Explore, Explain, and Debug

Human-Centered Interpretable Machine Learning

Przemysław Biecek

Model explanations

Why should we care?

Apple Card and Goldman Sachs accused of gender discrimination in credit card algorithm

Posted earlier today at 6:38am

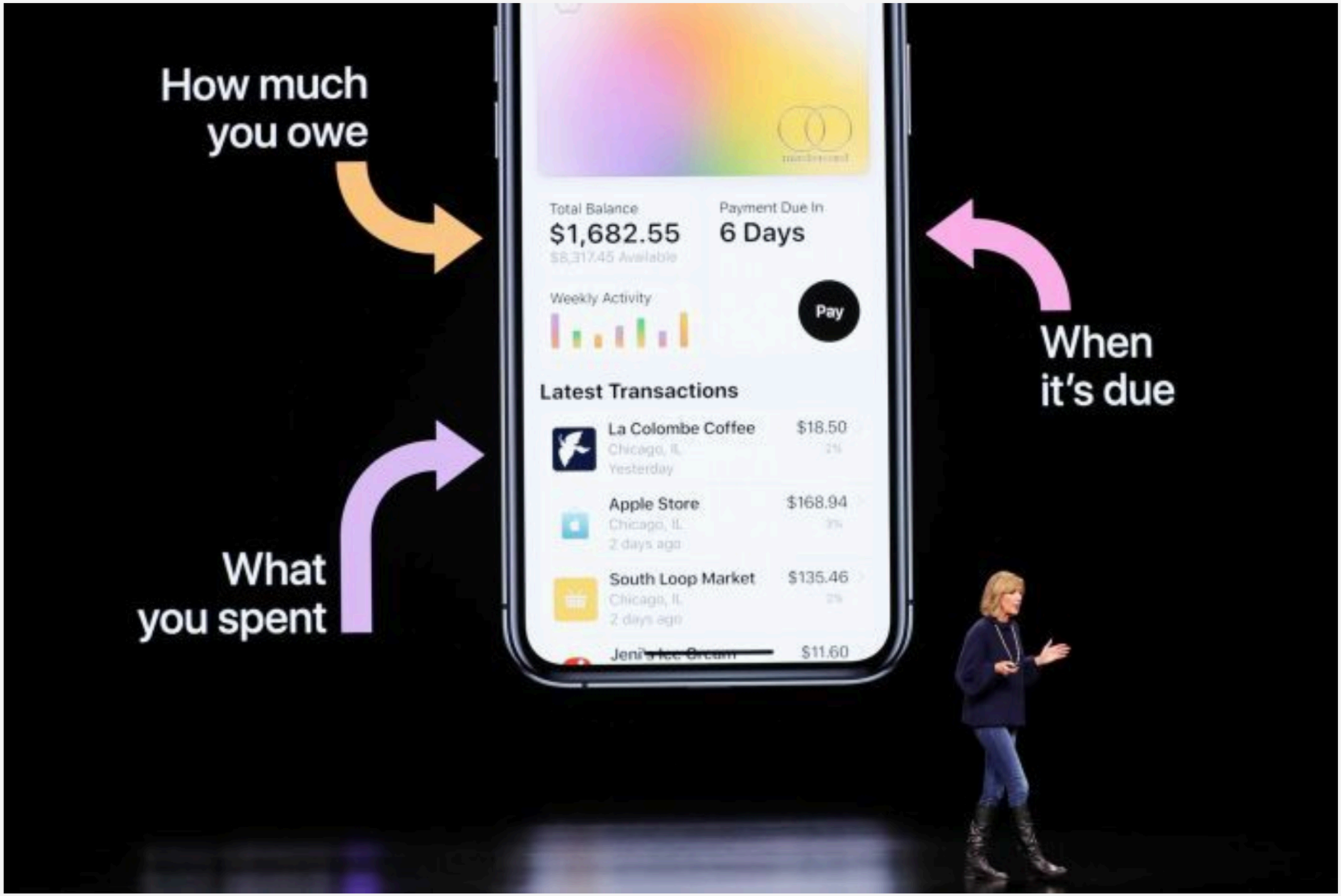


PHOTO: Apple Card has caused a Twitter storm and even prompted a regulator inquiry. (Reuters: Stephen Lam)

BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / A MONTH AGO

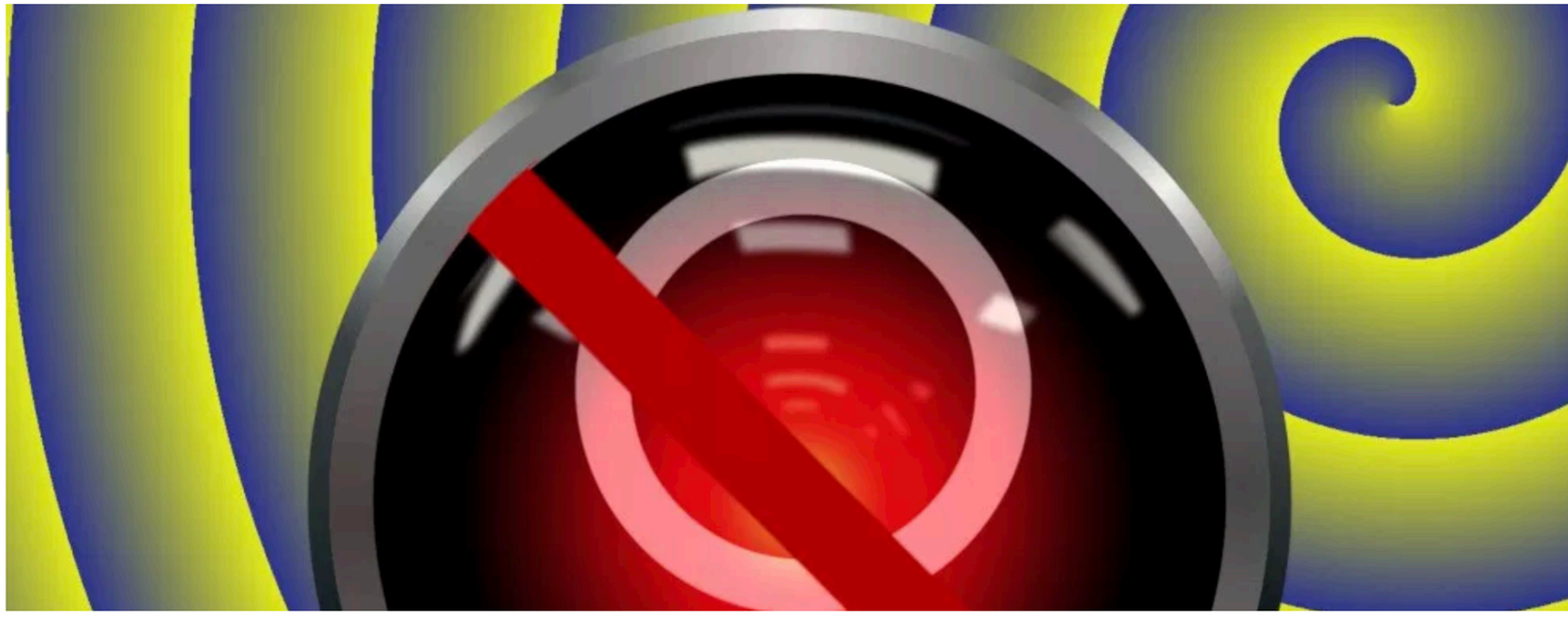
Amazon scraps secret AI showed bias against wom

Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's specialists uncovered a big problem: their

> The group created 500 computer models focused on specific job functions and locations. They taught each to recognize some 50,000 terms that showed up on past candidates' resumes. The algorithms learned to assign little significance to skills that were common across IT applicants, such as the ability to write various computer codes, the people said.
>
> Instead, the technology favored candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured," one person said.

Amazon trained a sexism-fighting, resume-screening AI with sexist hiring data, so the bot became sexist

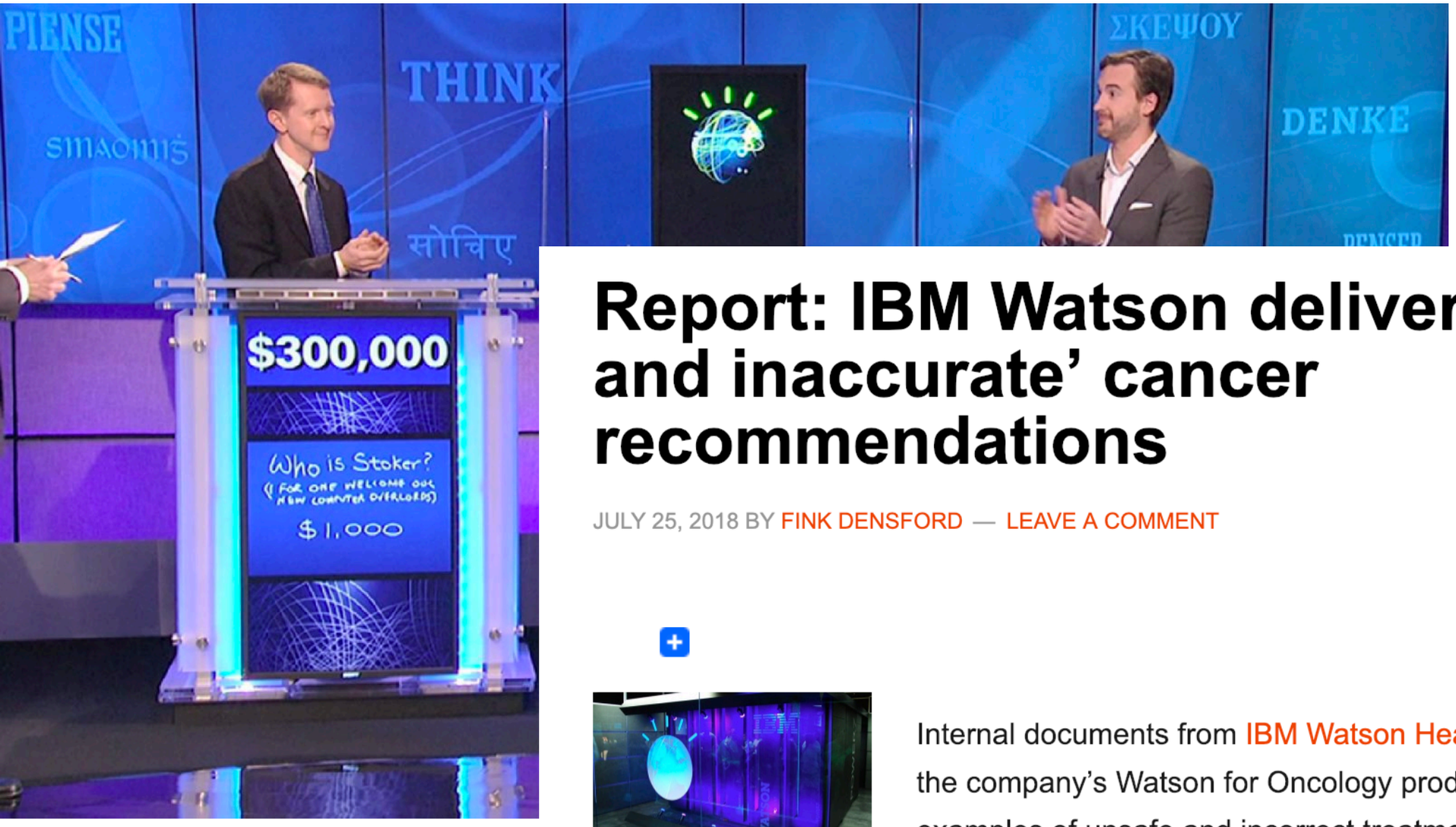


TECH AMAZON ARTIFICIAL INTELLIGENCE

21

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"



Report: IBM Watson delivered ‘unsafe and inaccurate’ cancer recommendations

JULY 25, 2018 BY FINK DENSFORD — LEAVE A COMMENT

Internal documents from IBM Watson Health (NYSE:IBM) indicate that the company’s Watson for Oncology product often returns “multiple examples of unsafe and incorrect treatment recommendations,” according to a new report from *STAT News*.

The documents come from slides presented last year by IBM Watson Health’s deputy chief health officer, according to the report, and include feedback from customers that indicated the product is “often inaccurate” and that its recommendations bring to light “serious questions about the process for building content and the underlying technology.”

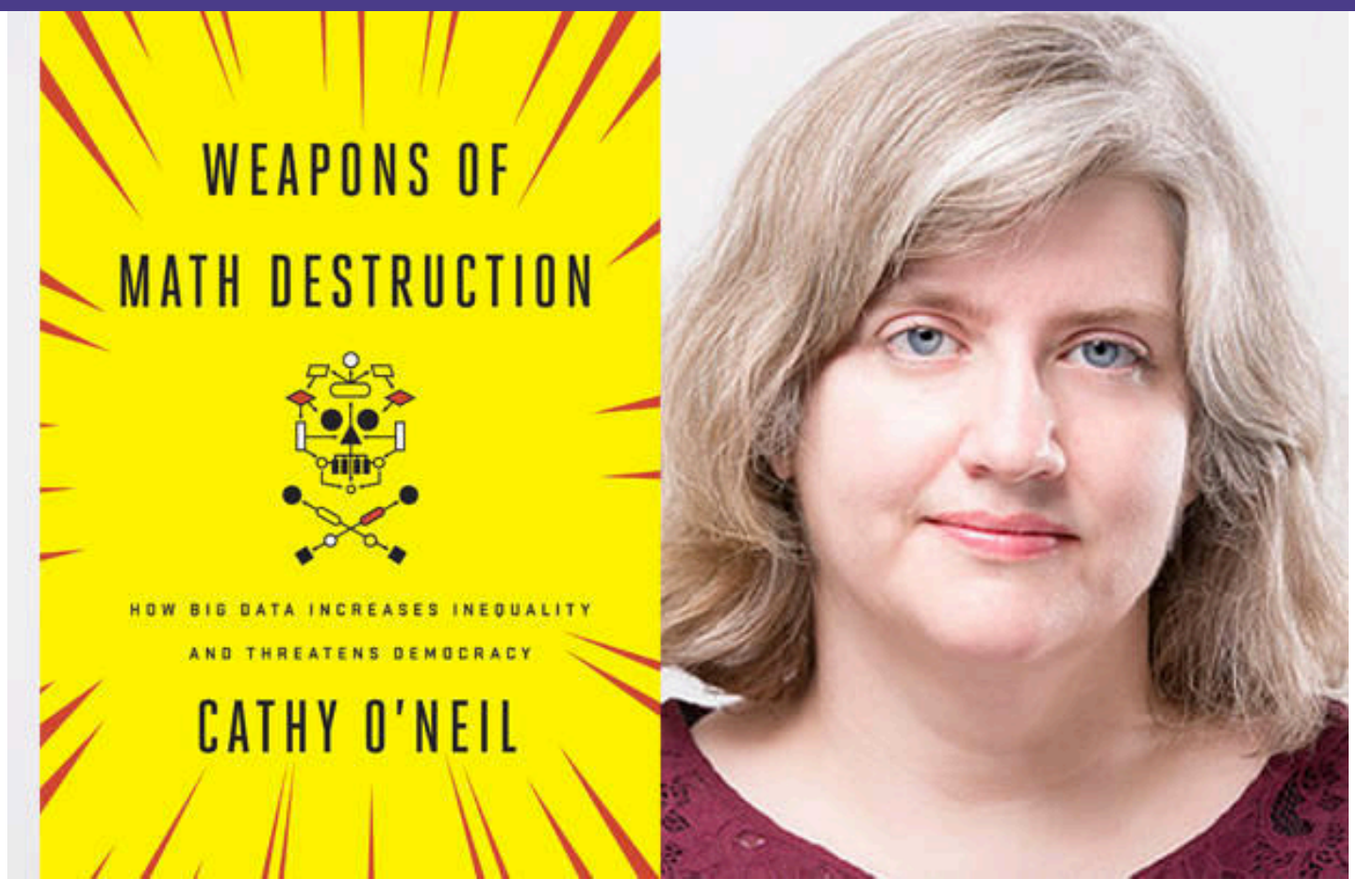
The issues were blamed on training the Watson product received by IBM engineers and physicians at the Memorial Sloan Kettering Cancer Center, which included “synthetic,” or hypothetical patients and cases, instead of real patient data, *STAT* reports.

https://www.massdevice.com/report-ibm-watson-delivered-unsafe-and-inaccurate-cancer-recommendations/

Why do we need explanations for complex models?

