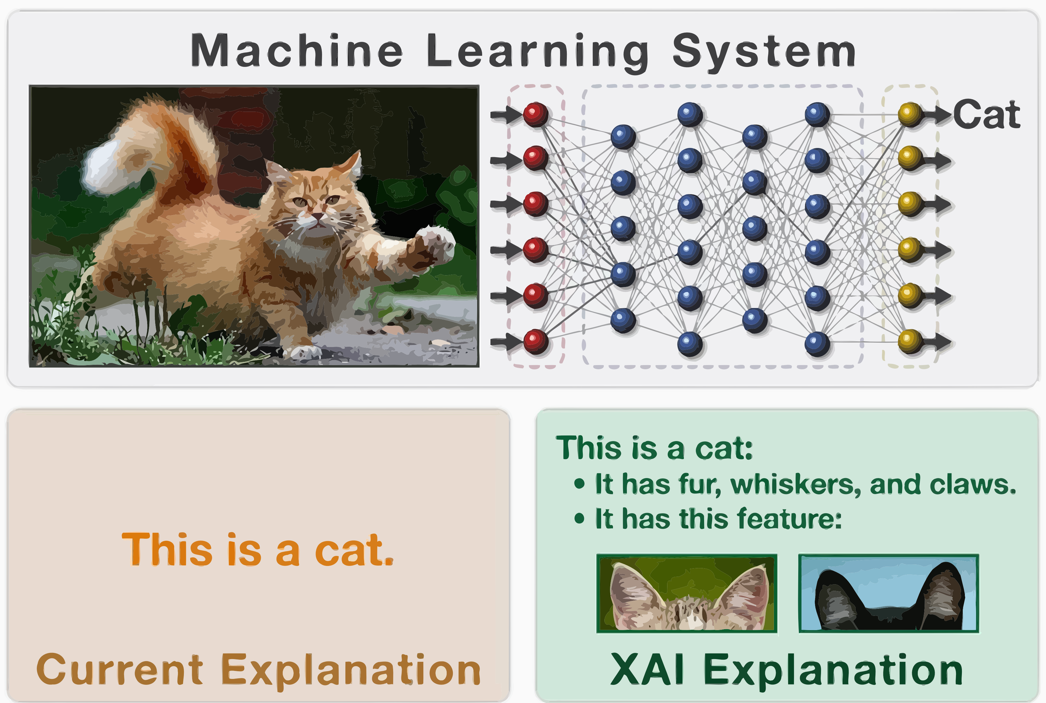
SAT-Based Rigorous Explanations for Decision Lists

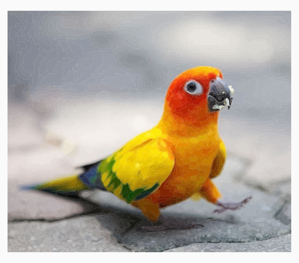
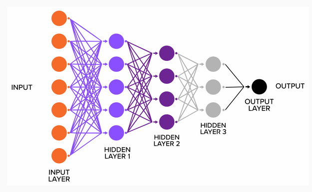
Alexey Ignatiev[1] and Joao Marques-Silva[2]

[1] Monash University, Australia [2] ANITI, IRIT, CNRS, France

eXplainable AI



Why? Status Quo

	A parrot 	Machine learning algorithm 
Learns random phrases	✓	✓
Doesn't understand s**t about what it learns	✓	✓
Occasionally speaks nonsense	✓	✓

Interpretable Models

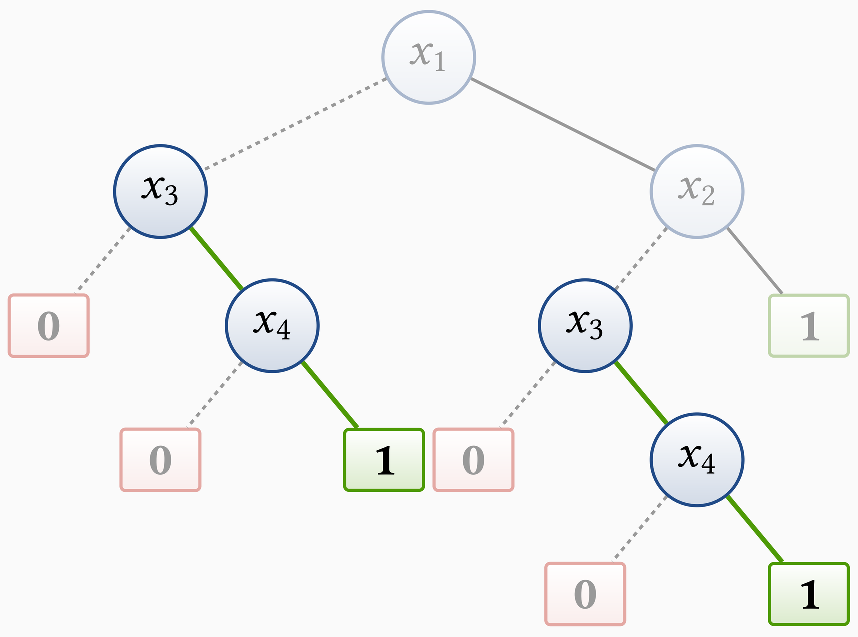
rule-based models

"*transparent*" and easy to interpret

come in handy in XAI

but...

