Explainable AI
– recent results and open problems

Lars Kai Hansen

DTU Compute, Technical University of Denmark

Co-author Laura Rieger

Outline

Motivation for opening the black box

- Trust, debugging, legal, scientific applications
- Explanation as an ill-posed task
- **Objectives viz. Explainable Expert Systems**

Function level visualization

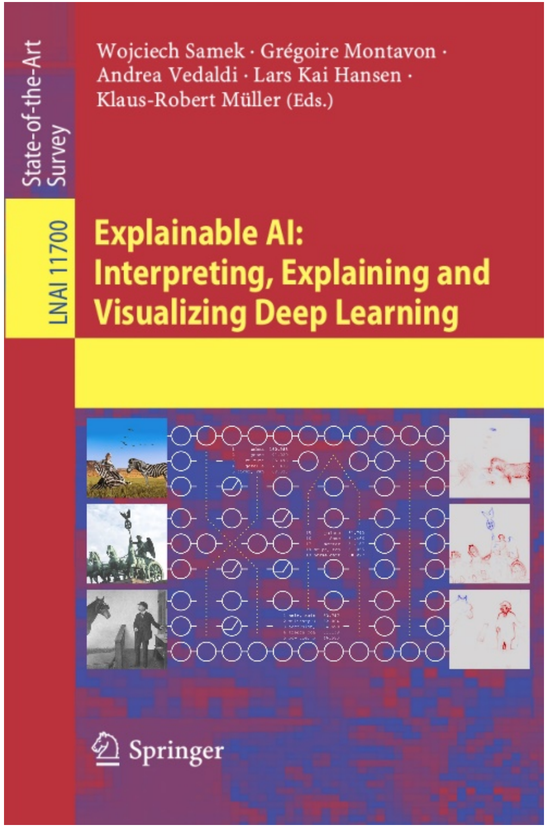
- Robustness vs methods, networks, training sets
- **Uncertainty quantification**

Decision explanations

- New result: Evaluation by simple counterfactuals
- New result: Better performance by model averaging
- New result: Resilience to “fairwashing” through model averaing

Open problems

- Evaluation?
- Human in the loop?
- Causal modelling?



Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K. and Müller, K.R. eds., 2019. *Explainable AI: interpreting, explaining and visualizing deep learning* Vol. 11700. Springer Nature.

Opening the black box - motivations

Trust & debugging

AI as a collaborator / teacher – **AI social competences**

Verification, performance optimization…

Align values – fairness, reduce biases, adversarial risks ...

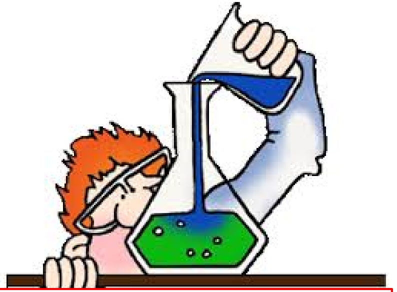
Legal requirements - *"right to explanation"*

General data protection regulatory May 26, 2018, DPOs

Scientific applications of machine learning

learning from machine learning solutions,

causal mechanisms,



Explanation is an (interesting) ill-posed task

Existence? - Unclear objectives, no canonical evaluation metrics

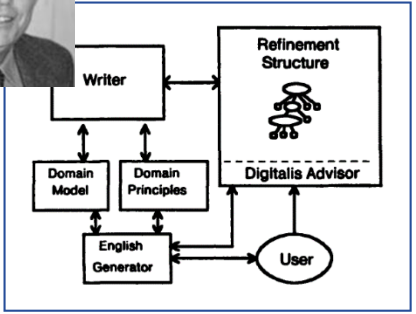
Uniqueness? – model uncertainty, robustness

Goodman, B. and Flaxman, S., 2016. European Union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*.
Wachter, S., Mittelstadt, B. and Floridi, L., 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law, 7(2), pp.76-99.

Explainability - objectives

Second generation AI

Swartout and Moore (1993)



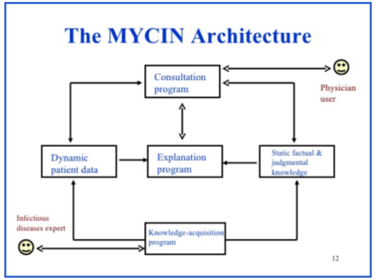
XPLAIN

Fidelity
The explanation must be a reasonable representation of what the system actually does.

Understandability
Involves multiple usability factors including terminology, user competencies, levels of abstraction and interactivity.

Sufficiency
Should be able to explain function and terminology and be detailed enough to justify decision (causal explanations)

Low Construction overhead & **Efficiency**:
The explanation should not dominate the cost of designing the AI.
The explanation system should not slow down the AI significantly.

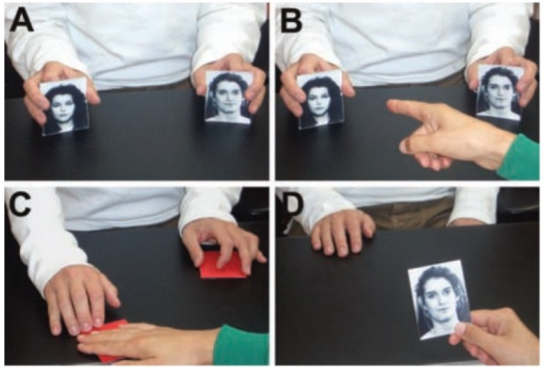


Swartout, W. R. and Moore, J. D. 1993. Explanation in second generation expert systems. In Second generation expert systems, pages 543–585. Springer.
Shortliffe, E.H. et al., 1975. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Computers and biomedical research, 8(4), pp.303-320. (antibiotics administration)
Swartout, W.R., 1983. Xplain: A system for creating and explaining expert consulting programs (No. ISI/RS-83-4). (digitalis therapy heart issues)

