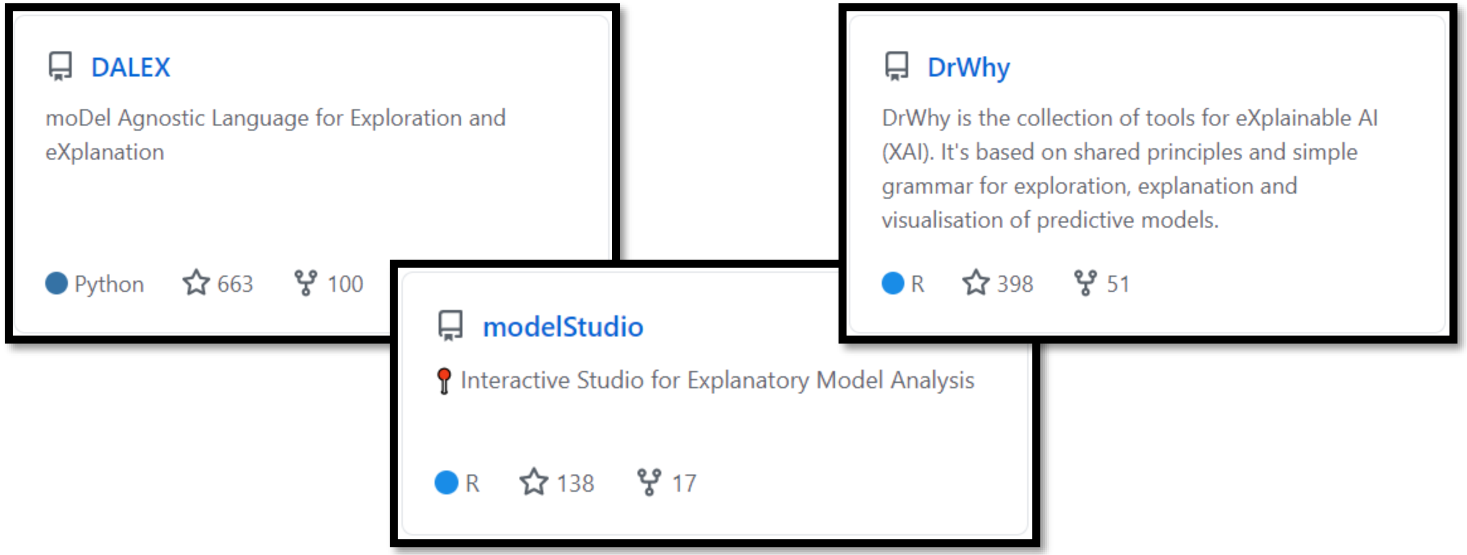
Tools for Explainable Artificial Intelligence

Hubert Baniecki

X-Europe Webinar, 11.2020



DALEX

moDel Agnostic Language for Exploration and eXplanation

● Python ☆ 663 ⑂ 100

modelStudio

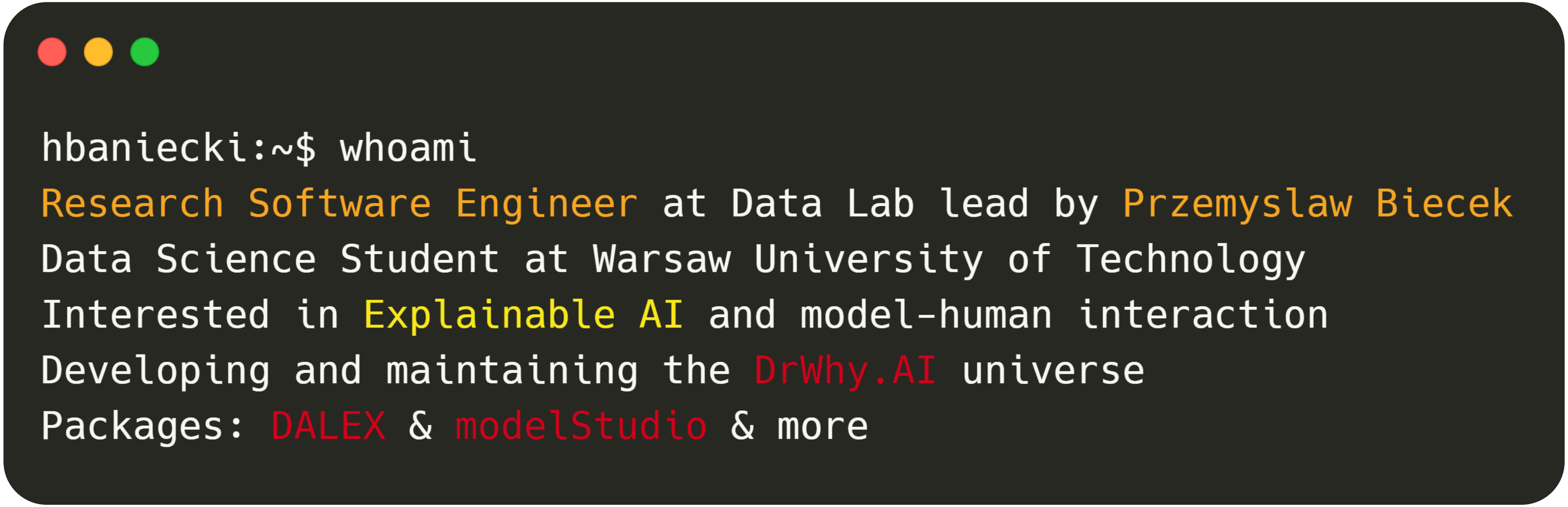
📍 Interactive Studio for Explanatory Model Analysis

● R ☆ 138 ⑂ 17

DrWhy

DrWhy is the collection of tools for eXplainable AI (XAI). It's based on shared principles and simple grammar for exploration, explanation and visualisation of predictive models.

● R ☆ 398 ⑂ 51

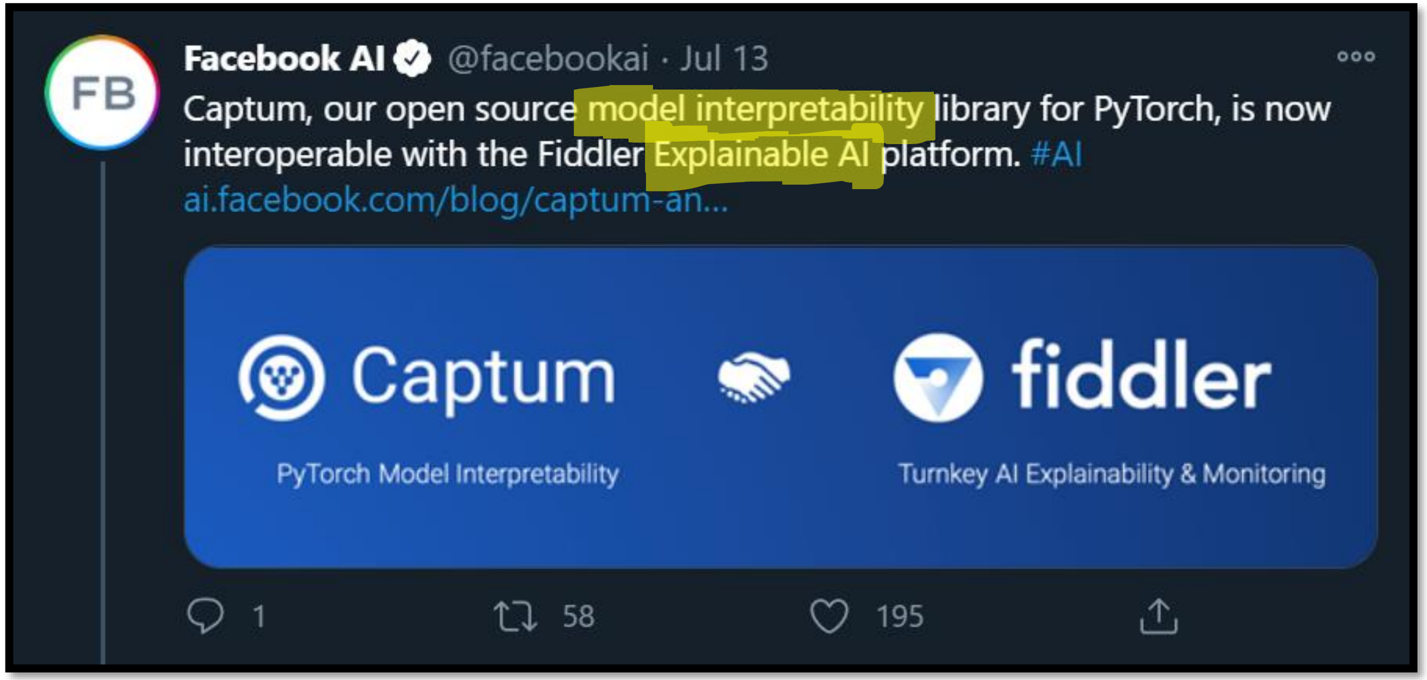


```
hbaniecki:~$ whoami
Research Software Engineer at Data Lab lead by Przemyslaw Biecek
Data Science Student at Warsaw University of Technology
Interested in Explainable AI and model-human interaction
Developing and maintaining the DrWhy.AI universe
Packages: DALEX & modelStudio & more
```

https://linkedin.com/in/hbaniecki

The semantics of Explainable AI (XAI)

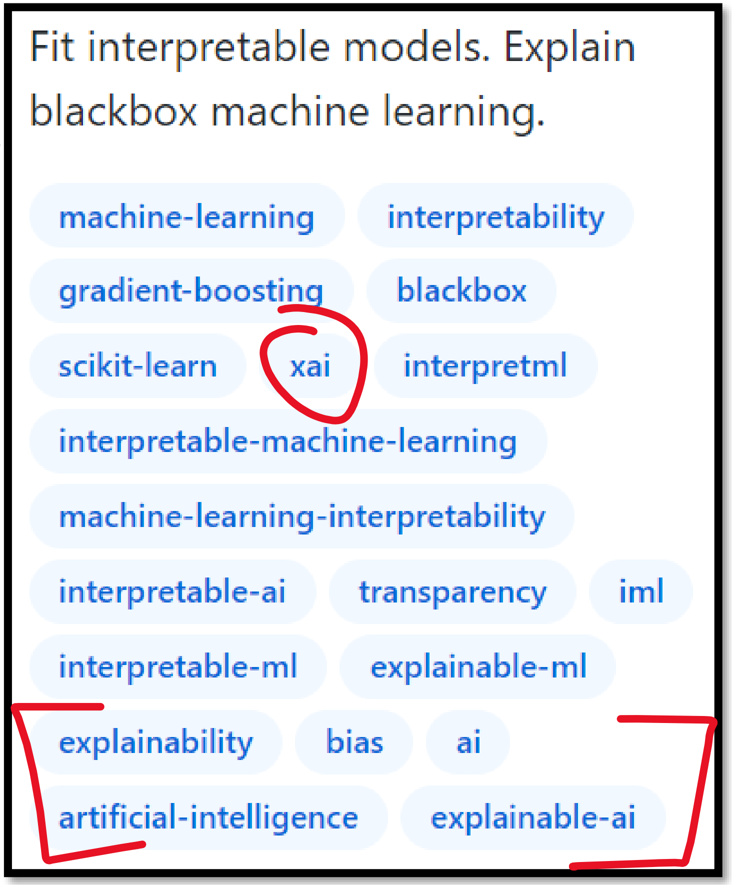
- **IBM**: A set of capabilities and methods used to describe an AI model, its expected impact and potential biases.
- **Microsoft**: *model interpretability*
-



Facebook AI @facebookai · Jul 13

Captum, our open source model interpretability library for PyTorch, is now interoperable with the Fiddler Explainable AI platform. #AI ai.facebook.com/blog/captum-an...

FB
Captum
PyTorch Model Interpretability
fiddler
Turnkey AI Explainability & Monitoring
1
58
195



Fit interpretable models. Explain blackbox machine learning.

https://linkedin.com/in/hbaniecki

Explainable AI (XAI)

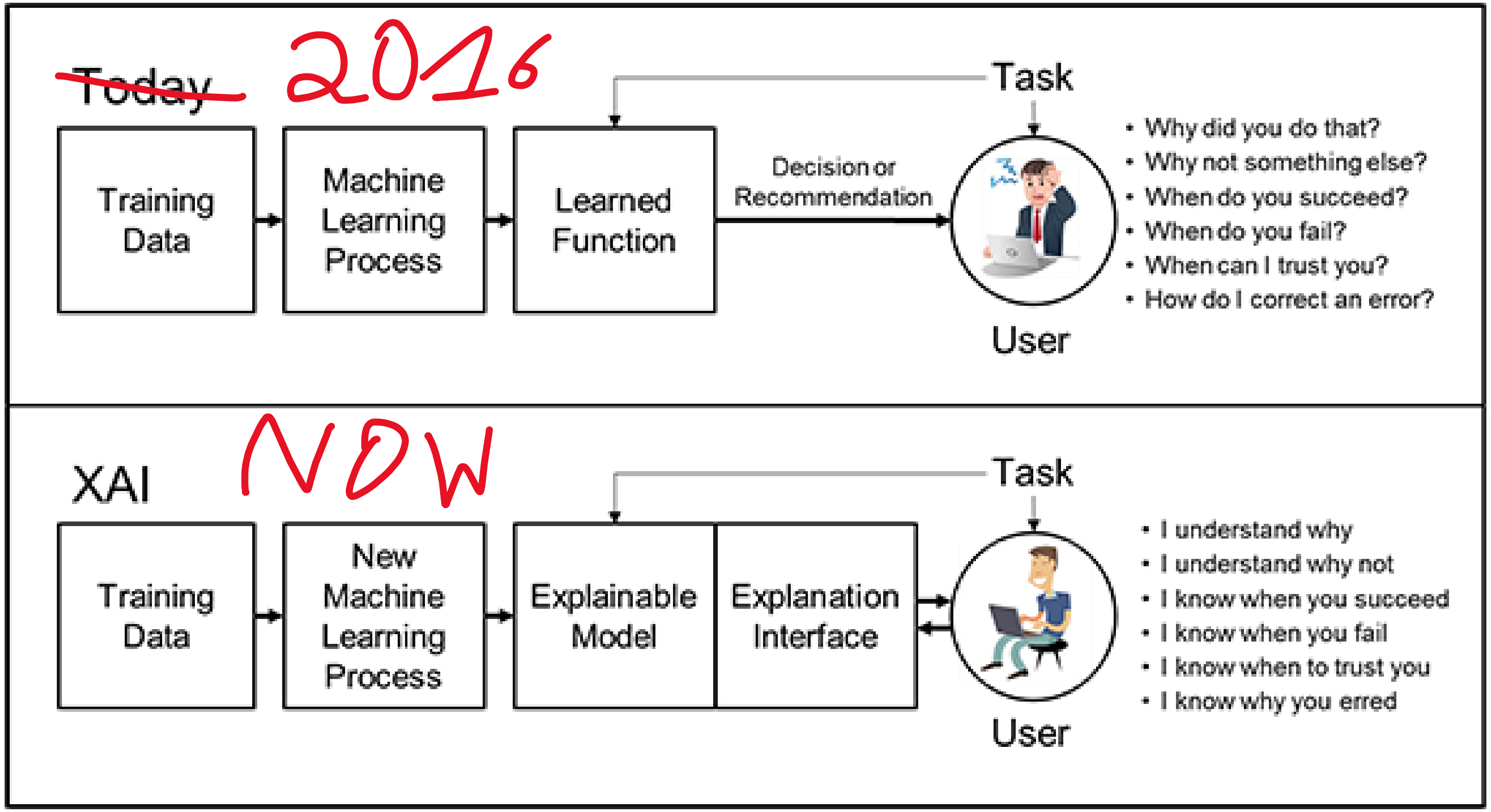
Google:
Tools and frameworks to understand and interpret your machine learning models.

https://cloud.google.com/explainable-ai

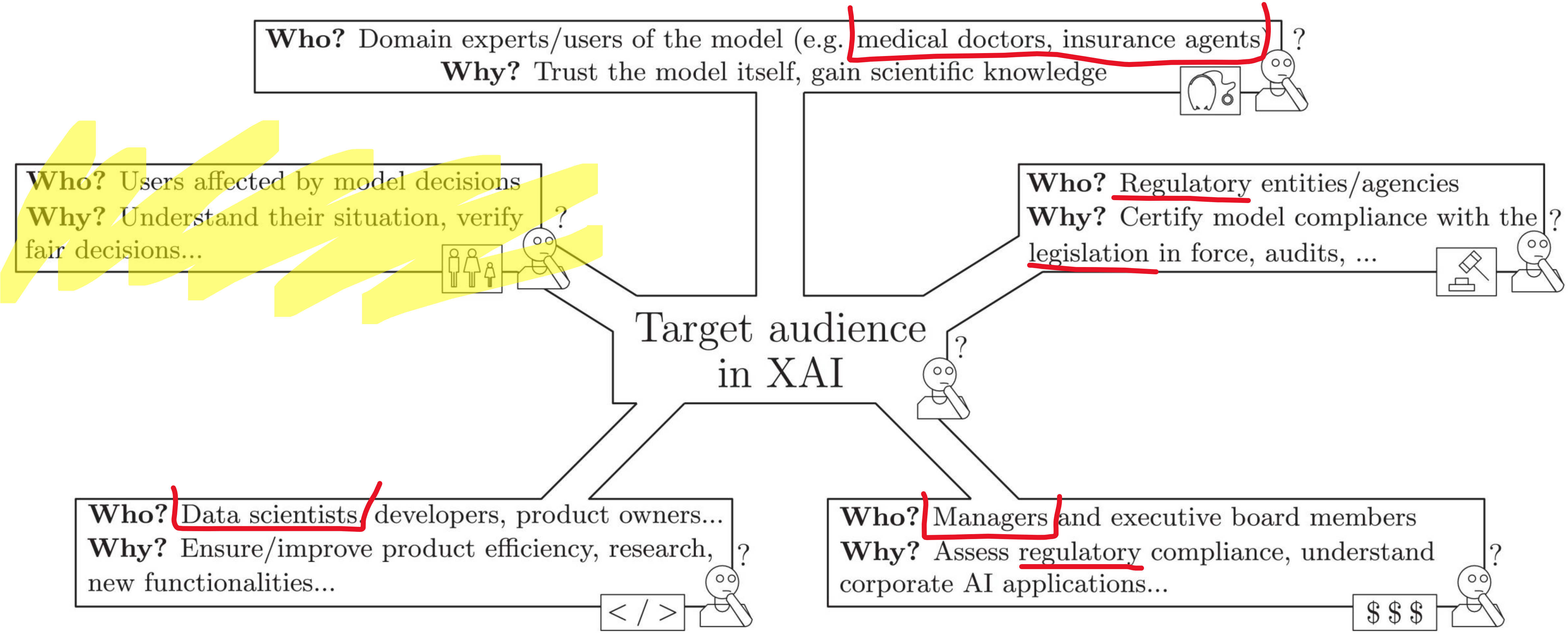
Invest in XAI research 2017-21

DARPA (Defense Advanced Research Projects Agency)

Government agency



https://www.darpa.mil/program/explainable-artificial-intelligence



Target audience in XAI

Who? Domain experts/users of the model (e.g. medical doctors, insurance agents)
Why? Trust the model itself, gain scientific knowledge

Who? Users affected by model decisions
Why? Understand their situation, verify fair decisions...

