Reasoning-Based Learning of Interpretable ML Models

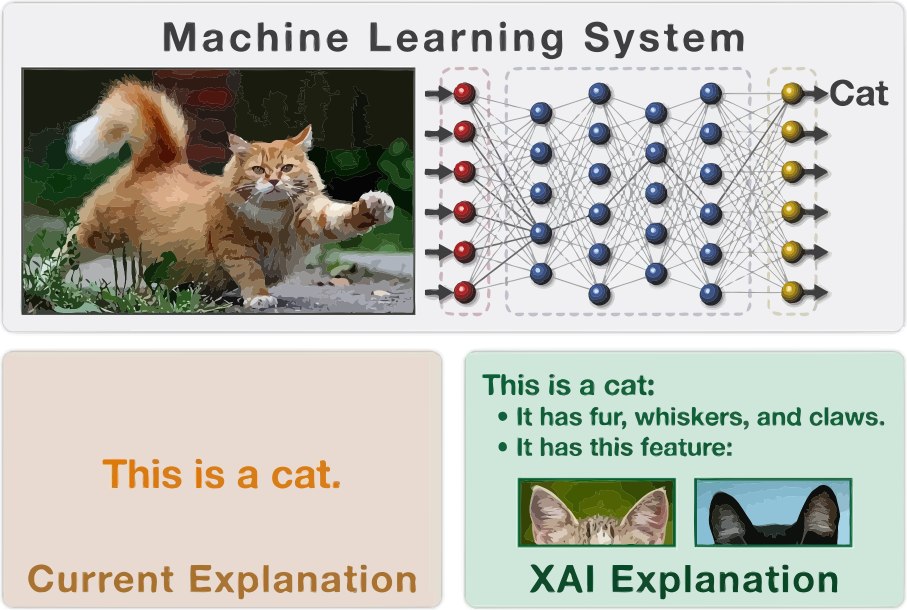
Alexey Ignatiev[1], Joao Marques-Silva[2], Nina Narodytska[3], and Peter J. Stuckey[1]

August 19-26, 2021 | **IJCAI**

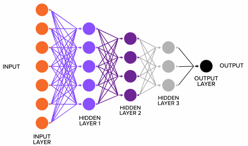
[1] Monash University, Melbourne, Australia
[2] IRIT, CNRS, Toulouse, France
[3] VMware Research, CA, USA

eXplainable AI

eXplainable AI



Why? Status quo...

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

interpretable ML models

e.g. decision trees, lists, sets

interpretable ML models

e.g. decision trees, lists, sets

posthoc explanation of ML models "on the fly"