Can we trust human explanations?- *"choice blindness"*

Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task

Petter Johansson,[1*] Lars Hall,[1*†] Sverker Sikström,[1] Andreas Olsson[2]

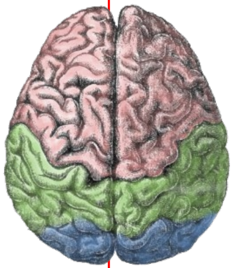


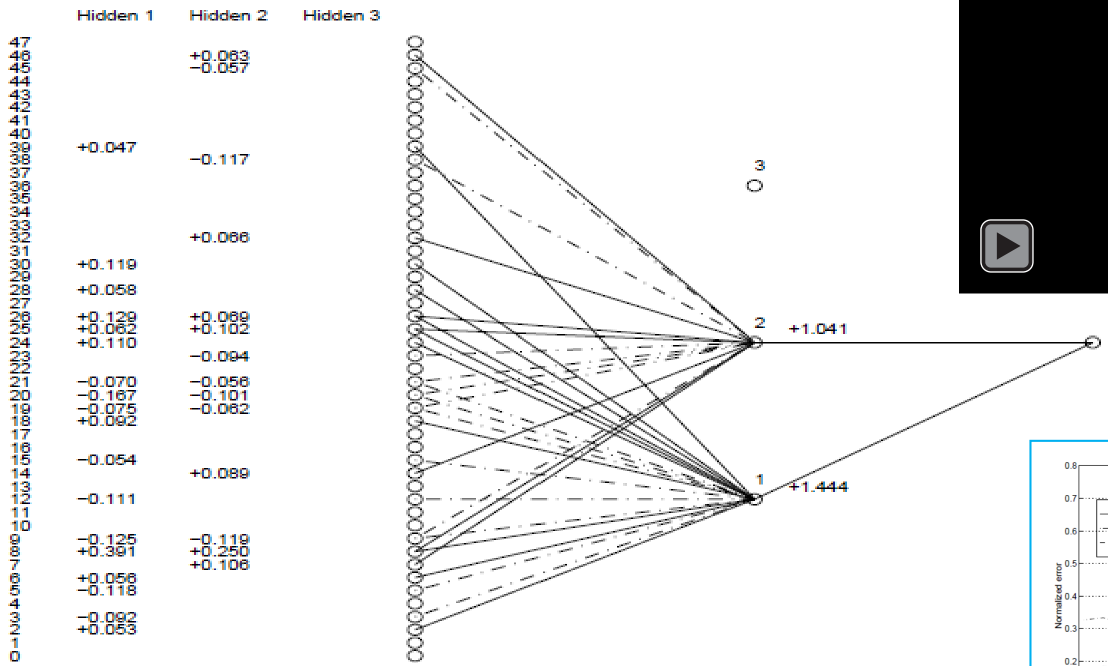
"Even when they were given unlimited time to deliberate upon their choice no more than 30% of all manipulated trials were detected.
But not only were the participants often blind to the manipulation of their choices, they also offered introspectively derived reasons for preferring the alternative they were given instead.

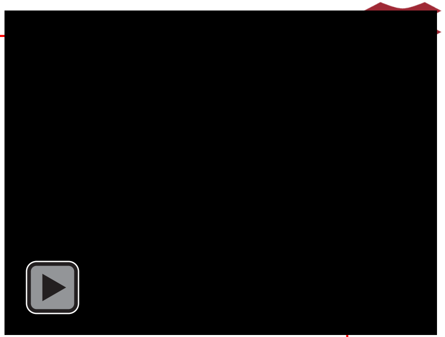
In addition to this, manipulated and non-manipulated reports were compared on a number of different dimensions, such as the level of emotionality, specificity and certainty expressed, but no substantial differences were found"

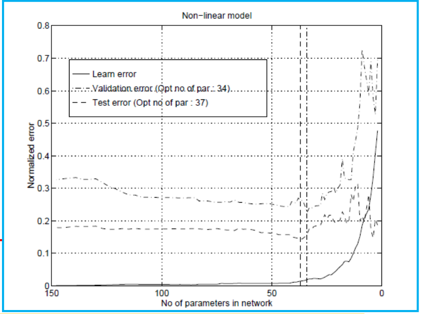
Johansson, P., Hall, L., Sikström, S. and Olsson, A., 2005. Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), pp.116-119.
Johansson, P., Hall, L., Sikström, S., 2008. From change blindness to choice blindness. Psychologia, 51(2), pp.142-155.

Saliency map for a neural network for decoding PET brain scans (1994-95)









LeCun, Y., Denker, J.S. and Solla, S.A., 1990. Optimal brain damage. In Advances in neural information processing systems (pp. 598-605).
Lautrup, B, Hansen, LK, Law, I., Mørch, N, Svarer, C, Strother, S Massive weight sharing: a cure for extremely ill-posed problems.
In *Workshop on supercomputing in brain research: From tomography to neural networks*. 137-144 (1994).
Mørch N, Kjems U, Hansen LK, Svarer C, Law I, Lautrup B, Strother S: Visualization of Neural Networks Using Saliency Maps.
In Proc. 1995 IEEE International Conference on Neural Networks, Perth, Australia, (2):2085-2090 (1995).

