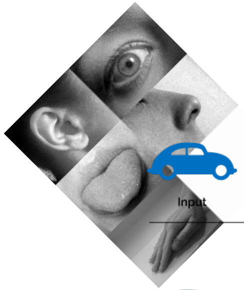
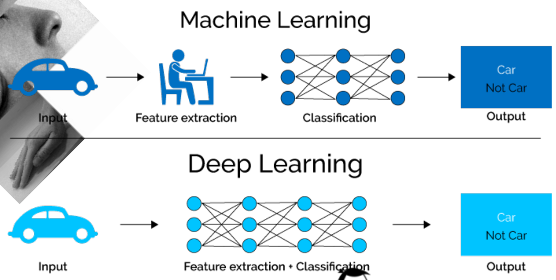
XAI - explainable AI

Lars Kai Hansen
DTU Compute, Technical University of Denmark

Wednesday	13:10 - 14:00	Intro & perspectives
	14:15 - 16:20	Explain deep learning & hands-on

AI hypotheses





Intelligent systems have active senses

- Seek relevant data (1).. causal discovery, embodied
- Attention

Intelligent systems are learning systems

- Machine learning … debugging, transfer learning
- Active learning … ask questions, intervention

Intelligent systems have social competences

- Communication, know own limitations, values
- Understand knowledge graphs, emotion

Intelligente systems perform "live"

- Global coordination – level C1/C2 conscious (2)
- Real-time operation –budget awareness (3)

(1) Bajcsy, R., 1988. Active perception. Proceedings of the IEEE, 76(8), pp.966-1005.
(2) Dehaene, S., Lau, H. and Kouider, S., 2017. What is consciousness, and could machines have it?. *Science*, *358*(6362), pp.486-492.
(3) Little, D.Y.J. and Sommer, F.T., 2013. Learning and exploration in action-perception loops. *Frontiers in neural circuits*, *7*, p.37.

the Cognitive Systems platform

Machine learning

-Ole Winther, Morten Mørup, Søren Hauberg, Jes Frellsen,

Computational social science

-Sune Lehmann, Jakob Eg Larsen

Cognitive science

-Sid Kouider, Ivana Konvalinka, Tobias Andersen,

Co-author network: Stanford, MIT, UCLA, UC London, ENS Paris,…

Top conferences …NeurIPS, AISTATS, ICLR, ICML

International peer review panel 2008, 2013, 2018

"Cutting edge - international leader"

Widex: *"…game changer for the hearing aid business"*

"WIDEX EVOKE will forever change what people expect from hearing aids."

Start-ups Peergrade, Spektral Experience, Corti, Unumed, BrainCapture

DABAI - open source ML workflows + Danish resources

We promote Safe & Sustainable AI

Safe AI = secure – test & verified software and hardware, adversarials

Safe AI = open source – methods, code, hardware, check and evolution

Safe AI = self-conscious – understands own role

Safe AI = can keep a secret – privacy by design

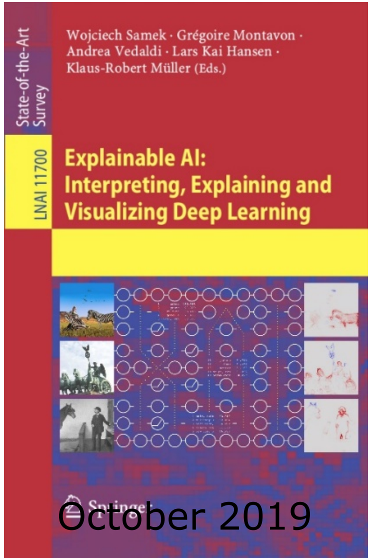
Safe AI = has calibrated values – debug for stereotypes, biases

Safe AI = is accountable

- transparent, communicating, ”right to explanation”

