Machine Learning and Sparsity

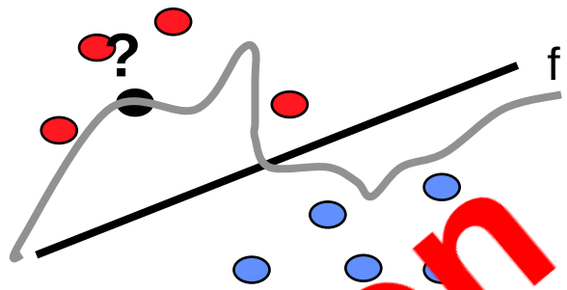
Klaus-Robert Müller !!et al.!!

Today‘s Talk

- sensing, sparse models and generalization
- interpretabilty and sparse methods
- explaining for nonlinear methods

Sparse Models & Generalization?

Machine Learning in a nutshell



Typical scenario: learning from data

- given data set **X** and labels **Y** (generated by some joint probabilty distribution p(x,y))
- **LEARN/INFER** underlying **unknown** mapping

Y = f(X); error(f) = $\frac{1}{N}\sum_{i=1}^{N} |f(\mathbf{x}_i) - y_i|^2 + \lambda\, |\mathbf{Pf}|^2$

Example: understand chemical compound space, distinguish brain states …

BUT: how to do this optimally with good performance on **unseen** data?

Most popular techniques: **kernel methods** and (deep) **neural networks**.

Generalization

Machine Learning for chemical compound space

Ansatz:

$$\{Z_I, \mathbf{R}_I\} \xmapsto{\text{ML}} E$$

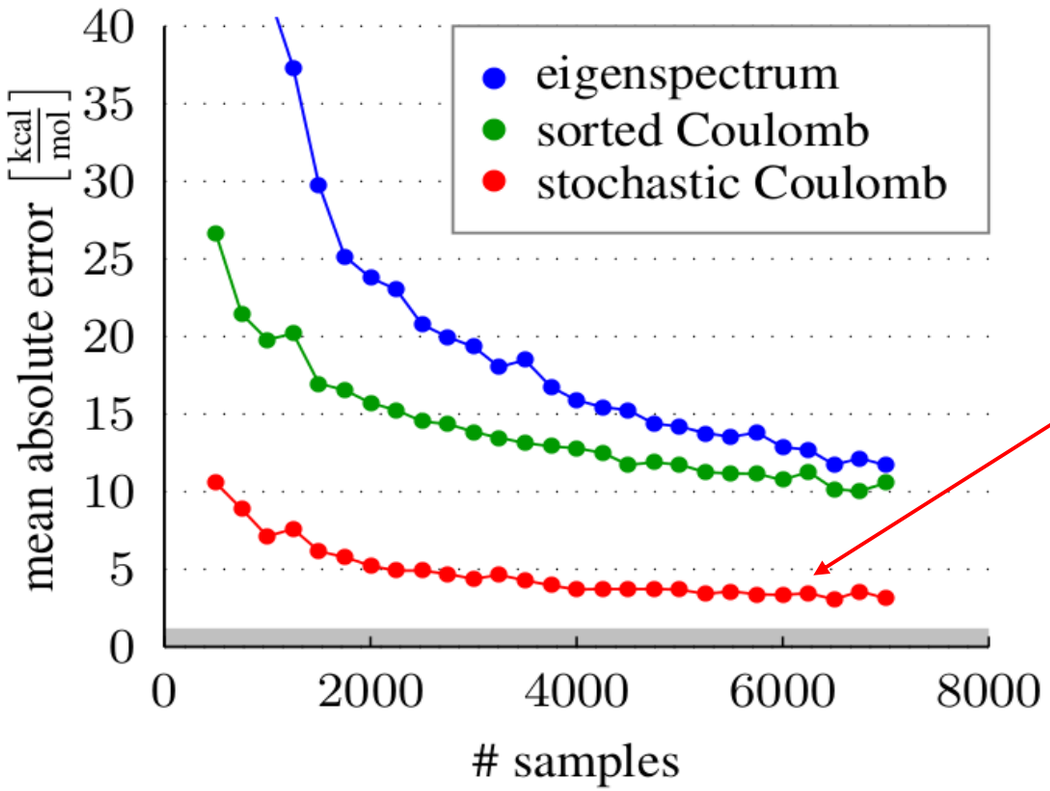
instead of

$$\hat{H}(\{Z_I, \mathbf{R}_I\}) \xmapsto{\Psi} E$$

$$\hat{H}\Psi = E\Psi$$

[Rupp, Tkatchenko, Müller & v Lilienfeld 2012, Hansen et al 2013, 2015, Snyder et al 2012, 2015, Montavon et al 2013]

Predicting Energy of small molecules with ML: Results



March 2012 KRR
Rupp et al., PRL
9.99 kcal/mol
(kernels + eigenspectrum)

2013/2015 KRR et al.
Hansen et al., JCTC
3.51 kcal/mol (Coulomb matrices)
But L_1 7.8 kcal/mol

2015 Hansen et al 1.3kcal/mol at **10 million** times faster than the state of the art

Prediction considered chemically accurate when MAE is below
1 kcal/mol

Dataset available at http://quantum-machine.org

Compressed Sensing and Generalization are different goals!

L2 better at Generalization unless truth is sparse!

[cf. Ng 2004, Braun et al. 2008]

Sparse Model = Interpretable Model?

