JCSE Journal of Computing Science and Engineering

Invited Paper

Journal of Computing Science and Engineering, Vol. 7, No. 2, June 2013, pp. 112-121

Decoding Brain States during Auditory Perception by Supervising Unsupervised Learning

Anne K. Porbadnigk[†]
Machine Learning Laboratory, Berlin Institute of Technology, Berlin, and DFG Research Training Group 'Sensory Computation in Neural Systems', GRK 1589/1, Berlin, Germany
anne.k.porbadnigk@tu-berlin.de

Nico Görnitz[†]
Machine Learning Laboratory, Berlin Institute of Technology, Berlin, Germany
nico.goernitz@tu-berlin.de

Marius Kloft
Courant Institute of Mathematical Sciences, New York, and Memorial Sloan-Kettering Cancer Center, New York, NY, USA
mkloft@cs.nyu.edu

Klaus-Robert Müller*
Machine Learning Laboratory, Berlin Institute of Technology, Berlin, Germany
Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea
klaus-robert.mueller@tu-berlin.de

Abstract

The last years have seen a rise of interest in using electroencephalography-based brain computer interfacing methodology for investigating non-medical questions, beyond the purpose of communication and control. One of these novel applications is to examine how signal quality is being processed neurally, which is of particular interest for industry, besides providing neuroscientific insights. As for most behavioral experiments in the neurosciences, the assessment of a given stimulus by a subject is required. Based on an EEG study on speech quality of phonemes, we will first discuss the information contained in the neural correlate of this judgement. Typically, this is done by analyzing the data along behavioral responses/labels. However, participants in such complex experiments often guess at the threshold of perception. This leads to labels that are only partly correct, and oftentimes random, which is a problematic scenario for using supervised learning. Therefore, we propose a novel supervised-unsupervised learning scheme, which aims to differentiate true labels from random ones in a data-driven way. We show that this approach provides a more crisp view of the brain states that experimenters are looking for, besides discovering additional brain states to which the classical analysis is blind.

Category: Smart and intelligent computing

Keywords: Brain-computer interfacing; EEG; Semi-supervised learning; Systematic label noise; Anomaly detection

Open Access http://dx.doi.org/10.5626/JCSE.2013.7.2.112 http://jcse.kiise.org

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received Received 26 April 2013, **Accepted** Accepted 8 May 2013

*Corresponding Author

[†]These Authors contributed equally.

Copyright © 2013. The Korean Institute of Information Scientists and Engineers

pISSN: 1976-4677 eISSN: 2093-8020

I. INTRODUCTION

Brain-computer interfacing (BCI) aims at establishing a novel communication channel between man and machine [1-3]. To this end, brain signals need to be measured non-invasively (electroencephalography [EEG], magnetoencephalography [MEG], functional magnetic resonance imaging [fMRI] or near-infrared spectroscopy [NIRS]), or invasively (multi-unit activity [MUA], electrocorticography [ECoG]). Subsequently, the underlying cognitive states are decoded, and then the user-machine interaction loop is closed by providing real-time feedback about a particular cognitive brain state to the user. From the non-clinical perspective [4], the BCI has also become a very attractive research topic in recent years, since human cognitive states and intentions can be decoded directly at their very origin: the human brain, and not indirectly through behavioral correlates. Note, however, that we should think of the neural correlate as complementary to the behavioral one. A certain cognitive processing might not be visible in the behavioral signal, but may be clearly detected from the neural correlate, e.g., in non-conscious processing [5]. This is particularly interesting in mental state monitoring (e.g., mental workload [6]) or in visual [7, 8] and auditory perception tasks [5, 9, 10]. In this paper, we will report about the latter, as an example of a complex cognitive task, where the decoding of brain states is particularly challenging.

BCI technology has advanced significantly with the advent of robust machine learning techniques [3, 11-16] that by now have become a standard in the field. Brain data is characterized by non-stationarity and significant variability, both between trials and between subjects. Oftentimes, signals are high-dimensional, with only relatively few samples available for fitting models to the data, and finally, the signal-to-noise ratio (SNR) is highly unfavorable. In fact, even what is signal and what is noise are typically ill-defined, respectively (cf. [3, 11-16]). Due to this variability, machine learning methods have become the tool of choice for the analysis of single-trial brain data. In contrast, classical neurophysiological analysis methods apply averaging methods, such as taking grand averages over trials, subjects and sessions, to get rid of various sources of variability. This approach investigates the average brain, and can answer generic questions of neurophysiological interest, but it is rather blind to the wealth of the dynamics and behavioral variability available only to single-subject, single-trial analysis methods.

In the following, we will focus on behavioral and neural data from the speech signal quality judgements of subjects and their respective neural correlates, as measured by an EEG-BCI. Of particular concern to us is the question of whether or not a participant has behaviorally noticed the loss of quality in a transmitted signal, and whether and how this is reflected in the respective neural correlate. Answering this question is crucial for any provider of signal quality (audio or visual), in order to find the right balance between customer satisfaction and profitability. However, the behavioral ratings of stimuli given by participants are particularly spurious, since the loss of quality is oftentimes at the threshold of perception, so that the participants' assessments can be unreliable or even close to random guessing. In other words, the label data resulting from such studies lack ground truth. Thus, the label noise is not independent, but—at a perceptual level—it consists of a mix of random labels, and only a few informative ones, and there is no way to tell which is which. This systematic label noise makes the decoding of the respective cognitive brain state hard. So the challenge in our experiment is to decipher, despite a high level of dependent label noise, whether or not the participant has processed a loss of quality on a neural level.

Previous attempts to solve this question (using EEG data for assessing quality perception) employed fully *supervised* machine learning approaches only, using the behavioral responses of participants as labels [5, 7, 8, 17]. This has the obvious drawback that classification is mislead by those spurious labels, exacerbating the problem of finding the true 'neural labels'. Additionally, supervised approaches are biased towards finding the exact two classes specified by behavioral labels ('target is noticed', 'target is missed'), ignoring that there may be additional states of mind (e.g., 'participant not on task'). Finally, psychophysical experiments can lead to highly unbalanced classes (when targets are unlikely to be noticed), which also challenges supervised learning approaches in the presence of high noise.

In the following, we will first introduce the paradigm of the EEG experiment, and explain why the conventional approach is biased. In the third section, we present our supervised-unsupervised learning approach that aims at inferring the correct (neural) labels. The results are presented in the fourth section, followed by a discussion.