Safe AI = understands social relations

Safe AI = is sustainable – understands budgets & footprints



October 2019

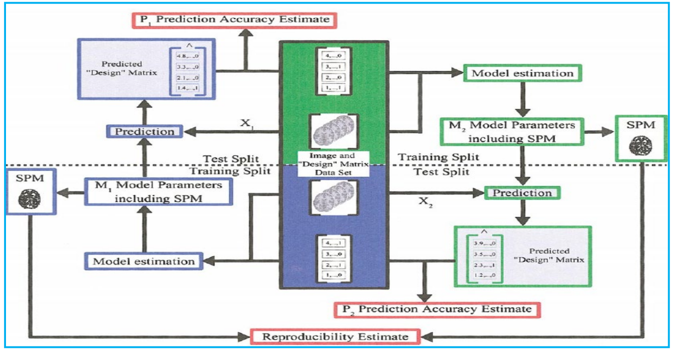
Outline

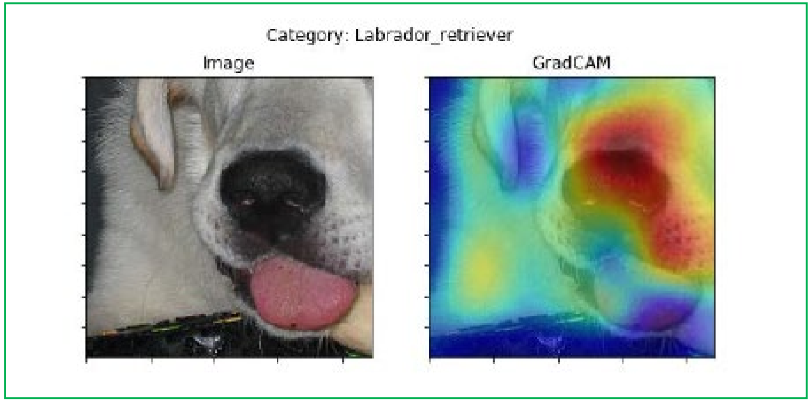
Why explain?
- Trust, debugging, legal, scientific applications

Background
- Explanation as an ill-posed task
- interpretation vs explanation,
- Objectives from Explainable Expert Systems

Function level visualization
- NPAIRS, PR-curves,
- Robustness vs methods, networks, training sets
- Uncertainty quantification

Decision explanations
- Consensus inference
- New results using aggregation





Why explain AI? - motivations

Trust & debugging

AI as a collaborator / teacher - social competences

Verification, performance optimization…

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

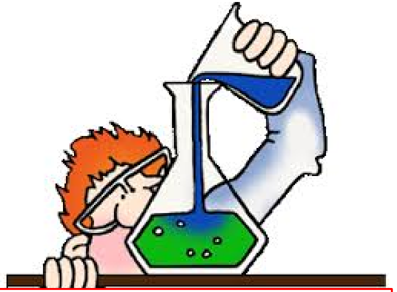
Legal requirement - "right to explanation"

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

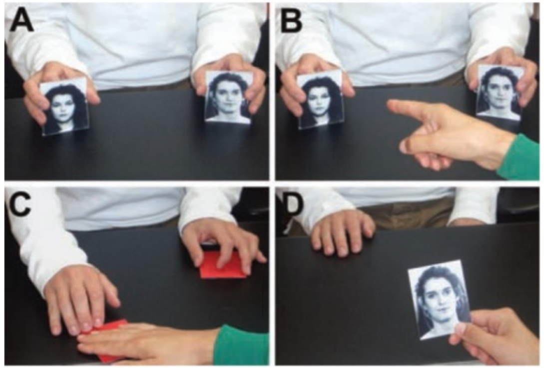
Uniqueness? – model uncertainty, robustness

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.

How well do you explain?- “*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”

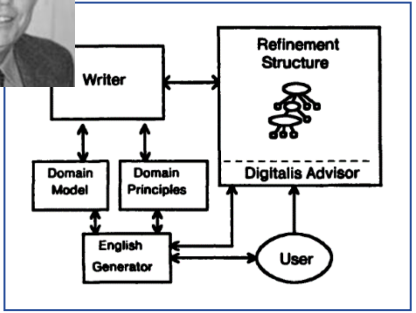
Johansson, P., Hall, L., Sikström, S. and Olsson, A., 2005. Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), pp.116-119.
Johansson, P., Hall, L., Sikström, S., 2008. From change blindness to choice blindness. Psychologia, 51(2), pp.142-155.

Explainability - objectives

WR Swartout, and JD Moore (1993)

Fidelity

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



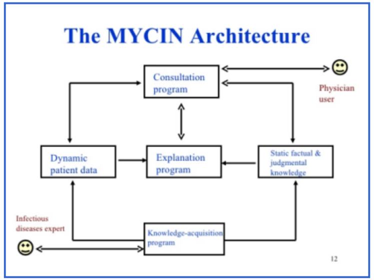
XPLAIN

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.



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)

Terminology Explanation vs. interpretability

Turner (2016)

- Explanation= single decisions (communication).
- Interpretability = understanding the mechanism (causal)

Guidotti et al. (2018)

- “Which are the real problems requiring interpretable models and explainable predictions?”

Doshi-Velez and Kim (2017)

- “Interpret means to explain or to present in understandable terms. In the context of ML systems, we define interpretability as the ability to explain or to present in understandable terms to a human.”
- “We argue that the need for interpretability stems from an incompleteness in the problem formalization, creating a fundamental barrier to optimization and evaluation.”

Gilpin et al. (2018)

- “…interpretability, loosely defined as the science of comprehending what a model did”
- “While interpretability is a substantial first step, these mechanisms need to *also* be complete, with the capacity to defend their actions, provide relevant responses to questions, and be audited. Although interpretability and explainability have been used interchangeably, we argue there are important reasons to distinguish between them.”

R Turner, 2016, model explanation system. In *Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on* (pp. 1-6). IEEE.
R Guidotti et al, 2018. A survey of methods for explaining black box models. ACM Computing Surveys (CSUR), 51(5), p.93.
Doshi-Velez, F. and Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
Gilpin et al., 2018. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. arXiv preprint arXiv:1806.00069.

