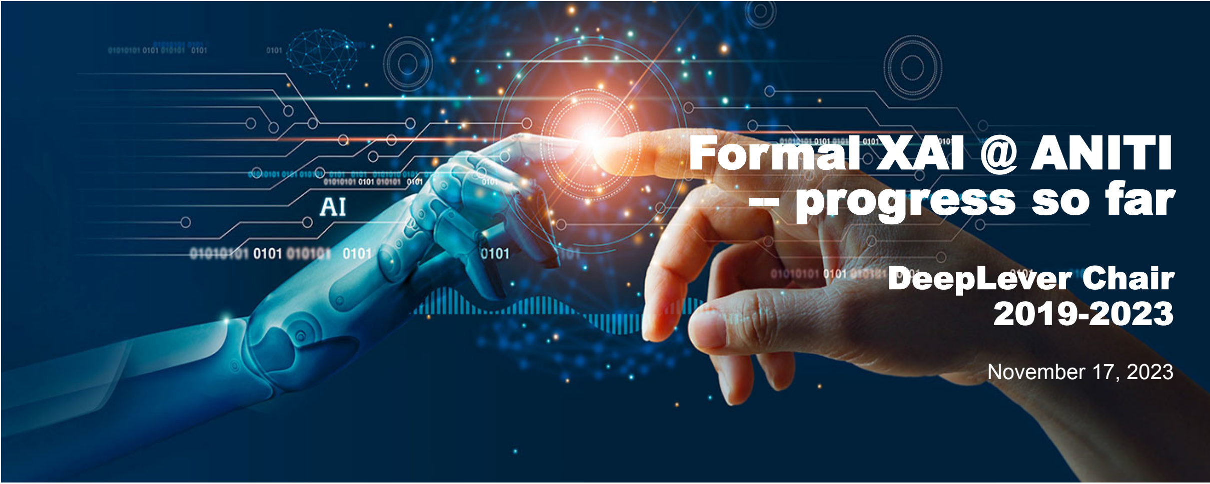


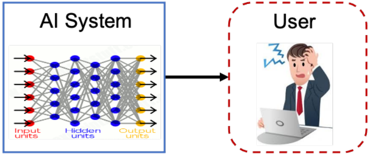
Formal XAI @ ANITI -- progress so far

DeepLever Chair 2019-2023

November 17, 2023

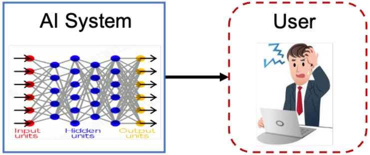
ANITI
Université de Toulouse

XAI & high-risk uses

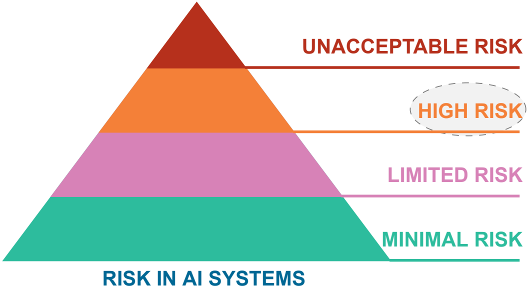


XAI: to help humans understand ML models

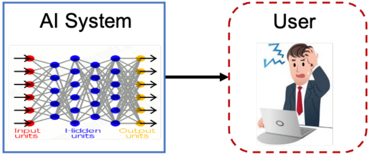
XAI & high-risk uses



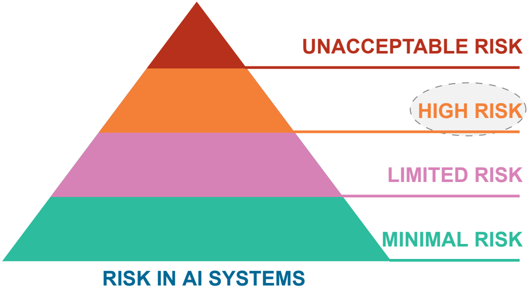
XAI: to help humans understand ML models



XAI & high-risk uses



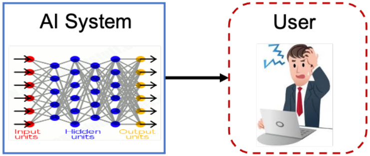
XAI: to help humans understand ML models

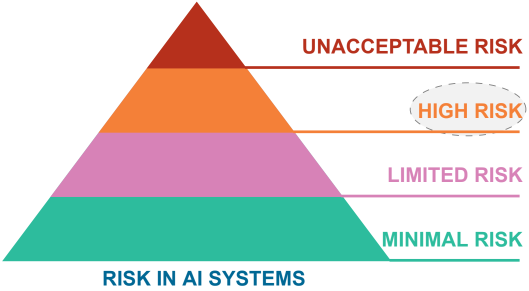


Many examples of high-risk uses: [Pro21]

- Credit worthiness & Law enforcement
- Management and operation of critical infrastructure
- Biometric identification and categorization of people; ...

XAI & high-risk uses





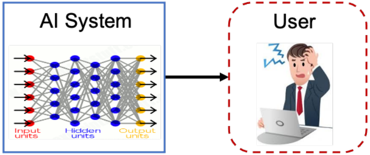
XAI: to help humans understand ML models

Many examples of high-risk uses: [Pro21]

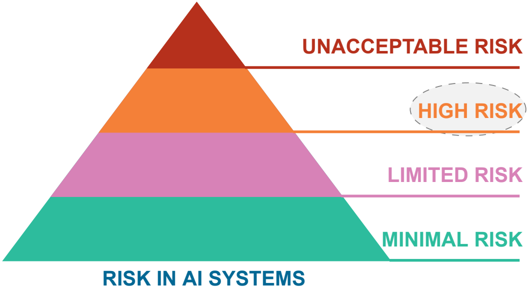
- Credit worthiness & Law enforcement
- Management and operation of critical infrastructure
- Biometric identification and categorization of people; ...

And also safety-critical uses !

XAI & high-risk uses -- focus of DeepLever Chair



XAI: to help humans understand ML models



Many examples of high-risk uses: [Pro21]

- Credit worthiness & Law enforcement
- Management and operation of critical infrastructure
- Biometric identification and categorization of people; ...

And also safety-critical uses !

Pervasive hallmarks of non-formal XAI

[RSG16, LL17, RSG18, Rud19]

“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

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

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

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

Pervasive hallmarks of non-formal XAI

LIME, SHAP; Anchor; Interpretability, ...

[RSG16, LL17, RSG18, Rud19]

“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

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

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

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

Pervasive hallmarks of non-formal XAI

LIME, SHAP; Anchor; Interpretability, ...

[RSG16, LL17, RSG18, Rud19]

"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

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

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

Anchors: High-Precision Model-Agnostic Explanations

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

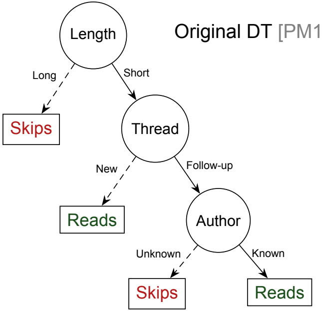
Sameer Singh
University of California, Irvine
sameer@uci.edu

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

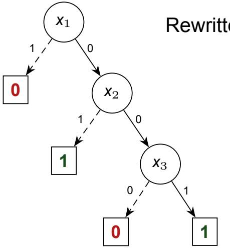
∴ We have disproved **ALL** these hallmarks. More detail later

What is an explanation?

Original DT [PM17]



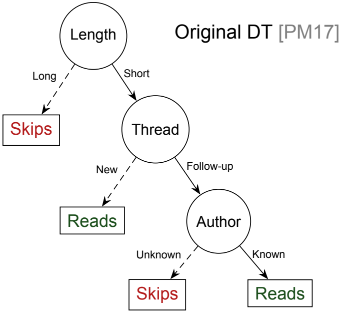
Rewritten DT

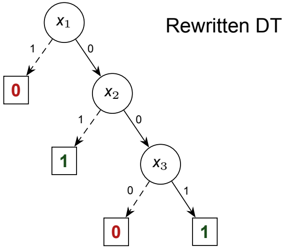


Mapping
$x_1 = 1$ iff Length $=$ Long
$x_2 = 1$ iff Thread $=$ New
$x_3 = 1$ iff Author $=$ Known
$\kappa(\cdot) = 1$ iff $\kappa'(\cdots) =$ Reads
$\kappa(\cdot) = 0$ iff $\kappa'(\cdots) =$ Skips

- What is an explanation?

What is an explanation?

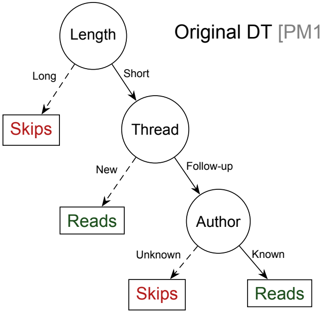




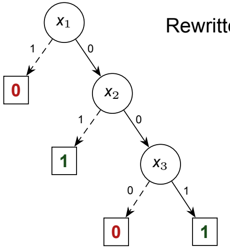
Mapping
$x_1 = 1$ iff Length $=$ Long
$x_2 = 1$ iff Thread $=$ New
$x_3 = 1$ iff Author $=$ Known
$\kappa(\cdot) = 1$ iff $\kappa'(\cdots) =$ Reads
$\kappa(\cdot) = 0$ iff $\kappa'(\cdots) =$ Skips

- What is an explanation?
  - Answer to question "**Why** (the prediction)?" is a **rule**: **IF <COND> THEN** $\kappa(\mathrm{x}) = c$

What is an explanation?