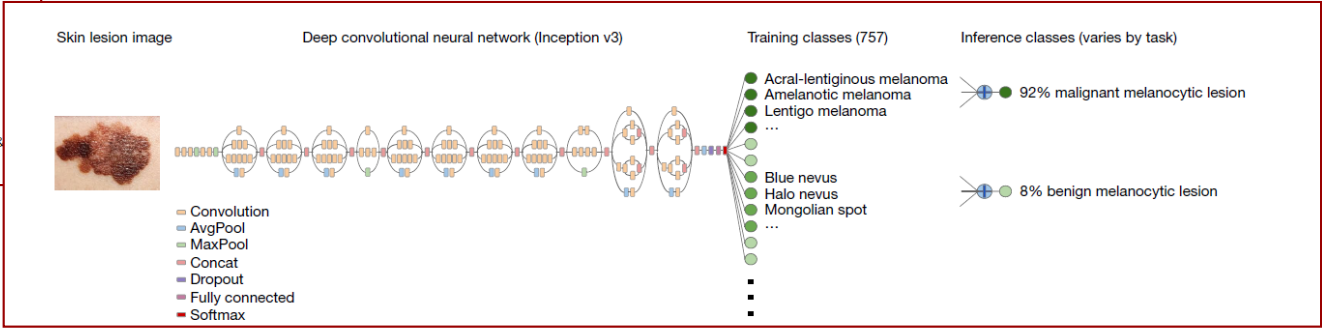
Dermatologist-level classification of skin cancer with deep neural networks

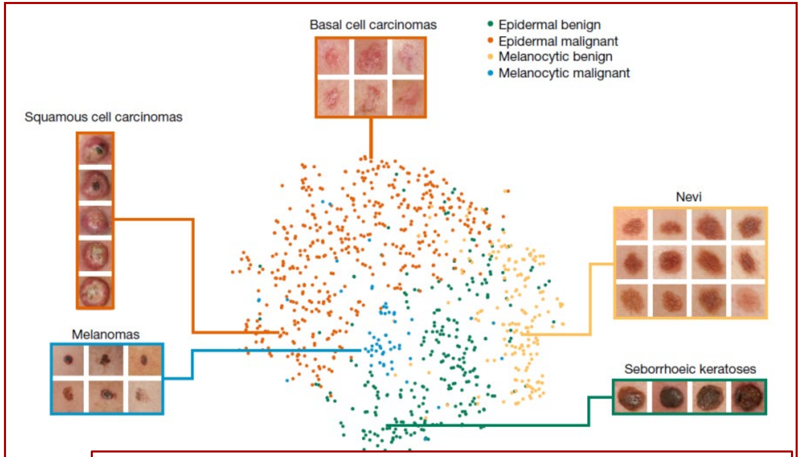
nature
International journal of science

Letter | Published: 25 January 2017

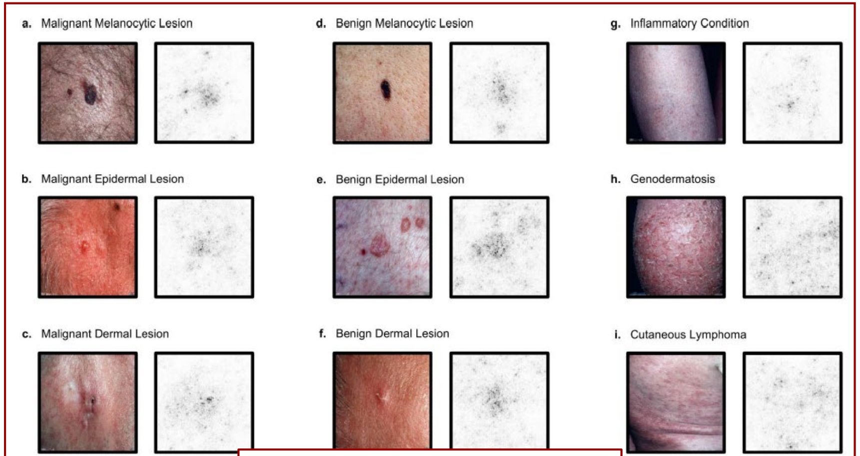
Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun





t-Distributed Stochastic Neighbor Embedding (*t-SNE*) plot of embedding



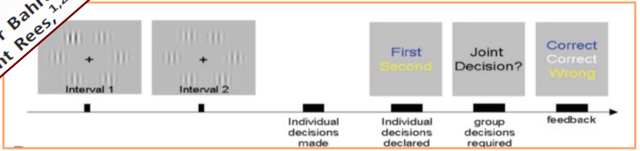
L1 sensitivity map

Inspiration from cognitive science:
Communicating uncertainty improves group inference

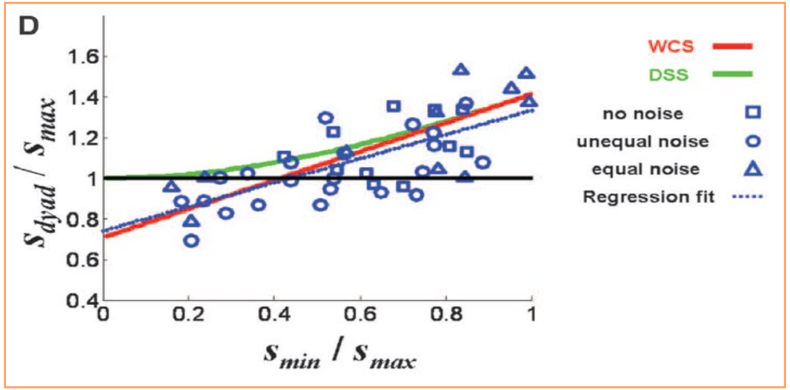
Optimally Interacting Minds

Bahador Bahrami,[1,2,3*] Karsten Olsen,[3] Peter E. Latham,[4] Andreas Roepstor...
Geraint Rees,[1,2] Chris D. Frith[2,3]

