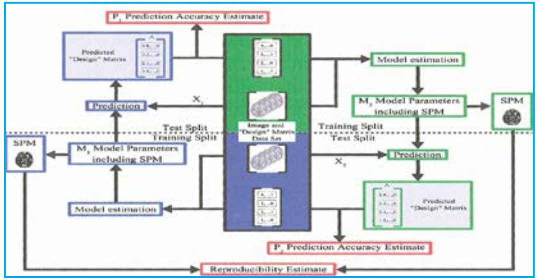
Opening the Black Box

How to Interpret Machine Learning Functions and Their Decisions

Lars Kai Hansen and Laura Rieger

Outline


P_1 Prediction Accuracy Estimate
Predicted "Design" Matrix
Prediction
X_1
Test Split
Training Split
Image and "Design" Matrix Data Set
SPM
M_2 Model Parameters including SPM
Model estimation
Model estimation
M_1 Model Parameters including SPM
SPM
Training Split
Test Split
Prediction
X_2
Predicted "Design" Matrix
P_2 Prediction Accuracy Estimate
Reproducibility Estimate

Why open the black box?

- Trust,debugging, legal, scientific applications

Principles

- Interpretation vs explanation, desiderata from "Explainable Expert Systems"
- ML Function visualizations

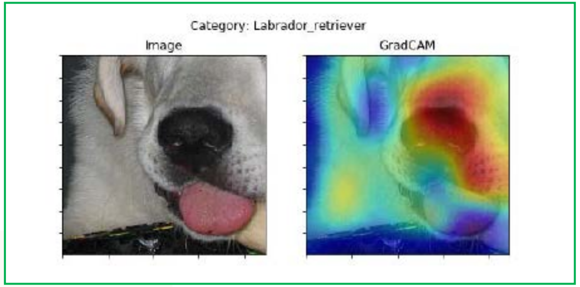
Function level visualization - Prediction & reproducibility evaluation

- NPAIRS, PR-curves,

Break

Decision level visualizations

- Methods for deep learning, examples from object recognition



Demonstration

- Robustness vs methods, networks, training sets

Keras
A deep learning library

Why open the black box? Multiple motivations

Trust

An AI that communicates its decisions is inherently more trustworthy

Debugging

Verification, performance optimization…

Align values - reduce biases, adversarial risks …

Legal - "right to explanation"

General data protection regulatory May 26, 2018

Scientific applications of machine learning

learning from machine learning solutions, causal mechanisms, why …

Explainability - General desiderata

Swartout, W. R. and Moore, J. D. (1993)

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

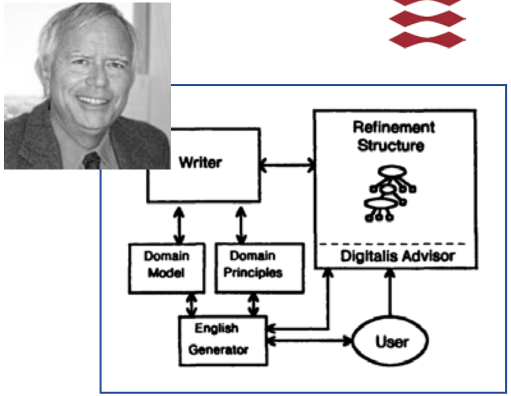
Understandability Involves multiple usability factors including terminology, user competencies, levels of abstraction and interactivity.

Uncertainty!

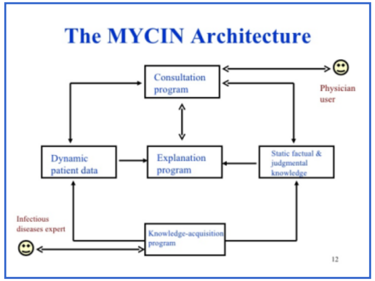
Sufficiency Should be able to explain function and terminology and be detailed enough to justify decision.

Low Construction overhead The explanation should not dominate the cost of designing the AI.

Efficiency: The explanation system should not slow down the AI significantly.

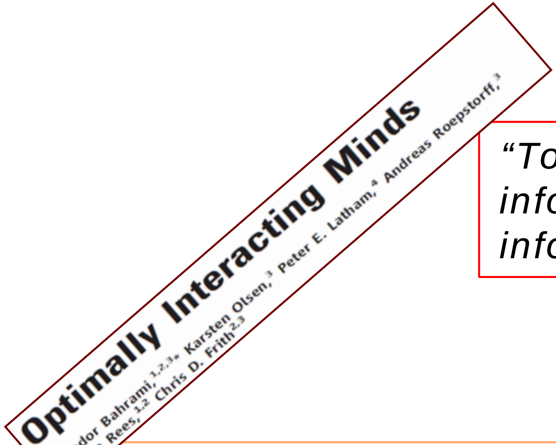


XPLAIN



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)

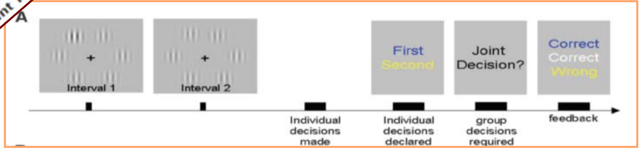
Communicating uncertainty improves group inference

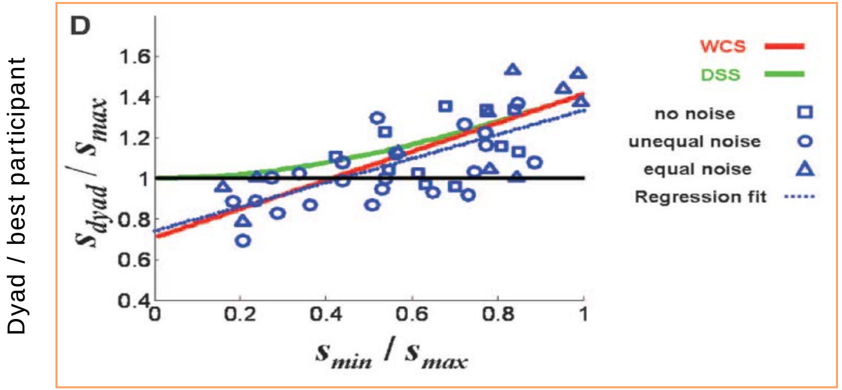


Optimally Interacting Minds

Bahador Bahrami,[1,2,3]* Karsten Olsen,[3] Peter E. Latham,[4] Andreas Roepstorff,[3] 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…"*





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*

Explanation vs interpretability in recent literature

- **Turner (2016)**
  - Explanation= single decisions. Interpretability = understanding the mechanism

- **Guidotti et al. (2018)**
  - "Which are the real problems requiring <u>interpretable models</u> and <u>explainable predictions</u>?"

- **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."

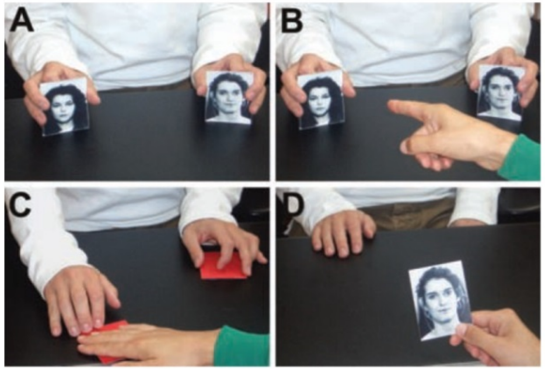
- **Gilpin et al. (2018)**
  - "…interpretability, loosely defined as the science of comprehending what a model did"
  - "While interpetability is a substantial first step, these mechanisms need to *also* be complete, with the capacity to defend their actions, provide relevant responses to questions, and be audited. Although interpretability and explainability have been used interchangeably, we argue there are important reasons to distinguish between them."
  - "Explainable models are interpretable by default, but the reverse is not always true"**.**

R Turner, 2016, September. A 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.

How reliable are human explanations?

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.