Original DT [PM17]



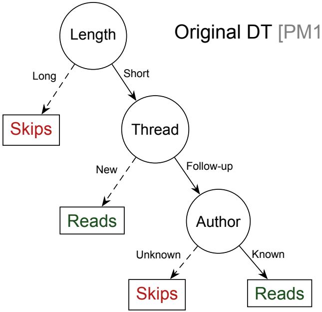
Rewritten DT

Mapping
$x_1 = 1$ iff Length = Long
$x_2 = 1$ iff Thread = New
$x_3 = 1$ iff Author = Known
$\kappa(\cdot) = 1$ iff $\kappa'(\cdots) =$ Reads
$\kappa(\cdot) = 0$ iff $\kappa'(\cdots) =$ Skips

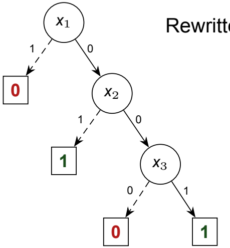
- What is an explanation?
  - Answer to question "**Why** (the prediction)?" is a **rule**: **IF <COND> THEN** $\kappa(\mathbf{x}) = c$
  - **Explanation**: set of literals (or just features) in **<COND>**; **irreducibility matters !**

What is an explanation?

Original DT [PM17]



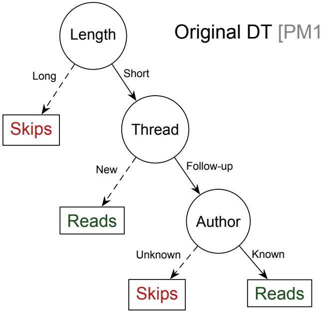
Rewritten DT



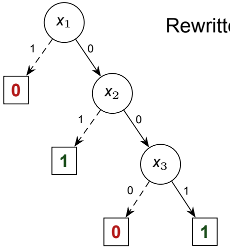
Mapping
$x_1 = 1$ iff Length = Long
$x_2 = 1$ iff Thread = New
$x_3 = 1$ iff Author = Known
$\kappa(\cdot) = 1$ iff $\kappa'(\cdots) =$ Reads
$\kappa(\cdot) = 0$ iff $\kappa'(\cdots) =$ Skips

- What is an explanation?
  - Answer to question "**Why** (the prediction)?" is a **rule**: **IF <COND> THEN** $\kappa(\mathbf{x}) = c$
  - **Explanation**: set of literals (or just features) in **<COND>**; **irreducibility matters !**
  - **E.g.**: explanation for $\mathbf{v} = (\neg x_1, \neg x_2, x_3)$?

What is an explanation?



Original DT [PM17]

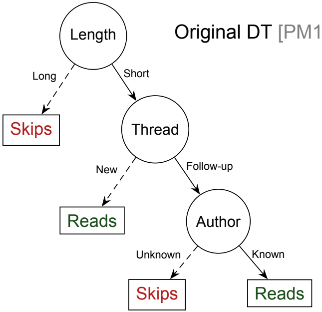


Rewritten DT

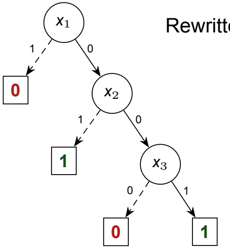
Mapping
$x_1 = 1$ iff Length $=$ Long
$x_2 = 1$ iff Thread $=$ New
$x_3 = 1$ iff Author $=$ Known
$\kappa(\cdot) = 1$ iff $\kappa'(\cdots) =$ Reads
$\kappa(\cdot) = 0$ iff $\kappa'(\cdots) =$ Skips

- What is an explanation?
  - Answer to question "**Why** (the prediction)?" is a **rule**: **IF <COND> THEN** $\kappa(\mathbf{x}) = c$
  - **Explanation**: set of literals (or just features) in **<COND>**; **irreducibility matters !**
  - **E.g.**: explanation for $\mathbf{v} = (\neg x_1, \neg x_2, x_3)$?
    - It is the case that, IF $\neg x_1 \wedge \neg x_2 \wedge x_3$ THEN $\kappa(\mathbf{x}) = 1$

What is an explanation?



Original DT [PM17]



Rewritten DT

Mapping
$x_1 = 1$ iff Length $=$ Long
$x_2 = 1$ iff Thread $=$ New
$x_3 = 1$ iff Author $=$ Known
$\kappa(\cdot) = 1$ iff $\kappa'(\cdots) =$ Reads
$\kappa(\cdot) = 0$ iff $\kappa'(\cdots) =$ Skips

- What is an explanation?
  - Answer to question "**Why** (the prediction)?" is a **rule**: **IF <COND> THEN $\kappa(\mathbf{x}) = c$**
  - **Explanation**: set of literals (or just features) in **<COND>**; **irreducibility matters !**
  - **E.g.**: explanation for $\mathbf{v} = (\neg x_1, \neg x_2, x_3)$?
    - It is the case that, IF $\neg x_1 \wedge \neg x_2 \wedge x_3$ THEN $\kappa(\mathbf{x}) = 1$
    - Explanation is $\{\neg x_1, \neg x_2, x_3\}$ or simply $\{1, 2, 3\}$

Formal explanations

Formal XAI in classification:

- Explanations rigorously defined

Formal explanations

Formal XAI in classification:

- Explanations rigorously defined
- Explanation for **Why?** question:
  - Minimal set of features sufficient for ensuring prediction $c = \kappa(\mathbf{v})$
  - I.e. pick minimal $\mathcal{X} \subseteq \mathcal{F}$ s.t.

$$\forall(\mathbf{z} \in \mathbb{F}).\, [\wedge_{i \in \mathcal{X}}(z_i = v_i) \rightarrow (\kappa(\mathbf{z}) = c)]$$

Formal explanations

Formal XAI in classification:

- Explanations rigorously defined
- Explanation for **Why?** question:
  - Minimal set of features sufficient for ensuring prediction $c = \kappa(\mathbf{v})$
  - I.e. pick minimal $\mathcal{X} \subseteq \mathcal{F}$ s.t.

$$\forall(\mathbf{z} \in \mathbb{F}). \left[\wedge_{i\in\mathcal{X}}(z_i = v_i) \rightarrow (\kappa(\mathbf{z}) = c)\right]$$

Represents a rule:
IF $\wedge_{i\in\mathcal{X}} (z_i = v_i)$ THEN $(\kappa(\mathbf{z}) = c)$

Formal explanations

Formal XAI in classification:

- Explanations rigorously defined
- Explanation for **Why?** question:
  - Minimal set of features sufficient for ensuring prediction $c = \kappa(\mathbf{v})$
  - I.e. pick minimal $\mathcal{X} \subseteq \mathcal{F}$ s.t.

$$\forall(\mathbf{z} \in \mathbb{F}). \left[\wedge_{i \in \mathcal{X}}(z_i = v_i) \rightarrow (\kappa(\mathbf{z}) = c)\right]$$

Represents a rule:
IF $\wedge_{i \in \mathcal{X}}(z_i = v_i)$ THEN $(\kappa(\mathbf{z}) = c)$

- Explanation for **Why Not?** question:
  - Minimal set of features sufficient for changing prediction $c = \kappa(\mathbf{v})$
  - I.e. pick minimal $\mathcal{Y} \subseteq \mathcal{F}$ s.t.

$$\exists(\mathbf{z} \in \mathbb{F}). \left[\wedge_{i \notin \mathcal{Y}}(z_i = v_i) \wedge (\kappa(\mathbf{z}) \neq c)\right]$$

Formal explanations

Formal XAI in classification:

- Explanations rigorously defined
- Explanation for **Why?** question:
  - Minimal set of features sufficient for ensuring prediction $c = \kappa(\mathbf{v})$
  - I.e. pick minimal $\mathcal{X} \subseteq \mathcal{F}$ s.t.

$$\forall(\mathbf{z} \in \mathbb{F}).\left[\wedge_{i \in \mathcal{X}}(z_i = v_i) \rightarrow (\kappa(\mathbf{z}) = c)\right]$$

Represents a rule:
IF $\wedge_{i \in \mathcal{X}} (z_i = v_i)$ THEN $(\kappa(\mathbf{z}) = c)$

- Explanation for **Why Not?** question:
  - Minimal set of features sufficient for changing prediction $c = \kappa(\mathbf{v})$
  - I.e. pick minimal $\mathcal{Y} \subseteq \mathcal{F}$ s.t.

