

Abduction-Based Explanations for Machine Learning Models

Alexey Ignatiev¹, Nina Narodytska², Joao Marques-Silva¹

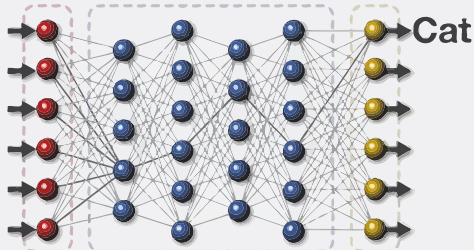
January 30, 2019

¹ Faculty of Science, University of Lisbon, Portugal

² VMWare Research, CA, USA

What is eXplainable AI (XAI)?

Machine Learning System



This is a cat.

Current Explanation

This is a cat:

- It has fur, whiskers, and claws.
- It has this feature:



XAI Explanation

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)

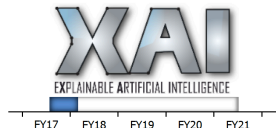
European Union regulations on algorithmic decision-making and a “right to explanation”

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



XAI controversy



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



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



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

NOVAX

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

heuristic approaches exist

(e.g. LIME or Anchor)

heuristic approaches exist

(e.g. LIME or Anchor)



- local explanations

heuristic approaches exist

(e.g. LIME or Anchor)



- local explanations
- no guarantees

heuristic approaches exist

(e.g. LIME or Anchor)

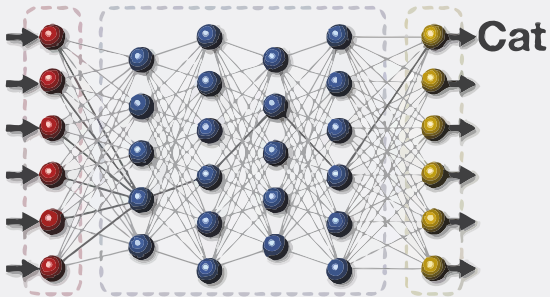


- local explanations
- no guarantees

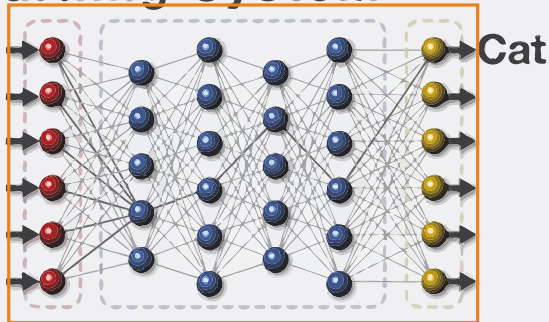


(un-)reliable?

