Machine Learning and Big Data

Klaus-Robert Müller et al.

Election of the Pope: 2005

[from Wiegand]

Election of the Pope: 2013

[from Wiegand]

Today‘s Talk

Remarks

- big data vs. small data (expensive!)
- Machine Learning & Database Management Systems:

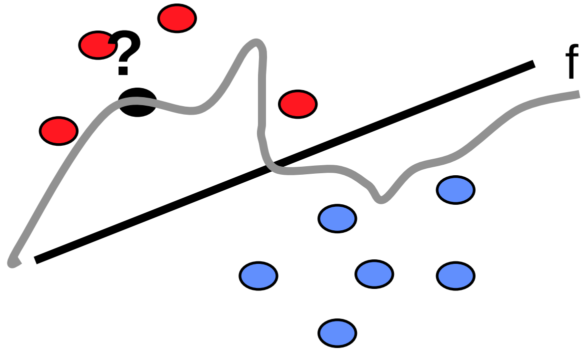
  Berlin Big Data Center

- ML: Kernel Methods and Deep networks

Applications of Big Data

- big data in neuroscience: BCI et al.
- physics & materials

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

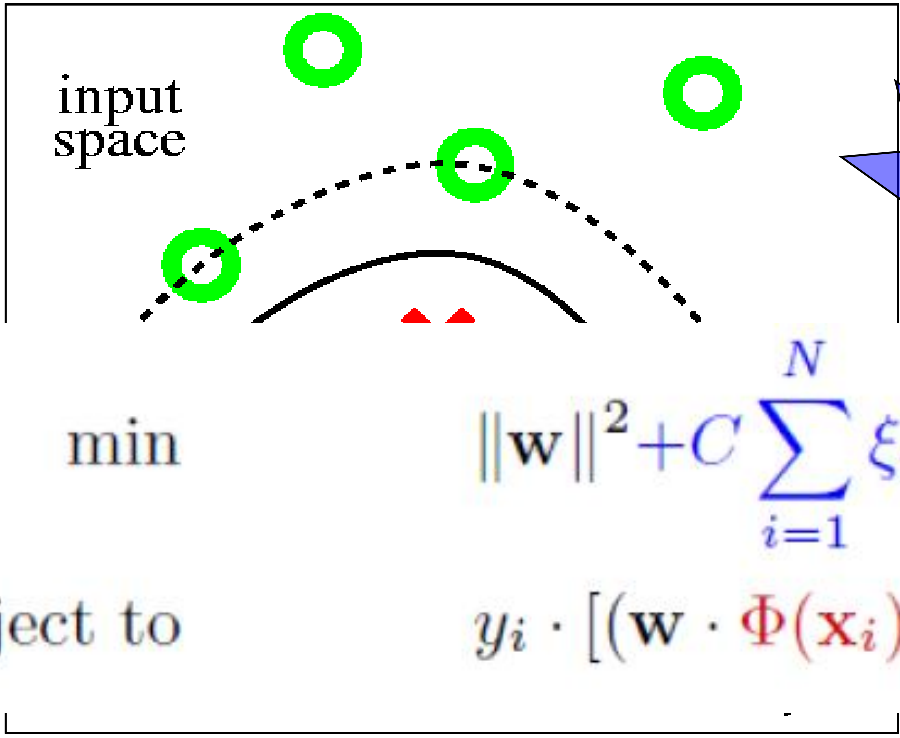
$$Y = f(X)$$

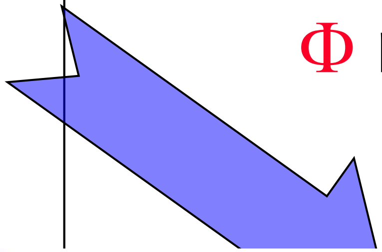
Example: cancer detection, find trends in social media, distinguish brain states ...

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

Support Vector Machines in a nutshell

$$f(\mathbf{x}) = \operatorname{sgn}\left(\mathbf{w} \cdot \Phi(\mathbf{x}) + b\right)$$

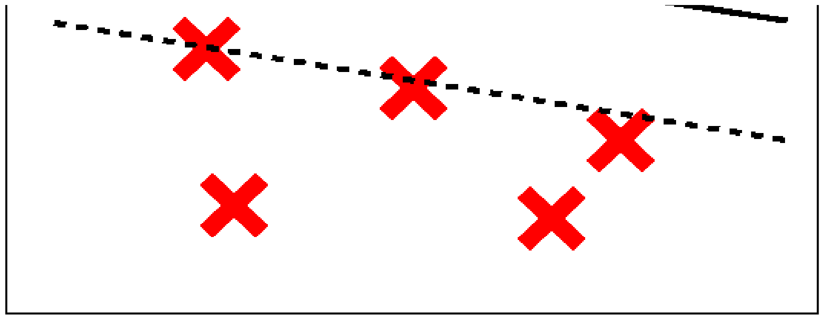




Φ rsp. $K(x,y) = \Phi(x) \bullet \Phi(y)$

$$\min \quad \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} {\xi_i}^p$$

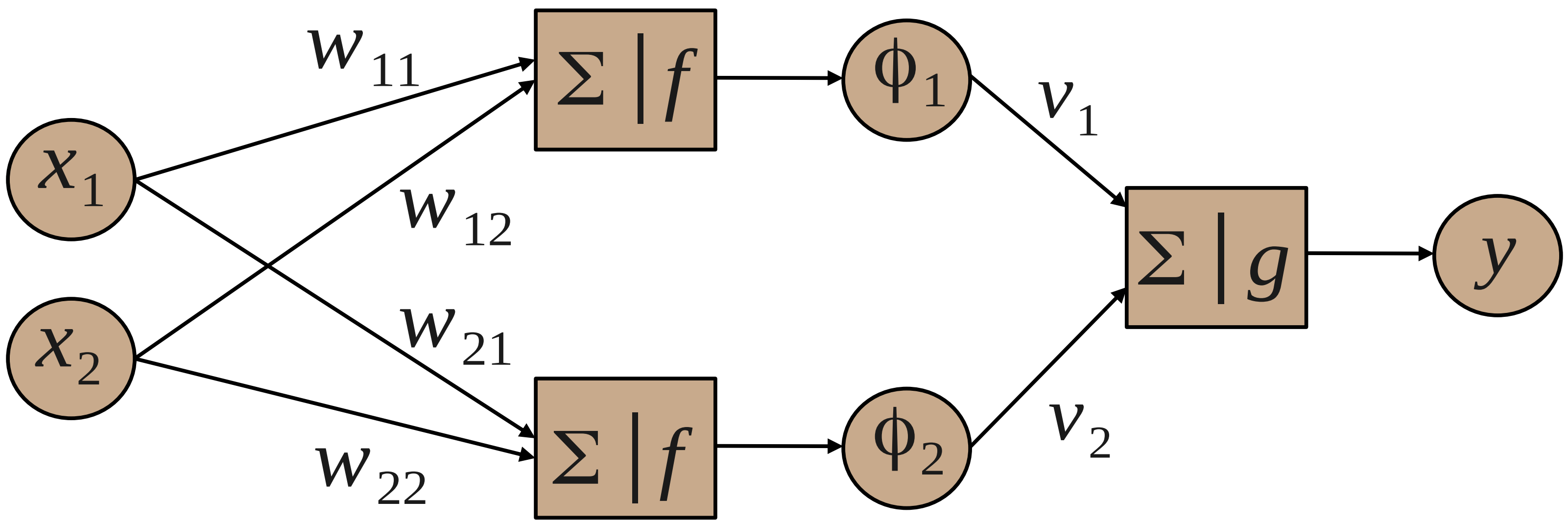
$$\text{subject to} \quad y_i \cdot [(\mathbf{w} \cdot \Phi(\mathbf{x}_i)) + b] \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \quad \text{for } i = 1 \ldots N$$



good theory
non-linear decision by implicitely mapping the data into feature space by SV **kernel** function **K**