II. EEG EXPERIMENT & CLASSICAL ANALYSIS

Understanding which levels of quality loss are still perceived by users is a crucial question for any provider of signal quality. Conventionally, behavioral tests are used for this purpose, asking participants directly for their rating. Recent work has proposed to complement this approach by also recording a user's neural response to a stimulus, as the neural response may differ from the behavioral response [7, 8, 17].

A. Paradigm and Stimuli

Eleven participants (mean age 25 years) took part in this study, for whom both behavioral and neural response were recorded using 64-channel EEG. Participants performed an auditory discrimination task, in which they

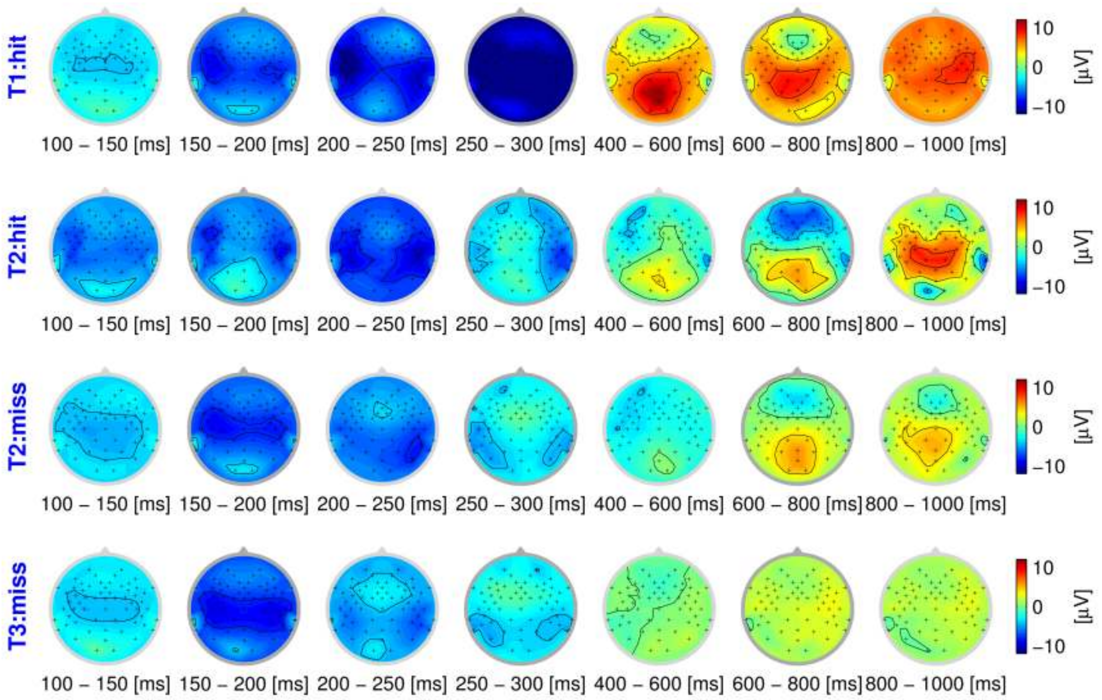


Fig. 1. Scalp distribution of event-related potentials for different stimuli in seven time intervals, grouped by their behavioral label (hit/miss; participant vp = 1). The maps represent a top view on the head, with nose pointing upwards.

had to press a button whenever they detected an auditory stimulus of degraded quality (target). Stimuli were presented in an oddball paradigm, using the undisturbed phoneme /a/ as non-target (NT, 70% of stimuli). Among these stimuli of high quality, the participant had to find instances when the phoneme was superimposed with signal-correlated noise. Participants were instructed to indicate by button press, if they noticed a deviation in the stimulus. Four noisy target stimuli were used, T1-T4, consisting of the phoneme /a/ superimposed with decreasing levels of signal-correlated noise (targets, 6% per class). In an additional 6% of trials, the phoneme /i/ was presented as control stimulus (C, target). The noise levels of the target stimuli (T1-T4) were chosen separately for each participant, in order to account for individual differences in sensitivity to noise, aiming at perception rates of 100%, 75%, 25%, and 0%, respectively. For this purpose, a pretest was run; the resulting SNRs for the deviant stimuli were set to 5, 21, 24, and 28 dB on average (mean perception rate in the experiment: 99%, 46%, 22%, and 7%). The disturbed auditory stimuli were created using a modulated noise reference unit (MNRU [18]). Target stimuli that were detected by the participant are referred to as 'hits' (true positives), and the others as 'misses' (false positives).

Each stimulus had a duration of 160 ms, with 1000 ms stimulus onset asynchrony. Per participant, 8 to 12 blocks were recorded, with 300 stimuli each. A parallel port computer keyboard was employed for recording the button presses of the participants. For stimulus presentation, in-ear headphones by Sennheiser, Germany were used. EEG was recorded using a Brain Products GmbH (Munich, Germany) EEG system, with 64 electrodes (AF3-4, 7-8; FAF1-2; Fz, 3-10; Fp1-2; FFC1-2, 5-8; FT7-10; FCz, 1-6; CFC5-8; Cz, 3-6; CCP7-8; CP1-2, 5-6; T7-8; TP7-10; P3-4, Pz, 7-8; POz; O1-2 and the right mastoid), and a BrainAmp (Brain Products GmbH) EEG amplifier. Electrodes were placed according to the international 10-10 system. The tip of the nose was chosen as a reference site, and a forehead ground electrode. EEG data were sampled at a rate of 100 Hz. In the following, we investigate event-related potentials (ERPs), i.e., the differential signal between the voltage at a given electrode position and the reference electrode.

B. Taking Wrong Labels at Face Value

The behavioral responses of the participants provide labels for each trial, seemingly indicating whether the stimulus was perceived as disturbed or not. However, these labels can be assumed to be confounded with label noise to a large degree, in particular at the threshold of perception (stimulus T2). As a first step, we take these spurious labels as ground truth, and analyze the ERPs in these groups. If the behavioral response indicates that the quality degradation is processed (hits), the resulting ERP activation pattern can be characterized by two components: early sensory and late cognitive processing stages.

Fig. 1 shows the spatial distribution of the ERPs as scalp distributions (head seen from above, nose pointing upwards), averaged over seven time intervals. The data of one participant (vp = 1) is examined here exemplarily. The top row shows the averaged neural response to a strong degradation that was noticed behaviorally (T1 hit). The four early intervals represent sensory processing of the stimulus (100–300 ms post-stimulus), which is reflected in a temporal negativity above the auditory cortices. In contrast, the last three intervals can be assumed to reflect cognitive processing (400–1000 ms post-stimulus). This elicits an occipital positivity, commonly referred to as P3 component. This component is elicited as a neural reaction to deviating stimuli in an oddball paradigm [19].

