Towards Actionable XAI

Artificial Intelligence Doctoral Academy

Dr. Sebastian Lapuschkin

Dept. of Artificial Intelligence
Fraunhofer Heinrich Hertz Institute

Speaker Intro
Dr. rer. nat. Sebastian Lapuschkin (né Bach)

[google scholar]

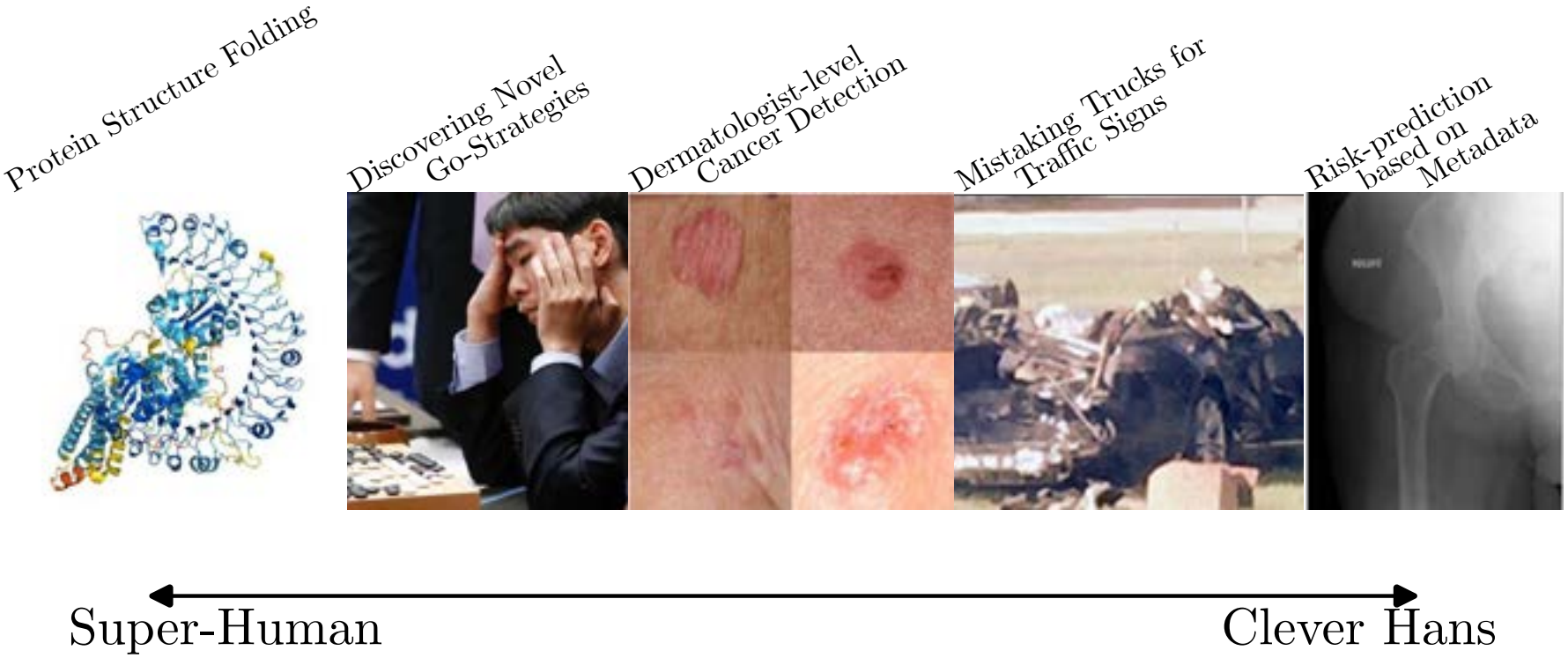
- **2022 - Method: CRP & RelMax: XAI 2.0 [Achtibat *et al.* 2022]**
- 2022 - Method: ClArC [Anders, Weber, *et al.* 2022]
- 2021 - Head of XAI Group at Fraunhofer HHI
- 2019 - Method: SpRAy [Lapuschkin, Wäldchen, *et al.* 2019]
- 2018 - PhD Machine Learning / XAI at TU Berlin
- 2017 - Method: DTD [Montavon *et al.* 2017]
- 2016 - Method: MoRF [Samek, Binder, *et al.* 2017]
- 2016 - 1st Encounter: Clever Hans [Lapuschkin, Binder, Montavon, Müller, *et al.* 2016a]
- 2015 - Method: LRP [Bach, Binder, Montavon, *et al.* 2015]
- 2013 - MSc CS (ML+XAI) at TU Berlin

Fraunhofer Heinrich Hertz Institute

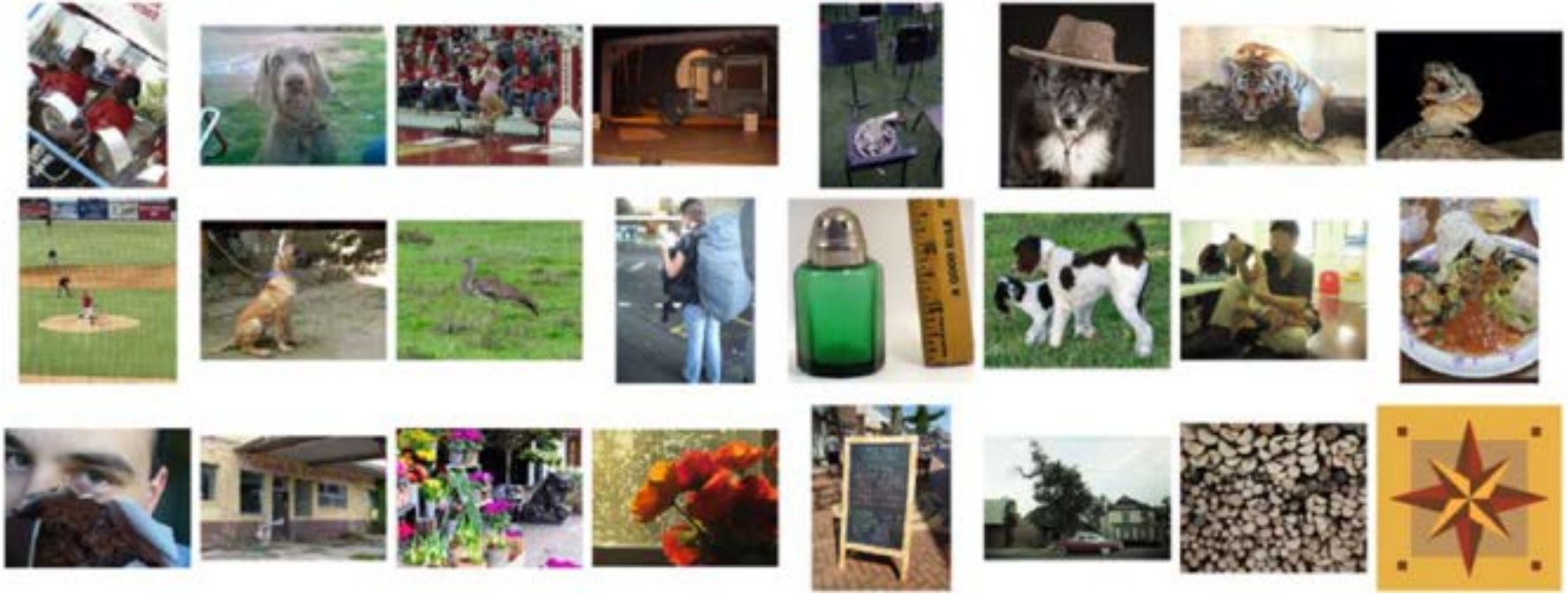
Roadmap of this Talk

- Brief Intro: Machine Learning & Artificial Intelligence
- Local XAI, Applications & Limitations
- Towards Actionable XAI with
  Concept Relevance Propagation and **Relevance Maximization**

The Power Spectrum of AI



Data is Major Driver in AI



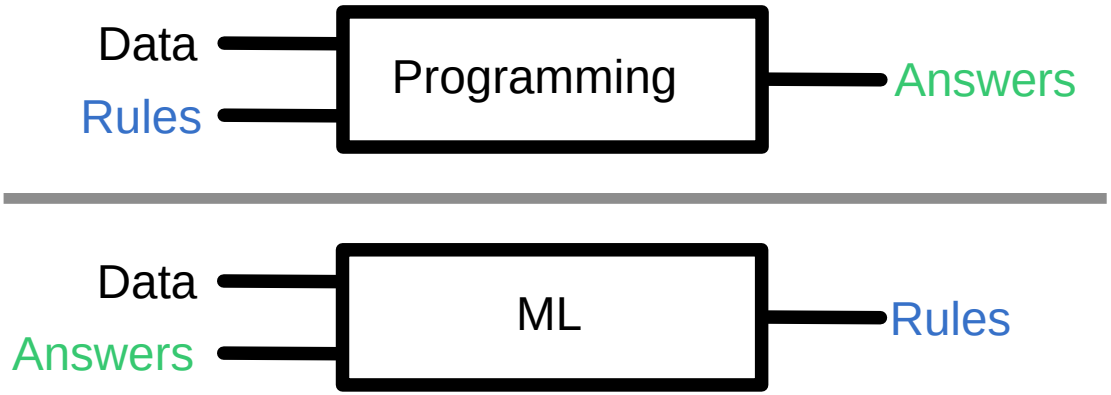
Source: ImageNet [Russakovsky *et al.* 2015]

However, not All Data is Flawless
and all Data is not Flawless

Left: Unwanted Correlations in Pascal Datasets [Everingham *et al.* 2007]

Right: Labelling Errors in ImageNet [Stock *et al.* 2018]

AI vs "Programming"
Avoid Issues by Hand-crafting a Well-behaved System?

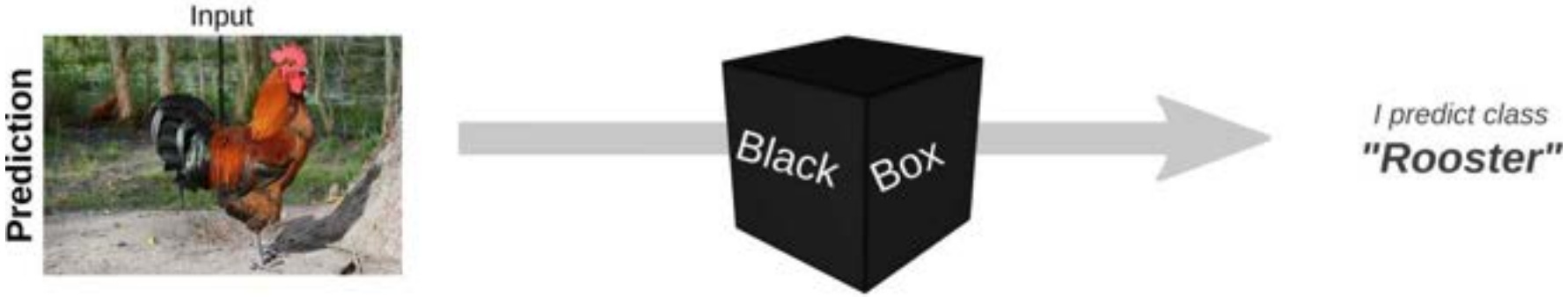



```
.is_cat()
 .is_cute()
 .has_whiskers()
 .can_purr()
 .has_ears()
   .cat_ears()
   .not_dog_ears()
      ?
. . .
```

Would likely fail at the complexity of typical current-day ML problems.

And, would we even be aware of these data issues?

Trade-off: Transparency vs Performance (?)



So What Does the Model Learn to Predict on?

"boat"!
but why?

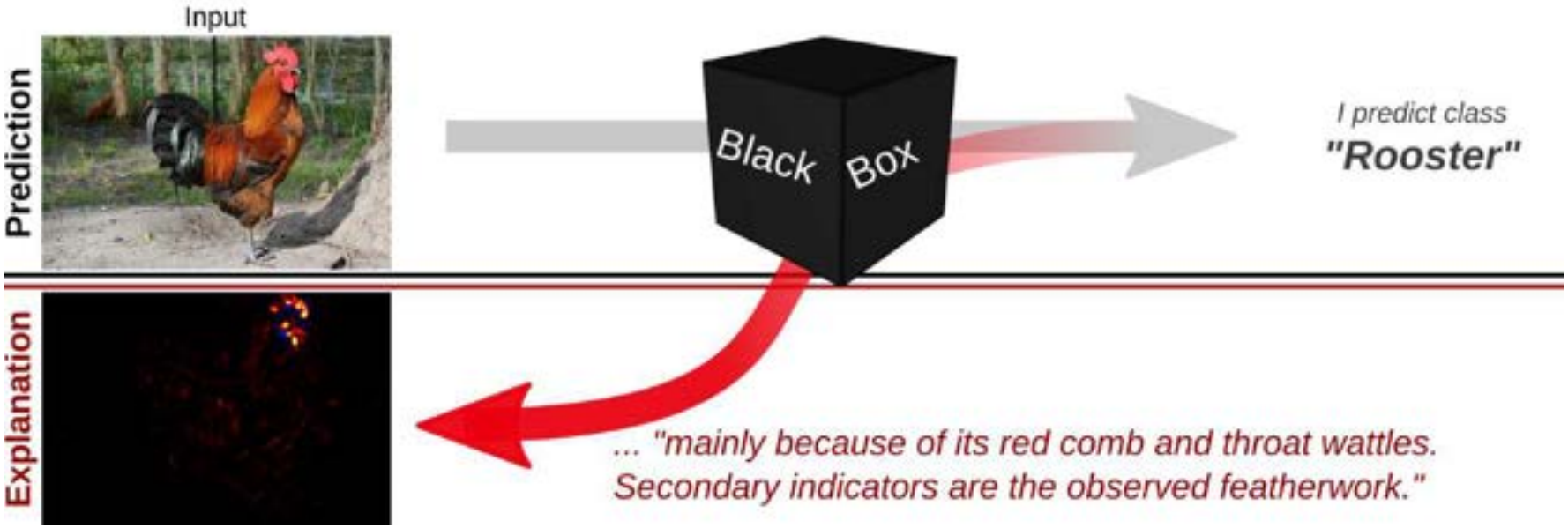
"horse"!
but why?

