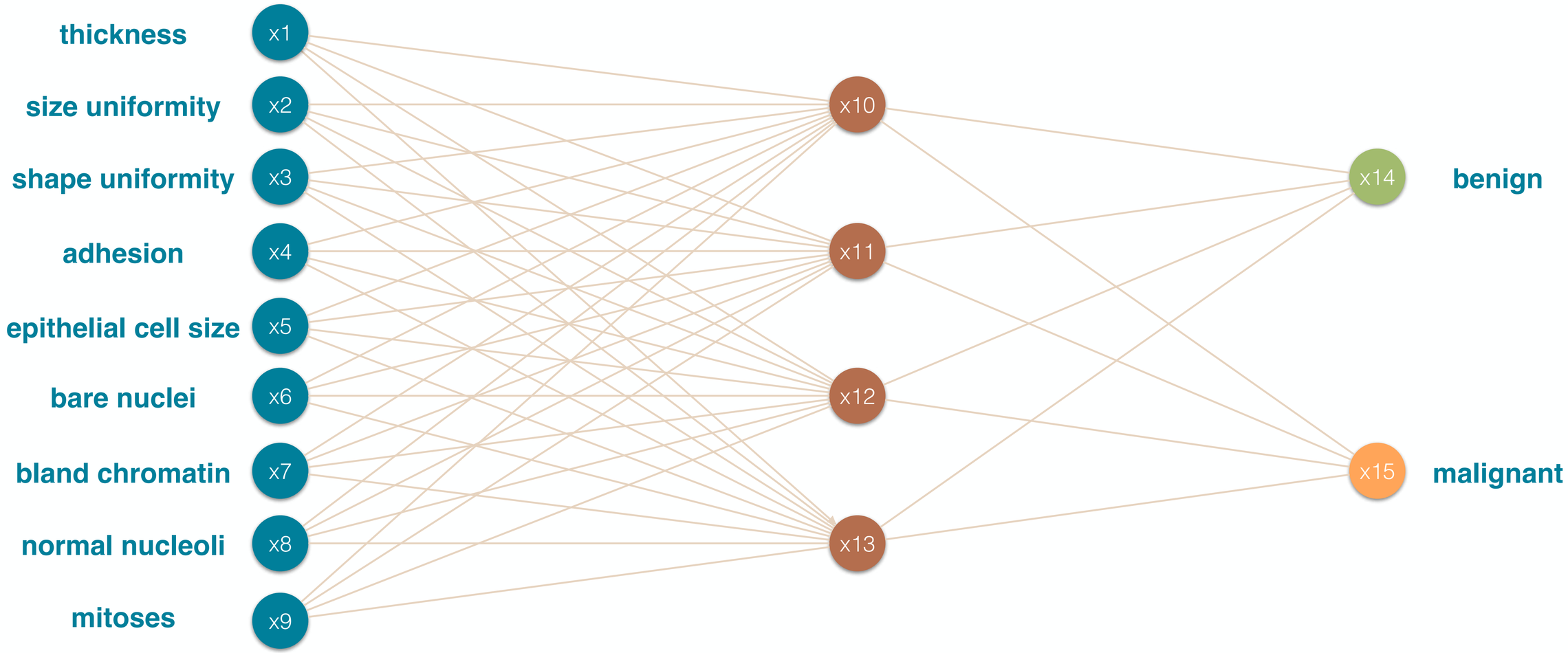
Faster Verified Explanations for Neural Networks

SAIF Days 2026

Caterina Urban (joint work with **Alessandro De Palma** and **Greta Dolcetti**)
Inria & École Normale Supérieure | Université PSL

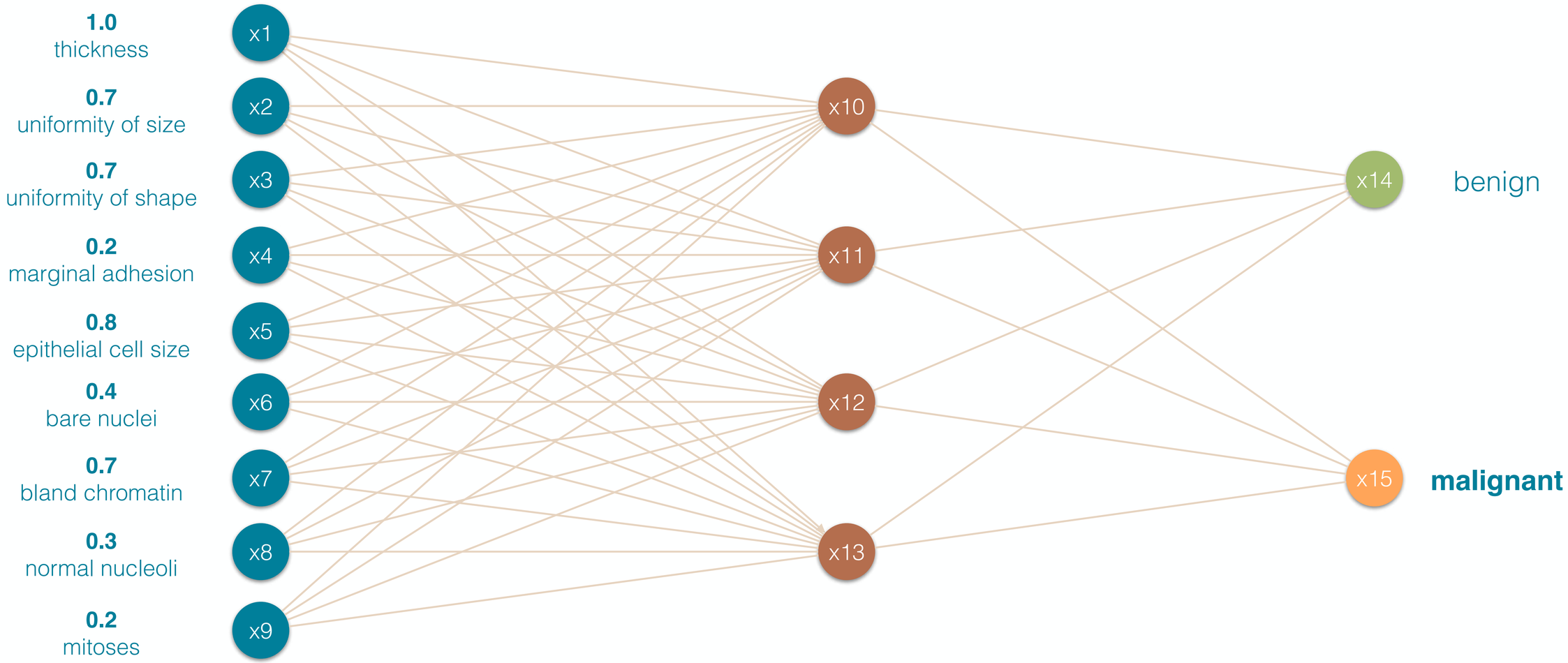
Optimal Robust Explanations

Abductive Explanations (AXps)



