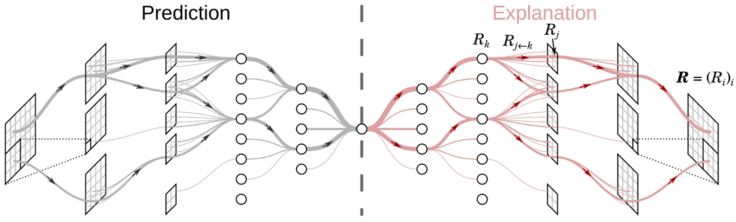
From Concepts to Prototypes

Towards Minimal Effort Post-Hoc Interpretability

Sebastian Lapuschkin, Fraunhofer HHI

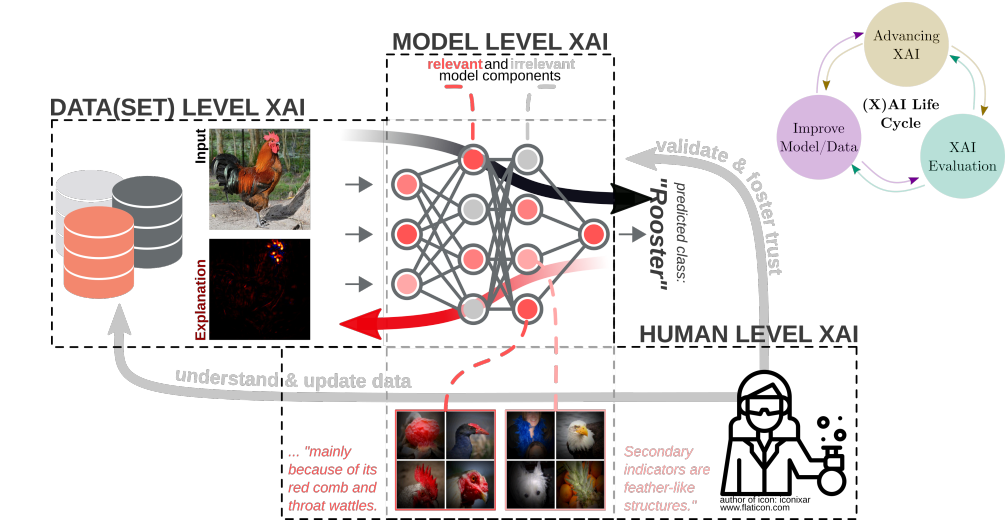
January 11-13 2024

Machine Teaching for Humans workshop, Valencia



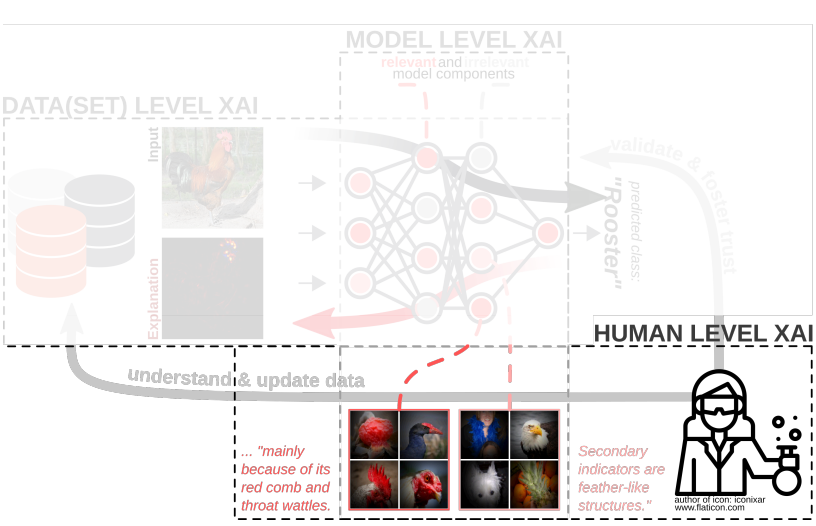
Brief Intro & Research Objectives

Focus: XAI Method Development



Brief Intro & Research Objectives

Topic Today: Advances in Human Level XAI

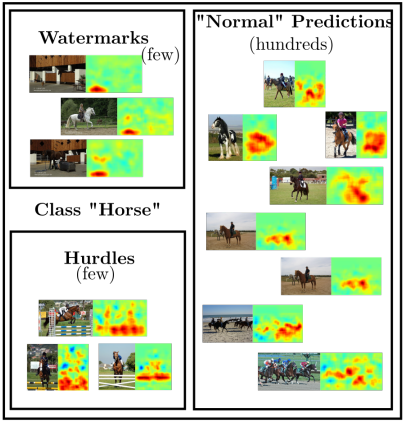


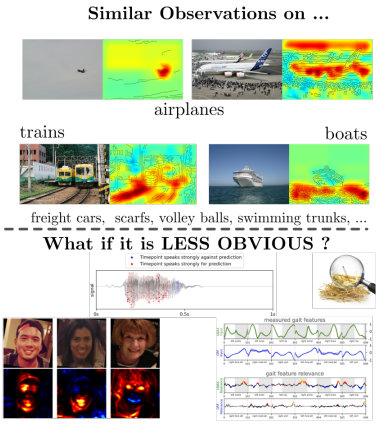
Advancing XAI for Humans

Past Observations on AI Behavior

Models Learn Consistent Strategies

... which, however, may be tricky or impossible to spot.



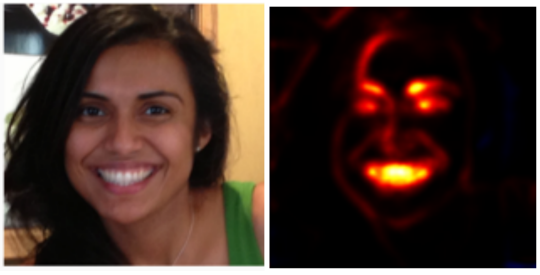


Manual assessments back in the day. Can we address this more smartly?

Sources: [Lapuschkin, Binder, Montavon, et al. 2016], [Lapuschkin, Binder, Müller, et al. 2017], [Horst et al. 2019], [Becker et al. 2023]

Previous Work

Human Alignment: Addressing the Limitations of Local XAI



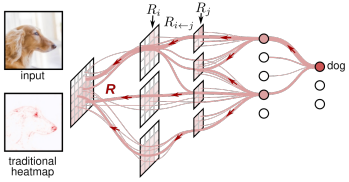
Age and Sex recognition on images of faces [Lapuschkin, Binder, Müller, et al. 2017]

- **Interpretation 1:**
  "laughing is relevant"
- **Interpretation 2:**
  "color of teeth is relevant"
- **Interpretation 3:**
  "size of teeth is relevant"
- ...