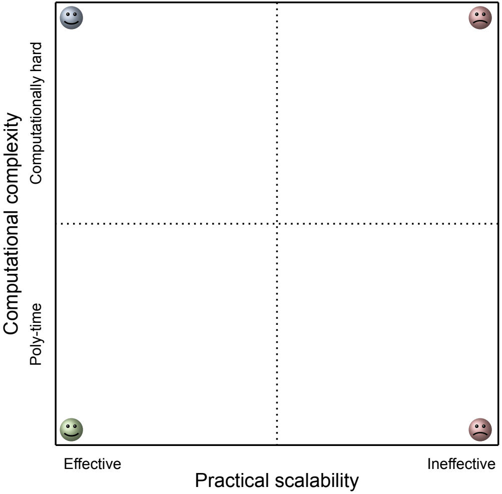
$$\exists(\mathbf{z} \in \mathbb{F}).\left[\wedge_{i \notin \mathcal{Y}}(z_i = v_i) \wedge (\kappa(\mathbf{z}) \neq c)\right]$$

- Duality results, e.g. between XPs for **Why?** and **Why Not?** questions [INAM20, INM19a]
- More problems: enumeration, membership, preferences, ...

Progress in formal XAI -- until 2022

[INM19b, IIM20, MGC+20, MGC+21, HIIM21, IM21, IMS21, CM21, IIM22, HII+22, IISMS22]

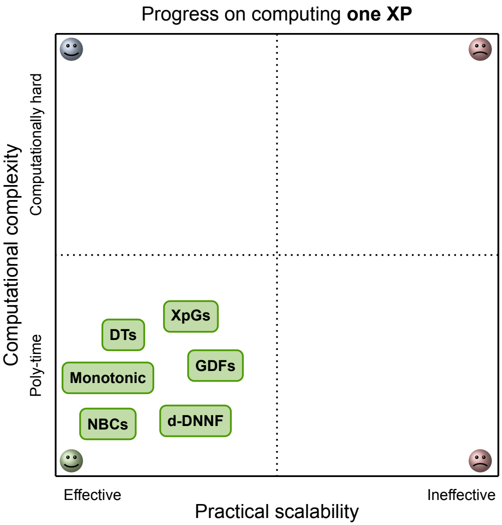
Progress on computing **one XP**



- Advances between 2019 and 2022:

Progress in formal XAI -- until 2022

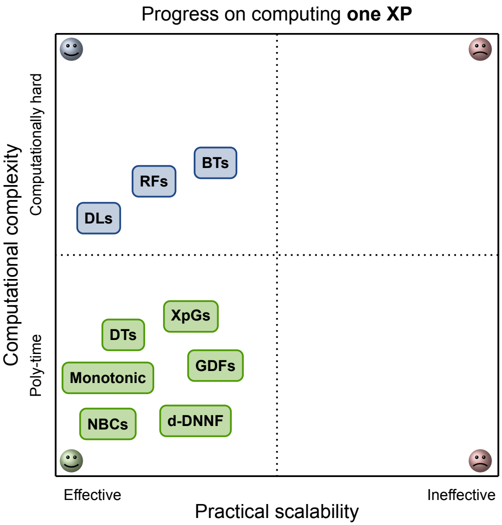
[INM19b, IIM20, MGC+20, MGC+21, HIIM21, IM21, IMS21, CM21, IIM22, HII+22, IISMS22]



- Advances between 2019 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]

Progress in formal XAI -- until 2022

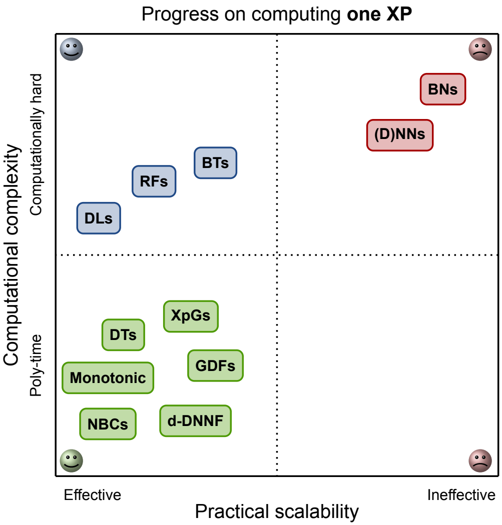
[INM19b, IIM20, MGC$^+$20, MGC$^+$21, HIIM21, IM21, IMS21, CM21, IIM22, HII$^+$22, IISMS22]



- Advances between 2019 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]
  - Comp. hard, but effective (efficient in practice):
    - Random forests (RFs) [IM21, IISMS22]
    - Decision lists (DLs) [IMS21]
    - Boosted trees (BTs) [INM19b, IISMS22]

Progress in formal XAI -- until 2022

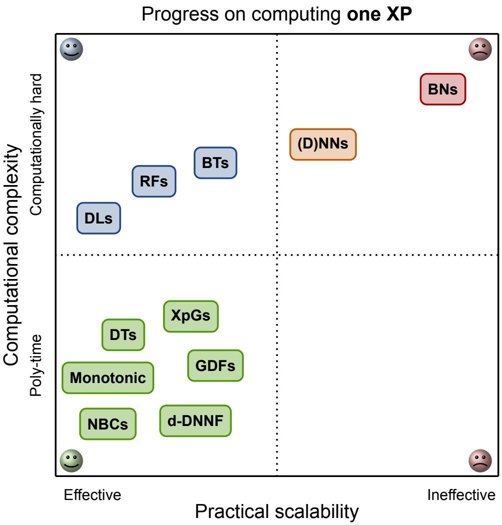
[INM19b, IIM20, MGC+20, MGC+21, HIIM21, IM21, IMS21, CM21, IIM22, HII+22, IISMS22]



- Advances between 2019 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]
  - Comp. hard, but effective (efficient in practice):
    - Random forests (RFs) [IM21, IISMS22]
    - Decision lists (DLs) [IMS21]
    - Boosted trees (BTs) [INM19b, IISMS22]
  - Comp. hard, and ineffective (hard in practice):
    - Neural networks (NNs) [INMS19]
    - Bayesian networks (BNs) [SCD18]

Progress in formal XAI -- recent progress

[INM19b, IIM20, MGC+20, MGC+21, HIIM21, IM21, IMS21, CM21, IIM22, HII+22, IISMS22, HM23a]



- Advances up until 2023:
  - 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]
  - Comp. hard, but effective (efficient in practice):
    - Random forests (RFs) [IM21, IISMS22]
    - Decision lists (DLs) [IMS21]
    - Boosted trees (BTs) [INM19b, IISMS22]
  - Comp. hard, but some practical scalability:
    - Neural networks (NNs) [HM23a]
  - Comp. hard, and ineffective (hard in practice):
    - Bayesian networks (BNs) [SCD18]

Dataset	(#F	#C	#I)	RF			CNF		SAT oracle				AXp (RFxpl)				Anchor	
				D	#N	%A	#var	#cl	MxS	MxU	#S	#U	Mx	m	avg	%w	avg	%w
ann-thyroid	(21	3	718)	4	2192	98	17854	29230	0.12	0.15	2	18	0.36	0.05	0.13	96	0.32	4
appendicitis	(7	2	43)	6	1920	90	5181	10085	0.02	0.02	4	3	0.05	0.01	0.03	100	0.48	0
banknote	(4	2	138)	5	2772	97	8068	16776	0.01	0.01	2	2	0.03	0.02	0.02	100	0.19	0
biodegradation	(41	2	106	5	4420	88	11007	23842	0.31	1.05	17	22	2.27	0.04	0.29	97	4.07	3
heart-c	(13	2	61)	5	3910	85	5594	11963	0.04	0.02	6	7	0.07	0.01	0.04	100	0.85	0
ionosphere	(34	2	71)	5	2096	87	7174	14406	0.02	0.02	22	11	0.11	0.02	0.03	100	12.43	0
karhunen	(64	10	200)	5	6198	91	36708	70224	1.06	1.41	35	29	14.64	0.65	2.78	100	28.15	0
letter	(16	26	398	8	44304	82	28991	68148	1.97	3.31	8	8	6.91	0.24	1.61	70	2.48	30
magic	(10	2	381)	6	9840	84	29530	66776	0.51	1.84	6	4	2.13	0.07	0.14	99	0.91	1
new-thyroid	(5	3	43)	5	1766	100	17443	28134	0.03	0.01	3	2	0.08	0.03	0.05	100	0.36	0
pendigits	(16	10	220)	6	12004	95	30522	59922	2.40	1.32	10	6	4.11	0.14	0.94	96	3.68	4
ring	(20	2	740	6	6188	89	19114	42362	0.27	0.44	11	9	1.25	0.05	0.25	92	7.25	8
segmentation	(19	7	42)	4	1966	90	21288	35381	0.11	0.17	8	10	0.53	0.11	0.31	100	4.13	0
shuttle	(9	7	116	3	1460	99	18669	29478	0.11	0.08	2	7	0.34	0.05	0.14	99	0.42	1
sonar	(60	2	42)	5	2614	88	9938	20537	0.04	0.06	36	24	0.43	0.04	0.09	100	23.02	0
spectf	(44	2	54)	5	2306	88	6707	13449	0.07	0.06	20	24	0.34	0.02	0.07	100	8.12	0
texture	(40	11	550)	5	5724	87	34293	64187	0.79	0.63	23	17	3.24	0.19	0.93	100	28.13	0
twonorm	(20	2	740	5	6266	94	21198	46901	0.08	0.08	12	8	0.28	0.06	0.10	100	5.73	0
vowel	(13	11	198)	6	10176	90	44523	88696	1.66	2.11	8	5	4.52	0.15	1.15	66	1.67	34
waveform-40	(40	3	500	5	6232	83	30438	58380	0.50	0.86	15	25	7.07	0.11	0.88	100	11.93	0
wpbc	(33	2	78)	5	2432	76	9078	18675	1.00	1.53	20	13	5.33	0.03	0.65	79	3.91	21

