Making sense of data and models

AIMS Lecture Series

AI in Medicine and Science

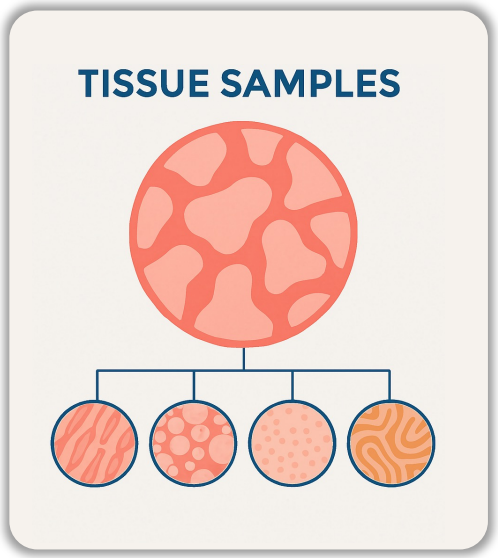
Nils Philipp Walter

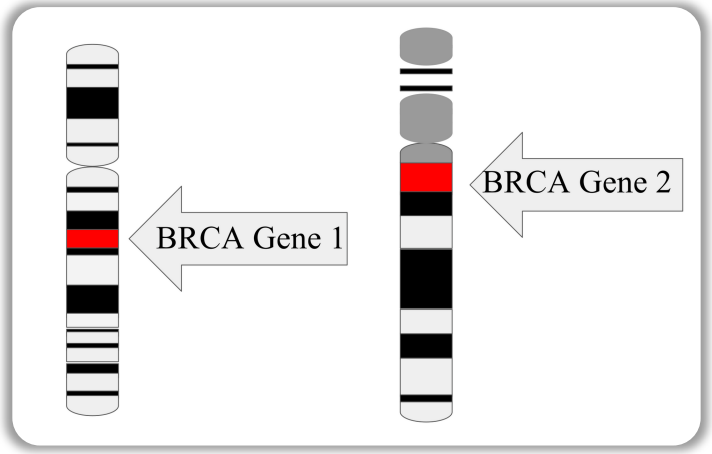
14.04.2025

Common questions in biology

Disclaimer: I am computer scientist

What is the difference between cancerous and benign tissue?



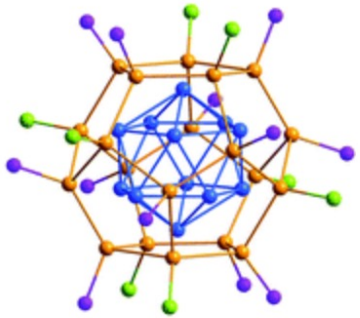


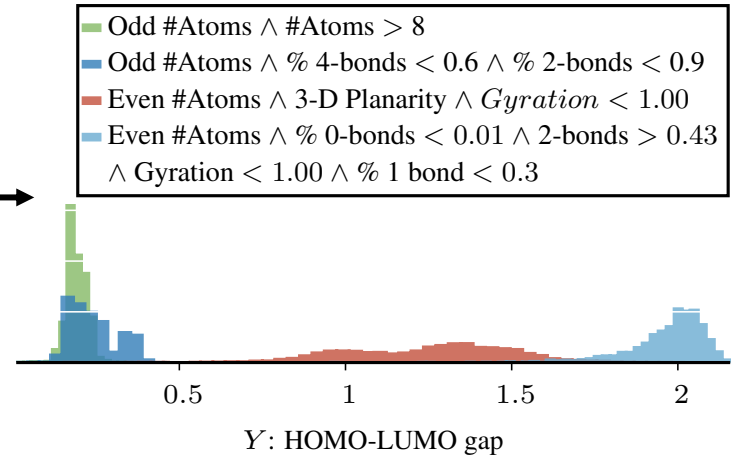
Ref. [1,2,3]

Common questions in physics

Disclaimer: I am computer scientist

What materials have exceptional conductivity?

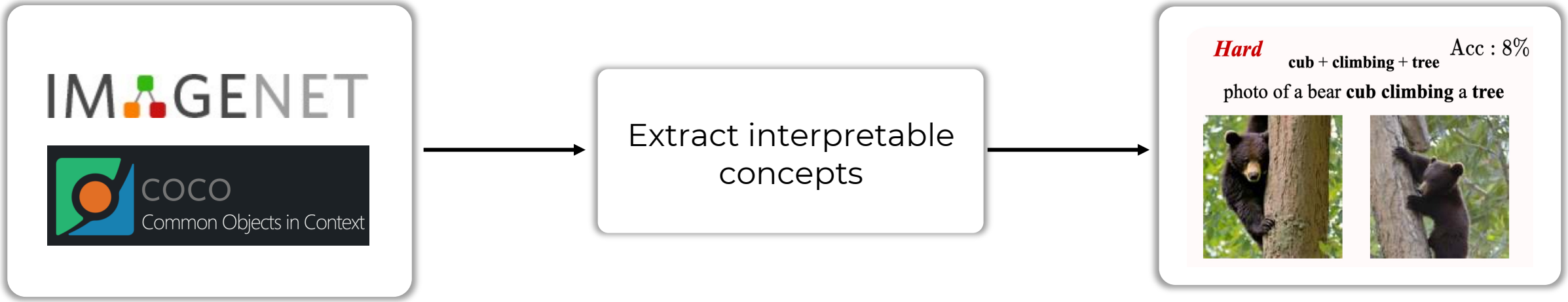




Ref. [4]

Common questions in explainability

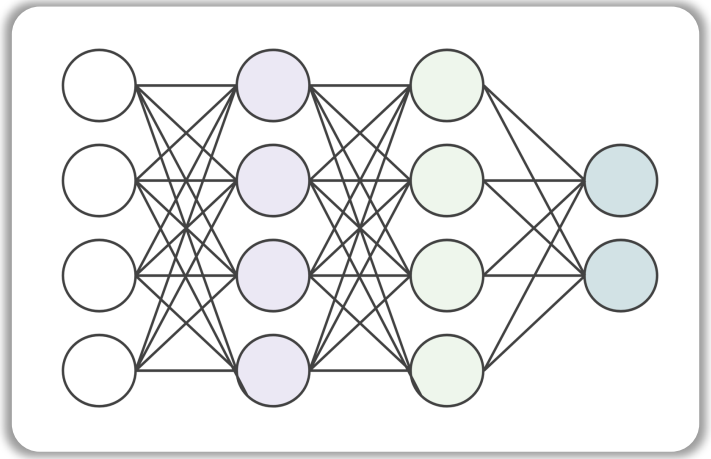
What are common failure modes and biases?



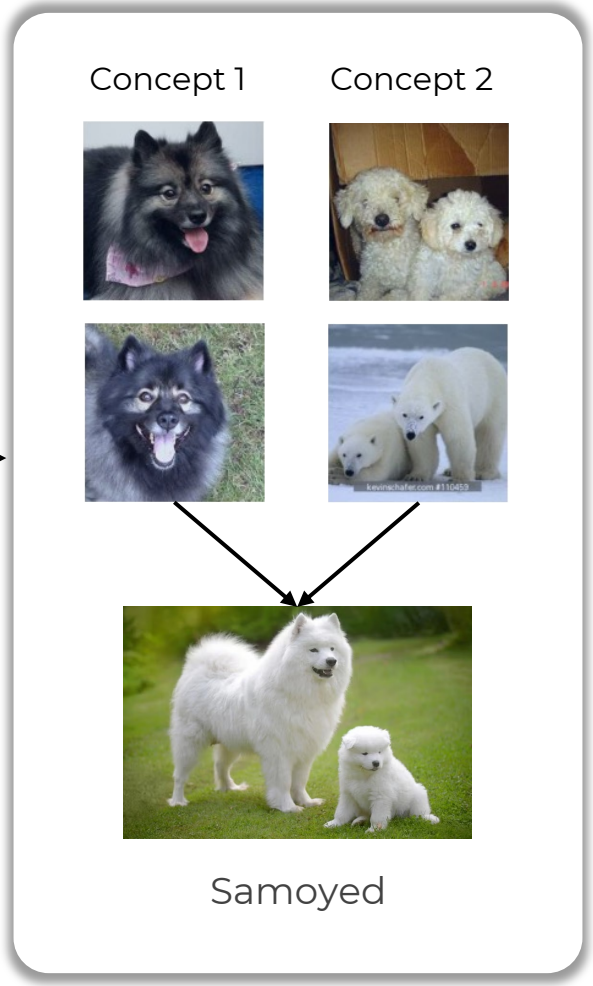
Ref. [5,6,7]

Common questions in interpretability

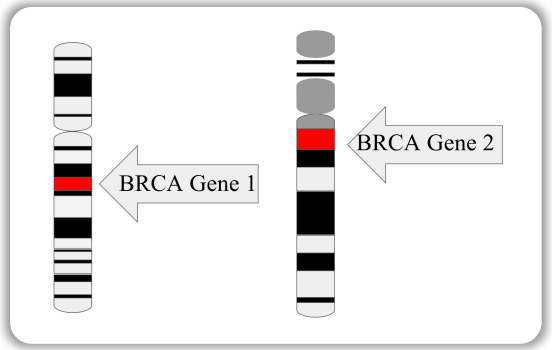
What internal reasoning influences predictions?