Previous Work

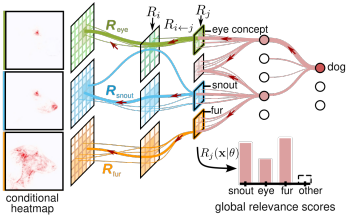
LRP, CRP & RelMax

a traditional explanation (LRP)



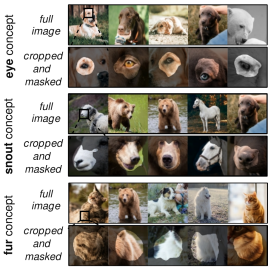
$$R_{i\leftarrow j}^{(l-1,l)}(\mathbf{x}) = \frac{z_{ij}}{z_j} R_j^l(\mathbf{x})$$

b concept-conditioned explanation (CRP)



$$R_{i\leftarrow j}^{(l-1,l)}(\mathbf{x}|\theta \cup \theta_l) = \frac{z_{ij}}{z_j} \sum_{c_l \in \theta_l} \delta_{jc_l} R_j^l(\mathbf{x}|\theta)$$

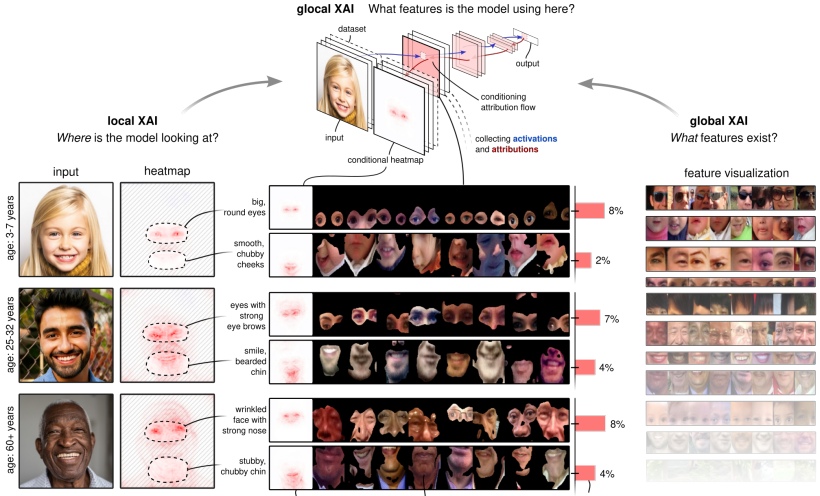
c concept reference samples (RelMax)



Source: [Achtibat et al. 2023]

Previous Work

LRP, CRP & RelMax



Source: [Achtibat et al. 2023]

Previous Work

LRP, CRP & RelMax

Reduan

Max

nature machine intelligence

Explore content ∨ About the journal ∨ Publish with us ∨

nature › nature machine intelligence › articles › article

Article | Open access | Published: 20 September 2023

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

Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek ✉ & Sebastian Lapuschkin ✉

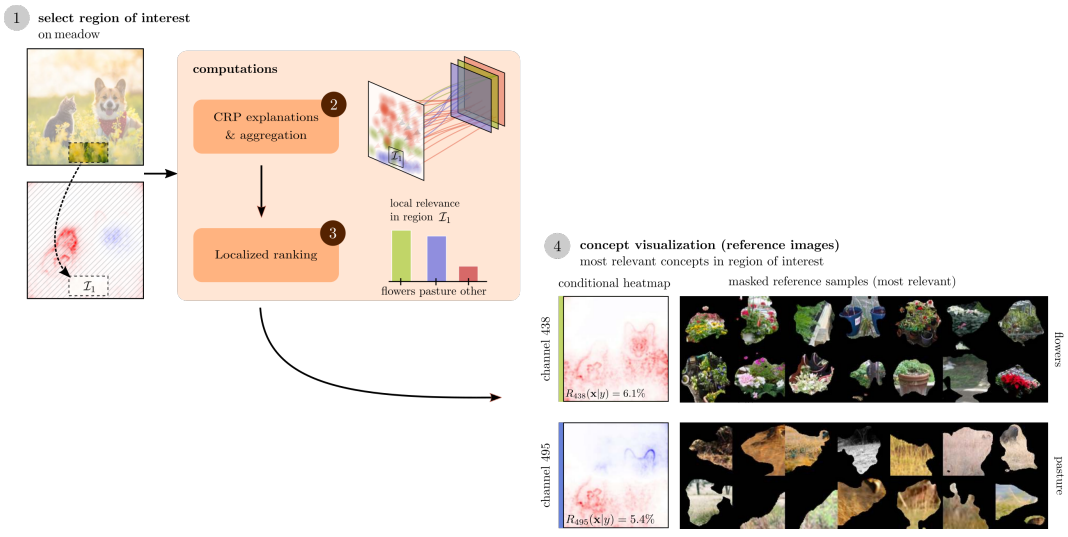
Nature Machine Intelligence **5**, 1006–1019 (2023) | Cite this article

11k Accesses | **3** Citations | **26** Altmetric | Metrics

Source: [Achtibat et al. 2023]

Novel Tools for Actionable (X)AI

Local Analysis: "*what* is here?" [Achtibat et al. 2023]

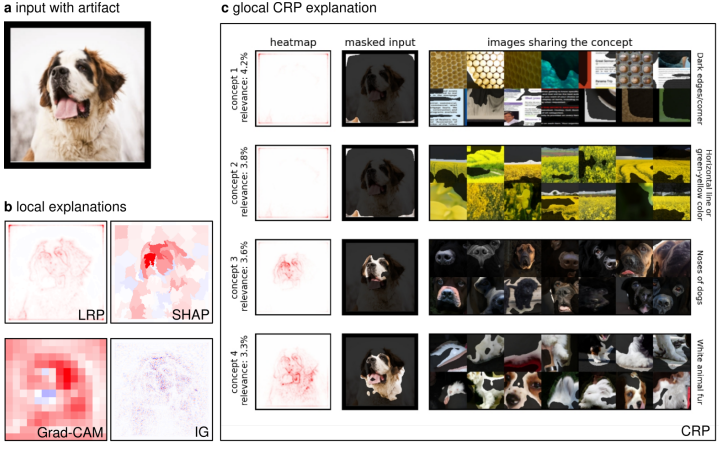


Novel Tools for Actionable (X)AI

Inverse Search: "*what else* is impacted by this?" [Achtibat et al. 2023]

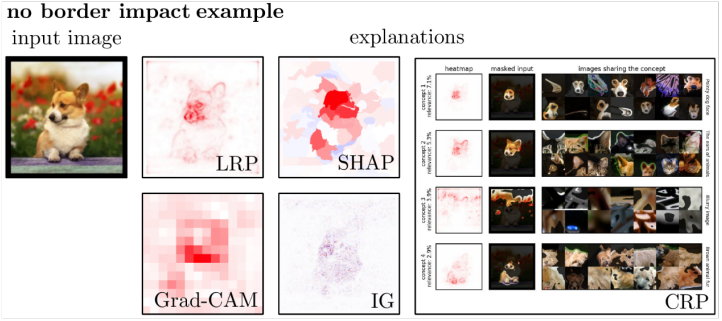
Human User Study

Users Agree! Concept-based is more informative.



