EXPLAINABLE ARTIFICIAL INTELLIGENCE

L193 – Lecture 1 – Lent 2025

UNIVERSITY OF CAMBRIDGE

WHO WE ARE

Prof. Mateja Jamnik

Professor of Artificial Intelligence

mateja.jamnik@cl.cam.ac.uk

Dr. Zohreh Shams

Lead Researcher
Leap Labs

zs315@cam.ac.uk

Mateo Espinosa Zarlenga

Fourth-year PhD Student

me466@cam.ac.uk

WHEN MACHINES MEET THE REAL-WORLD

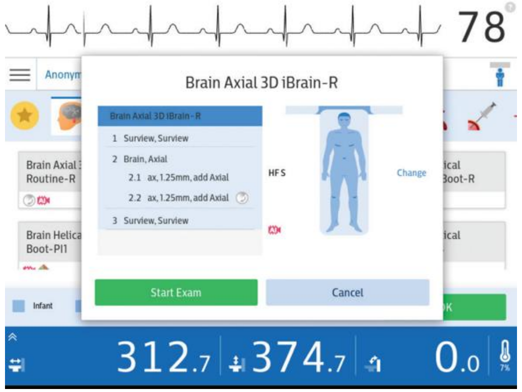
Machine learning is increasingly getting intertwined with **our day-to-day experience**:

- speech, medical diagnosis, credit risk, screening CVs, content recommendations, autonomous vehicles, law, search engines, chatbots, image generation…

Often ML models are **black-box**.

Self-driving cars
(e.g., Waymo, Tesla)

Court Rulings
(e.g., COMPAS)

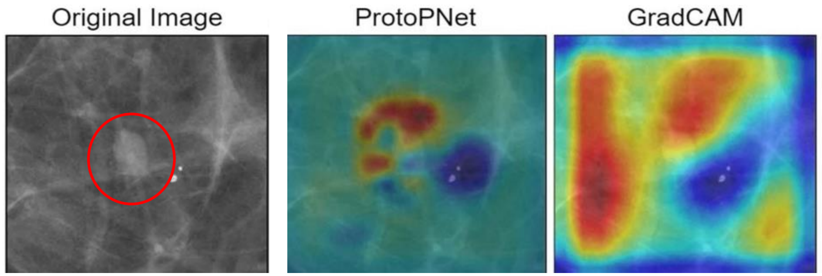


Healthcare
(e.g., Phillips Machines)

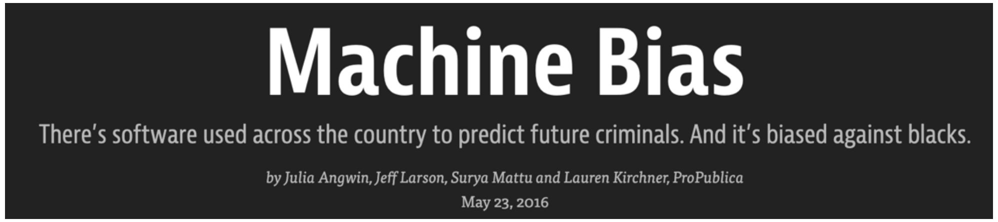
ChatBots
(e.g., ChatGPT, Gemini)

WHY IS EXPLAINABILITY IMPORTANT?

Things **CAN and WILL go south** when using black-box models in high-stakes tasks.



DNNs in computer-aided mammography focused mostly on healthy tissue rather than tumour!



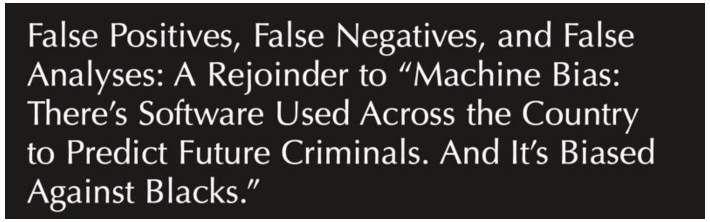
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

VS



False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks."

ProPublica claims black-box **COMPAS is racially biased**

Further studies show that **the analysis might've been mistaken**!

[1] Adapted from Barnett et al. "*A case-based interpretable deep learning model for classification of mass lesions in digital mammography.*" Nature Machine Intelligence (2021).
[2] Angwin, Julia, et al. "Machine bias." *Ethics of Data and Analytics*. Auerbach Publications, 2016. 254-264.
[3] Flores, Anthony W., Kristin Bechtel, and Christopher T. Lowenkamp. "False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks." *Fed. Probation* 80 (2016): 38.

WHY IS EXPLAINABILITY IMPORTANT?

The list keeps going on and on…

IBM Watson AI criticised after giving 'unsafe' cancer treatment advice

Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

Why Amazon's Automated Hiring Tool Discriminated Against Women

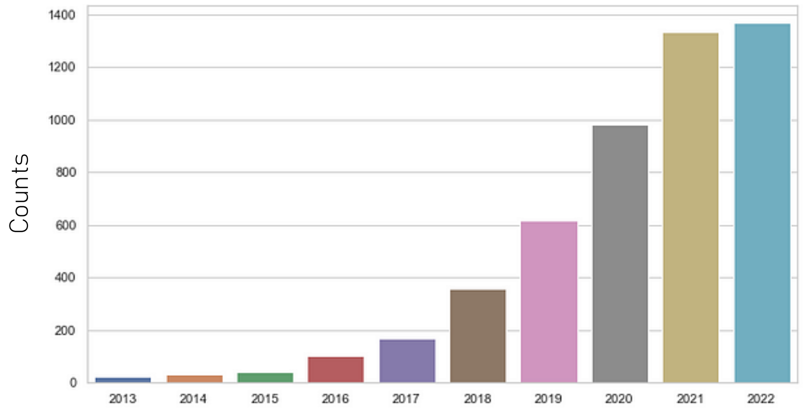
Predictive policing algorithms are racist. They need to be dismantled.

https://www.telegraph.co.uk/technology/2018/07/27/ibm-watson-ai-criticised-giving-unsafe-cancer-treatment-advice/
https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html
https://www.aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against
https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

WHO CARES ABOUT EXPLAINABILITY?

Academia

Number of XAI Papers Published



Jacovi, Alon. "Trends in explainable AI (XAI) literature." *Medium* (2023).

WHO CARES ABOUT EXPLAINABILITY?

Academia

Explainable Artificial Intelligence (XAI)

(DARPA 2016)

TAILOR

(EU Horizon Program)

Industry

Generative-AI-related risks that caused negative consequences for organizations,[1] % of respondents

The State of AI in 2024 (McKinsey)

WHO CARES ABOUT EXPLAINABILITY?

The Public

CIO BLOG

Companies Grapple With AI's Opaque Decision-Making Process

Uber, Xerox's PARC, Capital One among organizations investigating how AI solves problems

How to Build Artificial Intelligence We Can Trust

Computer systems need to understand time, space and causality. Right now they don't.

