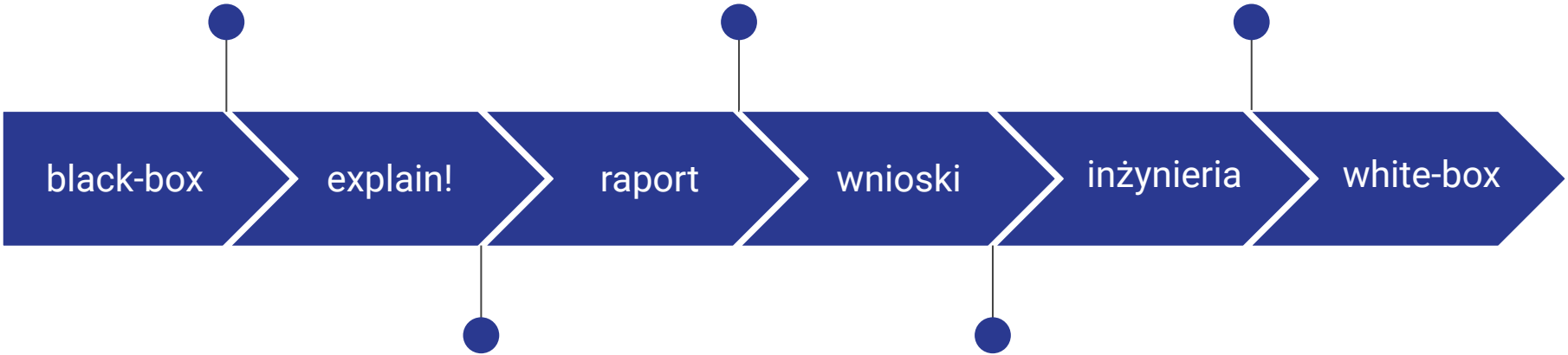
*wrap your Python model https://github.com/ModelOriented/DALEXtra

2019

Realny problem rozwiązany czarną skrzynką

Praktyka:
znajomość bibliotek +
tworzenie wyjaśnień +
stworzenie raportu +
reprodukowalność +
zapisanie + ...

Inżynieria cech do wyjaśnialnego modelu



Potrzeba wyjaśnienia modelu: kto? co? jak?

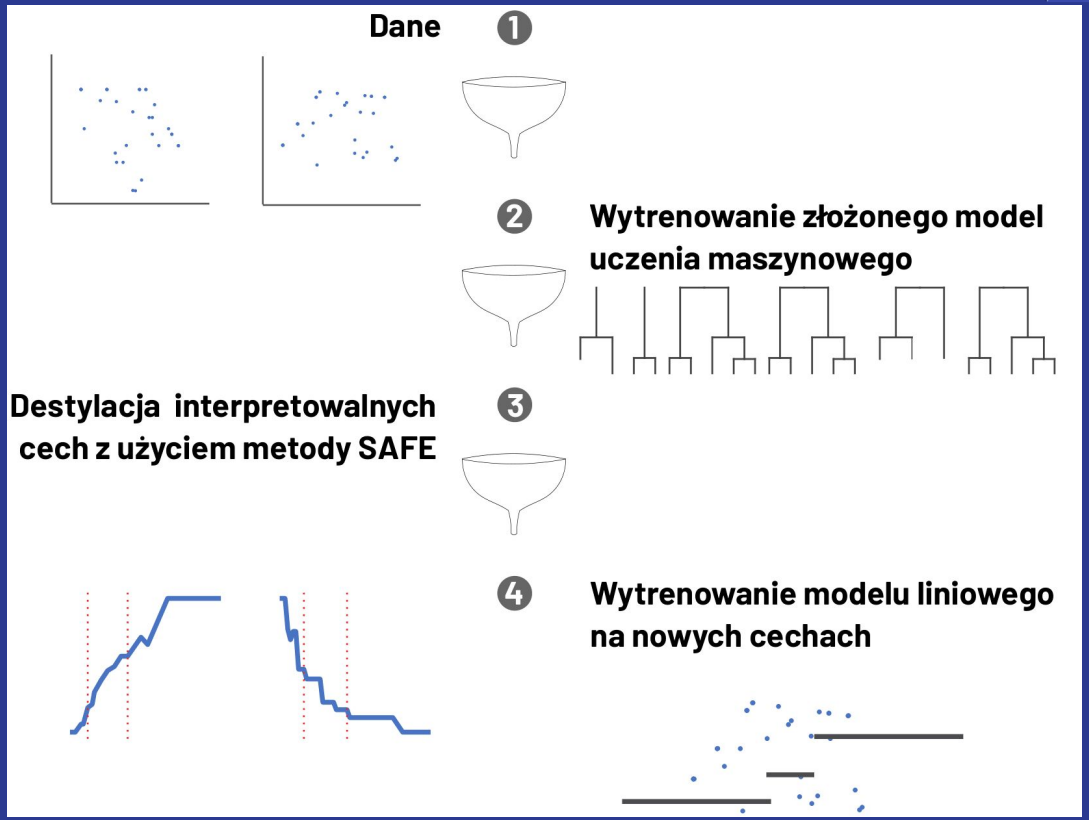
Teoria:
znajomość wyjaśnień +
wybranie poprawnych +
świadome wnioskowanie + ...

