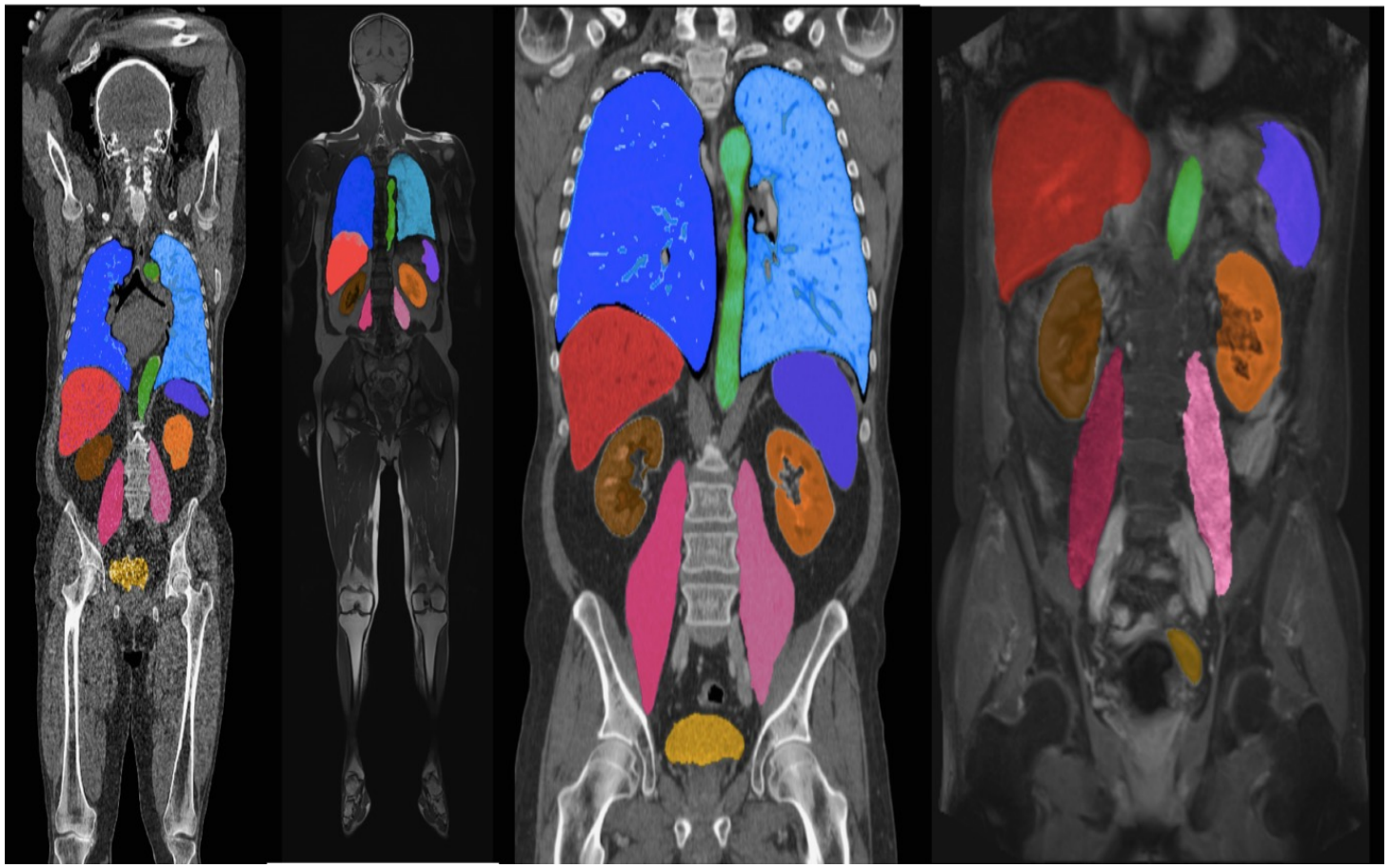
Hes·so VALAIS WALLIS
School of Management & Tourism

Explainable AI
(with a focus on medical AI)



AI4media

Mara Graziani
HES-SO & IBM research
Henning Müller
HES-SO & UNIGE

11.7.2022

Outline

- Who we are, where we are
- Our research on medical data analysis
  - Clinical decision support
- Several ways to say (almost) the same thing
  - Interpretability, explainability, transparency, …
  - Ways to classify interpretability/explainability
- Explaining classical machine learning
- Explaining deep learning
- Discussions and conclusions

Mara Graziani

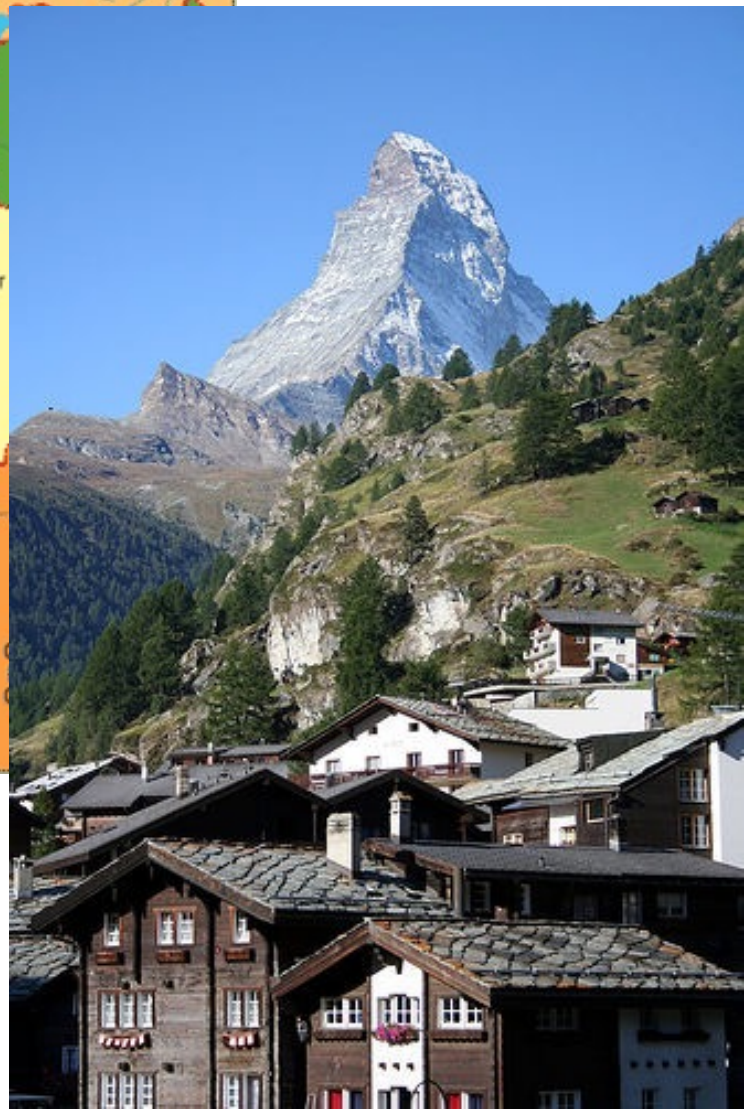
- Bachelor of IT engineering of La Sapienza, Rome, Italy (2015)
- Master in Machine Learning from the University of Cambridge, UK (2017)
- PhD on Interpretability of deep learning at the University of Geneva, Switzerland (2017-2021)
- Work at IBM research in Zürich (since 2021)
- Work on Interpretability received the best thesis award from the IEEE Technical Committee on Computational Life Sciences (2022)

Henning Müller

- Medical informatics studies in Heidelberg, Germany (1992-1997)
  - Exchange with Daimler Benz research, USA
- PhD in image processing, image retrieval, Geneva, Switzerland (1998-2002)
  - Exchange with Monash University, Melbourne, AUS
- Professor in radiology and medical informatics at the University of Geneva (2014-)
- Professor in Computer Science at the HES-SO, Sierre, Switzerland (2007-)
- Visiting faculty at Martinos Center (2015-2016)
- Member of the Swiss National Research Council

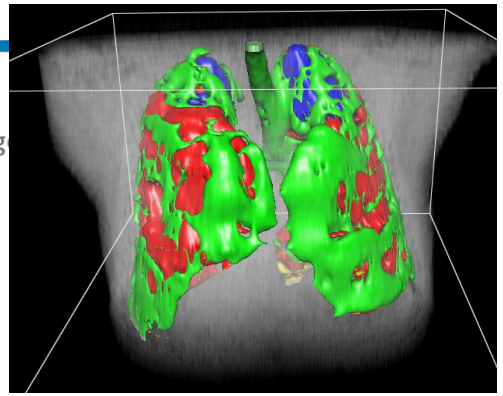
Where we are


Switzerland
GERMANY
FRANCE
SWITZERLAND
ITALY
LIECH.
Basell
Baden
Winterthur
Zürich
St.Gallen
Horgen
Biel
Lucerne
(Luzern)
Neuchâtel
Burghdorf
Bern
Steffisburg
Fribourg
Thun
Interlaken
Chur
Lausanne
Montreux
Sierre
Geneva
(Genève)
Martigny
Champex
Locarno
©maps.com



Team work

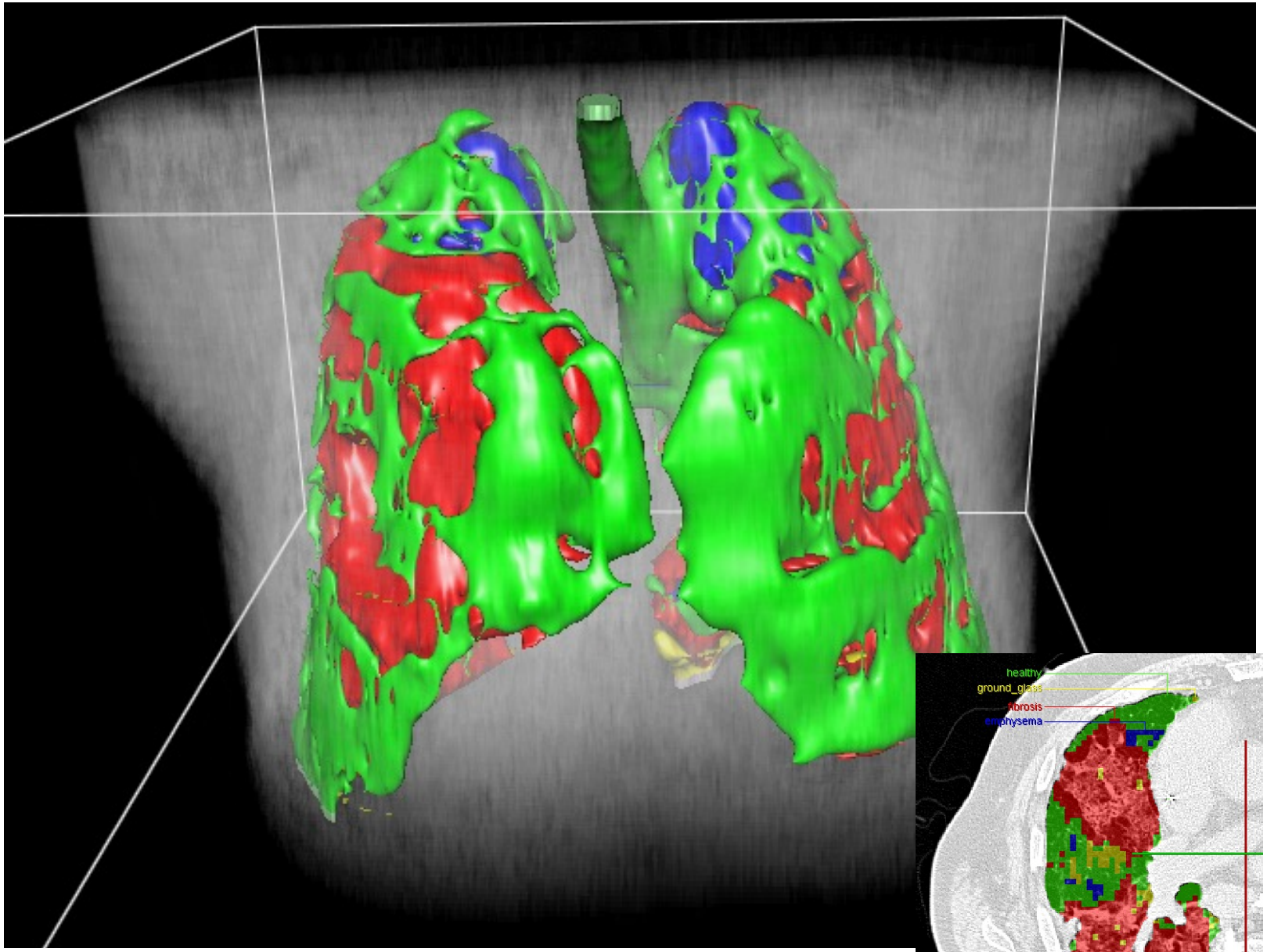
Our research directions

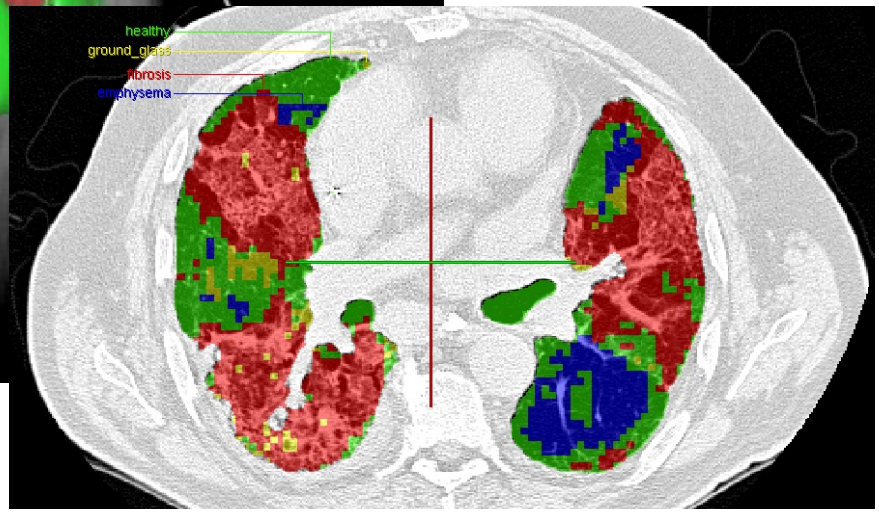


- Multimodal medical data analysis
  - Text analysis and retrieval in several languages
  - Combined text and image retrieval/analysis
  - Signal and image analysis for prosthesis control
  - Scientific challenges and evaluation methods
    - The ImageCLEF benchmark
- Clinical decision support systems
  - Supply clinicians with quantitative information to aid decisions (not to replace them)
    - Radiomics mainly in oncology (treatment, prognosis)
    - Histopathology to guide physicians to relevant areas
    - Explainability and Interpretability

3D ROIs in lung CT

Hes·so VALAIS WALLIS
School of Management & Tourism






DOES IT HURT
WHEN I PRESS HERE ?
© Denis Pessin
PESSIN

Motivations for the course

- Digital medicine is a reality
- ML is everywhere in medicine
  - Particularly in discussions and future potential
  - Broad, real-world use is relatively rare
- Results of ML need to be integrated with other knowledge, particularly in the medical field
  - The physician is responsible for the final decisions!
  - Outcomes that can not be explained/understood can not be integrated well
  - Explainability can help to judge potential bias
    - In connection with uncertainty, causality, …

