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

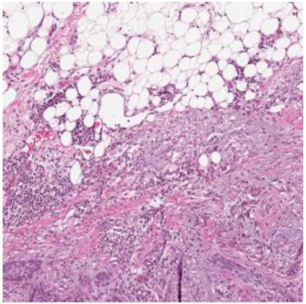
Outline

Outline

Why do we need XAI?

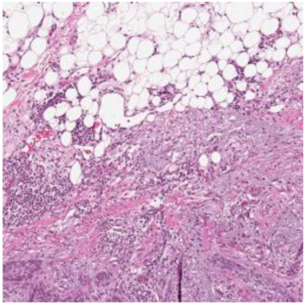
Why do we need XAI?

Why do we need XAI?


(Pocevičiūtė et al. 2020, Fig. 6)

Why do we need XAI?


©Patrick Fallon/Imago


(Pocevičiūtė et al. 2020, Fig. 6)

Where do we need XAI?

Use-cases:

- End users:
  - (appropriate!) **Trust**, informed consent
  - Onboarding
  - **Recourse**
- For developers and expert users:
  - **Debugging**
  - Knowledge retrieval
- For assessors:
  - Compliance *with law and standards*
  - **Assessment**, *e.g., wrt. safety, fairness*

Where do we need XAI?

Use-cases:

- End users:
  - (appropriate!) **Trust**, informed consent
  - Onboarding
  - **Recourse**
- For developers and expert users:
  - **Debugging**
  - Knowledge retrieval
- For assessors:
  - Compliance *with law and standards*
  - **Assessment**, *e.g., wrt. safety, fairness*

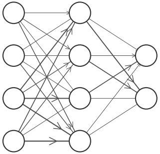
Application Fields:

Wherever automated decisions influence human well-being!

- **Ranking** systems *(credits, applications, …)*
- **Medical** assistant systems
- Automated **driving**
- **Military** decision systems
- **HMI** in Production
- …

Looking inside is hard: DNNs as example.

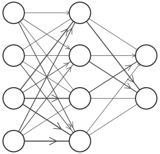
$$f_\theta : \mathbb{R}^n \longrightarrow \mathbb{R}^m$$



Feedforward DNN

Looking inside is hard: DNNs as example.

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



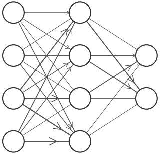
Feedforward DNN

Main challenges:

- **HUGE:** *YOLOv9, DETR:* ***>50 M*** *param.s*
  Llama 3.2: ***90 B*** *param.s*

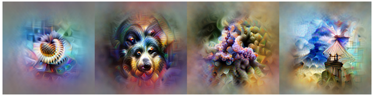
Looking inside is hard: DNNs as example.

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



Feedforward DNN

(Olah et al. 2017)

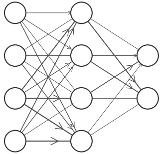


Main challenges:

- **HUGE**: *YOLOv9, DETR:* ***>50 M*** *param.s*
  Llama 3.2: ***90 B*** *param.s*
- **Automatically** learned representations:

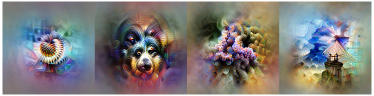
Looking inside is hard: DNNs as example.

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



Feedforward DNN

(Olah et al. 2017)



(Marco Túlio Ribeiro et al. 2016, Fig. 11)

Husky **image**
misclassified as *Wolf*

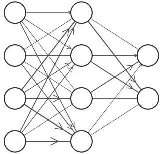
features most influential
for the decision

Main challenges:

- **HUGE**: *YOLOv9, DETR: **>50 M** param.s*
  *Llama 3.2: **90 B** param.s*
- **Automatically** learned representations:
  - **distributed** $\Rightarrow$ hard to read

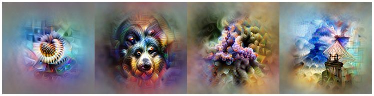
Looking inside is hard: DNNs as example.

