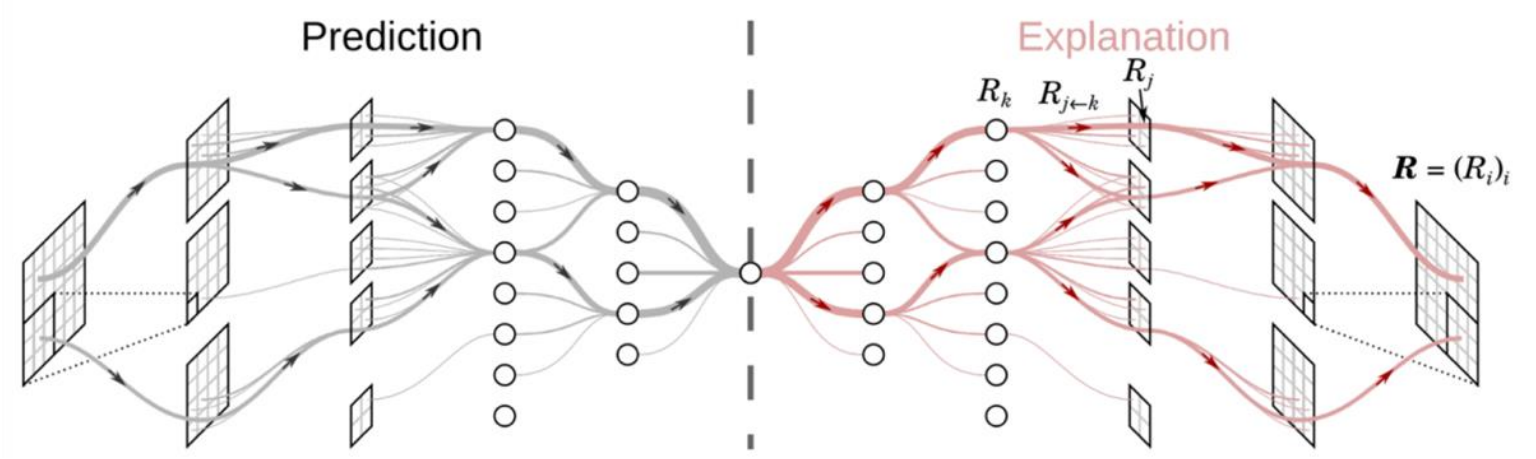
From Feature Attributions to Next-Generation Explainable AI

Wojciech Samek

TU Berlin & Fraunhofer HHI



Syllabus

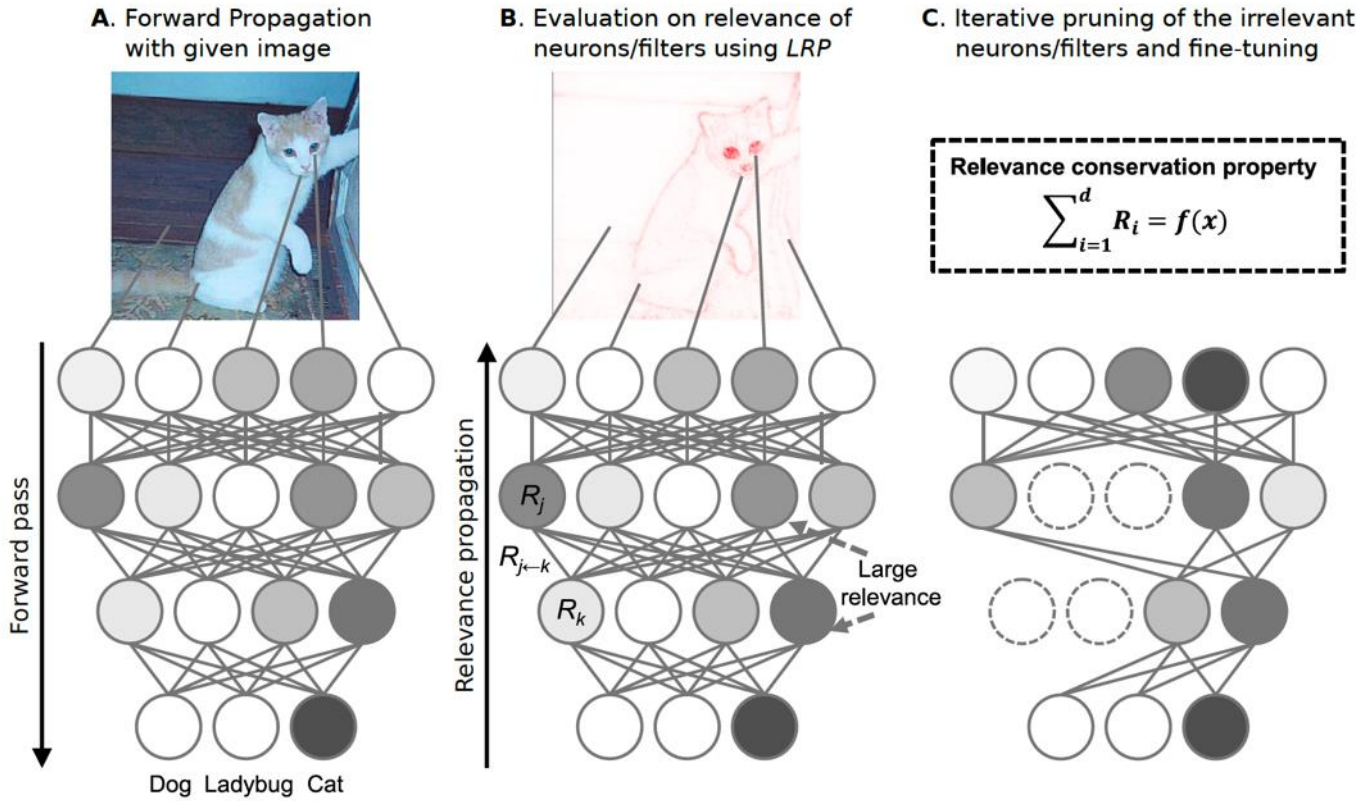
Part I: First Generation XAI
- What to explain
- Explaining by attribution
- DTD Framework
- Evaluating explanations

Part II: New Developments
- Concepts and prototypes
- XAI for LLMs
- Non-Interpretable domains
- Beyond classification

Part III: Beyond Explaining
- XAI-Based model surgery
- Reveal and revise
- Explanatory interactive ML
- Future of XAI

XAI-Based Model Surgery

XAI-Based Pruning



(Yeom et al. 2019)

XAI-Based Pruning

VGG-16	Scene 15 @ 55%					Event 8 @ 55%					Cats & Dogs @ 60%				
	U	W	T	G	L	U	W	T	G	L	U	W	T	G	L
Loss	2.09	2.27	1.76	1.90	**1.62**	0.85	1.35	1.01	1.18	**0.83**	0.19	0.50	0.51	0.57	**0.44**
Accuracy	88.59	82.07	83.00	82.72	**83.99**	95.95	90.19	91.79	90.55	**93.29**	99.36	97.90	97.54	97.19	**98.24**
Params	119.61	56.17	53.10	53.01	**49.67**	119.58	56.78	48.48	50.25	**47.35**	119.55	47.47	51.19	57.27	**43.75**
FLOPs	15.50	8.03	**4.66**	4.81	6.94	15.50	8.10	5.21	**5.05**	7.57	15.50	7.02	3.86	**3.68**	6.49

	Oxford Flower 102 @ 70%					CIFAR-10 @ 30%				
	U	W	T	G	L	U	W	T	G	L
Loss	3.69	3.83	3.27	3.54	**2.96**	1.57	1.83	1.76	1.80	**1.71**
Accuracy	82.26	71.84	72.11	70.53	**74.59**	91.04	93.36	93.29	93.05	**93.42**
Params	119.96	39.34	41.37	42.68	**37.54**	119.59	**74.55**	97.30	97.33	89.20
FLOPs	15.50	5.48	**2.38**	2.45	4.50	15.50	11.70	**8.14**	8.24	9.93

Pruning criteria
U = Unprunned
W = Weight
T = Taylor
G = Gradient
L = LRP

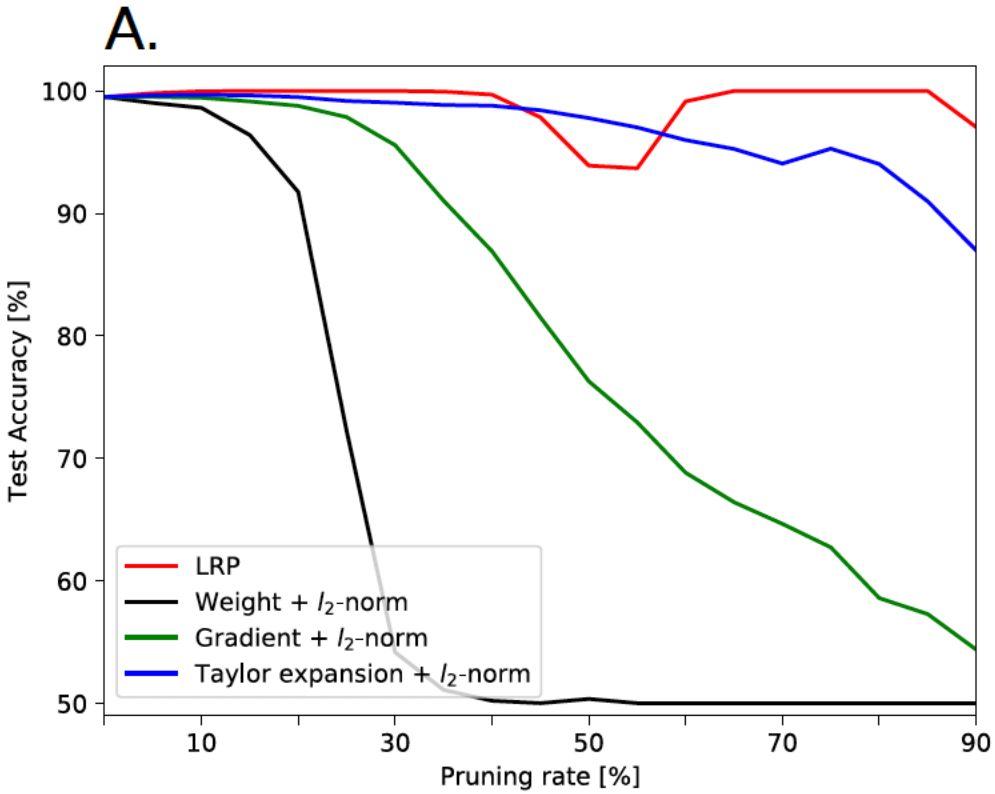
Observations:

- small weights can be relevant
- LRP conservation property act as normalizer (better comparability across layers)

(Yeom et al. 2019)

With fine-tuning

XAI-Based Pruning



Idea: Take generic model (VGG, 1000 classes) and make it smaller and more specialized (cats vs. dogs only) by pruning.

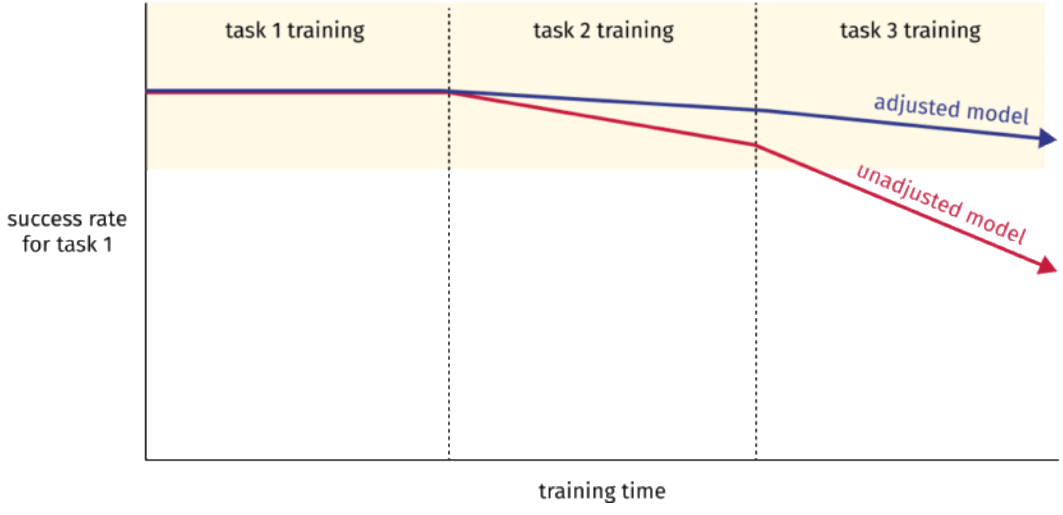
Only 10 samples per class (domain adaptation scenario)

(Yeom et al. 2019)

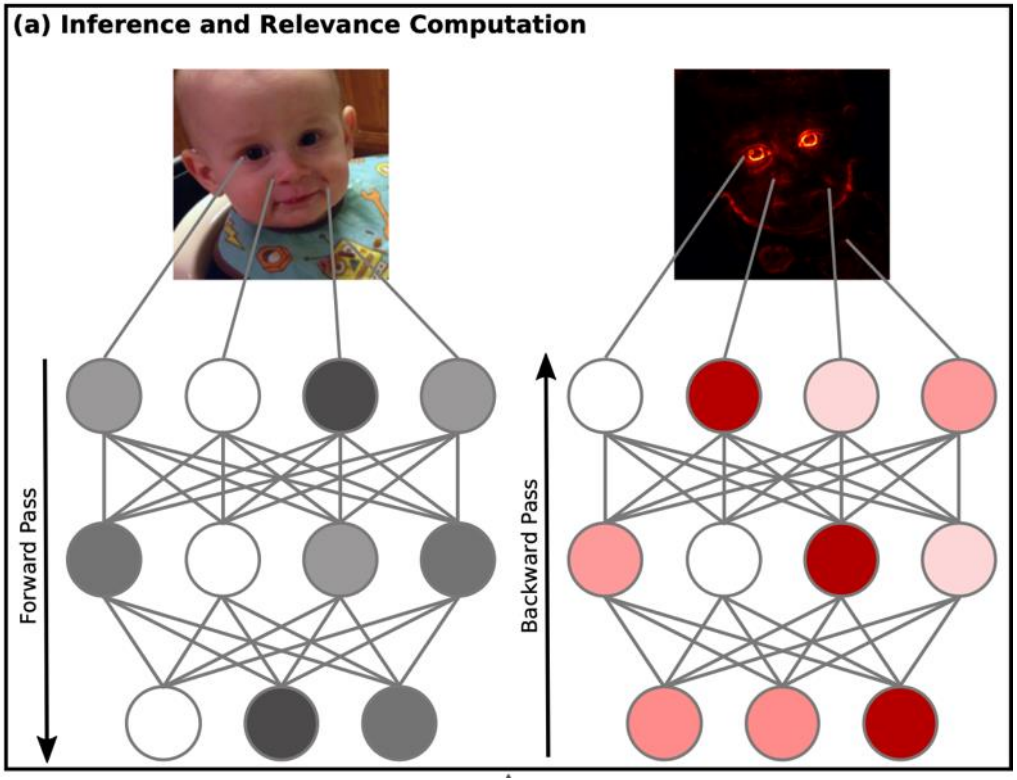
Without fine-tuning

Catastrophic Forgetting

Catastrophic forgetting is the tendency of an artificial neural network to abruptly and drastically forget previously learned information upon learning new information.



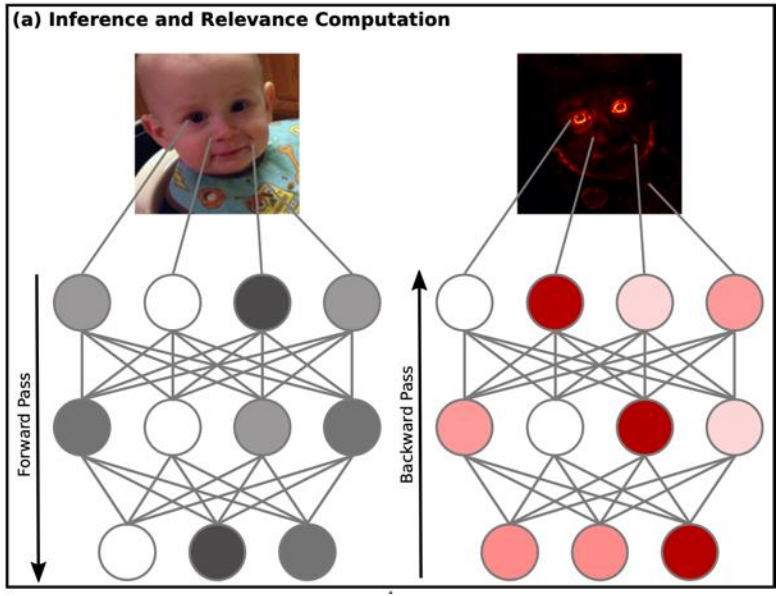
Tackling Catastrophic Forgetting with XAI

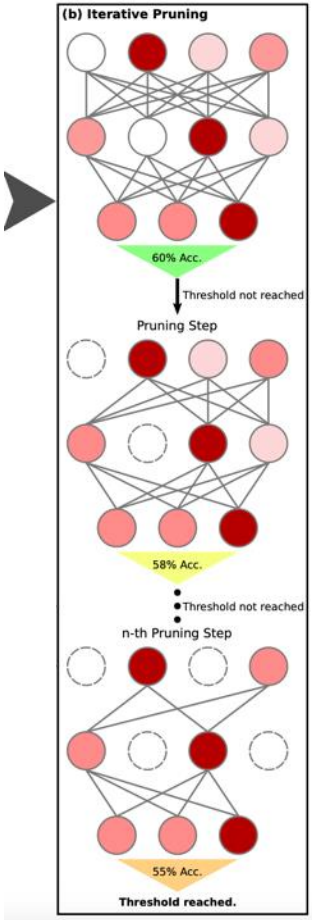


1. Step: Identify relevant subnetworks

(Ede et al. 2022)

Tackling Catastrophic Forgetting with XAI

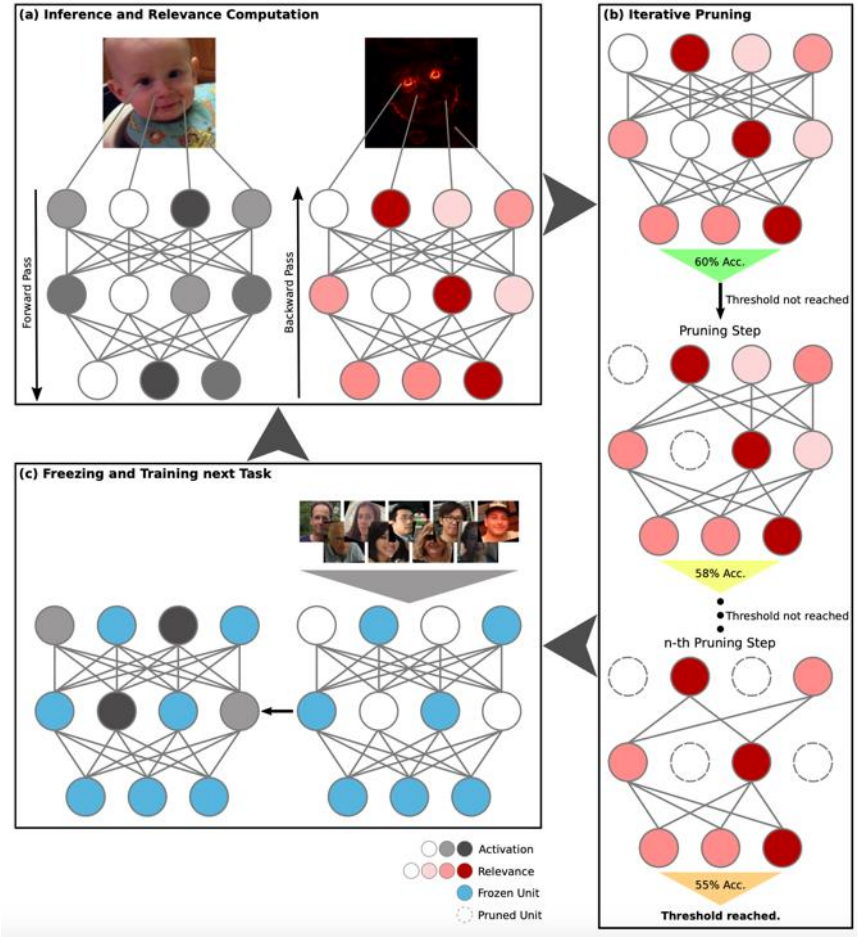




1. Step: Identify relevant subnetworks.
2. Prune model

(Ede et al. 2022)

