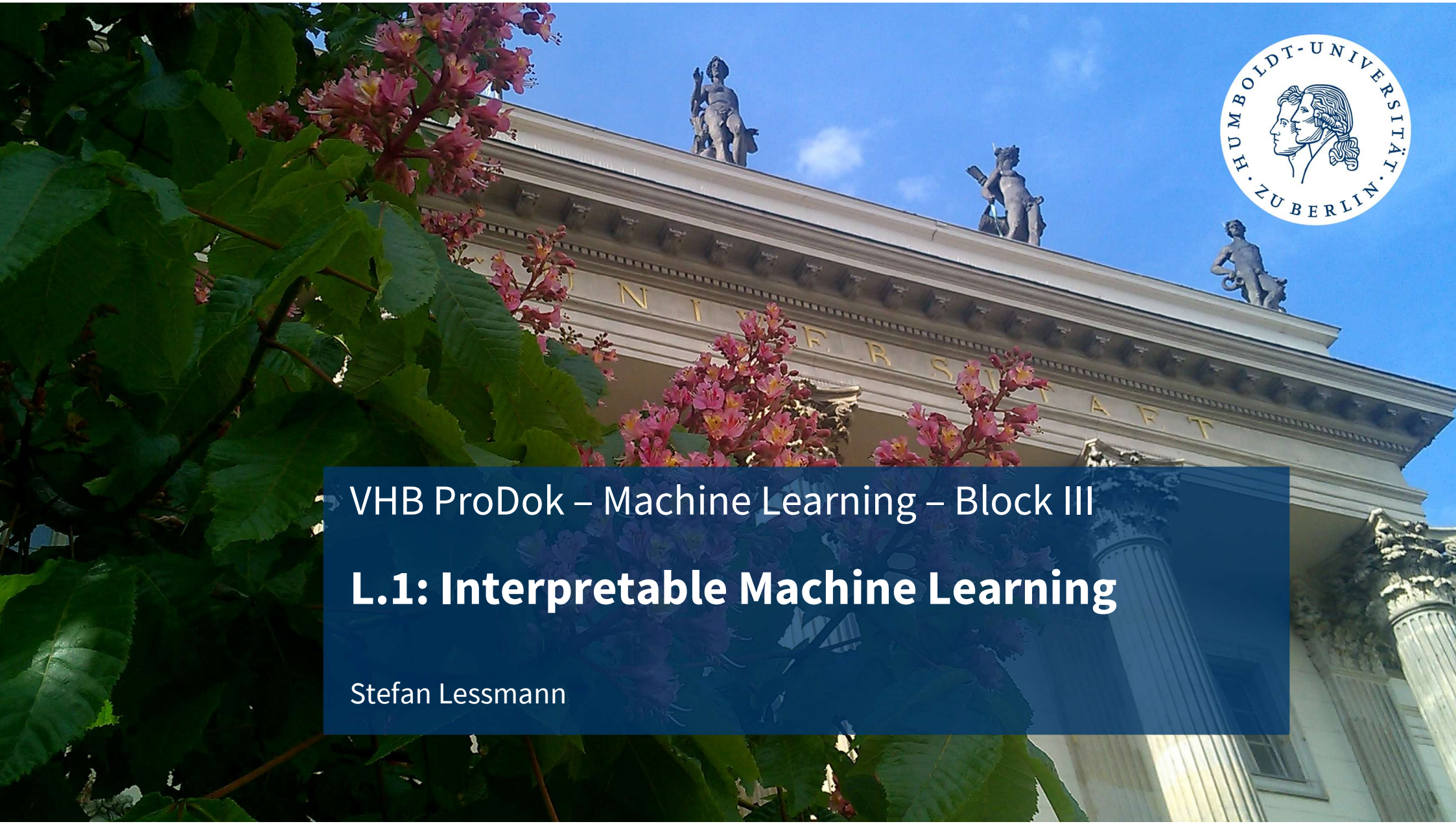
VHB ProDok – Machine Learning

Block III: Selected Topics in Machine Learning Research

Stefan Lessmann

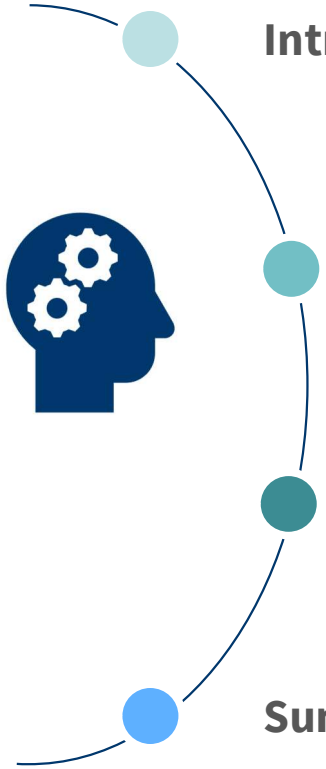


VHB ProDok – Machine Learning – Block III

L.1: Interpretable Machine Learning

Stefan Lessmann

Agenda



Introduction

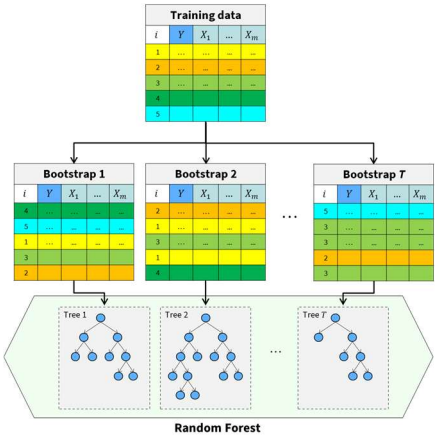
Global Explanation Methods
Surrogate models, permutation importance, and partial dependence

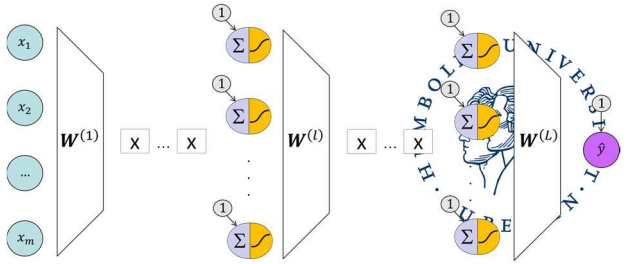
Local and Example Based Explanations
LIME, SHAP values, counterfactual and adversarial examples

Summary

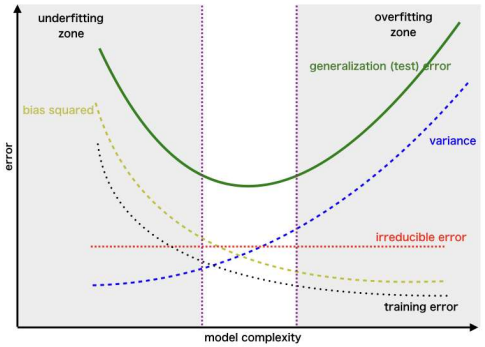
Introduction

Where We Stand Today





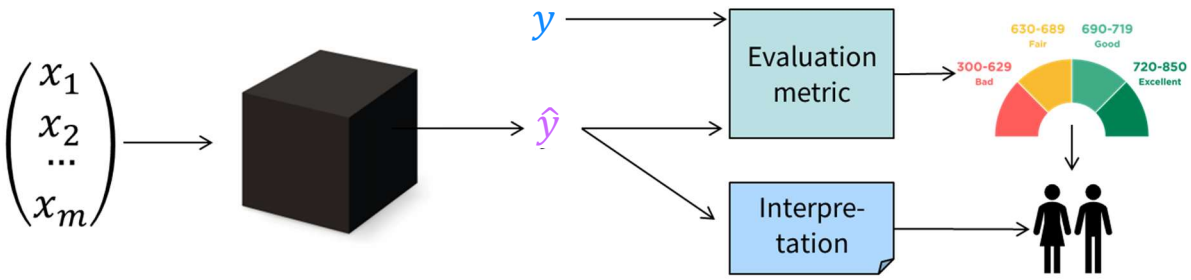
- **Various advanced ML algorithms**
  - Tree-based ensembles, Random Forest, XGB
  - Neural networks and deep learning
- **Advanced learners promise improved forecast accuracy**
  - Capture nonlinear dependencies, feature interactions, etc.
  - Bias-variance trade-off provides theory-grounding
    - Start from a model with low bias
    - Reduce complexity to control overfitting
- **Modeling practices**
  - Split sampling & cross-validation
  - Hyperparameter tuning & model selection



But is accuracy enough?

Why We Need XAI

- **ML&AI increasingly used in mission-critical applications**
- **Measures of (predictive) performance provide an incomplete picture**
- **Gain insight into model-inferred patterns to identify ways to improve**
- **Confirm model acts in a manner that agrees with domain knowledge**
- **Auditing and robustness**
- **Comply with legal requirements and codes of conduct**
  - Explain algorithmic decisions (e.g. EU GDPR Recital 71)
  - Verify that model outputs are not biased (e.g. do not discriminate against protected groups)



Lipton (2016)

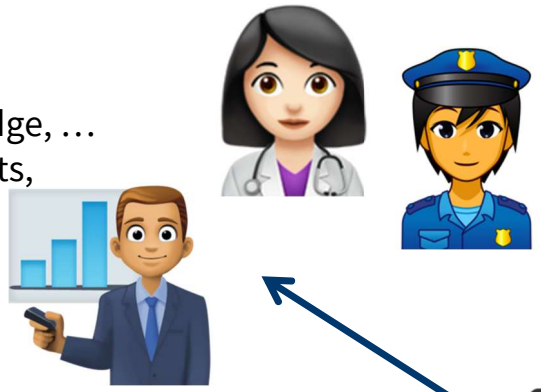
What Are We Talking About?