Machine Learning System



Machine Learning System

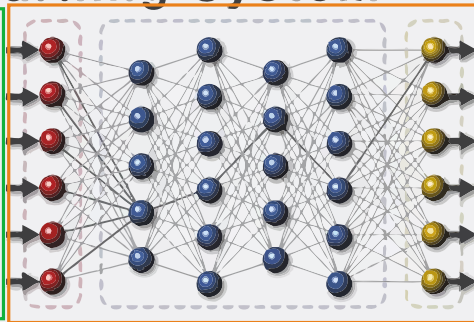


formula \mathcal{F}

Machine Learning System



cube \mathcal{C}



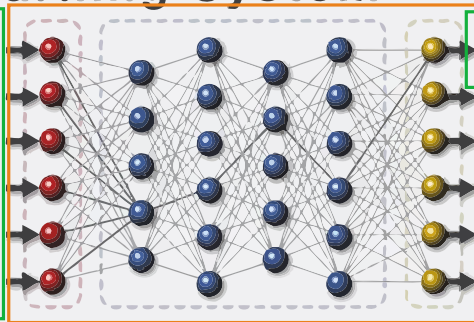
formula \mathcal{F}

Cat

Machine Learning System



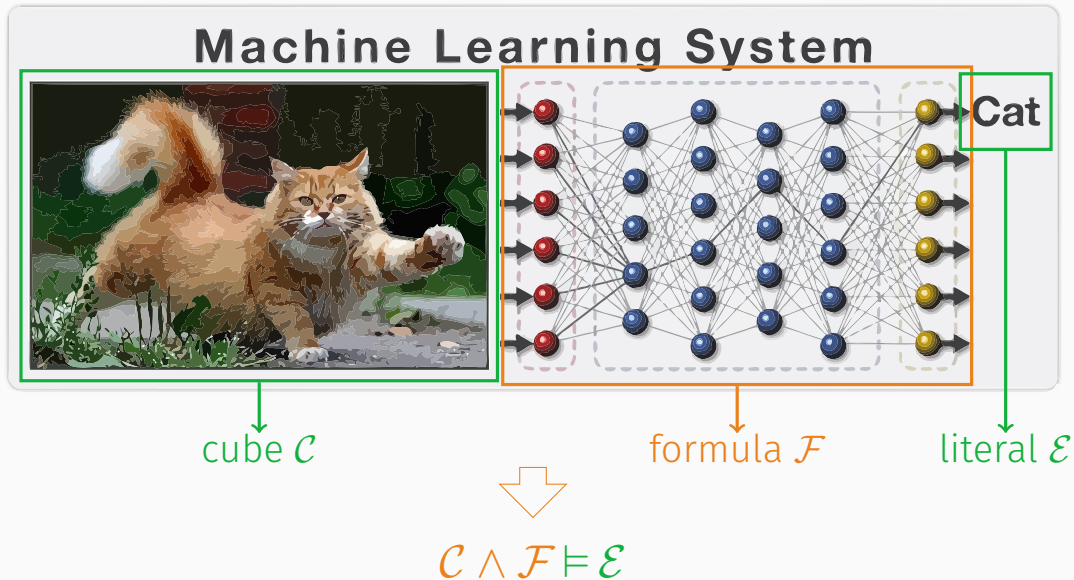
cube \mathcal{C}



formula \mathcal{F}

Cat

literal \mathcal{E}



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

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

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

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

and

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

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

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

and

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



iterative explanation procedure

1. $C_m \wedge \mathcal{F} \not\models \perp$

1. $C_m \wedge \mathcal{F} \not\models \perp$ — *tautology*

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

1. $\mathcal{C}_m \wedge \mathcal{F} \not\models \perp$ — *tautology*
2. $\mathcal{C}_m \wedge \mathcal{F} \models \mathcal{E} \iff \mathcal{C}_m \models (\mathcal{F} \rightarrow \mathcal{E})$

1. $C_m \wedge \mathcal{F} \not\models \perp$ — *tautology*
2. $C_m \wedge \mathcal{F} \models \mathcal{E} \iff C_m \models (\mathcal{F} \rightarrow \mathcal{E})$



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

Computing one subset-minimal explanation

Input: \mathcal{F} under \mathcal{M} , initial cube \mathcal{C} , prediction \mathcal{E}

Output: *Subset-minimal* explanation \mathcal{C}_m

```
1 begin  
2   for  $l \in \mathcal{C}$  :  
3     if  $\text{Entails}(\mathcal{C} \setminus \{l\}, \mathcal{F} \rightarrow \mathcal{E}, \mathcal{M})$  :  
4        $\mathcal{C} \leftarrow \mathcal{C} \setminus \{l\}$   
5   return  $\mathcal{C}$   
6 end
```

cardinality-minimal explanations can be computed
(following *implicit-hitting set* based approach¹)

¹Ignatiev, A.; Morgado, A.; and Marques-Silva, J. 2016. *Propositional abduction with implicit hitting sets*. In ECAI, 1327–1335

cardinality-minimal explanations can be computed

(following *implicit-hitting set* based approach¹)



see the paper

¹Ignatiev, A.; Morgado, A.; and Marques-Silva, J. 2016. *Propositional abduction with implicit hitting sets*. In ECAI, 1327–1335

Experimental setup

- implementation in Python
 - supports SMT solvers through PySMT
 - Yices2 used
 - supports CPLEX 12.8.0

²Fischetti, M., and Jo, J. 2018. *Deep neural networks and mixed integer linear optimization*. Constraints 23(3):296–309.

Experimental setup

- implementation in Python
 - supports SMT solvers through PySMT
 - Yices2 used
 - supports CPLEX 12.8.0
- ReLU-based neural networks²
 - one *hidden* layer with $i \in \{10, 15, 20\}$ neurons

²Fischetti, M., and Jo, J. 2018. *Deep neural networks and mixed integer linear optimization*. Constraints 23(3):296–309.

Experimental setup

- implementation in Python
 - supports **SMT** solvers through PySMT
 - **Yices2** used
 - supports **CPLEX 12.8.0**
- **ReLU-based** neural networks²
 - one *hidden* layer with $i \in \{10, 15, 20\}$ neurons
- benchmarks selected from:
 - **UCI** Machine Learning Repository
 - **Penn** Machine Learning Benchmarks
 - **MNIST** Digits Database

²Fischetti, M., and Jo, J. 2018. *Deep neural networks and mixed integer linear optimization*. Constraints 23(3):296–309.