Opening the black box - mapping ML functions

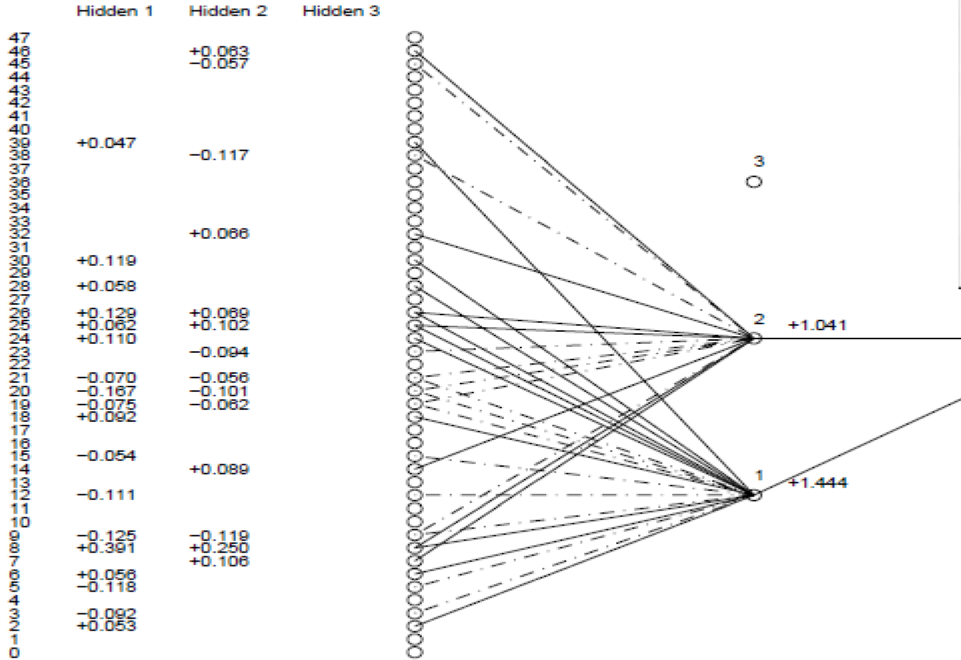
- A significant objective in scientific applications
- Legal requirement – minimal explanation? (Floridi et al.)

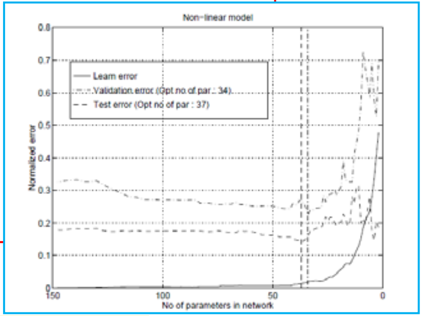
> The GDPR states that data controllers must notify consumers how their data will be used, including "*the existence of automated decision-making, and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject*.“
>
> https://gdpr-info.eu/art-15-gdpr/

- Saliency maps
  - Saliency defined in Le Cun et al. (1990) for network pruning, can be used to inputs as well –the estimated cost of removing input $\sim \; \Sigma_i \, H_i \, w_i^2$
- Sensitivity maps
  - Zurada et al. (1994) $\sim \; < (\, d \log(p) \, / \, dx \,)^2 >$ (average over data)

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.
LeCun, Y., Denker, J.S. and Solla, S.A., 1990. Optimal brain damage. In Advances in neural information processing systems (pp. 598-605).
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.

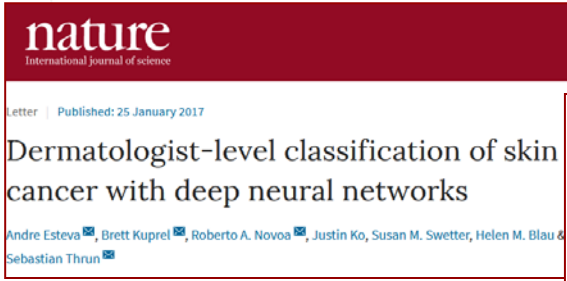
Saliency map for a neural network for decoding PET brain scans (1994-95)




Non−linear model
Learn error
Validation error (Opt no of par : 34)
Test error (Opt no of par : 37)
Normalized error
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
150
100
50
0
No of parameters in network

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

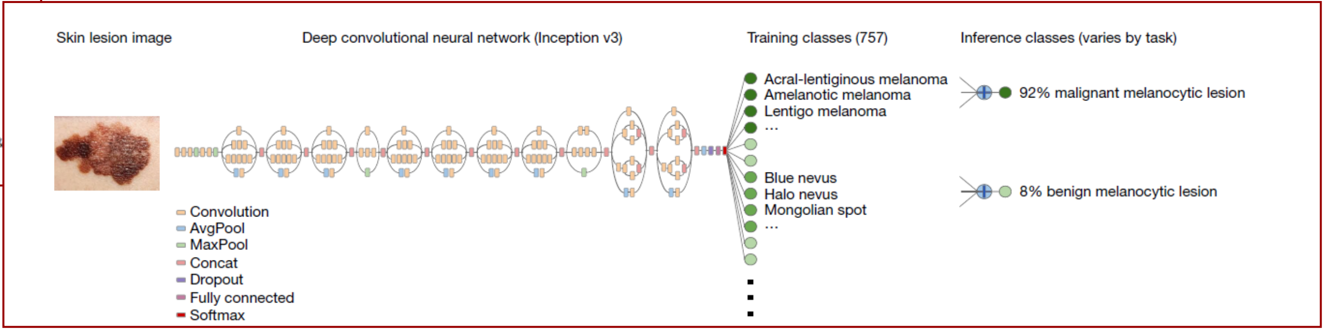


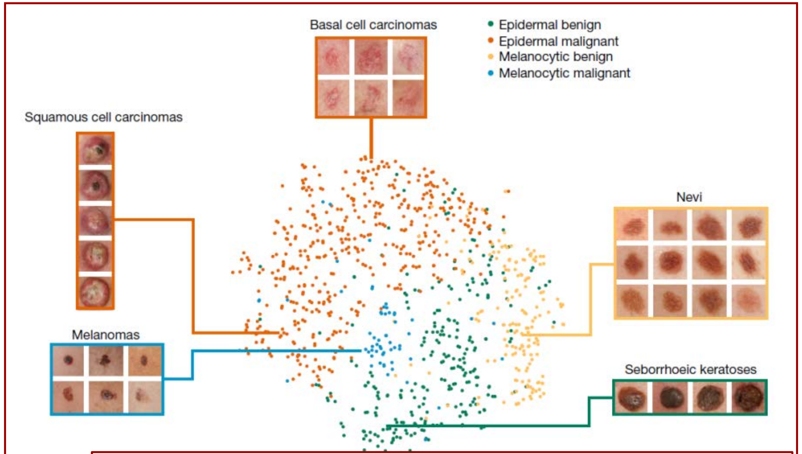
nature
International journal of science

Letter | Published: 25 January 2017

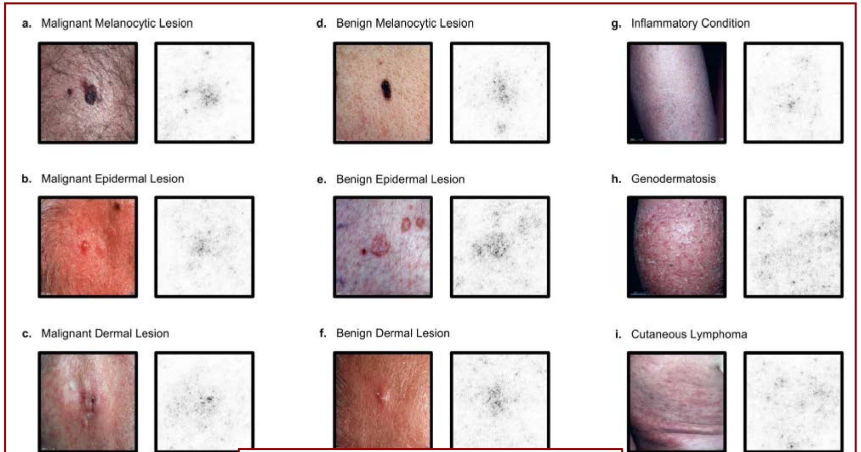
Dermatologist-level classification of skin cancer with deep neural networks