Transparency, explainability, and interpretability of machine learning models

- **Differences are often subtle**
- **Transparency is often understood as algorithmic transparency**
  - What are the inner mechanics of a learning algorithm?
  - How does a learning algorithm craft a model
- **Some use interpretability and explainability interchangeably (e.g., Molnar, 2023)**
- **When distinguishing interpretability and explainability**
  - Many sources associate interpretability with a solid understanding of the inner workings of a model
  - Explainability then refers to the degree to which humans can understand model outputs and how they follow from model inputs (see, e.g., this Amazon whitepaper)
    - Note that this understanding of interpretability goes in the direction of transparency
    - Whereby we could still distinguish between the interpretability of a model and the transparency of an algorithm
- **Exemplary definitions**
  - Miller (2017) *Interpretability is the degree to which a human can understand the cause of a decision.*
  - Kim et al. (2016) *Interpretability is the degree to which a human can consistently predict the model's result.*

Stakeholders and Their Perspectives

xAI outputs must match recipients‘ requirements and ML understanding



End-user
- Business unit, MD, Judge, …
- Trust, gain new insights, justifiability, impli-cations

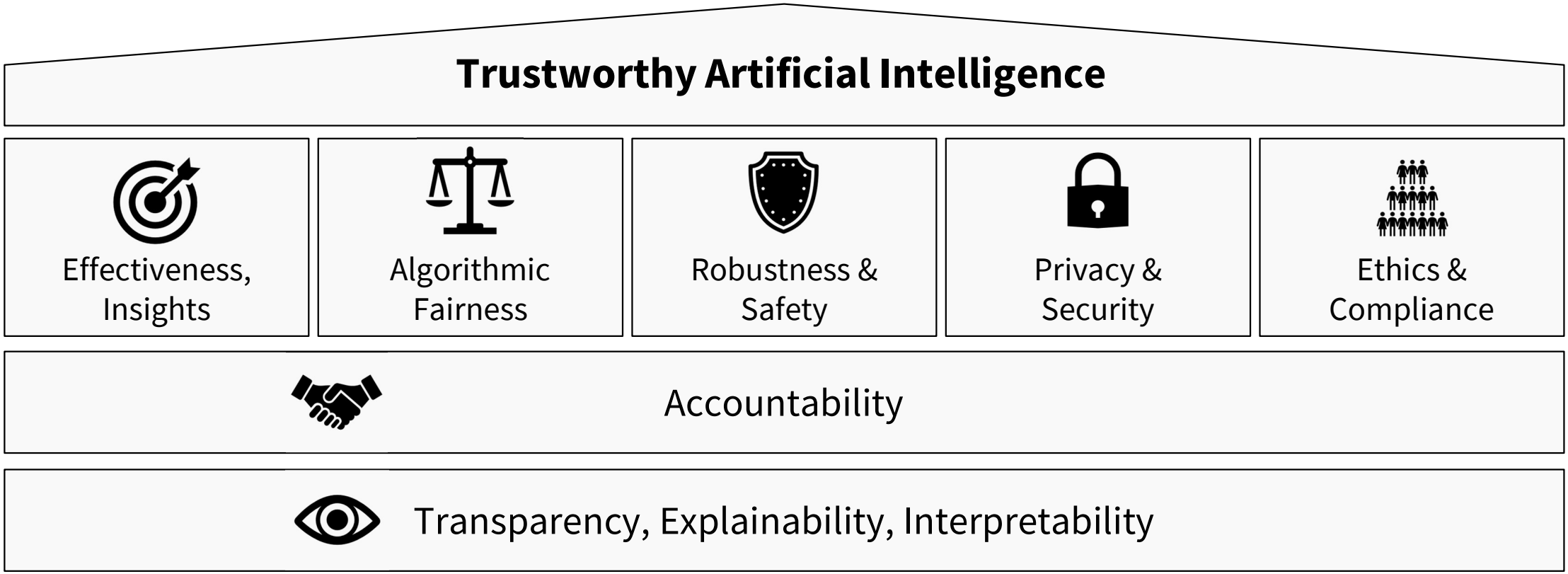
Law-/policy-maker
- EU, GDPR, financial surveillance, …
- Fairness, compliance, auditing

Developer
- Business analyst, data scientist, ML engineer …
- Ensure model quality, improve model

Affected subjects
- Clients, applications, defandants, …
- Grasp case, verify accuracy & fairness of decisions

The Bigger Picture

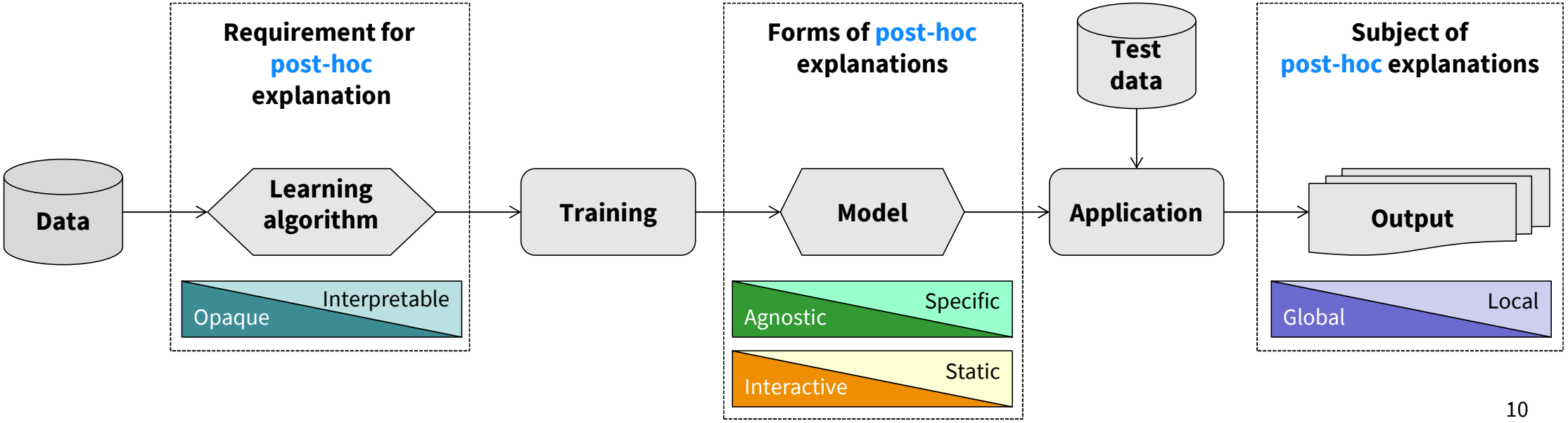
Explainable AI in context



Sources: Barredo et al. (2020); Doran et al. (2017); Doshi-Velez & Kim (2017); Langer et al. (2021)
https://www.research.ibm.com/artificial-intelligence/trusted-ai/

Scope of XAI and Interpretable Machine Learning

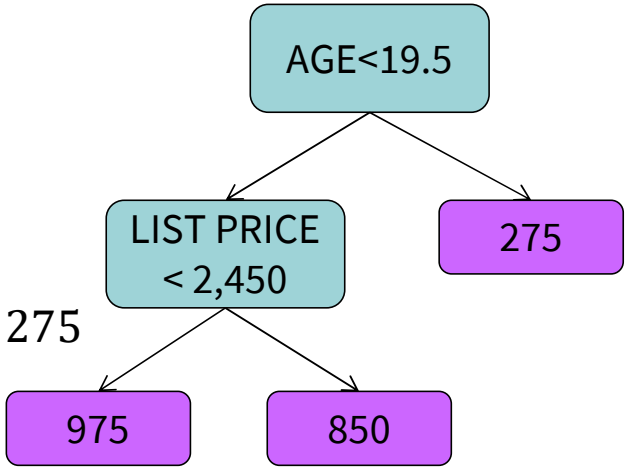
- **Several different approaches to *explain* opaque ML models**
  - Statistics and visualizations of feature importance
  - (Local) approximations using an intrinsically interpretable model
  - Illustrative examples and adjustments of cases‘ feature values that affect model outputs
- **An established stream of research concerns post-hoc explanation methods**



Intrinsically Interpretable Models

- **Linear models**
  - Linear and additive function of feature values and their weights
  - GLMs* consider transformations of the features and/or the target
    - Interpretability ranking: linear regression > GLMs > GAMs
    - $\hat{y}(\boldsymbol{x}) = \phi\big(E(y|\boldsymbol{x})\big) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p)$
  - Interpretability decreases with the number of features
- **Classification and regression trees**
  - Set of decision rules
  - $If\ AGE < 19.5\ Then\ (If\ LIST\ PRICE < 2{,}450\ Then\ 975\ Else\ 850)\ Else\ 275$
  - Interpretability decreases with depth
- **How these models are interpretable**
  - Understand how, in general, feature values cause model predictions
  - Understand how the forecast of a specific data instance emerges

$$\text{Resale price forecast} = bias + w_1 \text{LIST PRICE} + w_2 \text{AGE} + \cdots$$



Global Explanation Methods

Surrogate models, permutation importance, and partial dependence

Surrogate Models / Pedagogical Rule Extraction

Use a white-box model to explain a black-box model

i	PRODUCT	LIST PRICE	AGE	INDUSTRY	…	**RESALE PRICE [$]**
1	Dell XPS 15’	2,500	36	Mining	…	347
2	Dell XPS 17’	3,000	36	Health	…	538
3	HP Envy 17’	1,300	24	Office	…	121
4	HP EliteBook 850	1,900	36	Mining	…	172
5	Lenovo Yoga 13’	1,100	12	Office	…	266

$f^O(\boldsymbol{X})$

$\hat{Y}^O$
315
475
144
201
188

i	PRODUCT	LIST PRICE	AGE	INDUSTRY	…	$\hat{Y}^O$
1	Dell XPS 15’	2,500	36	Mining	…	315
2	Dell XPS 17’	3,000	36	Health	…	475
3	HP Envy 17’	1,300	24	Office	…	144
4	HP EliteBook 850	1,900	36	Mining	…	201
5	Lenovo Yoga 13’	1,100	12	Office	…	188

$$\hat{Y}^I = bias + w_1 \text{PRODUCT} + w_2 \text{LISTPRICE} + w_3 \text{AGE} + w_4 \text{INDUSTRY} + \cdots$$

1) Learn opaque model f^O and obtain forecasts

2) Form training set for intrinsically interpretable model f^I : Forecasts of opaque model $\hat{Y}^O$ become the target variable

3) Train f^I (e.g., linear regression) on the new training set

4) The fitted, interpretable model f^I approximates how the opaque model relates features to predictions.