Linear Models

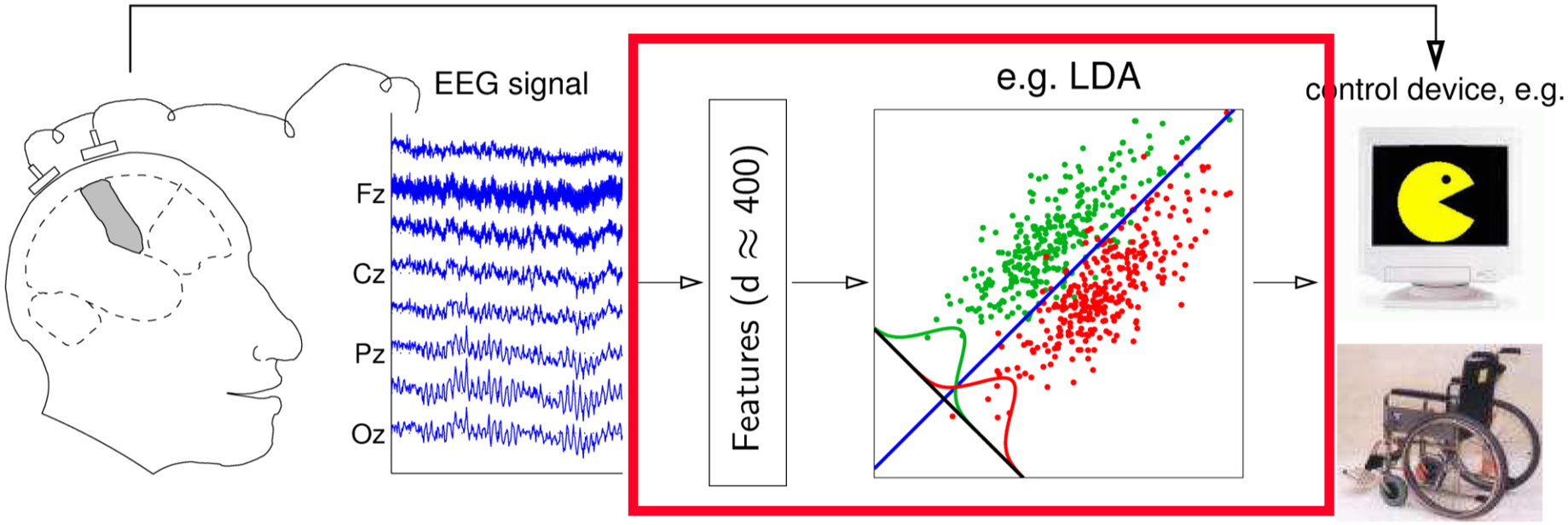
$$f(x) = w^T x$$

$$\text{error(f)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}|f(\mathbf{x}_i) - y_i|^2 + \lambda\,|\mathbf{Pf}|^2$$

Regularizer $||w||_1$

Neuroscience

Noninvasive Brain-Computer Interface



DECODING

BCI: Translation of human intentions into a technical control signal
without using activity of muscles or peripheral nerves

Brain Computer Interfacing: ‚Brain Pong‘

Berlin Brain Computer Ínterface

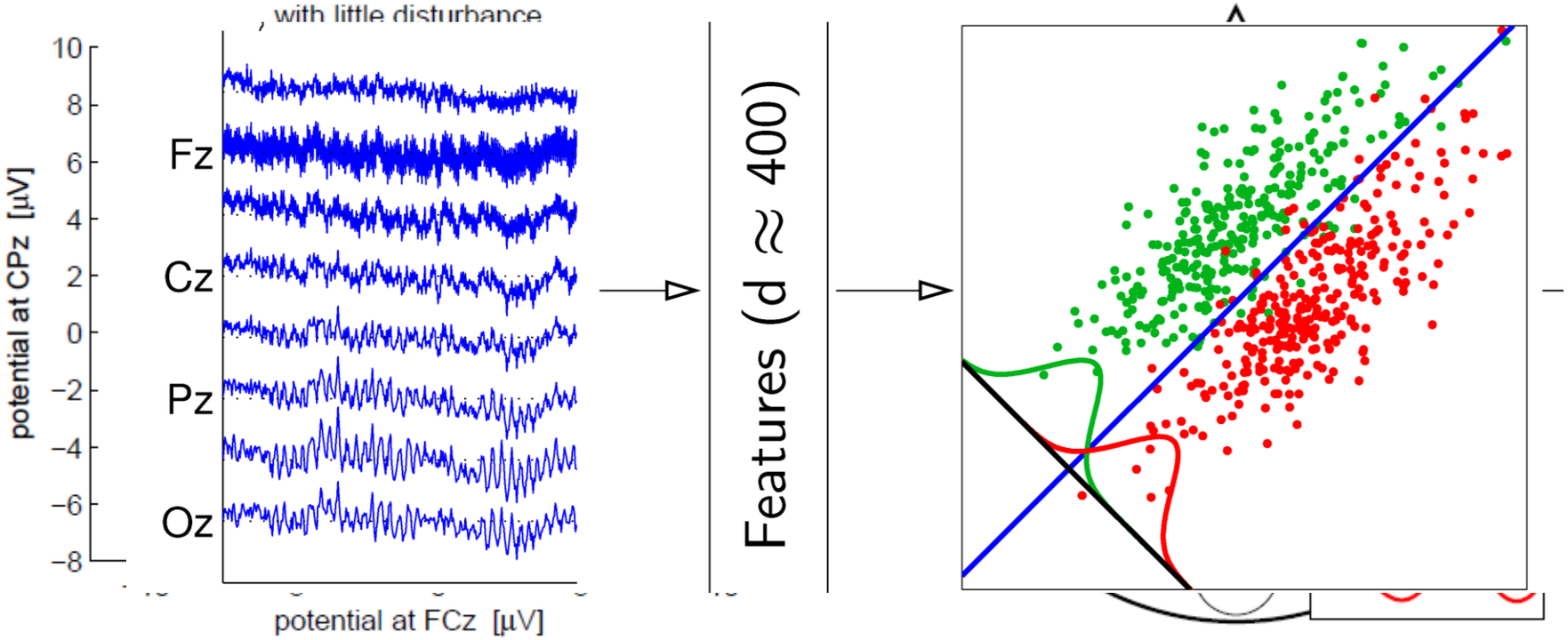
- ML reduces patient training from 300h -> 5min (BBCI)

Applications

- help/hope for patients (ALS, stroke…)
- neuroscience
- neurotechnology (**better** video coding in cooperation with HHI, gaming, monitoring, driving)

