On Human-Centered AI in Medicine

Andreas Holzinger

Human-Centered AI Lab (Holzinger Group)

Institute for Medical Informatics/Statistics, Medical University Graz, Austria

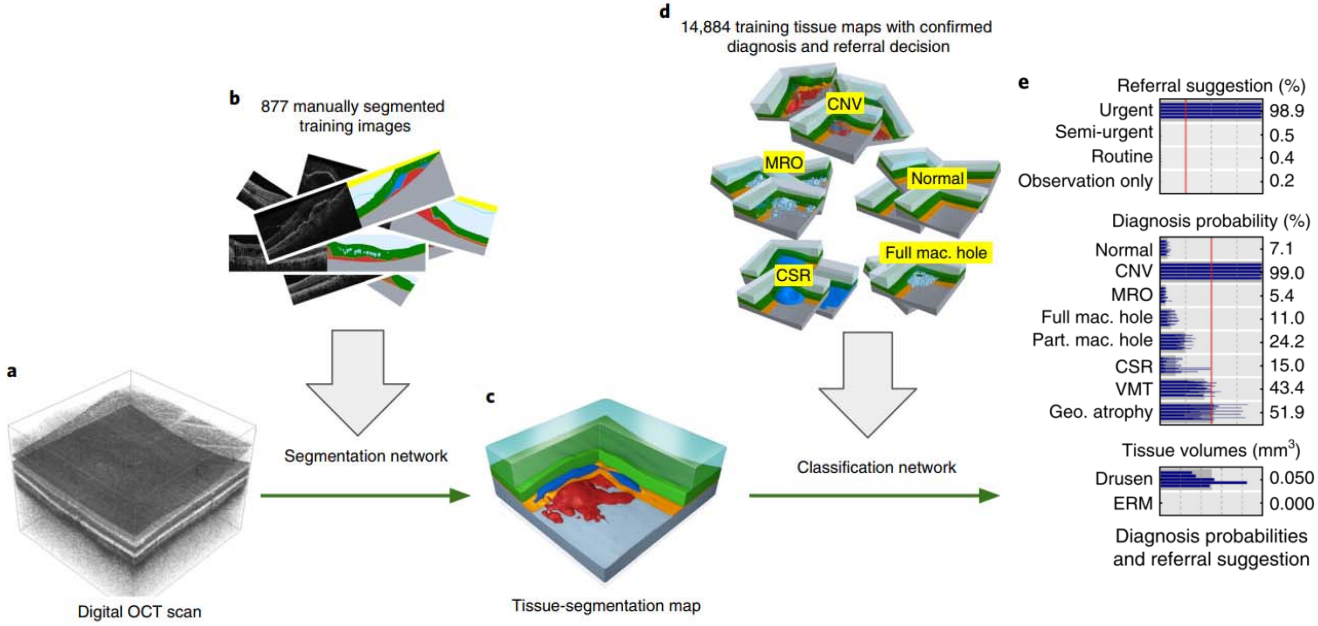
and

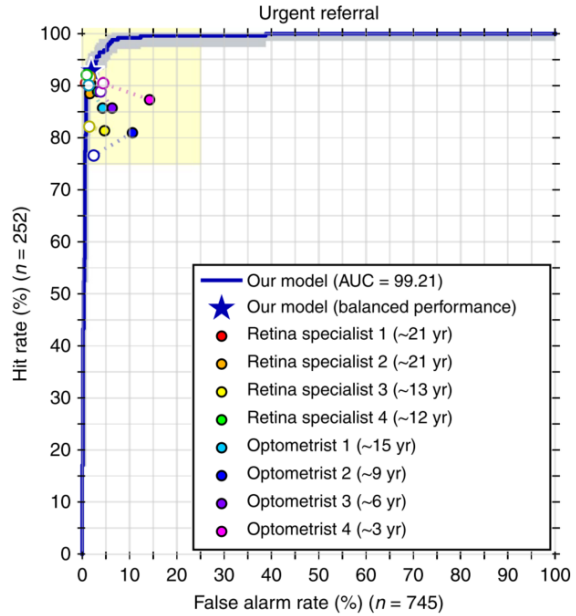
Explainable AI-Lab, Alberta Machine Intelligence Institute, Edmonton, Canada

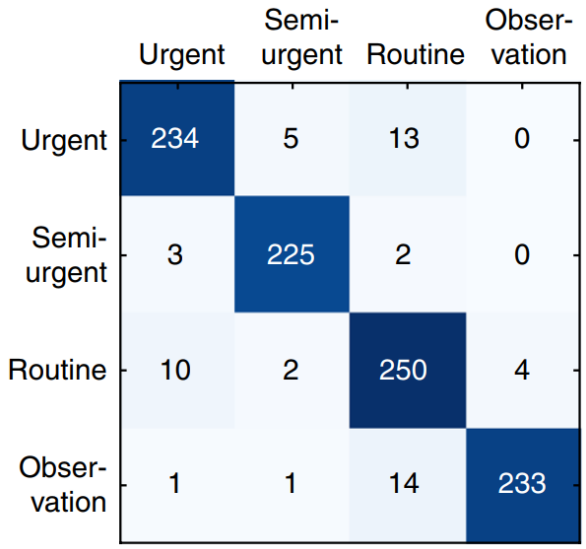
HCAI HUMAN-CENTERED.AI

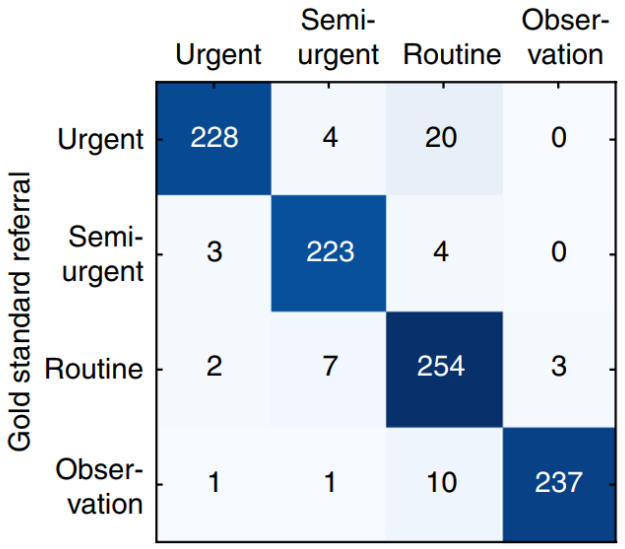
Towards Human-level AI in retinopathy ...

Jeffrey De Fauw et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24, (9), 1342-1350



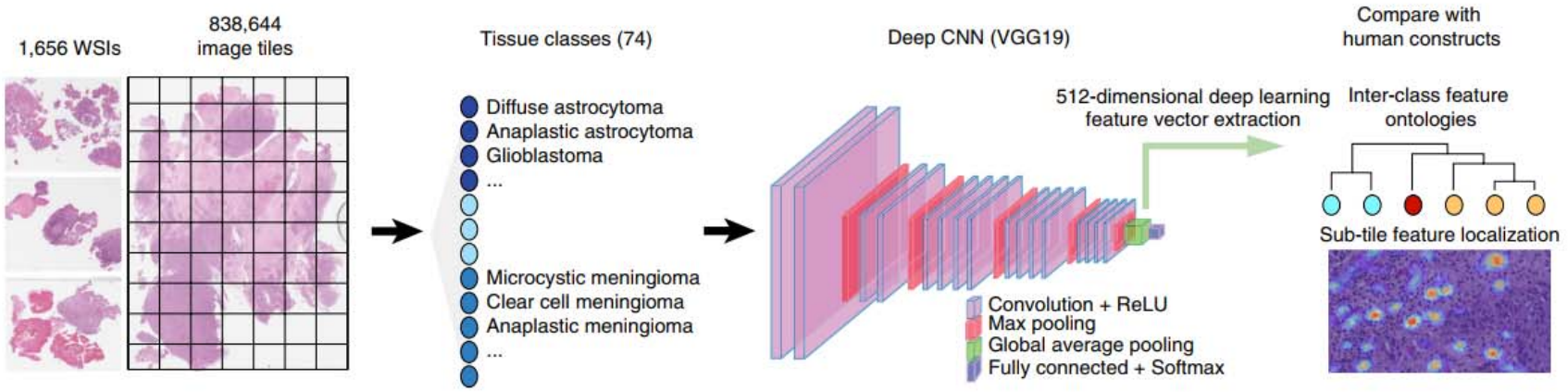


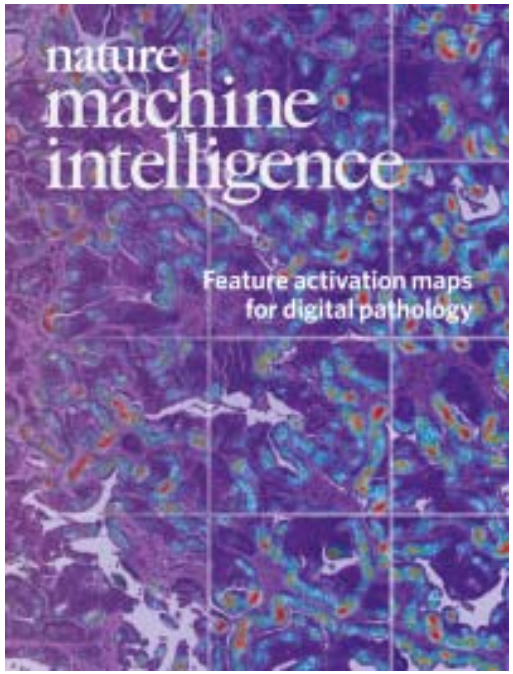




Towards Human-level AI in pathology ...