"horse"!
but why?

"horse"!
but why?

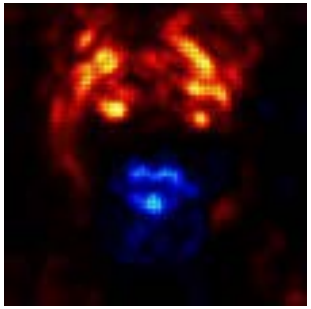
Critical to know esp. in high-risk scenarios, e.g. medical domain, autonomous driving, finance, ...

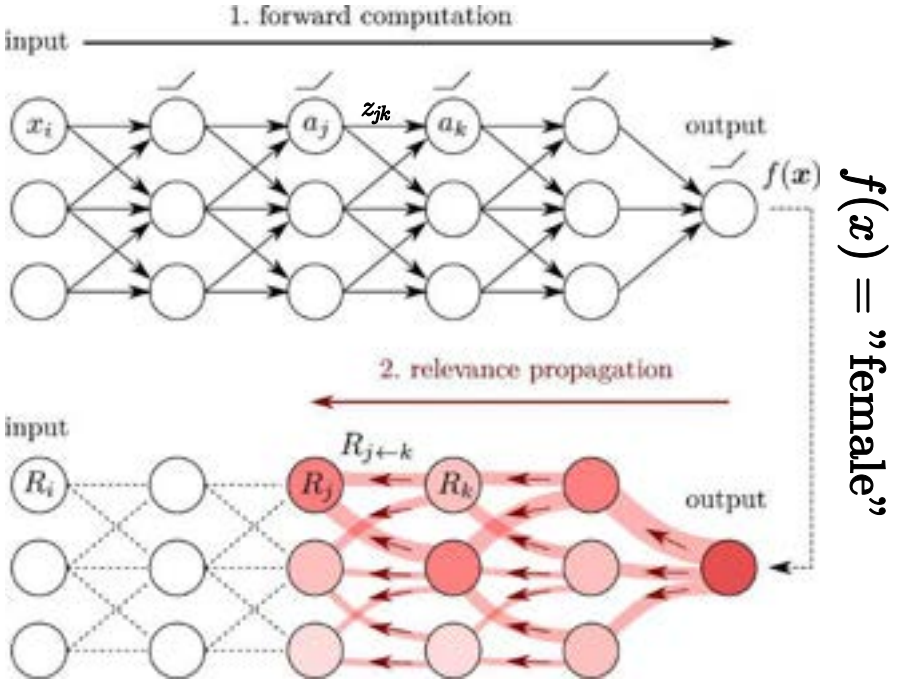
Enter Local XAI



Countless approaches: Saliency [Morch *et al.* 1995], DeconvNet [Zeiler *et al.* 2014], LRP [Bach, Binder, Montavon, *et al.* 2015], DTD [Montavon *et al.* 2017], IG [Sundararajan *et al.* 2017], SHAP [Lundberg *et al.* 2017], Grad-CAM [Selvaraju *et al.* 2017], SmoothGrad [Smilkov *et al.* 2017], PredDiff [Zintgraf *et al.* 2017], MPert [Fong *et al.* 2017], ...

How does XAI work?
Layer-wise Relevance Propagation in brief





LRP:

(1) decompose

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

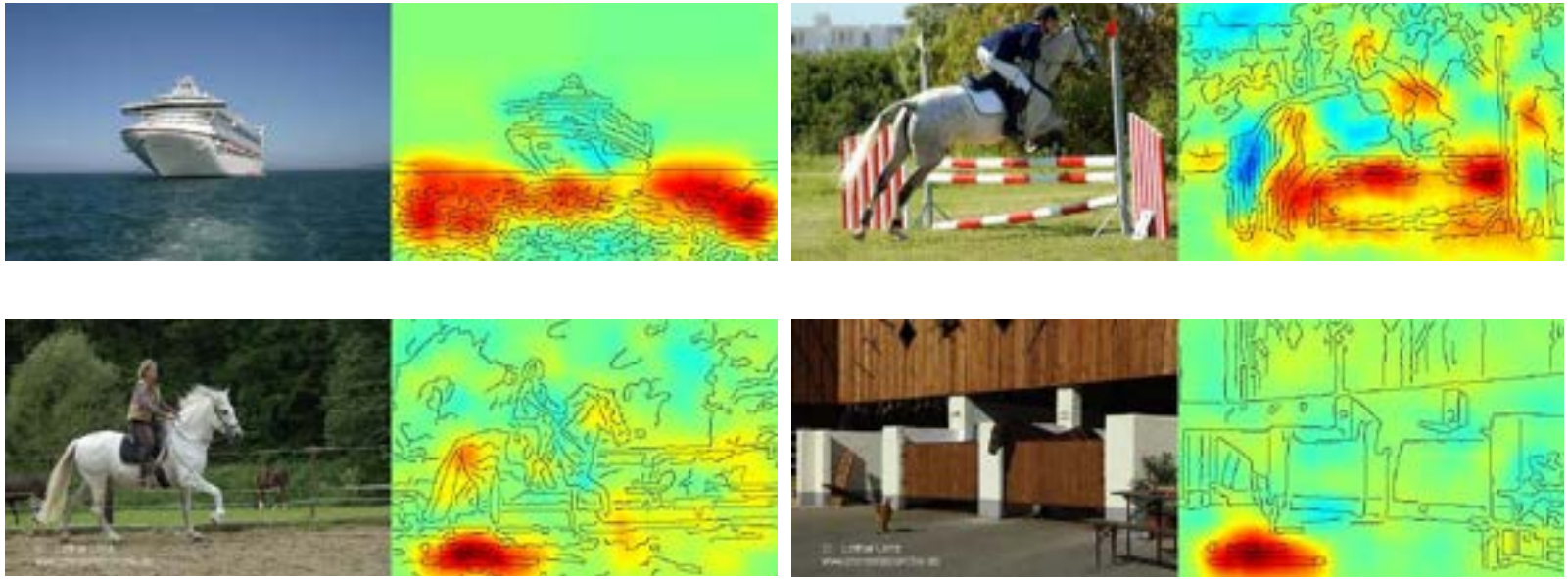
(2) aggregate

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

see [Bach, Binder, Montavon, *et al.* 2015], [Kohlbrenner *et al.* 2020], [Samek, Montavon, *et al.* 2021].

Software: [Lapuschkin, Binder, Montavon, Müller, *et al.* 2016b],[Alber *et al.* 2019],[Anders, Neumann, *et al.* 2021]

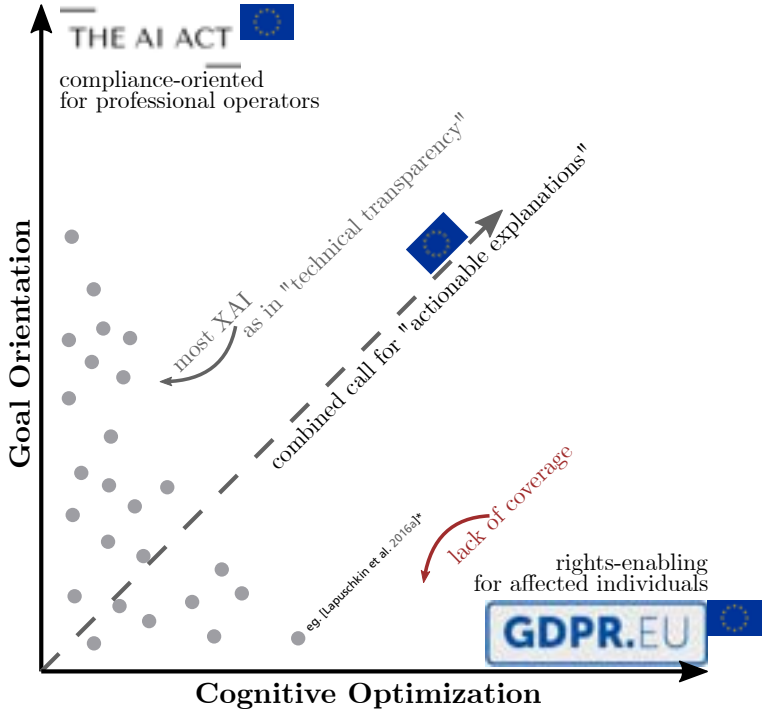
XAI in the Past
Provision of Hints into Predictions and Datasets



Pretty obvious results: XAI *may* assist in uncovering model behavior [Lapuschkin, Binder, Montavon, Müller, *et al.* 2016a], as a basic form of *knowledge discovery*, eg for fixing data sources (actionability!)

Here, on (former [Chatfield *et al.* 2011]) SoA model on the PASCAL VOC [Everingham *et al.* 2007] benchmark.

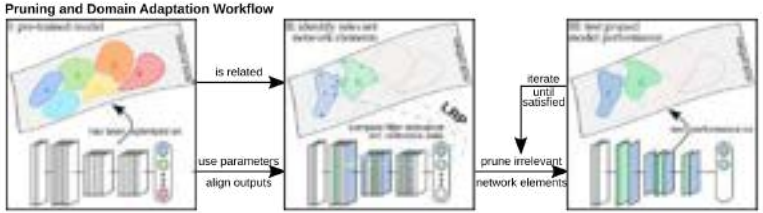
Defining Actionable XAI:
"Explanations allow the Stakeholder to Act"



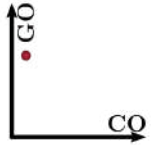
Based on [Hacker *et al.* 2022]: ``*Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond*''

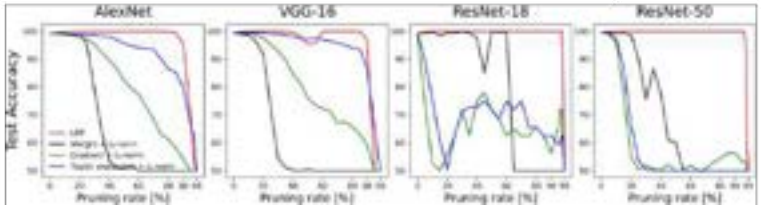
Recent Demonstrations of (Actionable) XAI

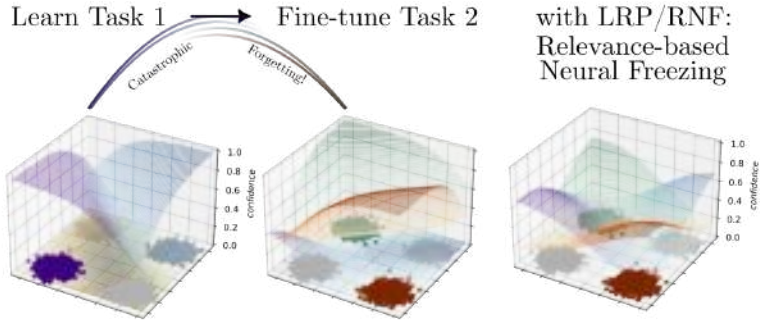
XAI for increasing model efficiency



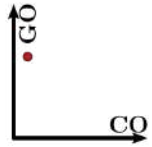
Improving Model Efficiency by Pruning / Quantizing irrelevant filters, neurons or weights with XAI.
(Yeom et al., 2021), (Becking et al., 2022)







Prevent Catastrophic Forgetting in Neural Network fine-tuning by reducing learning-elasticity of task-relevant network elements with XAI.
(Ede et al., 2022)

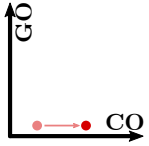


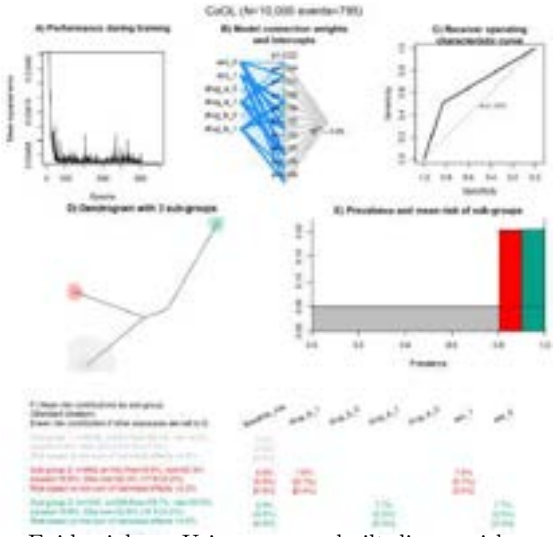
Refs: [Yeom *et al.* 2021] [Becking *et al.* 2021] [Ede *et al.* 2022]: Goal oriented, no / limited cognitive optimization.