Opinion **Artificial intelligence**

Beware the rise of the black box algorithm

FORBES › INNOVATION › ENTERPRISE TECH

Building Trust In AI: The Case For Transparency

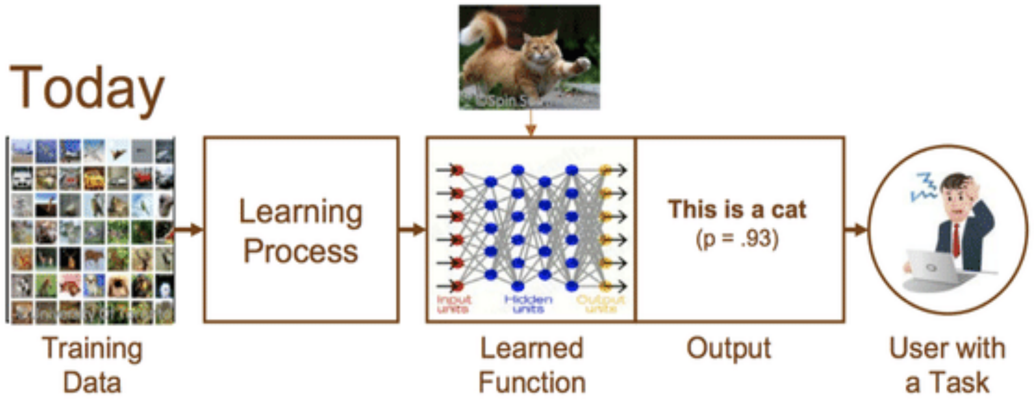
WHO calls for safe and ethical AI for health

16 May 2023 | Departmental update |Reading time: 2 min (507 words)

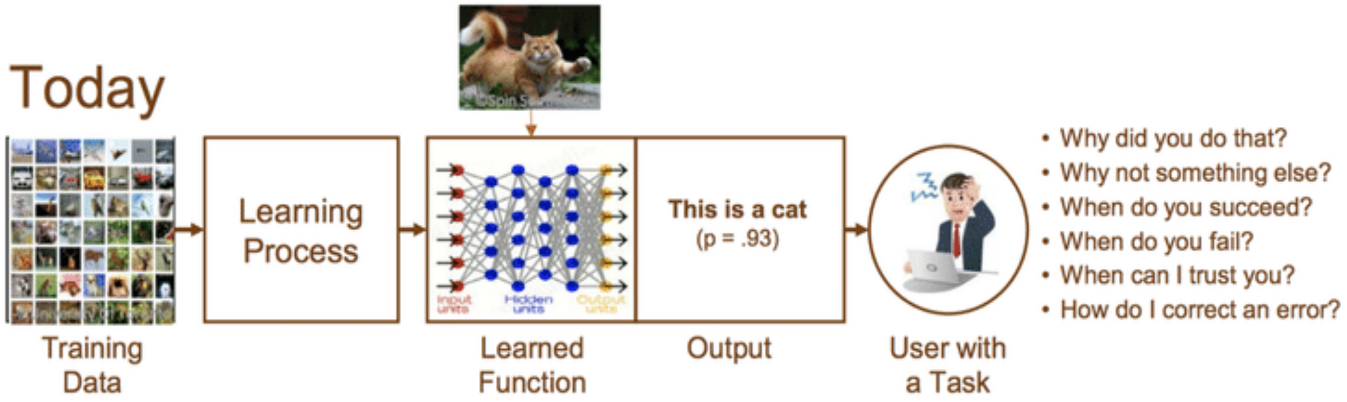
Why businesses need explainable AI—and how to deliver it

September 29, 2022 | Article

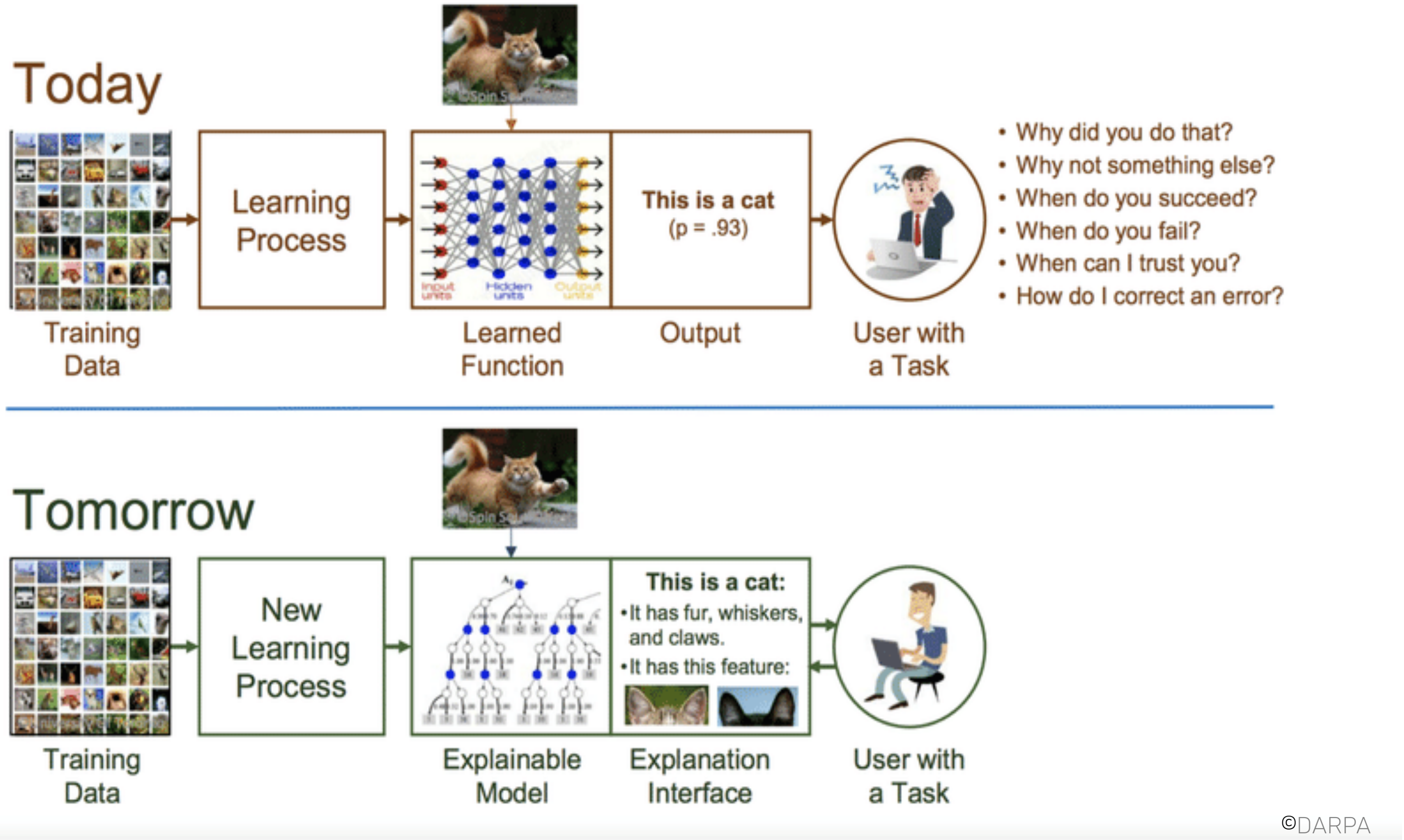
EXPLAINABLE WHAT?



