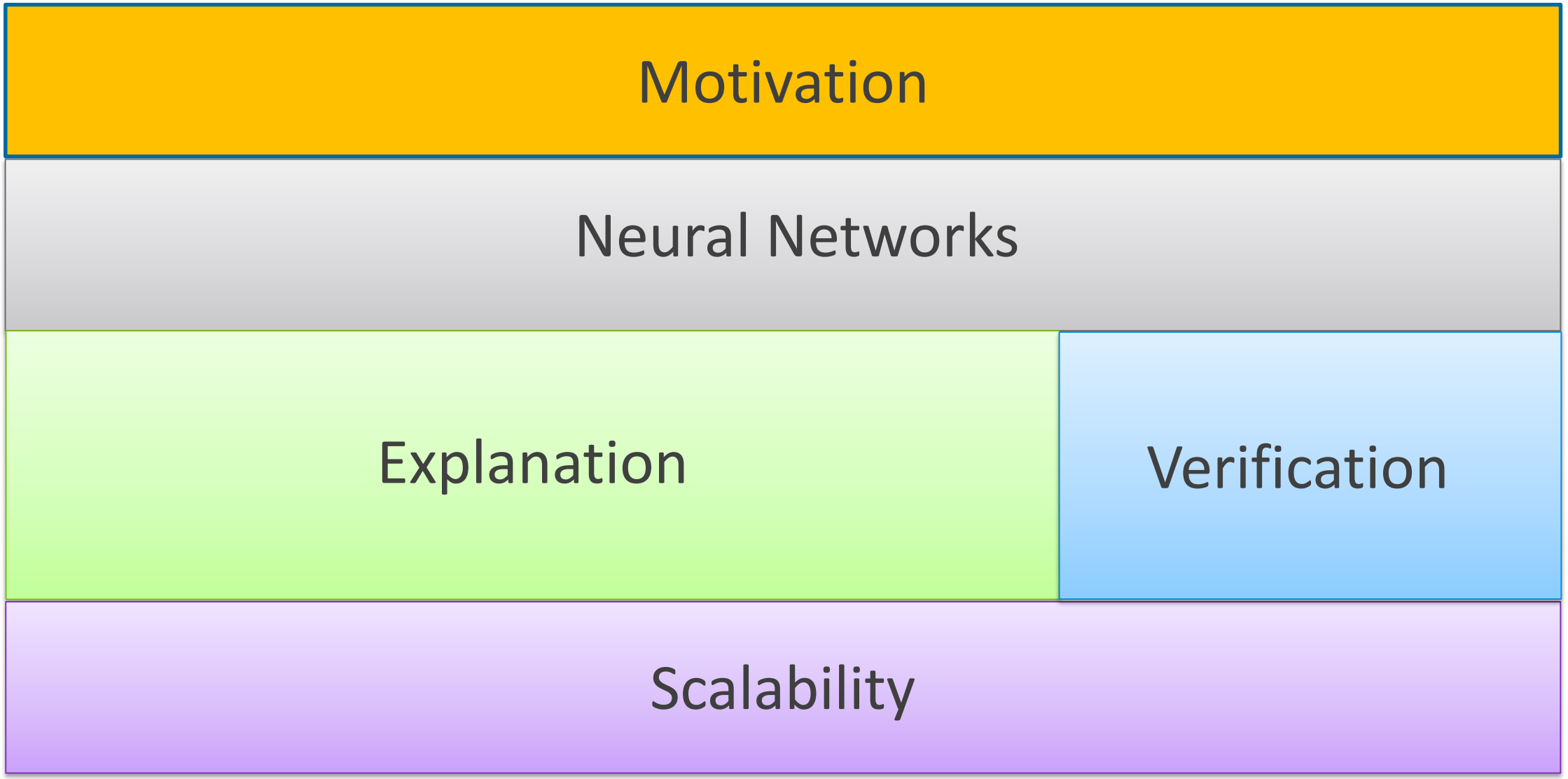
Verification and Explanation of Deep Neural Networks

Nina Narodytska, VMware Research

Joint work with Alexey Ignatiev, Joao Marques-Silva, Aditya Shrotri, Kuldeep Meel

vmware®

Outline


Motivation
Neural Networks
Explanation
Verification
Scalability

Outline

Motivation

The age of ML

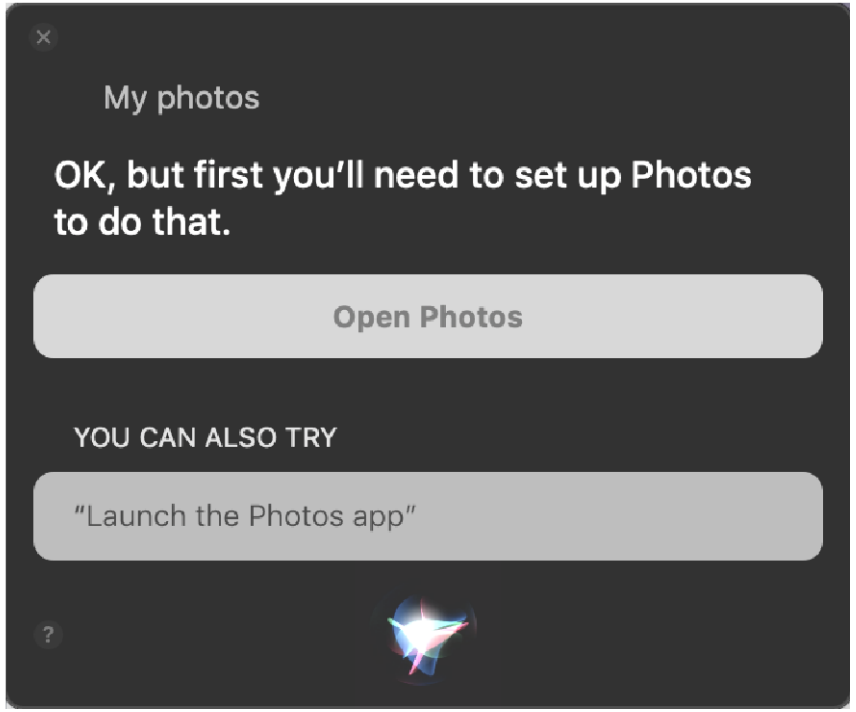
source: wikipedia

DeepMind

AlphaGo

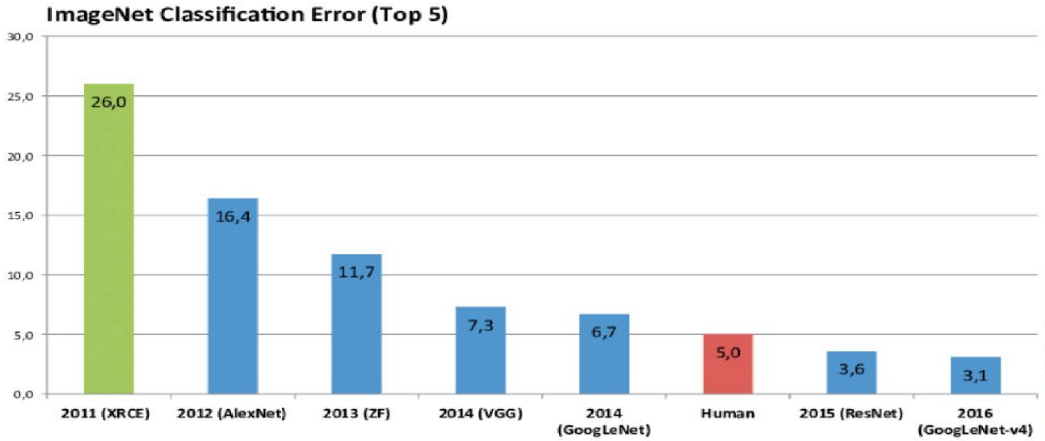
AlphaGo Zero & **Alpha Zero**

Image & Speech Recognition

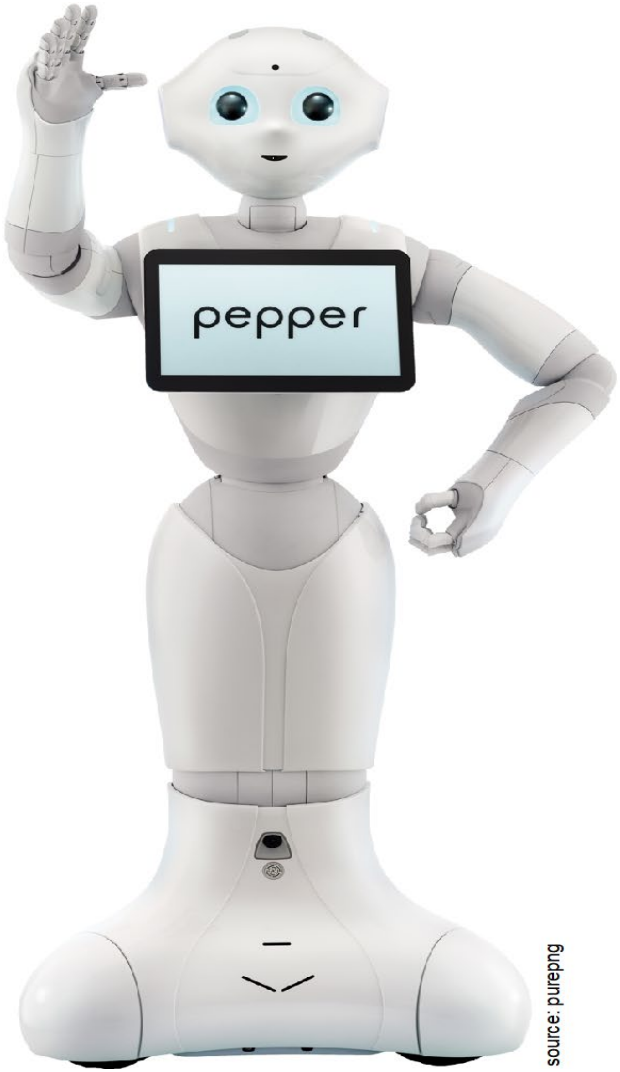


The age of ML

source: wikipedia



source: M. P. Kumar



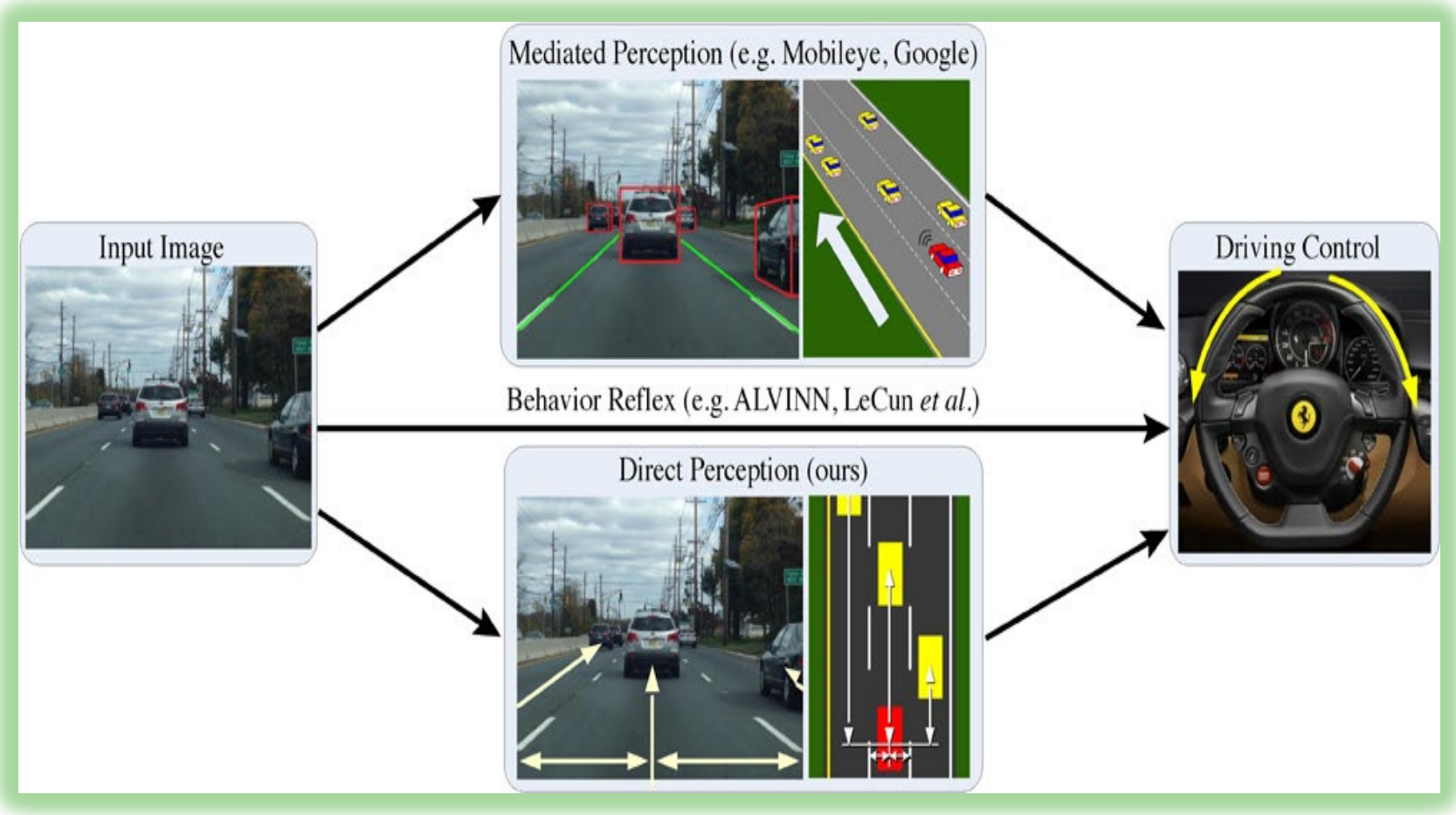
source: purepng

The age of ML

Audi AI:ME, concept level 4 autonomous car, April 2019

A **Level 4** autonomous car is one defined as a car that can completely drive itself, from start to finish, within a specifically-designated area.

The age of ML


Mediated Perception (e.g. Mobileye, Google)
Input Image
Driving Control
Behavior Reflex (e.g. ALVINN, LeCun et al.)
Direct Perception (ours)

The age of ML

"use" vs "not use"

Audi AI:ME, concept level 4 autonomous car, April 2019

The age of ML

Does ML-controller comply with a specification?

"use" vs "not use"

Audi AI:ME, concept level 4 autonomous car, April 2019

The age of ML

Does ML-controller comply with a specification?

Why does ML controller drive a car into a wall?

“use” vs “not use”

Audi AI:ME, concept level 4 autonomous car, April 2019

The age of ML

Verification/robustness

Why does ML controller drive a car into a wall?

“use” vs “not use”

Audi AI:ME, concept level 4 autonomous car, April 2019

The age of ML

Verification/robustness

Explainability/interpretability

"use" vs "not use"

