Vorabfragen

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Deep Learning, Machine Learning und Künstliche Intelligenz – SS 26

Gelesen von Prof. Dr. Daniel Gaida

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik

Lernraum III

Explainable AI (XAI) – Erklärbare KI

- Warum möchten wir, dass uns eine KI ihre Entscheidung erklärt?
  - Für wen ist die Erklärung?
- Was ist Erklärbare KI?

Lernraum III

Explainable AI (XAI) – Erklärbare KI

- Warum möchten wir, dass uns eine KI ihre Entscheidung erklärt?
  - Für wen ist die Erklärung?
- Was ist Erklärbare KI?

- Intrinsisch erklärbare Machine Learning Modelle
- Post-hoc-Methoden
  - LIME
  - SHAP

-

Lernraum III

Explainable AI (XAI) – Erklärbare KI

- Warum möchten wir, dass uns eine KI ihre Entscheidung erklärt?
  - Für wen ist die Erklärung?
- Was ist Erklärbare KI?

- Intrinsisch erklärbare Machine Learning Modelle
- Post-hoc-Methoden
  - LIME
  - SHAP

- Erklärungen heutiger Systeme

Lernziele von heute und Fragen zur Überprüfung der Lernziele

WAS:

Die Studierenden können einfache XAI-Methoden mit Hintergrundwissen nutzen und bewerten.

Lernziele von heute und Fragen zur Überprüfung der Lernziele

WAS:

Die Studierenden können einfache XAI-Methoden mit Hintergrundwissen nutzen und bewerten.

WOMIT:

Dies geschieht durch die Vorstellung typischer Anwendungsfelder für Explainability, den Vergleich intrinsischer und post-hoc Verfahren (z. B. Entscheidungsbaum vs. LIME) und die Analyse derer Ergebnisse.

Lernziele von heute und Fragen zur Überprüfung der Lernziele

WAS:
Die Studierenden können einfache XAI-Methoden mit Hintergrundwissen nutzen und bewerten.

WOMIT:
Dies geschieht durch die Vorstellung typischer Anwendungsfelder für Explainability, den Vergleich intrinsischer und post-hoc Verfahren (z. B. Entscheidungsbaum vs. LIME) und die Analyse derer Ergebnisse.

WOZU:
Die Studierenden sind dadurch in der Lage, die Auswahl und Anwendung von Erklärungen als Teil menschenzentrierter KI-Systeme kritisch zu gestalten und für unterschiedliche Zielgruppen angemessene Erklärungen zu identifizieren.

Lernziele von heute und Fragen zur Überprüfung der Lernziele

WAS:
Die Studierenden können einfache XAI-Methoden mit Hintergrundwissen nutzen und bewerten.

WOMIT:
Dies geschieht durch die Vorstellung typischer Anwendungsfelder für Explainability, den Vergleich intrinsischer und post-hoc Verfahren (z. B. Entscheidungsbaum vs. LIME) und die Analyse derer Ergebnisse.

WOZU:
Die Studierenden sind dadurch in der Lage, die Auswahl und Anwendung von Erklärungen als Teil menschenzentrierter KI-Systeme kritisch zu gestalten und für unterschiedliche Zielgruppen angemessene Erklärungen zu identifizieren.

s. Übungsaufgaben von heute

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Warum möchten wir, dass uns eine KI ihre Entscheidung erklärt?

- Haben Sie schon mal eine Entscheidung durch ein KI-System erhalten, die Sie nicht verstanden haben?

Warum möchten wir, dass uns eine KI ihre Entscheidung erklärt?

- Haben Sie schon mal eine Entscheidung durch ein KI-System erhalten, die Sie nicht verstanden haben?

Beispiele zur Diskussion:

- ChatGPT liefert „falsche“ Antwort, aber sagt nicht warum
- Kredit-Scoring: Warum abgelehnt?
- Studienberatung: Empfehlung unklar
- YouTube/Spotify/Amazon/Netflix: Warum wird mir das empfohlen?

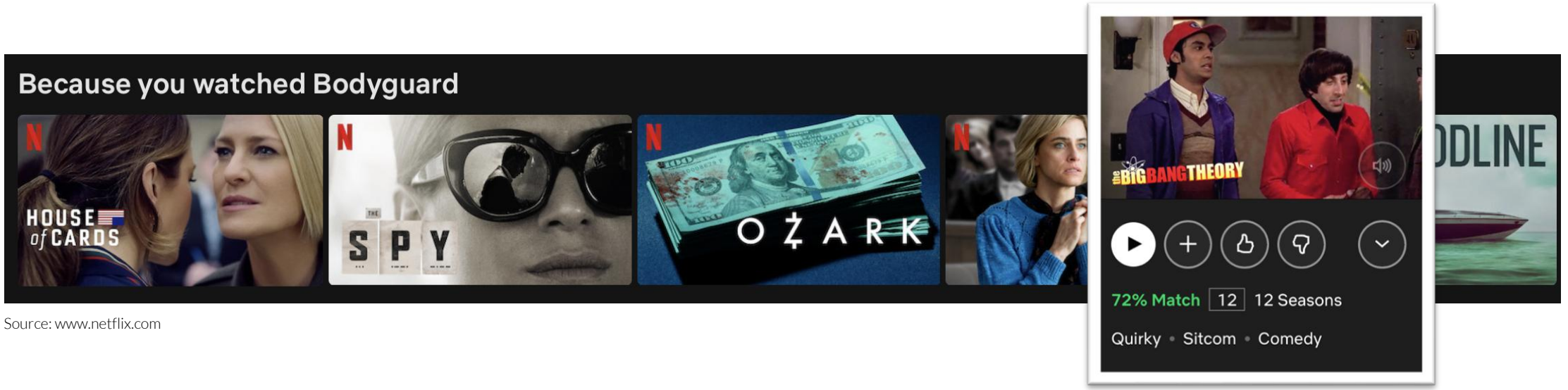
Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Warum möchten wir, dass uns eine KI ihre Entscheidung erklärt?

Diskussion

Wie erklärt Netflix, warum ein Film/eine Serie dem Nutzer empfohlen wird?

Glauben Sie, dass diese Erklärung den Nutzern hilft?

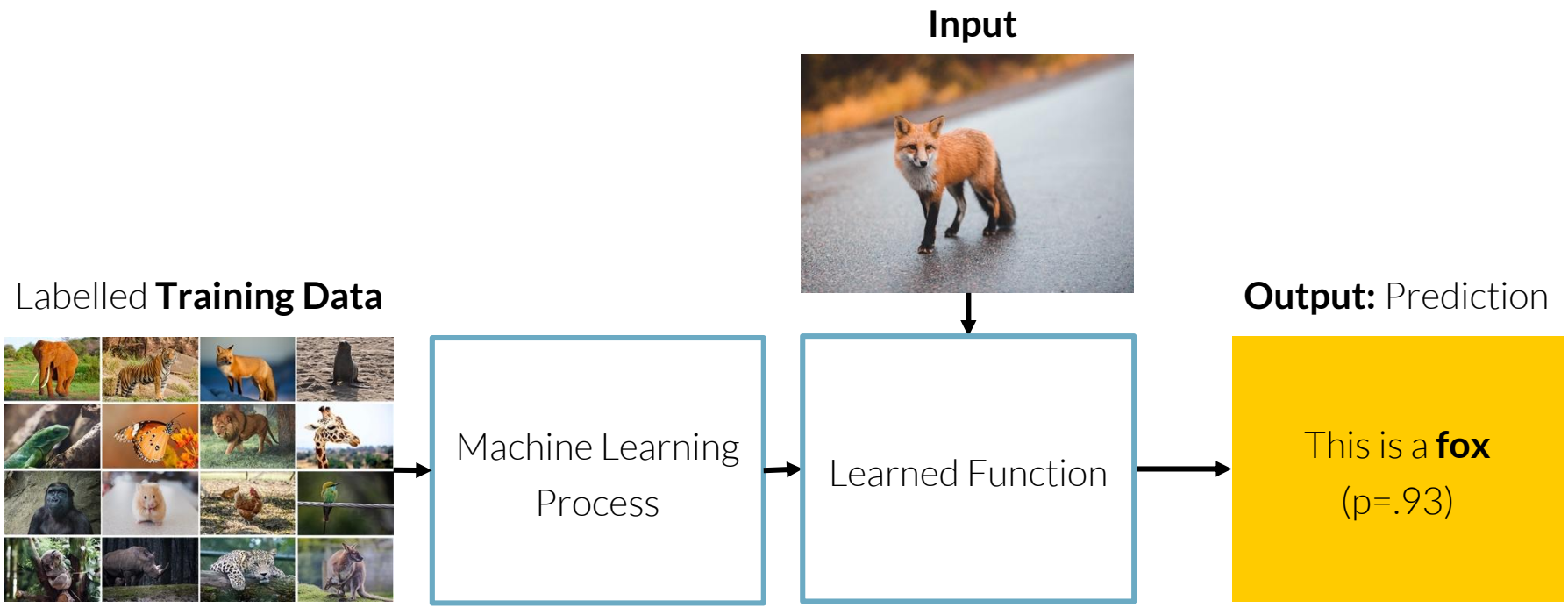


Source: www.netflix.com

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

The Black Box Problem of Machine Learning

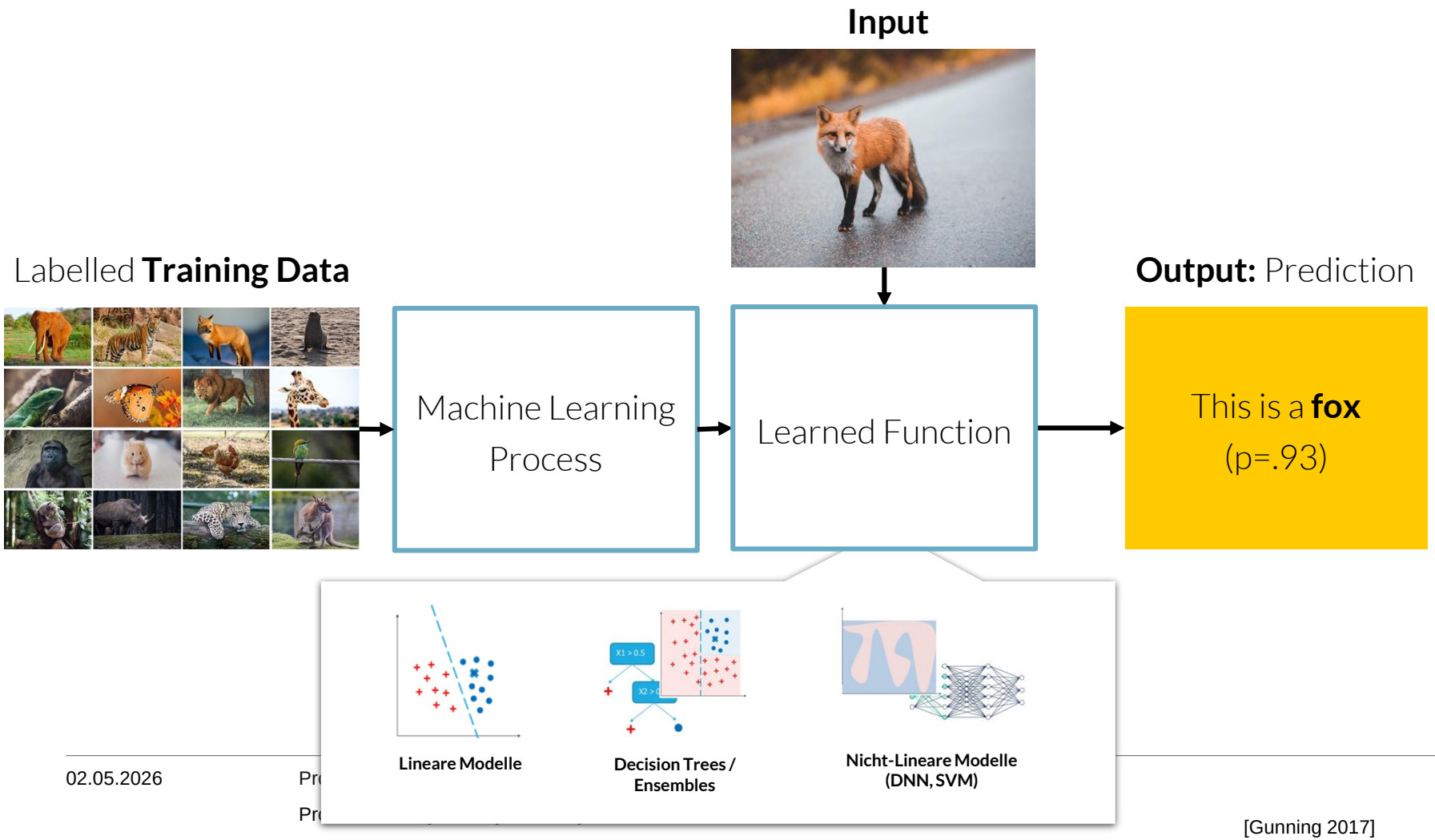
Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020



[Gunning 2017]

The Black Box Problem of Machine Learning

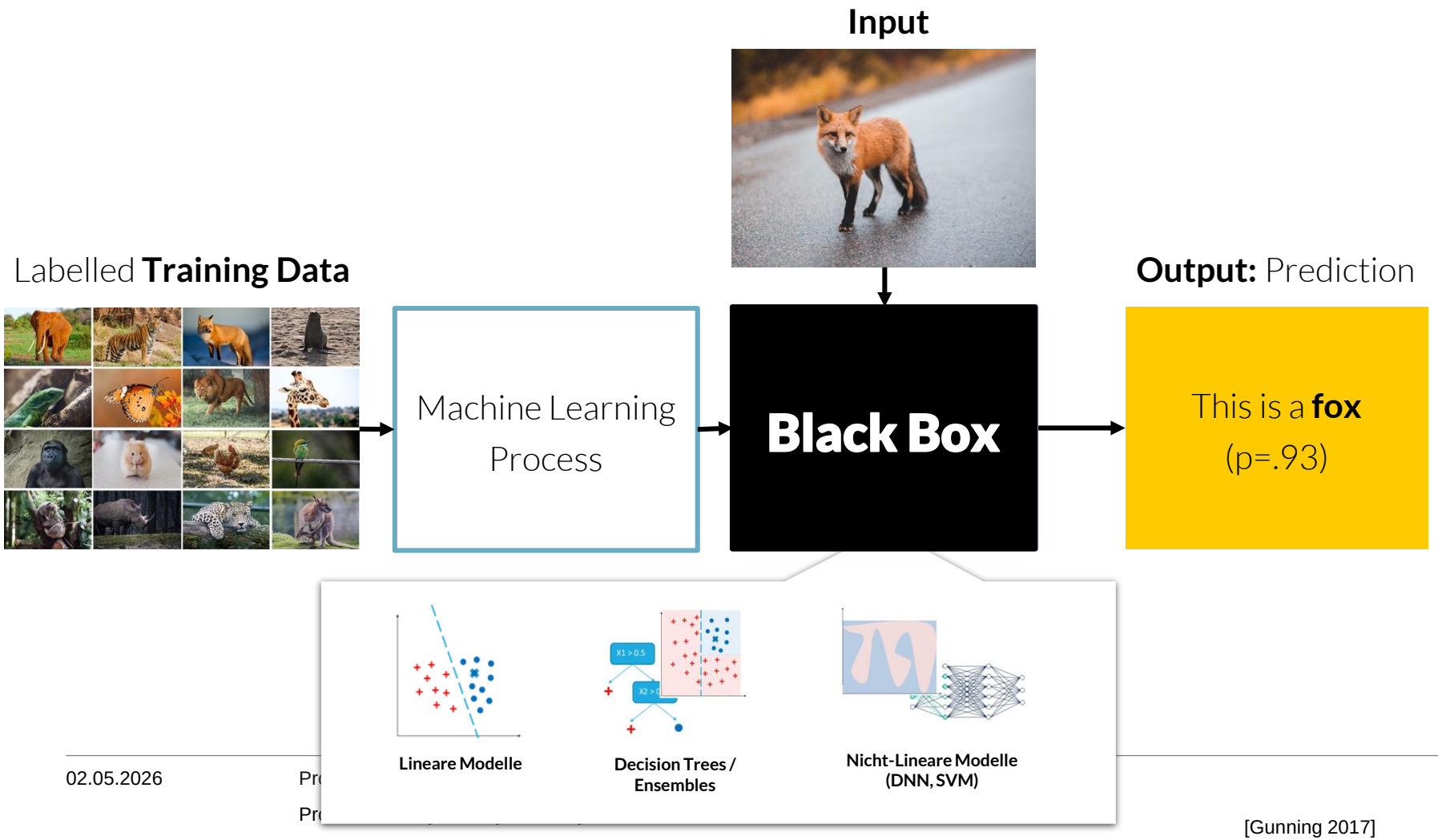
Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020



[Gunning 2017]

The Black Box Problem of Machine Learning

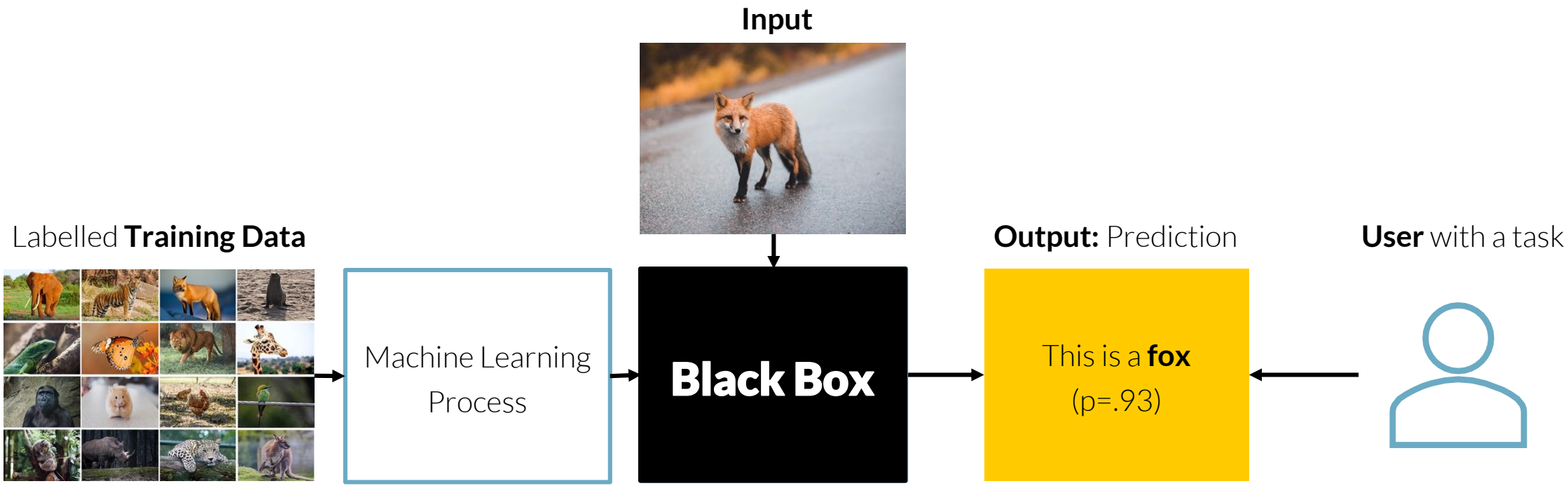
Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020



[Gunning 2017]

The Black Box Problem of Machine Learning

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

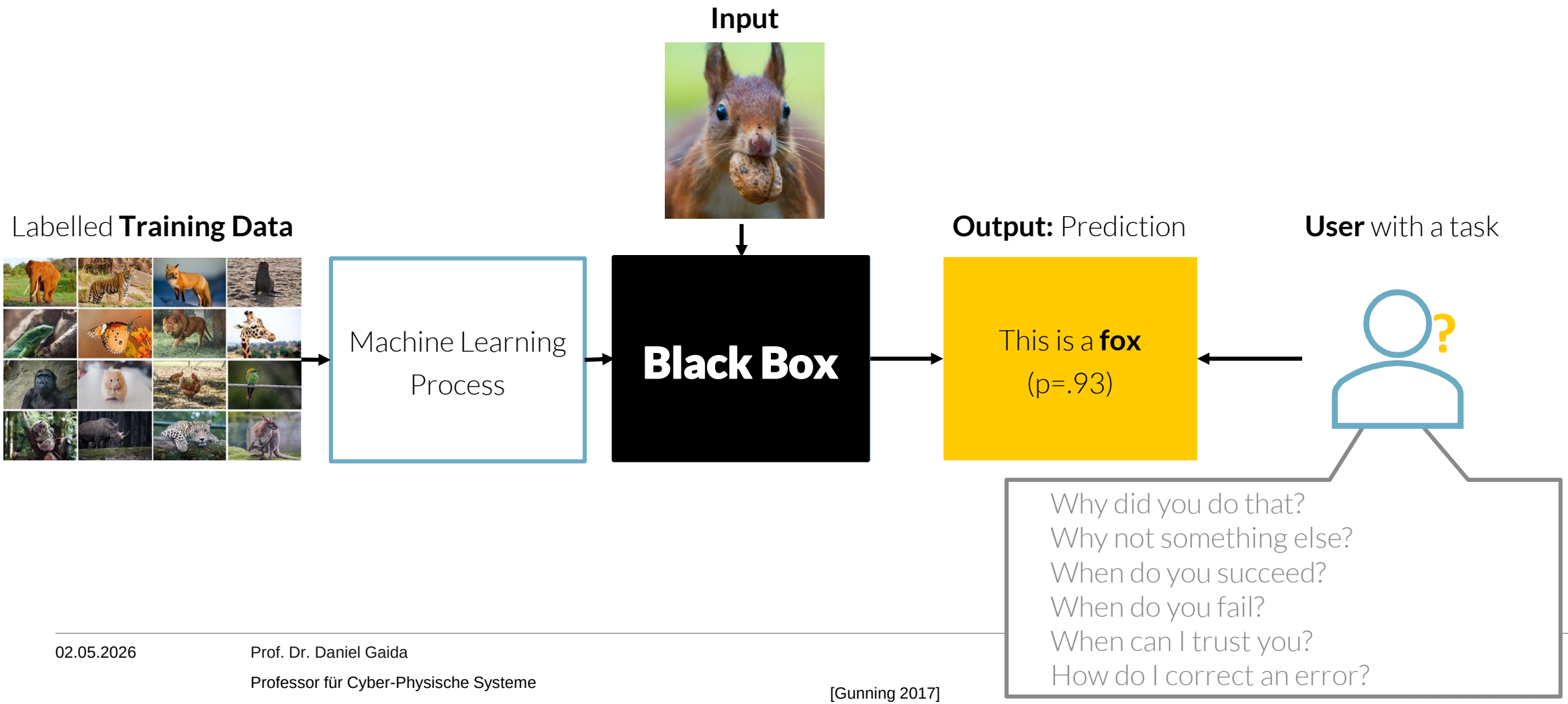


[Gunning 2017]