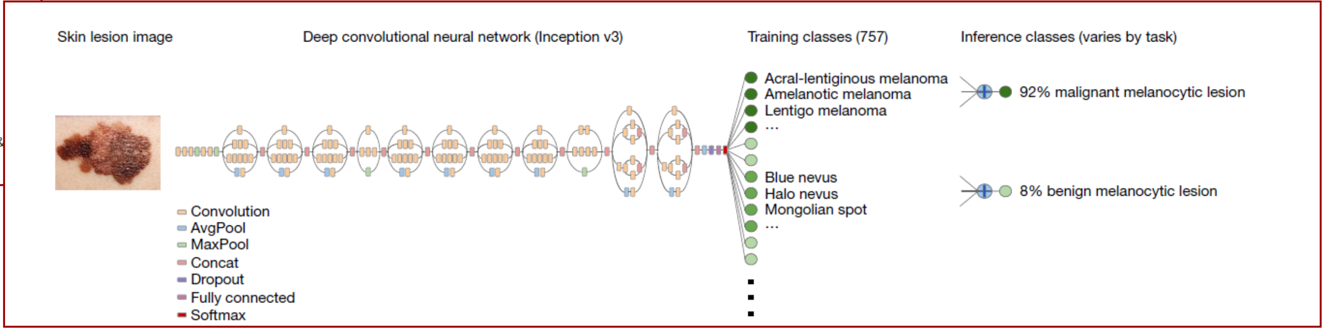
Dermatologist-level classification of skin cancer with deep neural networks

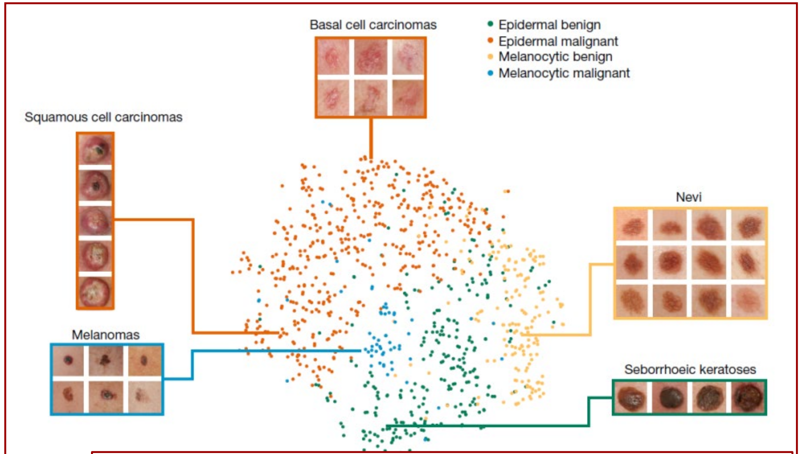
nature
International journal of science

Letter | Published: 25 January 2017

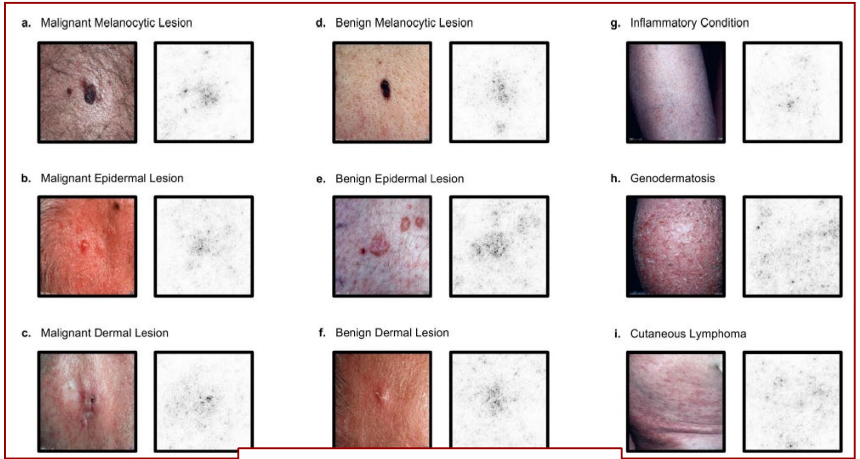
Dermatologist-level classification of skin cancer with deep neural networks