Who? Regulatory entities/agencies
Why? Certify model compliance with the legislation in force, audits, ...

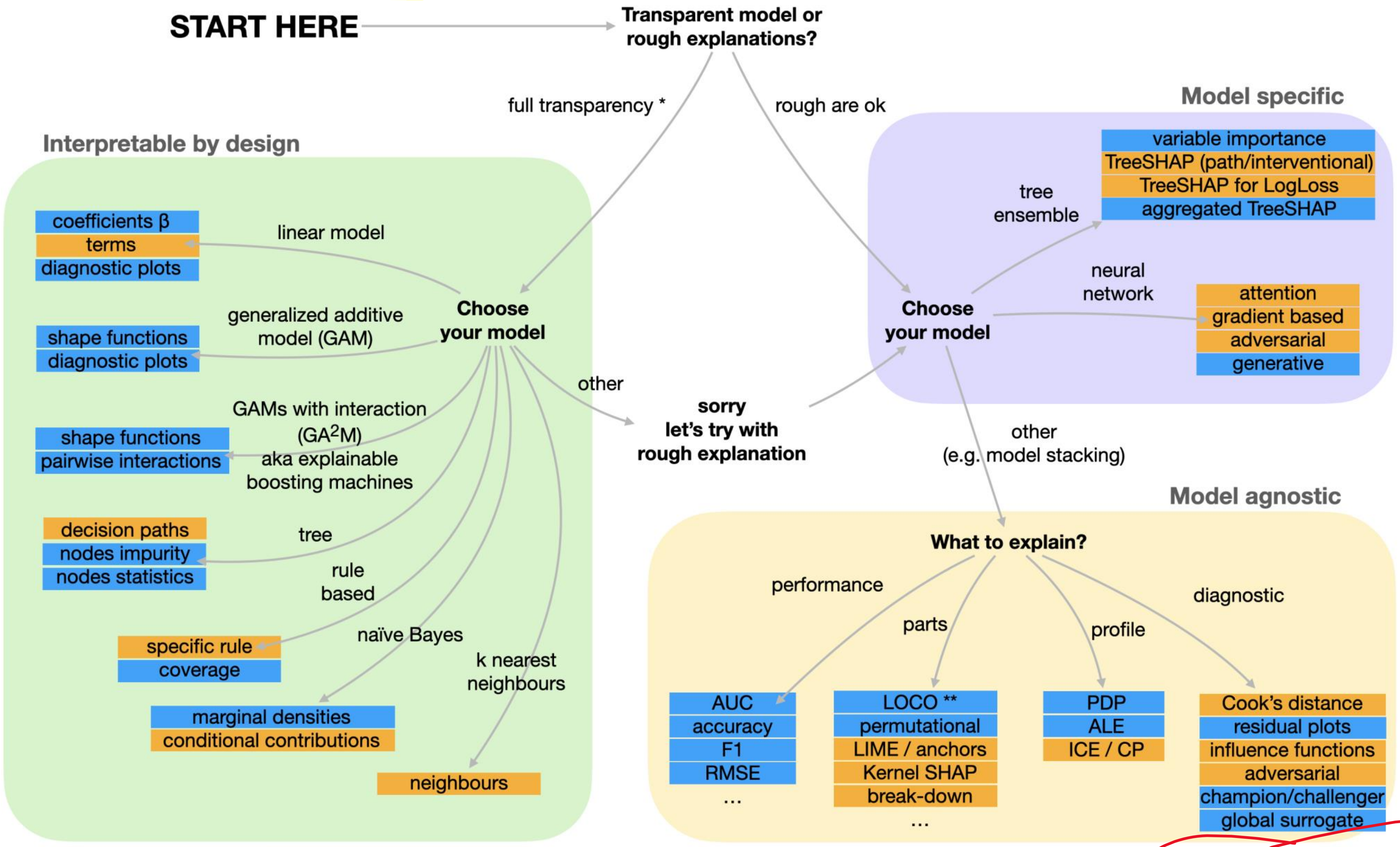
Who? Data scientists, developers, product owners...
Why? Ensure/improve product efficiency, research, new functionalities...

Who? Managers and executive board members
Why? Assess regulatory compliance, understand corporate AI applications...

Arrieta, A. B. et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*.

https://doi.org/10.1016/j.inffus.2019.12.012

Maksymiuk, S., Gosiewska, A., & Biecek, P. (2020). „Landscape of R packages for XAI”. *arXiv preprint*



START HERE

Transparent model or rough explanations?

full transparency *

rough are ok

Interpretable by design

Choose your model

linear model
- coefficients β
- terms
- diagnostic plots

generalized additive model (GAM)
- shape functions
- diagnostic plots

GAMs with interaction (GA^2M) aka explainable boosting machines
- shape functions
- pairwise interactions

tree
- decision paths
- nodes impurity
- nodes statistics

rule based
- specific rule
- coverage

naïve Bayes
- marginal densities
- conditional contributions

k nearest neighbours
- neighbours

other

sorry let's try with rough explanation

Model specific

Choose your model

tree ensemble
- variable importance
- TreeSHAP (path/interventional)
- TreeSHAP for LogLoss
- aggregated TreeSHAP

neural network
- attention
- gradient based
- adversarial
- generative

other (e.g. model stacking)

Model agnostic

What to explain?

performance
- AUC
- accuracy
- F1
- RMSE
- …

parts
- LOCO **
- permutational
- LIME / anchors
- Kernel SHAP
- break-down
- …

profile
- PDP
- ALE
- ICE / CP

diagnostic
- Cook's distance
- residual plots
- influence functions
- adversarial
- champion/challenger
- global surrogate

local explanations

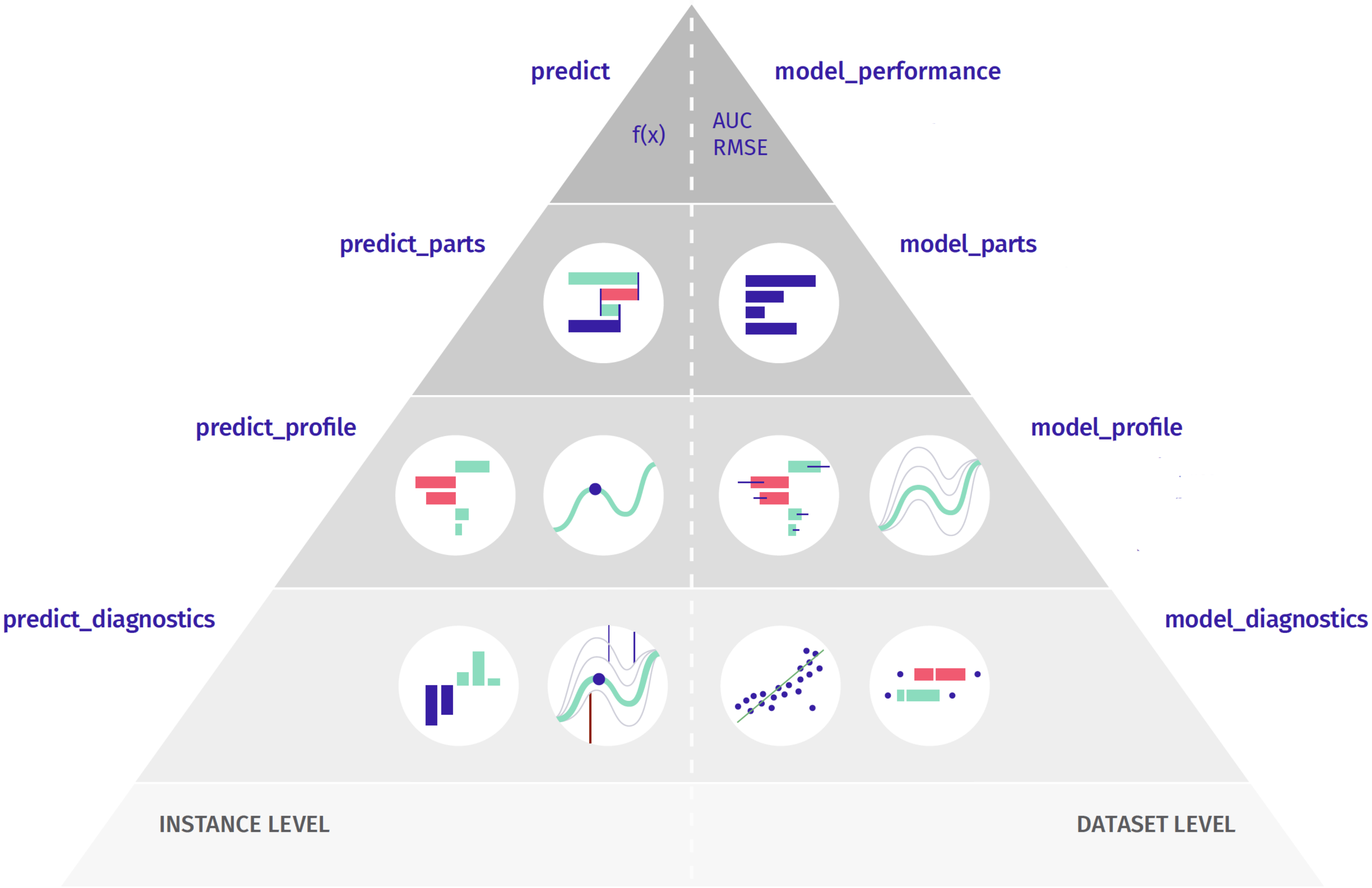
global explanations

* transparent models may be complex

** needs retraining

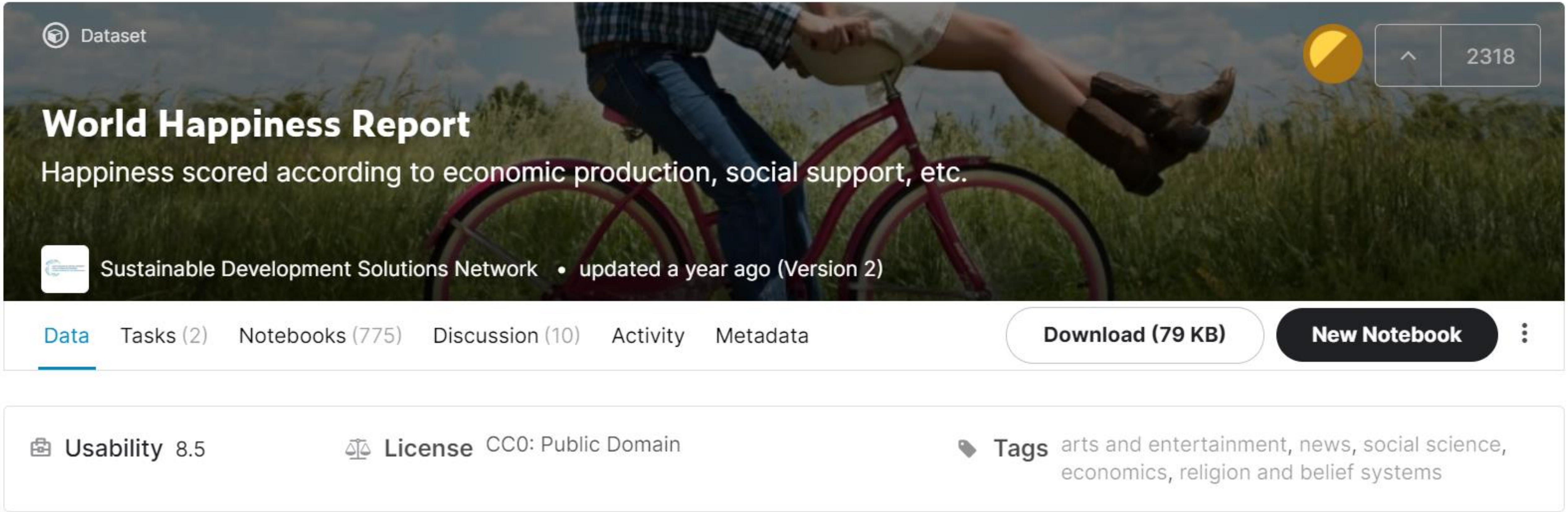
Find more at xai-tools.drwhy.ai

DALEX: moDel Agnostic Language for Exploration and eXplanation



http://dalex.drwhy.ai

Machine learning predictive task



World Happiness Report

Happiness scored according to economic production, social support, etc.

Sustainable Development Solutions Network • updated a year ago (Version 2)

Usability 8.5

License CC0: Public Domain

Tags arts and entertainment, news, social science, economics, religion and belief systems

