UNIVERSITÄT ZU LÜBECK
INSTITUTE FOR SOFTWARE ENGINEERING
AND PROGRAMMING LANGUAGES

Erklärbare KI

Warum wir sie brauchen und wie wir dort hinkommen

Dr. Gesina Schwalbe

KI-Landeskonferenz SH 2024

Outline

Warum brauchen wir XAI?

Wie kann man KI erklären?

Herausforderungen und Zusammenfassung

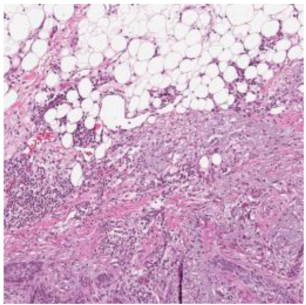
Gliederung

Warum brauchen wir XAI?

Warum brauchen wir XAI?

Warum brauchen wir XAI?

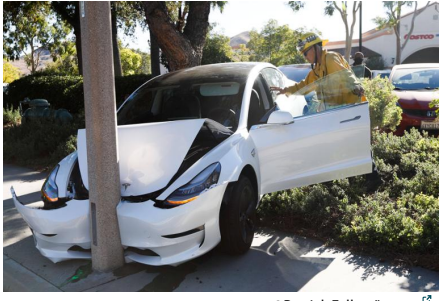
SCHUFA-
BonitätsAuskunft



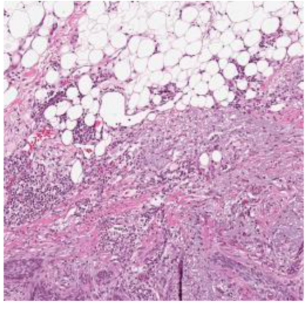
(Pocevičiūtė u. a. 2020, Fig. 6)

Warum brauchen wir XAI?

SCHUFA-
BonitätsAuskunft



©Patrick Fallon/Imago



(Pocevičiūtė u. a. 2020, Fig. 6)

Wo brauchen wir Erklärungen?

Anwendungsfälle:

- Endnutzende:
  - (angemessenes!) **Vertrauen**, informierte Zustimmung
  - Einarbeitung
  - **Regressansprüche**
- Entwicklung und Fachanwendung:
  - **Debugging**
  - Knowledge retrieval
- Gutachten:
  - Compliance *(Gesetze, Standards)*
  - **Assessments**, *(Safety, Fairness,…)*

Wo brauchen wir Erklärungen?

Anwendungsfälle:

- Endnutzende:
  - (angemessenes!) **Vertrauen**, informierte Zustimmung
  - Einarbeitung
  - **Regressansprüche**
- Entwicklung und Fachanwendung:
  - **Debugging**
  - Knowledge retrieval
- Gutachten:
  - Compliance *(Gesetze, Standards)*
  - **Assessments**, *(Safety, Fairness,…)*

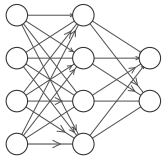
Anwendungsfelder:

Sobald automatisierte Entscheidungen Wohlergehen von Menschen beeinflussen!

- **Ranking**-Systeme *(Kredite, Bewerbungen, …)*
- **Medizinische** Assistenzsysteme
- Automatisiertes **Fahren**
- **Militärische** Entscheidungssysteme
- **HMI** in der Produktion
- …

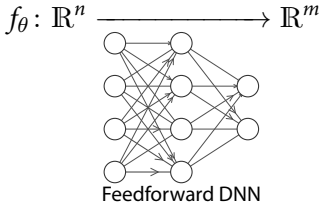
Hineinschauen ist schwer: Am Beispiel von DNNs.

$f_\theta : \mathbb{R}^n \longrightarrow \mathbb{R}^m$



Feedforward DNN

Hineinschauen ist schwer: Am Beispiel von DNNs.

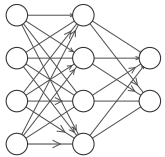


Herausforderungen:

- **Größe:** *YOLOv9, DETR:* ***>50 M*** *Param.*
  Llama 3.2: ***90 B*** *Param.*

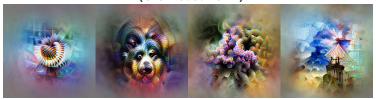
Hineinschauen ist schwer: Am Beispiel von DNNs.