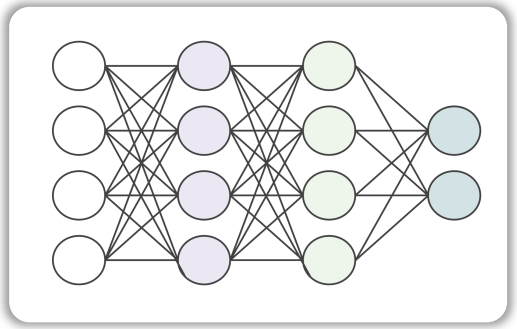


Extract interpretable concepts



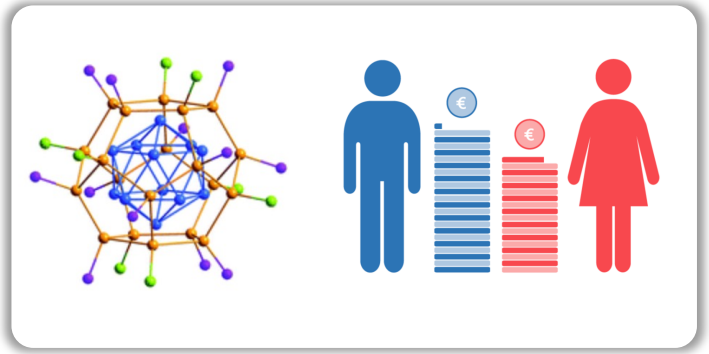
Common questions in ML

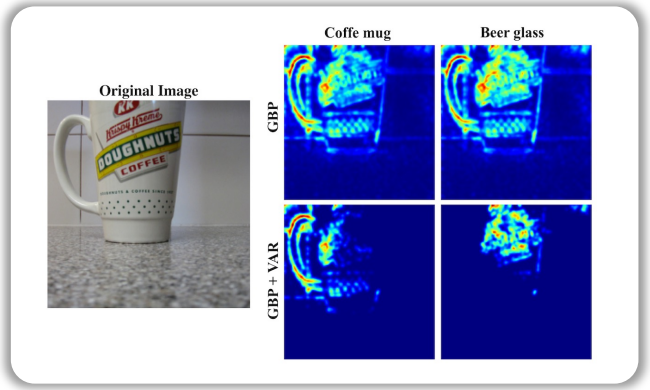




Data

Models

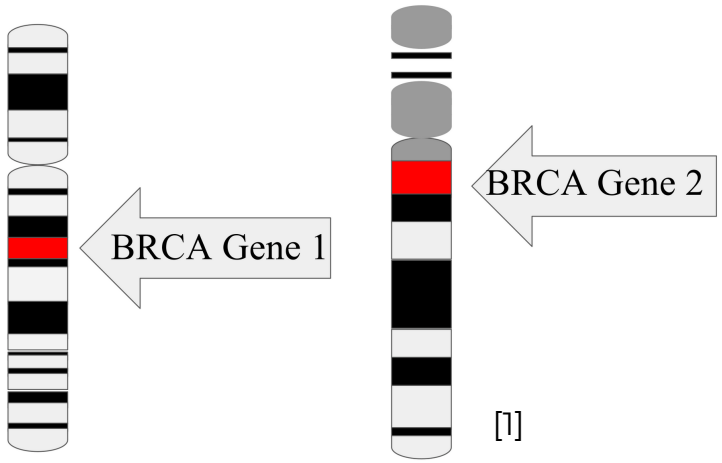




Understanding genetic data

Breast Cancer Data (BRCA)

- (Binarized) sequences of tissue samples
- Labeled with types of BRCA
- Very high-dimensional, low #samples

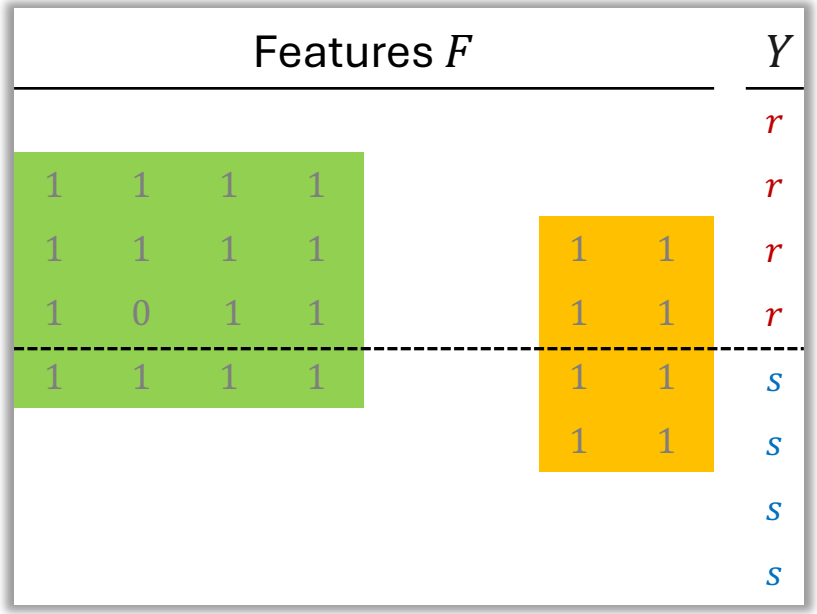


[1]

Goal: Find class-specific interactions of genes

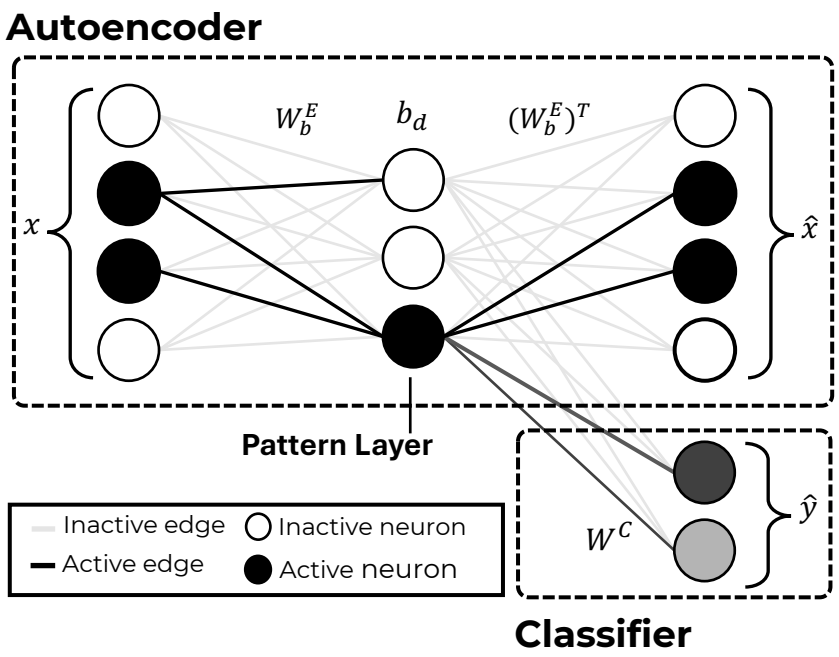
Class-specific pattern:

1. Occurs more often in class k
2. Is most predictive for class k



DiffNaps – In a nutshell

Main Idea: Compression + Classification → Class-specific pattern



Forward Pass

1. Binary input x
2. Binarize weight matrix W_b^E
3. Compute hidden activation z

$$z = \lambda_E(W_b^E\, x)\,,$$

where λ_E is a binary activation function.

4. Compute reconstruction $\hat{x}$

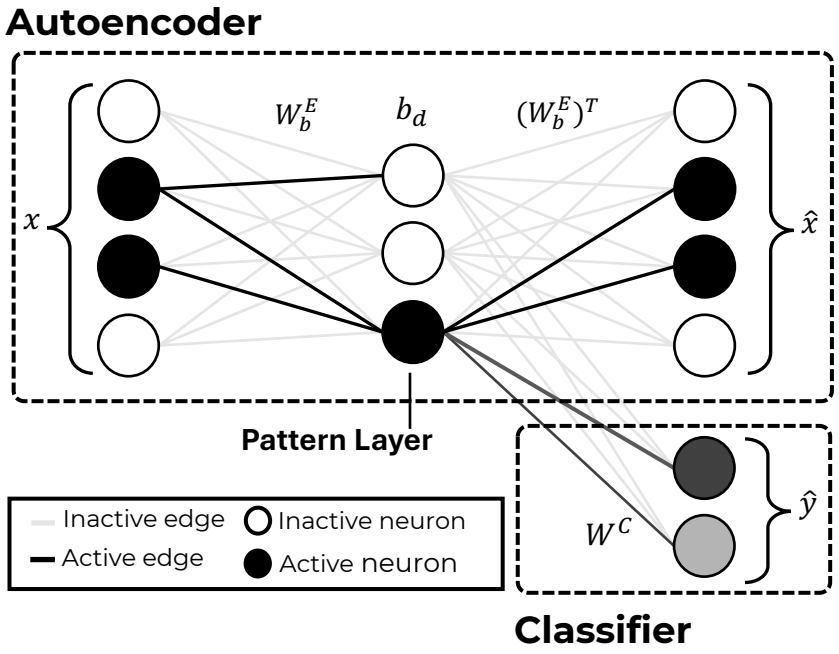
$$\hat{x} = \lambda_D((W_b^E)^T z)$$

5. Compute classification $\hat{y}$

$$\hat{y} = \mathrm{softmax}(W^C z)$$

Jointly optimize reconstruction and classification accuracy

DiffNaps – Extract patterns



Continuous **Discrete**

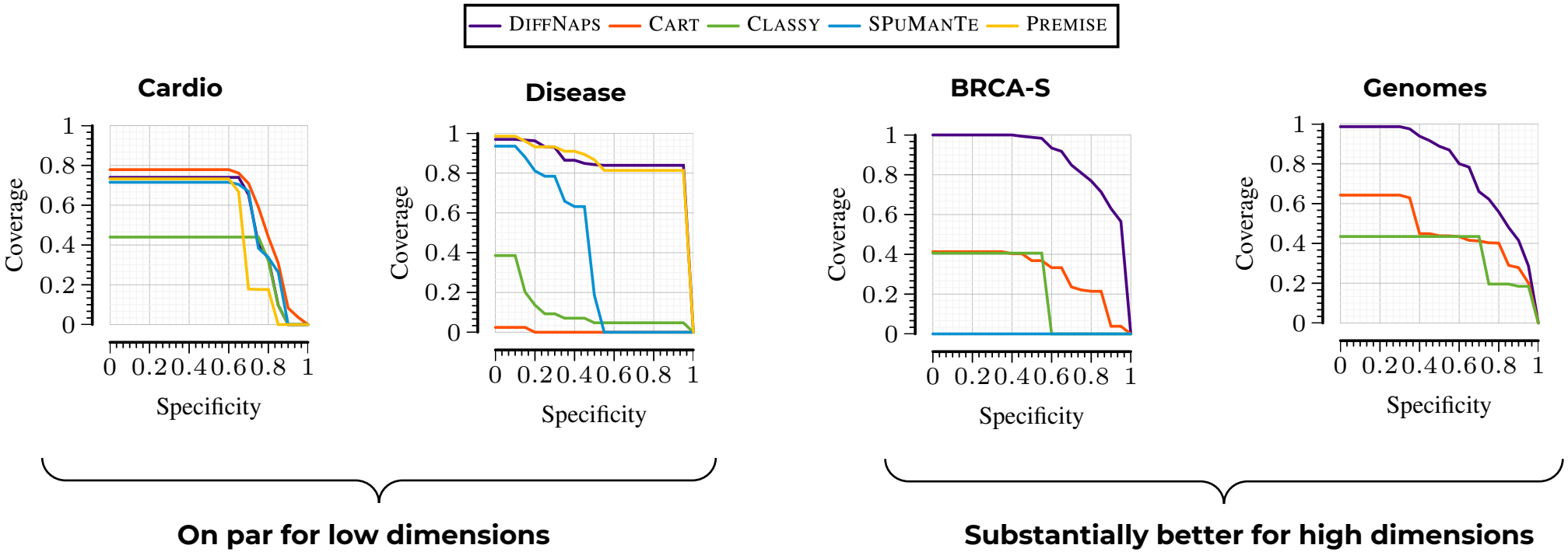
$$W^E = \begin{bmatrix} 0.02 & 0.87 & 0.02 & 0.87 \\ 0.8 & 0.01 & 0.04 & 0.9 \\ 0.0 & 0.99 & 0.8 & 0.09 \end{bmatrix} \xrightarrow{\mathcal{B}\tau_E(\cdot)} W_b^E = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

$$W^C = \begin{bmatrix} 0.07 & 0.01 & 0.9 \\ 0.02 & 0.9 & 0.2 \end{bmatrix} \xrightarrow{\mathcal{B}_{\tau_C}(\cdot)} W_b^C = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Class-specific patterns: $P^1 = \{\{2,3\}\}$ resp. $P^2 = \{\{1,4\}\}$