NATURE MACHINE INTELLIGENCE ARTICLES



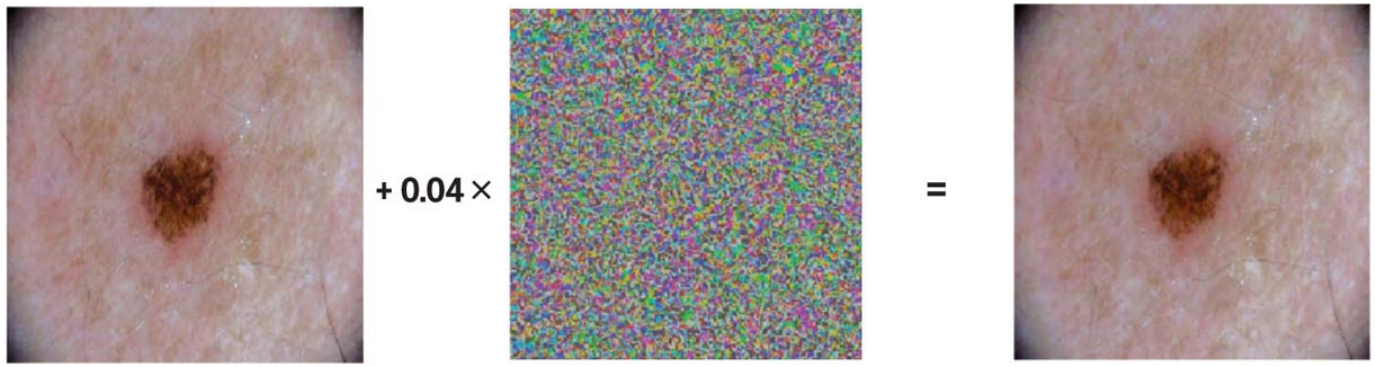


Kevin Faust, Sudarshan Bala, Randy Van Ommeren, Alessia Portante, Raniah Al Qawahmed, Ugljesa Djuric & Phedias Diamandis 2019. Intelligent feature engineering and ontological mapping of brain tumour histomorphologies by deep learning. Nature Machine Intelligence, 1, (7), 316-321, doi:10.1038/s42256-019-0068-6.

13. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at http://arxiv.org/abs/1409.1556 (2014).
14. Holzinger, A. et al. Causability and explainabilty of artificial intelligence in medicine. *WIRES Data Min. Knowl. Discov.* **9**, e1312 (2019).
15. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at http://arxiv.org/abs/1702.08608(2017).

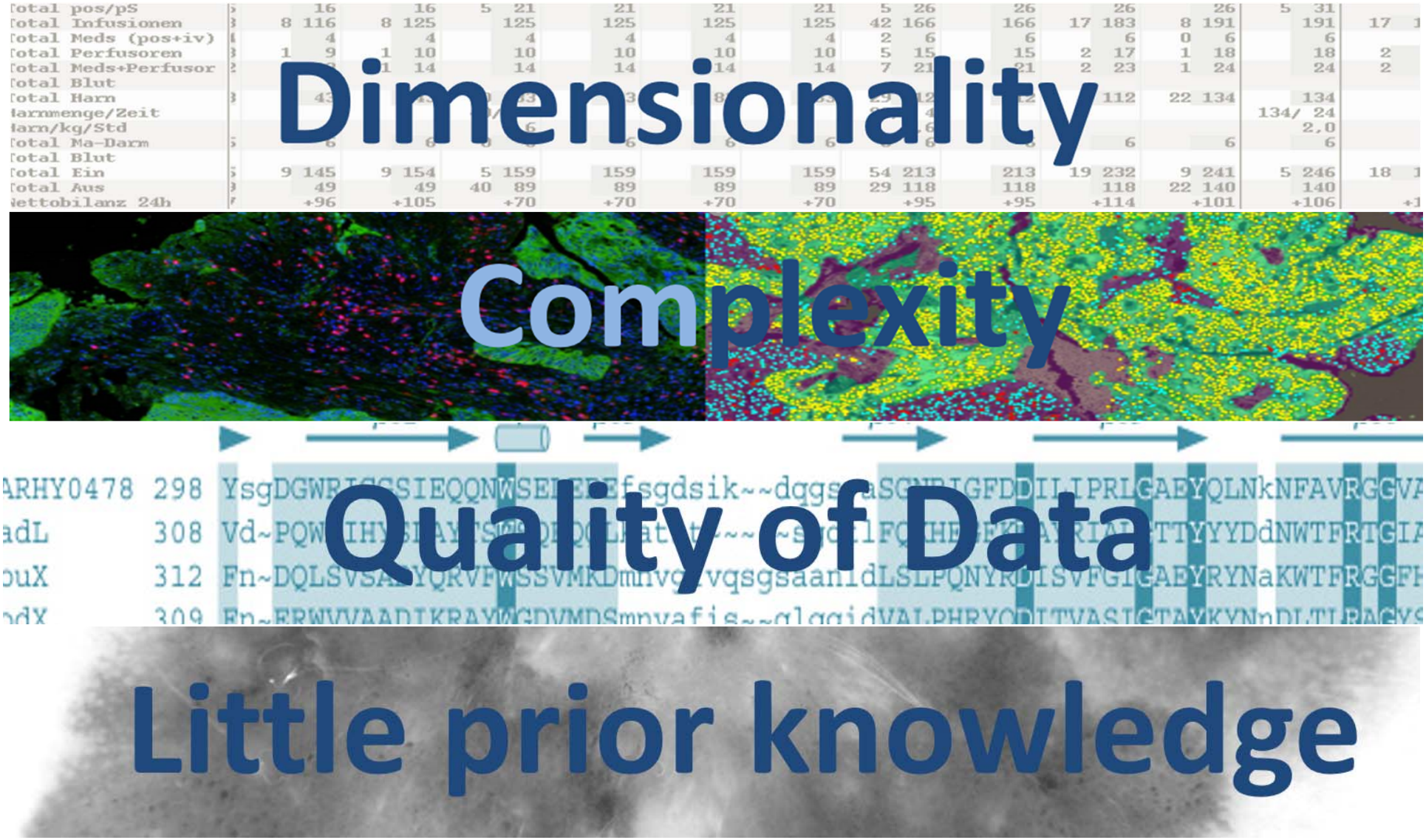
Why can AI solve some tasks better than humans ?
How did the AI even came to these results ?
What if the input data changes counterfactually ?

Robustness & Interpretability



Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam & Isaac S. Kohane 2019. Adversarial attacks on medical machine learning. Science, 363, (6433), 1287-1289, doi:10.1126/science.aaw4399.

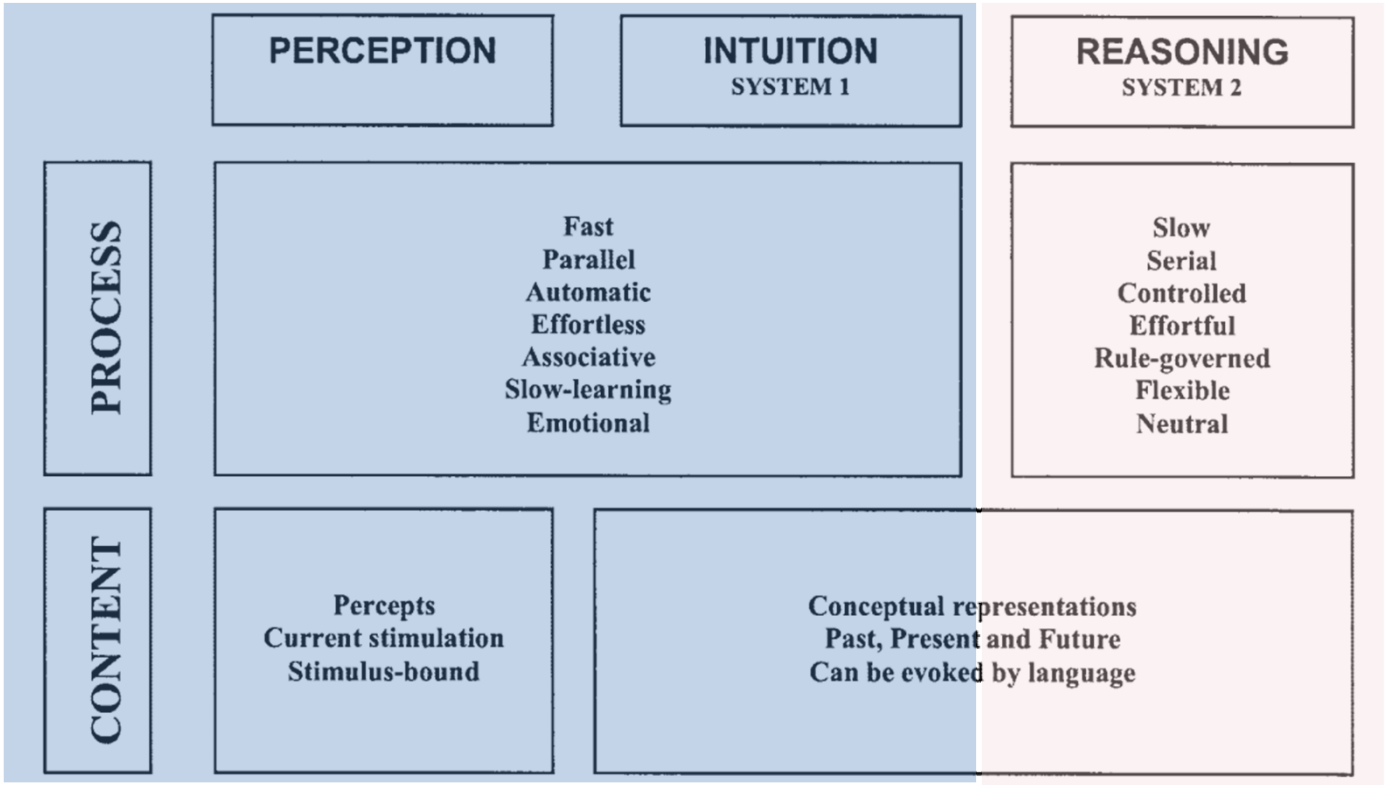
Data, Data, Data, ...

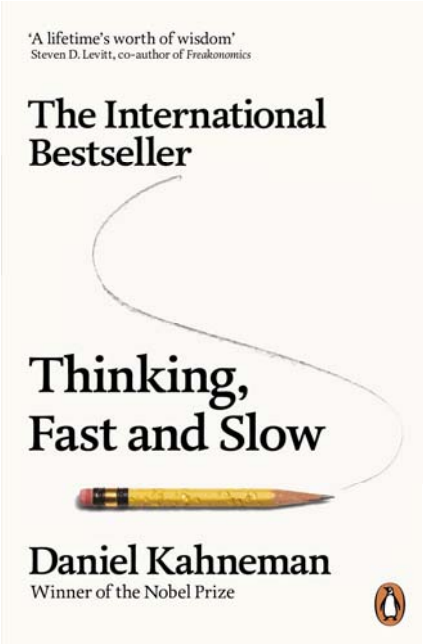


Andreas Holzinger, Matthias Dehmer & Igor Jurisica 2014. Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. *Springer/Nature BMC Bioinformatics,* 15, (S6), I1, doi:10.1186/1471-2105-15-S6-I1.

