PADERBORN UNIVERSITY

Erklärbare Künstliche Intelligenz

Dr. Stefan Heindorf

Data Science Junior Research Group

31.08.2023

Künstliche Intelligenz

Bildverarbeitung

- Bildklassifikation (Gesichtserkennung, Straßenschilder, medizinische Bilder)
- Texterkennung (z. B. Scan nach Word konvertieren)
- Bilderzeugung (z. B. DALL-E, Midjourney)

Natürliche Sprachverarbeitung

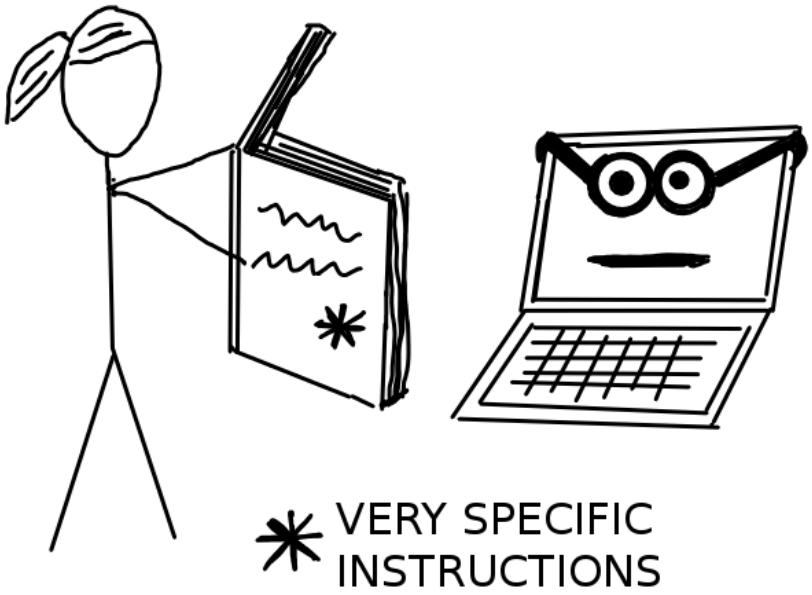
- Maschinelle Übersetzung (z. B. DeepL, Google Translate)
- Chatbots (z. B. Chat-GPT, Amazon Alexa)
- Dokumentenklassifikation (z. B. Spam / kein Spam, Dokumentenklassifizierung)

Brisante Entscheidungen

- Medizinische Diagnosen (z. B. basierend auf Symptomen, Röntgenbildern)
- Finanzwesen (z. B. Kreditvergabe basierend auf Historie)
- Justiz (z. B. COMPAS: Risiko für Rückfälligkeit von Straftätern einschätzen)

Kerntechnologie für künstliche Intelligenz:
Maschinelles Lernen

Klassische Programmierung



Mit maschinellem Lernen

Maschinelles Lernen

Beispiel: Fahrradverleih

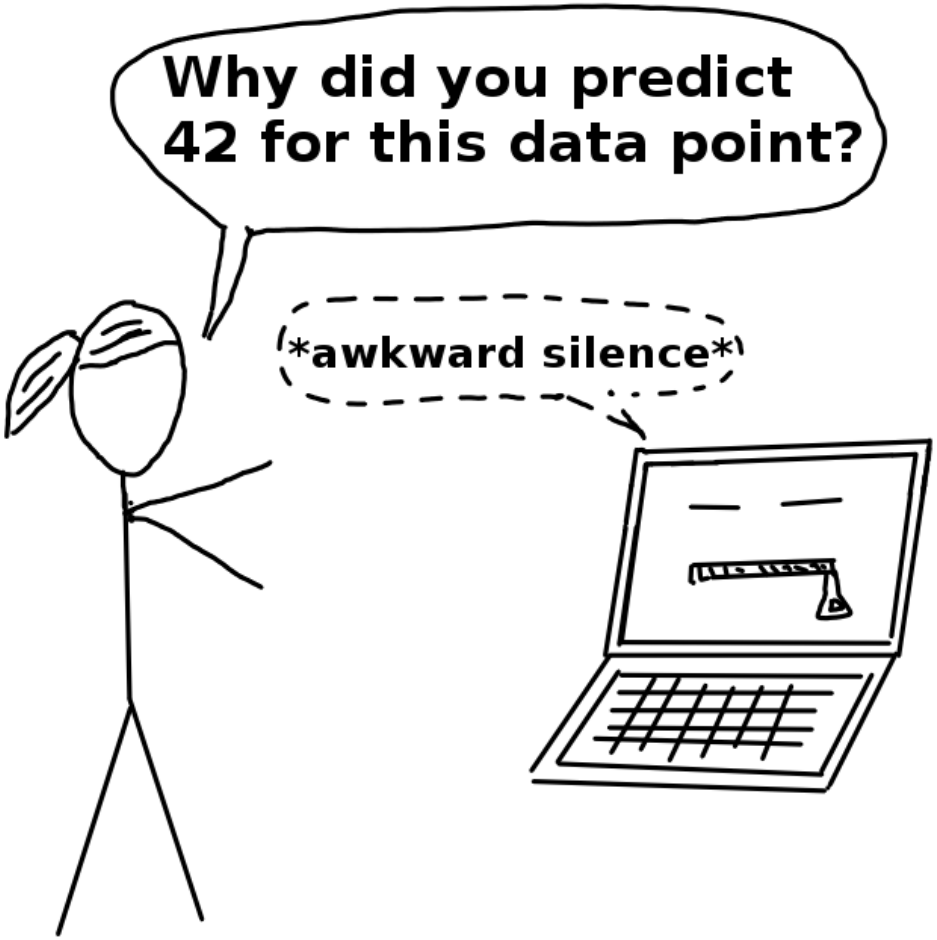
Training

Outlook	Temperature	Humidity	Windy	Rented bicycles
Sunny	Hot	High	False	4000
Sunny	Hot	Normal	True	5000
Overcast	Hot	Normal	False	3000
Rainy	Cool	High	False	1000

Vorhersage

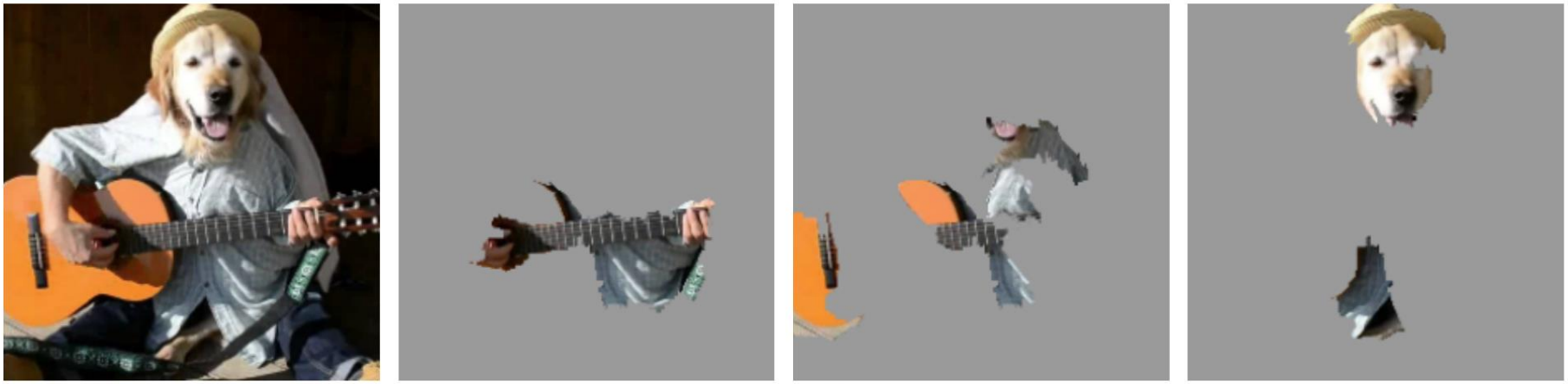
Outlook	Temperature	Humidity	Windy	Rented bicycles?
Sunny	Mild	High	True	?
Sunny	Cool	Normal	True	?

Bedarf der Erklärbarkeit



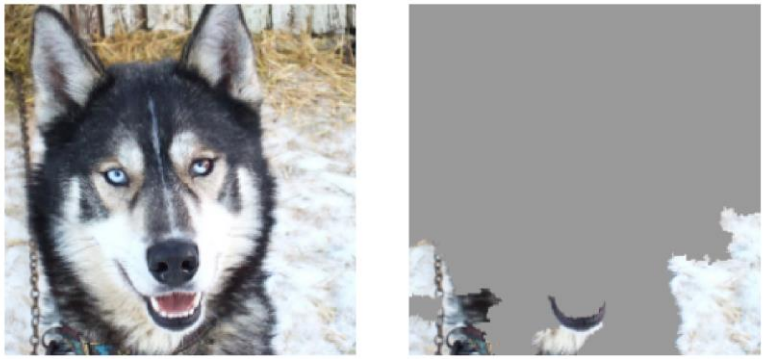
Bedarf der Erklärbarkeit

Beispiel: Bildklassifizierung [Riebeiro et al. 2016]



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)