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

The Black Box Problem of Machine Learning

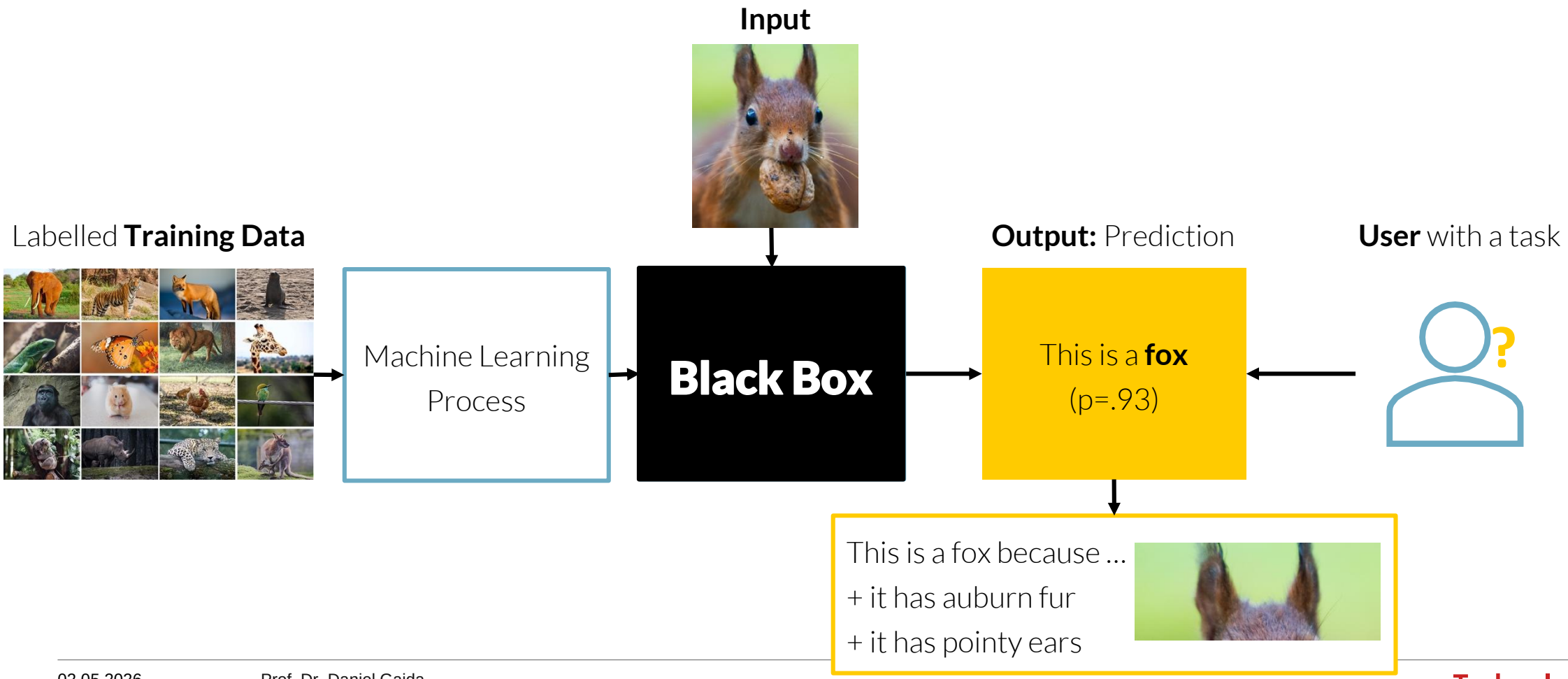
Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020



[Gunning 2017]

The Black Box Problem of Machine Learning

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020



[Gunning 2017]

Erklärungen für KI-Entwickler

Modell-Validierung

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Classified as Dog - correct

Classified as Wolf - correct

Erklärungen für KI-Entwickler

Modell-Validierung

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Source: Kateryna Babaieva | Pexels

Classified as Wolf – false (Husky)

[Ribeiro et al. 2016]

Erklärungen für KI-Entwickler

Modell-Validierung

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Source: Kateryna Babaieva | Pexels

Classified as Wolf – false (Husky)

Source : Kateryna Babaieva | Pexels, adapted after [Ribeiro et al. 2016]

LIME-Explanation (idealised)

[Ribeiro et al. 2016]

Warum möchten wir, dass uns eine KI ihre Entscheidung erklärt?

- Menschliche Neugierde und Lernen
  - Erklärungen stillen unser Bedürfnis nach Wissen und helfen uns, unerwartete Ergebnisse zu verstehen.
-
  -
  -
-
  -

Warum möchten wir, dass uns eine KI ihre Entscheidung erklärt?

- Menschliche Neugierde und Lernen
  - Erklärungen stillen unser Bedürfnis nach Wissen und helfen uns, unerwartete Ergebnisse zu verstehen.
- Sinn in der Welt finden
  - Erklärungen helfen uns, Widersprüche zwischen Erwartung und Realität aufzulösen.
  - Beispiel: Person wird ein Kreditantrag abgelehnt, obwohl sie eine gute Bonität hat. Erst durch eine Erklärung kann diese Diskrepanz verstanden werden.
-
  -

Warum möchten wir, dass uns eine KI ihre Entscheidung erklärt?

- Menschliche Neugierde und Lernen
  - Erklärungen stillen unser Bedürfnis nach Wissen und helfen uns, unerwartete Ergebnisse zu verstehen.
- Sinn in der Welt finden
  - Erklärungen helfen uns, Widersprüche zwischen Erwartung und Realität aufzulösen.
  - Beispiel: Person wird ein Kreditantrag abgelehnt, obwohl sie eine gute Bonität hat. Erst durch eine Erklärung kann diese Diskrepanz verstanden werden.
- Soziale Akzeptanz von KI erhöhen
  - Transparente Systeme schaffen Vertrauen und erleichtern die Integration von KI in unseren Alltag.

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Quelle: https://christophm.github.io/interpretable-ml-book/interpretability.html

Was ist Explainable AI?

Definition:

- Explainable AI (XAI) umfasst alle Ansätze, die darauf zielen, **menschlich verständliche Gründe** für die Entscheidungen eines KI-Systems bereitzustellen.

Was ist Explainable AI?

Definition:

- Explainable AI (XAI) umfasst alle Ansätze, die darauf zielen, **menschlich verständliche Gründe** für die Entscheidungen eines KI-Systems bereitzustellen.

Begriff	Bedeutung
Whitebox-Modell	Modell ist grundsätzlich durchschaubar (z. B. kleiner Entscheidungsbaum)
Blackbox-Modell	Struktur + Logik sind für Menschen nicht direkt verständlich (z. B. Deep Learning)
Erklärung	Zusatzinformationen zur Entscheidung – verbal, visuell oder symbolisch (z.B. Regeln)

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Transparenz vs Erklärbarkeit vs Interpretierbarkeit

Transparency
Provides documentation for the system's decisions and actions (details about model architecture, training data, optimizations, etc.) to ensures compliance and accountability

Examples: Model Cards, Datasheets for Datasets, Fairness Indicators, Algorithmic Impact Assessments

Interpretability
An interpretable model provides both visibility into its mechanisms and insight into how it arrives at its predictions.
Provides insights into what features are important, how they are related, or what rules/patterns are learned.

Examples: Inherently Interpretable Models - Decision Trees

Explainability
Aims to make any AI system, including opaque deep learning models, more explainable.
Involves developing techniques to explain the outputs/decisions of black-box AI models [usually] after they are trained.
Examples: Post-hoc Explanations - SHAP, Saliency Maps, Concept Activation Vectors

Quelle: Developing Explainable AI (XAI), Duke University, Coursera Kurs, https://www.coursera.org/learn/explainable-ai-xai/home/module/1

Transparenz

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Why you're seeing an ad

When you see an ad from Google's network, you can see more details:

- **Google services**, like Google Search, YouTube, or Gmail: Click **Info** ⓘ › **Why This Ad**.
- **Non-Google websites and apps** that partner with Google to show ads: Click **AdChoices**.
- **For some ads on Google's network**, you can click **Paid for by** to learn additional information about the advertiser.

Reasons you might see an ad

- **Your info:**
  - Info in your Google Account, like your age range and gender
  - Your general location
- **Your activity:**
  - Your current search query
  - Previous search activity
  - Your activity while you were signed in to Google
  - Your previous interactions with ads
  - Types of websites you visit

Source: https://adssettings.google.com

Right to Explanation in the GDPR

Datenschutz-Grundverordnung (DSGVO)

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Article 22

The data subject shall have the right **not to be subject to a decision based solely on automated processing**, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

[…]

…, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

Right to Explanation in the GDPR

Datenschutz-Grundverordnung (DSGVO)

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Article 22

The data subject shall have the right **not to be subject to a decision based solely on automated processing**, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

[…]

…, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

… trifft der für die Verarbeitung Verantwortliche geeignete Maßnahmen, um die Rechte und Freiheiten sowie die berechtigten Interessen der betroffenen Person zu wahren, zumindest das Recht, ein menschliches Eingreifen seitens des Verantwortlichen zu erwirken, ihren Standpunkt darzulegen und **die Entscheidung anzufechten**. [deepl]

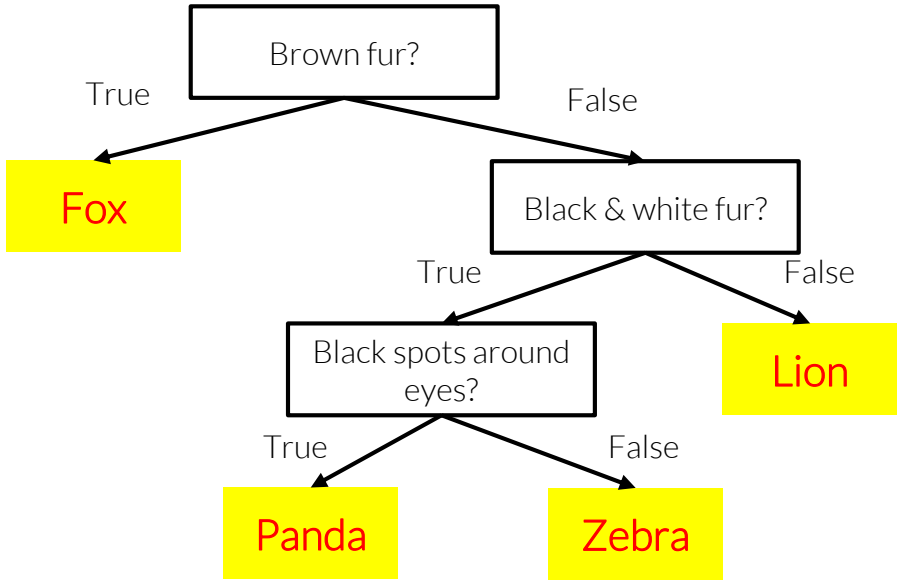
Quelle: https://gdpr-info.eu/art-22-gdpr/

Methoden der Erklärbarkeit

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Intrinsisch erklärbare Modelle (Whitebox)

Selbsterklärende Modelle, die die Interpretierbarkeit direkt in die Struktur integrieren



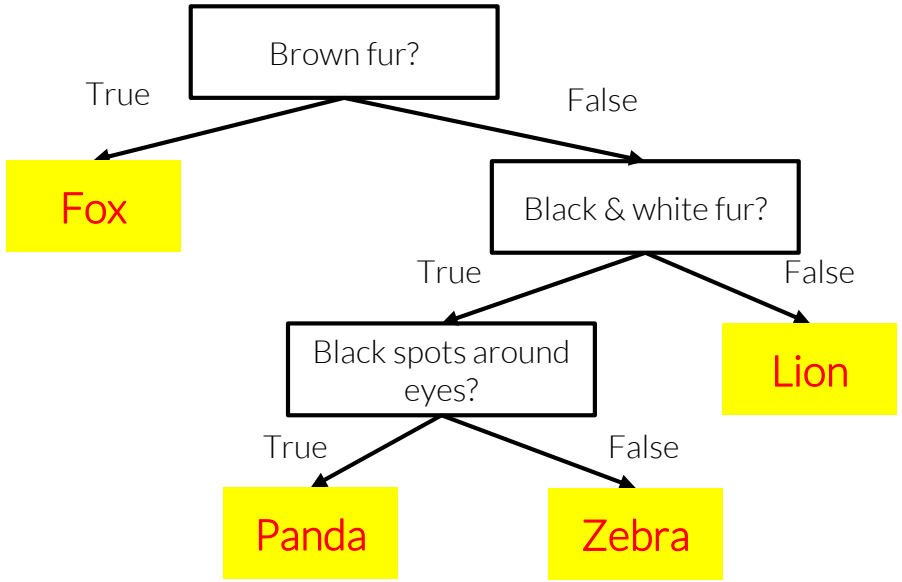
Post-hoc-Methoden (Blackbox-Erklärer)

Methoden der Erklärbarkeit

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

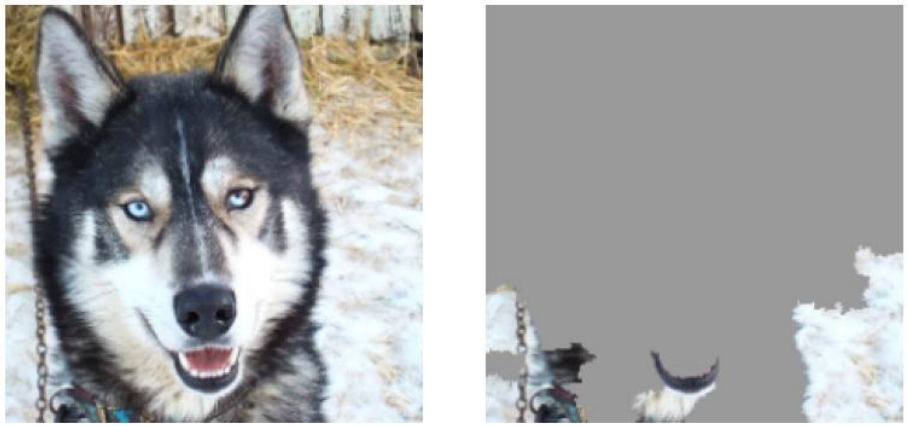
Intrinsisch erklärbare Modelle (Whitebox)

Selbsterklärende Modelle, die die Interpretierbarkeit direkt in die Struktur integrieren



Post-hoc-Methoden (Blackbox-Erklärer)

Ein zweites Modell ist erforderlich, das Erklärungen für das trainierte Blackbox Modell liefert



Source: [Ribeiro et al. 2016]

Methoden der Erklärbarkeit

Intrinsisch erklärbare Modelle (Whitebox)

- Lineare Regression
- Logistische Regression
- Entscheidungsbaum (klein)

Beispiele

Lineare Regression

- Bsp.: Vorhersage Mietpreis (MP) von Wohnung:
- MP(DZ, L) = -1.5*DZ - 0.5*L + 1050 €
- → Merkmal DZ (Distanz zum Zentrum/km) und L (Lärm/dB)

Methoden der Erklärbarkeit

Intrinsisch erklärbare Modelle (Whitebox)

- Lineare Regression
- Logistische Regression
- Entscheidungsbaum (klein)

Beispiele

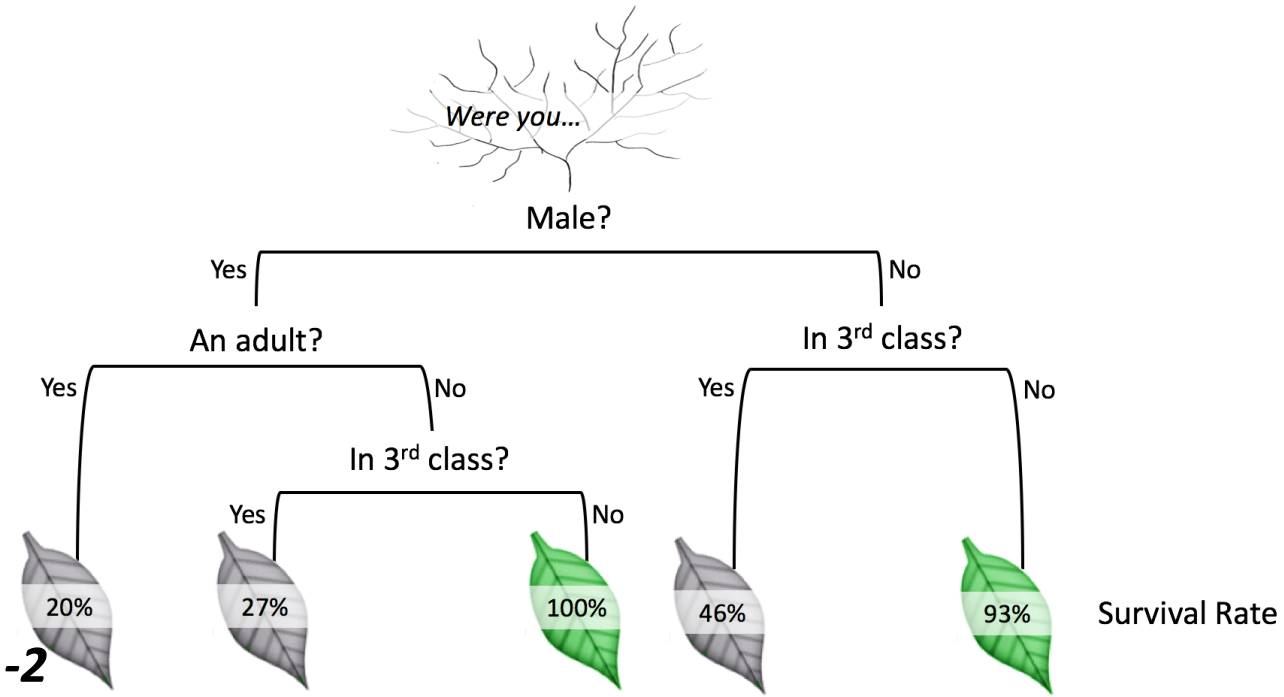
Lineare Regression

- Bsp.: Vorhersage Mietpreis (MP) von Wohnung:
- MP(DZ, L) = -1.5*DZ - 0.5*L + 1050 €
- → Merkmal DZ (Distanz zum Zentrum/km) und L (Lärm/dB) verringern Mietpreis.

Methoden der Erklärbarkeit

Intrinsisch erklärbare Modelle (Whitebox)

- Lineare Regression
- Logistische Regression
- Entscheidungsbaum (klein)



Beispiele

Lineare Regression

- Bsp.: Vorhersage Mietpreis (MP) von Wohnung:
- MP(DZ, L) = -1.5*DZ - 0.5*L + 1050 €
- → Merkmal DZ (Distanz zum Zentrum/km) und L (Lärm/dB) verringern Mietpreis.

