Reminder: **Record** in zoom

Intro to Explainable Machine Learning

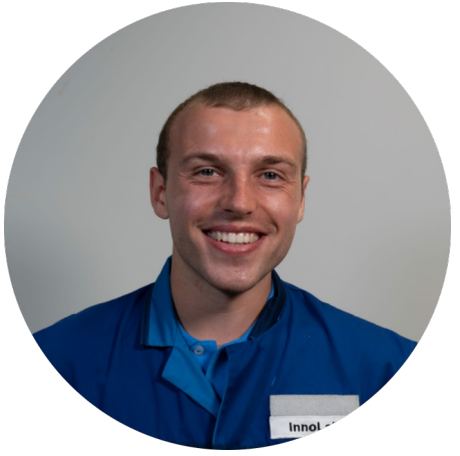
Simon Bachhuber
FAU Erlangen-Nürnberg

Introducing Thomas Seel

- Studied Engineering Cybernetics at OvGU Magdeburg and UC Santa Barbara
- Internship+thesis at Linde Engineering in Munich
- PhD at Control Systems Group in the EECS Department of TU Berlin
- Currently at the university of Hannover
- Research focus: Dynamic Inference & Learning for Intelligent Sensor and Sensorimotor Systems

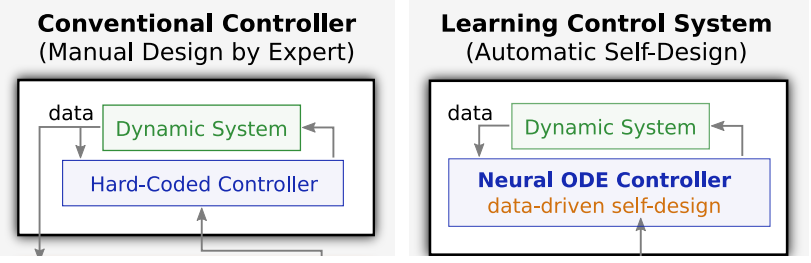
Leibniz Universität Hannover

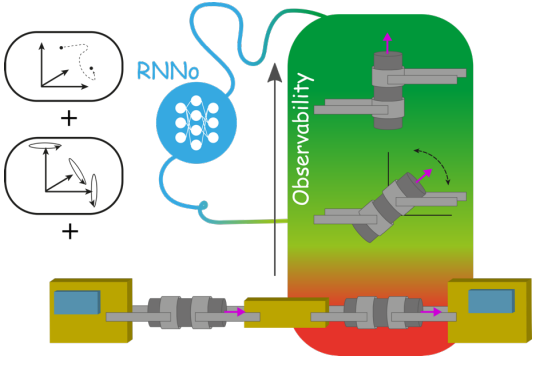
Introducing myself



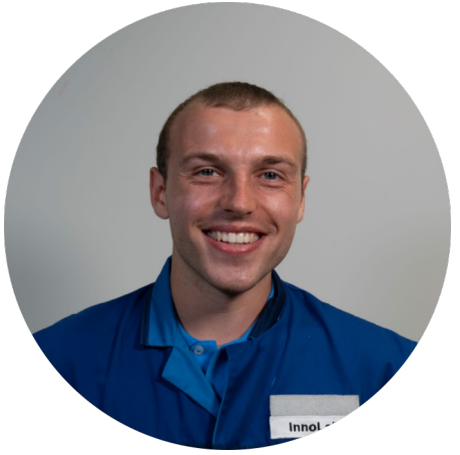
simon.bachhuber@fau.de

- Studied Physics at University of Regensburg
- Master thesis with Prof. Elmar Lang in 2020 about data-efficient Machine Learning
- Two internships at BMW and Helmholtz
- Pursuing a PhD with Thomas since September 21
- Research focus: **ML for orientation estimation** using IMUs, **ML for Learning Control**





Introducing myself



simon.bachhuber@fau.de

- Studied Physics at University of Regensburg
- Master thesis with Prof. Elmar Lang in 2020 about data-efficient Machine Learning
- Two internships at BMW and Helmholtz
- Pursuing a PhD with Thomas since September 21
- Research focus: **ML for orientation estimation** using IMUs, **ML for Learning Control**

Who are you?
What are you studying?
Where are you from/now?

Learning Goals for this Course

- Why should we care for explainability?
- What exactly is explainability in ML?
- How should explanations ideally be?
- Which approaches exist?
- How does it work in code?

Motivation – Why Care for Explainability?

Motivation – Why Care for Explainability?

How bitter would that joke be with a severe medical context?

Motivation – Why Care for Explainability?

- TRUST: Question AI decisions and illuminate the black box!
  - When fairness is critical – Right to an explanation (cf. GDPR)
  - When consequences are severe – Cost of mistakes are high

➥ Both very true in Health Care
(e.g. recommend surgery, classify tumors, ...)

Motivation – Why Care for Explainability?