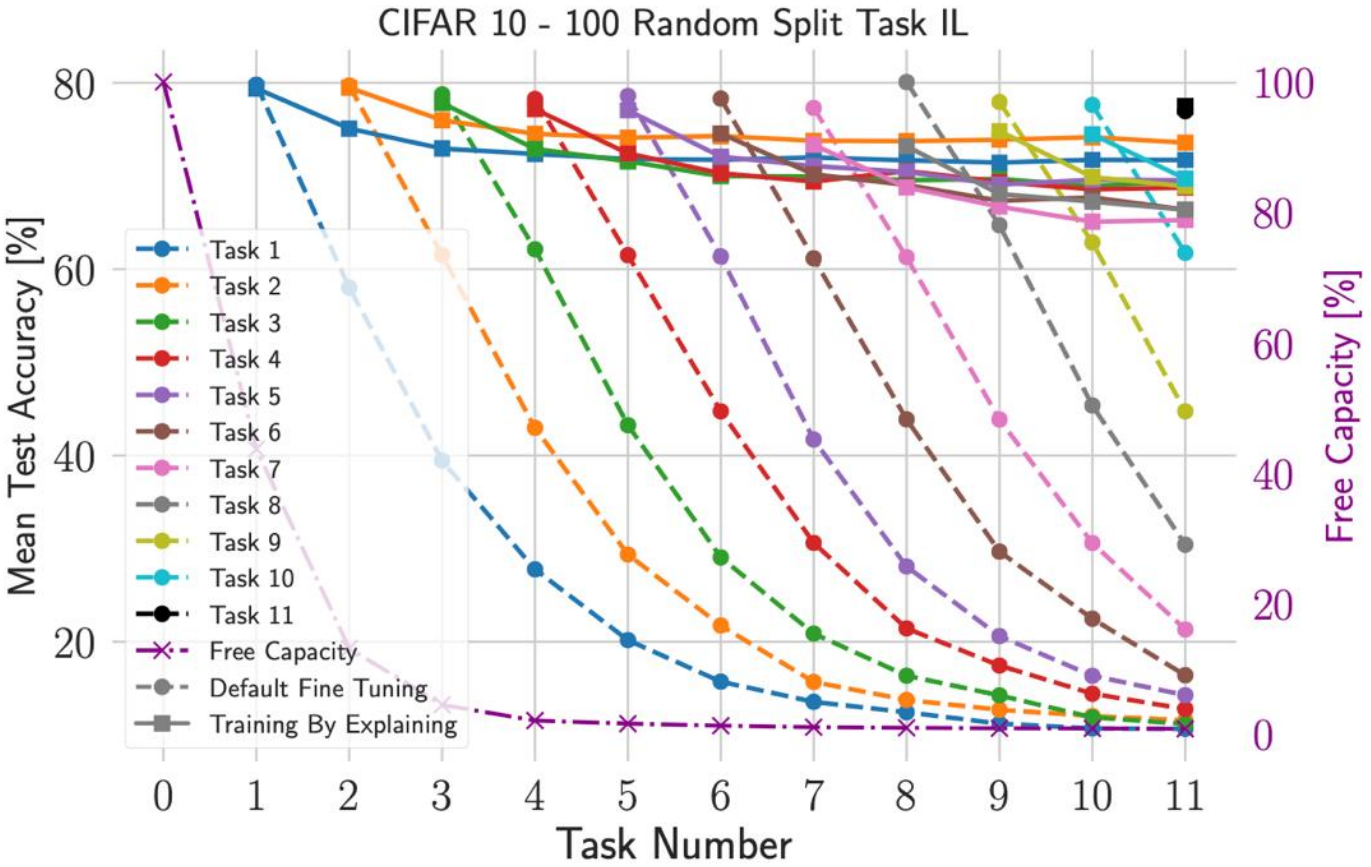
Tackling Catastrophic Forgetting with XAI



1. Step: Identify relevant subnetworks.
2. Prune model
3. Freeze and Train next Task

(Ede et al. 2022)

Tackling Catastrophic Forgetting with XAI



(Ede et al. 2022)

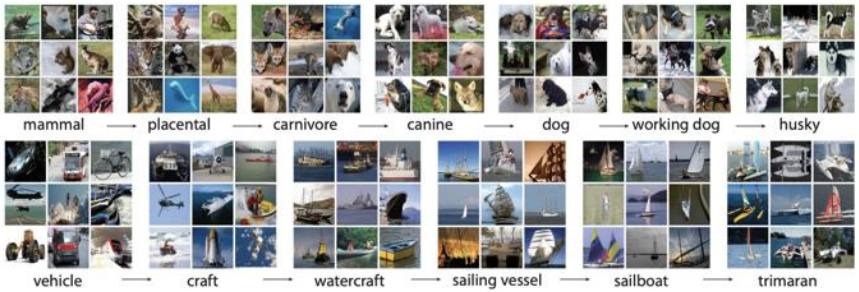
9th Take Home Message

Understanding the model better opens up a lot of opportunities beyond "just explaining" its predictions.

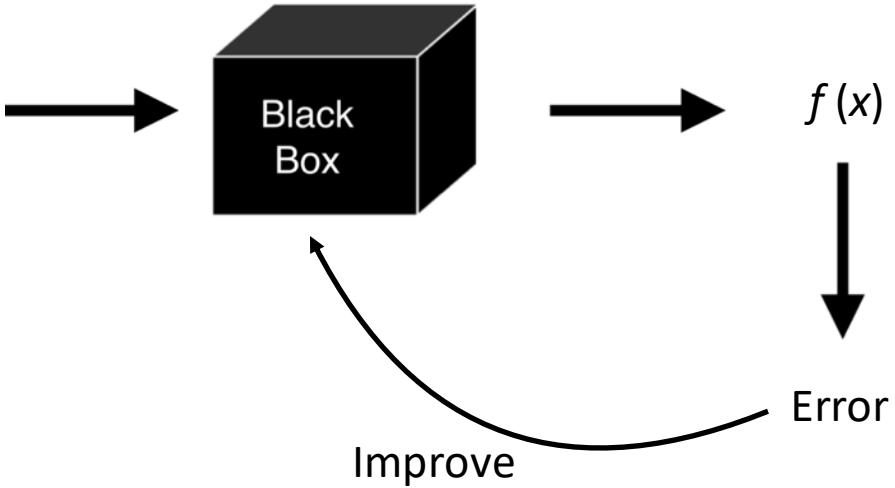
Reveal and Revise

Model Training

Training data *x* with labels *y*

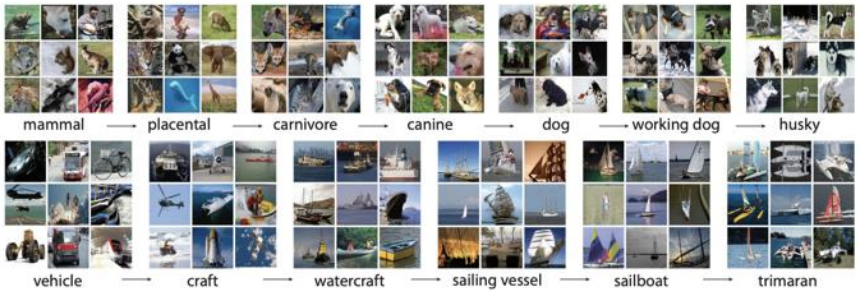


Predictions

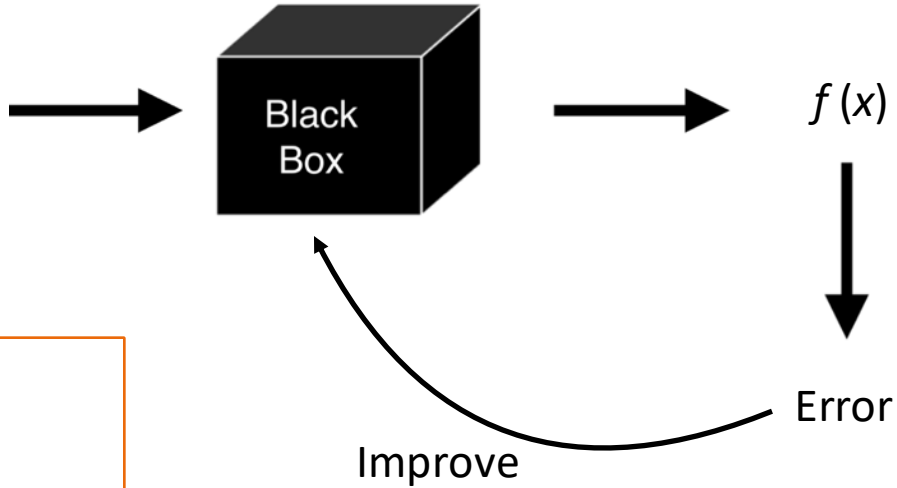


Model Training

Training data x with labels y



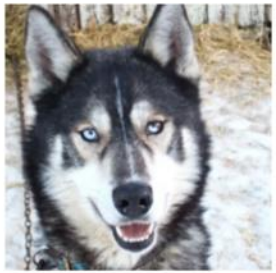
Predictions



$$\min_{f \in \mathcal{F}} \int_{x,y} \|f(\boldsymbol{x}) - y\|^2 dp(\boldsymbol{x}, y)$$

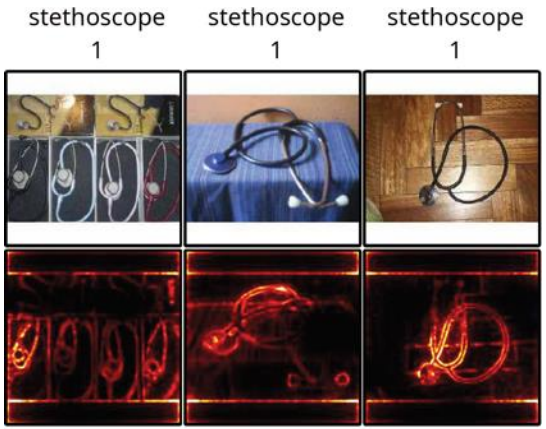
Is minimizing the error a guarantee for the model to work well in practice?

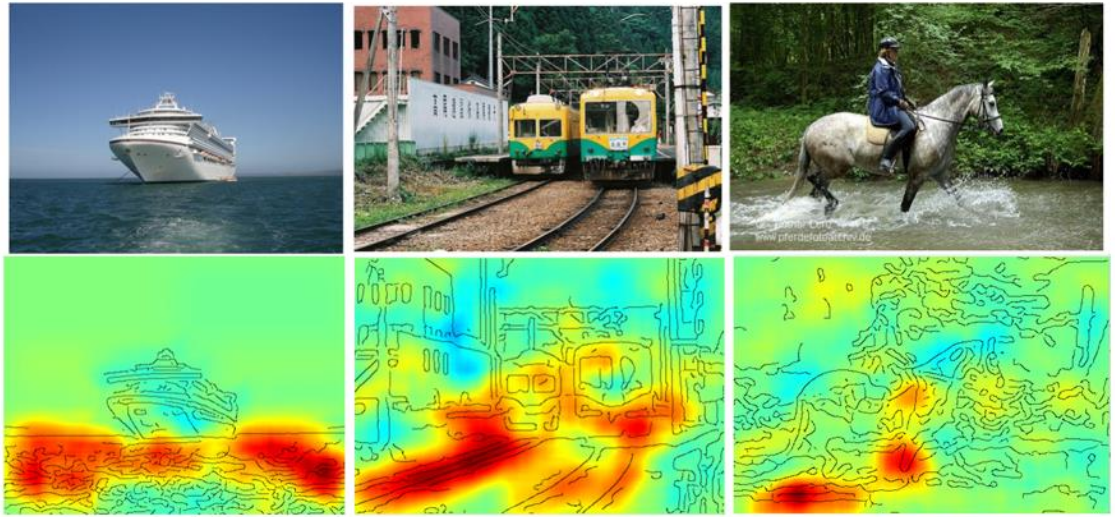
Clever Hans

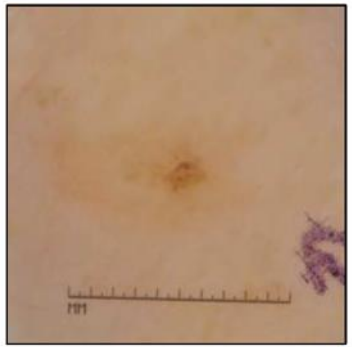


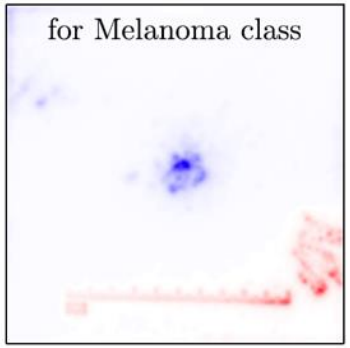
(a) Husky classified as wolf

(b) Explanation









XAI-Based Model Improvement

XAI-Bases model improvement aims to improve the:

Reasoning: alignment with human, better generalization (no Clever Hans), no bias

XAI-Based Model Improvement

XAI-Bases model improvement aims to improve the:

Reasoning: alignment with human, better generalization (no Clever Hans), no bias

Other Goals:

Performance: better test performance

Convergence: faster learning

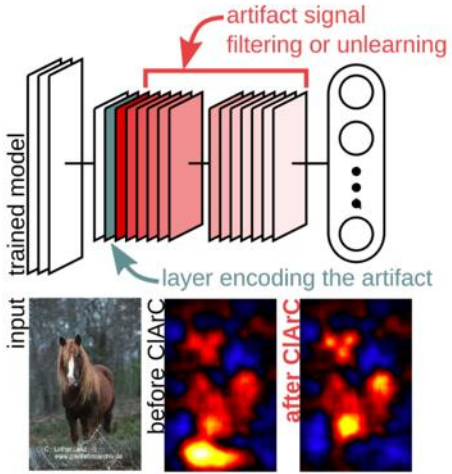
Robustness: robut against adversarial attackes and changes in distribution

Efficiency: less (labeled) data, faster inference, less energy

Equality: better handle unbalenced data, minority classes

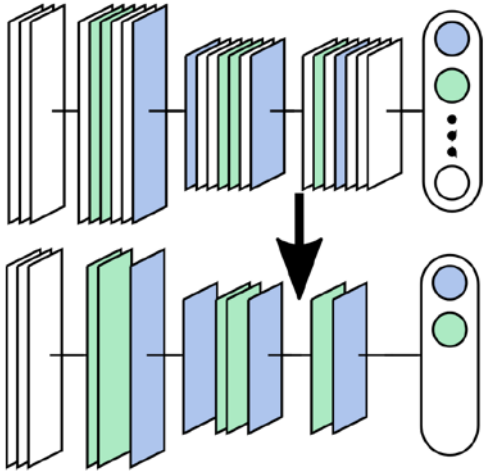
XAI-Based Model Improvement

With HiTL



``semantic'' model improvement, e.g., [Schramowski *et al.* 2020; Anders, Weber, *et al.* 2022]

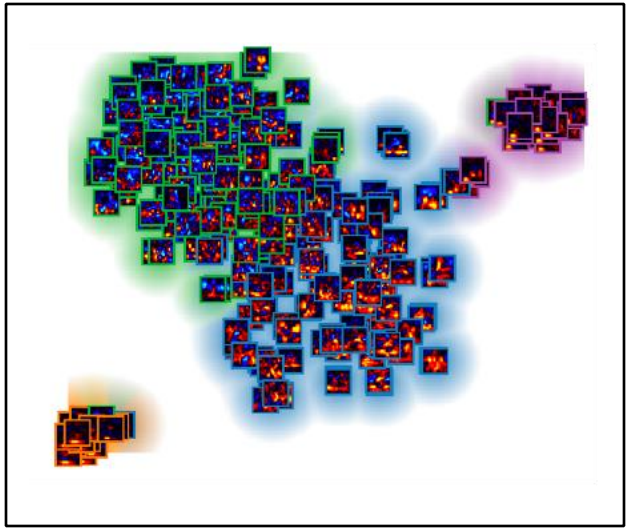
Without HiTL



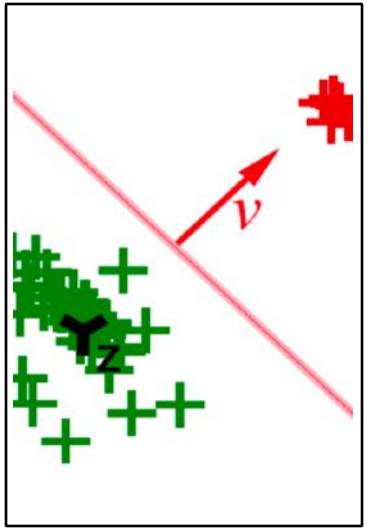
data-based model improvement, e.g., [Yeom *et al.* 2021; Becking *et al.* 2021; J. Sun, Lapuschkin, Samek, and Binder 2022]

Two Steps of Model Improvement

Analyse explanations

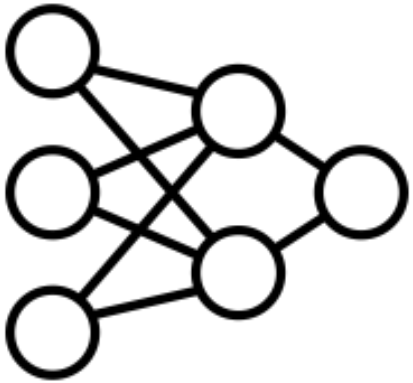


Artifact modelling



Clever Hans?

