Introduction to Brain-Computer Interface Design: Theory

Christian A. Kothe
SCCN, UCSD

Slides Available At

- ftp://sccn.ucsd.edu/pub/bcilab/embc2012/slides.zip
- Also on the DVDs

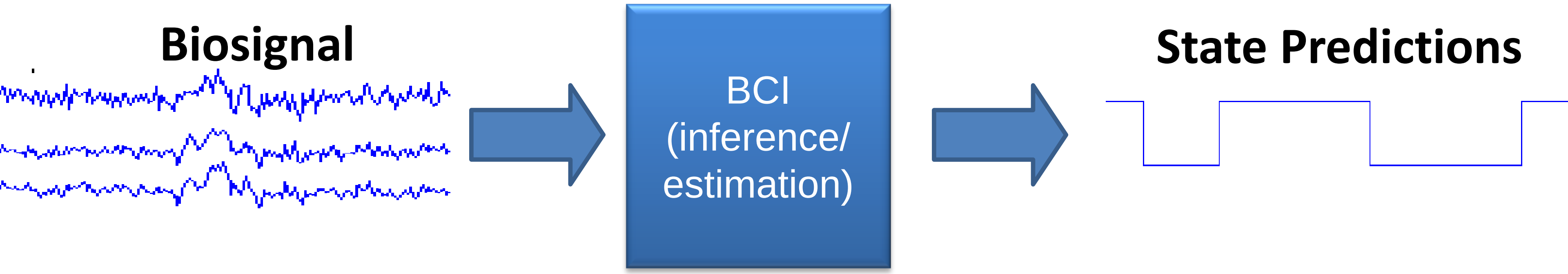
Outline

1. High-level View
2. Application Areas and Examples
3. Basic Theoretical Principles and Framework
4. Analyzing ERP-like Processes
5. Analyzing Oscillatory Processes
6. Evaluating Performance and Results
7. Further Reading

1 High-Level View

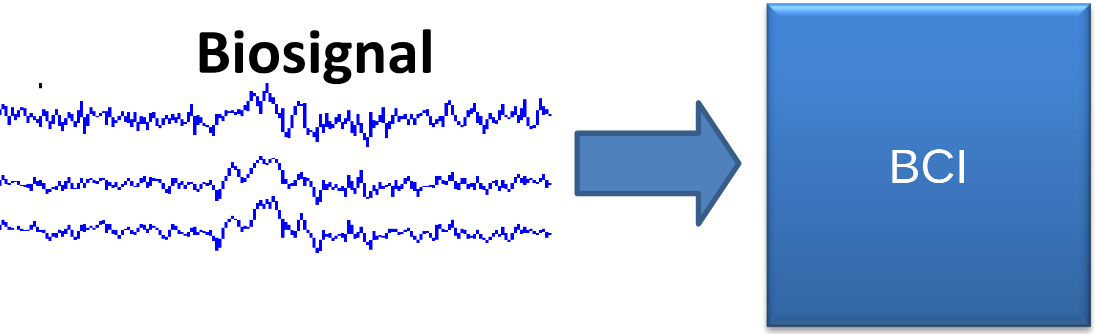
BCI: Our Working Definition

- "A system which takes a biosignal measured from a person and predicts (in real time / on a single-trial basis) some abstract aspect of the person's cognitive state."



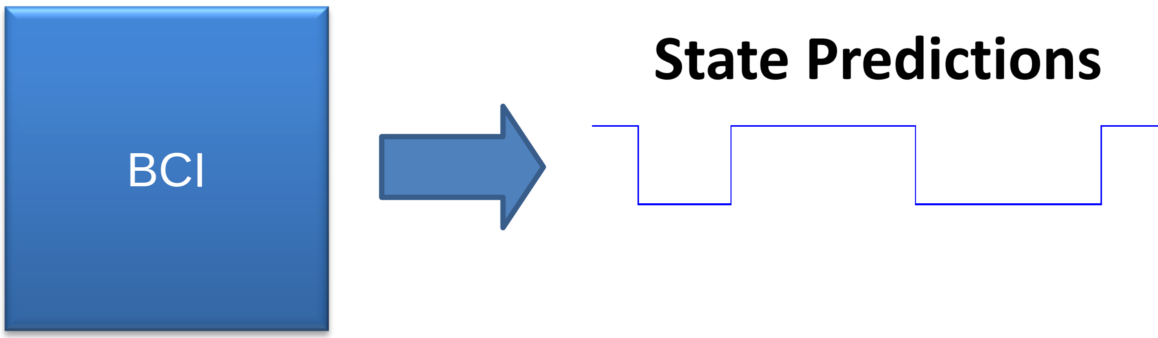
Biosignals and other Inputs

- **Brain Signals:** EEG, fNIRS, MEG, fMRI, ECoG, ...
- **Peripheral Measures:** ECG, EMG, EOG, GSR, Respiration, Gaze/Pupillometry, Motion Capture
- **Context Information:** Program/System State, Vehicle Speed, ...



BCI Estimates/Predictions

- Any aspect of the physical brain state that can be recovered from observable signals
- **Tonic state:** degree of "relaxation", cognitive load,…
- **Phasic state:** switching attention, type of imagined movement, …
- **Event-related state:** surprised/not surprised, committed error, event noticed/not noticed, …



2 Application Areas and Examples

Communication and Control for the Severely Disabled

- Severe Disabilities: Tetraplegia, Locked-in syndrome
- **Speller Programs, Wheelchairs, Robots, …**

P300 Speller

KU Leuven

Brain2Robot (Fraunhofer FIRST)

Other Health Uses

- **Sleep Stage Recognition, Neurorehabilitation**

iBrain

Takata et al., 2011

Operator Monitoring

- **Braking Intent, Lane-Change Intent, Workload, Fatigue, Alertness, Attention, …**

Haufe et al., 2011

The MITRE Corp., 2011

Entertainment, Social, etc.

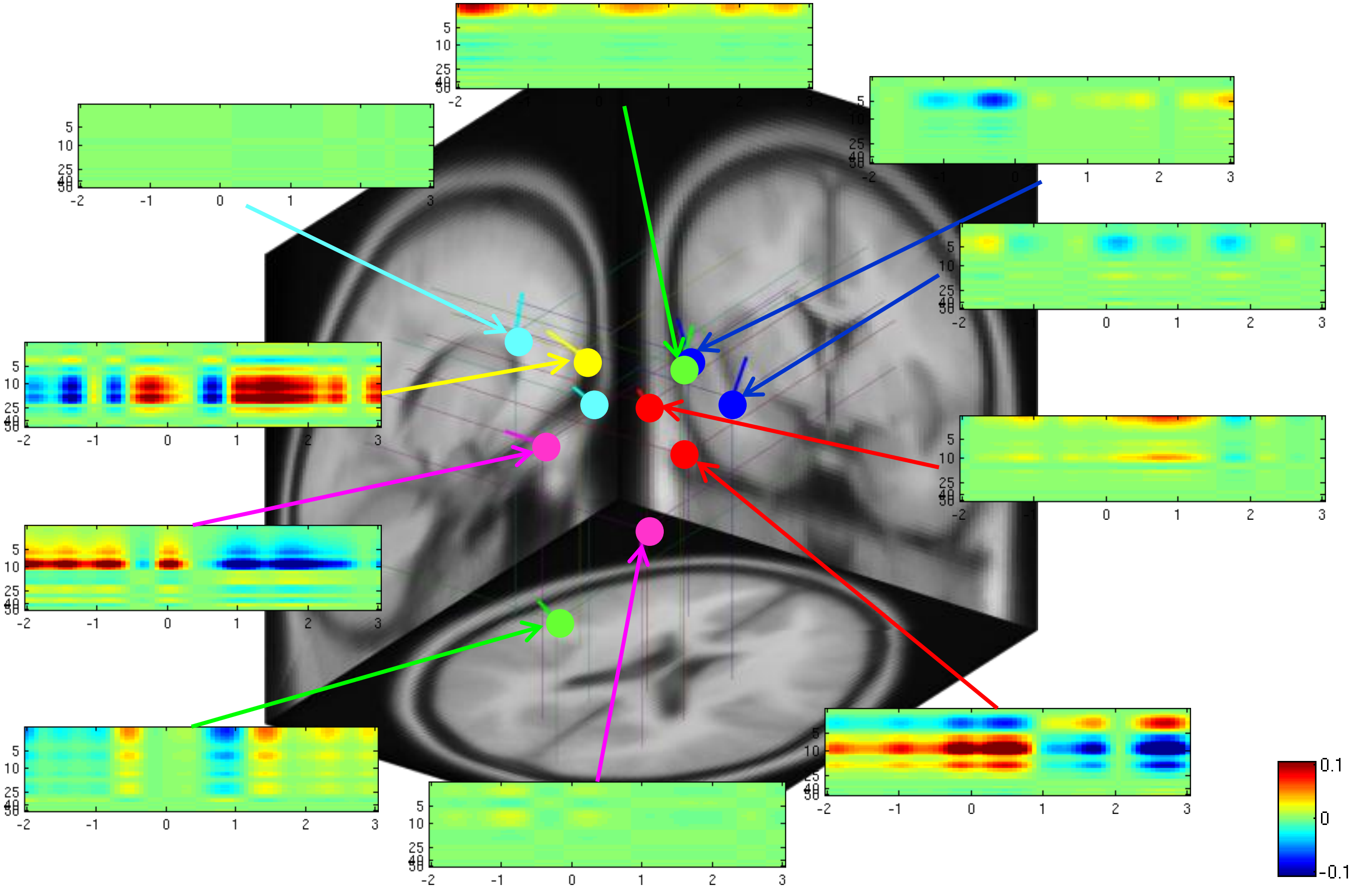
- **Control by Thought, Mood Assessment/Display**

Jedi Game Prototype

necomimi “neurowear”

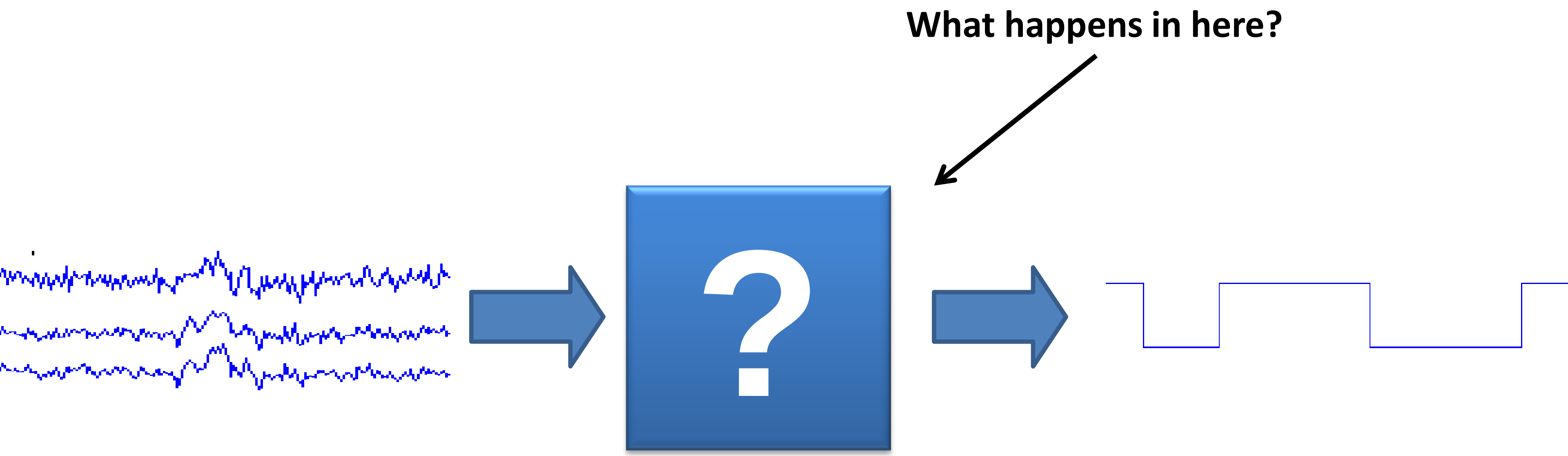
Neuroscience

- **Multivariate Pattern Analysis / Brain Imaging**



Kothe et al., 2011

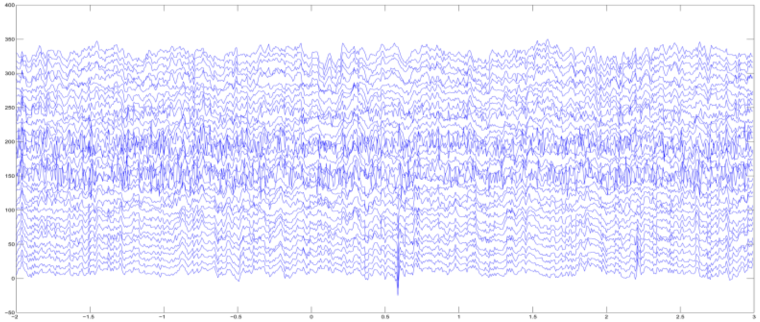
3 Basic Theoretical Principles and Framework



Component 1: Predictive Mapping

Central Predictive Mapping

- A BCI (with limited memory of the past) can be viewed as a mathematical function *f*:

y = f(**X**); **X**=  y= “subj. excited” (+1)
“subj. not excited” (-1)

- The functional form is arbitrary, for example

$$y = \mathrm{sign}(\mathrm{var}(\boldsymbol{WX}) + b)$$

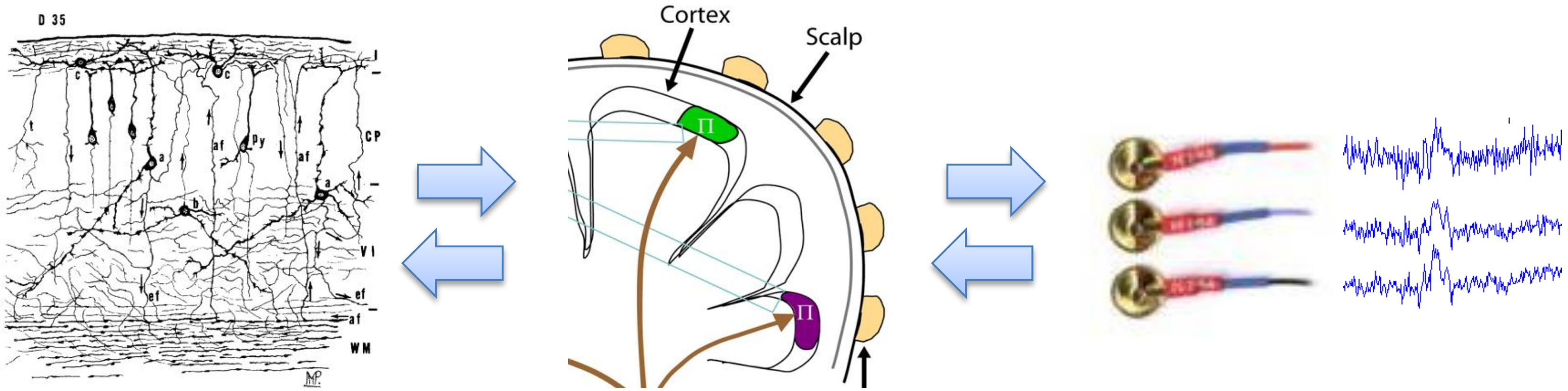
- The mapping involves free parameters, here **W** and b, and data from a *sliding window* **X**

Functional Form

- Reflects the relationship between observation (data segment **X**) and desired output (cognitive state parameter y)

Functional Form

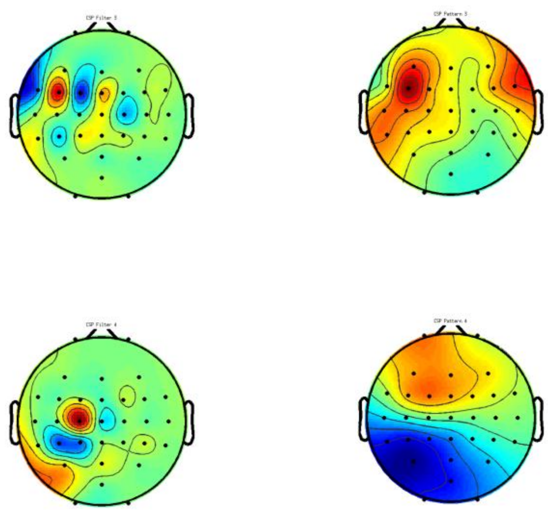
- Reflects the relationship between observation (data segment **X**) and desired output (cognitive state parameter y)
- Based on some assumed generative mechanism (forward model) – or ad hoc



- Note: Functional form is the inverse mapping!