Objectives of the course

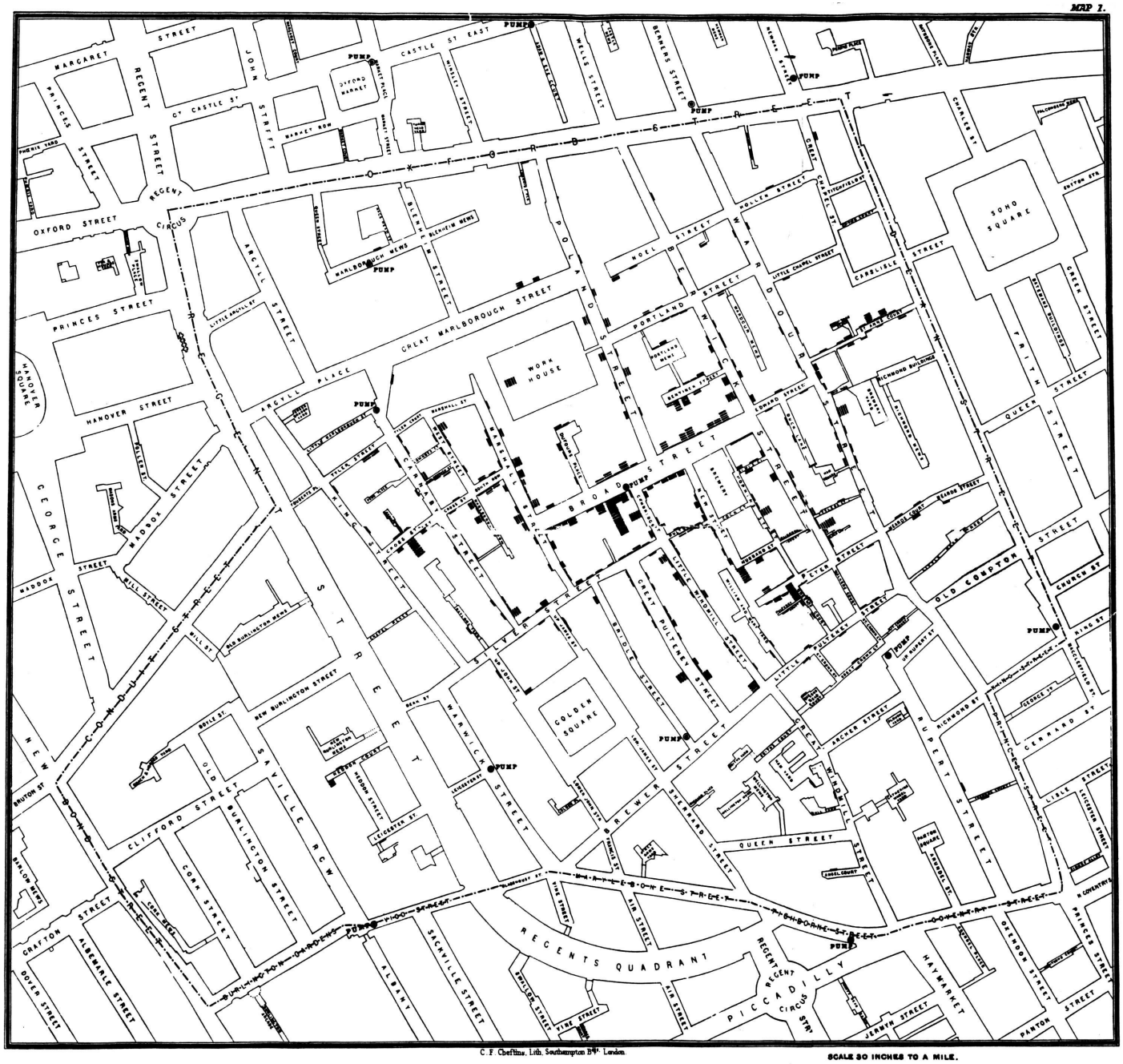
- Get an understanding of what explainability and interpretability are
  - Plus many related terms and their definitions
- Understand types of interpretability
  - Global, local, post-hoc, …
- Understand the difference of classical ML and DL in terms of explaining decisions
- Understand some of the techniques developed
  - Including what they are good for and what the differences are

Hes·so VALAIS WALLIS
School of Management & Tourism

Medical data analysis

Systematic medical data analysis

- Broad Stree
- Prevailing o
  transmitted
- Physician J
  water or oth
  - Water can
    filtered at
  - He noted



Google Flu Trends

DAVID LAZER AND RYAN KENNEDY OPINION 10.01.15 07:00 AM

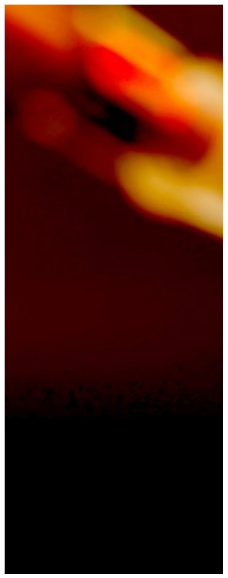
WHAT WE
THE EPIC
FLU T

POLICY FOR

The P

David Lazer[1]

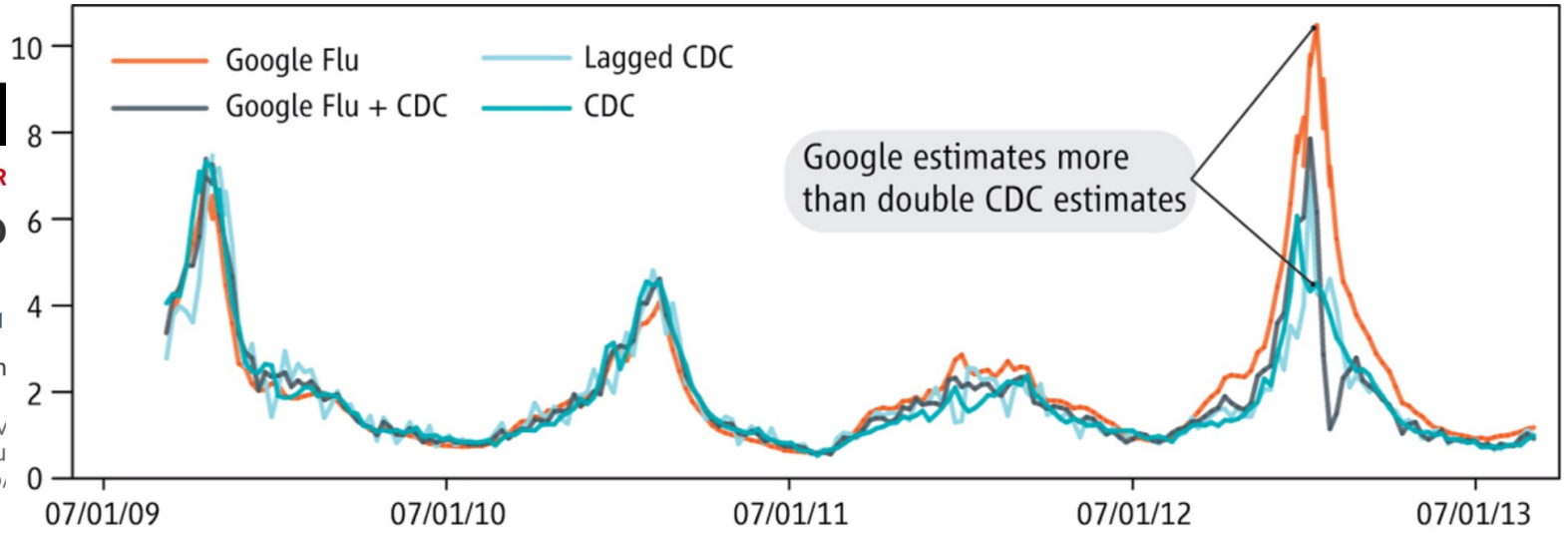
+ See all auth

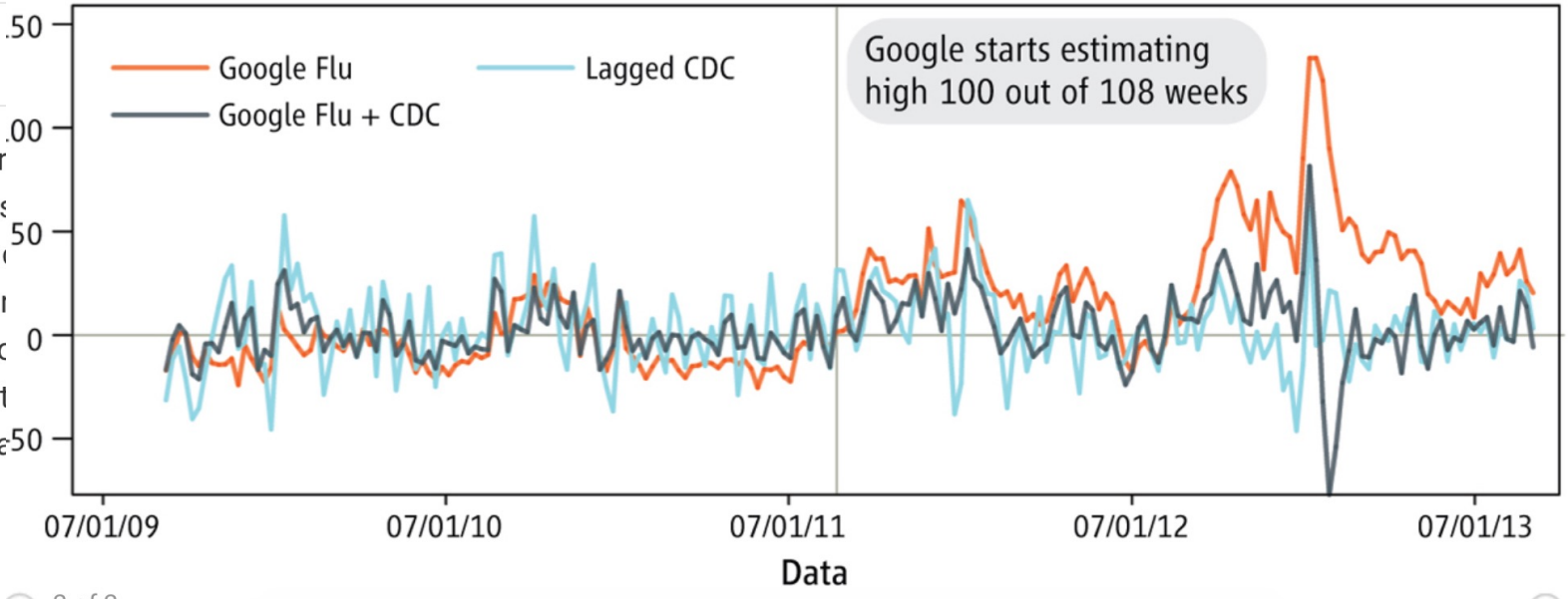


Science 14 M
Vol. 343, Issu
DOI: 10.1126/

Article

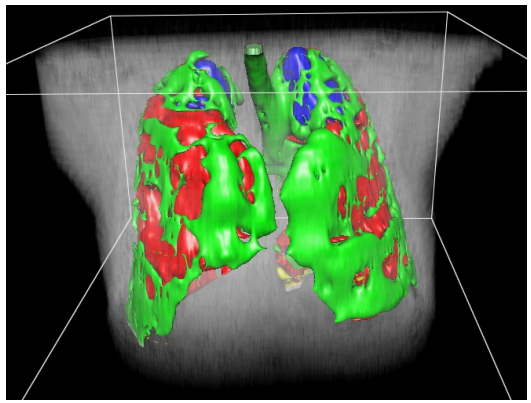
In Februar
executives
was predi
the Center
reports fr
was built t
(*3*, *4*), wha



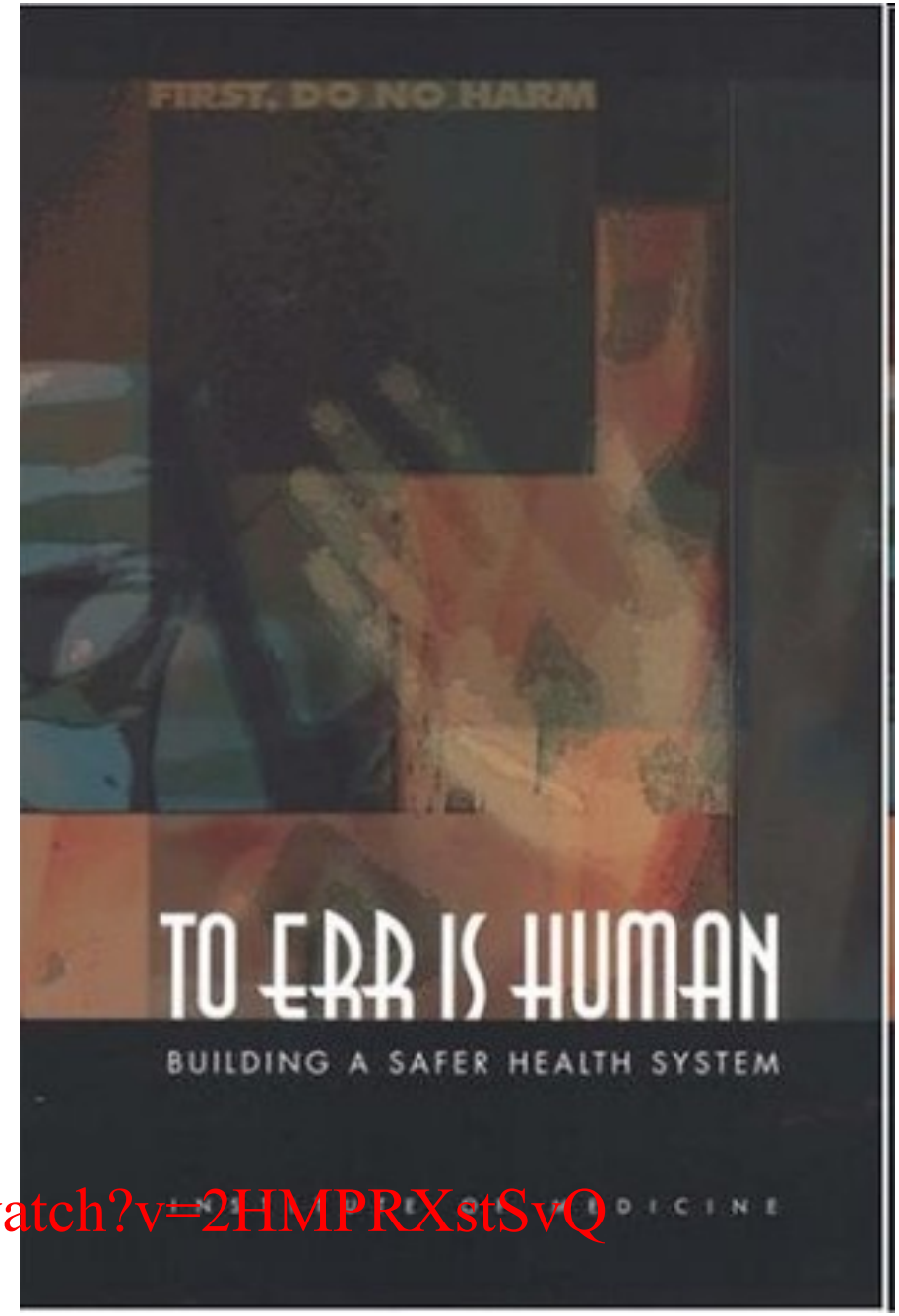


Imaging applications: CADx, CADe

- Computer-Aided Diagnosis (CADx)
- Computer-Aided Detection (CADe)
  - Finding locations of lesions
- Computer-Aided Decision Support
- Many tools are in this area
  - Finding similar patients (retrieval)
  - Finding criteria for or against specific diseases (rules)
  - Prediction of findings such as tissue types
  - Predicting a probability for a diagnosis using machine learning



Why decision support?