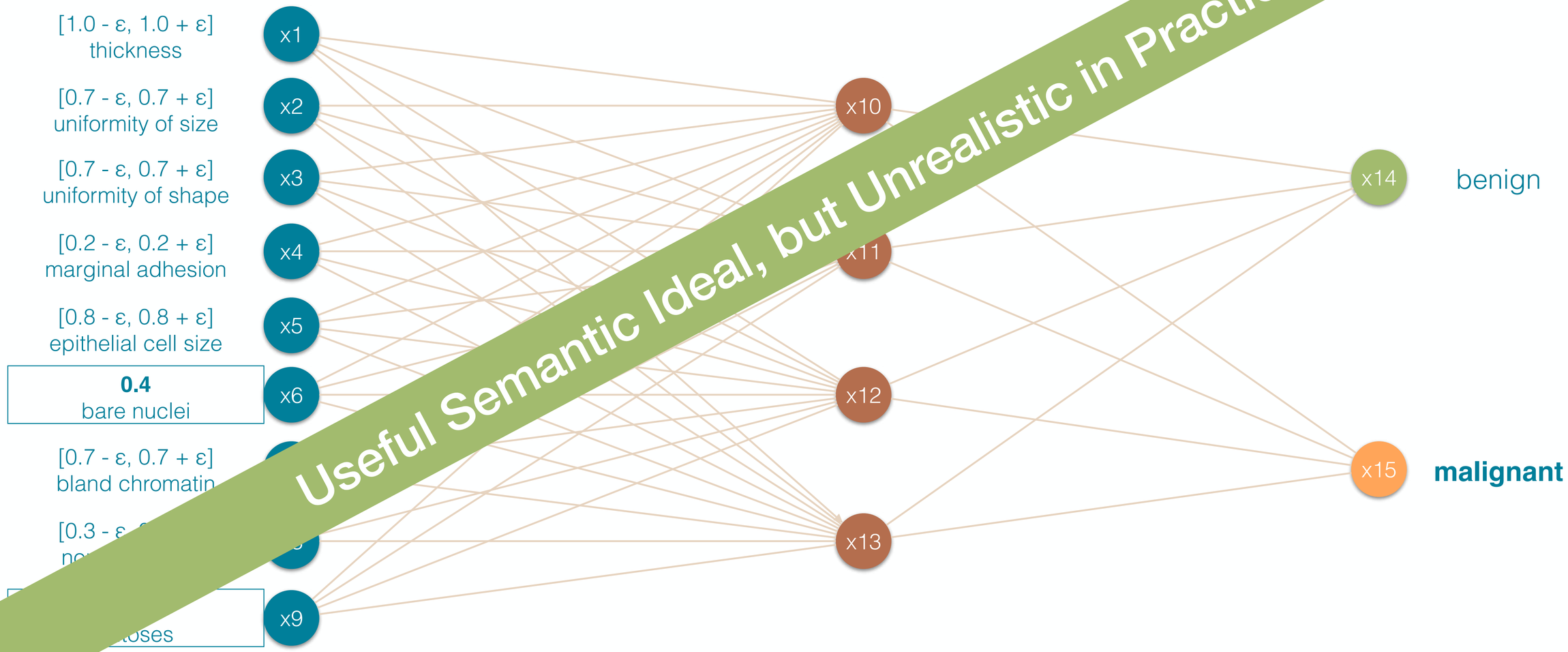
Optimal Robust Explanations

Abductive Explanations (AXps)



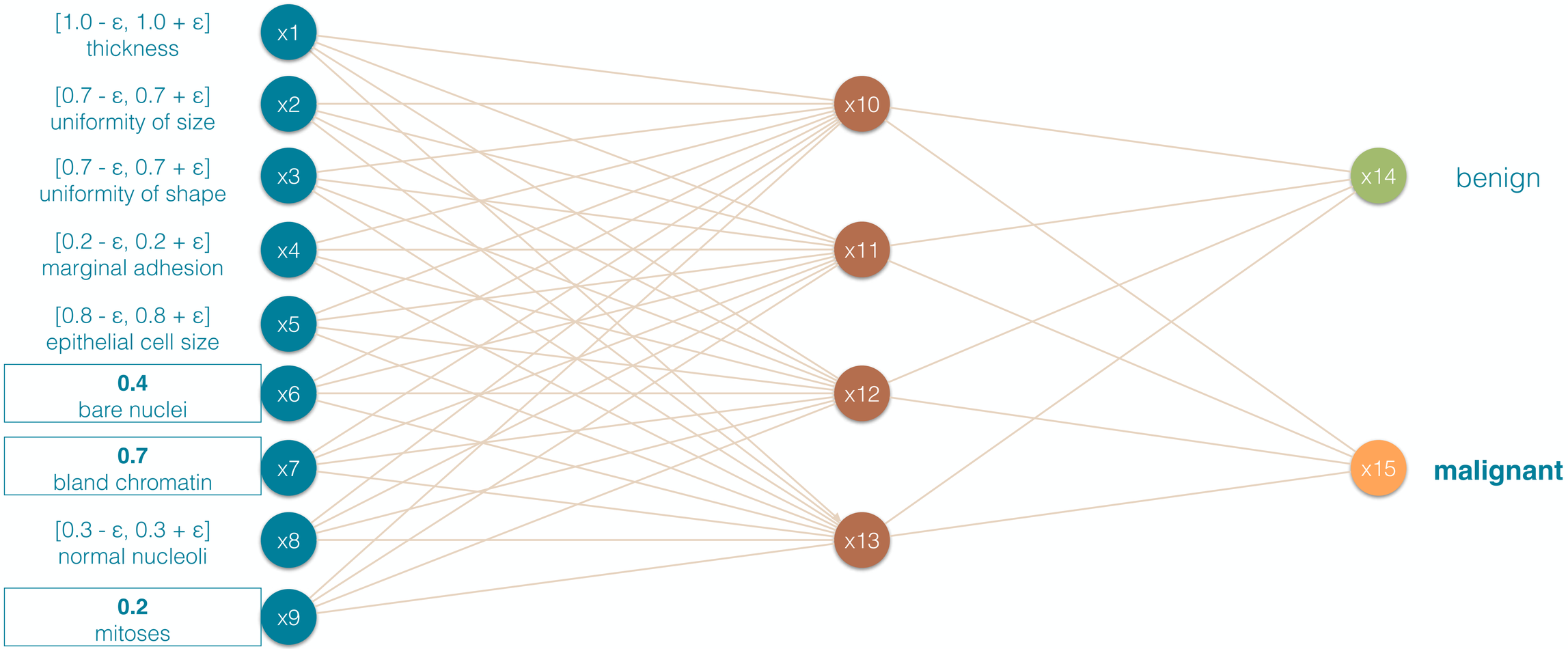
Optimal Robust Explanations

Abductive Explanations (AXps)