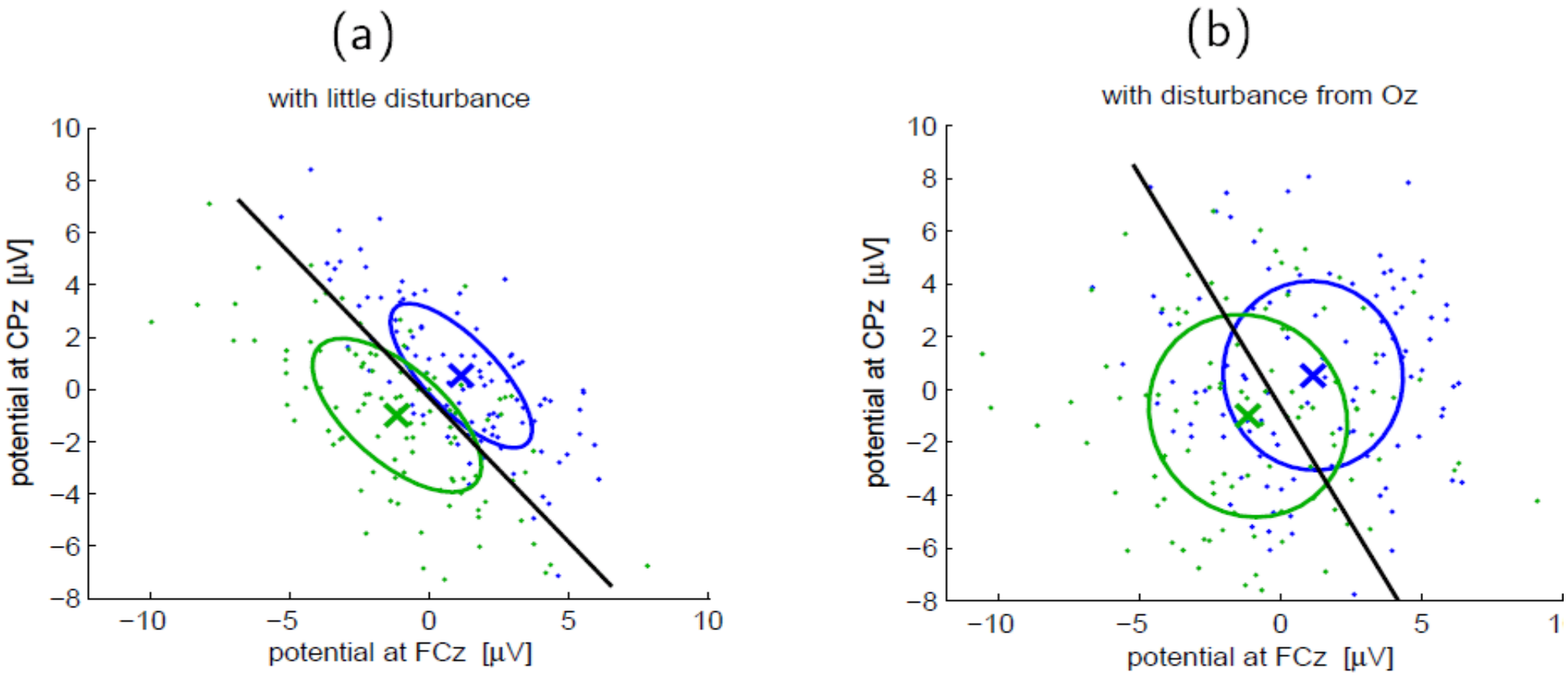
Breakthrough: *›let the machines learn‹*

Understanding spatial filters



$W^T x = \tilde{s}$	$x = As + \epsilon$
Latent factors $\tilde{s}$	Activation patterns A