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



Feedforward DNN

(Olah et al. 2017)



(Marco Túlio Ribeiro et al. 2016, Fig. 11)

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

features most influential
for the decision

Main challenges:

- **HUGE**: *YOLOv9, DETR: **>50 M** param.s*
  *Llama 3.2: **90 B** param.s*
- **Automatically** learned representations:
  - **distributed** $\Rightarrow$ hard to read
  - only learns **correlations** $\Rightarrow$ hard to anticipate

$\Rightarrow$ black-box.

Looking isn't enough: We need Understanding.

Definition (Understanding)

successful update of mental model; can be
mechanistical = how it works, or
functional = what is its purpose.

Looking isn't enough: We need Understanding.

Definition (Understanding)

successful update of mental model; can be *mechanistical* = how it works, or *functional* = what is its purpose.

Definition (Levels of transparency)

Levels of *transparency* of a model:

- simulatable = understandable as a whole
- decomposable into simulatable parts
- algorithmically transparent = mathematical understanding

EU AI Act

Preamble (72)

*[…] **transparency should be required for high-risk AI systems** […]. High-risk AI systems should be designed in a manner to enable deployers to **understand** how the AI system works, […]*

Article 13

*1. **High-risk AI systems** shall be designed and developed in such a way as to ensure that their operation is **sufficiently transparent** to enable deployers **to interpret** a system's output and **use it appropriately**. […]*

What we want: Explainable AI

Definition (Explainable decision system)

There exists a mechanism providing an explanation (= *explanator*) to a human (= *explainee*) allowing them to *understand* one of (= *explanandum*)

- the **model** resp. parts thereof,
- evidence for a model output, or
- the context of the system's reasoning.

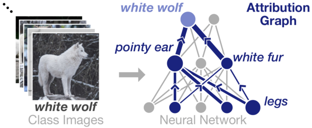
What we want: Explainable AI

Definition (Explainable decision system)

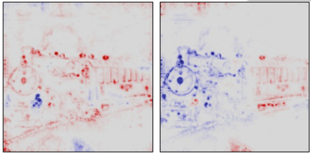
There exists a mechanism providing an explanation (= *explanator*) to a human (= *explainee*) allowing them to *understand* one of (= *explanandum*)

- the **model** resp. parts thereof,
- evidence for a model output, or
- the context of the system's reasoning.

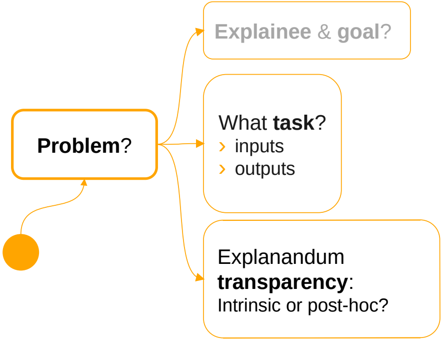
⇒ **How** does it work? (*global*)



⇒ **Why** this decision? (*local*)
Why not another? (*contrastive*))

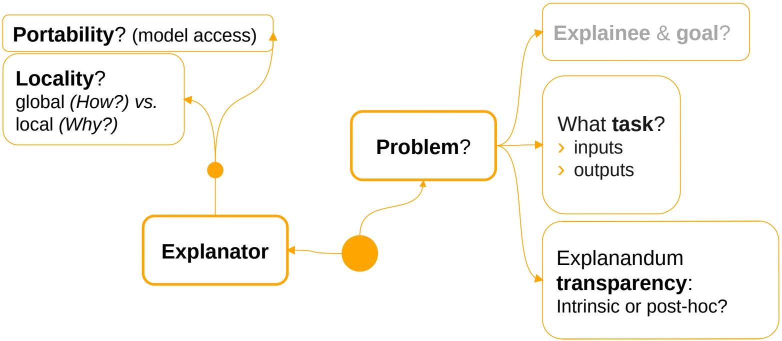


What to consider? A taxonomy.