Recent Demonstrations of (Actionable) XAI

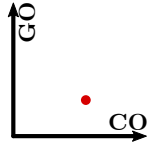
XAI for knowledge discovery

Large-scale analysis of attributions
with Spectral Relevance Analysis:
Exploring model behavior for systematic patterns.
(Lapuschkin et al., 2019)





Epidemiology: Using purpose-built disease risk predictors
to identify (novel) sufficient cause combinations for
understanding deseases.
(Rieckmann et al., 2022)

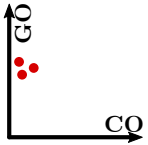


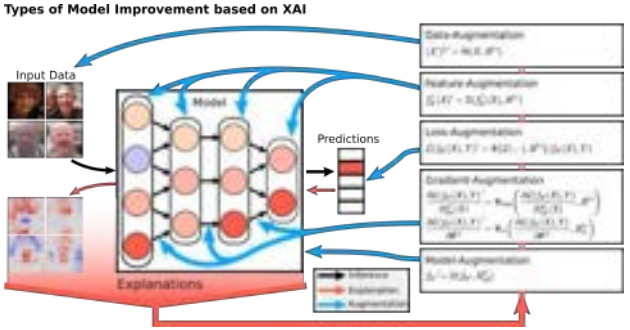
Refs: [Lapuschkin, Wäldchen, *et al.* 2019] [Rieckmann *et al.* 2022]: Some cognitive optimization

Recent Demonstrations of (Actionable) XAI

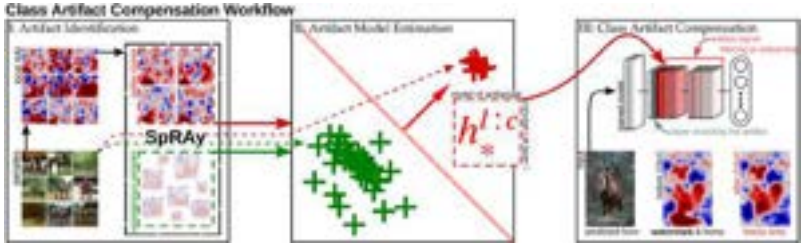
XAI for fixing model behavior

Guide or Restrict model behavior with ground truth knowledge during training:
e.g. RRR (Ross et al., 2017), CAIPI (Teso et al., 20219),
CD-EP (Rieger et al., 2020), XIL (Schramowski et al., 2020)

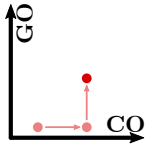




Overview paper on XAI-based
model-improvement: (Weber et al., 2022)



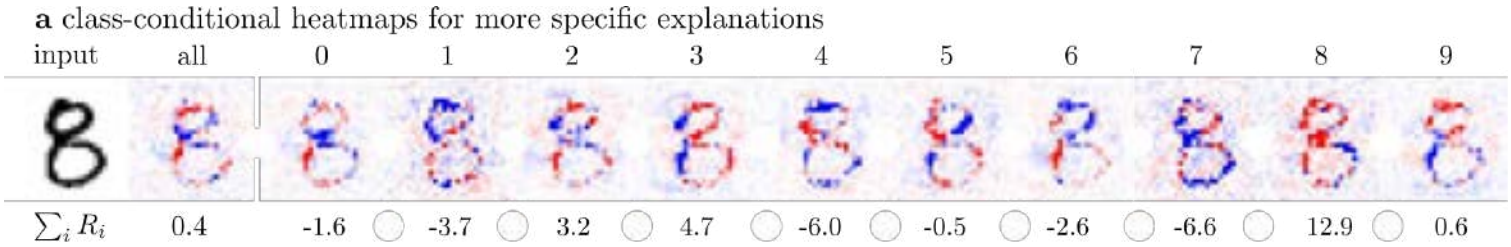
Class Artifact Compensation:
First identify systematically used features, and then unlearn or suppress if (un)desired.
(Anders et al., 2022) (Pahde et al., 2022)



Refs: [Ross *et al.* 2017] [Teso *et al.* 2019] [Rieger *et al.* 2019] [Schramowski *et al.* 2020]
[Anders, Weber, *et al.* 2022] [Pahde *et al.* 2022] [Weber *et al.* 2022]

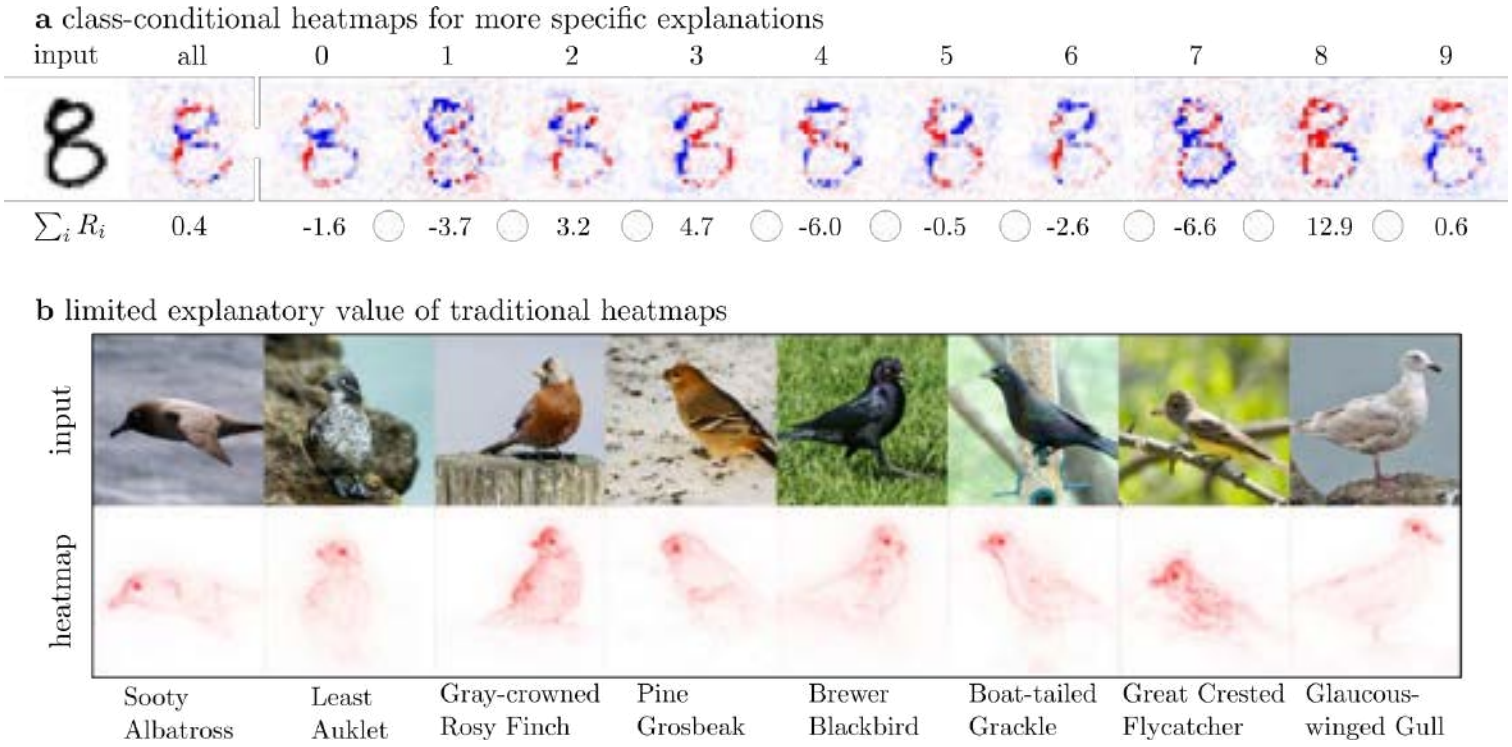
Limitations of Local XAI

What about Actionability via Cognitive Optimization?

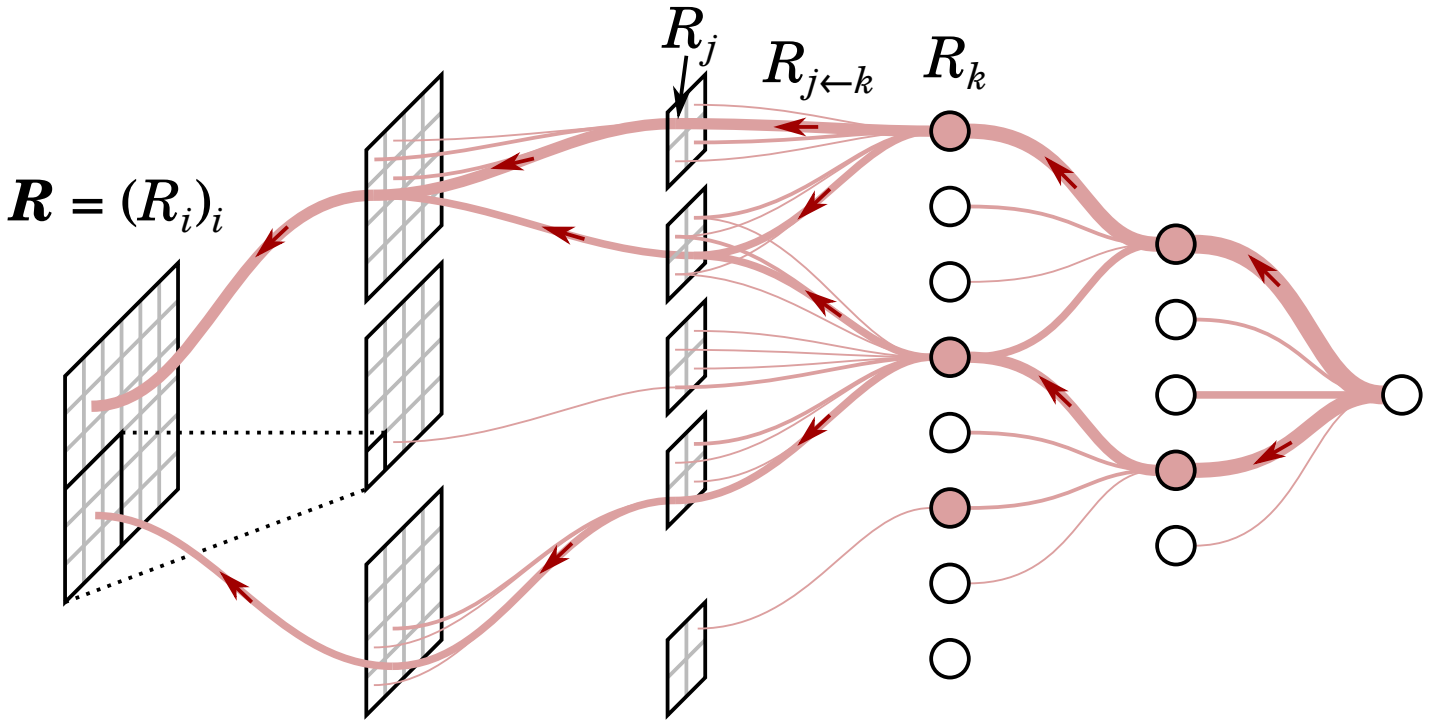


Limitations of Local XAI

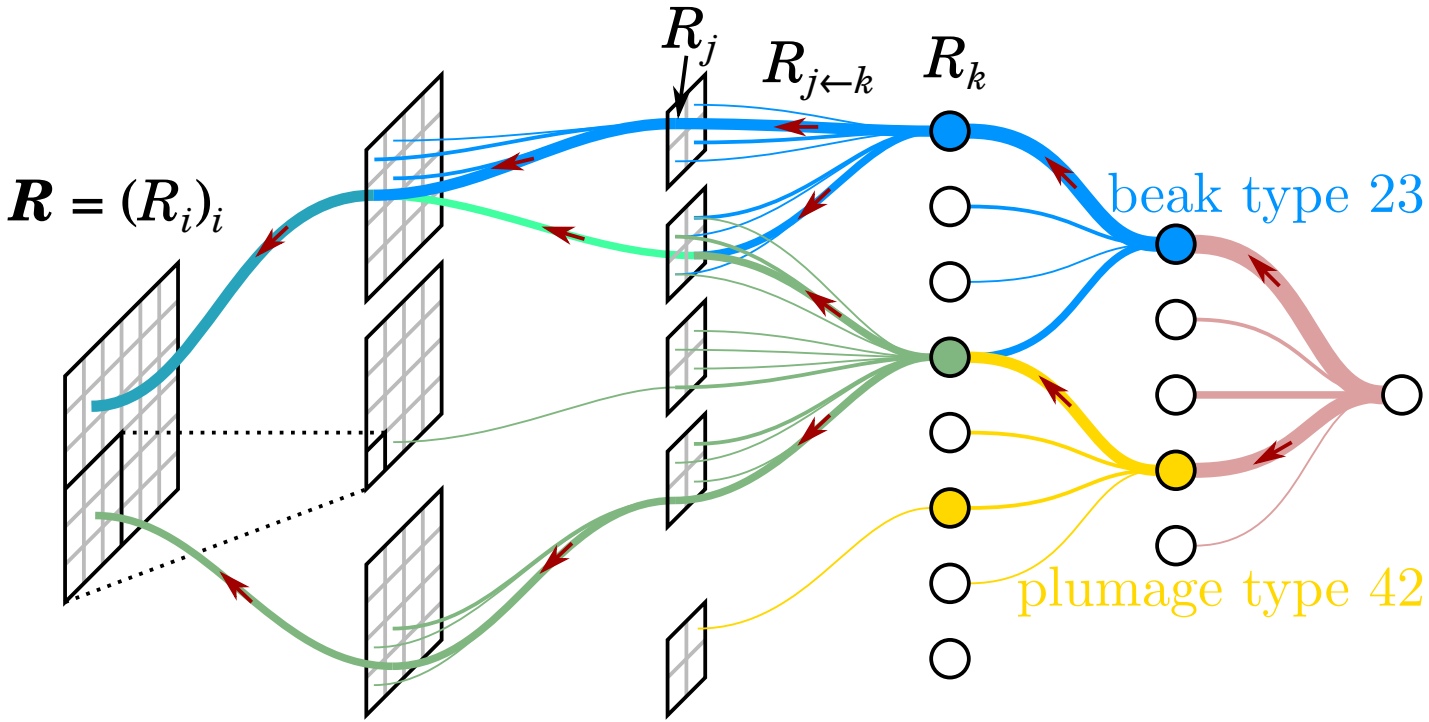
What about Actionability via Cognitive Optimization?



