Sweden AI Factory

Explainable AI (XAI)

What to explain and how to check it

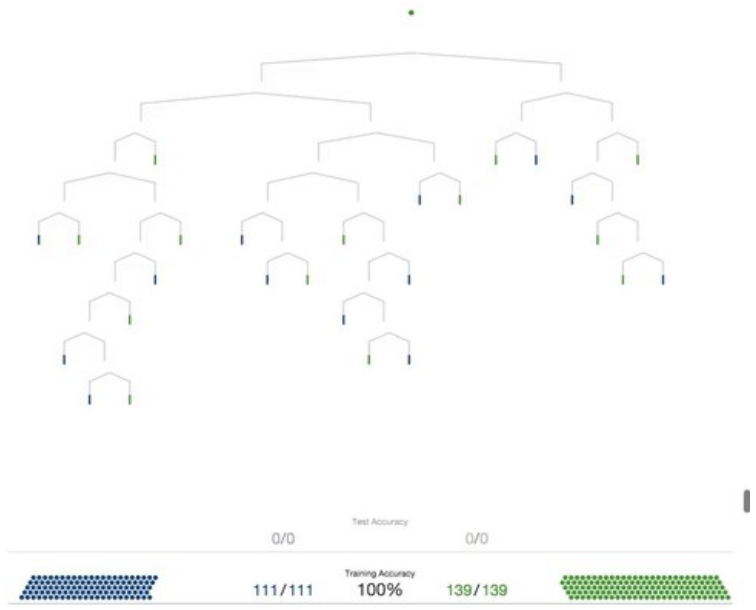
Nishat I Mowla, PhD

Senior Researcher, RISE
Team Lead of Trustworthy AI & Testing Frameworks,
Sweden AI Factory



AI image / Limitless Visions / Adobe Stock

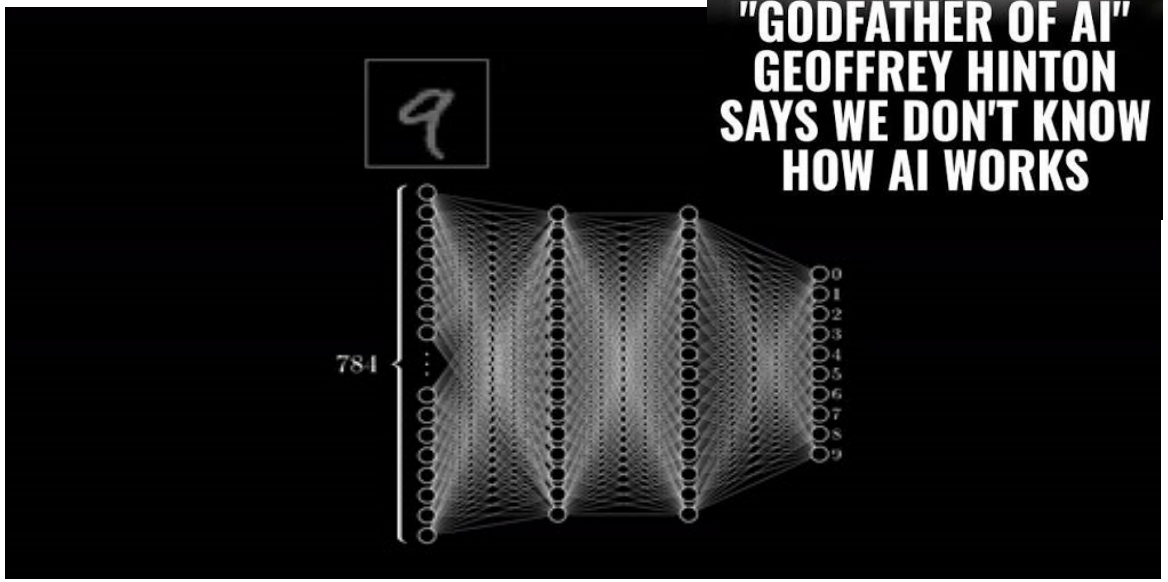
AI capability & transparency



Random Forest

A set of decision trees using a rule you can read, trace and argue with. Every prediction has a path.





Deep Neural Network

A deep neural network involves millions of weights. There is no path to point at, only an output.

Sweden AI Factory

The regulatory floor

Under the EU AI Act, transparency for high-risk systems is an obligation on the provider, not just a nice-to-have anymore.

AI ACT ARTICLE 13

Transparency requirements

High-risk systems must be designed so their operation is sufficiently transparent for deployers to **interpret the output and use it appropriately.** Instructions must cover **capabilities, limitations, expected performance**, and **how to read the output** and the **logs.**

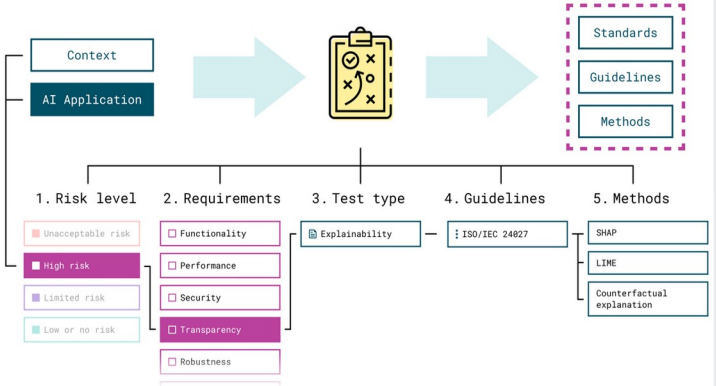
AI ACT ARTICLE 86

Right to an explanation

A person subject to a decision made with a **high-risk system** that significantly affects their **health, safety or fundamental rights** should be able to obtain a clear and meaningful explanation of the role the system played in that decision.

