Rigorous Explanations for Machine Learning Models

Joao Marques-Silva

(joint work with A. Ignatiev and N. Narodytska)

University of Lisbon, Portugal

AITP 2019 Conference

Obergurgl, Austria

April 2019

Progress in automated reasoning

- Automated reasoners (AR):
  - SAT
  - ILP

Progress in automated reasoning

- Automated reasoners (AR):
  - SAT
  - ILP
  - ASP
  - SMT
  - FOL

Progress in automated reasoning

- Automated reasoners (AR):
  - SAT
  - ILP
  - ASP
  - SMT
  - FOL

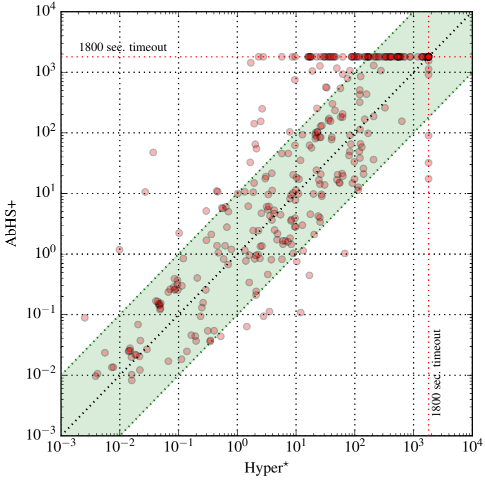
  - Reasoners as oracles
  - Reasoners within reasoners

Progress in automated reasoning & our work

- Automated reasoners (AR):
  - SAT
  - ILP
  - ASP
  - SMT
  - FOL

  - Reasoners as oracles
  - Reasoners within reasoners

Propositional abduction

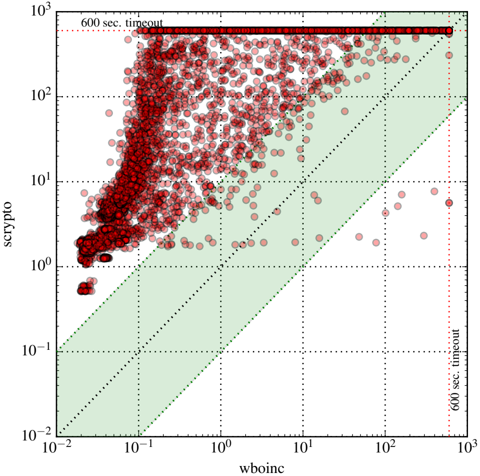


Progress in automated reasoning & our work

- Automated reasoners (AR):
  - SAT
  - ILP
  - ASP
  - SMT
  - FOL

  - Reasoners as oracles
  - Reasoners within reasoners

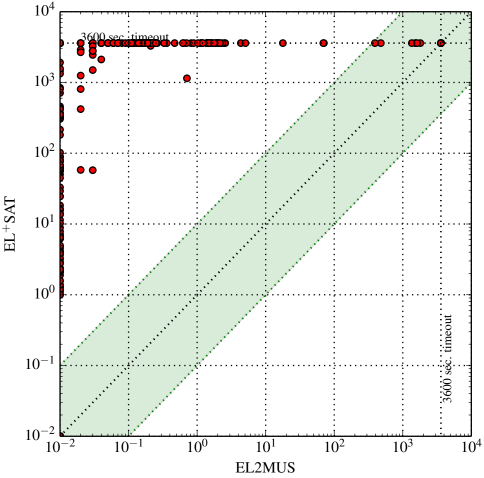
Model-based diagnosis



Progress in automated reasoning & our work

- Automated reasoners (AR):
  - SAT
  - ILP
  - ASP
  - SMT
  - FOL
  - Reasoners as oracles
  - Reasoners within reasoners

Axiom pinpointing for $\mathcal{EL}+$



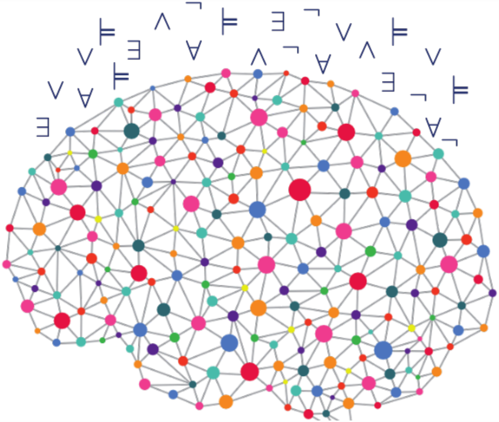
The question: how can AR improve ML's robustness?

Moshe Vardi

Machine learning and logic: Fast and slow thinking

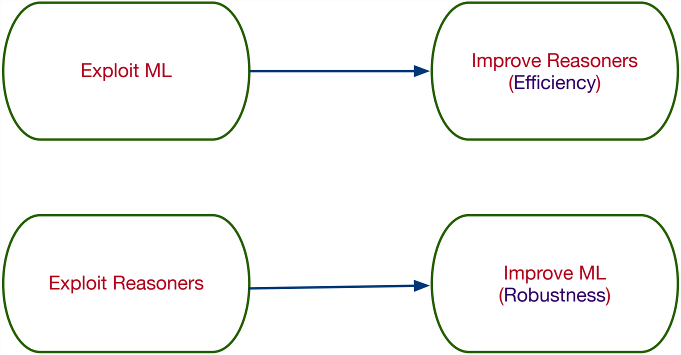
ABSTRACT. There is a recent perception that computer science is undergoing a Kuhnian paradigm shift, with CS as a model-driven science being replaced as a data-driven science. I will argue that, in general new scientific theories refine old scientific theories, rather than replace them. Thus, data-driven CS and model-driven CS complement each other, just as fast thinking and slow thinking complement each other in human thinking, as explicated by Daniel Kahneman. I will then use automated vehicles as an example, where in recent years, car makers and tech companies have been racing to be the first to market. In this rush there has been little discussion of how to obtain scalable standardization of safety assurance, without which this technology will never be commercially deployable. Such assurance requires formal methods, and combining machine learning with logic is the challenge of the day.