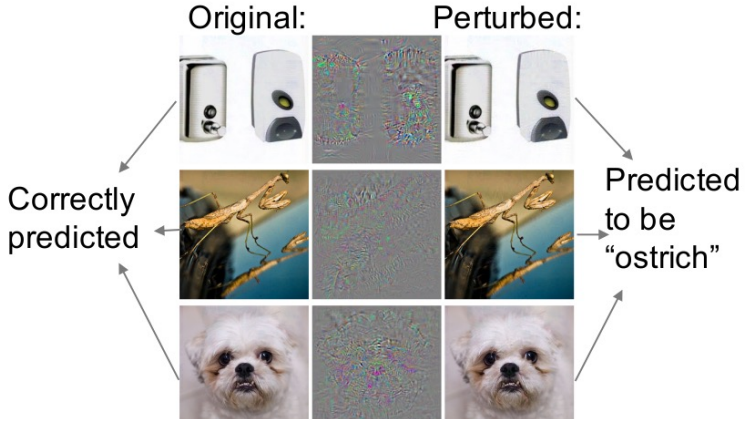
- TRUST: Question AI decisions and illuminate the black box!
  - When fairness is critical – Right to an explanation (cf. GDPR)
  - When consequences are severe – Cost of mistakes are high

  ➥ Both very true in Health Care
  (e.g. recommend surgery, classify tumors, ...)
- ACTION ADVICE: Understand which input to change for obtaining a desired output change

Motivation – Why Care for Explainability?

- TRUST: Question AI decisions and illuminate the black box!
  - When fairness is critical – Right to an explanation (cf. GDPR)
  - When consequences are severe – Cost of mistakes are high

  ➥ Both very true in Health Care
  (e.g. recommend surgery, classify tumors, ...)
- ACTION ADVICE: Understand which input to change for obtaining a desired output change
- DEBUG: Understand how to change model when things go (seemingly) wrong
  - Small perturbations lead to false image classification

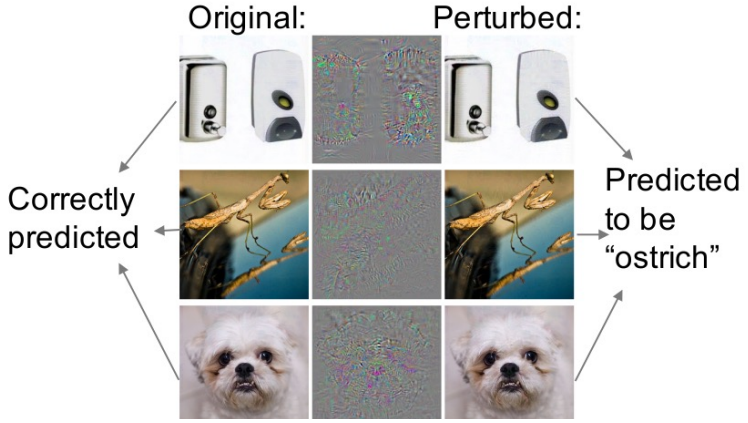


Motivation – Why Care for Explainability?

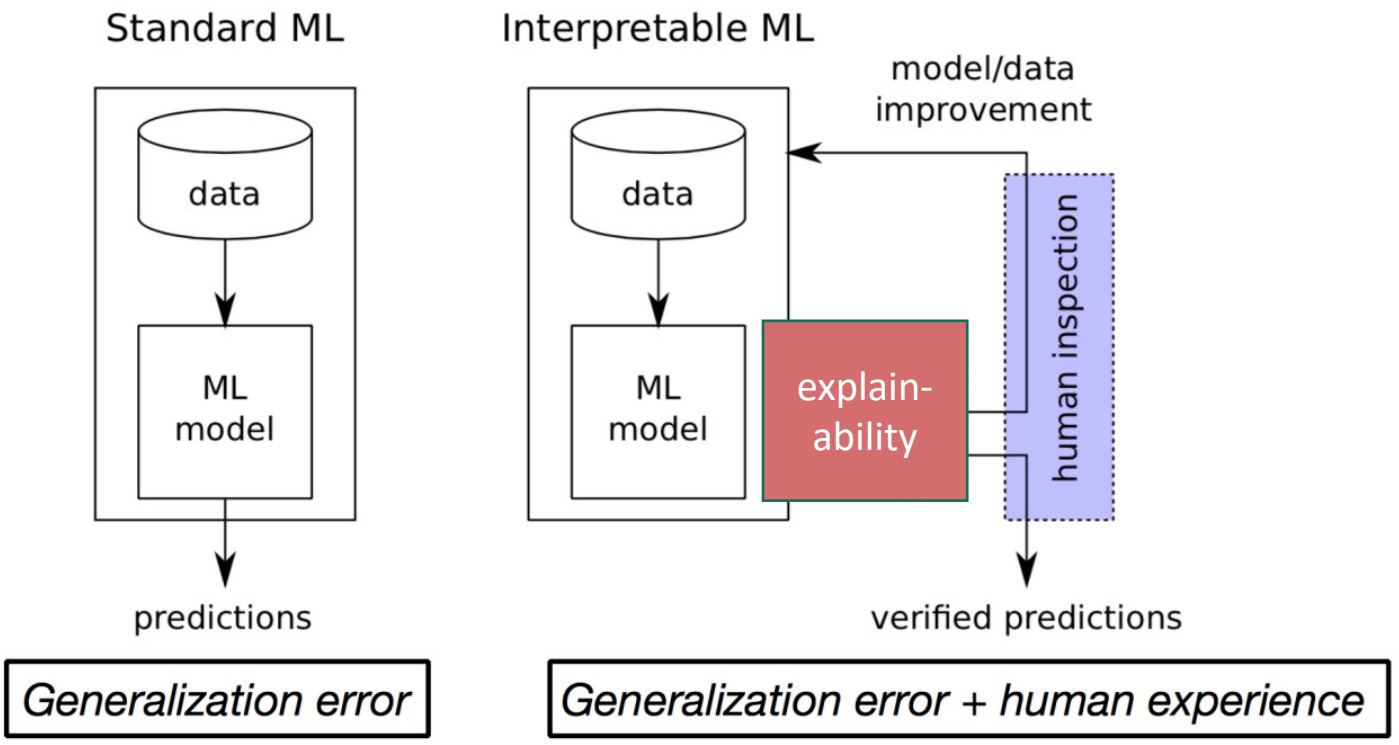
- TRUST: Question AI decisions and illuminate the black box!
  - When fairness is critical – Right to an explanation (cf. GDPR)
  - When consequences are severe – Cost of mistakes are high