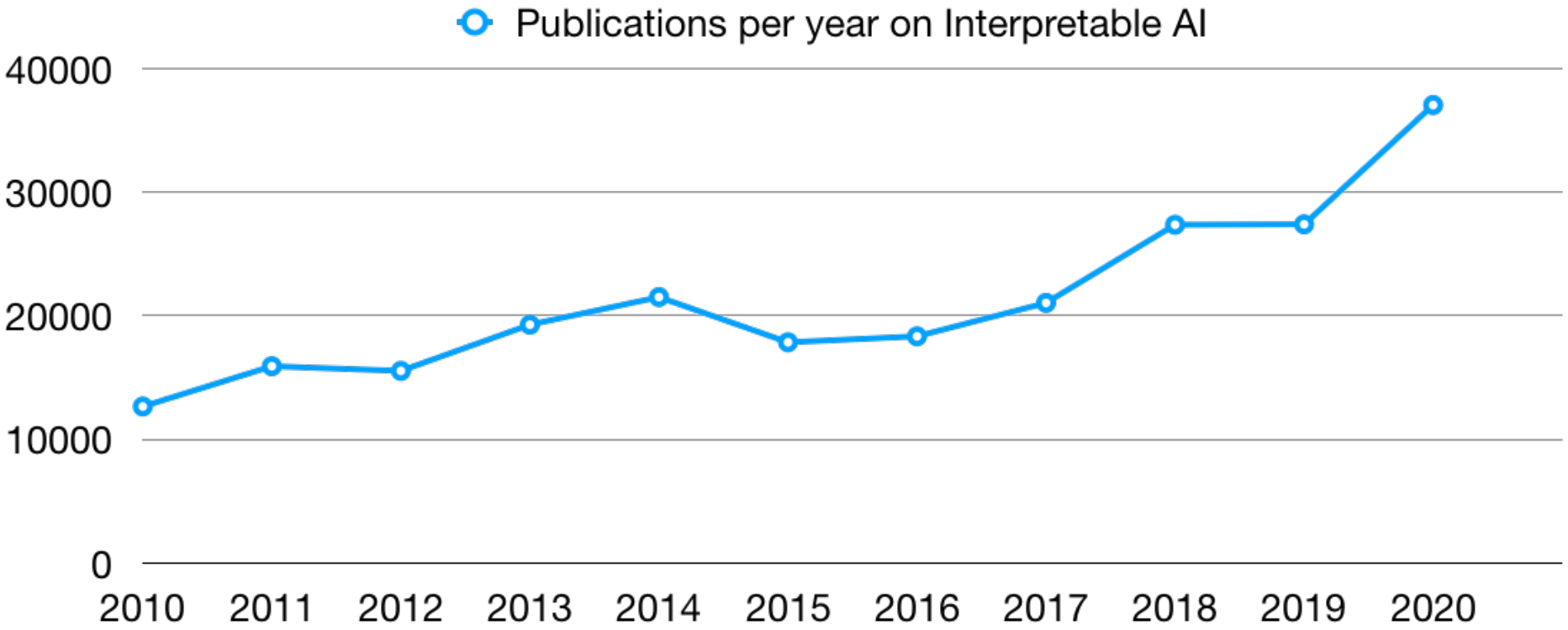
Geoff Hinton on radiology
https://www.youtube.com/watch?v=2HMPRXstSvQ

Terms and definitions

"Interpretability is the ability to explain or to present in understandable terms to a human*"

[Kim et al., 2018]

Trends of XAI



Source: app.dimensions.ai
Accessed on April 12, 2021
Criteria: “Interpretable AI OR explainable AI OR interpretability OR explainability OR XAI” in full data

Definitions

- Interpretability
  - Interpretability is the degree to which a human can understand the cause of a decision (Miller)
  - Interpretability is the degree to which a human can consistently predict the model's result (Kim)
- Explainability (wikipedia)
  - Explainable AI (XAI) is artificial intelligence (AI) in which the results of the solution can be understood by humans. It contrasts with the concept of the "black box". XAI may be an implementation of the social right to explanation.

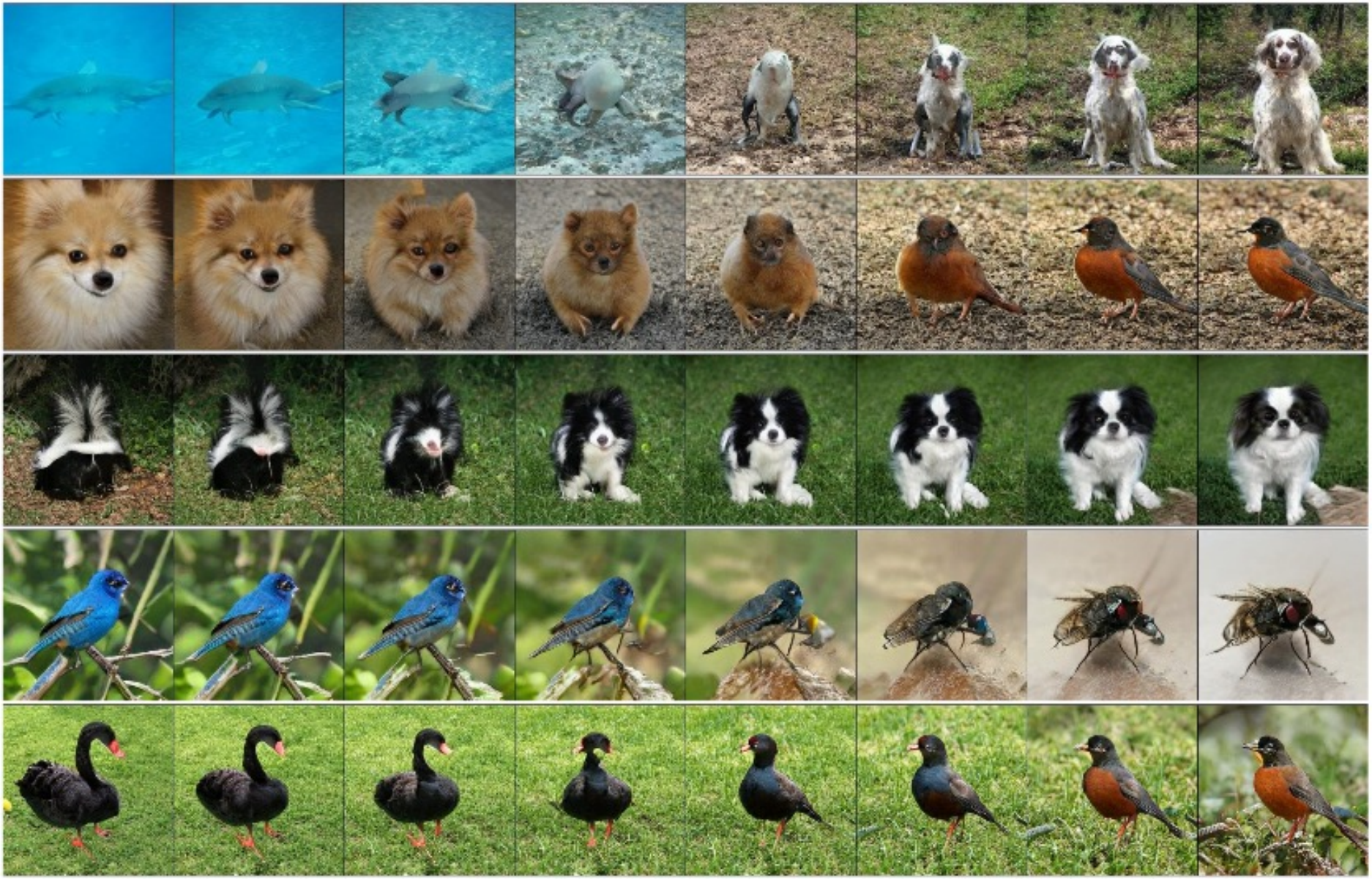
Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017).

Been Kim, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016).

What changed in machine learning?

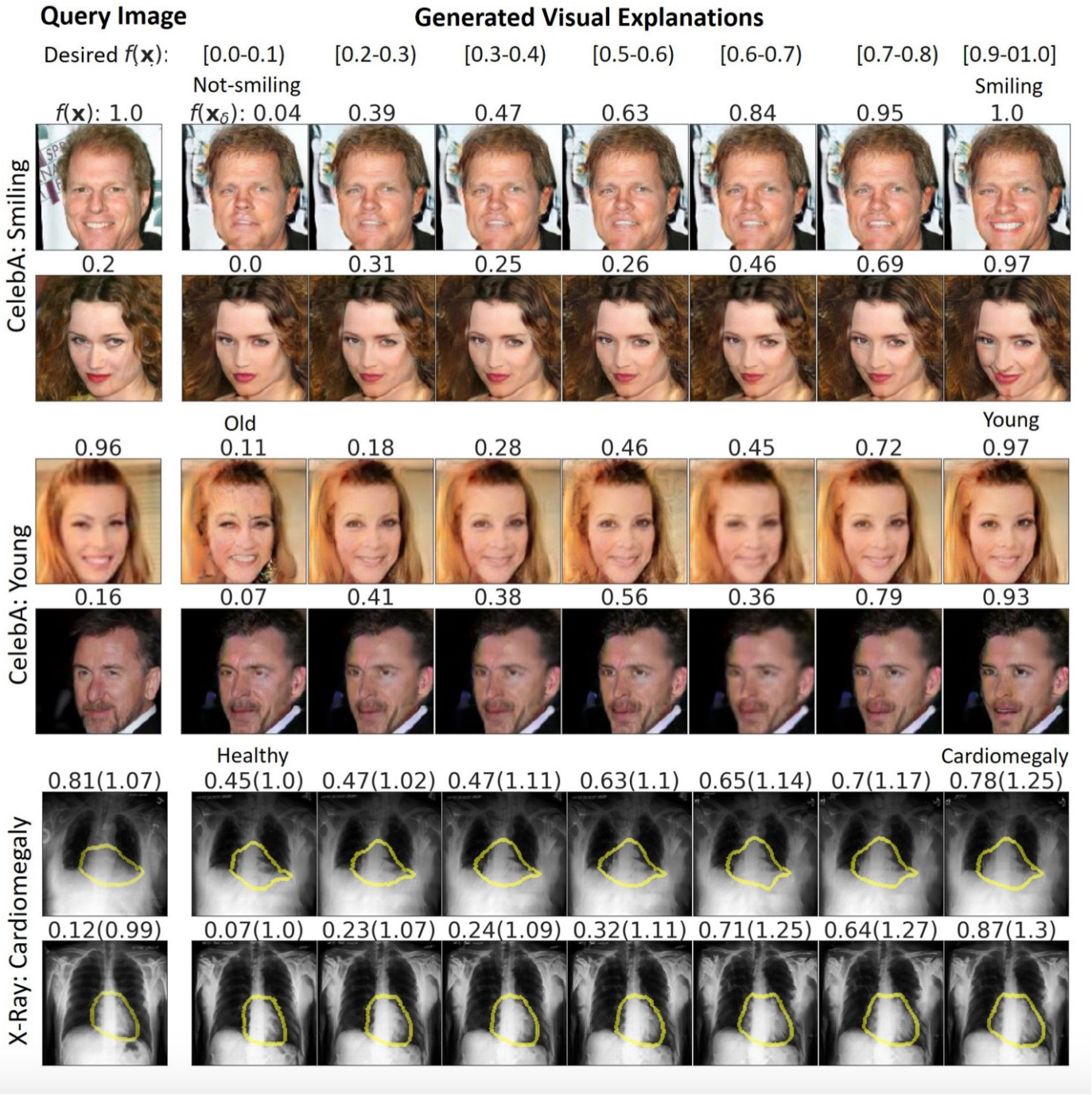
- Classical machine learning
  - Features extracted (handcrafted) model human knowledge and can often be understood
  - Distance measures/kNN/SVMs all have a clear reasoning and can be understood
  - Decision trees are very easy to understand
- Deep neural networks
  - Inner working is a black box
    - Millions of parameters!!
    - Non-deterministic outcomes
  - Best results in most scientific challenges
  - How can we make sure that DL can be understood?

A new situation with generative models



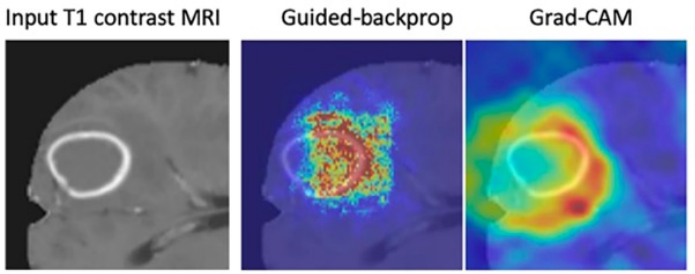
PyTorch BigGAN

Exaggerate features



Singla, Sumedha, et al. "Explanation by progressive exaggeration." *arXiv preprint arXiv:1911.00483* (2019).

Explainability of Deep Learning



- Make decisions understandable and remove the black box image
- Make sure that decisions are sound (data bias)
  - Explain why things may not be working
- In medicine it is particularly important to make sure that results can be explained & reproduced
  - High impact of wrong decisions
- There are many approaches on explainability
- Examples:
  - 2D projections, PCA, TSNE
  - Class activation maps, saliency, …

Many definitions

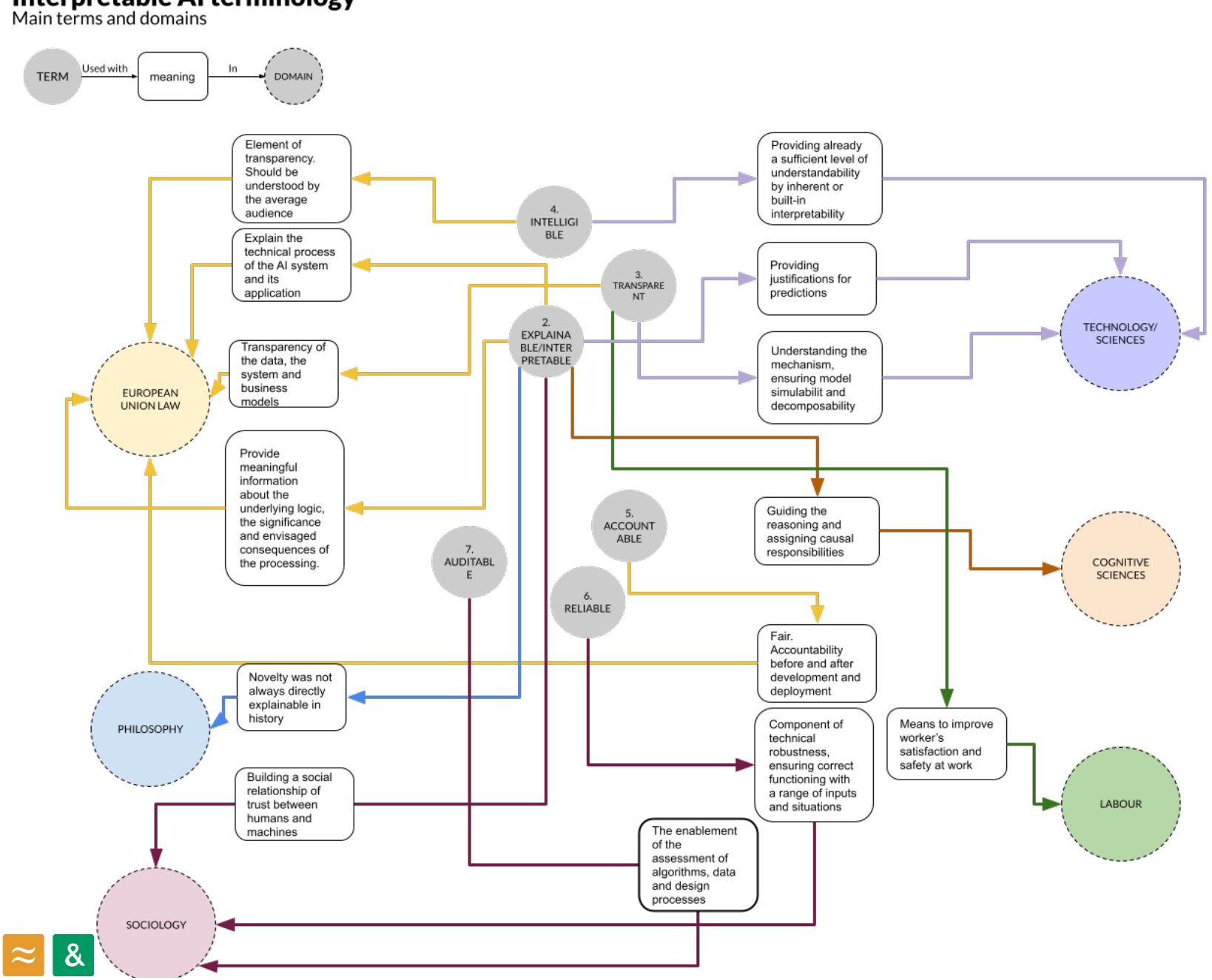
Interpretable	Explainable	Transparent	Intelligible=Understandable	Causality	
The degree to which an observer can understand the cause of a decision. Explanation is thus one mode in which an observer may obtain understanding, but clearly, there are additional modes that one can adopt, such as making decisions that are inherently easier to understand or via introspection. I equate 'interpretability' with 'explainability'	An explanation is an assignment of causal responsibility — Josephson and Josephson [81]	Transparency is seen as Lipton's model decomposability. So different from intelligibility. While there are many ways to increase trust and transparency of intelligent agents,two complementary approaches will form part of many trusted autonomous systems: (1) generating decisions ... (2) explaining the decisions.	x	The notion of 'counterfactual' is important in causality to generate the so-called conterfactual explanations.	Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." 2019
Interpretability is not a monolithic concept, but in fact reflects several distinct ideas.	Explanation is post-hoc interpretability. Post-hoc explanations can be verbal, by example, visual.	Understanding the mechanism by which the model works. Also related to simulatability (can you simulate the model's decisions?) and decomposability (what does each model's component do?)	Understandable models are some-times called transparent.	One might hope, however, that byinterpreting supervised learning models, we could gener-ate hypotheses that scientists could then test experimen-tally.	Lipton Z., The Mythos of Model Interpretability
Variety of definitions. Decide what the intended meaning of something is. To explain the meaning of something.	Improving the user's mental model of how a system works. Answering *Why* and *Why not* questions. In other cases, equated to interpretable.	transparency "explains how the system works"; level to which a systemprovides informationabout itsinternal workings or structure "; learly describing the modelstructure ,equations ,parameter values , andassumptions	Intelligible AI systems would need to communicate very complex computational processes to various types of users, autonomously	x	Clinciu, Miruna-Adriana, and Helen Hastie. "A Survey of Explainable AI Terminology." 2019
Divided in model-based and post-hoc. Model-based interpretability is the construction of models that readily provide insight intothe relationships they have learned.	Used as a synonim of interpreting	only mentioned referring to transparent feature engineering	x	Interpretability explanations are not causal explanations	Interpretable machine learning: definitions,methods, and applications. Murdoch, et al., 2019
Used as a synonim of Intelligible: The term interpretable machine learning (IML) often refers to research on models and algorithmsthat are considered as inherently interpretable while explainableAI (XAI) often refers to the generation of (post-hoc) explanationsor means of introspection for black-box model	Post-hoc explanations as in Lipton and Miller	explain how the system works	x	x	Chromik, Michael, and Martin Schuessler. "A Taxonomy for Human Subject Evaluation of Black-
Interoretability is a passive characteristic of a model referring to the level at which it makes sense for a human observer.This feature is also expressed as transparency.	explainability is an active characteristic of a model, any action or procedure to clarify the internal model functions.	As in Lipton, described by Simulability, Decomposability and Algorithmic Transparency.	Understandability is characterized by no means of understanding the internal model functioning. So in this taxonomy understandable is different from intelligible (and also from	Goal pursued by a smaller audience, interest in inferring causal relations from the extensive prior knowledge on the problem and the outcome.	Arrieta, Alejandro Barredo et al. "Explainable Artificial Intelligence (XAI): Concepts
We note that even though "Explainable" is a keyword in the XAI appellation, in ML community the term "interpretable" is more used than "Explainable".	explainability is a powerful tool for justifying AI based decisions	x	Furthermore, it should be noted that none of the aforementioned variation terms (understandable, comprehensible, intelligible…) is enough specific to enable	x	Adadi, Amina & Berrada, Mohammed. (2018). Peeking inside the black-box: A survey
Ability to explain or to present in understandableterms to a human.	x	x	x	Causality implies that the predicted change in output due to a perturbation will occur in the real system	Towards A Rigorous Science of Interpretable Machine LearningFinale