Basic Ingredient: Spatial Filter

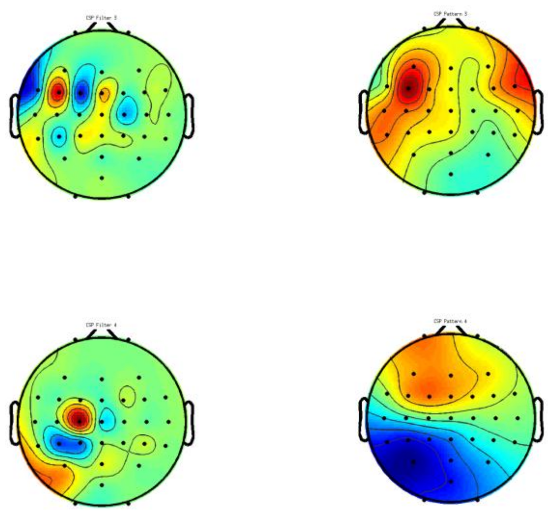
- Linear inverse of volume conduction effect between sources **S** and channels **X**
  $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$ (forward)
  $\boldsymbol{S} = \boldsymbol{W}\boldsymbol{X}$ (inverse)



W **$A=W^{-1}$**

Basic Ingredient: Spatial Filter

- Linear inverse of volume conduction effect between sources **S** and channels **X**
  $\boldsymbol{X} = \boldsymbol{AS}$ (forward)
  $\boldsymbol{S} = \boldsymbol{WX}$ (inverse)

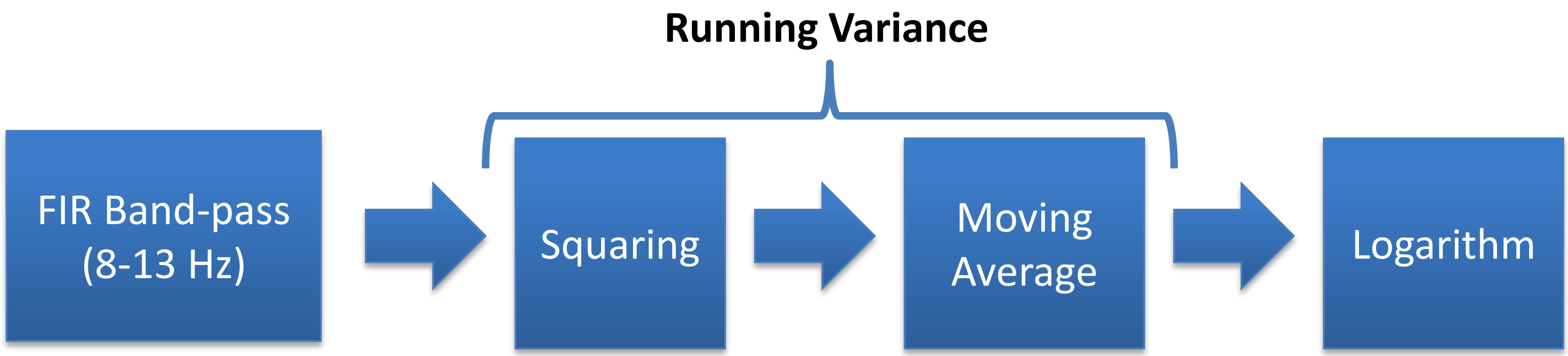


c.f. $y = \mathrm{sign}(\mathrm{var}(\boldsymbol{WX}) + b)$

Component 2: Signal Processing

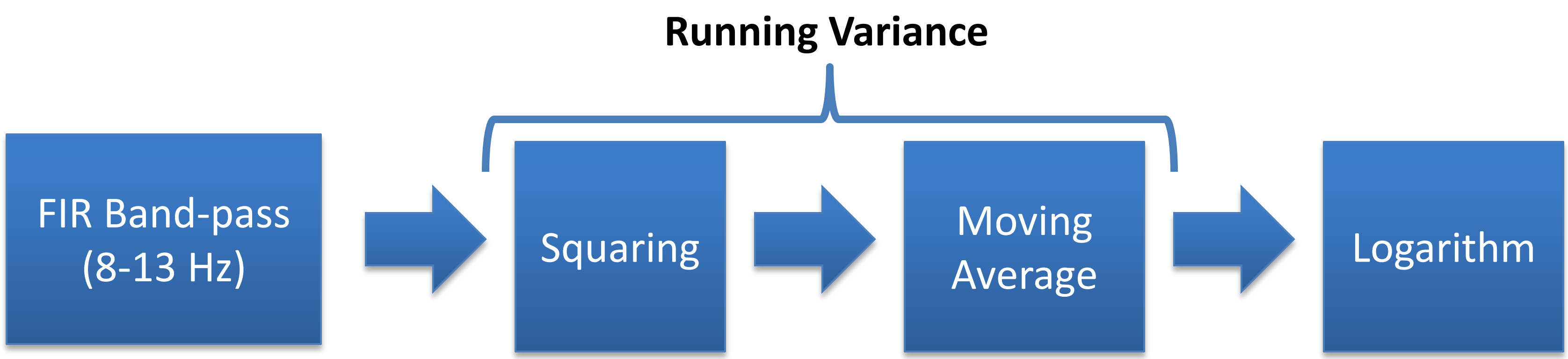
Role of Signal Processing

- BCIs can be constructed from Signal Processing blocks (digital filters):



Role of Signal Processing

- BCIs can be constructed from Signal Processing blocks (digital filters):

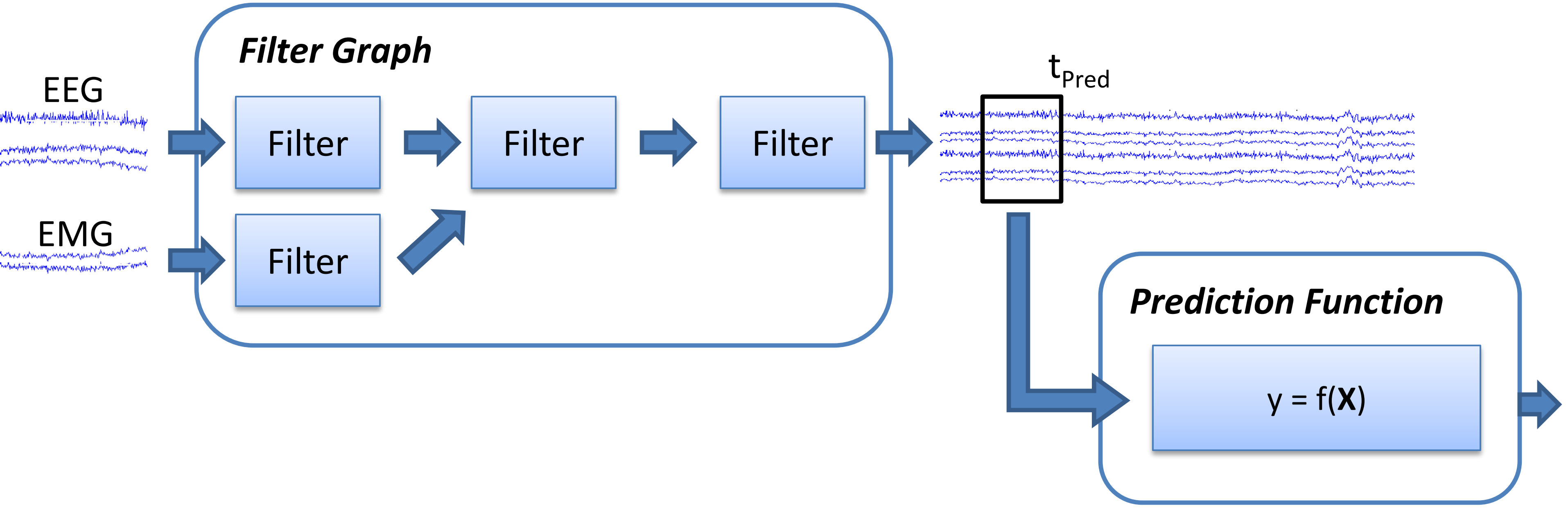


- This produces the same output as the following functional-style description (***T*** is a temporal filter matrix) :

$$f(\boldsymbol{X}) \;\; := \;\; y = \log var(\boldsymbol{XT})$$

Role of Signal Processing

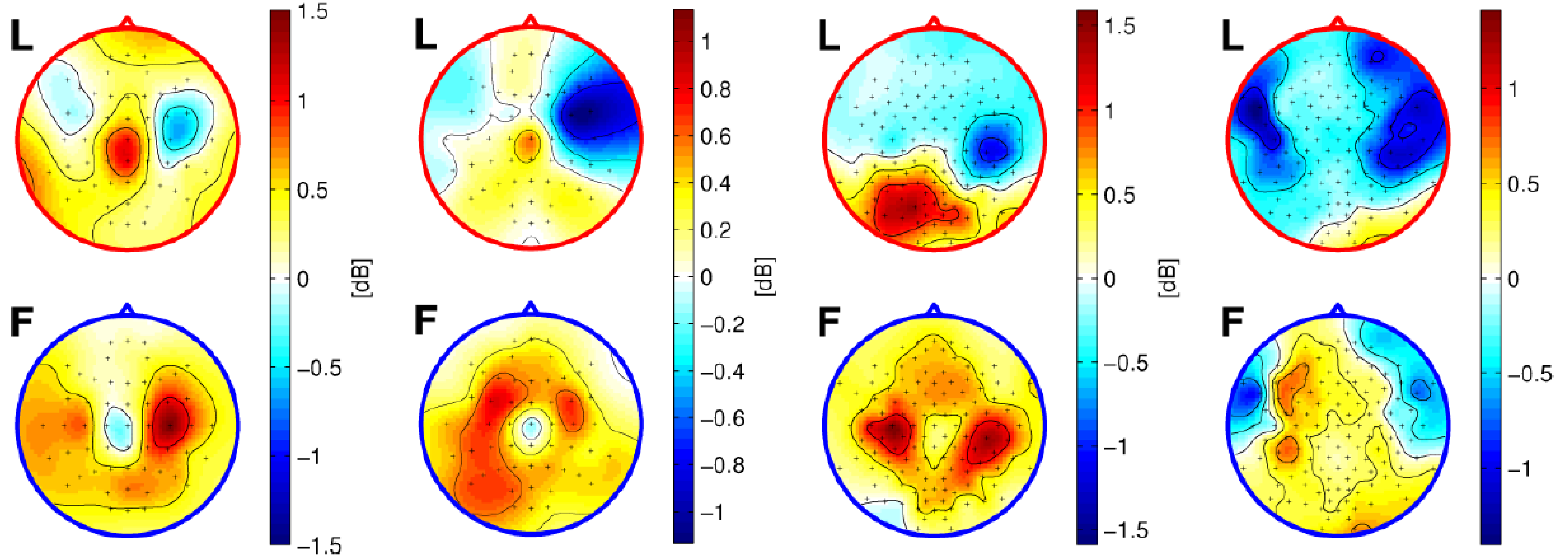
- Both frameworks are complementary, rather than contradictory, and are in practice often used *in combination*, e.g. to minimize computational costs



Component 3: Machine Learning

The Problem of Unknown Parameters

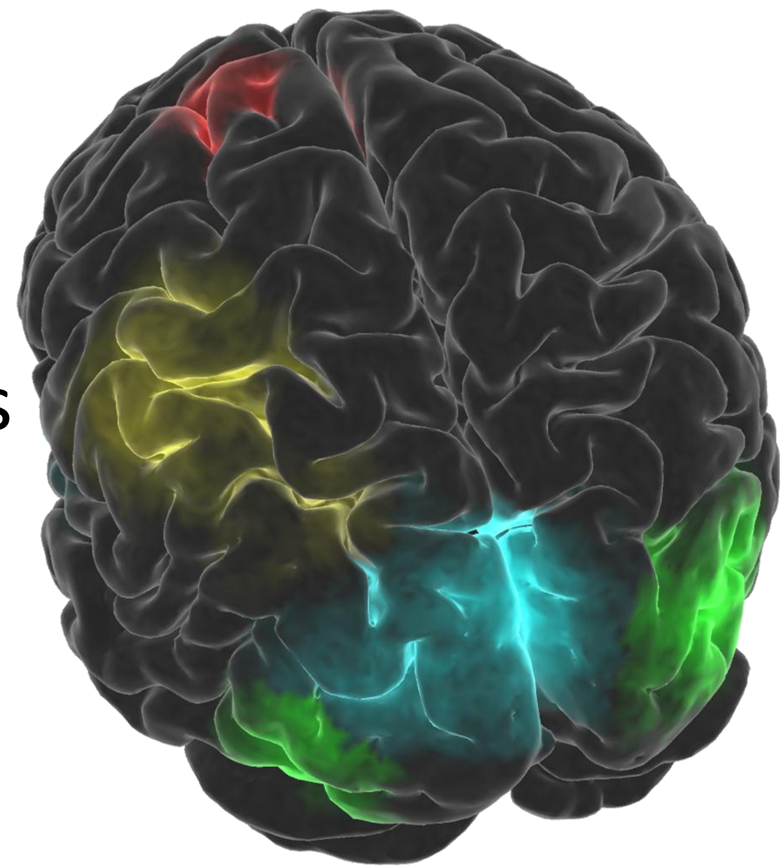
- Processing depends on unknown parameters (person-specific, task-specific, otherwise variable) – e.g., per-sensor weights as below:



Blankertz et al. 2007

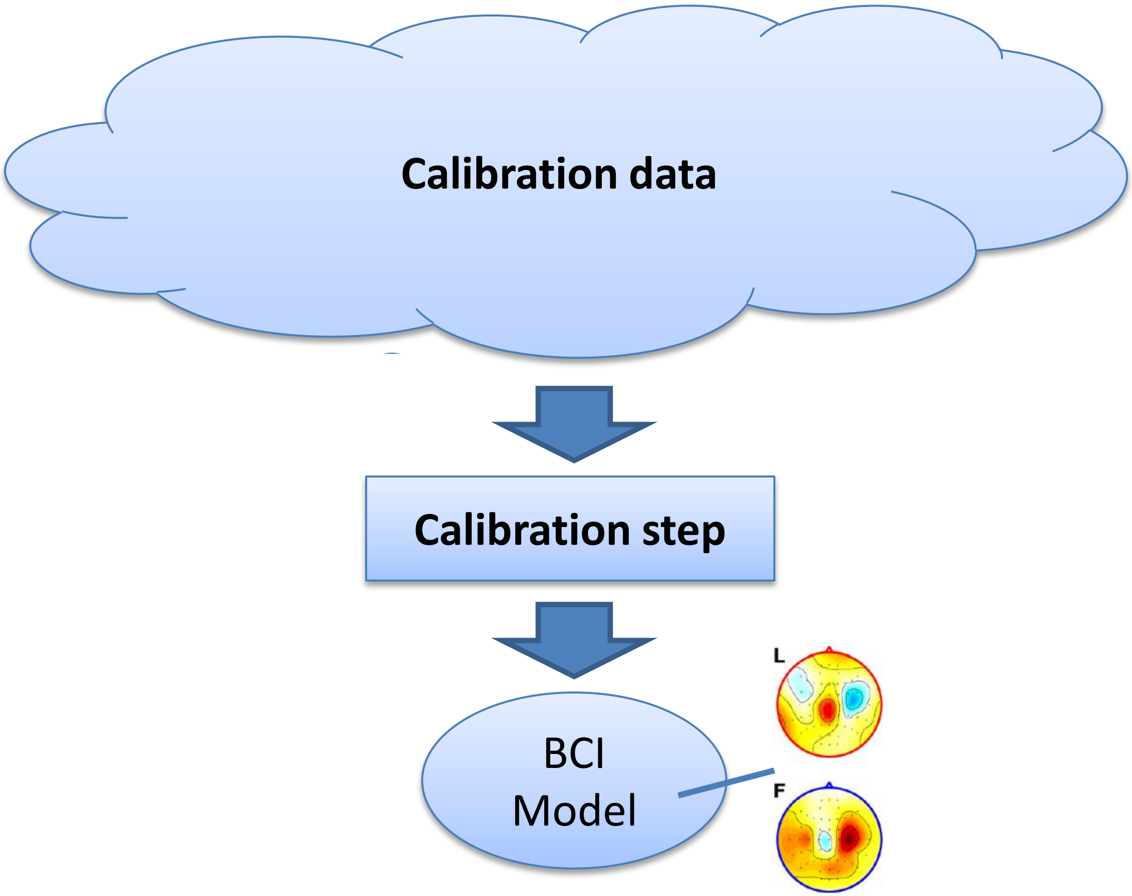
Reasons for Parameter Uncertainty

- Folding of cortex differs between any two persons (even identical twins)
- Relevant functional map differs across individuals
- Sensor locations differ across recording sessions
- Brain dynamics are non-stationary at all time scales