Problem:



Problem: Superposition of Feature Attributions
Opposes Cognitive Optimization!



Enter NextGen XAI
Respecting the Complexity of Deep Learning Inference

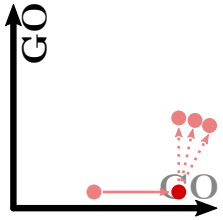
e.g. ProtoPNet [C. Chen *et al.* 2019] ...

we focus on CRP & RelMax
[Achtibat *et al.* 2022] for reasons*

*) reasons: it is our own work ;) ,

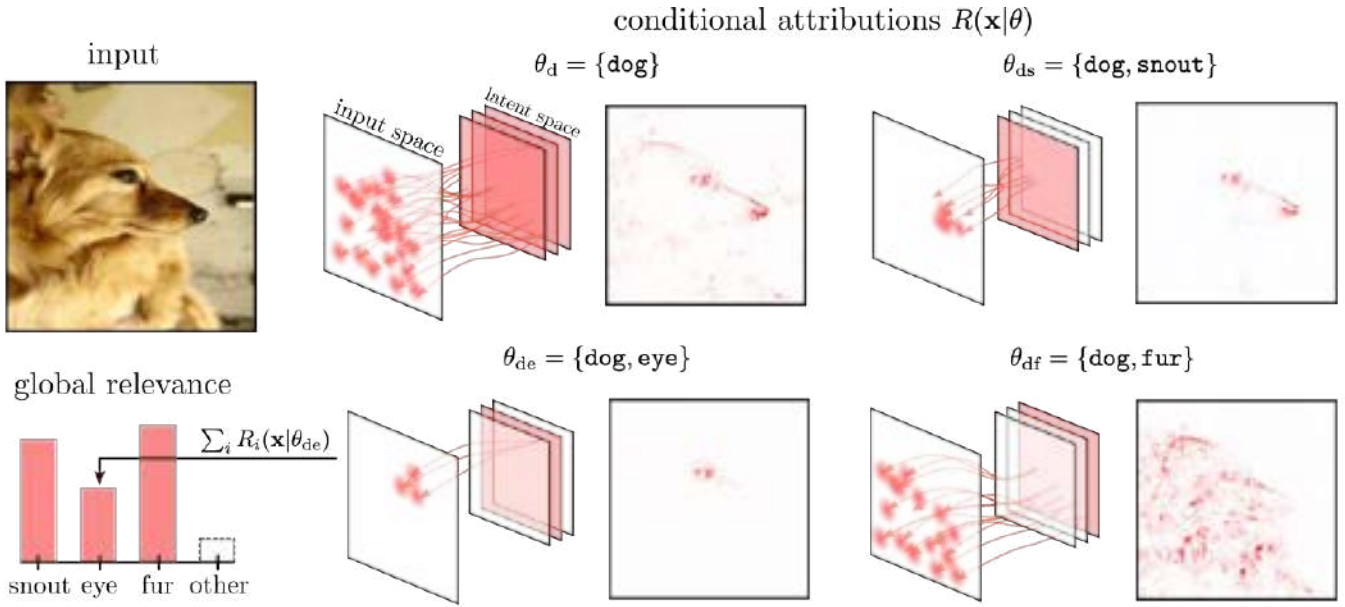
but also;
completely post-hoc, no extra requirements,
delivers **global-local** explanations,

widely applicable & highly efficient!



[I] Concept Relevance Propagation

Respect Concurrency (plus temporarily make some minor assumptions)

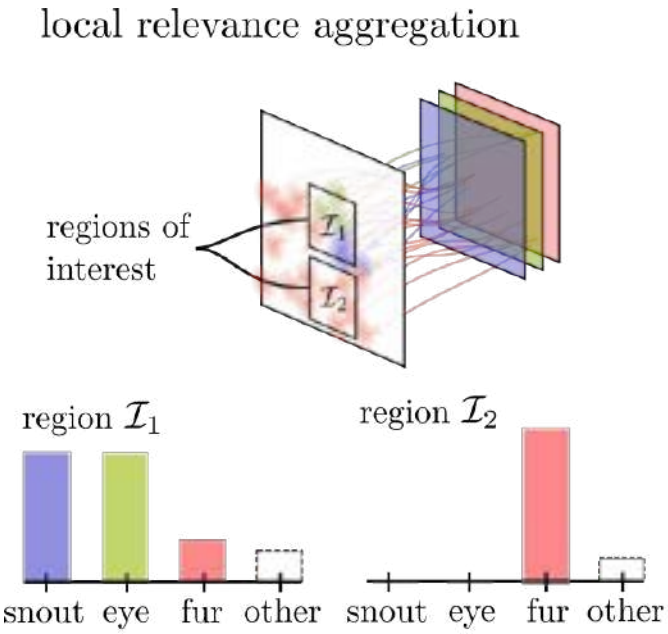


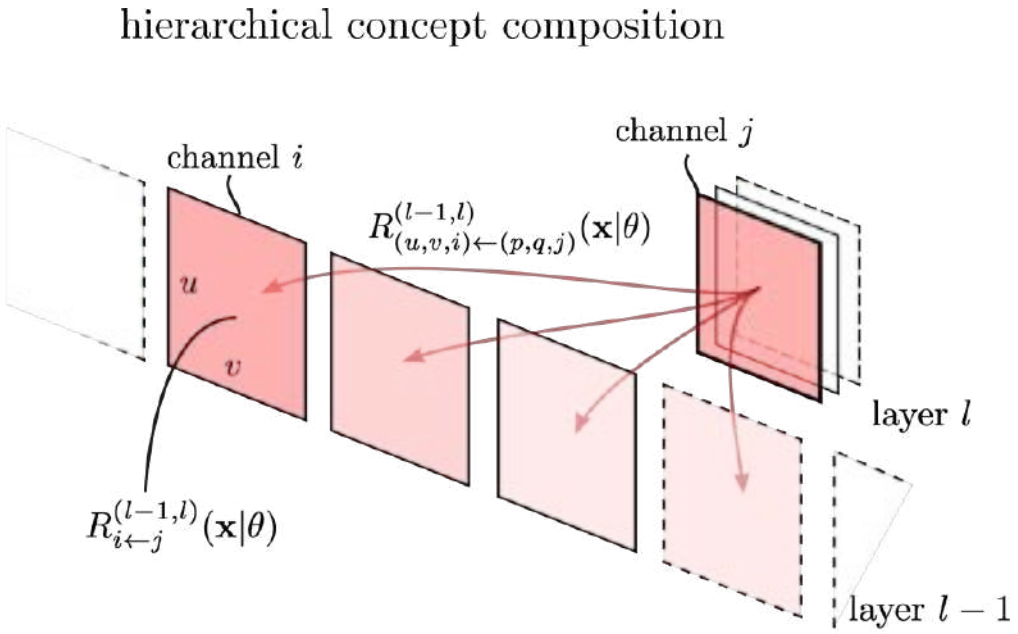
CRP:

$$R_{j \leftarrow k}(x|\theta) = \frac{z_{jk}}{z_k} \cdot \sum_{c_l \in \theta} \delta_{kc_l} R_k(x|\theta)$$

extends the LRP-decomposition step

CRP enables Novel Analyses and Insights!





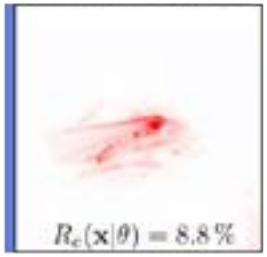
Conservativity of CRP enables meaningful aggregation strategies

CRP enables Novel Analyses and Insights!

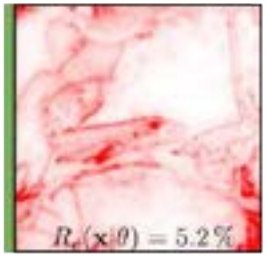
input image

LRP heatmap

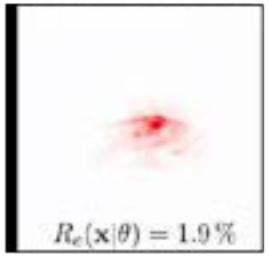
conditional CRP heatmaps of the top-5 most relevant latent concepts



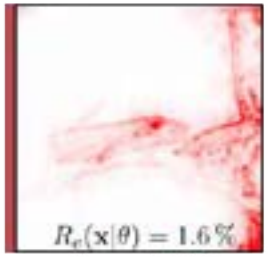
channel 469
lizardy features ?



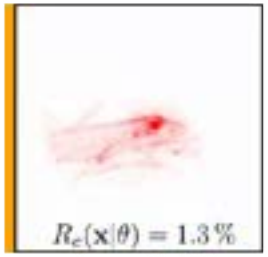
channel 35
background / vegetation ?



channel 89
(lizard) head / eye ?



channel 316
lizard and / on branches ?



channel 161
more lizardy features ?

CRP refines the "where".

(at pretty much ZERO COST)

What about the "what"?

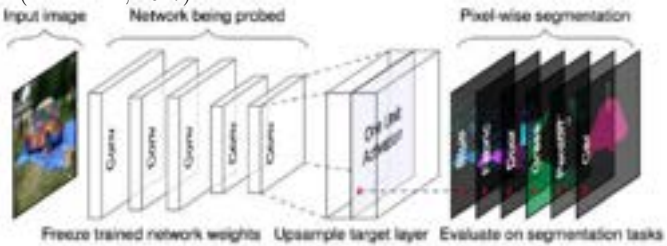
Addressing the "What"-Question
A (very) Brief Review of Global XAI

(Olah et al., 2017)
generative approaches
(early works starting 2009)

All Methods based on *Unit Activations*.
Many introduce *Additional Requirements*.

(Zeiler et al., 2014) feature visualization

(Bau et al., 2017) Network Dissection



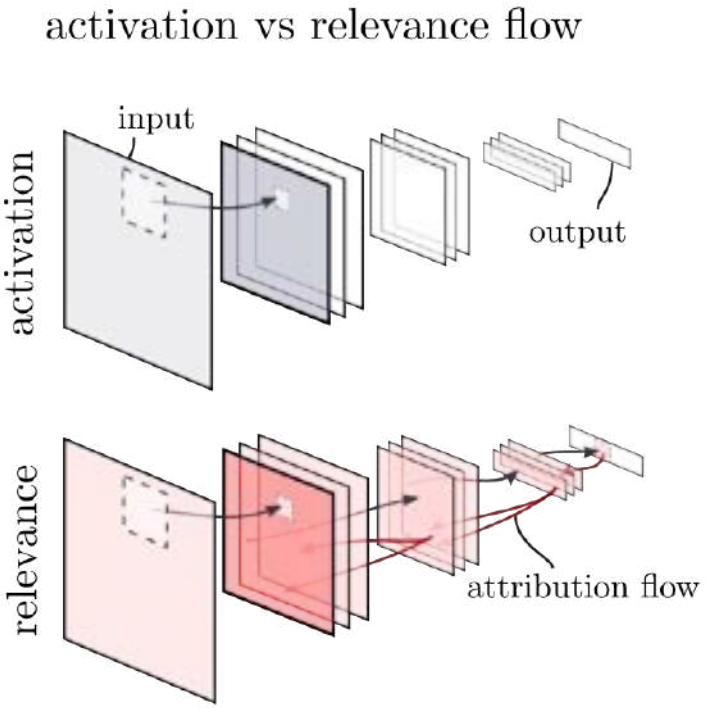
(Chen et al., 2020) data-based activation maximization

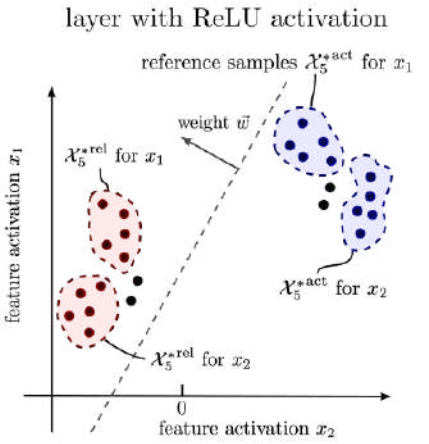
Consensus: DNNs learn abstract, human-understandable features in latent space

Further Reading: [Zeiler *et al.* 2014] [Olah *et al.* 2017] [Kim *et al.* 2018] [Bau *et al.* 2017] [Hohman *et al.* 2019] ...

