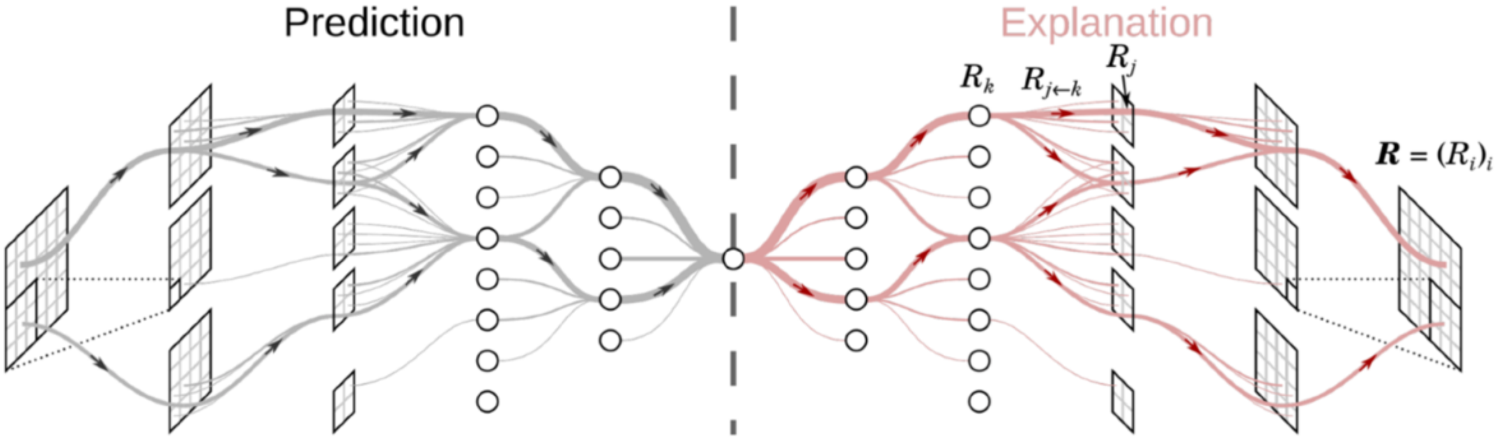
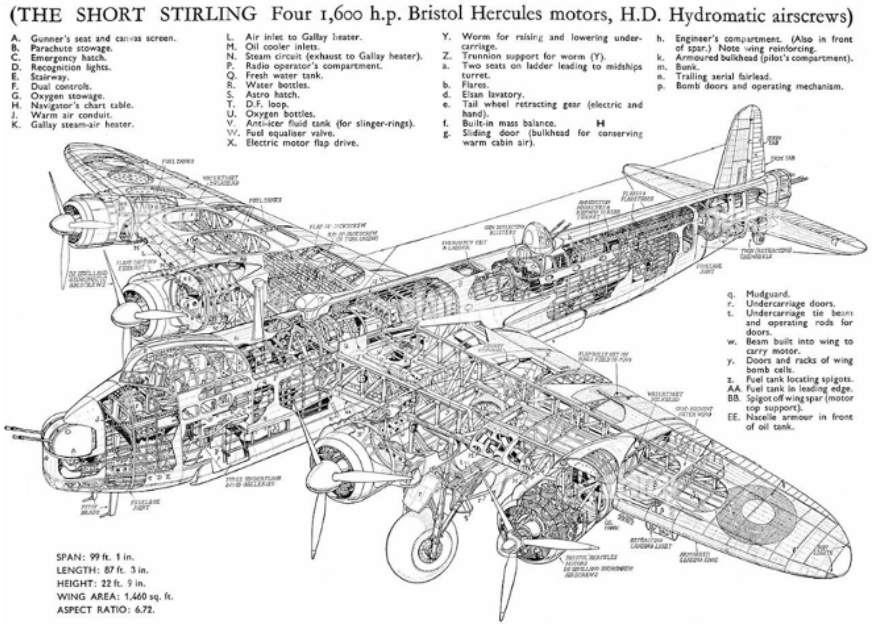
Inspecting AI Like Engineers: From Explanation to Validation with SemanticLens

Wojciech Samek
TU Berlin & Fraunhofer HHI



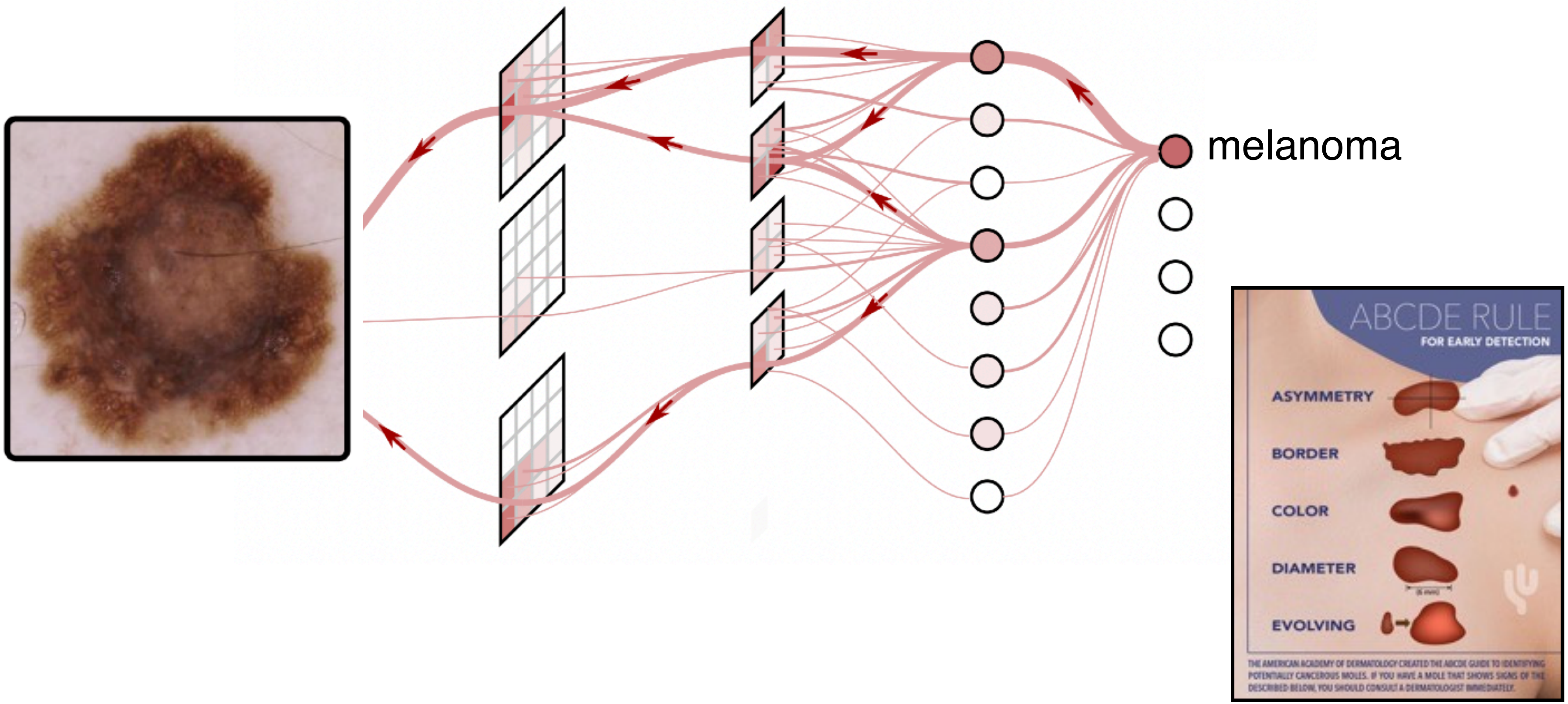
Inspecting AI Like Engineers

__Human-Engineered System__: Built of known components



__Modern AI__: Function of components unknown

Does my AI model follow the ABCDE rule ?



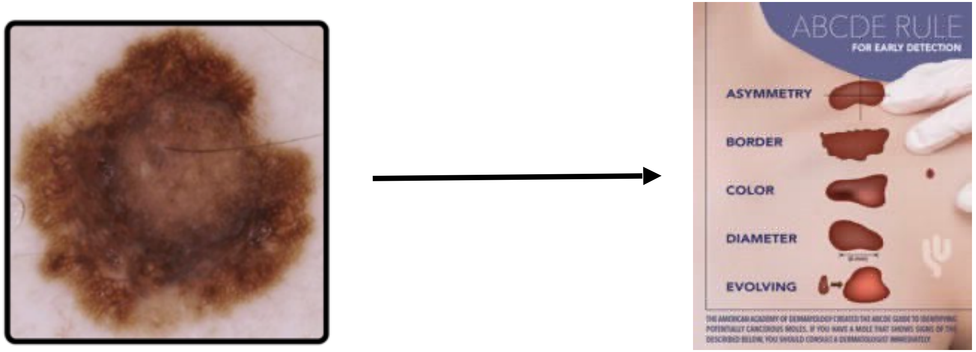
To trust or not to trust AI; that is the question

We need to to understand the “Black Box” at component-level

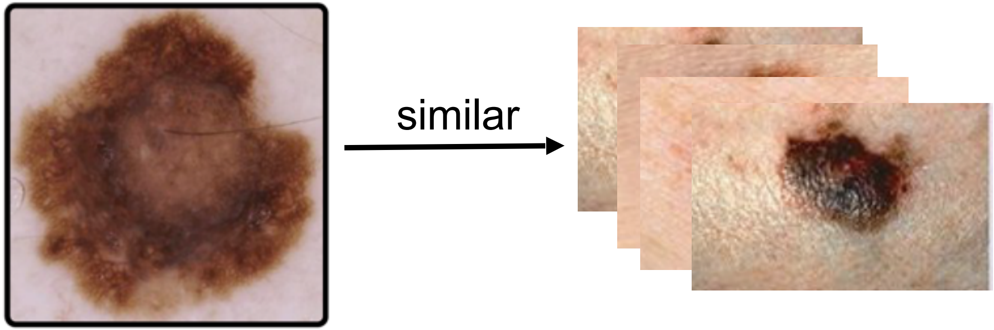
Explainable AI Research

Relevance-Based: Where does AI look at ?

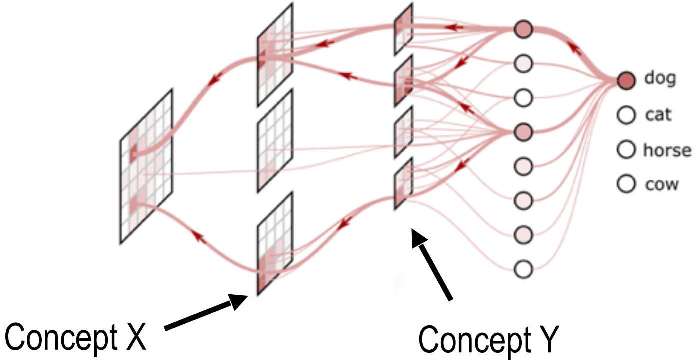
Concept-Based: Which concepts / rules does AI use?



Example-based: Are there similar cases ?



Model-based: What does model represent internally ?



First Wave of XAI: "Understand Prediction"

First Wave of Explainable AI

Zintgraf'17
Pred Diff

Ribeiro'16
LIME

Bach'15
LRP

Zhou'16
GAP

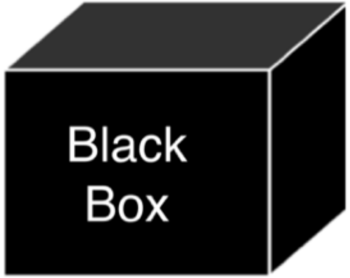
Zeiler'14
Occlusion

Fong'17
M Perturb

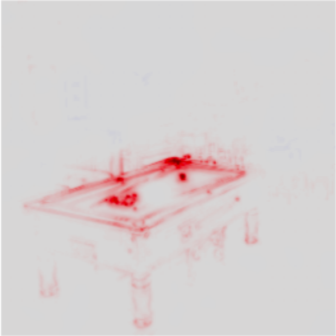
Zeiler'14
Deconv

Kindermanns'17
PatternNet

classify:
"pool table"



explain:
"heatmap"



Zhang'16
Excitation BP

Landecker'13
Contrib Prop

Lungberg'17
SHAP

Shrikumar'17
DeepLIFT

Springenberg'14
Guided BP

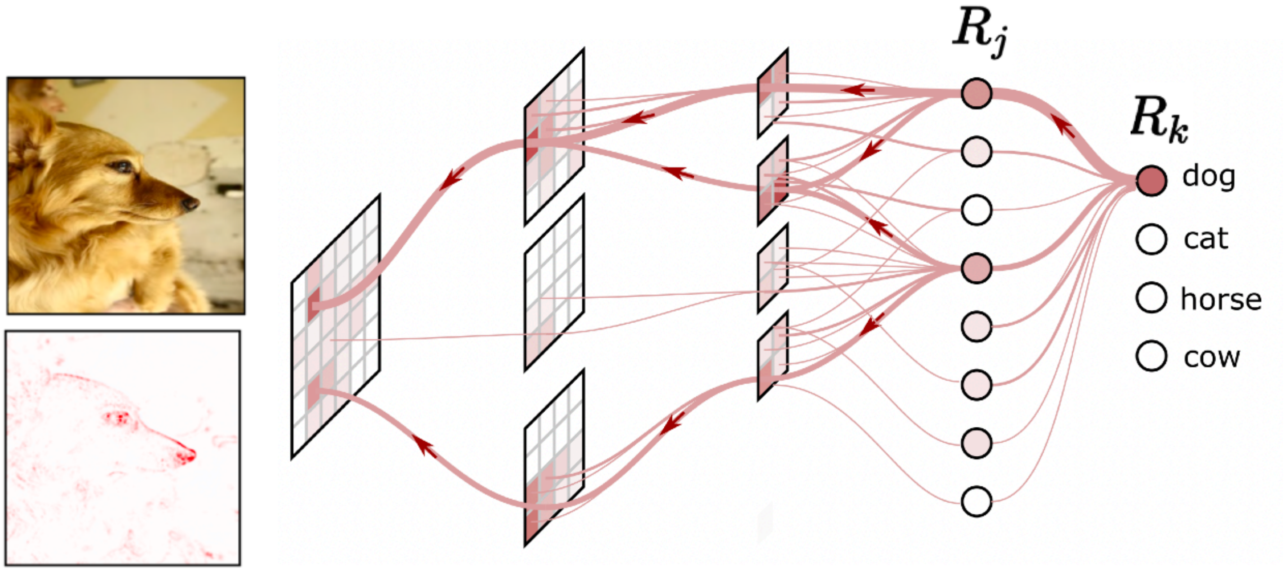
Selvaraju'17
Grad-CAM

Symonian'13
Gradient

Sundarajan'17
Int Grad

Montavon'17
Deep Taylor

Layer-wise Relevance Propagation (LRP)



(1) decompose

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

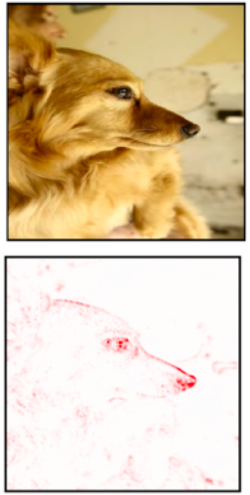
(2) aggregate

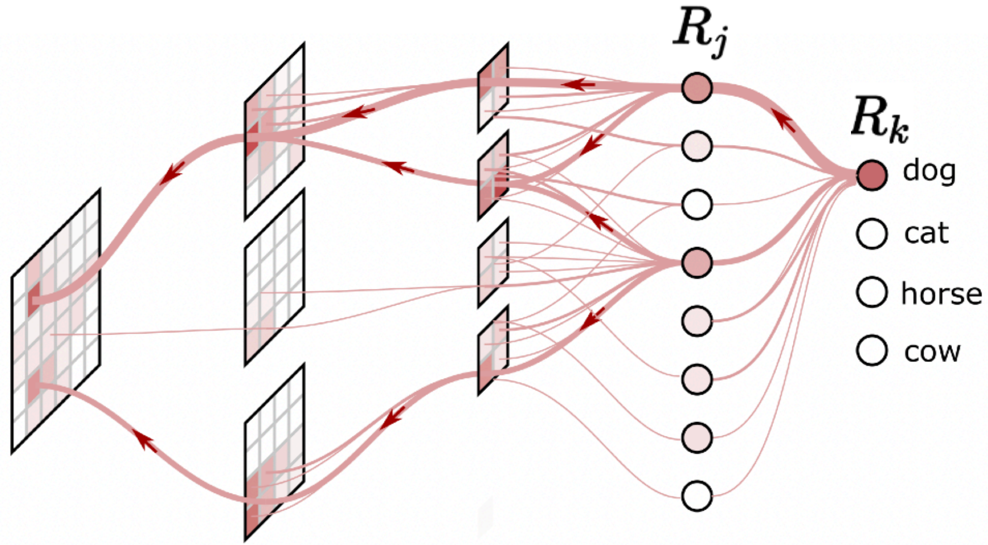
$$R_j = \sum R_{j\leftarrow k}$$

zjk measures how much has j contributed to activation of k

(Bach et al. 2015)

Layer-wise Relevance Propagation (LRP)





Advantages

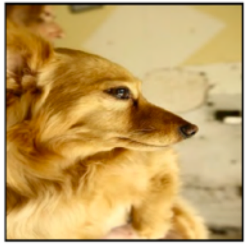
- efficient & faithful
- relevance values for all elements of NN
- applicable to non-differentiable layers (no gradient shattering)

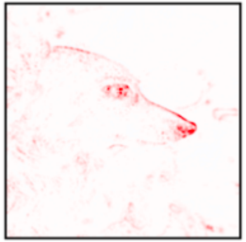
Layer-wise relevance conservation

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

Which redistribution rule is the right one (i.e. how to best measure zjk)?

