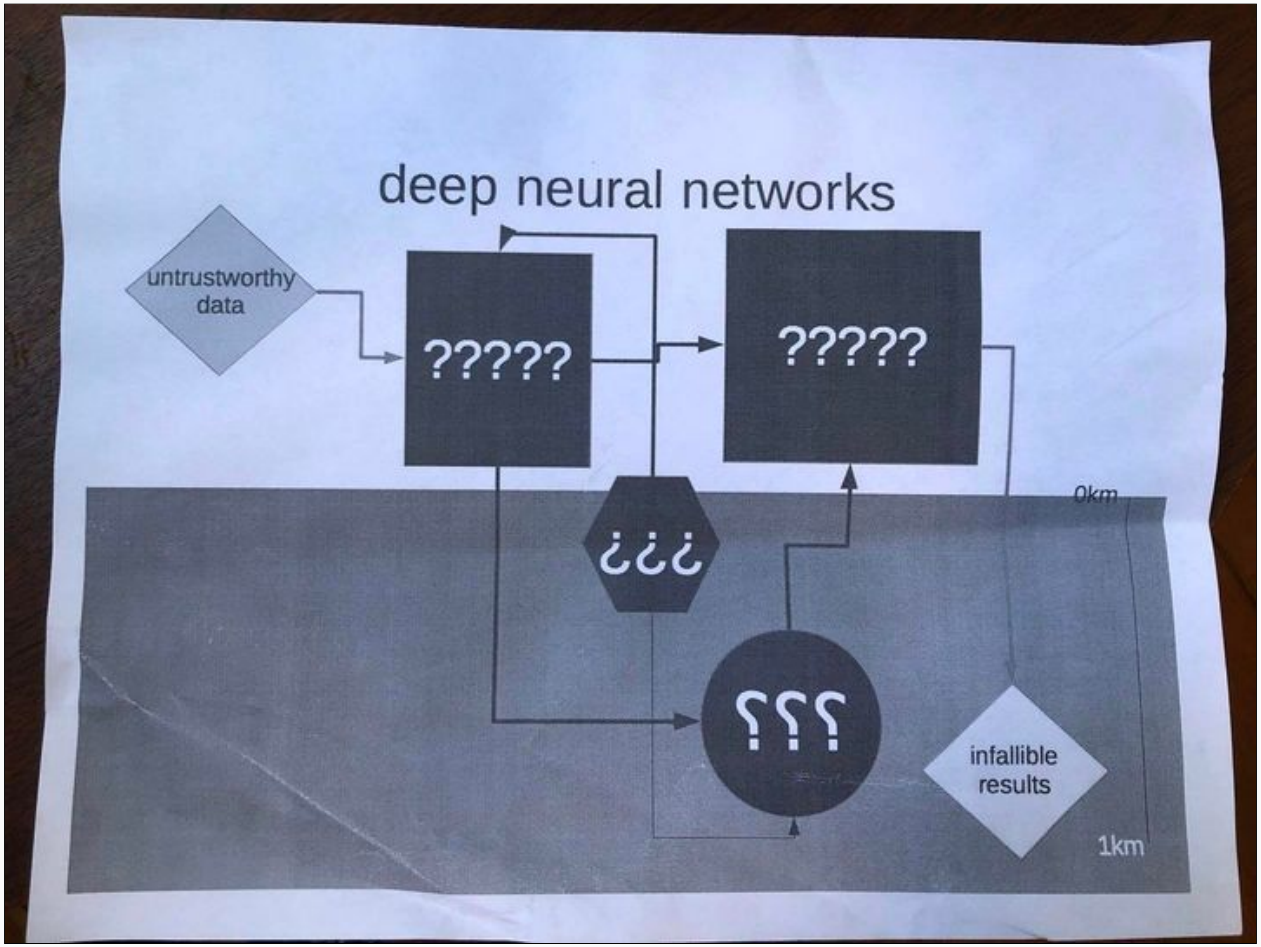
Detecting individual-level deviations in brain morphology with Layerwise Relevance Propagation

Esten Høyland Leonardsen

June 23, 2022

Explainable AI: Motivation



Explainable AI: Motivation

Explainable AI: Motivation

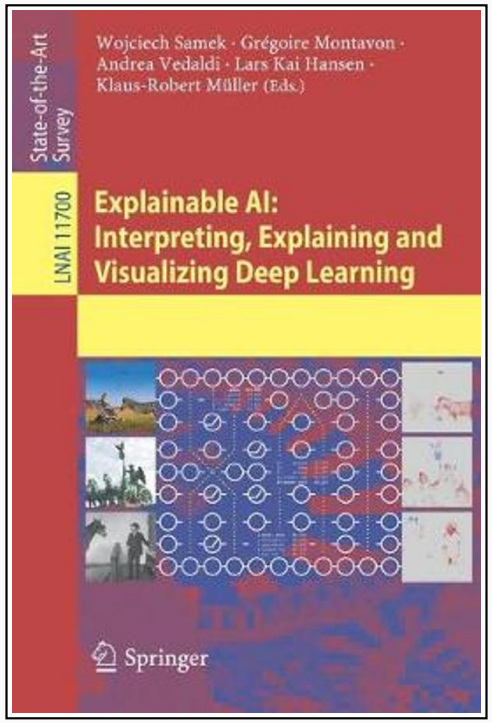
```
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "<stdin>", line 2, in print_model
  File "<stdin>", line 4, in print_layer
  File "<stdin>", line 4, in print_layer
  File "<stdin>", line 4, in print_layer
  [Previous line repeated 8 more times]
  File "<stdin>", line 3, in print_layer
MemoryError
```

Explainable AI: Motivation

"Relying on devices whose logic is opaque violates principles of medical ethics."

Kundu, S. AI in medicine must be explainable. Nat Med 27, 1328 (2021)

Explainable AI: Overview



Samek, et al., Explainable AI: interpreting, explaining and visualizing deep learning. Springer Nature, 2019

Explainable AI: Overview

Recipient

End user

Domain expert

Policy makers

Society

Developer

Level

Explain the model as a whole

Explain single decisions

Produce stereotypical examples

Explainable AI: Overview

Recipient

End user

Domain expert

Policy makers

Society

Developer

Level

Explain the model as a whole

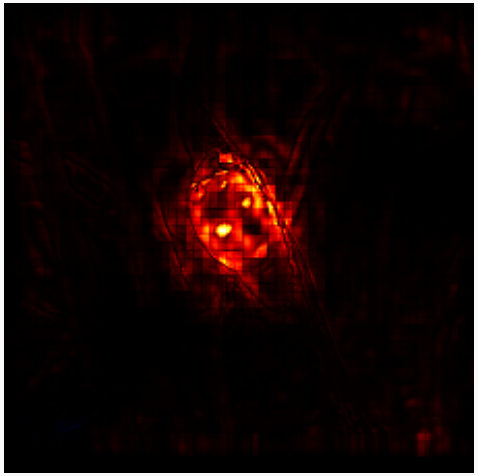
Explain single decisions

Produce stereotypical examples

Explainable AI: Overview

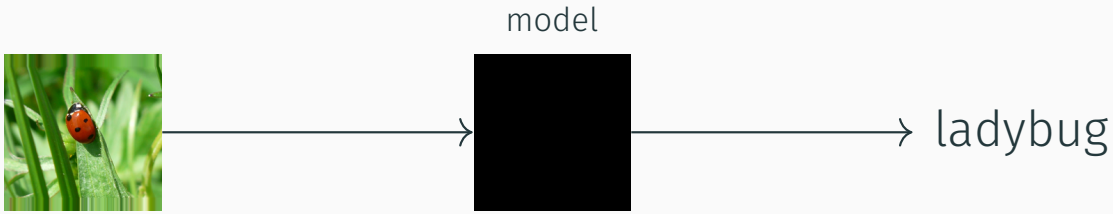
This is a
ladybug because
of the red
back with
the black dots

Explainable AI: Overview

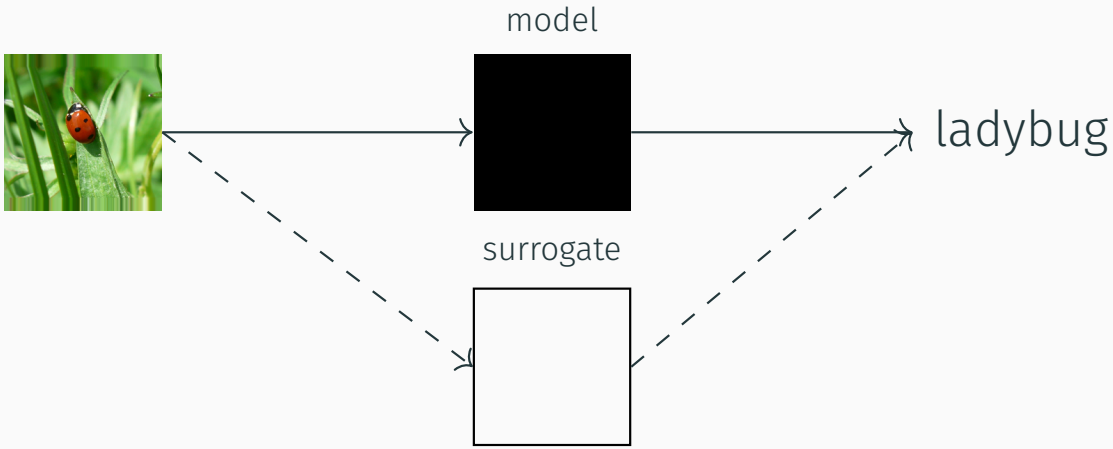


https://lrpserver.hhi.fraunhofer.de/image-classification

Explainable AI: Surrogate models

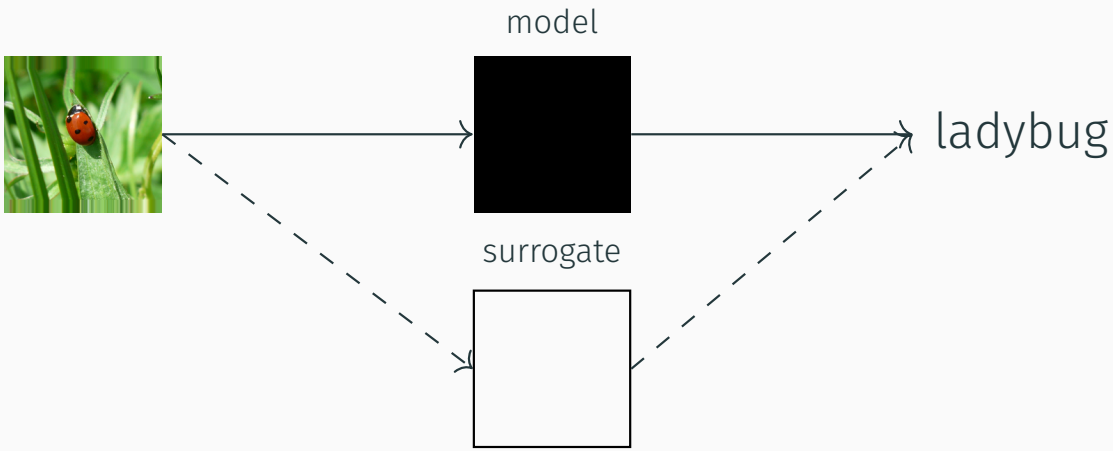


Explainable AI: Surrogate models

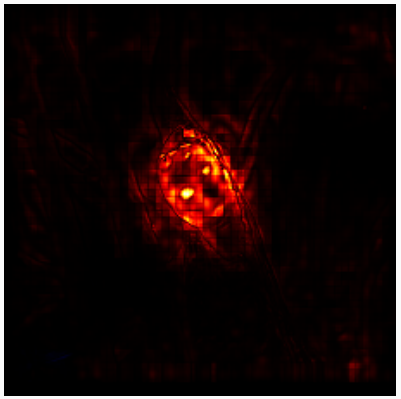


$$y \backsim x_1 w_1 + x_2 w_2 + \ldots + x_n w_n$$

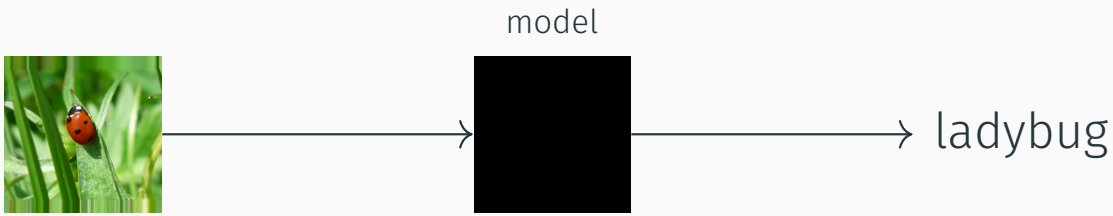
Explainable AI: Surrogate models



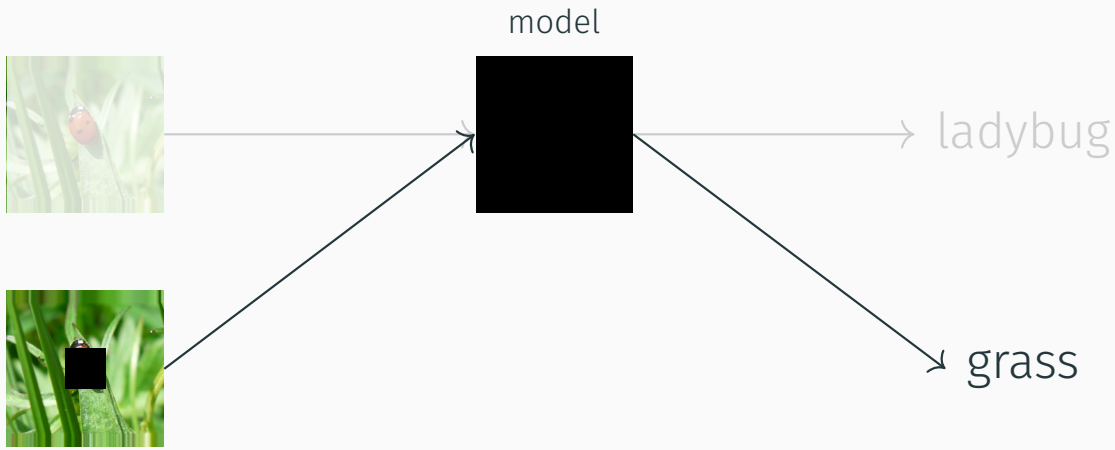
$$y \backsim x_1 w_1 + x_2 w_2 + \ldots + x_n w_n$$