[II] Relevance Maximization

or why Activation Maximization (alone) is not the Answer

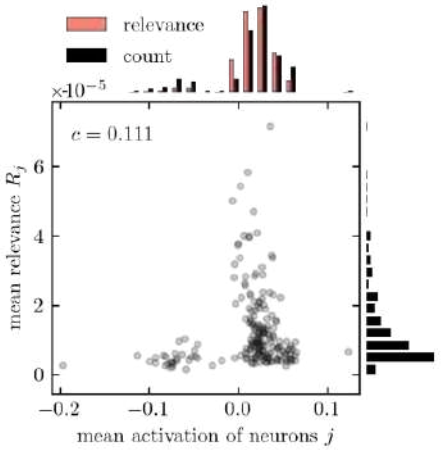




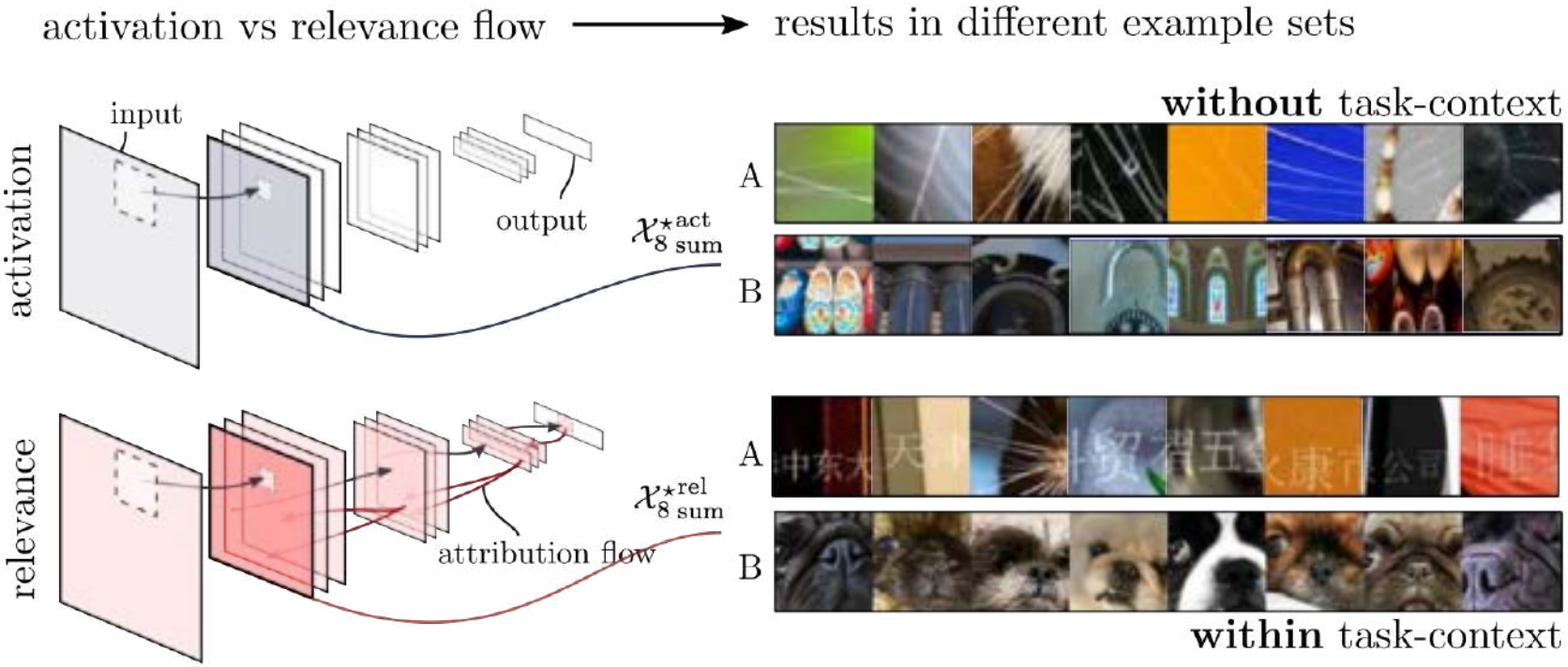
activation = stimulation
without task-context

relevance = usefulness
within a task-context

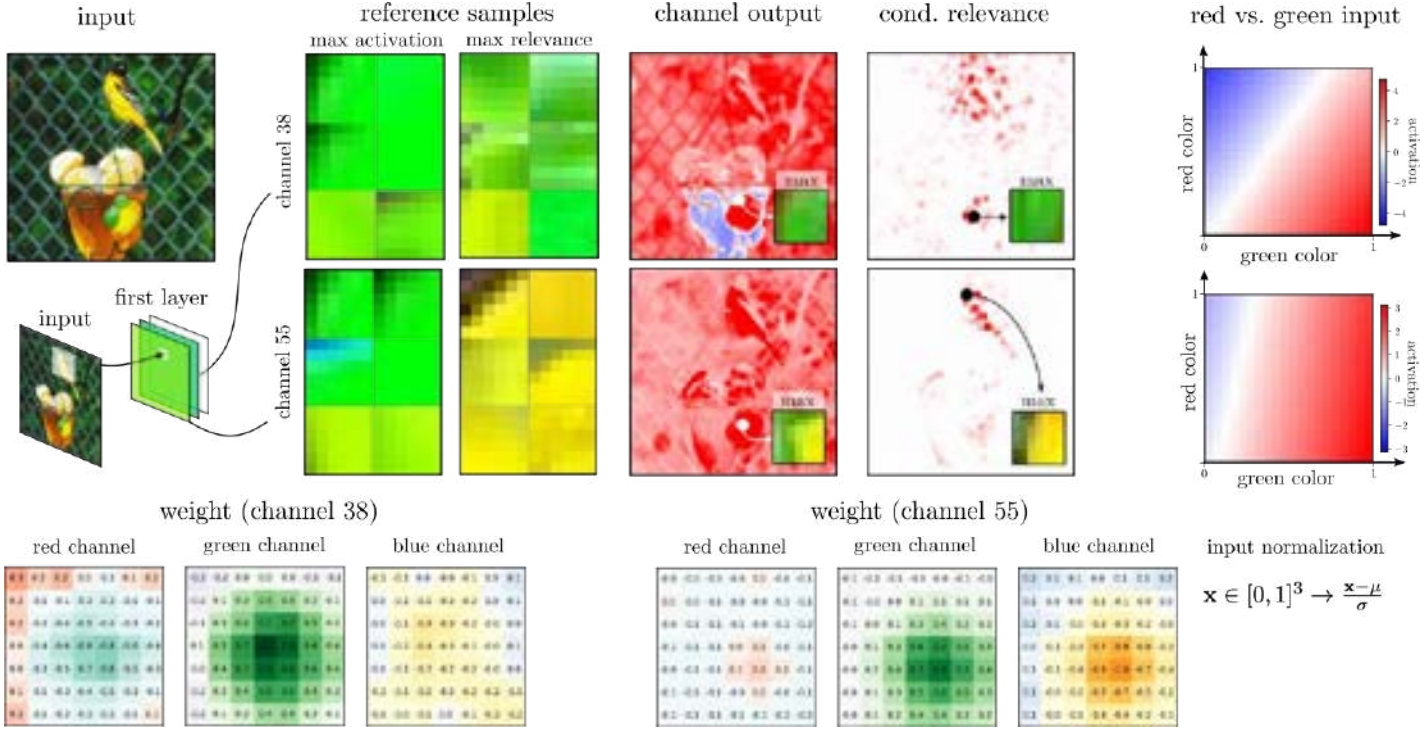
no correlation between
activation and relevance



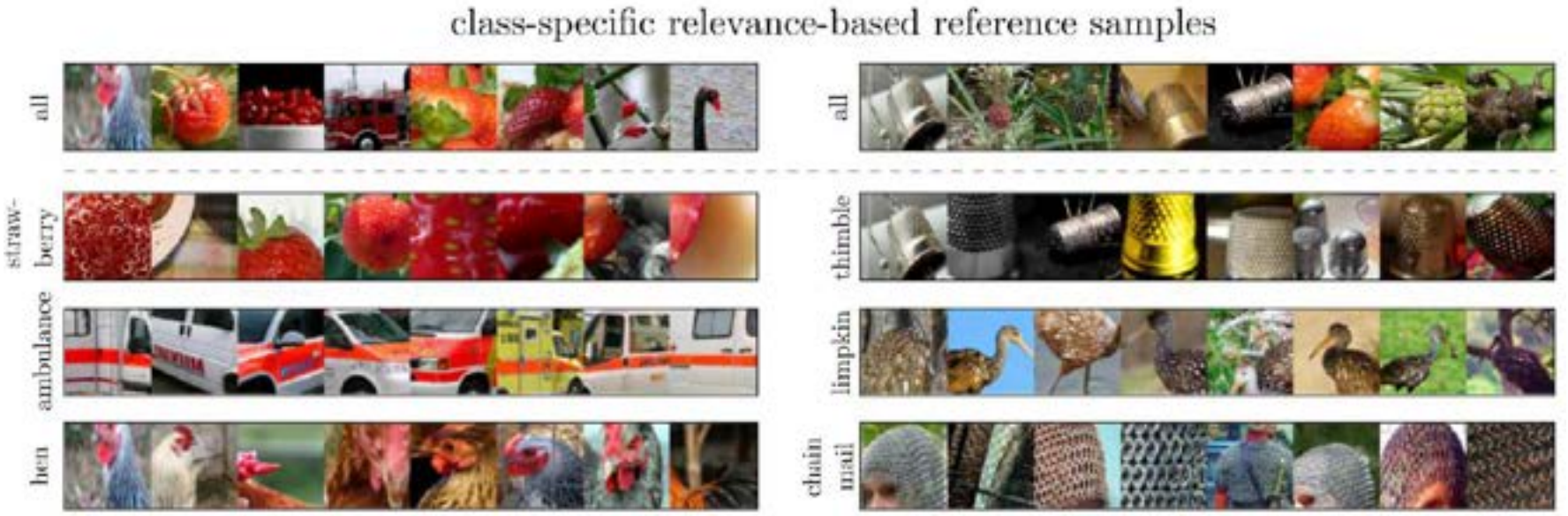
[II] Relevance Maximization
or why Activation Maximization (alone) is not the Answer



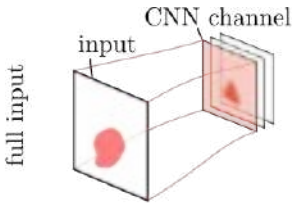
[II] Relevance Maximization
as a Stable Measure of Utility



[II] Relevance Maximization
as a Contextual Measure



Explaining the Explanations
for Cognitive Optimization



typical scenario in literature:

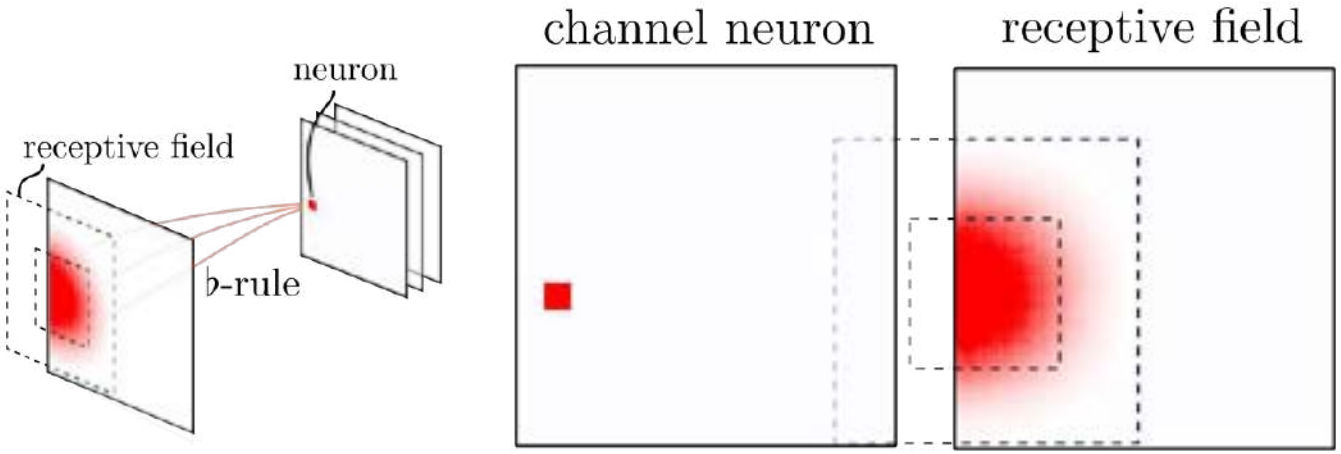
provide **full input-sized** explanatory examples.

?

which one is / are the relevant feature(s) ?