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

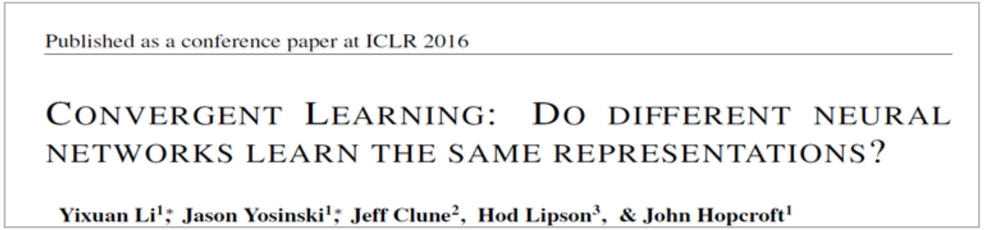




t-Distributed Stochastic Neighbor Embedding (*t-SNE*) plot of embedding


a. Malignant Melanocytic Lesion
d. Benign Melanocytic Lesion
g. Inflammatory Condition
b. Malignant Epidermal Lesion
e. Benign Epidermal Lesion
h. Genodermatosis
c. Malignant Dermal Lesion
f. Benign Dermal Lesion
i. Cutaneous Lymphoma

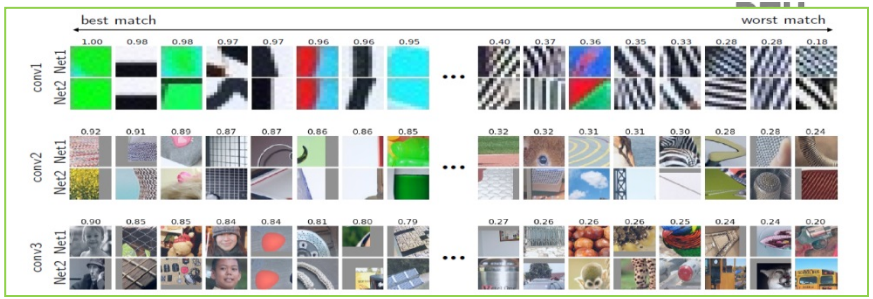
L1 sensitivity map



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]



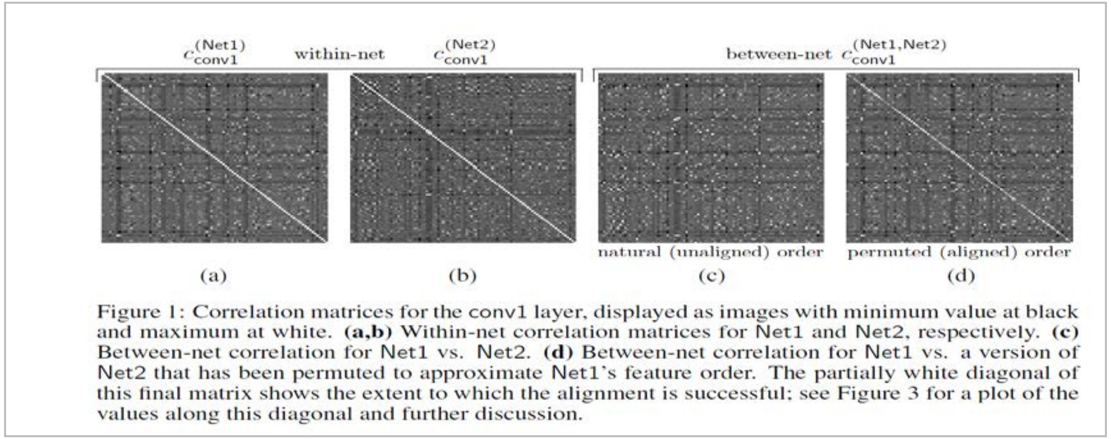
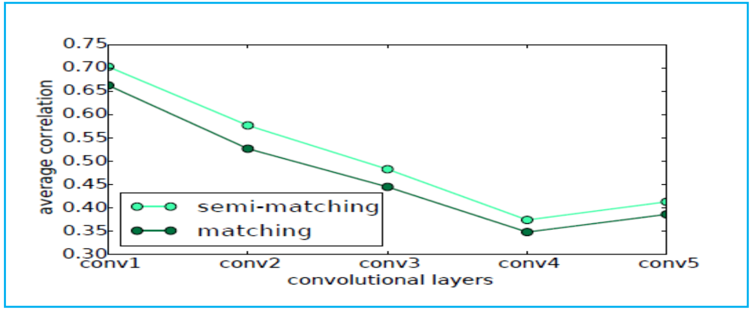
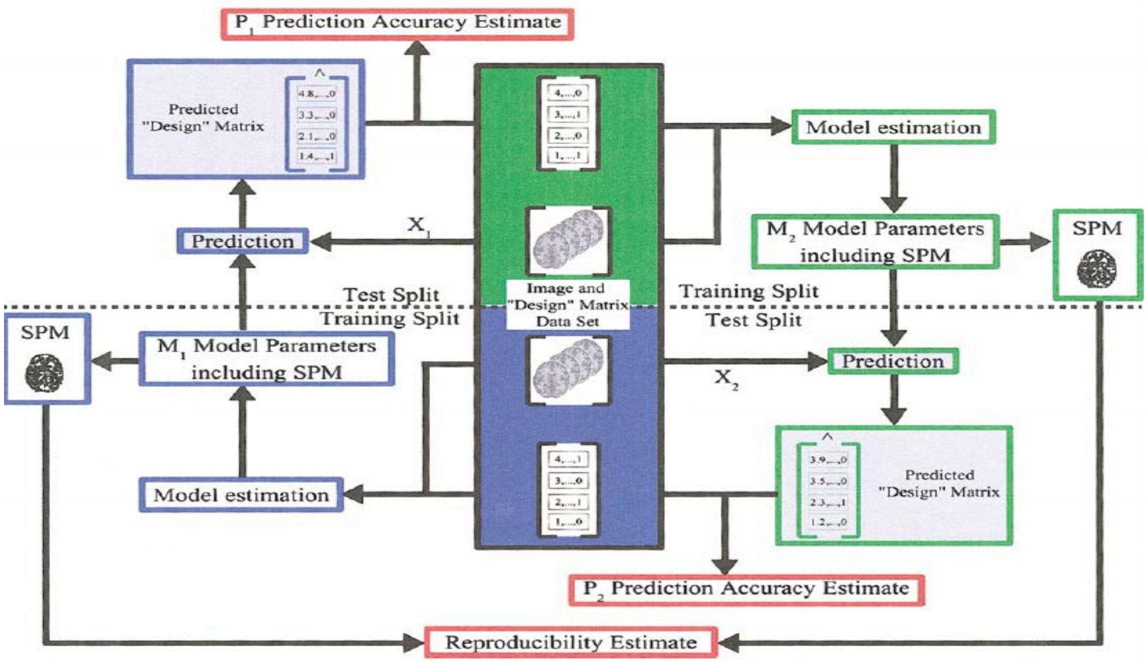


Figure 1: Correlation matrices for the conv1 layer, displayed as images with minimum value at black and maximum at white. **(a,b)** Within-net correlation matrices for Net1 and Net2, respectively. **(c)** Between-net correlation for Net1 vs. Net2. **(d)** Between-net correlation for Net1 vs. a version 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)