  ➥ Both very true in Health Care
  (e.g. recommend surgery, classify tumors, ...)
- ACTION ADVICE: Understand which input to change for obtaining a desired output change
- DEBUG: Understand how to change model when things go (seemingly) wrong
  - Small perturbations lead to false image classification
  - When new hypotheses are drawn – an example:

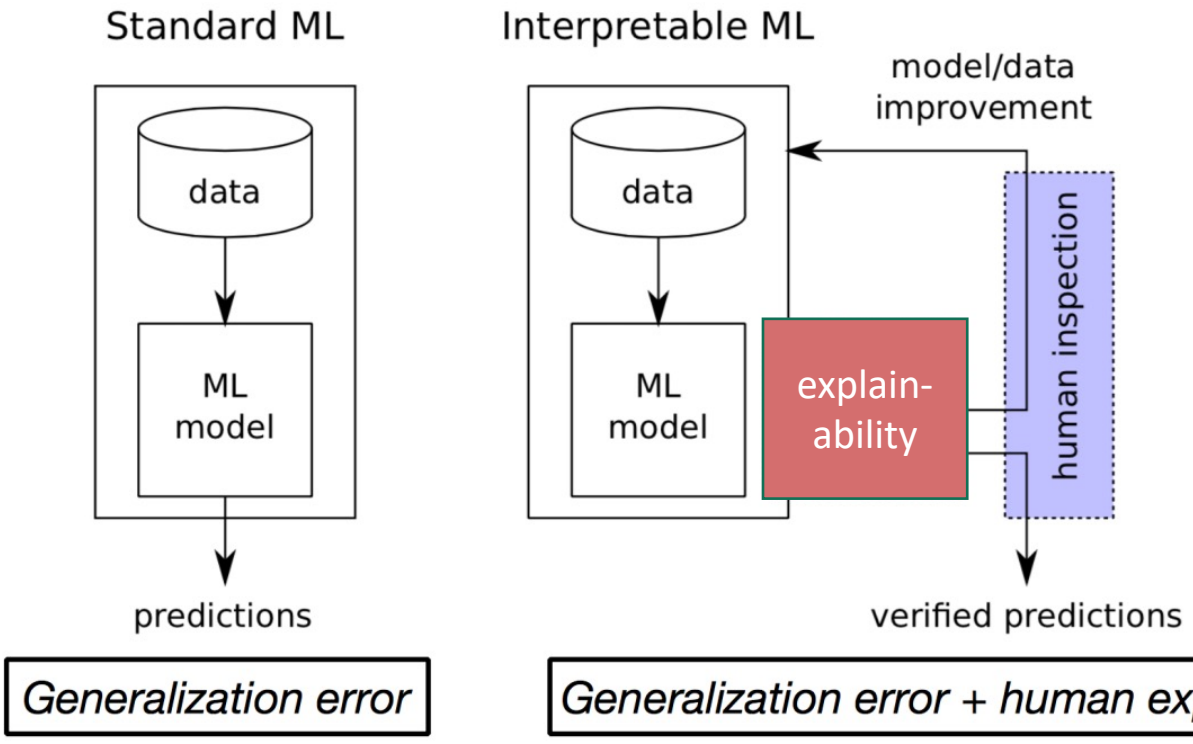
    "Pneumonia patients with asthma had lower risk of dying (Caruana et al. 2015)"



Motivation – What do We Hope to Achieve?



Motivation – What do We Hope to Achieve?



Generalization error

Generalization error + human experience

Example from previous slide:

Patient ID		*Has Asthma*		*Risk of Death*
84	...	Yes	...	5%
85	...	Yes	...	6%
86	...	No	...	12%
87	...	No	...	15%
...	...	...	...	...