Correct model

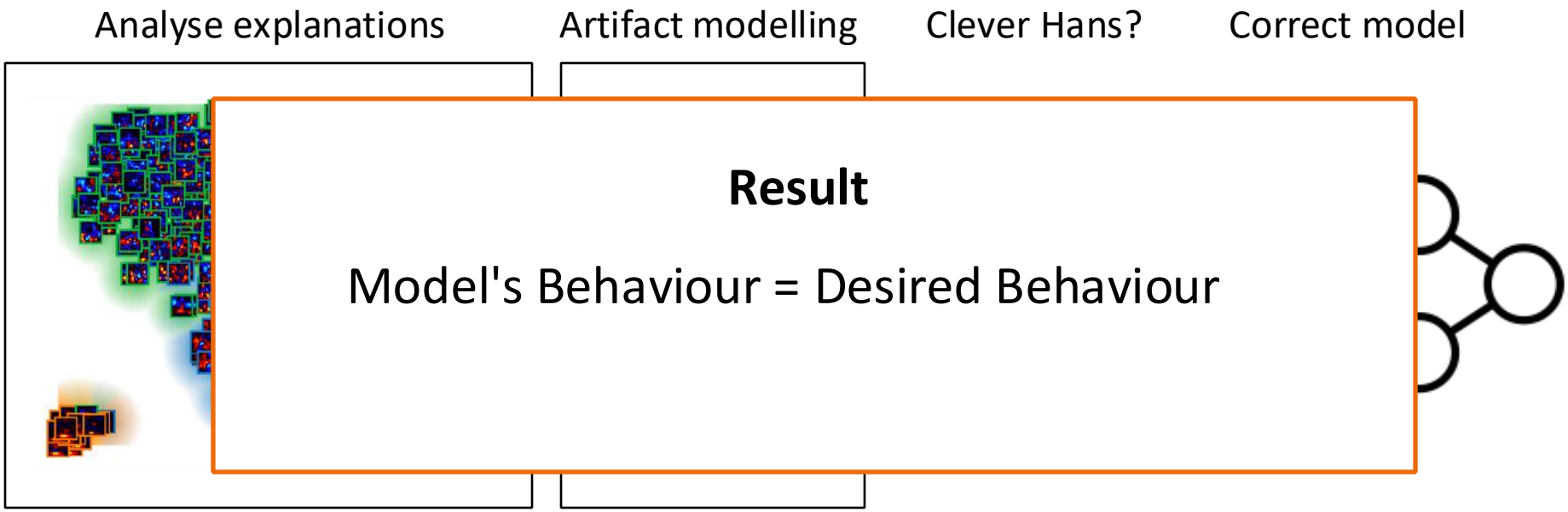


Step 1: Reveal

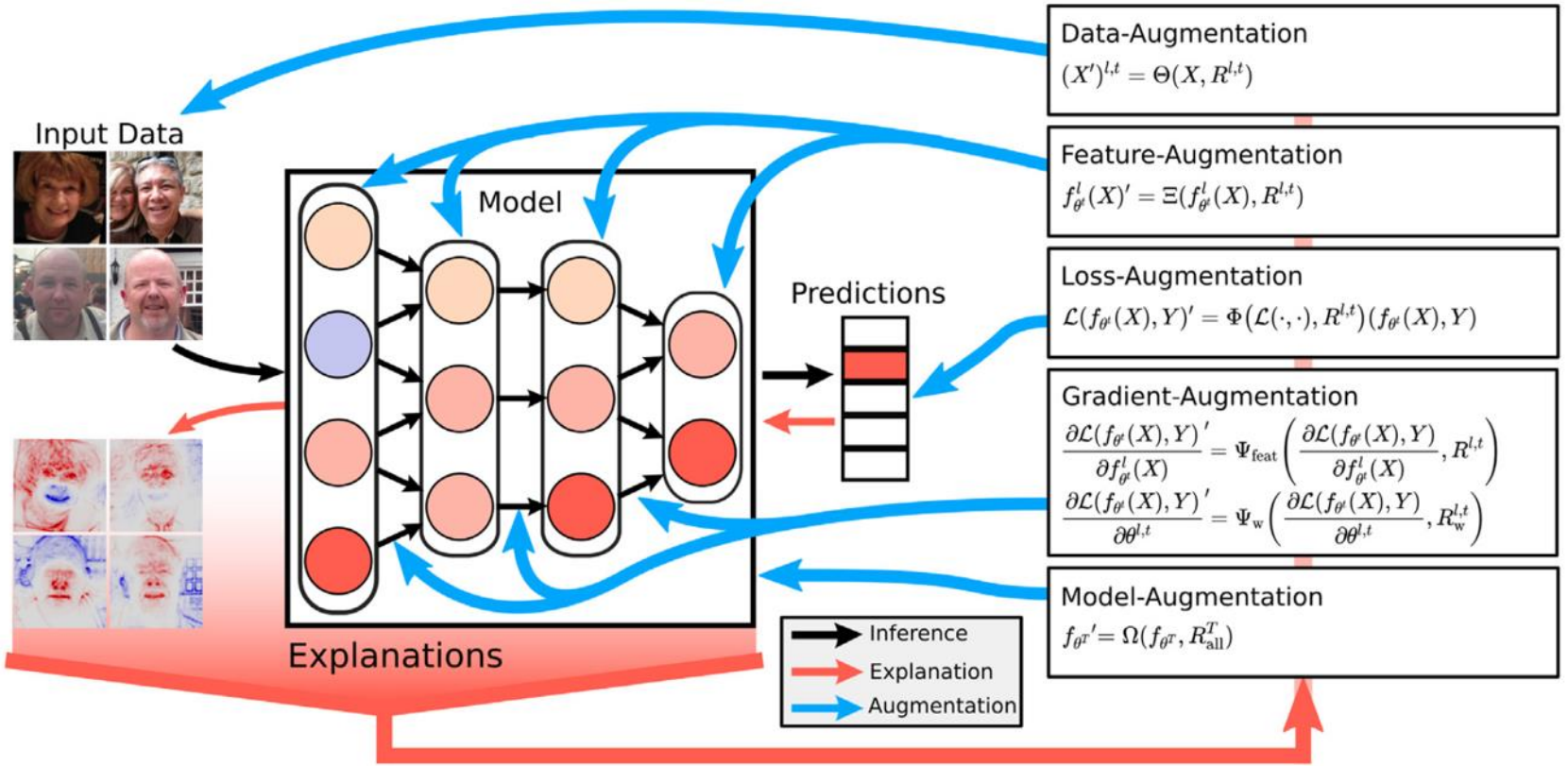
(or given a priori)

Step 2: Revise

Two Steps of Model Improvement



Where to Integrate Explanations?



(Weber et al. 2023)

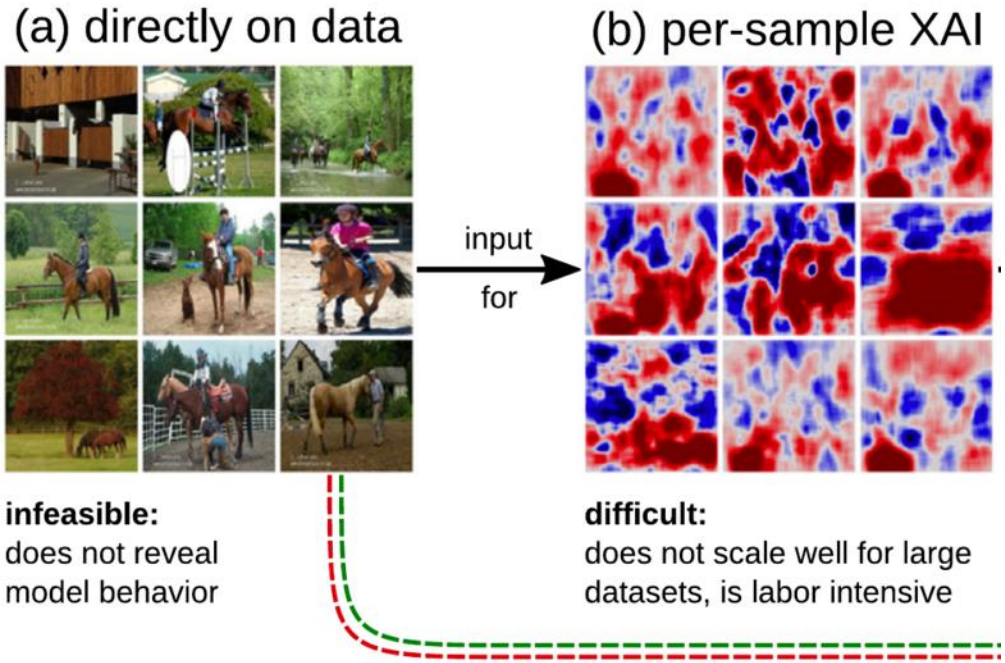
Step 1a: Detect the Artifact with Explanations (Reveal)

Reveal Clever Hans

(a) directly on data

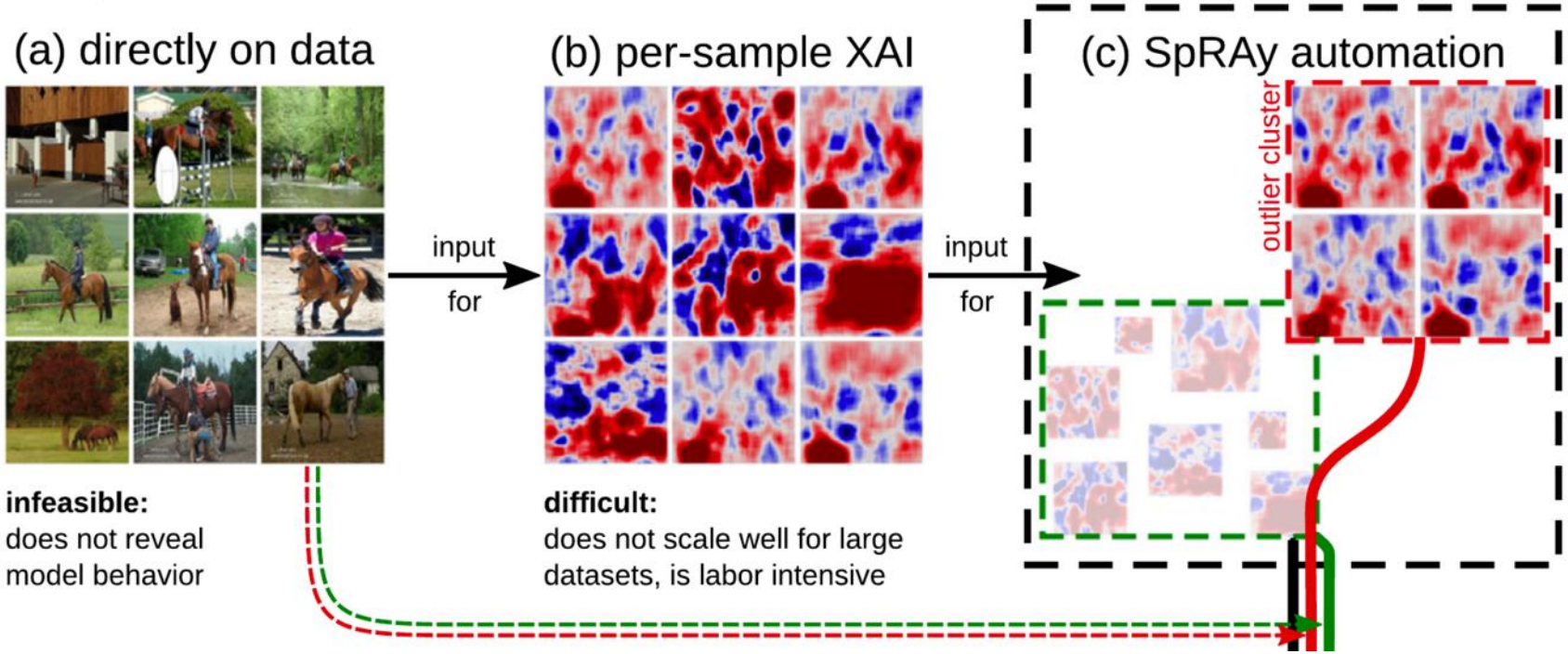
infeasible:
does not reveal
model behavior

Reveal Clever Hans



Reveal Clever Hans

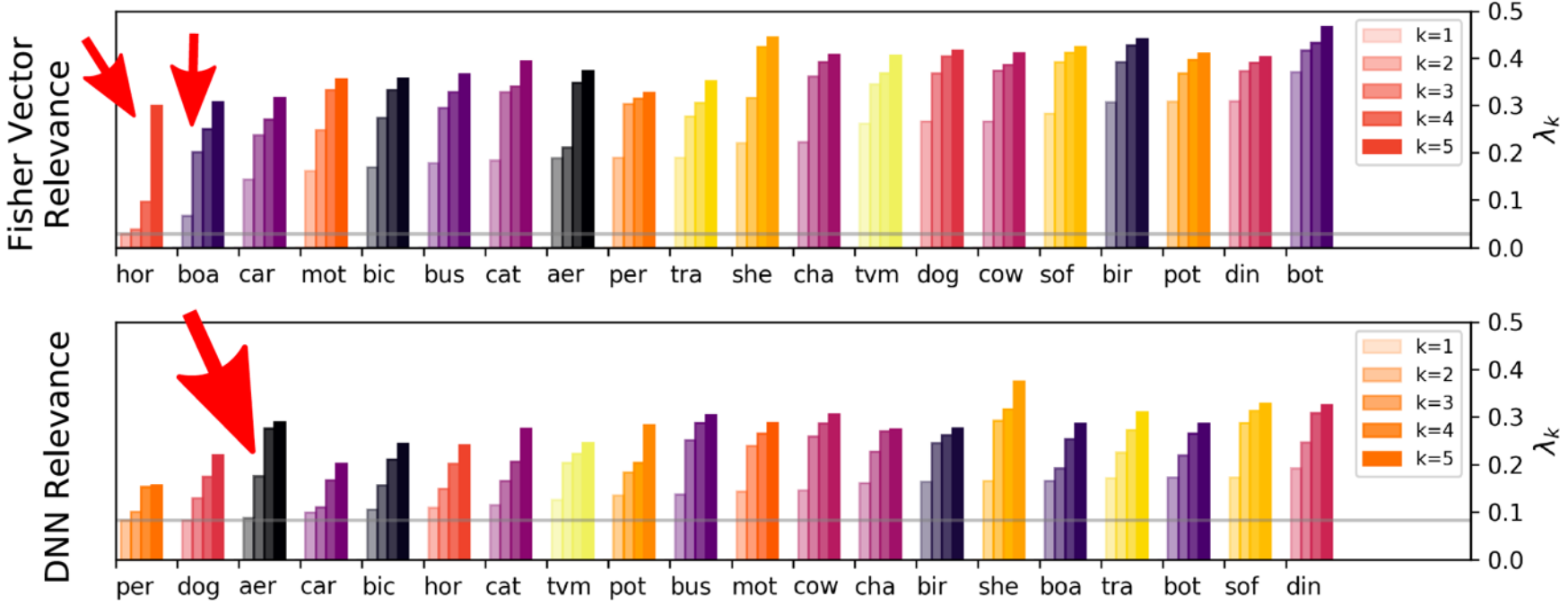
Idea: Clustering of explanations



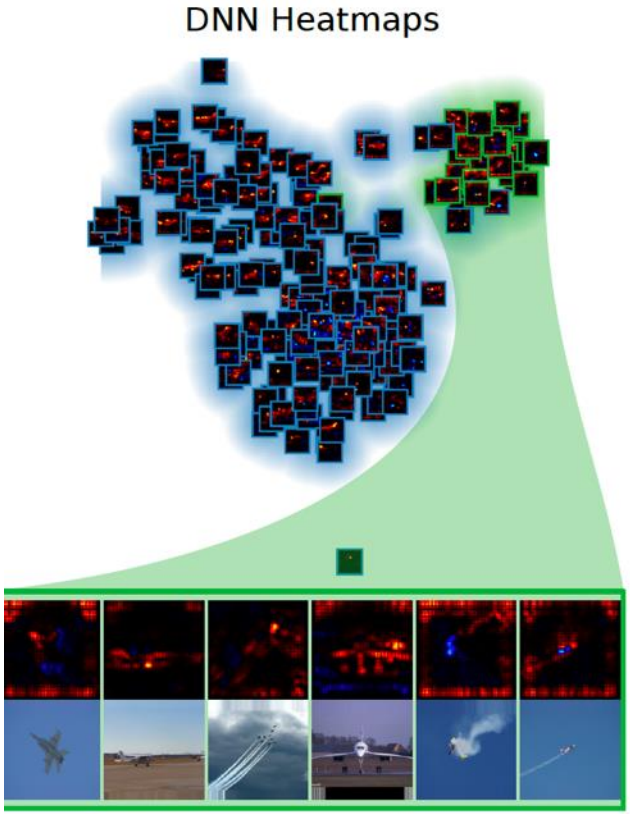
SpRAy = Spectral Relevance Analysis

Spectral Clustering

A strong increase in the difference between two successive eigenvalues (**eigengap**) indicates we



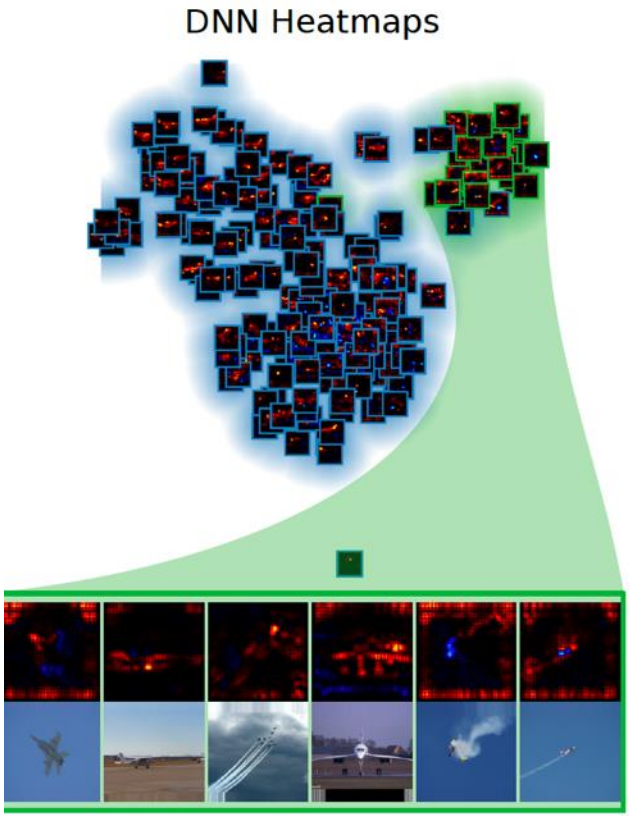
Aeroplane Cluster

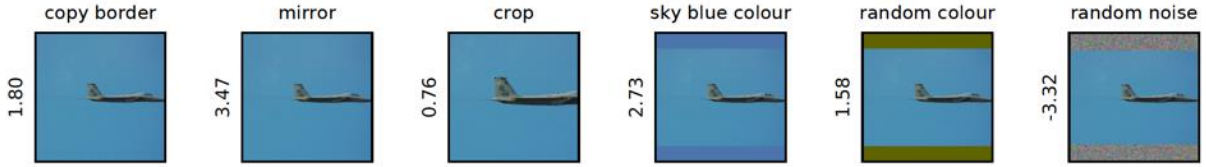


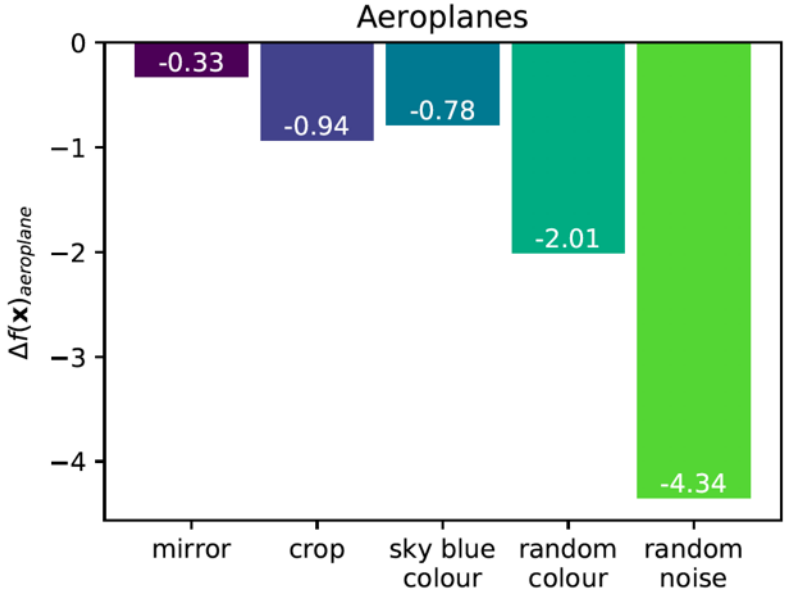
Observation: Artifact at the border.