Solution: Calibration

- *Calibration / training data* can be used to estimate parameters, during a separate *calibration step*

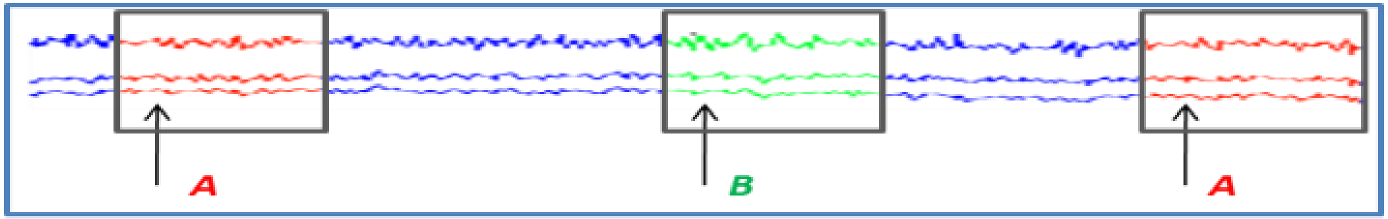


Calibration Data

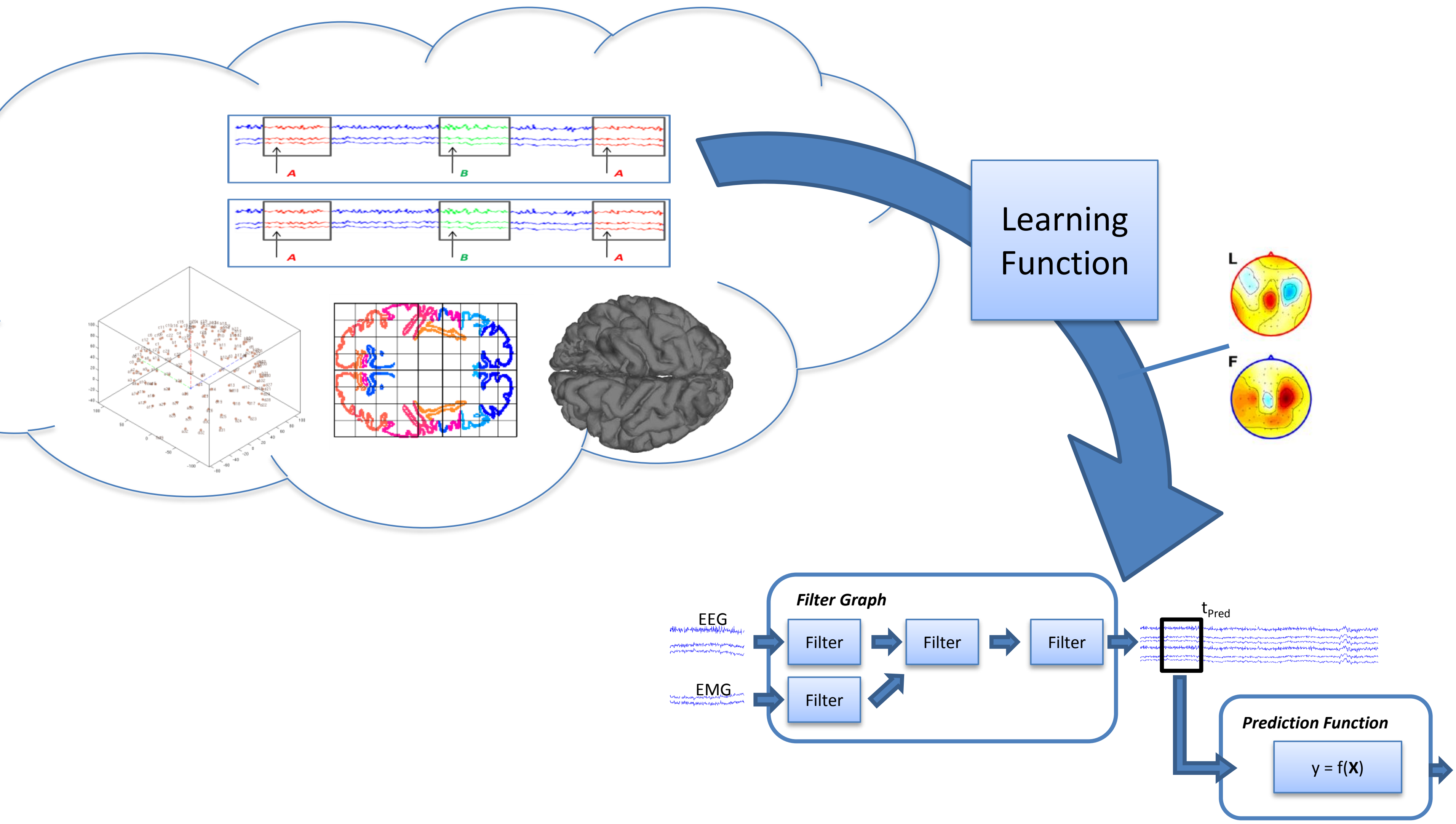
- Many possible kinds of data could be used
- Best known type of calibration data: ***example data***, i.e. examples of EEG of a person being excited, not excited, etc.
- Collected in a special *calibration recording* (before actual online use of the BCI)

Calibration Recording

- Similar to standard psychological experiments:
  - continuous EEG (or other)
  - multiple trials/blocks (capturing variation)
  - randomized (eliminating confounds)
  - event markers to encode cognitive state conditions of interest, e.g., stimuli/responses (called "*target markers*" in BCILAB)
- Can also be used for offline performance tests

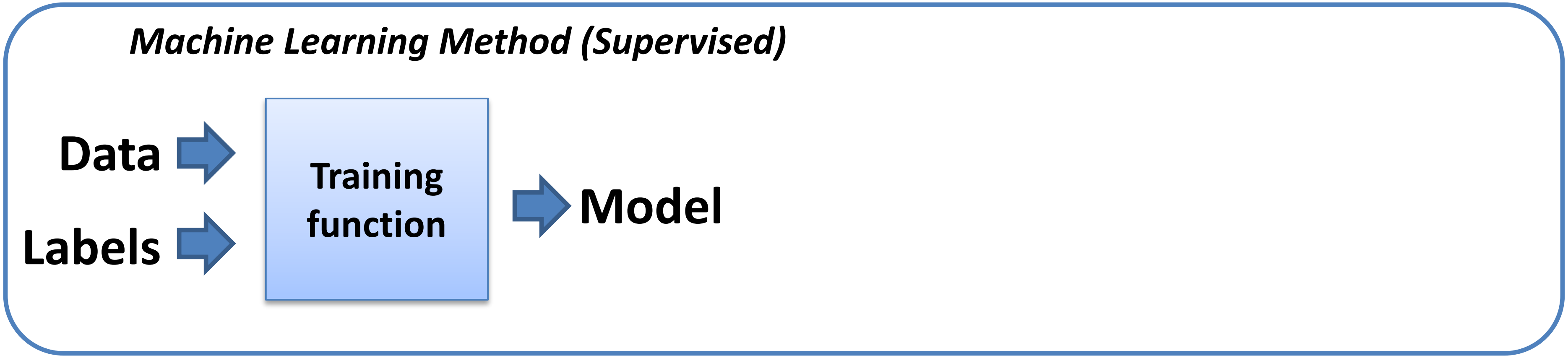


Big Picture


A
B
A
A
B
A
Learning Function
L
F
Filter Graph
EEG
EMG
Filter
Filter
Filter
Filter
t_Pred
Prediction Function
y = f(X)

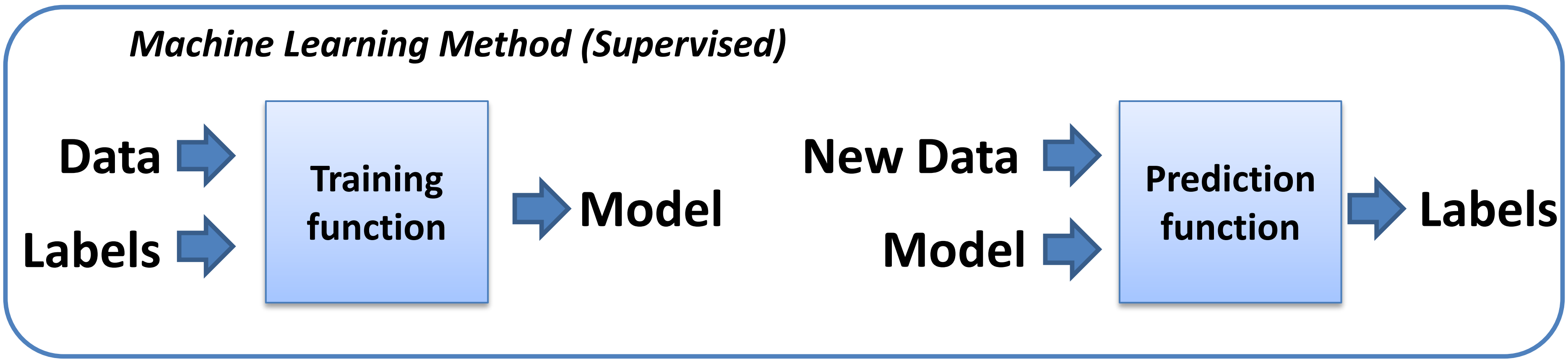
Machine Learning Framework

- Large field with 100s of algorithms
- Most methods conform to a common framework of a *training function* and a *prediction function*



Machine Learning Framework

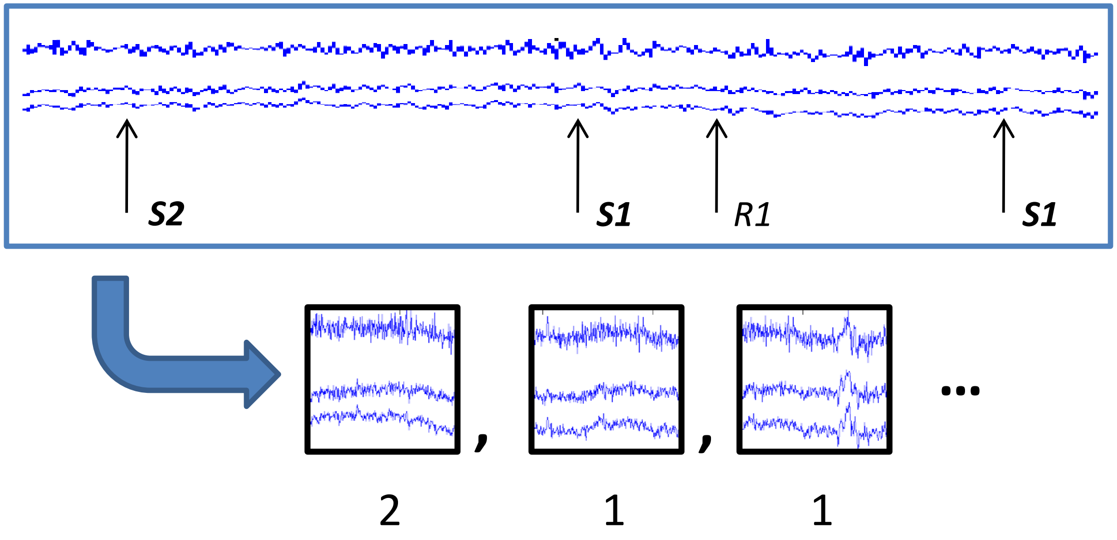
- Large field with 100s of algorithms
- Most methods conform to a common framework of a *training function* and a *prediction function*



- Intermediate model parameters capture the learned relationship

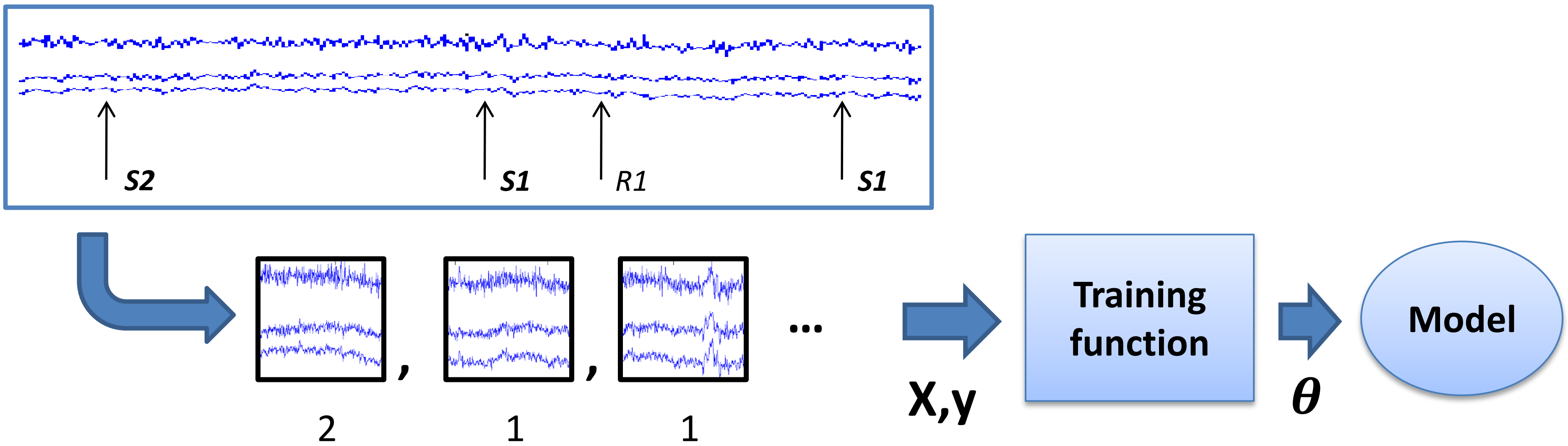
Using Machine Learning

- Often, one trial segment (sample) is extracted for every target marker in the calibration recording and is used as *training exemplar* $\boldsymbol{X}_k$
- Its associated label y_k can be deduced from the target marker



Using Machine Learning

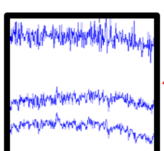
- The training function computes a parameter (here $\boldsymbol{\theta}$) of the prediction function *such that the performance of the prediction function on the given example data is optimal*



Detour: Feature Extraction

- **Caveat:** Off-the-shelf machine learning methods often do not work very well when applied to raw signal segments of the calibration recording
  - too high-dimensional (too many parameters to fit)
  - too complex structure to be captured (too much modeling freedom)

1000s of degrees of freedom!



