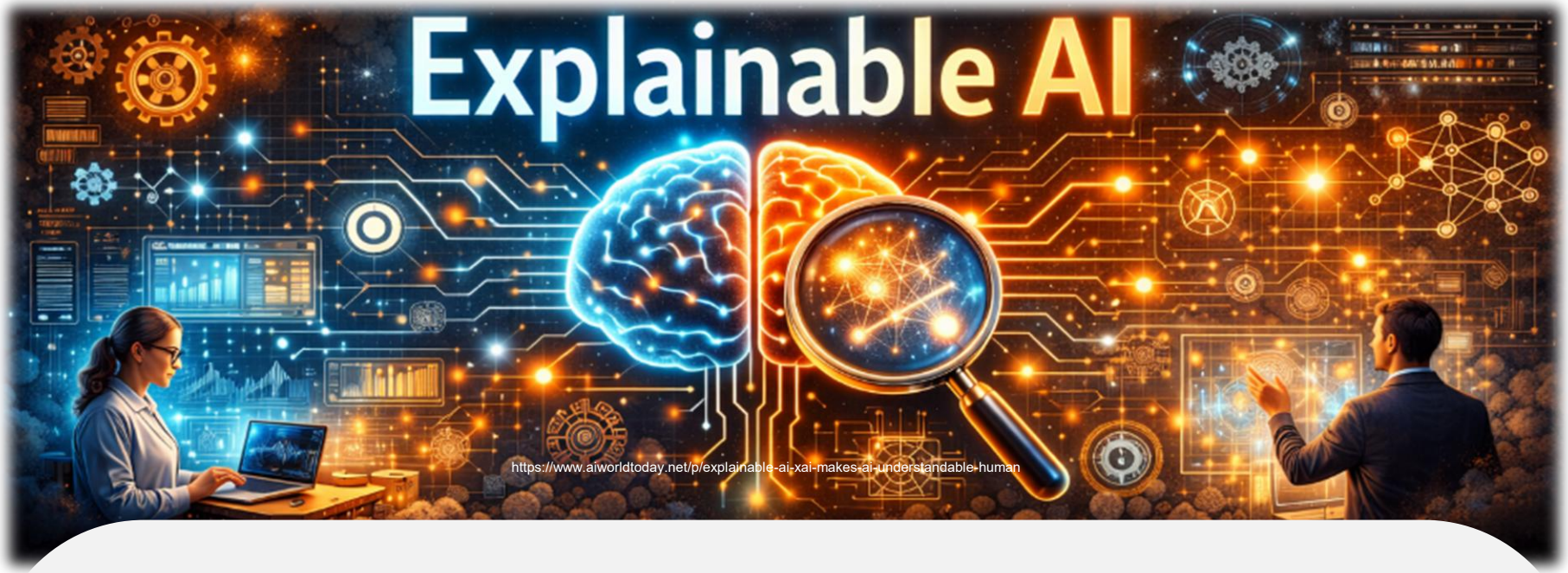
AI Factory Austria AI:AT



Introduction to Explainable AI (XAI)

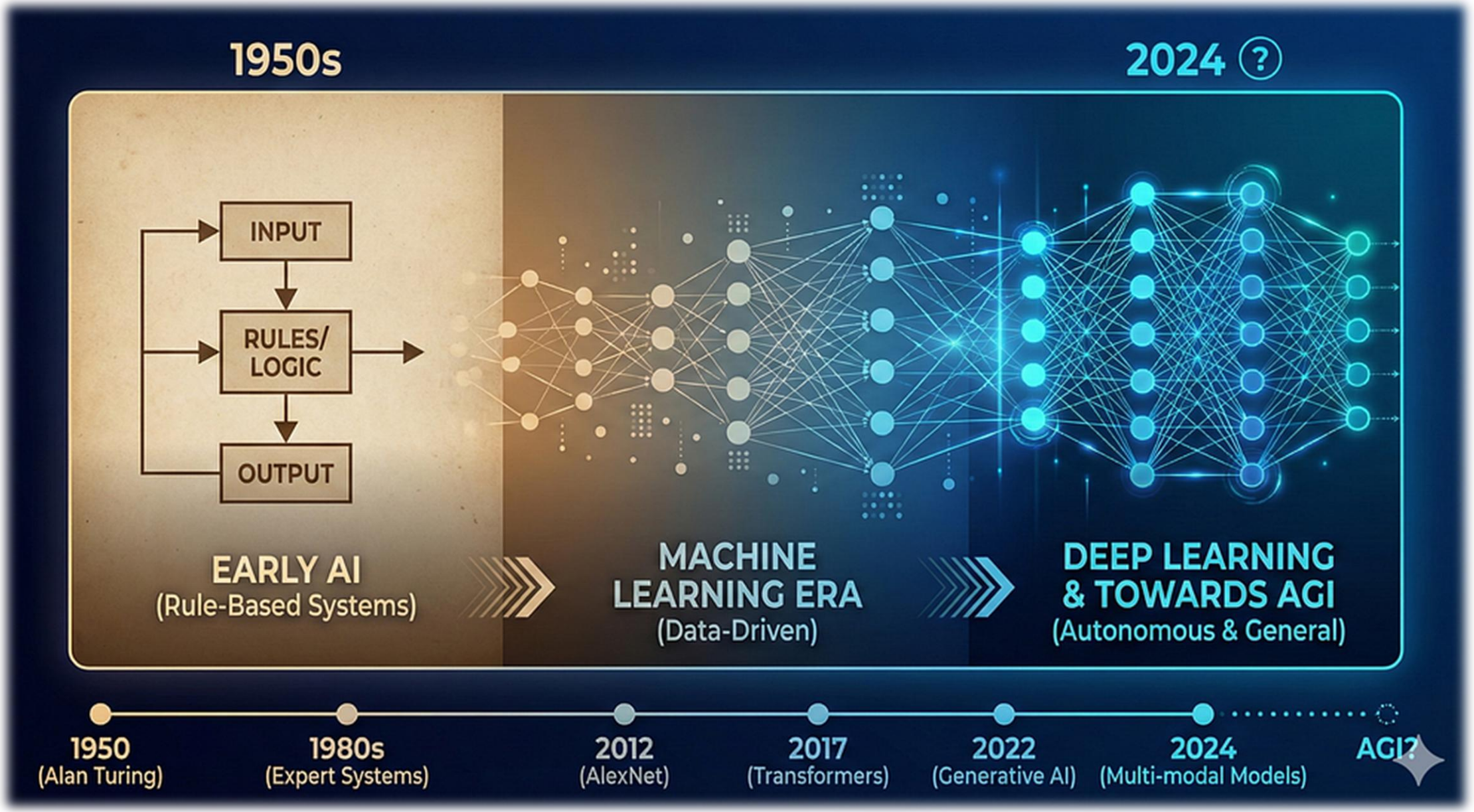
Motivation, Concepts & Challenges

Anahid Wachsenegger
AIT Austrian Institute of Technology GmbH

12.01.2026

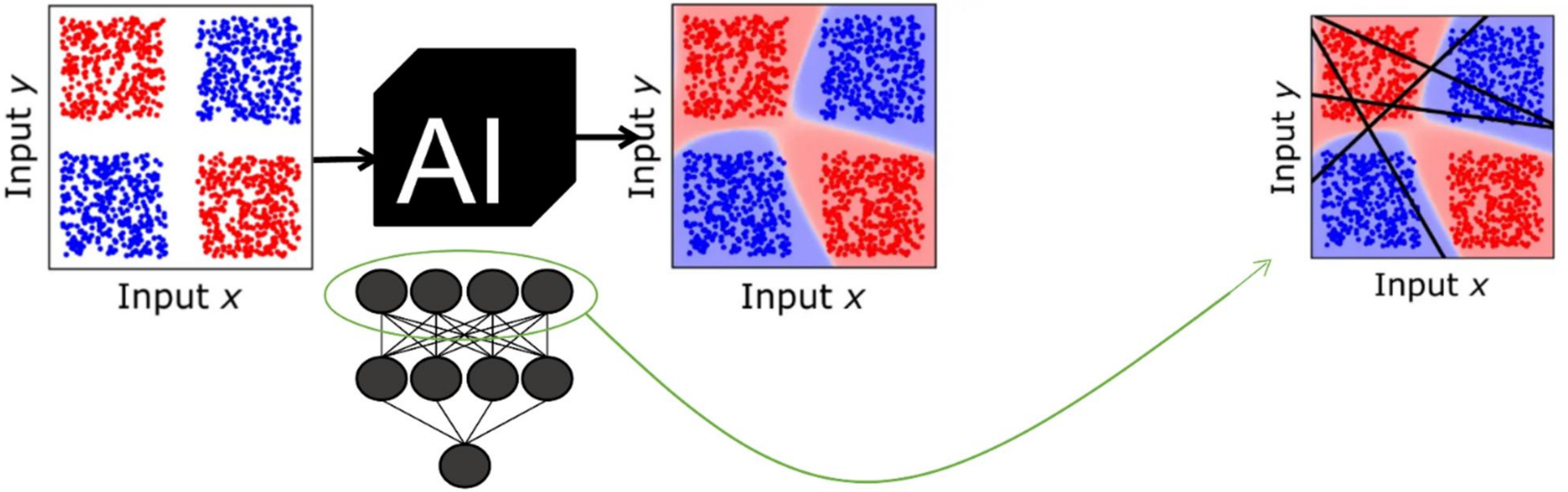
If AI makes decisions … Who explains them?