$f_\theta : \mathbb{R}^n \longrightarrow \mathbb{R}^m$



Feedforward DNN

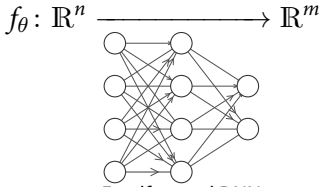
(Olah u. a. 2017)



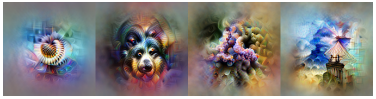
Herausforderungen:

- **Größe:** *YOLOv9, DETR:* ***>50 M*** *Param.*
  Llama 3.2: ***90 B*** *Param.*
- **Automatisch** erlernte Repräsentationen:
  - **verteilt** $\Rightarrow$ schwer zu lesen

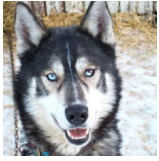
Hineinschauen ist schwer: Am Beispiel von DNNs.



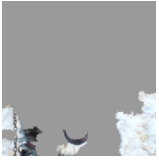
(Olah u. a. 2017)



(Marco Túlio Ribeiro u. a. 2016, Fig. 11)



***Husky* image** misclassified as *Wolf*



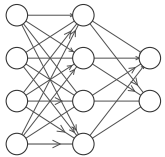
features most influential for the decision

Herausforderungen:

- **Größe:** *YOLOv9, DETR:* ***>50 M** Param.*
  Llama 3.2: ***90 B** Param.*
- **Automatisch** erlernte Repräsentationen:
  - **verteilt** ⇒ schwer zu lesen
  - nur **Korrelationen** ⇒ schwer einzuschätzen

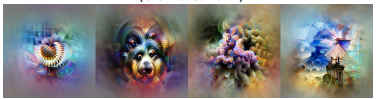
Hineinschauen ist schwer: Am Beispiel von DNNs.

$f_\theta : \mathbb{R}^n \longrightarrow \mathbb{R}^m$

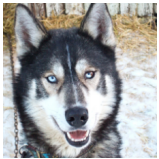


Feedforward DNN

(Olah u. a. 2017)



(Marco Túlio Ribeiro u. a. 2016, Fig. 11)



***Husky* image** misclassified as *Wolf*

features most influential for the decision

Herausforderungen:

- **Größe:** *YOLOv9, DETR:* ***>50 M*** *Param.*
  Llama 3.2: ***90 B*** *Param.*
- **Automatisch** erlernte Repräsentationen:
  - **verteilt** ⇒ schwer zu lesen
  - nur **Korrelationen** ⇒ schwer einzuschätzen

⇒ black-box.

Hineinschauen reicht nicht: Es braucht Verständnis.

Definition (Verstehen)

erfolgreiche Aktualisierung des mentalen Modells; entweder
mechanistisch = wie es funktioniert, oder
funktional = was ist der Zweck.

Hineinschauen reicht nicht: Es braucht Verständnis.

Definition (Verstehen)

erfolgreiche Aktualisierung des mentalen Modells; entweder
mechanistisch = wie es funktioniert, oder
funktional = was ist der Zweck.

Definition (Stufen von Transparenz)

- simulierbar = verständlich als Ganzes
- zerlegbar in simulierbare Teile
- algorithmisch transparent = mathematisch verständlich

EU AI Act

Preamble (72)

[Es] sollte für Hochrisiko-KI-Systeme Transparenz vorgeschrieben werden *[…]. [Sie] sollten so gestaltet sein, dass die Betreiber in der Lage sind, zu* ***verstehen****, wie das KI-System funktioniert. […]*

Artikel 13

1. ***Hochrisiko-KI-Systeme*** *werden so konzipiert und entwickelt, dass ihr Betrieb* ***hinreichend transparent*** *ist, damit die Betreiber die Ausgaben eines Systems angemessen* ***interpretieren und verwenden können****. […]*

Was wir wollen: Erklärbare KI (XAI)

Definition (Erklärb. Entscheidungssys.)

Es gibt einen Mechanismus, der
eine Erklärung erzeugt (= *Explanator*)
für eine Person (= *Explainee*), die dadurch
eines von (= *Explanandum*) *versteht*:

- die **Information** in Modell/-teilen,
- Einfluss von Modell**ausgaben**, oder
- den **Schlussfolgerung**sprozess.

(Lacave u. a. 2001)

Was wir wollen: Erklärbare KI (XAI)

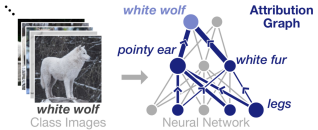
Definition (Erklärb. Entscheidungssys.)