[e.g. Vapnik 95, Muller et al 2001, Schölkopf & Smola 2002, Montavon et al 2013]

Multilayer networks



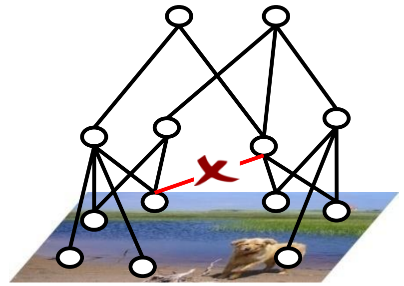
$$\phi_1 = f\left(x_1 w_{11} + x_2 w_{12} + b_1\right)$$
$$\phi_2 = f\left(x_1 w_{21} + x_2 w_{22} + b_2\right)$$
$$y = g\left(\phi_1 v_1 + \phi_2 v_2 + c\right)$$

Matrix form:

$$y = g\left(V \cdot f\left(W \cdot x\right)\right)$$

State of the art in ML: Kernel Methods and Deep Neural Networks

- Kernel methods have been the major ML algorithm for a decade
- recently deep learning has become the hot ML method: Why?
- Deep net architecture can be structured



- Representation is learned
- Multiscale information is included
- highly successful in practice, but WHY?
- parallelization is possible and **GPU** implementation available
- remark: **more data** **(big data)** and statistical estimators 1/N

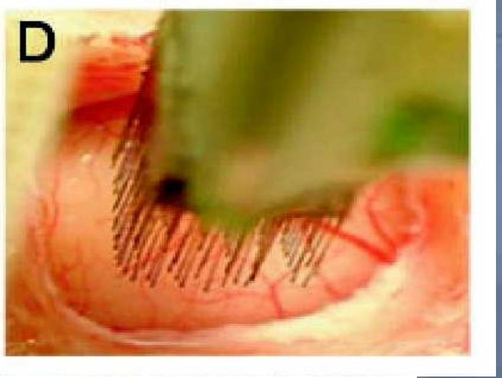
Toward Brain Computer Interfacing

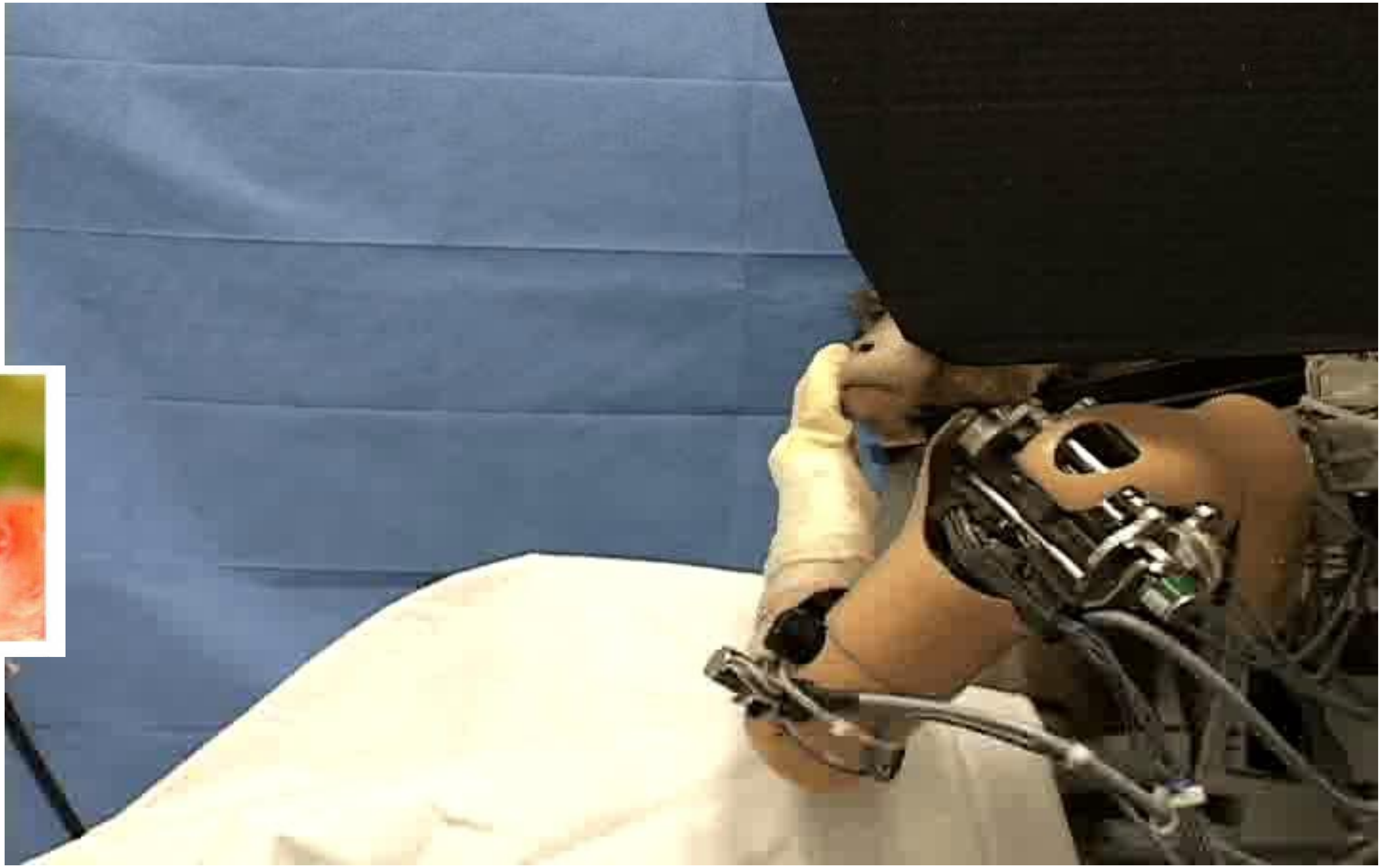
berlin
brain computer
interface

CHARITÉ CAMPUS BENJAMIN FRANKLIN

Klaus-Robert Müller, Siamac Fazli, Jan Mehnert, Stefan Haufe, Frank Meinecke, Paul von Bünau, Franz Kiraly, Felix Biessmann, Sven Dähne, Johannes Höhne, Michael Tangermann, Carmen Vidaure, Gabriel Curio, Benjamin Blankertz et al.