rule-based models

rule-based models

⬇

“*transparent*” **and easy to interpret**

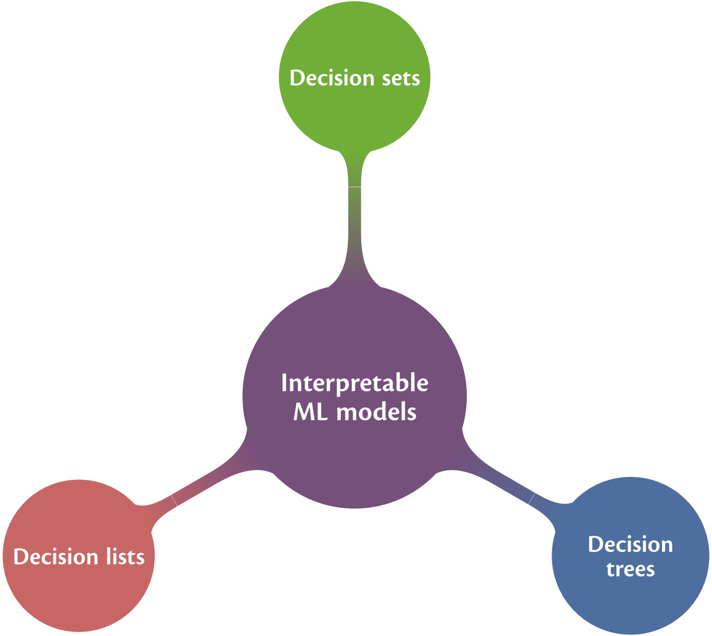
rule-based models

↓

“*transparent*” and easy to interpret

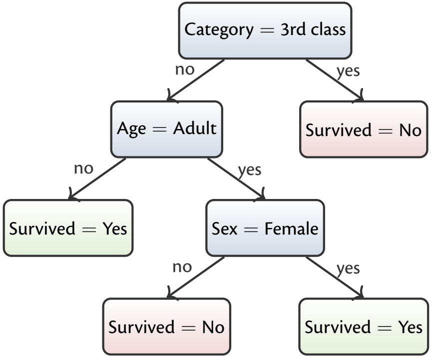
↓

come in handy in XAI


Decision sets
Interpretable ML models
Decision lists
Decision trees

Decision trees

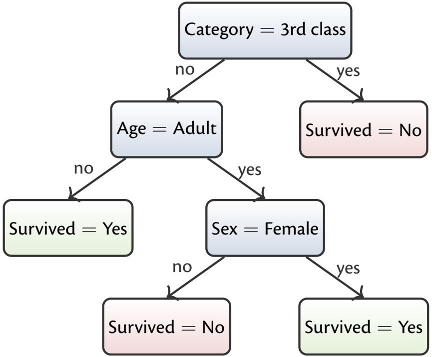
Decision trees: *perfect* and *sparse*



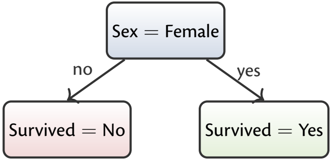
perfect DT for *Titanic* dataset

(training accuracy 78.25%)

Decision trees: *perfect* and *sparse*



perfect DT for *Titanic* dataset
(training accuracy 78.25%)



sparse DT for *Titanic* dataset
(training accuracy 33.05%)

Reasoning-based approaches to decision trees

	model		unbounded	engine					
	perfect	sparse	depth	MIP	CP	SAT	MaxSAT	DP	B-n-B
Nijssen et al., 2007		✔						✔	
Bessiere et al., 2009	✔				✔	✔			
Bertsimas et al., 2017		✔		✔					
Verwer et al., 2017		✔		✔					
Narodytska et al., 2018	✔		✔			✔			
Verwer et al., 2019		✔		✔					
Hu et al., 2019		✔	✔					✔	✔
Zhu et al., 2020		✔		✔+					
Janota et al., 2020	✔		✔			✔			
Avellaneda et al., 2020	✔		✔			✔+			
Hu et al., 2020	✔		✔				✔+		
Verhaeghe et al., 2020		✔			✔			✔	
Aglin et al., 2020		✔						✔	✔
Demirovic et al., 2020		✔						✔+	

Decision lists

Decision lists: *perfect* and *sparse*

IF	$\text{Age} = \text{Adult} \wedge \text{Sex} \neq \text{Female}$	**THEN**	$\text{Survived} = \text{No}$
ELSE IF	$\text{Category} \neq \text{3rd class}$	**THEN**	$\text{Survived} = \text{Yes}$
		ELSE	$\text{Survived} = \text{No}$

smallest size perfect DL for *Titanic* dataset

(training accuracy 78.25%)

Decision lists: *perfect* and *sparse*

IF	Age $=$ Adult $\wedge$ Sex $\neq$ Female	**THEN**	Survived $=$ No
ELSE IF	Category $\neq$ 3rd class	**THEN**	Survived $=$ Yes
		ELSE	Survived $=$ No

smallest size perfect DL for *Titanic* dataset

(training accuracy 78.25%)

IF	Category $=$ 1st class	**THEN**	Survived $=$ Yes
		ELSE	Survived $=$ No

sparse DL for *Titanic* dataset

(training accuracy 70.69%)

