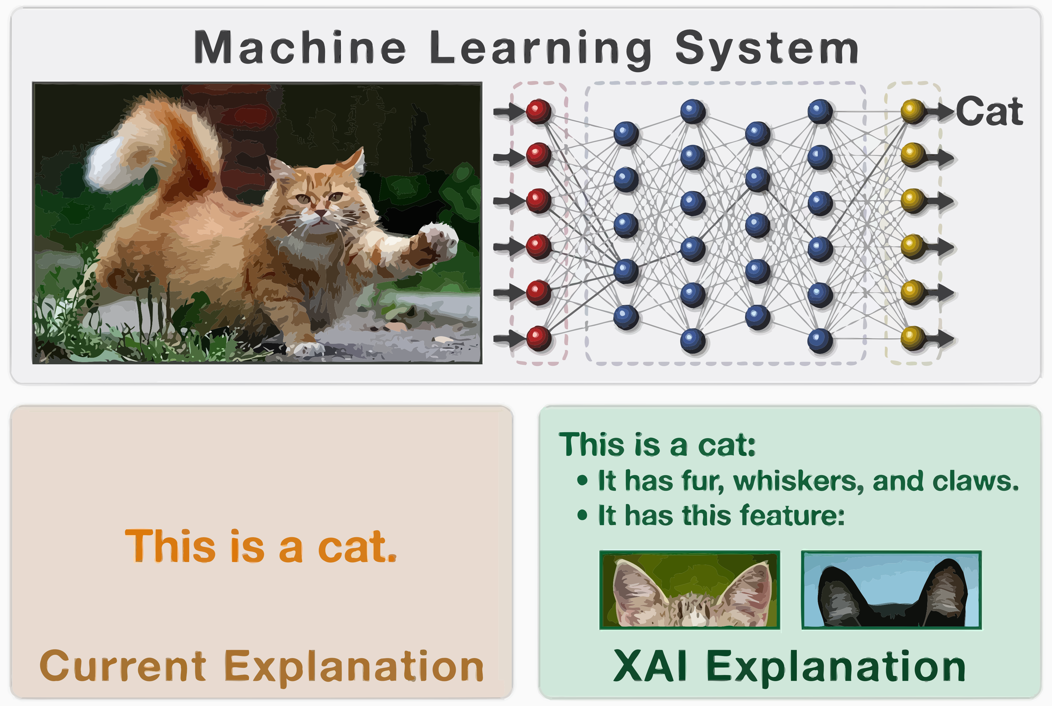
Reasoning-Based Learning of Interpretable ML Models

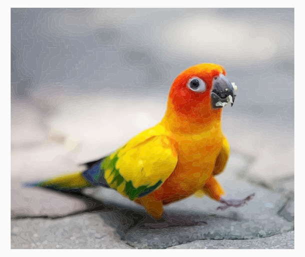
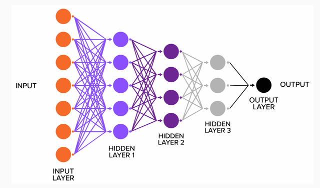
Alexey Ignatiev[1], Joao Marques-Silva[2], Nina Narodytska[3], and Peter J. Stuckey[1]

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

eXplainable AI



Why? Status Quo

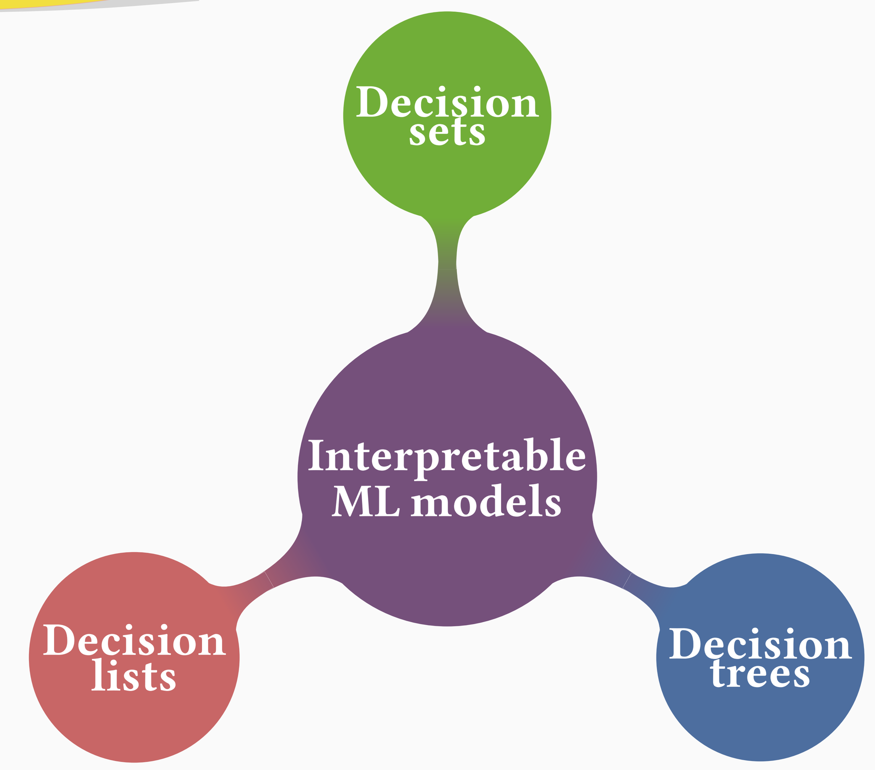
	A parrot	Machine learning algorithm
		
Learns random phrases	✔	✔
Doesn't understand s**t about what it learns	✔	✔
Occasionally speaks nonsense	✔	✔