Results for RFs in 2021 (with SAT) [IM21]

Dataset	(#F	#C	#I)	RF			CNF		SAT oracle				AXp (RFxpl)				Anchor	
				D	#N	%A	#var	#cl	MxS	MxU	#S	#U	Mx	m	avg	%w	avg	%w
ann-thyroid	(21	3	718)	4	2192	98	17854	29230	0.12	0.15	2	18	0.36	0.05	0.13	96	0.32	4
appendicitis	(7	2	43)	6	1920	90	5181	10085	0.02	0.02	4	3	0.05	0.01	0.03	100	0.48	0
banknote	(4	2	138)	5	2772	97	8068	16776	0.01	0.01	2	2	0.03	0.02	0.02	100	0.19	0
biodegradation	(41	2	106	5	4420	88	11007	23842	0.31	1.05	17	22	2.27	0.04	0.29	97	4.07	3
heart-c	(13	2	61)	5	3910	85	5594	11963	0.04	0.02	6	7	0.07	0.01	0.04	100	0.85	0
ionosphere	(34	2	71)	5	2096	87	7174	14406	0.02	0.02	22	11	0.11	0.02	0.03	100	12.43	0
karhunen	(64	10	200)	5	6198	91	36708	70224	1.06	1.41	35	29	14.64	0.65	2.78	100	28.15	0
letter	(16	26	398	8	44304	82	28991	68148	1.97	3.31	8	8	6.91	0.24	1.61	70	2.48	30
magic	(10	2	381)	6	9840	84	29530	66776	0.51	1.84	6	4	2.13	0.07	0.14	99	0.91	1
new-thyroid	(5	3	43)	5	1766	100	17443	28134	0.03	0.01	3	2	0.08	0.03	0.05	100	0.36	0
pendigits	(16	10	220)	6	12004	95	30522	59922	2.40	1.32	10	6	4.11	0.14	0.94	96	3.68	4
ring	(20	2	740	6	6188	89	19114	42362	0.27	0.44	11	9	1.25	0.05	0.25	92	7.25	8
segmentation	(19	7	42)	4	1966	90	21288	35381	0.11	0.17	8	10	0.53	0.11	0.31	100	4.13	0
shuttle	(9	7	116	3	1460	99	18669	29478	0.11	0.08	2	7	0.34	0.05	0.14	99	0.42	1
sonar	(60	2	42)	5	2614	88	9938	20537	0.04	0.06	36	24	0.43	0.04	0.09	100	23.02	0
spectf	(44	2	54)	5	2306	88	6707	13449	0.07	0.06	20	24	0.34	0.02	0.07	100	8.12	0
texture	(40	11	550)	5	5724	87	34293	64187	0.79	0.63	23	17	3.24	0.19	0.93	100	28.13	0
twonorm	(20	2	740	5	6266	94	21198	46901	0.08	0.08	12	8	0.28	0.06	0.10	100	5.73	0
vowel	(13	11	198)	6	10176	90	44523	88696	1.66	2.11	8	5	4.52	0.15	1.15	66	1.67	34
waveform-40	(40	3	500	5	6232	83	30438	58380	0.50	0.86	15	25	7.07	0.11	0.88	100	11.93	0
wpbc	(33	2	78)	5	2432	76	9078	18675	1.00	1.53	20	13	5.33	0.03	0.65	79	3.91	21

Rigorous & faster than Anchor !

Results for NNs in 2019 (w/ SMT/MILP) [INMS19]

Dataset			Minimal explanation			Minimum explanation		
			size	SMT (s)	MILP (s)	size	SMT (s)	MILP (s)
australian	(14)	m	1	0.03	0.05	—	—	—
		a	8.79	1.38	0.33	—	—	—
		M	14	17.00	1.43	—	—	—
backache	(32)	m	13	0.13	0.14	—	—	—
		a	19.28	5.08	0.85	—	—	—
		M	26	22.21	2.75	—	—	—
breast-cancer	(9)	m	3	0.02	0.04	3	0.02	0.03
		a	5.15	0.65	0.20	4.86	2.18	0.41
		M	9	6.11	0.41	9	24.80	1.81
cleve	(13)	m	4	0.05	0.07	4	—	0.07
		a	8.62	3.32	0.32	7.89	—	5.14
		M	13	60.74	0.60	13	—	39.06
hepatitis	(19)	m	6	0.02	0.04	4	0.01	0.04
		a	11.42	0.07	0.06	9.39	4.07	2.89
		M	19	0.26	0.20	19	27.05	22.23
voting	(16)	m	3	0.01	0.02	3	0.01	0.02
		a	4.56	0.04	0.13	3.46	0.3	0.25
		M	11	0.10	0.37	11	1.25	1.77
spect	(22)	m	3	0.02	0.02	3	0.02	0.04
		a	7.31	0.13	0.07	6.44	1.61	0.67
		M	20	0.88	0.29	20	8.97	10.73

Results for NNs in 2019 (w/ SMT/MILP) [INMS19]

First rigorous approach for **explaining** NNs !

			Minimal explanation			Minimum explanation		
			size	SMT (s)	MILP (s)	size	SMT (s)	MILP (s)
australian	(14)	m	1	0.03	0.05	—	—	—
		a	8.79	1.38	0.33	—	—	—
		M	14	17.00	1.43	—	—	—
backache	(32)	m	13	0.13	0.14	—	—	—
		a	19.28	5.08	0.85	—	—	—
		M	26	22.21	2.75	—	—	—
breast-cancer	(9)	m	3	0.02	0.04	3	0.02	0.03
		a	5.15	0.65	0.20	4.86	2.18	0.41
		M	9	6.11	0.41	9	24.80	1.81
cleve	(13)	m	4	0.05	0.07	4	—	0.07
		a	8.62	3.32	0.32	7.89	—	5.14
		M	13	60.74	0.60	13	—	39.06
hepatitis	(19)	m	6	0.02	0.04	4	0.01	0.04
		a	11.42	0.07	0.06	9.39	4.07	2.89
		M	19	0.26	0.20	19	27.05	22.23
voting	(16)	m	3	0.01	0.02	3	0.01	0.02
		a	4.56	0.04	0.13	3.46	0.3	0.25
		M	11	0.10	0.37	11	1.25	1.77
spect	(22)	m	3	0.02	0.02	3	0.02	0.04
		a	7.31	0.13	0.07	6.44	1.61	0.67
		M	20	0.88	0.29	20	8.97	10.73

Results for NNs in 2019 (w/ SMT/MILP) [INMS19]

First rigorous approach for **explaining** NNs !

			Minimal explanation			Minimum explanation		
			size	SMT (s)	MILP (s)	size	SMT (s)	MILP (s)
australian	(14)	m	1	0.03	0.05	—	—	—
		a	8.79	1.38	0.33	—	—	—
		M	14	17.00	1.43	—	—	—
backache	(32)	m	13	0.13	0.14	—	—	—
		a	19.28	5.08	0.85	—	—	—
		M	26	22.21	2.75	—	—	—
breast-cancer	(9)	m	3	0.02	0.04	3	0.02	0.03
		a	5.15	0.65	0.20	4.86	2.18	0.41
		M	9	6.11	0.41	9	24.80	1.81
cleve	(13)	m	4	0.05	0.07	4	—	0.07
		a	8.62	3.32	0.32	7.89	—	5.14
		M	13	60.74	0.60	13	—	39.06
hepatitis	(19)	m	6	0.02	0.04	4	0.01	0.04
		a	11.42	0.07	0.06	9.39	4.07	2.89
		M	19	0.26	0.20	19	27.05	22.23
voting	(16)	m	3	0.01	0.02	3	0.01	0.02
		a	4.56	0.04	0.13	3.46	0.3	0.25
		M	11	0.10	0.37	11	1.25	1.77
spect	(22)	m	3	0.02	0.02	3	0.02	0.04
		a	7.31	0.13	0.07	6.44	1.61	0.67
		M	20	0.88	0.29	20	8.97	10.73

Scales to (a few) **tens** of neurons...