In our study, a P3 can be expected to occur when a participant notices that the quality of a stimulus is degraded. Generally speaking, the stronger the degradation, the higher the amplitude of the EEG signal, in particular that of the P3 component. This effect becomes obvious when comparing the first two rows of the figure, with a much weaker activation during late intervals for stimulus T2 (weak degradation), compared to T1 (strong degradation). In contrast, the last row shows the neural processing of a stimulus with a subtle degradation that is not noticed on a behavioral level (T3 miss). While sensory processing still causes activity in the early intervals, there is no notable cognitive component.

C. Conclusion

While the topography of the averaged ERPs seems to show a consistent picture so far, the presence of label noise becomes very obvious for the stimulus at the threshold of perception (T2). As the ratings of participants become unreliable to the point of guessing, grouping according to behavioral labels becomes conspicuously confounded, as can be seen in the second and third row of Fig. 1. Even though the participant gave different ratings in these cases, the neural activation is strikingly similar. While the presence of label noise is obvious for this stimulus, the labels of the other classes can be expected to be confounded as well, just to a lesser degree. In the following, we will infer the correct labels in a data-driven way by using a novel learning methodology, in order to obtain an unbiased view of the EEG data.

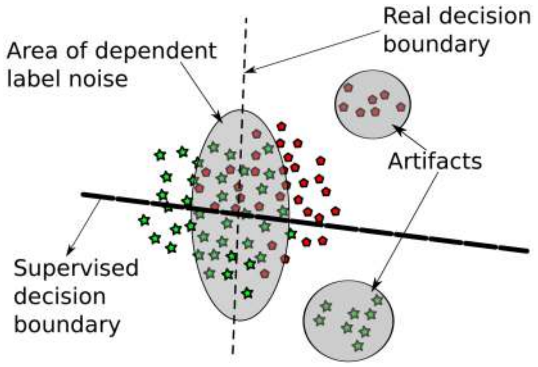


Fig. 2. The examined setting consists of various kinds of spurious data: artifacts caused by ill-behaving electrodes, as well as noisy dependent behavioral labels mixed in between trustworthy data. Hence, supervised methods are prone to find wrong decision boundaries.

III. INFERRING THE CORRECT LABELS

Our supervised-unsupervised learning approach aims at finding the true labels based solely on the EEG signal, without taking behavioral labels into account. Technically, we propose a two-step procedure to tackle the dependent label noise problem: (1) outlier detection to remove misleading trials, and (2) a subsequent distinction between labels that are random guesses and the ones that are informative. In the following, we will introduce this novel approach and exemplify it on data from the highly challenging EEG study on speech signal quality assessment, which we introduced in the previous section.

A. Abstract Learning Scenario

We consider a learning scenario where we have varying confidence in the labels (some are more trustworthy than others). In the considered setup, this stems from two sources: first, some data points are just artifacts, and second, some data points are labeled with very high error rate. The presence of falsely labeled data can hamper learning an accurate decision hyperplane, as illustrated in Fig. 2. Furthermore, some of the settings considered include only a single behavioral class label and hence make supervised learning impossible. As a remedy, we propose a two-step learning approach based on supervised-unsupervised learning, as follows:

1. *Artifact removal*: in the first step, as depicted in Fig. 3a, we remove measurement noise, which can in

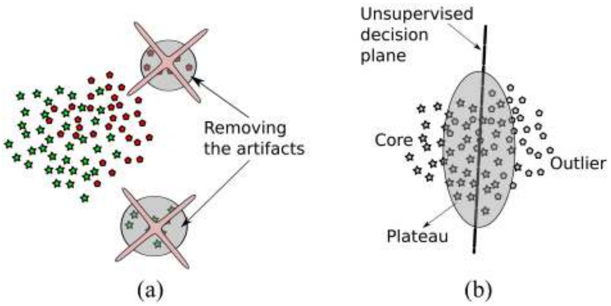


Fig. 3. In a first step, we get rid of the well-identifiable artifacts (a). In the second step, we train unsupervised, disregarding the spurious label information, to achieve a decision boundary based on data evidence (b).

general stem from broken sensors or, in our case, from faulty electrodes. The steps taken are described in Algorithm 1 lines 1–4.

2. *Handling label noise*: since we are explicitly not trusting the labels given, we train unsupervised on the remaining data points in the second step, resulting in classes that diverge maximal in the data (not necessarily in the labels). The setting is depicted in Fig. 3b, and described in detail in Algorithm 1 lines 6–8.

B. Sparsity-Inducing One-Class Learning

The first step of our approach is based on the paradigms of support vector learning [20-22] and density level set estimation [23]; that is, we are given n data points $x_1,...,x_n$, where x_i lies in some input space $\mathbb{R}^d$, and the goal is to find a model $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and a density level-set $D_\rho = \{x : f(x) \geq \rho\}$ encompassing the normal data, i.e., $x \in D_\rho$, while for outliers $x' \notin D_\rho$ holds. In this paper, we consider linear models of the form

$$f(x) = w^\top(x) \tag{1}$$

A popular density level set estimator is the so-called one-class SVM [24]

$$\begin{aligned} \min_{w,\rho,\xi} \quad & \Omega(w) + \frac{1}{\nu n}\|\xi\|_1 - \rho \\ \text{s.t.} \quad & w^\top x_i \geq \rho - \xi_i,\ \xi_i \geq 0,\ \forall i \in \{1,...,n\} \end{aligned} \tag{2}$$

where $\Omega(w)$ is a smooth regularizer, and $\nu \in]0, 1]$ is a hyperparameter controlling the ‘size’ of the level set (the lower ν, the larger the level set). Once the optimal parameters w^* and ρ^* are found, these are plugged into (1), and new instances x are classified according to sign($f(x) - \rho^*$). The learning machine (2) has been intensively studied for the choice of the regularizer $\Omega(w) := w^\top w$, which leads to *dense* optimal weight vectors w^*, i.e., the entries of w^* are strictly different from zero (except in pathological cases), and thus hinder feature *selection* (which is required for the application we have in mind). In contrast, we build the methodology used in this paper on more general regularizers of the form