(a) Husky classified as wolf (b) Explanation

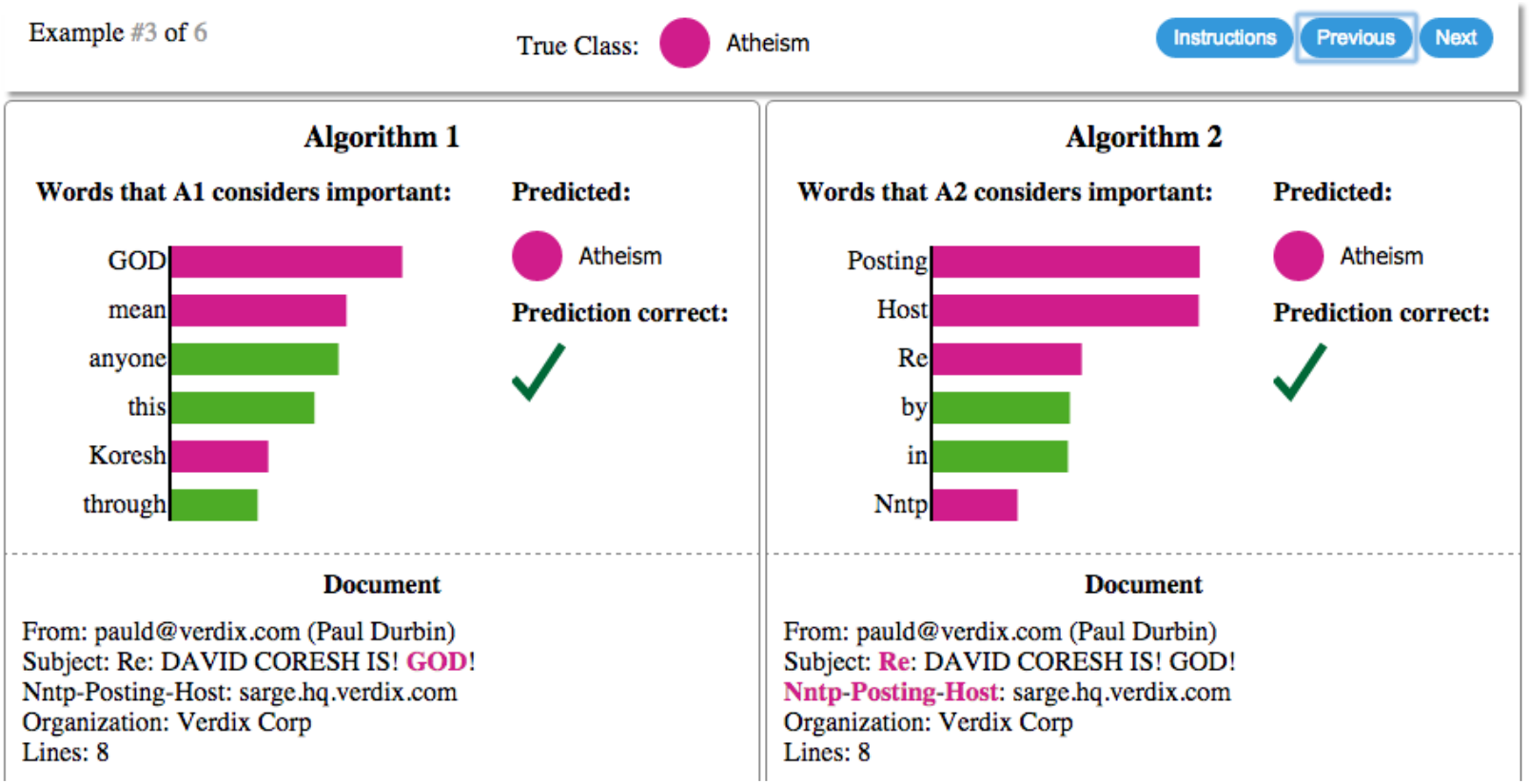
Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Bedarf der Erklärbarkeit

Beispiel: Textklassifizierung [Riebeiro et al. 2016]

- **Aufgabe:** Kategorie eines Dokuments vorhersagen
- Welcher Algorithmus ist besser?

88% Genauigkeit **94% Genauigkeit**



Wann ist Erklärbarkeit wichtig?

[Doshi-Velez and Kim 2017, Molnar 2020]

Unvollständigkeit in der Problemformalisierung

- Vorhersage (nach Zielfunktion) löst das Ursprungsproblem nur teilweise
- ➢ Modell muss auch erklären, wie es zur Vorhersage kam

Kausalität

- Werden nur kausale Beziehungen gelernt?

Zuverlässigkeit und Robustheit

- Kleine Änderungen der Eingabe → kleine Änderungen der Vorhersagen ?

Vertrauen

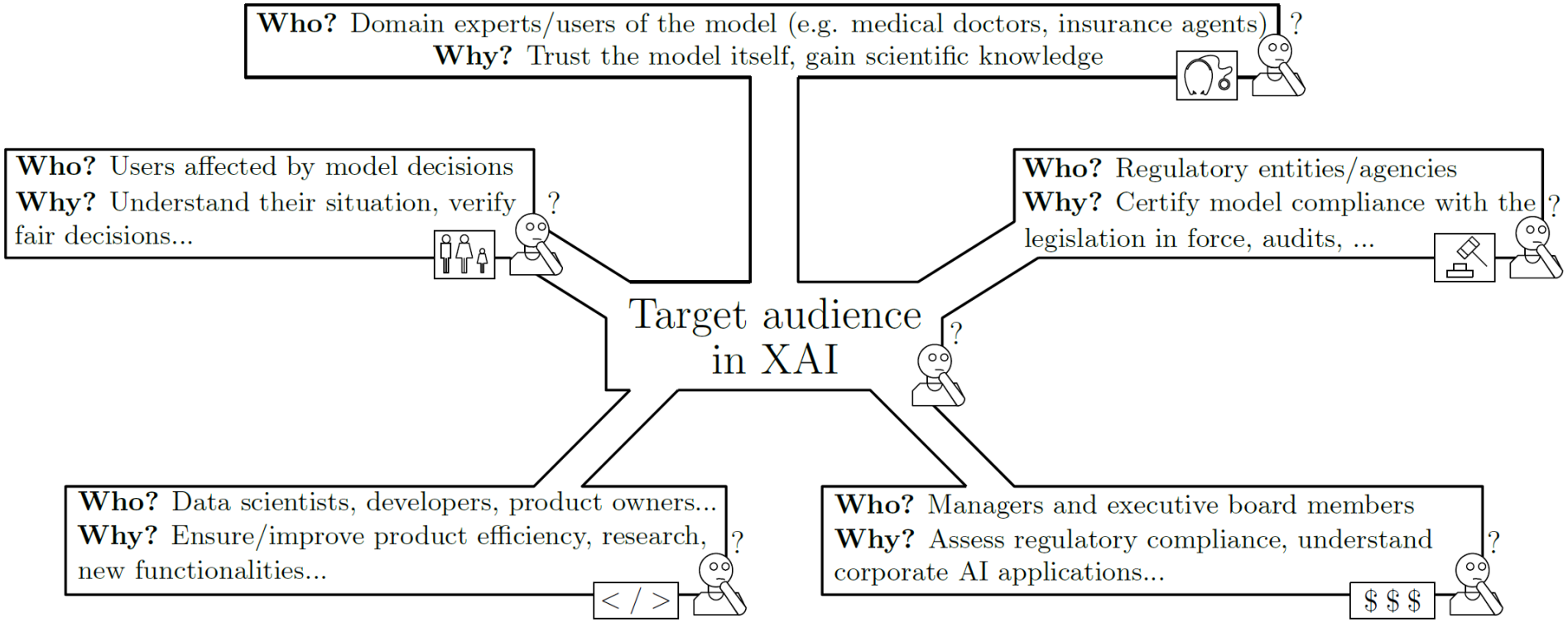
- Menschen vertrauen lieber einem System, das seine Entscheidungen erklärt, als einer Blackbox

Fairness

- Sicherstellen, dass Vorhersagen unvoreingenommen sind
- Keine Diskriminierung unterrepräsentierter Gruppen

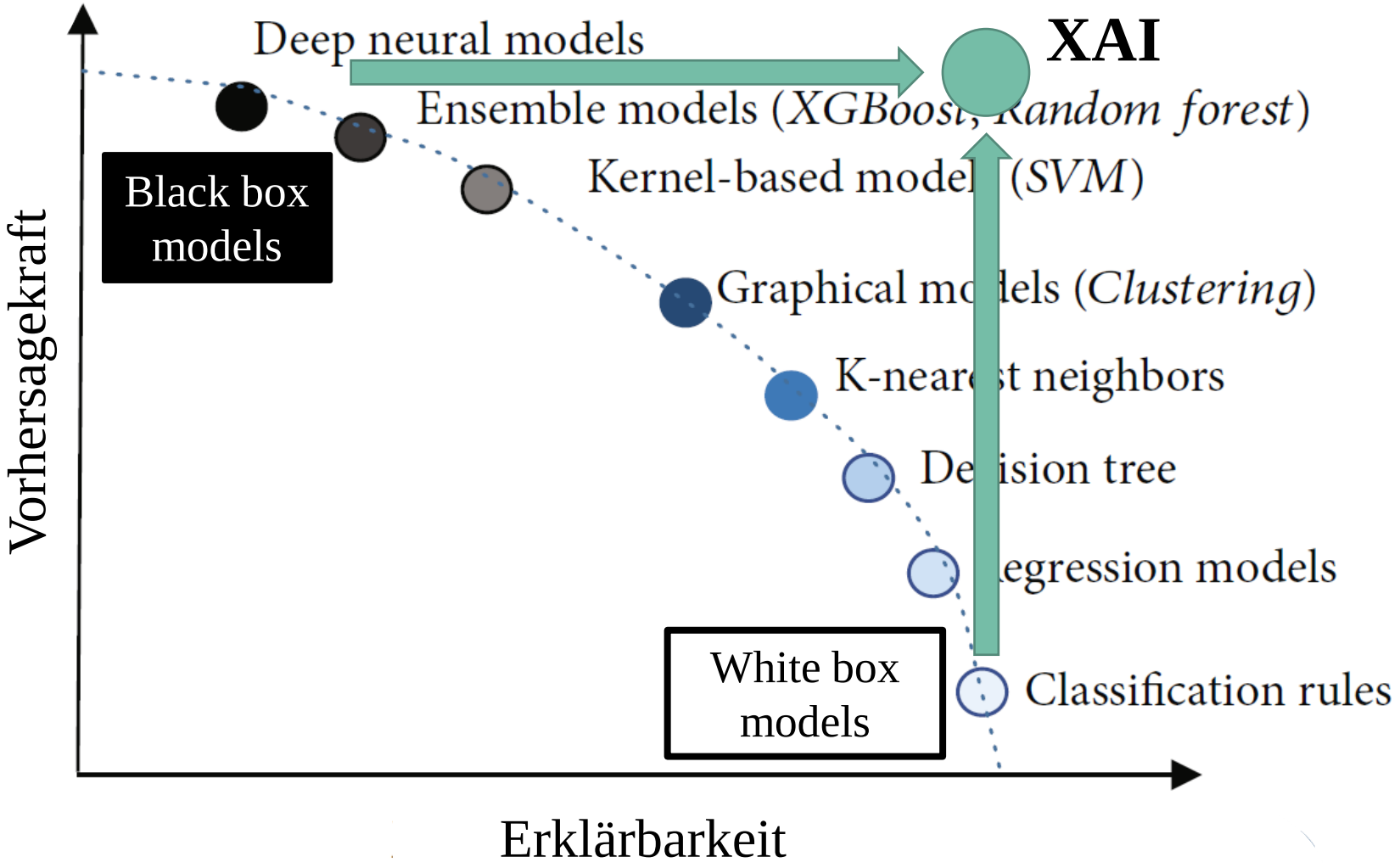
Erklärbare KI: Wer braucht dies und warum?

