Explanatory Model Analysis

Przemysław Biecek
/ˈpʂɛ.mɛk/

QuantUniversity
Summer School
31/08/2021

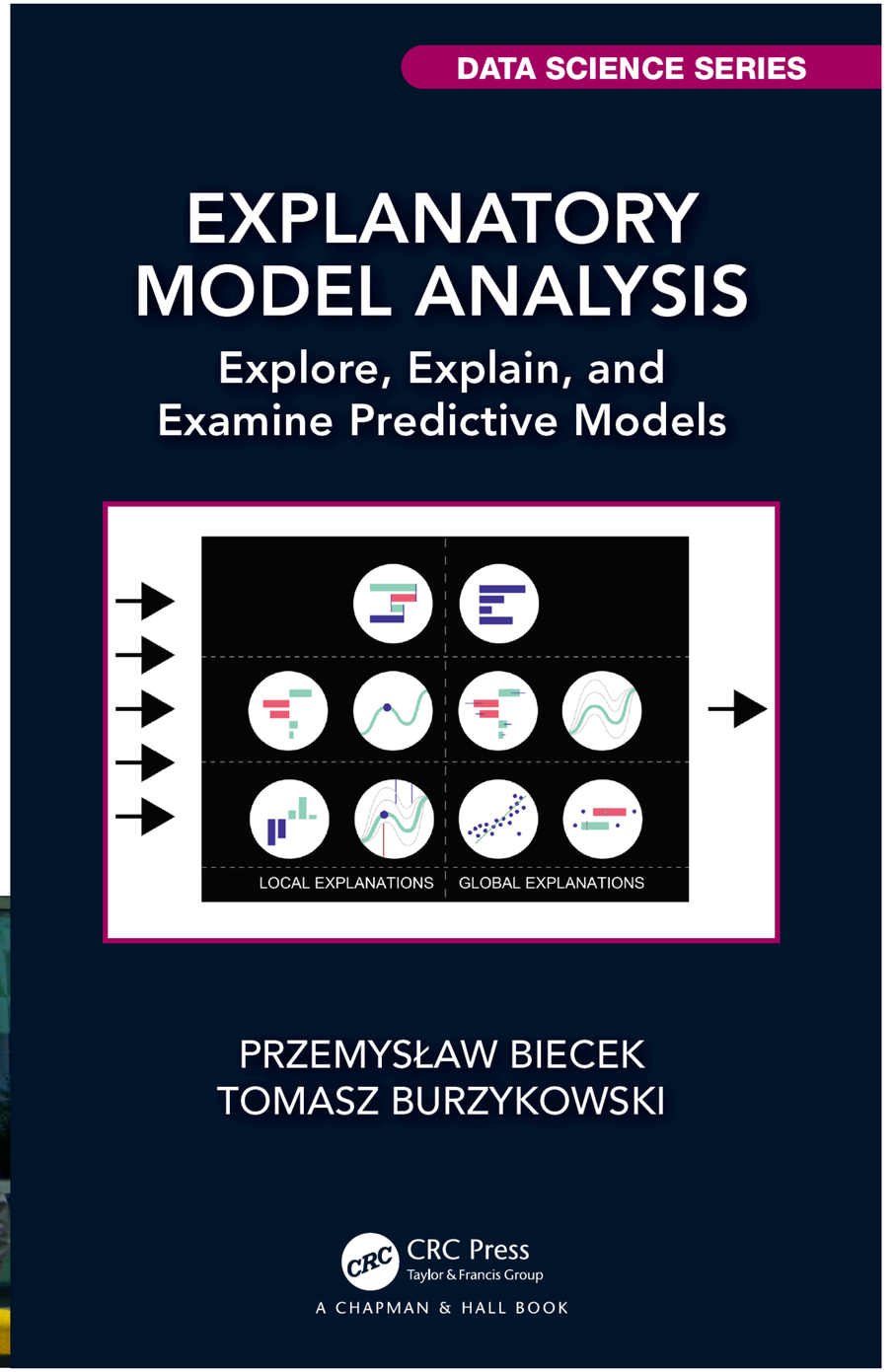
Agenda

- Why do we need EMA?
- What the EMA process looks like?
- What an example application looks like?
- How interaction helps with EMA?

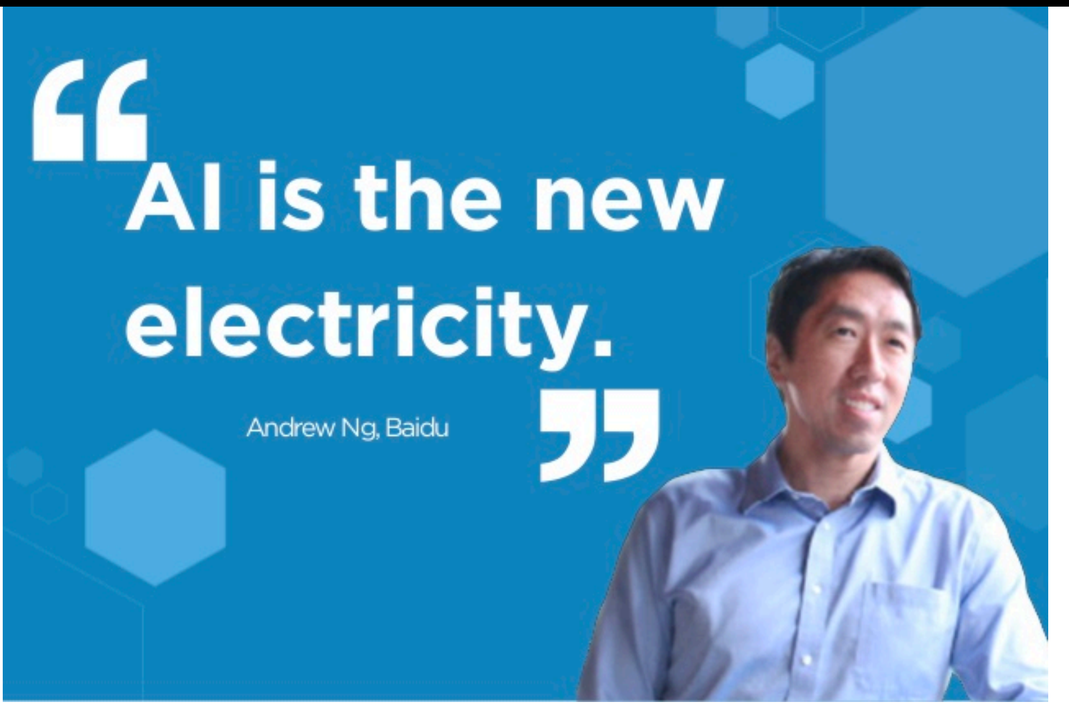
About me

Head of **MI2 Data Lab** at Warsaw University of Technology - team that develops tools and methods for **Responsible ML.**

Research area: Human Oriented AI Evidence based Machine Learning



Why do we need Explanatory Model Analysis?



https://www.slideshare.net/jaypod/digitaltransformation50soundbites/19-Data_is_the_new_oilClive
https://www.newworldai.com/forget-the-hype-what-every-business-leader-needs-to-know-about-artificial-intelligence-now/

Predictive models are like nuclear power,
it has huge potential but you have to be
very careful

governmentCIO MEDIA & RESEARCH

Topics Video Events Podcasts

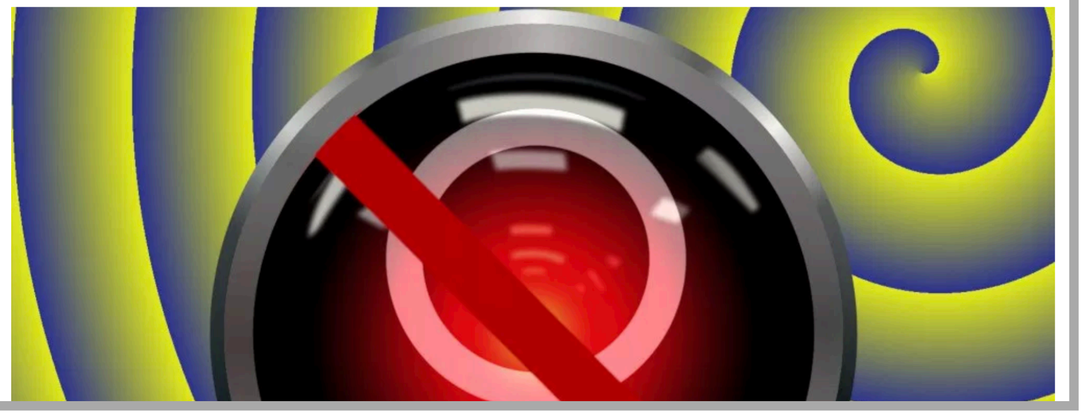
Machine Learning is Creating a Crisis in Science

The adoption of machine-learning techniques is contributing to a worrying number of research findings that cannot be repeated by other researchers.

Kevin McCaney
Wed, 02/27/2019 - 11:28

Photo credit: metamorworks/iStock

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



Forbes Billionaires Innovation Leadership Money Consumer Industry L

61,215 views | Mar 23, 2014, 09:00am

Why Google Flu Is A Failure

Steven Salzberg Contributor
Pharma & Healthcare

It seemed like such a good idea at the time.

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

List of failures in AI applications

Read more at: https://romanlutz.github.io/ResponsibleAI/

Speech Detection

- Oh dear... AI models used to flag hate speech online are, er, racist against black people
- The Risk of Racial Bias in Hate Speech Detection
- Toxicity and Tone Are Not The Same Thing: analyzing the new Google API on toxicity, PerspectiveAPI.
- Voice Is the Next Big Platform, Unless You Have an Accent
- Google's speech recognition has a gender bias
- Fair Speech report by Stanford Computational Policy Lab, also covered in Speech recognition algorithms may also have racial bias
- Automated moderation tool from Google rates People of Color and gays as "toxic"
- Someone made an AI that predicted gender from email addresses, usernames. It went about as well as expected

Image Labelling & Face Recognition

- Google Photos identified two black people as 'gorillas'
- When It Comes to Gorillas, Google Photos Remains Blind
- The viral selfie app ImageNet Roulette seemed fun – until it called me a racist slur
- Google Is Investigating Why it Trained Facial Recognition on 'Dark Skinned' Homeless People
- Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification
- Machines Taught by Photos Learn a Sexist View of Women
- Tenants sounded the alarm on facial recognition in their buildings. Lawmakers are listening.
- Google apologizes after its Vision AI produced racist results

Public Benefits & Health

- A health care algorithm affecting millions is biased against black patients
- What happens when an algorithm cuts your health care
- China Knows How to Take Away Your Health Insurance
- Foretelling the Future: A Critical Perspective on the Use of Predictive Analytics in Child Welfare

Lending & Credit approval

- Gender Bias Complaints against Apple Card Signal a Dark Side to Fintech
- Exploring Racial Discrimination in Mortgage Lending: A Call for Greater Transparency
- DFS Issues Guidance to Life Insurers on Use of "External Data" in Underwriting Decisions

Hiring

- Amazon scraps secret AI recruiting tool that showed bias against women
- Automated Employment Discrimination
- Help wanted: an examination of hiring algorithms, equity, and bias
- All the Ways Hiring Algorithms Can Introduce Bias
- Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices
- Help Wanted - An Examination of Hiring Algorithms, Equity, and Bias
- Wanted: The 'perfect babysitter.' Must pass AI scan for respect and attitude.

Employee evaluation

- Houston Schools Must Face Teacher Evaluation Lawsuit
- How Amazon automatically tracks and fires warehouse workers for 'productivity'

Pre-trial risk assessment and criminal sentencing

- Machine Bias
- How We Analyzed the COMPAS Recidivism Algorithm
- GitHub repository for COMPAS analysis

List of Artificial Intelligence Incidents

Read more at: https://incidentdatabase.ai/

INCIDENT DATABASE

Star 68

Discover • Submit

Welcome to the Artificial Intelligence Incident Database

Search all incident reports

Search

Entering text above will search across more than 1200 incident reports

Latest Incident Report

Police Are Telling ShotSpotter to Alter Evidence From Gunshot-Detecting AI

Aug 16, 2021, 2:00 a.m.

The central assumption of the machine learning models:

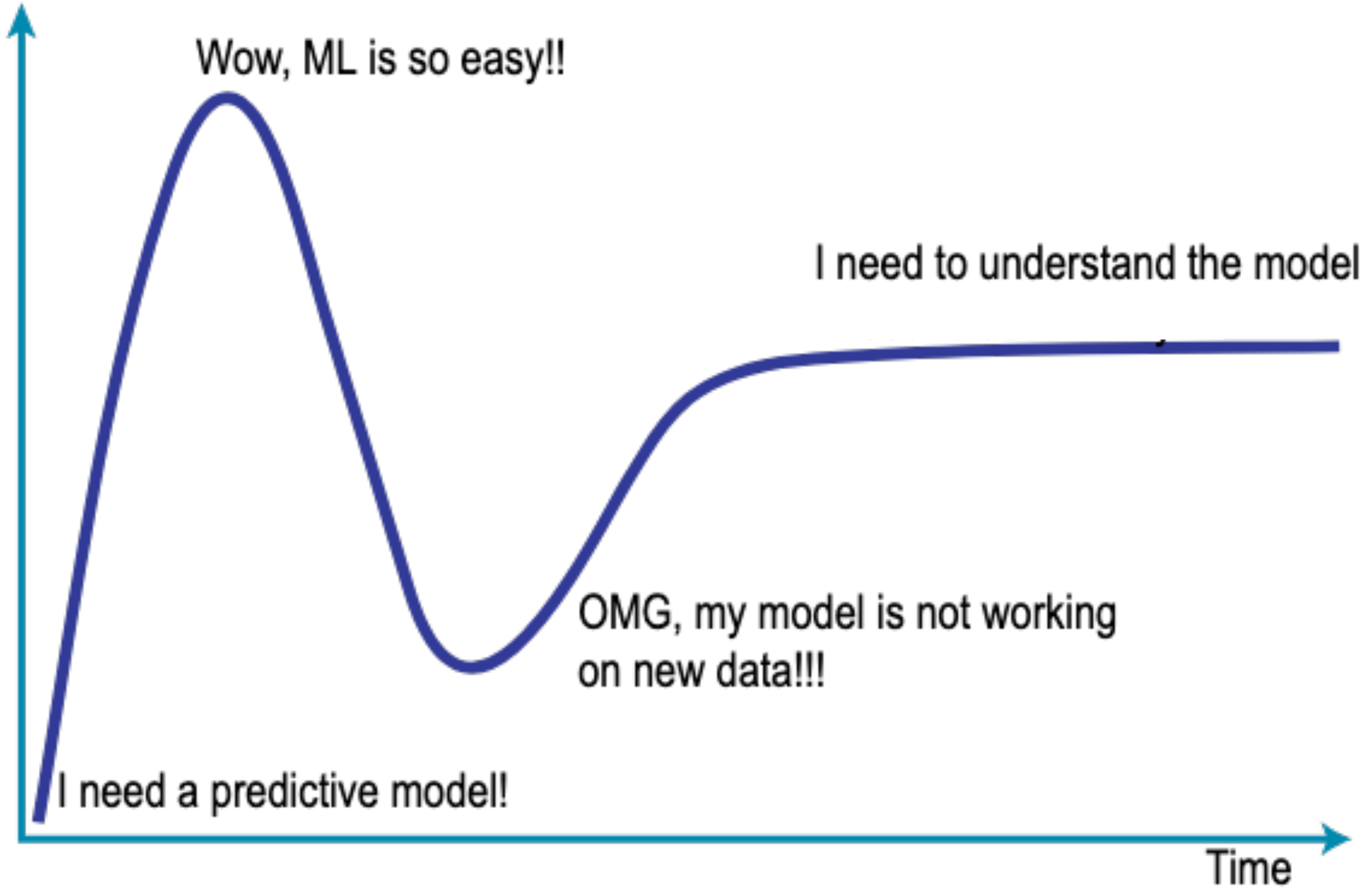
The future will be similar to the past

The central assumption of the machine learning models:

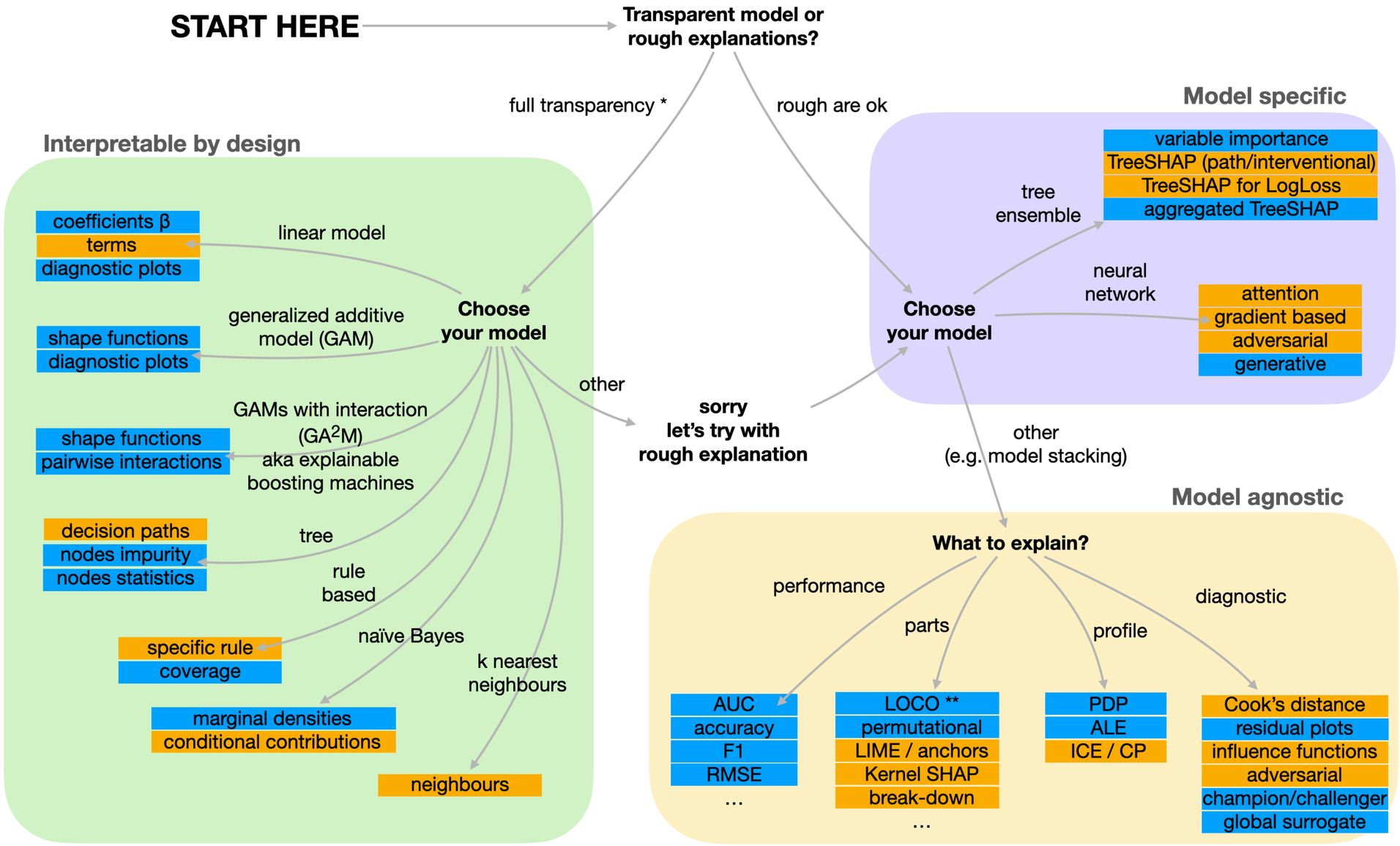
COVID19

The future will be similar to the past

Hype Cycle for Predictive Models



Plenty of tools for a better understanding of models



https://arxiv.org/abs/2009.13248

The three faces of model understanding

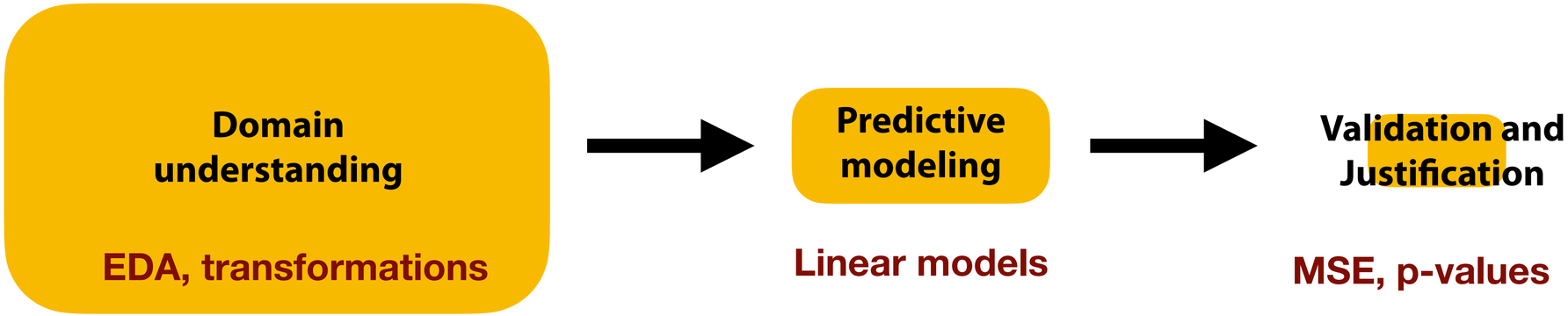
different names for different needs, similar tools and methods



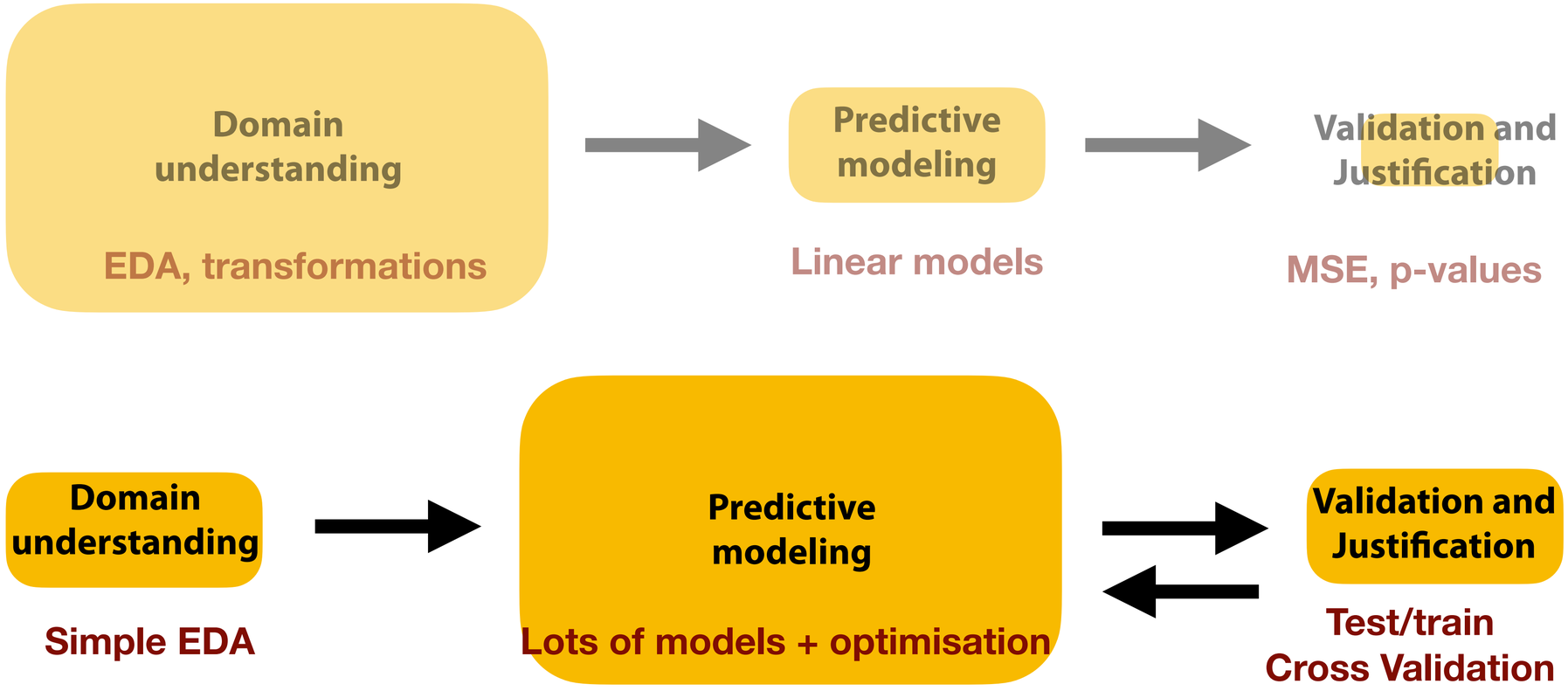
	eXplainable Artificial Inteligence XAI	**Interpretable Machine Learning IML**	**Explanatory Model Analysis EMA**
Main stakeholders	users	auditors	model developers
Goal	increase trust	verify assumptions	decrease trust
Typical application	understandable explanations	increased transparency	model debugging
Part of the process	delivering predictions	external model validation	internal critical analysis
Why do we need it	right to explanation	do no harm	life long performace

What the EMA process looks like?

