Rigorous Explainability by Feature Attribution

From SHAP to nuSHAP

Joao Marques-Silva

ICREA & Univ. Lleida, Spain

KR @ FLoC'26, July 2026

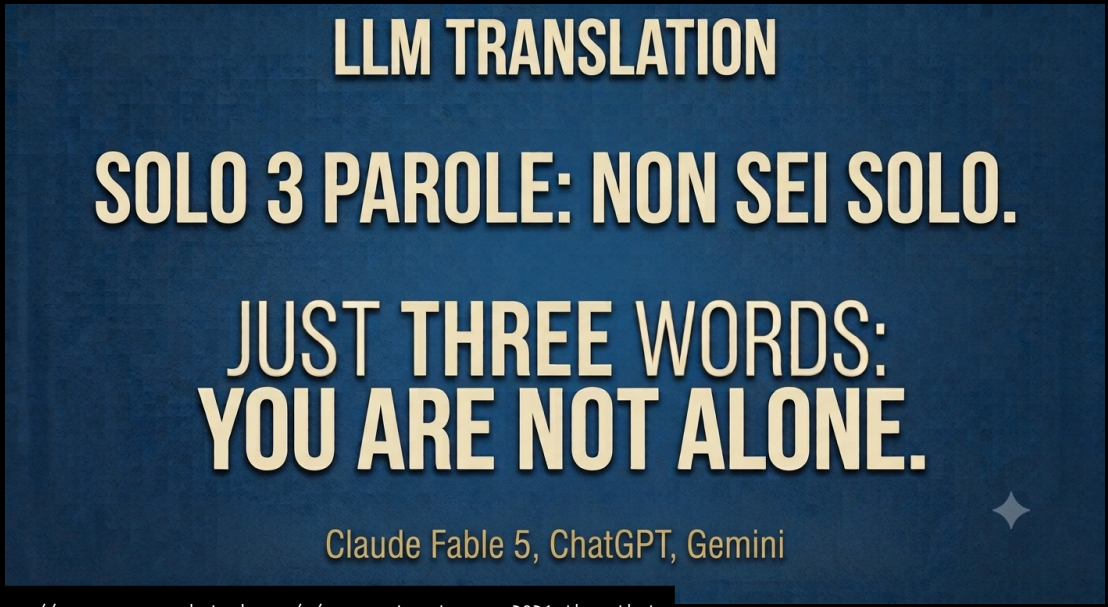
> *"All truths are easy to understand once they are discovered; the point is to discover them."*
>
> (G. Galilei)

ADVANCES OF AI IN RECENT YEARS: A TRANSFORMATIVE DECADE

SOURCE: INDUSTRY ANALYSIS AND RESEARCH (2024)

But news of human-level intelligence are greatly exaggerated...

But news of human-level intelligence are greatly exaggerated...



https://garymarcus.substack.com/p/you-cant-get-more-2026-than-that

Also, hallucinations are a critical weakness...

AI Hallucinations

When AI agents produce erroneous or misleading historical information

When did Angola become an independent country?

AI Agent

AI Response (Incorrect)

Angola has been an independent country since **1962.**

This is incorrect.
The independence date is wrong.

Historical Fact (Correct)

Angola has been an independent country since **11 November 1975.**

11 November 1975
Angola gained independence from Portugal.

Key Takeaway: AI systems can confidently generate incorrect historical information. **Always verify historical facts with reliable sources.**

Some clues of LLM hallucinations in recent papers...

VERGE: Formal Refinement and Guidance Engine for Verifiable LLM Reasoning

Wrong titles, authors, author list...

27 Jan 2026

Anton Belov, Mikoláš Janota, Inês Lynce, and Joao Marques-Silva. 2012. On computing minimal correction subsets. In *International Joint Conference on Artificial Intelligence*, pages 615–622.

Mark H Liffiton and Ammar Malik. 2008. Algorithms for computing minimal unsatisfiable subsets of constraints. *Journal of Automated Reasoning*, 40(1):1–33.

Some clues of LLM hallucinations in recent papers...

Concept-Based Abductive and Contrastive Explanations for Behaviors of Vision Models

Wrong titles, authors, author list...

7 May 2026

Alexey Ignatiev, Nina Narodytska, and João Marques-Silva. Formal explanations of neural network predictions. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 12344–12354, 2019b.

Alexey Ignatiev, Hadi Izza, João Marques-Silva, Nina Narodytska, and Mate Soos. Using MaxSAT for efficient explanations of tree ensembles. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2022)*, volume 36, pp. 3776–3784, 2022.

Hadi Izza and João Marques-Silva. On explaining random forests with SAT. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021)*, pp. 2584–2590, 2021.

THE IMPORTANCE OF EXPLAINABLE AI (XAI) FOR BUILDING TRUST

KEY BENEFITS OF XAI FOR TRUST

1. ENHANCED ACCOUNTABILITY
CLEAR RESPONSIBILITY FOR DECISIONS

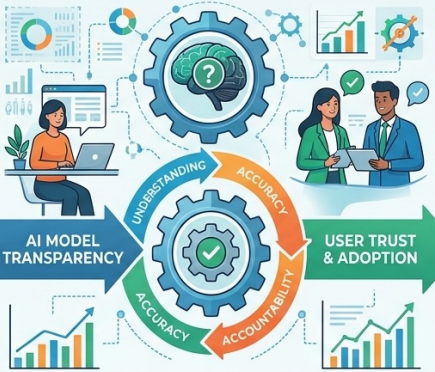
2. IMPROVED FAIRNESS & EQUITY
DETECTING AND MITIGATING BIAS

3. REGULATORY COMPLIANCE
MEETING ETHICAL & LEGAL REQUIREMENTS

4. USER UNDERSTANDING
MAKING AI REASONING AUDITABLE

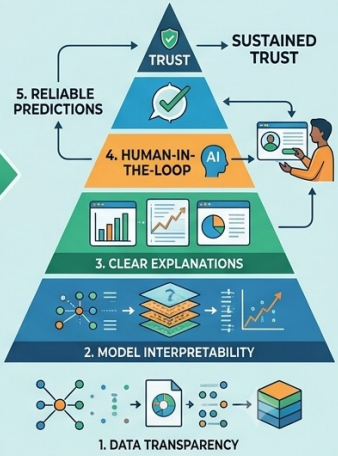
5. SYSTEM DEBUGGING & RELIABILITY & EDGE CASES

IDENTIFYING ERRORS & EDGE CASES



HOW XAI COUNTERS THE "BLACK BOX"

BUILDING TRUST THROUGH EXPLAINABILITY



XAI FOSTERS CONFIDENCE, REDUCES RISKS, AND ENSURES ETHICAL & RESPONSIBLE AI DEPLOYMENT

SUMMARY OF GLOBAL REGULATIONS ON EXPLAINABLE AI (XAI)

EXISTING FRAMEWORKS AS OF JULY 2026

EUROPEAN UNION (EU)

EU AI ACT
(In Force, 2026)

- Risk-based approach
- **High-Risk Systems:** Strict XAI requirements (transparency, (transparency, right to explanation)
- **GPAI Models:** Specific transparency (Technical documentation)

CANADA

AIDA (Artificial Intelligence & Data Act)
(Implementation Stage)

- Focus on **High-Impact AI**
- Requirements for **explainability to individuals and responsible management**
- **Preventing Harmful Bias**

ETC. (Global Trends)

EMERGING FRAMEWORKS & GLOBAL STANDARDS

- Increasing international cooperation (G7, OECD)
- Work on global standards (ISO/IEC)
- Common themes of transparency and explainability
- Diverse national approaches (Brazil, India, etc.)

UNITED STATES (USA)

STATE-LEVEL & SECTORAL APPROACH

- **State Privacy Laws** (e.g., CA, CO, TX) with XAI requirements for automated decisions
- **Federal Sectoral Guidance** (FTC, EO 14110 directives)
- Fragmented but growing federal efforts

AUSTRALIA

EXISTING LAWS & GUIDELINES
(No omnibus AI law yet)

- Approach based on **Privacy Act** (updated AI rules) and other laws
- **AI Ethics Framework** (Voluntary Principles) and regulator guidance
- Focus on **safety & transparency**

CHINA

INTERRELATED SECTORAL & CONTENT REGULATIONS

- **Algorithmic Recommendation Regulations** (2021/2022)
- **Generative AI Interim Measures** (2023)
- **Focus on content control, safety assessments, and preventing discrimination**
- **PIPL** provisions for automated decisions

THE PROGRESS OF EXPLAINABLE AI (XAI) SINCE 2016

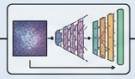
XAI DEVELOPMENT TRAJECTORY

EPOCH 3: 2022-PRESENT: MULTIMODAL, GENERATIVE, & DEPLOYED XAI

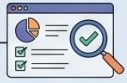
FOCUS: ADVANCED complexity, real-world deployment, and safety.

2022: Explainable LLMs
Chain-of-thought, rationalization

2023: Multimodal XAI
Combined Text+Image Explanation



2024: Generative AI transparency
Diffusion process visualization

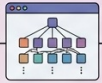


2024: Deployed XAI
Auditing, Governance, Regulatory Compliance (e.g., EU AI Act)

EPOCH 2: 2019-2021: INTERPRETABLE ARCHITECTURES & TRANSPARENCY

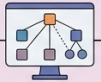
FOCUS: INTRINSIC interpretability and sercentricity.

2019: attention-based models
Self-Attention Visualization



2020: counterfactual explanations
Self-Attention Visualization

2020: counterfactual explanations
What-If Analysis

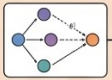


2020: conceptually interpretable models

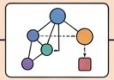
2021: XAI Evaluation Metrics
Benchmarks and Human Studies

FOCUS: INTRINSIC interpretability and user-centricity.

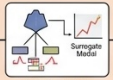
EPOCH 1: 2016-2018: POST-HOC FOUNDATIONS



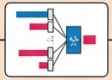
2016: LIME
Local Surrogate Models



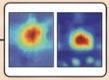
2017: CHAIP
Saliency Maps for CNNs



2016: LIME
Local Surrogate Models



2017: SHAP
Shapley Values / Feature Attribution

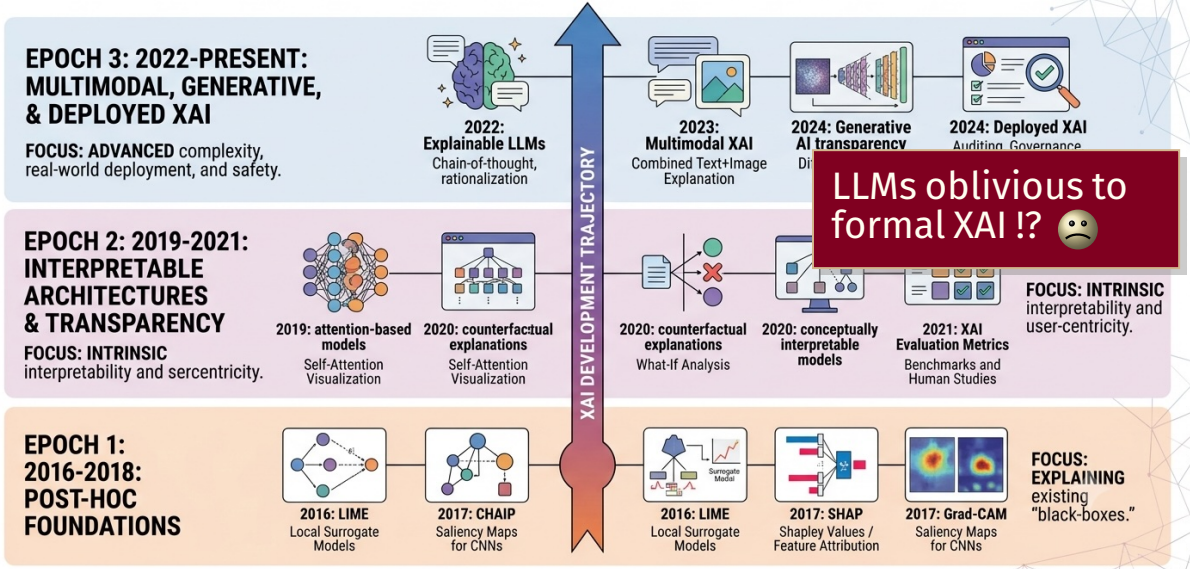


2017: Grad-CAM
Saliency Maps for CNNs

FOCUS: EXPLAINING existing "black-boxes."

Source: Synthesis of major XAI research trends.

THE PROGRESS OF EXPLAINABLE AI (XAI) SINCE 2016



Source: Synthesis of major XAI research trends.

EXPLAINABLE AI (XAI)

Making AI decisions understandable, trustworthy and actionable.

WHAT IS XAI?

Explainable AI aims to open the black box of AI systems by providing human-understandable reasons for their predictions.

WHY IT MATTERS

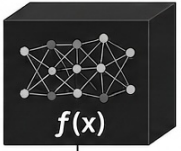
- Builds trust and transparency
- Supports fairness and accountability
- Helps detect errors and bias
- Enables better decisions and user acceptance

INPUT DATA

Age	45
Income	$60K
Education	Master
Credit Score	720
Debt	Low
...	...

BLACK BOX MODEL

(Complex ML/DL Model)



PREDICTION

Loan Approved

(Probability: 0.87)

XAI GOAL

Provide explanations for why the model made a specific prediction. Different explanations answer different needs.

TWO MAIN APPROACHES TO EXPLAIN A BLACK BOX

1 FEATURE SELECTION (Explanations as Rules)

Identify a small set of important features and generate human-readable rules that approximate the black box locally (for this prediction) or globally.

HOW IT WORKS

Select the most relevant features and learn simple rules that mimic the black box.

EXPLANATION (AS RULES)

IF Credit Score ≥ 700
AND Debt = Low
AND Income ≥ $50K

THEN Loan Approved
(Confidence: High)

Benefits: Easy to understand • Actionable • Supports compliance • Useful for communication and decision support

2 FEATURE ATTRIBUTION (Attribution Scores)

Quantify the contribution of each input feature to the specific prediction.

HOW IT WORKS

Estimate how much each feature pushed the prediction up or down from a baseline.

EXPLANATION (ATTRIBUTION SCORES)

Benefits: Shows influence of all features • Detects bias & spurious patterns • Supports debugging & model improvement

DIFFERENT EXPLANATIONS, DIFFERENT PURPOSES – SAME GOAL: UNDERSTAND AND TRUST AI

- **For Users & Decisions** — Use rules to justify and communicate decisions.
- **For Compliance** — Provide transparent, auditable justifications.
- **For Model Improvement** — Use attributions to find bias, errors, and issues.
- **For Domain Experts** — Combine insights with domain knowledge.

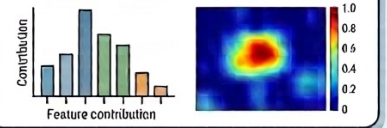
XAI builds trust and enables responsible, human-centered AI.

CATEGORIZING NON-SYMBOLIC XAI

FEATURE ATTRIBUTION

Method: Explains the **contribution** of each feature to a prediction.

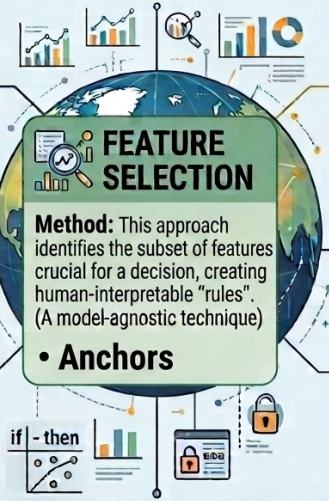
- **LIME** (Local Interpretable Model-agnostic Explanations)
- **SHAP** (SHapley Additive exPlanations)
- **Saliency Maps** (for visual models)



FEATURE SELECTION

Method: This approach identifies the subset of features crucial for a decision, creating human-interpretable "rules". (A model-agnostic technique)

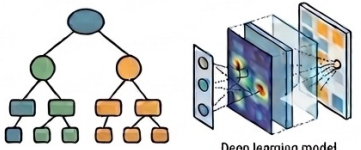
- **Anchors**



INTRINSIC INTERPRETABILITY

Method: Designing models with understandable **internal** logic.

- **DTs** (Decision Trees)
- **DLs** (Decision Lists), ...



⚠ Core premises of non-symbolic XAI disproved ⚠

[MH24, HM24, IIM22, Ign20]

[RSG16, LL17, RSG18, Rud19]

LIME

"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

X

SHAP

A Unified Approach to Interpreting Model Predictions

X

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

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x
nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

X

Cynthia Rudin

Intrinsic Interpretability

Anchors: High-Precision Model-Agnostic Explanations

X

Anchor

Marco Tulio Ribeiro
University of Washington
marcotcr@cs.washington.edu

Sameer Singh
University of California, Irvine
sameer@uci.edu

Carlos Guestrin
University of Washington
guestrin@cs.washington.edu

⚠ Core premises of non-symbolic XAI disproved ⚠

[MH24, HM24, IIM22, Ign20]

[RSG16, LL17, RSG18, Rud19]

LIME

"Why Should I Trust You?"
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

X

SHAP

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

X

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0048-x

nature machine intelligence

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

Cynthia Rudin

X

Intrinsic Interpretability

Anchors: High-Precision Model-Agnostic Explanations

Marco Tulio Ribeiro
University of Washington
marcotcr@cs.washington.edu

Sameer Singh
University of California, Irvine
sameer@uci.edu

Carlos Guestrin
University of Washington
guestrin@cs.washington.edu

X

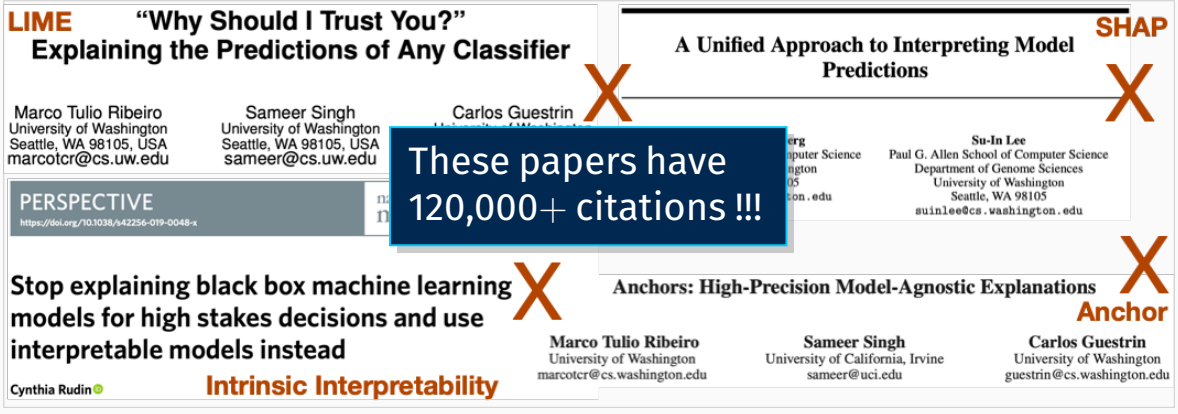
Anchor

Rigorous explainability with LIME/SHAP/Anchors is a myth !!!

⚠ Core premises of non-symbolic XAI disproved ⚠

[MH24, HM24, IIM22, Ign20]

[RSG16, LL17, RSG18, Rud19]



Rigorous explainability with LIME/SHAP/Anchors is a myth !!!

My invited talks at FLoC 2026...

- Invited lecture at SAT/SMT/AR Summer School:
  "Problem Solving with SAT Oracles" Jul 13
- Invited talk at the SAT 2026 conference:
  "Trustable Explainable AI: SAT to the Rescue" Jul 21
- Invited talk at the KR 2026 conference:
  "Rigorous Explainability by Feature Attribution: From SHAP to nuSHAP" Jul 23
- Invited talk at the PCCR 2026 workshop:
  "Logic-Based Explainable AI" Jul 25

My invited talks at FLoC 2026...