Detour: Feature Extraction

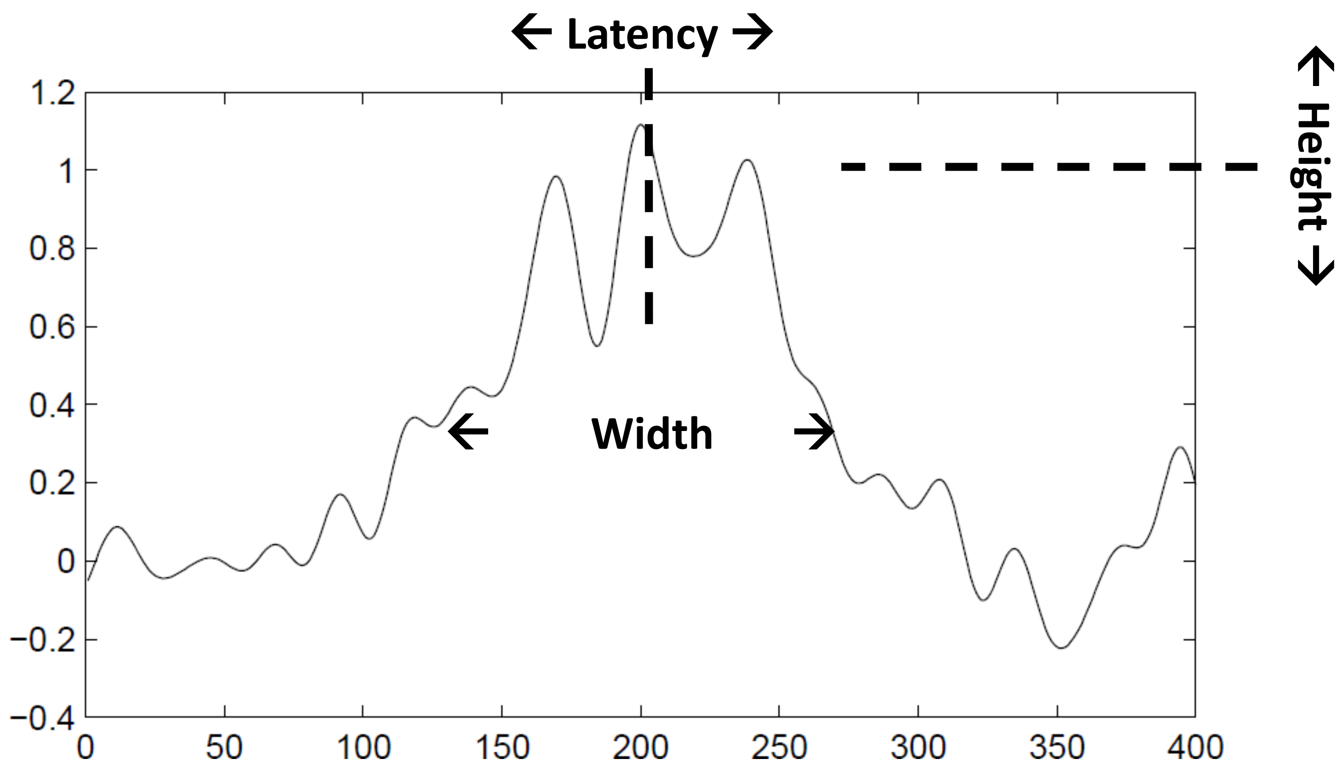
- **Solution**: Introduce additional mapping (called *“feature extraction”)* from raw signal segments onto feature vectors which extracts the *key features* of a raw observation
  - output is usually of lower dimensionality
  - hopefully statistically “better” distributed (easier to handle for machine learning)

Example for Feature Extraction

- **Task:** A person is presented with a sequence of 300 images (one ever 2 seconds). Half of the images are exciting, the other half are not. One channel of EEG (at Cz location) is recorded.
- **Question:** How to design a BCI that can determine whether a person is shown an exciting or a non-exciting image?
- **Approach:** For each trial k, cut out an epoch $\mathbf{X}_k$ of 1s length, extract a short vector of features $\mathbf{f}_k$, and assign a label y_k in {E,NE}. Use machine learning to find an optimal statistical mapping from $\mathbf{f}_k$ onto y_k.

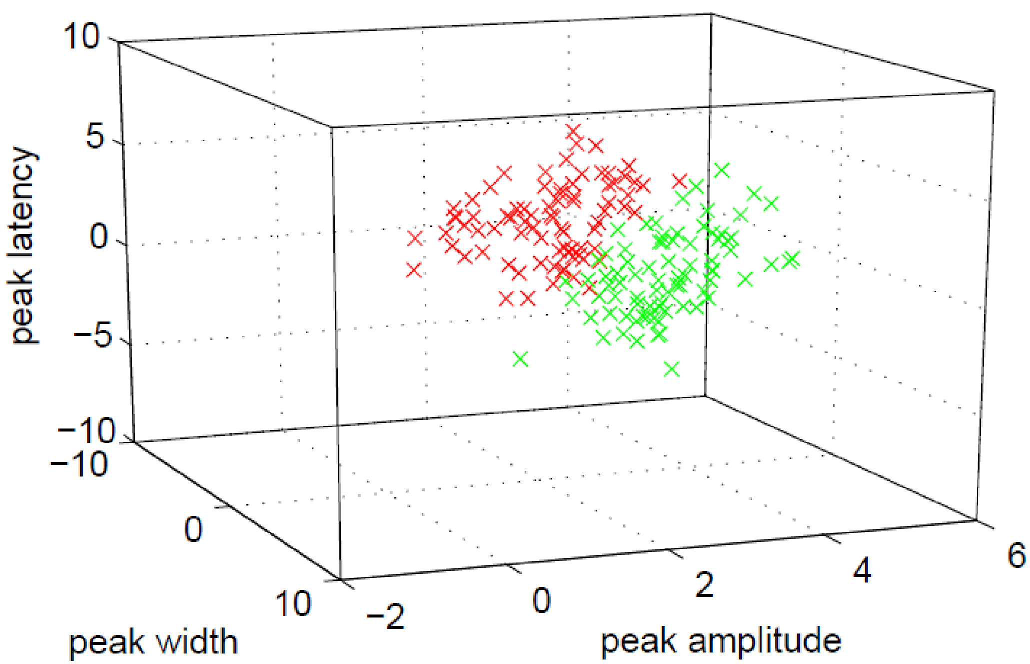
Example:
Features of an ERP Peak

- A supposed characteristic peak in a time window (relative to an event) could be characterized by three parameters:



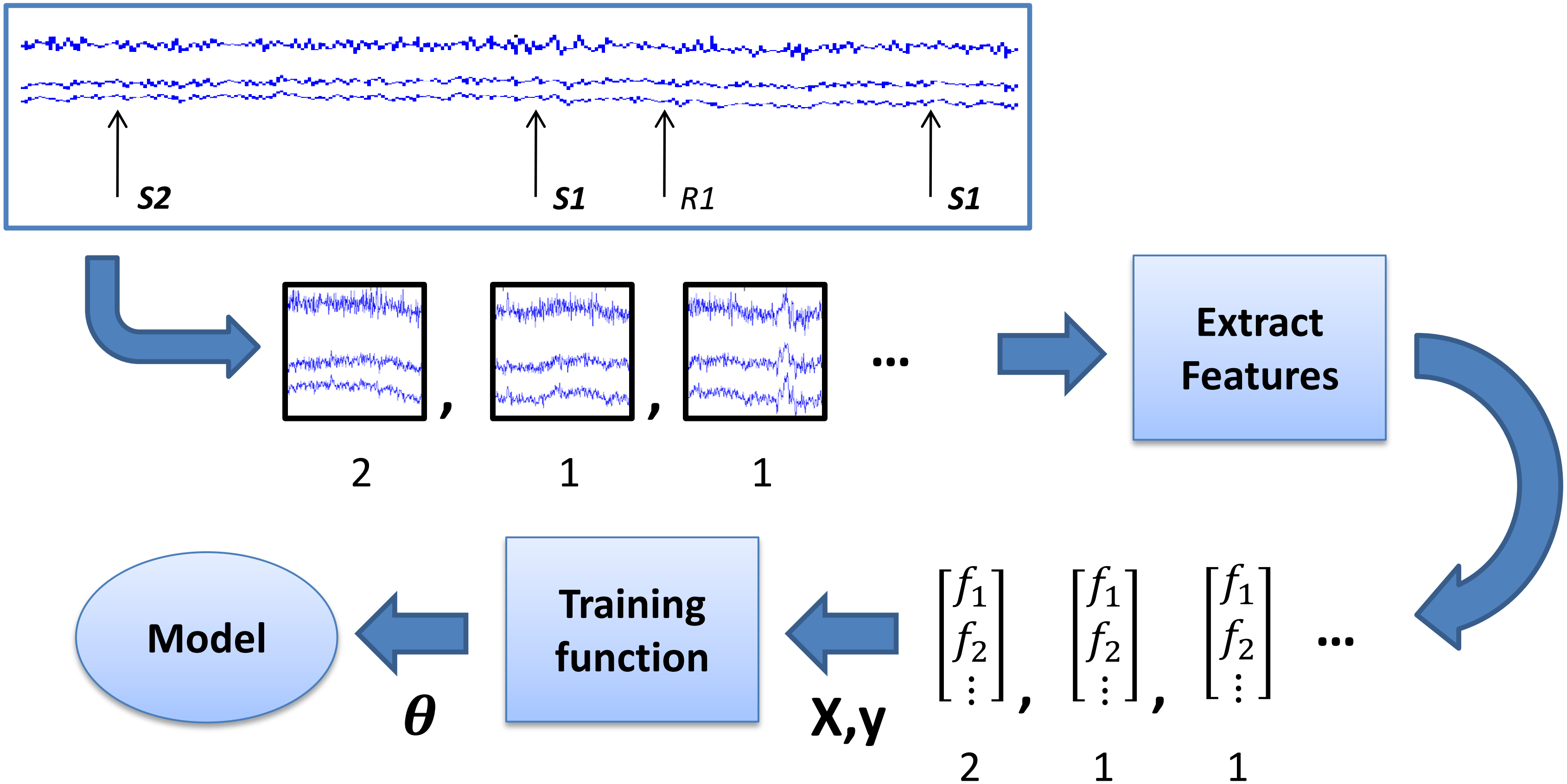
Resulting Feature Space

- Plotting the 3-element feature vectors for all exciting trials in red, and non-exciting trials in green, we obtain two distributions in a 3d space:



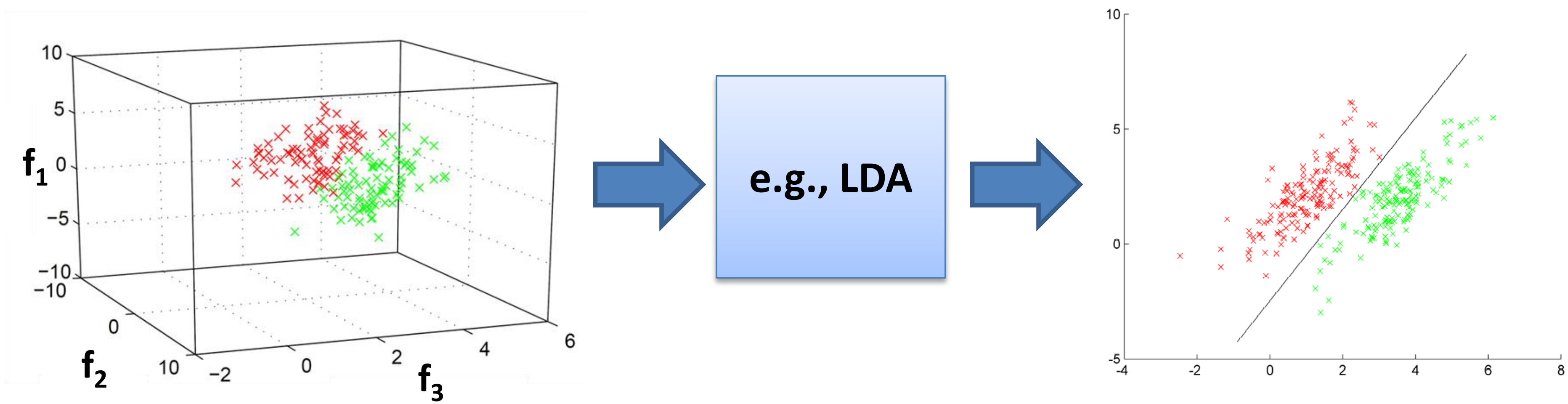
ML with Feature Extraction

- Including the feature extraction, the analysis process is as follows:



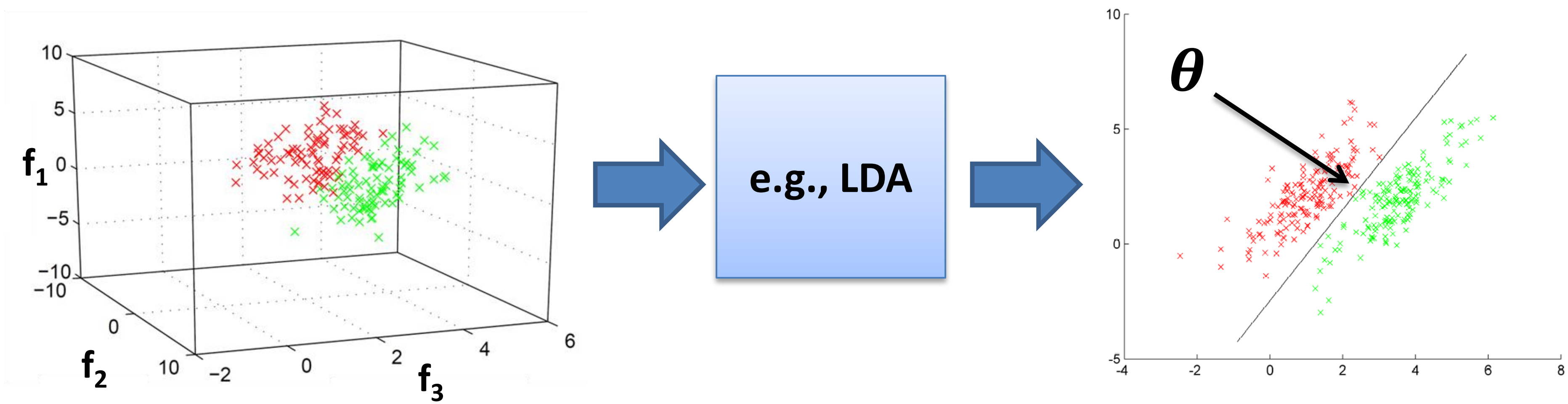
Using Machine Learning

- The feature vectors are passed on to a machine learning function (e.g., Linear Discriminant Analysis)



Using Machine Learning

- The feature vectors are passed on to a machine learning function (e.g., Linear Discriminant Analysis)
- … which determines a parametric predictive mapping

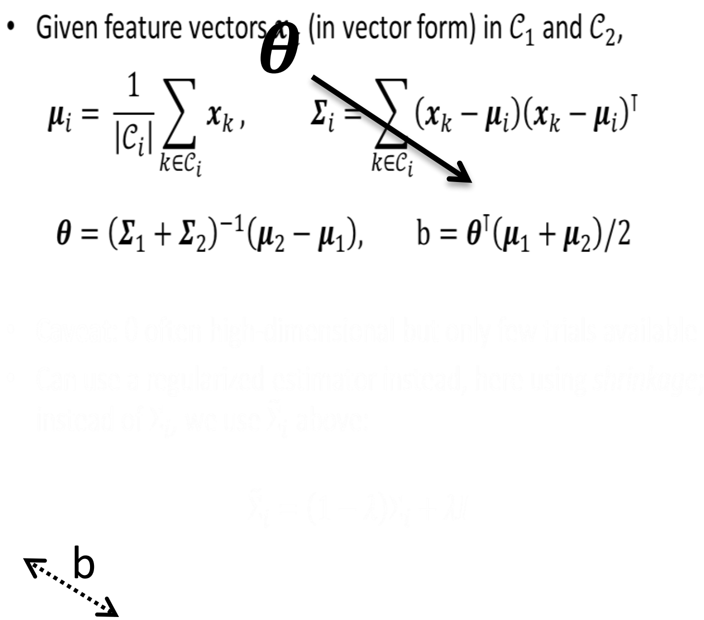


LDA In a Nutshell

- Given feature vectors $\boldsymbol{x}_k$ (in vector form) in $\mathcal{C}_1$ and $\mathcal{C}_2$,

$$\boldsymbol{\mu}_i = \frac{1}{|\mathcal{C}_i|} \sum_{k \in \mathcal{C}_i} \boldsymbol{x}_k \,, \qquad \boldsymbol{\Sigma}_i = \sum_{k \in \mathcal{C}_i} (\boldsymbol{x}_k - \boldsymbol{\mu}_i)(\boldsymbol{x}_k - \boldsymbol{\mu}_i)^\intercal$$

$$\boldsymbol{\theta} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1), \qquad \mathrm{b} = \boldsymbol{\theta}^\intercal(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$$