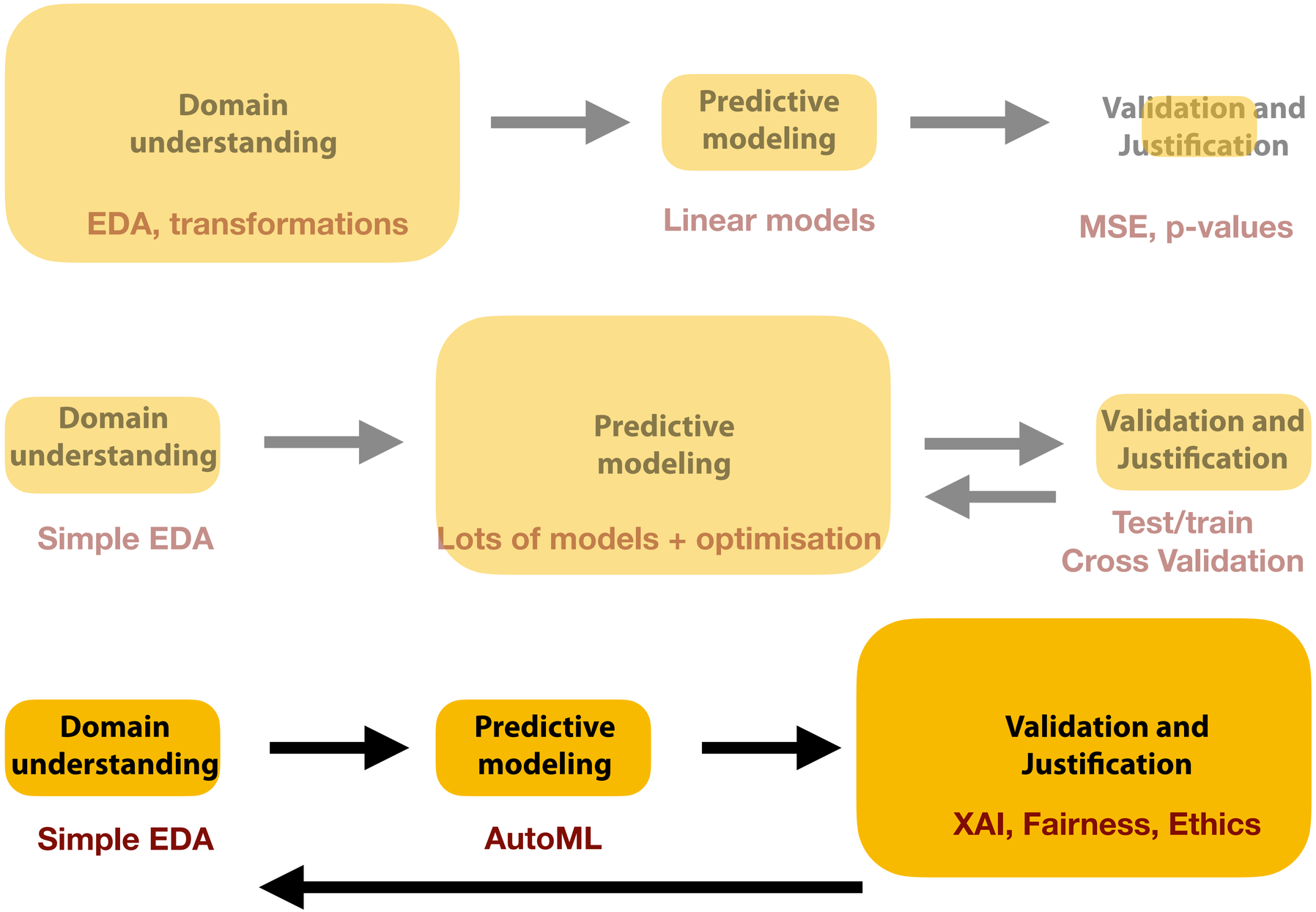
Shift in our focus: Statistics

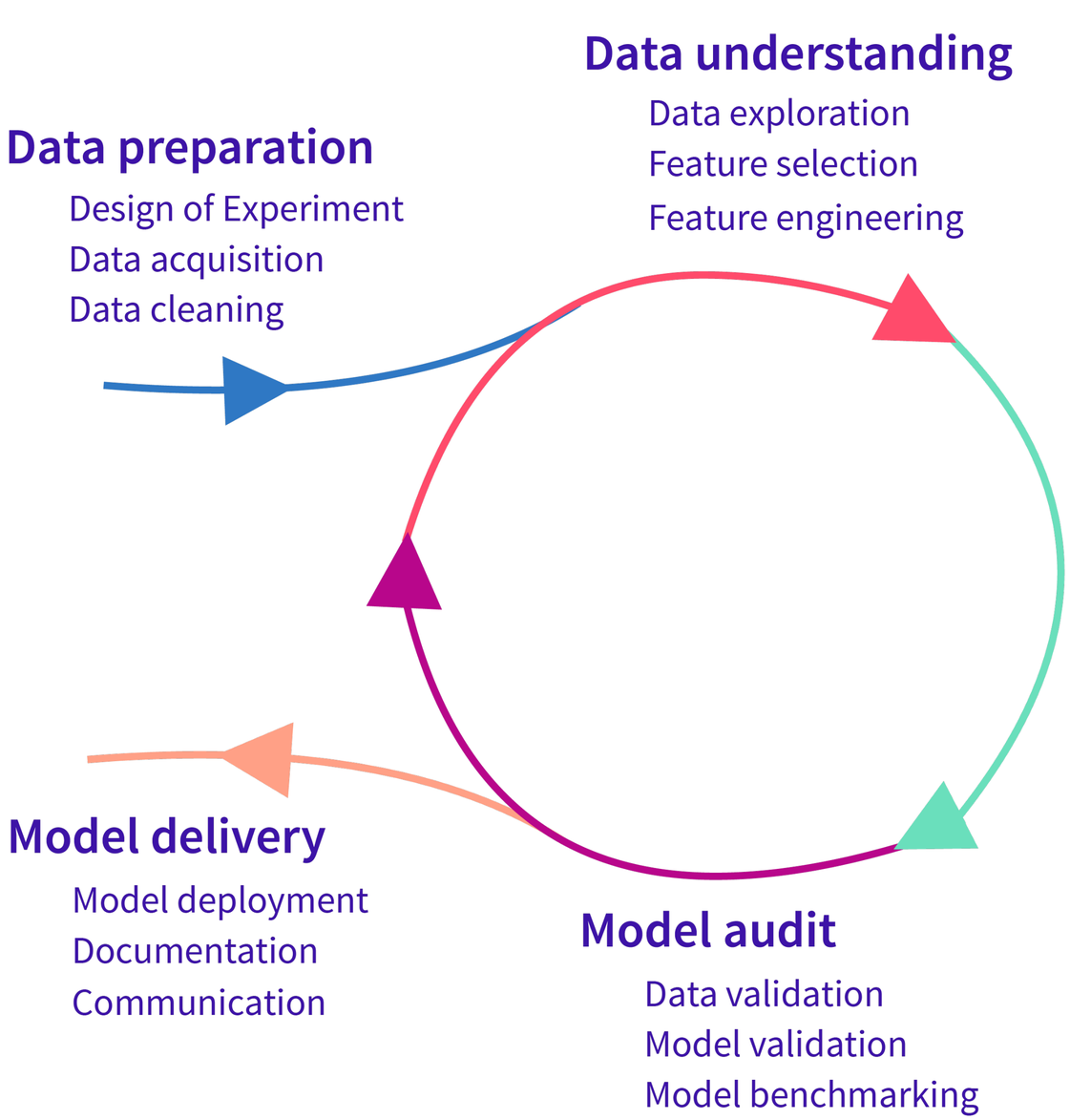


Shift in our focus: Machine Learning



Shift in our focus: Human-Model interactions



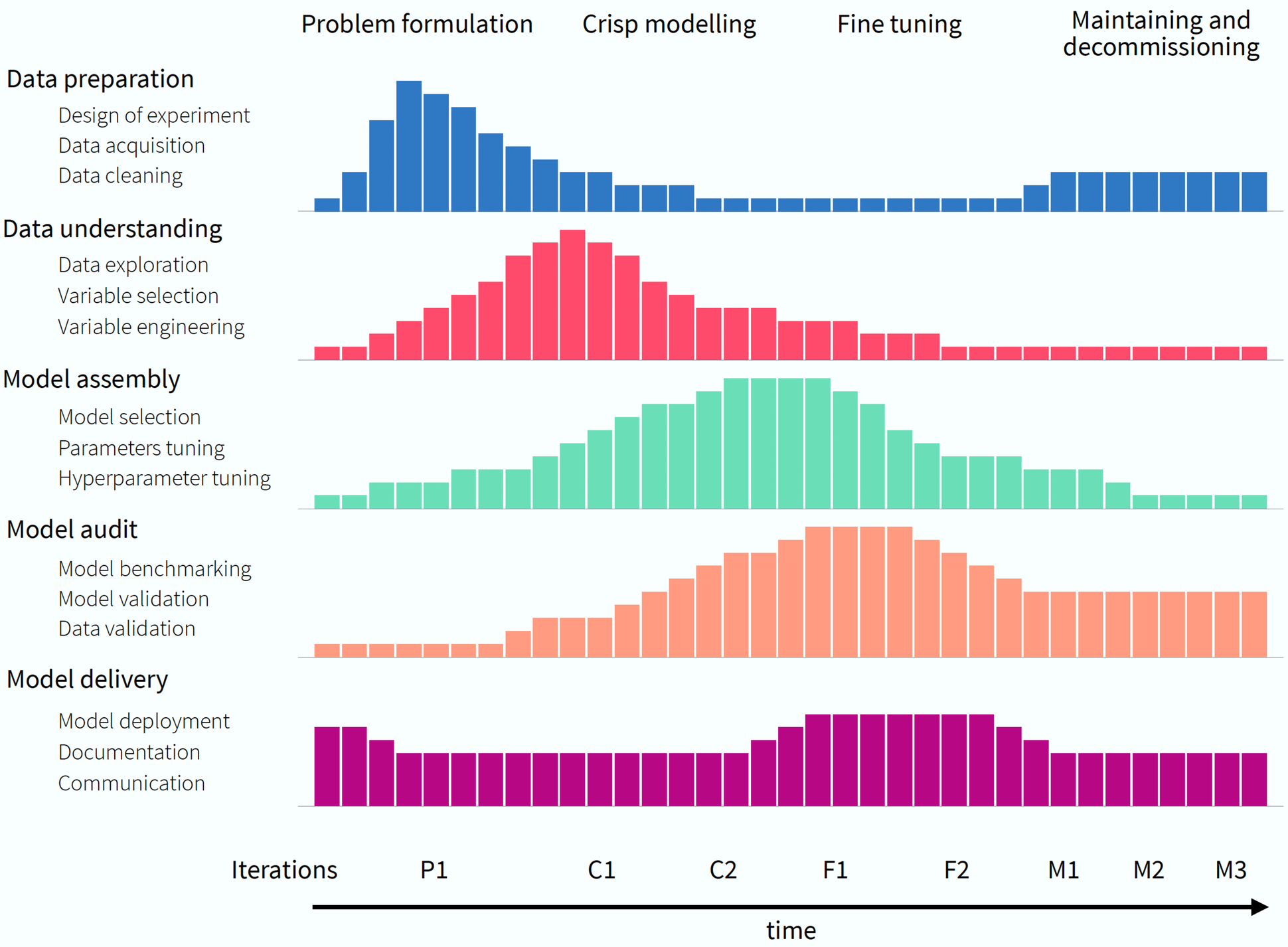


explore

explain

debug

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



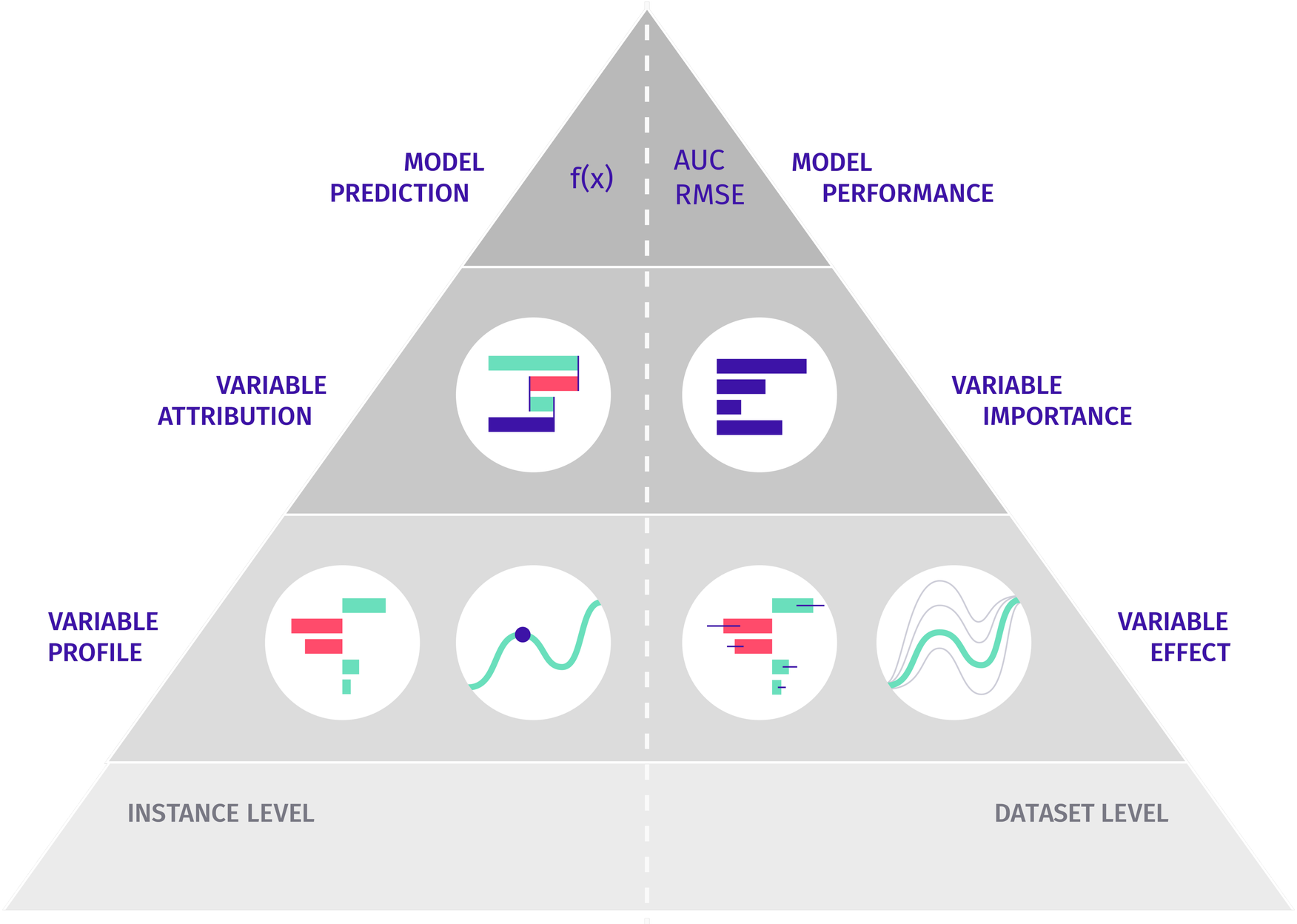
explore

explain

debug

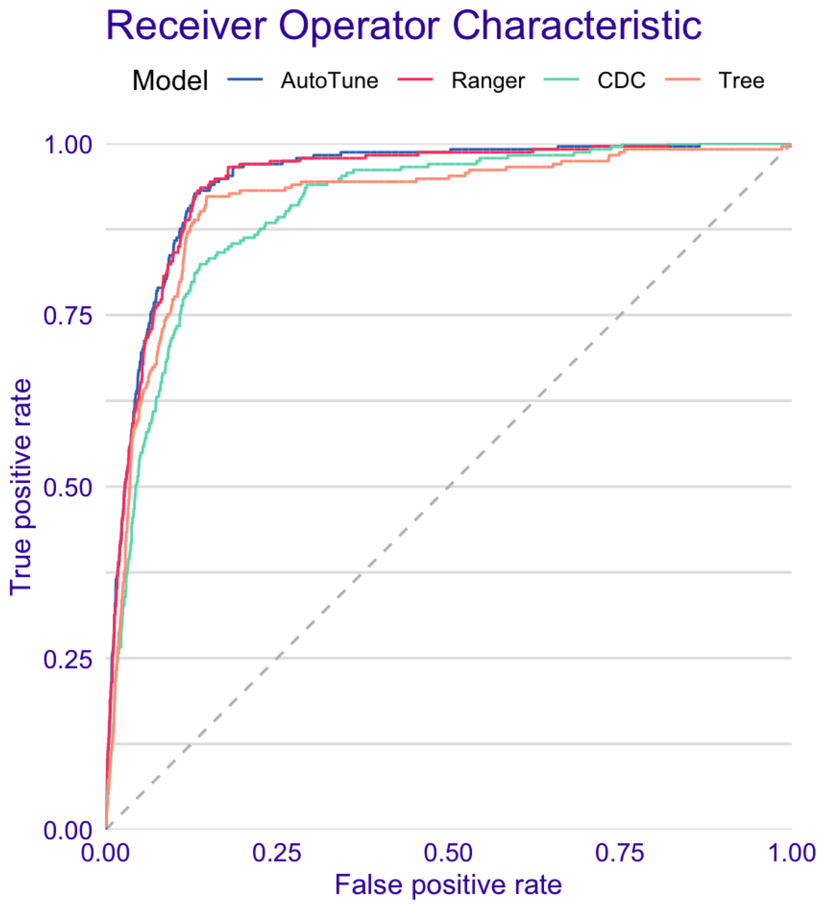
https://ema.drwhy.ai/

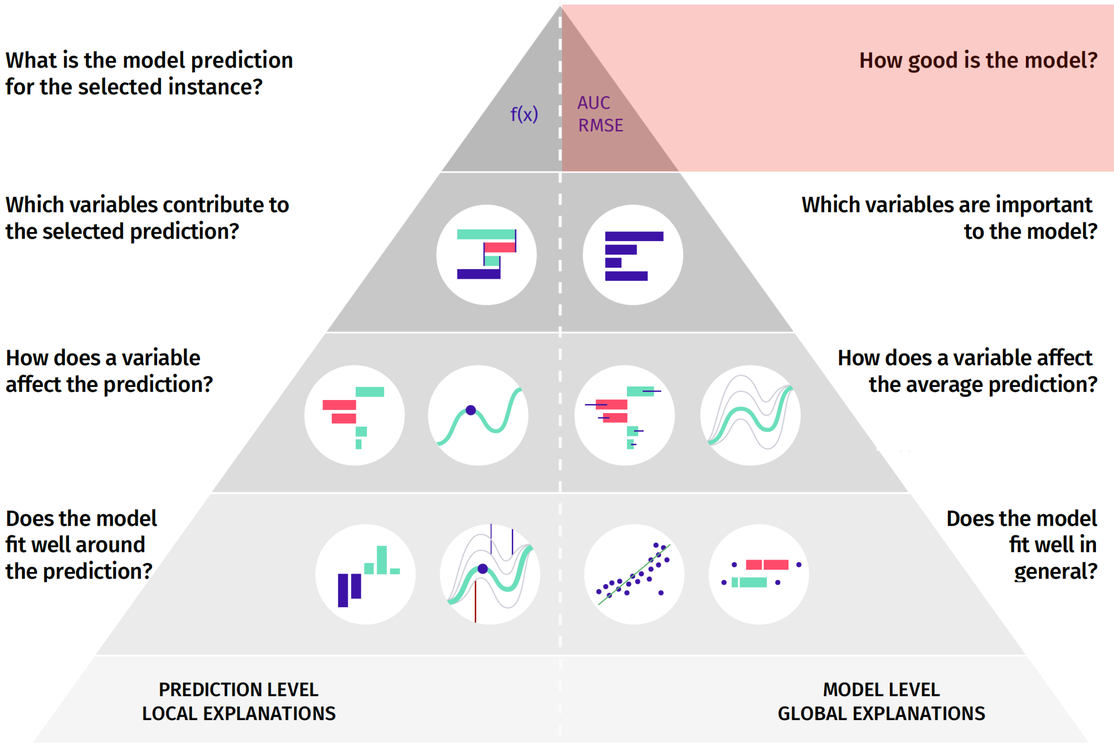
Model Exploration Stack



Explore Model Performance

MSE, RMS, AUC, F1 and many others





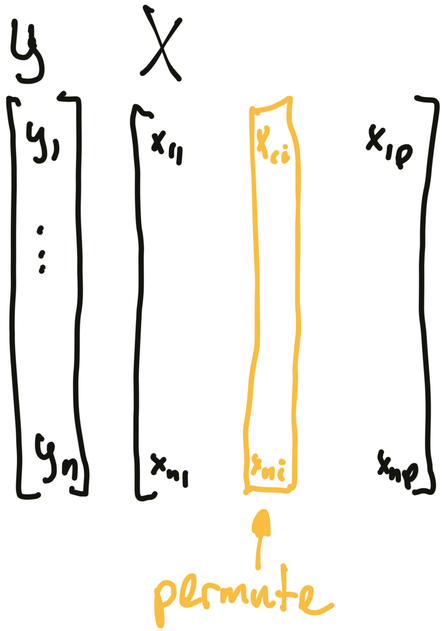
Morning sickness / pregnancy	Pregnant	Not Pregnant	
Has sickness	**TP** = 39	**FP** = 150	PPV = Prec = 20.6%
Has not	**FN** = 61	**TN** = 850	NPV = 93.3%
	Sensitivity = Recall = 39%	Specificity = 85%	F1 = 33.8%