Audi AI:ME, concept level 4 autonomous car, April 2019

Robustness of ML models

Interpretation of ML models

Robustness of ML models

Interpretation of ML models

Robustness of ML: misclassification

Original image

88% tabby cat

Robustness of ML: misclassification

Original image + Perturbation =

88% tabby cat

Robustness of ML: misclassification

Original image + Perturbation = Perturbed image

88% tabby cat

Robustness of ML: misclassification

Original image + Perturbation = Perturbed image

88% tabby cat

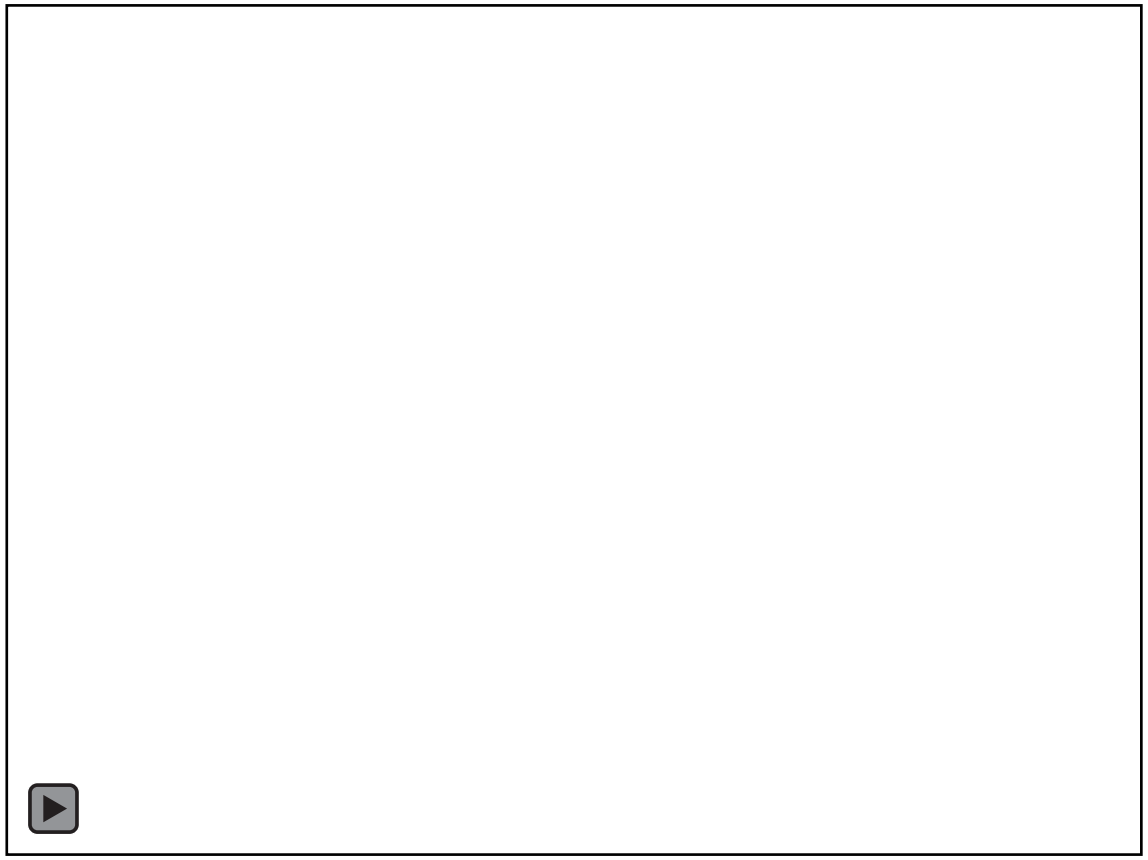
99% guacamole

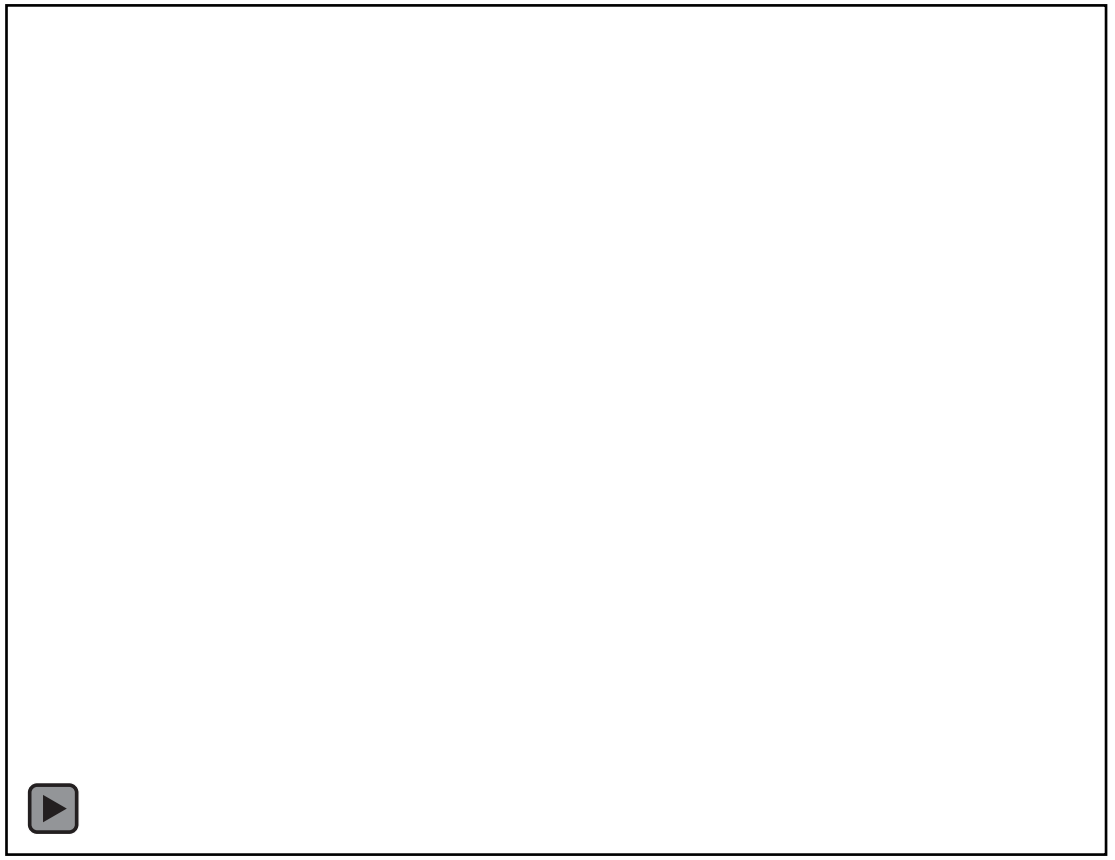
Robustness of ML: STOP sign attack

Eykholt et al'18

Aung et al'17

Robustness of ML: Autonomous car attack

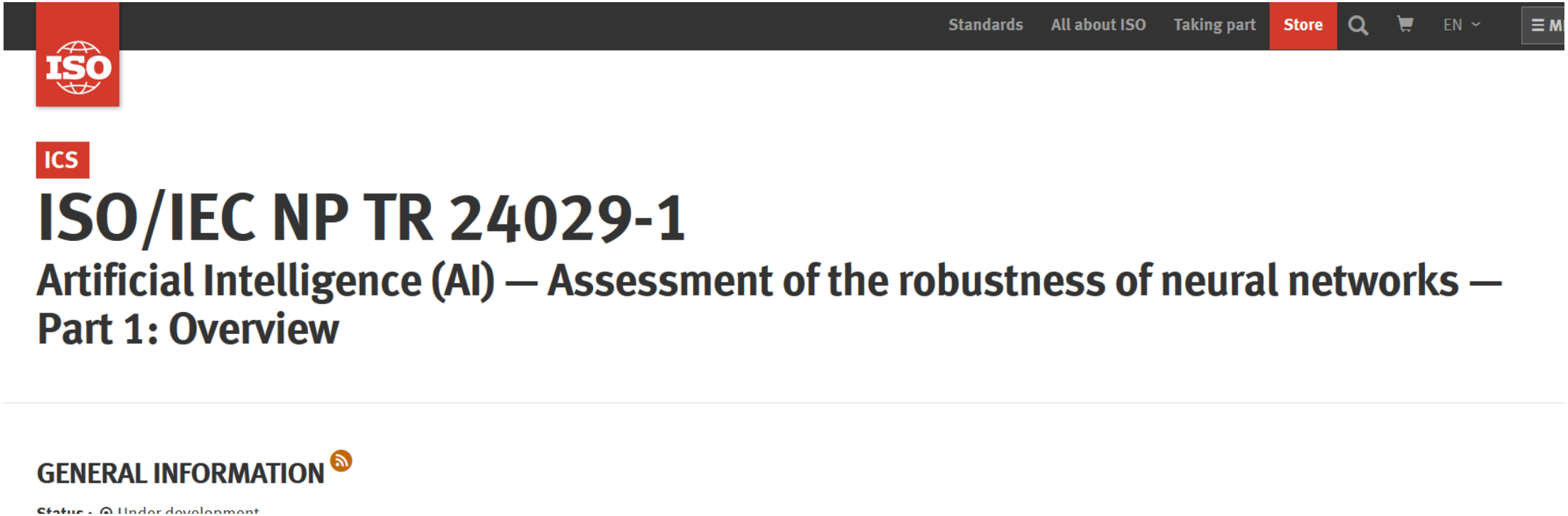




Adversarial machine learning,
Y. Vorobeychik, B. Li

Standardization: ISO

Standardization: ISO

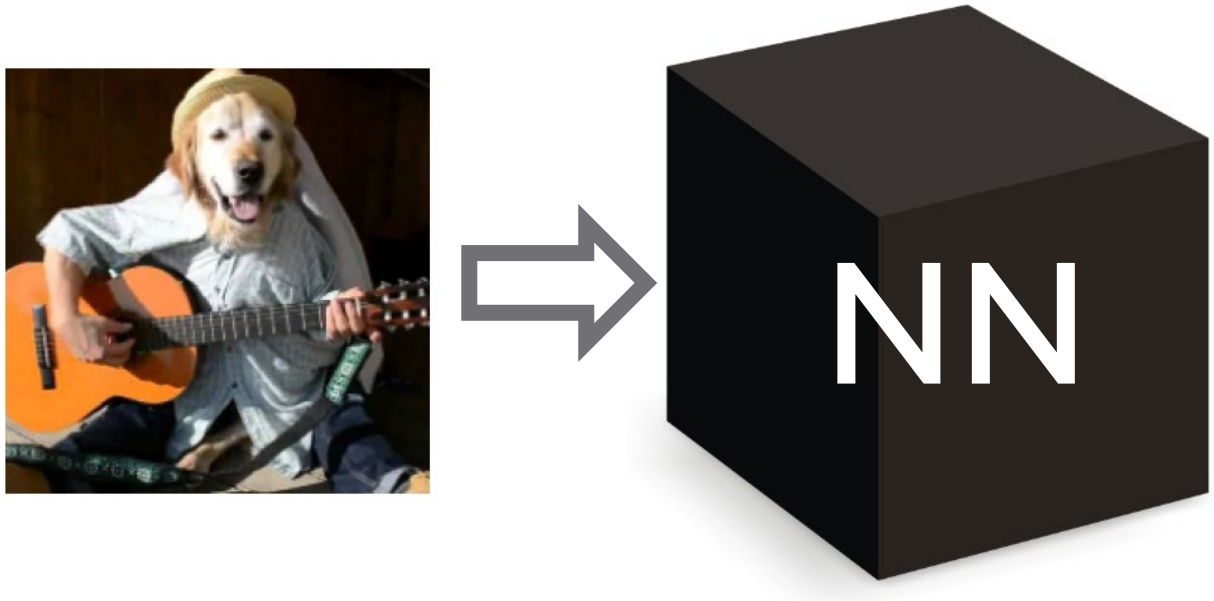


Robustness of ML models

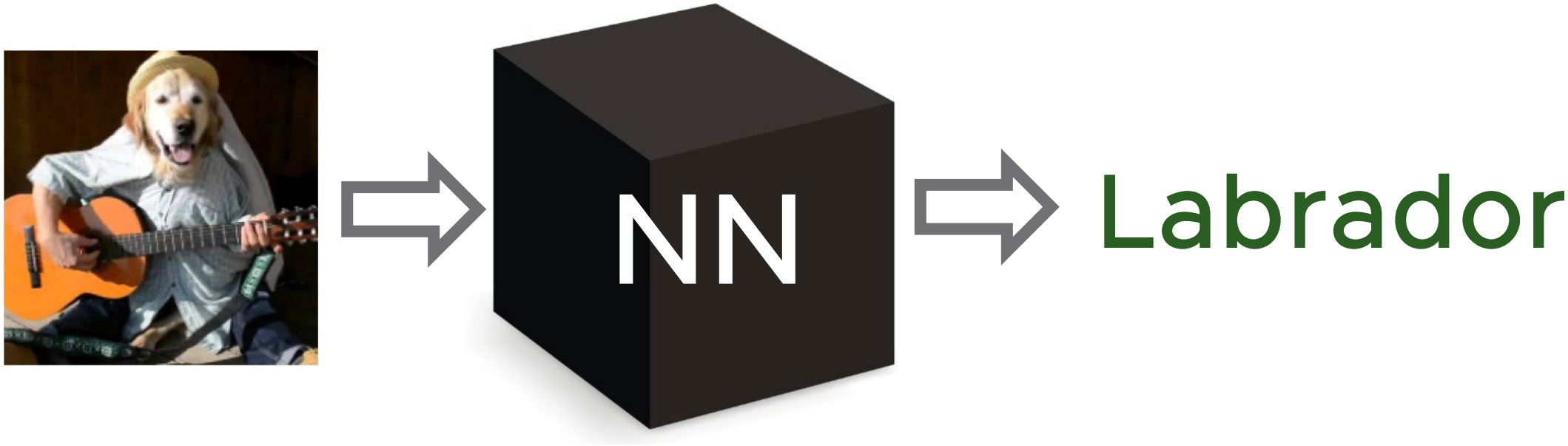
Interpretation of ML models

Interpretation of ML models

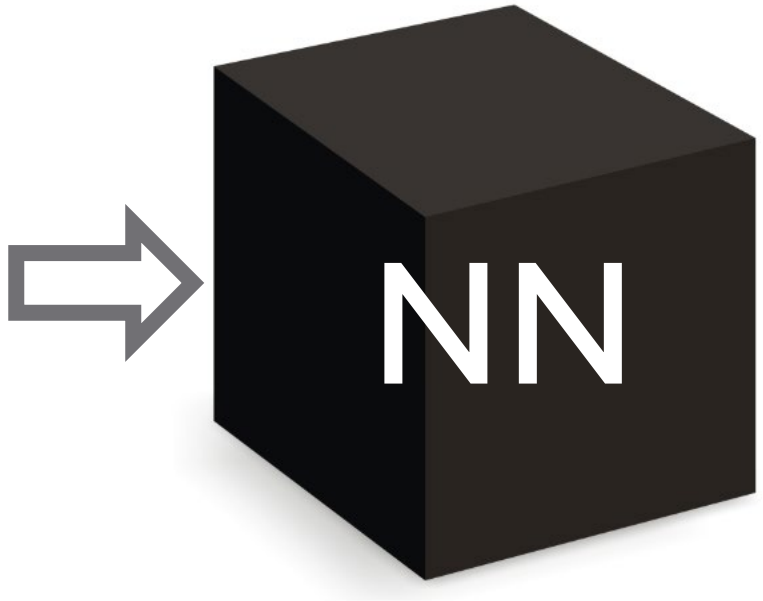
Interpretation of ML models



Interpretation of ML models



Interpretation of ML models



Labrador

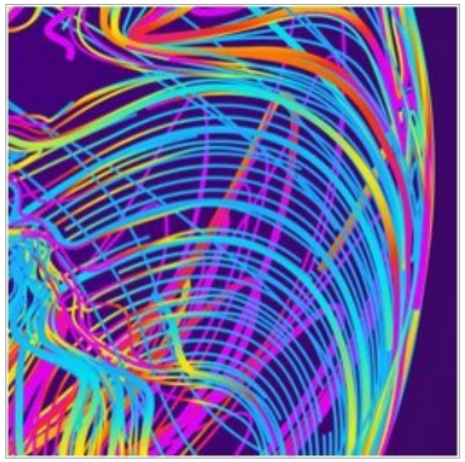
WHY?

Interpretation of ML models

The Challenge of Crafting Intelligible Intelligence

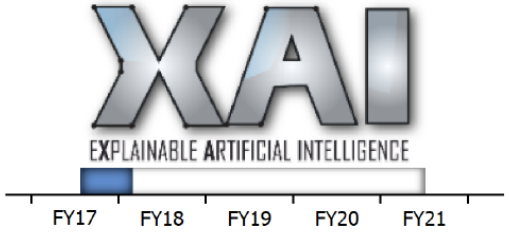
By Daniel S. Weld, Gagan Bansal
Communications of the ACM, June 2019, Vol. 62 No. 6, Pages 70-79
10.1145/3282486
Comments (1)

VIEW AS: SHARE:



Artificial Intelligence (ai) systems have reached or exceeded human performance for many circumscribed tasks. As a result, they are increasingly deployed in mission-critical roles, such as credit scoring, predicting if a bail candidate will commit another crime, selecting the news we read on social networks, and self-driving cars. Unlike other mission-critical software, extraordinarily complex AI systems are difficult to test: AI decisions are context specific and often based on thousands or millions of factors. Typically, AI behaviors are generated by searching vast action spaces or learned by the opaque optimization of mammoth neural networks operating over prodigious amounts of training data. Almost by definition, no clear-cut method can accomplish these AI tasks.

Explainable Artificial Intelligence (XAI)



David Gunning
DARPA/I2O
Program Update November 2017

DARPA

Standardization: GDPR

Standardization: GDPR

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.

Verification and explanation of ML models are important topics!

Verification and explanation of ML models are important topics!

Robustness of ML models

Verification and explanation of ML models are important topics!

Robustness of ML models

Verification and explanation of ML models are important topics!

Robustness of ML models

Interpretation of ML models

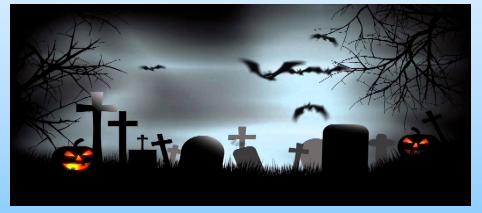
Verification and explanation of ML models are important topics!

Robustness of ML models

Interpretation of ML models

Verification and explanation of ML models are important topics!

Robustness of ML models

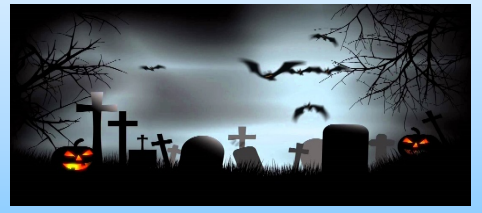


?

Interpretation of ML models

Verification and explanation of ML models are important topics!

Robustness of ML models



?