- Invited lecture at SAT/SMT/AR Summer School:
  "Problem Solving with SAT Oracles" Jul 13
- Invited talk at the SAT 2026 conference:
  "Trustable Explainable AI: SAT to the Rescue" Jul 21
- Invited talk at the KR 2026 conference:
  "Rigorous Explainability by Feature Attribution: From SHAP to nuSHAP" Jul 23
- Invited talk at the PCCR 2026 workshop:
  "Logic-Based Explainable AI" Jul 25

Feature selection

My invited talks at FLoC 2026...

- Invited lecture at SAT/SMT/AR Summer School:
  "Problem Solving with SAT Oracles" Jul 13
- Invited talk at the SAT 2026 conference:
  "Trustable Explainable AI: SAT to the Rescue" Jul 21
- Invited talk at the KR 2026 conference:
  "Rigorous Explainability by Feature Attribution: From SHAP to nuSHAP" Jul 23
- Invited talk at the PCCR 2026 workshop:
  "Logic-Based Explainable AI" Jul 25

Feature selection

Talks on formal XAI cover the myths of non-symbolic XAI

My invited talks at FLoC 2026... so, rationale for these talks?

- Invited lecture at SAT/SMT/AR Summer School:
  "Problem Solving with SAT Oracles" — Jul 13
- Invited talk at the SAT 2026 conference:
  "Trustable Explainable AI: SAT to the Rescue" — Jul 21
- Invited talk at the KR 2026 conference:
  "Rigorous Explainability by Feature Attribution: From SHAP to nuSHAP" — Jul 23
- Invited talk at the PCCR 2026 workshop:
  "Logic-Based Explainable AI" — Jul 25

Feature selection

Talks on formal XAI cover the myths of non-symbolic XAI

Quoting M. Vardi, from a talk some years ago...

Quoting M. Vardi, from a talk some years ago...

Outline

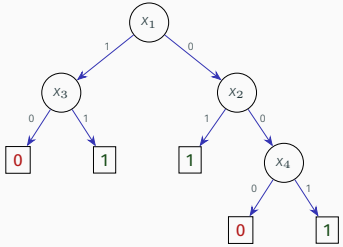
Definitions – classification problems

- Set of features $\mathcal{F} = \{1, 2, \ldots, m\}$, each feature i taking values from domain $\mathbb{D}_i$
  - Features can be categorical, discrete or real-valued
  - Feature space: $\mathbb{F} = \Pi_{i=1}^{m} \mathbb{D}_i$
- Set of classes $\mathcal{K} = \{c_1, \ldots, c_K\}$
- ML model $\mathbb{M}$ computes a (non-constant) classification function $\kappa : \mathbb{F} \to \mathcal{K}$
- Instance $(\mathbf{v}, c)$, for point $\mathbf{v} = (v_1, \ldots, v_m) \in \mathbb{F}$, with prediction $c = \kappa(\mathbf{v})$, $c \in \mathcal{K}$
  - **Goal:** compute explanations for $(\mathbf{v}, c)$

Definitions – classification problems

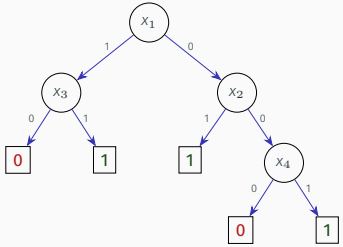
- Set of features $\mathcal{F} = \{1, 2, \ldots, m\}$, each feature i taking values from domain $\mathbb{D}_i$
  - Features can be categorical, discrete or real-valued
  - Feature space: $\mathbb{F} = \Pi_{i=1}^{m} \mathbb{D}_i$
- Set of classes $\mathcal{K} = \{c_1, \ldots, c_K\}$
- ML model $\mathbb{M}$ computes a (non-constant) classification function $\kappa : \mathbb{F} \to \mathcal{K}$
- Instance $(\mathbf{v}, c)$, for point $\mathbf{v} = (v_1, \ldots, v_m) \in \mathbb{F}$, with prediction $c = \kappa(\mathbf{v})$, $c \in \mathcal{K}$
  - **Goal:** compute explanations for $(\mathbf{v}, c)$
- **Framework can be extended to regression problems** [MS24]
  - In this talk: κ replaced by ρ for regression problems...

Symbolic explanations – intuition



- Domains: $\mathbb{D}_i = \{0, 1\}, i = 1, 2, 3, 4$
- Feature space: $\mathbb{F} = \{0, 1\}^4$
- Classes: $\mathcal{K} = \{0, 1\}$
- κ defined by DT
- Instance: $((0, 0, 1, 1), 1)$

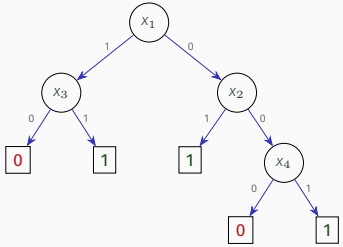
Symbolic explanations – intuition



- Domains: $\mathbb{D}_i = \{0, 1\}, i = 1, 2, 3, 4$
- Feature space: $\mathbb{F} = \{0, 1\}^4$
- Classes: $\mathcal{K} = \{0, 1\}$
- κ defined by DT
- Instance: $((0, 0, 1, 1), 1)$

- **Abductive explanation** answers question "**Why** (the prediction)?" as a **rule**: IF <COND> THEN $\kappa(\mathbf{x}) = c$
  - E.g. IF $\neg x_1 \wedge \neg x_2 \wedge x_4$ THEN $\kappa(\mathbf{x}) = 1$

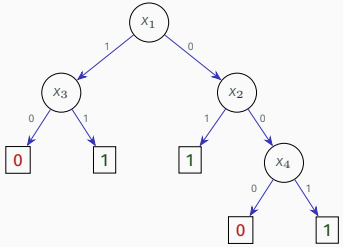
Symbolic explanations – intuition



- Domains: $\mathbb{D}_i = \{0, 1\}, i = 1, 2, 3, 4$
- Feature space: $\mathbb{F} = \{0, 1\}^4$
- Classes: $\mathcal{K} = \{0, 1\}$
- κ defined by DT
- Instance: $((0, 0, 1, 1), 1)$

- **Abductive explanation** answers question "**Why** (the prediction)?" as a **rule**: IF <COND> THEN $\kappa(\mathbf{x}) = c$
  - E.g. IF $\neg x_1 \wedge \neg x_2 \wedge x_4$ THEN $\kappa(\mathbf{x}) = 1$
- **Contrastive explanation** answers question "**Why Not** (some other prediction)?"
  - E.g. $\neg x_1 \wedge \neg x_2 \wedge x_3 \wedge \neg x_4$ is consistent with $\kappa(\mathbf{x}) = 0$

Symbolic explanations – intuition



- Domains: $\mathbb{D}_i = \{0, 1\}, i = 1, 2, 3, 4$
- Feature space: $\mathbb{F} = \{0, 1\}^4$
- Classes: $\mathcal{K} = \{0, 1\}$
- κ defined by DT
- Instance: $((0, 0, 1, 1), 1)$

- **Abductive explanation** answers question "**Why** (the prediction)?" as a **rule**: IF <COND> THEN $\kappa(\mathbf{x}) = c$
  - E.g. IF $\neg x_1 \wedge \neg x_2 \wedge x_4$ THEN $\kappa(\mathbf{x}) = 1$
- **Contrastive explanation** answers question "**Why Not** (some other prediction)?"
  - E.g. $\neg x_1 \wedge \neg x_2 \wedge x_3 \wedge \neg x_4$ is consistent with $\kappa(\mathbf{x}) = 0$
- Explanations represented by sets of features (other details left implicit)
  - E.g. $\{1, 2, 4\}$ or $\{4\}$

Formal explanations – definition

- Instance $(\mathbf{v}, c)$, i.e. $c = \kappa(\mathbf{v})$

Formal explanations – definition

- Instance $(\mathbf{v}, c)$, i.e. $c = \kappa(\mathbf{v})$
- **Weak abductive explanation** (WAXp) – answer to **Why?** question: [SCD18, INM19a]

$$\text{WeakAXp}(\mathcal{X}) \quad :=\quad \forall(\mathbf{x} \in \mathbb{F}).\ \bigwedge\nolimits_{j \in \mathcal{X}} (x_j = v_j) \rightarrow (\kappa(\mathbf{x}) = c)$$

Formal explanations – definition

- Instance $(\mathbf{v}, c)$, i.e. $c = \kappa(\mathbf{v})$
- **Weak abductive explanation** (WAXp) – answer to **Why?** question: [SCD18, INM19a]

$$\text{WeakAXp}(\mathcal{X}) \quad :\equiv \quad \forall(\mathbf{x} \in \mathbb{F}). \bigwedge\nolimits_{j \in \mathcal{X}} (x_j = v_j) \rightarrow (\kappa(\mathbf{x}) = c)$$

- **Weak contrastive explanation** (WCXp) – answer to **Why Not?** question: [Mil19, INAM20]

$$\text{WeakCXp}(\mathcal{Y}) \quad :\equiv \quad \exists(\mathbf{x} \in \mathbb{F}). \bigwedge\nolimits_{j \notin \mathcal{Y}} (x_j = v_j) \land (\kappa(\mathbf{x}) \neq c)$$

Formal explanations – definition

- Instance $(\mathbf{v}, c)$, i.e. $c = \kappa(\mathbf{v})$
- **Weak abductive explanation** (WAXp) – answer to **Why?** question: [SCD18, INM19a]

$$\text{WeakAXp}(\mathcal{X}) \quad := \quad \forall(\mathbf{x} \in \mathbb{F}).\ \bigwedge\nolimits_{j \in \mathcal{X}} (x_j = v_j) \rightarrow (\kappa(\mathbf{x}) = c)$$

- **Weak contrastive explanation** (WCXp) – answer to **Why Not?** question: [Mil19, INAM20]

I.e. adversarial example in ML

$$\text{WeakCXp}(\mathcal{Y}) \quad := \quad \exists(\mathbf{x} \in \mathbb{F}).\ \bigwedge\nolimits_{j \notin \mathcal{Y}} (x_j = v_j) \land (\kappa(\mathbf{x}) \neq c)$$

Formal explanations – definition

- Instance $(\mathbf{v}, c)$, i.e. $c = \kappa(\mathbf{v})$
- **Weak abductive explanation** (WAXp) – answer to **Why?** question: [SCD18, INM19a]

$$\text{WeakAXp}(\mathcal{X}) \quad := \quad \neg\left[\exists(\mathbf{x} \in \mathbb{F}).\bigwedge\nolimits_{j\in\mathcal{X}}(x_j = v_j) \land (\kappa(\mathbf{x}) \neq c)\right]$$

- **Weak contrastive explanation** (WCXp) – answer to **Why Not?** question: [Mil19, INAM20]

I.e. adversarial example in ML

$$\text{WeakCXp}(\mathcal{Y}) \quad := \quad \exists(\mathbf{x} \in \mathbb{F}).\bigwedge\nolimits_{j\notin\mathcal{Y}}(x_j = v_j) \land (\kappa(\mathbf{x}) \neq c)$$

Formal explanations – definition

- Instance $(\mathbf{v}, c)$, i.e. $c = \kappa(\mathbf{v})$
- **Weak abductive explanation** (WAXp) – answer to **Why?** question: [SCD18, INM19a]

$$\text{WeakAXp}(\mathcal{X}) \quad := \quad \neg \left[\exists (\mathbf{x} \in \mathbb{F}). \bigwedge\nolimits_{j \in \mathcal{X}} (x_j = v_j) \land (\kappa(\mathbf{x}) \neq c) \right]$$

- **Weak contrastive explanation** (WCXp) – answer to **Why Not?** question: [Mil19, INAM20]

I.e. adversarial example in ML

$$\text{WeakCXp}(\mathcal{Y}) \quad := \quad \exists (\mathbf{x} \in \mathbb{F}). \bigwedge\nolimits_{j \notin \mathcal{Y}} (x_j = v_j) \land (\kappa(\mathbf{x}) \neq c)$$

- We seek **subset-minimal** sets (i.e. with **no** redundancy):
  - $\text{AXp}(\mathcal{X}) := \text{WeakAXp}(\mathcal{X}) \land \forall (\mathcal{X}' \subsetneq \mathcal{X}). \neg \text{WeakAXp}(\mathcal{X}')$
  - $\text{CXp}(\mathcal{Y}) := \text{WeakCXp}(\mathcal{Y}) \land \forall (\mathcal{Y}' \subsetneq \mathcal{Y}). \neg \text{WeakCXp}(\mathcal{Y}')$

Formal explanations – definition

- Instance $(\mathbf{v}, c)$, i.e. $c = \kappa(\mathbf{v})$
- **Weak abductive explanation** (WAXp) – answer to **Why?** question: [SCD18, INM19a]

$$\text{WeakAXp}(\mathcal{X}) \quad := \quad \neg\left[\exists(\mathbf{x} \in \mathbb{F}).\bigwedge\nolimits_{j\in\mathcal{X}}(x_j = v_j) \land (\kappa(\mathbf{x}) \neq c)\right]$$

- **Weak contrastive explanation** (WCXp) – answer to **Why Not?** question: [Mil19, INAM20]

I.e. adversarial example in ML

$$\text{WeakCXp}(\mathcal{Y}) \quad := \quad \exists(\mathbf{x} \in \mathbb{F}).\bigwedge\nolimits_{j\notin\mathcal{Y}}(x_j = v_j) \land (\kappa(\mathbf{x}) \neq c)$$

- We seek **subset-minimal** sets (i.e. with **no** redundancy):
  - $\text{AXp}(\mathcal{X}) := \text{WeakAXp}(\mathcal{X}) \land \forall(\mathcal{X}' \subsetneq \mathcal{X}).\neg\text{WeakAXp}(\mathcal{X}')$
  - $\text{CXp}(\mathcal{Y}) := \text{WeakCXp}(\mathcal{Y}) \land \forall(\mathcal{Y}' \subsetneq \mathcal{Y}).\neg\text{WeakCXp}(\mathcal{Y}')$
- **Each AXp is minimal hitting set (MHS) of CXps and vice-versa** [INAM20, INM19b]
  - Builds on work of R. Reiter in MBD [Rei87]

Formal explanations – definition

- Instance $(\mathbf{v}, c)$, i.e. $c = \kappa(\mathbf{v})$
- **Weak abductive explanation** (WAXp) – answer to **Why?** question: [SCD18, INM19a]

$$\text{WeakAXp}(\mathcal{X}) \quad := \quad \neg\left[\exists(\mathbf{x} \in \mathbb{F}).\bigwedge\nolimits_{j\in\mathcal{X}}(x_j = v_j) \wedge (\kappa(\mathbf{x}) \neq c)\right]$$

- **Weak contrastive explanation** (WCXp) – answer to **Why Not?** question: [Mil19, INAM20]

I.e. adversarial example in ML

$$\text{WeakCXp}(\mathcal{Y}) \quad := \quad \exists(\mathbf{x} \in \mathbb{F}).\bigwedge\nolimits_{j\notin\mathcal{Y}}(x_j = v_j) \wedge (\kappa(\mathbf{x}) \neq c)$$

- We seek **subset-minimal** sets (i.e. with **no** redundancy):
  - $\text{AXp}(\mathcal{X}) := \text{WeakAXp}(\mathcal{X}) \wedge \forall(\mathcal{X}' \subsetneq \mathcal{X}).\neg\text{WeakAXp}(\mathcal{X}')$
  - $\text{CXp}(\mathcal{Y}) := \text{WeakCXp}(\mathcal{Y}) \wedge \forall(\mathcal{Y}' \subsetneq \mathcal{Y}).\neg\text{WeakCXp}(\mathcal{Y}')$

Goal is to encode NNs, RFs, BTs, etc.

- **Each AXp is minimal hitting set (MHS) of CXps and vice-versa** [INAM20, INM19b]
  - Builds on work of R. Reiter in MBD [Rei87]

Formal explanations – definition

- Instance $(\mathbf{v}, c)$, i.e. $c = \kappa(\mathbf{v})$
- **Weak abductive explanation** (WAXp) – answer to **Why?** question: [SCD18, INM19a]

$$\text{WeakAXp}(\mathcal{X}) \quad := \quad \neg\left[\exists(\mathbf{x} \in \mathbb{F}).\bigwedge\nolimits_{j\in\mathcal{X}}(x_j = v_j) \wedge (\kappa(\mathbf{x}) \neq c)\right]$$

- **Weak contrastive explanation** (WCXp) – answer to **Why Not?** question: [Mil19, INAM20]

I.e. adversarial example in ML

$$\text{WeakCXp}(\mathcal{Y}) \quad := \quad \exists(\mathbf{x} \in \mathbb{F}).\bigwedge\nolimits_{j\notin\mathcal{Y}}(x_j = v_j) \wedge (\kappa(\mathbf{x}) \neq c)$$

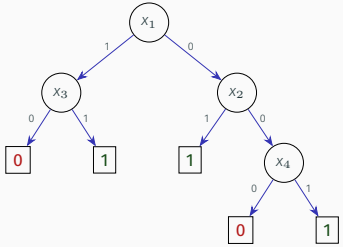
- We seek **subset-minimal** sets (i.e. with **no** redundancy):
  - $\text{AXp}(\mathcal{X}) := \text{WeakAXp}(\mathcal{X}) \wedge \forall(\mathcal{X}' \subsetneq \mathcal{X}).\neg\text{WeakAXp}(\mathcal{X}')$
  - $\text{CXp}(\mathcal{Y}) := \text{WeakCXp}(\mathcal{Y}) \wedge \forall(\mathcal{Y}' \subsetneq \mathcal{Y}).\neg\text{WeakCXp}(\mathcal{Y}')$

Goal is to encode NNs, RFs, BTs, etc.

- **Each AXp is minimal hitting set (MHS) of CXps and vice-versa** [INAM20, INM19b]
  - Builds on work of R. Reiter in MBD [Rei87]

Consistency decided with automated reasoners (SAT, SMT, ...)