Further information: Molnar (2023)

Multi-Stage Models

Combine interpretable white-box model with black-box model

i	Product	List price	Age	Industry	…	Resale price [$]
1	Dell XPS 15’	2,500	36	Mining	…	347
2	Dell XPS 17’	3,000	36	Health	…	538
3	HP Envy 17’	1,300	24	Office	…	121
4	HP EliteBook 850	1,900	36	Mining	…	172
5	Lenovo Yoga 13’	1,100	12	Office	…	266

1) Estimate white-box model

$$Y = f^{WB}(X) + \epsilon^{WB}$$

Apply WB model

$\hat{Y}^{WB} = f^{WB}(X)$	ϵ^{WB}
315	347-315=32
475	538-475=63
144	121-144=-23
201	172-201=-29
188	266-188=78

2) Calculate residuals

i	Product	List price	Age	Industry	…	ϵ^{WB}
1	Dell XPS 15’	2,500	36	Mining	…	32
2	Dell XPS 17’	3,000	36	Health	…	63
3	HP Envy 17’	1,300	24	Office	…	-23
4	HP EliteBook 850	1,900	36	Mining	…	-29
5	Lenovo Yoga 13’	1,100	12	Office	…	78

3) Fit black-box model to residuals

$$\epsilon^{WB} = f^{BB}(X) + \epsilon^{BB}$$

Apply BB model

$\hat{Y}^{BB} = f^{BB}(X)$	$\hat{Y} = \hat{Y}^{WB} + \hat{Y}^{BB}$	ϵ
29	344=315+29	347-344=3
66	541=475+66	538-541=-3
-18	126=144-18	121-126=-5
-26	175=201-26	172-175=-3
76	276=188+76	266-264=2

4) Integrated both models’ forecasts

$$\hat{Y} = \hat{Y}^{WB} + \hat{Y}^{BB} = f^{WB}(X) + f^{BB}(X)$$

See, e.g., Kraus & Feuerriegel (2019)

Permutation-Based Feature Importance

A feature is important if corrupting it hurts the model

- **Learner-agnostic way to judge the relevance of features**
  - Assess model performance before and after corrupting one feature
  - If performance decreases substantially, the feature was important
- **Produces a feature importance score to rank-order features**

PRODUCT	LIST PRICE	AGE	…	**RESALE PRICE**
Dell XPS 15’	2,500	36	…	347
Dell XPS 15’	2,500	24	…	416
Dell XPS 17’	3,000	36	…	538
HP Envy 17’	1,300	24	…	121
Lenovo Yoga 13’	1,100	12	…	266
…	…	…	…	…

Further information: Fisher et al. (2019)

Permutation-Based Feature Importance

A feature is important if accuracy decreases after corrupting that feature

- **Proposed in Breiman's (2001) paper on random forest**
- **Easily extendible to any type of predictive model**
- **Algorithm based on Fisher et al (2019)**
  - Input:
    Trained model f, feature matrix X, target vector y, error measure L(y,f).
  - Estimate the original model error e^{orig} = L(y, f(X)) (eg MSE)
  - For each feature j = 1,…,p do:
    - Generate feature matrix X^{perm} by permuting feature j in the data X. This breaks the association between feature j and the true outcome y.
    - Estimate error e^{perm} = L(Y,f(X^{perm})) based on the predictions of the permuted data.
    - Calculate permutation feature importance $FI^j = e^{perm}/e^{orig}$. Alternatively, the difference can be used: $FI^j = e^{perm} - e^{orig}$
  - Sort features by descending FI.

Permutation-Based Feature Importance

A feature is important if accuracy decreases after corrupting that feature

- **Proposed in Breiman's (2001) paper on random forest**
- **Easily extendible to any type of predictive model**
- **Algorithm based on Fisher et al (2019)**
  - Input:
    Trained model f, feature matrix X, target vector y,
  - Estimate the original model error e^{orig} = L(y, f(X)) (
  - For each feature j = 1,…,p do:
    - Generate feature matrix X^{perm} by permuting feature j in th association between feature j and the true outcome y.
    - Estimate error e^{perm} = L(Y,f(X^{perm})) based on the predictio
    - Calculate permutation feature importance $FI^j = e^{perm}/e^{orig}$. A difference can be used: $FI^j = e^{perm} - e^{orig}$
  - Sort features by descending FI.

Research note:

- Many papers have studied feature importance scoring algorithms and caution again potential bias
- For example, the SKLEARN implementation of Random Forest incorporates a feature importance measure that is known for its bias
  - See Strobl et al. (2007)
  - To be fair, the SKLEARN document-ation includes a warning
- Permutation-based importance is also vulnerable (see, e.g. Hooker et al. 2021)
- No free lunch

Formal Perspective on Partial Dependence Plot Analysis

Examine the marginal effect of a feature on model predictions

- **Proposed together with GBM in Friedman (2001)**
  - Depict how predictions change with changes in a feature when keeping everything else constant
  - Predictions equate to forecasts of the target/class membership probabilities in regression/classification
- **Formal definition of the partial dependence function**

$$\hat{f}_S(\boldsymbol{x}_S) = E_{\boldsymbol{X}_C}\left[\hat{f}(\boldsymbol{x}_S, \boldsymbol{X}_C)\right] = \int \hat{f}(\boldsymbol{x}_S, \boldsymbol{X}_C) d\mathbb{P}(\boldsymbol{X}_C)$$

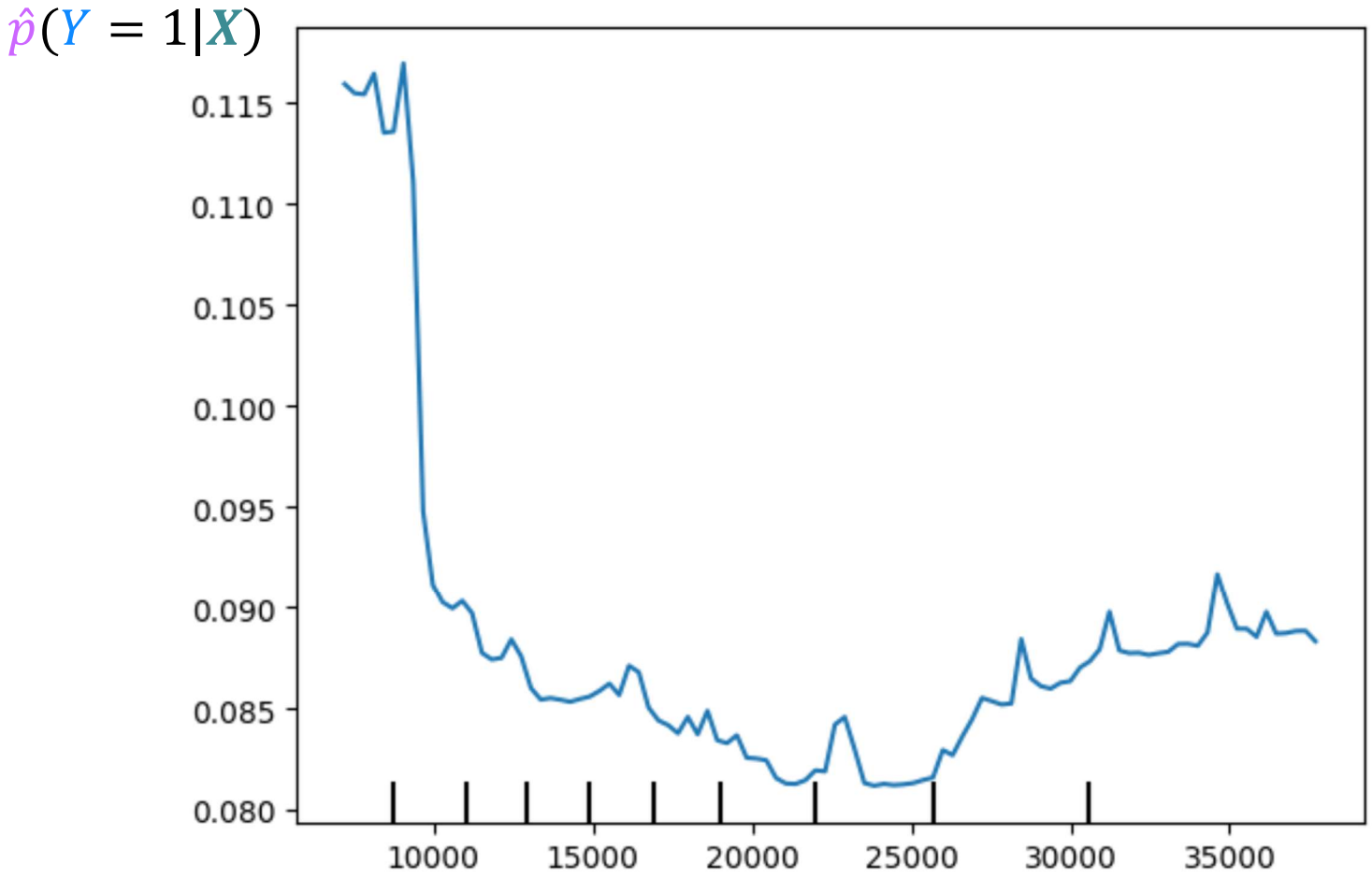
  - With $\boldsymbol{x}_S$ denoting the features for which partial dependence is plotted, $\boldsymbol{X}_C$ the remaining features, which we treat as random variables, and $\hat{f}$ the machine learning model
  - The features in $\boldsymbol{x}_S, \boldsymbol{X}_C$ make up the whole feature space where $|S| \leq 2$
- **Estimate partial dependence function by calculating averages in the training set**

$$\hat{f}_S(\boldsymbol{x}_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}\left(\boldsymbol{x}_S, \boldsymbol{X}_C^{(i)}\right)$$