Recent results for NNs (w/ Marabou [KHI+19]) [HM23a]

DNN	points	|AXp|	#Calls	Time	#TO	|AXp|	#Calls	Time	#TO
			$\epsilon = 0.1$				$\epsilon = 0.05$		
ACASXU_1_5	#1	3	5	185.9	0	2	5	113.8	0
	#2	2	5	273.8	0	1	5	33.2	0
	#3	0	5	714.2	0	0	5	4.3	0
ACASXU_3_1	#1	0	5	2219.3	0	0	5	14.2	0
	#2	2	5	4263.5	1	0	5	1853.1	0
	#3	1	5	581.8	0	0	5	355.9	0
ACASXU_3_2	#1	3	5	13739.3	2	1	5	6890.1	1
	#2	3	5	226.4	0	2	5	125.1	0
	#3	2	5	1740.6	0	2	5	173.6	0
ACASXU_3_5	#1	4	5	43.6	0	2	5	59.4	0
	#2	3	5	5039.4	0	2	5	4303.8	1
	#3	2	5	5574.9	1	2	5	2660.3	0
ACASXU_3_6	#1	1	5	6225.0	1	0	5	51.0	0
	#2	3	5	4957.2	1	2	5	1897.3	0
	#3	1	5	196.1	0	1	5	919.2	0
ACASXU_3_7	#1	3	5	6256.2	0	4	5	26.9	0
	#2	4	5	311.3	0	1	5	6958.6	1
	#3	2	5	7756.5	1	1	5	7807.6	1
ACASXU_4_1	#1	2	5	12413.0	2	1	5	5090.5	1
	#2	1	5	5035.1	1	0	5	2335.6	0
	#3	4	5	1237.3	0	4	5	1143.4	0
ACASXU_4_2	#1	4	5	15.9	0	4	5	12.1	0
	#2	3	5	1507.6	0	1	5	111.3	0
	#3	2	5	5641.6	2	0	5	1639.1	0

Recent results for NNs (w/ Marabou [KHI+19]) [HM23a]

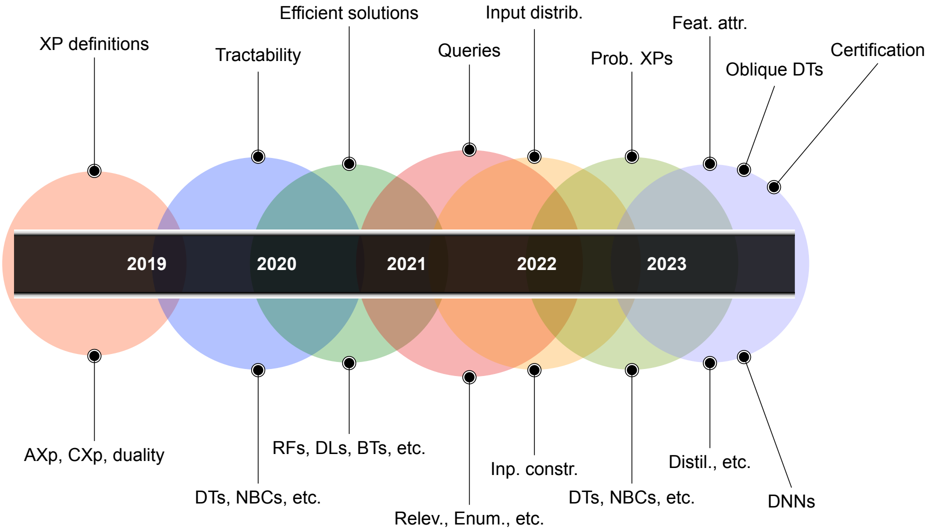
DNN	points	|AXp|	#Calls	Time	#TO	|AXp|	#Calls	Time	#TO
			$\epsilon = 0.1$				$\epsilon = 0.05$		
ACASXU_1_5	#1	3	5	185.9	0	2	5	113.8	0
	#2	2	5	273.8	0	1	5	33.2	0
	#3	0	5	714.2	0	0	5	4.3	0
ACASXU_3_1	#1	0	5	2219.3	0	0	5	14.2	0
	#2	2	5	4263.5	1	0	5	1853.1	0
	#3	1	5	581.8	0	0	5	355.9	0
ACASXU_3_2	#1	3	5	13739.3	2	1	5	6890.1	1
	#2	3	5	226.4	0	2	5	125.1	0
	#3	2	5	1740.6	0	2	5	173.6	0
ACASXU_3_5	#1	4	5	43.6	0	2	5	59.4	0
	#2	3	5	5039.4	0	2	5	4303.8	1
	#3	2	5	5574.9	1	2	5	2660.3	0
ACASXU_3_6	#1	1	5	6225.0	1	0	5	51.0	0
	#2	3	5	4957.2	1	2	5	1897.3	0
	#3	1	5	196.1	0	1	5	919.2	0
ACASXU_3_7	#1	3	5	6256.2	0	4	5	26.9	0
	#2	4	5	311.3	0	1	5	6958.6	1
	#3	2	5	7756.5	1	1	5	7807.6	1
ACASXU_4_1	#1	2	5	12413.0	2	1	5	5090.5	1
	#2	1	5	5035.1	1	0	5	2335.6	0
	#3	4	5	1237.3	0	4	5	1143.4	0
ACASXU_4_2	#1	4	5	15.9	0	4	5	12.1	0
	#2	3	5	1507.6	0	1	5	111.3	0
	#3	2	5	5641.6	2	0	5	1639.1	0

Scales to a few **hundred** neurons

DeepLever: publicly available explainers

1. Naive bayes and linear classifiers: https://github.com/jpmarquessilva/expxlc
2. Monotone classifiers: https://github.com/jpmarquessilva/xmono
3. Decision trees: https://github.com/yizza91/xpg
4. Tractable circuits: https://github.com/XuanxiangHuang/Xddnnf
5. Decision lists: https://github.com/alexeyignatiev/minds
6. Random forests: https://github.com/yizza91/RFxpl
7. Tree ensembles (+ boosted trees): https://github.com/alexeyignatiev/xreason
8. Decision trees (probabilistic Xps): https://github.com/yizza91/praxp
9. ...

The emergence of formal explainability -- timeline



And disproved pervasive hallmarks of non-formal XAI

[RSG16, LL17, RSG18, Rud19]

"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

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

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

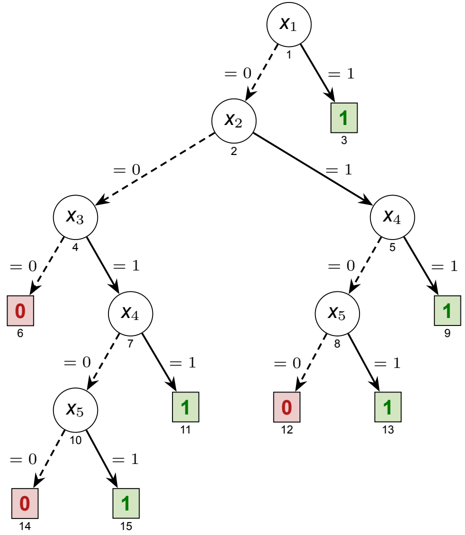
Anchors: High-Precision Model-Agnostic Explanations

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

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

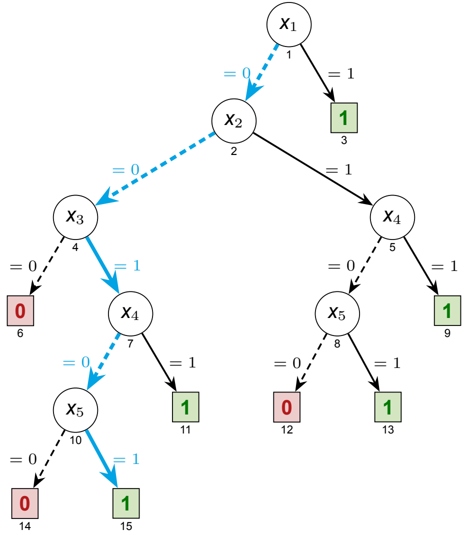
Carlos Guestrin
University of Washington
guestrin@cs.washington.edu

Interpretable models NOT interpretable -- DTs



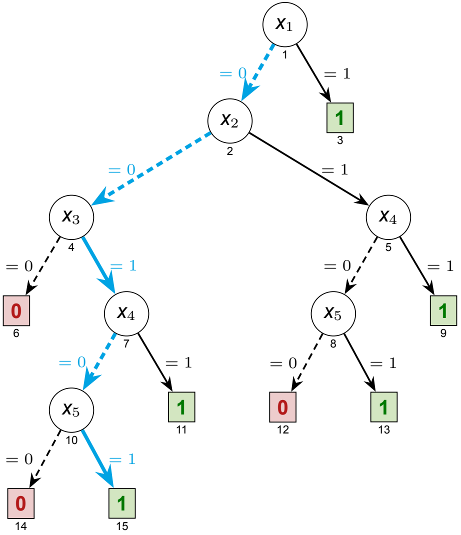
- Case of **optimal** decision tree (DT) [HRS19]
- Explanation for $(0, 0, 1, 0, 1)$, with prediction 1?

