Explainable AI

Prof. Damien Garreau

Julius-Maximilians-Universität Würzburg

Winter Term 2024–2025

1. Course organization

Organization of the course

- **Wuestudy Course ID:** 08134600
- **Name on Wuecampus:** Explainable AI
- **Who?**
  - **Lectures:** myself
  - **Exercises:** M. Taimeskhanov
- **Lectures** = slides (on Moodle after the lecture)
- **Exercise sessions** = experiments (coding in Python, bring your laptop!)
- **Schedule**:
  - lectures on Wednesdays, 4-5:30pm
  - exercise sessions on Wednesdays, 2-3:30pm *(starts next week)*
- **Room:** SE 2, CAIDAS building

Evaluation

- do not forget to register to the exam
- **Evaluation:**
  - written exam at the end of the semester (definitions, pseudo-code, limitations,...)
  - exercise sessions $\rightarrow$ bonus points
- **How does the bonus work?**
  - attend the exercise sessions
  - send the notebook to Magamed at the end of the session
  - global grade $\rightarrow$ up to 10% bonus
- **Examples:** (based on 10 sessions)
  - exam = 76%, I attended all exercise sessions and made a good effort for each: I get full bonus and my final grade is $76 + 10 = 86\%$
  - exam = 96%, I attended all exercise sessions and made a good effort for each: I get full bonus and my final grade is $96 + 10 = 100\%$
  - exam = 76%, I skipped two sessions and during one session I was not paying attention and handed out a subpar notebook: bonus = 7.5%, final grade = 83.5%

Goals and pre-requisites

- **Pre-requisites:**
  - machine learning fundamentals (training set, loss, basic algorithms)
  - deep learning (usual datasets, training a network)
  - I will recall everything we need when working with specific applications (*e.g.*, images)
- **Goals of the lecture:**
  - know the XAI current **landscape**
  - know about the taxonomy
  - learn about the key **methods**
  - **re-implement** (= code) these methods
  - **apply** them on concrete examples
  - know the **limitations** of some of these methods

Related seminars

- **Seminar Selected Topics in XAI:**
  - also proposed by me
  - recent paper presentation + implementation
- **Seminar Interpretability and Explainability in Graph Learning:**
  - proposed by Prof. Scholtes
  - graph learning
- **Joint information event:**
  - Monday, October 21, 2pm, SE I, CAIDAS building
  - also by zoom (see announcement on Moodle)

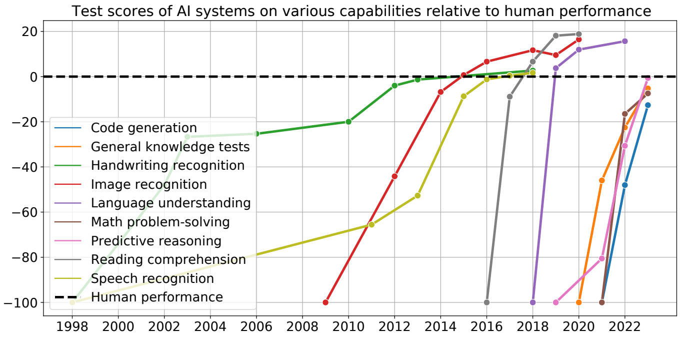
Tentative plan

- **Introduction: motivation and taxonomy**
- **Interpretable-by-design models**
- **Ad-hoc methods**
- **Perturbation-based approaches**
- **Gradient-based approaches**
- **Class Activation Maps**
- **Concept-based XAI**
- **XAI for time series**
- **Attention-based / generative models**
- **Multimodal data**

2. Introduction

AI today

- **AI today:** state-of-the-art surpasses human performance in several applications

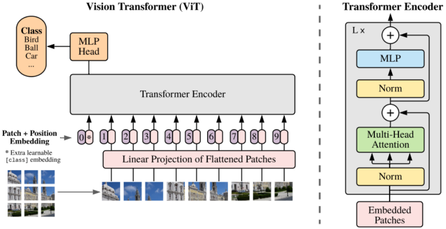


- **Figure:** test scores across different domains, figure courtesy of G. Lopardo[1]

[1] data from Kiela et al., *Plotting Progress in AI*, Contextual AI blog, 2023

How is this possible?