Human User Study

Users Agree! Concept-based is more informative, but also more work and effort. **And work sucks.**

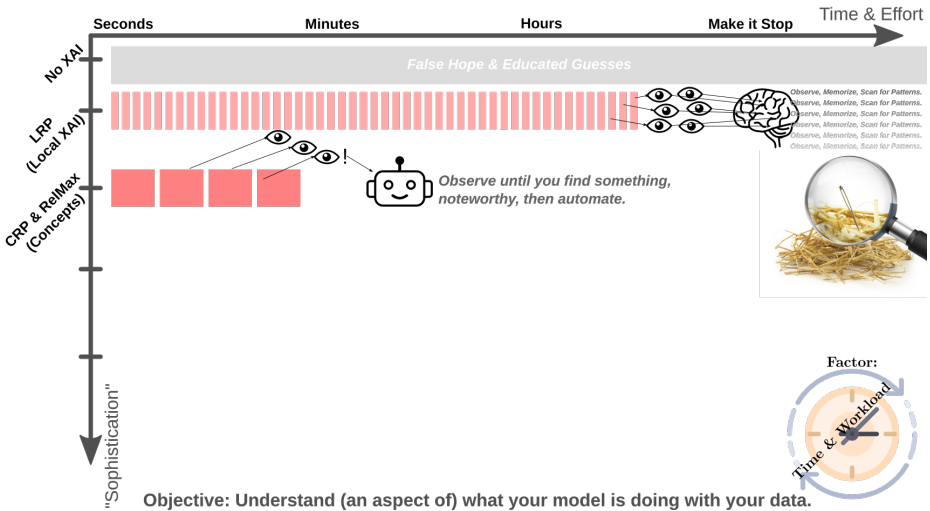


extended study results with secondary measures

Method	Accuracy (%)	F1-Score (%)	Confidence (%)	Clarity (%)
IG	51.7 ± 1.9	52.7 ± 2.9	$\mathbf{77.0 \pm 1.6}$	70.7 ± 1.6
LRP	56.6 ± 2.9	61.6 ± 2.4	74.3 ± 1.3	$\mathbf{71.8 \pm 1.7}$
SHAP	58.3 ± 2.7	62.2 ± 2.4	74.2 ± 1.6	67.7 ± 1.8
Grad-CAM	63.7 ± 3.4	67.4 ± 2.3	70.5 ± 1.8	64.9 ± 1.9
CRP (ours)	$\mathbf{80.9 \pm 3.4}$	$\mathbf{82.3 \pm 1.8}$	76.1 ± 1.7	64.1 ± 1.6

How much Manual Work does XAI Require?

We have made progress over the years, but can we do better?



Spectral Relevance Analysis (SpRAy)

Proposed in [Lapuschkin, Wäldchen, et al. 2019], extended in [Anders et al. 2022]

nature communications

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature communications › articles › article

Article | Open access | Published: 11 March 2019

Unmasking Clever Hans predictors and assessing what machines really learn

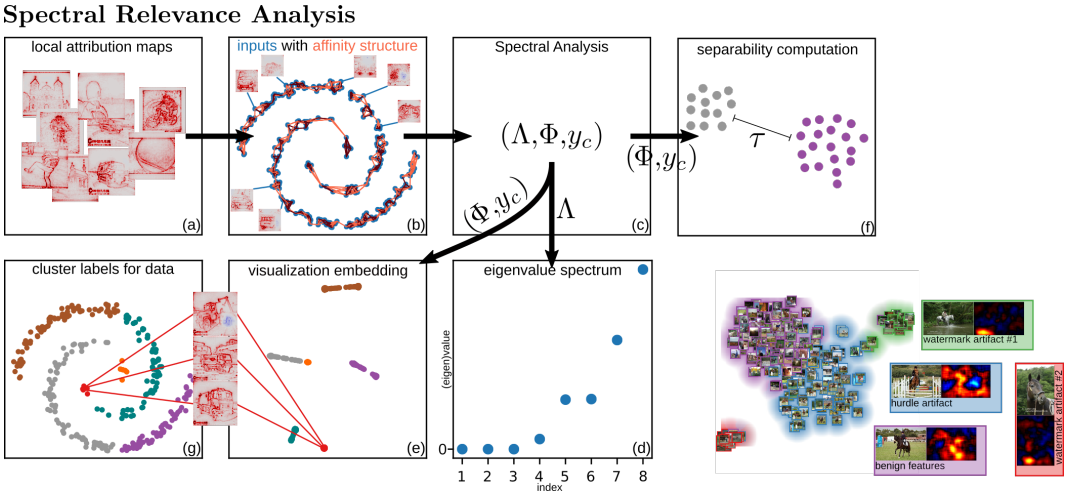
Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek ✉ & Klaus-Robert Müller ✉

Nature Communications **10**, Article number: 1096 (2019) | Cite this article

50k Accesses | **526** Citations | **168** Altmetric | Metrics

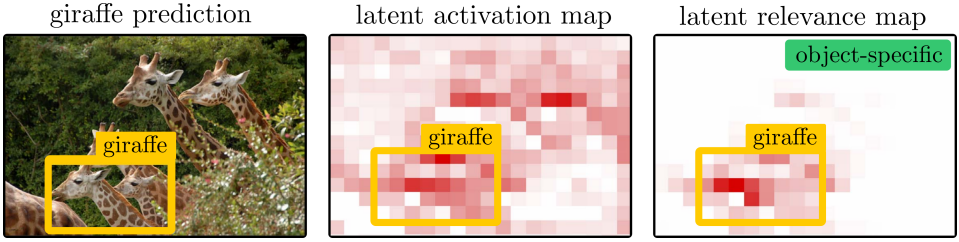
Spectral Relevance Analysis (SpRAy)

Proposed in [Lapuschkin, Wäldchen, et al. 2019], extended in [Anders et al. 2022]



Spectral Relevance Analysis (SpRAy)

Processing Relevance Instead of Features

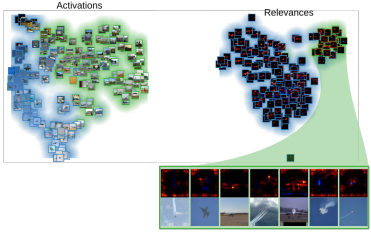


Source: [Dreyer, Achtibat, Wiegand, et al. 2023].
In Latent Space, Relevances naturally filter out irrelevant Activations (from the model's point of view).
We therefore analyze the data via the model.

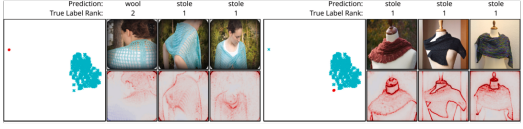
Spectral Relevance Analysis (SpRAy)

Various Findings Reveal Consistent Model Strategies

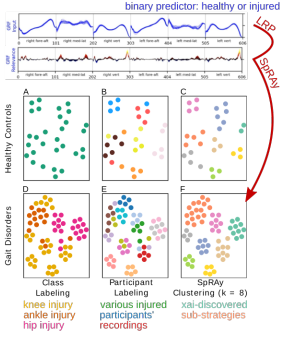
[Lapuschkin et al. 2019]:
Activations cluster based on appearance.
Relevances cluster based on reliance on processing artifact.



