Interpretability of Machine Learning Systems for Medical Imaging

Mara Graziani

Stéphane Marchand-Maillet

Henning Müller

UNIVERSITÉ DE GENÈVE
FACULTY OF SCIENCE
Department of Informatics

Hes·so // VALAIS WALLIS

PROCESS

European Commission
Horizon 2020
European Union funding for Research & Innovation

April 2nd, 2019

BACKGROUND

2013-2015 **B.En. in IT Engineering** at Sapienza

2015-2016 **1y** of **MSc in AI and Robotics** at Sapienza

2016-2017 **MPhil in Machine Learning, Speech and Language Technology** at Univ. of Cambridge

2017- now PhD in Computer Science at Univ. of Geneva and HES-SO Valais started in **November 2017**

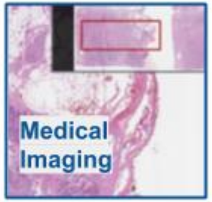
Theme: *Interpretability of Deep Learning for Medical Imaging*

FUNDING PROJECT

Providing Computing solutions for ExaScale challengeS

PROCESS

European Commission
Horizon 2020
European Union funding
for Research & Innovation






Goal: Train Deep Learning (DL) models on large scale Medical Imaging (MI) datasets

Main application: Breast Lymph-Node Histopathology (BLN)

Main dataset: Camelyon 2017 and 2016 challenges

DL Interpretability is one of the tasks

DEEP LEARNING FOR MI: SCENARIO

Data explosion

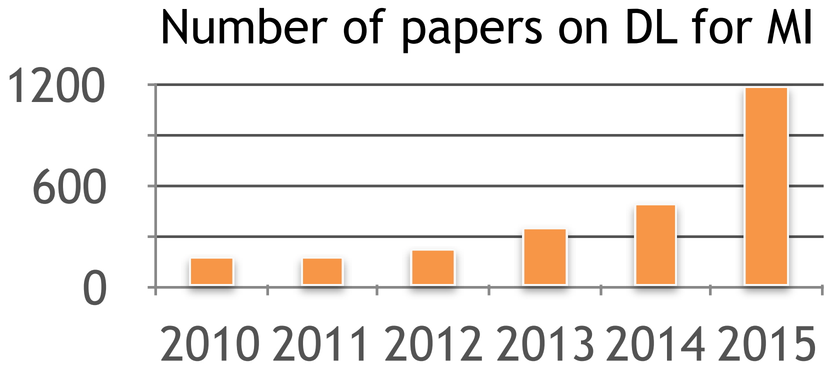
25 exabytes by 2020

[Jensen PB. et al.,2012]

2.5 Pb/y for mammography in U.S.

[Wittenburg et al., 2010]

Need for a model that scales



[Litjens et al., 2017]

Data explosion

30%
of worldwide

aphy in

ding the wave]

Need for a mode

120

60

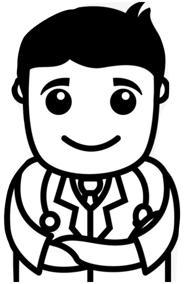
Still very challenging:

- 2D, 3D+, multimodal
- multi-scale
- acquisition variability
- subjectivity in diagnoses
- …

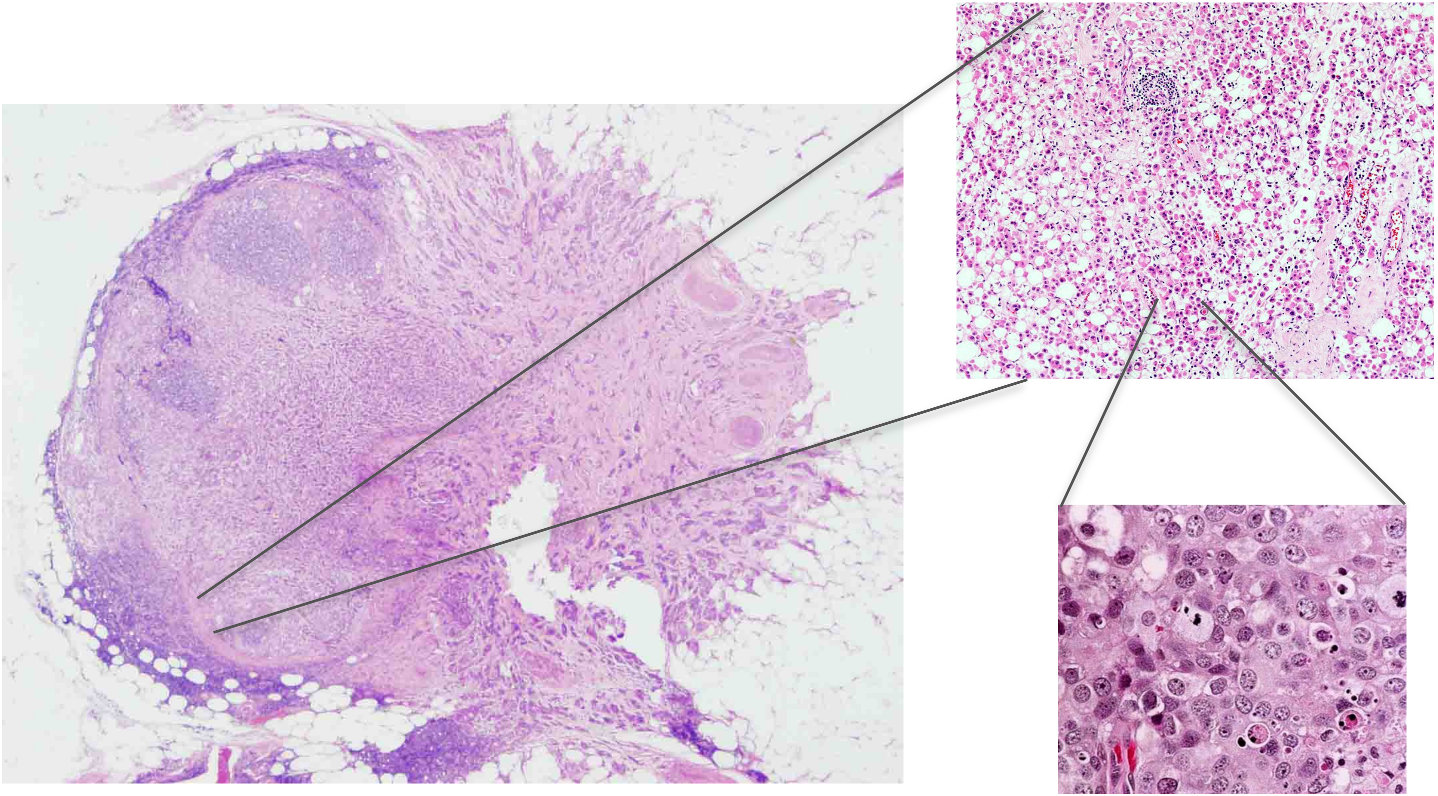
Often seen as a black box, especially by non-experts!