2020

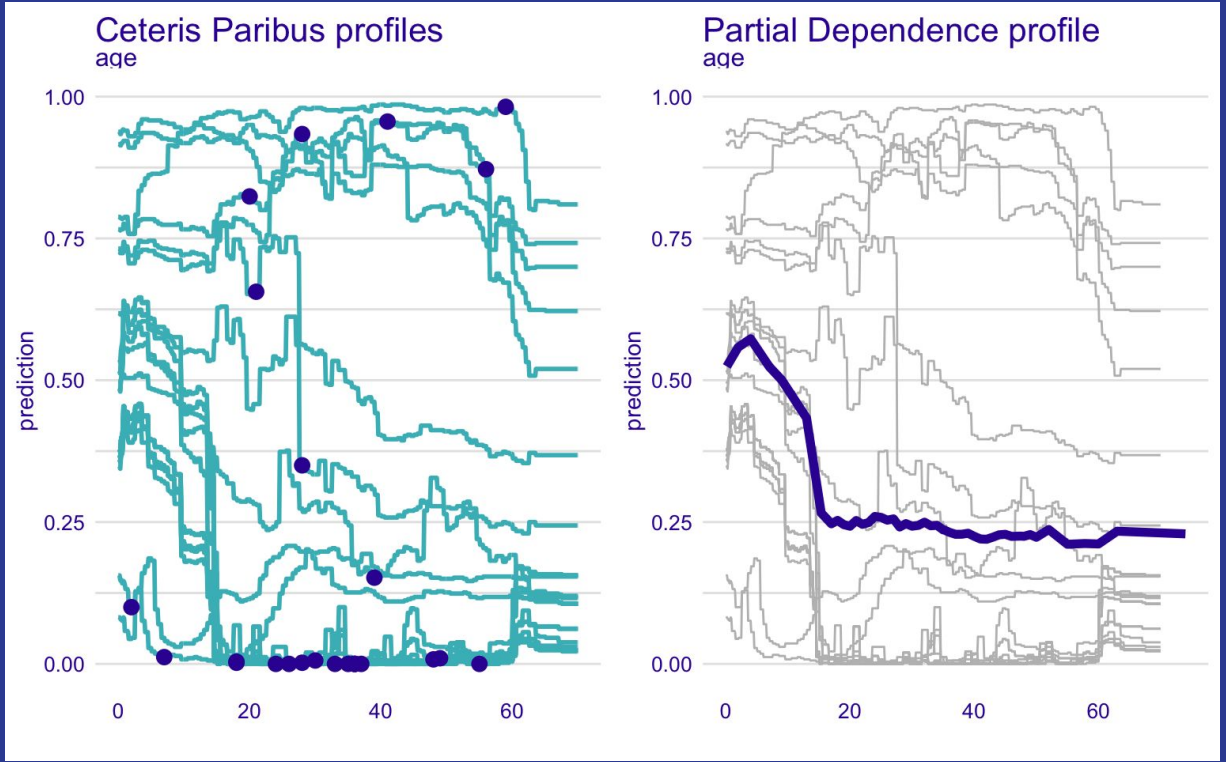
Realny problem rozwiązany czarną skrzynką

Automatyczna inżynieria cech do wyjaśnialnego modelu

SAFE


Dane
1
2
Wytrenowanie złożonego model uczenia maszynowego
Destylacja interpretowalnych cech z użyciem metody SAFE
3
4
Wytrenowanie modelu liniowego na nowych cechach

PDP


Ceteris Paribus profiles
age
prediction
1.00
0.75
0.50
0.25
0.00
0
20
40
60
Partial Dependence profile
age
prediction
1.00
0.75
0.50
0.25
0.00
0
20
40
60

Hierarchical Clustering

Bardzo użyteczne linki:

- List of XAI everything
  https://github.com/jphall663/Awesome-machine-learning-interpretability
- List of XAI papers & tools
  https://github.com/lopusz/awesome-interpretable-machine-learning
- XAI ebook https://pbiecek.github.io/ema/
- IML ebook https://christophm.github.io/interpretable-ml-book/

Użyteczne linki:

- modelStudio
  https://github.com/ModelOriented/modelStudio
- rSAFE https://github.com/ModelOriented/rSAFE
- pSAFE https://github.com/ModelOriented/SAFE
- 1st SAFE paper https://arxiv.org/abs/1902.11035
- 2nd SAFE paper https://arxiv.org/abs/2002.04267