Invasive BCI at it’s best

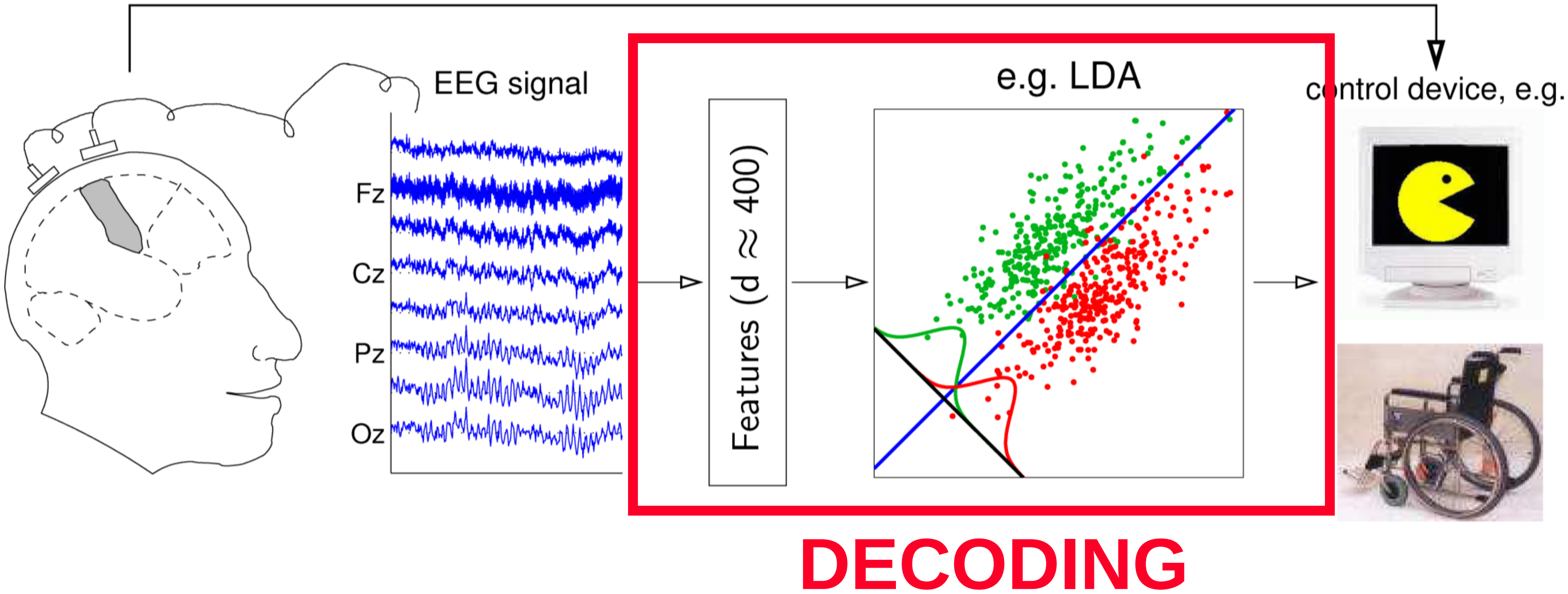




Remark: 24*1000*
3600*30000 ~ 2tb/day

[From Schwartz]

Noninvasive Brain-Computer Interface



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

BCI for communcation

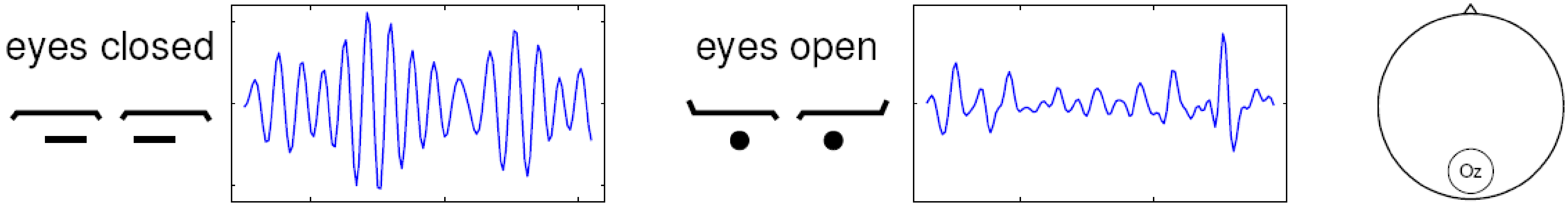
‚Brain Pong‘ with BBCI

Remark: 3*100*
3600*1000 ~ 1-2Gb/Experiment

Towards imaginations: Modulation of Brain Rhythms

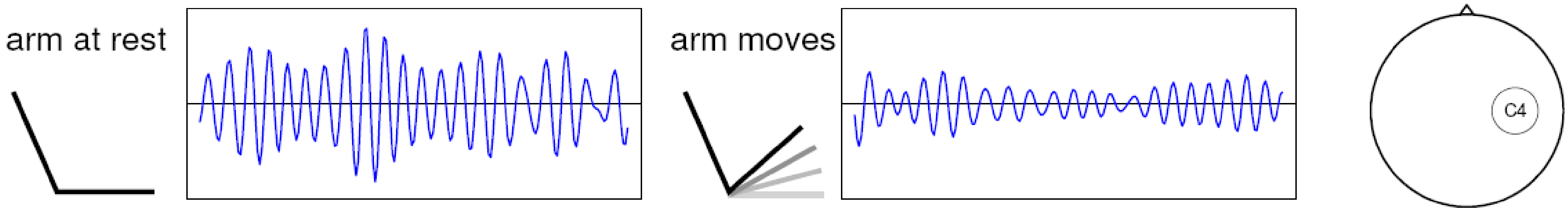
Most rhythms are idle rhythms, i.e., they are **attenuated** during activation.

- α-rhythm (around 10 Hz) in visual cortex:



Single channel

- μ-rhythm (around 10 Hz) in motor and sensory cortex:



IMAGINATION of left arm

BBCI paradigms

Leitmotiv: *›let the machines learn‹*

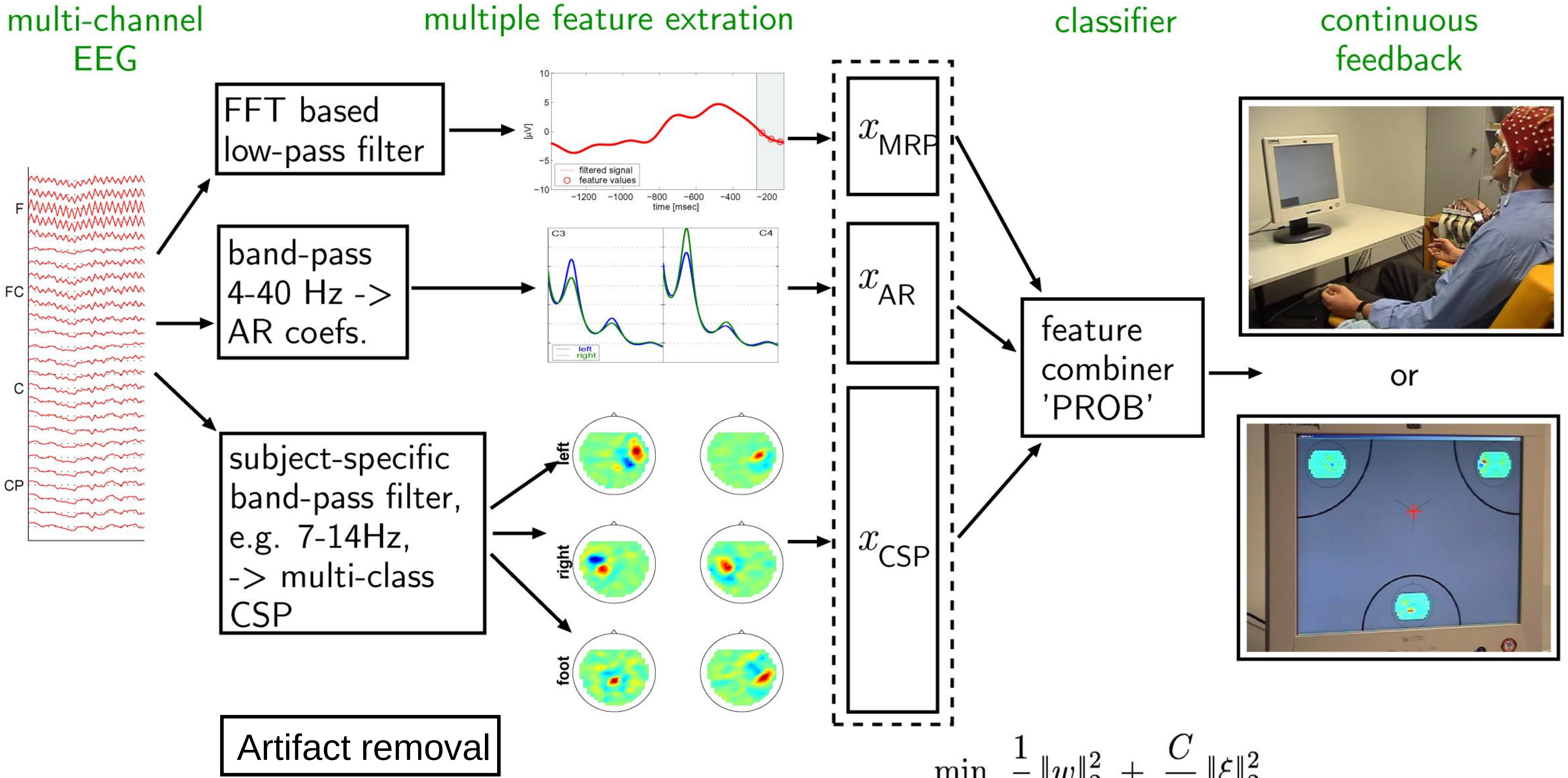
- healthy subjects *untrained* for BCI

A: training <10min: right/left hand **imagined** movements

→ infer the respective brain acivities (ML & SP)