Cathy O'Neil:

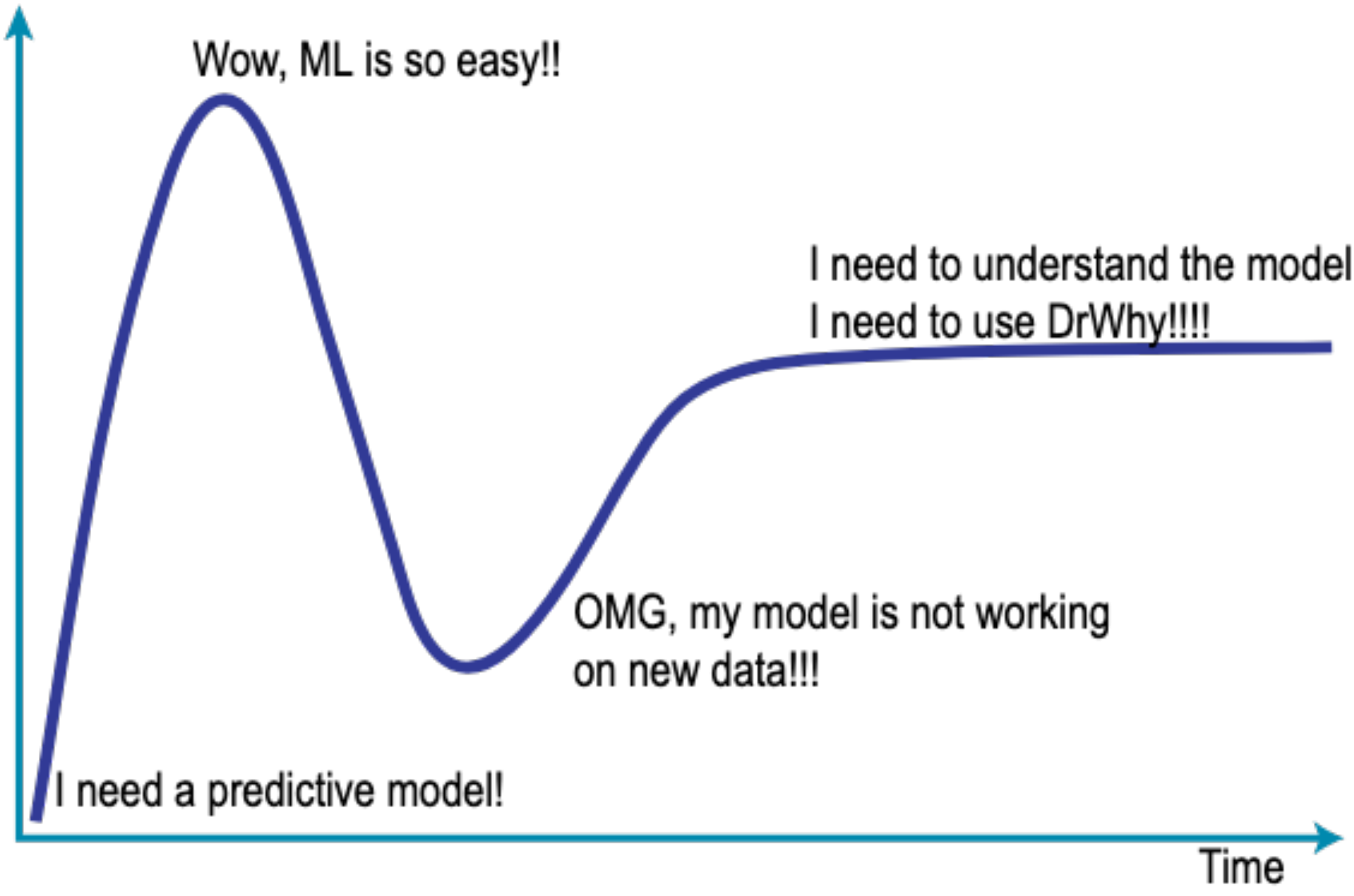
The era of blind faith in ~~big data~~ black boxes must end



- "You don't see a lot of skepticism," she says. "The algorithms are like shiny new toys that we can't resist using. We trust them so much that we project meaning on to them."
- Ultimately algorithms, according to O'Neil, reinforce discrimination and widen inequality, "using people's fear and trust of mathematics to prevent them from asking questions".

https://www.theguardian.com/books/2016/oct/27/cathy-oneil-weapons-of-math-destruction-algorithms-big-data

Hype Cycle for Predictive Models



http://drwhy.ai

Right to explanation

From Wikipedia, the free encyclopedia

In the regulation of algorithms, particularly artificial intelligence and its subfield of machine learning, a **right to expl explanation**) is a right to be given an explanation for an output of the algorithm. Such rights primarily refer to indivi explanation for decisions that significantly affect an individual, particularly legally or financially. For example, a pers is denied may ask for an explanation, which could be "Credit bureau X reports that you declared bankruptcy last ye considering you too likely to default, and thus we will not give you the loan you applied for."

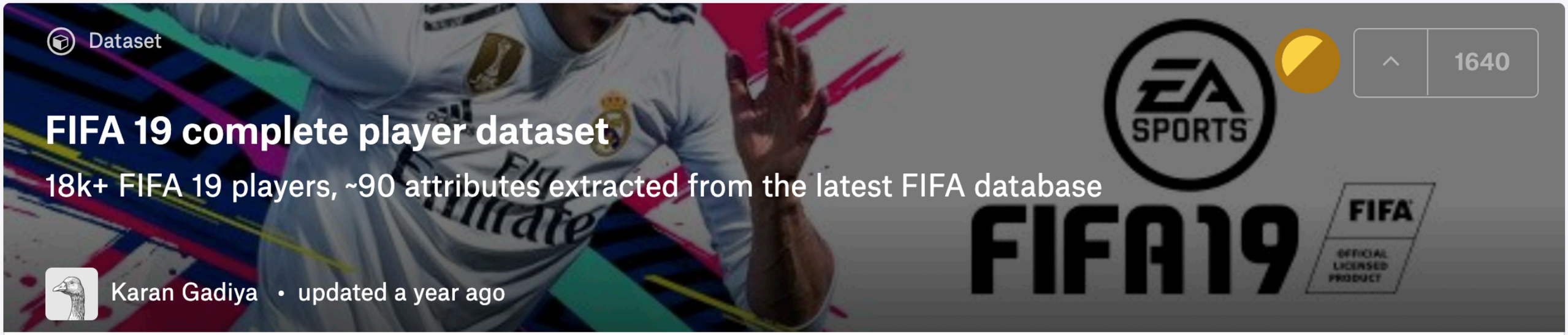
Some such legal rights already exist, while the scope of a general "right to explanation" is a matter of ongoing deba

Contents [hide]

How to explain a complex model?

Use case

FIFA 2019



Data | Kernels (304) | Discussion (21) | Activity | Metadata | Download (9 MB) | New Notebook

Your Dataset download has started.
Show your appreciation with an upvote

1640

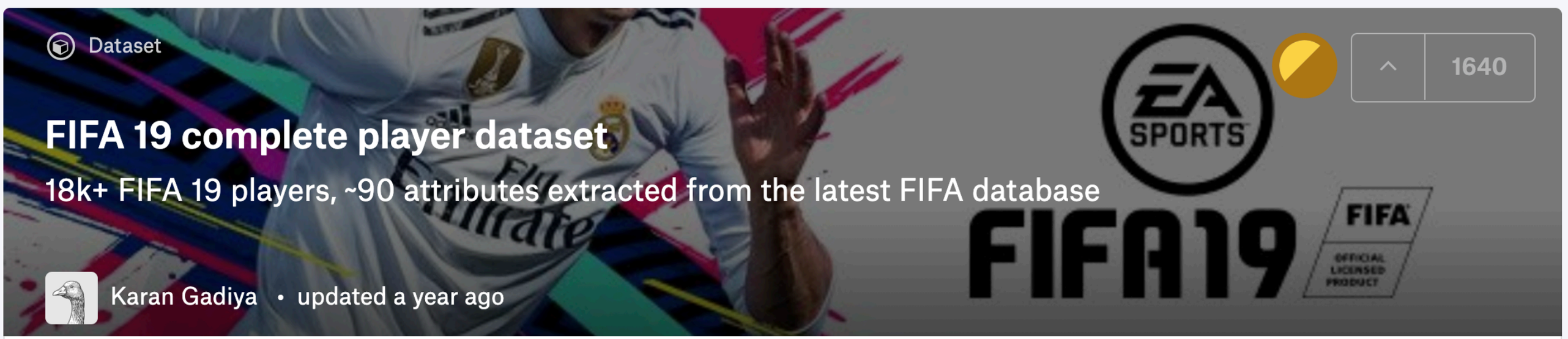
Usability 10.0 | **License** CC BY-NC-SA 4.0 | **Tags** data visualization, feature engineering, random forest, sports, regression analysis

Description

Context

Football analytics

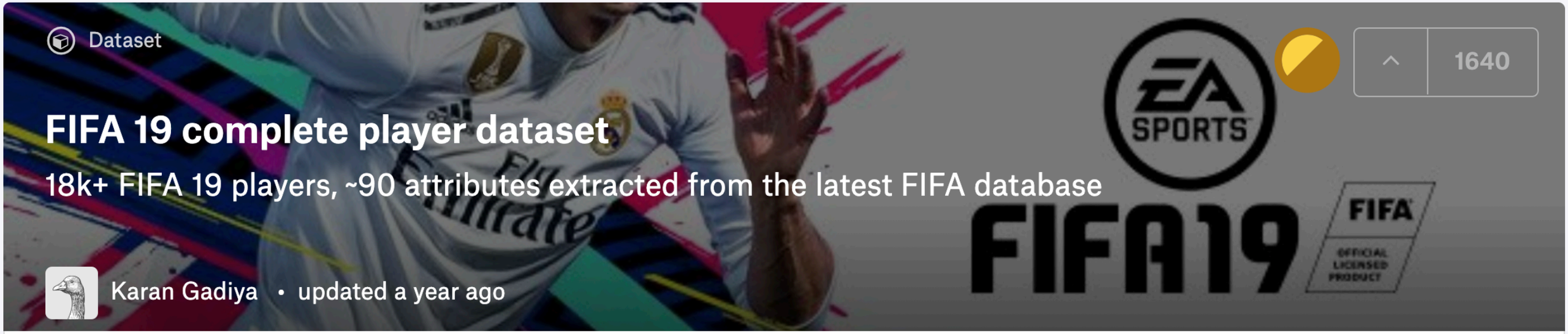
Source: https://www.kaggle.com/karangadiya/fifa19/data



Your Dataset download has started.
Show your appreciation with an upvote

1640

	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	Weak Foot
1	L. Messi	31	Argentina	94	94	FC Barcelona	€110.5M	€565K	Left	4
2	Cristiano Ronaldo	33	Portugal	94	94	Juventus	€77M	€405K	Right	4
3	Neymar Jr	26	Brazil	92	93	Paris Saint-Germain	€118.5M	€290K	Right	5
4	De Gea	27	Spain	91	93	Manchester United	€72M	€260K	Right	3
5	K. De Bruyne	27	Belgium	91	92	Manchester City	€102M	€355K	Right	5
6	E. Hazard	27	Belgium	91	91	Chelsea	€93M	€340K	Right	4
7	L. Modrić	32	Croatia	91	91	Real Madrid	€67M	€420K	Right	4
8	L. Suárez	31	Uruguay	91	91	FC Barcelona	€80M	€455K	Right	4
9	Sergio Ramos	32	Spain	91	91	Real Madrid	€51M	€380K	Right	3
10	J. Oblak	25	Slovenia	90	93	Atlético Madrid	€68M	€94K	Right	3
11	R. Lewandowski	29	Poland	90	90	FC Bayern München	€77M	€205K	Right	4
12	T. Kroos	28	Germany	90	90	Real Madrid	€76.5M	€355K	Right	5
13	D. Godín	32	Uruguay	90	90	Atlético Madrid	€44M	€125K	Right	3



Your Dataset download has started.
Show your appreciation with an upvote

1640

	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Preferred Foot	Weak Foot
1	L. Messi	31	Argentina	94	94	FC Barcelona	€110.5M	€565K	Left	4
2	Cristiano Ronaldo	33	Portugal	94	94	Juventus	€77M	€405K	Right	4
3	Neymar Jr	26	Brazil	92	93	Paris Saint-Germain	€118.5M	€290K	Right	5
4	De Gea	27	Spain	91	93	Manchester United	€72M	€260K	Right	3
5	K. De Bruyne	27	Belgium	91	92	Manchester City	€102M	€355K	Right	5
6	E. Hazard	27	Belgium	91	91	Chelsea				
7	L. Modrić	32	Croatia	91	91	Real Mad				
8	L. Suárez	31	Uruguay	91	91	FC Barcel				
9	Sergio Ramos	32	Spain	91	91	Real Mad				
10	J. Oblak	25	Slovenia	90	93	Atlético M				
11	R. Lewandowski	29	Poland	90	90	FC Bayern				
12	T. Kroos	28	Germany	90	90	Real Mad				
13	D. Godín	32	Uruguay	90	90	Atlético M				

```
library("gbm")
fifa_gbm <- gbm(Value~.,
        data              = fifa19,
        n.trees           = 250,
        interaction.depth = 3)
```

wikipedia

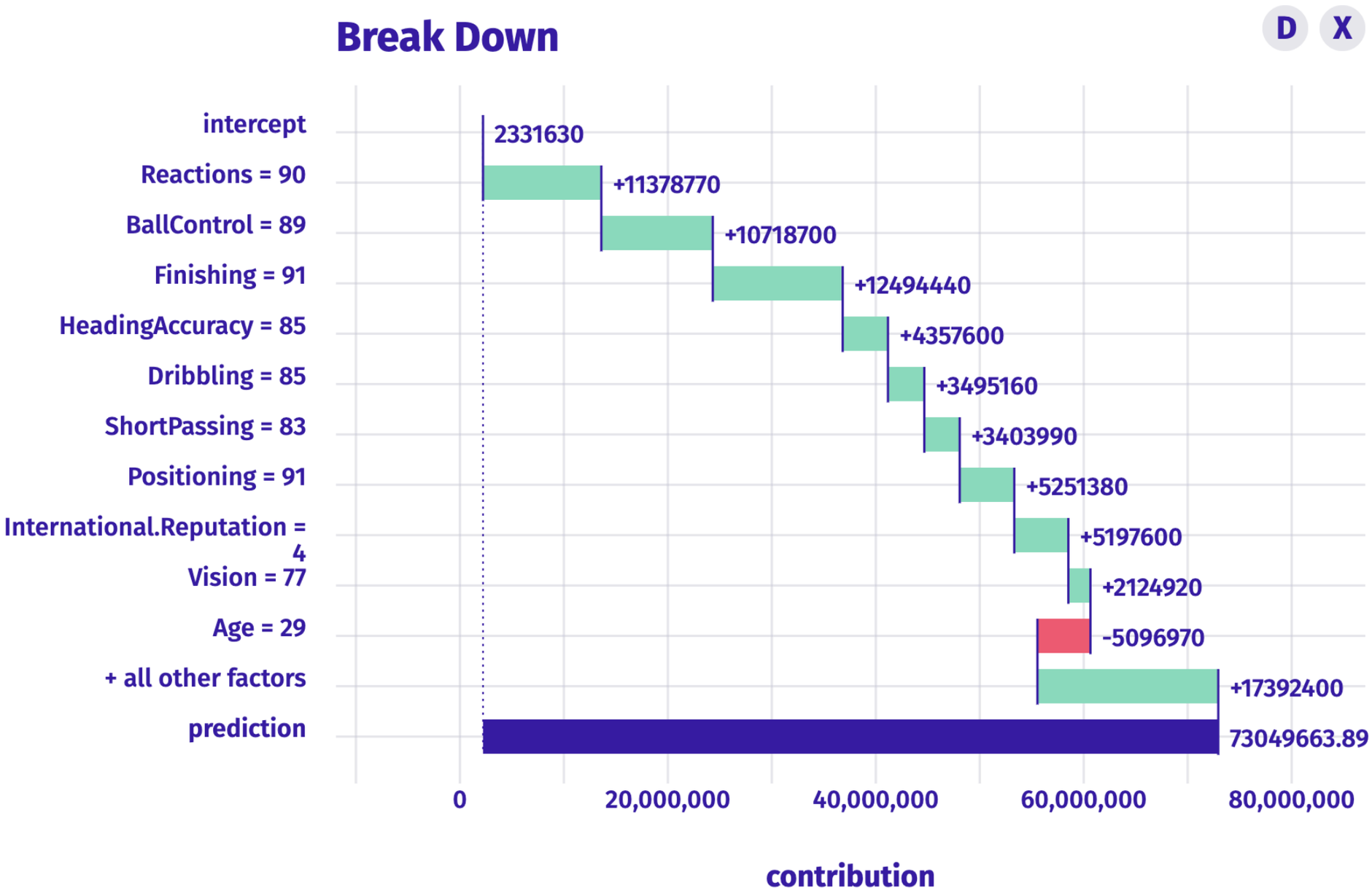
Full name	Robert Lewandowski
Date of birth	21 August 1988
Place of birth	Warsaw, Poland
Playing position	Striker