[Arrietta et al. 2020]

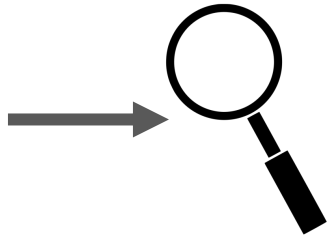


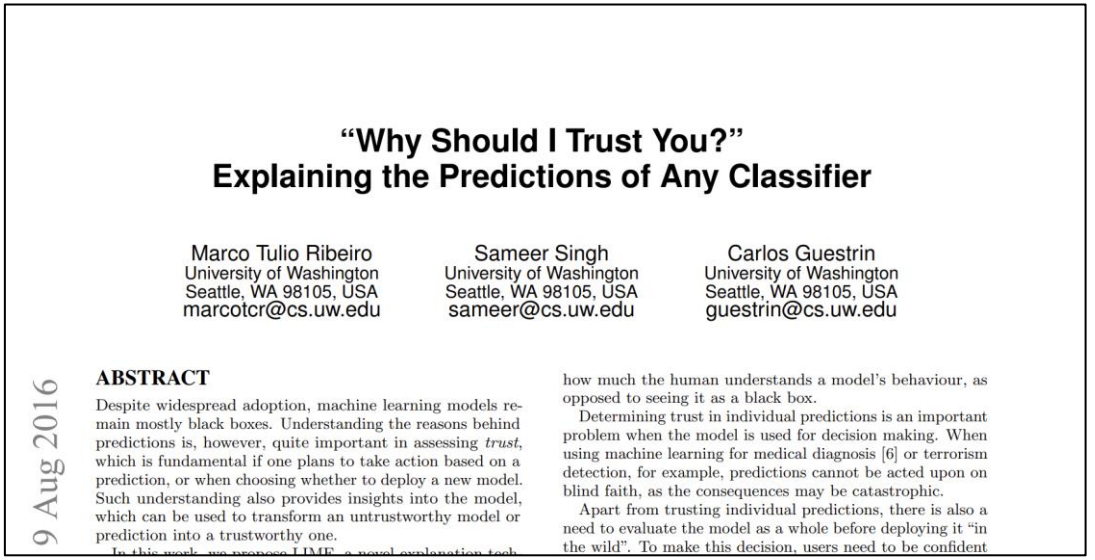
Wie kann man KI erklären?

Zielkonflikt zwischen Erklärbarkeit und Vorhersagekraft



Black box model





“Why Should I Trust You?”
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation tech-

how much the human understands a model's behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

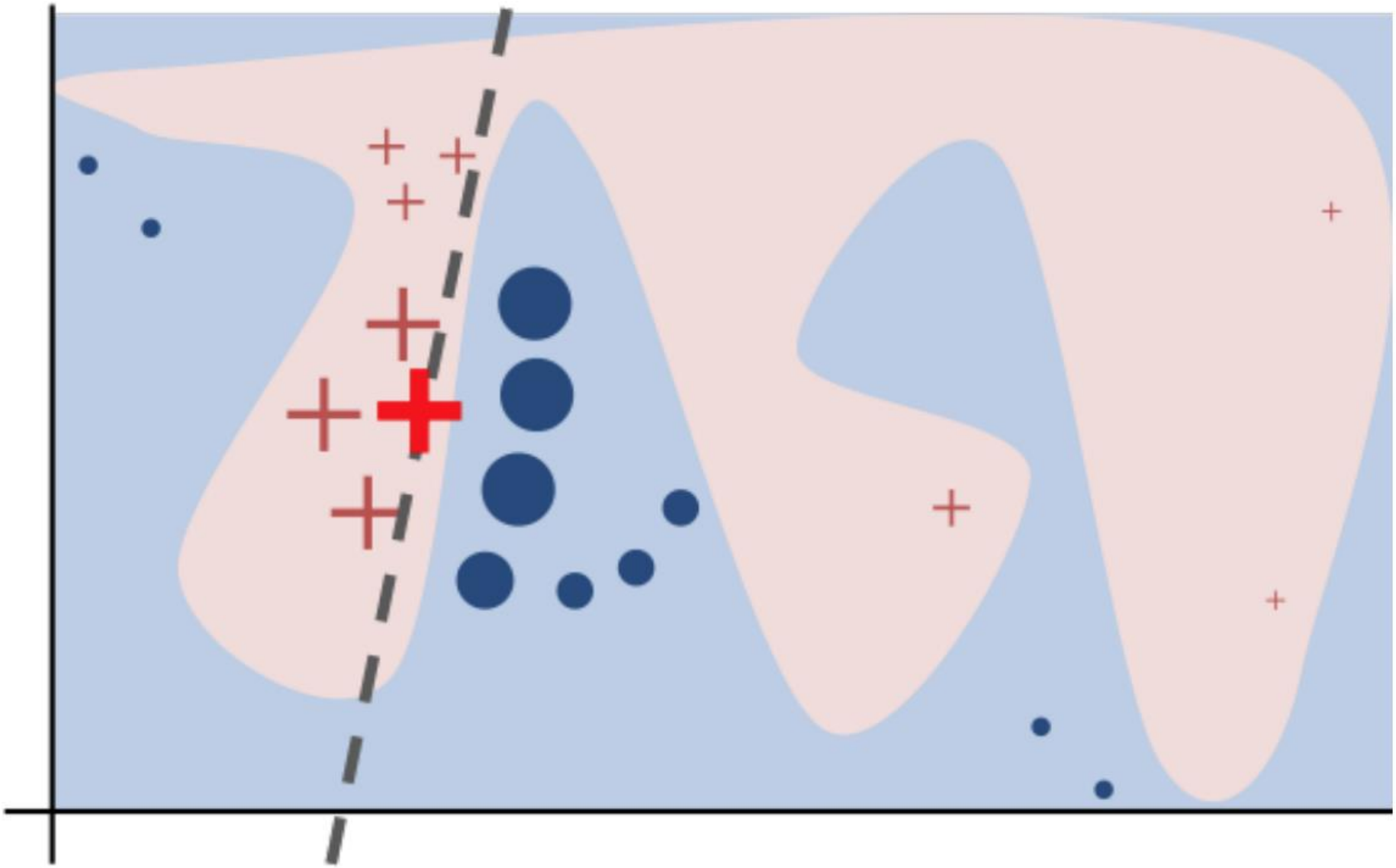
Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it “in the wild”. To make this decision, users need to be confident

LIME

Local Interpretable Model-agnostic Explanations

LIME

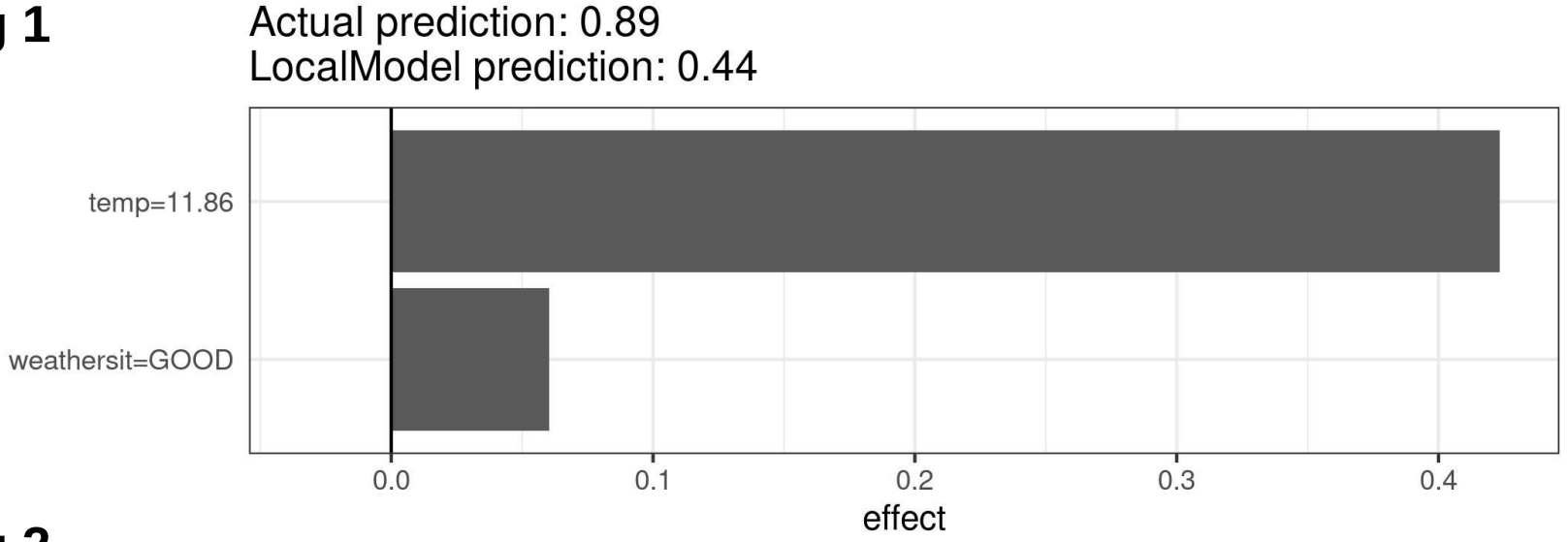
Local interpretable model-agnostic explanations [Ribeiro, 2016]