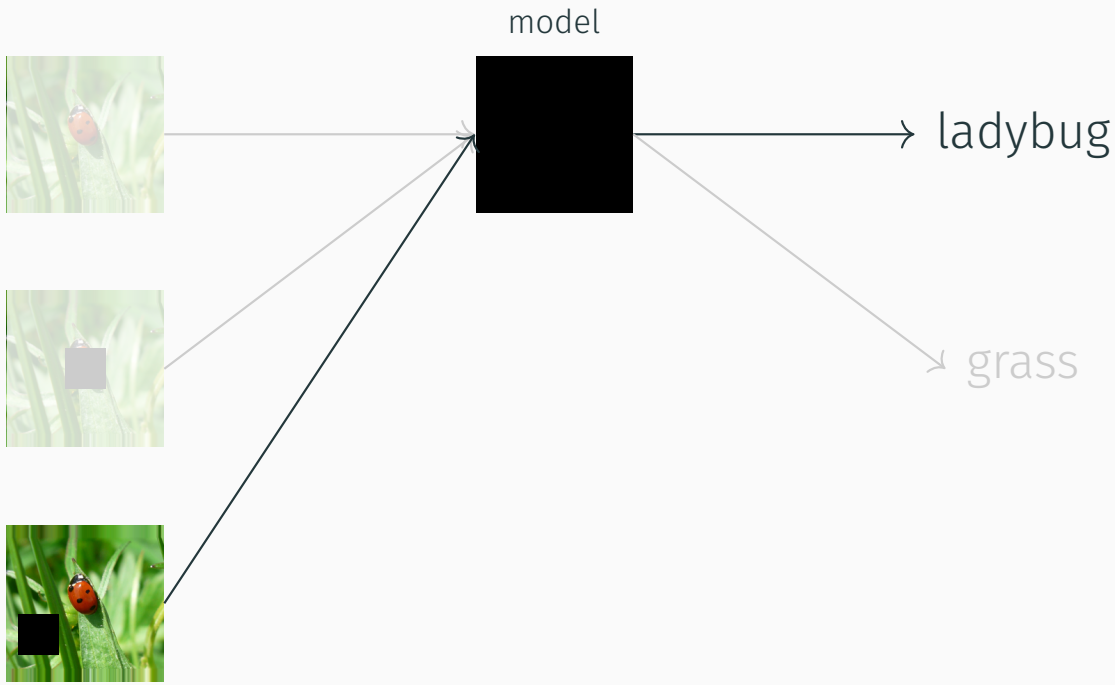
Explainable AI: Occlusion



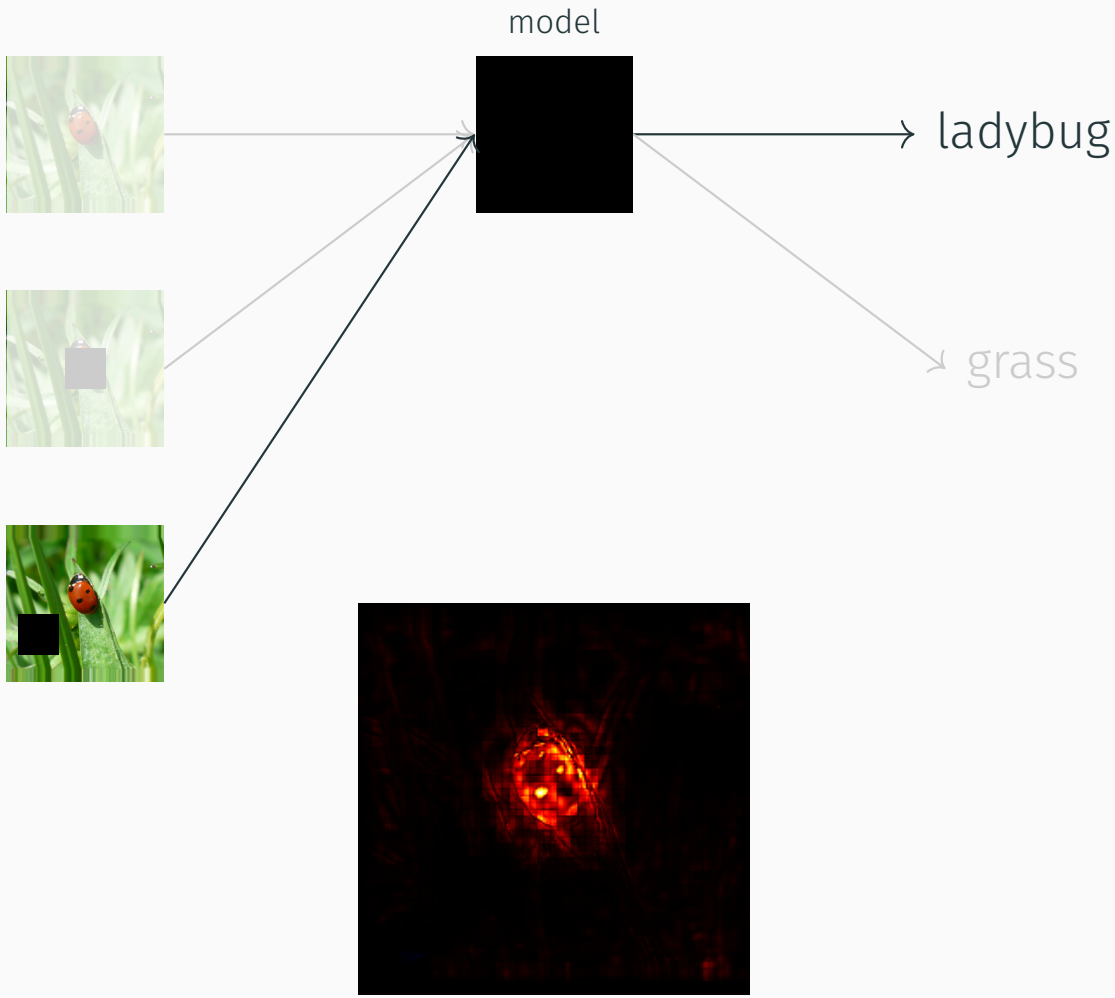
Explainable AI: Occlusion



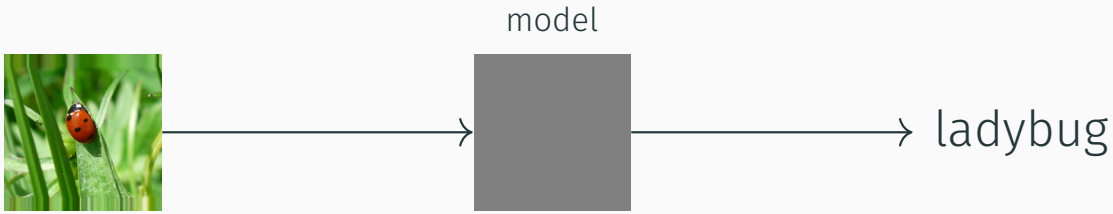
Explainable AI: Occlusion



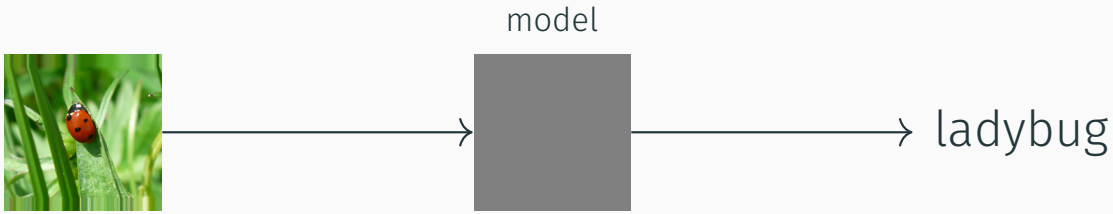
Explainable AI: Occlusion



Explainable AI: Saliency mapping

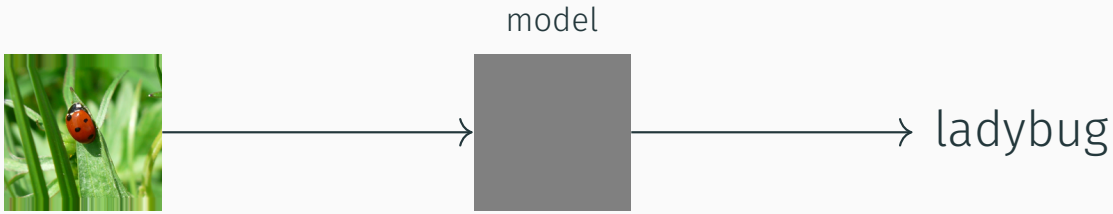


Explainable AI: Saliency mapping



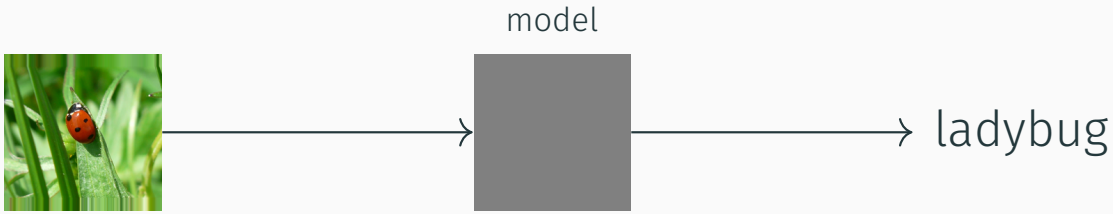
$$y = \sum \ldots \sum x_{i,j,k} W$$

Explainable AI: Saliency mapping

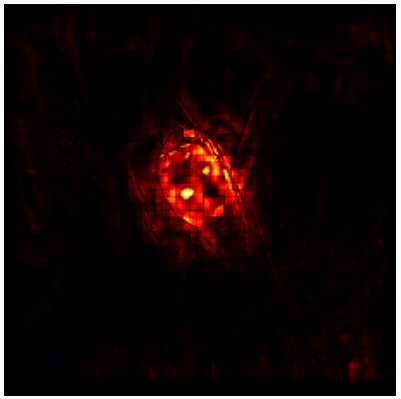


$$y = \sum \ldots \sum x_{i,j,k} w$$

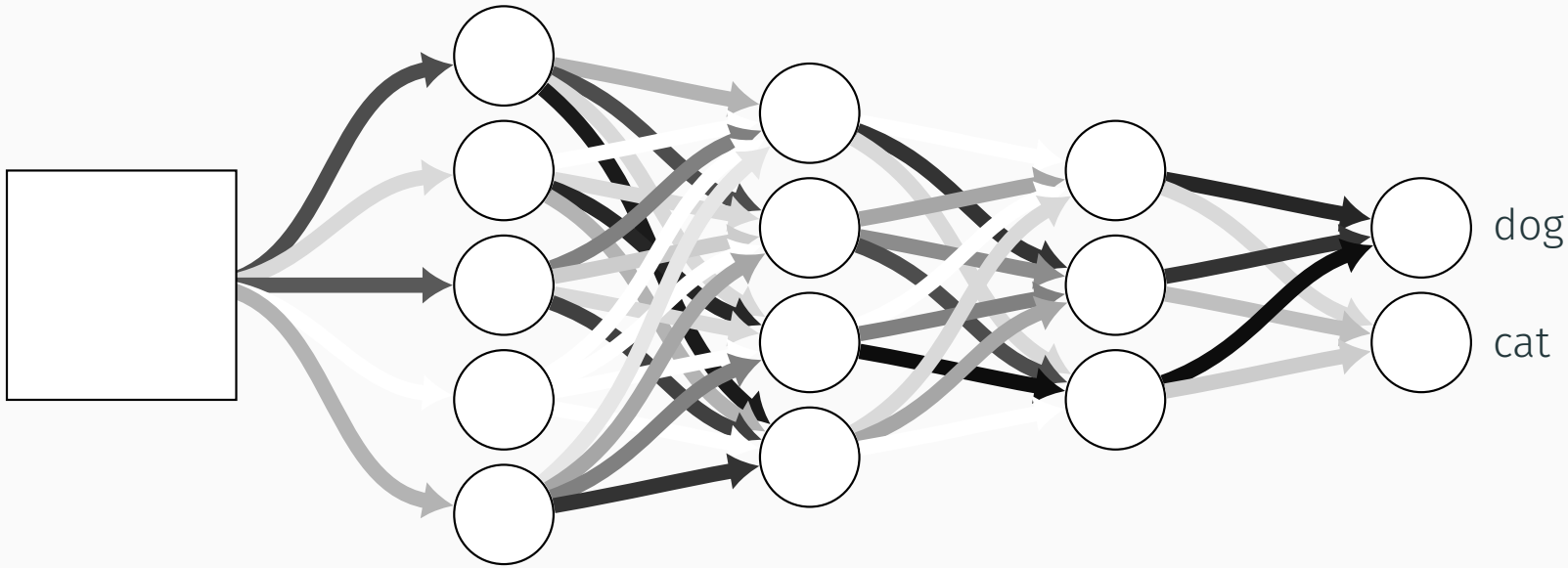
Explainable AI: Saliency mapping



$$y = \sum \ldots \sum x_{i,j,k} w$$



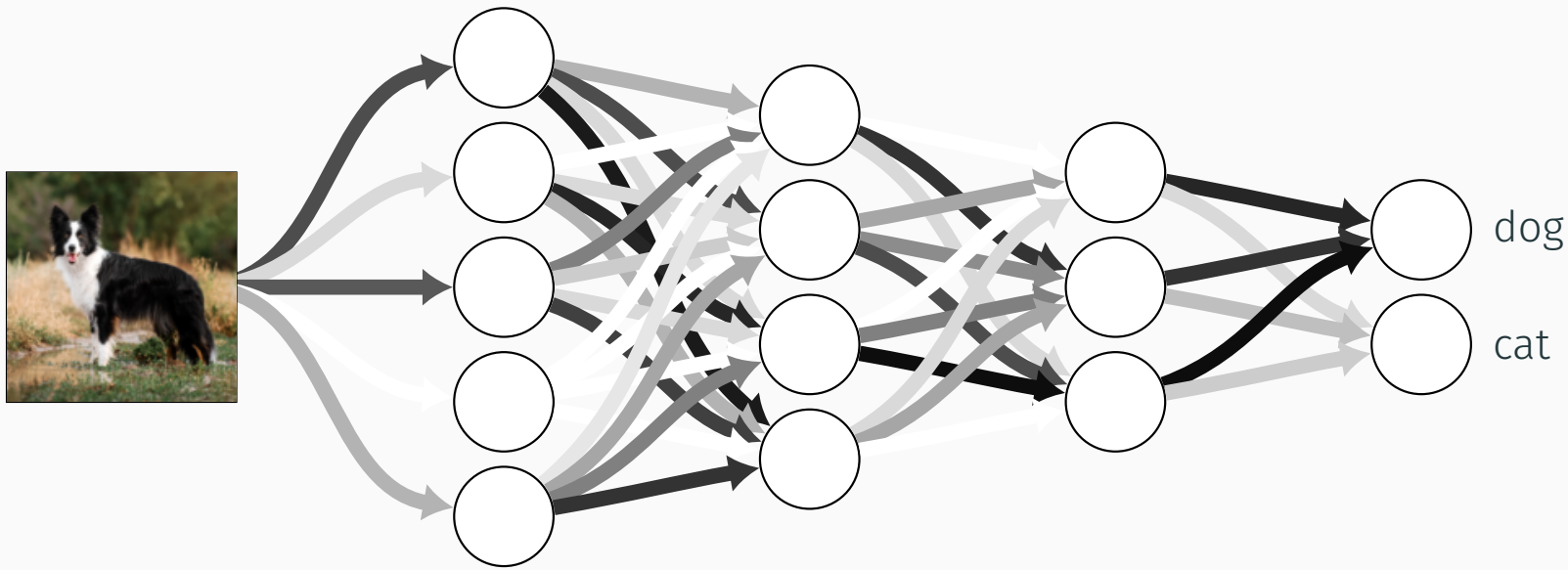
Explainable AI: Layerwise Relevance Propagation



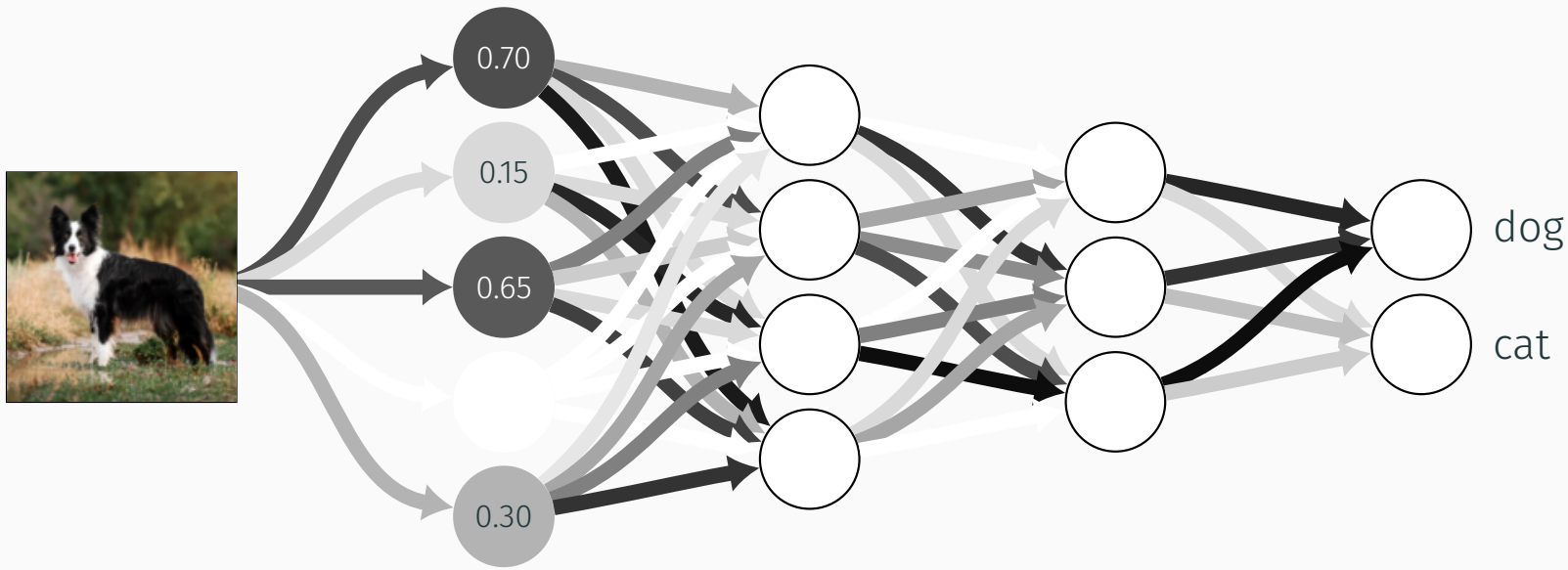
Imagenet-trained VGG19

https://keras.io/api/applications/

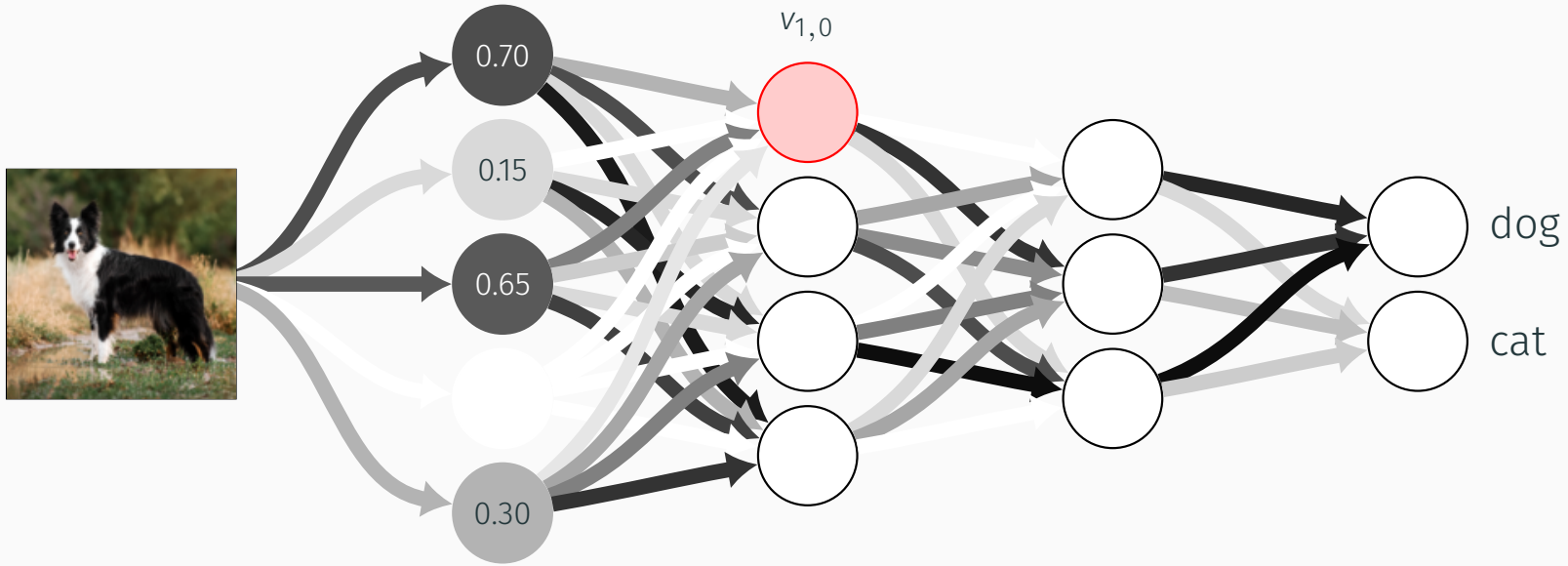
Explainable AI: Layerwise Relevance Propagation



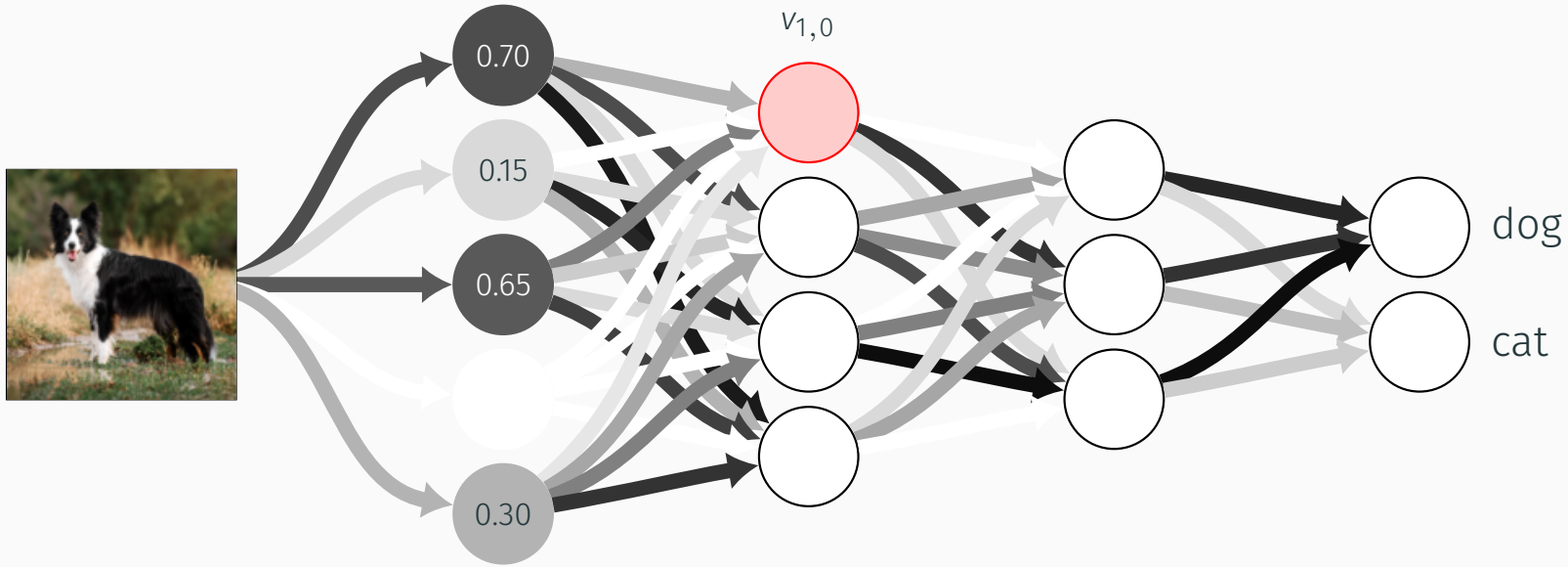
Explainable AI: Layerwise Relevance Propagation