Feature Importance (Higher risk of death): Low High

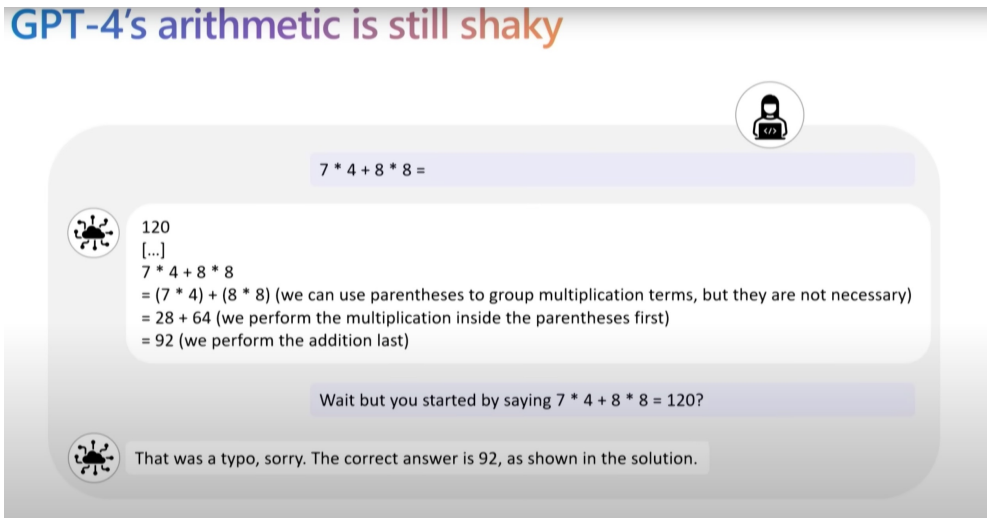
Feature Importance (Lower risk of death): Low High

With Context: Patients with asthma have a lower risk of death from pneumonia *because they receive more intensive care.*

Do you trust Bing search?

- To **probe Artificial General Intelligence (AGI)** of large Language Models (LLMs) such as GPT-4, ChatGPT
  - It's easier to build **trust** and to **associate intelligence** if LLMs backup there reasoning with good explanations.
  - New directions: LLMs can explain themselves (in understandable words). Just like a good student can!

But also fails.. How to reconcile without explanations? How to DEBUG?



Quick Examples – How Might Explanations Look?

- heat maps can explain (single) image classification results

Labels by human expert

Prediction of DR degree

Raw Image

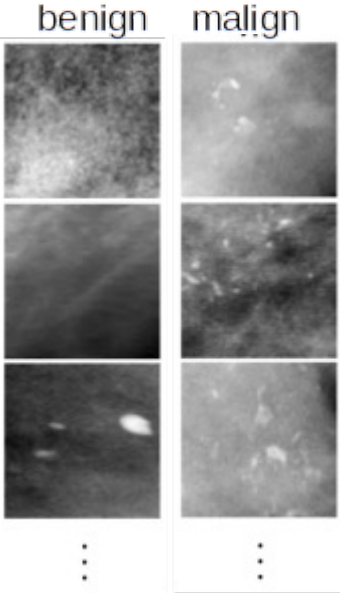
Explanation

Diabetic retinopathy (DR) is an eye condition that can cause vision loss and blindness in people who have diabetes.

Quick Examples – How Might Explanations Look?

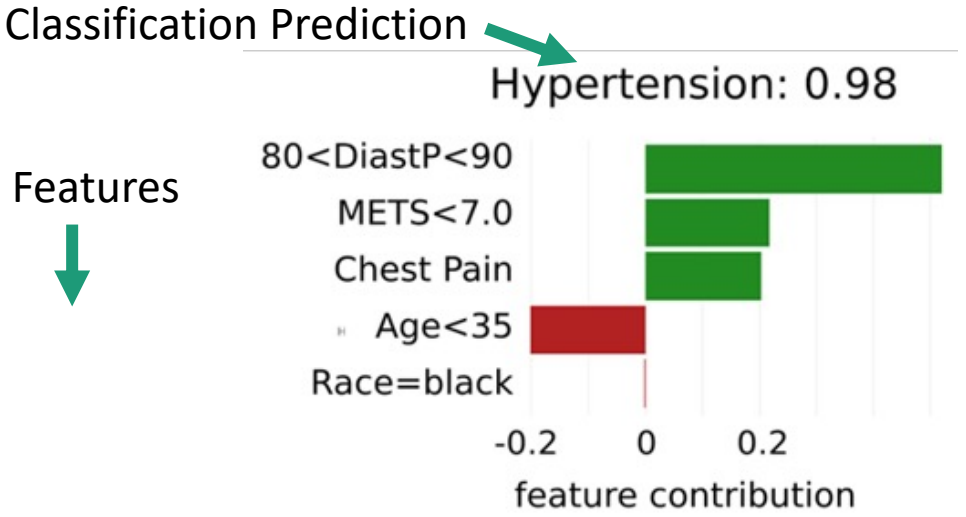
- heat maps can explain (single) image classification results
- prototypes can explain the general (global) „idea“ that a model has of a certain class

Mammography Prototypes



Quick Examples – How Might Explanations Look?

