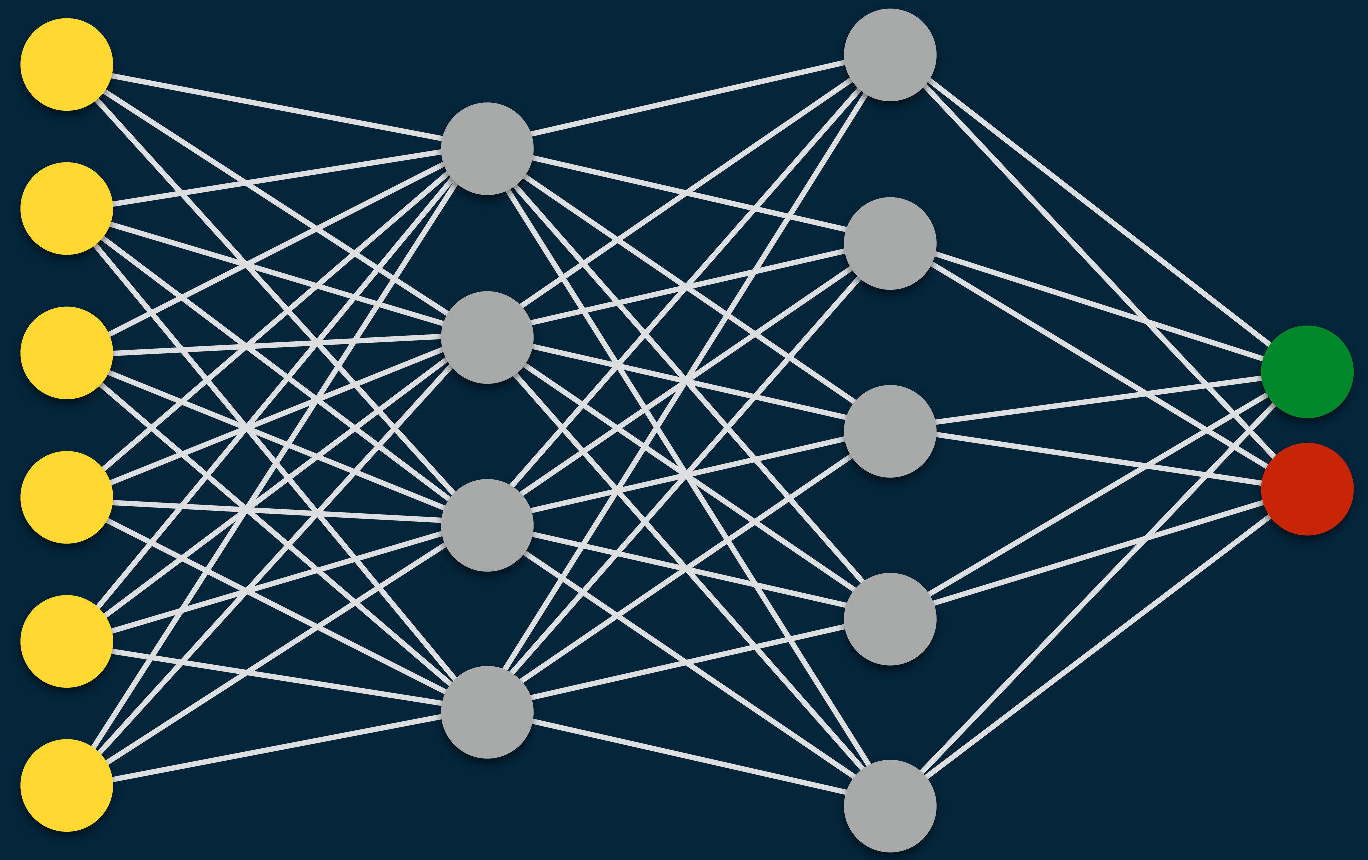
Machine Learning Interpretability and Verification



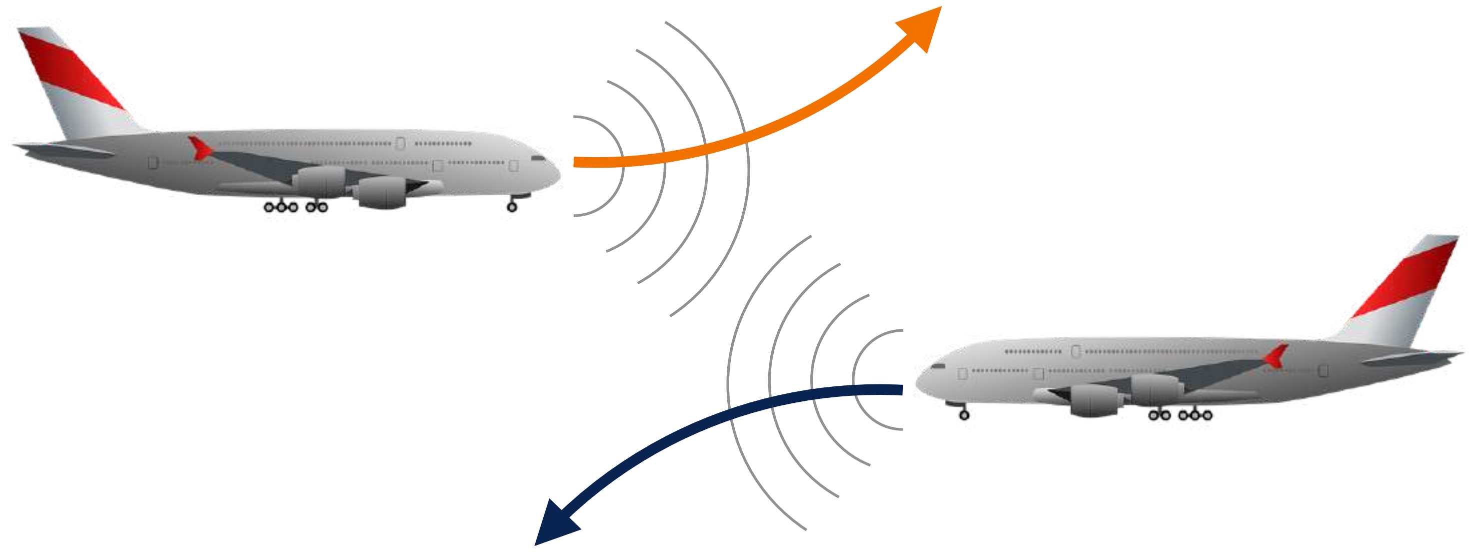
Caterina Urban
Inria & École Normale Supérieure

Machine Learning in High-Stakes Systems

data

ML software

perform tasks that are impossible using explicit programming

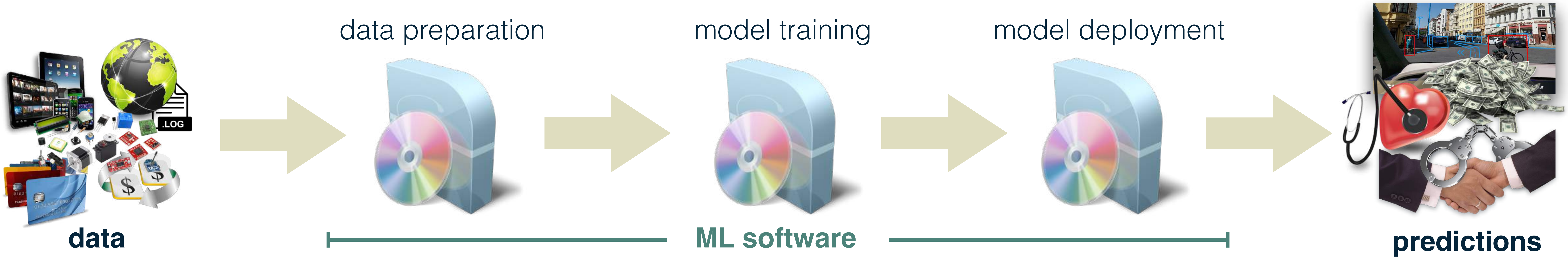


act as surrogate model

automate decision-making

Machine Learning Development Process

Machine Learning Pipeline



Models Only Give Probabilistic Guarantees

data

https://xkcd.com/2451/

model deployment

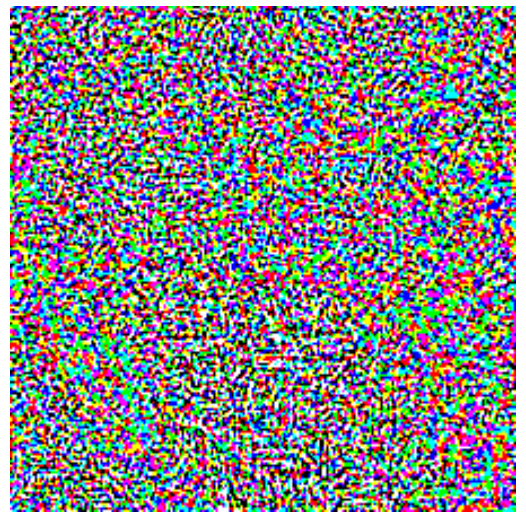
predictions

not sufficient for guaranteeing
an acceptable failure rate
under all circumstances

Models Only Give Probabilistic Guarantees

Adversarial Examples

$+\,.007\,\times$



$=$

$\boldsymbol{x}$	$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$	$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"panda"	"nematode"	"gibbon"
57.7% confidence	8.2% confidence	99.3 % confidence

Published as a conference paper at ICLR 2015

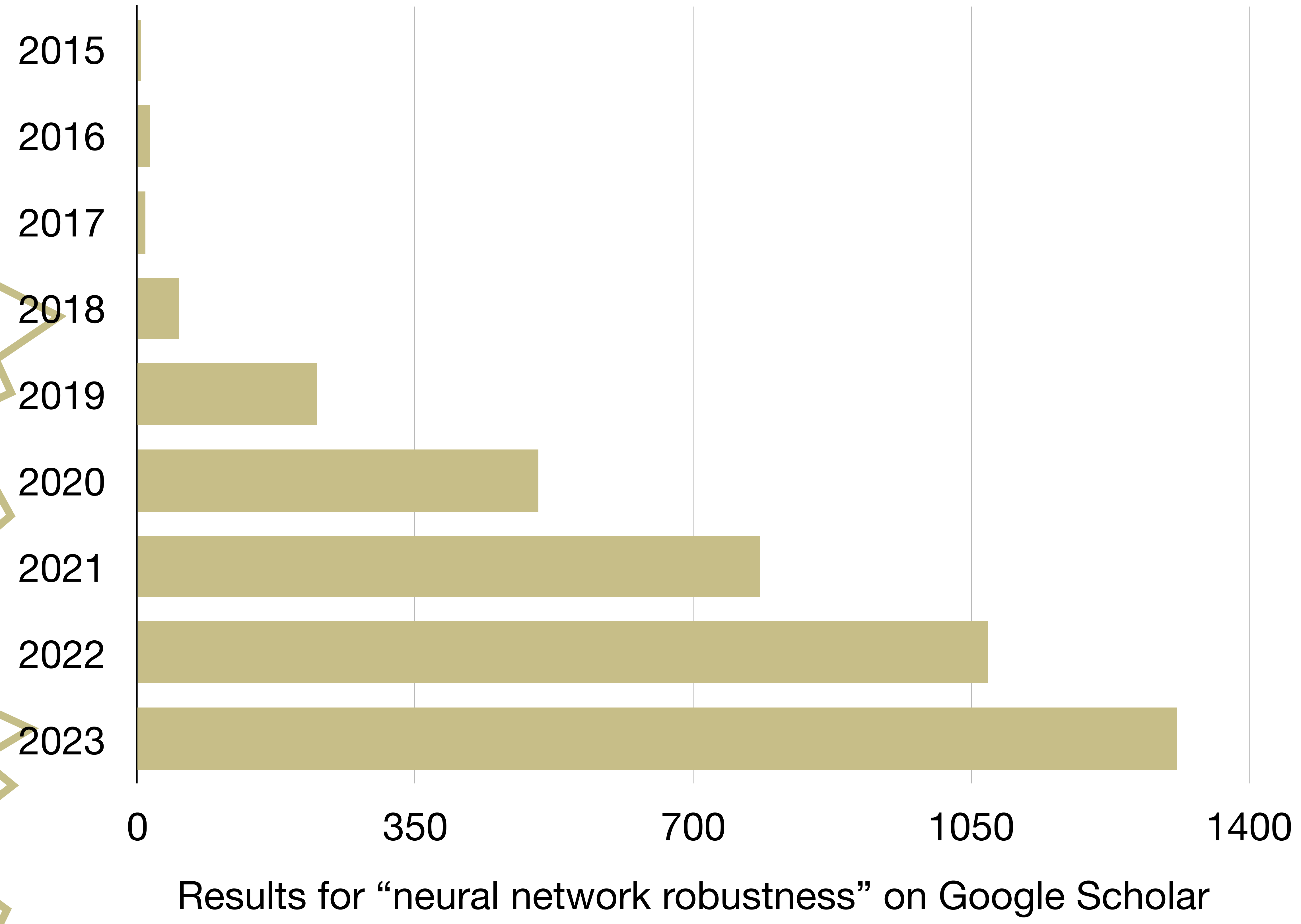
EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy
Google Inc., Mountain View, CA
{goodfellow,shlens,szegedy}@google.com


FORMAL
METHODS
COMMUNITY
TRADITIONAL
SAFETY-CRITICAL
SOFTWARE
MACHINE
LEARNING
SOFTWARE

Models Only Give Probabilistic Guarantees

Local Robustness Verification

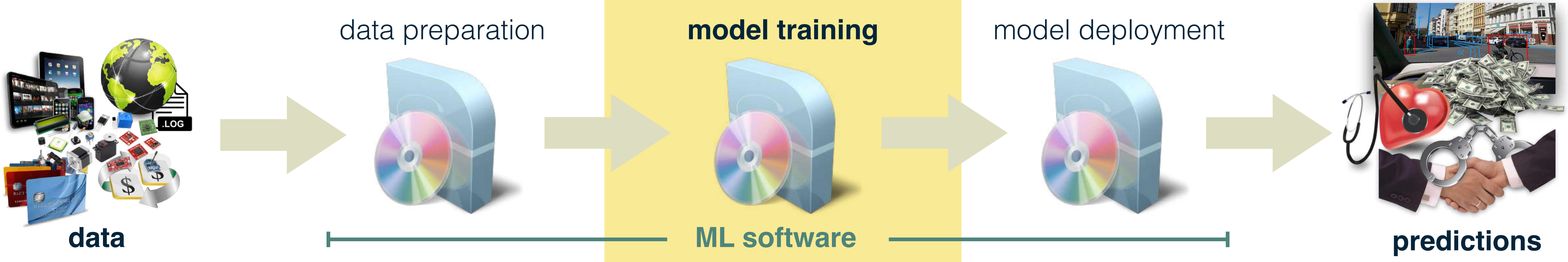


Machine Learning Community

Formal Methods Community

Machine Learning Development Process

Machine Learning Pipeline



Robust Training

Minimizing the Worst-Case Loss for Each Input

loss function (e.g, cross-entropy)

expected value

$$\min_{\boldsymbol{\theta}} \; \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \left[\max_{\mathbf{x}'\in\mathcal{C}(\mathbf{x})} \mathcal{L}(f(\boldsymbol{\theta}, \mathbf{x}'), \mathbf{y}) \right]$$

model parameters

data label

neural network

data point

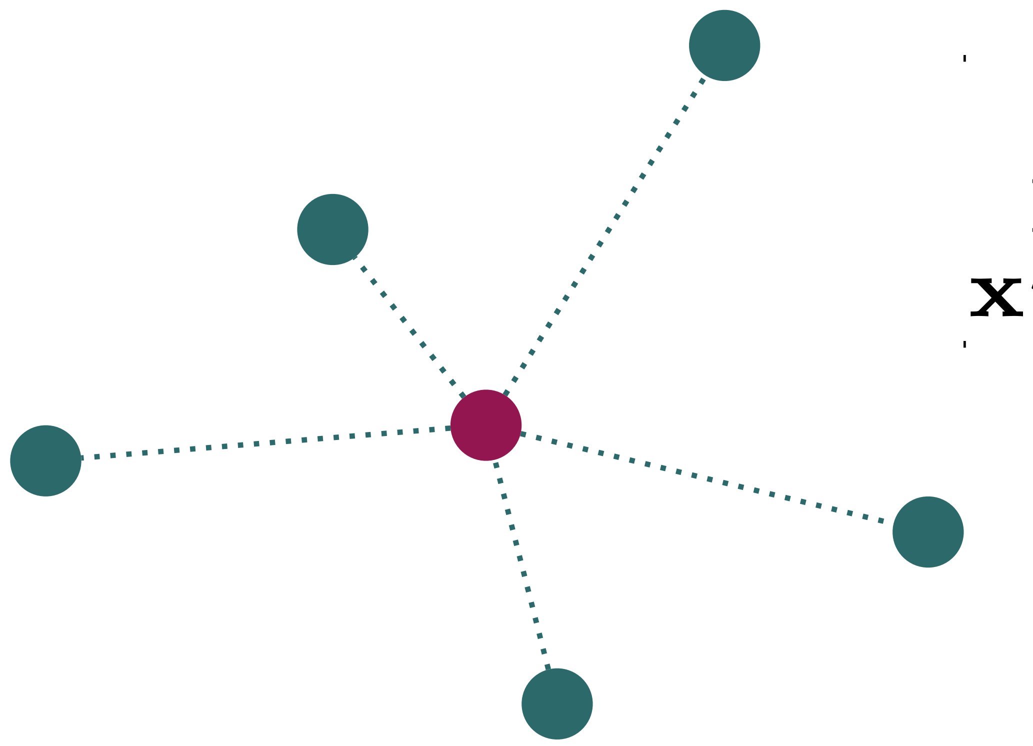
perturbation domain

Robust Training

Minimizing the Worst-Case Loss for Each Input

Adversarial Training

Minimizing a **Lower-Bound** on the Worst-Case Loss



$$\max_{\mathbf{x}' \in \mathcal{C}(\mathbf{x})} \mathcal{L}(f(\boldsymbol{\theta}, \mathbf{x}'), \mathbf{y})$$

$$\geq$$

$$\mathcal{L}(f(\boldsymbol{\theta}, \mathbf{x}_{\mathrm{adv}}), y)$$

Machine Learning Community

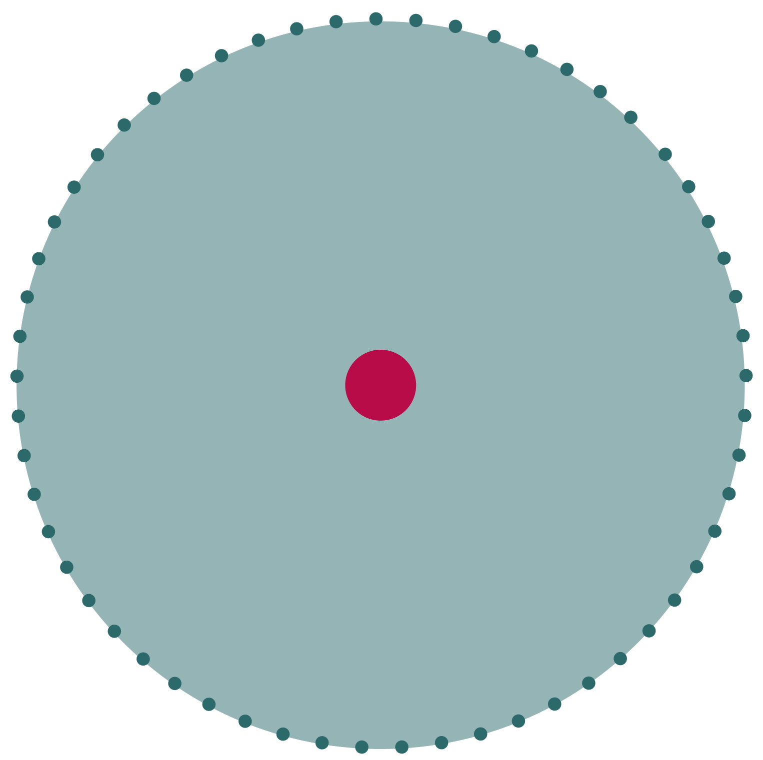
Certified Training

Minimizing an **Upper-Bound** on the Worst-Case Loss

$$\max_{\mathbf{x}' \in \mathcal{C}(\mathbf{x})} \mathcal{L}(f(\boldsymbol{\theta}, \mathbf{x}'), \mathbf{y})$$

$$\leq$$

$$\mathcal{L}_{\mathrm{ver}}(f(\boldsymbol{\theta}, \mathbf{x}), y)$$



Formal Methods Community

Robust Training

Certified Training

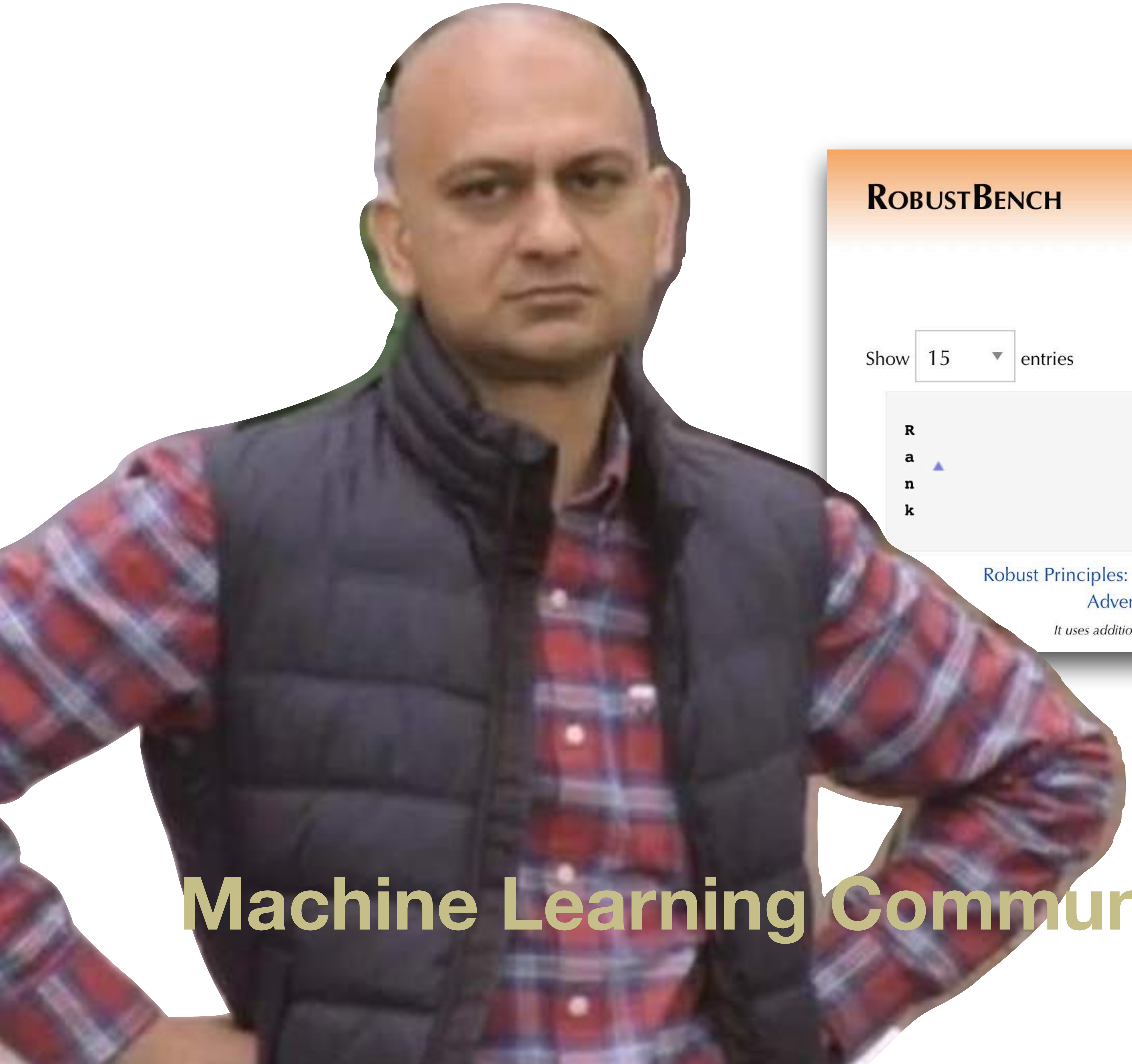


Table 7: Comparison of the standard (Acc.), adversarial (Adv. Acc), and certified (Cert. Acc.) accuracy for different certified training methods on the full CIFAR-10 test set. We use MN-BAB (Ferrari et al., 2022) to compute all certified and adversarial accuracies.