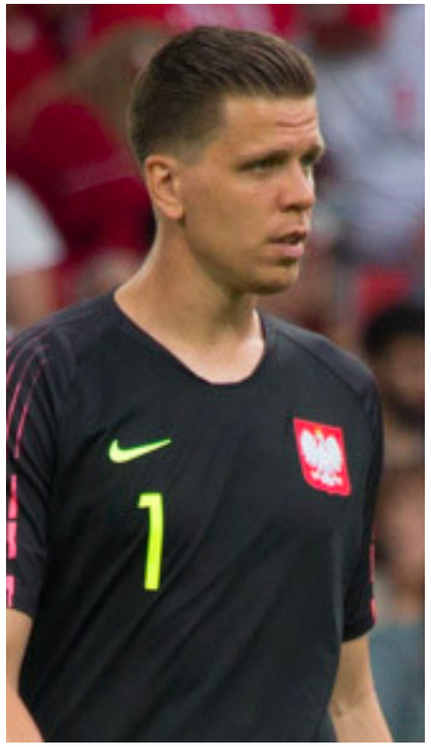
Prediction from GBM model: 73 049 663 EUR

How? Why? Really?

wikipedia

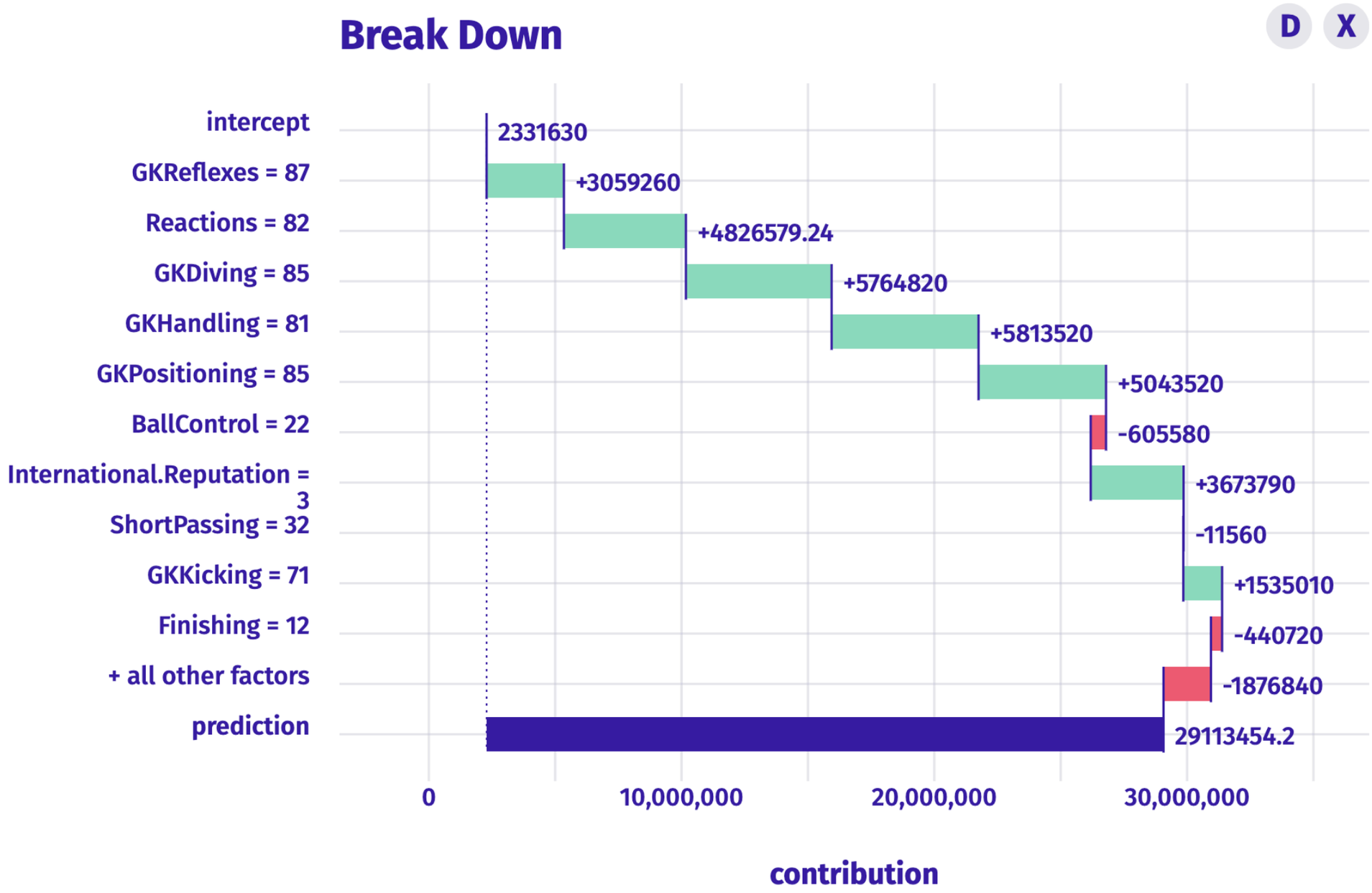
Full name	Robert Lewandowski
Date of birth	21 August 1988
Place of birth	Warsaw, Poland
Playing position	Striker





wikipedia

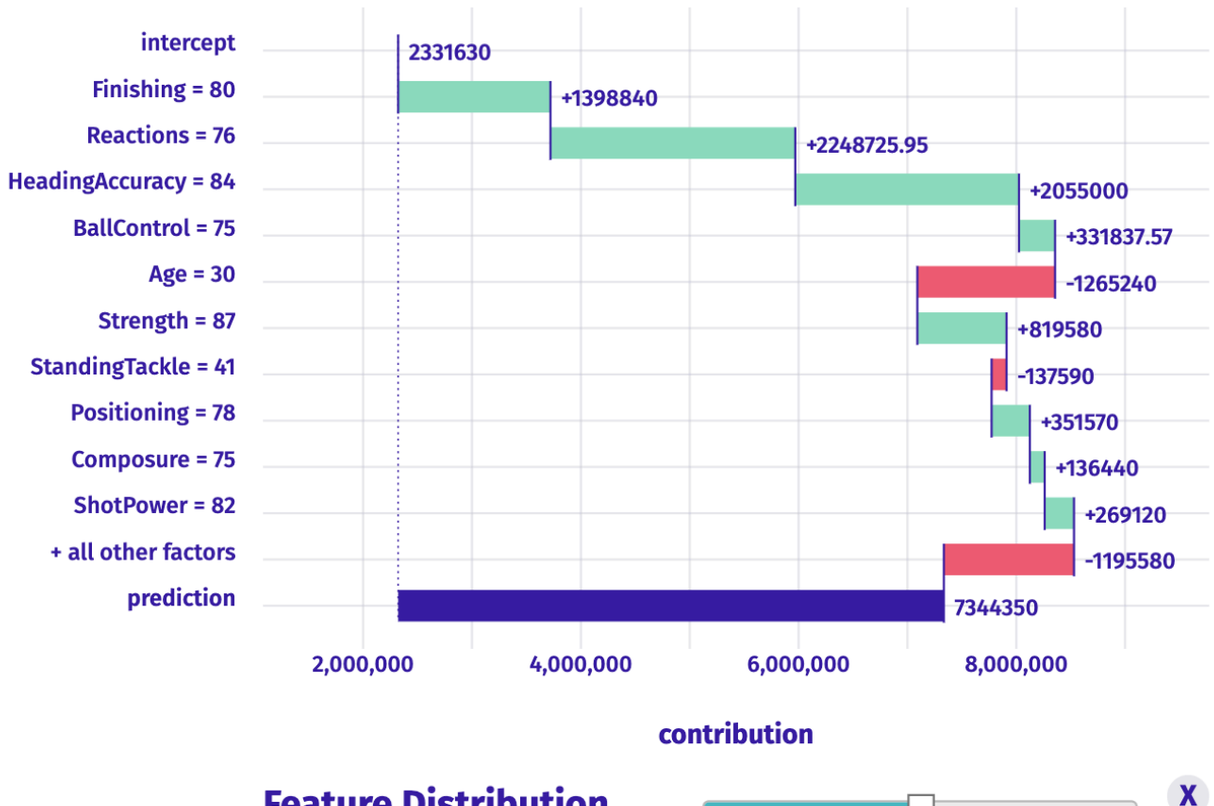
Full name	Wojciech Szczęsny
Date of birth	18 April 1990
Place of birth	Warsaw, Poland
Playing position	Goalkeeper



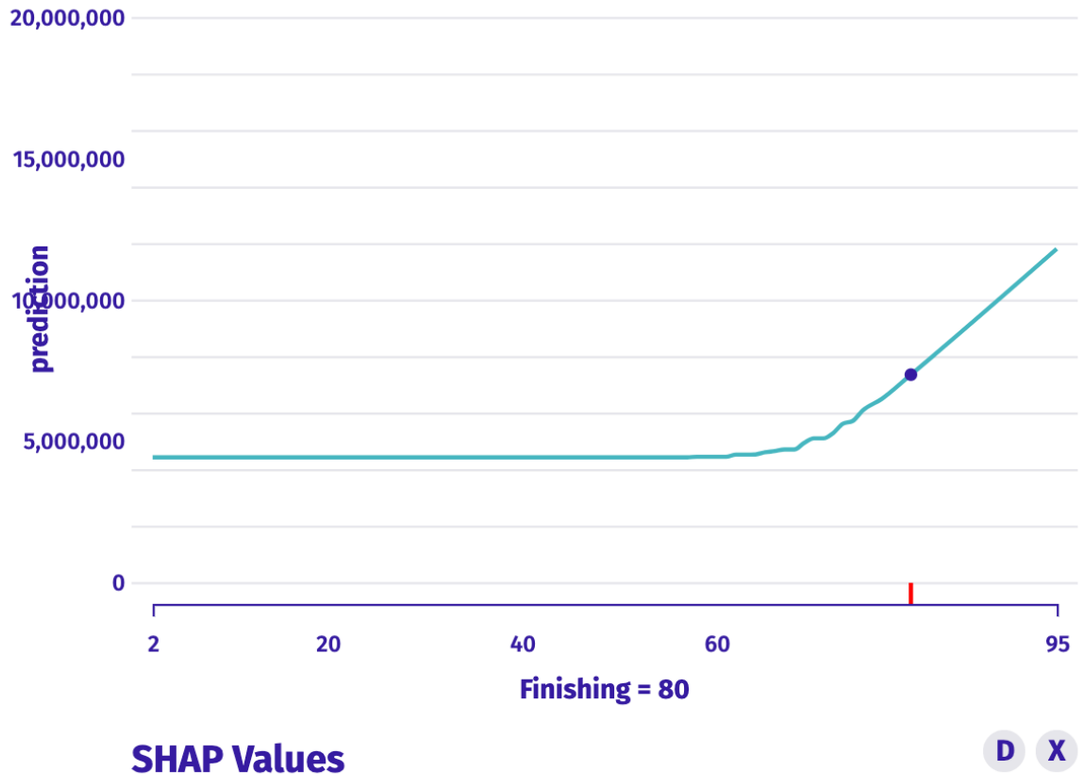
Interactive Model Studio for FIFA 2019 GBM model