Entscheidungsbaum:

- Bsp.: Titanic: „Wenn Sie weiblich waren und nicht 3. Klasse gereist sind, dann hätten Sie mit Wahrscheinlichkeit von 93 % überlebt.“

https://algobeans.com/2016/07/27/decision-trees-tutorial/

Methoden der Erklärbarkeit

Intrinsisch erklärbare Modelle (Whitebox)

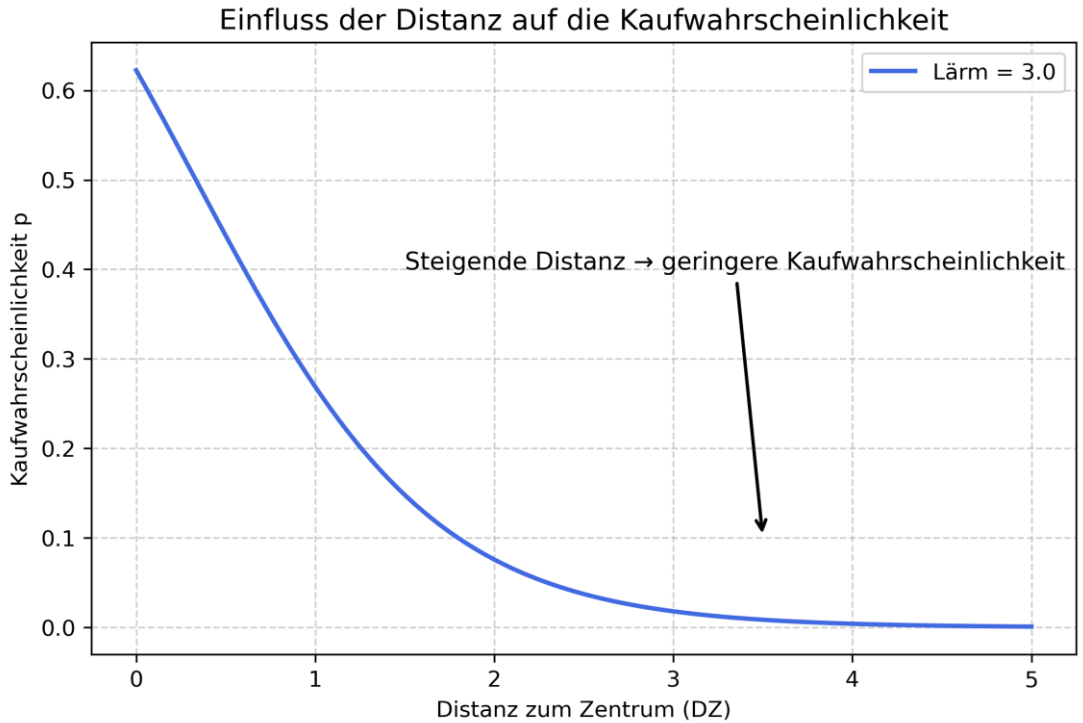
Beispiele

Logistische Regression

- Bsp.: Vorhersage Kaufentscheidung von Wohnung (Ja/Nein):
- $Wkt = \frac{1}{1+\exp(-(-1.5DZ-0.5L+2.0))}$
- DZ (Distanz zum Zentrum/km)
- L (Lärm/dB)
-

Methoden der Erklärbarkeit

Intrinsisch erklärbare Modelle (Whitebox)



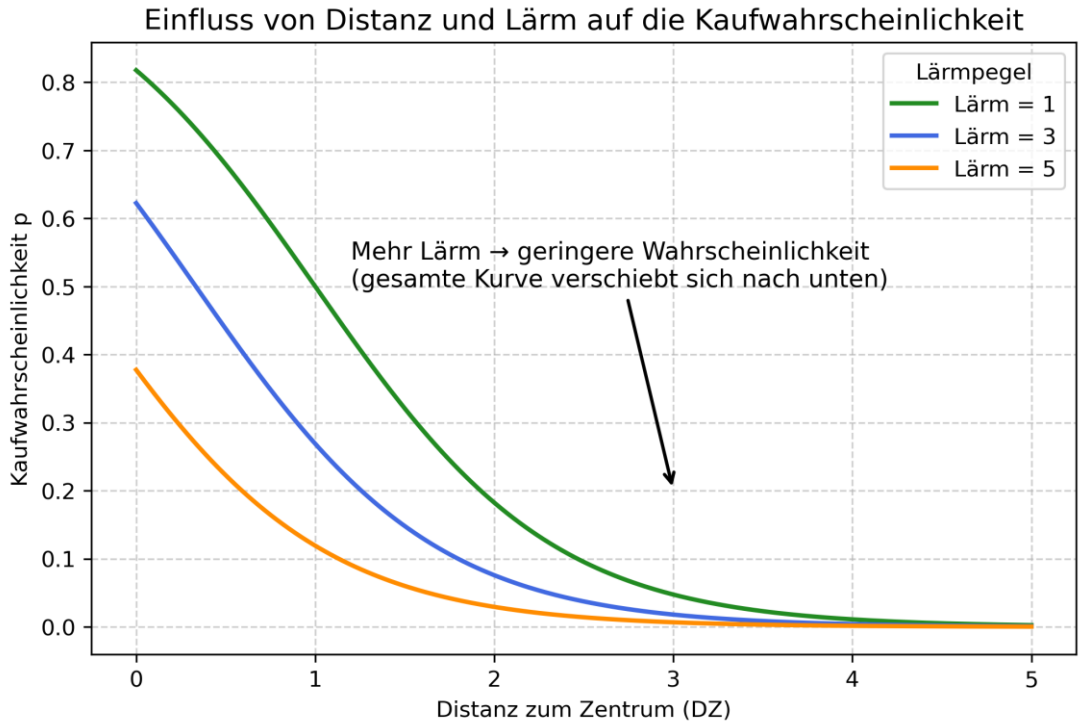
Beispiele

Logistische Regression

- Bsp.: Vorhersage Kaufentscheidung von Wohnung (Ja/Nein):
- $Wkt = \frac{1}{1+\exp(-(-1.5DZ-0.5L+2.0))}$
- DZ (Distanz zum Zentrum/km)
- L (Lärm/dB)
- Sigmoid-Funktion: $\frac{1}{1+e^{-(\omega_0+\omega_1 \cdot x)}}$

Methoden der Erklärbarkeit

Intrinsisch erklärbare Modelle (Whitebox)



Beispiele

Logistische Regression

- Bsp.: Vorhersage Kaufentscheidung von Wohnung (Ja/Nein):
- $Wkt = \frac{1}{1+\exp(-(-1.5DZ-0.5L+2.0))}$
- DZ (Distanz zum Zentrum/km)
- L (Lärm/dB)
- Sigmoid-Funktion: $\frac{1}{1+e^{-(\omega_0+\omega_1 \cdot x)}}$

Methoden der Erklärbarkeit

Intrinsisch erklärbare Modelle (Whitebox)

- Lineare Regression
- Logistische Regression
- Entscheidungsbaum (klein)

- Vorteil:
  - Einfach verständlich
- Nachteil:
  - Eingeschränkt in Komplexität / Performance

Beispiele

Lineare Regression

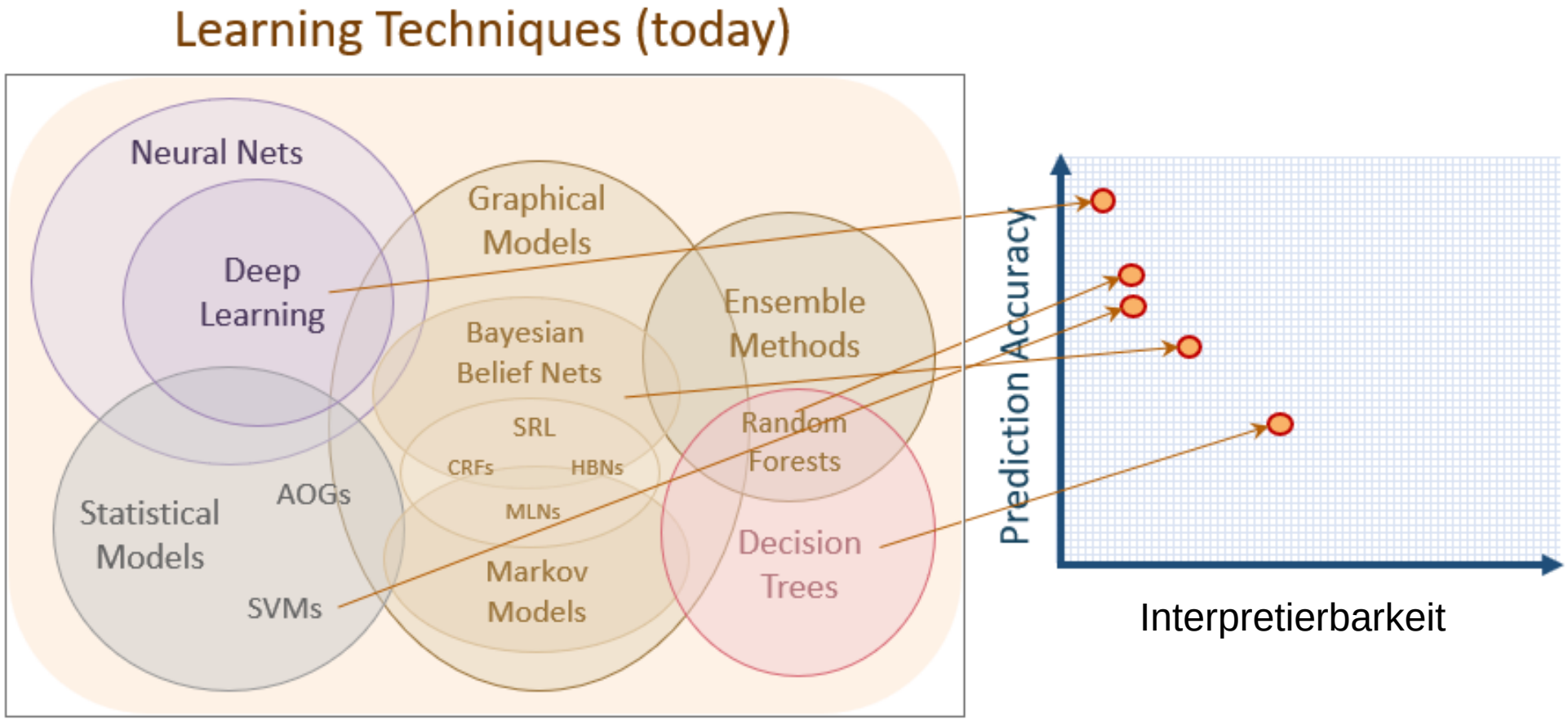
- Bsp.: Vorhersage Mietpreis (MP) von Wohnung:
- MP(DZ, L) = -1.5*DZ - 0.5*L + 1050 €
- → Merkmal DZ (Distanz zum Zentrum/km) und L (Lärm/dB) verringern Mietpreis.

Entscheidungsbaum:

- Bsp.: Titanic: „Wenn Sie weiblich waren und nicht 3. Klasse gereist sind, dann hätten Sie mit Wahrscheinlichkeit von 93 % überlebt.“

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Kompromiss zwischen Interpretierbarkeit und Genauigkeit



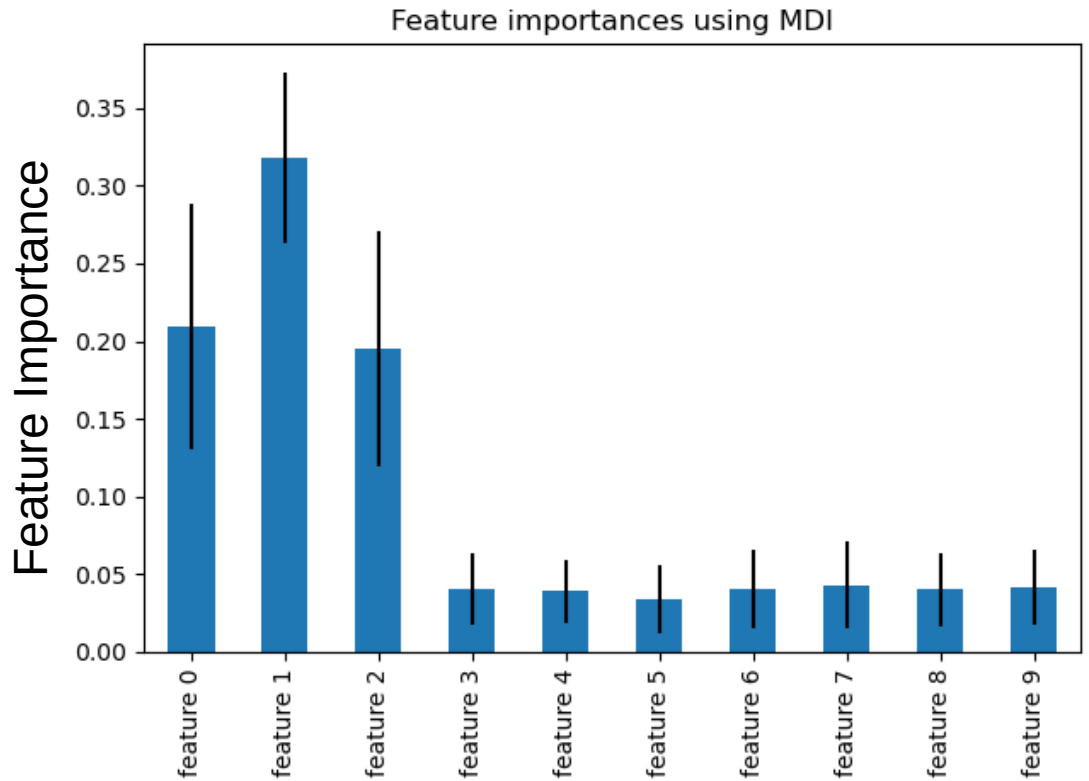
Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Image & Content: [Gunning 2017]

Entscheidungsbäume und Feature Importance

- Trainierter Entscheidungsbaum und Random Forest liefern Feature Importance zurück
- Desto größer der Wert der Importance eines Features, desto wichtiger ist es



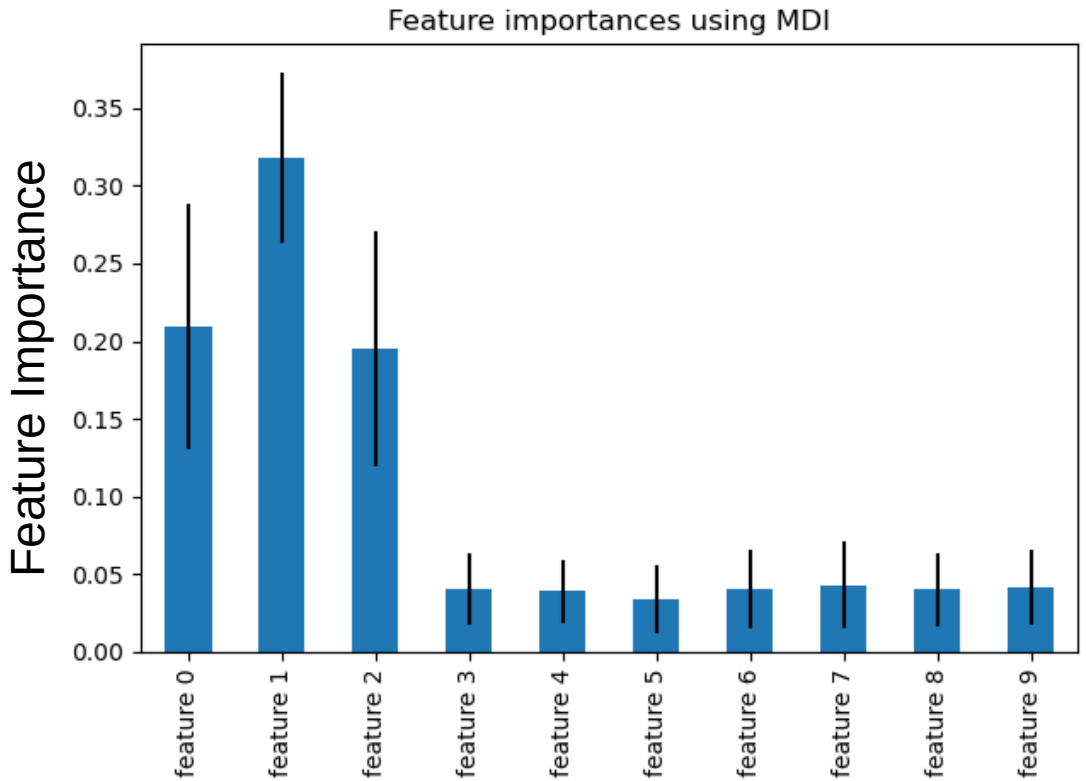
Quelle: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

Entscheidungsbäume und Feature Importance

- Trainierter Entscheidungsbaum und Random Forest liefern Feature Importance zurück
- Desto größer der Wert der Importance eines Features, desto wichtiger ist es

Code:

- importances = forest.feature_importances_



Quelle: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik

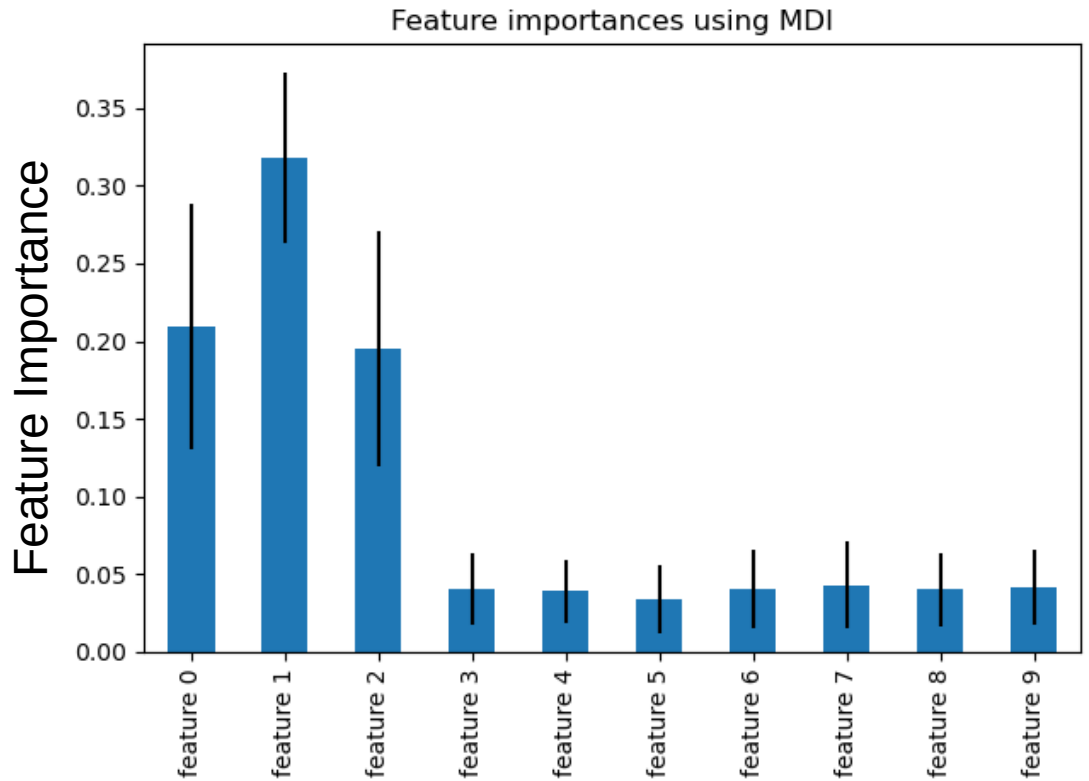
Entscheidungsbäume und Feature Importance

- Trainierter Entscheidungsbaum und Random Forest liefern Feature Importance zurück
- Desto größer der Wert der Importance eines Features, desto wichtiger ist es

Code:

- importances = forest.feature_importances_

s. Übung



Quelle: https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Fragen?

- Warum möchten wir, dass uns eine KI ihre Entscheidung erklärt?
  - Für wen ist die Erklärung?
- Was ist Erklärbare KI?
  - Erklärbarkeit
  - Interpretierbarkeit

- Intrinsisch erklärbare Machine Learning Modelle

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Lernraum III

Explainable AI (XAI) – Erklärbare KI

- Warum möchten wir, dass uns eine KI ihre Entscheidung erklärt?
  - Für wen ist die Erklärung?
- Was ist Erklärbare KI?

- Intrinsisch erklärbare Modelle
- **Post-hoc-Methoden**
  - **LIME**
  - **SHAP**

-

Lernraum III

Explainable AI (XAI) – Erklärbare KI

- Warum möchten wir, dass uns eine KI ihre Entscheidung erklärt?
  - Für wen ist die Erklärung?
- Was ist Erklärbare KI?