XAI | Black-Box, Opaque, Complex models



https://surendrallam.medium.com/the-evolution-of-ai-a-journey-from-machine-learning-to-agi-b5c23ebfea26

XAI | Motivation



https://towardsdatascience.com/why-we-will-never-open-deep-learnings-black-box-4c27cd335118#:~:text=What%20is%20the%20black%20box,is%20doing%20on%20its%20own

Let‘s take a step back

How do humans explain?

Dimensions of explanations

WHO

(**Who** is the explanation for?)

- Domain expert
- Decision subject (end user)
- Auditor / regulator
- Developer / data scientist

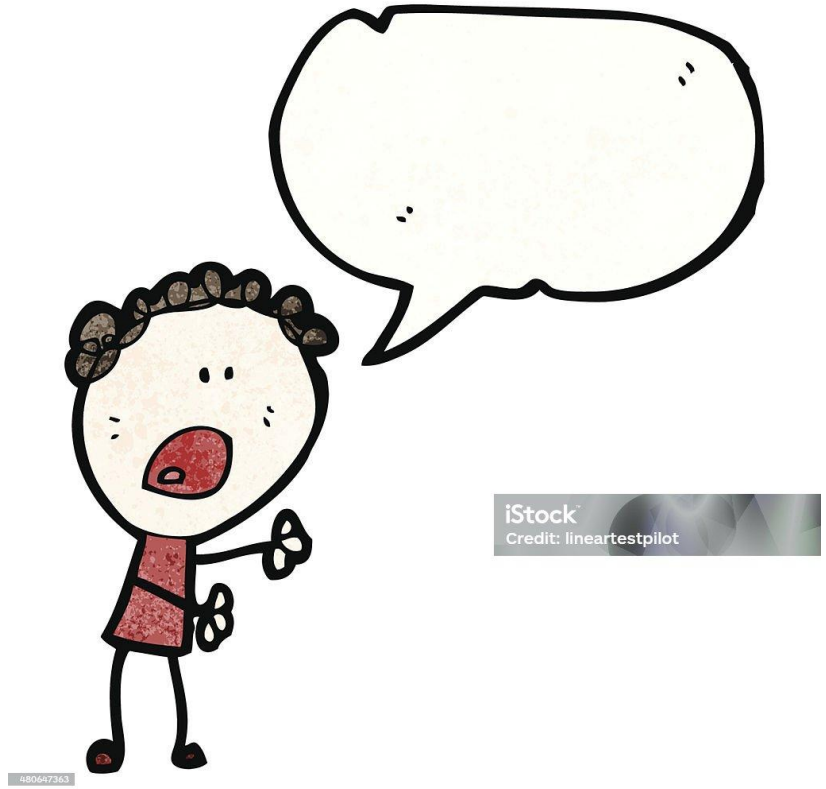
WHAT

(**What** Is Being Explained?)

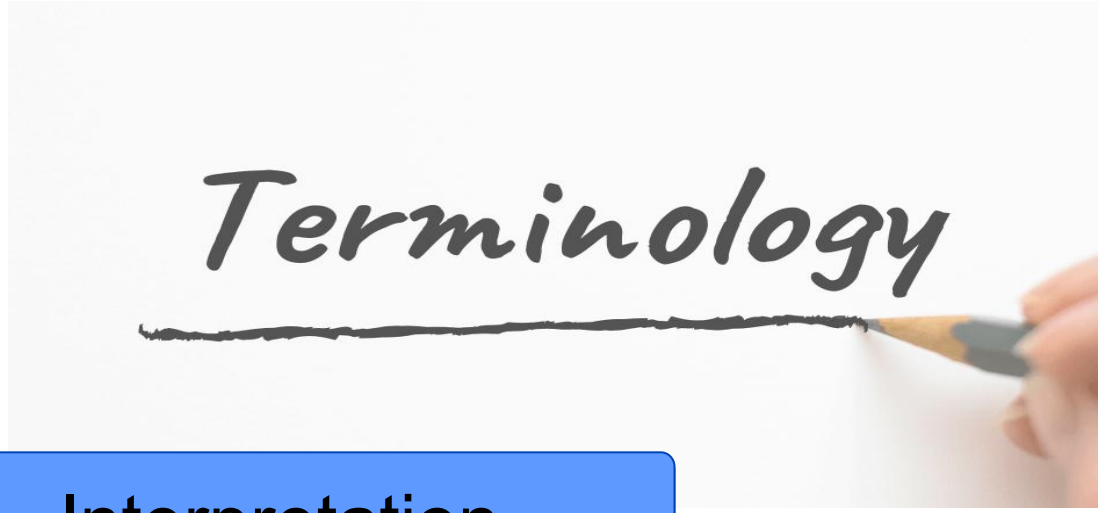
- Decision / Outcome explanation
- Model transparency explanation
- Scope of the explanation
  - Global (behavior)
  - Local (instance-based)

How do humans explain?

- A good explanation is Coherent
  - Parts of the explanation fit together
  - They are compatible with the existing beliefs and are consistent with the evidence
- A good explanation is Complete
  - No Gap is in the explanation
- A good explanation is Articulate
  - Preference for complex explanations (multiple causal paths; explanation length)
- Good explanation has Alternatives
  - Explanations might shift our mental model and generate more questions
  - -> no single explanation is always the answer



XAI | Interpretable or Explainable?



Explanation

”Statement, fact, or situation that tells you why something happened; a reason given for something”

Interpretation

“the particular way in which something is understood or explained”

XAI | Interpretable or Explainable?

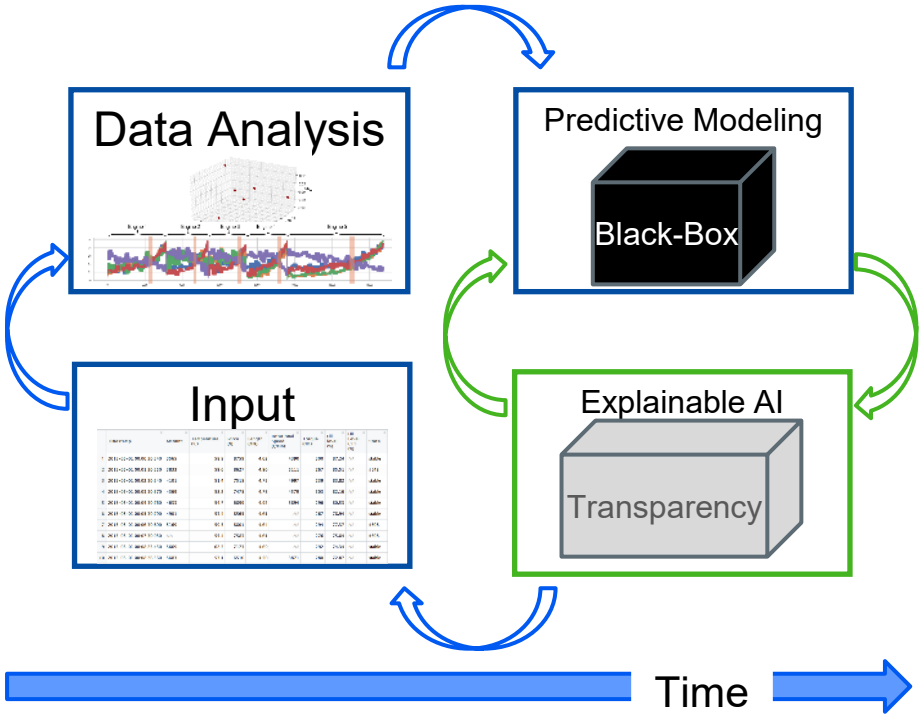
Explainability

”... the extent where the feature values of an instance are related to its model prediction in such a way that humans understand.”

Interpretability

“... defined as the amount of consistently predicting a model’s result without trying to know the reasons behind the scene.”