ϵ_∞	Training Method	Source	Acc. [%]	Adv. Acc. [%]	Cert. Acc. [%]
2/255	COLT	Balunovic & Vechev (2020)	78.42	**66.17**	61.02
	CROWN-IBP	Zhang et al. (2020)†	71.27	59.58	58.19
	IBP	Shi et al. (2021)	-	-	-
	SABR	this work	**79.52**	65.76	**62.57**
8/255	COLT	Balunovic & Vechev (2020)	51.69	31.81	27.60
	CROWN-IBP	Zhang et al. (2020)†	45.41	33.33	33.18
	IBP	Shi et al. (2021)	48.94	35.43	**35.30**
	SABR	this work	**52.00**	**35.70**	35.25

ROBUSTBENCH Leaderboards Paper FAQ Contribute Model Zoo 🚀

Leaderboard: CIFAR-10, $\ell_\infty = 8/255$, untargeted attack

Show 15 entries

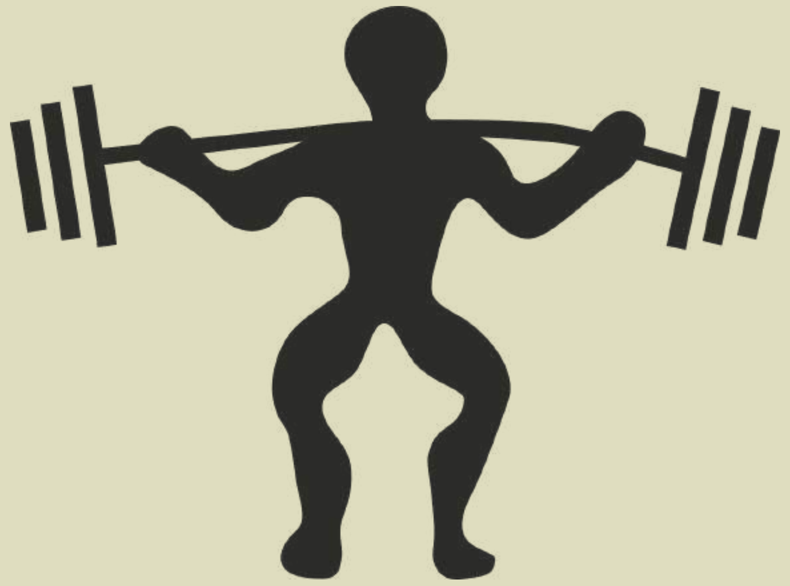
Search: Papers, architectures, v

Rank	Method	Standard accuracy	AutoAttack robust accuracy	Best known robust accuracy	AA eval. potentially unreliable	Extra data	Architecture	Venue
	Robust Principles: Architectural Design Principles for Adversarially Robust CNNs *It uses additional 50M synthetic images in training.*	93.27%	71.07%	71.07%	×	×	RaWideResNet-70-16	BMVC 2023

Machine Learning Community

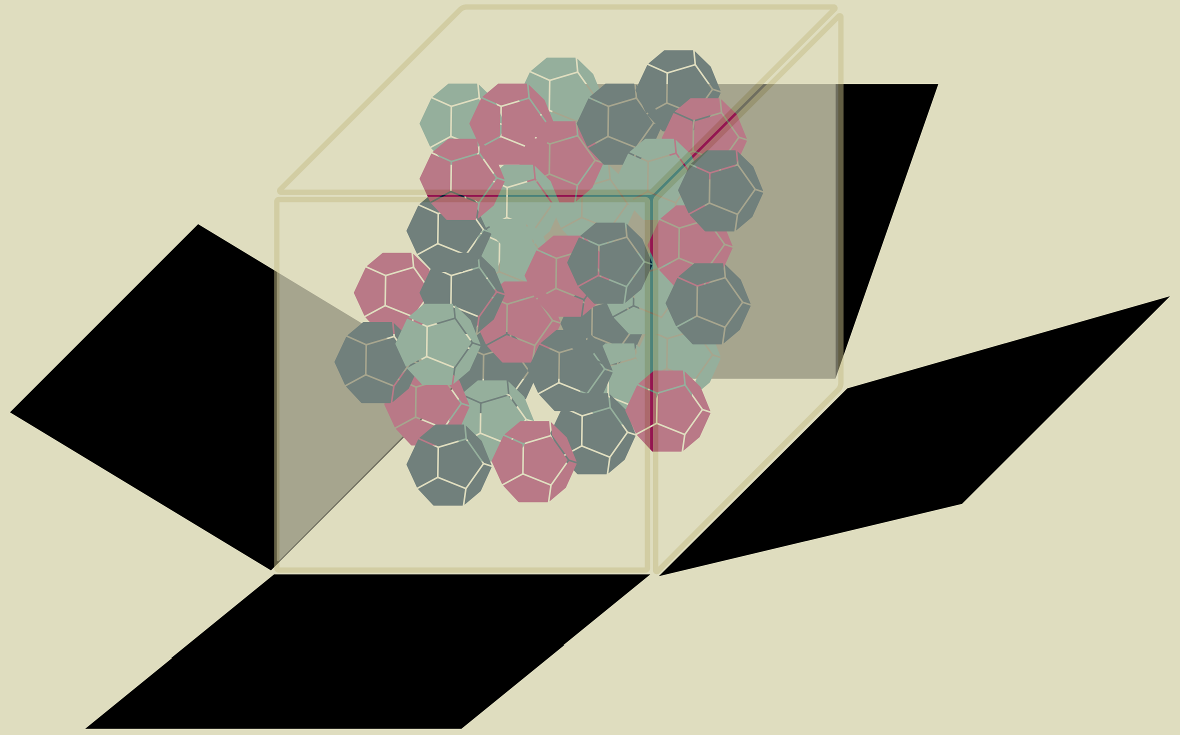
Formal Methods Community

Can we make formal methods interesting for the machine learning community?



Training

CIKM 2021

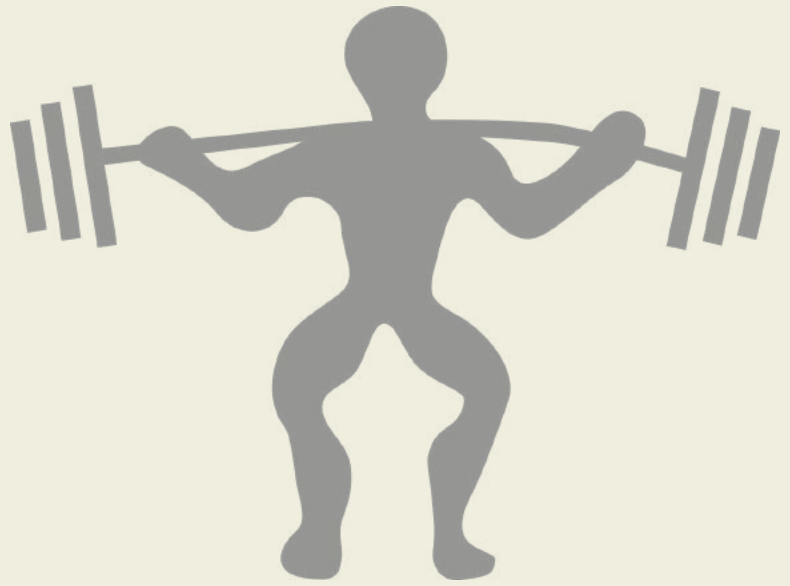


Interpretability

VMCAI 2024

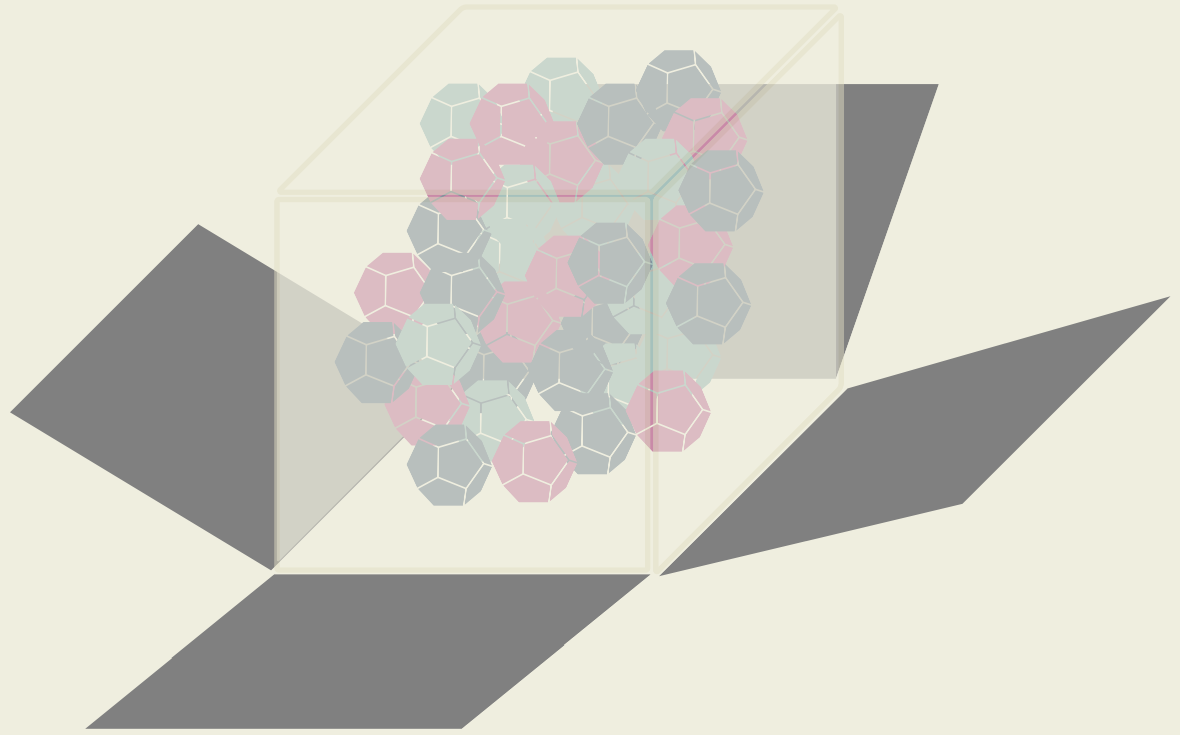
Verification

NFM 2023



Training

CIKM 2021



Interpretability

VMCAI 2024

Verification

NFM 2023

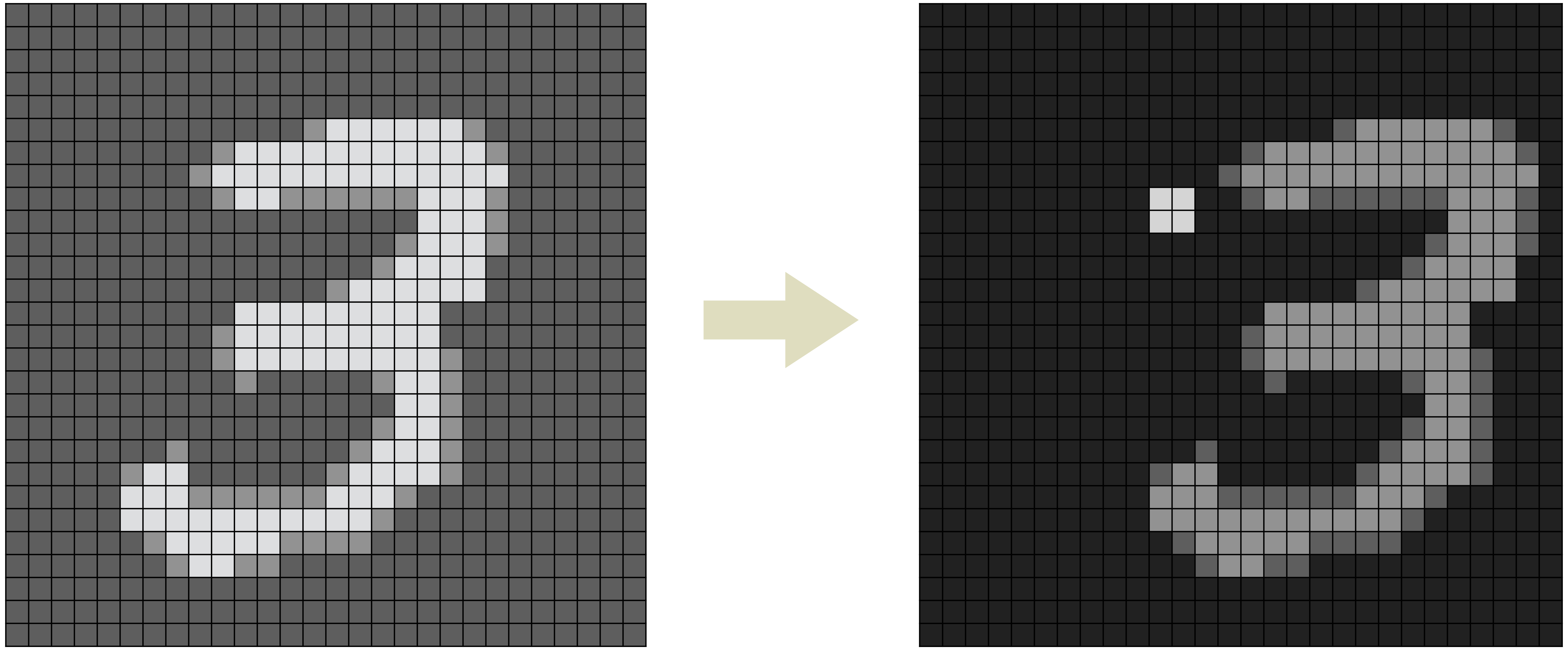
Using formal methods
for robustness verification

Using formal methods
interpretability-aware
robustness verification

Local Robustness Verification

Combinations of Semantic Perturbations

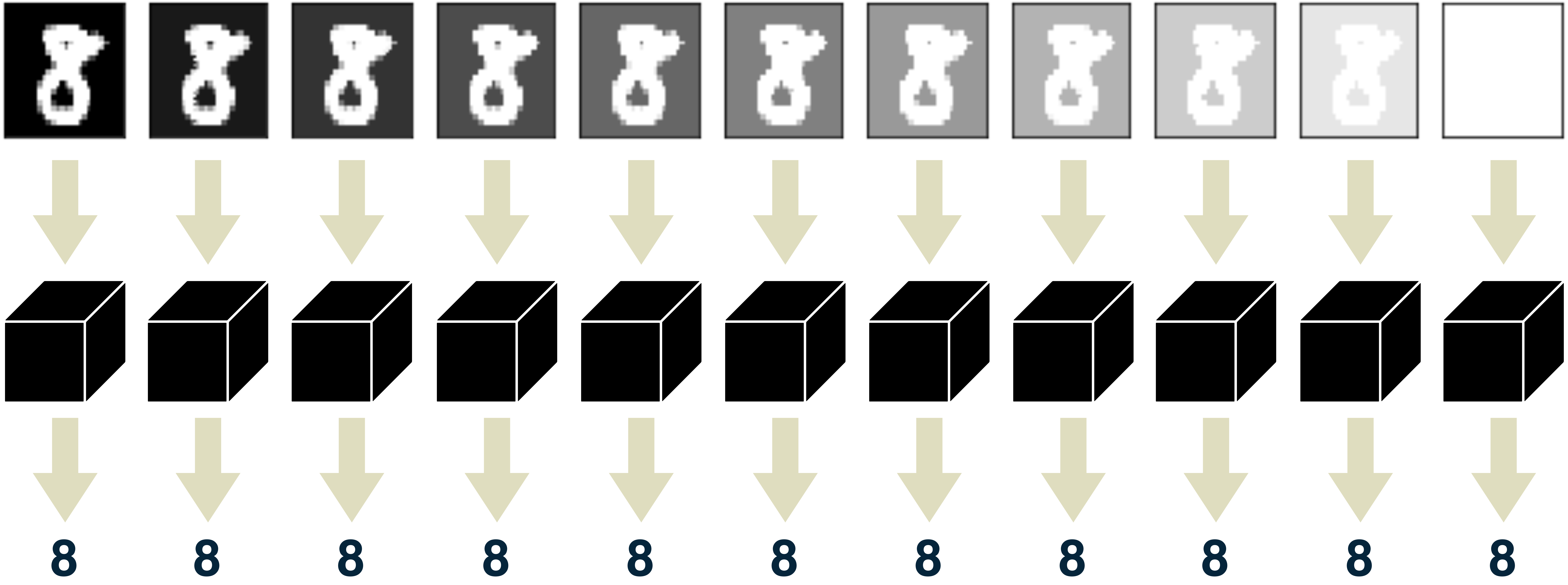
- Brightness Change
- Patch Placement
- Object Translation



Goal: Identifying Safe Ranges of Perturbation Parameters

Local Robustness Verification

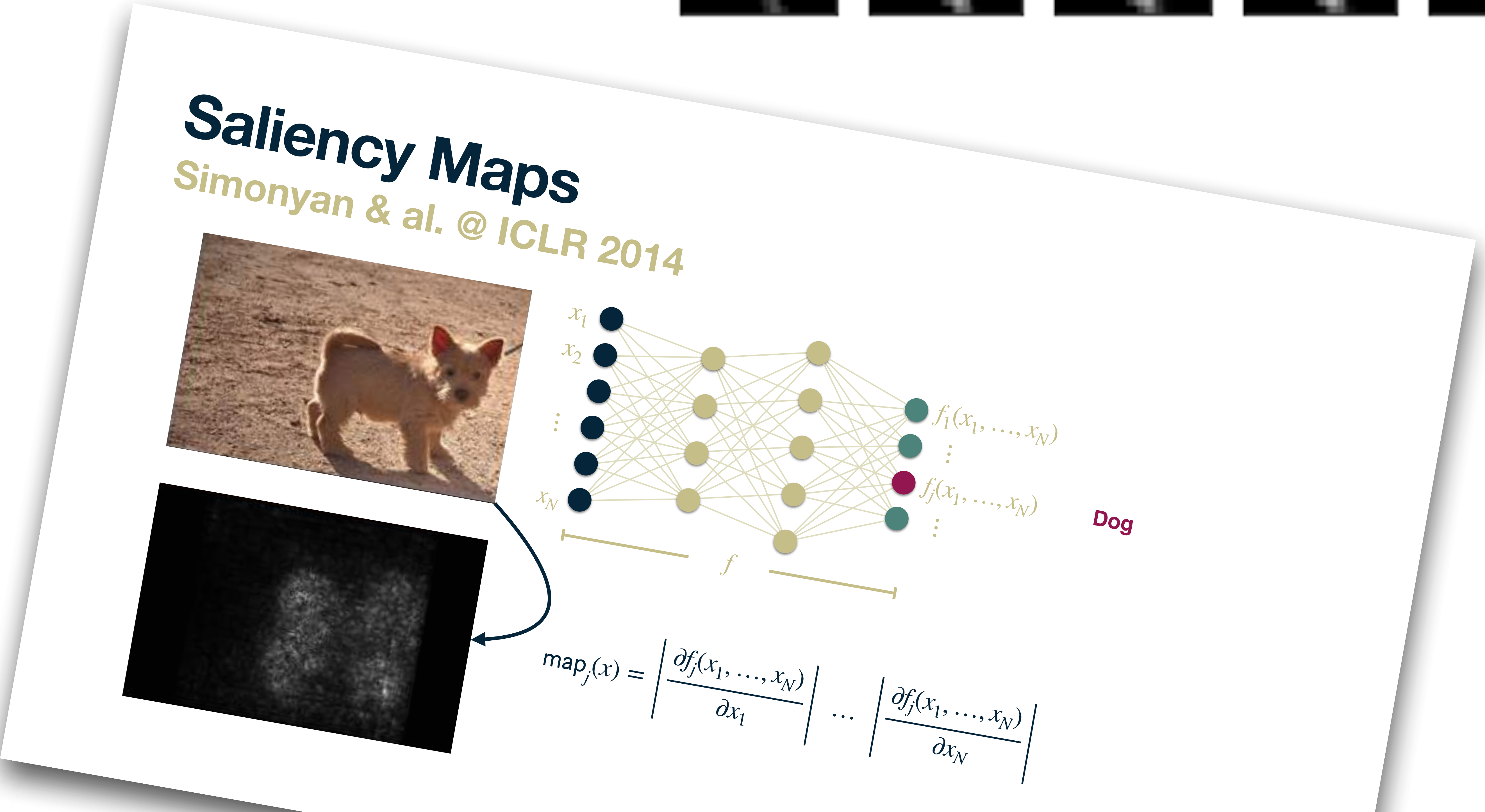
Classification Robustness is NOT ENOUGH!