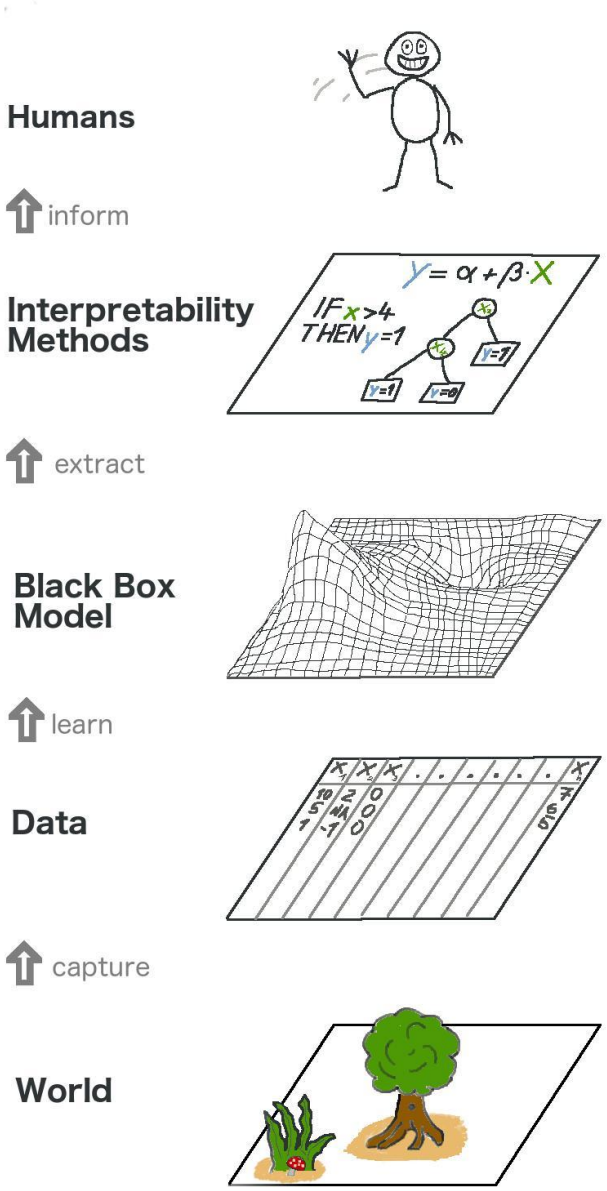
- Intrinsisch erklärbare Modelle
- **Post-hoc-Methoden**
  - **LIME**
  - **SHAP**

- Erklärungen heutiger Systeme

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Methoden der Erklärbarkeit

Bildquelle: https://christophm.github.io/interpretable-ml-book/overview.html

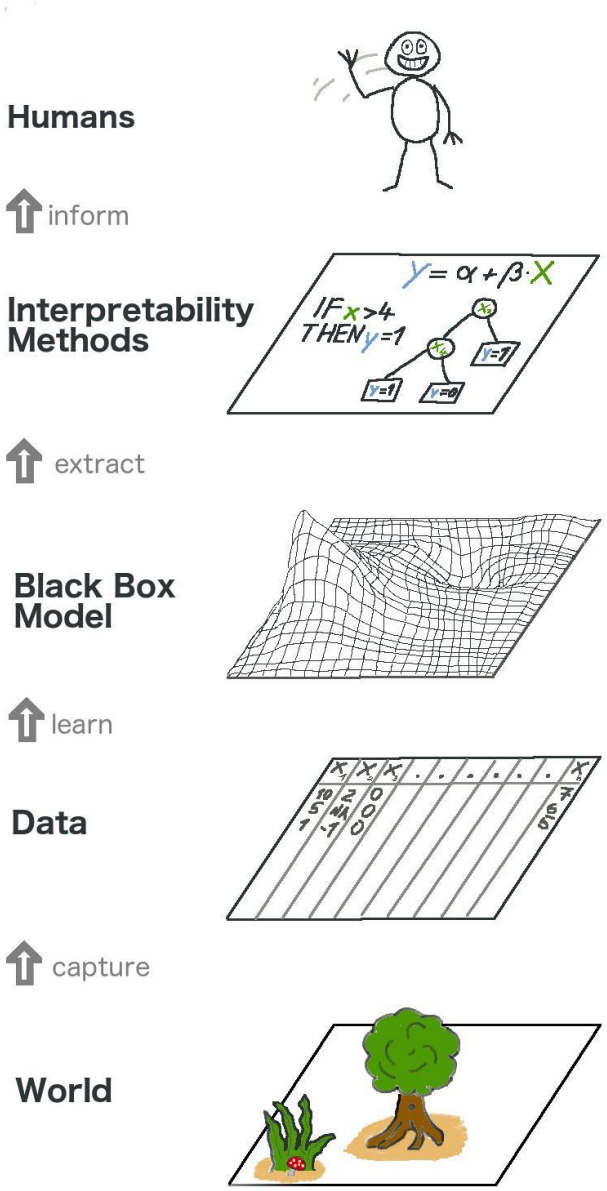


Post-hoc-Methoden (Blackbox-Erklärer)

Ein zweites Modell ist erforderlich, das Erklärungen für das trainierte Blackbox Modell liefert

Methoden der Erklärbarkeit

Bildquelle: https://christophm.github.io/interpretable-ml-book/overview.html



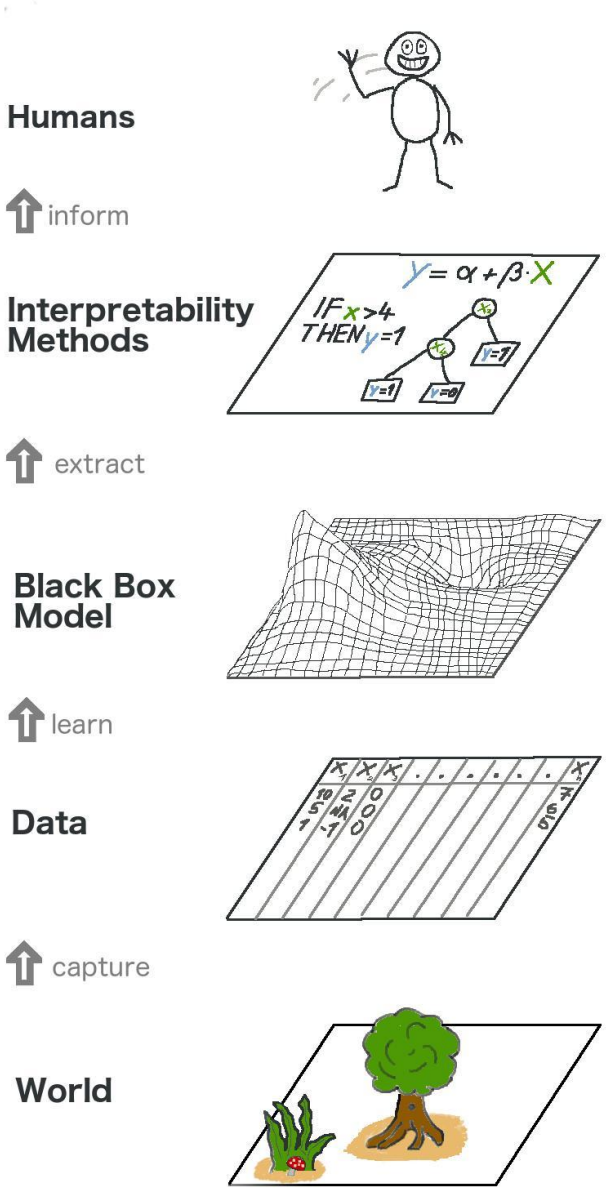
Post-hoc-Methoden (Blackbox-Erklärer)

Ein zweites Modell ist erforderlich, das Erklärungen für das trainierte Blackbox Modell liefert

- Ziel: Erklärung ohne Zugriff auf die Struktur des ersten Modells
-

Methoden der Erklärbarkeit

Bildquelle: https://christophm.github.io/interpretable-ml-book/overview.html



Post-hoc-Methoden (Blackbox-Erklärer)

Ein zweites Modell ist erforderlich, das Erklärungen für das trainierte Blackbox Modell liefert

- Ziel: Erklärung ohne Zugriff auf die Struktur des ersten Modells
- Methoden:
  - LIME
  - SHAP
  - Saliency Maps

Lokale vs. globale Erklärbarkeit

Lokale Erklärbarkeit

- Warum hat das Modell diese eine Entscheidung getroffen?
- Fokus: einzelne Vorhersage (z. B. Kredit für Person X).
- Hilft Einzelpersonen

Globale Erklärbarkeit

Prof. Dr. Daniel Gaida

Professor für Cyber-Physische Systeme

Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik

Technology Arts Sciences TH Köln

Lokale vs. globale Erklärbarkeit

Lokale Erklärbarkeit

- Warum hat das Modell diese eine Entscheidung getroffen?
- Fokus: einzelne Vorhersage (z. B. Kredit für Person X).
- Hilft Einzelpersonen

- Beispiel:
  - Kredit nicht erhalten, da Einkommen zu klein und letzter Kredit noch nicht zurückgezahlt

Globale Erklärbarkeit

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Lokale vs. globale Erklärbarkeit

Lokale Erklärbarkeit

- Warum hat das Modell diese eine Entscheidung getroffen?
- Fokus: einzelne Vorhersage (z. B. Kredit für Person X).
- Hilft Einzelpersonen

- Beispiel:
  - Kredit nicht erhalten, da Einkommen zu klein und letzter Kredit noch nicht zurückgezahlt

Globale Erklärbarkeit

- Wie funktioniert das Modell insgesamt?
- Fokus: generelle Regeln, Muster, Einflussgrößen.
- Hilft Entwickler*innen, Regulierungsbehörden oder Manager*innen, das Modell zu verstehen und zu verbessern.

-
  -

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Lokale vs. globale Erklärbarkeit

Lokale Erklärbarkeit

- Warum hat das Modell diese eine Entscheidung getroffen?
- Fokus: einzelne Vorhersage (z. B. Kredit für Person X).
- Hilft Einzelpersonen

- Beispiel:
  - Kredit nicht erhalten, da Einkommen zu klein und letzter Kredit noch nicht zurückgezahlt

Globale Erklärbarkeit

- Wie funktioniert das Modell insgesamt?
- Fokus: generelle Regeln, Muster, Einflussgrößen.
- Hilft Entwickler*innen, Regulierungsbehörden oder Manager*innen, das Modell zu verstehen und zu verbessern.

- Beispiel:
  - Das Einkommen ist das wichtigste Merkmal, gefolgt von Merkmal …

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Lokale vs. globale Erklärbarkeit

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Lokale Erklärbarkeit – eher schlechtes Beispiel, da es nichts erklärt

Source: mail.google.com

Post-hoc-Methoden (Blackbox-Erklärer)

Lokale modellagnostische Post-hoc-Methoden

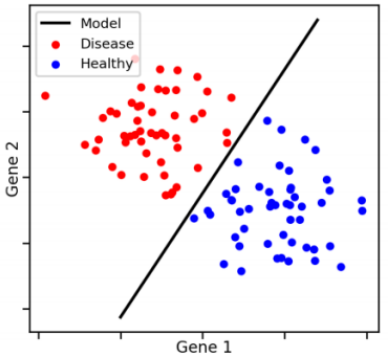
- LIME

Lokale und globale modellagnostische Post-hoc-Methoden

- SHAP

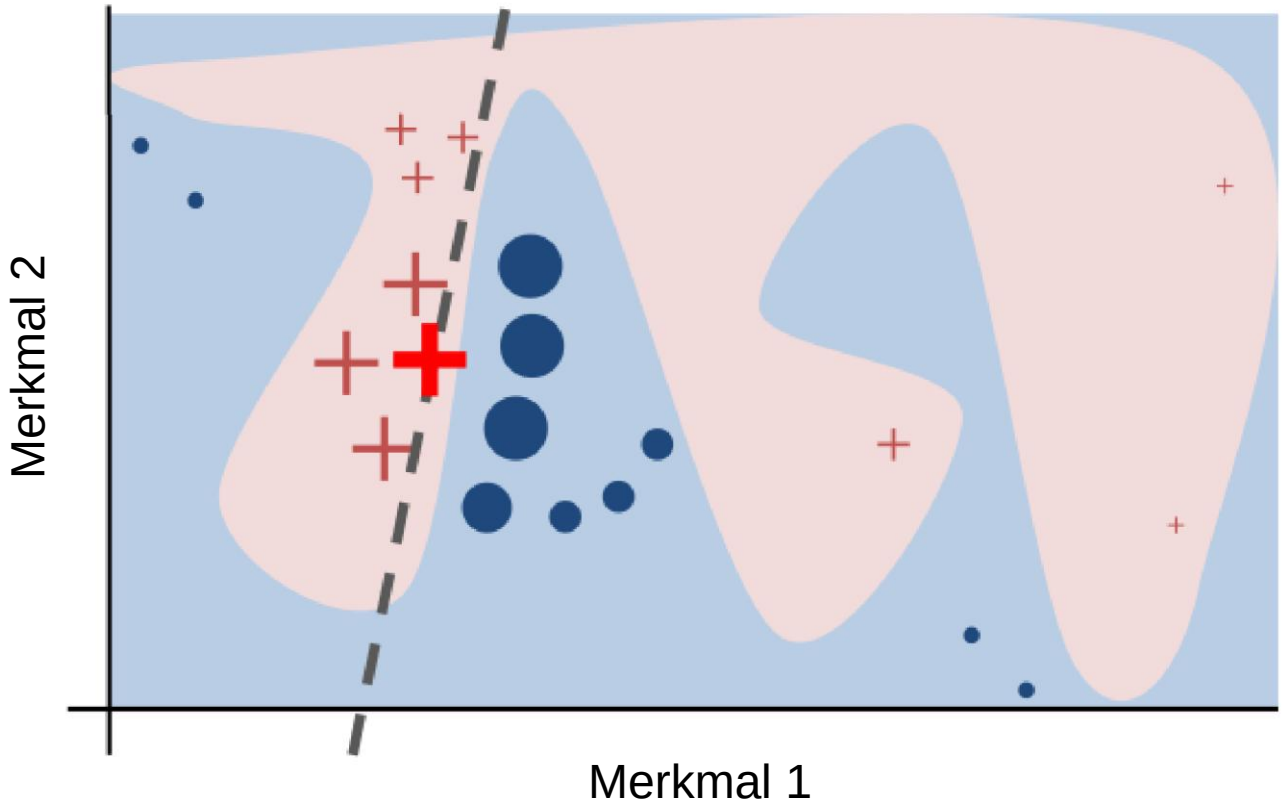
Post-hoc-Methoden (Blackbox-Erklärer)

LIME (Local Interpretable Model-agnostic Explanations)



LIME: Lokale Erklärbarkeit für Blackbox-Modelle

- LIME erzeugt eine lineare Approximation rund um eine konkrete Vorhersage
-
-
-

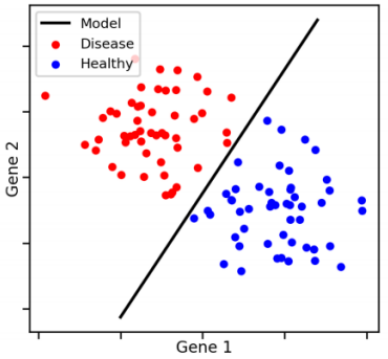


Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik

Bildquelle: Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "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).

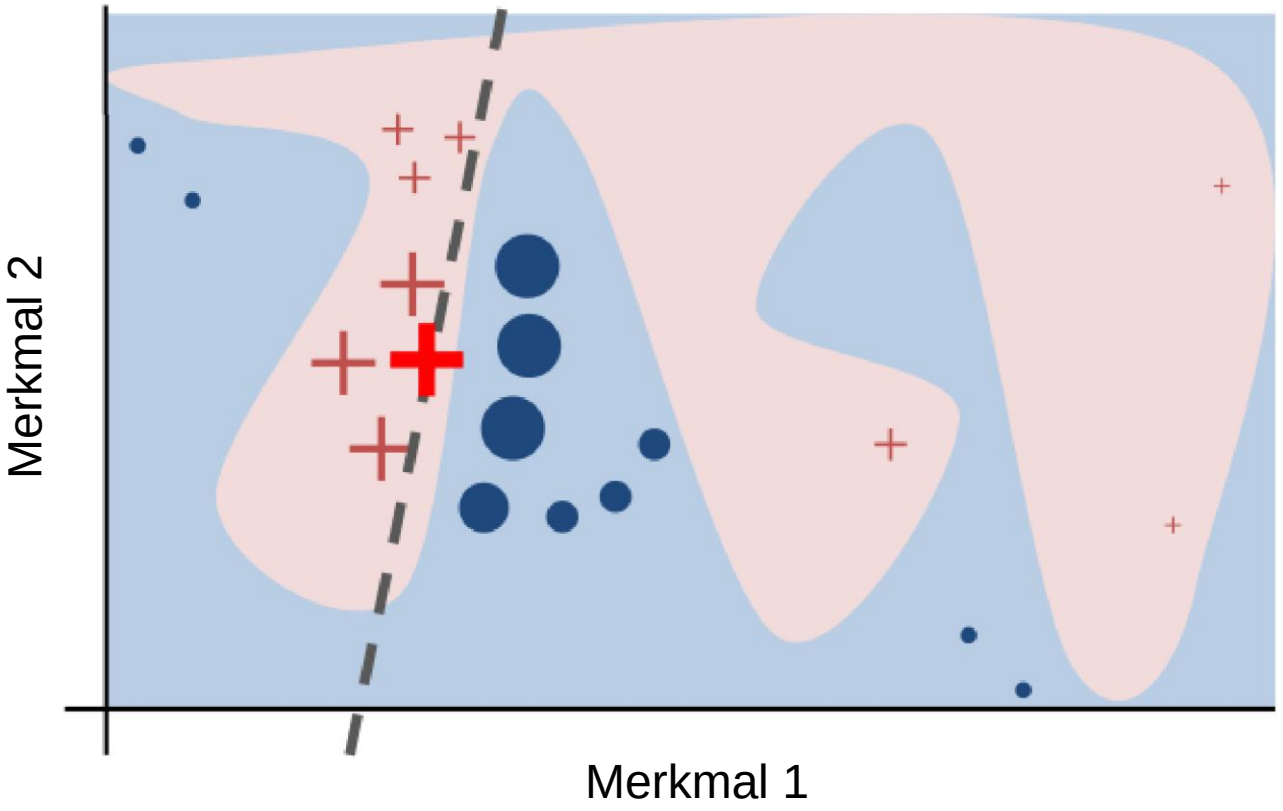
Post-hoc-Methoden (Blackbox-Erklärer)

LIME (Local Interpretable Model-agnostic Explanations)



LIME: Lokale Erklärbarkeit für Blackbox-Modelle

- LIME erzeugt eine lineare Approximation rund um eine konkrete Vorhersage
- Datenpunkte sind Variationen des Datenpunktes für den wir eine Erklärung haben möchten
-
-

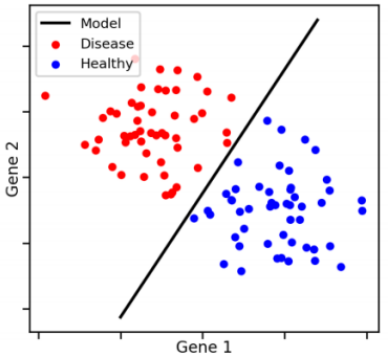


Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik

Bildquelle: Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "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).

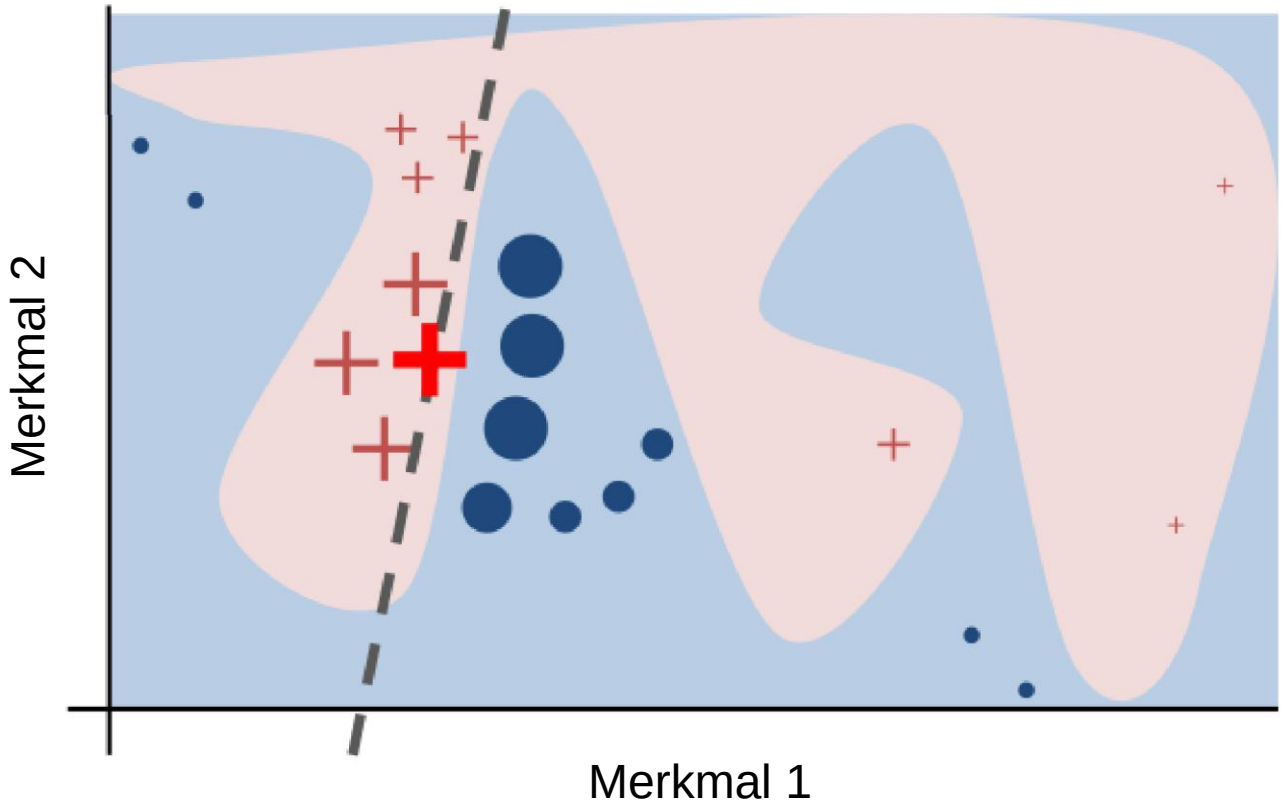
Post-hoc-Methoden (Blackbox-Erklärer)

LIME (Local Interpretable Model-agnostic Explanations)



LIME: Lokale Erklärbarkeit für Blackbox-Modelle

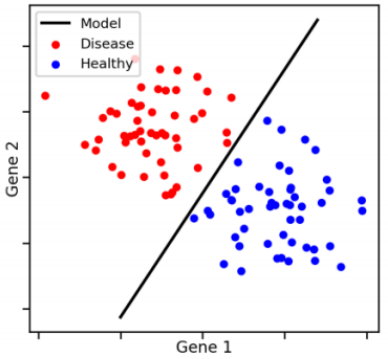
- LIME erzeugt eine lineare Approximation rund um eine konkrete Vorhersage
- Datenpunkte sind Variationen des Datenpunktes für den wir eine Erklärung haben möchten
- Variationen werden durch Blackbox-Modell klassifiziert
-



Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik

Bildquelle: Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "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).