B: online feedback session

BBCI Set-up

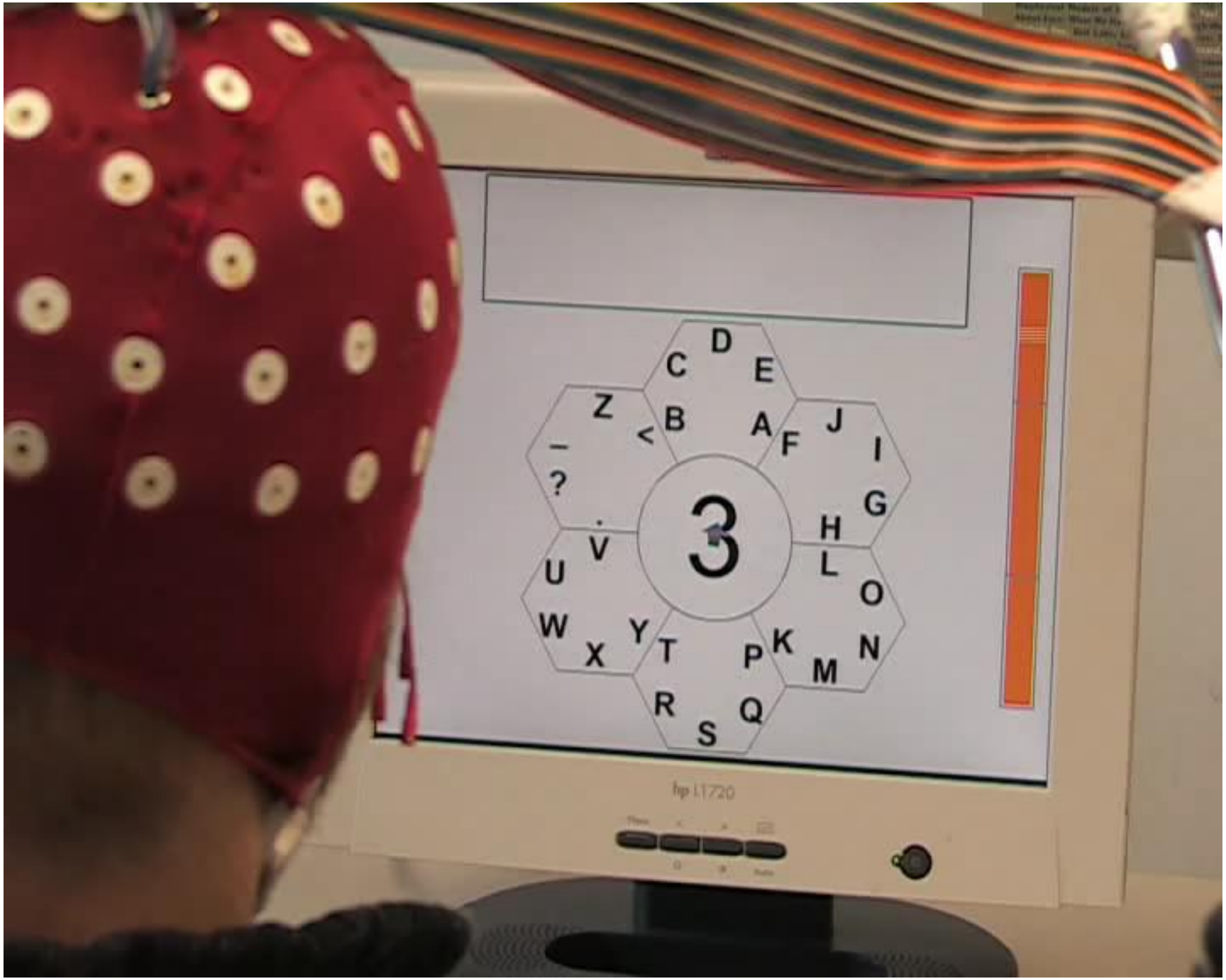


$$\min_{w,b,\xi} \; \frac{1}{2}\|w\|_2^2 + \frac{C}{K}\|\xi\|_2^2$$

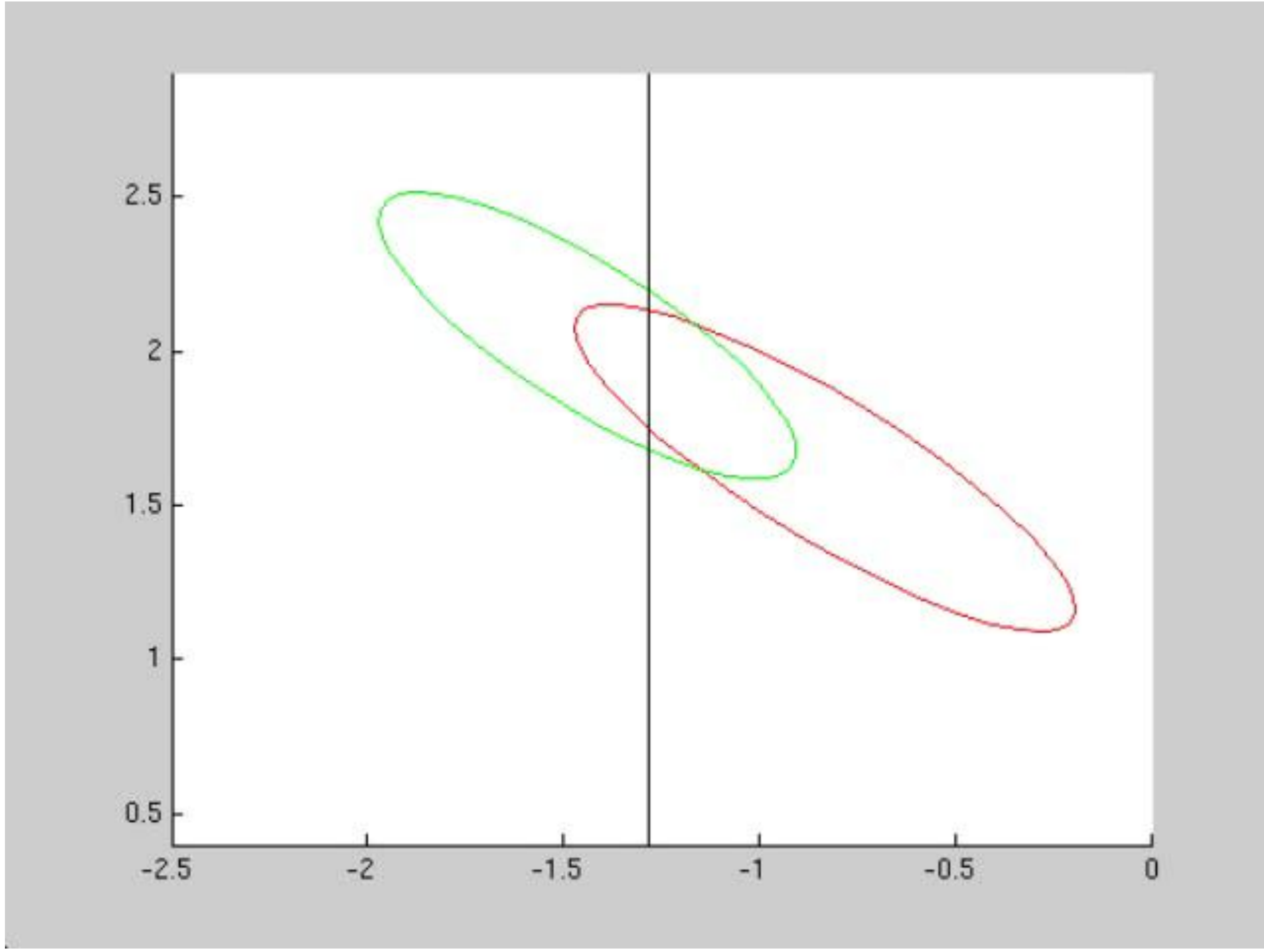
$$\text{subject to} \qquad y_k(w^\top x_k + b) = 1 - \xi_k \quad \text{for } k = 1, \ldots, K$$

[cf. Müller et al. 2001, 2007, 2008, Dornhege et al. 2003, 2007, Blankertz et al. 2004, 2005, 2006, 2007, 2008]

Spelling with BBCI: a communication for the disabled



Shifting distributions within experiment



Conclusion BCI

- BBCI: Untrained, Calibration < 10min, data analysis <<5min, BCI experiment
- 5-8 let/min mental typewriter CeBit 06,10. Brain2Robot@Medica 07, lNdW 09,11
- Machine Learning and modern data analysis is of central importance for BCI **et al**
- Applications: communication vs. measuring **(DECODING)**

  Rehabilitation: **TOBI EU IP, stroke**

  Computational Neuroscience: **Bernstein Centers Berlin**

  Man Machine Interaction: **brain@work**
- **Patient** studies

FOR INFORMATION SEE:
www.bbci.de

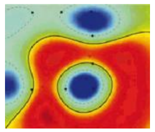
And now for something completely different

[Montavon et al 13, Rupp et al 2012 ….]

ML4Physics @ IPAM 2011

Klaus-Robert Müller, Matthias Rupp

Anatole von Lilienfeld and Alexandre Tkachenko et al



Machine Learning for chemical compound space

Ansatz:

$$\{Z_I, \mathbf{R}_I\} \xrightarrow{\mathrm{ML}} E$$

instead of

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

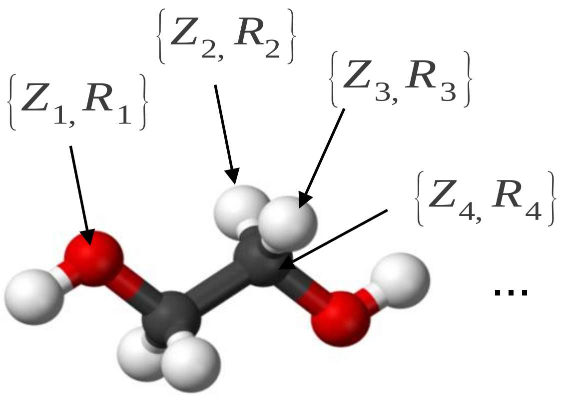
$$\hat{H}\Psi = E\Psi$$

[from von Lilienfeld]

Coulomb representation of molecules

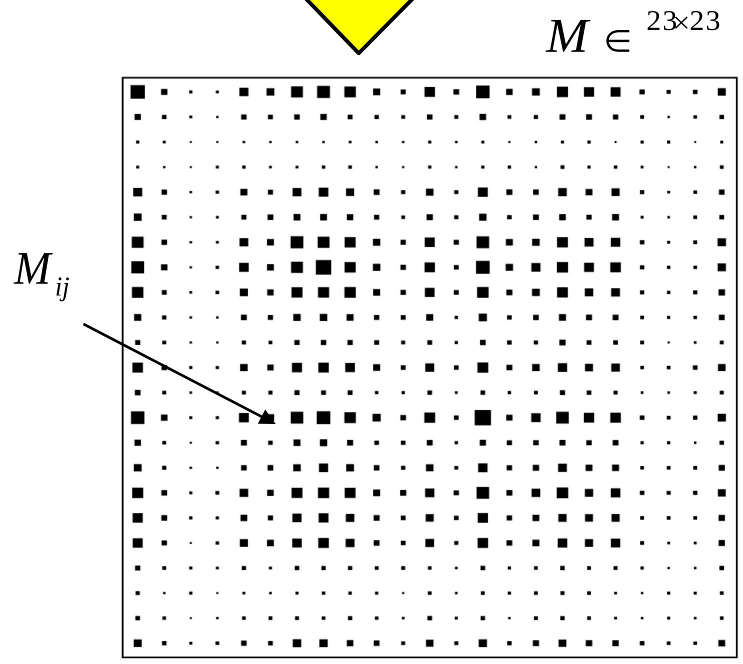
$$M_{ii} = Z_i^{2.4}$$

$$M_{ij} = \frac{Z_i Z_j}{\left\| R_i - R_j \right\|}$$



+ phantom atoms

$\{0, R_{21}\}$ $\{0, R_{22}\}$ $\{0, R_{23}\}$



Coulomb Matrix (Rupp, Müller et al 2012, PRL)

$$d(\mathbf{M}, \mathbf{M}') = \sqrt{\sum_{IJ} |M_{IJ} - M'_{IJ}|^2}$$

Kernel ridge regression

Distances between **M** define Gaussian kernel matrix **K**

$$k(\mathbf{M}, \mathbf{M}') \;=\; \exp\left(-\frac{d(\mathbf{M}, \mathbf{M}')^2}{2\sigma^2}\right)$$