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



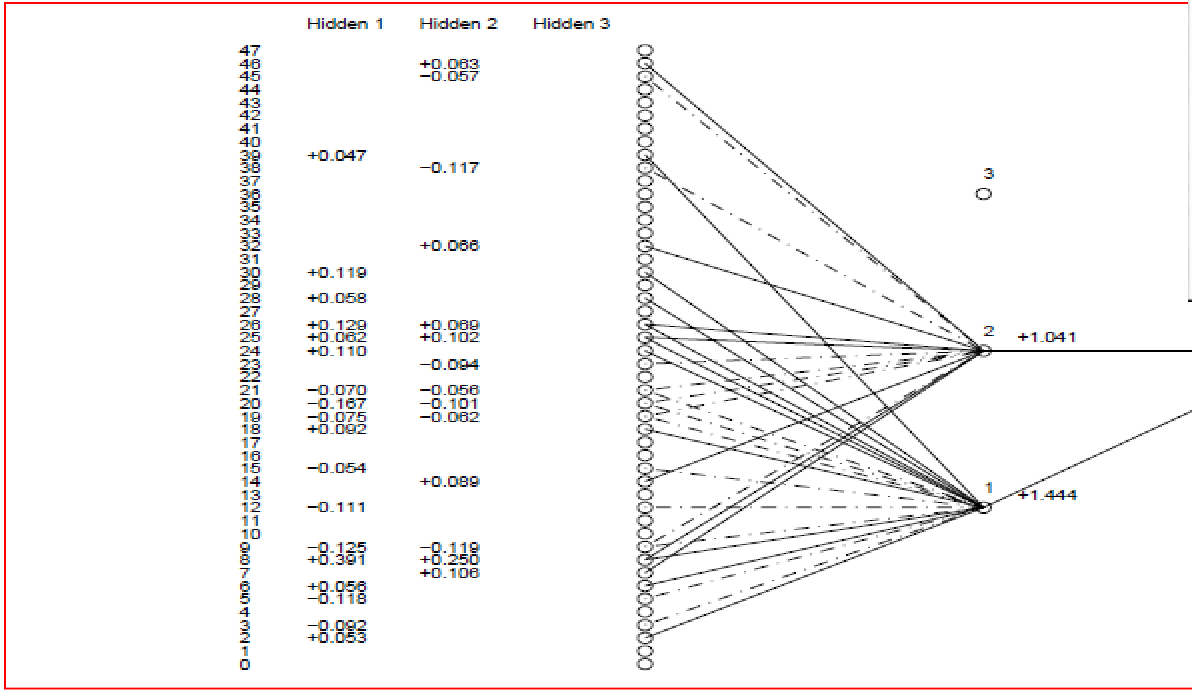


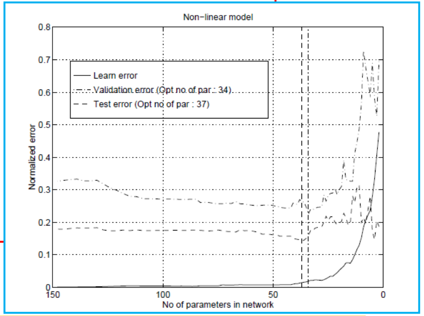
t-Distributed Stochastic Neighbor Embedding (*t-SNE*) plot of embedding



L1 sensitivity map

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





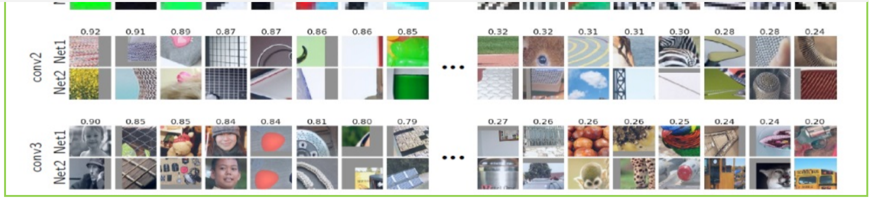
LeCun, Y., Denker, J.S. and Solla, S.A., 1990. Optimal brain damage. In Advances in neural information processing systems (pp. 598-605).
Lautrup, B, Hansen, LK, Law, I., Mørch, N, Svarer, C, Strother, S Massive weight sharing: a cure for extremely ill-posed problems. In *Workshop on supercomputing in 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).

Uniqueness of DNN?

Published as a conference paper at ICLR 2016

CONVERGENT LEARNING: DO DIFFERENT NEURAL NETWORKS LEARN THE SAME REPRESENTATIONS?

Yixuan Li[1]*, Jason Yosinski[1]*, Jeff Clune[2], Hod Lipson[3], & John Hopcroft[1]