Resulting Predictive Map

- LDA generates parameters of a linear mapping

$$\mathrm{y} = \boldsymbol{\theta x} - b$$

- For classification, the mapping is actually *non-linear*:

$$\mathrm{y} = \mathrm{sign}(\boldsymbol{\theta x} - b)$$

LDA Assumptions

- Gaussian noise distribution for each class of trials
- Noise covariance is independent of class (i.e., identical for both groups of trials)
- Optimal in the limit of infinite data
- Note: LDA can also be generalized to multiple classes

4 Analyzing ERP-like Processes

(properly)

Experimental Task

- **Flanker Task:** The experiment consists of a sequence of ca. 330 trials with inter-trial interval of 2s +/- 1.5s
- At the beginning of each trial, an arrow is presented centrally (pointing either left or right)
- The arrow is flanked by congruent or incongruent “flanker” arrows (coming slightly earlier):

← ← → ← ←

- The subject is asked to press the left/right button, according to the central arrow, and makes frequent errors (25%)

Consideration

- The peak ERP features discussed so far were chosen *for a single channel* of EEG
- **Problem:** with multiple channels all channels measure almost the same signal properties, thus little information gain to expect
- **Solution:** *Learn* a spatial filter and use multiple channels to *computationally focus* on source processes of interest, then extract *source signal features*

Consideration

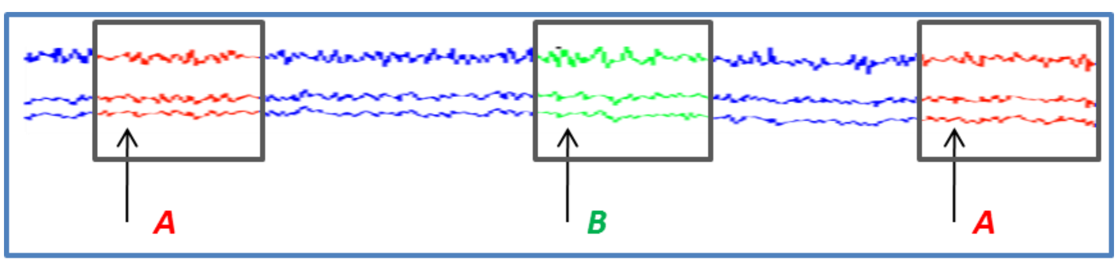
- This can be done automatically by a linear classifier when applied to multiple channels
- Works only for source-signal features that are a *linear transform* of channel-signal features
- The classifier must produce the *same solution under rotation and scaling* (not all do, but e.g., LDA does)

Approach

- Calibration recording is band-pass filtered between 0.5Hz and 15Hz
  - lower edge removes drifts
  - upper edge cuts off high-frequency noise

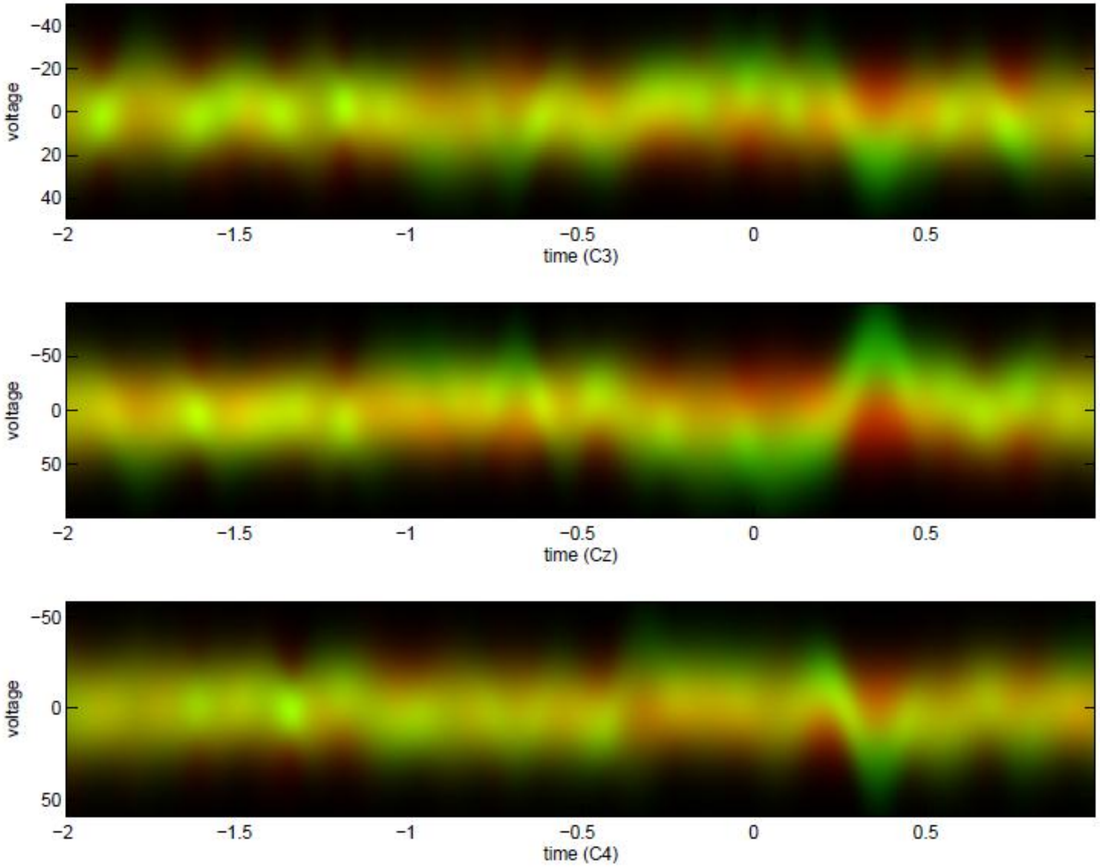
Approach

- Calibration recording is band-pass filtered between 0.5Hz and 15Hz
  - lower edge removes drifts
  - upper edge cuts off high-frequency noise
- Epochs are extracted for each trial and label is set to A for incorrect trials and B for corrects

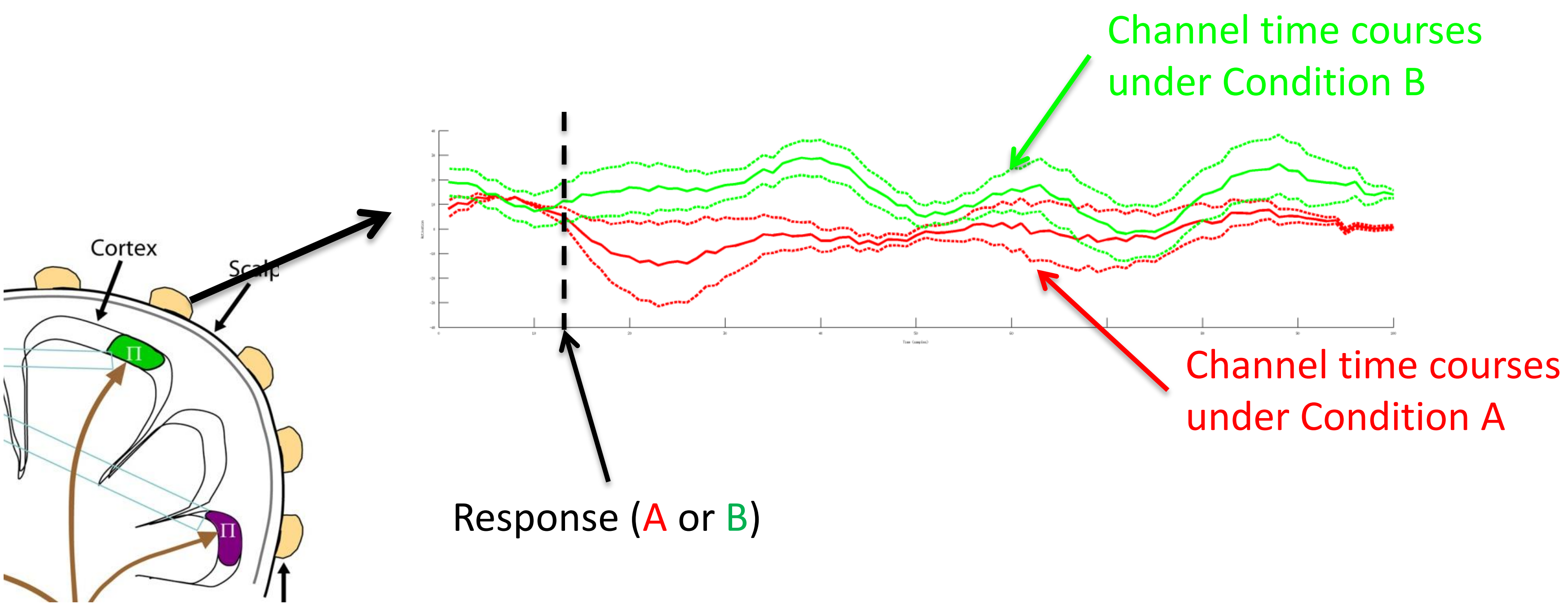


Actual Data

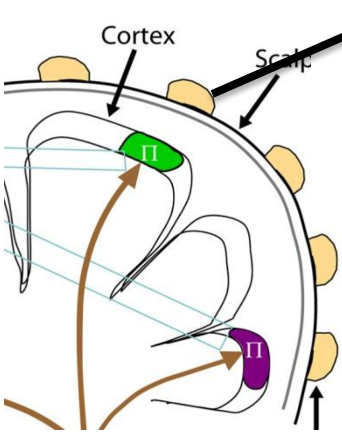
- Time courses for all trials super-imposed (color-coded by class) – but here different task

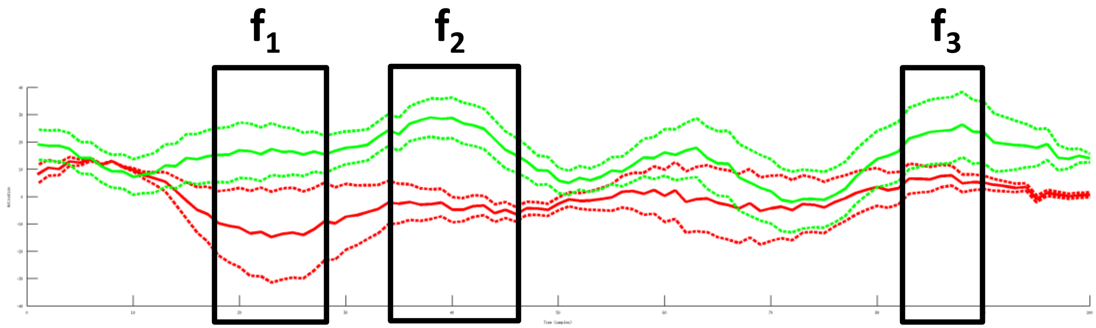


Extracted Epochs

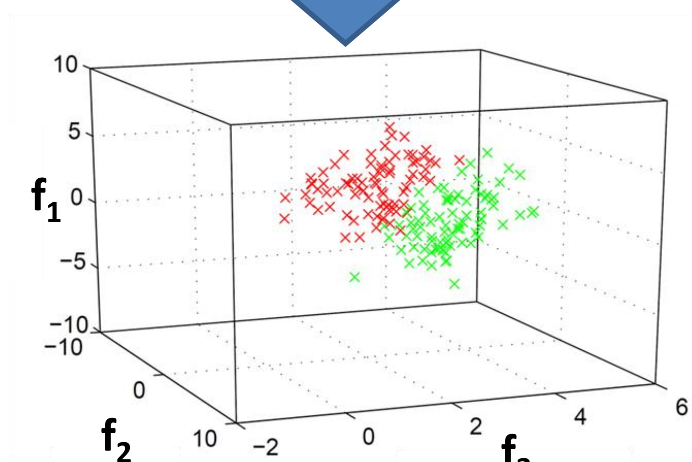


Extracting Linear Features



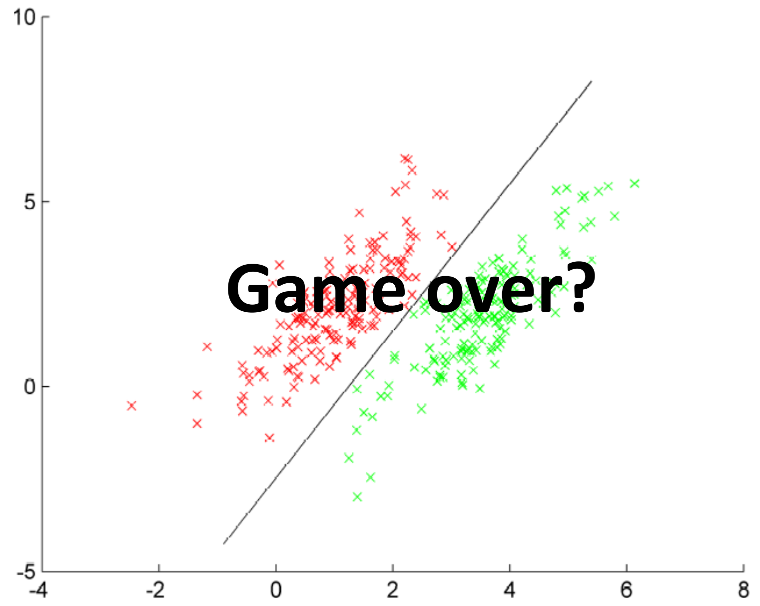


For each trial segment, calculate signal mean in 3 time sub-windows (→ 3-dim feature vector)



Problem with LDA

- Multi-channel features are usually too high-dimensional for LDA to handle with few trials!