- **One possible reason:** complexity of the models
- complexity = architecture + parameter count
- non-linearities, skip-connection, attention mechanism,...

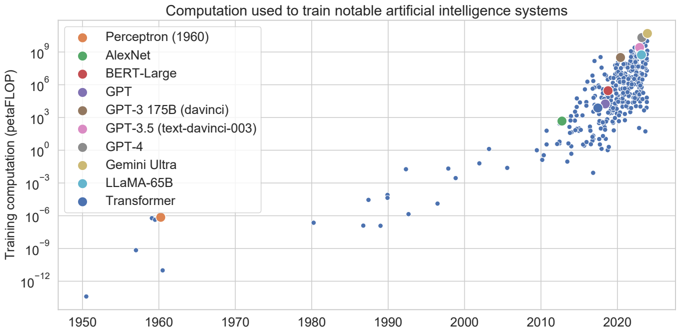


- **Figure:** vision transformer architecture[2]

[2]Dosovitskiy et al., *An image is worth* 16 × 16 *words: Transformers for image recognition at scale*, ICLR, 2021

How is this possible?

- **Consequence of model complexity:** explosion of required computing power



- **Figure:** floating point operations needed for training[3]

[3] data from *Parameter, Compute and Data Trends in Machine Learning*, Epoch AI, 2024

A motivating example

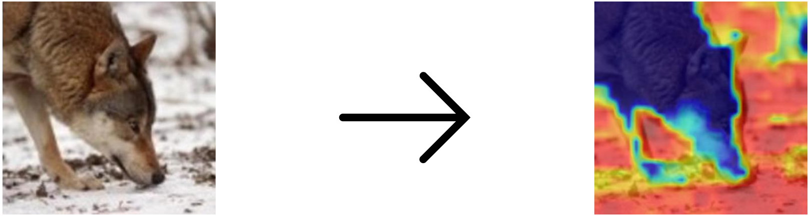
- **Consequence of complexity:** we cannot understand how individual decisions are taken
- **Motivating example:** model trained to classify husky *vs* wolf to a good accuracy[4]
- **Problem:** all images of wolves have snow in the background!
- model learns to classify "snow background" as "wolf" *(it is easier)*
- Now what happens when we feed a wolf without snow in the background to the model?

 $\longmapsto$ "husky" (!)

[4] Ribeiro, Singh, Guestrin, *"Why should I trust you?": Explaining the predictions of any classifier*, ACM SIGKDD, 2016

Motivation: debugging

- the field of **Explainable AI (XAI)** aims to provide tools addressing this issue
- **Example:** Ablation-CAM[5] explanation for an actual wolf (with snowy background)



- seeing this after training f would allow us not to release the problematic model *in the wild*
- we would fix this issue, and therefore **improve the model**

[5] Desai and Ramaswamy, *Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization*, WACV, 2020

Motivation: detecting hidden biases

- **Example:** consider a program filtering resumes for hiring at a large corporation
- if this program learns to systematically reject applications from female candidate...
- we want to know about it!
- **Spoiler alert:** this really happened:

TECH / AMAZON / ARTIFICIAL INTELLIGENCE

Amazon reportedly scraps internal AI recruiting tool that was biased against women

Illustration by Alex Castro / The Verge

/ The secret program penalized applications that contained the word "women's"

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Oct 10, 2018, 1:09 PM GMT+2

0 Comments (0 New)

Motivation: trust

- **Motivating example:** predicting pneumonia risk
- **More precisely:** predict the probability of death for patients in the next 30 days
- **Goal:** focus on the riskiest patients
- **State-of-the-art (in 2005)**:[6] neural nets
- to give concrete numbers: AUC $= 0.86$ for neural nets *vs* 0.77 for logistic regression
- although better accuracy, neural nets were considered too risky...
- ...and logistic regression was used instead (!)
- **More precisely:** another model, rule-based, learned that

$$\mathrm{HasAsthma}(x) \Rightarrow \mathrm{LowerRisk}(x) \tag{A}$$

[6] Cooper et al., *Predicting dire outcomes of patients with community acquired pneumonia*, Journal of Biomedical Informatics, 2005