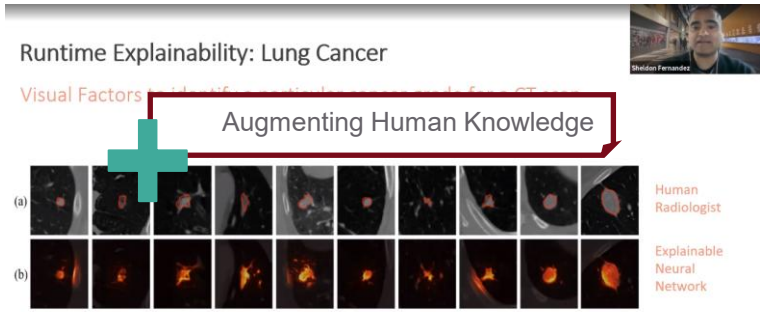
XAI | Machine Learning Workflow



Ribeiro et. Al., " Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD*, 2016.

XAI | Machine Learning Workflow

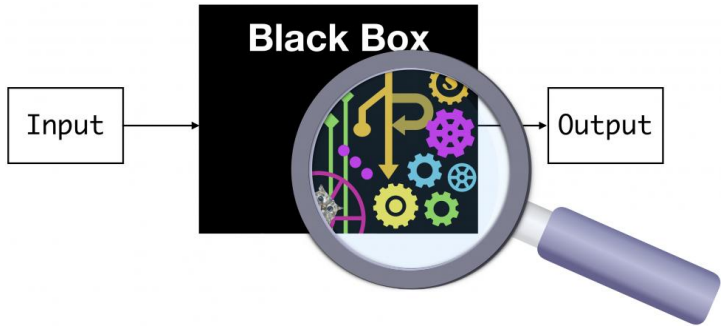
- ✓ Curiosity / learning
- ✓ Understanding the model's success and failing
- ✓ Is the model/data biased?
- ✓ How can we increase trust and acceptance of using such systems?
- ✓ How can I improve my model's performance?
- ✓ Is my model fair? Is it safe (privacy)? Is it reliable (robust)?
- ✓ Correlation vs Causality



Ribeiro et. Al., " Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD*, 2016.
GTC 2021 explainable-ML talks

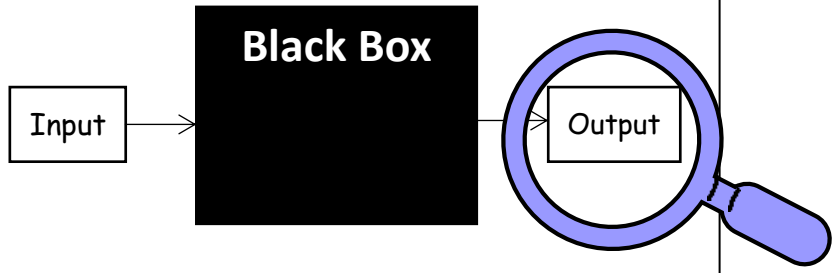
XAI | Specific vs Agnostic?

Model-Specific



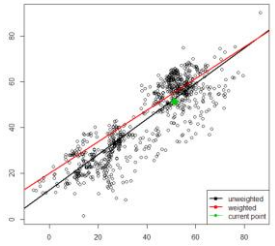
Intrinsically Explainable (transparency)

Model-Agnostic

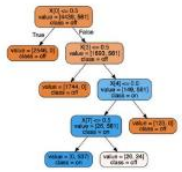


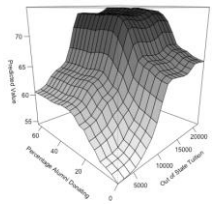
Post-Hoc Explainability

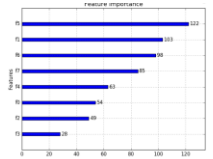
Model internals (e.g. Weights)



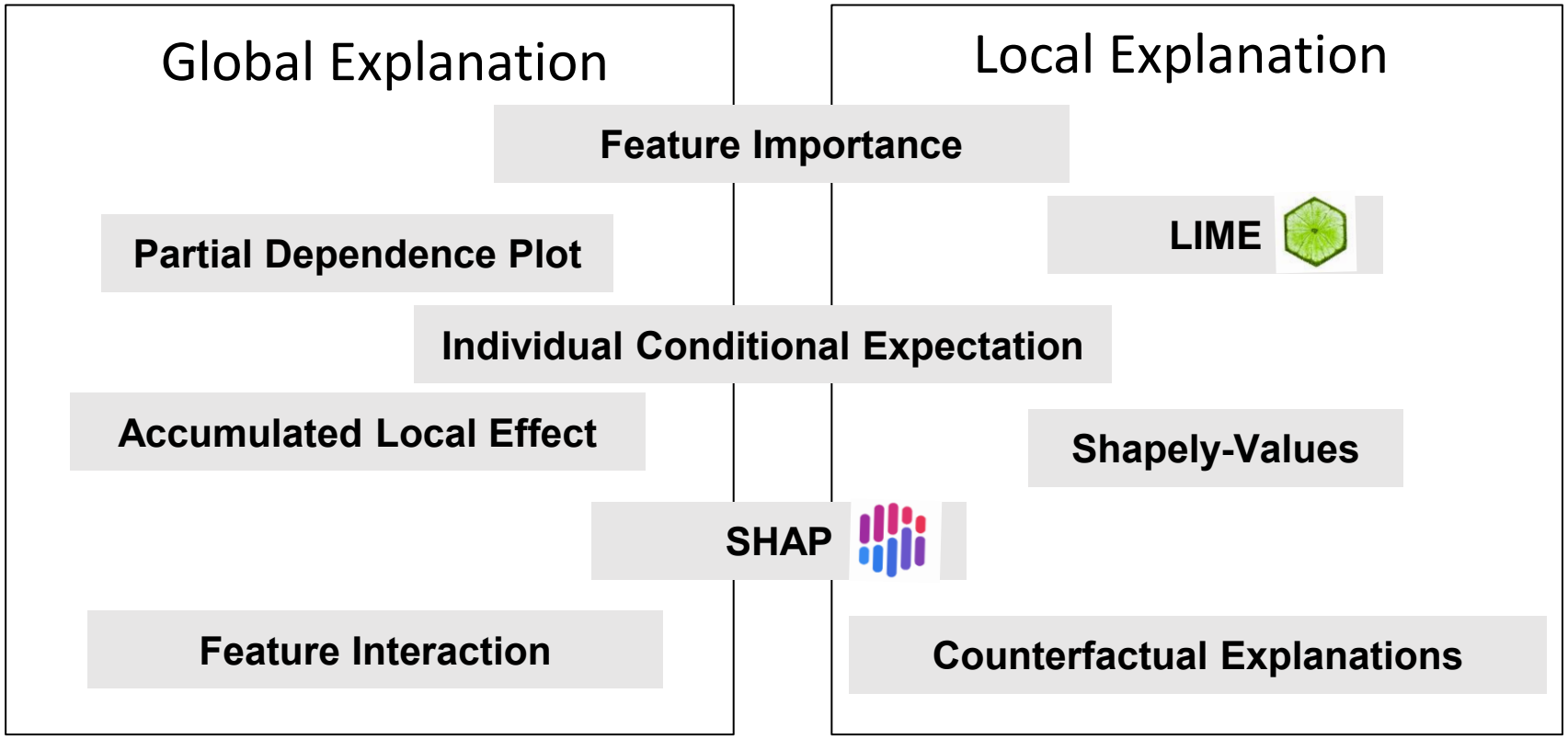
- Feature summary
- Surrogate models







XAI | Local vs Global?