[Anders et al. 2022]:
SpRAy in Latent Space: Identify also hard to localize artifacts.
Here: Combinations of factors?



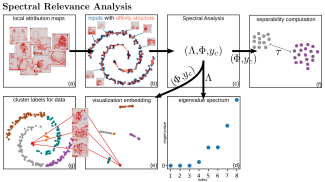
[Slijepcevic et al. 2021]:
Gait DNN distinguishes different variants of (unlabelled) injury types



Sources: [Lapuschkin, Wäldchen, et al. 2019] [Slijepcevic et al. 2021] [Anders et al. 2022]

Spectral Relevance Analysis (SpRAy)

Summary

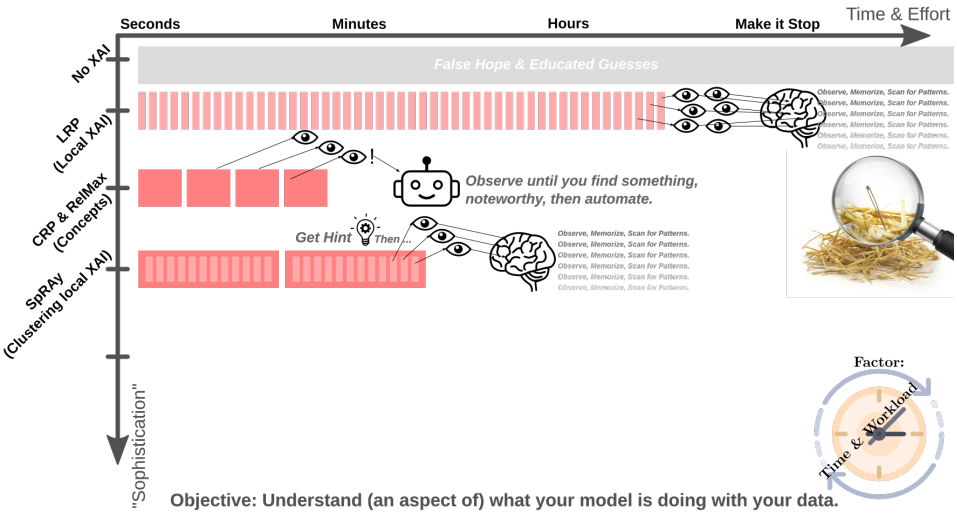


- **Pros:**
  Semi-automatic, Analytical Properties of Λ, Representative Embedding Φ, Latent Space Compatible.
- **Cons:**
  Fidgety parameters, bound to (mostly Euclidean Space) Affinity: no connection to particular Latent Concepts or other classes*, only combined `fit_transform()`[1], struggles with large datasets.

[1] speaking in `sklearn` API: ie not applicable during test time for *single* instances

How much Manual Work does XAI Require?

We have tried, a few years back...



Prototypical Concept-based Explanation (PCX)

Simpler, Better & Powerfuller [Dreyer, Achtibat, Samek, et al. 2023]

Max

Understanding the (Extra-)Ordinary: Validating Deep Model Decisions with Prototypical Concept-based Explanations

Maximilian Dreyer[1] Reduan Achtibat[1] Wojciech Samek[1,2,3,†] Sebastian Lapuschkin[1,†]

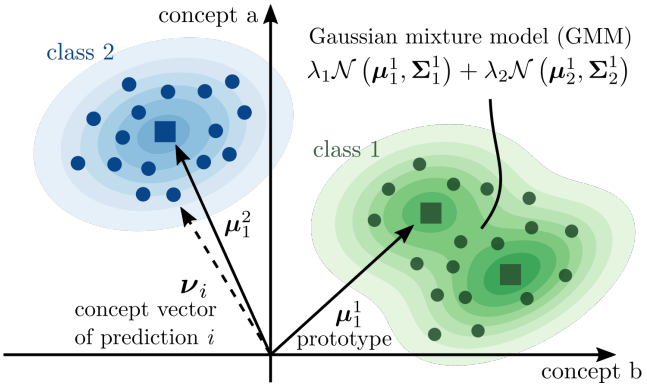
[1] Fraunhofer Heinrich-Hertz-Institute, 10587 Berlin, Germany
[2] Technische Universität Berlin, 10587 Berlin, Germany
[3] BIFOLD – Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany
[†] corresponding authors: {wojciech.samek,sebastian.lapuschkin}@hhi.fraunhofer.de

Reduan

Prototypical Concept-based Explanation (PCX)

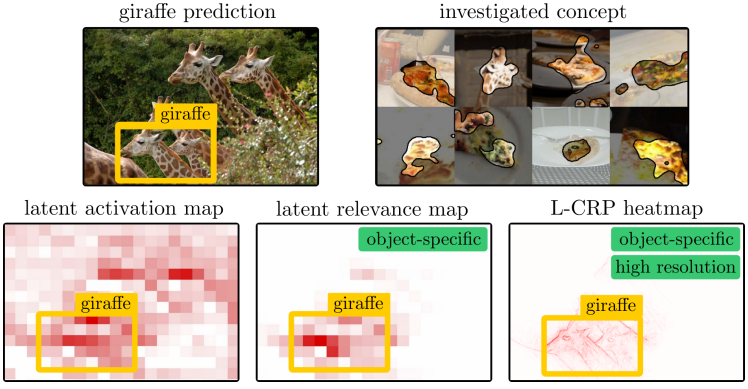
Assumptions & Intuition

assumption: relevance vectors form distinct clusters in latent space



Prototypical Concept-based Explanation (PCX)

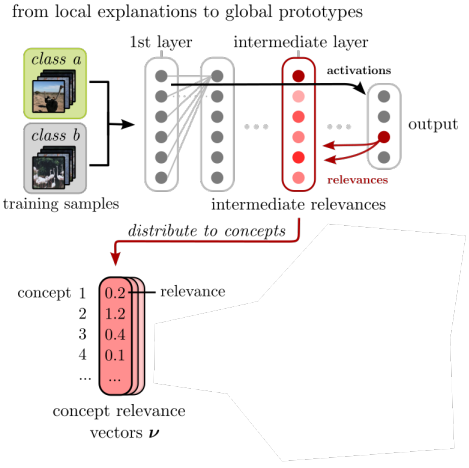
Relevances Naturally Filter out Irrelevant Activations, from the Model's Point of View