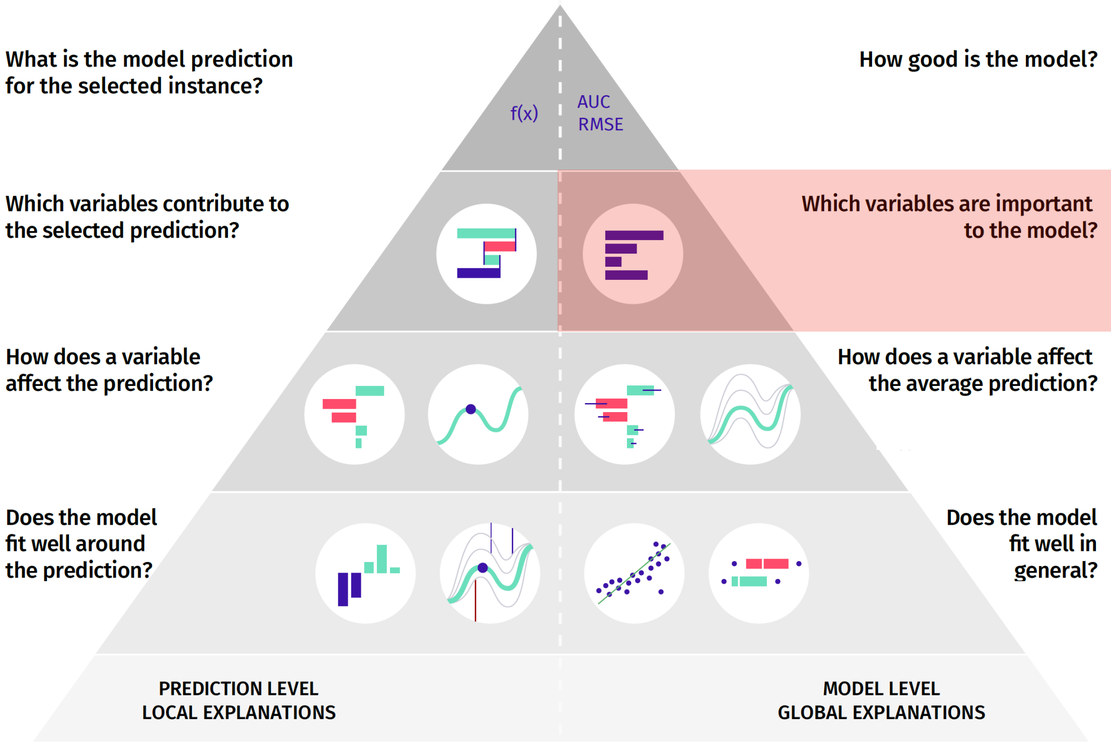
Explore Variable Importance

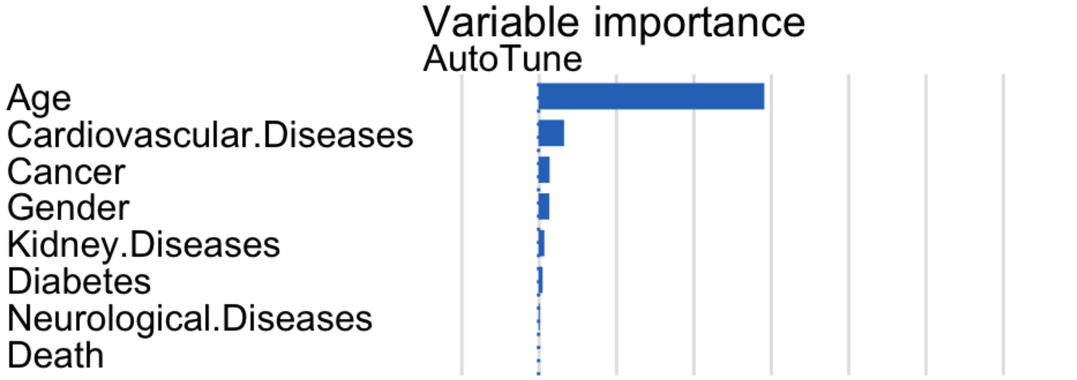
Permutational Variable Importance

Model specific assessments

$$VI(i) = L(f, X^{perm(i)}, y) - L(f, X, y)$$







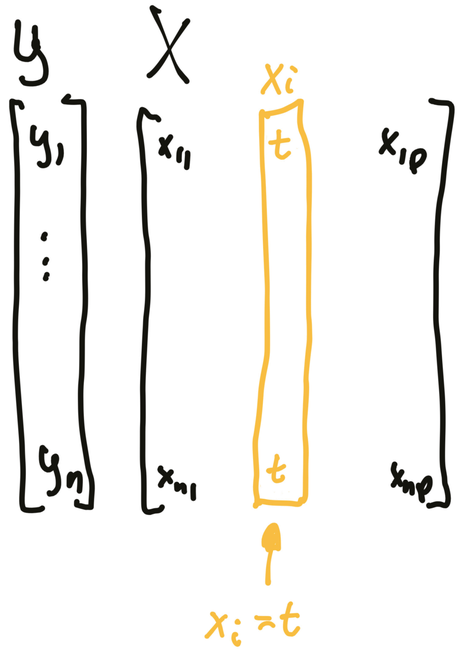
Explore Variable Effects

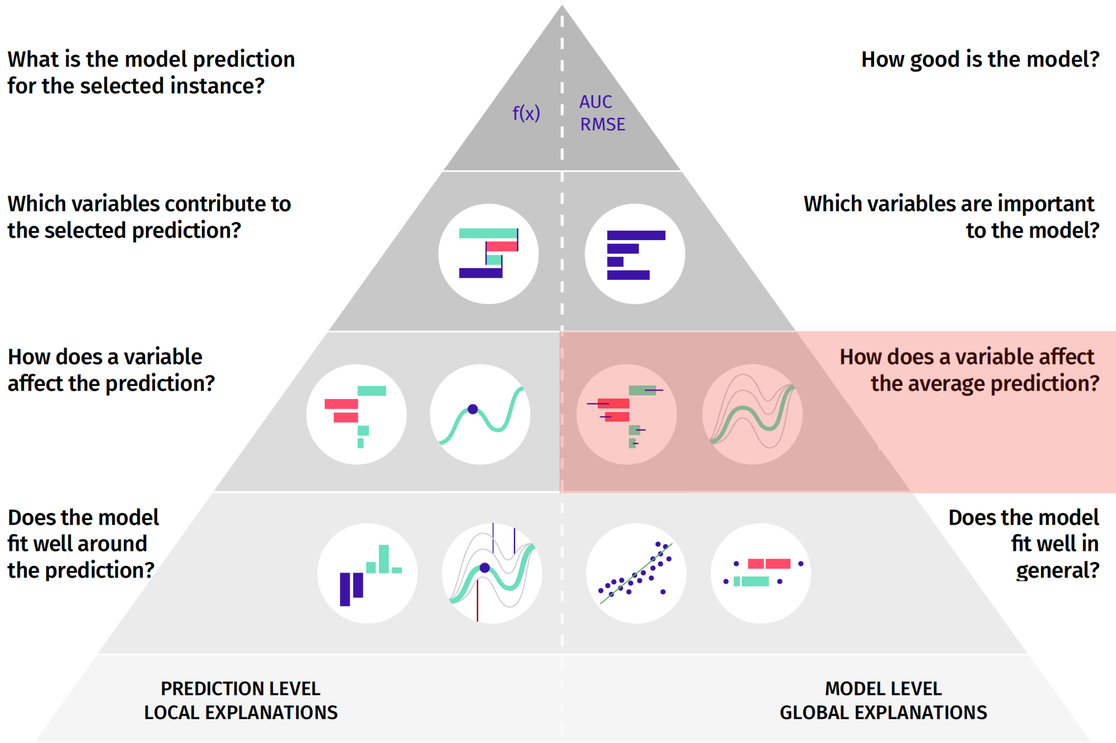
Partial Dependence Plots

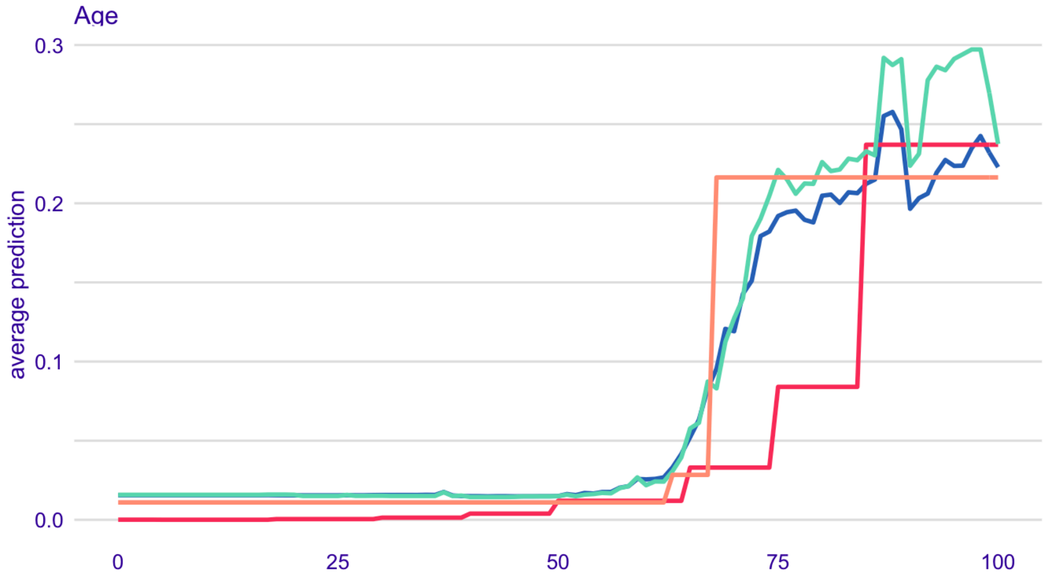
Accumulated Local Effects

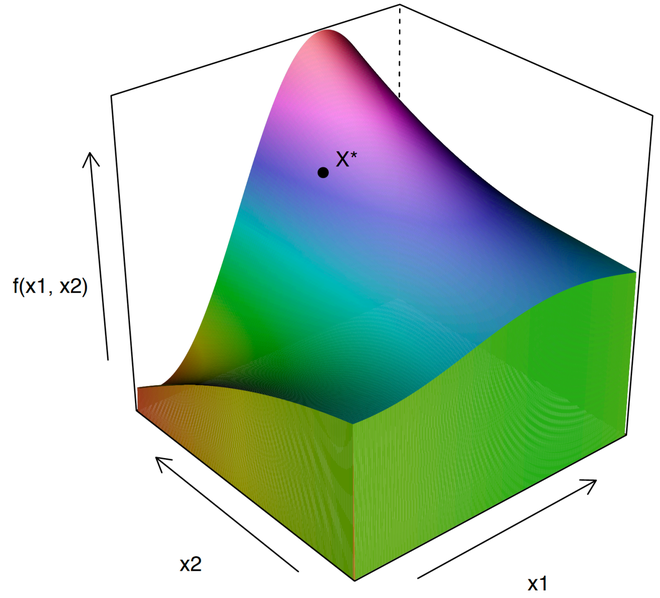
Shapley Values Curves

$$PD(i,t) = E\left[f(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_p)\right]$$

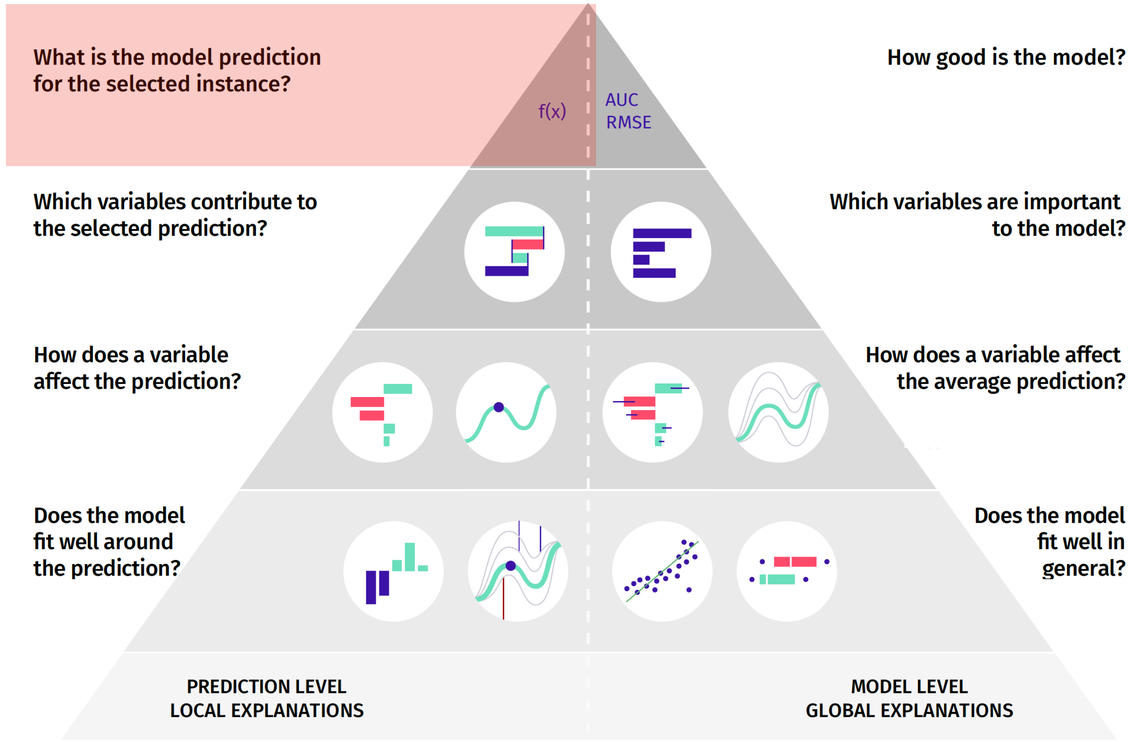




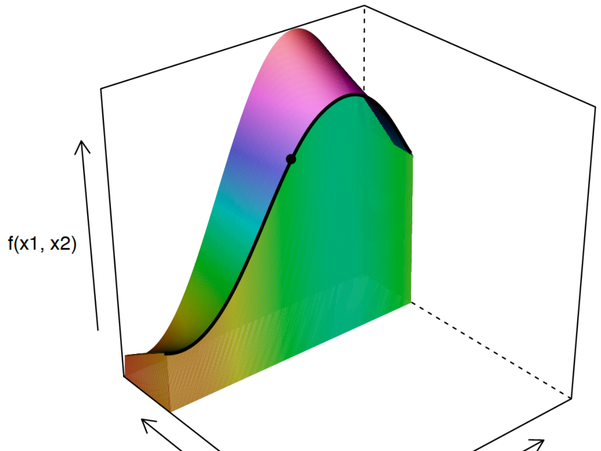




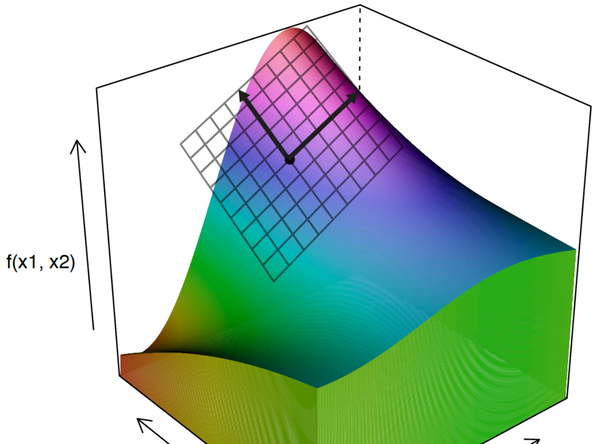
Model Exploration Stack



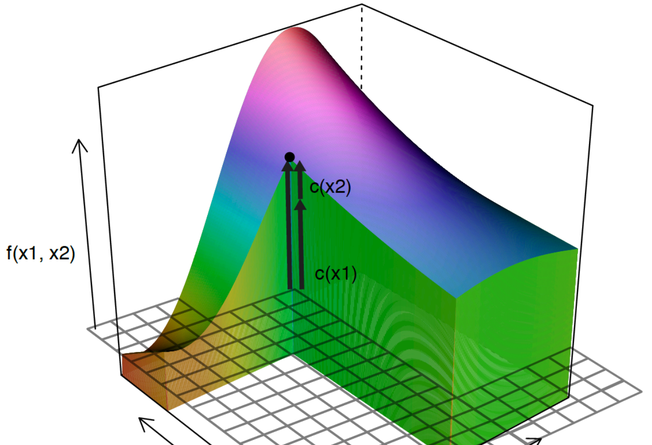
Ceteris Paribus / Individual Conditional Expectations



Local approximations (e.g. LIME)



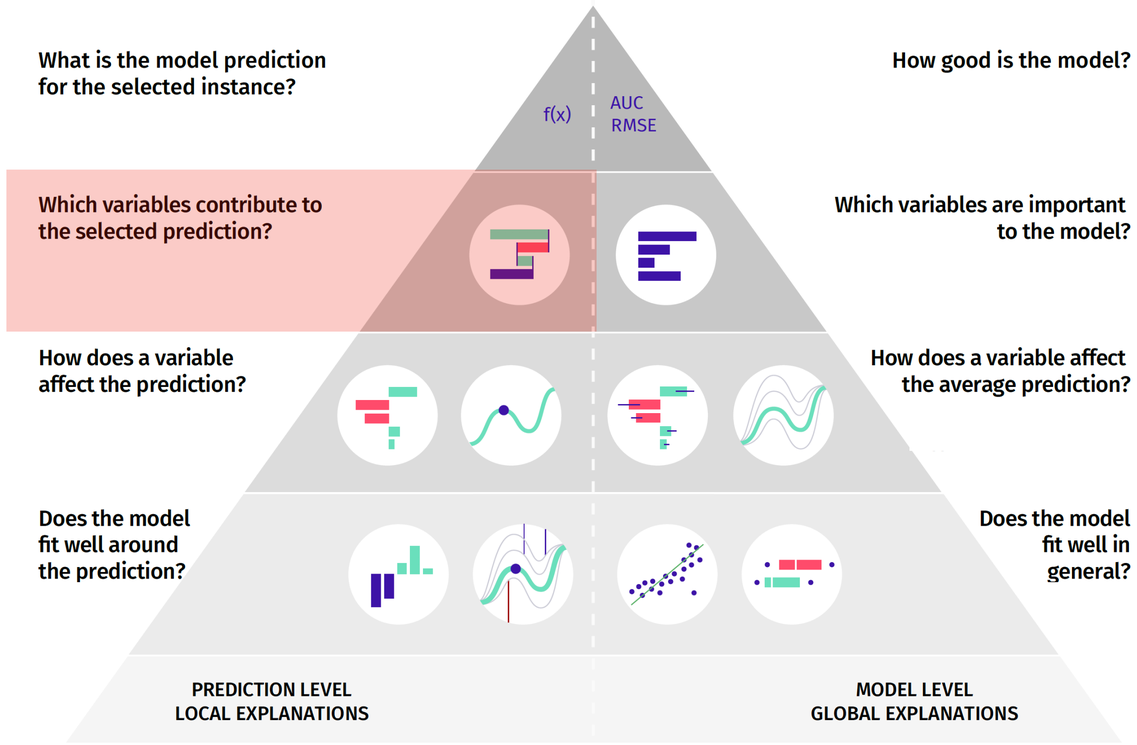
Local decompositions (e.g. SHAP)

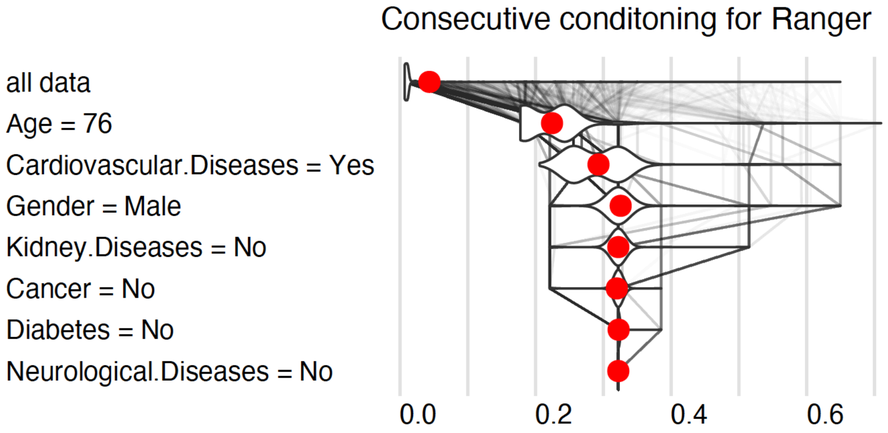


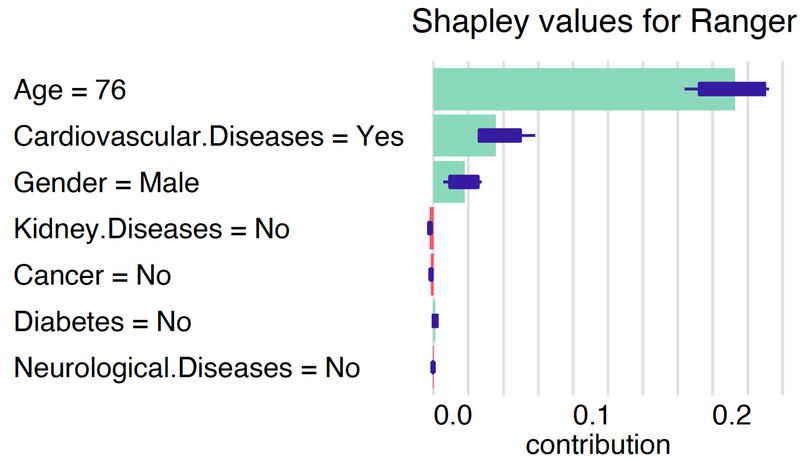
Explore Local Contributions

LIME, *SHAP, BD

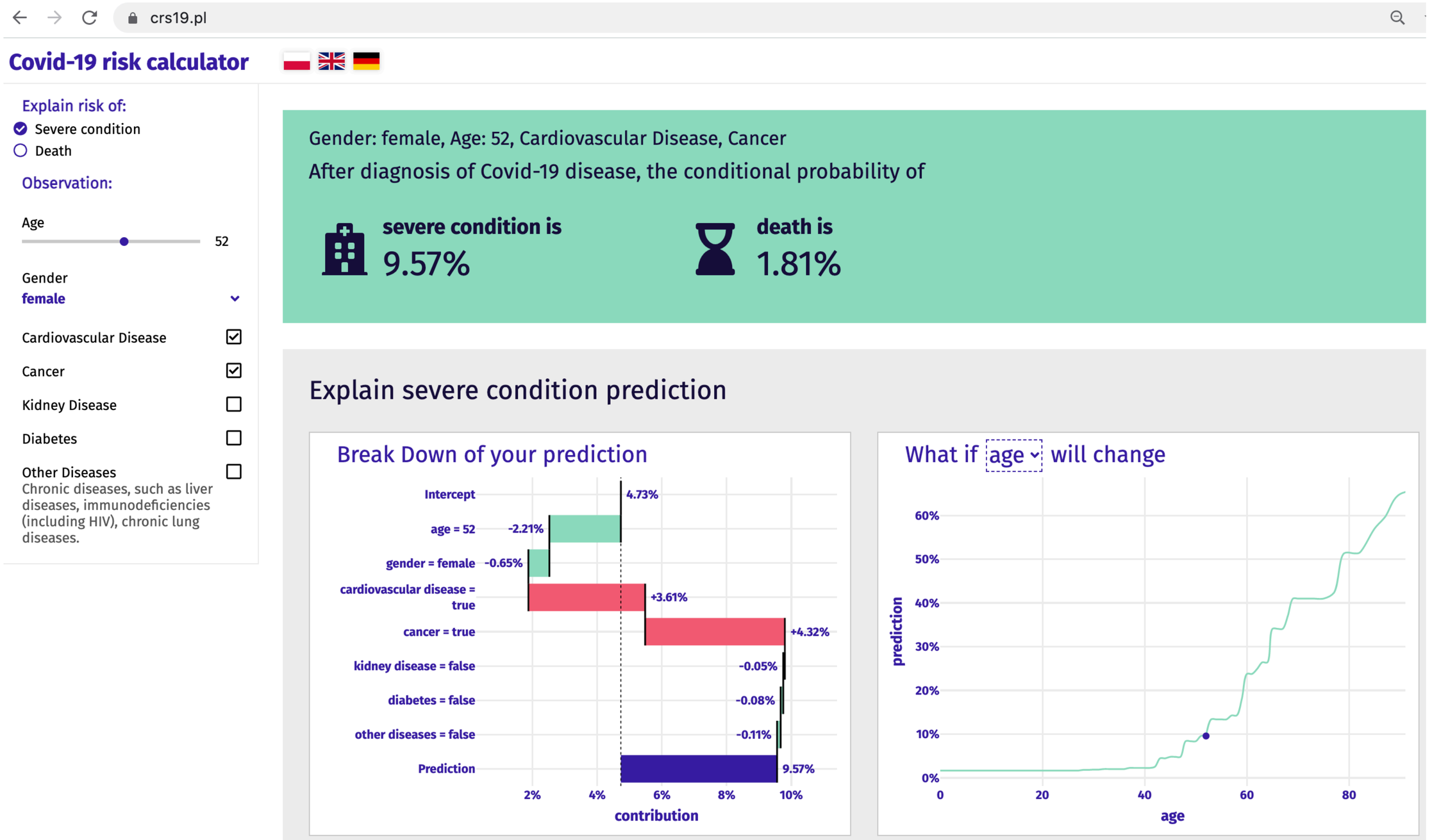
$$
\begin{aligned}
\mu &= \mathrm{E}\left[f(X)\right], \\
\mu_{x_1} &= \mathrm{E}\left[f(X)|X_1 = x_1^*\right], \\
\mu_{x_1,x_2} &= \mathrm{E}\left[f(X)|X_1 = x_1^*, X_2 = x_2^*\right], \\
&\ldots \\
\mu_{x_1,x_2,\ldots,x_p} &= \mathrm{E}\left[f(X)|X_1 = x_1^*, X_2 = x_2^*, \ldots, X_p = x_p^*\right] = f(x^*).
\end{aligned}
$$