Layer-wise Relevance Propagation (LRP)





Layer-wise rele

$\sum_i R_i = \ldots$

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-ϵ [7]	$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k$	middle layers	✓
LRP-γ	$R_j = \sum_k \frac{a_j (w_{jk} + \gamma w_{jk}^+)}{\sum_{0,j} a_j (w_{jk} + \gamma w_{jk}^+)} R_k$	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	$\times^\star$
flat [30]	$R_j = \sum_k \frac{1}{\sum_j 1} R_k$	lower layers	$\times$
w^2-rule [36]	$R_i = \sum_j \frac{w_{ij}^2}{\sum_i w_{ij}^2} R_j$	first layer ($\mathbb{R}^d$)	✓
$z^{\mathcal{B}}$-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 $\alpha = 1, \beta = 0$.)

ages

nt & faithful

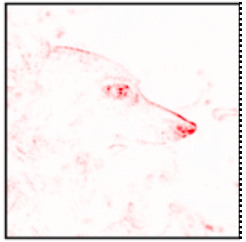
nce values for
ents of NN

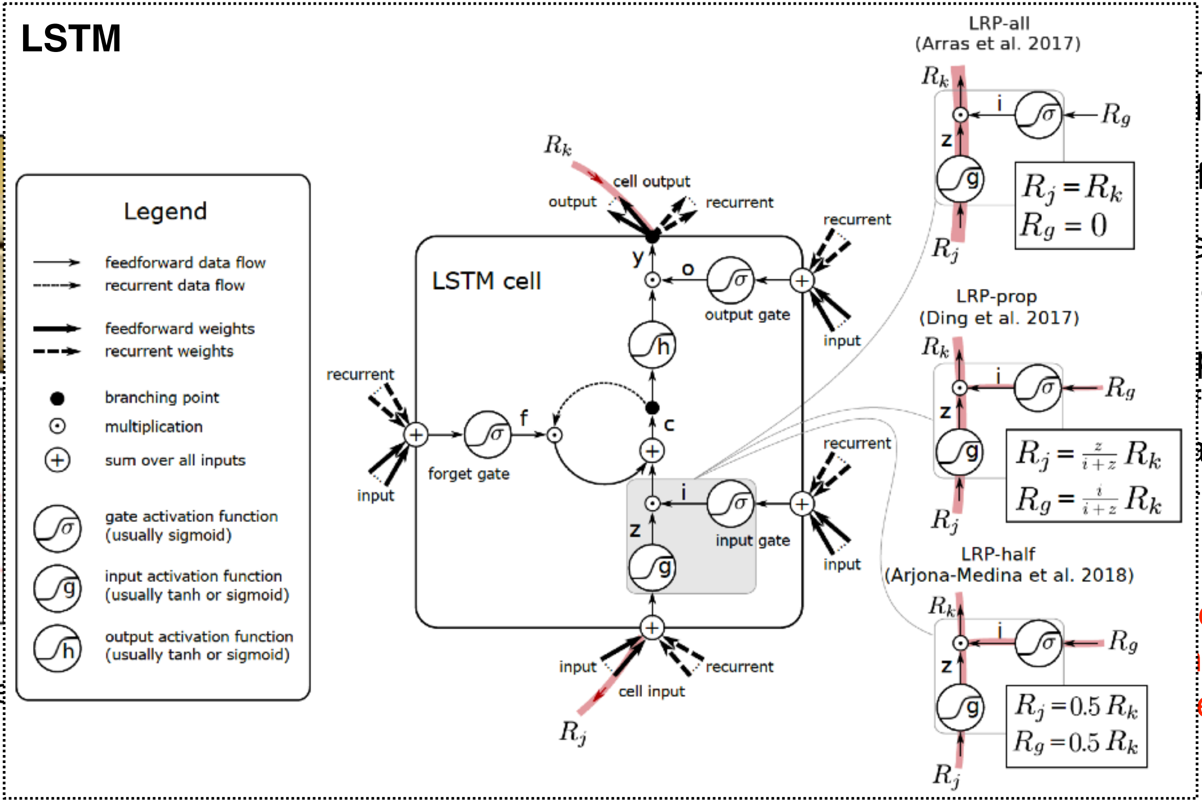
able to non-
tiable layers (no
t shattering)

edistribution rule
ht one (i.e. how
neasure zjk)?

(Montavon et al. 2017)

Layer-wise Relevance Propagation (LRP)





Layer-wise re

$\sum_i R_i = .$

lvantages

fficient & faithful

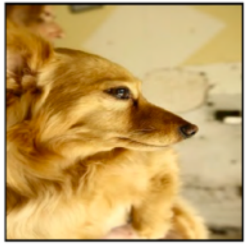
elevance values for
elements of NN

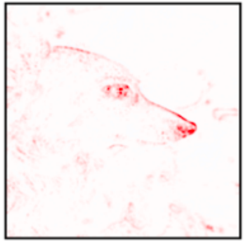
pplicable to non-
ferentiable layers (no
adient shattering)

ch redistribution rule
e right one (i.e. how
est measure zjk)?

(Arras et al. 2019)

Layer-wise Relevance Propagation (LRP)





Layer-wise rele

$\sum_i R_i = \ldots$

Transformers

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

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

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

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

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

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

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

Proposition 3.4 *Decomposing LayerNorm or R a Taylor decomposition (4) with reference poi bias or distributing the bias uniformly) yields the identity relevance propagation rule:*

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

tages

ent & faithful

ance values for
nents of NN

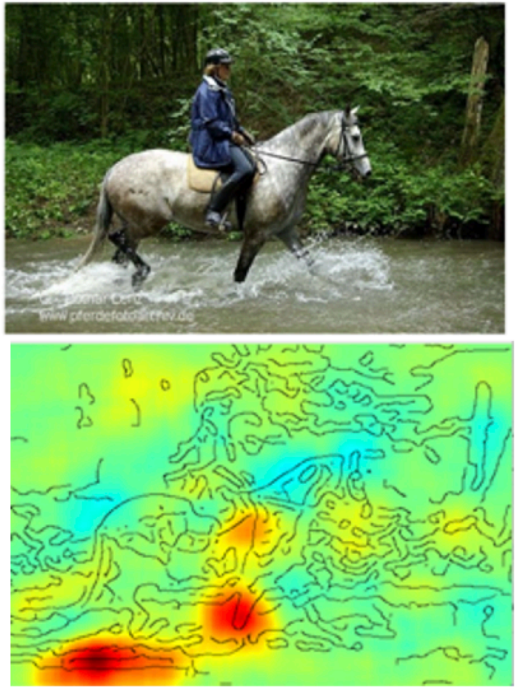
cable to non-
ntiable layers (no
nt shattering)

edistribution rule
ght one (i.e. how
measure zjk)?

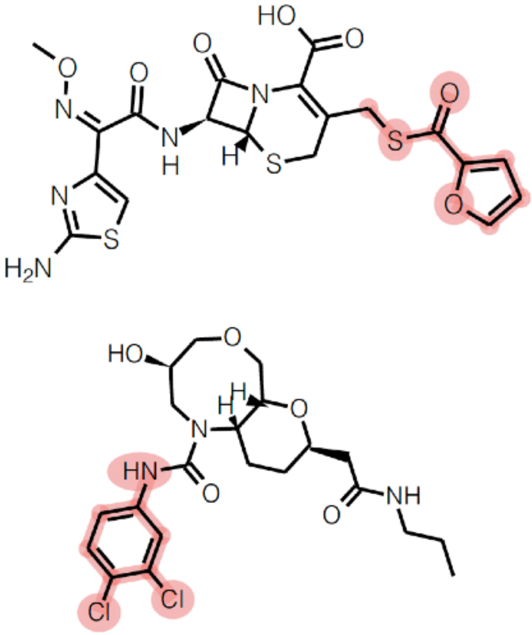
(Achtibat et al. 2024)

What Can We Do ?

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

Trust in LLMs
(Achtibat et al., ICML, 2024)

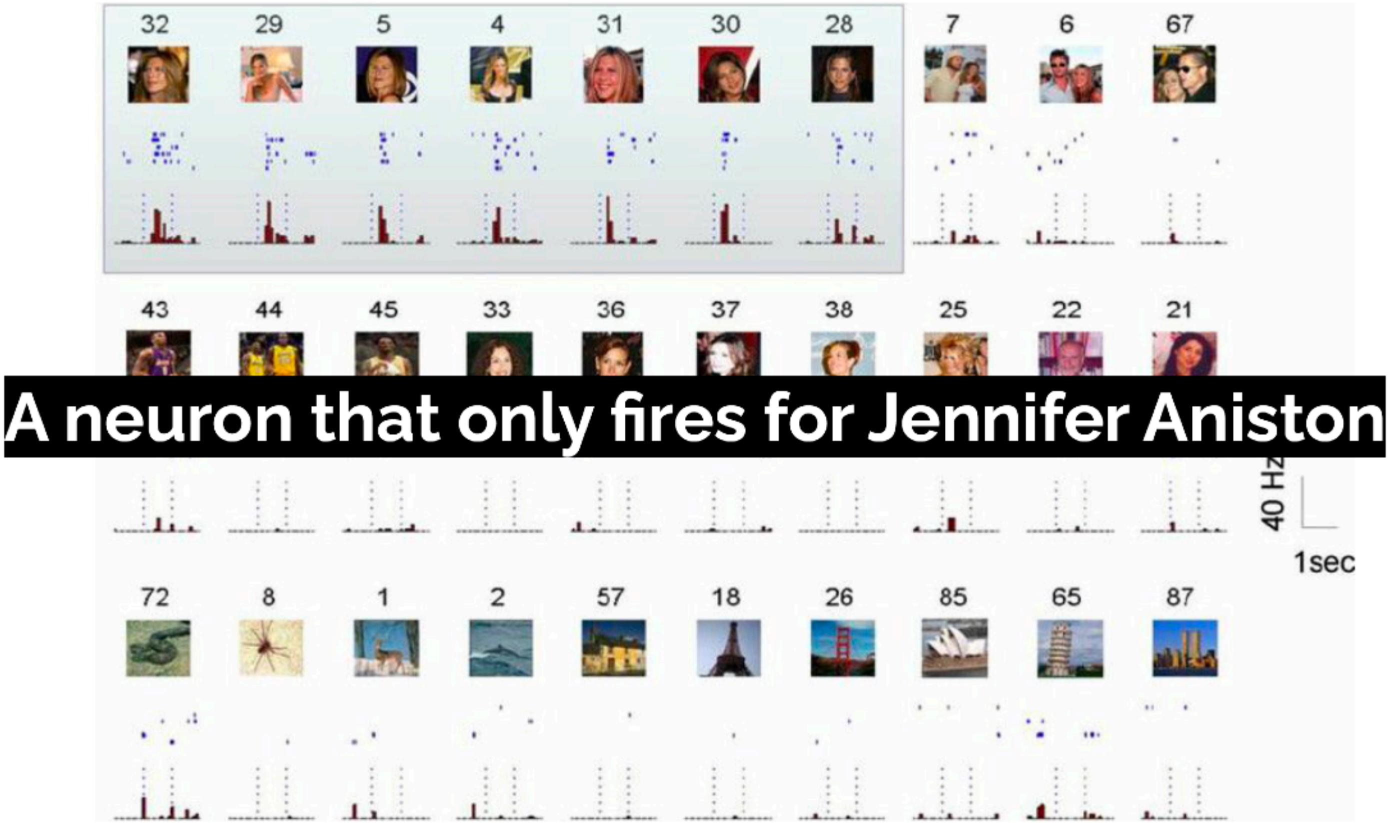
Question: In what country is Normandy located?
Answer: France

AttnLRP

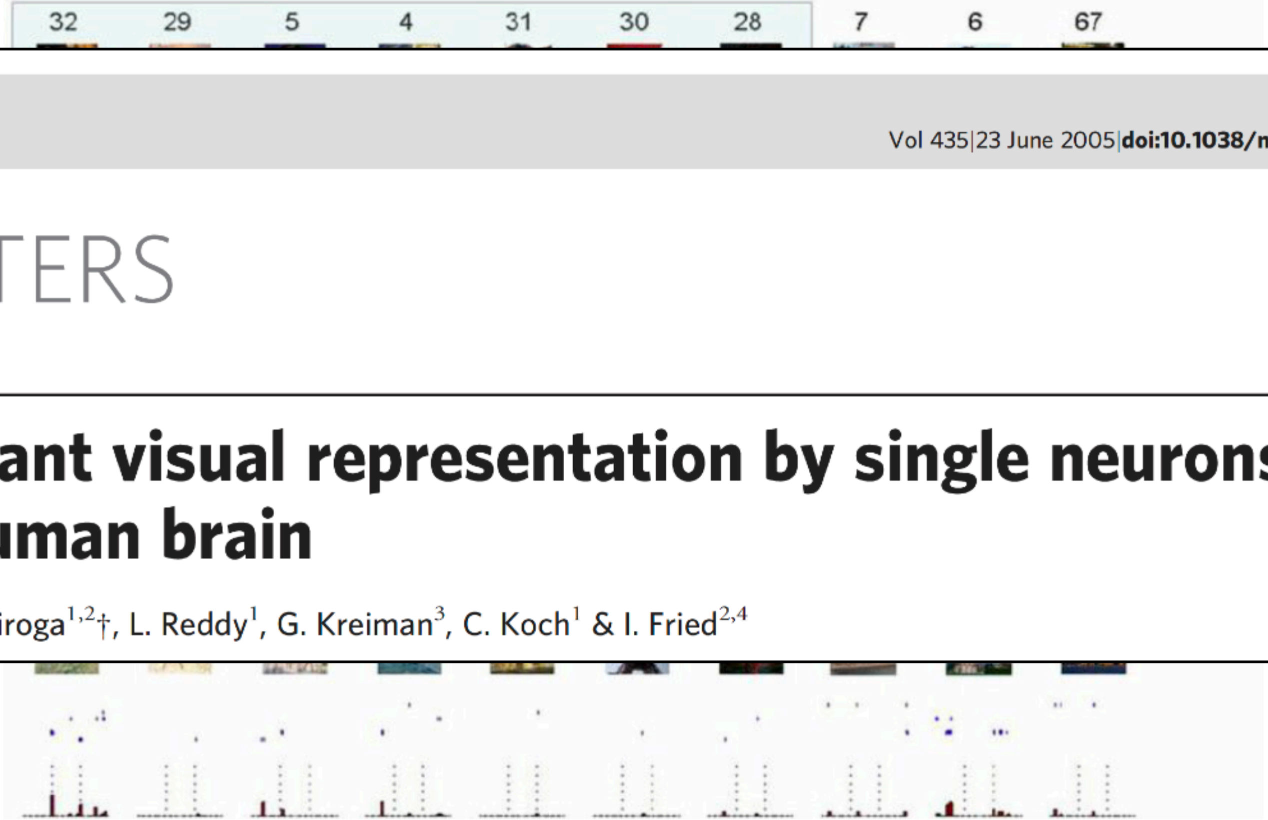
The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries gave their name to Normandy, a region in France. They were descended from Norse ("Norman" comes from "Norseman") raiders and pirates from Denmark, Iceland and Norway who, under their 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.