(Lapuschkin et al. 2019)

Aeroplane Cluster

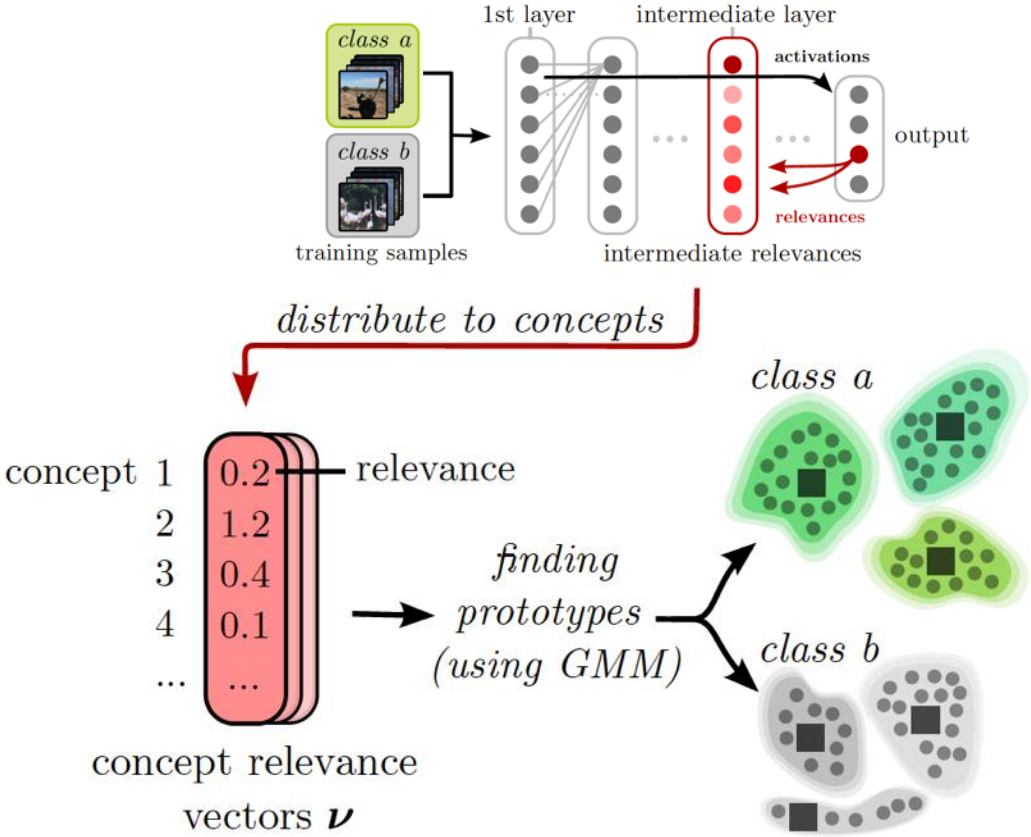






Limitation of Input Space Analysis: Natural invariances (rotation, translation, scaling) are difficult to capture.

Prototypical Concept-based Explanations

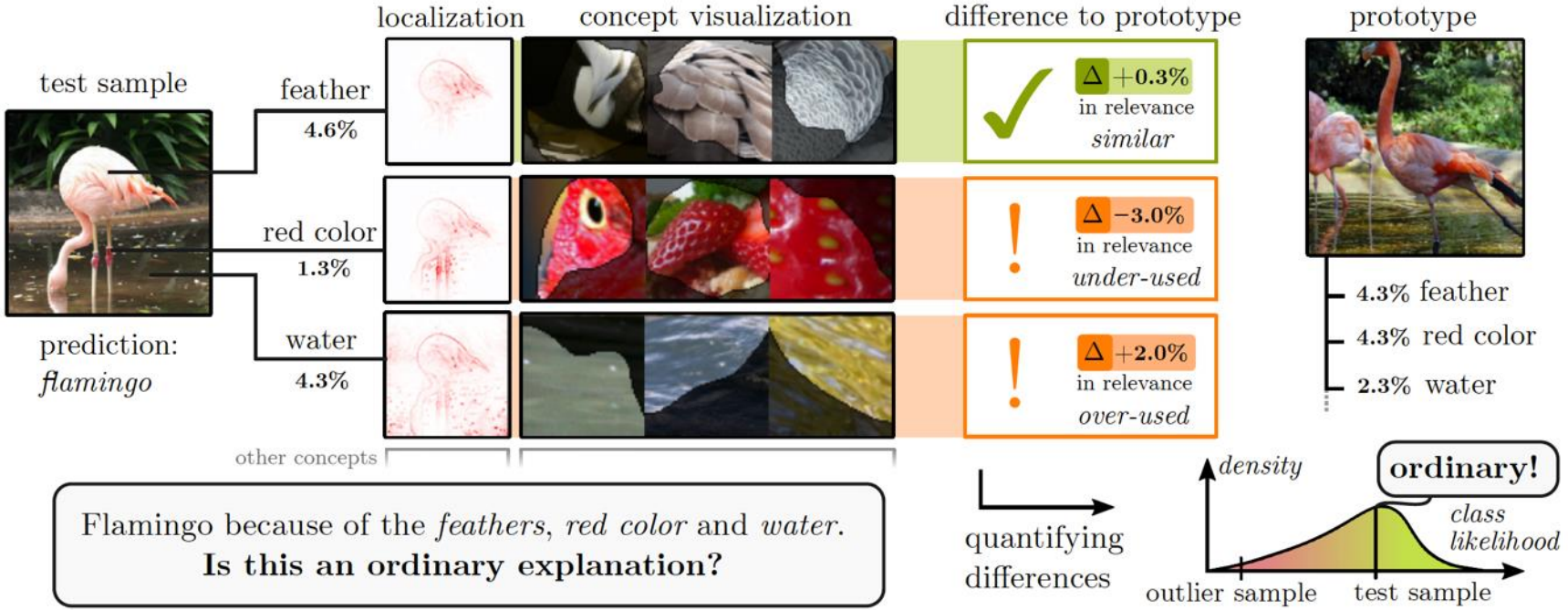


Quantifying the (Extra-)Ordinary:

$$L^k(\boldsymbol{\nu}) = \log p^k(\boldsymbol{\nu})$$

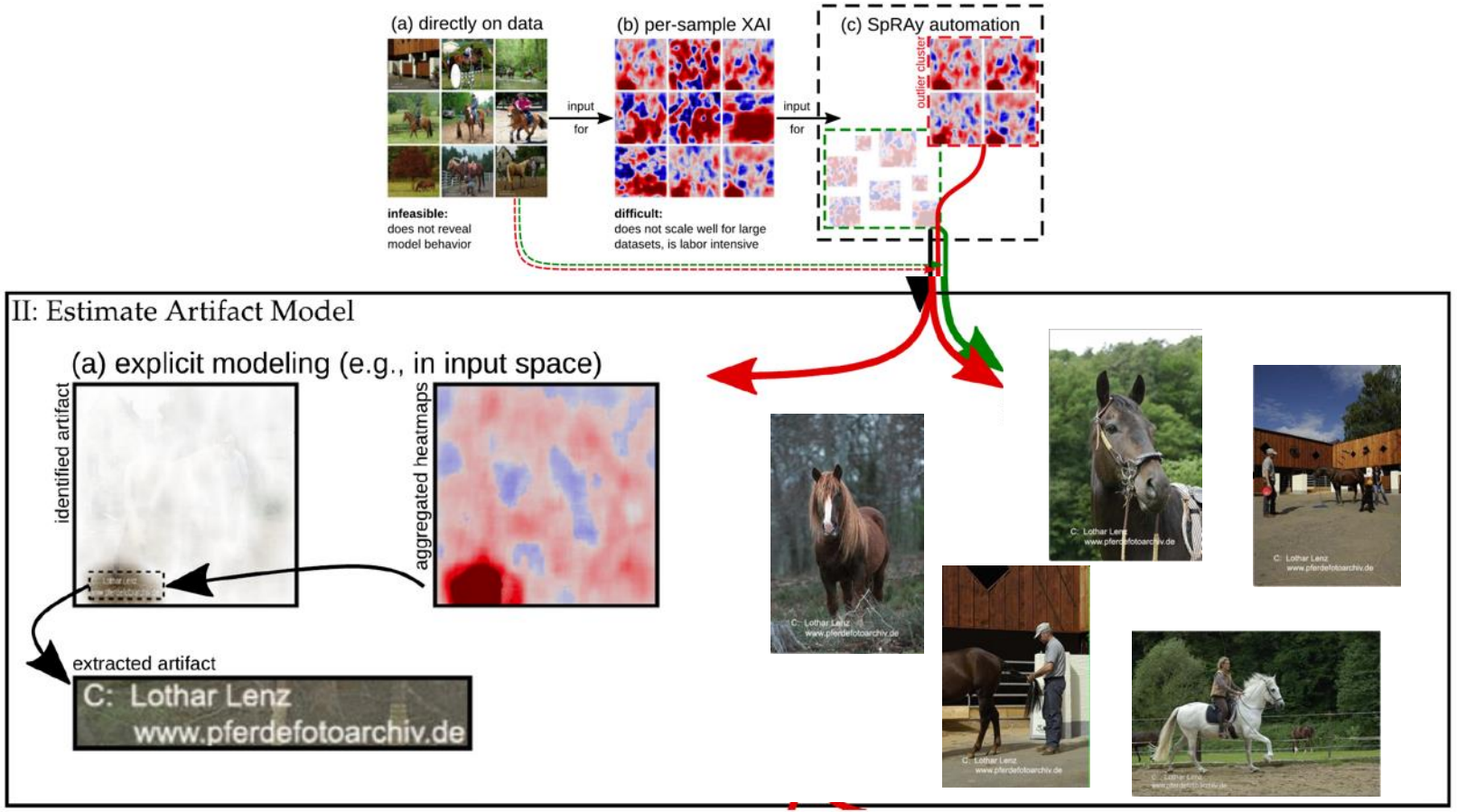
$$p_i^k(\boldsymbol{\nu}) = \frac{1}{(2\pi)^{\frac{n}{2}} \det(\boldsymbol{\Sigma}_i^k)^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{\nu}-\boldsymbol{\mu}_i^k)^\top (\boldsymbol{\Sigma}_i^k)^{-1}(\boldsymbol{\nu}-\boldsymbol{\mu}_i^k)}$$

Prototypical Concept-based Explanations

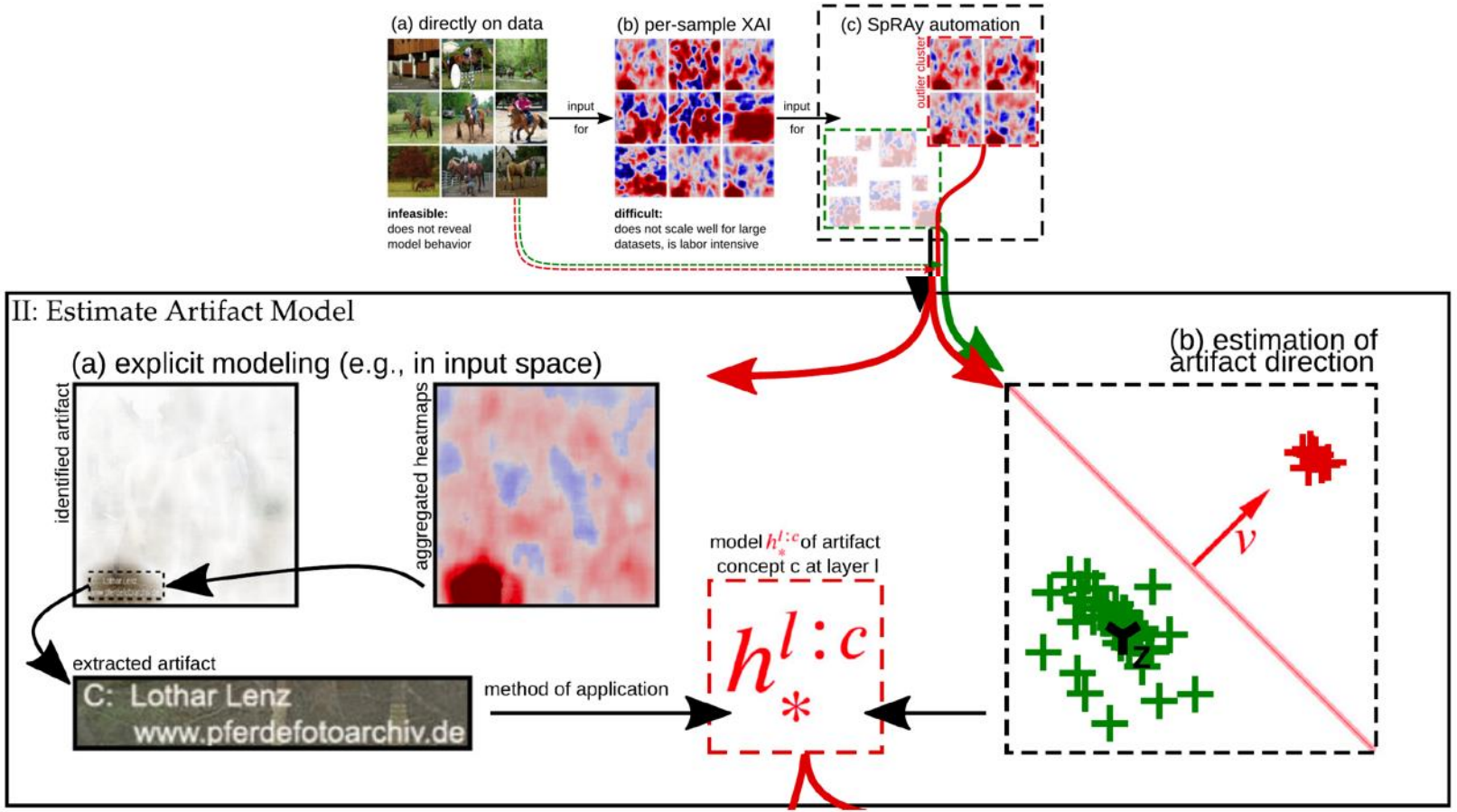


Step 1b: Model the Artifact

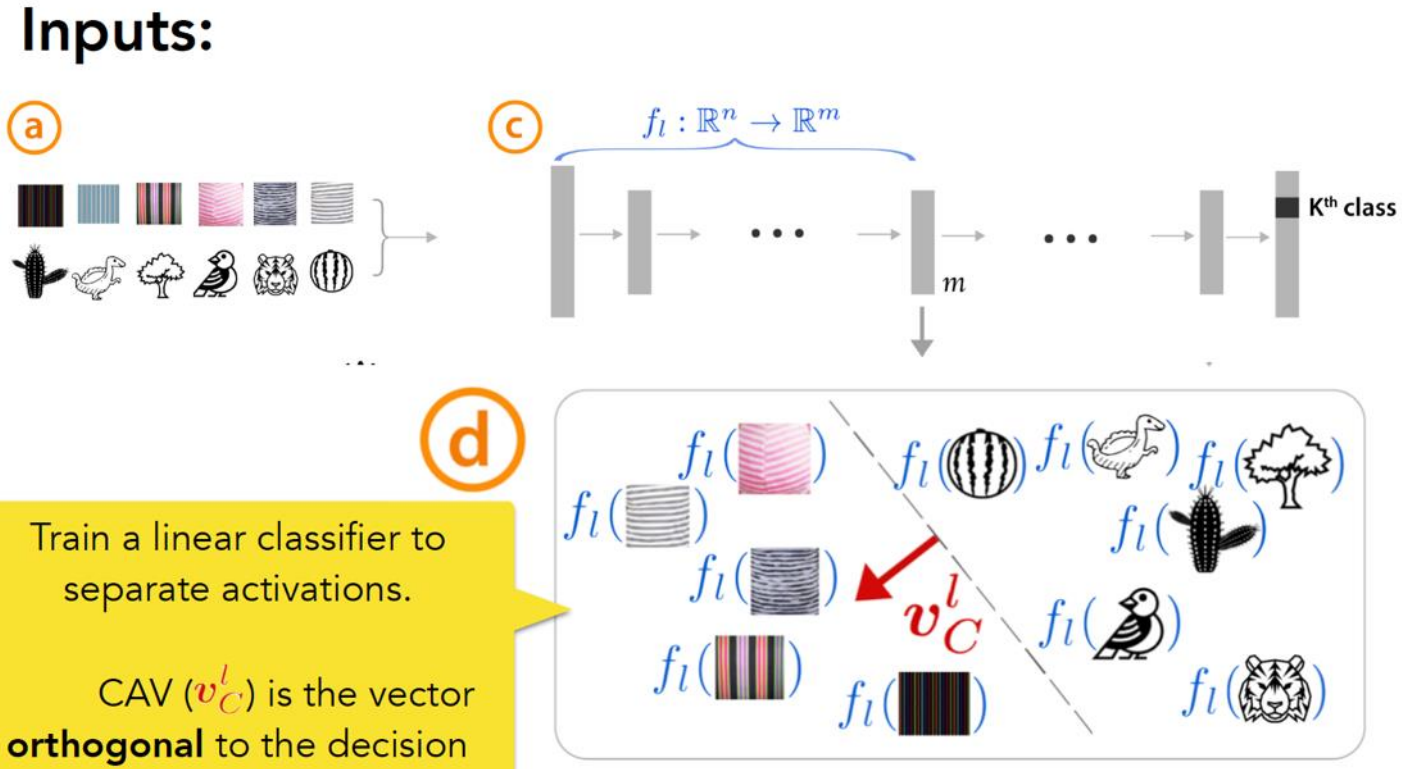
Estimate Artifact Model



Estimate Artifact Model



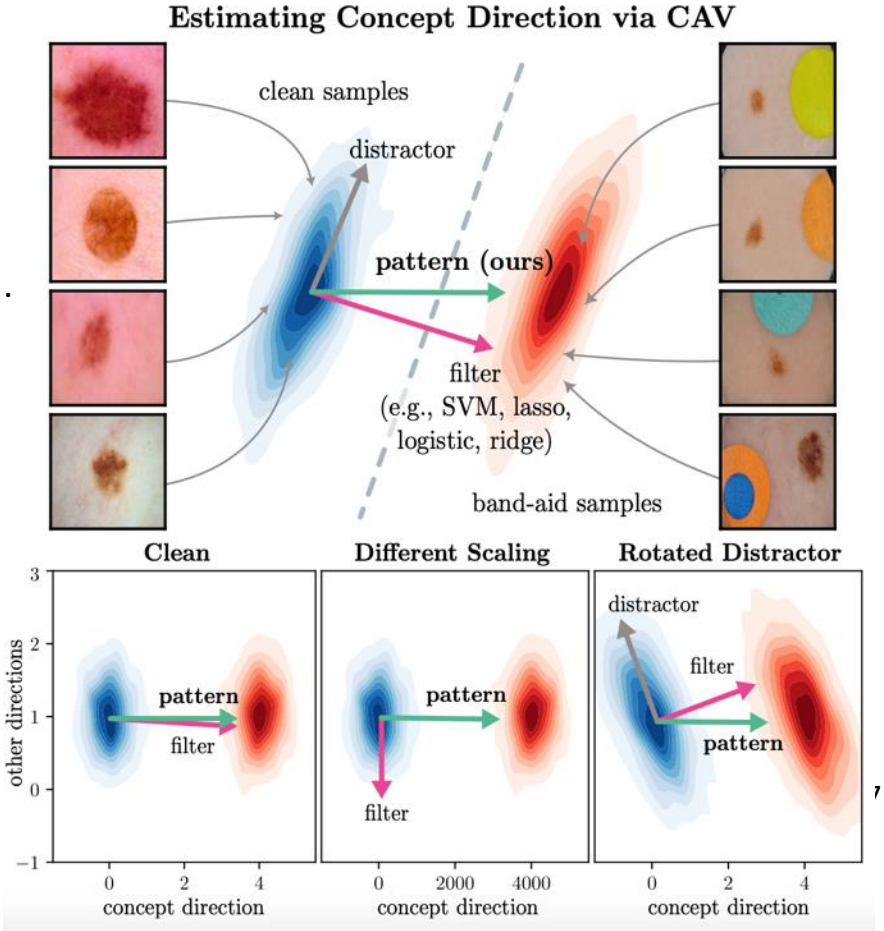
Revisited: Concept Activation Vector (CAV)