What an example application looks like?



https://crs19.pl/

Explanatory Model Analysis
for predicting mortality for persons infected with SARS-COV-2 virus

A sample of more than 52,000 people with a positive PCR test for Covid-19.
Data from epidemiological surveillance of NIZP-PZH from November 2020.
Very detailed history (over 51 variables).

```
##                                      Stratified by Death
##                                       No             Yes
##   n                                    9487            513
##   Gender = Male (%)                    4554 (48.0)     271 (52.8)
##   Age (mean (SD))                     44.19 (18.32)  74.44 (13.27)
##   Cardiovascular.Diseases = Yes (%)    839 ( 8.8)     273 (53.2)
##   Diabetes = Yes (%)                   260 ( 2.7)      78 (15.2)
##   Neurological.Diseases = Yes (%)      127 ( 1.3)      57 (11.1)
##   Kidney.Diseases = Yes (%)            111 ( 1.2)      62 (12.1)
##   Cancer = Yes (%)                     158 ( 1.7)      68 (13.3)
##   Hospitalization = Yes (%)           2344 (24.7)     481 (93.8)
##   Fever = Yes (%)                     3314 (34.9)     335 (65.3)
##   Cough = Yes (%)                     3062 (32.3)     253 (49.3)
##   Weakness = Yes (%)                  2282 (24.1)     196 (38.2)
```

CDC model

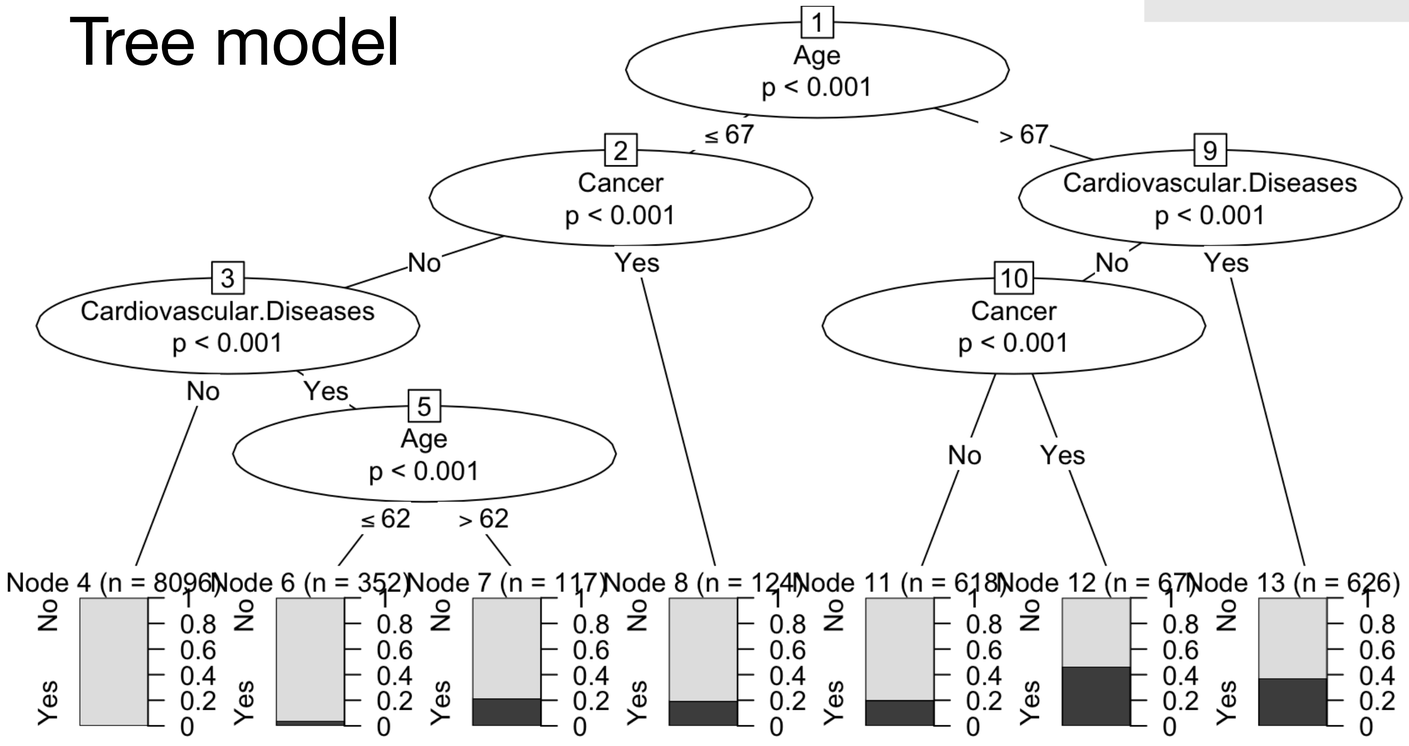
Risk for COVID-19 Infection, Hospitalization, and Death By Age Group

Updated Feb. 18, 2021 Print

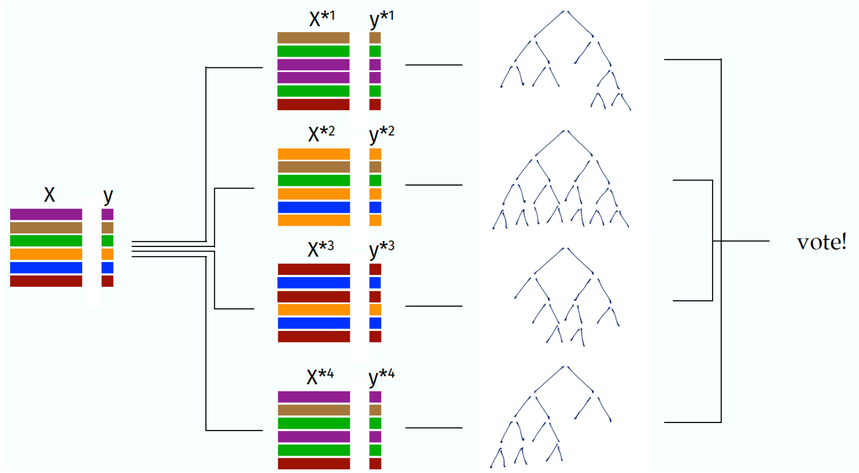
Rate ratios compared to 5-17 year olds[1]

	0—4 years old	5—17 years old	18—29 years old	30—39 years old	40—49 years old	50—64 years old	65—74 years old	75—84 years old	85+ years old
Cases[2]	<1x	Reference group	2x	2x	2x	2x	1x	1x	2x
Hospitalization[3]	2x	Reference group	6x	10x	15x	25x	40x	65x	95x
Death[4]	2x	Reference group	10x	45x	130x	440x	1300x	3200x	8700x

Tree model



Random Forest model



Explore Variable Importance

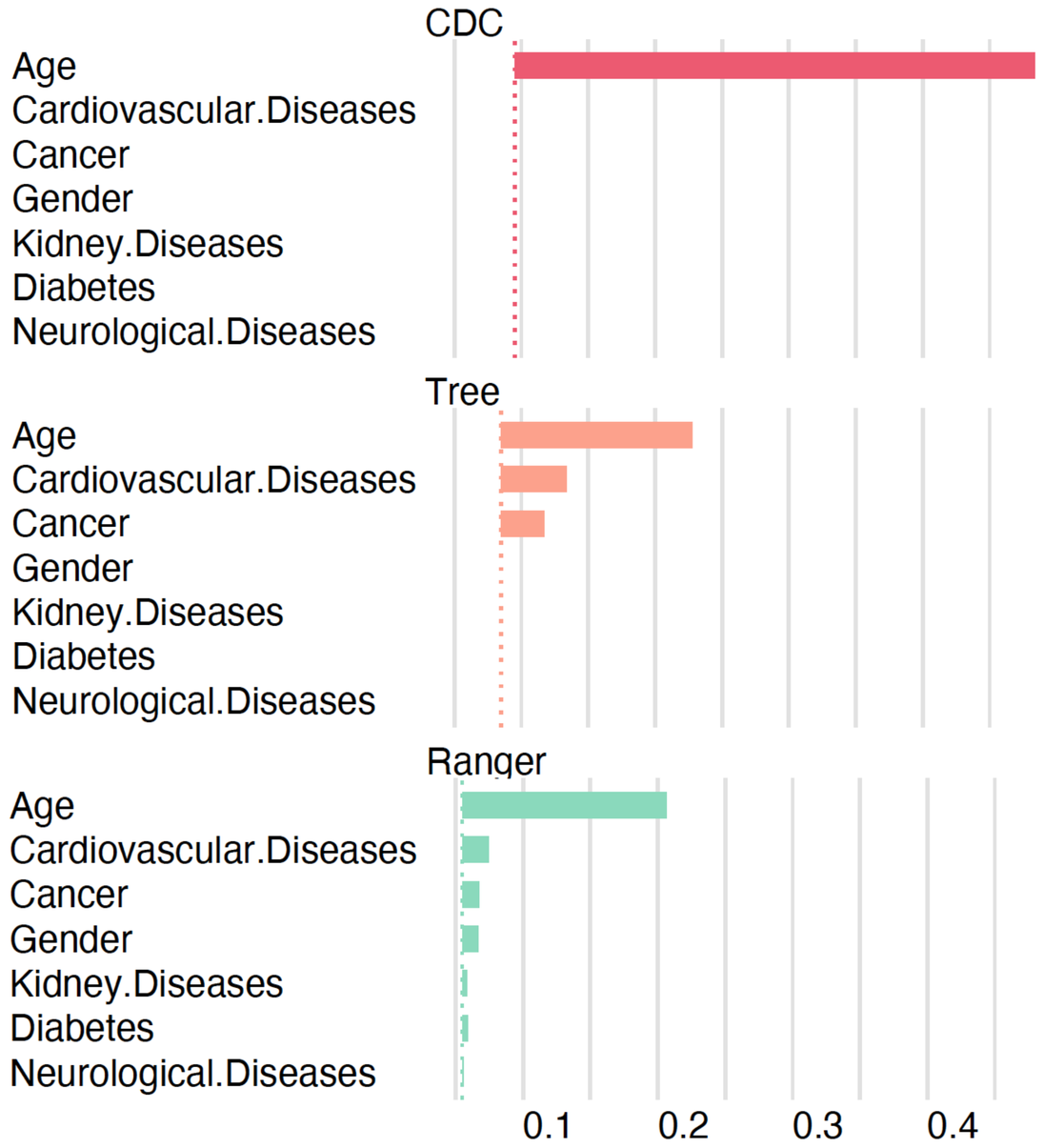
AUC
RMSE

How good is the model?

ROC curve
LIFT, Gain charts

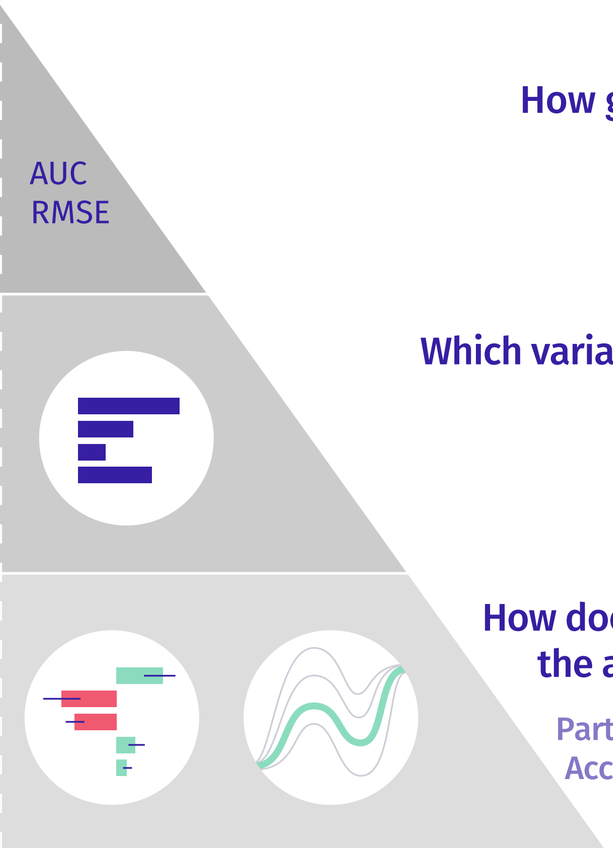
Which variables are important to the model?

Permutational
Variable Importance




```
fi_rms <- model_parts(exp_rms)

plot(fi_rms)
```

Explore Variable Effects



How good is the model?

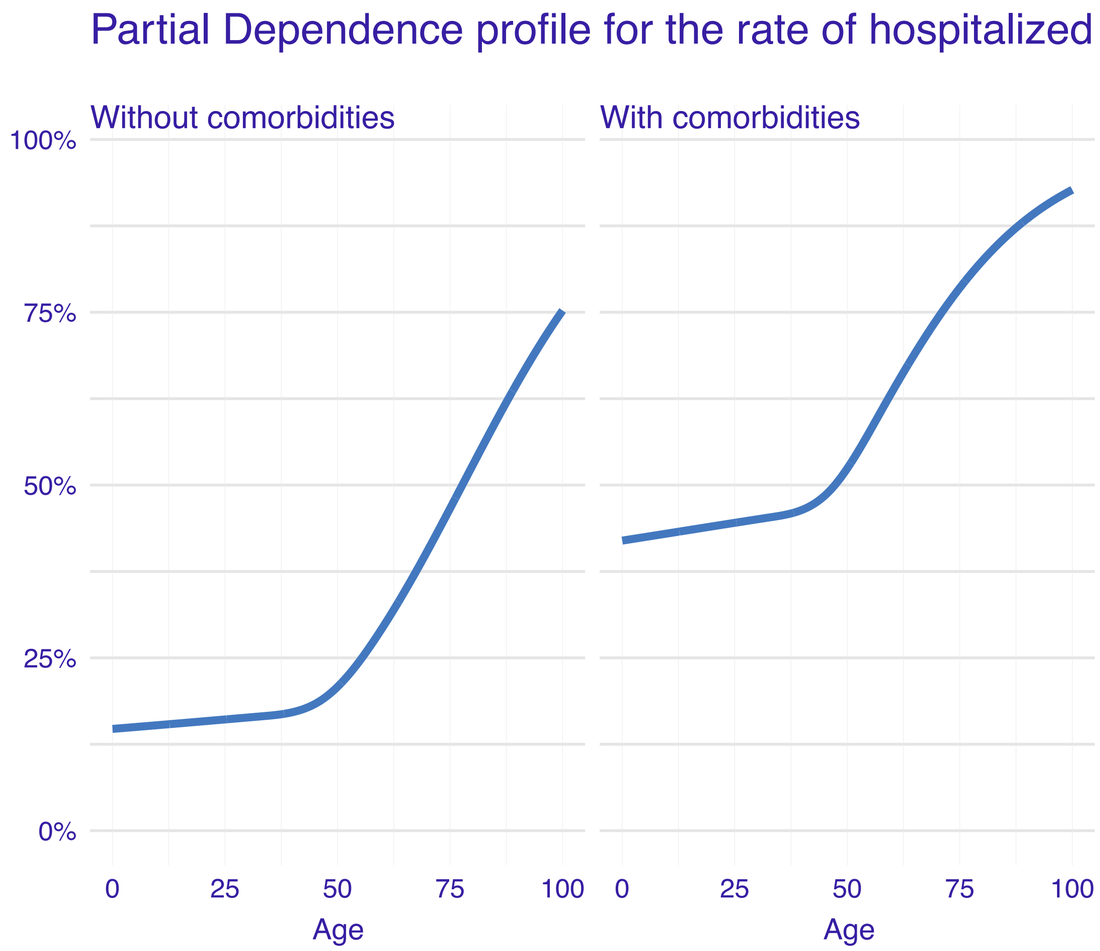
ROC curve
LIFT, Gain charts

Which variables are important to the model?