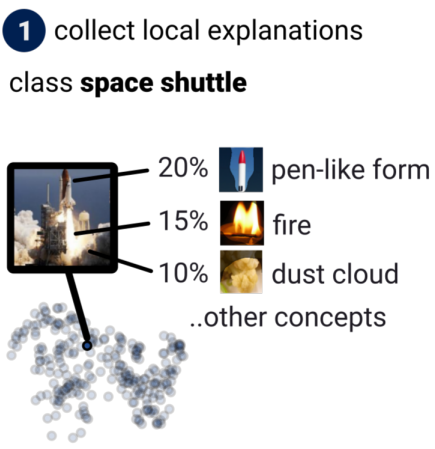


Source: [Dreyer, Achtibat, Wiegand, et al. 2023]

Prototypical Concept-based Explanation (PCX)

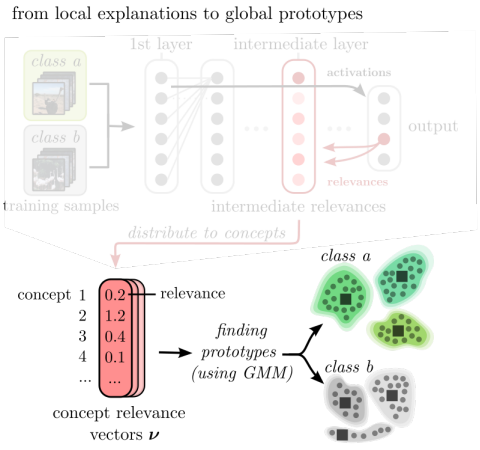
The PCX Preprocessing Pipeline

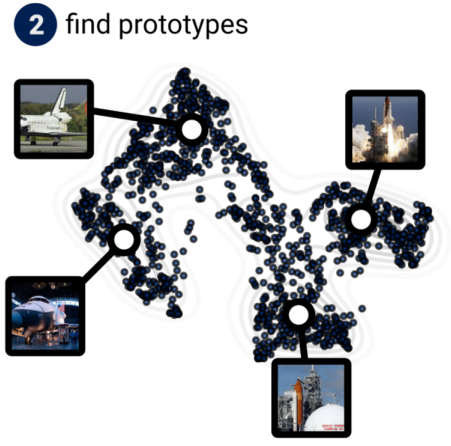




Prototypical Concept-based Explanation (PCX)

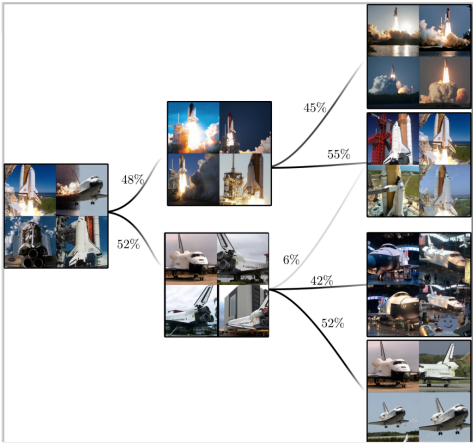
The PCX Preprocessing Pipeline

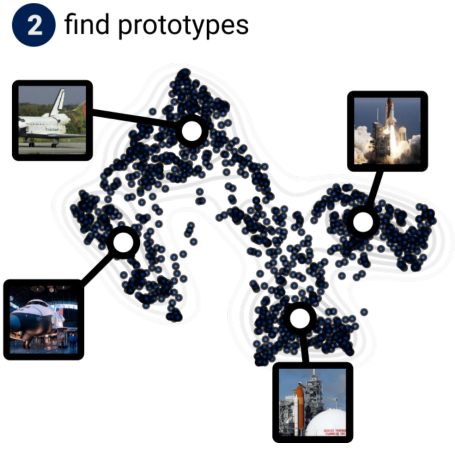




Prototypical Concept-based Explanation (PCX)

The PCX Preprocessing Pipeline





Tracking sample placement while increasing cluster count iteratively explores subclass hierarchies.

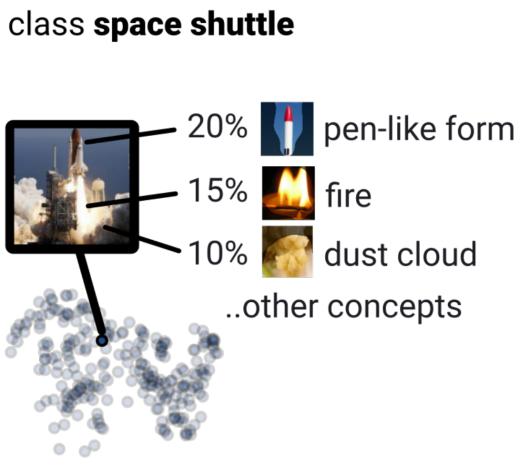
Prototypical Concept-based Explanation (PCX)

Key Questions

- $\mathbf{Q}_0$ What *are* Prototypes?
- $\mathbf{Q}_1$ What global insights can we gain with prototypes?
- $\mathbf{Q}_2$ How can we evaluate prototypes?
- $\mathbf{Q}_3$ How can we use prototypes to validate predictions and ensure safety?

$\mathbf{Q}_0$ What *are* Prototypes?

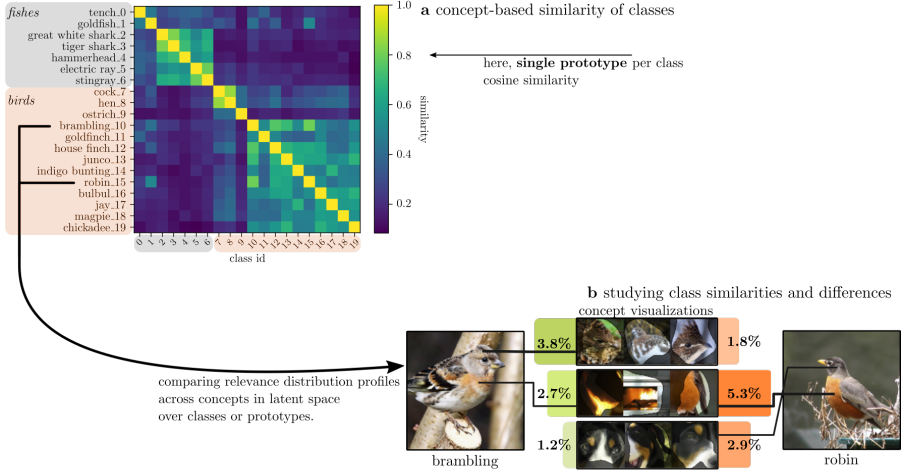
A brief disambiguation



We understand Prototypes as archetypical **Compositions of Concepts as used by the model**, rather than eg Parts of Instances as in ProtoPNet [Chen et al. 2019].

Q_1 What global insights can we gain with prototypes?