Es gibt einen Mechanismus, der
eine Erklärung erzeugt (= *Explanator*)
für eine Person (= *Explainee*), die dadurch
eines von (= *Explanandum*) *versteht*:

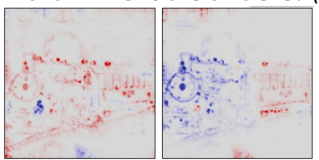
- die **Information** in Modell/-teilen,
- Einfluss von Modell**ausgaben**, oder
- den **Schlussfolgerung**sprozess.

(Lacave u. a. 2001)

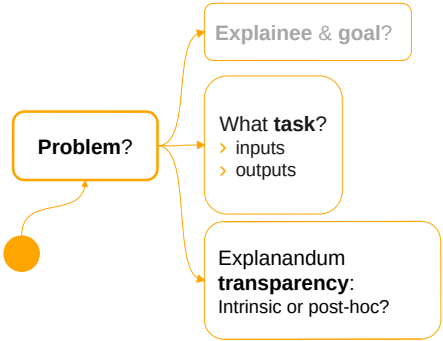
⇒ **Wie** funktioniert es? *(global)*



⇒ **Warum** diese Entscheidung? *(local)*
Warum nicht die andere? *(kontrastiv)*)

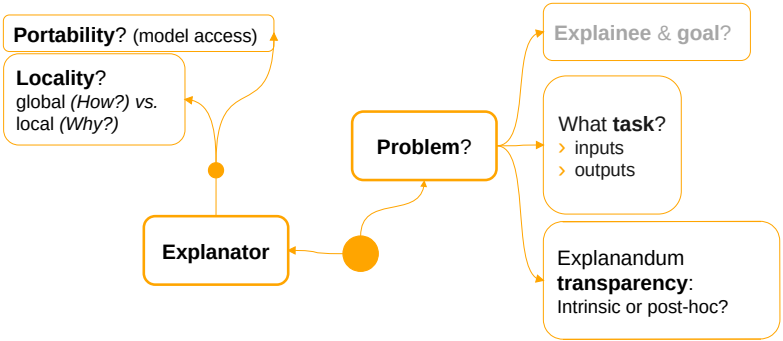


Was sollte bedacht werden? Eine Taxonomie.



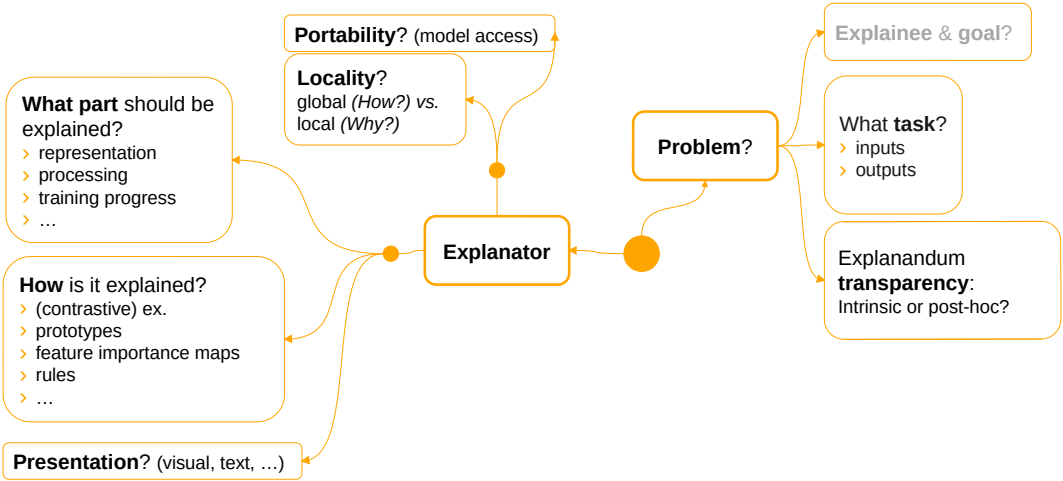
Gesina Schwalbe u. a. (Jan. 2023). „A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts“. In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00867-8. (Besucht am 09. 01. 2023)

Was sollte bedacht werden? Eine Taxonomie.