Explainable AI: Layerwise Relevance Propagation

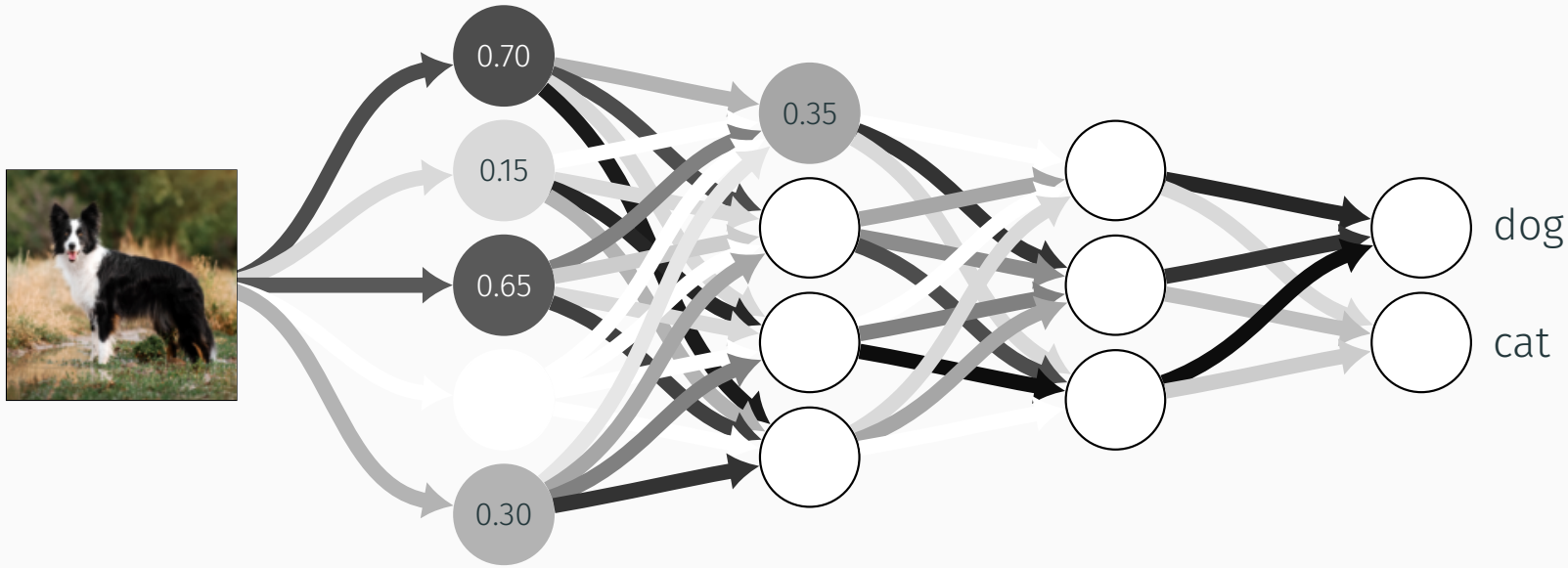


Explainable AI: Layerwise Relevance Propagation

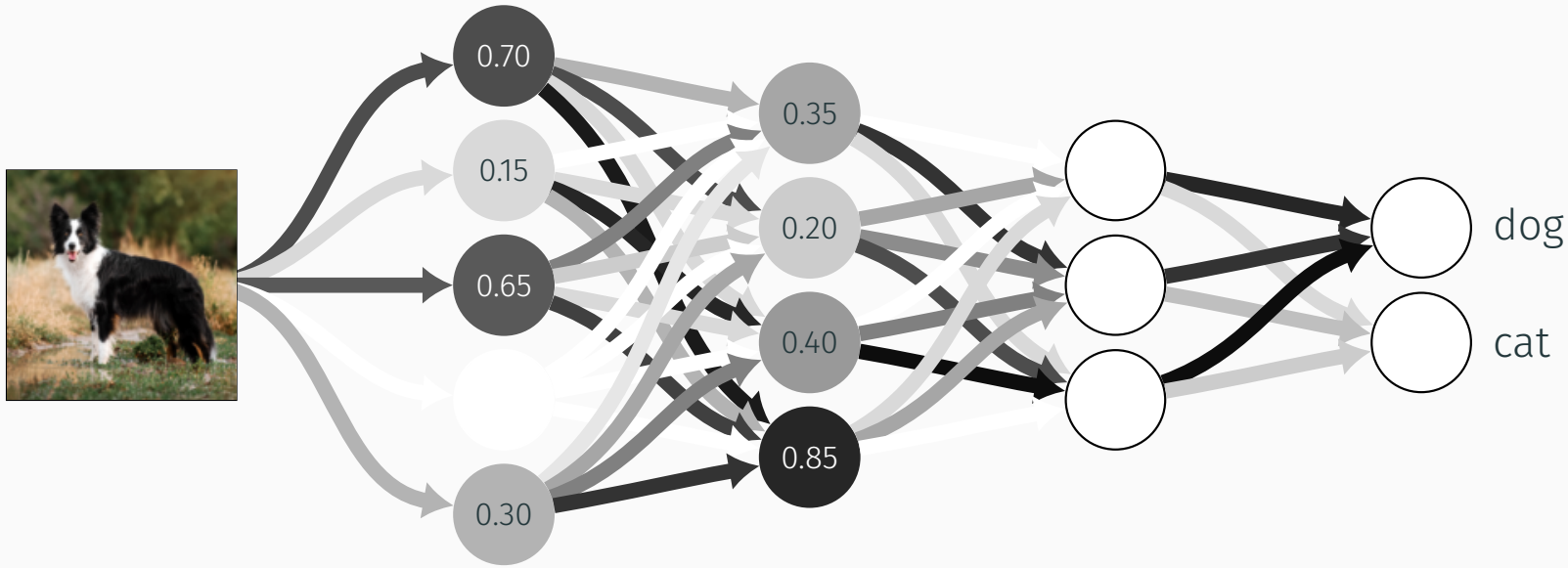


$$v_{1,0} = \sum v_{0,j} * w_{j,1}$$

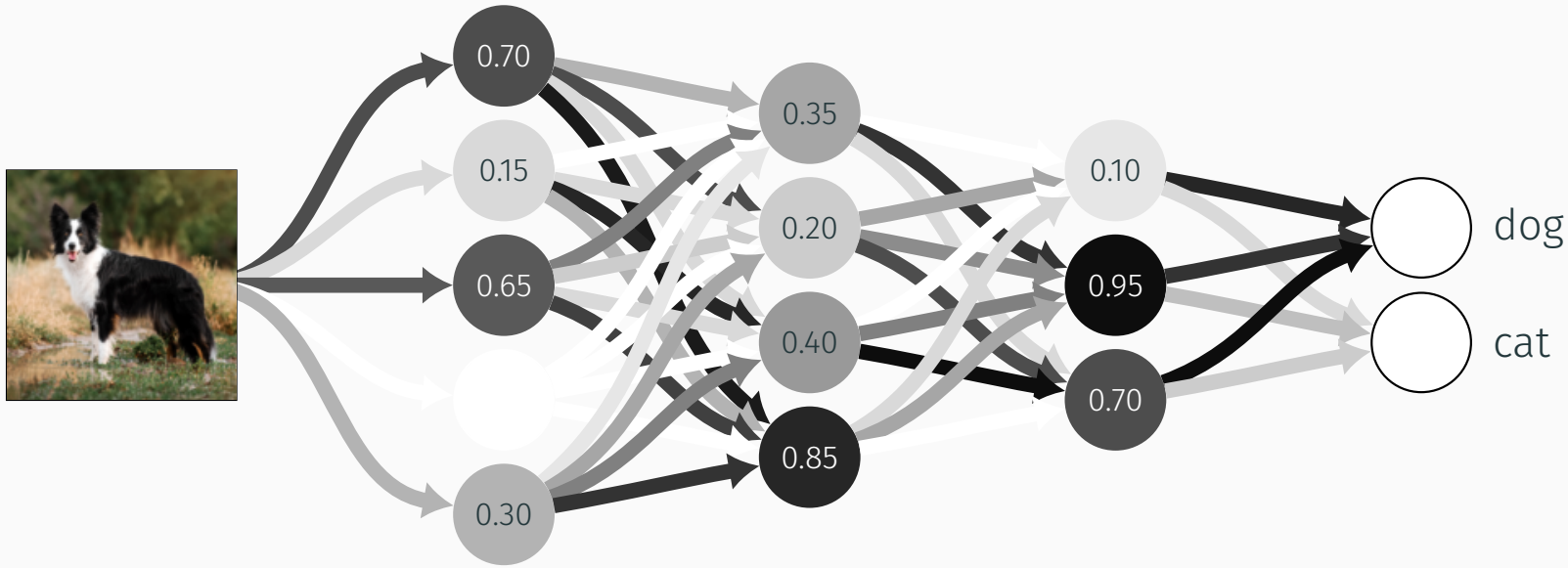
Explainable AI: Layerwise Relevance Propagation



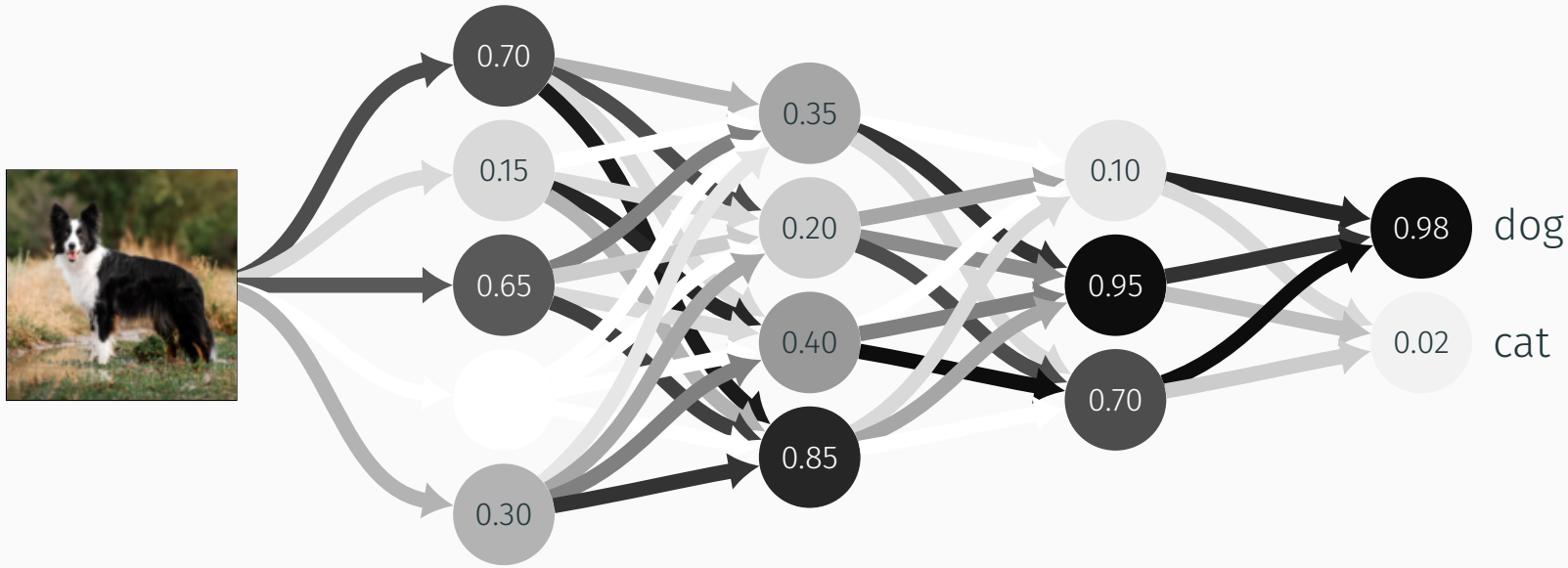
Explainable AI: Layerwise Relevance Propagation



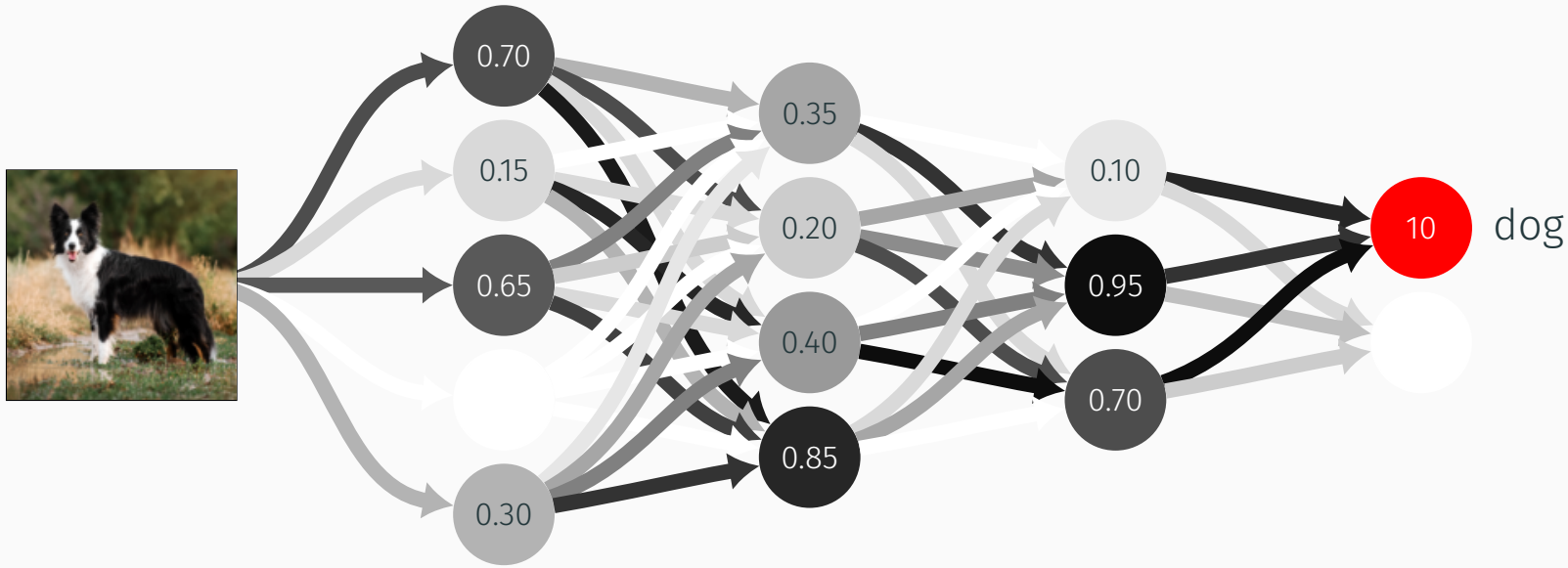
Explainable AI: Layerwise Relevance Propagation



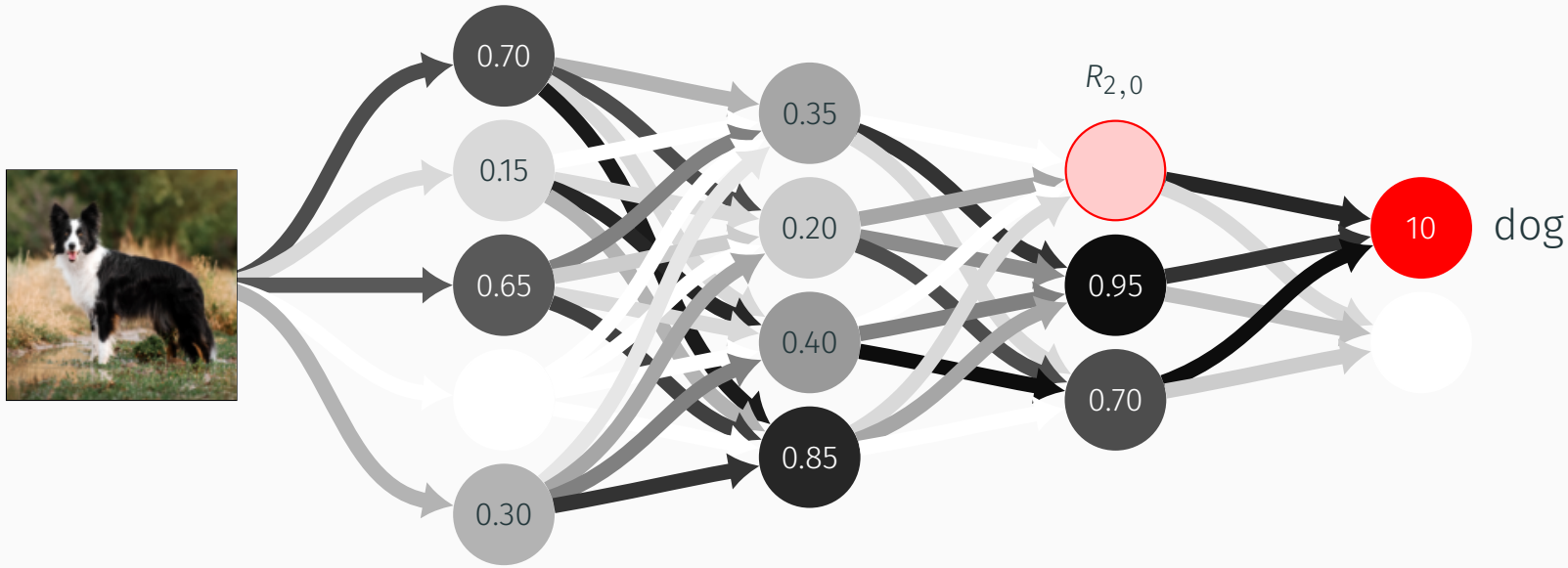
Explainable AI: Layerwise Relevance Propagation



