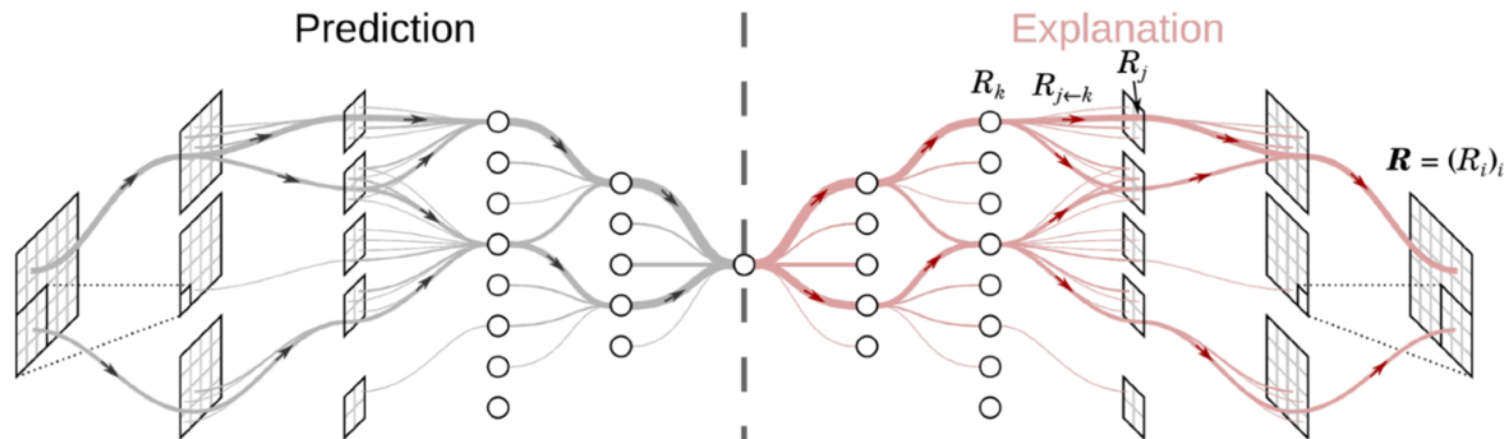


# Explainable AI for LLMs

Wojciech Samek

TU Berlin & Fraunhofer HHI



# Today's LLMs: Impress

Please tell me how to cut bread.



LLM



(source: <https://www.shutterstock.com>)

# Today's LLMs: Impress

Please tell me how to cut bread



LLM

Cutting bread might seem simple, but there's an art to doing it well! Here's a step-by-step guide to get clean, even slices without squishing the loaf:

- 1. Choose the Right Knife**
- 2. Place the Bread on a Cutting Board**
- 3. Position Your Hand Correctly**
- 4. Start Slicing**
- 5. Slice Thickness**

**Extra Tips**

# Today's LLMs: Hallucinate

Have there been any reports of implantation of a subcutaneous defibrillator immediately after cardiac surgery and please provide references



LLM

# Today's LLMs: Hallucinate

Have there been any reports of implantation of a subcutaneous defibrillator immediately after cardiac surgery and please provide references



LLM

Yes, there have been reports of implantation of subcutaneous defibrillators (S-ICDs) immediately after cardiac surgery. Here are some references:

1. 'Immediate postoperative implantation of a subcutaneous defibrillator in a patient undergoing cardiac surgery.' *J Interv Card Electrophysiol*. 2016 Feb;45(2):203–5. doi: 10.1007/s10840-015-0087-6. Epub 2015 Dec 17. PMID: 26676620.
2. 'Immediate postoperative subcutaneous implantable cardioverter-defibrillator implantation in patients undergoing cardiac surgery: feasibility, safety and outcomes.' *Europace*. 2017 Jul 1;19(7):1150–1156. doi: 10.1093/europace/euw235. PMID: 27614059.
3. 'Immediate postoperative implantation of a subcutaneous defibrillator in a patient undergoing cardiac surgery.' *J Arrhythm*. 2017 Dec;33(6):639–641. doi: 10.1016/j.joa.2017.06.008. Epub 2017 Jul 5. PMID: 29276432.

# Today's LLMs: Hallucinate

Have there been any reports of implantation of a subcutaneous defibrillator immediately after cardiac surgery and please provide references



LLM

**there are no  
such papers**

Yes, there have been reports of implantation of subcutaneous defibrillators (S-ICDs) immediately after cardiac surgery. Here are some references:

1. 'Immediate postoperative implantation of a subcutaneous defibrillator in a patient undergoing cardiac surgery.' *J Interv Card Electrophysiol*. 2016 Feb;45(2):203–5. doi: 10.1007/s10840-015-0087-6. Epub 2015 Dec 17. PMID: 26676620.
2. 'Immediate postoperative subcutaneous implantable cardioverter-defibrillator implantation in patients undergoing cardiac surgery: feasibility, safety and outcomes.' *Europace*. 2017 Jul 1;19(7):1150–1156. doi: 10.1093/europace/euw235. PMID: 27614059.
3. 'Immediate postoperative implantation of a subcutaneous defibrillator in a patient undergoing cardiac surgery.' *J Arrhythm*. 2017 Dec;33(6):639–641. doi: 10.1016/j.joa.2017.06.008. Epub 2017 Jul 5. PMID: 29276432.

---

**ChatGPT hallucinating: can it get any more humanlike?**

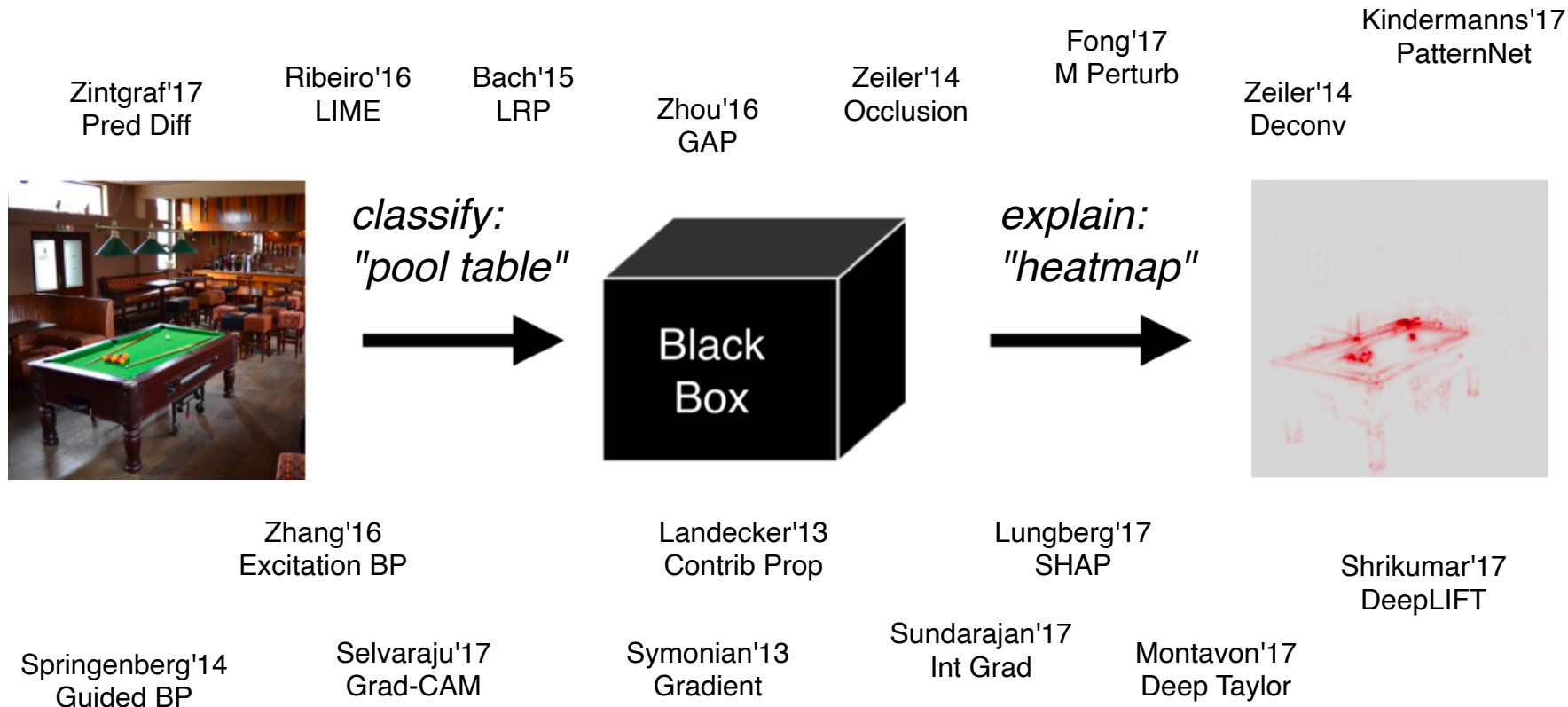
---

To trust or not to trust  
AI; that is the question

We need to solve the  
“Black Box” problem!



# First Wave of Explainable AI

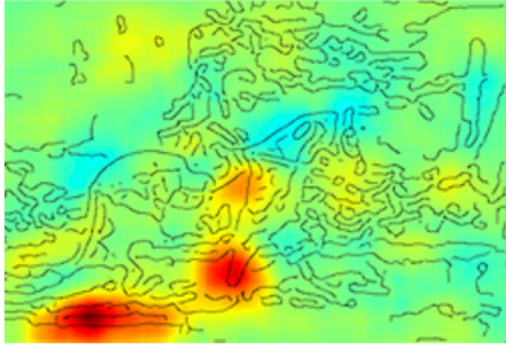




# Why Explaining?

## Debug models

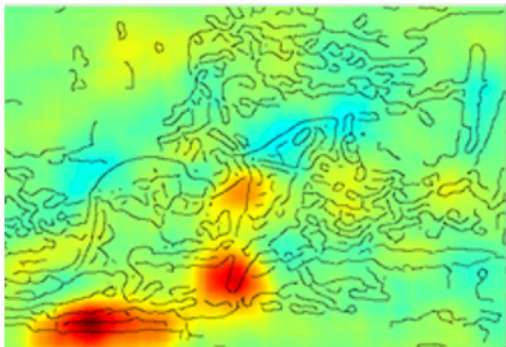
(Lapuschkin et al. Nat Comm, 2019)



# Why Explaining?

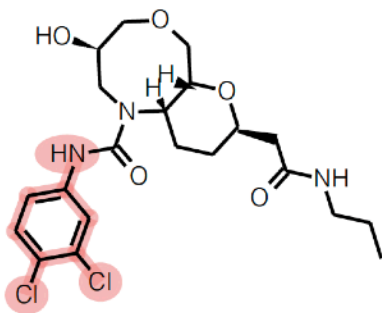
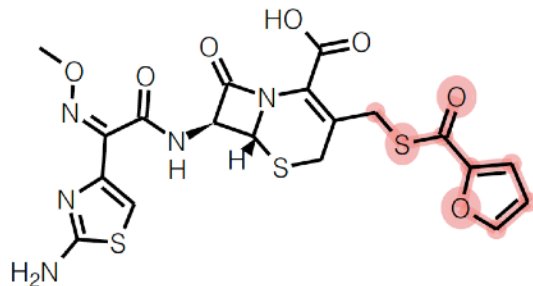
## Debug models

(Lapuschkin et al. Nat Comm, 2019)



## New insights

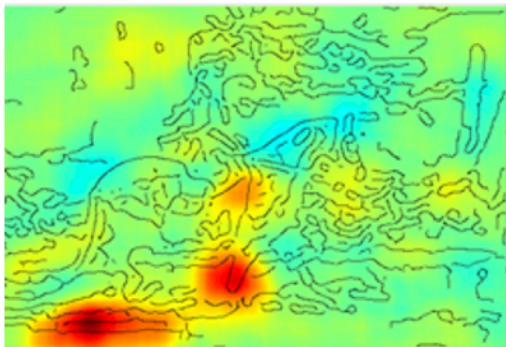
(Wong et al. Nature, 2023)



# Why Explaining?

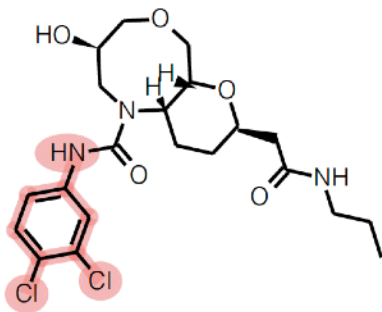
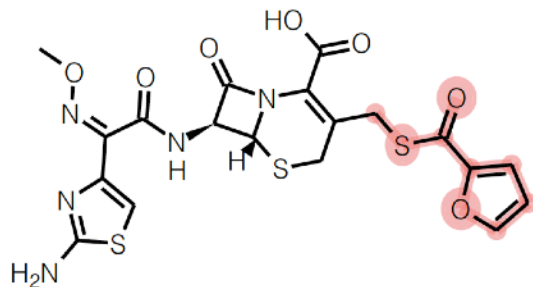
## Debug models

(Lapuschkin et al. Nat Comm, 2019)



## New insights

(Wong et al. Nature, 2023)



## "BLUE XAI"

(Biecek & Samek, ICML, 2024)

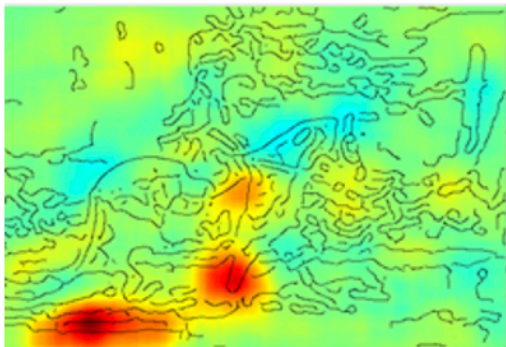
Human-values oriented

- Responsible models
- Legal issues
- Trust in predictions
- Ethical issues

# Why Explaining?

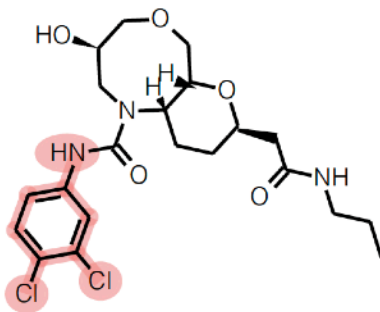
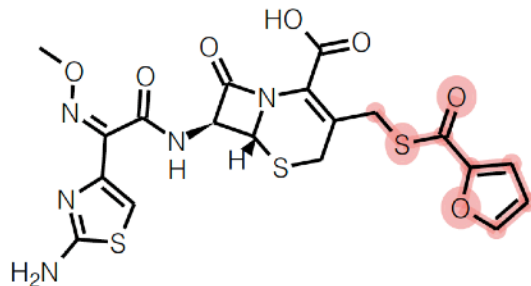
## Debug models

(Lapuschkin et al. Nat Comm, 2019)



## New insights

(Wong et al. Nature, 2023)



## "BLUE XAI"

(Biecek & Samek, ICML, 2024)

Human-values oriented

- Responsible models
- Legal issues
- Trust in predictions
- Ethical issues

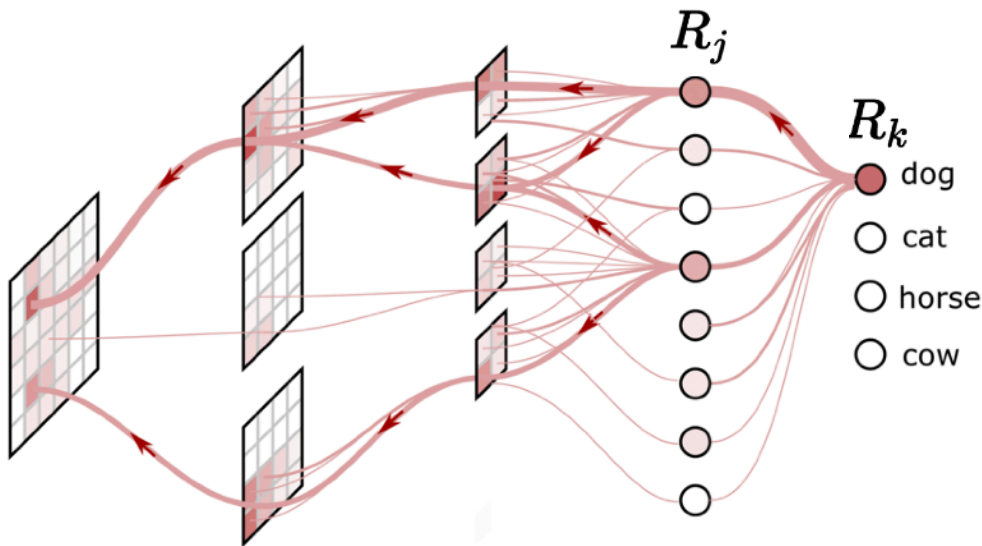
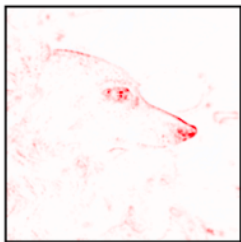
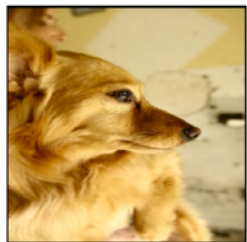
## Provide citations in LLMs

What is the capital of Germany?