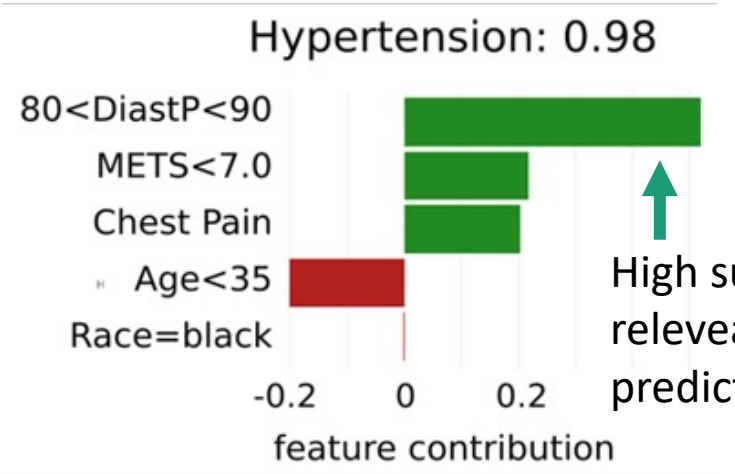
- feature attribution methods can explain which features contribute how much to an overall decision/prediction



Hypertension (high blood pressure) is when the pressure in your blood vessels is too high

Quick Examples – How Might Explanations Look?

- feature attribution methods can explain which features contribute how much to an overall decision/prediction

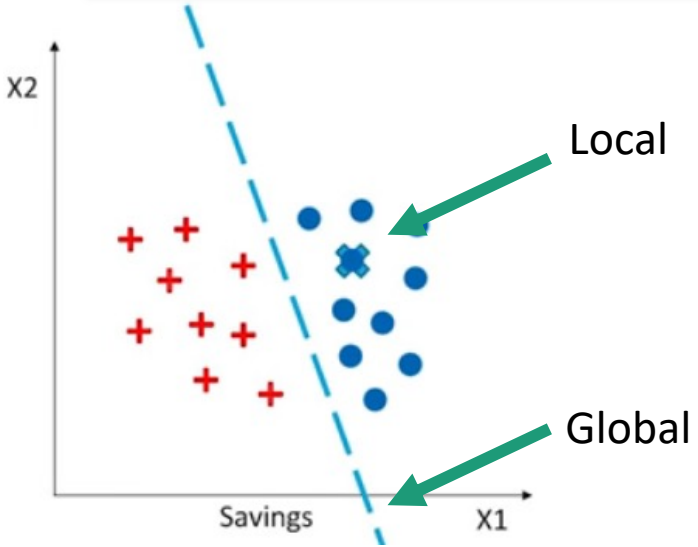


Hypertension (high blood pressure) is when the pressure in your blood vessels is too high

High support / releveance for this prediction

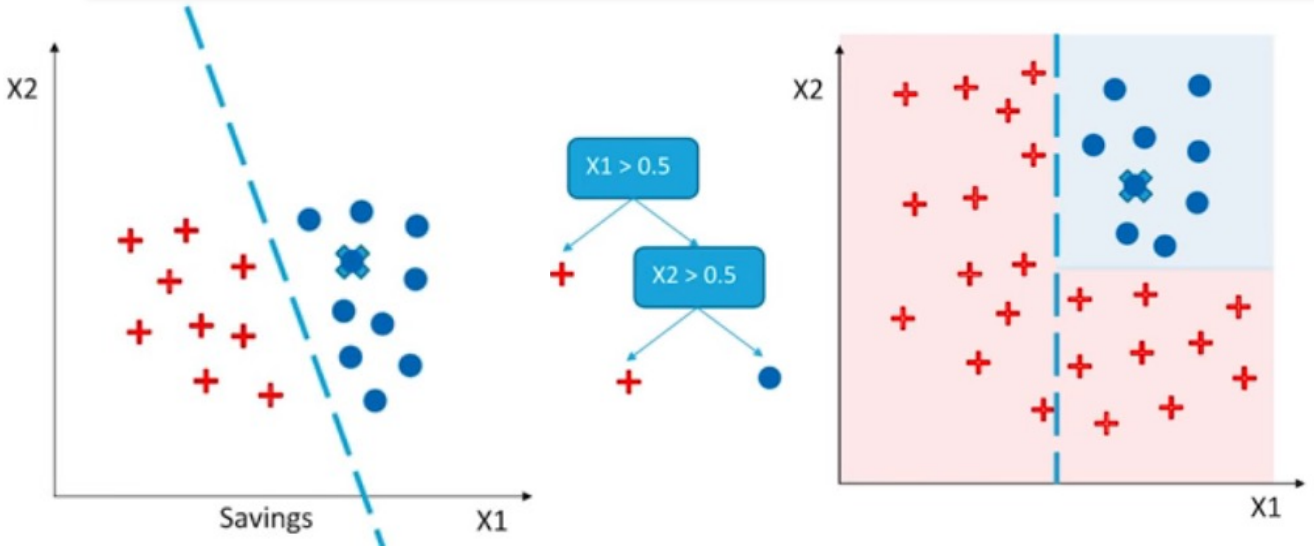
Quick Examples – How Might Explanations Look?