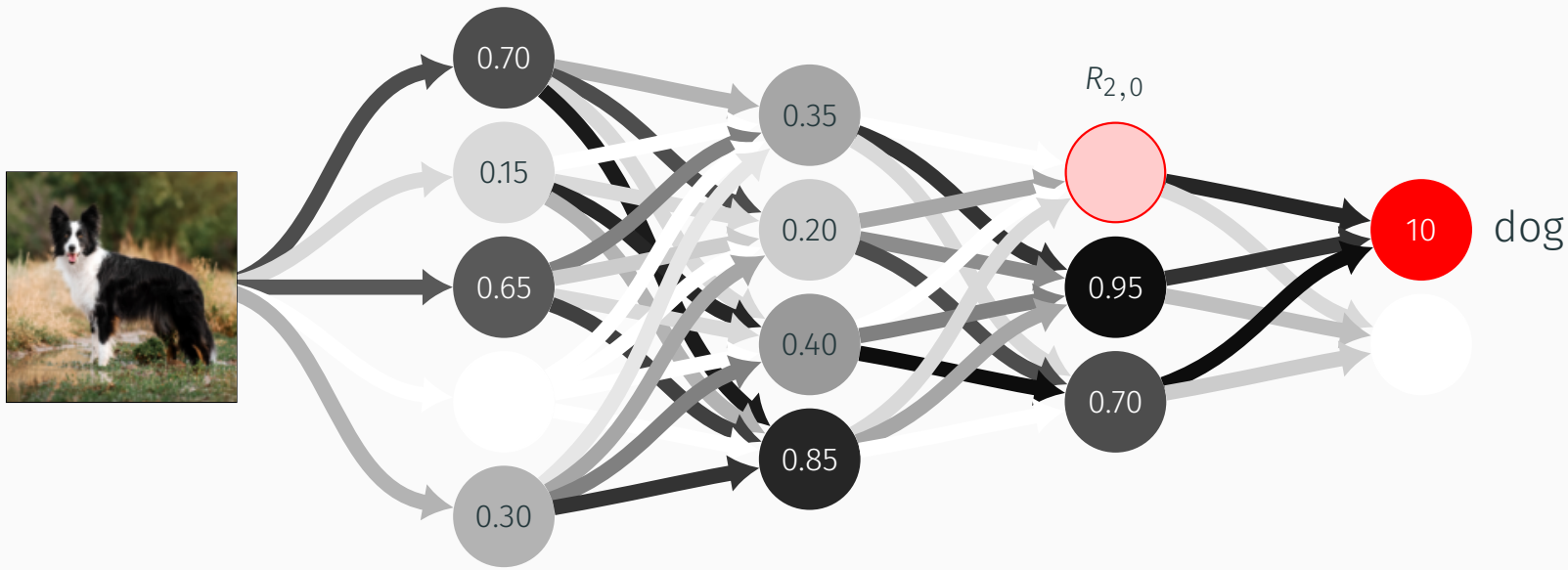
Explainable AI: Layerwise Relevance Propagation



Explainable AI: Layerwise Relevance Propagation

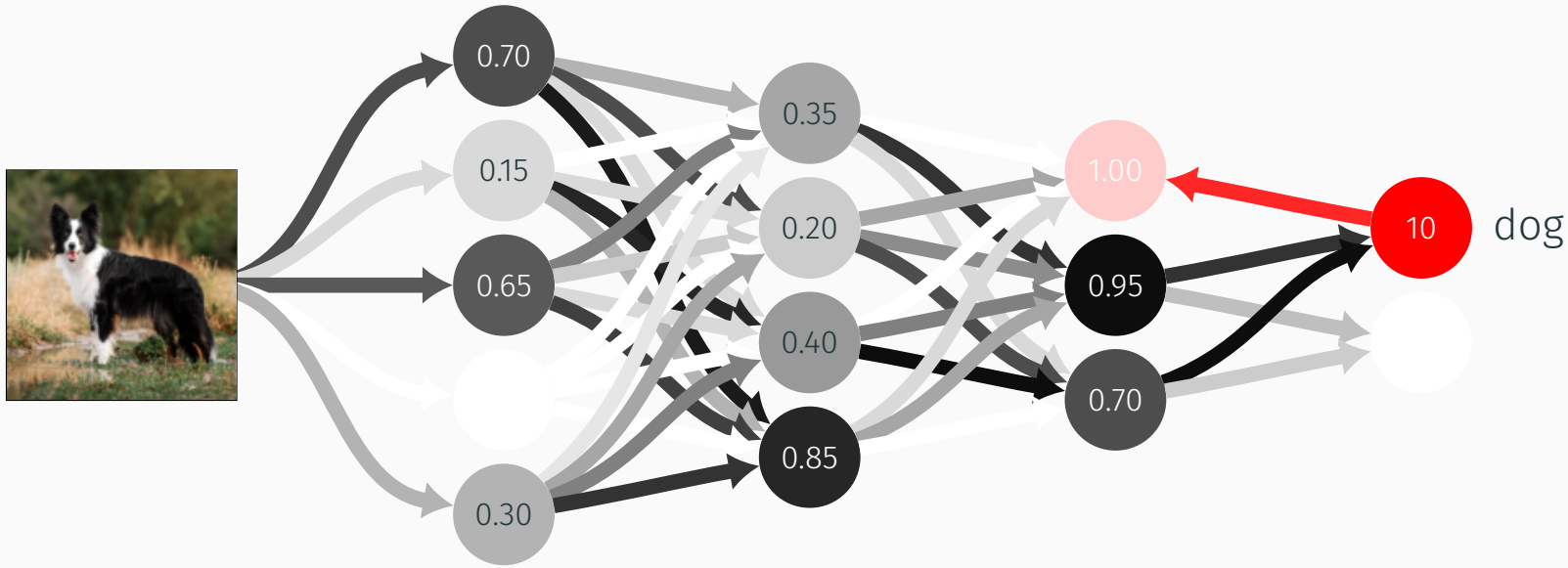


Explainable AI: Layerwise Relevance Propagation

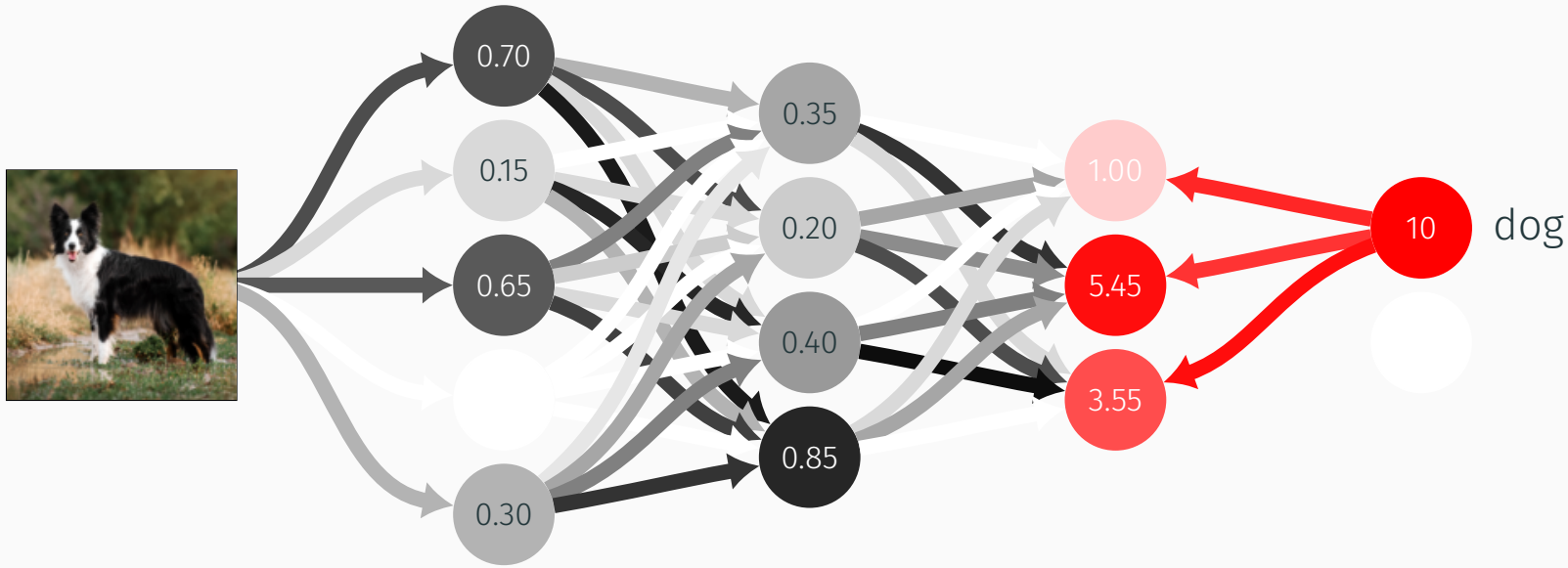


$$R_{2,0} = \frac{v_{2,0} w_{0,0}}{\sum v_{2,i} w_{i,0}}$$

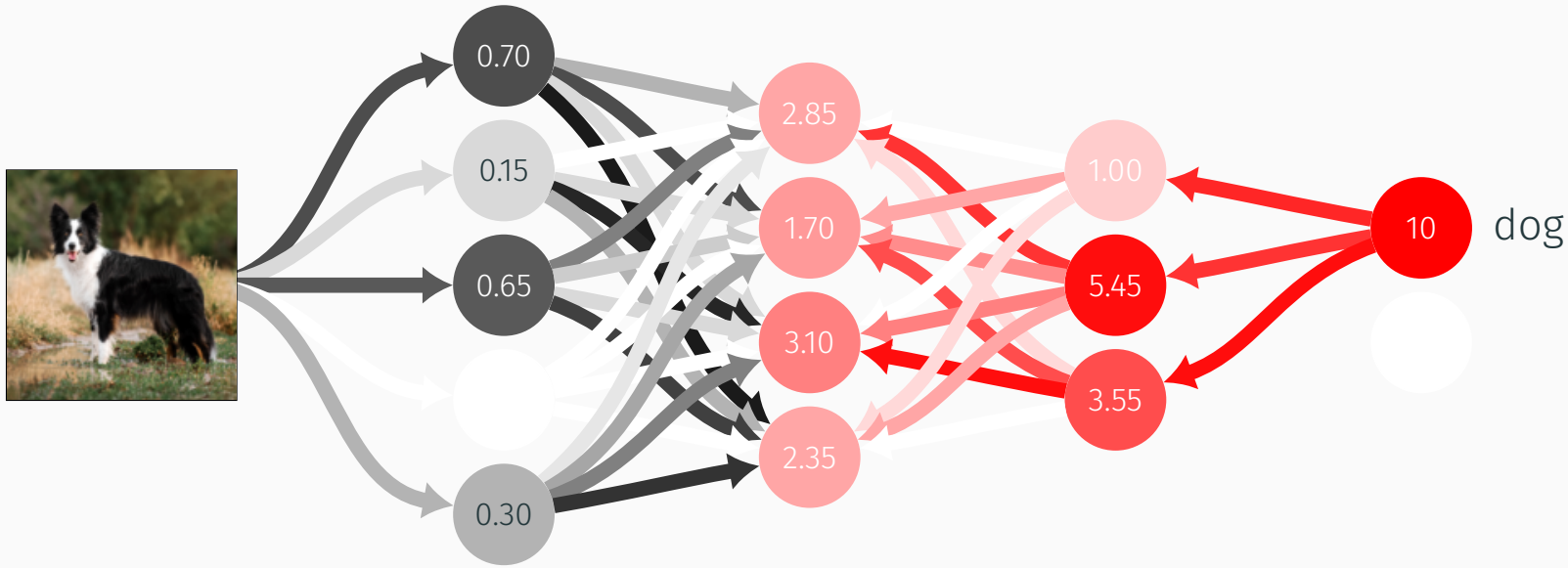
Explainable AI: Layerwise Relevance Propagation



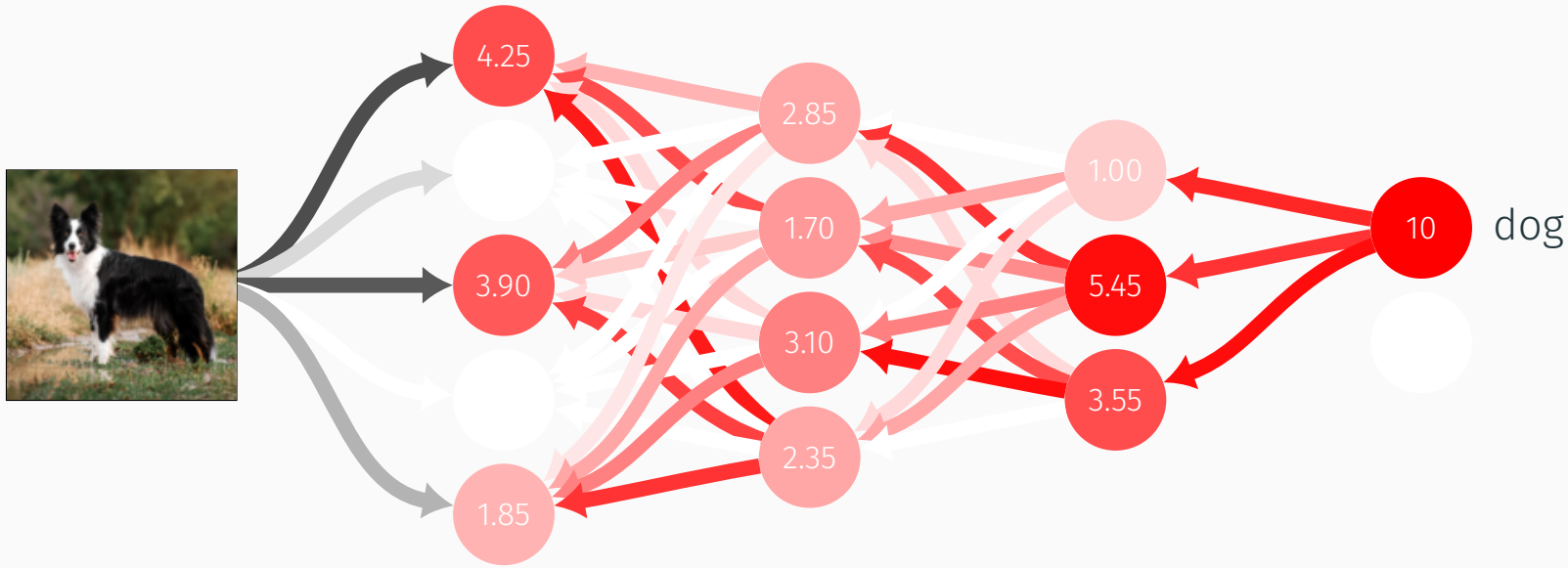
Explainable AI: Layerwise Relevance Propagation



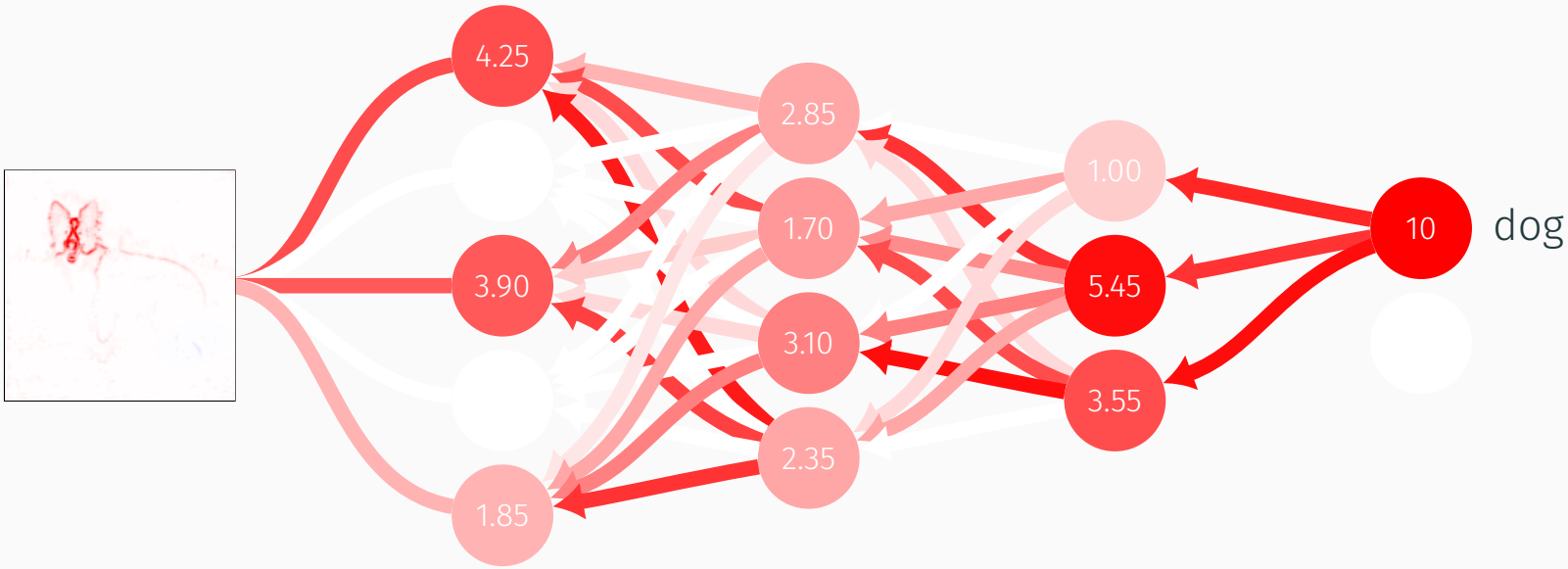
Explainable AI: Layerwise Relevance Propagation



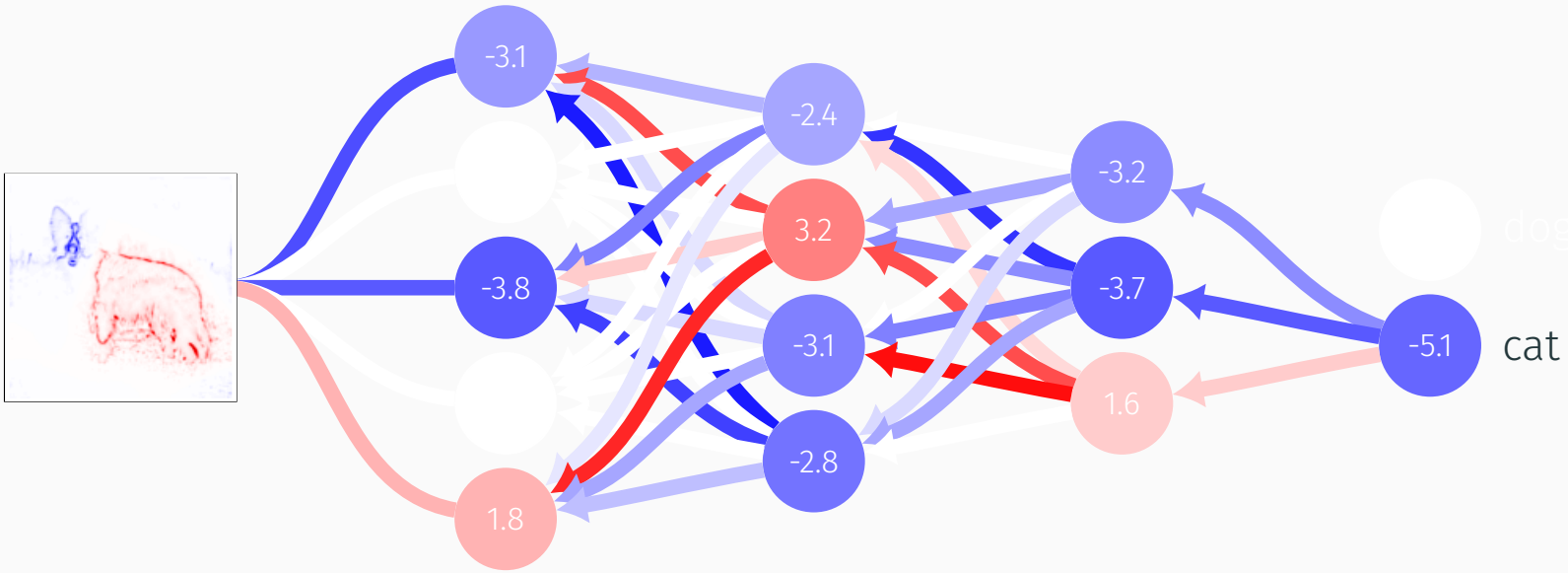
Explainable AI: Layerwise Relevance Propagation