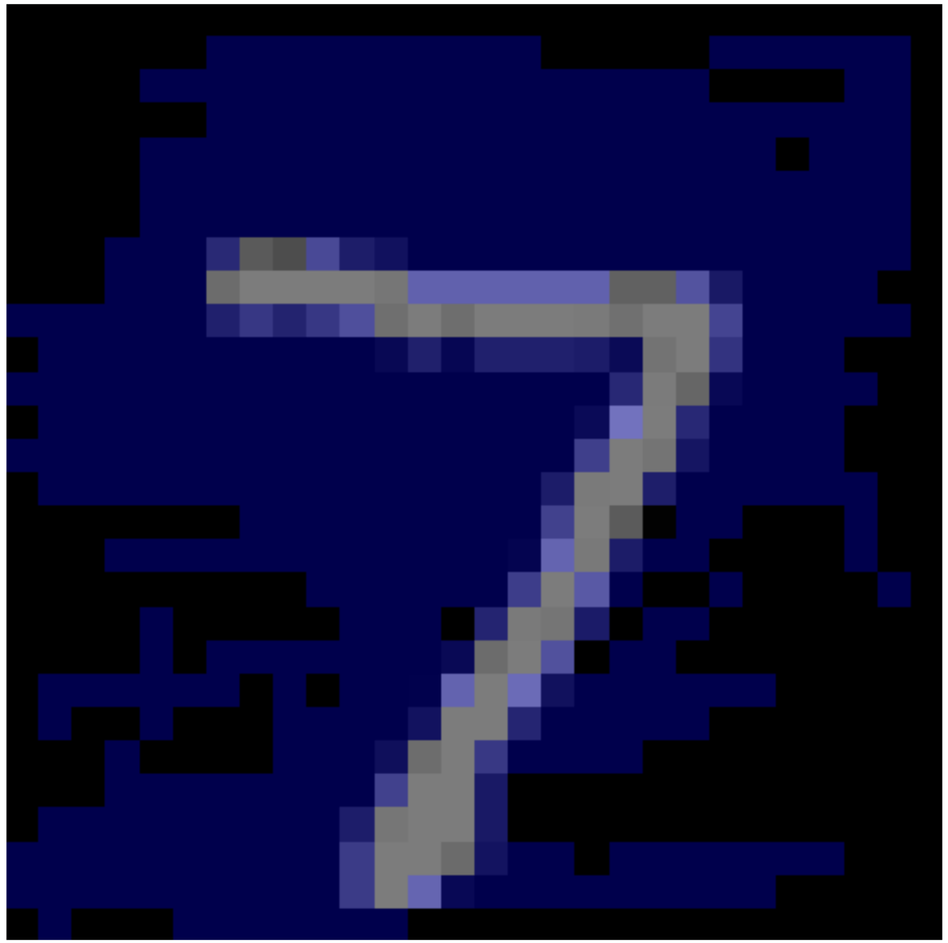
~~Optimal~~ Robust Explanations

Weak Abductive Explanations



~~Optimal~~ Robust Explanations

Weak Abductive Explanations



Computing ~~Optimal~~ Robust Explanations

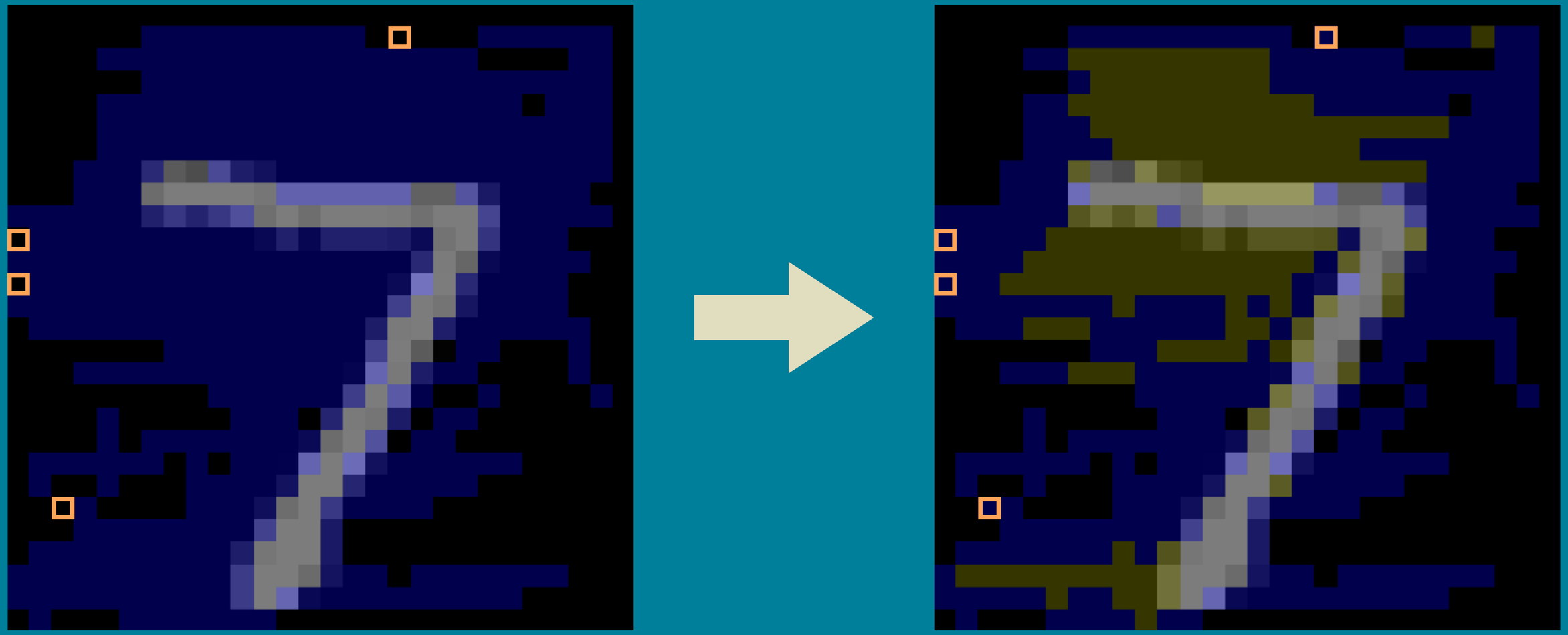
Finding Counterfactuals Rapidly Becomes Infeasible