DT Interpretability Issue



instance v = (1, 0, 1, 1) — 4 literals in the path

actual explanation $x_3 = 1 \wedge x_4 = 1$ — 2 literals

Same Issue with DL Interpretability

R_0:	IF	$x_1 = 0 \wedge x_3 = 0$	THEN $f = 0$
R_1:	ELSE IF	$x_1 = 0 \wedge x_3 = 1 \wedge x_4 = 0$	THEN $f = 0$
R_2:	ELSE IF	$x_1 = 0 \wedge x_3 = 1 \wedge x_4 = 1$	THEN $f = 1$
R_3:	ELSE IF	$x_1 = 1 \wedge x_2 = 0 \wedge x_3 = 0$	THEN $f = 0$
R_4:	ELSE IF	$x_1 = 1 \wedge x_2 = 0 \wedge x_3 = 1 \wedge x_4 = 0$	THEN $f = 0$
R_5:	ELSE IF	$x_1 = 1 \wedge x_2 = 0 \wedge x_3 = 1 \wedge x_4 = 1$	THEN $f = 1$
R_6:	ELSE IF	$x_1 = 1 \wedge x_2 = 1$	THEN $f = 1$
R_{DEF}:	ELSE		THEN $f = 1$

instance v = (1, 0, 1, 1) — rule R_5 fires the prediction

actual AXp — $x_3 = 1 \wedge x_4 = 1$ — 2 literals

Rigorous Explanations

classifier $\tau : \mathbb{F} \to \mathcal{K}$, **instance v s.t.** $\tau(\mathbf{v}) = c$

abductive explanation $\mathcal{X}$

$$\forall (\mathbf{x} \in \mathbb{F}) . \bigwedge_{j \in \mathcal{X}} (x_j = v_j) \rightarrow (\tau(\mathbf{x}) = c)$$

contrastive explanation $\mathcal{Y}$

$$\exists (\mathbf{x} \in \mathbb{F}) . \bigwedge_{j \notin \mathcal{Y}} (x_j = v_j) \wedge (\tau(\mathbf{x}) \neq c)$$

Explanation Duality

$\mathbb{F} = \{0, 1, 2\}^5$ $\quad \mathcal{K} = \{\ominus, \oplus\}$

R_0:	IF	$x_1 = 1 \wedge x_2 = 1$	THEN $\ominus$
R_1:	ELSE IF	$x_3 \neq 1$	THEN $\oplus$
R_{DEF}:	ELSE		THEN $\ominus$

observe $\tau(1, 1, 1, 1, 1) = \ominus$

AXps $\mathbb{X} = \{\{1, 2\}, \{3\}\}$

CXps $\mathbb{Y} = \{\{1, 3\}, \{2, 3\}\}$

minimal hitting set duality!

Problems

SAT query:

$\exists(\mathbf{x} \in \mathbb{F}). \quad \tau(\mathbf{x}) = c$

DLSAT is *NP*-complete

IM query:

$\forall(\mathbf{x} \in \mathbb{F}). \quad \rho(\mathbf{x}) \rightarrow \tau(\mathbf{x}) = c$

No polytime algorithm for DLIM, unless $P = NP$

Explanation Complexity

decision lists:

finding an AXp is not polytime unless $P = NP$

decision sets:

finding an AXp is D^P-complete

in contrast to decision trees!

Propositional Encoding

rule $j \in \mathfrak{R}$ fires:

$$\varphi(j) \triangleq \left(\bigwedge_{k \in \mathfrak{R},\, \mathfrak{o}(k) < \mathfrak{o}(j)} \neg \mathfrak{l}(k) \right) \wedge \mathfrak{l}(j)$$