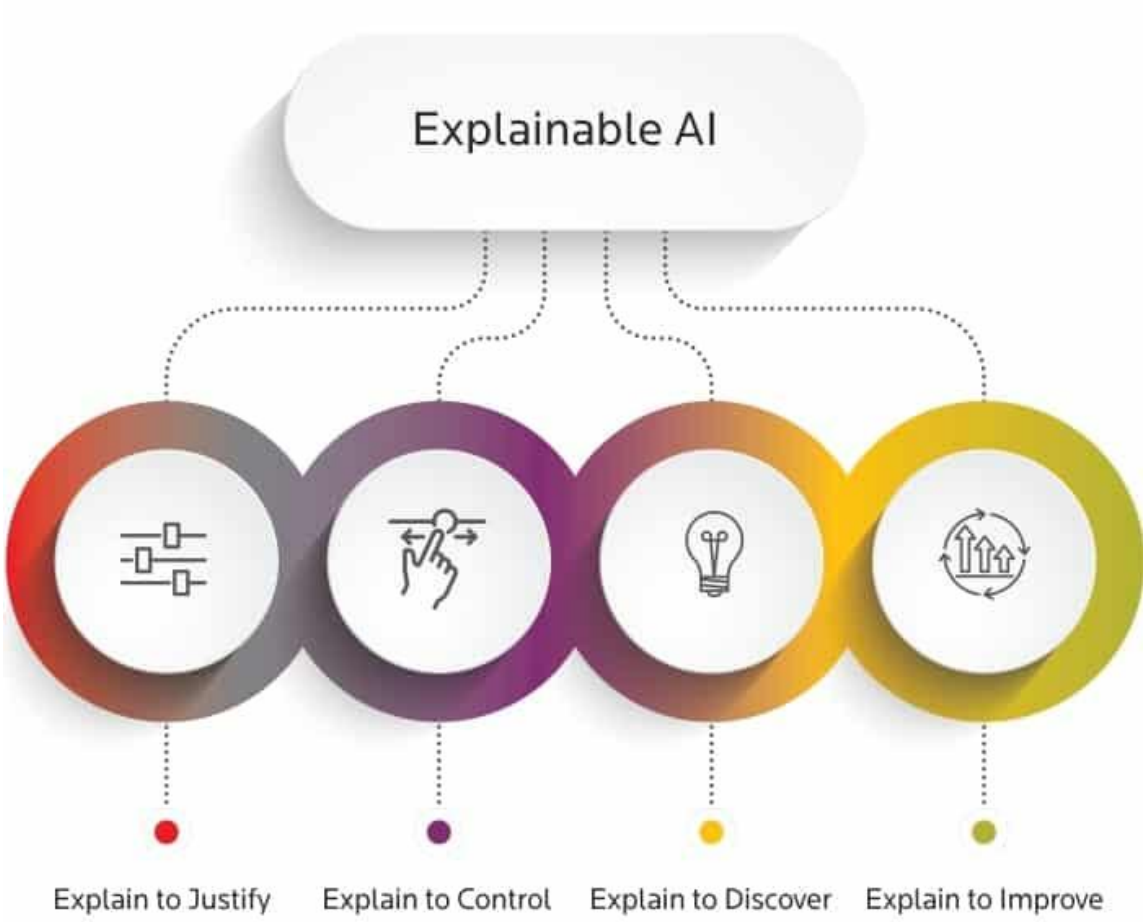


XAI | Terminology Wrap Up

XAI Terminology	
Can it explain a particular Model?	Model Agnostic Model Specific
Does it explain a particular sample? Or the entire model?	Global explanation Local explanation
When does it occure?	Pre-Model In-Model Post-Model
Does it mimic the model?	Surrogate

https://www.researchgate.net/publication/342358253_Explainable_Deep_Learning_Models_in_Medical_Image_Analysis

XAI | Tools & Libraries



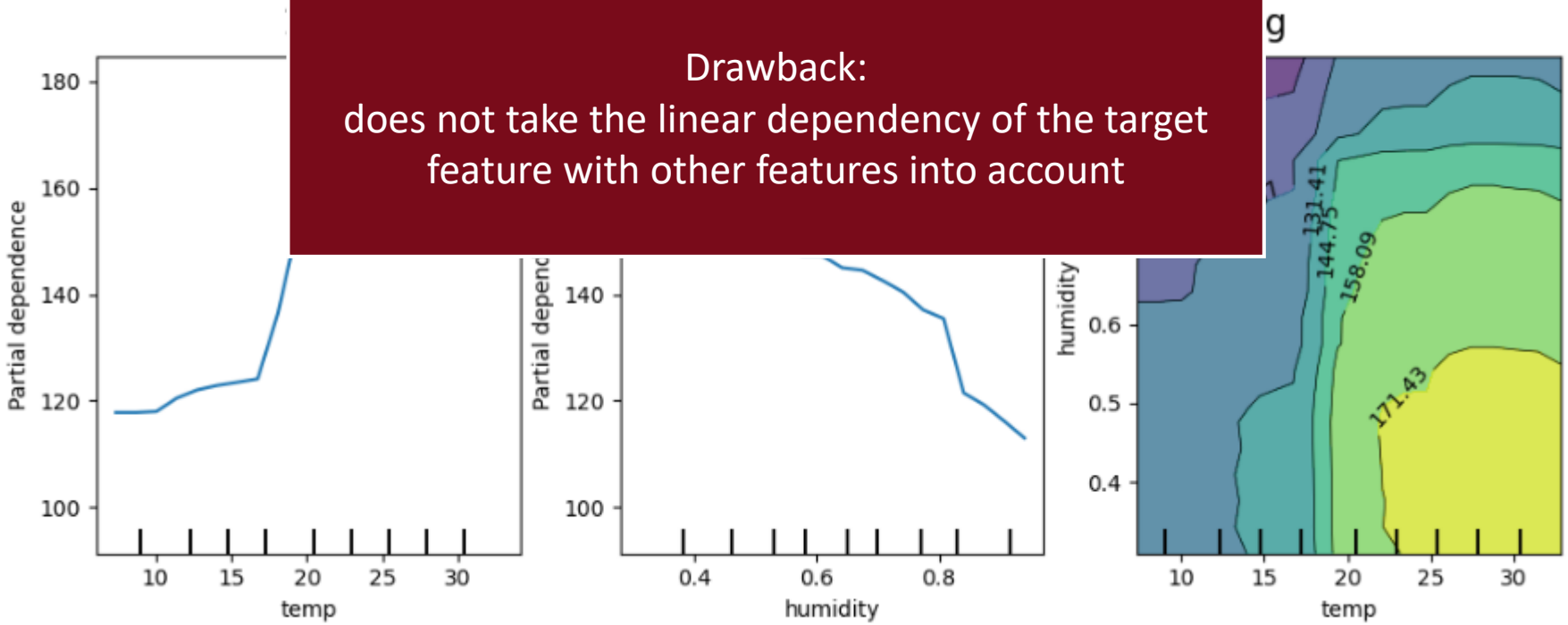
XAI | Partial Dependance Plot

Calculate Partial Dependence of a feature on the predictions

→

observe the feature's effect on the average predictions

Drawback:
does not take the linear dependency of the target feature with other features into account



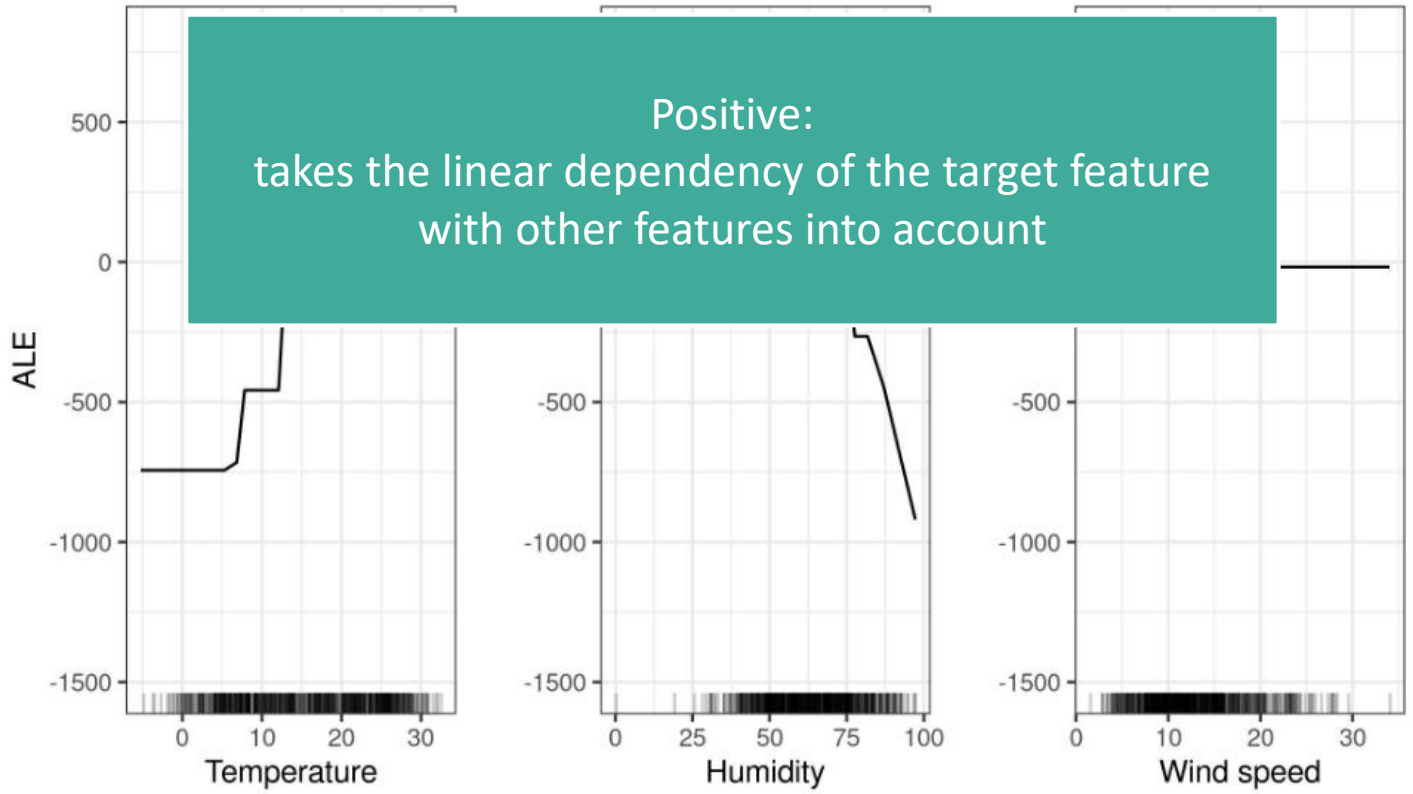
https://scikit-learn.org/stable/modules/partial_dependence.html#:~:text=Partial%20dependence%20plots%20(PDP)%20show,the%20'complement'%20features).