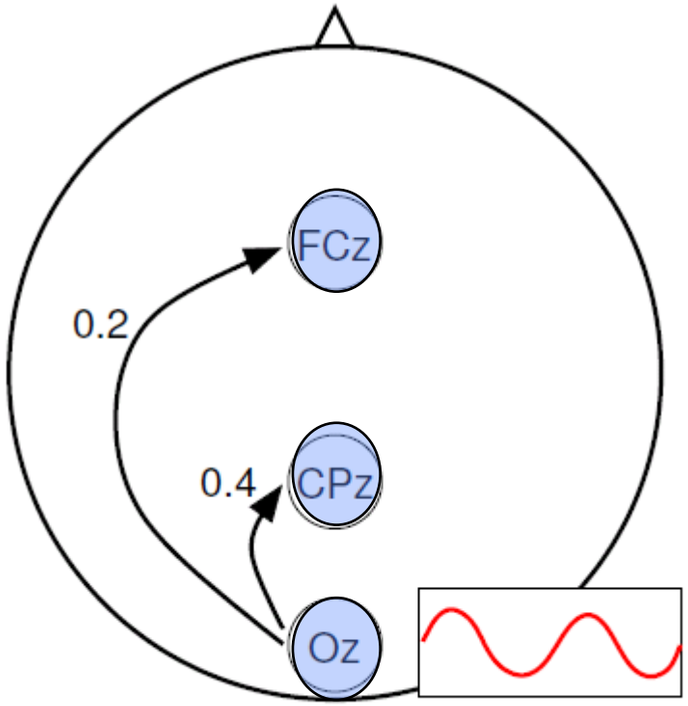
Understanding spatial filters II



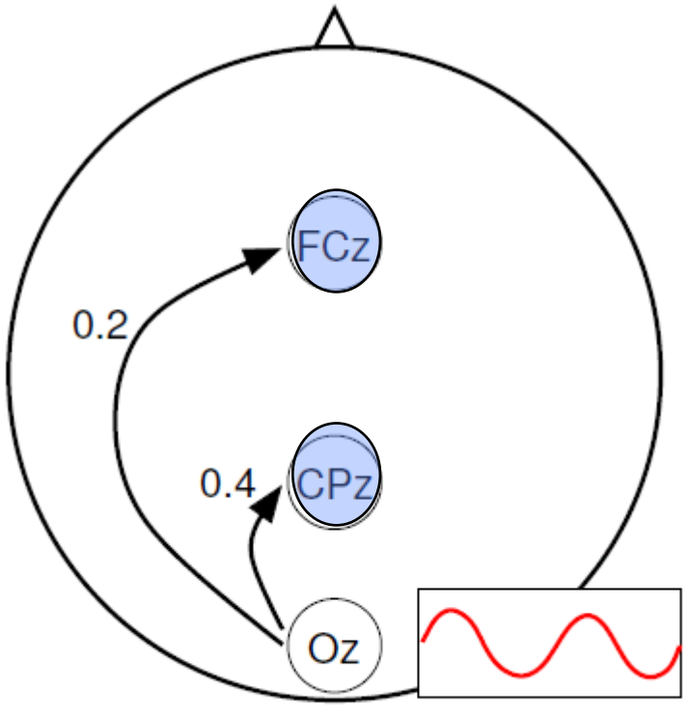
Two channel classification of (a): 15% error, (b): 37% error

When disturbing channel Oz is added to the data (3D): 16% error. Here, channel Oz is required for good classification although itself is not discriminative.

Understanding spatial filters



Filter W

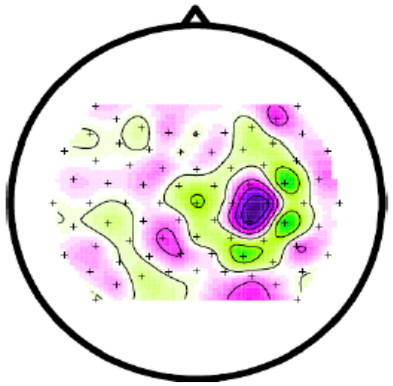


Pattern

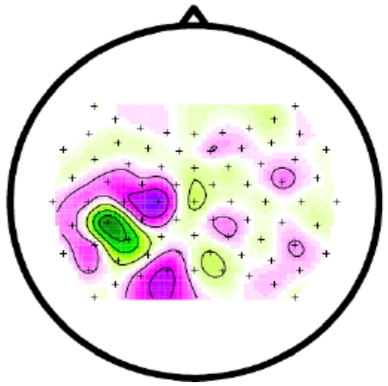
CSP Analysis

Filter W

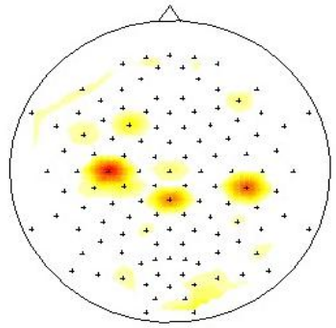
CSP filter **'left'**



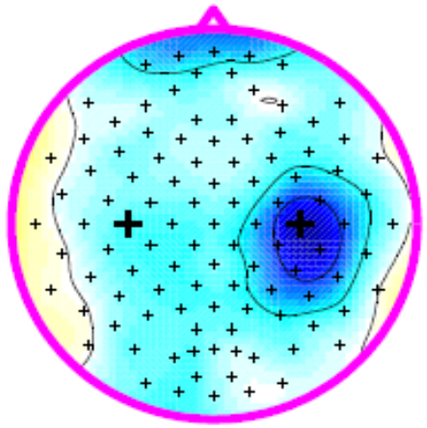
CSP filter **'right'**

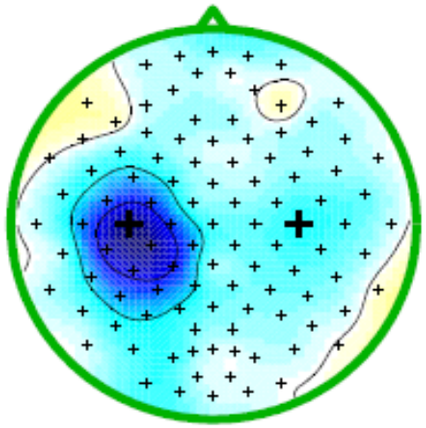


Sparse Filter



Pattern $A \propto \Sigma_x W$

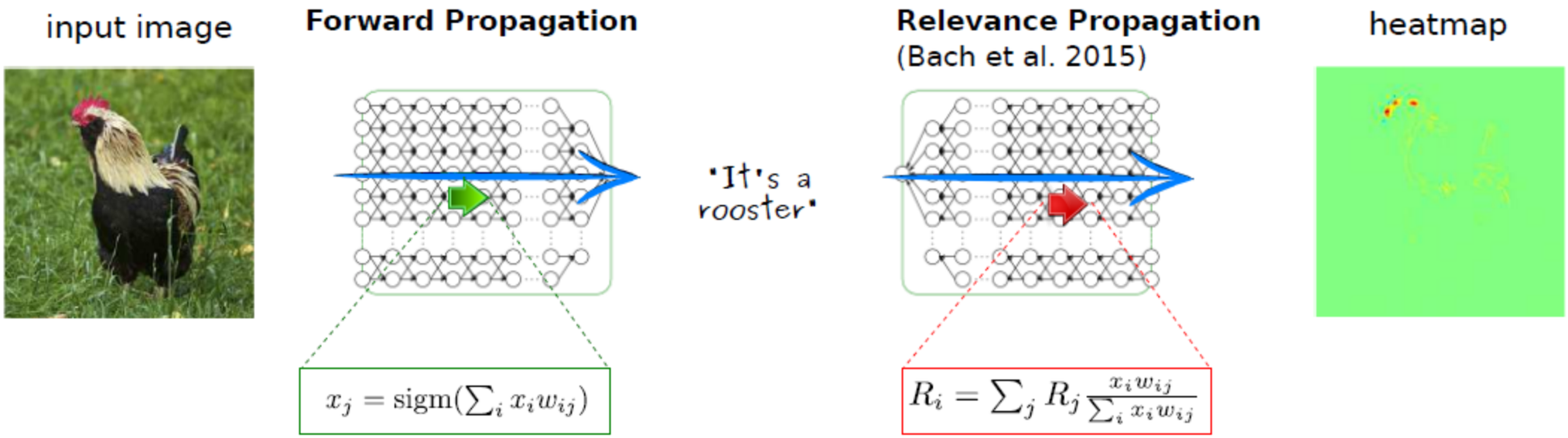




[cf. Blankertz et al 2011, Haufe et al. 2014]