Second Wave of XAI: "Understand Model"

Interpreting the Model



Interpreting the Model



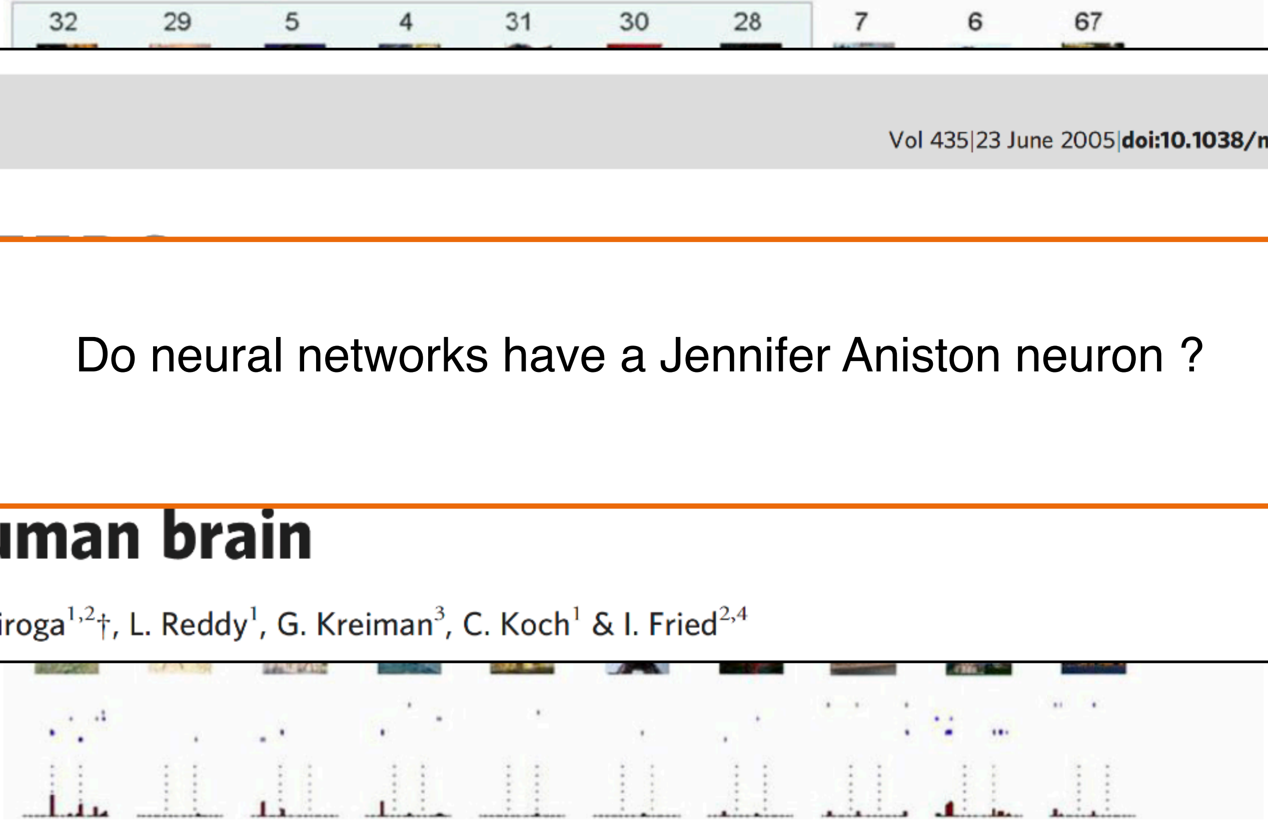
nature Vol 435|23 June 2005|doi:10.1038/nature03687

LETTERS

Invariant visual representation by single neurons in the human brain

R. Quian Quiroga[1,2]†, L. Reddy[1], G. Kreiman[3], C. Koch[1] & I. Fried[2,4]

Interpreting the Model



nature Vol 435|23 June 2005|doi:10.1038/nature03687

LET

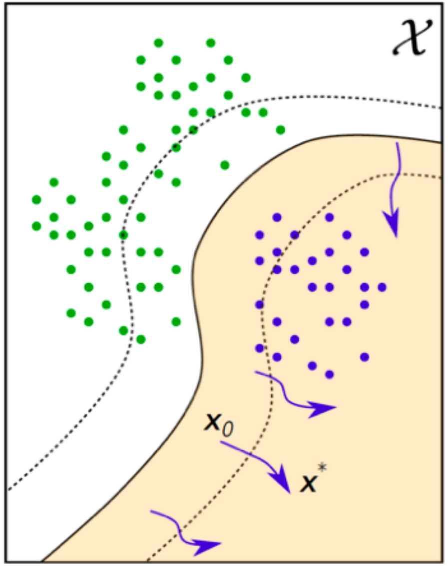
Invari ... in the human brain

R. Quian Quiroga[1,2]†, L. Reddy[1], G. Kreiman[3], C. Koch[1] & I. Fried[2,4]

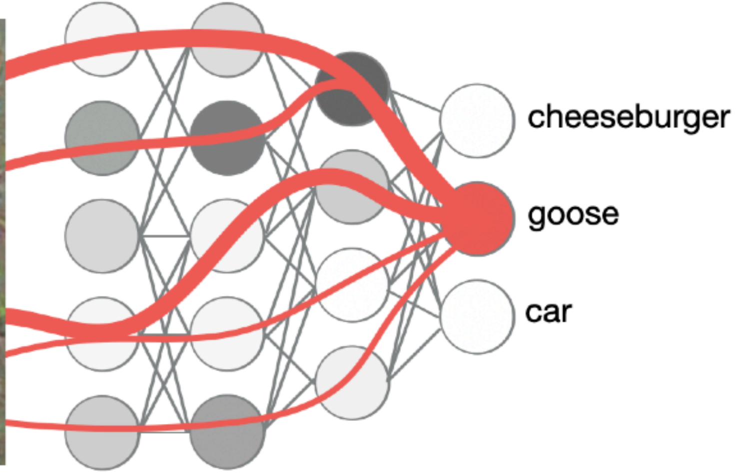
Do neural networks have a Jennifer Aniston neuron ?

Activation Maximization

Find the input pattern that maximizes class probability.



simple regularizer
(Simonyan et al. 2013)

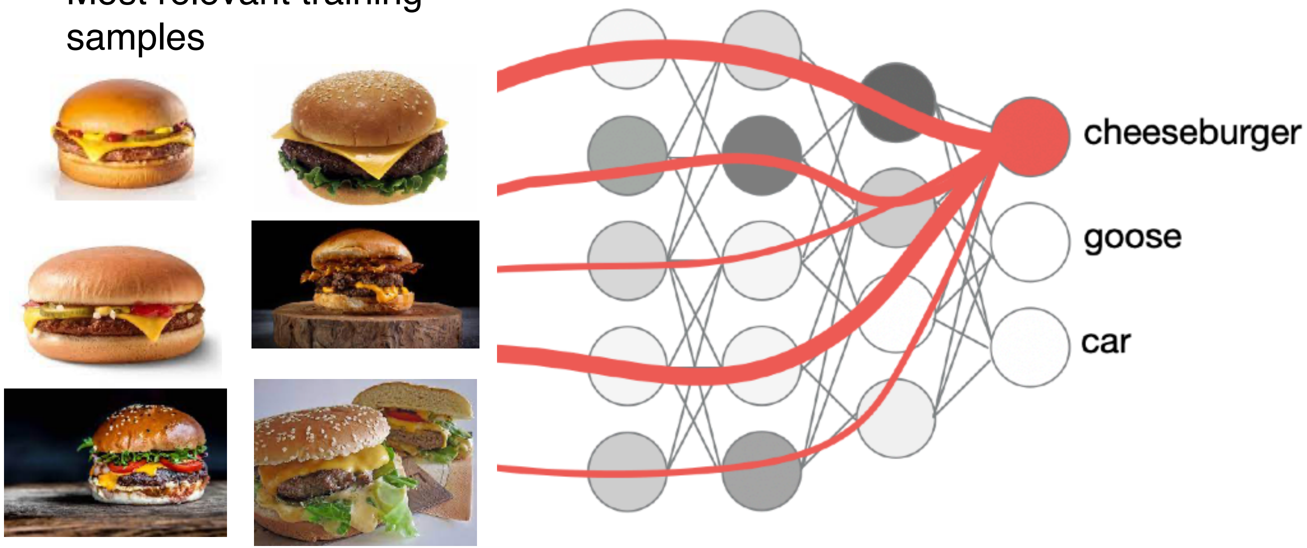


$$\max_{x \in \mathcal{X}} p_\theta(\omega_c \mid x) + \lambda \Omega(x)$$

Data-Based Activation Maximization

Find training samples, which maximially activate (output) neuron.

Most relevant training samples



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

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

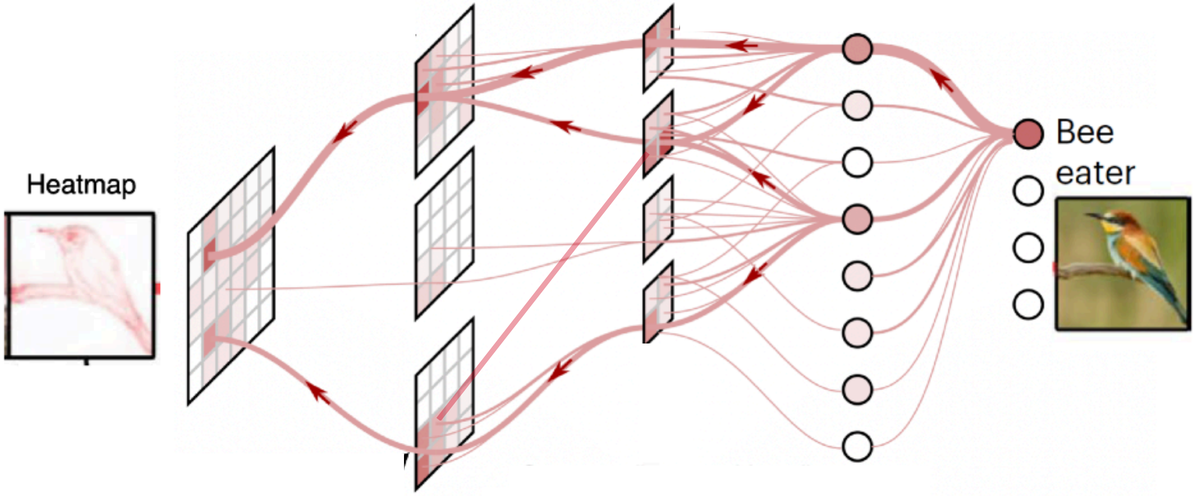
Check for updates

Reduan Achtibat [1,4], Maximilian Dreyer [1,4], Ilona Eisenbraun [1], Sebastian Bosse [1], Thomas Wiegand [1,2,3], Wojciech Samek [1,2,3] & Sebastian Lapuschkin [1]

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

Explainability 2.0: Where, What and How

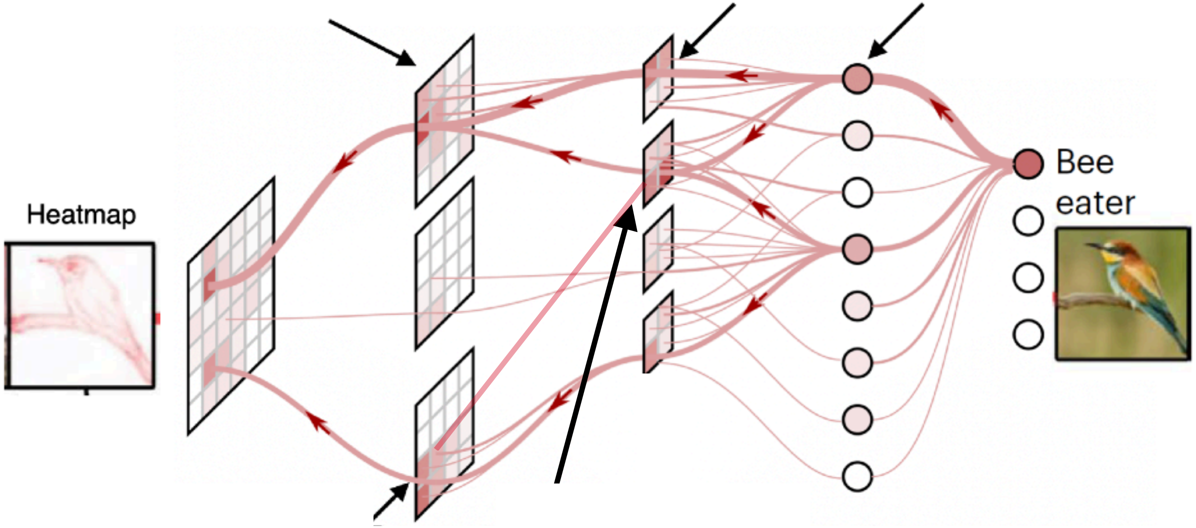
Known: Hidden layers encode semantic concepts.



Goal: Explain in terms of these concepts.

Explainability 2.0: Where, What and How

Known: Hidden layers encode semantic concepts.

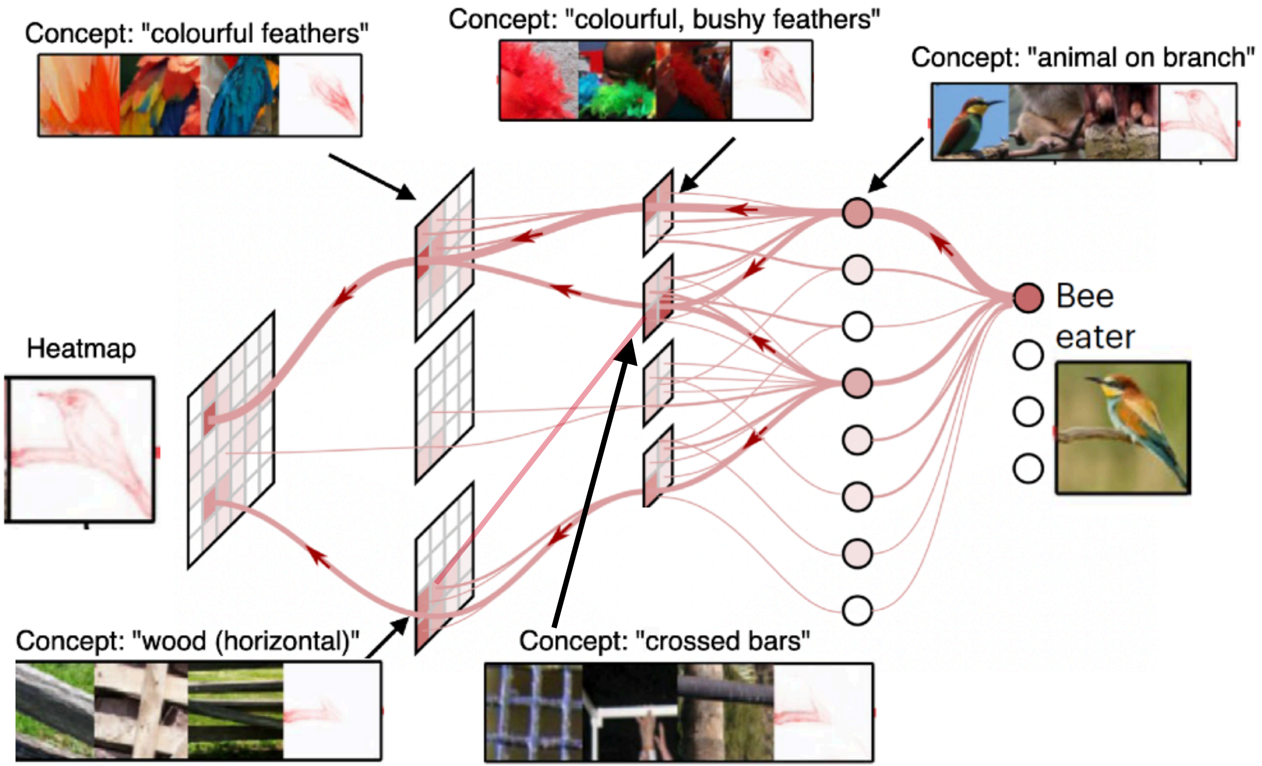


Which neurons are relevant ?

—> LRP

Explainability 2.0: Where, What and How

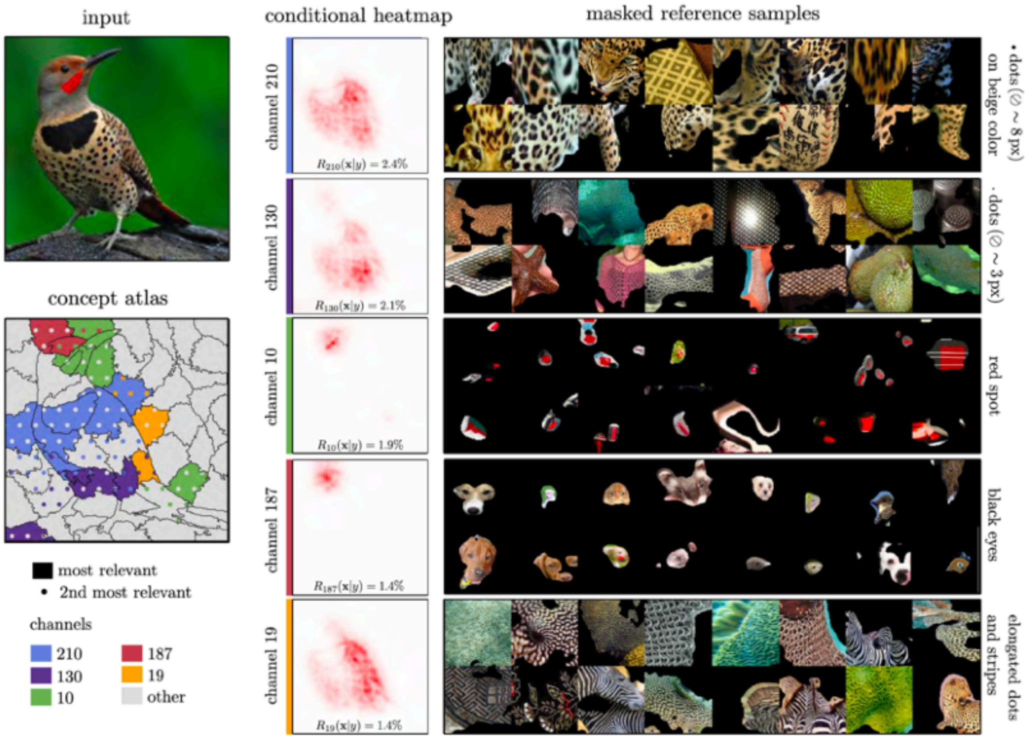
Known: Hidden layers encode semantic concepts.