M. Vardi, MLMFM'18 Summit



Machine learning vs. automated reasoning

Machine learning vs. automated reasoning



Our work ...

- Focus on classification problems

Our work ...

- Focus on classification problems
- Globally correct (ie rigorous) **explanations** for predictions made

Our work ...

- Focus on classification problems
- Globally correct (ie rigorous) **explanations** for predictions made

- Disclaimer: first inroads into ML & XAI;
  comments welcome

Outline

Successes & Pitfalls of ML

Explainable AI

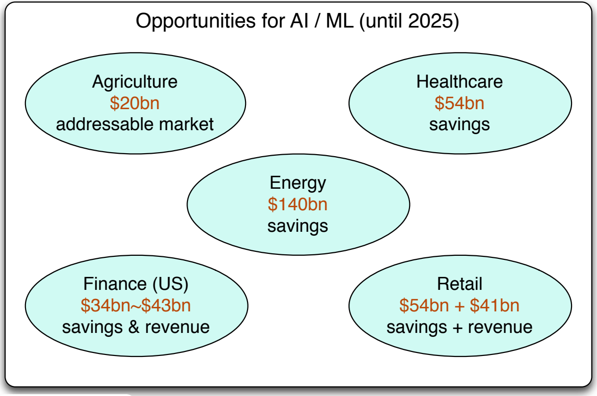
Explanations with Abductive Reasoning

Results

Some ML successes & expectations

- IBM Watson
- Deepmind AlphaGo
  - & AlphaZero
- Image Recognition
- Speech Recognition
- Financial Services
- Medical Diagnosis
- ...

Circa 2017



Source: Goldman-Sachs

Many more applications expected

source: Google

Many more applications expected

source: Google

source: Wikipedia

ⓒDARPA

But ML models are **brittle**

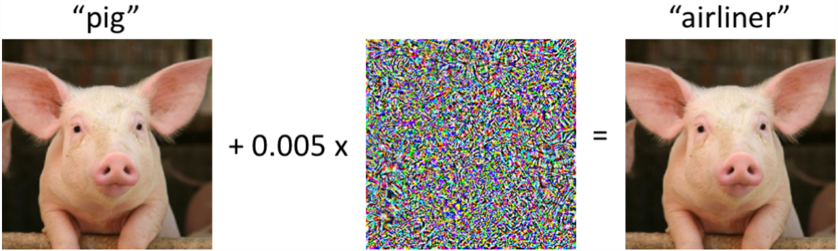
Eykholt et al'18

Aung et al'17

But ML models are **brittle**

Eykholt et al'18

Aung et al'17



Source: http://gradientscience.org/intro_adversarial/

Also, some ML models are **interpretable**

decision|rule lists|sets
decision trees

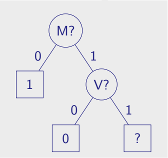
Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1

Also, some ML models are **interpretable**

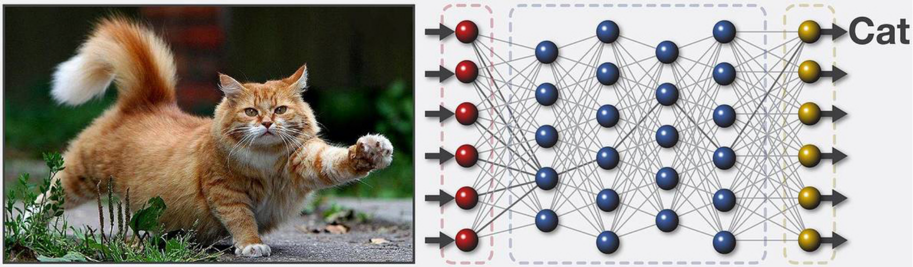
decision|rule lists|sets
decision trees

if $\neg$Meeting **then** Hike
if $\neg$Vacation **then** $\neg$Hike

Ex.	Vacation (V)	Concert (C)	Meeting (M)	Expo (E)	Hike (H)
e_1	0	0	1	0	0
e_2	1	0	0	0	1
e_3	0	0	1	1	0
e_4	1	0	0	1	1
e_5	0	1	1	0	0
e_6	0	1	1	1	0
e_7	1	1	0	1	1



But other ML models are **not** (interpretable)...



ⓒ DARPA

Why does the NN predict a cat?

