Machine Learning and Signal Processing Tools for BCI

CHARITÉ CAMPUS BENJAMIN FRANKLIN

Klaus-Robert Müller, Benjamin Blankertz, Gabriel Curio et al.

BBCI team:

Gabriel Curio
Florian Losch
Volker Kunzmann
Frederike Holefeld
Vadim Nikulin@Charite

Andreas Ziehe
Florin Popescu
Christian Grozea
Steven Lemm
Motoaki Kawanabe
Guido Nolte@FIRST

Yakob Badower@Pico Imaging
Marton Danozcy

berlin
brain computer
interface

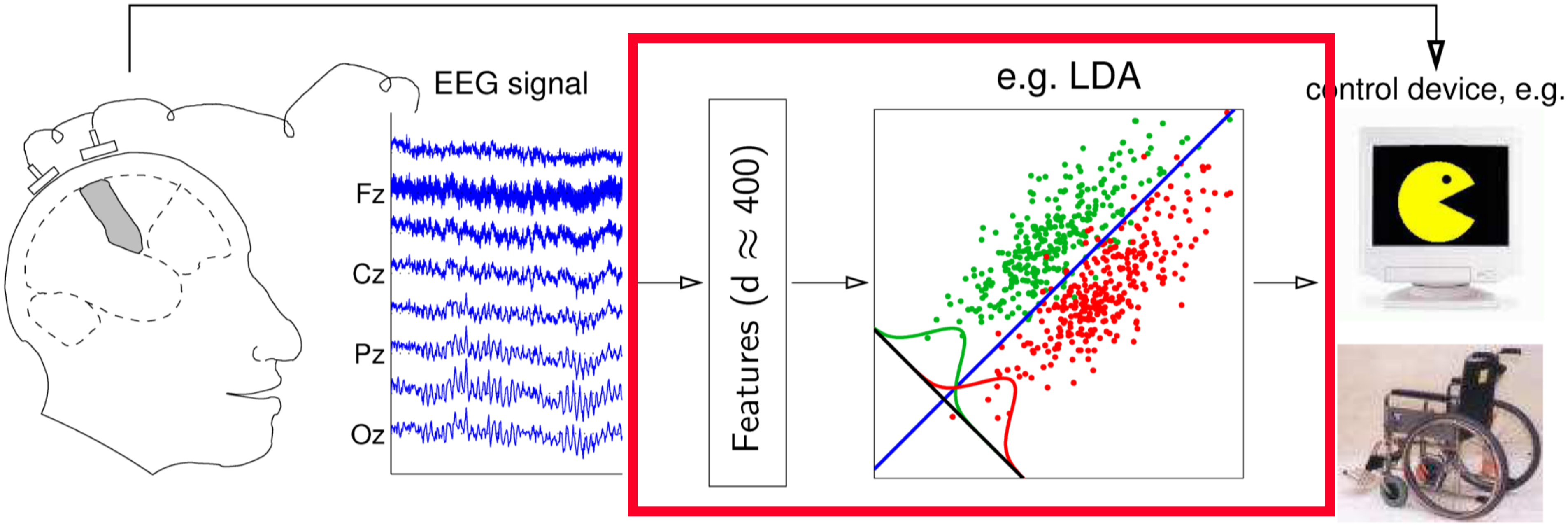
Benjamin Blankertz
Michael Tangermann
Claudia Sannelli
Carmen Vidaurre
Bastian Venthur
Siamac Fazli
Martijn Schreuder
Matthias Treder
Stefan Haufe
Thorsten Dickhaus
Frank Meinecke
Paul von Bünau
Marton Danozcy
Felix Biessmann
Klaus-Robert Müller@TUB

Matthias Krauledat
Guido Dornhege
Roman Krepki@industry

Collaboration with: U Tübingen, Bremen, Albany, TU Graz, EPFL, Daimler, Siemens, MES, MPIs, U Tokyo, TIT, RIKEN, Bernstein Focus Neurotechnology, Bernstein Center for Computational Neuroscience Berlin, picoimaging, Columbia, CUNY

Funding by: EU, BMBF and DFG

Noninvasive Brain-Computer Interface

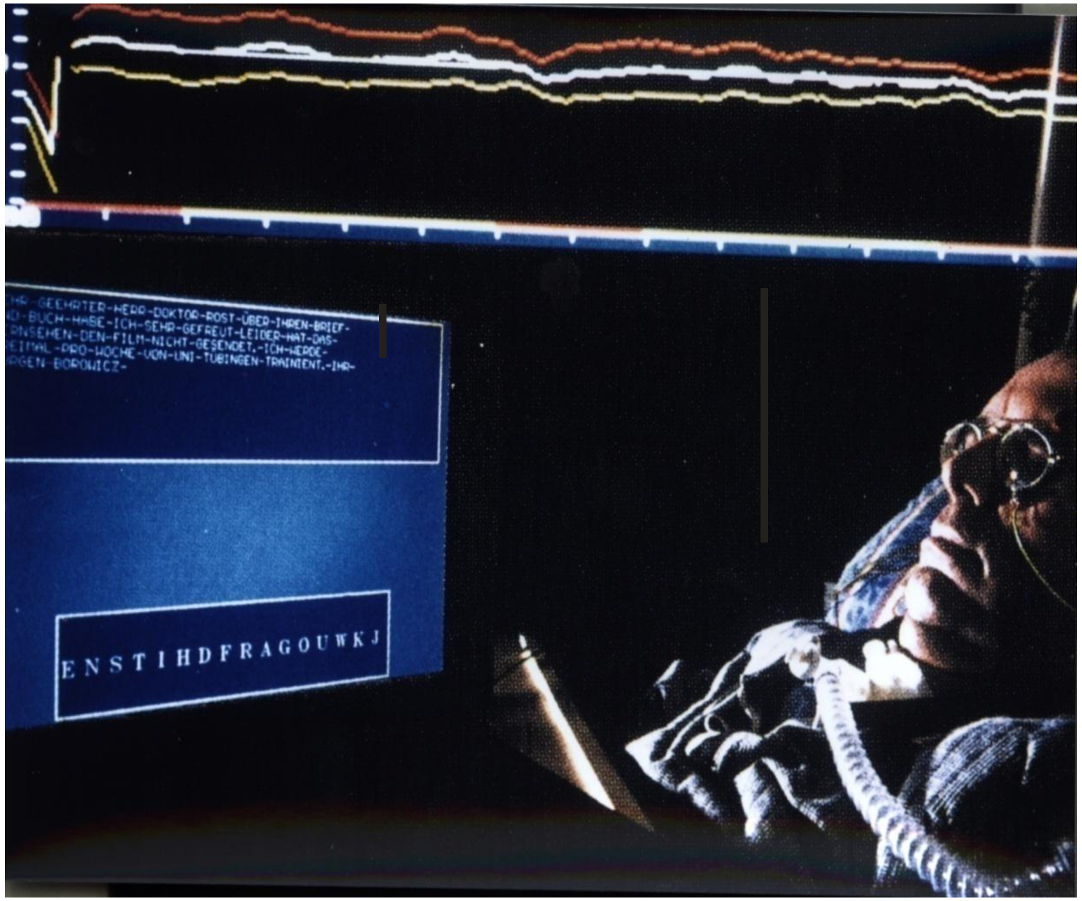


DECODING

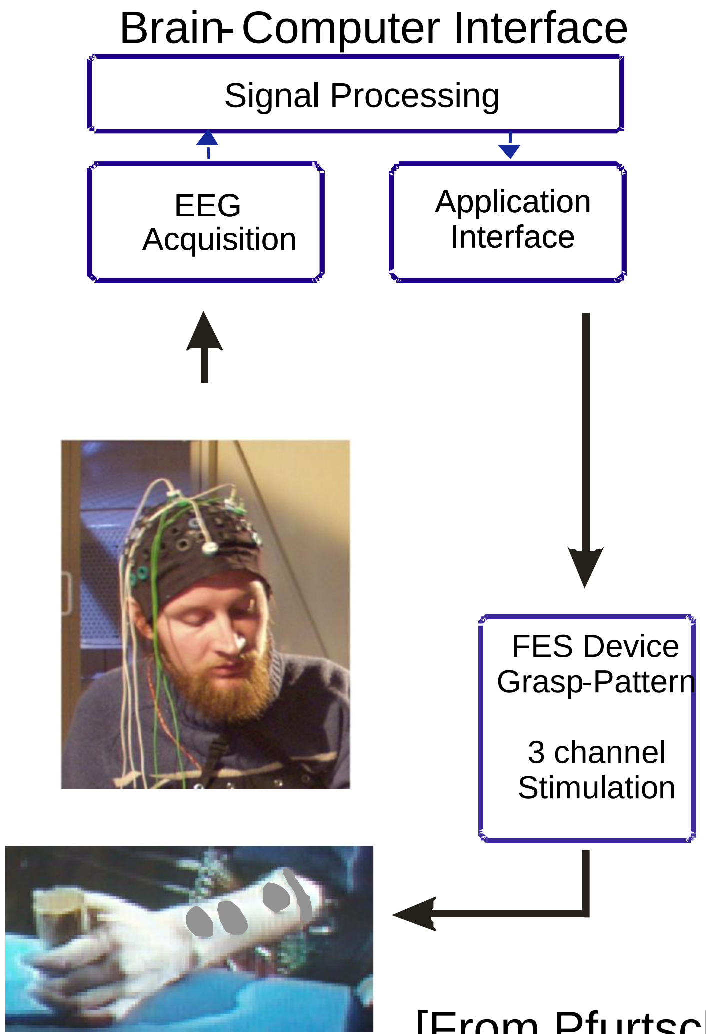
BCI: Translation of human intentions into a technical control signal
without using activity of muscles or peripheral nerves

‚Brain Pong‘ with BBCI

Noninvasive BCI: clinical applications



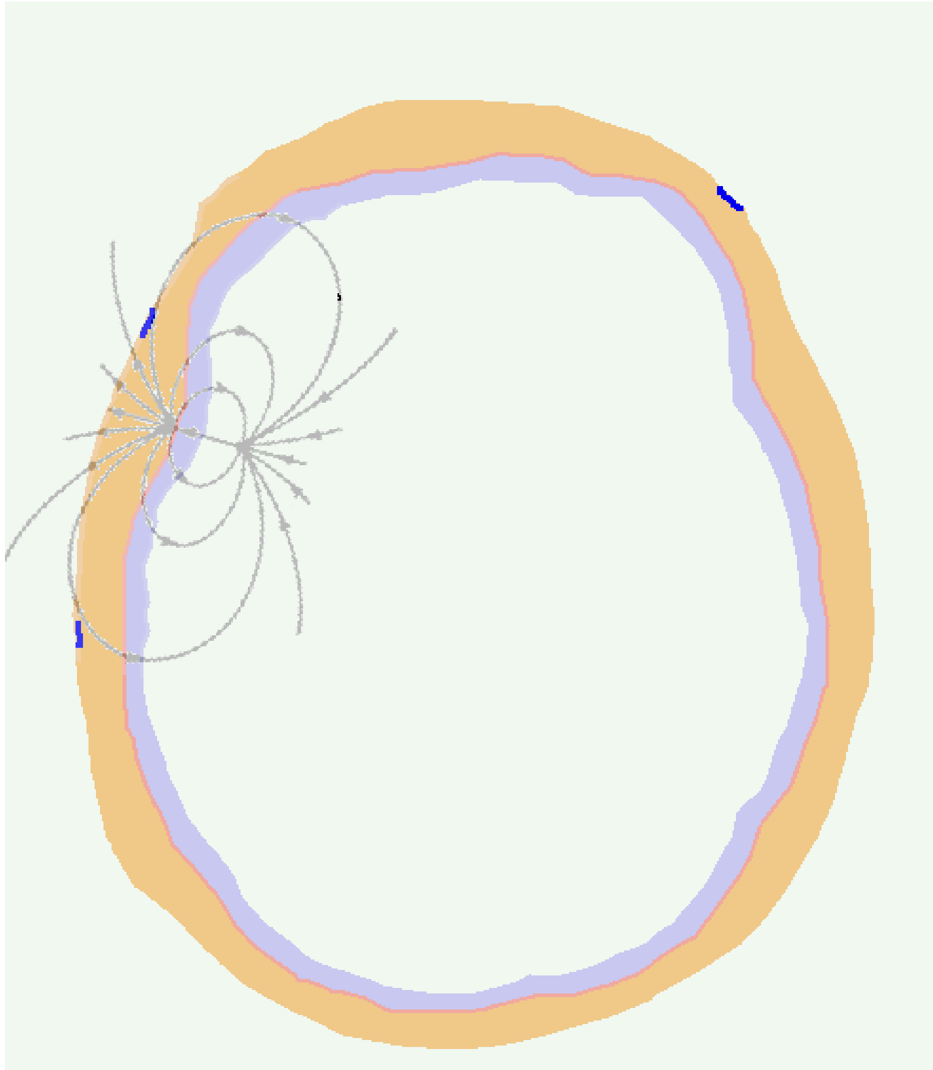
[From Birbaumer et al.]



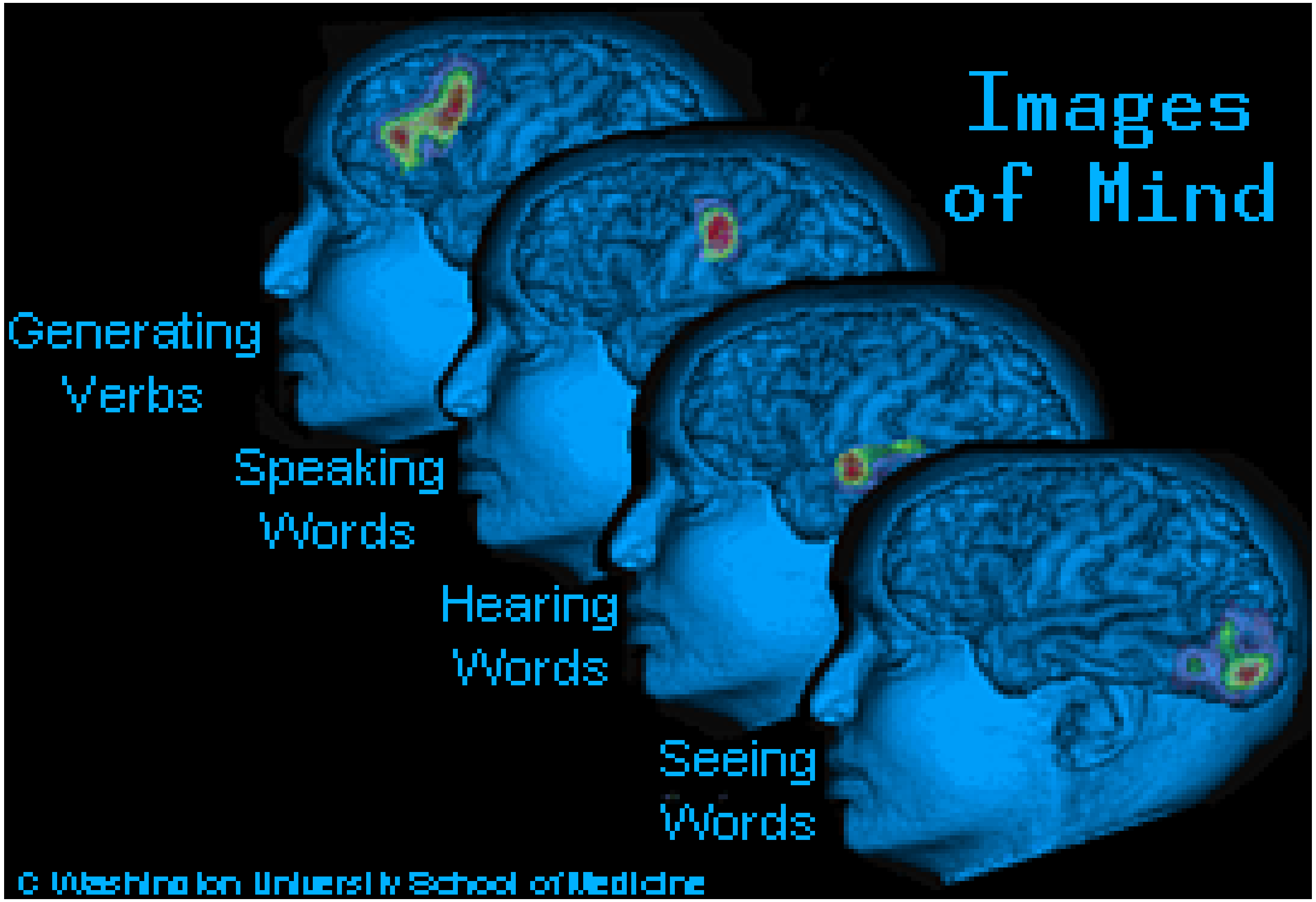
[From Pfurtscheller et al.]

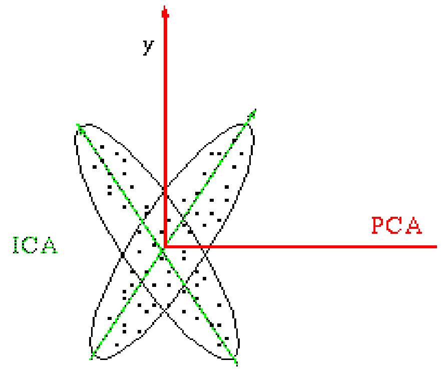
BBCI: Leitmotiv: *›let the machines learn‹*

EEG based noninvasive BCI



The cerebral cocktail party problem





- use ICA/NGCA projections for artifact and noise removal
- feature extraction and selection

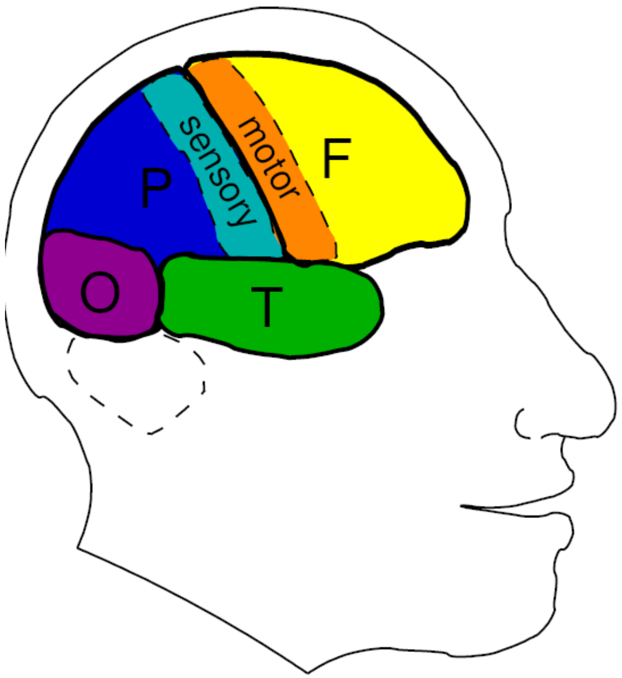
[cf. Ziehe et al. 2000, Blanchard et al. 2006]

BBCI paradigms

Leitmotiv: *›let the machines learn‹*

- healthy subjects (BCI *untrained*) perform "imaginary" movements (ERD/ERS)

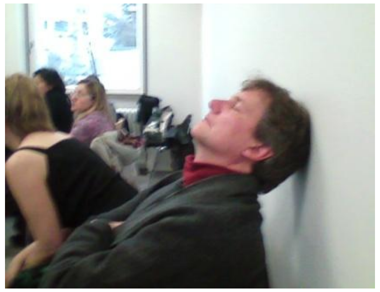
- instruction: imagine
  - squezzing a ball,
  - kicking a ball,
  - feel touch

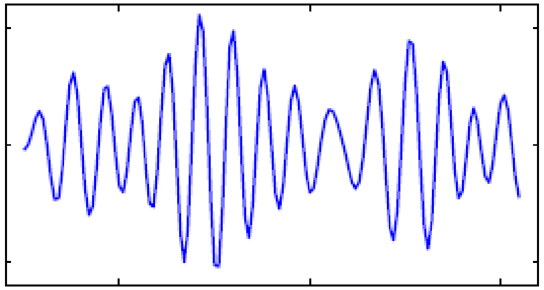


Towards imaginations: Modulation of Brain Rhythms

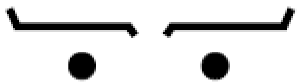
Most rhythms are idle rhythms, i.e., they are **attenuated** during activation.