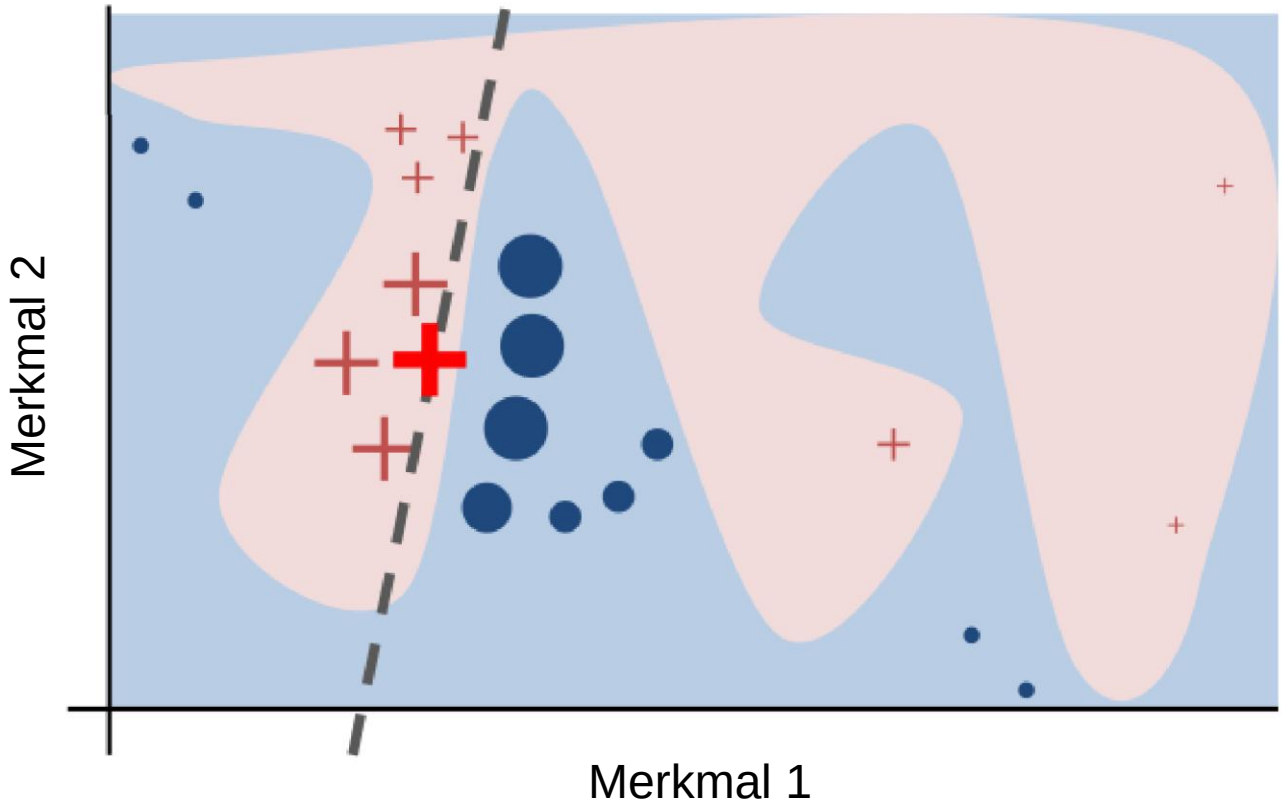
Post-hoc-Methoden (Blackbox-Erklärer)
LIME (Local Interpretable Model-agnostic Explanations)



LIME: Lokale Erklärbarkeit für Blackbox-Modelle

- LIME erzeugt eine lineare Approximation rund um eine konkrete Vorhersage
- Datenpunkte sind Variationen des Datenpunktes für den wir eine Erklärung haben möchten
- Variationen werden durch Blackbox-Modell klassifiziert

- Idee: Verhalten des Blackbox-Modells lokal erklären, nicht global verstehen



Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik

Bildquelle: Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "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).

Post-hoc-Methoden (Blackbox-Erklärer)

LIME (Local Interpretable Model-agnostic Explanations)

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Intuition

1) Divide input into **interpretable components** that "make sense" to humans (e.g. words or parts of image)

Original Image

Interpretable Components

Superpixel

Image & Content: [Ribeiro et al. 2016]

Post-hoc-Methoden (Blackbox-Erklärer)

LIME (Local Interpretable Model-agnostic Explanations)

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Intuition

1) Divide input into **interpretable components** that "make sense" to humans (e.g. words or parts of image)

2) **Generate random perturbations** of data set

Image & Content: [Ribeiro et al. 2016]

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Post-hoc-Methoden (Blackbox-Erklärer)

LIME (Local Interpretable Model-agnostic Explanations)

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Intuition

1) Divide input into **interpretable components** that "make sense" to humans (e.g. words or parts of image)
2) **Generate random perturbations** of data set
3) **Predict classes for** these **perturbations** using your black box model

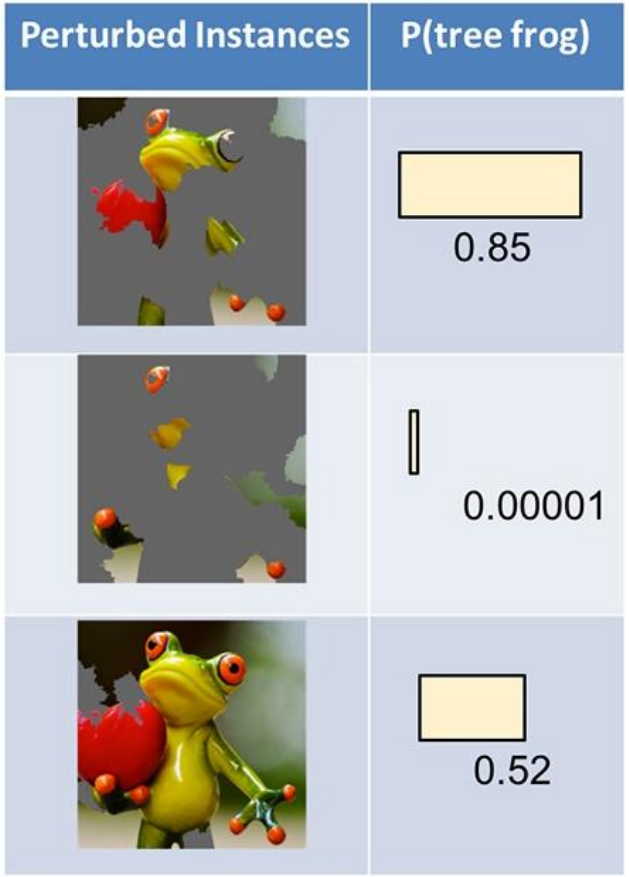


Image & Content: [Ribeiro et al. 2016]

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Post-hoc-Methoden (Blackbox-Erklärer)

LIME (Local Interpretable Model-agnostic Explanations)

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Intuition

1) Divide input into **interpretable components** that "make sense" to humans (e.g. words or parts of image)
2) **Generate random perturbations** of data set
3) **Predict classes for** these **perturbations** using your black box model
4) **Weight** the perturbations (importance) according to their proximity to the original input.

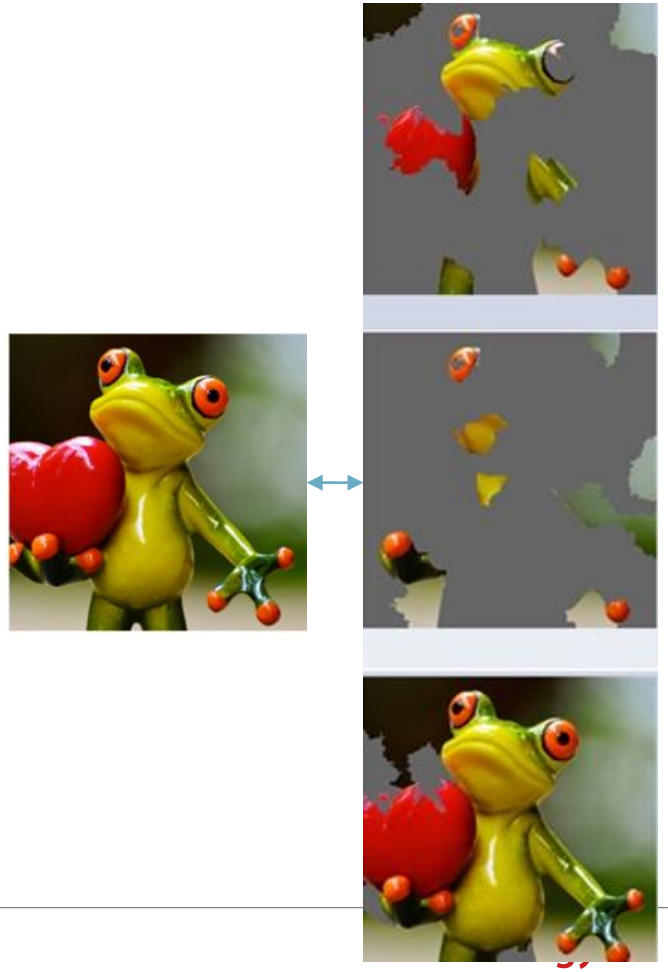


Image & Content: [Ribeiro et al. 2016]

Post-hoc-Methoden (Blackbox-Erklärer)

LIME (Local Interpretable Model-agnostic Explanations)

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Intuition

1) Divide input into **interpretable components** that "make sense" to humans (e.g. words or parts of image)
2) **Generate random perturbations** of data set
3) **Predict classes for** these **perturbations** using your black box model
4) **Weight** the perturbations (importance) according to their proximity to the original input.
5) **Train a weighted, interpretable model** on the dataset with the variations.

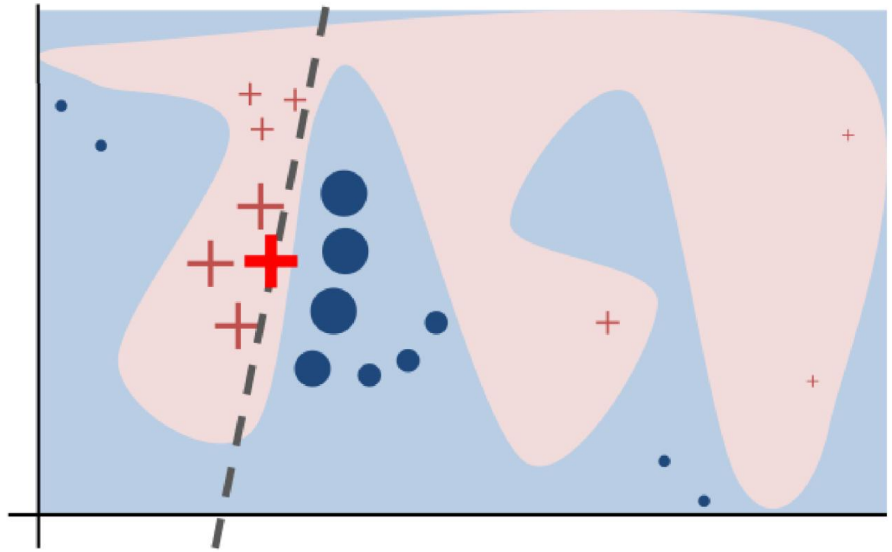


Image & Content: [Ribeiro et al. 2016]

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Post-hoc-Methoden (Blackbox-Erklärer)

LIME (Local Interpretable Model-agnostic Explanations)

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Intuition

1) Divide input into **interpretable components** that "make sense" to humans (e.g. words or parts of image)
2) **Generate random perturbations** of data set
3) **Predict classes for** these **perturbations** using your black box model
4) **Weight** the perturbations (importance) according to their proximity to the original input.
5) **Train a weighted, interpretable model** on the dataset with the variations.
6) Explain the prediction by **interpreting the local model**.
   → Wir schauen uns die Superpixels an, die zu den größten Gewichte gehören

Image & Content: [Ribeiro et al. 2016]

Post-hoc-Methoden (Blackbox-Erklärer)

LIME (Local Interpretable Model-agnostic Explanations)

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Practical Example:

https://colab.research.google.com/github/arteagac/arteagac.github.io/blob/master/blog/lime_image.ipynb

Post-hoc-Methoden (Blackbox-Erklärer)

LIME (Local Interpretable Model-agnostic Explanations)

Wie funktioniert LIME?

- Auswahl eines konkreten Datenpunkts („Warum diese Entscheidung?“)
- Erzeugung von leichten Variationen dieses Datenpunkts („Perturbationen“)
- Vorhersage mit dem Blackbox-Modell für jede Variation
- Training eines einfachen Modells (z. B. lineare Regression) mit diesen lokalen Variationen
- Visualisierung der Feature-Gewichte für die Erklärung

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Post-hoc-Methoden (Blackbox-Erklärer)

LIME (Local Interpretable Model-agnostic Explanations)

Wie funktioniert LIME?

- Auswahl eines konkreten Datenpunkts („Warum diese Entscheidung?“)
- Erzeugung von leichten Variationen dieses Datenpunkts („Perturbationen“)
- Vorhersage mit dem Blackbox-Modell für jede Variation
- Training eines einfachen Modells (z. B. lineare Regression) mit diesen lokalen Variationen
- Visualisierung der Feature-Gewichte für die Erklärung

- Beispiel s. Übung.

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Post-hoc-Methoden (Blackbox-Erklärer)
LIME (Local Interpretable Model-agnostic Explanations)

Vorteile

- Modellunabhängig
  - funktioniert mit beliebigen Blackbox-Modelle (z. B. Random Forests, Deep Learning)
- Erzeugt verständliche Erklärungen
- Flexibel einsetzbar
  - Textklassifikation, Bildklassifikation, tabellarische Daten

Grenzen

-
-
-

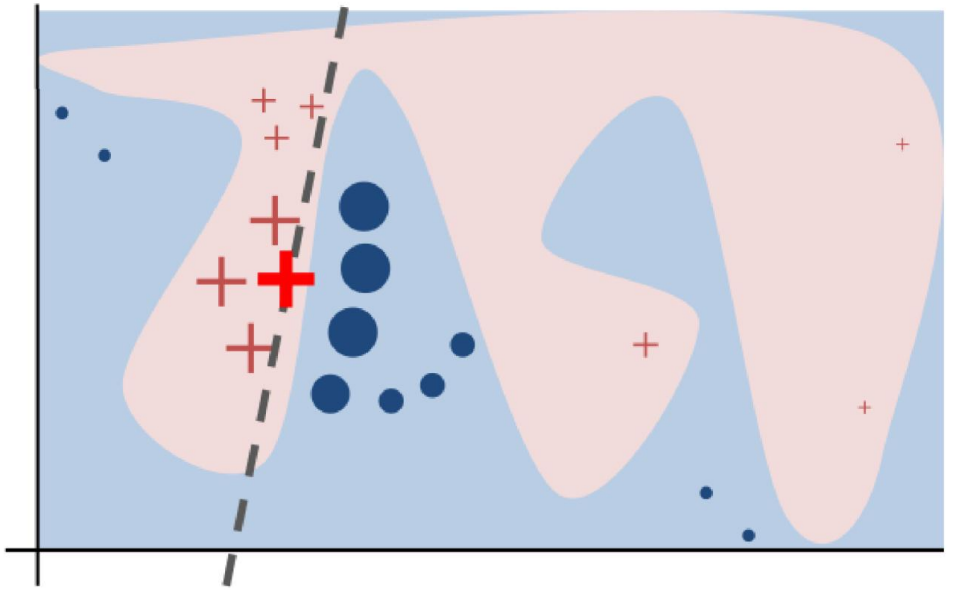
Post-hoc-Methoden (Blackbox-Erklärer)
LIME (Local Interpretable Model-agnostic Explanations)

Vorteile

- Modellunabhängig
  - funktioniert mit beliebigen Blackbox-Modelle (z. B. Random Forests, Deep Learning)
- Erzeugt verständliche Erklärungen
- Flexibel einsetzbar
  - Textklassifikation, Bildklassifikation, tabellarische Daten

Grenzen

- Ergebnisse können instabil sein (bei ähnlichem Input → andere Erklärung)
- Nur lokal interpretierbar (Lokale Erklärbarkeit)
- Erklärungen nicht immer eindeutig verständlich für Laien



Post-hoc-Methoden (Blackbox-Erklärer)

SHAP (SHapley Additive exPlanations)

SHAP: Fairness und Relevanz durch Spieltheorie

- SHAP basiert auf Shapley-Werten aus der Spieltheorie
  - Der Shapley-Wert ist eine Methode zur Zuweisung von Auszahlungen an Spieler in Abhängigkeit von ihrem Beitrag zur Gesamtauszahlung. Die Spieler kooperieren in einer Koalition und erhalten aus dieser Kooperation einen bestimmten Gewinn.

Quelle: https://christophm.github.io/interpretable-ml-book/overview.html

Post-hoc-Methoden (Blackbox-Erklärer)

SHAP (SHapley Additive exPlanations)

SHAP: Fairness und Relevanz durch Spieltheorie

- SHAP basiert auf Shapley-Werten aus der Spieltheorie
  - Der Shapley-Wert ist eine Methode zur Zuweisung von Auszahlungen an Spieler in Abhängigkeit von ihrem Beitrag zur Gesamtauszahlung. Die Spieler kooperieren in einer Koalition und erhalten aus dieser Kooperation einen bestimmten Gewinn.
  - Analogie:
    - Merkmal = Spieler; Gewinn = korrekte Vorhersage
- Jedes Merkmal liefert einen Beitrag zur Vorhersage – wie Spieler zum Teamerfolg
-
  -
  -

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik

Quelle: https://christophm.github.io/interpretable-ml-book/overview.html

Post-hoc-Methoden (Blackbox-Erklärer)

SHAP (SHapley Additive exPlanations)

SHAP: Fairness und Relevanz durch Spieltheorie

- SHAP basiert auf Shapley-Werten aus der Spieltheorie
  - Der Shapley-Wert ist eine Methode zur Zuweisung von Auszahlungen an Spieler in Abhängigkeit von ihrem Beitrag zur Gesamtauszahlung. Die Spieler kooperieren in einer Koalition und erhalten aus dieser Kooperation einen bestimmten Gewinn.
  - Analogie:
    - Merkmal = Spieler; Gewinn = korrekte Vorhersage
- Jedes Merkmal liefert einen Beitrag zur Vorhersage – wie Spieler zum Teamerfolg

- Vorgehensweise:
  - Es wird je ein Merkmal weggelassen und mit den anderen Merkmalen je ein Modell trainiert
  - Die Vorhersagen der verschiedenen Modelle werden linear kombiniert (ähnlich zu LIME – dort hatten wir viele Vorhersagen mit dem gleichen Modell, hier haben wir je eine Vorhersage mit vielen Modellen)

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik

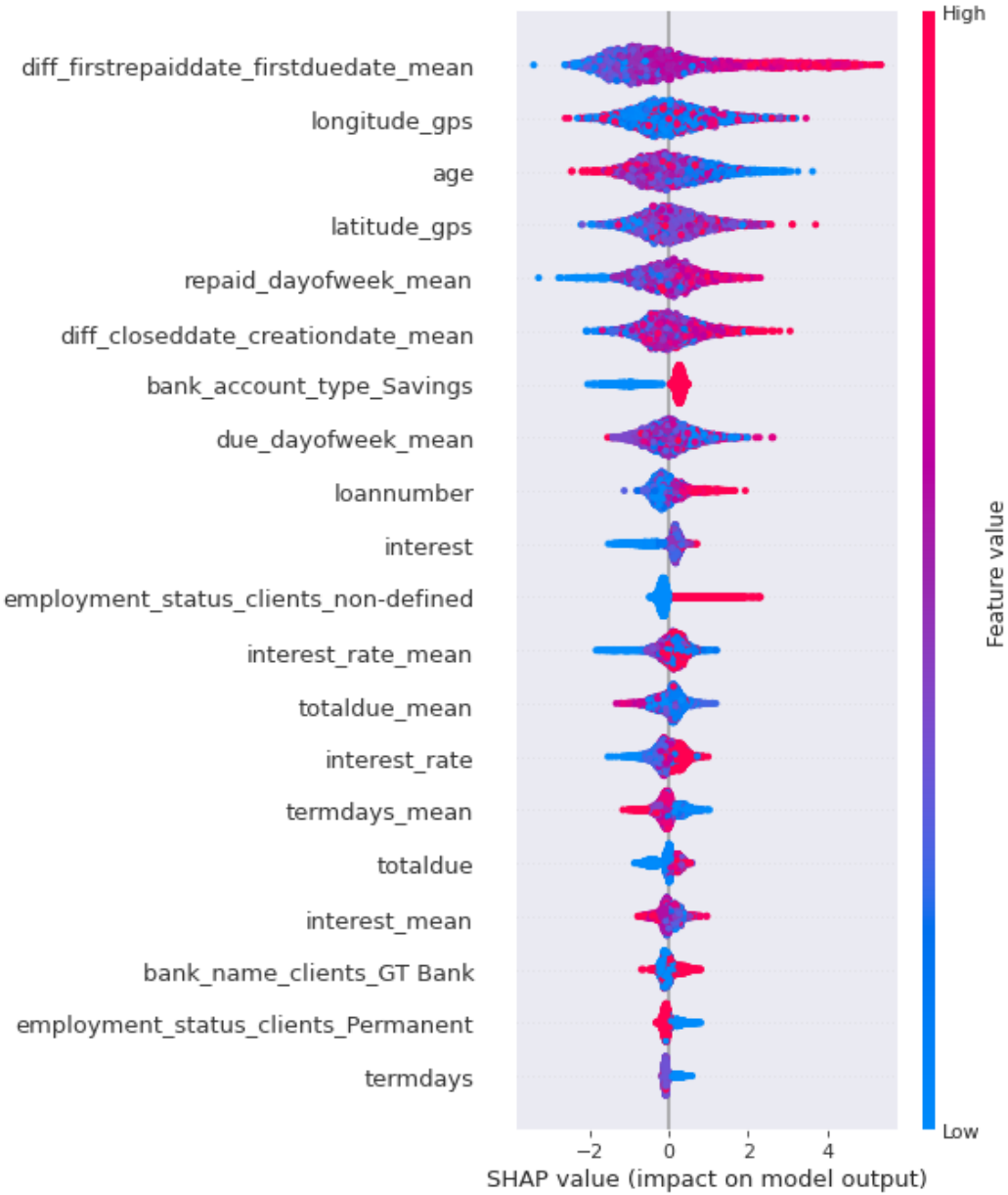
Quelle: https://christophm.github.io/interpretable-ml-book/overview.html

Post-hoc-Methoden (Blackbox-Erklärer)

SHAP (SHapley Additive exPlanations)

Bildquelle: Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. *ACM Computing Surveys*, *57*(2), 1-44.