See, eg, [Z. Chen *et al.* 2020]

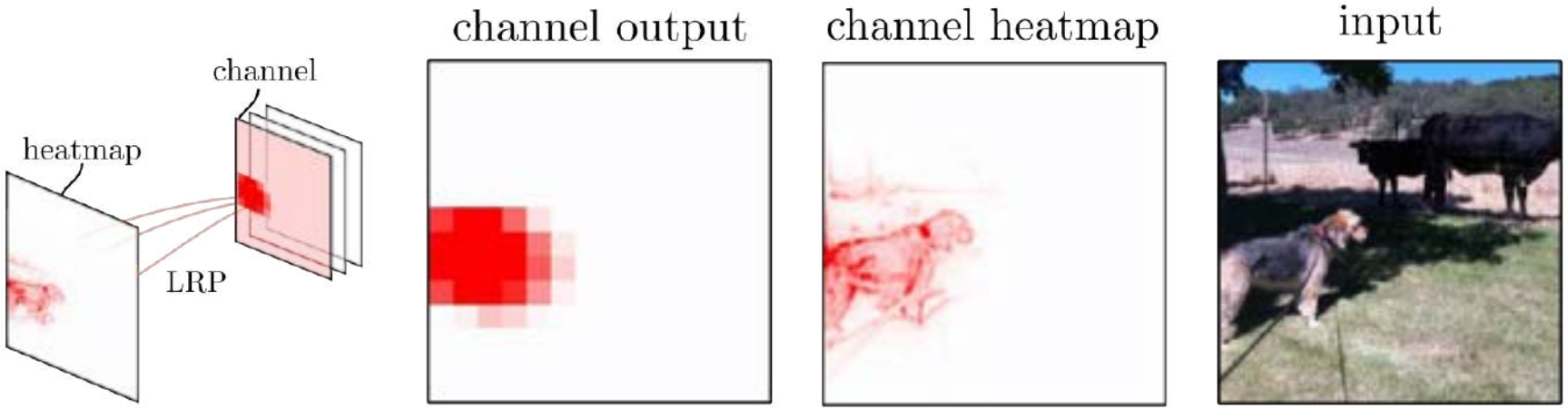
Explaining the Explanations
for Cognitive Optimization



Receptive Field Computation may be complex [Araujo *et al.* 2019],
but can be made easy with XAI [Bach, Binder, Müller, *et al.* 2016] [Kohlbrenner *et al.* 2020].

Explaining the Explanations
for Cognitive Optimization

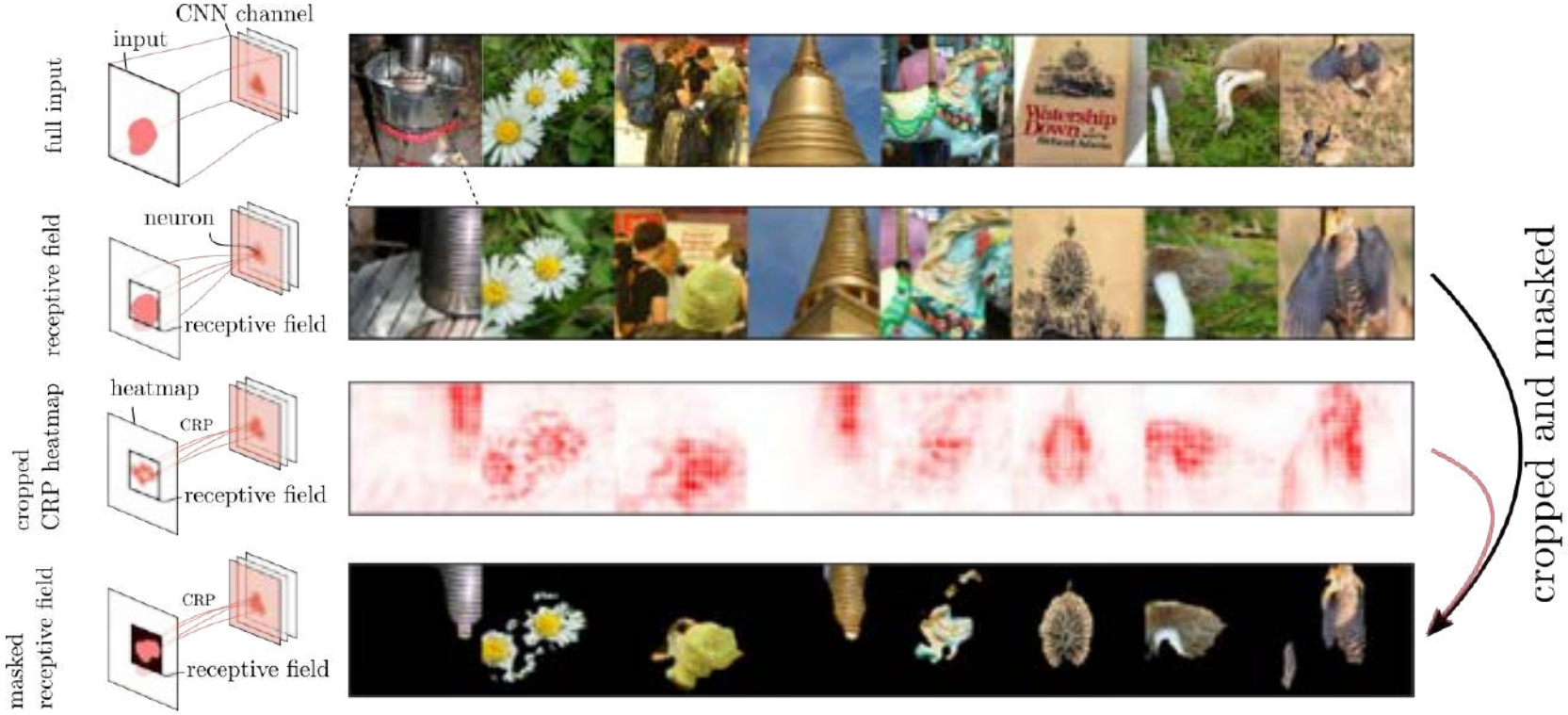
Explaining the Explanations
for Cognitive Optimization



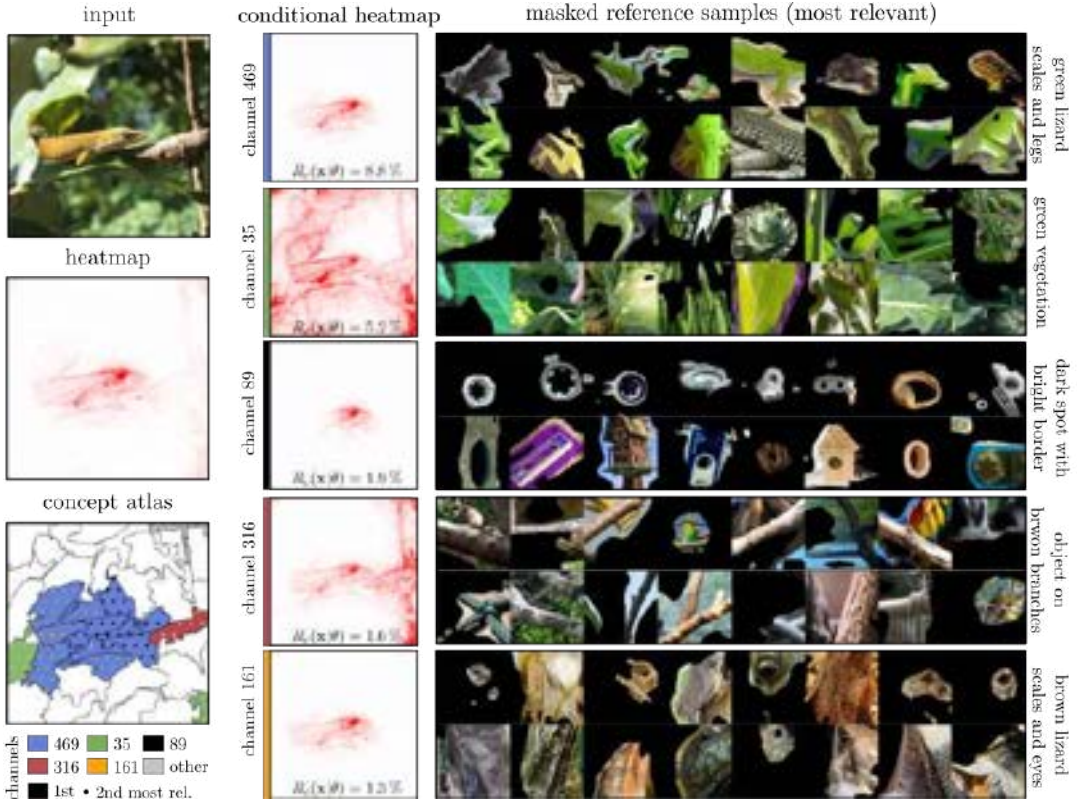
Latent features can be explained with LRP/CRP, if treated as (concept) detectors.

Explaining the Explanations
for Cognitive Optimization

Explaining the Explanations
for Cognitive Optimization

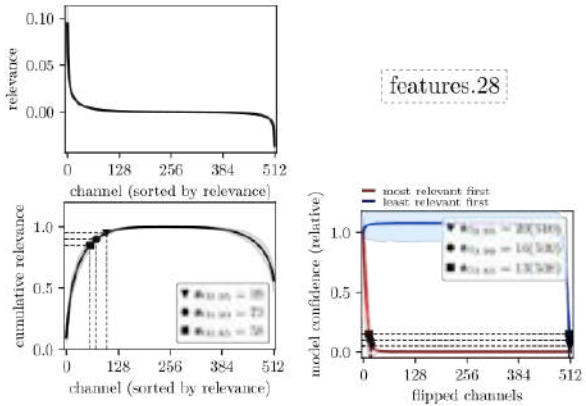


The Concept Atlas: Combining CRP and RelMax

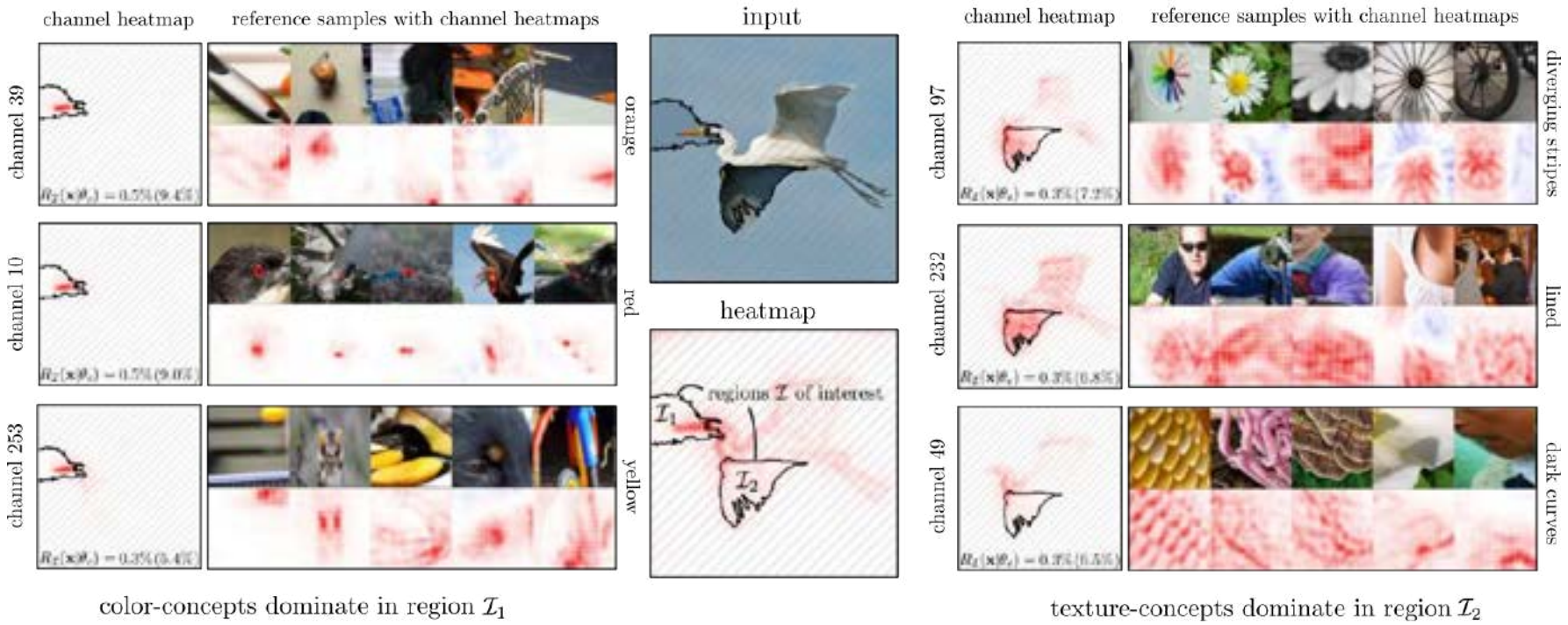


but how much info is needed to understand?