Pattern-CAV: New Noise Resilient CAV

Filter-Pattern Problem (Haufe et al. 2014):

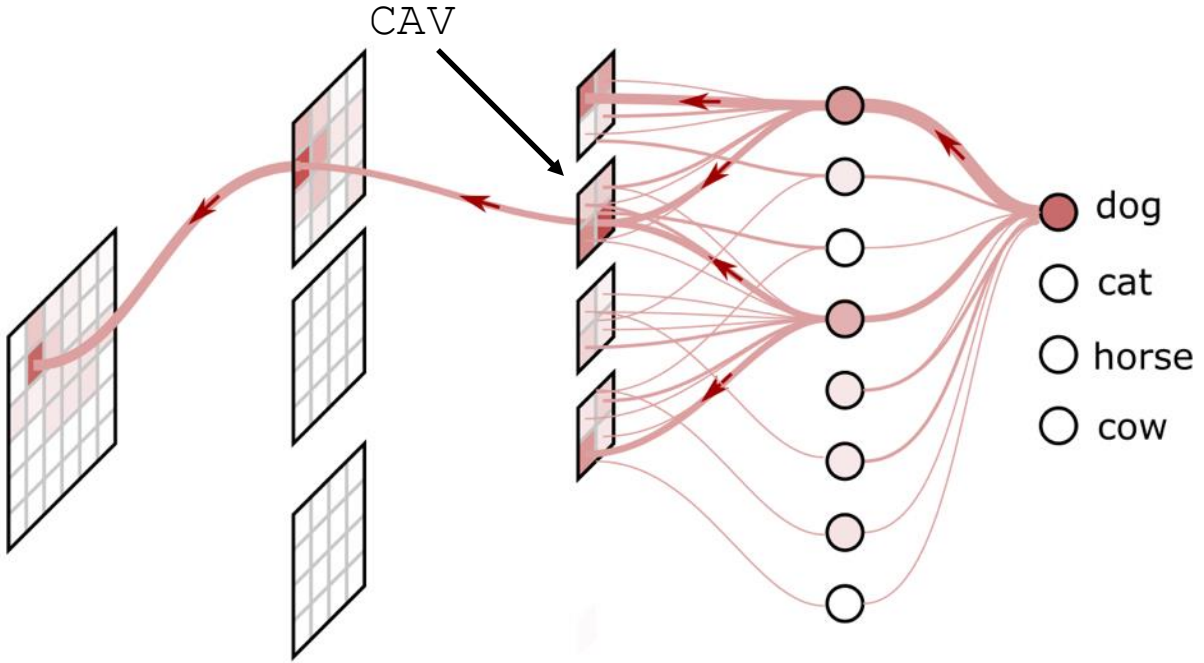
Significant influence of distractor (i.e., non-signal) directions contained in the data, which are picked up by filters (i.e., weights) of linear models to optimize class-separability



Navigating Neural Space: Revisiting Concept Activation Vectors to Overcome Directional Divergence

Frederik Pahde[1], Maximilian Dreyer[1], Leander Weber[1], Moritz Weckbecker[1], Christopher J. Anders[2,3], Thomas Wiegand[1,2,3], Wojciech Samek[1,2,3,†], Sebastian Lapuschkin[1,†]

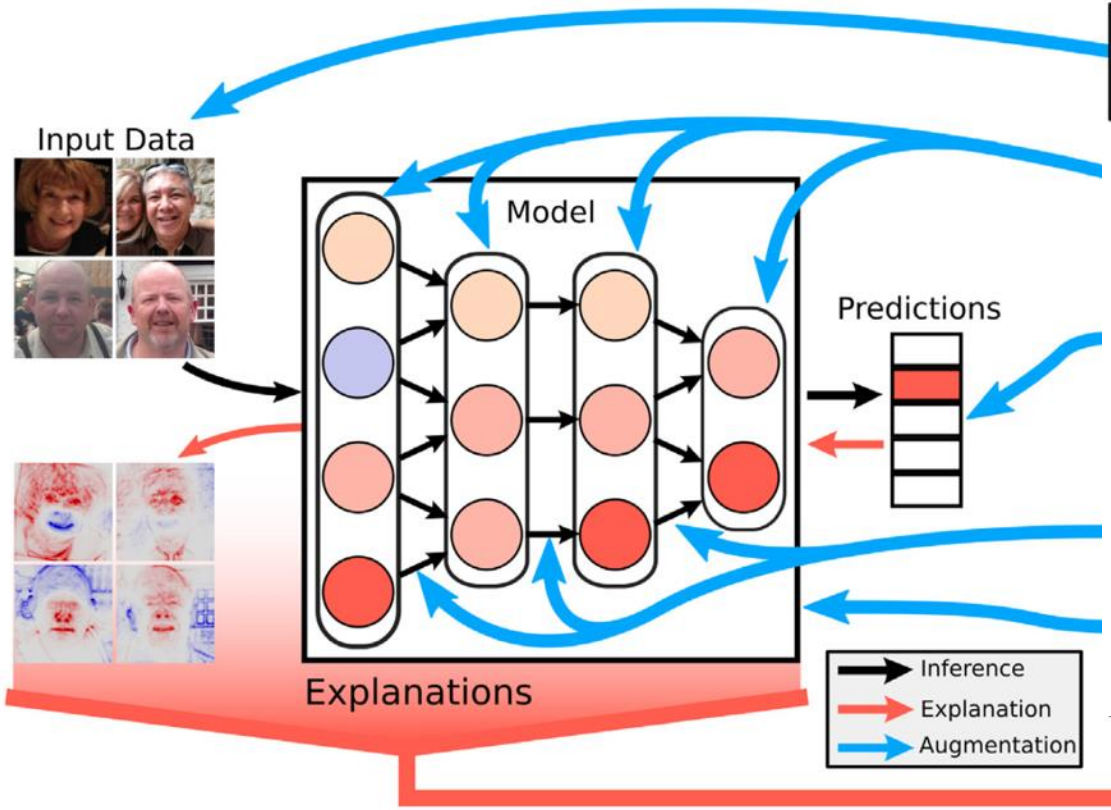
Back to Input Space



LRP / CRP can be used to localize CAV in input space.

Step 2: Correct the Model Behaviour (Revise)

XAI-Based Model Improvement



Generate new samples depending on explanations or alter the distribution of existing data

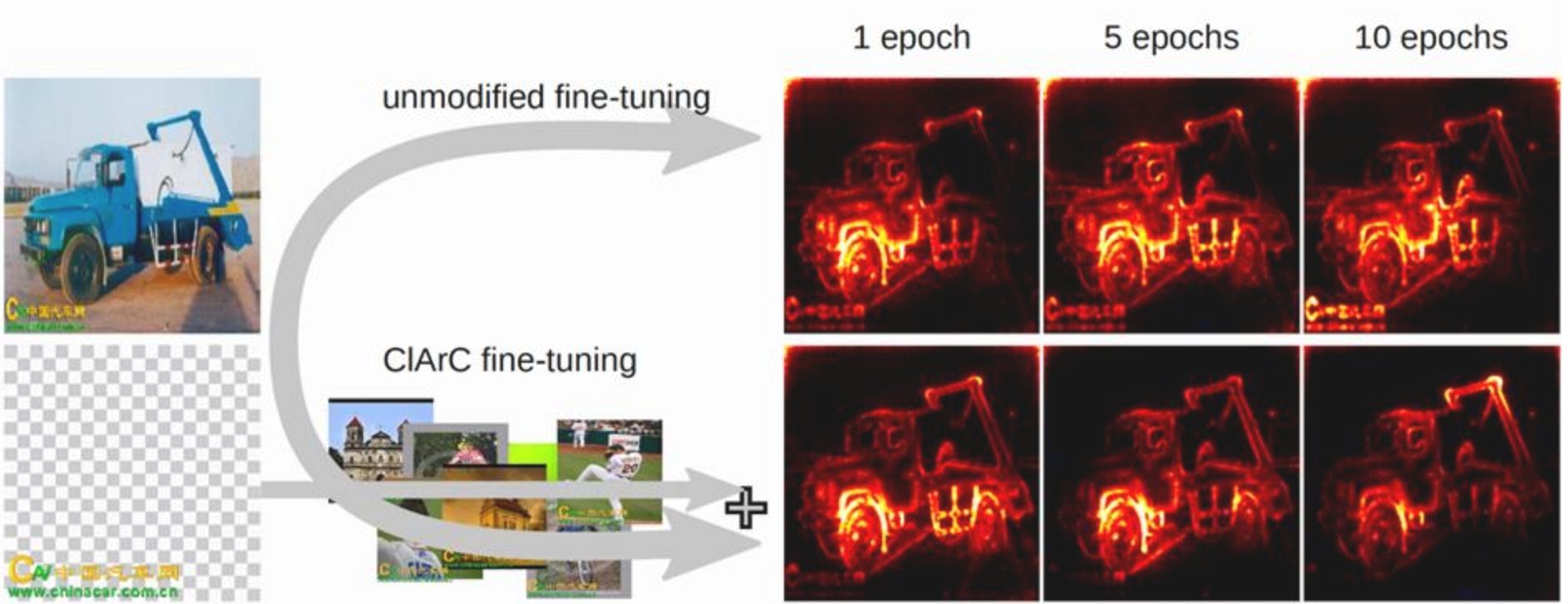
Data Augmentation

Isolate artefact

Data Augmentation

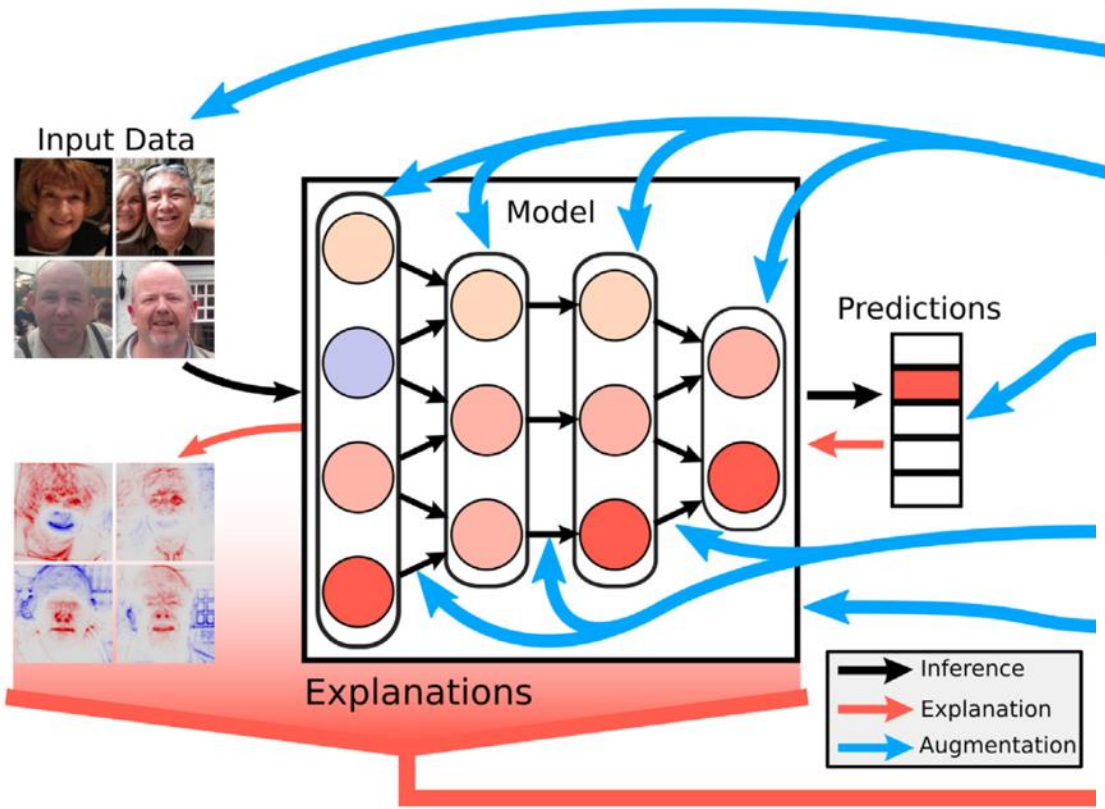
Isolate artefact, add to *other/all* classes

Data Augmentation



(Anders et al. 2022)

XAI-Based Model Improvement



Data-Augmentation

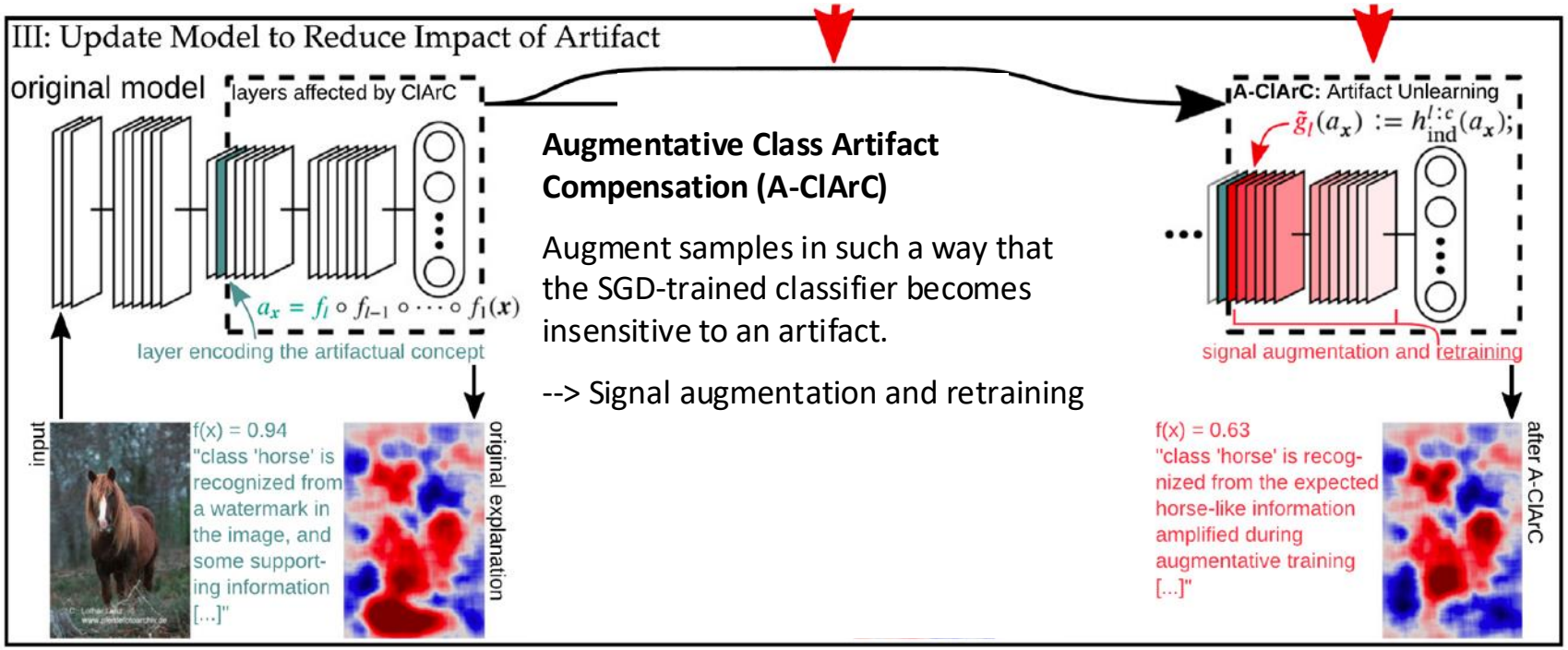
$(X')^{l,t} = \Theta(X, R^{l,t})$

Feature-Augmentation

$f^l_{\theta'}(X)' = \Xi(f^l_{\theta'}(X), R^{l,t})$

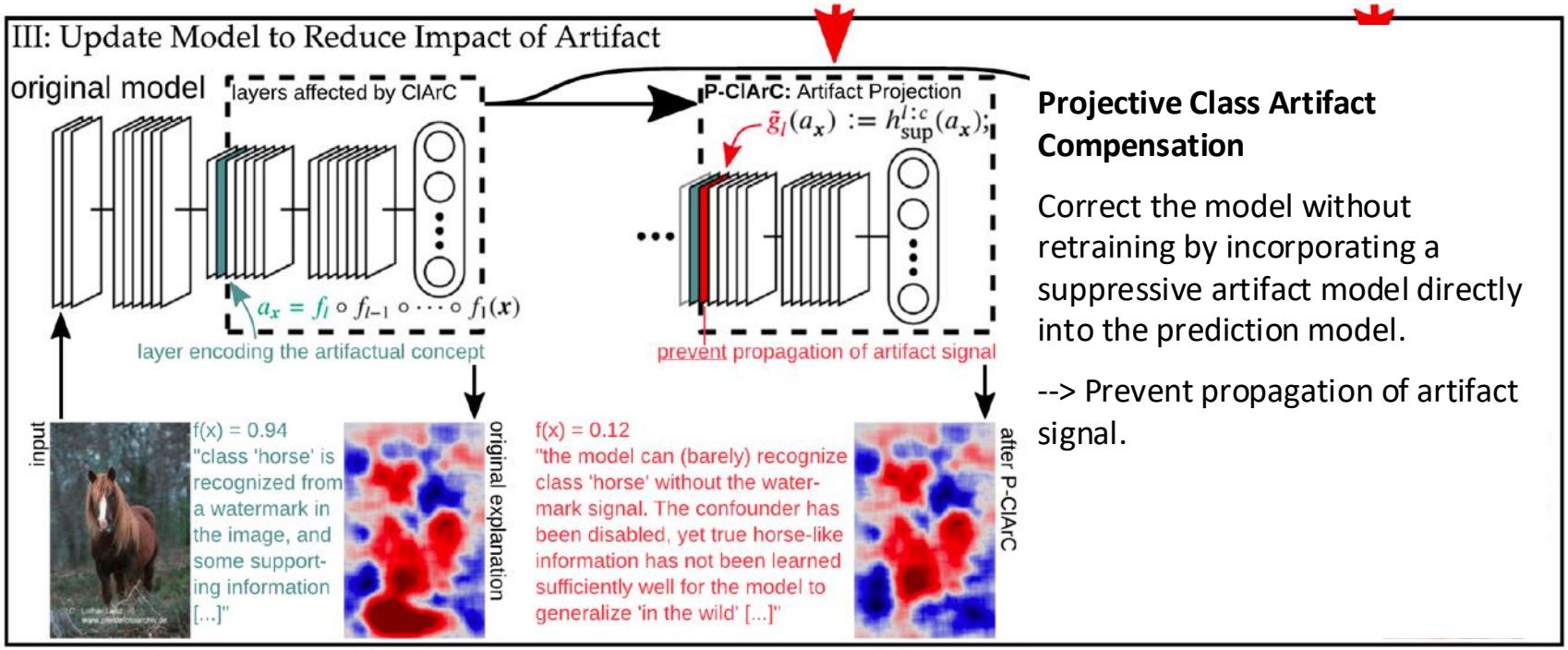
Explanations can be leveraged to scale, mask, or transform intermediate features

Feature Augmentation



inductive removal with **A-ClArC**.