$$\Omega(w) := \|w\|_p,$$

where $\|w\|_p = (\sum_{i=1}^{d}|w_i|^p)^{1/p}$ denotes the Minkowski ℓ_p-norm, focusing on the limiting case $p = 1$, which is likely to lead to sparse solutions: suppose we minimize an objective function $g(w)$ subject to $\|w\|_1 \leq 1$; then, the optimal solution is attained when the level sets of the objective function ‘hit’ the norm constraint. If the objective function is convex, the point of intersection is usually at one of the corners of the constraint, and thus has sparse coordinates. In linear methods, each dimension in the solution often corresponds to a measurable cause. The benefit of having a sparse solution vector lies in the fact that the solution now becomes interpretable. Since we have no ground truth, interpretability is mandatory. The resulting division into three classes, as depicted in Fig. 3b, seems from a non-sparse viewpoint somehow arbitrary. However, in a sparse setting, we will encounter three completely separated classes, where the plateau class is orthogonal to the core and outlier class. Hence, examples belonging to the plateau class will lie on the decision boundary. An elegant way to solve Equation (2) for $\Omega(w) = \|w\|_1$ is to set $w = w^+ - w^-$ substituting $\|w\|_1 = w^+ + w^-$, and to optimize over $w^+, w^- \geq 0$, instead of w. To enhance numerical stability of sparse one-class learning, we propose to consider the following sparsity-inducing one-class learning formulation:

$$\begin{aligned} \min_{w,\rho,\xi} \quad & \|w\|_1 + C\|\xi\|_1 \\ \text{s.t.} \quad & w^\top x_i \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i \in \{1,...,n\} \end{aligned} \tag{3}$$

(which is reminiscent of the very well known 2-class C-SVM, given by Cortes and Vapnik [25], or sparse Fisher, by Mika et al. [26]). The following theorem shows that Equation (2) is an exact re-formulation of Equation (3). Note that it is sufficient to consider the cases $C \geq \frac{1}{n}$ and $\nu \leq 1$, because otherwise, we trivially have $w^* = 0$.

Theorem 1. *Let $\Omega(w) = \|w\|_1$ and denote the optimal solution of (2) and (3) by $(w_\nu^*, \rho_\nu^*, \xi_\nu^*)$ and $(\tilde{w}_C^*, \tilde{\xi}_C^*)$, , respectively. Then, for any $\nu \in]0, 1]$, setting $C := \frac{1}{\nu n}$, it holds that*

$$w_\nu^* = \rho_\nu^* \tilde{w}_C,$$

i.e., the weight vectors output by (2) and (3) are, besides a scaling factor, equivalent.

Proof. Let (w^*, ρ^*, ξ^*) be optimal in (2). It follows that (w^*, ρ^*) is optimal in the corresponding unconstrained formulation:

$$(w^*, \rho^*) = \underset{w,\rho}{\operatorname{argmin}}\ \|w\|_1 + \frac{1}{\nu n}\sum_{i=1}^{n}\max(0, \rho - w^\top x_i) - \rho.$$

Note that thus $w^* = \operatorname{argmin}_w \|w\|_1 + \frac{1}{\nu n_*}\sum_{i=1}^{n} \max(0, \rho^* - w^\top x_i)$. Now denote $\tilde{w}^* := \operatorname{argmin}_{\tilde{w}} \rho^* \|\tilde{w}\|_1 + \frac{1}{\nu n}\sum_{i=1}^{n} \max(0, \rho^* - \rho^* \tilde{w}^\top x_i)$. By a variable substitution $w = \rho^* \tilde{w}$, we observe that $w^* = \rho^* \tilde{w}^*$ and hence w^*/ρ^* is optimal in $\min_{\tilde{w}} \|\tilde{w}\|_1 + \frac{1}{\nu n}\sum_{i=1}^{n} \max(0, 1 - \tilde{w}^\top x_i)$ (because ρ^* is positive), which, setting $C := \frac{1}{\nu n}$ is the unconstrained version of (3) (and thus equivalent). Thus w^*/ρ^* is optimal in (3), which proofs the assertion. □

C. Remark

For reasons that will become clear later, we wish to also include negatively labeled instances $\hat{x}_1, ..., \hat{x}_m$ (i.e., instances of which we already know that they are outliers) into the learning machine (3). A simple and effective way of doing so is to constrain the negatively labeled instances to lie outside of the density level set: $w^\top \hat{x}_i \leq 1 + \hat{\xi}_i$, $\hat{\xi}_i \geq 0$, $\forall i \in \{1,...,m\}$. This formulation is a *semi-supervised* extension of Equation (3), which constrains the negatively labeled examples $\hat{x}_1$ to lie outside of the density level set (besides a tolerance $\hat{\xi}_i$).

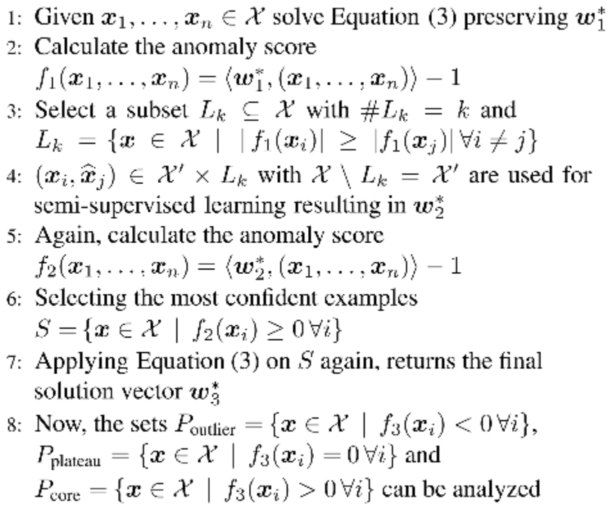
D. Processing Pipeline

In this section, we describe in detail the proposed supervised-unsupervised processing pipeline. The motivation behind this approach is that, even though it may seem that other methods (e.g., kernelized methods) could be more suitable for this problem, EEG data is well separable by linear classification (for a comparison of linear vs. nonlinear methods, cf. [27]). As discussed previously, the missing ground truth compels us to rely solely on interpretability of the results, which can be achieved easily by applying *linear* and *sparse* methods.

The inspection of the results of applying step 1 shows that there is a high chance of finding trials confounded by measurement noise (faulty electrodes) characterized by high amplitudes and/or drifts, which we denote as artifacts. Therefore, we deliberately force the method to exclude such examples and search for other features, by including the highest-ranked data points as outliers in a semi-supervised manner. Typically, we chose five examples of each end of the spectrum, to explicitly retain outlier labels (Algorithm 1 lines 3 and 4).

As illustrated in Fig. 3a, we divide into three classes: core, plateau and outlier class. These classes occur naturally, when applying the sparse one-class methods described in the previous section. Examples belonging to the plateau class are orthogonal to the core and outlier class; these data points lie on the decision boundary. Hence, for division, simple thresholding is sufficient.