Sample of ongoing efforts

- **Verification of NNs**:
  - Sound vs. unsound vs. complete [M.P. Kumar, VMCAI'19]
  - E.g. Reluplex: dedicated reasoning within SMT solver

- **Explanations for non-interpretable (ie black-box) models**:
  - Until recently, most approaches heuristic-based

Outline

Successes & Pitfalls of ML

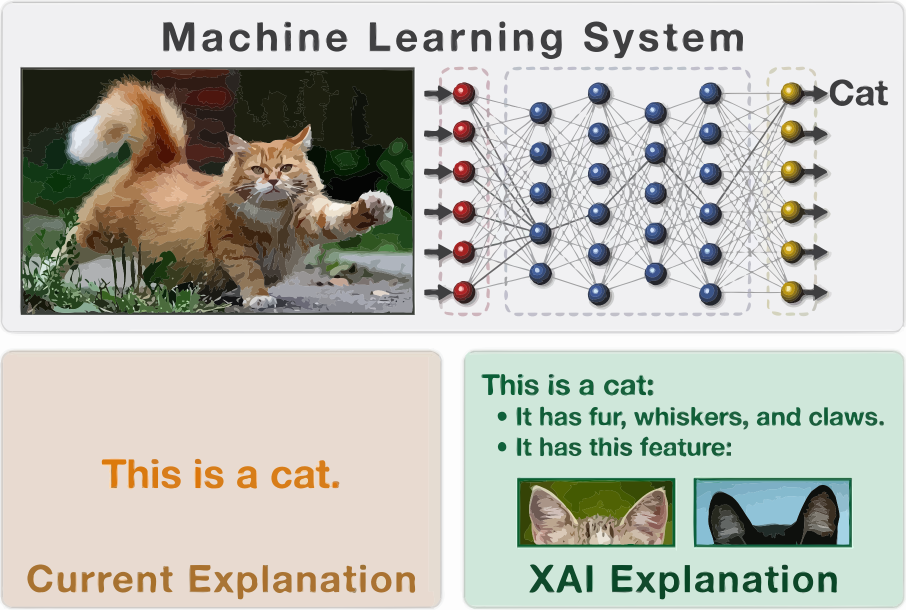
Explainable AI

Explanations with Abductive Reasoning

Results

What is eXplainable AI (XAI)?

What is eXplainable AI (XAI)?



©DARPA

Why XAI?

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

Why XAI?

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

European Union regulations on algorithmic decision-making and a “right to explanation”

Bryce Goodman,[1*] Seth Flaxman,[2]

Why XAI?

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman,[1*] Seth Flaxman,[2]

■ *We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine-learning algorithms. Slated to take effect as law across the European Union in 2018, it will place restrictions on automated individual decision making (that is, algorithms that make decisions based on user-level predictors) that "significantly affect" users. When put into practice, the law may also effectively create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision that significantly affects them. We argue that while this law may pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks that avoid discrimination and enable explanation.*

Why XAI?

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman,[1*] Seth Flaxman,[2]

SUMMIT ON MACHINE LEARNING MEETS FORMAL METHODS

■ *We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine-learning algorithms. Slated to take effect as law across the European Union in 2018, it will place restrictions on automated individual decision making (that is, algorithms that make decisions based on user-level predictors) that "significantly affect" users. When put into practice, the law may also effectively create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision that significantly affects them. We argue that while this law may pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks that avoid discrimination and enable explanation.*

Why XAI?

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

(Text with EEA relevance)

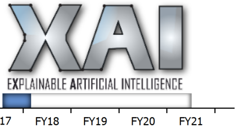
European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman,[1*] Seth Flaxman,[2]

SUMMIT ON MACHINE LEARNING MEETS FORMAL METHODS

■ *We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine-learning algorithms. Slated to take effect as law across the European Union in 2018, it will place restrictions on automated individual decision making (that is, algorithms that make decisions based on user-level predictors) that "significantly affect" users. When put into practice, the law may also effectively create a right to explanation, whereby a user can ask for an explanation of an algorithmic decision that significantly affects them. We argue that while this law may pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks that avoid discrimination and enable explanation.*

Explainable Artificial Intelligence (XAI)



David Gunning
DARPA/I2O
Program Update November 2017

Relevancy of XAI

MIT Technology Review

The Dark Secret at the Heart of AI
Will Knight
April 11, 2017

THE WALL STREET JOURNAL. WSJ

Inside DARPA's Push to Make Artificial Intelligence Explain Itself
Sara Castellanos and Steven Norton
August 10, 2017

The New York Times Magazine

Can A.I. Be Taught to Explain Itself?
Cliff Kuang
November 21, 2017

Intelligent Machines Are Asked to Explain How Their Minds Work
Richard Waters
July 11, 2017

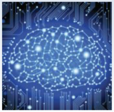
FT FINANCIA

The Register

You better explain yourself, mister: DARPA's mission to make an accountable AI
Dan Robinson
September 29, 2017

ExecutiveBiz

Charles River Analytics-Led Team Gets DARPA Contract to Support Artificial Intelligence Program
Ramona Adams
June 13, 2017



Entrepreneur

Elon Musk and Mark Zuckerberg Are Arguing About AI -- But They're Both Missing the Point
Artur Kiulian
July 28, 2017

Team investigates artificial intelligence, machine learning in DARPA project
Lisa Daigle
June 14, 2017

Military EMBEDDED SYSTEMS

FAST COMPANY

Why The Military And Corporate America Want To Make AI Explain Itself
Steven Melendez
June 22, 2017

NOVA NEXT

Ghosts in the Machine
Christina Couch
October 25, 2017

Jane's

DARPA's XAI seeks explanations from autonomous systems
Geoff Fein
November 16, 2017

COMPUTERWORLD

Oracle quietly researching 'Explainable AI'
George Nott
May 5, 2017

SCIENTIFIC AMERICAN

Demystifying the Black Box That Is AI
Ariel Bleicher
August 9, 2017

How AI detectives are cracking open the black box of deep learning
Paul Voosen
July 6, 2017