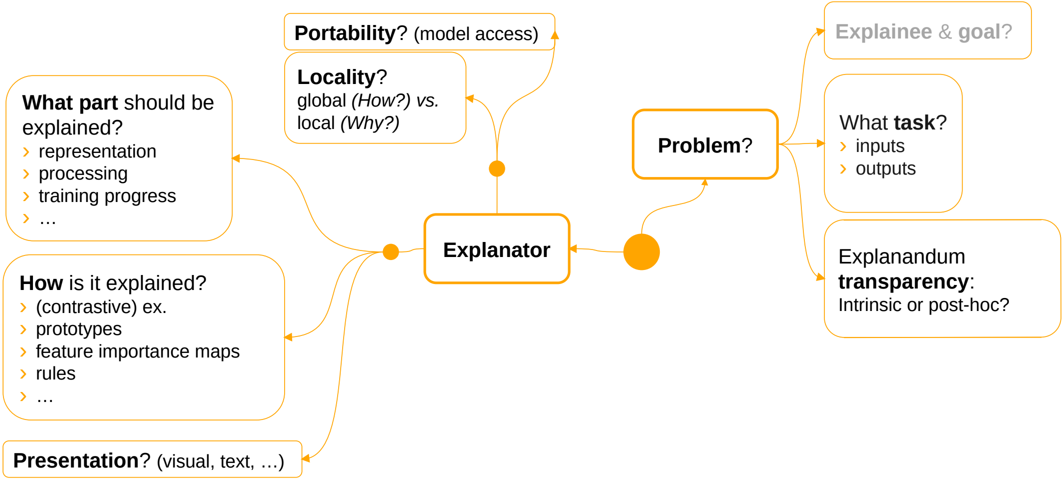
Gesina Schwalbe et al. (Jan. 2023). "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts". In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00867-8. (Visited on 01/09/2023)

What to consider? A taxonomy.



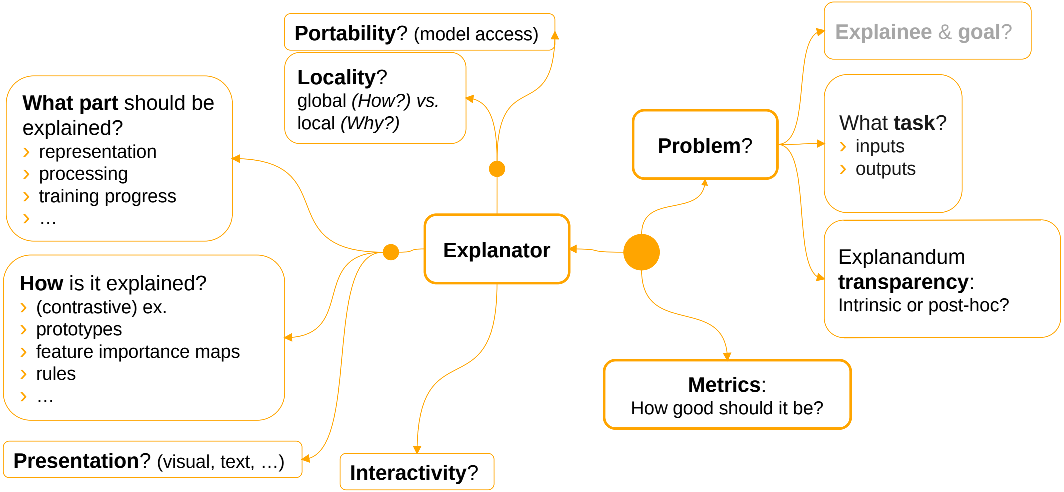
Gesina Schwalbe et al. (Jan. 2023). "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts". In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00867-8. (Visited on 01/09/2023)

What to consider? A taxonomy.



Gesina Schwalbe et al. (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. (Visited on 01/09/2023)

What to consider? A taxonomy.