NPAIRS: Reproducibility of parameters



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

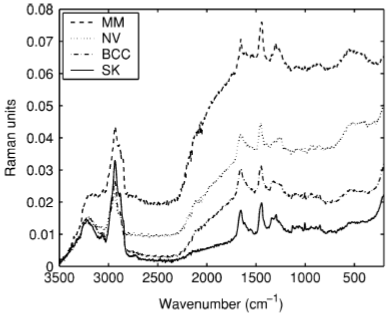
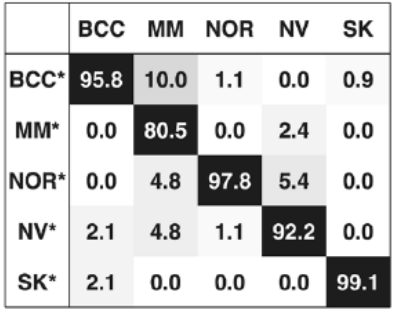
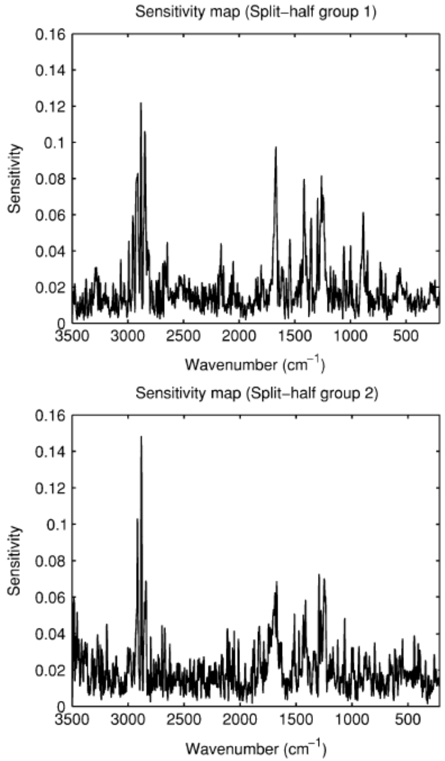


Fig. 1. Examples of the NIR-FT Raman spectra of benign and malignant skin lesions and tumors: BCC, MM, NV, and SK.



	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



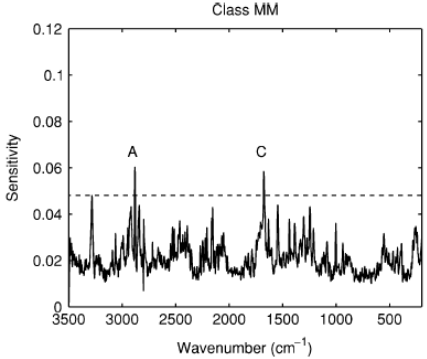


Fig. 10. Sensitivity maps for the MM class. Dashed line indicates 95% confidence interval. Sensitivity map seems more noisy than the BCC sensitivity map in Fig. 9. Region marked A represents the CH^- vibrations in the lipids and proteins around 2940 cm^{-1} and region marked C reflects the amide I band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

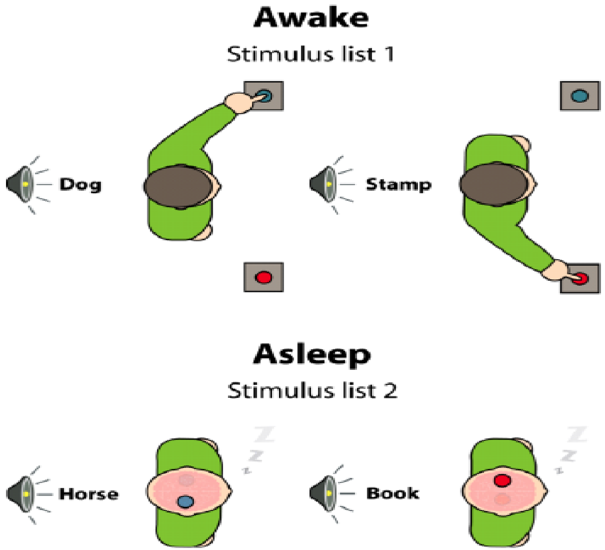
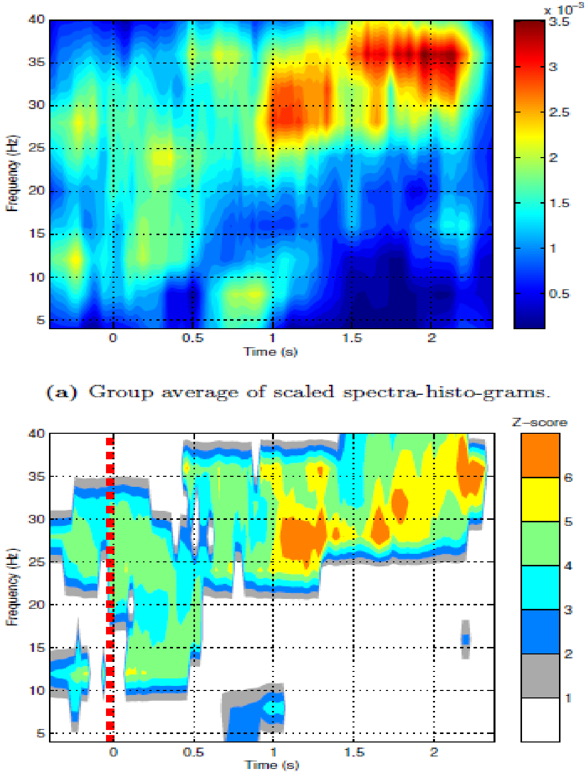


Figure 3.1: Before falling asleep subjects had to classify a word presented to them through headphones every 6 to 9 seconds as either animals or objects. This task allowed the mapping of each specific category with a specific motor response. This induction of a category-response mapping just before the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

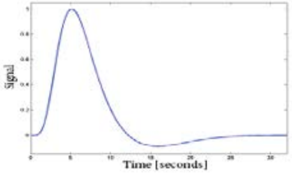


(a) Group average of scaled spectra-histo-grams.

(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.

BOLD - functional MRI

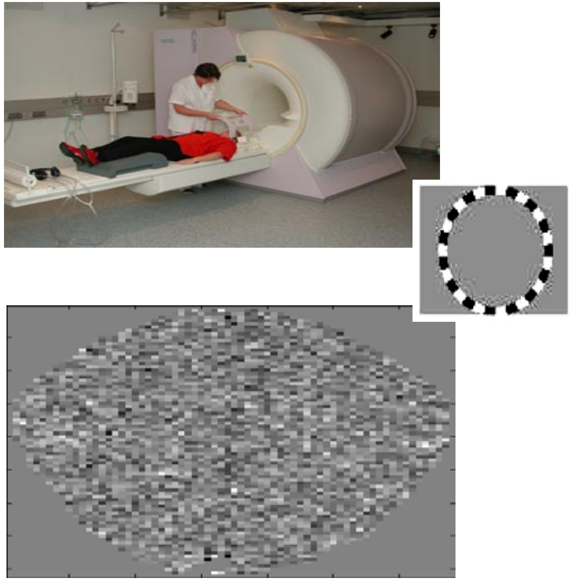


Indirect measure of neural activity - hemodynamics

A cloudy window to the human brain

Challenges:

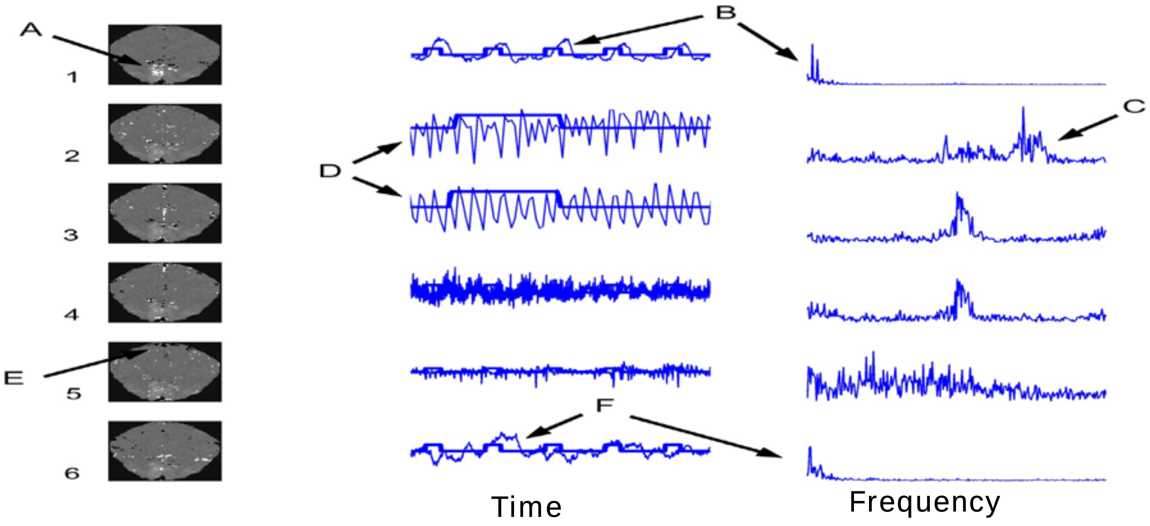
- Signals are multi-dimensional mixtures
- No simple relation between measures and brain state -”what is signal and what is noise”?