Local Robustness Verification

Input Image

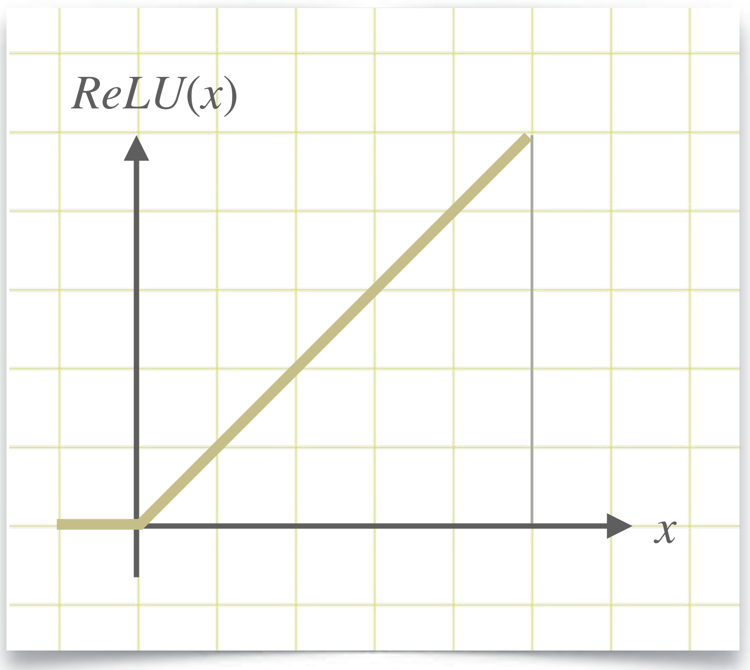
Saliency Map



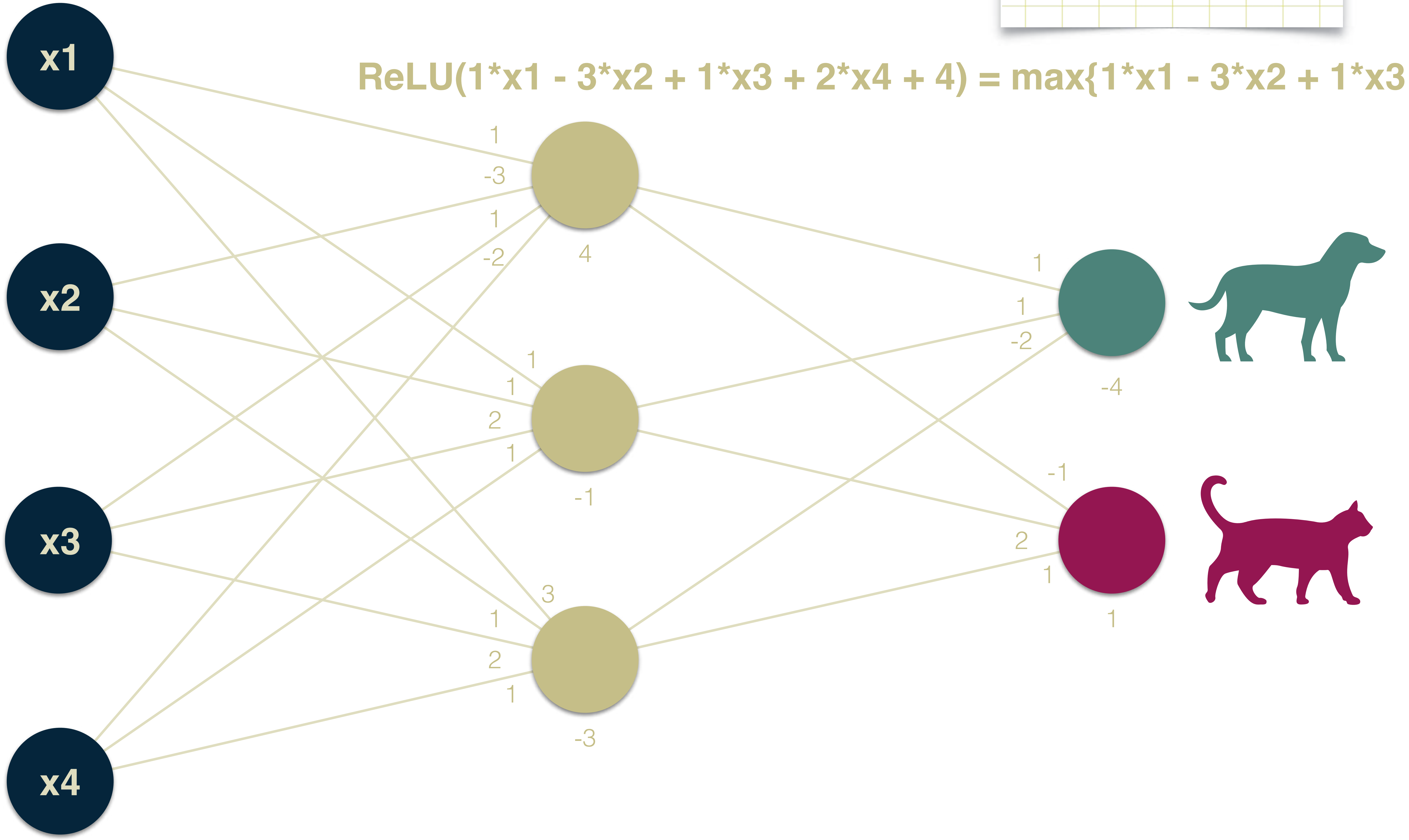
Local Robustness Verification

Saliency Map Robustness

(A Very Small) Example

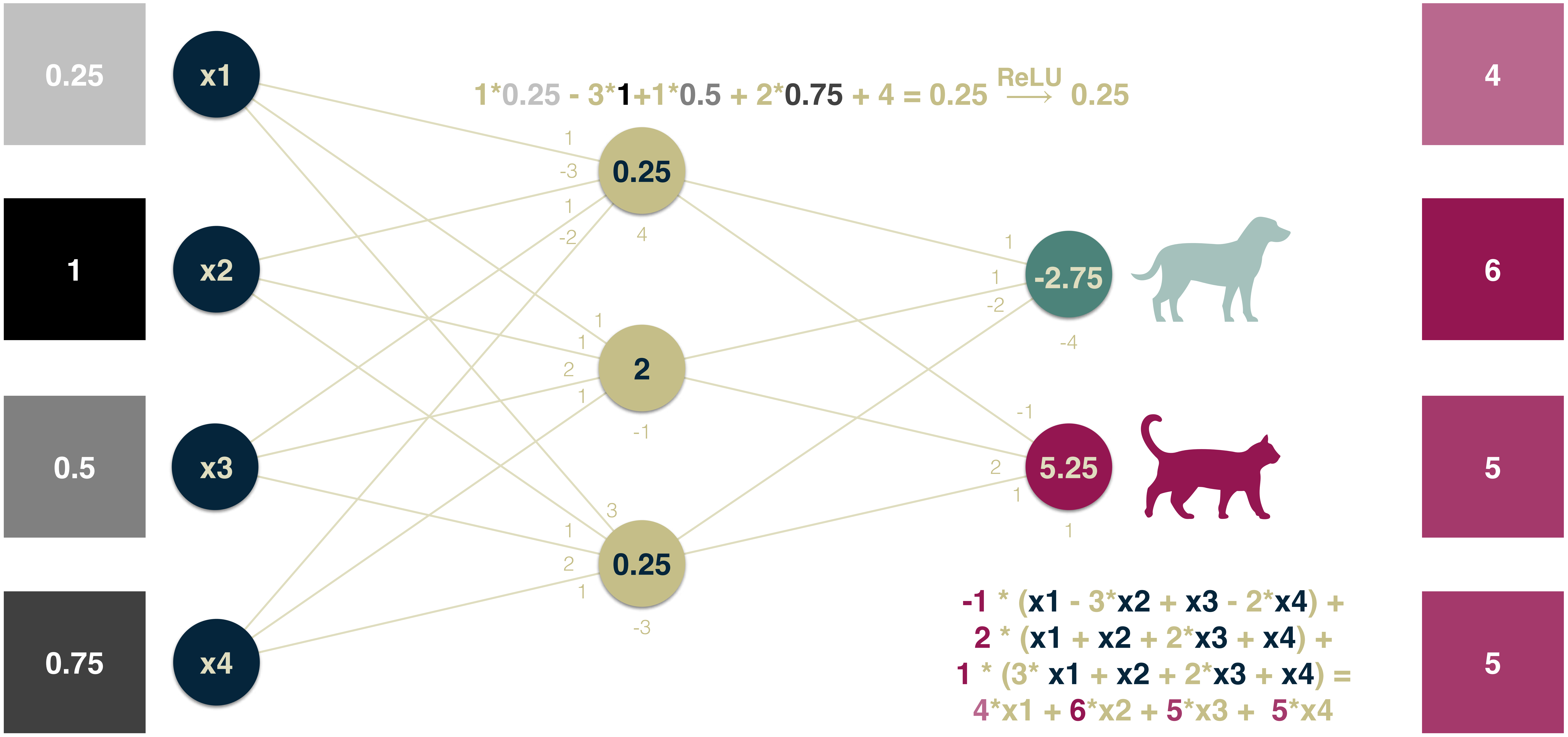


ReLU(1*x1 - 3*x2 + 1*x3 + 2*x4 + 4) = max{1*x1 - 3*x2 + 1*x3 + 2*x4 + 4, 0}



(A Very Small) Example

Saliency Maps

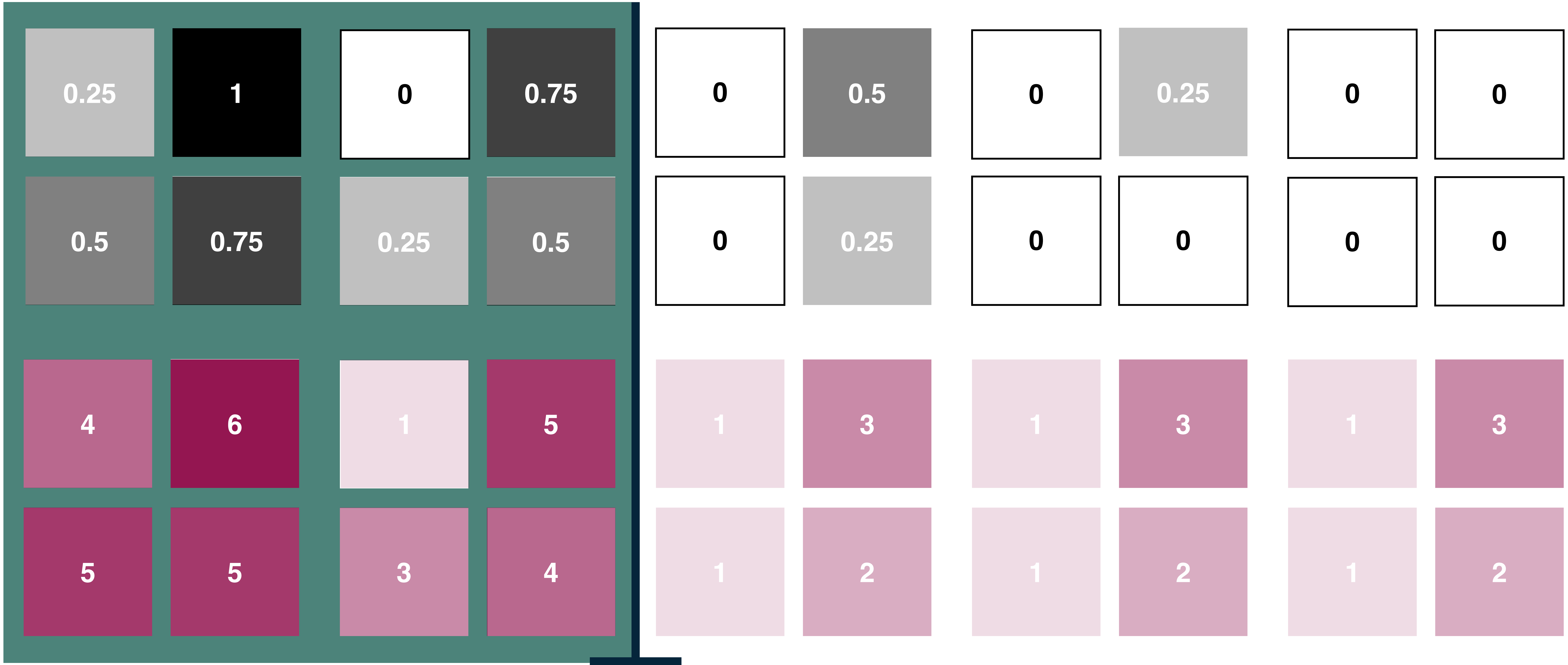


(A Very Small) Example

Semantic Perturbations

(A Very Small) Example

Saliency Map Robustness

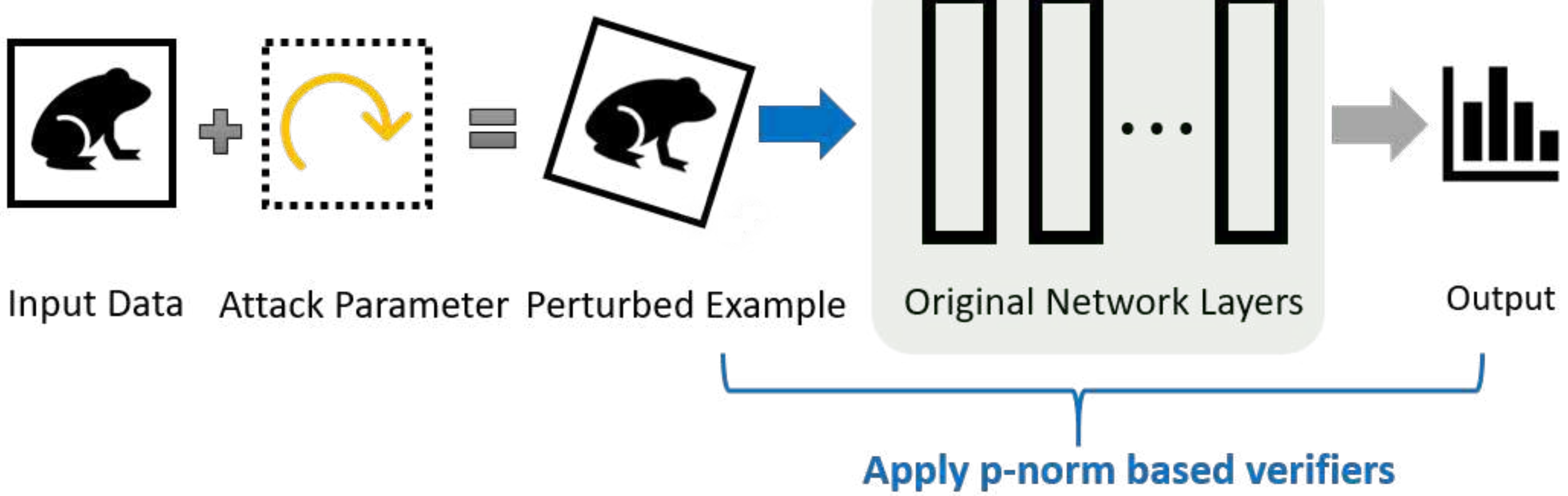


distance = 3.87 $\delta = 4$ distance = 6.56 distance = 6.56 distance = 6.56

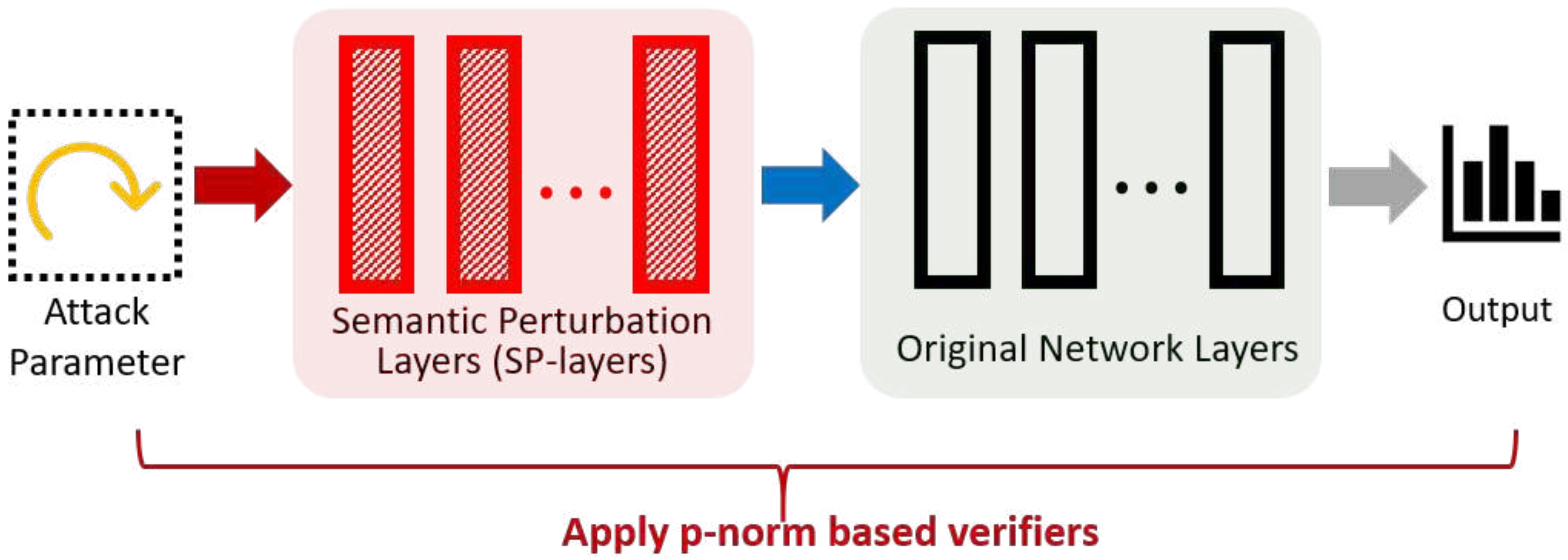
Encoding Semantic Perturbations

Mohapatra & al. @ CVPR 2020

- **Naïve verification approach:**

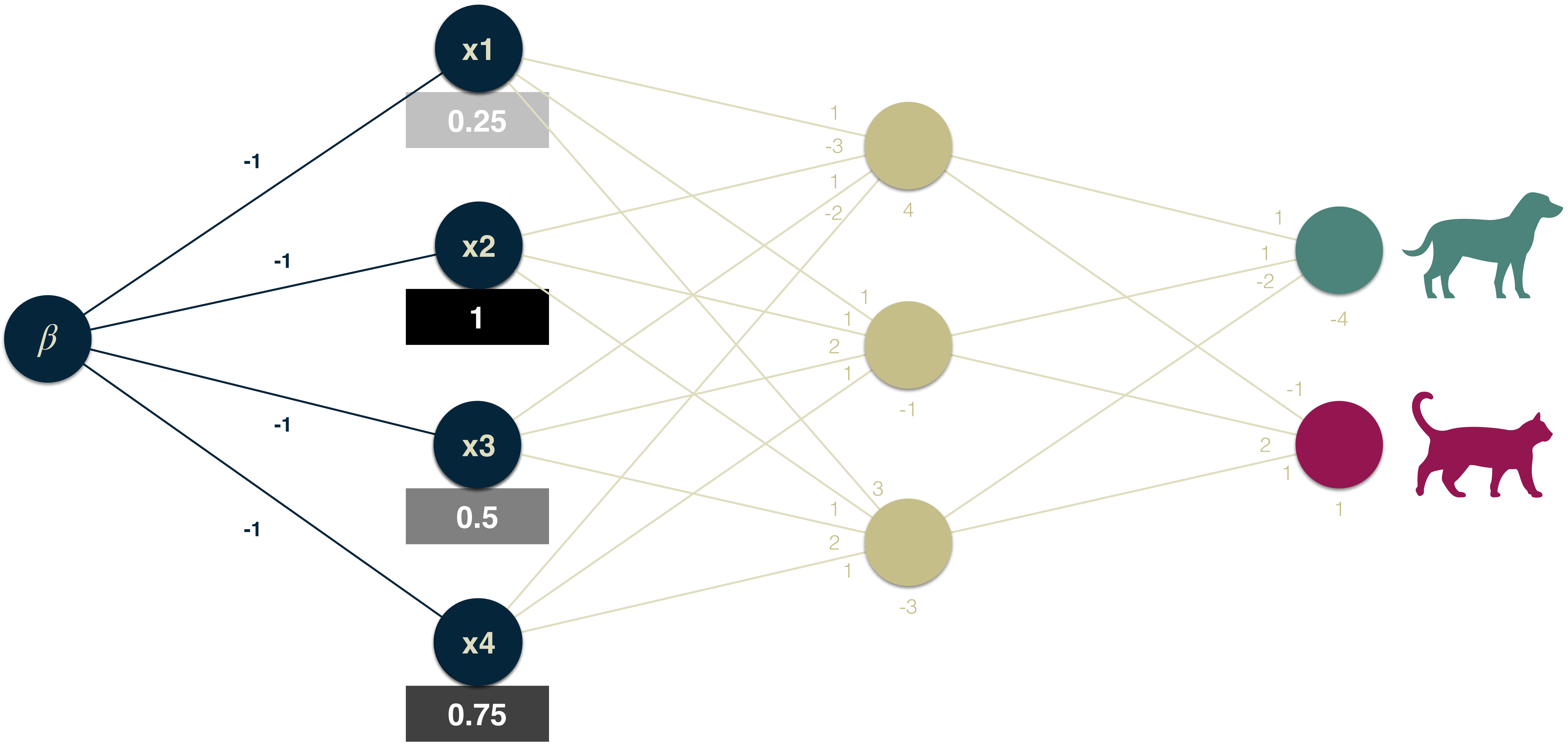


- **Our Semantify-NN:**



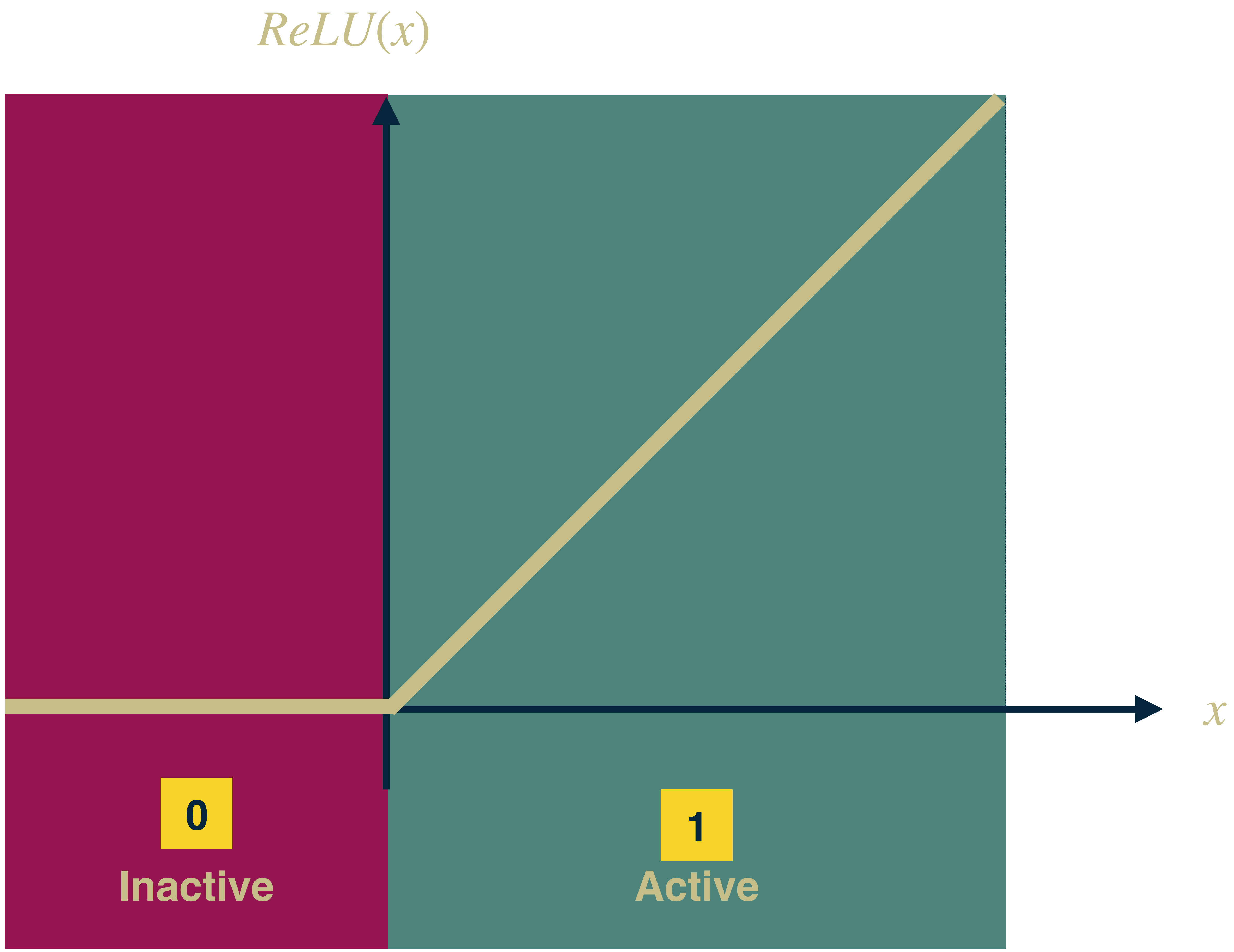
(A Very Small) Example