- 1) learning from few **data**
- 2) extracting **knowledge**
- 3) **generalize**
- 4) fight the curse of **dimensionality**
- 5) disentangle the **underlying explanatory factors of data**, i.e.
- 6) **causal understanding** of the data in the **context** of an application domain

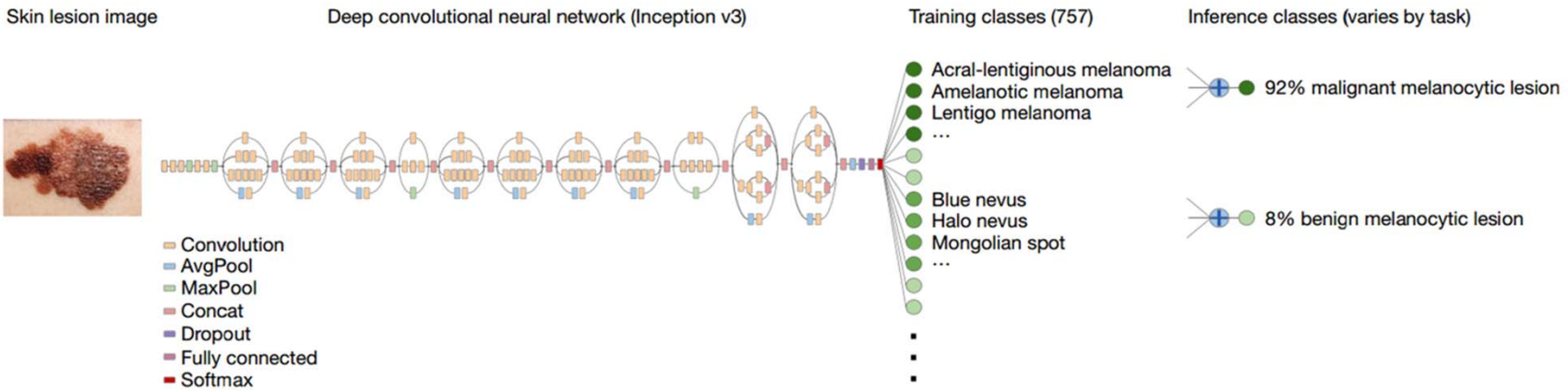
What is the current state-of-the-art of deep learning ?

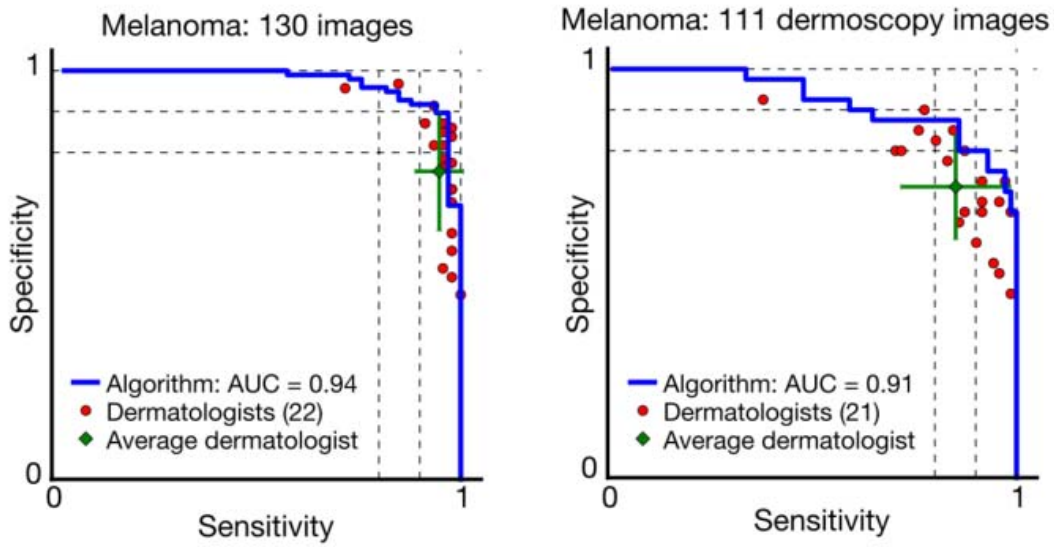




Amos Tversky & Daniel Kahneman 1974. Judgment under uncertainty: Heuristics and biases. *Science,* 185, (4157), 1124-1131, doi:10.1126/science.185.4157.1124.

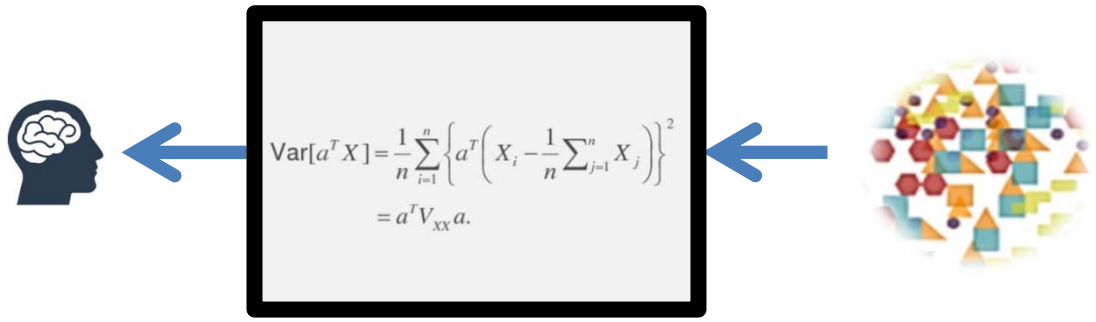
Success: Towards Human-Level AI in dermatology ...



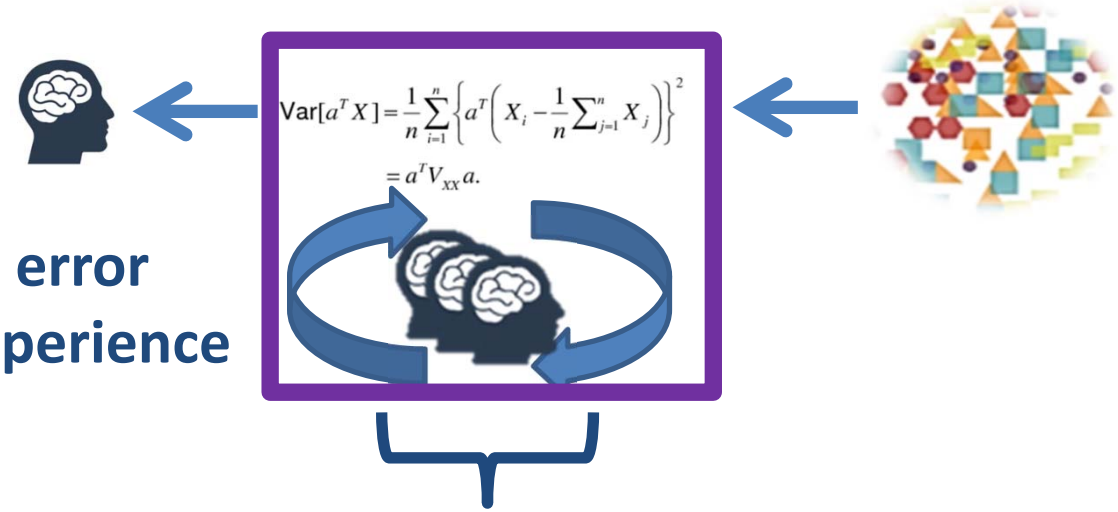


Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118, doi:10.1038/nature21056.

What should the human in the loop do ?



Generalization error



Generalization error
plus human experience

iML = human can (sometimes) bring in implicit knowledge

Andreas Holzinger et al. 2018. Interactive machine learning: experimental evidence for the human in the algorithmic loop. Springer/Nature Applied Intelligence, doi:10.1007/s10489-018-1361-5.

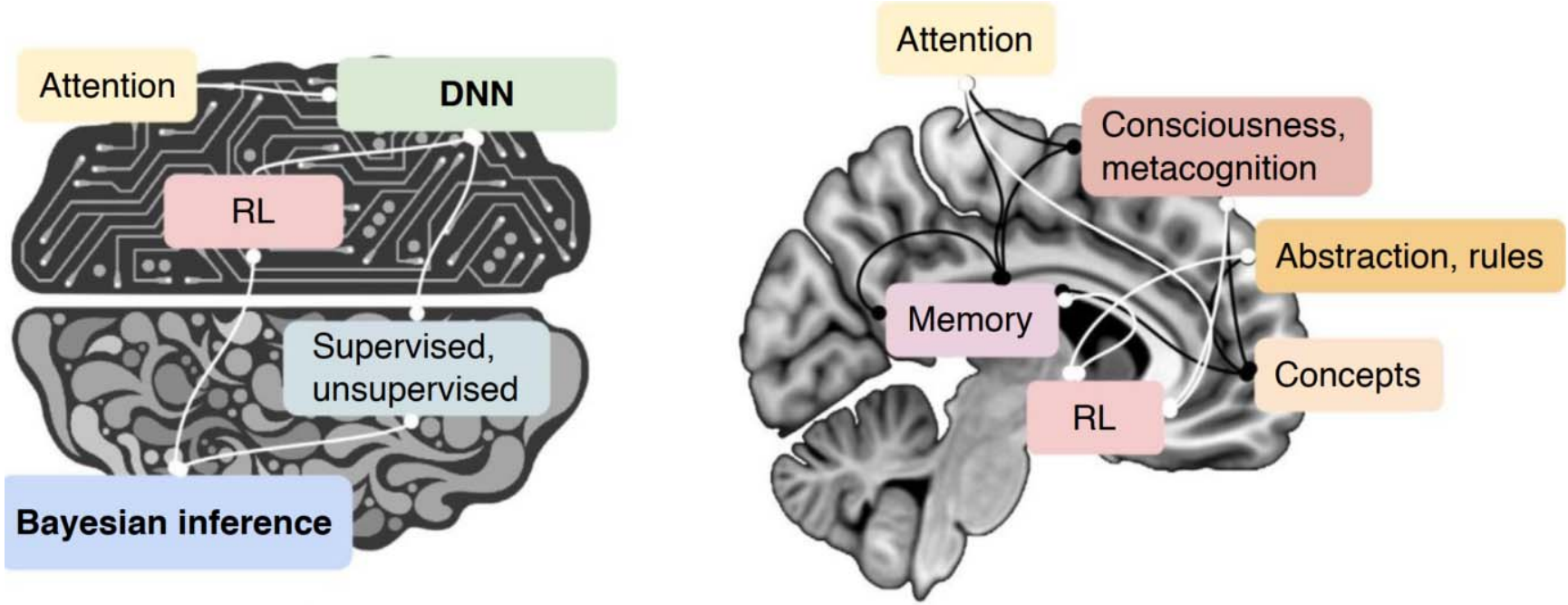
How do humans generalize from a few examples ?

(Sometimes – **not** always!) humans are able ...

- to understand the context
- to make inferences from little, noisy, incomplete data sets
- to learn relevant representations
- to find shared underlying explanatory factors,
- in particular between $P(x)$ and $P(Y|X)$, with a causal link between $Y \rightarrow X$

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths & Noah D. Goodman 2011. How to grow a mind: Statistics, structure, and abstraction. *Science,* 331, (6022), 1279-1285, doi:10.1126/science.1192788.