Explainable AI: Layerwise Relevance Propagation



Explainable AI: Layerwise Relevance Propagation



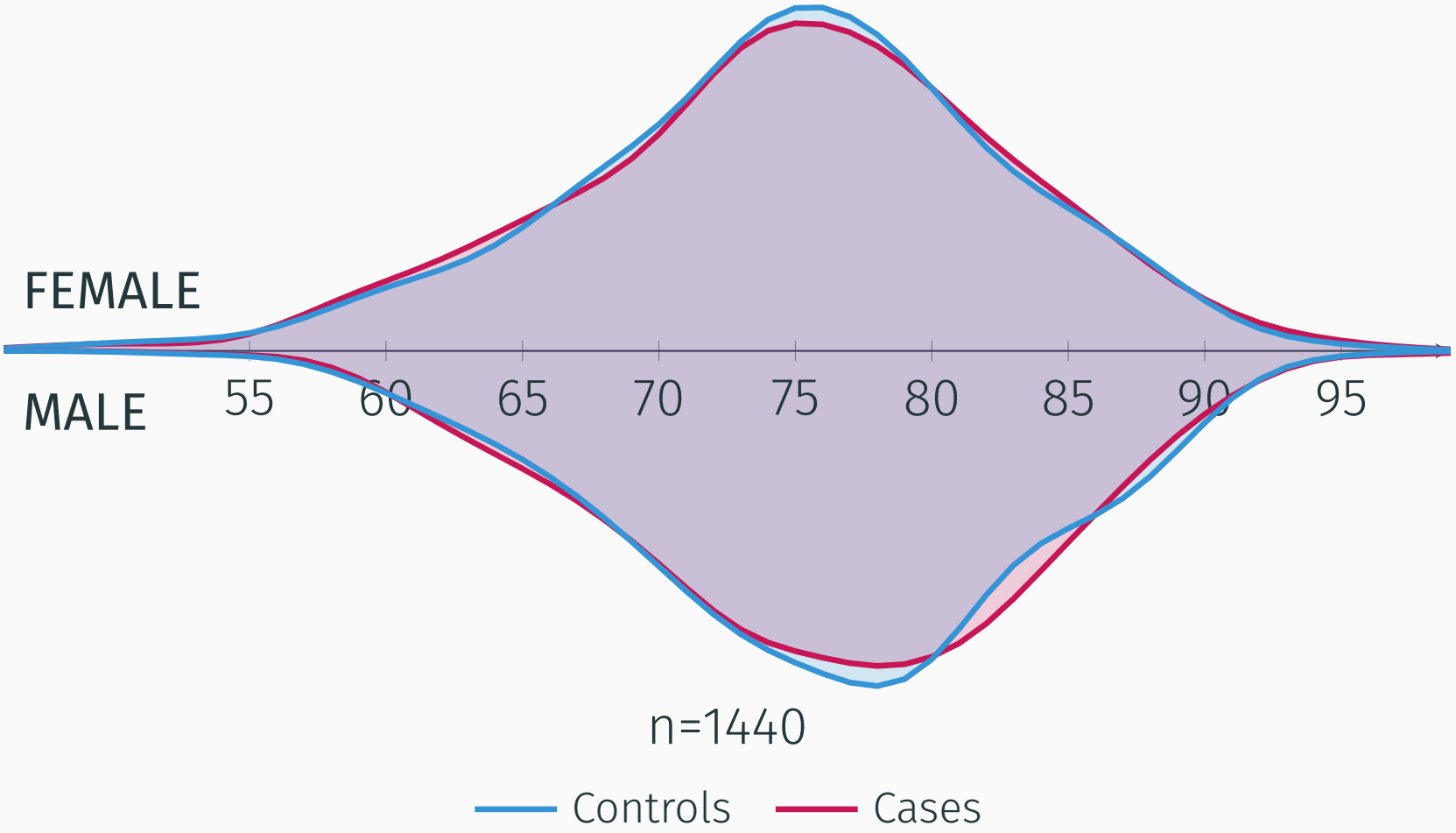
Explainable AI: Layerwise Relevance Propagation

$$\text{LRP-0:} \qquad R_j^l = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k^{(l+1)}$$

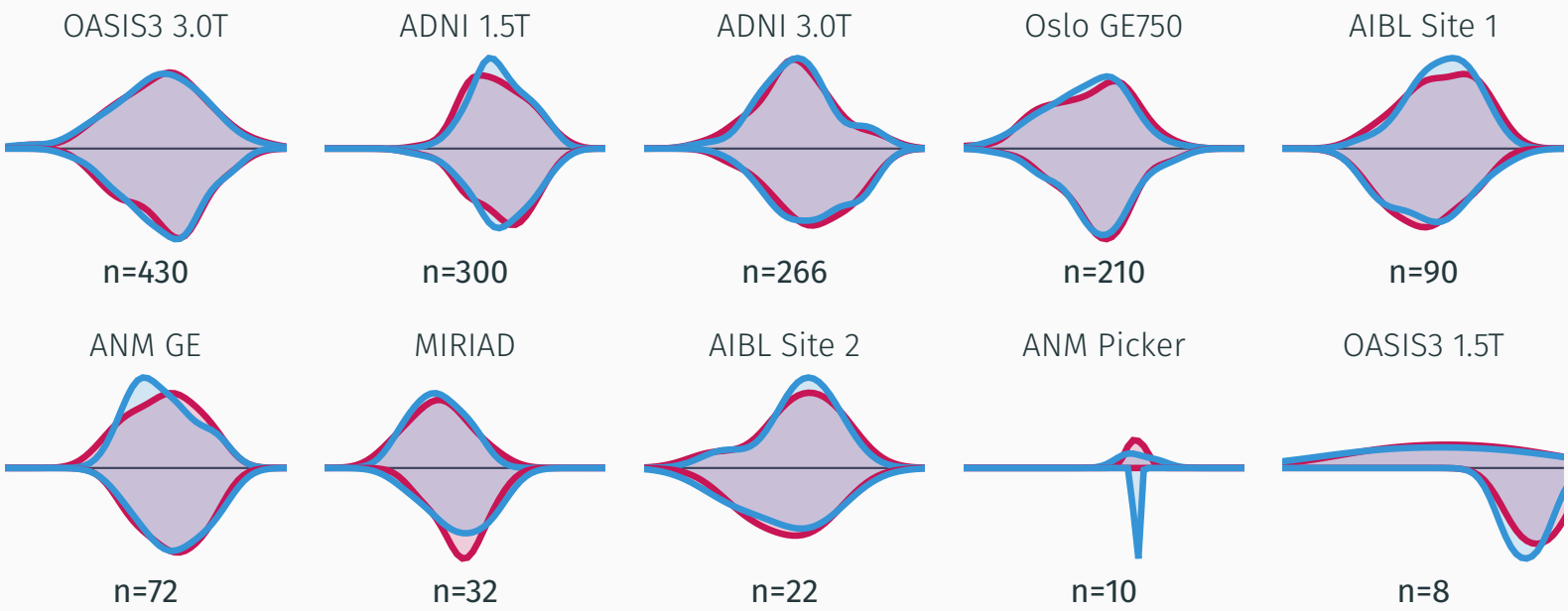
$$\text{LRP-}\epsilon\text{:} \qquad R_j^l = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk} + sign(a_j w_{jk}) * \epsilon} R_k^{(l+1)}$$

$$\text{LRP-}\alpha\beta\text{:} \qquad R_j^l = \sum_k \alpha \frac{a_j w_{jk}^+}{\sum_{0,j} a_j w_{jk}} - \beta \frac{a_j w_{jk}^-}{\sum_{0,j} a_j w_{jk}} R_k^{(l+1)}$$

Dementia: Dataset



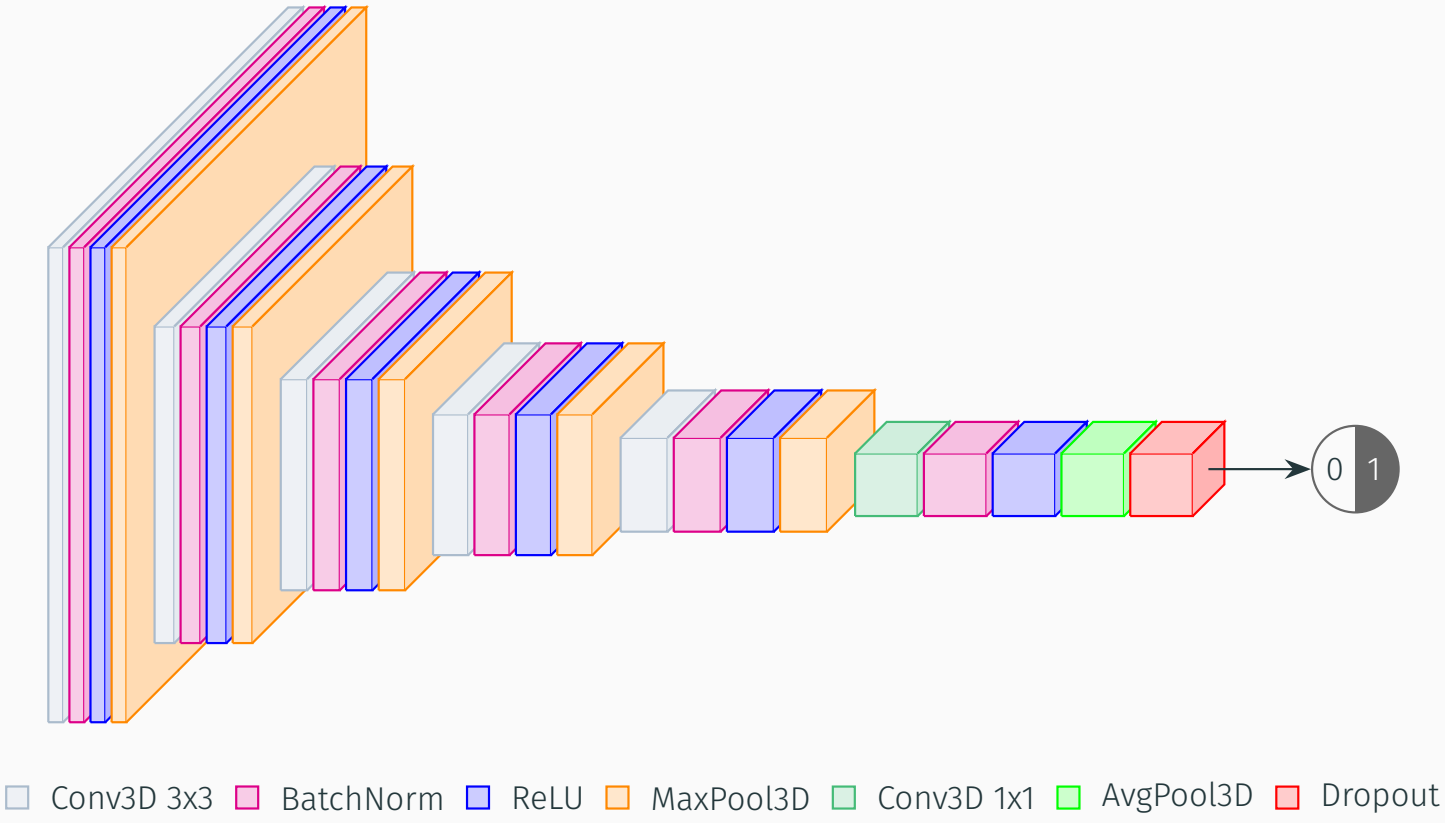
Dementia: Dataset



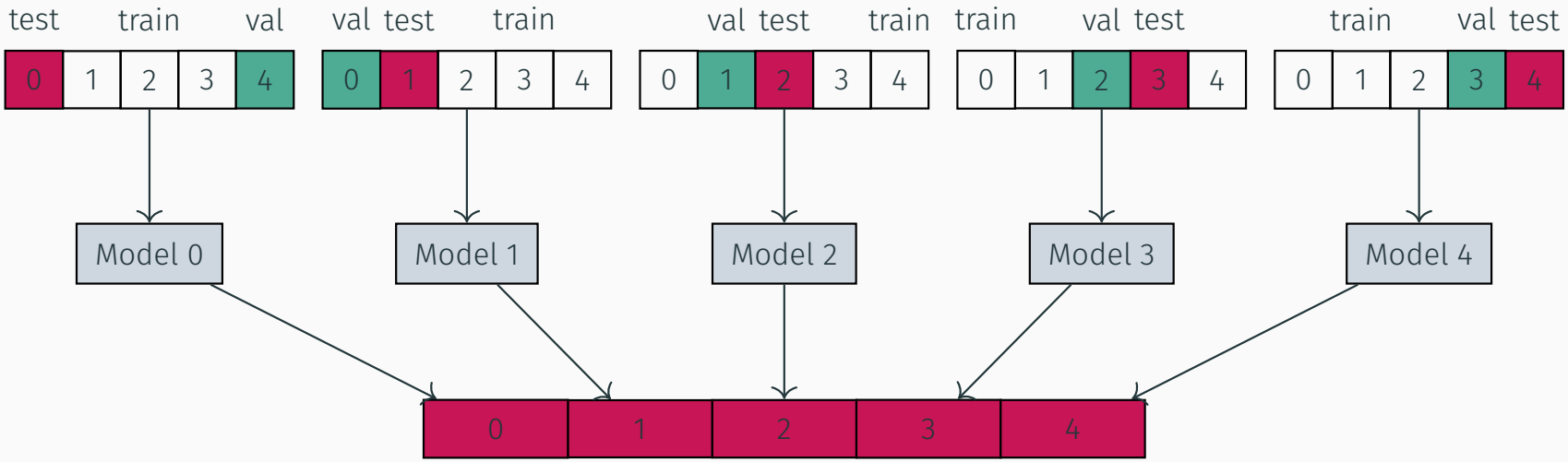
Dementia: Dataset

Dataset	Controls	Patients
AddNeuroMed	MMSE ≥ 24	MMSE < 19
ADNI	Group = CN	Group = AD
AIBL	Group = DXNORM	Group $\in$ {DXAD, DXOTHDEM}
CADDementia	?	?
Demgen	-	DX $\in$ {AD, OtherDem, UnspecDem, VaD}
MIRIAD	Group = Control	Group = AD
OASIS3	NORMCOG = 1	NORMCOG = 0 & DEMENTED = 1
StrokeMRI	Group = Control	-
TOP	diagnosis = CTRL	-