I.Goodfellow at AAAI 2019,
“Adversarial Machine Learning”

Interpretation of ML models

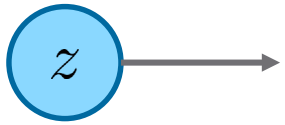
Outline

Motivation

Neural Networks

Neural Networks

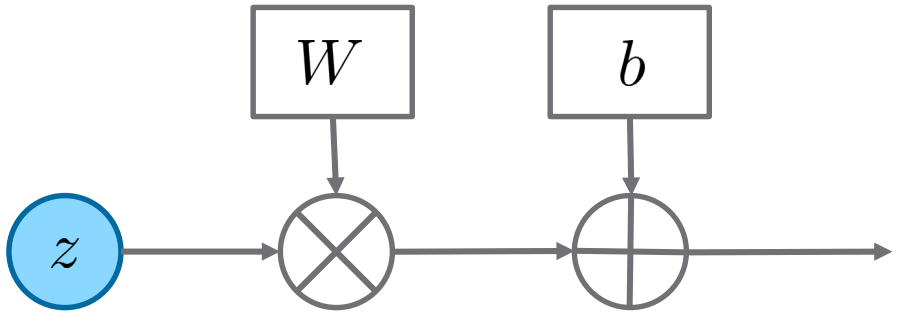
Neural Networks



Input

- features
- images

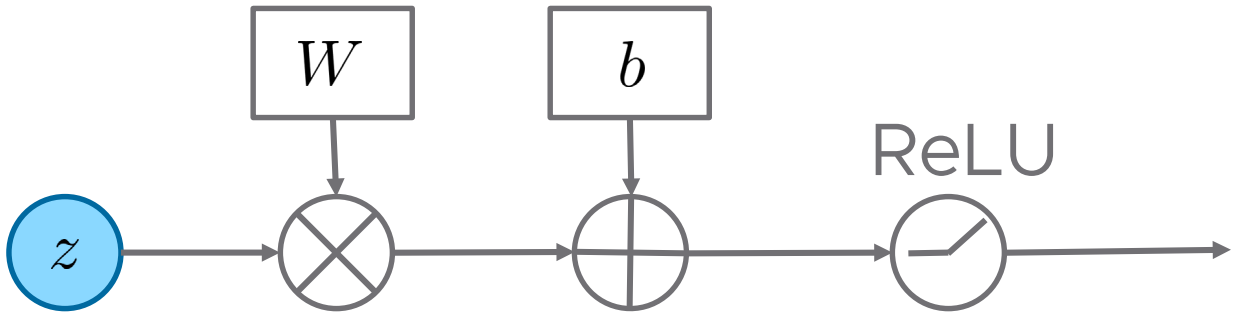
Linear transformation



Input

- features
- images

Non-Linear transformation

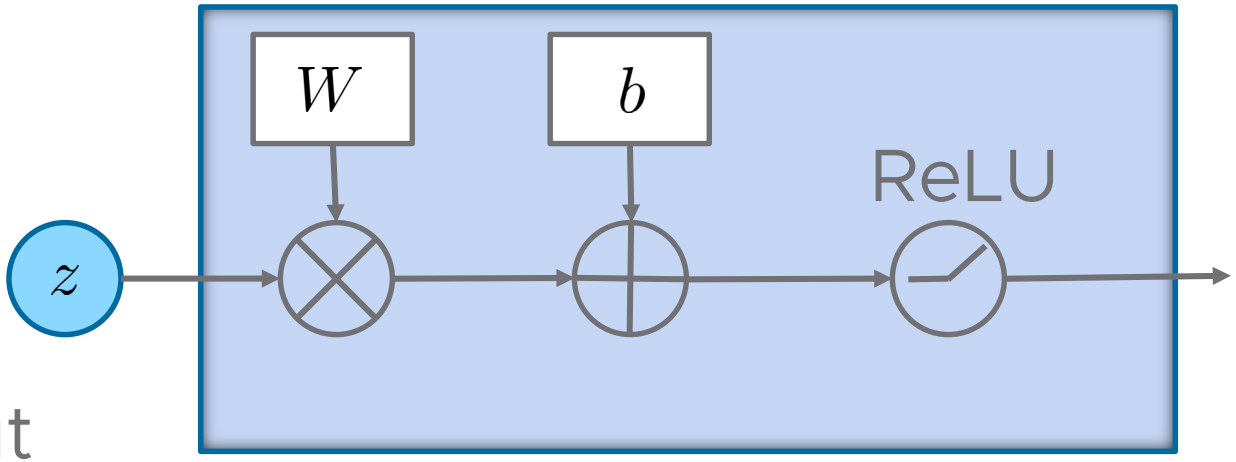


$$max(0, Wz + b)$$

Input

- features
- images

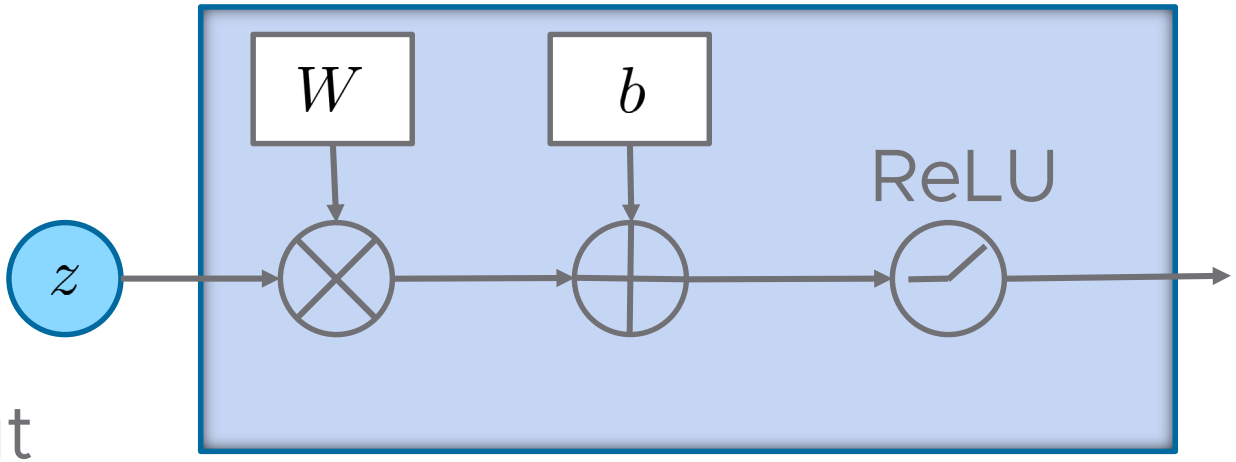
Non-Linear transformation



Input

- features
- images

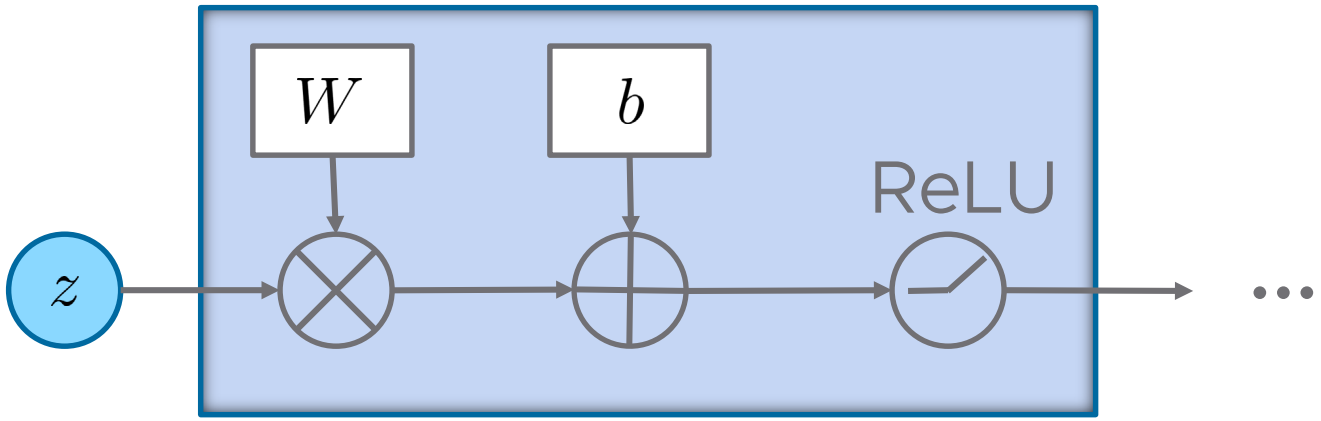
Block structure



Input

- features
- images

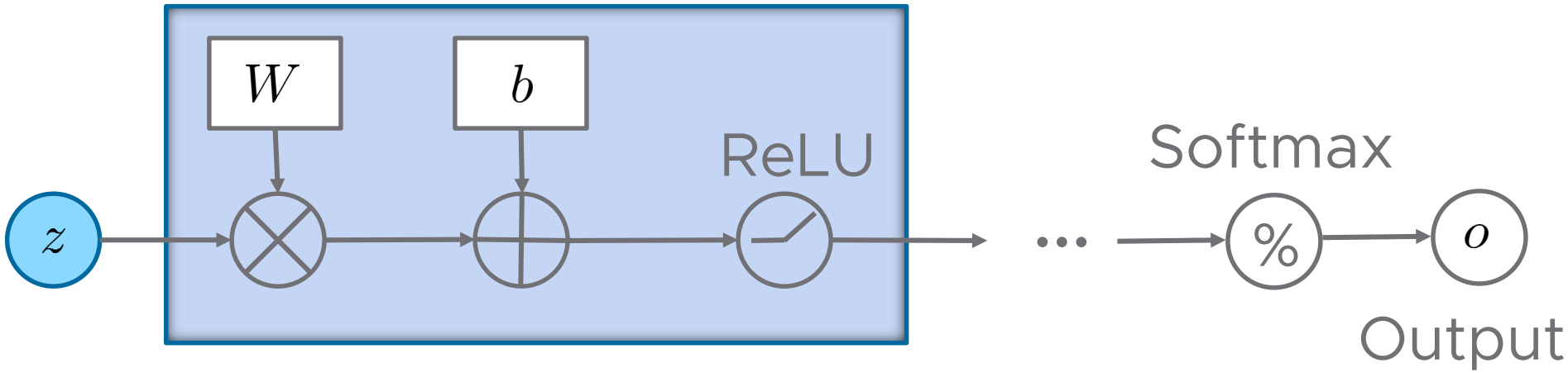
Repeated block structure



Input

- features
- images

Repeated block structure



Input

- features
- images

$$\frac{e^{a_i}}{\sum_{j=1}^{k} e^{a_j}}, i \in \{1, \ldots, k\}$$

Outline

Motivation

Neural Networks

Explanation

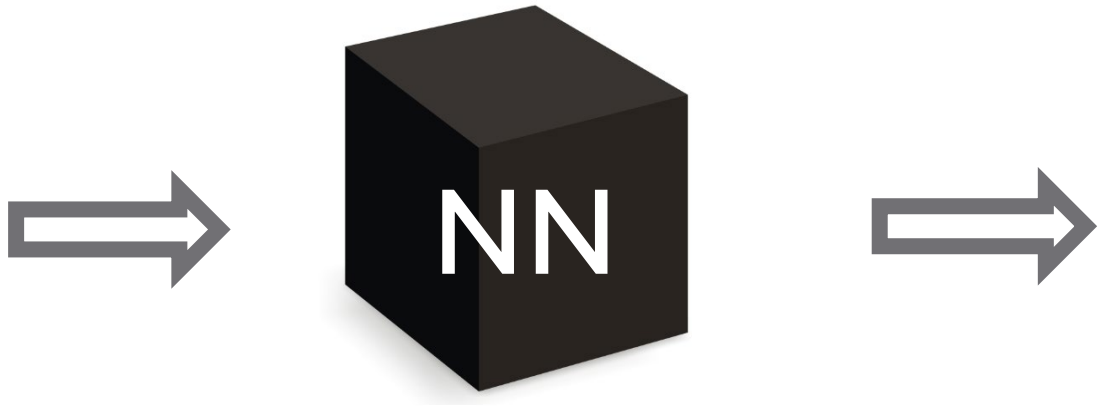
Explanations

Interpretable Explanations (Discussion)

Explanations

Labrador

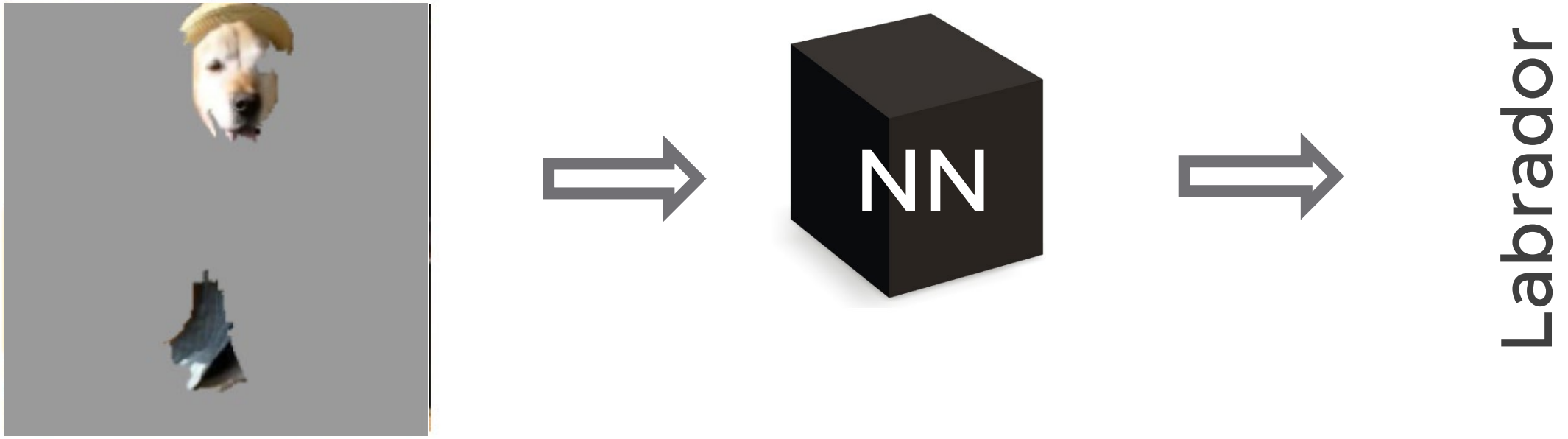
Explanations



Labrador

How would you explain this decision?

Explanations



How would you explain this decision?

Explanations

From: salem@pangea.Stanford.EDU (Bruce Salem)
Subject: Re: Science and theories

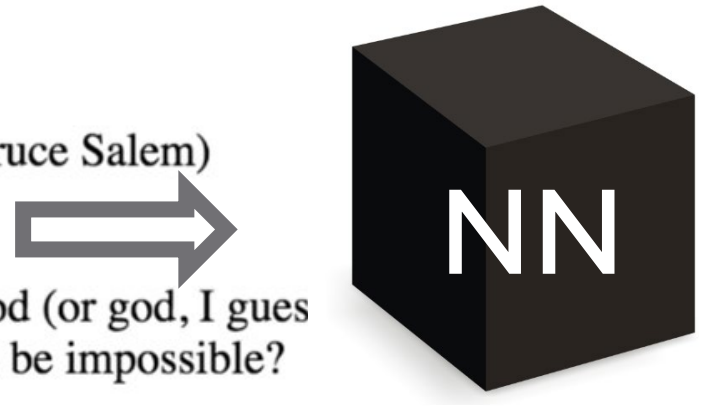
How is it possible for us to believe in God (or god, I gues
when science has shown his existence to be impossible?
I think atheism is the way to go forward.

Atheism

Explanations

From: salem@pangea.Stanford.EDU (Bruce Salem)
Subject: Re: Science and theories

How is it possible for us to believe in God (or god, I gues
when science has shown his existence to be impossible?
I think atheism is the way to go forward.



Atheism

How would you explain this decision?

Explanations

From: salem@pangea.Stanford.EDU (Bruce Salem)
Subject: Re: Science and theories

How is it possible for us to believe in God (or god, I guess)
when science has shown his existence to be impossible?
I think atheism is the way to go forward.

Atheism

Atheism Christianity

How would you explain this decision?

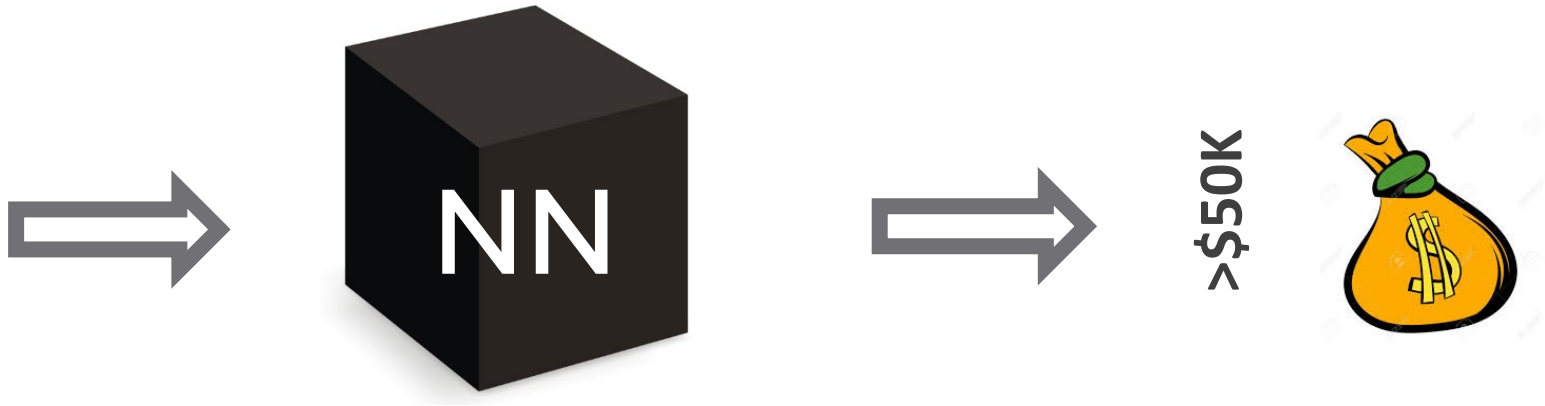
Explanations

Age is (37, 48],
White,
Male,
Married

>$50K

Explanations

Age is (37, 48],
White,
Male,
Married



How would you explain this decision?

Explanations

Age is (37, 48],
White,
Male,
Married

>$50K

How would you explain this decision?