  - With $\boldsymbol{X}_C^{(i)}$ denoting the actual feature values from the data set and n the number of instances

Partial Dependence Plot (PDP)

A visualization of a feature's marginal effect on the model prediction



$X :=$ CREDIT AMOUNT

Partial Dependence Plot (PDP)

Examine the marginal effect of a feature on model predictions

Raw data

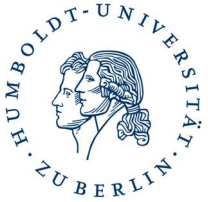
i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2	1	0.55	$25,000	$65,000	< 2 years
3	1	0.33	$50,000	$55,000	>5 years
4	0	0.53	$25,000	$45,000	< 2 years
5	0	0.79	$50,000	$111,000	2 - 5 years

- **Study how changes in one (or at most two) feature(s) affect the output of the predictive model when keeping everything else constant**
- **Model output varies with the specific type of predictive model**
  - Regression models forecast the actual value of the target variable
  - Classification models estimate class-membership probabilities

Further information: Friedman et al. (2001)

Partial Dependence Plot (PDP)

Examine the marginal effect of a feature on model predictions



Raw data

i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2	1	0.55	$25,000	$65,000	< 2 years
3	1	0.33	$50,000	$55,000	>5 years
4	0	0.53	$25,000	$45,000	< 2 years
5	0	0.79	$50,000	$111,000	2 - 5 years

- **Say we consider the feature CREDIT AMOUNT**

- **Study how changes in one (or at most two) feature(s) affect the output of the predictive model when keeping everything else constant**
- **Model output varies with the specific type of predictive model**
  - Regression models forecast the actual value of the target variable
  - Classification models estimate class-membership probabilities

Further information: Friedman et al. (2001)

Partial Dependence Plot (PDP)

Examine the marginal effect of a feature on model predictions

Raw data

i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	\$25,000	\$75,000	2 - 5 years
2	1	0.55	\$25,000	\$65,000	< 2 years
3	1	0.33	\$50,000	\$55,000	>5 years
4	0	0.53	\$25,000	\$45,000	< 2 years
5	0	0.79	\$50,000	\$111,000	2 - 5 years

- **Say we consider the feature CREDIT AMOUNT**
- **Step 1: Expand the data set**
  - Include each case as many time as there are distinct credit amounts
    - Two values in our example, 25K and 50K
    - Heuristics to accelerate calculations

Further information: Friedman et al. (2001)

Partial Dependence Plot (PDP)

Examine the marginal effect of a feature on model predictions

Raw data

i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2					
3					
4					
5					

i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2	1	0.55	$25,000	$65,000	< 2 years
3	1		$25,000	$55,000	>5 years
4	0	0.53	$25,000	$45,000	< 2 years
5	0		$25,000	$111,000	2 - 5 years
1	1		$50,000	$75,000	2 - 5 years
2	1		$50,000	$65,000	< 2 years
3	1	0.33	$50,000	$55,000	>5 years
4	0		$50,000	$45,000	< 2 years
5	0	0.79	$50,000	$111,000	2 - 5 years

- **Say we consider the feature CREDIT AMOUNT**
- **Step 1: Expand the data set**
  - Include each case as many time as there are distinct credit amounts
    - Two values in our example, 25K and 50K
    - Heuristics to accelerate calculations

Further information: Friedman et al. (2001)

Partial Dependence Plot (PDP)

Examine the marginal effect of a feature on model predictions

Raw data

i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2					
3					
4					
5					

i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2	1	0.55	$25,000	$65,000	< 2 years
3	1		$25,000	$55,000	>5 years
4	0	0.53	$25,000	$45,000	< 2 years
5	0		$25,000	$111,000	2 - 5 years
1	1		$50,000	$75,000	2 - 5 years
2	1		$50,000	$65,000	< 2 years
3	1	0.33	$50,000	$55,000	>5 years
4	0		$50,000	$45,000	< 2 years
5	0	0.79	$50,000	$111,000	2 - 5 years

- **Say we consider the feature CREDIT AMOUNT**
- **Step 1: Expand the data set**
  - Include each case as many time as there are distinct credit amounts
    - Two values in our example, 25K and 50K
    - Heuristics to accelerate calculations
  - Predict new, synthetic data instances

Further information: Friedman et al. (2001)

Partial Dependence Plot (PDP)

Examine the marginal effect of a feature on model predictions

Raw data

i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2					
3					
4					
5					

i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2	1	0.55	$25,000	$65,000	< 2 years
3	1	0.27	$25,000	$55,000	>5 years
4	0	0.53	$25,000	$45,000	< 2 years
5	0	0.61	$25,000	$111,000	2 - 5 years
1	1	0.55	$50,000	$75,000	2 - 5 years
2	1	0.61	$50,000	$65,000	< 2 years
3	1	0.33	$50,000	$55,000	>5 years
4	0	0.57	$50,000	$45,000	< 2 years
5	0	0.79	$50,000	$111,000	2 - 5 years

- **Say we consider the feature CREDIT AMOUNT**
- **Step 1: Expand the data set**
  - Include each case as many time as there are distinct credit amounts
    - Two values in our example, 25K and 50K
    - Heuristics to accelerate calculations
  - Predict new, synthetic data instances

Further information: Friedman et al. (2001)

Partial Dependence Plot (PDP)

Examine the marginal effect of a feature on model predictions

Raw data

i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2					
3					
4					
5					

i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2	1	0.55	$25,000	$65,000	< 2 years
3	1		$25,000	$55,000	>5 years
4	0	0.53	$25,000	$45,000	< 2 years
5	0	0.61	$25,000	$111,000	2 - 5 years
1	1	0.55	$50,000	$75,000	2 - 5 years
2	1	0.61	$50,000	$65,000	< 2 years
3	1		$50,000	$55,000	>5 years
4	0	0.57	$50,000	$45,000	< 2 years
5	0	0.79	$50,000	$111,000	2 - 5 years

Average

Average

- **Say we consider the feature CREDIT AMOUNT**
- **Step 1: Expand the data set**
  - Include each case as many time as there are distinct credit amounts
    - Two values in our example, 25K and 50K
    - Heuristics to accelerate calculations
  - Predict new, synthetic data instances
- **Step 2: Average model predictions**

Further information: Friedman et al. (2001)

Partial Dependence Plot (PDP)

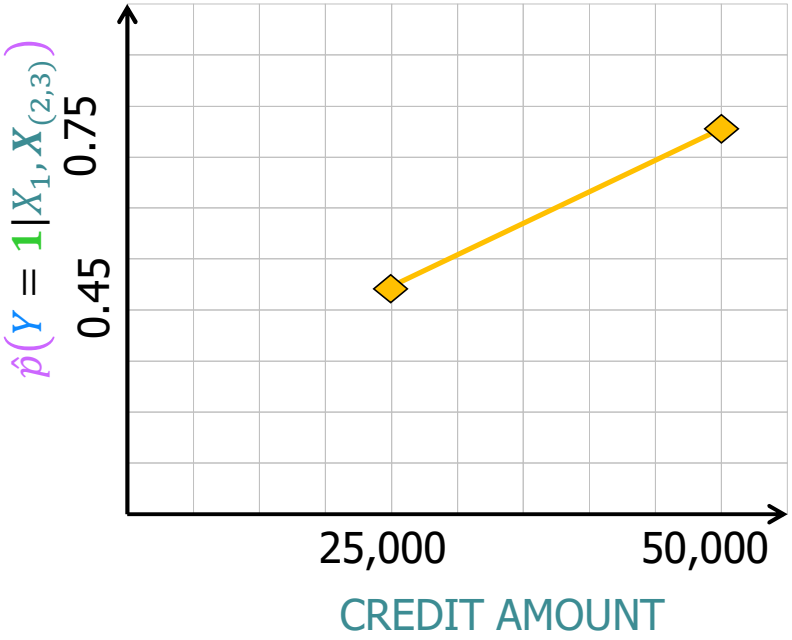
Examine the marginal effect of a feature on model predictions

Raw data

i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1	0.45	$25,000	$75,000	2 - 5 years
2					
3					
4					
5					

i	Y	$\hat{p}(Y=1\|\boldsymbol{X})$	CREDIT AMOUNT X_1	SALARY X_2	DURATION X_3
1	1				years
2	1				years
3	1				ears
4	0				years
5	0	0.61	$25,000	$111,000	2 - 5 years
1	1	0.55	$50,000	$75,000	2 - 5 years
2	1	0.61	$50,000	$65,000	< 2 years
3	1	0.33	$50,000	$55,000	>5 years
4	0	0.57	$50,000	$45,000	< 2 years
5	0	0.79	$50,000	$111,000	2 - 5 years

