Trends in Internal Auditing and Corporate Governance

Part 6: Explainable Artificial Intelligence (XAI)

SCAN ME

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

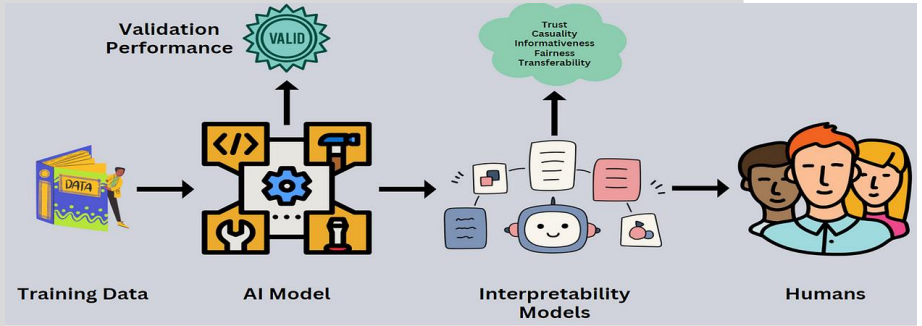
CENTER FOR INTERNAL AUDITING EXCELLENCE
UNIVERSITÄT DUISBURG-ESSEN
LEHRSTUHL FÜR INTERNE REVISION

Univ. Prof. Dr. Marc Eulerich | Lehrstuhl für Interne Revision | 18. Juli 2025

1. **Understand the concept and importance of Explainable AI (XAI)**
2. **Learn about the different XAI methods**
3. **Explore the relevance of XAI for Internal Audit**
4. **Discuss real-world examples of XAI applications in audit contexts**
5. **Apply XAI concepts in a short case study exercise**

Motivation and Relevance

- AI systems are increasingly used across different businesses and industries
- Internal Audit functions must understand and evaluate these systems
- Trust & Transparency: Black-box models lack transparency -> Accountability and trust are at stake
- Regulatory pressure mandates explainability (e.g., EU AI Act, GDPR)
- Improved Stakeholder Communication
- Root Cause Analysis



Paper Discussion

Explainable Artificial Intelligence (XAI) in auditing (Zhang et al.; 2022)

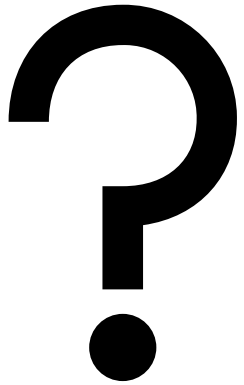
International Journal of Accounting Information Systems

https://www.sciencedirect.com/science/article/pii/S1467089522000240?via%3Dihub

Why this paper?

- Comprehensive overview, starting point
- Covers XAI methods like LIME, SHAP, Counterfactuals
- Practical examples and theoretical insights

What is Explainable AI?



Definition of Explainable AI (XAI)

Methods that enable users to understand and trust AI model outputs (Gunning; 2017)

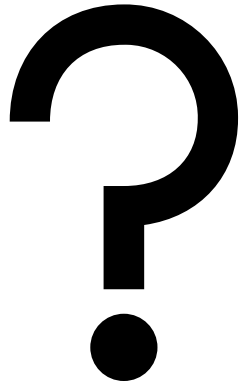
Key Component:

Interpretability: How to make sense of model decisions

- “Interpretability is the degree to which a human can understand the cause of a decision.” (Biran and Cotton; 2017)
- “A method is interpretable if a user can correctly and efficiently predict the method’s results.” (Kim et al.; 2016)
- “Interpretability is about mapping an abstract concept from the models into an understandable form.” (Roscher et al.; 2020)

Importance of Interpretability

If a machine learning model performs well, **why do we not just trust the model** and ignore **why** it made a certain decision?



Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Why are we so focused on understanding AI when we don't even fully understand humans?

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

What about understanding people and their decisions?

Source: https://www.whitehouse.gov/administration/donald-j-trump/

Source: https://de.wikipedia.org/wiki/Elon_Musk

Why do we apply different standards (e.g., whether AI in medicine, such as in tumor classification or surgery, or in self-driving cars, makes error-free decisions)? Do we have higher expectations for AI than for humans?

Shouldn't it simply be tested whether the AI, on average, performs better than the average human? Why do we have higher expectations?

Reasons why we want to understand AI:

- Humans have always had the desire to understand things
- Control and monitoring
- Trust
- Responsibility (legal and ethical accountability)
- Greater acceptance of human errors than machine errors
- Fear of the unknown
- Curiosity
- A mirror of humanity
- Etc.

To achieve this, various tools are used, e.g. statistical models and calculations.

But what are possible approaches to understanding human decisions?

Peer reviews, verification of sources, argumentation & justifications, etc.

Importance of Interpretability

The need for interpretability arises from an incompleteness in problem formalization (Doshi-Velez and Kim 2017), which means that for certain problems or tasks it is not enough to get the prediction (the what). The model must also explain how it arrived at the prediction (the why), because a correct prediction only partially solves your original problem. The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks.” (Doshi-Velez and Kim 2017)

- Imagine a self-driving car automatically detects cyclists based on a deep learning system. You want to be 100% sure that the abstraction the system has learned is error-free because running over cyclists is very bad.
- How effective some drug will be for a patient.
- Advertising follows me on the Internet because I recently bought a washing machine, and I know that for the next days I’ll be followed by advertisements for washing machines.
- A wolf versus dog classifier relied on snow in the background instead of image regions that showed the animals (Ribeiro, Singh, and Guestrin 2016).

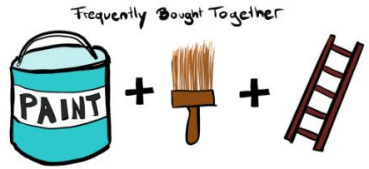


Figure 2.1: Illustration of recommended products that are frequently bought together.

Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

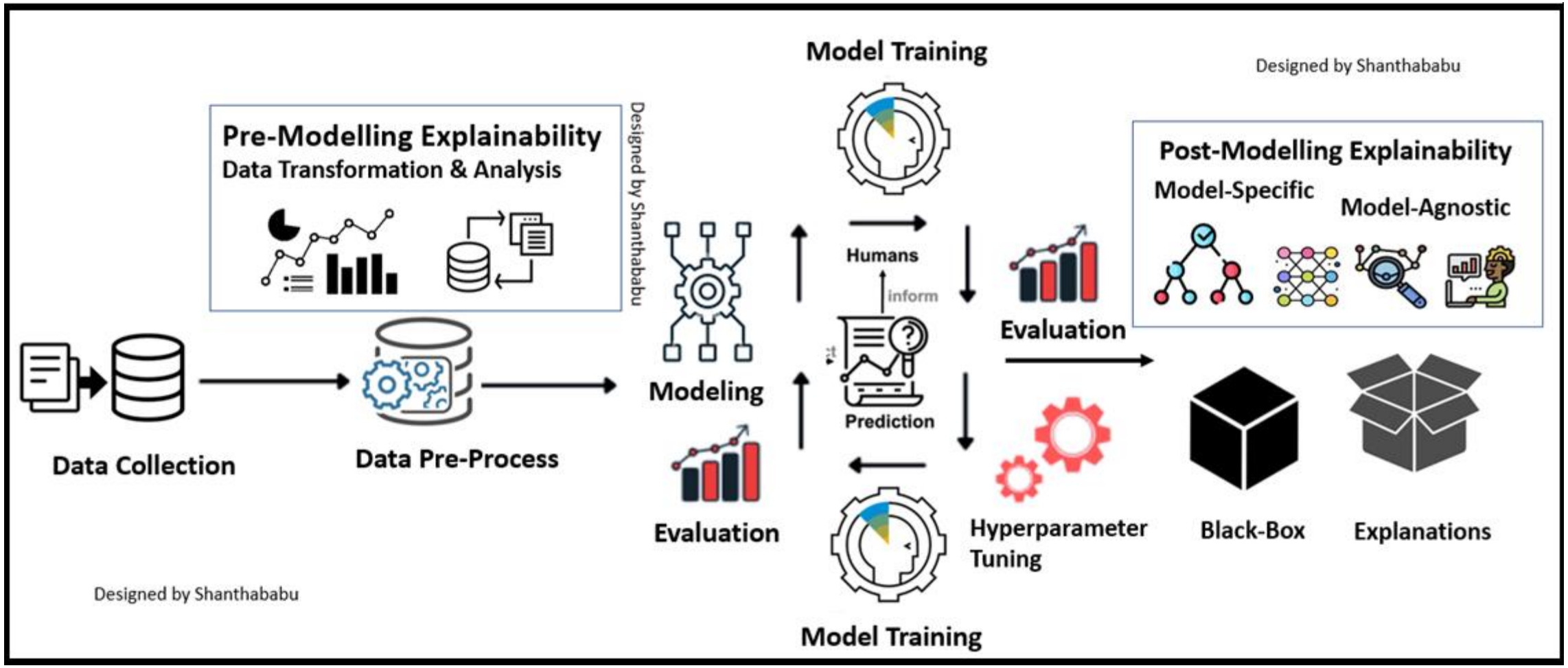
Importance of Interpretability

If you can ensure that the machine learning model can explain decisions, you can also check the following traits more easily (Doshi-Velez and Kim 2017):

- **Fairness**: Ensuring that predictions are unbiased and don't implicitly or explicitly discriminate against underrepresented groups.
- **Privacy**: Ensuring that sensitive information in the data is protected.
- **Reliability or Robustness**: Ensuring that small changes in the input don't lead to large changes in the prediction.
- **Causality**: Ensure that only causal relationships are picked up.
- **Trust**: It's easier for humans to trust a system that explains its decisions compared to a black box.

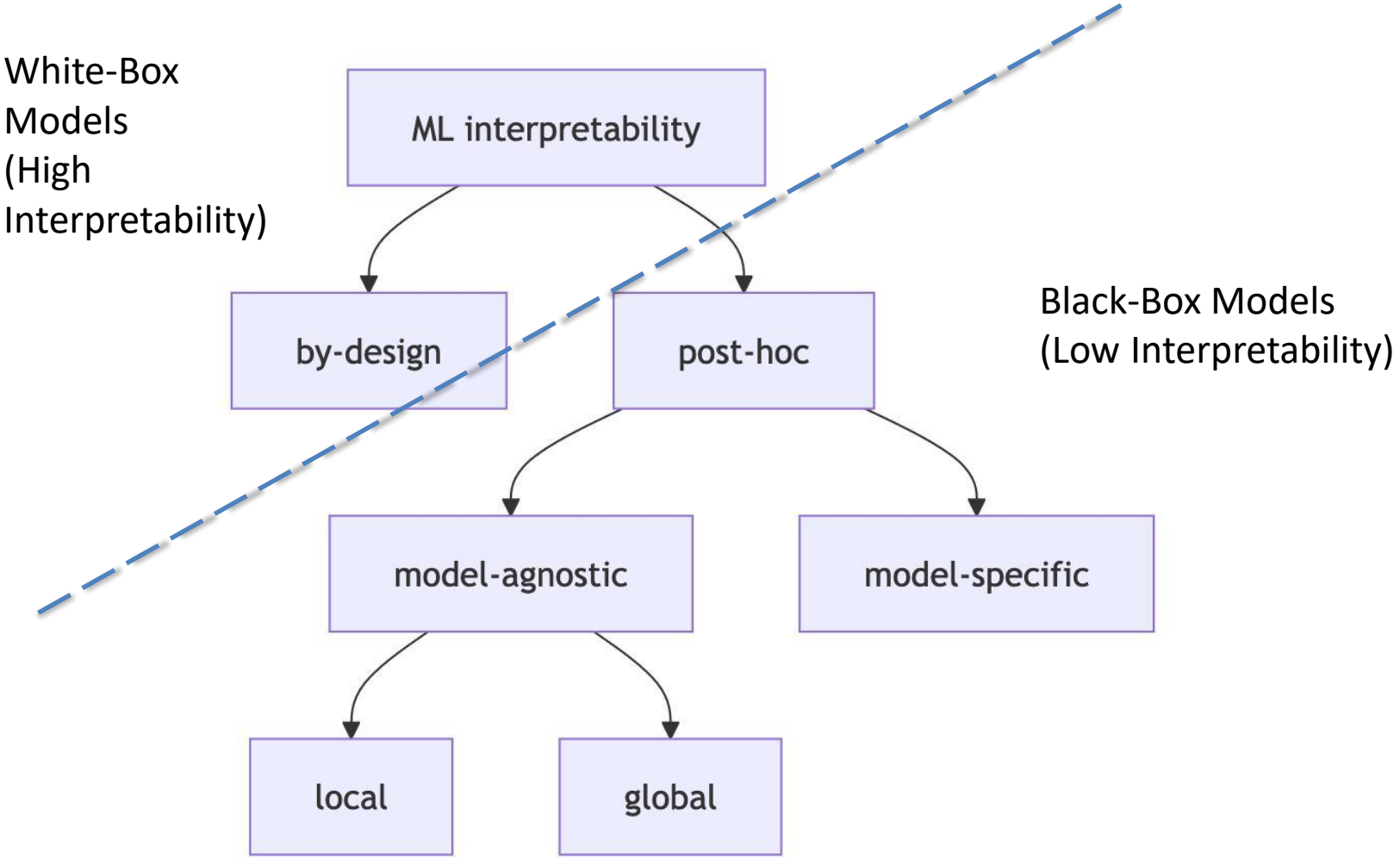
Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Model



Source: Shanthababu, 2023 (https://www.datasciencecentral.com/explainable-artificial-intelligence-xai-for-ai-ml-engineers/)

Methods Overview



Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Interpretable Models by Design

- Linear regression: Fit a linear model by minimizing the sum of squared errors.
- Logistic regression: Extend linear regression for classification using a nonlinear transformation.
- Linear model extensions: Add penalties, interactions, and nonlinear terms for more flexibility.
- Decision trees: Recursively split data to create tree-based models.
- Decision rules: Extract if-then rules from data.
- RuleFit: Combine tree-based rules with Lasso regression to learn sparse rule-based models