Algorithm 1 Processing Pipeline



```
1: Given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$ solve Equation (3) preserving $\boldsymbol{w}_1^*$
2: Calculate the anomaly score
   $f_1(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \langle \boldsymbol{w}_1^*, (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \rangle - 1$
3: Select a subset $L_k \subseteq \mathcal{X}$ with $\#L_k = k$ and
   $L_k = \{\boldsymbol{x} \in \mathcal{X} \mid |f_1(\boldsymbol{x}_i)| \geq |f_1(\boldsymbol{x}_j)| \forall i \neq j\}$
4: $(\boldsymbol{x}_i, \widehat{\boldsymbol{x}}_j) \in \mathcal{X}' \times L_k$ with $\mathcal{X} \setminus L_k = \mathcal{X}'$ are used for
   semi-supervised learning resulting in $\boldsymbol{w}_2^*$
5: Again, calculate the anomaly score
   $f_2(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \langle \boldsymbol{w}_2^*, (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \rangle - 1$
6: Selecting the most confident examples
   $S = \{\boldsymbol{x} \in \mathcal{X} \mid f_2(\boldsymbol{x}_i) \geq 0 \forall i\}$
7: Applying Equation (3) on $S$ again, returns the final
   solution vector $\boldsymbol{w}_3^*$
8: Now, the sets $P_{\text{outlier}} = \{\boldsymbol{x} \in \mathcal{X} \mid f_3(\boldsymbol{x}_i) < 0 \forall i\}$,
   $P_{\text{plateau}} = \{\boldsymbol{x} \in \mathcal{X} \mid f_3(\boldsymbol{x}_i) = 0 \forall i\}$ and
   $P_{\text{core}} = \{\boldsymbol{x} \in \mathcal{X} \mid f_3(\boldsymbol{x}_i) > 0 \forall i\}$ can be analyzed
```

E. EEG Features

Based on the time series of the ERPs, we first reduced the dimensionality of the data (cf. [12]). Hence, we calculated the mean of the ERP signal within the seven neurophysiologically plausible intervals shown in Fig. 1 (for each electrode and trial). For this, the EEG signal from 61 recorded electrodes was used (omitting the Fp and EO electrodes). Thus, the dimensionality of the data was reduced from 6400 (100 data points × 61 electrodes) to 427 features (7 data points × 61 electrodes). These features were then used as input for the processing pipeline.

IV. RESULTS

The supervised-unsupervised learning approach groups the trials into three classes: a core class, an outlier class and a plateau class. These three classes can be seen exemplarily in Fig. 4 for one participant (vp = 1) and the stimulus at the threshold of perception (T2). Again, the scalp distribution of ERPs are shown in the seven intervals, which were also used as input features. Remarkably, the core class (row 2) finds a very typical representation of hits with distinct auditory processing (first intervals), and a strong P3 component (last two intervals), suggesting that the degradation was processed consciously. This pattern is subdued in the plateau class (row 3), where the auditory cortices still show a strong activation, but only a very subtle P3 is visible, indicating that the degradation was processed on a sensory level, but not noticed by the participant. Finally, there is virtually no activation in early or late components for the outlier class (row 4), suggesting, at most, subliminal processing of the stimulus. This distinction is more cogent by far than that based on behavioral labels, where two classes were assumed (hit/miss) that were obviously confounded (middle rows of Fig. 1). Not only does the algorithm find plausible classes, it also does so on the basis of neurophysiologically plausible features: as can be seen in the top row of the Fig. 4, the active features reflect the bi-temporal neural activity in early processing stages (auditory) and the occipital activity in late processing stages (cognitive). Across all participants and stimuli, the trials grouped into the core class show a distinct representation of how the stimulus is processed, including both sensory and cognitive components (‘neural hit’) or only sensory processing (‘neural miss’). For obvious degradations (C, T1), it is always the ‘neural hit’ that is found, while the algorithm rather assigns ‘neural misses’ to this class for subtle degradations. This is reasonable; as neural misses can be assumed to be predominant in those classes (the same is true for hits). In almost all cases (participants/stimuli), the outlier class represents trials that reflect a mental state other than these clear hit/miss patterns. Mostly, these are trials with very subdued activation (60% of trials show an amplitude lower

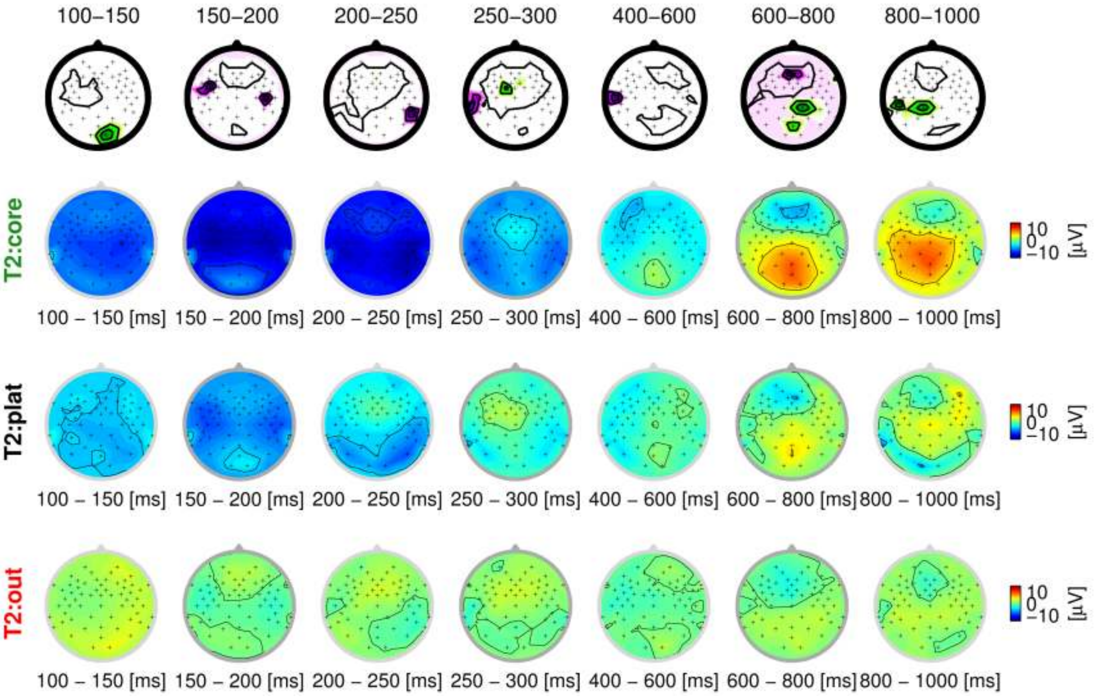
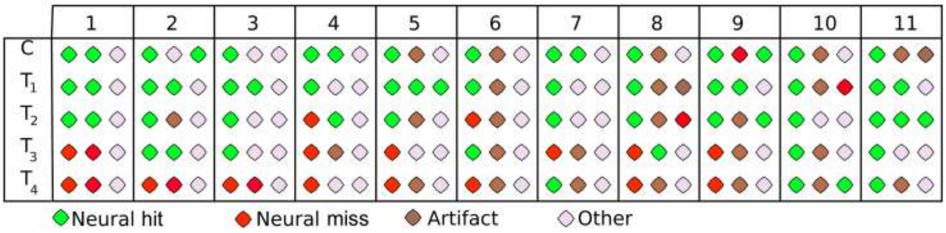