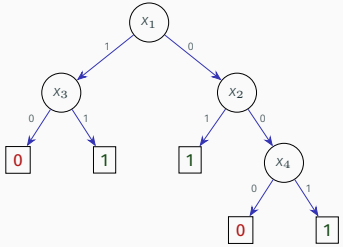
An example



- Domains: $\mathbb{D}_i = \{0, 1\}, i = 1, 2, 3, 4$
- Feature space: $\mathbb{F} = \{0, 1\}^4$
- Classes: $\mathcal{K} = \{0, 1\}$
- κ defined by DT
- Instance: $((0, 0, 1, 1), 1)$

- $\{1, 2, 4\}$ is a weak AXp: $\forall(\mathbf{x} \in \mathbb{F}).(\neg x_1 \land \neg x_2 \land x_4) \rightarrow (\kappa(\mathbf{x}) = 1)$
  - It is not subset-minimal; $\{1, 2, 4\}$ is **not** an AXp
- $\{1, 4\}$ is also a weak AXp, i.e. $\forall(\mathbf{x} \in \mathbb{F}).(\neg x_1 \land x_4) \rightarrow (\kappa(\mathbf{x}) = 1)$
  - And, it is subset-minimal; $\{1, 4\}$ is an AXp
- $\{4\}$ is a weak CXp, i.e. $\exists(\mathbf{x} \in \mathbb{F}).(\neg x_1 \land \neg x_2 \land x_3) \land (\kappa(\mathbf{x}) \neq 1)$
  - E.g. $(0, 0, 1, 0)$
  - And, it is subset-minimal; $\{4\}$ is a CXp

An example

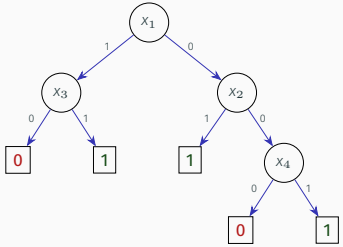


- Domains: $\mathbb{D}_i = \{0, 1\}, i = 1, 2, 3, 4$
- Feature space: $\mathbb{F} = \{0, 1\}^4$
- Classes: $\mathcal{K} = \{0, 1\}$
- κ defined by DT
- Instance: $((0, 0, 1, 1), 1)$

- $\{1, 2, 4\}$ is a weak AXp: $\forall(\mathbf{x} \in \mathbb{F}).(\neg x_1 \land \neg x_2 \land x_4) \rightarrow (\kappa(\mathbf{x}) = 1)$
  - It is not subset-minimal; $\{1, 2, 4\}$ is **not** an AXp
- $\{1, 4\}$ is also a weak AXp, i.e. $\forall(\mathbf{x} \in \mathbb{F}).(\neg x_1 \land x_4) \rightarrow (\kappa(\mathbf{x}) = 1)$
  - And, it is subset-minimal; $\{1, 4\}$ is an AXp
- $\{4\}$ is a weak CXp, i.e. $\exists(\mathbf{x} \in \mathbb{F}).(\neg x_1 \land \neg x_2 \land x_3) \land (\kappa(\mathbf{x}) \neq 1)$
  - E.g. $(0, 0, 1, 0)$
  - And, it is subset-minimal; $\{4\}$ is a CXp

AXps	$\{1, 4\}, \{3, 4\}$
CXps	$\{\{4\}, \{1, 3\}\}$

An example



- Domains: $\mathbb{D}_i = \{0, 1\}, i = 1, 2, 3, 4$
- Feature space: $\mathbb{F} = \{0, 1\}^4$
- Classes: $\mathcal{K} = \{0, 1\}$
- κ defined by DT
- Instance: $((0, 0, 1, 1), 1)$

- $\{1, 2, 4\}$ is a weak AXp: $\forall(\mathbf{x} \in \mathbb{F}).(\neg x_1 \land \neg x_2 \land x_4) \rightarrow (\kappa(\mathbf{x}) = 1)$
  - It is not subset-minimal; $\{1, 2, 4\}$ is **not** an AXp
- $\{1, 4\}$ is also a weak AXp, i.e. $\forall(\mathbf{x} \in \mathbb{F}).(\neg x_1 \land x_4) \rightarrow (\kappa(\mathbf{x}) = 1)$
  - And, it is subset-minimal; $\{1, 4\}$ is an AXp
- $\{4\}$ is a weak CXp, i.e. $\exists(\mathbf{x} \in \mathbb{F}).(\neg x_1 \land \neg x_2 \land x_3) \land (\kappa(\mathbf{x}) \neq 1)$
  - E.g. $(0, 0, 1, 0)$
  - And, it is subset-minimal; $\{4\}$ is a CXp

AXps	$\{1, 4\}, \{3, 4\}$
CXps	$\{\{4\}, \{1, 3\}\}$

Hitting-set duality used ubiquitously!

Logic-based abduction vs. abductive explanations

Logic-based abduction

[EG95]

- Tuple: $\mathcal{P} = \langle V, H, M, R \rangle$
  - V: Finite set of propositional variables
  - H: Set of hypotheses
  - $M \subseteq V$: Set of manifestations
  - R: Consistent theory
- A solution for $\mathcal{P}$ is subset-minimal $S \subseteq H$ such that,
  1. $R \cup S \not\models \bot$, and
  2. $R \cup S \models M$

Logic-based abduction vs. abductive explanations

Logic-based abduction [EG95]

- Tuple: $\mathcal{P} = \langle V, H, M, R \rangle$
  - V: Finite set of propositional variables
  - H: Set of hypotheses
  - $M \subseteq V$: Set of manifestations
  - R: Consistent theory
- A solution for $\mathcal{P}$ is subset-minimal $S \subseteq H$ such that,
  1. $R \cup S \not\models \bot$, and
  2. $R \cup S \models M$

AXps formulated as logic-based abduction [INM19a]

- Define $h_j \rightarrow (x_j = v_j), j \in \mathcal{F}$, and $o \leftrightarrow (\kappa(\mathbf{x}) = c)$; and let $H \triangleq \{h_j \mid j \in \mathcal{F}\}$, $M \triangleq \{o\}$
- Let $R \triangleq \wedge_{j \in \mathcal{F}} \text{Encode}(h_j \rightarrow (x_j = v_j) \wedge \text{Encode}(o \leftrightarrow (\kappa(\mathbf{x}) = c))$, with $V \triangleq \text{vars}(R)$, $M \triangleq \{o\}$
- Then, a solution S to $\mathcal{P} = \langle V, H, M, R \rangle$ is such that
  1. $R \cup S \not\models \bot$: holds trivially, any partial input consistent with ML model
  2. $R \cup S \models M$: prediction sufficiency holds
- Thus, an AXp is a solution to a logic-based abduction problem

Logic-based abduction vs. abductive explanations

Logic-based abduction [EG95]

- Tuple: $\mathcal{P} = \langle V, H, M, R \rangle$
  - V: Finite set of propositional variables
  - H: Set of hypotheses
  - $M \subseteq V$: Set of manifestations
  - R: Consistent theory
- A solution for $\mathcal{P}$ is subset-minimal $S \subseteq H$ such that,
  1. $R \cup S \not\models \bot$, and
  2. $R \cup S \models M$

AXps formulated as logic-based abduction [INM19a]

- Define $h_j \rightarrow (x_j = v_j), j \in \mathcal{F}$, and $o \leftrightarrow (\kappa(\mathbf{x}) = c)$; and let $H \triangleq \{h_j \,|\, j \in \mathcal{F}\}$, $M \triangleq \{o\}$
- Let $R \triangleq \wedge_{j \in \mathcal{F}} \text{Encode}(h_j \rightarrow (x_j = v_j) \wedge \text{Encode}(o \leftrightarrow (\kappa(\mathbf{x}) = c))$, with $V \triangleq \text{vars}(R)$, $M \triangleq \{o\}$
- Then, a solution S to $\mathcal{P} = \langle V, H, M, R \rangle$ is such that
  1. $R \cup S \not\models \bot$: holds trivially, any partial input consistent with ML model
  2. $R \cup S \models M$: prediction sufficiency holds
- Thus, an AXp is a solution to a logic-based abduction problem

Can also be formulated as MUSes/MCSes, or prime implicants, or ...

Logic-based abduction vs. abductive explanations

Logic-based abduction

AXps, PI-explanations, sufficient reasons, ... [EG95]

- Tuple: $\mathcal{P} = \langle V, H, M, R \rangle$
  - V: Finite set of propositional variables
  - H: Set of hypotheses
  - $M \subseteq V$: Set of manifestations
  - R: Consistent theory
- A solution for $\mathcal{P}$ is subset-minimal $S \subseteq H$ such that,
  1. $R \cup S \not\models \bot$, and
  2. $R \cup S \models M$

AXps formulated as logic-based abduction [INM19a]

- Define $h_j \rightarrow (x_j = v_j), j \in \mathcal{F}$, and $o \leftrightarrow (\kappa(\mathbf{x}) = c)$; and let $H \triangleq \{h_j \,|\, j \in \mathcal{F}\}$, $M \triangleq \{o\}$
- Let $R \triangleq \wedge_{j \in \mathcal{F}} \text{Encode}(h_j \rightarrow (x_j = v_j) \wedge \text{Encode}(o \leftrightarrow (\kappa(\mathbf{x}) = c))$, with $V \triangleq \text{vars}(R)$, $M \triangleq \{o\}$
- Then, a solution S to $\mathcal{P} = \langle V, H, M, R \rangle$ is such that
  1. $R \cup S \not\models \bot$: holds trivially, any partial input consistent with ML model
  2. $R \cup S \models M$: prediction sufficiency holds
- Thus, an AXp is a solution to a logic-based abduction problem

Can also be formulated as MUSes/MCSes, or prime implicants, or ...

Logic-based abduction vs. abductive explanations

Logic-based abduction [EG95]

AXps, PI-explanations, sufficient reasons, ...

- Tuple: $\mathcal{P} = \langle V, H, M, R \rangle$
  - V: Finite set of propositional variables
  - H: Set of hypotheses
  - $M \subseteq V$: Set of manifestations
  - R: Consistent theory
- A solution for $\mathcal{P}$ is subset-minimal $S \subseteq H$ such that,
  1. $R \cup S \not\models \bot$, and
  2. $R \cup S \models M$

Should we stop splitting hairs and work to making an impact in ML?

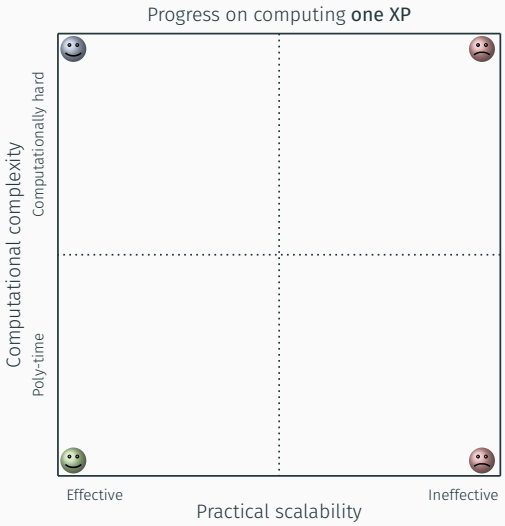
AXps formulated as logic-based abduction [INM19a]

- Define $h_j \rightarrow (x_j = v_j), j \in \mathcal{F}$, and $o \leftrightarrow (\kappa(\mathbf{x}) = c)$; and let $H \triangleq \{h_j \mid j \in \mathcal{F}\}$, $M \triangleq \{o\}$
- Let $R \triangleq \wedge_{j \in \mathcal{F}} \text{Encode}(h_j \rightarrow (x_j = v_j) \wedge \text{Encode}(o \leftrightarrow (\kappa(\mathbf{x}) = c))$, with $V \triangleq \text{vars}(R)$, $M \triangleq \{o\}$
- Then, a solution S to $\mathcal{P} = \langle V, H, M, R \rangle$ is such that
  1. $R \cup S \not\models \bot$: holds trivially, any partial input consistent with ML model
  2. $R \cup S \models M$: prediction sufficiency holds
- Thus, an AXp is a solution to a logic-based abduction problem

Can also be formulated as MUSes/MCSes, or prime implicants, or ...

Progress in symbolic XAI – until 2022

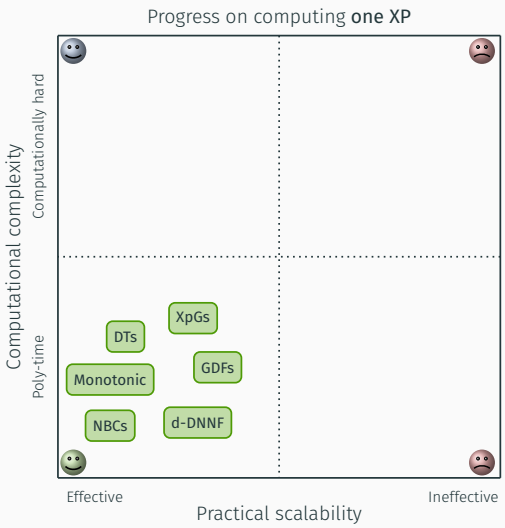
[INM19c, IIM20, MGC+20, MGC+21, HIIM21, IMS21, IM21, CM21, IIM22, HII+22, IISMS22]



- Advances between 2018 and 2022:

Progress in symbolic XAI – until 2022

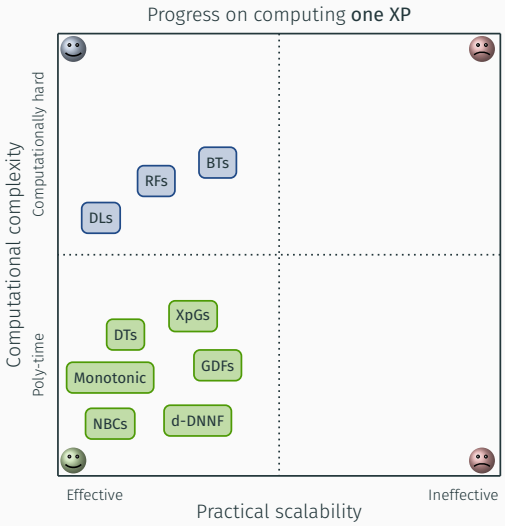
[INM19c, IIM20, MGC+20, MGC+21, HIIM21, IMS21, IM21, CM21, IIM22, HII+22, IISMS22]



- Advances between 2018 and 2022:
  - Polynomial-time:
    - Naive-Bayes classifiers (NBCs) [MGC+20]
    - Decision trees (DTs) [IIM20, HIIM21, IIM22]
    - XpG's: DTs, OBDDs, OMDDs, etc. [HIIM21]
    - Monotonic classifiers [MGC+21]
    - Propositional languages (e.g. d-DNNF, ...) [HII+22]
    - Additional results [CM21, HII+22, CM23, BKR+25]

Progress in symbolic XAI – until 2022

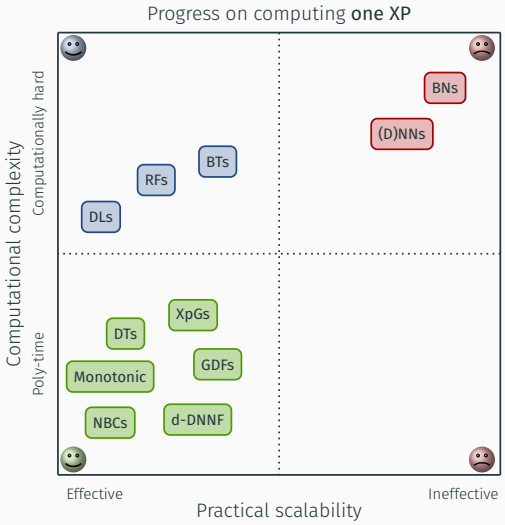
[INM19c, IIM20, MGC+20, MGC+21, HIIM21, IMS21, IM21, CM21, IIM22, HII+22, IISMS22]



- Advances between 2018 and 2022:
  - Polynomial-time:
    - Naive-Bayes classifiers (NBCs) [MGC+20]
    - Decision trees (DTs) [IIM20, HIIM21, IIM22]
    - XpG's: DTs, OBDDs, OMDDs, etc. [HIIM21]
    - Monotonic classifiers [MGC+21]
    - Propositional languages (e.g. d-DNNF, ...) [HII+22]
    - Additional results [CM21, HII+22, CM23, BKR+25]
  - Comp. hard, but effective (efficient in practice):
    - Random forests (RFs) [IMS21, IISMS22]
    - Decision lists (DLs) [IM21]
    - Boosted trees (BTs) [INM19c, IISMS22]

Progress in symbolic XAI – until 2022

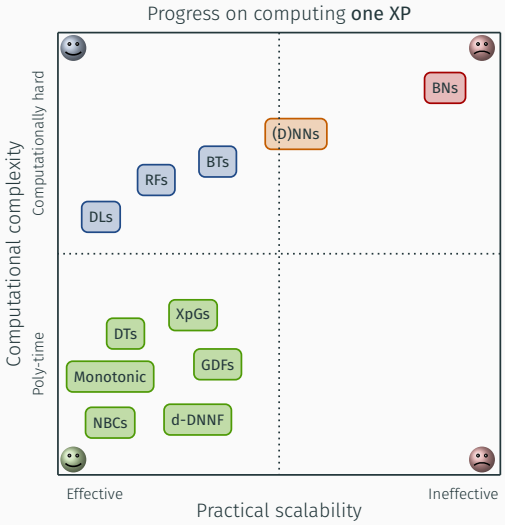
[INM19c, IIM20, MGC+20, MGC+21, HIIM21, IMS21, IM21, CM21, IIM22, HII+22, IISMS22]



- Advances between 2018 and 2022:
  - Polynomial-time:
    - Naive-Bayes classifiers (NBCs) [MGC+20]
    - Decision trees (DTs) [IIM20, HIIM21, IIM22]
    - XpG's: DTs, OBDDs, OMDDs, etc. [HIIM21]
    - Monotonic classifiers [MGC+21]
    - Propositional languages (e.g. d-DNNF, ...) [HII+22]
    - Additional results [CM21, HII+22, CM23, BKR+25]
  - Comp. hard, but effective (efficient in practice):
    - Random forests (RFs) [IMS21, IISMS22]
    - Decision lists (DLs) [IM21]
    - Boosted trees (BTs) [INM19c, IISMS22]
  - Comp. hard, and ineffective (hard in practice):
    - Neural networks (NNs) [INM19a]
    - Bayesian networks (BNs) [SCD18]

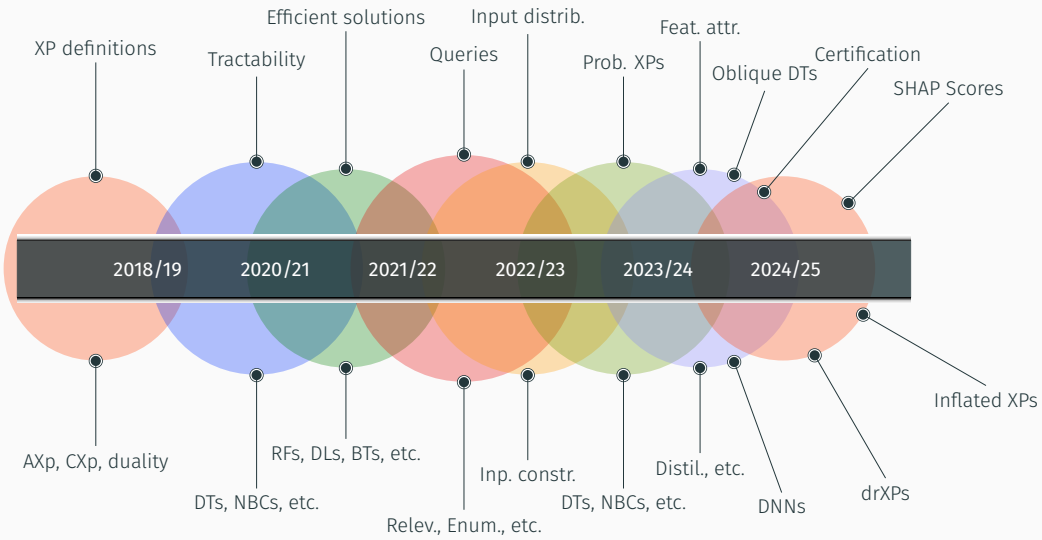
Progress in symbolic XAI – recent progress

[INM19c, IIM20, MGC+20, MGC+21, HIIM21, IMS21, IM21, CM21, IIM22, HII+22, IISMS22, HM23a, IHM+24a]

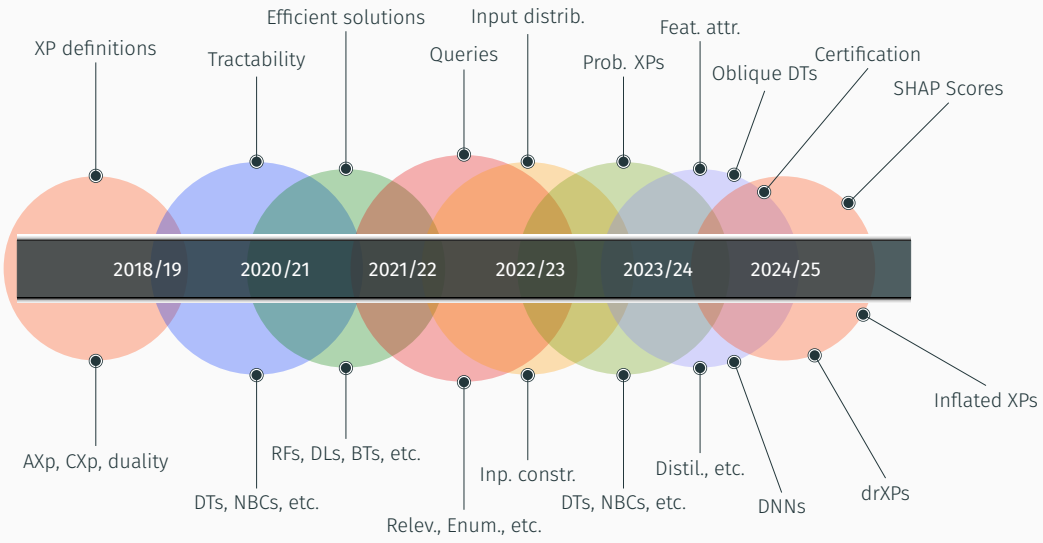


- Advances up until 2024:
  - Polynomial-time:
    - Naive-Bayes classifiers (NBCs) [MGC+20]
    - Decision trees (DTs) [IIM20, HIIM21, IIM22]
    - XpG's: DTs, OBDDs, OMDDs, etc. [HIIM21]
    - Monotonic classifiers [MGC+21]
    - Propositional languages (e.g. d-DNNF, ...) [HII+22]
    - Additional results [CM21, HII+22, CM23, BKR+25]
  - Comp. hard, but effective (efficient in practice):
    - Random forests (RFs) [IMS21, IISMS22]
    - Decision lists (DLs) [IM21]
    - Boosted trees (BTs) [INM19c, IISMS22]
  - Comp. hard, but some practical scalability:
    - Neural networks (NNs) [HM23a, IHM+24a, IHM+24b]
  - Comp. hard, and ineffective (hard in practice):
    - Bayesian networks (BNs) [SCD18]

The emergence of symbolic explainability – timeline



The emergence of symbolic explainability – timeline



Since 2025: Corrected SHAP Scores, Sample-based Xps, Certification, Rule-ensembles, other ML models, etc.