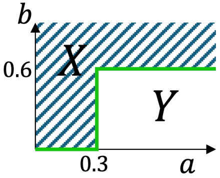
Gesina Schwalbe et al. (Jan. 2023). "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts". In: *Data Mining and Knowledge Discovery*. ISSN: 1573-756X. DOI: 10.1007/s10618-022-00867-8. (Visited on 01/09/2023)

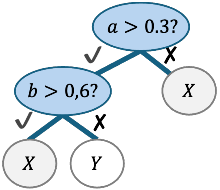
Outline

Inherently Interpretable Models

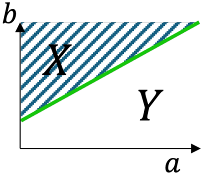
If possible, make it interpretable right away. (Rudin 2019)

Decision **Trees**



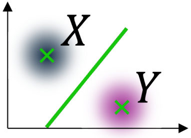


Linear Models



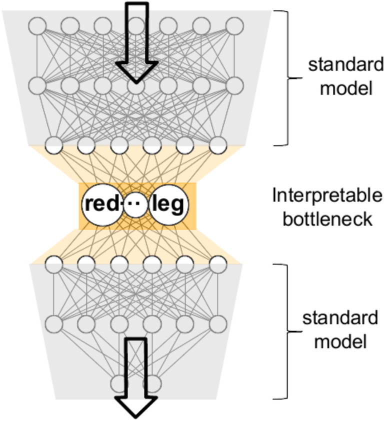
$$f(x) = \alpha a + \beta b$$

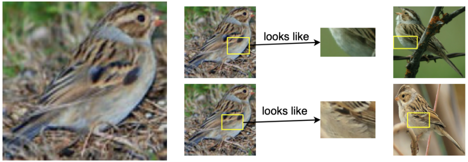
Clusters / **Prototypes**



Modularize: Blended Models

Modularize: Blended Models





(Chen et al. 2019)

Towards understanding representations: Feature visualization

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

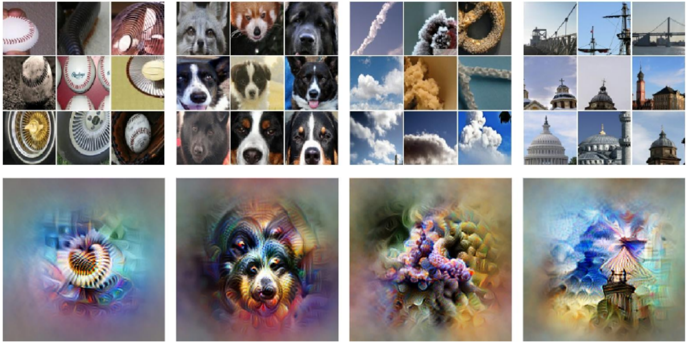
(Olah et al. 2017, Fig. 5)

Towards understanding representations: Feature visualization

Question: What does the output of a network unit/part (e.g., 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*

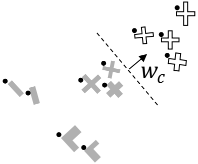
(Olah et al. 2017, Fig. 5)