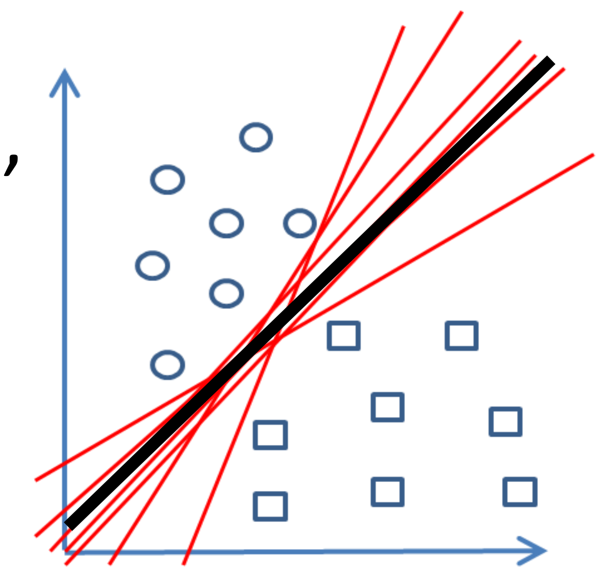


Problem with LDA

- Multi-channel features are usually too high-dimensional for LDA to handle with few trials!
- There is a simple generalization to LDA called *shrinkage LDA* that can handle such feature spaces

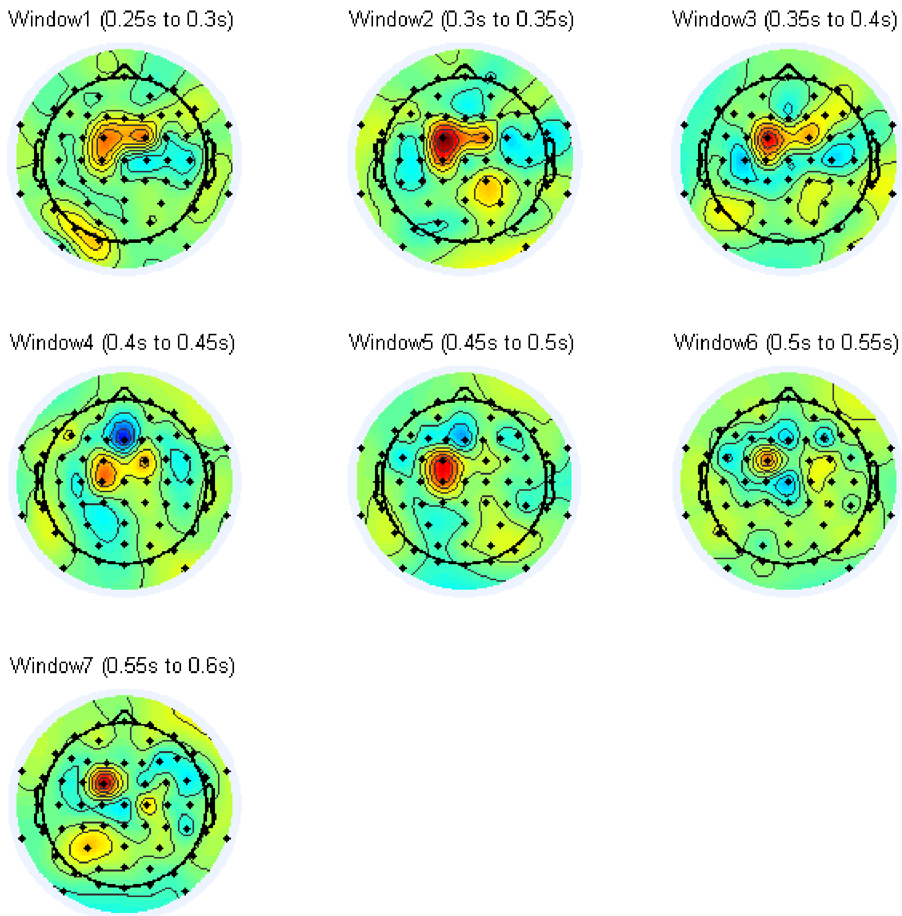
Problem with LDA

- Multi-channel features are usually too high-dimensional for LDA to handle with few trials!
- There is a simple generalization to LDA called *shrinkage LDA* that can handle such feature spaces
- *Many alternative methods* for high-dimensional data exist (e.g., Support Vector Machines, Regularized Logistic Regression)



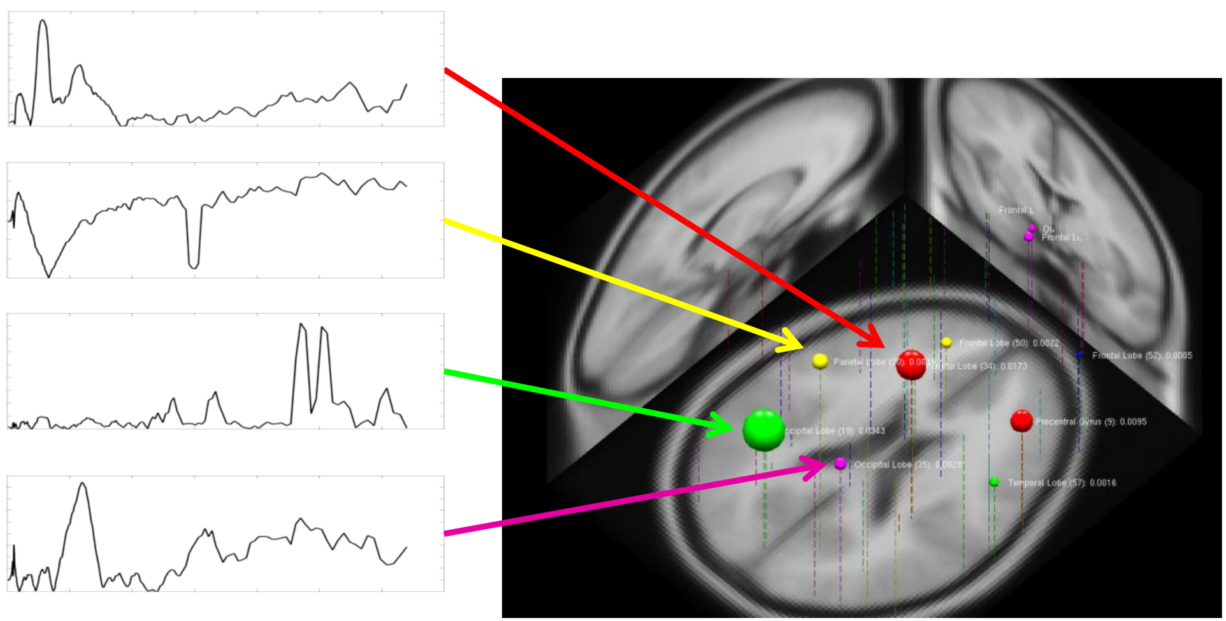
Resulting Spatial Filters

- Topographically mapped, the following filters emerge:



A Note on Interpretability

- Spatial filters are not very interpretable
- When the classifier is applied to localizable features (e.g., on *independent components*), the weights assigned by it *are also localized*
- Example:



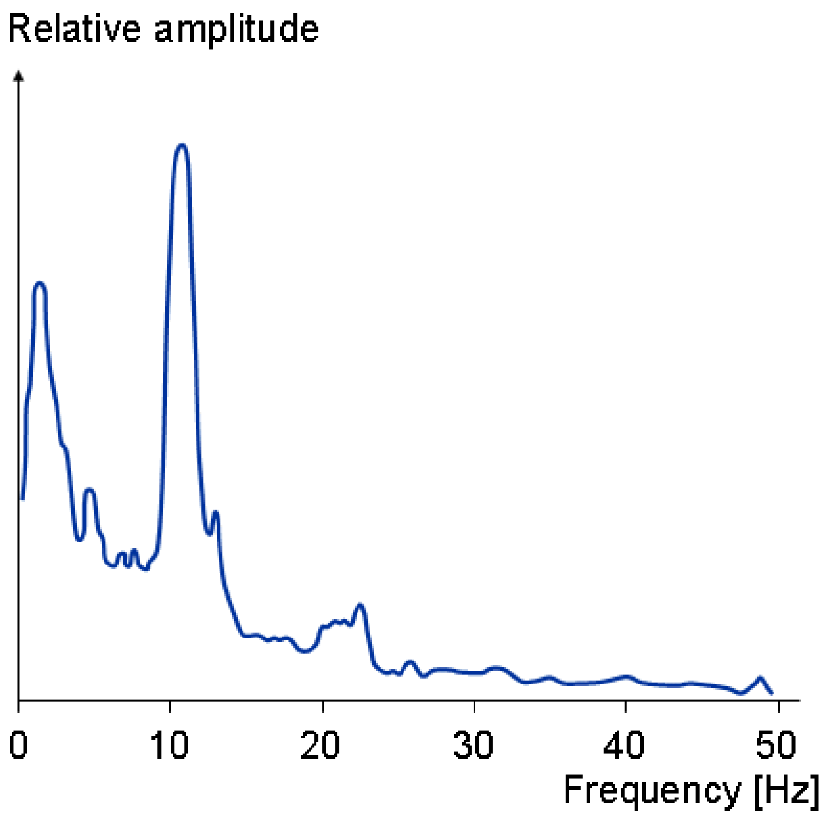
How Good is This Approach?

- Source activation S can be recovered from sensor measurements by a linear mapping if (linear) volume conduction is invertible ($\boldsymbol{S} = \boldsymbol{W}\boldsymbol{X}$)
- Assuming a jointly Gaussian noise process and a noise distribution that is independent of the condition (A/B), LDA approximates the *optimal linear mapping*
- Shrinkage LDA on these features yields **state-of-the-art ERP performance**! (although the assumptions are not entirely true)

5 Analyzing Oscillatory Processes

Oscillatory Processes

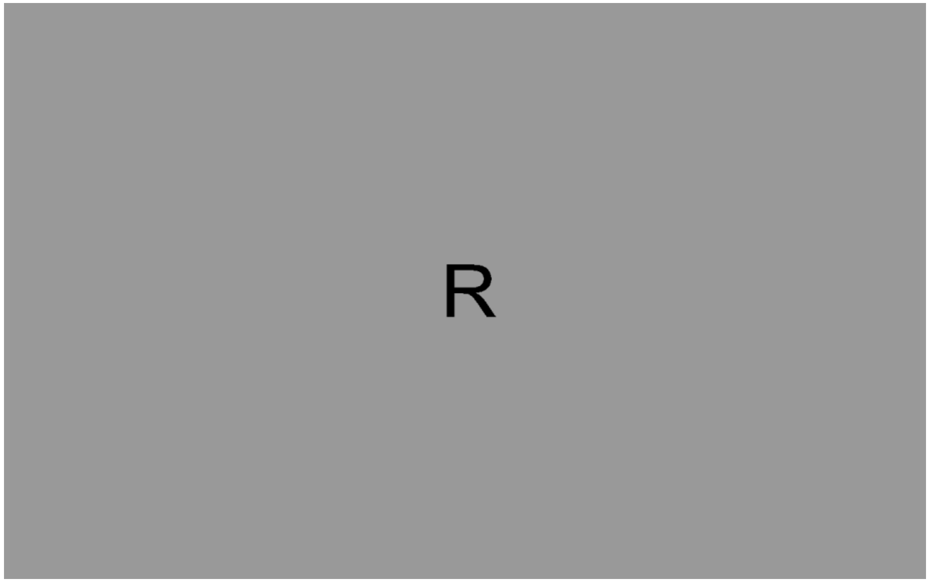
- **Best example:** cortical idle rhythms, e.g. occipital alpha, motor cortex alpha+beta



Malmivuo and Plonsey, 1995

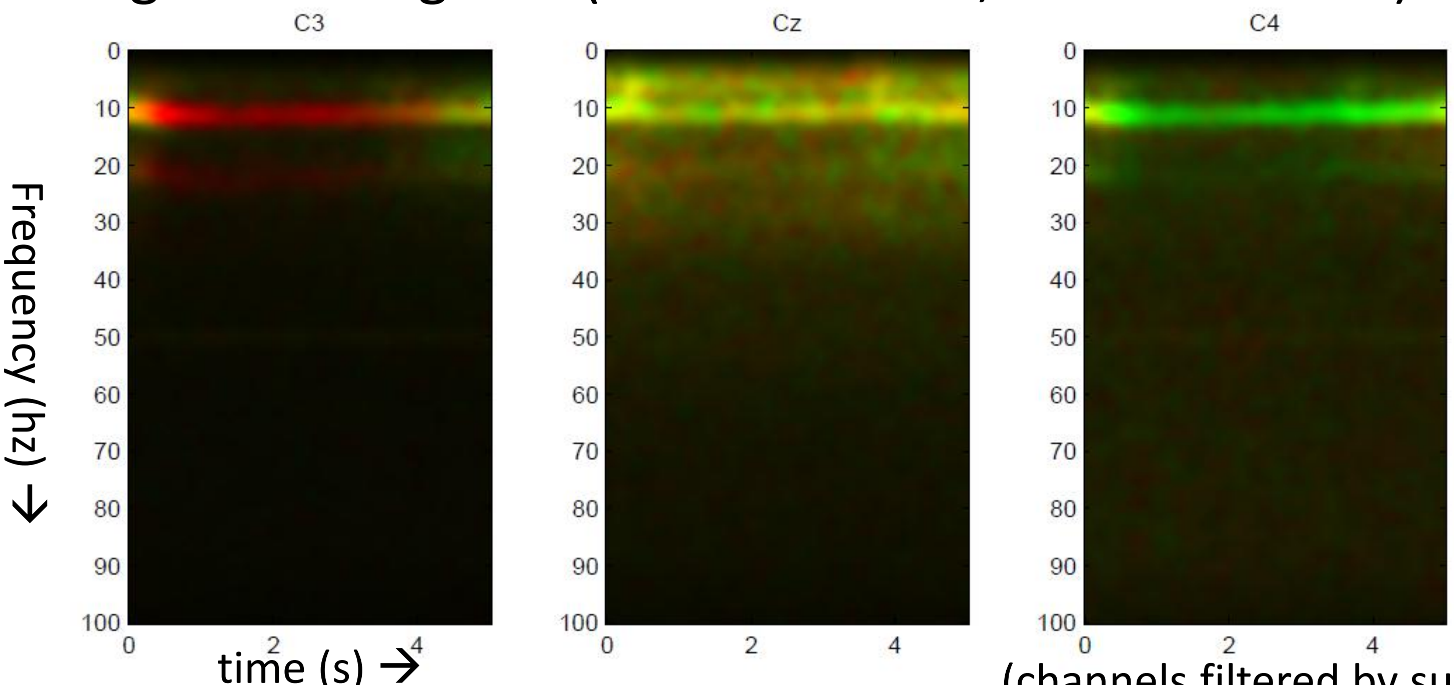
Sample Experimental Task

- The experiment consists of 160 trials (pause at ½ the experiment). Each trial begins with a letter (either L or R) displayed for 3s. The subject is instructed to subsequently imagine either a left-hand or a right-hand movement. Each trial ends with a blank screen displayed for 3.5s.