A key ingredient – feature relevancy

[HIIM21]

- Consider instance $(\mathbf{v}, c)$
- Sets of all AXp's & CXp's:

$$\mathbb{A} \coloneqq \{\mathcal{X} \subseteq \mathcal{F} \,|\, \mathsf{AXp}(\mathcal{X})\}$$

$$\mathbb{C} \coloneqq \{\mathcal{X} \subseteq \mathcal{F} \,|\, \mathsf{CXp}(\mathcal{X})\}$$

A key ingredient – feature relevancy

- Consider instance $(\mathbf{v}, c)$
- Sets of all AXp's & CXp's:

$$\mathbb{A} \coloneqq \{\mathcal{X} \subseteq \mathcal{F} \,|\, \mathsf{AXp}(\mathcal{X})\}$$

$$\mathbb{C} \coloneqq \{\mathcal{X} \subseteq \mathcal{F} \,|\, \mathsf{CXp}(\mathcal{X})\}$$

- Features occurring in some AXp in $\mathbb{A}$ and in some CXp in $\mathbb{C}$:

$$F_{\mathbb{A}} \coloneqq \bigcup_{\mathcal{X} \in \mathbb{A}} \mathcal{X}$$

$$F_{\mathbb{C}} \coloneqq \bigcup_{\mathcal{X} \in \mathbb{C}} \mathcal{X}$$

A key ingredient – feature relevancy

- Consider instance $(\mathbf{v}, c)$
- Sets of all AXp's & CXp's:

$$\mathbb{A} \coloneqq \{\mathcal{X} \subseteq \mathcal{F} \,|\, \mathsf{AXp}(\mathcal{X})\}$$

$$\mathbb{C} \coloneqq \{\mathcal{X} \subseteq \mathcal{F} \,|\, \mathsf{CXp}(\mathcal{X})\}$$

- Features occurring in some AXp in $\mathbb{A}$ and in some CXp in $\mathbb{C}$:

$$F_{\mathbb{A}} \coloneqq \bigcup_{\mathcal{X} \in \mathbb{A}} \mathcal{X}$$

$$F_{\mathbb{C}} \coloneqq \bigcup_{\mathcal{X} \in \mathbb{C}} \mathcal{X}$$

- **Claim**: $F_{\mathbb{A}} = F_{\mathbb{C}}$
  - I.e. a feature exists in some AXp iff it exists in some CXp

A key ingredient – feature relevancy

[HIIM21]

- Consider instance $(\mathbf{v}, c)$
- Sets of all AXp's & CXp's:

$$\mathbb{A} \coloneqq \{\mathcal{X} \subseteq \mathcal{F} \,|\, \mathsf{AXp}(\mathcal{X})\}$$

$$\mathbb{C} \coloneqq \{\mathcal{X} \subseteq \mathcal{F} \,|\, \mathsf{CXp}(\mathcal{X})\}$$

- Features occurring in some AXp in $\mathbb{A}$ and in some CXp in $\mathbb{C}$:

$$F_{\mathbb{A}} \coloneqq \bigcup_{\mathcal{X} \in \mathbb{A}} \mathcal{X}$$

$$F_{\mathbb{C}} \coloneqq \bigcup_{\mathcal{X} \in \mathbb{C}} \mathcal{X}$$

- **Claim**: $F_{\mathbb{A}} = F_{\mathbb{C}}$
  - I.e. a feature exists in some AXp iff it exists in some CXp
- A feature $i \in \mathcal{F}$ is **relevant** if $i \in F_{\mathbb{A}}$ (and so, if $i \in F_{\mathbb{C}}$)
  - A feature is relevant if it is included in some AXp (or CXp)

A key ingredient – feature relevancy

[HIIM21]

- Consider instance $(\mathbf{v}, c)$
- Sets of all AXp's & CXp's:

$$\mathbb{A} \coloneqq \{\mathcal{X} \subseteq \mathcal{F} \,|\, \mathsf{AXp}(\mathcal{X})\}$$

$$\mathbb{C} \coloneqq \{\mathcal{X} \subseteq \mathcal{F} \,|\, \mathsf{CXp}(\mathcal{X})\}$$

- Features occurring in some AXp in $\mathbb{A}$ and in some CXp in $\mathbb{C}$:

$$F_{\mathbb{A}} \coloneqq \bigcup_{\mathcal{X} \in \mathbb{A}} \mathcal{X}$$

$$F_{\mathbb{C}} \coloneqq \bigcup_{\mathcal{X} \in \mathbb{C}} \mathcal{X}$$

- **Claim**: $F_{\mathbb{A}} = F_{\mathbb{C}}$
  - I.e. a feature exists in some AXp iff it exists in some CXp
- A feature $i \in \mathcal{F}$ is **relevant** if $i \in F_{\mathbb{A}}$ (and so, if $i \in F_{\mathbb{C}}$)
  - A feature is relevant if it is included in some AXp (or CXp)
- A feature $i \in \mathcal{F}$ is **irrelevant** if $i \notin F_{\mathbb{A}}$ (and so, if $i \notin F_{\mathbb{C}}$)
  - A feature is irrelevant if it is not included in any AXp (or CXp)

An example

- Consider the classifier:

$$\kappa(x_1, x_2, x_3, x_4) = \bigvee_{i=1}^{4} x_i$$

  - $(\mathbf{v}, c) = ((0, 0, 0, 1), 1)$

An example

- Consider the classifier:

$$\kappa(x_1, x_2, x_3, x_4) = \bigvee_{i=1}^{4} x_i$$

  - $(\mathbf{v}, c) = ((0, 0, 0, 1), 1)$

- $\mathbb{A} = \{\{4\}\} = \mathbb{C}$
  - Why?
    - If 4 fixed, then prediction *must* be 1
    - If 4 is allowed to change, then prediction changes
    - Values of 1, 2, 3 not used to fix/change the prediction

An example

- Consider the classifier:

$$\kappa(x_1, x_2, x_3, x_4) = \bigvee_{i=1}^{4} x_i$$

  - $(\mathbf{v}, c) = ((0, 0, 0, 1), 1)$

- $\mathbb{A} = \{\{4\}\} = \mathbb{C}$
  - Why?
    - If 4 fixed, then prediction *must* be 1
    - If 4 is allowed to change, then prediction changes
    - Values of 1, 2, 3 not used to fix/change the prediction

- Feature 4 is relevant, since it is included in one (and the only) AXp/CXp

An example

- Consider the classifier:

$$\kappa(x_1, x_2, x_3, x_4) = \bigvee_{i=1}^{4} x_i$$

  - $(\mathbf{v}, c) = ((0, 0, 0, 1), 1)$

- $\mathbb{A} = \{\{4\}\} = \mathbb{C}$
  - Why?
    - If 4 fixed, then prediction *must* be 1
    - If 4 is allowed to change, then prediction changes
    - Values of 1, 2, 3 not used to fix/change the prediction

- Feature 4 is relevant, since it is included in one (and the only) AXp/CXp

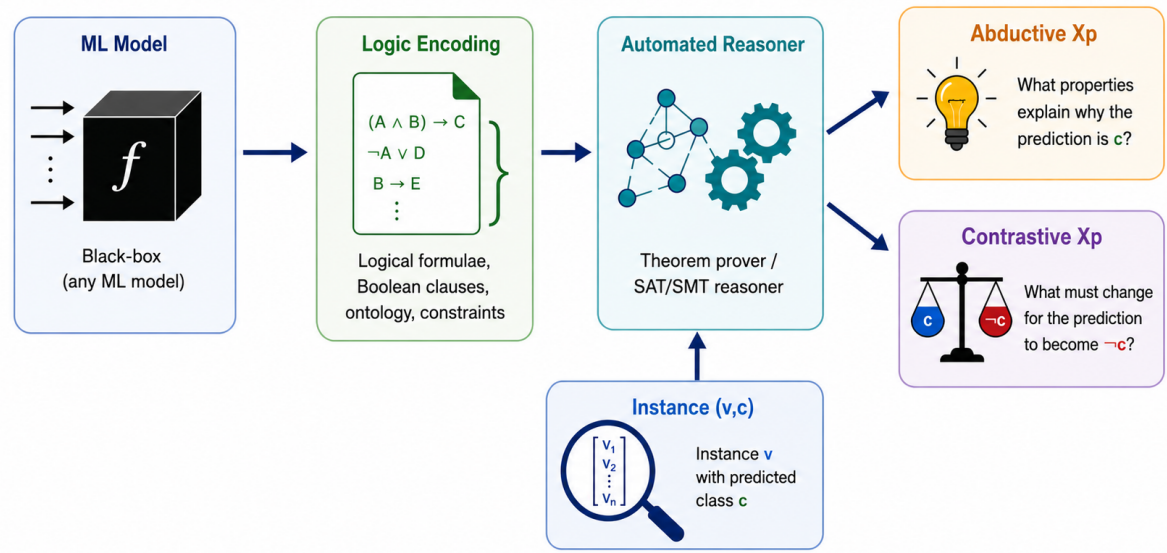
- Features 1, 2, 3 are irrelevant, since there are not included in any AXp/CXp

  - Obs: irrelevant features are absolutely unimportant!

    We could propose some other explanation by adding features 1, 2 or 3 to AXp $\{4\}$, but prediction would remain unchanged for any value assigned to those features

    - And we aim for irreducibility (Occam's razor is a mainstay of AI/ML)

Model-Aware Explanations

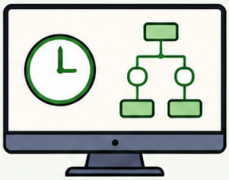


Model-Aware Explanations – Poly-Time

ML Model

Black-box
(any ML model)

Poly-time Algorithm



Polynomial-time
algorithm that computes
explanations

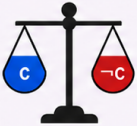
Instance (v, c)

Instance **v**
with predicted
class **c**

Abductive Xp

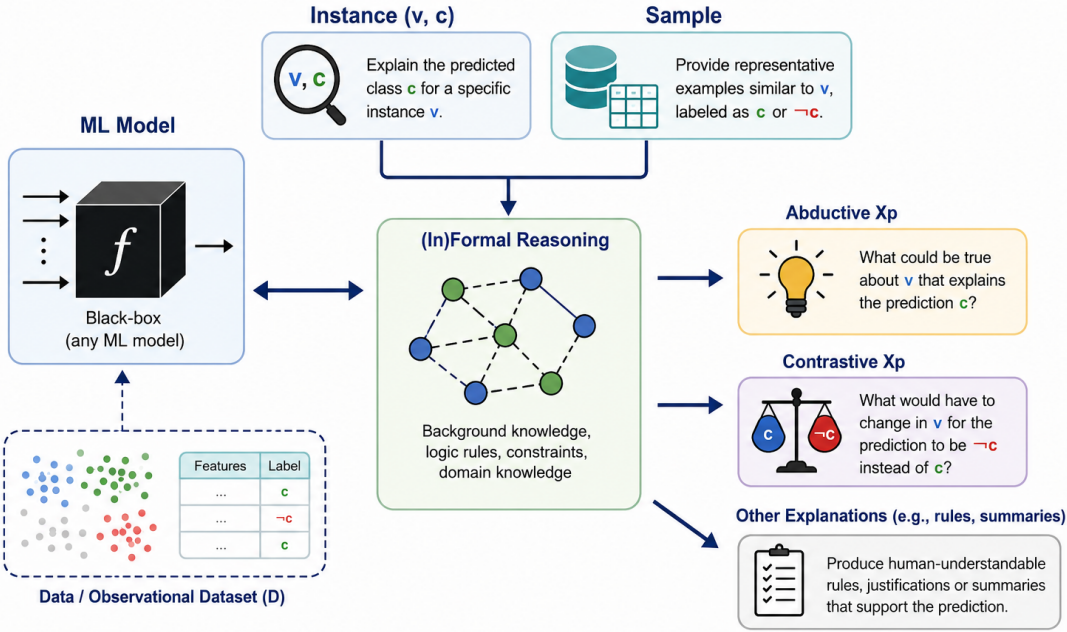
What properties
explain why the
prediction is **c**?

Contrastive Xp



What must change
for the prediction
to become ¬c?

Model-Agnostic Explanations



Another key ingredient – model-agnostic sample-based XAI

[MLM25]

- A sample $\mathbb{S}$ is a subset of feature space $\mathbb{F}$; a prediction p_i is associated with each $\mathbf{v}_i \in \mathbb{S}$
- Target instance: $(\mathbf{v}, c)$, WLOG included in $\mathbb{S}$

Another key ingredient – model-agnostic sample-based XAI

[MLM25]

- A sample $\mathbb{S}$ is a subset of feature space $\mathbb{F}$; a prediction p_i is associated with each $\mathbf{v}_i \in \mathbb{S}$
- Target instance: $(\mathbf{v}, c)$, WLOG included in $\mathbb{S}$
- **Sample-based weak abductive explanation** (sbWAXp, WAXp)

$$\text{WeakAXp}(\mathcal{X}) \quad := \quad \forall(\mathbf{x} \in \mathbb{S}). \bigwedge\nolimits_{j \in \mathcal{X}} (x_j = v_j) \rightarrow (\kappa(\mathbf{x}) = c)$$

- **Sample-based weak contrastive explanation** (sbWCXp, WCXp)

$$\text{WeakCXp}(\mathcal{Y}) \quad := \quad \exists(\mathbf{x} \in \mathbb{S}). \bigwedge\nolimits_{j \notin \mathcal{Y}} (x_j = v_j) \land (\kappa(\mathbf{x}) \neq c)$$

Another key ingredient – model-agnostic sample-based XAI

[MLM25]

- A sample $\mathbb{S}$ is a subset of feature space $\mathbb{F}$; a prediction p_i is associated with each $\mathbf{v}_i \in \mathbb{S}$
- Target instance: $(\mathbf{v}, c)$, WLOG included in $\mathbb{S}$
- **Sample-based weak abductive explanation** (sbWAXp, WAXp)

$$\text{WeakAXp}(\mathcal{X}) \quad := \quad \forall(\mathbf{x} \in \mathbb{S}). \bigwedge\nolimits_{j \in \mathcal{X}} (x_j = v_j) \rightarrow (\kappa(\mathbf{x}) = c)$$

- **Sample-based weak contrastive explanation** (sbWCXp, WCXp)

$$\text{WeakCXp}(\mathcal{Y}) \quad := \quad \exists(\mathbf{x} \in \mathbb{S}). \bigwedge\nolimits_{j \notin \mathcal{Y}} (x_j = v_j) \land (\kappa(\mathbf{x}) \neq c)$$

- Each AXp (resp. CXp) is a subset-minimal WAXp (resp. WCXp) with respect to the predicate

Another key ingredient – model-agnostic sample-based XAI

[MLM25]

- A sample $\mathbb{S}$ is a subset of feature space $\mathbb{F}$; a prediction p_i is associated with each $\mathbf{v}_i \in \mathbb{S}$
- Target instance: $(\mathbf{v}, c)$, WLOG included in $\mathbb{S}$
- **Sample-based weak abductive explanation** (sbWAXp, WAXp)

$$\text{WeakAXp}(\mathcal{X}) \quad := \quad \forall(\mathbf{x} \in \mathbb{S}). \bigwedge\nolimits_{j \in \mathcal{X}} (x_j = v_j) \rightarrow (\kappa(\mathbf{x}) = c)$$

- **Sample-based weak contrastive explanation** (sbWCXp, WCXp)

$$\text{WeakCXp}(\mathcal{Y}) \quad := \quad \exists(\mathbf{x} \in \mathbb{S}). \bigwedge\nolimits_{j \notin \mathcal{Y}} (x_j = v_j) \land (\kappa(\mathbf{x}) \neq c)$$

- Each AXp (resp. CXp) is a subset-minimal WAXp (resp. WCXp) with respect to the predicate
- Literals with relational operators other than $=$ can be used
  - Same applies to the case of model-aware AXps/CXps

[IISM24]

Another key ingredient – model-agnostic sample-based XAI

[MLM25]

- A sample $\mathbb{S}$ is a subset of feature space $\mathbb{F}$; a prediction p_i is associated with each $\mathbf{v}_i \in \mathbb{S}$
- Target instance: $(\mathbf{v}, c)$, WLOG included in $\mathbb{S}$
- **Sample-based weak abductive explanation** (sbWAXp, WAXp)

$$\text{WeakAXp}(\mathcal{X}) \quad := \quad \forall(\mathbf{x} \in \mathbb{S}). \bigwedge\nolimits_{j \in \mathcal{X}} (x_j = v_j) \rightarrow (\kappa(\mathbf{x}) = c)$$

- **Sample-based weak contrastive explanation** (sbWCXp, WCXp)

$$\text{WeakCXp}(\mathcal{Y}) \quad := \quad \exists(\mathbf{x} \in \mathbb{S}). \bigwedge\nolimits_{j \notin \mathcal{Y}} (x_j = v_j) \land (\kappa(\mathbf{x}) \neq c)$$

- Each AXp (resp. CXp) is a subset-minimal WAXp (resp. WCXp) with respect to the predicate
- Literals with relational operators other than $=$ can be used
  - Same applies to the case of model-aware AXps/CXps

[IISM24]

NOTE: the sample can be your training data and/or you can sample the ML model as much as you wish!

Example of a sample, with instance $((1, 1, 1, 1), 1)$

#	x_1	x_2	x_3	x_4	$\kappa(\cdot)$
00	0	0	0	0	0
02	0	0	1	0	0
03	0	0	1	1	0
06	0	1	1	0	0
08	1	0	0	0	1
12	1	1	0	0	1
15	1	1	1	1	1

- How can we change prediction, given the **sample**?
- Move from row id 15 to one row taken out of $\{00, 02, 03, 06\}$

- Reduced matrix, with instances having prediction other than 1

Example of a sample, with instance $((1, 1, 1, 1), 1)$

#	x_1	x_2	x_3	x_4	$\kappa(\cdot)$
00	0	0	0	0	0
02	0	0	1	0	0
03	0	0	1	1	0
06	0	1	1	0	0
08	1	0	0	0	1
12	1	1	0	0	1
15	1	1	1	1	1

- How can we change prediction, given the **sample**?
- Move from row id 15 to one row taken out of $\{00, 02, 03, 06\}$

- Reduced matrix, with instances having prediction other than 1

$$\begin{matrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \end{matrix}$$

Example of a sample, with instance $((1, 1, 1, 1), 1)$

#	x_1	x_2	x_3	x_4	$\kappa(\cdot)$
00	0	0	0	0	0
02	0	0	1	0	0
03	0	0	1	1	0
06	0	1	1	0	0
08	1	0	0	0	1
12	1	1	0	0	1
15	1	1	1	1	1

- How can we change prediction, given the **sample**?
- Move from row id 15 to one row taken out of $\{00, 02, 03, 06\}$

- What do we need to change to allow changing the prediction, which are the WCXps?

$$
\begin{array}{cccc}
0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 \\
0 & 1 & 1 & 0
\end{array}
$$

Example of a sample, with instance $((1, 1, 1, 1), 1)$

#	x_1	x_2	x_3	x_4	$\kappa(\cdot)$
00	0	0	0	0	0
02	0	0	1	0	0
03	0	0	1	1	0
06	0	1	1	0	0
08	1	0	0	0	1
12	1	1	0	0	1
15	1	1	1	1	1

- How can we change prediction, given the **sample**?
- Move from row id 15 to one row taken out of $\{00, 02, 03, 06\}$

- What do we need to change to allow changing the prediction, which are the WCXps?