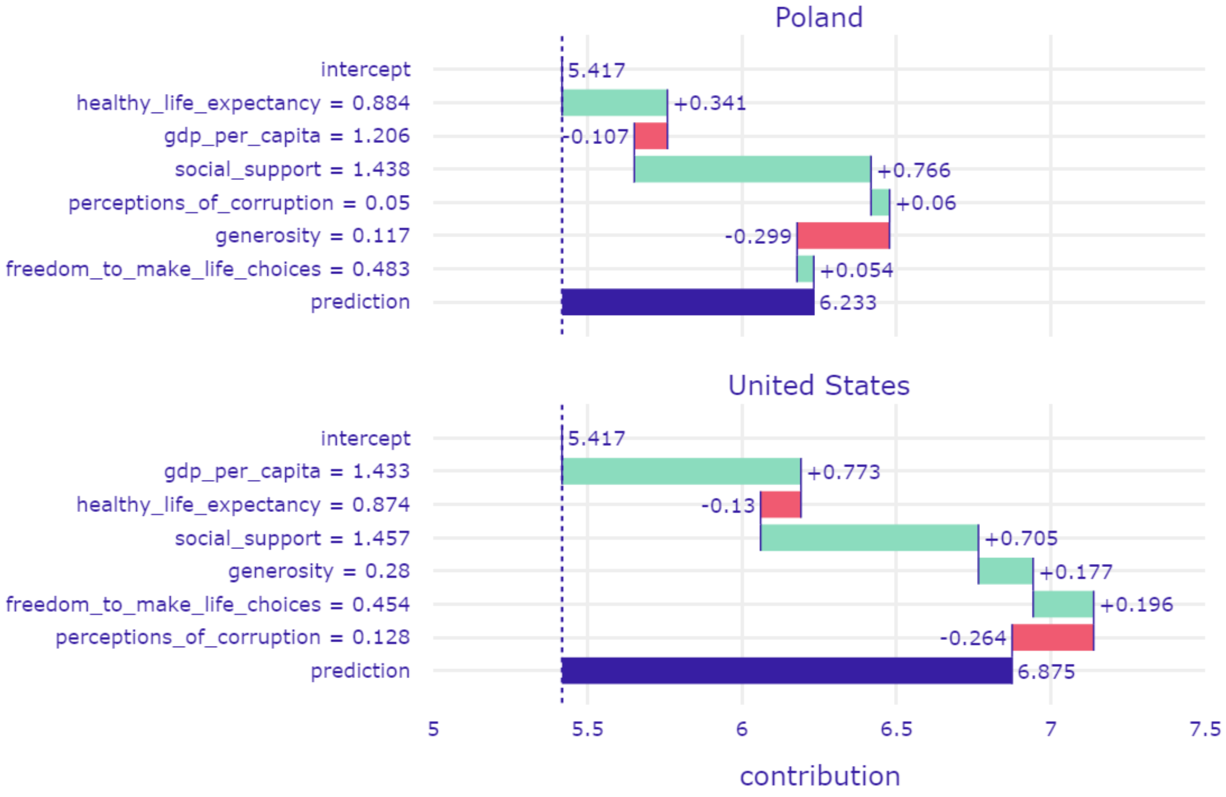
GDP, life expectancy, freedom, social => country happiness score [0, 10]

https://www.kaggle.com/unsdsn/world-happiness

parts

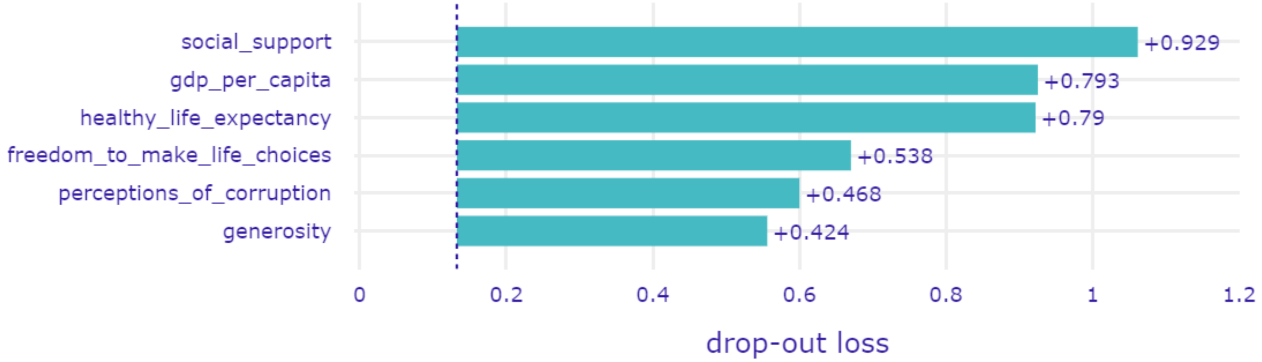
predict_parts

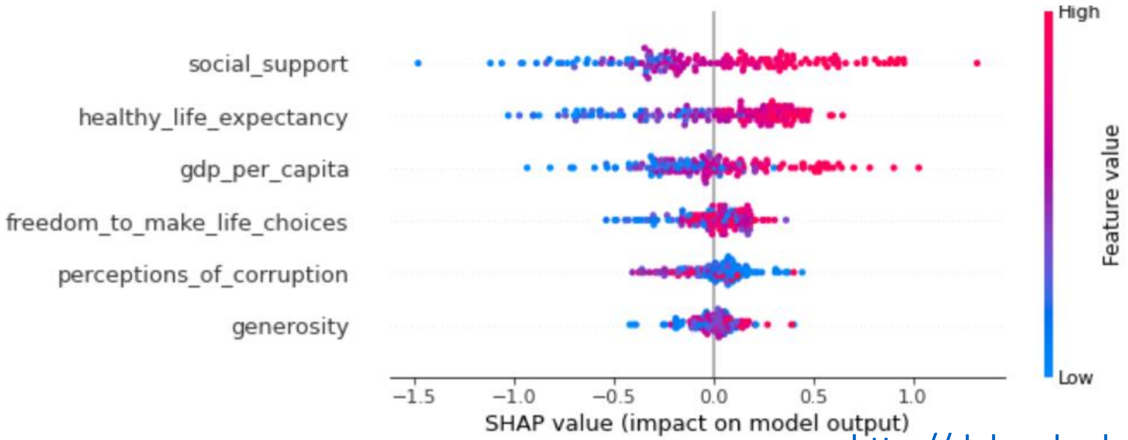
Break Down



model_parts

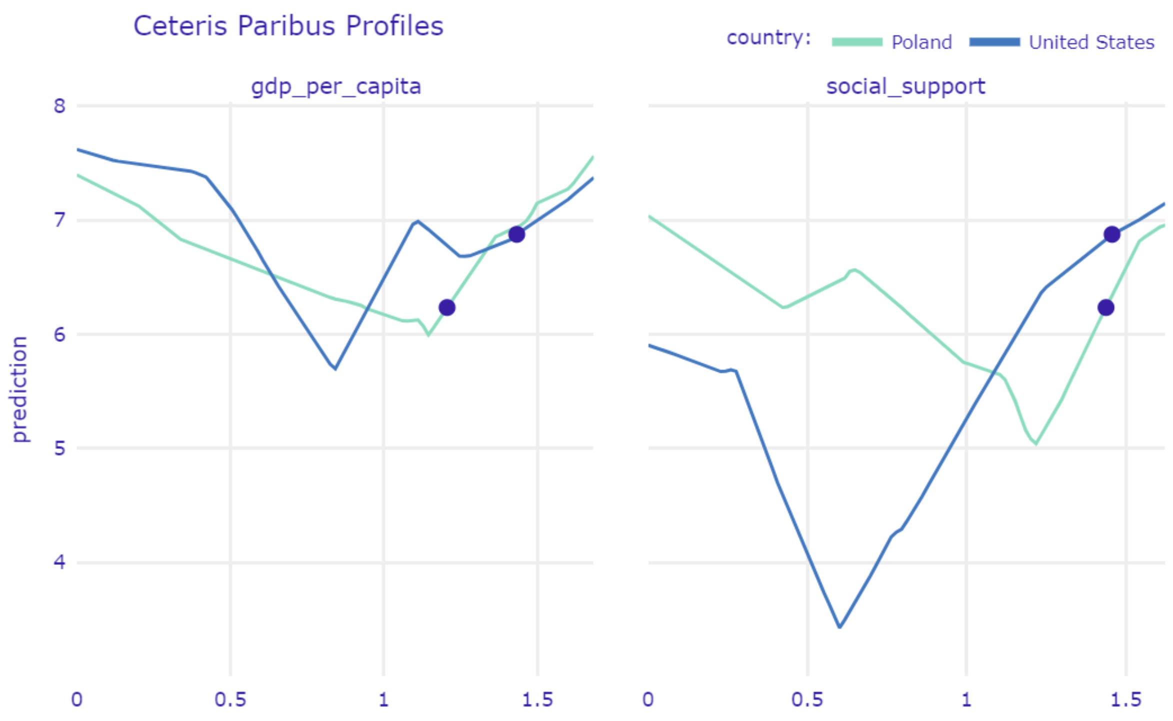
Permutational Importance



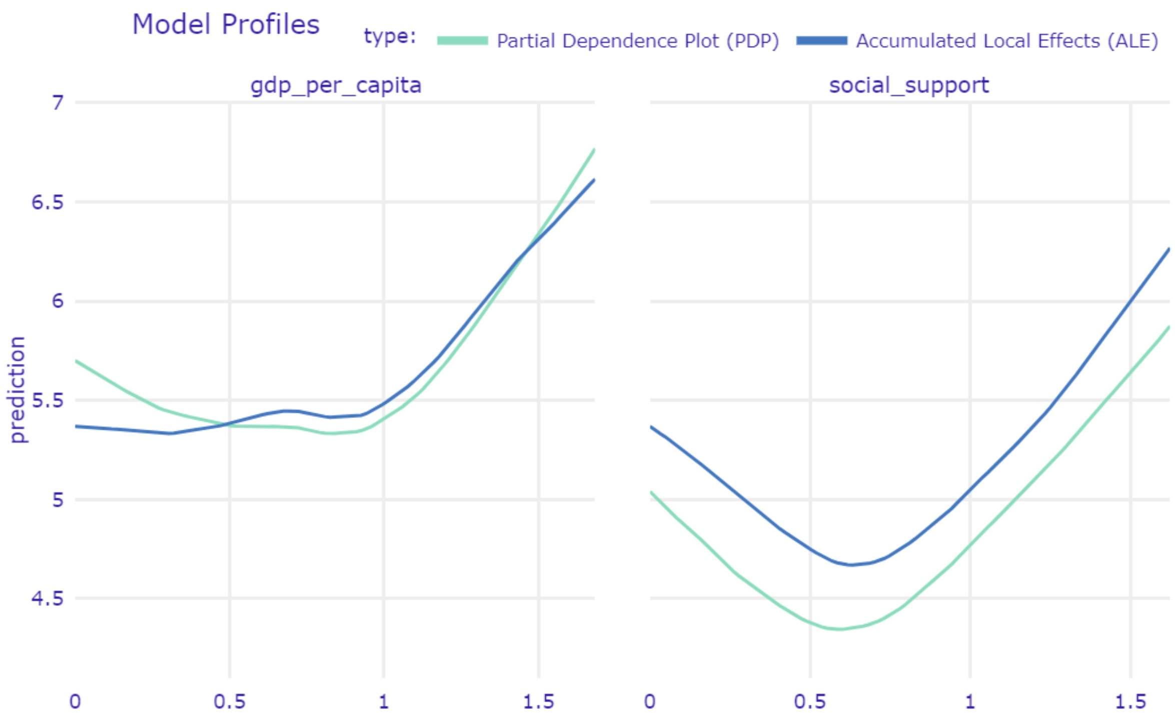


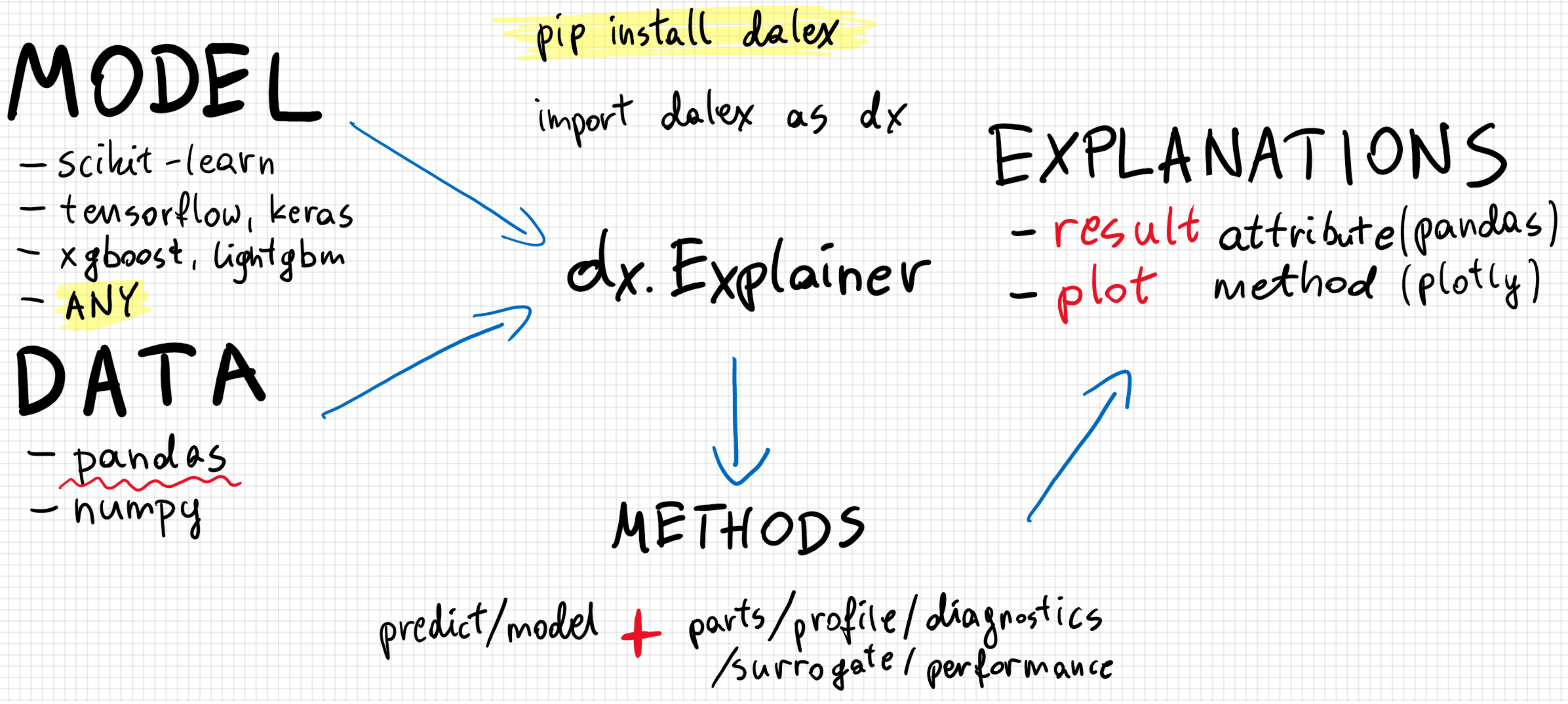
profile

predict_profile



model_profile





```
pip install dalex

import dalex as dx
```

MODEL

- scikit-learn
- tensorflow, keras
- xgboost, lightgbm
- ANY

DATA

- pandas
- numpy

dx.Explainer

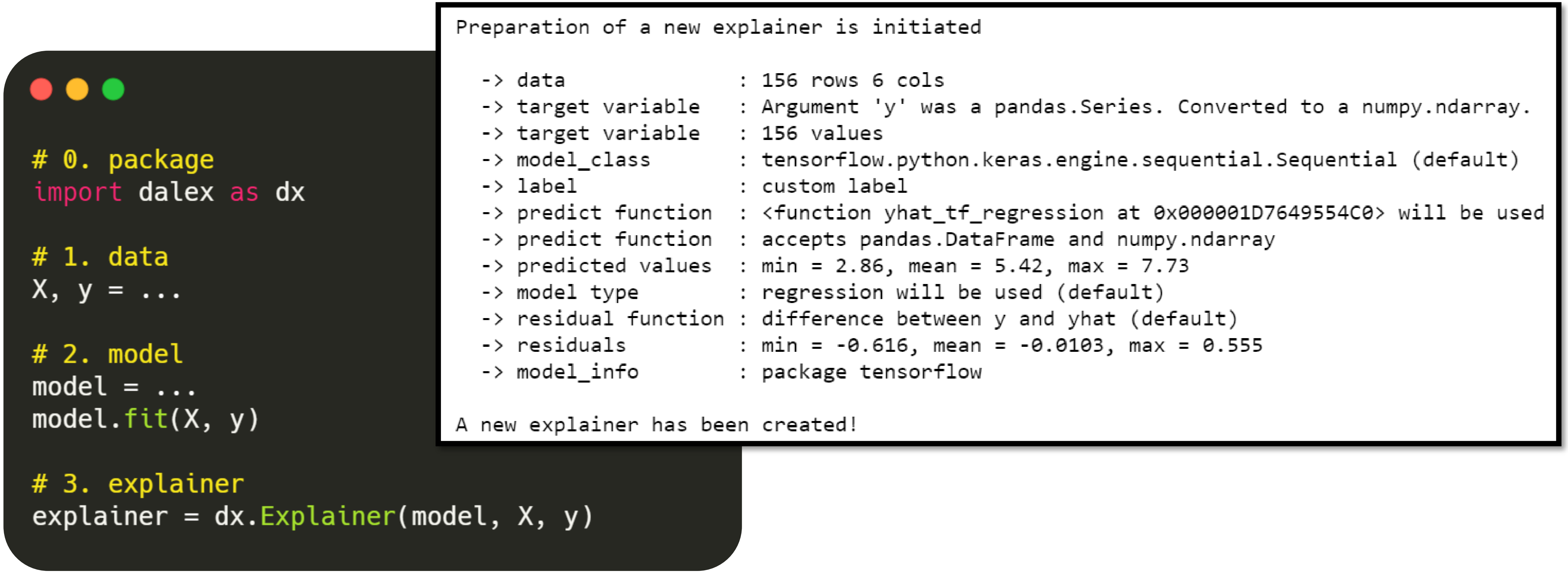
METHODS

predict/model + parts/profile/diagnostics /surrogate/performance

EXPLANATIONS

- result attribute (pandas)
- plot method (plotly)