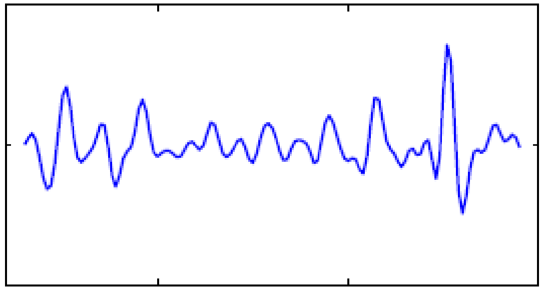
- α-rhythm (around 10 Hz) in visual cortex:

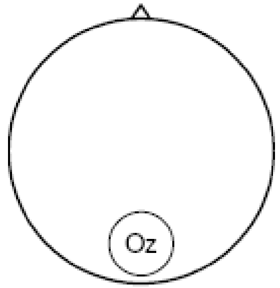




eyes open



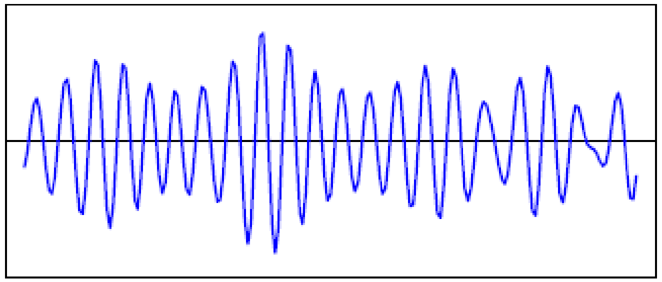




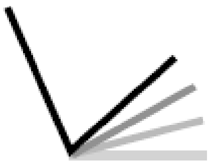
Single channel

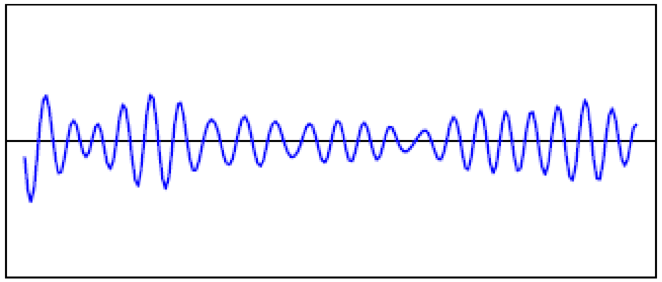
- μ-rhythm (around 10 Hz) in motor and sensory cortex:

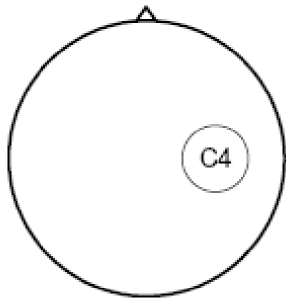
arm at rest



arm moves

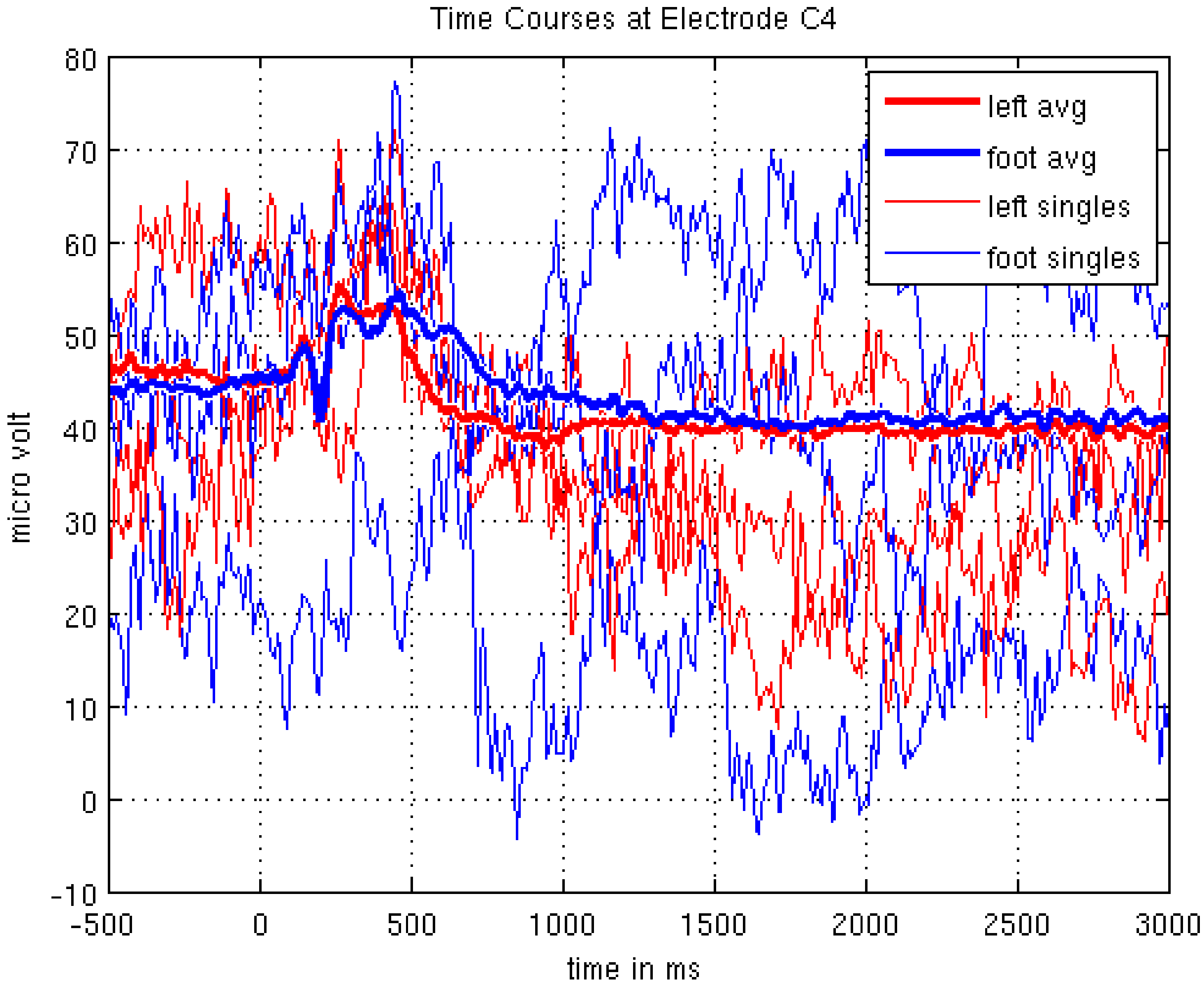






IMAGINATION of left arm

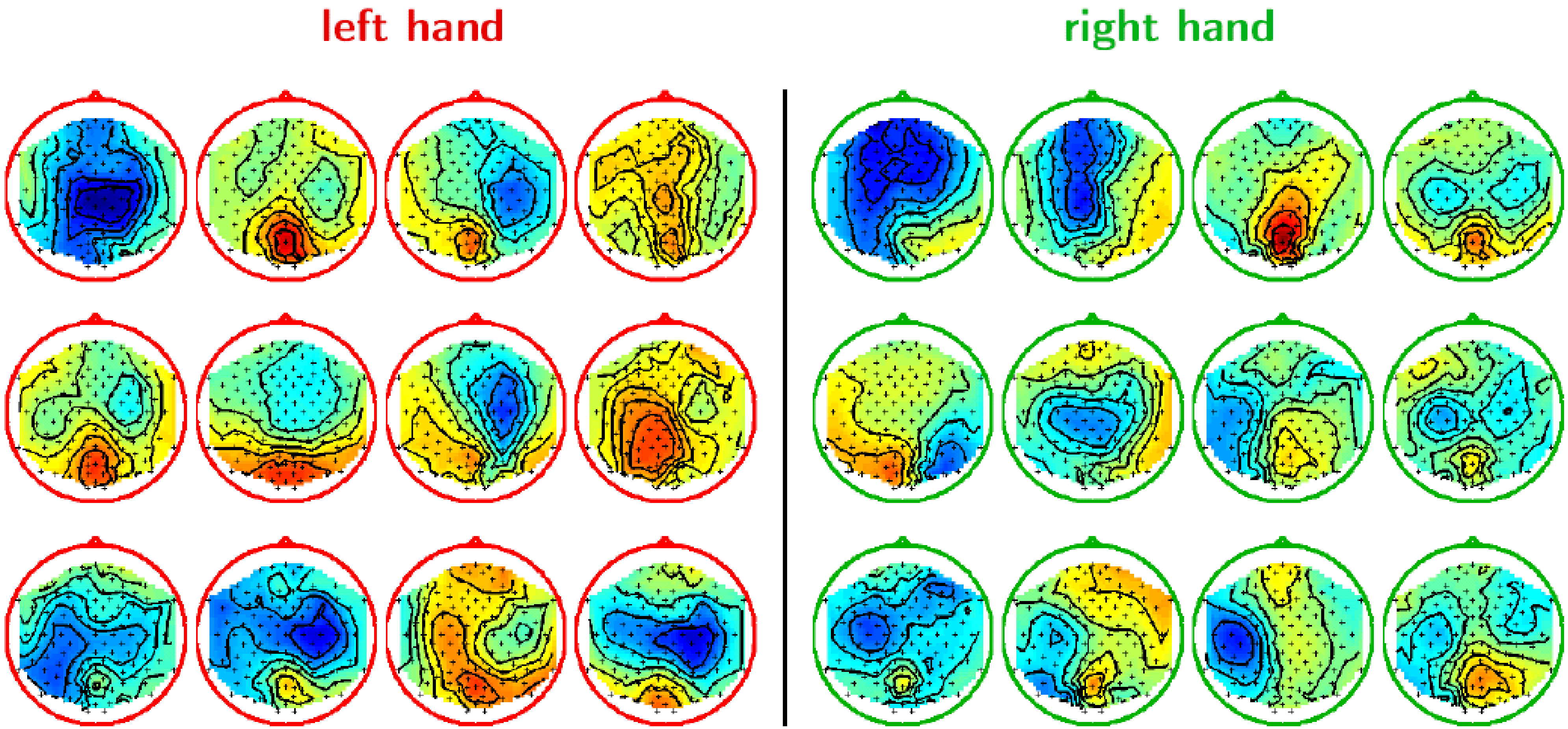
Variance I: Single-trial vs. Averaging



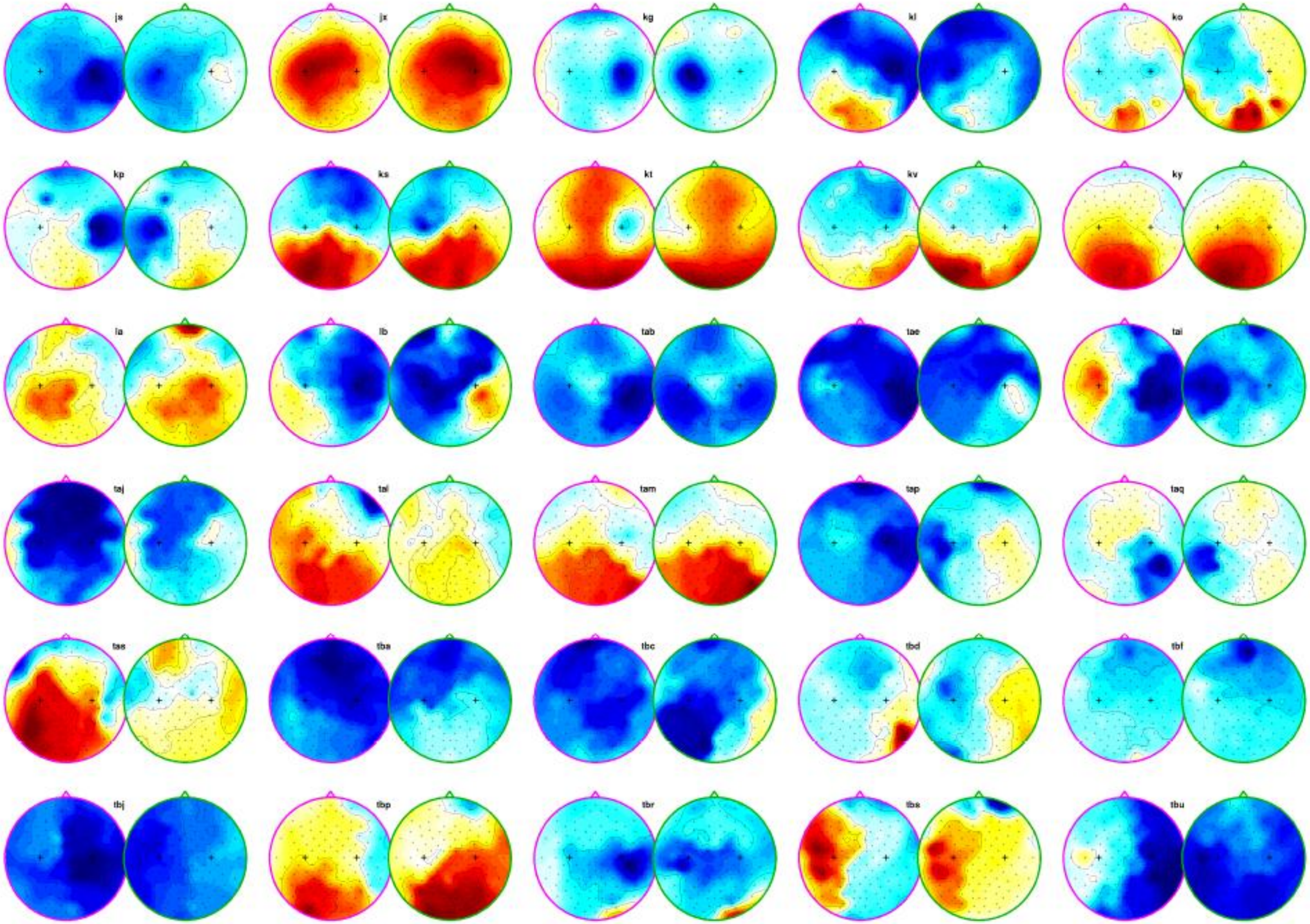
Single channel

Variance II: Trial to trial variability

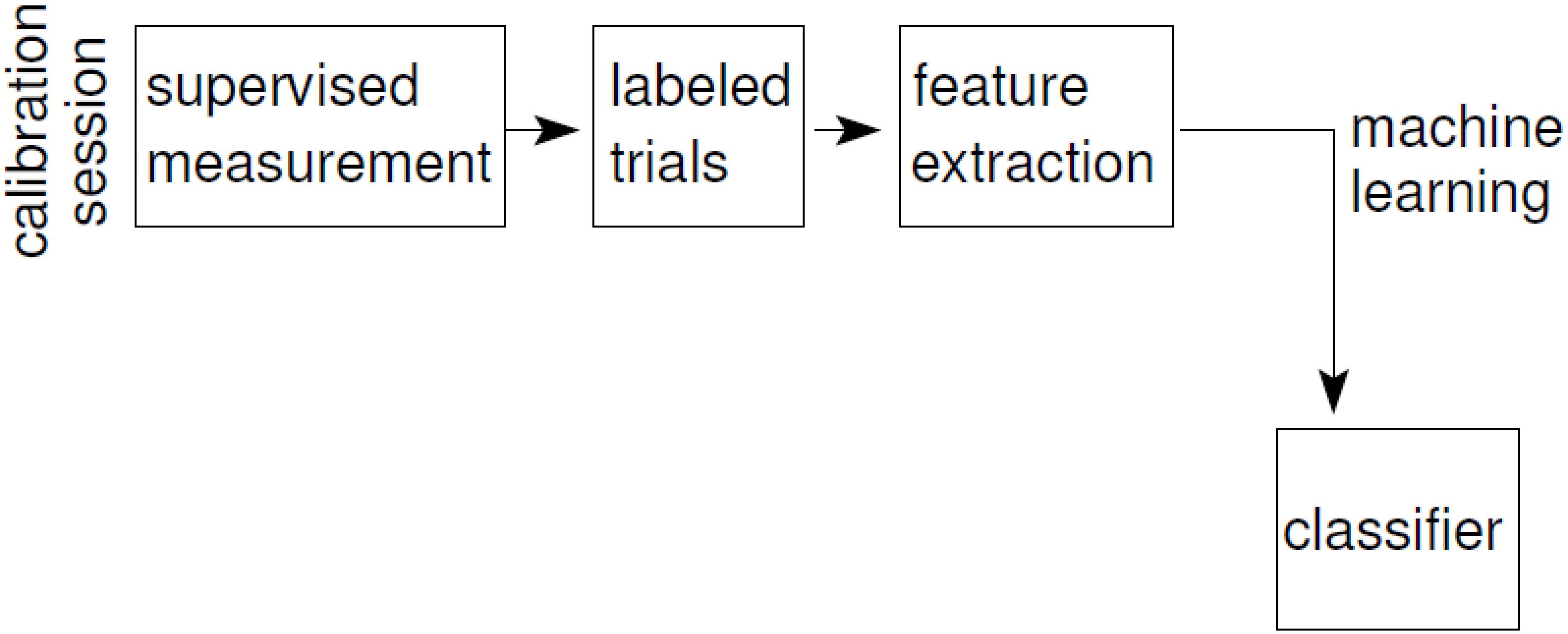
- Experiment: One subject imagined **left** vs. **right** hand movements.
- Topographies show power in the **alpha band** during trials of 3.5 s.
- They exhibit an extreme diversity, although recorded from **one subject** on **one day**.



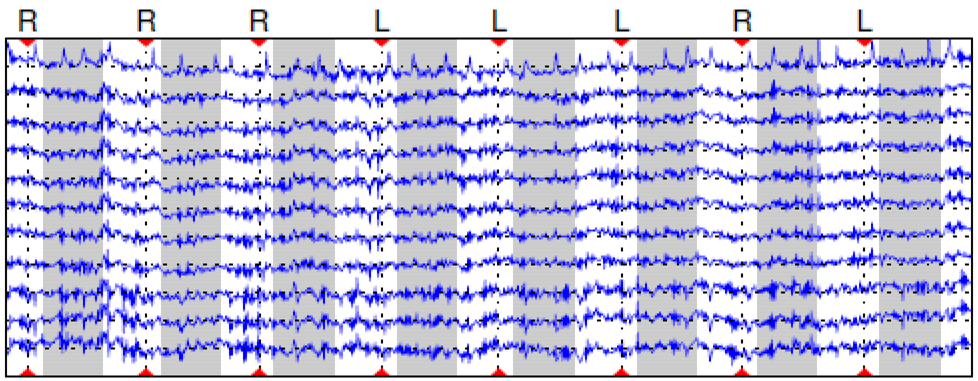
Variance III: inter subject variability [l vs r]



BCI with machine learning: training



offline: calibration (10–20 minutes)



collect training samples

BBCI paradigms

Leitmotiv: *›let the machines learn‹*

- healthy subjects *untrained* for BCI

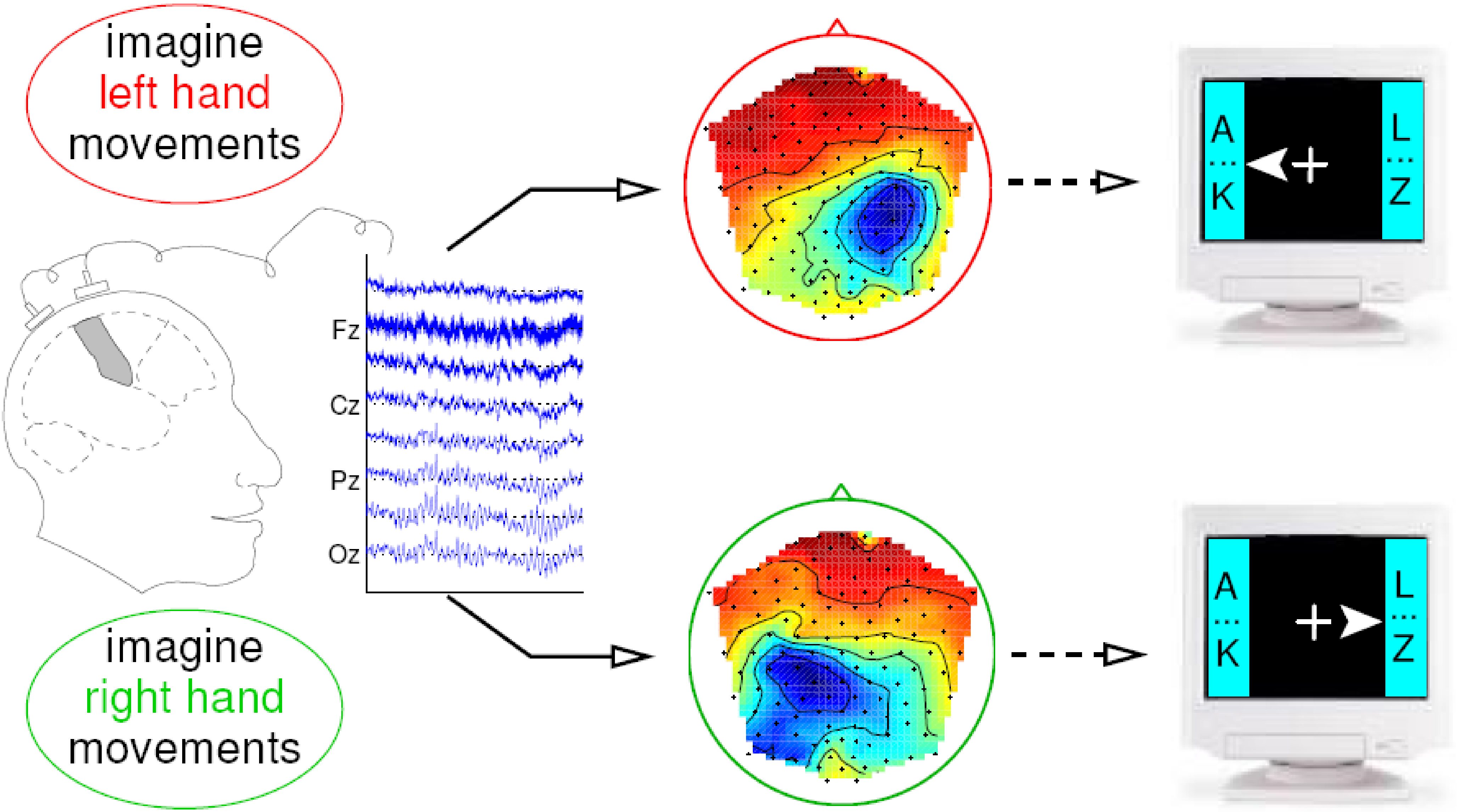
A: training 20min: right/left hand **imagined** movements

$\longrightarrow$ infer the respective brain acivities (ML & SP)