Reasoning-based approaches to decision lists

	model		criterion		optimality	classification		engine				symmetry
	perfect	sparse	rules	literals	guarantee	binary	arbitrary	MIP	SAT	MaxSAT	B-n-B	breaking
Angelino et al., 2017a		✔	✔			✔		✔				
Angelino et al., 2017b		✔	✔		✔	✔					✔	✔
Yu et al., 2020	✔	✔	✔	✔	✔		✔		✔	✔		✔

Decision sets

Decision sets: *perfect* and *sparse*

IF Category $=$ 3rd class	**THEN** Survived $=$ No
IF Age $=$ Adult $\wedge$ Sex $\neq$ Female	**THEN** Survived $=$ No
IF Category $\neq$ 3rd class $\wedge$ Age $\neq$ Adult	**THEN** Survived $=$ Yes
IF Category $\neq$ 3rd class $\wedge$ Sex $=$ Female	**THEN** Survived $=$ Yes

smallest size perfect DS for *Titanic* dataset

(training accuracy 78.25%)

Decision sets: *perfect* and *sparse*

IF	Category $=$ 3rd class	**THEN** Survived $=$ No
IF	Age $=$ Adult $\wedge$ Sex $\neq$ Female	**THEN** Survived $=$ No
IF	Category $\neq$ 3rd class $\wedge$ Age $\neq$ Adult	**THEN** Survived $=$ Yes
IF	Category $\neq$ 3rd class $\wedge$ Sex $=$ Female	**THEN** Survived $=$ Yes

smallest size perfect DS for *Titanic* dataset

(training accuracy 78.25%)

IF	Category $=$ 3rd class	**THEN** Survived $=$ No
IF	Sex $\neq$ Female	**THEN** Survived $=$ No
IF	Category $\neq$ 3rd class $\wedge$ Sex $=$ Female	**THEN** Survived $=$ Yes

sparse DS for *Titanic* dataset

(training accuracy 77.57%)

Reasoning-based approaches to decision sets

	model		criterion			explicit repr.		setup		engine			
	perfect	sparse	rules	lex	literals	single class	all classes	single run	two phases	IP	SAT	MaxSAT	LS
Kamath et al., 1992	✔		✔			✔		✔		✔			
Lakkaraju et al., 2016		✔	✔				✔	✔					✔
Ignatiev et al., 2018	✔		✔	✔			✔	✔			✔	✔	
Malioutov et al., 2018		✔			✔-	✔		✔				✔	
Dash et al., 2018		✔	✔			✔		✔		✔			
Ghosh et al., 2019		✔			✔-	✔		✔				✔	
Ghosh et al., 2020		✔+			✔-	✔		✔				✔	
Yu et al., 2020	✔	✔			✔		✔	✔			✔	✔	
Ignatiev et al., 2021	✔		✔		✔		✔		✔	✔	✔	✔	

Additional considerations

Additional considerations 1

- **Comparing to heuristic methods**
  - **higher accuracy** *but*

Additional considerations 1

- **Comparing to heuristic methods**
  - **higher accuracy** *but*
  - **higher training time**
    - evolution of reasoning methods!

Additional considerations 1

- **Comparing to heuristic methods**
  - **higher accuracy** *but*
  - **higher training time**
    - evolution of reasoning methods!

- **Other interpretable models**
  - **learning OBDDs**
    - SAT-based inference

Additional considerations 1

- **Comparing to heuristic methods**
  - **higher accuracy** *but*
  - **higher training time**
    - evolution of reasoning methods!
- **Other interpretable models**
  - **learning OBDDs**
    - SAT-based inference
- **Perfect vs. sparse models**
  - **pros of perfect models:**
    - ***highest possible accuracy***

Additional considerations 1

- **Comparing to heuristic methods**
  - **higher accuracy** *but*
  - **higher training time**
    - evolution of reasoning methods!