Model	MNIST, ϵ=0.25			CIFAR-10, ϵ=8/255		
	Counterfactuals	Unknowns	Time	Counterfactuals	Unknowns	Time
CNN-3	0.00	247.80	45m	0.00	461.00	2h 30m

Model	MNIST, ϵ=0.25			CIFAR-10, ϵ=16/255		
	Counterfactuals	Unknowns	Time	Counterfactuals	Unknowns	Time
CNN-7	0.00	452.00	3h 59m	0.00	730.67	7h 5m

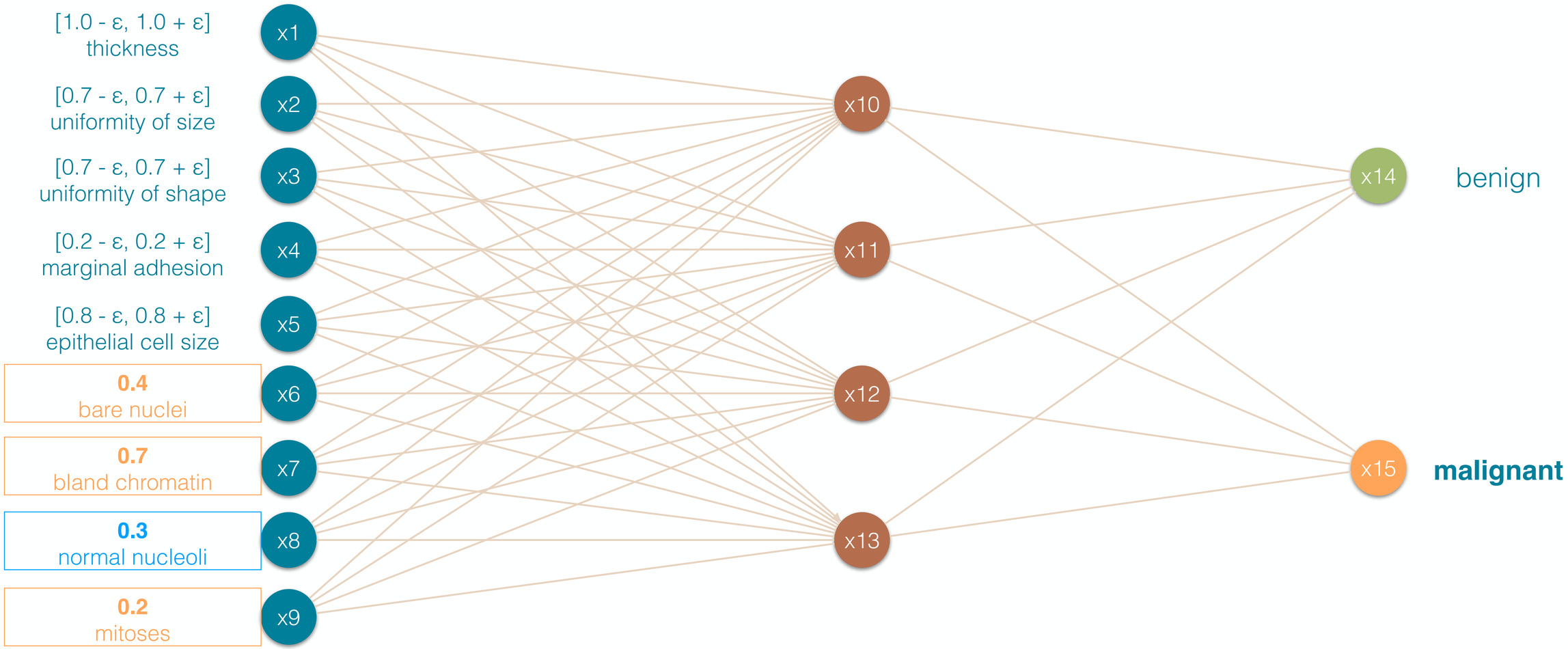
Verifier-Optimal Robust Explanations

Weak Abductive Explanations



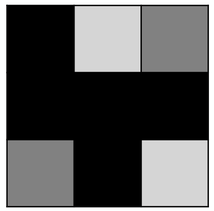
Verifier-Optimal Robust Explanations

Weak Abductive Explanations

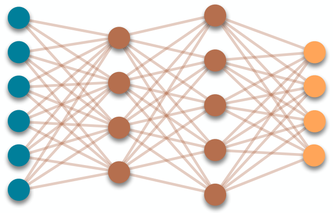


Neural Network Verification

Local Robustness



$\mathbf{x} \in \mathbb{R}^d$



f

4

$y_f(\mathbf{x})$

$(\mathbf{x}, \mathscr{A}, \epsilon, \mathbf{f})$

v

1 $\forall x' \in \mathscr{B}^{\epsilon}_{\mathscr{A}} : y_f(x') = y_f(x)$

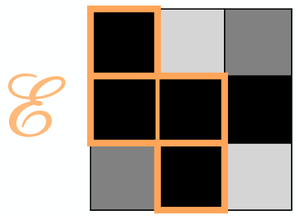
0 🤷‍♀️

-1 $\exists x' \in \mathscr{B}^{\epsilon}_{\mathscr{A}} : y_f(x') \neq y_f(x)$

$$B^{\epsilon}_{\mathscr{A}}(\mathbf{x}) \stackrel{\text{def}}{=} \{\mathbf{x}' \in \mathbb{R}^d \mid \mathbf{x}'_{\overline{\mathscr{A}}} = \mathbf{x}_{\overline{\mathscr{A}}} \wedge \left\| \mathbf{x}'_{\mathscr{A}} - \mathbf{x}_{\mathscr{A}} \right\|_{\infty} \leq \epsilon\}$$