TR = 333 ms

Independent components – ”disentanglement”

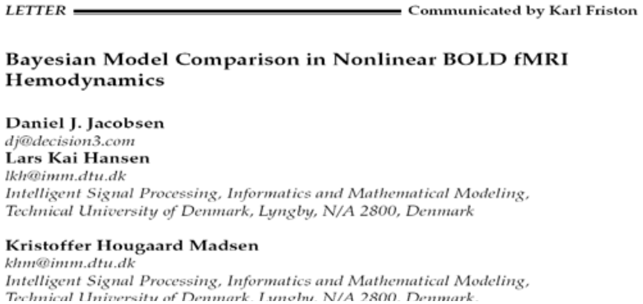
deconstruct signal in space x time components



McKeown, Hansen, Sejnowski, Curr. Op. in Neurobiology (2003)

Main message: Models should be predictive and informative

BOLD fMRI: Is hemodynamic de-convolution feasible?



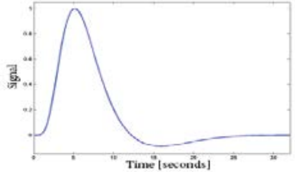
LETTER Communicated by Karl Friston

Bayesian Model Comparison in Nonlinear BOLD fMRI Hemodynamics

Daniel J. Jacobsen
dj@decision3.com
Lars Kai Hansen
lkh@imm.dtu.dk
Intelligent Signal Processing, Informatics and Mathematical Modeling, Technical University of Denmark, Lyngby, N/A 2800, Denmark

Kristoffer Hougaard Madsen
khm@imm.dtu.dk
Intelligent Signal Processing, Informatics and Mathematical Modeling, Technical University of Denmark, Lyngby, N/A 2800, Denmark

Neural Computation **20, 738–755** (2008)



Balloon model: Non-linear relations between stimulus and physiology – described by four non-linear differential eqs.

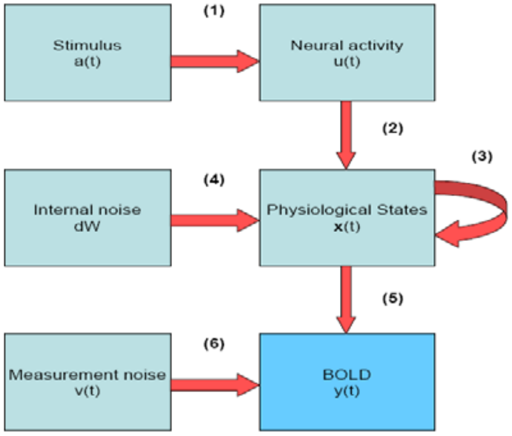


Figure 3.1: Overview diagram of hemodynamic models.

Bayesian averaging with split-1/2 resampling loop
to establish generalizability and reproducibility

$$R(M) = -\frac{1}{K}\sum_{i=1}^{K}\int p(\theta \mid D_i^1, M)\log\frac{p(\theta \mid D_i^1, M)}{p(\theta \mid D_i^2, M)}d\theta,$$

BOLD hemodynamics: Resampling-Bayes model selection

Model A: Sustained neural input   vs   Model B: Fading input

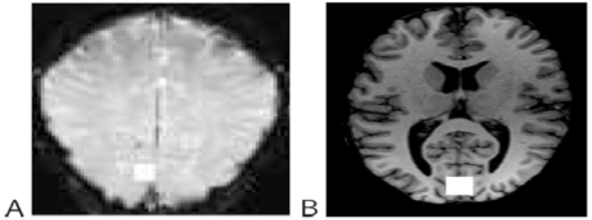


Figure 3: Regions of interest, marked with white squares. (A) Data set 1; T2* weighted image slice parallel to the calcarine sulcus. (B) Data set 2; MPRAGE (magnetization prepared rapid gradient echo) horizontal slice.

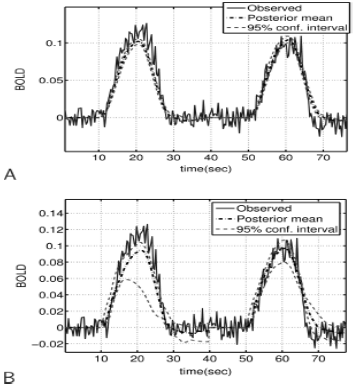


Figure 4: Prediction of data set 1. (A) Model A. (B) Model B. Note that the confidence interval is an empirical confidence interval for the mean prediction, based on the MCMC samples.

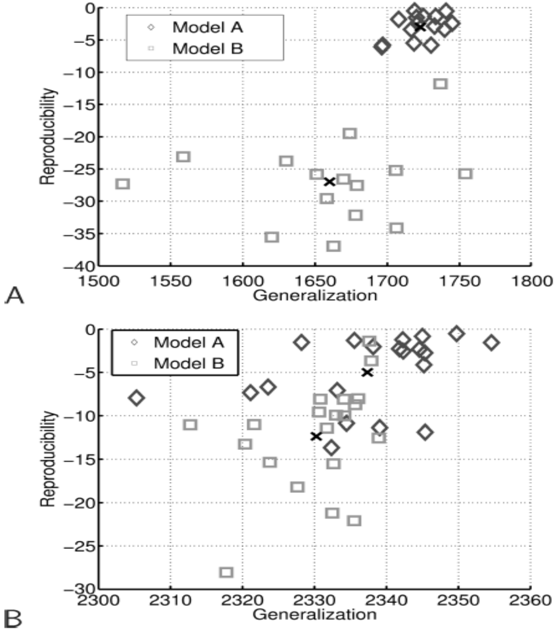


Figure 6: Generalization and reproducibility for real data; crosses mark the mean. (A) Data set 1. (B) Data set 2.

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

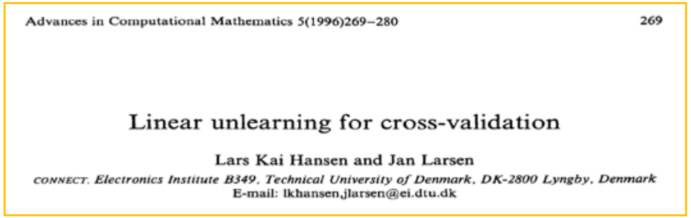
Asymptotic theory investigates the sampling fluctuations in the limit N -> ∞

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...