Many terms exist for ML concepts

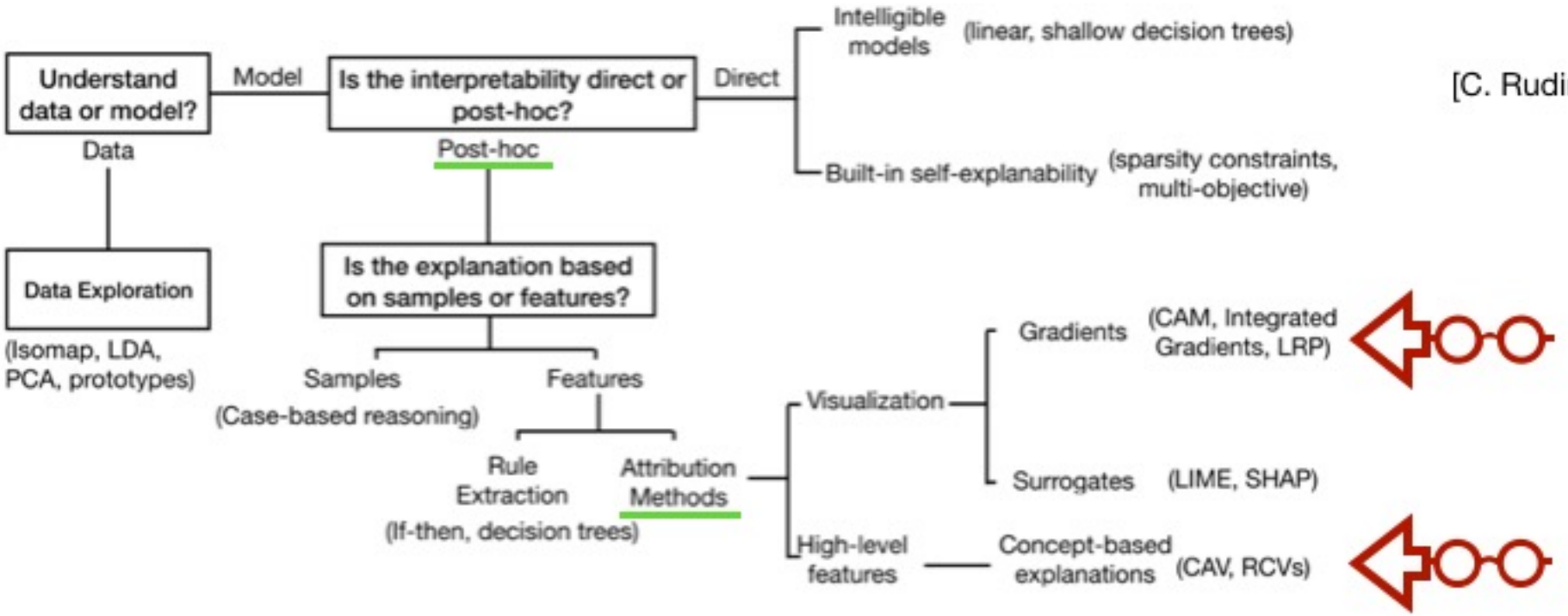
- Understandable, intelligible
  - May depend on the exact user
- Interpretable, explainable
- Transparency, fairness, bias
- Accountable, reliable, trustable, robustness
- Causality, uncertainty
- …

Global taxonomy initiative: https://taxonomyinterpretableai.wordpress.com/

Terminology in various domains



Multiple types of interpretability



[C. Rudin, 2019]

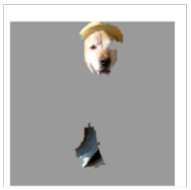
Post-hoc attribution methods are the most common in MIA [Singh et al., 2020]:

- No need of re-training: keeping initial performance
- Easy to apply with off-the-shelf toolboxes
- Model agnostic

Preserve performance + **Improve interpretability**

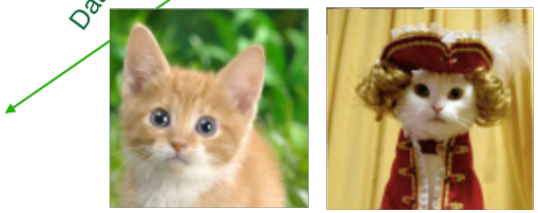
3 dimensions of interpretability

Prediction:
- Why a certain input led to a certain output
- What areas of the input were used to produce the output
- What high-level concepts were involved to reach the output

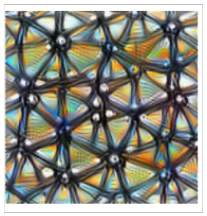


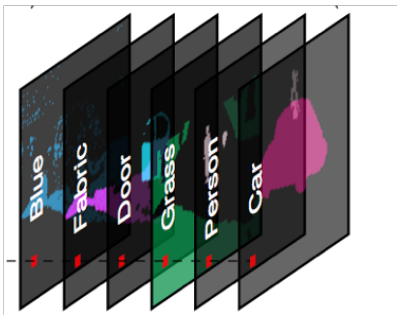
Data:
- What dimensions of the data are the most relevant for the task
- Detection of biases, e.g. prototypes and criticism

UMAP

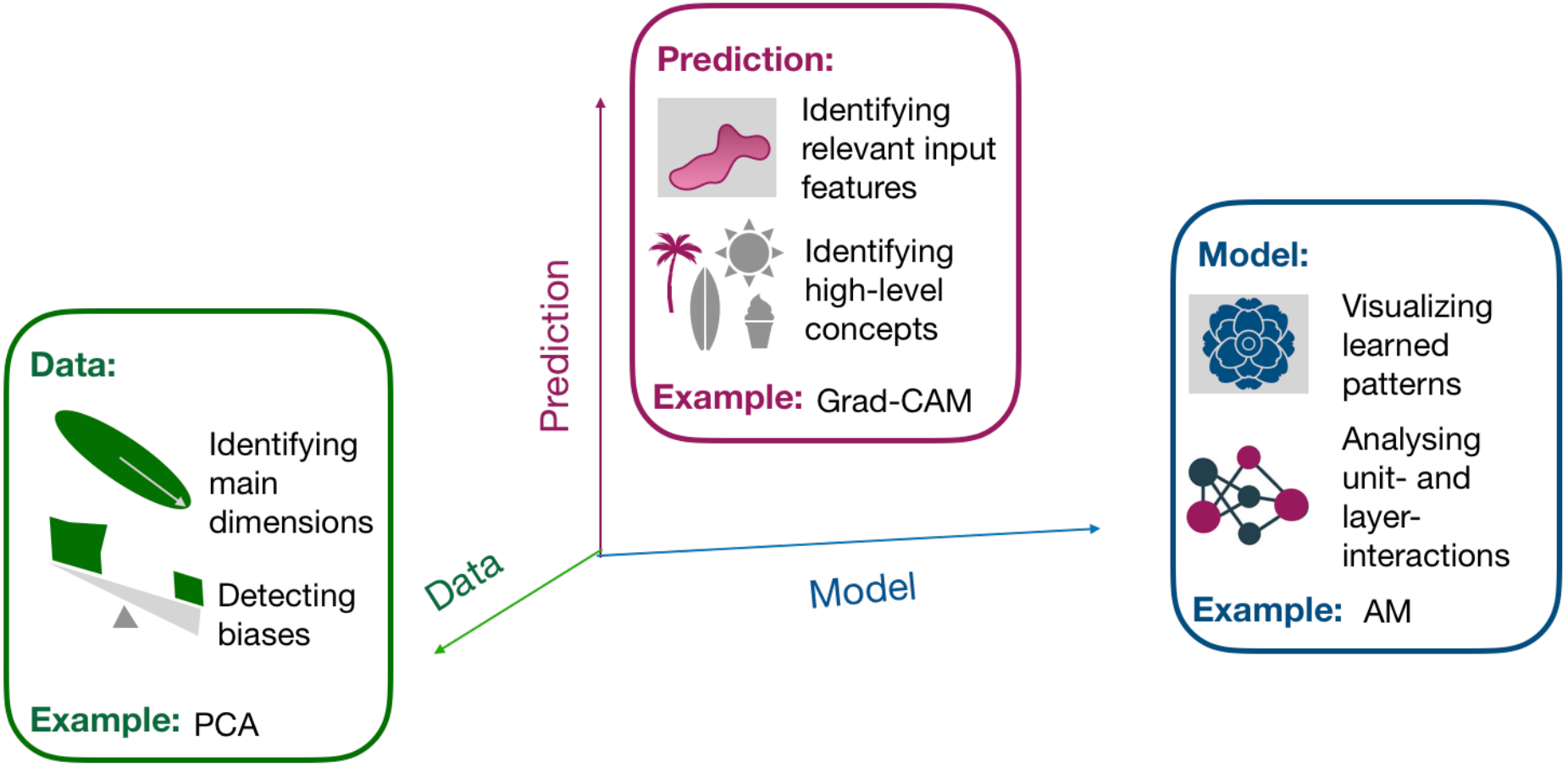


Model:
- How does the model behave, what kind of patterns does it focus on
- How do the nodes and layers interact
- How do we compare across models





[Adapted from talk by Samek and Müller, 2017]



Hes·so VALAIS WALLIS
School of Management & Tourism

Explaining classical machine learning

Steps in visual decision support

Hes·so VALAIS WALLIS
School of Management & Tourism

Resource creation

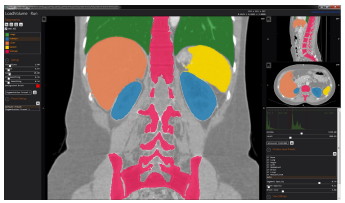
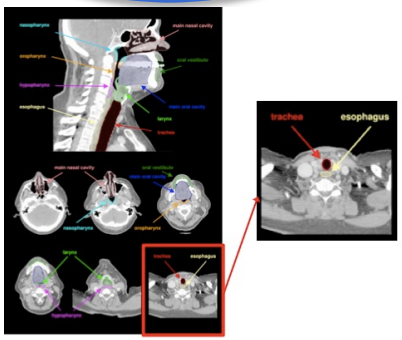
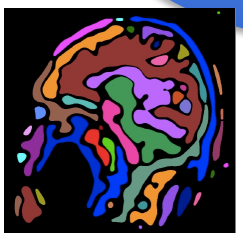


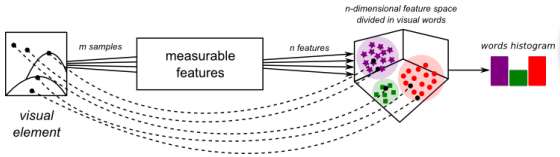
Image pre-treatment



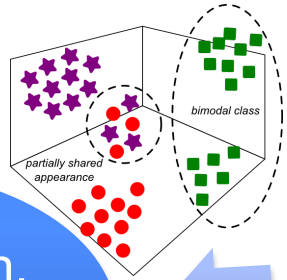
Visual feature extraction



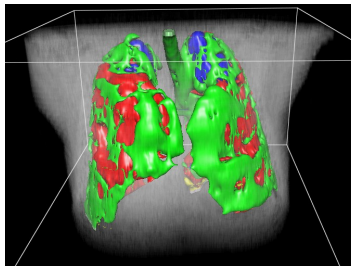
Feature modeling



Multimodal fusion



Classification, detection, retrieval



Visualization, results presentation

Types of visual features

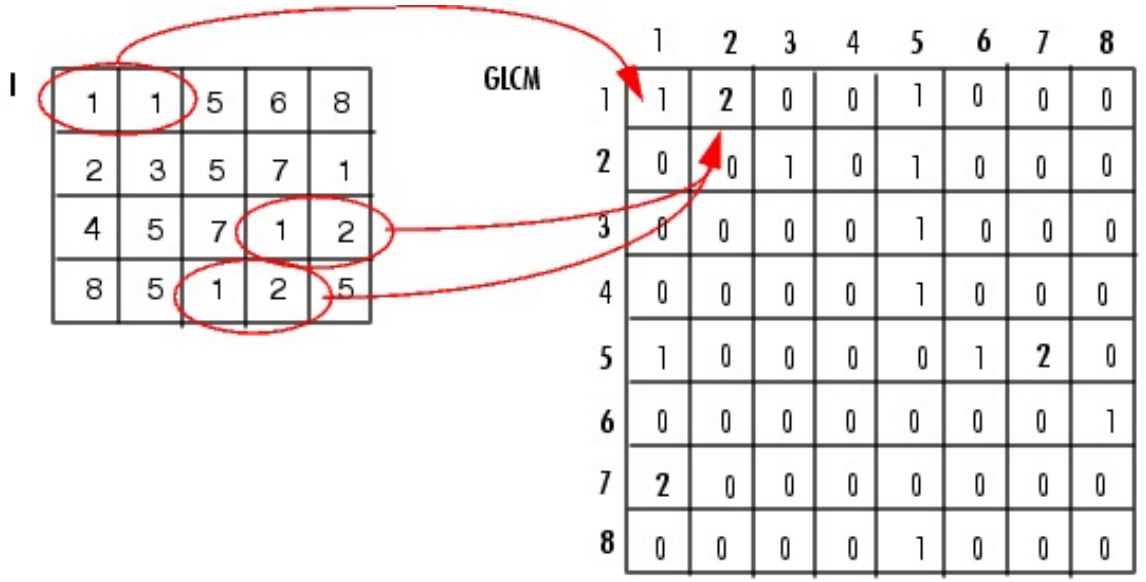
- Handcrafted vs. partially learned vs. fully learned
  - Deep learning vs. traditional approaches
- Classifications of visual features
  - Low level vs. mid level vs. semantic/high-level
    - Higher levels via feature modeling (visual words) or latent semantic techniques, sometimes matching words and pictures
- Type of information that is modeled
  - Shape vs. grey level/color vs. texture
- Local vs. global features
    - Local based on segmentation or partitioning
- 2D vs. 3D vs. nD (3D +time, protocols)

Explaining classical learning

- Handcrafted features often have an understandable meaning
  - Average grey level in the lung in a CT
  - Shape heterogeneity
- Not all features can be interpreted easily, though
  - Even absolute values of features based on coocurrence matrices are hard to interpret
- Distance measures have a clear meaning
  - SVMs are maybe harder
- Decision trees are easy to understand