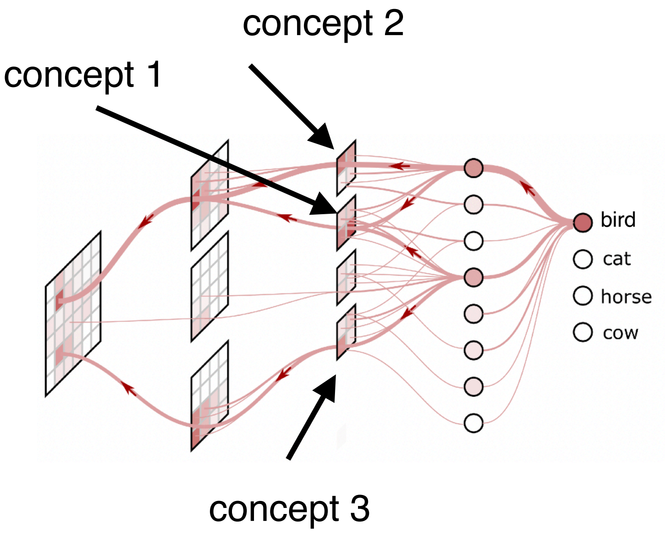


Which neurons are relevant ?
—> LRP

What are they encoding ?
—> Activation Maximization

Concept Relevance Propagation (CRP)



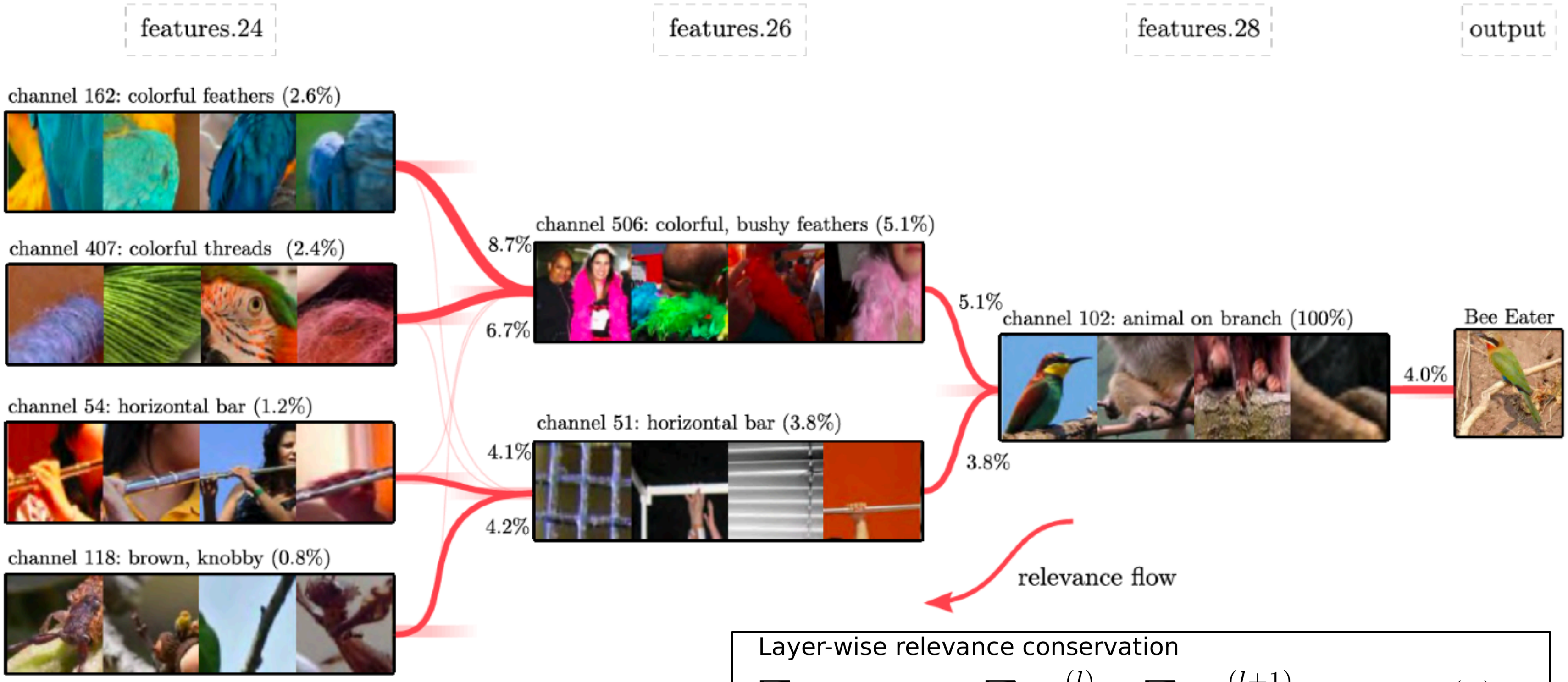


Step 1: Find relevant concepts

Step 2: Compute conditional explanation (*where*)

Step 3: Visualize relevant samples (*what*)

Concept Composition



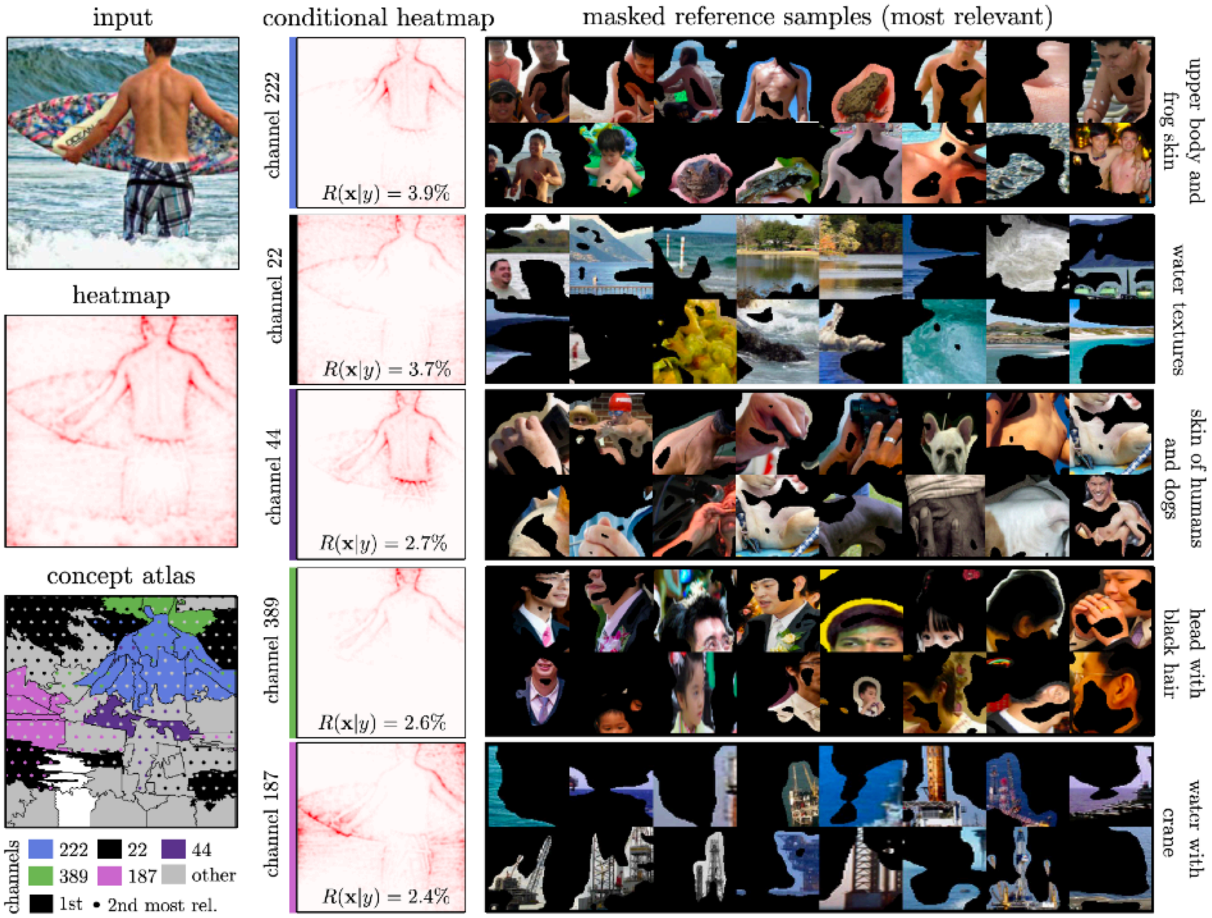
Layer-wise relevance conservation

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

Identifying Clever Hans

Prediction: swimming trunk

Relevant concepts: skin, body, hair, water



Third Wave of XAI: "Understand Everything"

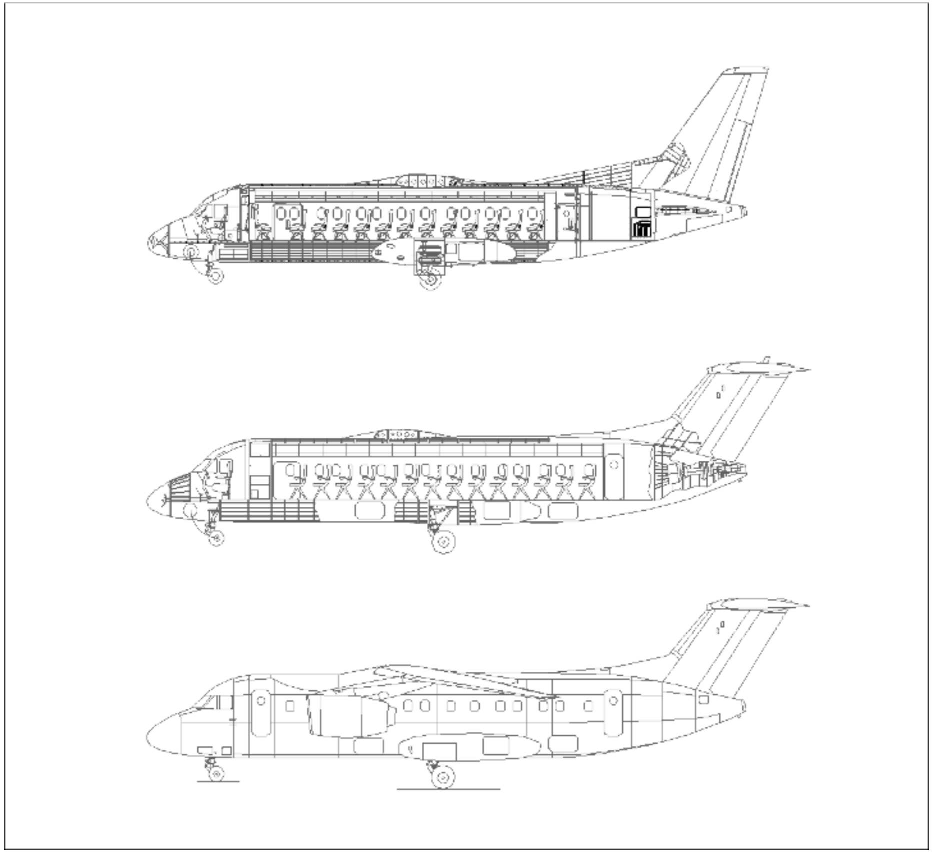
SemanticLens

Mechanistic understanding and validation of large AI models with SemanticLens

Maximilian Dreyer[1*] **Jim Berend**[1*] **Tobias Labarta**[1] **Johanna Vielhaben**[1]
Thomas Wiegand[1,2,3] **Sebastian Lapuschkin**[1] **Wojciech Samek**[1,2,3]
[1]Fraunhofer Heinrich Hertz Institute [2]Technische Universität Berlin
[3]BIFOLD – Berlin Institute for the Foundations of Learning and Data
{wojciech.samek,sebastian.lapuschkin}@hhi.fraunhofer.de

https://arxiv.org/pdf/2501.05398

Technical systems designed by humans



Technical systems designed by humans

- constructed step by step

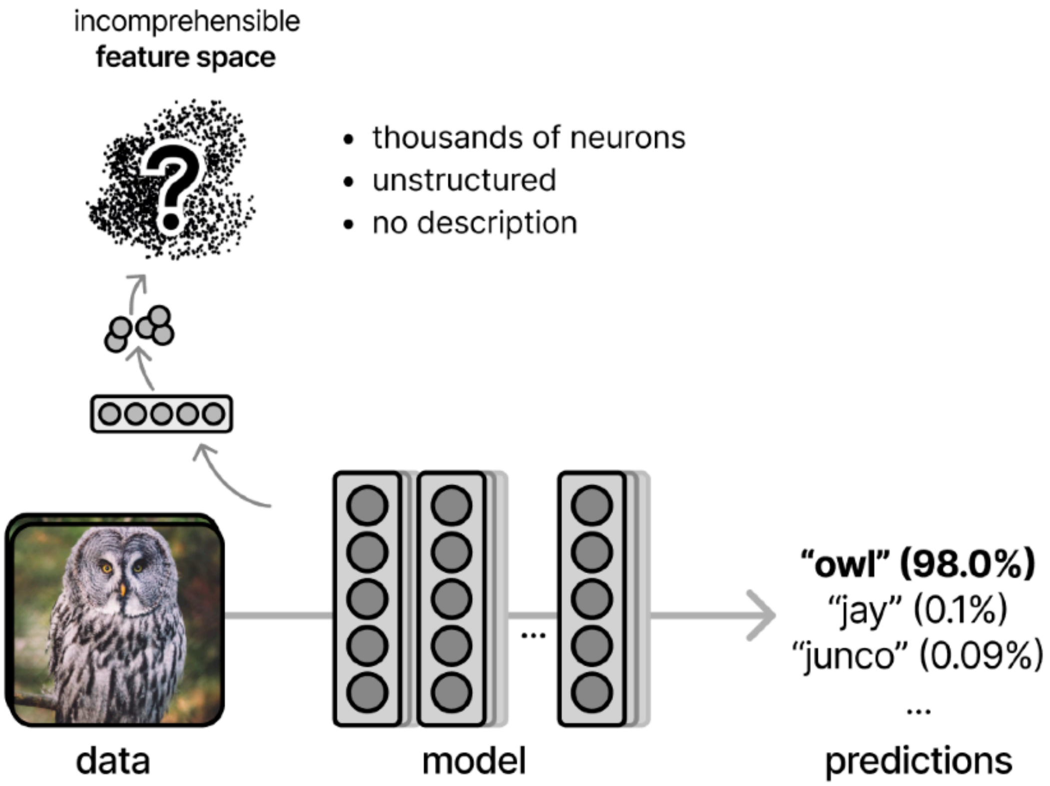
- modular

- each component serving a specific, well-understood function

- can be validated and certified

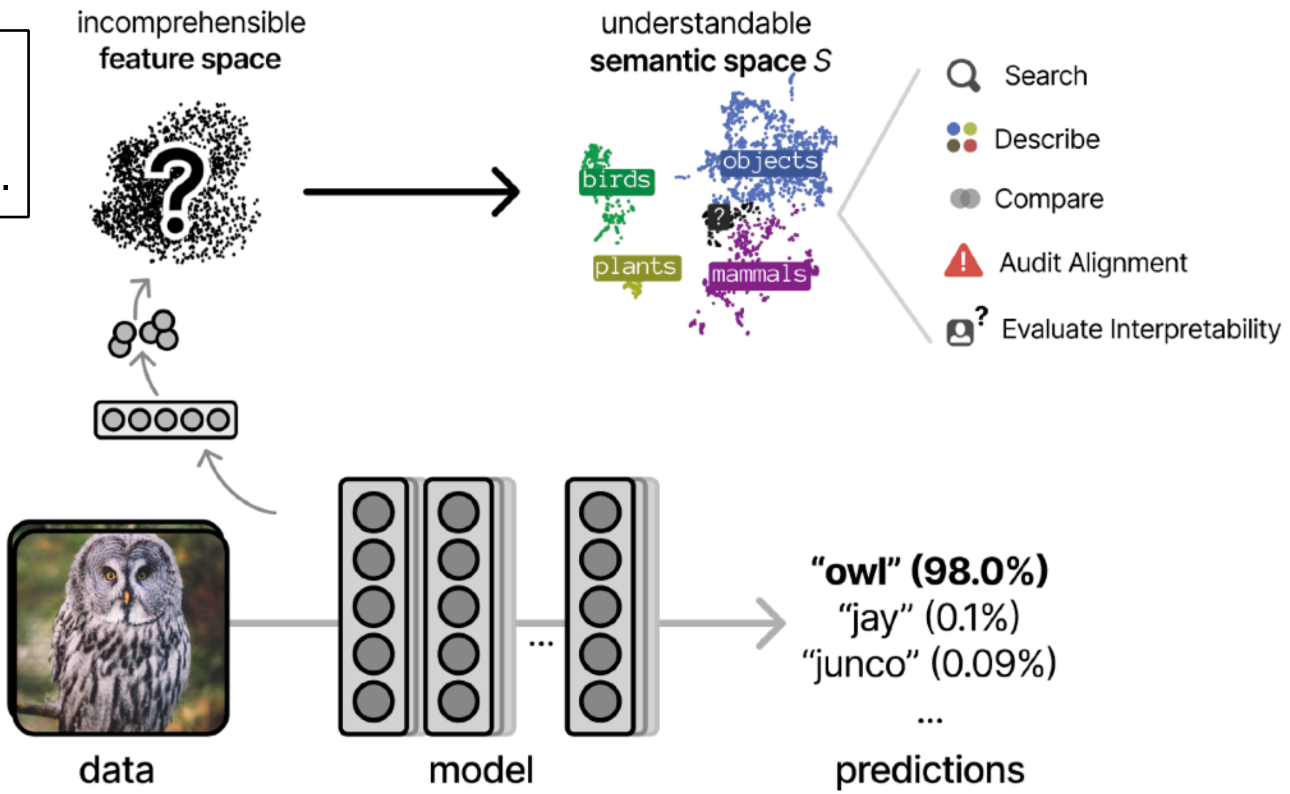
...