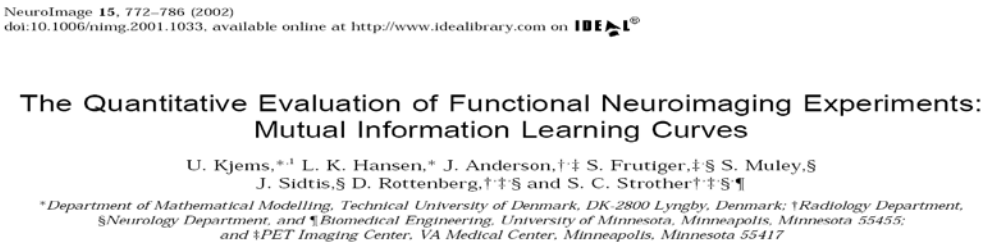
Advances in Computational Mathematics 5(1996)269–280 269

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



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

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

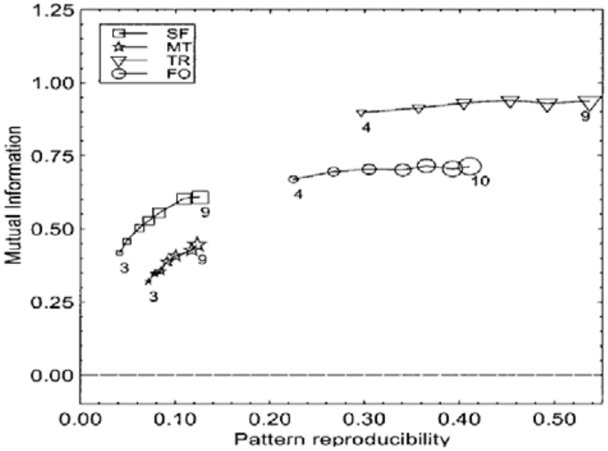


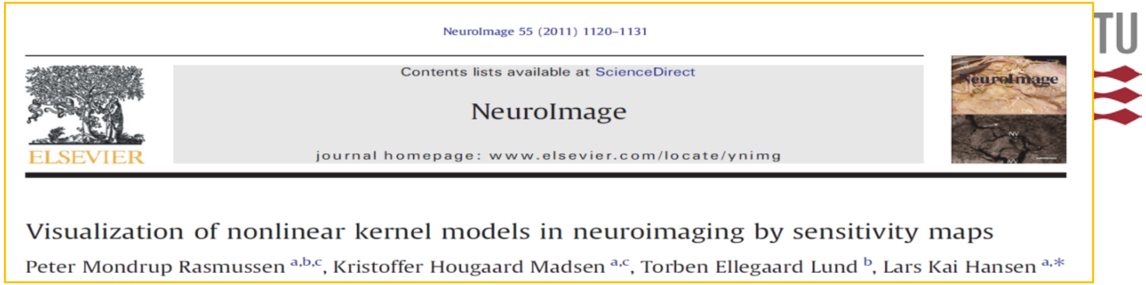
FIG. 3. Plot of scan/label mutual information versus reproducibility signal/noise for the four data sets, for varying numbers of subjects in the training set. There were 2 labels/4 scans per subject (balanced data set; Setup 1, Table 1) corresponding to the dashed solid line in Fig. 4. We see that both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

Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers. 1999 Mar 26;10(3):61-74.

Rasmussen, P. M., Madsen, K. H., Lund, T. E., Hansen, L. K. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. *NeuroImage*, *55*(3):1120-1131 (2011).

‘SVM mind reading’



NeuroImage 55 (2011) 1120–1131

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Visualization of nonlinear kernel models in neuroimaging by sensitivity maps

Peter Mondrup Rasmussen [a,b,c], Kristoffer Hougaard Madsen [a,c], Torben Ellegaard Lund [b], Lars Kai Hansen [a,*]

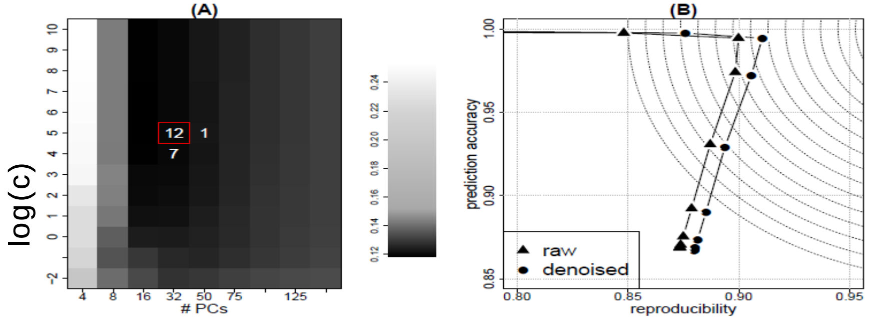
Non-linear kernel machines, SVM

Local voting +/-

$$s(n') \approx \sum_{n=1}^{N} \alpha(n) K(x_n, x_{n'})$$

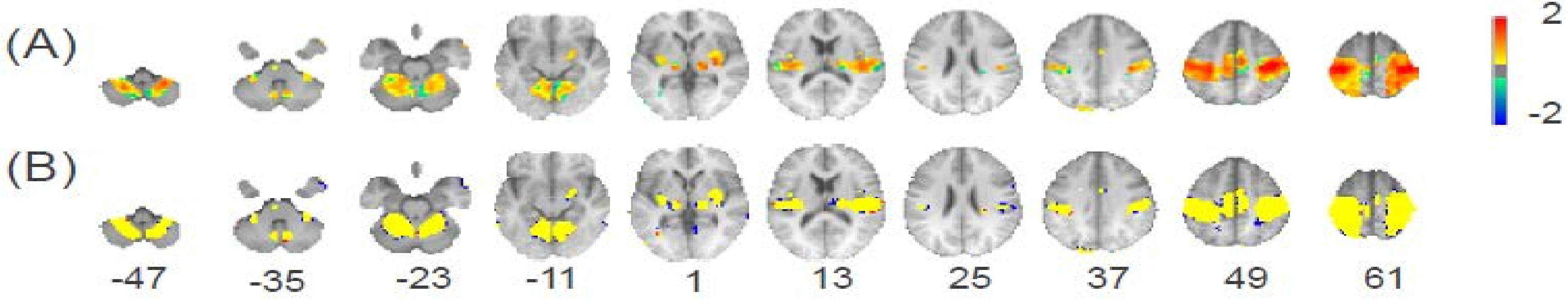
$$K(x_n, x_{n'}) = \exp\left\{-\frac{\|x_n - x_{n'}\|^2}{2c}\right\}$$

Does denoising help fMRI decoding?



(A)Comparison of resampling
z-score = z-kPCA – z-Raw

(B) FDR corrected:
yellow: consensus,
blue: only kPCA,
red: only raw



PM Rasmussen, TJ Abrahamsen, KH Madsen, LK Hansen: Nonlinear denoising and analysis of neuroimages with kernel principal component analysis and pre-imageestimation, NeuroImage 60(3):1807-1818 (2012).

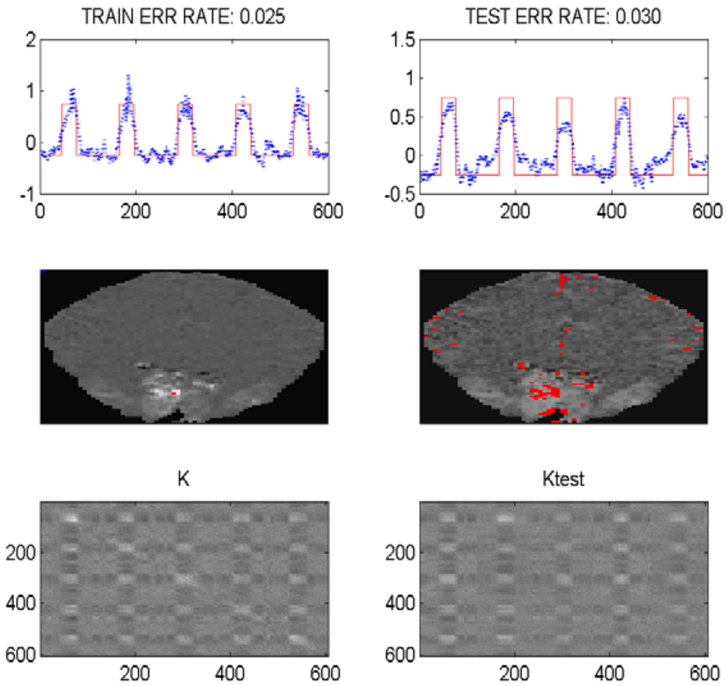
Initial dip data: Visual stimulus (TR 0.33s)

- Gaussian kernel, sparse kernel regression

- Sensitivity map computed for whole slice

- Error rates about 0.03

- How to set
  - Kernel width?
  - Sparsity?