Interpretabilty in Nonlinear Methods

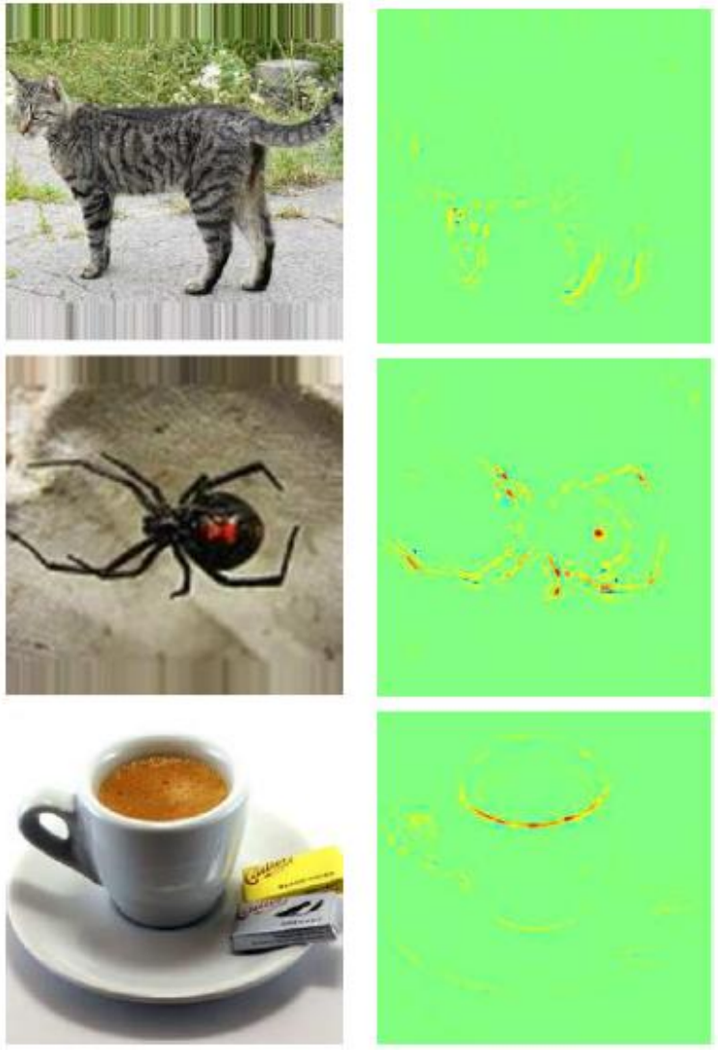
Explaining Predictions Pixel-wise



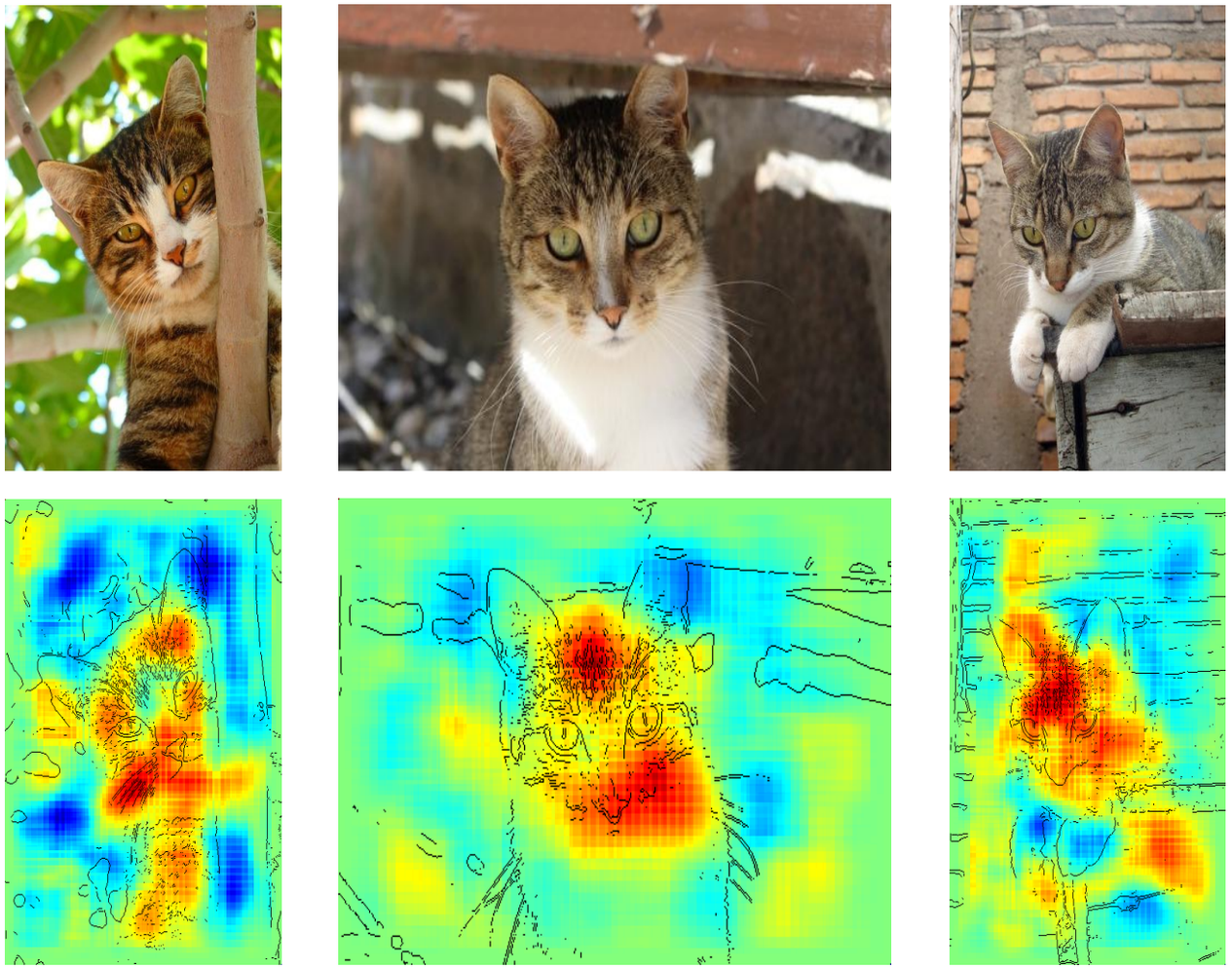
- The total relevance $\sum_p R_p$ (number of red pixels in the heatmap) corresponds to the amount of evidence $f(\mathbf{x})$ for the predicted class. ($\Rightarrow$ Relevance does not get *lost* or *created* out of nothing.)
- This equivalence is ensured by the *layer-wise conservation* property of the relevance propagation formula.