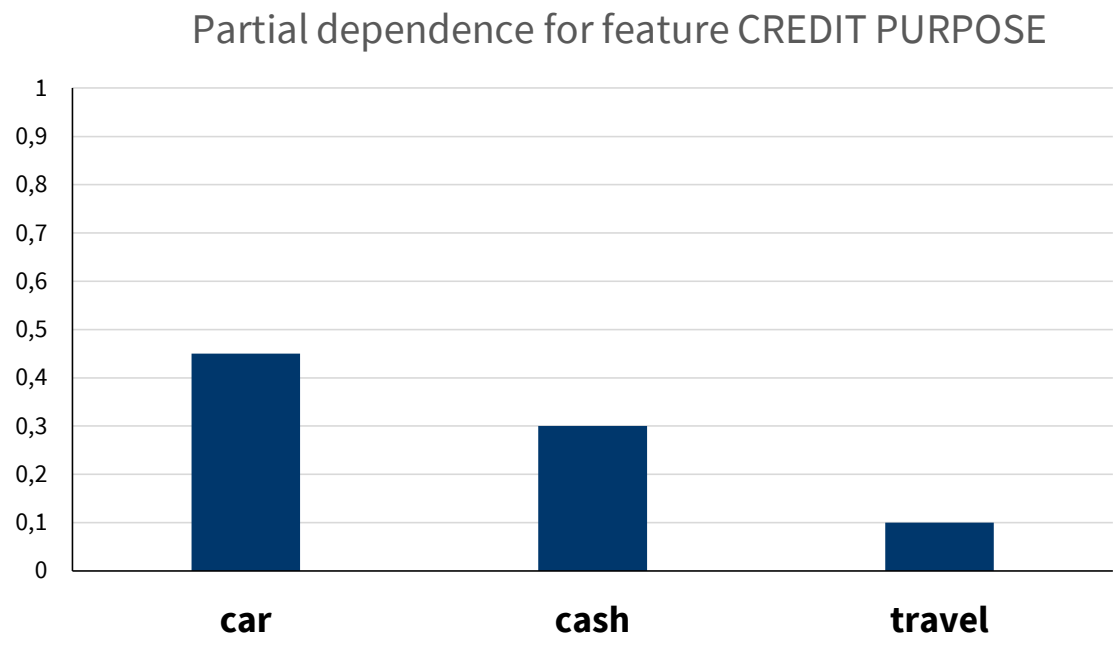
$\hat{p}(Y=1\|X_1, \boldsymbol{X}_{(2,3)})$	CREDIT AMOUNT X_1
0.45	$25,000
0.75	$50,000



Partial Dependence Plot (PDP)

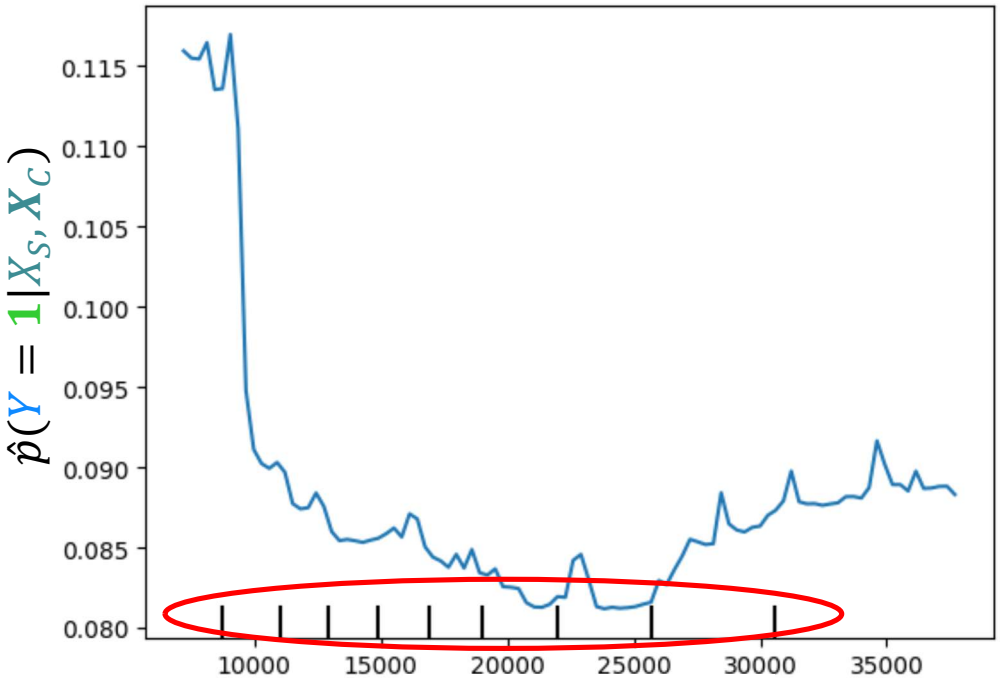
Categorical features: force all data instances to have the same level

- **For categorical features, force all data instances to have the same level**
  - Consider feature credit purpose with levels *car*, *cash*, and *travel*
  - To compute value for *car*, replace category level for all data instances with this value and average over model forecasts
  - Proceed in the same way with the other levels



Partial Dependence Plot (PDP)

Concluding remarks and discussion



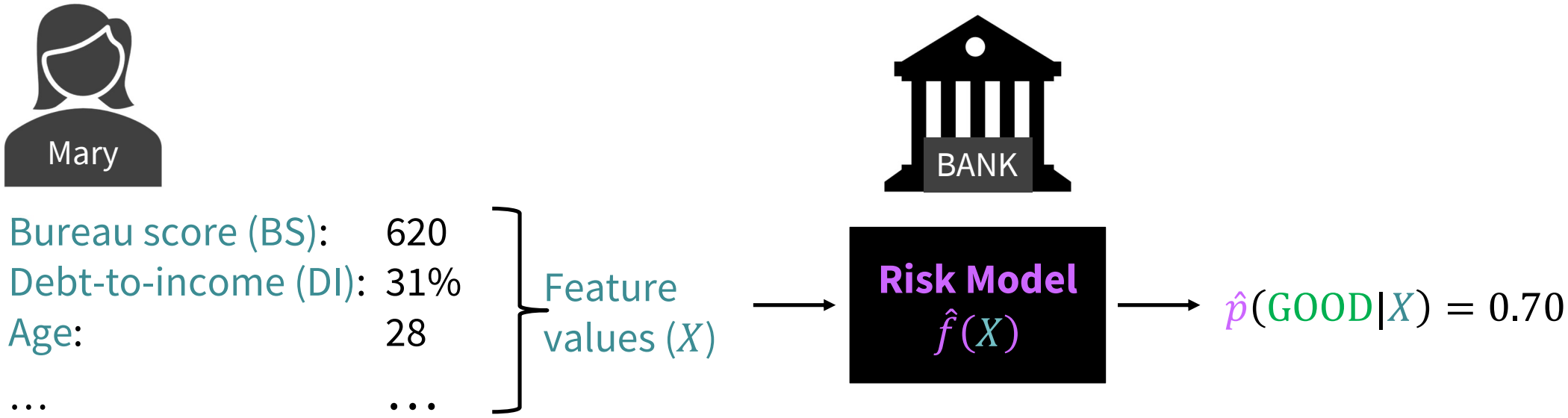
- **Easy to understand and implement**
- **Causal interpretation of the model prediction**
- **Make sure to examine feature distribution**
- **Does not scale to more than two features**
- **Assume that features in set S and set C are not correlated**
  - Correlation can lead to implausible data instances during averaging
  - For ex. Very small INCOME paired with very large CREDIT AMOUNT
  - Possible remedy: Accumulated Local Effect plots
- **Heterogeneous effects may be overlooked**
  - Say feature has pos. correlation with outcome for males and neg. correlation with females
  - Average effect of the feature is zero → PDP plot shows a horizontal line
  - Individual Conditional Expectation Curves (ICE) can uncover interaction effects

Local and Example Based Explanations

LIME, SHAP values, counterfactual and adversarial examples

Local Explanation Setting: Explain prediction (decision) to a Subject

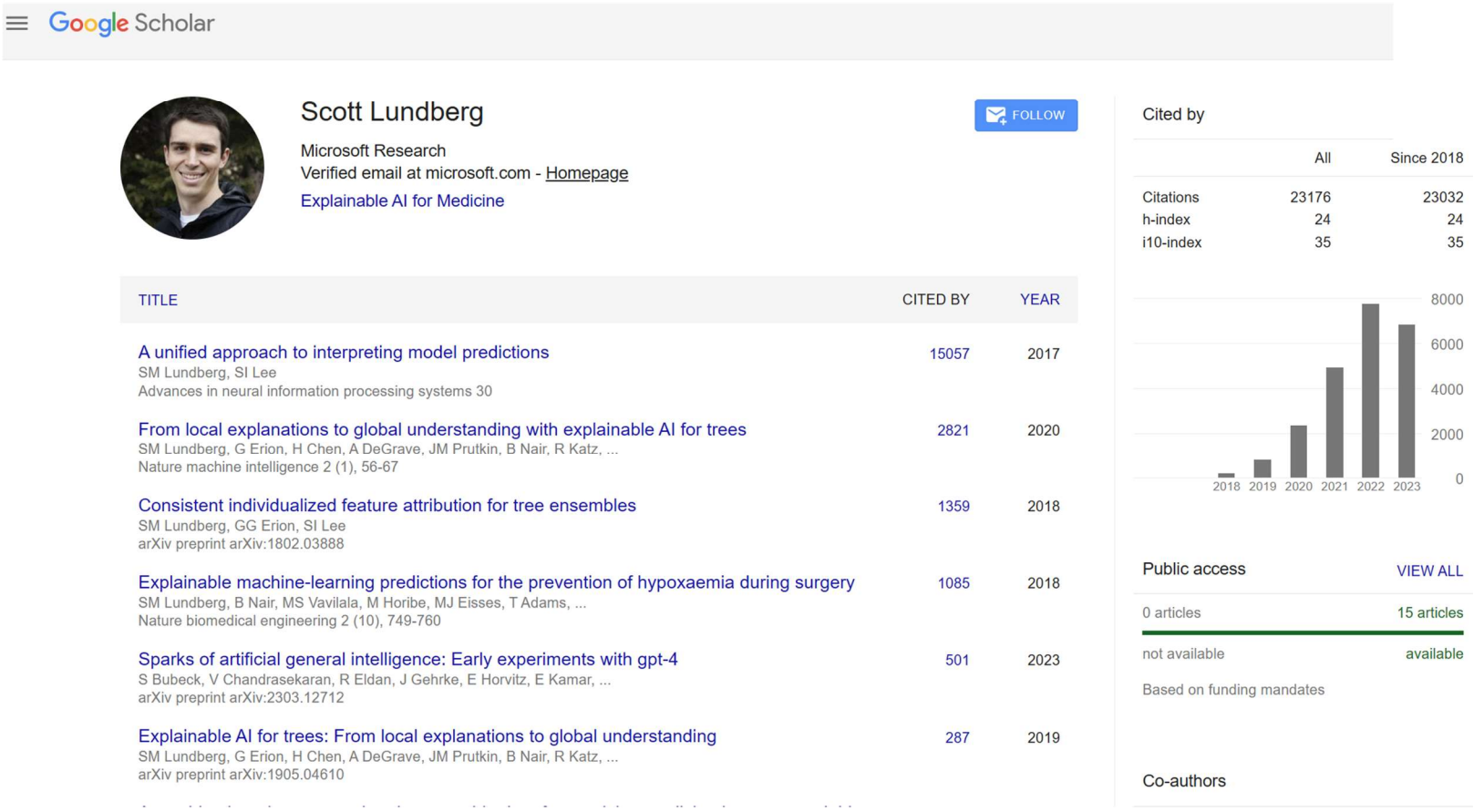
Credit scoring use case: explain approval decision to applicant