Permutational
Variable Importance

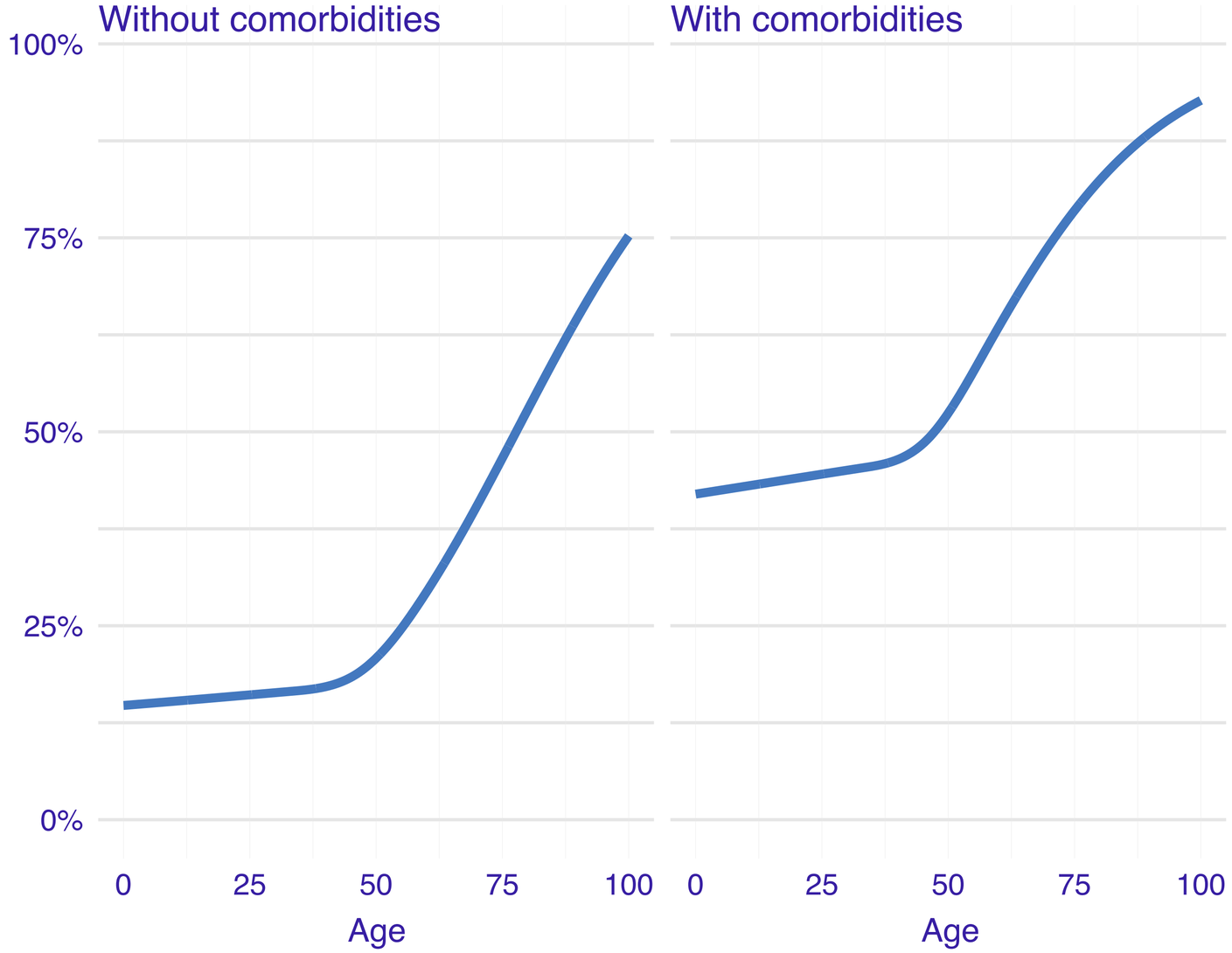
How does a variable affect the average prediction?

Partial Dependence Profile
Accumulated Local Effects



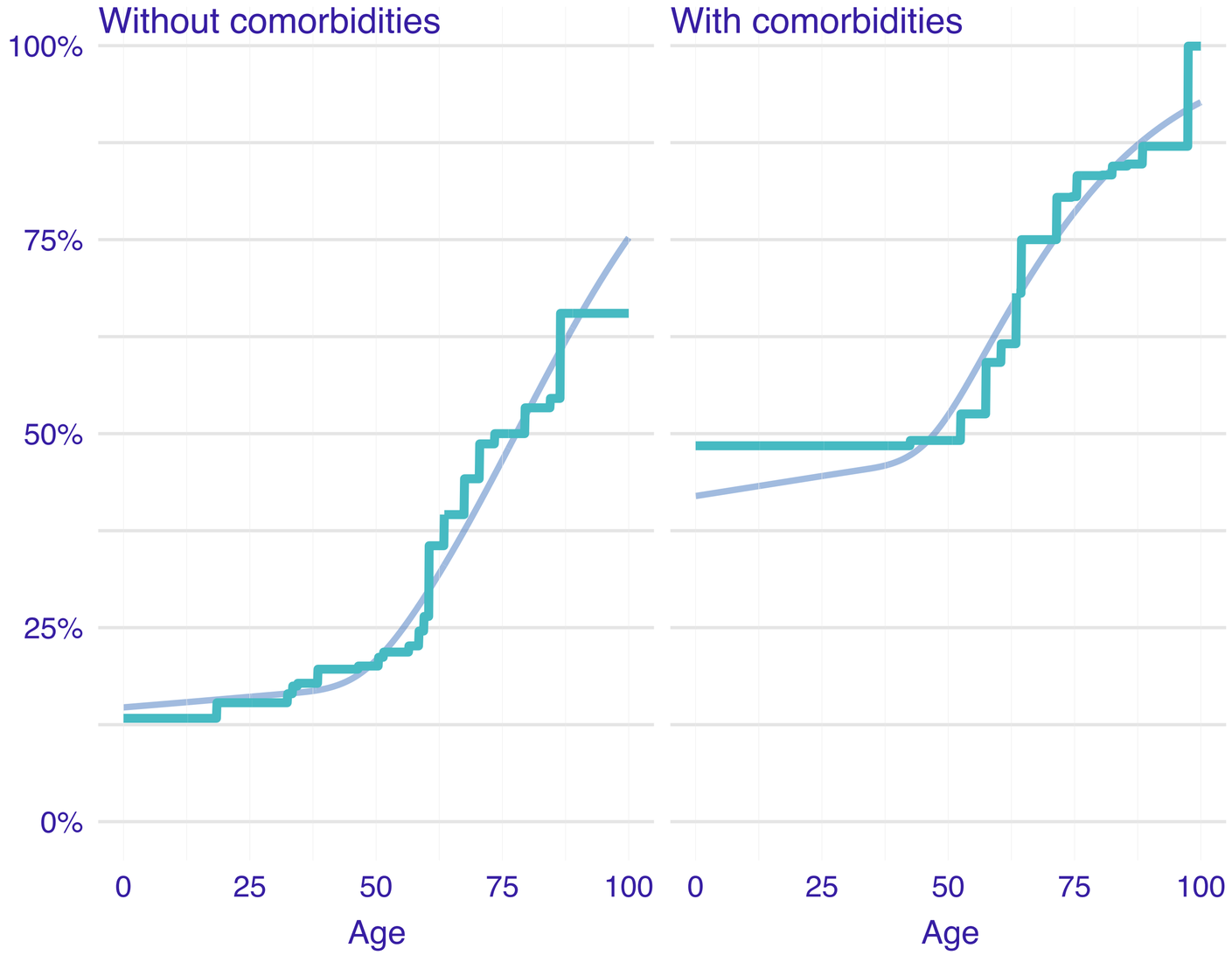

```
pdp_rms <- model_profile(exp_rms)

plot(pdp_rms)
```

Partial Dependence profile for the rate of hospitalized



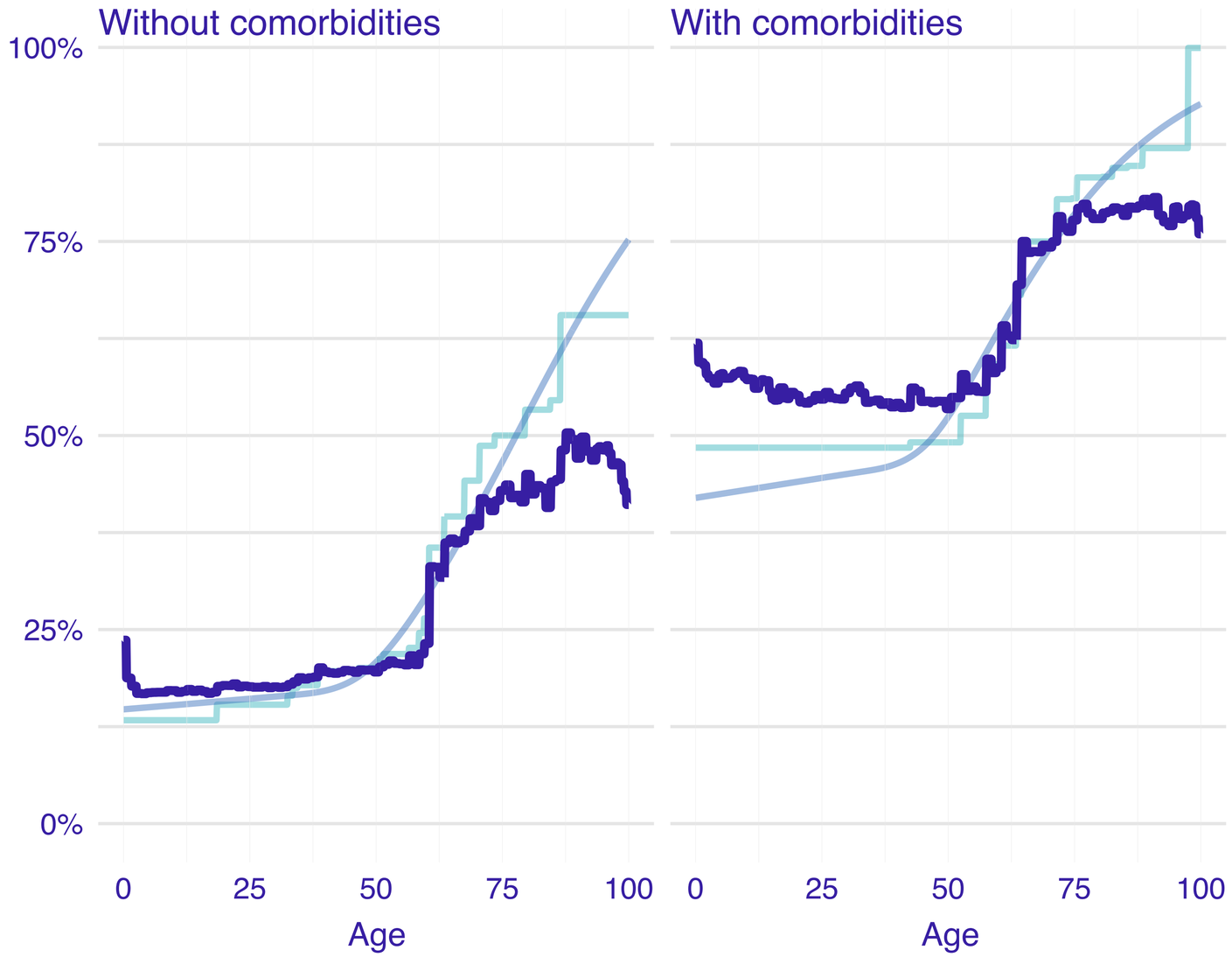
Models for Age+Comordibilities	AUC
Logistic regression with linear tail-restricted cubic spline	0.7490

Partial Dependence profile for the rate of hospitalized



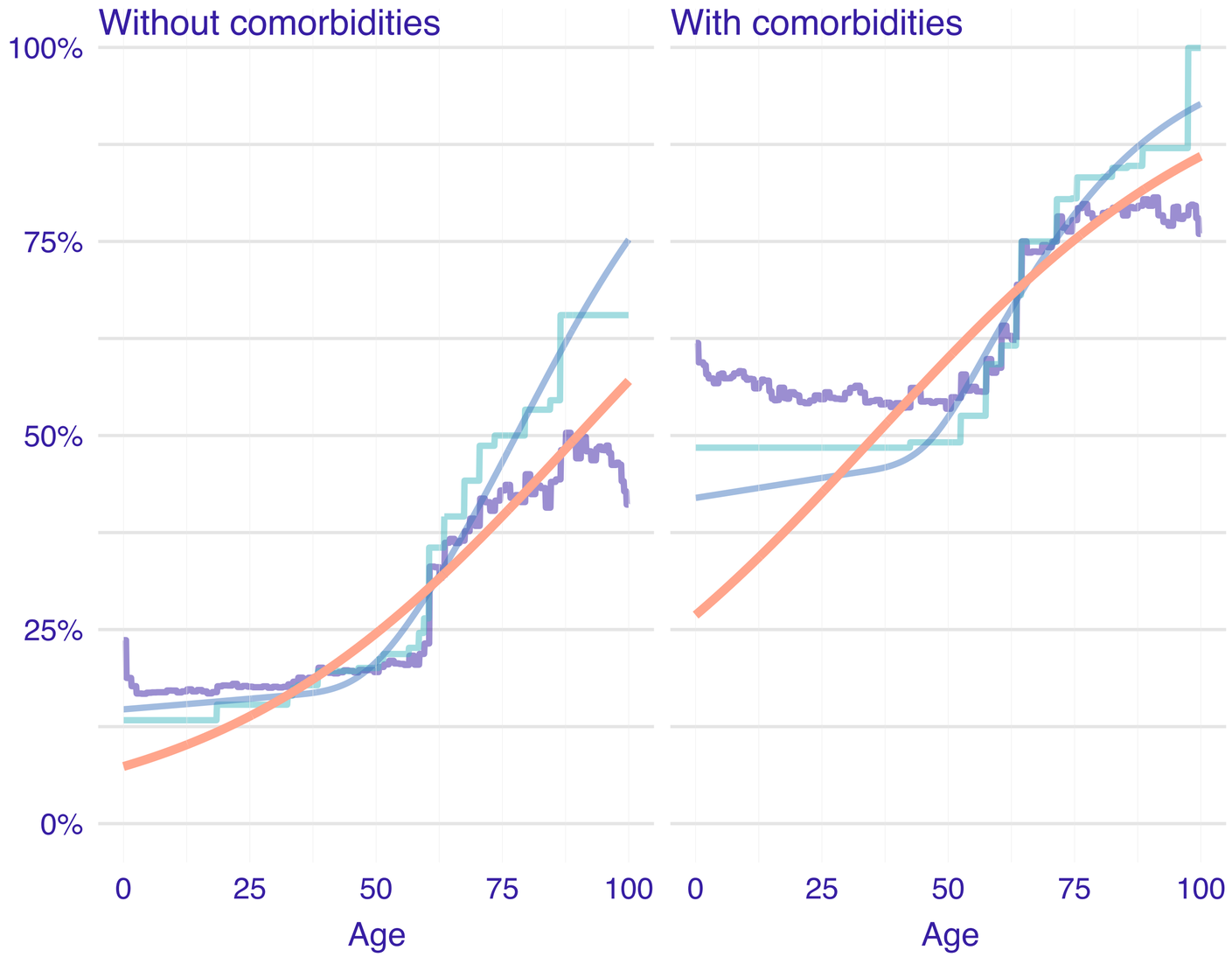
Models for Age+Comordibilities	AUC
Logistic regression with linear tail-restricted cubic spline	0.7490
Gradient boosting with monotonicity contraints	0.7629

Partial Dependence profile for the rate of hospitalized



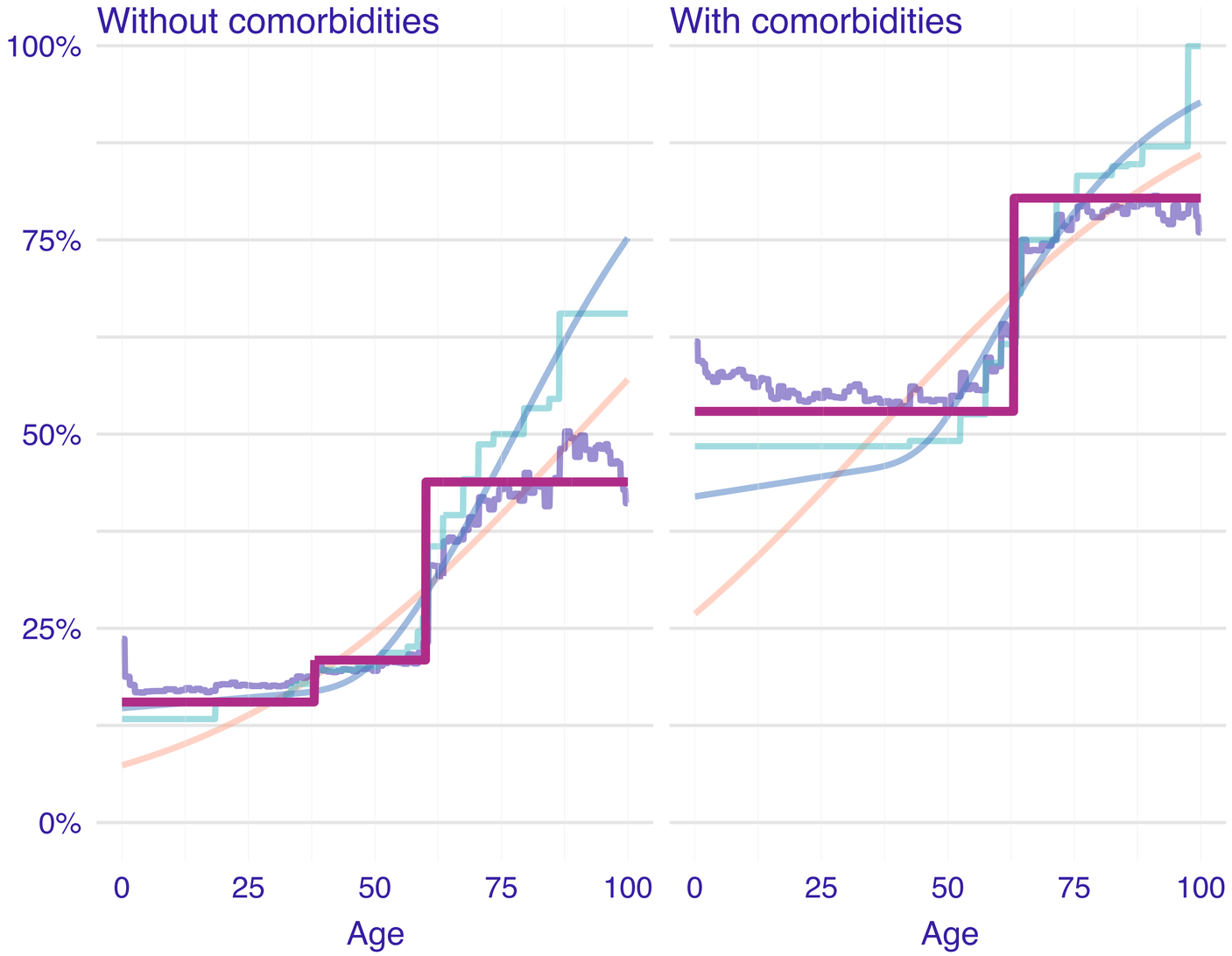
Models for Age+Comordibilities	AUC
Logistic regression with linear tail-restricted cubic spline	0.7490
Gradient boosting with monotonicity contraints	0.7629
Random forest	0.7568