A MAPPING

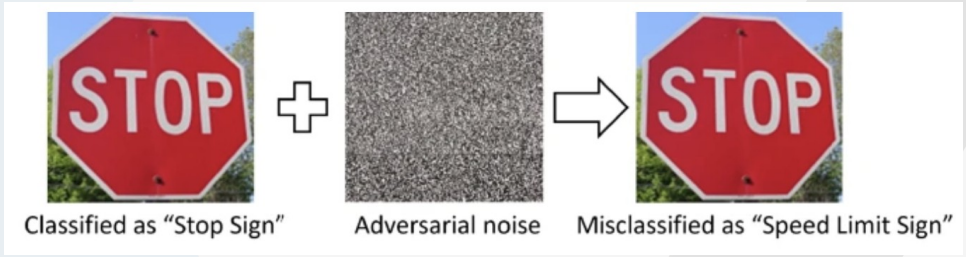
AI Act risk to requirements



Sources: The EU AI Act was published in the Official Journal of the European Union on 12 July 2024

Sweden AI Factory

XAI in AI pipeline



Gao, Z., Saremi, A., & Gu, Z. (2026). Performance measurement and analysis of certifiable defenses against adversarial patch attacks. Journal of Signal Processing Systems, 98, Article 22.

Exiting AI:

Training Data → Machine learning process → Learned function → Decision making → Cyber analyst

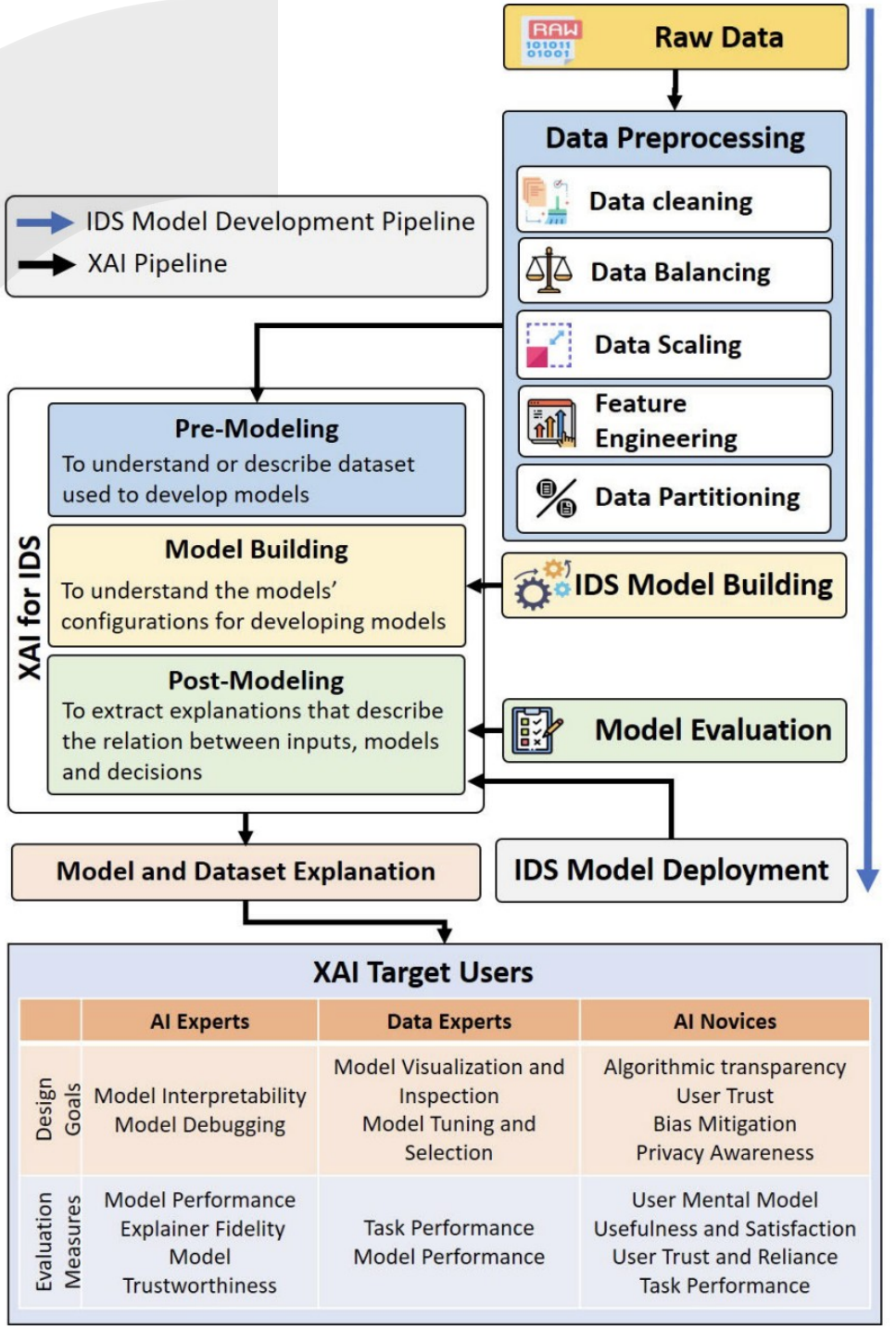
Limitations
- Why AI chose this decision?
- When AI succeeds or fails?
- How reliable is the decision?