1 1 1 1
1 1 0 1
1 1 0 0
1 0 0 1

Example of a sample, with instance $((1,1,1,1),1)$

#	x_1	x_2	x_3	x_4	$\kappa(\cdot)$
00	0	0	0	0	0
02	0	0	1	0	0
03	0	0	1	1	0
06	0	1	1	0	0
08	1	0	0	0	1
12	1	1	0	0	1
15	1	1	1	1	1

- How can we change prediction, given the **sample**?
- Move from row id 15 to one row taken out of $\{00, 02, 03, 06\}$

- Which WCXps are subset-minimal, i.e. which are CXps?

1 1 1 1
1 1 0 1
1 1 0 0
1 0 0 1

Example of a sample, with instance $((1, 1, 1, 1), 1)$

#	x_1	x_2	x_3	x_4	$\kappa(\cdot)$
00	0	0	0	0	0
02	0	0	1	0	0
03	0	0	1	1	0
06	0	1	1	0	0
08	1	0	0	0	1
12	1	1	0	0	1
15	1	1	1	1	1

- How can we change prediction, given the **sample**?
- Move from row id 15 to one row taken out of $\{00, 02, 03, 06\}$

- Which WCXps are subset-minimal, i.e. which are CXps?

1 1 0 0
1 0 0 1

Example of a sample, with instance $((1, 1, 1, 1), 1)$

#	x_1	x_2	x_3	x_4	$\kappa(\cdot)$
00	0	0	0	0	0
02	0	0	1	0	0
03	0	0	1	1	0
06	0	1	1	0	0
08	1	0	0	0	1
12	1	1	0	0	1
15	1	1	1	1	1

- How can we change prediction, given the **sample**?
- Move from row id 15 to one row taken out of $\{00, 02, 03, 06\}$

- Two CXps: $\{1, 2\}$ and $\{1, 4\}$

1 1 0 0
1 0 0 1

Example of a sample, with instance $((1,1,1,1),1)$

#	x_1	x_2	x_3	x_4	$\kappa(\cdot)$
00	0	0	0	0	0
02	0	0	1	0	0
03	0	0	1	1	0
06	0	1	1	0	0
08	1	0	0	0	1
12	1	1	0	0	1
15	1	1	1	1	1

- How can we change prediction, given the **sample**?
- Move from row id 15 to one row taken out of $\{00, 02, 03, 06\}$

- Two CXps: $\{1,2\}$ and $\{1,4\}$

1 1 0 0
1 0 0 1

- Number of (W)CXps not larger than sample, and so polynomial on the input!

Example of a sample, with instance $((1, 1, 1, 1), 1)$

#	x_1	x_2	x_3	x_4	$\kappa(\cdot)$
00	0	0	0	0	0
02	0	0	1	0	0
03	0	0	1	1	0
06	0	1	1	0	0
08	1	0	0	0	1
12	1	1	0	0	1
15	1	1	1	1	1

- How can we change prediction, given the **sample**?
- Move from row id 15 to one row taken out of $\{00, 02, 03, 06\}$

- Two CXps: $\{1, 2\}$ and $\{1, 4\}$

1 1 0 0
1 0 0 1

- Number of (W)CXps not larger than sample, and so polynomial on the input!

$\therefore$ sample-based explainability can use the algorithms developed for DTs!

Sample-based formal explanations

[MLM25]

- Set of CXps is linear on size of sample [MLM25]
- Set of CXps computed in polynomial time [MLM25]
- One AXp computed in polynomial time [CA23]

Sample-based formal explanations

[MLM25]

- Set of CXps is linear on size of sample [MLM25]
- Set of CXps computed in polynomial time [MLM25]
- One AXp computed in polynomial time [CA23]
- Computing one smallest AXp is NP-hard [RS23, MLM25]
  - MaxSAT encoding used for DTs can be used verbatim for smallest sample-based explanations

Sample-based formal explanations

[MLM25]

- Set of CXps is linear on size of sample [MLM25]
- Set of CXps computed in polynomial time [MLM25]
- One AXp computed in polynomial time [CA23]
- Computing one smallest AXp is NP-hard [RS23, MLM25]
  - MaxSAT encoding used for DTs can be used verbatim for smallest sample-based explanations
- Enumeration of AXps is in quasi-polynomial time; or just use a SAT oracle

Sample-based formal explanations

[MLM25]

- Set of CXps is linear on size of sample [MLM25]
- Set of CXps computed in polynomial time [MLM25]
- One AXp computed in polynomial time [CA23]
- Computing one smallest AXp is NP-hard [RS23, MLM25]
  - MaxSAT encoding used for DTs can be used verbatim for **smallest** sample-based explanations
- Enumeration of AXps is in quasi-polynomial time; or just use a SAT oracle
- And formal explanations are rigorous!
  - **Obs:** Anchors can produce **incorrect** explanations (talk on Tue...)

Outline

What are Shapley values?

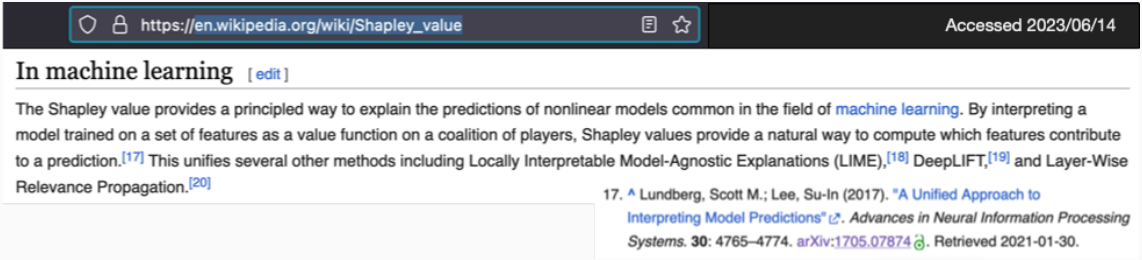
- First proposed in game theory in the early 50s by L. S. Shapley [Sha53]
  - Measures the contribution of each player to a cooperative game (v, N)

What are Shapley values?

- First proposed in game theory in the early 50s by L. S. Shapley [Sha53]
  - Measures the contribution of each player to a cooperative game (v, N)
- Application in XAI since the 2000s [LC01, SK10, SK14, DSZ16, LL17, ABBM21, VLSS21, VLSS22, ABBM23]
  - Popularized by SHAP [LL17]
  - Used for feature attribution, i.e. relative feature importance

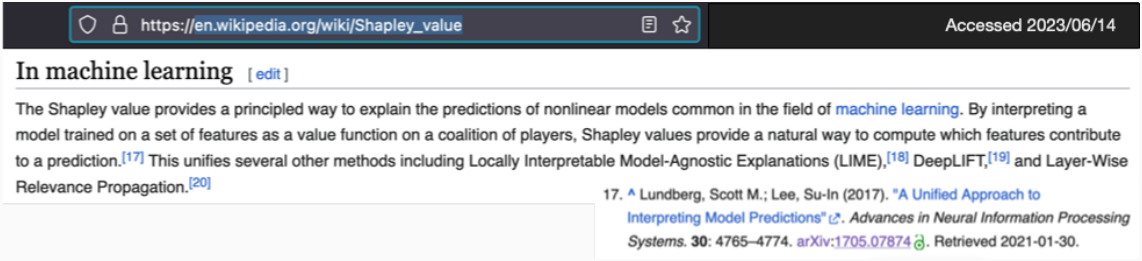
What are Shapley values?

- First proposed in game theory in the early 50s by L. S. Shapley [Sha53]
  - Measures the contribution of each player to a cooperative game (v, N)
- Application in XAI since the 2000s [LC01, SK10, SK14, DSZ16, LL17, ABBM21, VLSS21, VLSS22, ABBM23]
  - Popularized by SHAP [LL17]
  - Used for feature attribution, i.e. relative feature importance
- Shapley values are becoming ubiquitous in XAI...



What are Shapley values?

- First proposed in game theory in the early 50s by L. S. Shapley [Sha53]
  - Measures the contribution of each player to a cooperative game (v, N)
- Application in XAI since the 2000s [LC01, SK10, SK14, DSZ16, LL17, ABBM21, VLSS21, VLSS22, ABBM23]
  - Popularized by SHAP [LL17]
  - Used for feature attribution, i.e. relative feature importance
- Shapley values are becoming ubiquitous in XAI...



- Q: Do Shapley values for XAI really provide a rigorous measure of feature importance?

How are Shapley values used in explainability?

- Instance: $(\mathbf{v}, c)$

How are Shapley values used in explainability?

- Instance: $(\mathbf{v}, c)$
- $\Upsilon\colon 2^{\mathcal{F}} \to 2^{\mathbb{F}}$ defined by, [ABBM21, ABBM23]

$$\Upsilon(\mathcal{S}) = \{\mathbf{x} \in \mathbb{F} \mid \wedge_{i \in \mathcal{S}} x_i = v_i\}$$

$\Upsilon(\mathcal{S})$ gives points in feature space having the features in $\mathcal{S}$ fixed to their values in $\mathbf{v}$

How are Shapley values used in explainability?

- Instance: $(\mathbf{v}, c)$
- $\Upsilon\colon 2^{\mathcal{F}} \to 2^{\mathbb{F}}$ defined by, [ABBM21, ABBM23]

$$\Upsilon(\mathcal{S}) = \{\mathbf{x} \in \mathbb{F} \mid \wedge_{i \in \mathcal{S}}\, x_i = v_i\}$$

  $\Upsilon(\mathcal{S})$ gives points in feature space having the features in $\mathcal{S}$ fixed to their values in $\mathbf{v}$

- $\phi\colon 2^{\mathcal{F}} \to \mathbb{R}$ defined by,

$$\phi(\mathcal{S}) = 1/2^{|\mathcal{F} \setminus \mathcal{S}|} \sum_{\mathbf{x} \in \Upsilon(\mathcal{S})} \kappa(\mathbf{x}) \; = v_e(\mathcal{S})$$

  $\phi(\mathcal{S})$ represents the expected value of the classifier on the points given by $\Upsilon(\mathcal{S})$

How are Shapley values used in explainability?

- Instance: $(\mathbf{v}, c)$
- $\Upsilon\colon 2^{\mathcal{F}} \to 2^{\mathbb{F}}$ defined by, [ABBM21, ABBM23]

$$\Upsilon(\mathcal{S}) = \{\mathbf{x} \in \mathbb{F} \mid \wedge_{i \in \mathcal{S}} x_i = v_i\}$$

  $\Upsilon(\mathcal{S})$ gives points in feature space having the features in $\mathcal{S}$ fixed to their values in $\mathbf{v}$

- $\phi\colon 2^{\mathcal{F}} \to \mathbb{R}$ defined by,

$$\phi(\mathcal{S}) = 1/2^{|\mathcal{F}\setminus\mathcal{S}|} \sum_{\mathbf{x} \in \Upsilon(\mathcal{S})} \kappa(\mathbf{x}) \ = v_e(\mathcal{S})$$

  $\phi(\mathcal{S})$ represents the expected value of the classifier on the points given by $\Upsilon(\mathcal{S})$

- Sc: $\mathcal{F} \to \mathbb{R}$ defined by,

$$\mathrm{Sc}(i) = \sum_{\mathcal{S} \subseteq (\mathcal{F}\setminus\{i\})} |\mathcal{S}|!(|\mathcal{F}| - |\mathcal{S}| - 1)!/|\mathcal{F}|! \times (\phi(\mathcal{S} \cup \{i\}) - \phi(\mathcal{S}))$$

  For all subsets of features, excluding i, compute the expected value of the classifier, with and without i fixed, weighted by $\frac{1}{n}\binom{n}{|\mathcal{S}|}^{-1}$

  - **Obs:** Uniform distribution assumed; it suffices for our purposes

How are Shapley values used in explainability?

Marginal contribution (in SHAP lingo)!

- Instance: $(\mathbf{v}, c)$
- Υ: $2^{\mathcal{F}} \to 2^{\mathbb{F}}$ defined by, [ABBM21, ABBM23]

$$\Upsilon(\mathcal{S}) = \{\mathbf{x} \in \mathbb{F} \mid \wedge_{i \in \mathcal{S}} x_i = v_i\}$$

$\Upsilon(\mathcal{S})$ gives points in feature space having the features in $\mathcal{S}$ fixed to their values in $\mathbf{v}$

- ϕ: $2^{\mathcal{F}} \to \mathbb{R}$ defined by,

$$\phi(\mathcal{S}) = 1/2^{|\mathcal{F} \setminus \mathcal{S}|} \sum_{\mathbf{x} \in \Upsilon(\mathcal{S})} \kappa(\mathbf{x}) \ = v_e(\mathcal{S})$$

$\phi(\mathcal{S})$ represents the expected value of the classifier on the points given by $\Upsilon(\mathcal{S})$

- Sc: $\mathcal{F} \to \mathbb{R}$ defined by,

$$\mathrm{Sc}(i) = \sum_{\mathcal{S} \subseteq (\mathcal{F} \setminus \{i\})} |\mathcal{S}|!(|\mathcal{F}| - |\mathcal{S}| - 1)!/|\mathcal{F}|! \times (\phi(\mathcal{S} \cup \{i\}) - \phi(\mathcal{S}))$$

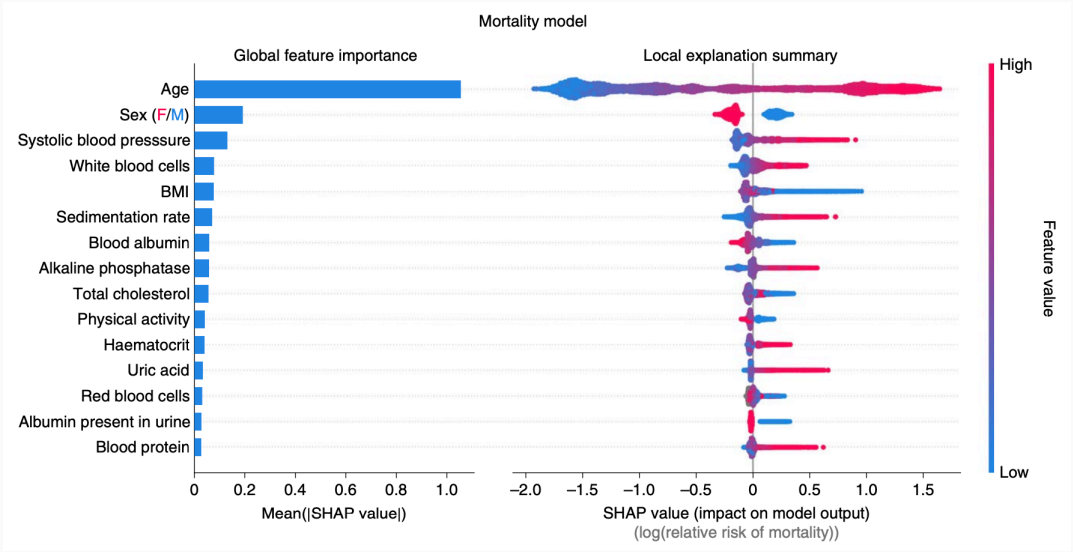
For all subsets of features, excluding i, compute the expected value of the classifier, with and without i fixed, weighted by $\frac{1}{n}\binom{n}{|\mathcal{S}|}^{-1}$

- **Obs:** Uniform distribution assumed; it suffices for our purposes

How are Shapley values computed in practice?

- Exact evaluation is computationally (very) hard [VLSS21, ABBM21, VLSS22, ABBM23, HM24]
- SHAP proposes a sample-based approach; with no guarantees of rigor [LL17]
  - Recent experiments revealed little to no correlation between Shapley values and SHAP's results [HM23b]
- Polynomial-time algorithm for deterministic decomposable boolean circuits [ABBM21]
- Polynomial-time algorithm for boolean functions represented with a truth-table [HM23b]

How are Shapley values computed in practice? example SHAP output...



What do Shapley values tell in terms of feature importance?

- [SK10] reads:
  "*According to the 2nd axiom, if two features values have an identical influence on the prediction they are assigned contributions of equal size. The 3rd axiom says that if a* ***feature has no influence*** *on the prediction* ***it is assigned a contribution of 0****.*"
  (Obs: the axioms refer to the axiomatic characterization of Shapley values.)

What do Shapley values tell in terms of feature importance?

- [SK10] reads:
  "*According to the 2nd axiom, if two features values have an identical influence on the prediction they are assigned contributions of equal size. The 3rd axiom says that if a* ***feature has no influence*** *on the prediction* ***it is assigned a contribution of 0****.*"
  (Obs: the axioms refer to the axiomatic characterization of Shapley values.)

- [SK10] also reads:
  "*When viewed together, these properties ensure that* ***any effect the features might have on the classifiers output will be reflected in the generated contributions****, which effectively deals with the issues of previous general explanation methods.*"

What do Shapley values tell in terms of feature importance?

- [SK10] reads:
  "*According to the 2nd axiom, if two features values have an identical influence on the prediction they are assigned contributions of equal size. The 3rd axiom says that if a* ***feature has no influence*** *on the prediction* ***it is assigned a contribution of 0****.*"
  (Obs: the axioms refer to the axiomatic characterization of Shapley values.)

- [SK10] also reads:
  "*When viewed together, these properties ensure that* ***any effect the features might have on the classifiers output will be reflected in the generated contributions****, which effectively deals with the issues of previous general explanation methods.*"

- **Obs:** Shapley values are defined axiomatically, i.e. **no** immediate relationship with AXp's/CXp's or with feature (ir)relevancy

What do Shapley values tell in terms of feature importance?

- [SK10] reads:
  *"According to the 2nd axiom, if two features values have an identical influence on the prediction they are assigned contributions of equal size. The 3rd axiom says that if a **feature has no influence** on the prediction **it is assigned a contribution of 0**."*
  (Obs: the axioms refer to the axiomatic characterization of Shapley values.)