Feature Augmentation



Projective Class Artifact Compensation

Correct the model without retraining by incorporating a suppressive artifact model directly into the prediction model.

--> Prevent propagation of artifact signal.

inductive removal with **A-ClArC**. *suppressive* removal with **P-ClArC**.

(Anders et al. 2022)

Augmentative Class Artifact Compensation

Pushing samples around, to artifact reference point z

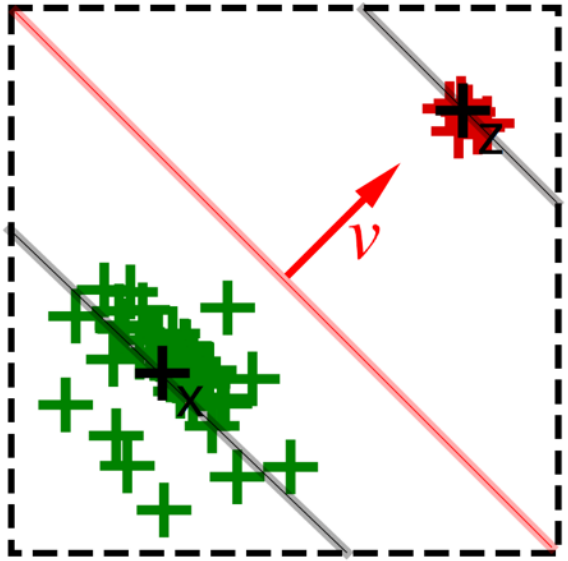
activation vector

artifact CAV

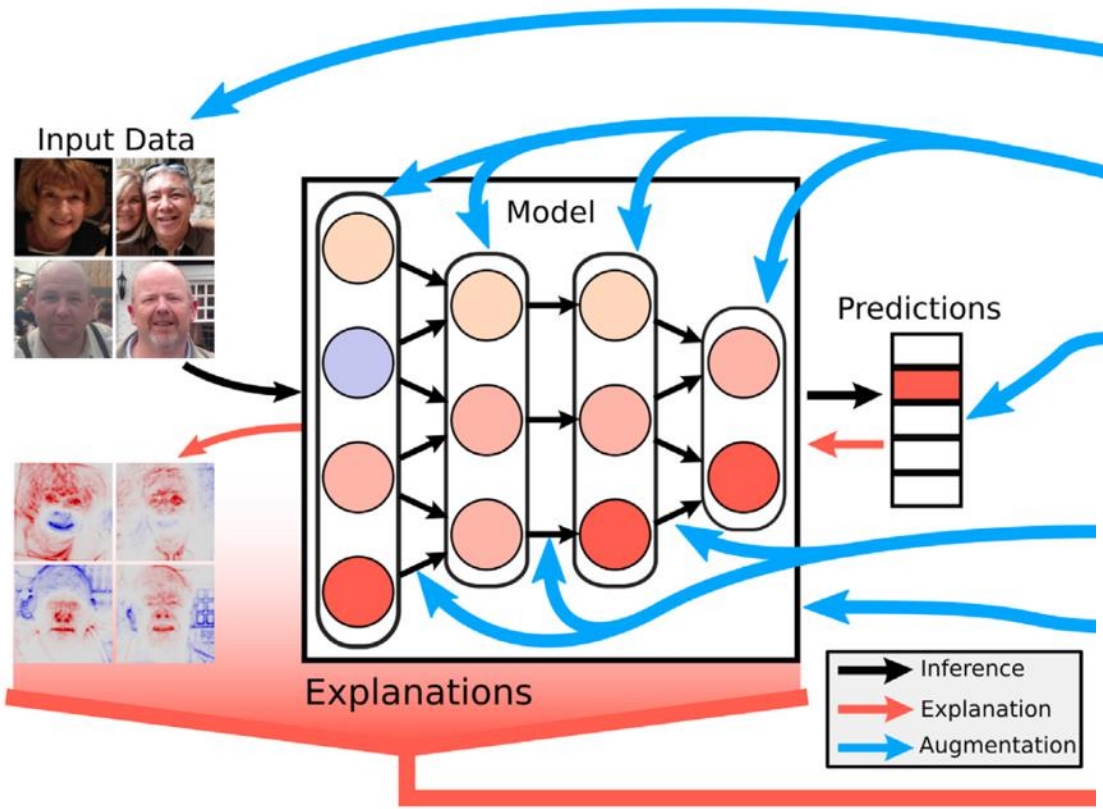
$$h_{ind}(x) = x + v \cdot (-v^\top x + v^\top z)$$

$$z = \frac{1}{|X^+|} \sum_{x^+ \in X^+} x^+$$

pushig samples over the decision boundary (e.g. SVM CAV)



XAI-Based Model Improvement



Data-Augmentation

$(X')^{l,t} = \Theta(X, R^{l,t})$

Feature-Augmentation

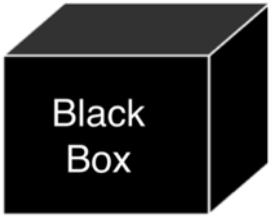
$f^l_{\theta'}(X)' = \Xi(f^l_{\theta'}(X), R^{l,t})$

Loss-Augmentation

$\mathcal{L}(f_{\theta'}(X), Y)' = \Phi\big(\mathcal{L}(\cdot,\cdot), R^{l,t}\big)(f_{\theta'}(X), Y)$

Augmenting the loss function based on explanations can help specify which behavior is desired, using explanations as a feedback.

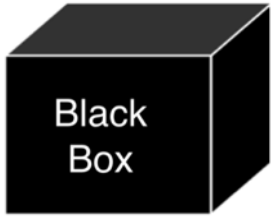
Right for the Right Reason



"Look at the horse!"

Otherwise you will be penalized

Right for the Right Reason



"Look at the horse!"

Otherwise you will be penalized

"explanation" (sensitivity)

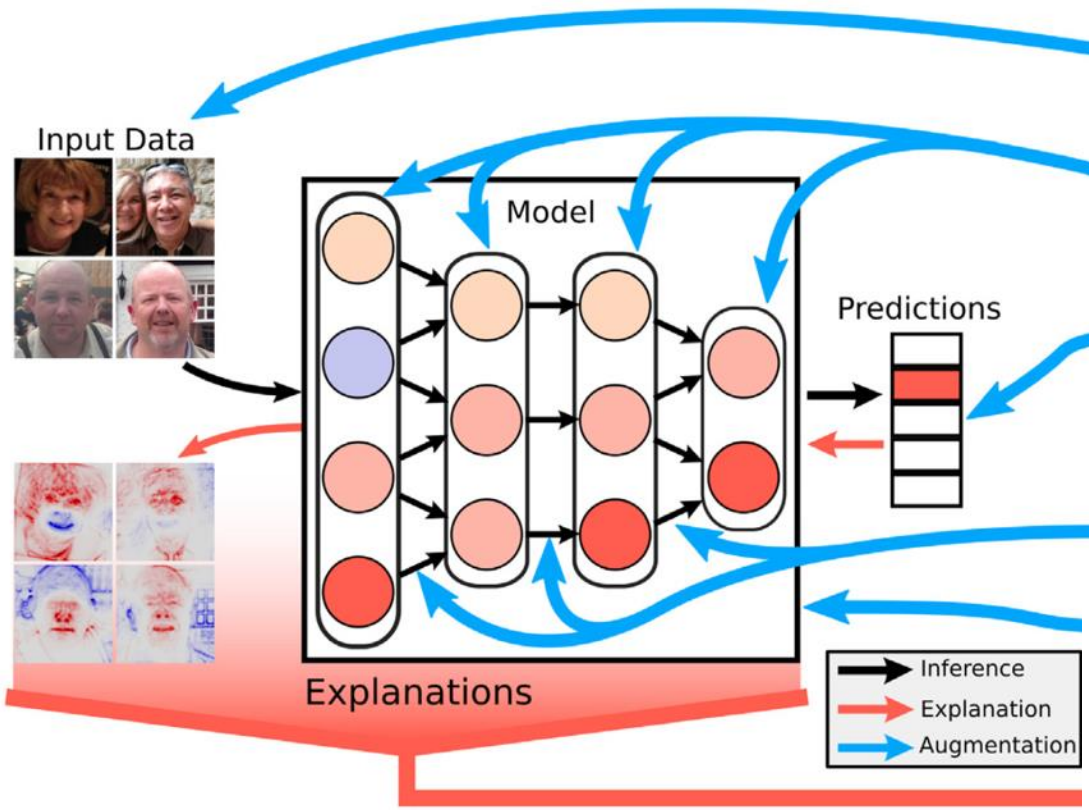
binary mask

$$L(\theta, X, y, A) = \underbrace{\sum_{n=1}^{N}\sum_{k=1}^{K} -y_{nk}\log(\hat{y}_{nk})}_{\text{Right answers}}$$

$$+ \underbrace{\lambda_1 \sum_{n=1}^{N}\sum_{d=1}^{D}\left(A_{nd}\frac{\partial}{\partial x_{nd}}\sum_{k=1}^{K}\log(\hat{y}_{nk})\right)^2}_{\text{Right reasons}} + \underbrace{\lambda_2 \sum_{i}\theta_i^2}_{\text{Regular}},$$

(Ross et al. 2017)

XAI-Based Model Improvement



Data-Augmentation

$$(X')^{l,t} = \Theta(X, R^{l,t})$$

Feature-Augmentation

$$f^l_{\theta'}(X)' = \Xi(f^l_{\theta'}(X), R^{l,t})$$

Loss-Augmentation

$$\mathcal{L}(f_{\theta'}(X), Y)' = \Phi\big(\mathcal{L}(\cdot,\cdot), R^{l,t}\big)(f_{\theta'}(X), Y)$$

Gradient-Augmentation

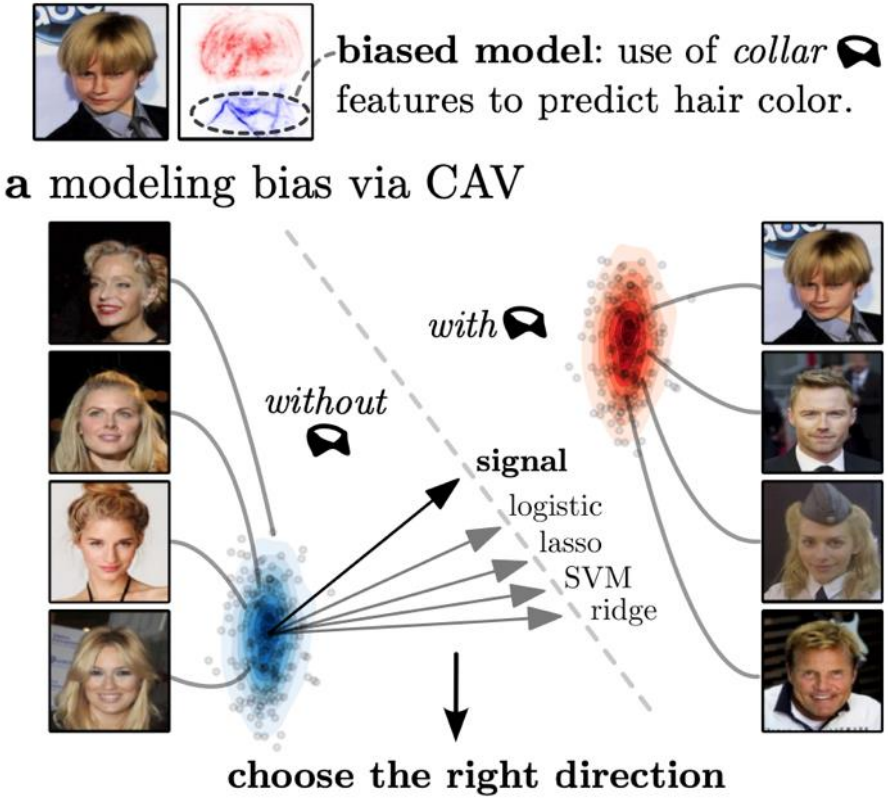
$$\frac{\partial \mathcal{L}(f_{\theta'}(X), Y)'}{\partial f^l_{\theta'}(X)} = \Psi_{\text{feat}}\left(\frac{\partial \mathcal{L}(f_{\theta'}(X), Y)}{\partial f^l_{\theta'}(X)}, R^{l,t}\right)$$

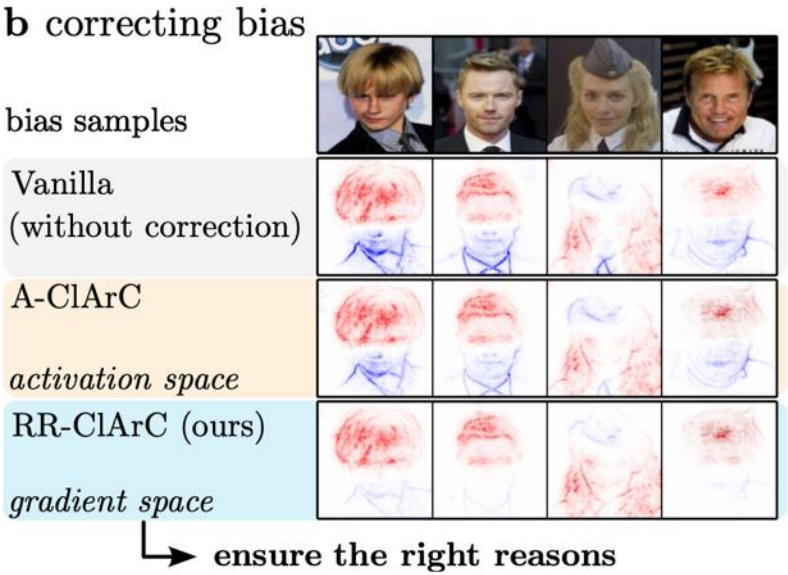
$$\frac{\partial \mathcal{L}(f_{\theta'}(X), Y)'}{\partial \theta^{l,t}} = \Psi_{\text{w}}\left(\frac{\partial \mathcal{L}(f_{\theta'}(X), Y)}{\partial \theta^{l,t}}, R^{l,t}_{\text{w}}\right)$$

Intermediate feature gradients can be scaled, masked, or transformed based on explanation.

Parameter gradients can be augmented directly by computing parameter-wise relevance scores.

RR-ClArC





(Dreyer et al. 2024)

RR-ClArC

$$L_{\mathrm{RR}}(\mathbf{x}) = \left(\nabla_{\mathbf{a}} \tilde{f}(\mathbf{a}(\mathbf{x})) \cdot \mathbf{h} \right)^2$$

sensitivity w.r.t. latent features

bias direction (artifact CAV)

Interpretation: Loss penalizes the model for changing the output when slightly adding or removing activations along the bias direction h.