Say the bank decides to reject the application based on the model forecast. By GDPR, the subjects of algorithmic decisions have the right to an explanation.
How do we explain Mary why she was rejected?

SHapley Additive exPlanations

Lundberg & Lee (2017, 2020)



Google Scholar

Scott Lundberg

FOLLOW

Microsoft Research
Verified email at microsoft.com - Homepage
Explainable AI for Medicine

TITLE	CITED BY	YEAR
A unified approach to interpreting model predictions SM Lundberg, SI Lee Advances in neural information processing systems 30	15057	2017
From local explanations to global understanding with explainable AI for trees SM Lundberg, G Erion, H Chen, A DeGrave, JM Prutkin, B Nair, R Katz, ... Nature machine intelligence 2 (1), 56-67	2821	2020
Consistent individualized feature attribution for tree ensembles SM Lundberg, GG Erion, SI Lee arXiv preprint arXiv:1802.03888	1359	2018
Explainable machine-learning predictions for the prevention of hypoxaemia during surgery SM Lundberg, B Nair, MS Vavilala, M Horibe, MJ Eisses, T Adams, ... Nature biomedical engineering 2 (10), 749-760	1085	2018
Sparks of artificial general intelligence: Early experiments with gpt-4 S Bubeck, V Chandrasekaran, R Eldan, J Gehrke, E Horvitz, E Kamar, ... arXiv preprint arXiv:2303.12712	501	2023
Explainable AI for trees: From local explanations to global understanding SM Lundberg, G Erion, H Chen, A DeGrave, JM Prutkin, B Nair, R Katz, ... arXiv preprint arXiv:1905.04610	287	2019

Cited by

	All	Since 2018
Citations	23176	23032
h-index	24	24
i10-index	35	35

Public access VIEW ALL

0 articles 15 articles

not available available

Based on funding mandates

Co-authors

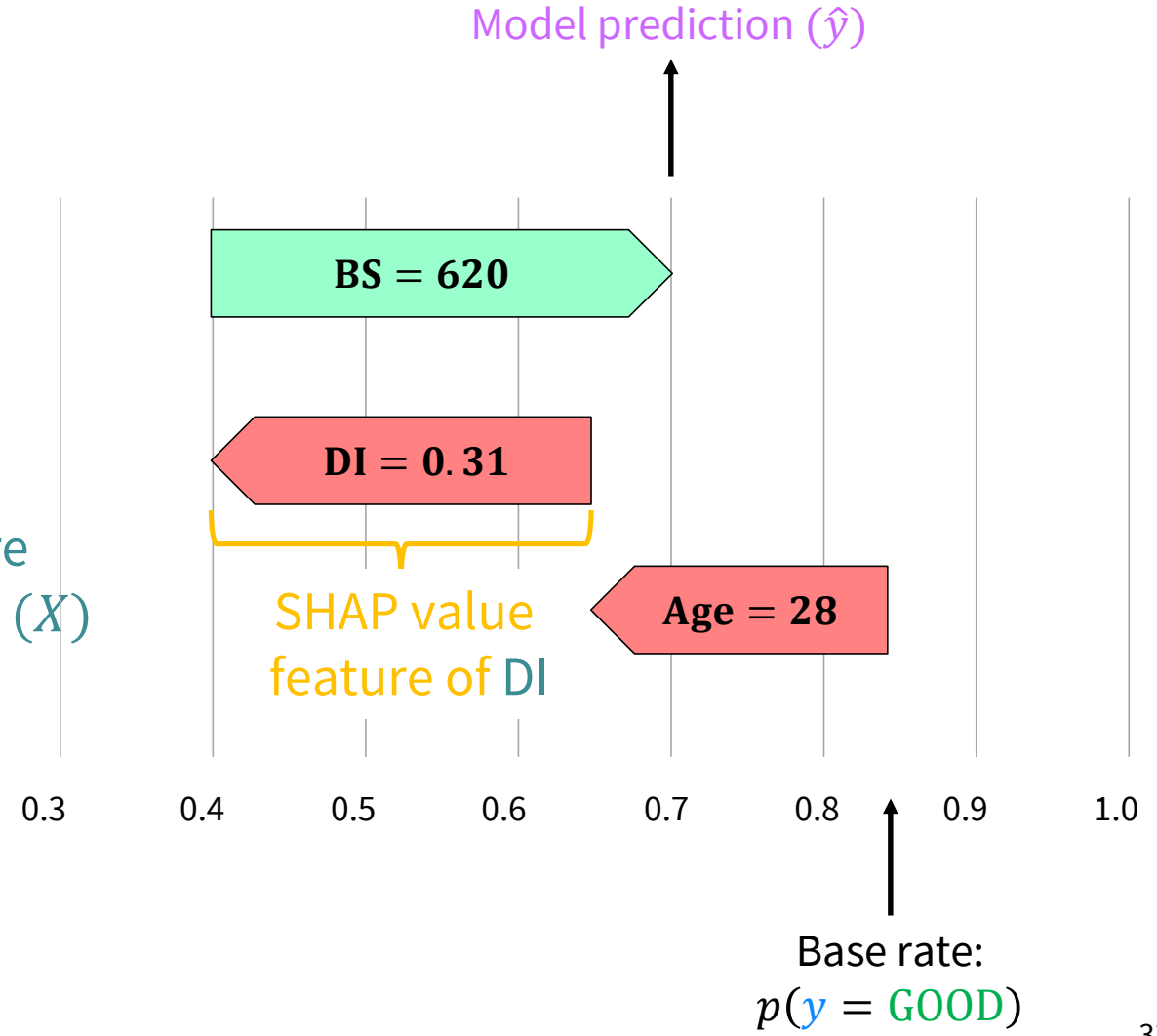
https://github.com/shap/shap

SHAP Facilitates Model-Agnostic Local Explanations

Credit scoring example

Bureau score (BS): 620
Debt-to-income (DI): 31%
Age: 28
…

Feature values (X)



Model estimate:

$$\hat{f}(X) = \hat{p}(y = \text{GOOD}|X) = 0.70$$

Shapley Value

Fair payout of a player in a cooperative game

- **Introduced by Lloyd Shapley**
  - Famous economist specializing in game theory
  - (Shared) Winner of the 2023 Nobel Prize for economic sciences
- **Captures the marginal contribution of a player to a coalition in a cooperative game**
  - Payout of the collation with the player included minus the payout with that player not playing
  - Calculated as the weighted average of a player's contributions to all the coalitions that the player could join.
- **Based on well-defined axioms, which conceptualize fairness**
  - Efficiency, symmetry, null player, additivity, (consistency)
  - Mathematically provable that the Shapley value is the only value satisfying all axioms

Lloyd Shapley in 1980
Source: Wikipedia

Key Properties of the Shapley Value

- **Symmetry**
  - Two players are considered interchangeable if they make the same contributions to all coalitions.
  - If two players are interchangeable, then they must be given an equal share of the total payout.
- **Null player property**
  - If a player makes zero marginal contribution to all coalitions, then s/he receives none of the payout.
- **Additivity**
  - If combining multiple games, a player's overall contribution is the sum of the contributions for the individual games.
  - This axiom assumes that any games played are independent.
- **Efficiency**
  - Total payout is distributed among the players
- **(Consistency)**

Illustration of the Shapley Value

Weighted average of a player's contributions to all coalitions s/he could join

Assume a game with three players. When playing together, they achieve a total payout of 10,000. The question is how to distribute this payout fairly.

	Payout	Players		
		1	2	3
	10,000	X	X	X

Illustration of the Shapley Value

Weighted average of a player's contributions to all coalitions s/he could join

Assume a game with three players. When playing together, they achieve a total payout of 10,000. The question is how to distribute this payout fairly. To answer this, we consider all possible coalitions and the corresponding payouts.

	Payout	Players		
		1	2	3
1	10,000	X	X	X
2	7,500	X	X	
3	7,500	X		X
4	5,000		X	X
5	5,000	X		
6	5,000		X	
7	0			X
8	0			

Illustration of the Shapley Value

Weighted average of a player's contributions to all coalitions s/he could join

To calculate a player's Shapely value, we need to compute the marginal contribution that the player made to a coalition. We do this for all coalitions the player could join.