- [SK10] also reads:
  *"When viewed together, these properties ensure that **any effect the features might have on the classifiers output will be reflected in the generated contributions**, which effectively deals with the issues of previous general explanation methods."*

- **Obs:** Shapley values are defined axiomatically, i.e. **no** immediate relationship with AXp's/CXp's or with feature (ir)relevancy
  - **Qs**: can we have **irrelevant** features with a non-zero Shapley value, and/or **relevant** features with a Shapley of zero?
    - Recall: **relevant** features occur in **some** AXp/CXp; **irrelevant** features do **not** occur in **any** AXp/CXp

Shapley values vs. feature (ir)relevancy – identified issues

[HM23b, HM23c, HM23d, MH23, HM24, MH24]

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:

Shapley values vs. feature (ir)relevancy – identified issues

[HM23b, HM23c, HM23d, MH23, HM24, MH24]

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:
  - Issue I1 occurs if,

$$\text{Irrelevant}(i) \land (\text{Sv}(i) \neq 0)$$

Shapley values vs. feature (ir)relevancy – identified issues

[HM23b, HM23c, HM23d, MH23, HM24, MH24]

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:
  - Issue I1 occurs if,

$$\text{Irrelevant}(i) \land (\text{Sv}(i) \neq 0)$$

  - Issue I2 occurs if,

$$\text{Irrelevant}(i_1) \land \text{Relevant}(i_2) \land (|\text{Sv}(i_1)| > |\text{Sv}(i_2)|)$$

Shapley values vs. feature (ir)relevancy – identified issues

[HM23b, HM23c, HM23d, MH23, HM24, MH24]

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:
  - Issue I1 occurs if,
    $$\text{Irrelevant}(i) \land (\text{Sv}(i) \neq 0)$$
  - Issue I2 occurs if,
    $$\text{Irrelevant}(i_1) \land \text{Relevant}(i_2) \land (|\text{Sv}(i_1)| > |\text{Sv}(i_2)|)$$
  - Issue I3 occurs if,
    $$\text{Relevant}(i) \land (\text{Sv}(i) = 0)$$

Shapley values vs. feature (ir)relevancy – identified issues [HM23b, HM23c, HM23d, MH23, HM24, MH24]

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:
  - Issue I1 occurs if,
    $$\text{Irrelevant}(i) \land (\text{Sv}(i) \neq 0)$$
  - Issue I2 occurs if,
    $$\text{Irrelevant}(i_1) \land \text{Relevant}(i_2) \land (|\text{Sv}(i_1)| > |\text{Sv}(i_2)|)$$
  - Issue I3 occurs if,
    $$\text{Relevant}(i) \land (\text{Sv}(i) = 0)$$
  - Issue I4 occurs if,
    $$[\text{Irrelevant}(i_1) \land (\text{Sv}(i_1) \neq 0)] \land [\text{Relevant}(i_2) \land (\text{Sv}(i_2) = 0)]$$

Shapley values vs. feature (ir)relevancy – identified issues [HM23b, HM23c, HM23d, MH23, HM24, MH24]

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:
  - Issue I1 occurs if,
    $$\text{Irrelevant}(i) \land (\text{Sv}(i) \neq 0)$$
  - Issue I2 occurs if,
    $$\text{Irrelevant}(i_1) \land \text{Relevant}(i_2) \land (|\text{Sv}(i_1)| > |\text{Sv}(i_2)|)$$
  - Issue I3 occurs if,
    $$\text{Relevant}(i) \land (\text{Sv}(i) = 0)$$
  - Issue I4 occurs if,
    $$[\text{Irrelevant}(i_1) \land (\text{Sv}(i_1) \neq 0)] \land [\text{Relevant}(i_2) \land (\text{Sv}(i_2) = 0)]$$
  - Issue I5 occurs if,
    $$[\text{Irrelevant}(i) \land \forall_{1 \leq j \leq m, j \neq i} (|\text{Sv}(j)| < |\text{Sv}(i)|)]$$

Shapley values vs. feature (ir)relevancy – identified issues [HM23b, HM23c, HM23d, MH23, HM24, MH24]

- Boolean classifier, instance $(\mathbf{v}, c)$, and some $i, i_1, i_2 \in \mathcal{F}$:
  - Issue I1 occurs if,
    $$\text{Irrelevant}(i) \land (\text{Sv}(i) \neq 0)$$
  - Issue I2 occurs if,
    $$\text{Irrelevant}(i_1) \land \text{Relevant}(i_2) \land (|\text{Sv}(i_1)| > |\text{Sv}(i_2)|)$$
  - Issue I3 occurs if,
    $$\text{Relevant}(i) \land (\text{Sv}(i) = 0)$$
  - Issue I4 occurs if,
    $$[\text{Irrelevant}(i_1) \land (\text{Sv}(i_1) \neq 0)] \land [\text{Relevant}(i_2) \land (\text{Sv}(i_2) = 0)]$$
  - Issue I5 occurs if,
    $$[\text{Irrelevant}(i) \land \forall_{1 \leq j \leq m, j \neq i} (|\text{Sv}(j)| < |\text{Sv}(i)|)]$$

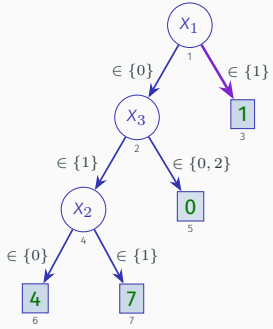
Any of these issues is a cause of (**serious**) concern per se**!**

Some stats – all boolean functions with 4 variables

[HM23b, HM23c, HM23d, MH23, HM24, MH24]

Issue-related metric	Value	Recap issue
# of functions	65536	
# number of instances	1048576	
# of I1 issues	781696	
# of functions with I1 issues	65320	
% I1 issues / function	99.67	[Irrelevant(i) $\wedge$ (Sv(i) $\neq 0$)]
# of I2 issues	105184	
# of functions with I2 issues	40448	
% I2 issues / function	61.72	[Irrelevant(i_1) $\wedge$ Relevant(i_2) $\wedge$ ($\lvert$Sv(i_1)$\rvert$ > $\lvert$Sv(i_2)$\rvert$)]
# of I3 issues	43008	
# of functions with I3 issues	7800	
% I3 issues / function	11.90	[Relevant(i) $\wedge$ (Sv(i) $= 0$)]
# of I4 issues	5728	
# of functions with I4 issues	2592	
% I4 issues / function	3.96	[Irrelevant(i_1) $\wedge$ (Sv(i_1) $\neq 0$)] $\wedge$ [Relevant(i_2) $\wedge$ (Sv(i_2) $= 0$)]
# of I5 issues	1664	
# of functions with I5 issues	1248	
% I5 issues / function	1.90	[Irrelevant(i) $\wedge \forall_{1 \leq j \leq m, j \neq i}$ ($\lvert$Sv(j)$\rvert$ < $\lvert$Sv(i)$\rvert$)]

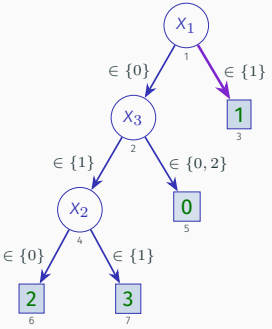
Previous results do matter! Let's go non-boolean...



DT1

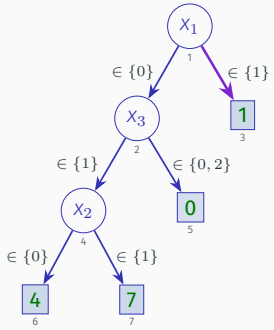
row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations



DT2

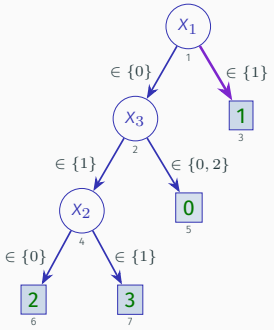
Instance $((1, 1, 2), 1)$ – which feature matters the most for prediction 1?



DT1

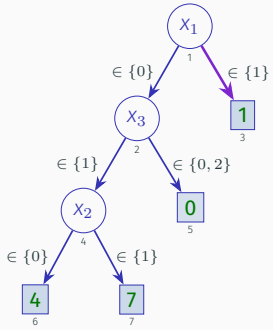
row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations



DT2

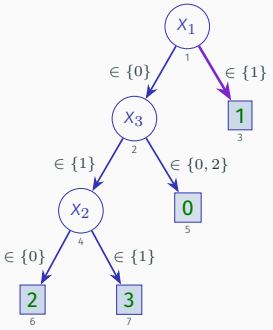
Computing XPs – make sense...



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

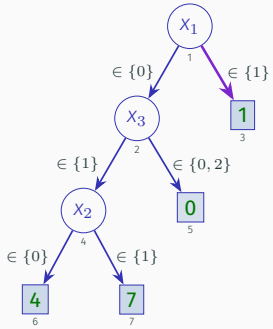
Tabular representations



DT2

XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

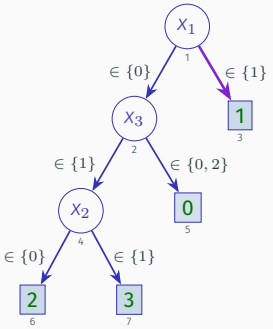
Computing XPs, AEs – also make sense...



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations

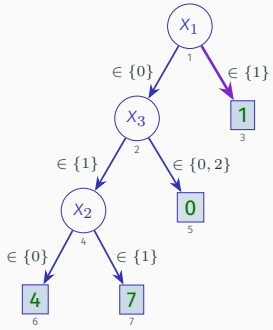


DT2

XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

Adversarial examples	
DT	minimal AEs
DT1	$\{1\}$
DT2	$\{1\}$

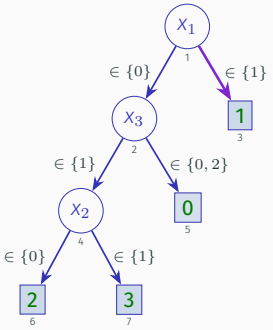
Computing XPs, AEs & Svs



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations



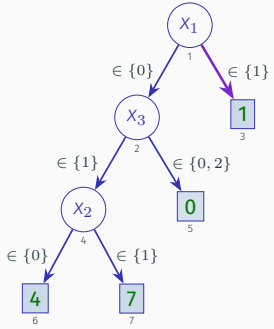
DT2

XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

Adversarial examples	
DT	minimal AEs
DT1	$\{1\}$
DT2	$\{1\}$

Shapley values			
DT	Sc(1)	Sc(2)	Sc(3)
DT1	0.000	0.083	-0.500
DT2	0.278	0.028	-0.222

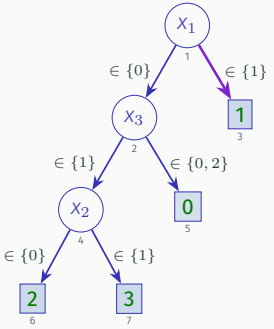
Computing XPs, AEs & Svs – what!?!?



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations



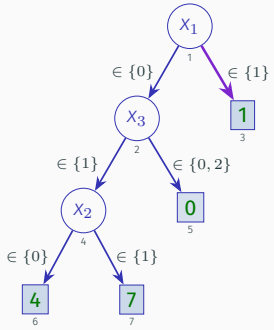
DT2

XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

Adversarial examples	
DT	minimal AEs
DT1	$\{1\}$
DT2	$\{1\}$

Shapley values in theory				
DT	Sc(1)	Sc(2)	Sc(3)	
DT1	0.000	0.083	-0.500	!!!
DT2	0.278	0.028	-0.222	!!

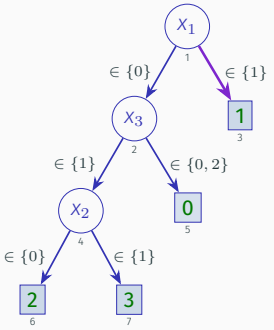
Computing XPs, AEs & Svs – what!?!?



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations



DT2

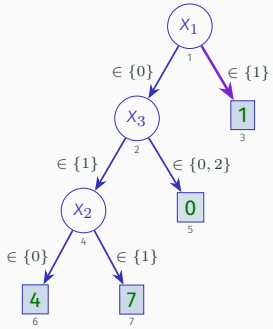
XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

Adversarial examples	
DT	minimal AEs
DT1	$\{1\}$
DT2	$\{1\}$

Shapley values in theory				
DT	Sc(1)	Sc(2)	Sc(3)	
DT1	0.000	0.083	-0.500	!!!
DT2	0.278	0.028	-0.222	!!

Computing XPs, AEs & Svs – what!?!?

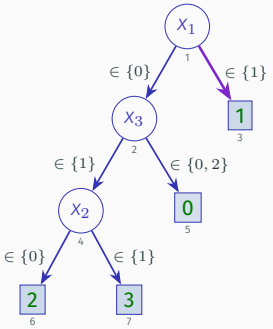
∴ Shapley values can mislead human decision-makers !



DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations



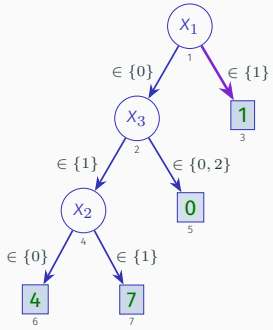
DT2

XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

Adversarial examples	
DT	minimal AEs
DT1	$\{1\}$
DT2	$\{1\}$

Shapley values in theory				
DT	Sc(1)	Sc(2)	Sc(3)	
DT1	0.000	0.083	-0.500	!!!
DT2	0.278	0.028	-0.222	!!

Computing XPs, AEs & Svs – what!?!?

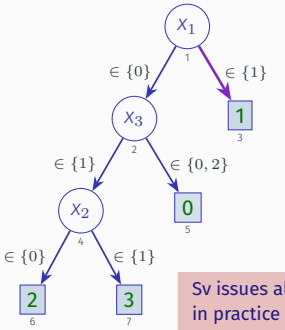


DT1

row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	0	0
2	0	0	1	4	2
3	0	0	2	0	0
4	0	1	0	0	0
5	0	1	1	7	3
6	0	1	2	0	0
7	1	0	0	1	1
8	1	0	1	1	1
9	1	0	2	1	1
10	1	1	0	1	1
11	1	1	1	1	1
12	1	1	2	1	1

Tabular representations

$\therefore$ Shapley values can mislead human decision-makers !



Sv issues also occur in practice [HM23d]

DT2

XPs: AXps/CXps

DT	AXps	CXps
DT1	$\{1\}$	$\{1\}$
DT2	$\{1\}$	$\{1\}$

Adversarial examples

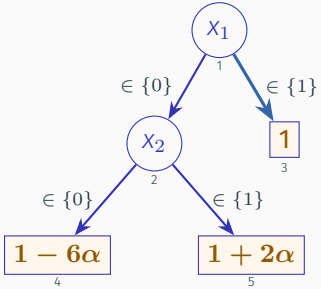
DT	minimal AEs
DT1	$\{1\}$
DT2	$\{1\}$

Shapley values in theory

DT	Sc(1)	Sc(2)	Sc(3)	
DT1	0.000	0.083	-0.500	!!!
DT2	0.278	0.028	-0.222	!!

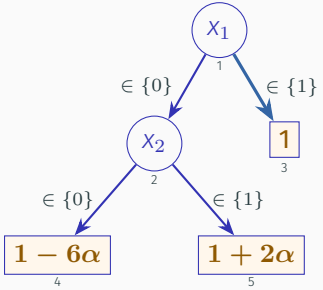
Another example – arbitrary mistakes!

[LHAM24]



Another example – arbitrary mistakes!

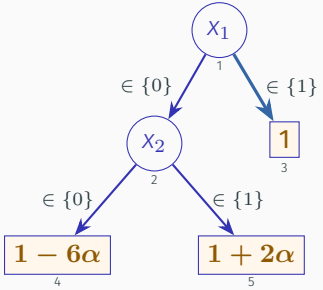
[LHAM24]



- Instance: $((1, 1), 1)$
- Obs: $\alpha \neq 0$

Another example – arbitrary mistakes!

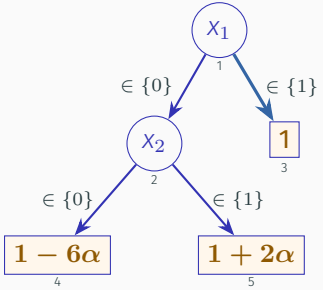
[LHAM24]



- Instance: $((1, 1), 1)$
- Obs: $\alpha \neq 0$
- $\mathrm{Sc}(1) = 0$
- $\mathrm{Sc}(2) = \alpha$

Another example – arbitrary mistakes!

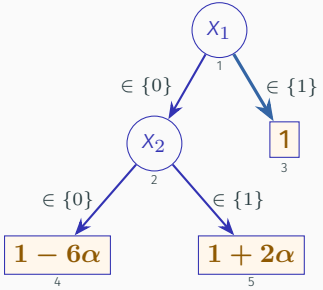
[LHAM24]



- Instance: $((1,1),1)$
- Obs: $\alpha \neq 0$
- $\mathrm{Sc}(1) = 0$
- $\mathrm{Sc}(2) = \alpha$ (you can pick the α...)

Another example – arbitrary mistakes!

[LHAM24]

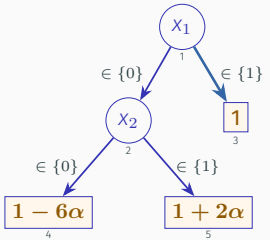


- Instance: $((1,1),1)$
- Obs: $\alpha \neq 0$
- $\mathrm{Sc}(1) = 0$
- $\mathrm{Sc}(2) = \alpha$ (you can pick the α...)

Note: example devised in 2024 by O. Letoffe, PhD student at IRIT

More detail

row	x_1	x_2	$\rho(\mathbf{x})$	$\rho_a(\mathbf{x})$ $\alpha = 1/2$	$\rho_b(\mathbf{x})$ $\alpha = 1/4$
1	0	0	$1-6\alpha$	-2	$-1/2$
2	0	1	$1+2\alpha$	2	$3/2$
3	1	0	1	1	1
4	1	1	1	1	1



$\mathcal{S}$	rows($\mathcal{S}$)	$v_e(\mathcal{S})$
$\emptyset$	$1,2,3,4$	$1-\alpha$
$\{x_1\}$	$3,4$	1
$\{x_2\}$	$2,4$	$1+\alpha$
$\{x_1,x_2\}$	4	1

$i = 1$

$\mathcal{S}$	$v_e(\mathcal{S})$	$v_e(\mathcal{S}\cup\{1\})$	$\Delta_1(\mathcal{S})$	$\varsigma(\mathcal{S})$	$\varsigma(\mathcal{S})\times\Delta_1(\mathcal{S})$
$\emptyset$	$1-\alpha$	1	α	$1/2$	$\alpha/2$
$\{2\}$	$1+\alpha$	1	$-\alpha$	$1/2$	$-\alpha/2$
			$\mathsf{Sc}_E(1)$	$=$	0

$i = 2$

$\mathcal{S}$	$v_e(\mathcal{S})$	$v_e(\mathcal{S}\cup\{2\})$	$\Delta_2(\mathcal{S})$	$\varsigma(\mathcal{S})$	$\varsigma(\mathcal{S})\times\Delta_2(\mathcal{S})$
$\emptyset$	$1-\alpha$	$1+\alpha$	2α	$1/2$	α
$\{1\}$	1	1	0	$1/2$	0
			$\mathsf{Sc}_E(2)$	$=$	α

SHAP scores also fail with regression models

[LHM24b]

- Let $\mathcal{F} = \{1, 2\}$, $\mathbb{D}_1 = \mathbb{D}_2 = \mathbb{D} = [-1/2, 3/2]$, $\mathbb{F} = \mathbb{D} \times \mathbb{D}$
  - Also, let $\mathbb{D}^+ = [1/2, 3/2]$ and $\mathbb{D}^- = \mathbb{D} \setminus \mathbb{D}^+$

- Regression model maps to real values, i.e. $\mathbb{K} = \mathbb{R}$:

$$\rho_2(x_1, x_2) = \begin{cases} x_1 & \text{if } x_1 \in \mathbb{D}^+ \\ x_2 - 2 & \text{if } x_1 \notin \mathbb{D}^+ \wedge x_2 \notin \mathbb{D}^+ \\ x_2 + 1 & \text{if } x_1 \notin \mathbb{D}^+ \wedge x_2 \in \mathbb{D}^+ \end{cases}$$

SHAP scores also fail with regression models

[LHM24b]

- Let $\mathcal{F} = \{1, 2\}$, $\mathbb{D}_1 = \mathbb{D}_2 = \mathbb{D} = [-1/2, 3/2]$, $\mathbb{F} = \mathbb{D} \times \mathbb{D}$
  - Also, let $\mathbb{D}^+ = [1/2, 3/2]$ and $\mathbb{D}^- = \mathbb{D} \setminus \mathbb{D}^+$

- Regression model maps to real values, i.e. $\mathbb{K} = \mathbb{R}$:

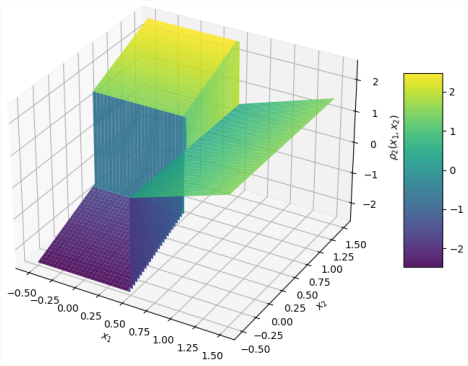
$$\rho_2(x_1, x_2) = \begin{cases} x_1 & \text{if } x_1 \in \mathbb{D}^+ \\ x_2 - 2 & \text{if } x_1 \notin \mathbb{D}^+ \wedge x_2 \notin \mathbb{D}^+ \\ x_2 + 1 & \text{if } x_1 \notin \mathbb{D}^+ \wedge x_2 \in \mathbb{D}^+ \end{cases}$$

- Average values unchanged (wrt previous example), and so $\mathrm{Sc}(1) = 0$ and $\mathrm{Sc}(2) = \alpha$

SHAP scores also fail with regression models

[LHM24b]

- Let $\mathcal{F} = \{1, 2\}$, $\mathbb{D}_1 = \mathbb{D}_2 = \mathbb{D} = [-1/2, 3/2]$, $\mathbb{F} = \mathbb{D} \times \mathbb{D}$
  - Also, let $\mathbb{D}^+ = [1/2, 3/2]$ and $\mathbb{D}^- = \mathbb{D} \setminus \mathbb{D}^+$



- Regression model maps to real values, i.e. $\mathbb{K} = \mathbb{R}$:

$$\rho_2(x_1, x_2) = \begin{cases} x_1 & \text{if } x_1 \in \mathbb{D}^+ \\ x_2 - 2 & \text{if } x_1 \notin \mathbb{D}^+ \wedge x_2 \notin \mathbb{D}^+ \\ x_2 + 1 & \text{if } x_1 \notin \mathbb{D}^+ \wedge x_2 \in \mathbb{D}^+ \end{cases}$$

- Average values unchanged (wrt previous example), and so $\mathrm{Sc}(1) = 0$ and $\mathrm{Sc}(2) = \alpha$

SHAP scores fail even when Lipschitz continuity holds!