XAI | Accumulated Local Effect

Calculate Accumulated Local effect of a feature on the predictions

→

observe the changes of the predictions within an interval

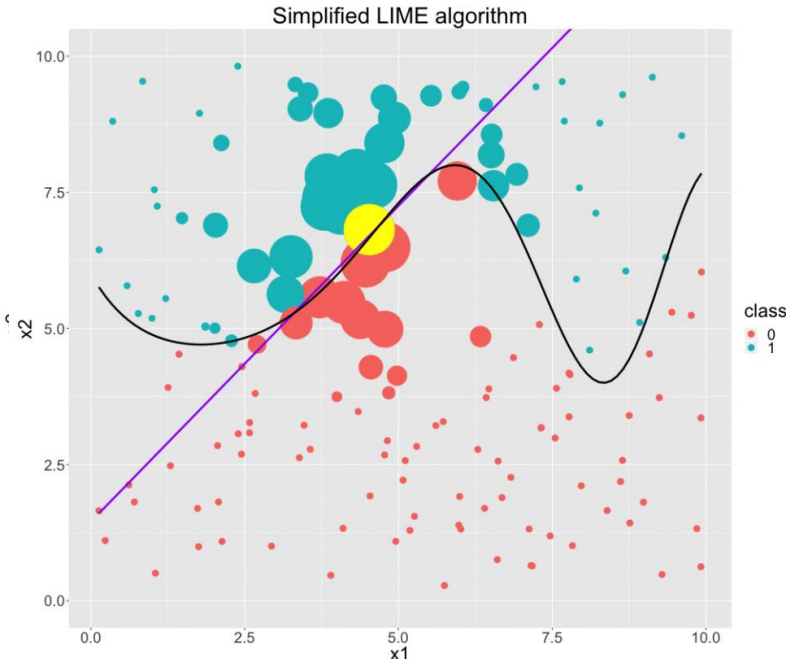


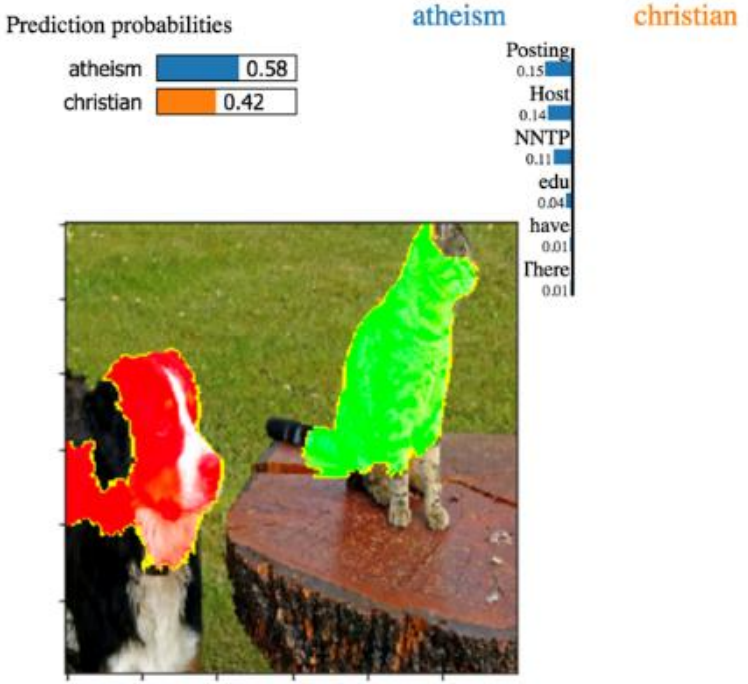
Positive:
takes the linear dependency of the target feature with other features into account

https://christophm.github.io/interpretable-ml-book/ale.html

XAI | Local Interpretable Model Agnostic Explanations

- LIME

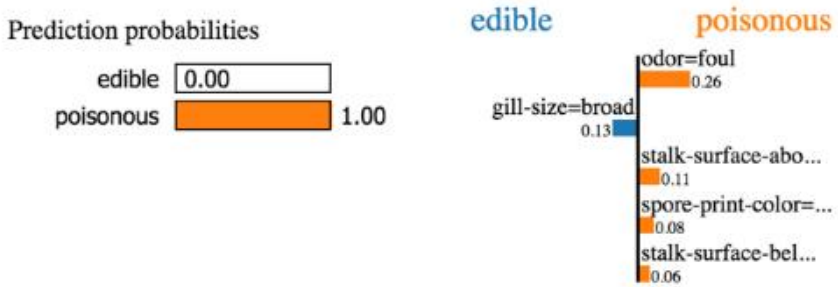




Text with highlighted words
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.



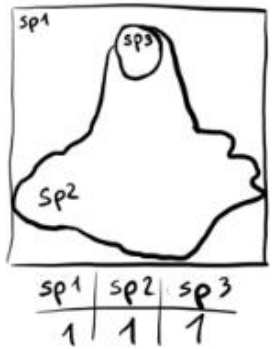
Feature	Value
odor=foul	True
gill-size=broad	True
stalk-surface-above-ring=silky	True
spore-print-color=chocolate	True
stalk-surface-below-ring=silky	True

XAI | SHapely Additive exPlanations?

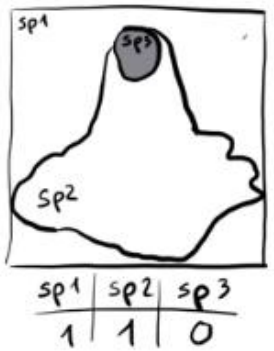
- SHAP

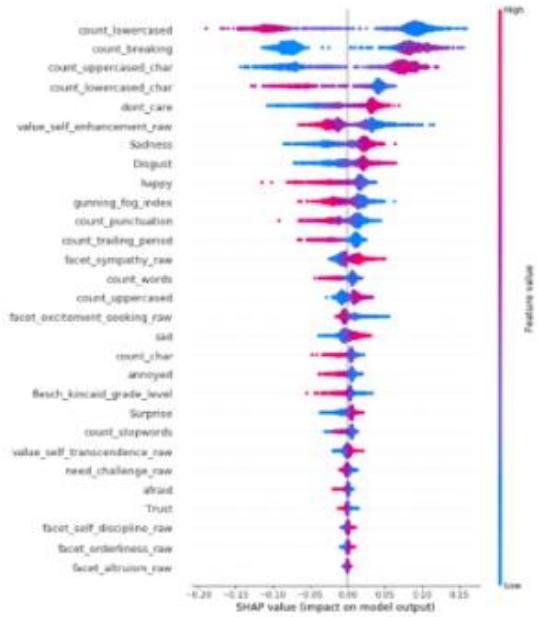
Coalitions of super pixels $\xrightarrow{h_x(z')}$ Image