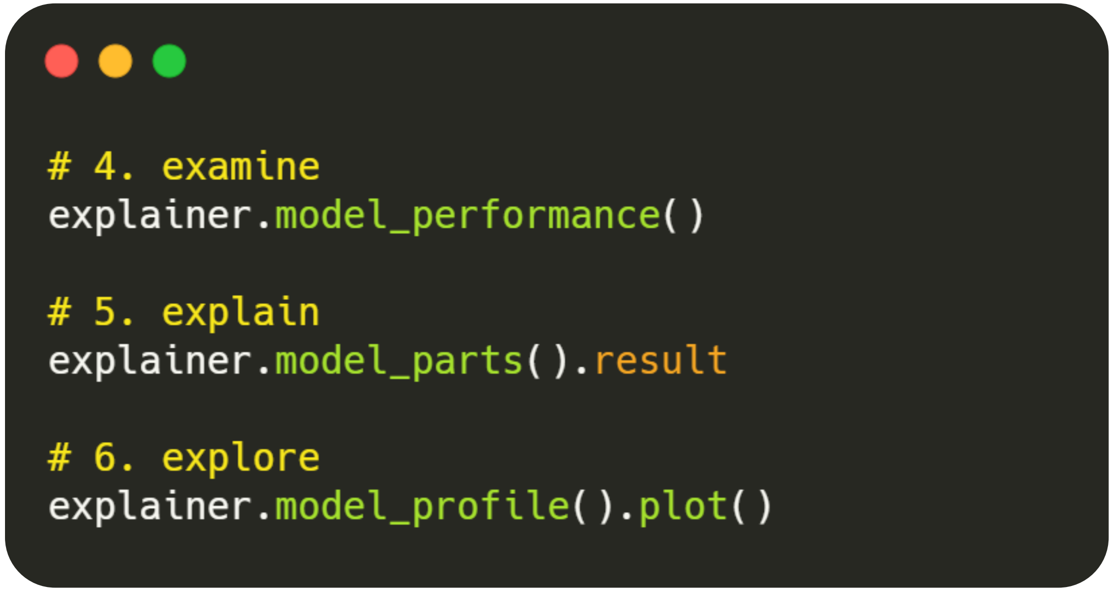
Explainer



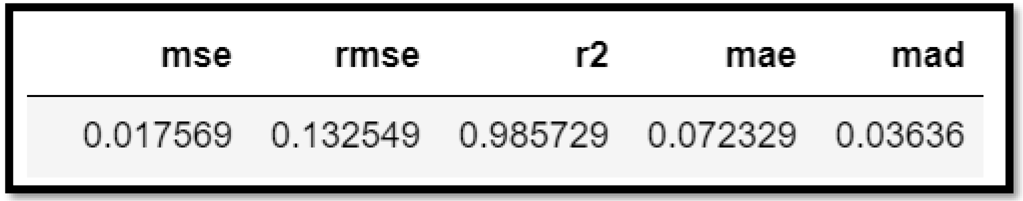
```python
# 0. package
import dalex as dx

# 1. data
X, y = ...

# 2. model
model = ...
model.fit(X, y)

# 3. explainer
explainer = dx.Explainer(model, X, y)
```

```
Preparation of a new explainer is initiated

  -> data              : 156 rows 6 cols
  -> target variable   : Argument 'y' was a pandas.Series. Converted to a numpy.ndarray.
  -> target variable   : 156 values
  -> model_class       : tensorflow.python.keras.engine.sequential.Sequential (default)
  -> label             : custom label
  -> predict function  : <function yhat_tf_regression at 0x000001D7649554C0> will be used
  -> predict function  : accepts pandas.DataFrame and numpy.ndarray
  -> predicted values  : min = 2.86, mean = 5.42, max = 7.73
  -> model type        : regression will be used (default)
  -> residual function : difference between y and yhat (default)
  -> residuals         : min = -0.616, mean = -0.0103, max = 0.555
  -> model_info        : package tensorflow

A new explainer has been created!
```

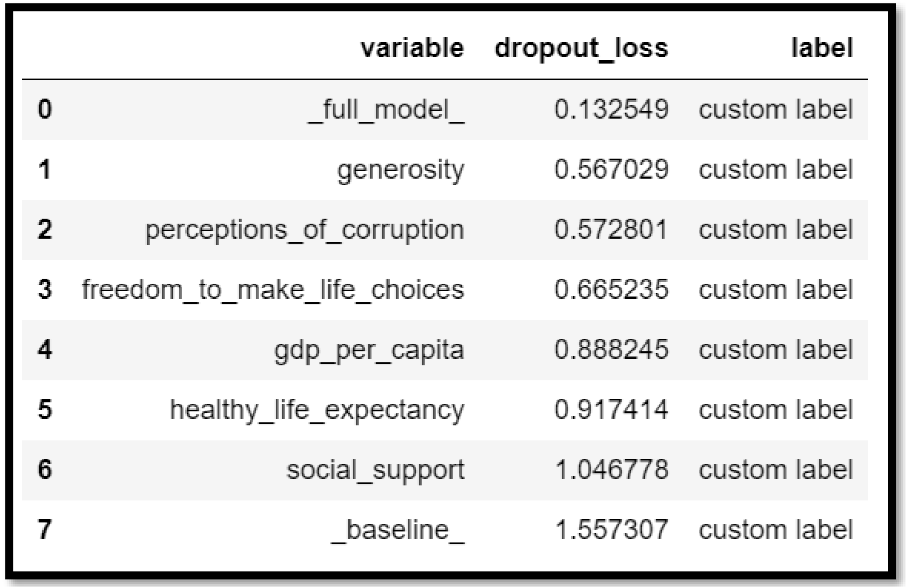
model



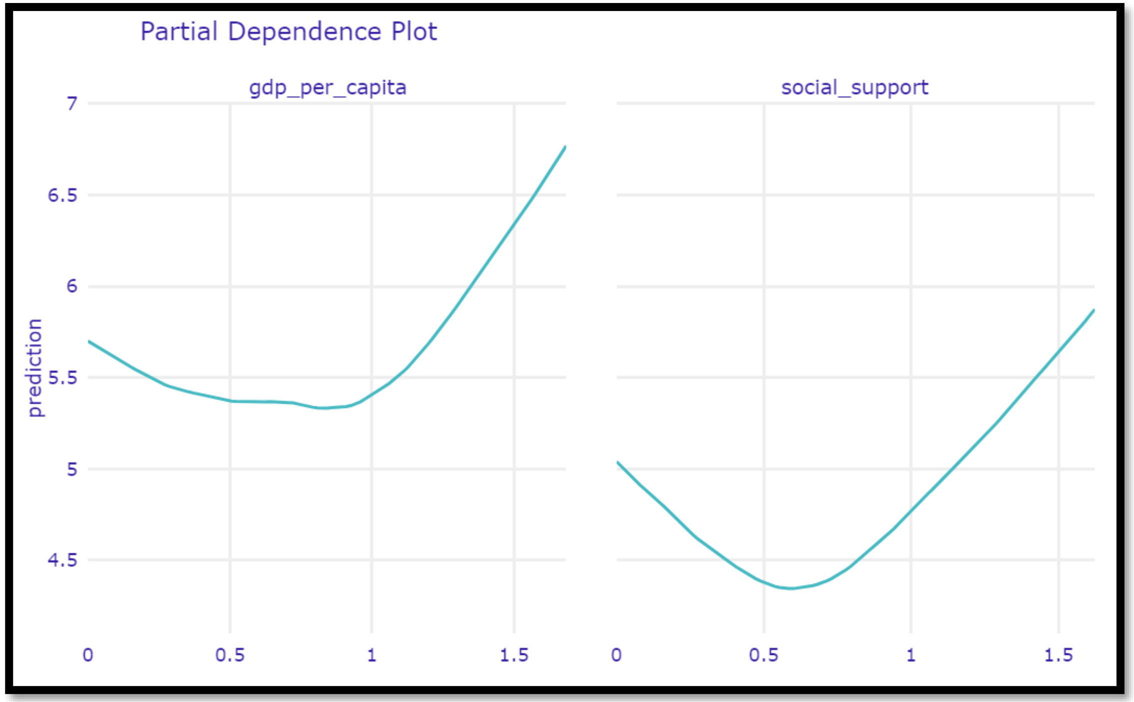
```
# 4. examine
explainer.model_performance()

# 5. explain
explainer.model_parts().result

# 6. explore
explainer.model_profile().plot()
```



mse	rmse	r2	mae	mad
0.017569	0.132549	0.985729	0.072329	0.03636

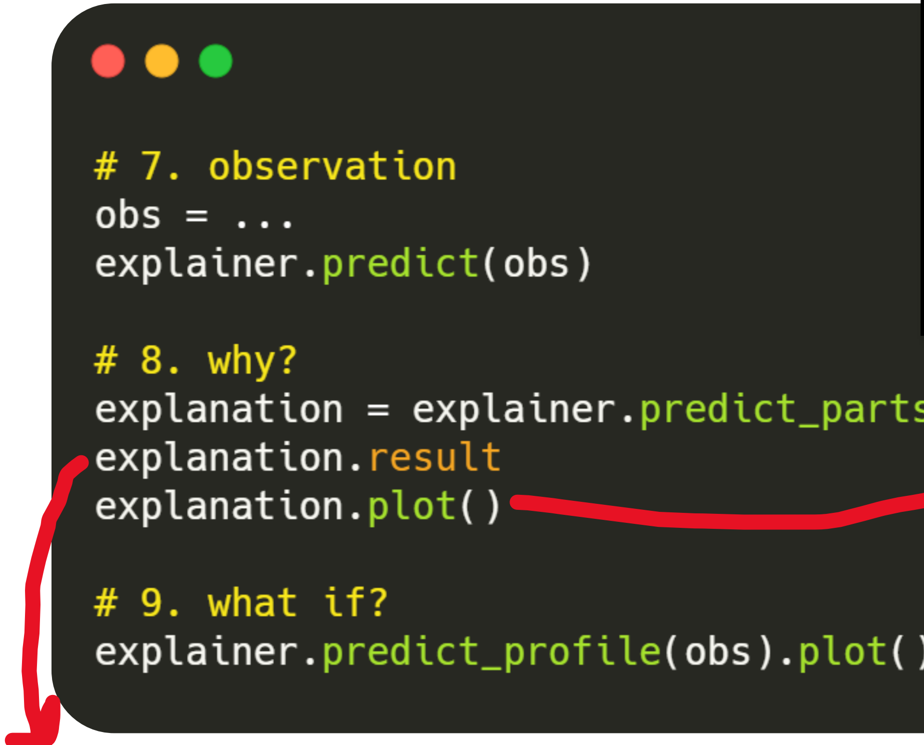


	variable	dropout_loss	label
0	_full_model_	0.132549	custom label
1	generosity	0.567029	custom label
2	perceptions_of_corruption	0.572801	custom label
3	freedom_to_make_life_choices	0.665235	custom label
4	gdp_per_capita	0.888245	custom label
5	healthy_life_expectancy	0.917414	custom label
6	social_support	1.046778	custom label
7	_baseline_	1.557307	custom label

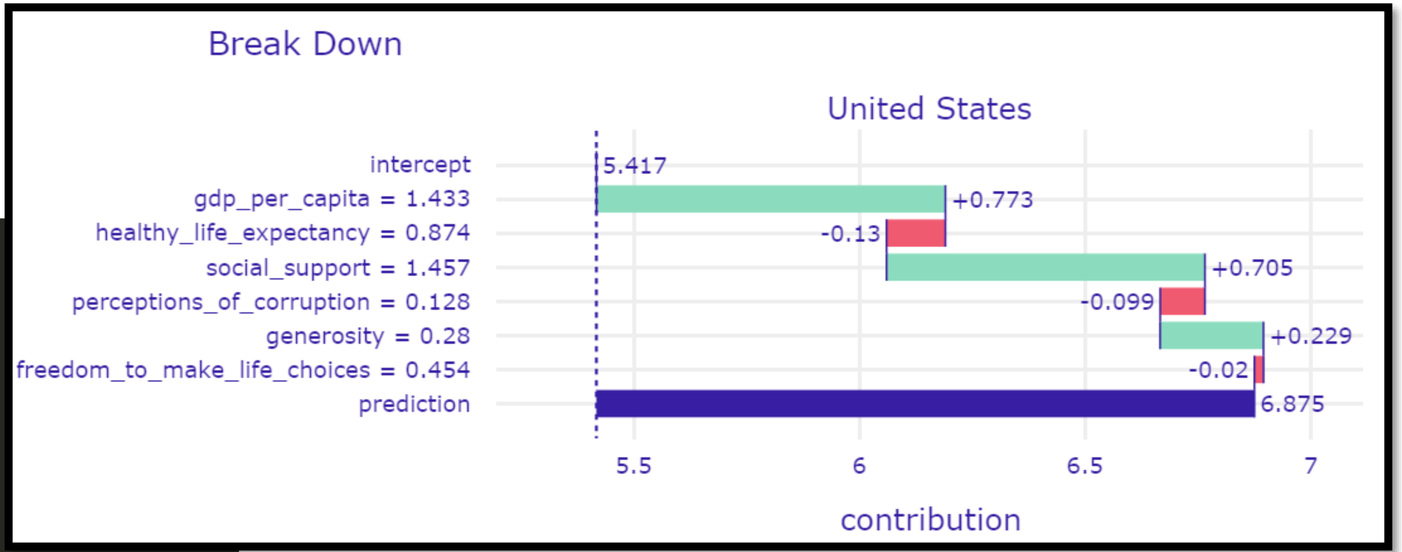


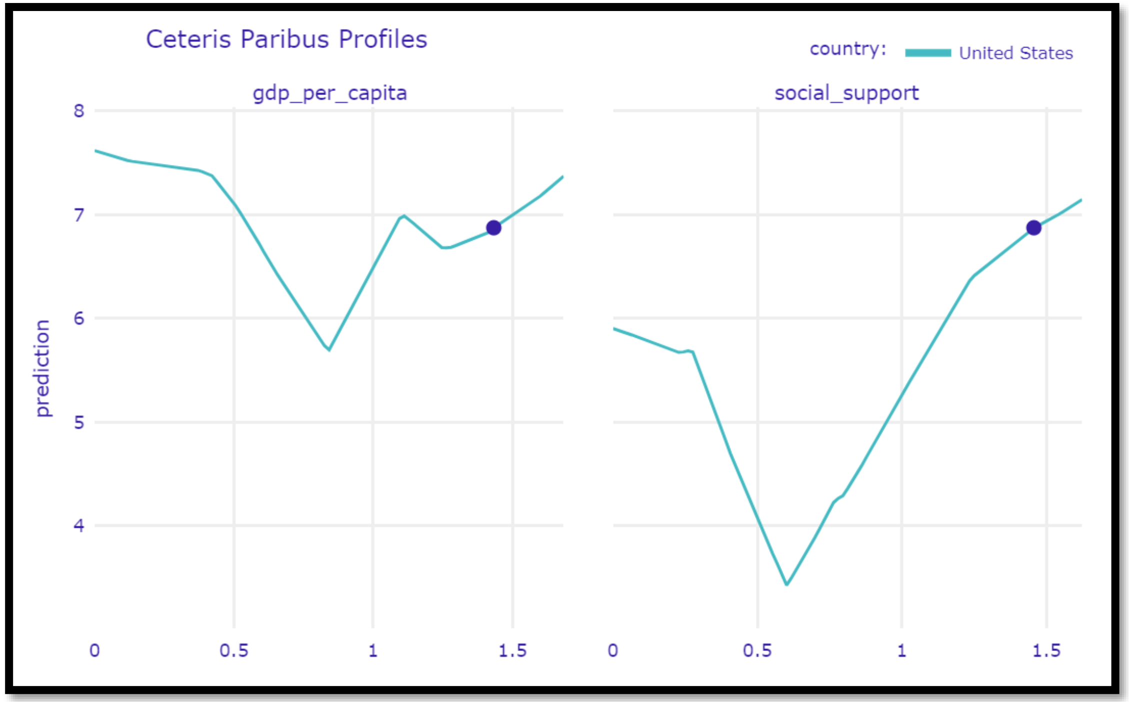
Partial Dependence Plot

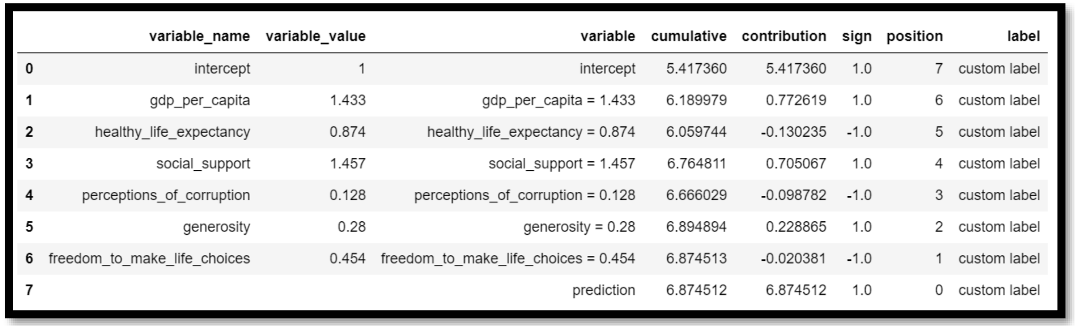
predict



```
# 7. observation
obs = ...
explainer.predict(obs)

# 8. why?
explanation = explainer.predict_parts(obs)
explanation.result
explanation.plot()

# 9. what if?
explainer.predict_profile(obs).plot()
```