[LHM24b]

$$\rho_3(x_1, x_2) = \begin{cases} x_1 & \text{if } x_2 \leq 1 \wedge \alpha x_1 \leq \alpha \\ (1 + 4|\alpha|)x_1 - 4|\alpha| & \text{if } x_2 \leq 1 \wedge \alpha x_1 \geq \alpha \\ 28|\alpha|x_1x_2 + (1 - 28|\alpha|)x_1 - 28|\alpha|x_2 + 28|\alpha| & \text{if } x_2 \geq 1 \wedge \alpha x_1 \leq \alpha \\ -4|\alpha|x_1x_2 + (1 + 8|\alpha|)x_1 + 4|\alpha|x_2 - 8|\alpha| & \text{if } x_2 \geq 1 \wedge \alpha x_1 \geq \alpha \end{cases}$$

- As before, average values unchanged, and so $\mathrm{Sc}(1) = 0$ and $\mathrm{Sc}(2) = \alpha$

SHAP scores fail even when Lipschitz continuity holds!

[LHM24b]

$$\rho_3(x_1, x_2) = \begin{cases} x_1 & \text{if } x_2 \leq 1 \wedge \alpha x_1 \leq \alpha \\ (1 + 4|\alpha|)x_1 - 4|\alpha| & \text{if } x_2 \leq 1 \wedge \alpha x_1 \geq \alpha \\ 28|\alpha|x_1x_2 + (1 - 28|\alpha|)x_1 - 28|\alpha|x_2 + 28|\alpha| & \text{if } x_2 \geq 1 \wedge \alpha x_1 \leq \alpha \\ -4|\alpha|x_1x_2 + (1 + 8|\alpha|)x_1 + 4|\alpha|x_2 - 8|\alpha| & \text{if } x_2 \geq 1 \wedge \alpha x_1 \geq \alpha \end{cases}$$

- As before, average values unchanged, and so $\mathrm{Sc}(1) = 0$ and $\mathrm{Sc}(2) = \alpha$

NOTE: **SHAP scores also fail with arbitrary differentiability!**

Outline

Corrected SHAP scores & feature importance scores

[LHM24a, LHAM24]

- Is the theory of Shapley values incorrect?

Corrected SHAP scores & feature importance scores

[LHM24a, LHAM24]

- Is the theory of Shapley values incorrect? No!

Corrected SHAP scores & feature importance scores

[LHM24a, LHAM24]

- Is the theory of Shapley values incorrect? No!
- What is flawed is the characteristic function used in XAI [SK10, SK14, LL17]
  - In XAI: characteristic function uses the expected value
  - This defines the *marginal contribution* in SHAP lingo...

Corrected SHAP scores & feature importance scores

[LHM24a, LHAM24]

- Is the theory of Shapley values incorrect? No!
- What is flawed is the characteristic function used in XAI [SK10, SK14, LL17]
  - In XAI: characteristic function uses the expected value
  - This defines the *marginal contribution* in SHAP lingo...
- Replace characteristic function based on expected values by new characteristic function based on AXps/WAXps [LHM24a]
  - Resulting scores are (still) Shapley values & identified issues no longer observed

Corrected SHAP scores & feature importance scores

[LHM24a, LHAM24]

- Is the theory of Shapley values incorrect? No!
- What is flawed is the characteristic function used in XAI [SK10, SK14, LL17]
  - In XAI: characteristic function uses the expected value
  - This defines the *marginal contribution* in SHAP lingo...
- Replace characteristic function based on expected values by new characteristic function based on AXps/WAXps [LHM24a]
  - Resulting scores are (still) Shapley values & identified issues no longer observed
- Observed tight connection between feature attribution and power indices from a priori voting power

An initial compromise

[LHAM24]

- Replace the characteristic function used for SHAP scores:

$$v_e(\mathcal{S}) \quad := \quad \mathbf{E}[\tau(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

An initial compromise

[LHAM24]

- Replace the characteristic function used for SHAP scores:

$$v_e(\mathcal{S}) \quad := \quad \mathbf{E}[\tau(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

- Recall the similarity predicate:

$$\sigma(\mathbf{x}) = \begin{cases} 1 & \text{if } (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \\ 0 & \text{otherwise} \end{cases}$$

An initial compromise

[LHAM24]

- Replace the characteristic function used for SHAP scores:

$$\upsilon_e(\mathcal{S}) \quad := \quad \mathbf{E}[\tau(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

- Recall the similarity predicate:

$$\sigma(\mathbf{x}) = \begin{cases} 1 & \text{if } (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \\ 0 & \text{otherwise} \end{cases}$$

- The new characteristic function becomes:

$$\upsilon_s(\mathcal{S}) \quad := \quad \mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

An initial compromise

[LHAM24]

- Replace the characteristic function used for SHAP scores:

$$\upsilon_e(\mathcal{S}) \quad := \quad \mathbf{E}[\tau(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

- Recall the similarity predicate:

$$\sigma(\mathbf{x}) = \begin{cases} 1 & \text{if } (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \\ 0 & \text{otherwise} \end{cases}$$

- The new characteristic function becomes:

$$\upsilon_s(\mathcal{S}) \quad := \quad \mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

- Issues with non-boolean classifiers disappear; issues with boolean classifiers remain

An initial compromise

[LHAM24]

- Replace the characteristic function used for SHAP scores:

$$\upsilon_e(\mathcal{S}) \quad := \quad \mathbf{E}[\tau(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

- Recall the similarity predicate:

$$\sigma(\mathbf{x}) = \begin{cases} 1 & \text{if } (\kappa(\mathbf{x}) = \kappa(\mathbf{v})) \\ 0 & \text{otherwise} \end{cases}$$

- The new characteristic function becomes:

$$\upsilon_s(\mathcal{S}) \quad := \quad \mathbf{E}[\sigma(\mathbf{x}) \,|\, \mathbf{x}_{\mathcal{S}} = \mathbf{v}_{\mathcal{S}}]$$

- Issues with non-boolean classifiers disappear; issues with boolean classifiers remain
- Developed SSHAP prototype using SHAP's code base [LHM24a]

Fixing the known issues of SHAP scores

Fixing the known issues of SHAP scores

- New characteristic function (based on WAXps):

$$v_a(\mathcal{S}) := \begin{cases} 1 & \text{if WAXp}(\mathcal{S}) \\ 0 & \text{otherwise} \end{cases}$$

Fixing the known issues of SHAP scores

- New characteristic function (based on WAXps):

$$v_a(\mathcal{S}) := \begin{cases} 1 & \text{if WAXp}(\mathcal{S}) \\ 0 & \text{otherwise} \end{cases}$$

- Known issues of SHAP scores guaranteed **not** to occur (more later)

Fixing the known issues of SHAP scores

- New characteristic function (based on WAXps):

$$\upsilon_a(\mathcal{S}) := \begin{cases} 1 & \text{if } \mathsf{WAXp}(\mathcal{S}) \\ 0 & \text{otherwise} \end{cases}$$

- Known issues of SHAP scores guaranteed **not** to occur (more later)

- **Corrected** SHAP scores reveal tight connection between XAI by feature selection (i.e. WAXps) and feature attribution

Outline

- Shapley values estimated with well-known algorithm – CGT [CGT09]
  - Strong theoretical guarantees on approximation of Shapley values
  - With enough sampling, given by ϵ and α:

$$\text{Prob}[|\hat{\text{Sc}}(i) - \text{Sc}(i)| \leq \epsilon] \geq 1 - \alpha$$

  - E.g. we used $\epsilon = 0.0015$ and $\alpha = 0.015$

- Shapley values estimated with well-known algorithm – CGT [CGT09]
  - Strong theoretical guarantees on approximation of Shapley values
  - With enough sampling, given by ϵ and α:

$$\text{Prob}[|\hat{\text{Sc}}(i) - \text{Sc}(i)| \leq \epsilon] \geq 1 - \alpha$$

  - E.g. we used $\epsilon = 0.0015$ and $\alpha = 0.015$

- Replace expected value with novel characteristic function: test for WAXp
  - Model-aware WAXp checking unrealistic for highly complex models
    - And CGT may require a very large number of WAXp checks

A glimpse of NuSHAP

[MHL25]

- Shapley values estimated with well-known algorithm – CGT [CGT09]
  - Strong theoretical guarantees on approximation of Shapley values
  - With enough sampling, given by ϵ and α:

$$\text{Prob}[|\hat{\text{Sc}}(i) - \text{Sc}(i)| \leq \epsilon] \geq 1 - \alpha$$

  - E.g. we used $\epsilon = 0.0015$ and $\alpha = 0.015$

- Replace expected value with novel characteristic function: test for WAXp
  - Model-aware WAXp checking unrealistic for highly complex models
    - And CGT may require a very large number of WAXp checks
  - NuSHAP uses model-agnostic (and sample-based) WAXp checking
    - Polynomial-time algorithm
    - Flexibility in the sample chosen, e.g. training data

NuSHAP vs. *SHAP

- Compute SHAP scores with NuSHAP and SHAP
- Compare features' rankings using Rank-Biased Overlap (RBO) [WMZ10]
  - Give weight to differences in first k elements of ranking
  - E.g. $\text{RBO}([0,1,2,3,4,5,6,7,8,9],[0,1,2,3,4,5,6,7,8,9],k=5)=1.0$
- Results for different datasets and ML models (using scikit, XGBoost):
  - LR: (adult, corral); DT: (iris, mux6); kNN: (connect_4, spambase, spectf); BT: (clean1, coil2000, dna); CNN: (MNIST)

NuSHAP vs. *SHAP

- Compute SHAP scores with NuSHAP and SHAP
- Compare features' rankings using Rank-Biased Overlap (RBO) [WMZ10]
  - Give weight to differences in first k elements of ranking
  - E.g. $\text{RBO}([0,1,2,3,4,5,6,7,8,9],[0,1,2,3,4,9,8,7,6,5],k=5)=1.0$
- Results for different datasets and ML models (using scikit, XGBoost):
  - LR: (adult, corral); DT: (iris, mux6); kNN: (connect_4, spambase, spectf); BT: (clean1, coil2000, dna); CNN: (MNIST)

NuSHAP vs. *SHAP

- Compute SHAP scores with NuSHAP and SHAP
- Compare features' rankings using Rank-Biased Overlap (RBO) [WMZ10]
  - Give weight to differences in first k elements of ranking
  - E.g. $\text{RBO}([0,1,2,3,4,5,6,7,8,9],[4,3,2,1,0,5,6,7,8,9],k=5)=0.42$
- Results for different datasets and ML models (using scikit, XGBoost):
  - LR: (adult, corral); DT: (iris, mux6); kNN: (connect_4, spambase, spectf); BT: (clean1, coil2000, dna); CNN: (MNIST)

NuSHAP vs. *SHAP

- Compute SHAP scores with NuSHAP and SHAP
- Compare features' rankings using Rank-Biased Overlap (RBO) [WMZ10]
  - Give weight to differences in first k elements of ranking
  - E.g. $\text{RBO}([0,1,2,3,4,5,6,7,8,9],[5,6,7,8,9,0,1,2,3,4],k=5)=0.0$
- Results for different datasets and ML models (using scikit, XGBoost):
  - LR: (adult, corral); DT: (iris, mux6); kNN: (connect_4, spambase, spectf); BT: (clean1, coil2000, dna); CNN: (MNIST)

NuSHAP vs. *SHAP

- Compute SHAP scores with NuSHAP and SHAP
- Compare features' rankings using Rank-Biased Overlap (RBO) [WMZ10]
  - Give weight to differences in first k elements of ranking
  - E.g. $\text{RBO}([0,1,2,3,4,5,6,7,8,9],[5,6,7,8,9,0,1,2,3,4],k=5)=0.0$
- Results for different datasets and ML models (using scikit, XGBoost):
  - LR: (adult, corral); DT: (iris, mux6); kNN: (connect_4, spambase, spectf); BT: (clean1, coil2000, dna); CNN: (MNIST)

		adult	corral	iris	mux6	connect_4	spambase	spectf	clean1	coil2000	dna	MNIST
Min	nuSHAP vs. SHAP	0.08	0.17	0.31	0.32	0.0	0.01	0.0	0.0	0.0	0.0	0.0
	nuSHAP vs. |SHAP|	0.05	0.12	0.27	0.32	0.0	0.05	0.0	0.0	0.0	0.03	0.0
Max	nuSHAP vs. SHAP	0.96	0.96	0.94	0.97	0.9	0.94	0.91	0.69	0.69	0.88	0.06
	nuSHAP vs. |SHAP|	0.88	0.97	0.94	0.95	0.77	0.94	0.91	0.88	0.69	0.88	0.06
Mean	nuSHAP vs. SHAP	0.37	0.53	0.84	0.7	0.21	0.41	0.2	0.12	0.05	0.17	0.0
	nuSHAP vs. |SHAP|	0.31	0.5	0.84	0.69	0.19	0.42	0.19	0.17	0.08	0.43	0.0

NuSHAP vs. *SHAP – run times

- Compute SHAP scores with NuSHAP and SHAP
- Compare features' rankings using Rank-Biased Overlap (RBO) [WMZ10]
  - Give weight to differences in first k elements of ranking
  - E.g. $\text{RBO}([0,1,2,3,4,5,6,7,8,9],[5,6,7,8,9,0,1,2,3,4],k=5)=0.0$
- Results for different datasets and ML models (using scikit, XGBoost):
  - LR: (adult, corral); DT: (iris, mux6); kNN: (connect_4, spambase, spectf); BT: (clean1, coil2000, dna); CNN: (MNIST)

		adult	corral	iris	mux6	connect_4	spambase	spectf	clean1	coil2000	dna	MNIST
Min	nuSHAP vs. SHAP	0.08	0.17	0.31	0.32	0.0	0.01	0.0	0.0	0.0	0.0	0.0
	nuSHAP vs. |SHAP|	0.05	0.12	0.27	0.32	0.0	0.05	0.0	0.0	0.0	0.03	0.0
Max	nuSHAP vs. SHAP	0.96	0.96	0.94	0.97	0.9	0.94	0.91	0.69	0.69	0.88	0.06
	nuSHAP vs. |SHAP|	0.88	0.97	0.94	0.95	0.77	0.94	0.91	0.88	0.69	0.88	0.06
Mean	nuSHAP vs. SHAP	0.37	0.53	0.84	0.7	0.21	0.41	0.2	0.12	0.05	0.17	0.0
	nuSHAP vs. |SHAP|	0.31	0.5	0.84	0.69	0.19	0.42	0.19	0.17	0.08	0.43	0.0

Tool SHAP produces results of very poor quality!

NuSHAP vs. *SHAP – run times

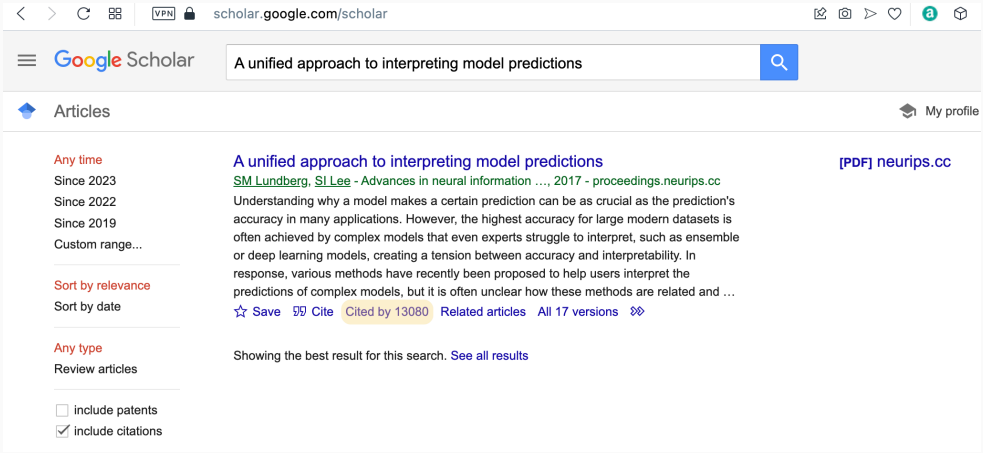
- Compute SHAP scores with NuSHAP and SHAP
- Compare features' rankings using Rank-Biased Overlap (RBO) [WMZ10]
  - Give weight to differences in first k elements of ranking
  - E.g. $\text{RBO}([0,1,2,3,4,5,6,7,8,9],[5,6,7,8,9,0,1,2,3,4],k=5)=0.0$
- Results for different datasets and ML models (using scikit, XGBoost):
  - LR: (adult, corral); DT: (iris, mux6); kNN: (connect_4, spambase, spectf); BT: (clean1, coil2000, dna); CNN: (MNIST)

	adult	corral	iris	mux6	connect_4	spambase	spectf	clean1	coil2000	dna	MNIST
SHAP	3.4	**0.1**	**0.0**	**0.0**	21.7	**0.5**	**0.7**	6.8	28.2	23.0	281.3
nuSHAP	**1.9**	1.5	1.5	1.5	**4.5**	2.7	2.9	**2.7**	**1.7**	**4.5**	**48.9**
#Samples	68045.5	63.7	955.4	63.9	9202.5	13654.1	36459.5	3929.6	2960.8	2756.8	2929.9

Outline

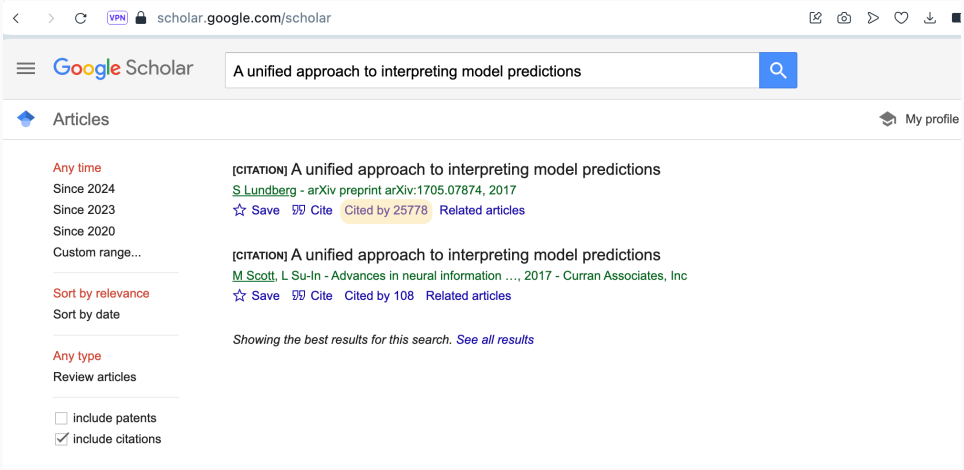
Can non-symbolic XAI's myths be stopped?