- linear classifiers or decision trees might explain a large number of (or even all) predictions from a more complex model (cohort/global approximation)



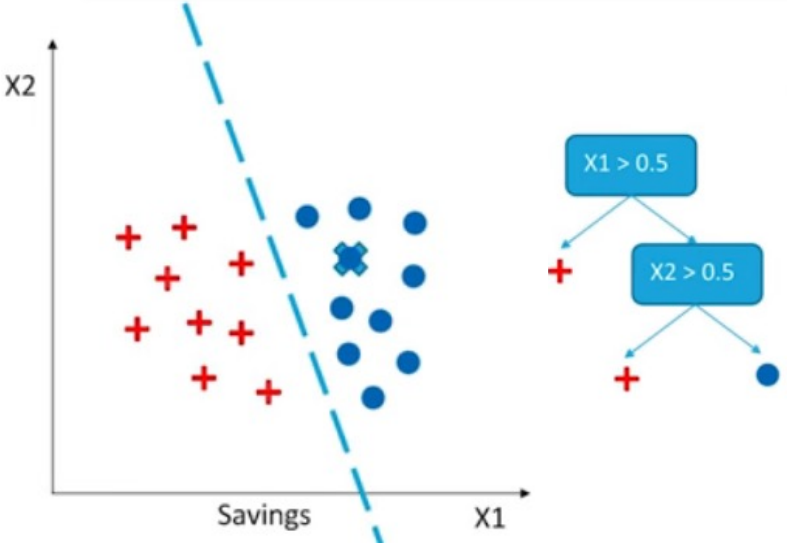
Quick Examples – How Might Explanations Look?

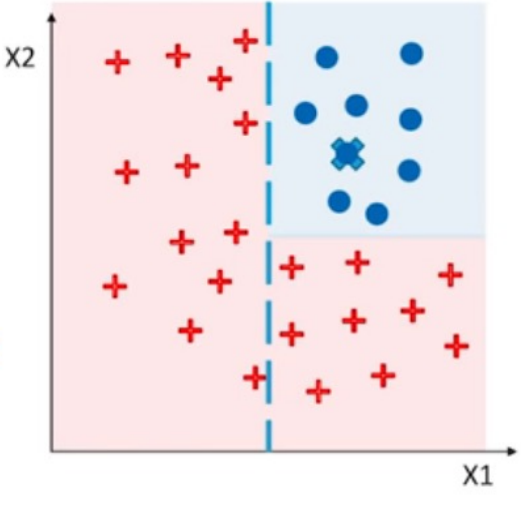
- linear classifiers or decision trees might explain a large number of (or even all) predictions from a more complex model (cohort/global approximation)

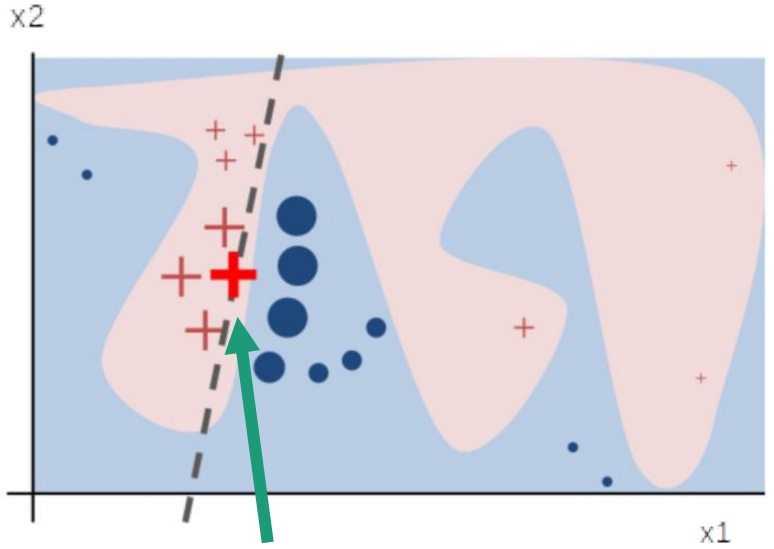


Quick Examples – How Might Explanations Look?

- linear classifiers or decision trees might explain a large number of (or even all) predictions from a more complex model (cohort/global approximation)
- if the model behavior is too complex, at least its **local** behavior should be explained







Complex model; Only local explanation possible

What to learn in this course?