"To come to an optimal joint decision, individuals must share information with each other and, importantly, weigh that information by its reliability…"



Dyad / best participant



Ratio of participant detection "slopes"

For interactive decisions …
communication of internal uncertainty helps: "dyad benefit"

Bahrami B, Olsen K, Latham PE, Roepstorff A, Rees G, Frith CD. Optimally interacting minds. Science. 2010 Aug 27;329(5995):1081-5.
Navajas, J., Niella, T., Garbulsky, G., Bahrami, B. and Sigman, M., 2017. Deliberation increases the wisdom of crowds. *arXiv preprint arXiv:1703.00045*

NPAIRS: Sensitivity map w/ uncertainty estimates

NeuroImage **15**, 772–786 (2002)
doi:10.1006/nimg.2001.1033, available online at http://www.idealibrary.com on IDEAL®

The Quantitative Evaluation of Functional Neuroimaging Experiments: Mutual Information Learning Curves

U. Kjems,*[,1] L. K. Hansen,* J. Anderson,†,‡ S. Frutiger,‡,§ S. Muley,§
J. Sidtis,§ D. Rottenberg,†,‡,§ and S. C. Strother†,‡,§,¶

*Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark; †Radiology Department, §Neurology Department, and ¶Biomedical Engineering, University of Minnesota, Minneapolis, Minnesota 55455; and ‡PET Imaging Center, VA Medical Center, Minneapolis, Minnesota 55417

$$m_j = \left\langle \left(\frac{\partial \log p(s|x)}{\partial x_j} \right)^2 \right\rangle$$

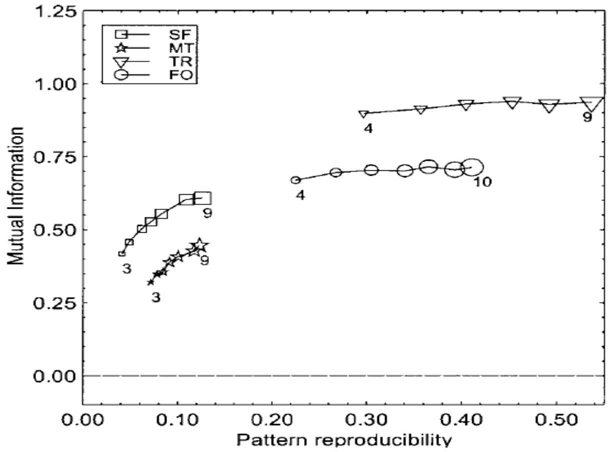
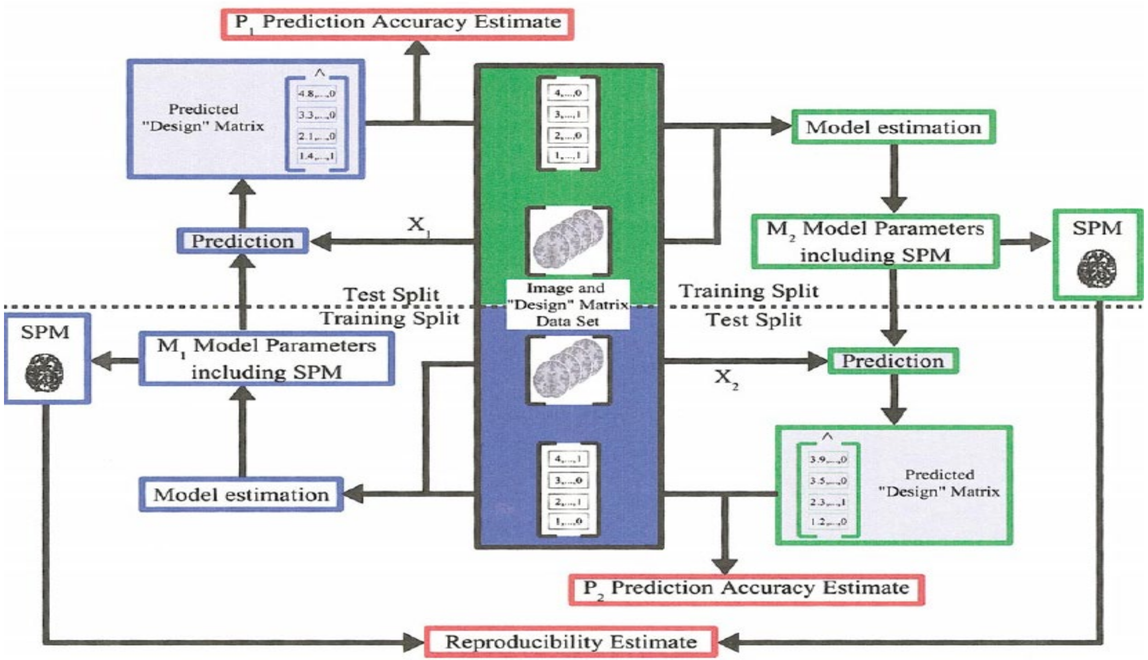


FIG. 3. Plot of scan/label mutual information versus reproducibility signal/noise for the four data sets, for varying numbers of subjects in the training set. There were 2 labels/4 scans per subject (balanced data set; Setup 1, Table 1) corresponding to the dashed solid line in Fig. 4. We see that both measures indicate improved performance of the model as the number of subjects increases.