**ChatGPT:** The capital of Germany is Berlin.

WIKIPEDIA The Free Encyclopedia		List of national capitals
Beirut	 Lebanon	Asia
Belgrade	 Serbia	Europe
Belmopan	 Belize	North America
Berlin	 Germany	Europe

# LRP: Faithful + Comput. Efficient + Latent Attribution



(1) decompose

$$R_{j \leftarrow k} = \frac{z_{jk}}{z_k} R_k$$



(2) aggregate

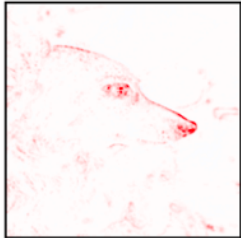
$$R_j = \sum R_{j \leftarrow k}$$

**Theoretical Interpretation:**  
Deep Taylor Decomposition

$z_{jk}$  measures how much  $j$  has contributed to activation of  $k$

(Bach et al. 2015)

# LRP: Faithful + Comput. Efficient + Latent Attribution



## Convolutional NN

Name	Formula	Usage	DTD
LRP-0 [7]	$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$	upper layers	✓
LRP- $\epsilon$ [7]	$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k$	middle layers	✓
LRP- $\gamma$	$R_j = \sum_k \frac{a_j (w_{jk} + \gamma w_{jk}^+)}{\sum_{0,j} a_j (w_{jk} + \gamma w_{jk}^+)} R_k$	lower layers	✓
LRP- $\alpha\beta$ [7]	$R_j = \sum_k \left( \alpha \frac{(a_j w_{jk})^+}{\sum_{0,j} (a_j w_{jk})^+} - \beta \frac{(a_j w_{jk})^-}{\sum_{0,j} (a_j w_{jk})^-} \right) R_k$	lower layers	×*
flat [30]	$R_j = \sum_k \frac{1}{\sum_j 1} R_k$	lower layers	×
$w^2$ -rule [36]	$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$	first layer ( $\mathbb{R}^d$ )	✓
$z^B$ -rule [36]	$R_i = \sum_j \frac{x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-}{\sum_i x_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-} R_j$	first layer (pixels)	✓

(\* DTD interpretation only for the case  $\alpha = 1, \beta = 0$ .)

compose

$$= \frac{z_{jk}}{z_k} R_k$$



aggregate

$$\sum R_{j \leftarrow k}$$

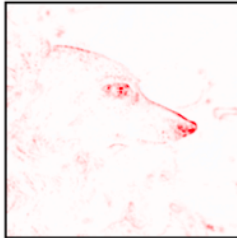
activation of k

Theoretical Interp

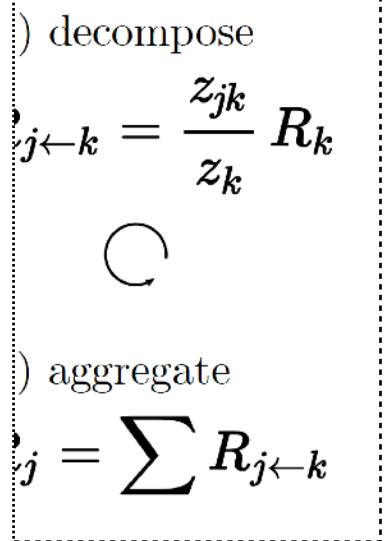
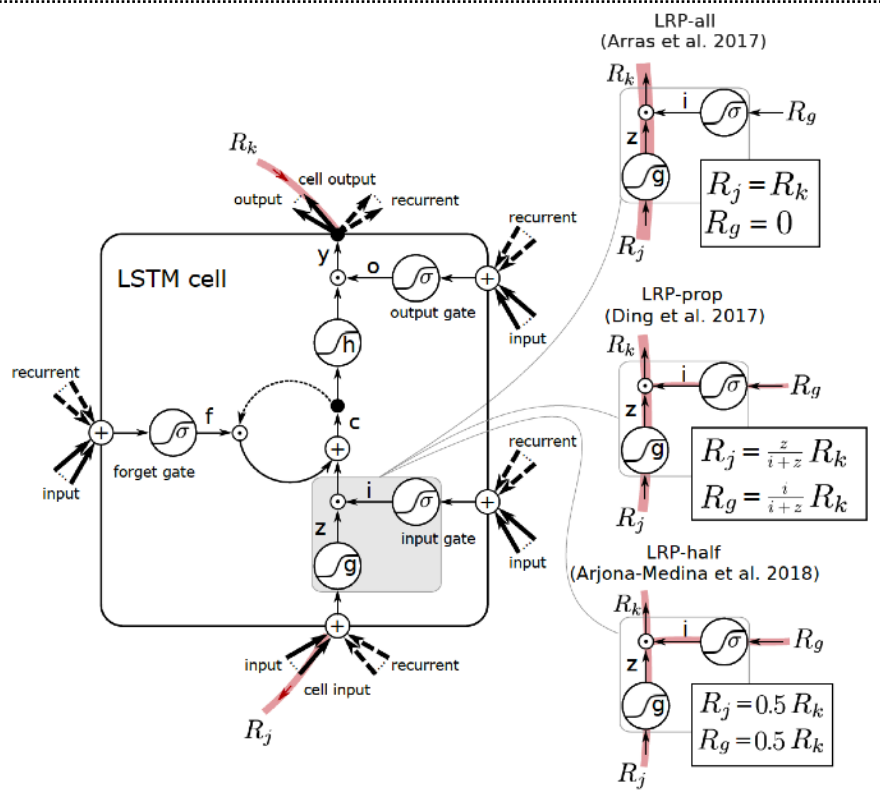
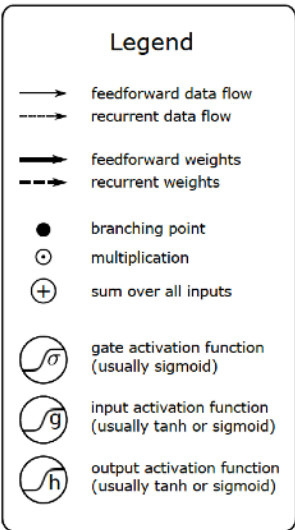
Deep Taylor Decomposition

(Montavon et al. 2017)

# LRP: Faithful + Comput. Efficient + Latent Attribution



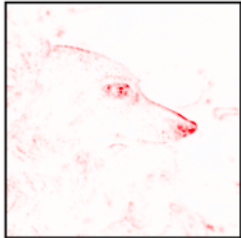
## LSTM



to activation of k

Theoretical Int  
Deep Taylor Dec

# LRP: Faithful + Comput. Efficient + Latent Attribution



## Transformers

*"How to meaningfully  
redistribute relevance  
through these layers?"*

$$\mathbf{A} = \text{softmax} \left( \frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}} \right)$$

$$\mathbf{O} = \mathbf{A} \cdot \mathbf{V}$$

$$\text{softmax}_j(\mathbf{x}) = \frac{e^{x_j}}{\sum_k e^{x_k}}$$

$$\text{LayerNorm}(\mathbf{x}) = \frac{x_j - \mathbb{E}[\mathbf{x}]}{\sqrt{\text{Var}[\mathbf{x}] + \varepsilon}} \gamma_j + \beta_j$$

$$\text{RMSNorm}(\mathbf{x}) = \frac{x_j}{\sqrt{\frac{1}{N} \sum_k x_k^2 + \varepsilon}} \gamma_j$$

compose

$$= \frac{z_{jk}}{z_k} R_k$$



aggregate

$$= \sum R_{j \leftarrow k}$$

Theoretical Interpretation

Deep Taylor Decomposition

activation of k



# AttnLRP: Attention-Aware Layer-wise Relevance Propagation for Transformers

Reduan Achtibat<sup>1</sup>    Sayed Mohammad Vakilzadeh Hatefi<sup>1</sup>    Maximilian Dreyer<sup>1</sup>  
Aakriti Jain<sup>1</sup>    Thomas Wiegand<sup>1,2,3</sup>    Sebastian Lapuschkin<sup>1,†</sup>    Wojciech Samek<sup>1,2,3,†</sup>

<sup>1</sup> Fraunhofer Heinrich-Hertz-Institute, 10587 Berlin, Germany

<sup>2</sup> Technische Universität Berlin, 10587 Berlin, Germany

<sup>3</sup> BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany

**Proposition 3.1** *Decomposing the softmax function by a Taylor decomposition (4) at reference point  $\mathbf{x}$  yields the following relevance propagation rule:*

$$R_i^l = x_i(R_i^{l+1} - s_i \sum_j R_j^{l+1}) \quad (13)$$

where  $s_j$  denotes the  $j$ -th output of the softmax function.

The hidden bias term error, consequently at

**Proposition 3.2** *Decomposing element-wise multiplication with  $N$  input variables of the form*

$$f_j(\mathbf{x}) = \prod_i^N x_i$$

*by Shapley (with baseline zero) or Taylor decomposition (4) at reference point  $\mathbf{x}$  (without bias or distributing the bias uniformly) yields the following uniform relevance propagation rule:*

$$R_{i \leftarrow j}(x_i) = \frac{1}{N} R_j. \quad (14)$$

<https://proceedings.mlr.press/v235/achtibat24a.html>

<https://github.com/rachtibat/LRP-eXplains-Transformers>

**Proposition 3.4** *Decomposing LayerNorm or RMSNorm by a Taylor decomposition (4) with reference point  $\mathbf{0}$  (without bias or distributing the bias uniformly) yields the identity relevance propagation rule:*

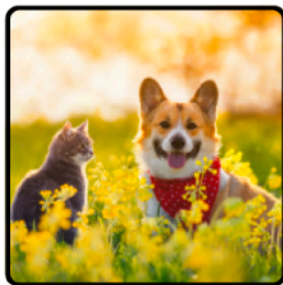
$$R_i^{l-1} = R_i^l \quad (19)$$

**Proposition 3.3** *Decomposing matrix multiplication with a sequential application of the uniform rule (14) and the  $\varepsilon$ -rule (8) yields the following relevance propagation rule:*

$$R_{ji}^{l-1}(\mathbf{A}_{ji}) = \sum_p \mathbf{A}_{ji} \mathbf{V}_{ip} \frac{R_{jp}^l}{2 \mathbf{O}_{jp} + \varepsilon} \quad (15)$$

# Comparison

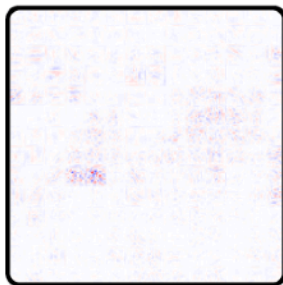
explanation for "dog"



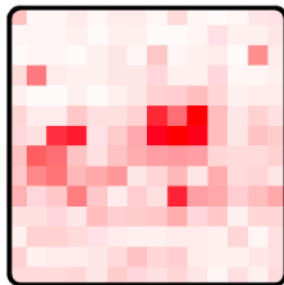
input



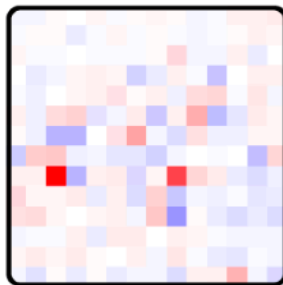
AttnLRP(ours)



SmoothGrad



Grad×AttnRoll



AtMan

faithfulness



computational  
efficiency



latent  
attributions

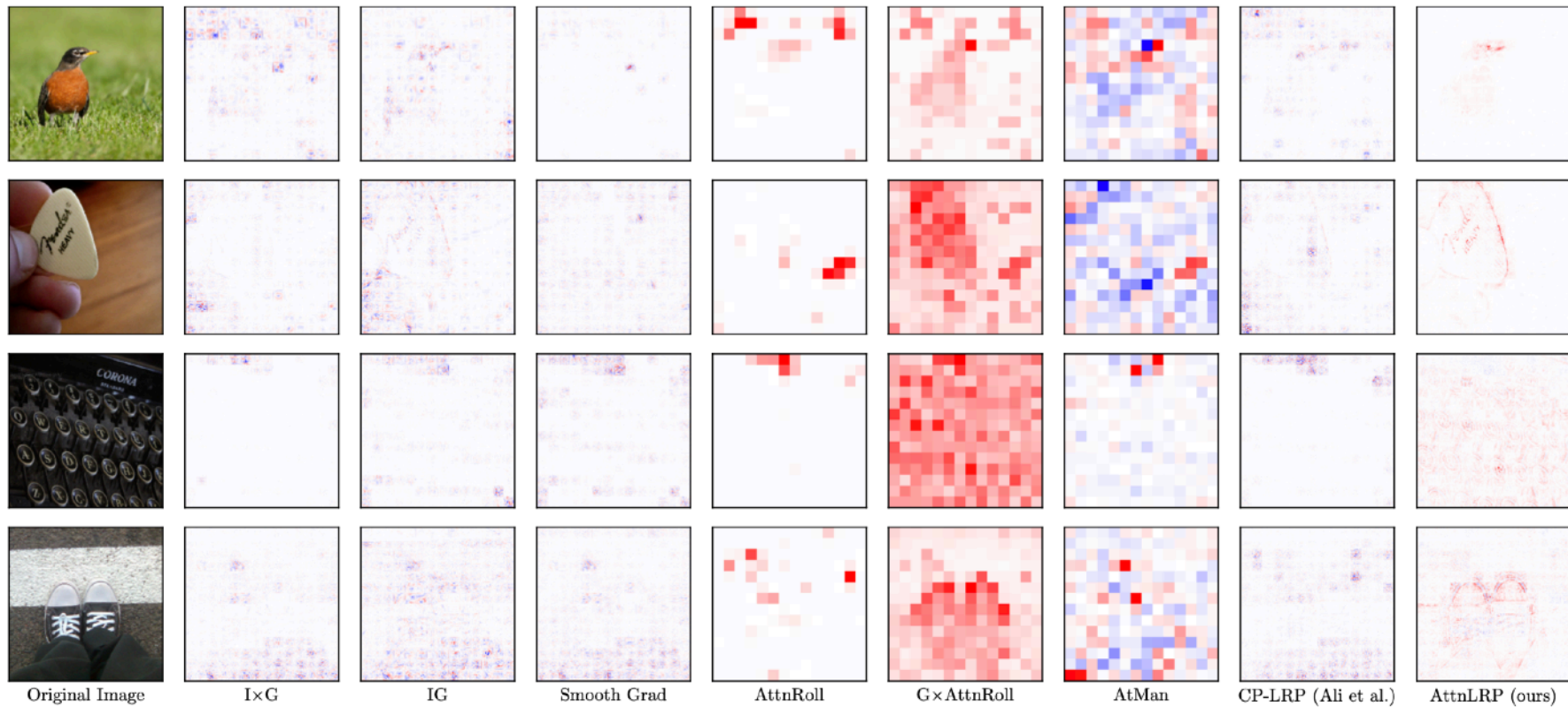


# Faithfulness

Table 1. Faithfulness scores as area between the least and most relevant order perturbation curves (Blücher et al., 2024) on different models and datasets. To assess plausibility, the (top-1) accuracy along with the IoU in parentheses are depicted for SQuAD v2. Methods marked with (\*) have been proposed here. Additional results for ViT-L-16 and ViT-L-32 are in Appendix Table B.6.