[Bach, Binder, Klauschen, Montavon, Müller & Samek, PLOS ONE 2015]

Explaining Predictions Pixel-wise

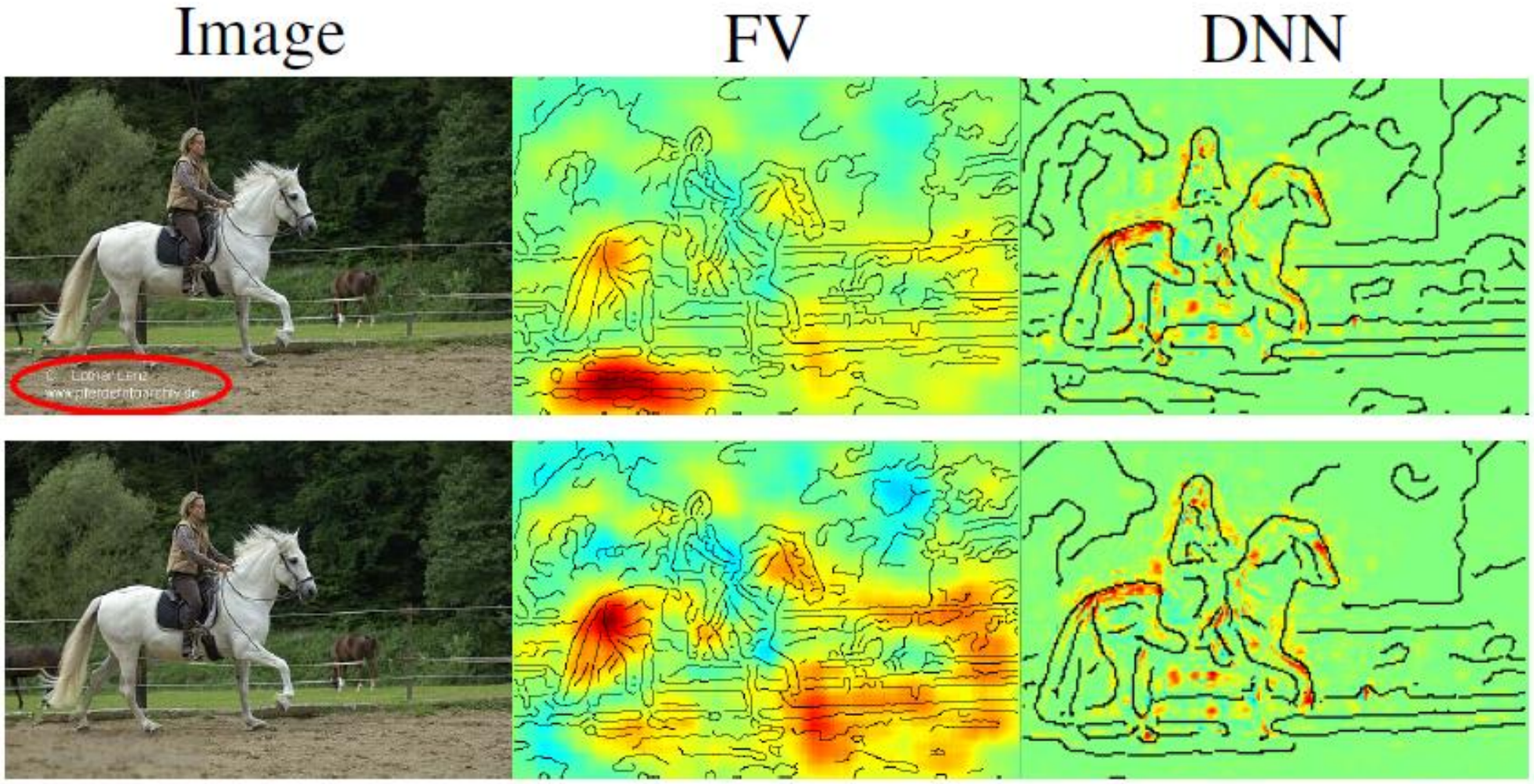


Neural networks



Kernel methods

Understanding Models is only possible if we explain



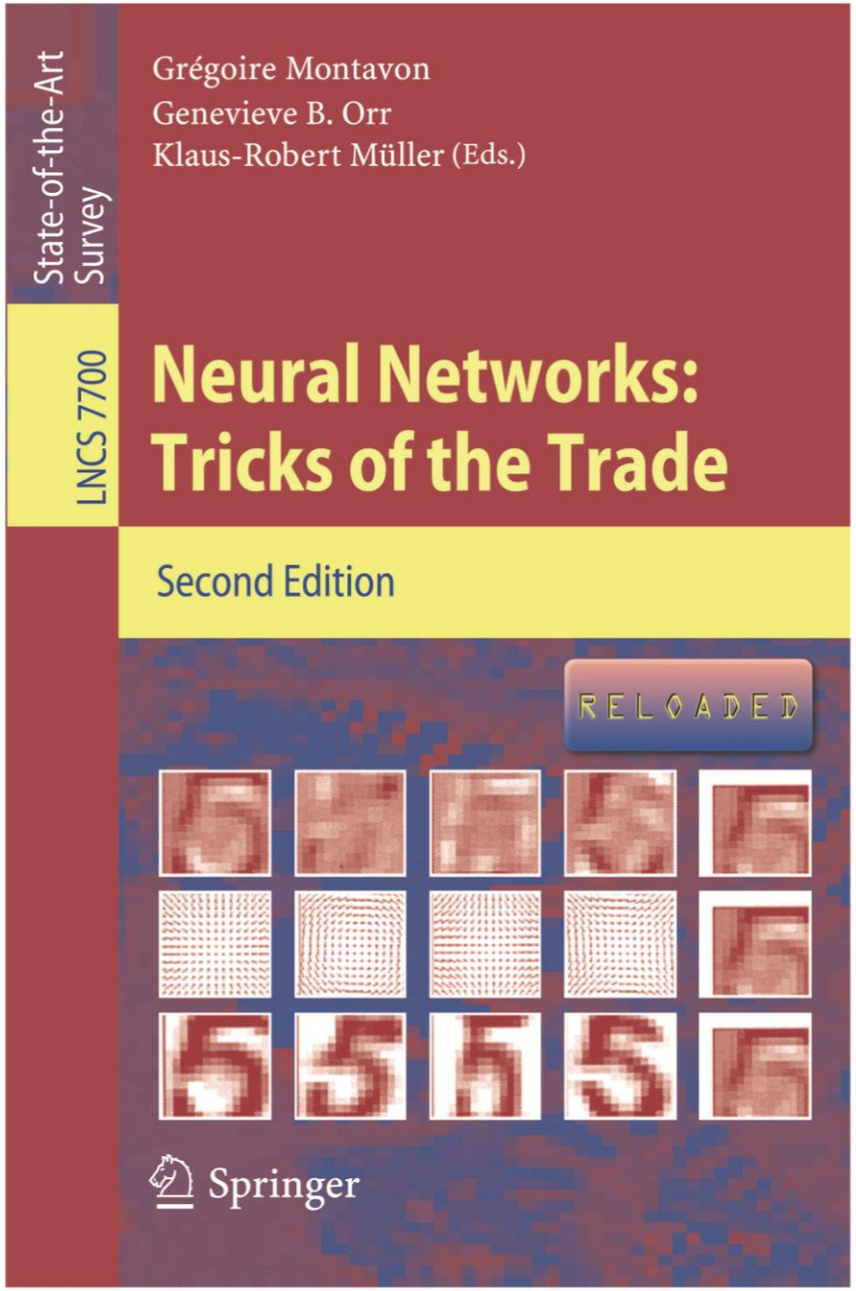
Fischer

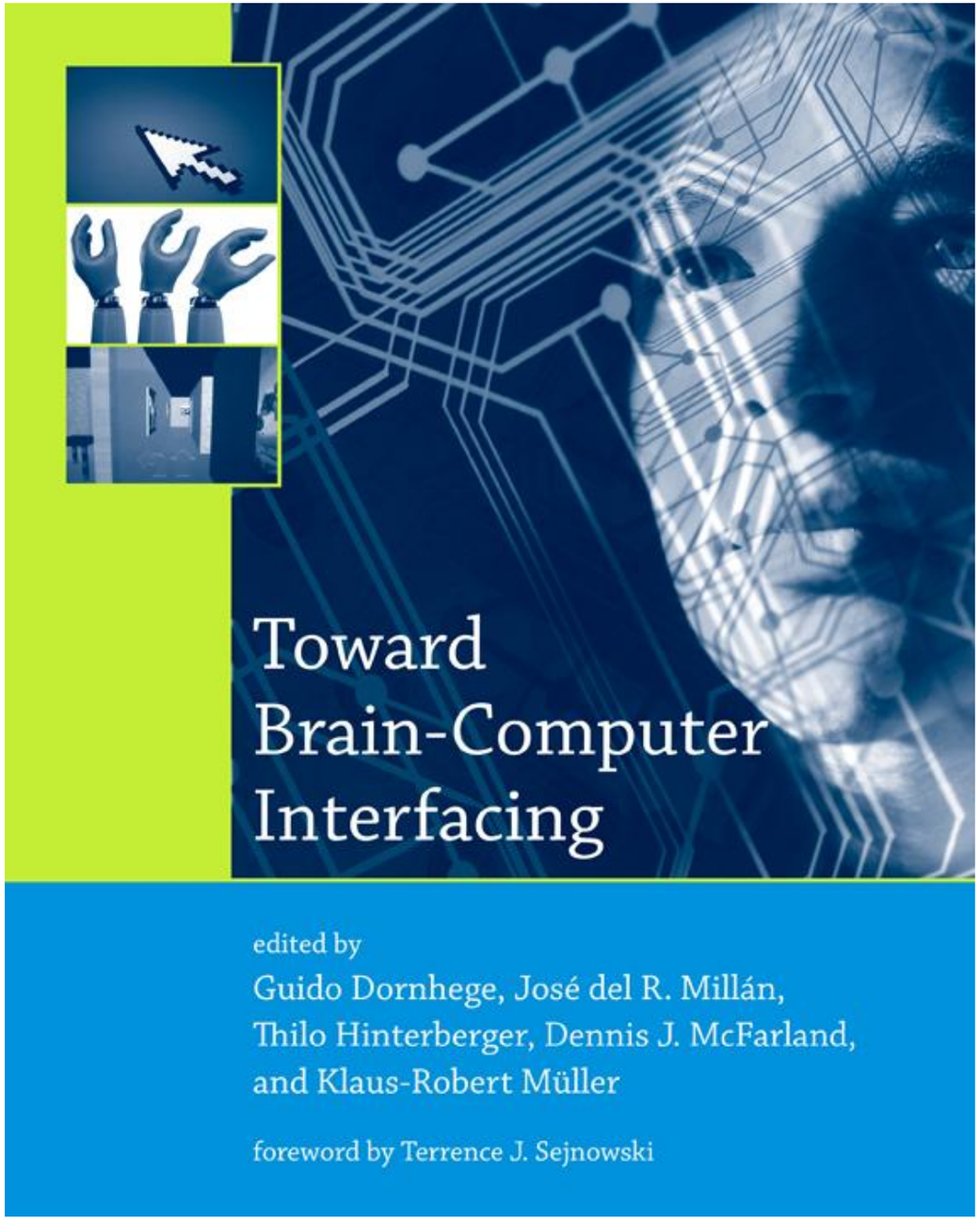
Neural networks

Conclusion

- ML & modern data analysis of central importance in daily life, sciences & industry
- ML and compressed sensing follow different goals: sensing vs **generalization**!
- Sparse models are not necessarily good for understanding: example sparse linear models and Brain Computer Interface application.
- challenge: learn about application from **nonlinear** ML model: towards better **understanding**