Explainable AI:

Training Data → New machine learning process → Explainable model ⇄ Cyber analyst

Expectations
- Explain why AI chose this decision
- Explain when AI succeeds or fails
- Explains when to rely on AI decision

Mowla, Nishat I., Kyi Thar, Sarder Fakhrul Abedin, Aamir Mahmood, Zhu Han, Mikael Gidlund, Kabir Fahria et al. "ACHILLES: A Machine Learning Framework for Explainable and Generalized Automotive Intrusion Detection System." *IEEE Transactions on Intelligent Transportation Systems* (2026).

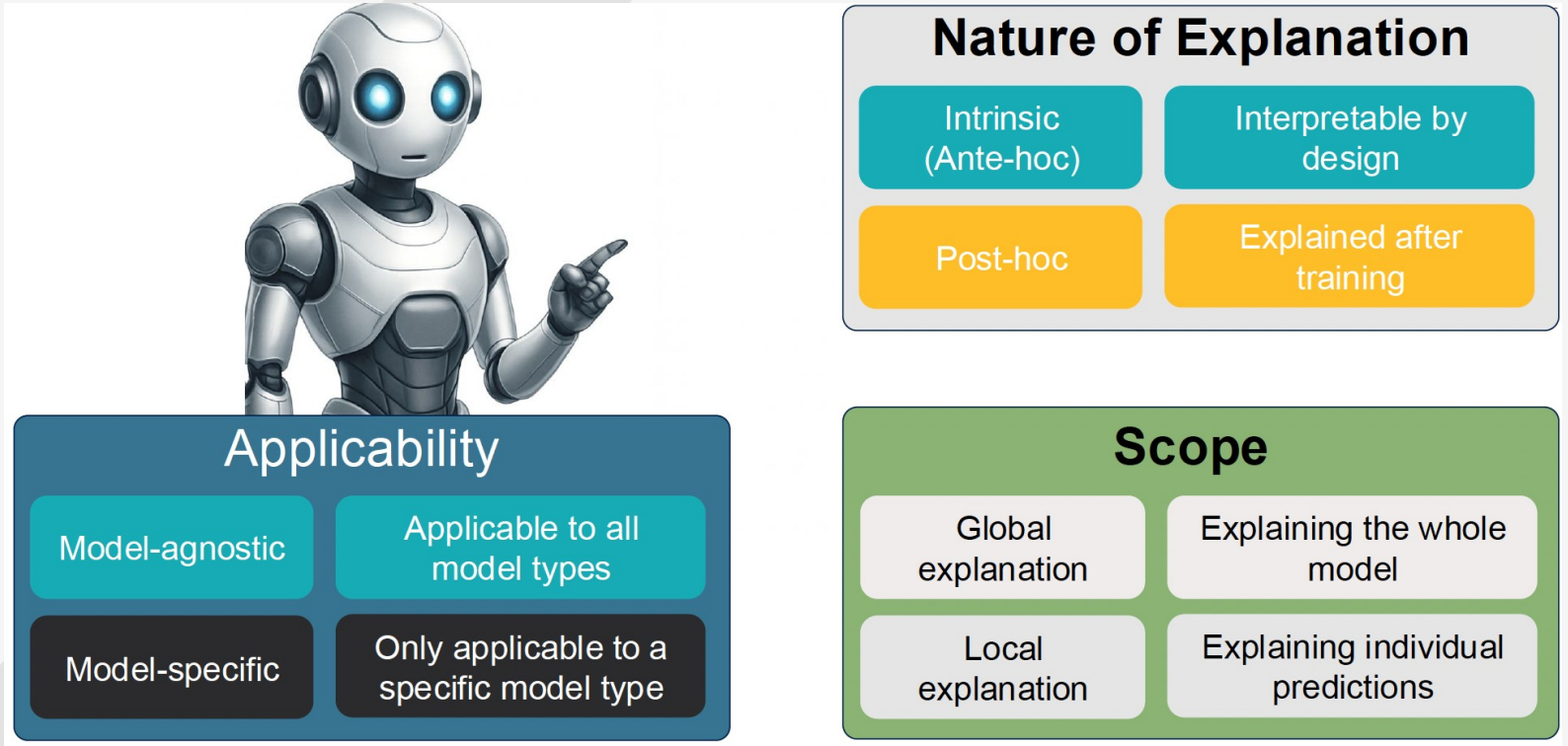


1. **Debug and improve**
2. **Trust and adoption**
3. **Fairness and safety**
4. **Accountability and understanding**

Sweden AI Factory

Existing XAI methods

Nature of Explanation · Applicability · Scope.