Interpretable models NOT interpretable – DTs



- Case of **optimal** decision tree (DT) [HRS19]
- Explanation for $(0, 0, 1, 0, 1)$, with prediction 1?
  - Clearly,
    IF $\neg x_1 \wedge \neg x_2 \wedge x_3 \wedge \neg x_4 \wedge x_5$ THEN $\kappa(\mathbf{x}) = 1$

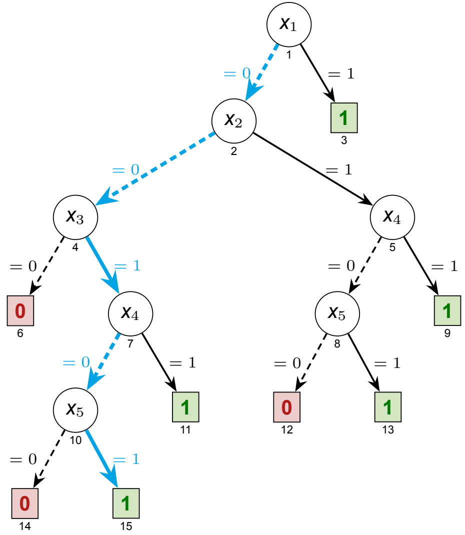
Interpretable models NOT interpretable -- DTs



- Case of **optimal** decision tree (DT) [HRS19]
- Explanation for $(0, 0, 1, 0, 1)$, with prediction 1?
  - Clearly,
    IF $\neg x_1 \wedge \neg x_2 \wedge x_3 \wedge \neg x_4 \wedge x_5$ THEN $\kappa(\mathbf{x}) = 1$
  - But, x_1, x_2, x_4 are irrelevant for the prediction:

x_3	x_5	x_1	x_2	x_4	$\kappa(\mathbf{x})$
1	1	0	0	0	1
1	1	0	0	1	1
1	1	0	1	0	1
1	1	0	1	1	1
1	1	1	0	0	1
1	1	1	0	1	1
1	1	1	1	0	1
1	1	1	1	1	1

Interpretable models NOT interpretable – DTs



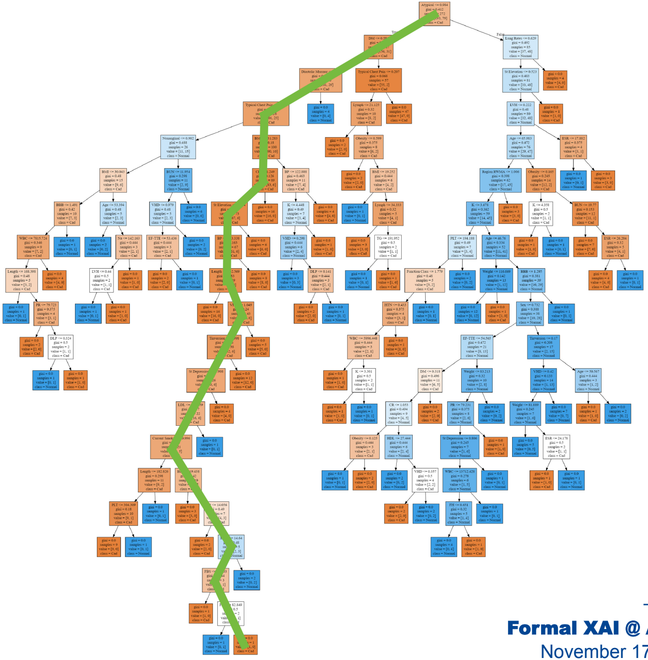
- Case of **optimal** decision tree (DT) [HRS19]
- Explanation for $(0, 0, 1, 0, 1)$, with prediction 1?
  - Clearly,
    IF $\neg x_1 \wedge \neg x_2 \wedge x_3 \wedge \neg x_4 \wedge x_5$ THEN $\kappa(\mathbf{x}) = 1$
  - But, x_1, x_2, x_4 are irrelevant for the prediction:

x_3	x_5	x_1	x_2	x_4	$\kappa(\mathbf{x})$
1	1	0	0	0	1
1	1	0	0	1	1
1	1	0	1	0	1
1	1	0	1	1	1
1	1	1	0	0	1
1	1	1	0	1	1
1	1	1	1	0	1
1	1	1	1	1	1

$\therefore$ one AXp is $\{3, 5\}$
Compare with $\{1, 2, 3, 4, 5\}$...

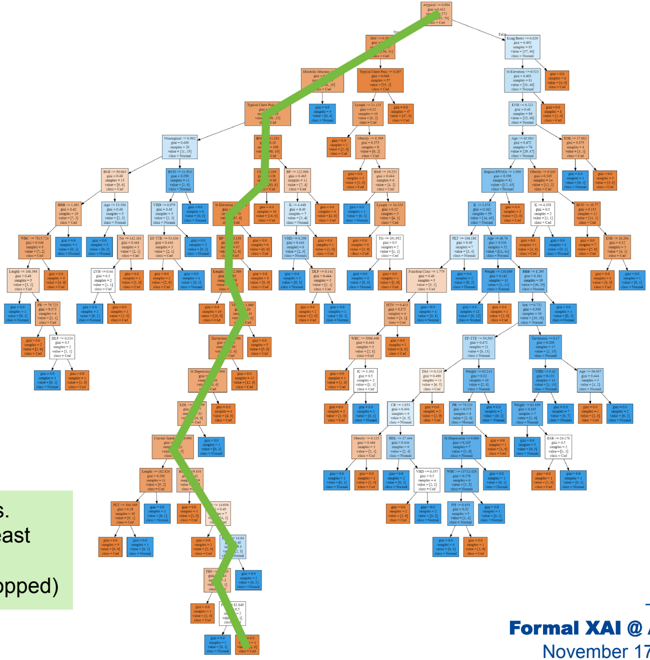
Interpretable models NOT interpretable -- large DTs

[GZM20]



Interpretable models NOT interpretable -- large DTs

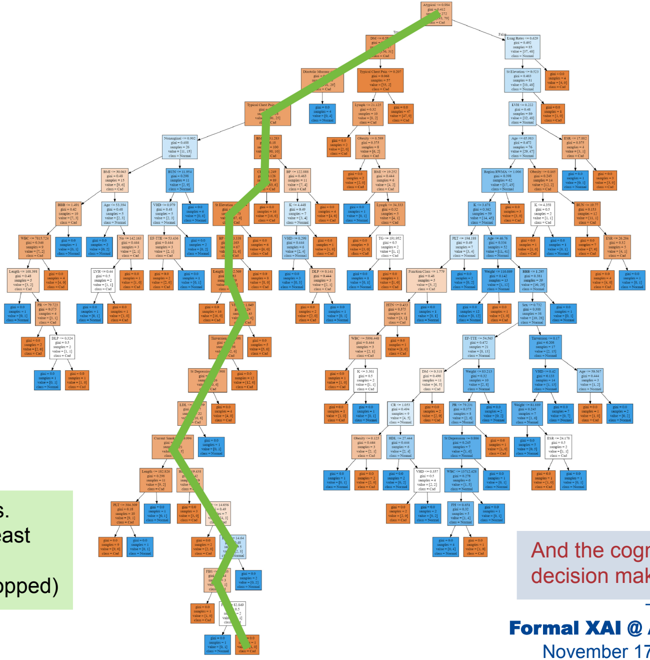
[GZM20]



Path with 19 internal nodes.
By manual inspection, at least
10 literals are redundant!
(And at least 9 features dropped)

Interpretable models NOT interpretable -- large DTs

[GZM20]



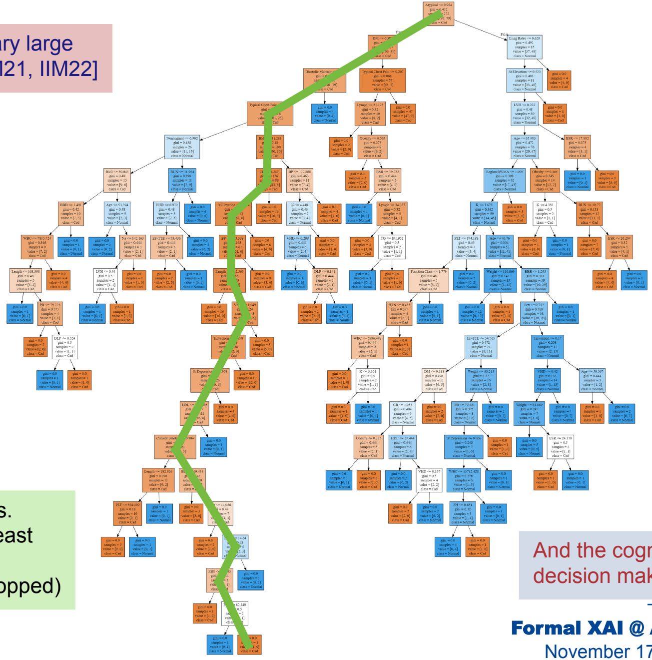
Path with 19 internal nodes.
By manual inspection, at least 10 literals are redundant!
(And at least 9 features dropped)

And the cognitive limits of human decision makers are well-known [Mil56]

Interpretable models NOT interpretable -- large DTs

[GZM20]

Redundancy can be arbitrary large on path length [IIM20, HIIM21, IIM22]



Path with 19 internal nodes.
By manual inspection, at least 10 literals are redundant!
(And at least 9 features dropped)

And the cognitive limits of human decision makers are well-known [Mil56]

Model-agnostic explanations are incorrect often!

- **Errors in model-agnostic explanations known since 2019** [INM19b, Ign20, YIS+23]

Model-agnostic explanations are incorrect often!

- **Errors in model-agnostic explanations known since 2019** [INM19b, Ign20, YIS+23]
- Results for boosted trees, due to non-scalability with NNs [CG16]

Model-agnostic explanations are incorrect often!

- **Errors in model-agnostic explanations known since 2019** [INM19b, Ign20, YIS+23]
- Results for boosted trees, due to non-scalability with NNs [CG16]
- Some results for Anchors [RSG18]

Dataset	% Incorrect	% Redundant	% Correct
adult	80.5%	1.6%	17.9%
lending	3.0%	0.0%	97.0%
rcdv	99.4%	0.4%	0.2%
compas	84.4%	1.7%	13.9%
german	99.7%	0.2%	0.1%

Model-agnostic explanations are incorrect often!