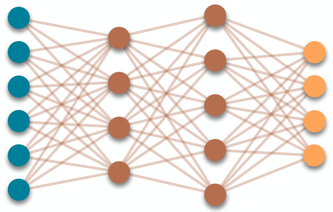
~~Optimal~~ Robust Explanations

Weak Abductive Explanations



$\mathscr{E}$

$\mathbf{x} \in \mathbb{R}^d$



f

4

$\mathbf{y_f(x)}$

$(\mathbf{x}, \overline{\mathscr{E}}, \epsilon, \mathbf{f})$

1

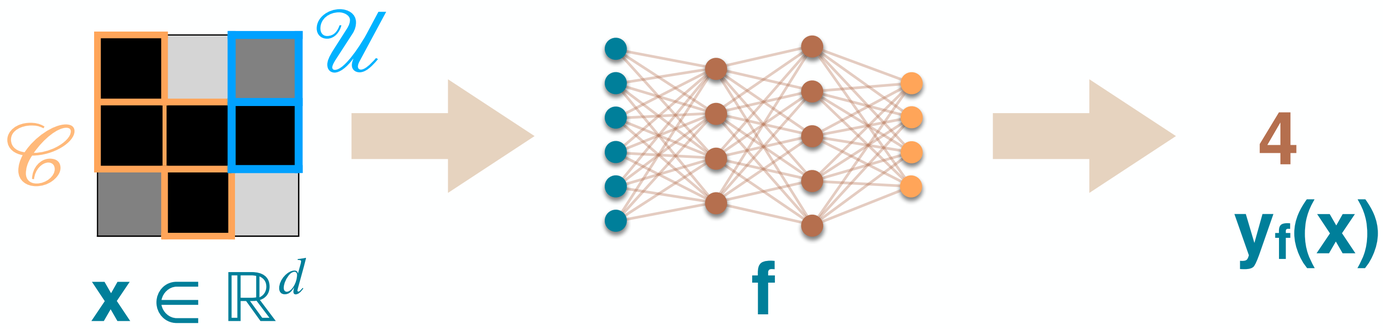
$\forall i \in \mathscr{E}: \ (\mathbf{x}, \overline{\mathscr{E}} \cup \{i\}, \epsilon, \mathbf{f})$

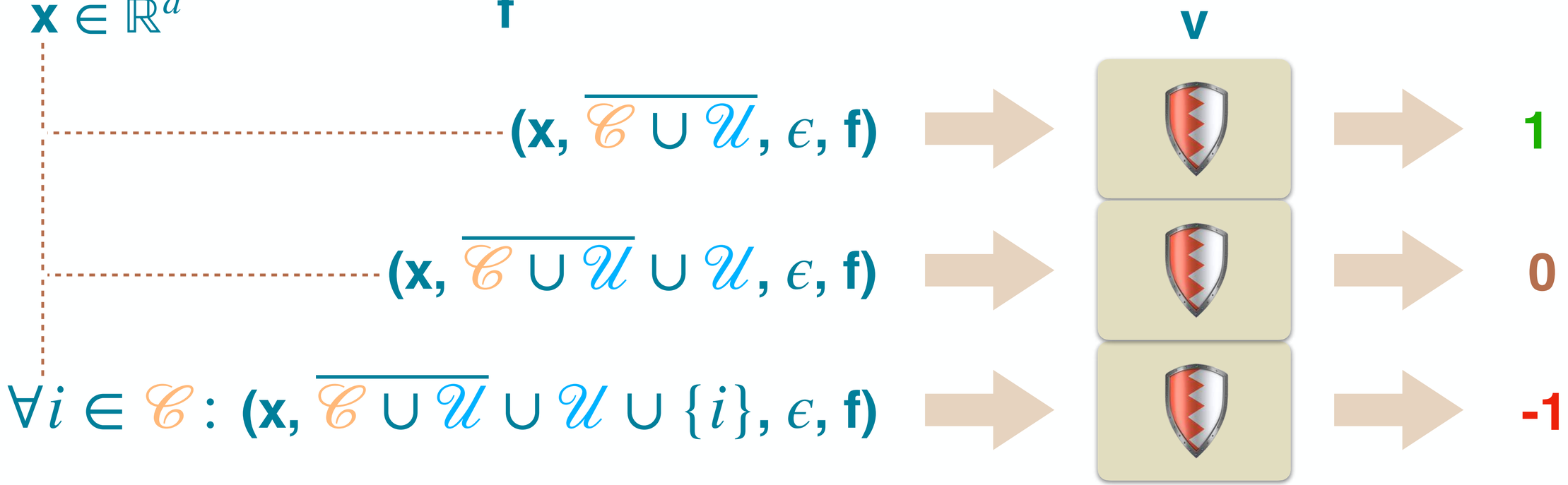
-1 or **0**

v

Verifier-Optimal Robust Explanations

Weak Abductive Explanations



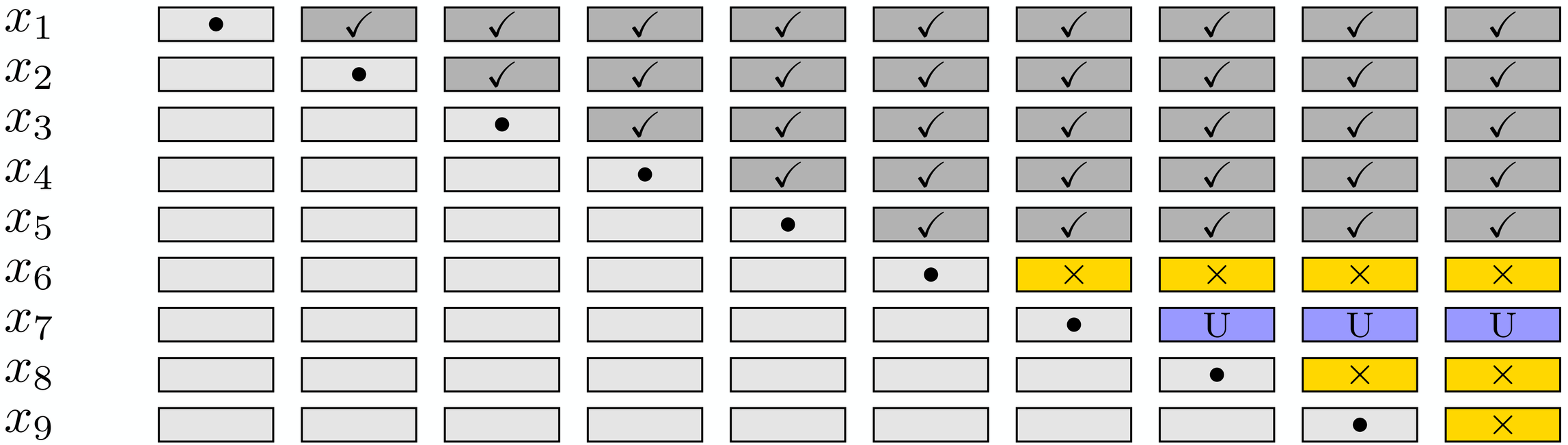


Computing Verifier-Optimal Robust Explanations

Drop (i.e., Free) Input Features While AXp Condition Holds

ADD TO $\overline{\mathscr{C} \cup \mathscr{U}}$

LOCAL ROBUSTNESS IN $B^{\epsilon}_{\overline{\mathscr{C} \cup \mathscr{U}}}(\mathbf{x})$