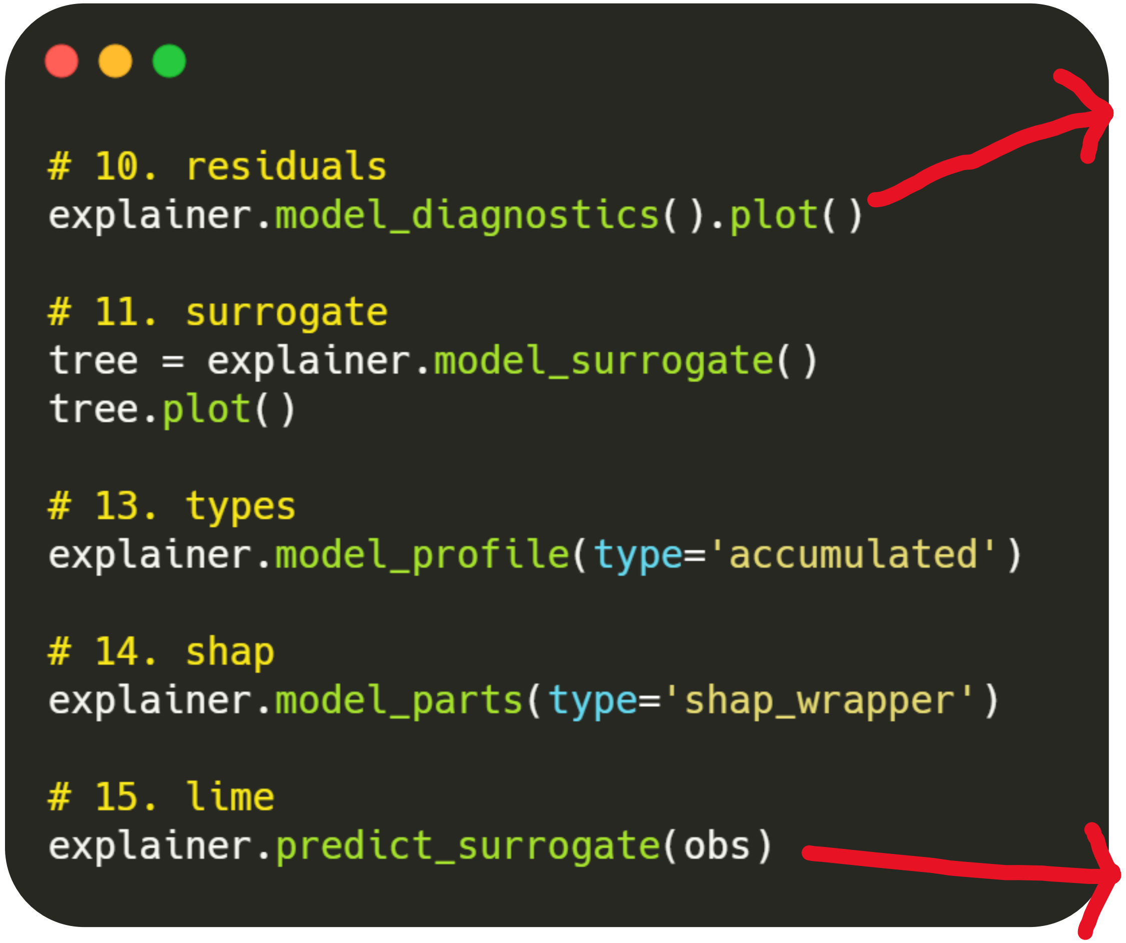




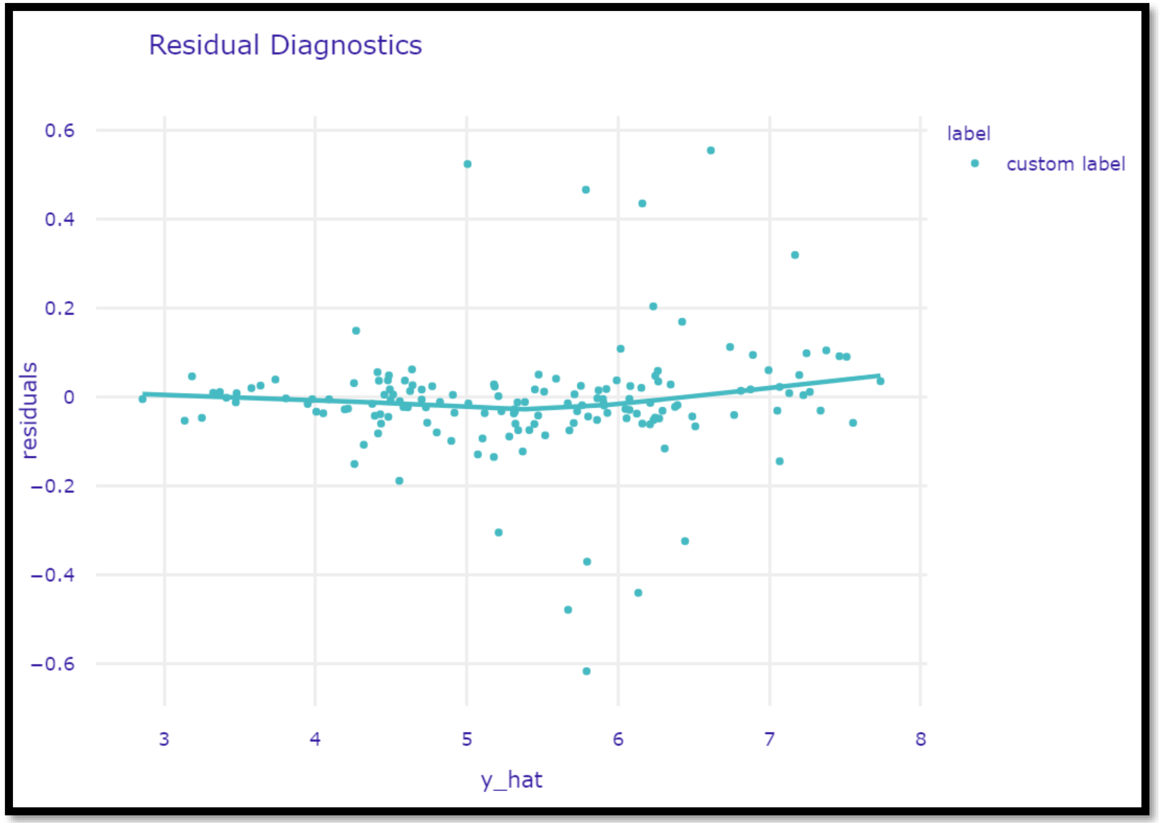


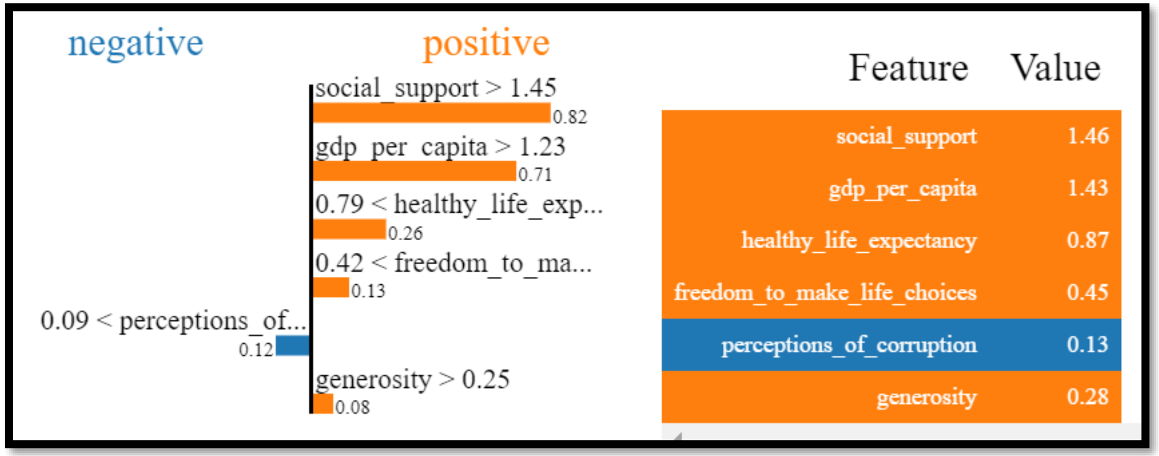
	variable_name	variable_value	variable	cumulative	contribution	sign	position	label
0	intercept	1	intercept	5.417360	5.417360	1.0	7	custom label
1	gdp_per_capita	1.433	gdp_per_capita = 1.433	6.189979	0.772619	1.0	6	custom label
2	healthy_life_expectancy	0.874	healthy_life_expectancy = 0.874	6.059744	-0.130235	-1.0	5	custom label
3	social_support	1.457	social_support = 1.457	6.764811	0.705067	1.0	4	custom label
4	perceptions_of_corruption	0.128	perceptions_of_corruption = 0.128	6.666029	-0.098782	-1.0	3	custom label
5	generosity	0.28	generosity = 0.28	6.894894	0.228865	1.0	2	custom label
6	freedom_to_make_life_choices	0.454	freedom_to_make_life_choices = 0.454	6.874513	-0.020381	-1.0	1	custom label
7			prediction	6.874512	6.874512	1.0	0	custom label

more!

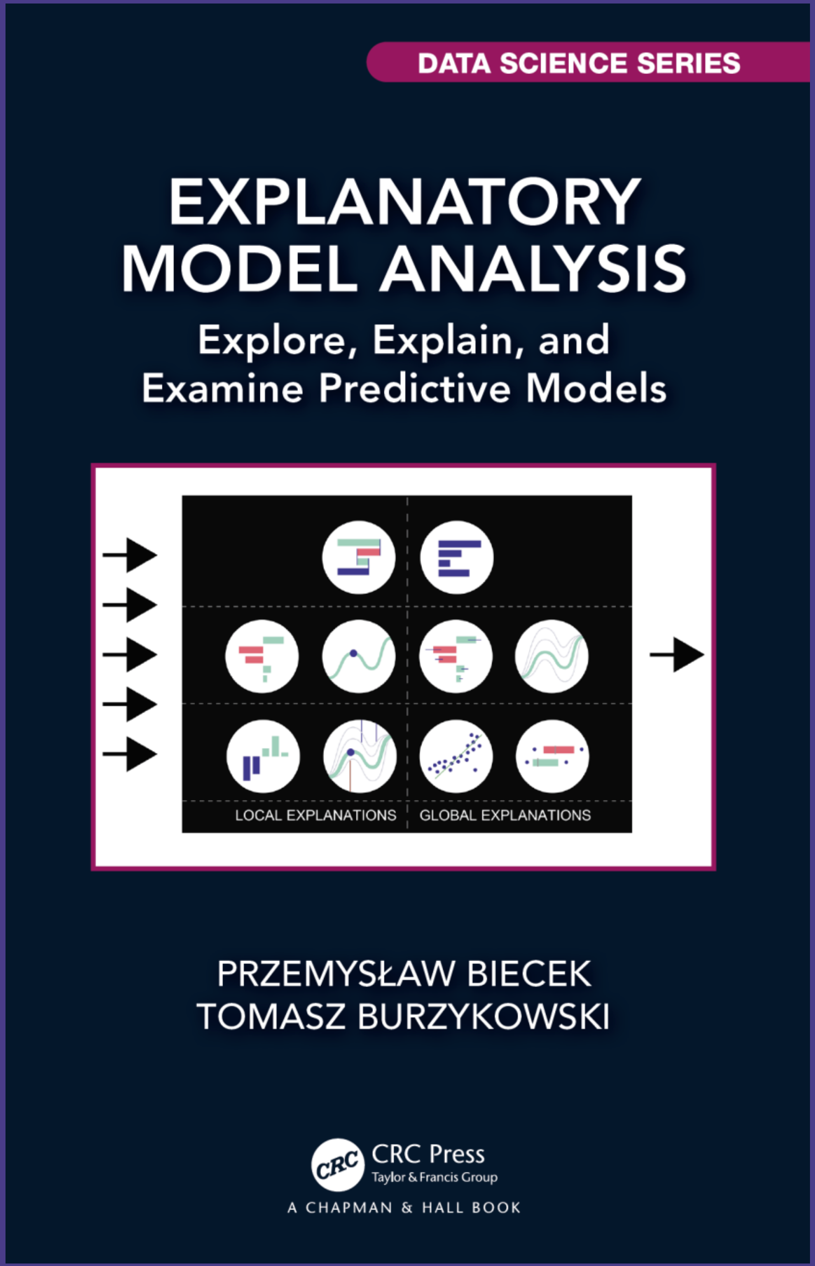


```
# 10. residuals
explainer.model_diagnostics().plot()

# 11. surrogate
tree = explainer.model_surrogate()
tree.plot()

# 13. types
explainer.model_profile(type='accumulated')

# 14. shap
explainer.model_parts(type='shap_wrapper')

# 15. lime
explainer.predict_surrogate(obs)
```





Feature	Value
social_support	1.46
gdp_per_capita	1.43
healthy_life_expectancy	0.87
freedom_to_make_life_choices	0.45
perceptions_of_corruption	0.13
generosity	0.28



DATA SCIENCE SERIES

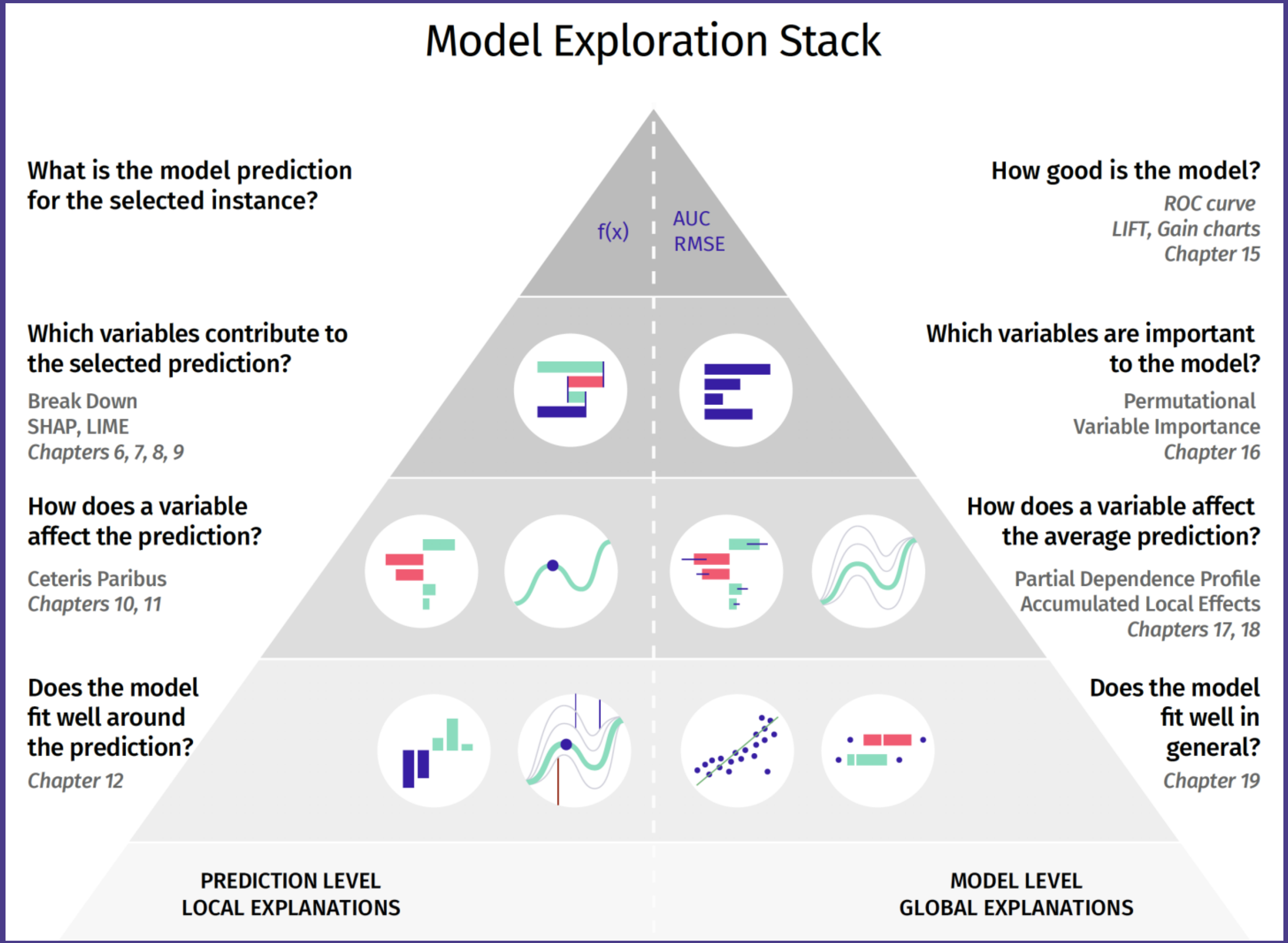
EXPLANATORY MODEL ANALYSIS

Explore, Explain, and Examine Predictive Models

PRZEMYSŁAW BIECEK
TOMASZ BURZYKOWSKI

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

https://pbiecek.github.io/ema/



Model Exploration Stack

What is the model prediction for the selected instance?