- **Errors in model-agnostic explanations known since 2019** [INM19b, Ign20, YIS+23]
- Results for boosted trees, due to non-scalability with NNs [CG16]
- Some results for Anchors [RSG18]

Dataset	% Incorrect	% Redundant	% Correct
adult	80.5%	1.6%	17.9%
lending	3.0%	0.0%	97.0%
rcdv	99.4%	0.4%	0.2%
compas	84.4%	1.7%	13.9%
german	99.7%	0.2%	0.1%

Obs: most often, Anchor's rules are **NOT** rules...

Model-agnostic explanations are incorrect often!

- **Errors in model-agnostic explanations known since 2019** [INM19b, Ign20, YIS+23]
- Results for boosted trees, due to non-scalability with NNs [CG16]
- Some results for Anchors [RSG18]

Obs: most often, Anchor's rules are **NOT** rules...

Dataset	% Incorrect	% Redundant	% Correct
adult	80.5%	1.6%	17.9%
lending	3.0%	0.0%	97.0%
rcdv	99.4%	0.4%	0.2%
compas	84.4%	1.7%	13.9%
german	99.7%	0.2%	0.1%

- **Obs:** Results are not positive even if we count how often prediction changes [NSM+19]
  - In this case, BNNs were used, to allow for model counting...

Model-agnostic explanations are incorrect often!

- **Errors in model-agnostic explanations known since 2019** [INM19b, Ign20, YIS+23]
- Results for boosted trees, due to non-scalability with NNs [CG16]
- Some results for Anchors [RSG18]

Dataset	% Incorrect	% Redundant	% Correct
adult	80.5%	1.6%	17.9%
lending	3.0%	0.0%	97.0%
rcdv	99.4%	0.4%	0.2%
compas	84.4%	1.7%	13.9%
german	99.7%	0.2%	0.1%

Obs: most often, Anchor's rules are **NOT** rules...

- **Obs:** Results are not positive even if we count how often prediction changes [NSM+19]
  - In this case, BNNs were used, to allow for model counting...
- Feature attribution also assessed, with similar results [INM19b, NSM+19, Ign20, YIS+23]

How wrong can model-agnostic explanations be?

- Another possible scenario:

How wrong can model-agnostic explanations be?

- Another possible scenario:

Incorrect explanations (XPs):

Classifier for deciding bank loans

Two samples:

Bessie $:= (v_1, \mathbf{Y})$, Clive $:= (v_2, \mathbf{N})$

Explanation X: age $= 45$, salary $=$ 50K

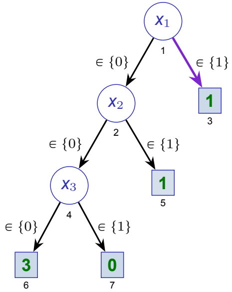
X is consistent with Bessie $:= (\mathbf{v}_1, \mathbf{Y})$

X is consistent with Clive $:= (\mathbf{v}_2, \mathbf{N})$

$\therefore$ different outcomes **&** same explanation **!?**

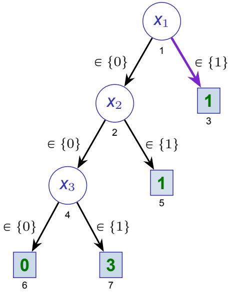
Exact SHAP scores can mislead...

[HM23b, HM23c, HM23d, MH23]



DT1

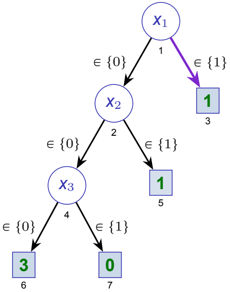
row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	3	0
2	0	0	1	0	3
3	0	1	0	1	1
4	0	1	1	1	1
5	1	0	0	1	1
6	1	0	1	1	1
7	1	1	0	1	1
8	1	1	1	1	1



DT2

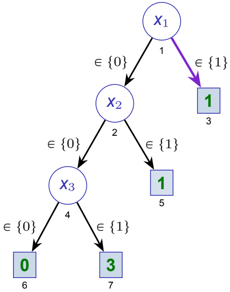
Instance $((1,1,1),1)$. Which features matter?

[HM23b, HM23c, HM23d, MH23]



DT1

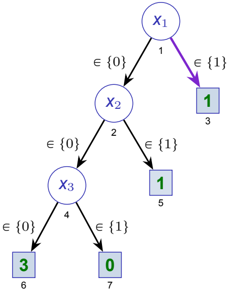
row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	3	0
2	0	0	1	0	3
3	0	1	0	1	1
4	0	1	1	1	1
5	1	0	0	1	1
6	1	0	1	1	1
7	1	1	0	1	1
8	1	1	1	1	1



DT2

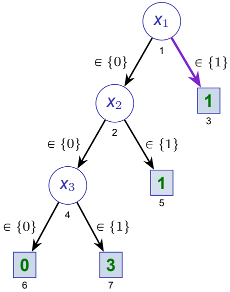
Instance $((1,1,1),1)$. Which features matter? Say 1 & 2?

[HM23b, HM23c, HM23d, MH23]



DT1

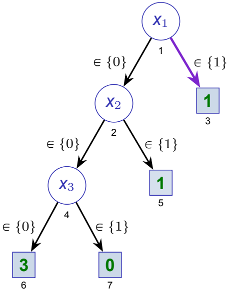
row #	x_1	x_2	x_3	$\kappa_1(\mathbf{x})$	$\kappa_2(\mathbf{x})$
1	0	0	0	3	0
2	0	0	1	0	3
3	0	1	0	1	1
4	0	1	1	1	1
5	1	0	0	1	1
6	1	0	1	1	1
7	1	1	0	1	1
8	1	1	1	1	1



DT2

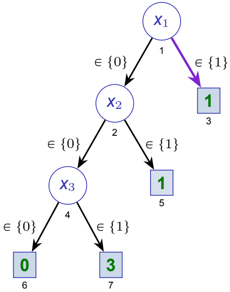
AXps/CXps OK

[HM23b, HM23c, HM23d, MH23]



DT1

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

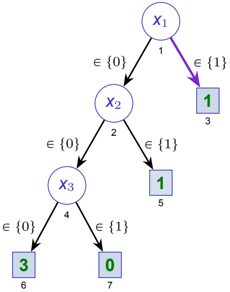


DT2

XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\}, \{2\}$	$\{1, 2\}$
DT2	$\{1\}, \{2\}$	$\{1, 2\}$

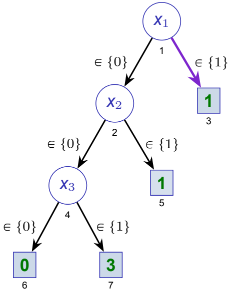
AXps/CXps OK, AExs OK

[HM23b, HM23c, HM23d, MH23]



DT1

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



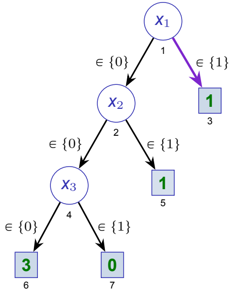
DT2

XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\},\{2\}$	$\{1,2\}$
DT2	$\{1\},\{2\}$	$\{1,2\}$

Adversarial Examples	
DT	l_0-minimal AEs
DT1	$\{1,2\}$
DT2	$\{1,2\}$

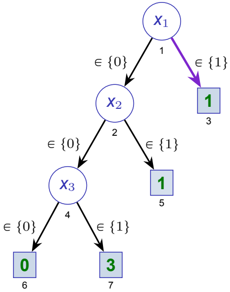
AXps/CXps OK, AExs OK, Svs ...

[HM23b, HM23c, HM23d, MH23]



DT1

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



DT2

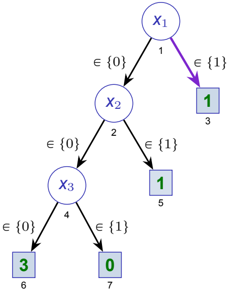
XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\},\{2\}$	$\{1,2\}$
DT2	$\{1\},\{2\}$	$\{1,2\}$

Adversarial Examples	
DT	l_0-minimal AEs
DT1	$\{1,2\}$
DT2	$\{1,2\}$

SHAP Scores			
DT	Sv(1)	Sv(2)	Sv(3)
DT1	0.000	0.000	-0.125
DT2	-0.125	-0.125	0.125

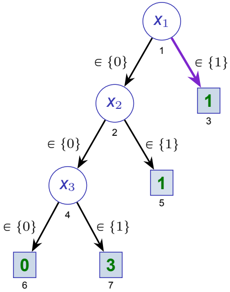
AXps/CXps OK, AExs OK, Svs not OK!!!

[HM23b, HM23c, HM23d, MH23]



DT1

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



DT2

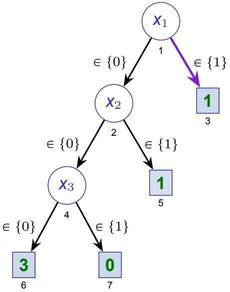
XPs: AXps/CXps		
DT	AXps	CXps
DT1	$\{1\},\{2\}$	$\{1,2\}$
DT2	$\{1\},\{2\}$	$\{1,2\}$

Adversarial Examples	
DT	l_0-minimal AEs
DT1	$\{1,2\}$
DT2	$\{1,2\}$

SHAP Scores			
DT	Sv(1)	Sv(2)	Sv(3)
DT1	0.000	0.000	-0.125
DT2	-0.125	-0.125	0.125

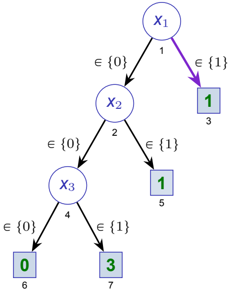
AXps/CXps OK, AExs OK, Svs not OK!!!