[Litjens et al., 2017]

INTERPRETABILITY IN MI

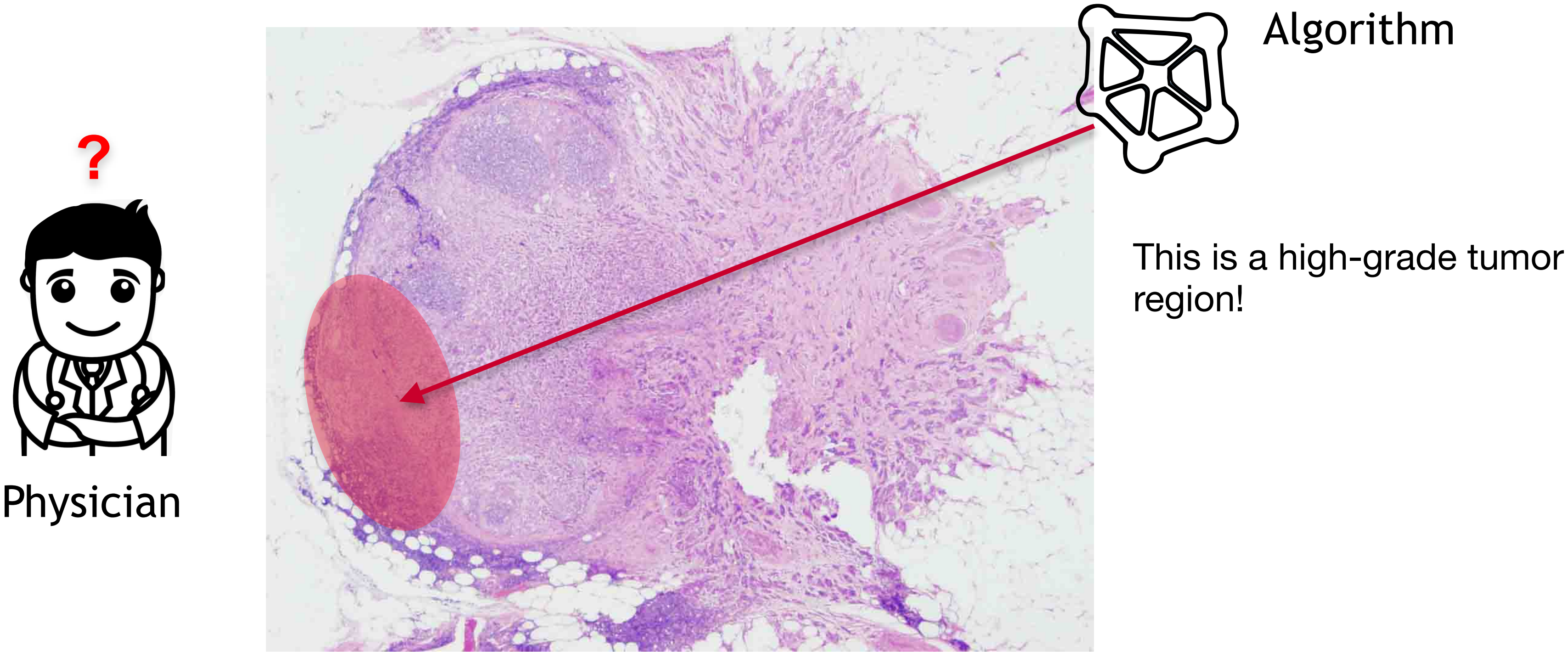


Physician



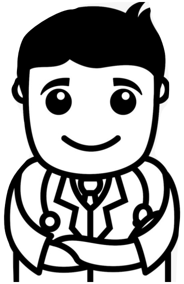
~ 200K x 100K pixels

INTERPRETABILITY IN MI

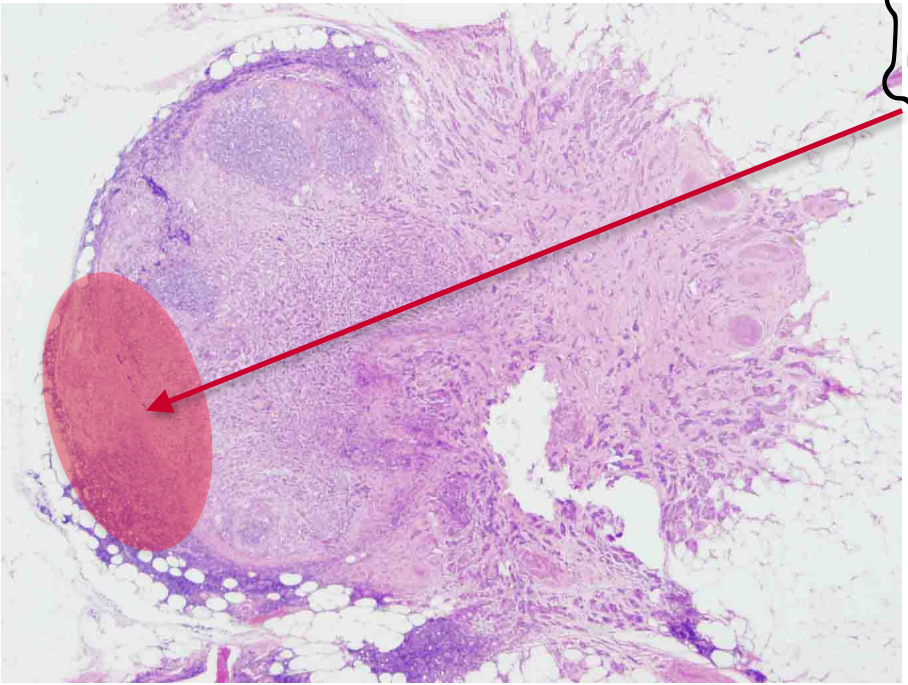


INTERPRETABILITY IN MI

OK!



Physician



Algorithm
With
Interpretability

This is a high-grade tumor region:

1. The cells are 10% larger than non-tutor average
2. The nuclei texture appears vesicular (contrast is 20% larger than average)

WHAT IS INTERPRETABILITY?

HYP. 1: Interpretability is defined as the ability to explain or to present in understandable terms to a human*.