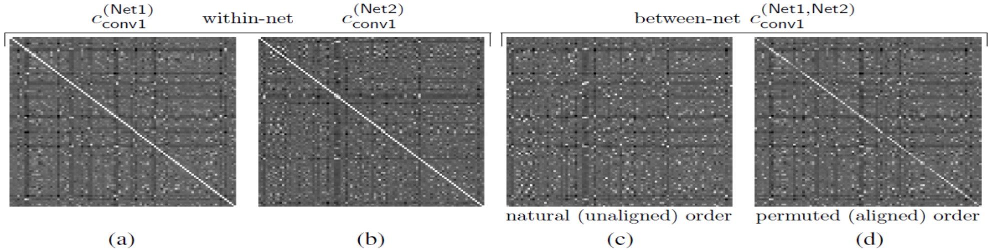
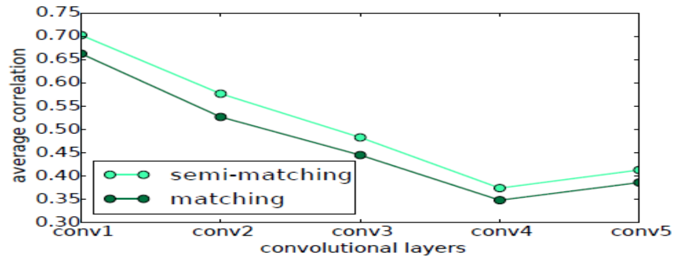
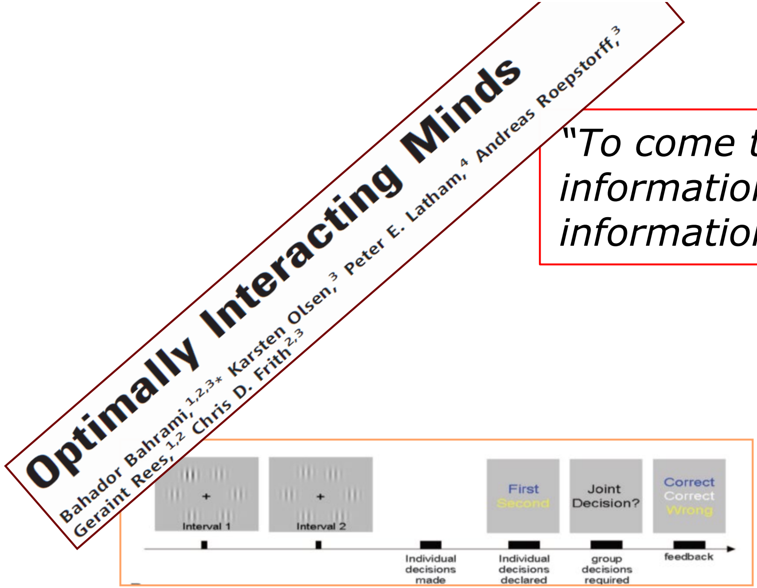


Figure 1: Correlation matrices for the conv1 layer, displayed as images with minimum value at black and maximum at white. **(a,b)** Within-net correlation matrices for Net1 and Net2, respectively. **(c)** Between-net correlation for Net1 vs. Net2. **(d)** Between-net correlation for Net1 vs. a version of Net2 that has been permuted to approximate Net1's feature order. The partially white diagonal of this final matrix shows the extent to which the alignment is successful; see Figure 3 for a plot of the values along this diagonal and further discussion.

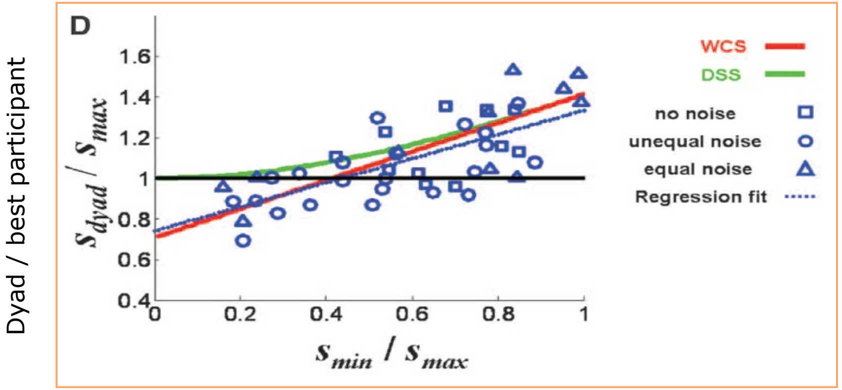


L, Yixuan, J Yosinski, J Clune, H Lipson, J Hopcroft. "Convergent Learning: Do different neural networks learn the same representations?." *arXiv preprint arXiv:1511.07543* (2015)

Communicating uncertainty improves group inference



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



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*

Reproducibility of parameters/visualization? ...hints from asymptotic theory

Asymptotic theory investigates the sampling fluctuations in the limit N -> ∞

Cross-validation good news: The ensemble average predictor is equivalent to training on all data (Hansen & Larsen, 1996)

Simple asymptotics for parametric and semi-parametric models

(Some results available also for non-parametric e.g. kernel machines)

In general: Asymptotic predictive performance has **bias and variance components**, there is proportionality between parameter fluctuation and the variance component...

Advances in Computational Mathematics 5(1996)269–280 269

Linear unlearning for cross-validation

Lars Kai Hansen and Jan Larsen
CONNECT, *Electronics Institute B349, Technical University of Denmark, DK-2800 Lyngby, Denmark*
E-mail: lkhansen,jlarsen@ei.dtu.dk