Science AAAS

Relevancy of XAI & hundreds(?) of recent papers

MIT Technology Review

The Dark Secret at the Heart of AI
Will Knight
April 11, 2017

THE WALL STREET JOURNAL. WSJ

Inside DARPA's Push to Make Artificial Intelligence Explain Itself
Sara Castellanos and Steven Norton
August 10, 2017

The New York Times Magazine

Can A.I. Be Taught to Explain Itself?
Cliff Kuang
November 21, 2017

Intelligent Machines Are Asked to Explain How Their Minds Work
Richard Waters
July 11, 2017

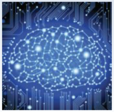
FT FINANCIA

The Register

You better explain yourself, mister: DARPA's mission to make an accountable AI
Dan Robinson
September 29, 2017

ExecutiveBiz

Charles River Analytics-Led Team Gets DARPA Contract to Support Artificial Intelligence Program
Ramona Adams
June 13, 2017



Entrepreneur

Elon Musk and Mark Zuckerberg Are Arguing About AI -- But They're Both Missing the Point
Artur Kiulian
July 28, 2017

Team investigates artificial intelligence, machine learning in DARPA project
Lisa Daigle
June 14, 2017

Military EMBEDDED SYSTEMS

NOVA NEXT

Ghosts in the Machine
Christina Couch
October 25, 2017

FAST COMPANY

Why The Military And Corporate America Want To Make AI Explain Itself
Steven Melendez
June 22, 2017

Jane's

DARPA's XAI seeks explanations from autonomous systems
Geoff Fein
November 16, 2017

COMPUTERWORLD

Oracle quietly researching 'Explainable AI'
George Nott
May 5, 2017

SCIENTIFIC AMERICAN

Demystifying the Black Box That Is AI
Ariel Bleicher
August 9, 2017

How AI detectives are cracking open the black box of deep learning
Paul Voosen
July 6, 2017

Science AAAS

How to XAI?

Main challenge: **black-box models**

Heuristic approaches, e.g. **LIME** & **Anchor** [Guerreiro et al., KDD'16, AAAI'18]

- Compute **local** explanations ...

How to XAI?

Main challenge: **black-box models**

Heuristic approaches, e.g. **LIME** & **Anchor** [Guerreiro et al., KDD'16, AAAI'18]

- Compute **local** explanations ...
- ... offer **no** guarantees

How to XAI?

Main challenge: **black-box models**

Heuristic approaches, e.g. **LIME** & **Anchor** [Guerreiro et al., KDD'16, AAAI'18]

- Compute **local** explanations ...
- ... offer **no** guarantees

Recent efforts on **rigorous** approaches

- **Compilation**-based, e.g. for BNCs [Shih,Choi&Darwiche, IJCAI'18]
  - Issues with scalability
- **Abduction**-based, e.g. for NNs [Ignatiev,Narodytska,M.-S., AAAI'19]
  - Issues with scalability

How to XAI?

Main challenge: **black-box models**

Heuristic approaches, e.g. **LIME** & **Anchor** [Guerreiro et al., KDD'16, AAAI'18]

- Compute **local** explanations ...
- ... offer **no** guarantees

Recent efforts on **rigorous** approaches

- **Compilation**-based, e.g. for BNCs [Shih,Choi&Darwiche, IJCAI'18]
  - Issues with scalability
- **Abduction**-based, e.g. for NNs [Ignatiev,Narodytska,M.-S., AAAI'19]
  - Issues with scalability, but less significant

Some current challenges

- For heuristic methods: lack of rigor (more later)

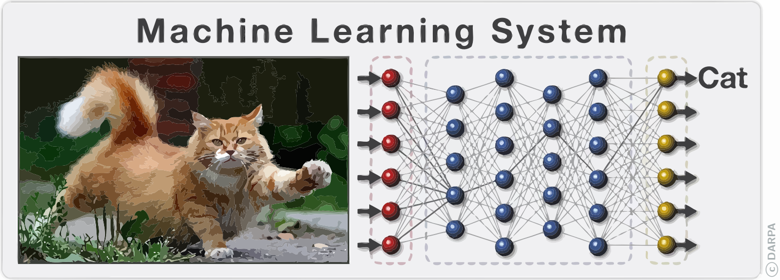
Some current challenges

- For heuristic methods: lack of rigor (more later)

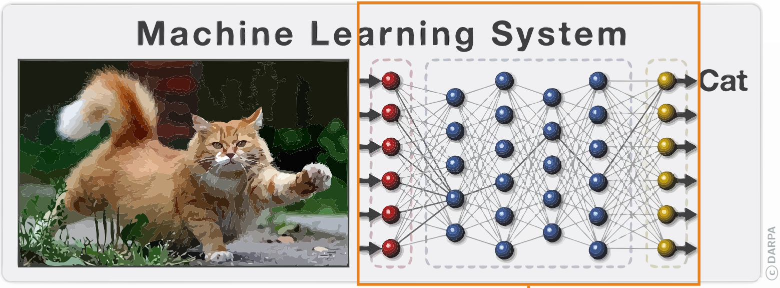
- For rigorous methods: scalability, scalability, scalability...