How good is the model?
ROC curve
LIFT, Gain charts
Chapter 15

Which variables contribute to the selected prediction?
Break Down
SHAP, LIME
Chapters 6, 7, 8, 9

Which variables are important to the model?
Permutational
Variable Importance
Chapter 16

How does a variable affect the prediction?
Ceteris Paribus
Chapters 10, 11

How does a variable affect the average prediction?
Partial Dependence Profile
Accumulated Local Effects
Chapters 17, 18

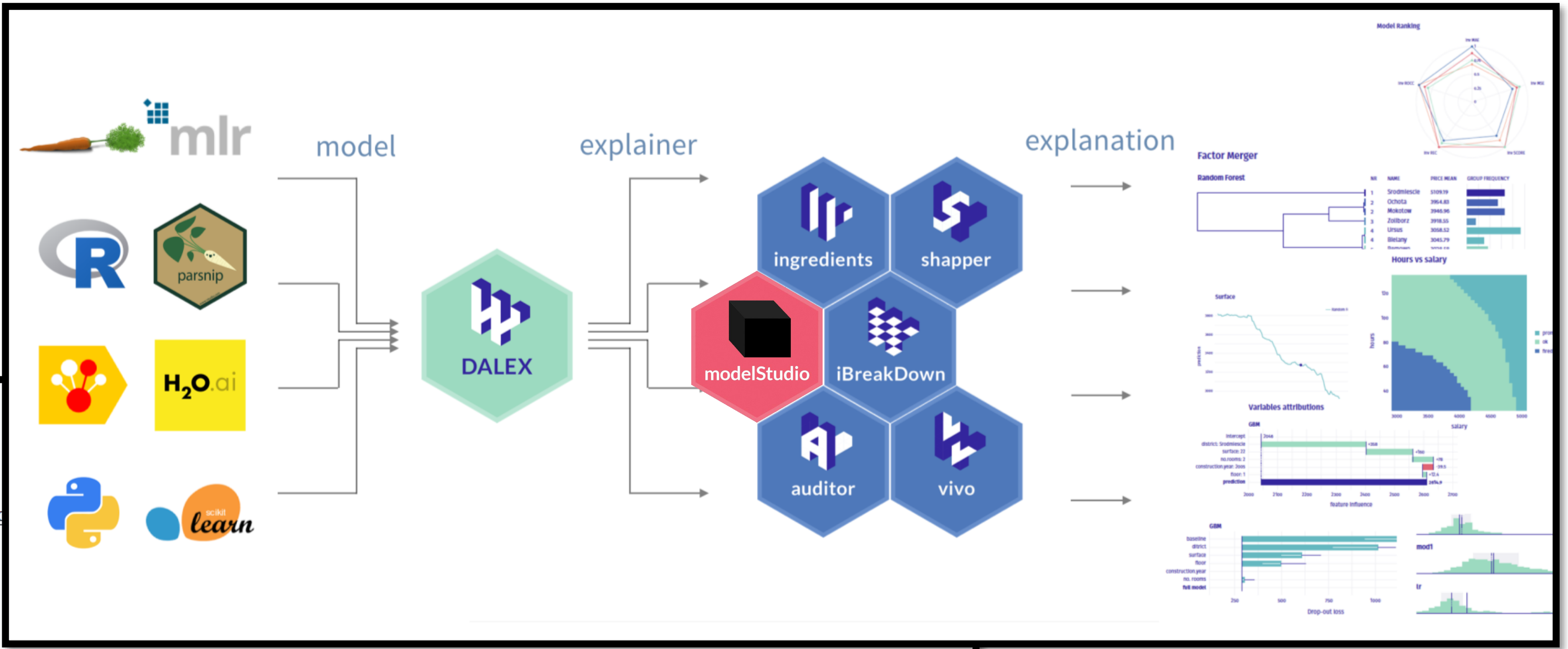
Does the model fit well around the prediction?
Chapter 12

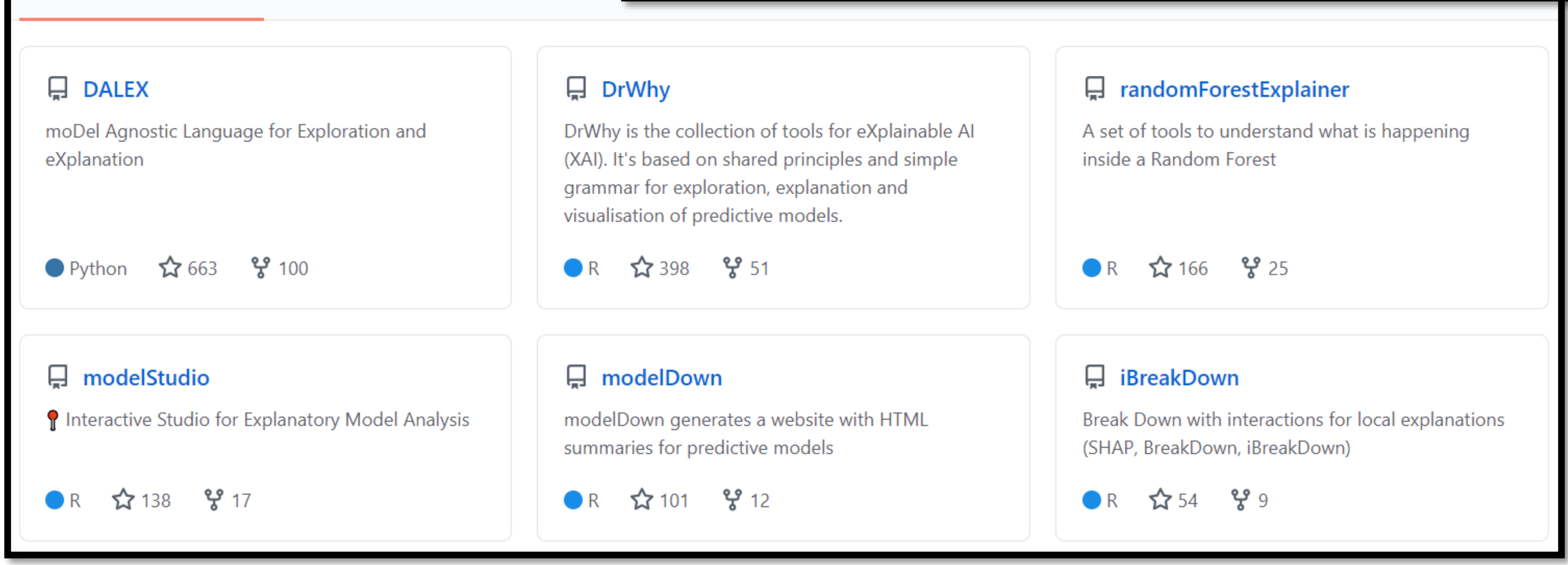
Does the model fit well in general?
Chapter 19

PREDICTION LEVEL
LOCAL EXPLANATIONS

MODEL LEVEL
GLOBAL EXPLANATIONS

DrWhy.AI





Model Oriented

MI2DataLab @ Warsaw University of Technolog

Repositories 46 | Packages | People 21

DALEX
moDel Agnostic Language for Exploration and eXplanation
● Python ☆ 663 100

DrWhy
DrWhy is the collection of tools for eXplainable AI (XAI). It's based on shared principles and simple grammar for exploration, explanation and visualisation of predictive models.
● R ☆ 398 51

randomForestExplainer
A set of tools to understand what is happening inside a Random Forest
● R ☆ 166 25

modelStudio
📍 Interactive Studio for Explanatory Model Analysis
● R ☆ 138 17

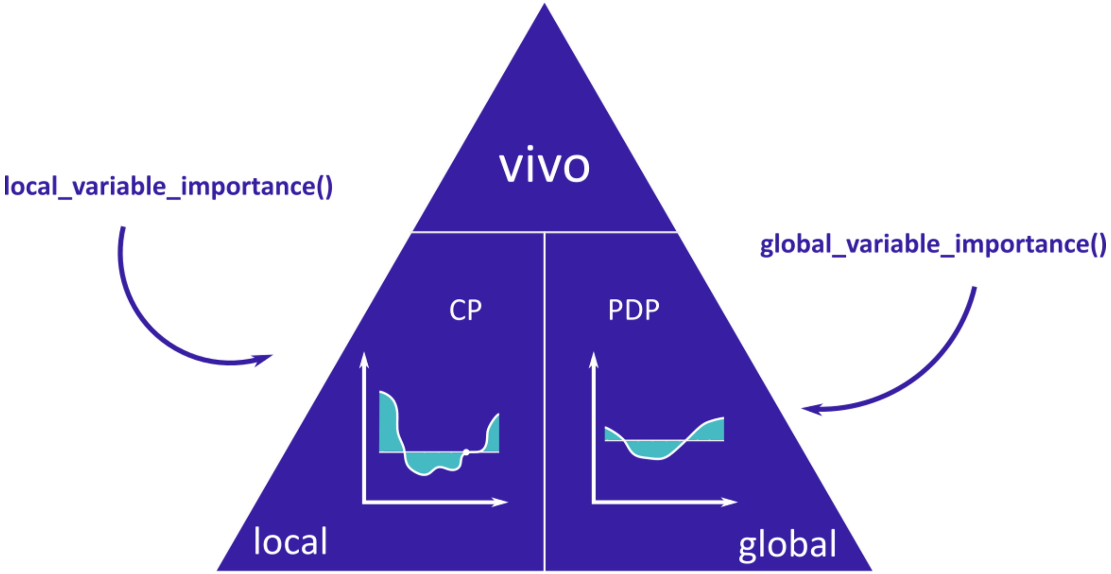
modelDown
modelDown generates a website with HTML summaries for predictive models
● R ☆ 101 12

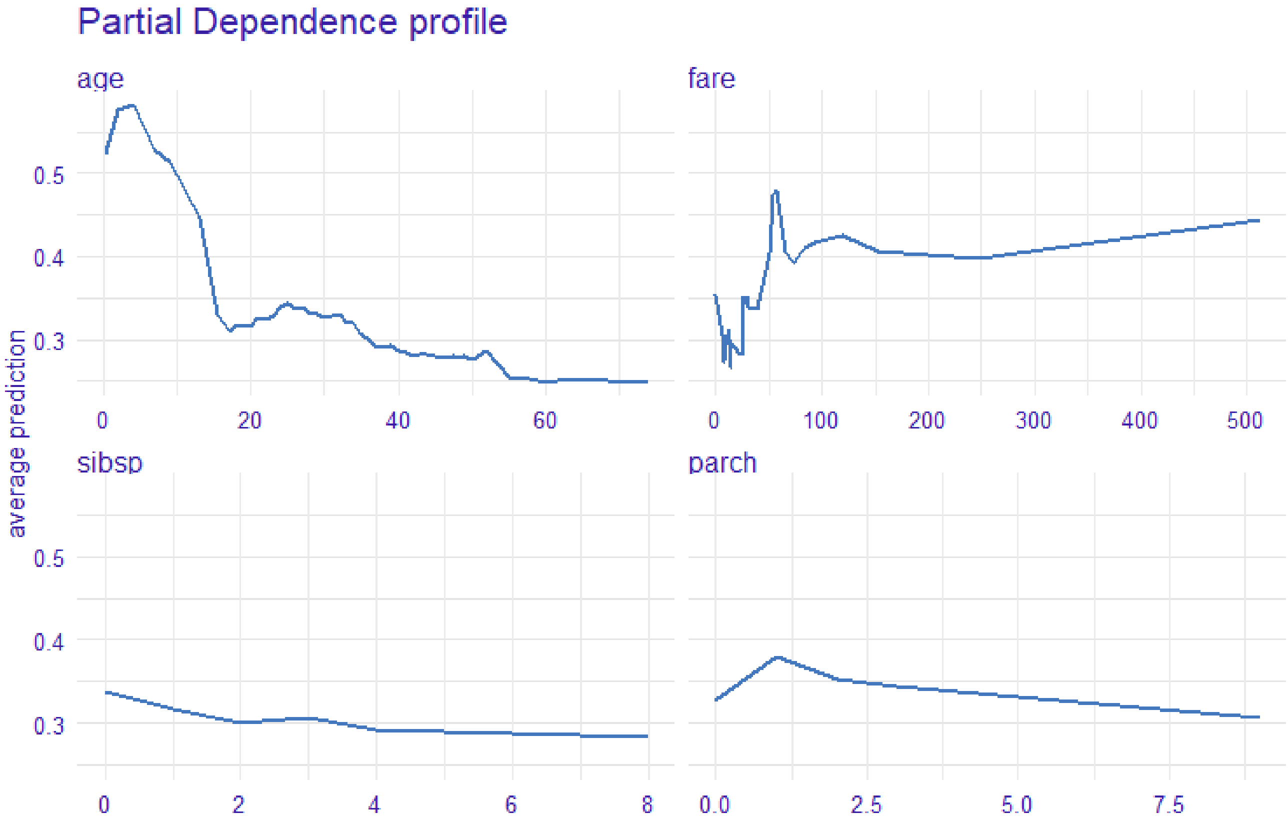
iBreakDown
Break Down with interactions for local explanations (SHAP, BreakDown, iBreakDown)
● R ☆ 54 9

http://drwhy.ai/

vivo

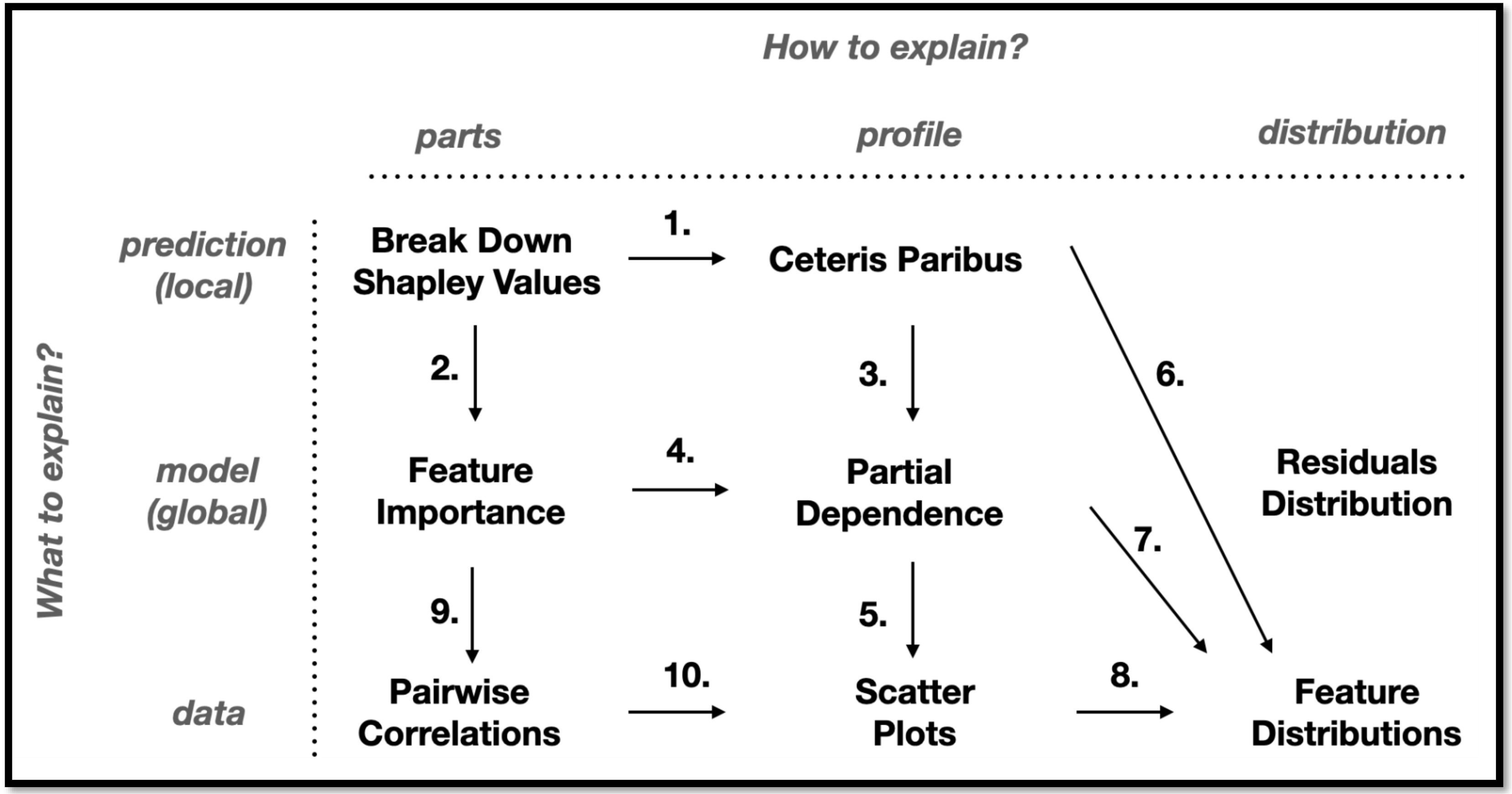
- alternative, model-agnostic way of calculating variable importance
- based on the Ceteris Paribus and Partial Dependence Profiles
- faster, no random component





https://github.com/ModelOriented/vivo

Interactive XAI



How to explain?

parts *profile* *distribution*

What to explain?

prediction (local)

Break Down Shapley Values

1.

Ceteris Paribus

2.

3.

6.

model (global)

Feature Importance

4.

Partial Dependence

Residuals Distribution

7.

9.

5.

data

Pairwise Correlations

10.