Instance x

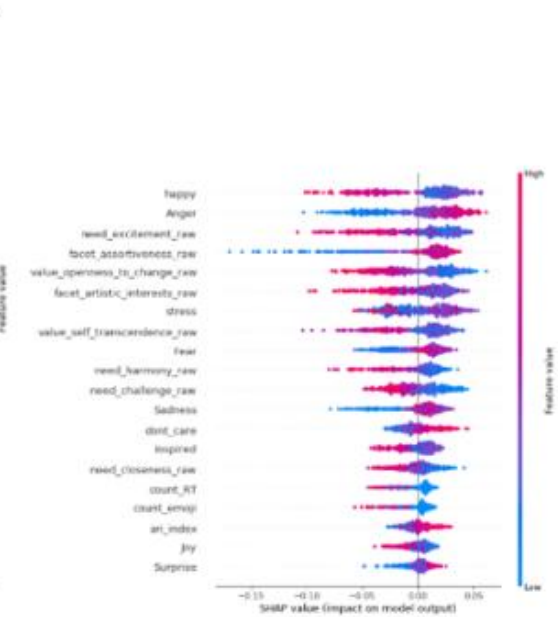


Instance x with absent features

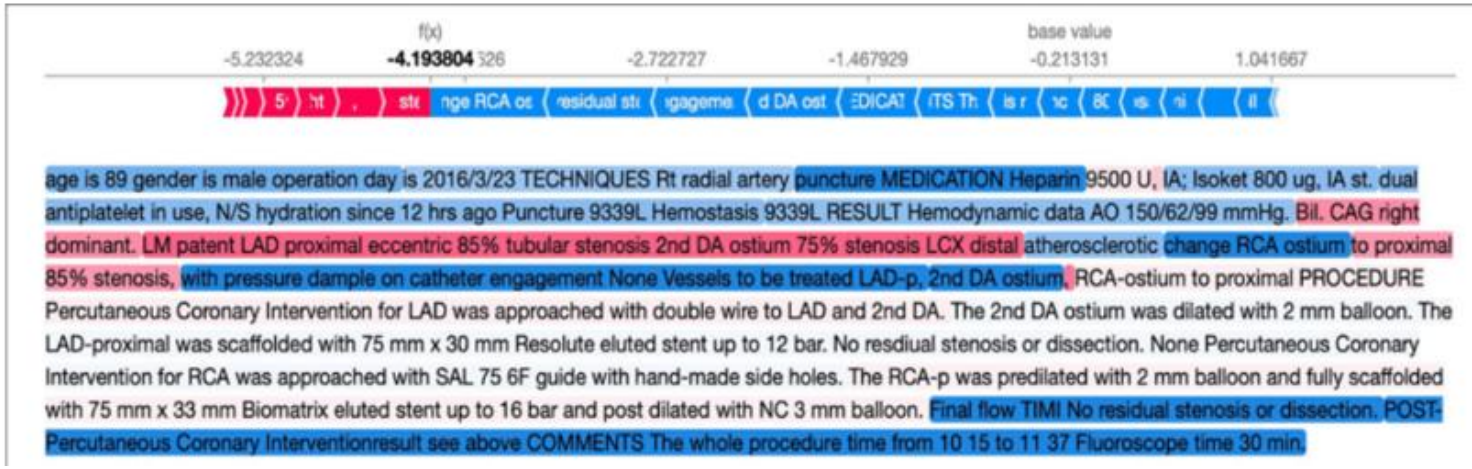




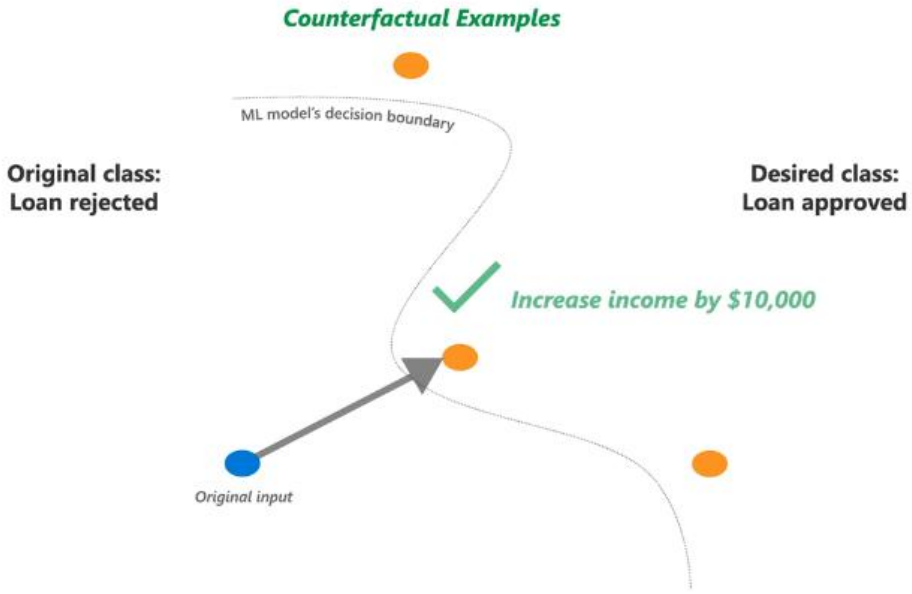
(a) PolitiFact



(b) PAN



XAI | Diverse Counterfactual Explanations (DiCE)




```
# visualize the resutls
dice_exp.visualize_as_dataframe()
```

Query instance (original outcome : 0)

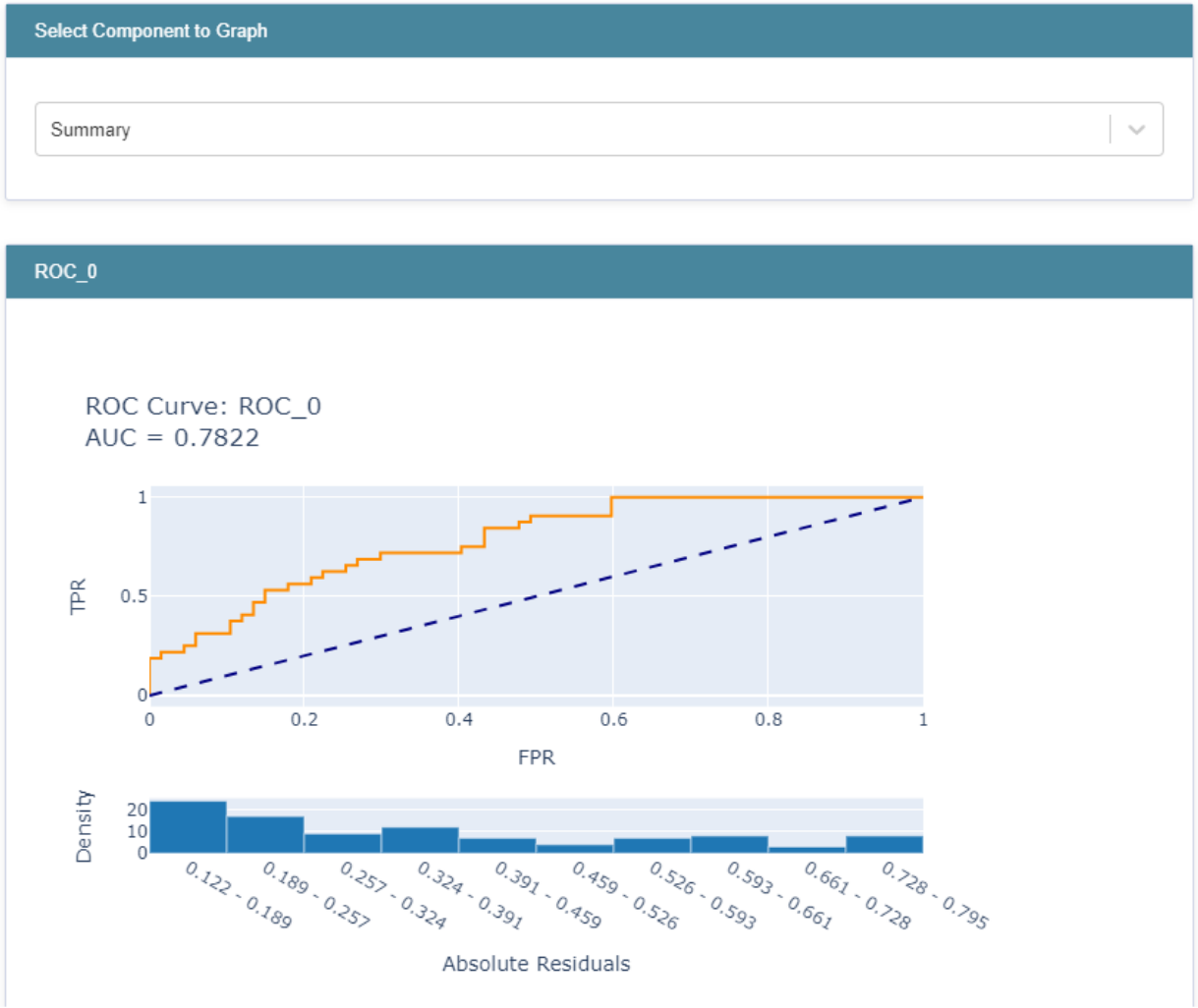
	pclass	sex	age	sibsp	parch	fare	embarked	title	family_size	survived
0	3	male	22.0	1	0	7.25	S	Mr	1	0.136179

Diverse Counterfactual set (new outcome : 1)

	pclass	sex	age	sibsp	parch	fare	embarked	title	family_size	survived
0	1	male	22.0	8	0	11.89	C	Mr	1	0.662
1	3	male	22.0	1	0	0.00	S	Master	1	0.968
2	1	female	35.7	2	0	25.43	S	Mr	1	0.871
3	1	female	16.6	1	0	2.54	S	Mrs	1	0.996