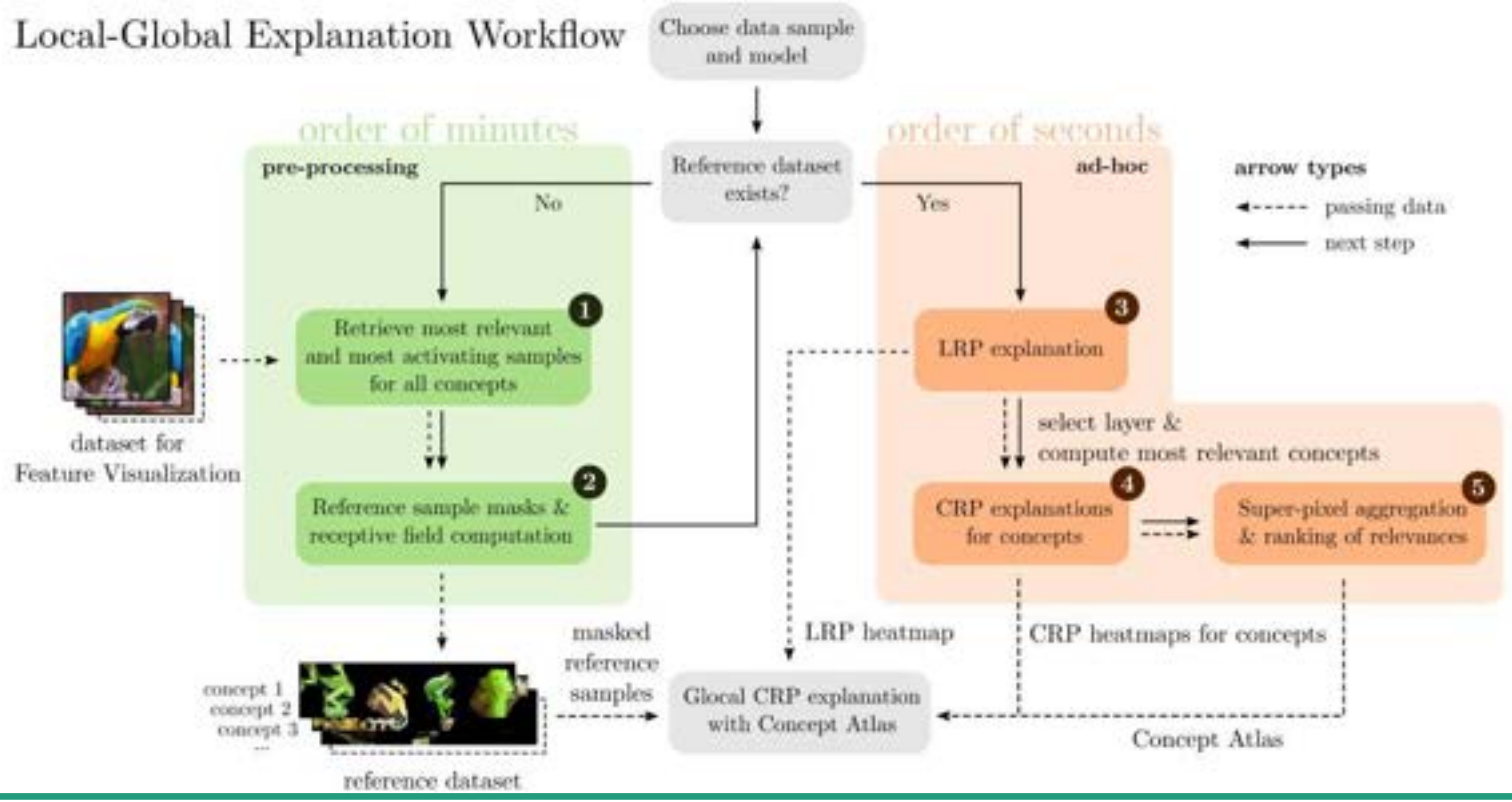
not all that much, fortunately!



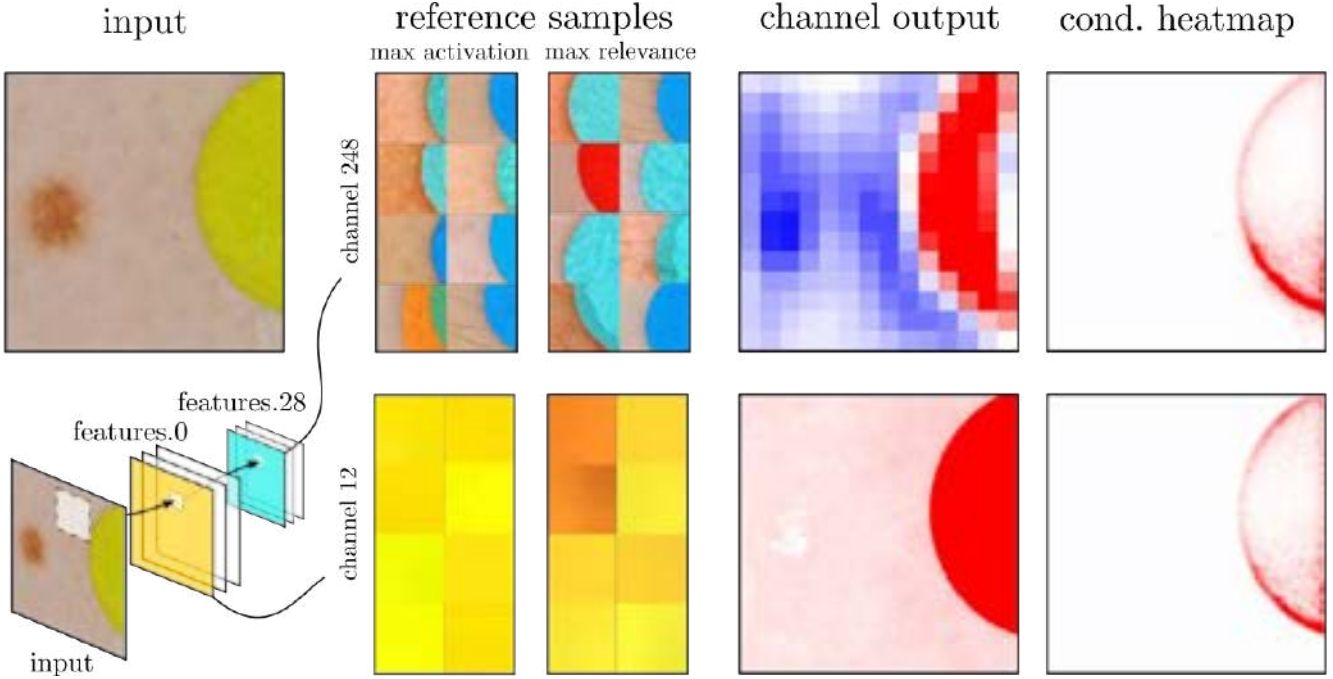
Local Analyses as a Bottom-Up Complement



Run-time Cost of CRP & RelMax

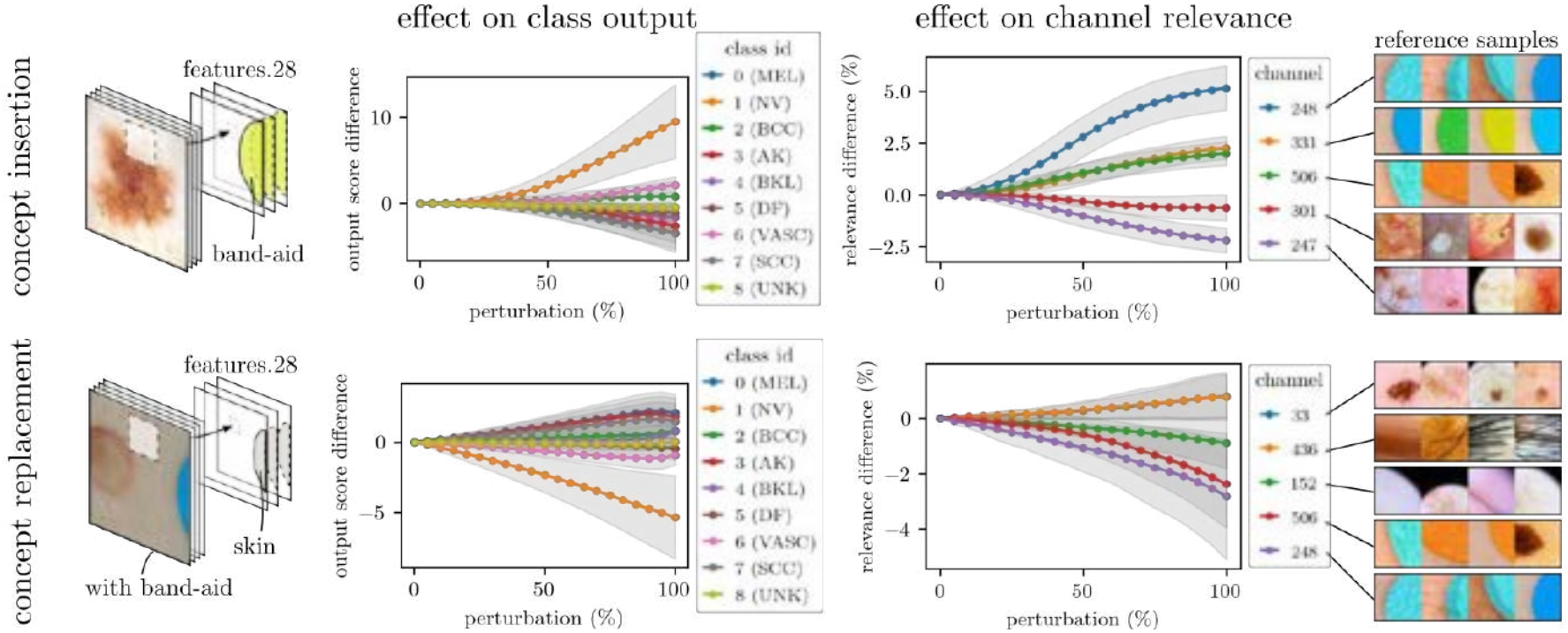


Towards Actionability
Investigating Suspicious Encodings

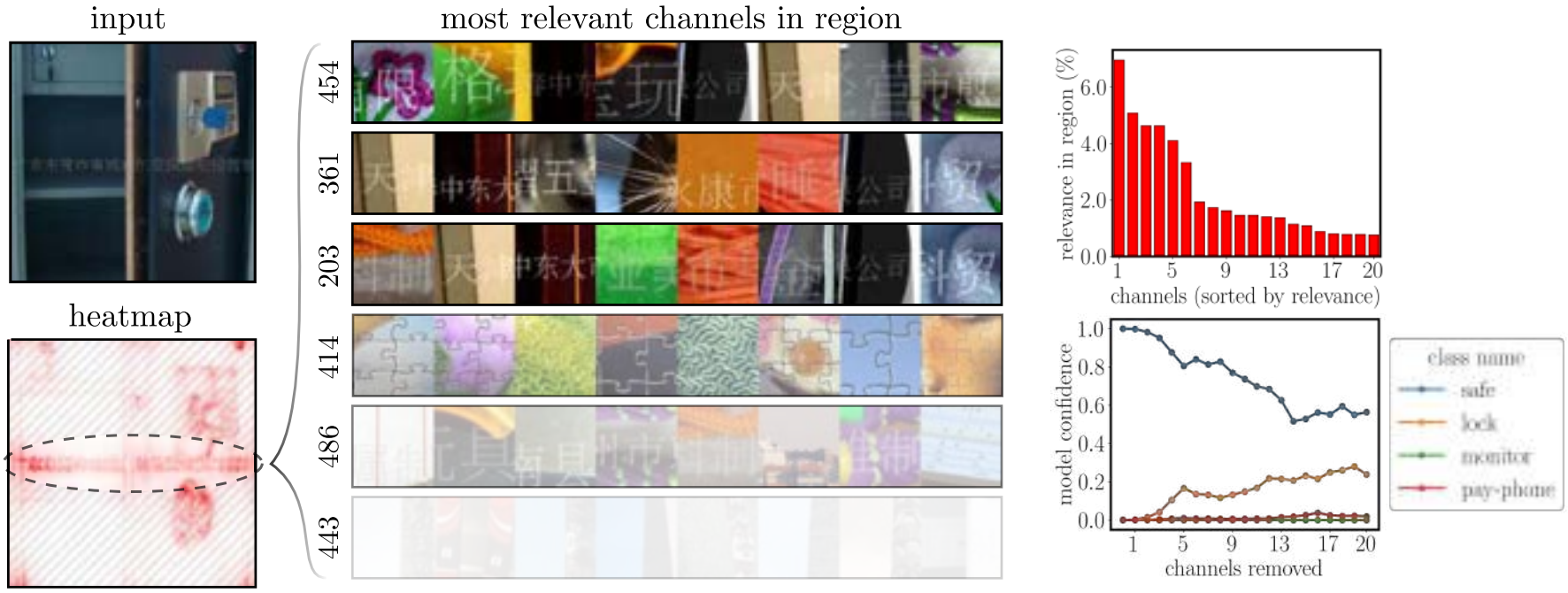


Example: ISIC [Combalia *et al.* 2019][Codella *et al.* 2018][Tschandl *et al.* 2018] has known problems [Rieger *et al.* 2019][Anders, Weber, *et al.* 2022]. **How to identify, evaluate & *precisely* fix post-hoc?**