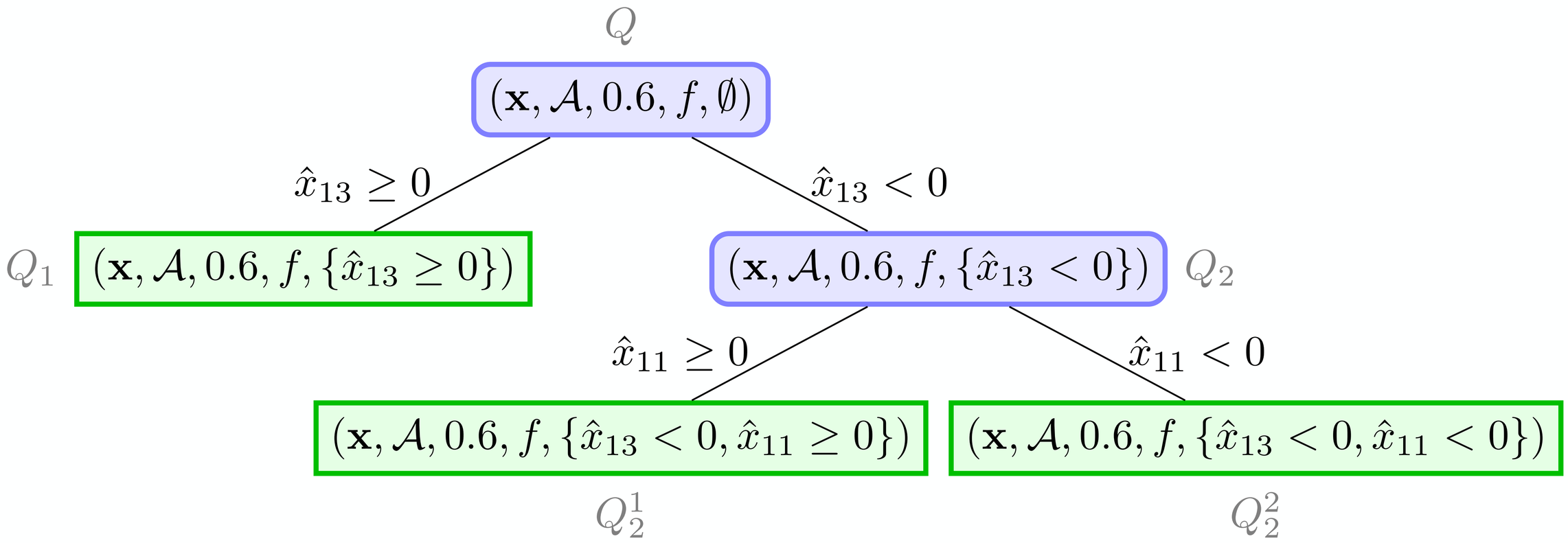
Computing Verifier-Optimal Robust Explanations

FaVeX (Simplified)

1: **function** FaVeX($f, \mathbf{x}, \epsilon, v, \vec{\mathcal{A}}$)
2: $\quad\mathcal{R}_\mathbf{x}, \mathcal{U}_\mathbf{x}, \mathcal{C}_\mathbf{x} \leftarrow \emptyset, \emptyset, \emptyset$
3: $\quad$batches $\leftarrow \{\vec{\mathcal{A}}\}$
4: $\quad$**for** $B \in$ batches **do**
5: $\quad\quad$batches $\leftarrow$ batches $\setminus \{B\}$
6: $\quad\quad$**if** $|B| > 1$ **then** ▷ *batch robustness query*
7: $\quad\quad\quad$result $\leftarrow v(\mathbf{x}, \mathcal{R}_\mathbf{x} \cup \mathcal{U}_\mathbf{x} \cup B, \epsilon, f)$
8: $\quad\quad\quad$**if** result $== 1$ **then** $\mathcal{R}_\mathbf{x} \leftarrow \mathcal{R}_\mathbf{x} \cup B$
9: $\quad\quad\quad$**else**
10: $\quad\quad\quad\quad\vec{\mathcal{A}}_1, \vec{\mathcal{A}}_2 \leftarrow$ halve$(\vec{\mathcal{A}})$
11: $\quad\quad\quad\quad$batches $\leftarrow$ batches $\cup \{\vec{\mathcal{A}}_1, \vec{\mathcal{A}}_2\}$
12: $\quad\quad$**else** ▷ *single robustness query*
13: $\quad\quad\quad$result $\leftarrow v(\mathbf{x}, \mathcal{R}_\mathbf{x} \cup \mathcal{U}_\mathbf{x} \cup B, \epsilon, f)$
14: $\quad\quad\quad$**if** result $== 1$ **then** $\mathcal{R}_\mathbf{x} \leftarrow \mathcal{R}_\mathbf{x} \cup B$
15: $\quad\quad\quad$**else if** result $=$ -1 **then** $\mathcal{C}_\mathbf{x} \leftarrow \mathcal{C}_\mathbf{x} \cup B$ **else** $\mathcal{U}_\mathbf{x} \leftarrow \mathcal{U}_\mathbf{x} \cup B$
16: $\quad$**return** $\mathcal{U}_\mathbf{x}, \mathcal{C}_\mathbf{x}$