Motor Cortex ERD/ERS

- **Event-Related Synchronization / Desynchronization:** attentuation of motoric idle rhythms in response to an event
- Average spectrogram for left-hand movement imagination in red **+** average spectrogram for right-hand movement imagination in green (160 trials each, stimulus at t=0)



(channels filtered by surface Laplacian)

The Problem In Oscillatory BCIs

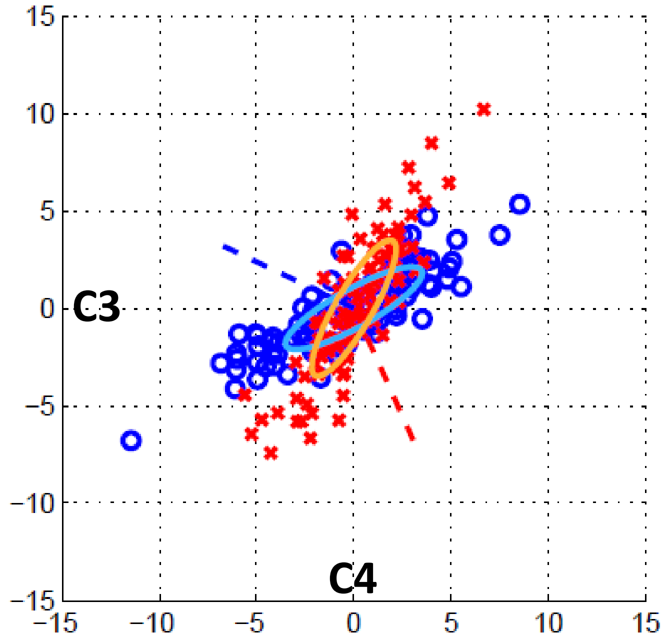
- Calculating the power or amplitude of an oscillation requires a *squaring of the signal*
- This is *after spatial filtering*, i.e. the spatial filter must be adapted such that the squared signal (or its variance) is maximally informative
- If multiple source amplitudes are involved, they need to be weighted by *another learned linear mapping* (after squaring)

Common Spatial Patterns

- Most popular algorithm in BCI field for learning spatial filters for oscillatory processes
- **Assumptions:**
  - Frequency band and time window are known
  - band-passed signal is jointly Gaussian within the time window
  - Source activity constellation differs between two classes

Common Spatial Patterns

- Below: Different EEG signals for a single left-hand epoch vs. a single right-hand epoch (band-passed to 7-30 Hz)
- Signal activation is scatter-plotted for channels C3 and C4:

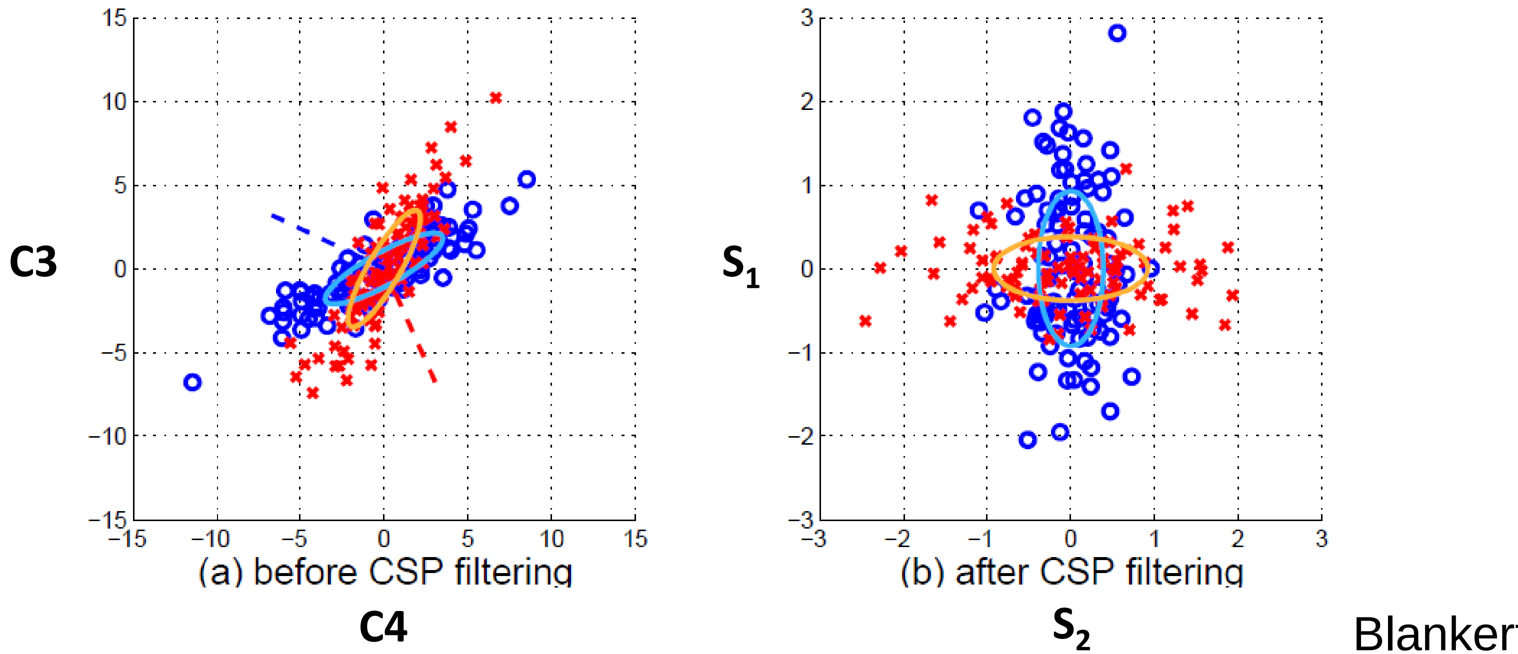


(a) before CSP filtering

Blankertz 2009

Common Spatial Patterns

- **Goal:** Design spatial filters (i.e., linear transforms) such that the signal's variance along the filtered direction is maximal for one condition while minimal for the other
- Ideally find multiple filters with that property



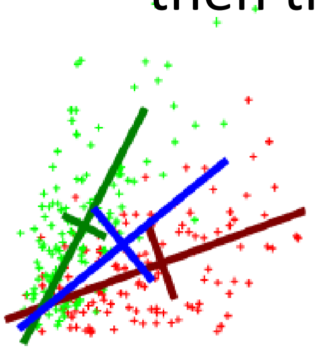
Blankertz 2009

Swartz Center for Computational Neuroscience

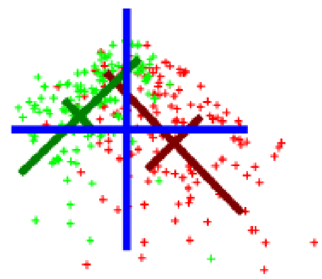
One Way to Compute It

- **Geometric Approach:** An intuitive approach is a three-step procedure:
  1. Determine a *whitening* transform ***U*** for the average of both covariance matrices (blue) using PCA
  2. Apply it to one of the point clouds and calculate its principal components ***P*** (green)
  3. The spatial filter operation ***W*** is to first whiten by ***U*** and then transform by ***P***$^{-1}$, i.e. $\boldsymbol{W} = \boldsymbol{P}^{-1}\boldsymbol{U}$ so then $\boldsymbol{S} = \boldsymbol{W}\boldsymbol{X}$

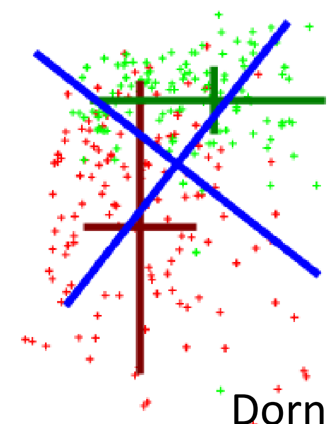
1.



2.



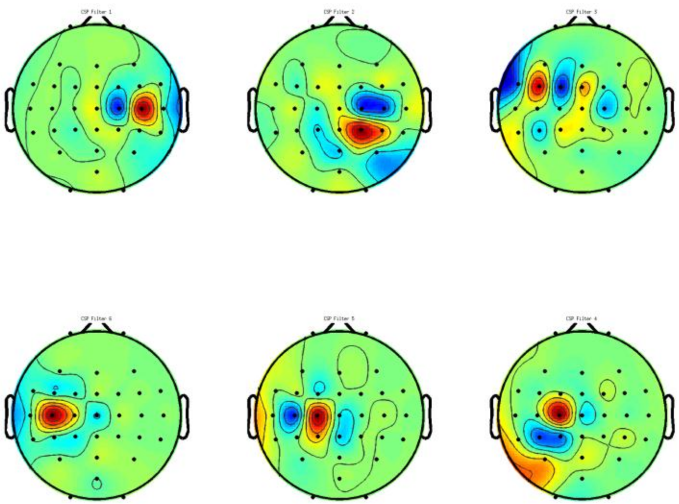
3.



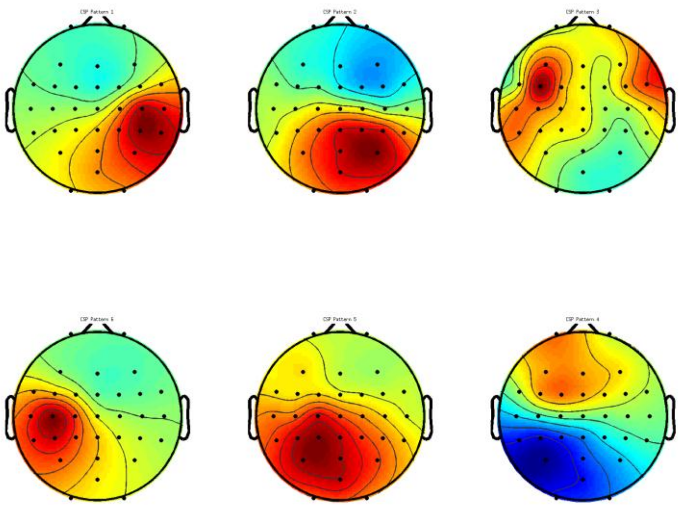
Dornhege, 2004

Resulting Spatial Filters

- Produces well-adapted filters (left) and occasionally roughly dipolar filter inverses (right)
- Note that typically only filters for the k top and k bottom eigenvalues are retained



W



W^{-1}

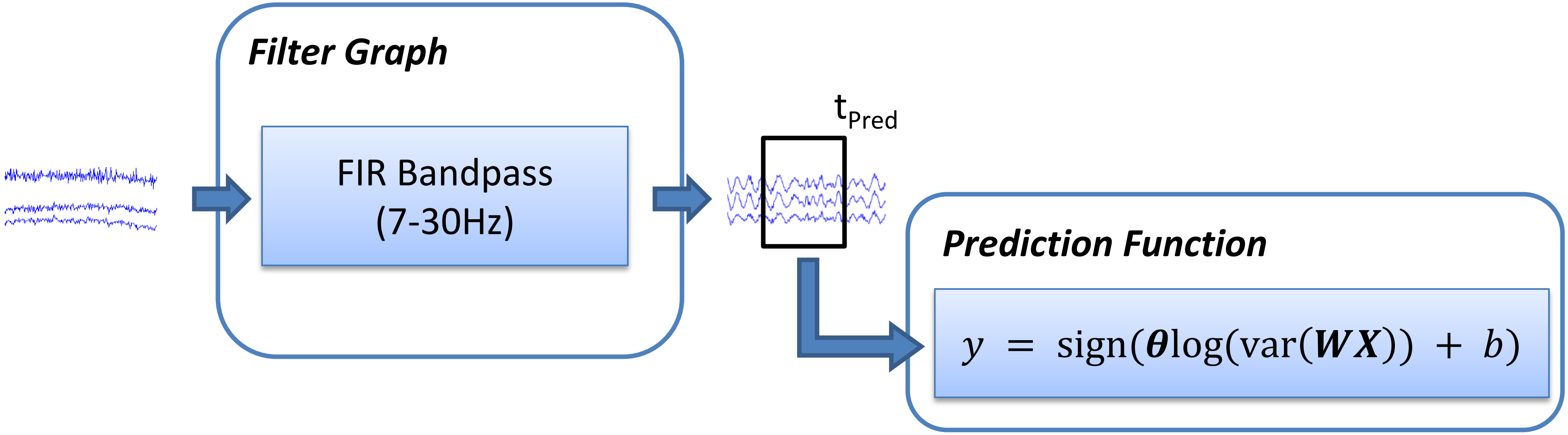
CSP Prediction Function

- The CSP Prediction function amounts to:
  - Spatial filtering
  - Log-variance calculation
  - Application of a linear (or non-linear) classifier

$$y = \mathrm{sign}(\boldsymbol{\theta} \log(\mathrm{var}(\boldsymbol{W}\boldsymbol{X})) + b)$$

Putting it all Together

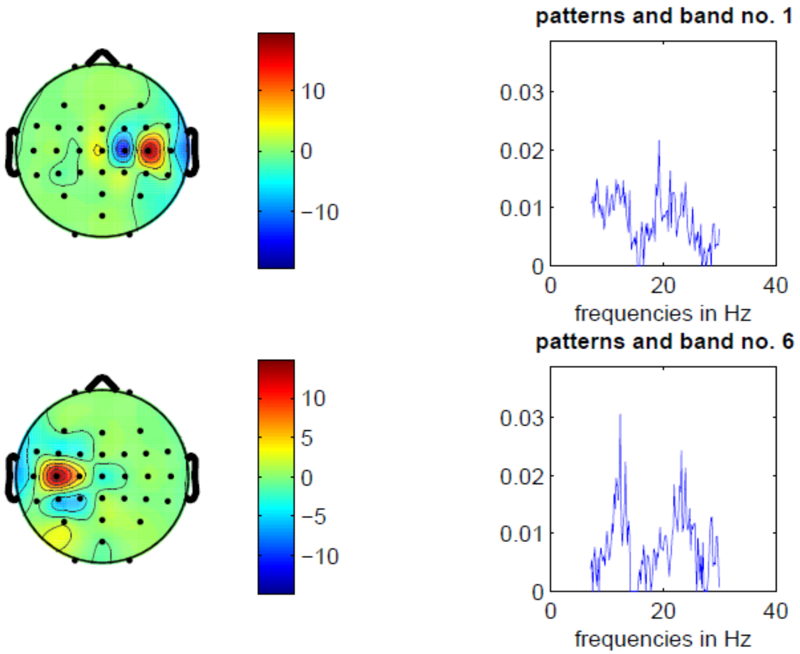
- A CSP-based BCI typically operates on a band-pass filtered signal
- Choice of the frequency band is not trivial
- The online window length does not need to correspond to the training window length



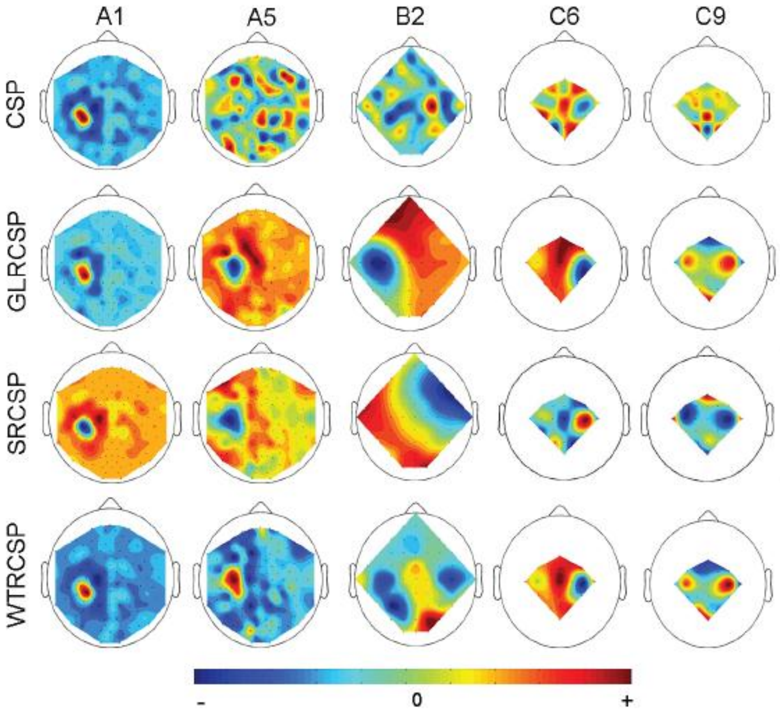
Alternatives to CSP

- Dozens of extensions (Spec-CSP, FBCSP, TRCSP, …)

Spectrally Weighted CSP (Spec-CSP)



Some Regularized CSP Variants



Lotte et al., 2010

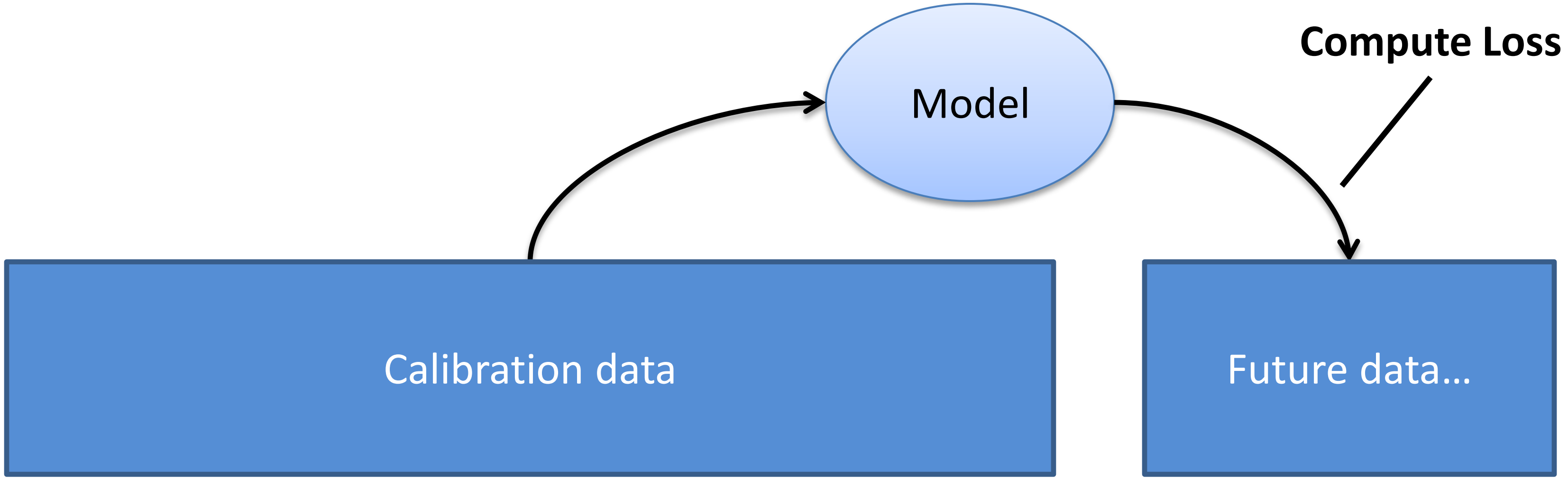
Alternatives to CSP

- Other ways to calculate spatial filters: ICA, Beamforming, Stationary Subspace Analysis, Dictionary Learning, …
- “Second-order trick”: using a linear classifier applied to the *covariance matrix* of the data epoch, but requires *large-scale machine learning methods*
- Some types of neural networks / graphical models