- **Other interpretable models**
  - **learning OBDDs**
    - SAT-based inference

- **Perfect vs. sparse models**
  - **pros of perfect models:**
    - ***highest possible accuracy***
  - **pros of sparse models:**
    - ***smaller size***
    - easier to compute
    - smaller explanations

Additional considerations 1

- **Comparing to heuristic methods**
  - **higher accuracy** *but*
  - **higher training time**
    - evolution of reasoning methods!
- **Other interpretable models**
  - **learning OBDDs**
    - SAT-based inference
- **Perfect vs. sparse models**
  - **pros of perfect models:**
    - ***highest possible accuracy***
  - **pros of sparse models:**
    - ***smaller size***
    - easier to compute
    - smaller explanations
- **Model expressivity and size**
  - **DLs are more succinct than DTs**

Additional considerations 1

- **Comparing to heuristic methods**
  - **higher accuracy** *but*
  - **higher training time**
    - evolution of reasoning methods!
- **Other interpretable models**
  - **learning OBDDs**
    - SAT-based inference
- **Perfect vs. sparse models**
  - **pros of perfect models:**
    - ***highest possible accuracy***
  - **pros of sparse models:**
    - ***smaller size***
    - easier to compute
    - smaller explanations
- **Model expressivity and size**
  - **DLs are more succinct than DTs**
  - **DLs are more succinct than DNFs**
    - *a special case of DSs*

Additional considerations 1

- **Comparing to heuristic methods**
  - **higher accuracy** *but*
  - **higher training time**
    - evolution of reasoning methods!
- **Other interpretable models**
  - **learning OBDDs**
    - SAT-based inference
- **Perfect vs. sparse models**
  - **pros of perfect models:**
    - ***highest possible accuracy***
  - **pros of sparse models:**
    - *smaller size*
    - easier to compute
    - smaller explanations
- **Model expressivity and size**
  - **DLs are more succinct than DTs**
  - **DLs are more succinct than DNFs**
    - *a special case of DSs*
  - **how to categorise DSs?**
    - ***empirically**, less succinct than DLs!*

Additional considerations 1

- **Comparing to heuristic methods**
  - **higher accuracy** *but*
  - **higher training time**
    - evolution of reasoning methods!
- **Other interpretable models**
  - **learning OBDDs**
    - SAT-based inference
- **Perfect vs. sparse models**
  - **pros of perfect models:**
    - ***highest possible accuracy***
  - **pros of sparse models:**
    - ***smaller size***
    - easier to compute
    - smaller explanations

- **Model expressivity and size**
  - **DLs are more succinct than DTs**
  - **DLs are more succinct than DNFs**
    - *a special case of DSs*
  - **how to categorise DSs?**
    - ***empirically**, less succinct than DLs!*
  - **OBDDs vs. other models?**

Additional considerations 2

- **Fairness and other constraints**
  - **model properties can be *enforced***
    - in the form of **constraints**
    - *easy to plug in!*

Additional considerations 2

- **Fairness and other constraints**
  - **model properties can be *enforced***
    - in the form of **constraints**
    - *easy to plug in!*
  - **fairness constraints**
    - learning ***fair DTs and DSs***
    - **accuracy** vs. **fairness**

Additional considerations 2

- **Fairness and other constraints**
  - **model properties can be *enforced***
    - in the form of **constraints**
    - *easy to plug in!*
  - **fairness constraints**
    - learning ***fair DTs and DSs***
    - **accuracy** vs. **fairness**

- **Intepretability**
  - **empirical considerations:**
    - $|\mathrm{XP}|$ for *perfect DSs* $<$ $|\mathrm{XP}|$ for *perfect DLs*

Additional considerations 2

- **Fairness and other constraints**
  - **model properties can be *enforced***
    - in the form of **constraints**
    - *easy to plug in!*
  - **fairness constraints**
    - learning ***fair DTs and DSs***
    - **accuracy** vs. **fairness**