Gray-Level Co-Occurrence Matrices

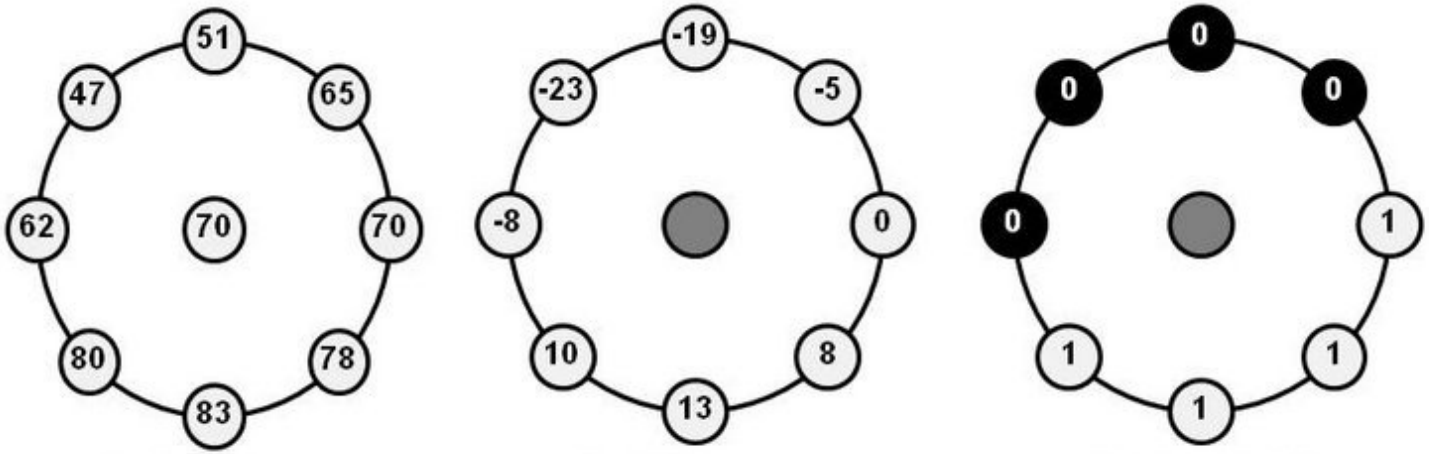
- Statistical descriptor for texture properties of an image by comparing neighboring pixels
  - Direction and distance
    - Features extracted from several matrices in general
  - Extract features from matrix
    - Entropy
    - Contrast
    - Correlation
    - …



Local binary patterns

The value of the LBP code of a pixel (x_c, y_c) is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \qquad s(x) = \begin{cases} 1, if\ x \geq 0; \\ 0, otherwise. \end{cases}$$



1. Sample | 2. Difference | 3. Threshold

1*1 + 1*2 + 1*4 + 1*8 + 0*16 + 0*32 + 0*64 + 0*128 = 15

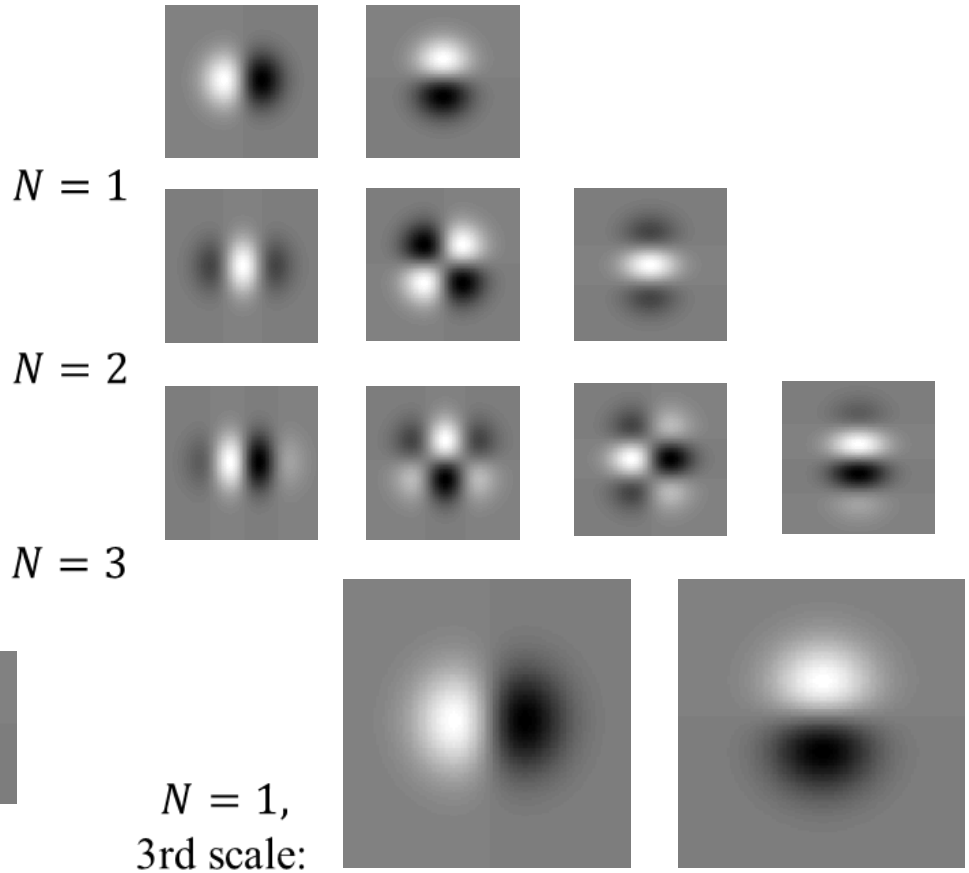
4. Multiply by powers of two and sum

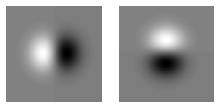
Ojala, T., Pietik
Feature Distributions. Pattern Recognition 19(3):51-59.

ased on

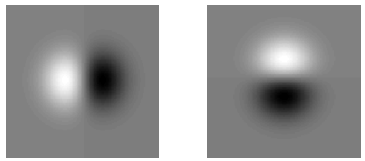
The Riesz transform

- The Riesz transform implements Nth-order directional derivatives:
  - At multiple scales:





$N = 1$,
1st scale:



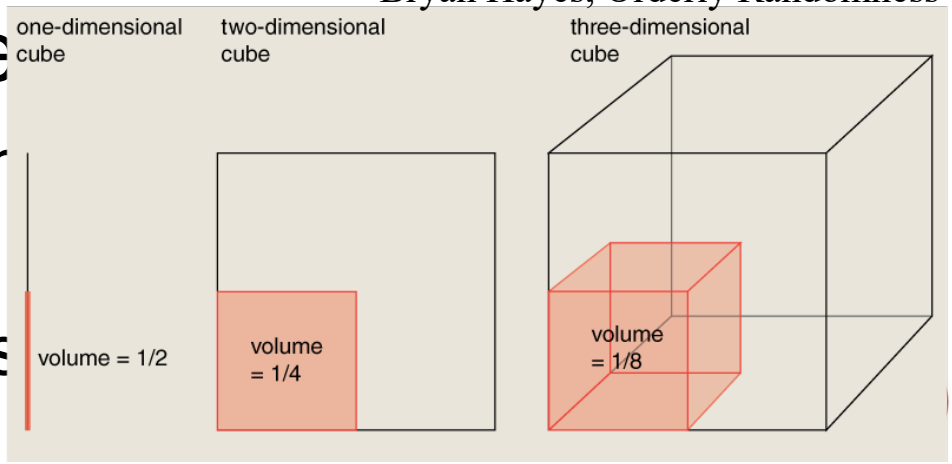
$N = 1$,
2nd scale:

$N = 1$,
3rd scale:

Curse of dimensionality

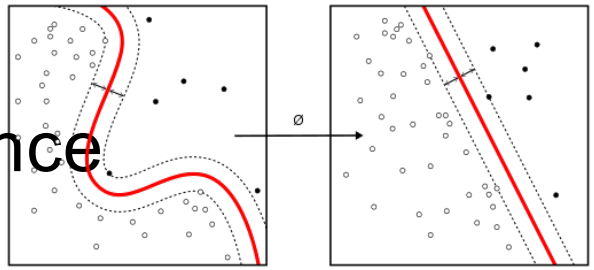
- "The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience." Wikipedia
- Increasing numbers of features mean that generalization requires exponential data amounts
- Volume of a space increase with more dimensions, so data ge
  - Distance between all item becomes similar
  - Automatic classification is then hard (using kNN)

Bryan Hayes, Orderly Randomness



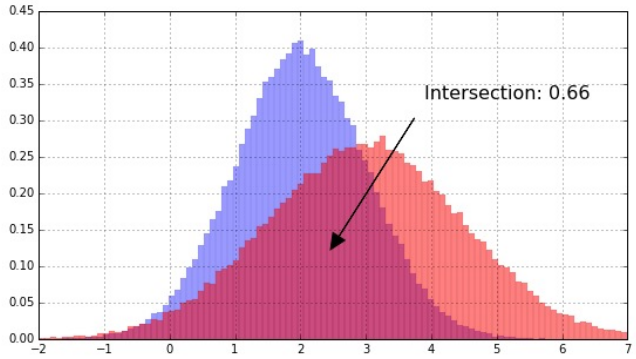
Dimensionality reduction

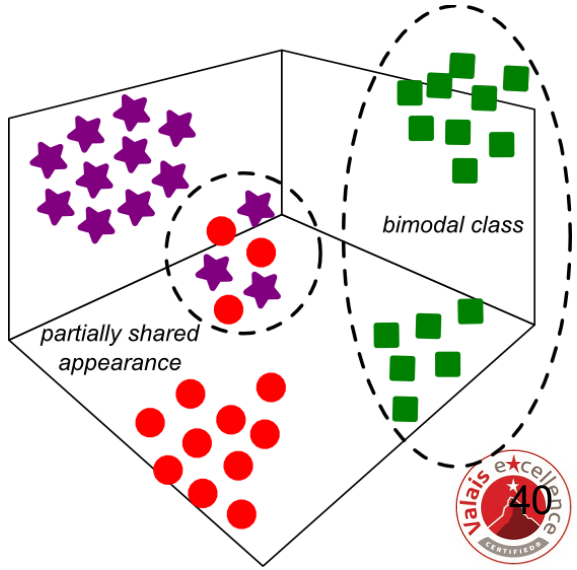
- Feature selection
- Principal component analysis (PCA)
  - Linear mapping of data onto fewer dimensions
  - Mapping to 2D, 3D allows to visualize data
- Kernel PCA
  - Nonlinear space, maximizing variance
- Linear Discriminant Analysis (LDA)
  - Finding a linear combination to best separate classes
- …



Distance measures

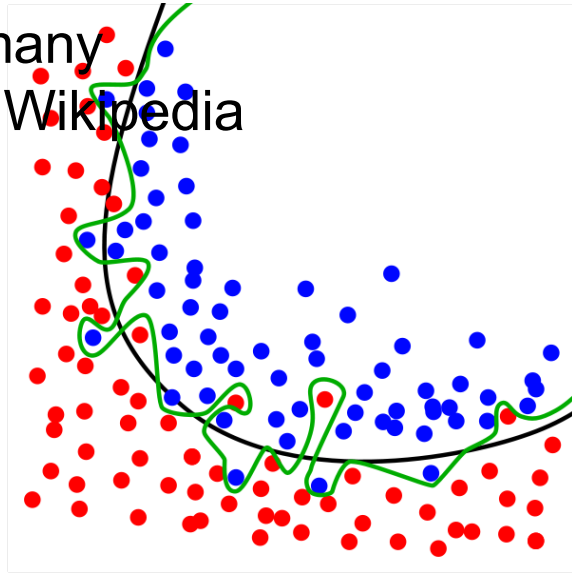
- Visual features represent structures in an n-dimensional space
  - Hopefully our visual features separate the items well
- Many distance metrics exist
  - Histogram intersection
  - City block, Manhattan distance
  - Euclidean distance
  - Earth Movers distance
  - Mahalanobis, Bhattacharyya, …





Overfitting

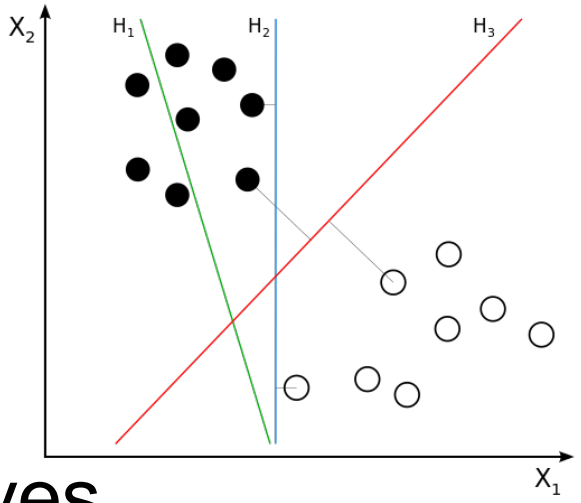
- “In overfitting, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations.” Wikipedia



- Over fit models do not generalize
  - Not good on new or unseen data
- Real risk in learning with many parameters or training on test data
  - Manual tuning to get good results
- A model should perform well on unseen data!
  - Methods such as testing on unseen data can help

Machine learning approaches

- Key nearest neighbors (kNN), simplest approach
  - Parameter free, local approach
- Decision trees
  - Random forests
- Support Vector machines (SVMs)
- Neural networks
  - More on deep learning later
- Boosting
- Linear classifiers such as naïve Bayes



Hes·so VALAIS WALLIS
School of Management & Tourism

Explaining deep learning

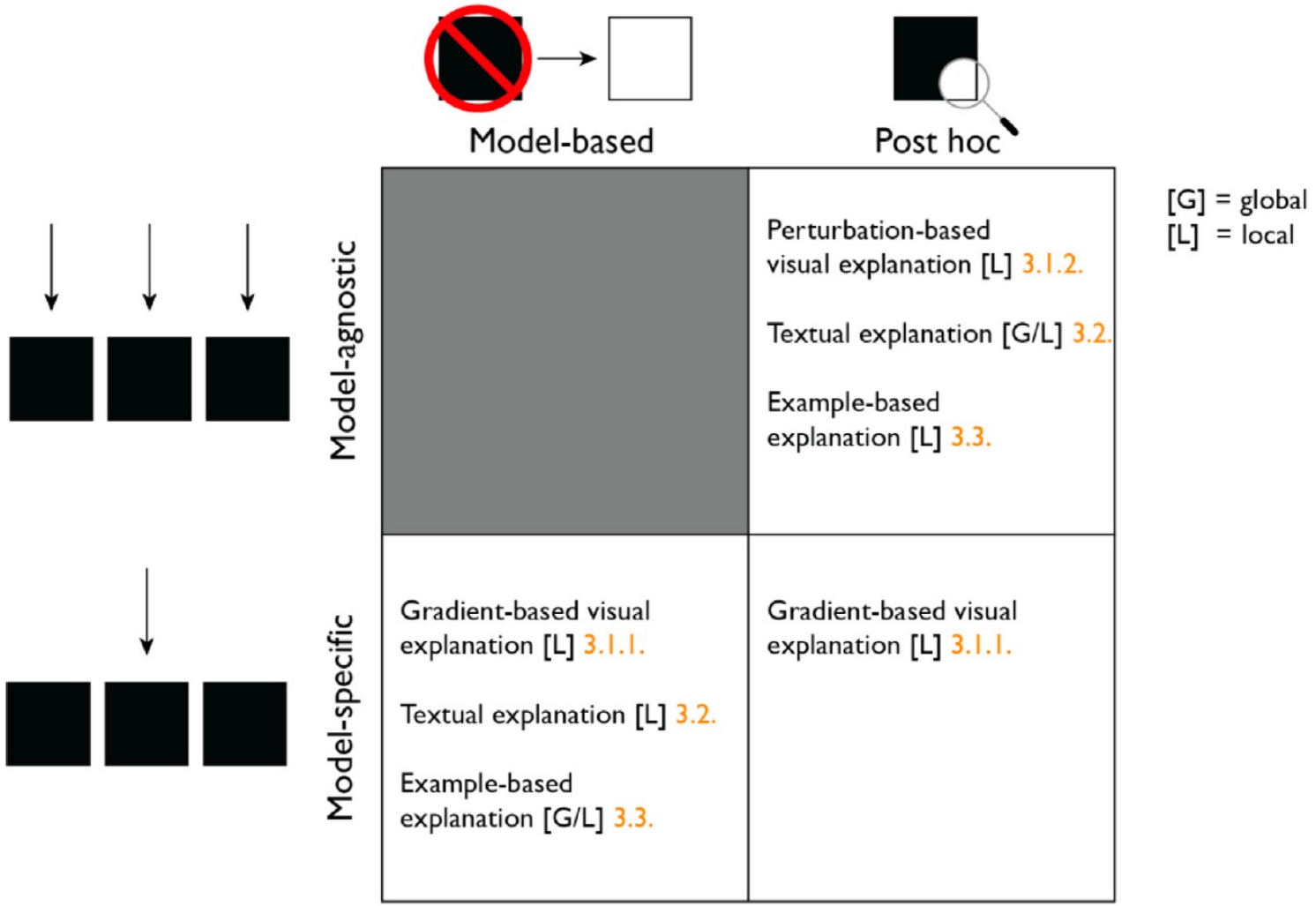
Interpretability of Deep Learning

- Make decisions understandable and remove the black box image
- Make sure that decisions are sound
- Explain why things may not be working
- In medicine it is particularly important to make sure that results can be explained
  - High impact of wrong decisions
  - Integrate information from many sources
- There are several notions of what interpretability is
- Examples:
  - 2D projections, PCA, TSNE
  - Class activation maps, saliency, …