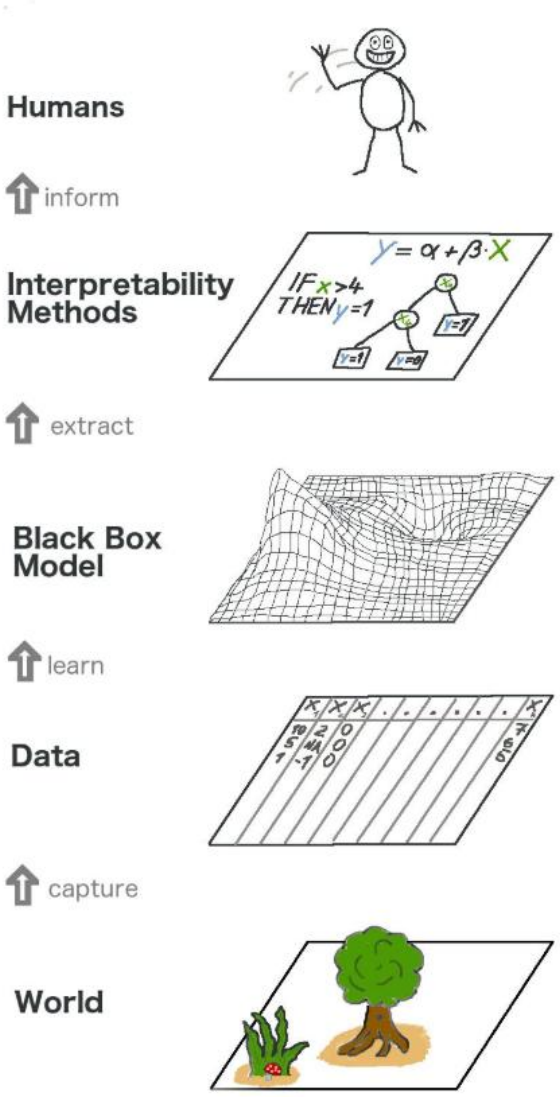
Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

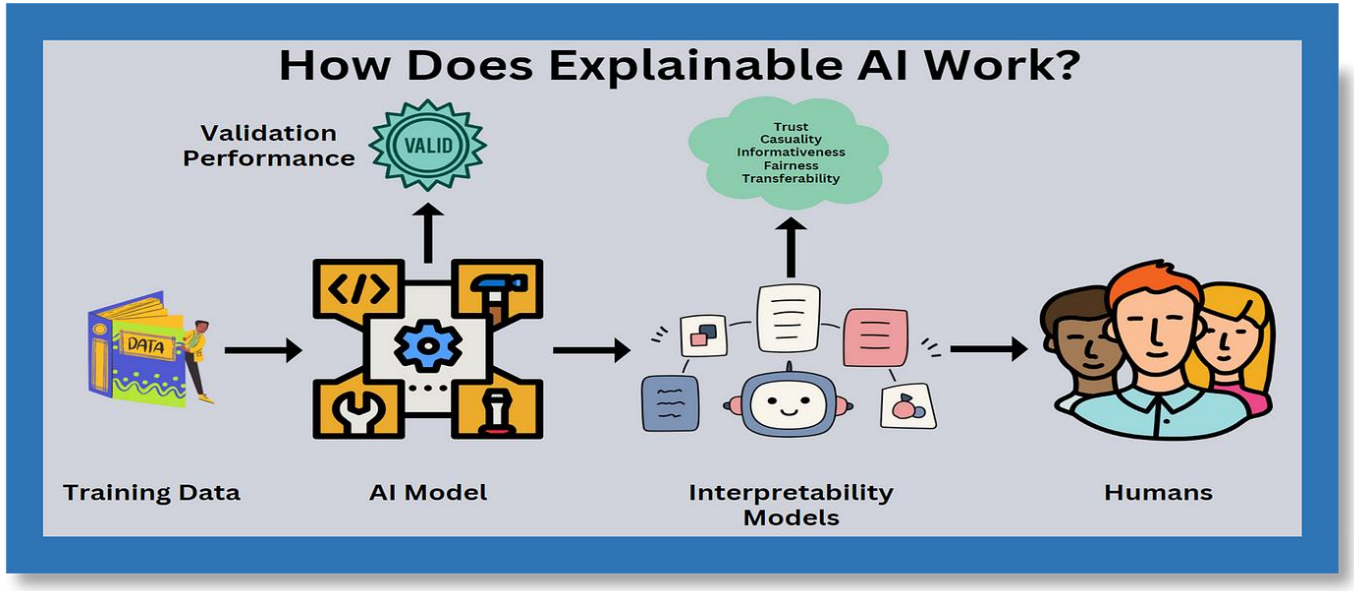
Interpretable Models by Design

- The model is entirely interpretable. Example: a small decision tree can be visualized and understand easily. Or a linear regression model with not too many coefficients. “Entirely interpretable” is a tough requirement, and again a bit fuzzy at the same time. My stance is that the term entirely interpretable may only be used for the simplest of models such as very sparse linear regression or very short trees, if at all.
- Parts of the model are interpretable. While a regression model with hundreds of features may not be “entirely interpretable”, we can still interpret the individual coefficients associated with the features. Or if you have a huge decision list, you can still inspect individual rules.

Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Black-Box Model and How Does Explainable AI Work





Source: Dabass, 2024 (https://python.plainenglish.io/demystifying-explainable-ai-a-beginners-guide-with-examples-37ea8c86ed16)

Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

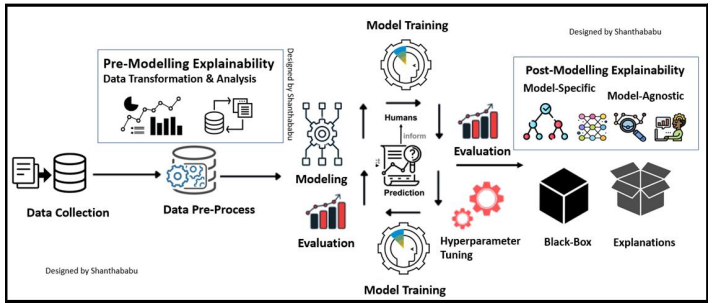
Post-hoc Interpretability

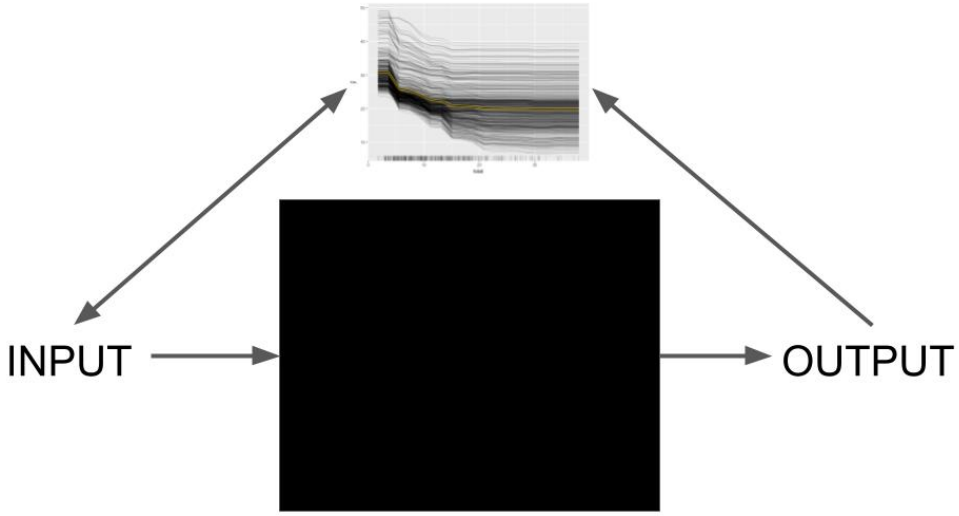
- **Model-agnostic:** We **ignore what's inside** the model and only analyze how the model **output** changes with respect to changes in the feature inputs. For example, permuting a feature and measuring how much the model error increases.
- **Model-specific:** We analyze **parts of the model** to better understand it. This can be analyzing which types of images a neuron in a neural network responds to the most, or the Gini importance in random forests.

Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Model-Agnostic:

- Model-agnostic methods work by the SIPA (sampling, intervention, prediction, aggregation) principle: sample from the data, perform an intervention on the data, get the predictions for the manipulated data, and aggregate the results (Scholbeck et al. 2020). An example is permutation feature importance: We take a data sample, intervene on the data by permuting it, get the model predictions, and compute the model error again and compare it to the original loss (aggregation). What makes these methods model-agnostic is that they don't need to "look inside" the model, like reading out coefficients or weights





Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

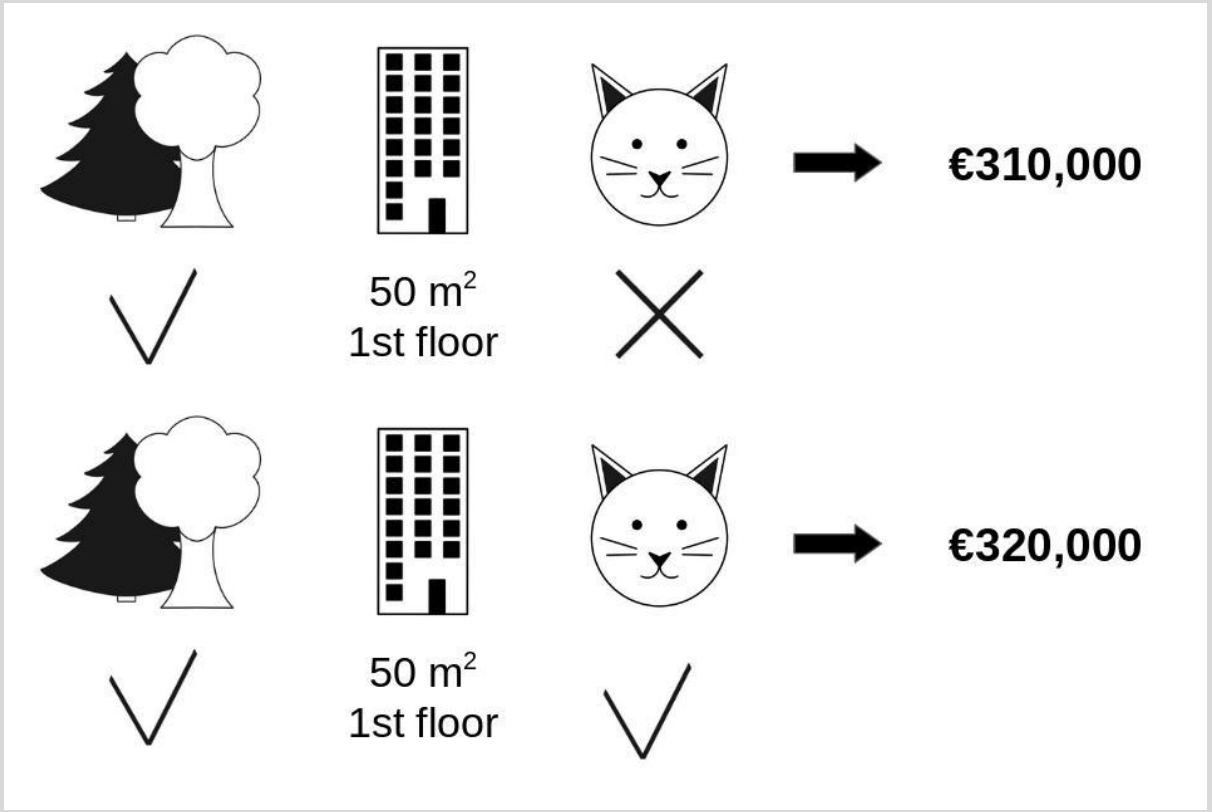
Local Model-Agnostic Post Hoc Methods

Local interpretation methods explain individual predictions.

- Ceteris paribus plots show how changing a feature changes a prediction.
- Individual conditional expectation curves show how changing one feature changes the prediction of multiple data points.
- Local surrogate models (LIME) explain a prediction by replacing the complex model with a locally interpretable model.
- Scoped rules (anchors) are rules that describe which feature values “anchor” a prediction, meaning that no matter how many of the other features you change, the prediction remains fixed.
- Counterfactual explanations explain a prediction by examining which features would need to be changed to achieve a desired prediction.
- **Shapley values** fairly assign the prediction to individual features.
- **SHAP** is a computation method for Shapley values but also suggests global interpretation methods based on combinations of Shapley values across the data.

Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

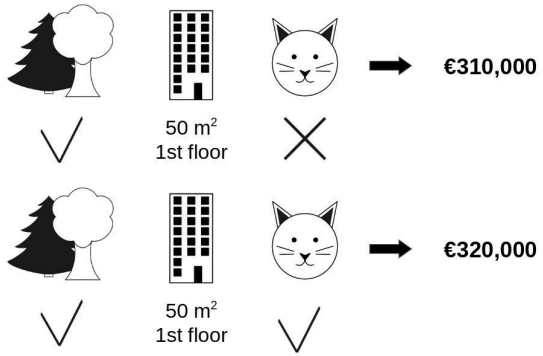
Shapley Values & SHAP (SHapley Additive exPlanations)



- One sample repetition to estimate the contribution of cat-banned to the prediction when added to the coalition of park-nearby and area-50.
- The contribution of cat-banned was €310,000 - €320,000 = -€10,000.

Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Shapley Values & SHAP (SHapley Additive exPlanations)



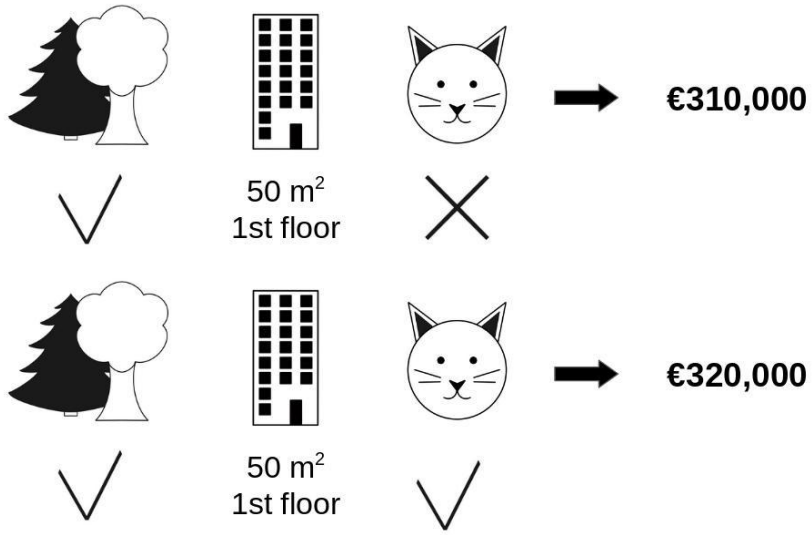
- The interpretation of the Shapley value for feature j is: The value of the j-th feature contributed to the prediction of this particular instance (house-prices) compared to the average prediction for the dataset. The Shapley value works for both classification (if we are dealing with probabilities) and regression.

or in other words:

- We repeat this computation for all possible coalitions. The Shapley value is the average of all the marginal contributions to all possible coalitions. The computation time increases exponentially with the number of features. One solution to keep the computation time manageable is to compute contributions for only a few samples of the possible coalitions.

Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Shapley Values & SHAP (SHapley Additive exPlanations)



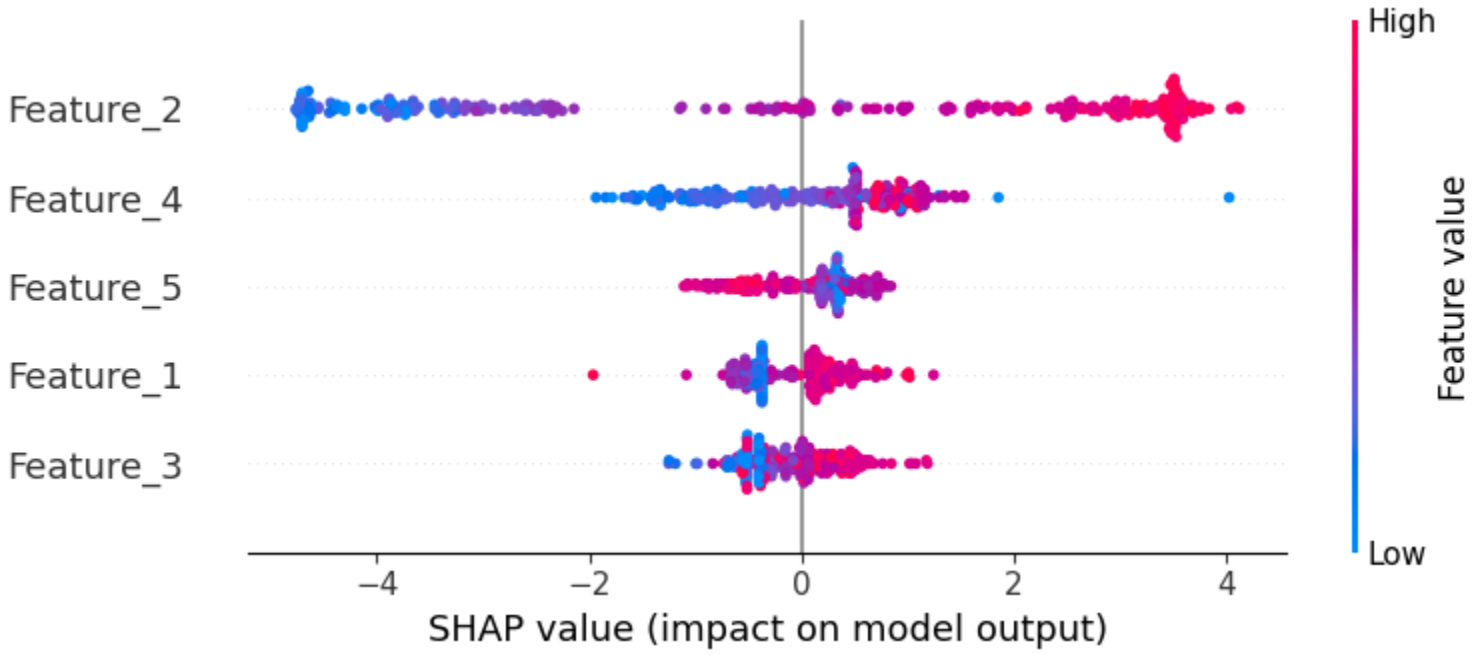
- {} (empty coalition)
- {park-nearby}
- {area-50}
- {floor-2nd}
- {park-nearby,area-50}
- {park-nearby,floor-2nd}
- {area-50,floor-2nd}
- {park-nearby,area-50,floor-2nd}

- For each of these coalitions, we compute the predicted apartment price with and without the feature value cat-banned and take the difference to get the marginal contribution.

Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Shapley Values & SHAP (SHapley Additive exPlanations)

- SHAP Plot result interpretation



Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Shapley Values & SHAP (SHapley Additive exPlanations)

- SHAP Plot result interpretation

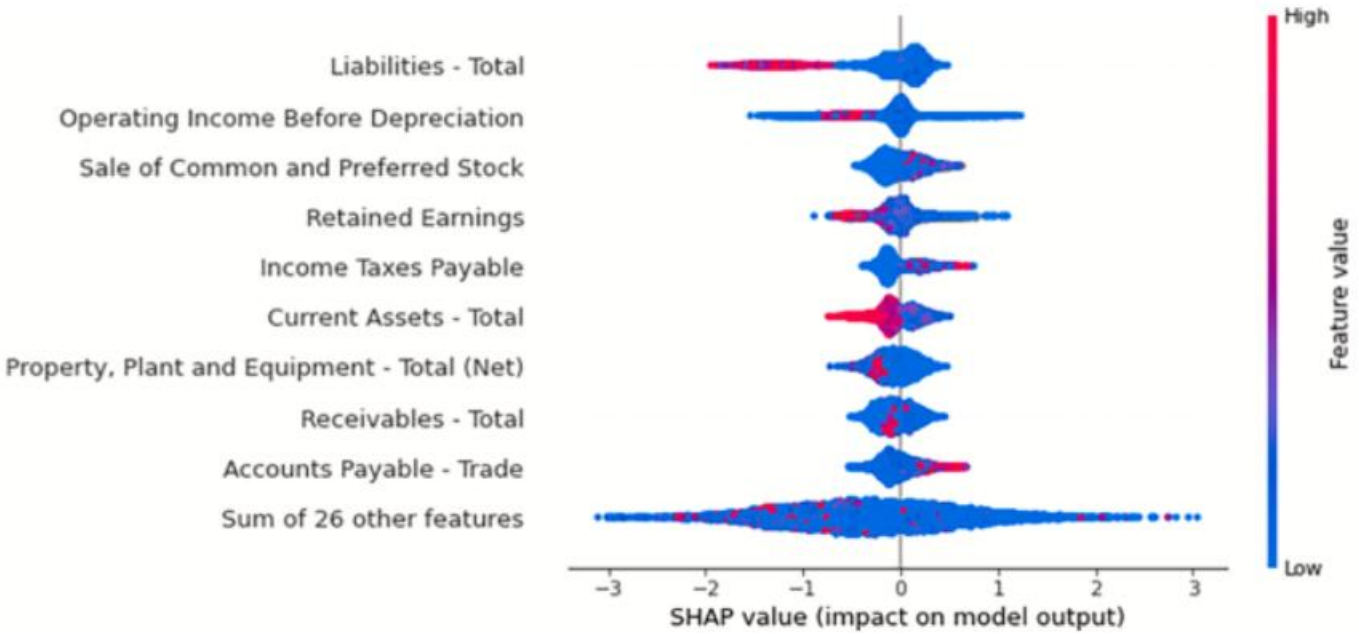


Fig. 8. Scatter Plot of SHAP.

Zhang et al.(2022)

Shapley Values & SHAP (SHapley Additive exPlanations)

There are as well other ways to estimate Shapley values for explaining predictions:

- KernelSHAP
- Permutation Method
- TreeSHAP.

Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Global Model-Agnostic Post-Hoc Methods

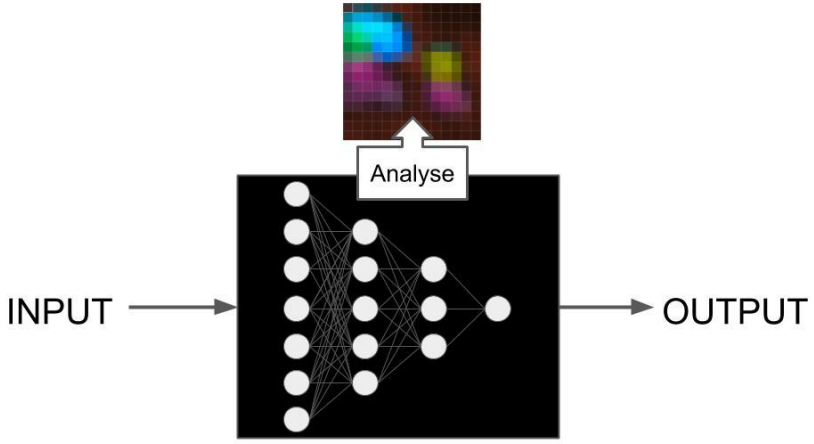
Global methods describe the **average behavior** of a machine learning model across a dataset.

- The partial dependence plot is a feature effect method.
- Accumulated local effect plots also visualize feature effects, designed also for correlated features.
- Feature interaction (H-statistic) quantifies the extent to which the prediction is the result of joint effects of the features.
- Functional decomposition is a central idea of interpretability and a technique for decomposing prediction functions into smaller parts.
- Permutation feature importance measures the importance of a feature as an increase in loss when the feature is permuted.
- Leave one feature out (LOFO) removes a feature and measures the increase in loss after retraining the model without that feature.
- Surrogate models replace the original model with a simpler model for interpretation.
- Prototypes and criticisms are representative data points of a distribution and can be used to improve interpretability.

Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Model-Specific Post-Hoc Methods

- As the name implies, post-hoc model-specific methods are applied after model training but only work for specific machine learning models
- To interpret the behavior and predictions of neural networks, we need specific interpretation methods.



Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

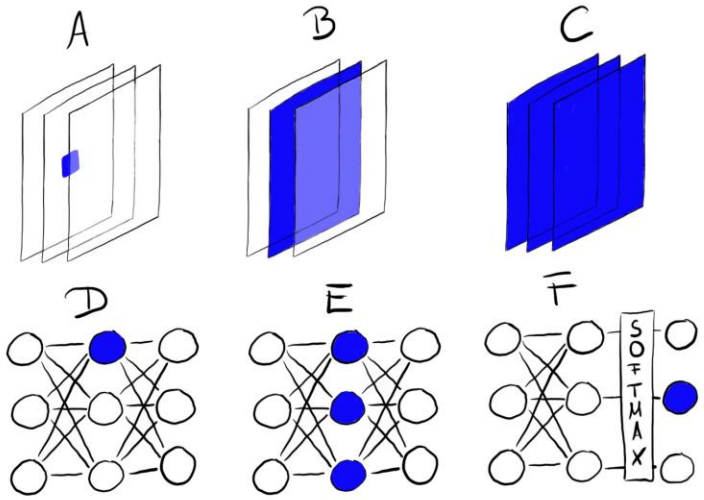
Model-Specific Post-Hoc Methods

- The neural network part covers the following techniques that answer different questions:
- **Learned Features:** What features did the neural network learn?
- Saliency Maps: How did each pixel contribute to a particular prediction?
- **Detecting Concepts**: Which concepts did the neural network learn?
- Adversarial Examples: How can we fool the neural network?
- Influential Instances: How influential was a training data point for a given prediction?

Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Learned Features – Feature Visualization / Attention Visualization

The approach of making the learned features explicit is called Feature Visualization. Feature visualization for a unit of a neural network is done by **finding the input that maximizes the activation of that unit**. “Unit” refers either to individual neurons, channels (also called feature maps), entire layers, or the final class probability in classification (or the corresponding pre-softmax neuron, which is recommended).



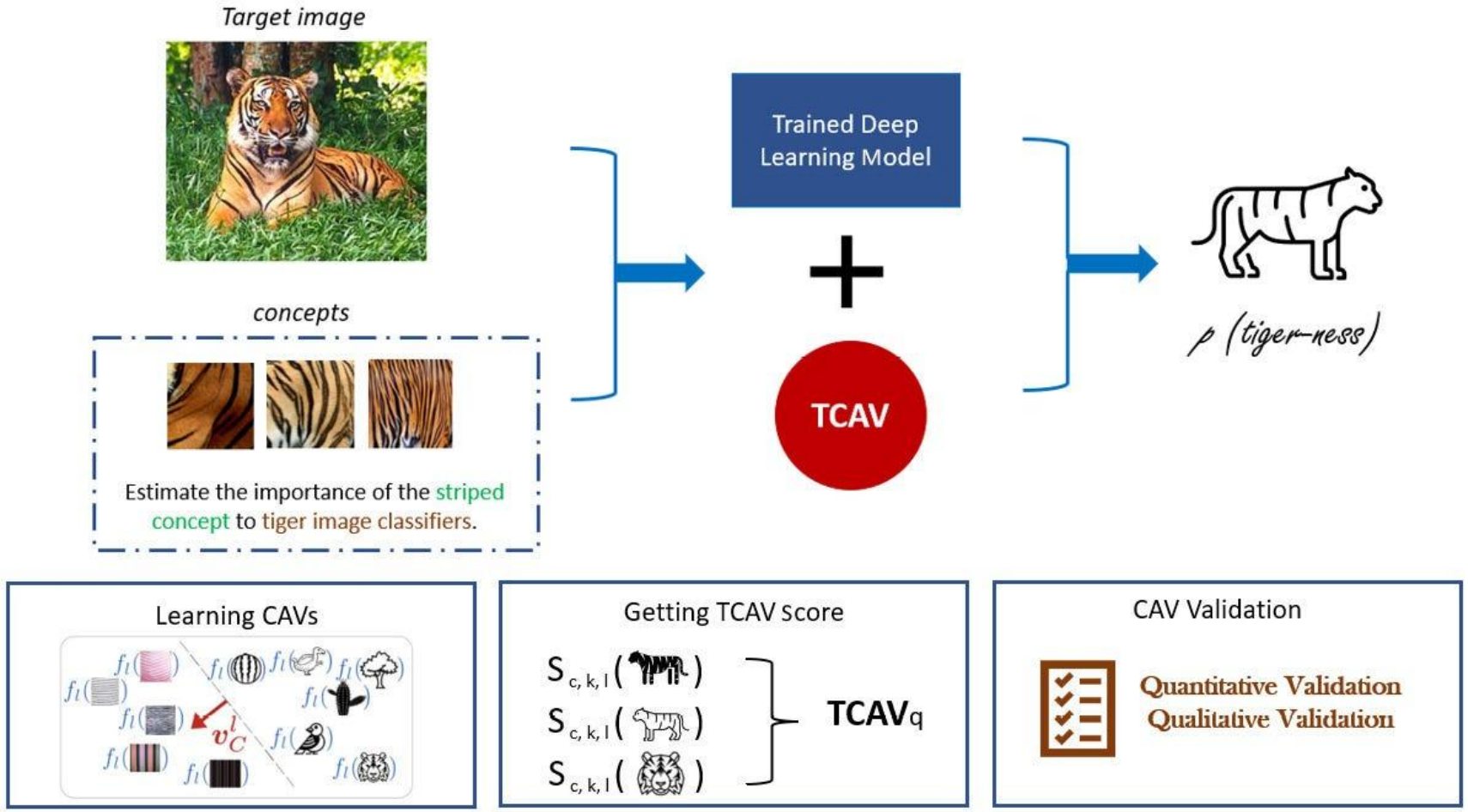
Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Detecting Concept
TCAV: Testing with Concept Activation Vectors