Concept Embedding Models

Goal: association

semantic concepts,
e.g., `isHead`

⟷

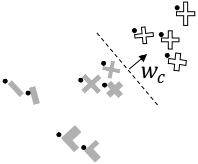


concept activation vectors w_c
(CAVs) in DNN latent space

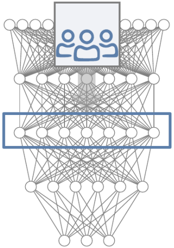
Concept Embedding Models

Goal: association

semantic concepts,
e.g., `isHead`



concept activation vectors w_c
(CAVs) in DNN latent space

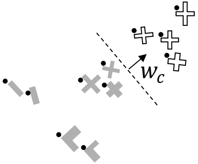


Concept Embedding Models

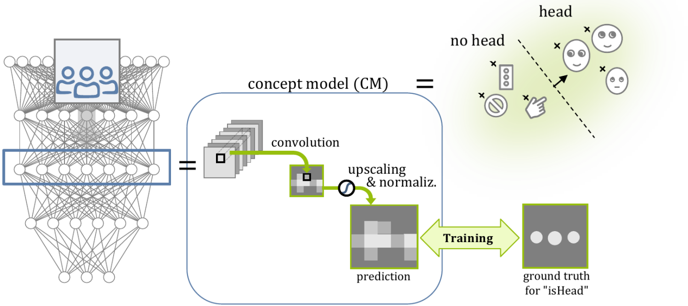
Goal: association

semantic concepts,
e.g., `isHead`

$\longleftrightarrow$



concept activation vectors w_c
(CAVs) in DNN latent space



Explainable Surrogates

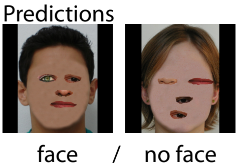
Idea: **Approximate** DNN (parts) by an interpretable model.

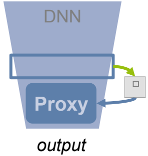
Explainable Surrogates

Idea: **Approximate** DNN (parts) by an interpretable model.

Examples:

- **Decision tree or rules,**
  CA-ILP (Rabold et al. 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)
```

Explainable Surrogates

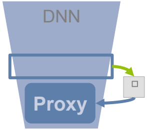
Idea: **Approximate** DNN (parts) by an interpretable model.

Examples:

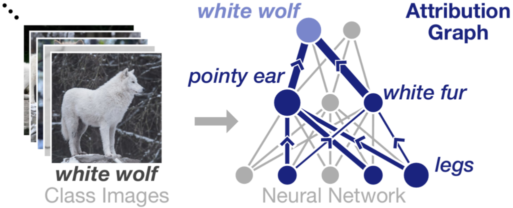
- **Decision tree or rules,**
  CA-ILP (Rabold et al. 2020)
- Dependency / flow **graphs,**
  e.g., SUMMIT (Hohman et al. 2020)

Predictions

face / no face