Encoding Semantic Perturbations



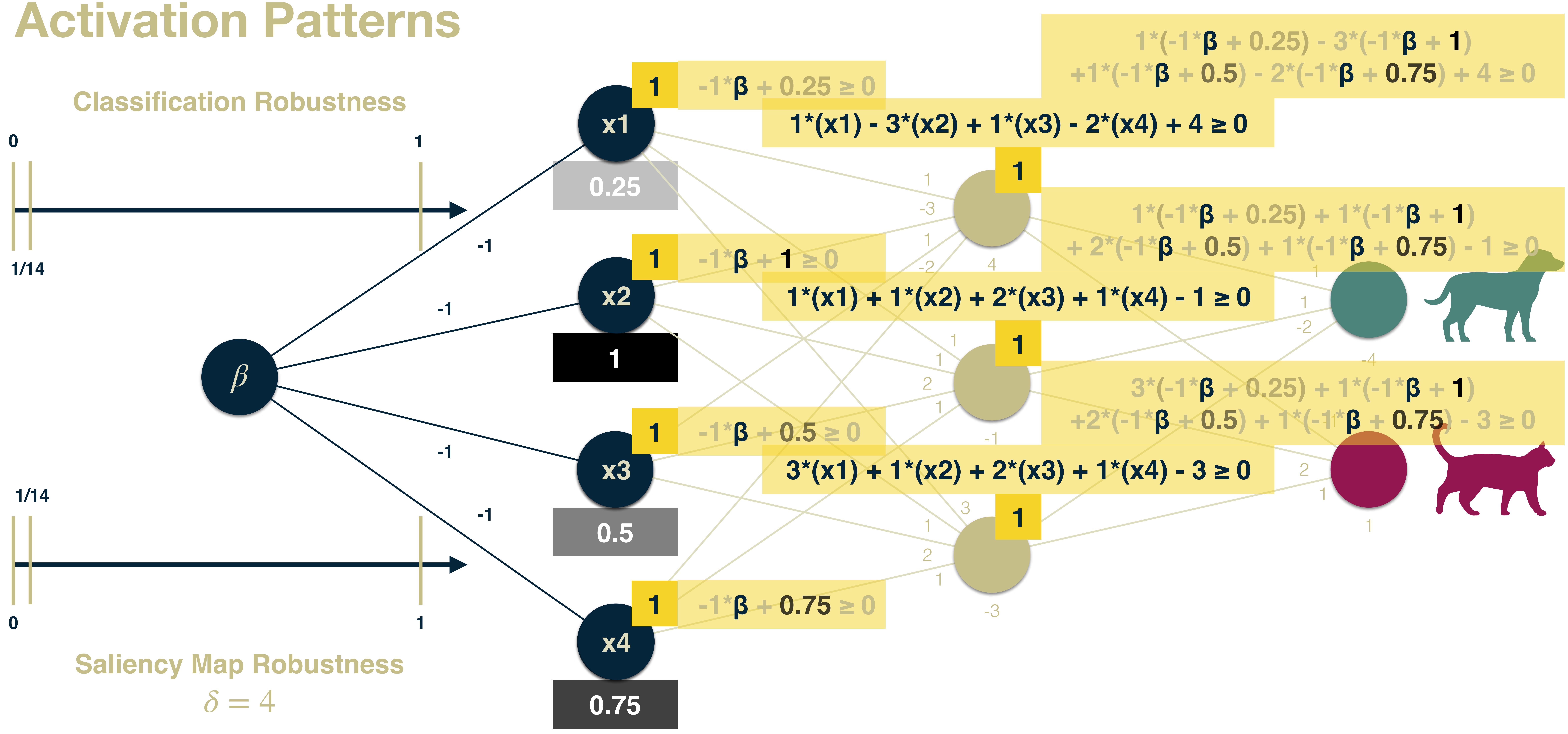
Naïve Breadth-First Search

Activation Patterns



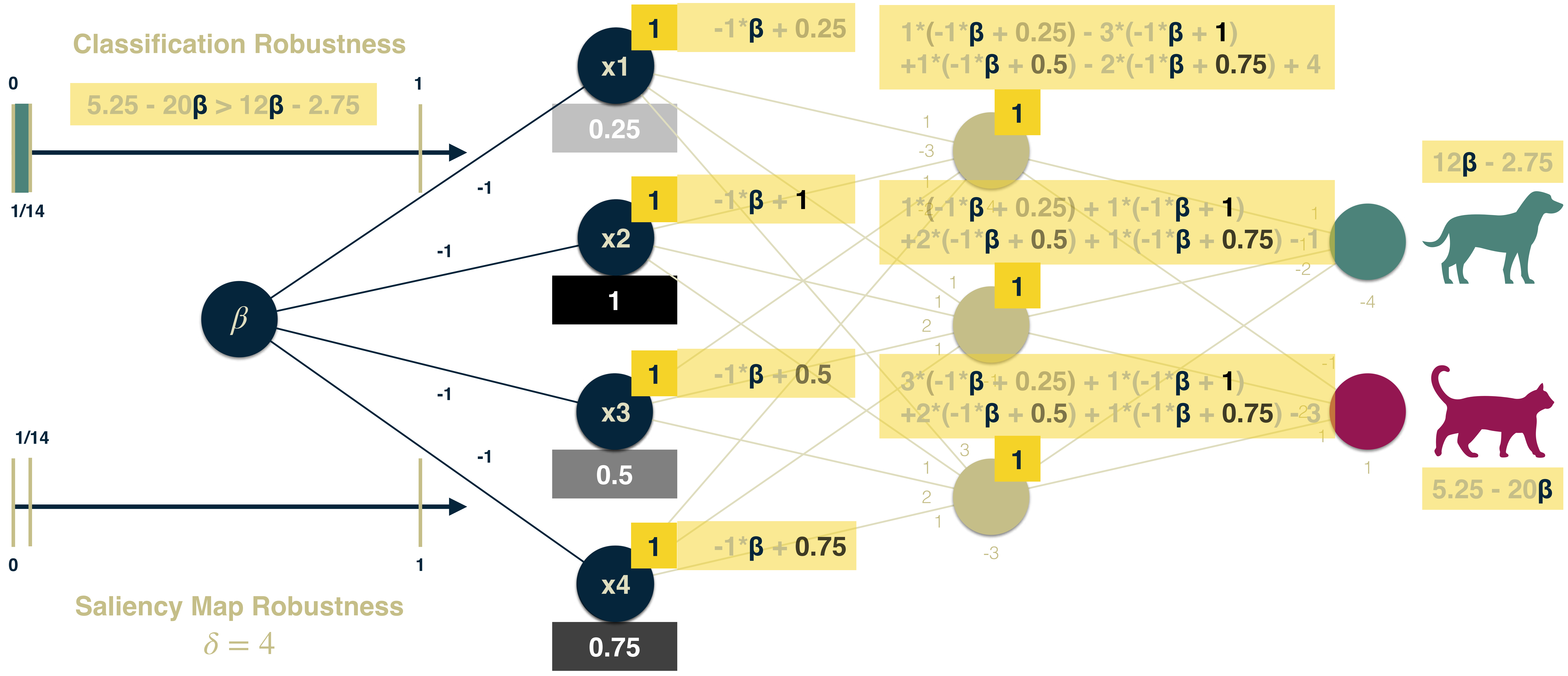
(A Very Small) Example

Activation Patterns



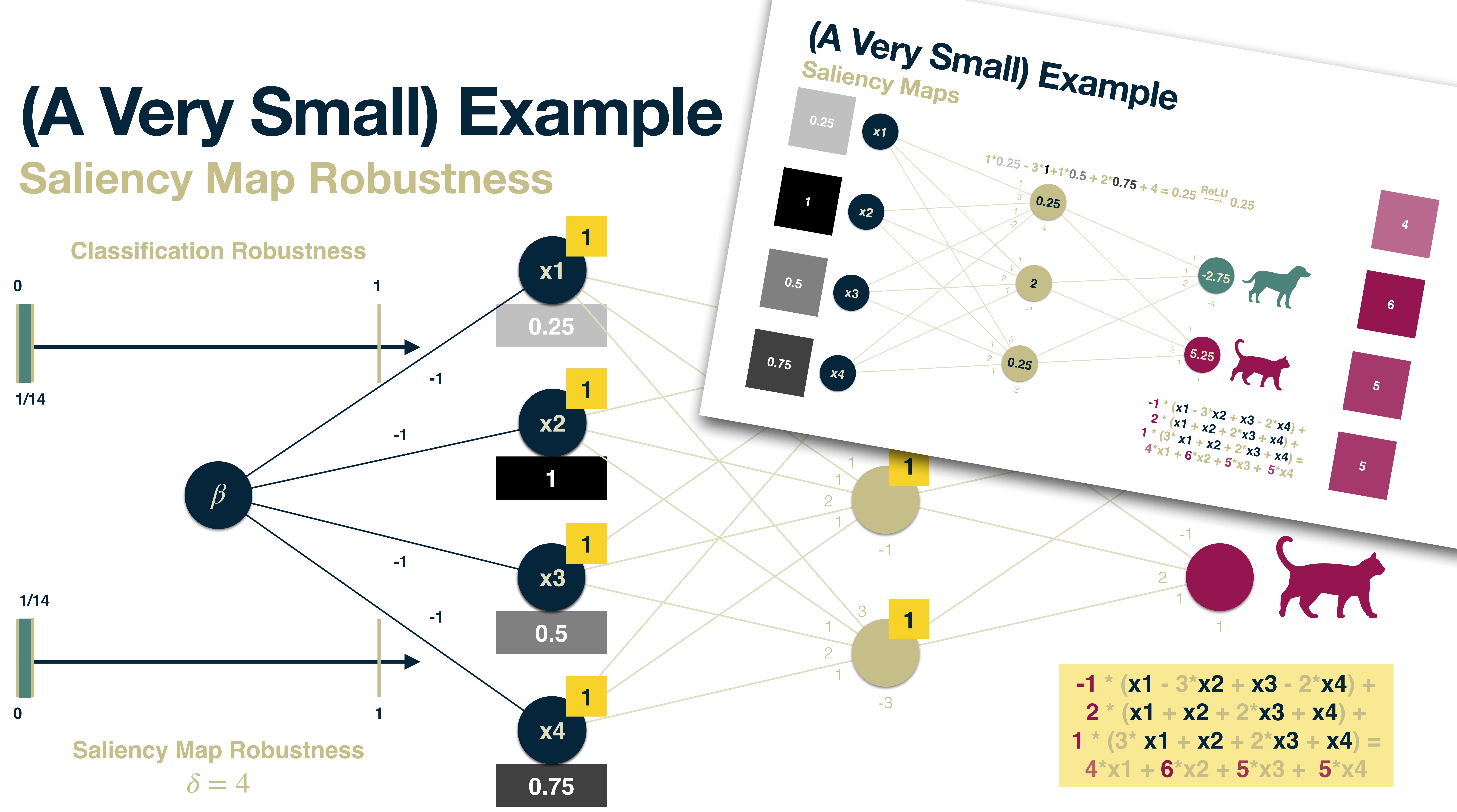
(A Very Small) Example

Classification Robustness

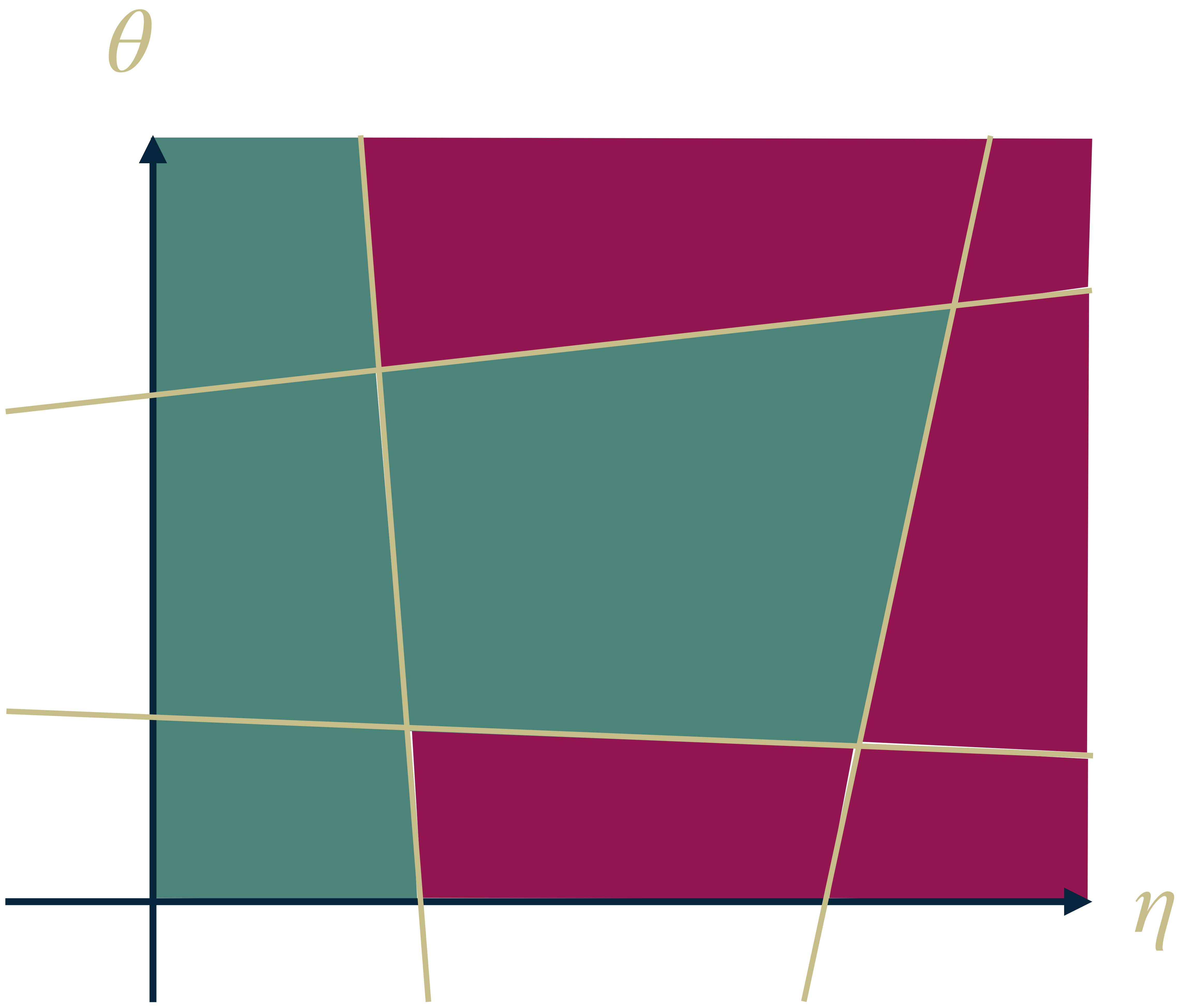


(A Very Small) Example

Saliency Map Robustness

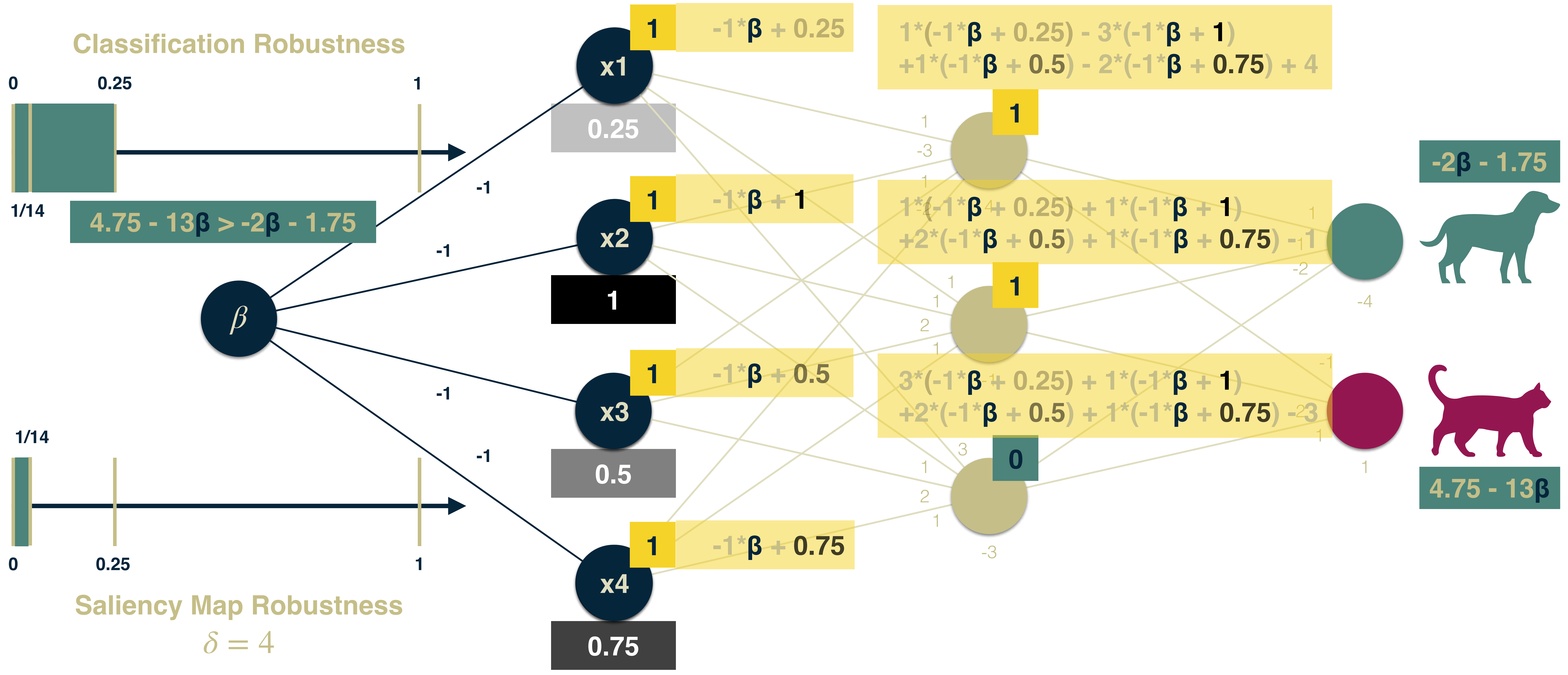


Naïve Breadth-First Search



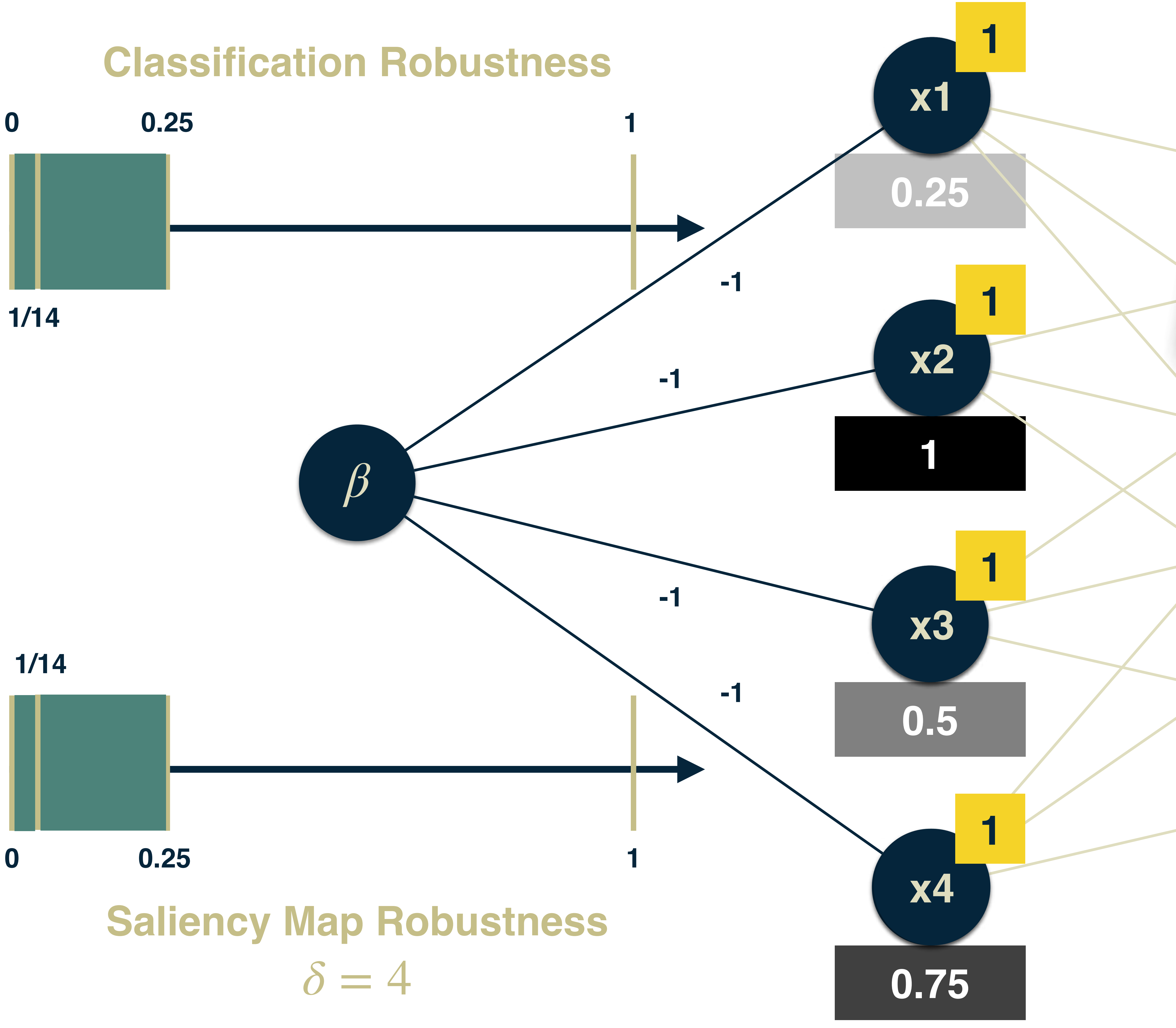
(A Very Small) Example

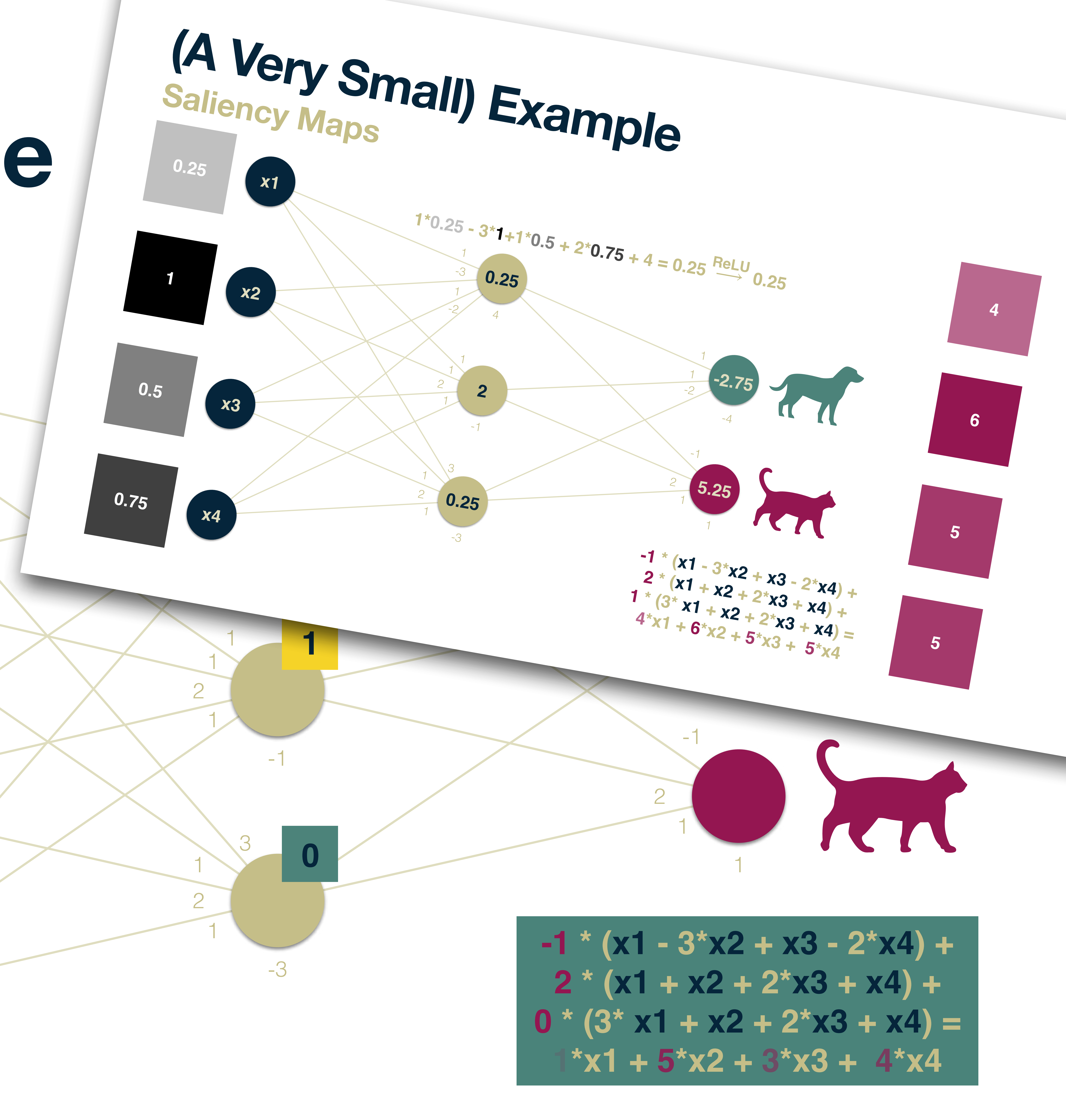
Classification Robustness



(A Very Small) Example

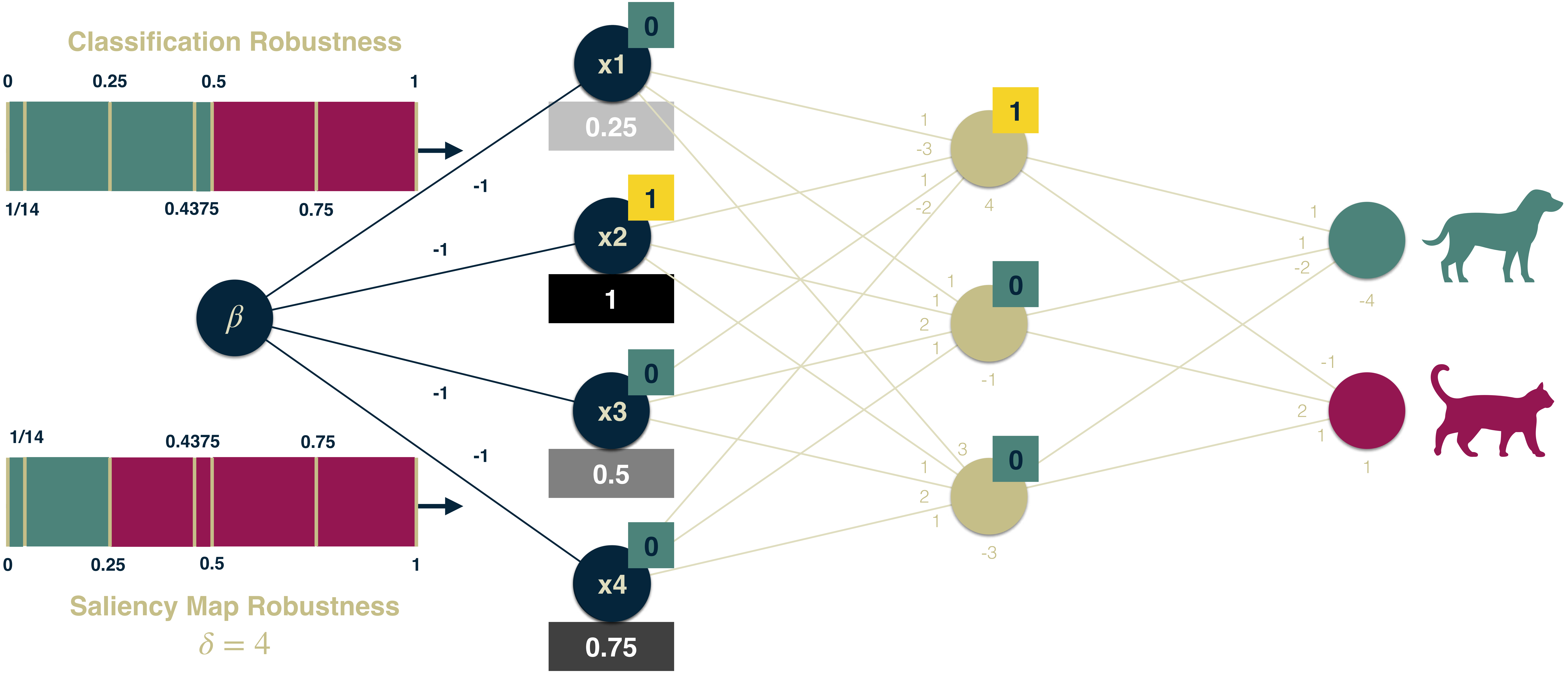
Saliency Map Robustness



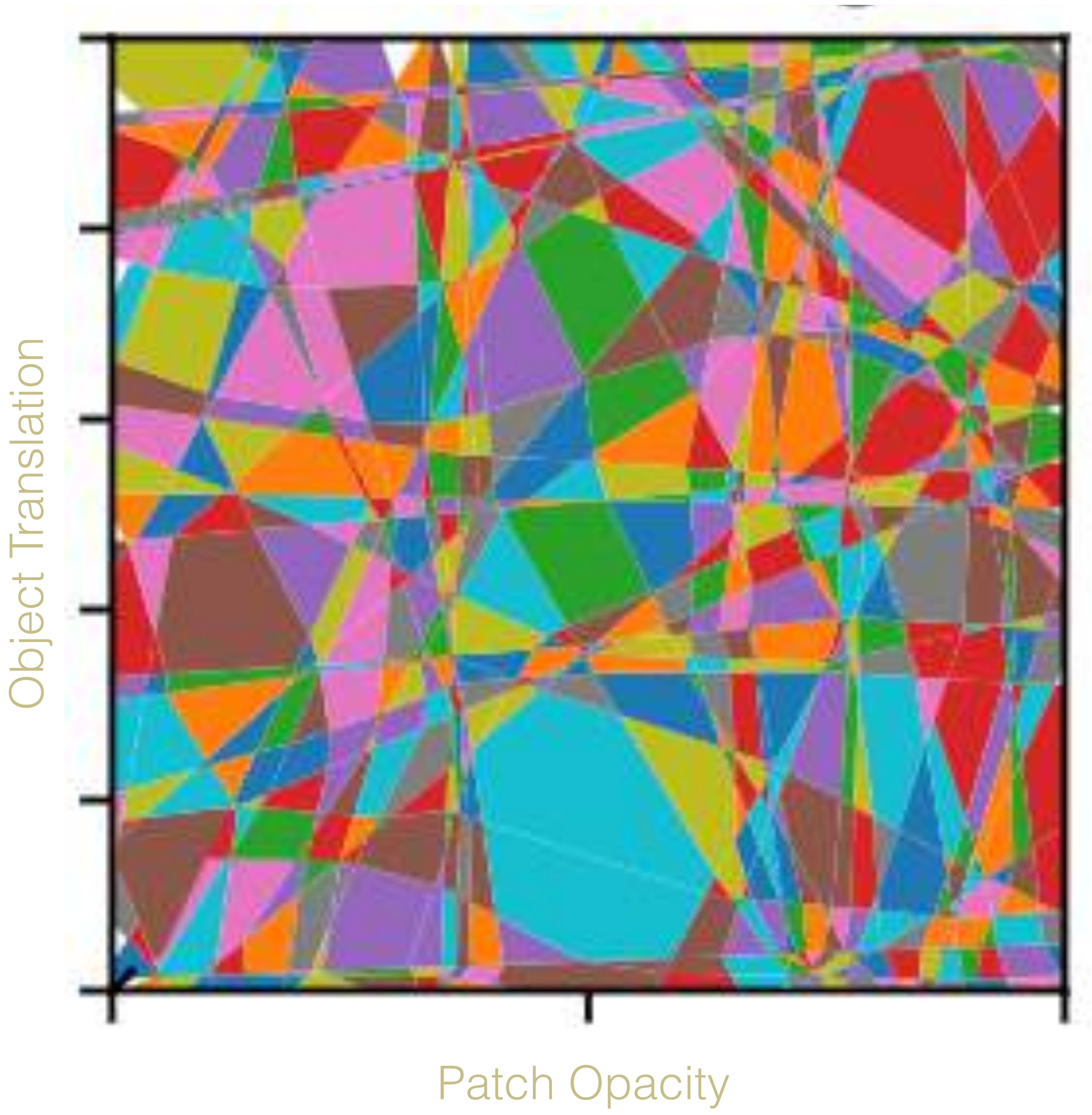


(A Very Small) Example

Naïve Breadth-First Search

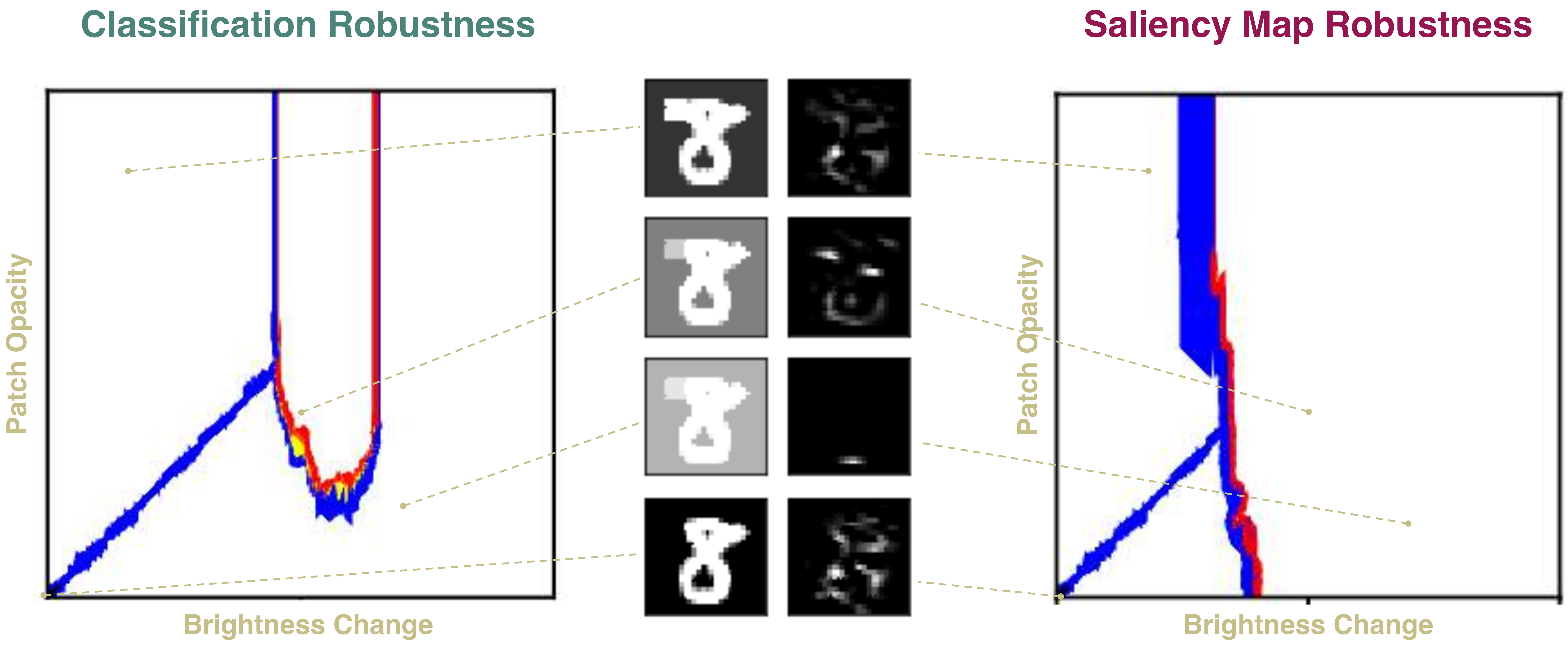


Naïve Breadth-First Search



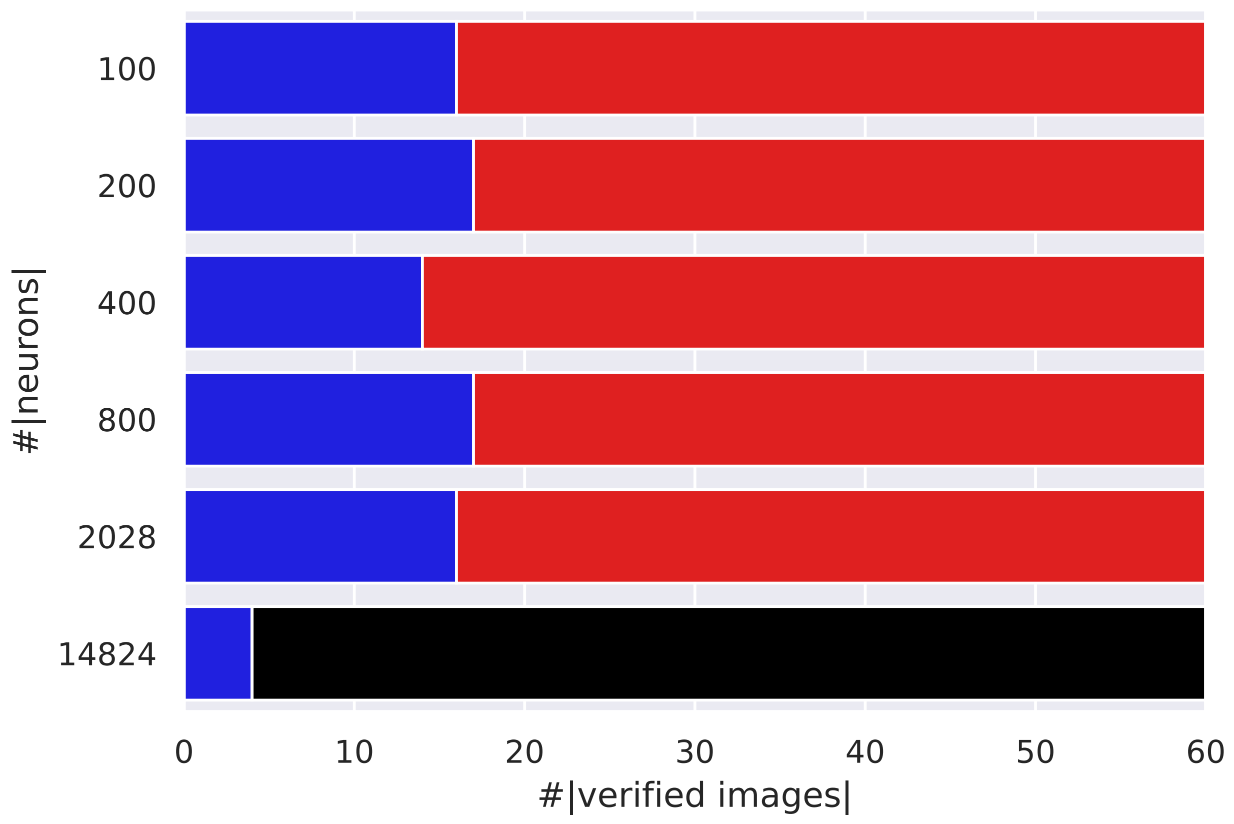
TOO MANY ACTIVATION PATTERNS!