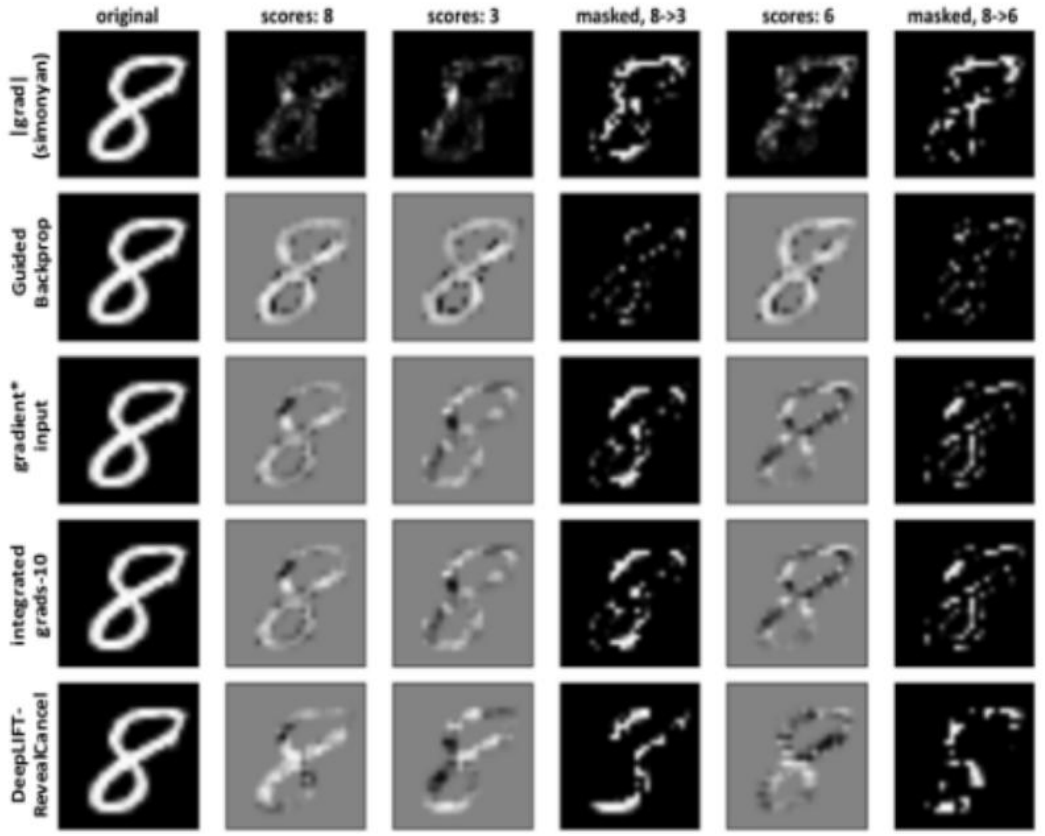
XAI | Explainable Boosting Machines

- EBM is a glassbox model indented to have comparable accuracy to ML models such as Random Forest and Boosted Trees as well as interpretability capabilities.
- Type of Generalized Additive Models (GAMS)
  - generate predictions by combining multiple functions, each representing the influence of a single predictor. These functions can be linear or non-linear
  - The model is built by adding together the contributions of each predictor. This is in contrast to multiplicative models where predictors interact with each other.

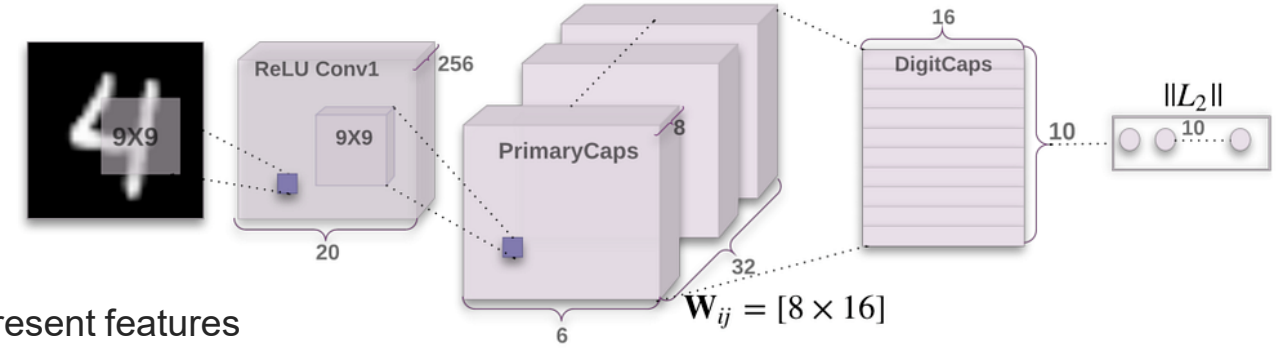


XAI | Deep Learning Important FeaTures (DeepLift)

- Learning Important Features Through Propagating Activation Differences (ICML 2017)
- Gradient-based interpretability methods
- Goal: give example specific explanations
- For a given example and an output, give importance score to individual inputs
- This is done by backpropagating the contributions of all neurons in the network to every feature of the input
- Covers only CNNs classifiers – code example here
  - https://github.com/kundajelab/deeplift/blob/master/examples/mnist/MNIST_replicate_figures.ipynb



XAI | Capsule Networks



- Capsule Networks
  - Use capsules (vectors of neurons) instead of scalar neurons to represent features
  - Encode both feature presence and properties (e.g., pose, orientation, scale)
  - Employ dynamic routing by agreement to model part, whole relationships
  - Are more **structurally interpretable** than standard CNNs due to explicit hierarchical representations

- Capsule activations are interpretable:
  - Vector length indicates confidence, while vector components encode meaningful feature attributes
- Routing coefficients act as explanations:
  - They show which lower-level features contributed to which high-level (class) capsules
- Feature-to-decision traceability:
  - Decisions can be explained by following the routing paths from input capsules to class capsules
- Less reliance on post-hoc methods:
  - Interpretability is largely built into the architecture, rather than added afterward (e.g., via saliency maps)

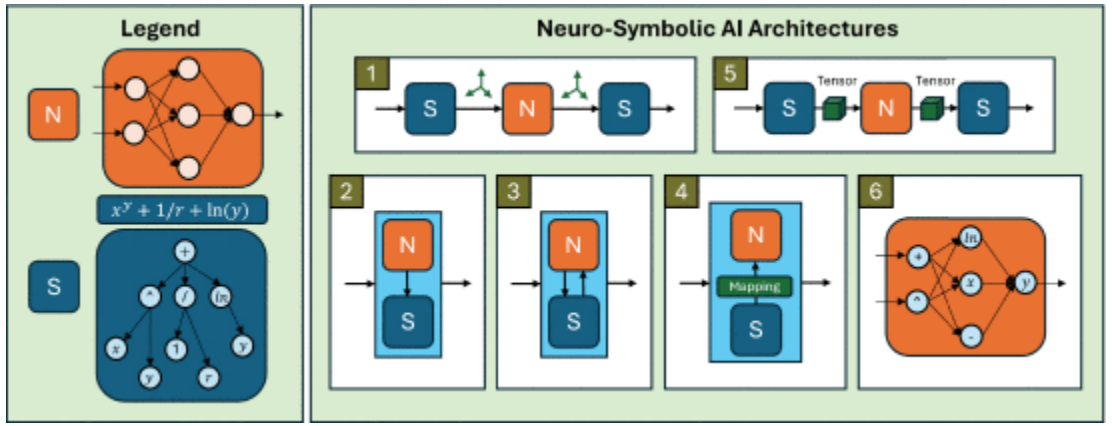
https://medium.com/ai%C2%B3-theory-practice-business/understanding-hintons-capsule-networks-part-i-intuition-b4b559d1159b

XAI | Reinforcement Learning

- **Definition**: learning to choose actions through interaction to maximize long-term reward
- **What is explained?** Action choices, policies, and long-term strategies rather than single predictions
  - Key explanation targets:
    - Why a specific action was chosen in a given state
    - How decisions contribute to future rewards (expected outcomes)
    - Which state features most influenced the policy
- Main XAI approaches:
  - **Policy explanation:** interpret or approximate policies (e.g., rules, trees)
  - **Value-based explanation:** expose Q-values, rewards, and expected returns
  - **Trajectory explanation:** explain sequences of actions over time
  - **Counterfactuals:** “What would happen if the agent acted differently?”
- Why XRL is different? Decisions are sequential and temporal & Explanations must reflect environment dynamics and reward structure

XAI | Neuro-Symbolic AI

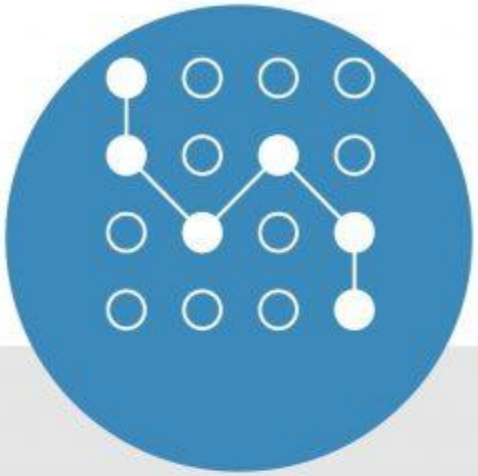
- This AI paradigm that integrates neural learning with symbolic reasoning to combine pattern recognition with explicit knowledge and logic.
- NSAI's explainability comes from exposing symbolic logic and rule-based reasoning integrated with neural network outputs, making decisions traceable and interpretable in structured terms



XAI

Explainable Data

What data was used to train the model and why?