Fig. 4. Weights of features (filter) assigned in the last step of the Algorithm (top row). Scalp plots of the trials that are grouped into the core, plateau and outlier class (participant, vp = 1; T2) (bottom rows).



	1	2	3	4	5	6	7	8	9	10	11
C	🟢 🟢 ◇	🟢 ◇ 🟢	🟢 ◇ ◇	🟢 🟢 ◇	🟢 🟤 ◇	🟢 🟤 ◇	🟢 🟢 ◇	🟢 🟤 ◇	🟢 🔴 🟢	🟢 🟤 ◇	🟢 🟤 🟤
T_1	🟢 🟢 ◇	🟢 🟢 ◇	🟢 🟢 ◇	🟢 ◇ ◇	🟢 🟢 🟢	🟢 🟤 ◇	🟢 ◇ ◇	🟢 🟤 🟤	🟢 🟢 ◇	🟢 🟤 🔴	🟢 🟢 ◇
T_2	🟢 🟢 ◇	🟢 🟤 ◇	🟢 ◇ ◇	🔴 🟢 ◇	🟢 🟤 ◇	🔴 🟤 ◇	🟢 ◇ ◇	🟢 🟤 🔴	🟢 🟤 🟢	🟢 ◇ ◇	🟢 🟢 🟢
T_3	🔴 🔴 ◇	🟢 🟢 ◇	🟢 ◇ ◇	🔴 🟤 ◇	🔴 ◇ ◇	🟢 🟤 ◇	🔴 🟤 ◇	🔴 🟢 ◇	🔴 🟤 ◇	🟢 🟤 ◇	🟢 ◇ ◇
T_4	🔴 🔴 ◇	🔴 🔴 ◇	🔴 🔴 ◇	🔴 ◇ ◇	🔴 🟤 ◇	🔴 🟤 ◇	🟢 🟤 ◇	🔴 🟤 ◇	🔴 🟤 ◇	🟢 🟤 🟢	🟢 🟤 ◇

🟢Neural hit 🔴Neural miss 🟤Artifact ◇Other

Fig. 5. Overview over all participants (x-axis) and stimuli (y-axis): neural pattern of core, plateau and outlier classes (column 1–3), based on visual inspection.

than +/-5 µV on average), which indicates that the stimulus was processed at a subliminal level, at most. Finally, the plateau class, where the EEG signal is orthogonal to the features chosen by the algorithm, contains a cluster of trials that differ most widely among participants. These either reflect measurement noise or eye artifacts (40%), a subdued pattern of neural hits/misses (30%), or a mental state other than that (20%). Fig. 5 summarizes these results, based on visual inspection.

The motivation behind our approach is to find a coherent way to handle dependent label noise that is composed of a mixture of random labels and accurate ones. Fig. 6 provides an insight into these ratios, as far as our approach can reveal them. The behavioral perception rate is shown in black, i.e., the percentage of trials that were labeled as hits by the participants. As can be seen, the perception rate is high (almost 100%) for stimuli C/T1 and then drops markedly for stimuli T2–T4 (left to right). Underneath these values, the figure shows which percentage of these behavioral hits is assigned to the core, plateau or outlier class (ratios shown in gray, orange and white). This could be interpreted as the quantitative mixture of random labels and accurate ones.

V. DISCUSSION

Robustly analyzing EEG signals, despite their high non-stationarity (cf. [2, 28-30]), their multimodal nature, and the obviously noisy signal characteristics [2], is a major challenge that necessitates machine learning. However, in complex cognitive tasks in particular, the behavioral ratings given by participants are often unreliable, thus introducing label noise. Although in practice, independent

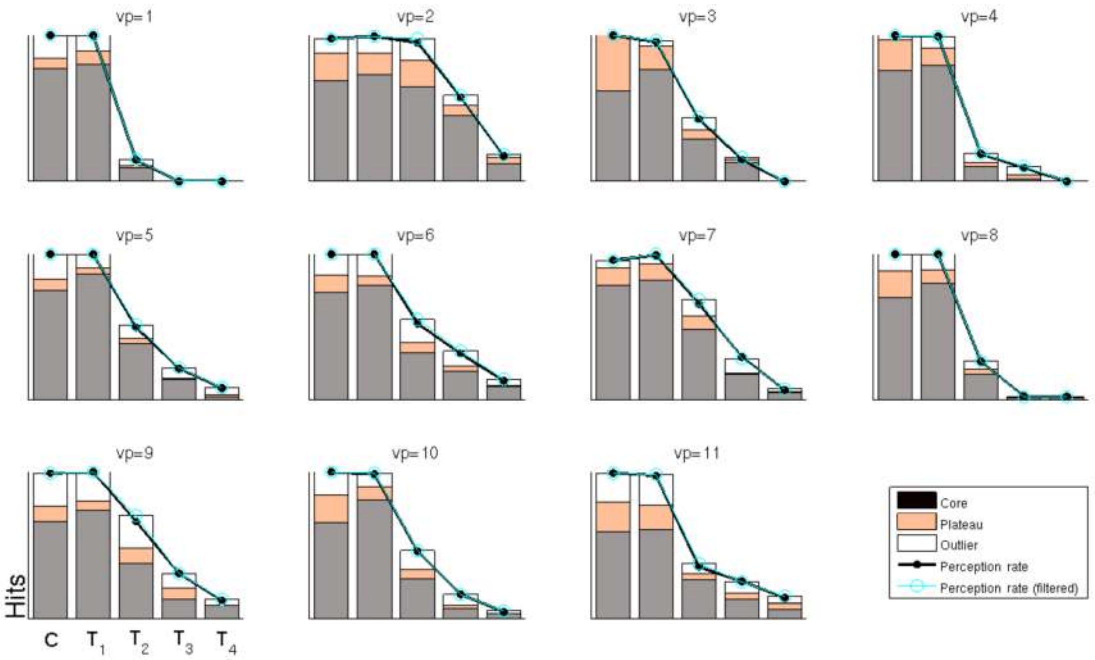


Fig. 6. Behavioral perception rate for all participants and target stimuli (C, T1–T4 from left to right), with the ratio of how many of these trials are grouped into core, plateau and outlier class (gray, orange, and white box).