Outline

From ML model to logic

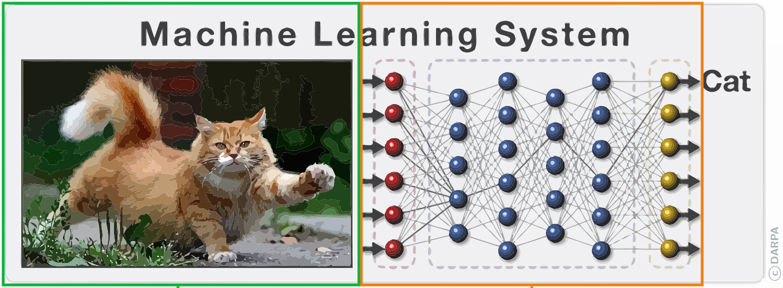


From ML model to logic



formula $\mathcal{F}$

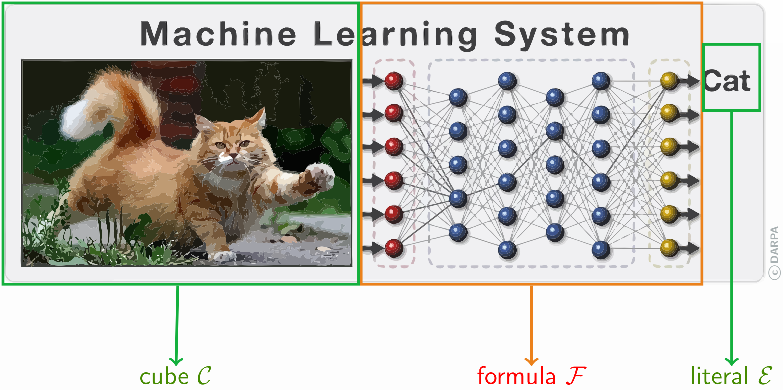
From ML model to logic



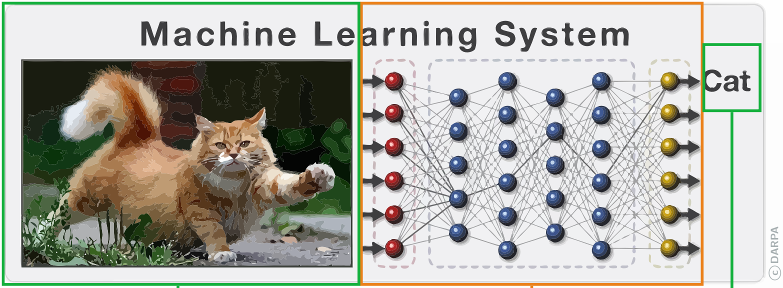
cube $\mathcal{C}$

formula $\mathcal{F}$

From ML model to logic



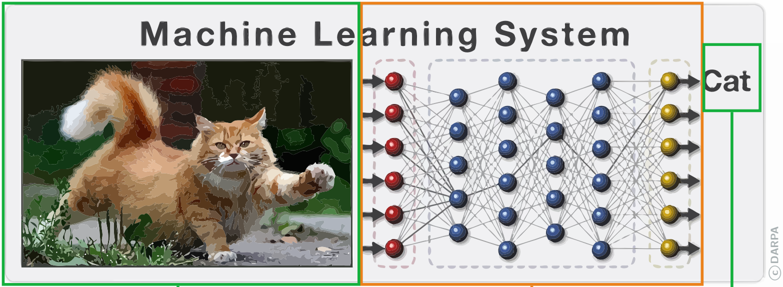
From ML model to logic



cube $\mathcal{C}$ formula $\mathcal{F}$ literal $\mathcal{E}$

$$\mathcal{C} \wedge \mathcal{F} \models \mathcal{E}$$

From ML model to logic



cube $\mathcal{C}$

formula $\mathcal{F}$

literal $\mathcal{E}$

$$\mathcal{C} \wedge \mathcal{F} \vDash \mathcal{E}$$

Must be able to encode ML model
E.g. SMT, ILP, etc.

Abductive explanations of ML models

given a **classifier** $\mathcal{F}$, a **cube** $\mathcal{C}$ and a **prediction** $\mathcal{E}$,

Abductive explanations of ML models

given a **classifier** $\mathcal{F}$, a **cube** $\mathcal{C}$ and a **prediction** $\mathcal{E}$,
compute a (**subset- or cardinality-**) minimal $\mathcal{C}_m \subseteq \mathcal{C}$ s.t.

Abductive explanations of ML models

given a **classifier** $\mathcal{F}$, a **cube** $\mathcal{C}$ and a **prediction** $\mathcal{E}$,
compute a (**subset- or cardinality-**) minimal $\mathcal{C}_m \subseteq \mathcal{C}$ s.t.

$$\mathcal{C}_m \wedge \mathcal{F} \not\models \perp$$

and

$$\mathcal{C}_m \wedge \mathcal{F} \models \mathcal{E}$$

Abductive explanations of ML models

given a **classifier** $\mathcal{F}$, a **cube** $\mathcal{C}$ and a **prediction** $\mathcal{E}$,

compute a (**subset- or cardinality-**) minimal $\mathcal{C}_m \subseteq \mathcal{C}$ s.t.

$$\mathcal{C}_m \wedge \mathcal{F} \not\models \bot$$

and

$$\mathcal{C}_m \wedge \mathcal{F} \models \mathcal{E}$$

$\Downarrow$

iterative explanation procedure

Computing primes

1. $\mathcal{C}_m \wedge \mathcal{F} \not\models \bot$

Computing primes

1. $\mathcal{C}_m \wedge \mathcal{F} \not\models \bot$ — **tautology**

Computing primes

1. $\mathcal{C}_m \wedge \mathcal{F} \not\models \bot$ — **tautology**
2. $\mathcal{C}_m \wedge \mathcal{F} \models \mathcal{E}$

Computing primes

$$
\begin{array}{lccc}
\textbf{1.} & \mathcal{C}_m \wedge \mathcal{F} \not\models \bot & \textbf{—} & \textbf{tautology} \\
\textbf{2.} & \mathcal{C}_m \wedge \mathcal{F} \models \mathcal{E} & \Leftrightarrow & \mathcal{C}_m \models (\mathcal{F} \rightarrow \mathcal{E})
\end{array}
$$