B: online feedback session

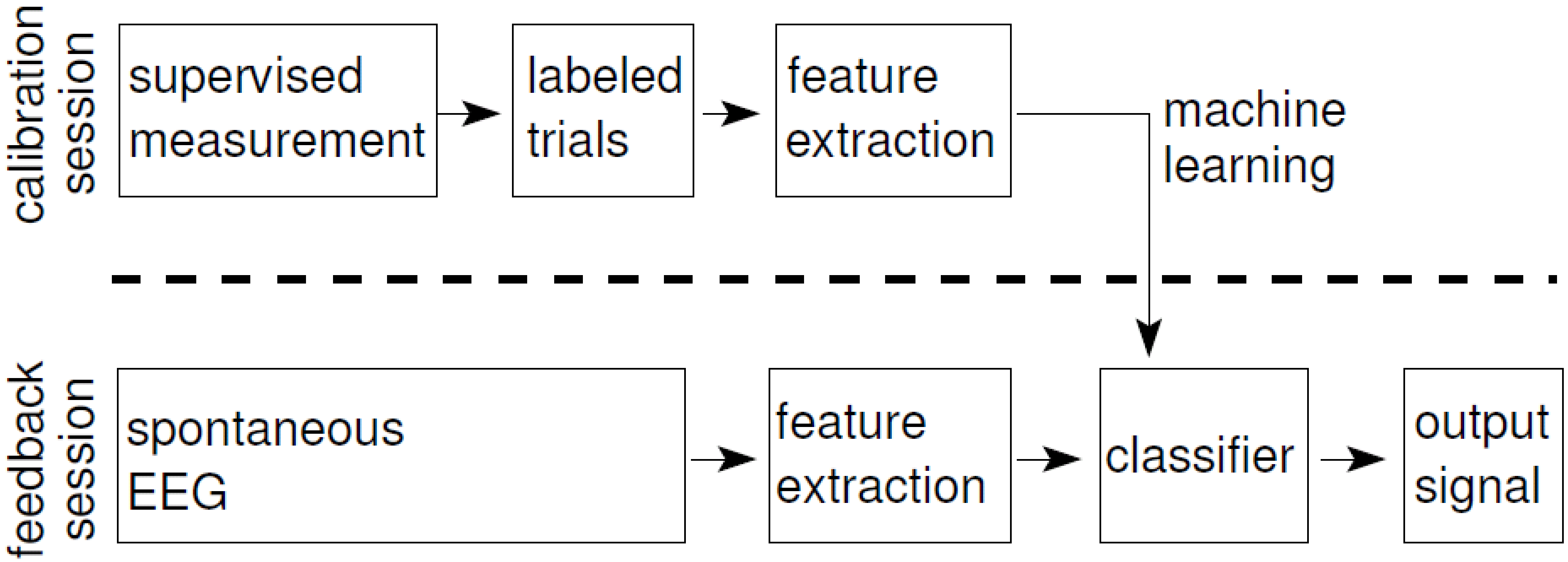
Playing with BCI: training session (20 min)

Machine learning approach to BCI: infer prototypical pattern

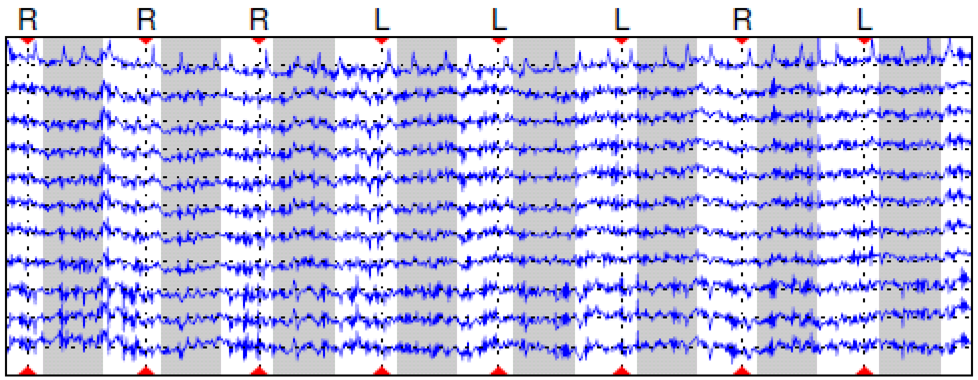


Inference by CSP Algorithm

BCI with machine learning: feedback

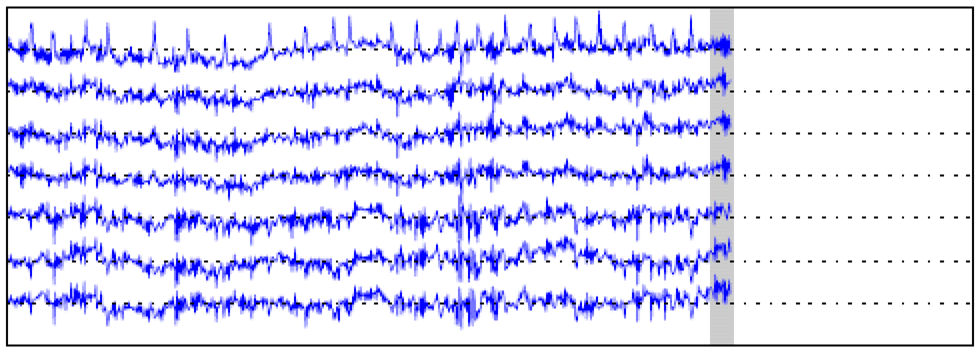


offline: calibration (10–20 minutes)



collect training samples

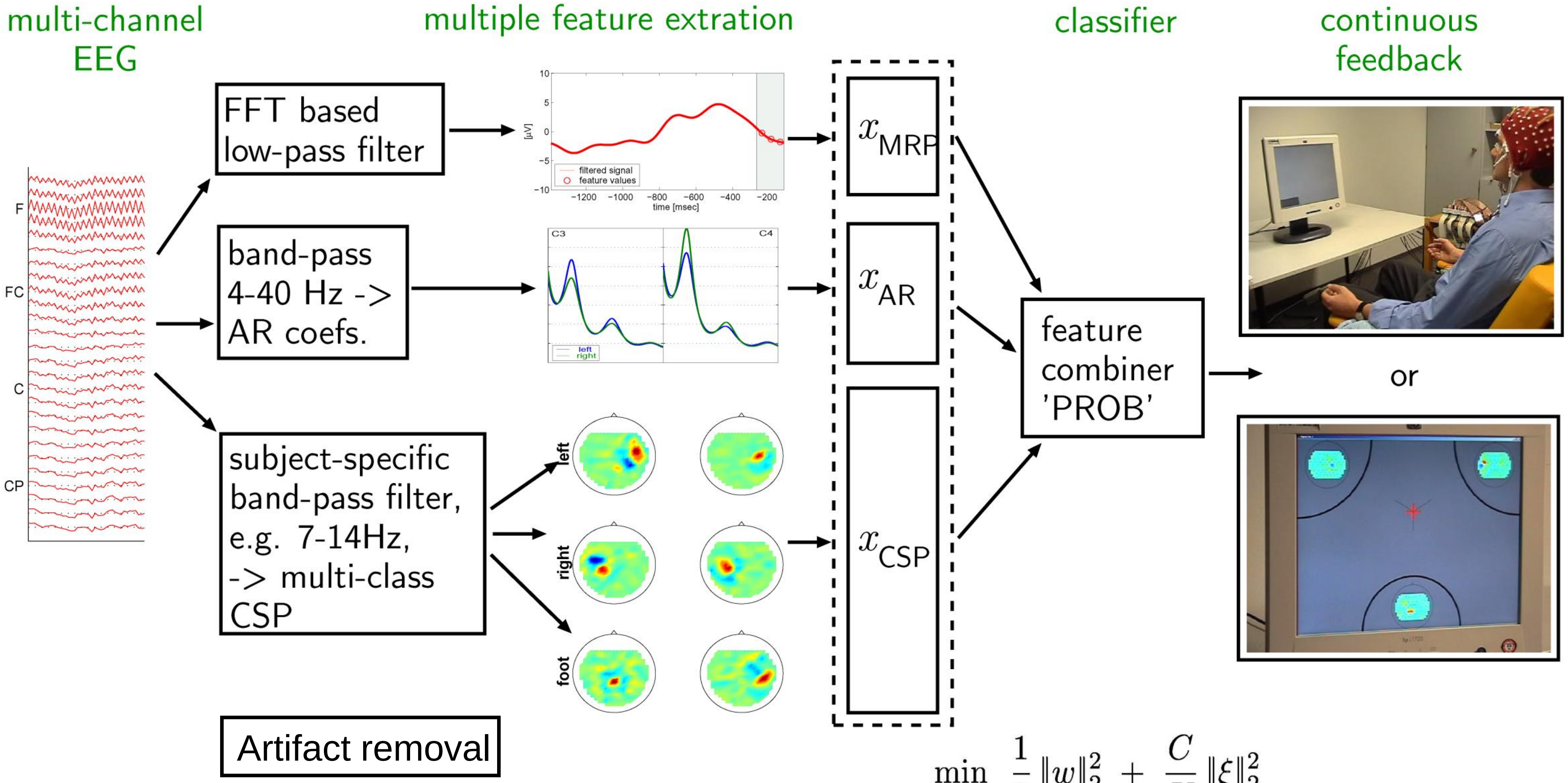
online: feedback (up to 6 hours)



classification of sliding windows ($\leq$ 1s)

Lecture Blankertz here

BBCI Set-up

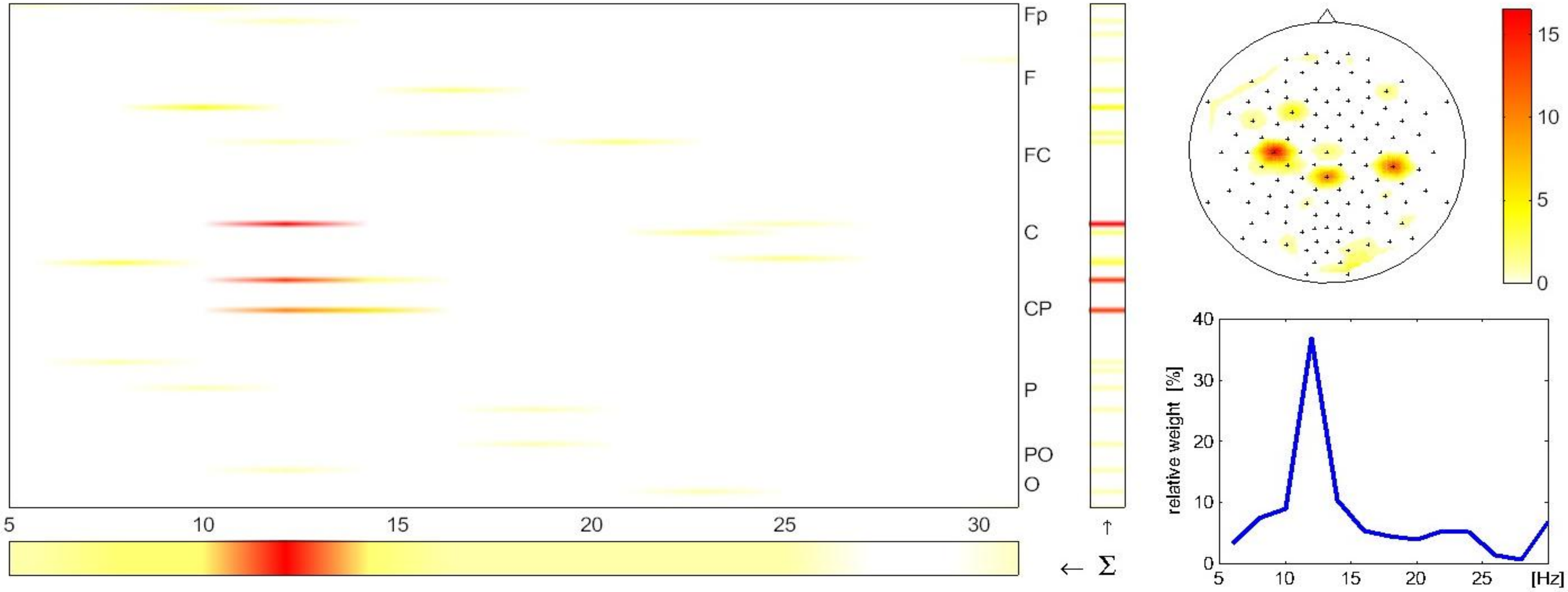


$$\min_{w,b,\xi} \; \frac{1}{2}\,\|w\|_2^2 \; + \; \frac{C}{K}\,\|\xi\|_2^2$$

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

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

What can Machine Learning tell us about physiology?



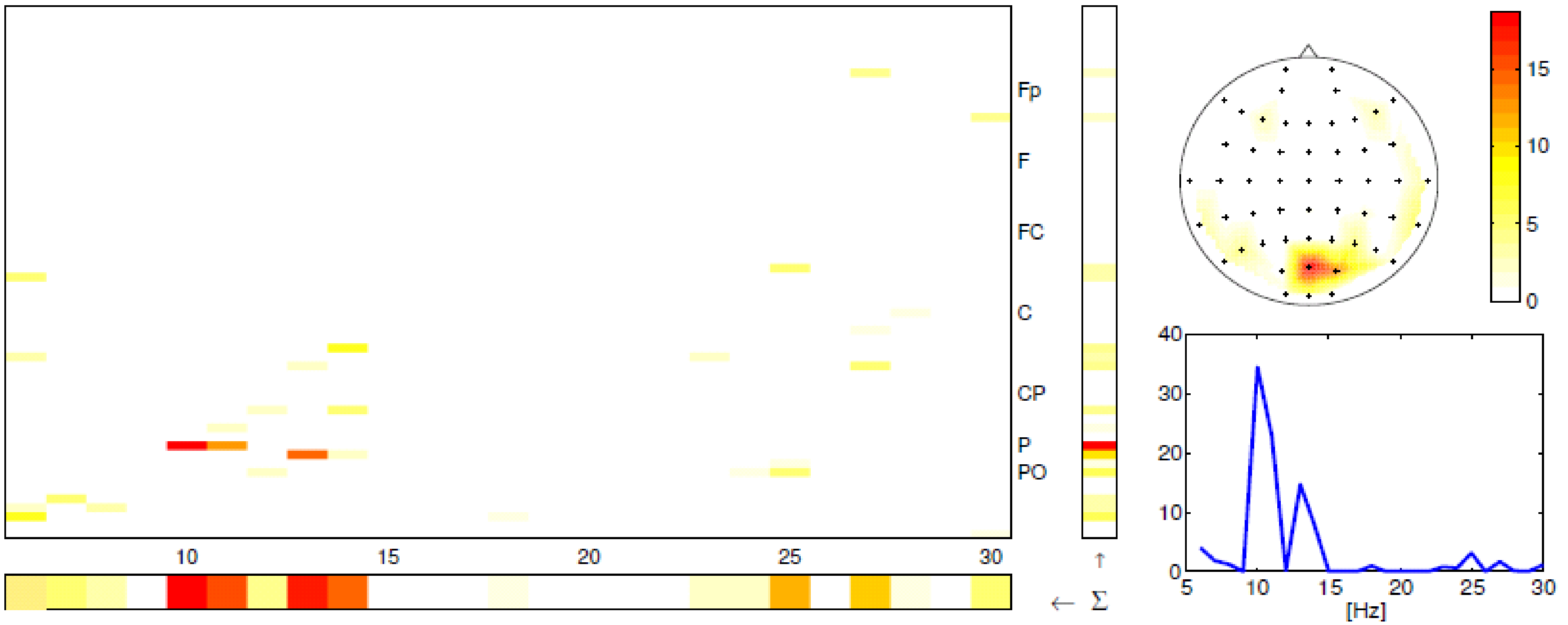
$$\min_{w,b,\xi} \; \frac{1}{2}\|w\|_1 \; + \; \frac{C}{K}\|\xi\|_1$$

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

[cf. Blankertz et al. 2001, 2006]

ML for knowledge discovery

Results for a Linear Programm Machine (LPM) for the classification **stress** vs. **no stress** periods



$$\min_{w,b,\xi} \ {}^1\!/_2 \, \|w\|_1 \ + \ {}^C\!/_K \, \|\xi\|_1 \quad \text{subject to}$$

$$y_k(w^\top x_k + b) \geqslant 1 - \xi_k, \quad \text{and} \quad \xi_k \geqslant 0 \text{ for } k = 1, \ldots, K.$$

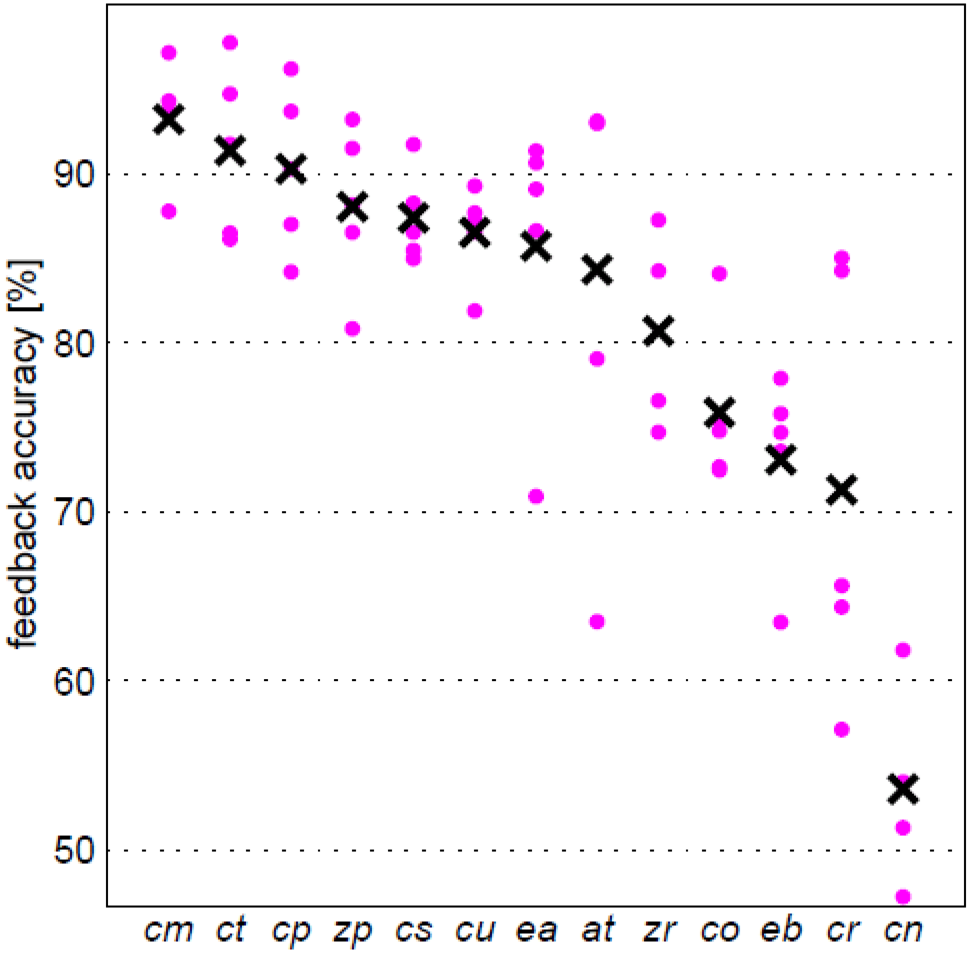
Results: Exploring the limits of untrained users

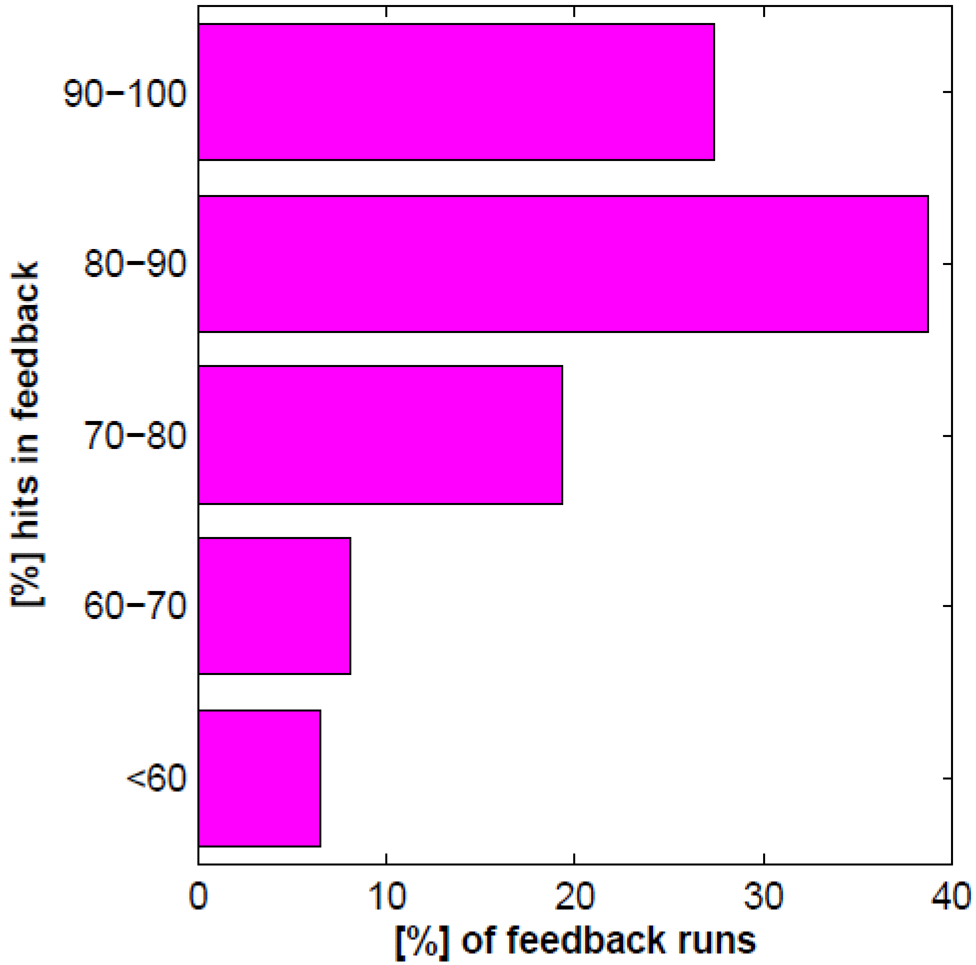
In **1D Cursor Control** 3/6 subjects achieved an information trafer rate (ITR) of more than 30 bits per minute ([2]).

sbj.	**cls.**	acc [%]		cursor ctrl	
		cal.	fb.	overall	peak
al	LF	98.0	98.0	24.4	35.4
ay	LR	97.6	95.0	22.6	31.5
aw	RF	95.4	80.5	5.9	11.0
aa	LR	78.2	88.5	17.4	37.1
av	LF	78.1	90.5	9.0	24.5
$\varnothing$		89.5	90.5	15.9	27.9

- Subject *al* spelled 135 letters in 30 minutes, i.e. 4.5 chars/min with a simple binary speller.
- With the advanced BBCI text entry system Hex-o-Spell he achieved up to 8 chars/min, see [3].