Kyi Thar, Mid Sweden University, workshop on explainable AI (Categorization of Explanation).

Sweden AI Factory

Intrinsic or post-hoc

When is the explanation created: together with the model, or afterwards?

A Intrinsic (ante-hoc)

The model is the explanation

The **model's structure is readable**. Nothing is approximated, because there is no second artefact. What you read is exactly what runs.

TYPICAL MODELS

Linear & logistic regression · shallow decision trees

B Post-hoc

A second artefact explains the first

Train whatever performs best, then build an explanation on top. Maximum freedom in modelling, at the price of an **explanation that only approximates the model**.

TYPICAL METHODS

SHAP · LIME · counterfactuals

Sweden AI Factory

Global or local

Does the explanation describe the model, or one decision?

A Global

The model as an object

Which features matter overall, what the learned relationships look like, where the model is confident.

The view a developer, a reviewer or an auditor needs.

TYPICAL METHODS

Permutation importance · mean SHAP · global surrogate trees

B Local

One row, one output

Why this applicant, this scan, this packet and what would have to be different for another answer.

The view the affected person and the operator need.

TYPICAL METHODS

SHAP values · LIME · counterfactuals · saliency maps

Sweden AI Factory

Model-agnostic or model-specific

What can the method actually be attached to?

A Model-agnostic

Needs only inputs and outputs

Treats the model as a **black box it can query**. Portable across model families, survives a model swap, and usually **slow** because it has to probe the model thousands of times per explanation.

TYPICAL METHODS

KernelSHAP · LIME · permutation importance · counterfactuals

B Model-specific

Exploits internal structure

Uses the **gradients, the tree structure or the activations**. Orders of magnitude **faster**, often exact at the cost of only working on one family of models.

TYPICAL METHODS

TreeSHAP · DeepSHAP · gradient based counterfactuals

Sweden AI Factory

Five forms of explanation

1

Feature attribution

How much did **each input push the output**, and **in which direction**? The default form, and the most quantitative.

SHAP · LIME · Integrated Gradients

2

Rules & surrogates

A simple readable model that **mimics the complex one**: globally, or just in the neighbourhood of one point.

Surrogate trees

3

Example-based

Explain by pointing at data: the **most similar cases**, and the training points that most influenced this decision.

k-NN

4

Counterfactual

The smallest change to the **input that flips the outcome**. The only form that gives the affected person a **route to act**.

Wachter et al. · Diverse counterfactual explanations (DiCE)