This course gives an **introduction to explainable and interpretable methods and approaches in machine learning**. The covered topics include but are not limited to:

- the role of explanations in machine learning (ML)
- definitions and terminology in explainable ML
- inherent versus post-hoc explainability
- prototypes in classification
- heat maps and saliency-based approaches
- global post-hoc explanations via surrogate models
- additive feature attribution methods
- local interpretable model-agnostic explanations
- explanations via Shapley values
- advanced methods from recent literature
- plausability, faithfulness, comprehensibility and consistency of explanations

Recommended prior knowledge

Participants should be familiar with fundamental methods and concepts in machine learning. They should, for example, have completed one of the following courses

- Machine Learning for Engineers
- Maschinelles Lernen für Zeitreihen
- Pattern Recognition
- Deep Learning

Basic knowledge of

- Linear Algebra and Analysis
- Probability theory
- Python (for computer exercises)
- Your OS (for computer exercises)

Reminder

- “Confirmation1“ object on studOn

Recommended literature

- C. Molnar. “Interpretable Machine Learning – A Guide for Making Black Box Models Explainable” https://christophm.github.io/interpretable-ml-book/
- A. Thampi. “Interpretable AI – Building explainable machine learning systems”, Manning, https://www.manning.com/books/interpretable-ai
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (Editors). “Explainable AI: Interpreting, Explaining and Visualizing Deep Learning”, Springer, 2019.
- HJ Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, M. van Gerven (Editors) . “Explainable and Interpretable Models in Computer Vision and Machine Learning”, Springer, 2018.
- Biran, Or, and Courtenay Cotton. "Explanation and justification in machine learning: A survey." In IJCAI-17 Workshop on ExplainableAI (XAI), p. 8. 2017, http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf.
- Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint, 2017, https://arxiv.org/abs/1702.08608.
- R Guidotti, A Monreale, F Turini, D Pedreschi, F Giannotti. "A survey of methods for explaining black box models." arXiv preprint, 2018, https://arxiv.org/abs/1802.01933.

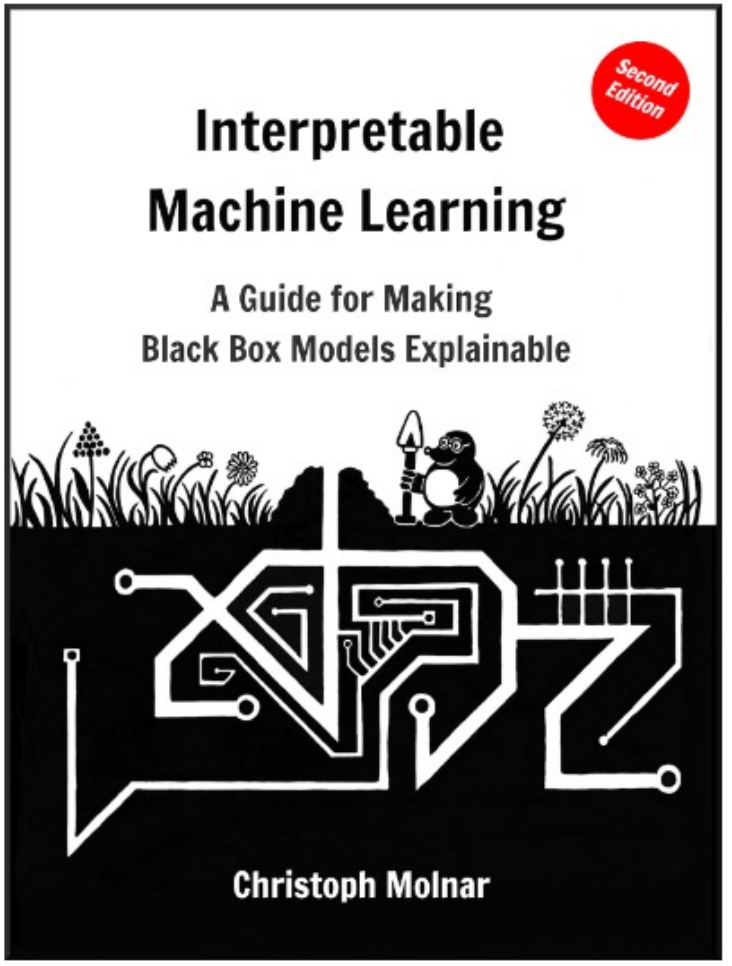
Reference Book

We will in large parts of this lecture **follow the book**

- C. Molnar. “Interpretable Machine Learning – A Guide for Making Black Box Models Explainable” https://christophm.github.io/interpretable-ml-book/

It is excellent, and imo currently the most complete and best reference available.

It is available **for free** as an online book.



Organizational Aspects

- 5.0 ECTS / 4 SWS / 60 h contact + 90 h independent
- Two timeslots per week, alternating between
  - a set of lectures for each topic/chapter
  - computer exercises on that topic/chapter
- Computer exercises are in Python, with close-to-complete templates
- Guest lecture(s) by expert(s) from the field
- One journal club session (pitch a paper of choice)
- Please feedback! participate! co-create!

Thursday timeslot

Thursday timeslot at
8:30 instead of 8:15

Preliminary Schedule

- There will be a preliminary schedule regarding the content and alternation between lectures and computer exercises (CEs) on studOn.
- It will be updated as we go and valid for the current week (and hopefully upcoming week).

Week from...	Lecture/Computer exercise
2023-04-17	Lecture
2023-04-24	W: Lecture T: CE Warmup
2023-05-01	CE1
2023-05-08	W: Lecture T: CE2
2023-05-15	W: CE2 T: Holiday
2023-05-22	Lecture
2023-05-29	CE3
2023-06-05	W: Lecture T: Holiday
2023-06-12	CE4
2023-06-19	W: Guest Lecture 1 T: Guest Lecture 2
2023-06-26	Lecture
2023-07-03	Journal Club
2023-07-10	Mock Exam / Q&A
2023-07-17	TBD

Grading Explainable Machine Learning

The grade in explainable Machine Learning is based on the **(final) exam, four computer exercises, and a journal club contribution.**

The exam will be held at the end of the semester in presence (no hybrid / online options!). **It consists of (usually) 60 multiple-choice questions**. For each question there exists exactly one correct answer. There are no negative points. **The duraction is 60 minutes**.