Computing primes

1. $\mathcal{C}_m \wedge \mathcal{F} \not\models \bot$ — **tautology**

2. $\mathcal{C}_m \wedge \mathcal{F} \models \mathcal{E} \quad \Leftrightarrow \quad \mathcal{C}_m \models (\mathcal{F} \rightarrow \mathcal{E})$

$\Downarrow$

$\mathcal{C}_m$ is a **prime implicant** of $\mathcal{F} \rightarrow \mathcal{E}$

Computing one minimal explanation

- **One** subset-minimal explanation:

Input: $\mathcal{F}$ under $\mathcal{M}$, initial cube $\mathcal{C}$, prediction $\mathcal{E}$
Output: **Subset-minimal** explanation $\mathcal{C}_m$

```
begin
   for l ∈ C :
      if Entails(C \ {l}, F → E) :
         C ← C \ {l}
   return C
end
```

Computing one minimal explanation

- **One** subset-minimal explanation:

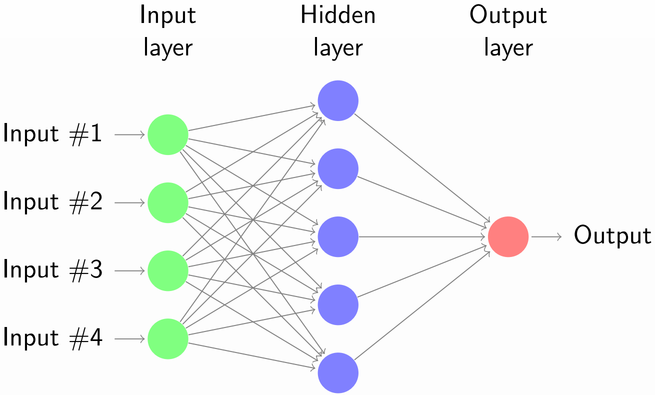
  Input: $\mathcal{F}$ under $\mathcal{M}$, initial cube $\mathcal{C}$, prediction $\mathcal{E}$
  Output: **Subset-minimal** explanation $\mathcal{C}_m$

```
begin
    for l ∈ C :
        if Entails(C \ {l}, F → E) :
            C ← C \ {l}
    return C
end
```

- **One** cardinality-minimal explanation:
  - Harder than computing subset-minimal explanation
  - Exploit **implicit hitting set dualization**
  - Details in earlier papers

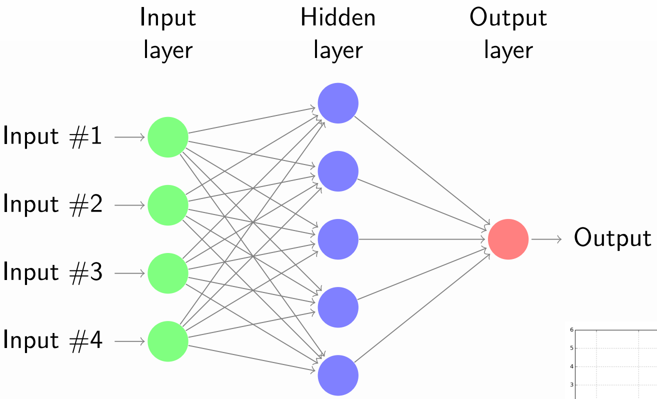
Outline

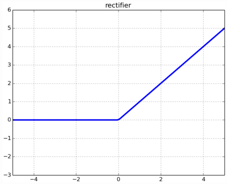
Encodings NNs



- Each layer (except first) viewed as a block
  - Compute $\mathbf{x}'$ given input $\mathbf{x}$, weights matrix $\mathbf{A}$, and bias vector $\mathbf{b}$
  - Compute output $\mathbf{y}$ given $\mathbf{x}'$ and activation function

Encodings NNs





- Each layer (except first) viewed as a block
  - Compute $\mathbf{x}'$ given input $\mathbf{x}$, weights matrix $\mathbf{A}$, and bias vector $\mathbf{b}$
  - Compute output $\mathbf{y}$ given $\mathbf{x}'$ and activation function
- Each unit uses a ReLU activation function

Encoding NNs using MILP

Computation for a NN ReLU block:

$$\mathbf{x}' = \mathbf{A} \cdot \mathbf{x} + \mathbf{b}$$
$$\mathbf{y} = \max(\mathbf{x}', \mathbf{0})$$

Encoding NNs using MILP

Computation for a NN ReLU block:

$$\mathbf{x}' = \mathbf{A} \cdot \mathbf{x} + \mathbf{b}$$
$$\mathbf{y} = \max(\,\mathbf{x}', \mathbf{0})$$

Block encoded as follows: [Fischetti&Jo, CJ'18]

$$\sum_{j=1}^{n} a_{i,j} x_j + b_i = y_i - s_i$$
$$z_i = 1 \rightarrow y_i \leq 0$$
$$z_i = 0 \rightarrow s_i \leq 0$$
$$y_i \geq 0, s_i \geq 0, z_i \in \{0, 1\}$$

– Simpler encodings not as effective [Katz et al. CAV'17]

Outline

Experimental setup

- Implementation in Python
  - Supports SMT solvers through PySMT
    - Yices2 used
  - Supports CPLEX 12.8.0