5

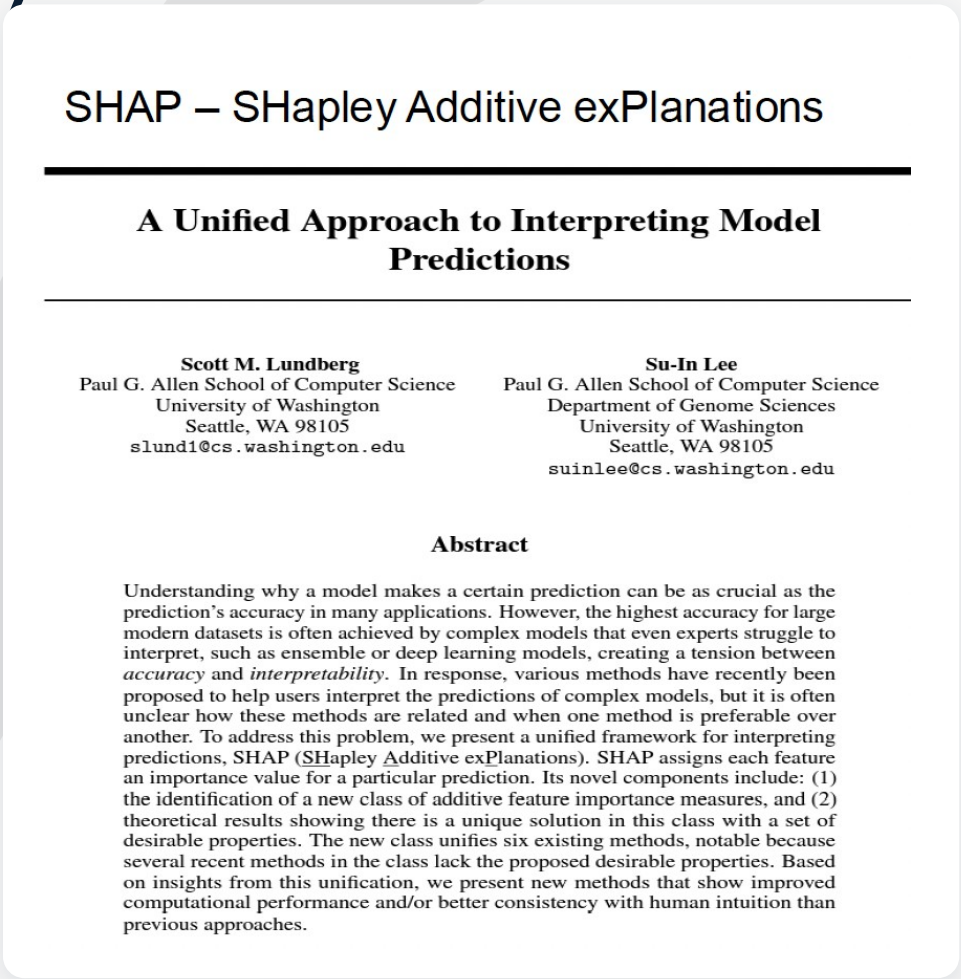
Visual & saliency

Where in an image, a spectrum or a sequence the evidence sits, as a map over the input rather than a table.

Grad-CAM · attention maps

Sweden AI Factory

SHAP- Shapley Additive exPlanations



SHAP – SHapley Additive exPlanations

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. NeurIPS 30. · Documentation: shap.readthedocs.io

Sweden AI Factory

The idea: a fair split of the payout

- Four people enter a competition as a team and win $10,000
- They did not contribute equally. How should the prize be split?
- Shapley (1953): each player gets their average marginal contribution across every order in which the team could have formed

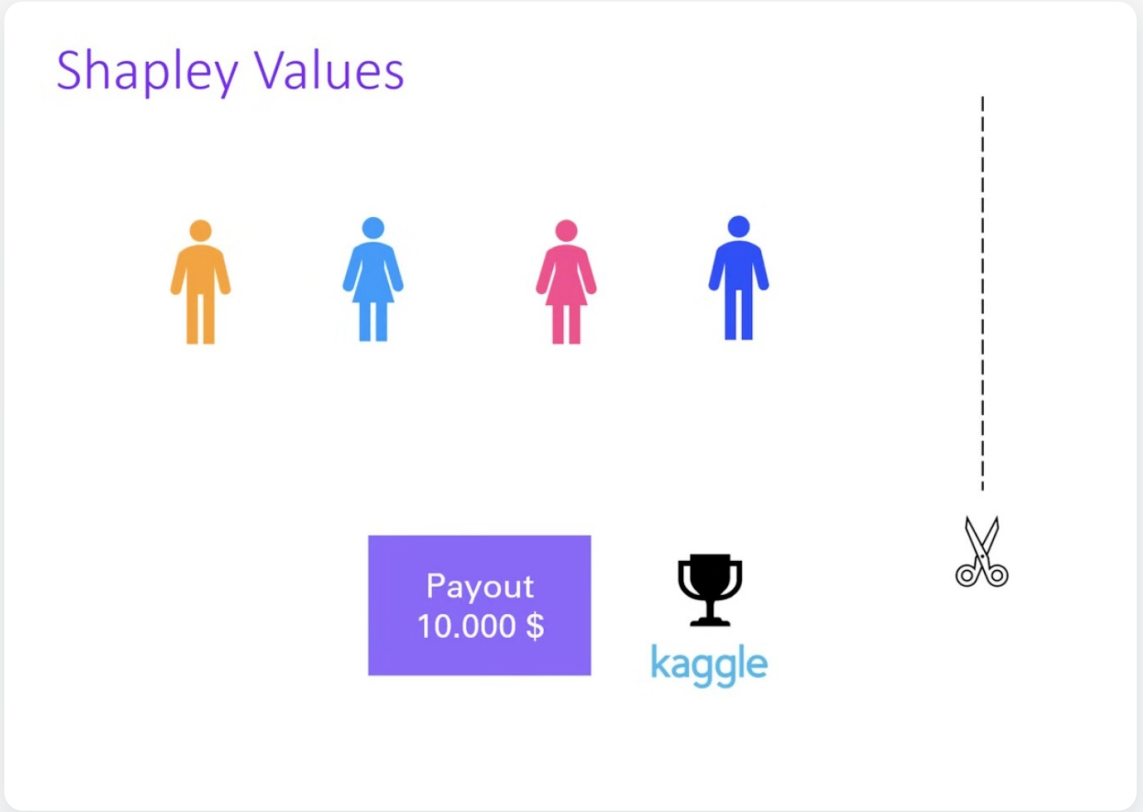
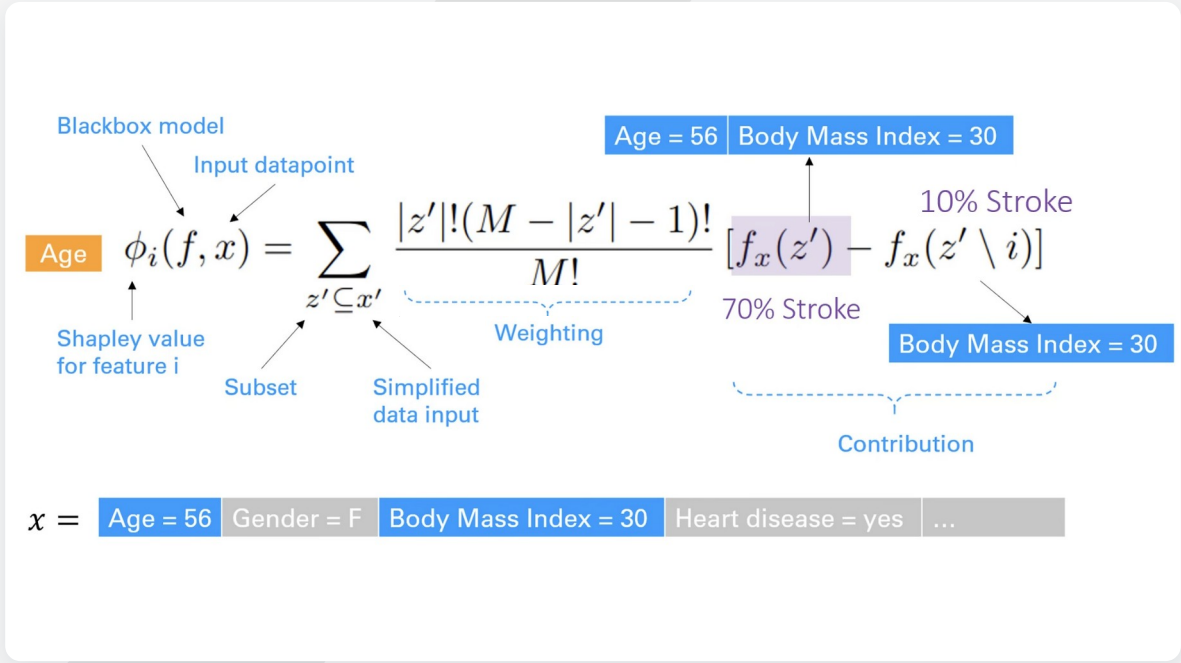