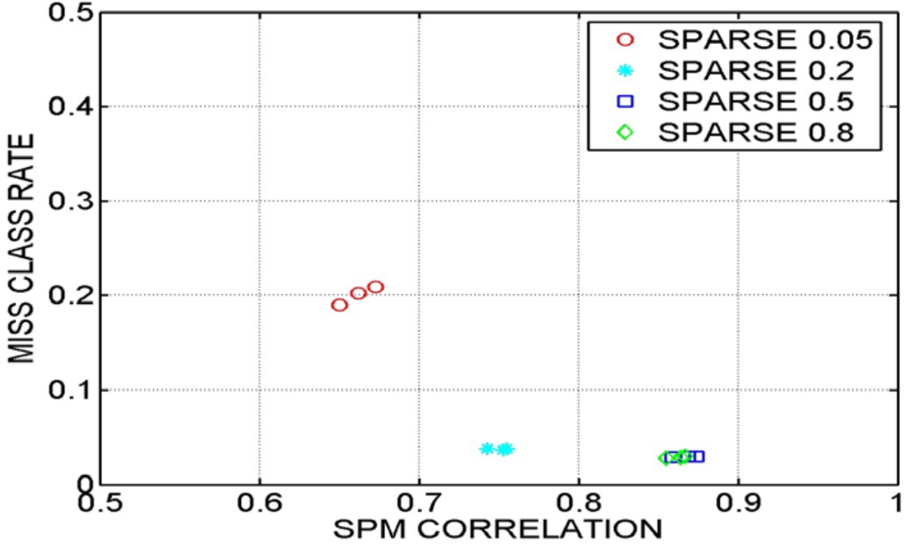
Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers. 1999 Mar 26;10(3):61-74.

Rasmussen, P. M., Madsen, K. H., Lund, T. E., Hansen, L. K. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. *NeuroImage*, *55*(3):1120-1131 (2011).

Initial dip data: Visual stimulus (TR 0.33s)

Select hyperparameters of kernel machine using NPAIRS resampling

- Degree of sparsity
- Kernel scale parameter



Sensitivity maps for non-linear kernel regression

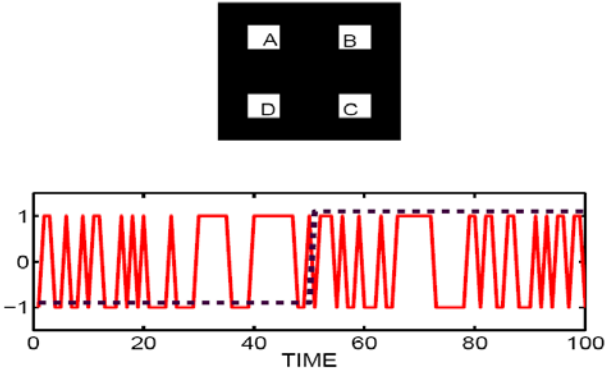


Fig. 1. XOR-image set define by four activated regions (A,B,C,D). Initially we let regions (A,B,D) be activated by random sequence taking values ±1, as shown in example in the bottom panel (full curve). The target signal, also taking values $t_n = pm1$, and is also indicated in the bottom panel (dashed line). The region (C) is activated with an XOR-sequence relative to (A) and t_n, so that $C_n = A_n * t_n$, hence, in the active state the two regions (A,C) are randomly, but identically activated, while in the resting condition, they are random, but opposite

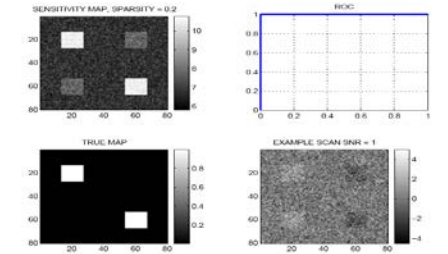


Fig. 2. XOR-image set define by four activated regions. The results of analyzing a image set with $N = 400$ examples. The image signal-to-noise ratio is $SNR = 1$, i.e., the additive noise is unit variance. The target function has in addition been contaminated by 10% random label noise. The four subplots show: The sensitivity map (upper left), the near-perfect receiver operating curve (ROC, upper right), the true activation map (lower left), and a random example of the simulated brain images. We modeled the data set using the kernel regression method. The linear model was estimated using the so-called least angle elastic net method (LARSEN) with a degree of sparsity of 0.2, i.e., using $N = 0.2 \times 400 = 80$ support vectors.

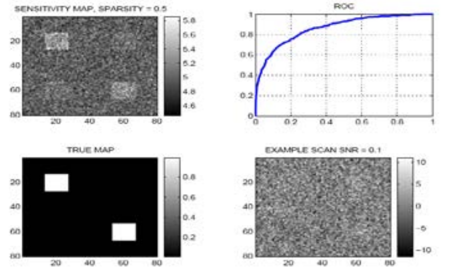
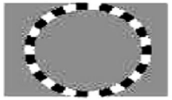


Fig. 4. XOR-image set define by four activated regions. Similar to figure 2, however the image signal-to-noise ratio is $SNR = 0.1$.

Non-linearity in fMRI?

Visual stimulus ”XOR”: half checker board no/left/right/both



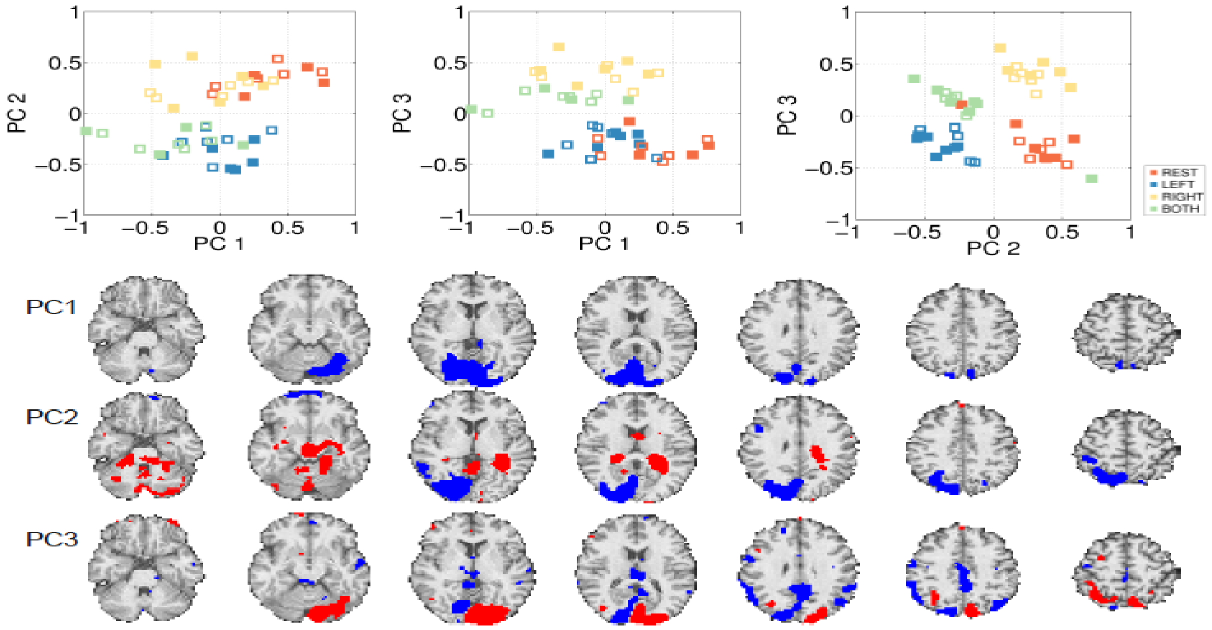
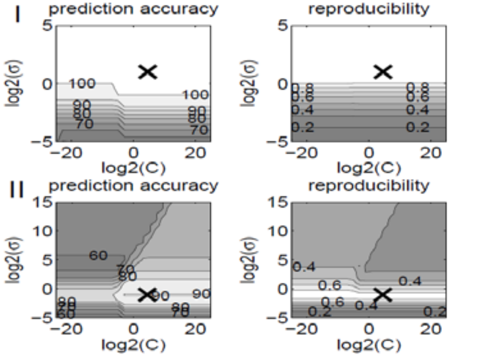