Methods	ViT-B-16	LLaMa 2-7b		Mixtral 8x7b	Flan-T5-XL
	ImageNet $\uparrow$	IMDB $\uparrow$	Wikipedia $\uparrow$	SQuAD v2 $\uparrow$	SQuAD v2 $\uparrow$
Random	$0.01 \pm 0.01$	$-0.01 \pm 0.05$	$-0.07 \pm 0.13$	0.03 (0.09)	0.03 (0.08)
Input $\times$ Grad (Simonyan et al., 2014)	$0.80 \pm 0.03$	$0.12 \pm 0.05$	$0.18 \pm 0.13$	0.56 (0.35)	0.60 (0.39)
IG (Sundararajan et al., 2017)	$1.54 \pm 0.03$	$1.23 \pm 0.05$	$4.05 \pm 0.13$	0.68 (0.44)	0.10 (0.16)
SmoothGrad (Smilkov et al., 2017)	$-0.04 \pm 0.03$	$0.25 \pm 0.05$	$-2.22 \pm 0.14$	0.47 (0.24)	0.05 (0.09)
GradCAM (Chefer et al., 2021b)	$0.27 \pm 0.04$	$-0.82 \pm 0.05$	$2.01 \pm 0.15$	0.82 (0.72)	0.81 (0.70)
AttnRoll (Abnar and Zuidema, 2020)	$1.31 \pm 0.03$	$-0.64 \pm 0.05$	$-3.49 \pm 0.15$	0.05 (0.10)	0.02 (0.08)
Grad $\times$ AttnRoll (Chefer et al., 2021a)	$2.60 \pm 0.03$	$1.61 \pm 0.05$	$9.79 \pm 0.14$	0.91 (0.40)	<b>0.94</b> (0.53)
AtMan (Deb et al., 2023)	$0.70 \pm 0.02$	$-0.20 \pm 0.05$	$3.31 \pm 0.15$	0.86 ( <b>0.83</b> )	0.88 (0.80)
KernelSHAP (Lundberg and Lee, 2017)	$4.71 \pm 0.03$	-	-	-	-
CP-LRP ( $\epsilon$ -rule, Ali et al. (2022))	$2.53 \pm 0.02$	$1.72 \pm 0.04$	$7.85 \pm 0.12$	0.50 (0.40)	0.91 (0.83)
CP-LRP ( $\gamma$ -rule for ViT, as proposed here)*	$6.06 \pm 0.02$	-	-	-	-
AttnLRP (ours)*	<b><math>6.19 \pm 0.02</math></b>	<b><math>2.50 \pm 0.05</math></b>	<b><math>10.93 \pm 0.13</math></b>	<b>0.96</b> (0.72)	<b>0.94</b> ( <b>0.84</b> )

# Results



# Results

Question: In what country is Normandy located?

Answer: France

## AttnLRP

The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.

## AtMan

The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.

## Integrated Gradient

The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.

Evaluation on the  
Mixtral 8x7b model

## Input x Gradient

The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.

## Grad-CAM

The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.

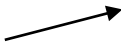
## CP-LRP

The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would gradually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerged initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.

# Comparison with other LRP Rules

Methods	Softmax	Matrix Multiplication	Layer Normalization
(Ding et al., 2017)	Identity rule ⇒ unstable (Appendix A.2.1)	0-LRP (bi-linear) ⇒ violates conservation	not available
(Voita et al., 2021)	Taylor decomposition at $\mathbf{x}$ (distributes the bias uniformly) ⇒ unstable (Appendix A.2.1)	$z^+$ -LRP (bi-linear) ⇒ violates conservation	Taylor decomposition at $\mathbf{x}$ (distributes the bias uniformly) ⇒ unstable (Appendix A.2.1)
(Chefer et al., 2021b)	Identity rule ⇒ unstable (Appendix A.2.1)	0-LRP & post-hoc normalization (bi-linear) ⇒ ensures conservation	Identity rule ⇒ ensures conservation & faithful
(Ali et al., 2022)	Regarded as constant ⇒ stable & no attribution inside attention module	0-LRP (linear only) ⇒ ensures conservation	Identity rule ⇒ ensures conservation & faithful
AttnLRP	Taylor decomposition at $\mathbf{x}$ (with bias) ⇒ stable & faithful	$\epsilon$ -LRP & uniform rule (bi-linear) ⇒ ensures conservation & faithful	Identity rule ⇒ ensures conservation & faithful

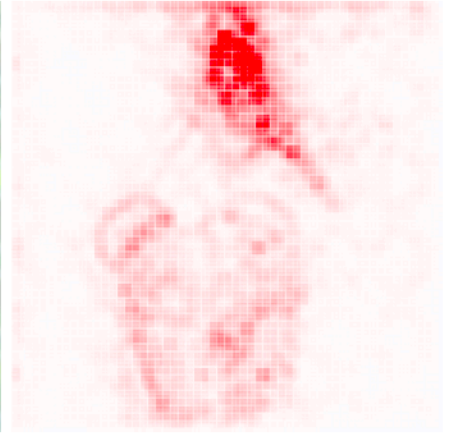
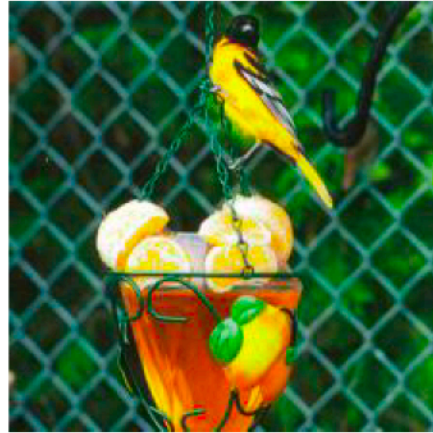
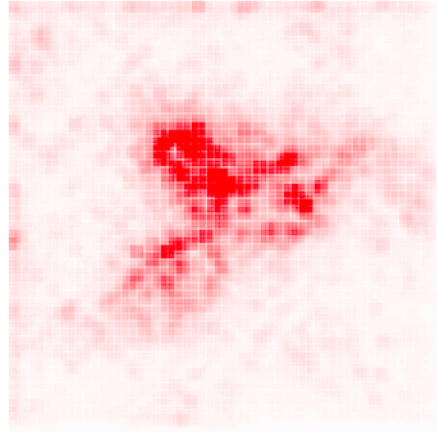
fully within Deep Taylor Decomposition Framework



Towards Concept-Level XAI

# Interpretation Gap

Local XAI functions as **marker** for important features



Does it help understanding?

Local XAI tells us "where" important features occur.

However, it neither tells us "what" this feature is nor how it is used by the model.



# Explainability 2.0: Where, What and How

nature machine intelligence



Article


<https://doi.org/10.1038/s42256-023-00711-8>

## From attribution maps to human-understandable explanations through Concept Relevance Propagation

Received: 7 June 2022

Accepted: 31 July 2023

Published online: 20 September 2023

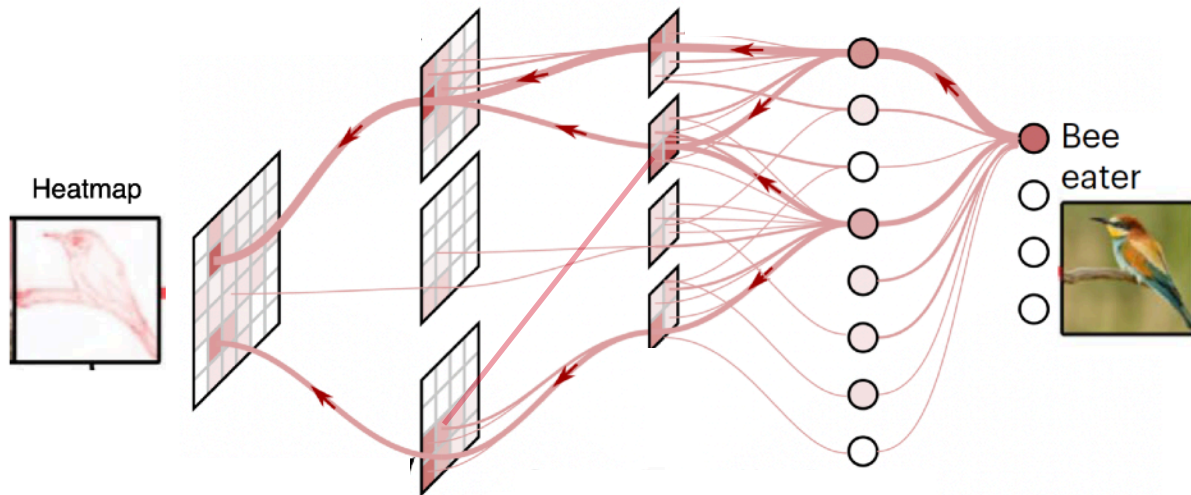
 Check for updates

Reduan Aichtibat<sup>1,4</sup>, Maximilian Dreyer<sup>1,4</sup>, Ilona Eisenbraun<sup>1</sup>, Sebastian Bosse<sup>1</sup>,  
Thomas Wiegand<sup>1,2,3</sup>, Wojciech Samek<sup>1,2,3</sup>✉ & Sebastian Lapuschkin<sup>1</sup>✉

<https://doi.org/10.1038/s42256-023-00711-8>

# Explainability 2.0: Where, What and How

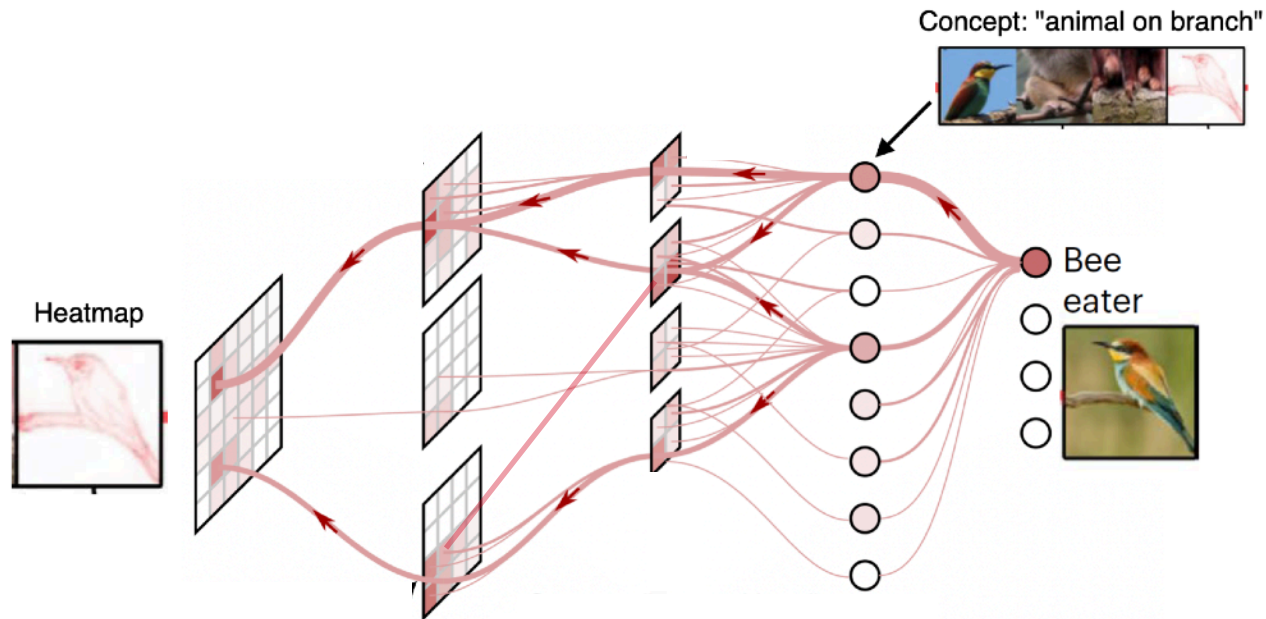
Known: Hidden layers encode semantic concepts.



Goal: Explain in terms of these concepts.

# Explainability 2.0: Where, What and How

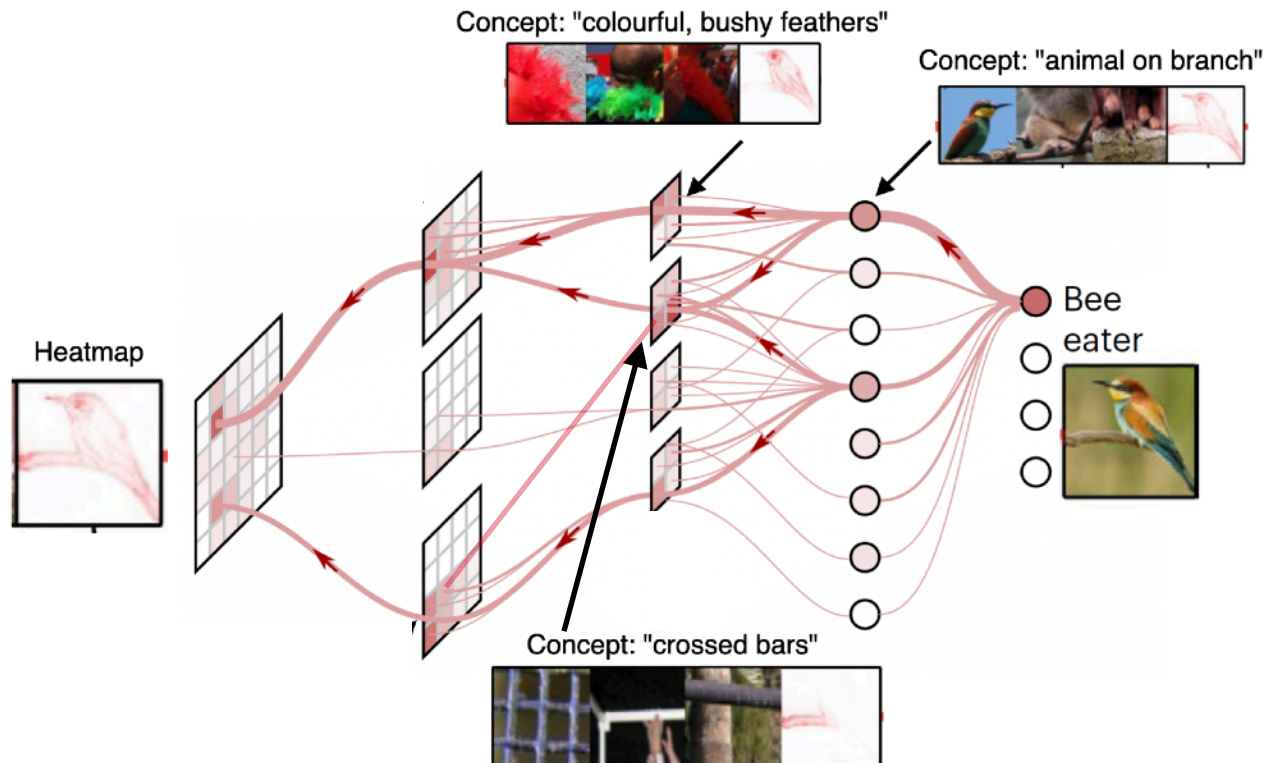
Known: Hidden layers encode semantic concepts.