Experimental setup

- Implementation in Python
  - Supports SMT solvers through PySMT
    - Yices2 used
  - Supports CPLEX 12.8.0
- **ReLU-based** neural networks [Fischetti&Jo CJ'18]
  - One *hidden* layer with $i \in \{10, 15, 20\}$ neurons
  - Pick NN that achieves *good* accuracy

Experimental setup

- Implementation in Python
  - Supports SMT solvers through PySMT
    - Yices2 used
  - Supports CPLEX 12.8.0
- **ReLU-based** neural networks [Fischetti&Jo CJ'18]
  - One *hidden* layer with $i \in \{10, 15, 20\}$ neurons
  - Pick NN that achieves *good* accuracy
- Benchmarks selected from:
  - **UCI** Machine Learning Repository
  - **Penn** Machine Learning Benchmarks
  - **MNIST** Digits Database

Experimental setup

- Implementation in Python
  - Supports SMT solvers through PySMT
    - Yices2 used
  - Supports CPLEX 12.8.0
- **ReLU-based** neural networks [Fischetti&Jo CJ'18]
  - One *hidden* layer with $i \in \{10, 15, 20\}$ neurons
  - Pick NN that achieves *good* accuracy
- Benchmarks selected from:
  - **UCI** Machine Learning Repository
  - **Penn** Machine Learning Benchmarks
  - **MNIST** Digits Database
- Machine configuration:
  - Intel Core i7 2.8GHz, 8GByte
  - Time limit — 1800s
  - Memory limit — 4GByte

Sample of experimental results

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

Sample of experimental results

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

Sample of experimental results

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

Sample of experimental results

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

Comparing quality to compilation-based BNC

[Shih,Choi&Darwiche, IJCAI'18]

- *"Congressional Voting Records"* dataset

Comparing quality to compilation-based BNC

[Shih,Choi&Darwiche, IJCAI'18]

- *"Congressional Voting Records"* dataset
- `(0 1 0 1 1 1 0 0 0 0 0 0 1 1 0 1)` — data sample **(16 features)**

Comparing quality to compilation-based BNC

[Shih,Choi&Darwiche, IJCAI'18]

- *"Congressional Voting Records"* dataset
- (0 1 0 1 1 1 0 0 0 0 0 0 1 1 0 1) — data sample **(16 features)**

smallest size explanations computed by:

- (0 1 1 0 0 0 1 1 0) — **9 literals**
- (0 1 1 1 0 0 1 1 0) — **9 literals**

Comparing quality to compilation-based BNC

[Shih,Choi&Darwiche, IJCAI'18]

- *"Congressional Voting Records"* dataset
- `(0 1 0 1 1 1 0 0 0 0 0 0 1 1 0 1)` — data sample **(16 features)**

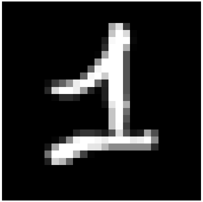
smallest size explanations computed by:

- `(    0 1 1   0 0 0       1 1 0  )` — **9 literals**
- `(    0 1 1 1   0 0       1 1 0  )` — **9 literals**

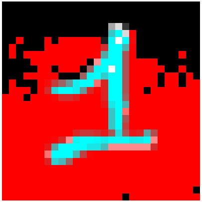
subset-minimal explanations computed by **our approach**:

- `(      1         0   0       0  )` — **4 literals**
- `(      1         0   0           )` — **3 literals**
- `(    0 1         0   0       0  )` — **5 literals**
- `(    0 1         0   0         1)` — **5 literals**

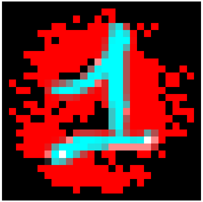
There are many explanations of different quality



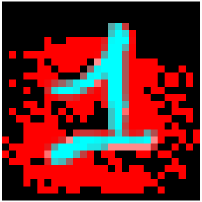
(a) digit 1



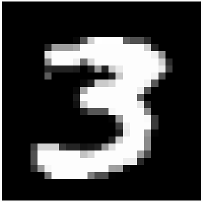
(b) simple expl.



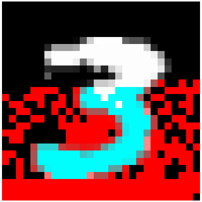
(c) central pixels



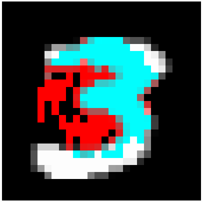
(d) light pixels



(a) digit 3



(b) simple expl.



(c) central pixels

(d) light pixels

Outline

Successes & Pitfalls of ML

Explainable AI

Explanations with Abductive Reasoning

Results

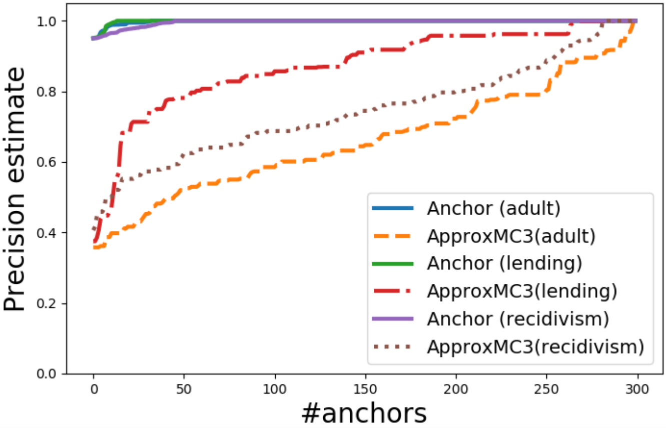
- Assessing Local Explanations – Recent Work