The sensitivity map & the PR plot

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

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

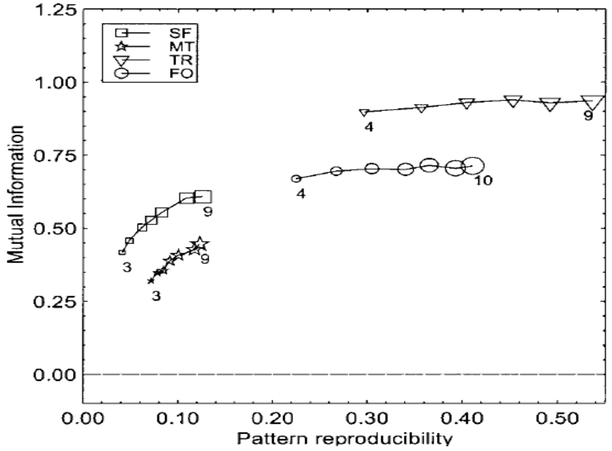
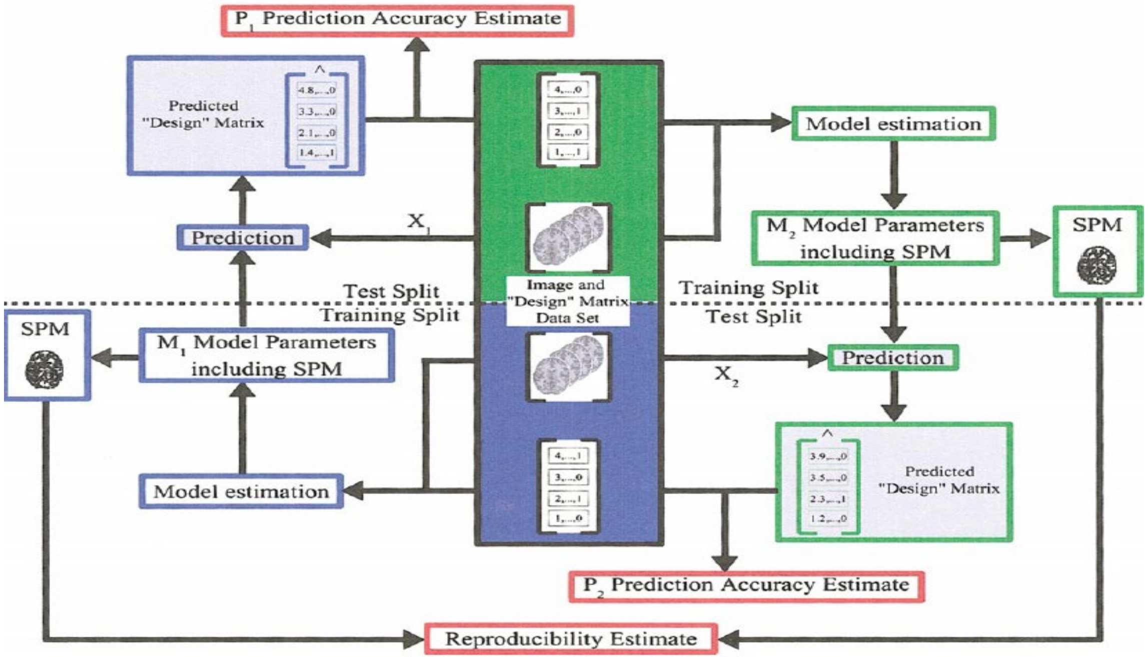


FIG. 3. Plot of scan/label mutual information versus reproducibility signal/noise for the four data sets, for varying numbers of subjects in the training set. There were 2 labels/4 scans per subject (balanced data set; Setup 1, Table 1) corresponding to the dashed solid line in Fig. 4. We see that both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

The sensitivity map measures the impact of a specific feature/location 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

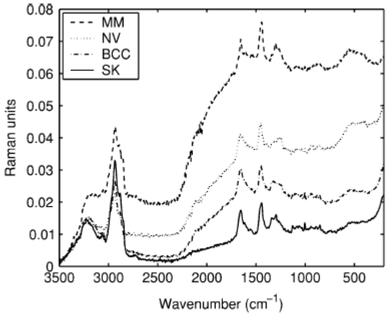
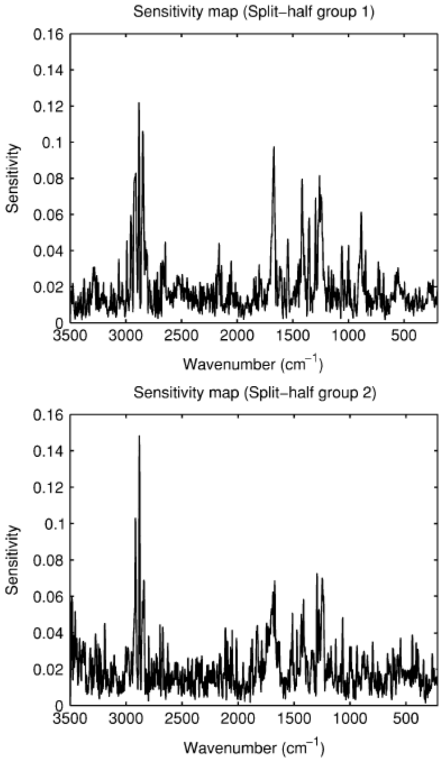


Fig. 1. Examples of the NIR-FT Raman spectra of benign and malignant skin lesions and tumors: BCC, MM, NV, and SK.