Explanations

State-of-the-art methods

Workflow

Workflow

Workflow

(Age is (37, 48], White, Male, Married)

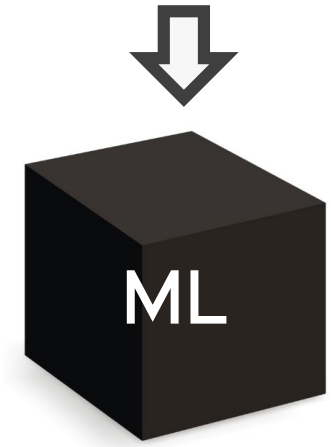
Workflow

(Age is (37, 48], White, Male, Married)

>$50K

Workflow

(Age is (37, 48], White, Male, Married)

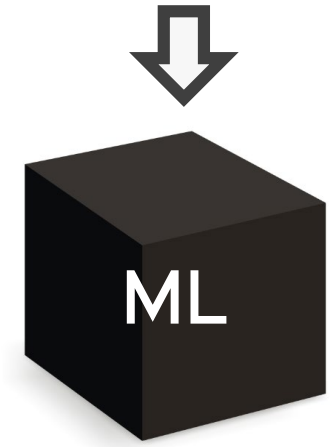


>$50K

Explainer

Workflow

(Age is (37, 48], White, Male, Married)



>$50K

Explainer

White, Male

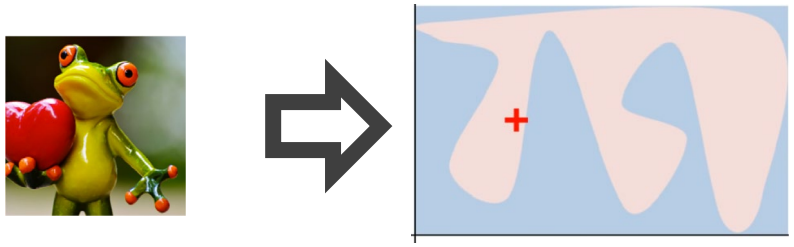
State-of-the-art

Heuristic approaches (e.g. LIME)

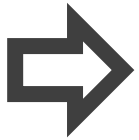
Why Should I Trust You?" Explaining the Predictions of Any Classifier, Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, KDD'16

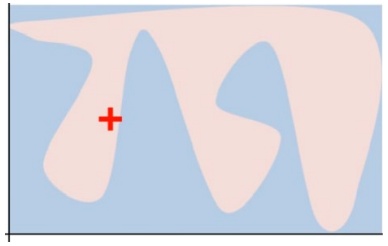
LIME: build a local neighborhood

LIME: build a local neighborhood

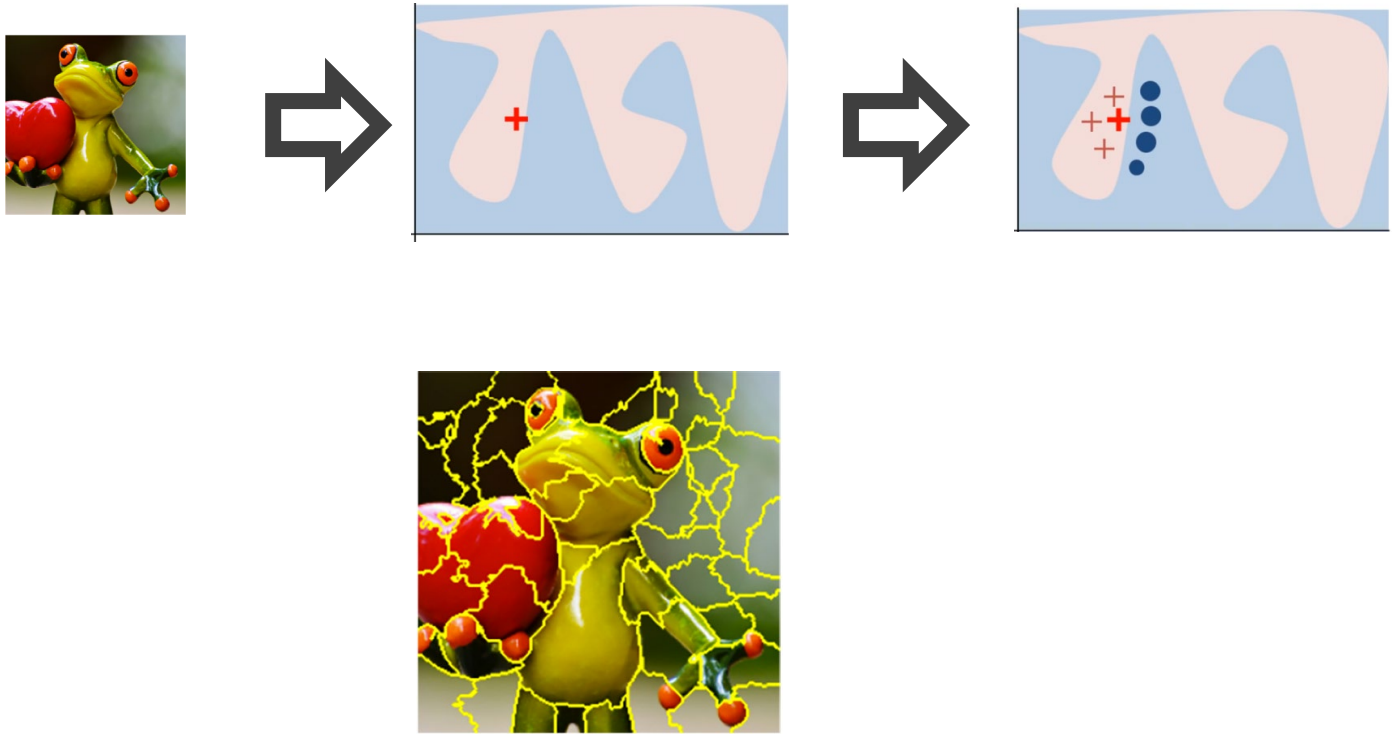


LIME: build a local neighborhood

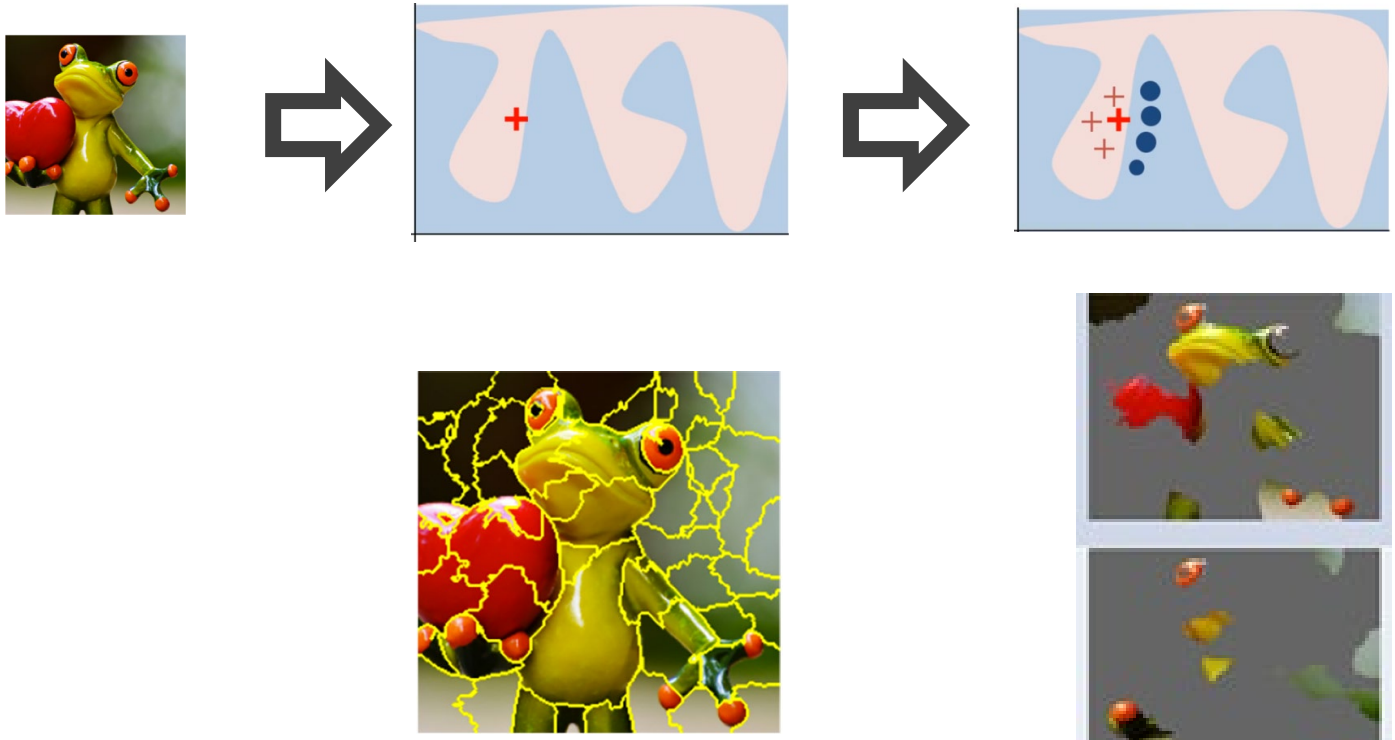




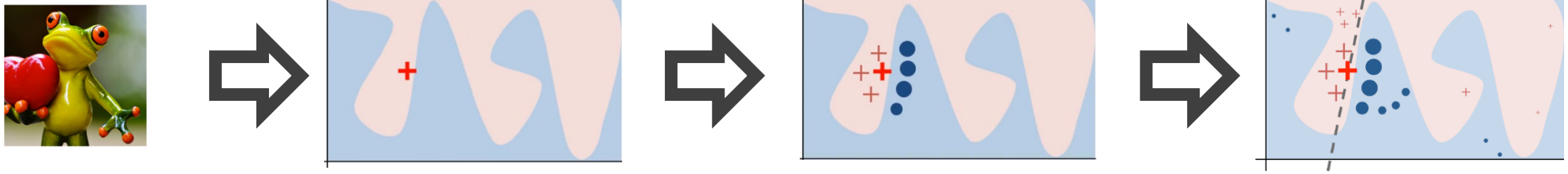
LIME: build a local neighborhood

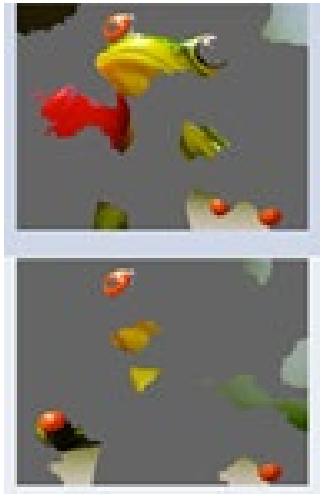


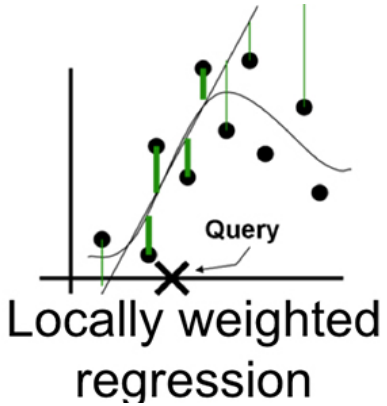
LIME: build a local neighborhood



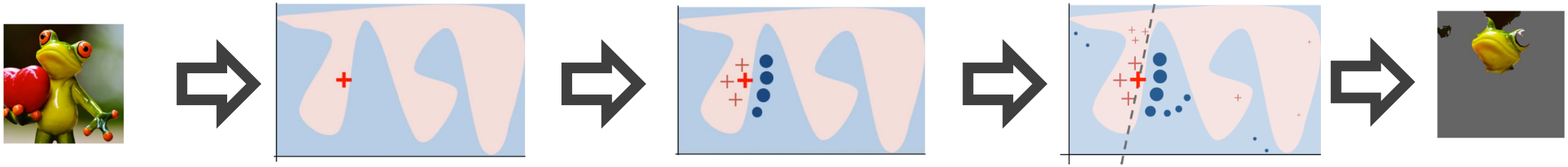
LIME: build a local neighborhood

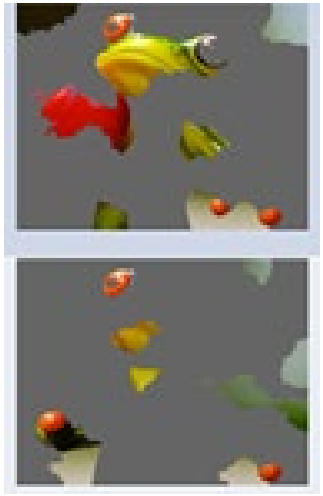


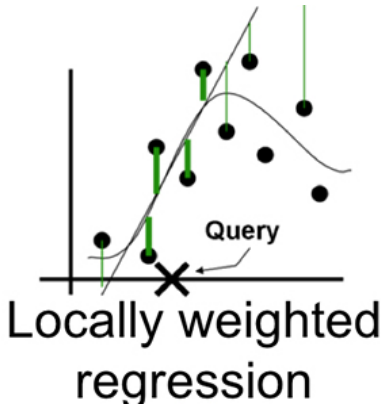




LIME: build a local neighborhood





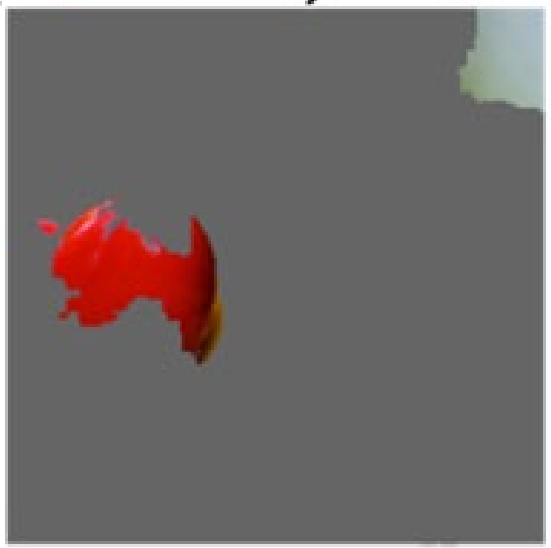


LIME: build a local neighborhood

LIME: build a local neighborhood

P() = 0.07

P() = 0.05



Heuristic approaches

1. local explanations

Heuristic approaches

1. local explanations
2. no guarantees about quality

Heuristic approaches

1. local explanations
2. no guarantees about quality
3. robustness of explanations

Explanations

Logic-based approach to explanations

Abduction-Based Explanations for Machine Learning Models, AAAI'19
Alexey Ignatiev, Nina Narodytska and Joao Marques-Silva

Andy Shih and Arthur Choi and Adnan Darwiche
A Symbolic Approach to Explaining Bayesian Network Classifiers, IJCAI'18
Compiling Bayesian Network Classifiers into Decision Graphs, AAAI'19

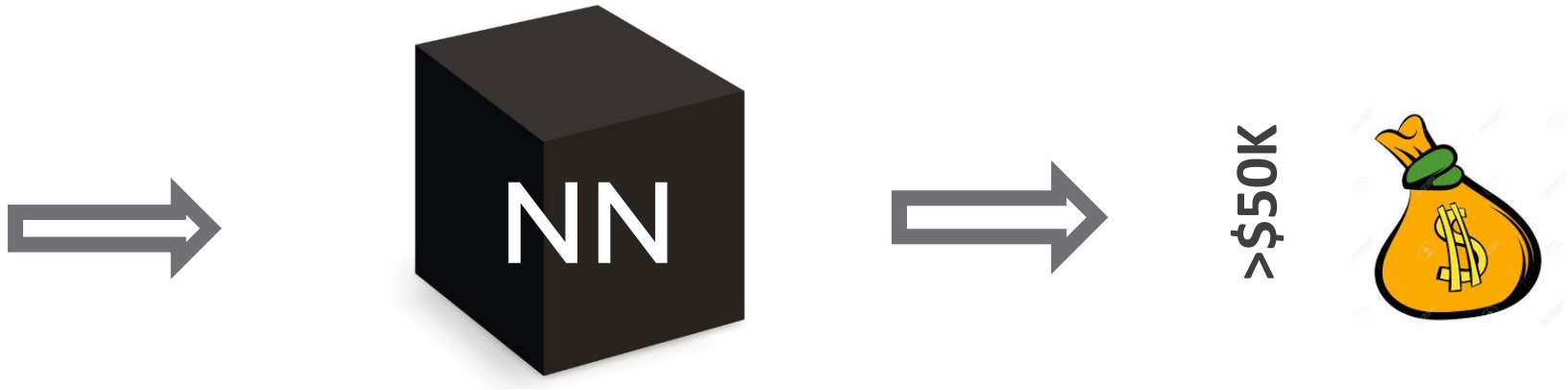
Propositional abduction problem

The main idea comes from work on model diagnosis to explain failures of the systems from a given set of hypotheses

R. Reiter. A theory of diagnosis from first principles. Artif. Intell., 32

Propositional abduction problem

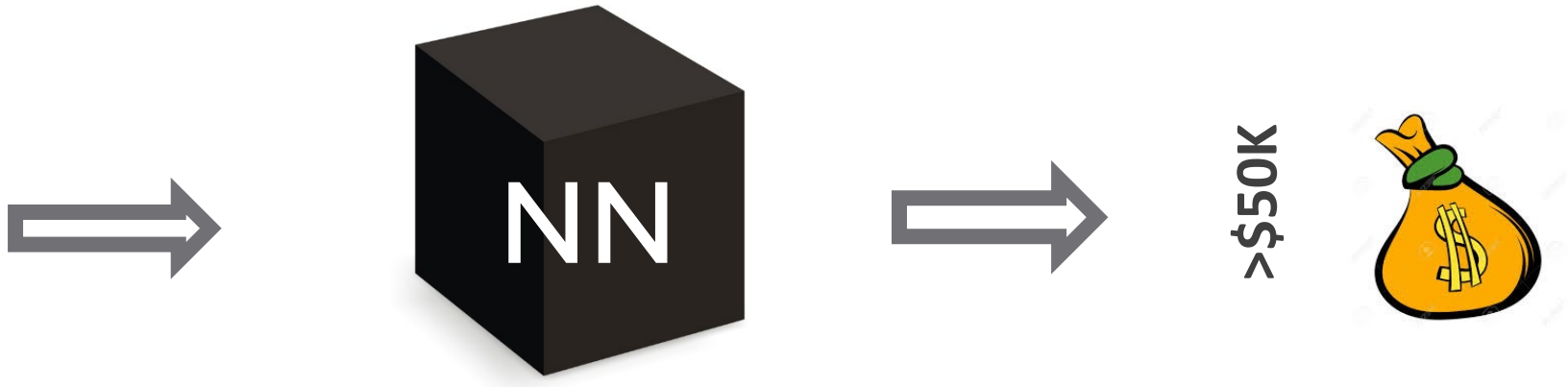
Age is (37, 48],
White,
Male,
Married



Propositional abduction problem

Age is (37, 48],
White,
Male,
Married

$\text{cube } Z$



Propositional abduction problem

Age is (37, 48],
White,
Male,
Married

$\text{cube } Z$

$\text{model } \mathbf{M}$

>$50K

Propositional abduction problem

Age is (37, 48],
White,
Male,
Married

cube Z

model $\mathbf{M}$

prediction p

Propositional abduction problem

Given a classifier $\mathbf{M}$, represented by some logic encoding, a cube Z and a prediction p, compute a subset-minimal $\mathcal{Z} \subseteq Z$ s.t.

$$1.(\mathcal{Z} \wedge \mathbf{M} \not\models \bot)$$

$$2.(\mathcal{Z} \wedge \mathbf{M} \models p)$$

Propositional abduction problem

Given a classifier $\mathbf{M}$, represented by some logic encoding, a cube Z and a prediction p, compute a subset-minimal $\mathcal{Z} \subseteq Z$ s.t.