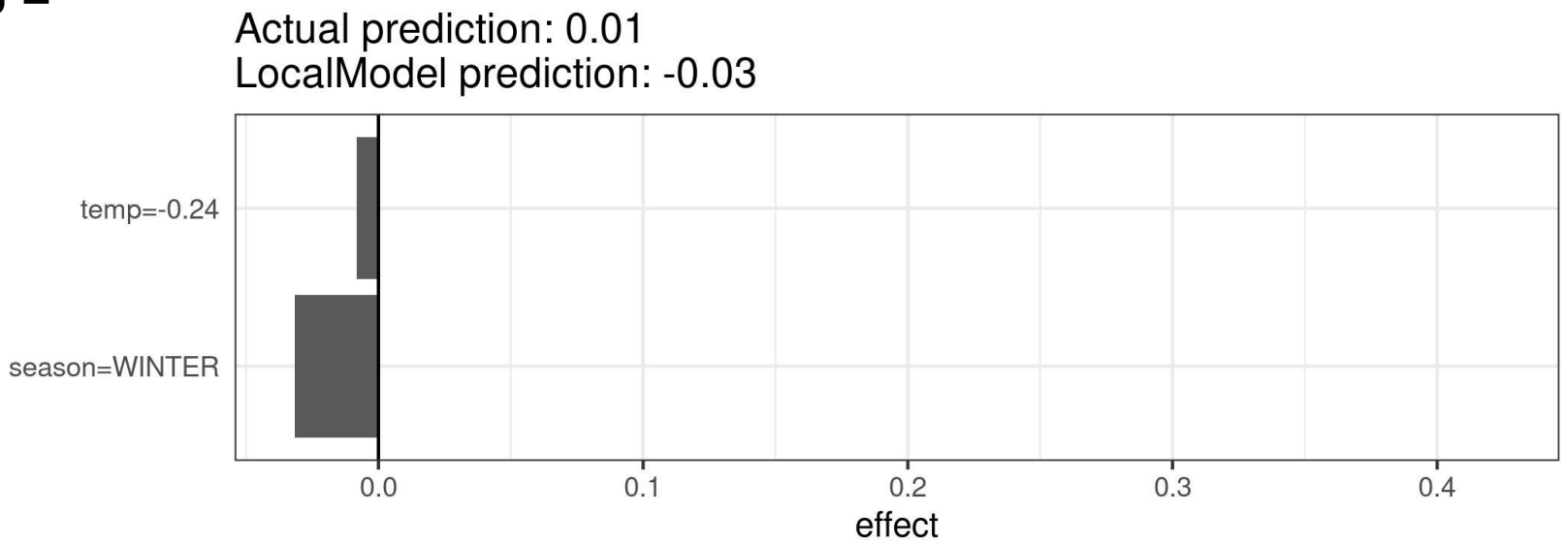
LIME: Beispiel für tabellarische Daten

Fahrradverleih: Vorhersage, ob viele oder wenig Fahrräder verliehen werden

Tag 1



Tag 2



LIME: Beispiel für Textdaten

Dokumentenklassifizierung

y=sci.med (probability **0.989**, score **3.945**) top features

Contribution?	Feature
+8.958	Highlighted in text (sum)
-5.013	<BIAS>

from: brian@ucsd.edu (brian kantor) subject: re: help for kidney stones ..... organization: the avant-garde of the now, ltd. lines: 12 nntp-posting-host: ucsd.edu as i recall from my bout with kidney stones, there isn't any medication that can do anything about them except relieve the pain.

y=alt.atheism (probability **0.000**, score **-8.709**) top features

Contribution?	Feature
+1.743	Highlighted in text (sum)
-10.453	<BIAS>

from: brian@ucsd.edu (brian kantor) subject: re: help for kidney stones ..... organization: the avant-garde of the now, ltd. lines: 12 nntp-posting-host: ucsd.edu as i recall from my bout with kidney stones, there isn't any medication that can do anything about them except relieve the pain.

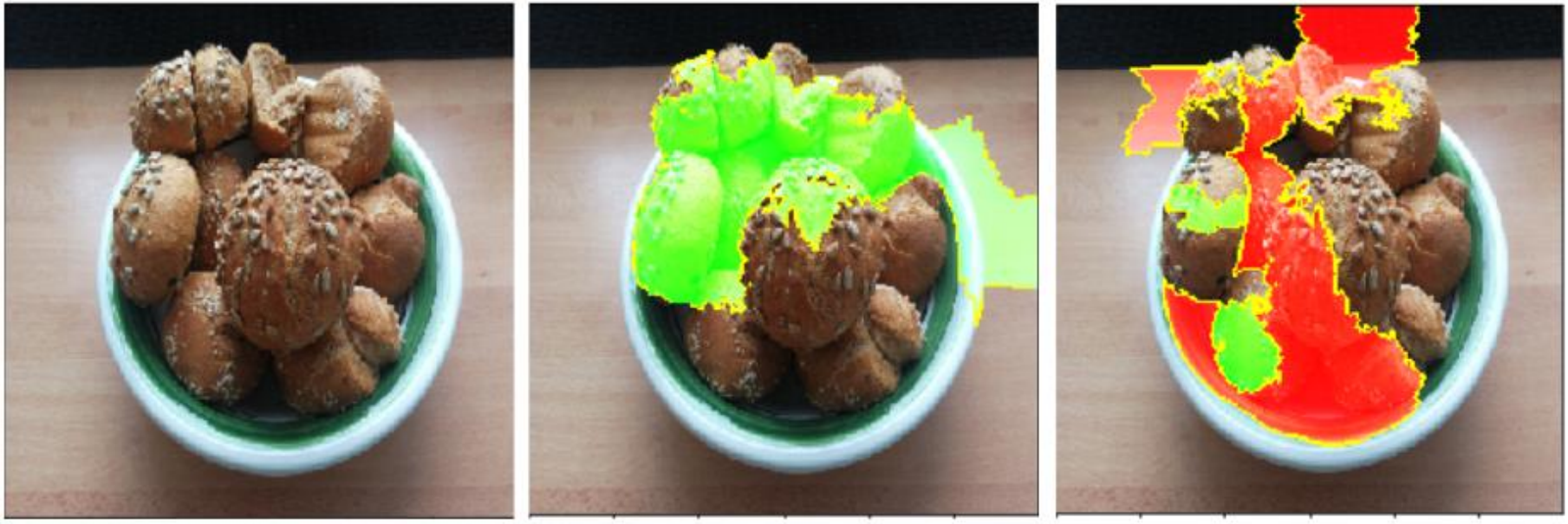
Source: https://github.com/TeamHG-Memex/eli5/blob/master/notebooks/Debugging%20scikit-learn%20text%20classification%20pipeline.ipynb

LIME: Beispiel für Bildverarbeitung for images

Bildklassifikation

Bagel? (77%)

Erdbeere? (4%)



LIME

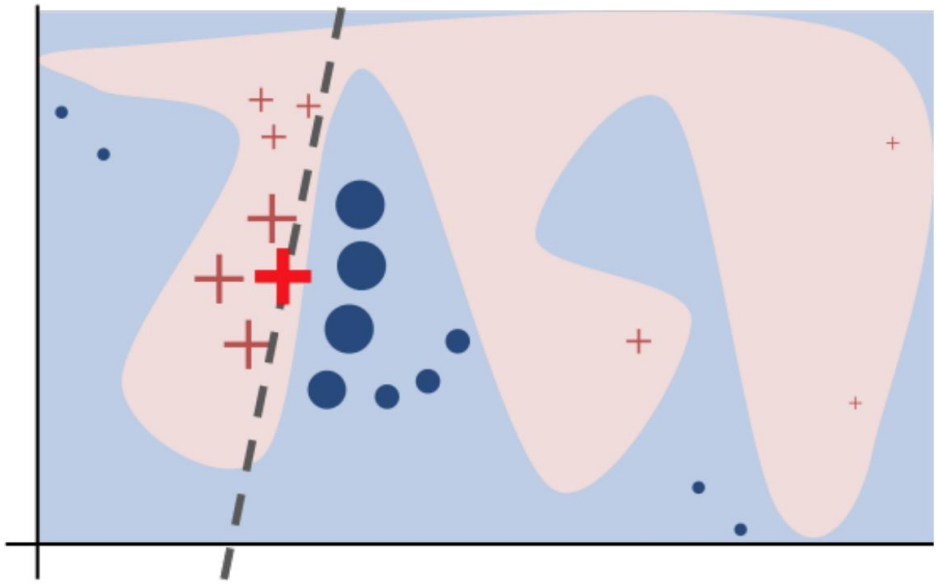
Vorteile & Nachteile

Vorteile

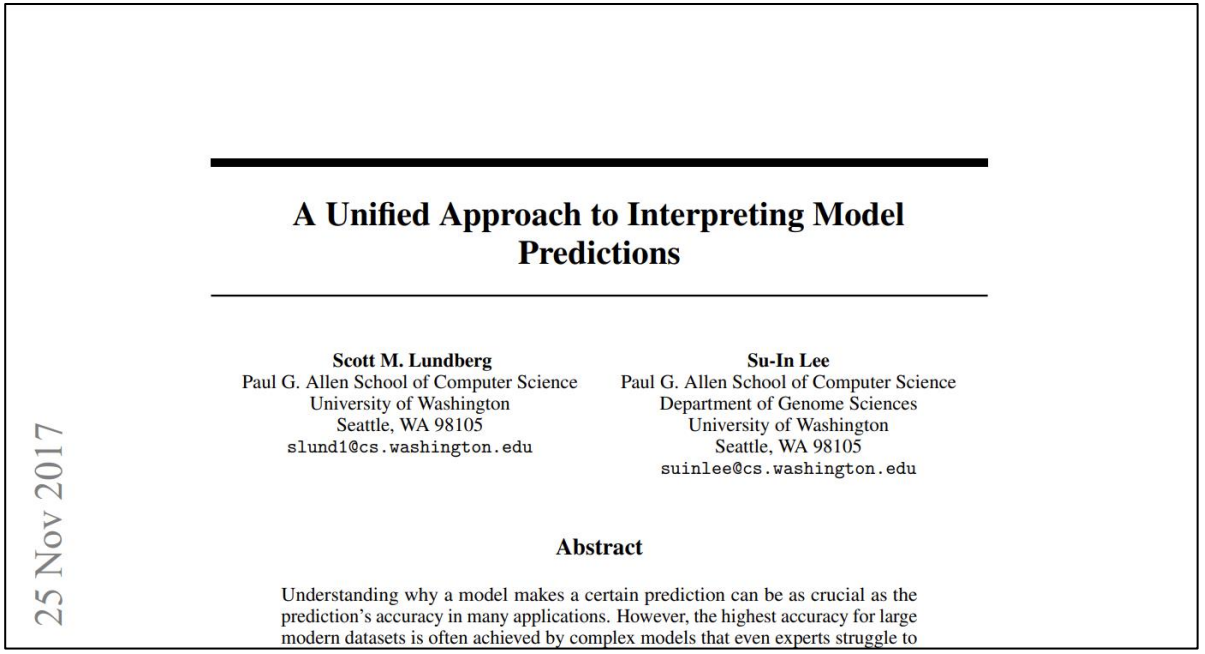
- Erzeugt kurze, benutzerfreundliche Erklärungen
- LIME funktioniert für tabellarische Daten, Text und Bilder