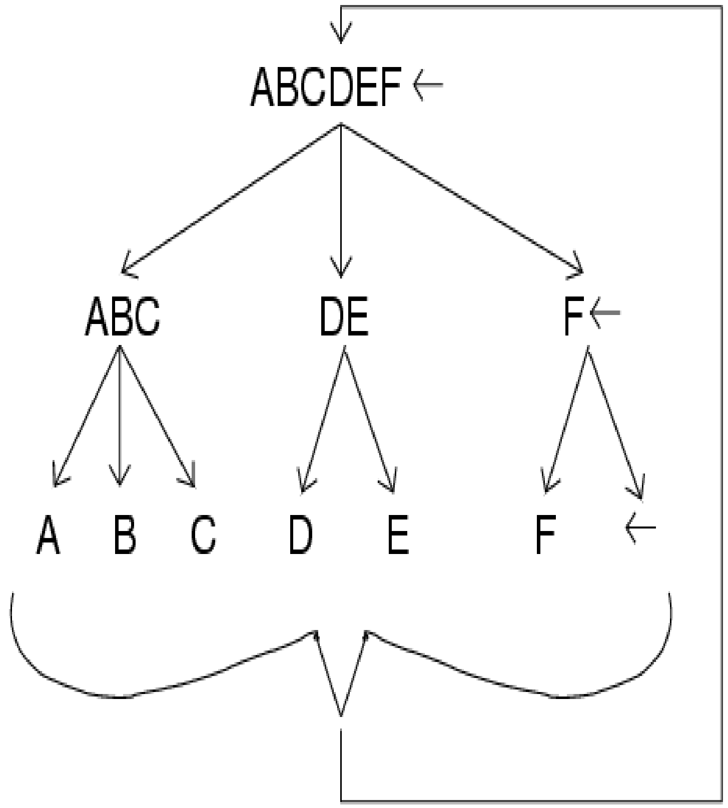
BCI: 1st session performance for novices



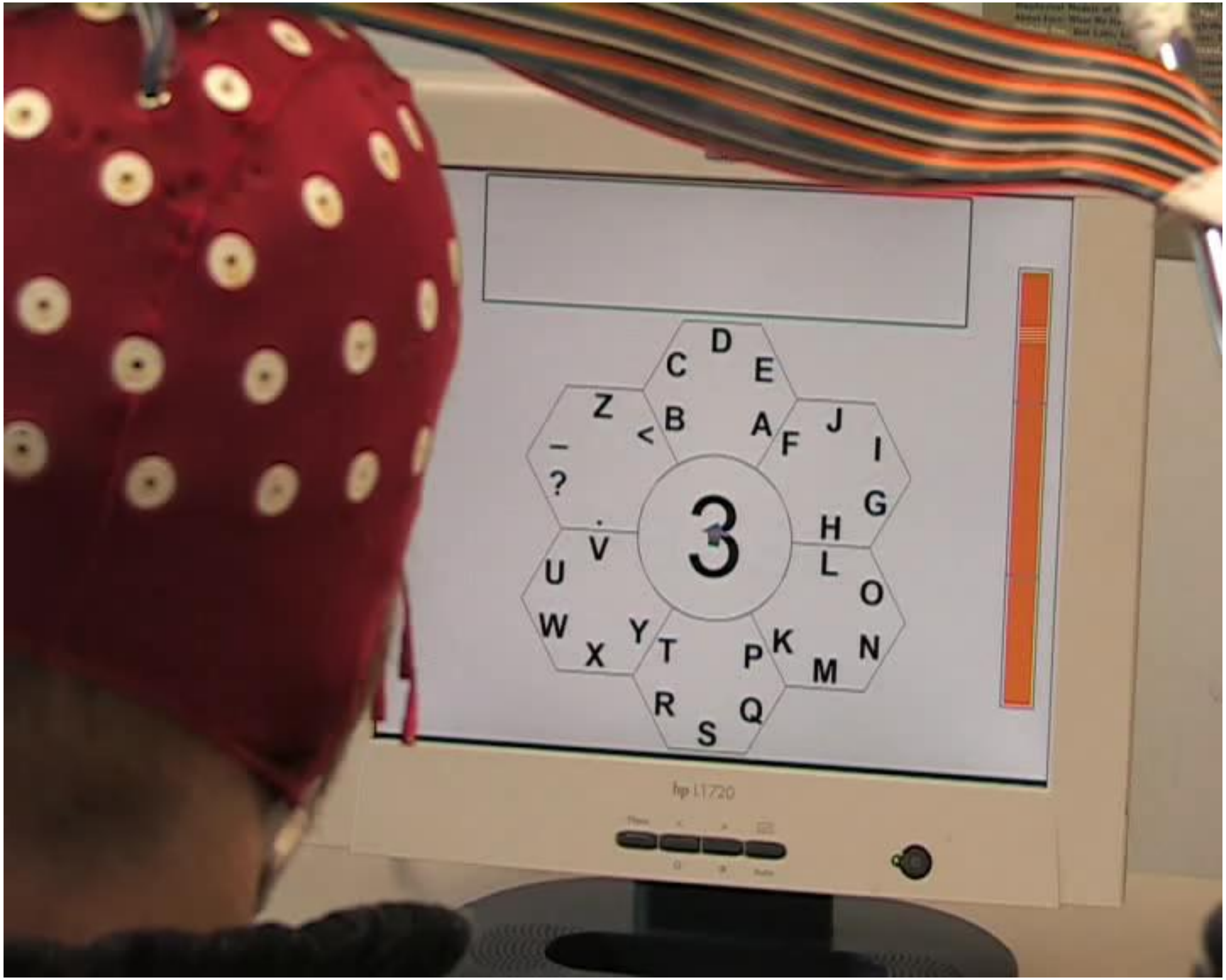


- Performance (1D cursor control) is top in international comparison.
- Still, there is a non-negligeble portion of *illiterates*.
- Also for the majority of subjects, performance is critically low for most applications.
- BCI systems based on evoked potentials typically have less *illiterates*.

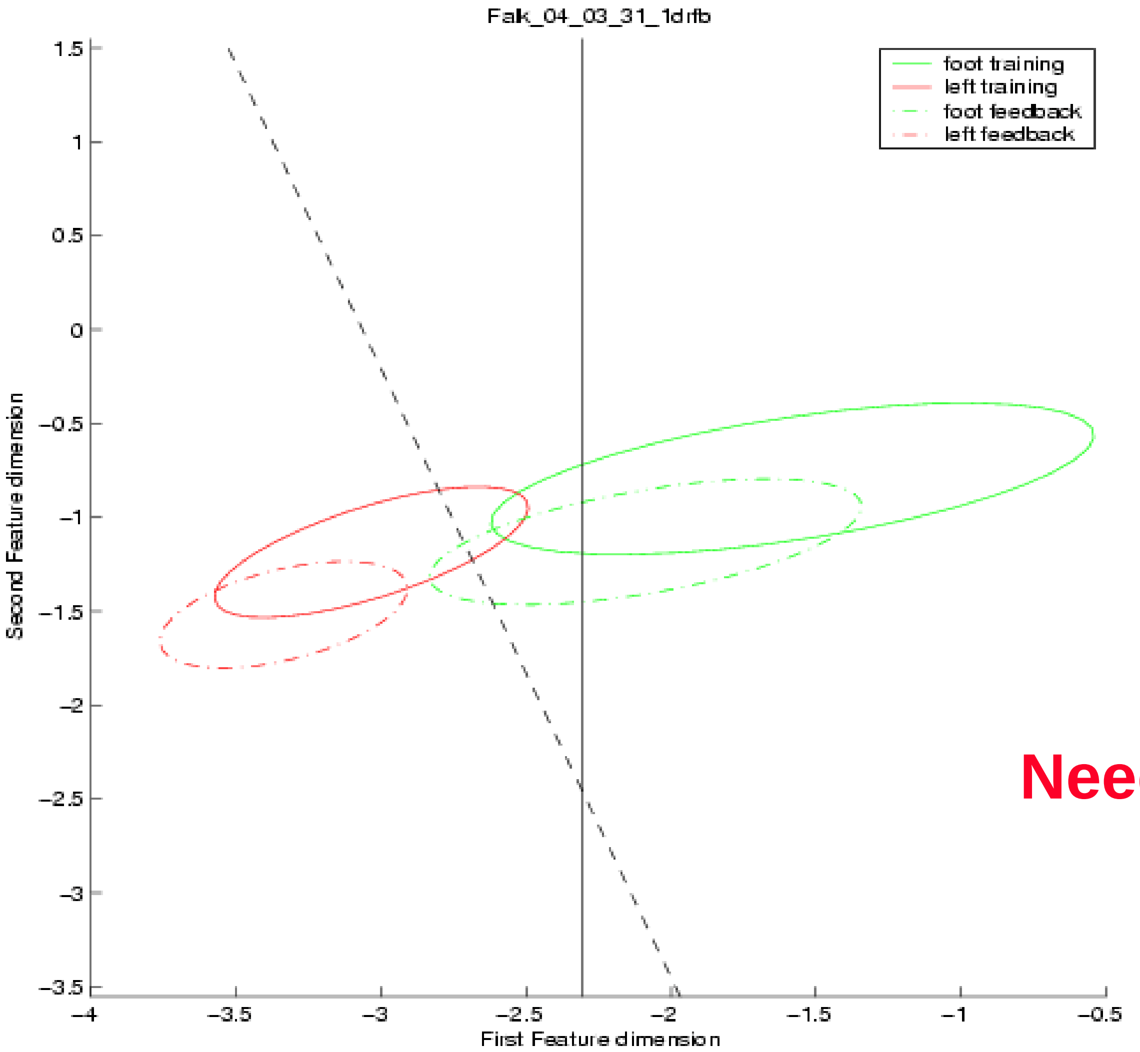
Spelling with BBCI: a communication for the disabled I



Spelling with BBCI: a communication for the disabled II



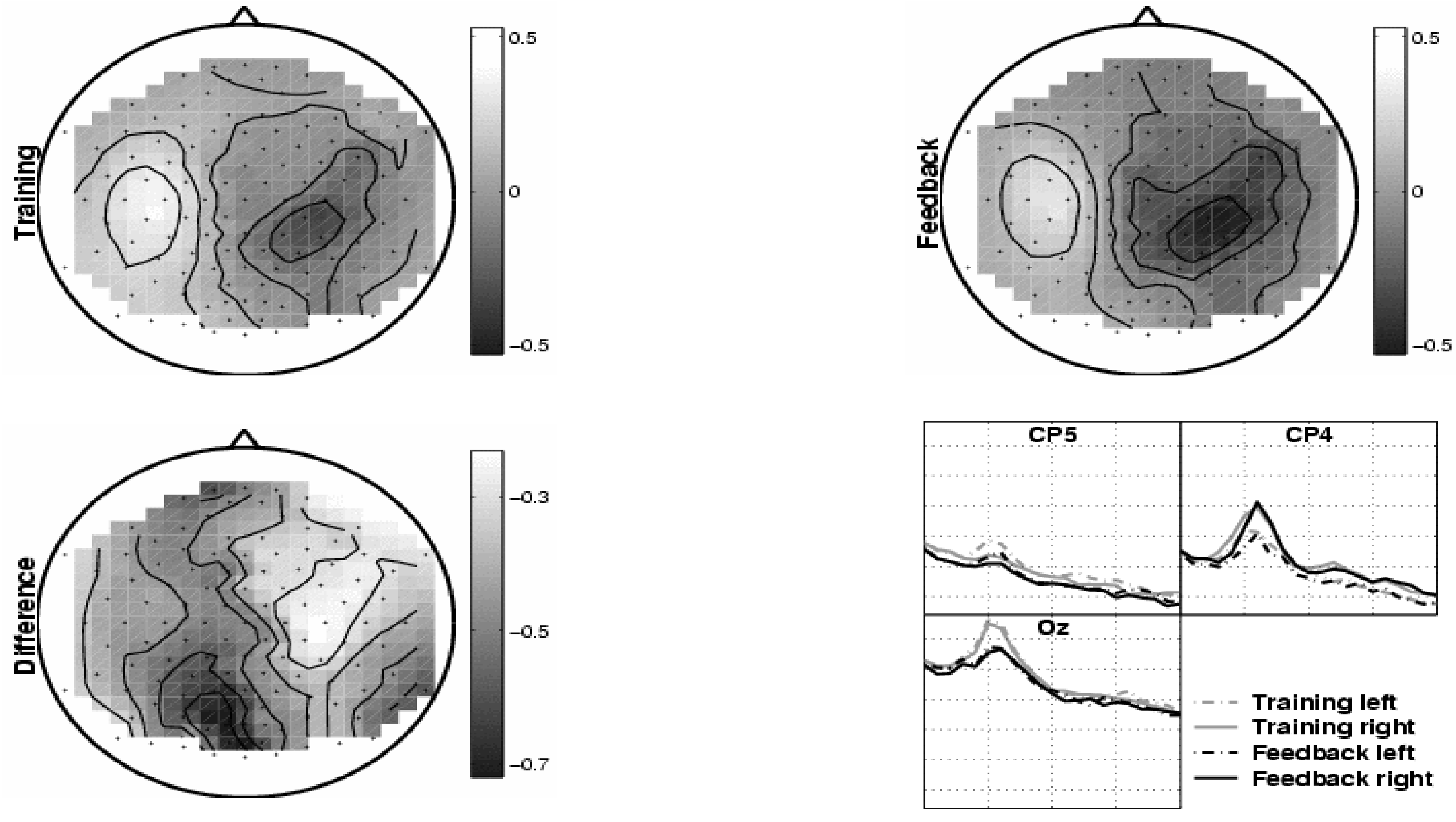
Variance IV: covariate shift: from training to feedback



Need for adaptation !

[cf. Sugiyama & Müller 2005, Shenoy et al. 2005, Sugiyama et al. 2007]

Neurophysiological analysis



[cf. Krauledat et al. submitted]

Weighted Linear Regression for covariate shift compensation

Given training samples

$$\{(\boldsymbol{x}_i, y_i) \mid y_i = f(\boldsymbol{x}_i) + \epsilon_i\}_{i=1}^{n}$$

for some function f and linearly independent basis functions $\Phi = \{\varphi_i(\boldsymbol{x})\}_{i=1}^{p}$, find

$\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_p^*)^\top$ which minimizes

$$\min_{\{\alpha_i\}_{i=1}^{p}} \left[\sum_{i=1}^{n} w(\boldsymbol{x}_i) \left(\hat{f}(\boldsymbol{x}_i) - y_i \right)^2 + \langle \boldsymbol{R}\boldsymbol{\alpha}, \boldsymbol{\alpha} \rangle \right].$$

$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{p} \alpha_i \varphi_i(\boldsymbol{x})$, choosing $w(\boldsymbol{x}_i) = \frac{p_{fb}(\boldsymbol{x}_i)}{p_{tr}(\boldsymbol{x}_i)}$ yields **unbiased** estimator even under convariate shift

[cf. Sugiyama & Müller 2005, Sugiyama et al. 2007]

BCI Illiteracy

Percentage of ~20% of naïve users:

BCI accuracy does not reach level criterion,
i.e., control not accurate enough to control applications

Screening Study (N=80):

Cat I: good calibration (cb), good feedback (fb)

Cat II: good cb, no good fb

Cat III: no good cb

design a predictor !

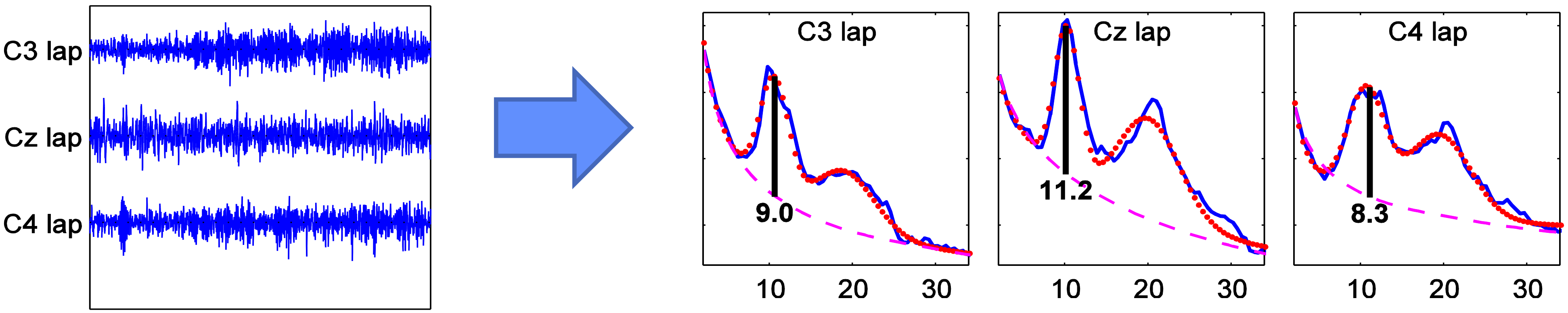
SMR-Predictor

Calculate the power spectral density (PSD) in three Laplacian channels C3, Cz, C4 under rest cond.

Model each resulting curve by $g = g_1 + g_2$, with

$$g_1 = g_1\,(x, \lambda, \boldsymbol{k}) \quad = k_1 + k_2 \,/\, x^{\lambda} \qquad \text{(estimated noise)}$$

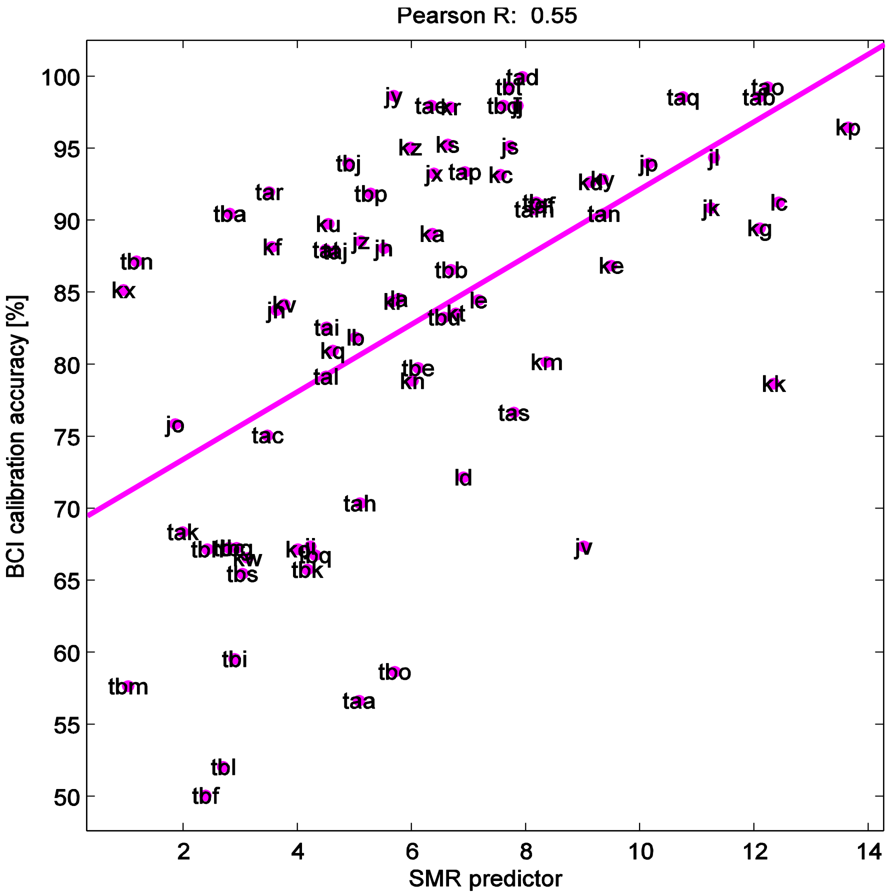
$$g_2 = g_2\,(x, \mu, \sigma, \boldsymbol{k}) = k_3\, \varphi(x\,;\, \mu_1\,,\, \sigma_1) + k_4\, \varphi\,(x\,;\, \mu_2\,,\, \sigma_2) \qquad \text{(2 peaks)}$$



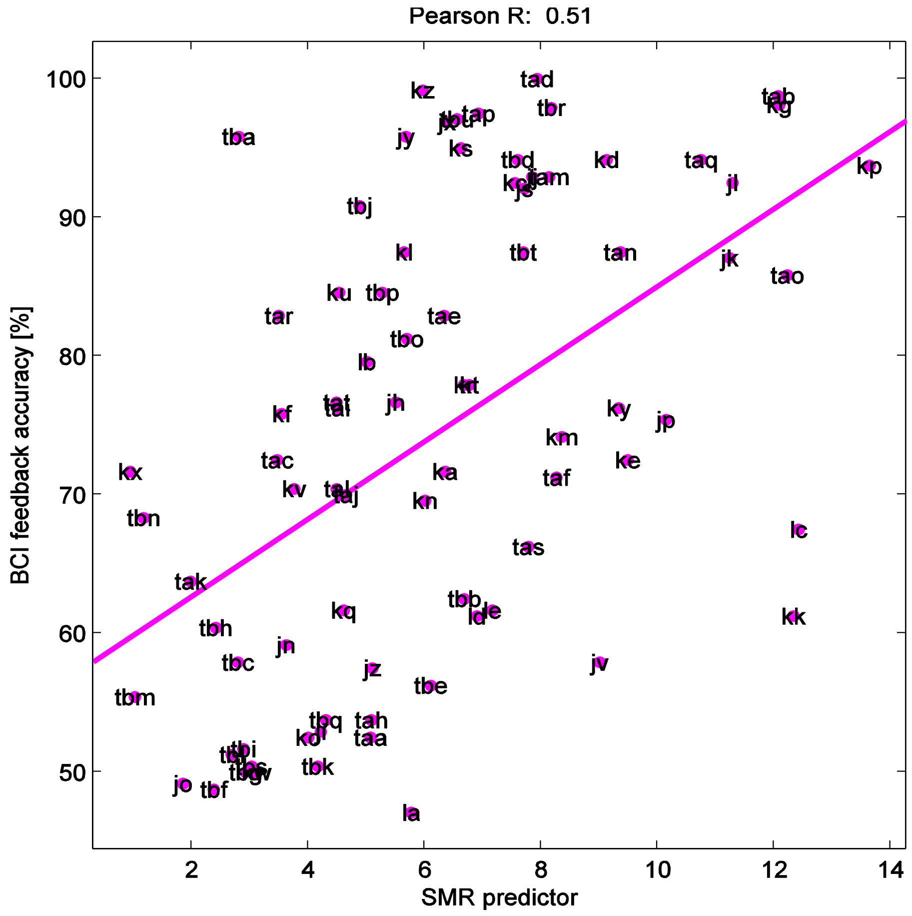
Proposed predictor: Average height of the larger peak

SMR-Predictor (Results)

Pearson R = 0.55



Pearson R = 0.51



As much as $R^2 = 30\%$ (calibration) and $R^2 = 26\%$ (feedback) of variability in BCI accuracy can be explained by the SMR predictor in our study sample!