$$1.(\mathcal{Z} \wedge \mathbf{M} \not\models \bot)$$

$$2.(\mathcal{Z} \wedge \mathbf{M} \models p)$$

Formalization of NN

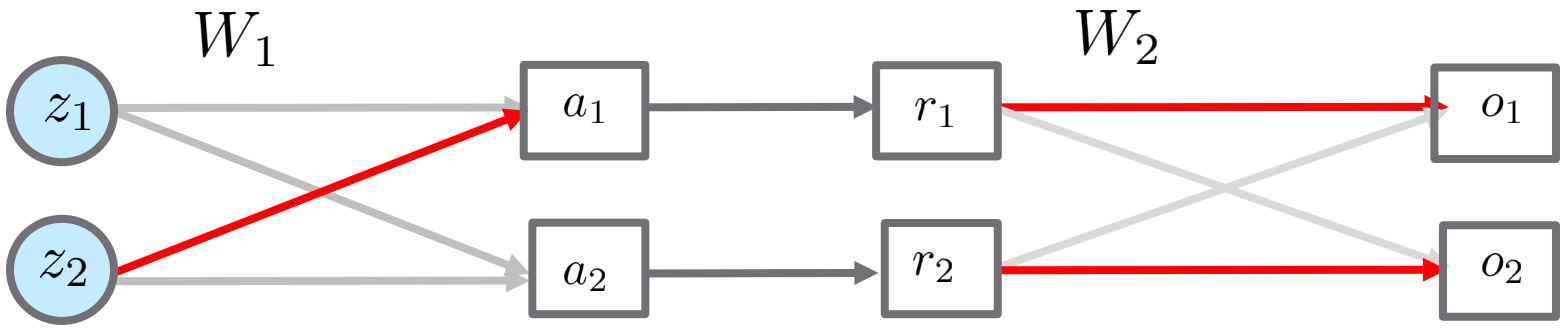
Formalization of NN

$$W_1 = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}$$

ReLU

$max(0, a_i)$

$$W_2 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$



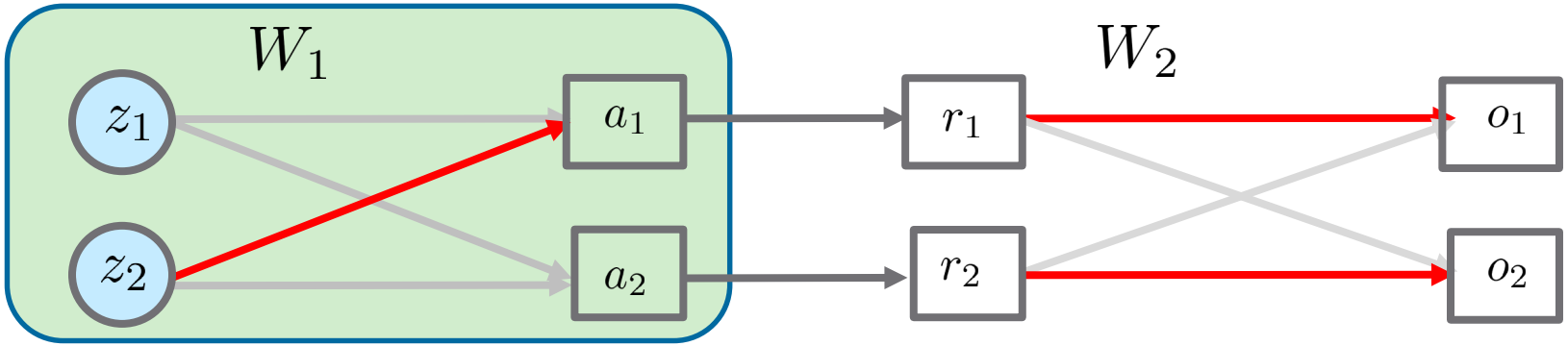
Formalization of NN

$$W_1 = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}$$

ReLU

$max(0, a_i)$

$$W_2 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$



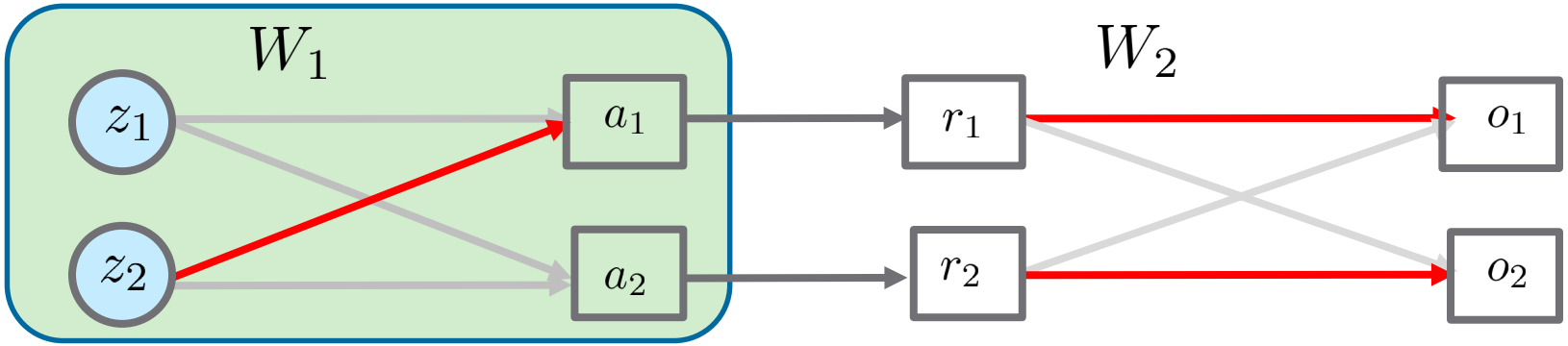
Formalization of NN

$$W_1 = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}$$

ReLU

$max(0, a_i)$

$$W_2 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$



$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

$$a_1 = -z_1 + z_2$$

$$a_2 = -z_1 - z_2$$

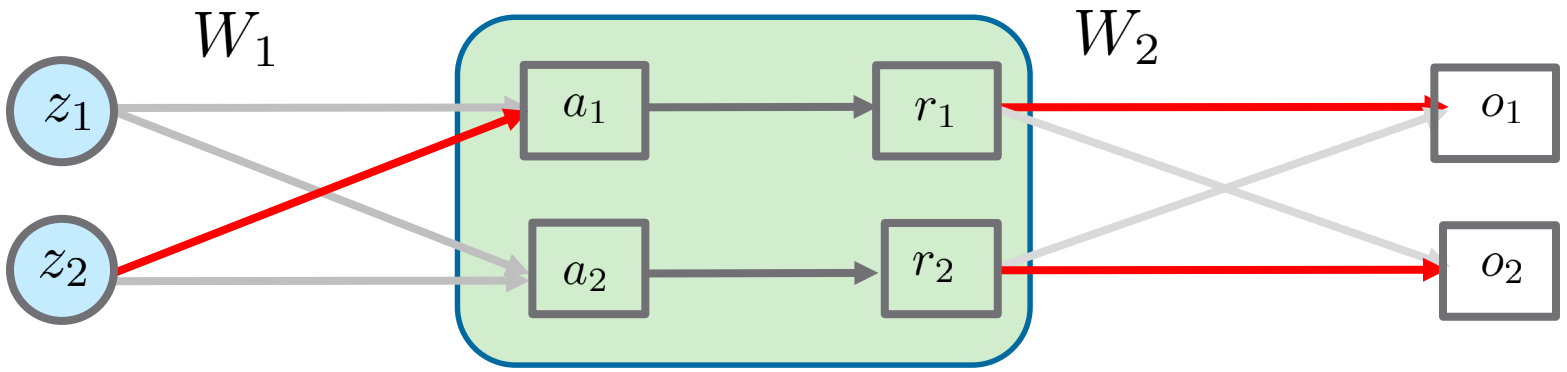
Formalization of NN

$$W_1 = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}$$

ReLU

$max(0, a_i)$

$$W_2 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$



$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

$$a_1 = -z_1 + z_2$$

$$a_2 = -z_1 - z_2$$

$$r_1 = max(0, a_1)$$

$$r_2 = max(0, a_2)$$

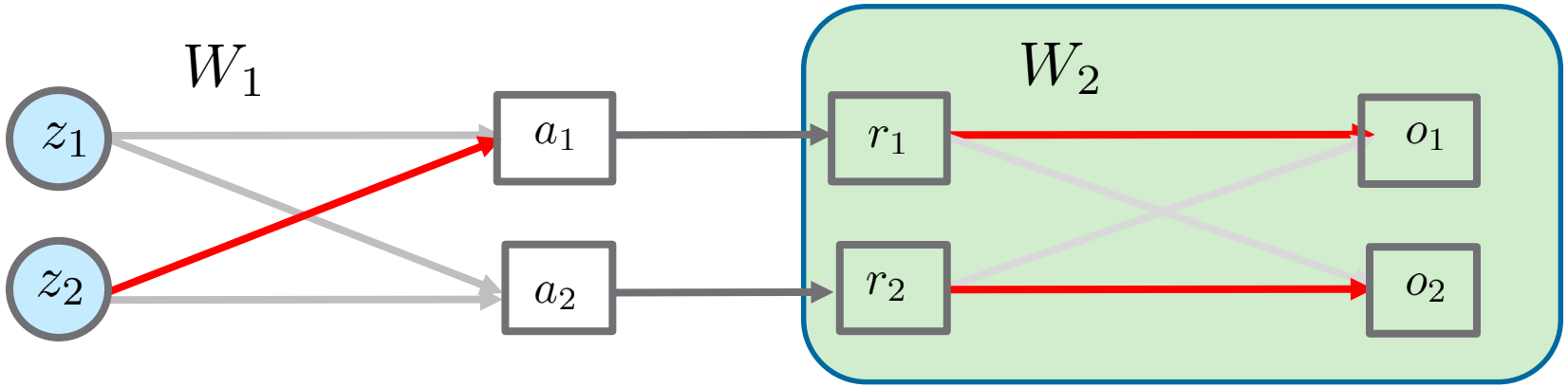
Formalization of NN

$$W_1 = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}$$

ReLU

$max(0, a_i)$

$$W_2 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$



$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

$$a_1 = -z_1 + z_2$$

$$a_2 = -z_1 - z_2$$

$$r_1 = max(0, a_1)$$

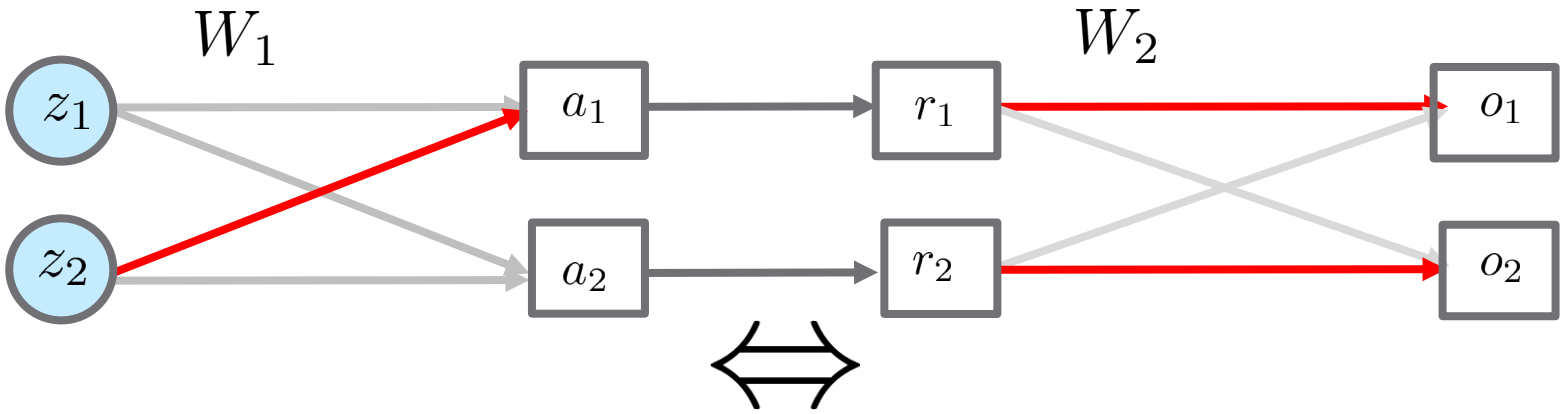
$$r_2 = max(0, a_2)$$

$$\begin{bmatrix} o_1 \\ o_2 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$$

$$o_1 = r_1 - r_2$$

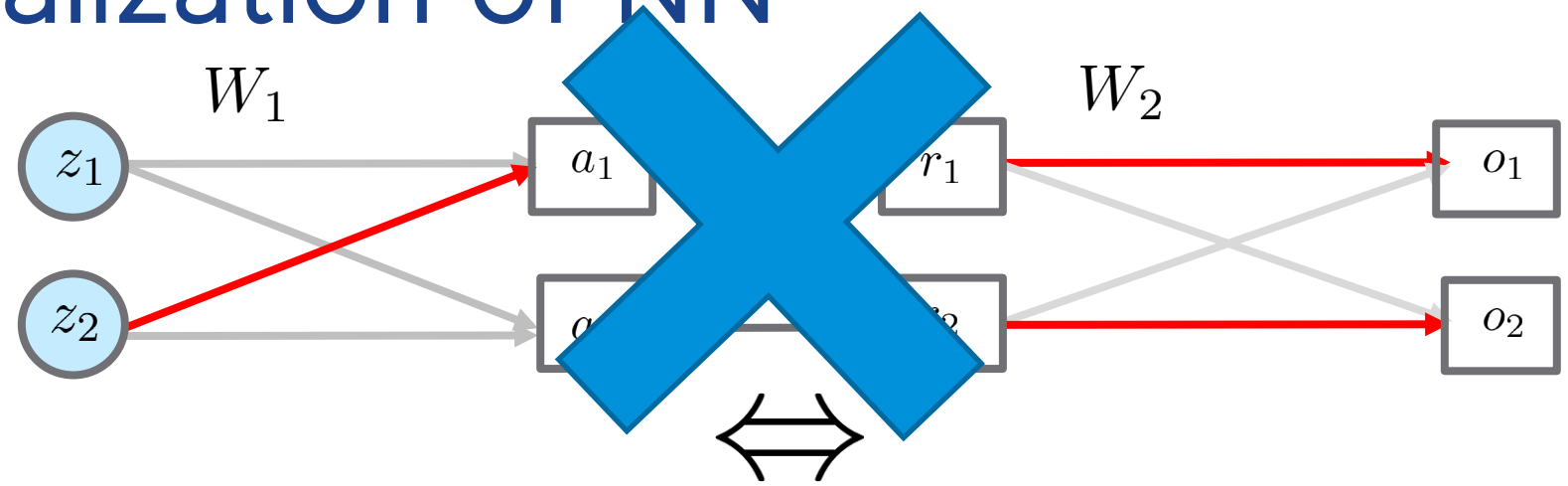
$$o_2 = -r_1 + r_2$$

Formalization of NN



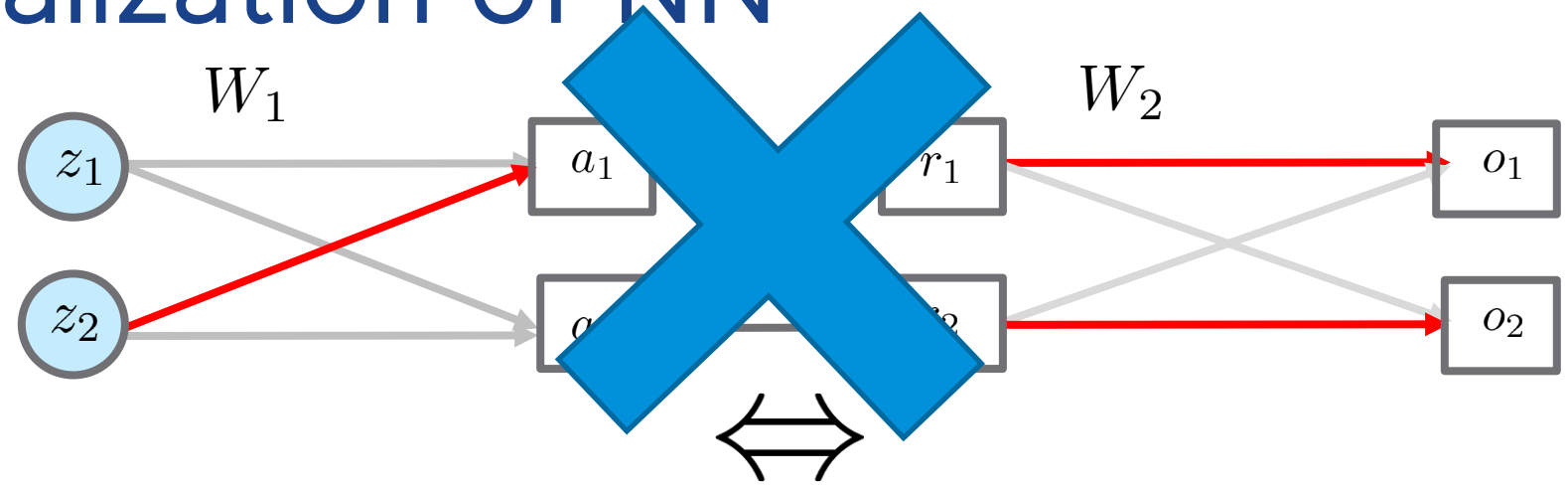
$$a_1 = -z_1 + z_2$$
$$a_2 = -z_1 - z_2$$
$$r_1 = max(0, a_1)$$
$$r_2 = max(0, a_2)$$
$$o_1 = r_1 - r_2$$
$$o_2 = -r_1 + r_2$$
$$-1 \leq z_1 \leq 1$$
$$-1 \leq z_2 \leq 1$$