Ádám Szalai

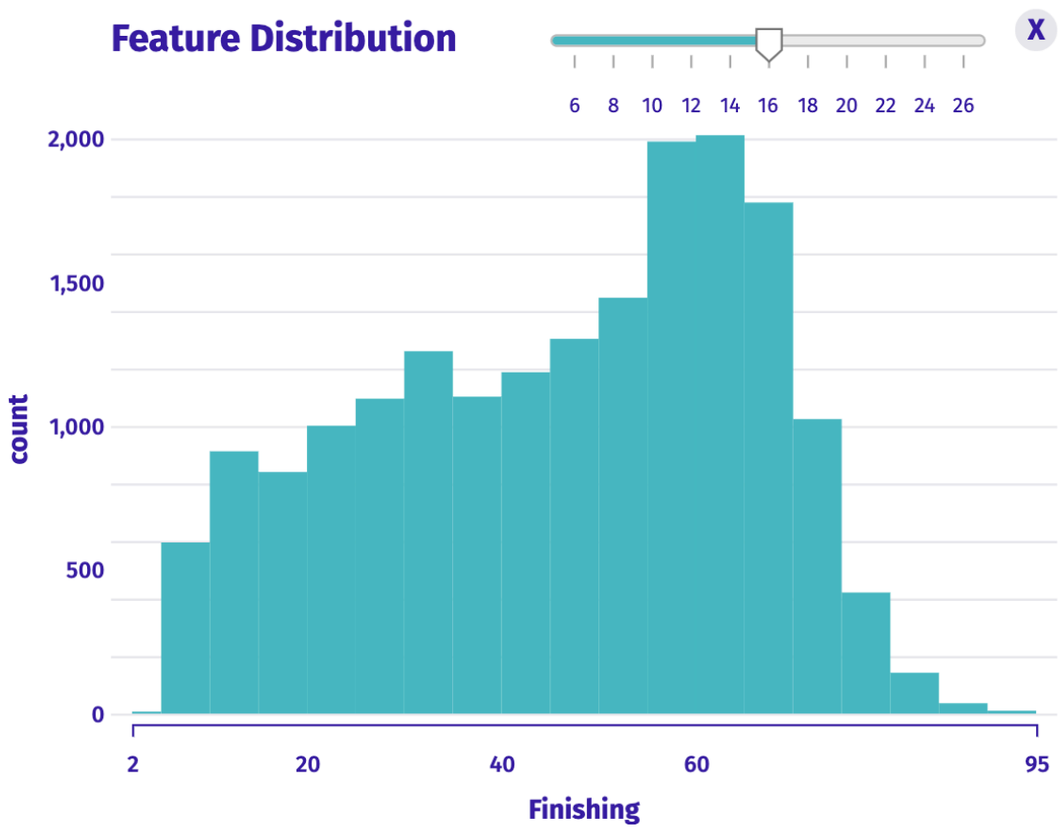
Break Down



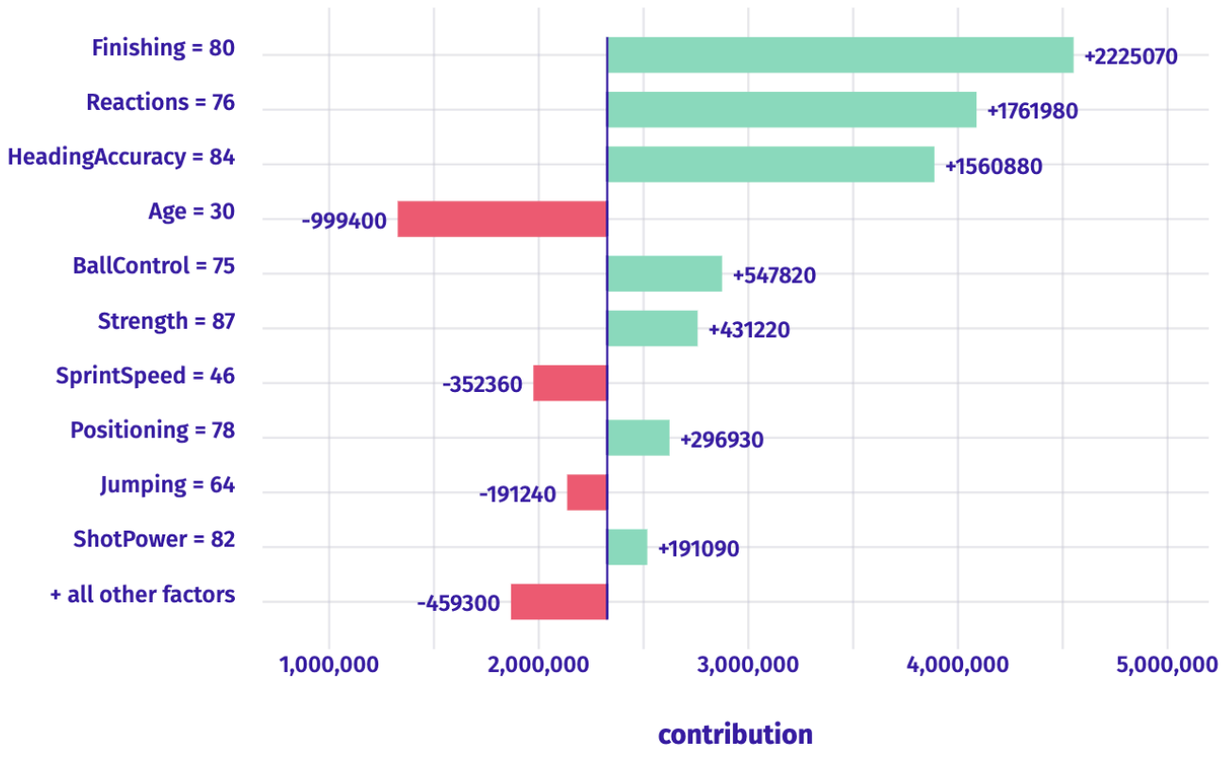
Ceteris Paribus



Feature Distribution



SHAP Values



https://pbiecek.github.io/explainFIFA19/

Use case

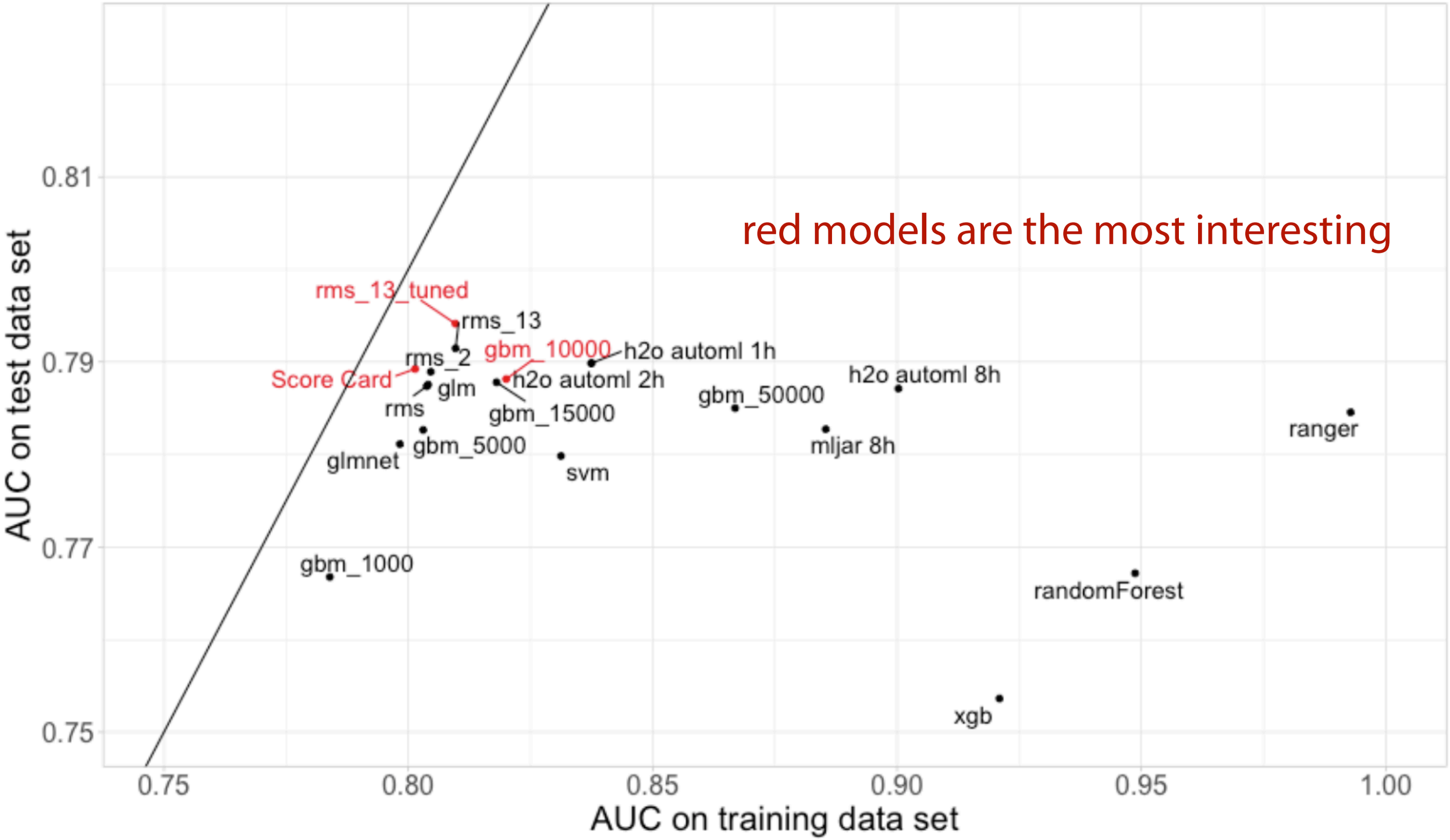
Credit Scoring

Data set for study: xML Challenge by FICO

- Explainable Machine Learning Challenge by FICO (2019)
- Focus: Home Equity Line of Credit (HELOC) Dataset
- Customers requested a credit line in the range of $5,000 - $150,000
- Task is to predict whether they will repay their HELOC account within 2 years
- Number of observations: 2,615
- Variables: 23 covariates (mostly numeric) and 1 target variable (risk performance "good" or "bad")

https://buecker.ms/slides/crccxvi/slides.html#1

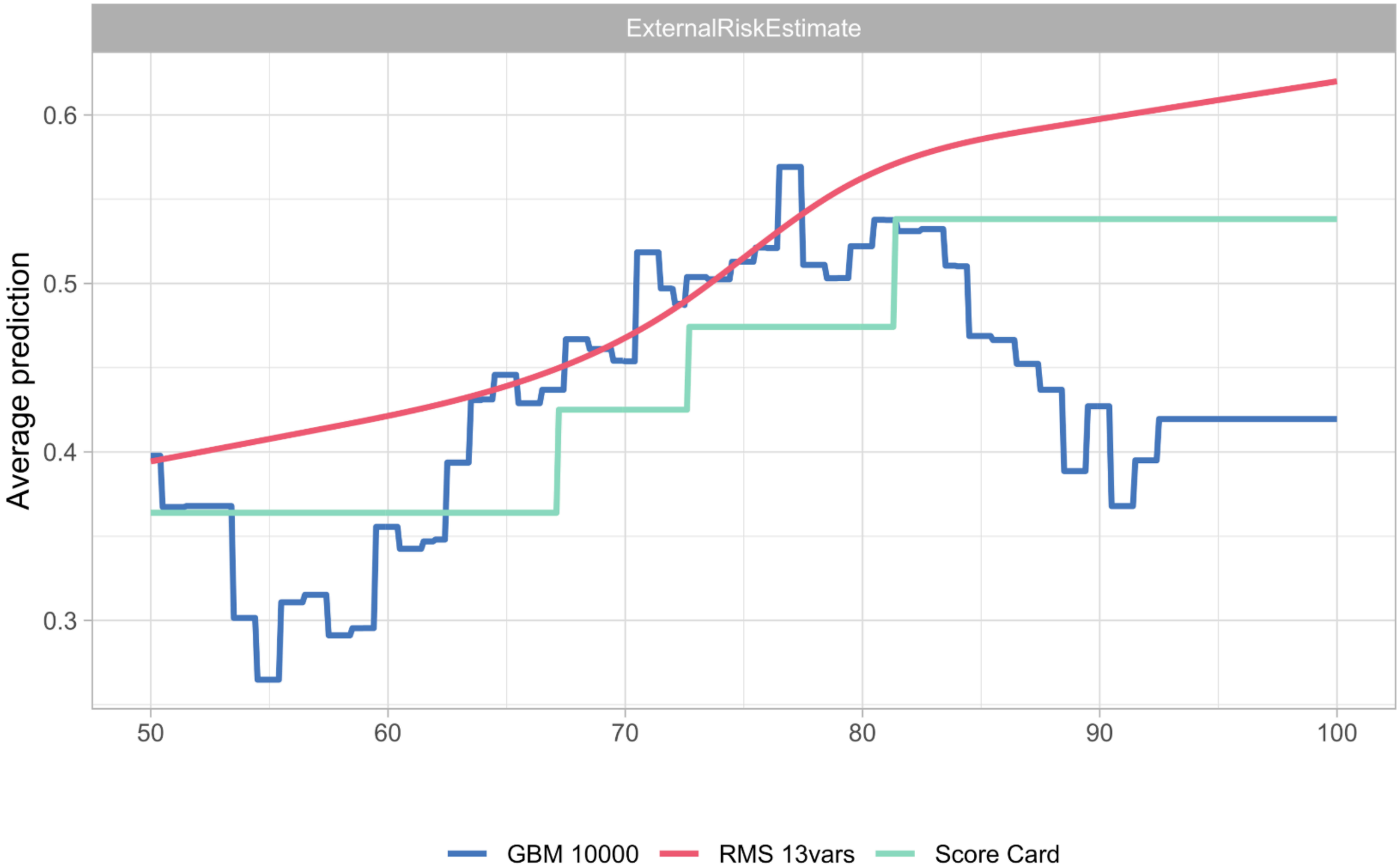
Performance for selected modeling methods



https://buecker.ms/slides/crccxvi/slides.html#1

Champion vs. Challenger

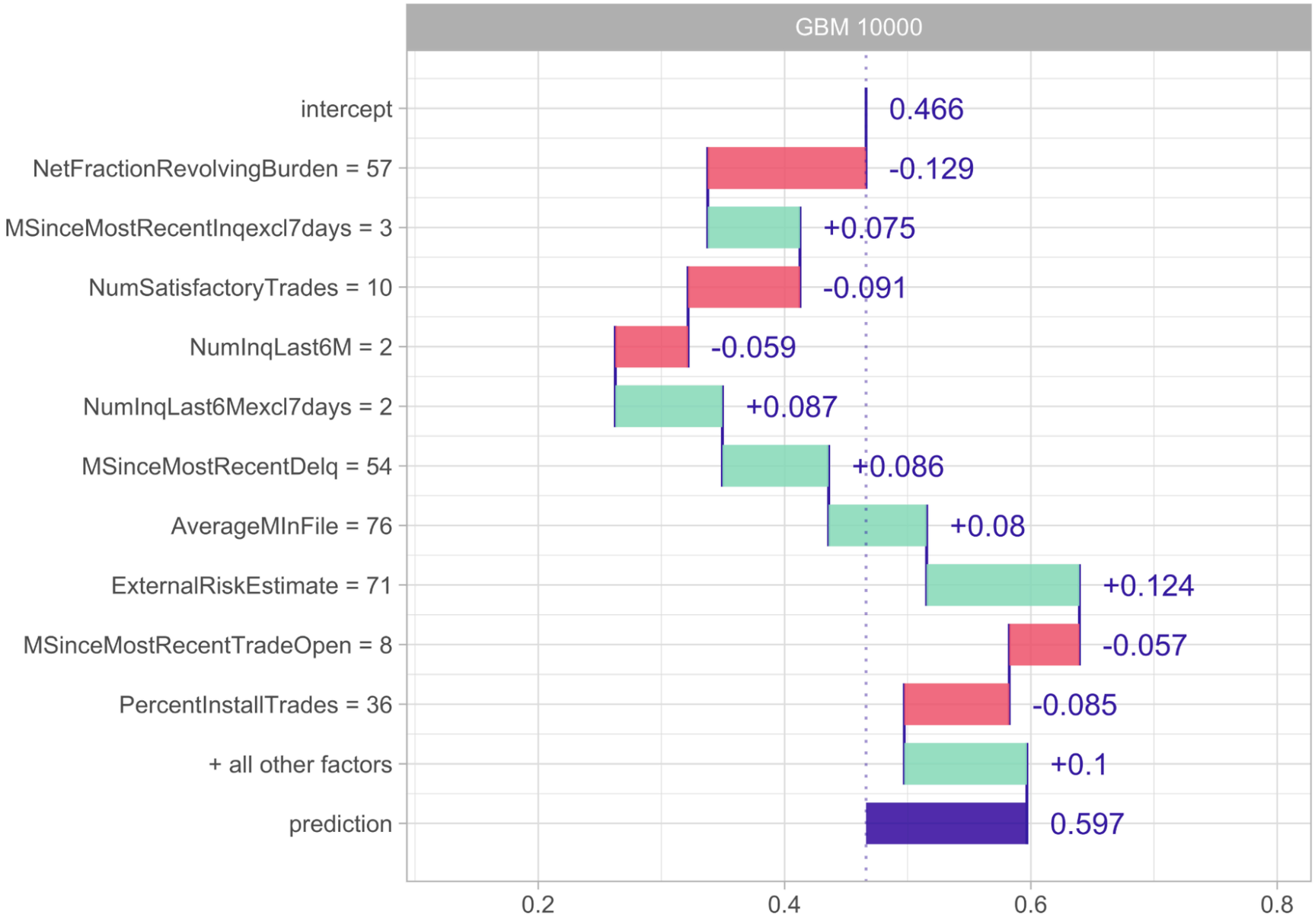
Partial Dependency Plot for the most important feature



https://buecker.ms/slides/crccxvi/slides.html#1

Single customer

Model agnostic: Variable contribution break down



- Such instance-lev explorations can be performed in a model-agnostic w
- Unfortunately, fo non-additive mod variable contributions depend on the ordering of varia

https://buecker.ms/slides/crccxvi/slides.html#1

FH MÜNSTER
University of Applied Sciences

MSB FB Wirtschaft
Münster School of Business

Transparency of Machine Learning Models in Credit Scoring

CRC Conference XVI

Michael Bücker, Gero Szepannek, Przemyslaw Biecek,
Alicja Gosiewska and Mateusz Staniak

28 August 2019

https://buecker.ms/slides/crccxvi/slides.html#1

The most common questions for XAI

XAI-BOT:

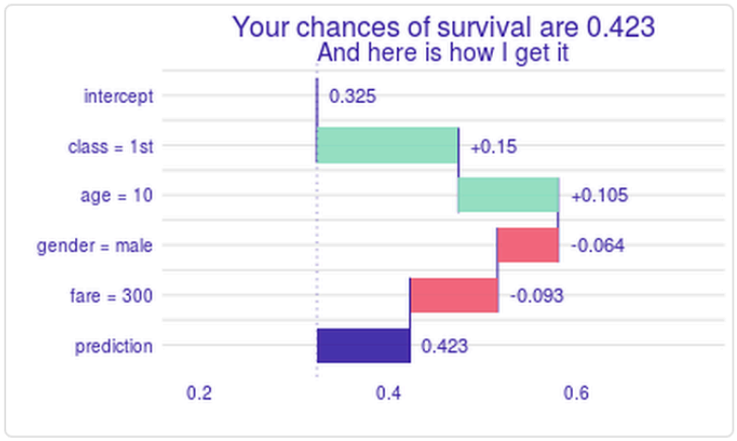
What would you ask for?

Dialogflow Bot APP 12:38 PM
Your chance of survival equals 0.423. It's close to a toss of a coin!

Przemyslaw Biecek 12:38 PM
Why ?

Dialogflow Bot APP 12:38 PM
Creating a plot. It may take a few seconds...

Break down plot
This chart illustrates the contribution of variables to the final prediction (16 kB)



Przemyslaw Biecek 12:39 PM
what if I paid more?

Dialogflow Bot APP 12:39 PM
Creating a plot. It may take a few seconds...

Ceteris Paribus plot
This plot illustrates how the prediction changes when fare is changed and everything else is fixed (17 kB)

https://kmichael08.github.io

XAI-BOT:

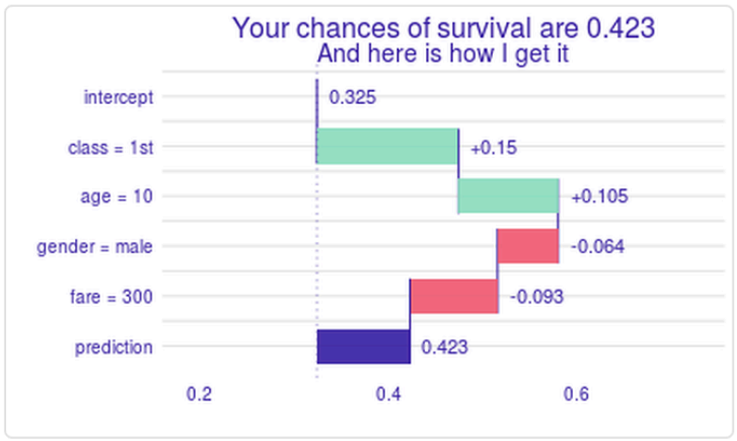
What would you ask for?

Dialogflow Bot APP 12:38 PM
Your chance of survival equals 0.423. It's close to a toss of a coin!

Przemyslaw Biecek 12:38 PM
Why ?

Dialogflow Bot APP 12:38 PM
Creating a plot. It may take a few seconds...

Break down plot
This chart illustrates the contribution of variables to the final prediction (16 kB)



Przemyslaw Biecek 12:39 PM
what if I paid more?

Dialogflow Bot APP 12:39 PM
Creating a plot. It may take a few seconds...

Ceteris Paribus plot
This plot illustrates how the prediction changes when fare is changed and everything else is fixed (17 kB)

Why?

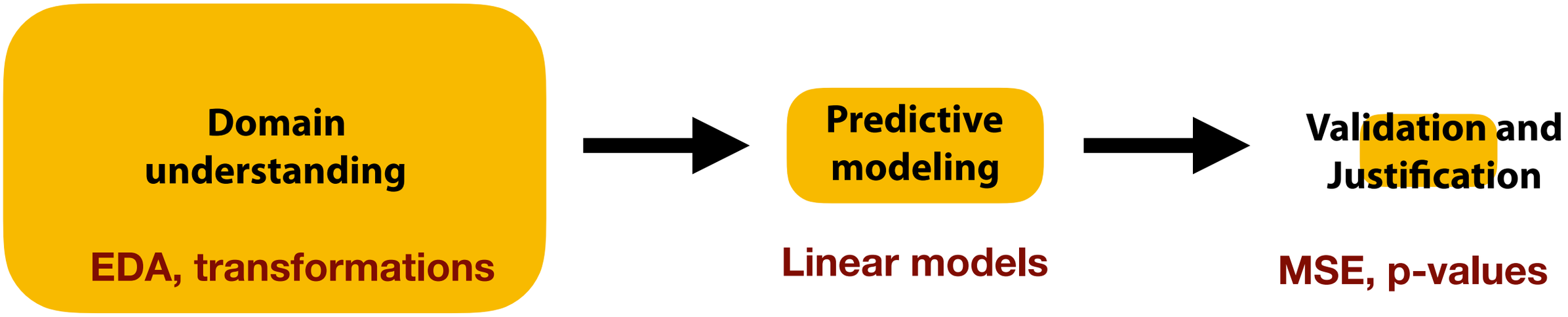
What If?