Predict energy as sum over weighted Gaussians

$$\boxed{E^{est}(\mathbf{M}) \;=\; \sum_i \alpha_i k(\mathbf{M}, \mathbf{M}_i) + b}$$

using weights that minimize error in training set

$$\min_{\alpha} \quad \sum_i \left(E^{est}(\mathbf{M}_i) - E_i^{ref}\right)^2 + \lambda \sum_i \alpha_i^2$$

$$\alpha \;=\; (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{E}^{ref}$$

Exact solution

As many parameters as molecules + 2 global parameters, characteristic length-scale or kT of system (σ), and noise-level (λ)

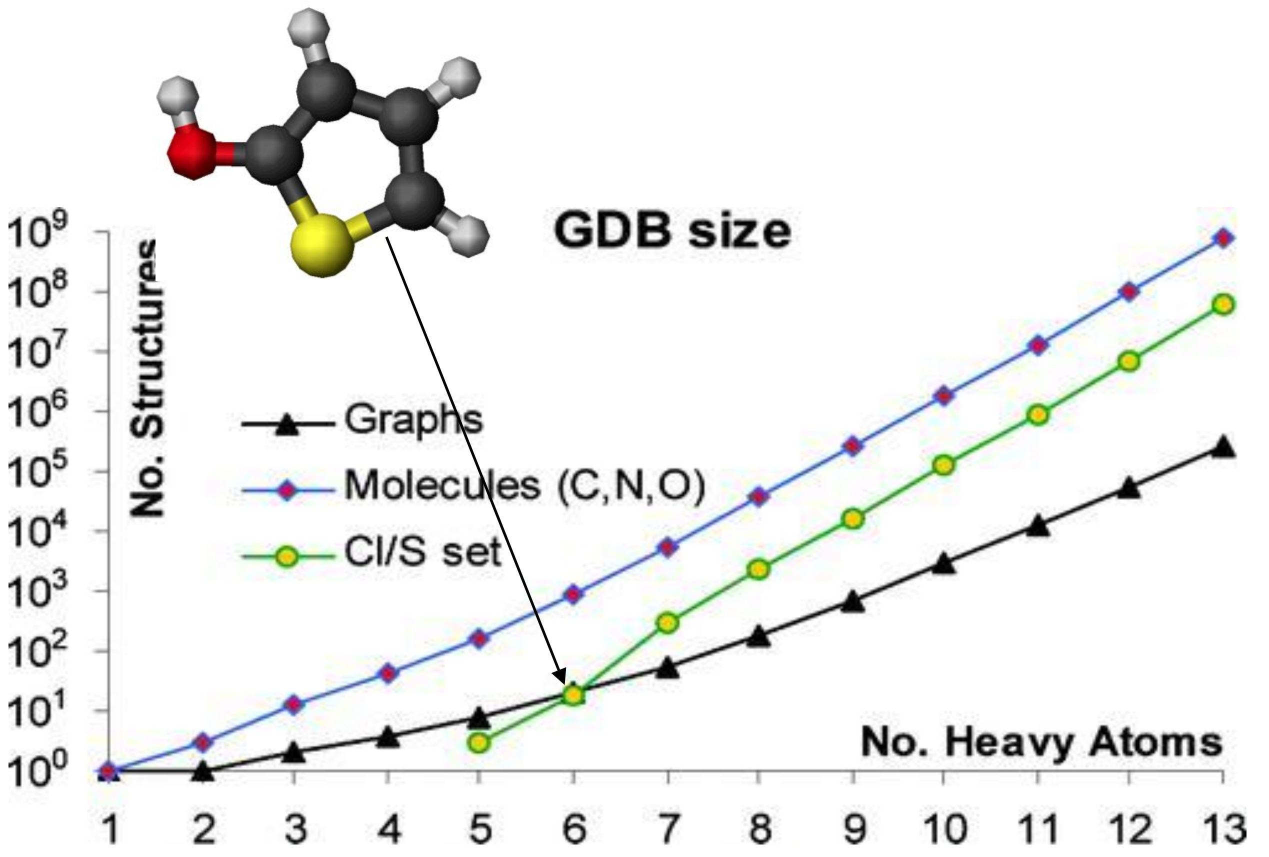
[from von Lilienfeld]

The data

GDB-13 database of all organic molecules (within stability & synthetic constraints) of 13 heavy atoms or less: 0.9B compounds

Table 1. Structure Generation Statistics for GDB-13

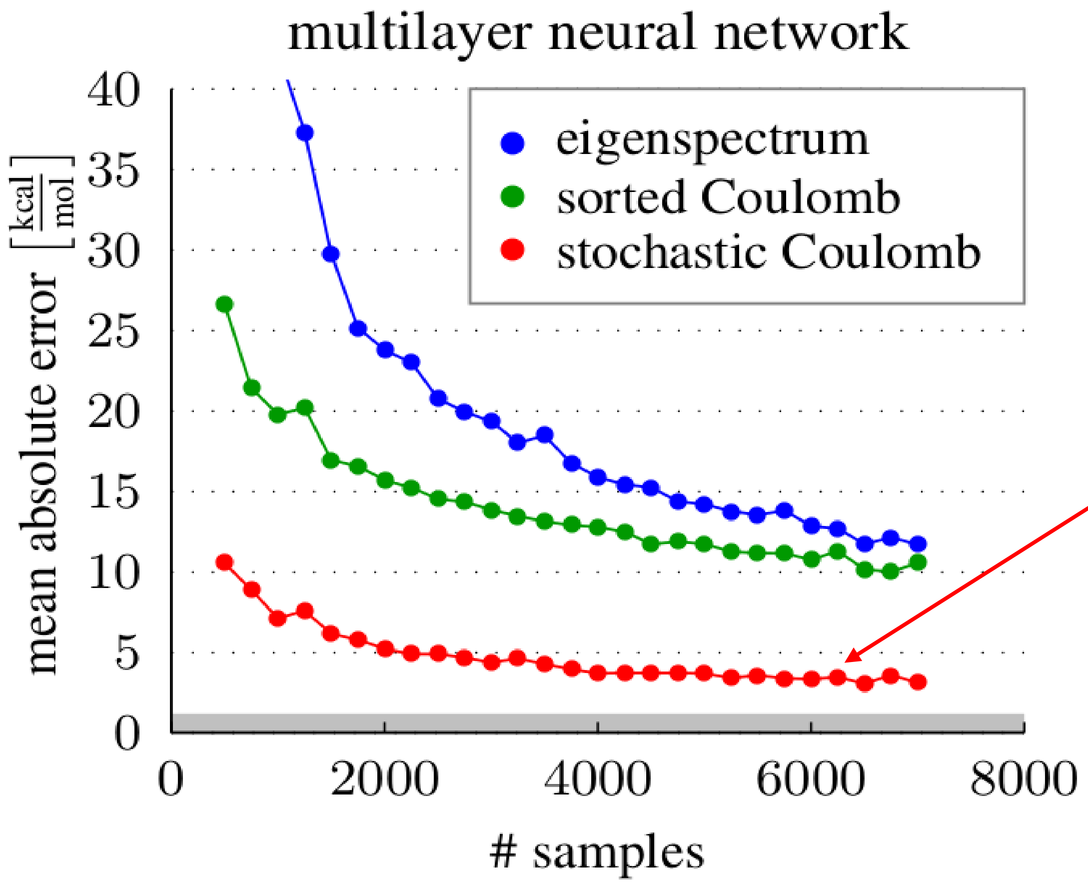
nodes[a]	graphs[b]	GDB[c]	CI/S[d]	CPU time (h)[e]
1	1	1	0	0.00
2	1	3	0	0.00
3	2	12	0	0.00
4	4	43	0	0.00
5	8	155	3	0.01
6	20	934	19	0.02
7	57	5726	315	0.05
8	194	37151	2438	0.33
9	706	255542	17056	2.68
10	2831	1784626	130465	25.26
11	12011	12961686	938704	223.49
12	53789	99821343	7240108	3023.79
13	250268	795244451	59027533	36606.45
Total	**319892**	**910111673**	**67356641**	**39882.08**