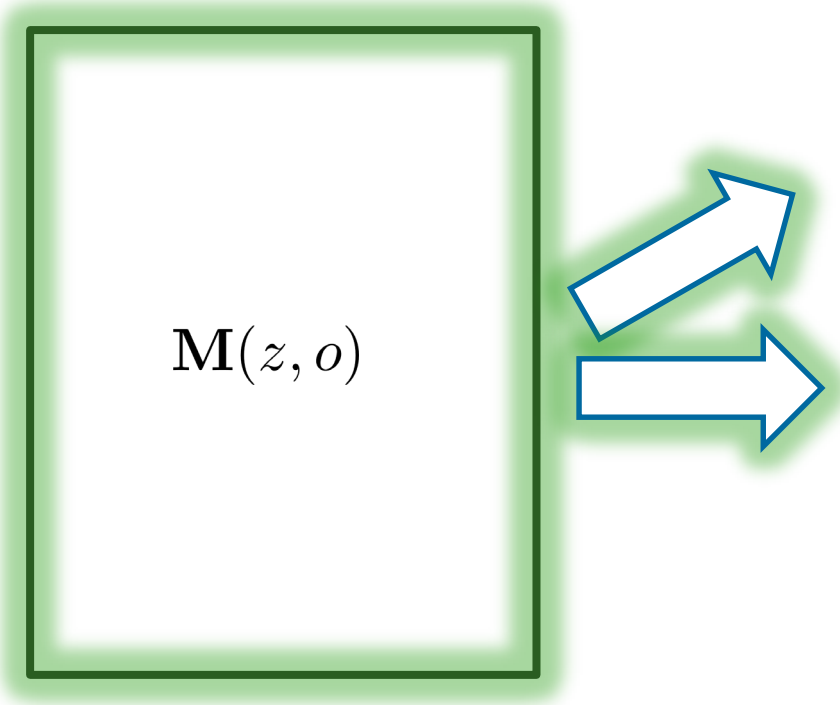
Formalization of NN



$$\mathbf{M}(z, o)$$

Formalization of NN





CP

MILP

Propositional abduction problem

Given a classifier $\mathbf{M}$, represented by some logic encoding, a cube Z and a prediction p, compute a subset-minimal $\mathcal{Z} \subseteq Z$ s.t.

$$1.(\mathcal{Z} \wedge \mathbf{M} \not\models \bot)$$

$$2.(\mathcal{Z} \wedge \mathbf{M} \models p)$$

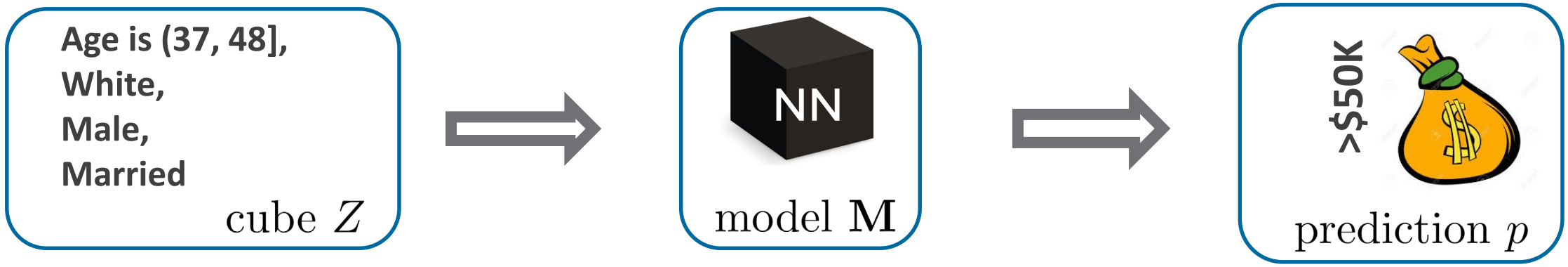
Propositional abduction problem

Given a classifier $\mathbf{M}$, represented by some logic encoding, a cube Z and a prediction p, compute a subset-minimal $\mathcal{Z} \subseteq Z$ s.t.

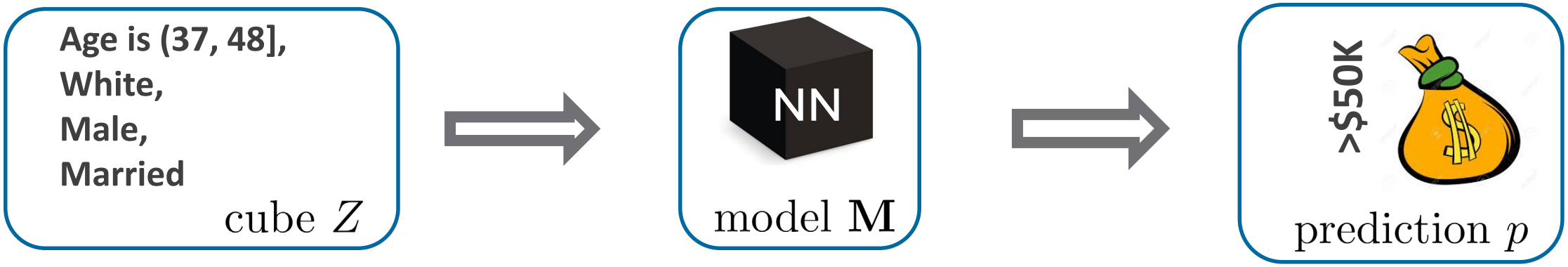
$$1.(\mathcal{Z} \wedge \mathbf{M} \not\models \bot)$$

$$2.(\mathcal{Z} \wedge \mathbf{M} \models p)$$

Propositional abduction problem

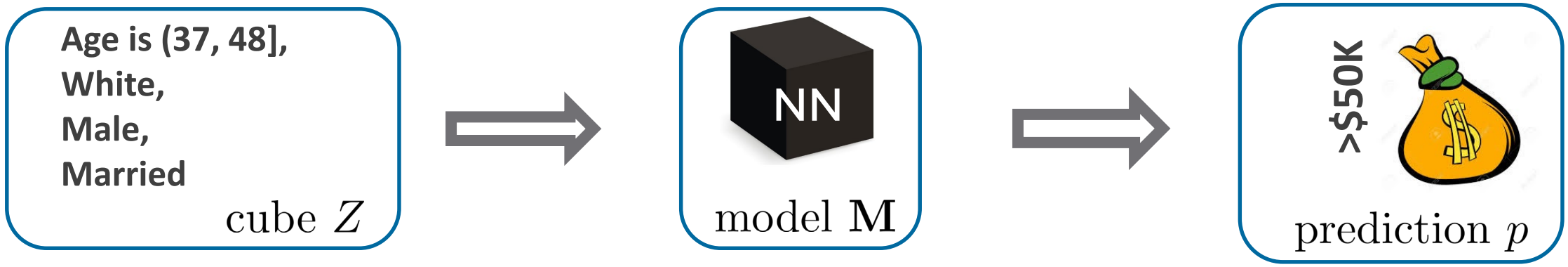


Propositional abduction problem



$$\mathcal{Z} \wedge \mathbf{M} \not\models \bot$$

Propositional abduction problem



$$Z \wedge \mathbf{M} \not\models \bot$$

By definition: $Z \wedge \mathbf{M} \models p$

Propositional abduction problem

Age is (37, 48],
White,
Male,
Married

cube Z

model $\mathbf{M}$

prediction p

$$Z \wedge \mathbf{M} \not\models \text{tautology}$$

By definition: $Z \wedge \mathbf{M} \models p$

Propositional abduction problem

Given a classifier $\mathbf{M}$, represented by some logic encoding, a cube Z and a prediction p, compute a subset-minimal $\mathcal{Z} \subseteq Z$ s.t.

$$1.(\mathcal{Z} \wedge \mathbf{M} \not\models \bot)$$

$$2.(\mathcal{Z} \wedge \mathbf{M} \models p)$$

Propositional abduction problem

Given a classifier $\mathbf{M}$, represented by some logic encoding, a cube Z and a prediction p, compute a subset-minimal $\mathcal{Z} \subseteq Z$ s.t.

$$1.(\mathcal{Z} \wedge \mathbf{M} \not\models \bot)$$

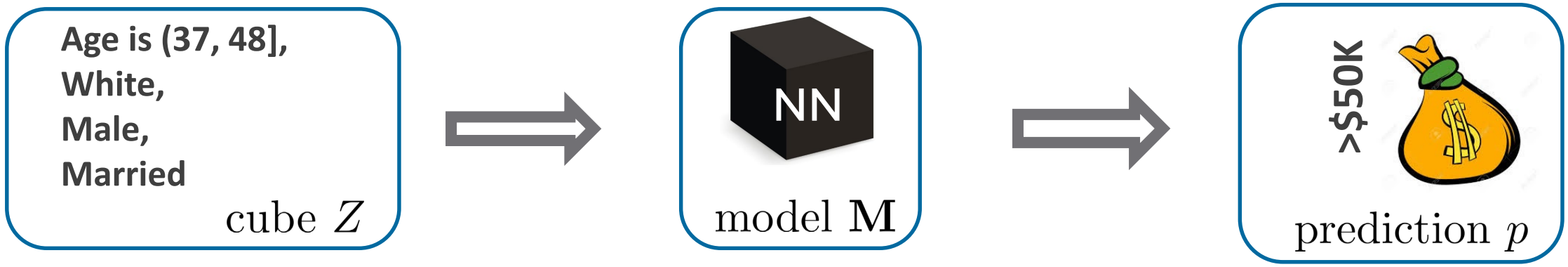
$$2.(\mathcal{Z} \wedge \mathbf{M} \models p)$$

Propositional abduction problem

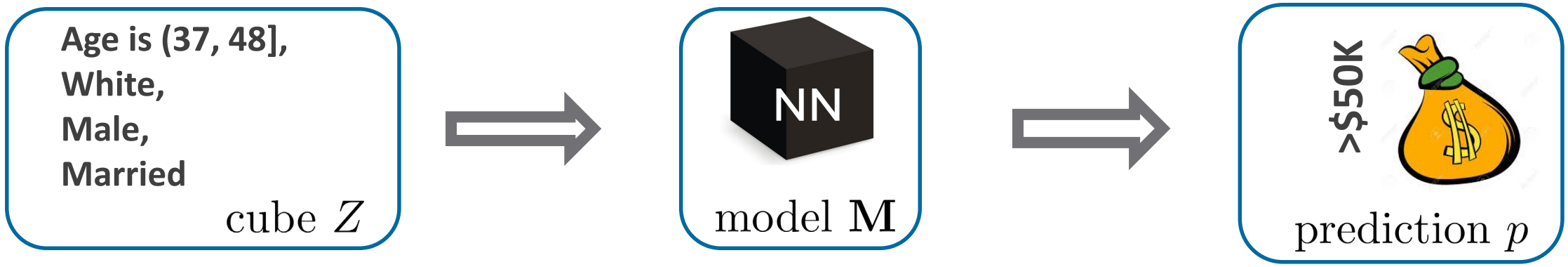
Given a classifier $\mathbf{M}$, represented by some logic encoding, a cube Z and a prediction p, compute a subset-minimal $\mathcal{Z} \subseteq Z$ s.t.

$$2.(\mathcal{Z} \wedge \mathbf{M} \models p)$$

Propositional abduction problem

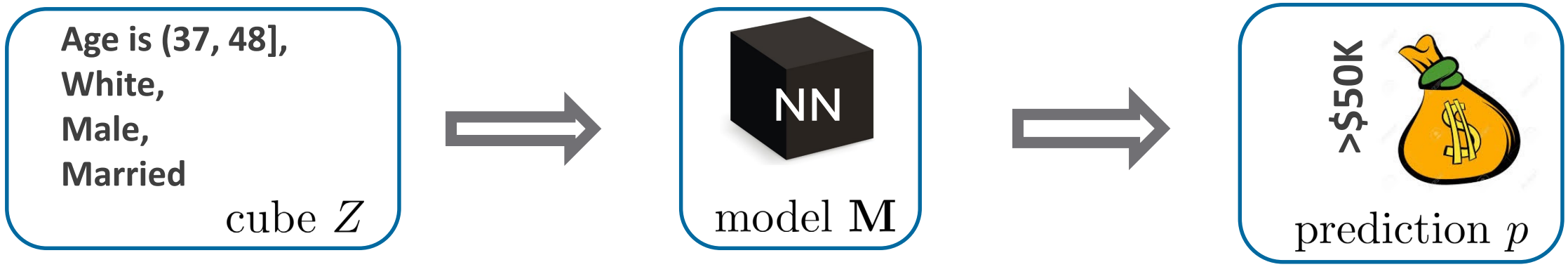


Propositional abduction problem



$$\mathcal{Z} \wedge \mathbf{M} \models p$$

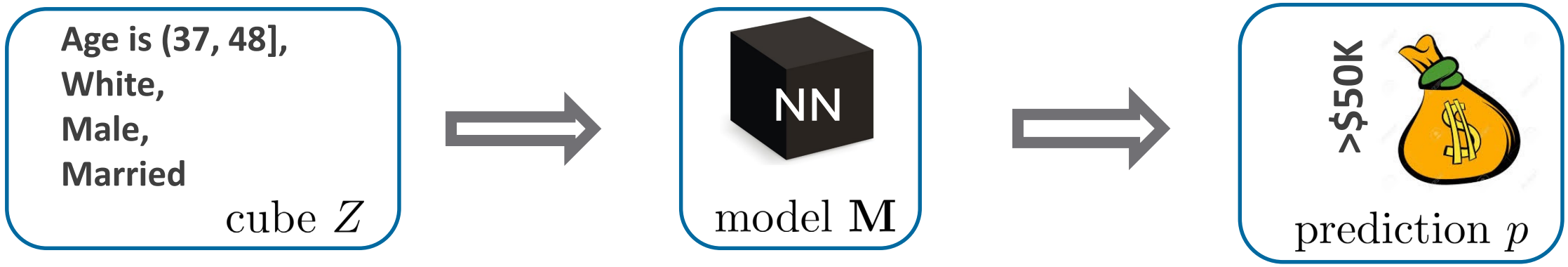
Propositional abduction problem



$$\mathcal{Z} \wedge \mathbf{M} \models p$$

$$\Leftrightarrow (\mathcal{Z} \models (\mathbf{M} \rightarrow p))$$

Propositional abduction problem



$$\mathcal{Z} \models (\mathbf{M} \rightarrow p)$$

Propositional abduction problem

$$\mathcal{Z} \models (\mathbf{M} \rightarrow p)$$

Propositional abduction problem

An **explanation** is a subset of input features so that changes to the rest of inputs do not affect the prediction.

Subset-minimal explanation

Input: M, initial cube Z, prediction p

```
1 begin
2     foreach l ∈ Z do
3         if Z \ {l} ⊨ M → p then
4             Z ← Z \ {l}
5     return Z
6 end
```

Algorithm 1: Computing a subset-minimal explanation

Subset-minimal explanation

Input: M, initial cube Z, prediction p

1 **begin**

2 **foreach** $l \in Z$ **do**

3 **if** $Z \setminus \{l\} \models \mathbf{M} \rightarrow p$ **then**

4 $Z \leftarrow Z \setminus \{l\}$

5 **return** Z

6 **end**

Algorithm 1: Computing a subset-minimal explanation

Logic-based approaches (Xplainer)

1. local explanations
2. no guarantees about quality
3. robustness of explanations

Logic-based approaches (Xplainer)

1. global explanations
2. no guarantees about quality
3. robustness of explanations

Logic-based approaches (Xplainer)

1. global explanations
2. provide guarantees about quality
3. robustness of explanations

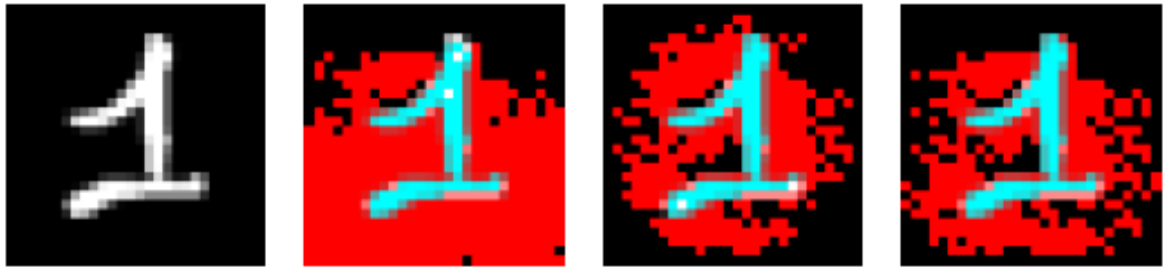
On Validating, Repairing and Refining Heuristic ML Explanations, CoRR report'19,
Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva

Logic-based approaches (Xplainer)

1. global explanations
2. provide guarantees about quality
3. evaluate robustness of explanations

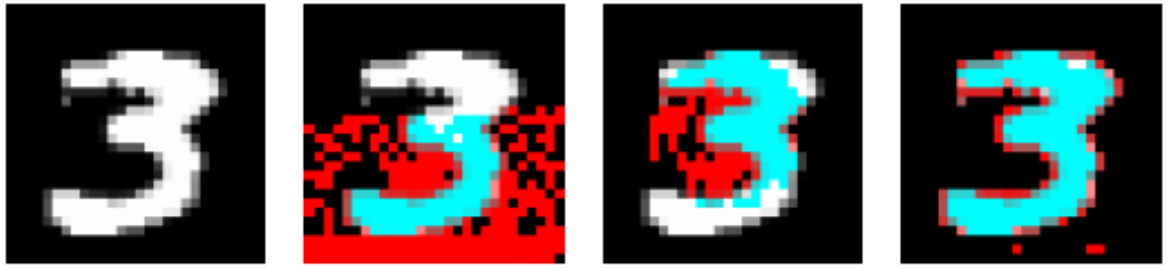
Assessing Heuristic Machine Learning Explanations with Model Counting, SAT'19
Nina Narodytska, Aditya A. Shrotri, Kuldeep S. Meel, Alexey Ignatiev, João Marques-Silva
(Used approximate model counting to evaluate quality of the explanations)

Xplainer: Subset-minimal explanation



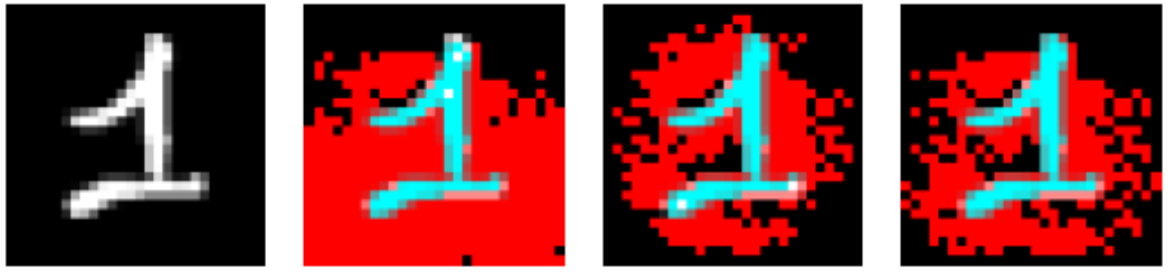
(a) (b) (c) (d)

Figure 1: Possible minimal explanations for digit one.