DiffNaps – Real World Data

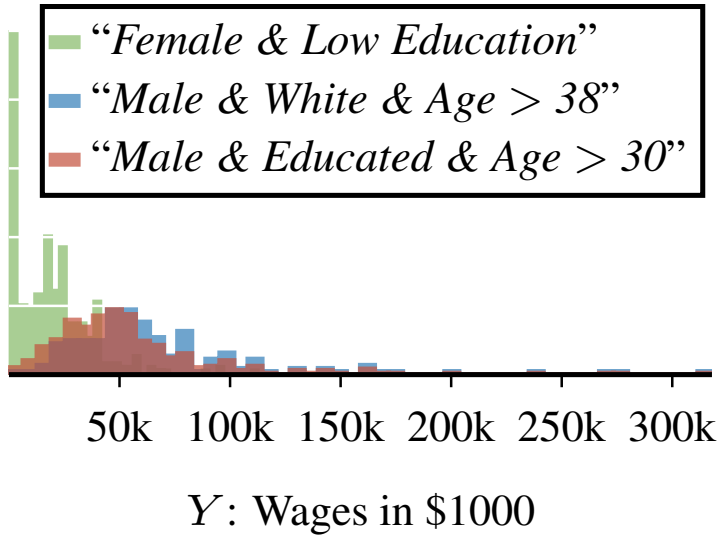
- **Cardio** ($m = 45$): Patient data annotated with heart disease [8]
- **Disease** ($m = 131$) : Various symptoms annotated with disease [9]
- **BRCA-S** ($m = 20k$): ncRNA annotated with cancer type [10] → Found biologically **relevant** patterns!
- **Genomes** ($m = 225k$): DNA annotated with ethnicity [11]



Discovering exceptional subgroups

Census Data

Sex	Height	Ethn	Education	Age	Income
♀	168	White	12	72	17k
♂	163	White	11	55	23k
♀	160	White	**5**	62	**1k**
♂	188	White	16	38	63k
♀	165	White	**9**	45	**4k**
♂	172	White	12	78	71k
♀	180	White	**8**	74	**1k**



Task – Subgroup Discovery:

1. Find **exceptional** subgroups
2. With an **interpretable** description

SyFlow – In a nutshell

Subgroup Discovery

1. Dependent on Pre-Discretization

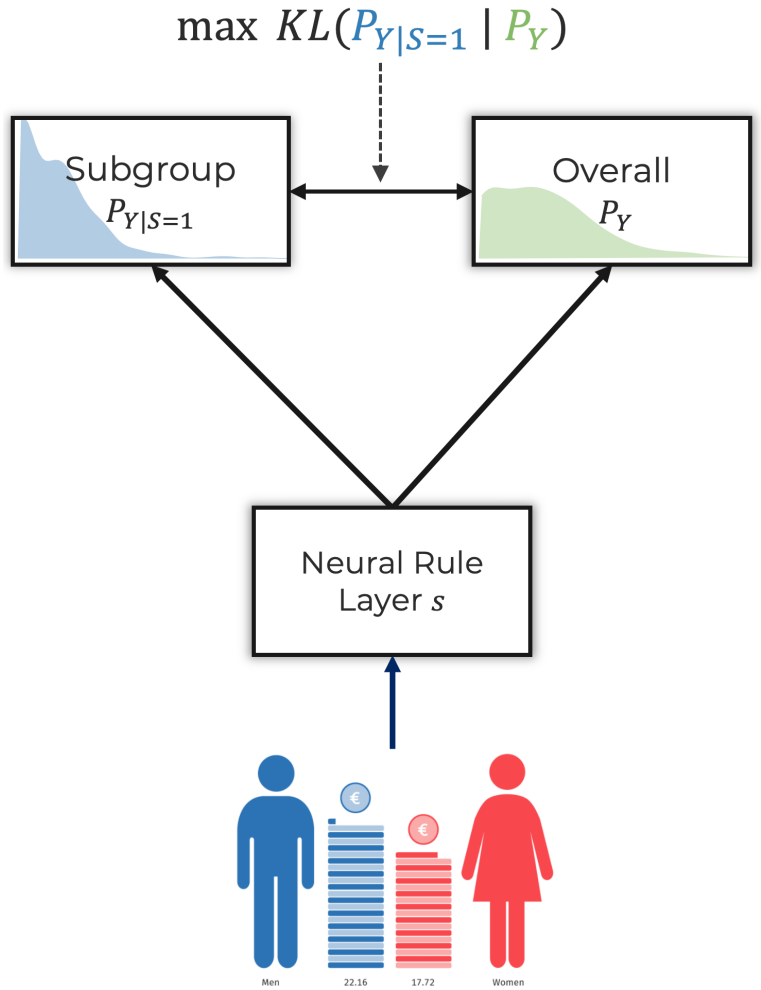
2. Strong assumptions on the target distribution

3. Combinatorial optimization

1. Learn predicates end-to-end
   → **Accurate Discretization**

2. Use Normalizing Flows (NFs)
   → **No assumptions**

3. Continous optimization
   → **Highly scalable**



Ref. [16]

SyFlow – Neural Rule Layer I

Goal: Find an crisp interpretable description

$$\sigma(x) = \neg Smoker \wedge 44 < Age < 64$$

Ingredients:

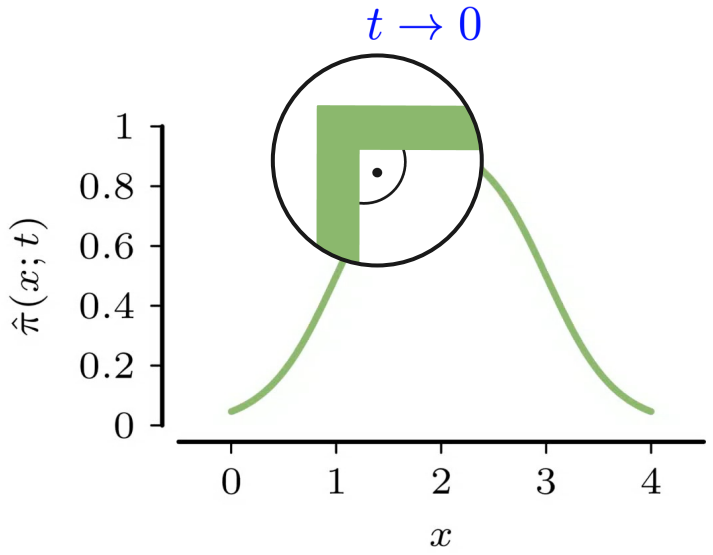
1. Differentiable binning predicate

$$\hat{\pi}(x_i; \alpha_i, \beta_i, \boldsymbol{t}) = \frac{e^{\frac{1}{t}(2x_i - \alpha_i)}}{e^{\frac{1}{t}x_i} + e^{\frac{1}{t}(2x_i - \alpha_i)} + e^{\frac{1}{t}(3x_i - \alpha_i - \beta_i)}}$$

- Differentiable analog of:

$$\pi(x_i; \alpha_i, \beta_i) = \begin{cases} 1 & \text{if } \alpha_i < x_i < \beta_i \\ 0 & \text{otherwise} \end{cases}$$

- Temperature t controls crispness



Theorem 1 *Given its lower and upper bounds $\alpha_i, \beta_i \in \mathbb{R}$, the soft predicate of Eq. (1) applied on $x \in R$ converges to the crisp predicate that decides whether $x \in (\alpha, \beta)$,*

$$\lim_{t \to 0} \hat{\pi}(x_i; \alpha_i, \beta_i, t) = \begin{cases} 1 & \text{if } \alpha_i < x_i < \beta_i \\ 0.5 & \text{if } x_i = \alpha_i \vee x_i = \beta_i \\ 0 & \text{otherwise} \end{cases} .$$

(Yang et. al. 2018)

SyFlow – Neural Rule Layer II

Ingredients:

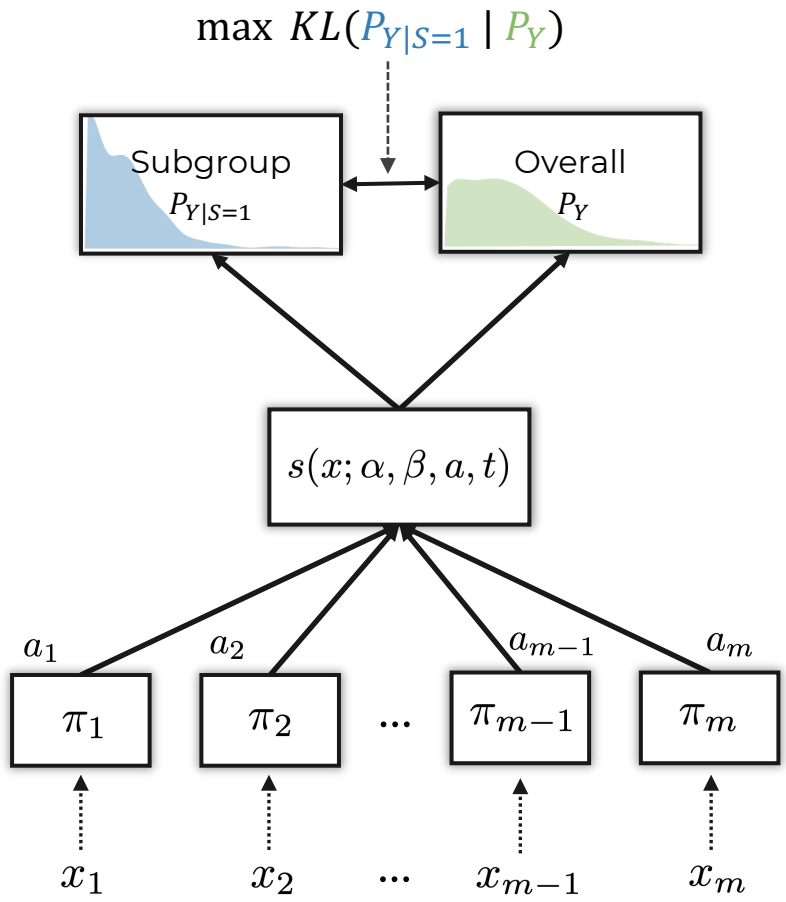
1. Differentiable binning predicate
2. Differentiable logical *AND*
   - Harmonic means behaves like an *AND*
     1. If one π_i is zero then evaluate to false
     2. If all π_i are one then evaluate to true
   - Implicit feature selection with a_i