Motivation: trust

- **Why?** this is an association which actually exists in the data:
- asthmatic patients with pneumonia usually admitted **directly** to the Intensive Care Unit
- there, the aggressive care received actually reduces risk w.r.t. general population
- if implemented, such rule would have a clear negative effect on the long run:
- asthmatic patients would receive worse care
- further reasoning was: if rule-based system learned (A), probably neural nets did as well
- experts could identify the problem by inspecting the model
- in critical applications, interpretability of the model is essential to gain trust
- **Interestingly,** in this setting, interpretable models can achieve near state-of-the-art accuracy[7]

[7] Caruana et al., *Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission*, KDD, 2015

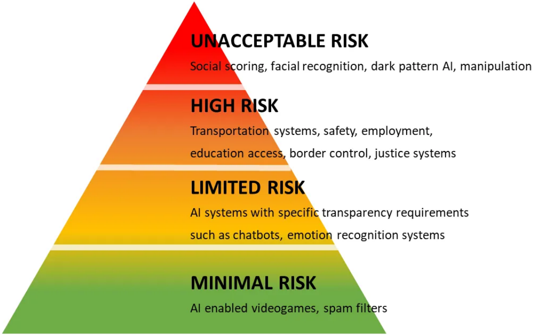
Motivation: legal requirement

- **In Europe:** General Data Protection Regulation (GDPR, 2016)
- **Articles 13 and 14:** when profiling takes place, user has the right to "meaningful information about the logic involved"
- opened a debate regarding the "right to explanation"[8]

[8]Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, International Data Privacy Law, 2017

Motivation: legal requirement

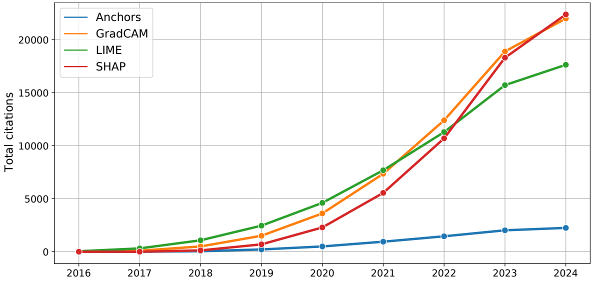
- somewhat clarified by EU AI Act (2024)
- **Main issue:** still no clear definition of “AI system” and “transparency requirements”



- **Figure:** risk classification according to the EU AI Act (figure credits Lori Witzel)

Perspectives

- despite many existing methods (which we will learn everything about), XAI is still a growing field of research
- the quest is not over!



- **Figure:** total citations for the most prominent XAI methods

Summary

- current machine learning models are "black-boxes"
- need XAI to:
  - improve the models
  - detect hidden biases
  - gain trust of users
  - comply to legal requirements
- XAI methods are essential to achieve social acceptability of machine learning algorithms

3. Taxonomy

First definitions

- two concurrent phrasings:
  - **Interpretability:** an interpretable model is transparent in its operation and provides information about the relationships between inputs and outputs
  - **Explainability:** ability to explain the decision-making process of a model in human-understandable terms
- **Slight nuance:** *interpretability* is more concerned with interpretable-by-design models, whereas *explainability* refers to tools making black-box models interpretable
- **Important:** no agreement within the machine learning community!
- we will use both terms interchangeably

Taxonomy: pre / post modeling

- one can focus on different steps of the machine learning pipeline
- **Pre-modeling:** making the model interpretable before training
  - **Example:** creating interpretable features, selecting only relevant features
  - **Pros:** easy to understand
  - **Cons:** quite restrictive (interpretable features do not always exist)
- **Explainable modeling:** creating *interpretable-by-design* models
  - **Example:** linear models, decision trees, concept-based models
  - **Pros:** easy to understand
  - **Cons:** restrictive (models are too simple)
- **Post-modeling:** model is already trained, **post-hoc** inspection (after the fact)
  - **Example:** gradient-based approaches
  - **Pros:** flexible (to retraining / fine-tuning)
  - **Cons:** hard to leverage insights (cannot modify the model)
  - $\rightarrow$ most frequent approach