[cf. Blankertz, Kübler, Müller et al. 2009]

Approach to „Cure“ BCI Illiteracy

Runs 1 - 3

fixed Laplace

supervised

- Direct feedback -> Unspecific LDA classifier.
- Each trial, perform adaptation of the cls.
- Features: log band power (alpha and beta).
- Laplacian channels C3, C4 and Cz.

Runs 4 - 6

CSP + sel.Lap.

- Compute CSP and sel. Laps. from runs 1-3.
- Fixed CSP filters, automated laps. selection.
- Each trial retrain the classifier.

Runs 7 - 8

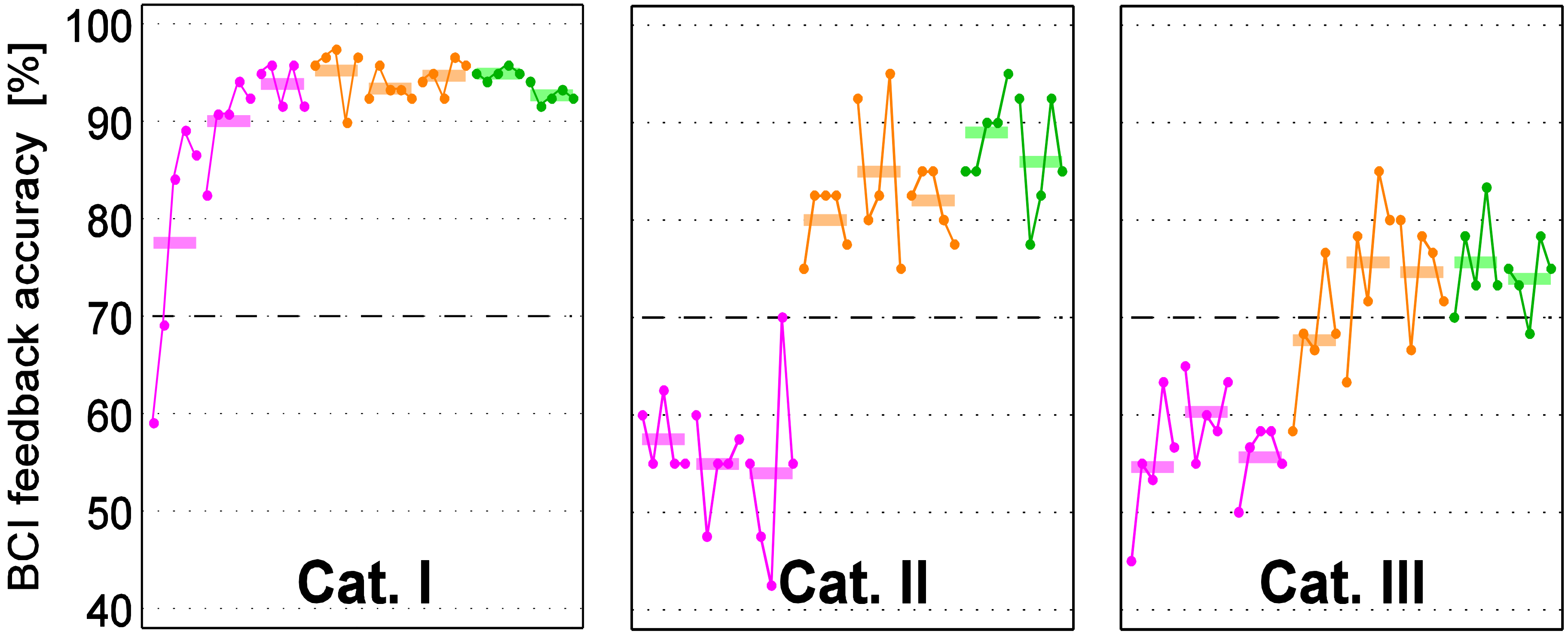
CSP

unsupervised

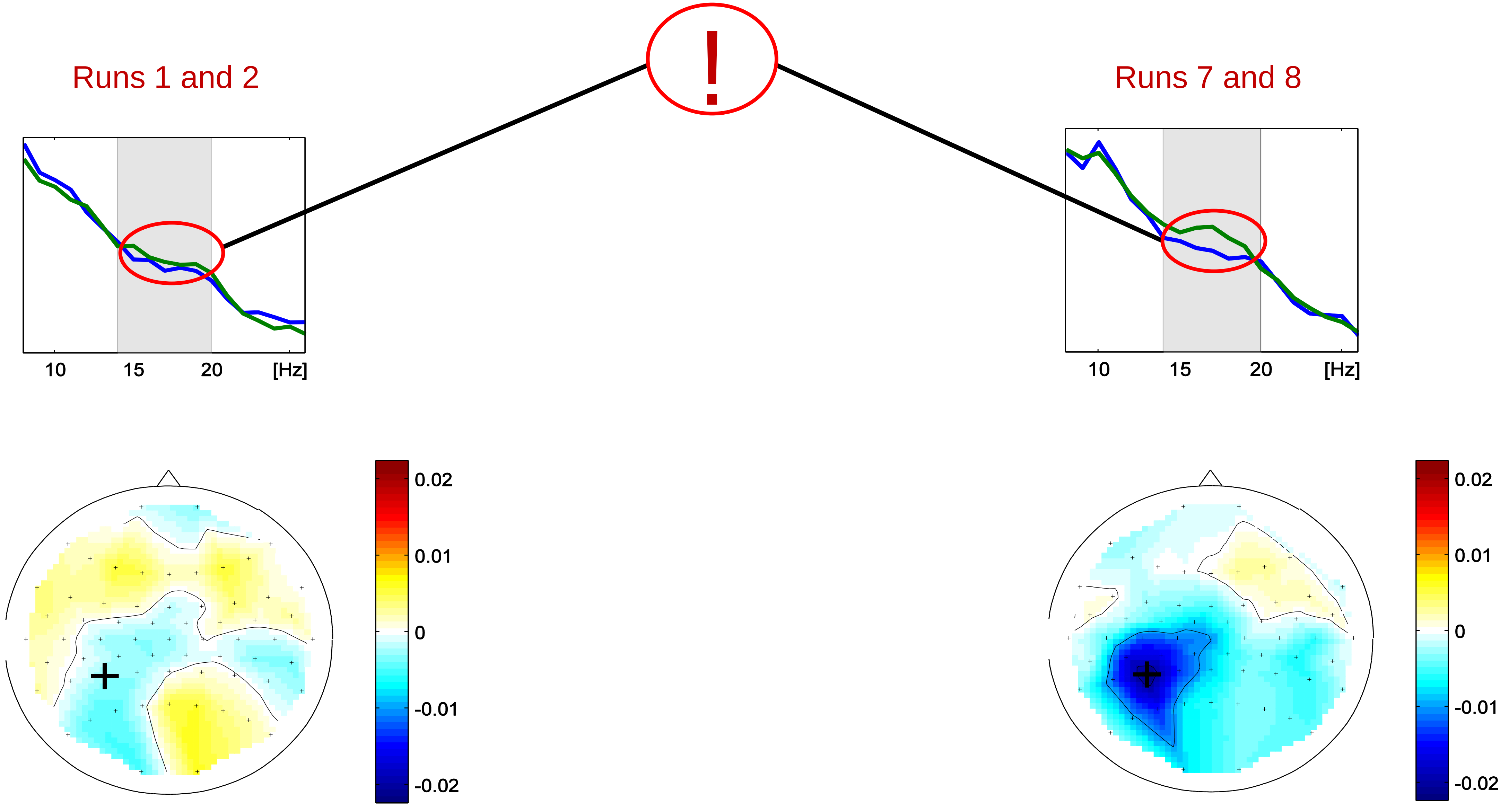
- Compute CSP from runs 4-6.
- Perform unsupervised adaptation of pooled mean.
- Update the bias of the classifier.

[cf. **Vidaurre**, Blankertz, Müller et al. in preparation]

Results (Grand Averages)

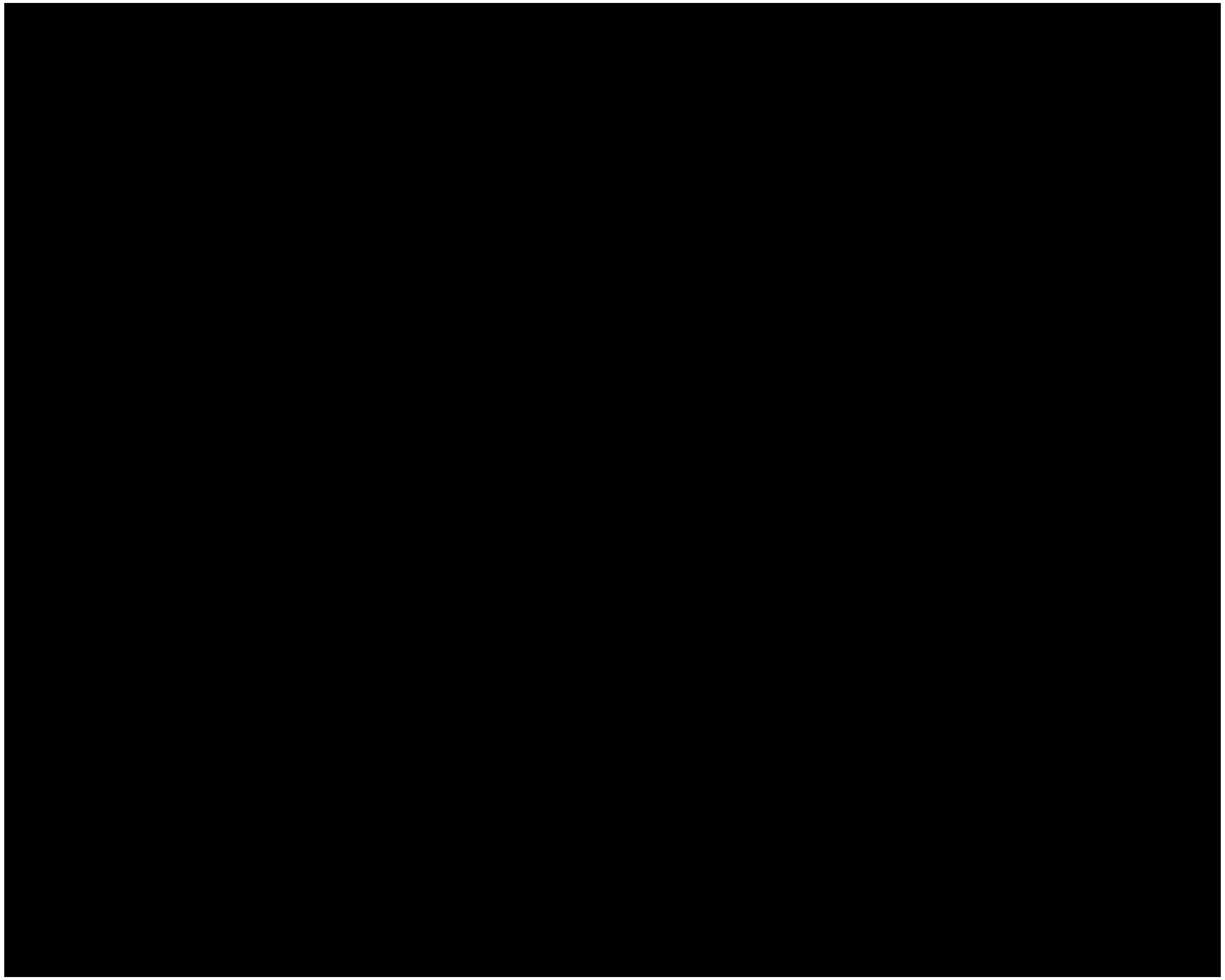


Example: one subject of Cat. III



[cf. Vidaurre, Blankertz, Müller et al. 2009]

Real Man Machine Interaction



[Tangermann, Müller et al 2009]

Harvest from BCI-gaming

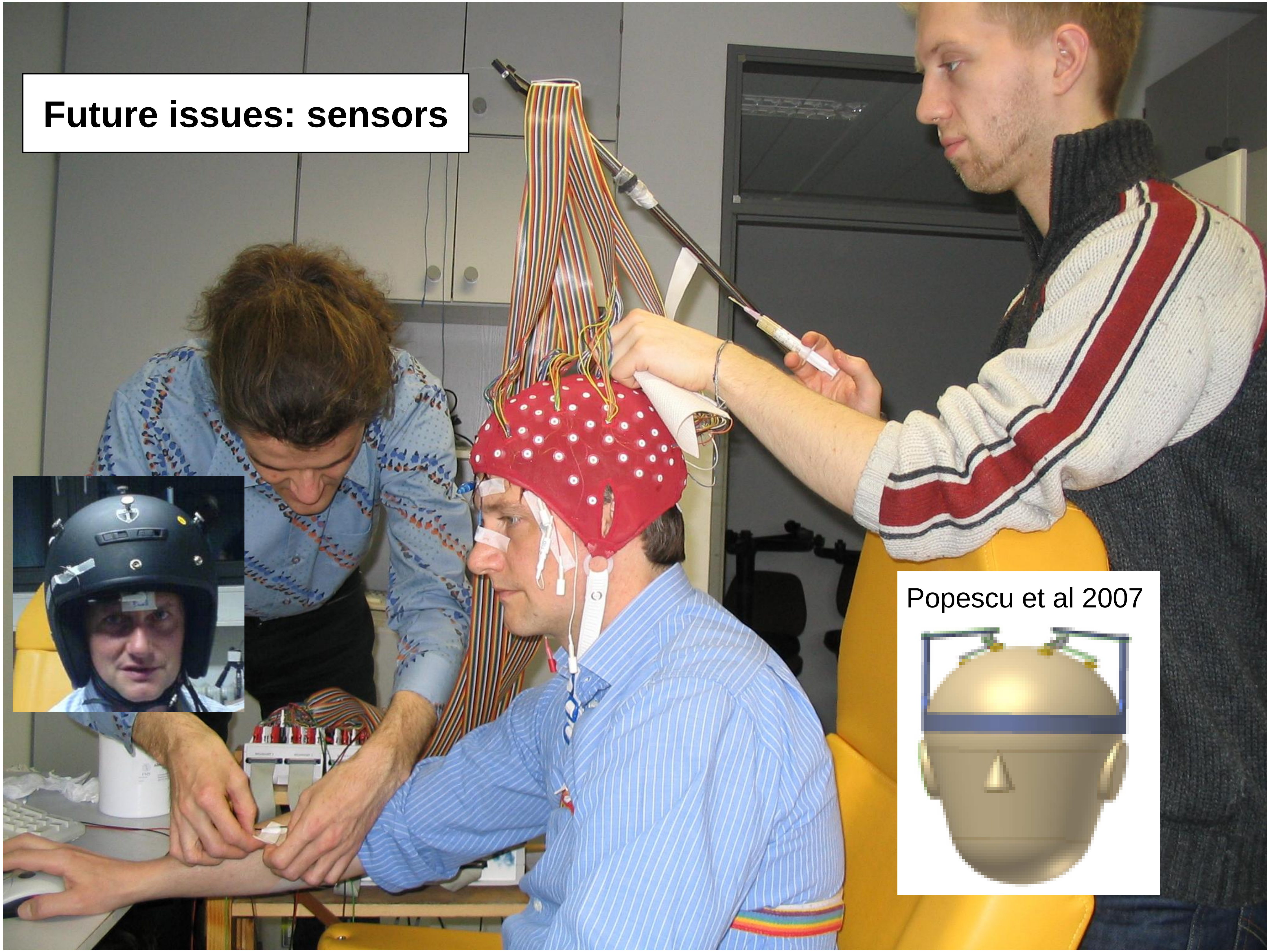
Limits of BCI-reaction time?

Research platform:
BCI gaming +
Machine Learning +
Single Trial Analysis (MSM)

Mental states
JUST BEFORE
success or failure?

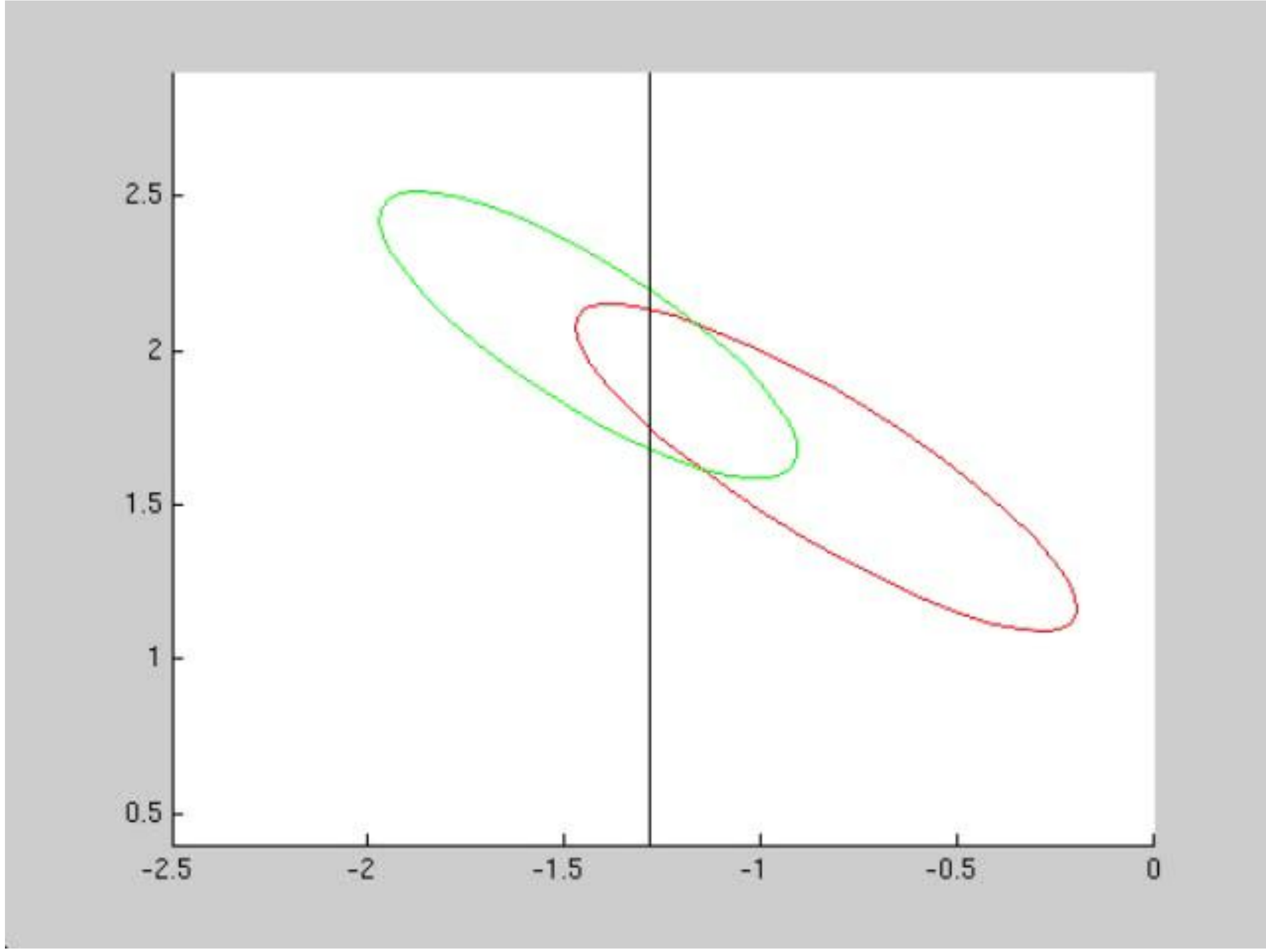
Dynamics:
limits of temporal
control precision?