During the semester you will submit computer exercises (CEs). **You can achieve at most 5% of bonus points per exercise.**

At the end of the semester we will discuss recent research in the form of a journal club, **your contribution there can achieve at most 5% of bonus points.**

In total you can achieve a maximum of 20% bonus points. Those bonus points are then added on top of your exam result.

The grading scheme is static. The final grade is calculated using both exam- and bonus points. Grade 1.0 is achieved at 85%, 1.3 at 80%, ..., 4.0 at 40%. To pass the lecture you have to achieve 4.0 or 40%. The ratio of exam- to bonus points is irrelevant.

When is the exam?

Examples:

- You achieve 70% in CE2, 50% in CE3, 50% in CE4 and 40/60 correct questions in final exam. You have 3.5%+2.5%+2.5%+66,6%=75,1%. You will receive the grade 1.7
- You achieve 0% in CE1/2/3/4, have no journal club contribution and 51/60 correct questions in final exam. You have 0%+0%+0%+85%=85%. You will receive the grade 1.0
- You achieve 50% in CE1, 100% in CE2/3/4, 100% in your journal club contribution and 15/60 correct questions in final exam. You have 2.5%+5%+5%+5%+5%=22.5% bonus points. You can have 20% bonus points at most. You have 20%+25%=45%. You will receive the grade 3.7
- The exam in summer semester 2023 will take place either
  - **a) on Friday, 28.07.23, or**
  - **b) on Wednesay, 02.08.23.**

What to expect from exercises?

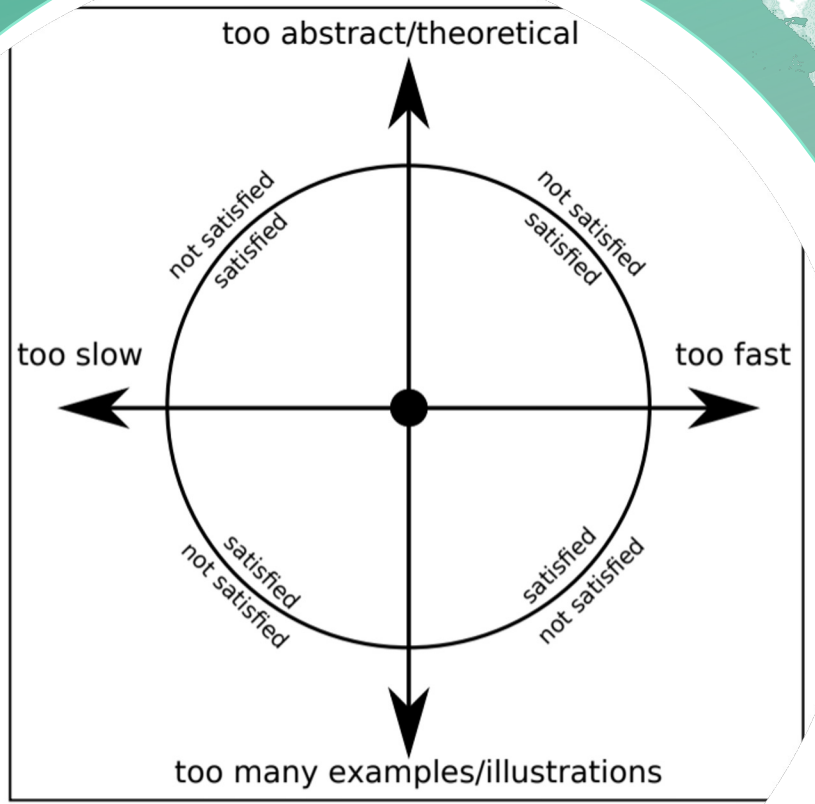
- CEs are **not required**. You can gain bonus points.
- There will be four computer exercises (CEs).
- In every CEs we will apply a subset of methods from the lectures on openly available datasets (we will even reproduce some of the plots that i have shown).
- To complete a CEs you will have to fill out a provided Jupyter Notebook with the correct code. You will see your current score at the end of the notebook (auto-grading).
- Every CEs will use two timeslots. In the first timeslot we will together start the notebook and you will continue on your own. The second timeslot is purely a long Q&A session. You **should work on your computer exercise during class**, and i will be there to answer questions you might have.

Outlook for next timeslots..

- This week: Lecture
- Next week:
  - W: Lecture
  - T: TBD

Dicuss in the forum

Ask questions and discuss with your fellow students in the studOn forum



feedback not covered by the above:

Please remember to give Feedback on studOn!

- Voice loud enough?
- Language clear?
- Zoom: Camera and mic works?
- You can also write what you like ☺

Thanks, and have a great day!