How can we model basic cognitive capacities as intuitive Bayes ?



Aurelio Cortese, Benedetto De Martino & Mitsuo Kawato (2019). The neural and cognitive architecture for learning from a small sample. Current opinion in neurobiology, 55, 133-141, doi:10.1016/j.conb.2019.02.011

$$P(h|x,T) = \frac{P(x|h,T)P(h|T)}{\sum_{h' \in H_T} P(x|h',T)P(h'|T)}$$

Joshua B. Tenenbaum, Thomas L. Griffiths & Charles Kemp 2006. Theory-based Bayesian models of inductive learning and reasoning. Trends in cognitive sciences, 10, (7), 309-318, doi:10.1016/j.tics.2006.05.009.

How does our mind get so much out of it ?

Salakhutdinov, R., Tenenbaum, J. & Torralba, A. 2012. One-shot learning with a hierarchical nonparametric Bayesian model. Journal of Machine Learning Research, 27, 195-207.

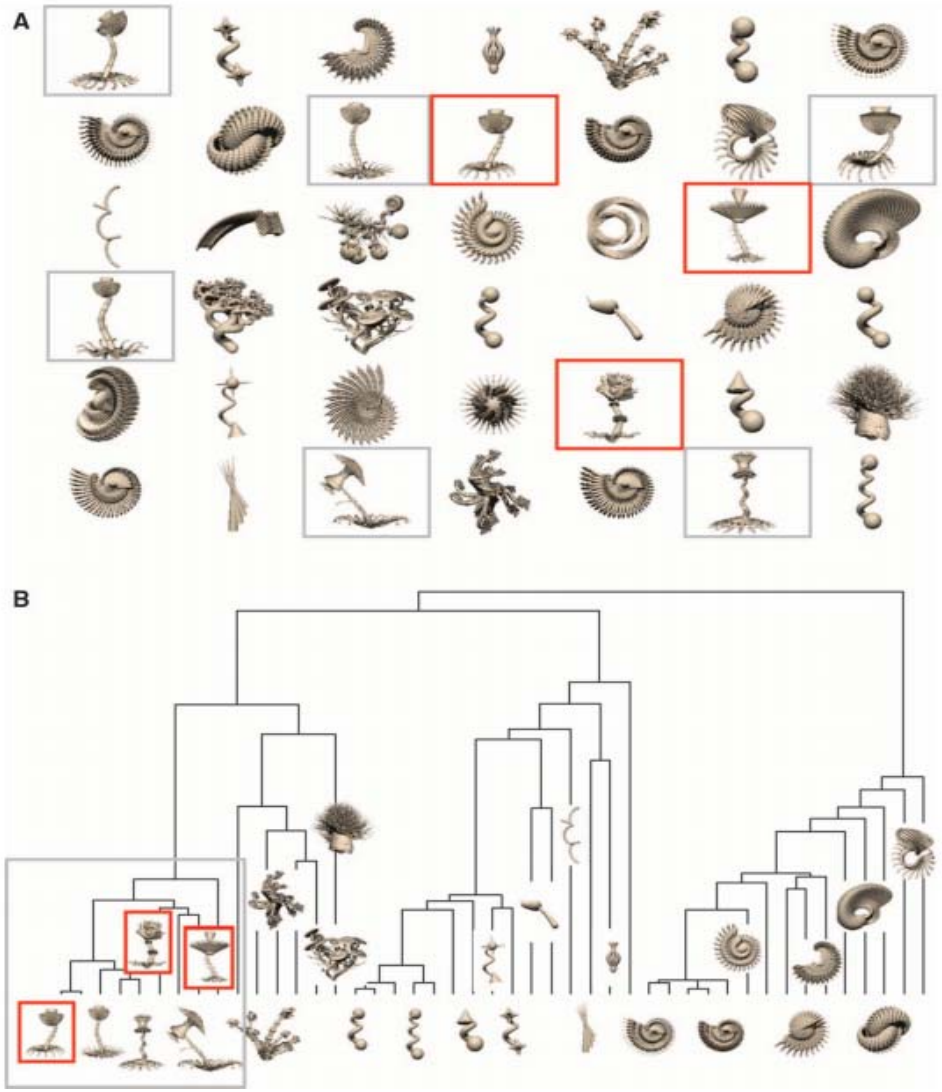
Learning words for objects – concepts from examples

Quaxl Quaxl

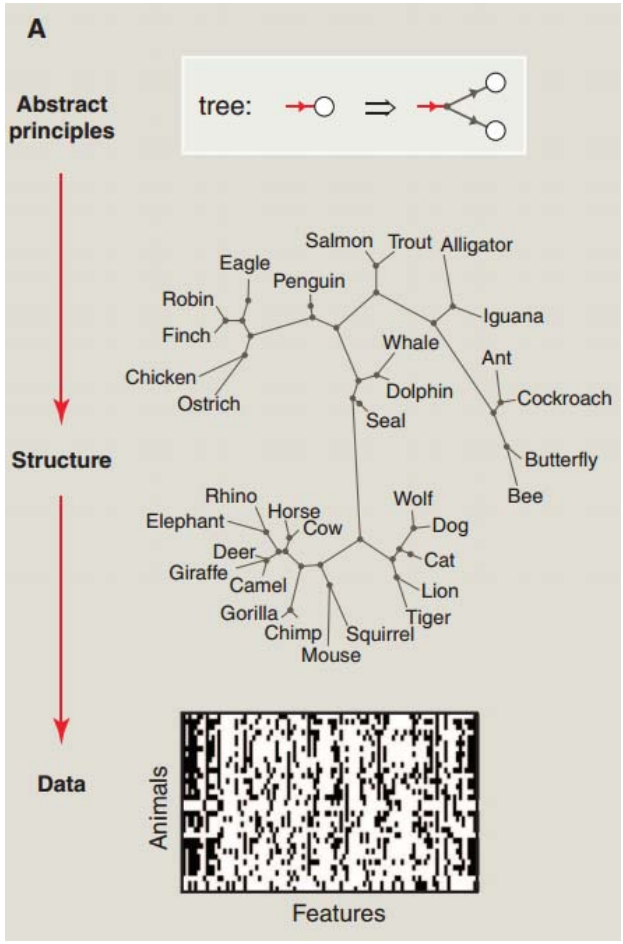
Quaxl

Salakhutdinov, R., Tenenbaum, J. & Torralba, A. 2012. One-shot learning with a hierarchical nonparametric Bayesian model. Journal of Machine Learning Research, 27, 195-207.

How do we understand our world ?

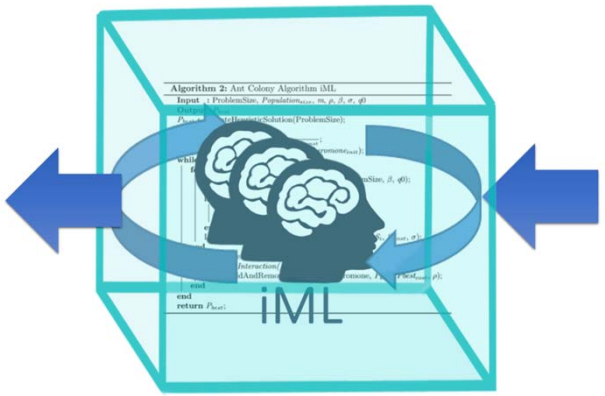


$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in H} P(d|h')P(h')} \propto P(d|h)P(h)$$

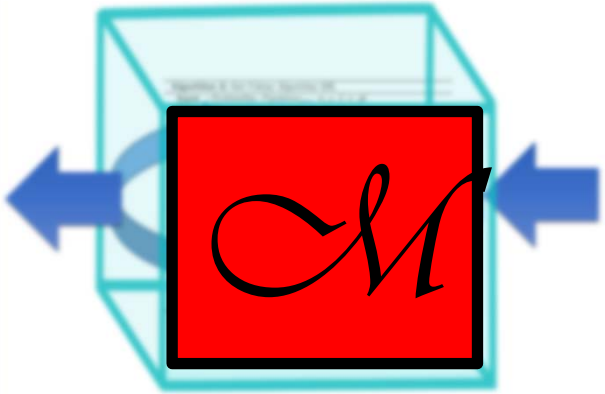


Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331, (6022), 1279-1285.

- **Interpretable Models, = ante-hoc -** the "glass-box" model itself is ante-hoc interpretable, e.g. Regression, Naïve Bayes, Decision Trees, Graphs …



- **Interpreting Black-Box Models, = post-hoc -** the model is not interpretable and needs a post-hoc interpretability method $\mathcal{M}$



Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis & Douglas B. Kell 2017. What do we need to build explainable AI systems for the medical domain? *arXiv:1712.09923*.

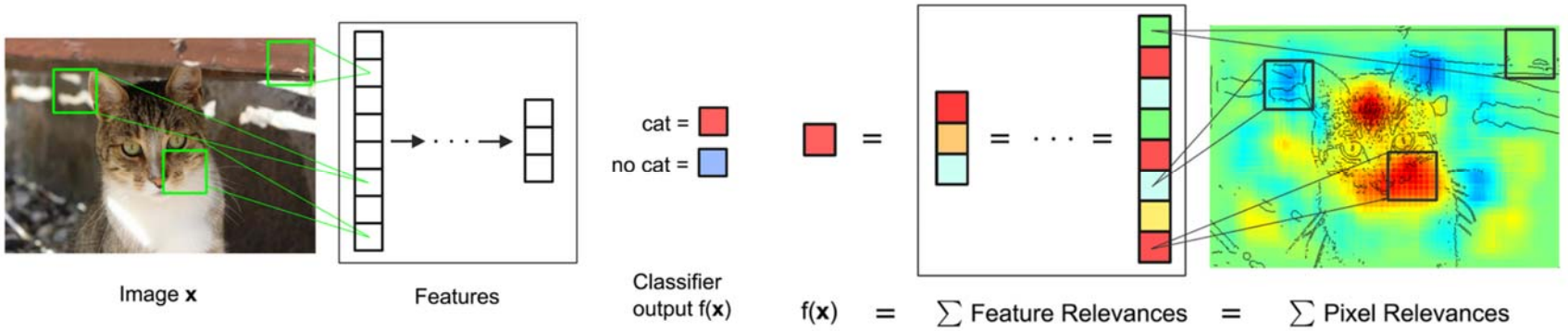
Examples of post-hoc methods of explainable AI