Dementia: Modelling

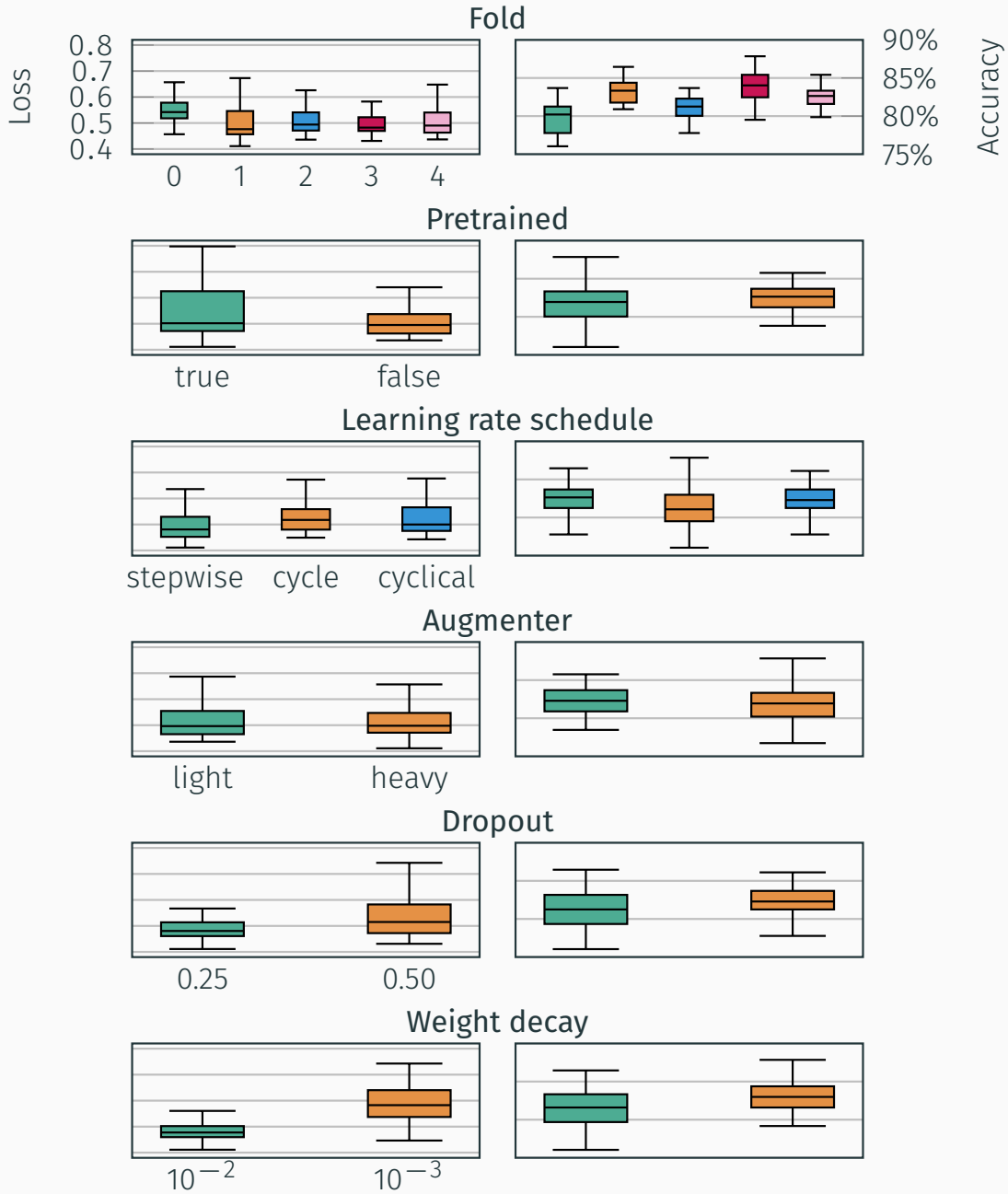


Dementia: Modelling

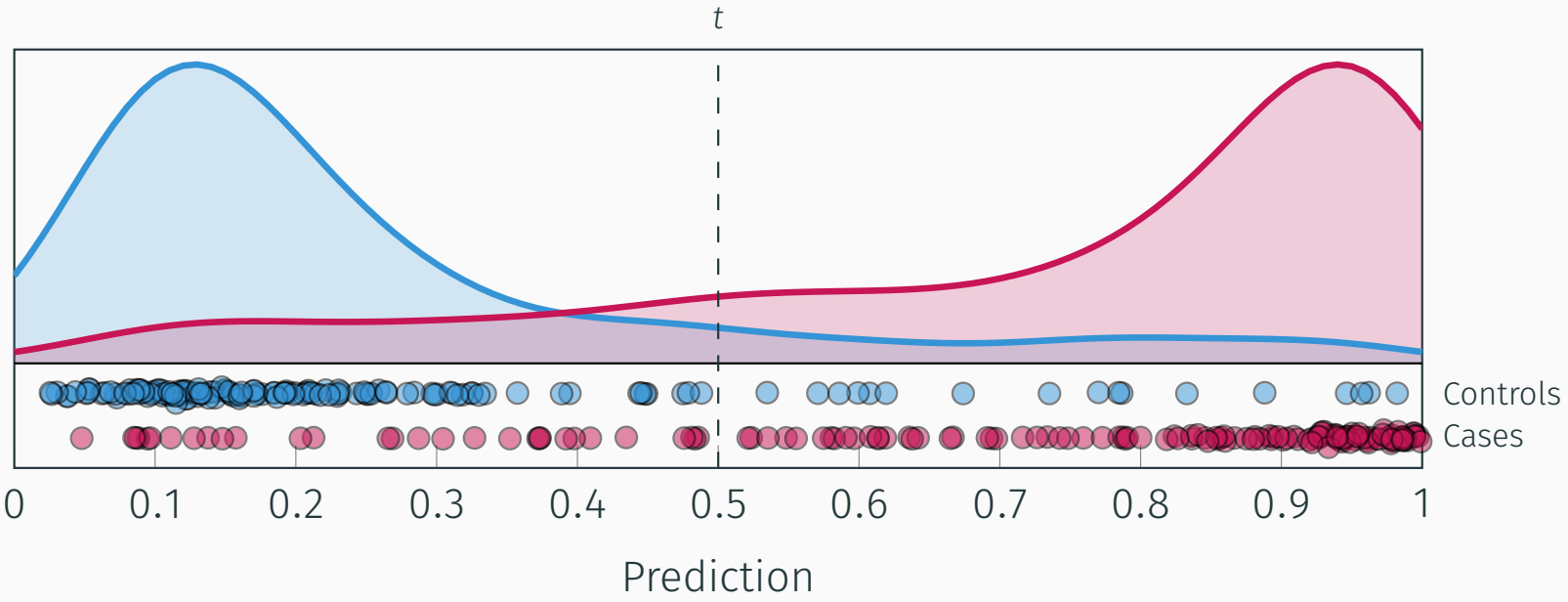


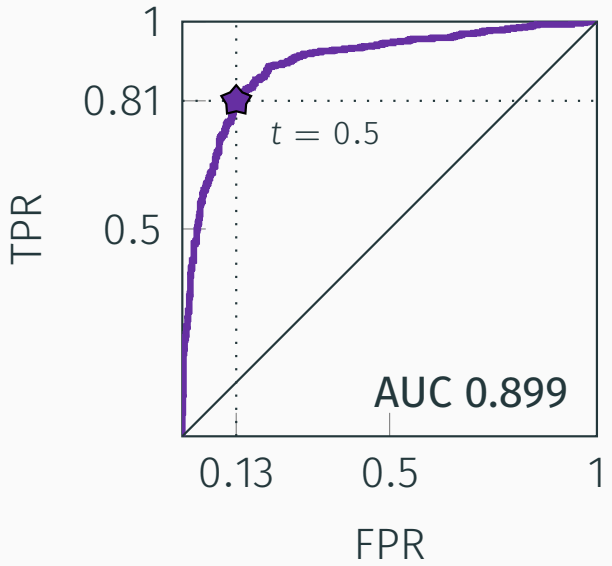
Out-of-sample predictions and relevance maps

Dementia: Modelling



Dementia: Predictive performance





		Predicted 0	Predicted 1
Observed	0	626	94
	1	138	582

Accuracy: 83.88%

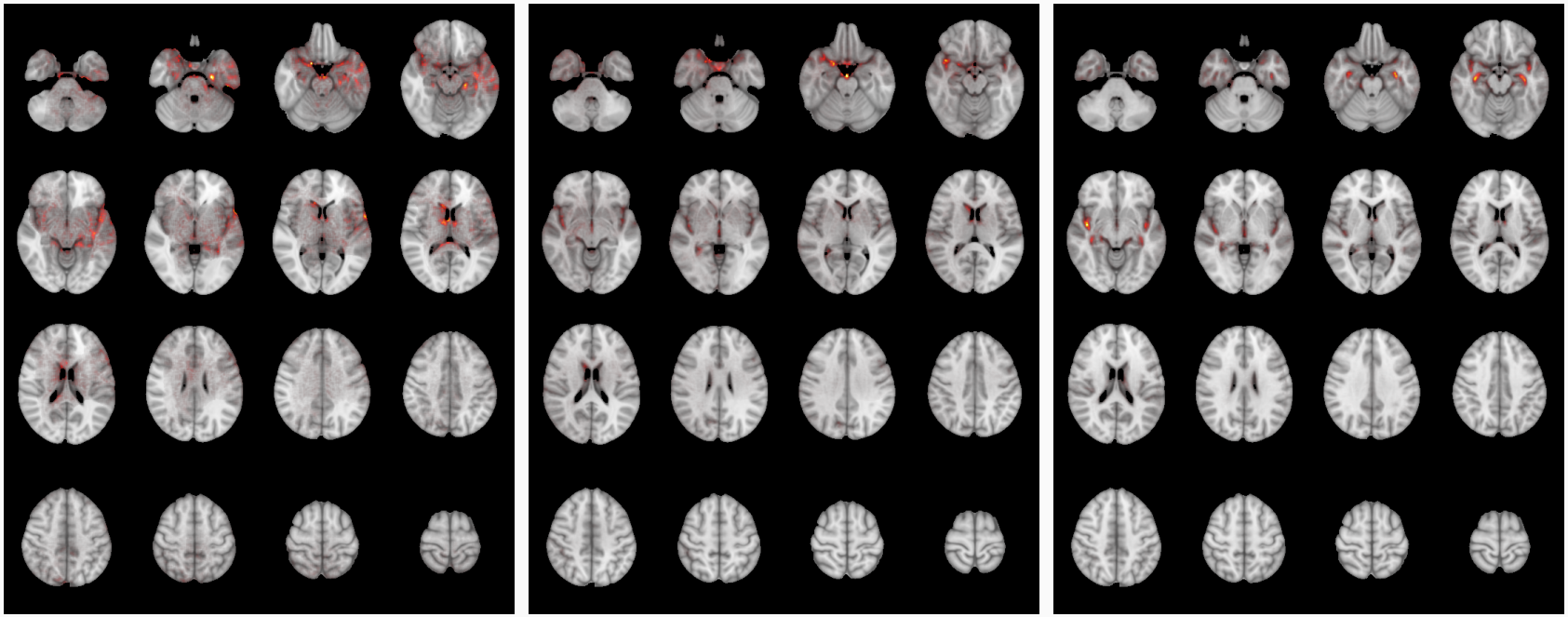
Dementia: Predictive performance

Site	Size	AUC	Accuracy
OASIS3 3.0T	430	0.841	76.9
ADNI 1.5T	300	0.915	87.0
ADNI 3.0T	266	0.951	88.3
Oslo GE750	210	0.915	82.8
AIBL Site 1	90	0.920	87.7
ANM GE	72	0.853	81.9
MIRIAD	32	1.00	100
AIBL Site 2	22	0.892	86.3
ANM Picker	10	0.840	80.0
OASIS3 1.5T	8	0.812	75.0

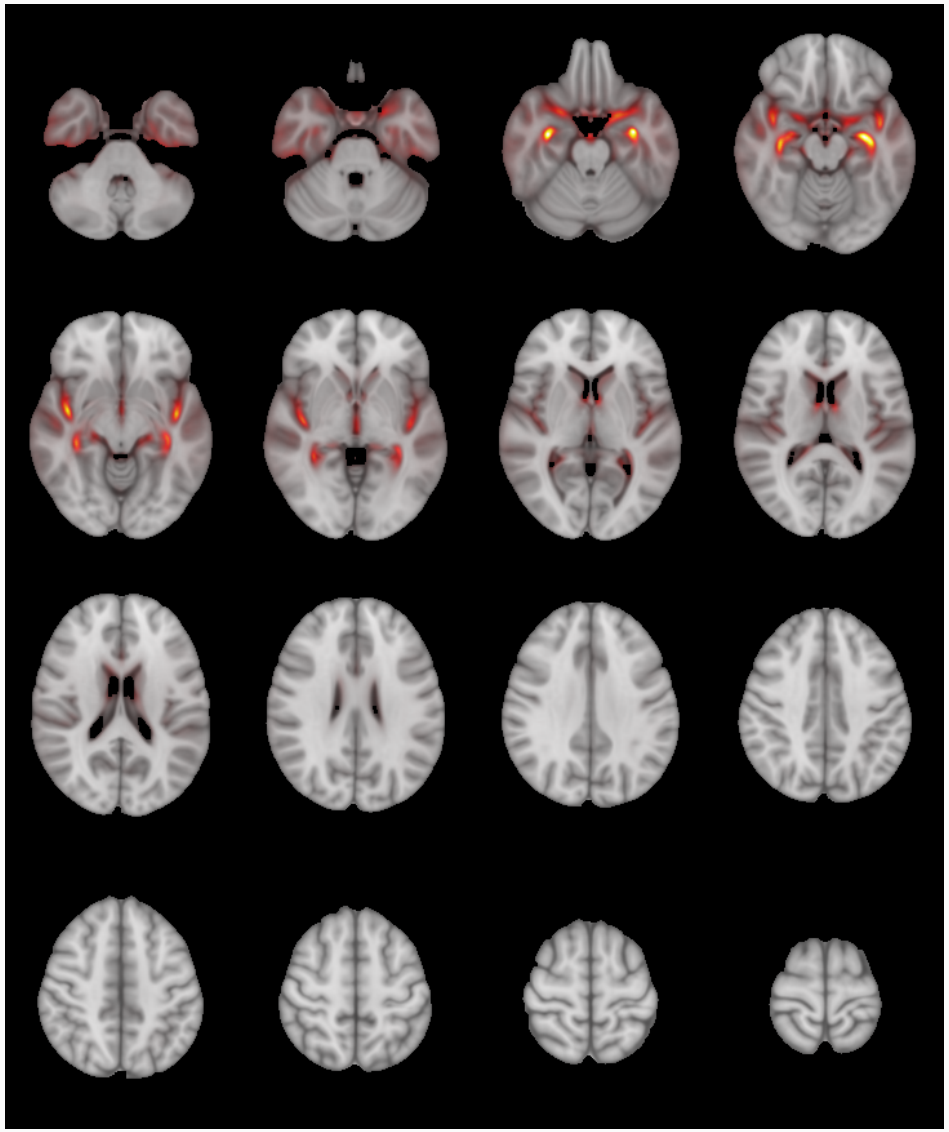
Dementia: Relevance maps

Layer	LRP Strategy
Input	-
Conv3D	{flat: True}
MaxPooling3D	-
Conv3D	{flat: True}
MaxPooling3D	-
Conv3D	$\{\alpha$: 1, β: 0$\}$
MaxPooling3D	-
Conv3D	$\{\alpha$: 1, β: 0$\}$
MaxPooling3D	-
Conv3D	$\{\alpha$: 1, β: 0$\}$
MaxPooling3D	-
Conv3D	$\{\alpha$: 1, β: 0$\}$
GlobalAveragePooling3D	-
Dropout	-
Dense	$\{\epsilon$: 0.25$\}$