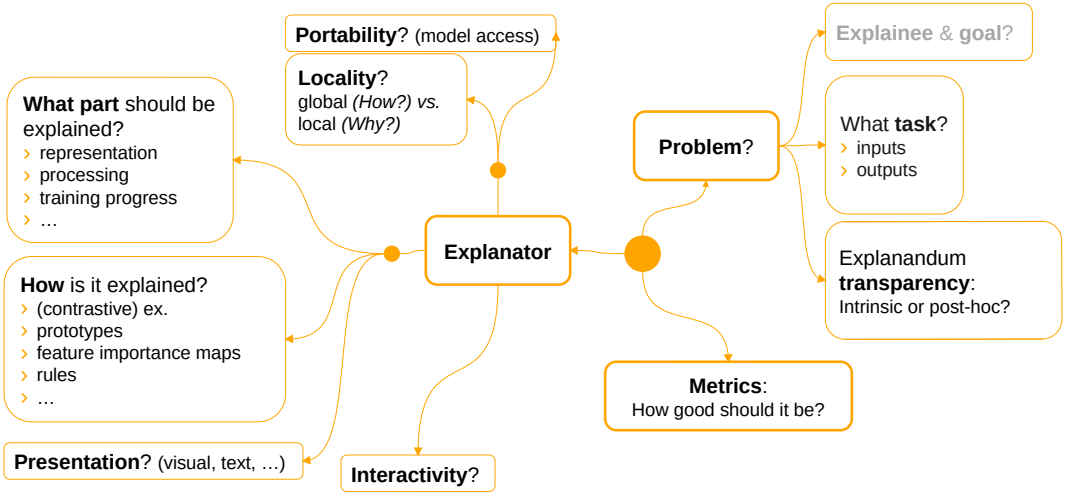
Gesina Schwalbe u. a. (Jan. 2023). „A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts“.
In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00867-8. (Besucht am 09. 01. 2023)

Was sollte bedacht werden? Eine Taxonomie.



Gesina Schwalbe u. a. (Jan. 2023). „A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts“. In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00867-8. (Besucht am 09. 01. 2023)

Was sollte bedacht werden? Eine Taxonomie.



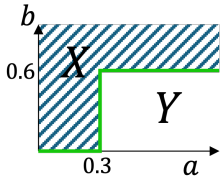
Gesina Schwalbe u. a. (Jan. 2023). „A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts". In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00867-8. (Besucht am 09. 01. 2023)

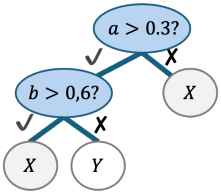
Gliederung

Inhärent transparente Modelle

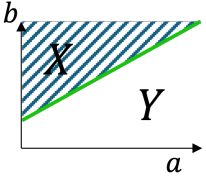
Wenn möglich, baue es von vorneherein transparent. (Rudin 2019)

Decision **Trees**



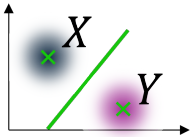


Linear Models



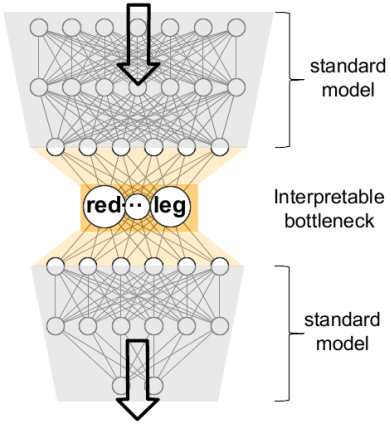
$$f(x) = \alpha a + \beta b$$

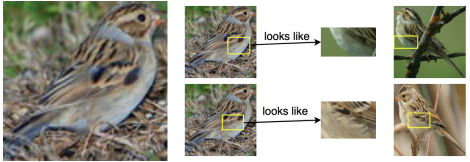
Clusters / **Prototypes**



Modularisierung: Mischmodelle

Modularisierung: Mischmodelle





(Chen u. a. 2019)

Ein erster Schritt: Feature Visualization

Question: What does the output of a network unit/part (z.B., neuron, channel) encode?

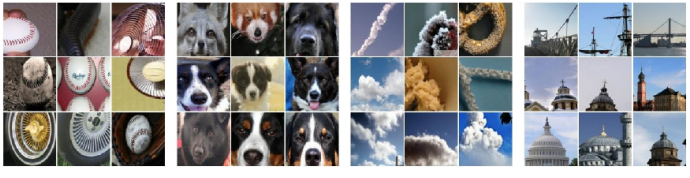
(Olah u. a. 2017, Fig. 5)

Ein erster Schritt: Feature Visualization

Question: What does the output of a network unit/part (z.B., neuron, channel) encode?