[Doshi-Velez et al., 2017]

$\mathbf{E}_m$: Explanation in the model representation space (input pixels, activations)

$\mathbf{E}_h$: Explanation in the human representation space (high-level concepts)

$$g : \mathbf{E}_m \rightarrow \mathbf{E}_h$$

[Kim et al., 2018]

The interpretability task can be solved post-hoc by a distinct model.

[Lipton, 2018]

* not all humans are familiar with Machine Learning

INTERPRETER MODEL

$$g : \mathbf{E}_m \rightarrow \mathbf{E}_h$$

The task of the Interpreter model is linking the representation spaces in an "interpretable" way. This interpretability task is solved on the representations learned by the network that solves the primary task (ex. classification of tumor regions) without the need of retraining.

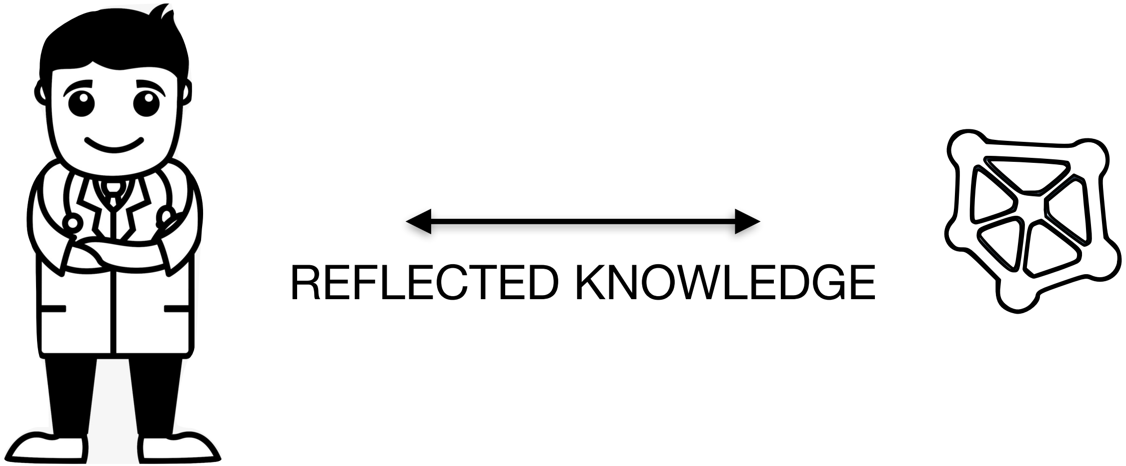
ASS. 1: If the interpreter is a non-complex model, as for ex. a linear model, we define g as linear interpretability.

[Kim et al., 2018]

* not all humans are familiar with Machine Learning

USER-CENTRIC INTERPRETABILITY FOR MI

From the medical imaging viewpoint, deep learning interpretability is applied to explain the decisions of a complex model in terms understandable by doctors. This eases the interaction and improves the quality of the diagnosis.

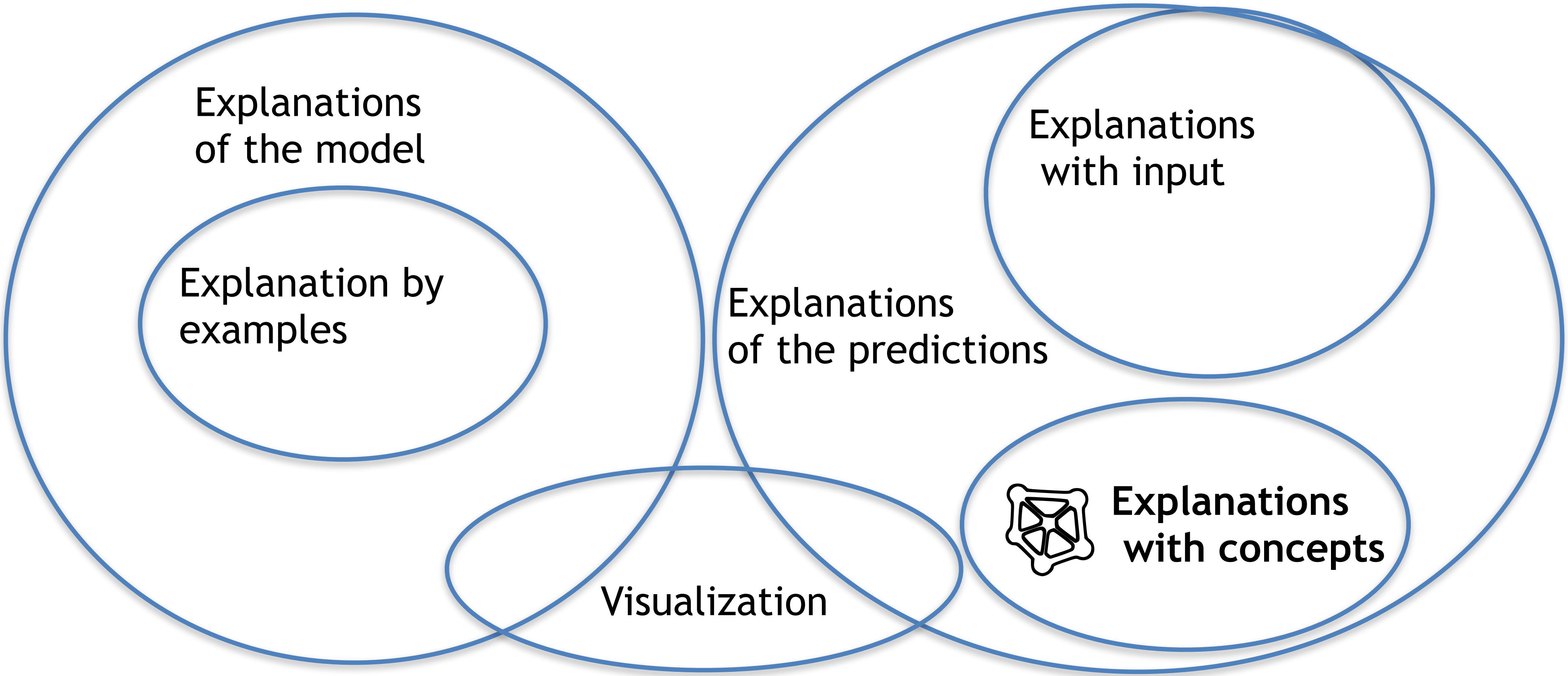


RESEARCH QUESTION

Can **domain-related concepts** (for example clinical measures) be learned post-hoc in the latent space and used to produce user-centric explanations of deep learning decisions?