Taxonomy: model-specific / model-agnostic

- **Model-specific (= ad-hoc):** rely on particular properties of the algorithm to explain
  - **Example:** look at the largest coefficients of a linear model
  - **Pros:** stay close to the true operating procedure
  - **Cons:** not very adaptive
- **Model-agnostic:** consider the model as a black-box
  - **Example:** perturbation-based approaches
  - **Pros:** applicable to any (or a large class of) model(s)
  - **Cons:** can only rely on queries to the model
  - $\rightarrow$ most frequent approach
- **Remark:** some methods fall *in-between*
- **Example:** taking a gradient $\rightarrow$ assuming it exists / it is non-zero

Taxonomy: scope

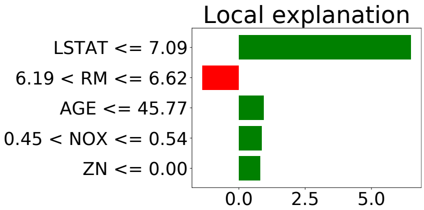
- "scope" = what is the scale of the explanations we provide?
- **Global:** cover all the input space of the model
  - **Example:** feature 3 is important for predicting the output
  - **Pros:** do it once and for all
  - **Cons:** usually complicate function, hard to summarize all of it
- **Sub-groups:** how the model behaves on group of observations / part of the input space
  - **Example:** if feature 3 lies between 4.9 and 5.2, it has a positive influence on the output
  - **Pros:** easier to understand
  - **Cons:** defining groups (clustering) is a challenging problem *per se*
- **Individual:** focus on a particular example ξ (= *local* explainability)
  - **Example:** feature 3 is important for predicting $f(\xi) = 0.9$
  - **Pros:** very easy to understand
  - **Cons:** have to re-compute explanation for each new example
  - $\rightarrow$ most frequent approach

Taxonomy: explanation type

- very important point: **how the explanation is presented to the user**
- depends on:
  - the **XAI method**
  - the **data-type** (tabular, text, image, graph, etc.)
  - the **intended user** (expert or not)
- non-exhaustive list:
  - **feature importance**
  - **local rules**
  - **visualizations**
  - explanation by **example**
  - ...

Feature importance: tabular data

- **Tabular data:** spreadsheet data
- feature importance gives one real-number per feature
- if > 0, positive contribution to the prediction
- **Example:** LIME explanation for squared meter price prediction (Boston housing dataset[9])

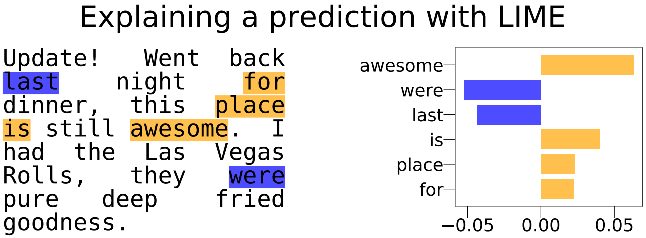


- **Remark:** even with few features, not all are displayed

[9] Harrison and Rubinfeld, *Hedonic housing prices and the demand for clean air*, Journal of environmental economics and management, 1978

Feature importance: text data

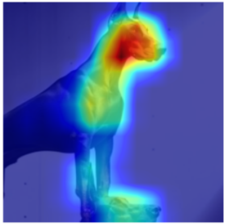
- **Text data:** sequence of words
- in practice, tokenized (and tokens $\neq$ words)
- highlight words in the document, can also display more precise explanation
- **Example:** LIME explanation for prediction of a positive sentiment on a Yelp review[10]



[10] courtesy of Mardaoui and Garreau, *An analysis of LIME for text data*, AISTATS, 2021