- 1) Gradients
- 2) Sensitivity Analysis
- 3) Decomposition Relevance Propagation (Pixel-RP, Layer-RP, Deep Taylor Decomposition, …)
- 4) Optimization (Local-IME – model agnostic, BETA transparent approximation, …)
- 5) Deconvolution and Guided Backpropagation
- 6) Model Understanding
  - Feature visualization, Inverting CNN
  - Qualitative Testing with Concept Activation Vectors TCAV
  - Network Dissection

Andreas Holzinger LV 706.315 From explainable AI to Causability, 3 ECTS course
https://human-centered.ai/explainable-ai-causability-2019 (course given since 2016)

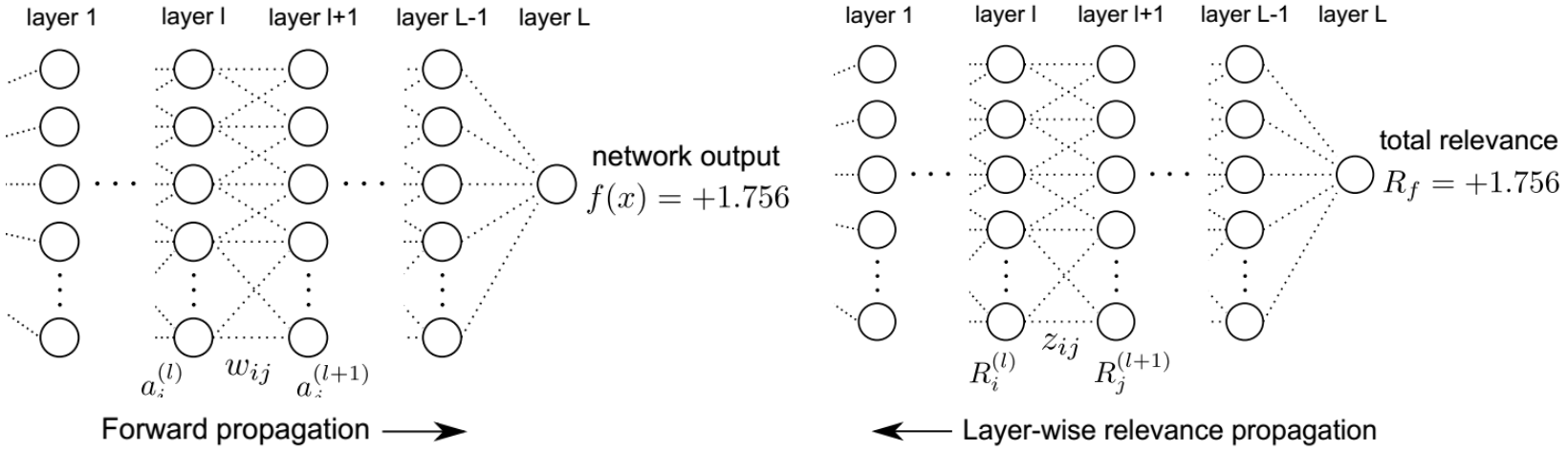
LRP Layer-Wise Relevance Propagation

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.



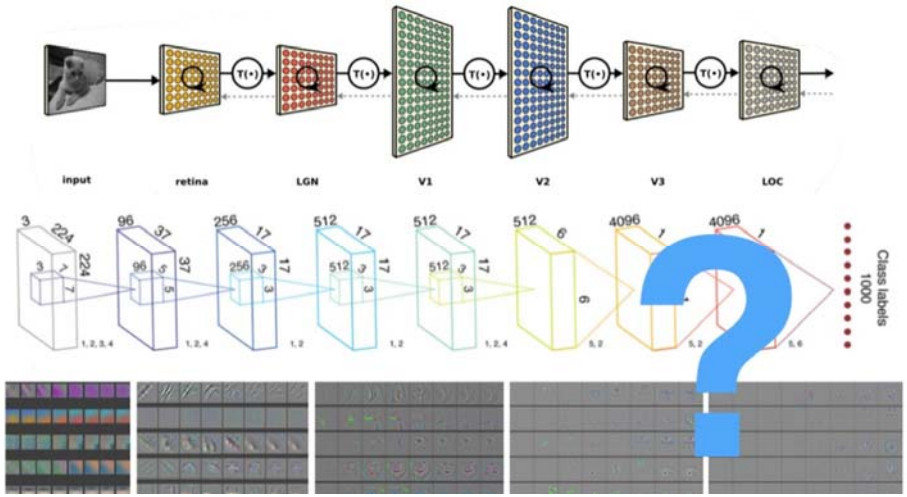
$$a_j^{(l+1)} = \sigma\left(\sum_i a_i^{(l)} w_{ij} + b_j^{(l+1)}\right)$$

$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j}} R_j^{(l+1)}$$

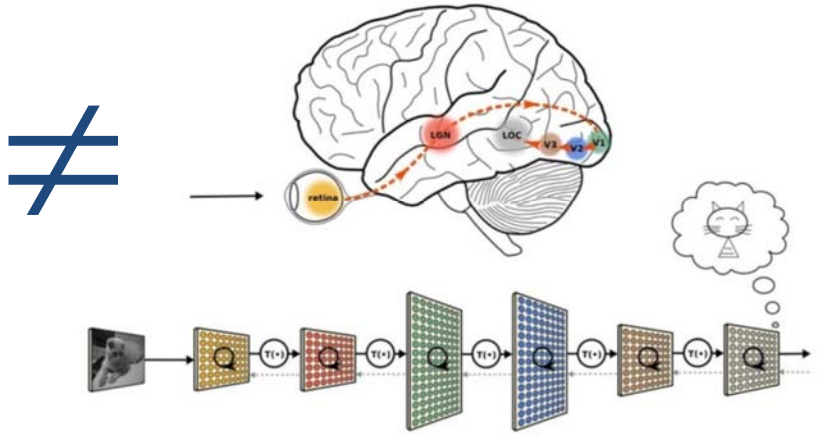


$$R_i = \left|\left|\frac{\partial}{\partial x_i} f(\mathbf{x})\right|\right| \qquad \sum_i R_i = \ldots = \sum_j R_j = \sum_k R_k = \ldots = f(\mathbf{x})$$

Testing with Concept Activation Vectors (CAV)



Yann Lecun, Yoshua Bengio & Geoffrey Hinton 2015. Deep learning. Nature, 521, (7553), 436-444, doi:10.1038/nature14539.



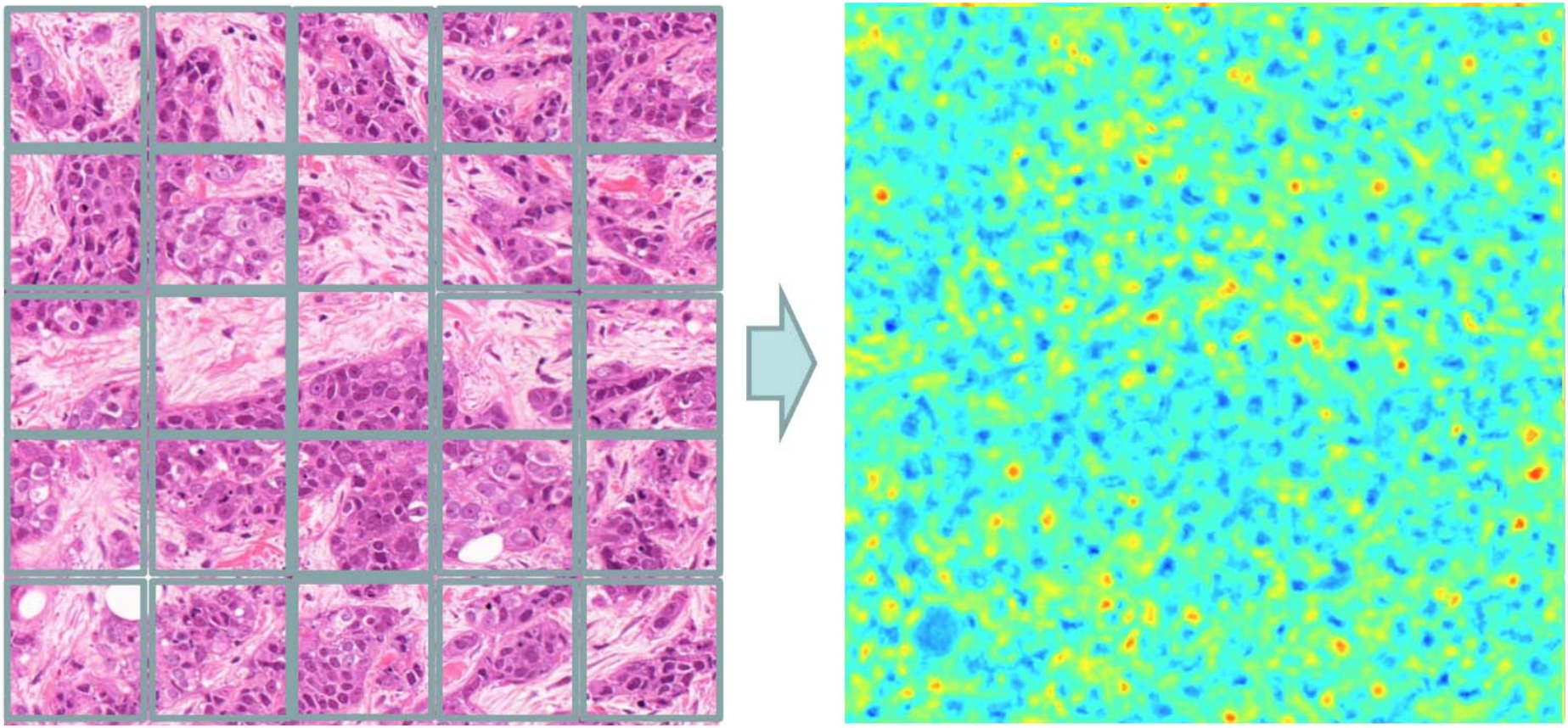
$$\frac{\partial h_k(x)}{\partial x_{a,b}}$$

Humans work in another vector space which is spanned by **implicit knowledge** vectors corresponding to an unknown set of human interpretable concepts.

$$S_{C,k,l}(x) = \lim_{\epsilon \to 0} \frac{h_{l,k}\left(f_l(x) + \epsilon v_C^l\right) - h_{l,k}\left(f_l(x)\right)}{\epsilon} = \nabla h_{l,k}\left(f_l(x)\right) \cdot v_C^l$$