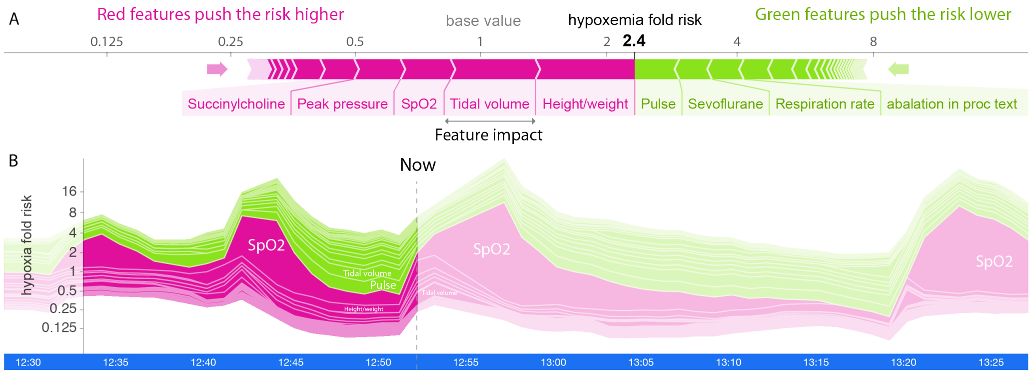
STATE OF ART - Post-hoc Interpretability for healthcare

Post-hoc explanations for DL models with medical applications



STATE OF ART - SHAP

Explanations with input features: Shapley Additive exPlanations (SHAP)*



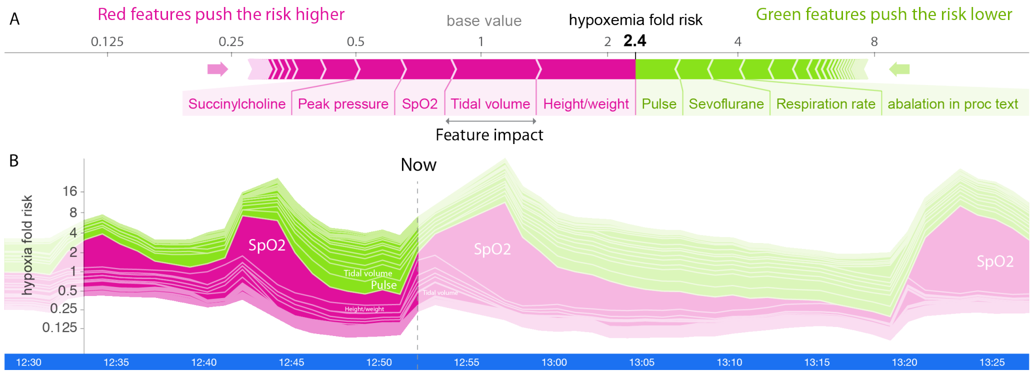
[Lundberg et al., 2017]

Attributes to each input feature the change in the expected model prediction when conditioning on that feature.

*Model agnostic, unifies six methods: LIME, deepLIFT, LRP, Shapley regression, Shapley sampling, quantitative input influence.

STATE OF ART - SHAP

Explanations with input features: Shapley Additive exPlanations (SHAP)

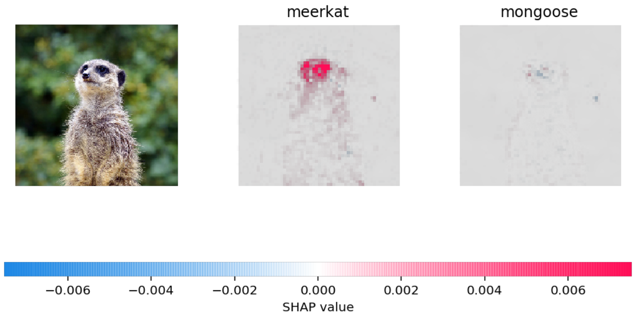


Attributes to each input feature the change in the expected model prediction when conditioning on that feature.

[Lundberg et al., 2017]

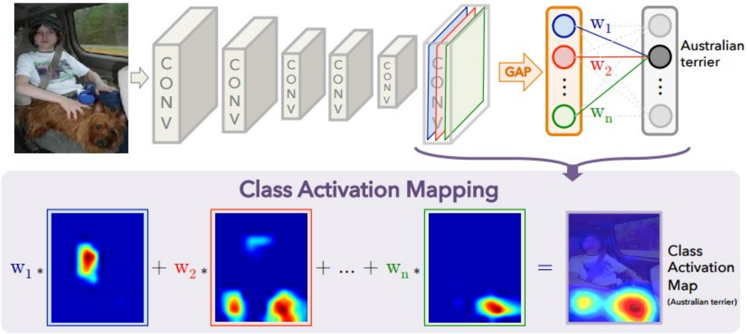
✔ Very good for non-visual inputs

✘ Difficult abstraction on images
Only qualitative evaluation on images

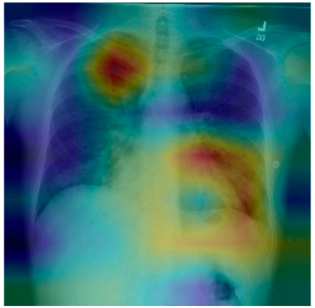


*Model agnostic, unifies six methods: LIME, deepLIFT, LRP, Shapley regression, Shapley sampling, quantitative input influence.

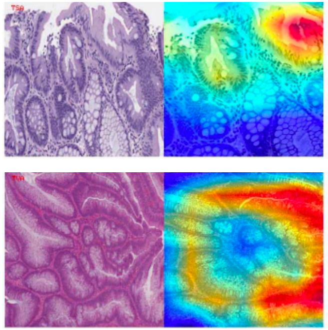
Explanations with input features: Class Activation Maps (CAM)



[Zhou et al., 2016]



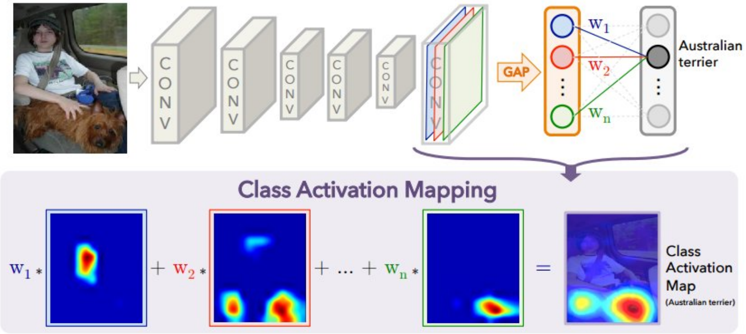
[Rajpurkar et al., 2017]