Blum & Reymond, *JACS* (2009)

[from von Lilienfeld]

Results



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

December 2012
Montavon et al., NIPS
3.51 kcal/mol
(Neural nets + Coulomb sets)

2015 Tkatchenko 1.3kcal/mol

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

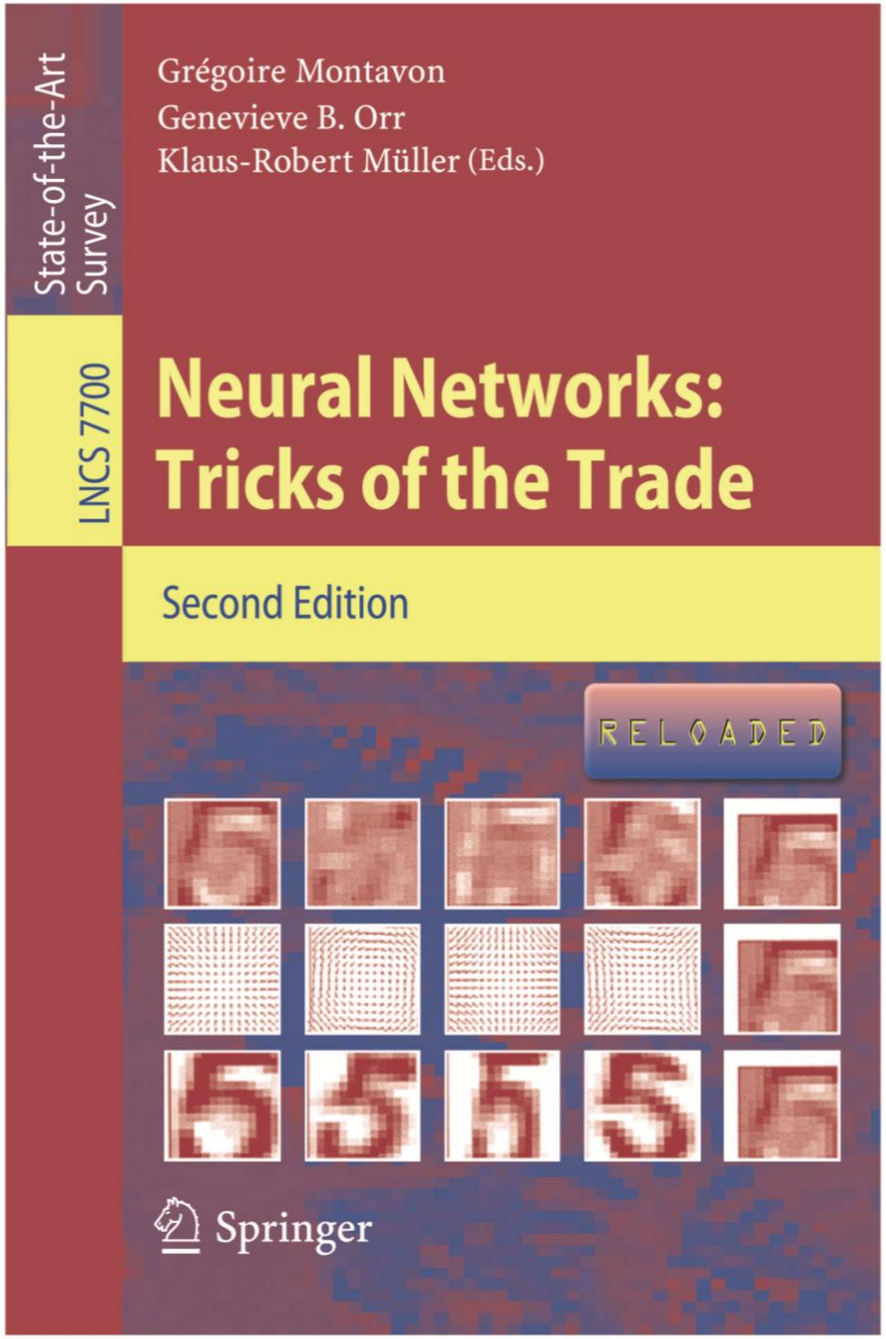
Dataset available at http://quantum-machine.org