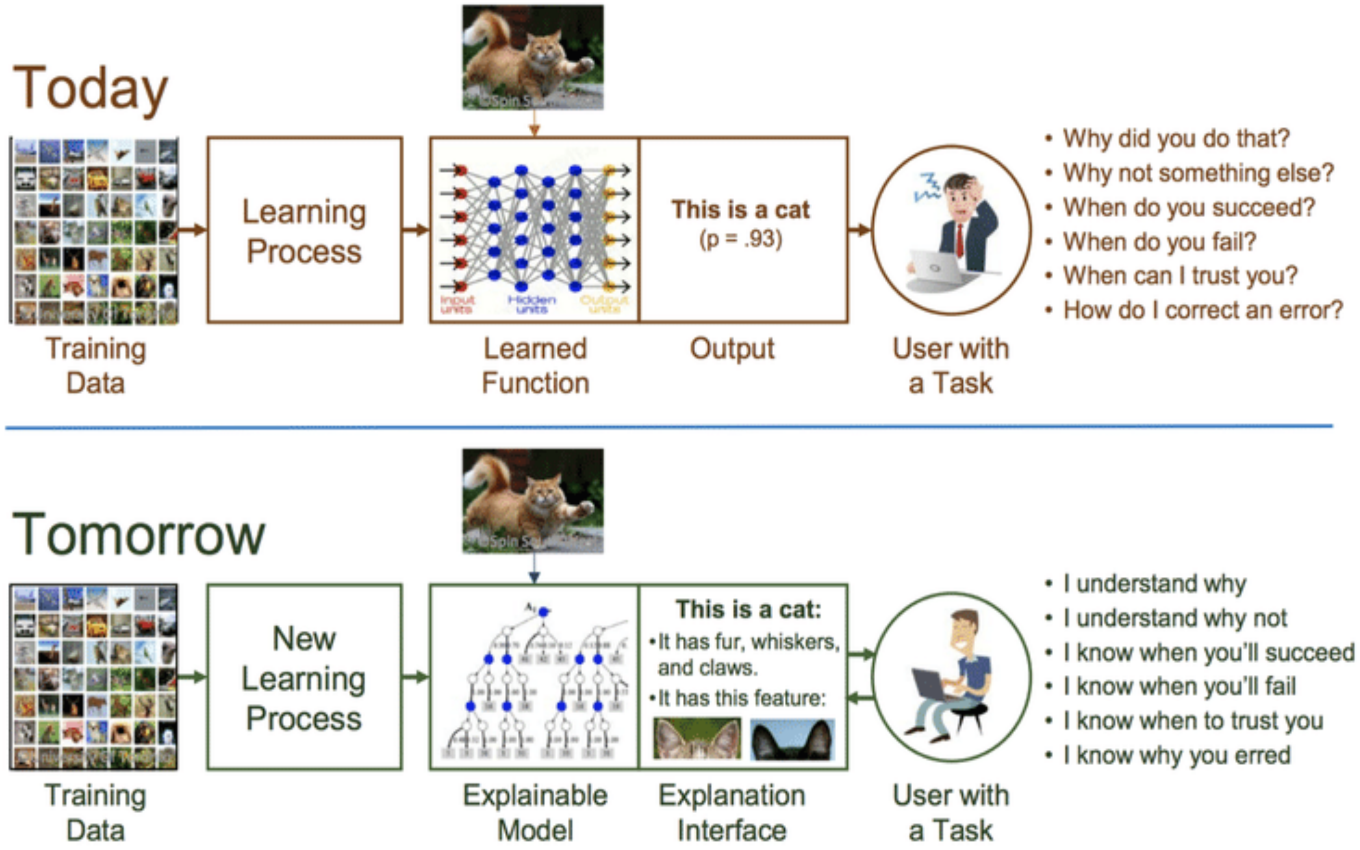
EXPLAINABLE WHAT?



EXPLAINABLE WHAT?



EXPLAINABLE WHAT?



Types of XAI questions:

- What does the prediction mean?
- How did the model make a prediction?
- Which features contributed to a certain prediction and how?
- How can a model learn or select features that are the most interpretable or informative?
- How much does each sample contribute to model training?

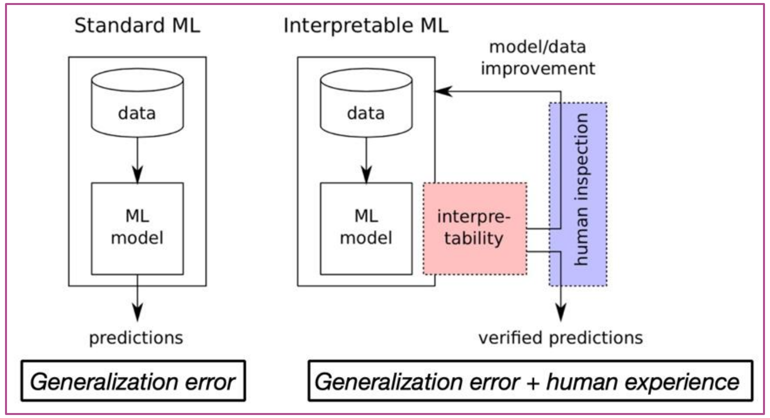
EXPLAINABILITY GIVES WHAT?

Debug and debias predictions

SHAP
(Lundberg et al., 2017)

Verify systems

Improve models



©Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Knowledge Discovery

Strategy Discovery
(Schut et al., 2023)

Theorem Discovery
(Davies et al., 2021)

EXPLAINABILITY GIVES A FOUNDATION FOR RESPONSIBLE AI

- **Competence**: XAI for improving/debugging models
- **Fairness**: XAI for removing unwanted bias
- **Safety**: XAI for making safer decisions
- **Usability**: XAI for actionable decision making
- **Human-AI collaboration**: XAI for better control and user interaction
- **Accountability**: XAI for enabling documentation and governance
- **Privacy**: XAI to preserve privacy

Legislation: anti-discrimination laws, GDPR (Article 22), EU AI Act, USA AI Bill of Rights, etc.