	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



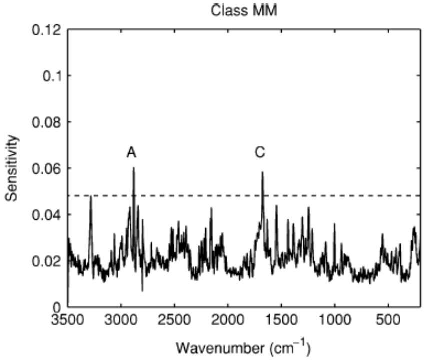


Fig. 10. Sensitivity maps for the MM class. Dashed line indicates 95% confidence interval. Sensitivity map seems more noisy than the BCC sensitivity map in Fig. 9. Region marked A represents the CH^{-} vibrations in the lipids and proteins around 2940 cm^{-1} and region marked C reflects the amide I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

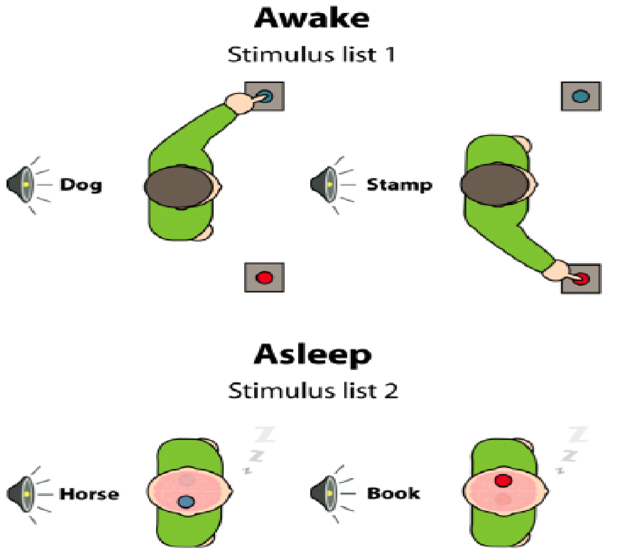
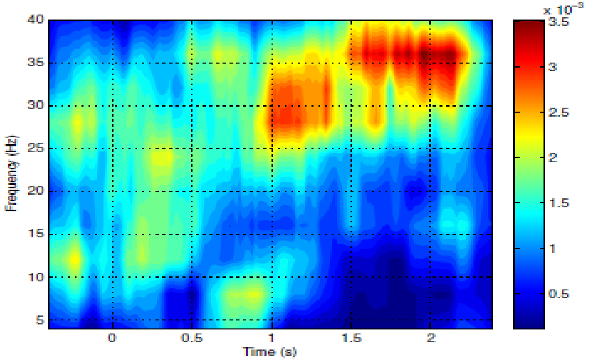
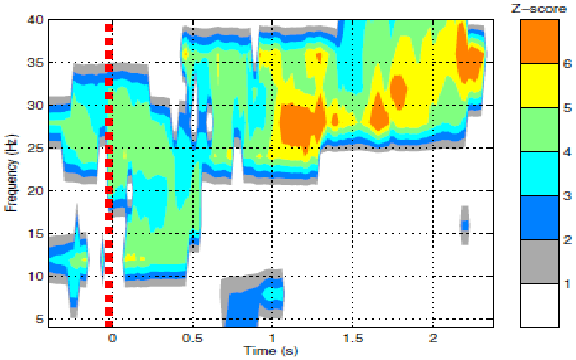


Figure 3.1: Before falling asleep subjects had to classify a word presented to them through headphones every 6 to 9 seconds as either animals or objects. This task allowed the mapping of each specific category with a specific motor response. This induction of a category-response mapping just before the 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)



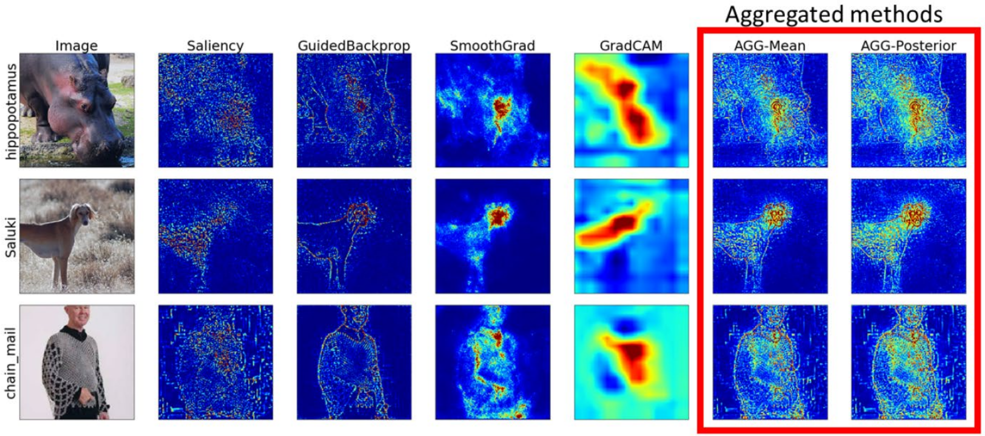
(a) Group average of scaled spectra-histo-grams.



(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 w/ Laura Rieger



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).