Examples
activating unit strongly

DeepDream Prototypes
= starting image optimized to activate unit strongly



Baseball—or stripes?
mixed4a, Unit 6

Animal faces—or snouts?
mixed4a, Unit 240

Clouds—or fluffiness?
mixed4a, Unit 453

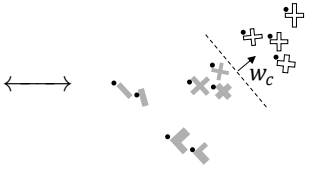
Buildings—or sky?
mixed4a, Unit 492

(Olah u. a. 2017, Fig. 5)

Concept Embedding Modelle

Ziel: Assoziation zw.

semantischen Konzepten, z.B., `istKopf`

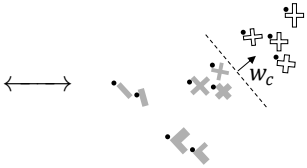


Konzeptaktivierungsvektoren w_c (CAVs) im Vektorraum der Zw.ausgabe

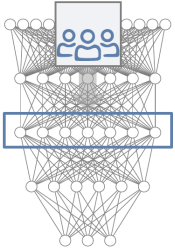
Concept Embedding Modelle

Ziel: Assoziation zw.

semantischen Konzepten, z.B., `istKopf`



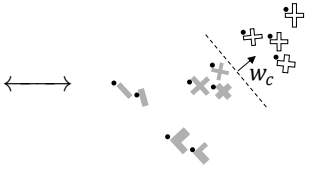
Konzeptaktivierungsvektoren w_c (CAVs) im Vektorraum der Zw.ausgabe



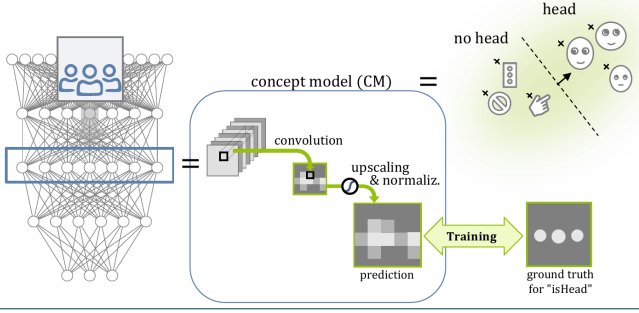
Concept Embedding Modelle

Ziel: Assoziation zw.

semantischen Konzepten, z.B., `istKopf`



Konzeptaktivierungsvektoren w_c (CAVs) im Vektorraum der Zw.ausgabe



Erklärbare Surrogate

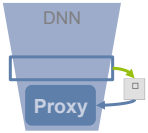
Idee: **Approximiere** DNN(-Teile) mittels transparenterer Modelle.

Erklärbare Surrogate

Idee: **Approximiere** DNN(-Teile) mittels transparenterer Modelle.

Beispiele:

Gesicht / kein Gesicht



- **Entscheidungsbäume or -regeln**, *CA-ILP* (Rabold u. a. 2020)

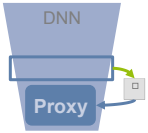
```
face(F) :- contains(F, A), isa(A, nose),
           contains(F, B), isa(B, mouth), top_of(A, B),
           contains(F, C), top_of(C, A)
```

Erklärbare Surrogate

Idee: **Approximiere** DNN(-Teile) mittels transparenterer Modelle.

Beispiele:

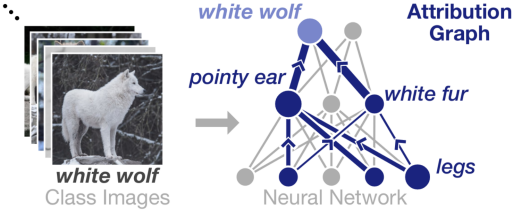
Gesicht / kein Gesicht



- **Entscheidungsbäume or -regeln**, *CA-ILP* (Rabold u. a. 2020)

```
face(F) :- contains(F, A), isa(A, nose),
           contains(F, B), isa(B, mouth), top_of(A, B),
           contains(F, C), top_of(C, A)
```

- Abhängigkeits- / Informationsfluss**graphen**, *z.B., SUMMIT* (Hohman u. a. 2020)



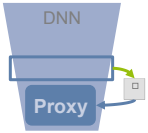
(Hohman u. a. 2020, Fig. 2)

Erklärbare Surrogate

Idee: **Approximiere** DNN(-Teile) mittels transparenterer Modelle.