Eine SHAP-Erklärung lesen



Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik

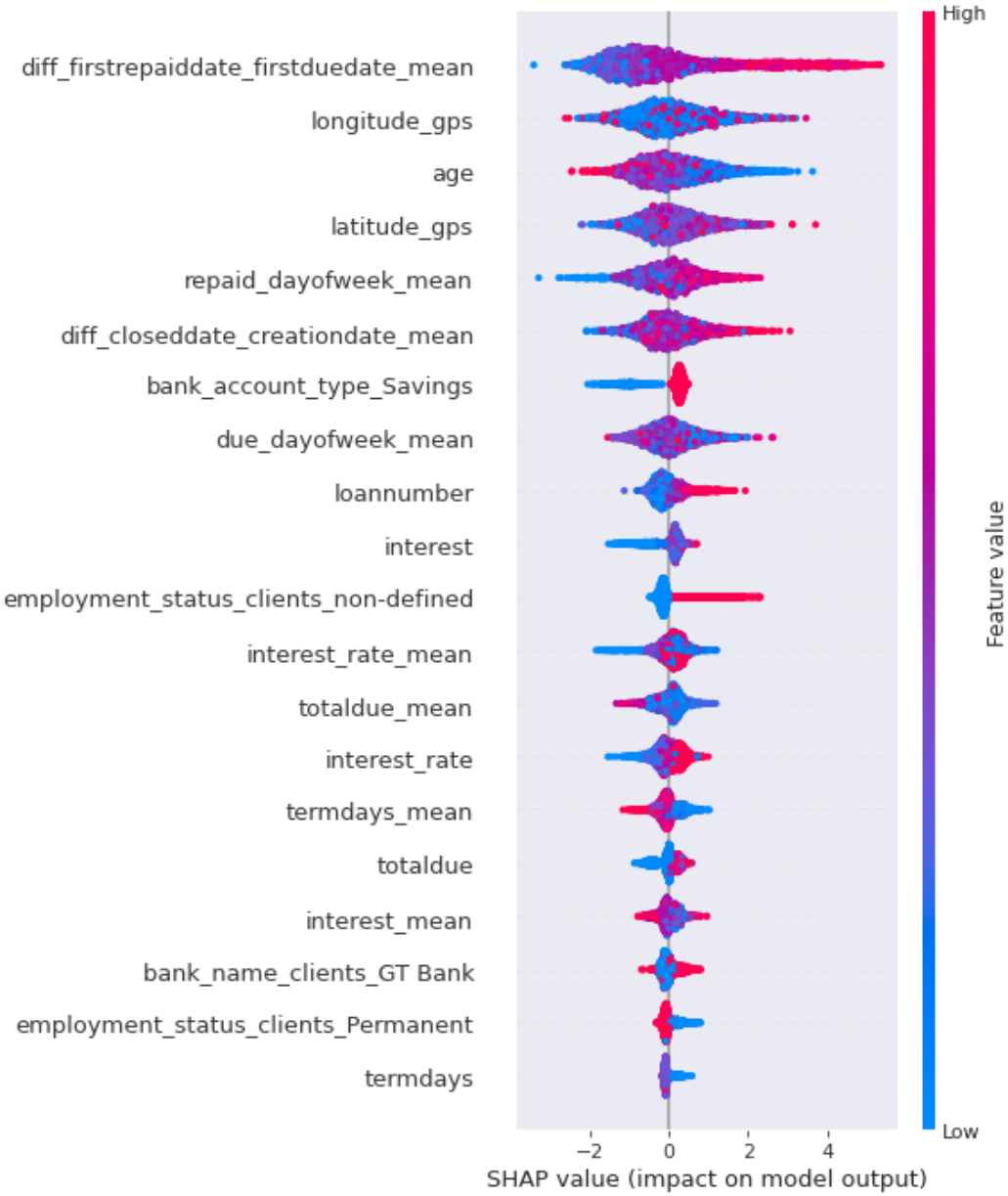
Post-hoc-Methoden (Blackbox-Erklärer)

SHAP (SHapley Additive exPlanations)

Bildquelle: Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. *ACM Computing Surveys*, *57*(2), 1-44.

Eine SHAP-Erklärung lesen

- Anwendung
  - Vorhersage ob man Kredit zurückzahlen wird (negative Klasse: 0) oder nicht (positive Klasse: 1)



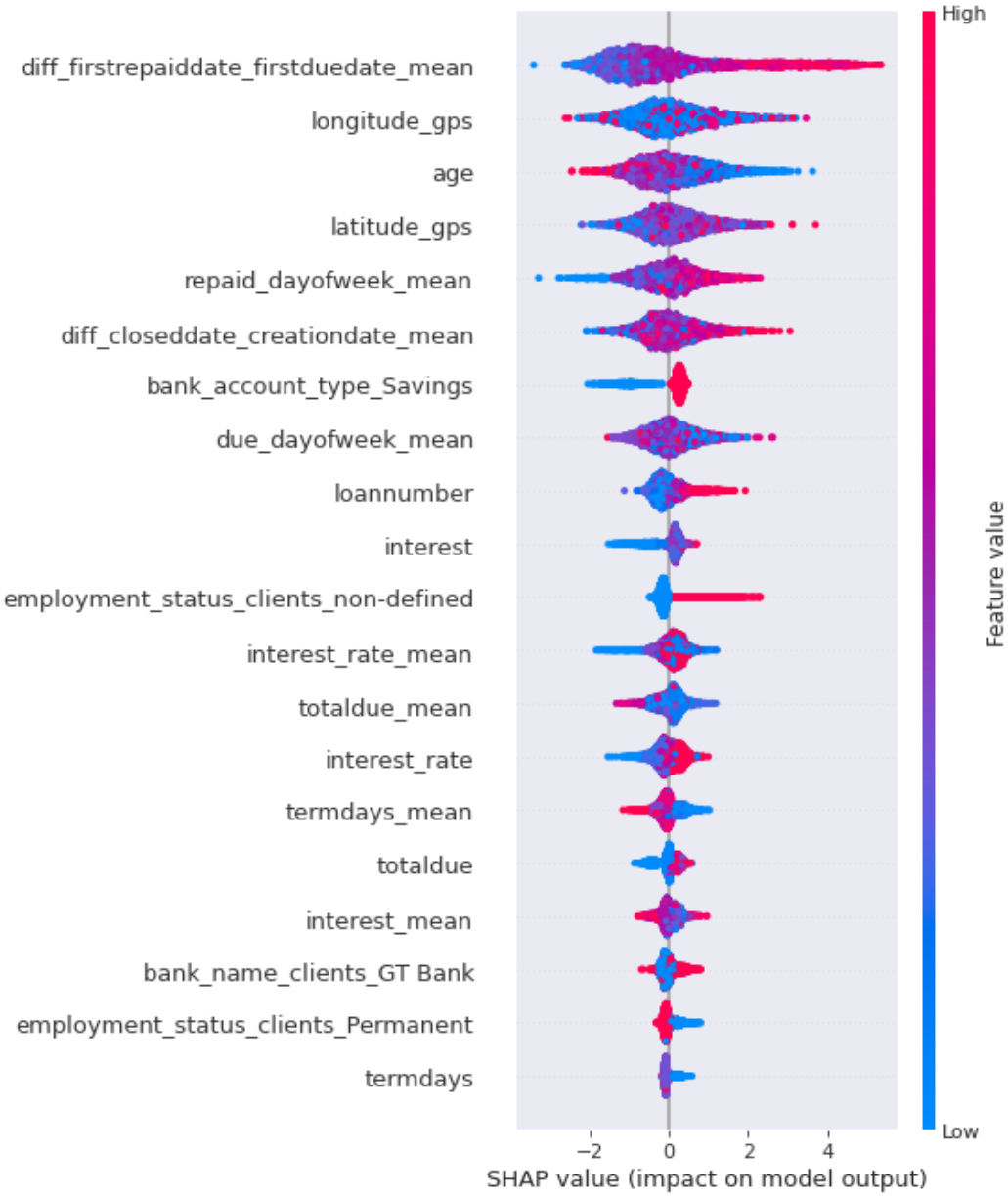
Post-hoc-Methoden (Blackbox-Erklärer)

SHAP (SHapley Additive exPlanations)

Bildquelle: Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. *ACM Computing Surveys*, *57*(2), 1-44.

Eine SHAP-Erklärung lesen

- Anwendung
  - Vorhersage ob man Kredit zurückzahlen wird (negative Klasse: 0) oder nicht (positive Klasse: 1)
  - Merkmale
    - *diff_firstrepaiddate_firstduedate_mean*: Durchschnittliche Verzögerung bei der ersten Zahlung früherer Kredite
    -
    -
    -



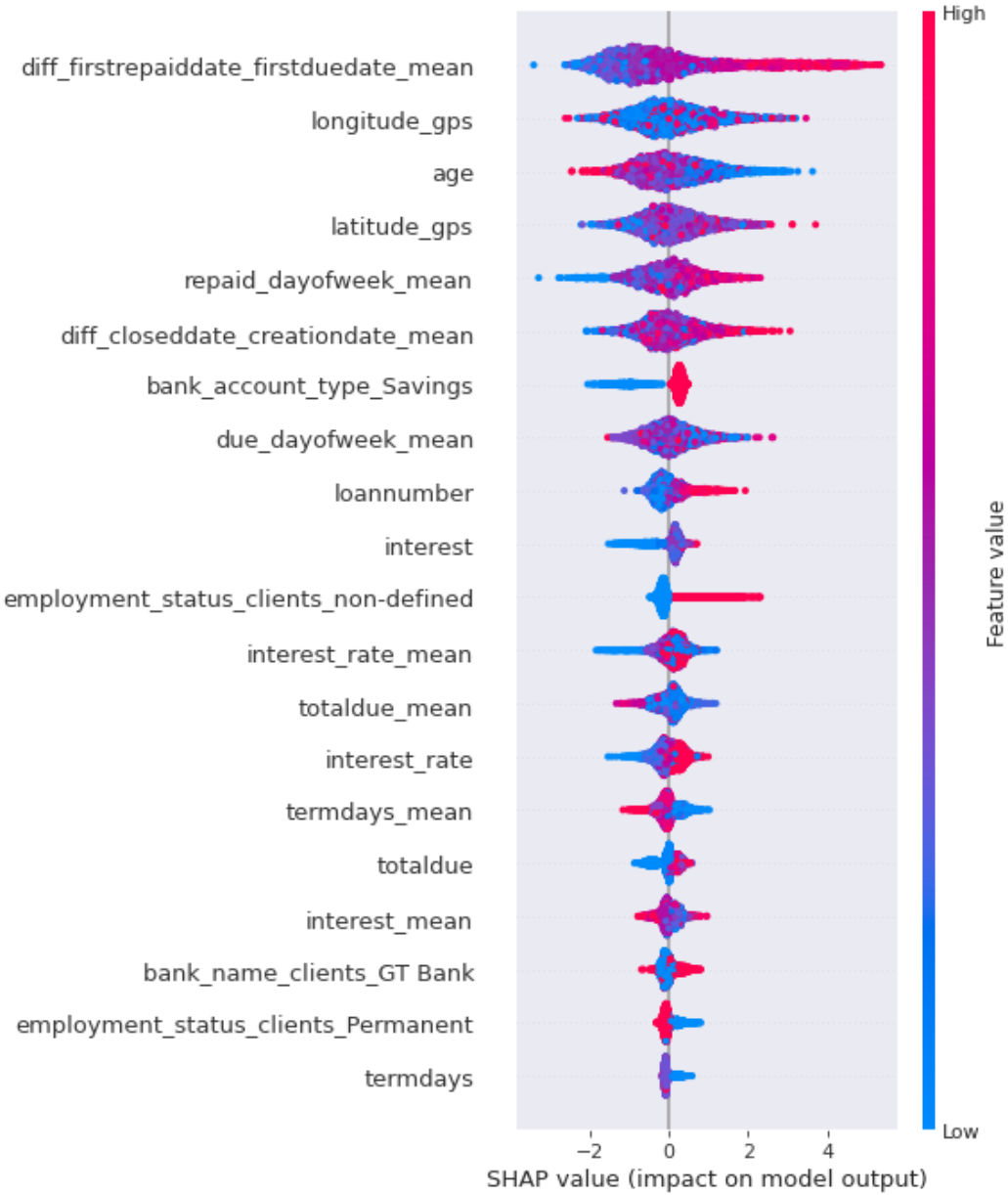
Post-hoc-Methoden (Blackbox-Erklärer)

SHAP (SHapley Additive exPlanations)

Bildquelle: Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. *ACM Computing Surveys*, *57*(2), 1-44.

Eine SHAP-Erklärung lesen

- Anwendung
  - Vorhersage ob man Kredit zurückzahlen wird (negative Klasse: 0) oder nicht (positive Klasse: 1)
  - Merkmale
    - *diff_firstrepaiddate_firstduedate_mean*: Durchschnittliche Verzögerung bei der ersten Zahlung früherer Kredite
    - GPS-Koordinaten des Wohnorts
    - Alter
    - …



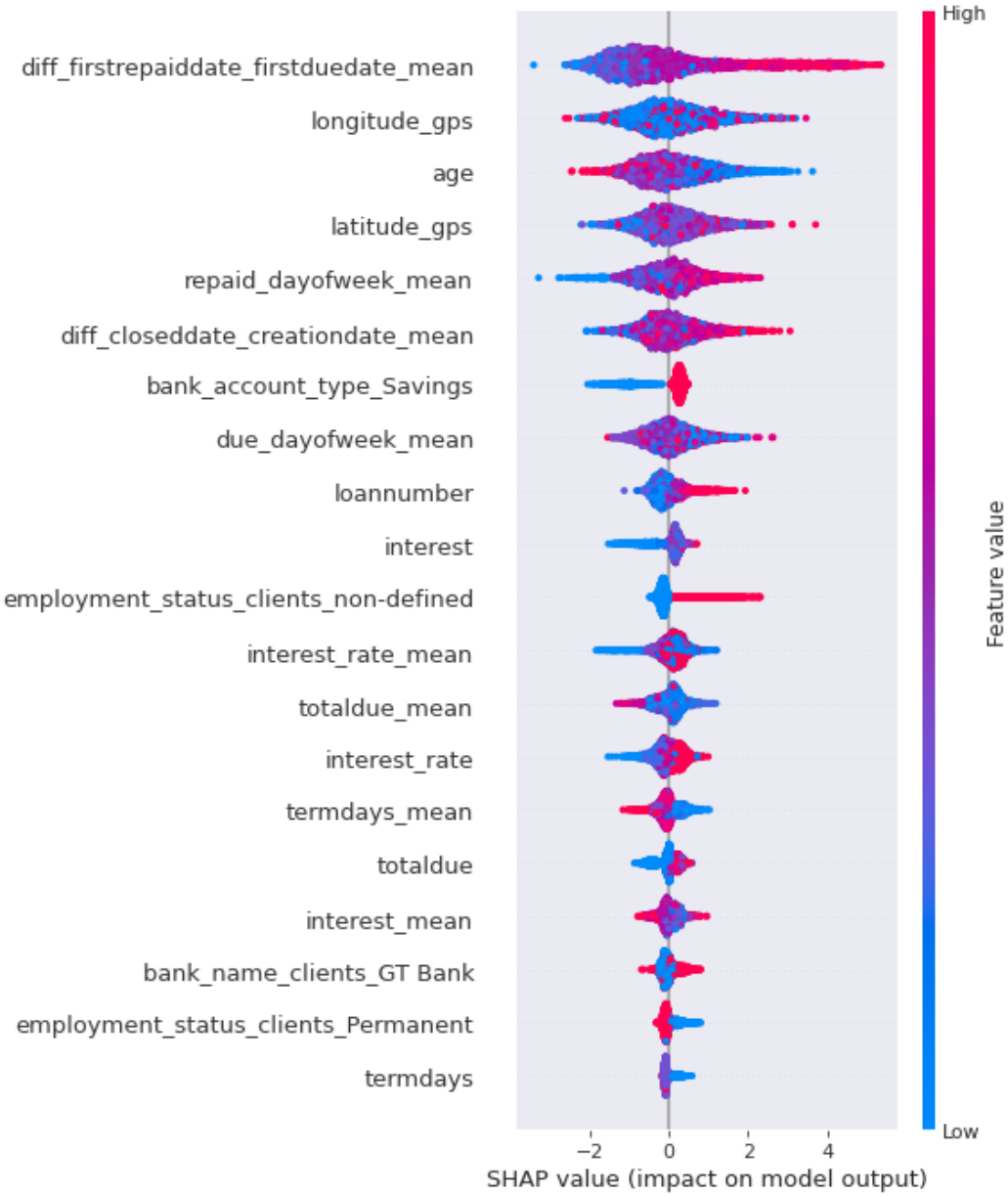
Post-hoc-Methoden (Blackbox-Erklärer)

SHAP (SHapley Additive exPlanations)

Bildquelle: Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. *ACM Computing Surveys*, *57*(2), 1-44.

Eine SHAP-Erklärung lesen

- SHAP Wert (numerische Approximation des Shapley Wertes): Wie stark beeinflusst jedes Feature die Vorhersage?
- Hier: SHAP Summary Plot



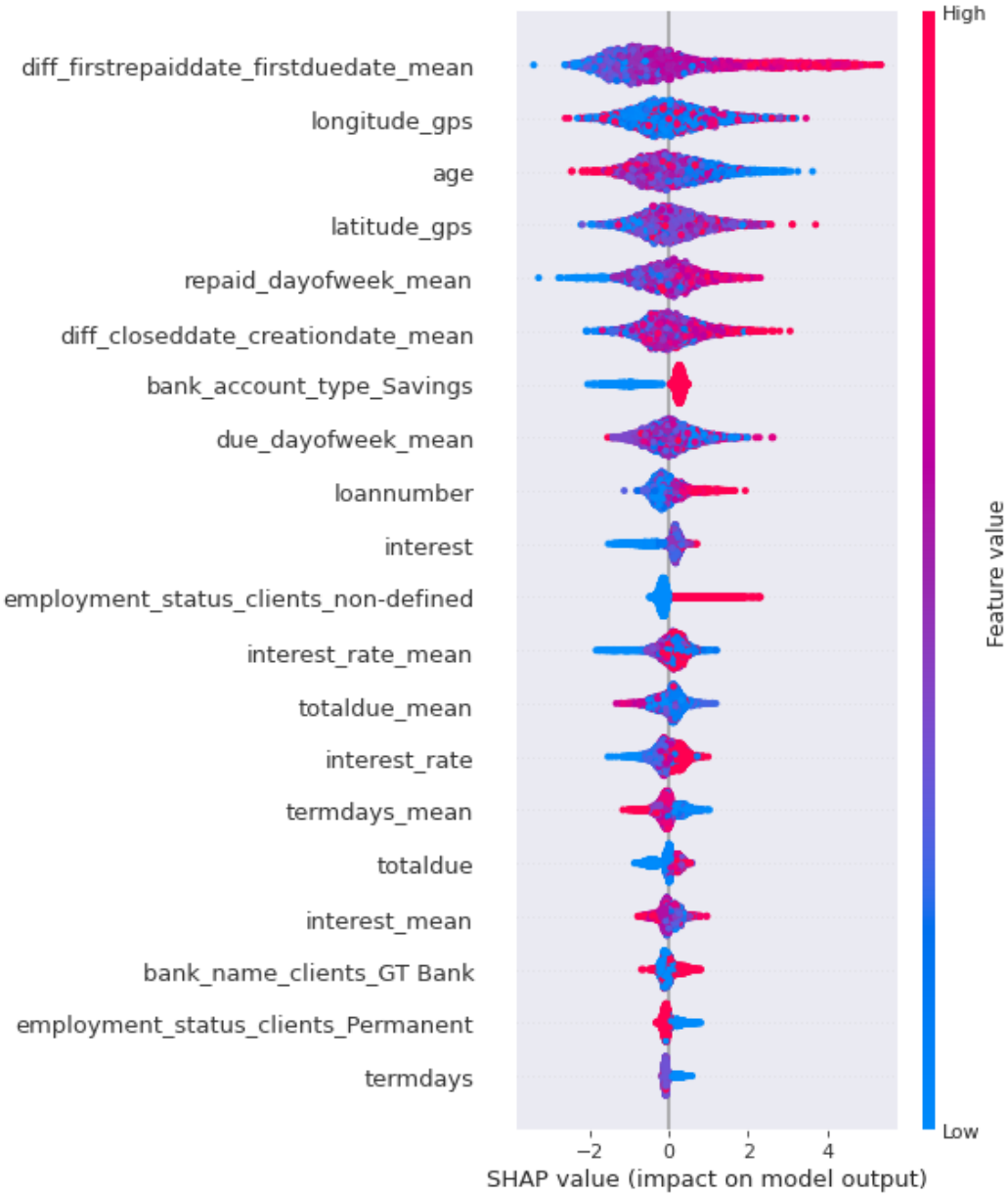
Post-hoc-Methoden (Blackbox-Erklärer)

SHAP (SHapley Additive exPlanations)

Bildquelle: Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. *ACM Computing Surveys*, *57*(2), 1-44.