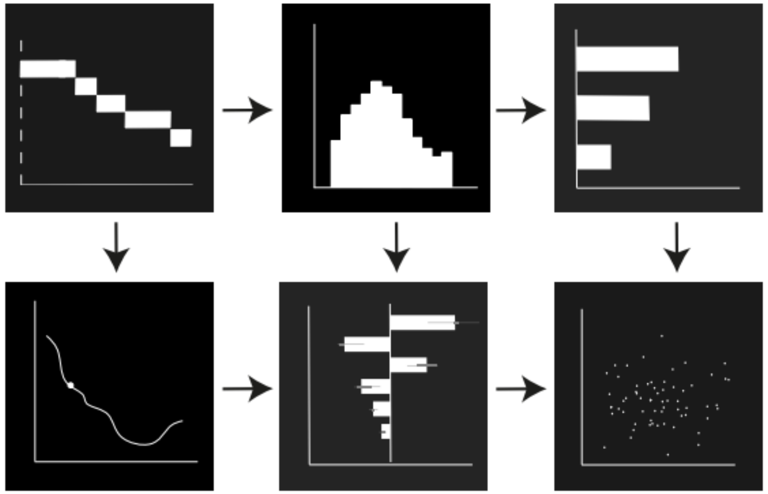
Scatter Plots

8.

Feature Distributions

II generation explanations

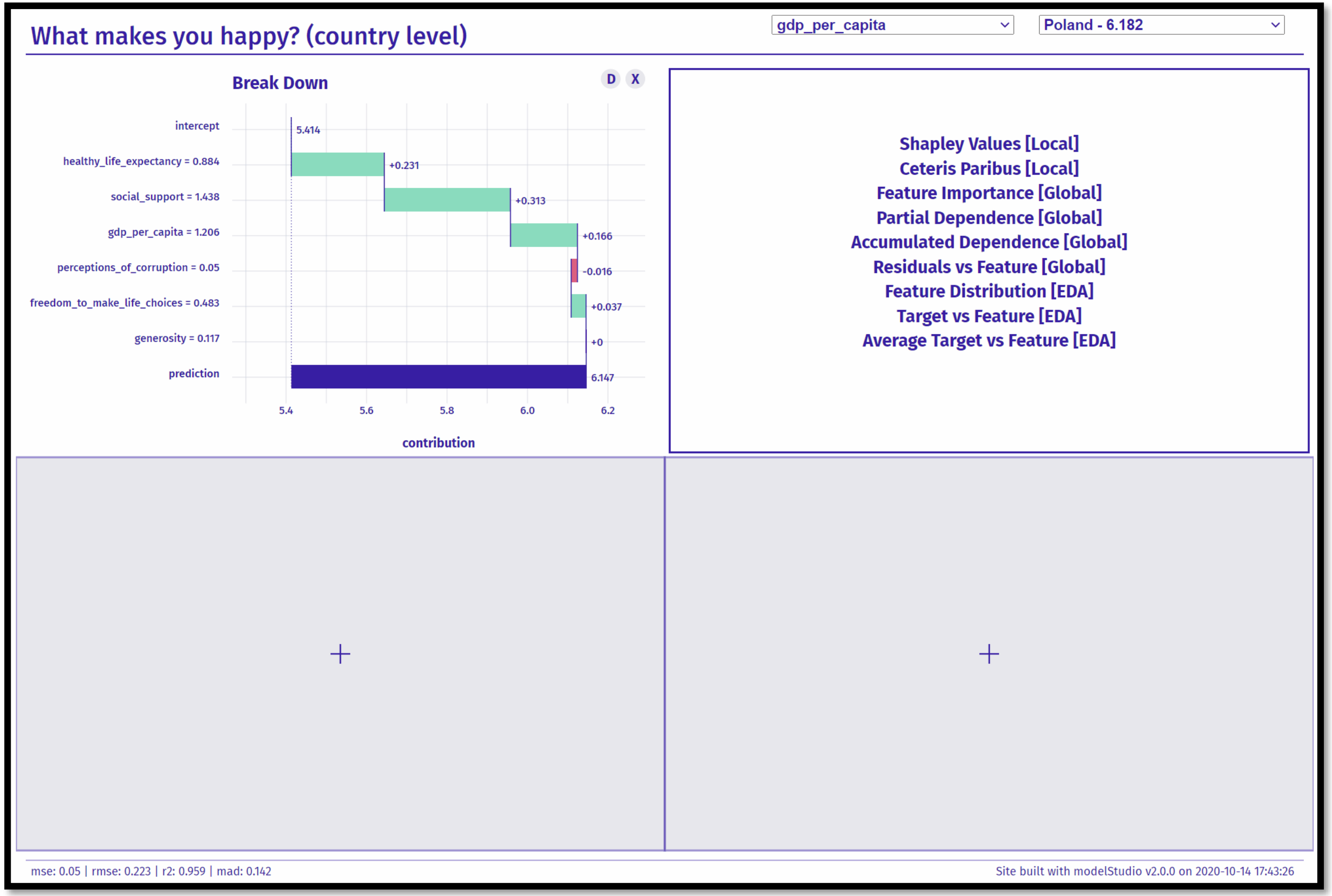
(interactive explanatory model analysis)



https://arxiv.org/abs/2005.00497

modelStudio

- creates a dashboard for interactive Explainable AI
- model explanation and data exploration
- automated calculations
- save & share your analysis



What makes you happy? (country level)

gdp_per_capita

Poland - 6.182

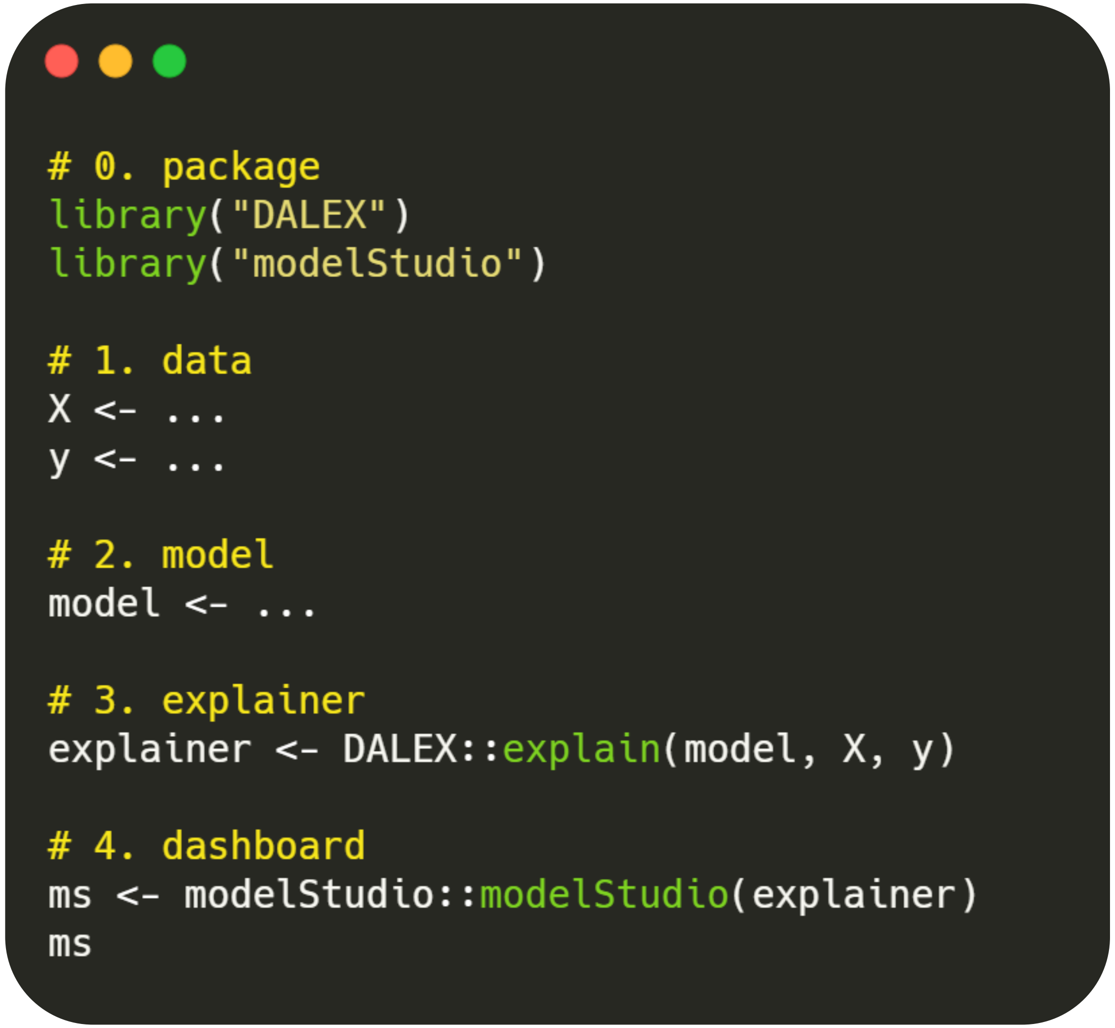
Break Down
D
X
intercept
healthy_life_expectancy = 0.884
social_support = 1.438
gdp_per_capita = 1.206
perceptions_of_corruption = 0.05
freedom_to_make_life_choices = 0.483
generosity = 0.117
prediction
5.414
+0.231
+0.313
+0.166
-0.016
+0.037
+0
6.147
5.4
5.6
5.8
6.0
6.2
contribution
Shapley Values [Local]
Ceteris Paribus [Local]
Feature Importance [Global]
Partial Dependence [Global]
Accumulated Dependence [Global]
Residuals vs Feature [Global]
Feature Distribution [EDA]
Target vs Feature [EDA]
Average Target vs Feature [EDA]
+
+

mse: 0.05 | rmse: 0.223 | r2: 0.959 | mad: 0.142

Site built with modelStudio v2.0.0 on 2020-10-14 17:43:26

https://github.com/ModelOriented/modelStudio

convenient



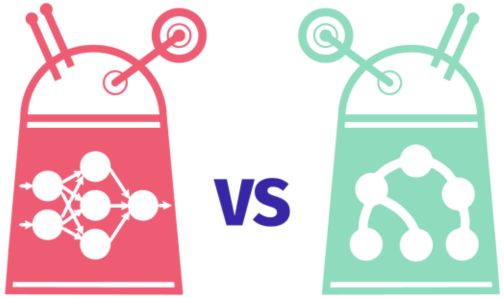
```
# 0. package
library("DALEX")
library("modelStudio")

# 1. data
X <- ...
y <- ...

# 2. model
model <- ...

# 3. explainer
explainer <- DALEX::explain(model, X, y)

# 4. dashboard
ms <- modelStudio::modelStudio(explainer)
ms
```

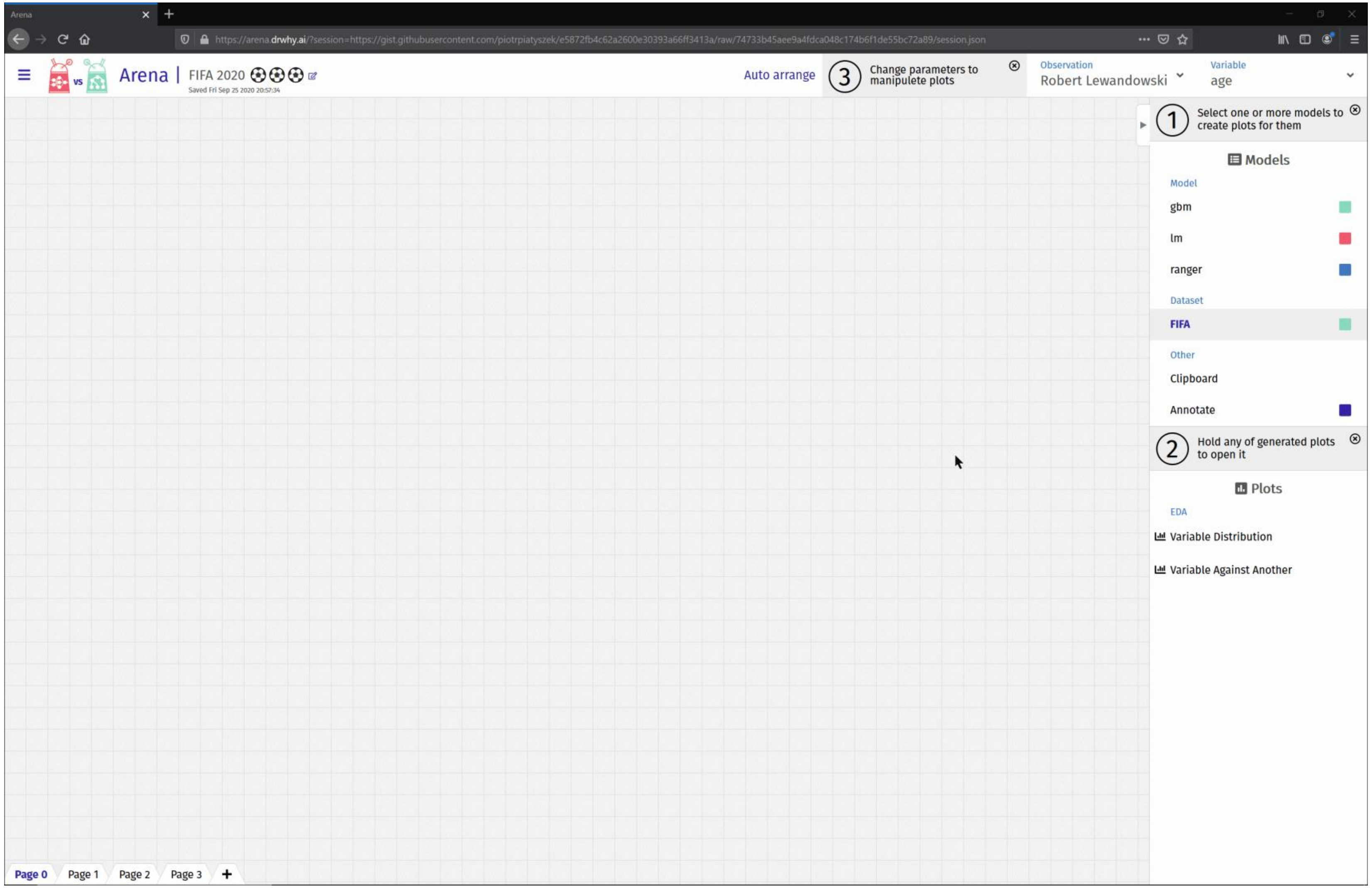
DEMO:

https://pbiecek.github.io/xai-happiness

Arena

- multiple models!
- multiple datasets!!
- more plots (e.g. fairness)
- pages & cache & …
- R & Python (next version)




Arena

https://arena.drwhy.ai/?session=https://gist.githubusercontent.com/piotrpiatyszek/e5872fb4c62a2600e30393a66ff3413a/raw/74733b45aee9a4fdca048c174b6f1de55bc72a89/session.json

Arena | FIFA 2020 ⚽⚽⚽

Saved Fri Sep 25 2020 20:57:34

Auto arrange

3 Change parameters to manipulete plots

Observation: Robert Lewandowski

Variable: age

1 Select one or more models to create plots for them

Models

Model
- gbm 🟩
- lm 🟥
- ranger 🟦

Dataset
- FIFA 🟩

Other
- Clipboard
- Annotate 🟦

2 Hold any of generated plots to open it

Plots

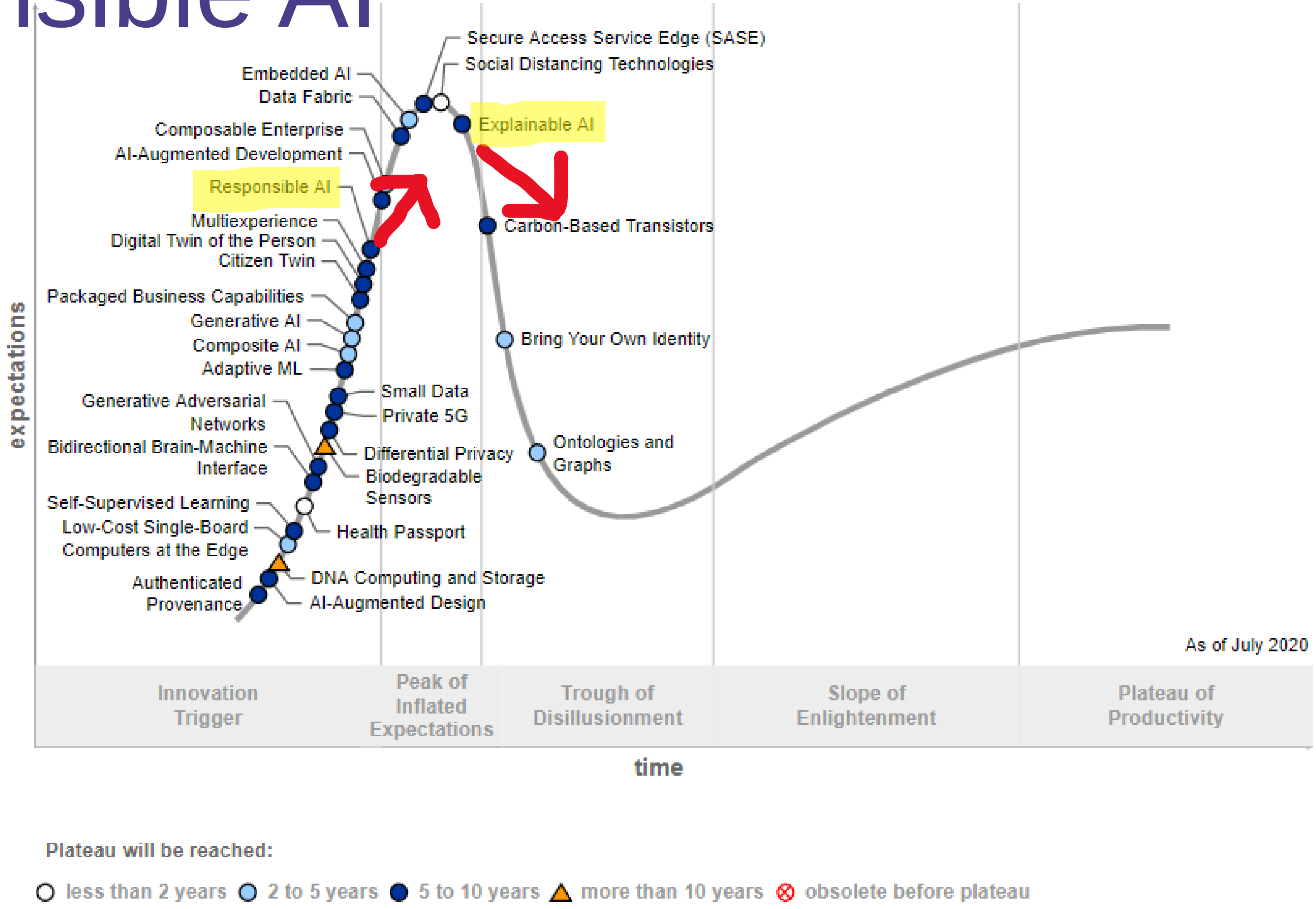
EDA
- Variable Distribution
- Variable Against Another