How good is the prediction?

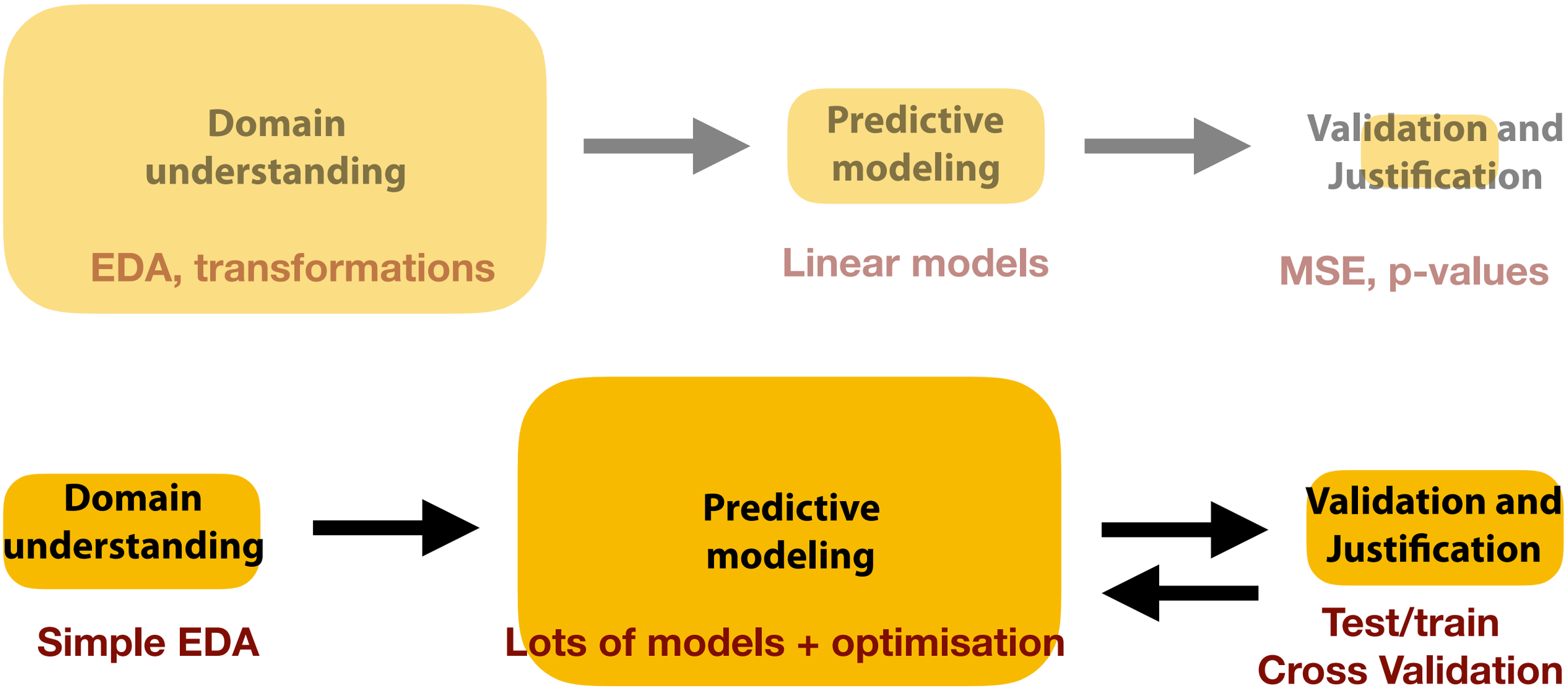
https://kmichael08.github.io

When to use XAI?

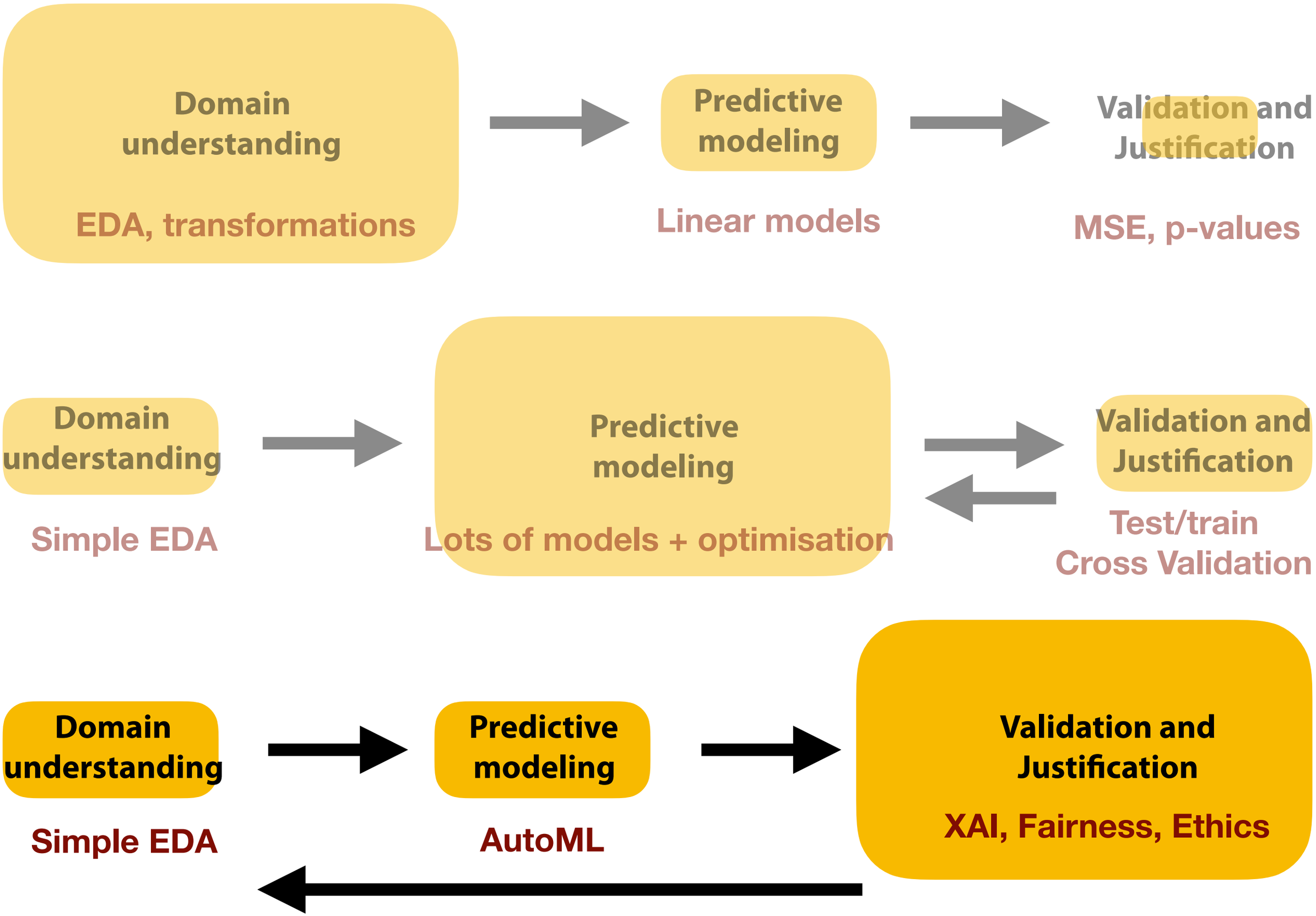
Shift in our focus: Statistics



Shift in our focus: Machine Learning

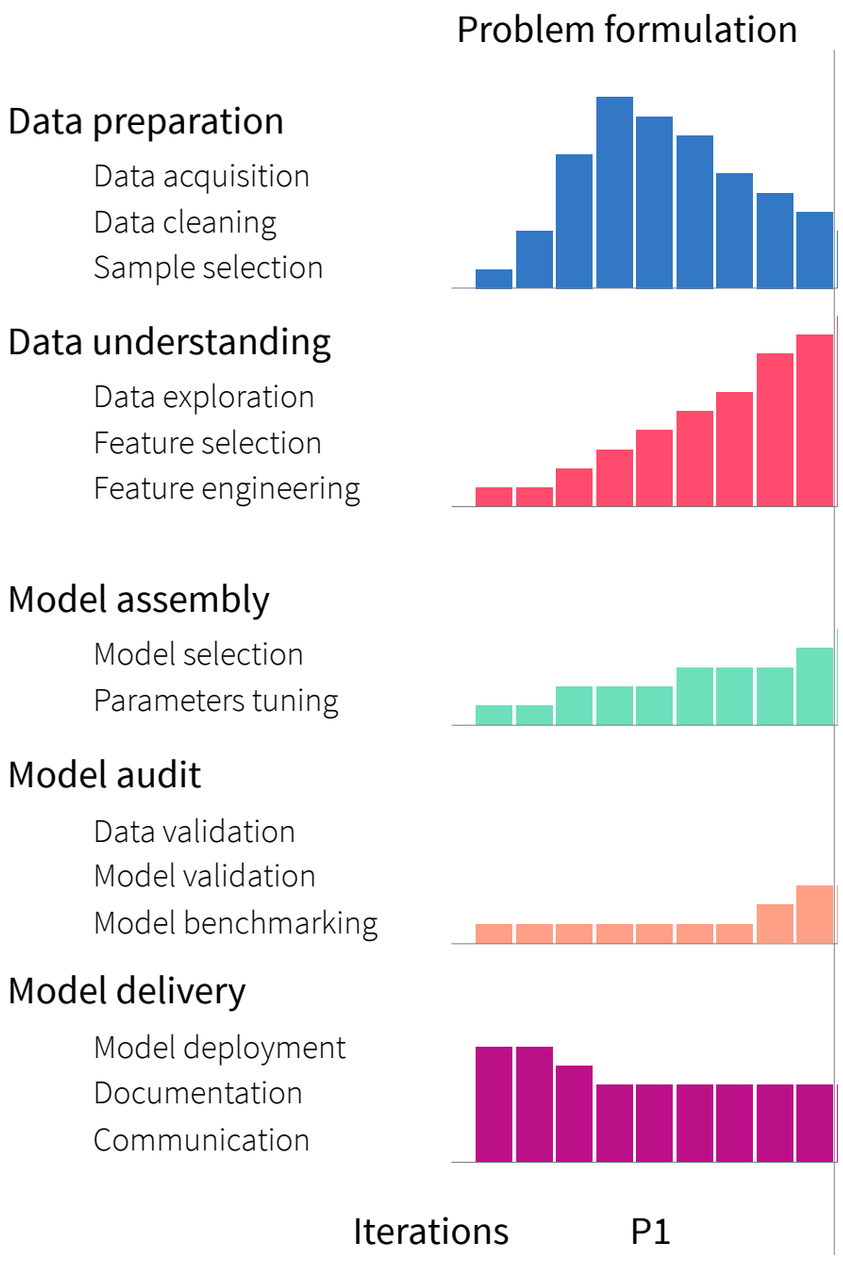


Shift in our focus: Human Oriented ML?