Optimization

1. Learn the overall distribution P_Y
2. Learn the subgroup distribution $P_{Y|S=1}$
3. Optimize classifier weights and bins

} Repeat for N steps

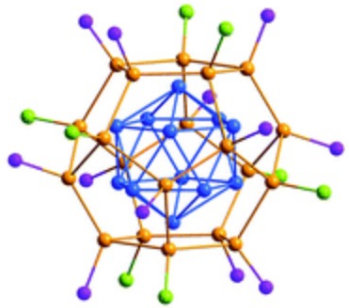
4. Output: Subgroup



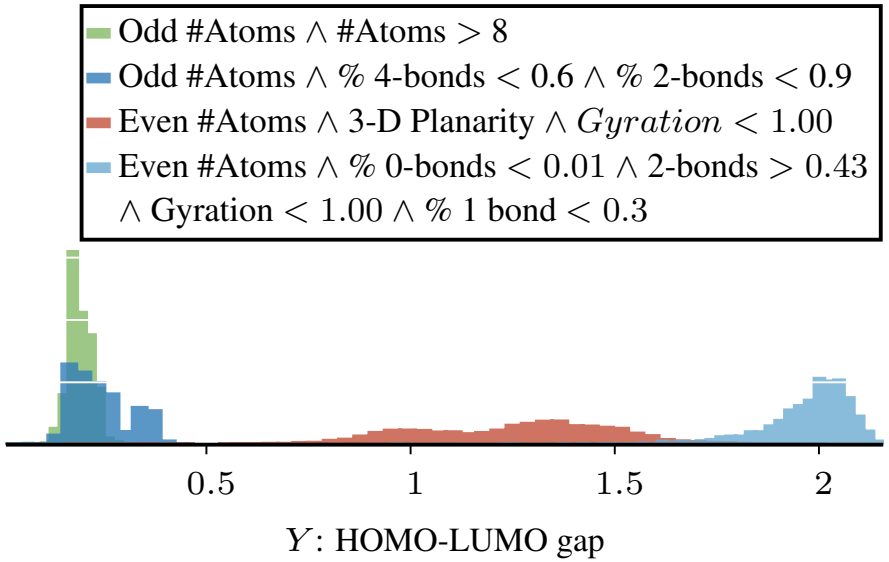
Fully differentiable!

Experiments – Materials Sciences

Gold Nanoclusters



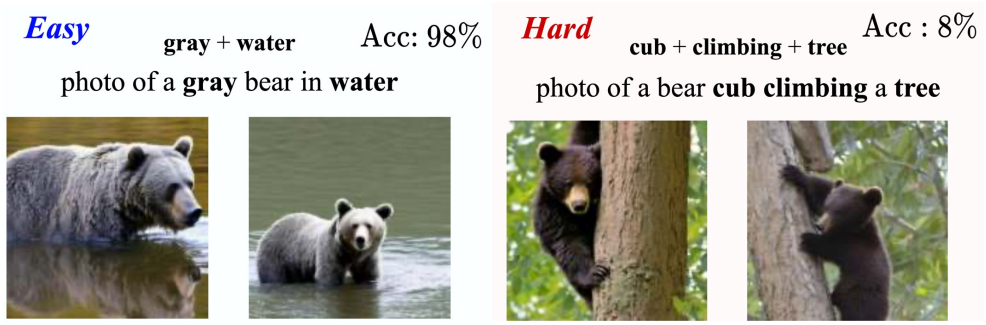
- Number of Atoms
- Even #Atoms
- 3-D Planarity



Target: HOMO-LUMO gap
~ stability and conductivity

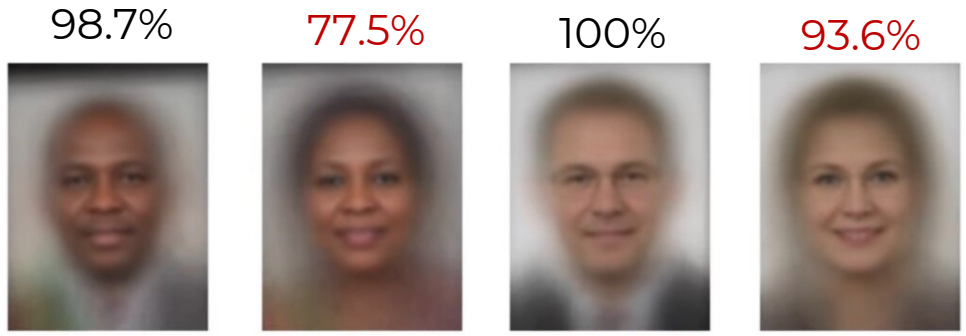
From Nanoclusters to Analyzing Model Outputs

When can predictions be trusted?

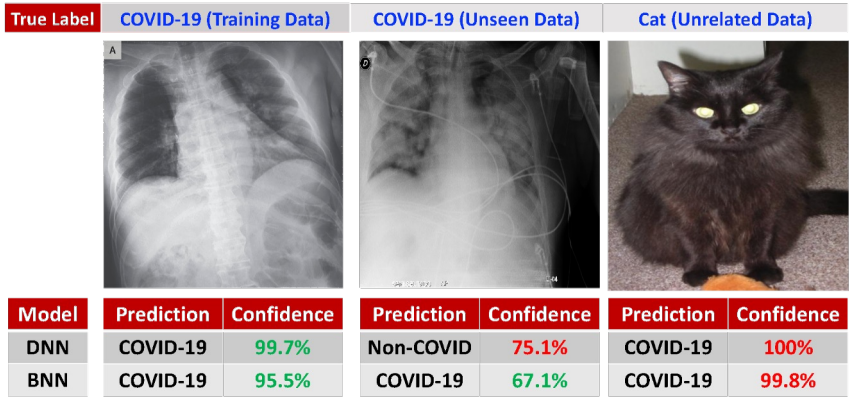


Aligning with instructions

Does the model act fair?



Ehm ... what?

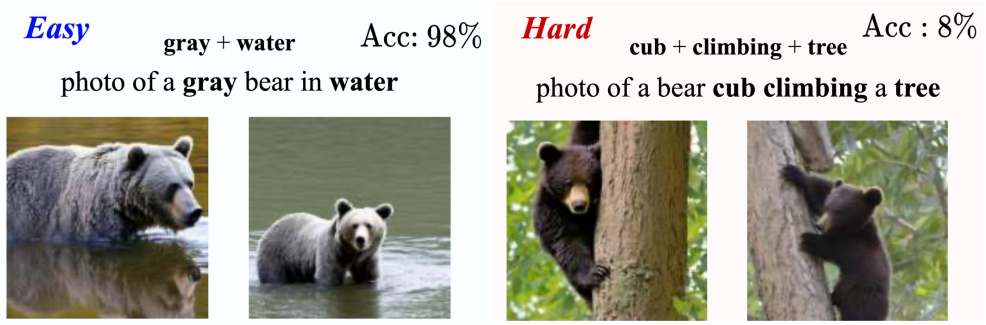


True Label	COVID-19 (Training Data)		COVID-19 (Unseen Data)		Cat (Unrelated Data)	
Model	Prediction	Confidence	Prediction	Confidence	Prediction	Confidence
DNN	COVID-19	99.7%	Non-COVID	75.1%	COVID-19	100%
BNN	COVID-19	95.5%	COVID-19	67.1%	COVID-19	99.8%

Ref. [17, 18, 19]

From Nanoclusters to Analyzing Model Outputs

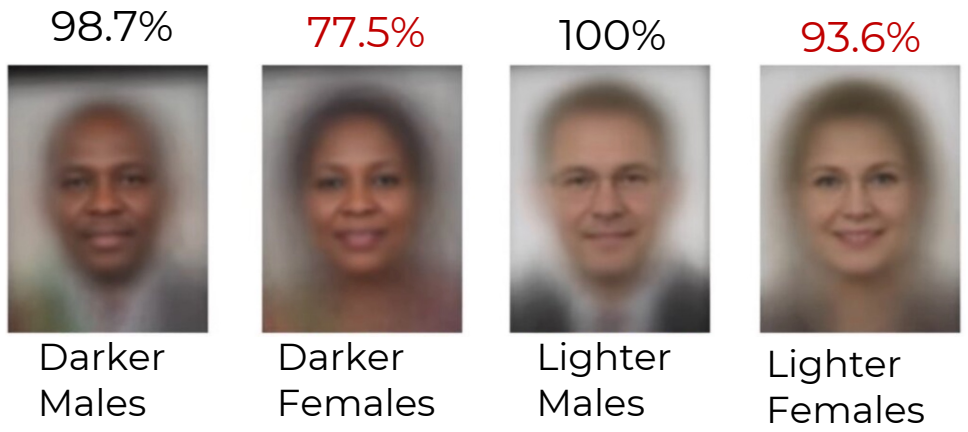
When can predictions be trusted?



Human interpretable feature extraction

1. Select a labeling model e.g. RAM or SAM
2. Forward the dataset and extract bounding boxes and labels
3. Forward the model you want to explain and record the target e.g. loss or prediction
4. Compile results into a table

Does the model act fair?