Nachteile

- “Korrekte” Definition der Nachbarschaft ist schwierig
- “Korrekte” Stichprobenauswahl ist schwierig



Lloyd Shapley: Nobelpreis für Wirtschaftswissen schaften 2012



A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

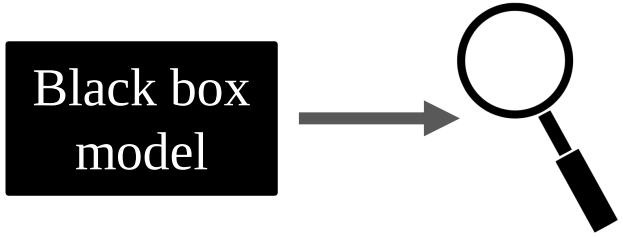
Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to

SHAP

Shapley Values



Shapley-Wert

Einführung

Shapley-Werte stammen aus der kooperativen Spieltheorie

Ein kooperatives Spiel ist definiert als

- Menge an Spielern: {a, b, c, …}
- Einer Funktion val, die kooperierenden Spielern einen Wert zuordnet

Beispiel: 3 Spieler

- $val(\{\}) = 0$
- $val(\{a\}) = 300$
- $val(\{b\}) = 300$
- $val(\{c\}) = 300$
- $val(\{a, b\}) = 700$
- $val(\{a, c\}) = 500$
- $val(\{b, c\}) = 400$
- $val(\{a, b, c\}) = 1{,}000$

Fragestellung

- Wie hoch ist der Beitrag von Spieler $\mathbf{a}$ zur Koalition $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$?

Shapley-Werte

Einführung

Idee

- Spieler betreten nach und nach den Raum
- Für jede mögliche Reihenfolge, berechne den Beitrag von Spieler a

Beispiel

- $[a, b, c]$: $val(\{a\}) - val(\{\}) = 300 - 0 = 300$
- $[a, c, b]$: $val(\{a\}) - val(\{\}) = 300 - 0 = 300$
- $[c, a, b]$: $val(\{a, c\}) - val(\{c\}) = 500 - 300 = 200$
- $[b, a, c]$: $val(\{a, b\}) - val(\{b\}) = 700 - 300 = 400$
- $[b, c, a]$: $val(\{a, b, c\}) - val(\{b, c\}) = 1{,}000 - 400 = 600$
- $[c, b, a]$: $val(\{a, b, c\}) - val(\{b, c\}) = 1{,}000 - 400 = 600$

- Durchschnittlicher Beitrag $2400 / 6 = 400$

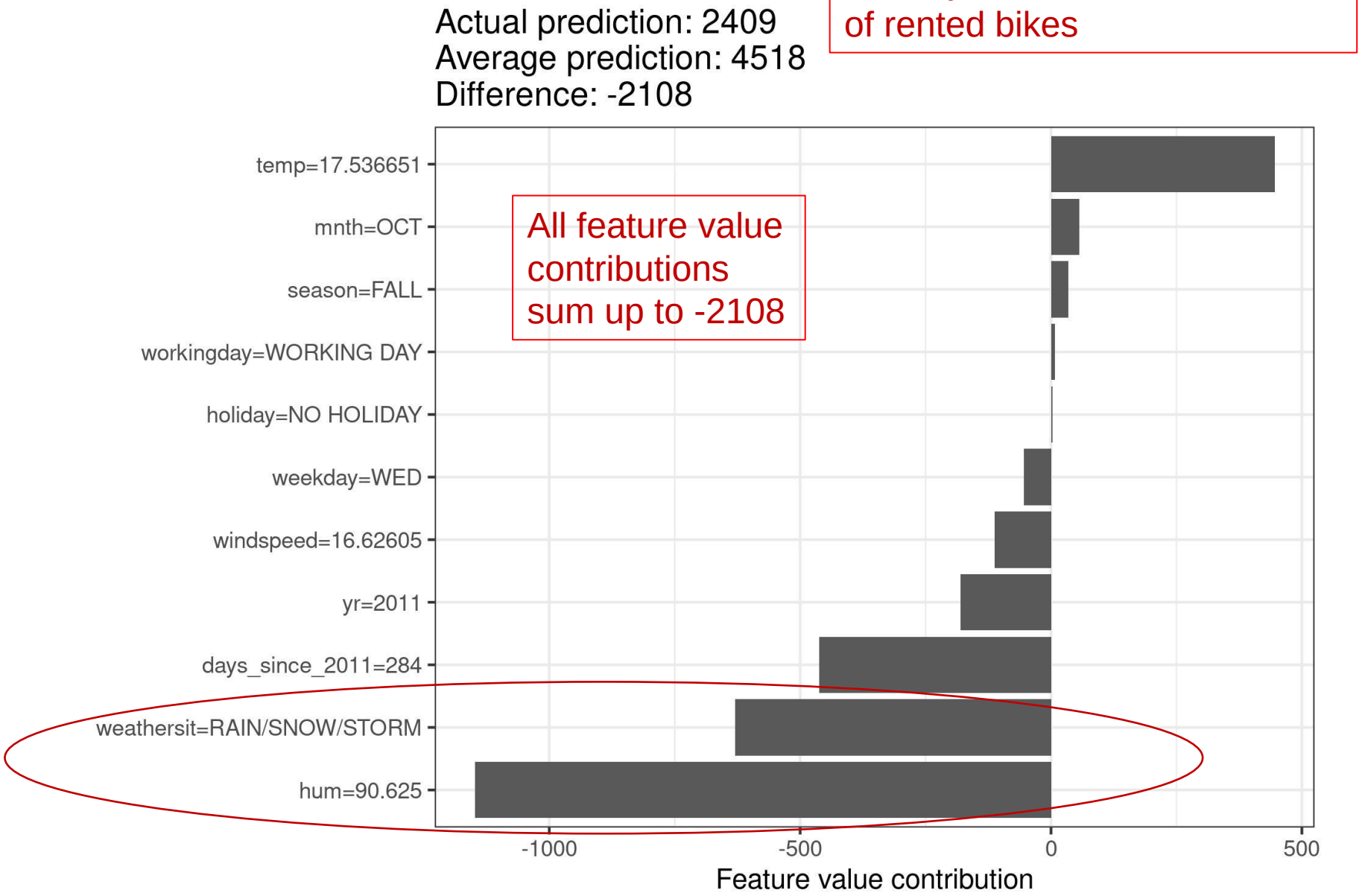
➢ Shapley-Wert von Spieler a 400

Maschinelles Lernen: Spieler entsprechen Merkmalen einer Instanz

Beispiel und Interpretation

Fahrradverleih

For day 285. Predicted number of rented bikes



Shapley-Werte
Vorteile & Nachteile

Vorteile

- Erzeugt eine vollständige Erklärung (Aufsummierung aller Shapley-Werte)
- Basiert auf solider Theorie

Nachteile

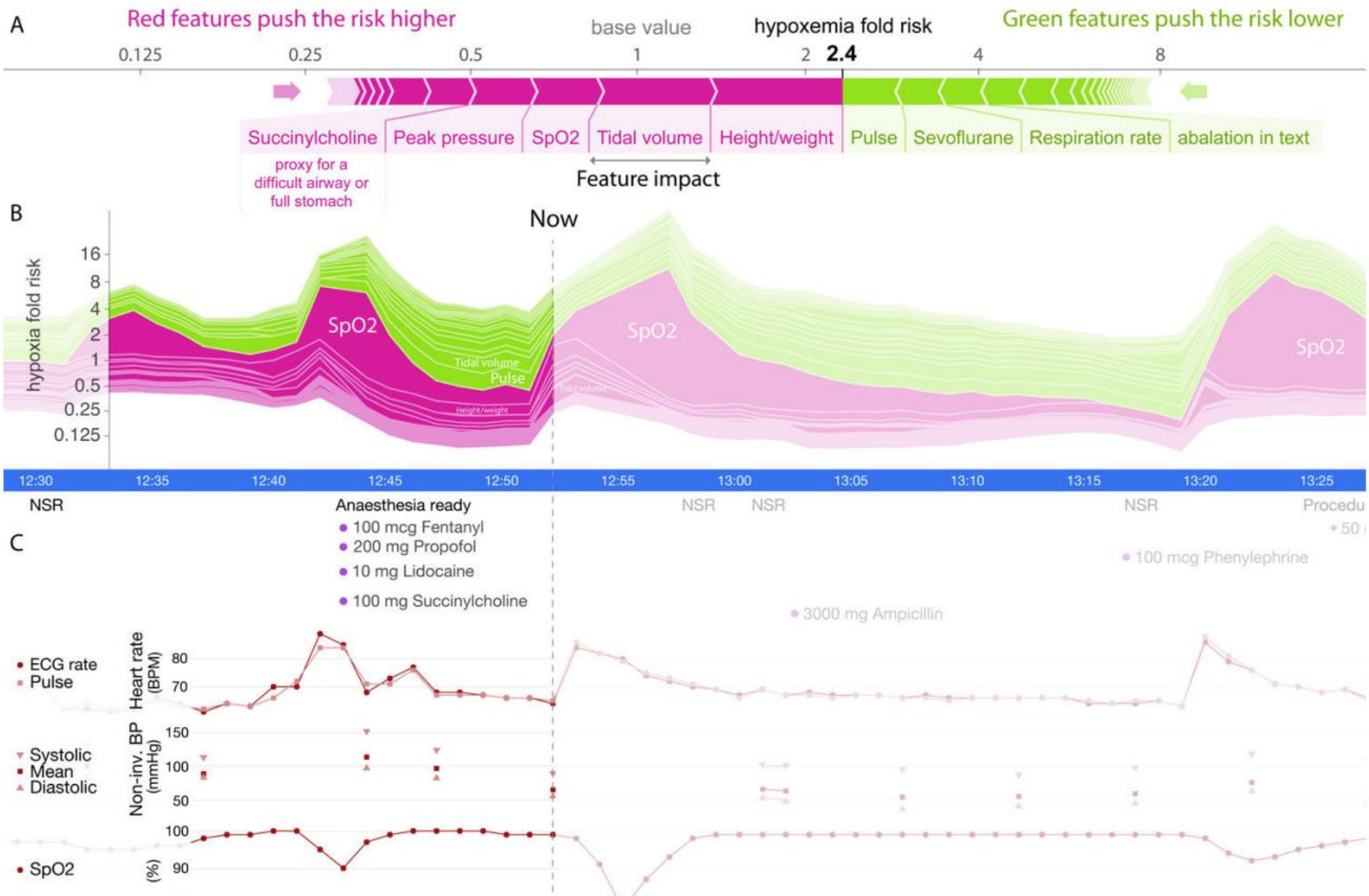
- Genaue Berechnung der Shapley-Werte erfordert viel Rechenzeit