- For example, the importance of a single pixel in an image usually does not convey much meaningful interpretation. Second, the expressiveness of a feature-based explanation is constrained by the number of features.
- The concept-based approach addresses both limitations mentioned above. A concept can be any abstraction, such as a color, an object, or even an idea. In other words, the concept-based approach can generate explanations that are not limited by the feature space of a neural network.
- For any given concept, TCAV measures the extent of that concept's influence on the model's prediction for a certain class. **For example, TCAV can answer questions such as how the concept of "striped" influences a model classifying an image as a "zebra."** Since TCAV describes the relationship between a concept and a class, instead of explaining a single prediction, it provides useful global interpretation for a model's overall behavior

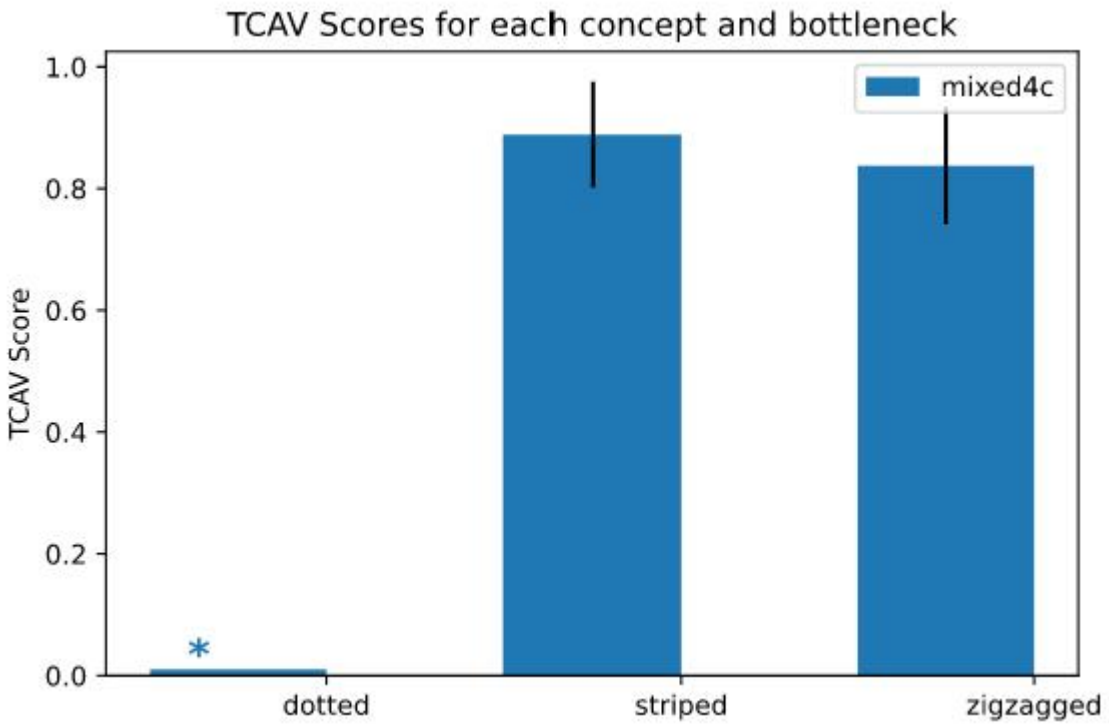
Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

Detecting Concept
TCAV: Testing with Concept Activation Vectors



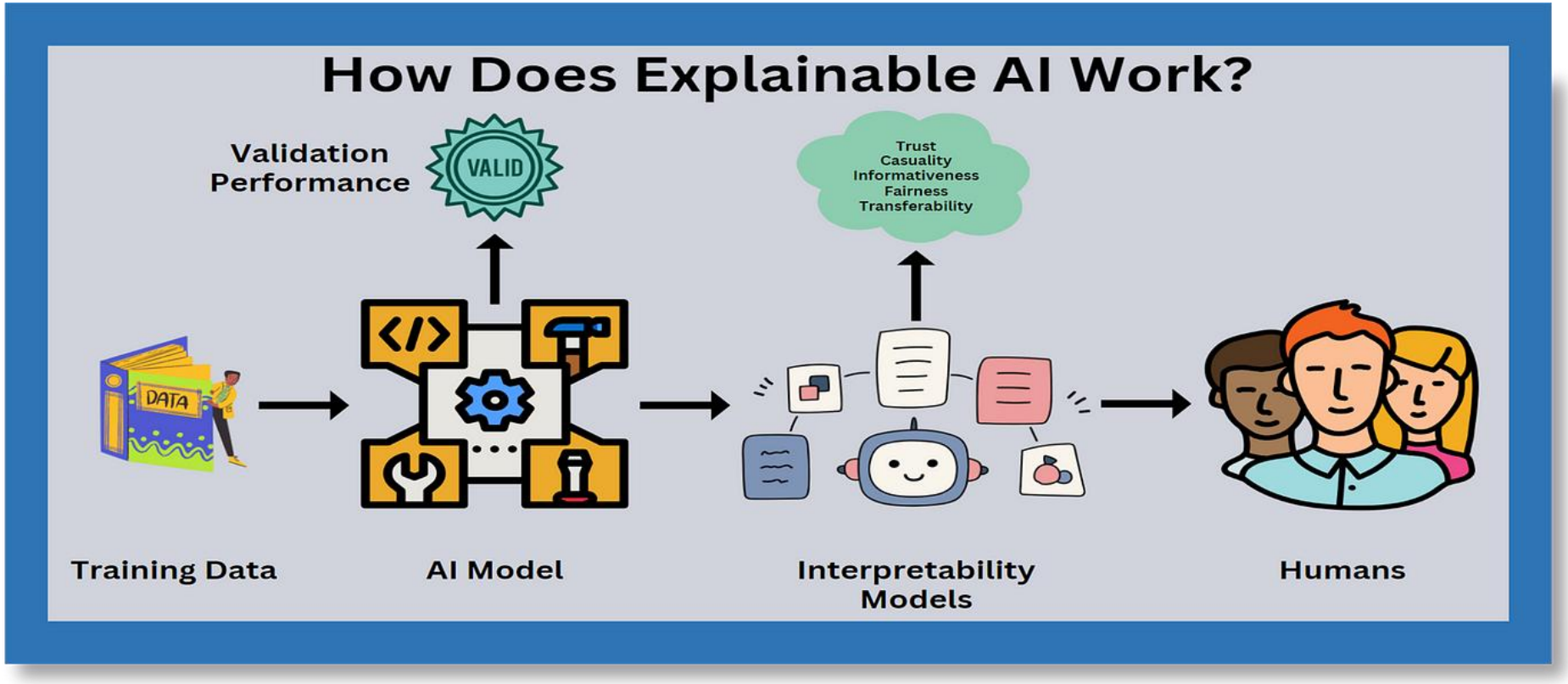
Source: https://medium.com/data-science/explainable-ai-with-tcav-from-google-ai-5408adf905e

Detecting Concept
TCAV: Testing with Concept Activation Vectors



Source: Molnar (Interpretable ML – A Guide for Making Black Box Models Explainable)

What is Missing?



Source: Dabass, 2024 (https://python.plainenglish.io/demystifying-explainable-ai-a-beginners-guide-with-examples-37ea8c86ed16)

Are there other Explainable AI methods that could be used in general and in Internal Audit?