```
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)
```



(Hohman et al. 2020, Fig. 2)

Explainable Surrogates

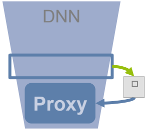
Idea: **Approximate** DNN (parts) by an interpretable model.

Examples:

- **Decision tree or rules**,
  CA-ILP (Rabold et al. 2020)
- Dependency / flow **graphs**,
  e.g., SUMMIT (Hohman et al. 2020)
- Local **linear** approximations

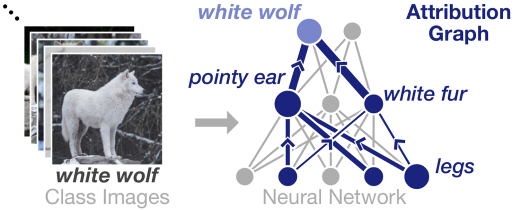
Predictions

face / no face



output

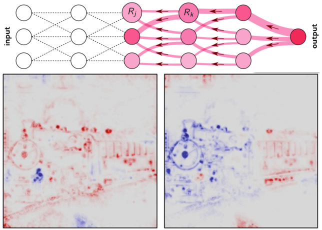
```
face(F) :- contains(F, A), isa(A, nose),
           contains(F, B), isa(B, mouth), top_of(A, B),
           contains(F, C), top_of(C, A)
```



(Hohman et al. 2020, Fig. 2)

Feature Importance Methods

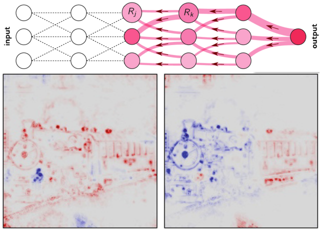
White-box:
Backpropagation or
Gradient based



(Montavon et al. 2019, Figs. 10.2-3)
e.g., LRP (Montavon et al. 2019),
SmoothGrad (Smilkov et al. 2017)

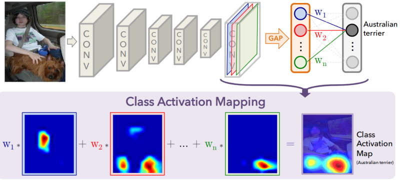
Feature Importance Methods

White-box:
Backpropagation or **Gradient** based



(Montavon et al. 2019, Figs. 10.2-3)
e.g., LRP (Montavon et al. 2019),
SmoothGrad (Smilkov et al. 2017)

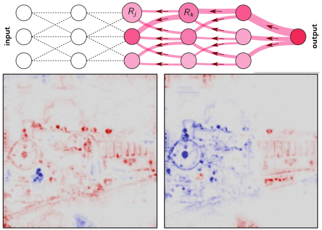
Gray-box:
Activation map based



(Zhou et al. 2016, Fig. 2)
e.g.,Grad-CAM (Selvaraju et al. 2017),
SIDU (Muddamsetty et al. 2020)

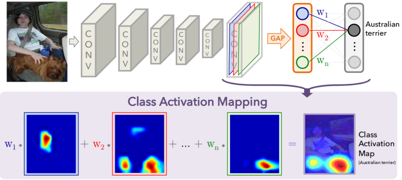
Feature Importance Methods

White-box:
Backpropagation or **Gradient** based



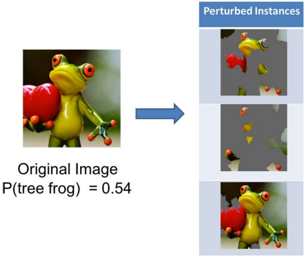
(Montavon et al. 2019, Figs. 10.2-3)
e.g., LRP (Montavon et al. 2019),
SmoothGrad (Smilkov et al. 2017)

Gray-box:
Activation map based



(Zhou et al. 2016, Fig. 2)
e.g.,Grad-CAM (Selvaraju et al. 2017),
SIDU (Muddamsetty et al. 2020)

Total black-box:
Perturbation based



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

Outline

So, how to choose?

1. Requirements: Know your goals & needs (taxonomy!)

*Loan decisions: E.g.,
local feature importance
for lay users' recourse*

So, how to choose?

1. Requirements: Know your goals & needs (taxonomy!)

   Loan decisions: E.g., local feature importance for lay users' recourse

2. Design:
   - Prefer **transparent** models where possible; balance explainability and accuracy

     linear model, decision tree?

   - Prefer **causal** and **contrastive** explanations

     What to change to get the loan?

So, how to choose?

1. Requirements: Know your goals & needs (taxonomy!)

Loan decisions: E.g., local feature importance for lay users' recourse

2. Design:
   - Prefer **transparent** models where possible; balance explainability and accuracy

     linear model, decision tree?

   - Prefer **causal** and **contrastive** explanations

     What to change to get the loan?

3. V&V: **Measure** it!
   - Understandability $\rightarrow$ **ask!**

     Could test users improve their chances?

   - Faithfulness, coverage
   - Scalability
   - …

So, how to choose?

1. Requirements: Know your goals & needs (taxonomy!)

   Loan decisions: E.g., local feature importance for lay users' recourse

2. Design:
   - Prefer **transparent** models where possible; balance explainability and accuracy

     linear model, decision tree?

   - Prefer **causal** and **contrastive** explanations

     What to change to get the loan?

3. V&V: **Measure** it!
   - Understandability $\rightarrow$ **ask!**

     Could test users improve their chances?

   - Faithfulness, coverage
   - Scalability
   - …

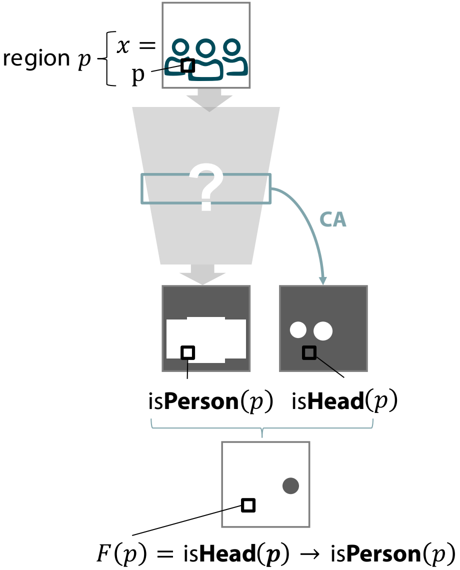