Types of explainability

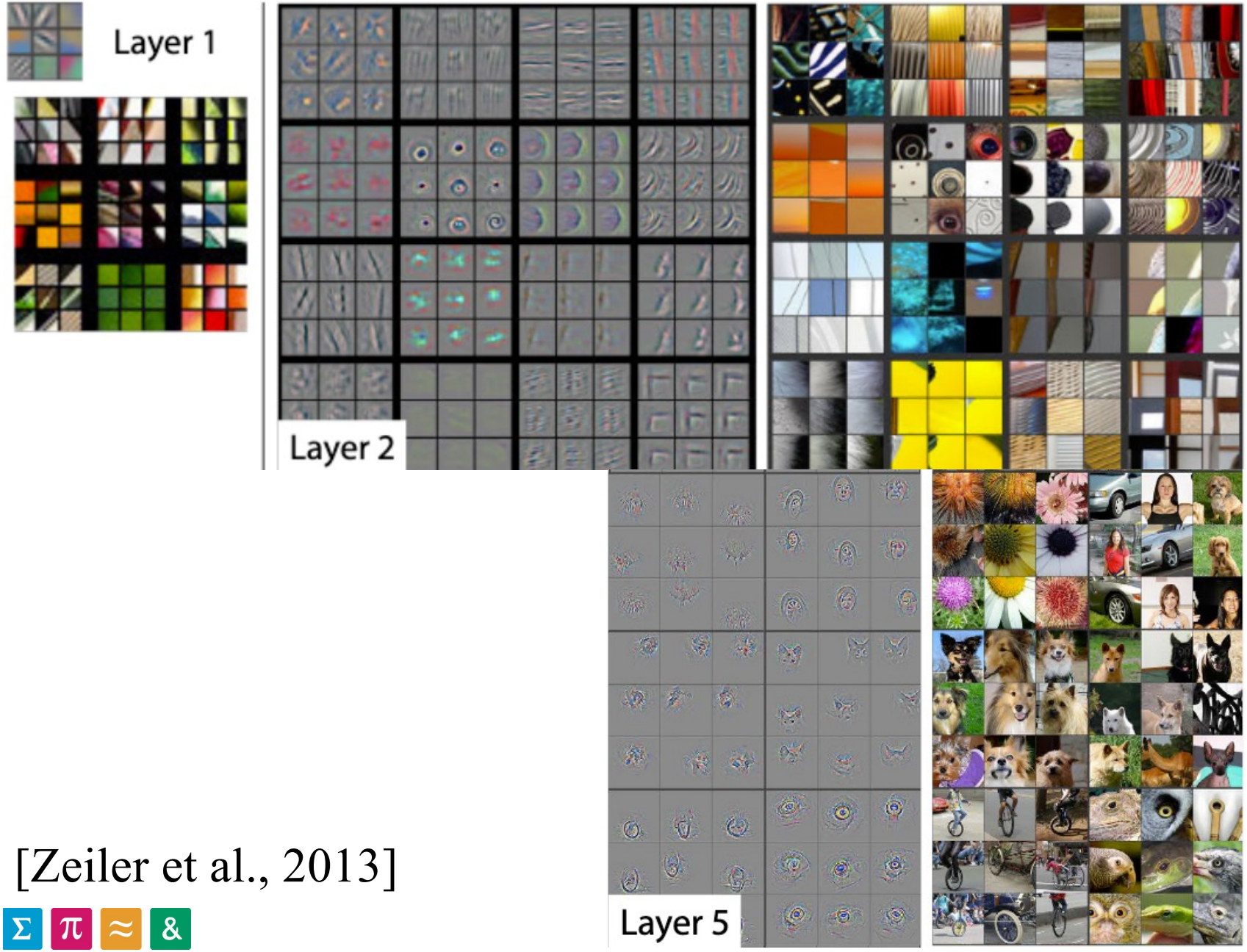
- Type of data: text, structured, images, signals
  - Much of it is for images as CNNs directly work on images
    - Pixel-level, small regions, large regions
    - Translate to concepts outside of the image domain
- Explaining data vs. explaining models
- Post-hoc explications vs. model-based
  - Or the creation of models that are understandable from the start
    - Often does not have the very best results

Types of interpretability



van der Velden, Bas HM, et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis." Medical Image Analysis (2022): 102470.

What a neuron, channel., layer looks at



[Zeiler et al., 2013]

Visualizations for interpretability

What pixels need to be changed the least to affect the output the most?

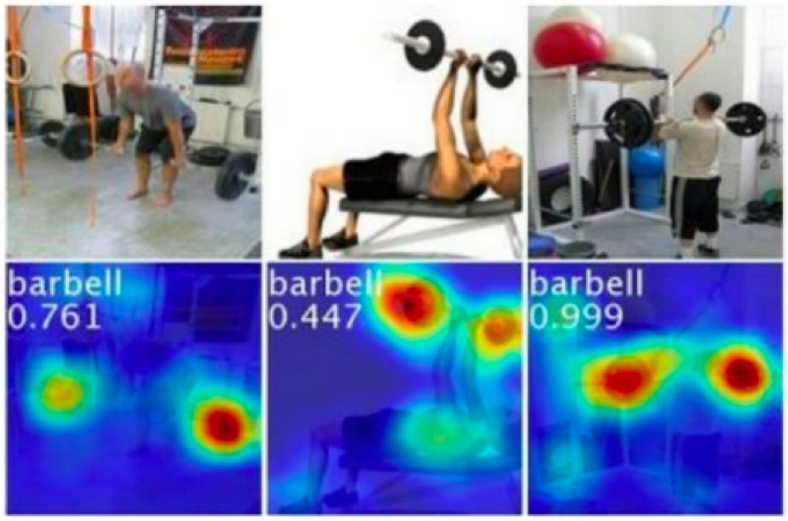
$$Saliency = \frac{\partial output}{\partial input}$$

[Erhan et al., 2009, Simonyan et a., 2013, Springenberg et al. 2015, Fong and Vedaldi, 2017, Sundararajan et al. 2017, Smilkov et al. 2017, many more...]

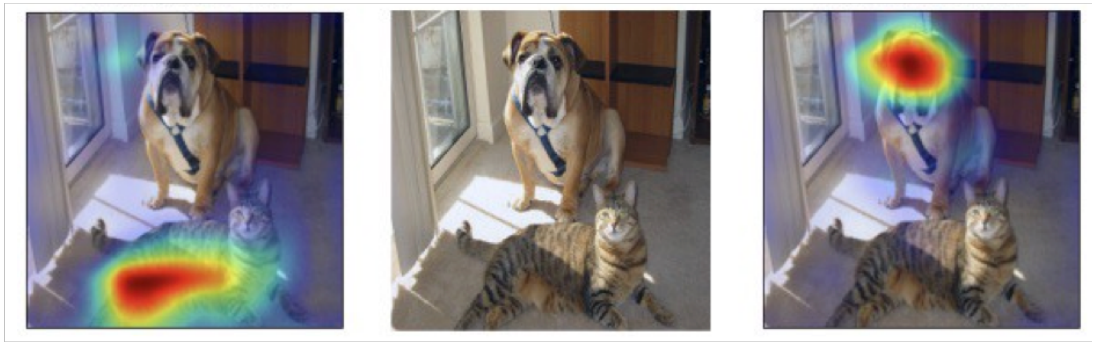
Class activation maps (CAMs)

[Zhou B. et al., 2016]

- Works on any classification network for images
- Extendable to regression
- Local interpretability method: per-sample explanations
- Post-hoc

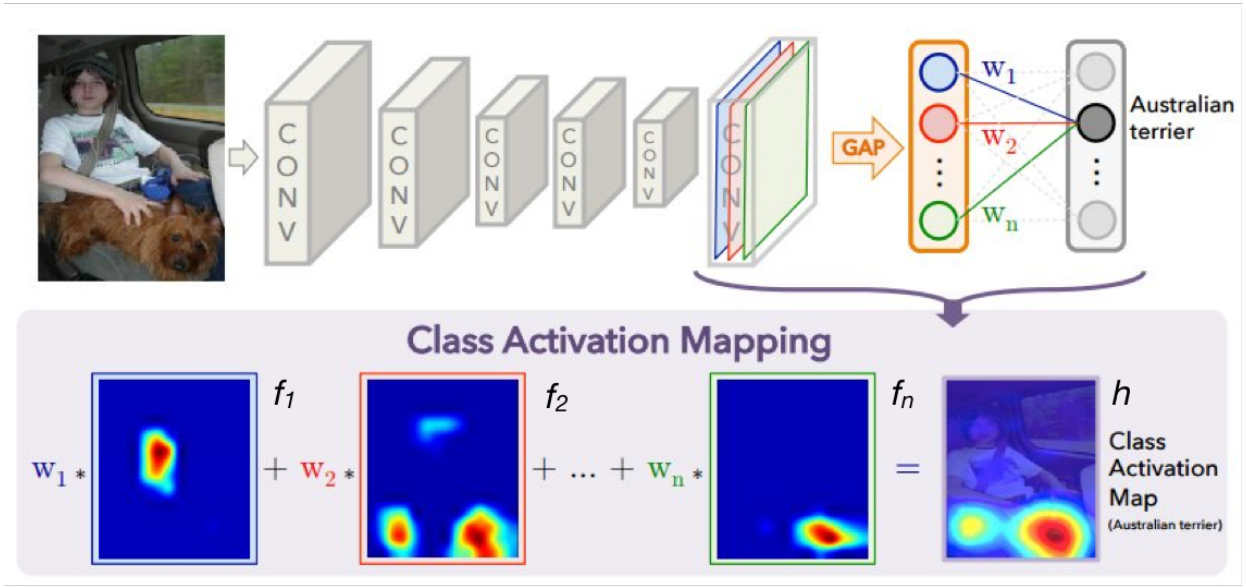


Class activation maps for one object class



Class activations maps

Output of convolution layers are multiple feature maps of dimensions w x h x c



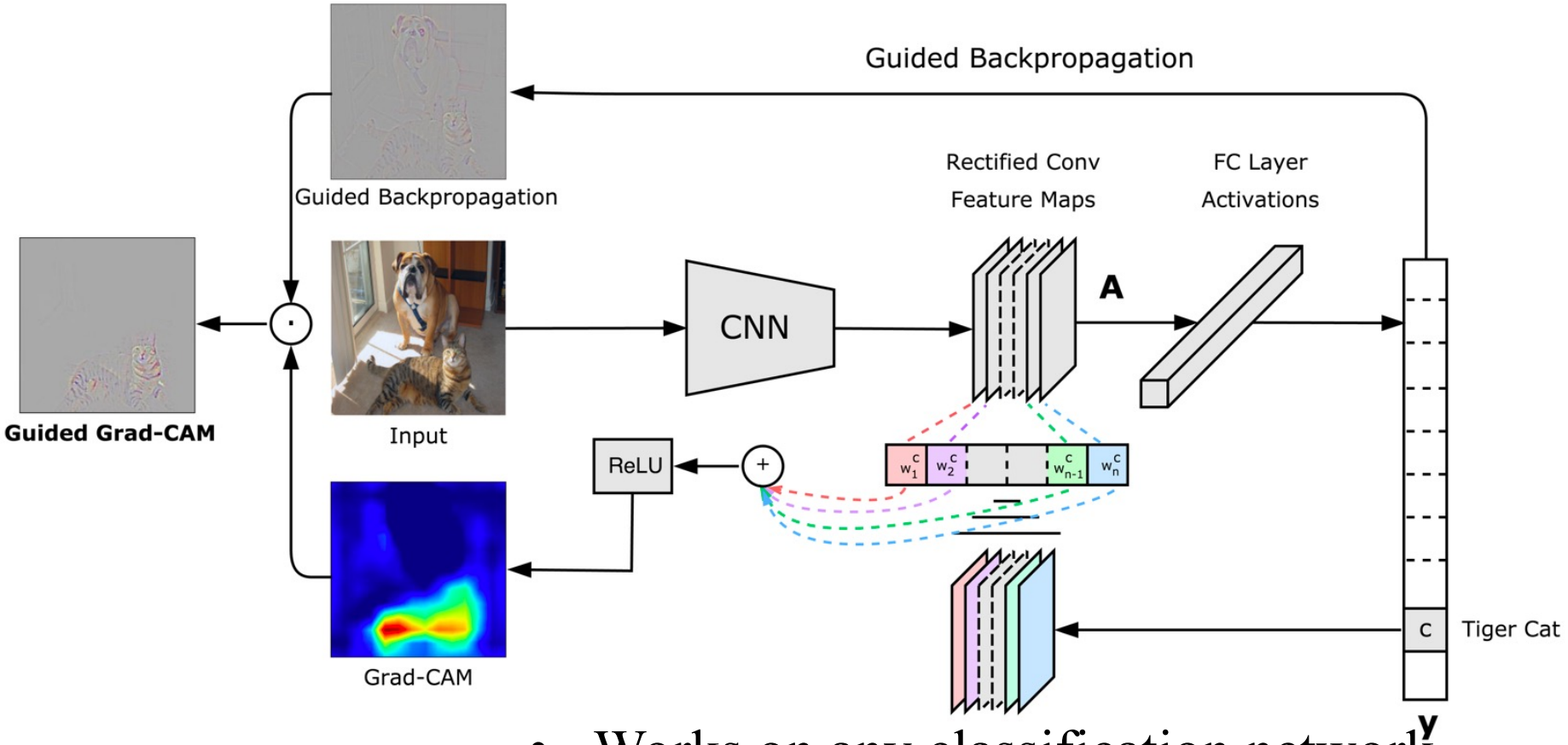
Global Average Pooling takes the spatial average (on dimensions w and h) of each feature map to obtain a vector of dimension c

Such vector is weighted by the last layer vector of weights W to determine the contributions to the classification layer

We can use this weight vector W to weight the feature maps as if we were not performing the pooling. The result will be a map of weighted contributions for each feature.

$$h = w_1 f_1 + w_2 f_2 + \ldots + w_n f_n$$

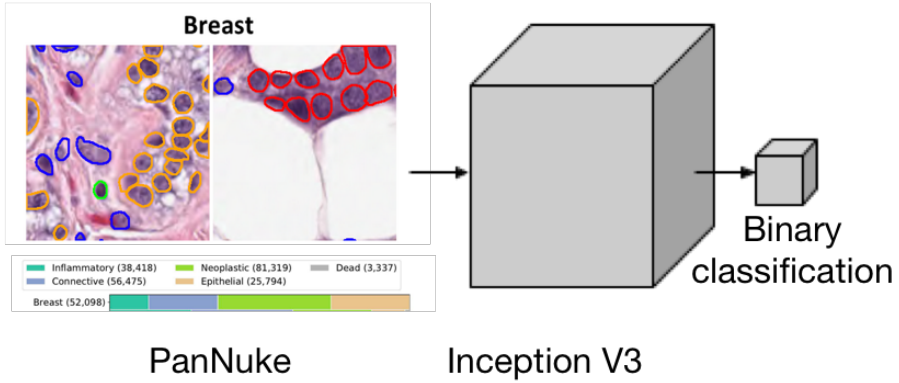
Gradient-weighted CAM



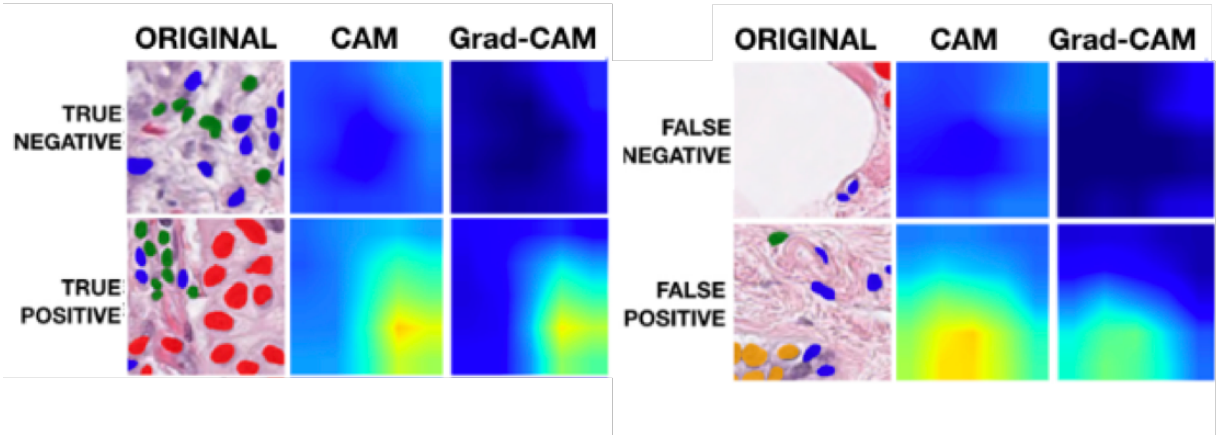
- Works on any classification network
- Extendable to regression
- Local interpretability method: per-sample explanations
- Post-hoc