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

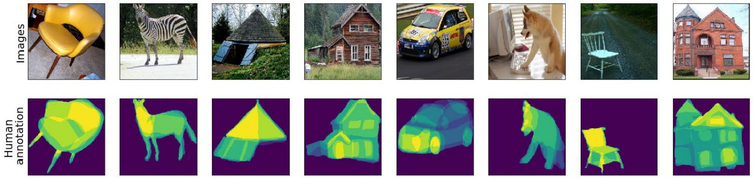


Figure 5. Example images and human-annotated heatmaps from (Mohseni & Ragan, 2018)

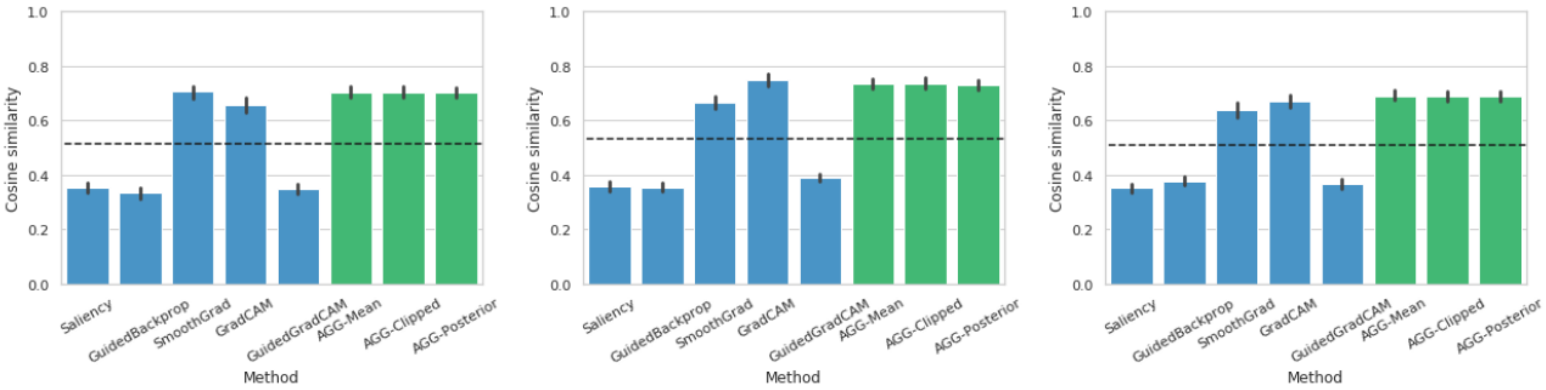


Figure 6. Averaged cosine similarity between human-assigned relevance and explanation methods reported on Inception(left), Xception (middle) and VGG19 (right). Aggregated methods in green. Dashed line is the average over all methods.

Evaluate explanations by simple counterfactuals

Existing approach "Pixel flipping"

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

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	**0.08 ± 0.00**	**0.31 ± 0.01**	**0.14 ± 0.01**
AGG-Posterior	**0.08 ± 0.00**	**0.31 ± 0.01**	**0.14 ± 0.01**
AGG-Clipped	0.14 ± 0.01	0.45 ± 0.02	0.27 ± 0.01

Here:

Identify meaningful

(sub-)objects by image segmentation

Grey out segments rather than individual pixels

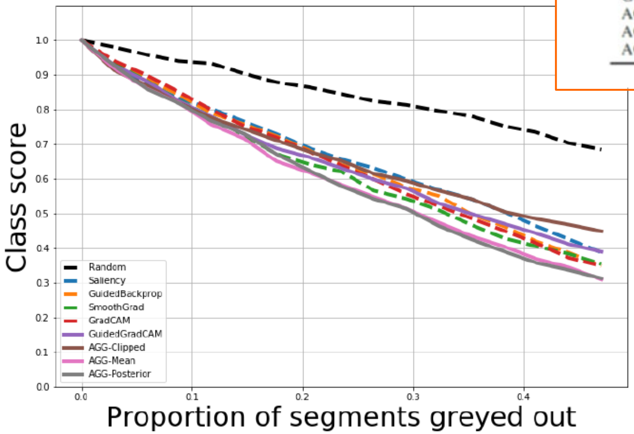
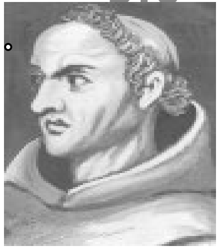


Figure 4. Quantitative evaluation: Decay of class scores with seg-

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. Explaining image classifiers by counterfactual generation ICLR19.

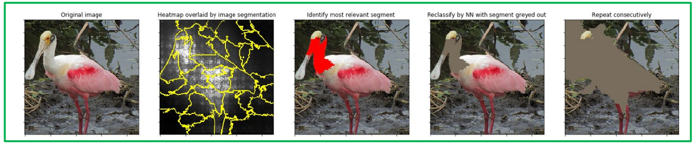
Conclusions

Do not multiply causes!



Explanability is not a new concept

– Yet, many open research problems, some at the interface to fairness
– Function visualization – quest for mechanisms
– Decision level explanations – causality?, counterfactuals



Visualize general ML functions with perturbation based methods

- saliency maps, sensitivity maps

NPAIRS resampling workflow

- quantification of performance and uncertainty