Geometric Boundary Search

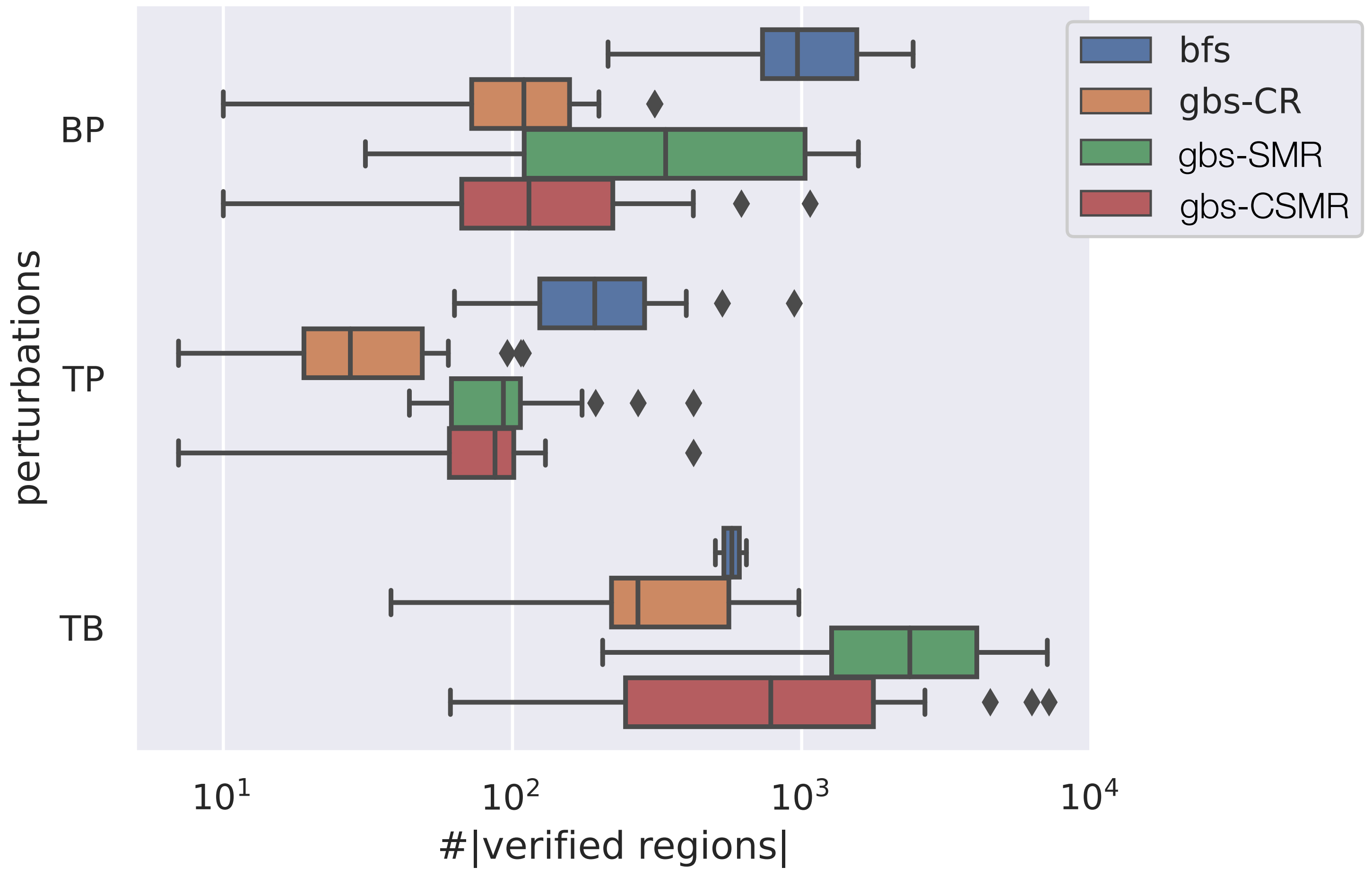


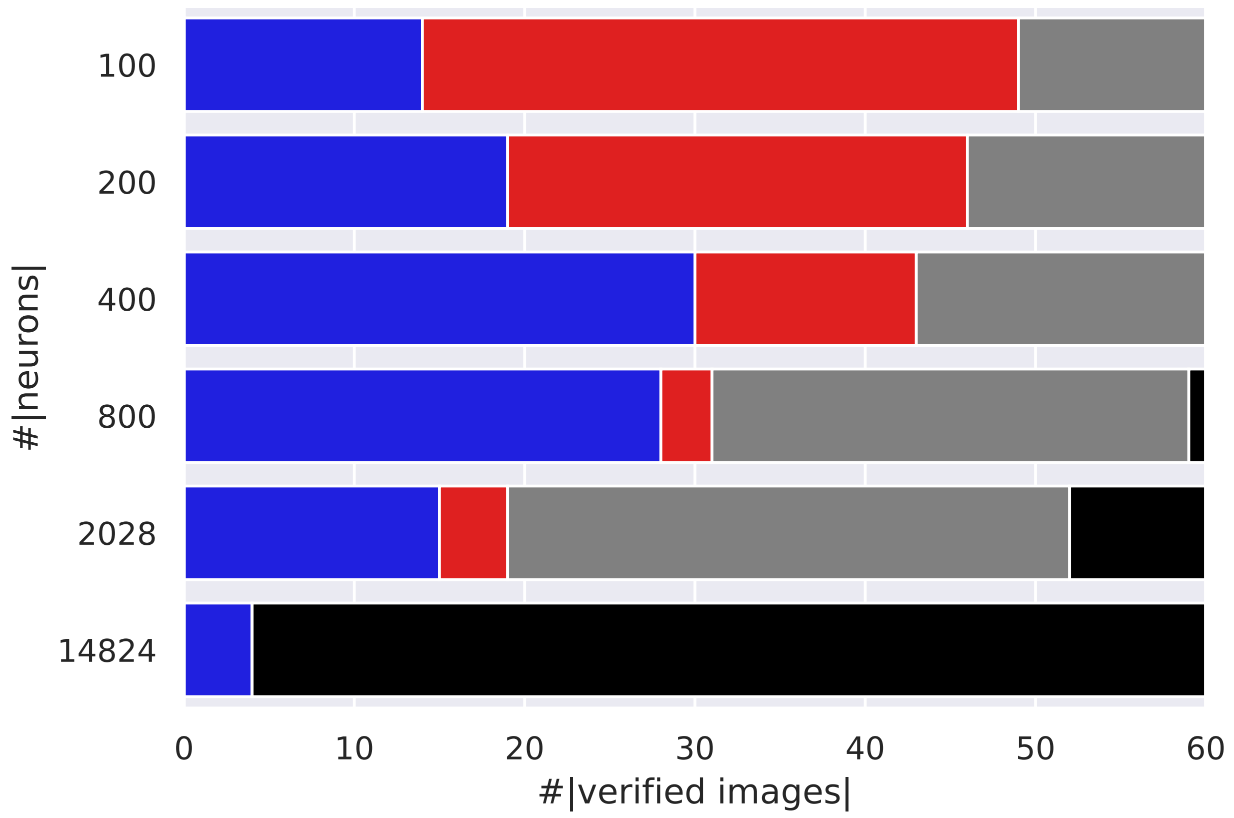
Geometric Boundary Search

Experimental Results

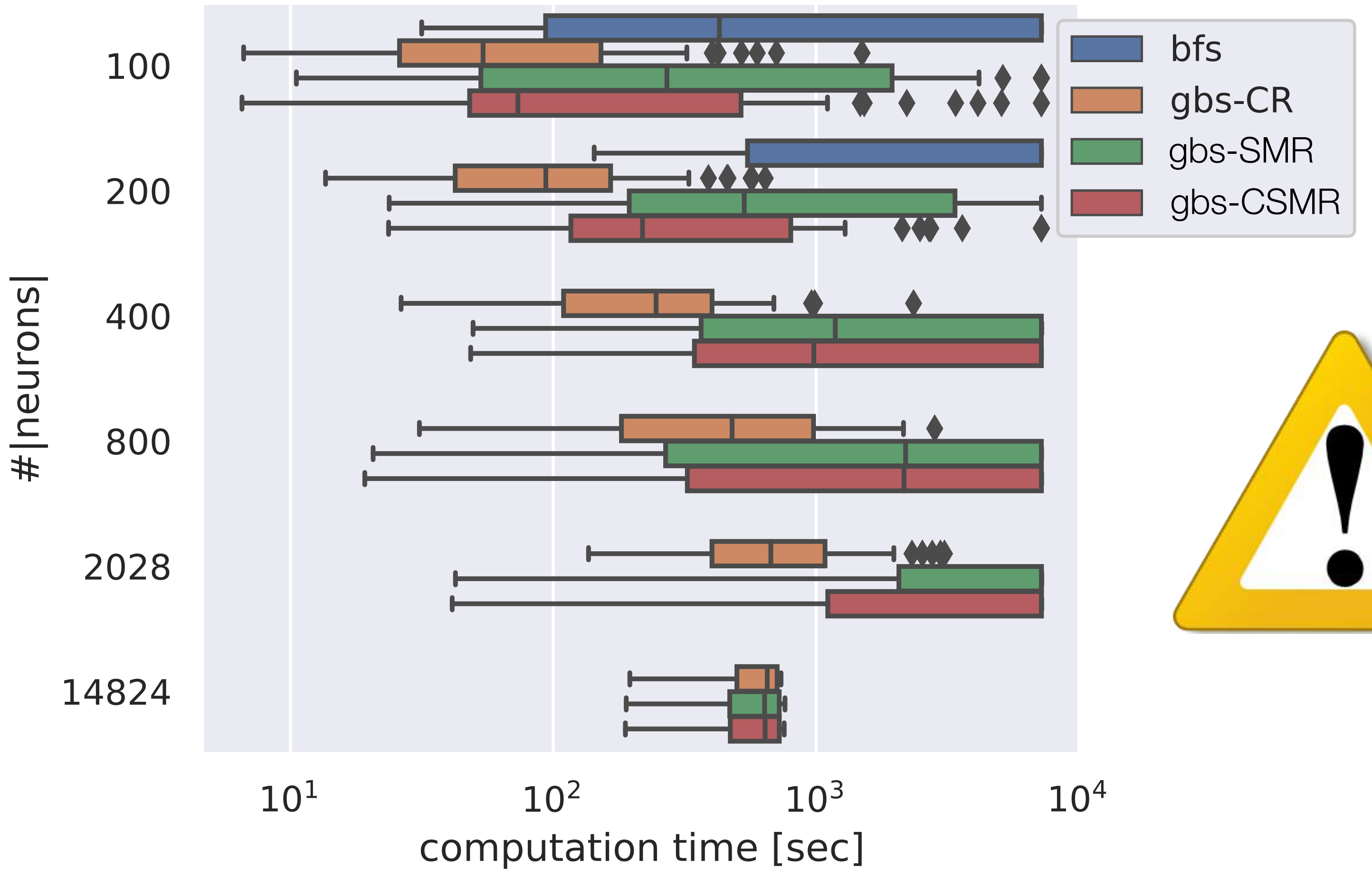


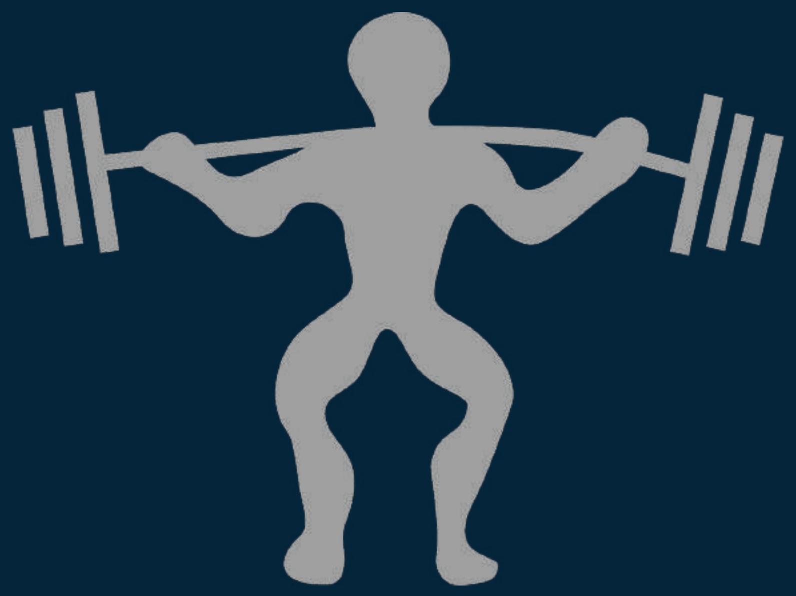
Classification Robustness





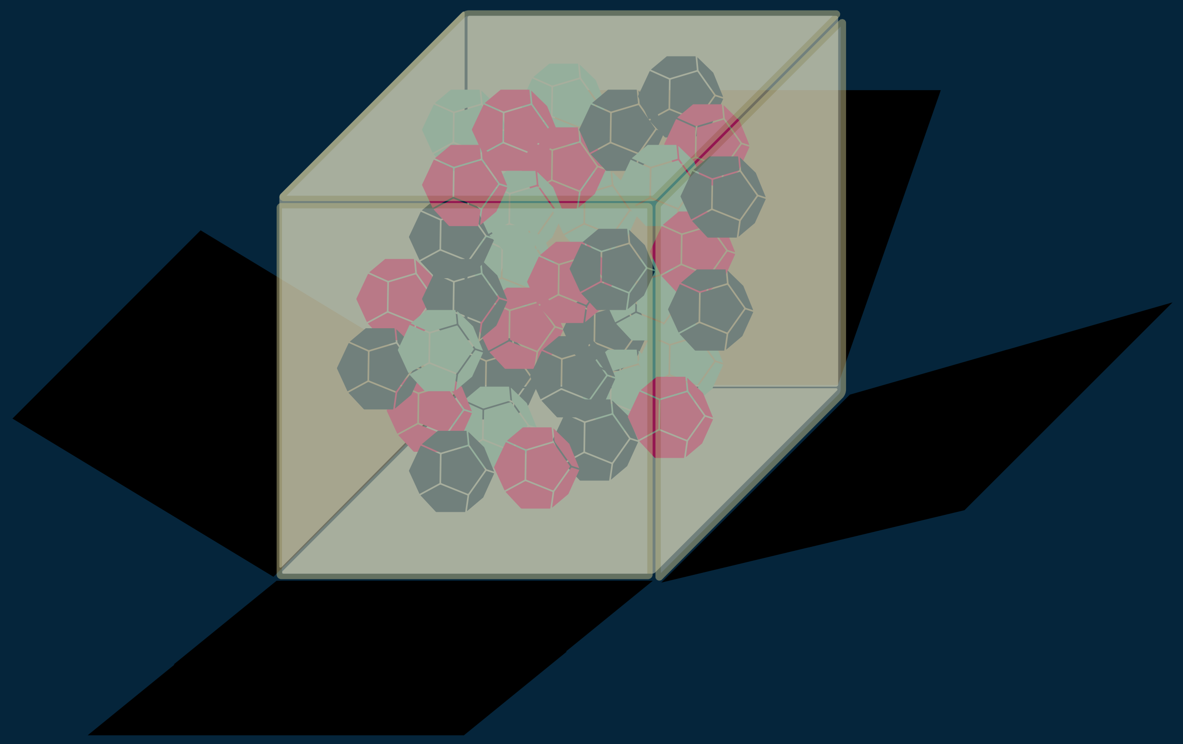
Saliency Map Robustness





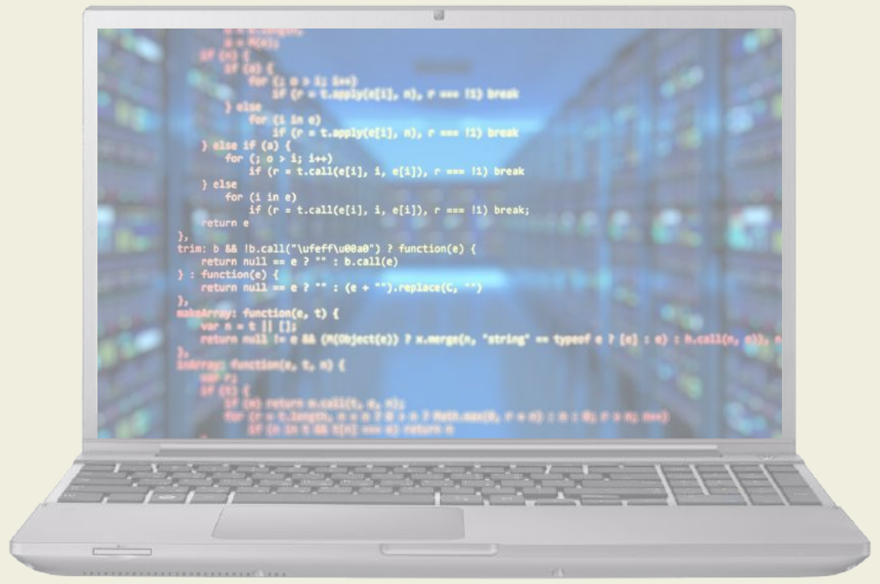
Training

CIKM 2021



Interpretability

VMCAI 2024

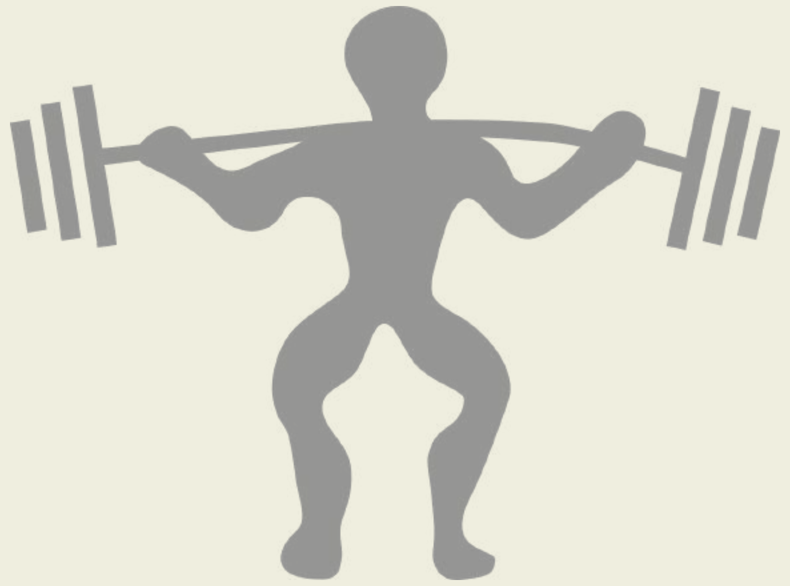


Verification

NFM 2023

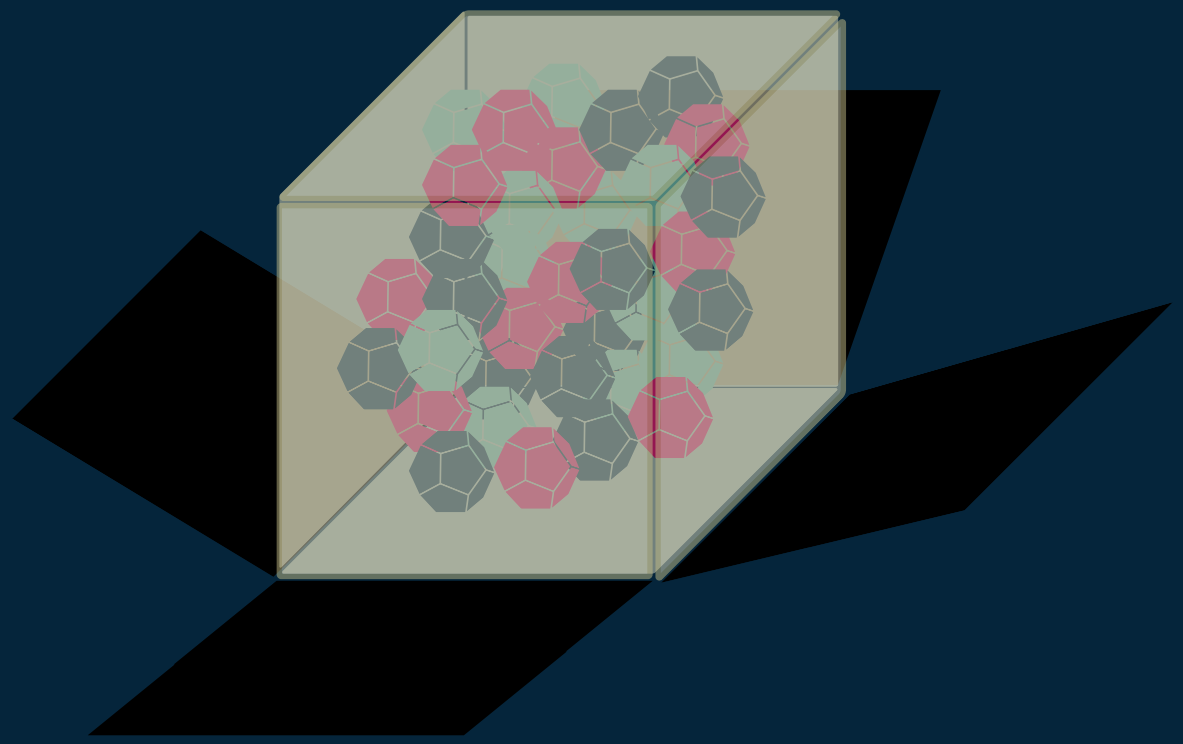
Using formal methods
for verification

Using formal methods
for something else
than verification



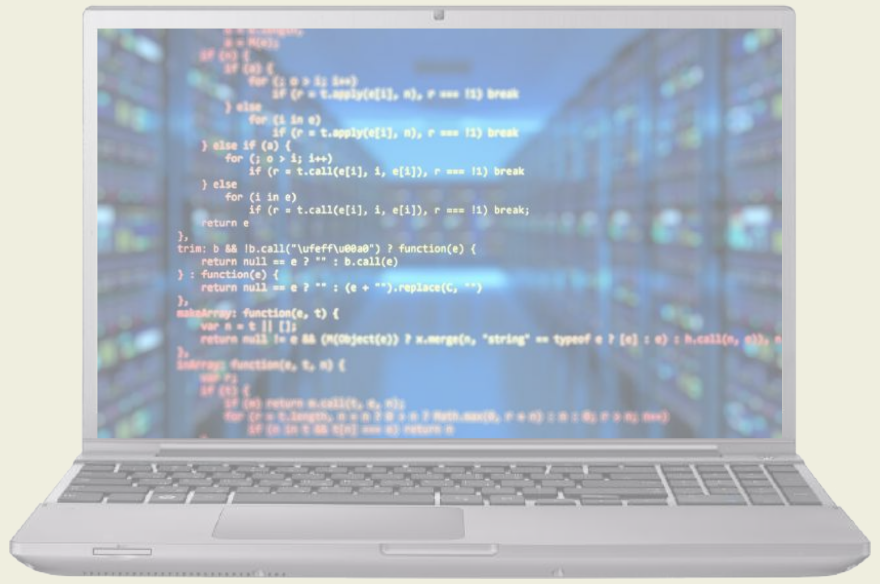
Training

CIKM 2021



Interpretability

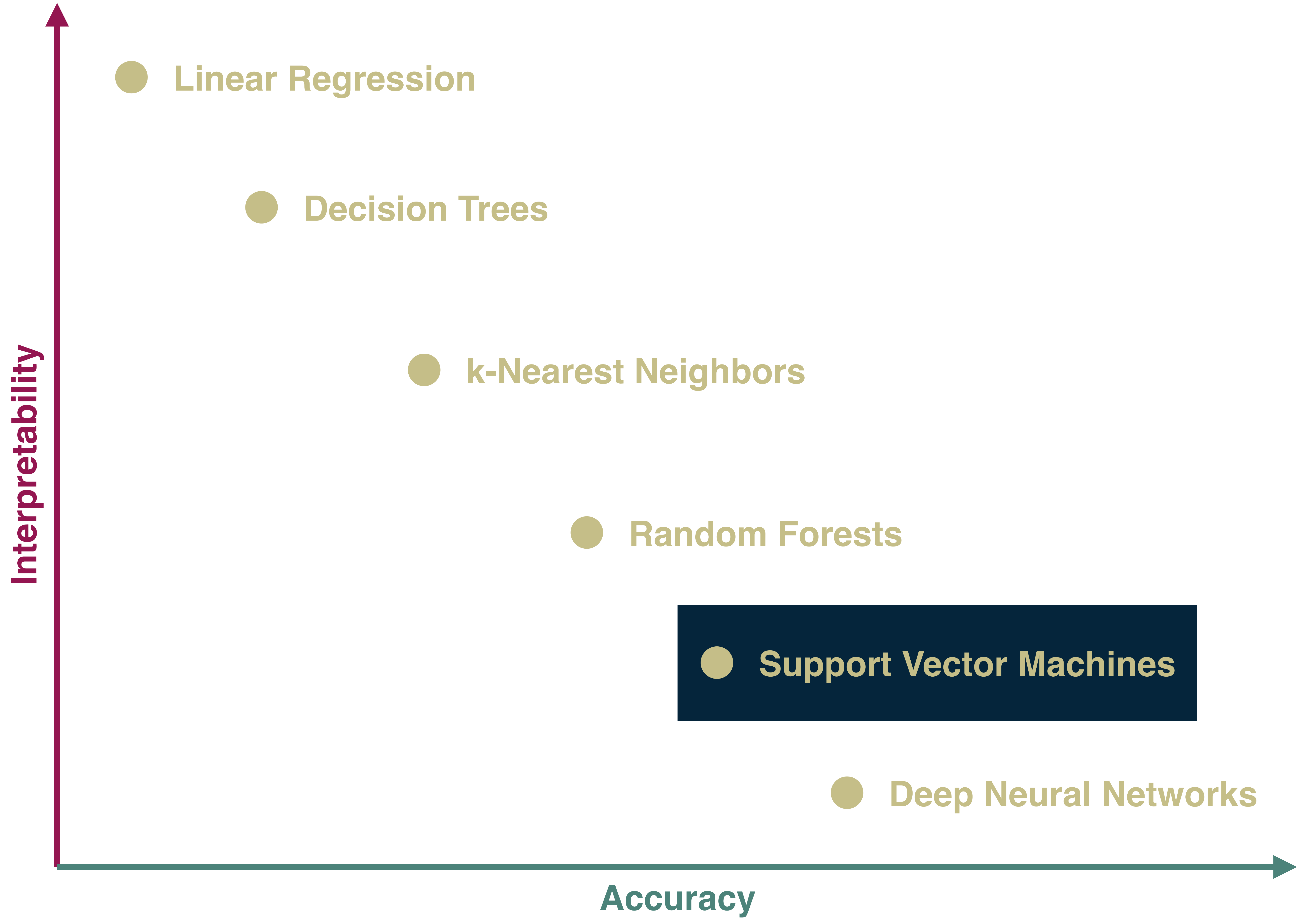
VMCAI 2024



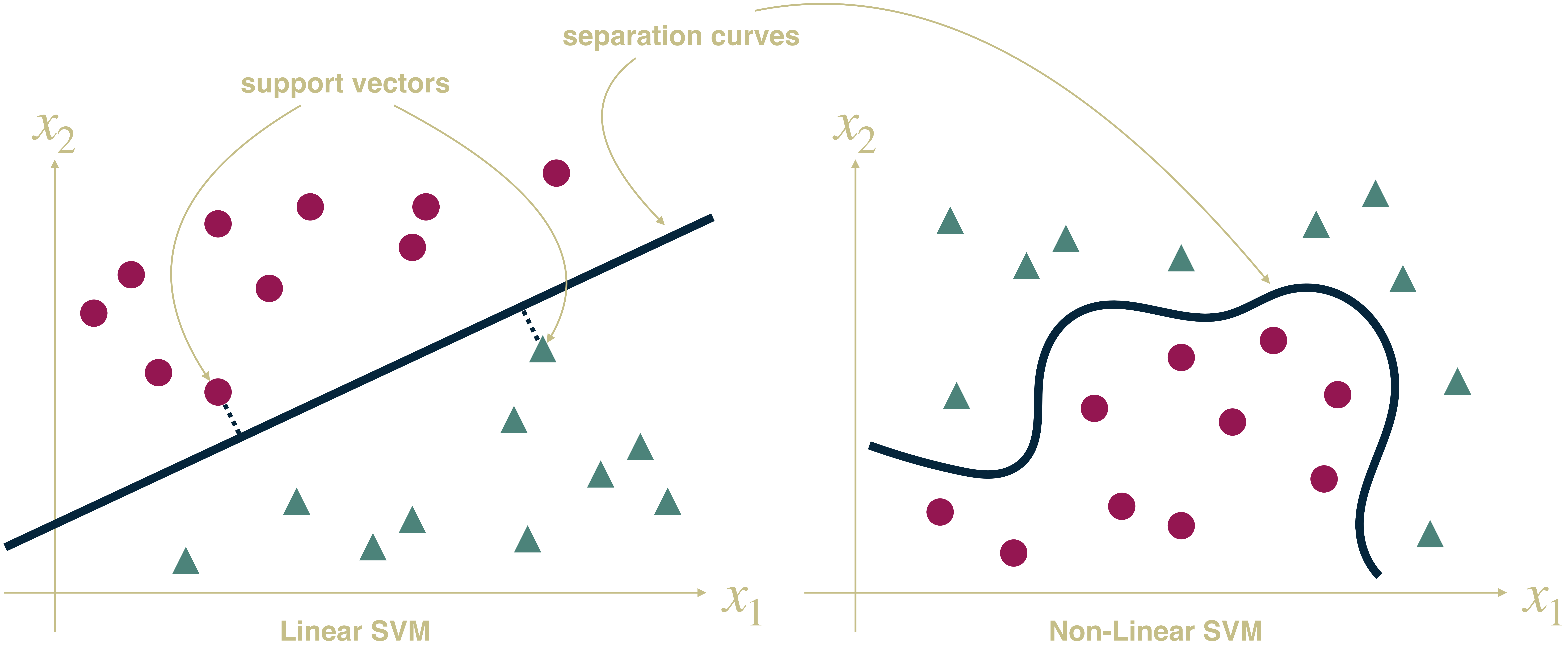
Verification

NFM 2023

Interpretability vs Performance



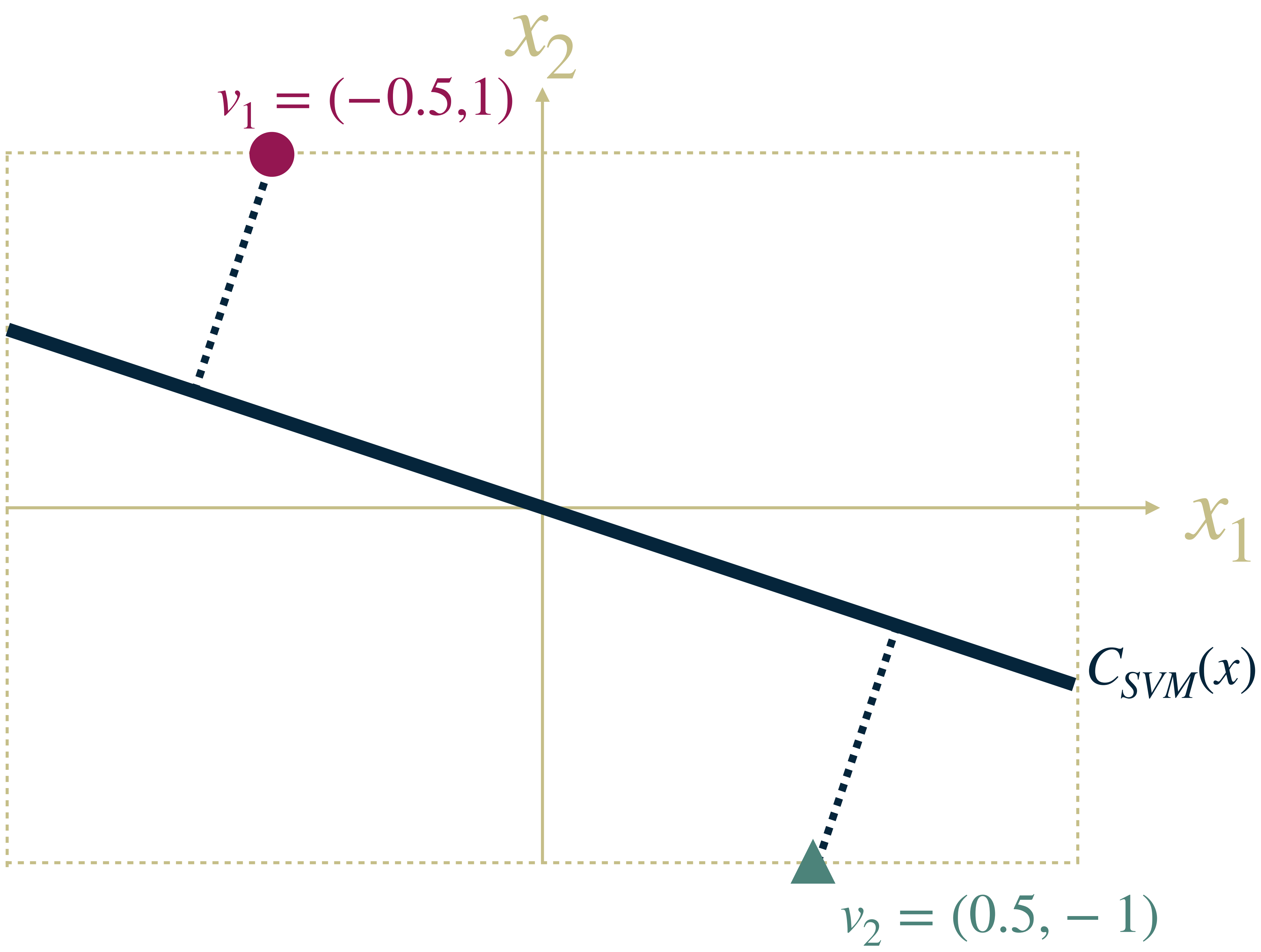
Support Vector Machines (SVMs)



Support Vector Machines (SVMs)

● $\mapsto -1$
▲ $\mapsto 1$

Example