The sensitivity map measures the impact that a given feature has on the predictive distribution

Zurada, J.M., Malinowski, A. and Cloete, I., 1994, June. Sensitivity analysis for minimization of input data dimension for feedforward neural network. In Circuits and Systems, 1994. ISCAS'94., 1994 IEEE International Symposium on (Vol. 6, pp. 447-450). IEEE.

NPAIRS Workflow: Performance and reproducibility estimates



NeuroImage: Hansen et al (1999), Lange et al. (1999), Hansen et al (2000), Strother et al (2002), Kjems et al. (2002), LaConte et al (2003), Strother et al (2004), Mondrup et al (2011), Andersen et al (2014)
Brain and Language: Hansen (2007)

Detection of Skin Cancer by Classification of Raman Spectra

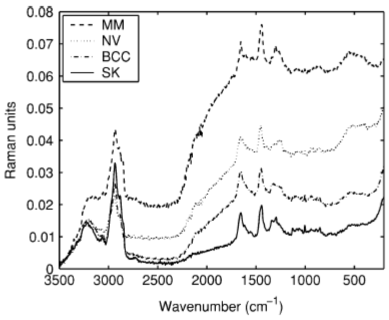
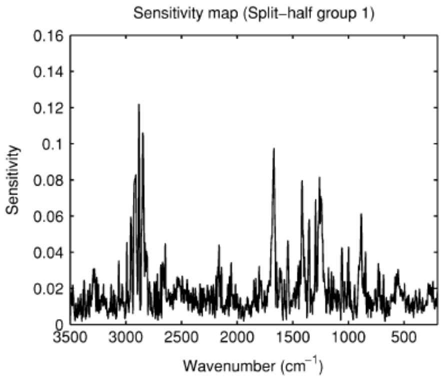
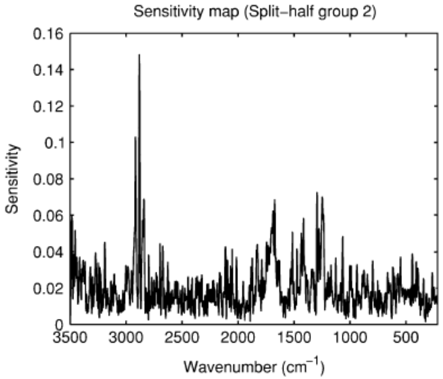


Fig. 1. Examples of the NIR-FT Raman spectra of benign and malignant skin lesions and tumors: BCC, MM, NV, and SK.

	BCC	MM	NOR	NV	SK
BCC*	95.8	10.0	1.1	0.0	0.9
MM*	0.0	80.5	0.0	2.4	0.0
NOR*	0.0	4.8	97.8	5.4	0.0
NV*	2.1	4.8	1.1	92.2	0.0
SK*	2.1	0.0	0.0	0.0	99.1





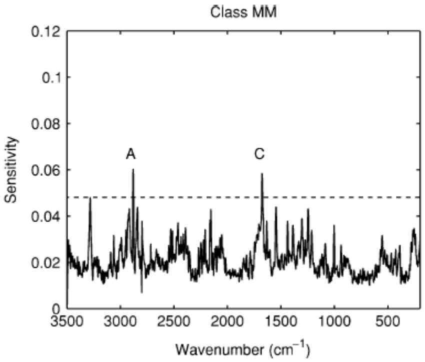


Fig. 10. Sensitivity maps for the MM class. Dashed line indicates 95% confidence interval. Sensitivity map seems more noisy than the BCC sensitivity map in Fig. 9. Region marked A represents the CH^{-} vibrations in the lipids and proteins around 2940 cm^{-1} and region marked C reflects the amide I band of proteins 1600–1800 cm^{-1}.

Sigurdsson, S., Philipsen, P.A., Hansen, L.K., Larsen, J., Gniadecka, M. and Wulf, H.C., 2004. Detection of skin cancer by classification of Raman spectra. *IEEE transactions on biomedical engineering*, *51*(10), pp.1784-1793.

EEG mind reading Mapping time-frequency response

2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

DEEP CONVOLUTIONAL NEURAL NETWORKS FOR INTERPRETABLE ANALYSIS OF EEG SLEEP STAGE SCORING

Albert Vilamala[1], Kristoffer H. Madsen[1,2] and Lars K. Hansen[1]

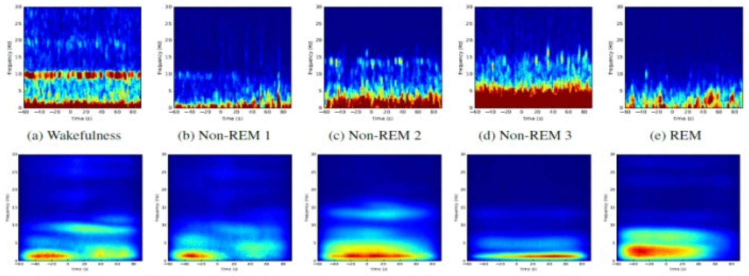


Fig. 1. Sensitivity analysis for subject 7. Top row shows characteristic sleep stage spectra; bottom row presents per-class sensitivity maps.