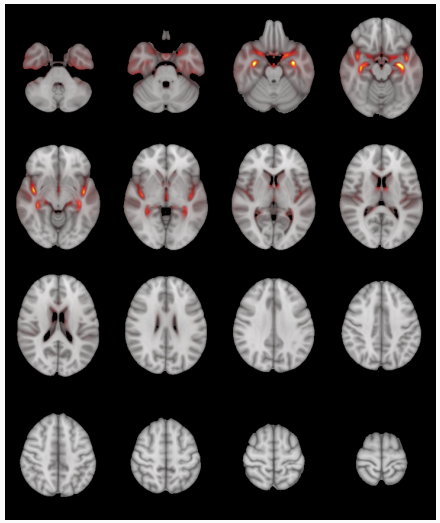
Dementia: Relevance maps



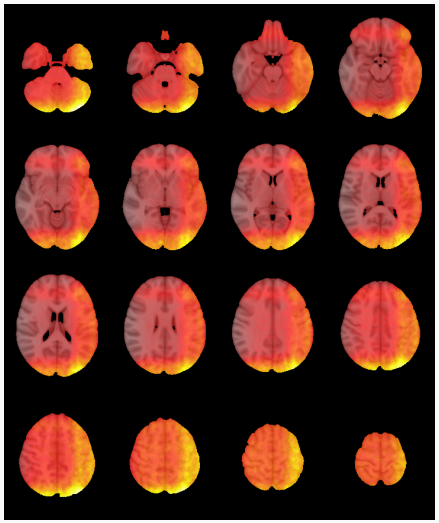
Dementia: Relevance maps



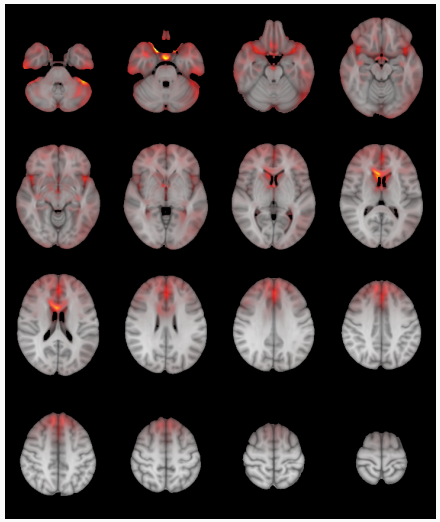
Dementia: Relevance maps



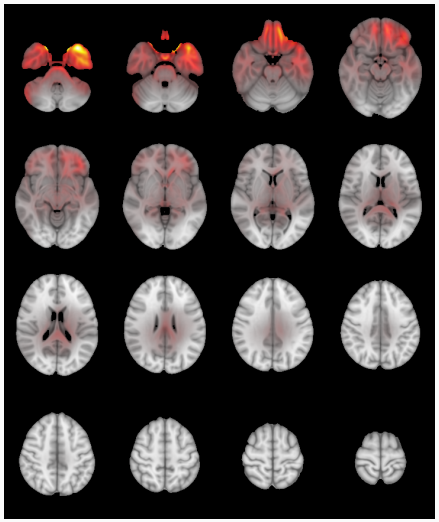
Dementia model



Dementia model with randomized images

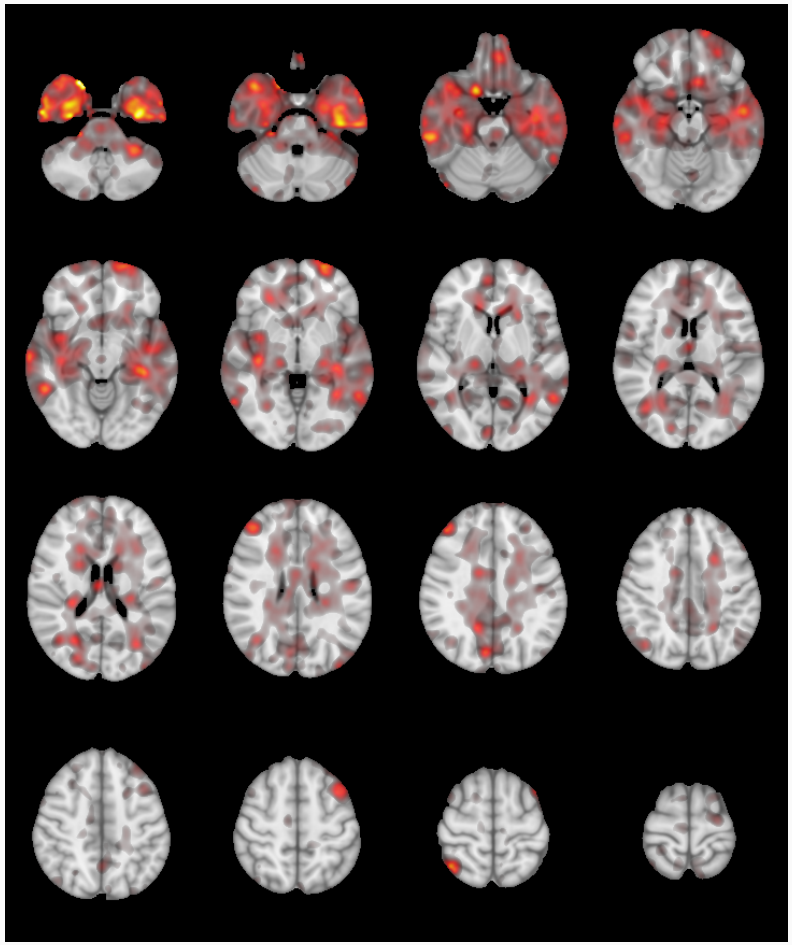


Sex model



Model with randomized weights

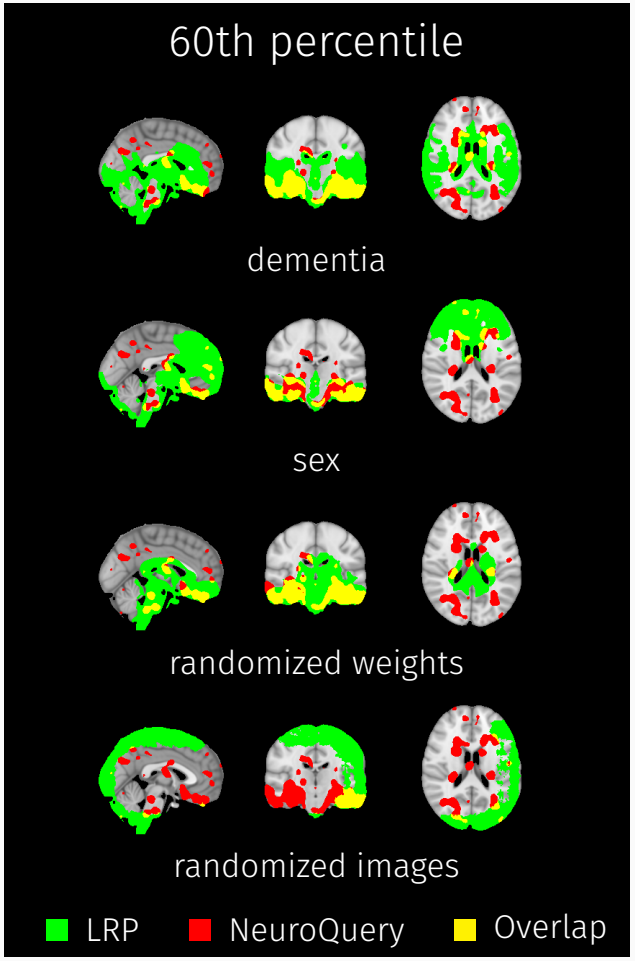
Dementia: Relevance maps



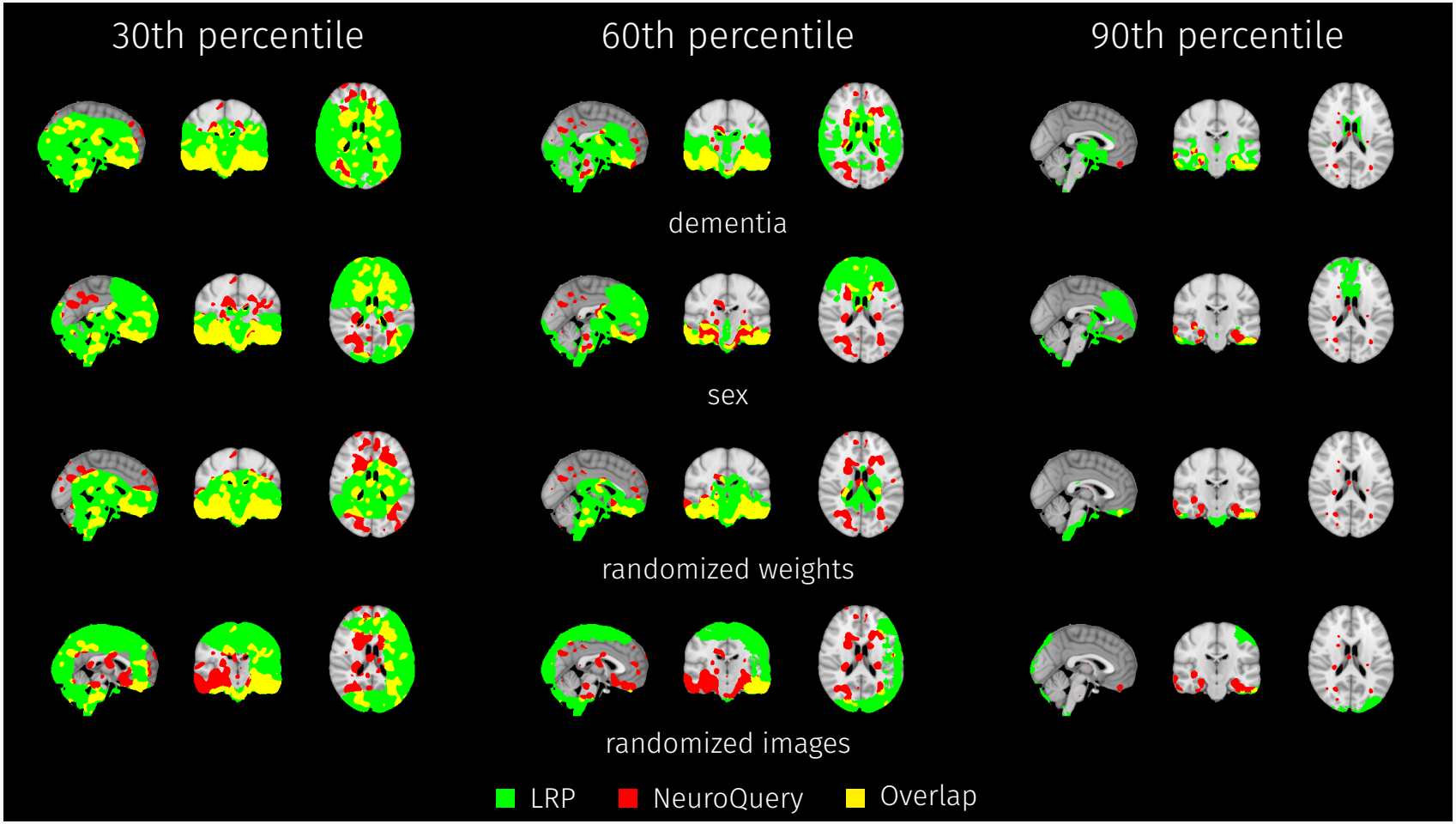
Neuroquery

https://neuroquery.org/

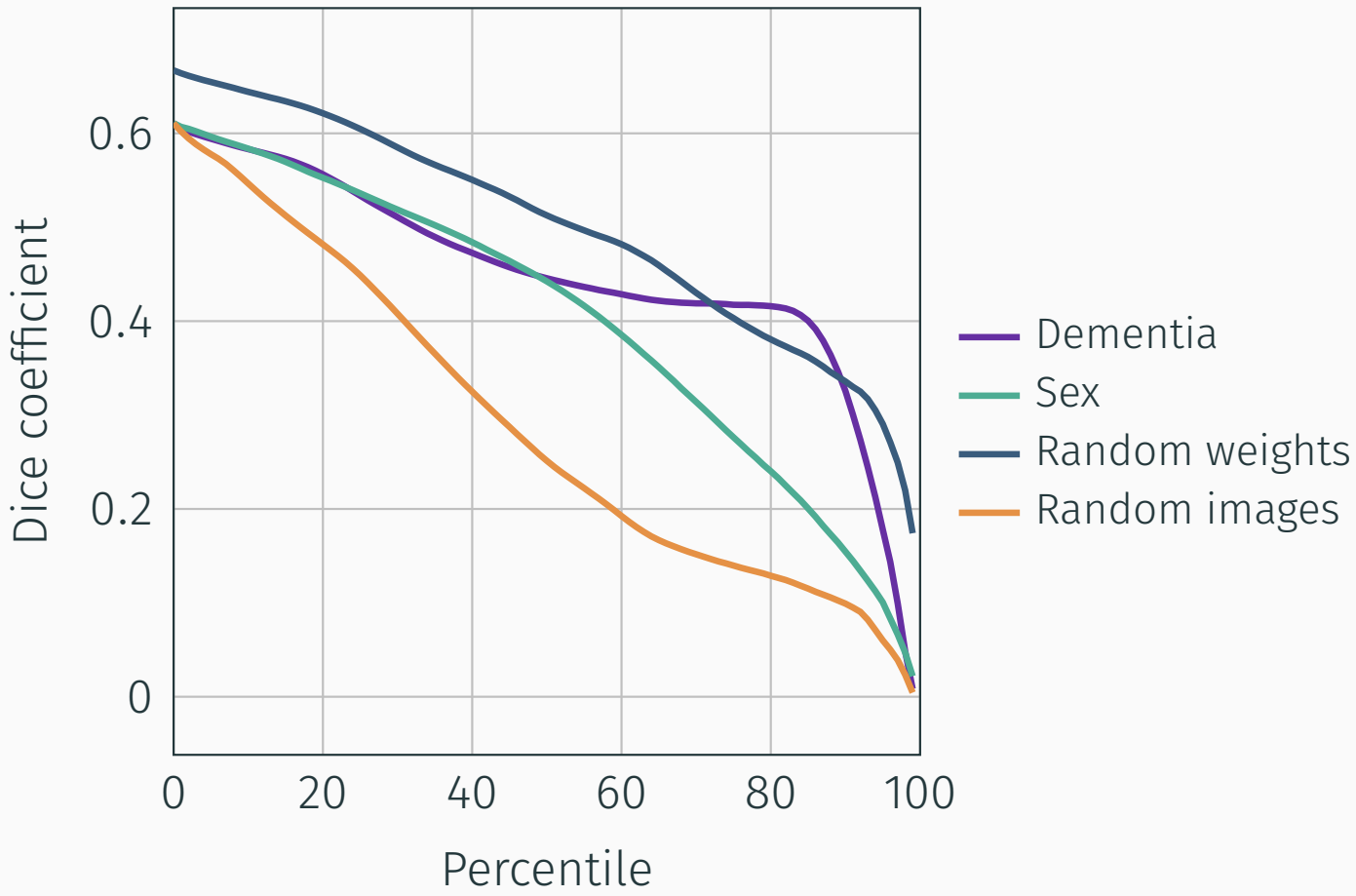
Dementia: Relevance maps



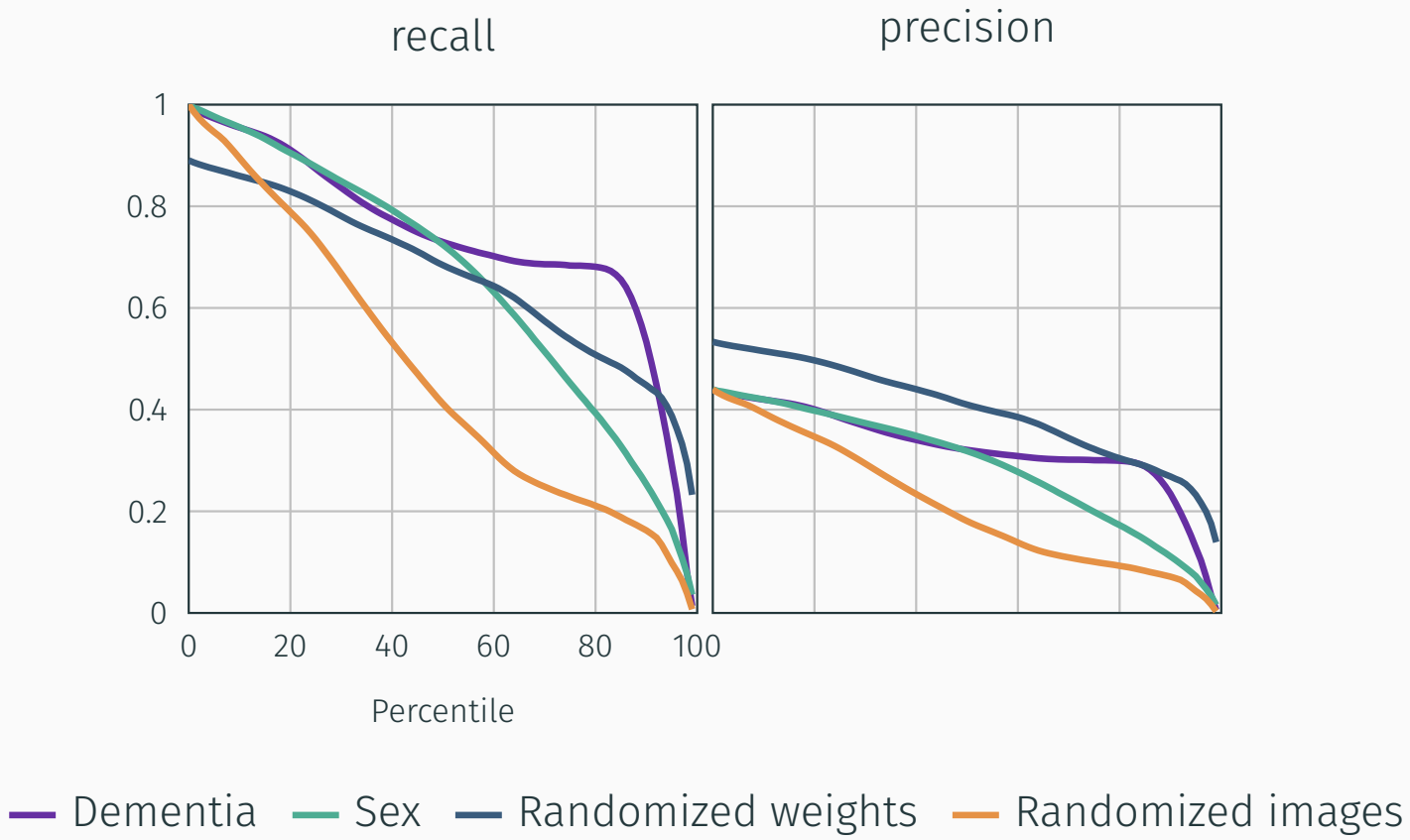
Dementia: Relevance maps



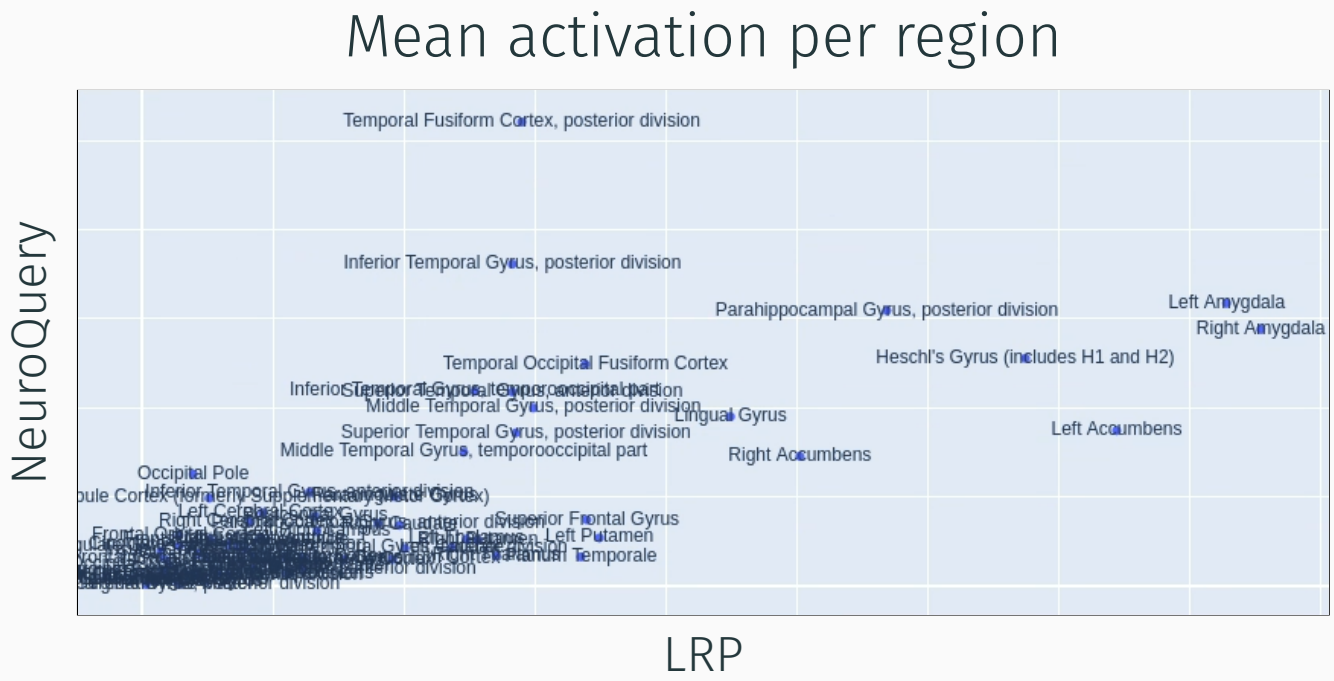
Dementia: Relevance maps



Dementia: Relevance maps

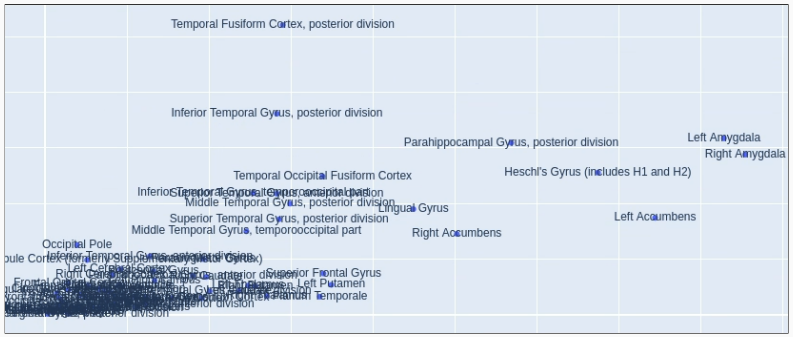


Dementia: Relevance maps

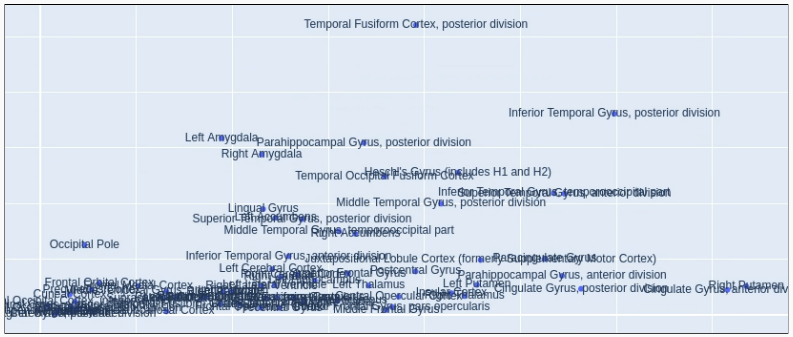