General Data Protection Regulations (GDPR, 2016):

- "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling,…" (Art. 22)
- The data subject has the right to "meaningful information about the logic involved" in the decision. (Art. 13 and 15)

EU AI Act (2024):

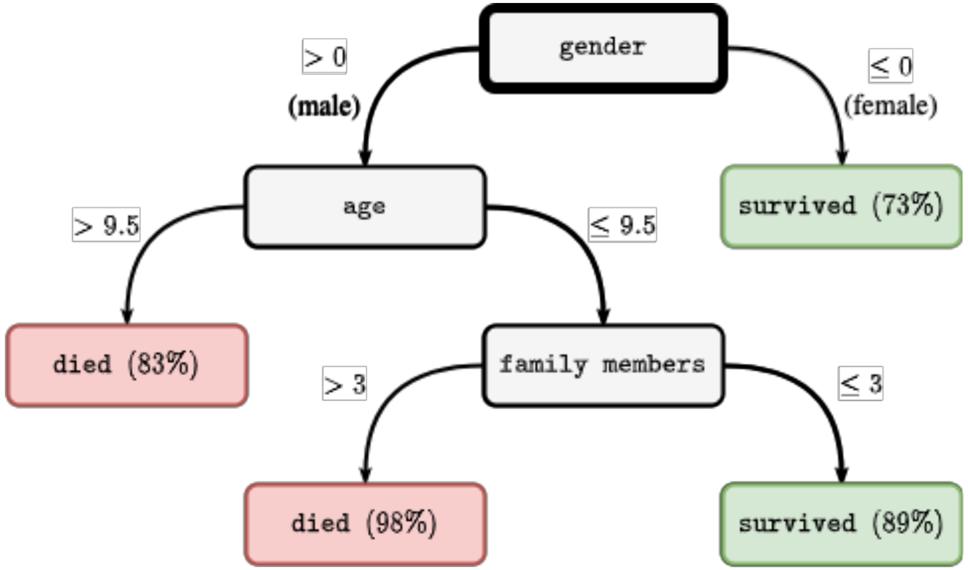
- "Any affected person subject to a decision which is taken by.. a high-risk AI system … shall have the right to obtain from the deployer clear and meaningful explanations (Art. 86)

EXAMPLE: DECISION TREES

Model is interpretable, because prediction can be explained with a rule

Explanation:

```
If a passenger was male and
under 9.5 years of age and
there were 3 or fewer members
in their family, then there
was an 89% chance that they
survived.
```



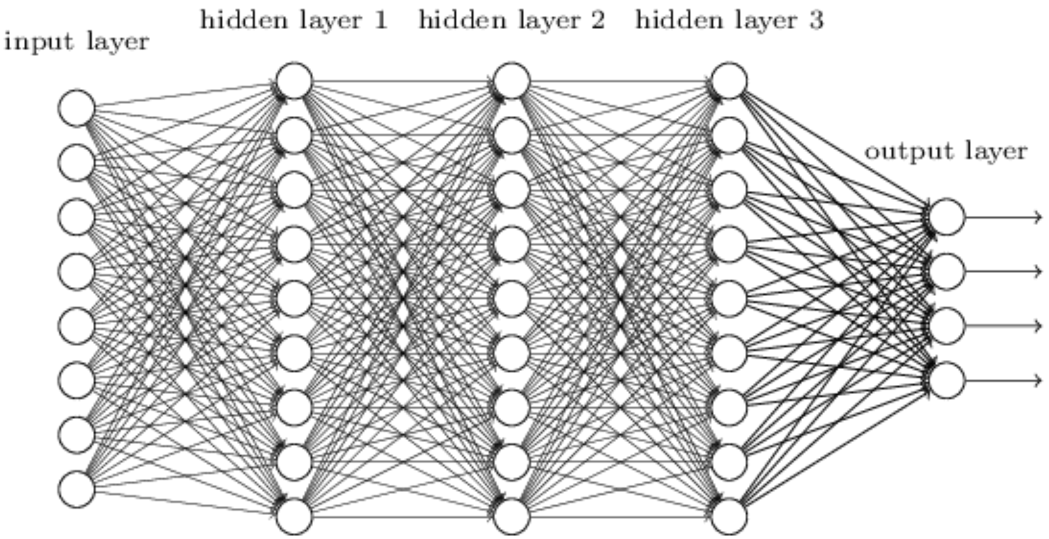
$$\texttt{IF}\ \Big((\texttt{gender} > 0) \wedge (\texttt{age} \leq 9.5) \wedge (\texttt{family members} \leq 3)\Big)\ \texttt{THEN survived}$$

"Women and children first"

[1] Diagram inspired by "Survival of passengers on the Titanic." Wikimedia Commons 2022.

EXAMPLE: DEEP NEURAL NETWORKS

- Deep Neural Nets (DNNs) are "**black-box"** models
- Predictions can be explained mathematically
- But their evaluation is highly non—linear, so it is difficult to understand what factors determined a prediction
  - Even more complicated with modern architectures (hundreds of layers + attention + convolutions + normalisation layers + etc…)
- **Explanation**: current approaches explain some of these factors in terms of
  - Data
  - Important features, combinations of features
  - Rules from approximations of DNNs
  - Influential examples, counterexamples



[1] Michael A. Nielsen, "Neural Networks and Deep Learning", Determination Press, 2015

WHAT THIS MODULE IS ABOUT

- Definition of an explanation: what does it mean to explain a model?
- Explainability methods for black-box models: we focus on **deep neural networks**, although some methods we will discuss are applicable to other ML models
- Taxonomy of XAI approaches: how is the XAI field divided? What are its active research areas?
- Survey of XAI methods: Feature importance, concept-based methods, prototypical explanations, self-explaining DNNs, influence functions, mechanistic interpretability

WHAT THIS MODULE IS NOT ABOUT

XAI tackles many areas, but we do not cover them in this module:

- "Traditional" inherently interpretable models (i.e., non-DNN interpretable models)
- Bias and fairness of data or decisions: is the prediction based on biased features?
- Privacy: how is data processed, is it anonymised for training?
- Transparency: can we inspect the way decisions are made?
- Planning: which actions are responsible for a plan?

ROUGH ROADMAP FOR NEXT FEW WEEKS

Format

1. Lectures: Fridays in weeks 1, 2, 3, 5, 7 in LT2
2. Practicals in lab (hands-on of XAI methods and exercises): Fri week 4 in SW02 2-4pm, Tue week 6 in SW02 3-5pm
3. Presentations: Fridays in weeks 2, 3, 5, 7, 8 in LT2

Topics Covered (focus on XAI for Deep Neural Networks)

- Overview and taxonomy of XAI
- Feature attribution methods
- Saliency methods
- Concept-based explainability
- Self-explaining architectures

Assessment

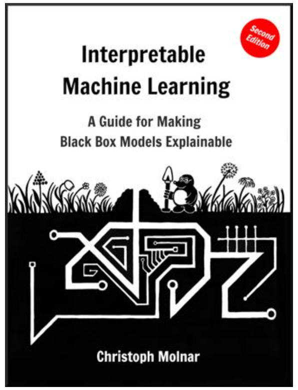
- ✓ **10%** Jupyter practical 1 due on 24 February 2025 2pm
- ✓ **10%** Jupyter practical 2 due on 11 March 2025 2pm
- ✓ **10%** Paper presentation
- ✓ **70%** Mini project (implement, modify, experiment, combine approach from a research paper) due on 28 March 2025 4pm

Course web page https://www.cl.cam.ac.uk/teaching/2425/L193/

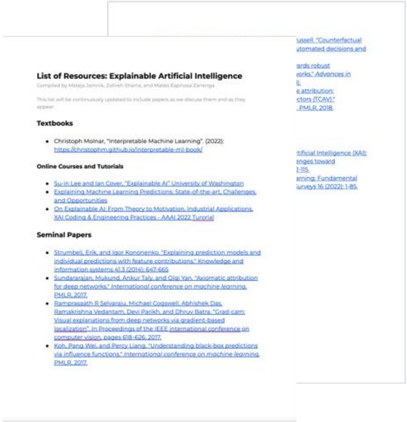
Submission On Moodle via course web page

READING MATERIAL

No official textbook, but some resources:



"Interpretable Machine Learning" Christoph Molnar
https://christophm.github.io/interpretable-ml-book/



List of Resources: Explainable Artificial Intelligence

Compiled by Mateja Jamnik, Zohreh Shams, and Mateo Espinosa Zarlenga

This list will be continuously updated to include papers as we discuss them and as they appear.