weights

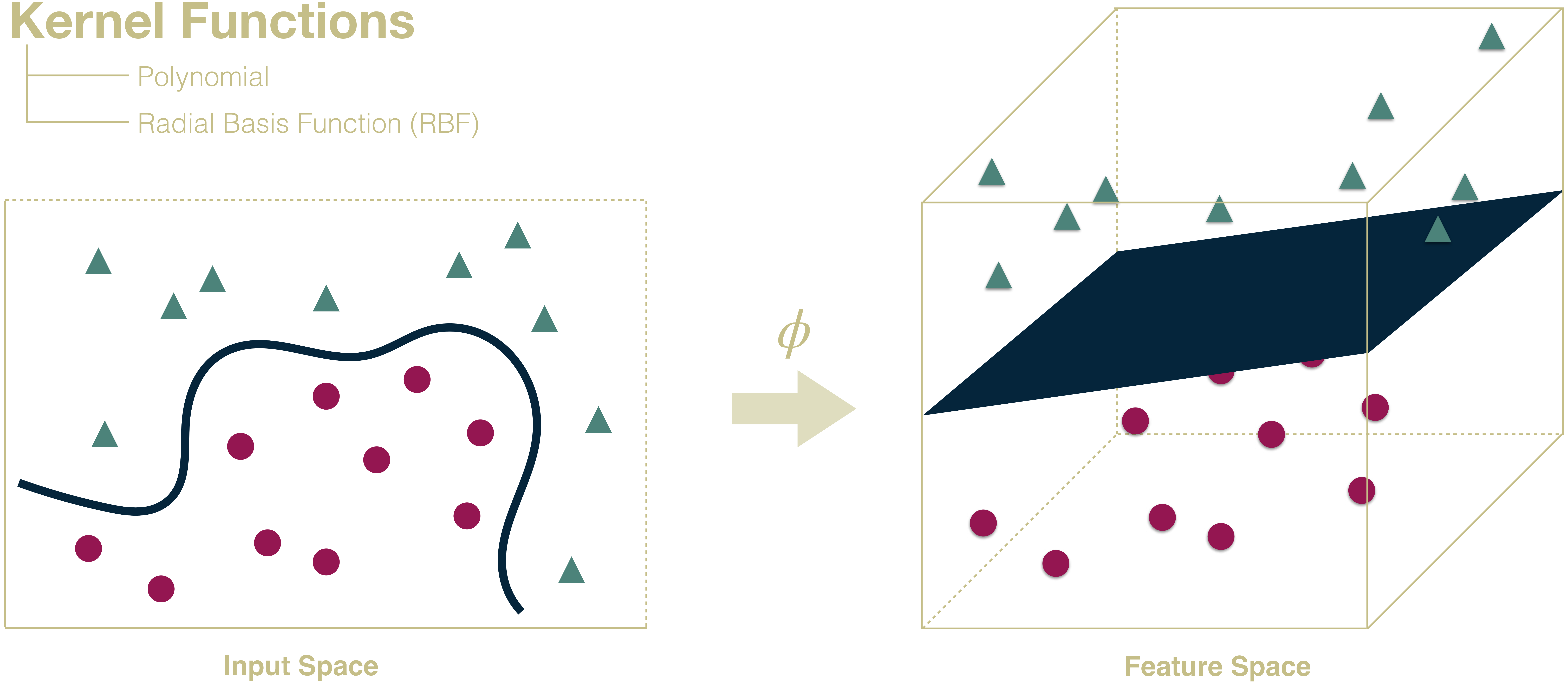
$$C_{SVM}(x) = sgn\left(SVM(x)\right) = sgn\left(-1 * 0.5(v_1 \cdot x) \ + \ 1 * 0.5(v_2 \cdot x)\right)$$

$$= sgn\left(0.5x_1 - x_2\right)$$

Non-Linear SVMs

Kernel Functions

- Polynomial
- Radial Basis Function (RBF)



Feature Importance

Measuring Contribution of Input Features to Prediction

	Local	Global	Model-		Performance	Effect
			Specific	Agnostic	-Based	
Permutation Feature Importance (PFI)		X		X	X	
Partial Dependence (PD) Plots		X		X		X
Individual Conditional Expectation (ICE) Plots		X		X		X
Accumulated Local Effects (ALE) Plots		X		X		X
Local Interpretable Model-Agnostic Explanations (LIME)	X			X		X
SHapley Additive exPlanations (SHAP)	X			X		X
Individual Conditional Importance (ICI) Curves	X			X	X	
Partial Importance (PI) Curves	X			X	X	
Shapley Feature Importance (SFIMP)		X		X	X	
Input Gradients	X			X	X	X
Abstract Feature Importance (AFI)	X	X	X			X

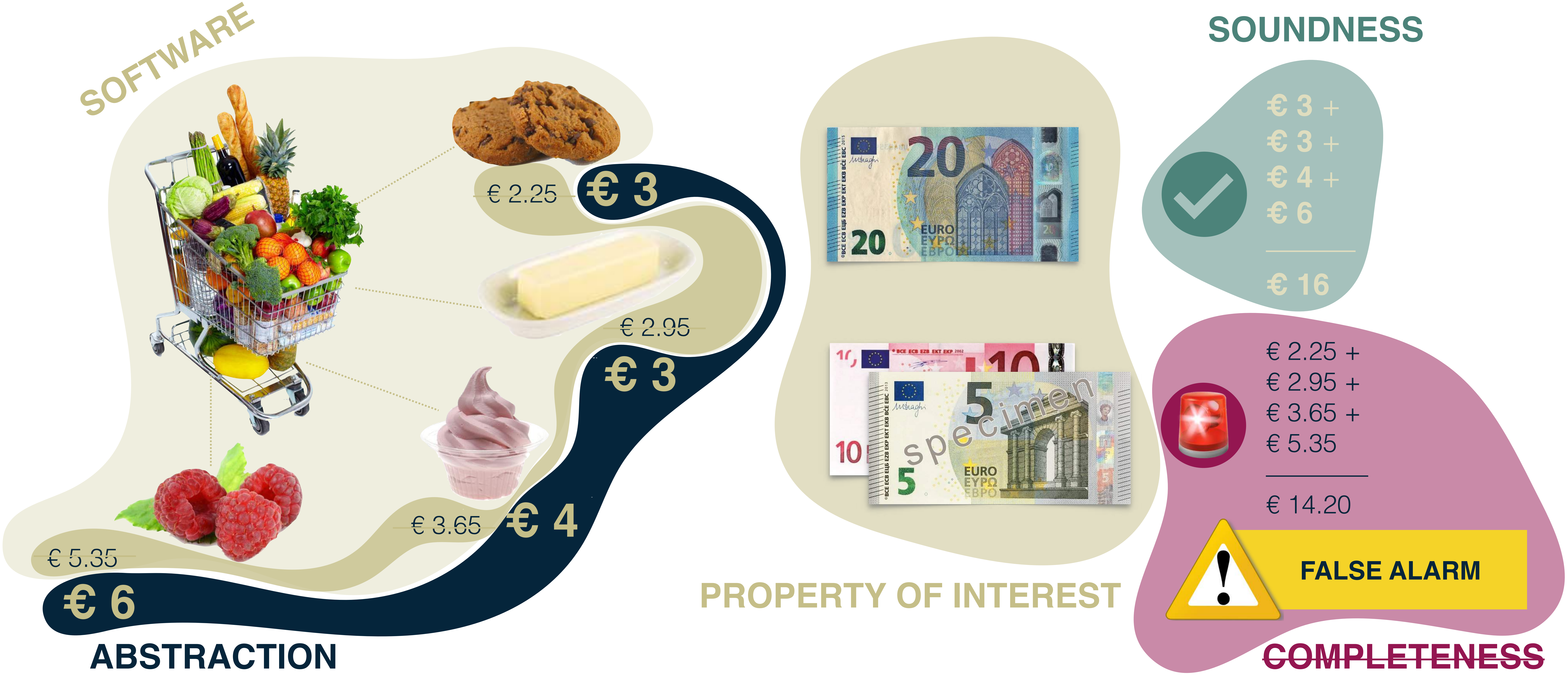
Abstract Feature Importance

Why Another Feature Importance Measure?