Future issues: sensors



Popescu et al 2007

Future Issues: Shifting distributions within experiment



Conclusion

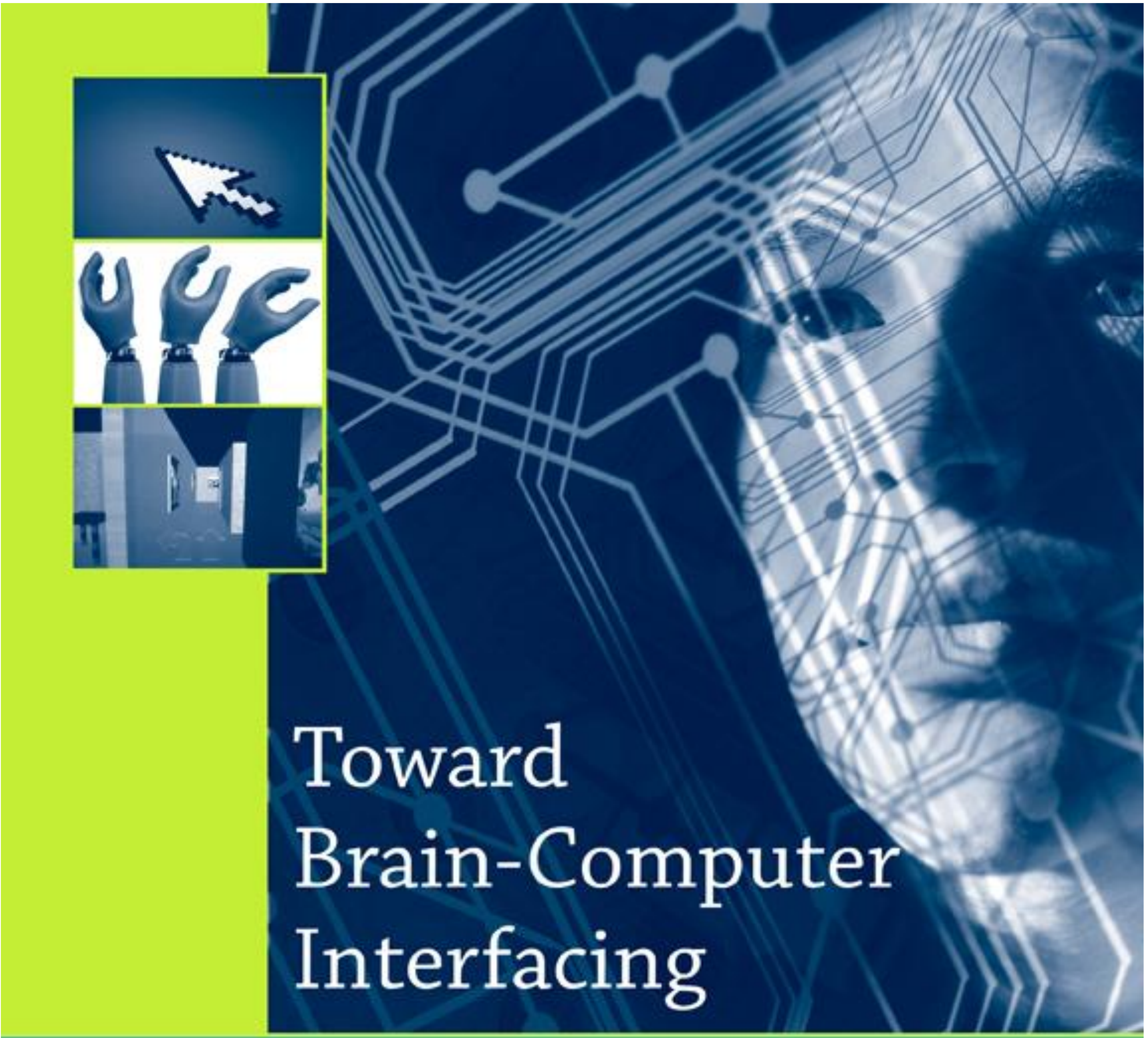
- BBCI: non-invasive with high Information transfer rates for the **Untrained**
- BBCI: Untrained, Calibration < 20min, data analysis <<10min, BCI experiment
- 5-8 letters/min mental typewriter on CeBit 06. Brain2Robot@Medica 07, lNdW 09
- Machine Learning and modern data analysis is of central importance for BCI **et al**
- Applications:

  Rehabilitation: **TOBI EU IP**

  Computational Neuroscience: **Bernstein Centers Berlin**

  Man Machine Interaction: using BCI as a measuring device: **brain@work**
- BBCI Sensors, software: **IDA spinoffs**
- towards no training, non-cooperative behavior
- ‚illiterates‘, nonstationarity, wireless EEG

FOR INFORMATION SEE:
www.bbci.de



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

BBCI team:

Gabriel Curio
Florian Losch
Volker Kunzmann
Frederike Holefeld
Vadim Nikulin@Charite

Andreas Ziehe
Florin Popescu
Christian Grozea
Steven Lemm
Motoaki Kawanabe
Guido Nolte@FIRST

Yakob Badower@Pico Imaging
Marton Danozcy

Benjamin Blankertz
Michael Tangermann
Claudia Sannelli
Carmen Vidaurre
Bastian Venthur
Siamac Fazli
Martijn Schreuder
Matthias Treder
Stefan Haufe
Thorsten Dickhaus
Frank Meinecke
Paul von Bünau
Marton Danozcy
Felix Biessmann
Klaus-Robert Müller@TUB

Matthias Krauledat
Guido Dornhege
Roman Krepki@industry

Collaboration with: U Tübingen, Bremen, Albany, TU Graz, EPFL, Daimler, Siemens, MES, MPIs, U Tokyo, TIT, RIKEN, Bernstein Focus Neurotechnology, Bernstein Center for Computational Neuroscience Berlin, picoimaging, Columbia, CUNY

Funding by: EU, BMBF and DFG

Overview of BCI Competitions

BCI competition I	BCI competition II
December 2001 – June 2002	December 2003 – June 2004
3 datasets	6 datasets
10 submissions	59 submissions
[Sajda et al., 2003]	[Blankertz et al., 2004]

BCI Competition III

- Dec 12th 2004 – May 31st 2005
- announcement of the results: between June 14th and 19th 2005
- 8 datasets from 5 different BCI groups with different tasks

For BCI IV Competition see www.bbci.de

berlin
brain computer
interface

FOR INFORMATION SEE: www.bbci.de

Machine Learning open source software initiative: MLOSS see

www.jmlr.org

Machine Learning and Signal Processing Tools for Brain-Computer Interfacing

Benjamin Blankertz[1,2]
Klaus-Robert Müller[1]

[1]Machine Learning Laboratory, Berlin Institute of Technology
[2]Fraunhofer FIRST (IDA)

blanker@cs.tu-berlin.de

08-Jul-2009

Topic of This Section

Method:

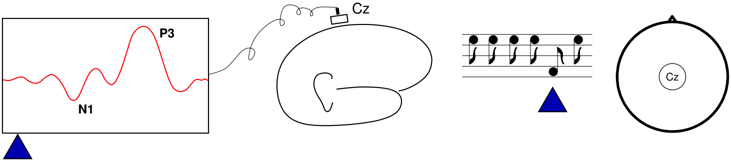
- classification of **spatio-temporal** features;
- *shrinkage* of the sample covariance matrix to counterbalance the estimation bias

Application:

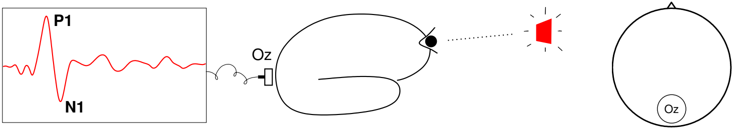
- classification of single-trial ERPs in an attention-based speller

Some Neurophysiological Background

An infrequent stimulus in a series of standard stimuli evokes a P300 component at central scalp position *if attended*:



The presentation of a visual stimulus elicits a Visual Evoked Potential (VEP) in visual cortex *if focused*:



Experimental Design

Classic Matrix Speller

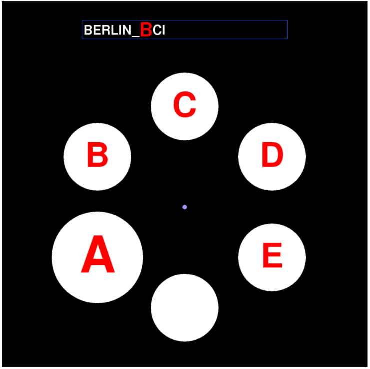
Attention-based Hex-o-Spell

See **Poster W07** (Treder et al.) for a investigation of *overt* vs. *covert* attention and a comparison of those two speller designs.

Experimental Design

Classic Matrix Speller

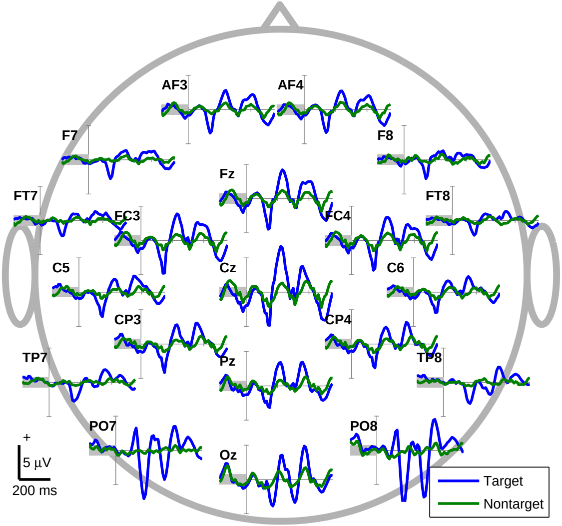
Attention-based Hex-o-Spell



See **Poster W07** (Treder et al.) for a investigation of *overt* vs. *covert* attention and a comparison of those two speller designs.

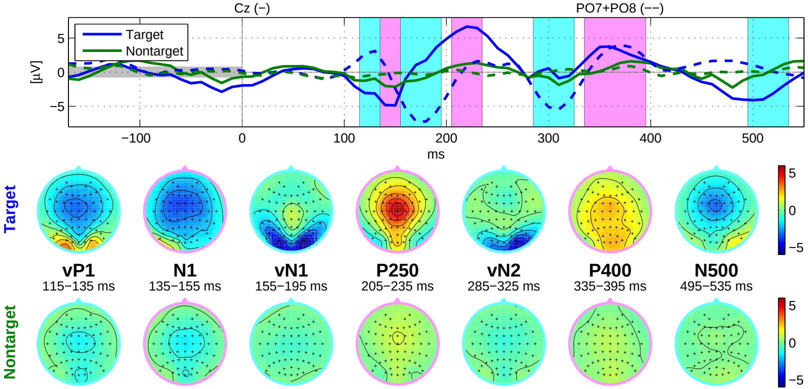
Single-subject ERPs in Hex-o-Spell

Data set for illustration of classification methods:



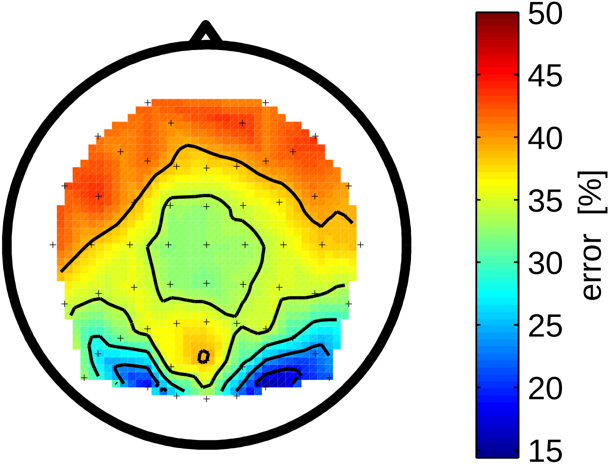
Topographies of ERP Components

There are several ERP components that can be used to determine the attended symbol:



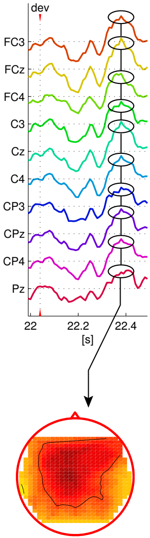
Classification of Temporal Features

As a first step: classification on raw time courses (115–535 ms) in single channels. The result is displayed as scalp map:

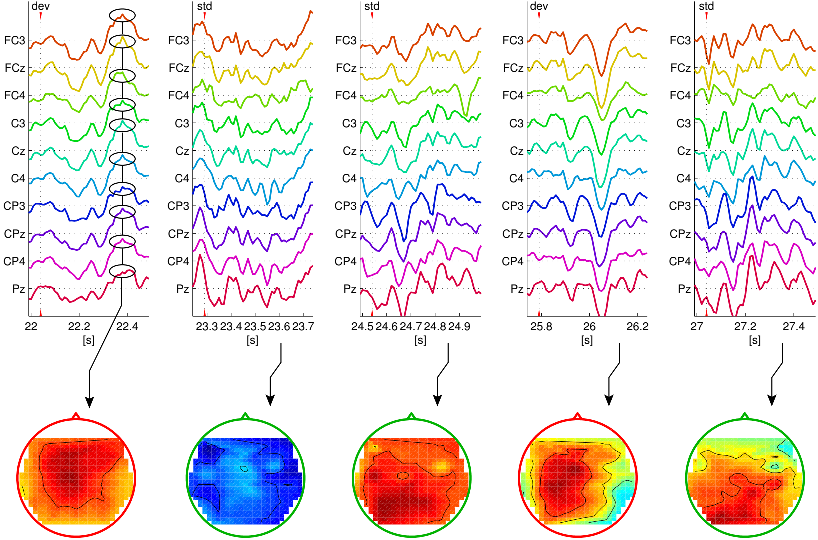


BBCI

Extraction of Spatial Features



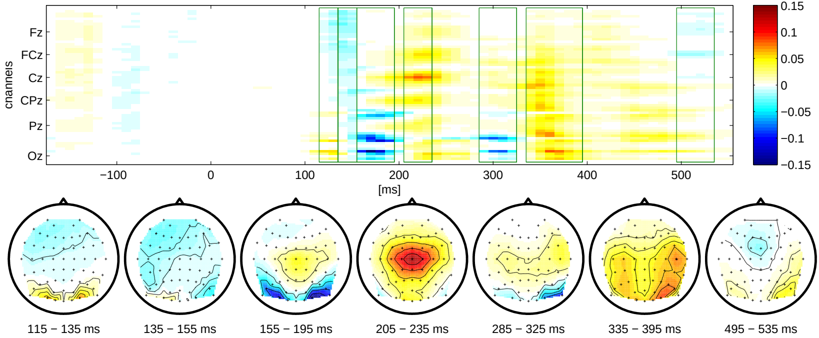
Extraction of Spatial Features



BBCI

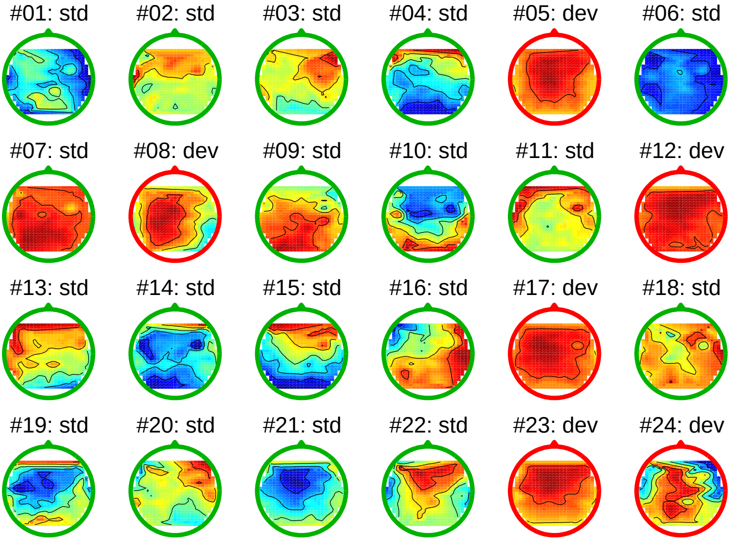
The r^2-Matrix of Differences

The temporal and spatial structure of the difference between ERPs of different conditions can be investigated by the signed r^2-matrix:



$$r(x) := \frac{\sqrt{N_1 \cdot N_2}}{N_1 + N_2} \, \frac{\mathrm{mean}\{x_i \,|\, y_i = 1\} - \mathrm{mean}\{x_i \,|\, y_i = 2\}}{\mathrm{std}\{x_i\}}$$

Spatial Features



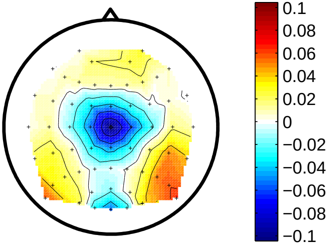
Linear Classifier as Spatial Filter

A linear classifier that was trained on *spatial features* can also be regarded as a **spatial filter**.
Let $\mathbf{w}$ be the LDA weight vector and $\mathbf{X} \in \mathbb{R}^{\#\text{chans} \times \#\text{time points}}$ be continuous EEG signals. Then

$$\mathbf{X}_f := \mathbf{w}^\top \mathbf{X} \quad \in \mathbb{R}^{1 \times \#\text{time points}}$$

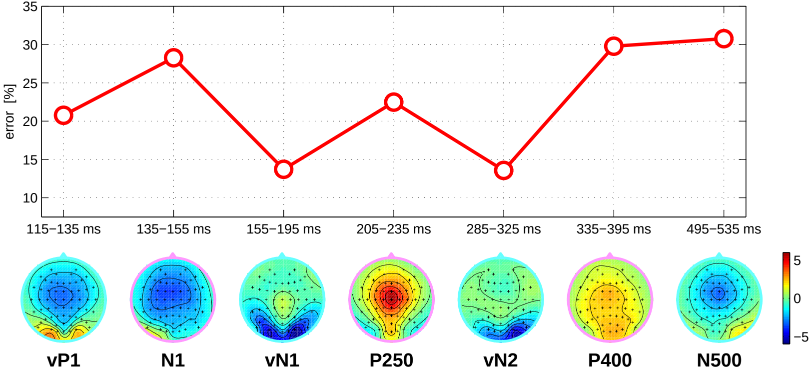
is the result of spatial filtering: each channel of $\mathbf{X}$ is weighted with the corresponding component of $\mathbf{w}$ and summed up.

The weight vector of the classifier can be display as scalp map:

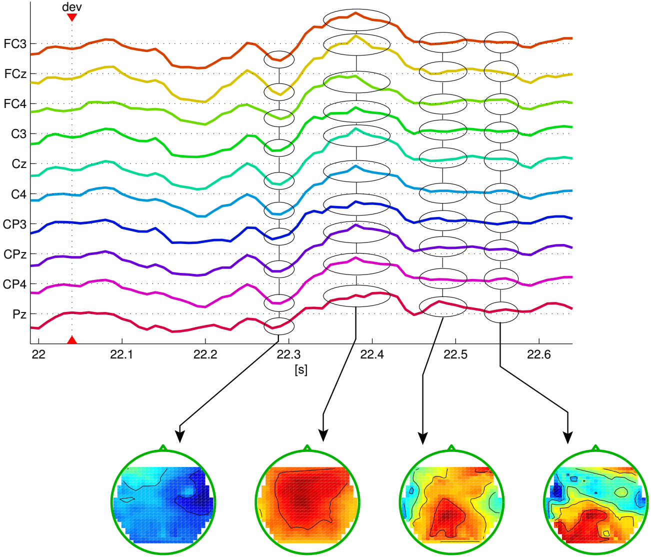


BBCI

Classification Results for Spatial Features



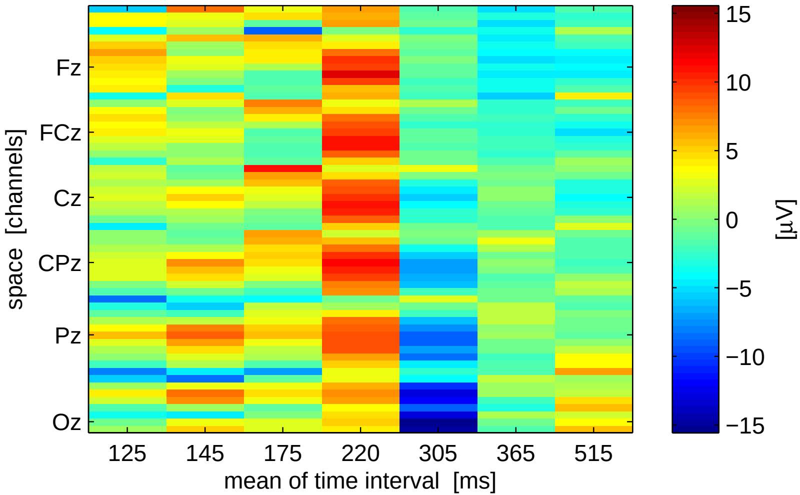
Extraction of Spatio-Temporal Features



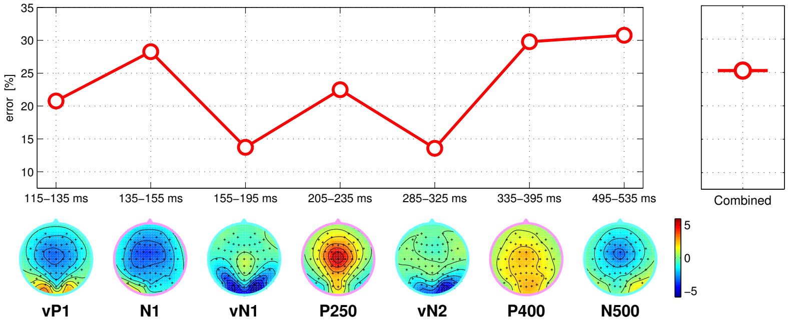
BBCI

Spatio-Temporal Features

Spatio-temporal features are typically high-dimensional (here 59 EEG channels $\times$ 7 time intervals $=$ 413 dimensional features):



Classification Result for Spatio-Temporal Features



Although information was added, classification on the concatenated feature becomes worse: *overfitting*.