	Payout	Players			Marginal contribution of player 1
		1	2	3	
1	10,000	X	X	X	
2	7,500	X	X		
3	7,500	X		X	
4	5,000		X	X	5,000
5	5,000	X			
6	5,000		X		2,500
7	0			X	7,500
8	0				5,000

Illustration of the Shapley Value

Weighted average of a player's contributions to all coalitions s/he could join

To calculate a player's Shapely value, we need to compute the marginal contribution that player made to a coalition. We do this for all coalitions the player could join. Further, we must consider the probability of the player making those contributions to a coalition of three players.

	Payout	Players			Marginal contribution of player 1
		1	2	3	
1	10,000	X	X	X	
2	7,500	X	X		
3	7,500	X		X	
4	5,000		X	X	5,000
5	5,000	X			
6	5,000		X		2,500
7	0			X	7,500
8	0				5,000

Illustration of the Shapley Value

Weighted average of a player's contributions to all coalitions s/he could join

Recall that we aim to distribute the total payout among three players. Assume the players join sequentially. Then, we have six alternative ways to form the grand coalition with all players.

Possible orders of players joining the grand coalition	Player 1	Player 2	Player 3
	1	2	3
	1	3	2
	2	1	3
	3	1	2
	2	3	1
	3	2	1

In general, we have $F!$ alternative ways to form the grand coalition, where F denotes the number of players.

Illustration of the Shapley Value

Weighted average of a player's contributions to all coalitions s/he could join

To calculate a player's Shapely value, we need to compute the marginal contribution that the player made to a coalition. We do this for all coalitions the player could join. Further, we must consider the probability of the player making those contributions to a coalition of three players.

	Payout	Players			Marginal contribution of player 1
		1	2	3	
1	10,000	X	X	X	
2	7,500	X	X		
3	7,500	X		X	
4	5,000		X	X	5,000
5	5,000	X			
6	5,000		X		2,500
7	0			X	7,500
8	0				5,000

Possible orders of joining	Player 1	Player 2	Player 3
	1	2	3
	1	3	2
	2	1	3
	3	1	2
	2	3	1
	3	2	1

Illustration of the Shapley Value

Weighted average of a player's contributions to all coalitions s/he could join

To calculate a player's Shapely value, we need to compute the marginal contribution that the player made to a coalition. We do this for all coalitions the player could join. Further, we must consider the probability of the player making those contributions to a coalition of three players.

	Payout	Players			Marginal contribution of player 1	Probability of player 1 making contribution
		1	2	3		
1	10,000	X	X	X		
2	7,500	X	X			
3	7,500	X		X		
4	5,000		X	X	5,000	2/6 = 1/3
5	5,000	X				
6	5,000		X		2,500	1/6
7	0			X	7,500	1/6
8	0				5,000	2/6 = 1/3

Illustration of the Shapley Value

Weighted average of a player's contributions to all coalitions s/he could join

Finally, we are ready to compute the Shapley value of player 1, ϕ_1, as the expected marginal contribution, where we obtain the expectation by probability weighting:

$$\phi_1 = 5{,}000^*1/3 + 2{,}500^*1/6 + 7{,}500^*1/6 + 5{,}000^*1/3 = 5{,}000$$

	Payout	Players			Marginal contribution of player 1	Probability of player 1 making contribution
		1	2	3		
1	10,000	X	X	X		
2	7,500	X	X			
3	7,500	X		X		
4	5,000		X	X	5,000	2/6 = 1/3
5	5,000	X				
6	5,000		X		2,500	1/6
7	0			X	7,500	1/6
8	0				5,000	2/6 = 1/3

Illustration of the Shapley Value

Weighted average of a player's contributions to all coalitions s/he could join

We proceed analogously to compute the Shapley values of player 2 and player 3.

$\phi_2 = \ldots \quad = 3{,}750$

$\phi_3 = \ldots \quad = 1{,}250$

	Payout	Players			Player2		Player 3	
		1	2	3	Marginal contribution	Probability	Marginal contribution	Probability
1	10,000	X	X	X				
2	7,500	X	X					
3	7,500	X		X				
4	5,000		X	X				
5	5,000	X						
6	5,000		X					
7	0			X				
8	0							

Generalization and Formalization of the Shapley Value

Weighted average of a player's contributions to all coalitions s/he could join

$$\phi_j = \sum_{S \subseteq \{1,2,\ldots,m\} \setminus \{j\}} \frac{|S|!\,(m - |S| - 1)!}{m!} [val(S \cup \{j\}) - val(S)]$$

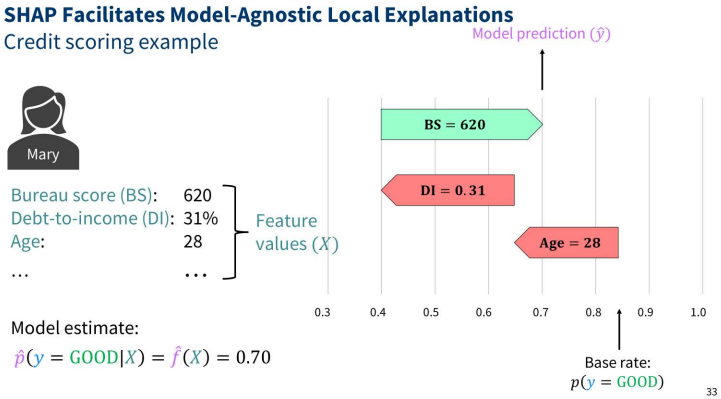
- **Main ideas**
  - Consider all coalitions to which a player j can make a marginal contribution
  - Marginal contribution of j equals the payout of the coalition with j included less the payout without j
  - Shapley value (i.e., fair payout) of j equals the weighted average of their marginal contributions
- **Notation**

ϕ_j	Shapley value of player j
m	Total number of players
S	Subset of coalitions not including player j
$val(\cdot)$	Value function facilitating computing the payout of a coalition
$\lvert S\rvert!$ / $\mathrm{m} - \lvert S\rvert! - 1$	Number of ways players can join a coalition before / after player j

Shapley Value for Machine Learning

Frame explanation task as a *game*

- **Players**
  - Recall our context: local explanation
  - Focus is to explain model prediction for one instance
  - Feature values of that instance are the players
- **Payout**
  - Feature values determine ML model output
  - So model prediction equals payout



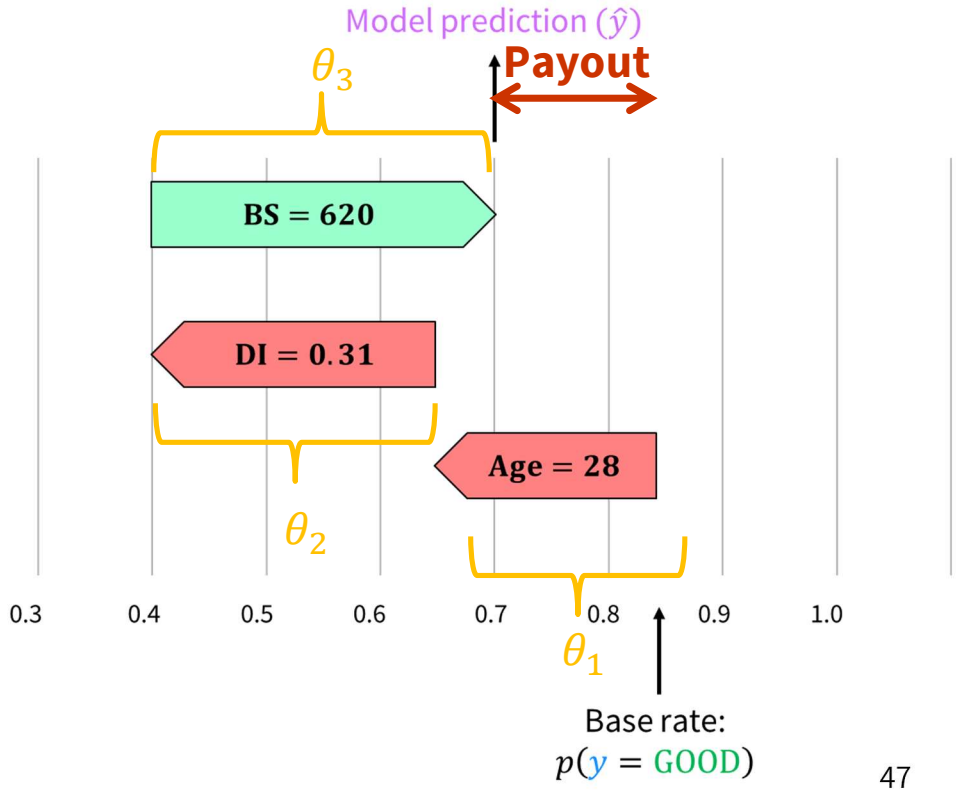
	Payout	Players		
		BS	DI	AGE

Shapley Value for Machine Learning

Frame explanation task as a *game*

	Payout $\hat{p}(y_i = \text{GOOD} \mid X_i) - p(y = \text{GOOD})$	Players / Feature values		
		BS	DI	AGE
Mary	0.70 – 0.85 = -0.15	620	31%	28

- **Players**
  - Recall our context: local explanation
  - Focus is to explain model prediction for one instance
  - The feature values of that instance are the players
- **Payout**
  - The feature values determine model prediction
  - We define the payout to be the difference between the model prediction and the base rate
- **The SHAP value θ_j of a feature j**
  - Explain difference between model prediction for one instance and the *average* model prediction
  - Tells us how much the feature value contributed to this difference (i.e., payout)