Computing Verifier-Optimal Robust Explanations

Branch-and-Bound Verification



Computing Verifier-Optimal Robust Explanations

Branch-and-Bound Verification with Branching Save/Reuse

1: **function** BAB$(a, (\mathbf{x}, \mathcal{A}, \epsilon, f), \mathbb{C})$ ▷ $a\colon \hat{\mathcal{Q}} \to \mathbb{R}, \mathbb{C} \in \mathcal{P}(\mathcal{P}(\mathcal{C}))$
2: **for** $C \in \mathbb{C}$ **do** unresolved $\leftarrow \{(\mathbf{x}, \mathcal{A}, \epsilon, f, C)\}$ ▷ *Reuse*
3: $\mathbb{C} \leftarrow \emptyset$
4: **for** $Q \in$ unresolved **do**
5: **if** CEX$(\mathbf{x}, \mathcal{A}, \epsilon, f)$ **then**
6: **for** $Q' \in$ unresolved **do** $\mathbb{C} \leftarrow \mathbb{C} \cup \{Q'_C\}$ ▷ *Save* unresolved
7: **return** $-1, \mathbb{C}$
8: unresolved $\leftarrow$ unresolved $\setminus \{Q\}$
9: **if** a(Q) > 0 **then** $\mathbb{C} \leftarrow \mathbb{C} \cup \{Q_C\}$ ▷ *Save* Q
10: **else if** $\neg$TIMEOUT **then**
11: $Q_1, Q_2 \leftarrow$ SPLIT(Q)
12: unresolved $\leftarrow$ unresolved $\cup \{Q_1, Q_2\}$
13: **else**
14: **for** $Q' \in$ unresolved **do** $\mathbb{C} \leftarrow \mathbb{C} \cup \{Q'_C\}$ ▷ *Save* unresolved
15: **return** $0, \mathbb{C}$
16: **return** $1, \mathbb{C}$

Computing Verifier-Optimal Robust Explanations

Restricted-Space Counterfactual Search

Figure 7 Illustration of full-space vs. restricted-space' counterfactual search. Full-space search (left) explores the entire perturbation region $B^{\epsilon}_{\mathcal{A}\cup B}(\mathbf{x})$. Restricted-space search (right) begins from the output $\mathbf{z}$ of the previous search (which may not be a counterfactual) and varies only the newly added feature(s), thus exploring a much smaller subset of $B^{\epsilon}_{\mathcal{A}\cup B}(\mathbf{x})$.

Computing Verifier-Optimal Robust Explanations

CNN-3

	MNIST, ϵ=0.25			CIFAR-10, ϵ=8/255		
	Counterfactuals	Unknowns	Time	Counterfactuals	Unknowns	Time
Standard	0.00	247.80	45m	0.00	461.00	2h 30m
Verifier-Optimal	**160.30**	94.40	**10m**	**210.40**	251.70	**19m**

160.30 +
94.40 =
254.70

210.40 +
251.70 =
462.10

Computing Verifier-Optimal Robust Explanations

CNN-3

Ablations

Table 6 On `CNN-3`, FaVeX computes `OVAL`-optimal robust explanations faster than previous computation strategies [65, 64] while at the same time finding more counter-factuals. Results are averaged over the first 10 images of the test set. Bold entries correspond to the smallest runtime and the largest number of provided counterfactuals.

	MNIST, $\epsilon = 0.25$			CIFAR-10, $\epsilon = \frac{8}{255}$		
Method	$\lvert\mathcal{C}_\mathbf{x}\rvert$	$\lvert\mathcal{U}_\mathbf{x}\rvert$	time [s]	$\lvert\mathcal{C}_\mathbf{x}\rvert$	$\lvert\mathcal{U}_\mathbf{x}\rvert$	time [s]
Sequential	129.10	125.50	1164.89	209.30	253.10	1287.96
Binary Search	134.20	120.40	1026.67	209.60	252.70	1367.25
BinS + Incr	135.20	119.50	1040.82	209.60	252.50	1374.40
BinS + Incr + RSA	157.70	96.60	727.94	**211.90**	250.20	1334.00
FaVeX	**160.30**	94.40	**612.75**	210.40	251.70	**1150.82**

Computing Verifier-Optimal Robust Explanations

CNN-7

	MNIST, ϵ=0.25			CIFAR-10, ϵ=16/255		
	Counterfactuals	Unknowns	Time	Counterfactuals	Unknowns	Time
Standard	0.00	452.00	239m	0.00	730.67	7h 5m
Verifier-Optimal	**207.33**	249.33	**74m**	**467.00**	266.33	**1h 49m**

207.33 +
249.33 =
456.66

467.00 +
266.33 =
733.33

Computing Verifier-Optimal Robust Explanations

CNN-7

Ablations

Table 7 On `CNN-7`, FaVeX computes `OVAL`-optimal robust explanations faster than previous computation strategies [65, 64] while at the same time finding more counter-factuals. Results are averaged over the first 3 images of the test set. Bold entries correspond to the smallest runtime and the largest number of provided counterfactuals.

	MNIST, $\epsilon = 0.25$			CIFAR-10, $\epsilon = \frac{16}{255}$		
Method	$\lvert\mathcal{C}_\mathbf{x}\rvert$	$\lvert\mathcal{U}_\mathbf{x}\rvert$	time [s]	$\lvert\mathcal{C}_\mathbf{x}\rvert$	$\lvert\mathcal{U}_\mathbf{x}\rvert$	time [s]
Sequential	142.33	315.00	7979.61	464.00	269.33	6868.53
Binary Search	148.33	308.67	9165.06	464.33	269.00	8284.12
BinS + Incr	151.33	305.00	9116.13	467.00	266.33	8231.01
BinS + Incr + RSA	196.33	260.33	5406.70	**468.00**	265.33	8304.52
FaVeX	**207.33**	249.33	**4446.53**	467.00	266.33	**6532.98**

Computing Verifier-Optimal Robust Explanations

Traversal Strategies

Table 10 Comparison of different traversal strategies for `OVAL`-optimal robust explanations on `CNN-3`. Results are averaged over the first 10 images of the test set. Bold entries correspond to the smallest $\mathcal{C}_\mathbf{x} \cup \mathcal{U}_\mathbf{x}$.