What Happens Inside the Model?



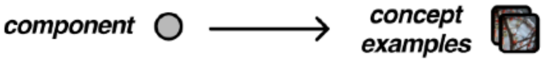
What Happens Inside the Model?

Idea: Transform model into comprehensible form.



SemanticLens

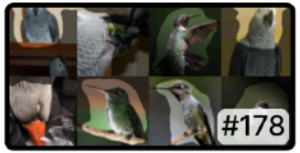
❶ Describing the Role of Components

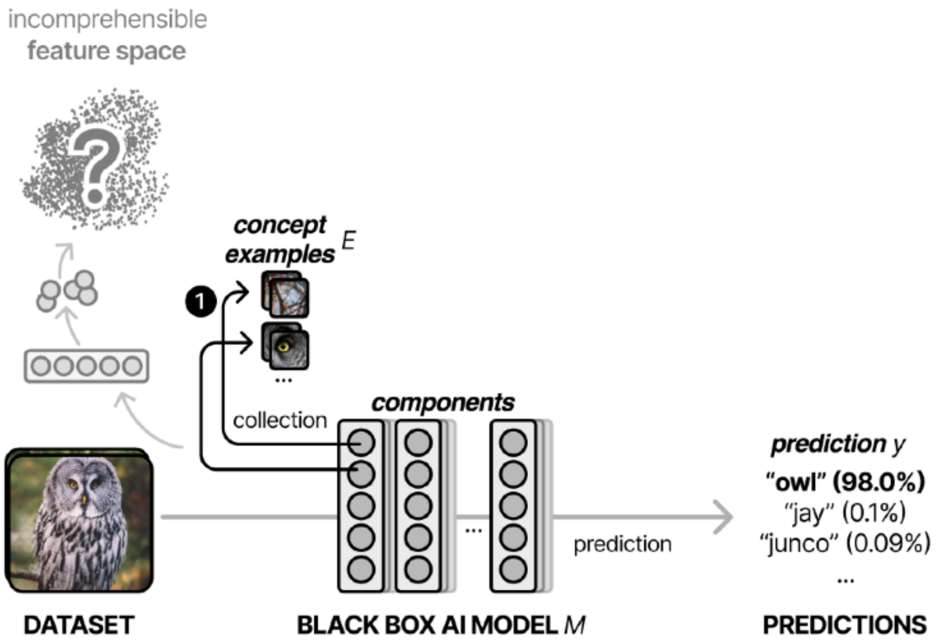


By collecting highly activating data samples + *CRP*.

concept examples

concept examples





SemanticLens

Model

Data

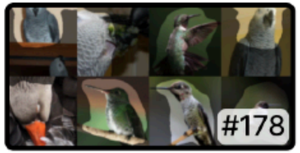
❶ Describing the Role of Components

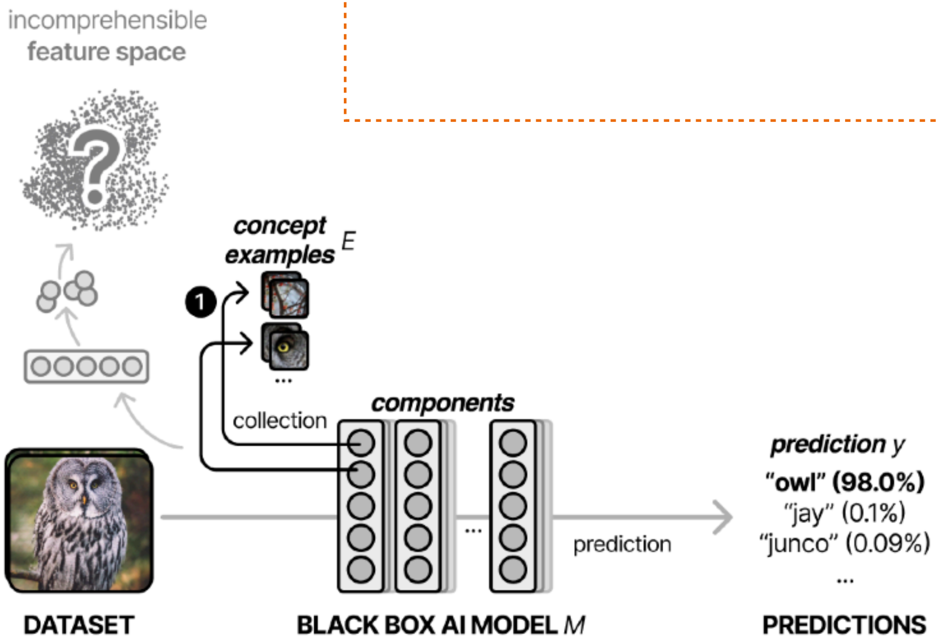
component ○ ⟶ *concept examples*

By collecting highly activating data samples + *CRP*.

concept examples

concept examples

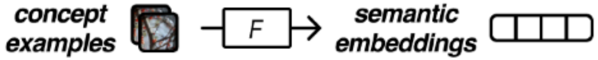




SemanticLens

❶ Describing the Role of Components

❷ Semantic Embedding



Via Foundation Model as *Semantic Domain Expert*:

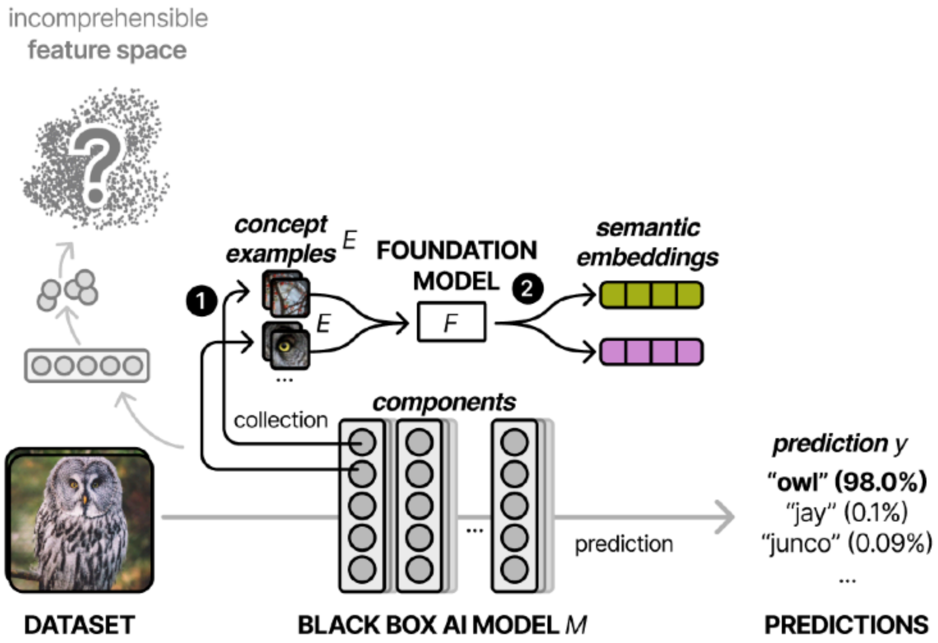
General Knowledge:

- CLIP
- Florence
- ...

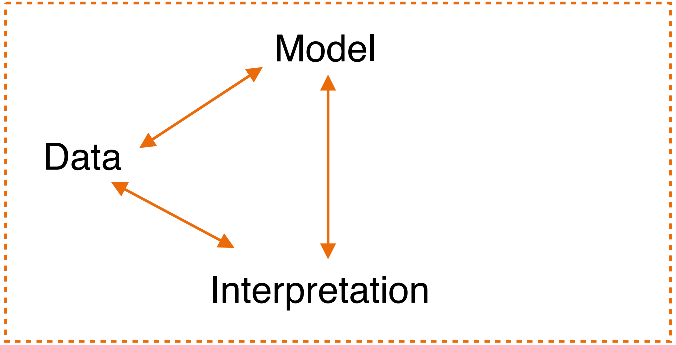
Medical Domain:

- WhyLesion-CLIP
- CXR-CLIP
- ...

No need for human in the loop anymore

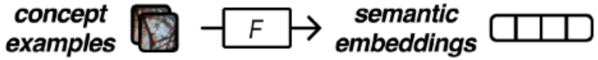


SemanticLens



❶ Describing the Role of Components

❷ Semantic Embedding

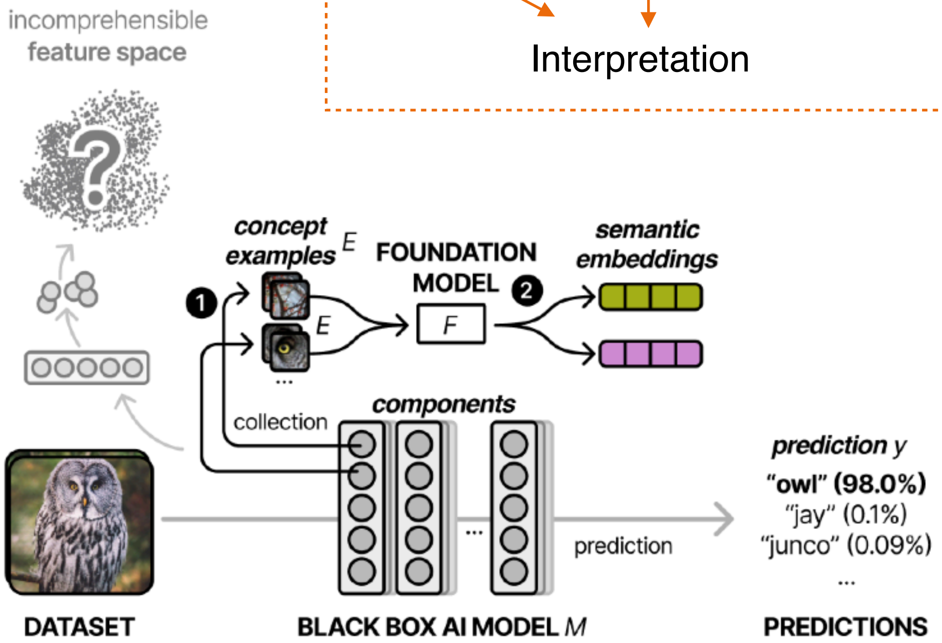


Via Foundation Model as *Semantic Domain Expert*:

General Knowledge:
- CLIP
- Florence
- ...

Medical Domain:
- WhyLesion-CLIP
- CXR-CLIP
- ...

No need for human in the loop anymore



SemanticLens

1 Describing the Role of Components

component → *concept examples*

2 Semantic Embedding

concept examples → F → *semantic embeddings*

3 Connect with Concept Relevance

semantic embeddings <-> *relevance scores* R

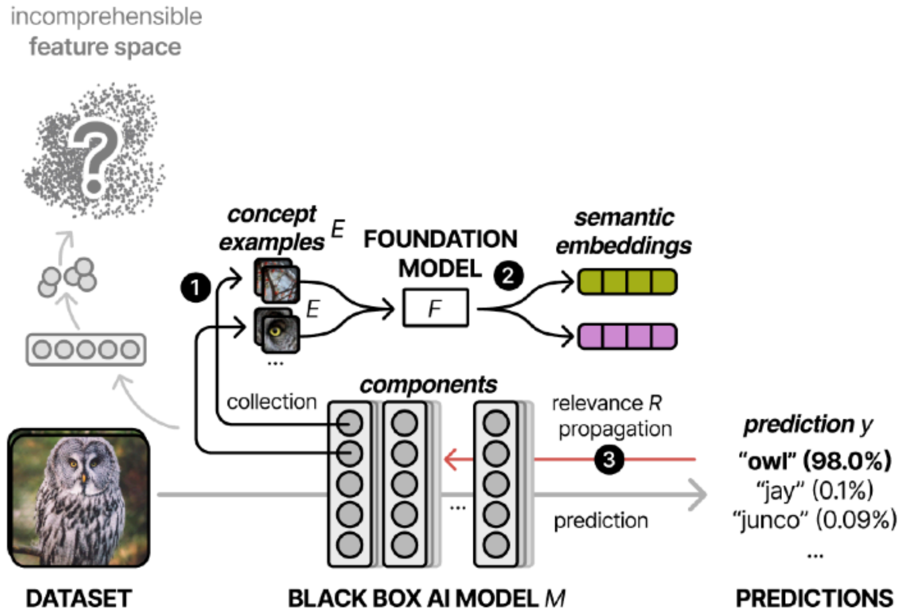
Retrieve component-level relevance scores with *CRP* for:

- output predictions

≈ ← "owl" (98.0%)

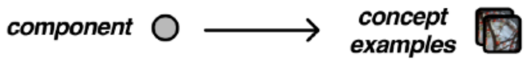
- upper-level components

≈ ← ≈

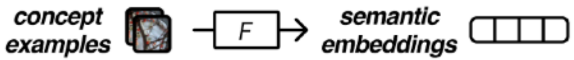


SemanticLens

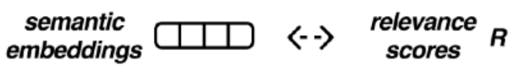
❶ Describing the Role of Components



❷ Semantic Embedding



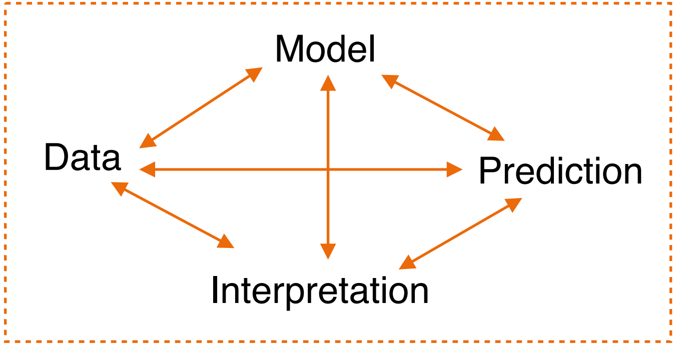
❸ Connect with Concept Relevance

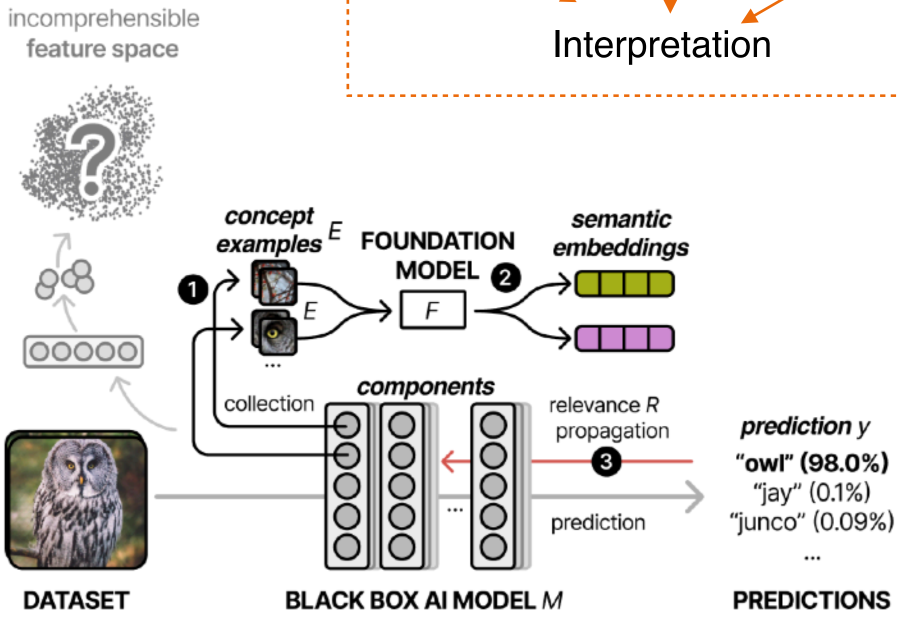


Retrieve component-level relevance scores with *CRP* for:

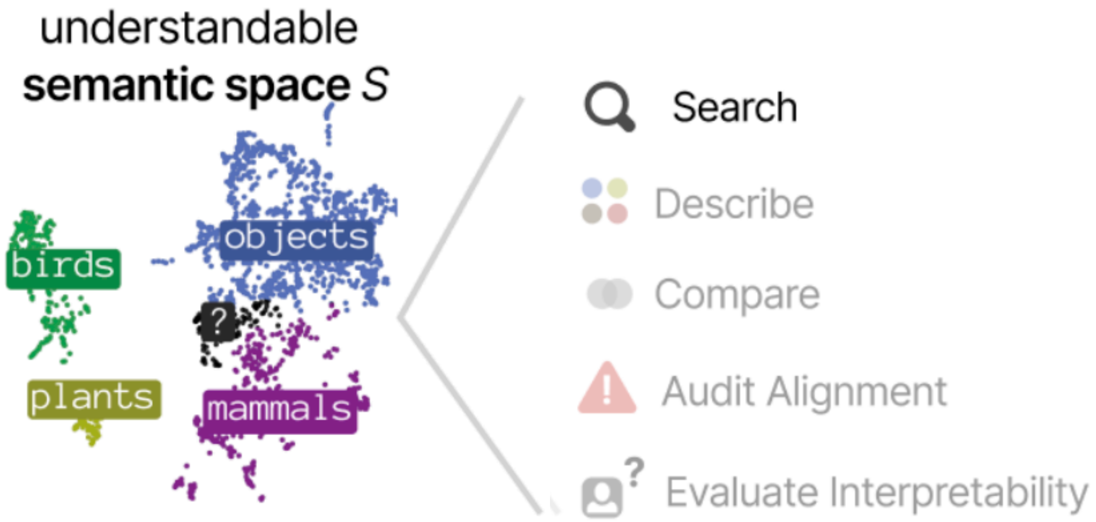
- output predictions

- upper-level components

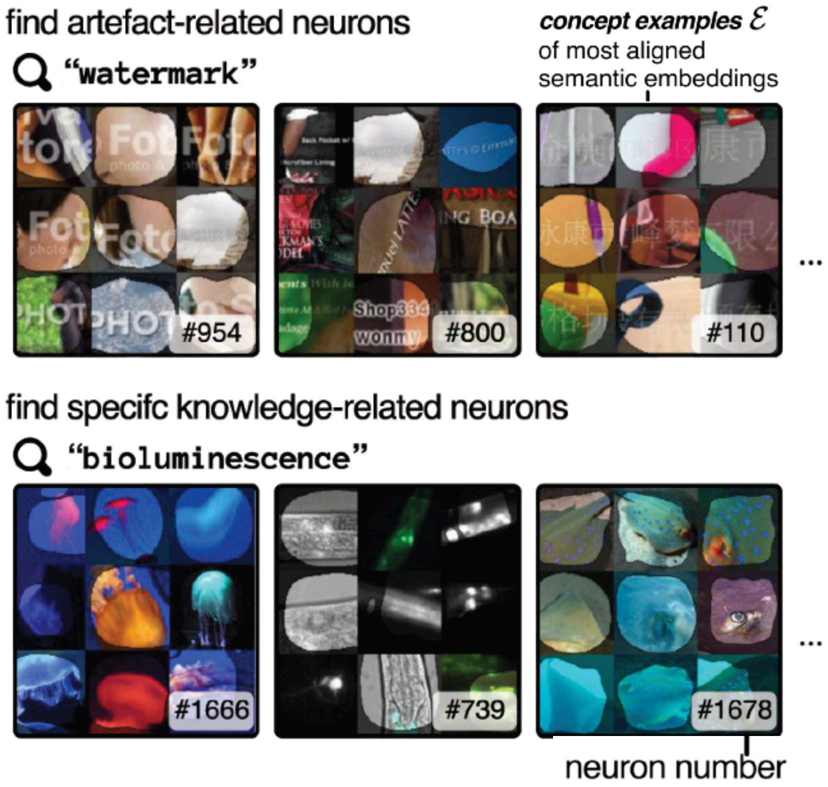




SemanticLens: What Can We do ?

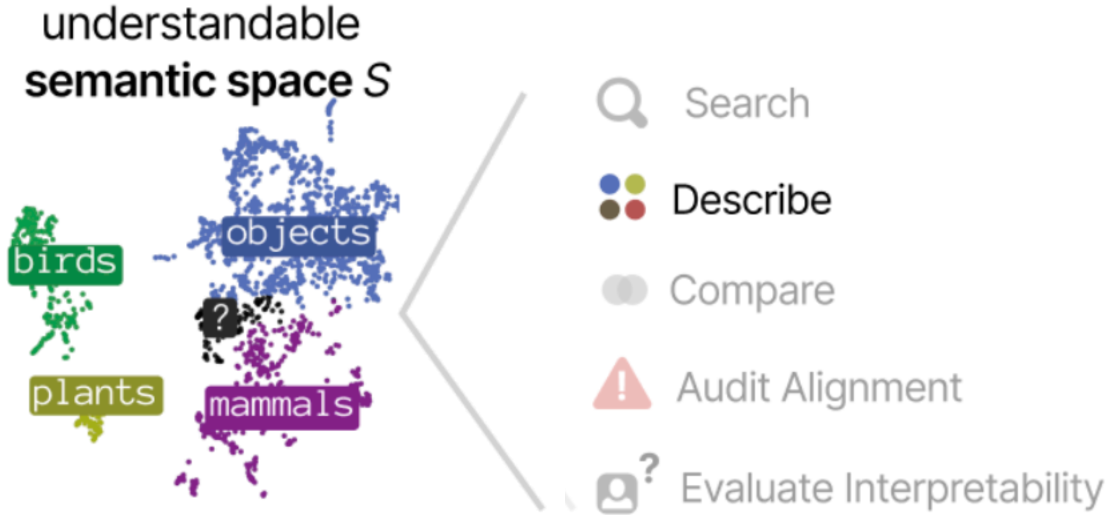


Search: Finding the Needle in the Haystack

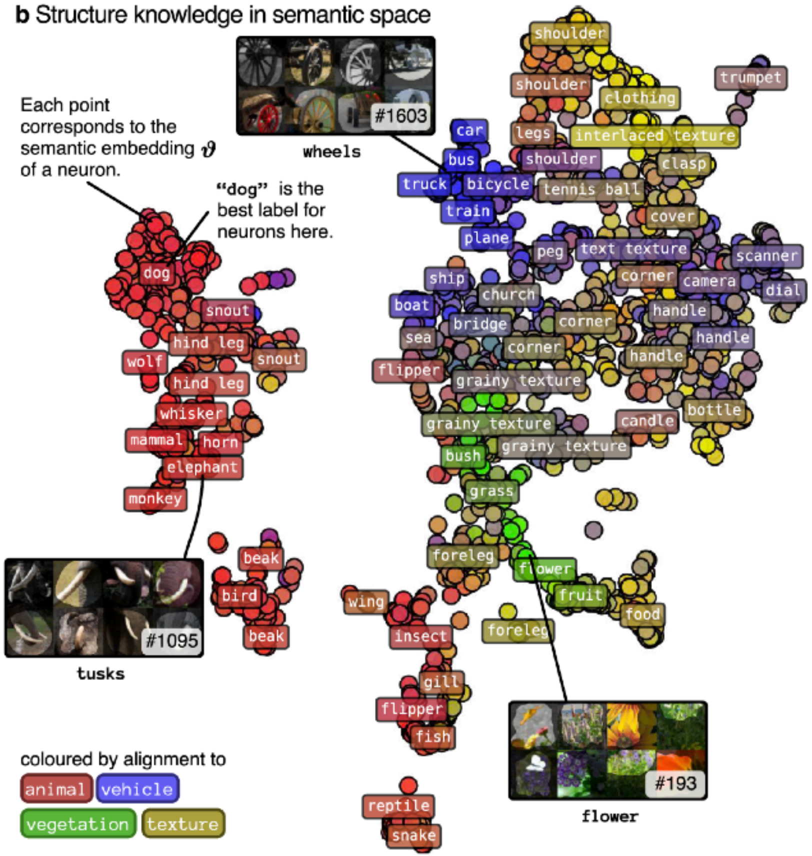


<u>Note</u>: CLIP models allow to measure similarity between image embeddings (here: neuron) and text embeddings (here: query).

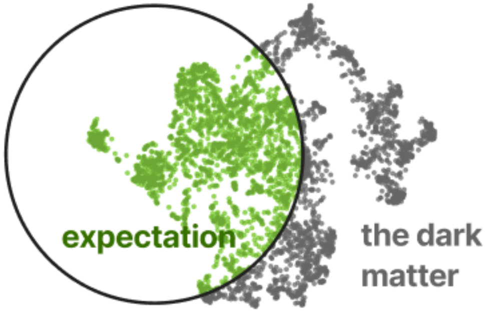
SemanticLens: What Can We do ?



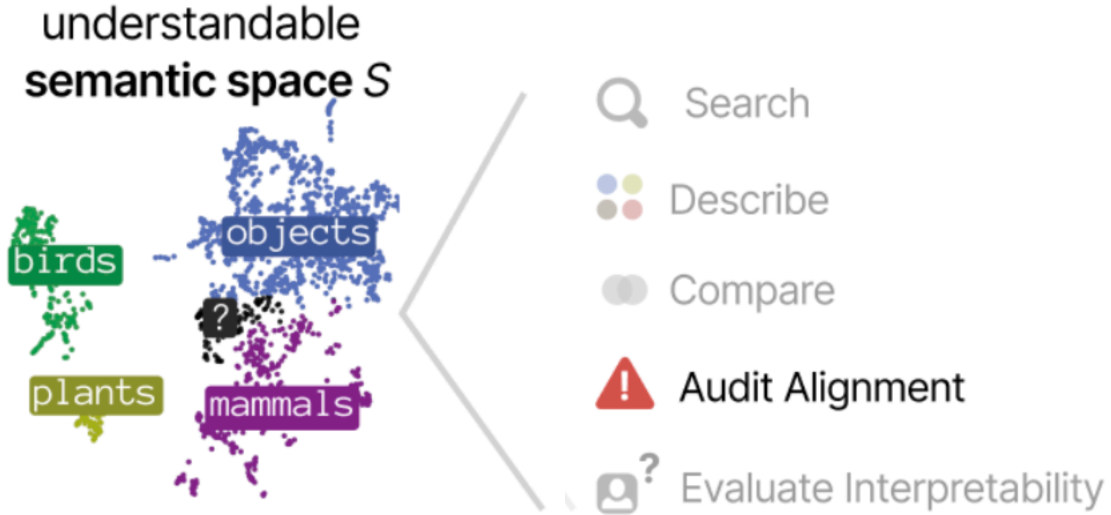
Describe: What Knowledge (does not) Exists ?



Also here we measure similarity between image embeddings (here: neuron) and text embeddings (here: label from a vocabulary of labels).



SemanticLens: What Can We do ?



A Tool for Auditing

1 define concepts for detecting Ox

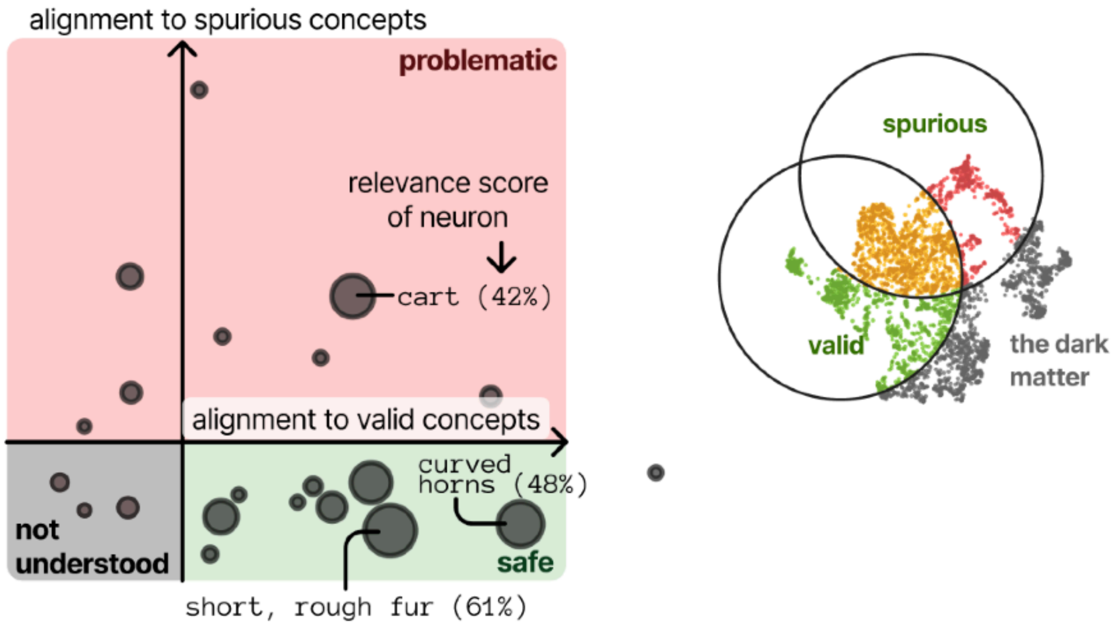
valid concepts:

large muscular body
curved horns
hooves
thick neck
short, rough fur
soft fur
long fur
brown coat
black coat
white coat
strong legs
long tail
wide muzzle

spurious concepts:

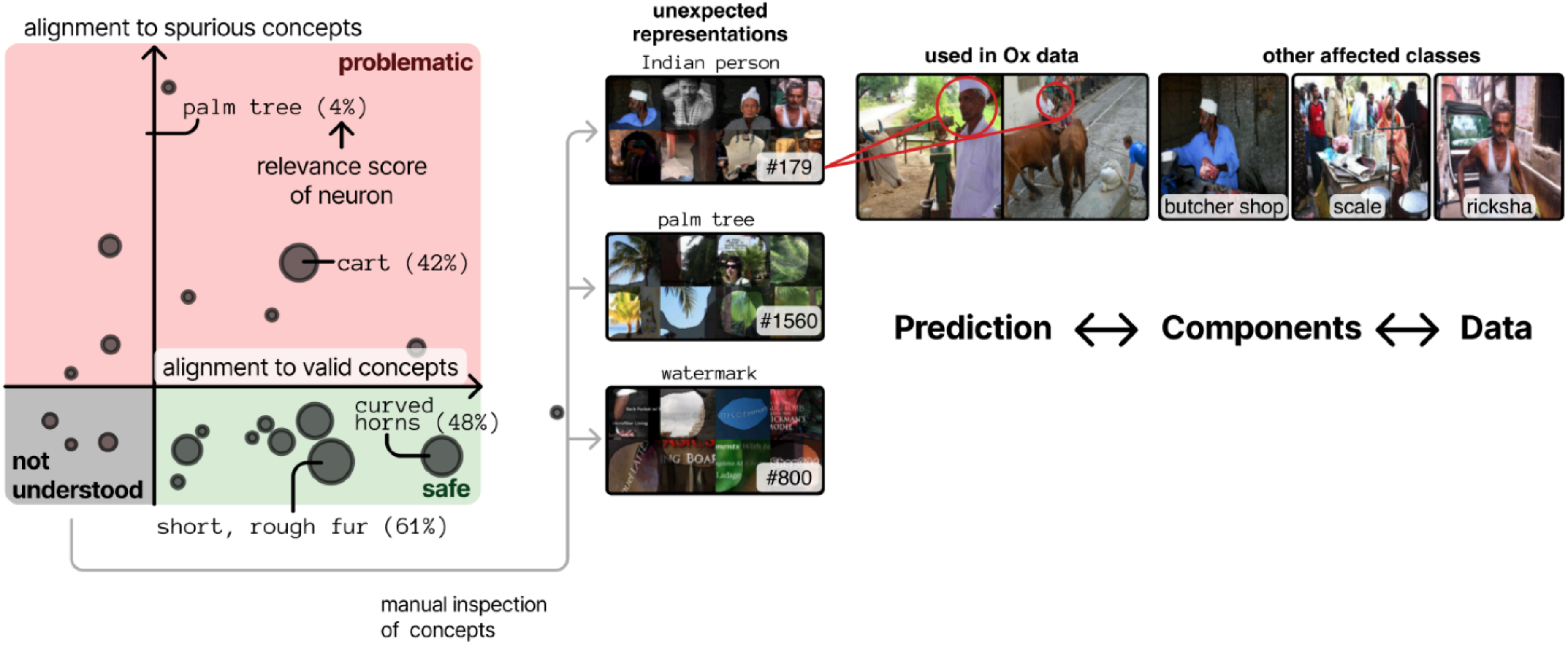
grassland
sky
tree
water
grain, straw
cart
wheel
mud, dirt
person
wooden

2 evaluate alignment of neurons



Note: Size of circle shows the “relevance” of this concept.