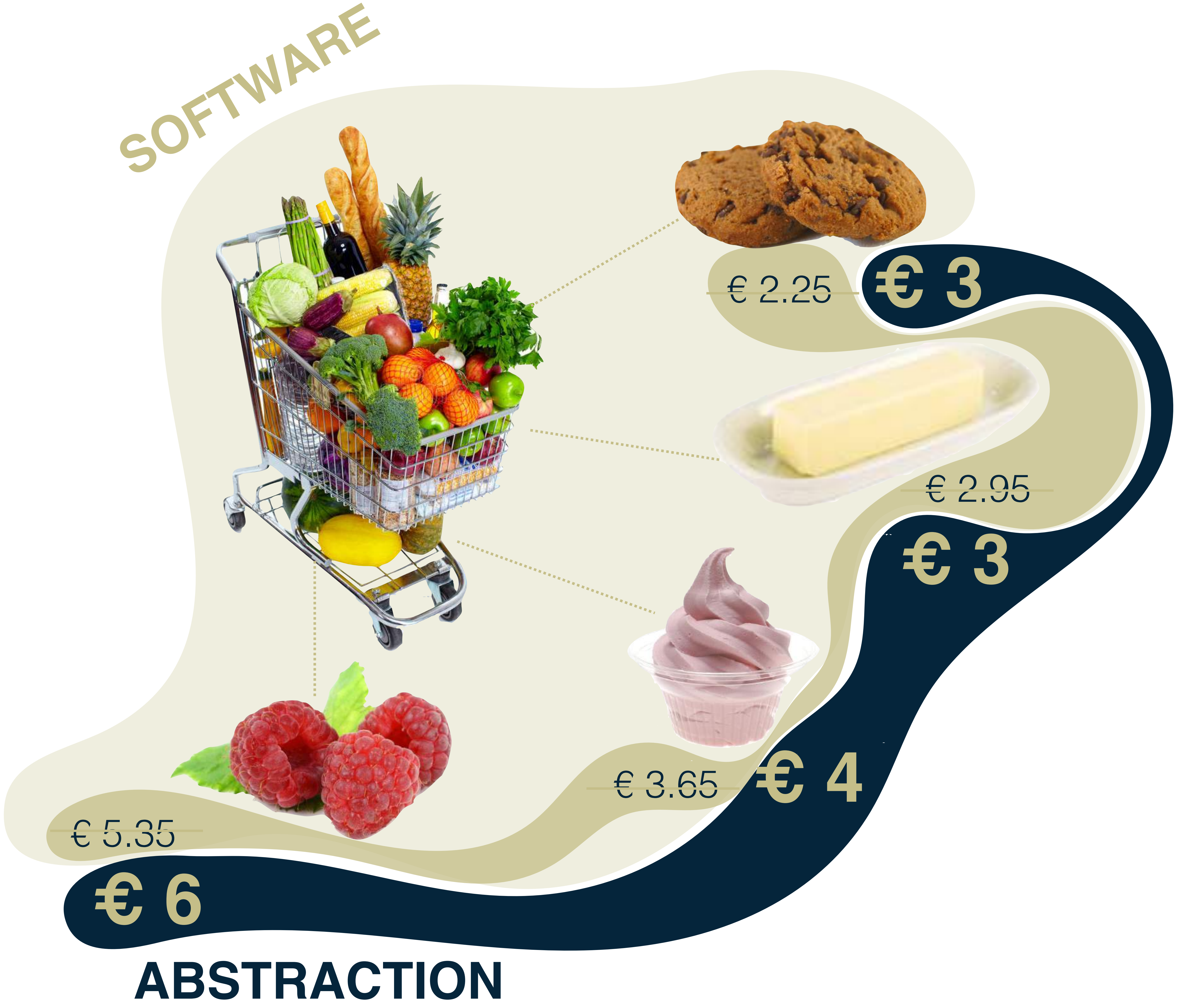
Permutation Feature Importance (PFI)	• **result may greatly vary depending on the dataset** • **resource intensive** when the number of fea… • **misleading result when features** … • **quality of the result h… …model accuracy**
Local Interpretable Model-Agnostic Explanations (L…	• **requi… …gful optimal neighborhood**: …ble and easily manipulable explanations • …es **that the decision boundary is linear at the local level**, but there is no theoretically guarantee that this is the case
S… …nations (SHAP)	• Shapley values estimations **depend on the dataset** • **assumes that features are independent** • **has a very high computational cost**, even for small models
Abstract Feature Importance (AFI)	• yields **a formally correct by construction approximation** • **does not depend** from a dataset nor the accuracy of the model • **extremely fast to compute**, whatever the number of features • supports both linear and non-linear kernel functions

“Make Sense” but Give No Guarantees

Abstract Interpretation



Abstract Interpretation



Abstract Interpretation of SVMs

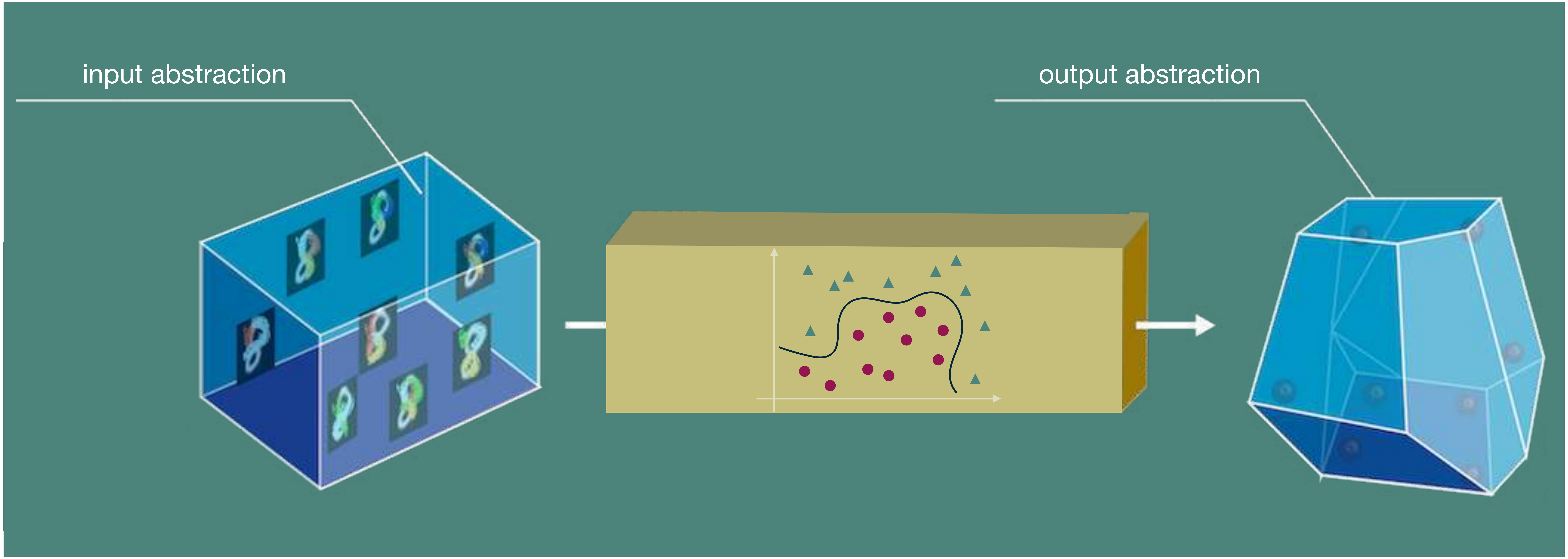


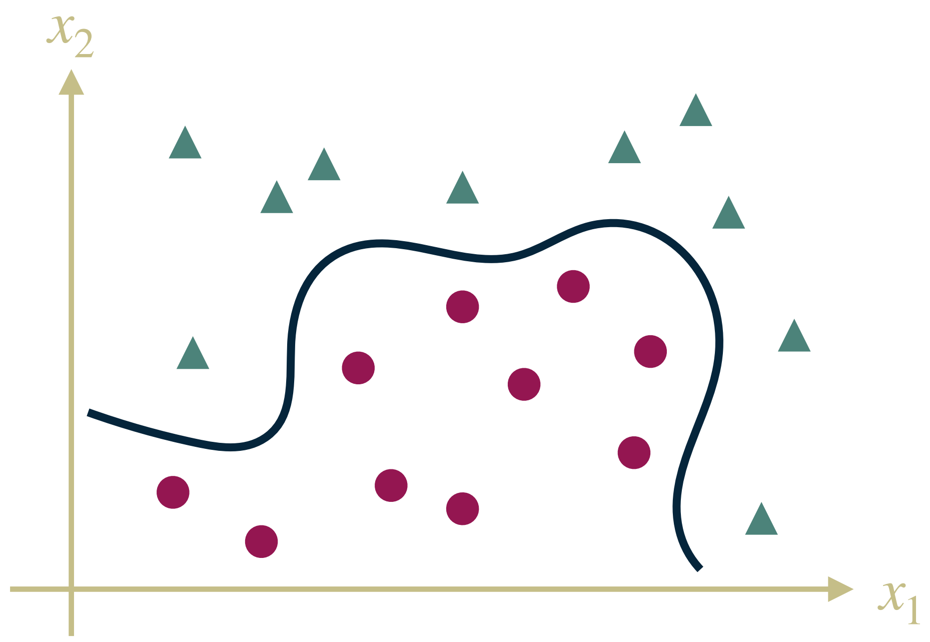
Image taken (and modified) from http://safeai.ethz.ch

Abstract Interpretation of SVMs

Reduced Affine Form (RAF) Abstraction

$$\mathbb{R}^n \xrightarrow{SVM(x)} \mathbb{R} \xrightarrow{sgn(SVM(x))} \{●, ▲\}$$

$$C_{SVM}(x)$$



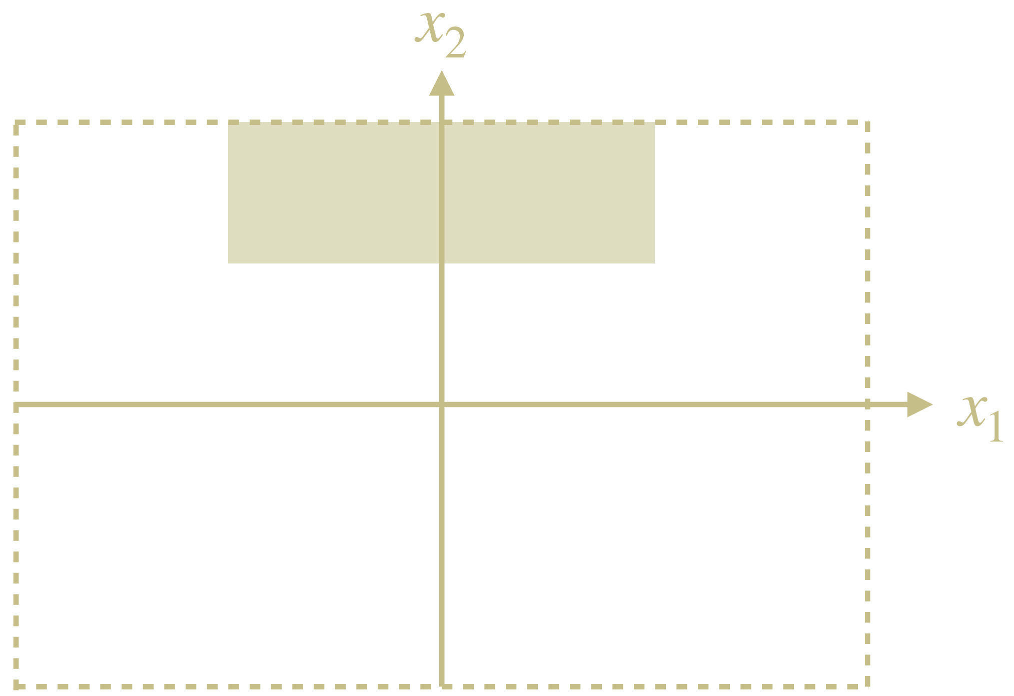
$$(RAF_n)^n \xrightarrow{SVM^\sharp(x^\sharp)} RAF_n \longrightarrow \{●, ▲\}$$

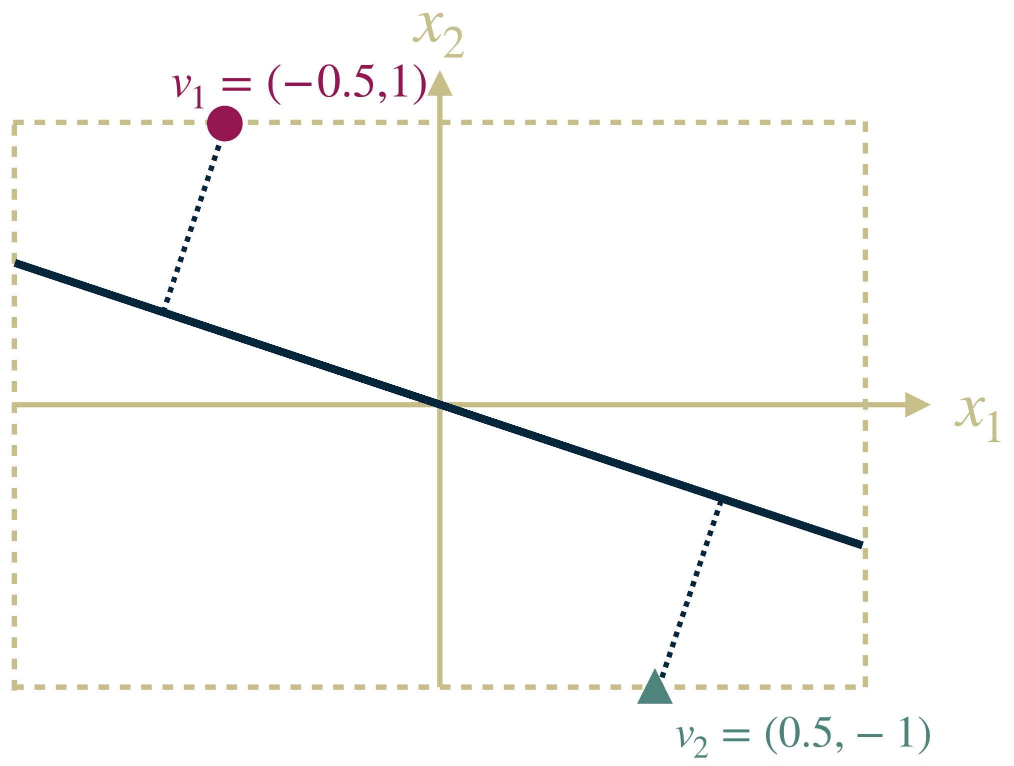
$$RAF_n \stackrel{\text{def}}{=} \left\{ a_0 + \sum_{i=1}^{n} a_i\epsilon_i + a_r\epsilon_r \mid a_0, a_1, \ldots a_n \in \mathbb{R}, a_r \in \mathbb{R}_{\geq 0} \right\} \cup \{ \top_{RAF} \}$$

Abstract Interpretation of SVMs

Example

$$\mathbb{R}^n \xrightarrow{SVM(x) = -0.5(v_1 \cdot x) + 0.5(v_2 \cdot x)} \mathbb{R} \overset{sgn\left(0.5x_1 - 1x_2\right)}{\dashrightarrow} \{●, ▲\}$$





$$(RAF_n)^n \xrightarrow{SVM^\sharp(x^\sharp)} RAF_n \dashrightarrow \{●, ▲\}$$

$(-0.5\epsilon_1,\ 0.75 - 0.25\epsilon_2)$

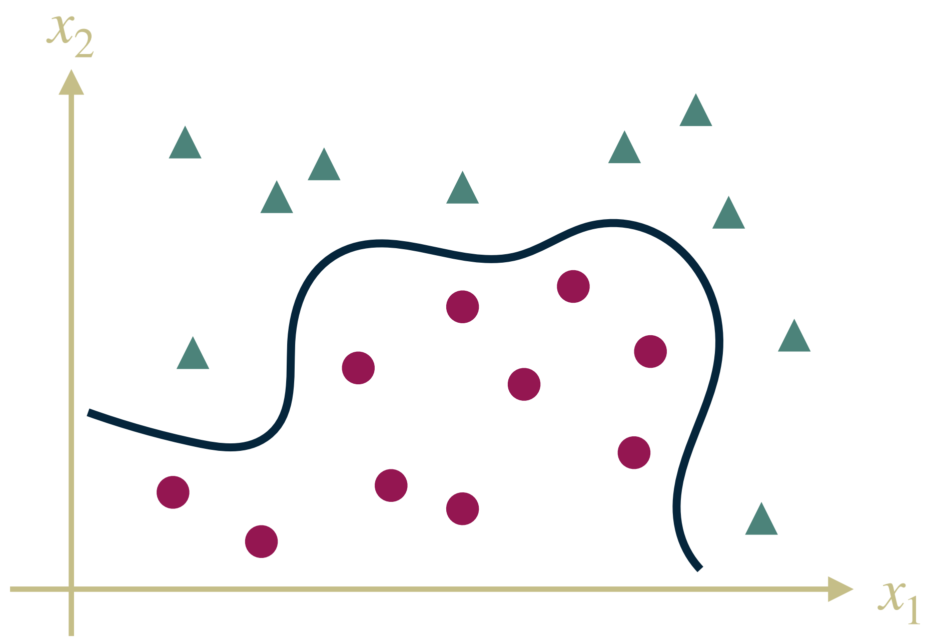
$-0.75 - 0.25\epsilon_1 + 0.25\epsilon_2$

$$SVM^\sharp((-0.5\epsilon_1,\ 0.75 - 0.25\epsilon_2))$$
$$= -0.5(-0.5(-0.5\epsilon_1) + 1(0.75 - 0.25\epsilon_2)) + 0.5(0.5(-0.5\epsilon_1) - 1(0.75 - 0.25\epsilon_2))$$
$$= -0.5(0.75 + 0.25\epsilon_1 - 0.25\epsilon_2) + 0.5(-0.75 - 0.25\epsilon_1 + 0.25\epsilon_2)$$
$$= -0.75 - 0.25\epsilon_1 + 0.25\epsilon_2$$

Abstract Feature Importance (AFI)

$$\mathbb{R}^n \xrightarrow{SVM(x)} \mathbb{R} \xrightarrow{sgn(SVM(x))} \{\bullet, \blacktriangle\}$$

$$C_{SVM}(x)$$



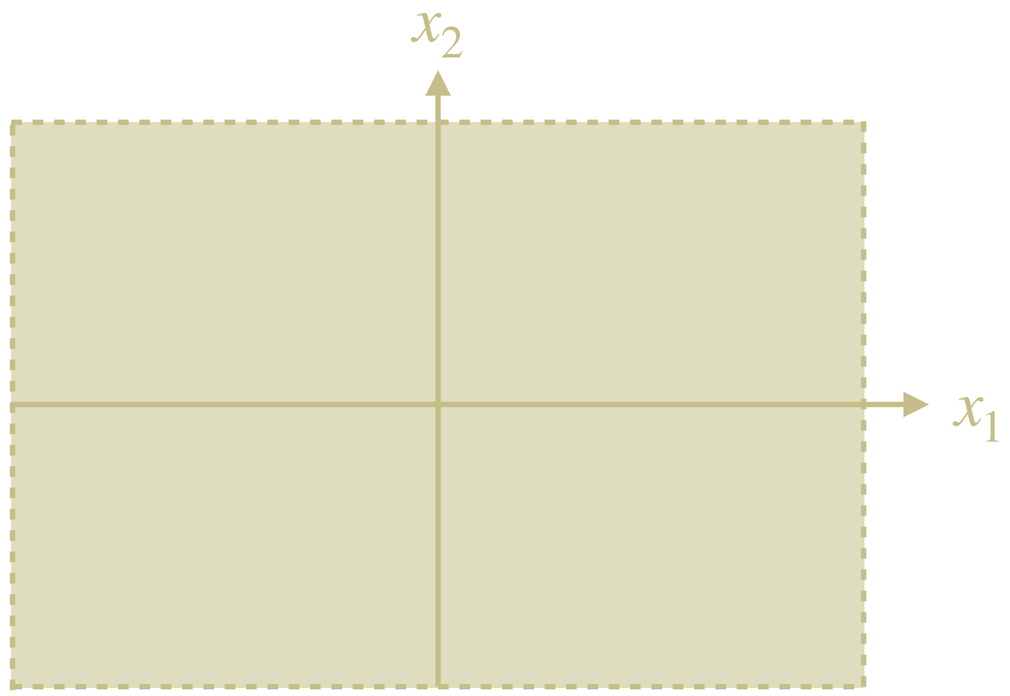
$$(RAF_n)^n \xrightarrow{SVM^\sharp(x^\sharp)} RAF_n \dashrightarrow \{\bullet, \blacktriangle\}$$