Dementia: Relevance maps

Dementia

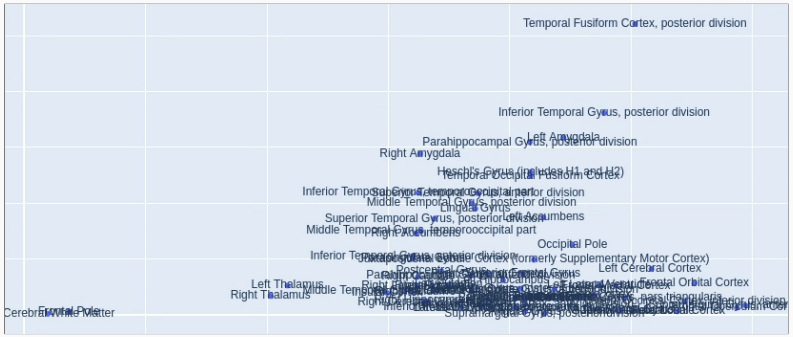


Sex



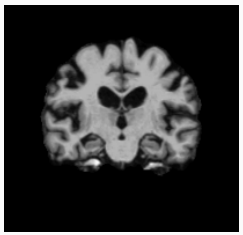
Randomized weights

Randomized images



Dementia: Relevance maps

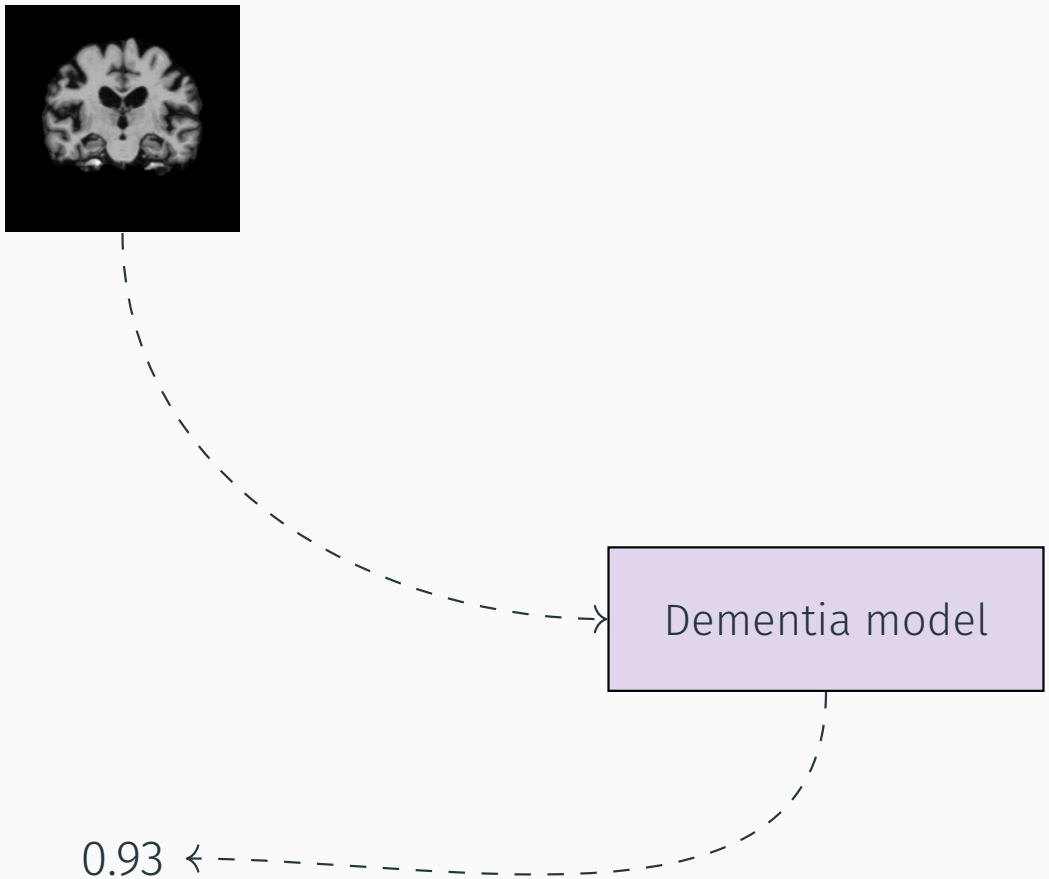
Iteration 0



Dementia model

Dementia: Relevance maps

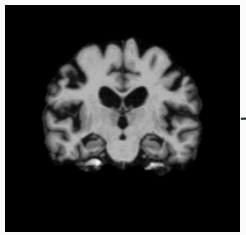
Iteration 0

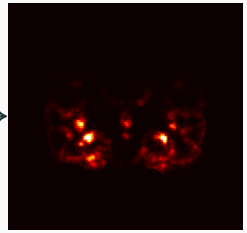


Dementia: Relevance maps

Iteration 0

Validation



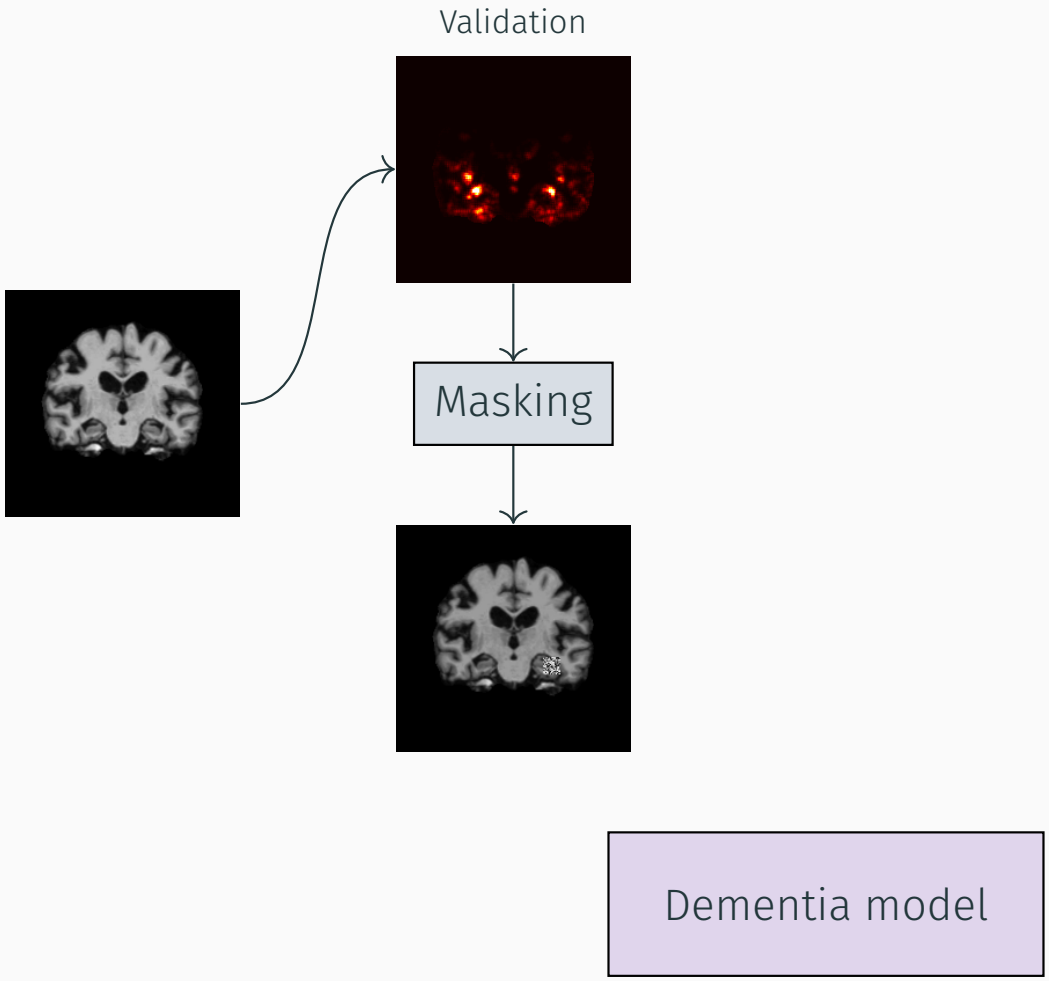


Dementia model

0.93

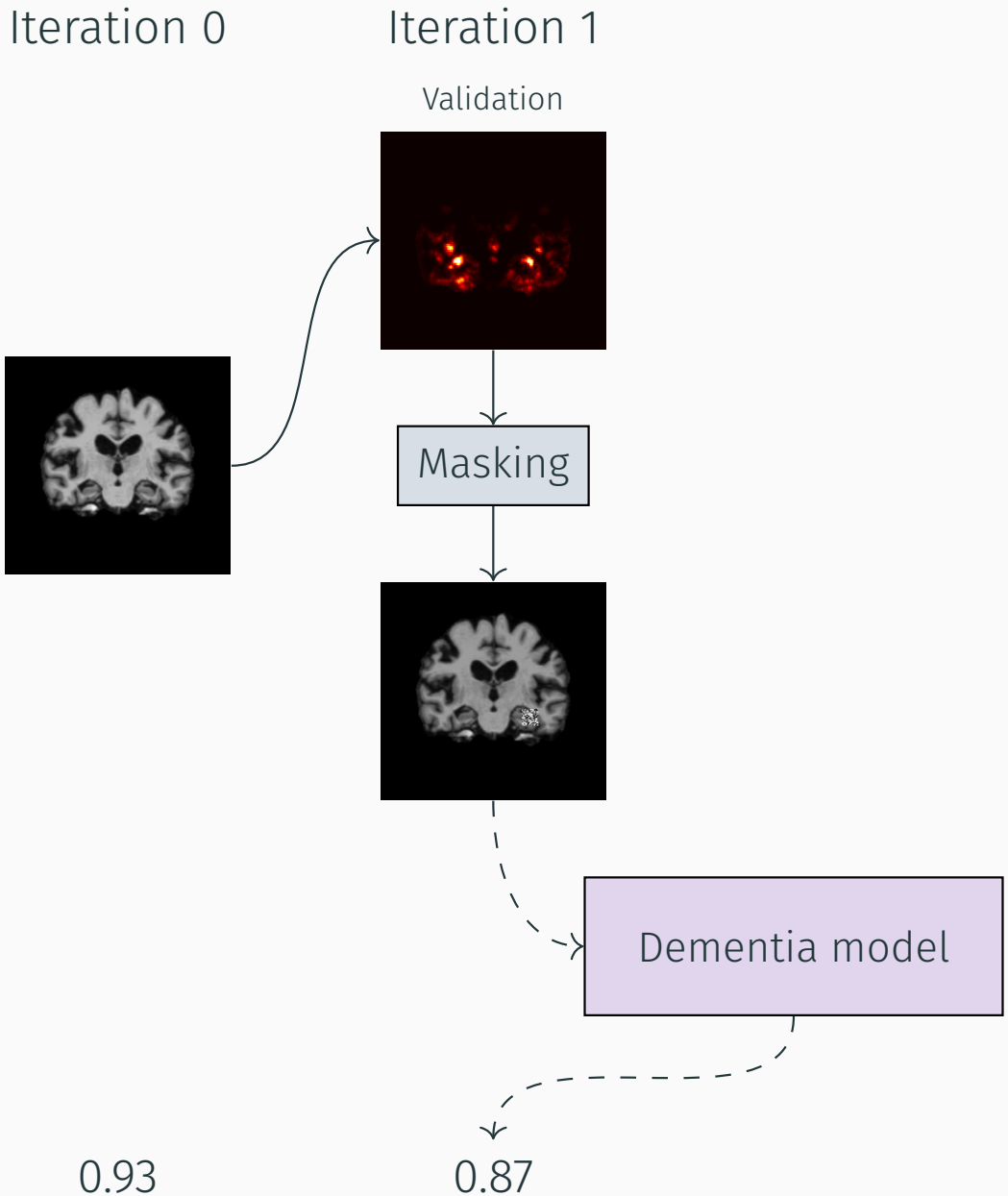
Dementia: Relevance maps

Iteration 0

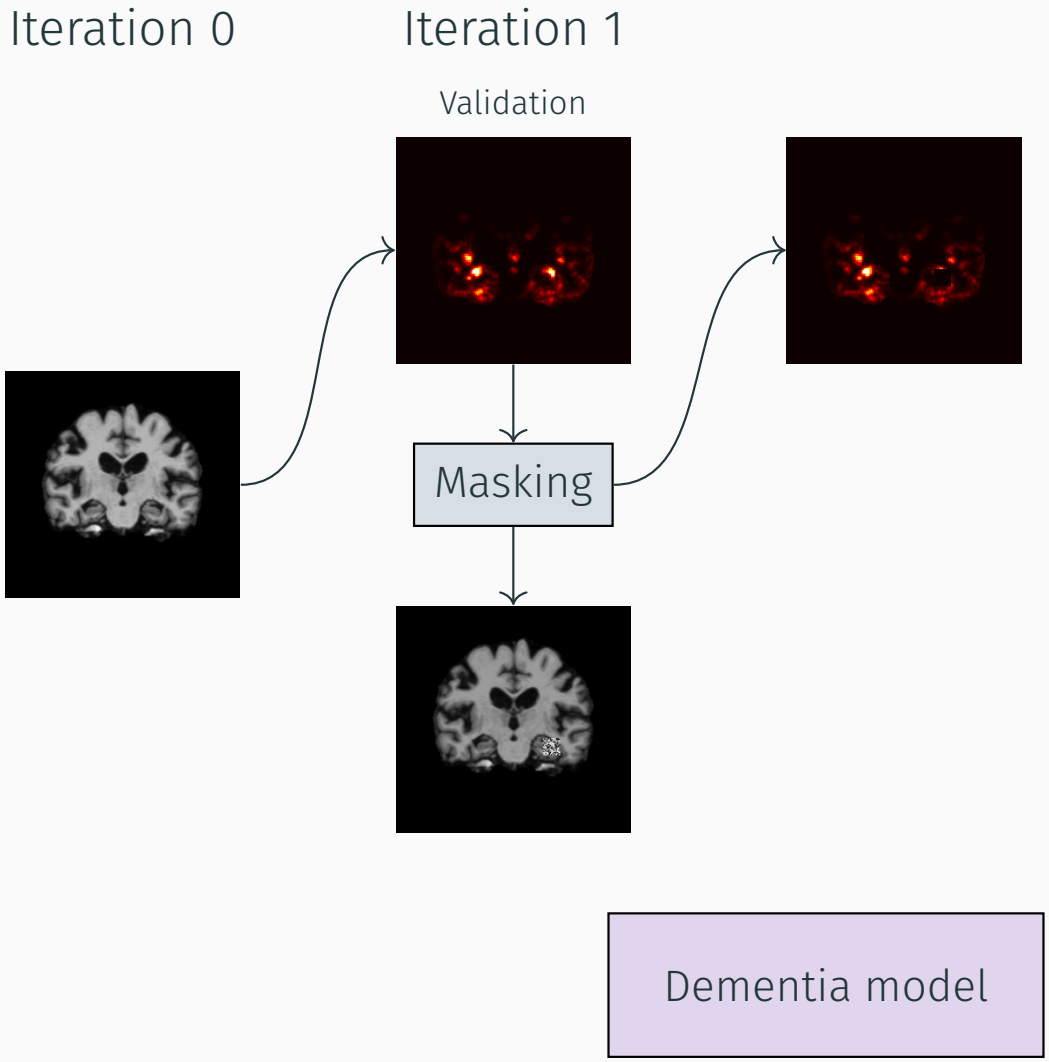


0.93

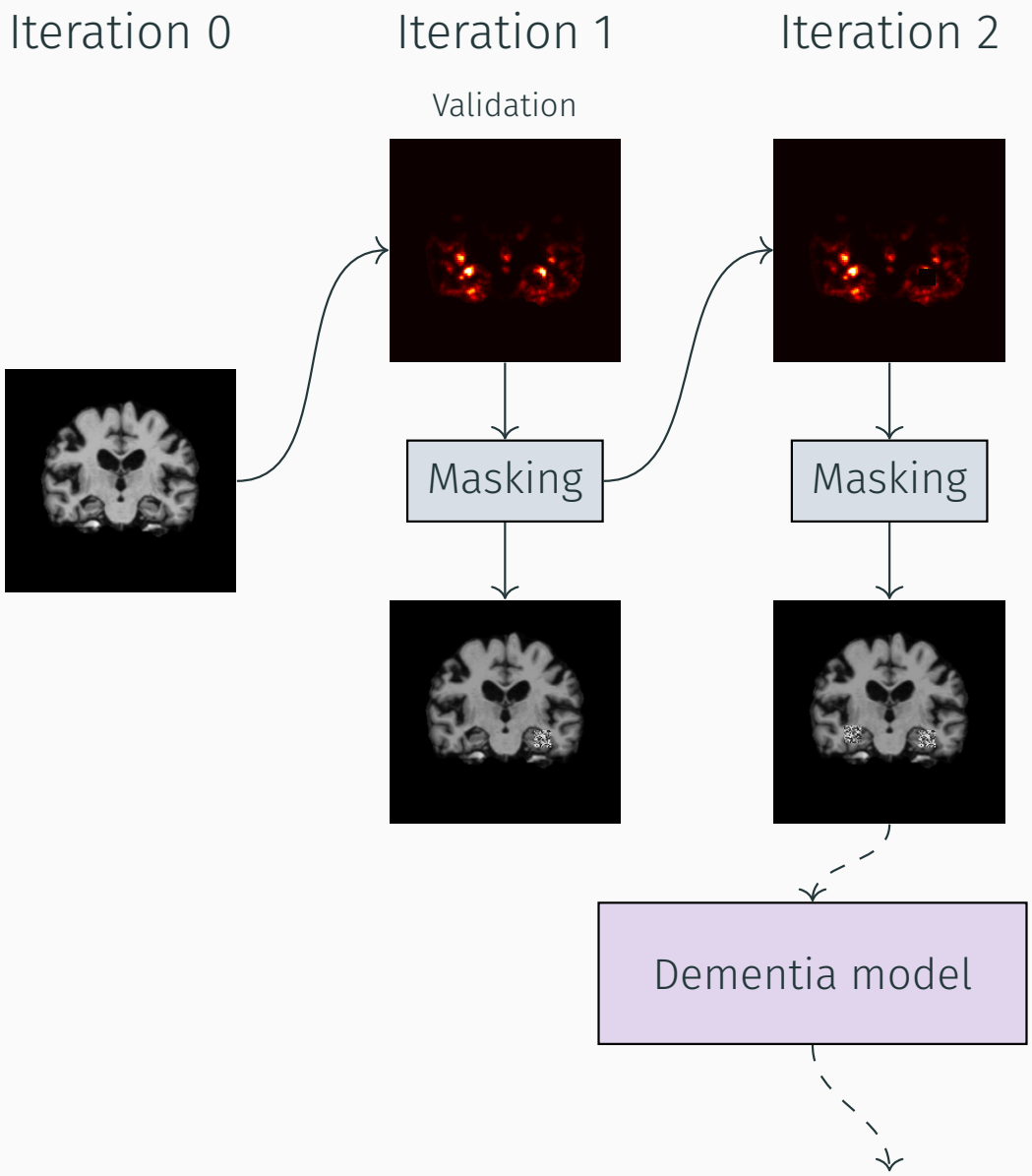
Dementia: Relevance maps



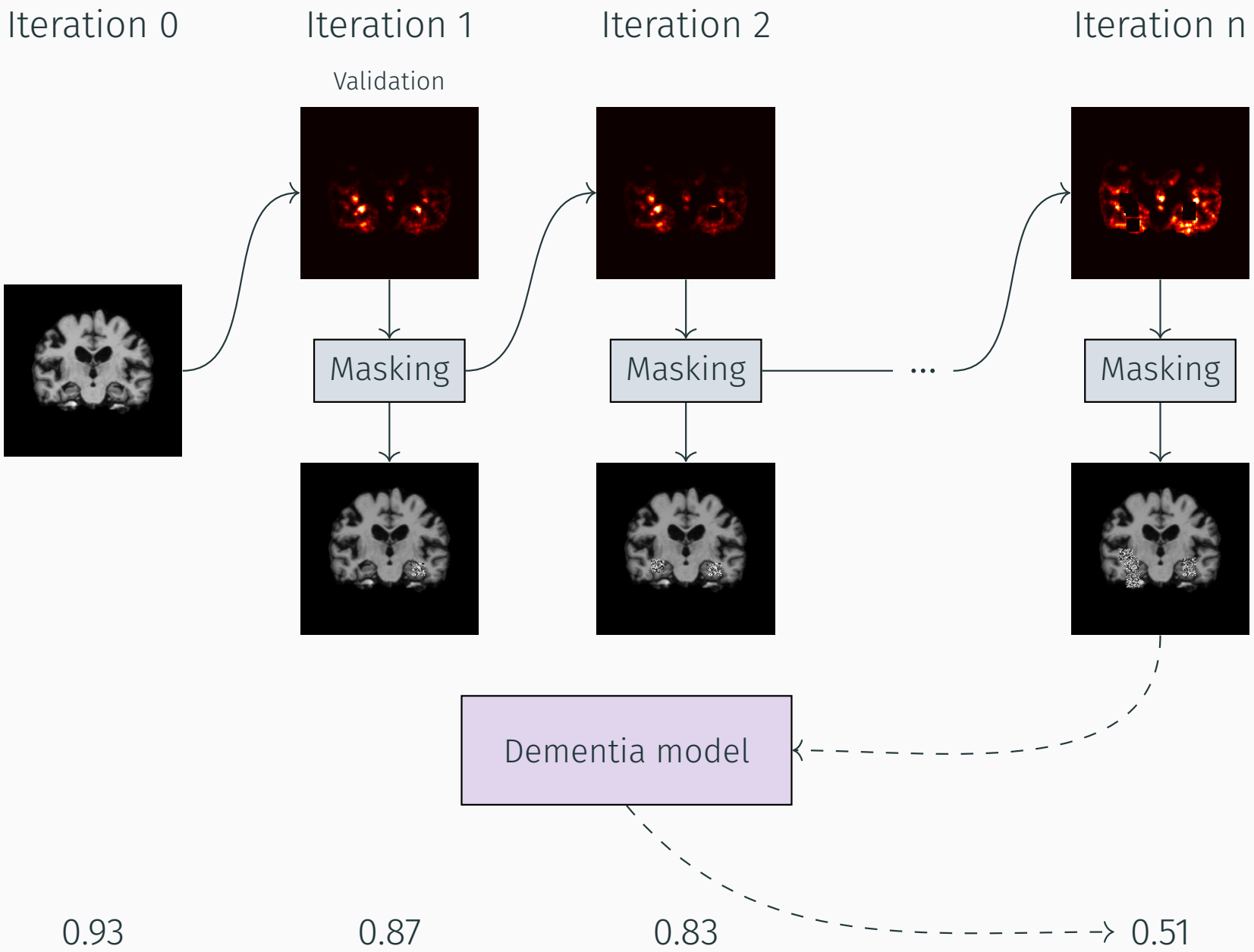
Dementia: Relevance maps



Dementia: Relevance maps



Dementia: Relevance maps



Dementia: Relevance maps