Illustration after DeepFindr, "SHAP values explained" (YouTube).

Sweden AI Factory

The Shapley value

A weighted average of the change in prediction when feature i joins a coalition.



→ *Kernel SHAP*

→ *Tree SHAP*

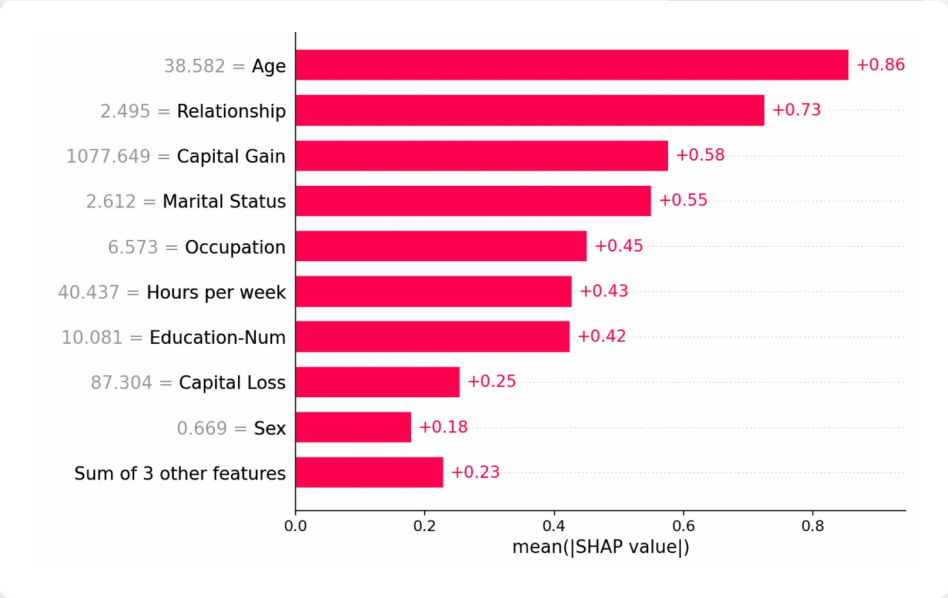
Exact computation is exponential in the number of features

Illustration after DeepFindr, "SHAP values explained" (YouTube).