Partial Dependence profile for the rate of hospitalized



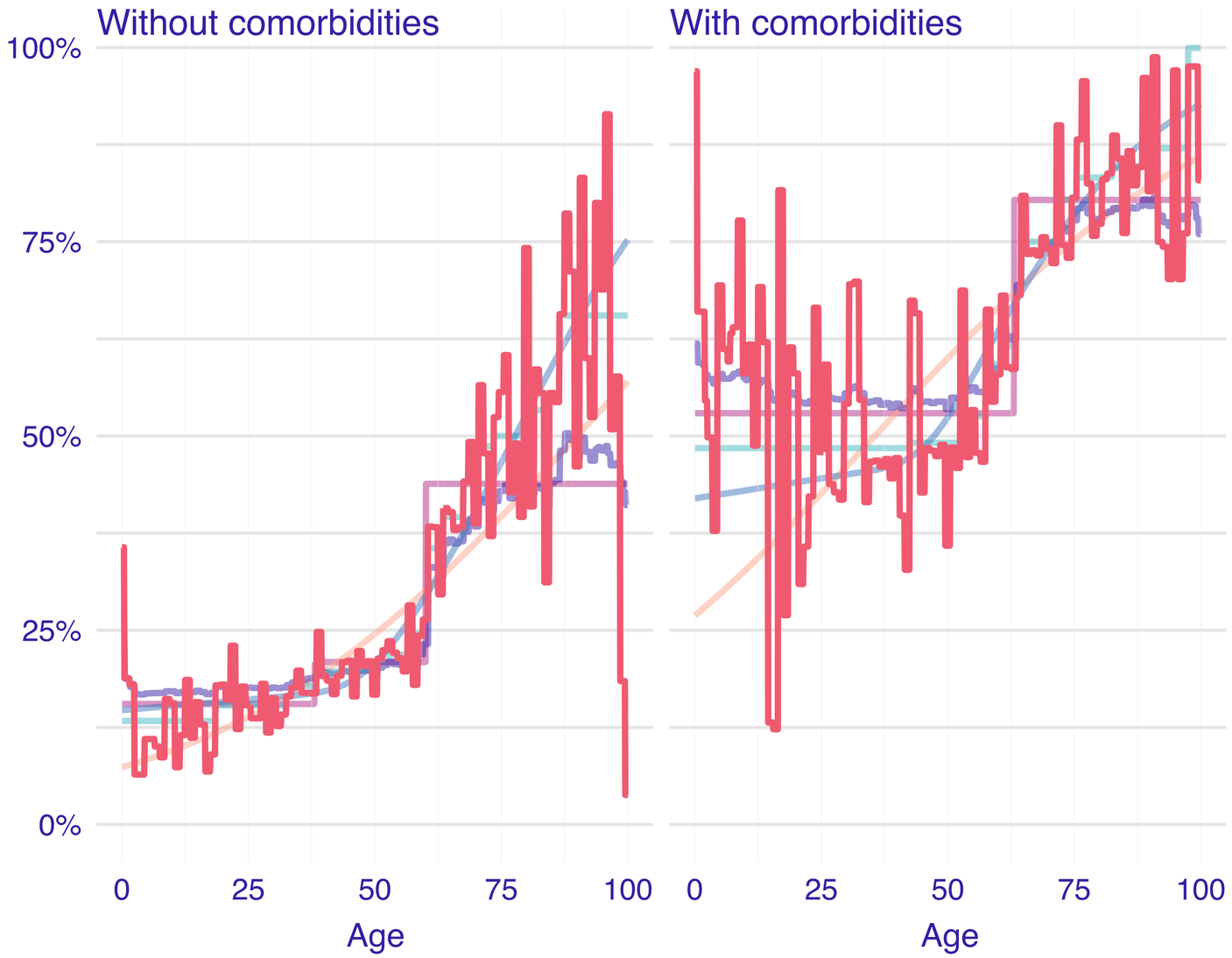
Models for Age+Comordibilities	AUC
Logistic regression with linear tail-restricted cubic spline	0.7490
Gradient boosting with monotonicity contraints	0.7629
Random forest	0.7568
Logistic regression	0.7484

Partial Dependence profile for the rate of hospitalized



Models for Age+Comordibilities	AUC
Logistic regression with linear tail-restricted cubic spline	0.7490
Gradient boosting with monotonicity contraints	0.7629
Random forest	0.7568
Logistic regression	0.7484
Classification tree	0.7403

Partial Dependence profile for the rate of hospitalized



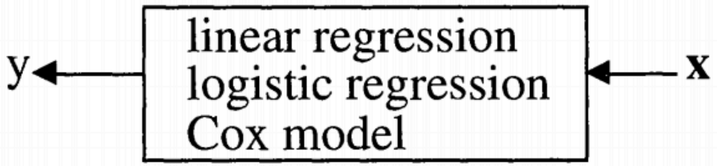
Models for Age+Comordibilities	AUC
Logistic regression with linear tail-restricted cubic spline	0.7490
Gradient boosting with monotonicity contraints	0.7629
Random forest	0.7568
Logistic regression	0.7484
Classification tree	0.7403
Gradient boosting	0.7597

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

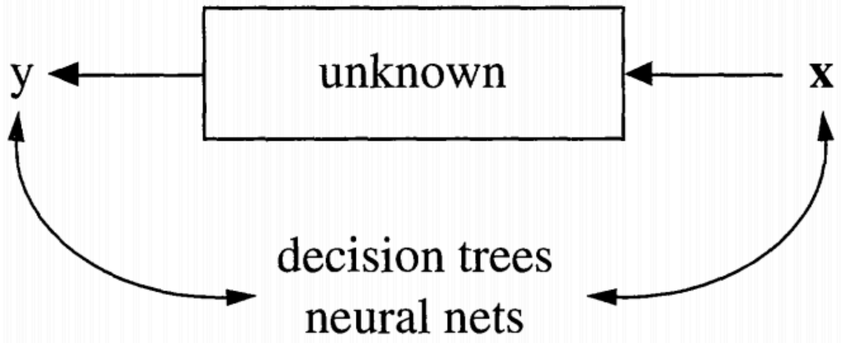
Leo Breiman

The Data Modeling Culture



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

8. RASHOMON AND THE MULTIPLICITY OF GOOD MODELS

Rashomon is a wonderful Japanese movie in which four people, from different vantage points, witness an incident in which one person dies and another is supposedly raped. When they come to testify in court, they all report the same facts, but their stories of what happened are very different.

What I call the Rashomon Effect is that there is often a multitude of different descriptions [equations $f(\mathbf{x})$] in a class of functions giving about the same minimum error rate. The most easily understood example is subset selection in linear regression. Suppose there are 30 variables and we want to find the best five variable linear regressions. There are about 140,000 five-variable subsets in competition. Usually we pick the one with the lowest residual sum-of-squares (RSS), or, if there is a test set,

Picture 1

$$y = 2.1 + 3.8x_3 - 0.6x_8 + 83.2x_{12} - 2.1x_{17} + 3.2x_{27},$$

Picture 2

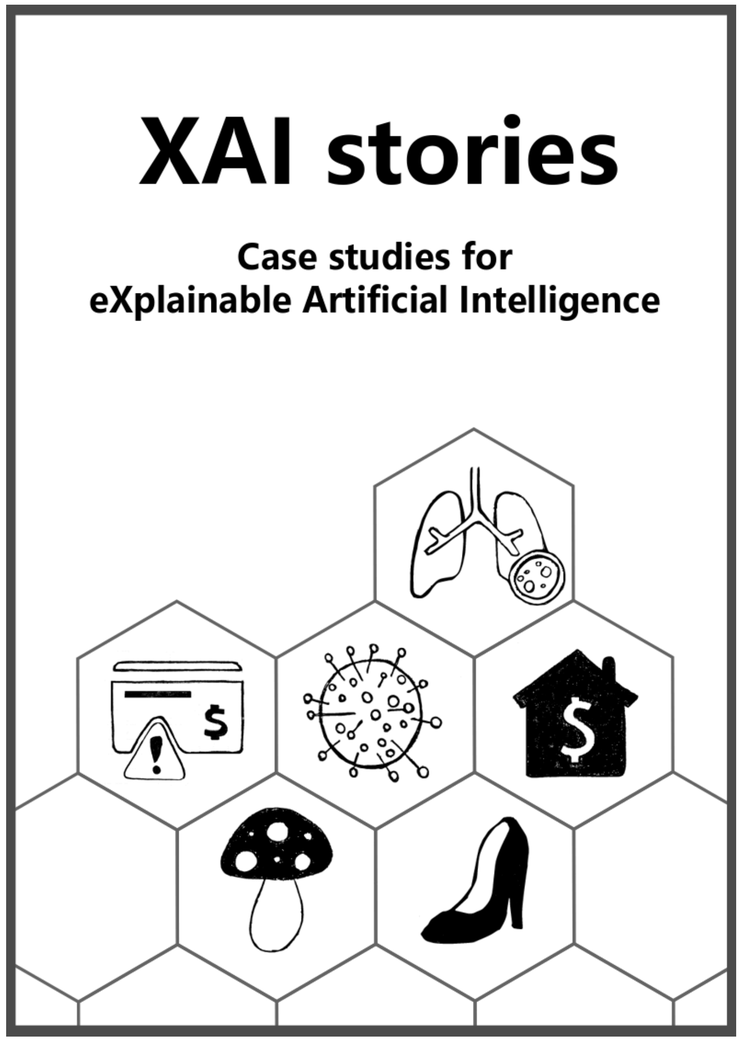
$$y = -8.9 + 4.6x_5 + 0.01x_6 + 12.0x_{15} + 17.5x_{21} + 0.2x_{22},$$

Picture 3

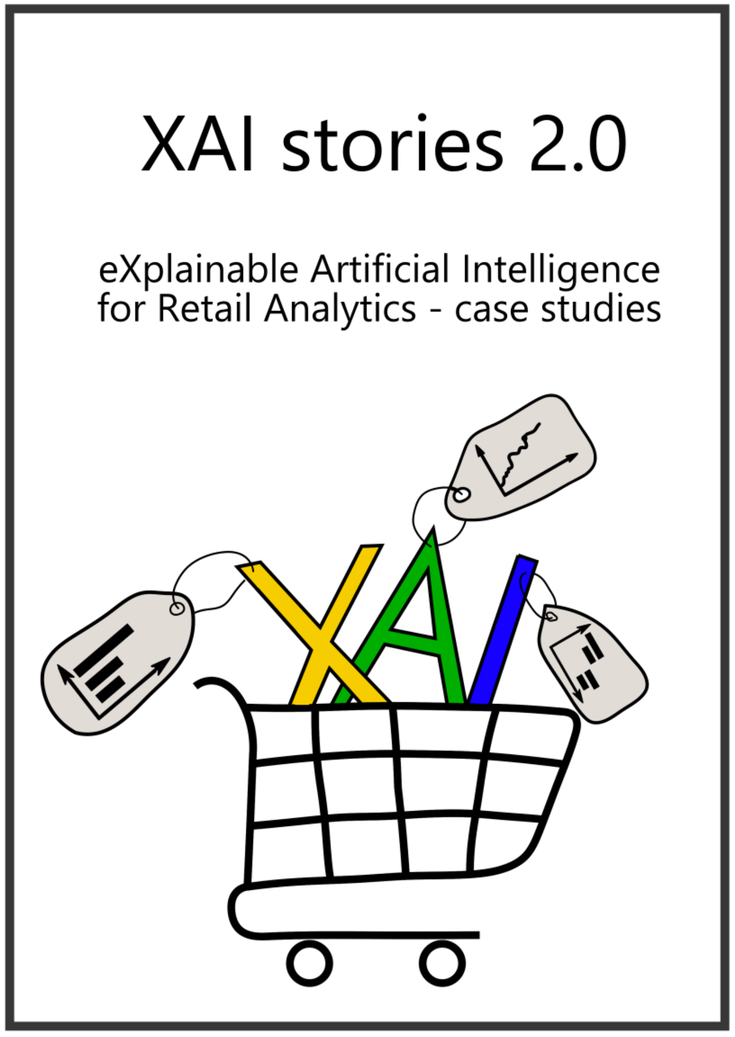
$$y = -76.7 + 9.3x_2 + 22.0x_7 - 13.2x_8 + 3.4x_{11} + 7.2x_{28}.$$

Which one is better? The problem is that each one tells a different story about which variables are important.

The Two Cultures, Leo Breiman



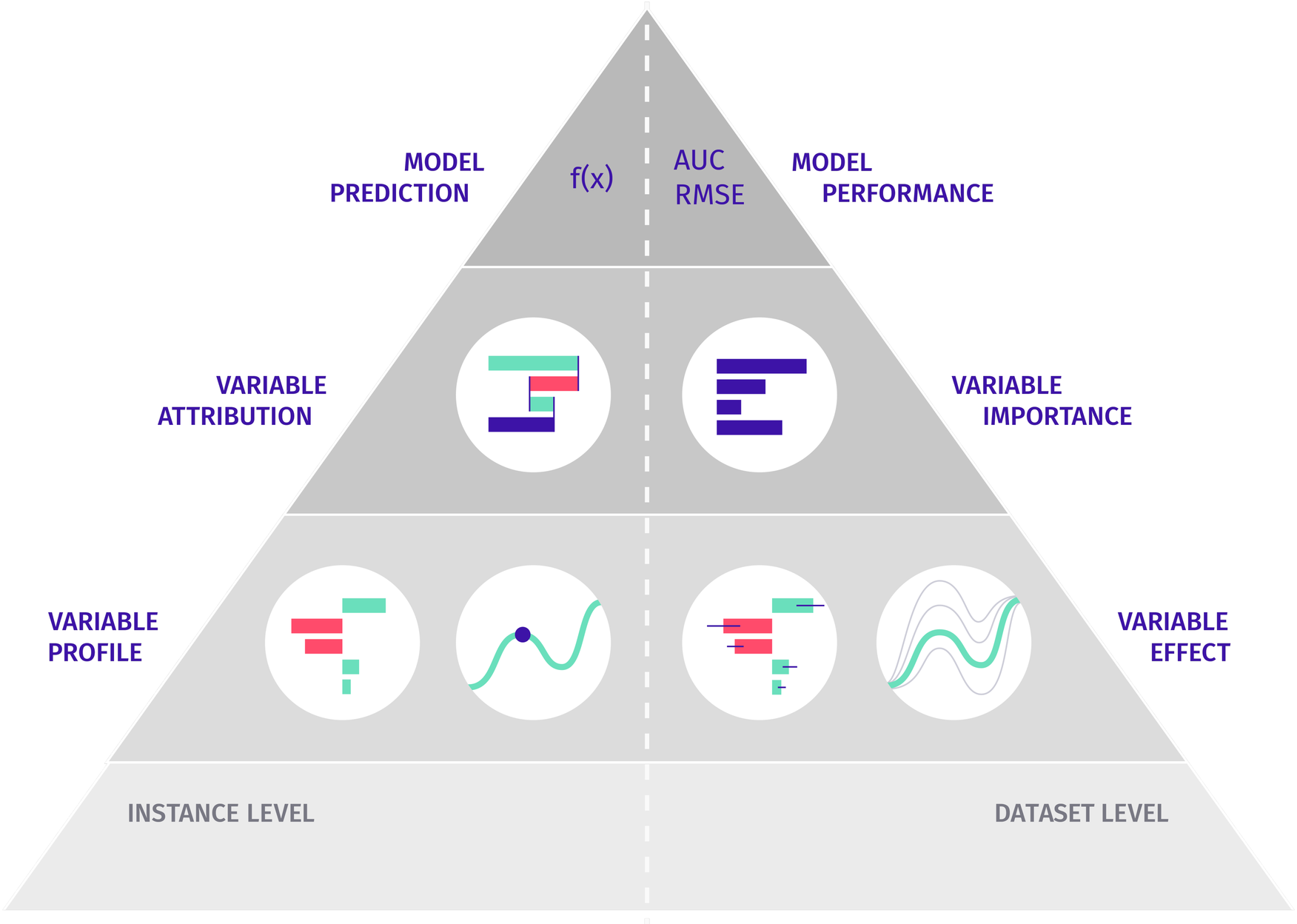
https://pbiecek.github.io/xai_stories/



https://pbiecek.github.io/xai_stories_2/

How interaction helps with EMA?