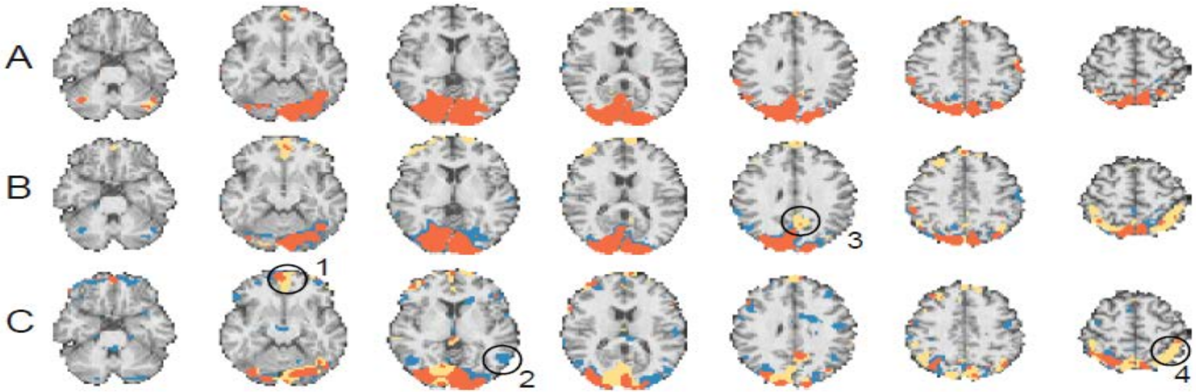
Figure 1: PCA analysis of the fMRI data set. An example of the three first PCs estimated from the training set in a NPAIRS split. The scatter plots show both training (filled markers) and test examples projected onto the PCs. The blue and red voxels on the brain slices corresponds to negative and positive PC loadings respectively. The maps are thresholded to show the 5 upper positive and negative percentiles.

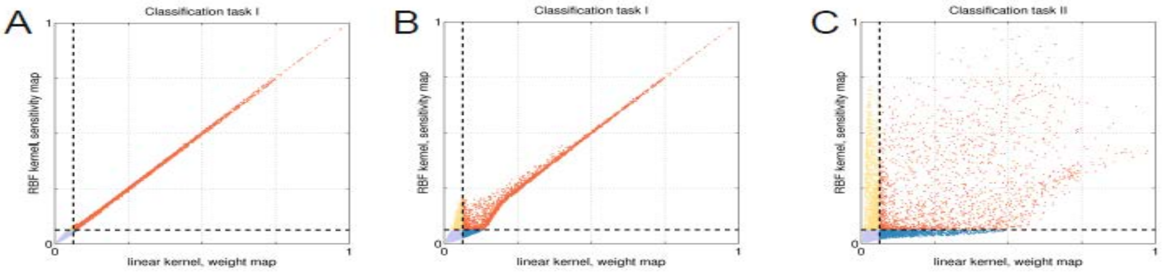
PM Rasmussen, KH Madsen, TE Lund, LK Hansen. Visualization of nonlinear kernel models in neuroimaging by sensitivity maps. *NeuroImage*, *55*(3):1120-1131 (2011).

Non-linearity in fMRI – detecting networks



PM Rasmussen et al. NeuroImage 55 (2011) 1120-





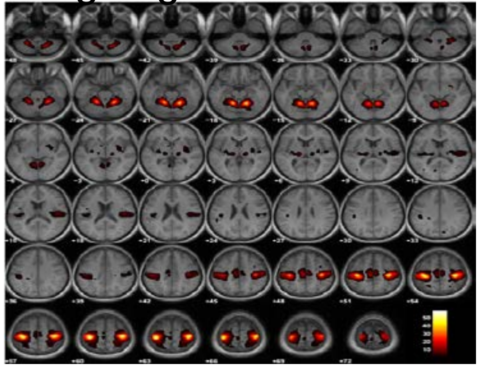
A: Easy problem-
(Left vs Right) and RBF
kernel is wide … i.e.
similar to linear kernel

B: Easy problem-
Pars optimized to yield
the best P-R

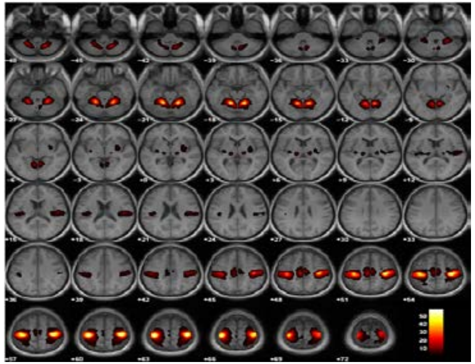
C : Hard XOR problem
Pars optimized to yield
The best P-R

Consistency across models (left-right finger tapping)

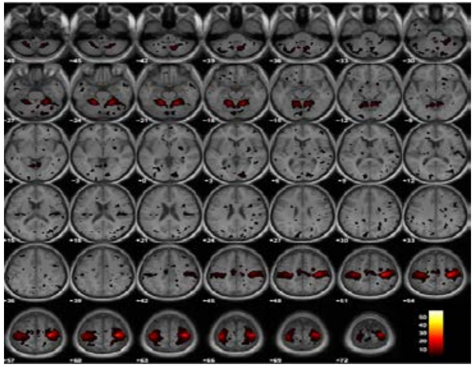
LogReg



SVM



RVM



Sparsity increasing

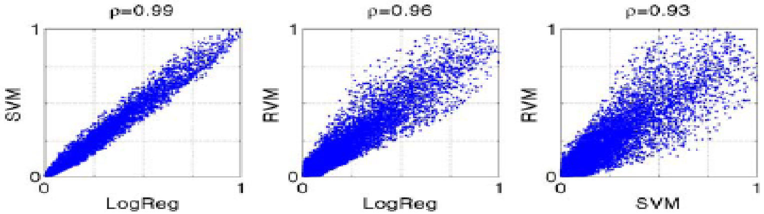
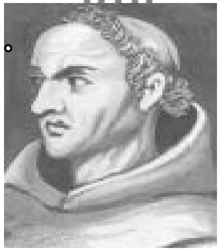


Figure 7: fMRI fingertapping experiment - consensus analysis. The plots show the extend of consensus in the average rSPI among the three models. The rSPI for LogReg was scaled by its maximum value. Hereafter the rSPIs from the SVM and RVM were transformed to match the histogram of that of LogReg. Correlation coefficients between histograms are found on top of the plots.

Rasmussen, P. M., Hansen, L. K., Madsen, K. H., Churchill, N. W., & Strother, S. C. (2012). Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition*, *45*(6), 2085-2100.

Conclusions

Do not multiply causes!



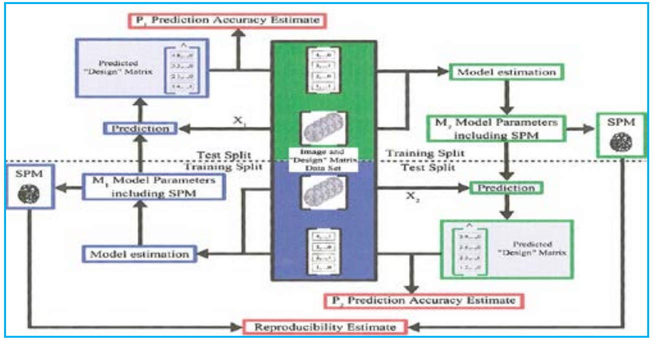
Scientific applications of machine learning have two equally important aims

- Generalizability
- Reproducible interpretation

We can visualize general ML functions with perturbation based methods (saliency maps, sensitivity maps etc)

NPAIRS split-half based framework for optimization of generalizability and robustness of visualizations

More complex mechanisms may be revealed with non-linear detectors - can still be visualized...


P_1 Prediction Accuracy Estimate
Predicted "Design" Matrix
Model estimation
X_1
Prediction
M_2 Model Parameters including SPM
SPM
Test Split
Training Split
Image and "Design" Matrix Data Set
Training Split
Test Split
SPM
M_1 Model Parameters including SPM
X_2
Prediction
Predicted "Design" Matrix
Model estimation
P_2 Prediction Accuracy Estimate
Reproducibility Estimate