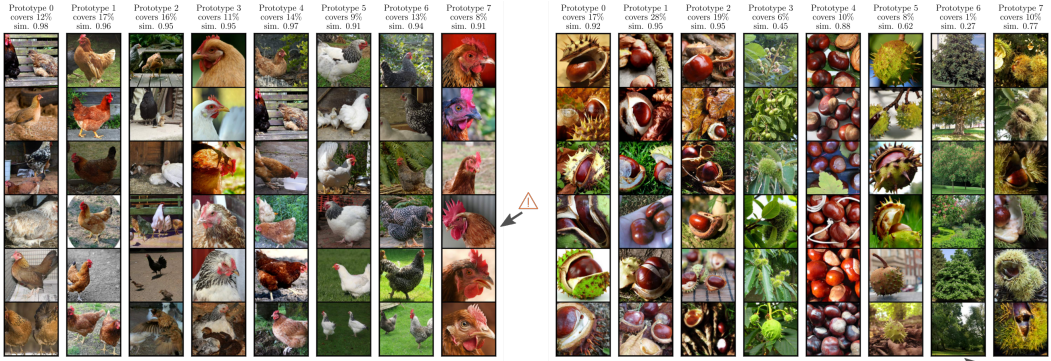
Prototypical Concept-based Explanations: Comparing Classes



Since we directly operate in any of the the NNs (here, VGG16, fist 20 ImageNet classes) latent spaces, the representatory basis of all classes is shared.
Indications for Sub- and Superclass exploration? More results in paper [Dreyer, Achtibat, Samek, et al. 2023]

$\mathbf{Q}_1$ What global insights can we gain with prototypes?

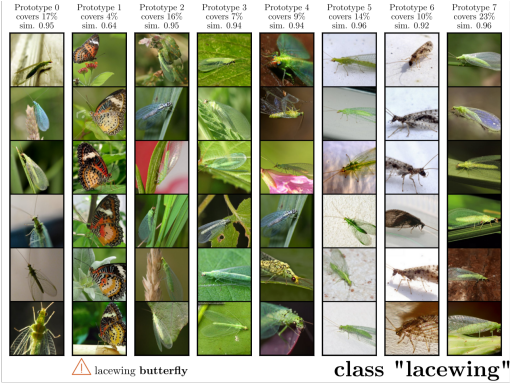
Prototypical Concept-based Explanations: Assessing Data Quality and Annotating Data

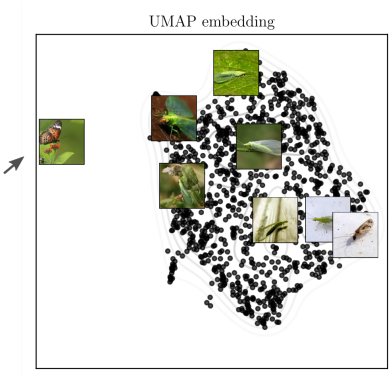


Understanding model sub-strategies using (eight) prototypes for class “hen” (left) and “buckeye” (right).

Q_1 What global insights can we gain with prototypes?

Prototypical Concept-based Explanations: Assessing Data Quality and Annotating Data

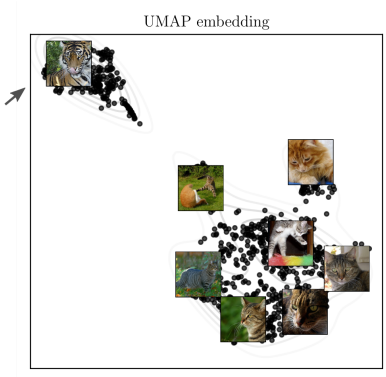




Spotting mislabelled subpopulations made easy. Here, “lacewing” from ImageNet.

$\mathbf{Q}_1$ What global insights can we gain with prototypes?

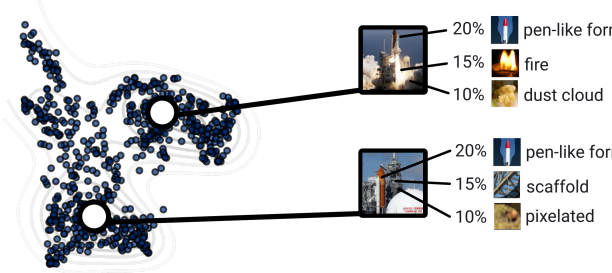
Prototypical Concept-based Explanations: Assessing Data Quality and Annotating Data



Spotting mislabelled subpopulations made easy. Here, “tiger cat” from ImageNet.

Q_1 What global insights can we gain with prototypes?

Prototypical Concept-based Explanations: Assessing Data Quality and Annotating Data



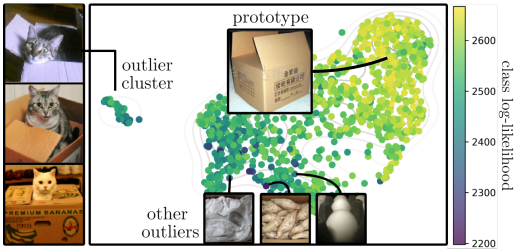
understand global model behavior quickly and in detail

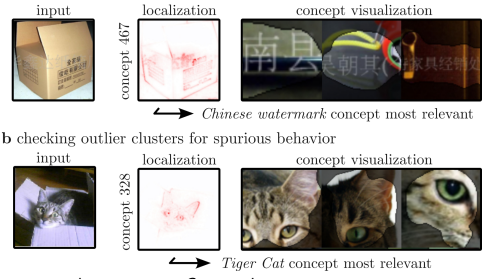
discover flaws of model & data

Since Prototypes directly reside in the NNs latent space, we have access to Concepts!
Prototypes can be checked, validated, marked based on composition.

Q_1 What global insights can we gain with prototypes?

Prototypical Concept-based Explanations: Assessing Data Quality and Annotating Data





Since Prototypes directly reside in the NNs latent space, we have access to Concepts!
Prototypes can be checked, validated, marked based on composition.

Q_1 What global insights can we gain with prototypes?

Prototypical Concept-based Explanations: Assessing Data Quality and Annotating Data

With PCX and CRP combined, we can immediately spot class compositions, opportunities for (re)annotations and even the absence of expected features, such as the "dome-like" properties in prototype 6 of class "pickelhaube".

Q_2 How can we evaluate prototypes?

We will most likely skip this, due to time and complexity. This is a brief summary.

Table 1. Evaluating different attribution methods for concept relevance scores used for prototypes. We show results on ImageNet for 20 classes using (VGG | ResNet | EfficientNet) architectures averaged over all layers, where higher (↑) values are better and best are bold.

	Faithfulness (↑)	Stability (↑)	Sparseness (↑)	Coverage (↑)
LRP (ε-rule) [6]	12.2 | **14.2** | 7.4	99.00 | 98.48 | 99.68	**37.1** | **36.6** | **37.0**	**79.1** | **81.8** | **85.4**
Input×Gradient [46]	12.2 | **14.2** | 6.7	99.00 | 98.44 | 95.45	**37.1** | **36.6** | 35.5	**79.1** | **81.8** | 67.5
LRP (composite) [35]	**12.6** | 13.6 | **7.5**	99.82 | 99.90 | **99.93**	21.0 | 22.8 | 14.0	59.0 | 68.9 | 58.7
GuidedBackProp [48]	12.0 | 13.0 | 6.0	99.89 | **99.93** | 96.11	31.1 | 30.9 | 31.8	60.7 | 73.1 | 67.4
Activation (max)	11.9 | 12.5 | 6.3	**99.92** | **99.93** | 99.92	7.1 | 4.9 | 9.8	54.5 | 60.1 | 49.9
Activation (mean)	11.1 | 13.1 | 5.9	99.86 | 99.90 | 99.30	11.4 | 12.2 | 24.0	51.3 | 61.6 | 55.2

Table taken from [Dreyer, Achtibat, Samek, et al. 2023].