Goal: Explain in terms of these concepts.

# Explainability 2.0: Where, What and How

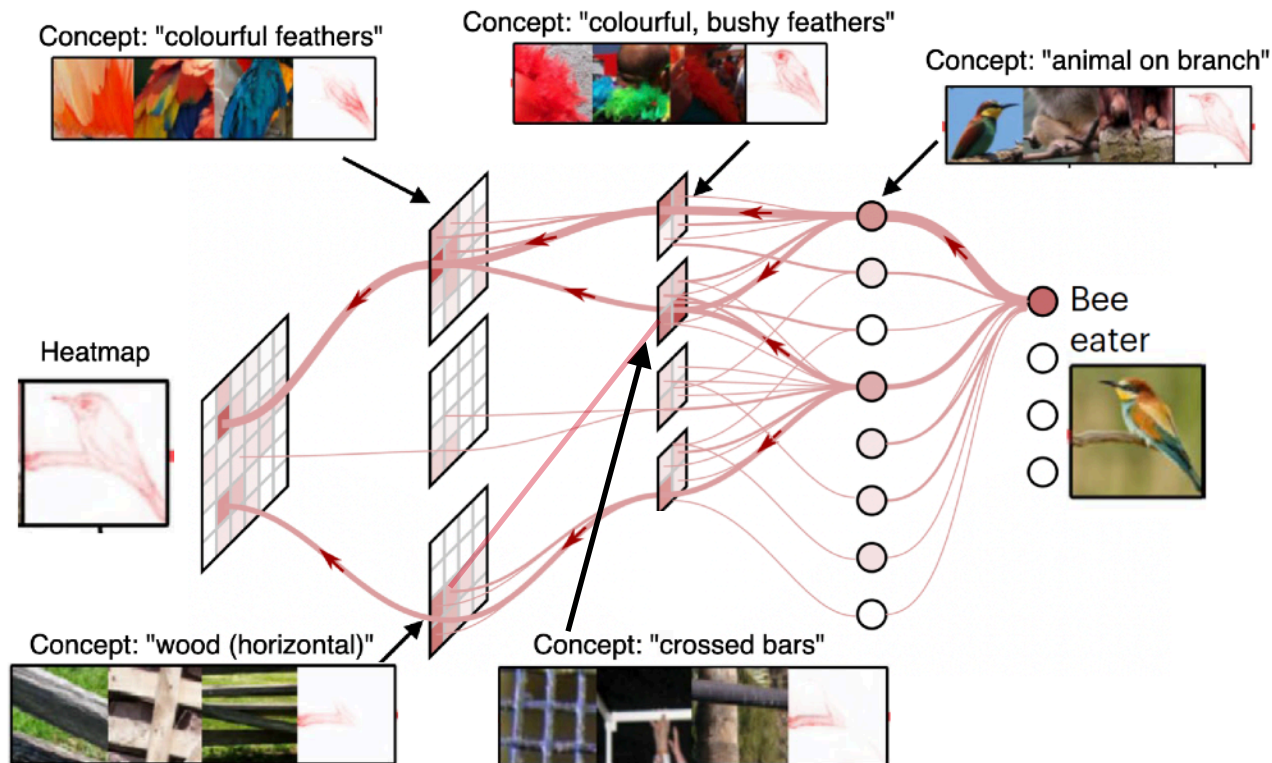
Known: Hidden layers encode semantic concepts.



Goal: Explain in terms of these concepts.

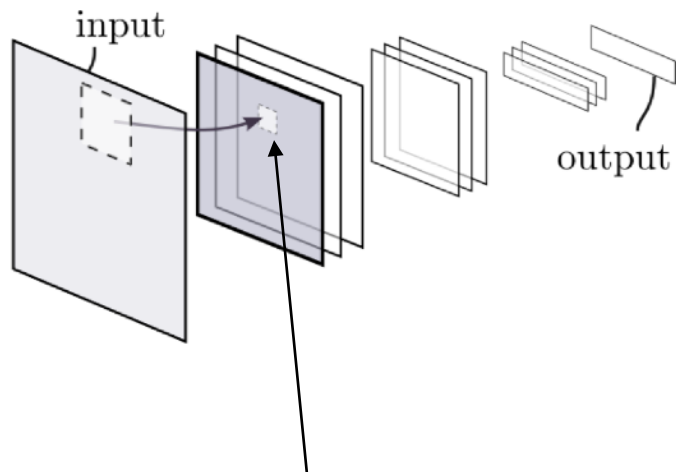
# Explainability 2.0: Where, What and How

Known: Hidden layers encode semantic concepts.



Goal: Explain in terms of these concepts.

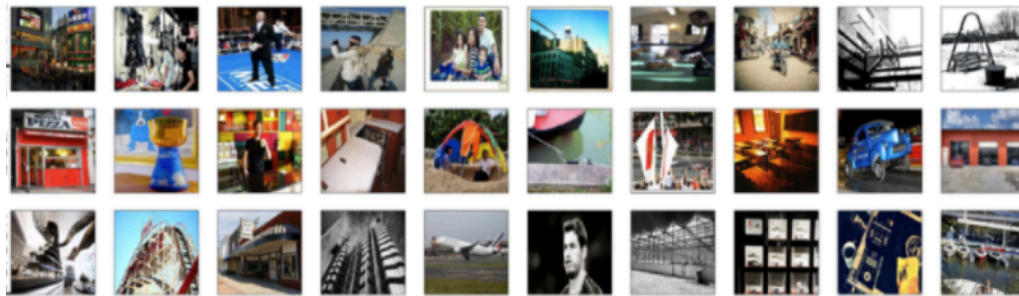
# Addressing the "What"-Question



What does this channel encode?

We can find out by activation maximization.

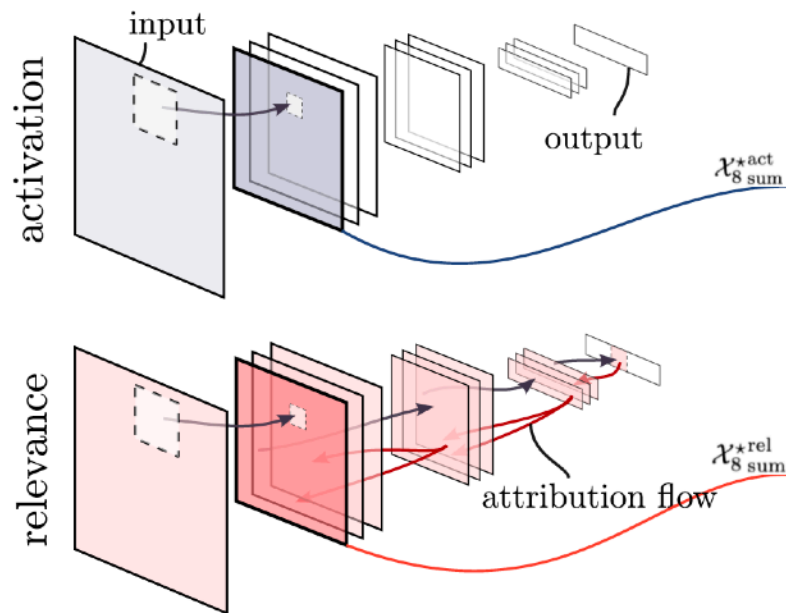
(Chen et al., 2020) data-based activation maximization



**Idea:** Image which maximally active the neuron contain the "concept" encoded by the neuron.

# From Activation To Relevance Maximization

activation vs relevance flow  $\longrightarrow$  result in different example sets



**without** task-context



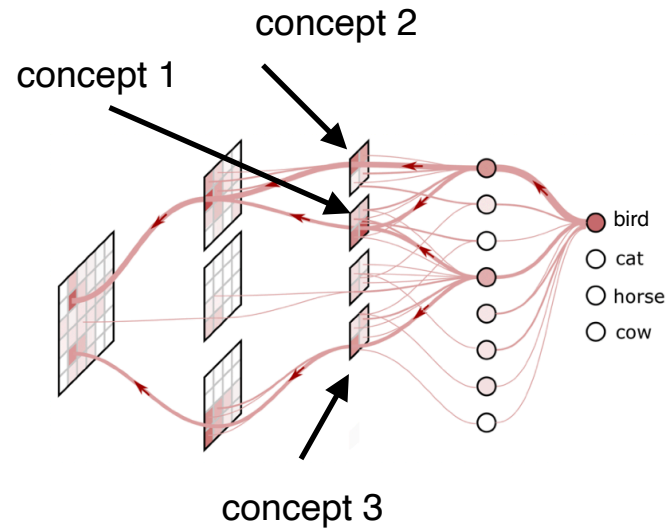
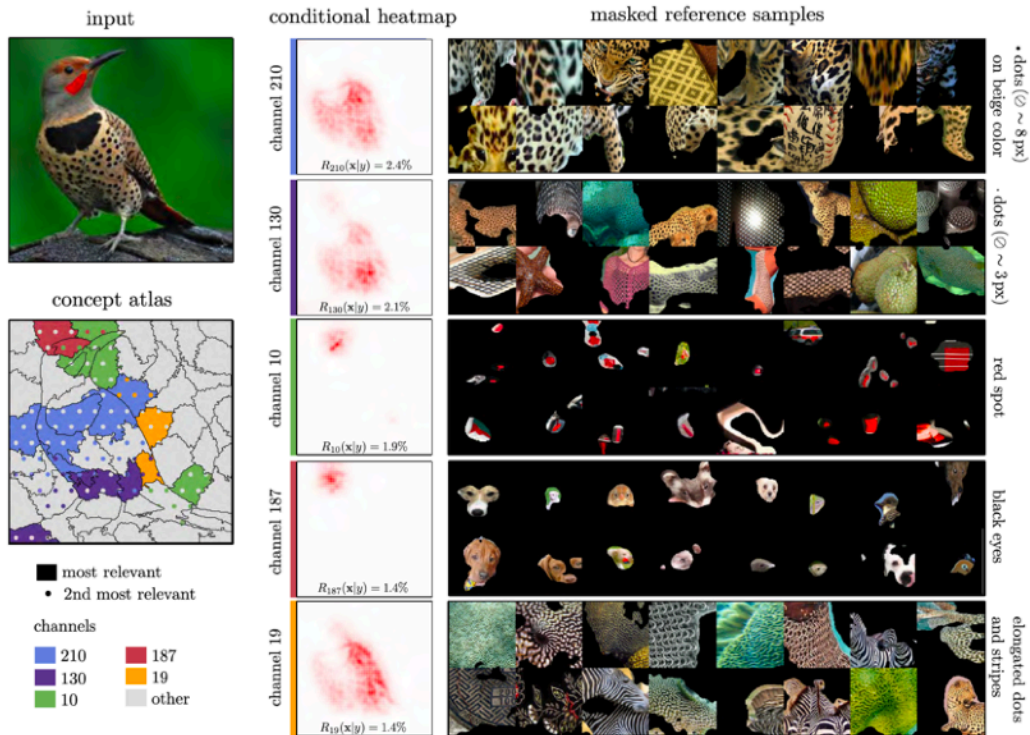
**AM:** Images which maximally "stimulate" channel  
(unrelated to prediction which we aim to explain)

**RM:** Images contain the concept encoded by the channel  
for which the channel is "most useful" to the model.



**within** task-context

# Concept Relevance Propagation (CRP)



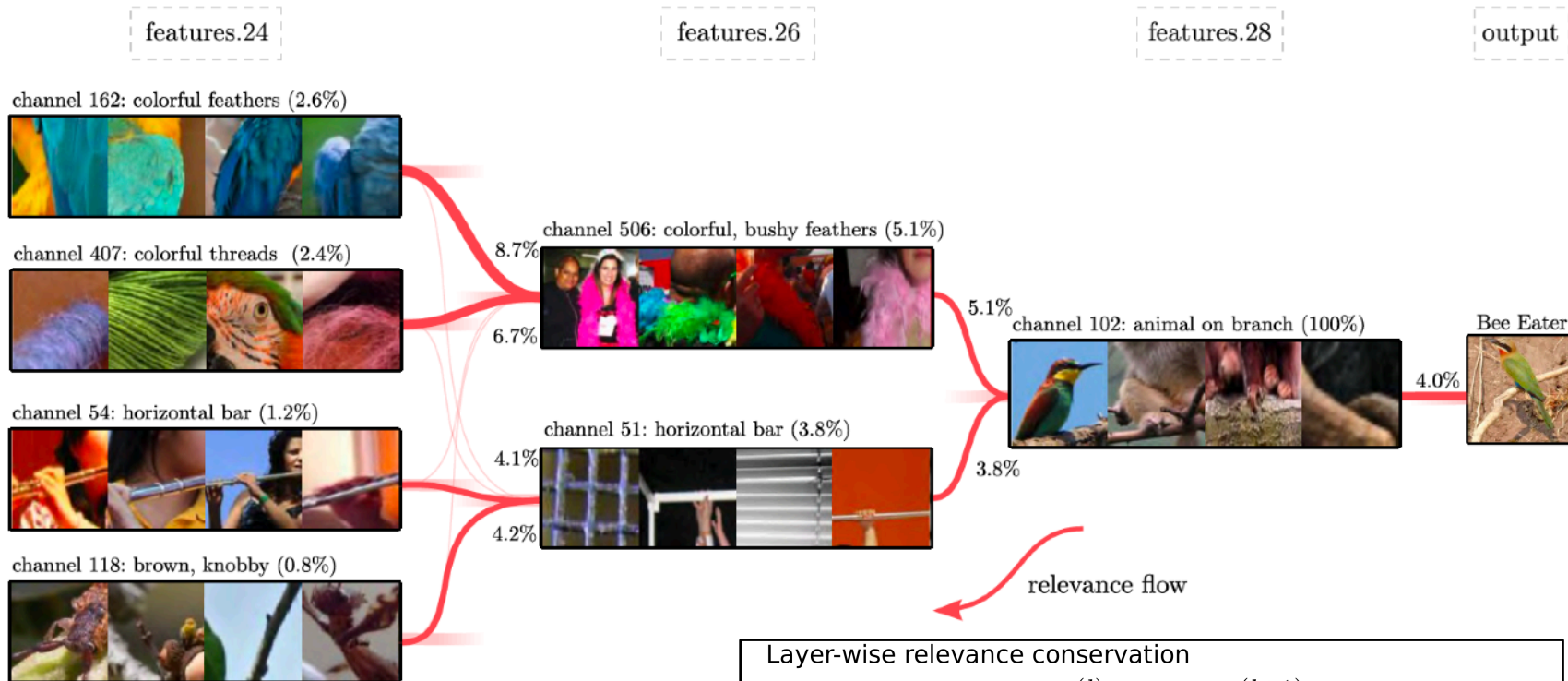
Step 1: Find relevant concepts

Step 2: Compute conditional explanation (*where*)

Step 3: Visualize relevant samples (*what*)