Textbooks

- Christoph Molnar, "Interpretable Machine Learning". (2022): https://christophm.github.io/interpretable-ml-book/

Online Courses and Tutorials

- Su-in Lee and Ian Cover, "Explainable AI" University of Washington
- Explaining Machine Learning Predictions: State-of-the-art, Challenges, and Opportunities
- On Explainable AI: From Theory to Motivation, Industrial Applications, XAI Coding & Engineering Practices - AAAI 2022 Tutorial

Seminal Papers

- Strumbelj, Erik, and Igor Kononenko. "Explaining prediction models and individual predictions with feature contributions." Knowledge and information systems 41.3 (2014): 647-665.
- Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." *International conference on machine learning*. PMLR, 2017.
- Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In Proceedings of the IEEE international conference on computer vision, pages 618-626, 2017.
- Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions." *International conference on machine learning*. PMLR, 2017.

List of Compiled Resources
https://t.ly/Zmfze

OVERVIEW OF THE WORLD OF XAI

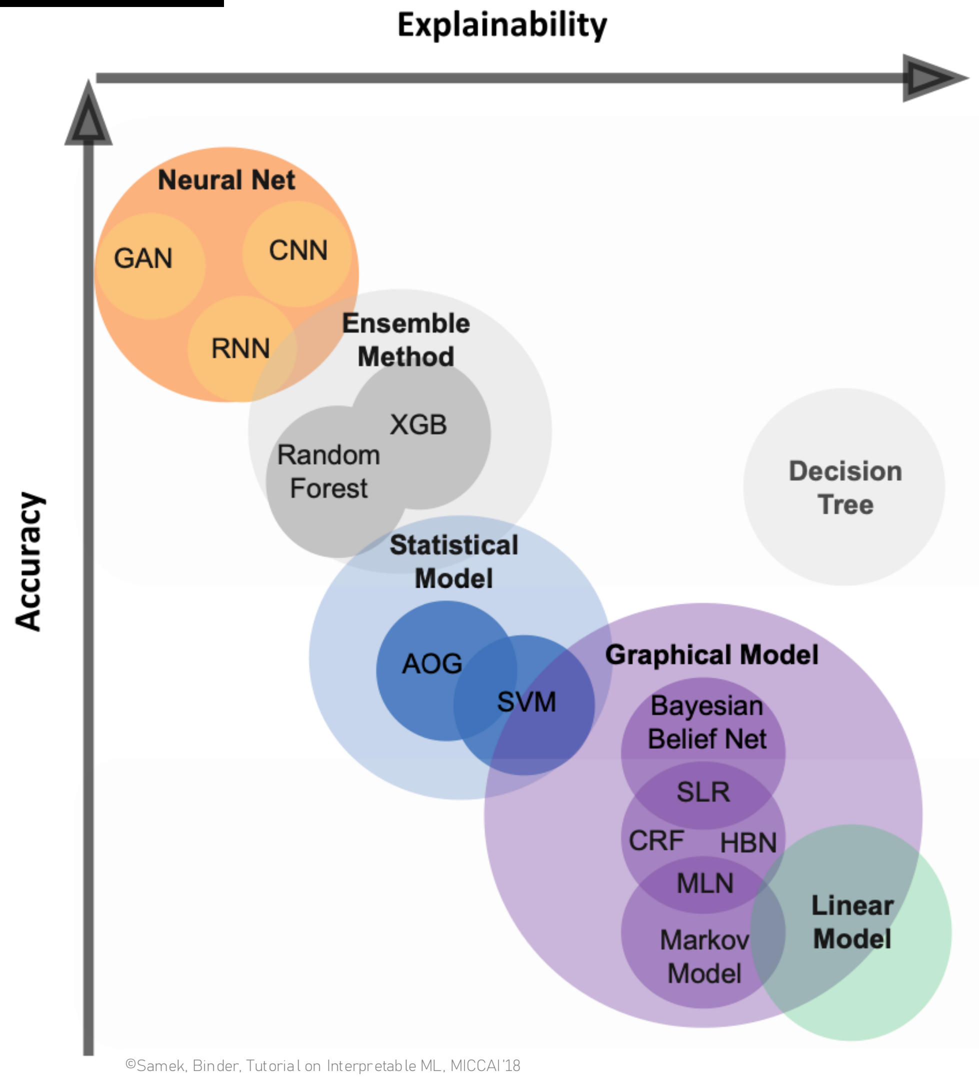
SOME DEFINITIONS

- Confusing nomenclature: explainable / interpretable / transparent models
  - Interpretability: the ability to explain or provide the meaning in understandable terms to humans
  - Explainability: a notion of explanation as an interface between humans and a decision maker that is both an accurate proxy of the decision maker and comprehensible to humans
  - Transparency: a model is transparent if by itself it is understandable.

A TAXONOMY FOR ML MODELS

- Inherently explainable/glass box models:
  - Linear models
  - Decision trees
  - Rule-based models