Color	Activity	Texture	Location	Size	Pred
Black	Hunting	Hairy	Tree	S	F
Black	Climbing	Hairy	Tree	S	F
Gray	Hunting	Hairy	**Water**	L	**T**
Black	Climbing	Hairy	Tree	M	T
Gray	Walking	Hairy	**Water**	M	**T**
Black	Climbing	Hairy	Tree	M	F
Gray	Walking	Hairy	**Water**	L	**T**

From Nanoclusters to Analyzing Model Outputs

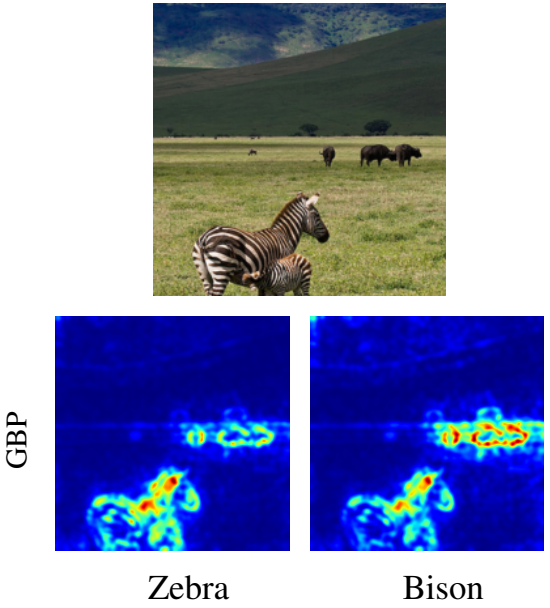
Human interpretable feature extraction

Color	Activity	Texture	Location	Size	Pred
Black	Hunting	Hairy	Tree	S	F
Black	Climbing	Hairy	Tree	S	F
Gray	Hunting	Hairy	**Water**	L	**T**
Black	Climbing	Hairy	Tree	M	T
Gray	Walking	Hairy	**Water**	M	**T**
Black	Climbing	Hairy	Tree	M	F
Gray	Walking	Hairy	**Water**	L	**T**

- This is not always possible
- May also be too coarse

Saliency maps

- Highlight important pixels for classification
- Mostly Gradient and perturbation-based i.e. depending on an infinitesimal change



VAR: Three simple steps to class-specific saliency maps

Step 1: Initial Attribution

Compute attribution for a set of classes $c \in \{1, \ldots, K\}$:

$$A_c = \text{Attribution}(x, c)$$

Where: x is the input and A_c is the attribution for class c

Step 2: Pixel-wise Softmax

Compute softmax across classes for each pixel position (i, j)

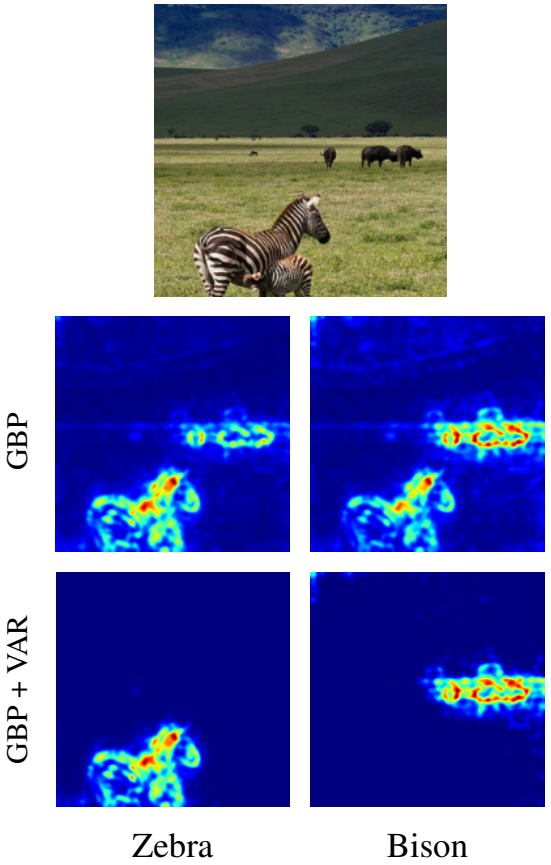
$$M_c(i,j) = \frac{e^{A_c(i,j)/\tau}}{\sum_{k=1}^{K} e^{A_k(i,j)/\tau}}$$

Step 3: Final Attribution

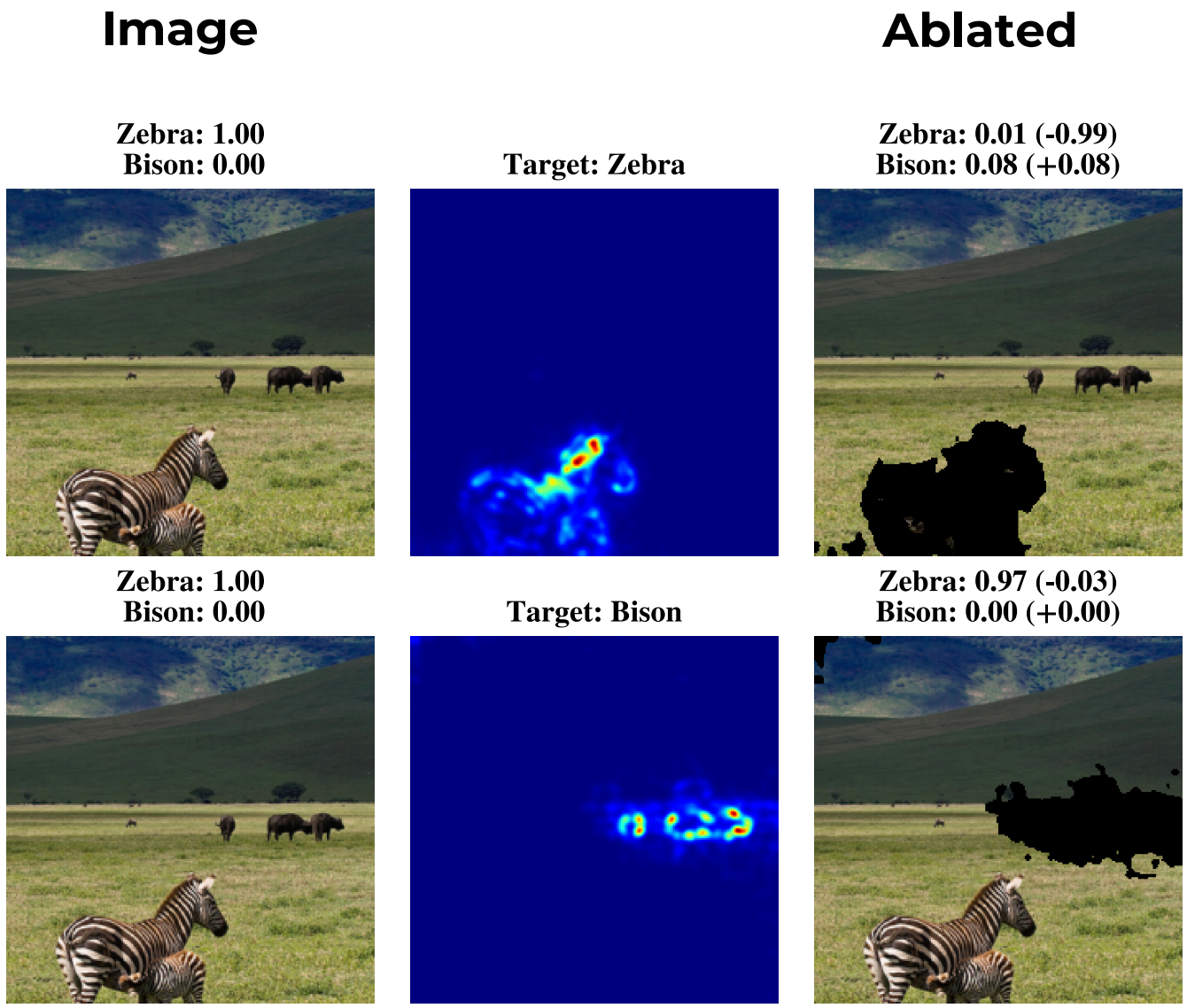
The final attribution for class c is computed as

$$V_c = A_c \odot M_c \odot \mathbb{1}_{M_c - \frac{1}{K} > 1\times10^{-3}}$$

Where: $\odot$ denotes element-wise multiplication, $\mathbb{1}$ is the indicator function, 1×10^{-3} is the threshold parameter



More than pretty pictures

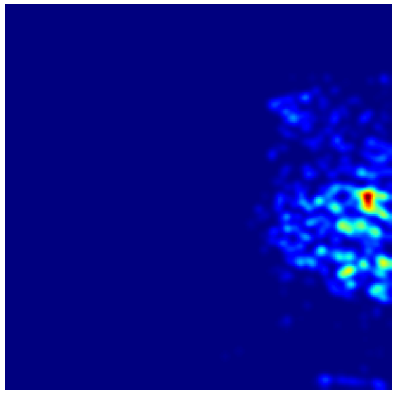


More than pretty pictures

Image

Echidna: 0.00
Colobus: 0.00

Target: Echidna

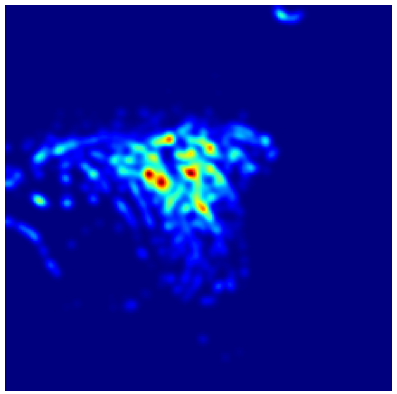


Ablated

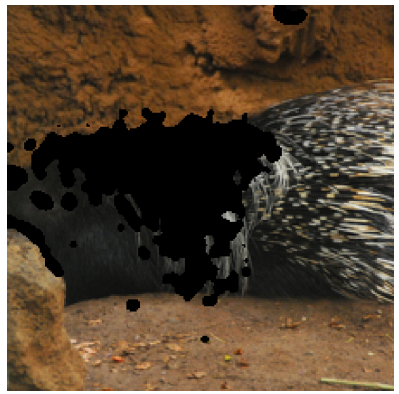
Echidna: 0.00 (-0.00)
Colobus: 0.41 (+0.41)

Echidna: 0.00
Colobus: 0.00

Target: Colobus



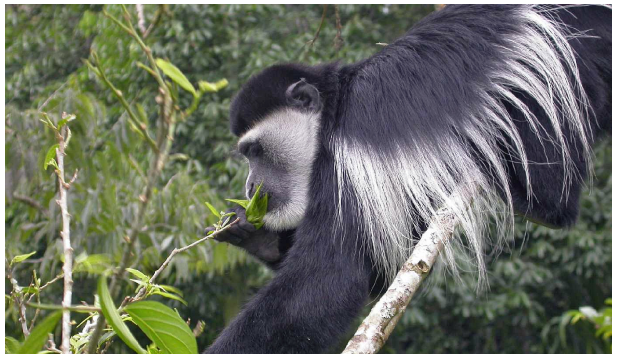
Echidna: 0.15 (+0.15)
Colobus: 0.00 (-0.00)