The Journal of Neuroscience, June 15, 2016 • 36(24):6583–6596 • 6583

Behavioral/Cognitive

Neural Markers of Responsiveness to the Environment in Human Sleep

Thomas Andrillon,[1,2] Andreas Trier Poulsen,[3] Lars Kai Hansen,[3] Damien Léger,[4] and Sid Kouider[1]

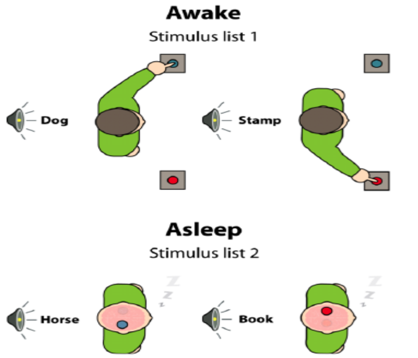
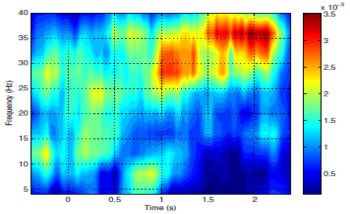
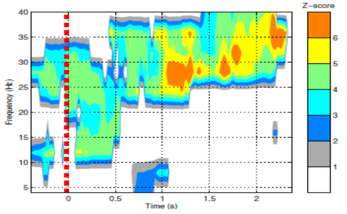


Figure 3.1: Before falling asleep subjects had to classify a word presented to them through headphones every 6 to 9 seconds as either animals or objects. This task allowed the mapping of each specific category with a specific motor response. This induction of a category-response mapping just before the onset of sleep is believed to promote the maintenance the task-set even after the sleep onset. Testing conditions encouraged the transition towards sleep while remaining engaged with the same task-set. For each subject one of two lists of words was presented during wakefulness and the other list during sleep ensuring actual abstract categorization rather than simple stimulus-response associations. (Source: Sid Kouider)



(a) Group average of scaled spectra-histo-grams.



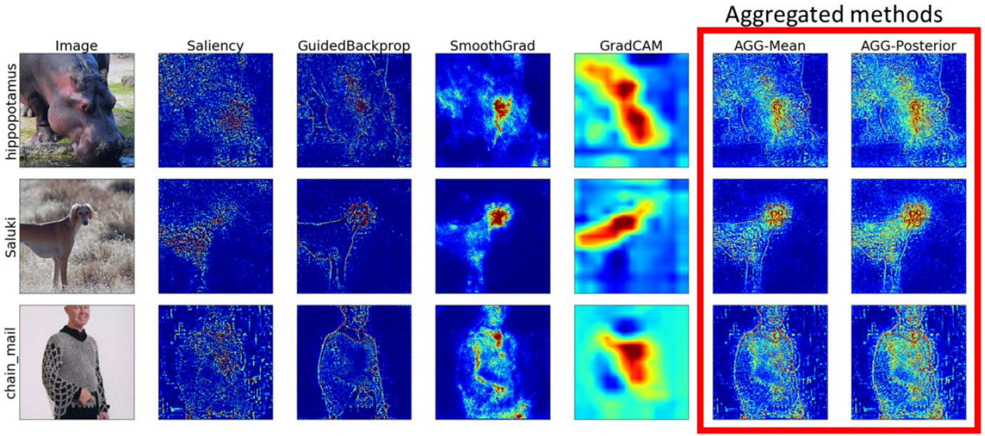
(b) Z-score.

Christian V Karsten (2012) Pattern Recognition in Electric Brain Signals- mind reading in the sleeping brain w./ Sid Kouider Paris. MSc Thesis DTU Informatics.
Andrillon, T., Poulsen, A.T., Hansen, L.K., Léger, D. and Kouider, S., 2016. Neural markers of responsiveness to the environment in human sleep. *Journal of Neuroscience*, *36*(24), pp.6583-6596.

Explain deep visual decisions – reducing uncertainty

Challenge

- 100+ proposals on how to explain image classification
- Do not agree on what to explain!



Aims:

Aggregate to reduce model uncertainy

Evaluate by counterfactual (what would happen if the image was different?)

Rieger, L. and Hansen, L.K., 2019. Aggregating explainability methods for neural networks stabilizes explanations. *arXiv:1903.00519*.
Chang, C.H., Creager, E., Goldenberg, A. and Duvenaud, D., 2018. Explaining image classifiers by counterfactual generation (ICLR19).

Epistemic /model uncertainty – consensus inference

Individual explainability methods come at idiosyncratic scales – non-parametric alignment of “gray scales”

Averaging, clipped and posterior weighted ensemble aggregation

–Reduce variance and model uncertainty

–Evaluation 1)– correlation with human annotations

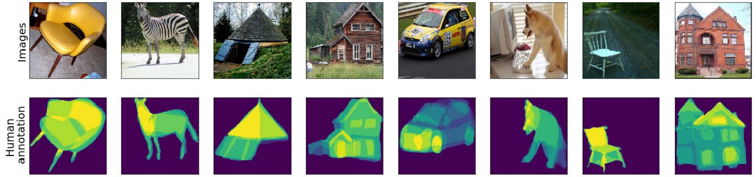


Figure 5. Example images and human-annotated heatmaps from (Mohseni & Ragan, 2018)