Beispiele:

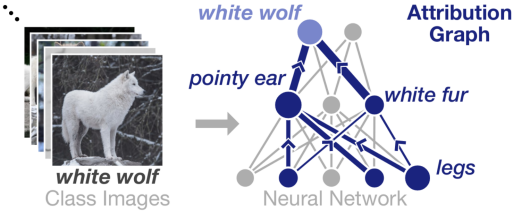
Gesicht / kein Gesicht



- **Entscheidungsbäume or -regeln**, *CA-ILP* (Rabold u. a. 2020)

```
face(F) :- contains(F, A), isa(A, nose),
           contains(F, B), isa(B, mouth), top_of(A, B),
           contains(F, C), top_of(C, A)
```

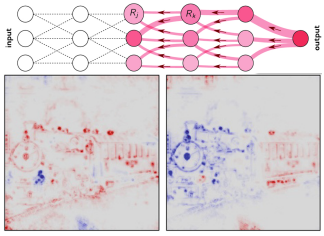
- Abhängigkeits- / Informationsfluss**graphen**, *z.B., SUMMIT* (Hohman u. a. 2020)
- Lokale **lineare** Approximationen



(Hohman u. a. 2020, Fig. 2)

Feature Importance Methoden

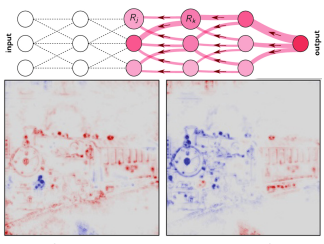
White-box:
backpropagation- oder
gradientbasiert



(Montavon u. a. 2019, Figs. 10.2-3)
z.B., LRP (Montavon u. a. 2019),
SmoothGrad (Smilkov u. a. 2017)

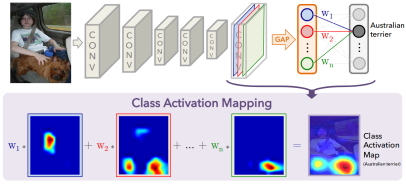
Feature Importance Methoden

White-box:
backpropagation- oder
gradientbasiert



(Montavon u. a. 2019, Figs. 10.2-3)
z.B., LRP (Montavon u. a. 2019),
SmoothGrad (Smilkov u. a. 2017)

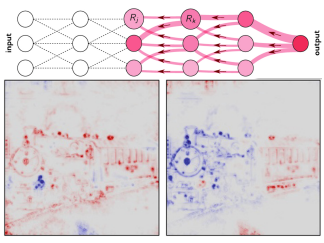
Gray-box:
Activation map basiert



(Zhou u. a. 2016, Fig. 2)
z.B.,Grad-CAM (Selvaraju u. a. 2017),
SIDU (Muddamsetty u. a. 2020)

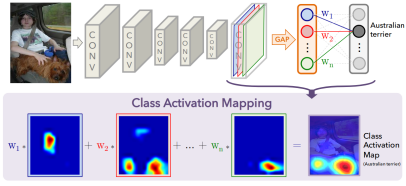
Feature Importance Methoden

White-box:
backpropagation- oder
gradientbasiert



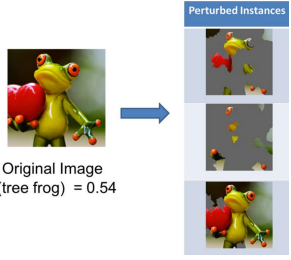
(Montavon u. a. 2019, Figs. 10.2-3)
z.B., LRP (Montavon u. a. 2019),
SmoothGrad (Smilkov u. a. 2017)

Gray-box:
Activation map basiert



(Zhou u. a. 2016, Fig. 2)
z.B.,Grad-CAM (Selvaraju u. a. 2017),
SIDU (Muddamsetty u. a. 2020)

Totale black-box:
Perturbationsbasiert



Original Image
P(tree frog) = 0.54

(Marco Tulio Ribeiro u. a. 2016, Fig. 4)
LIME (Marco Túlio Ribeiro u. a. 2016),
SHAP (Lundberg u. a. 2017),
RISE (Petsiuk u. a. 2018)

Gliederung

Wonach auswählen?

1. Anforderungen: Kenne Ziele & Bedarf (Taxonomie!)

Kreditscoring: z.B., lokale Feature Importance für Laien

Wonach auswählen?

1. Anforderungen: Kenne Ziele & Bedarf (Taxonomie!)

   Kreditscoring: z.B., lokale Feature Importance für Laien

2. Design:

   - Bevorzuge **transparente** Modelle wo möglich;

     balanciere Erklärbarkeit and Performanz aus

     Lineares Modell, Entscheidungsbaum?

   - Bevorzuge **kausale** and **kontrastive** Erklärungen

     Was ändern, um den Score zu bessern?

Wonach auswählen?