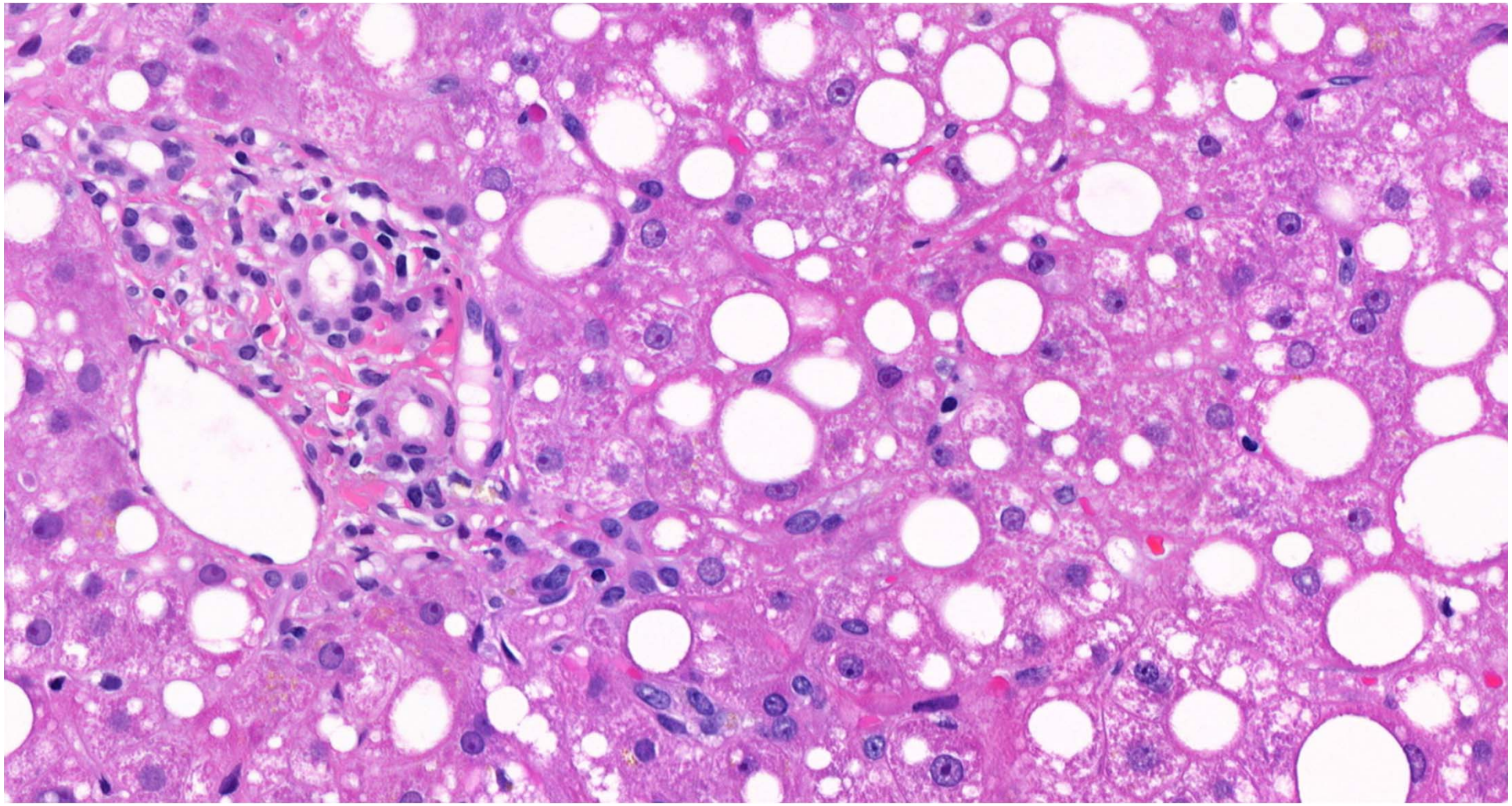
Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler & Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors, ICML, 2018. 2673-2682.

Example Histopathology



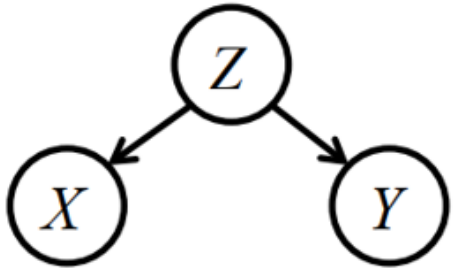
Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

How do human pathologists make diagnoses ?



What is ground truth ? Where is the ground truth ?

- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
  - Empirical evidence = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
  - Empirical inference = drawing conclusions from empirical data (observations, measurements)
  - Causal inference = drawing conclusions about a causal connection based on the conditions of the occurrence of an effect.

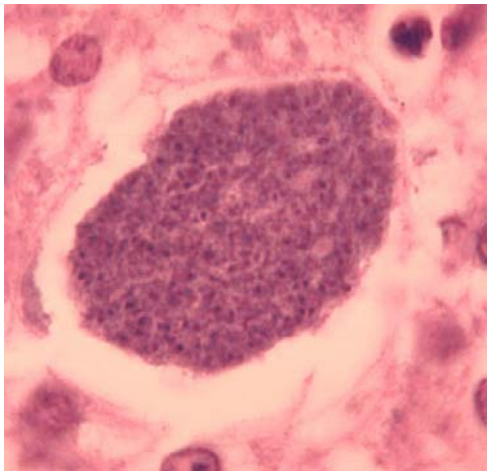


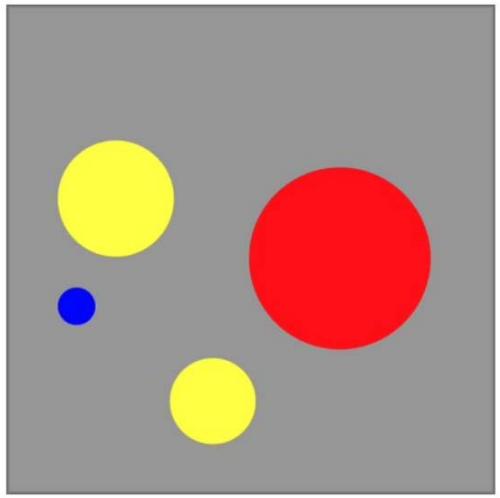
There is no correlation without causation

Hans Reichenbach (1956). The direction of time. New York: Dover.

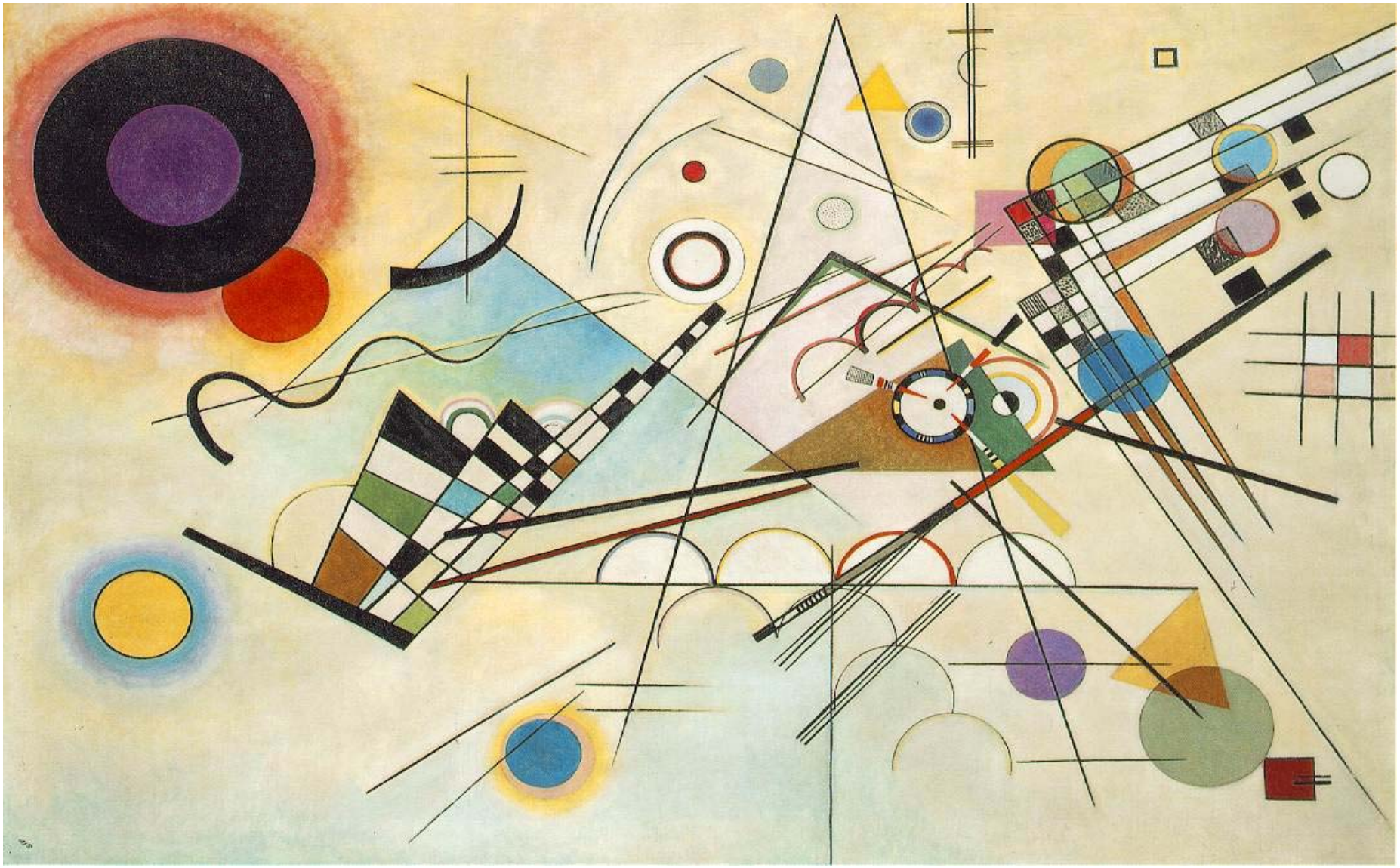
What problems do we face in the real (medical) world ?

- 1) ground truth is not always well defined, especially when making a medical diagnosis;
- 2) although human (scientific) models are often based on understanding causal mechanisms,
- today's successful machine models or algorithms are typically based on correlation or related concepts of similarity and distance!