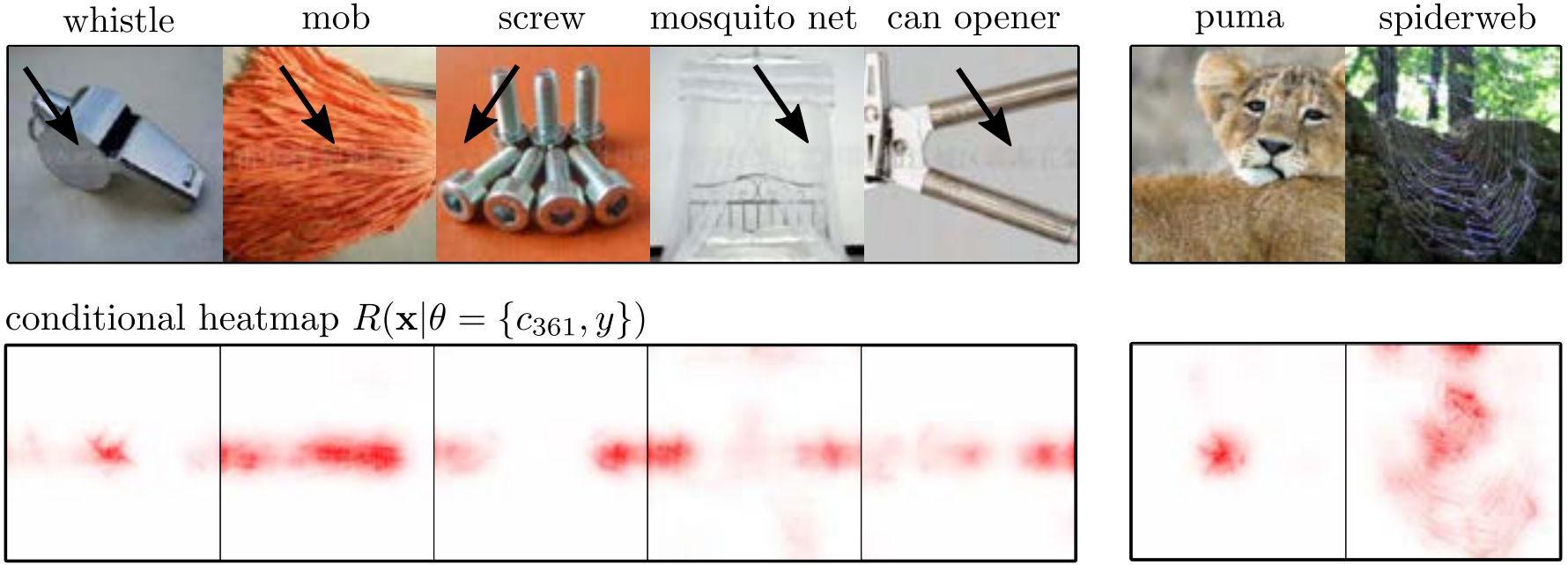
Towards Actionability
Investigating Suspicious Encodings



Towards Actionability
Concept-based Reverse Search

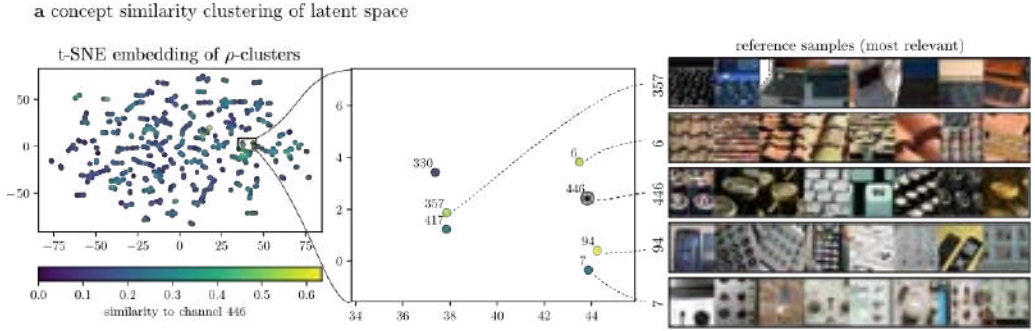


Towards Actionability
Concept-based Reverse Search



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

Towards Actionability
Outlook: Understanding Feature Subspaces & Interactions

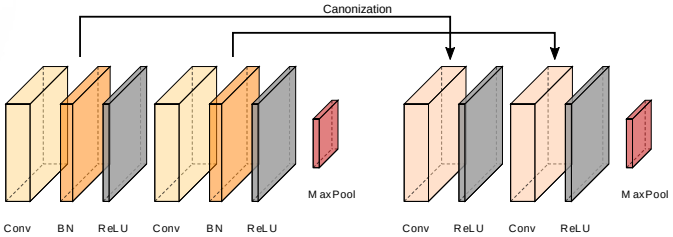


Global-Local XAI:
new tools for understanding and revising AI
empowering **novel applications** of XAI
via **precise informed interactions** with the AI

Other Noteworthy Mentions & Projects

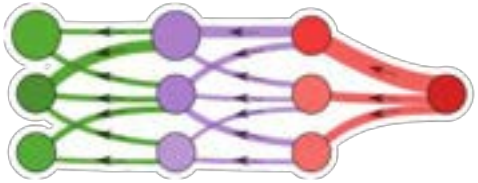
zennit

Canonization



Some of the presented works have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nos. **957059 (COPA EUROPE)** and **965221 (iToBoS).**

QUANTUS



ExplainableAI.jl

Refs: [Alber *et al.* 2019][Anders, Neumann, *et al.* 2021][Motzkus *et al.* 2022][Hedström *et al.* 2022][Hill 2022]

Ideas, Applications and Collaborations
Always Highly Welcome!

Get In Touch!

sebastian.lapuschkin@hhi.fraunhofer.de

For more details, awesome results and inspiration, read the full preprint of [Achtibat *et al.* 2022] at https://arxiv.org/abs/2206.03208

Want to try out CRP and RelMax with your PyTorch model?
Go to https://github.com/rachtibat/zennit-crp

References I

[1] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin, ``From" where" to" what": Towards human-understandable explanations through concept relevance propagation,'' *arXiv preprint arXiv:2206.03208*, 2022.

[2] C. J. Anders, L. Weber, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin, ``Finding and removing clever hans: Using explanation methods to debug and improve deep models,'' *Information Fusion*, vol. 77, pp. 261–295, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521001573.

[3] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, ``Unmasking clever hans predictors and assessing what machines really learn,'' *Nature Communications*, vol. 10, p. 1096, 2019.

[4] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, ``Explaining nonlinear classification decisions with deep taylor decomposition,'' *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

References II

[5] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, ``Evaluating the visualization of what a deep neural network has learned,'' *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[6] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek, ``Analyzing classifiers: Fisher vectors and deep neural networks,'' in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2912–2920.

[7] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, ``On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,'' *PLoS ONE*, vol. 10, no. 7, e0130140, 2015.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F.-F. Li, ``Imagenet large scale visual recognition challenge,'' *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[9] M. Everingham, L. Gool, C. Williams, J. Winn, and A. Zisserman, ``The pascal visual object classes challenge results,'' *"http://host.robots.ox.ac.uk/pascal/VOC/voc2007/workshop/everingham_cls.pdf"*, 2007.

References III

[10] P. Stock and M. Cissé, ``Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases,'' in *Proc. of European Conference on Computer Vision (ECCV)*, 2018, pp. 504–519.

[11] N. J. Morch, U. Kjems, L. K. Hansen, C. Svarer, I. Law, B. Lautrup, S. Strother, and K. Rehm, ``Visualization of neural networks using saliency maps,'' in *Proceedings of ICNN'95-International Conference on Neural Networks*, IEEE, vol. 4, 1995, pp. 2085–2090.

[12] M. D. Zeiler and R. Fergus, ``Visualizing and understanding convolutional networks,'' in *European conference on computer vision*, Springer, 2014, pp. 818–833.

[13] M. Sundararajan, A. Taly, and Q. Yan, ``Axiomatic attribution for deep networks,'' in *Proc. of International Conference on Machine Learning (ICML)*, 2017, pp. 3319–3328.

[14] S. M. Lundberg and S.-I. Lee, ``A unified approach to interpreting model predictions,'' in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765–4774.

References IV

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ``Grad-cam: Visual explanations from deep networks via gradient-based localization,'' in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[16] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, ``Smoothgrad: Removing noise by adding noise,'' *CoRR*, vol. abs/1706.03825, 2017.

[17] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, ``Visualizing deep neural network decisions: Prediction difference analysis,'' in *Proc. of International Conference on Learning Representations (ICLR)*, 2017.

[18] R. C. Fong and A. Vedaldi, ``Interpretable explanations of black boxes by meaningful perturbation,'' in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3449–3457.

[19] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin, ``Towards best practice in explaining neural network decisions with lrp,'' in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

References V

[20] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, ``Explaining deep neural networks and beyond: A review of methods and applications,'' *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[21] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek, ``The lrp toolbox for artificial neural networks,'' *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3938–3942, 2016.

[22] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, ``iNNvestigate Neural Networks!'' *Journal of Machine Learning Research*, vol. 20, 93:1–93:8, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-540.html.

[23] C. J. Anders, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin, ``Software for dataset-wide xai: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy,'' *CoRR*, vol. abs/2106.13200, 2021.

[24] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman, ``The devil is in the details: An evaluation of recent feature encoding methods.,'' in *BMVC*, vol. 2, 2011, p. 8.

References VI

[25] P. Hacker and J.-H. Passoth, ``Varieties of ai explanations under the law. from the gdpr to the aia, and beyond,'' in *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, Springer, 2022, pp. 343–373.

[26] S.-K. Yeom, P. Seegerer, S. Lapuschkin, A. Binder, S. Wiedemann, K.-R. Müller, and W. Samek, ``Pruning by explaining: A novel criterion for deep neural network pruning,'' *Pattern Recognition*, vol. 115, p. 107 899, 2021.

[27] D. Becking, M. Dreyer, W. Samek, K. Müller, and S. Lapuschkin, ``Ecq$x$: Explainability-driven quantization for low-bit and sparse dnns,'' *arXiv preprint arXiv:2109.04236*, 2021.

[28] S. Ede, S. Baghdadlian, L. Weber, A. Nguyen, D. Zanca, W. Samek, and S. Lapuschkin, ``Explain to not forget: Defending against catastrophic forgetting with xai,'' in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Springer, 2022, pp. 1–18.

References VII

[29] A. Rieckmann, P. Dworzynski, L. Arras, S. Lapuschkin, W. Samek, O. A. Arah, N. H. Rod, and C. T. Ekstrøm, ``Causes of outcome learning: A causal inference-inspired machine learning approach to disentangling common combinations of potential causes of a health outcome,'' *International Journal of Epidemiology*, May 2022, dyac078. [Online]. Available: https://doi.org/10.1093/ije/dyac078.

[30] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, ``Right for the right reasons: Training differentiable models by constraining their explanations,'' in *Proc. of Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 2662–2670.

[31] S. Teso and K. Kersting, ``Explanatory interactive machine learning,'' in *Proc. of the Conference on AI, Ethics and Society (AIES) 2019*, 2019, pp. 239–245.

[32] L. Rieger, C. Singh, W. J. Murdoch, and B. Yu, ``Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge,'' *CoRR*, vol. abs/1909.13584, 2019.

References VIII

[33] P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.-G. Luigs, A.-K. Mahlein, and K. Kersting, ``Making deep neural networks right for the right scientific reasons by interacting with their explanations,'' *Nature Machine Intelligence*, vol. 2, no. 8, pp. 476–486, 2020.

[34] F. Pahde, L. Weber, C. J. Anders, W. Samek, and S. Lapuschkin, ``Patclarc: Using pattern concept activation vectors for noise-robust model debugging,'' *arXiv preprint arXiv:2202.03482*, 2022.

[35] L. Weber, S. Lapuschkin, A. Binder, and W. Samek, ``Beyond explaining: Opportunities and challenges of xai-based model improvement,'' *arXiv preprint arXiv:2203.08008*, 2022.

[36] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, ``This looks like that: Deep learning for interpretable image recognition,'' *Advances in Neural Information Processing Systems*, vol. 32, pp. 8930–8941, 2019.

[37] C. Olah, A. Mordvintsev, and L. Schubert, ``Feature visualization,'' *Distill*, vol. 2, no. 11, e7, 2017.

References IX

[38] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, and R. Sayres, ``Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV),'' in *Proc. of International Conference on Machine Learning (ICML)*, 2018, pp. 2673–2682.

[39] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, ``Network dissection: Quantifying interpretability of deep visual representations,'' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6541–6549.

[40] F. Hohman, H. Park, C. Robinson, and D. H. P. Chau, ``SUMMIT: Scaling deep learning interpretability by visualizing activation and attribution summarizations,'' *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1096–1106, 2019.

[41] Z. Chen, Y. Bei, and C. Rudin, ``Concept whitening for interpretable image recognition,'' *Nature Machine Intelligence*, vol. 2, no. 12, pp. 772–782, 2020.

[42] A. Araujo, W. Norris, and J. Sim, ``Computing receptive fields of convolutional neural networks,'' *Distill*, vol. 4, no. 11, e21, 2019.

References X

[43] S. Bach, A. Binder, K.-R. Müller, and W. Samek, ``Controlling explanatory heatmap resolution and semantics via decomposition depth,'' in *2016 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2016, pp. 2271–2275.

[44] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, *et al.*, ``BCN20000: Dermoscopic lesions in the wild,'' *arXiv preprint arXiv:1908.02288*, 2019.

[45] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, *et al.*, ``Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC),'' in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2018, pp. 168–172.

[46] P. Tschandl, C. Rosendahl, and H. Kittler, ``The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,'' *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

References XI

[47] F. Motzkus, L. Weber, and S. Lapuschkin, ``Measurably stronger explanation reliability via model canonization,'' *arXiv preprint arXiv:2202.06621*, 2022, accepted for publication at ICIP 2022.

[48] A. Hedström, L. Weber, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M.-C. Höhne, ``Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations,'' *arXiv preprint arXiv:2202.06861*, 2022, accepted for publication in JMLR.

[49] A. Hill, *Explainableai.jl*, 2022. [Online]. Available: https://github.com/adrhill/ExplainableAI.jl.