- Black-box models:
  - Deep neural networks
  - Ensemble models

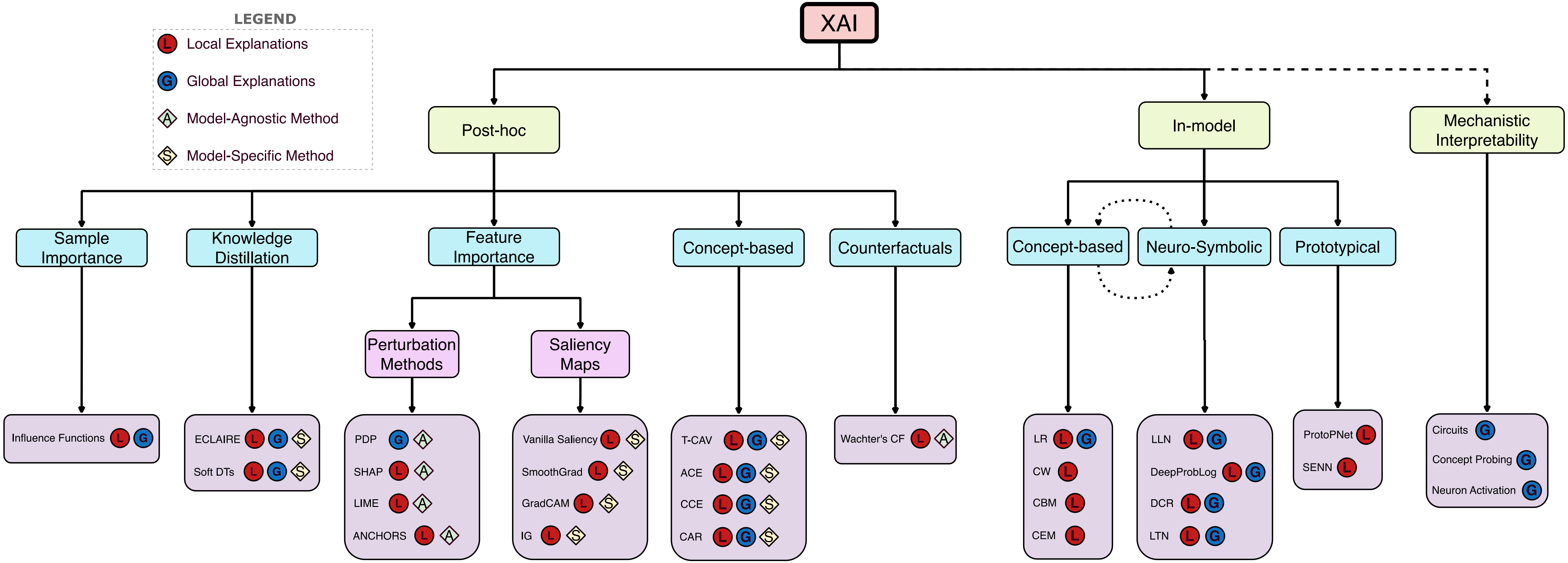


Accuracy vs Explainability

A TAXONOMY FOR XAI METHODS FOR BLACK-BOX MODELS

- When is explanation extracted: in-model (inherently interpretable), post-hoc
- Does it explain a particular sample or the whole model: **local** L, **global** G, **both**
- Does it depend on a particular model: **model-specific** S, **model-agnostic** A
- Does it explain the model or an approximation of the model: **visualisation, surrogate**

DIFFERENT TYPES OF EXPLANATIONS



DIFFERENT WAYS OF PRODUCING THOSE EXPLANATIONS

Explanation modes:

- Analytic statement: natural language descriptions of elements and context that support the decision
- Visualisations: highlight parts of data that support the decisions and allow user to make their own understanding
- Cases: give typical/illustrative examples that support the decision
- Rejections or alternative choice: counterfactuals or common misconceptions that argue against the alternative decisions

GLOBAL MODEL AGNOSTIC INTERPRETATION METHODS

MOTIVATION

- Explaining the average behaviour of a model
- Understanding and debugging the general mechanism of the model

How is attempted in practice

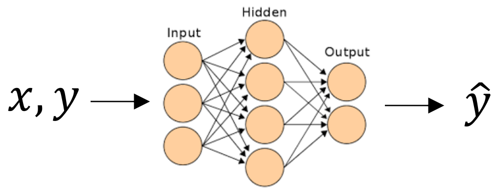
- Global Feature Importance
- Knowledge Distillation
- Prototypical Examples

KNOWLEDGE DISTILLATION

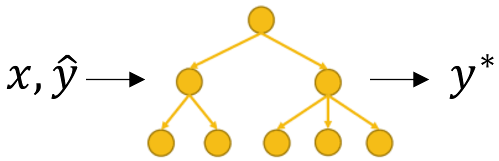
KNOWLEDGE DISTILLATION I

Intuition: an interpretable model is trained to approximate the predictions of a black model and then used to explain its predictions

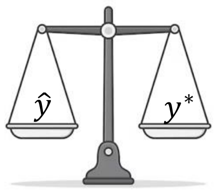
Step 1: train a black box model on some data x with labels y



Step 2: train an interpretable model on x and $\hat{y}$



Step 3: check the alignment of $\hat{y}$ and y^*



Step 4: use the well-aligned surrogate for interpreting $\hat{y}$

SURROGATE ALIGNMENT

R^2 measures the percentage of variance that is captured by the surrogate model

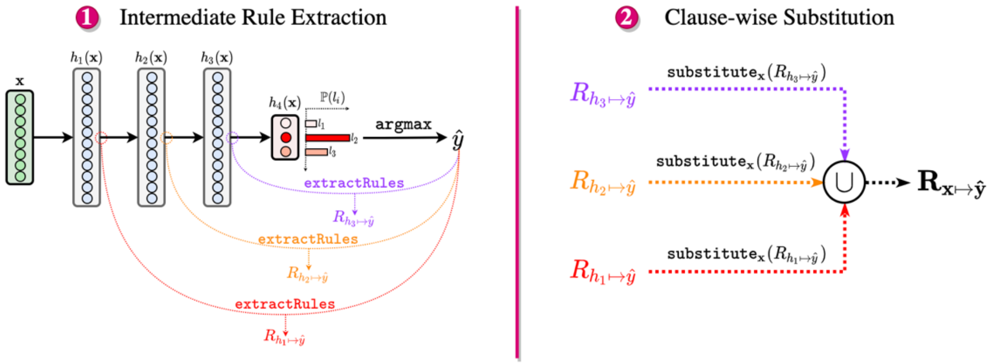
$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i^* - \widehat{y_i})^2}{\sum_{i=1}^{n}(\widehat{y_i} - \bar{\hat{y}})^2}$$

R^2 close to 1: surrogate is great
R^2 close to 0: surrogate is not good enough

- SSE: sum of squares error
- SST : sum of squares total
- y_i^*:surrogate model prediction for instance i
- $\hat{y}_i$: black-box model prediction for instance i
- $\bar{\hat{y}}$: mean of black-box model predictions

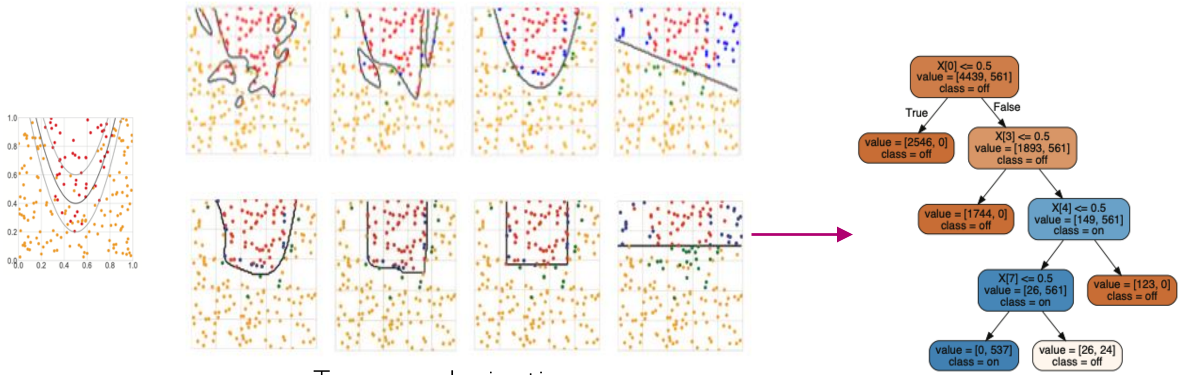
SURROGATE EXAMPLES

ECLAIRE: Rule extraction from pre-trained models



Soft DTs: Training models such that their decisions boundaries can be approximated with simple decision trees