Experimental setup

- implementation in Python
 - supports **SMT** solvers through PySMT
 - **Yices2** used
 - supports **CPLEX 12.8.0**
- **ReLU-based** neural networks²
 - one *hidden* layer with $i \in \{10, 15, 20\}$ neurons
- 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**

²Fischetti, M., and Jo, J. 2018. *Deep neural networks and mixed integer linear optimization*. Constraints 23(3):296–309.

Some of the 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

Some of the 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

Some of the 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

Some of the 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

Some of the 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 state of the art³

- “*Congressional Voting Records*” dataset

³Shih, A.; Choi, A.; and Darwiche, A. 2018. *A symbolic approach to explaining Bayesian network classifiers*. In IJCAI, 5103–5111

Comparing quality to state of the art³

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

³Shih, A.; Choi, A.; and Darwiche, A. 2018. *A symbolic approach to explaining Bayesian network classifiers*. In IJCAI, 5103–5111

Comparing quality to state of the art³

- “*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 ³:

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

³Shih, A.; Choi, A.; and Darwiche, A. 2018. *A symbolic approach to explaining Bayesian network classifiers*. In IJCAI, 5103–5111

Comparing quality to state of the art³

- “*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 ³:

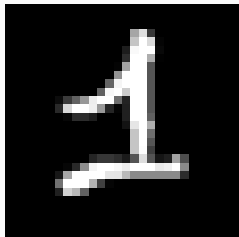
- (0 1 1 0 0 0 1 1 0) — 9 literals
- (0 1 1 1 0 0 1 1 0) — 9 literals

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

³Shih, A.; Choi, A.; and Darwiche, A. 2018. A symbolic approach to explaining Bayesian network classifiers. In IJCAI, 5103–5111

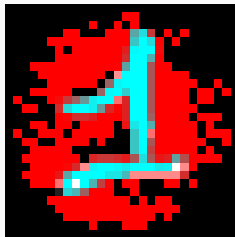
There are many explanations of different quality



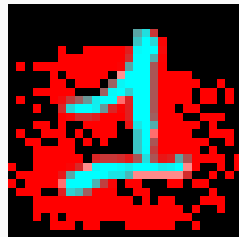
(a) digit 1



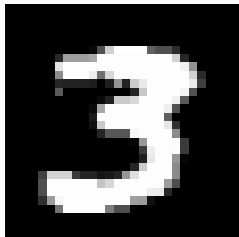
(b) simple expl.



(c) central pixels



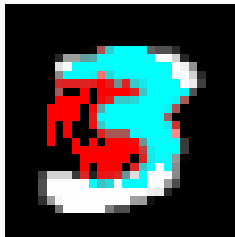
(d) light pixels



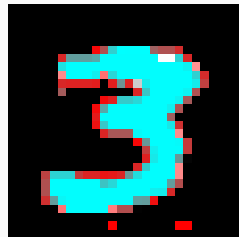
(a) digit 3



(b) simple expl.



(c) central pixels



(d) light pixels

- *principled* approach to XAI

- *principled* approach to XAI
- based on **abductive reasoning**

- *principled* approach to XAI
- based on **abductive reasoning**
- applies a *reasoning engine*, e.g. **SMT** or **MILP**

- *principled* approach to XAI
- based on **abductive reasoning**
- applies a *reasoning engine*, e.g. **SMT** or **MILP**
- provides **minimality guarantees**

- *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

- *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
- **other** ML models?

- *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

- **other** ML models?
- better scalability

- *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

- **other** ML models?
- better scalability
 - better *encodings*?

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

- **other** ML models?
- better scalability
 - better **encodings**?
 - more advanced **reasoners**?

- *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**

- **other** ML models?
- better scalability
 - better *encodings*?
 - more advanced *reasoners*?
 - *abstraction refinement*?

- *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

- **other** ML models?
- better scalability
 - better **encodings**?
 - more advanced **reasoners**?
 - **abstraction refinement**?
- explanation **enumeration**?

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**

- **other** ML models?
- better scalability
 - better **encodings**?
 - more advanced **reasoners**?
 - **abstraction refinement**?
- explanation **enumeration**? + **preferences**?

Questions?