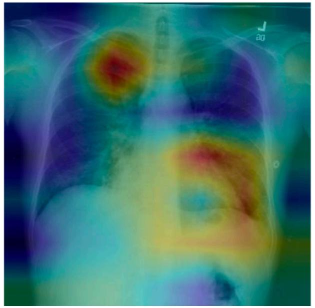
[Korbar et al., 2017]

STATE OF ART - CAM, gradCAM, guided CAM

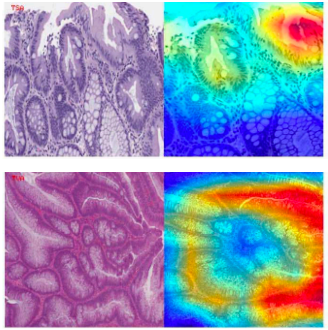
Explanations with input features: Class Activation Maps (CAM)



[Zhou et al., 2016]



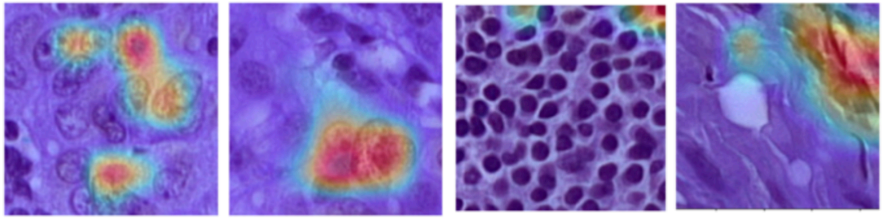
[Rajpurkar et al., 2017]



[Korbar et al., 2017]

- ✓ Direct visualization on the input image
- ✗ Not sharp
  Only qualitative evaluation
  Individual instances (local)

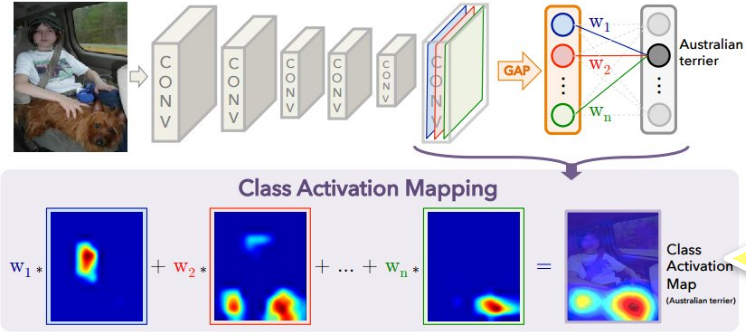
Experiments in the lab

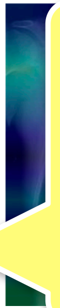


(a) tumor, $p = 0.994$ (b) tumor, $p = 0.999$ (c) non-tumor, $p = 4.7e-4$ (d) non-tumor, $p = 0.841$

STATE OF ART - CAM, gradCAM, guided CAM

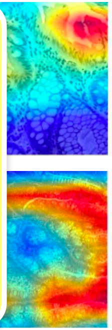
Explanations with input features: Class Activation Maps (CAM)





(Un)reliability of saliency methods
[Kindermans et al., 2017]

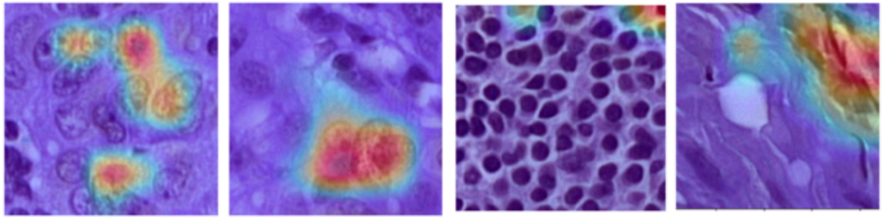
Sanity Checks for Saliency Maps
[Adebayo et al., 2018]



[Zhou et al., 2016] [Rajpurkar et al., 2017] [Korbar et al., 2017]

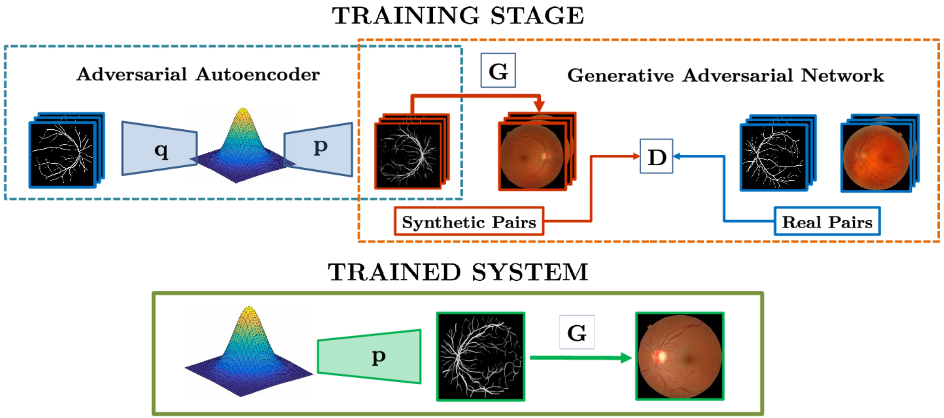
- ✓ Direct visualization on the input image
- ✗ Not sharp
  Only qualitative evaluation
  Individual instances (local)

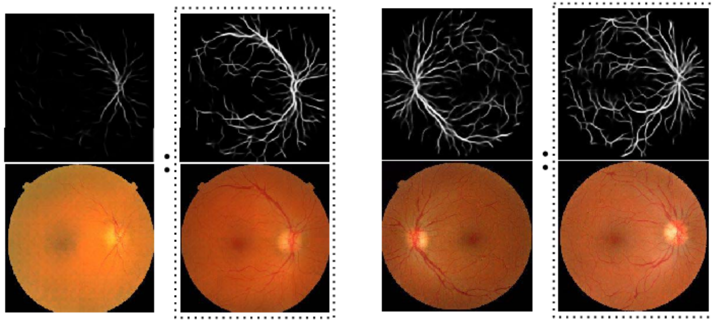
→ Experiments in the lab



(a) tumor, $p = 0.994$ (b) tumor, $p = 0.999$ (c) non-tumor, $p = 4.7e{-}4$ (d) non-tumor, $p = 0.841$