Page 0 | Page 1 | Page 2 | Page 3 | +

http://arena.drwhy.ai

Responsible AI (RAI)

Hype Cycle for Artificial Intelligence, 2020

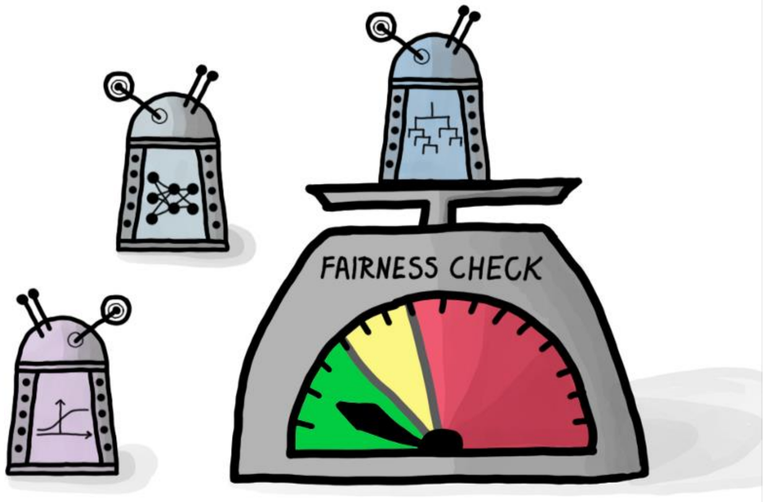


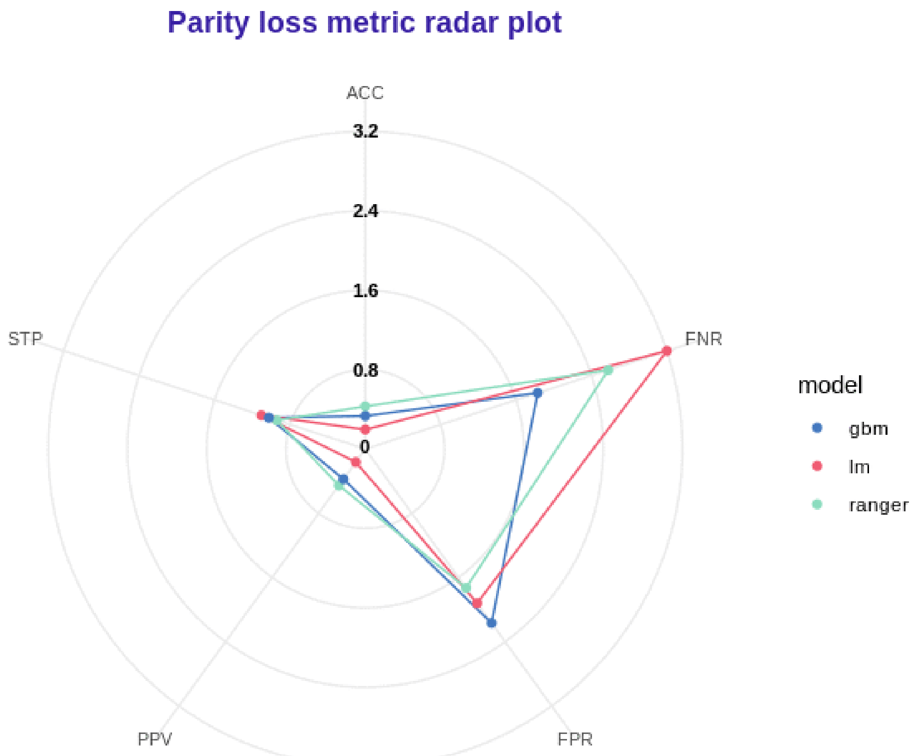
gartner.com/SmarterWithGartner

Source: Gartner

fairmodels

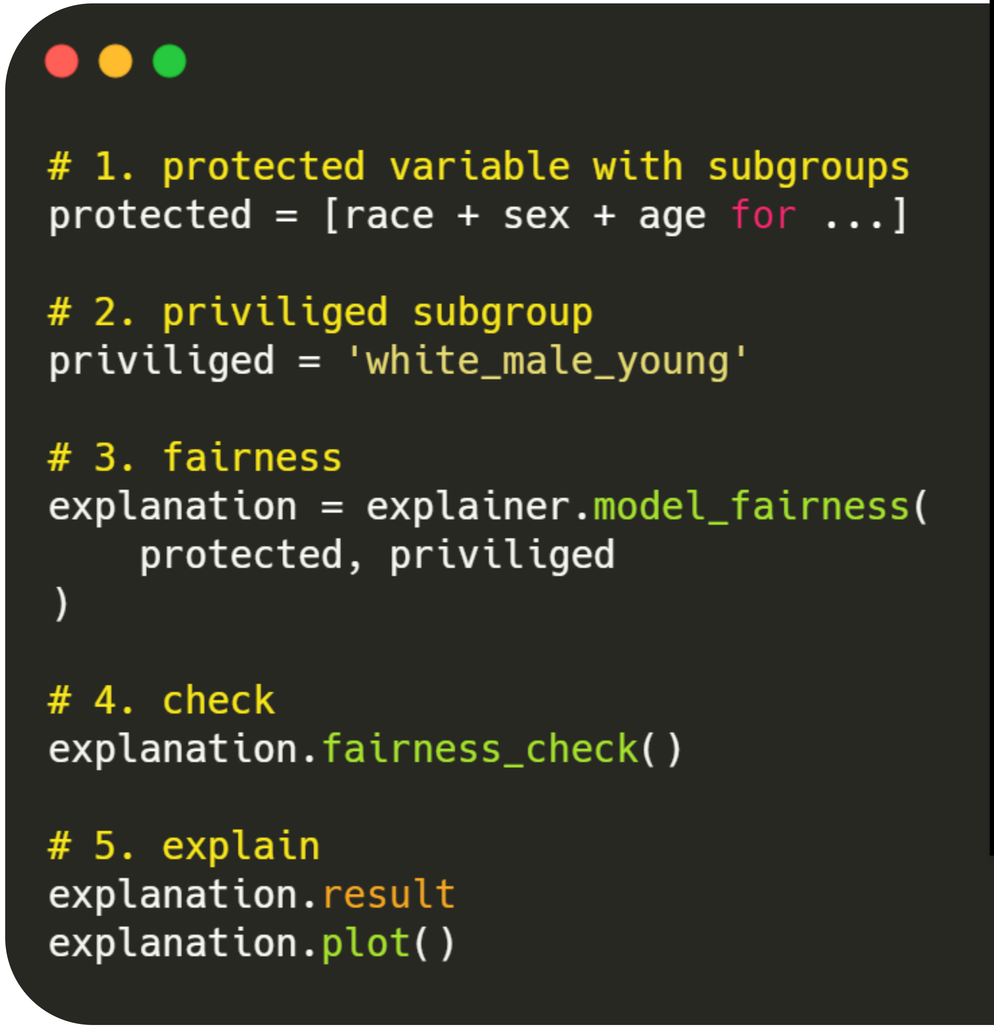
- check model fairness in respect to sensitive categorical variables
- pre- and post- bias mitigation
- compare measures for multiple models
- various techniques and visualisations



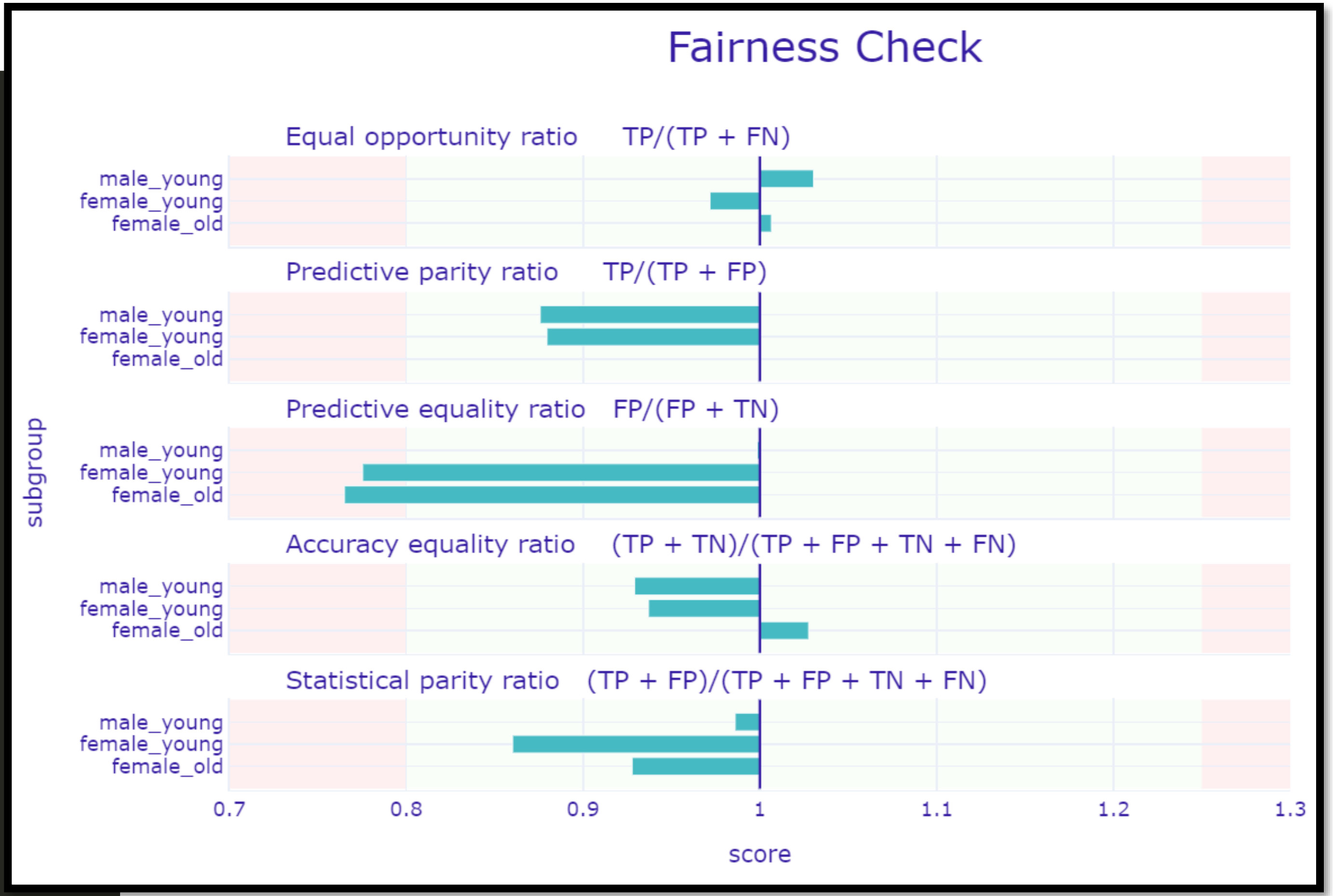


https://github.com/ModelOriented/fairmodels

fairness check



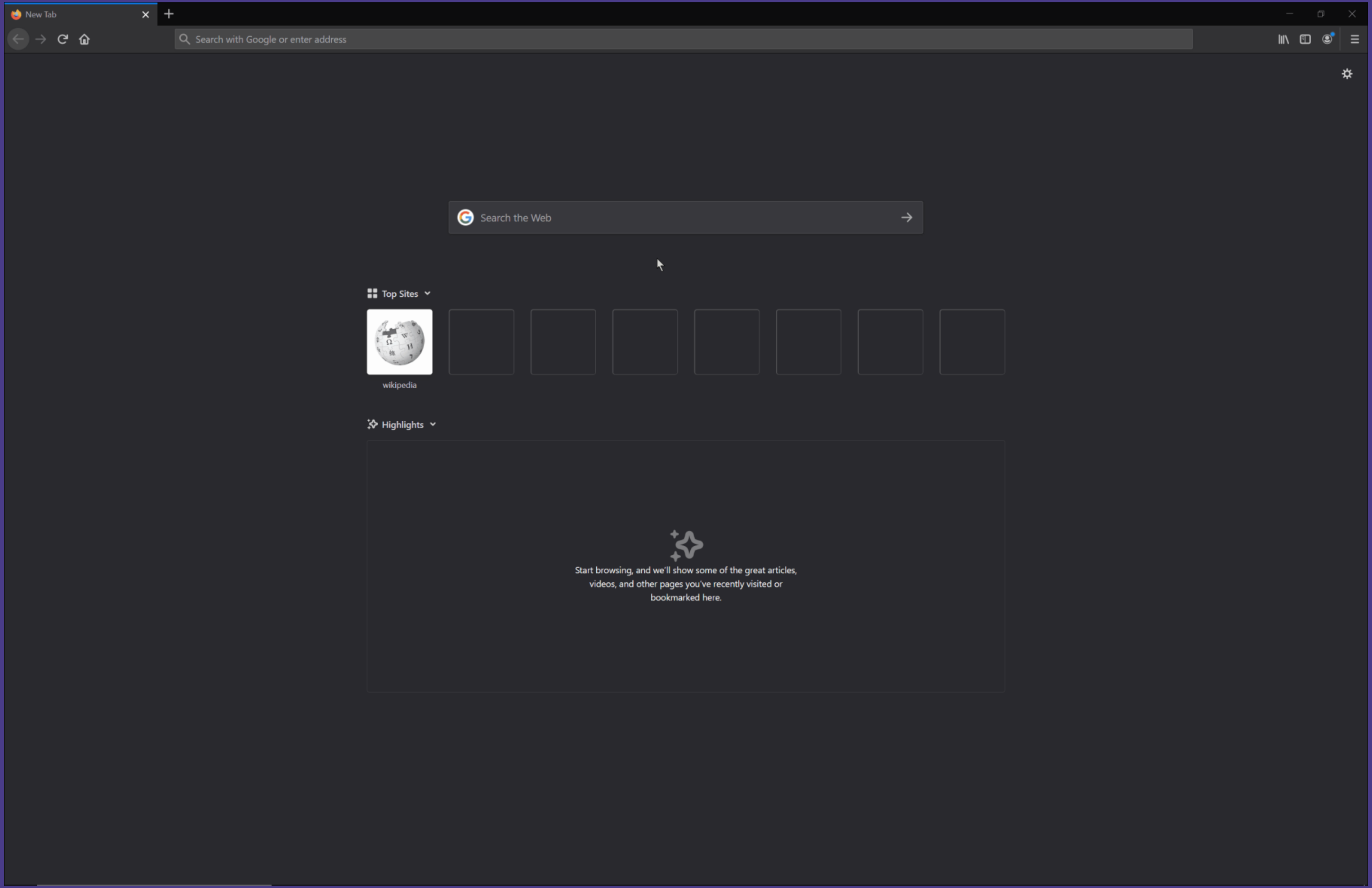
```
# 1. protected variable with subgroups
protected = [race + sex + age for ...]

# 2. priviliged subgroup
priviliged = 'white_male_young'

# 3. fairness
explanation = explainer.model_fairness(
    protected, priviliged
)

# 4. check
explanation.fairness_check()

# 5. explain
explanation.result
explanation.plot()
```



http://dalex.drwhy.ai

DrWhy.AI blog: Responsible ML

https://medium.com/responsibleml



Start browsing, and we'll show some of the great articles, videos, and other pages you've recently visited or bookmarked here.

Feedback apprieciated !

Contact me	linkedin.com/in/hbaniecki
DALEX	dalex.drwhy.ai
DrWhy.AI	drwhy.ai
Blog	medium.com/responsibleml