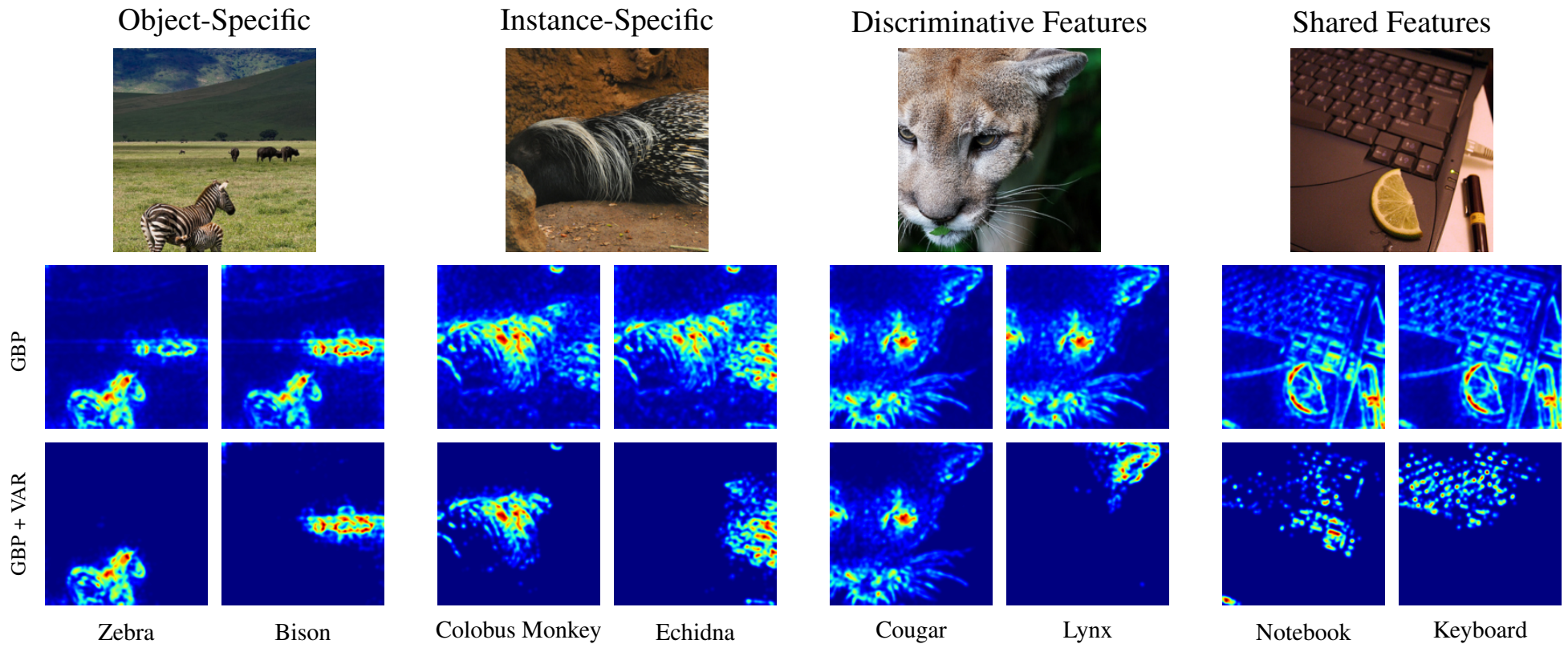
Classes

Echidna

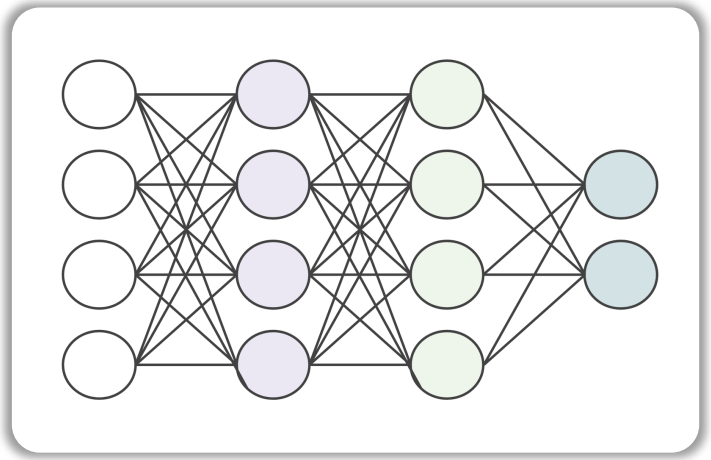
Colobus



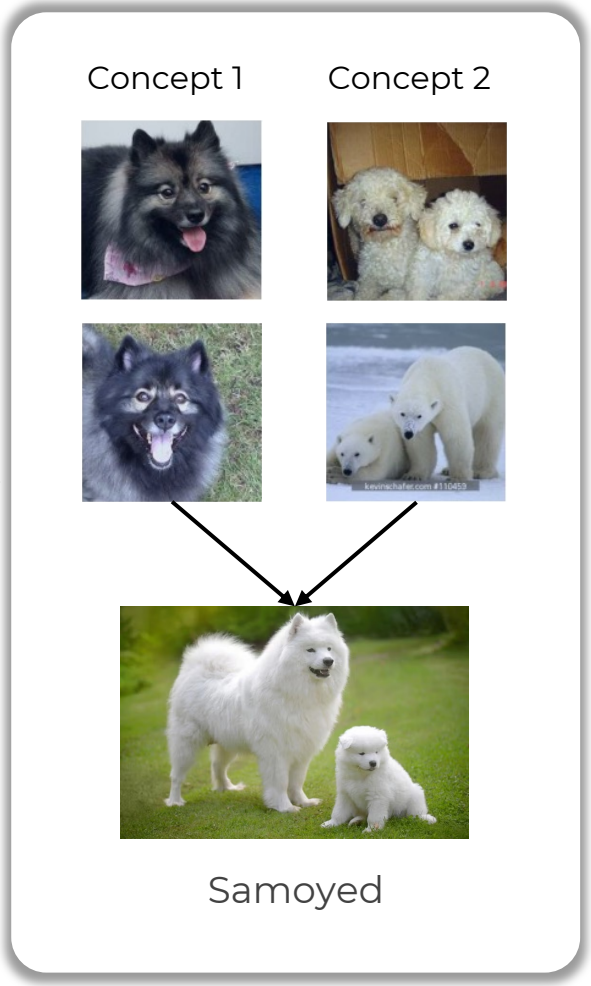
VAR: A Framework for Class-specific saliency maps



Understanding reasoning

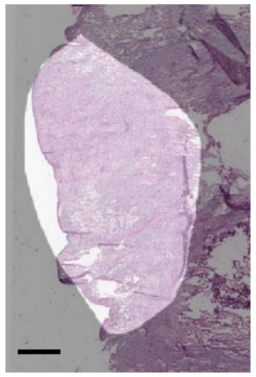


Extract interpretable concepts

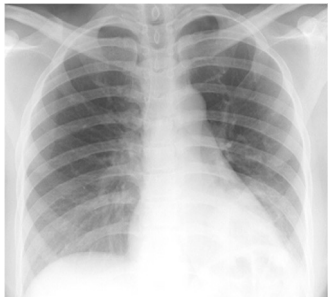


Why is it important?

Gaining new scientific insights



"... 45 out of 54 of the TCGA images **misclassified by at least one of the pathologists** were **assigned** to the **correct** cancer type **by the algorithm**".



COVID-Net

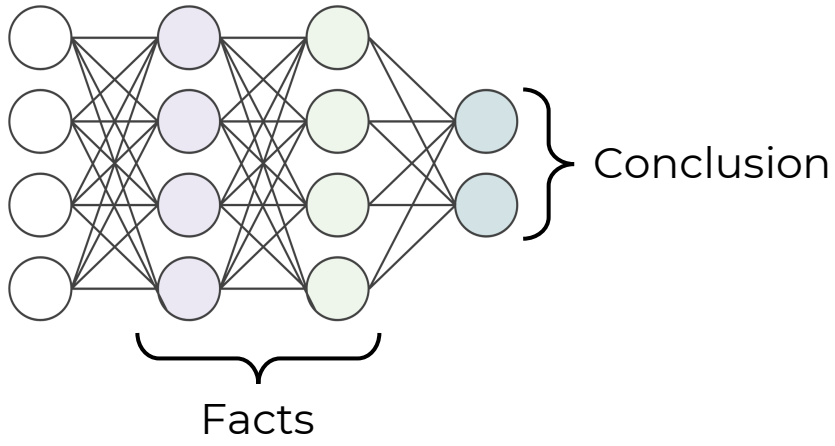
Gain insights into **critical factors** associated with **COVID** cases

What is **reasoning**?

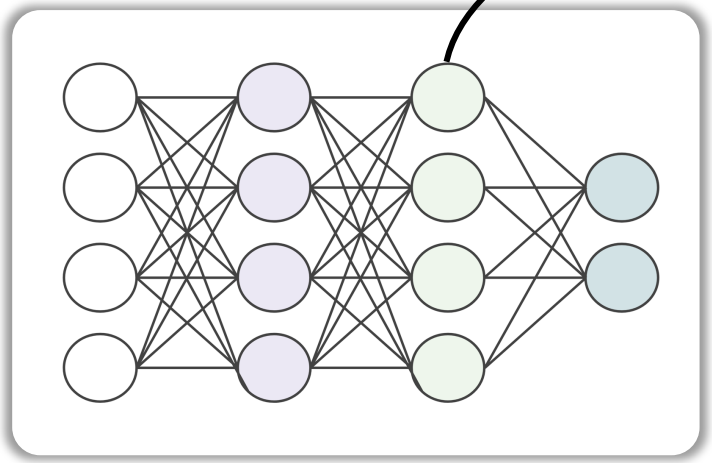
*„Reasoning is the process by which you reach a **conclusion** after thinking about all the **facts**."*

— Collins dictionary

i.e. a relation between **facts** and a **conclusion**

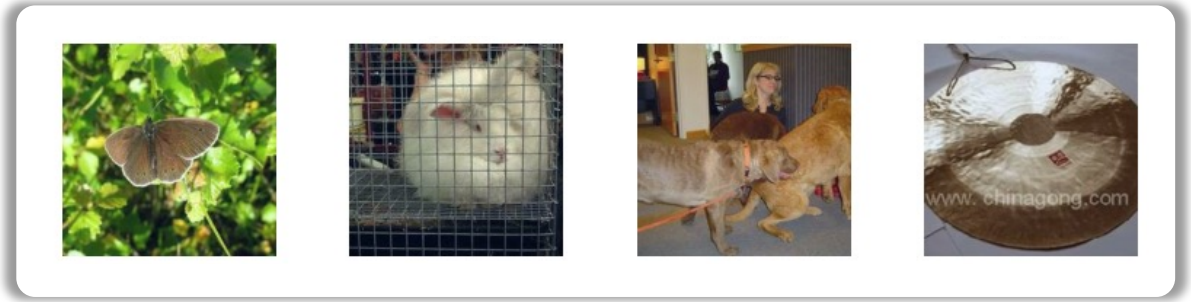


Extracting Features



Neuron activations								Y
0	1	1	1	1	1	1	1	c_1
1	1	0	0	0	0	1	1	c_1
1	1	1	1	0	1	0	1	c_1
0	1	1	1	0	1	0	0	c_2
0	1	1	1	0	1	0	0	c_2