BBCI

Bias in Estimating Covariances

Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ be n vectors drawn from a d-dimensional Gaussian distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$.
For classification μ and $\mathbf{\Sigma}$ have to be estimated from the data:

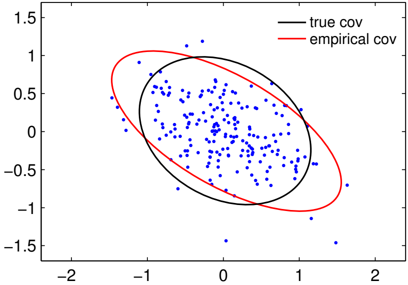
- $\hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$
- $\hat{\mathbf{\Sigma}} = \frac{1}{n-1} \sum_{k=1}^{n} (\mathbf{x}_k - \hat{\mu})(\mathbf{x}_k - \hat{\mu})^\top$

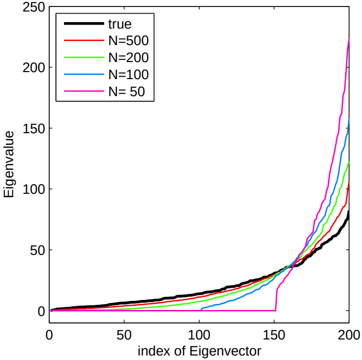
But, if the number of samples n is not large relative to the dimension d, the estimation is error-prone.
There is a systematical bias:

- Large Eigenvalues of $\hat{\mathbf{\Sigma}}$ are too large
- Small Eigenvalues of $\hat{\mathbf{\Sigma}}$ are too small

This affects, e.g., classification with LDA:
Normal vector of LDA: $w = \hat{\mathbf{\Sigma}}^{-1}(\mu_1 - \mu_2)$.

Bias in Estimating Covariances (2)





A Remedy for Classification

A simple way that can partly fix the bias is **shrinkage**: the empirical covariance matrix is modified to be more spherical.
In LDA the empirical covariance matrix $\hat{\mathbf{\Sigma}}$ is replaced by

$$\tilde{\mathbf{\Sigma}}(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}$$

for a $\gamma \in [0,\, 1]$ and ν defined as average Eigenvalue $\mathrm{trace}(\mathbf{S}_i)/d$.

A Remedy for Classification

A simple way that can partly fix the bias is **shrinkage**: the empirical covariance matrix is modified to be more spherical. In LDA the empirical covariance matrix $\hat{\mathbf{\Sigma}}$ is replaced by

$$\tilde{\mathbf{\Sigma}}(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}$$

for a $\gamma \in [0,\, 1]$ and ν defined as average Eigenvalue $\mathrm{trace}(\mathbf{S}_i)/d$. Since $\hat{\mathbf{\Sigma}}$ is positive semi-definite we can have an Eigenvalue decomposition $\hat{\mathbf{\Sigma}} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ with orthonormal $\mathbf{V}$ and diagonal $\mathbf{D}$. From

$$\tilde{\mathbf{\Sigma}} = (1-\gamma)\mathbf{V}\mathbf{D}\mathbf{V}^\top + \gamma\nu\mathbf{I} = \mathbf{V}\left((1-\gamma)\mathbf{D} + \gamma\nu\mathbf{I}\right)\mathbf{V}^\top$$

we see that

- $\tilde{\mathbf{\Sigma}}(\gamma)$ and $\hat{\mathbf{\Sigma}}$ have the same Eigenvectors (columns of $\mathbf{V}$)
- extreme Eigenvalues (large/small) are shrunk/extended towards the average ν.
- $\gamma = 0$ yields LDA without shrinkage, $\gamma = 1$ assumes spherical covariance matrices.

Modelselection

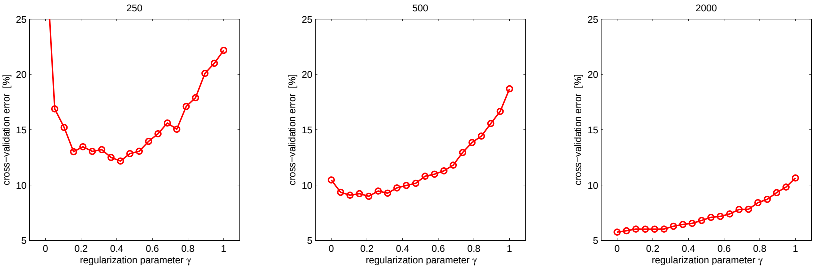
LDA with shrinkage of the empirical covariance matrix has one free parameter (γ), also called hyperparameter, that needs to be selected. There is no general way to do it.
Numerous strategies with different properties exist, e.g.

- empirical Bayes shrinkage estimator
- MDL: Minimum Description Length
- Model-selection based on cross-validation.
- ...

An easy (and also time-consuming) way is model-selection based on **cross-validation**.

Regularized LDA at Work

Cross-validation results for different sizes of training data (250, 500, 2000) for different values of the regularization parameter γ (x-axis). Features vectors have 250 dimensions.

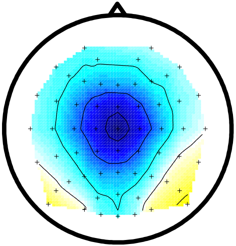


Investigating the Impact of Shrinkage

LDA: $w = \hat{\mathbf{\Sigma}}^{-1}(\mu_1 - \mu_2)$; **shrinkage:** $\tilde{\mathbf{\Sigma}}(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}$

$\gamma = 1$

$w \sim \mu_1 - \mu_2$

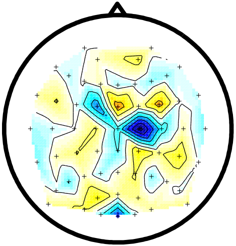


Investigating the Impact of Shrinkage

LDA: $w = \hat{\mathbf{\Sigma}}^{-1}(\mu_1 - \mu_2)$; **shrinkage:** $\tilde{\mathbf{\Sigma}}(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}$

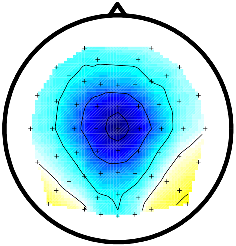
$\gamma = 0$

$w \sim \hat{\mathbf{\Sigma}}^{-1}(\mu_1 - \mu_2)$



$\gamma = 1$

$w \sim \mu_1 - \mu_2$

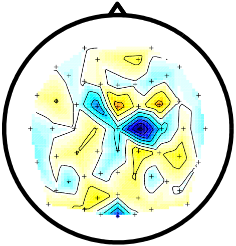


Investigating the Impact of Shrinkage

LDA: $w = \hat{\mathbf{\Sigma}}^{-1}(\mu_1 - \mu_2)$; **shrinkage:** $\tilde{\mathbf{\Sigma}}(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}$

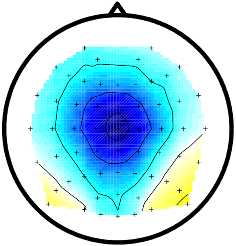
$\gamma = 0$

$w \sim \hat{\mathbf{\Sigma}}^{-1}(\mu_1 - \mu_2)$



$\gamma = 1$

$w \sim \mu_1 - \mu_2$

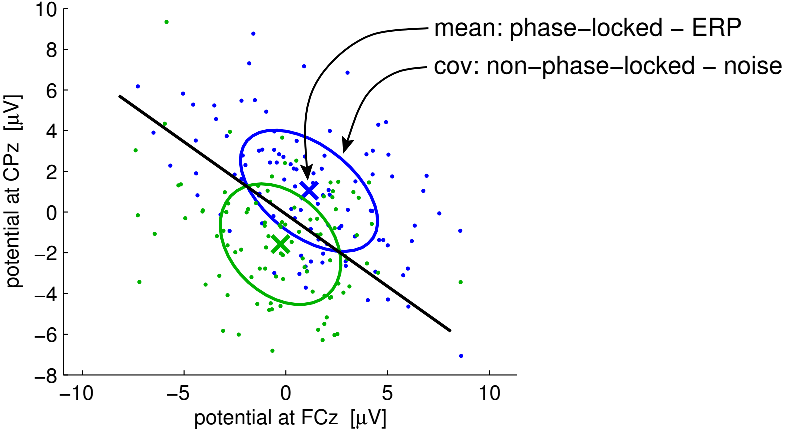


accounting for spatial structure of the noise

ERP and Noise

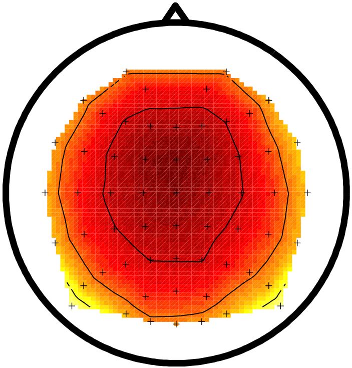
Simple assumption for ERPs: single trial $x_k(t)$ is composed of an ERP $s(t)$ and Gaussian 'noise' $\mathbf{n}_k(t)$:

$$\mathbf{x}_k(t) = \mathbf{s}(t) + \mathbf{n}_k(t) \qquad \text{for all trials } k = 1, \ldots, K$$

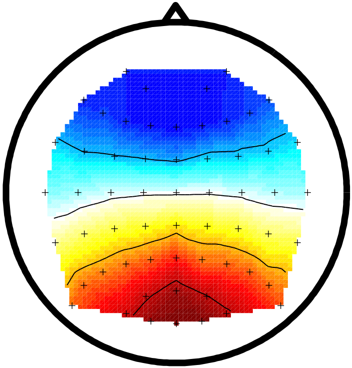


Spatial Structure of the Noise

The two strongest principal components of the noise (covariance matrix) in this data set:



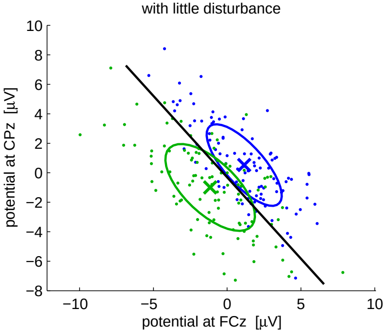
Trial-to-trial variation of P3



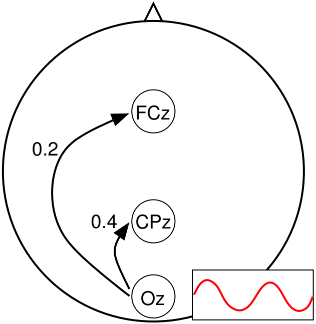
Visual alpha

Understanding Spatial Filters

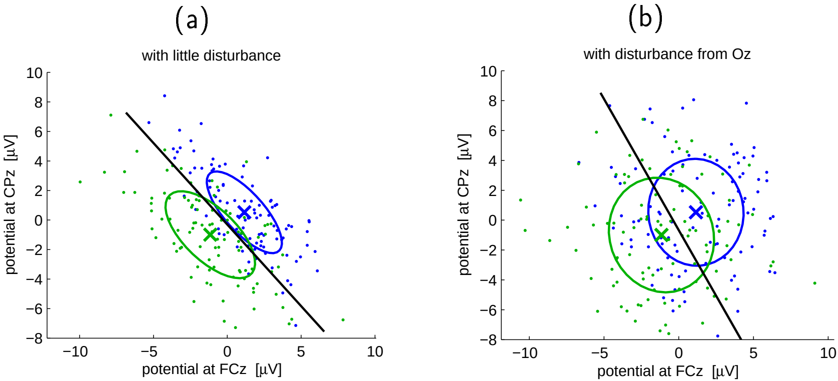
(a)



(b)



Understanding Spatial Filters

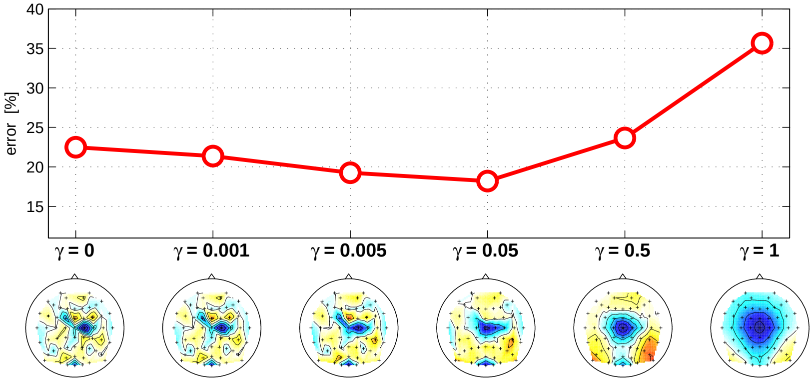


Two channel classification of (a): 15% error, (b): 37% error

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.

Impact of Shrinkage on the Spatial Filters

With increasing shrinkage, the spatial filters (classifier) look smoother, but classification may degrade with too much shrinkage.



Maps of spatial filters for different values of γ.

A Novel Analytical Method

Recently, a method to analytically calculate the optimal shrinkage parameter was published ([1]).

Thanks to *Nicole Krämer* for pointing the BBCI group to this method.

Optimal Selection of Shrinkage Parameter

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be n feature vectors and let $\hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$ be the empirical mean.

Aim: get a better estimate of the true covariance matrix $\mathbf{\Sigma}$ (especially in case $n < d$) than the sample covariance matrix $\hat{\mathbf{\Sigma}} = \frac{1}{n-1} \sum_{k=1}^{n} (\mathbf{x}_k - \hat{\mu})(\mathbf{x}_k - \hat{\mu})^\top$ by selecting a γ in

$$\tilde{\mathbf{\Sigma}}(\gamma) := (1 - \gamma)\hat{\mathbf{\Sigma}} + \gamma \nu \mathbf{I}.$$

Optimal Selection of Shrinkage Parameter

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ be n feature vectors and let $\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$ be the empirical mean.

Aim: get a better estimate of the true covariance matrix $\mathbf{\Sigma}$ (especially in case $n < d$) than the sample covariance matrix $\hat{\mathbf{\Sigma}} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\mu})(\mathbf{x}_k - \hat{\mu})^\top$ by selecting a γ in

$$\tilde{\mathbf{\Sigma}}(\gamma) := (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\nu\mathbf{I}.$$

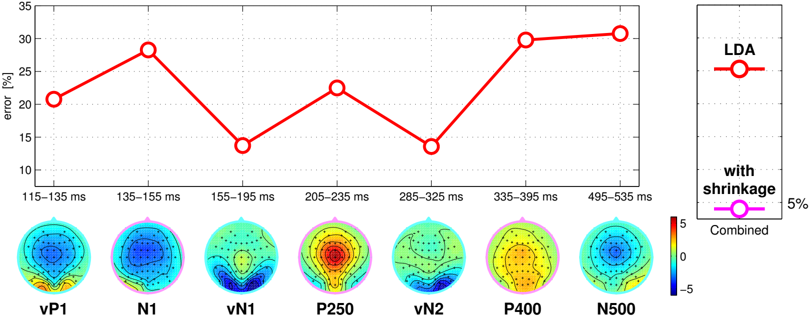
We denote by $(\mathbf{x}_k)_i$ resp. $(\hat{\mu})_i$ the i-th element of the vector $\mathbf{x}_k$ resp. $\hat{\mu}$. Furthermore we denote by s_{ij} the element in the i-th row and j-th column of $\hat{\mathbf{\Sigma}}$. We define

$$z_{ij}(k) = ((\mathbf{x}_k)_i - (\hat{\mu})_i)\,((\mathbf{x}_k)_j - (\hat{\mu})_j)$$

Then the optimal shrinkage parameter $\gamma^\star$ for which $\tilde{\mathbf{\Sigma}}(\gamma^\star) = \mathrm{argmin}_{\mathbf{S}} \|\mathbf{S} - \mathbf{\Sigma}\|_F^2$ can be analytically calculated ([2]) as

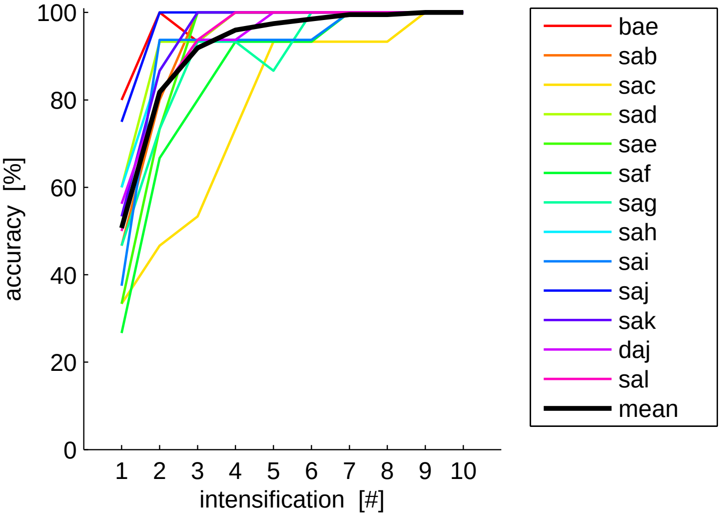
$$\gamma^\star = \frac{n}{(n-1)^2}\,\frac{\sum_{i,j=1}^{d} \mathrm{var}_k(z_{ij}(k))}{\sum_{i\neq j} s_{ij}^2 + \sum_i (s_{ii} - \nu)^2}$$

Result of Classification with Shrinkage



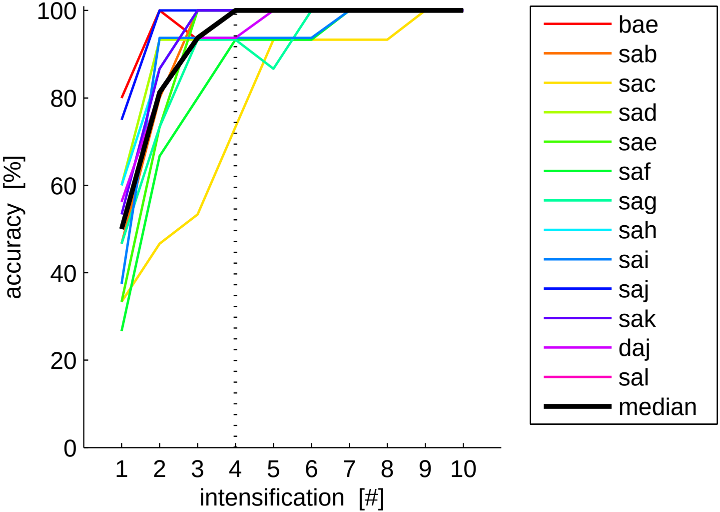
Using shrinkage the classification error could be drastically reduced to 4%.

Results for the Classic Matrix Speller



Accuracy in **letter selection**, chance level: 3.33%.

Results for the Classic Matrix Speller



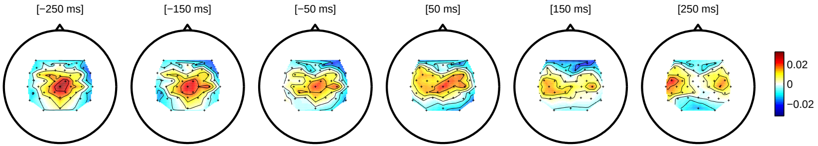
See Poster **W07** (Treder et al), and for other applications of this technique **A02** (Thurlings et al), **W10** (Höhne et al).

BBCI

Summary of Spatio-Temporal Classification

- Linear classification with shrinkage is a powerful method.
- Complete shrinkage ($\gamma = 1$) means neglecting the structure of the noise. In this case the classifier is the difference of the ERPs.
- The appropriateness of a linear separation depends on the way features are extracted and transformed.
- In contrast to non-linear classifiers, the weights of a linear classifier are informative.

The weights of the trained classifier can be visualized as a sequence of scalp topographies:



Topic of This Section

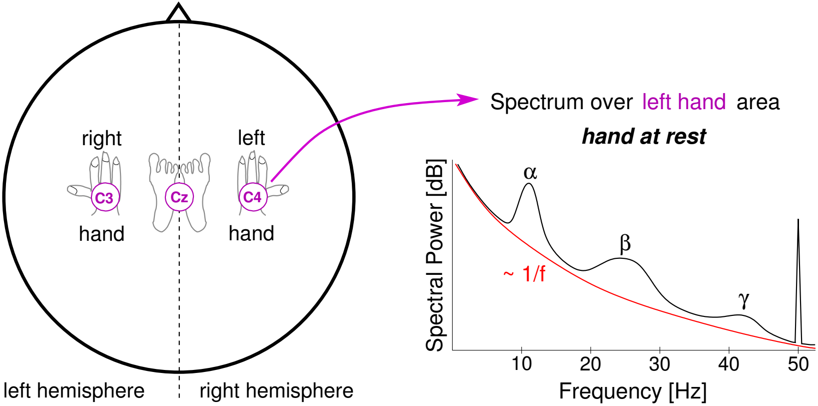
Method:

- Classification of **spectral** features, namely modulations of the amplitude in specific frequency bands.
- In particular, Common Spatial Pattern (CSP) analysis to classify different conditions that are characterized by a modulation of the amplitude of brain rhythms ([3, 4]).

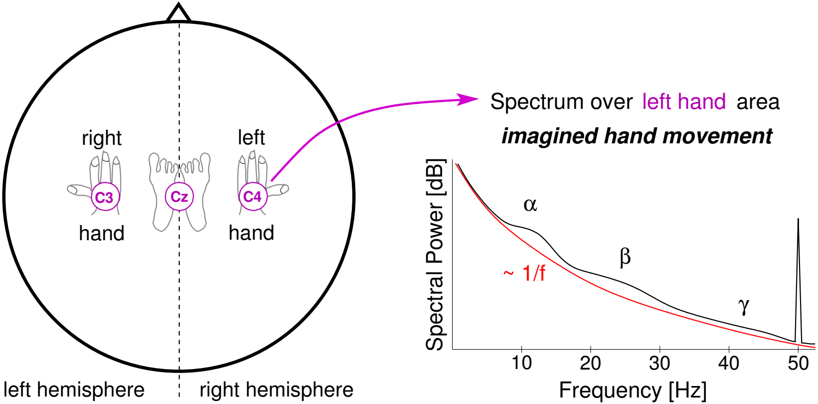
Application:

- Classification of motor imagery conditions in a BCI paradigm.