Sweden AI Factory

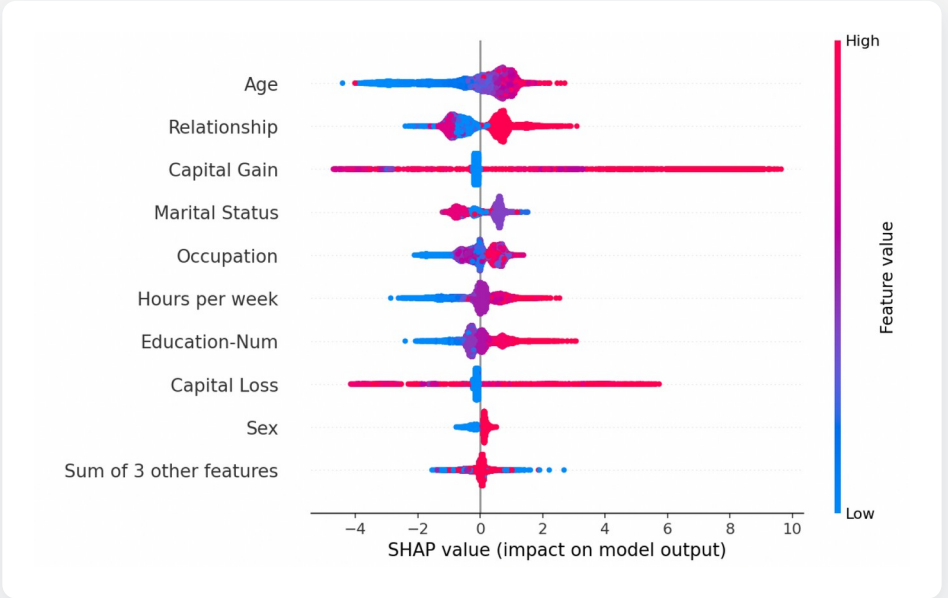
Typical SHAP explanations

Ranked — mean SHAP per feature



Magnitude of influence, direction discarded. Labels carry the feature value, so the ranking stays interpretable.

Beeswarm, one dot per case

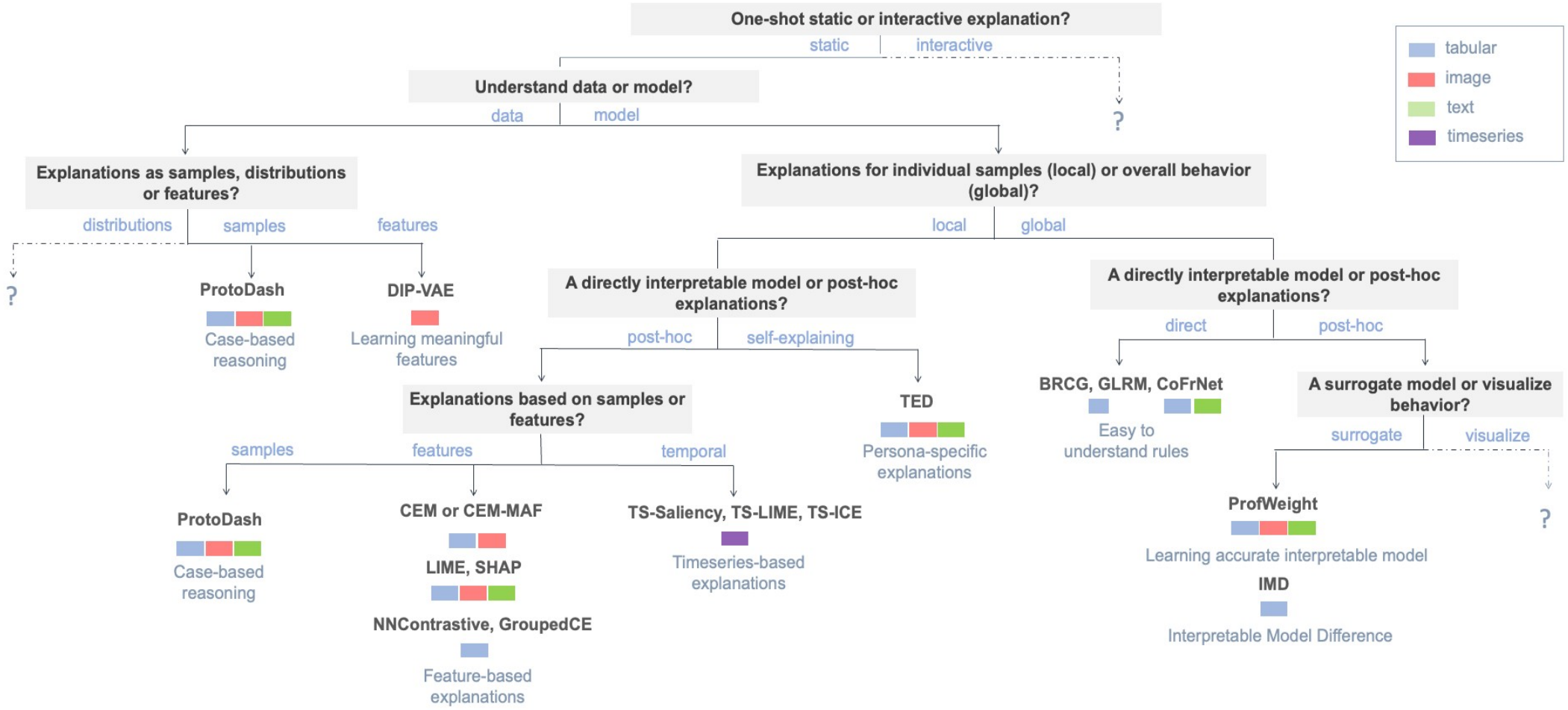


Horizontal position is the SHAP value, colour is the feature value. Spread shows where a feature is decisive and where it is not.

After Kyi Thar, Mid Sweden University — workshop on explainable AI.