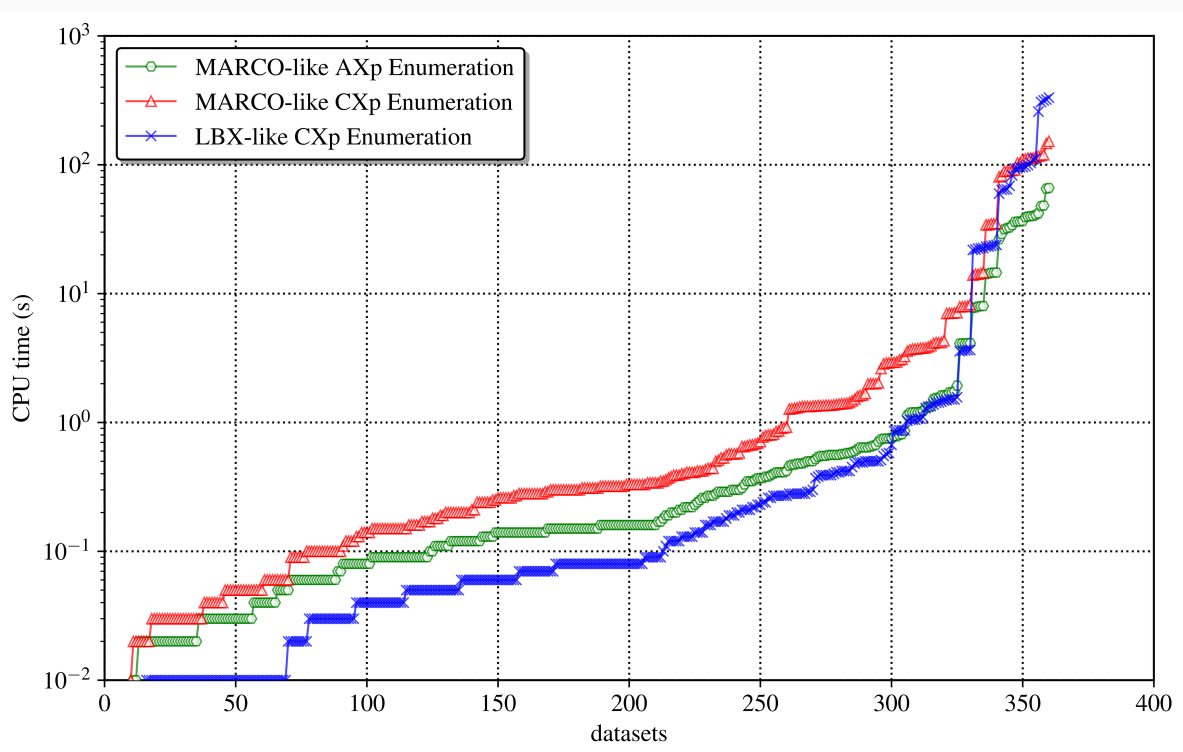
unsatisfiable $\mathcal{S} \wedge \mathcal{H}$ **s.t.**

$$\mathcal{S} \triangleq I_{\mathbf{v}} \qquad \mathcal{H} \triangleq \bigvee_{j \in \mathfrak{R},\, \mathfrak{c}(j) = \mathfrak{c}(i)} \varphi(j)$$

instance v, prediction $\mathfrak{c}(i)$:

AXps are MUSes **CXps are MCSes**

Raw Performance

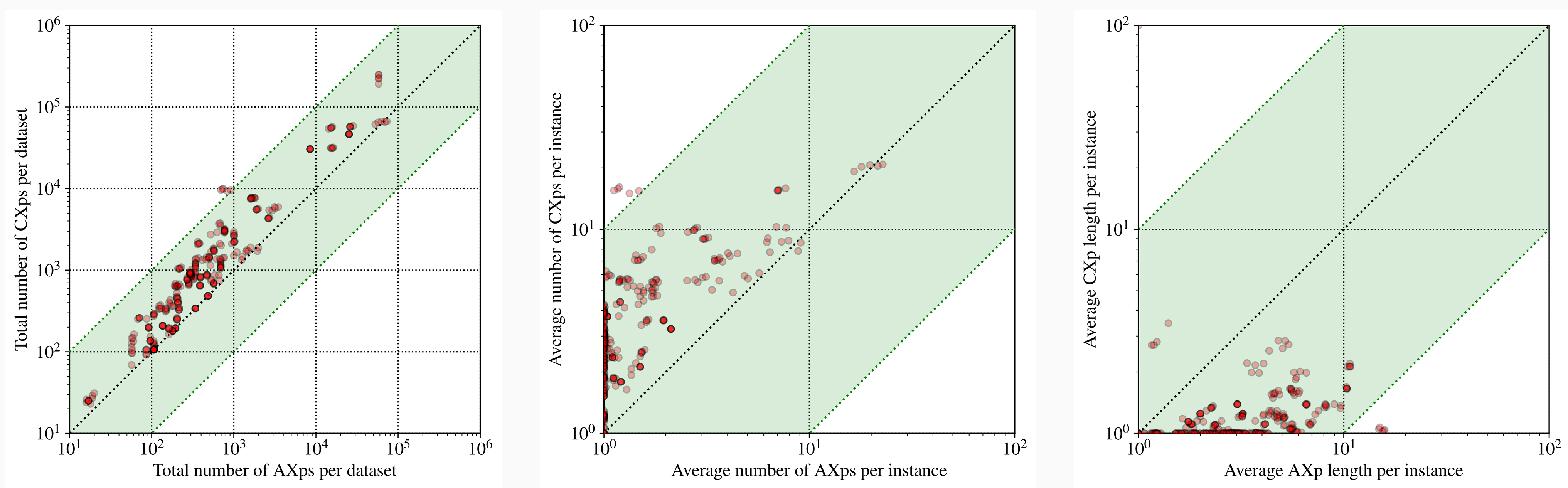


all tools finish complete XP enumeration within <1000 sec.

MARCO-like setup — targeting AXps may pay off

direct CXp enumeration is slower *(too many XPs?)*

AXps vs. CXps



16–72838 AXps	**vs.**	**23–248825 CXps**	*per dataset*
1–22.7 AXps	**vs.**	**1–20.8 CXps**	*per instance*
1–15.8 lits per AXp	**vs.**	**≤2.8 lits per CXp**	

LaTeX TikZposter