See also: www.quantum-machine.org, www.bbci.de


State-of-the-Art Survey
LNCS 7700
Grégoire Montavon
Genevieve B. Orr
Klaus-Robert Müller (Eds.)
Neural Networks: Tricks of the Trade
Second Edition
RELOADED
Springer


Toward Brain-Computer Interfacing
edited by
Guido Dornhege, José del R. Millán, Thilo Hinterberger, Dennis J. McFarland, and Klaus-Robert Müller
foreword by Terrence J. Sejnowski

Further Reading I

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7).

Bießmann, F., Meinecke, F. C., Gretton, A., Rauch, A., Rainer, G., Logothetis, N. K., & Müller, K. R. (2010). Temporal kernel CCA and its application in multimodal neuronal data analysis. *Machine Learning*, *79*(1-2), 5-27.

Blum, L. C., & Reymond, J. L. (2009). 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. Journal of the American Chemical Society, 131(25), 8732-8733.

Braun, M. L., Buhmann, J. M., & Müller, K. R. (2008). On relevant dimensions in kernel feature spaces. *The Journal of Machine Learning Research*, *9*, 1875-1908

Hansen, Katja, Grégoire Montavon, Franziska Biegler, Siamac Fazli, Matthias Rupp, Matthias Scheffler, O. Anatole von Lilienfeld, Alexandre Tkatchenko, and Klaus-Robert Müller. "Assessment and validation of machine learning methods for predicting molecular atomization energies." *Journal of Chemical Theory and Computation* 9, no. 8 (2013): 3404-3419.

Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., von Lilienfeld, O. A., Müller, K. R., & Tkatchenko, A. (2015). Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space, J. Phys. Chem. Lett. 6, 2326−2331.

Harmeling, S., Ziehe, A., Kawanabe, M., & Müller, K. R. (2003). Kernel-based nonlinear blind source separation. *Neural Computation*, *15*(5), 1089-1124.

Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, KR Muller (1999), Fisher discriminant analysis with kernels, Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop, 41-48.

Kloft, M., Brefeld, U., Laskov, P., Müller, K. R., Zien, A., & Sonnenburg, S. (2009). Efficient and accurate lp-norm multiple kernel learning. In *Advances in neural information processing systems* (pp. 997-1005).

Further Reading II

Laskov, P., Gehl, C., Krüger, S., & Müller, K. R. (2006). Incremental support vector learning: Analysis, implementation and applications. *The Journal of Machine Learning Research*, *7*, 1909-1936 Mika, S., Schölkopf, B., Smola, A. J., Müller, K. R., Scholz, M., & Rätsch, G. (1998). Kernel PCA and De-Noising in Feature Spaces. In *NIPS* (Vol. 4, No. 5, p. 7).

Müller, K. R., Mika, S., Rätsch, G., Tsuda, K., & Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, *12*(2), 181-201.

Montavon, G., Braun, M. L., & Müller, K. R. (2011). Kernel analysis of deep networks. *The Journal of Machine Learning Research*, *12*, 2563-2581.

Montavon, Grégoire, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole V. Lilienfeld, and Klaus-Robert Müller. "Learning invariant representations of molecules for atomization energy prediction." In *Advances in Neural Information Processing Systems*, pp. 440-448. 2012.

Montavon, G., Braun, M., Krueger, T., & Muller, K. R. (2013). Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment. *IEEE Signal Processing Magazine, 30*(4), 62-74.

Montavon, G., Orr, G. & Müller, K. R. (2012). Neural Networks: Tricks of the Trade, Springer LNCS 7700. Berlin Heidelberg.

Montavon, Grégoire, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. "Machine learning of molecular electronic properties in chemical compound space." New Journal of Physics 15, no. 9 (2013): 095003.

Snyder, J. C., Rupp, M., Hansen, K., Müller, K. R., & Burke, K. Finding density functionals with machine learning. Physical review letters, 108(25), 253002. 2012.

Further Reading III

Pozun, Z. D., Hansen, K., Sheppard, D., Rupp, M., Müller, K. R., & Henkelman, G., Optimizing transition states via kernel-based machine learning. The Journal of chemical physics, 136(17), 174101. 2012 .

K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, How to represent crystal structures for machine learning: Towards fast prediction of electronic properties Phys. Rev. B 89, 205118 (2014)

Rätsch, G., Onoda, T., & Müller, K. R. (2001). Soft margins for AdaBoost. *Machine learning*, *42*(3), 287-320.

Rupp, M., Tkatchenko, A., Müller, K. R., & von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. Physical review letters, 108(5), 058301.

Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. Neural computation, 10(5), 1299-1319.

Smola, A. J., Schölkopf, B., & Müller, K. R. (1998). The connection between regularization operators and support vector kernels. *Neural networks*, *11*(4), 637-649.

Schölkopf, B., Mika, S., Burges, C. J., Knirsch, P., Müller, K. R., Rätsch, G., & Smola, A. J. (1999). Input space versus feature space in kernel-based methods. IEEE Transactions on Neural Networks, 10(5), 1000-1017.

Tsuda, K., Kawanabe, M., Rätsch, G., Sonnenburg, S., & Müller, K. R. (2002). A new discriminative kernel from probabilistic models. *Neural Computation*, *14*(10), 2397-2414.

Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T., & Müller, K. R. (2000). Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, *16*(9), 799-807.

.