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Human SMEs / Human in the Loop
Expert judgment provides golden source of truth

Statistics SMEs

Using statistical methods on language (e.g. SHAP)

Chain-of-Thought (CoT) Reasoning and Constitutional AI
Shows the user the different steps in the decision-making process

LLM as a Judge
or
RAG as a Judge
or
Prompt Engineering (e.g. EY with EY VIA)

Digital Twin & Simulations

Benchmark Tests (e.g. TruthfulQA, EVALS, OpenXAI)

Model Cards & System Cards
https://openai.com/de-DE/index/gpt-4o-system-card/

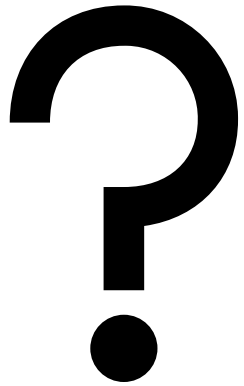
??

Why is Explainable AI Important for Internal Audit?

- Internal Audit functions must understand and evaluate these systems. They need tools to comprehend and assess certain AI applications in order to estimate their risks and impact.

- Or at least, Internal Audit should be able to verify whether appropriate processes and controls (e.g., XAI mechanisms) are in place to properly govern, monitor, and constrain the AI applications.

What are Real-World Examples of XAI application in the Internal Audit Context?



Limitations & Critiques

- XAI may oversimplify complex models
- Explanations can be misleading/complex
- Lack of standardization
- AI System Trade-off: Accuracy vs. Interpretability
- Explainability ≠ Transparency or Traceability: Transparent models are not automatically explainable in a social sense. Example: A model may be mathematically “transparent” but still not explainable to users or legal professionals.
- The use of risk assessment algorithms in the justice system (e.g., to estimate the likelihood of reoffending (Rückfallwahrscheinlichkeit) may be technically explainable. However, if the basis for decision-making is not publicly verifiable or politically legitimized, this creates a democratic explainability gap.
- …

Case Study 1 – LLM as a Judge – Human SMEs / Human in the Loop – 20 Min.

- Company REM: Real estate management (sale and rental of properties)
- The Analytics Department prepares monthly evaluations of various KPIs for management (C-level) and, based on these, provides recommendations. Management then makes the corresponding investment decisions (e.g., purchasing or selling real estate).
- As part of its review of core processes, the Internal Audit function examines particularly critical business processes and intends to conduct a sample-based audit of recent evaluations to assess their quality and accuracy. Internal Audit plans to use the LLM as a Judge methodology, supported by human subject-matter experts (SMEs) / a Human-in-the-Loop approach.
- Tasks:
  - Based on the file “house_prices.xlsx” (scan the QR code), create possible analyses (e.g., a histogram of house prices, a map or bar chart of the most expensive locations, proportion of vacant houses and vacancy costs, etc.). => Step of the Analytics Department
  - Manipulate certain data in the file “house_prices.xlsx” (scan the QR code) and take note of the changes. => Step of the Analytics Department
  - Use a (different) LLM model to check the quality and accuracy of the analyses and data. (Note: the focus should be on quantitative validation.) => Step of the Internal Audit Department

Case Study 1 – LLM as a Judge – Human SMEs / Human in the Loop – 20 Min.

https://gigamove.rwth-aachen.de/de/download/1b99e6da3f2eb0c2bd82d1c90fadeafe

Case Study 1 – LLM as a Judge – Human SMEs / Human in the Loop – 20 Min.

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Questions:

- What were the results of your analysis?
- **How well did LLM as a Judge perform?**
- **What worked well and why?**
- **What didn’t work well and why not?**
- Are there any limitations?
- Were additional methods used for validation?
- If so, which ones?
- **Can Internal Audit rely on the results of LLM as a Judge? Or what would need to be added or improved?**

Case Study 2 – SHAP – 20 Min.

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Company REM: Real estate management (sale and rental of properties)
- The Analytics Department produces monthly analyses and believes that the feature “sunny side” is the most important factor in determining house prices, and therefore has the strongest influence on the output. Based on this, the department recommended to management that a **new strategy** should be pursued: **to invest exclusively in houses with a sunny side.**
- The Internal Audit Department was commissioned by the CEO to independently review this recommendation. To do so, Internal Audit would like to apply the SHAP methodology.
- Tasks:
  - Create a SHAP plot based on the file “house_prices.xlsx” (scan the QR code) to validate the key drivers of house prices.

Case Study 2 – SHAP – 20 Min.

https://gigamove.rwth-aachen.de/de/download/1b99e6da3f2eb0c2bd82d1c90fadeafe

Case Study 2 – SHAP – 20 Min.

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Questions
  - **What is your conclusion?**
  - **What would you recommend to the Analytics Department and Management from the perspective of the Internal Audit function?**
  - Would you accept this model as an auditor?
  - Are the model's decisions explainable and trustworthy?
  - **Could this model also be used for LLMs? Why or why not? What might be possible alternatives?**

Summary and Takeaways

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Students developed a basic working knowledge of XAI tools.
- The concept and importance of XAI were understood and reflected upon.
- Participants familiarized themselves with various XAI methods.
- The relevance of XAI for Internal Audit was explored and evaluated.
- The trade-offs and limitations of different XAI approaches were discussed.
- Real-world examples of XAI applications in audit contexts were analyzed.
- XAI concepts were applied in a short case study exercise.

Advertising – Volunteers?

References

- Biran, Or, and Courtenay V. Cotton. 2017. “Explanation and Justification in Machine Learning: A Survey.” In Proceedings of the IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI). https://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf.
- Dabass, R. 2024. “Demystifying Explainable AI: A Beginner’s Guide with Examples.” Plain English - Python. https://python.plainenglish.io/demystifying-explainable-ai-a-beginners-guide-with-examples-37ea8c86ed16.
- Doshi-Velez, Finale, and Been Kim. 2017. “Towards a Rigorous Science of Interpretable Machine Learning.” arXiv Preprint arXiv:1702.08608.
- Gunning, D. 2017. “Explainable Artificial Intelligence (XAI).” DARPA/I2O Program Information. https://www.darpa.mil/program/explainable-artificial-intelligence.
- Google AI. 2018. “Explainable AI with TCAV from Google AI.” Medium – Towards Data Science. https://medium.com/data-science/explainable-ai-with-tcav-from-google-ai-5408adf905e.

References

- Kim, Been, Rajiv Khanna, and Oluwasanmi Koyejo. 2016. “Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability.” In Proceedings of the 30th International Conference on Neural Information Processing Systems, 2288–96. NIPS’16. Red Hook, NY, USA: Curran Associates Inc.
- Molnar, Christoph. 2022. Interpretable Machine Learning – A Guide for Making Black Box Models Explainable. Self-published.
- PwC Switzerland. 2025. “Explainable AI in Internal Audit – Managing Trust in Black Box Systems.” Internal Whitepaper. Zürich: PwC Schweiz.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016a. “Model-Agnostic Interpretability of Machine Learning.” arXiv Preprint arXiv:1606.05386.
- Roscher, Ribana, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. 2020. “Explainable Machine Learning for Scientific Insights and Discoveries.” IEEE Access 8: 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199.

Kontakt

Lehrstuhl für Interne Revision

E-Mail: hiwi.ircg@uni-due.de

Universität Duisburg-Essen
Mercator School of Management
Lehrstuhl für Interne Revision
Lotharstraße 65
47057 Duisburg, Deutschland

www.ircg.msm.uni-due.de

www.internalauditing.de

SSRN Working Papers at SSRN.com

Folgen Sie dem Lehrstuhl auf Lehrstuhl für Interne Revision