1. Anforderungen: Kenne Ziele & Bedarf (Taxonomie!)

2. Design:
   - Bevorzuge **transparente** Modelle wo möglich;

     balanciere Erklärbarkeit and Performanz aus
   - Bevorzuge **kausale** and **kontrastive** Erklärungen

3. V&V: **Nachmessen**!
   - Verständlichkeit → **nachfragen**!
   - Modelltreue, Abdeckung
   - Skalierbarkeit
   - …

Kreditscoring: z.B., lokale Feature Importance für Laien

Lineares Modell, Entscheidungsbaum?

Was ändern, um den Score zu bessern?

Können Testnutzer ihren Score bessern?

Wonach auswählen?

1. Anforderungen: Kenne Ziele & Bedarf (Taxonomie!)

   Kreditscoring: z.B., lokale Feature Importance für Laien

2. Design:

   - Bevorzuge **transparente** Modelle wo möglich;

     balanciere Erklärbarkeit and Performanz aus

     Lineares Modell, Entscheidungsbaum?

   - Bevorzuge **kausale** and **kontrastive** Erklärungen

     Was ändern, um den Score zu bessern?

3. V&V: **Nachmessen**!

   - Verständlichkeit → **nachfragen**!

     Können Testnutzer ihren Score bessern?

   - Modelltreue, Abdeckung
   - Skalierbarkeit
   - …

⇒ Viele Frameworks verfügbar! *für DNNs, z.B., captum.ai / Xplique*

Offene Herausforderungen

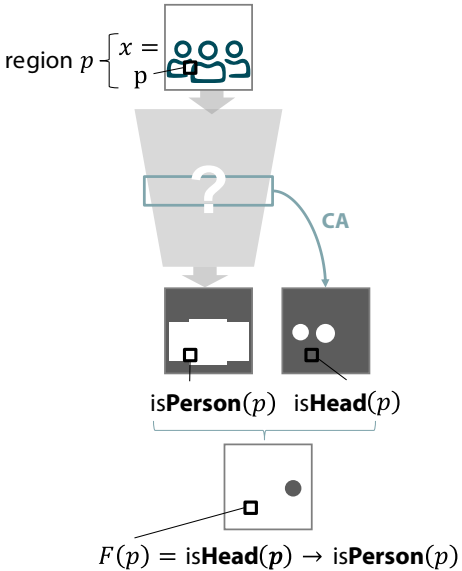
Eine (subjektive) Auswahl:

- **Evaluation**
- Reale Anwendungen: Performanz, Skalierbarkeit, …

Offene Herausforderungen

Eine (subjektive) Auswahl:

- **Evaluation**
- Reale Anwendungen: Performanz, Skalierbarkeit, …
- **Garantien** anstatt vagen Verständnisses
- „Es ist fehlerhaft; was nun?“
  → **Handlungsempfehlungen**
  → Verständnis von **strukturellen Mustern**



Conclusion

Takeaways:

- **Wir brauchen Erklärungen für** KI-Modelle, für Compliance und Qualität.
- **Es gibt viele Methoden** die zusammen viele Anwendungsfälle abdecken.

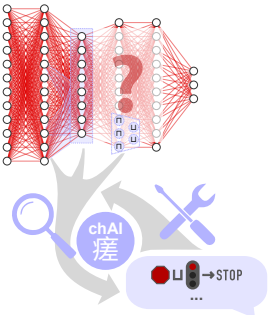
- Es gibt **keine eierlegende Wollmilchsau**! Beachte Trade-offs und Explainee-Bedürfnisse.

Conclusion

Takeaways:

- **Wir brauchen Erklärungen für** KI-Modelle,
  für Compliance und Qualität.
- **Es gibt viele Methoden**
  die zusammen viele Anwendungsfälle abdecken.

- Es gibt **keine eierlegende Wollmilchsau**!
  Beachte Trade-offs und Explainee-Bedürfnisse.



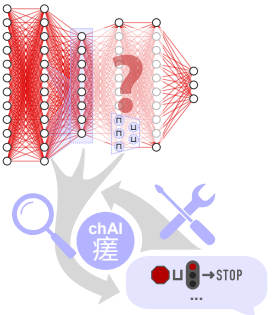
Ausblick:
Forschung läuft *(z.B., meine ☺)*,
Frameworks wachsen.

Conclusion

Takeaways:

- **Wir brauchen Erklärungen für** KI-Modelle, für Compliance und Qualität.
- **Es gibt viele Methoden** die zusammen viele Anwendungsfälle abdecken.

- Es gibt **keine eierlegende Wollmilchsau**! Beachte Trade-offs und Explainee-Bedürfnisse.