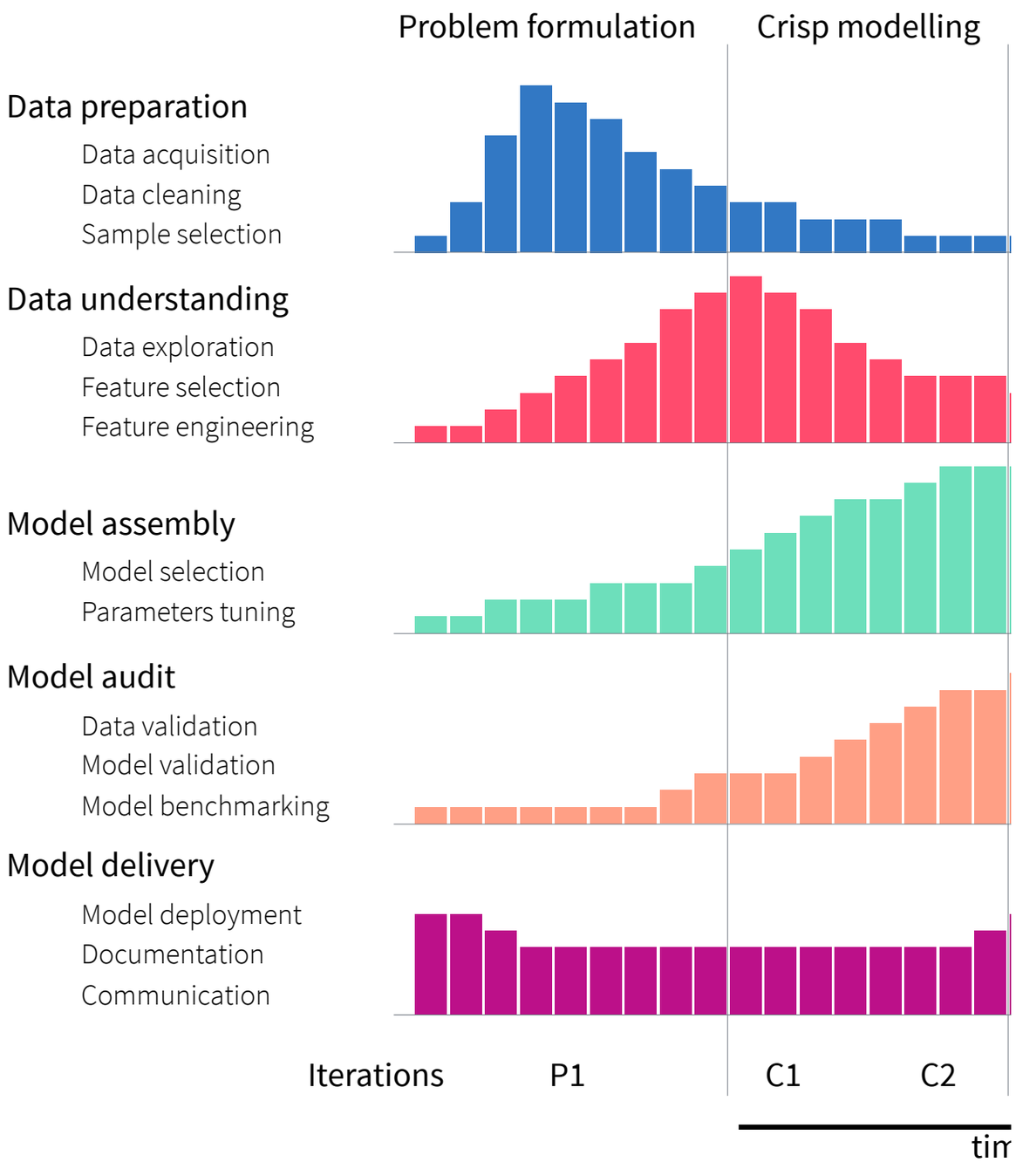
MDP - process for model development

MDP : : Model Development Process



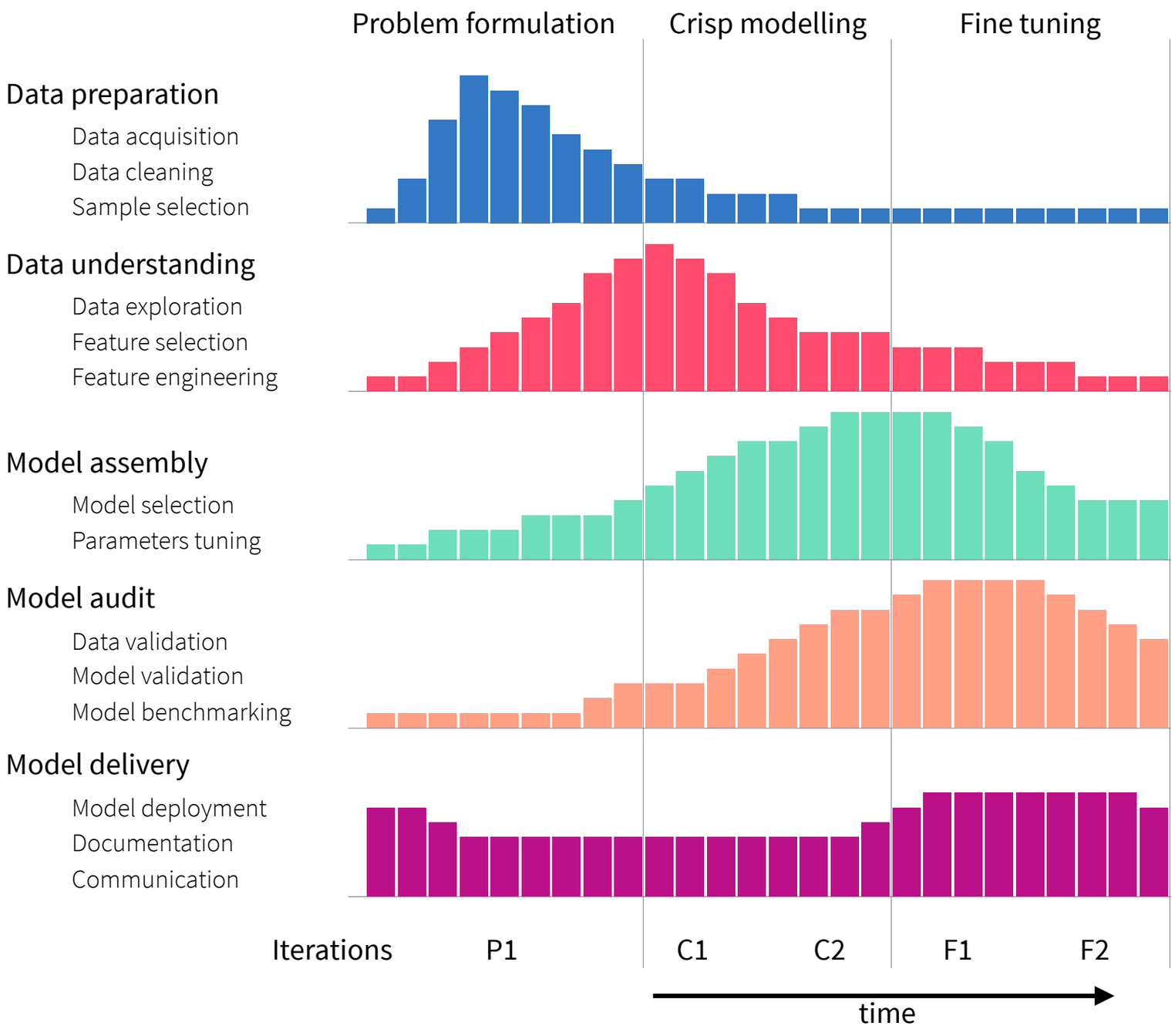
MDP - process for model development

MDP : : Model Development Process



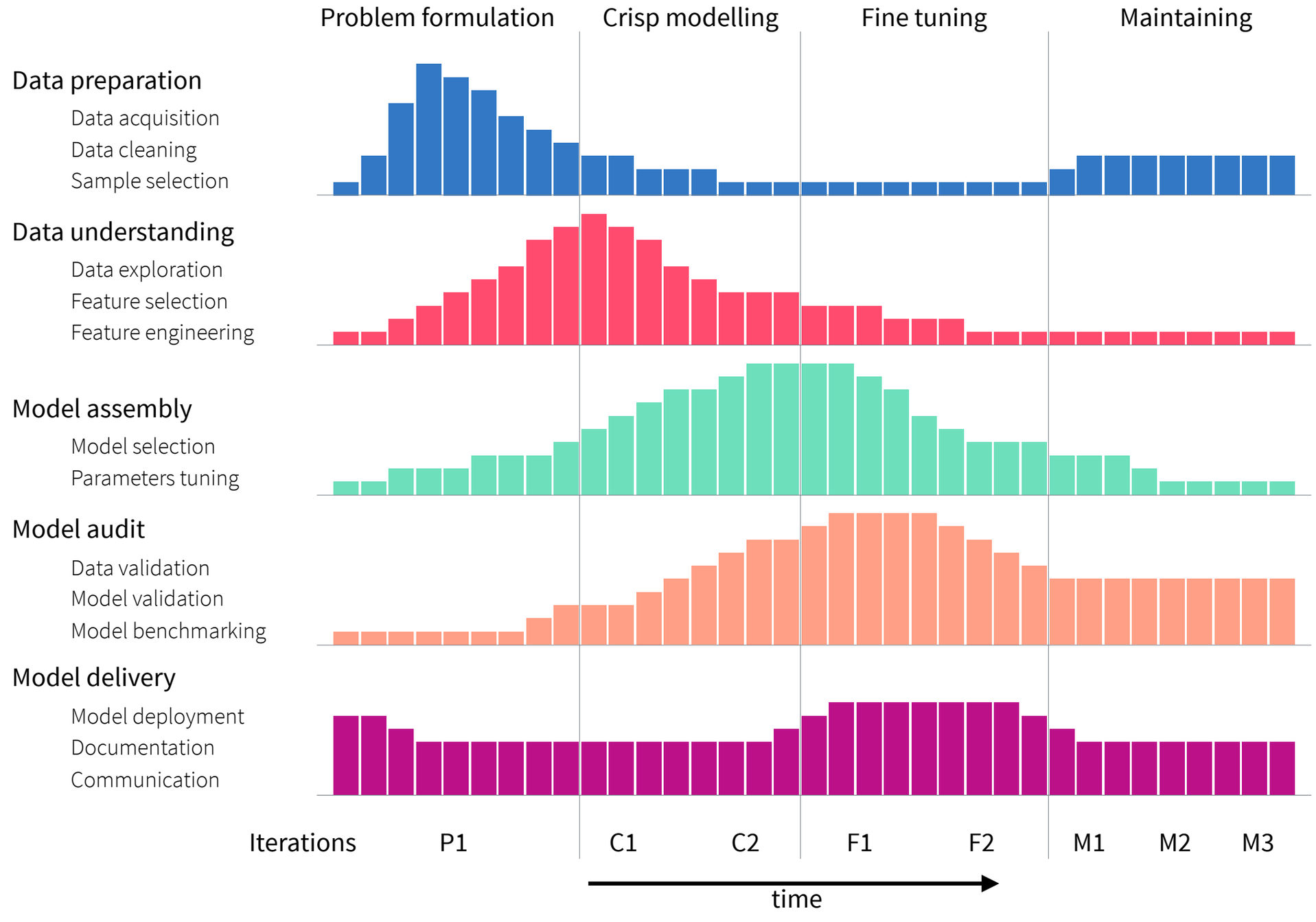
MDP - process for model development

MDP :: Model Development Process

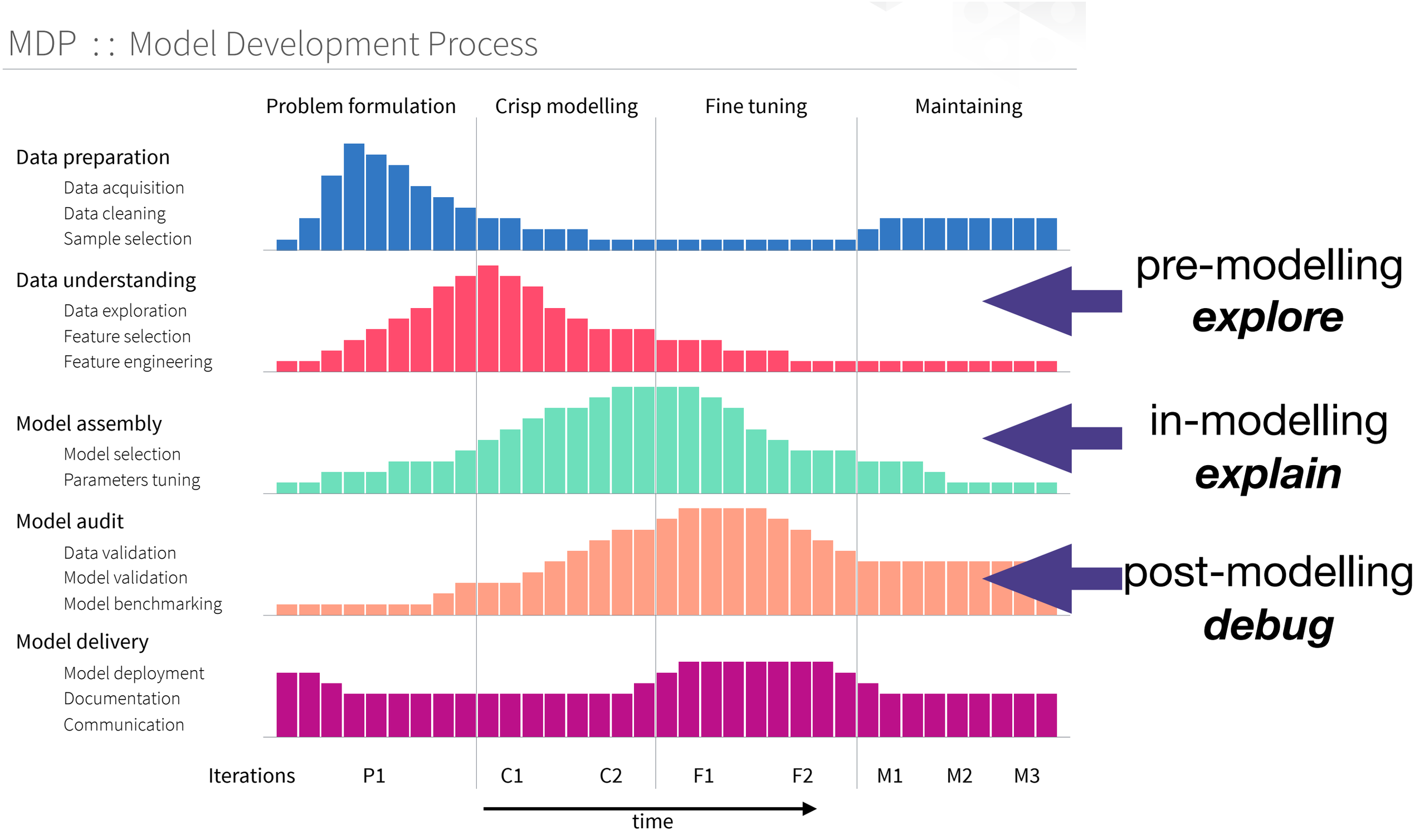


MDP - process for model development

MDP :: Model Development Process



pre-, in-, post- modelling explainability



Model specific / agnostic explainability

Model specific

+ Exploits the structure of a model
+ Allow for deep diagnostic

- Explanations cannot be easily compared between models
- You need to create an explainer for every possible model structure

Examples:

- randomForestExplainer
- EIX - lightgbm/xgboost explainer

Model agnostic

+ Explanations can be compared between different models
+ Can be used to any model (model is a black box)

- Explanations are (im most cases) approximations, they may be inaccurate or wrong
- It is easy to miss some quirks

Examples:
LIME / SHAP / Break Down / Partial Dependency Profiles / Permutational Feature Importance

Local / global explainability

Local explanations

+ Focused on a single observation
+ Good for dissecting of a single prediction
+ Good for debugging

- Feature effects may be different for different observations
- Tricky for correlated features

Examples:

- LIME / SHAP / Break Down
- Ceteris Paribus

Global explanations

+ Focused on a model
+ Good for global model summaries
+ Good for comparisons

- Local behaviour may be different
- For non-additive models they may be to simplistic

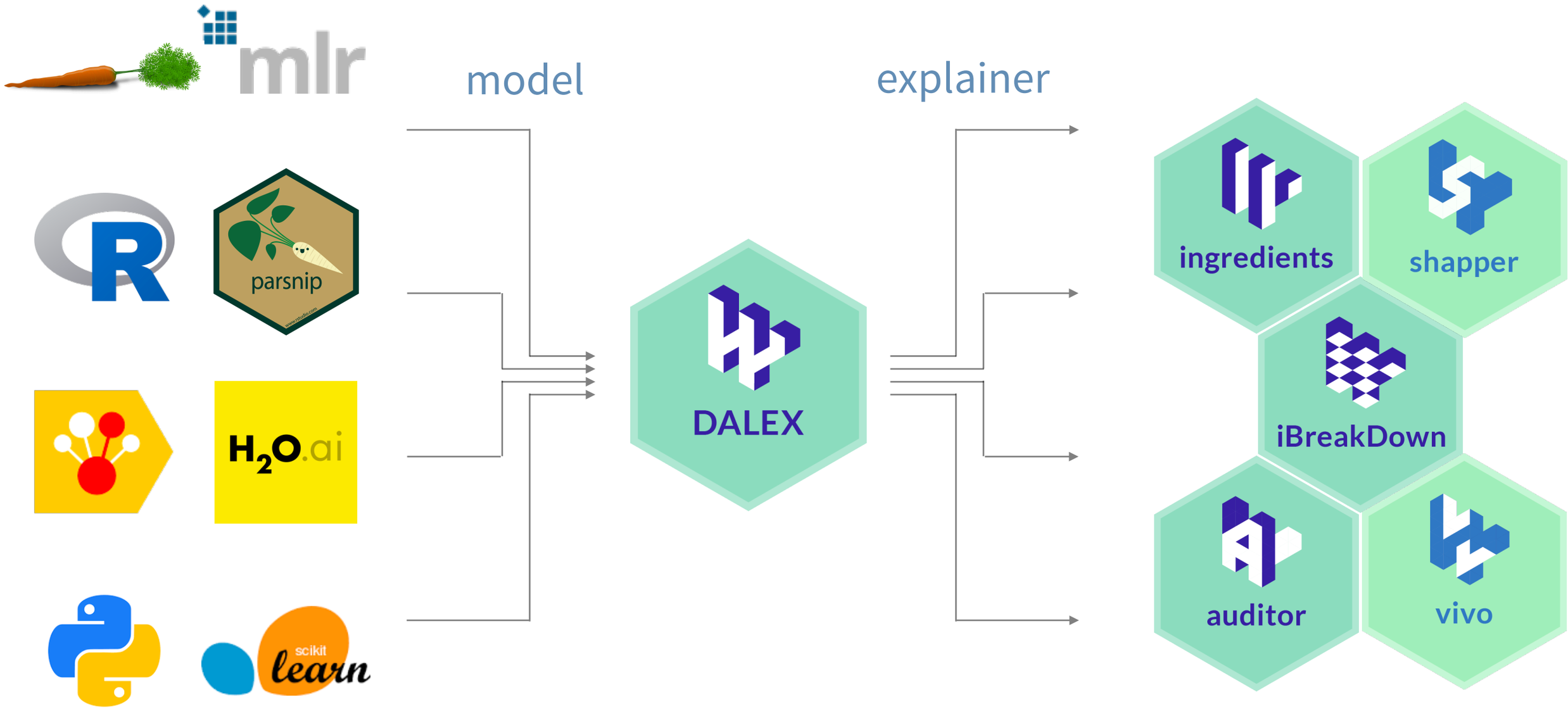
Examples:

- Feature Importance
- Partial Dependency Profiles

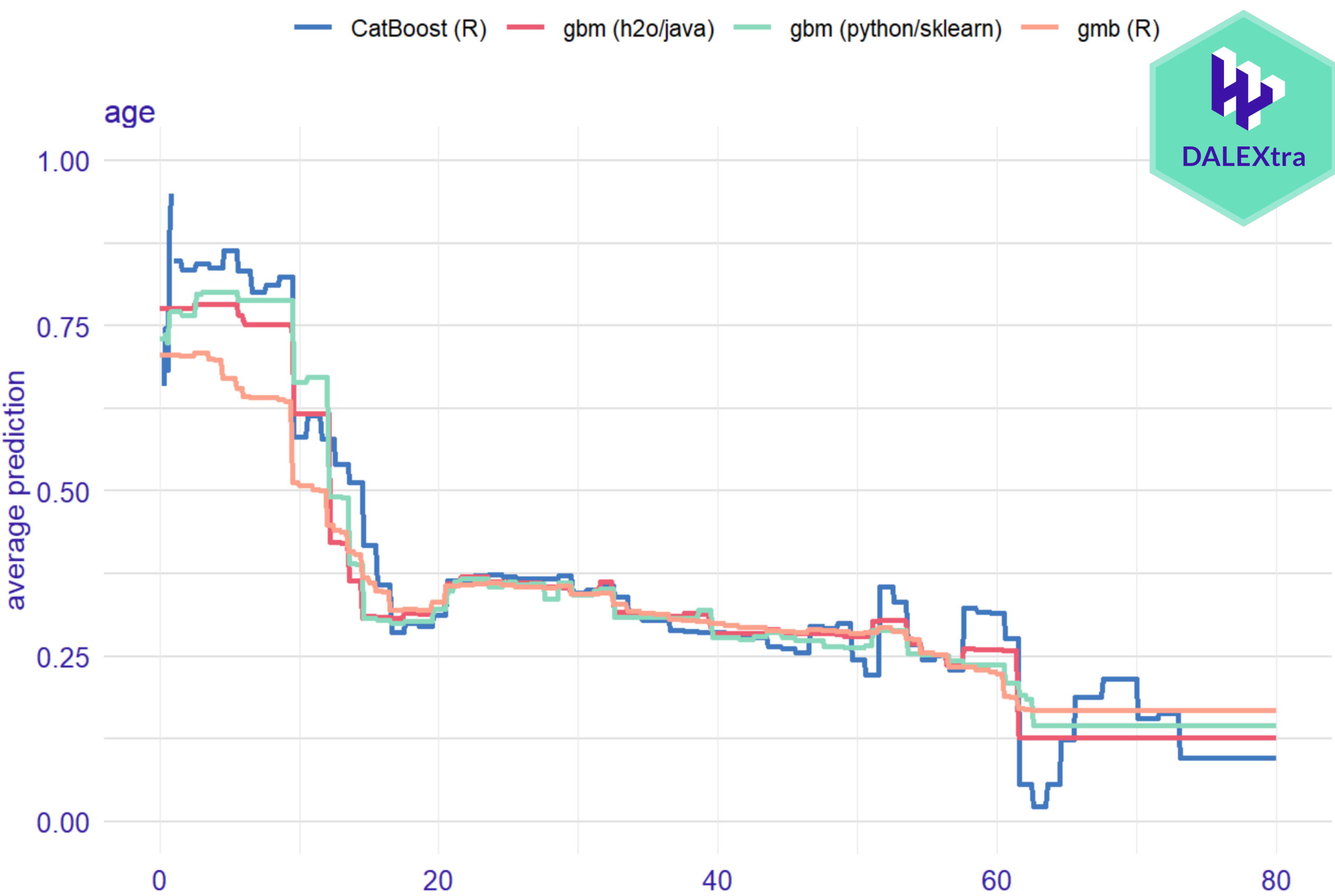
Make models not wars

DALEX wraps a model available in R or python or through REST API into a standardised container.