Explainable Predictions

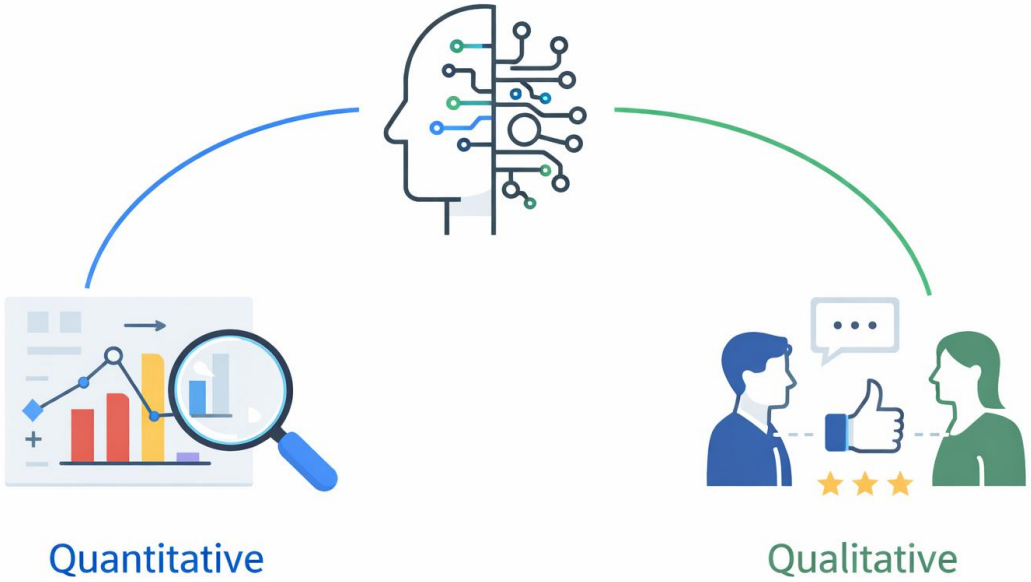
What features and weights were used for this particular prediction?

Explainable Algorithms

What are the individual layers and the thresholds used for a prediction?

Questions around AI explainability help us understand how data, predictions and algorithms influence decisions.

XAI | Evaluating explanations



- Quantitative
  - Fidelity score
  - Feature Importance
  - Consistency metrics (i.e., Stability, Robustness)
  - User Studies (I.e., Human Model agreements, comprehensibility)
  - Completeness – do they capture model behavior across different types of instances?
  - Consistency – aligns with the known domain knowledge
- Qualitative
  - Use feedbacks (on intuitiveness, understandability)
  - User Studies (trust, interactive tools, user confidence)

XAI | Challenges

- Challenge 1: Evaluation of model explainability and Interpretability
- Challenge 2: Some methods to explain black box models might be black box themselves
- Challenge 3: Non-consistent use of terminology (suggestions: Adadi et al., 2018 and Guidotti et al., 2018)
- Challenge 4: Lack of research for Explainable ML/AI for time series
- Challenge 5: Coverage of explanations w.r.t. black box models
- Challenge 6: Integration of domain knowledge into XAI (or even AI)

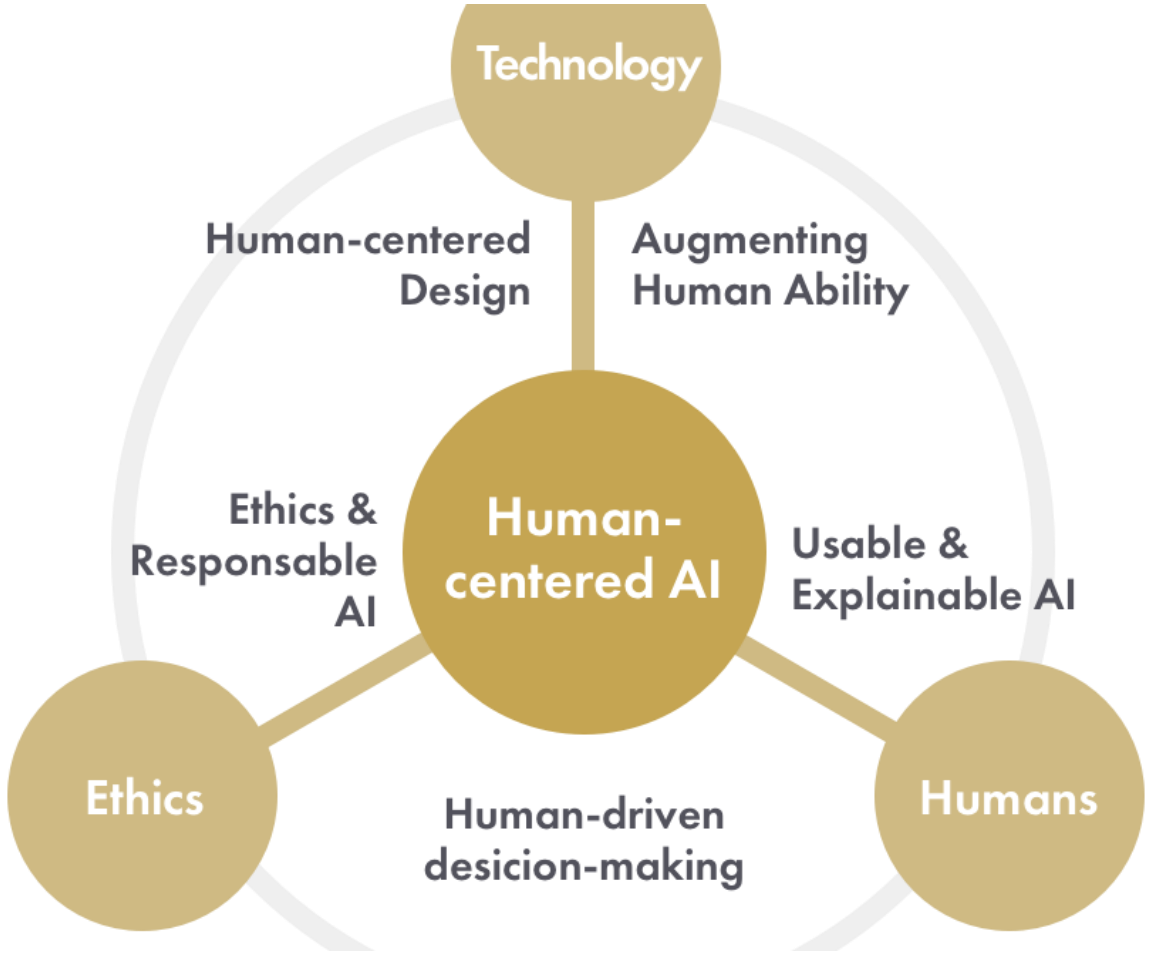
XAI | Summary

- AI makes gives suggestions and predictions, humans make decisions and assign meaning
- Interpretation is the goal, explanations are the means

- There is no single explanation
  - Good explanations depend on who, what, and scope

- XAI helps us judge trustworthiness
  - Not just accuracy, but reliability, fairness, and robustness

- Explanations support decisions, not replace them
  - Human judgment remains essential



Q & A

XAI | References

1. Doshi-Velez & Kim (2017): *Interpretability depends on the target audience*
2. https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html
3. https://link.springer.com/chapter/10.1007/978-3-031-04083-2_2
4. https://towardsdatascience.com/explainable-neural-networks-recent-advancements-part-3-6a838d15f2fb
5. https://arxiv.org/pdf/1802.07810.pdf
6. J. S. Kadyan, M. Sharma, S. Kadyan, S. Gupta, N. K. Hamid and B. Kiran Bala, "Explainable AI with Capsule Networks for Credit Risk Assessment in Financial Systems," *2025 International Conference on Next Generation Information System Engineering (NGISE)*, Ghaziabad, Delhi (NCR), India, 2025, pp. 1-6, doi: 10.1109/NGISE64126.2025.11085369, https://ieeexplore.ieee.org/abstract/document/11085369
7. https://ieeexplore.ieee.org/abstract/document/11124915

Contact

Anahid Wachsenegger

AIT Austrian Institute of Technology GmbH

anahid.wachsenegger@ait.ac.at

AI Factory Austria AI:AT
Schwarzenbergplatz 2
1010 Wien, Austria

training@ai-at.eu
info@ai-at.eu
ai-at.eu

@ai-factory-austria

Funded by

under discussion with

AI Factory Austria AI:AT has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101253078. The JU receives support from the Horizon Europe Programm of the European Union and Austria (BMIMI / FFG).