Ausblick:
Forschung läuft *(z.B., meine ☺)*,
Frameworks wachsen.

Fragen?

gesina.schwalbe@uni-luebeck.de
https://gesina.github.io
0000-0003-2690-2478

Literaturverzeichnis I

Chen, Chaofan u. a. (2019). „This Looks like That: Deep Learning for Interpretable Image Recognition". In: *Advances in Neural Information Processing Systems 32*. Bd. 32. Vancouver, BC, Canada, S. 8928–8939.

Hohman, Fred u. a. (Jan. 2020). „Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations". In: *IEEE Transactions on Visualization and Computer Graphics* 26.1, S. 1096–1106. ISSN: 1941-0506. DOI: 10.1109/TVCG.2019.2934659. (Besucht am 14. 02. 2021).

Lacave, Carmen und Francisco Dez (Mai 2001). „A Review of Explanation Methods for Bayesian Networks". In: *The Knowledge Engineering Review* 17. DOI: 10.1017/S026988890200019X.

Lundberg, Scott M und Su-In Lee (2017). „A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., S. 4765–4774. (Besucht am 30. 04. 2019).

Montavon, Grégoire u. a. (2019). „Layer-Wise Relevance Propagation: An Overview". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Lecture Notes in Computer Science. Springer International Publishing, S. 193–209. ISBN: 978-3-030-28954-6. DOI: 10.1007/978-3-030-28954-6_10. (Besucht am 20. 06. 2022).

Muddamsetty, Satya M., N. S. Jahromi Mohammad und Thomas B. Moeslund (Okt. 2020). „SIDU: Similarity Difference And Uniqueness Method for Explainable AI". In: *2020 IEEE International Conference on Image Processing (ICIP)*, S. 3269–3273. DOI: 10.1109/ICIP40778.2020.9190952.

Olah, Chris, Alexander Mordvintsev und Ludwig Schubert (Nov. 2017). „Feature Visualization". In: *Distill* 2.11, e7. ISSN: 2476-0757. DOI: 10.23915/distill.00007. (Besucht am 20. 05. 2019).

Petsiuk, Vitali, Abir Das und Kate Saenko (2018). „RISE: Randomized Input Sampling for Explanation of Black-Box Models". In: *Proc. British Machine Vision Conf.* BMVA Press, S. 151. (Besucht am 21. 01. 2019).

Pocevičiūtė, Milda, Gabriel Eilertsen und Claes Lundström (2020). „Survey of XAI in Digital Pathology". In: *Lecture Notes in Computer Science* 2020, S. 56–88. DOI: 10.1007/978-3-030-50402-1_4. arXiv: 2008.06353. (Besucht am 24. 02. 2021).

Rabold, Johannes, Gesina Schwalbe und Ute Schmid (2020). „Expressive Explanations of DNNs by Combining Concept Analysis with ILP". In: *KI 2020: Advances in Artificial Intelligence*. Lecture Notes in Computer Science. Springer International Publishing, S. 148–162. ISBN: 978-3-030-58285-2. DOI: 10.1007/978-3-030-58285-2_11.

Ribeiro, Marco Tulio, Sameer Singh und Carlos Guestrin (Aug. 2016). *Local Interpretable Model-Agnostic Explanations (LIME): An Introduction*. https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/. (Besucht am 20. 06. 2022).

Ribeiro, Marco Túlio, Sameer Singh und Carlos Guestrin (2016). „"Why Should I Trust You?: Explaining the Predictions of Any Classifier". In: *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*. KDD '16. ACM, S. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778.

Literaturverzeichnis II

Rudin, Cynthia (Mai 2019). „Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: *Nature Machine Intelligence* 1.5, S. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x. (Besucht am 25. 02. 2021).

Schwalbe, Gesina und Bettina Finzel (Jan. 2023). „A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts". In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00867-8. (Besucht am 09. 01. 2023).

Selvaraju, Ramprasaath R. u. a. (Okt. 2017). „Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *Proc. 2017 IEEE Int. Conf. Computer Vision*. IEEE, S. 618–626. ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.74. (Besucht am 14. 11. 2020).

Smilkov, Daniel u. a. (2017). „SmoothGrad: Removing Noise by Adding Noise". In: *CoRR* abs/1706.03825.

Zhou, Bolei u. a. (2016). „Learning Deep Features for Discriminative Localization". In: *Proc. 2016 IEEE Conf. Comput. Vision and Pattern Recognition*. IEEE Computer Society, S. 2921–2929. DOI: 10.1109/CVPR.2016.319.