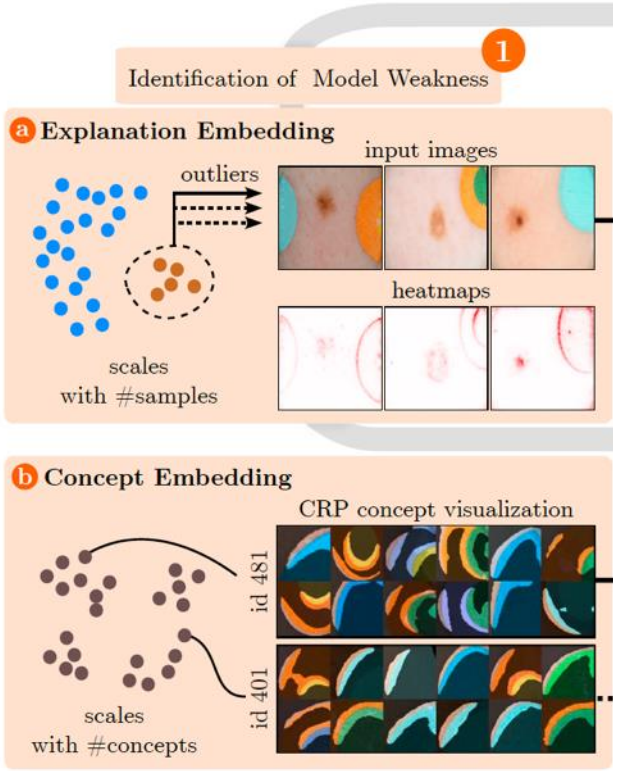
$$\lim_{\epsilon \to 0} \frac{\tilde{f}(\mathbf{a}(\mathbf{x}) + \epsilon \mathbf{h}) - \tilde{f}(\mathbf{a}(\mathbf{x}))}{\epsilon} \approx \nabla_{\mathbf{a}} \tilde{f}(\mathbf{a}(\mathbf{x})) \cdot \mathbf{h} \stackrel{!}{=} 0$$

RR-ClArC

architecture	method	Bone Age			ISIC			ImageNet			CelebA		
		clean	*biased*	TCAV	*clean*	*biased*	TCAV	*clean*	*biased*	TCAV	*clean*	*biased*	TCAV
VGG-16	*Vanilla*	78.8	49.8	0.86	76.2	34.9	0.84	68.7	43.5	0.63	93.7	82.8	0.37
	RRR	78.8	49.8	0.86	76.7	42.8	0.72	68.6	49.6	0.55	93.7	91.2	0.43
	P-ClArC	78.9	77.4	0.66	75.1	49.0	0.77	68.3	**62.6**	0.37	56.6	60.8	0.19
	A-ClArC	77.8	69.0	0.66	75.2	49.5	0.65	67.7	60.9	**0.49**	93.0	90.4	0.44
	RR-ClArC (ours)	78.8	**77.7**	**0.52**	74.3	**57.0**	**0.49**	68.5	**62.6**	**0.49**	93.6	**92.6**	**0.54**
ResNet-18	*Vanilla*	75.1	46.3	1.00	81.8	56.8	1.00	66.7	52.9	1.00	96.8	58.3	0.21
	RRR	74.5	47.9	1.00	78.7	61.1	1.00	66.4	59.1	0.08	95.5	74.7	0.92
	P-ClArC	75.0	70.7	**0.60**	60.8	59.9	1.00	67.0	61.7	0.80	96.5	64.4	0.06
	A-ClArC	74.8	57.4	0.34	77.1	65.0	0.98	65.0	63.3	0.88	96.1	62.9	0.38
	RR-ClArC (ours)	71.1	**74.2**	0.39	78.5	**71.2**	**0.76**	66.5	**64.0**	**0.55**	95.8	**75.3**	**0.61**
Efficient Net-B0	*Vanilla*	78.2	44.3	0.90	84.2	62.9	1.00	73.9	53.2	0.99	96.6	58.3	0.25
	RRR	78.4	49.6	0.79	83.1	68.7	0.85	73.9	59.1	0.66	95.4	75.6	**0.50**
	P-ClArC	65.2	35.1	0.02	19.7	29.6	1.00	74.1	54.6	0.21	96.8	55.0	0.05
	A-ClArC	78.0	54.2	0.64	77.7	72.8	0.68	71.4	69.9	0.90	96.7	60.6	0.24
	RR-ClArC (ours)	77.6	**70.3**	**0.53**	78.7	**75.6**	**0.54**	73.9	**70.8**	**0.56**	92.0	**77.6**	0.43

Reveal & Revise Example

Explain and Improve



Analysis in input space

Cluster heatmaps with SpRAy (Lapusckin et al. 2019)

Present cluster representative to human

Identify artifact clusters

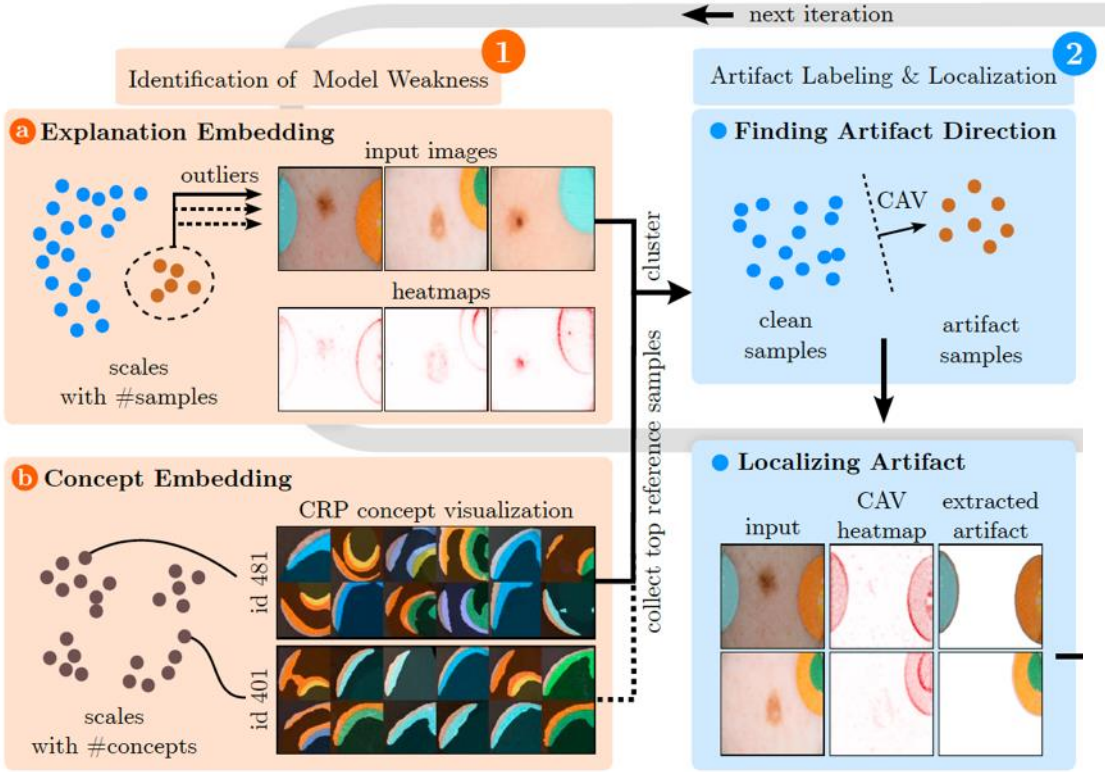
Analysis in concept space

Present concepts (or concept clusters) to human

Identify artifact concepts / clusters

[Pahde et al. 2023]

Explain and Improve



Compute "artifact CAV" h separating clean and outlier samples in latent space.

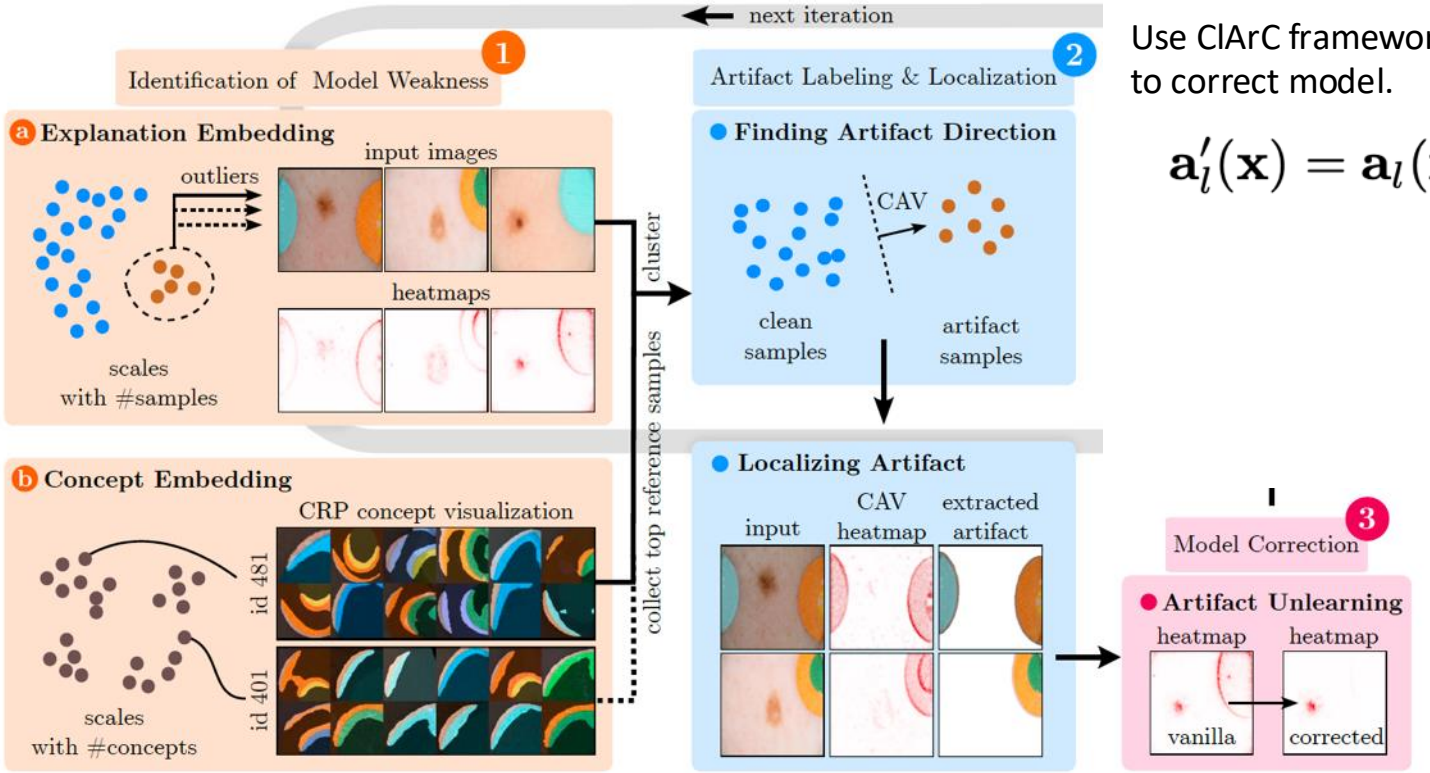
Localize the "artifact CAV" in input space (use LRP rule).

$$\mathbf{R}_l(\mathbf{x}) = \mathbf{a}_l(\mathbf{x}) \circ \mathbf{h}_l$$

→ explain artifact in input space using LRP.

[Pahde et al. 2023]

Explain and Improve



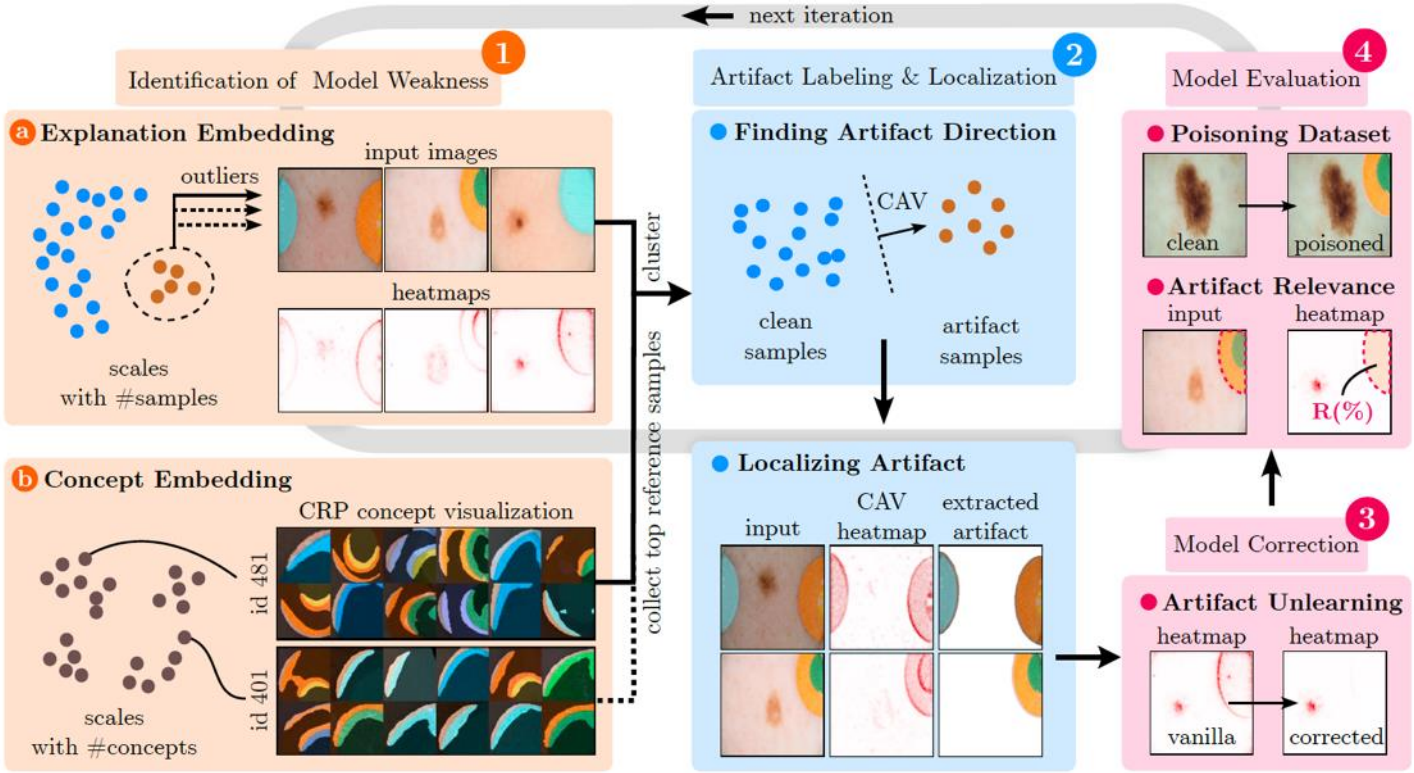
Latent space correction

Use ClArC framework (Anders et al. 2022) to correct model.

$$\mathbf{a}_l'(\mathbf{x}) = \mathbf{a}_l(\mathbf{x}) + \gamma(\mathbf{x})\mathbf{h}_l$$

[Pahde et al. 2023]

Explain and Improve

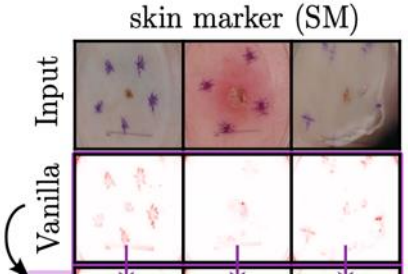


[Pahde et al. 2023]

Example

ISIC Dataset Artifacts:

- Skin Marker



--- Reveal Step ---

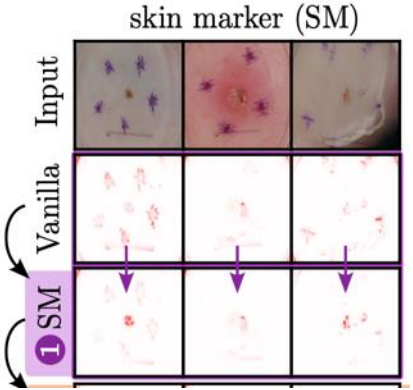
R2R iteration	corrected artifacts	↓ artifact relevance (%)			↑ F1 (%) *poisoned*		*original*	↑ accuracy (%) *poisoned*		*original*
0	-	18.4	45.5	24.2	61.3	59.7	60.5	73.9	71.8	71.5	68.7	**80.1**

[Pahde et al. 2023]

Example

ISIC Dataset Artifacts:

- Skin Marker



--- Revise Step ---

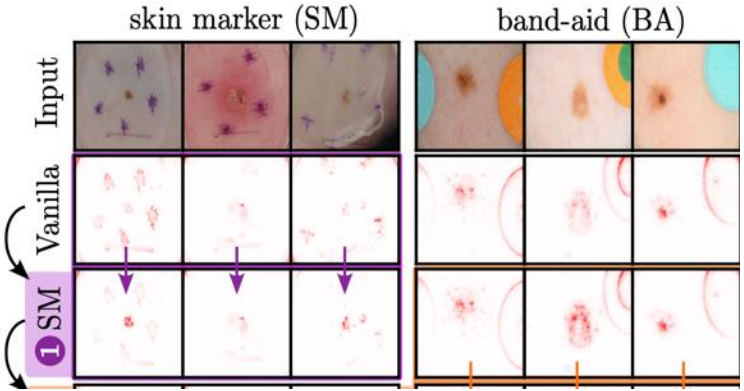
R2R iteration	corrected artifacts	↓ artifact relevance (%)			↑ F1 (%) *poisoned*			↑ F1 (%) *original*	↑ accuracy (%) *poisoned*			↑ accuracy (%) *original*
0	-	18.4	45.5	24.2	61.3	59.7	60.5	73.9	71.8	71.5	68.7	**80.1**
1	SM	13.1	35.0	21.3	61.6	61.0	60.7	73.8	72.2	72.6	68.4	80.0

[Pahde et al. 2023]

Example

ISIC Dataset Artifacts:

- Skin Marker
- Band-Aid



--- Reveal Step ---

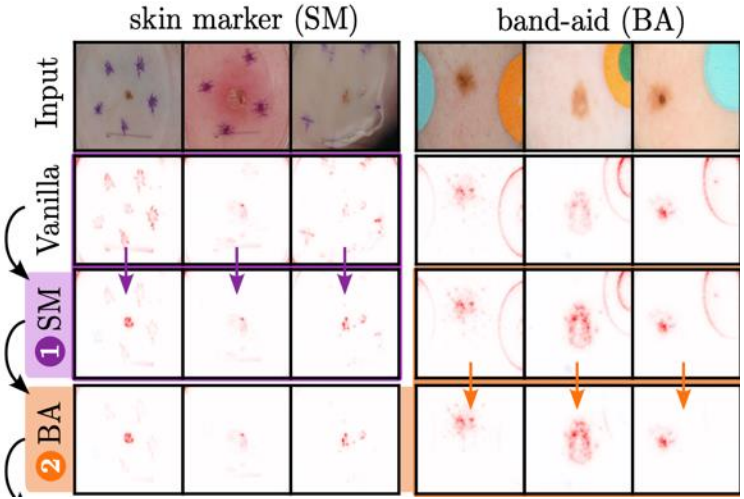
R2R iteration	corrected artifacts	↓ artifact relevance (%)			↑ F1 (%) *poisoned*			↑ F1 (%) *original*	↑ accuracy (%) *poisoned*			↑ accuracy (%) *original*
0	-	18.4	45.5	24.2	61.3	59.7	60.5	73.9	71.8	71.5	68.7	**80.1**
1	SM	13.1	35.0	21.3	61.6	61.0	60.7	73.8	72.2	72.6	68.4	80.0

[Pahde et al. 2023]

Example

ISIC Dataset Artifacts:

- Skin Marker
- Band-Aid



--- Revise Step ---

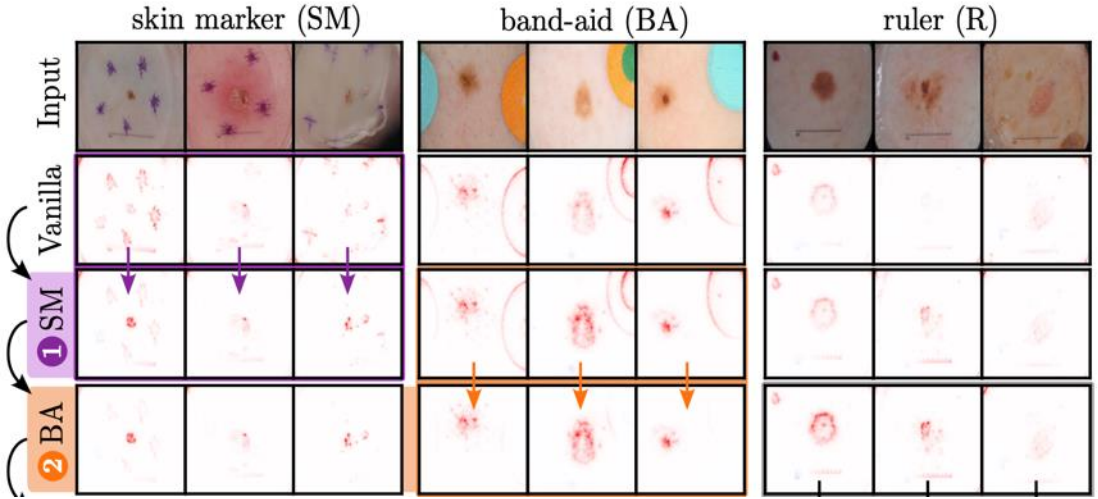
R2R iteration	corrected artifacts	↓ artifact relevance (%)			↑ F1 (%) *poisoned*			↑ F1 (%) *original*	↑ accuracy (%) *poisoned*			↑ accuracy (%) *original*
0	-	18.4	45.5	24.2	61.3	59.7	60.5	73.9	71.8	71.5	68.7	**80.1**
1	SM	13.1	35.0	21.3	61.6	61.0	60.7	73.8	72.2	72.6	68.4	80.0
2	SM, BA	**12.8**	16.8	16.8	61.5	**63.6**	61.1	73.9	72.3	**74.6**	68.6	79.7