Other tools for XAI may then work on the model despite its internal structure.



https://medium.com/@ModelOriented/xai-in-the-jungle-of-competing-frameworks-for-machine-learning-fa6e96a99644



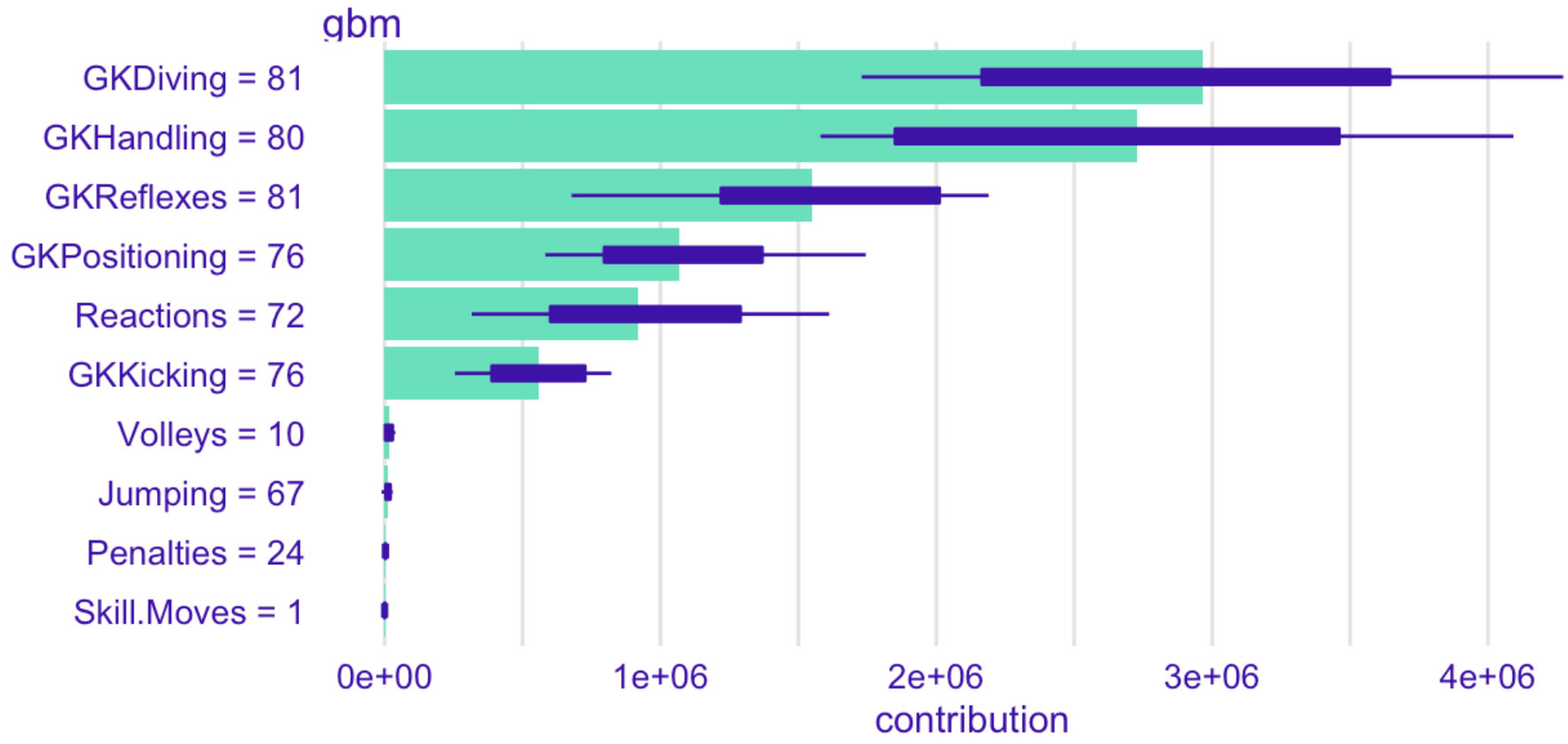
https://medium.com/@ModelOriented/xai-in-the-jungle-of-competing-frameworks-for-machine-learning-fa6e96a99644

Why?

SHAP （local）

Average Break Down profile / SHapley Additive exPlanations

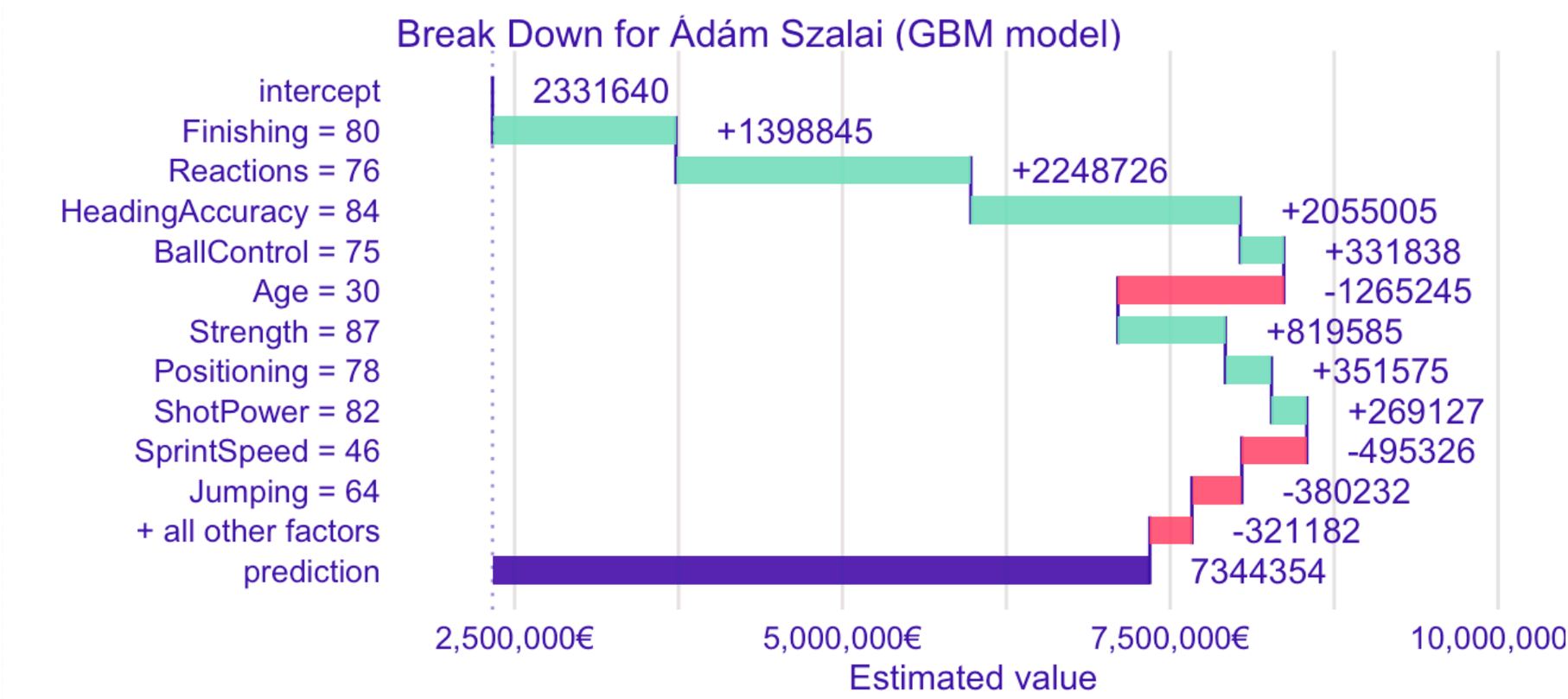
See: iBreakDown package
https://github.com/ModelOriented/iBreakDown



Break Down （local）

Waterfall like plots for additive and non-additive contributions

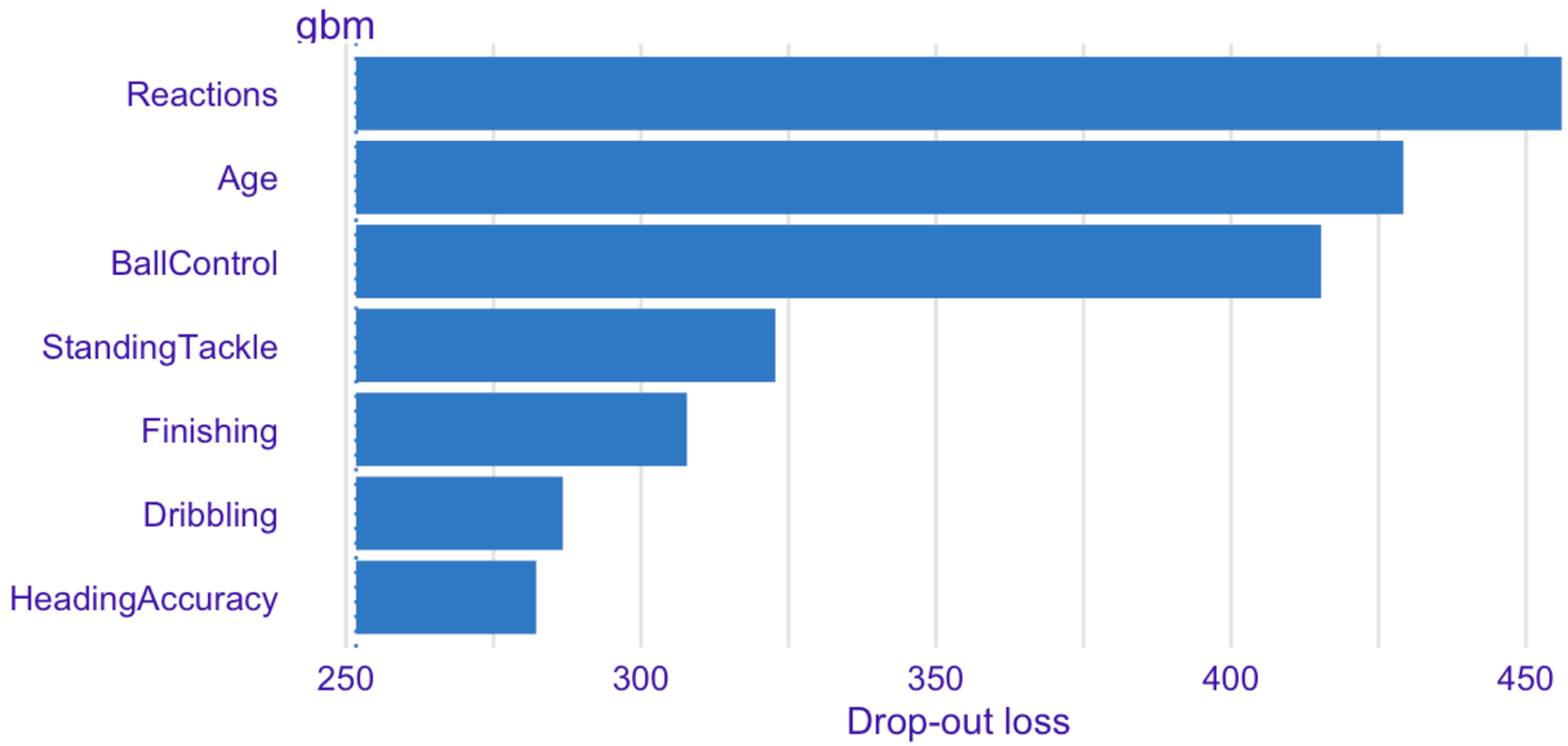
See: iBreakDown package
https://github.com/ModelOriented/iBreakDown



Feature Importance （ global ）

Permutation based Feature Importance

See: ingredients package
https://github.com/ModelOriented/ingredients

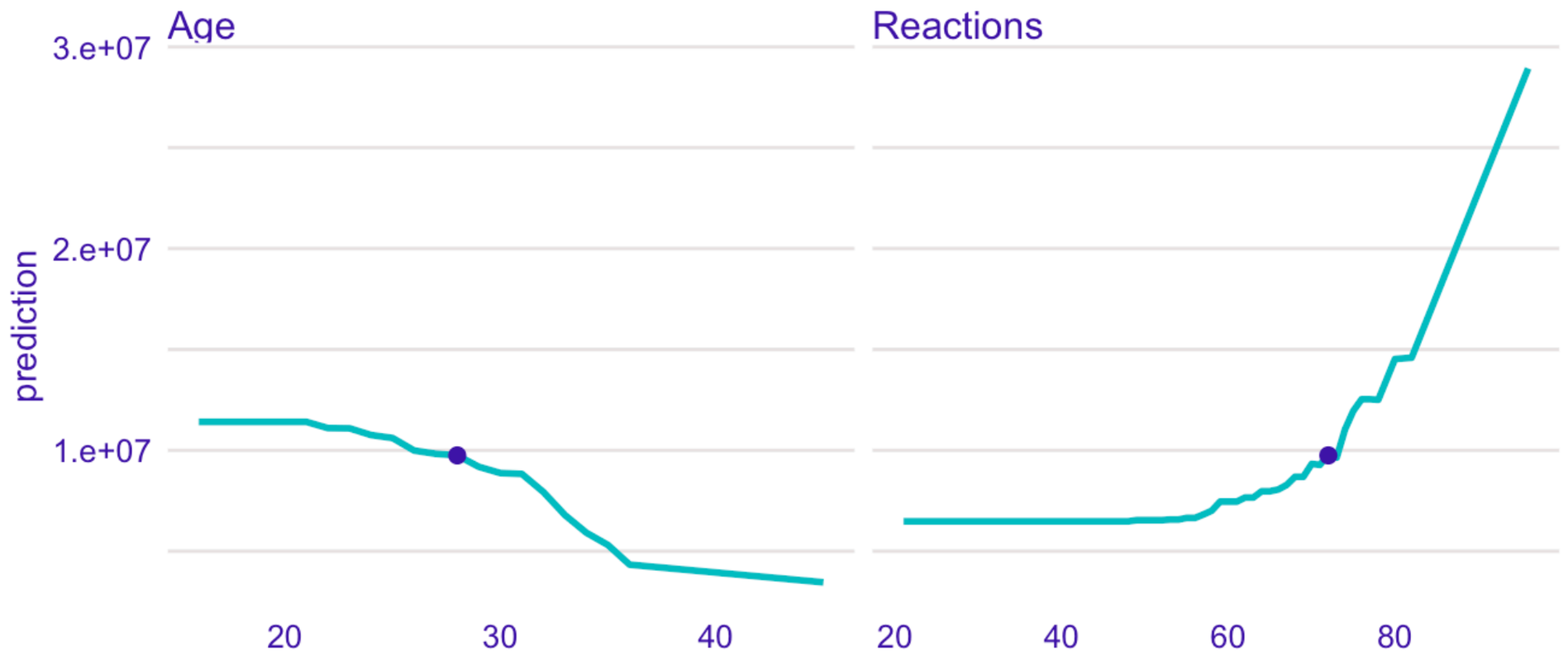


What if?

Ceteris Paribus （local）

Ceteris Paribus show model response profile as a function of a single variable.