Model Exploration Stack



MODEL PERFORMANCE

AUC

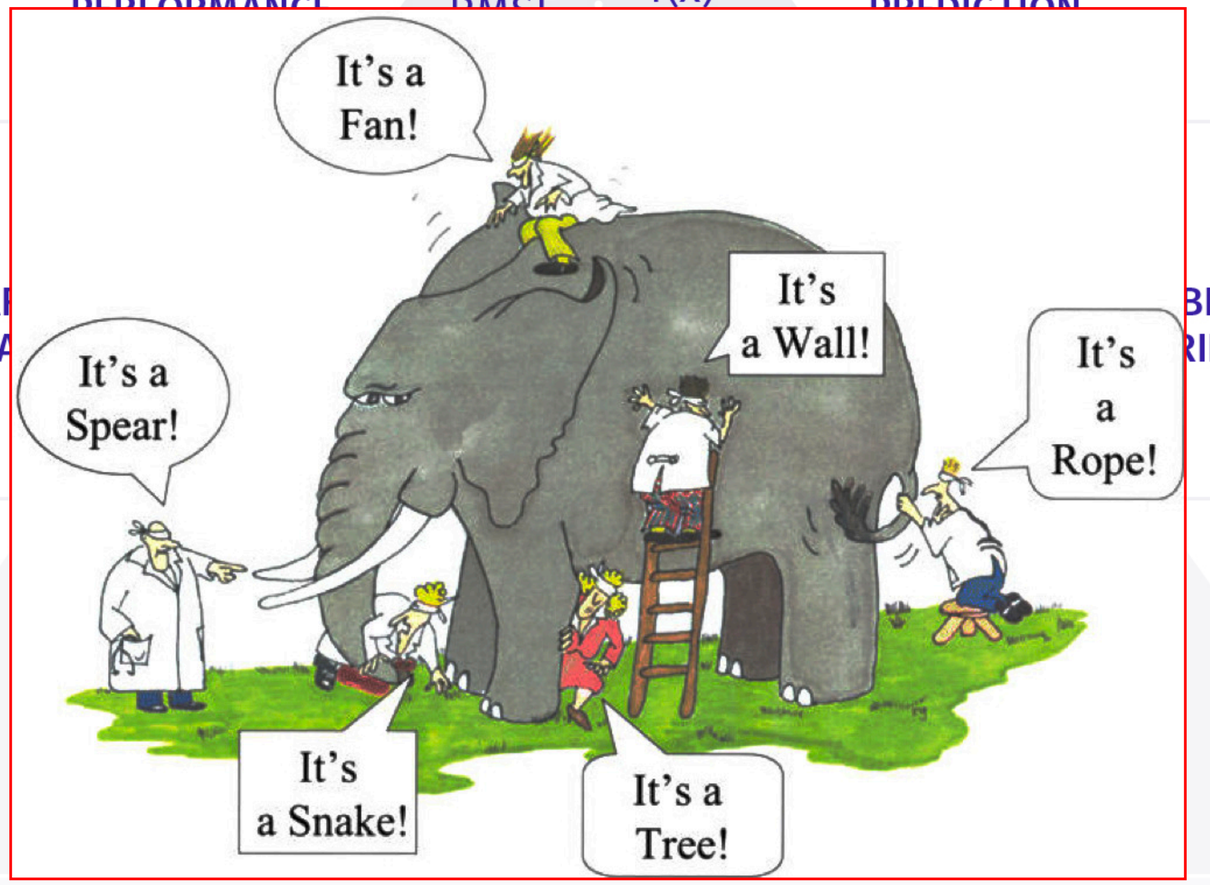
RMSE

f(x)

MODEL PREDICTION

VARIABLE IMPORTANCE

VARIABLE ATTRIBUTION



VARIABLE EFFECT

VARIABLE PROFILE

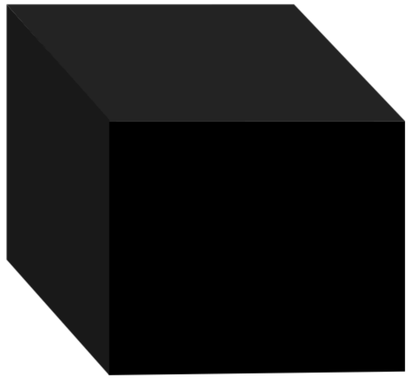
MODEL LEVEL

INSTANCE LEVEL

https://en.wikipedia.org/wiki/Blind_men_and_an_elephant

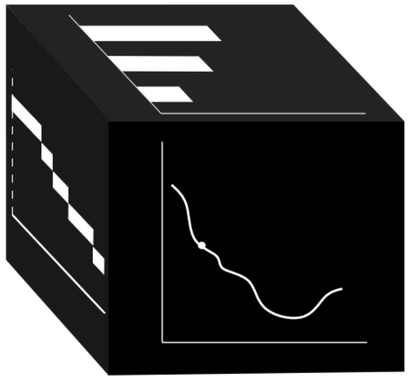
Interactive Explanatory Model Analysis that Opens Black-Box AI

Black Box Model



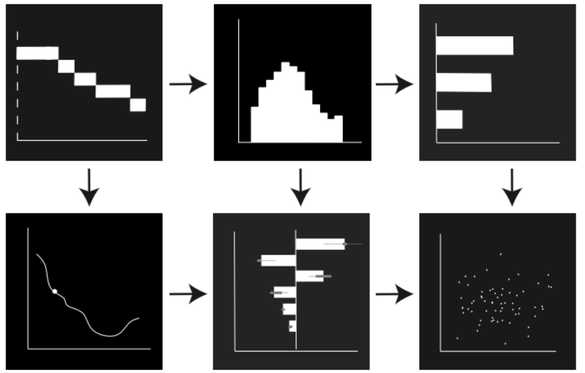
I generation explanations

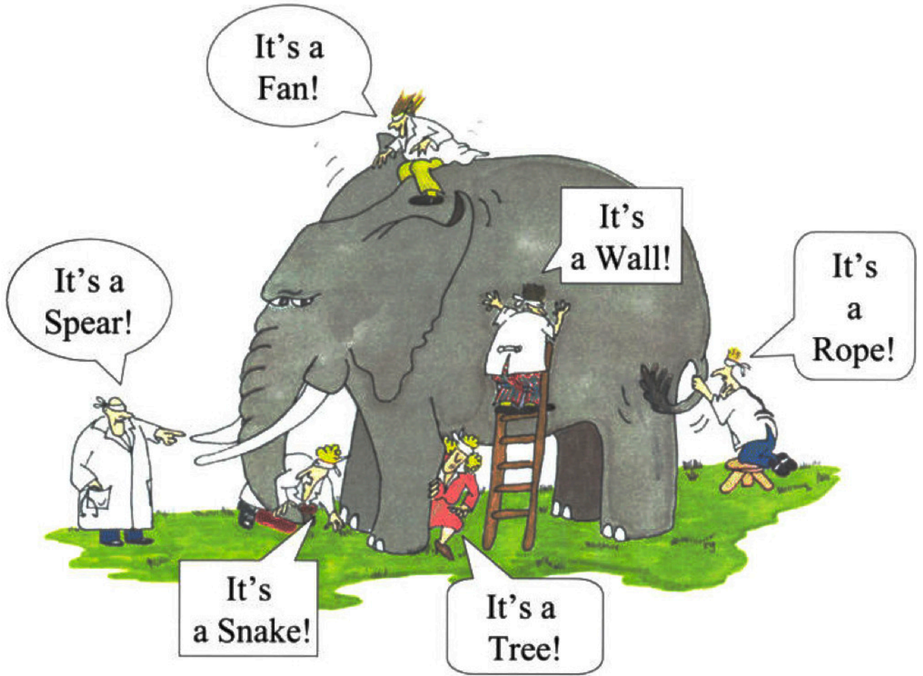
(single aspect
model exploration)



II generation explanations

(interactive explanatory
model analysis)

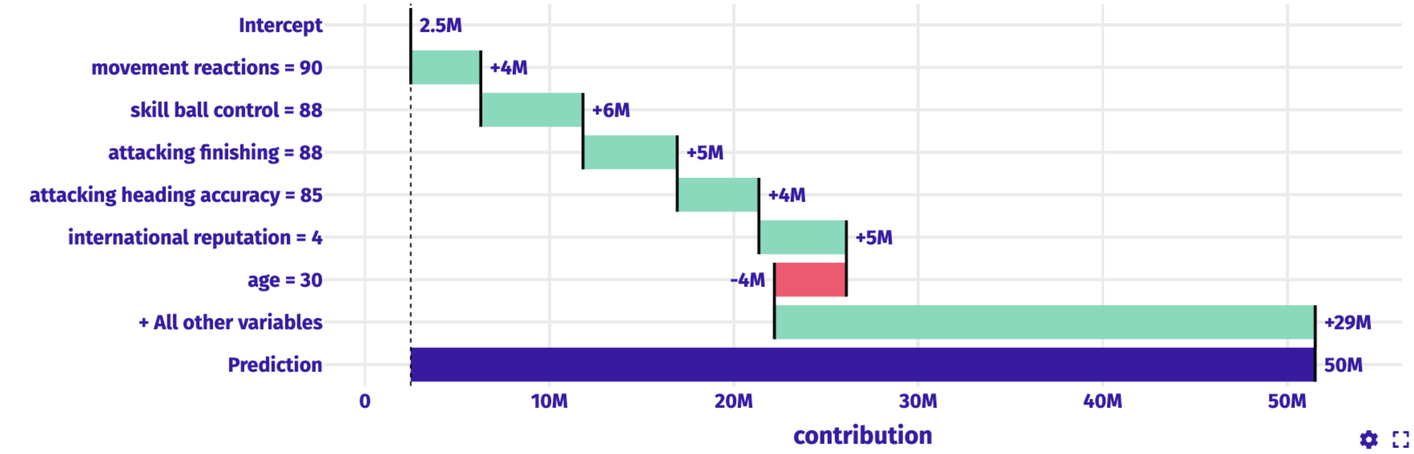




gbm

Break Down

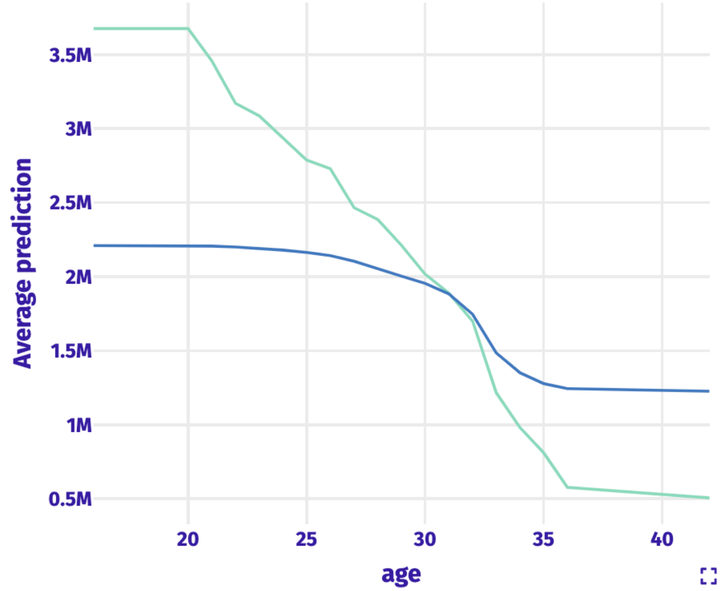
observation



gbm, ranger

Partial Dependence

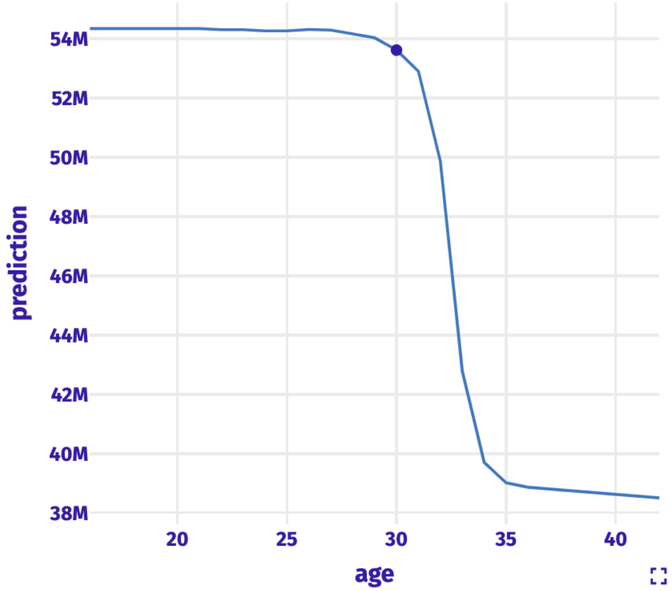
Split
variable



ranger

Ceteris Paribus

variable
observation



Models

- gbm
- lm
- ranger

Plots

Dataset Level

- Partial Dependence
- Accumulated Dependence
- Variable Importance

Model Performance

- Regression Error Characteristic
- Metrics
- Subsets Performance
- Funnel Plot

Observation Level

- Break Down
- Shapley Values

Models | Datasets | Other | Plots | Observation Details | Options

Explanatory Model Analysis

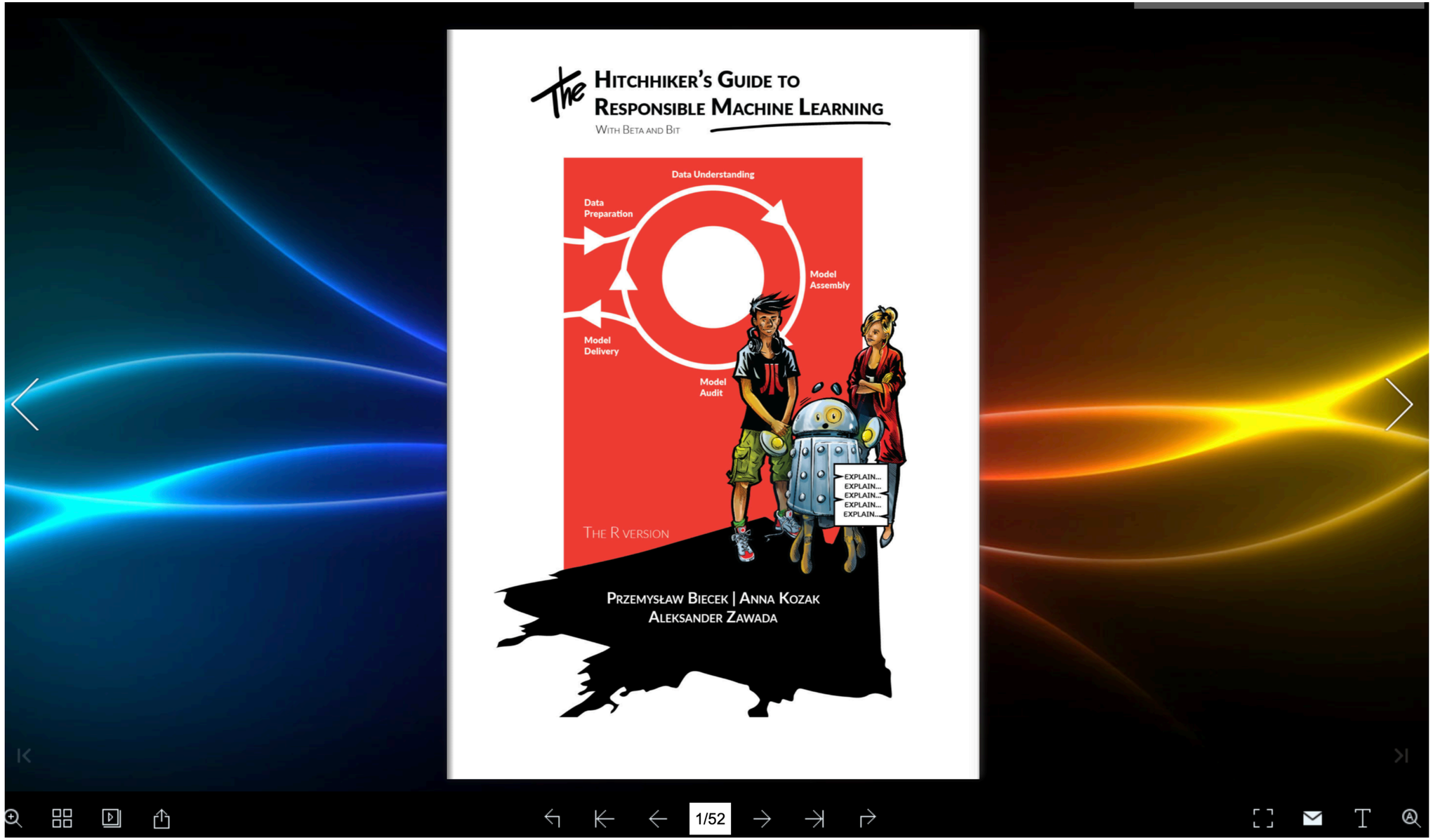
Explore, Explain, and Examine Predictive Models. With examples in R and Python.

Przemyslaw Biecek and Tomasz Burzykowski

2020-12-12

https://ema.drwhy.ai/

Preface



https://betaandbit.github.io/RML/

Thank you!

Questions?

Przemysław Biecek
/ˈpʂɛ.mɛk/

https://www.linkedin.com/in/pbiecek/

Special thanks go to:

Hubert Baniecki (modelStudio)
Ewa Baranowska (drifter)
Alicja Gosiewska (auditor)
Aleksandra Grudziąż (survxai)
Adam Izdebski (describe)
Ewelina Karbowiak (EIX)
Marcin Kosiński (archivist)
Ania Kozak (vivo)
Michał Kuźba (xaibot)
Szymon Maksymiuk (DALEXtra)
Magda Młynarczyk (cr17)
Aleksandra Paluszyńska (randomForestExplainer)
Kasia Pękała (triplot)
Piotr Piątyszek (Arena)
Hanna Piotrowska (DrWhy theme)
Adam Rydelek (xai2cloud)
Agnieszka Sitko (factorMerger)
Jakub Wiśniewski (fairModels)
... and others from MI2DataLab