label noise can be handled by most vanilla supervised learning algorithms, they can fail miserably in the case of *dependent* label noise. This set-up is rather common in behavioral experiments where a subject is required to assess a given stimulus; in this work we have analyzed data from speech signal quality judgements. Near perception threshold, the behavioral responses of subjects provide labels that are noisy, through a subjective assessment of the auditory signal. There are two reasons for this: (1) the subjects guess, i.e., the labels are random, and (2) a very weakly correlated perception of a change in audio signal quality is reported that gives rise to a faint structure in the noisy labels. Computing the neural correlates of behavior requires labels that reflect the task as cleanly as possible. To achieve this, we propose a novel supervised-unsupervised learning procedure, that first removes artifactual trials from the experiment before inferring which of the remaining labels are reliable and which are random (Fig. 2). Once these more reliable labels are in place, a better and more meaningful experimental evaluation of the neural correlates in our speech signal quality application can be performed. Moreover, our approach allows for defining groupings of trials that reflect more finely-grained cognitive states. Furthermore, it is an interesting point to note that in this manner, a neural correlate may occasionally be even more sensitive than the conscious behavioral one.

Future work will apply our method in the calibration phase of a BCI experiment, where subjects are asked to assume predefined brain states. Here, it is well-known that subjects occasionally do not comply with the instruction given [31] or do not maintain a certain cognitive state throughout the prescribed duration of a stimulus. Our algorithm could thus again contribute to a cleaning of the labels and thus to an increased robustness of the trained BCI system.

ACKNOWLEDGMENTS

We gratefully thank Benjamin Blankertz and Gabriel Curio for valuable comments and helpful suggestions. The data was recorded in a previous study [17], and we thank our colleagues for permission to use this data. This work was supported by the German Bundesministerium fur Bildung und Forschung (BMBF FKZ 01GQ0850 and 01IB001A), the German Science Foundation (DFG MU 987/6-1, RA 1894/1-1), and the World Class University Program of the Korea Research Foundation (R31-10008). We would like to acknowledge support for this project from the BMBF project ALICE, Autonomous Learning in Complex Environments (01IB10003B). Marius Kloft acknowledges a postdoctoral fellowship from the German Research Foundation (DFG).

REFERENCES

1. J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, “Brain-computer interfaces for communication and control,” *Clinical Neurophysiology*, vol. 113, no.

6, pp. 767-791, 2002.

2. A. Kübler and K. R. Müller, "An introduction to brain computer interfacing," in *Toward Brain-Computer Interfacing*, G. Dornhege, J. del R. Millan, T. Hinterberger, D. McFarland, and K. R. Müller, Editors, Cambridge, MA: MIT Press, 2007, pp. 1-25.
3. G. Dornhege, M. Krauledat, K. R. Müller, and B. Blankertz, "General signal processing and machine learning tools for BCI," in *Toward Brain-Computer Interfacing*, G. Dornhege, J. del R. Millan, T. Hinterberger, D. McFarland, and K. R. Müller, Editors, Cambridge, MA: MIT Press, 2007, pp. 207-234.
4. B. Blankertz, M. Tangermann, C. Vidaurre, S. Fazli, C. Sannelli, S. Haufe, C. Maeder, L. E. Ramsey, I. Sturm, G. Curio, and K. R. Müller, "The berlin brain-computer interface: non-medical uses of BCI technology," *Front Neuroscience*, vol. 4, pp. 198, 2010.
5. A. K. Porbadnigk, M. S. Treder, B. Blankertz, J. N. Antons, R. Schleicher, S. Möller, G. Curio, and K. R. Müller, "Single-trial analysis of the neural correlates of speech quality perception," *Journal of Neural Engineering*, 2013, submitted to publication.
6. K. R. Müller, M. Tangermann, G. Dornhege, M. Krauledat, G. Curio, and B. Blankertz, "Machine learning for real-time single-trial EEG-analysis: from brain-computer interfacing to mental state monitoring," *Journal of Neuroscience Methods*, vol. 167, no. 1, pp. 82-90, 2008.
7. A. K. Porbadnigk, S. Scholler, B. Blankertz, A. Ritz, M. Born, R. Scholl, K. R. Müller, G. Curio, and M. S. Treder, "Revealing the neural response to imperceptible peripheral flicker with machine learning," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Boston, MA, 2011, pp. 3692-3695.
8. S. Scholler, S. Bosse, M. S. Treder, B. Blankertz, G. Curio, K. R. Müller, and T. Wiegand, "Towards a direct measure of video quality perception using EEG," *IEEE Transactions on Image Processing*, vol. 21, no. 5, pp. 2619-2629, 2012.
9. A. K. Porbadnigk, J. N. Antons, M. S. Treder, B. Blankertz, R. Schleicher, S. Möller, and G. Curio, "ERP assessment of word processing under broadcast bit rate limitations," *Neuroscience Letters,* vol. 500, suppl. 1, p. e49, 2011.
10. J. N. Antons, R. Schleicher, S. Arndt, S. Möller, A. K. Porbadnigk, and G. Curio, "Analyzing speech quality perception using electro-encephalography," *IEEE Journal of Selected Topics in Signal Processing,* vol. 6, no. 6, pp. 721-731, 2012.
11. B. Blankertz, F. Losch, M. Krauledat, G. Dornhege, G. Curio, and K. R. Müller, "The berlin brain-computer interface: accurate performance from first-session in BCI-naive subjects," *IEEE Transactions on Bio-medical Engineering*, vol. 55, no. 10, pp. 2452-2462, 2008.
12. B. Blankertz, S. Lemm, M. S. Treder, S. Haufe, and K. R. Müller, "Single-trial analysis and classification of ERP components: a tutorial," *Neuroimage*, vol. 56, no. 2, pp. 814-825, 2011.
13. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process Magazine*, vol. 25, no. 1, pp. 41-56, 2008.
14. B. Blankertz, G. Dornhege, M. Krauledat, K. R. Müller, and G. Curio, "The non-invasive berlin brain-computer interface: fast acquisition of effective performance in untrained subjects," *Neuroimage*, vol. 37, no. 2, pp. 539-550, 2007.
15. B. Blankertz, G. Curio, and K. R. Müller, "Classifying single trial EEG: towards brain computer interfacing," In *Advances in Neural Information Processing Systems*, T. G. Diettrich, S. Becker, and Z. Ghahramani, Editors, Cambridge, MA: MIT Press, 2002, pp. 157-164.
16. S. Lemm, B. Blankertz, T. Dickhaus, and K. R. Müller, "Introduction to machine learning for brain imaging," *Neuroimage*, vol. 56, no. 2, pp. 387-399, 2011.
17. A. K. Porbadnigk, J. N. Antons, B. Blankertz, M. S. Treder, R. Schleicher, S. Möller, and G. Curio, "Using ERPs for assessing the (sub)conscious perception of noise," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Buenos Aires, Argentina, 2010, pp. 2690-2693.
18. International Telecommunication Union, "ITUR Recommendation P.810: modulated noise reference unit (MNRU)," 1996.
19. S. Sutton, M. Braren, J. Zubin, and E. John, "Evoked-potential correlates of stimulus uncertainty," *Science*, vol. 150, no. 3700, pp. 1187-1188, 1965.
20. V. Vapnik, The Nature of Statistical Learning Theory, New York, NY: Springer, 1995.
21. K. R. Müller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181-201, 2001.
22. G. Montavon, M. Braun, T. Krueger, and K. R. Müller, "Analyzing local structure in kernel-based learning: explanation, complexity and reliability assessment," *IEEE Signal Processing Magazine*, 2013, submitted to publication.
23. W. Polonik, "Measuring mass concentration and estimating density contour clusters: an excess mass approach," *Annals of Statistics*, vol. 23, no. 3, pp. 855-881, 1995.
24. B. Scholkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443-1471, 2001.
25. C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.
26. S. Mika, G. Ratsch, and K. R. Müller, "A mathematical programming approach to the kernel fisher algorithm," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Editors, Cambridge, MA: MIT Press, 2001, pp. 591-597.
27. K. R. Müller, C. W. Anderson, and G. E. Birch, "Linear and non-linear methods for brain-computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 165-169, 2003.
28. P. von Bunau, F. C. Meinecke, F. Kiraly, and K. R. Müller, "Finding stationary subspaces in multivariate time series," *Physical Review Letters*, vol. 103, no. 21, pp. 214101, 2009.
29. P. Shenoy, M. Krauledat, B. Blankertz, R. P. N. Rao, and K. R. Müller, "Towards adaptive classification for BCI," *Journal of Neural Engineering*, vol. 3, no. 1, pp. R13-R23, 2006.
30. C. Vidaurre, C. Sannelli, K. R. Müller, and B. Blankertz,