Interpretable Models

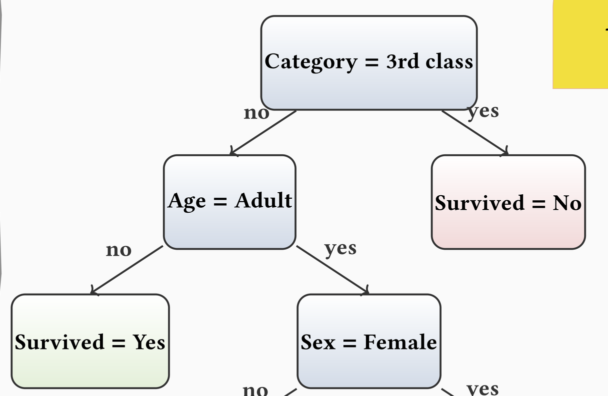
rule-based models

"*transparent*" and easy to interpret

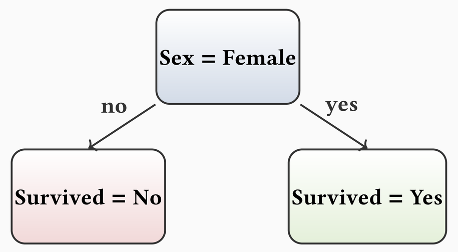
come in handy in XAI



Perfect and sparse DTs



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



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

Reasoning-based approaches to DTs

	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		✔						✔+	

Perfect and sparse DLs and DSs

IF	Age = Adult ∧ Sex ≠ Female	**THEN**	Survived = No
ELSE IF	Category ≠ 3rd class	**THEN**	Survived = Yes
		ELSE	Survived = No

smallest 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%)

IF	Category = 3rd class	**THEN**	Survived = No
IF	Age = Adult ∧ Sex ≠ Female	**THEN**	Survived = No
IF	Category ≠ 3rd class ∧ Age ≠ Adult	**THEN**	Survived = Yes
IF	Category ≠ 3rd class ∧ Sex = Female	**THEN**	Survived = Yes

smallest perfect DS for *Titanic* dataset
(training accuracy 78.25%)

IF	Category = 3rd class	**THEN**	Survived = No
IF	Sex ≠ Female	**THEN**	Survived = No
IF	Category ≠ 3rd class ∧ Sex = Female	**THEN**	Survived = Yes

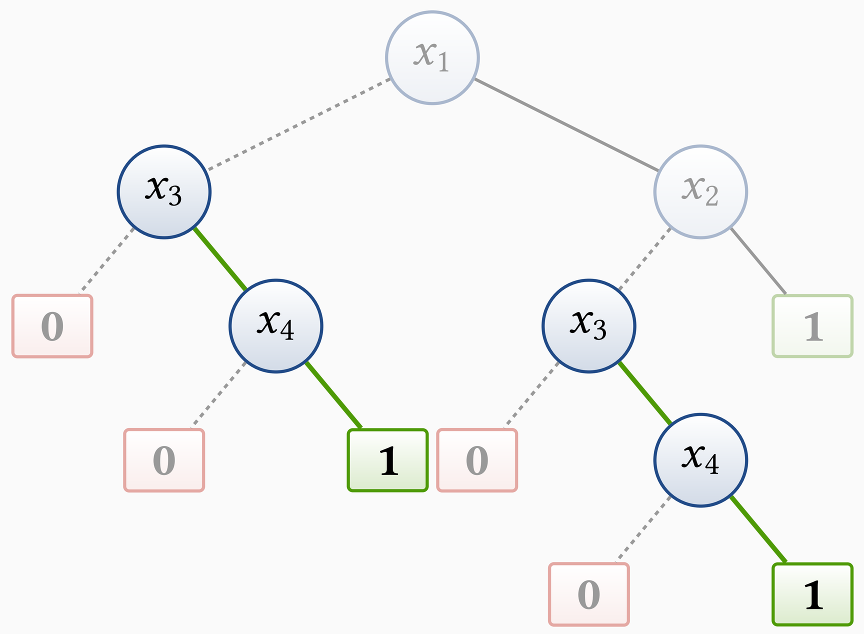
sparse DS for *Titanic* dataset
(training accuracy 77.57%)

Reasoning-based approaches to DLs and DSs

	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	✔	✔	✔	✔	✔		✔		✔	✔		✔

	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	✔		✔		✔		✔		✔	✔	✔	✔	

DT Interpretability Issue



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

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), i.e. rule R_5 fires the prediction
actual AXp: $x_3 = 1 \wedge x_4 = 1$, i.e. 2 literals

Additional remarks 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 remarks 2

- **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?**
- **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*
  - **DTs and DLs may be uninterpretable**
  - **AXps for DTs – in polytime!**
    - ***not the case for DLs and DSs!***

LaTeX TikZposter