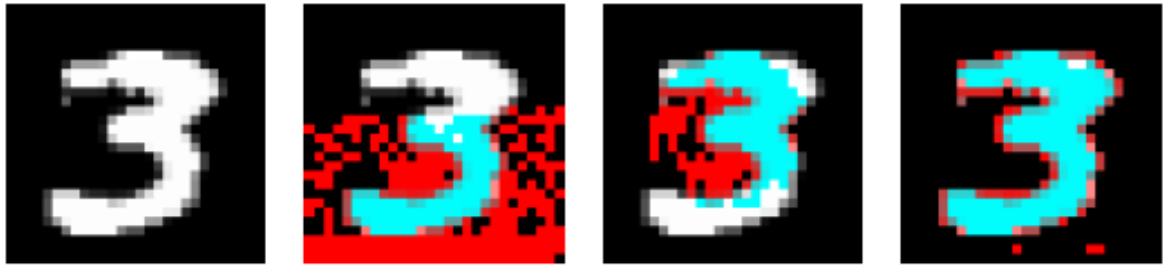
(a) (b) (c) (d)

Xplainer: Subset-minimal explanation



(a) (b) (c) (d)

Figure 1: Possible minimal explanations for digit one.



(a) (b) (c) (d)

Xplainer: Subset-minimal explanation

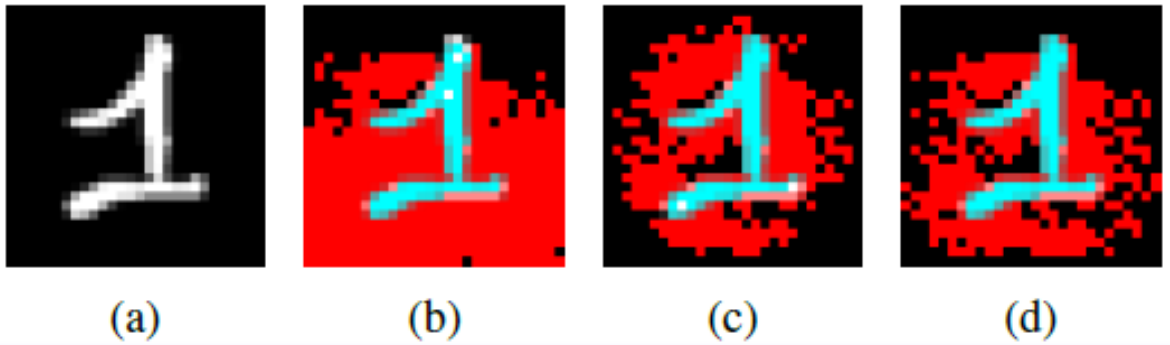
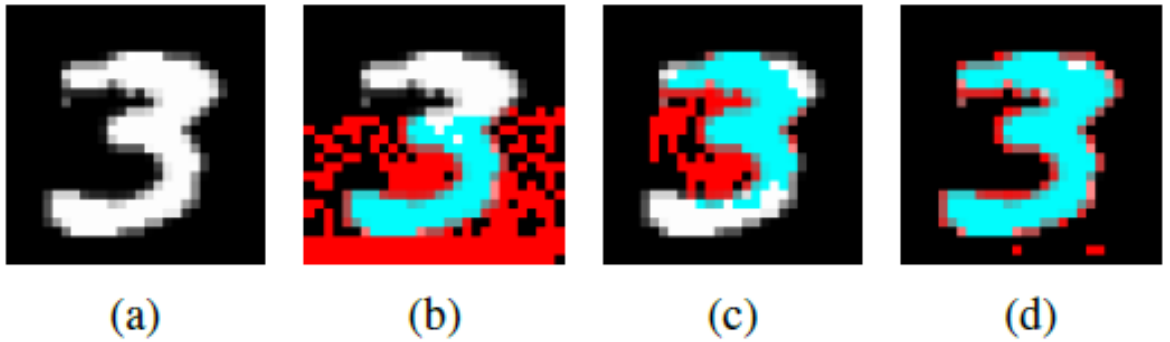
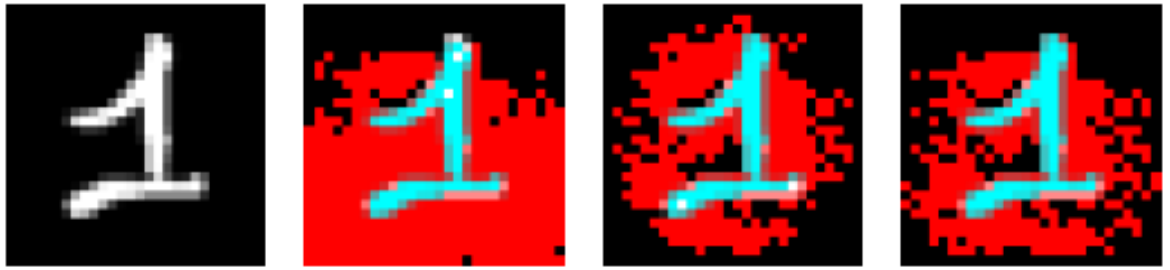


Figure 1: Possible minimal explanations for digit one.

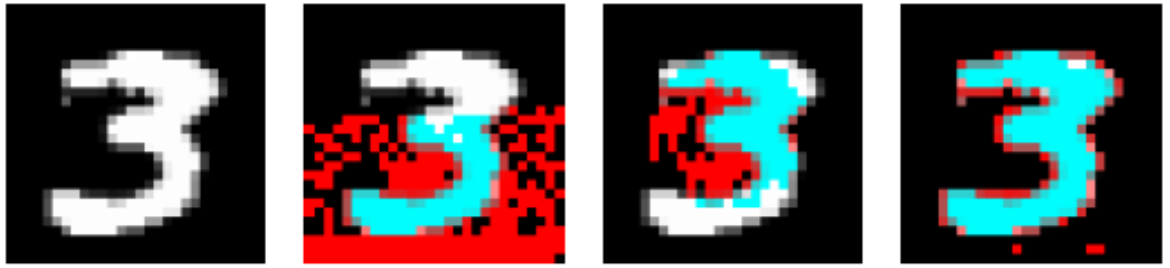


Xplainer: Subset-minimal explanation



(a) (b) (c) (d)

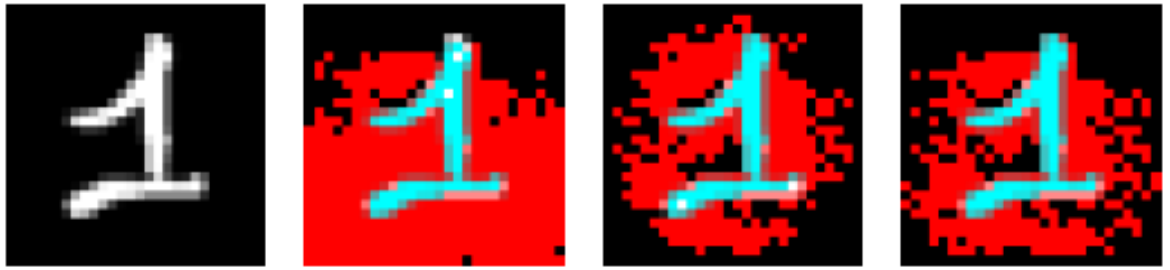
Figure 1: Possible minimal explanations for digit one.



(a) (b) (c) (d)

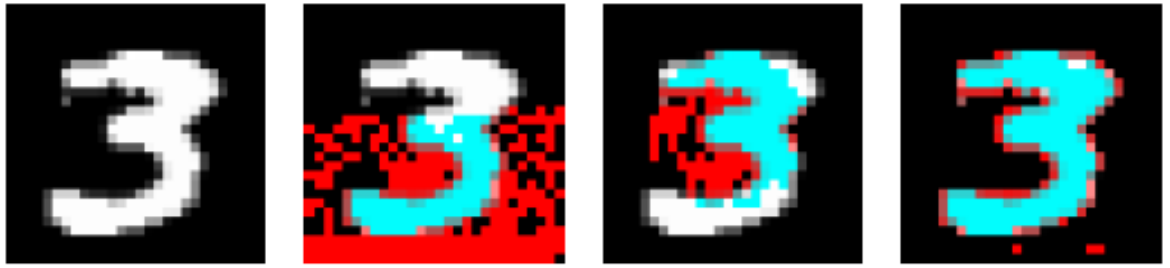
Explanations are not unique

Xplainer: Subset-minimal explanation



(a) (b) (c) (d)

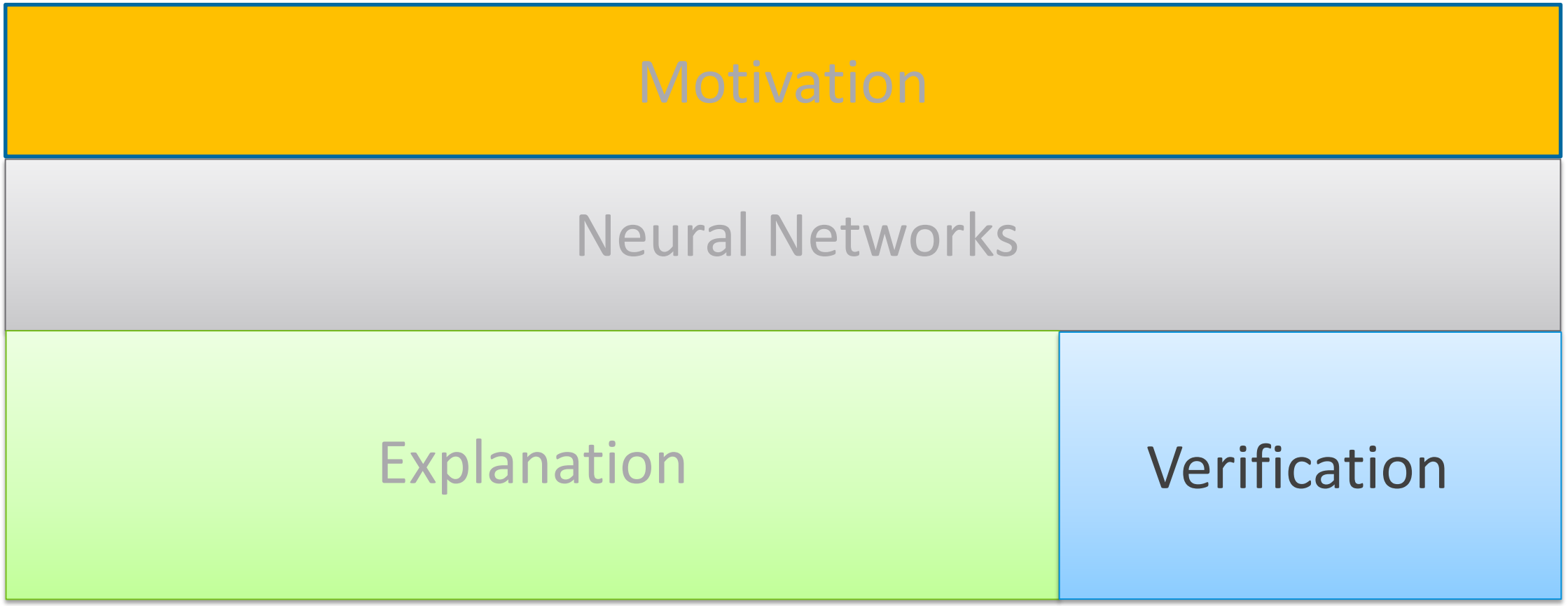
Figure 1: Possible minimal explanations for digit one.



(a) (b) (c) (d)

Some look more sensible than others!

Outline



Verification of NNs

1. Certification of Neural Networks
   (train a network that satisfies a property)

2. Verification of Neural Networks
   (complete, incomplete verification)

Verification of NNs

1. Certification of Neural Networks
   (train a network that satisfies a property)

2. Verification of Neural Networks
   (complete, incomplete verification)

Algorithms for Verifying Deep Neural Networks
Changliu Liu, Tomer Arnon, Christopher Lazarus, Clark Barrett, Mykel J. Kochenderfer

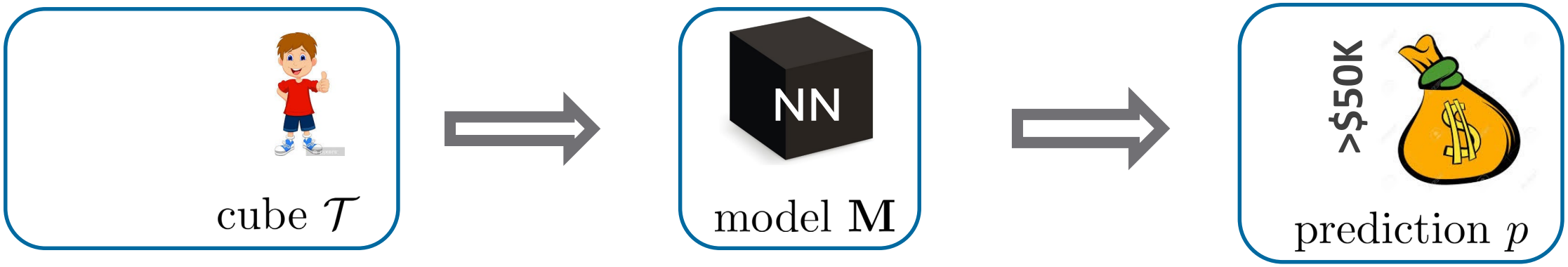
Logic-based approach

Logic-based approach

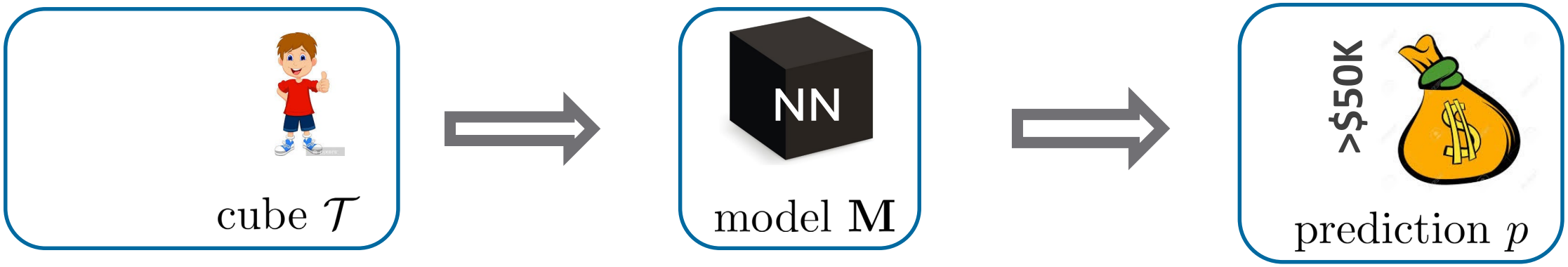
Given a classifier $\mathbf{M}$, a *counterexample* to a prediction p is a subset-minimal set of feature literals $\mathcal{T}$, such that

$$\mathcal{T} \models \vee_{t, t \neq p} (\mathbf{M} \rightarrow t)$$

Logic-based approach

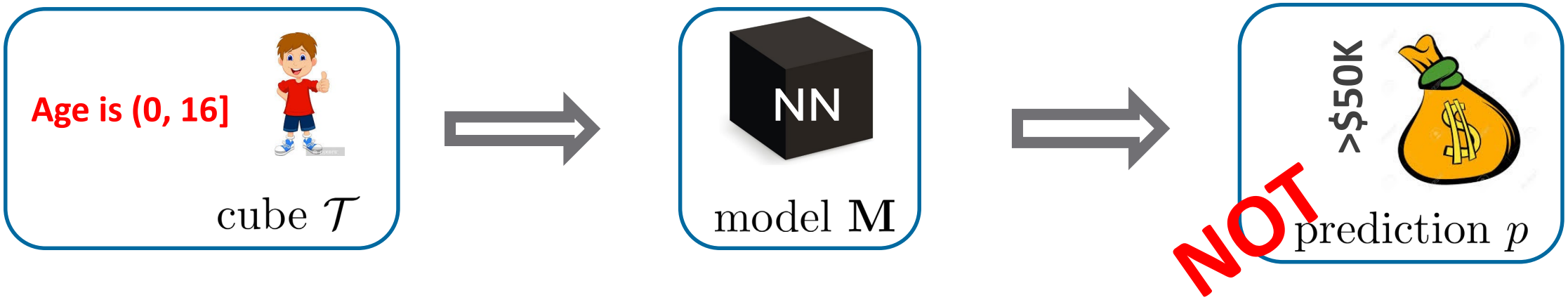


Logic-based approach



$$\mathcal{T} \models \vee_{t,t \neq p}(\mathbf{M} \rightarrow t)$$

Logic-based approach



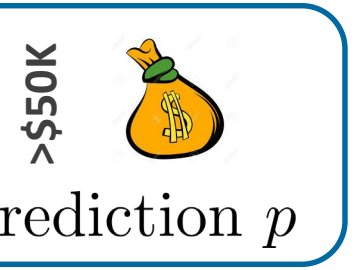
$$\mathcal{T} \models \vee_{t,t \neq p}(\mathbf{M} \rightarrow t)$$

Duality

$\mathcal{T}$ **breaks** $\mathcal{Z}$

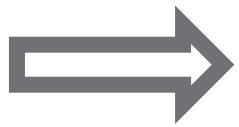
Duality

Age is (0, 16], cube $\mathcal{T}$



$\mathcal{T}$ **breaks** $\mathcal{Z}$

Age is (30, 65],
Male, cube $\mathcal{Z}$



Duality

$\mathcal{T}$ **breaks** $\mathcal{Z}$

Duality

Given a model $\mathbf{M}$, represented by some logic encoding, and a prediction p,

- every explanation $\mathcal{Z}$ of p *breaks every* counterexample of p, and
- every counterexample $\mathcal{T}$ of p *breaks every* explanation of p.