➢ Approximationsalgorithmus erforderlich

Anwendungen

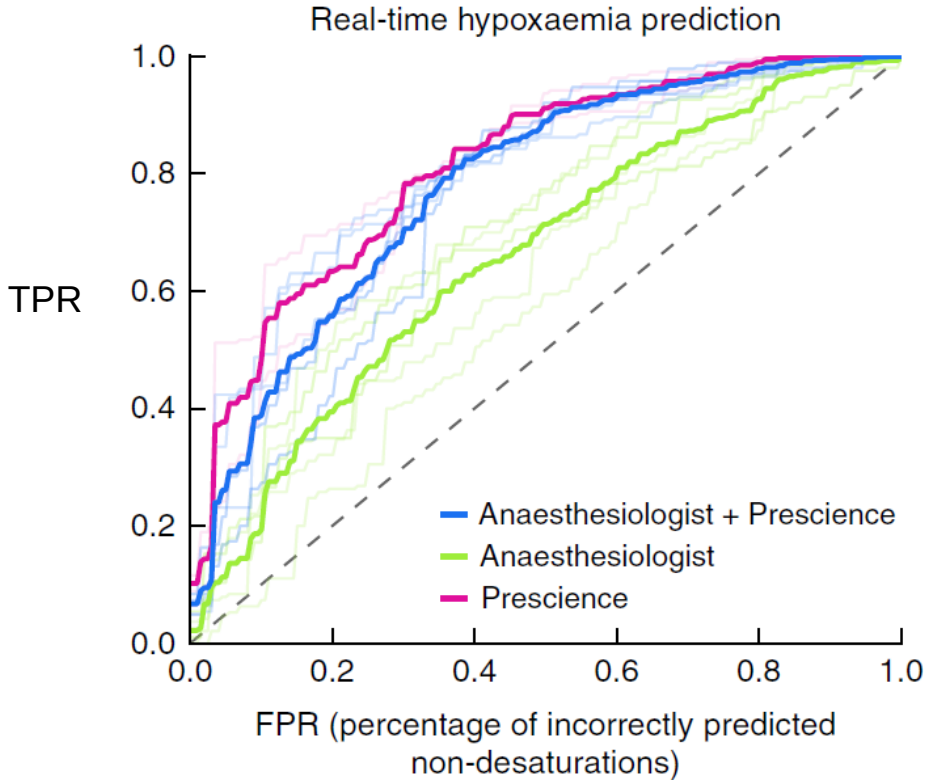
Vorhersage von Sauerstoffmangel während einer Operation

[Lundberg et al. 2018]



Vorteile und Ergebnisse

[Lundberg et al. 2018]



- Ohne Unterstützung: Anesthesisten konnten 15% der Problemfälle voraussehen
- Mit Unterstützung: Anesthesisten konnten 30% der Problemfälle voraussehen

➢ Über 2 Millionen zusätzliche Fälle pro Jahr könnten in den USA erkannt werden

➢ Etwa 20 % von ihnen würden davon profitieren (Änderungen der Medikation)

Vorhersage von Kreditrisiken

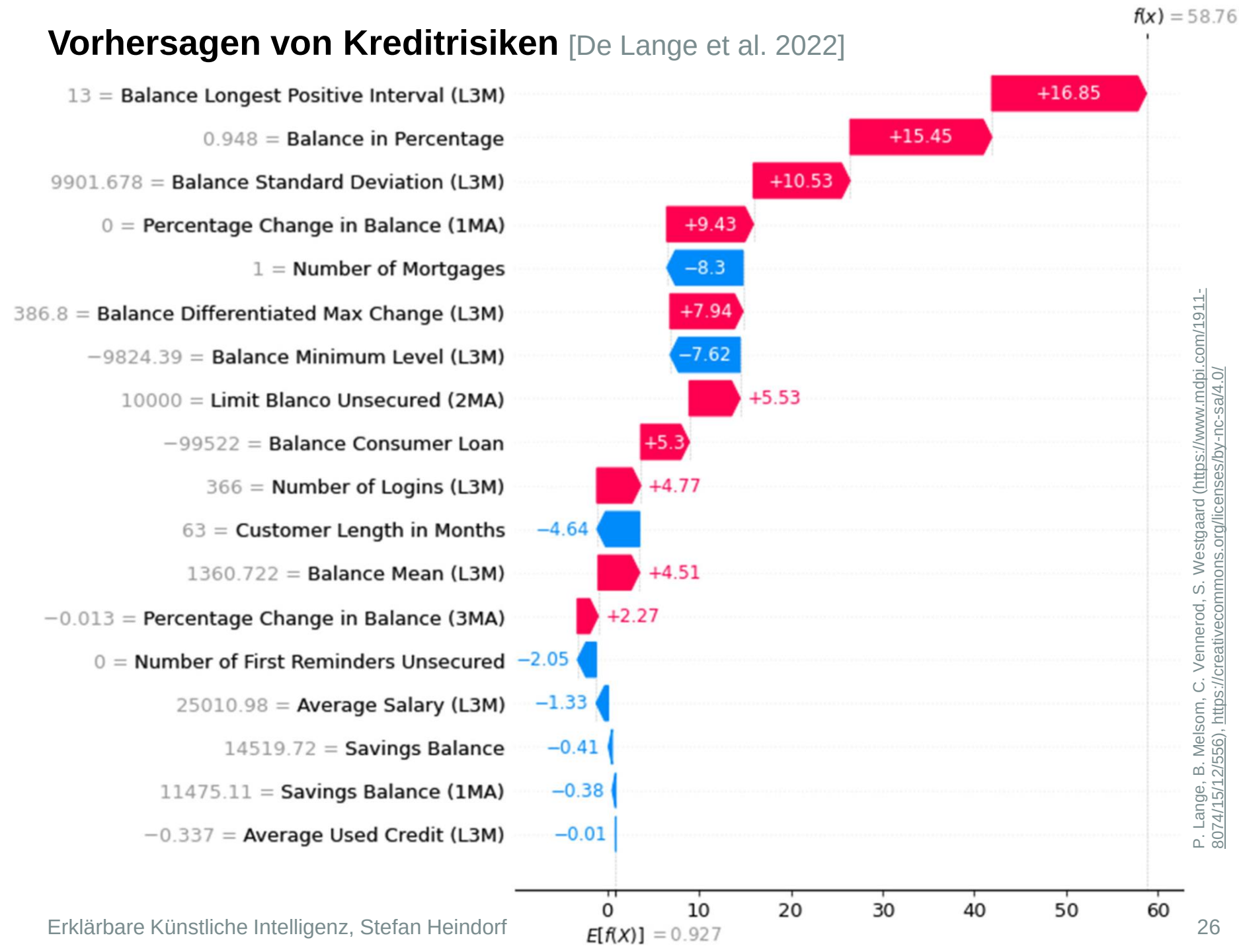
Derzeit verwenden die meisten Banken lineare Modelle

- Vorschriften, die Transparenz und Erklärbarkeit verlangen
- Finanzielle Risiken
- Vertrauen der Kunden

Datensatz

- Norwegische Bank mittlerer Größe
- Verbraucherkredite von 14,000 Kunden über 4 Jahre
- Aufgabe: Vorhersage von Zahlungsverzug in nächsten12 Monaten

Vorhersagen von Kreditrisiken [De Lange et al. 2022]



Vorhersagen von Kreditrisiken

[De Lange et al. 2022]