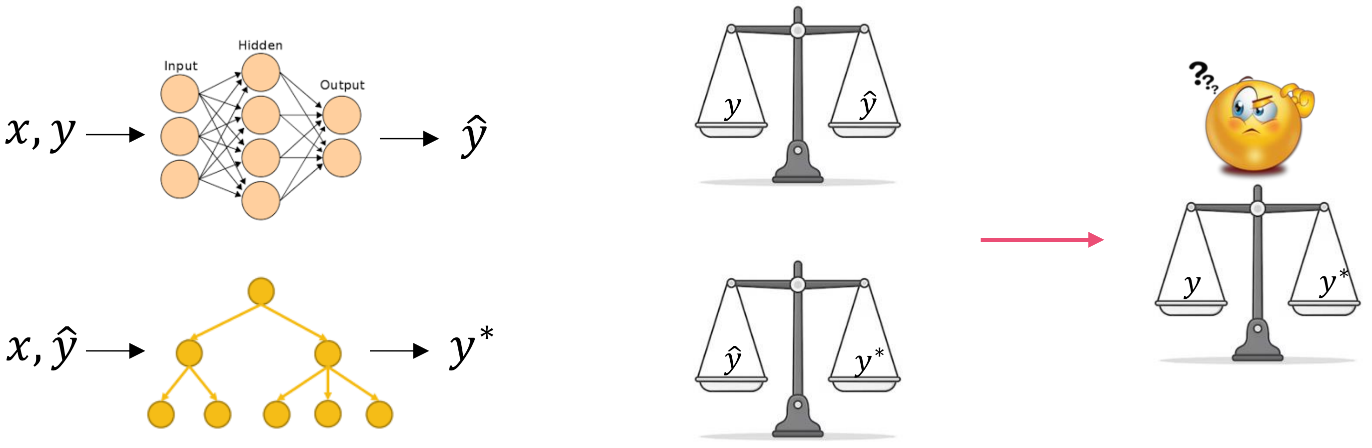
L2 regularisation



Tree regularisation

KNOWLEDGE DISTILLATION II

Ideal Scenario

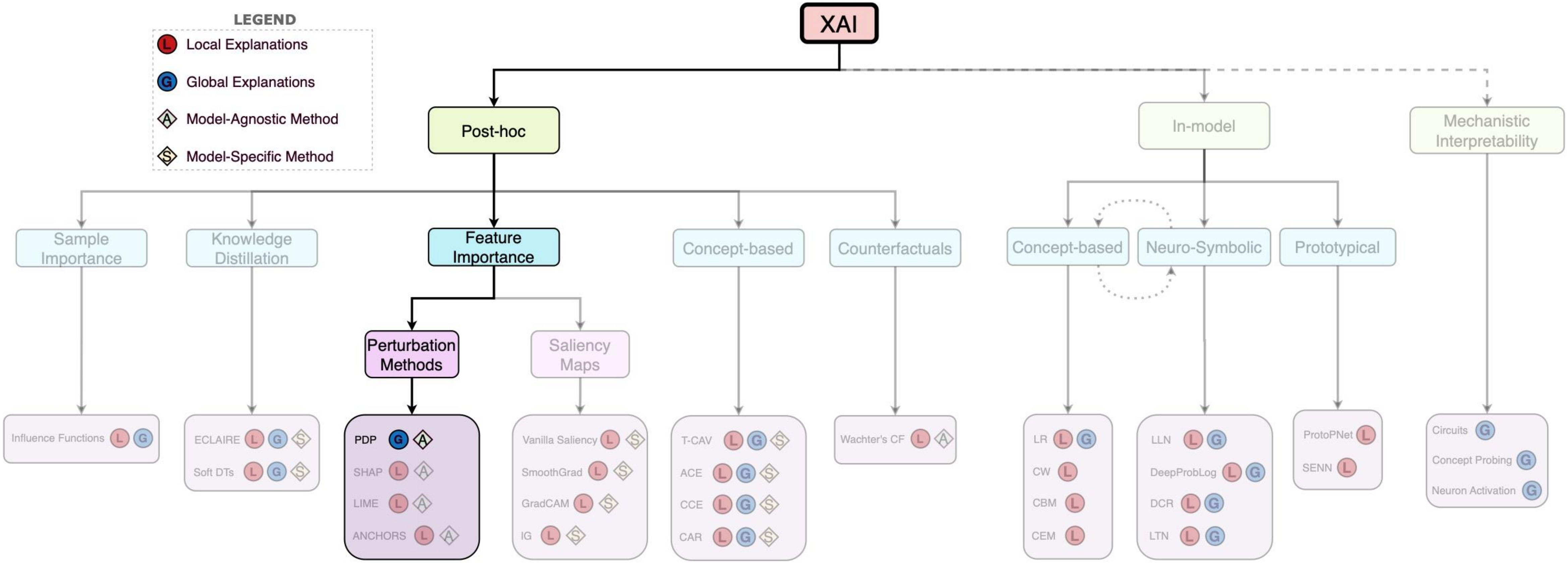


Knowledge distillation

If we can't approximate globally, can we look at features globally important or can we approximate locally?

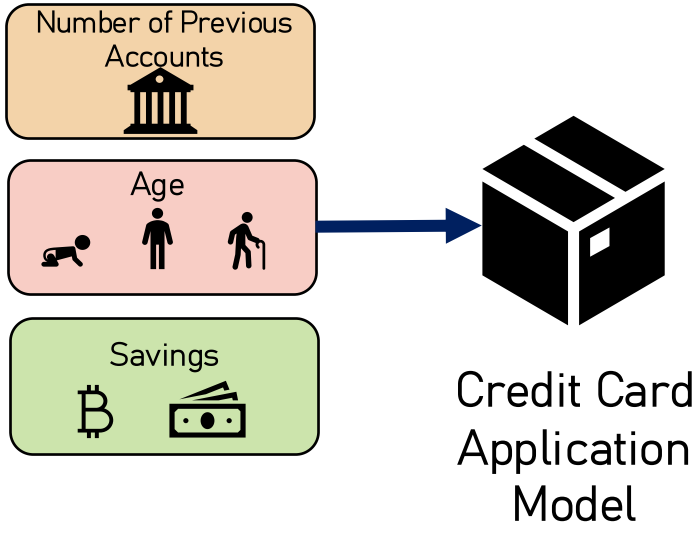
GLOBAL PERTURBATION METHODS

GLOBAL PERTURBATION METHODS



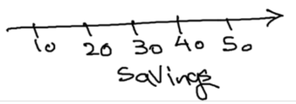
PARTIAL DEPENDENCE PLOT (PDP)

PDP [1] measures the marginal effect of a feature on the prediction of the model while holding other features constant.