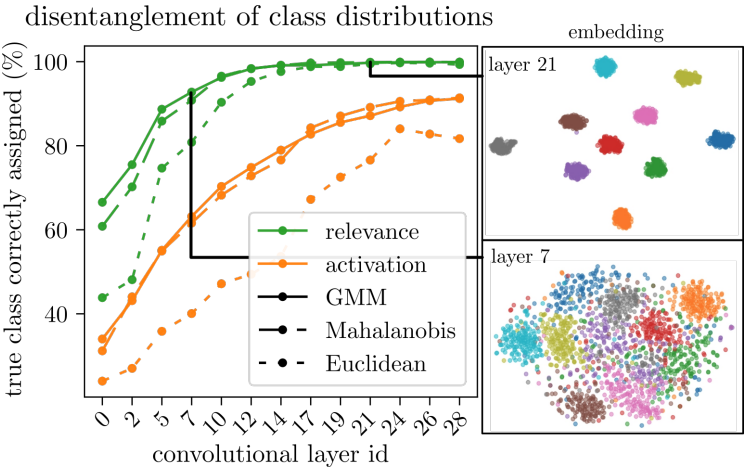
Overall Question: Which quantity yields a "good" basis for prototypes? Different aspects of goodness:

Faithfulness: How representative are the prototypes of the model's behavior? **Stability:** How stable/similar are prototypes across random subsets of data? **Sparseness:** How Sparse are the prototypes in terms of feature/concept alignment?

Coverage: NEW! How complete and correct is the (test) data modeled by the prototypes, in terms of true label assignment?

Q_2 How can we evaluate prototypes?

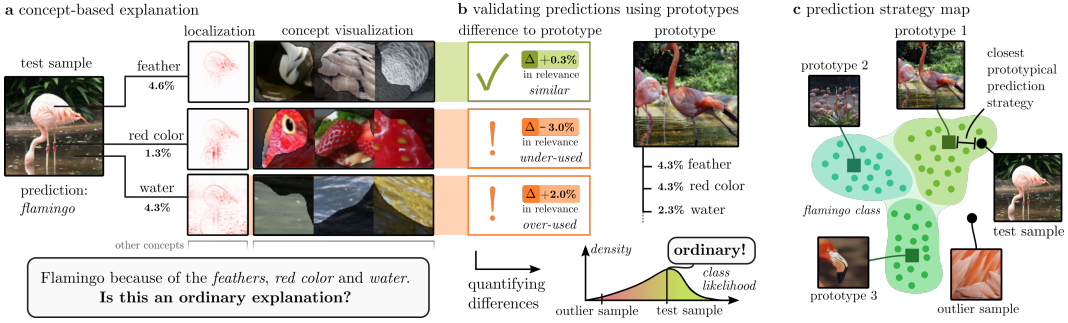
Coverage Visualized



Relevances naturally filter out irrelevant activations, from the models' task-contextual point of view, leading to cleaner, more distinct cluster for prototype identification and mapping.

Q_3 How can we validate predictions for safety?

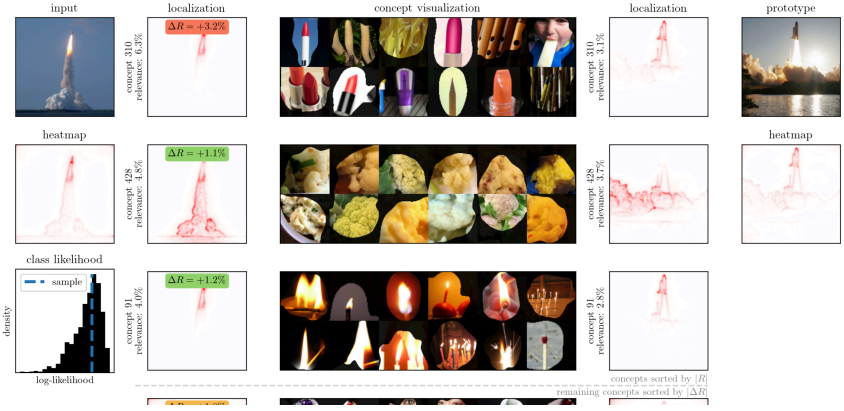
PCX is applicable *during test time*



Once the prototypes have been sighted and assessed, each test point can be (sub)categorized, re-labelled, related to, *positively validated* or rejected near-instantaneously and automatically.

Q_3 How can we validate predictions for safety?

Case Study: Space Shuttle



Read: Pretty ordinary sample, just with significantly more "pen-like shape" expression and stronger ~~"cauliflower"~~ "dust cloud" and "fire" concepts than the closest prototype.

Q_3 How can we validate predictions for safety?

Large Scale Evaluation: Out-of-Domain (OOD) Testing:

SoftMax output statistics

Latent Activations

Table 2. OOD detection results for (VGG|ResNet|EfficientNet) models trained on CUB-200. Higher AUC scores are better with best bold.

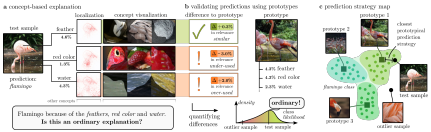
	LSUN	iSUN	Textures	SVHN	Average
MSP [20]	98.9 | 98.8 | 99.2	94.5 | 94.1 | 95.5	89.7 | 91.6 | 89.2	96.2 | 98.6 | 99.0	95.5
Energy [32]	47.7 | 99.8 | 100.0	65.6 | 96.2 | 99.8	63.9 | 95.0 | 94.7	52.5 | 99.7 | 100.0	84.6
Mahalanobis [30]	53.1 | 74.4 | 16.9	87.3 | 97.6 | 41.4	95.6 | 96.9 | 92.1	85.6 | 92.7 | 6.4	70.0
PCX-E (ours)	99.8 | 99.8 | 99.9	99.2 | 98.8 | 98.3	98.6 | 98.9 | 98.7	99.7 | 99.8 | 100.0	99.3
PCX-MD (ours)	99.9 | 99.9 | 100.0	99.5 | 99.6 | 99.3	98.8 | 99.3 | 99.3	99.7 | 100.0 | 100.0	**99.6**
PCX-GMM (ours)	99.9 | 99.9 | 100.0	99.5 | 99.6 | 99.3	98.8 | 99.3 | 99.3	99.7 | 100.0 | 100.0	**99.6**

Latent Space Relevance Scores

Gist: Given a model, and an in-domain and OOD test set, how good is your method in distinguishing the samples' origin ?