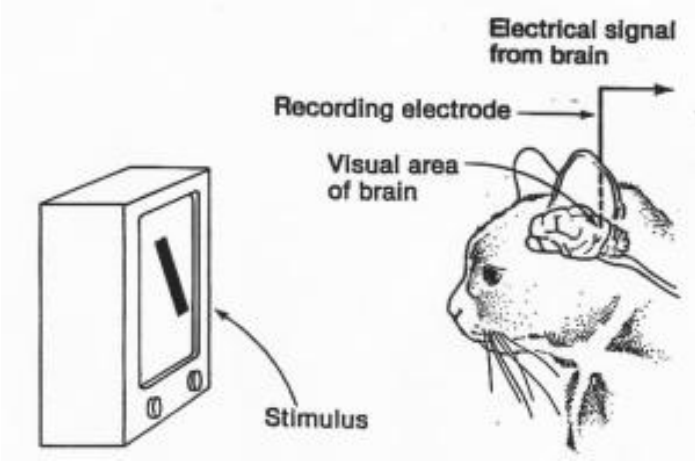


Wassily Kandinsky (1866 – 1944)

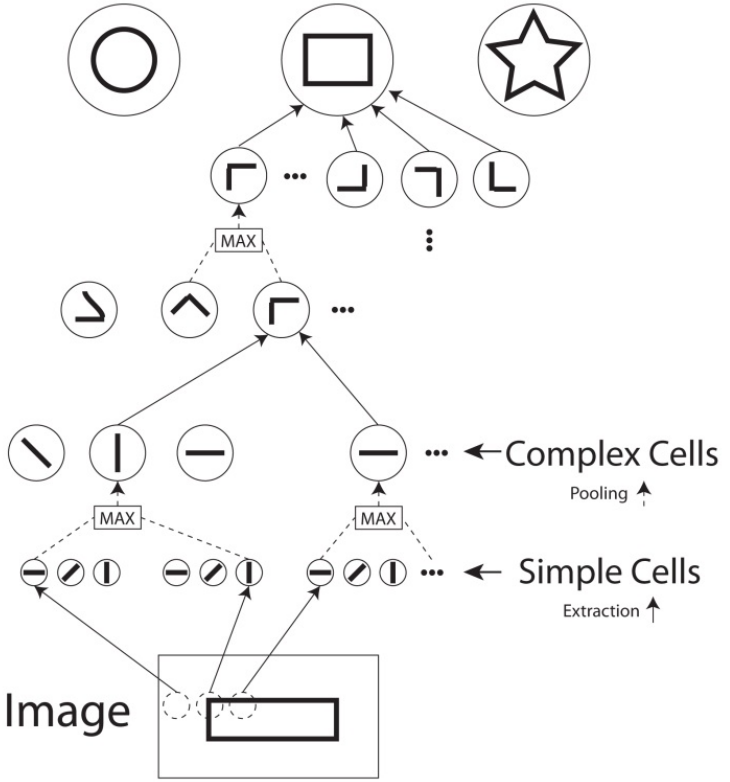


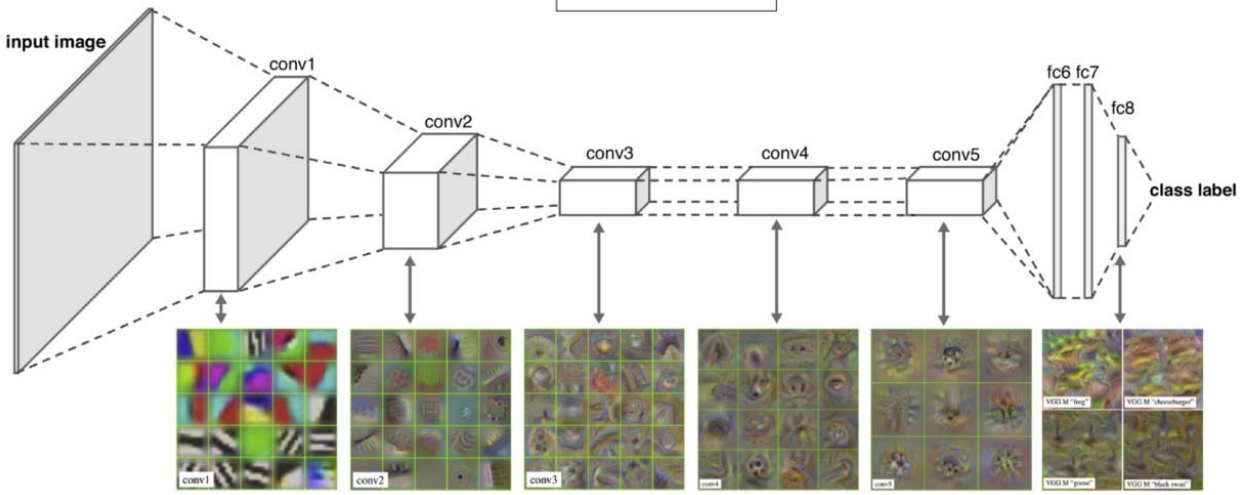
Komposition VIII, 1923, Solomon R. Guggenheim Museum, New York.Source: https://de.wikipedia.org/wiki/Wassily_Kandinsky
This images are in the public domain.

Hubel & Wiesel (1962): Our world is compositional !

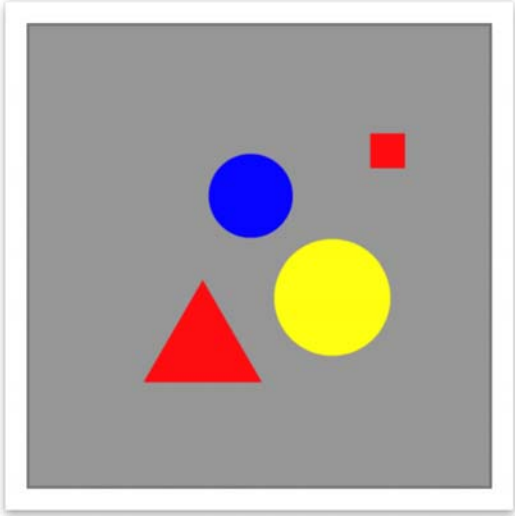


David H. Hubel & Torsten N. Wiesel 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160, (1), 106-154, doi:10.1113/jphysiol.1962.sp006837





Definition 1: A Kandinsky Figure is …



- … a square image containing 1 to n geometric objects.
- Each object is characterized by its shape, color, size and position within this square.
- Objects do not overlap and are not cropped at the border.
- All objects must be easily recognizable and clearly distinguishable by a human observer.

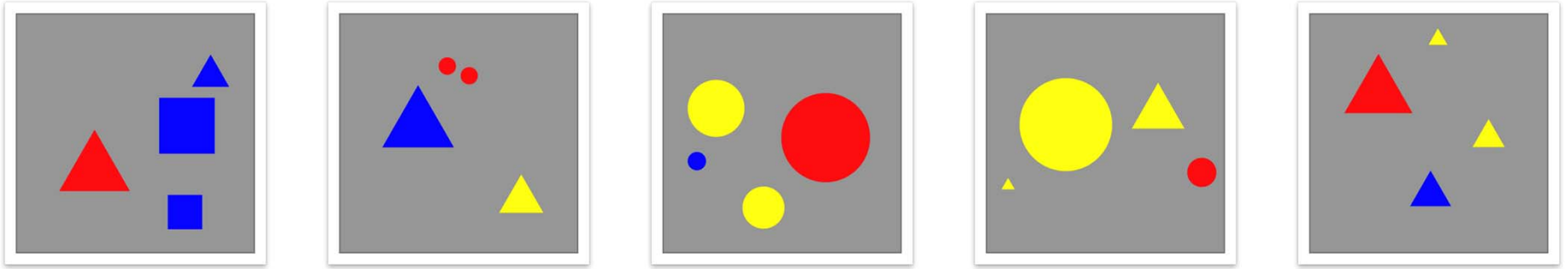
Definition 2 A statement $s(k)$

- about a Kandinsky Figure k is …
- either a mathematical function $s(k) \rightarrow B$; with B $(0,1)$
- or a $natural\ language\ statement$ which is true or false

- Remark: The evaluation of a natural language statement is always done in a specific context. In the followings examples we use **well known concepts from human perception** and linguistic theory.
- If $s(k)$ is given as an algorithm, it is essential that the function is a pure function, which is a computational analogue of a mathematical function.

Definition 3 A Kandinsky Pattern K …

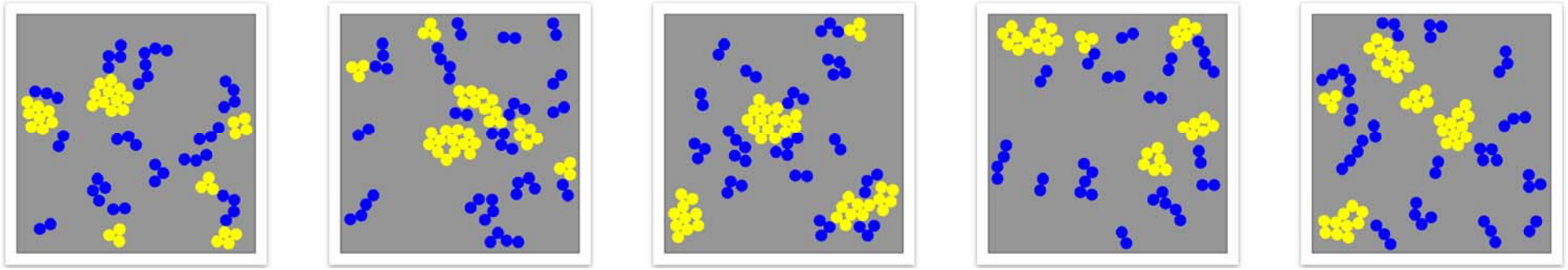
- … is defined as the subset of all possible Kandinsky Figures k with $s(k) \rightarrow 1$ or the natural language statement is true.
- $s(k)$ and a natural language statement are equivalent, if and only if the resulting Kandinsky Patterns contains the same Kandinsky Figures.
- $s(k)$ and the natural language statement are defined as the **Ground Truth** of a Kandinsky Pattern



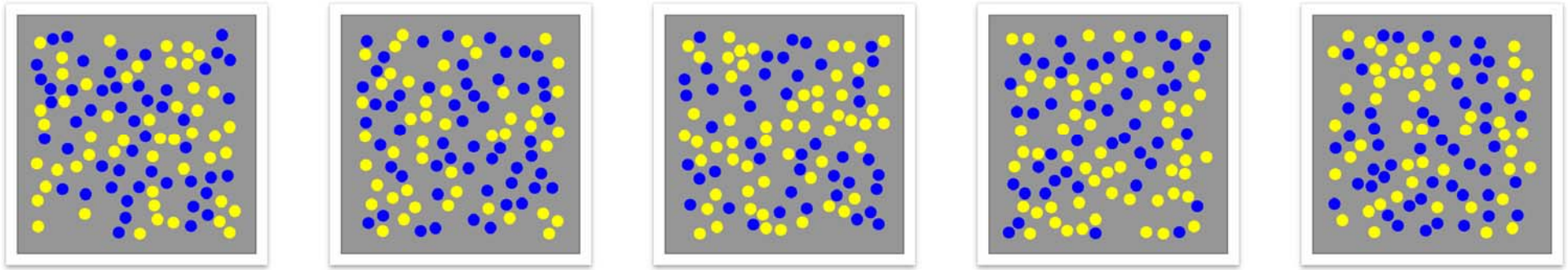
"… the Kandinsky Pattern has two pairs of objects with the same shape, in one pair the objects have the same color, in the other pair different colors, two pairs are always disjunct, i.e. they don't share a object …".