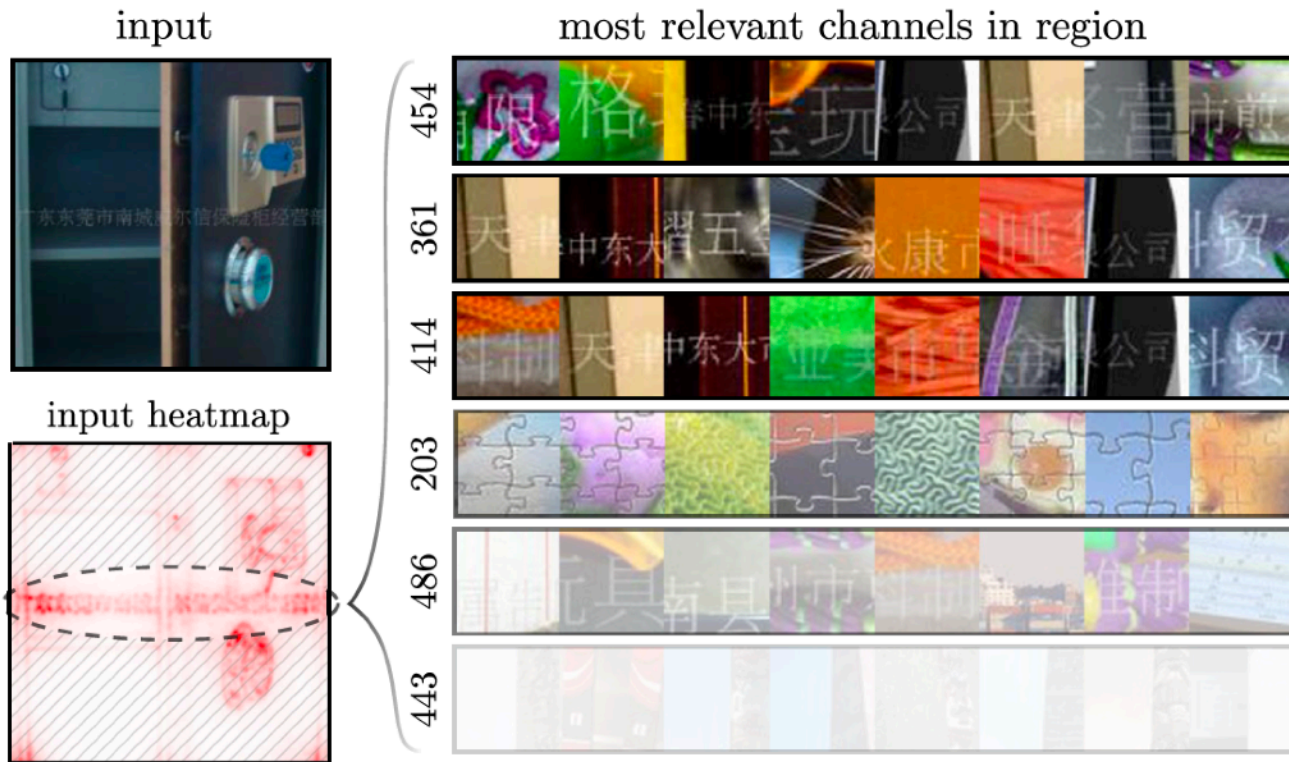
# Concept Composition



Layer-wise relevance conservation

$$\sum_i R_i = \dots = \sum_i R_i^{(l)} = \sum_j R_j^{(l+1)} = \dots = f(x)$$

# Concept-based Reverse Search



Reverse search: Find other images, where these channels are also relevant.

# Concept-based Reverse Search

whistle



mob



screw



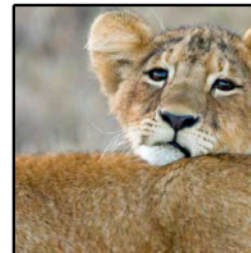
mosquito net



can opener



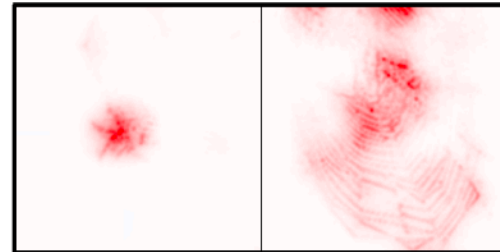
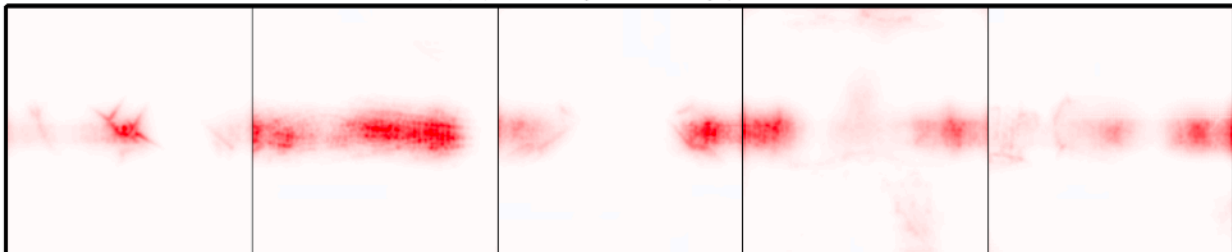
puma



spiderweb



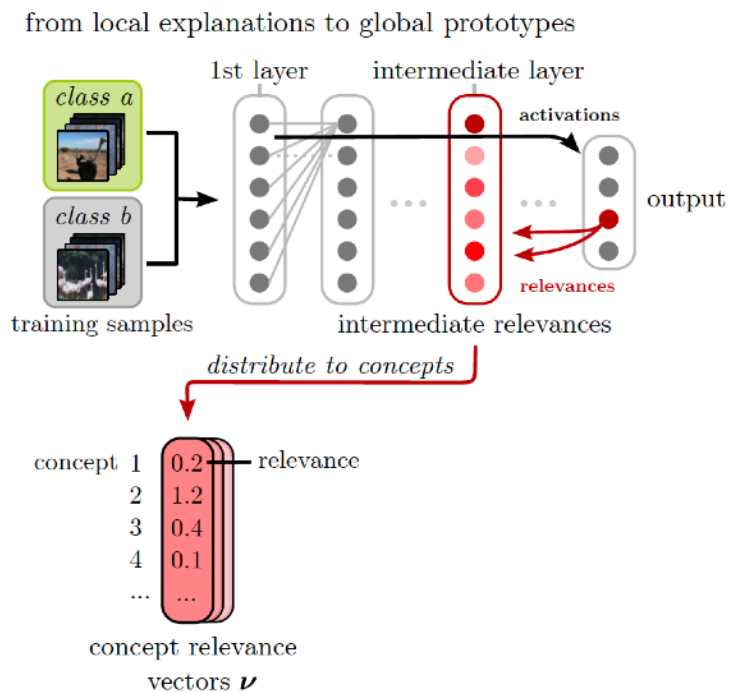
conditional heatmap  $R(\mathbf{x}|\theta = \{c_{361}, y\})$



**Fixing the Model:** Adapt encoding space globally [Anders, Weber, et al. 2022] or rather outcome-dependently?

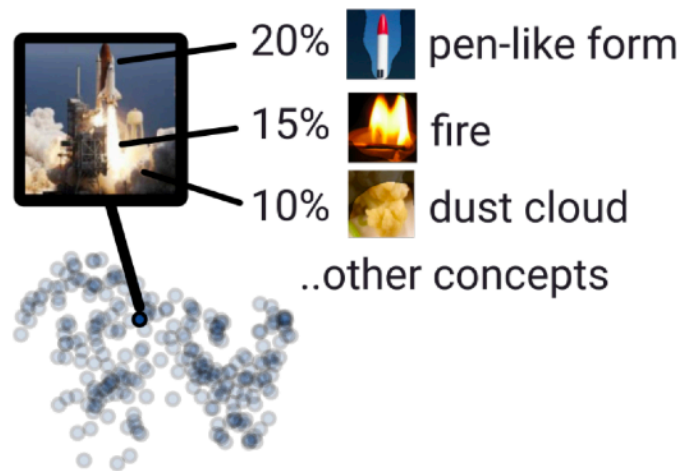
# From Individual Explanations to Understanding Global Behaviour

# Prototypical Concept-based Explanation (PCX)



1 collect local explanations

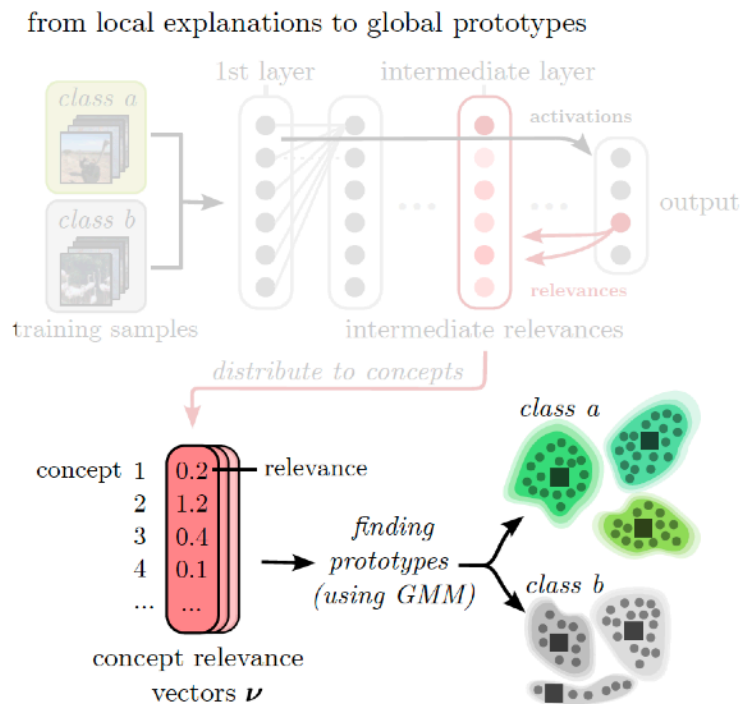
class **space shuttle**



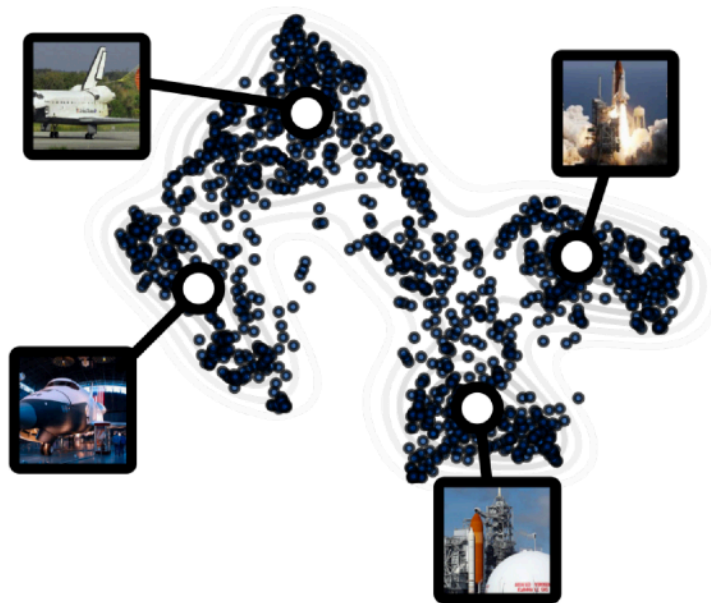
(Dreyer et al. 2024)

<https://doi.org/10.1109/CVPRW63382.2024.00353>

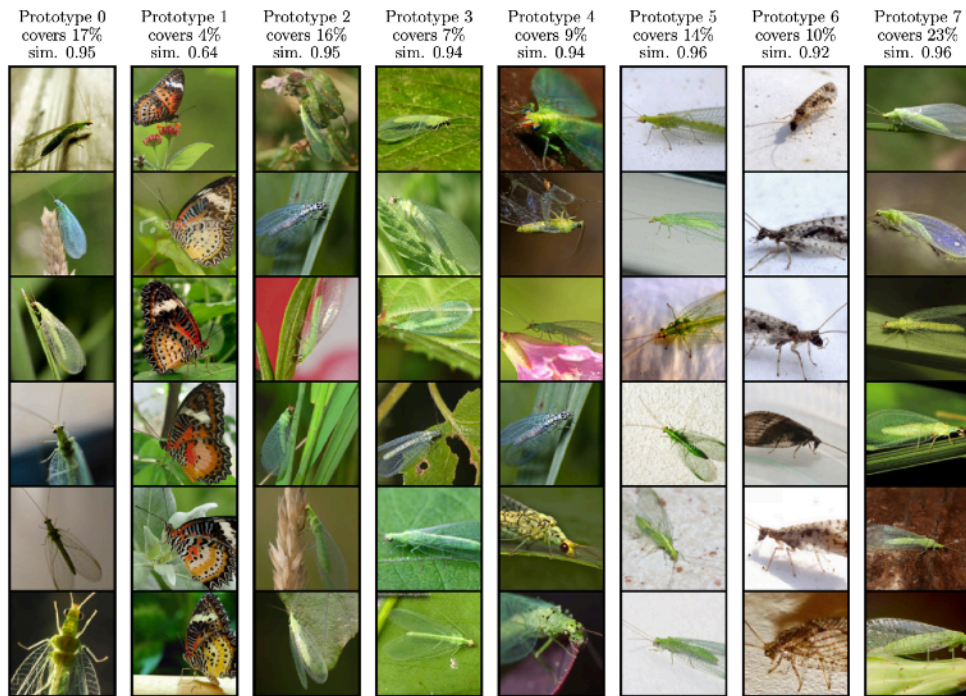
# Prototypical Concept-based Explanation (PCX)