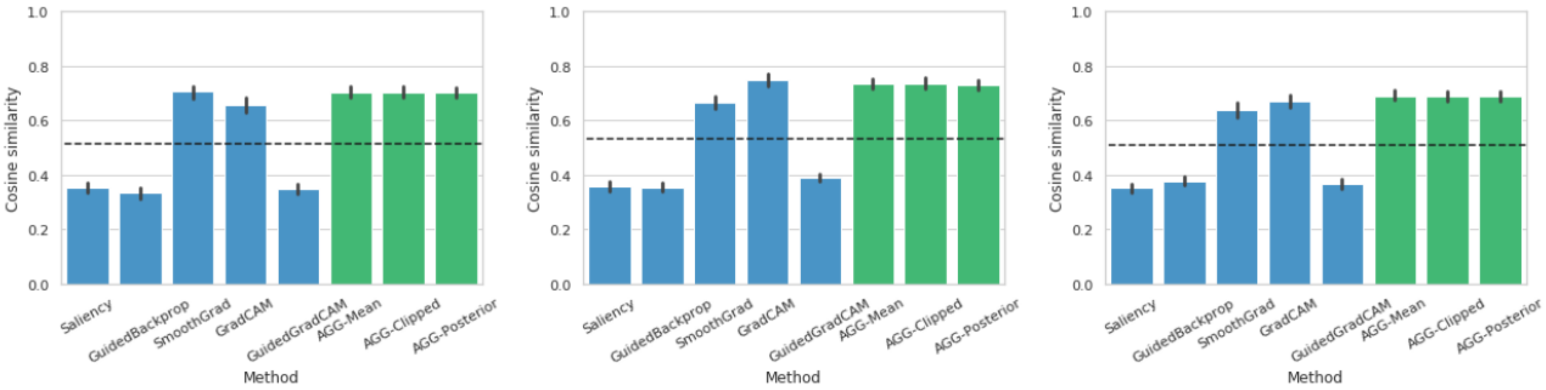
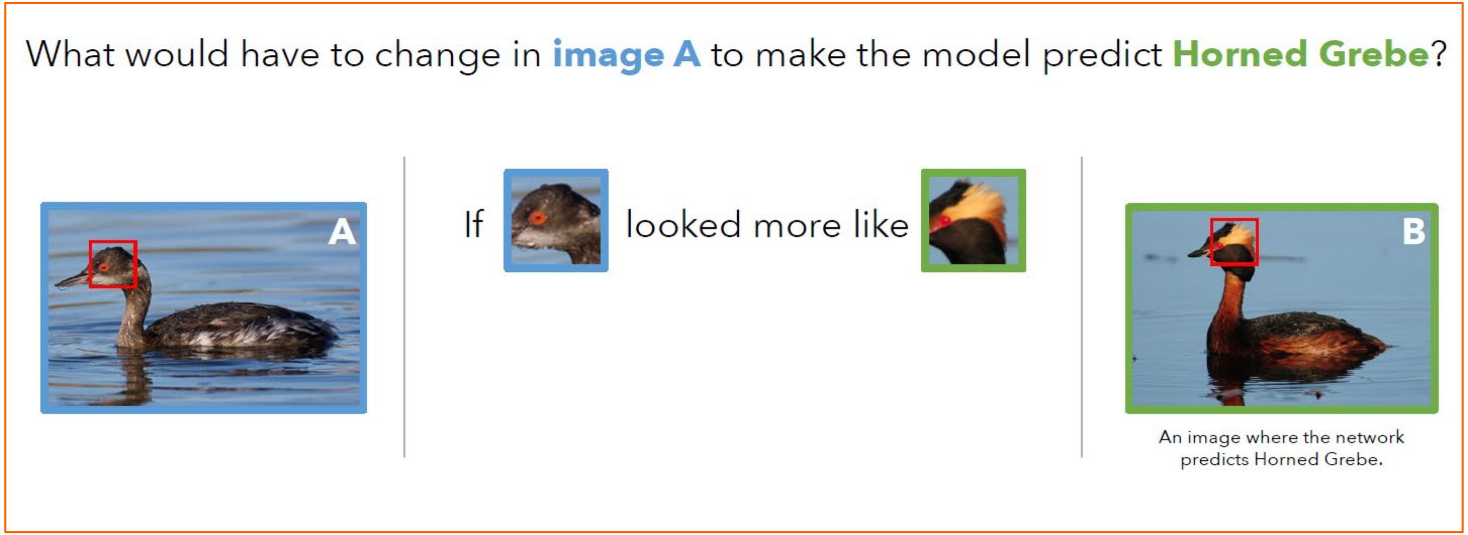


Figure 6. Averaged cosine similarity between human-assigned relevance and explanation methods reported on Inception(left), Xception (middle) and VGG19 (right). Aggregated methods in green. Dashed line is the average over all methods.

Open problem: Evaluation – counterfactuals?

Gyoal et al. (2019)   Users' think in terms of counterfactuals

*"Given a query image **A** for which a vision system predicts class **c**, a counterfactual visual explanation identifies how **A** could change such that the system would output a different specified class **c'** "*



Goyal Y, Wu Z, Ernst J, Batra D, Parikh D, Lee S: Counterfactual Visual Explanations. In *ICML 2019*.

IROF: Evaluate explanations by simple counterfactuals

Existing approach “Pixel flipping”

Saliency maps identify important pixels - grey out to understand how much performance deteriorates

Here:

Identify meaningful

(sub-)objects by image segmentation

Grey out segments rather than individual pixels

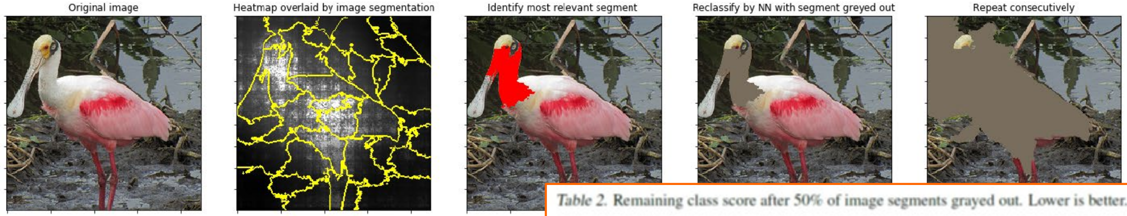
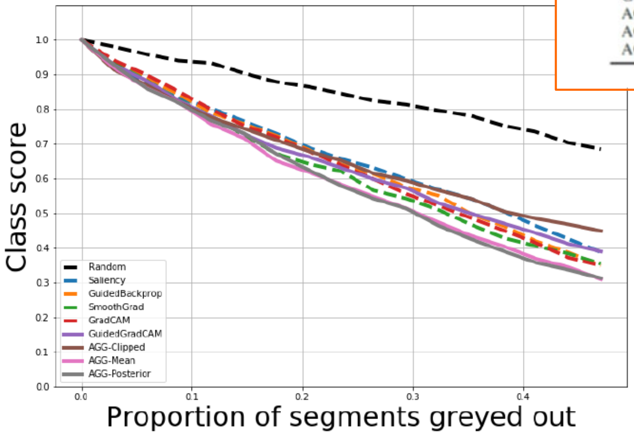


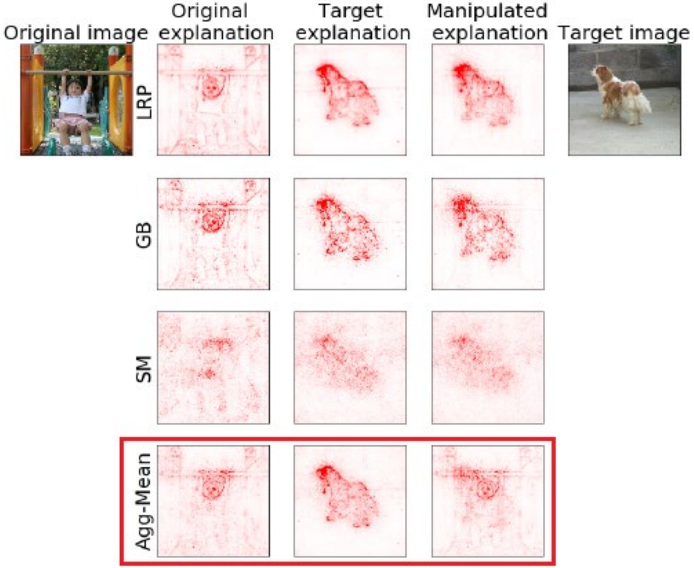
Table 2. Remaining class score after 50% of image segments grayed out. Lower is better.

	VGG19	Xception	Inception
Saliency	0.14 ± 0.01	0.39 ± 0.02	0.25 ± 0.01
Guided Backprop	0.00 ± 0.00	0.35 ± 0.02	0.20 ± 0.01
SmoothGrad	0.13 ± 0.01	0.35 ± 0.02	0.19 ± 0.01
Grad-CAM	0.09 ± 0.00	0.35 ± 0.01	0.22 ± 0.01
GuidedGrad-CAM	0.09 ± 0.00	0.35 ± 0.01	0.20 ± 0.01
AGG-Mean	$\mathbf{0.08 \pm 0.00}$	$\mathbf{0.31 \pm 0.01}$	$\mathbf{0.14 \pm 0.01}$
AGG-Posterior	$\mathbf{0.08 \pm 0.00}$	$\mathbf{0.31 \pm 0.01}$	$\mathbf{0.14 \pm 0.01}$
AGG-Clipped	0.14 ± 0.01	0.45 ± 0.02	0.27 ± 0.01



Rieger L, Hansen LK. IROF: a low resource evaluation metric for explanation methods. In *Workshop AI for Affordable Healthcare at ICLR 2020, Addis Ababa, Ethiopia*, 2020
Creager E, Goldenberg A, Duvenaud D: Explaining image classifiers by counterfactual generation ICLR19.

Attacks on explanations ”Fairwashing” – Exploit epistemic uncertainty



Fairwashing: the risk of rationalization
Aivodji et al. Proc ICML 2019.

“Fairwashing explanations with off-manifold detergent” Anders et al. Proc ICML 2020.

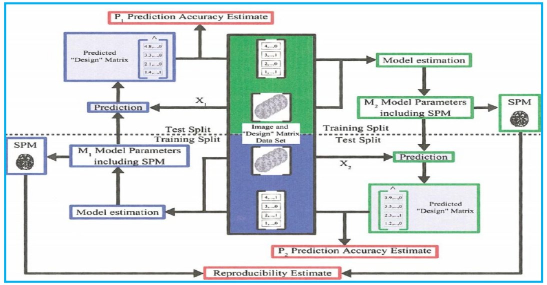
Effective defence:

Exploit epistemic uncertainty

Resilience by model averaging

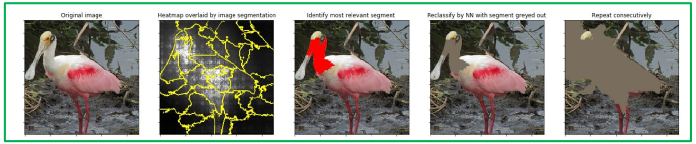
L Rieger, LK Hansen. “A simple defense against adversarial attacks on heatmap explanations.”
In proc ICML 2020 Workshop on Human Interpretability in ML (WHI)

Conclusions – ML is not black box – yet much to do…



Explainability is well established

- ✓ Function visualization – quest for mechanisms
- ✓ Decision level explanations – causality, counterfactuals
- ✓ Quantification of uncertainty
- ✓ Model averaging can improve performance
- ✓ Model averaging defends against fairwashing attacks



Many open problems

- Evaluation protocols?
- Explain with humans in the loop,
  competences?, visualize uncertainty?
- True counterfactuals require causal models