Sweden AI Factory

Choosing a method: a decision path

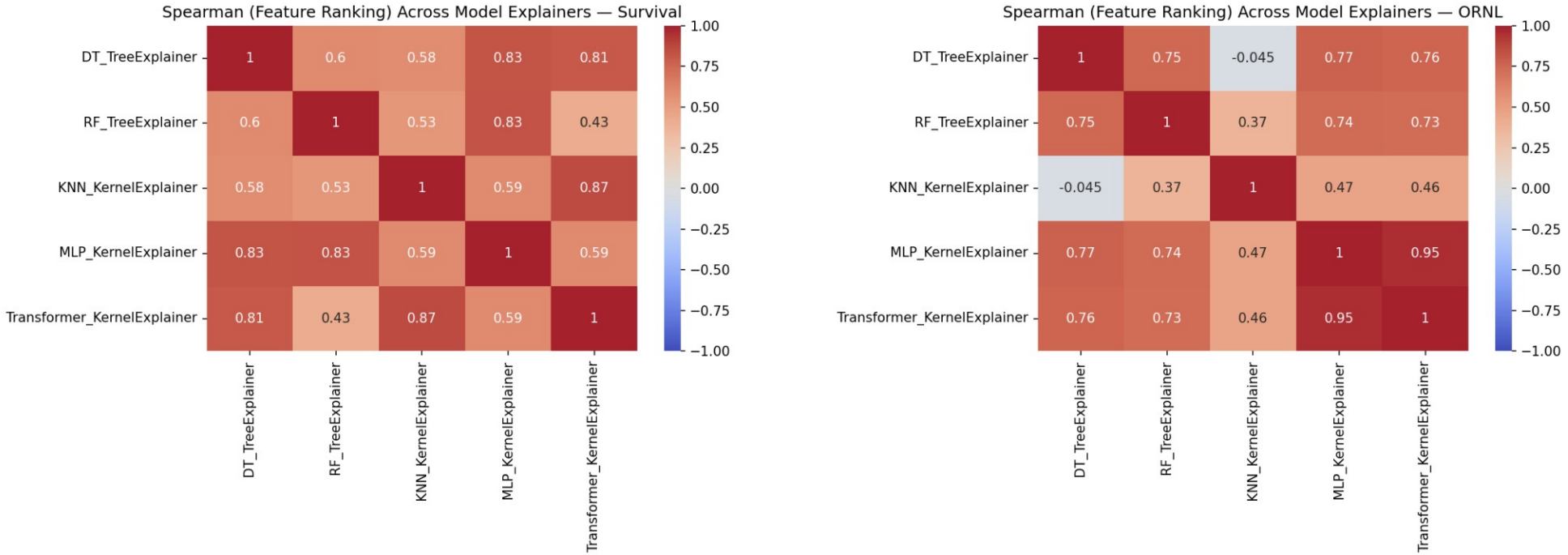


Source: IBM Research — AI Explainability 360.

Sweden AI Factory

Do explainers agree?

Five model families, five explainers, two automotive intrusion-detection datasets



(a) Survival dataset. (b) ORNL dataset.

Mowla, Rosell & Abedin (2026), ACDSA. Figure: Spearman rank correlation between model–explainer pairs.

Sweden AI Factory

Tooling landscape

Mature, open, and mostly Python

1

Attribution

TreeSHAP and DeepSHAP. Captum covers gradient methods for PyTorch.

Captum

2

Rules & surrogates

Local surrogates and rule extraction, plus interpretable model families you can train directly.

InterpretML · LNN

3

Counterfactual & recourse

Diverse, feasible counterfactuals with mutability constraints

CARLA – Counterfactual And Recourse Library

4

Toolkits

Curated collections with a taxonomy, runnable notebooks and one consistent API across method families.

AIX360 · InterpretML

5

Evaluation

Faithfulness, robustness, and stability: check that any of the above is true.

captum.metrics

Sweden AI Factory

Who is the user?

Affected person

"What in my case drove this and what would change it?"

Needs a sentence and a route. Local, contrastive, in plain language.

Recourse, not coefficients

Data scientist

"Is the model using the signal I think it is?"

Global importance, error analysis, leakage detection. The only audience most tools are designed for.

Debugging, not communication

Domain expert

"Should I act on this recommendation, now?"

Reasons plus calibrated confidence, at decision time, in the operator's own vocabulary.

Their real question is when to skip a check

Auditor / regulator

"Can this be documented and reproduced?"

Traceable evidence against a standard: versioned, archived, reproducible eighteen months later.

Provenance beats elegance

Sweden AI Factory

Checklist

1. Name the audience and the question before you pick a method.
2. Try an intrinsically interpretable model and measure what it actually costs you.
3. Record the model version, the explainer, and the background dataset.
4. Explain at least two viable models and report how much they agree.
5. Test faithfulness and stability, not just whether the plot looks reasonable.
6. Ship the artefact the audience can use: a sentence, a rule, or a counterfactual.



Pinterest. (2020, February 22). Every cat thinks it is a lion! [Pin].

Sweden AI Factory