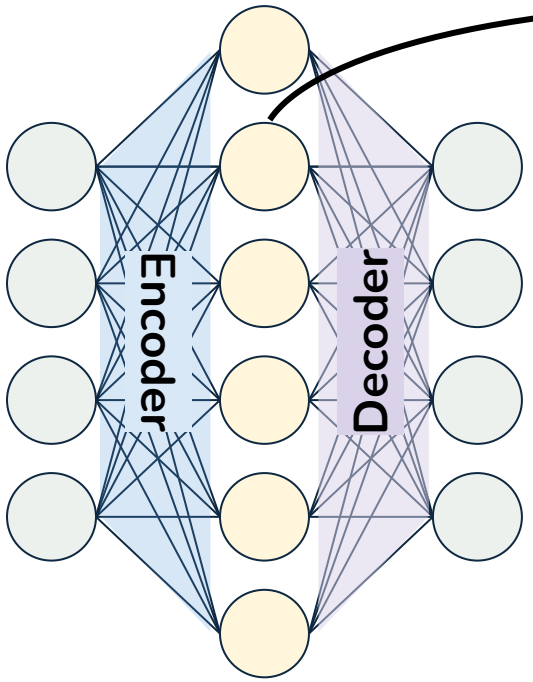
Neurons are polysemantic

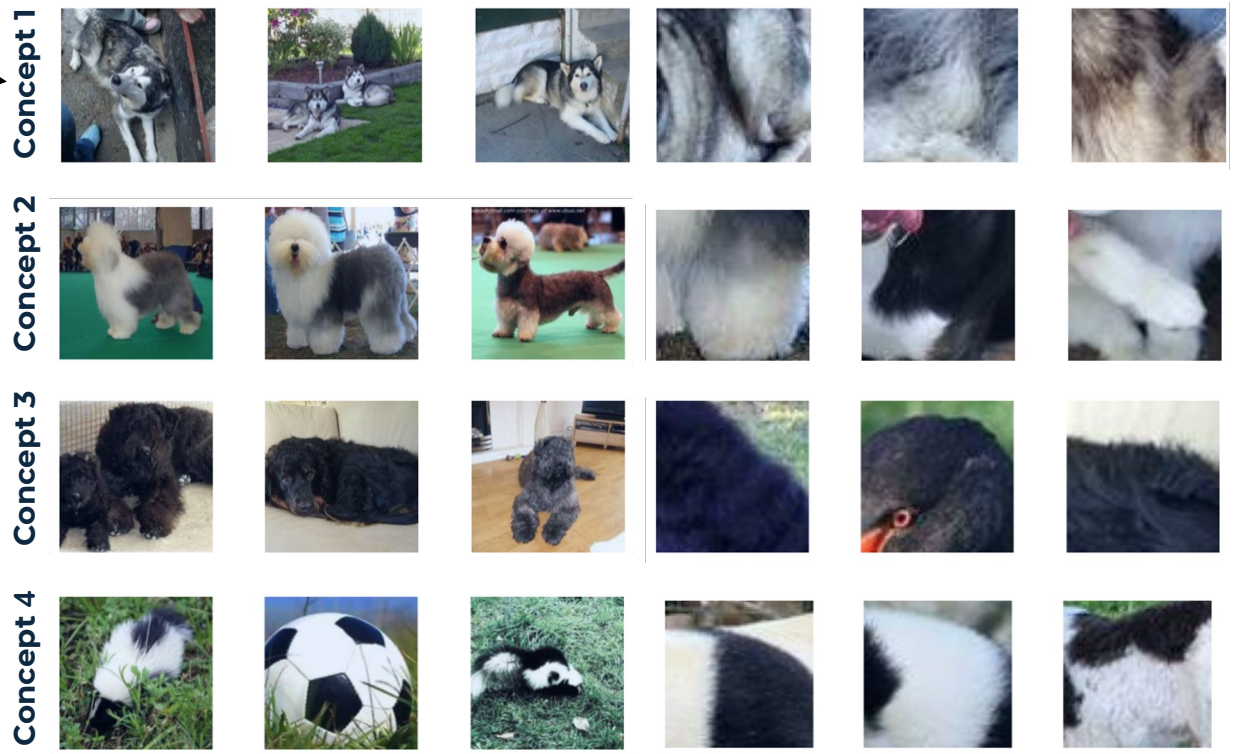


Slide credit: Ghada Said (masters student)

Extracting ~~Features~~ Concepts

Sparse Autoencoder





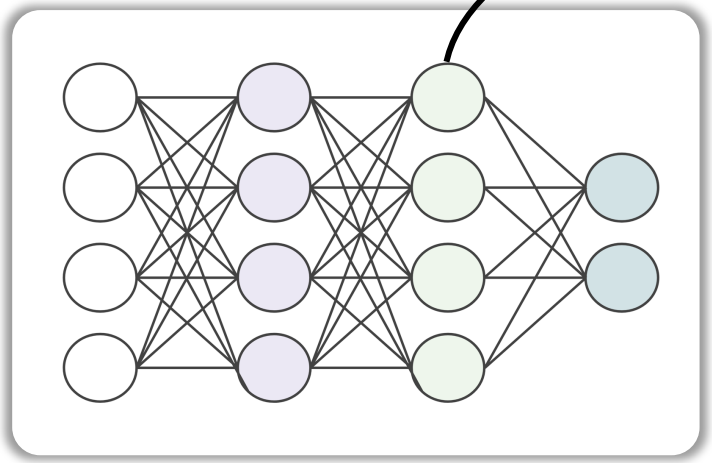
(Not perfect, but a big leap forward)

Loss Function: Reconstruction Error + Latents Sparsity *

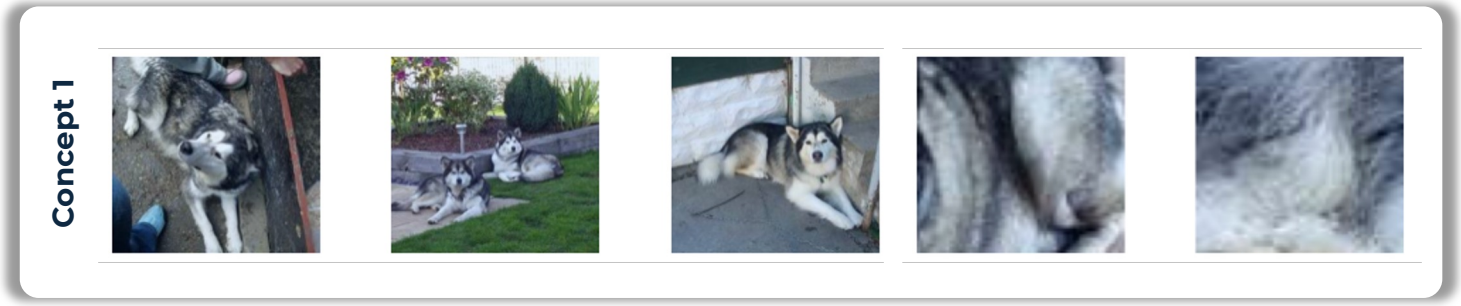
* Latent Sparsity = L1 weight x L1(Latents)

Slide credit: Ghada Said (masters student)

Extracting ~~Features~~ Concepts

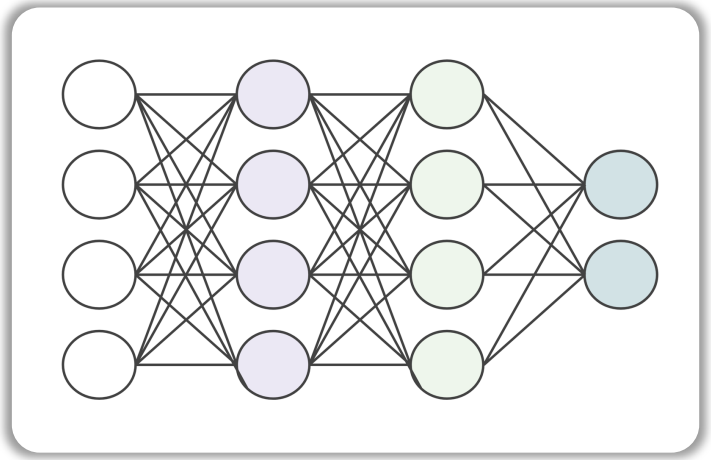


Concept ~~Feature~~ activations								Y
0	1	1	1	1	1	1	1	c_1
1	1	0	0	0	0	1	1	c_1
1	1	1	1	0	1	0	1	c_1
0	1	1	1	0	1	0	0	c_2
0	1	1	1	0	1	0	0	c_2

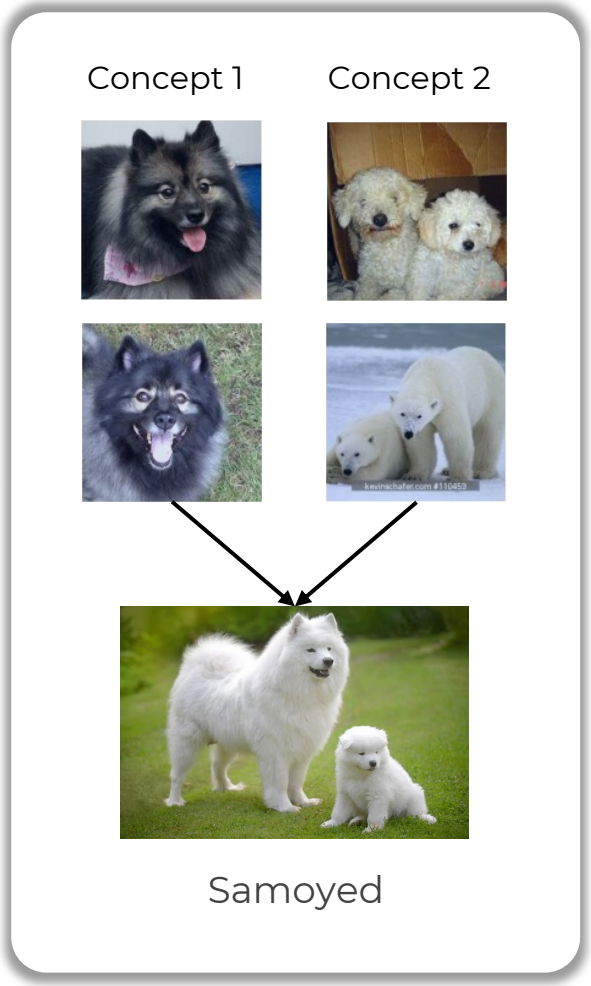


Slide credit: Ghada Said (masters student)