[Selvaraju et al., 2017]

Applied on digital pathology

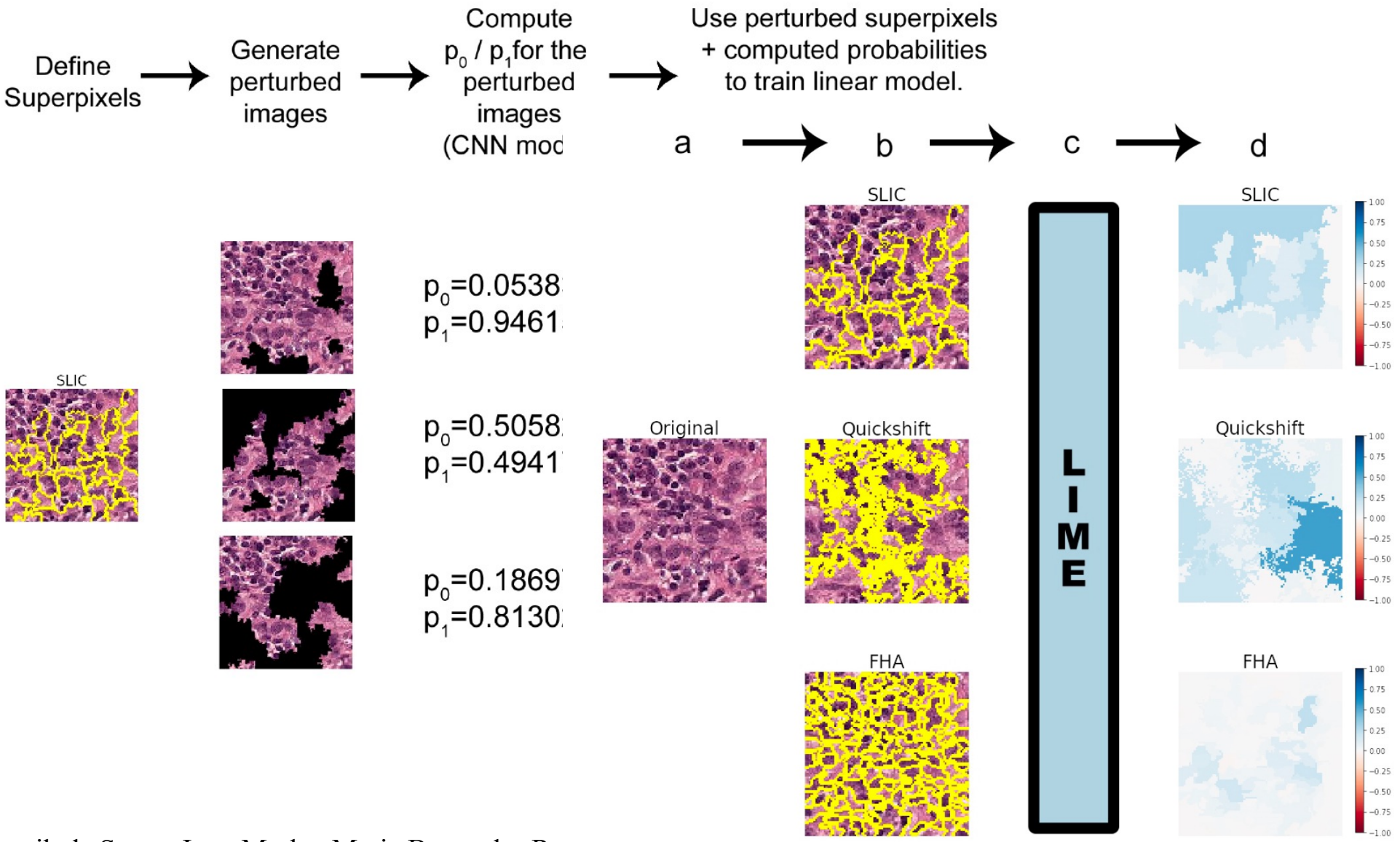


Remarks:
- Do not normalize the heatmaps
- Perform quantitative analyses too
- Structural similarity and intersection over union can be used as quantitative metrics



Visual inspection:
- By not normalizing the heatmaps we can see the low activations on negative images
- Positive images show higher activations on the areas of the “red” nuclei, that are nuclei presenting neoplasticity

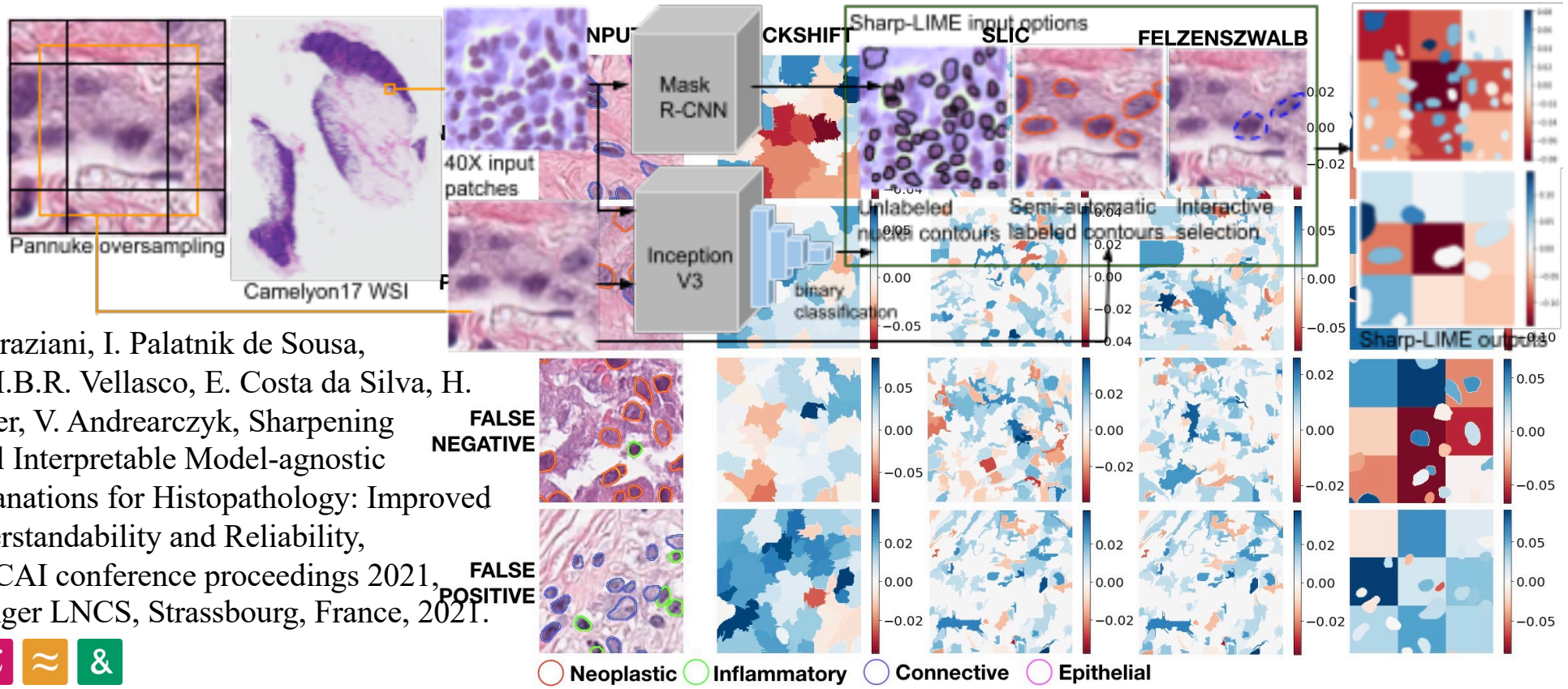
Local Interpretable Model-Agnostic Explanations (LIME)



Palatnik de Sousa, Iam, Marley Maria Bernardes Rebuzzi Vellasco, and Eduardo Costa da Silva. "Local interpretable model-agnostic explanations for classification of lymph node metastases." Sensors 19.13 (2019): 2969.

Sharp LIME vs. LIME

- Improve visualizations of regions that are relevant for the decision of a DNN
  - LIME is commonly used to highlight regions, but interpretations can be difficult



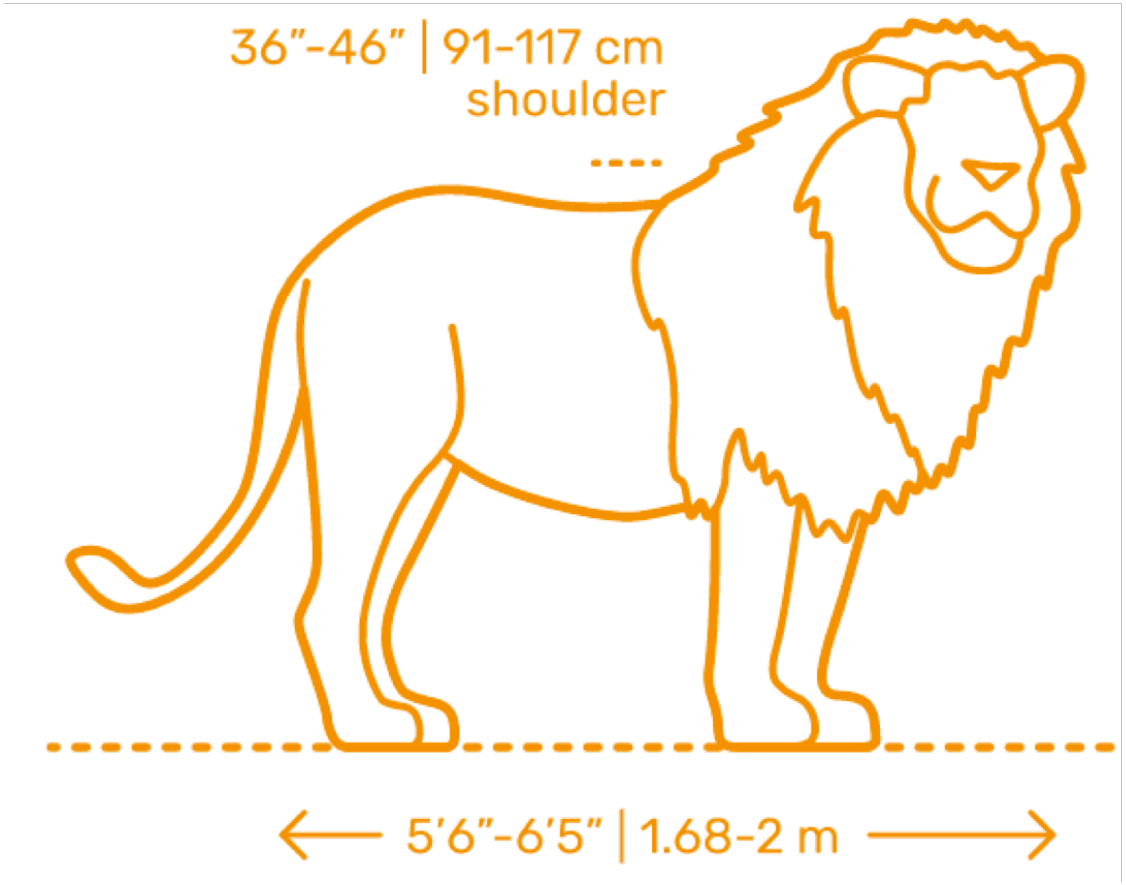
M. Graziani, I. Palatnik de Sousa, M. M.B.R. Vellasco, E. Costa da Silva, H. Müller, V. Andrearczyk, Sharpening Local Interpretable Model-agnostic Explanations for Histopathology: Improved Understandability and Reliability, MICCAI conference proceedings 2021, Springer LNCS, Strassbourg, France, 2021.

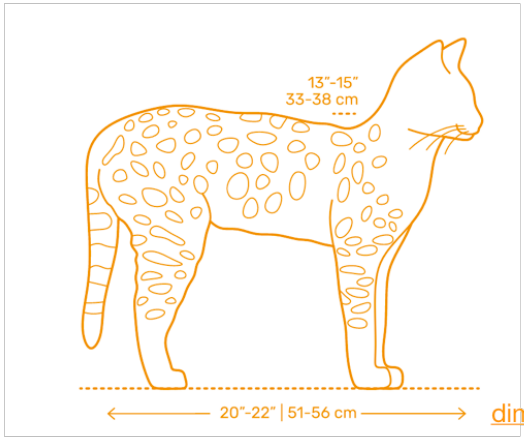
How to evaluate explainability

- It is subjective what people understand
  - Depends on the level of knowledge
- Concrete impact on the decision making?
  - Speed of the decision-making
  - Quality of the decisions
    - Possibly even confidence in the decisions
  - Satisfaction with the use of the system
- Impact that the decisions taken might have
  - Better patient treatment, survival time, quality of life, …
- Linked with human computer interaction