Eine SHAP-Erklärung lesen

- SHAP Wert (numerische Approximation des Shapley Wertes): Wie stark beeinflusst jedes Feature die Vorhersage?
- Hier: SHAP Summary Plot
- Merkmale sortiert nach SHAP Wert
-
-
-



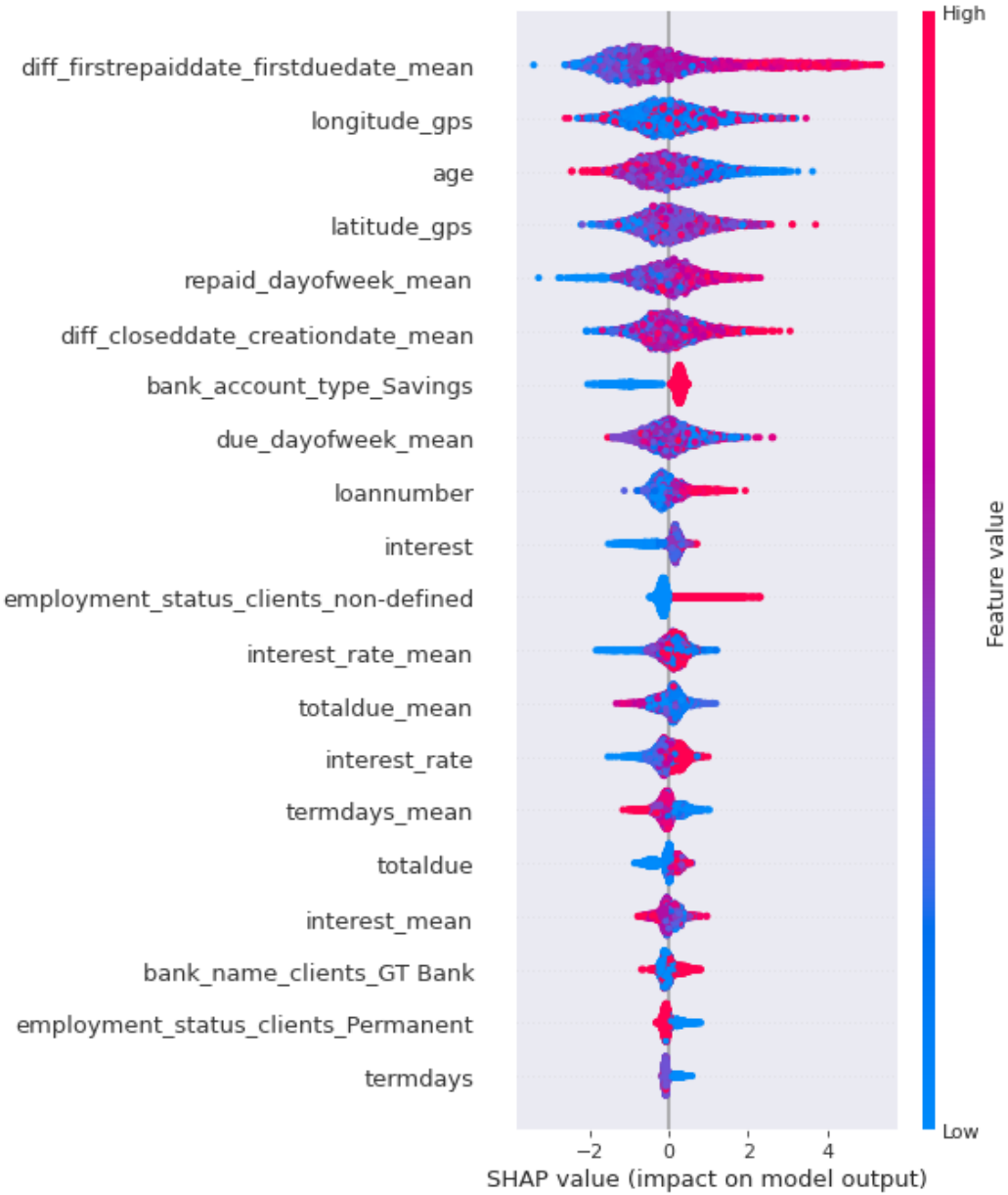
Post-hoc-Methoden (Blackbox-Erklärer)

SHAP (SHapley Additive exPlanations)

Bildquelle: Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. *ACM Computing Surveys*, *57*(2), 1-44.

Eine SHAP-Erklärung lesen

- SHAP Wert (numerische Approximation des Shapley Wertes): Wie stark beeinflusst jedes Feature die Vorhersage?
- Hier: SHAP Summary Plot
- Merkmale sortiert nach SHAP Wert
- Für jeden Datenpunkt (bspw. aus Trainingsdatensatz) bestimmen wir den SHAP Wert mittels Vorhersage der trainierten Modelle
-
-



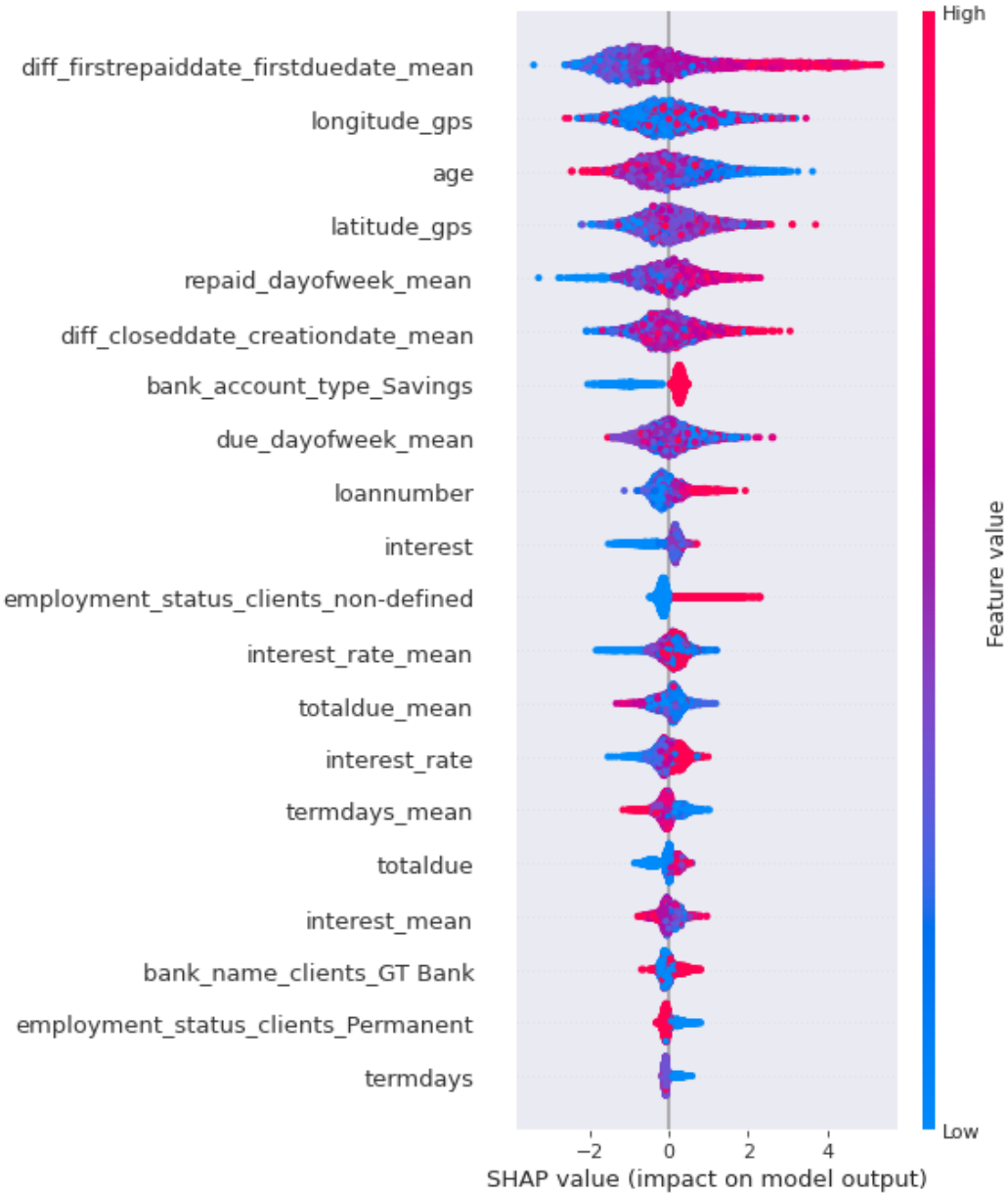
Post-hoc-Methoden (Blackbox-Erklärer)

SHAP (SHapley Additive exPlanations)

Bildquelle: Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. *ACM Computing Surveys*, *57*(2), 1-44.

Eine SHAP-Erklärung lesen

- SHAP Wert (numerische Approximation des Shapley Wertes): Wie stark beeinflusst jedes Feature die Vorhersage?
- Hier: SHAP Summary Plot
- Merkmale sortiert nach SHAP Wert
- Für jeden Datenpunkt (bspw. aus Trainingsdatensatz) bestimmen wir den SHAP Wert mittels Vorhersage der trainierten Modelle
- Farbkodierung: rot = hoher Featurewert, blau = niedriger



Post-hoc-Methoden (Blackbox-Erklärer)

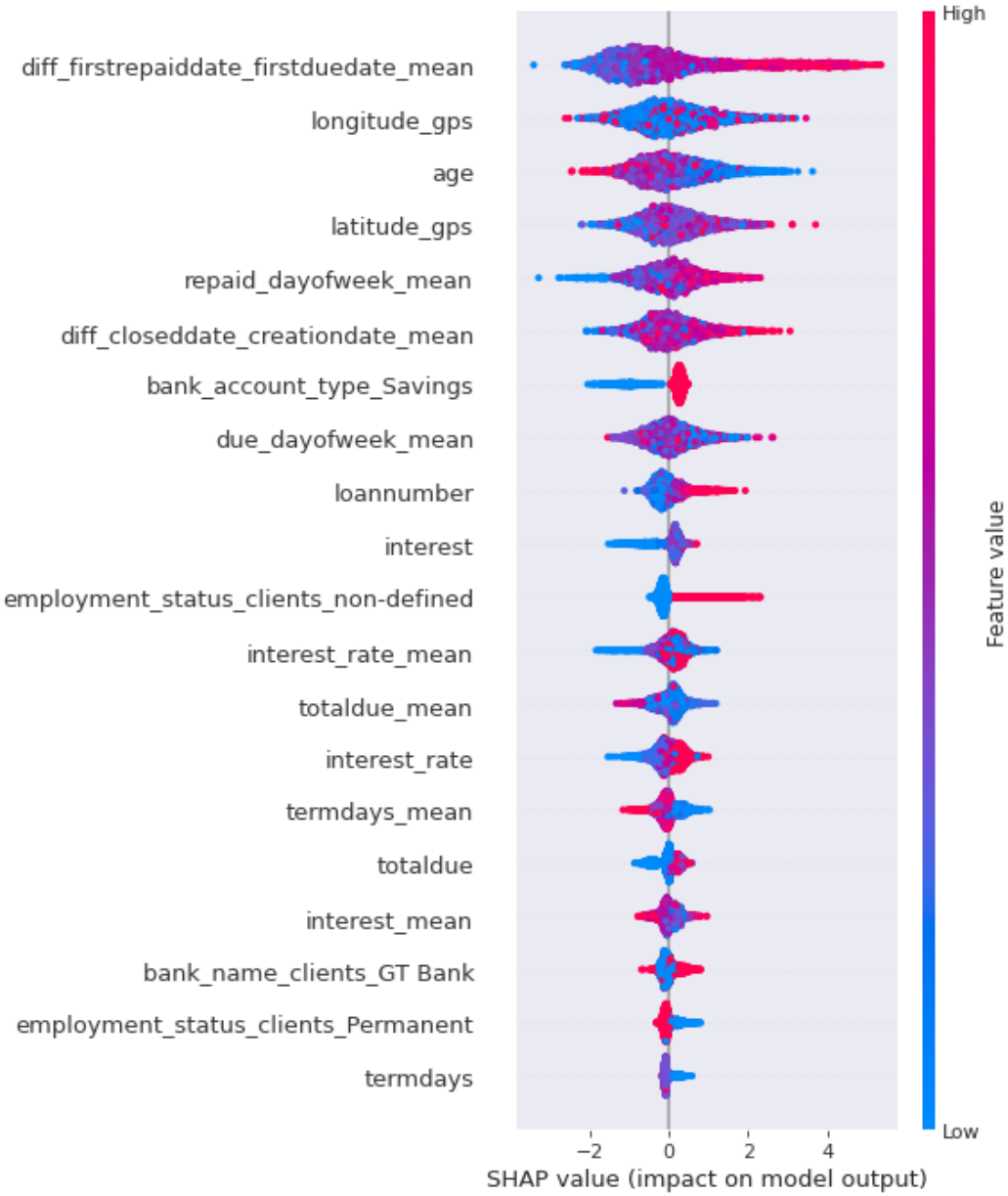
SHAP (SHapley Additive exPlanations)

Bildquelle: Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. *ACM Computing Surveys*, *57*(2), 1-44.

Eine SHAP-Erklärung lesen

- SHAP Wert (numerische Approximation des Shapley Wertes): Wie stark beeinflusst jedes Feature die Vorhersage?
- Hier: SHAP Summary Plot
- Merkmale sortiert nach SHAP Wert
- Für jeden Datenpunkt (bspw. aus Trainingsdatensatz) bestimmen wir den SHAP Wert mittels Vorhersage der trainierten Modelle
- Farbkodierung: rot = hoher Featurewert, blau = niedriger

- Hier erhalten wir eine globale Erklärung



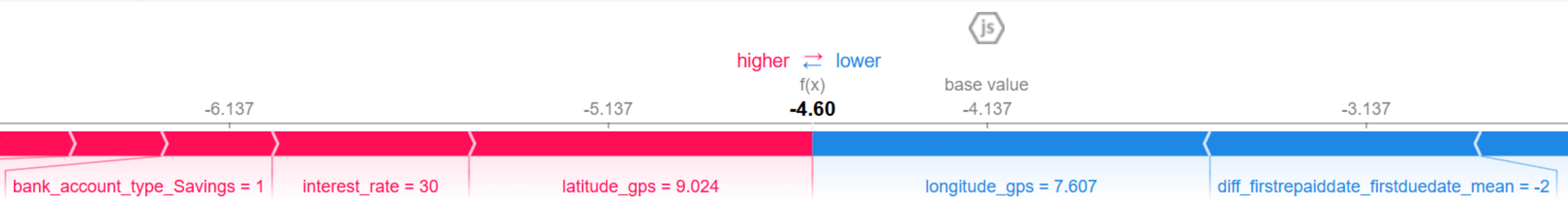
Bildquelle: Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. *ACM Computing Surveys*, *57*(2), 1-44.

Post-hoc-Methoden (Blackbox-Erklärer)
SHAP (SHapley Additive exPlanations)

Eine SHAP-Erklärung lesen

- Hier: SHAP Force Plot

- Farbkodierung:
  - rot = Kräfte, die zur positiven Klasse führen
  - blau = Kräfte, die zur negativen Klasse führen



Bildquelle: Bennetot, A., Donadello, I., El Qadi El Haouari, A., Dragoni, M., Frossard, T., Wagner, B., ... & Diaz-Rodriguez, N. (2024). A practical tutorial on explainable AI techniques. *ACM Computing Surveys*, *57*(2), 1-44.

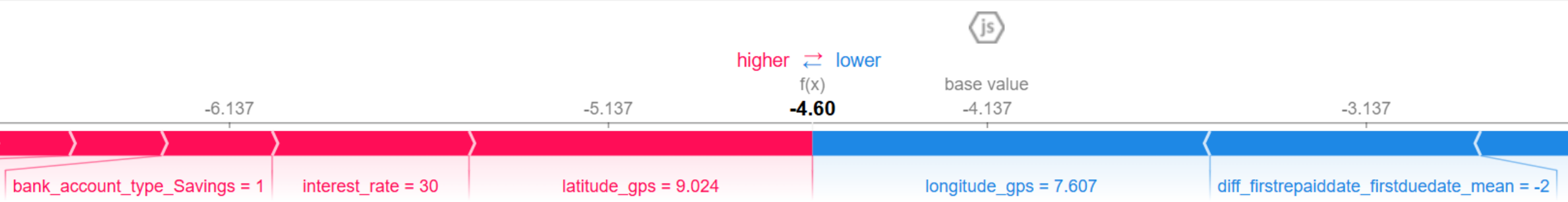
Post-hoc-Methoden (Blackbox-Erklärer)
SHAP (SHapley Additive exPlanations)

Eine SHAP-Erklärung lesen

- Hier: SHAP Force Plot

- Farbkodierung:
  - rot = Kräfte, die zur positiven Klasse führen
  - blau = Kräfte, die zur negativen Klasse führen

- Lokale Erklärbarkeit



Post-hoc-Methoden (Blackbox-Erklärer)
SHAP (SHapley Additive exPlanations)

Vorteile

- Basiert auf Spieltheorie
- Modellagnostisch: funktioniert für sehr viele ML-Modelle
- Man erhält Einsicht sowohl für die lokale als auch globale Erklärbarkeit

Herausforderungen

Post-hoc-Methoden (Blackbox-Erklärer)
SHAP (SHapley Additive exPlanations)

Vorteile

- Basiert auf Spieltheorie
- Modellagnostisch: funktioniert für sehr viele ML-Modelle
- Man erhält Einsicht sowohl für die lokale als auch globale Erklärbarkeit

Herausforderungen

- Komplexität / Rechenaufwand (v. a. bei Deep Learning)
- Interpretation für Endnutzer*innen kann erklärungsbedürftig sein
- Nur eingeschränkt visuell verständlich

Post-hoc-Methoden (Blackbox-Erklärer)
SHAP (SHapley Additive exPlanations)

Vorteile

- Basiert auf Spieltheorie
- Modellagnostisch: funktioniert für sehr viele ML-Modelle
- Man erhält Einsicht sowohl für die lokale als auch globale Erklärbarkeit

- Beispiel: s. Übung

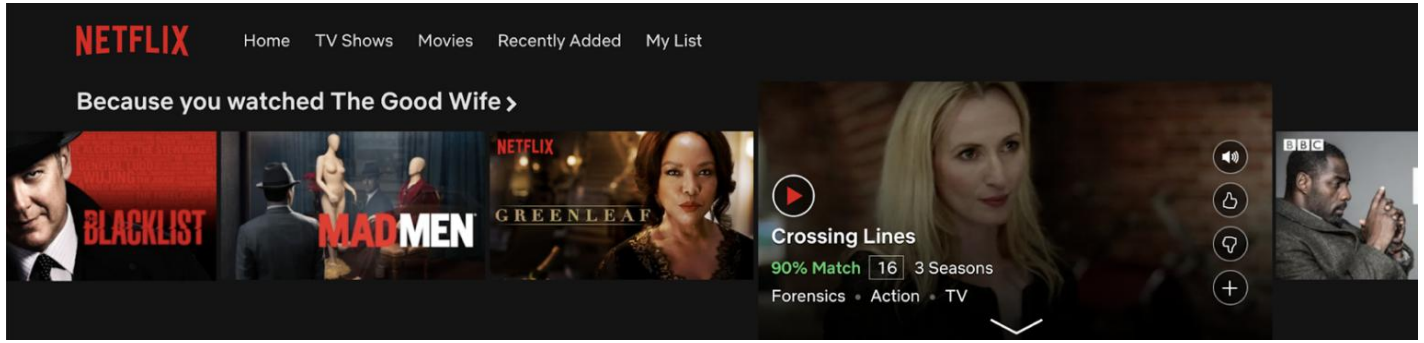
Herausforderungen

- Komplexität / Rechenaufwand (v. a. bei Deep Learning)
- Interpretation für Endnutzer*innen kann erklärungsbedürftig sein
- Nur eingeschränkt visuell verständlich