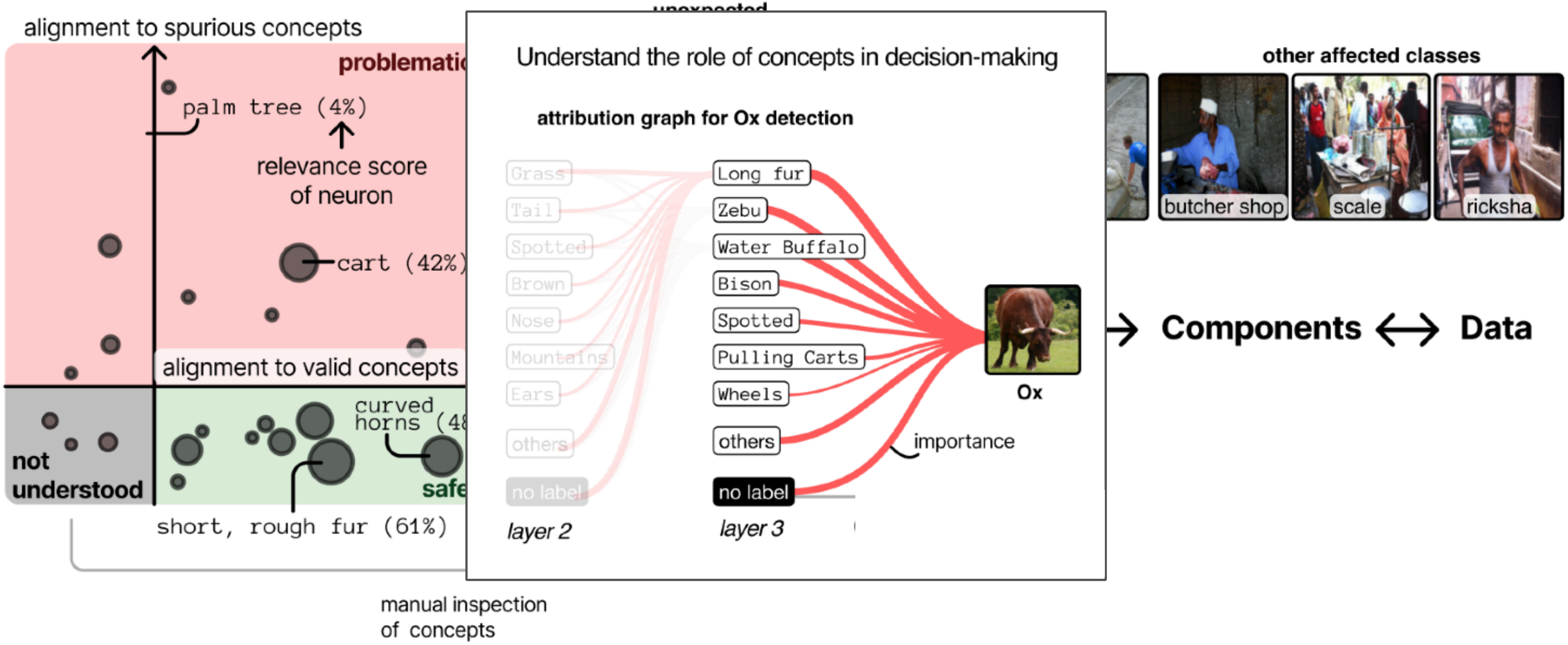
A Tool for Auditing



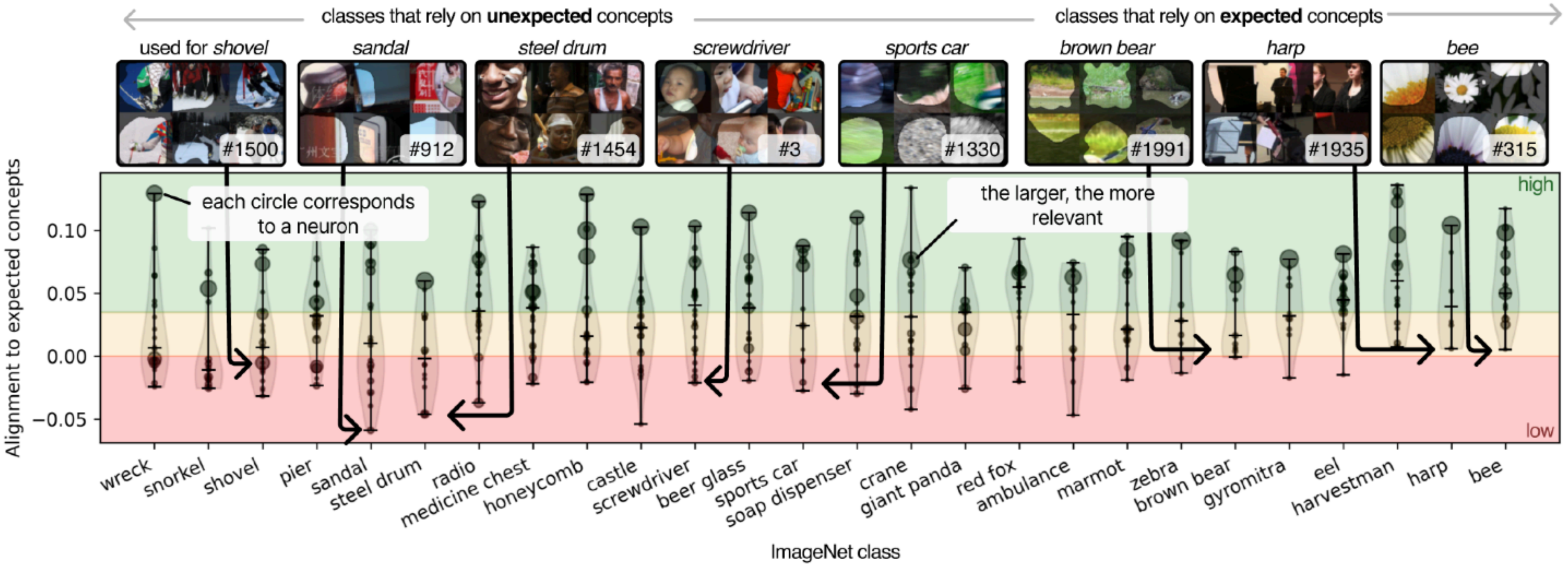
Prediction ⟷ Components ⟷ Data

For concepts which we do not understand (i.e., dark matter) we can go back to data for manual inspection.

A Tool for Auditing

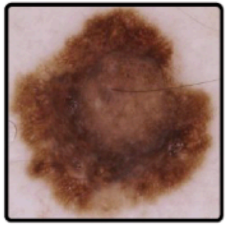


A Tool for Auditing

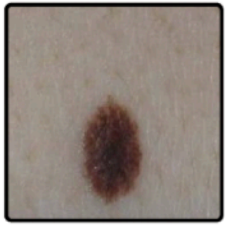


Audit Alignment: Medical Case

a Defining concepts for melanoma detection: ABCDE rule



Melanoma



Other (Regular)

valid concepts:

A. Asymmetry
- asymmetric lesion, ...

B. Border
- ragged border, ...

C. Color
- blue-white veil, ...

D. Diameter
- large lesion, ...

E. Evolving
- crusty surface, ...

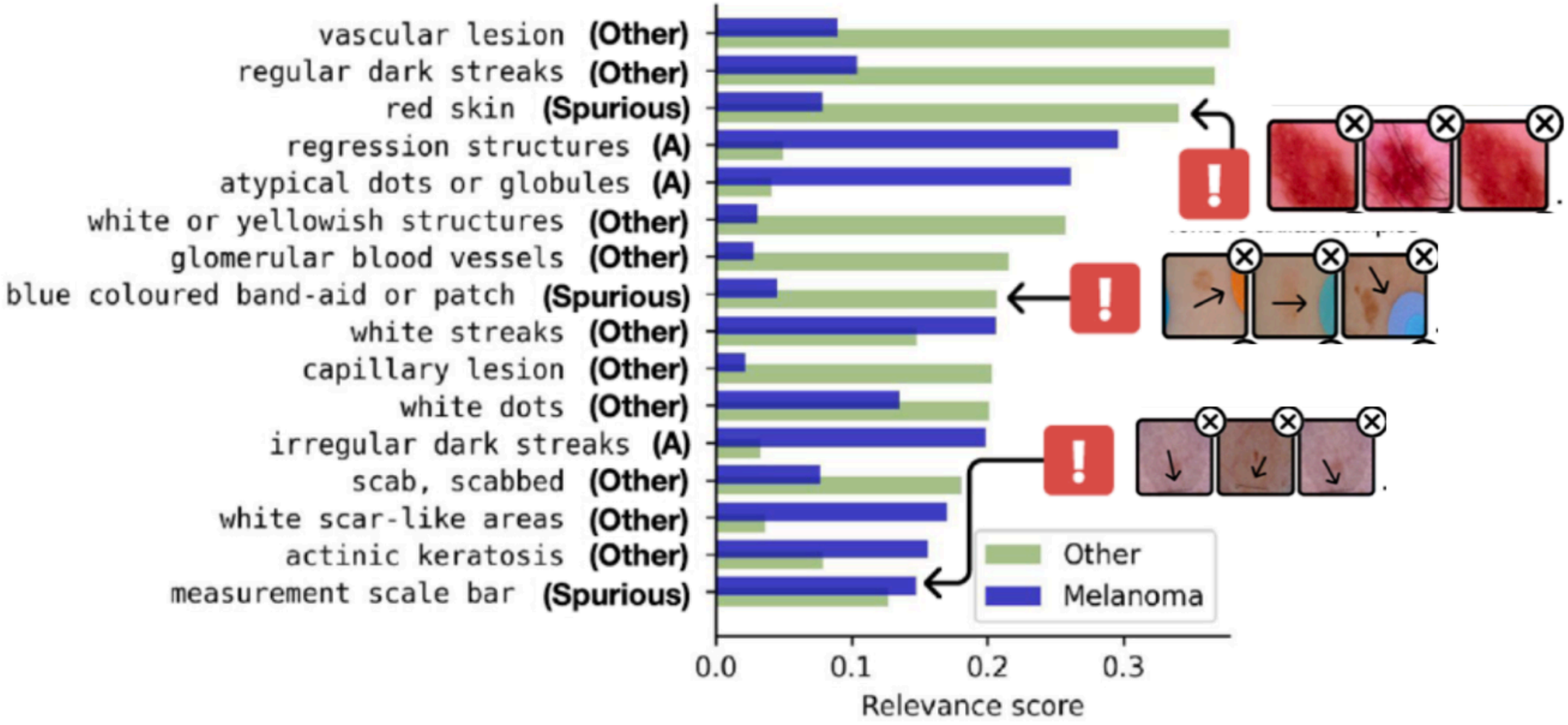
Other (Regular)
- even border, ...

Other (Irregular)
- white or yellowish structures, ...

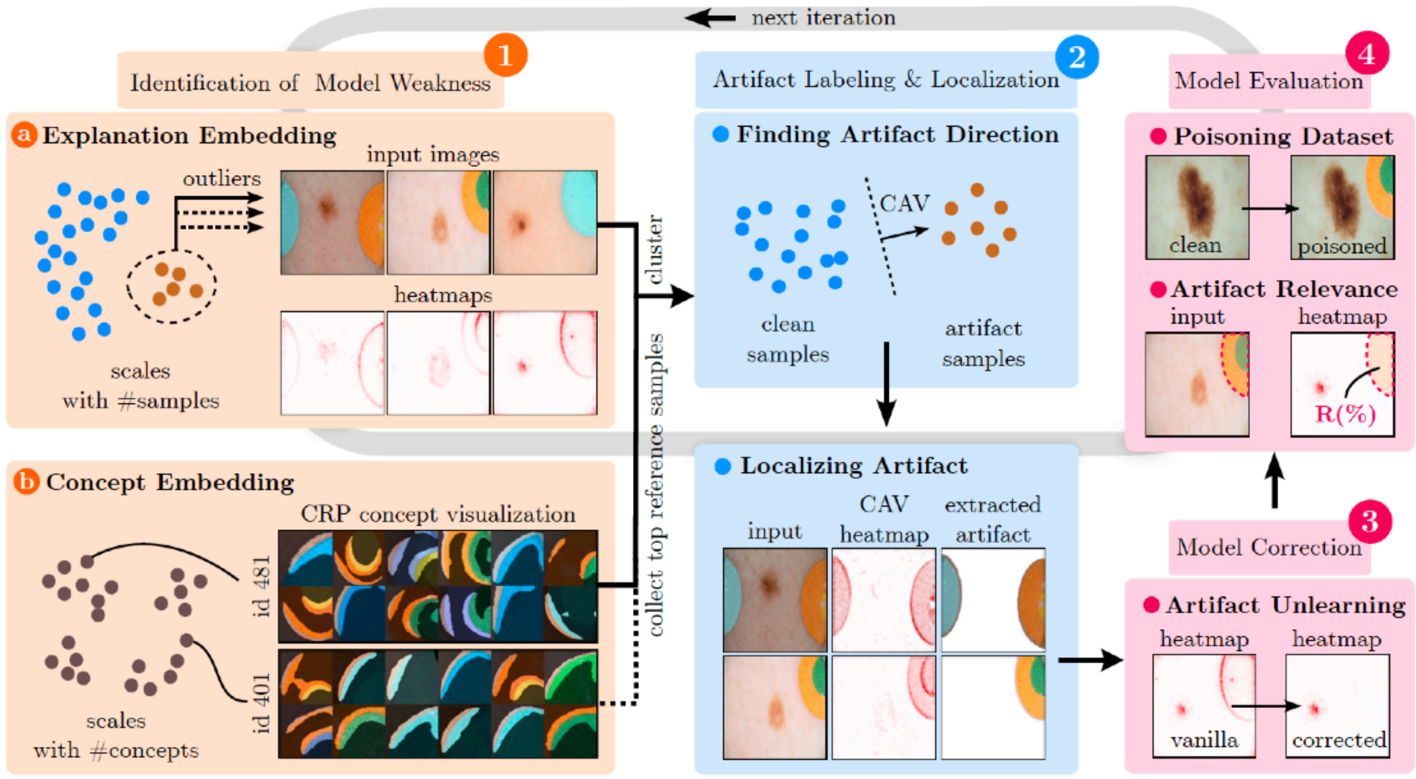
spurious concepts:

hairs
hairy
band-aid
blue-coloured band-aid
red skin
measurement scale bar
ruler
vignetting
purple ink
skin marker
...

Audit Alignment: Medical Case

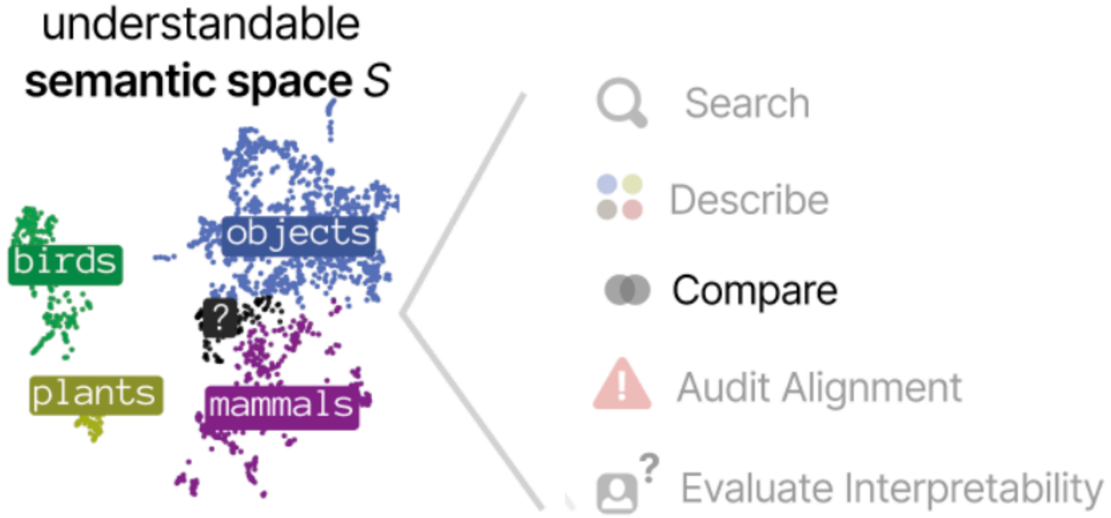


From Inspecting to Debugging



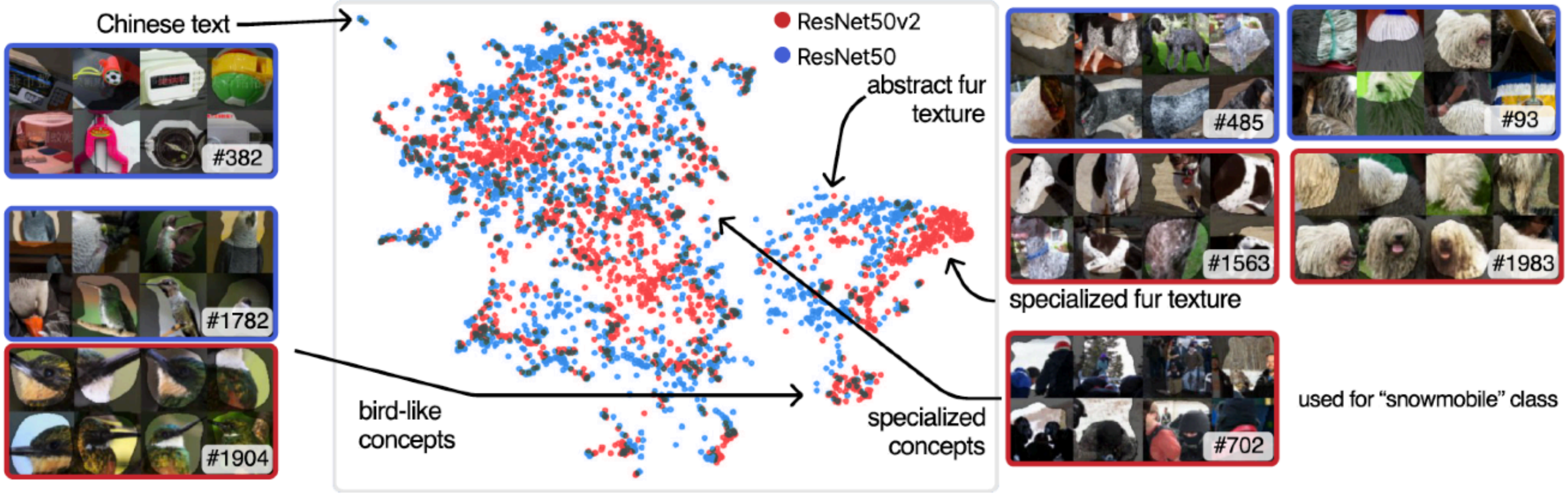
[Pahde et al. 2023]

SemanticLens: What Can We do ?