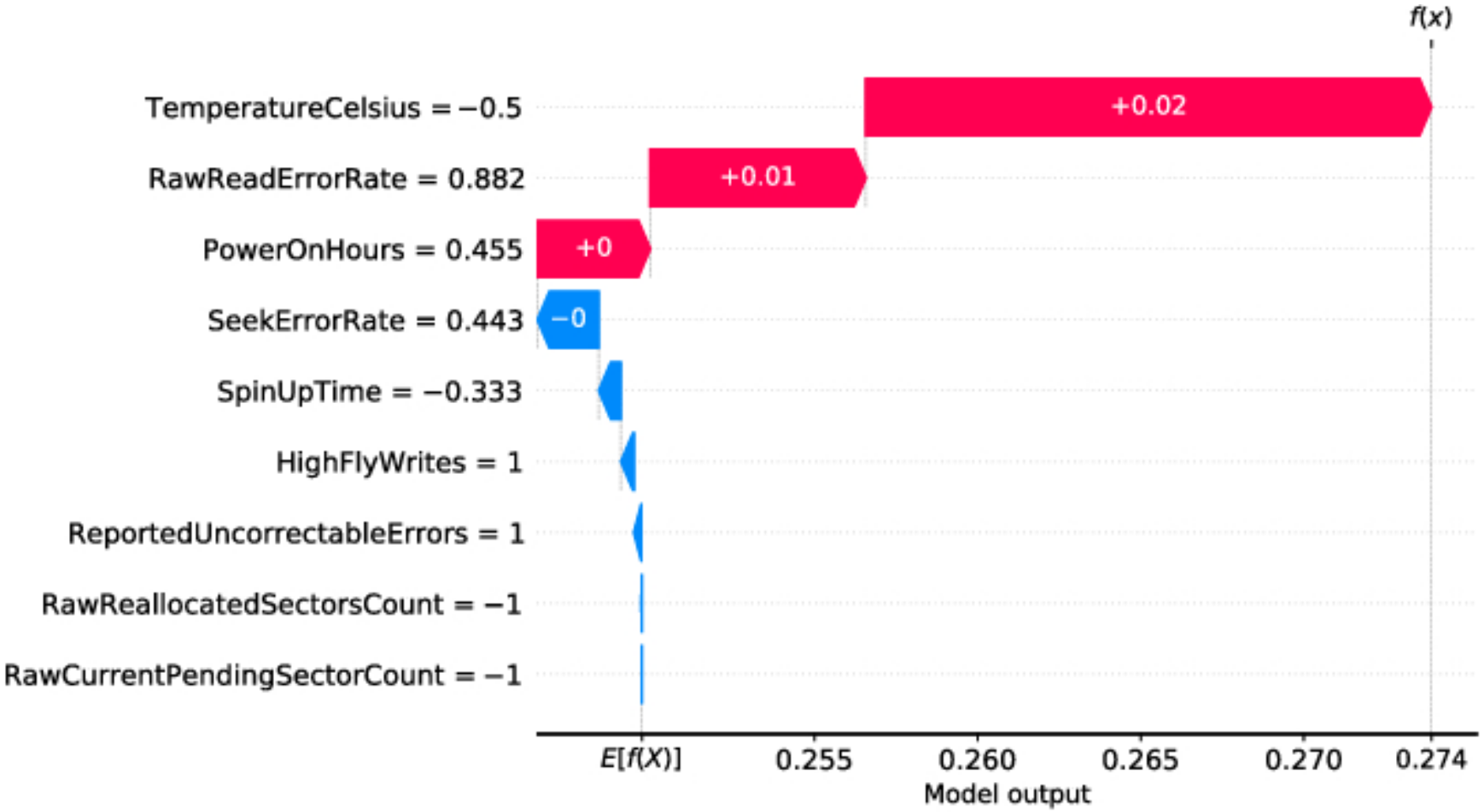
Evaluation

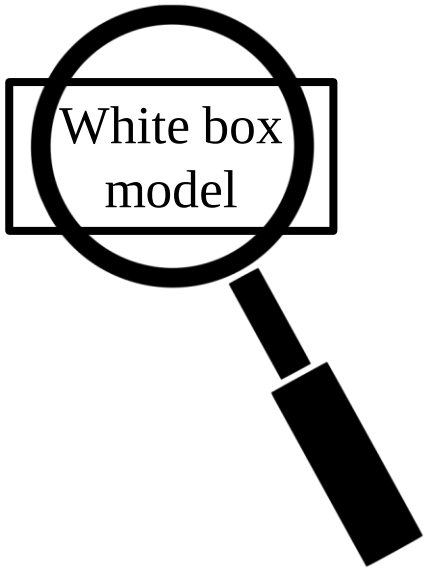
- Problem 1: Kredit wird nicht gewährt, obwohl die Person nicht in Verzug geraten würde
- Problem 2: Kredit wird gewährt, aber Person gerät in Zahlungsverzug

➢ Gesamt: Black-Box (LightGBM) übertrifft lineare Modelle deutlich: Die Bank würde etwa 20 - 30 Millionen NOK (1,7 - 2,6 Millionen EUR) zusätzlich verdienen

Vorhersage von Festplattenausfällen

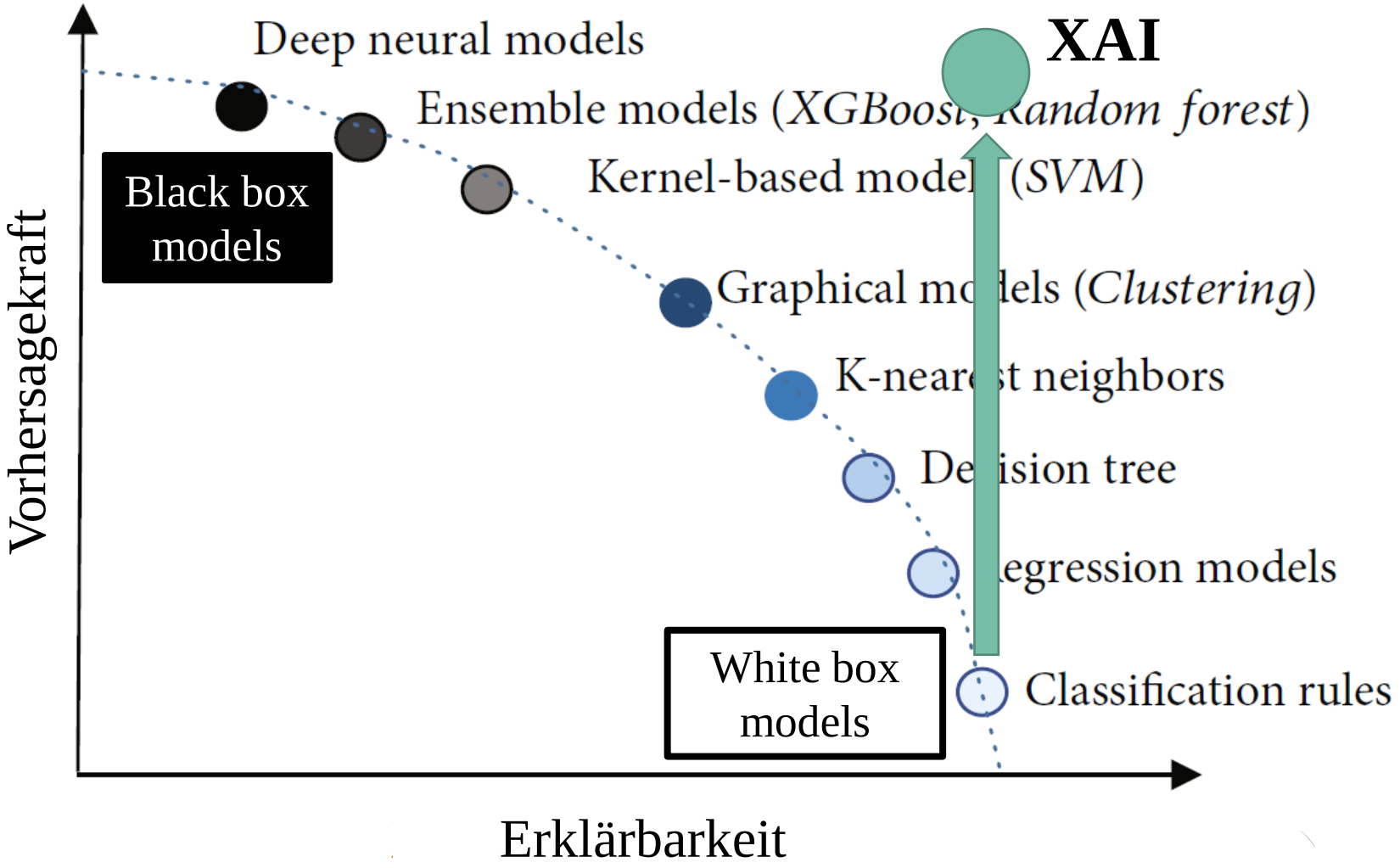
[Ferraro et al. 2023]





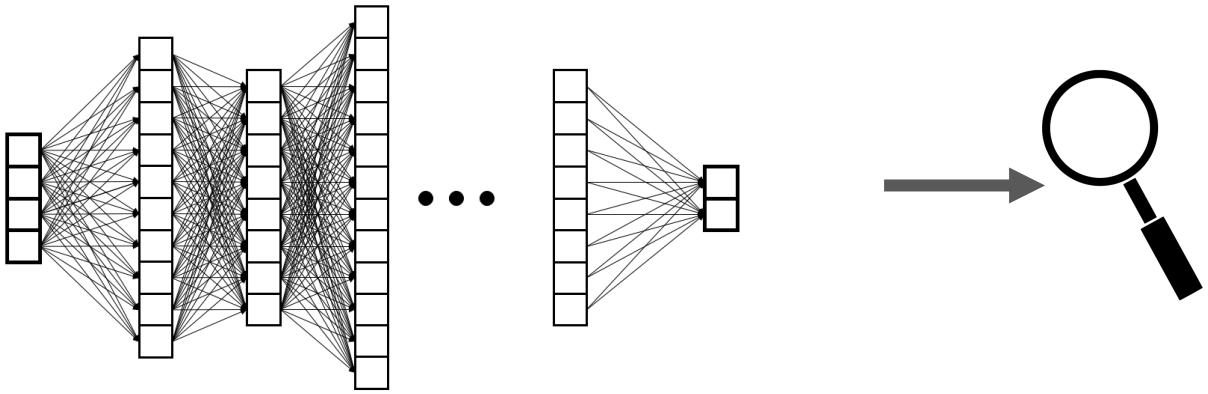
White-Box-Modelle

Wie kann man KI erklären?
Zielkonflikt zwischen Erklärbarkeit und Vorhersagekraft



Black-Box- vs. White-Box-Modelle

Black-Box Modell (z. B. tiefes neuronales Netz)



White-Box Model (z. B. Klassifizierungsregeln)

```
IF       age 18-20 and male                THEN predict arrest
ELSE IF  age 21-23 and 2-3 prior offense   THEN predict arrest
ELSE IF  3+ prior offenses                 THEN predict arrest
ELSE                                            predict no arrest
```

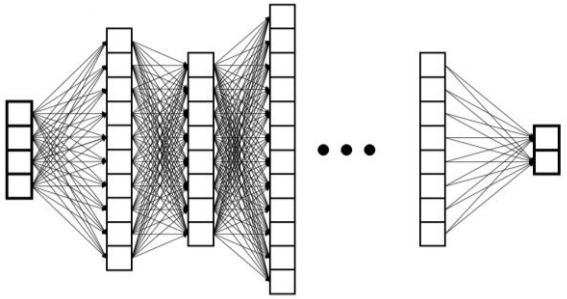
Black-Box- vs. White-Box-Modelle

[Rudin 2019]