$\Rightarrow$ Many frameworks available! *for DNNs, e.g., captum.ai / Xplique*

Further Challenges

A (subjective) selection:

- **Evaluation**
- Real-world applications: accuracy, scalability, …

Further Challenges



A (subjective) selection:

- **Evaluation**
- Real-world applications: accuracy, scalability, …
- **Guarantees** instead of vague understanding
- "It's broken; what now?"
  $\rightarrow$ **Actionability**
  $\rightarrow$ Better understanding of **structural patterns**

Conclusion

Takeaways:

- **We need explanations** for AI models,
  for legal and quality reasons.
- **There are many methods available**
  serving many use-cases.

- There is **no golden bullet**!
  Mind trade-offs and explainees' needs.

Conclusion

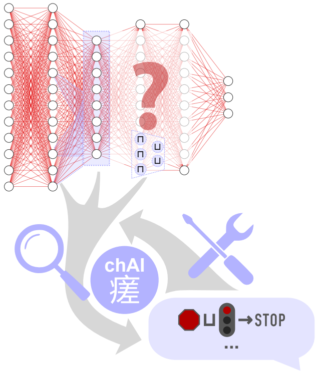
Takeaways:

- **We need explanations** for AI models,
  for legal and quality reasons.
- **There are many methods available**
  serving many use-cases.

- There is **no golden bullet!**
  Mind trade-offs and explainees' needs.

Outlook:

Active research is ongoing *(e.g., mine ☺)*,
frameworks are growing.



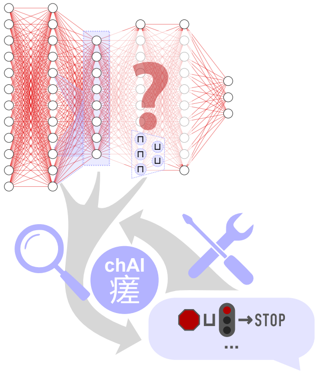
Conclusion

Takeaways:

- **We need explanations** for AI models,
  for legal and quality reasons.
- **There are many methods available**
  serving many use-cases.

- There is **no golden bullet**!
  Mind trade-offs and explainees' needs.

Outlook:
Active research is ongoing *(e.g., mine ☺)*,
frameworks are growing.



Questions?

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

References I

Chen, Chaofan et al. (2019). "This Looks like That: Deep Learning for Interpretable Image Recognition". In: *Advances in Neural Information Processing Systems 32*. Vol. 32. Vancouver, BC, Canada, pp. 8928–8939.

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

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pp. 4765–4774. (Visited on 04/30/2019).

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

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

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

Petsiuk, Vitali, Abir Das, and Kate Saenko (2018). "RISE: Randomized Input Sampling for Explanation of Black-Box Models". In: *Proc. British Machine Vision Conf.* BMVA Press, p. 151. (Visited on 01/21/2019).

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

Rabold, Johannes, Gesina Schwalbe, and 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, pp. 148–162. ISBN: 978-3-030-58285-2. DOI: 10.1007/978-3-030-58285-2_11.

Ribeiro, Marco Tulio, Sameer Singh, and 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/. (Visited on 06/20/2022).

Ribeiro, Marco Túlio, Sameer Singh, and 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, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778.

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

References II

Schwalbe, Gesina and 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. (Visited on 01/09/2023).

Selvaraju, Ramprasaath R. et al. (Oct. 2017). “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *Proc. 2017 IEEE Int. Conf. Computer Vision*. IEEE, pp. 618–626. ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.74. (Visited on 11/14/2020).

Smilkov, Daniel et al. (2017). “SmoothGrad: Removing Noise by Adding Noise”. In: *CoRR* abs/1706.03825.

Zhou, Bolei et al. (2016). “Learning Deep Features for Discriminative Localization”. In: *Proc. 2016 IEEE Conf. Comput. Vision and Pattern Recognition*. IEEE Computer Society, pp. 2921–2929. DOI: 10.1109/CVPR.2016.319.