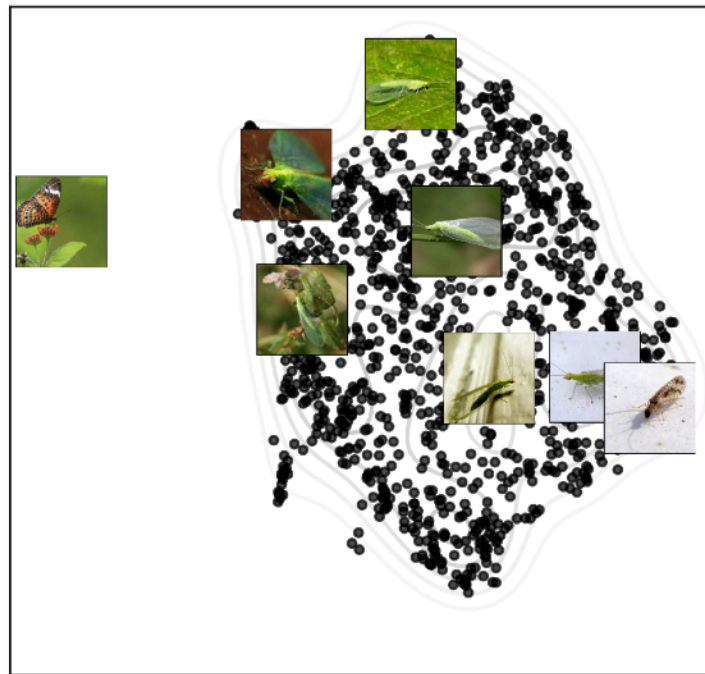
## 2 find prototypes



# Identifying Bugs in the Model / Data

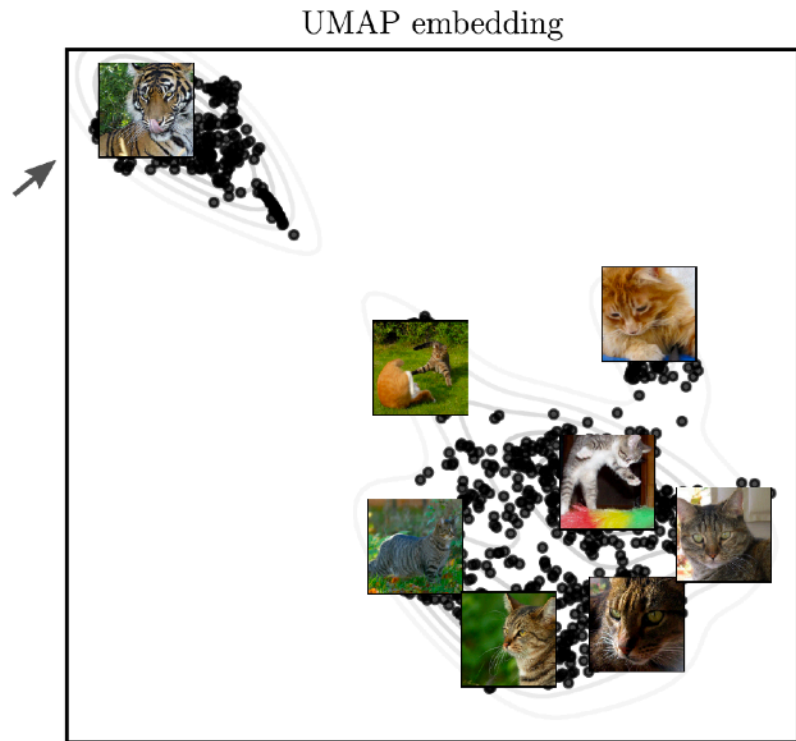
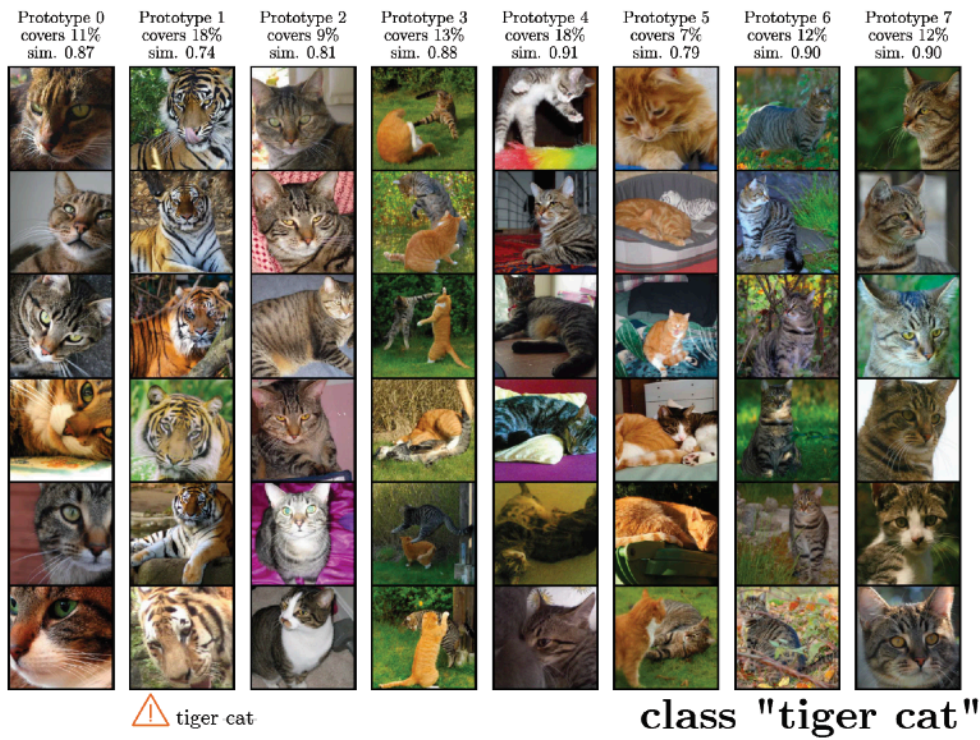


UMAP embedding



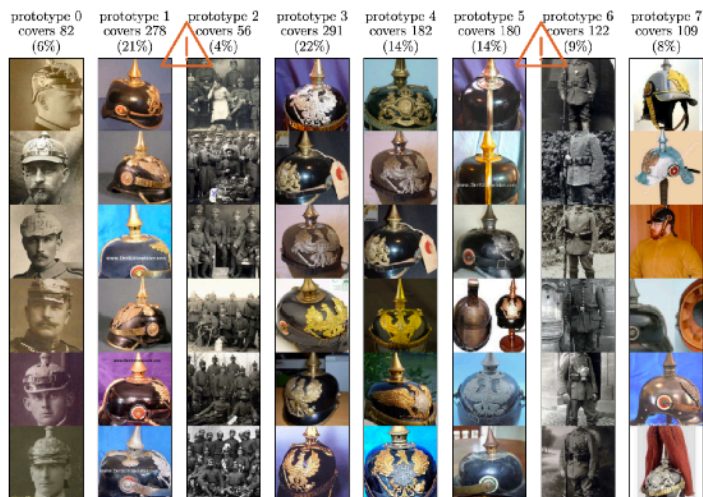
For the ImageNet class “lacewing” (VGG-16, layer features.28), there are also samples of Leopard Lacewing butterflies in the training data.

# Identifying Bugs in the Model / Data



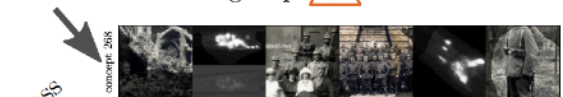


# class "pickelhaube"



prototypes

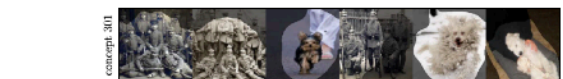
black & white & groups



brass



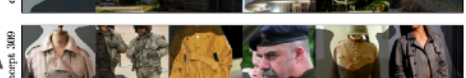
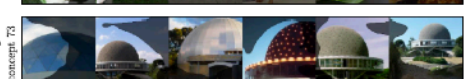
concepts



dome-like

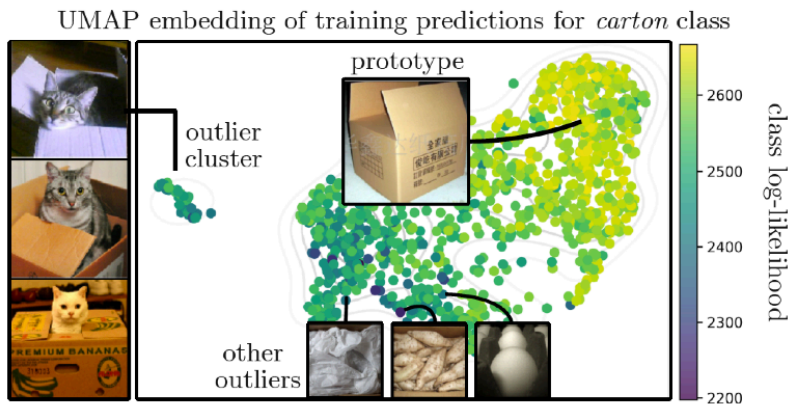


uniform  
coat

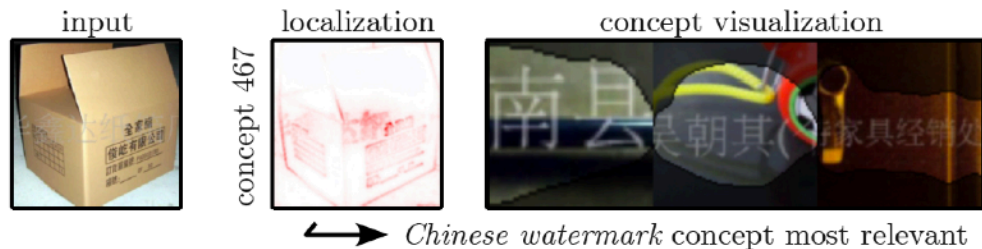


	prototype 0 covers 82 (6%)	prototype 1 covers 278 (21%)	prototype 2 covers 56 (4%)	prototype 3 covers 291 (22%)	prototype 4 covers 182 (14%)	prototype 5 covers 180 (14%)	prototype 6 covers 122 (9%)	prototype 7 covers 109 (8%)
concept 506	4.9	0.7	6.2	0.5	0.6	0.6	5.8	1.0
concept 6	4.9	5.1	4.1	5.3	6.1	3.2	4.3	3.2
concept 301	1.4	2.2	5.6	3.7	4.1	2.0	3.2	1.4
concept 115	0.9	2.5	0.5	4.6	3.5	3.8	0.3	1.6
concept 73	1.1	3.7	0.3	3.5	2.7	3.4	0.1	1.6
concept 300	2.8	1.4	2.7	1.5	1.3	1.3	3.1	1.8

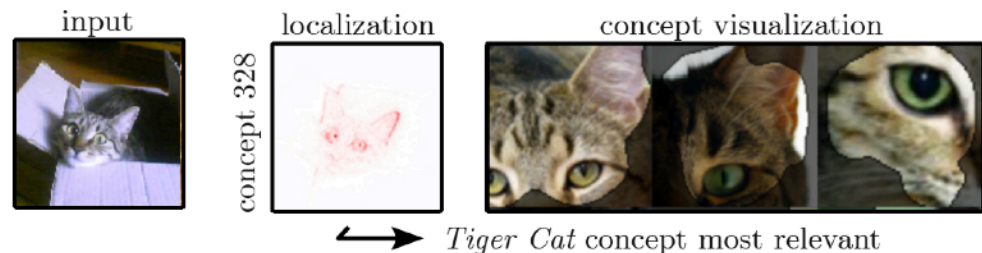
# Validating AI Predictions



a checking prototypes for spurious behavior



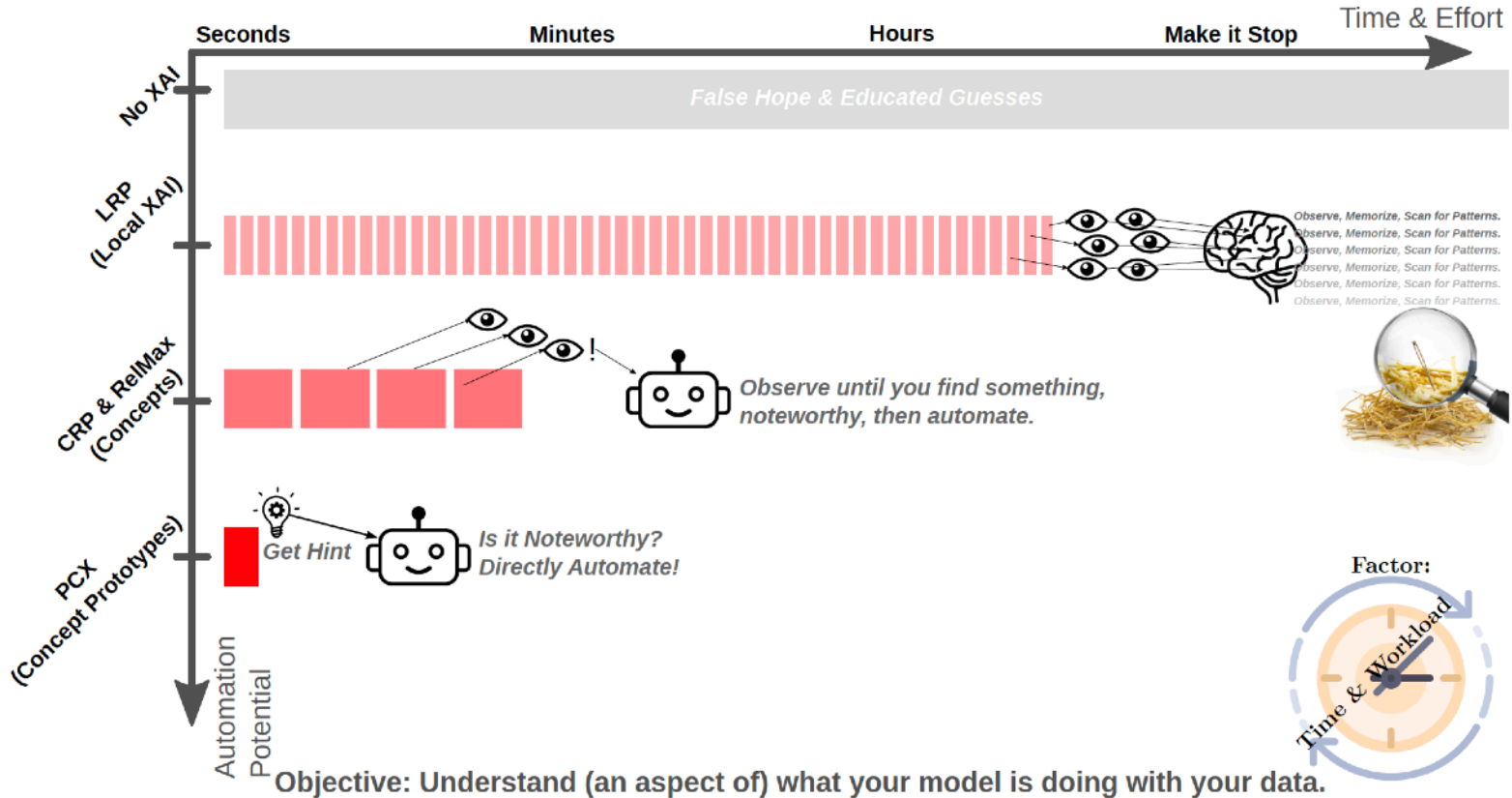
b checking outlier clusters for spurious behavior



Firstly, we examine the characteristic concepts of each prototype.

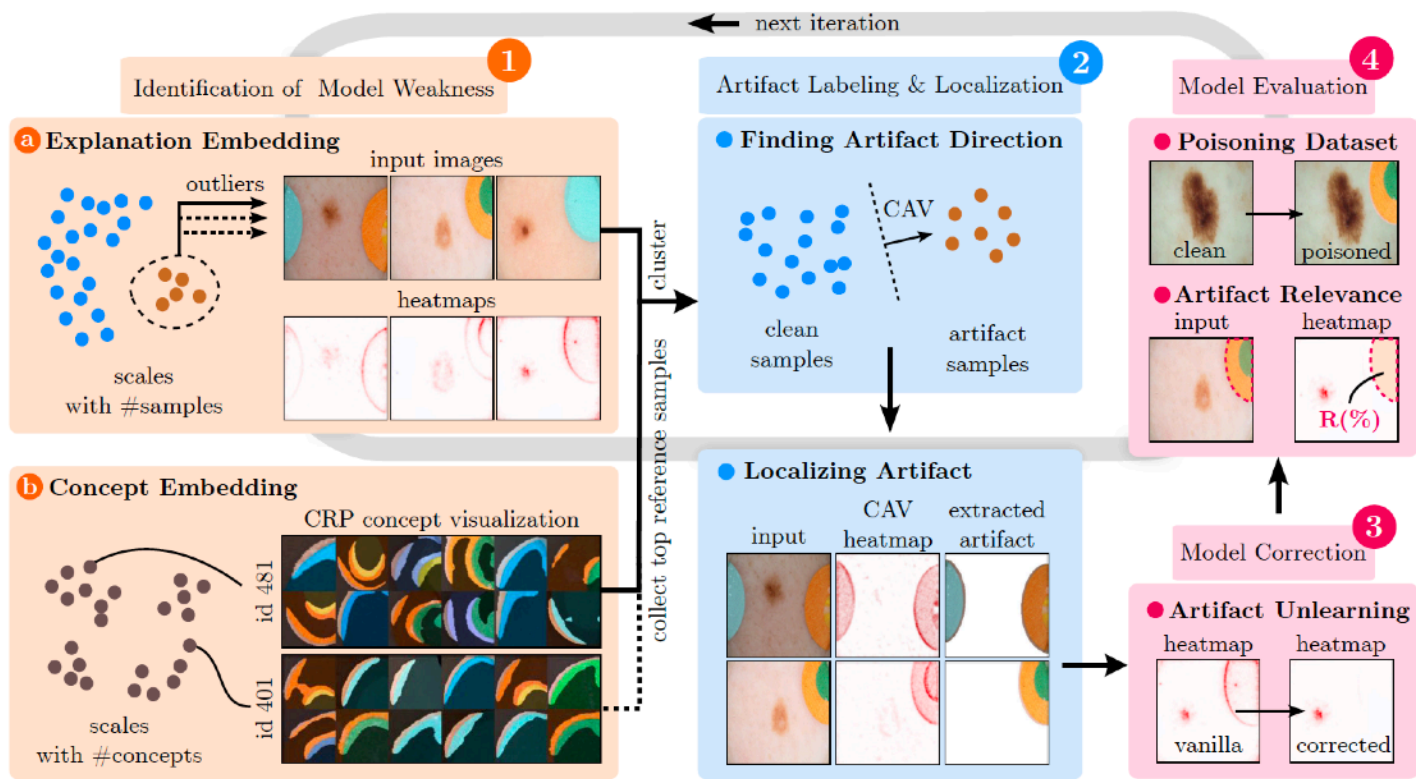
Secondly, clusters of training predictions that deviate strongly from prototypes can be studied for spurious behavior.