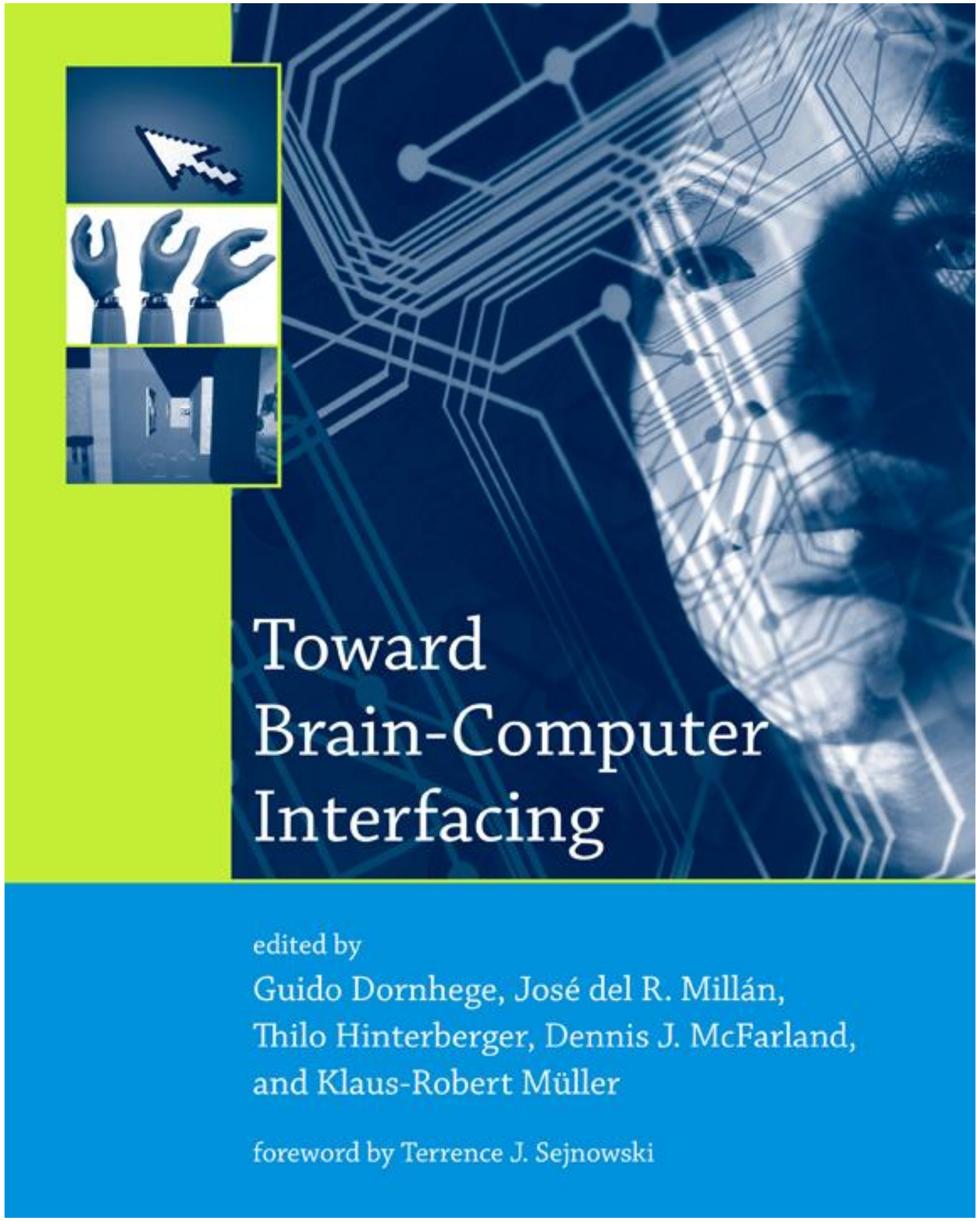
Conclusion

- Machine Learning is a versatile and ready to use tool for data analysis
- No single best ML algorithm, despite of hypes
- small data vs. big data
- data is not equal to information
- Big data= ML & Data Bases -> BBDC
- technical challenges: nonstationarity, heterogeneous complex data, streaming data, energy consumption, robustness, explanation
- trust & privacy vs. convenience: new legislative efforts needed NOW
- ML is a tool, there are applications of ML that are beneficial for mankind others are more unclear

BBDC

BERLIN BIG DATA CENTER


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

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.

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K. R., & Curio, G. (2007). The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects. *NeuroImage*, *37*(2), 539-550.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., & Muller, K. R. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine, 25*(1), 41-56.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K. R. (2011). Single-trial analysis and classification of ERP components—a tutorial. *NeuroImage*, *56*(2), 814-825.

Blythe, D. A., von Bunau, P., Meinecke, F. C., & Muller, K. (2012). Feature extraction for change-point detection using stationary subspace analysis. *, IEEE Transactions on Neural Networks and Learning Systems23*(4), 631-643.

von Bünau, P., Meinecke, F. C., Király, F. C., & Müller, K. R. (2009). Finding stationary subspaces in multivariate time series. *Physical Review Letters*, *103*(21), 214101.

Dähne, S., Bießman, F., Meinecke, F.C., Mehnert, J., Fazli, S., Müller, K.R., (2013). Integration of multivariate data streams with bandpower signals. IEEE Trans. Multimedia 15 (5), 1001–1013.

Dähne, S., Meinecke, F. C., Haufe, S., Höhne, J., Tangermann, M., Müller, K. R., & Nikulin, V. V. (2014). SPoC: A novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. *NeuroImage*, *86*, 111-122.

Dähne, S., Nikulin, V., Ramirez, D., Schreier, PJ, Müller, K. R., Haufe, S. (2014). Finding brain oscillations with power dependencies in neuroimaging data, Neuroimage 96, 334–348

Further Reading II

Dähne, S., Bießman, F., Samek, W., Haufe, S., Goltz, D., Gundlach, C., Villringer, A., Fazli, S., and Müller, K.-R. (2015). Multivariate machine learning methods for fusing functional multimodal neuroimaging data. Proceedings of the IEEE.accepted

Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K. R., & Blankertz, B. (2012). Enhanced performance by a hybrid NIRS–EEG brain computer interface. *Neuroimage*, *59*(1), 519-529.

Fazli, S., Dähne, S., Samek, W., Bießmann, F., and Müller, K.-R. (2015). Learning from more than one data source: data fusion techniques for sensorimotor rhythm-based Brain-Computer Interfaces. *Proceedings of the IEEE*. Accepted

Höhne, J., Holz, E., Staiger-Sälzer, P., Müller, K. R., Kübler, A., & Tangermann, M. (2014). Motor imagery for severely motor-impaired patients: evidence for brain-computer interfacing as superior control solution. PloS one, 9(8), e104854.

Király, F. J., von Bünau, P., Meinecke, F. C., Blythe, D. A., & Müller, K. R. (2012). Algebraic geometric comparison of probability distributions. *The Journal of Machine Learning Research*, *13*, 855-903.

Lemm, S., Blankertz, B., Dickhaus, T., & Müller, K. R. (2011). Introduction to machine learning for brain imaging. *Neuroimage*, *56*(2), 387-399.

Samek, W., Vidaurre, C., Müller, K. R., & Kawanabe, M. (2012). Stationary common spatial patterns for brain–computer interfacing. *Journal of neural engineering*, *9*(2), 026013.

Samek, W., Meinecke, F. C., & Müller, K. R. (2013). Transferring subspaces between subjects in brain-computer interfacing. *IEEE Trans on Biomedical Engineering*, *60*(8), 2289-2298.

Further Reading III

Samek, W., Kawanabe, M., & Muller, KR. (2014). Divergence-based framework for common spatial patterns algorithms. IEEE Rev Biomed Eng, 7, 50-72.

Tangermann, M., Krauledat, M., Grzeska, K., Sagebaum, M., Blankertz, B., Vidaurre, C., & Müller, K. R. (2008). Playing Pinball with non-invasive BCI. In NIPS (pp. 1641-1648).

Books

Dornhege, G. del Millan, J., McFarland, D., Hinterberger, T., & Muller, KR. (eds.) (2007). *Toward brain-computer interfacing*. MIT press.

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

Further reading: Physics and ML (see also quantum-machine.org)

Quantum machine

M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld: Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, Physical Review Letters, 108(5):058301, 2012

G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, O. A. von Lilienfeld, K.-R. Müller, Learning Invariant Representations of Molecules for Atomization Energy Prediction, Advances in Neural Information Processing Systems (NIPS), 2012

G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O.A. von Lilienfeld, Machine Learning of Molecular Electronic Properties in Chemical Compound Space, New Journal of Physics, 2013

K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, K.-R. Müller. Assessment and Validation of Machine Learning Methods for Predicting Molecular Energies, J. Chem. Theory Comput., 2013

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.

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)

Related papers (databases, quantum chemistry methods and simulations)

A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard III, W. M. Skid, UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations, J. Am. Chem. Soc., 114:10024, 1992

R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. K. Wegner and E. Willighagen, The Blue Obelisk - Interoperability in Chemical Informatics, J. Chem. Inf. Model., 46:991, 2006

L. C. Blum, J.-L. Reymond, 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13, J. Am. Chem. Soc., 131:8732, 2009