Prototypical Concept-based Explanation

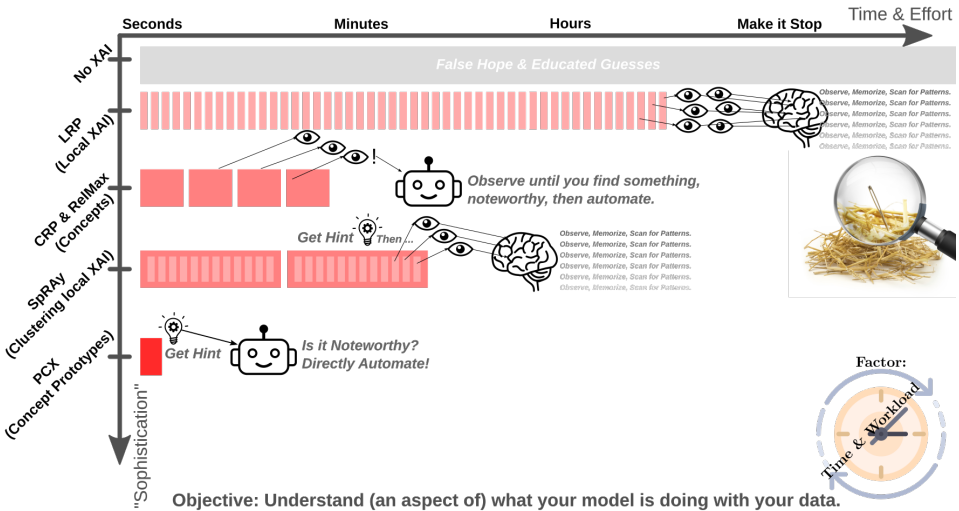
Summary



- **Pros:**
  Operates in "native" NN Latent Space: Prototypes comparable across classes, PCX compatible to CRP. Separate `fit()` and `predict()` functions: Applicable during test time. Increased Automation Potential, for eg. Data Annotation and Prediction Validation. Only one parameter k for the GMM.
- **Cons:**
  Struggles with small datasets. Open Challenges: Non-Image-Domains largely unexplored.

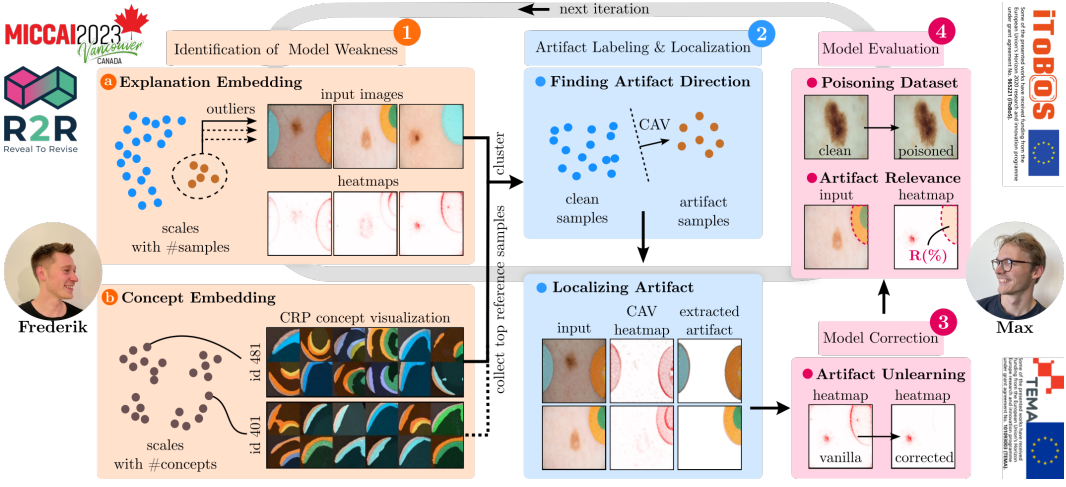
How much Manual Work does XAI Require?

Not much, these days!



What Can We Do With This?

Quickly Understand Model and Data. Annotate Data Beyond GT Categories. *Iterate Fast!*



Reveal to Revise [Pahde et al. 2023]. Successor Paper [Dreyer, Pahde, et al. 2023] accepted at AAAI'24.

Thank you for your attention.
Questions, feedback & input are welcome!

References I

- Achtibat, Reduan et al. (2023). "From attribution maps to human-understandable explanations through Concept Relevance Propagation". In: *Nature Machine Intelligence* 5.9, pp. 1006–1019.
- Anders, Christopher J. et al. (2022). "Finding and removing Clever Hans: Using explanation methods to debug and improve deep models". In: *Information Fusion* 77, pp. 261–295. ISSN: 1566-2535.
- Becker, Sören et al. (2023). "AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark". In: *Journal of the Franklin Institute*.
- Chen, Chaofan et al. (2019). "This Looks Like That: Deep Learning for Interpretable Image Recognition". In: *Advances in Neural Information Processing Systems* 32, pp. 8930–8941.

References II

- Dreyer, Maximilian, Reduan Achtibat, Wojciech Samek, et al. (2023). "Understanding the (Extra-)Ordinary: Validating Deep Model Decisions with Prototypical Concept-based Explanations". In: *arXiv preprint arXiv:2311.16681*.
- Dreyer, Maximilian, Reduan Achtibat, Thomas Wiegand, et al. (2023). "Revealing Hidden Context Bias in Segmentation and Object Detection through Concept-specific Explanations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3828–3838.
- Dreyer, Maximilian, Frederik Pahde, et al. (2023). "From Hope to Safety: Unlearning Biases of Deep Models by Enforcing the Right Reasons in Latent Space". In: *arXiv preprint arXiv:2308.09437*.
- Horst, Fabian et al. (2019). "Explaining the unique nature of individual gait patterns with deep learning". In: *Scientific reports* 9.1, p. 2391.

References III

- Lapuschkin, Sebastian, Alexander Binder, Grégoire Montavon, et al. (2016). "Analyzing classifiers: Fisher vectors and deep neural networks". In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2912–2920.
- Lapuschkin, Sebastian, Alexander Binder, Klaus-Robert Müller, et al. (2017). "Understanding and comparing deep neural networks for age and gender classification". In: *Proc. of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 1629–1638.
- Lapuschkin, Sebastian, Stephan Wäldchen, et al. (2019). "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn". In: *Nature Communications* 10, p. 1096.

References IV

- Pahde, Frederik et al. (2023). "Reveal to Revise: An Explainable AI Life Cycle for Iterative Bias Correction of Deep Models". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan et al. Cham: Springer Nature Switzerland, pp. 596–606. ISBN: 978-3-031-43895-0.
- Slijepcevic, Djordje et al. (2021). "Explaining machine learning models for clinical gait analysis". In: *ACM Transactions on Computing for Healthcare (HEALTH)* 3.2, pp. 1–27.