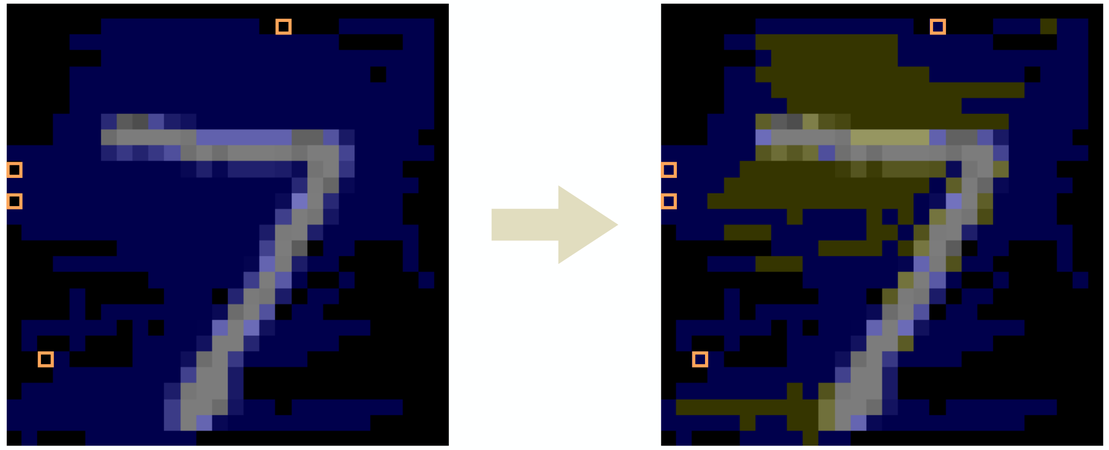
	MNIST, $\epsilon = 0.25$				CIFAR-10, $\epsilon = \frac{8}{255}$			
Method	$\lvert\mathcal{C}_\mathbf{x} \cup \mathcal{U}_\mathbf{x}\rvert$	$\lvert\mathcal{C}_\mathbf{x}\rvert$	$\lvert\mathcal{U}_\mathbf{x}\rvert$	time [s]	$\lvert\mathcal{C}_\mathbf{x} \cup \mathcal{U}_\mathbf{x}\rvert$	$\lvert\mathcal{C}_\mathbf{x}\rvert$	$\lvert\mathcal{U}_\mathbf{x}\rvert$	time [s]
VeriX	282.80	160.50	122.30	974.35	765.50	437.20	328.30	1925.09
VeriX+	**247.40**	155.20	92.20	576.67	468.90	262.00	206.90	915.45
α-FaVeX	289.90	181.20	108.70	696.31	**462.10**	210.40	251.70	1150.82
FaVeX-IBP	254.70	160.30	94.40	612.75	466.40	250.90	215.50	941.95

Table 11 Comparison of different traversal strategies for `OVAL`-optimal robust explanations on `CNN-7`. Results are averaged over the first 3 images of the test set. Bold entries correspond to the smallest $\mathcal{C}_\mathbf{x} \cup \mathcal{U}_\mathbf{x}$.

	MNIST, $\epsilon = 0.25$				CIFAR-10, $\epsilon = \frac{16}{255}$			
Method	$\lvert\mathcal{C}_\mathbf{x} \cup \mathcal{U}_\mathbf{x}\rvert$	$\lvert\mathcal{C}_\mathbf{x}\rvert$	$\lvert\mathcal{U}_\mathbf{x}\rvert$	time [s]	$\lvert\mathcal{C}_\mathbf{x} \cup \mathcal{U}_\mathbf{x}\rvert$	$\lvert\mathcal{C}_\mathbf{x}\rvert$	$\lvert\mathcal{U}_\mathbf{x}\rvert$	time [s]
VeriX	547.00	123.33	423.67	7763.91	945.00	728.67	216.33	4974.69
VeriX+	**429.33**	196.67	232.67	4394.67	735.67	522.00	213.67	5105.86
α-FaVeX	552.33	234.67	317.67	5344.59	733.33	467.00	266.33	6532.98
FaVeX-IBP	456.67	207.33	249.33	4446.53	**729.33**	512.67	216.67	5099.98

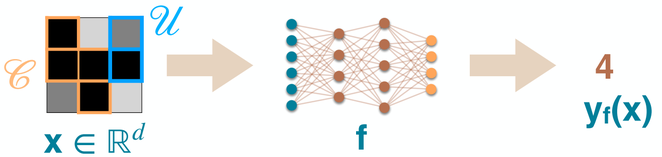
Verifier-Optimal Robust Explanations

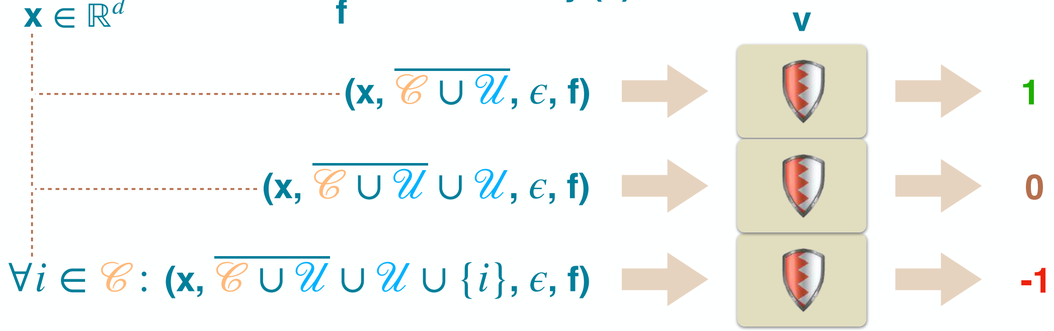
Weak Abductive Explanations



Verifier-Optimal Robust Explanations

Weak Abductive Explanations



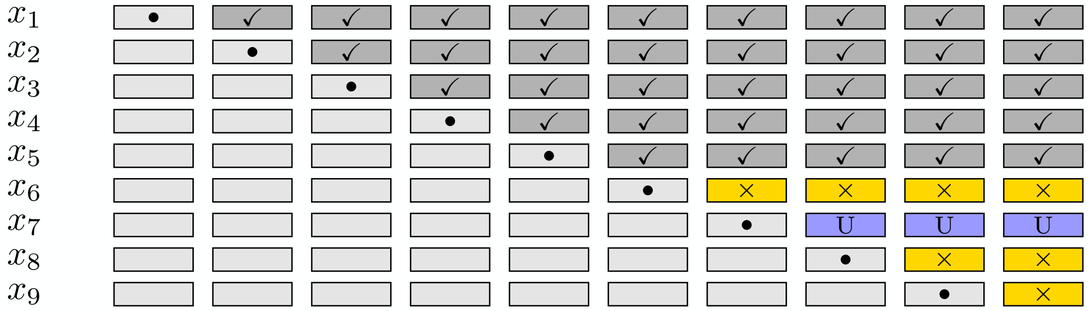


Computing Verifier-Optimal Robust Explanations

Drop (i.e., Free) Input Features While AXp Condition Holds

ADD TO $\overline{\mathcal{C} \cup \mathcal{U}}$

LOCAL ROBUSTNESS IN $B^{\epsilon}_{\overline{\mathcal{C} \cup \mathcal{U}}}(\mathbf{x})$



Computing Verifier-Optimal Robust Explanations

CNN-3

	MNIST, ϵ=0.25			CIFAR-10, ϵ=8/255		
	Counterfactuals	Unknowns	Time	Counterfactuals	Unknowns	Time
Standard	0.00	247.80	45m	0.00	461.00	2h 30m
Verifier-Optimal	**160.30**	94.40	**10m**	**210.40**	251.70	**19m**

160.30 + 94.40 = 254.70

210.40 + 251.70 = 462.10

THANKS!