[HM23b, HM23c, HM23d, MH23]



DT1

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



DT2

XPs: AXps/CXps

DT	AXps	CXps
DT1	$\{1\}, \{2\}$	$\{1, 2\}$
DT2	$\{1\}, \{2\}$	$\{1, 2\}$

Adversarial Examples

DT	l_0-minimal AEs
DT1	$\{1, 2\}$
DT2	$\{1, 2\}$

SHAP Scores

DT	Sv(1)	Sv(2)	Sv(3)
DT1	0.000	0.000	-0.125
DT2	-0.125	-0.125	0.125

SHAP [LL17] most often does **NOT** agree with SHAP scores... & SHAP scores are **misleading**...

A take-home message...

[RSG16, LL17, RSG18, Rud19]

“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

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

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

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

A take-home message...

[RSG16, LL17, RSG18, Rud19]

"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

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

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

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

For high-risk / safety-critical uses of AI/ML do **NOT** use non-formal XAI !

A take-home message...

[RSG16, LL17, RSG18, Rud19]

"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

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

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

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

For high-risk / safety-critical uses of AI/ML do **NOT** use non-formal XAI !

I.e. unsuitable for trustworthy AI !

Ongoing & future research

Ongoing & future research

- Distance-restricted AXps/CXps [HM23a]
  - Links with adversarial robustness

Ongoing & future research

- Distance-restricted AXps/CXps [HM23a]
  - Links with adversarial robustness
- Certification of formal explainability [HM23e]
  - Initial results for monotonic classifiers

Ongoing & future research

- Distance-restricted AXps/CXps [HM23a]
  - Links with adversarial robustness
- Certification of formal explainability [HM23e]
  - Initial results for monotonic classifiers
- More expressive explanations [IISM23]
  - Use rel. op. $\in$ instead of $=$

Ongoing & future research

- Distance-restricted AXps/CXps [HM23a]
  - Links with adversarial robustness
- Certification of formal explainability [HM23e]
  - Initial results for monotonic classifiers
- More expressive explanations [IISM23]
  - Use rel. op. $\in$ instead of $=$
- Understand the limitations of (exact) SHAP scores [HM23b, HM23c, HM23d, MH23]

Ongoing & future research

- Distance-restricted AXps/CXps [HM23a]
  - Links with adversarial robustness
- Certification of formal explainability [HM23e]
  - Initial results for monotonic classifiers
- More expressive explanations [IISM23]
  - Use rel. op. $\in$ instead of $=$
- Understand the limitations of (exact) SHAP scores [HM23b, HM23c, HM23d, MH23]
- Inference of input constraints [YIS+23]
  - Not all points in feature space may be meaningful

Ongoing & future research

- Distance-restricted AXps/CXps [HM23a]
  - Links with adversarial robustness
- Certification of formal explainability [HM23e]
  - Initial results for monotonic classifiers
- More expressive explanations [IISM23]
  - Use rel. op. $\in$ instead of $=$
- Understand the limitations of (exact) SHAP scores [HM23b, HM23c, HM23d, MH23]
- Inference of input constraints [YIS$^+$23]
  - Not all points in feature space may be meaningful
- Tractability results [CM23, CCM23]
  - E.g. oblique DTs

Ongoing & future research

- Distance-restricted AXps/CXps [HM23a]
  - Links with adversarial robustness
- Certification of formal explainability [HM23e]
  - Initial results for monotonic classifiers
- More expressive explanations [IISM23]
  - Use rel. op. $\in$ instead of $=$
- Understand the limitations of (exact) SHAP scores [HM23b, HM23c, HM23d, MH23]
- Inference of input constraints [YIS+23]
  - Not all points in feature space may be meaningful
- Tractability results [CM23, CCM23]
  - E.g. oblique DTs
- Reduced explanation size [IHI+23]
  - Given cognitive limits of human decision-makers [Mil56]

Q & A

Joint work with X. Huang, O. Létoffé, M. Cooper, N. Asher, Y. Izza, A. Ignatiev, N. Narodytska, J. Planes, A. Morgado, R. Bejar, et al.

References

Clément Carbonnel, Martin C. Cooper, and João Marques-Silva.
Tractable explaining of multivariate decision trees.
In *KR*, pages 127–135, 2023.

Tianqi Chen and Carlos Guestrin.
XGBoost: A scalable tree boosting system.
In *KDD*, pages 785–794, 2016.

Martin C. Cooper and Joao Marques-Silva.
On the tractability of explaining decisions of classifiers.
In *CP*, pages 21:1–21:18, 2021.

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

Mohammad M. Ghiasi, Sohrab Zendehboudi, and Ali Asghar Mohsenipour.
Decision tree-based diagnosis of coronary artery disease: CART model.
Comput. Methods Programs Biomed., 192:105400, 2020.

References II

- Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, Martin Cooper, Nicholas Asher, and Joao Marques-Silva.
  Tractable explanations for d-DNNF classifiers.
  In *AAAI*, February 2022.
- Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva.
  On efficiently explaining graph-based classifiers.
  In *KR*, pages 356–367, 2021.
- Xuanxiang Huang and João Marques-Silva.
  From robustness to explainability and back again.
  CoRR, abs/2306.03048, 2023.
- Xuanxiang Huang and João Marques-Silva.
  The inadequacy of Shapley values for explainability.
  CoRR, abs/2302.08160, 2023.
- Xuanxiang Huang and Joao Marques-Silva.
  A refutation of shapley values for explainability.
  CoRR, abs/2309.03041, 2023.

References III

- Xuanxiang Huang and Joao Marques-Silva.
  Refutation of shapley values for XAI – additional evidence.
  CoRR, abs/2310.00416, 2023.
- Aurélie Hurault and João Marques-Silva.
  Certified logic-based explainable AI - the case of monotonic classifiers.
  In *TAP*, pages 51–67, 2023.
- Xiyang Hu, Cynthia Rudin, and Margo I. Seltzer.
  Optimal sparse decision trees.
  In *NeurIPS*, pages 7265–7273, 2019.
- Alexey Ignatiev.
  Towards trustable explainable AI.
  In *IJCAI*, pages 5154–5158, 2020.
- Yacine Izza, Xuanxiang Huang, Alexey Ignatiev, Nina Narodytska, Martin C. Cooper, and João Marques-Silva.
  On computing probabilistic abductive explanations.
  Int. J. Approx. Reason., 159:108939, 2023.

References IV

- Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva.
  On explaining decision trees.
  CoRR, abs/2010.11034, 2020.
- Yacine Izza, Alexey Ignatiev, and João Marques-Silva.
  On tackling explanation redundancy in decision trees.
  J. Artif. Intell. Res., 75:261–321, 2022.
- Yacine Izza, Alexey Ignatiev, Peter J. Stuckey, and João Marques-Silva.
  Delivering inflated explanations.
  CoRR, abs/2306.15272, 2023.
- Alexey Ignatiev, Yacine Izza, Peter J. Stuckey, and Joao Marques-Silva.
  Using MaxSAT for efficient explanations of tree ensembles.
  In *AAAI*, February 2022.
- Yacine Izza and Joao Marques-Silva.
  On explaining random forests with SAT.
  In *IJCAI*, pages 2584–2591, 2021.

References V

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

Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and Joao Marques-Silva.
From contrastive to abductive explanations and back again.
In *AI*IA*, pages 335–355, 2020.

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

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

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

References VI

- Guy Katz, Derek A. Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljic, David L. Dill, Mykel J. Kochenderfer, and Clark W. Barrett.
  The marabou framework for verification and analysis of deep neural networks.
  In *CAV*, pages 443–452, 2019.
- Scott M. Lundberg and Su-In Lee.
  A unified approach to interpreting model predictions.
  In *NIPS*, pages 4765–4774, 2017.
- 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.
- Joao Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska.
  Explanations for monotonic classifiers.
  In *ICML*, pages 7469–7479, 2021.

References VII

- Joao Marques-Silva and Xuanxiang Huang.
  Explainability is NOT a game.
  CoRR, abs/2307.07514, 2023.
- George A Miller.
  The magical number seven, plus or minus two: Some limits on our capacity for processing information.
  Psychological review, 63(2):81–97, 1956.
- Nina Narodytska, Aditya A. Shrotri, Kuldeep S. Meel, Alexey Ignatiev, and Joao Marques-Silva.
  Assessing heuristic machine learning explanations with model counting.
  In *SAT*, pages 267–278, 2019.
- David Poole and Alan K. Mackworth.
  Artificial Intelligence - Foundations of Computational Agents.
  CUP, 2017.

EU Proposal.
European Artificial Intelligence Act – Proposal.
https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206, 2021.

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.

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

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.

References IX

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

- Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey, Nina Narodytska, and Joao Marques-Silva.
  Eliminating the impossible, whatever remains must be true: On extracting and applying background knowledge in the context of formal explanations.
  In *AAAI*, 2023.