# How much Manual Work does XAI Require?



# From Explainable to Trustworthy AI

# Reveal and Revise Framework

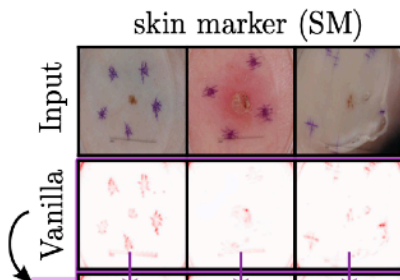


[Pahde et al. 2023]

# Example

ISIC Dataset Artifacts:

- Skin Marker



--- Reveal Step ---

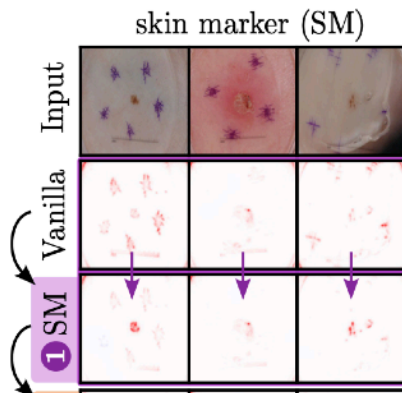
R2R iteration	corrected artifacts	↓ artifact relevance (%)			↑ F1 (%)			↑ accuracy (%)				
					<i>poisoned</i>	<i>original</i>		<i>poisoned</i>	<i>original</i>			
0	-	18.4	45.5	24.2	61.3	59.7	60.5	73.9	71.8	71.5	68.7	<b>80.1</b>

[Pahde et al. 2023]

# Example

ISIC Dataset Artifacts:

- Skin Marker



--- Revise Step ---

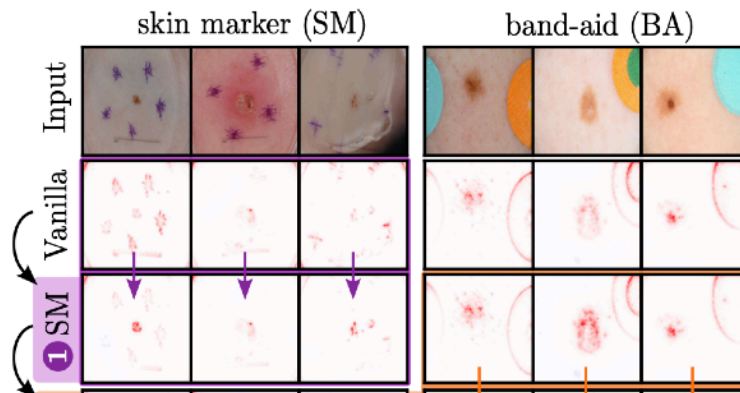
R2R iteration	corrected artifacts	↓ artifact relevance (%)			↑ F1 (%)				↑ accuracy (%)			
					<i>poisoned</i>		<i>original</i>		<i>poisoned</i>		<i>original</i>	
0	-	18.4	45.5	24.2	61.3	59.7	60.5	73.9	71.8	71.5	68.7	<b>80.1</b>
1	SM	13.1	35.0	21.3	61.6	61.0	60.7	73.8	72.2	72.6	68.4	80.0

[Pahde et al. 2023]

# Example

ISIC Dataset Artifacts:

- Skin Marker
- Band-Aid



--- Reveal Step ---

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

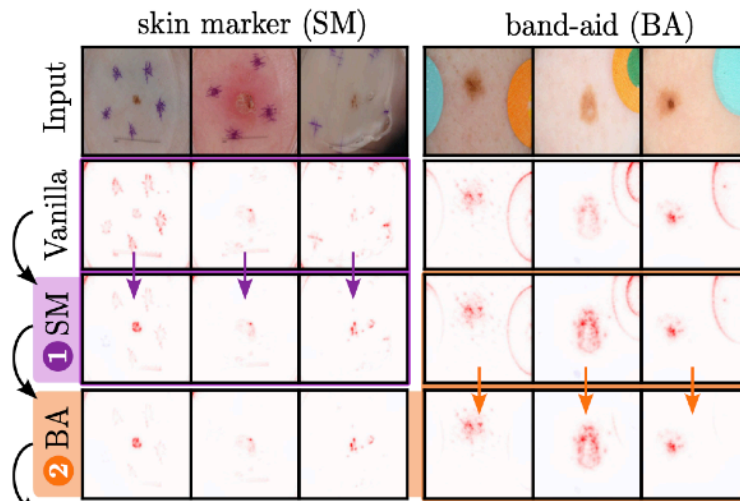
[Pahde et al. 2023]



# Example

ISIC Dataset Artifacts:

- Skin Marker
- Band-Aid



--- Revise Step ---

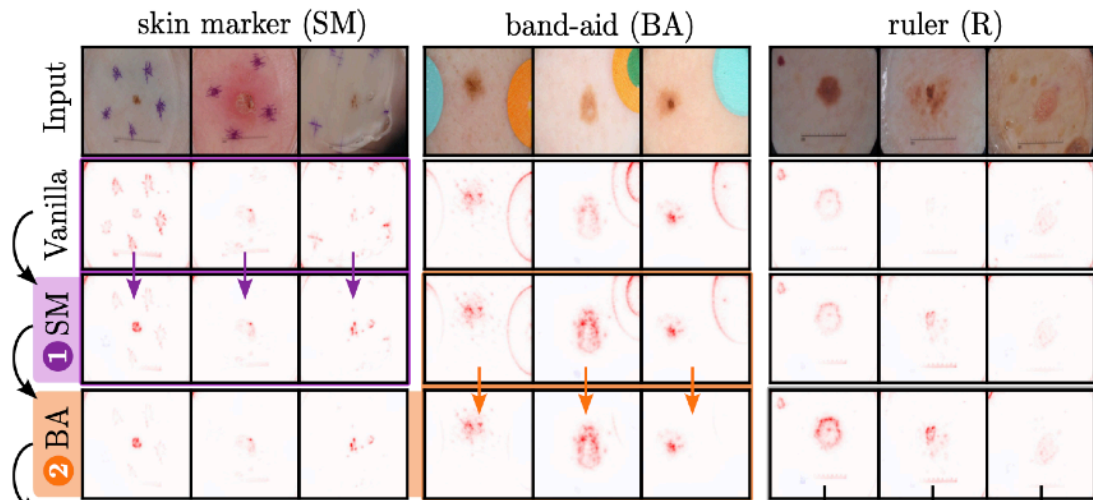
R2R iteration	corrected artifacts	↓ artifact relevance (%)			↑ F1 (%)			↑ accuracy (%)				
					<i>poisoned</i>	<i>original</i>		<i>poisoned</i>	<i>original</i>			
0	-	18.4	45.5	24.2	61.3	59.7	60.5	73.9	71.8	71.5	68.7	<b>80.1</b>
1	SM	13.1	35.0	21.3	61.6	61.0	60.7	73.8	72.2	72.6	68.4	80.0
2	SM, BA	<b>12.8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

[Pahde et al. 2023]

# Example

ISIC Dataset Artifacts:

- Skin Marker
- Band-Aid
- Ruler



--- Reveal Step ---

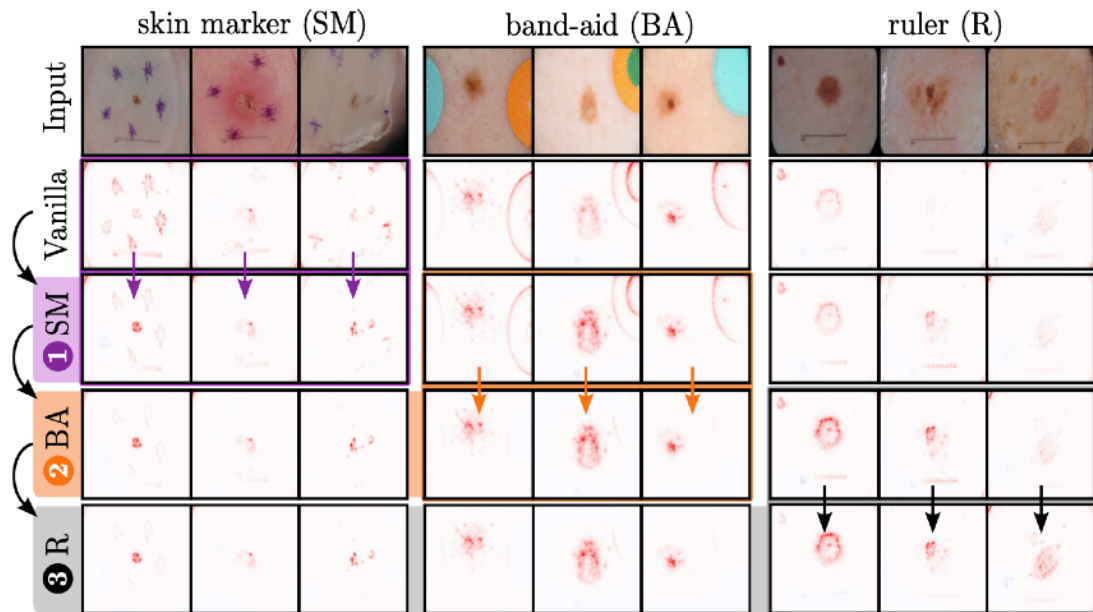
R2R iteration	corrected artifacts	↓ artifact relevance (%)			↑ F1 (%)			↑ accuracy (%)				
					<i>poisoned</i>	<i>original</i>		<i>poisoned</i>	<i>original</i>			
0	-	18.4	45.5	24.2	61.3	59.7	60.5	73.9	71.8	71.5	68.7	<b>80.1</b>
1	SM	13.1	35.0	21.3	61.6	61.0	60.7	73.8	72.2	72.6	68.4	80.0
2	SM, BA	<b>12.8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

[Pahde et al. 2023]

Back to LLMs

# Concept-Level Understanding

The San Diego Electric Railway (SDERy) was a mass transit system in Southern California, United States, using 600 volt DC streetcars and (in later years) buses.

Maintenance of way (commonly abbreviated to MOW) refers to the maintenance, construction, and improvement of rail infrastructure, including tracks, ballast, grade, and lineside infrastructure such as signals and signs.

France currently operates the second-largest European railway network, with a total of 29,901 kilometres of railway.

The MHR had, in 1846, amalgamated with the "Little" North Western Railway (NWR), which was taken over by the Midland Railway in 1874. Awdry, p.97 The rival London and North Western Railway (LNWR) built its own branch line to Morecambe in 1864, joining the main LNWR line at Hest Bank.

Some railway companies had a standard signalbox design, such as the London & North Western Railway, whereas others, such as the Great Eastern Railway had many different designs.

AttnLRP attributions on top 10 ActMax sentences collected over the Wikipedia summary dataset for neuron #256, inlayer 18. The knowledge neuron seems to activate for transport systems (railways in particular).

# Concept-Level Understanding

The volume presents six short stories, with the titular story featuring Yahiro, a substitute teacher, who begins having an affair with his student Kago.

In 2018, Derek Michael Boyce, a high school math and science teacher at the school, was arrested for having an inappropriate relationship with one of his students, a fifteen-year-old girl.

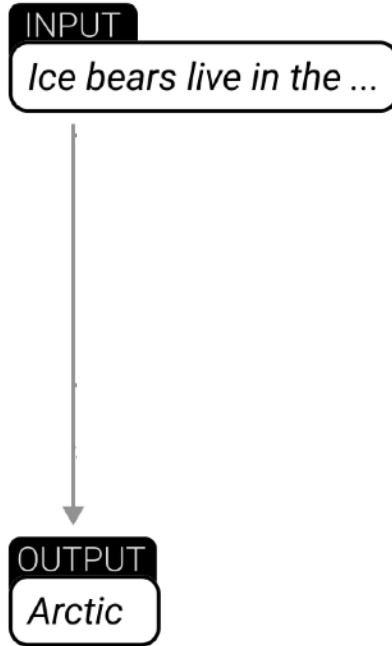
The film follows a school teacher as she suspects one of her students is suffering from personal problems in his home life, not knowing that the student is harboring an evil demon in his house.

During his time as a teacher Franco admitted to having sex with several of his students, which led to lawsuits and a \$2 million sexual-misconduct settlement in 2021.

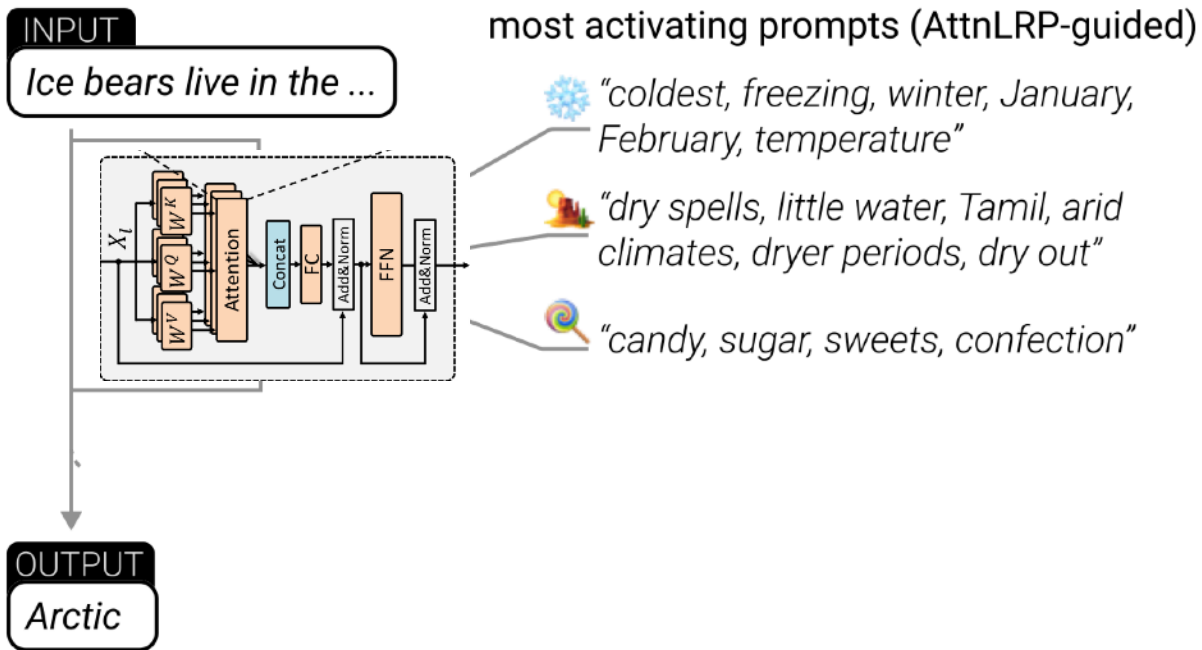
It tells the story of a schoolteacher who falls in love with one of his students, and moves away in order to escape his infatuation.

AttnLRP attributions on top 10 ActMax sentences collected over the Wikipedia summary dataset for neuron #2207, inlayer 20. The knowledge neuron is activating for 'teacher', in unusual context such as inappropriate behavior, sexual misconduct etc.