[Pahde et al. 2023]

Example

ISIC Dataset Artifacts:

- Skin Marker
- Band-Aid
- Ruler



--- Reveal Step ---

R2R iteration	corrected artifacts	↓ artifact relevance (%)			↑ F1 (%) *poisoned*		↑ F1 (%) *original*	↑ accuracy (%) *poisoned*		↑ accuracy (%) *original*
0	-	18.4	45.5	24.2	61.3	59.7	60.5	73.9	71.8	71.5	68.7	**80.1**
1	SM	13.1	35.0	21.3	61.6	61.0	60.7	73.8	72.2	72.6	68.4	80.0
2	SM, BA	**12.8**	16.8	16.8	61.5	**63.6**	61.1	73.9	72.3	**74.6**	68.6	79.7

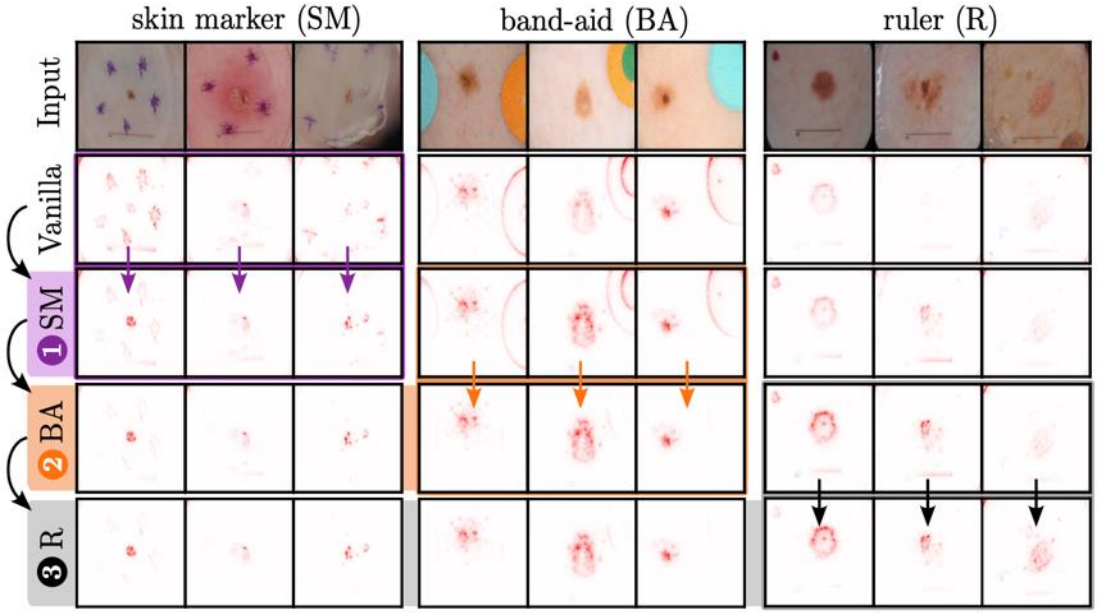
[Pahde et al. 2023]

Example

ISIC Dataset Artifacts:

- Skin Marker
- Band-Aid
- Ruler

--- Revise Step ---

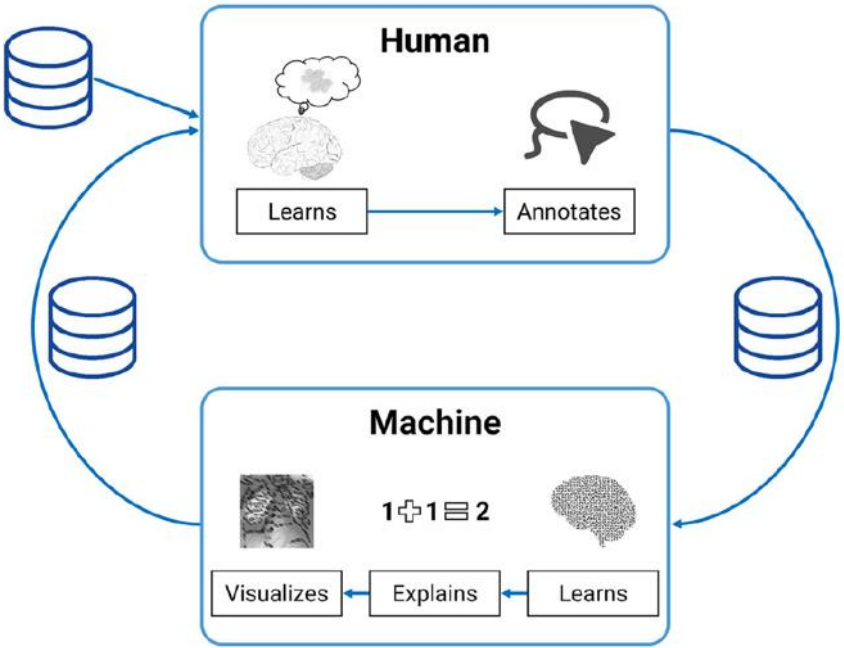


R2R iteration	corrected artifacts	↓ artifact relevance (%)			↑ F1 (%) *poisoned*			↑ F1 (%) *original*	↑ accuracy (%) *poisoned*			↑ accuracy (%) *original*
0	-	18.4	45.5	24.2	61.3	59.7	60.5	73.9	71.8	71.5	68.7	**80.1**
1	SM	13.1	35.0	21.3	61.6	61.0	60.7	73.8	72.2	72.6	68.4	80.0
2	SM, BA	**12.8**	16.8	16.8	61.5	**63.6**	61.1	73.9	72.3	**74.6**	68.6	79.7
3	SM, BA, R	14.6	**15.7**	**8.5**	**62.0**	63.4	**64.0**	**74.0**	**72.4**	74.5	**71.8**	79.9

[Pahde et al. 2023]

Explanatory Interactive ML

Explanatory Interactive ML (XIL)



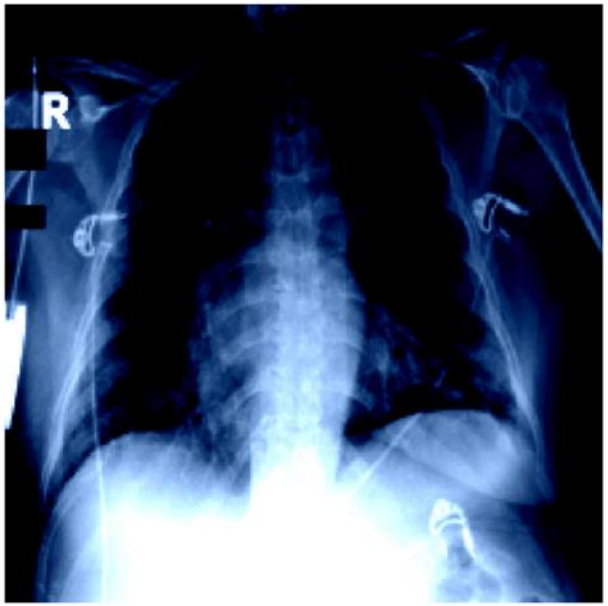
"In XIL, a learner can interactively query the user (or some other information source) to obtain the desired outputs of the data points. The interaction takes the following form. At each step, the learner considers a data point (labeled or unlabeled), predicts a label, and provides explanations of its prediction. The user responds by correcting the learner if necessary, providing a slightly improved – but not necessarily optimal – feedback to the learner."

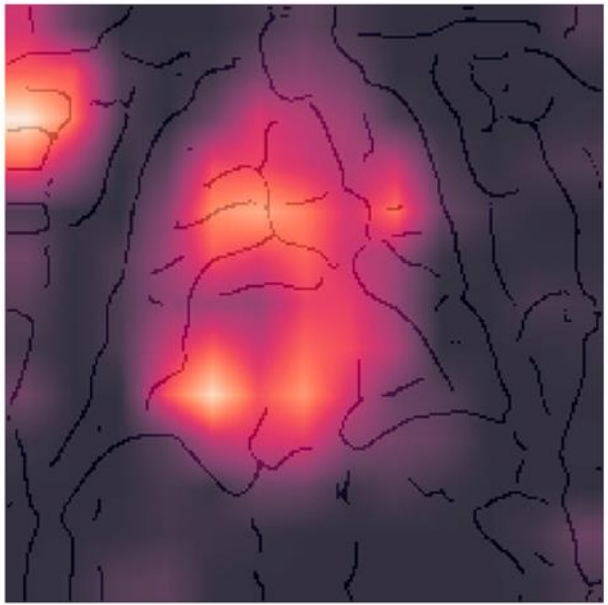
[Pfeuffer et al. 2023]

Example

1. Step: Reveal

Fig. 6 Loop 1 XAI. Original image on the left, XAI showing confounders with Grad-CAM method (Selvaraju et al. 2017) on an overlay of the edge-filtered image on the right (color figure online)





Example

2. Step: Data curation

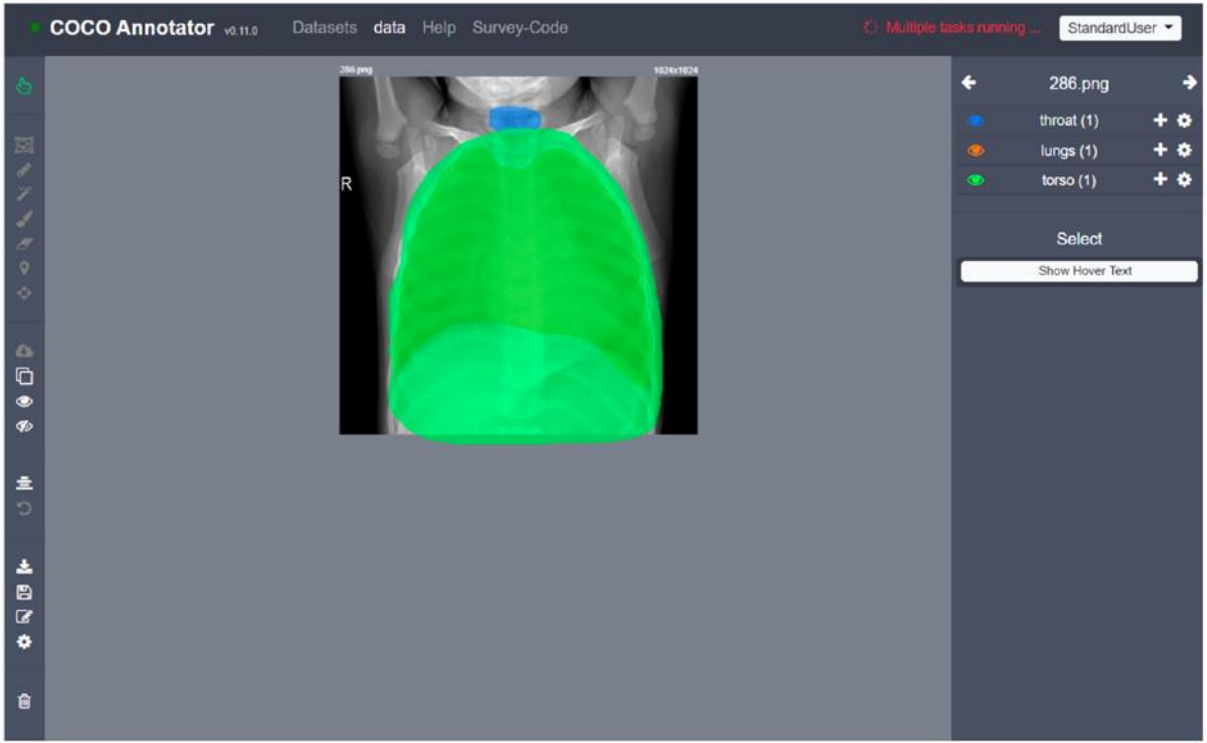


Fig. 7 The user interface of the customized COCO Annotator used by the crowdworkers to annotate the images. Using the tools on the left on each X-ray, the areas for the throat, lungs, and torso were highlighted (color figure online)

Example

3. Step: Refine data and retrain model

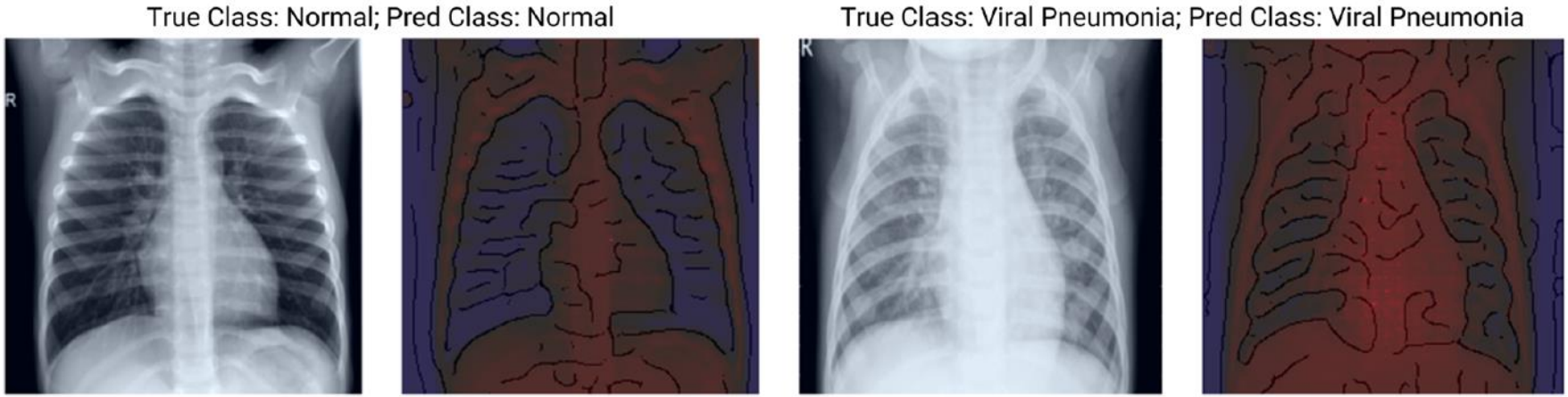


Fig. 9 XAI showing Explanations for Classification in loop 3a (using our customized red-blue filter solution based on VarGrad (Adebayo et al. 2018)) (color figure online)

10th Take Home Message

Integrating the human into XAI / XIL is crucial.

Concept-Level explanantions can be a useful interface here (especially for non-image data).

Future of XAI

Next-Generation Explanations

contextual and semantically rich:

--> explain whole inference (where, what, how)

adequate for receiver:

--> presented in comprehensible form (also for time series etc.)

actionable:

--> useful beyond visualization, e.g., for debugging, improving or auditing model.

of various types:

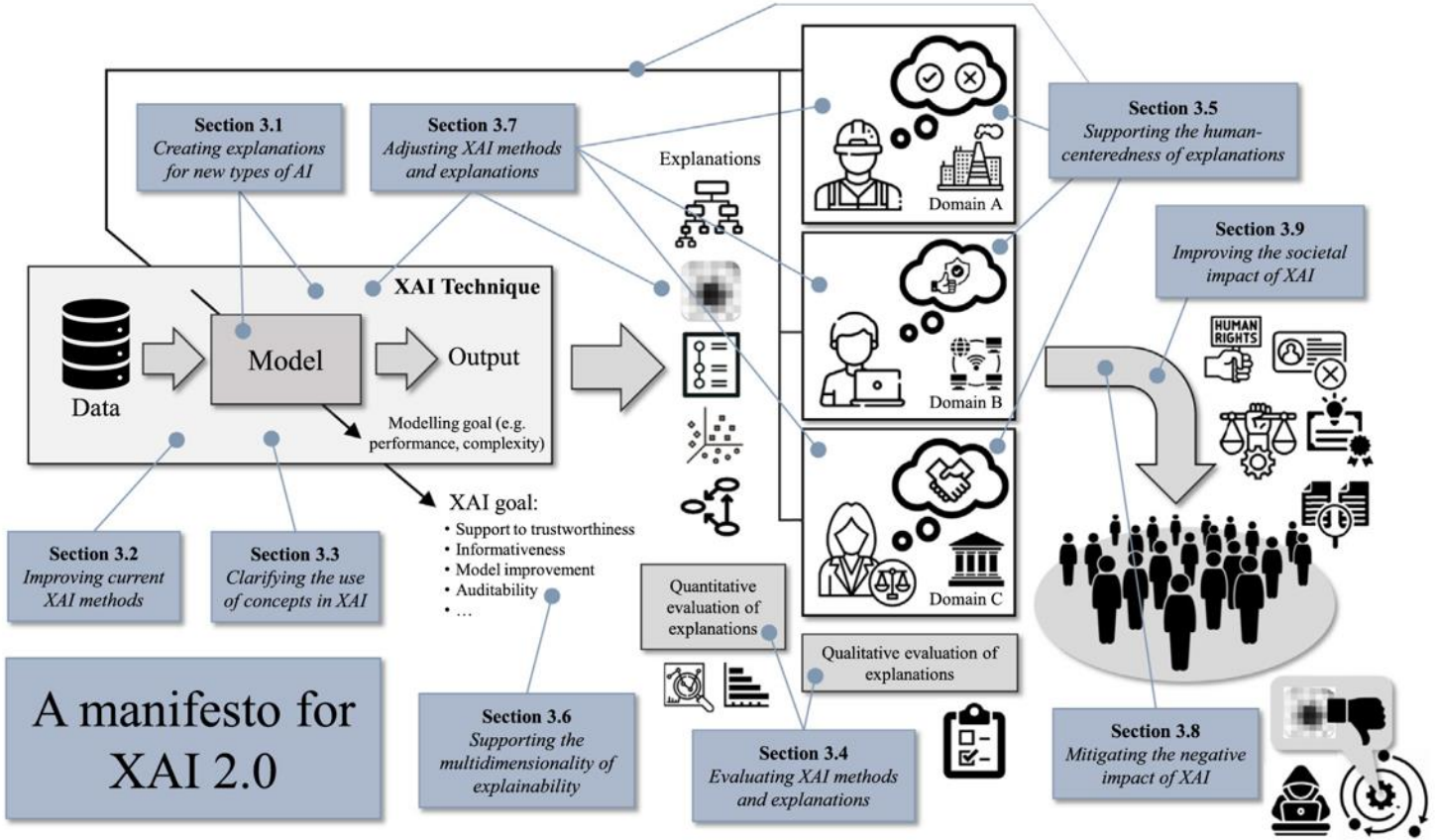
--> broader definition of explainability / transparency

Next-Generation Explanations

Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions

Luca Longo [1,2,*], Mario Brcic [3], Federico Cabitza [4,5], Jaesik Choi [6,7], Roberto Confalonieri [8], Javier Del Ser [9,10,11], Riccardo Guidotti [12], Yoichi Hayashi [13], Francisco Herrera [11], Andreas Holzinger [14], Richard Jiang [15], Hassan Khosravi [16], Freddy Lecue [17], Gianclaudio Malgieri [18], Andrés Páez [19,20], Wojciech Samek [21,22,23], Johannes Schneider [24], Timo Speith [25,26], Simone Stumpf [27]

Next-Generation Explanations



Thank you for your attention

http://www.heatmapping.org

- Tutorials
- Software
- Online Demos