Neurophysiology: Sensorimotor Rhythms



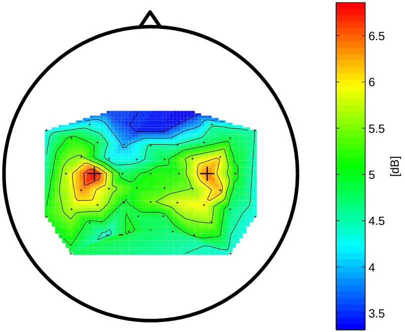
Neurophysiology: Sensorimotor Rhythms



Imagining a movement of a limb causes a local blocking of the corresponding sensorimotor rhythm (SMR), see [5, 6, 7].

Average Topography of Idle SMR

For each Laplace filtered channel in a relax recording, the strength of the local rhythm was estimated. The grand average over 80 participants is displayed as topographic mapping:

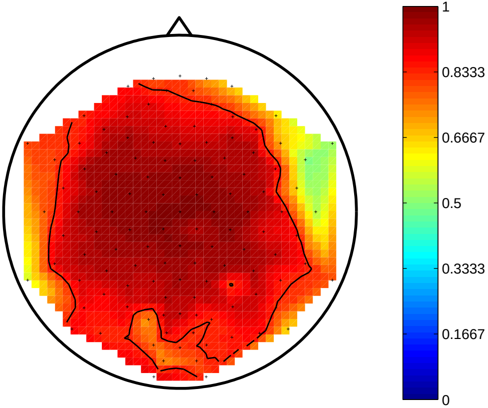


Conclusion

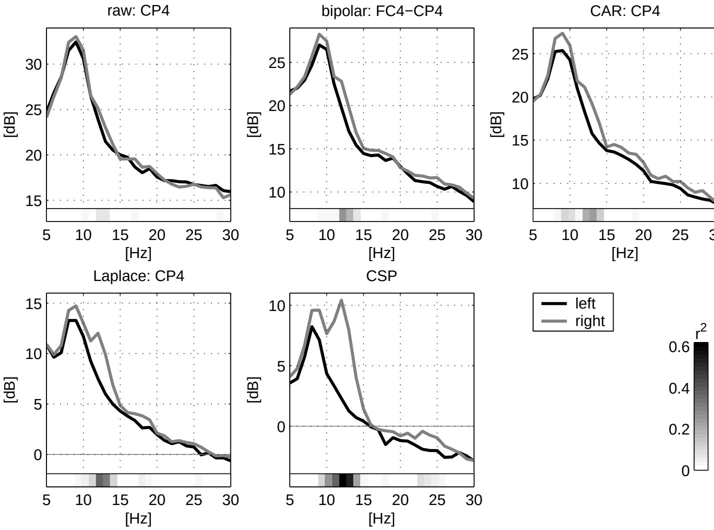
Locations C3 and C4 are good candidates to observe SMR modulations. These cover the sensorimotor areas of the right and the left hand.

Spatial Smearing

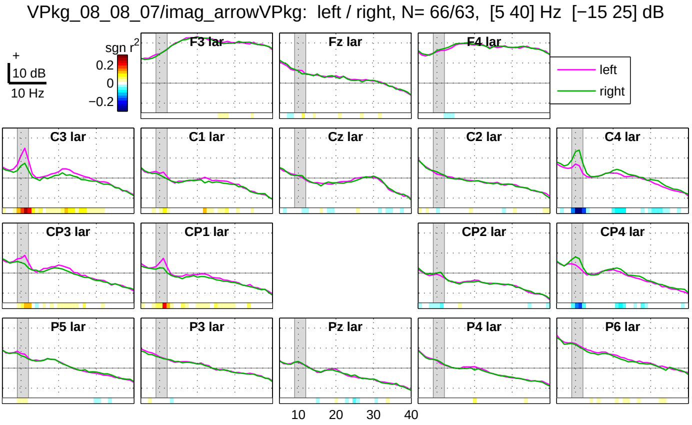
- Raw EEG scalp potentials are known to be associated with a large spatial scale owing to volumne conduction.
- In a simulation of Nunez et al [8] only half the contribution to one scalp electrode comes from sources within a 3 cm radius.



The Need for Spatial Filtering

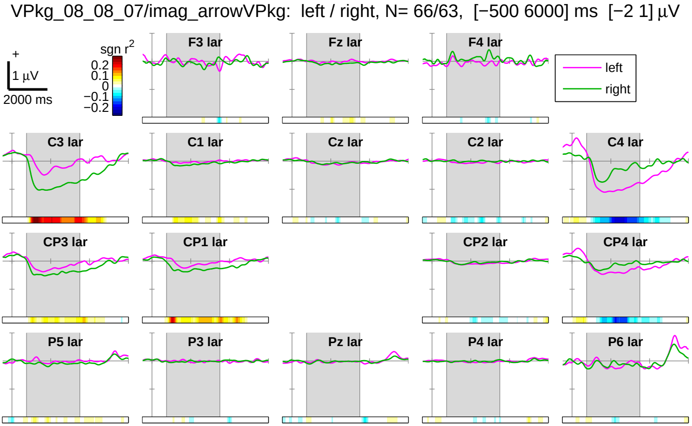


Analysis of Motor Imagery Conditions: Spectra



First step: determine a suitable frequency band that shows good discrimination between the conditions.

ERD Curves of Motor Imagery



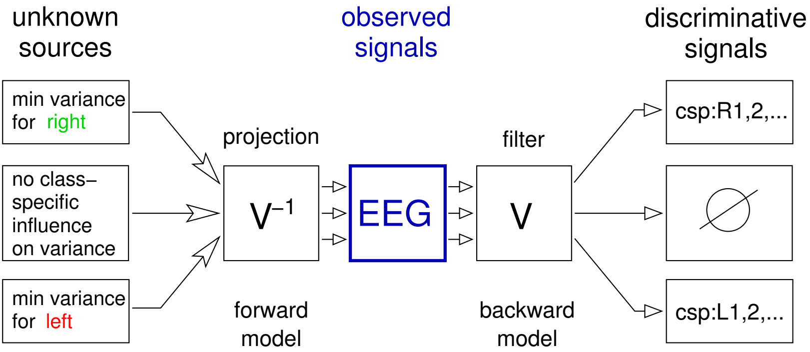
Second step: determine a suitable time interval during which discrimination is most prominent.

Remark: Simultaneous selection of frequency band and interval is more appropriate.

Common Spatial Pattern (CSP) Analysis

Goal: Find spatial filters that optimally capture modulations of brain rhythms

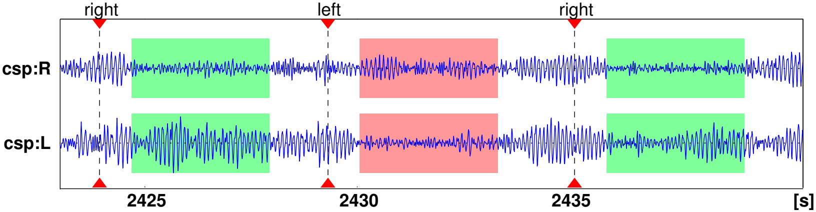
Observation: power of a brain rhythm $\sim$ variance of band-pass filtered signal.



BBCI

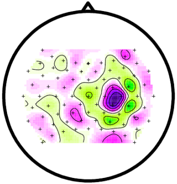
CSP Analysis

The goal of CSP:

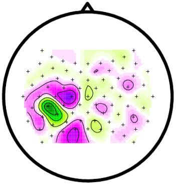


CSP analysis yields spatial filters that can be visualized:

CSP filter **'left'**

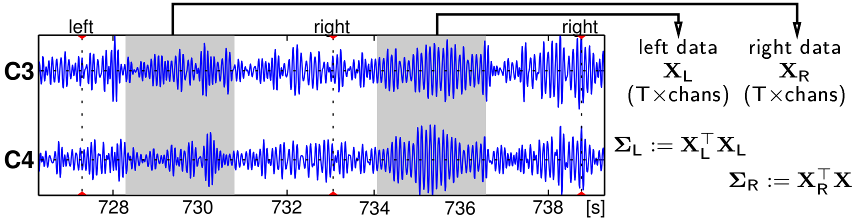


CSP filter **'right'**



CSP More Practical

EEG-signals during **motor imagery**, band-pass filtered (here 9–13 Hz):

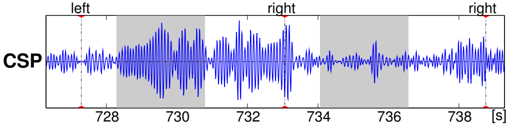


$$\mathbf{\Sigma}_\mathsf{L} := \mathbf{X}_\mathsf{L}^\top \mathbf{X}_\mathsf{L}$$

$$\mathbf{\Sigma}_\mathsf{R} := \mathbf{X}_\mathsf{R}^\top \mathbf{X}_\mathsf{R}$$

$$\mathbf{V}^\top \mathbf{\Sigma}_\mathsf{L} \mathbf{V} = D \quad \& \quad \mathbf{V}^\top (\mathbf{\Sigma}_\mathsf{L} + \mathbf{\Sigma}_\mathsf{R}) \mathbf{V} = \mathbf{I}$$

1) choose eigenvector $\mathbf{v}_i$ from $\mathbf{V}$ having a **large** eigenvalue d_i w.r.t. $\mathbf{\Sigma}_\mathsf{L}$.



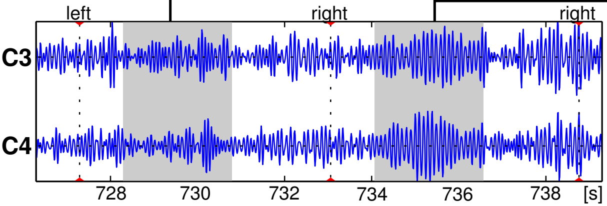
$$\mathrm{var}(\mathbf{X}_\mathsf{L} \mathbf{v}_i) = d_i \text{ large}$$

$$\mathrm{var}(\mathbf{X}_\mathsf{R} \mathbf{v}_i) = 1 - d_i \text{ small}$$

In Matlab: » `[V,D]= eig(Sigma1, Sigma1+Sigma2)`.

CSP More Practical

EEG-signals during **motor imagery**, band-pass filtered (here 9–13 Hz):



left data $\mathbf{X}_{\mathsf{L}}$ (T$\times$chans)

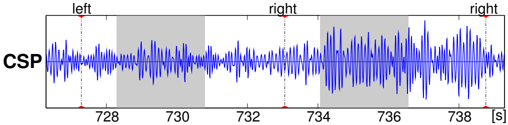
right data $\mathbf{X}_{\mathsf{R}}$ (T$\times$chans)

$\boldsymbol{\Sigma}_{\mathsf{L}} := \mathbf{X}_{\mathsf{L}}^{\top}\mathbf{X}_{\mathsf{L}}$

$\boldsymbol{\Sigma}_{\mathsf{R}} := \mathbf{X}_{\mathsf{R}}^{\top}\mathbf{X}_{\mathsf{R}}$

$$\mathbf{V}^{\top}\boldsymbol{\Sigma}_{\mathsf{L}}\mathbf{V} = D \quad \& \quad \mathbf{V}^{\top}(\boldsymbol{\Sigma}_{\mathsf{L}} + \boldsymbol{\Sigma}_{\mathsf{R}})\mathbf{V} = \mathbf{I}$$

2) choose eigenvector $\mathbf{v}_i$ from $\mathbf{V}$ having a **small** eigenvalue d_i w.r.t. $\boldsymbol{\Sigma}_{\mathsf{L}}$.



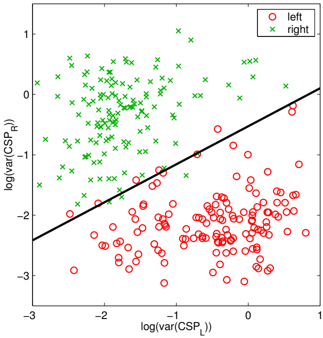
$\mathrm{var}(\mathbf{X}_{\mathsf{L}}\mathbf{v}_i) = d_i$ small

$\mathrm{var}(\mathbf{X}_{\mathsf{R}}\mathbf{v}_i) = 1 - d_i$ large

In Matlab: » `[V,D]= eig(Sigma1, Sigma1+Sigma2)`.

Training CSP-based Classification

To obtain features from the CSP filtered EEG, in each channel and trial, the variance across time is calculated and the logarithm is applied. This is a scatter plot of the resulting CSP features:



Here, only two dimensions are shown. Note, that applying the logarithm to the band power features makes the distribution more Gaussian and therefore enhances linear separability.

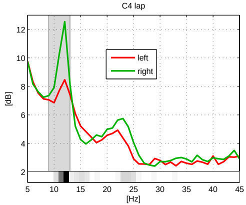
Training CSP-based Classification

- Determine most discriminative frequency band,
- band-pass filter EEG in that band,
- extract single trials using an appropriate time interval,
- calculate and select CSP filters,
- and apply them to EEG single trials,
- calculate the log variance within trials.

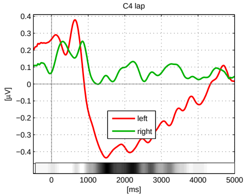
This results in a low dimensional feature vector for each trial (dimensionality equals number of selected CSP filters).

- Train a linear classifier like LDA on the features.
  (Since these features are low-dimensional, shrinkage is typically not necessary.)

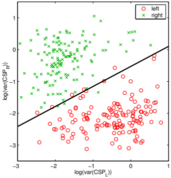
Summary: Training CSP-based Classification



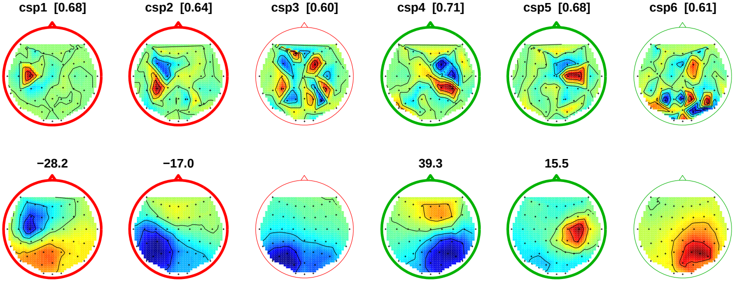
(1)



(2)



(4)



(3)

Applying CSP-based Classification

- Project EEG with spatial CSP filters and apply band-pass filter,
- calculate the variance in short windows (e.g. last 500 ms),
- take the logarithm,
- and apply the classifier weighting.

Remark: One nice feature of CSP is that the length of the classification window can be changed at runtime (i.e. during feedback).

For more theoretical considerations as well as practical hints see [3].

Section: Caveats in Validation

When machine learning techniques are used for classification of EEG single-trials, the expected performance of a method has to be evaluated carefully, and there are several possible pitfalls.

The estimation of generalization performance requires a training and a test set. The estimation is only proper

- if the test set was not used in any way to determine parameters of the method, and
- if the samples in the test set are independent from the samples in the training set.

Although these principles are quite obvious, it happens that they are violated.
Unfortunately, even some published journal articles lack a proper validation of the proposed methods.

Hall of Pitfalls in Single-Trial EEG Analysis

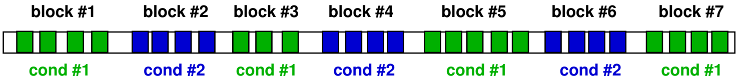
- preprocessing methods that use statistics of the whole data set like ICA, or normalization of features
(particularly severe for methods that use label information)
- features are selected on the whole data set, including trials that are later in the test set
- select parameters by cross validation on the whole data set and report the performance for the selected values
- artifacts/outliers are rejected from the whole data set
(resulting in a simplified test set)
- unsufficient validation for paradigms with block design

In this presentation we highlight the last issue.

Block Design

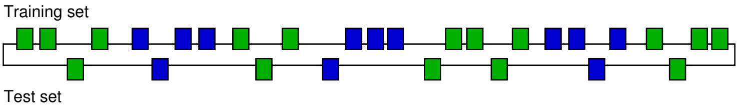
Assume the task is to discriminate between mental states in different conditions.

We say that an experiment has a block design, if the periods for which there is no alternation between conditions are longer than the intended change of states in online operation.

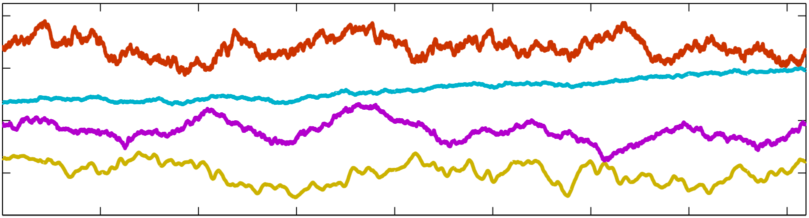


A problem arises, if the performance is estimated for such a data set by cross validation.

Slowly Changing Variables



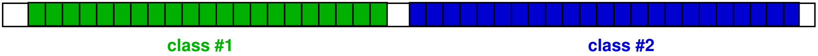
In EEG there are many slowly changing variables of background activity, therefore the single-trials are not independent. For an ordinary cross validation in a block design data set, the requirement of independence between training and test set is violated.



A Validation Test

To demonstrate impact of block design in cross validation, we perform cross validation in the following setting. Taking an arbitrary EEG data set, we assign **fake** labels (regardless of what happened during the recording) like this:

nBlocksPerClass=1:



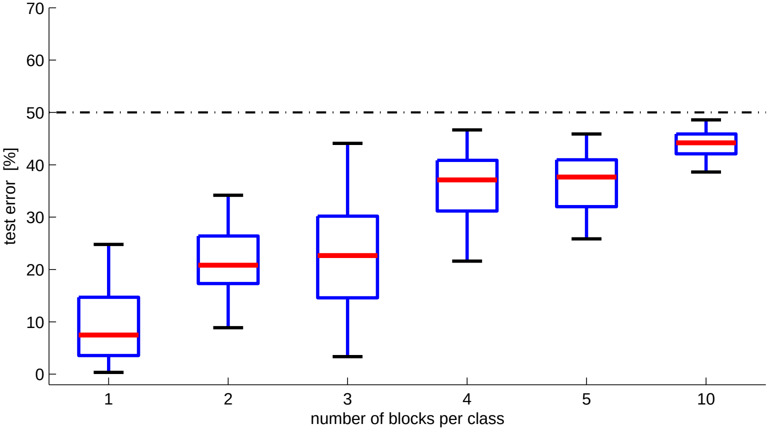
nBlocksPerClass=2:

nBlocksPerClass=3:

and so on.

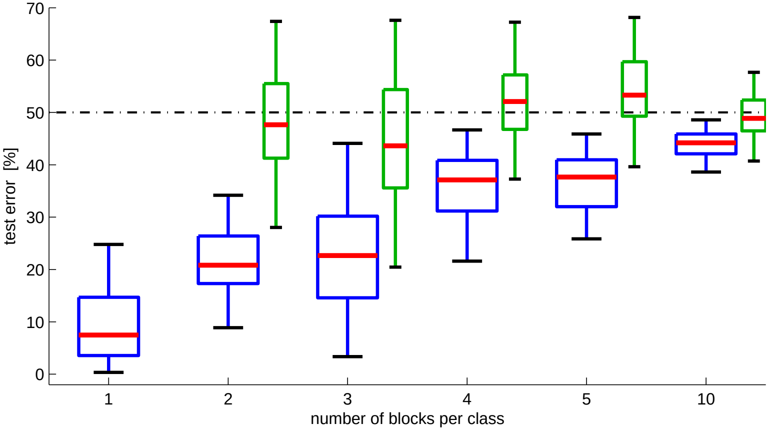
Results of the Validation Test

From each block single-trials are extracted of length 1s. This procedure was performed for 80 EEG data sets. Blue boxplots show the results of cross-validation:



Results of the Validation Test

From each block single-trials are extracted of length 1s. This procedure was performed for 80 EEG data sets. Blue boxplots show the results of cross-validation:



For comparison, results for **leave-one-block-out** validation are shown in green.

Further Comments and Summary

- The severeness of the underestimation of the true error depends on the complexity of the features and the classifier.
- Cross validation in block design data might also give the correct result – but alternative evaluation is required.
- The situation gets worse if trials are extracted from overlapping segments.
- The most realistic validation is to train the methods on the first $N-1$ runs and to evaluate on the last run.
- Leave-one-block-out and leave-one-run-out have larger standard errors than cross validation.

References

O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices", *J Multivar Anal*, 88(2): 365–411, 2004.

J. Schäfer and K. Strimmer, "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics", *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing Spatial Filters for Robust EEG Single-Trial Analysis", *IEEE Signal Proc Magazine*, 25(1): 41–56, 2008, URL `http://dx.doi.org/10.1109/MSP.2008.4408441`.

H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement", *IEEE Trans Rehab Eng*, 8(4): 441–446, 2000.

C. Neuper and W. Klimesch, eds., *Event-related Dynamics of Brain Oscillations*, Elsevier, 2006.

G. Pfurtscheller, C. Brunner, A. Schlögl, and F. L. da Silva, "Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks", *NeuroImage*, 31(1): 153–159, 2006.

G. Pfurtscheller and F. Lopes da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles", *Clin Neurophysiol*, 110(11): 1842–1857, 1999.

P. L. Nunez, R. Srinivasan, A. F. Westdorp, R. S. Wijesinghe, D. M. Tucker, R. B. Silberstein, and P. J. Cadusch, "EEG coherency I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales", *Electroencephalogr Clin Neurophysiol*, 103(5): 499–515, 1997.