# Concept-Level Understanding

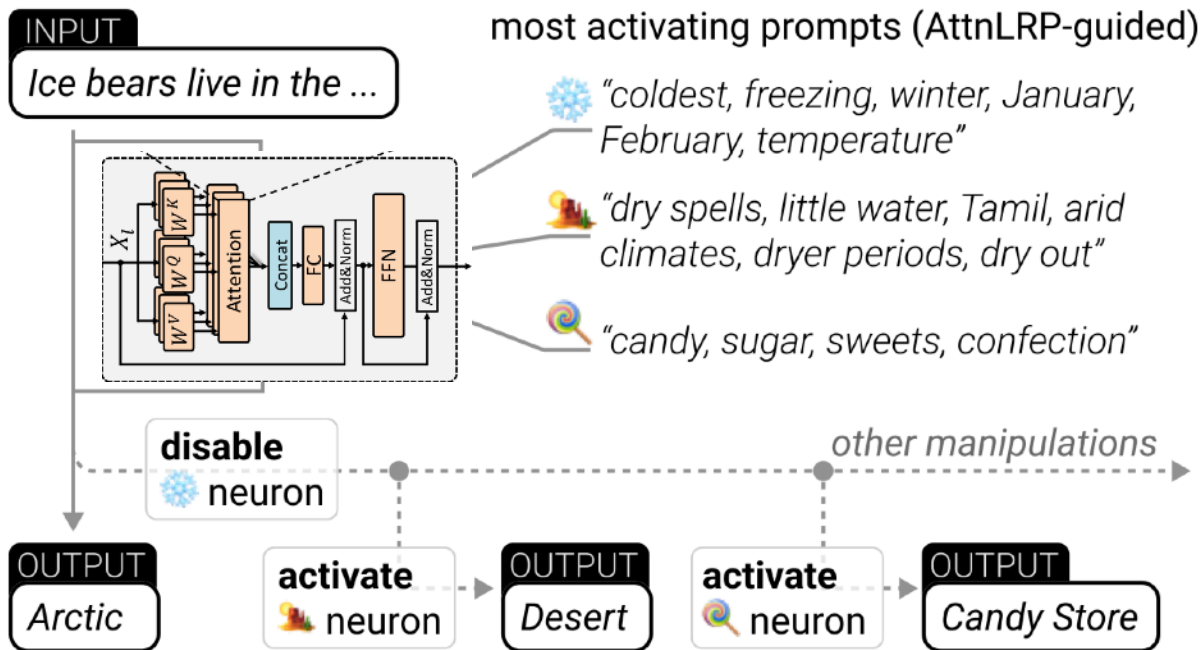


# Concept-Level Understanding

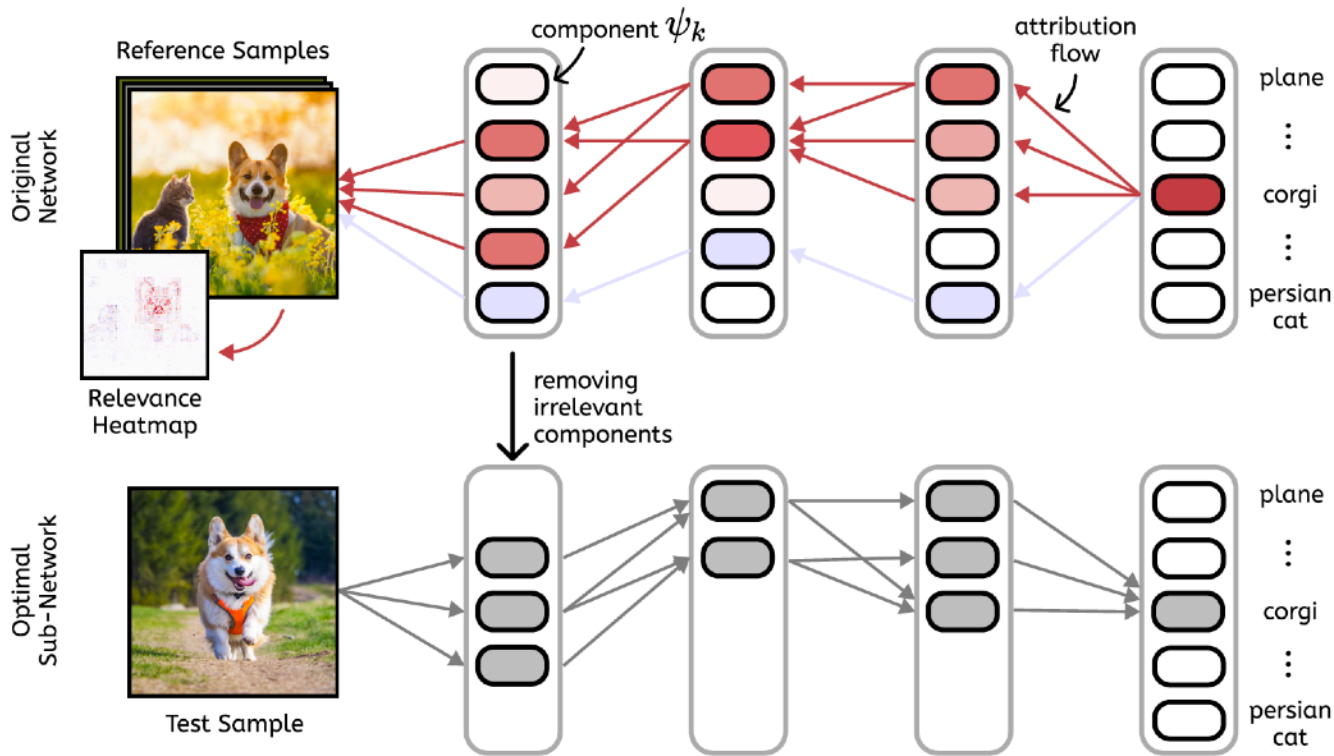




# Concept-Level Understanding



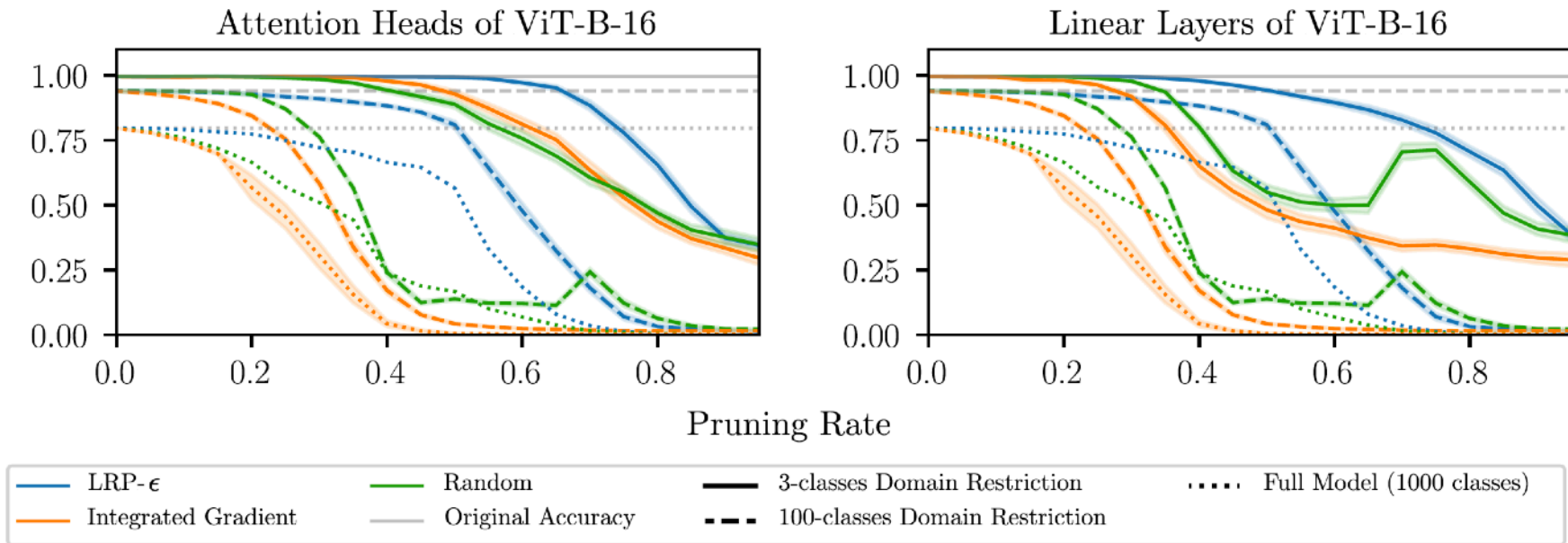
# Pruning By Explaining



(Hatefi et al. 2024)

<https://arxiv.org/pdf/2408.12568>

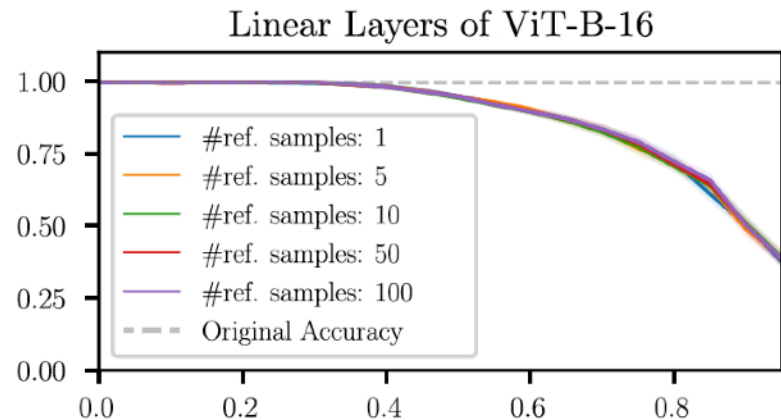
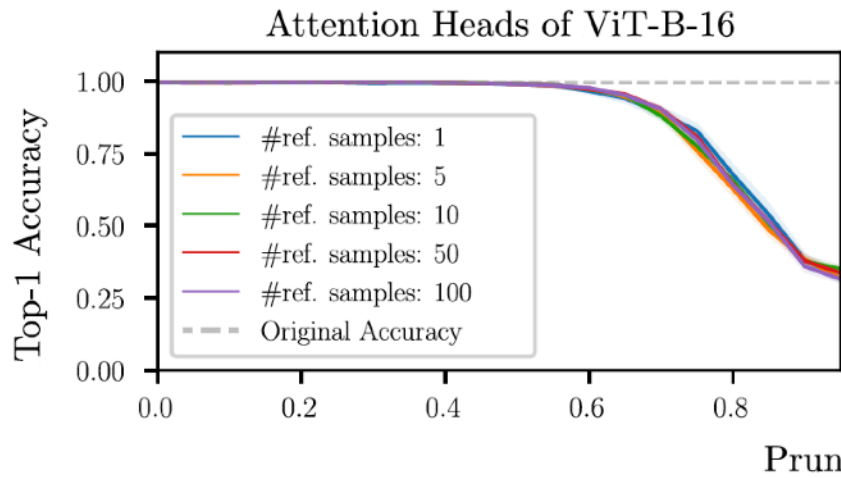
# Pruning By Explaining



(Hatefi et al. 2024)

<https://arxiv.org/pdf/2408.12568>

# Pruning By Explaining

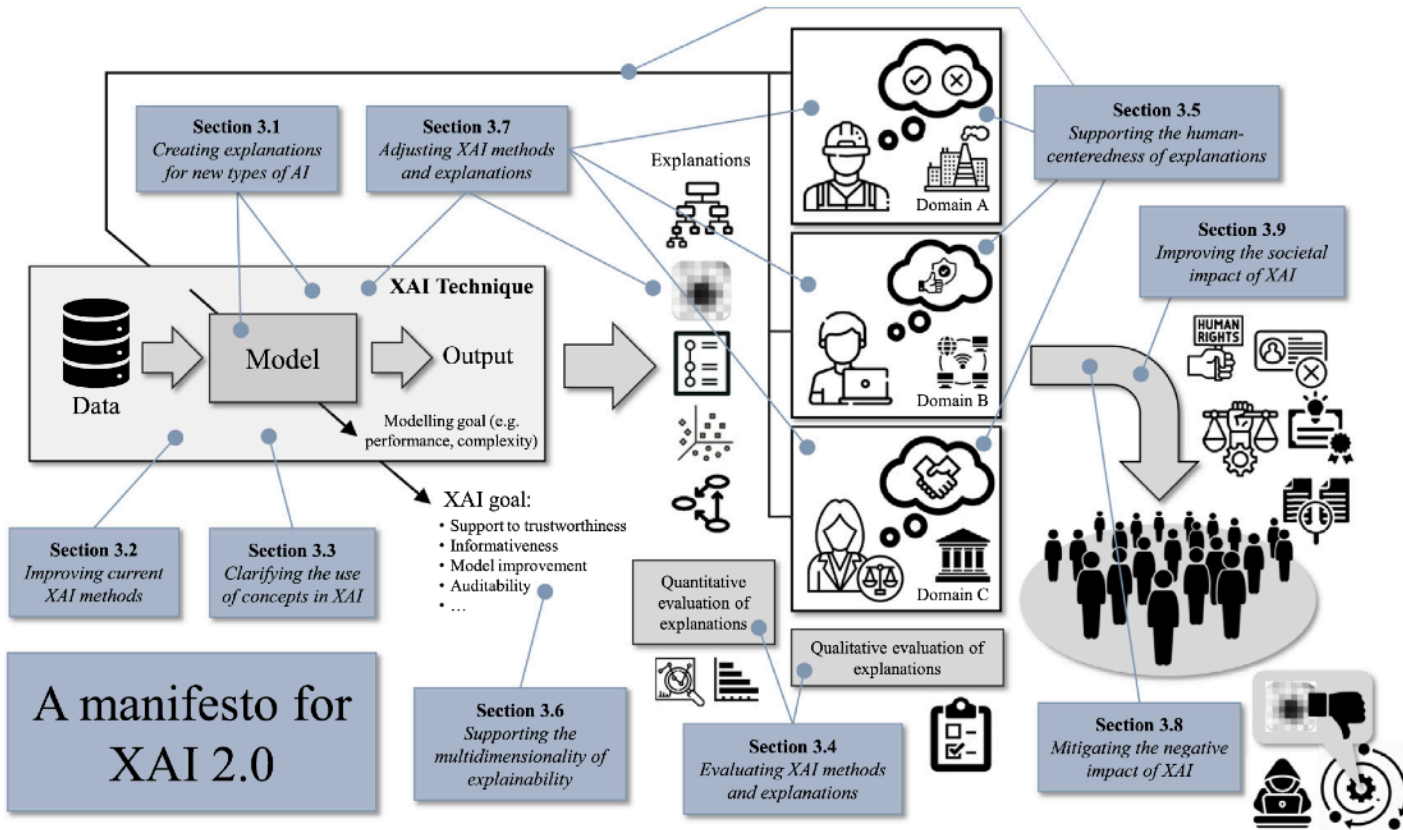


Required number of reference samples (per class) is very low.

(Hatefi et al. 2024)

<https://arxiv.org/pdf/2408.12568>

# Future Work



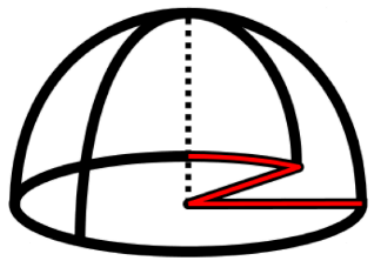
(Longo et al. 2024)

# Toolboxes

Benchmarking:

QUANTUS

<https://github.com/understandable-machine-intelligence-lab/Quantus>



**zennit**

<https://github.com/chr5tphr/zennit>

**iNNvestigate**

<https://github.com/albermax/innvestigate>



**ExplainableAI.jl**

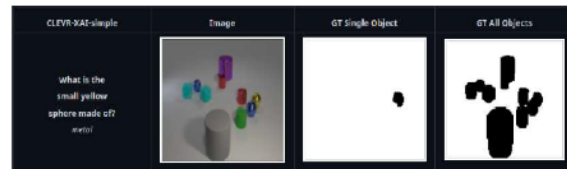
<https://github.com/adrhill/ExplainableAI.jl>



**quanda**

<https://github.com/dilyabareeva/quanda>

Benchmarking:  
**CLEVR-XAI**



<https://github.com/ahmedmagdiosman/clevr-xai>



<https://github.com/rachtibat/zennit-crp>



**THANK YOU  
FOR YOUR  
ATTENTION**