On Relating Explanations and Adversarial Examples, NeurIPS'2019
Alexey Ignatiev, Nina Narodytska, Joao Marques-Silva

Logic-based approach

Input: formula $\mathbf{M}$ and prediction p
Output: set $\mathbb{E}$ of all absolute explanations of prediction p

$(\mathbb{C}, \mathbb{E}, \mathcal{Z}) \leftarrow (\emptyset, \emptyset, \emptyset)$
do:
  if $\mathcal{Z} \models (\mathbf{M} \rightarrow p)$ **:**
    $\mathbb{E} \leftarrow \mathbb{E} \cup \{\mathcal{Z}\}$ `# ` $\mathcal{Z}$ `is an explanation; save it`
  else:
    $(\mathcal{T}, t) \leftarrow$ `ExtractInstance()` `# get an instance` $\mathcal{T}$ `with a prediction` t`,` $t \neq p$
    for $l \in \mathcal{T}$ **:**
      if $(\mathcal{T} \setminus \{l\}) \models (\mathbf{M} \rightarrow t)$ **:**
        $\mathcal{T} \leftarrow \mathcal{T} \setminus \{l\}$
    $\mathbb{C} \leftarrow \mathbb{C} \cup \{\mathcal{T}\}$ `# update` $\mathbb{T}$ `with a new counterexample` $\mathcal{T}$
  $\mathcal{E} \leftarrow$ `MinimumHS`$(\mathbb{C})$ `# get a new hitting set of` $\mathbb{C}$
while $\mathcal{Z} \neq \emptyset$
return $\mathbb{E}$

Algorithm 1: Duality-based computation of all absolute explanations

Logic-based approach

Input: formula $\mathbf{M}$ and prediction p
Output: set $\mathbb{E}$ of all absolute explanations of prediction p

$(\mathbb{C}, \mathbb{E}, \mathcal{Z}) \leftarrow (\emptyset, \emptyset, \emptyset)$
do:
  if $\mathcal{Z} \models (\mathbf{M} \rightarrow p)$ **:**
    $\mathbb{E} \leftarrow \mathbb{E} \cup \{\mathcal{Z}\}$ `# ` $\mathcal{Z}$ `is an explanation; save it`
  else:
    $(\mathcal{T}, t) \leftarrow$ `ExtractInstance()` `# get an instance` $\mathcal{T}$ `with a prediction` t, $t \neq p$
    for $l \in \mathcal{T}$ **:**
      if $(\mathcal{T} \setminus \{l\}) \models (\mathbf{M} \rightarrow t)$ **:**
        $\mathcal{T} \leftarrow \mathcal{T} \setminus \{l\}$
    $\mathbb{C} \leftarrow \mathbb{C} \cup \{\mathcal{T}\}$ `# update` $\mathbb{T}$ `with a new counterexample` $\mathcal{T}$
  $\mathcal{E} \leftarrow$ `MinimumHS`$(\mathbb{C})$ `# get a new hitting set of` $\mathbb{C}$
while $\mathcal{Z} \neq \emptyset$
return $\mathbb{E}$

Algorithm 1: Duality-based computation of all absolute explanations

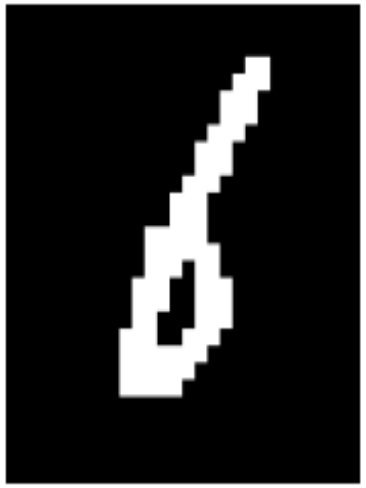
Logic-based approach

Input: formula $\mathbf{M}$ and prediction p
Output: set $\mathbb{E}$ of all absolute explanations of prediction p

$(\mathbb{C}, \mathbb{E}, \mathcal{Z}) \leftarrow (\emptyset, \emptyset, \emptyset)$
do:
  if $\mathcal{Z} \models (\mathbf{M} \rightarrow p)$ **:**
    $\mathbb{E} \leftarrow \mathbb{E} \cup \{\mathcal{Z}\}$ `# ` $\mathcal{Z}$ `is an explanation; save it`
  else:
    $(\mathcal{T}, t) \leftarrow$ `ExtractInstance()` `# get an instance` $\mathcal{T}$ `with a prediction` t, $t \neq p$
    for $l \in \mathcal{T}$ **:**
      if $(\mathcal{T} \setminus \{l\}) \models (\mathbf{M} \rightarrow t)$ **:**
        $\mathcal{T} \leftarrow \mathcal{T} \setminus \{l\}$
    $\mathbb{C} \leftarrow \mathbb{C} \cup \{\mathcal{T}\}$ `# update` $\mathbb{T}$ `with a new counterexample` $\mathcal{T}$
  $\mathcal{E} \leftarrow$ `MinimumHS`$(\mathbb{C})$ `# get a new hitting set of` $\mathbb{C}$
while $\mathcal{Z} \neq \emptyset$
return $\mathbb{E}$

Algorithm 1: Duality-based computation of all absolute explanations

Logic-based approach

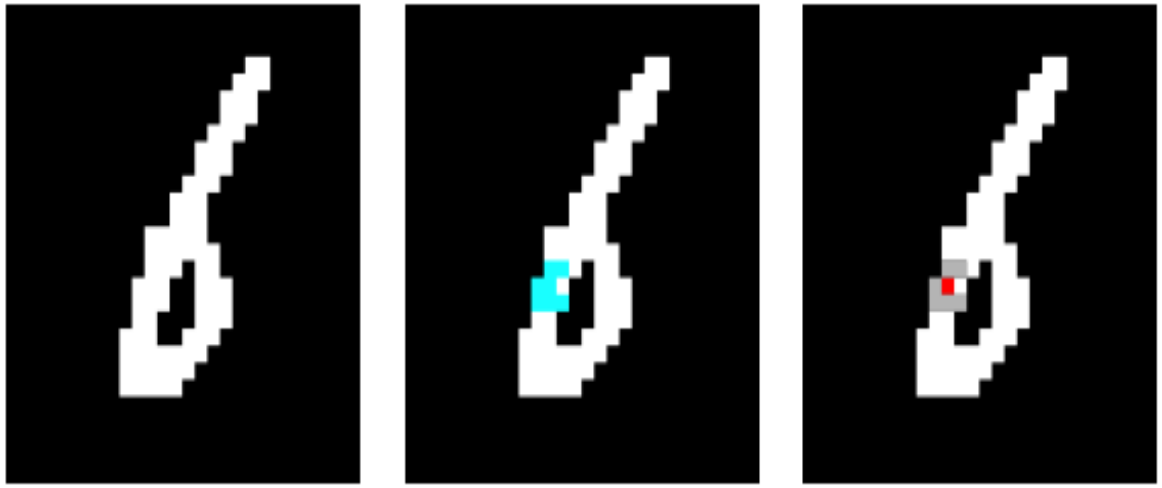


(a) digit “6”

Logic-based approach

(a) digit “6” (b) patch area

Logic-based approach

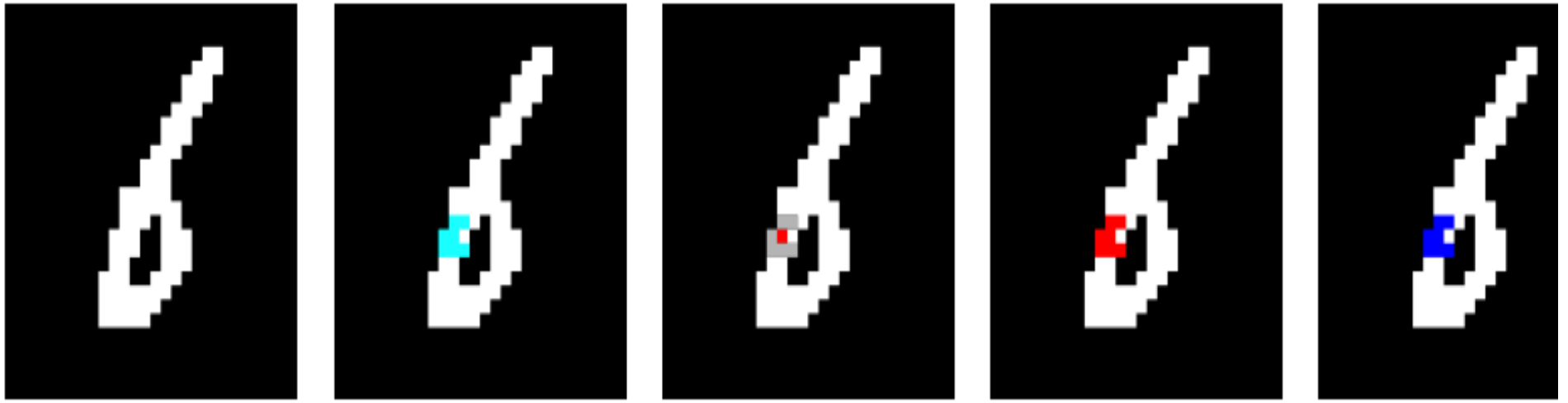


(a) digit “6” (b) patch area (c) an XP

Logic-based approach

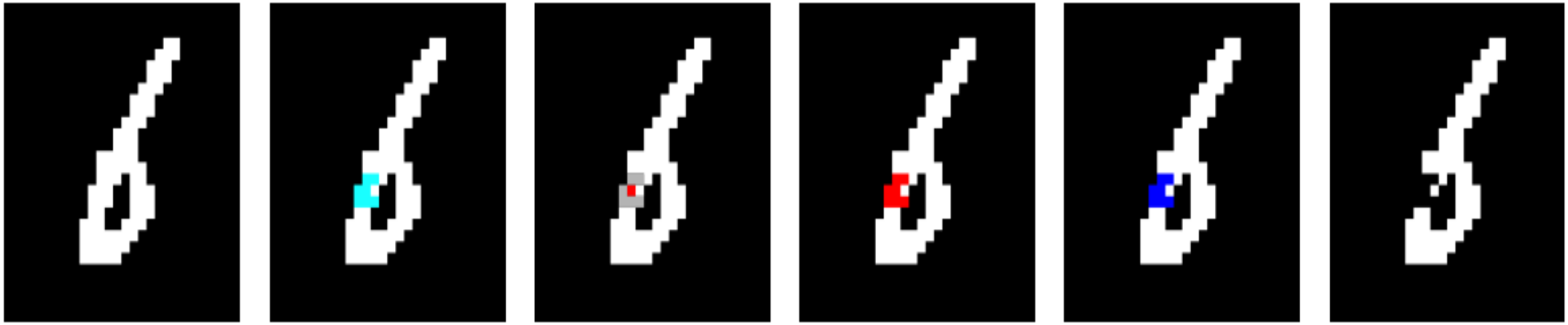
(a) digit “6” (b) patch area (c) an XP (d) all XP’s

Logic-based approach



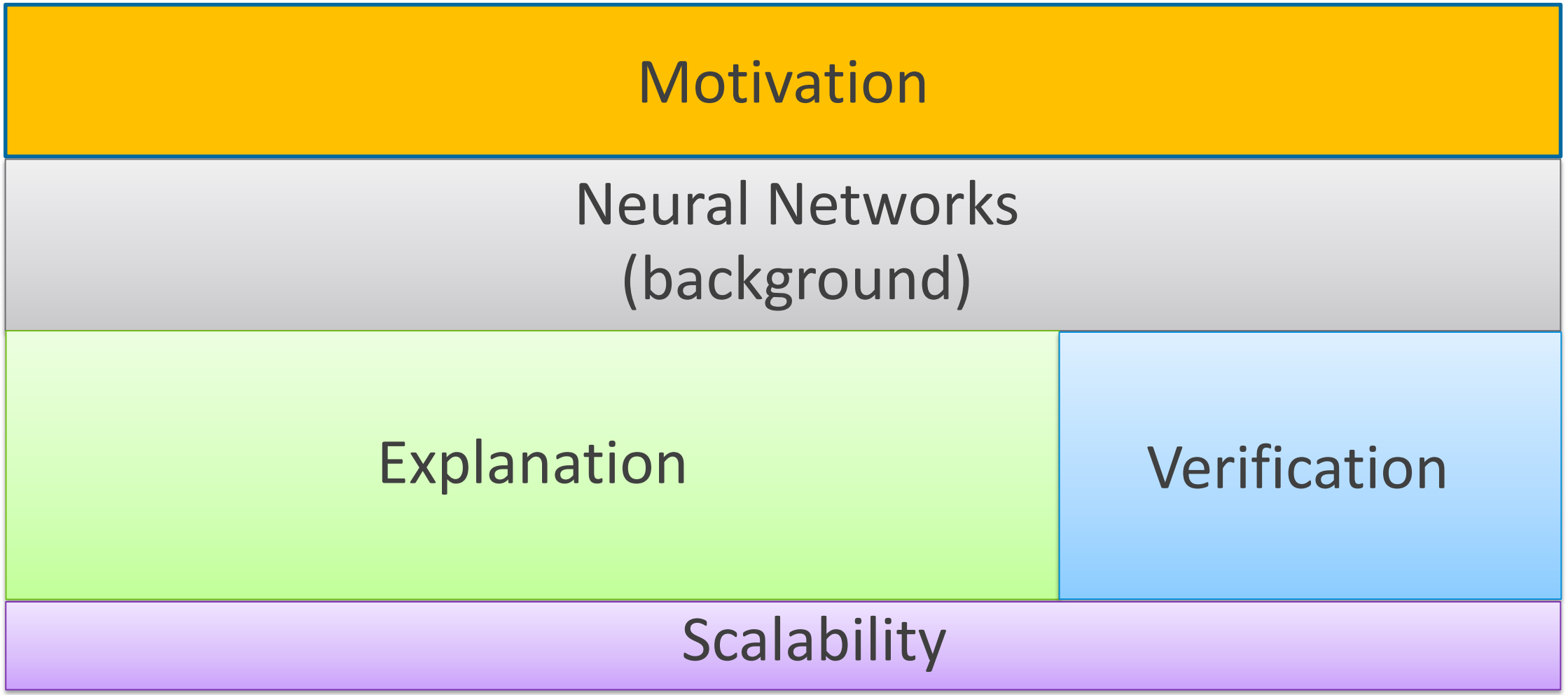
(a) digit “6” (b) patch area (c) an XP (d) all XP’s (e) a CE

Logic-based approach



(a) digit “6” (b) patch area (c) an XP (d) all XP’s (e) a CE (f) an AE (“5”)

Outline


Motivation
Neural Networks
(background)
Explanation
Verification
Scalability

Scalability

Scalability

Heuristic methods for NN explainability
do not suffer from scalability issues

Scalability

Heuristic methods for NN explainability
do not suffer from scalability issues

no guarantees about the quality of explanations

Scalability

Logic-based methods on explanability of NNs
do suffer from scalability issues

Scalability

Input: formula $\mathbf{M}$ and prediction p
Output: set $\mathbb{E}$ of all absolute explanations of prediction p

$(\mathbb{C}, \mathbb{E}, \mathcal{Z}) \leftarrow (\emptyset, \emptyset, \emptyset)$
do:
  if $\mathcal{Z} \models (\mathbf{M} \rightarrow p)$ **:**
    $\mathbb{E} \leftarrow \mathbb{E} \cup \{\mathcal{Z}\}$ `# ` $\mathcal{Z}$ `is an explanation; save it`
  else:
    $(\mathcal{T}, t) \leftarrow$ `ExtractInstance()` `# get an instance` $\mathcal{T}$ `with a prediction` t, $t \neq p$
    for $l \in \mathcal{T}$ **:**
      if $(\mathcal{T} \setminus \{l\}) \models (\mathbf{M} \rightarrow t)$ **:**
        $\mathcal{T} \leftarrow \mathcal{T} \setminus \{l\}$
    $\mathbb{C} \leftarrow \mathbb{C} \cup \{\mathcal{T}\}$ `# update` $\mathbb{T}$ `with a new counterexample` $\mathcal{T}$
  $\mathcal{E} \leftarrow$ `MinimumHS`$(\mathbb{C})$ `# get a new hitting set of` $\mathbb{C}$
while $\mathcal{Z} \neq \emptyset$
return $\mathbb{E}$

Algorithm 1: Duality-based computation of all absolute explanations

Scalability

It is an issue!

Scalability

Many papers on verification of NNs is battling scalability issue, e.g.

- add constraints to the training procedure
- use approximate reasoning
- simplify networks during the training

Scalability

Many papers on verification of NNs is battling scalability issue, e.g.

- add constraints to the training procedure
- use approximate reasoning
- simplify networks during the training

Scalability

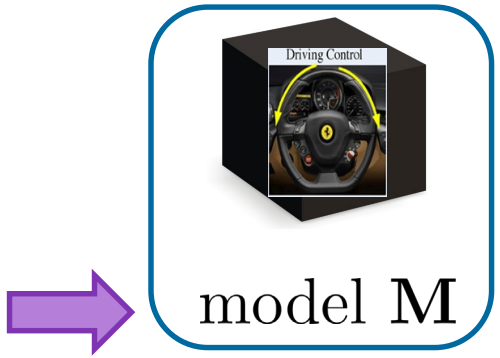
In Search for a SAT-friendly Binarized Neural Network Architecture

Hongce Zhang (summer @ VMware), Aarti Gupta, Toby Walsh

Training for Faster Adversarial Robustness Verification via Inducing ReLU Stability
Kai Y. Xiao, Vincent Tjeng, Nur Muhammad (Mahi) Shafiullah, Aleksander Madry

Scalability

Scalability



Scalability

model $\mathbf{M}$

Scalability

model $\mathbf{M}$

Audi AI:ME, concept level 4 autonomous car, April 2019

Scalability

Verification

model M

Explanation

Audi AI:ME, concept level 4 autonomous car, April 2019

Scalability

Verification

model $\mathbf{M}$

Explanation

Can we train a NN so that it is easier to analyze?

Conclusion

Conclusion

We showed a tight connection between explanations and counterexamples

Conclusion

There is hope to battle scalability issues!

Conclusion

Thanks!