- **Intepretability**
  - **empirical considerations:**
    - $|XP|$ for *perfect DSs* $<$ $|XP|$ for *perfect DLs*
    - $|XP|$ for *sparse DSs* $>$ $|XP|$ for *sparse DLs*

Additional considerations 2

- **Fairness and other constraints**
  - **model properties can be *enforced***
    - in the form of **constraints**
    - *easy to plug in!*
  - **fairness constraints**
    - learning ***fair DTs and DSs***
    - **accuracy** vs. **fairness**

- **Intepretability**
  - **empirical considerations:**
    - $|XP|$ for *perfect DSs* $<$ $|XP|$ for *perfect DLs*
    - $|XP|$ for *sparse DSs* $>$ $|XP|$ for *sparse DLs*
    - quality of sparsity metrics ***differs***

Additional considerations 2

- **Fairness and other constraints**
  - **model properties can be *enforced***
    - in the form of **constraints**
    - *easy to plug in!*
  - **fairness constraints**
    - learning ***fair DTs and DSs***
    - **accuracy** vs. **fairness**

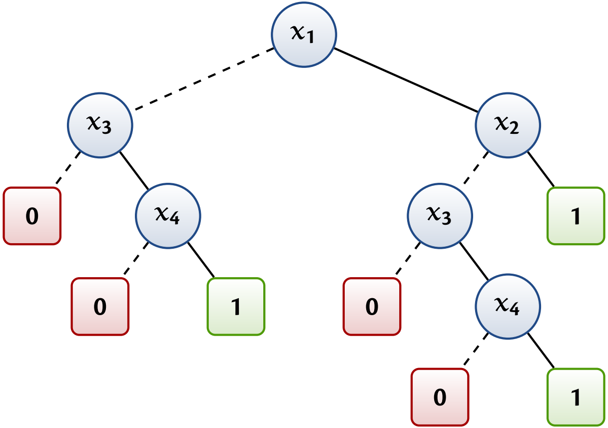
- **Intepretability**
  - **empirical considerations:**
    - $|XP|$ for *perfect DSs* $<$ $|XP|$ for *perfect DLs*
    - $|XP|$ for *sparse DSs* $>$ $|XP|$ for *sparse DLs*
    - quality of sparsity metrics ***differs***
  - **interpretability of DTs**

Interpretability of DTs – the issue

$$f(x_1, \ldots, x_n) = \bigvee_{i=1}^{n/2} x_{2i-1} \wedge x_{2i}, \textbf{ with } n = 4$$

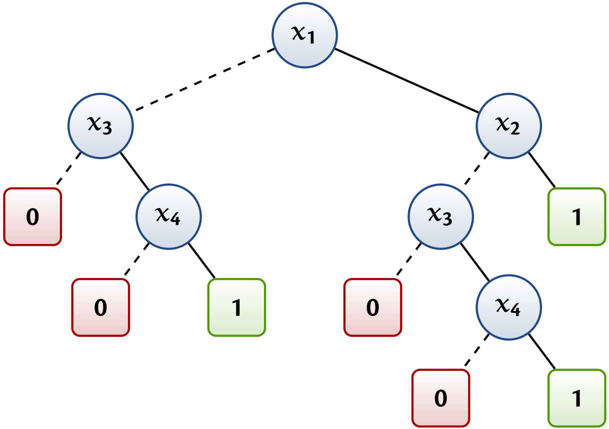
Interpretability of DTs – the issue

$$f(x_1, \ldots, x_n) = \bigvee_{i=1}^{n/2} x_{2i-1} \wedge x_{2i}, \text{ with } n = 4$$