Explanations with examples: Generative Adversarial Networks

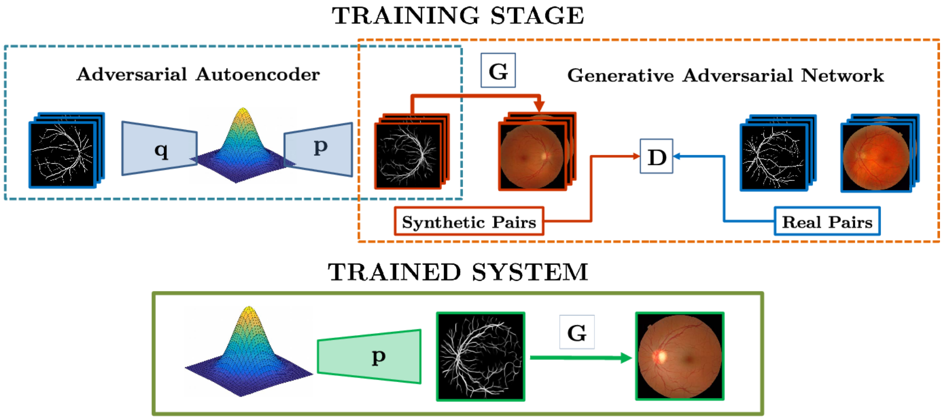


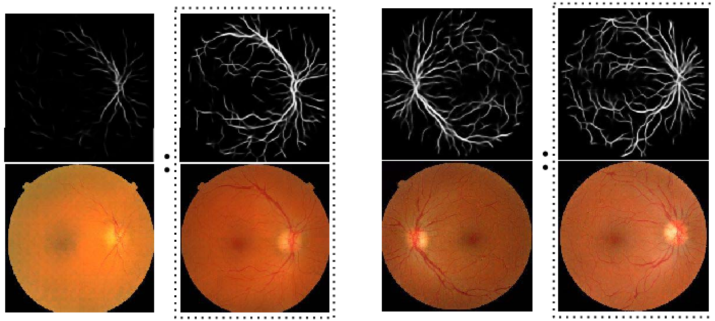


[Costa et al., 2018]

STATE OF ART - Generative nets, Activation Maximization

Explanations with examples: Generative Adversarial Networks

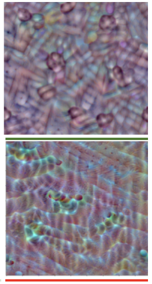




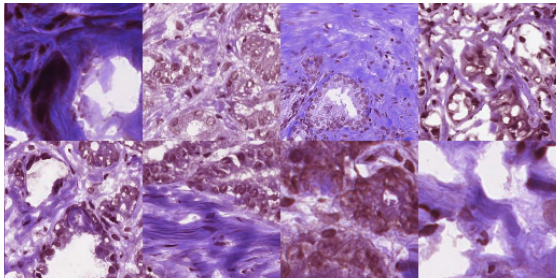
[Costa et al., 2018]

Experiments in the lab:

AM



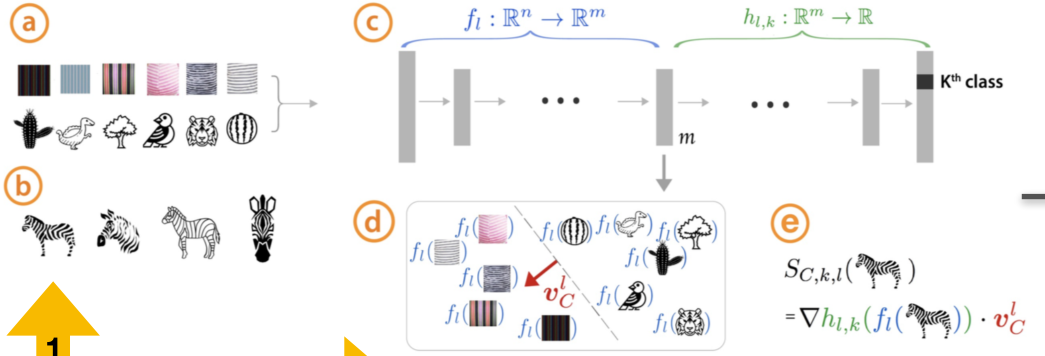
GANs



Reasoning by examples

Needs guidance on major structures
Difficult abstraction
Qualitative evaluation
Difficult to learn low represented pathology identified regions

Explanations with high-level concepts

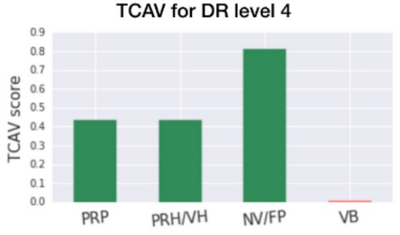


1

2 **Classification** in the activation space

Relevance scores

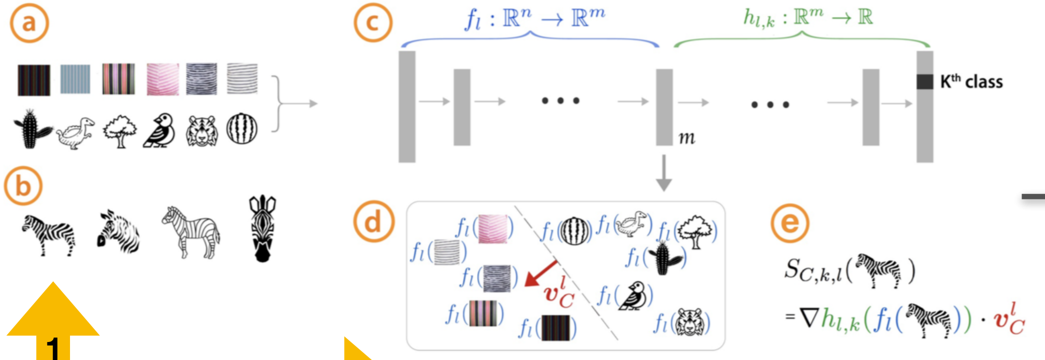
3



[Kim et al, 2018]

STATE OF ART - Testing with Concept Activation Vectors

Explanations with high-level concepts

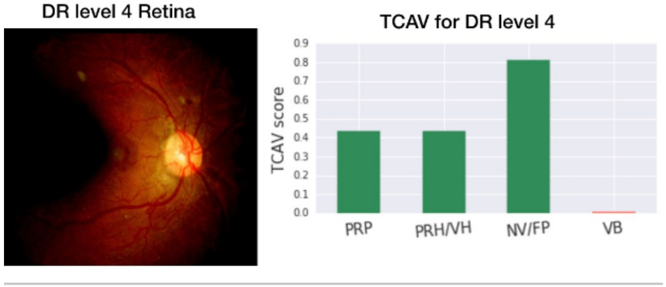


1

2 **Classification** in the activation space

Relevance scores

3



[Kim et al, 2018]