Assessing precision with model counting

- Evaluated Anchor [Guerreiro et al., AAAI18]
  - Anchor more accurate than LIME
  - Anchor computes accuracy estimate for each explanation

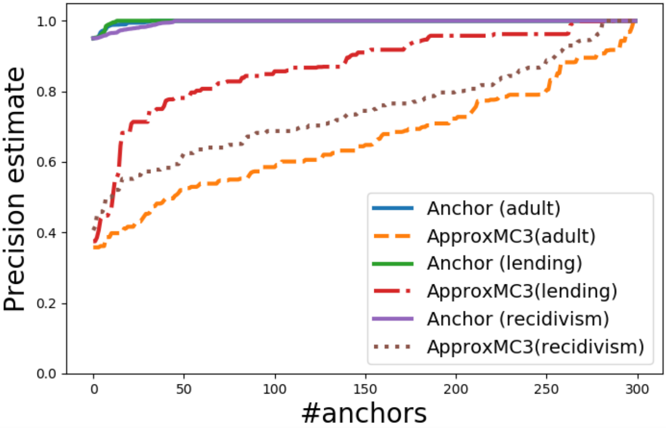
- Represented ML model as propositional formula
  - E.g. binarized NNs (BNNs)
  - Use (approximate) model counter to assess precision of ML model on explanation (anchor) computed by Anchor

Preliminary results



- Anchor often claims $\approx 99\%$ precision

Preliminary results



- Anchor often claims $\approx$ 99% precision; this cannot be confirmed

Summary and future work

- **Principled** approach to XAI

Summary and future work

- **Principled** approach to XAI
- Based on **abductive reasoning**

Summary and future work

- **Principled** approach to XAI
- Based on **abductive reasoning**
- Applies a **reasoning engine**, e.g. SMT or MILP

Summary and future work

- **Principled** approach to XAI
- Based on **abductive reasoning**
- Applies a **reasoning engine**, e.g. SMT or MILP
- Provides **minimality guarantees**

Summary and future work

- **Principled** approach to XAI
- Based on **abductive reasoning**
- Applies a **reasoning engine**, e.g. SMT or MILP
- Provides **minimality guarantees**
- Tested on ReLU-based NNs
- First results on precision of Anchor's explanations

Summary and future work

- **Principled** approach to XAI
- Based on **abductive reasoning**
- Applies a **reasoning engine**, e.g. SMT or MILP
- Provides **minimality guarantees**
- Tested on ReLU-based NNs
- First results on precision of Anchor's explanations

- **Other** ML models?

Summary and future work

- **Principled** approach to XAI
- Based on **abductive reasoning**
- Applies a **reasoning engine**, e.g. SMT or MILP
- Provides **minimality guarantees**
- Tested on ReLU-based NNs
- First results on precision of Anchor's explanations

- **Other** ML models?
- Address scalability:
  - Better **encodings**?
  - More advanced **reasoners**?

Summary and future work

- **Principled** approach to XAI
- Based on **abductive reasoning**
- Applies a **reasoning engine**, e.g. SMT or MILP
- Provides **minimality guarantees**
- Tested on ReLU-based NNs
- First results on precision of Anchor's explanations

- **Other** ML models?
- Address scalability:
  - Better **encodings**?
  - More advanced **reasoners**?
- Explanation **enumeration**? + **preferences**?

Questions?

References to our work

- A. Ignatiev, N. Narodytska, J. Marques-Silva:
  Abduction-Based Explanations for Machine Learning Models.
  AAAI 2019
- N. Narodytska, A. Ignatiev, F. Pereira, J. Marques-Silva:
  Learning Optimal Decision Trees with SAT.
  IJCAI 2018: 1362-1368
- A. Ignatiev, F. Pereira, N. Narodytska, J. Marques-Silva:
  A SAT-Based Approach to Learn Explainable Decision Sets.
  IJCAR 2018: 627-645

Additional references I

- M. T. Ribeiro, S. Singh, C. Guestrin:
  "Why Should I Trust You?": Explaining the Predictions of Any Classifier.
  KDD 2016: 1135-1144
- G. Katz, C. W. Barrett, D. L. Dill, K. Julian, M. J. Kochenderfer:
  Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks.
  CAV (1) 2017: 97-117
- M. T. Ribeiro, S. Singh, C. Guestrin:
  Anchors: High-Precision Model-Agnostic Explanations.
  AAAI 2018: 1527-1535
- A Shih, A. Choi, A. Darwiche:
  A Symbolic Approach to Explaining Bayesian Network Classifiers.
  IJCAI 2018: 5103-5111
- M. Fischetti, J. Jo:
  Deep neural networks and mixed integer linear optimization.
  Constraints 23(3): 296-309 (2018)

Additional references II

- B. Goodman, S. R. Flaxman:
  European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". AI Magazine 38(3): 50-57 (2017)
- A. M. Aung, Y. Fadila, R. Gondokaryono, L. Gonzalez:
  Building Robust Deep Neural Networks for Road Sign Detection.
  CoRR abs/1712.09327 (2017)
- K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song:
  Robust Physical-World Attacks on Deep Learning Visual Classification.
  CVPR 2018: 1625-1634
- A. Madry, L. Schmidt:
  A Brief Introduction to Adversarial Examples.
  `http://gradientscience.org/intro_adversarial/`, 2018
- M. P. Kumar:
  Tutorial: Neural Network Verification.
  VMCAI 2019 Winter School
  `http://mpawankumar.info/tutorials/vmcai2019/`, 2019