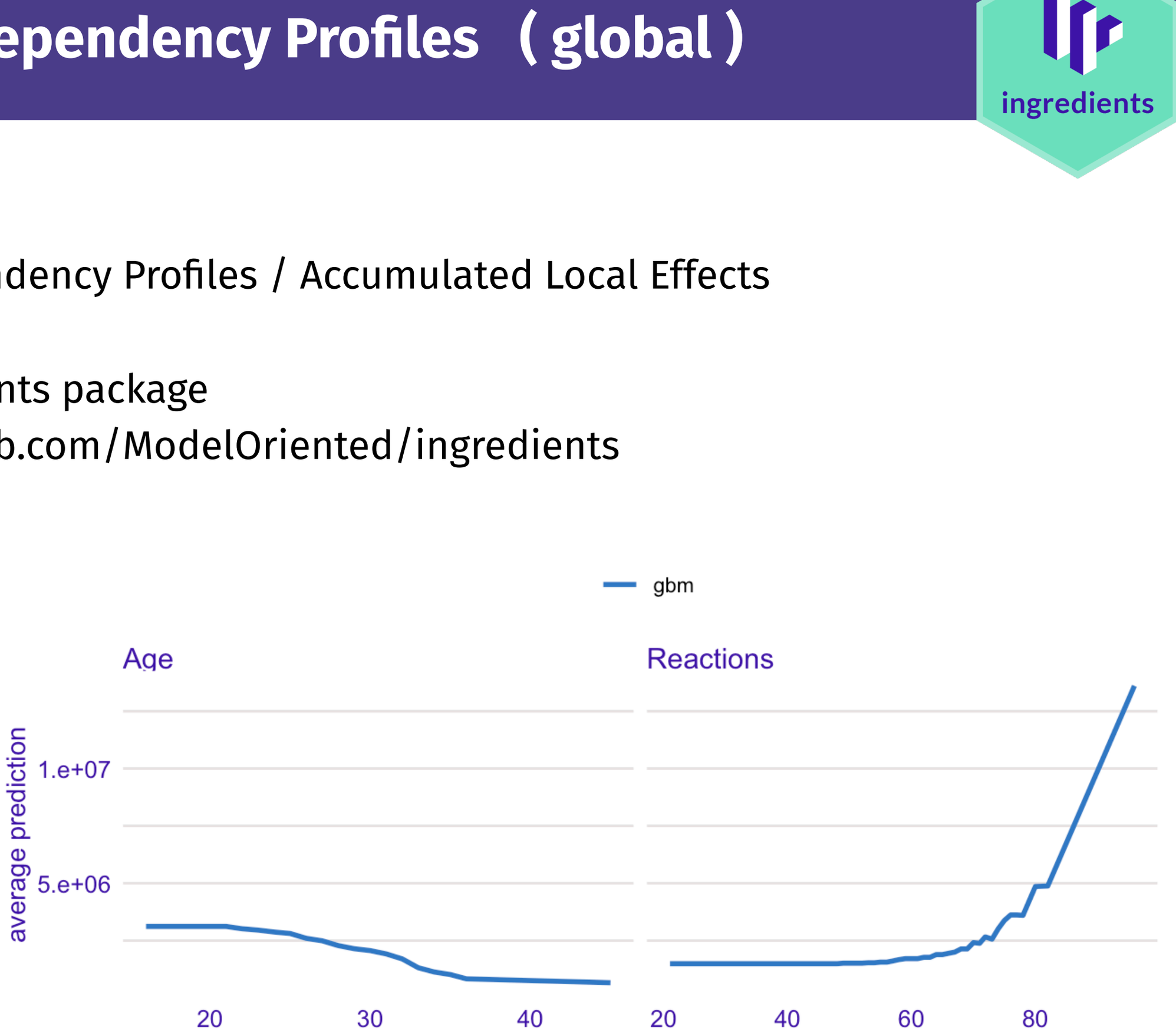
See: ingredients package
https://github.com/ModelOriented/ingredients



Partial Dependency Profiles （ global ）

Partial Dependency Profiles / Accumulated Local Effects

See: ingredients package
https://github.com/ModelOriented/ingredients

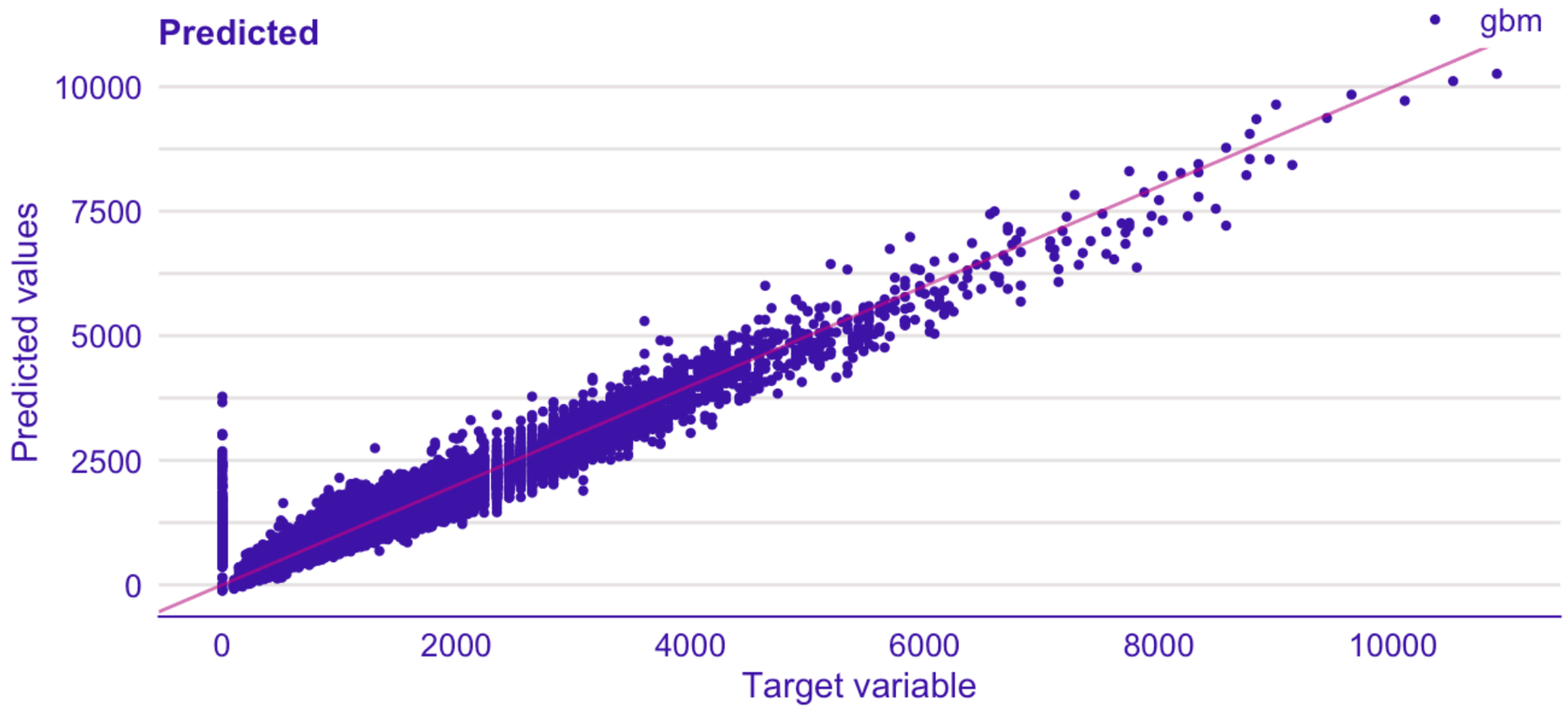


How good is the prediction?

Model Performance （ global ）

Functions for diagnosis of model residuals and model performance

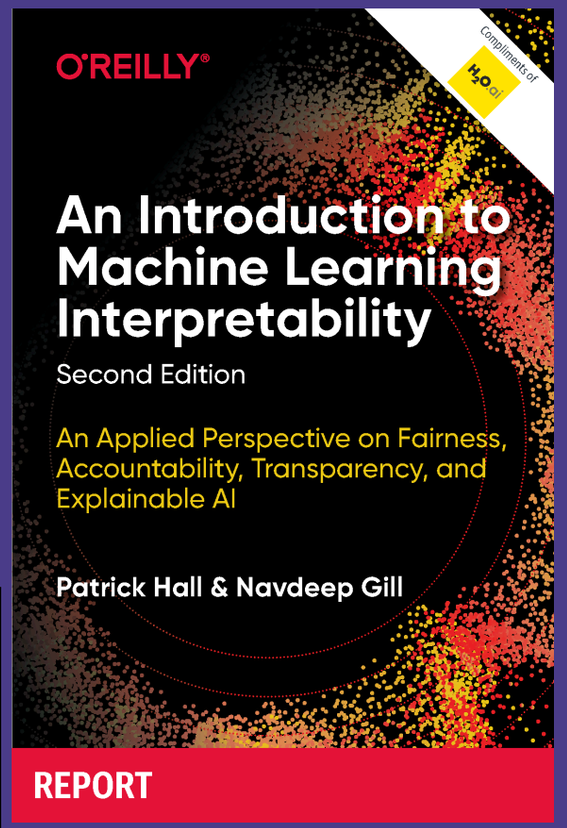
See: auditor package
https://github.com/ModelOriented/auditor



An Introduction to Machine Learning Interpretability

Navdeep Gill, Patrick Hall

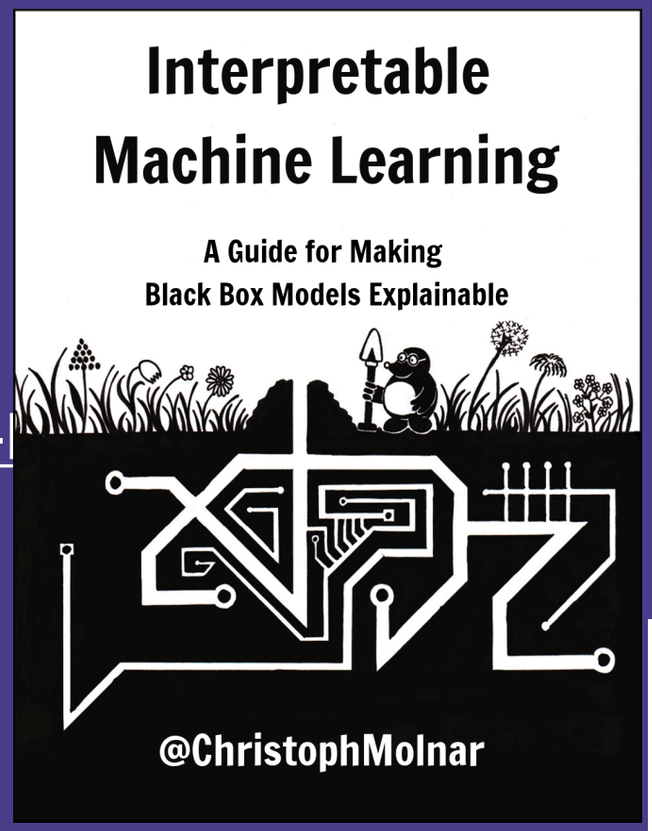
https://www.h2o.ai/oreilly-mli-booklet-2019/



Interpretable Machine Learning

Christoph Molnar

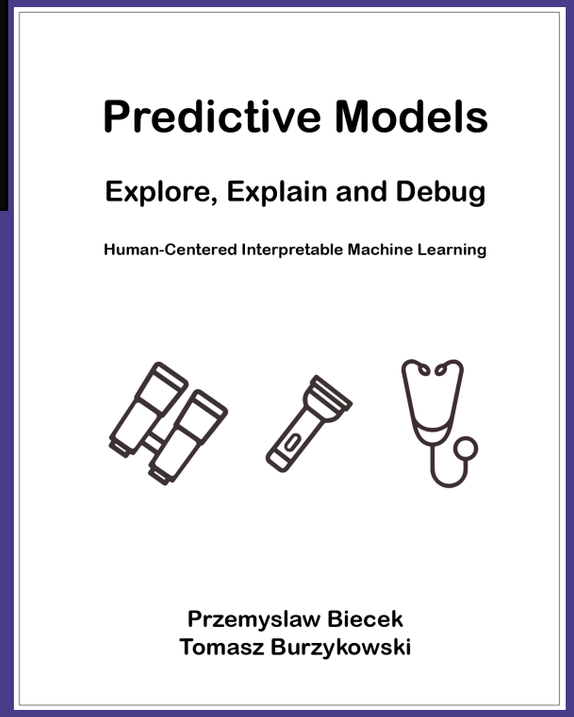
https://christophm.github.io/interpretable-ml-



Predictive Models: Explore, Explain, and Debug

Przemyslaw Biecek and Tomasz Burzykowski

https://pbiecek.github.io/PM_VEE/



Descriptive mAchine Learning EXplanations

Overview

The `DALEX` package (Descriptive mAchine Learning EXplanations) helps to understand how complex models are working. The main function `explain()` creates a wrapper around a predictive model. Wrapped models may then be explored and compared with a collection of local and global explainers. Recent developents from the area of Interpretable Machine Learning/eXplainable Artificial Intelligence.

The philosophy behind `DALEX` explanations is described in the Predictive Models: Explore, Explain, and Debug e-book. The `DALEX` package is a part of DrWhy.AI universe.

If you work with `scikitlearn`, `keras`, `H2O`, `mljar` or `mlr`, you may be interested in the DALEXtra package. It is an extension pack for `DALEX` with easy to use connectors to models created in these libraries.

https://github.com/ModelOriented/DALEX

Thank you!

Predictive models: Explore, Explain and Debug (locally)

Local methods are designed to better understand model behaviour around a single observation.

Prepare model explainer (Ch. 2)

The DALEX ::explain() function creates model adapters: objects with standardised structure that are used by other methods for model exploration and explanations.

```
library("DALEX")
explain(model, data, y, label,
        predict_func, residual_fun)
```

Models can be trained in different languages with various libraries. New libraries will emerge, existing libraries will change. Various models have different structures. This is why we need uniform adapters.

General workflow

Function explain() turns models into *explainers* - wrappers with uniform structure.

Specific functions turn *explainers* into *explanations.*

For *explanations* one can use generic functions: print - short text summary, plot - a ggplot2 plot, plotD3 - a D3 plot based on r2d3 package, describe - a text summary for an explanation.

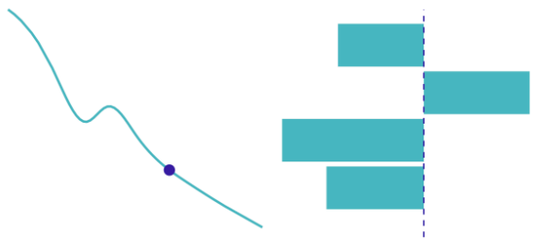
```
print(explanation)
plot(explanation)
plotD3(explanation)
describe(explanation)
```

Ceteris Paribus Profiles (Ch. 6)

How would the model response change for a particular observation if only a single feature is changed?

Best for:
What if questions.
A few interpretable features.
Be careful when:
Features are correlated.

```
library("ingredients")
ceteris_paribus(explainer,
        observation, variables)
```

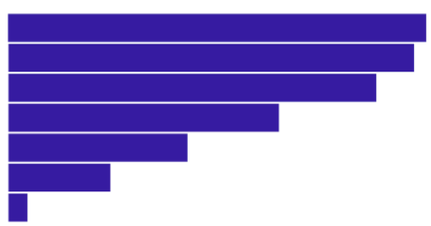


Profile Oscillations (Ch. 7)

How sensitive is the model response on individual features?

Best for:
Selection of important CP profiles.
Be careful when:
Features are correlated.

```
calculate_oscillations(explanation)
```

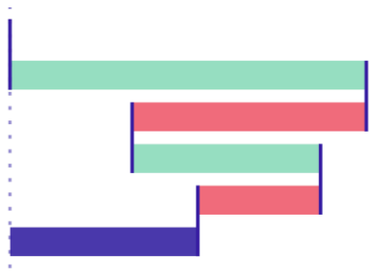


Break Down attributions (Ch. 9)

How the average model response change when new features are being fixed in the observation of interest?

Best for:
Why questions.
Moderate number of features.
Be careful when:
Features are correlated.

```
library("iBreakDown")
break_down(explainer, observation)
```

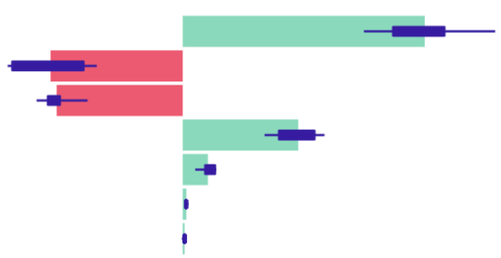


Shapley additive values (Ch. 11)

How the model response can be decompose into additive attributions.

Best for:
Why questions.
Moderate number of features.
Be careful when:
Features are correlated.
Model has interactions.

```
shap(explainer, observation)
```



Local Interpretable Model (Ch. 12)

Local Interpretable Model-Agnostic Explanations (LIME) shows sparse explanations for selected aspects.

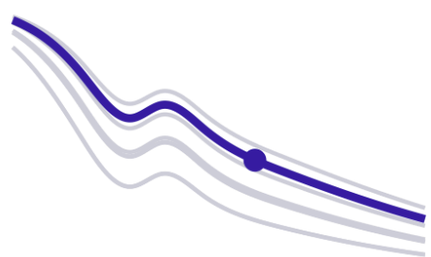
Best for:
Why questions.
Lots of non-interpretable features.
Be careful when:
Sparse explanations make no sense.

```
library("ingredients")
lime(explainer, observation)
```

Local diagnostics (Ch. 8)

Two or more explanations can be superimposed on a single plot.

Instance level analysis of local fit.
Diagnostic for local residuals.
Stability of predictions.



Learn more at **Predictive Models: Explore, Explain, and Debug** Przemysław Biecek, Tomasz Burzykowski, https://pbiecek.github.io/PM_VEE/