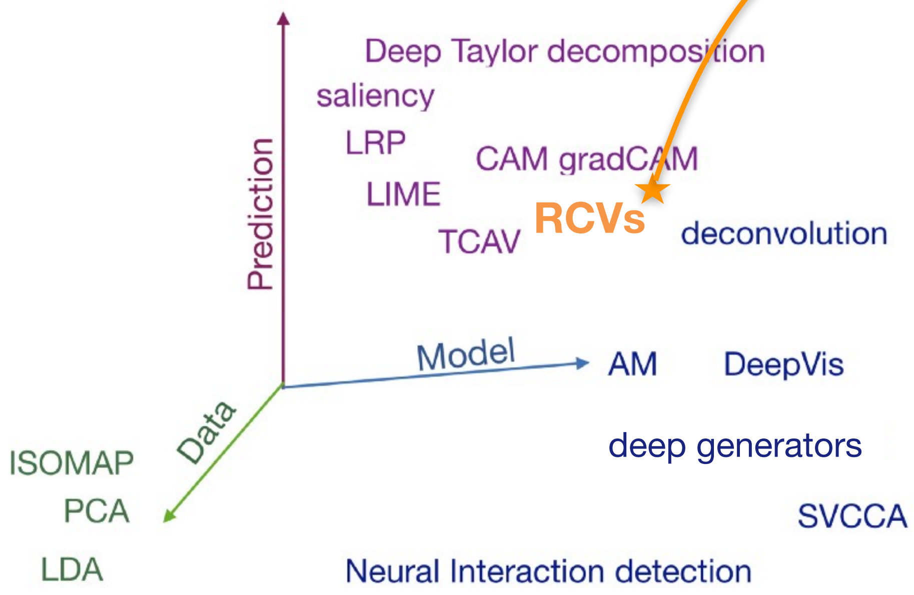
High abstraction
Quantitative evaluation

No support for continuous measures

Used as building block of my previous research papers!
[Graziani et al., 2018]
[Graziani et al., 2019]

STATE OF ART: THERE IS MUCH MORE!

Three dimensions of *Interpretability**, but more than 20K papers in the last two years



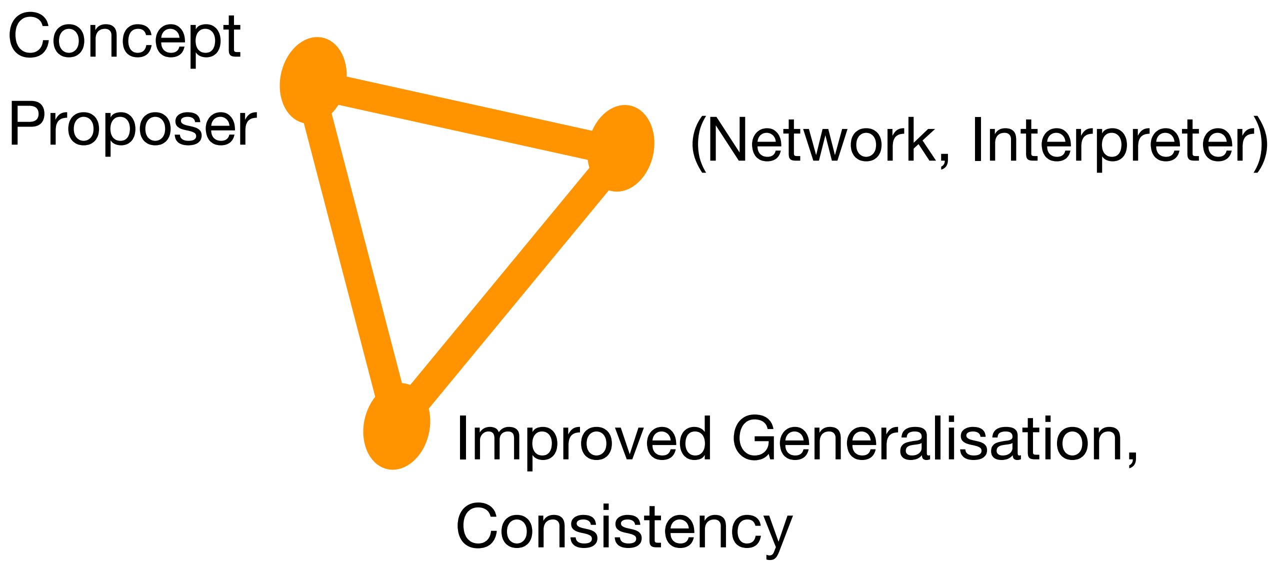
* *defined in [Montavon et. al., 2017]*

Previous work:

- ✓ "Regression Concept Vectors for Bidirectional Explanations for Histopathology", best paper award iMIMIC at MICCAI
  [Graziani et al., 2018]
- ✓ "Improved Interpretability for Computer-Aided assessment of Retinopathy of Prematurity", SPIE Medical Imaging
  [Graziani et al., 2019]
- ✓ "Concept Attribution with Regression Concept Vectors", to submit at IEEE TMM Special Issue on Multimedia Computing with Interpretable Machine Learning

PHD CONTRIBUTIONS (PROPOSAL)

Can **domain-related concepts** (for example clinical measures) be learned by a distinct model in the latent space and used to produce user-centric explanations of deep learning decisions?



CONCLUSION

WHEN: November 2017 - November 2021

WHAT: User-centric Interpretability of Deep Learning for Medical Imaging with domain-related concepts (ex. clinical measures)

HOW: Concept proposal, (Network, Interpreter), Model Improvements

WHY: This work could contribute in

- Identifying concepts and their relevance at the multi-scale level
- Reduce the impact of acquisition-dependent concepts (e.g. staining)
- Introduce objectivity and improve the interaction with Computer Aided Diagnostic systems

QUESTIONS?

REFERENCES

- Jensen, P. B., Jensen, L. J., & Brunak, S. “Mining electronic health records: towards better research applications and clinical care”, *Nature Reviews Genetics* 2012.
- Wittenburg, P., Van de Sompel, H., Vigen, J., Bachem, A., Romary, L., Marinucci, M., Lopez, D. R. "Riding the wave: How Europe can gain from the rising tide of scientific data." (2010).
- Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., and Sánchez, C. I. "A survey on deep learning in medical image analysis." *Medical image analysis* 42 (2017): 60-88.
- Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European conference on computer vision*. Springer, Cham, 2014.
- Raghu, Maithra, Zhang, C., Kleinberg, J., and Bengio, Sammy. "Transfusion: Understanding Transfer Learning with Applications to Medical Imaging”, *arXiv:1902.07208*, 2019
- Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in Neural Information Processing Systems*. 2017.
- Doshi-Velez, Finale, Kim, Been, "A Roadmap for a Rigorous Science of Interpretability.”, *CoRR abs/1702.08608*, 2017
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018, July). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning* (pp. 2673-2682).
- Lipton, Zachary C. "The Mythos of Model Interpretability." *Queue* 16.3 (2018): 30.