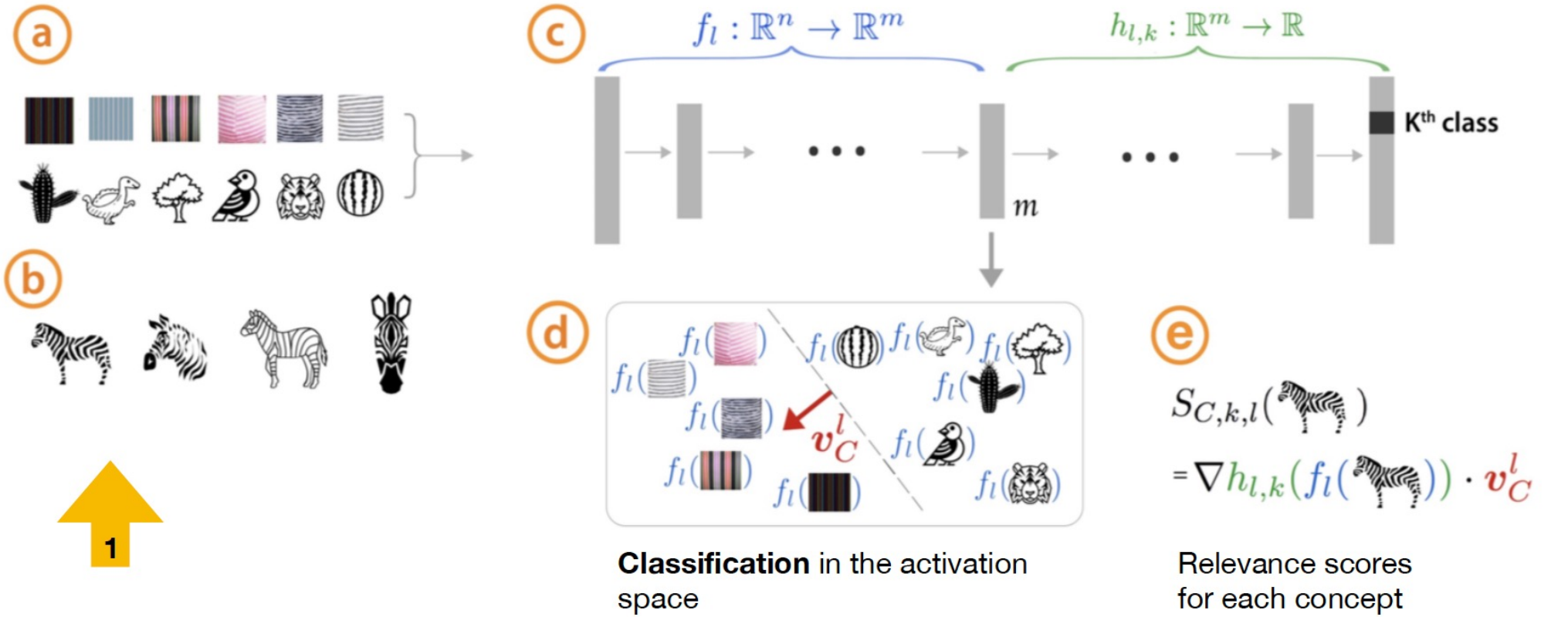
Concept attribution

Main idea: conceptual descriptions identify object categories





Concept Activation Vectors

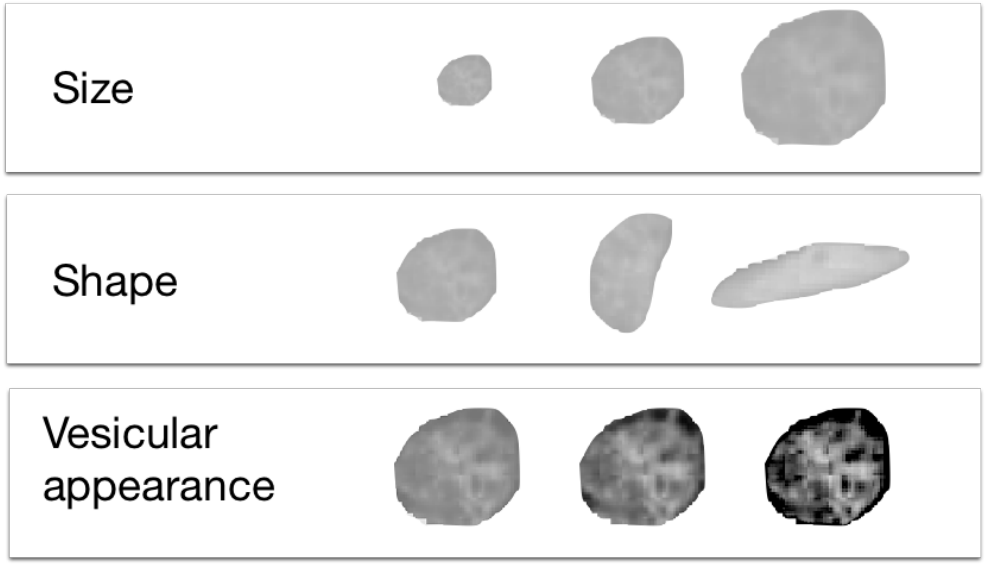


Classification in the activation space

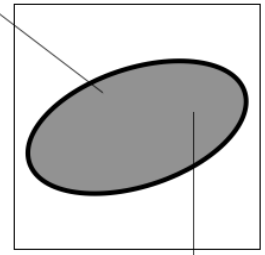
Relevance scores for each concept

[Kim, 2018]

Histopathology concepts (nuclei)



Segmentation (manual or automatic)

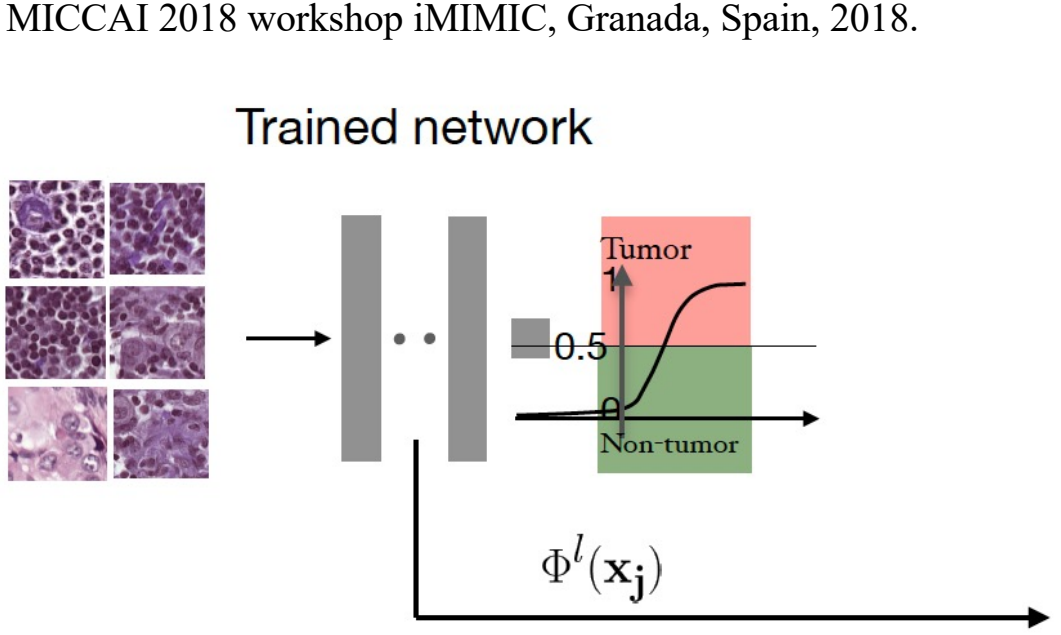


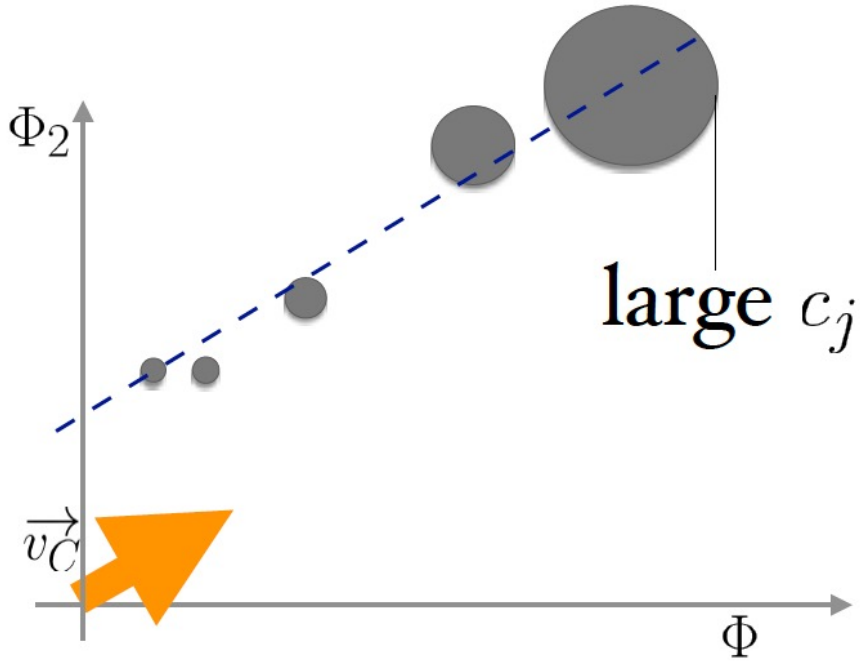
Handcrafted features, texture descriptors, shape, size, …

Regression concept vectors

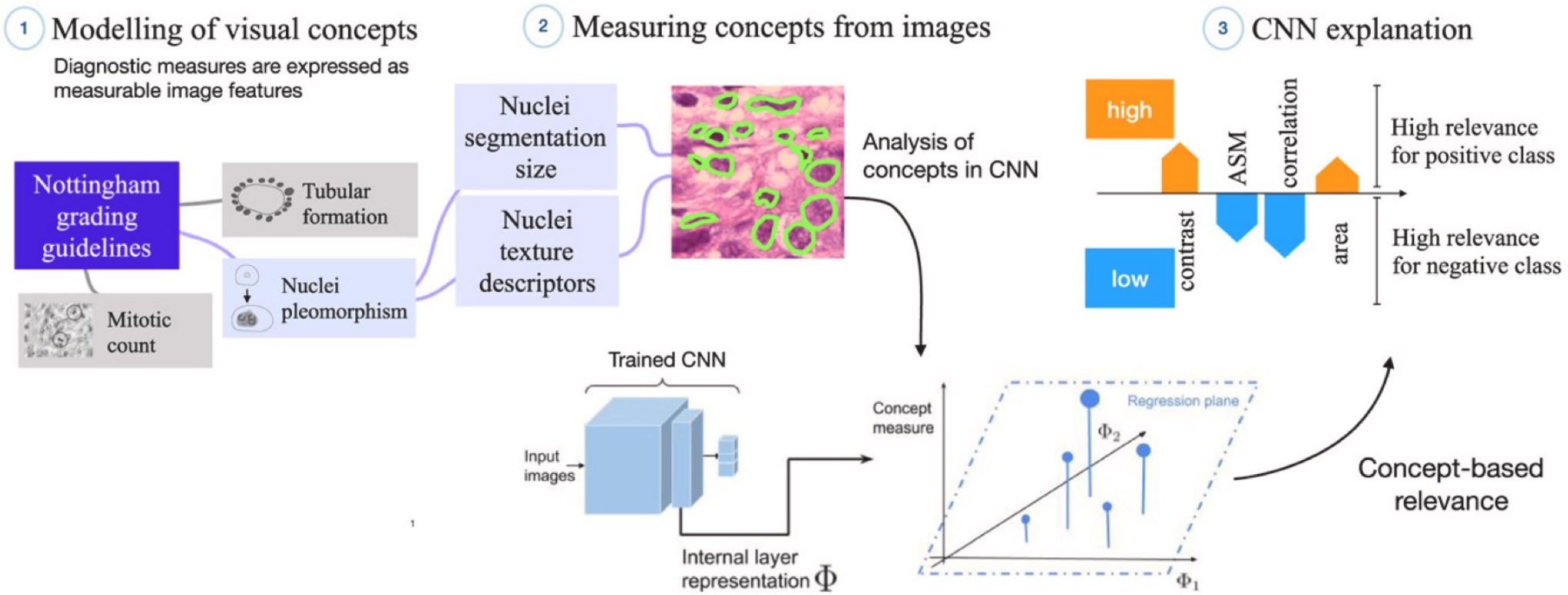
- Identify existing features and check how the decision layers correlate to these features
  - i.e.: nuclei size, heterogeneity
  - How much can a decision be explained with these

M. Graziani, V. Andreaczyk, H. Müller, Regression Concept Vectors for Bidirectional Explanations in Histopathology, MICCAI 2018 workshop iMIMIC, Granada, Spain, 2018.





Regression concept vectors



Criticism towards explainbility

- The opaque function of the black box remains the basis for the ML decision
  - And for the post-hoc explainability
- As the explainability is based on the black box, but not using the same model, it can not be perfect, is not focused on accuracy
- Often the outcomes are not easy to interpret
- Human factors are often ignored but are essential
- Uncertainty on decisions and also explications is needed

Babic, Boris, et al. "Beware explanations from AI in health care." Science 373.6552 (2021): 284-286.

Interactive explainability

- User tests are required to evaluate explainability
  - Comparing user groups
  - Addressing several groups with separate approaches
  - Evaluating different aspects: effectiveness, efficiency, satisfactions, confidence, …
- How can the explanations be adapted to a user with the feedback from the user?
  - Adapt to the type of user
  - Combine several types of interpretability

Next steps in explainable AI

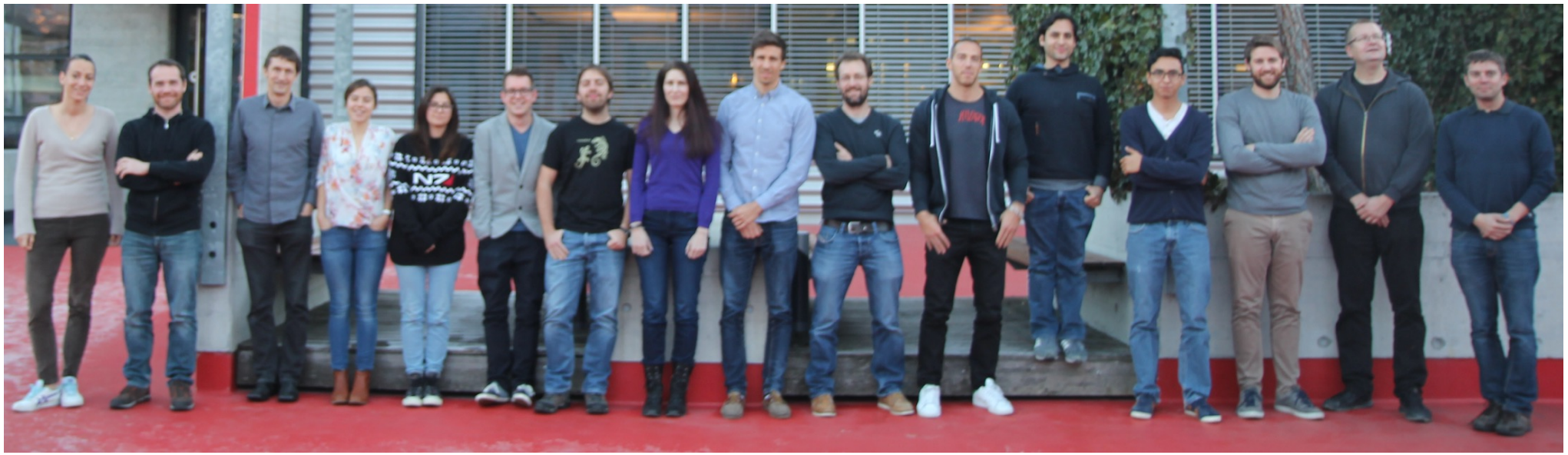
- Clearly, we need to better understand AI models and the ways in which they are build
  - Limiting complexity can help, so avoid ensembles, …
- Human factors need to be taken into account
  - Depending on the exact users of systems
- For medical XAI we need clinical trials
  - To really evaluate the impact of different approaches on the outcomes (decision making of clinicians)
- Approaches need to be compared and then the best improved (organize scientific challenges?)

Conclusions

- XAI/Interpretability is an important research topic in medical imaging and beyond
- It can help to keep the high quality of deep learning but reduce the black box character of it
  - Essential for real, clinical use
- Currently a tendency is to win scientific challenges and create ensembles for small gains
  - Very complex models
  - Maybe we should concentrate on simpler models?
- Robustness to changes in input may be more important than pure performance

Contact

- More information can be found at
  - http://medgift.hevs.ch/
  - http://publications.hevs.ch/
- Contact:
  - Henning.mueller@hevs.ch



References

- Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning. PMLR, 2018.
- Donaldson, Molla S., Janet M. Corrigan, and Linda T. Kohn, eds. "To err is human: building a safer health system." (2000).
- Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." arXiv Preprint arXiv:1706.07269. (2017).
- Been Kim, Rajiv Khanna, and Oluwasanmi O. Koyejo. "Examples are not enough, learn to criticize! Criticism for interpretability." Advances in Neural Information Processing Systems (2016).
- Singla, Sumedha, et al. "Explanation by progressive exaggeration." arXiv:1911.00483 (2019).
- 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.
- Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1.5 (2019): 206-215.
- Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- van der Velden, Bas HM, et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis." Medical Image Analysis (2022): 102470.
- Reyes, Mauricio, et al. "On the interpretability of artificial intelligence in radiology: challenges and opportunities." Radiology: artificial intelligence 2.3 (2020).

More references

- Babic, Boris, et al. "Beware explanations from AI in health care." Science 373.6552 (2021): 284-286.
- Holzinger, Andreas, et al. "Causability and explainability of artificial intelligence in medicine." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9.4 (2019): e1312.
- Graziani, Mara, et al. "Concept attribution: Explaining CNN decisions to physicians." Computers in biology and medicine 123 (2020): 103865.

More recommended reading

- https://arxiv.org/pdf/2112.03245.pdf
- https://proceedings.neurips.cc/paper/2021/file/251bd0442dfcc53b5a761e050f8022b8-Paper.pdf
- https://arxiv.org/pdf/2112.03245.pdf
- https://link.springer.com/chapter/10.1007/978-3-030-93736-2_40
- https://www.medrxiv.org/content/10.1101/2022.04.30.22274520v1.full.pdf
- https://www.nature.com/articles/s42256-020-00265-z#citeas
- https://arxiv.org/abs/2012.04456 (data exploratory analysis)
- https://openreview.net/forum?id=xNOVfCCvDpM
- https://arxiv.org/abs/2105.15164