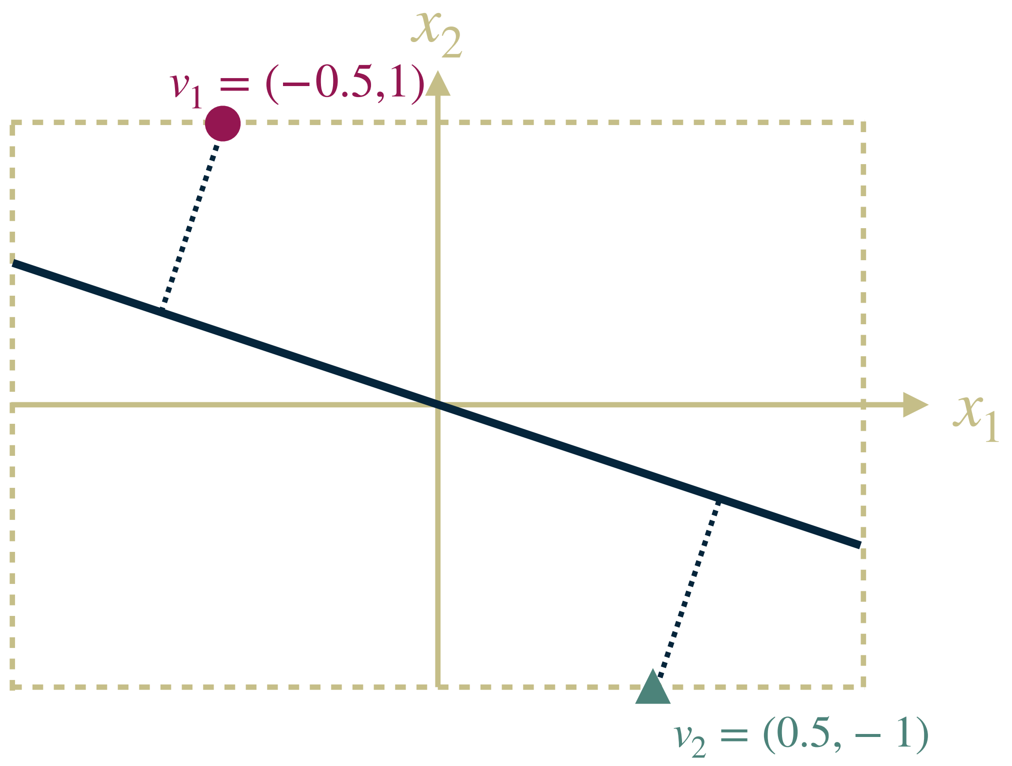
$$RAF_n \stackrel{\text{def}}{=} \left\{ a_0 + \sum_{i=1}^{n} a_i \epsilon_i + a_r \epsilon_r \mid a_0, a_1, \ldots a_n \in \mathbb{R}, a_r \in \mathbb{R}_{\geq 0} \right\} \cup \{ \top_{RAF} \}$$

Abstract Feature Importance (AFI)

Example

$\mathbb{R}^n \xrightarrow{SVM(x) = -0.5(v_1 \cdot x) + 0.5(v_2 \cdot x)} \mathbb{R} \xrightarrow{sgn\left(0.5x_1 - 1x_2\right)} \{●, ▲\}$





$(RAF_n)^n \xrightarrow{SVM^\sharp(x^\sharp)} RAF_n \dashrightarrow \{●, ▲\}$

(ϵ_1, ϵ_2)

$0.5\epsilon_1 - 1\epsilon_2$

$$
\begin{aligned}
&SVM^\sharp((\epsilon_1, \epsilon_2)) \\
&= -0.5(-0.5\epsilon_1 + 1\epsilon_2) + 0.5(0.5\epsilon_1 - 1\epsilon_2) \\
&= 0.25\epsilon_1 - 0.5\epsilon_2 + 0.25\epsilon_1 - 0.5\epsilon_2 \\
&= 0.5\epsilon_1 - \epsilon_2
\end{aligned}
$$

AFI vs PFI

German Dataset

		Grade for each feature										
Linear	Baseline (13.55s)	**5**	**5**	**5**	**6**	**6**	**7**	**7**	**7**	**7**	**8**	Distance
	AFI (0.01s)	**5**	**5**	**5**	**6**	**6**	**7**	8	**7**	**7**	**8**	**1.0**
	PFI (4.07s)	**5**	**5**	6	7	7	9	6	6	**7**	7	3.16
RBF	Baseline (17.98s)	**5**	**5**	**5**	**6**	**6**	**7**	**7**	**7**	**8**	**8**	Distance
	AFI (0.02s)	**5**	6	**5**	**6**	**6**	8	**7**	**7**	**8**	7	**1.73**
	PFI (6.23s)	6	7	**5**	**6**	7	8	**7**	6	7	5	4.24
Polynomial	Baseline (15.83s)	**5**	**5**	**5**	**6**	**7**	**7**	**7**	**7**	**7**	**8**	Distance
	AFI (0.01s)	7	6	7	7	5	**7**	6	6	5	**8**	**4.47**
	PFI (4.15s)	6	7	9	7	6	**7**	5	6	6	6	5.74

AFI vs PFI

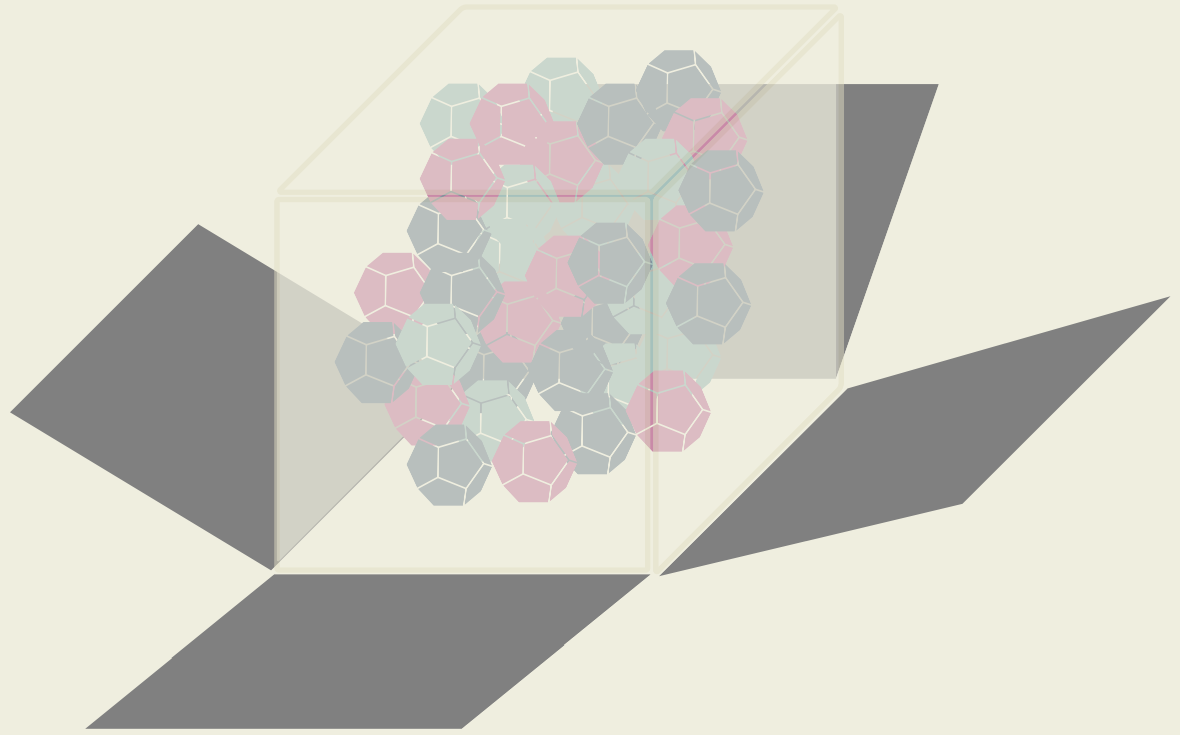
	Baseline	$N = 2k$ $\epsilon = 0.2$	$N = 10k$ $\epsilon = 0.2$	$N = 2k$ $\epsilon = 0.4$	$N = 10k$ $\epsilon = 0.4$	$N = 2k$ $\epsilon = 0.6$	$N = 5k$ $\epsilon = 0.6$	$N = 10k$ $\epsilon = 0.6$	$N = 2k$ $\epsilon = 0.8$	$N = 5k$ $\epsilon = 0.8$	$N = 10k$ $\epsilon = 0.8$
Adult	AFI (0.27s)	**0.0**	**0.0**	**1.0**	**0.0**	**1.0**	**1.41**	**1.0**	**1.0**	**1.41**	**1.0**
Linear	PFI (10009s)	2.45	2.45	2.24	2.45	2.24	**1.41**	2.24	2.24	**1.41**	2.24
Adult	AFI (0.48s)	**1.0**	**1.41**	**1.41**	**1.41**	**1.73**	**1.73**	**1.41**	**1.41**	**1.41**	**1.41**
RBF	PFI (25221s)	1.73	2.45	2.45	2.0	2.65	2.65	2.45	2.45	2.45	2.45
Adult	AFI (0.44s)	**1.0**	**1.0**	**0.0**	1.41	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Polynomial	PFI (9985s)	**1.0**	**1.0**	1.41	**1.0**	1.41	1.41	1.41	1.41	1.41	1.41
Compas	AFI (0.22s)	**1.41**	**1.41**	**1.73**	**1.73**	**1.41**	**1.73**	**1.41**	**1.41**	**1.41**	**1.73**
Linear	PFI (1953s)	1.73	1.73	2.0	2.0	2.24	2.0	2.24	2.24	2.24	2.83
Compas	AFI (0.27s)	**2.0**	**2.0**	**2.65**	**2.65**	**2.83**	**2.83**	**2.83**	**2.83**	**2.83**	**2.83**
RBF	PFI (6827s)	**2.0**	**2.0**	**2.65**	**2.65**	**2.83**	**2.83**	**2.83**	**2.83**	**2.83**	**2.83**
Compas	AFI (0.22s)	4.24	4.24	4.12	4.12	4.24	4.24	4.24	4.24	4.24	4.24
Polynomial	PFI (2069s)	**2.45**	**2.45**	**3.0**	**3.0**	**3.74**	**3.74**	**3.74**	**3.74**	**3.74**	**3.74**
German	AFI (0.01s)	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.41**	**1.73**	**1.41**
Linear	PFI (4.07s)	3.16	3.46	3.16	3.16	3.16	3.16	3.16	3.6	3.74	3.0
German	AFI (0.02s)	**1.73**	**1.0**	**1.73**	**1.73**	**2.0**	**1.41**	**1.73**	**1.73**	**2.0**	**2.24**
RBF	PFI (6.23s)	4.0	3.46	4.24	4.24	4.36	3.61	4.24	4.24	4.36	4.47
German	AFI (0.01s)	**4.90**	**4.12**	**4.47**	**3.87**	**3.87**	**4.24**	**3.46**	**3.46**	**3.46**	**3.46**
Polynomial	PFI (4.15s)	5.74	5.10	5.74	4.69	4.69	5.0	4.58	4.58	4.58	4.58

AFI vs LIME

Distance between LIME and ...	**Adult**			**Compas**			**German**		
	Lin.	RBF	Poly	Lin.	RBF	Poly	Lin.	RBF	Poly
AFI ($\epsilon = 0.1$)	2.42	2.04	2.98	1.67	1.06	3.05	2.62	2.03	**5.31**
AFI ($\epsilon = 0.2$)	1.68	1.32	2.67	1.63	0.17	2.73	2.21	2.00	5.41
AFI ($\epsilon = 0.3$)	1.39	0.51	2.58	**1.57**	0.14	**2.62**	1.92	2.05	5.45
AFI (Global)	**1.37**	**0.01**	**1.01**	**1.57**	**0.13**	3.16	**1.90**	**1.89**	5.53

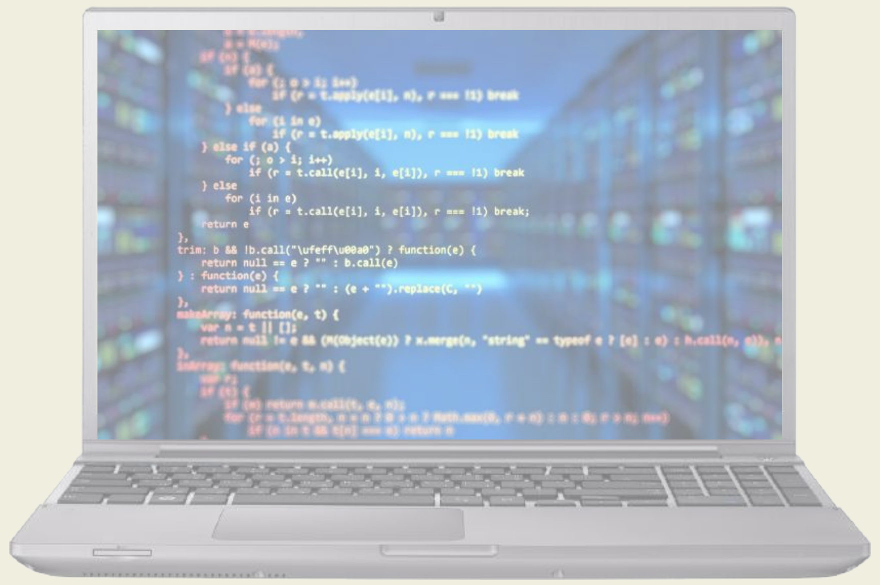
Training

CIKM 2021



Interpretability

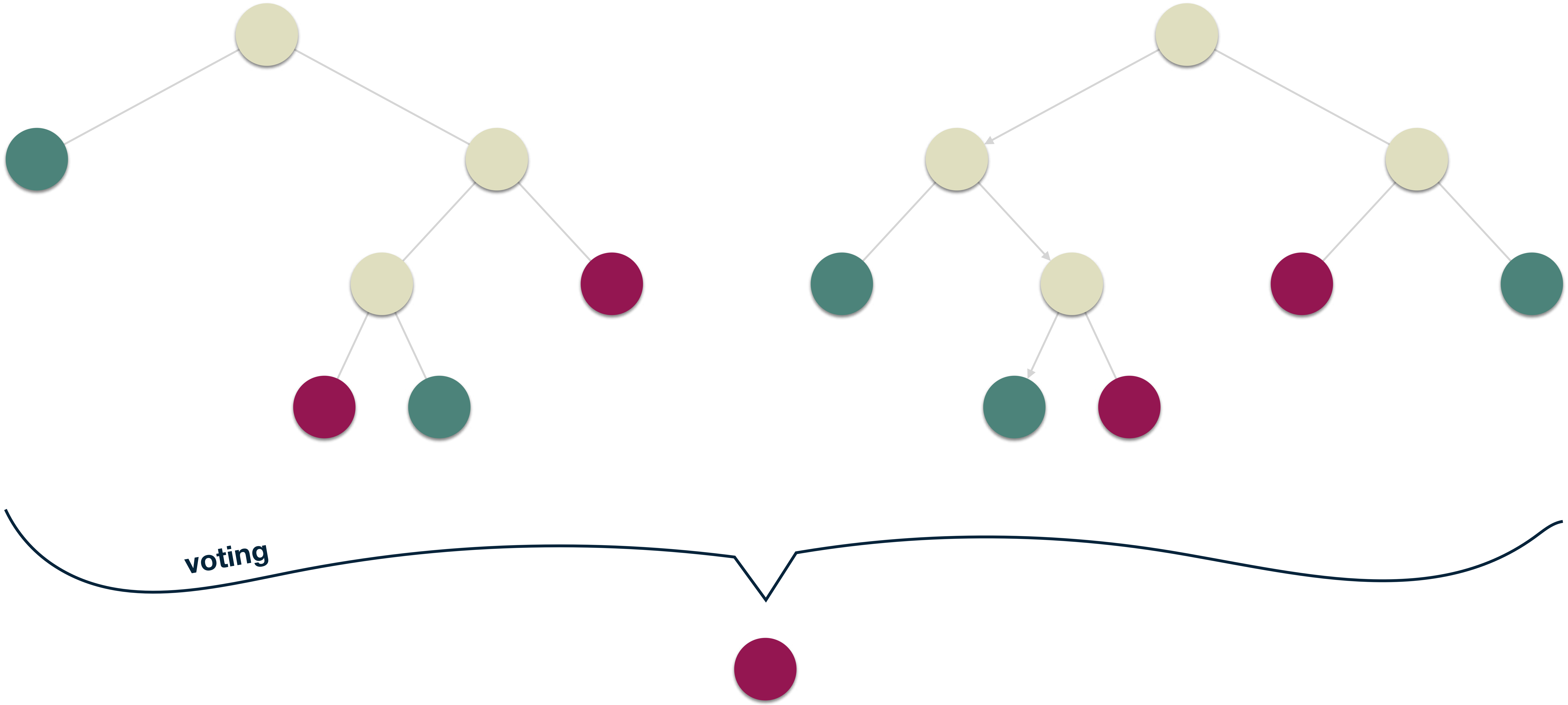
VMCAI 2024



Verification

NFM 2023

Random Forests



Robust Training

Minimizing the Worst-Case Loss for Each Input

Adversarial Training

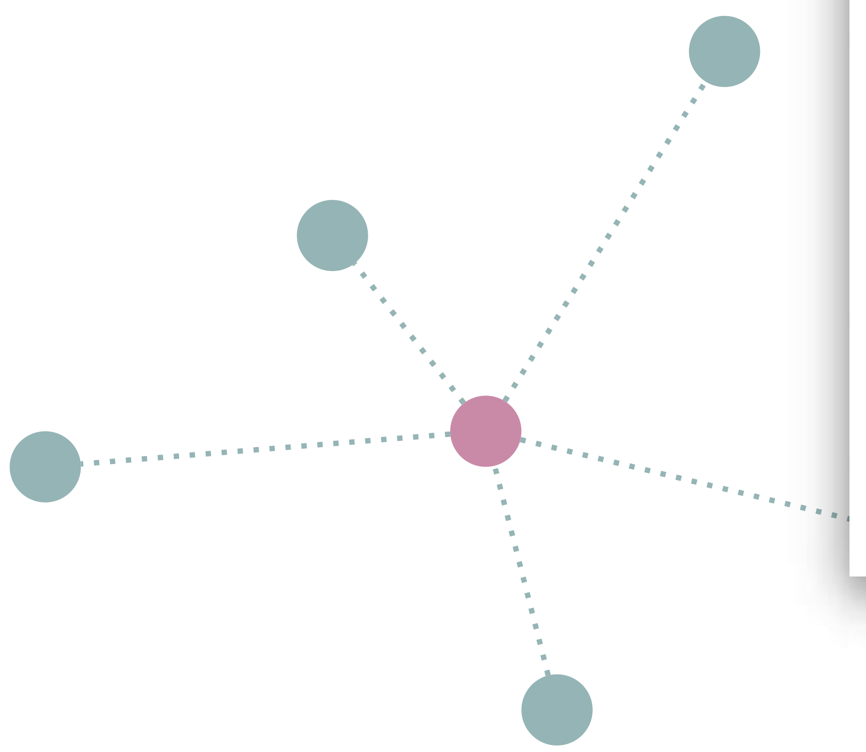
Minimizing a **Lower-Bound** on the Worst-Case Loss

Certified Training

Minimizing an **Upper-Bound** on the Worst-Case Loss

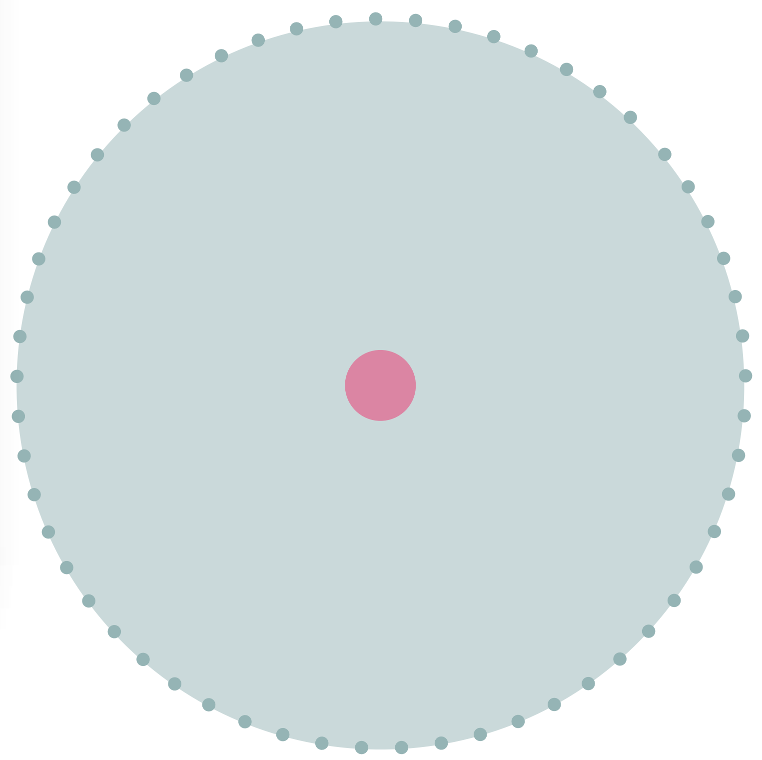
Hybrid Training

$$(1-\alpha)\mathcal{L}(f(\boldsymbol{\theta}, \mathbf{x}_{\mathrm{adv}}), y) + \alpha\ \mathcal{L}_{\mathrm{ver}}(f(\boldsymbol{\theta}, \mathbf{x}), y)$$



$$\mathcal{L}(f(\boldsymbol{\theta}, \mathbf{x}_{\mathrm{adv}}), y)$$

$$\mathcal{L}_{\mathrm{ver}}(f(\boldsymbol{\theta}, \mathbf{x}), y)$$



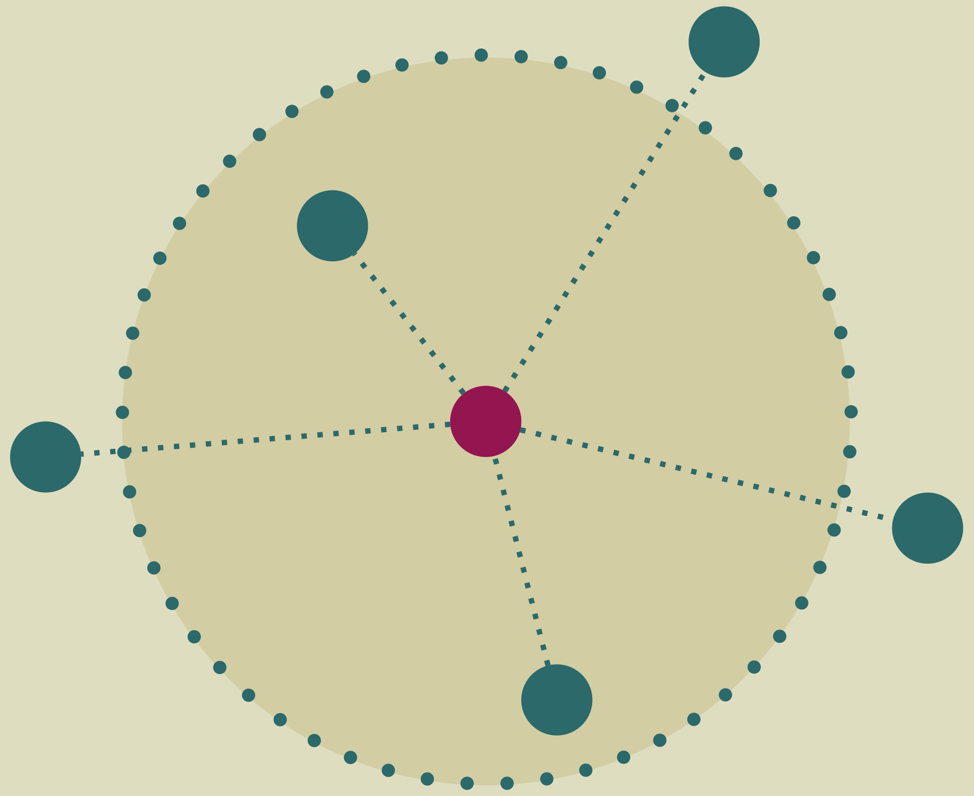
Machine Learning Community

Formal Methods Community

Hybrid Training

Random Forests

Dataset	FATT			Natural CART			CART with Hints		
	Accuracy %	Fairness %	Size	Accuracy %	Fairness %	Size	Accuracy %	Fairness %	Size
Adult	80.84	95.21	43	85.32	77.56	270	84.77	87.46	47
Compas	64.11	85.98	75	65.91	22.25	56	65.91	22.25	56
Crime	79.45	75.19	11	77.69	24.31	48	77.44	60.65	8
German	72.00	99.50	2	75.50	57.50	115	73.50	86.00	4
Health	77.87	97.03	84	83.85	79.98	2371	82.25	93.64	100
Average	74.85	**90.58**	**43**	**77.65**	52.32	572	76.77	70.00	**43**



Hybrid Training

CIKM 2021

$$a_0 + \sum_{i=1}^{n} a_i\epsilon_i + a_r\epsilon_r$$

Verification for Interpretability

VMCAI 2024

Interpretability for Verification

NFM 2023

THANKS!