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Erklärungen heutiger Systeme

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020



Source: www.netflix.com

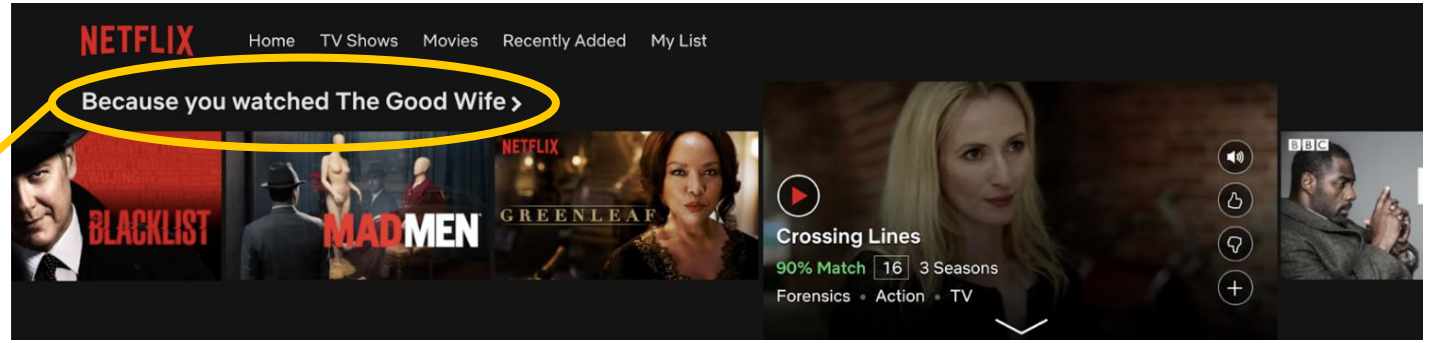
Source: www.amazon.de

Erklärungen heutiger Systeme

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

"Warum"

Erklärung

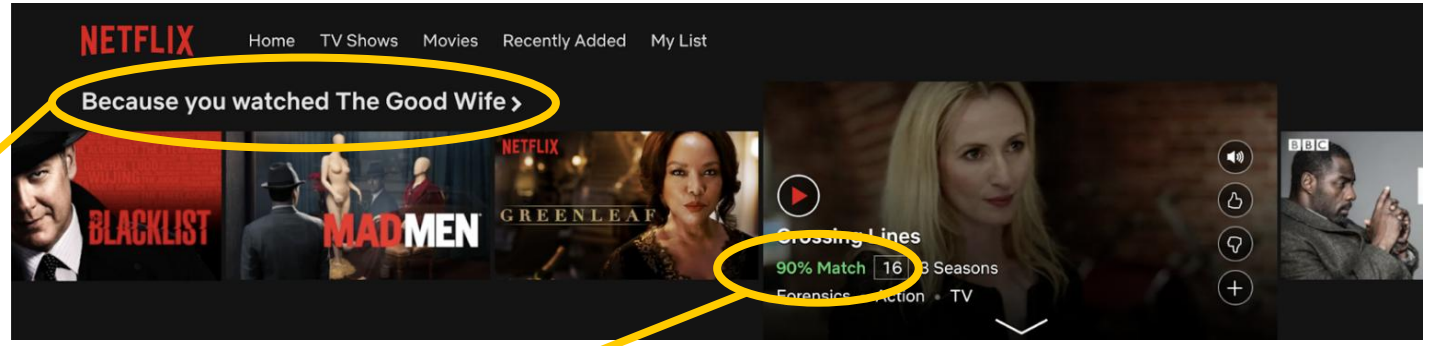


Source: www.netflix.com

Source: www.amazon.de

Erklärungen heutiger Systeme

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020



Source: www.netflix.com

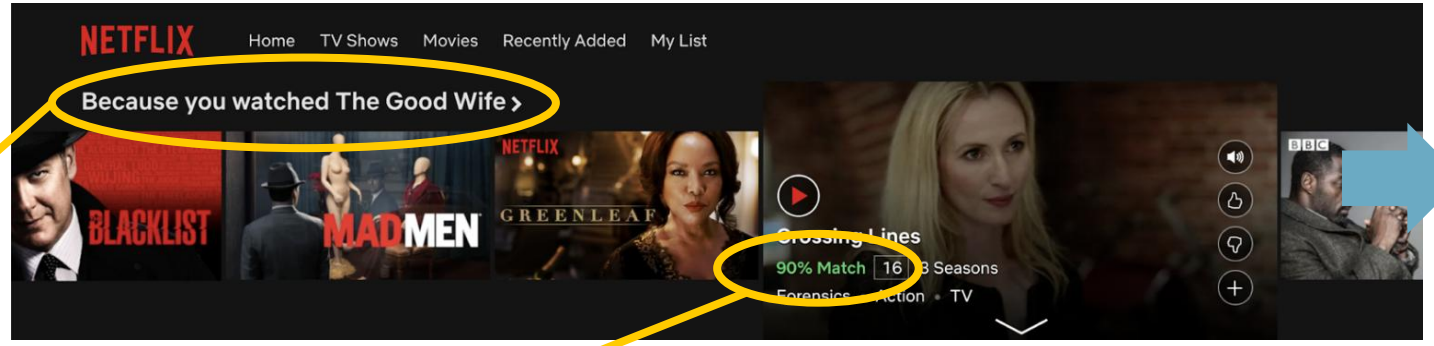
“Warum”
Erklärung

“Sicherheit”
Erklärung

Source: www.amazon.de

Erklärungen heutiger Systeme

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020



Source: www.netflix.com

Source: www.amazon.de

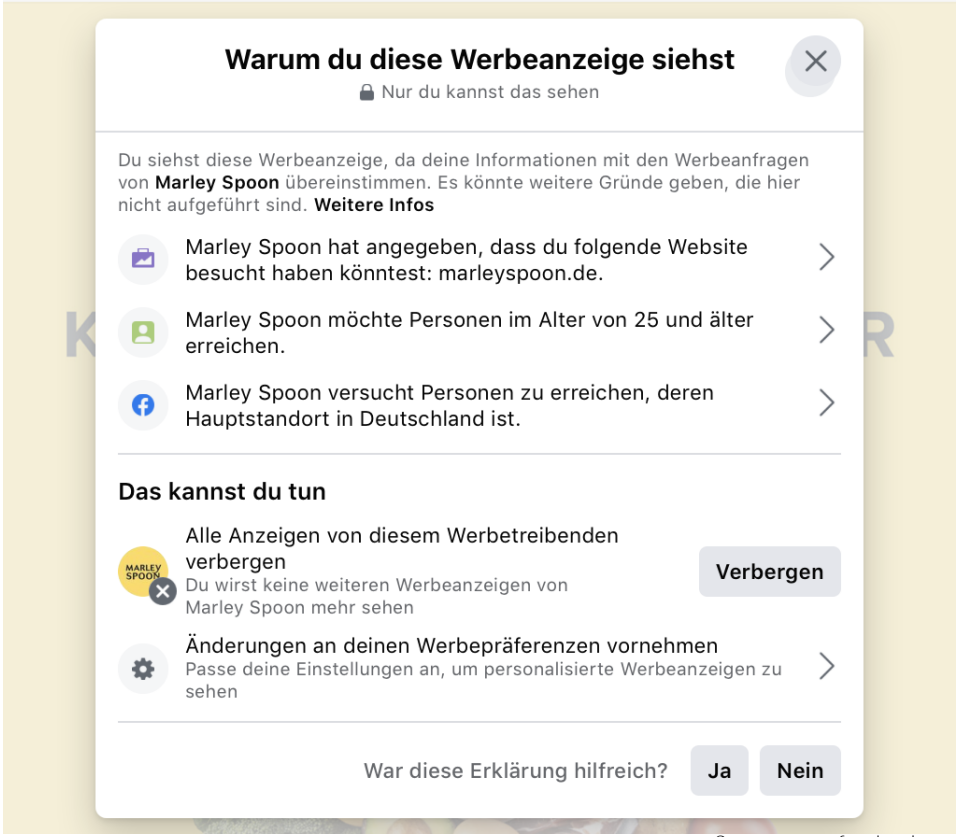
“Warum”
Erklärung

“Sicherheit”
Erklärung

Transparenz
Vertrauen
Effektivität
Überzeugungskraft

Erklärungen heutiger Systeme

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

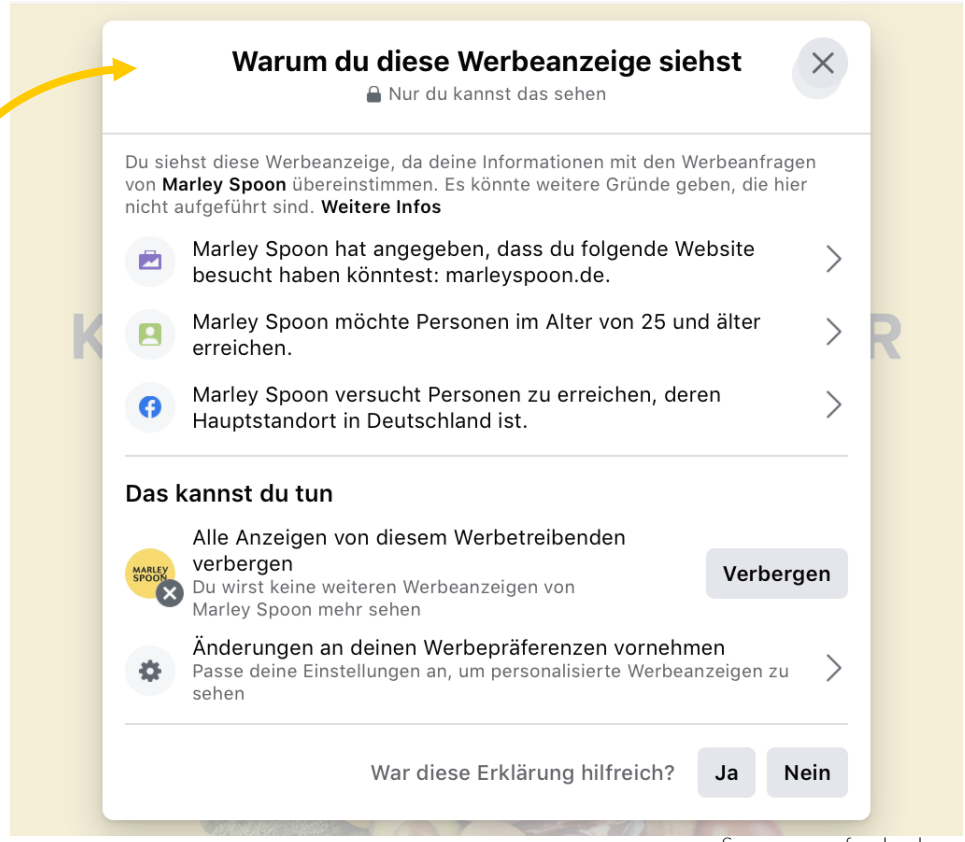


Source: www.facebook.com

Erklärungen heutiger Systeme

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

"Warum"
Erklärung

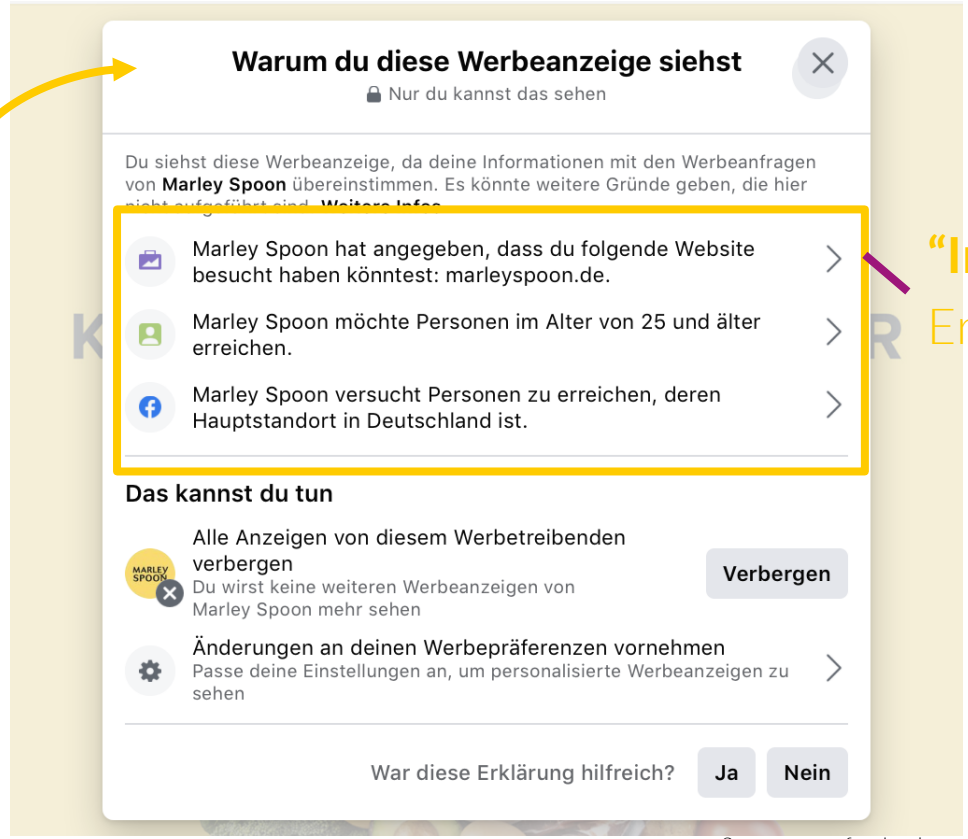


Source: www.facebook.com

Erklärungen heutiger Systeme

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

“Warum” Erklärung

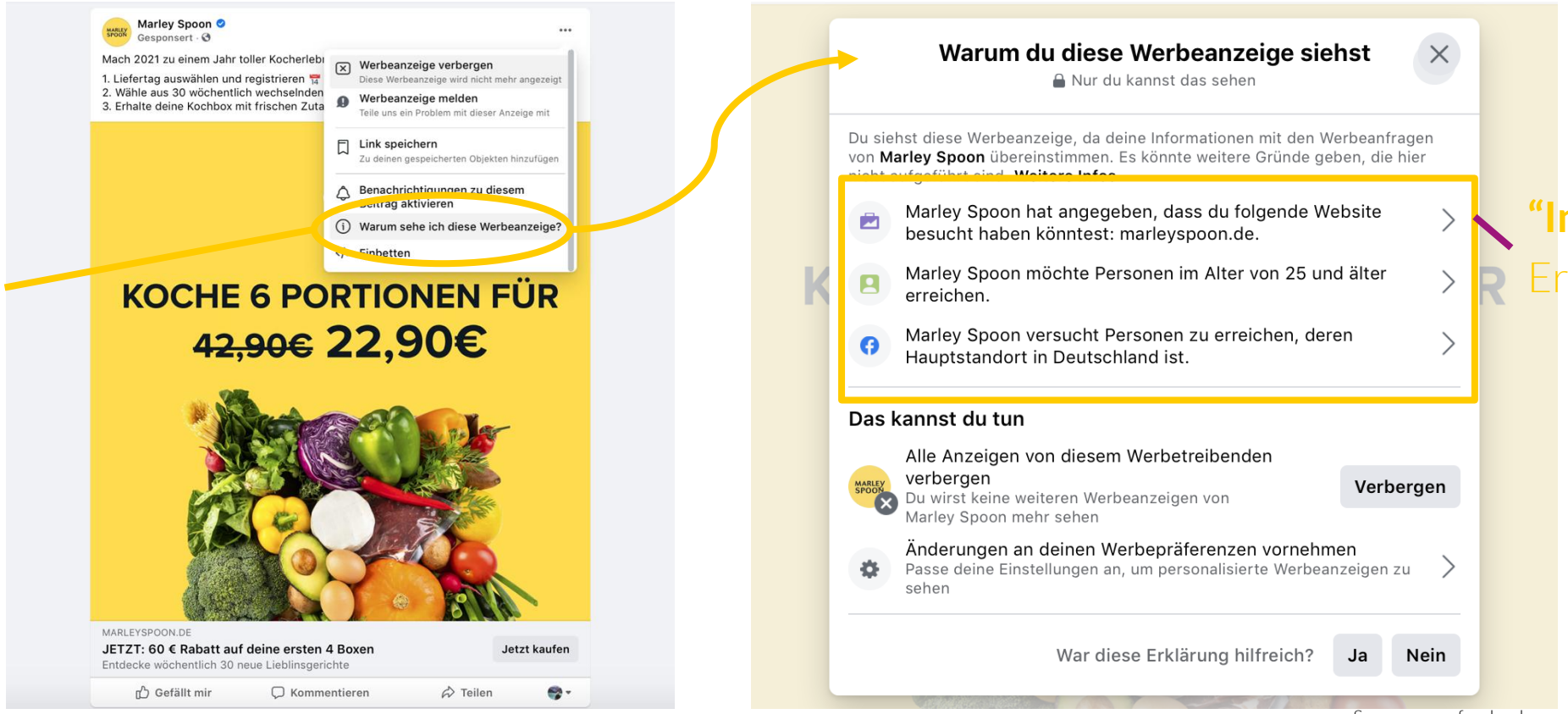


“Inputs” Erklärung

Source: www.facebook.com

Erklärungen heutiger Systeme

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020



Source: www.facebook.com

Transparenz
Überprüfbarkeit

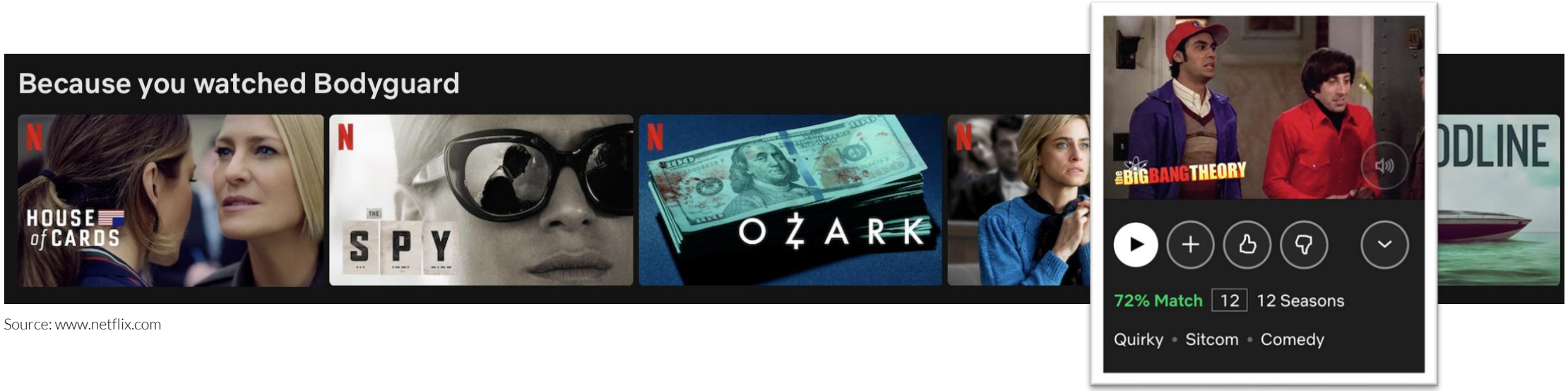
Zufriedenstellung

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Diskussion

Quelle: Introduction to Intelligent User Interfaces, Sarah Theres Völkel, LMU, 2020

Wie würden Sie die **Erklärung von Netflix verbessern**, warum ein bestimmter Film empfohlen wurde?



Source: www.netflix.com

Zusammenfassung

- XAI ist nicht nur eine technische Ergänzung, sondern zentral für menschenzentrierte Gestaltung.
- Wichtig ist nicht nur ob ein Modell erklärt, sondern wie, für wen und zu welchem Zweck.

Zusammenfassung

- XAI ist nicht nur eine technische Ergänzung, sondern zentral für menschenzentrierte Gestaltung.
- Wichtig ist nicht nur ob ein Modell erklärt, sondern wie, für wen und zu welchem Zweck.

- Sie können interpretierbare Modelle nutzen …
  - Lineare/Logistische Regression, kleine Entscheidungsbäume, …

-
  -
  -
  -

Prof. Dr. Daniel Gaida
Professor für Cyber-Physische Systeme
Fakultät für Informatik und Ingenieurwissenschaften - Institut für Informatik
Technology Arts Sciences TH Köln

Zusammenfassung

- XAI ist nicht nur eine technische Ergänzung, sondern zentral für menschenzentrierte Gestaltung.
- Wichtig ist nicht nur ob ein Modell erklärt, sondern wie, für wen und zu welchem Zweck.

- Sie können interpretierbare Modelle nutzen …
  - Lineare/Logistische Regression, kleine Entscheidungsbäume, …

- … oder BlackBox-Modelle versuchen zu erklären mit …
  - LIME
  - SHAP
  - …

Nächste Vorlesung und Fragen

- Trainingsstrategien (Strategy for DL Troubleshooting)
- Fragen?