Aufgabe: Vorhersage der Rückfälligkeit von Straftätern

Black-Box-Modell
(COMPAS)

- 130+ Merkmale
- Teure Lizenz



White-Box-Modell
(CORELS)

- 3 Merkmale (Alter, Vorstrafen, Geschlecht)
- Kostenlos

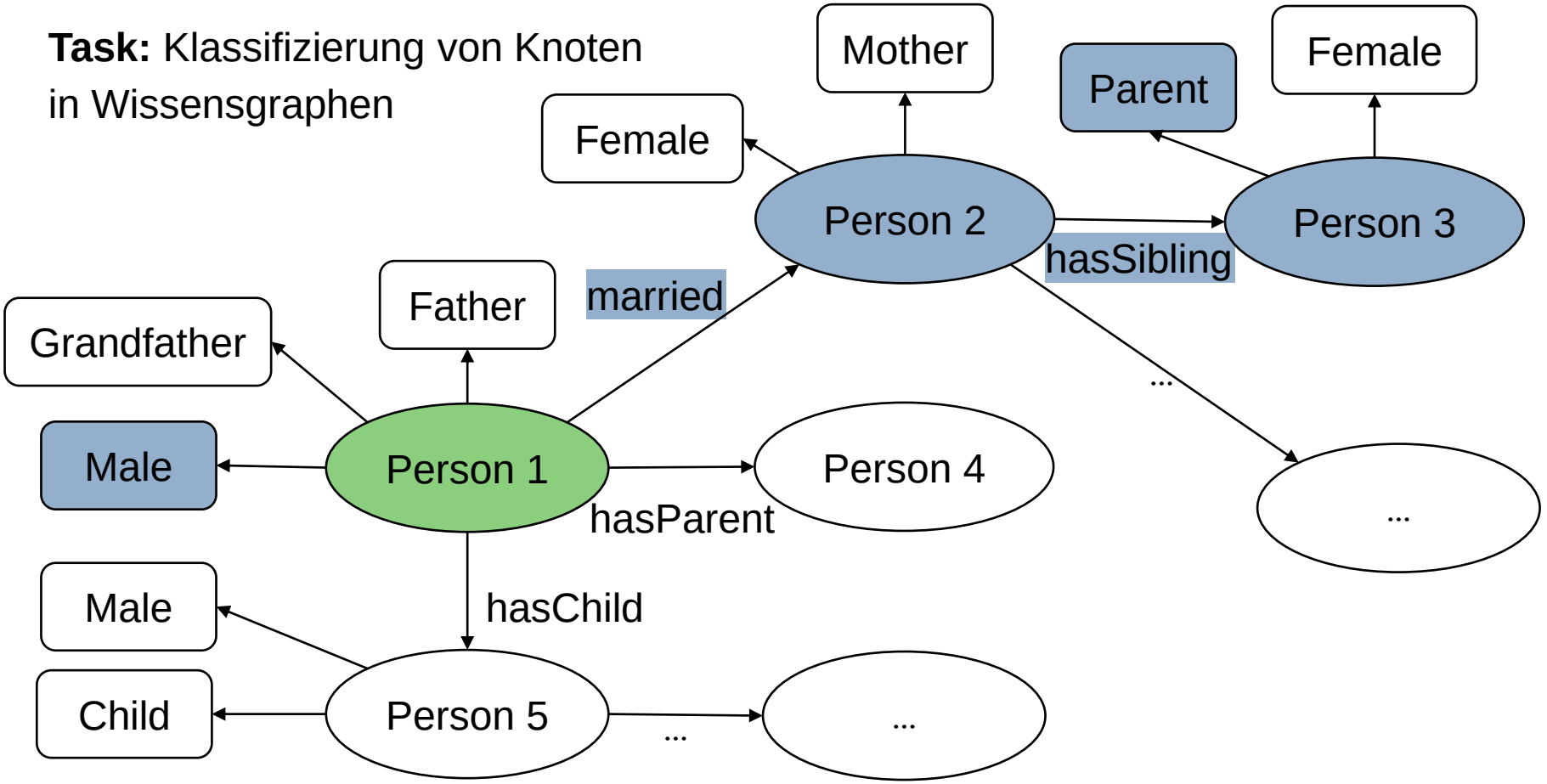
```
IF       age 18-20 and male               THEN predict arrest
ELSE IF  age 21-23 and 2-3 prior offense  THEN predict arrest
ELSE IF  3+ prior offenses                THEN predict arrest
ELSE                                           predict no arrest
```

→ Ungefähr gleiche Vorhersageleistung

EvoLearner: Konzeptlernen auf Wissensgraphen

[Heindorf et al. 2022]

Task: Klassifizierung von Knoten in Wissensgraphen



Konzept eines Onkels:

$$\texttt{Male} \sqcap ((\exists \texttt{married}.\exists \texttt{hasSibling}.\texttt{Parent}) \sqcup \exists \texttt{hasSibling}.\texttt{Parent}))$$

Plädoyer für White-Box-Modelle [Rudin 2019]

Cynthia Rudin

Probleme der Black-Box-Erklärverfahren

- Erklärungen spiegeln nicht 100% das Verhalten des Black-Box-Modells wieder
- Auch wenn nur manche Erklärungen falsch sind, gefährdet dies das Vertrauen
- Unternehmen können mit Black-Boxen Gewinne erwirtschaften und haben keinen Anreiz für White-Box-Modelle

Oft gibt es ein ebenso gutes White-Box-Modell. Insbesondere:

- Für tabellarische Daten
- Bei aufwendiger Identifikation guter Merkmale (Feature-Engineering)
- Für Entscheidungen, bei denen viel auf dem Spiel steht

Lösungsvorschläge

- Black-Box-Modelle verbieten, wenn gleichgute White-Box-Modelle existieren
- Verpflichtung, die Leistung von Black-Box-Modelle zusammen mit der Leistung von White-Box-Modells zu veröffentlichen

Erklärbare KI an der Universität Paderborn

Ausgewählte Projekte

CONSTRUCTING EXPLAINABILITY

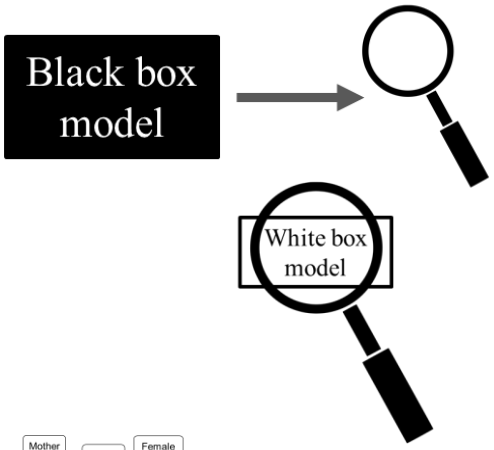
Data Science Junior Research Group

- Erklärbare KI für Wissensgraphen
- Erklärung von Black-Box Modellen
- Erklärung von White-Box Modellen
- Kontakt: heindorf@uni-paderborn.de

Zusammenfassung & weitere Informationen

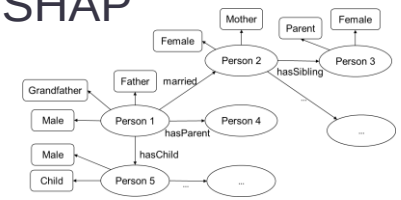
Zusammenfassung

- Bedarf der Erklärbarkeit
- Erklärung von Black-Box-Modellen: LIME, SHAP
- White-Box-Modelle: Klassifikationsregeln



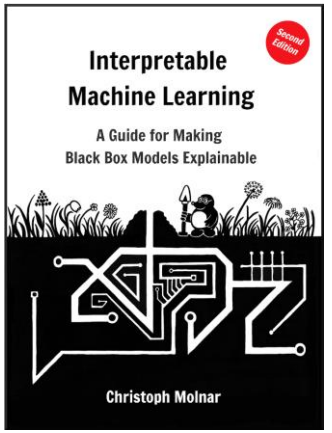
Ausblick

- Erklärungsmethoden jenseits von LIME und SHAP
- Erklärungsmethoden für Graphen
- Benutzerfreundliche Erklärungen



Weitere Informationen

- Buch
  - Interpretable Machine Learning (Molnar 2020) https://christophm.github.io/interpretable-ml-book/
- Code
  - LIME: https://github.com/marcotcr/lime
  - SHAP: https://github.com/shap/shap
  - EvoLearner: https://github.com/dice-group/Ontolearn



Referenzen

- **Arrieta**, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information fusion 58* (**2020**): 82-115.
- **De Lange**, Petter Eilif, Borger Melsom, Christian Bakke Vennerød, and Sjur Westgaard. "Explainable AI for Credit Assessment in Banks." *Journal of Risk and Financial Management* 15, no. 12 (**2022**): 556.
- **Doshi-Velez**, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." *arXiv preprint arXiv:1702.08608* (**2017**).
- **Ferraro**, Antonino, Antonio Galli, Vincenzo Moscato, and Giancarlo Sperlì. "Evaluating eXplainable artificial intelligence tools for hard disk drive predictive maintenance." *Artificial Intelligence Review* 56, no. 7 (**2023**): 7279-7314.
- **Heindorf**, Stefan, Lukas Blübaum, Nick Düsterhus, Till Werner, Varun Nandkumar Golani, Caglar Demir, and Axel-Cyrille Ngonga Ngomo. "Evolearner: Learning description logics with evolutionary algorithms." In *Proceedings of the ACM Web Conference 2022*, pp. 818-828. **2022**.
- **Lundberg**, Scott M., Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston et al. "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery." *Nature biomedical engineering* 2, no. 10 (**2018**): 749-760.

Referenzen

- **Molnar**, Christoph. *Interpretable machine learning*. **2020**.
- **Ribeiro**, Marco Tulio, Sameer Singh, and Carlos Guestrin. ""Why should I trust you?" Explaining the predictions of any classifier." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135-1144. **2016.**
- **Rudin**, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature machine intelligence* 1, no. 5 (**2019**): 206-215.
- **Wang**, Maonan, Kangfeng Zheng, Yanqing Yang, and Xiujuan Wang. "An explainable machine learning framework for intrusion detection systems." *IEEE Access* 8 (**2020**): 73127-73141.