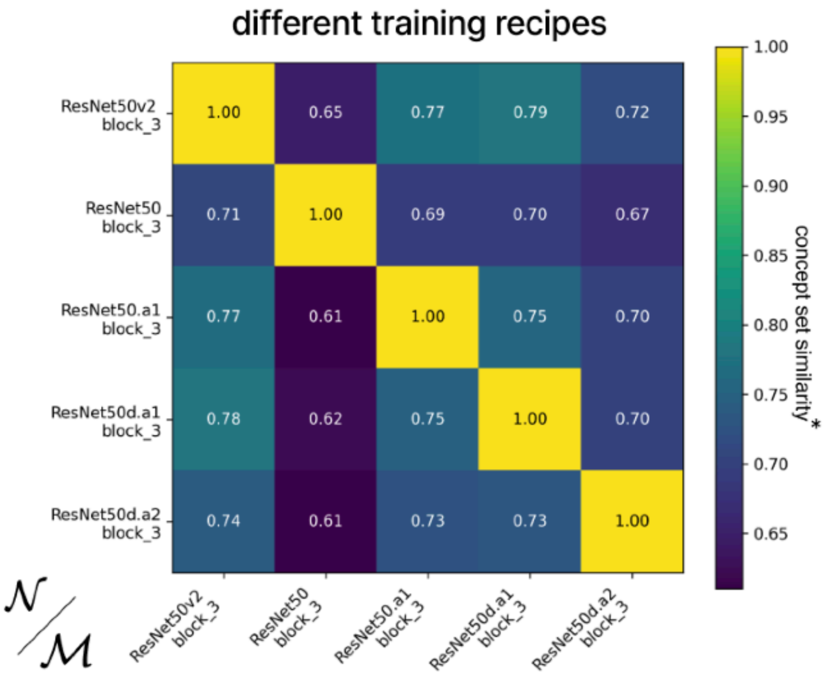
Compare: Identify Common and Unique Knowledge

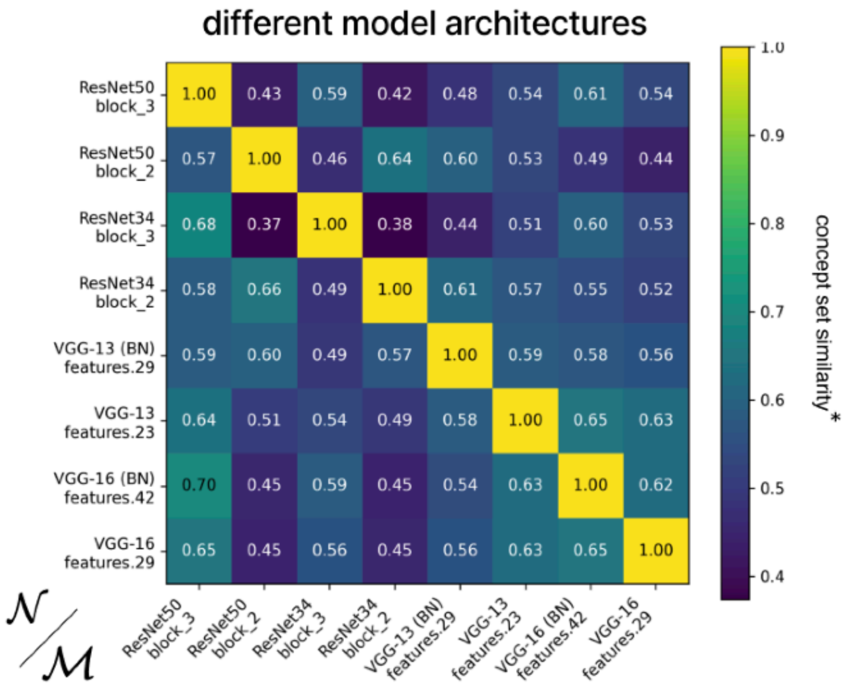
What concepts are shared between two models, and which are unique to each one?



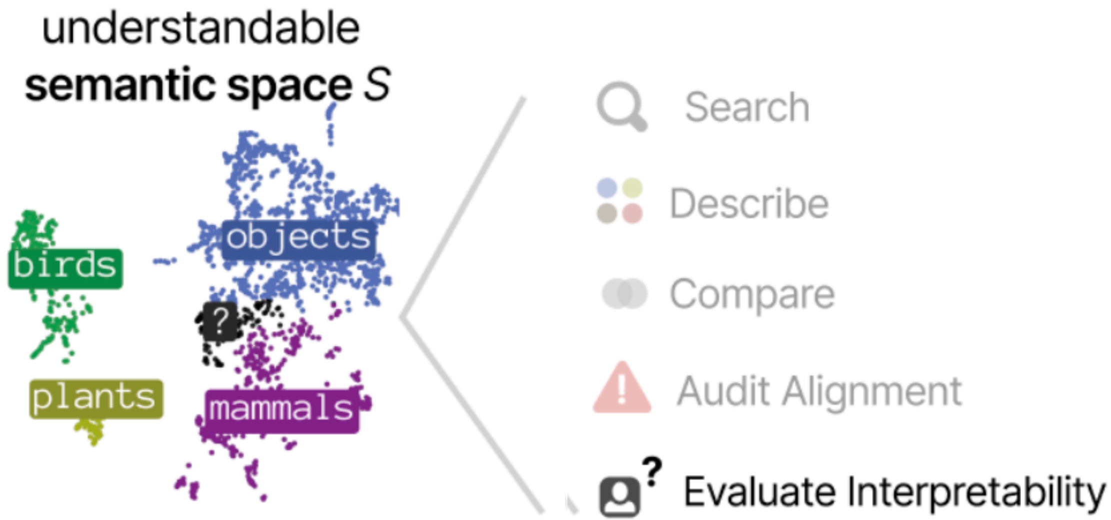
Note: Comparison can be done because components of different models maps into the same semantic space.

Compare: Identify Common and Unique Knowledge

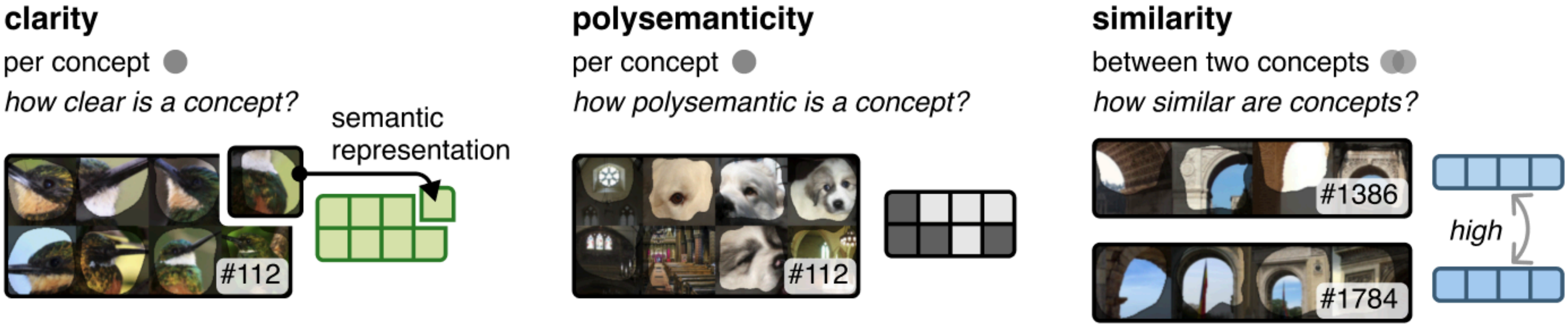




SemanticLens: What Can We do ?

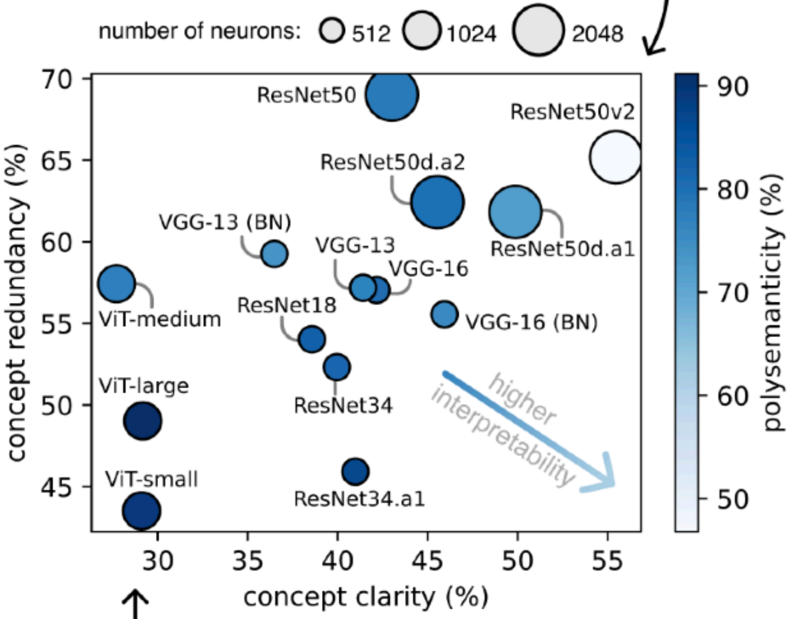


Evaluating Component Interpretability



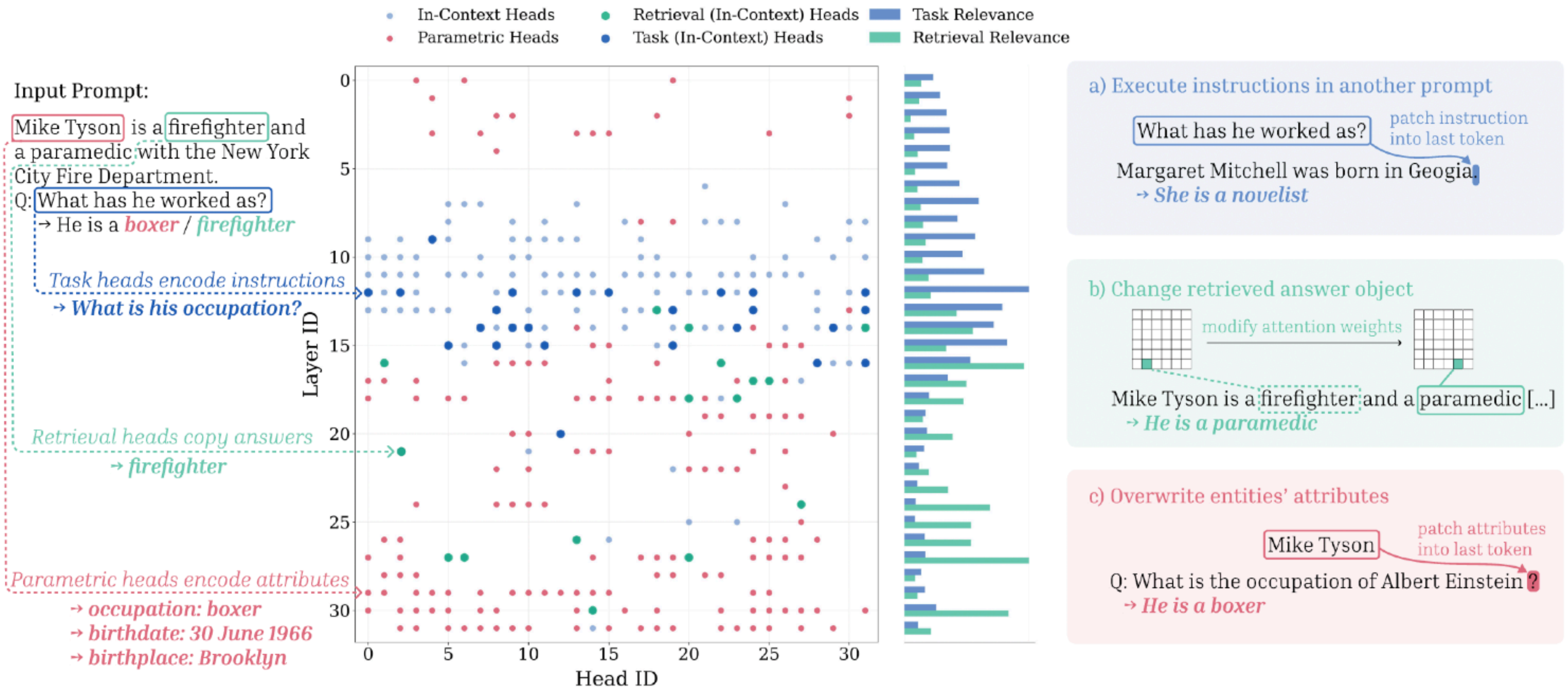
Evaluating Component Interpretability

well-trained, large CNNs have interpretable components

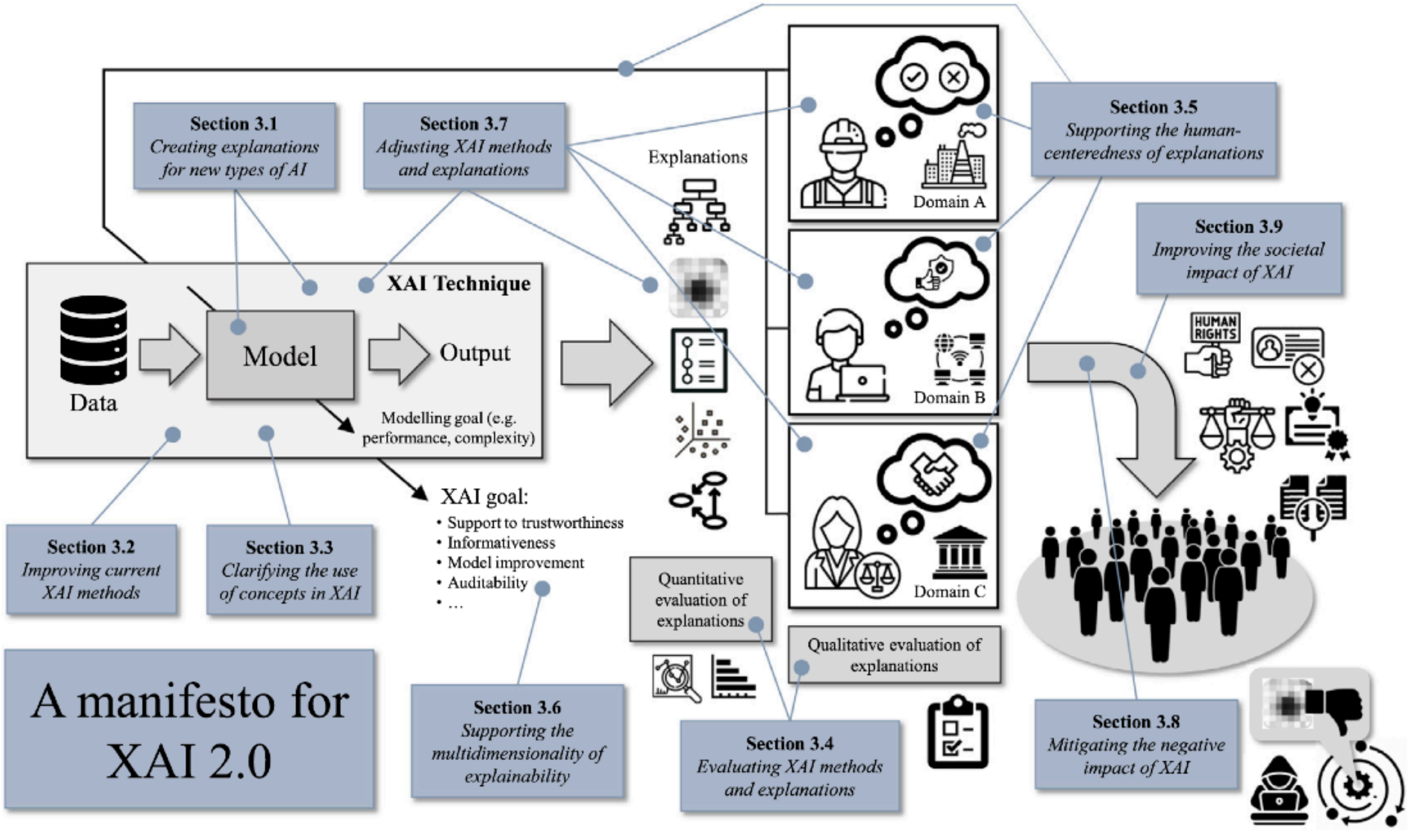


ViTs have no inherently intepretable components

Next Steps: Component-Level Understanding of LLMs



Future Work



(Longo et al. 2024)

https://doi.org/10.1016/j.inffus.2024.102301

Toolboxes

Benchmarking:
QUANTUS

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

SemanticLens

https://github.com/jim-berend/semanticlens

zennit

https://github.com/chr5tphr/zennit

Benchmarking:
CLEVR-XAI

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

CRP
CONCEPT
RELEVANCE
PROPAGATION
zennit-crp

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

iNNvestigate

https://github.com/albermax/innvestigate

quanda

https://github.com/dilyabareeva/quanda


THANK YOU
FOR YOUR
ATTENTION