REFERENCES

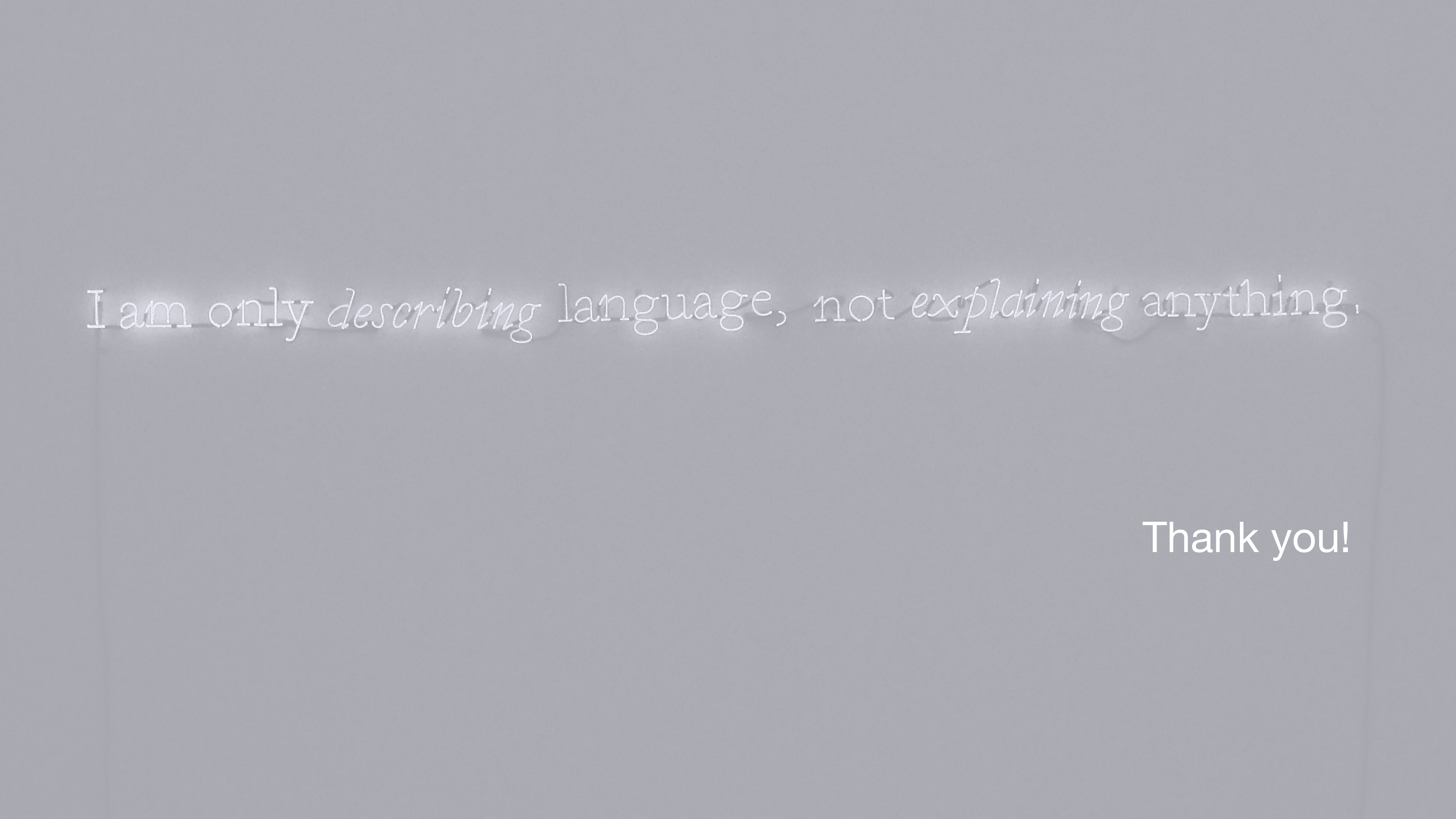
- Bolei Zhou*, David Bau*, Aude Oliva, and Antonio Torralba. “Interpreting Deep Visual Representations via Network Dissection.”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, June 2018
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K. and Lungren, M.P., CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning.
- Korbar, B., Olofson, A. M., Miraflor, A. P., Nicka, C. M., Suriawinata, M. A., Torresani, L.,and Hassanpour, S. "Looking under the hood: Deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017.
- Kindermans, P. J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S, Erhan D., and Kim, B. “The (un) reliability of saliency methods.” *arXiv preprint arXiv:1711.00867,* 2017.
- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems* (pp. 9525-9536). Costa, P., Galdran, A., Meyer, M. I., Niemeijer, M., Abràmoff, M., Mendonça, A. M., & Campilho, A. (2018). End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, *37*(3), 781-791.
- Zhu, H., Paschalidis, I. C., & Tahmasebi, A. (2018). “Clinical Concept Extraction with Contextual Word Embedding”. NeurIPS 2018

REFERENCES

- Cheng, M. Y., & Wu, H. T. (2013). "Local linear regression on manifolds and its geometric interpretation". *Journal of the American Statistical Association*, *108*(504), 1421-1434.
- Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." *The Journal of Machine Learning Research* 17.1 (2016): 2096-2030.

Own work:

- ✓ Graziani, Mara, Vincent Andrearczyk, and Henning Müller. "Regression Concept Vectors for Bidirectional Explanations in Histopathology." *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer, Cham, 2018. 124-132.
- ✓ Graziani, M., Brown, J. M., Andrearczyk, V., Yildiz, V., Campbell, J. P., Erdogmus, D.,Stratis, Y., and Müller, H. (2019, March). Improved interpretability for computer-aided severity assessment of retinopathy of prematurity. In *Medical Imaging 2019: Computer-Aided Diagnosis* (Vol. 10950, p. 109501R). International Society for Optics and Photonics.
- ✓ Graziani, M., Andrearczyk, V., Marchand-Maillet, S., and Müller, H. "Concept Attribution with Regression Concept Vectors", to submit at IEEE TMM Special Issue on Multimedia Computing with Interpretable Machine Learning



Thank you!