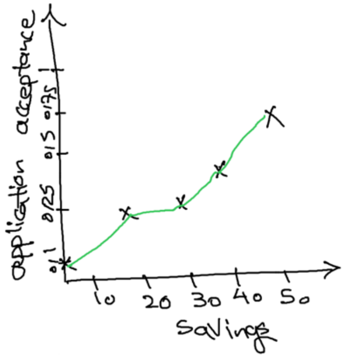
Step 1: select a feature

Step 2: define a grid over feature values



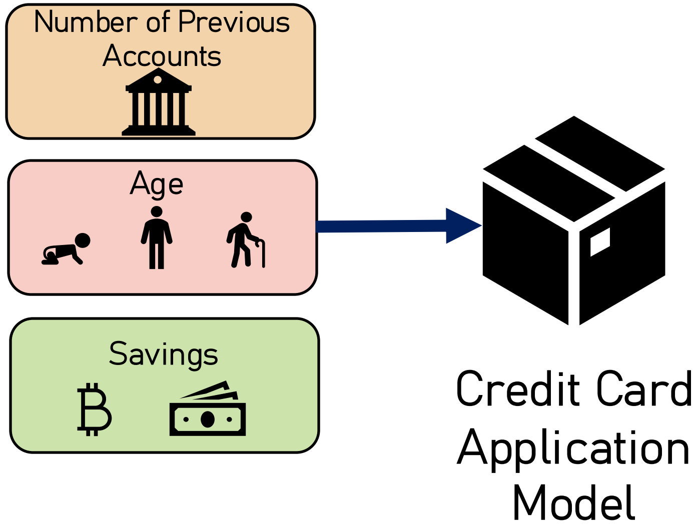
Step 3: replace all values of the feature with the grid value

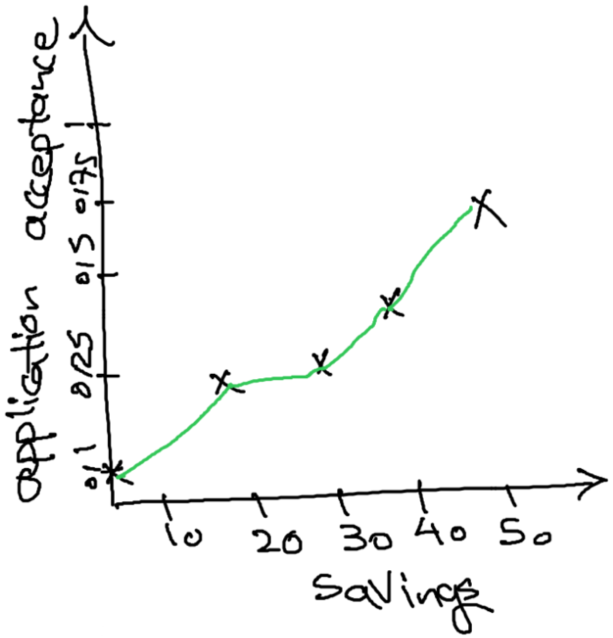
Step 4: calculate and average the prediction of the target

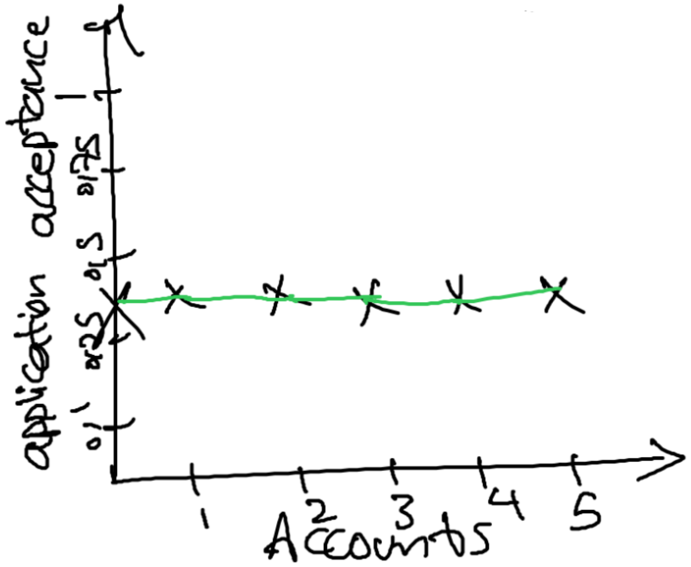


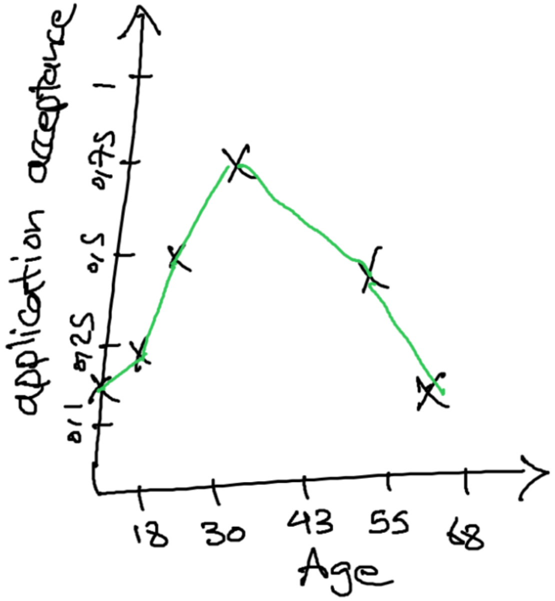
[1] Friedman, Jerome H. "Greedy function approximation: A gradient boosting machine." Annals of statistics (2001): 1189-1232

PDP EXAMPLE


Number of Previous Accounts
Age
Savings
Credit Card Application Model


application acceptance
0.1
0.25
0.5
0.75
1
10
20
30
40
50
Savings


application acceptance
0.1
0.25
0.5
0.75
1
1
2
3
4
5
Accounts


application acceptance
0.1
0.25
0.5
0.75
1
18
30
43
55
68
Age

PDP-BASED FEATURE IMPORTANCE

Intuition: the more the PDP varies the more important the feature is [1]

Formulation: How to measure flatness/variability?

Sample standard deviation for continuous features and the range divided by four for categorical ones

$$I(x_i) = \begin{cases} \sqrt{\frac{1}{k-1} \sum_{k=1}^{K} \left(\bar{f}_i \left(x_i^{(k)} \right) - \frac{1}{k} \sum_{k=1}^{K} \bar{f}_i \left(x_i^{(k)} \right) \right)^2} & x_i \text{ is continous} \\ \left(max_k \left(\bar{f}_i \left(x_i^{(k)} \right) \right) - min_k \left(\bar{f}_i \left(x_i^{(k)} \right) \right) \right) / 4 & x_i \text{ is categorical} \end{cases}$$

[1] Greenwell, Brandon M., Bradley C. Boehmke, and Andrew J. McCarthy. "A simple and effective model-based variable importance measure." arXiv preprint arXiv:1805.04755 (2018)

PDP SHORTCOMINGS

- Interactable for high dimensional data
- Does not factor in feature interactions
- It is defined over unique values of features, regardless of their frequency
- Any alternatives? Look up Accumulated Local Effects (ALE) [1] plots and Permutation Feature Importance [2]

[1] Apley, Daniel W., and Jingyu Zhu. "Visualizing the effects of predictor variables in black box supervised learning models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82.4 (2020): 1059-1086

[2] Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously." http://arxiv.org/abs/1801.01489 (2018).

QUESTIONS?

Let's have a little break...