Interpretability of DTs – the issue

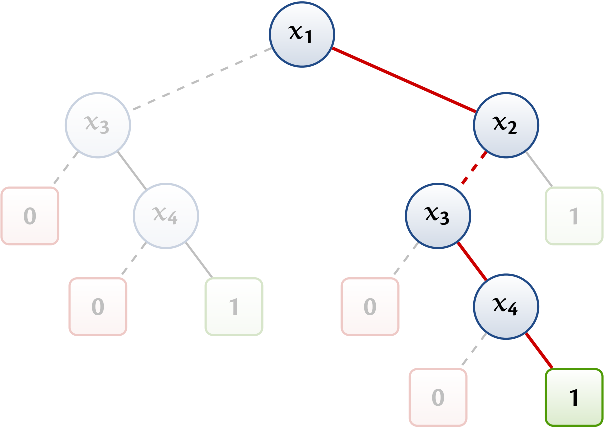
$$f(x_1, \ldots, x_n) = \bigvee_{i=1}^{n/2} x_{2i-1} \wedge x_{2i}, \text{ with } n = 4$$



instance $v = (1, 0, 1, 1)$, i.e. 4 literals in the path

Interpretability of DTs – the issue

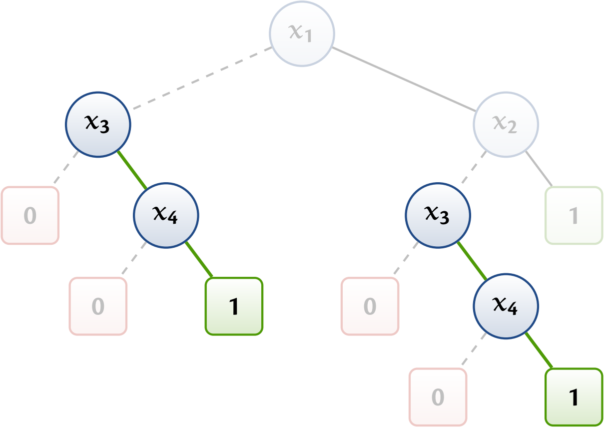
$$f(x_1,\ldots,x_n) = \bigvee_{i=1}^{n/2} x_{2i-1} \wedge x_{2i}, \text{ with } n = 4$$



instance $v = (1, 0, 1, 1)$, i.e. 4 literals in the path

Interpretability of DTs – the issue

$$f(x_1, \ldots, x_n) = \bigvee_{i=1}^{n/2} x_{2i-1} \wedge x_{2i}, \text{ with } n = 4$$



instance $v = (1, 0, 1, 1)$, i.e. 4 literals in the path

actual explanation $x_3 = 1 \wedge x_4 = 1$, i.e. 2 literals decision trees aren’t interpretable!

Additional considerations 2

- **Fairness and other constraints**
  - **model properties can be *enforced***
    - in the form of **constraints**
    - *easy to plug in!*
  - **fairness constraints**
    - learning ***fair DTs and DSs***
    - **accuracy** vs. **fairness**

- **Intepretability**
  - **empirical considerations:**
    - $|XP|$ for *perfect DSs* $<$ $|XP|$ for *perfect DLs*
    - $|XP|$ for *sparse DSs* $>$ $|XP|$ for *sparse DLs*
    - quality of sparsity metrics ***differs***
  - **DTs may be uninterpretable**
    - ***similar problem for DLs***

Additional considerations 2

- **Fairness and other constraints**
  - **model properties can be *enforced***
    - in the form of **constraints**
    - *easy to plug in!*
  - **fairness constraints**
    - learning ***fair DTs and DSs***
    - **accuracy** vs. **fairness**

- **Intepretability**
  - **empirical considerations:**
    - |XP| for *perfect DSs* **<** |XP| for *perfect DLs*
    - |XP| for *sparse DSs* **>** |XP| for *sparse DLs*
    - quality of sparsity metrics ***differs***
  - **DTs may be uninterpretable**
    - ***similar problem for DLs***
  - **AXps for DTs – in polytime!**
    - ***not the case for DLs and DSs!***

Thank you!