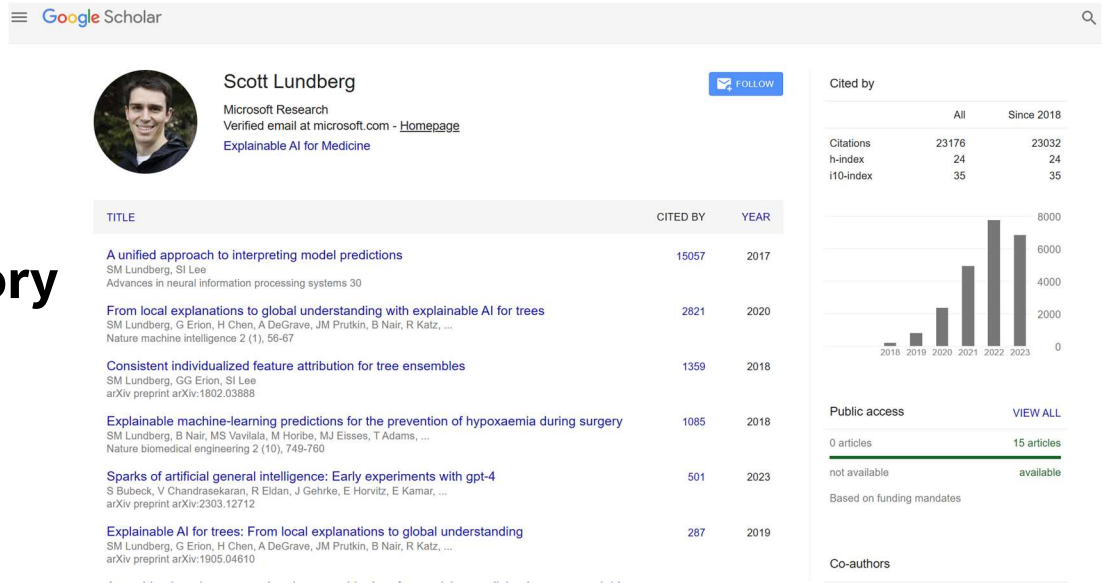
Generalization and Formalization of the Shapley Value for ML

$$\phi_j = \sum_{S \subseteq \{1,2,\ldots,m\} \setminus \{j\}} \frac{|S|!\,(m - |S| - 1)!}{m!} \left[\hat{f}_{S \cup \{j\}}\left(X_{S \cup \{j\}}\right) - \hat{f}_S\left(X_S\right) \right]$$

- **With**
  - ϕ_j Shapley value of feature j
  - m Number of features
  - S Subset of coalitions not including feature j
  - X Instance the prediction of which we aim to explain
  - $\hat{f}(X)$ Model output (aka prediction) for instance i with feature values $X_j, j = 1, \ldots, m$

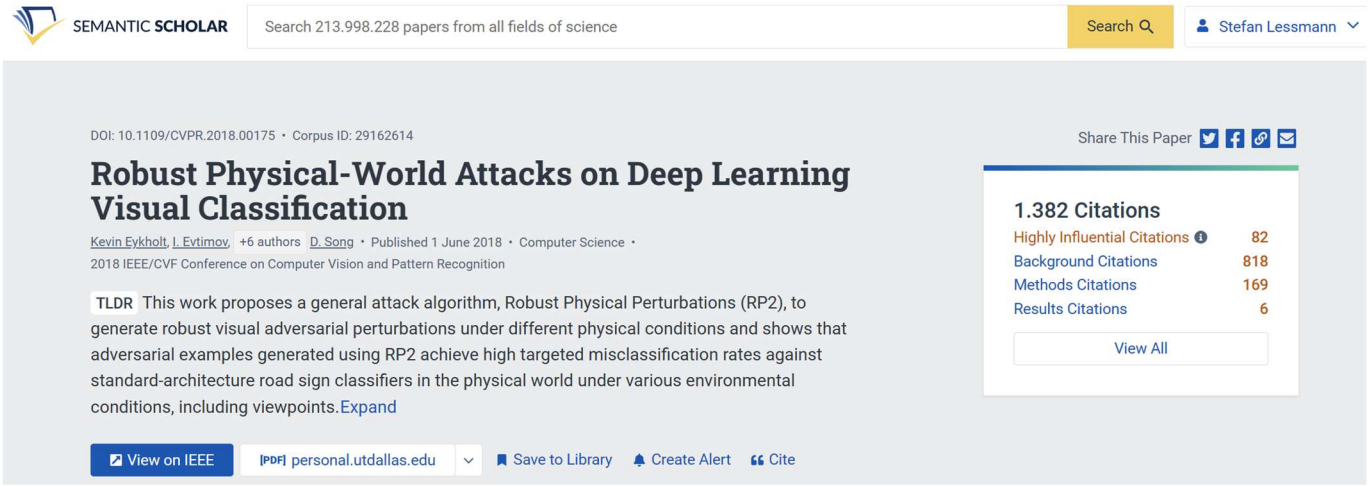
SHAP as a Quasi-Standard?

Lundberg & Lee (2017, 2020)



- **Shapley value long known in game theory**
  - Using Shapley value for explaining ML model predictions also around for long
  - E.g., Štrumbelj & Kononenko (2014)
- **Key contributions of SHAP**
  - Prior ideas to Shapely values were computationally intractable / restricted to small-scale settings
  - Lundberg & Lee (2017, 2020) address this problem
    - Model agnostic approach to approximate Shapely values (KernelSHAP)
    - Fast approximations for specific models (tree-based & deep learning)
  - Unify prior local explanation methods (e.g., LIME) in a unified framework
  - Software support (https://github.com/shap/shap) offering many great visualizations

Example-Based Explanation Methods



K. Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 1625-1634, doi: 10.1109/CVPR.2018.00175.

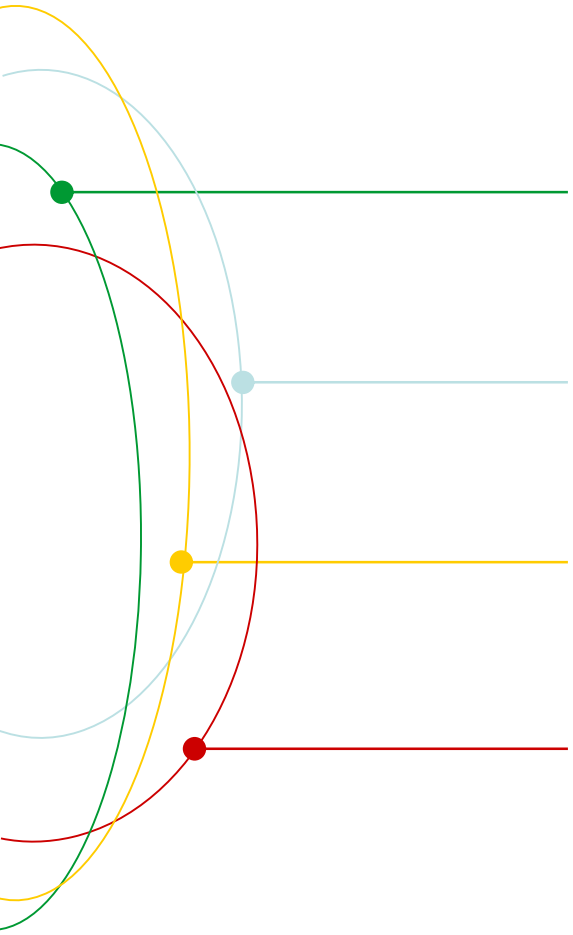
Example-Based Explanation Methods

Explain prediction by neighboring cases and/or feature manipulations

- **Anchors (Ribeiro et al., 2018)**
  - Local explanation method that finds a decision rule that “anchors” the prediction sufficiently
  - A rule anchors a prediction if changes in other feature values do not affect the prediction
- **Counterfactual examples**
  - Local explanation method that treats feature values as causes of an instance’s prediction
  - A counterfactual explanation describes the smallest change to the feature value(s) that changes the prediction to a predefined output
    - Counterfactual examples are new/artificial instances
    - Possibly many and contradictory counterfactuals
- **Adversarial examples**
  - As counterfactual examples but with the aim to deceive the model (e.g. min. change to make an error)
  - Much current research in the scope of computer vision (e.g. misinterpreting a sign in self-driving cars)

Summary

Summary



Learning goals

- Option space for interpreting ML models
- Understanding of specific approaches

Findings

- Crucial to report and diagnose ML-based systems
- Local versus global interpretation
- Example-based techniques
- Model-specific vs. –agnostic explanation methods
- Often use interpretable models as component
- Working of permutation importance, PDP, LIME

What next

- From ML to causal inference
- Uplift modeling use case

Literature

- Alvarez-Melis, D., & Jaakkola, T. S. (2018). *On the Robustness of Interpretability Methods*. ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden.
- Apley, D. W., & Zhu, J. (2016). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Archive Preprint,* arXiv:1612.08468v2.
- Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science,* 49(3), 312-329.
- Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research,* 20(177), 1-81.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*, 1189-1232.
- Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine,* 38(3), 50-57.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.
- Hooker, G., Mentch, L., & Zhou, S. (2021). Unrestricted Permutation forces Extrapolation: Variable Importance Requires at least One More Model, or There Is No Free Variable Importance. arXiv preprint, arXiv:1905.03151v2. https://doi.org/10.48550/arXiv.1905.03151
- Molnar, C. (2019) *Interpretable Machine Learning*. E-Book. https://christophm.github.io/interpretable-ml-book/
- Lipton, Z. C. (2016). The Mythos of Model Interpretability. *ICML 2016 Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, NY, USA.
- Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems, pp. 4765-4774.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56-67. https://doi.org/10.1038/s42256-019-0138-9
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM KDD2016), ACM: New York, NY, USA.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). *Anchors: High-Precision Model-Agnostic Explanations*. Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI'18).
- Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics,* 8(1), 1-20.

Thank you for your attention!

Stefan Lessmann

Chair of Information Systems
School of Business and Economics
Humboldt-University of Berlin, Germany

Tel. +49.30.2093.5742
Fax. +49.30.2093.5741

stefan.lessmann@hu-berlin.de
http://bit.ly/hu-wi

www.hu-berlin.de

Photo: Heike Zappe