What are domain concepts from histopathology ?

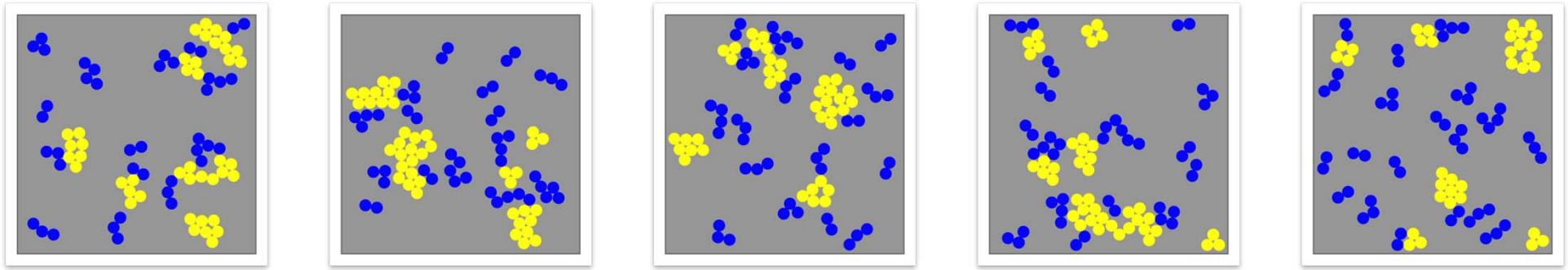
A) True (the cells are smaller and closer together – it is an tumor ...)

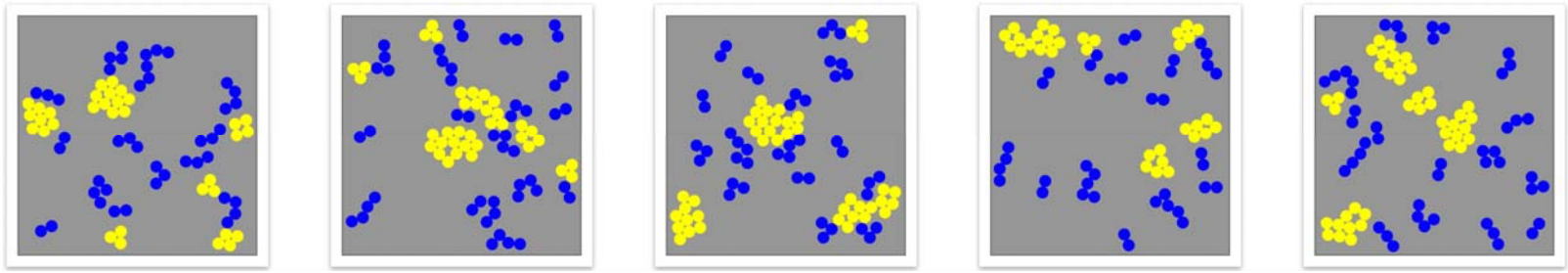


B) False



C) Counterfactual (What if the cells are slightly bigger ?)





Task 1: Explain KandinskyPatterns algorithmically, i.e. train a network which classifies Kandinsky Figures according to the ground truth …

Task 2: Explain the Kandinsky Pattern in natural language, that a human can understand …

Causability := provide the underlying explanatory factors of WHY the 92 % are true or the 8 % are false

Causality:
The art and science of cause and effect

Judea Pearl 2000. Causality: Models, Reasoning, and Inference, Cambridge: Cambridge University Press.

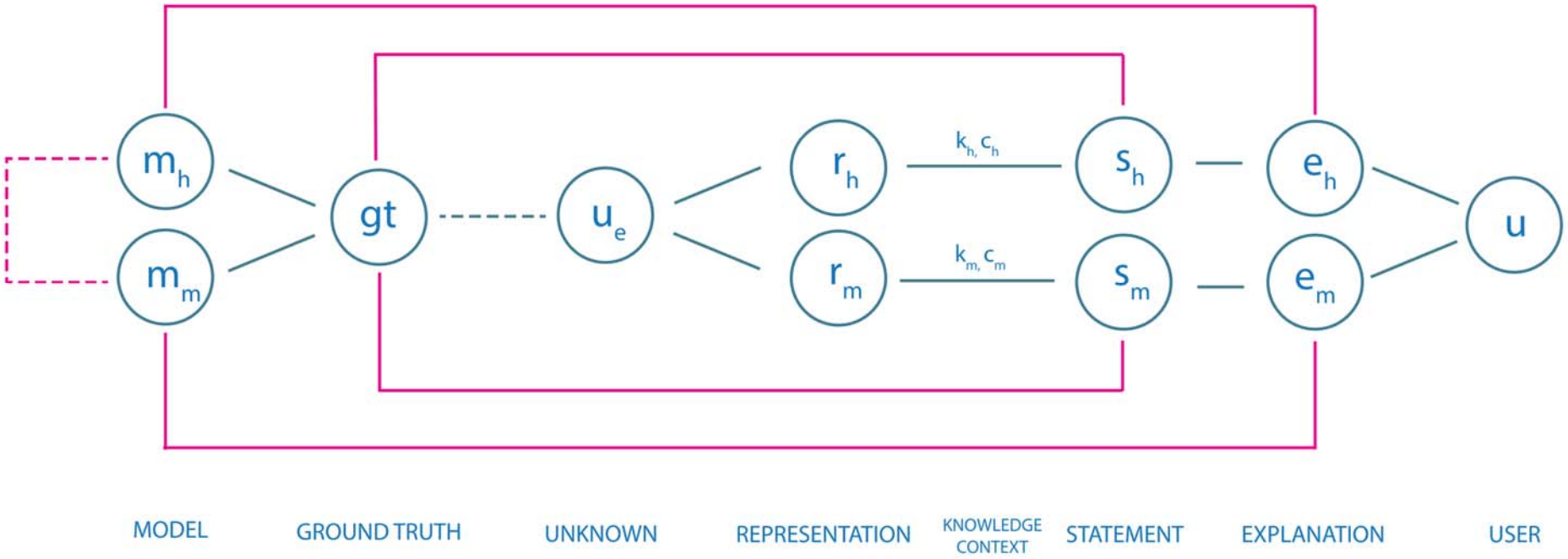
Causability: Mapping machine explanations with human understanding

Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal & Heimo Mueller 2019. Causability and Explainability of Artificial Intelligence in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9, (4), doi:10.1002/widm.1312.

Measuring the quality of Explanations: The Systems Causability Scale

Andreas Holzinger, Andre Carrington & Heimo Müller 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. KI - Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt, 34, (2), doi:10.1007/s13218-020-00636-z

Metrics of AI - System Causability Scale (SCS)



Andreas Holzinger, Andre Carrington & Heimo Müller 2020. Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. KI - Künstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt, 34, (2), doi:10.1007/s13218-020-00636-z.

what - to whom - how

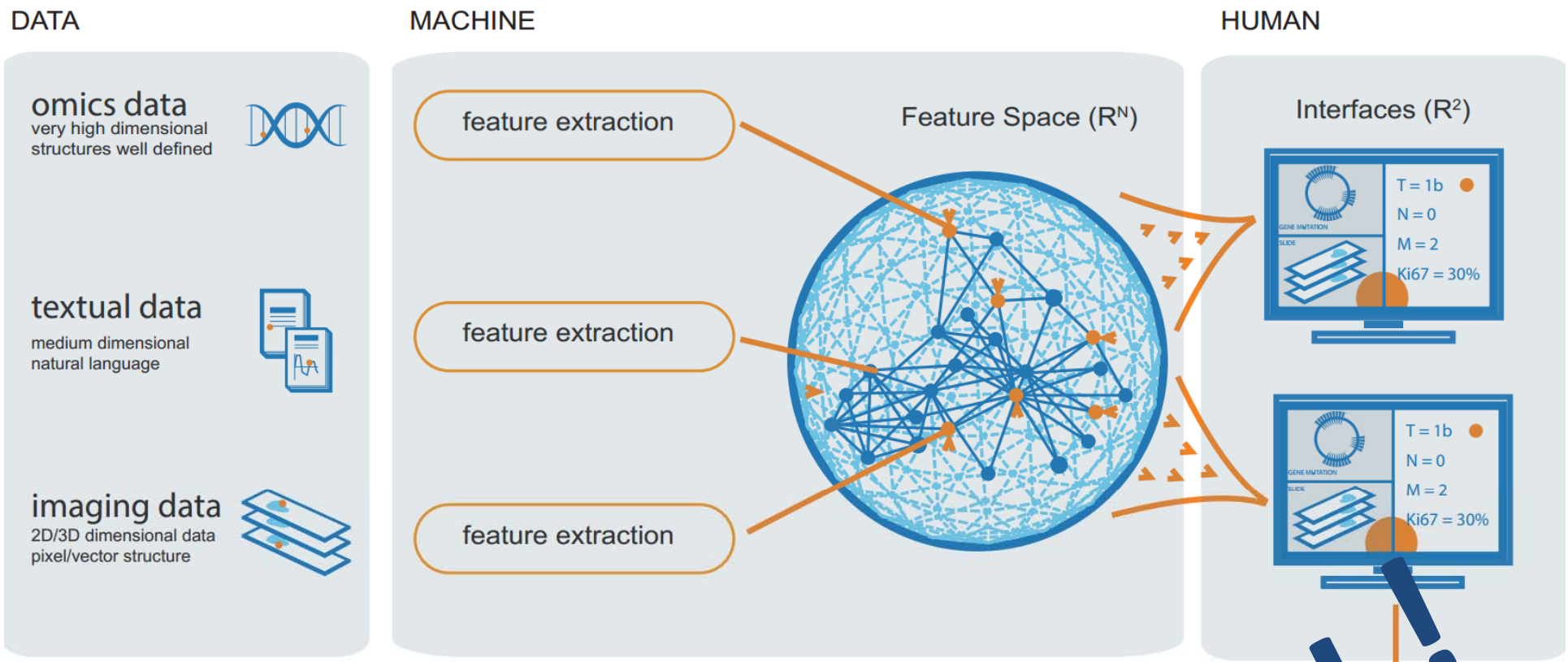
What is my take-home message to you ?

- Current AI does not generalize well,
- can not learn from few examples,
- can not infer causal relationships.

We need robust and interpretable AI …

- to reduce costs and limitations,
- to get causal explanations (why? - what if?)
- to understand machine decisions …

Multimodal Causability enabling why and what-if ...



Thank you !