"Machine-learning based co-adaptive calibration," *Neural Computation*, vol. 23, no. 3, pp. 791-816, 2011.

31. C. Sannelli, M. Braun, and K. R. Müller, "Improving BCI performance by task-related trial pruning," *Neural Networks*, vol. 22, no. 9, pp. 1295-1304, 2009.

Anne Porbadnigk

Anne Porbadnigk received her M.S. in Computer Science and Engineering from the University of Michigan (Ann Arbor MI, USA), in 2006, where she studied on a Fulbright scholarship. In 2009, she completed her German Diplom (M.S. equivalent) in Computer Science at the Karlsruhe Institute of Technology (Germany). She spent the academic year 2008/09 as a visiting researcher at Carnegie Mellon University (Pittsburgh PA, USA), advised by John R. Anderson. In 2009, she joined the Machine Learning Laboratory at the Berlin Institute of Technology (Germany), working with Klaus-Robert Müller, Benjamin Blankertz and Gabriel Curio. She is also affiliated with the Research Training Group 'Sensory Computation in Neural Systems'. The focus of her research is on machine learning approaches to computational neuroscience, in particular Brain Computer Interfaces.

Nico Görnitz

Nico Görnitz received a diploma (MSc equivalent) in computer engineering from the Technische Universität Berlin. Currently, he is enrolled as a PhD student in the machine learning program at the Technische Universität Berlin headed by Klaus-Robert Müller. From 2010-2011, he was also affiliated with the Friedrich Miescher Laboratory of the Max Planck Society in Tübingen, Germany where he was co-advised by Gunnar Rätsch. His primary research interests cover machine learning and its applications in computational biology, Brain Computer Interfaces and computer security. This includes especially structured output prediction, density level-set estimation, semi-supervised learning and anomaly detection as well as corresponding optimization methods.

Marius Kloft

Marius Kloft received a diploma in mathematics from Philipps-Universität Marburg with a thesis in algebraic geometry and a PhD in computer science from Technische Universität Berlin, where he was co-advised by Klaus-Robert Müller and Gilles Blanchard. During his PhD, he spent a year at UC Berkeley, where he was advised by Peter Bartlett. After research visits to the Friedrich-Miescher-Laboratory of the Max Planck Society, Tübingen, and Korea University, Seoul, he is now a postdoctoral fellow, jointly appointed at Courant Institute of Mathematical Sciences and Memorial Sloan-Kettering Cancer Center, New York, working with Mehryar Mohri and Gunnar Rätsch, respectively. His current research focus lies in the field of machine learning with non-i.i.d. data and applications to genomic cancer data.

Klaus-Robert Müller

Klaus-Robert Müller has been Professor for Computer Science at TU Berlin since 2006; at the same time he is directing the Bernstein Focus on Neurotechnology Berlin. He studied physics in Karlsruhe from 1984-89 and obtained his PhD in Computer Science at TU Karlsruhe in 1992. After a PostDoc at GMD-FIRST, Berlin, he was a Research Fellow at University of Tokyo from 1994-1995. From 1995 he built up the Intelligent Data Analysis (IDA) group at GMD-FIRST (later Fraunhofer FIRST) and directed it until 2008. 1999-2006 he was a Professor at University of Potsdam. In 1999, he was awarded the Olympus Prize by the German Pattern Recognition Society, DAGM and in 2006 he received the SEL Alcatel Communication Award. In 2012 he was elected to be a member of the German National Academy of Sciences – Leopoldina. His research interests are intelligent data analysis, machine learning, signal processing and Brain Computer Interfaces.