Feature importance: image data

- **Image data:** $H \times W$ pixels, C channels
- feature $=$ each channel of a pixel
- agglomerate the values for each channel, per pixel (take the norm) $\rightarrow$ heatmap
- **Example:** GradCAM[11] explanation for classification as "doberman" by a VGG16



[11] Selvaraju et al., *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*, ICCV, 2017

Local rules

- **Local rules:** simple *if-then-else* statement summarizing behavior of the model around ξ
- **Example:** explanation given by LORE[12] for the example

$$\xi = \{(\mathsf{age} = 22), (\mathsf{job} = \mathsf{none}), (\mathsf{amount} = 10^4), (\mathsf{car} = \mathsf{no})\}$$

  is the following simple rule:

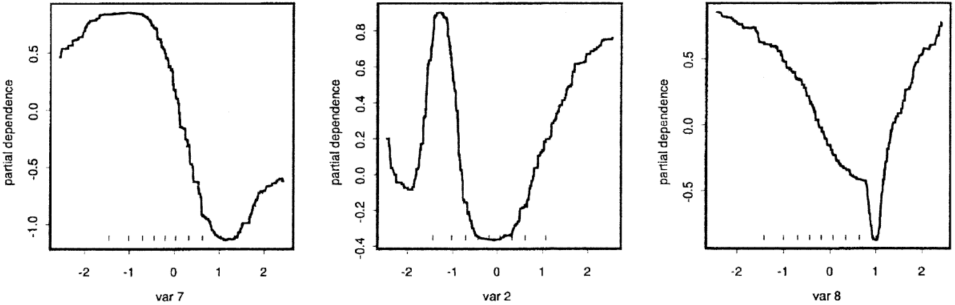
$$\{\mathsf{age} \leq 25, \mathsf{job} = \mathsf{none}, \mathsf{amount} > 5 \cdot 10^3\} \rightarrow \mathsf{deny}\,.$$

- easy to understand!

[12] Guidotti et al., *Local rule-based explanations of black-box decision systems*, preprint, 2018

Visualizations

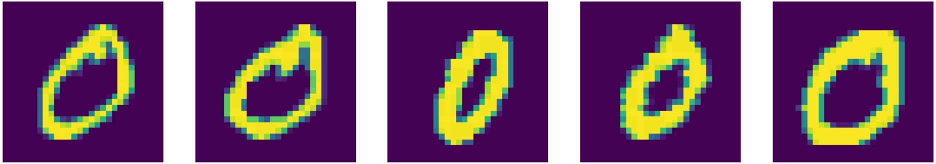
- **Visualizations:** useful for complex explanations
- typical for global feature importance
- **Example:** partial dependency plots[13] $\approx$ variation of output w.r.t. each feature



[13] Friedman, *Greedy function approximation: a gradient boosting machine*, The Annals of Statistics, 2001

Explanation by example: prototypes

- **Prototypes:** data points close to the example to explain with similar predictions
- **Example:** explaining k-nearest neighbors[14] prediction on MNIST

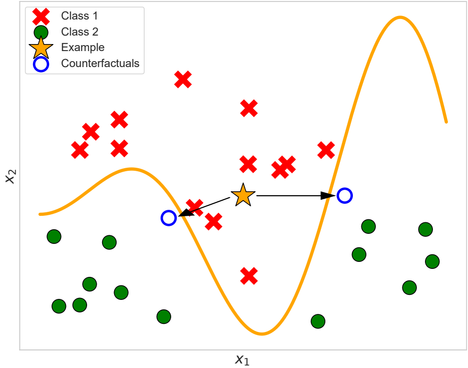


- **Beware:** simply plotting nearest neighbors is explaining the data, not the model (if we are talking about any model)

[14] Fix and Hodges, *Discriminatory analysis, nonparametric discrimination*, Tech. Report, 1951

Explanation by example: counterfactuals

- **Counterfactuals:** smallest perturbation of the input changing the decision
- **Example:** "What do I need to change for the bank to approve my loan?"



- **Remark:** similar to adversarial examples, but different goal
- we do not want to fool the model, rather explain its behavior