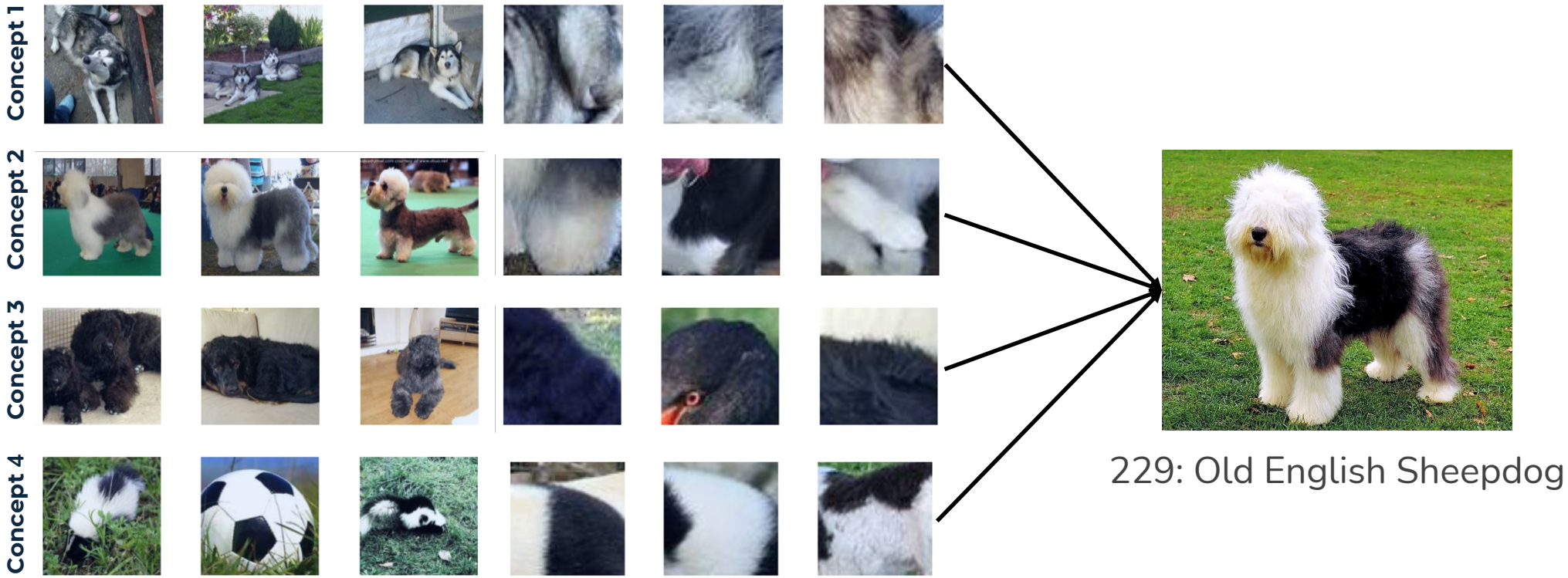
Understanding reasoning



Extract interpretable concepts

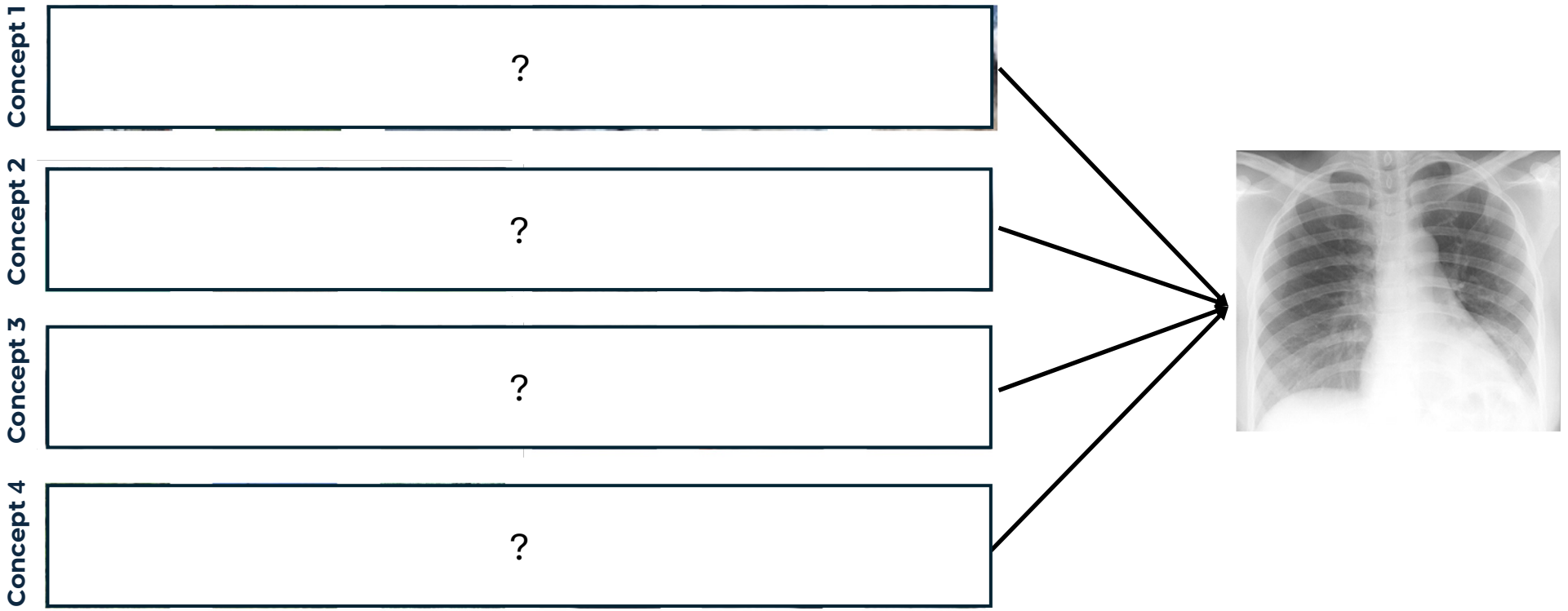


Connecting concepts



Slide credit: Ghada Said (masters student)

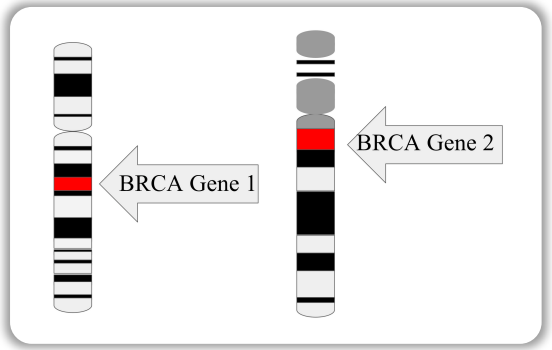
Connecting concepts



Slide credit: Ghada Said (masters student)

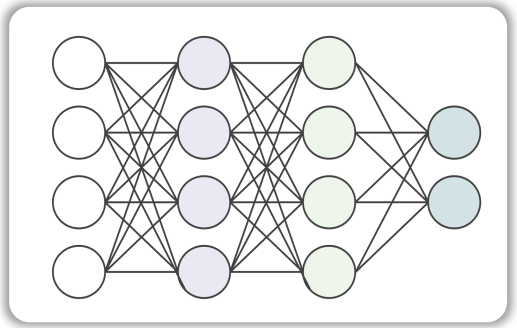
Conclusion

DIFFNAPS



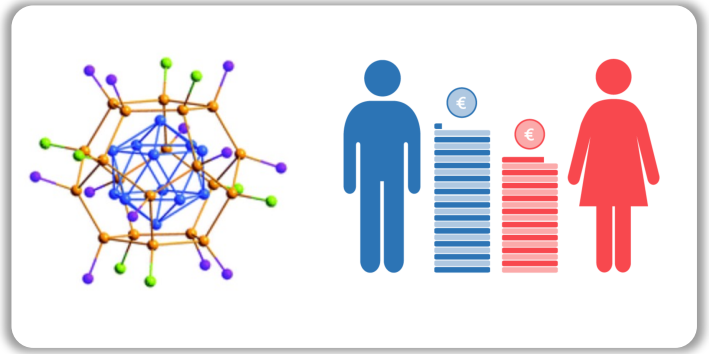
WIP

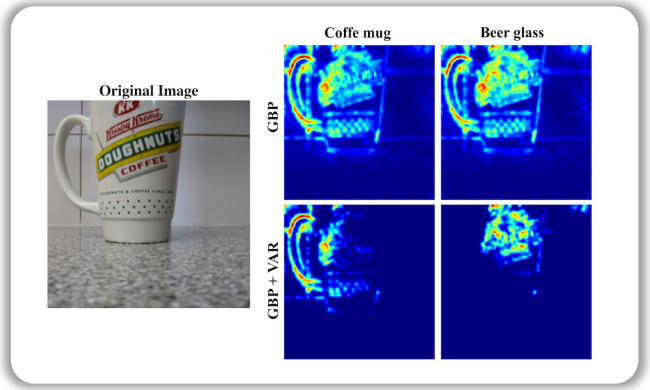
WIP



Data ⟵ ⟶ Models

SYFLOW





VAR

References

[1] ChatGPT
[2] ChatGPT
[3] https://en.wikipedia.org/wiki/BRCA_mutation
[4] Kenzler, S., & Schnepf, A. (2021). Metalloid gold clusters–past, current and future aspects. *Chemical Science*.
[5] https://www.image-net.org/
[6] https://cocodataset.org/#home
[7] Rezaei, Keivan, et al. "PRIME: Prioritizing Interpretability in Failure Mode Extraction." The Twelfth International Conference on Learning Representations.
[8] Ulianova, S. 2017. Cardiovascular Disease dataset. *https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset*.
[9] Patil, P.; and Rathod, P. 2020. Disease Symptom Prediction. *https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset*
[10] The Cancer Genome Atlas (TCGA). *https://www.cancer.gov/tcga*.
[11] The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature*.
[12] Breiman, L. 1984. Classification and regression trees. Routledge
[13] Proenca, H. M.; and van Leeuwen, M. 2020. Interpretable multiclass classification by MDL-based rule lists. *Information Sciences*.
[14] Pellegrina, L.; Riondato, M.; and Vandin, F. 2019. SPuManTE: Significant pattern mining with unconditional testing. *In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD*).
[15] Hedderich, M. A.; Fischer, J.; Klakow, D.; and Vreeken, J. 2022. Label-descriptive patterns and their application to characterizing classification errors. *In Proceedings of the International Conference on Machine Learning (ICML)*.
[16] https://www.destatis.de/DE/Presse/Pressemitteilungen/2020/03/PD20_097_621.html
[17] https://www.emmstech.org/blog/racial-and-gender-bias-found-in-facial-recognition-systems[9] Patil, P.; and Rathod, P. 2020.
[18] ChatGPT
[19] Mallick et al 2020
[20] https://lazoo.org/explore-your-zoo/our-animals/mammals/short-nosed-echidna/
[21] https://en.wikipedia.org/wiki/Black-and-white_colobus
[22] Coudray, N. et al. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning.
[23] https://alexswong.github.io/COVID-Net/