SHAP on 2023/05/31:



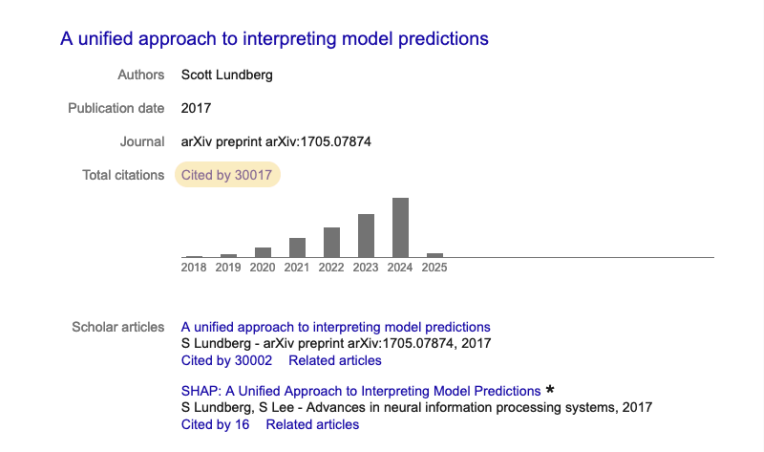
Can non-symbolic XAI's myths be stopped?

SHAP on 2024/09/15:



Can non-symbolic XAI's myths be stopped?

SHAP on 2025/01/12:

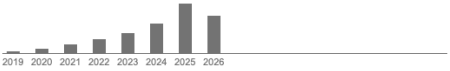


Can non-symbolic XAI's myths be stopped?

SHAP on 2026/07/16:

A unified approach to interpreting model predictions [PDF] from neurips.cc

Authors	Scott M Lundberg, Su-In Lee
Publication date	2017
Journal	Advances in neural information processing systems
Volume	30
Description	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.
Total citations	Cited by 64411



Can non-symbolic XAI's myths be stopped?

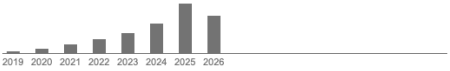
SHAP on 2026/07/16:

But: theoretical SHAP scores can mislead & results of tool SHAP of poor quality!

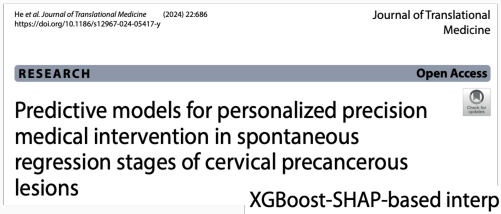
A unified approach to interpreting model predictions

[PDF] from neurips.cc

Authors	Scott M Lundberg, Su-In Lee
Publication date	2017
Journal	Advances in neural information processing systems
Volume	30
Description	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.
Total citations	Cited by 64411



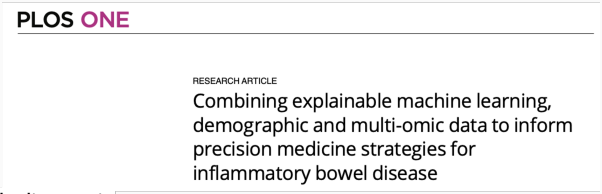
Many, many high-risk uses of SHAP & clones


He et al. Journal of Translational Medicine (2024) 22:686
https://doi.org/10.1186/s12967-024-05417-y

Journal of Translational Medicine

RESEARCH Open Access

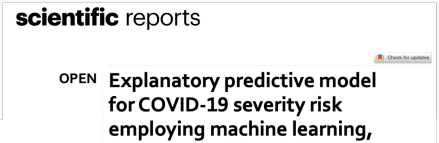
Predictive models for personalized precision medical intervention in spontaneous regression stages of cervical precancerous lesions


PLOS ONE

RESEARCH ARTICLE

Combining explainable machine learning, demographic and multi-omic data to inform precision medicine strategies for inflammatory bowel disease

XGBoost-SHAP-based interpretable diagnostic framework for alzheimer's disease

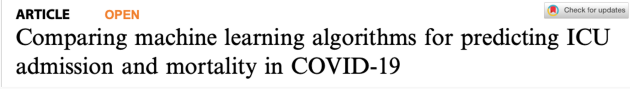

scientific reports

OPEN Explanatory predictive model for COVID-19 severity risk employing machine learning, shapley addition, and LIME

Hypothesis-free discovery of novel cancer predictors using machine learning

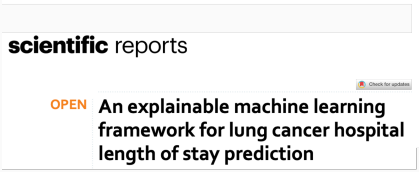
Explainable AI-based Deep-SHAP for mapping the multivariate relationships between regional neuroimaging biomarkers and cognition

npj | Digital Medicine

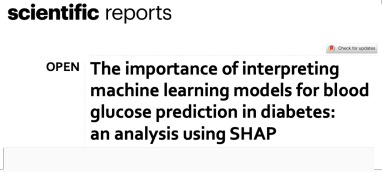

ARTICLE OPEN

Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19

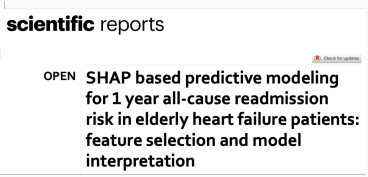
Interpretable prediction of 30-day mortality in patients with acute pancreatitis based on machine learning and SHAP


scientific reports

OPEN An explainable machine learning framework for lung cancer hospital length of stay prediction


scientific reports

OPEN The importance of interpreting machine learning models for blood glucose prediction in diabetes: an analysis using SHAP


scientific reports

OPEN SHAP based predictive modeling for 1 year all-cause readmission risk in elderly heart failure patients: feature selection and model interpretation

What next?

What next? Massive retraction (of ~~tens~~ hundreds of thousands) of papers??

thanks to ChatGPT...

What next? Massive retraction (of ~~tens~~ hundreds of thousands) of papers??

& beware of high-risk uses of SHAP! ☹

thanks to ChatGPT...

What’s the bottom line?

What's the bottom line?

- Use(r)s of non-symbolic XAI experience a persistent "*Don't Look Up*" moment...

What's the bottom line?

- Use(r)s of non-symbolic XAI experience a persistent "*Don't Look Up*" moment...

BTW, there are a multitude of proposed uses of LIME/SHAP in medicine... ⚠

Conclusions

- Brief contact with logic-based symbolic XAI
  - Rigorous definition of abductive/constrastive explanations
- Highlighted the flaws of Shapley values for XAI, i.e. the SHAP scores
- Detailed corrections to Shapley values for XAI
  - Right connection between logic-based XAI by feature selection and (corrected) XAI by feature attribution
- Overviewed nuSHAP
  - Still a prototype...
- Research directions?

Conclusions

- Brief contact with logic-based symbolic XAI
  - Rigorous definition of abductive/constrastive explanations
- Highlighted the flaws of Shapley values for XAI, i.e. the SHAP scores
- Detailed corrections to Shapley values for XAI
  - Right connection between logic-based XAI by feature selection and (corrected) XAI by feature attribution
- Overviewed nuSHAP
  - Still a prototype...
- Research directions?
  - Several ongoing collaborations...

Conclusions

- Brief contact with logic-based symbolic XAI
  - Rigorous definition of abductive/constrastive explanations
- Highlighted the flaws of Shapley values for XAI, i.e. the SHAP scores
- Detailed corrections to Shapley values for XAI
  - Right connection between logic-based XAI by feature selection and (corrected) XAI by feature attribution
- Overviewed nuSHAP
  - Still a prototype...
- Research directions?
  - Several ongoing collaborations...
  - ... EU funding agencies averse to sponsor our research on formal XAI... 😕

Conclusions

- Brief contact with logic-based symbolic XAI
  - Rigorous definition of abductive/constrastive explanations
- Highlighted the flaws of Shapley values for XAI, i.e. the SHAP scores
- Detailed corrections to Shapley values for XAI
  - Right connection between logic-based XAI by feature selection and (corrected) XAI by feature attribution
- Overviewed nuSHAP
  - Still a prototype...
- Research directions?
  - Several ongoing collaborations...
  - ... EU funding agencies averse to sponsor our research on formal XAI... 😕
    - Non-symbolic XAI methods, e.g. SHAP, rather popular ... 😕😕

Conclusions

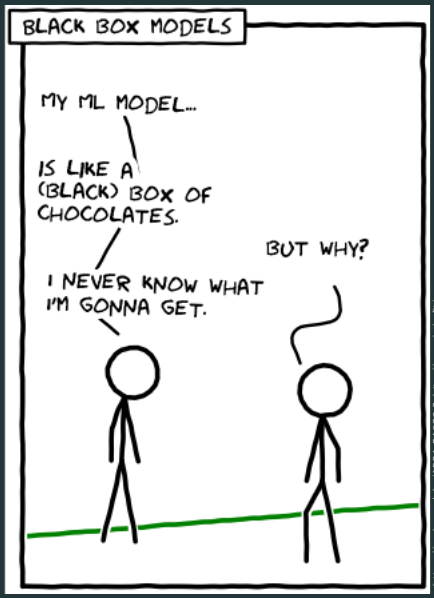
- Brief contact with logic-based symbolic XAI
  - Rigorous definition of abductive/constrastive explanations
- Highlighted the flaws of Shapley values for XAI, i.e. the SHAP scores
- Detailed corrections to Shapley values for XAI
  - Right connection between logic-based XAI by feature selection and (corrected) XAI by feature attribution
- Overviewed nuSHAP
  - Still a prototype...
- Research directions?
  - Several ongoing collaborations...
  - ... EU funding agencies averse to sponsor our research on formal XAI... 😕
    - Non-symbolic XAI methods, e.g. SHAP, rather popular ... 😕😕
    - **Challenge accepted!**

Q & A

"Research is seeing what everybody else has seen and thinking what nobody else has thought."

(A. Szent-Gyorgyi)

Acknowledgments:
A. Ignatiev, X. Huang, Y. Izza, O. Létoffé, M. Cooper, N. Asher, N. Narodytska, J. Planes, A. Morgado, R. Bejar, C. Mencía, P. Stuckey, R. Passos, F. Chiariello, M. Siala, J. Lefebre-Lobaina, A. Hurault, H. Hu et al.
And: Gemini & ChatGPT


BLACK BOX MODELS
MY ML MODEL...
IS LIKE A (BLACK) BOX OF CHOCOLATES.
I NEVER KNOW WHAT I'M GONNA GET.
BUT WHY?
https://arxiv.org/abs/1901.01686 & http://cmx.io/edit/

References i

[ABBM21] Marcelo Arenas, Pablo Barceló, Leopoldo E. Bertossi, and Mikaël Monet.
The tractability of SHAP-score-based explanations for classification over deterministic and decomposable boolean circuits.
In *AAAI*, pages 6670–6678, 2021.

[ABBM23] Marcelo Arenas, Pablo Barceló, Leopoldo E. Bertossi, and Mikaël Monet.
On the complexity of SHAP-score-based explanations: Tractability via knowledge compilation and non-approximability results.
J. Mach. Learn. Res., 24:63:1–63:58, 2023.

[BKR+25] Pablo Barceló, Alexander Kozachinskiy, Miguel Romero, Bernardo Subercaseaux, and José Verschae.
Explaining k-nearest neighbors: Abductive and counterfactual explanations.
Proc. ACM Manag. Data, 3(2):97:1–97:26, 2025.

[CA23] Martin C. Cooper and Leila Amgoud.
Abductive explanations of classifiers under constraints: Complexity and properties.
In *ECAI*, pages 469–476, 2023.

[CGT09] Javier Castro, Daniel Gómez, and Juan Tejada.
Polynomial calculation of the shapley value based on sampling.
Comput. Oper. Res., 36(5):1726–1730, 2009.

References ii

[CM21] Martin C. Cooper and Joao Marques-Silva.
On the tractability of explaining decisions of classifiers.
In *CP*, October 2021.

[CM23] Martin C. Cooper and João Marques-Silva.
Tractability of explaining classifier decisions.
Artif. Intell., 316:103841, 2023.

[DSZ16] Anupam Datta, Shayak Sen, and Yair Zick.
Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems.
In *IEEE S&P*, pages 598–617, 2016.

[EG95] Thomas Eiter and Georg Gottlob.
The complexity of logic-based abduction.
J. ACM, 42(1):3–42, 1995.

[HII+22] Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, Martin Cooper, Nicholas Asher, and Joao Marques-Silva.
Tractable explanations for d-DNNF classifiers.
In *AAAI*, February 2022.

References iii

[HIIM21] Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva.
On efficiently explaining graph-based classifiers.
In *KR*, November 2021.
Preprint available from `https://arxiv.org/abs/2106.01350`.

[HM23a] Xuanxiang Huang and João Marques-Silva.
From robustness to explainability and back again.
CoRR, abs/2306.03048, 2023.

[HM23b] Xuanxiang Huang and João Marques-Silva.
The inadequacy of Shapley values for explainability.
CoRR, abs/2302.08160, 2023.

[HM23c] Xuanxiang Huang and Joao Marques-Silva.
A refutation of shapley values for explainability.
CoRR, abs/2309.03041, 2023.

[HM23d] Xuanxiang Huang and Joao Marques-Silva.
Refutation of shapley values for XAI – additional evidence.
CoRR, abs/2310.00416, 2023.

References iv

[HM24] Xuanxiang Huang and João Marques-Silva.
On the failings of shapley values for explainability.
Int. J. Approx. Reason., 171:109112, 2024.

[Ign20] Alexey Ignatiev.
Towards trustable explainable AI.
In *IJCAI*, pages 5154–5158, 2020.

[IHM+24a] Yacine Izza, Xuanxiang Huang, António Morgado, Jordi Planes, Alexey Ignatiev, and João Marques-Silva.
Distance-restricted explanations: Theoretical underpinnings & efficient implementation.
In *KR*, 2024.

[IHM+24b] Yacine Izza, Xuanxiang Huang, Antonio Morgado, Jordi Planes, Alexey Ignatiev, and Joao Marques-Silva.
Distance-restricted explanations: Theoretical underpinnings & efficient implementation.
CoRR, abs/2405.08297, 2024.

[IIM20] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva.
On explaining decision trees.
CoRR, abs/2010.11034, 2020.

[IIM22] Yacine Izza, Alexey Ignatiev, and João Marques-Silva.
On tackling explanation redundancy in decision trees.
J. Artif. Intell. Res., 75:261–321, 2022.

References v

[IISM24] Yacine Izza, Alexey Ignatiev, Peter J. Stuckey, and João Marques-Silva.
Delivering inflated explanations.
In *AAAI*, pages 12744–12753, 2024.

[IISMS22] Alexey Ignatiev, Yacine Izza, Peter J. Stuckey, and Joao Marques-Silva.
Using MaxSAT for efficient explanations of tree ensembles.
In *AAAI*, February 2022.

[IM21] Alexey Ignatiev and Joao Marques-Silva.
SAT-based rigorous explanations for decision lists.
In *SAT*, pages 251–269, July 2021.

[IMS21] Yacine Izza and Joao Marques-Silva.
On explaining random forests with SAT.
In *IJCAI*, pages 2584–2591, July 2021.

[INAM20] Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and João Marques-Silva.
From contrastive to abductive explanations and back again.
In *AIxIA*, pages 335–355, 2020.

[INM19a] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva.
Abduction-based explanations for machine learning models.
In *AAAI*, pages 1511–1519, 2019.

References vi

[INM19b] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva.
On relating explanations and adversarial examples.
In *NeurIPS*, pages 15857–15867, 2019.

[INM19c] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva.
On validating, repairing and refining heuristic ML explanations.
CoRR, abs/1907.02509, 2019.

[LC01] Stan Lipovetsky and Michael Conklin.
Analysis of regression in game theory approach.
Applied Stochastic Models in Business and Industry, 17(4):319–330, 2001.

[LHAM24] Olivier Letoffe, Xuanxiang Huang, Nicholas Asher, and João Marques-Silva.
From SHAP scores to feature importance scores.
CoRR, abs/2405.11766, 2024.

[LHM24a] Olivier Letoffe, Xuanxiang Huang, and João Marques-Silva.
On correcting SHAP scores.
CoRR, abs/2405.00076, 2024.

[LHM24b] Olivier Letoffe, Xuanxiang Huang, and João Marques-Silva.
SHAP scores fail pervasively even when Lipschitz succeeds.
Under review, July 2024.

References vii

[LL17] Scott M. Lundberg and Su-In Lee.
A unified approach to interpreting model predictions.
In *NIPS*, pages 4765–4774, 2017.

[MGC+20] Joao Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska.
Explaining naive bayes and other linear classifiers with polynomial time and delay.
In *NeurIPS*, 2020.

[MGC+21] Joao Marques-Silva, Thomas Gerspacher, Martinc C. Cooper, Alexey Ignatiev, and Nina Narodytska.
Explanations for monotonic classifiers.
In *ICML*, pages 7469–7479, July 2021.

[MH23] Joao Marques-Silva and Xuanxiang Huang.
Explainability is NOT a game.
CoRR, abs/2307.07514, 2023.

[MH24] João Marques-Silva and Xuanxiang Huang.
Explainability is *Not* a game.
Commun. ACM, 67(7):66–75, 2024.

[MHL25] João Marques-Silva, Xuanxiang Huang, and Olivier Letoffe.
The explanation game - rekindled (extended version).
CoRR, abs/2501.11429, 2025.

References viii

[Mil19] Tim Miller.
Explanation in artificial intelligence: Insights from the social sciences.
Artif. Intell., 267:1–38, 2019.

[MLM25] João Marques-Silva, Jairo A. Lefebre-Lobaina, and Maria Vanina Martinez.
Efficient and rigorous model-agnostic explanations.
In *IJCAI*, pages 2637–2646, 2025.

[MS24] Joao Marques-Silva.
Logic-based explainability: Past, present and future.
In *ISoLA*, pages 181–204, 2024.

[Rei87] Raymond Reiter.
A theory of diagnosis from first principles.
Artif. Intell., 32(1):57–95, 1987.

[RS23] Cynthia Rudin and Yaron Shaposhnik.
Globally-consistent rule-based summary-explanations for machine learning models: Application to credit-risk evaluation.
J. Mach. Learn. Res., 24:16:1–16:44, 2023.

References ix

[RSG16] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin.
"why should I trust you?": Explaining the predictions of any classifier.
In *KDD*, pages 1135–1144, 2016.

[RSG18] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin.
Anchors: High-precision model-agnostic explanations.
In *AAAI*, pages 1527–1535. AAAI Press, 2018.

[Rud19] Cynthia Rudin.
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.
Nature Machine Intelligence, 1(5):206–215, 2019.

[SCD18] Andy Shih, Arthur Choi, and Adnan Darwiche.
A symbolic approach to explaining bayesian network classifiers.
In *IJCAI*, pages 5103–5111, 2018.

[Sha53] Lloyd S. Shapley.
A value for *n*-person games.
Contributions to the Theory of Games, 2(28):307–317, 1953.

References x

[SK10] Erik Strumbelj and Igor Kononenko.
An efficient explanation of individual classifications using game theory.
J. Mach. Learn. Res., 11:1–18, 2010.

[SK14] Erik Strumbelj and Igor Kononenko.
Explaining prediction models and individual predictions with feature contributions.
Knowl. Inf. Syst., 41(3):647–665, 2014.

[VLSS21] Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu.
On the tractability of SHAP explanations.
In *AAAI*, pages 6505–6513, 2021.

[VLSS22] Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu.
On the tractability of SHAP explanations.
J. Artif. Intell. Res., 74:851–886, 2022.

[WMZ10] William Webber, Alistair Moffat, and Justin Zobel.
A similarity measure for indefinite rankings.
ACM Trans. Inf. Syst., 28(4):20:1–20:38, 2010.