6 Evaluating Results

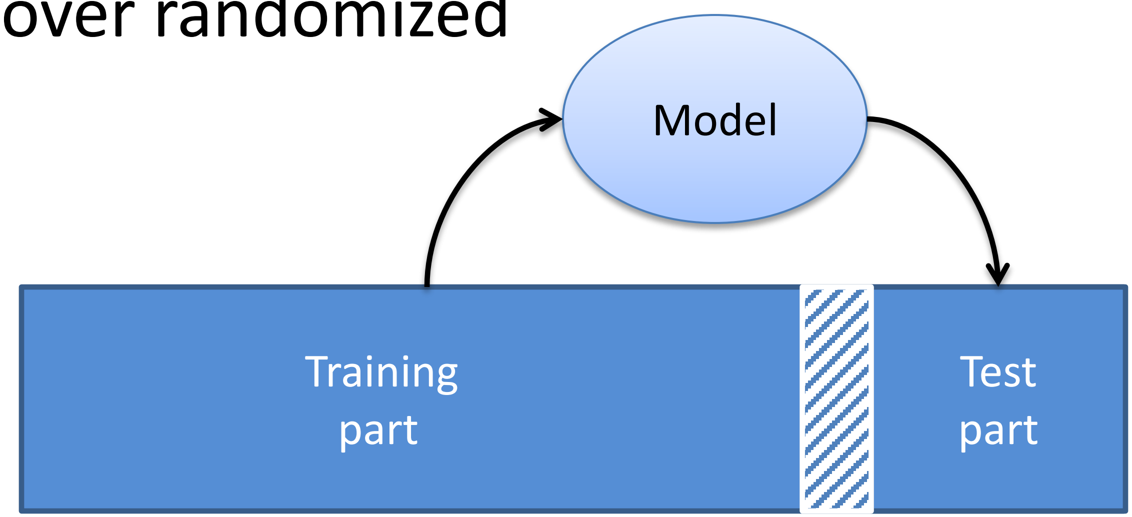
Overall Evaluation Strategies

- **When given calibration data and test data…**
- Estimate model parameters (spatial filters, statistics)
- Apply the model to new data (online / single-trial)
- Measure prediction performance or loss (e.g., mis-classification rate or mean-square error)



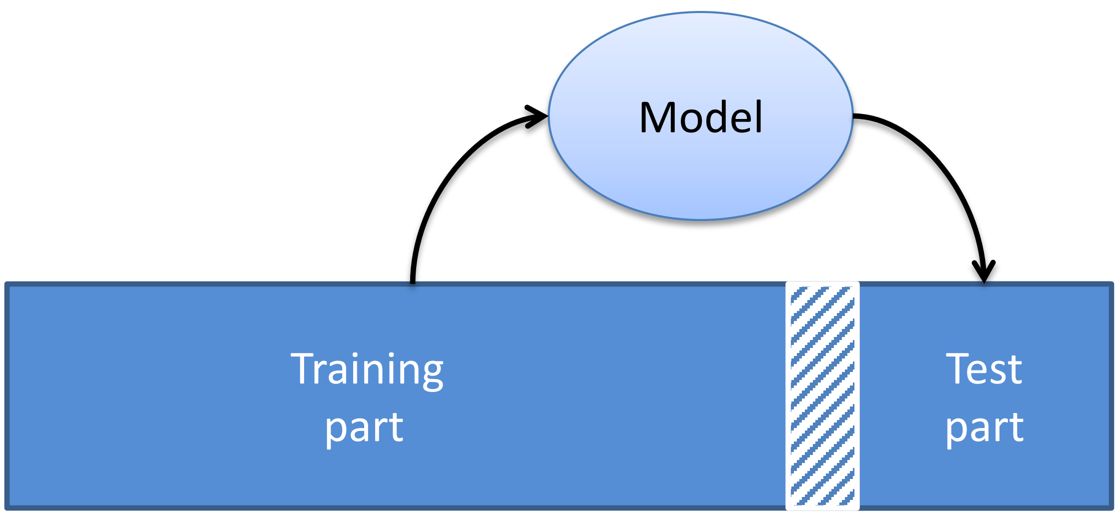
Overall Evaluation Strategies

- **What if there is no second data set?**
- split *one data set* repeatedly into training/test blocks systematically, a.k.a. *cross-validation*
- Each trial is used for testing once
- Time series data: Prefer block-wise cross-validation over randomized



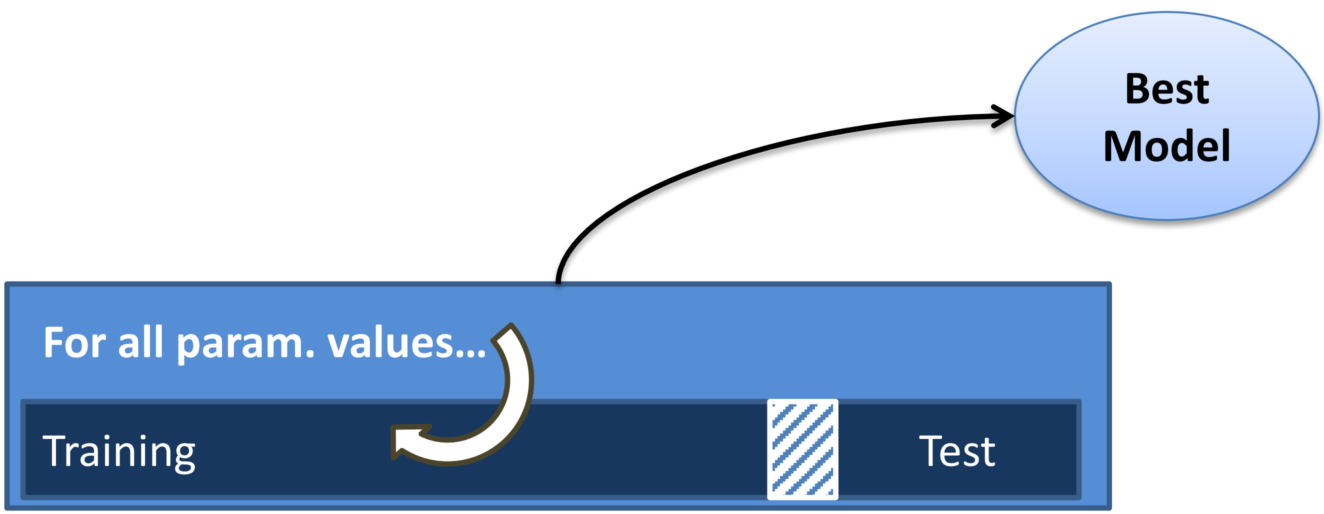
Overall Evaluation Strategies

- **Consideration:** Since neighboring trials are more closely related than training and future online data, *leave a margin of several trials/seconds between training and test*
- Standard splitting schemes: 5x, 10x



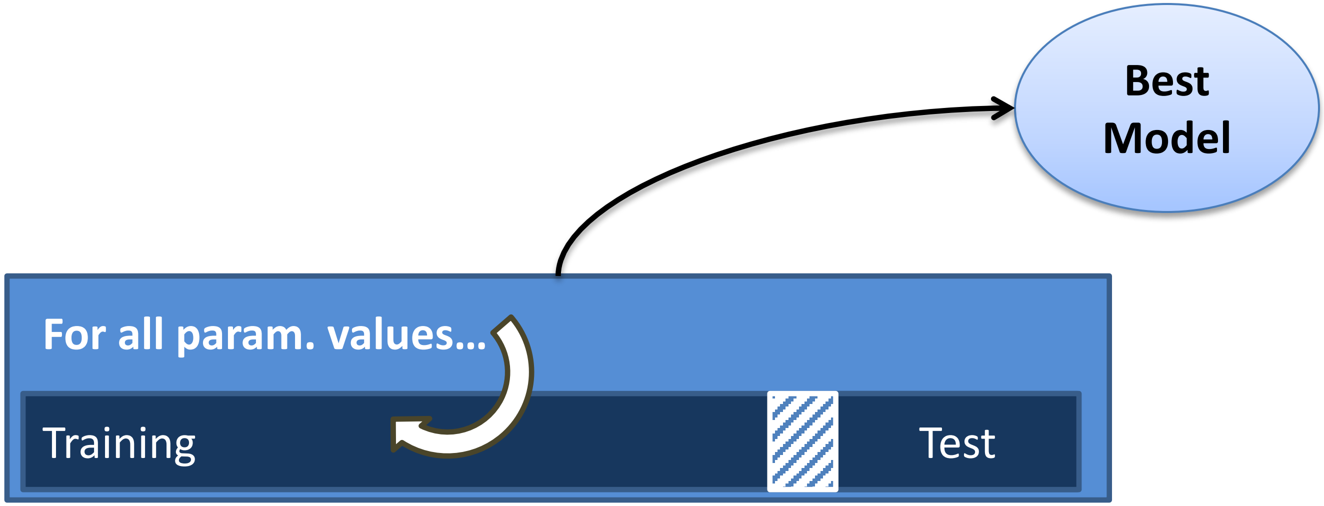
Overall Evaluation Strategies

- **Parameter search** can be done using cross-validation in a grid search (try all values of free parameters)
- Quite general (e.g. can search for best method)



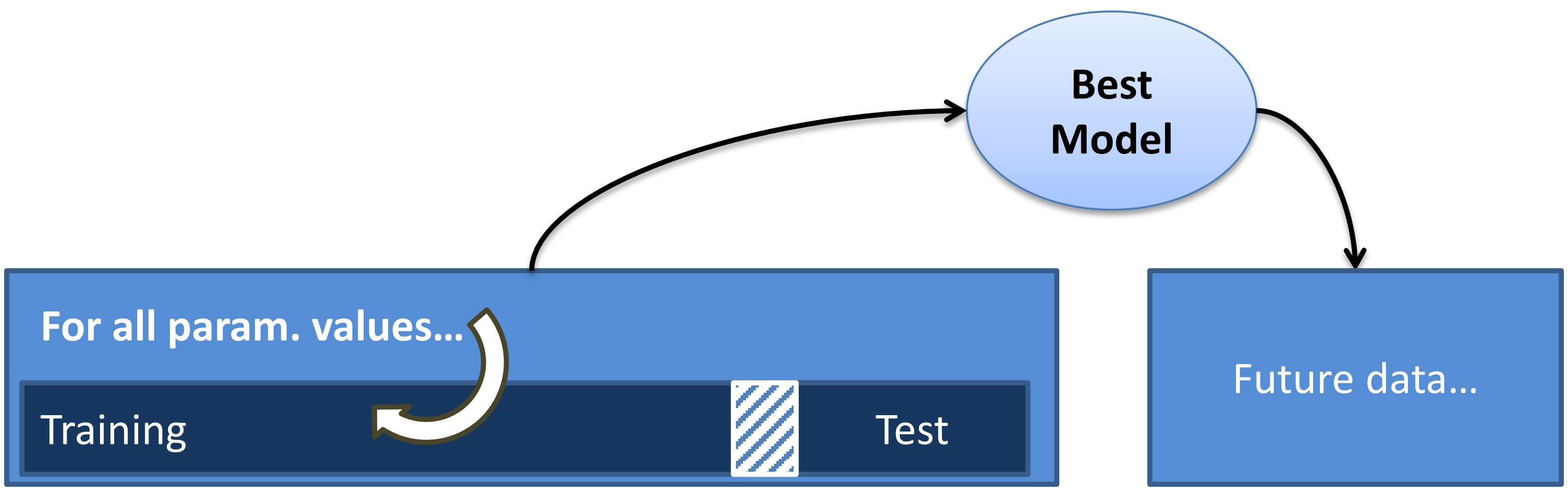
Overall Evaluation Strategies

- **Parameter search** can be done using cross-validation in a grid search (try all values of free parameters)
- Quite general (e.g. can search for best method)
- **However:** Cannot directly report "best performance" estimates (=cherry-picked)



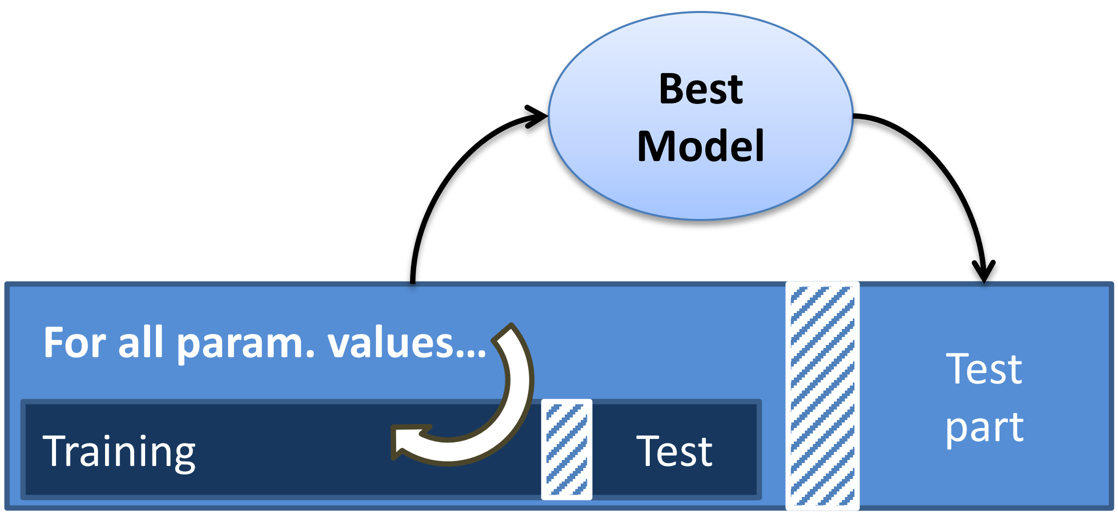
Overall Evaluation Strategies

- **Parameter search** can be done using cross-validation in a grid search (try all values of free parameters)
- Quite general (e.g. can search for best method)
- **However:** Cannot directly report "best performance" estimates (=cherry-picked), except on future data



Overall Evaluation Strategies

- **Parameter search** can be done using cross-validation in a grid search (try all values of free parameters)
- **Alternatively:** Parameter search can be nested *within* an outer cross-validation ("nested cross-validation")



7 Further Reading

BCI Papers Worth Reading

- B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Mueller, "Single-trial analysis and classification of ERP components - A tutorial", NeuroImage, vol. 56, no. 2, pp. 814–825, May 2011.
- F. Lotte and C. Guan, “Regularizing common spatial patterns to improve BCI designs: unified theory and new algorithms,” IEEE Transactions on Biomedical Engineering, vol. 58, no. 2, pp. 355-362, Feb. 2011.
- R. Tomioka and K.-R. Mueller, A regularized discriminative framework for EEG analysis with application to brain-computer interface", NeuroImage, vol. 49, no. 1, pp. 415–432, 2010.
- B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Mueller, 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, Aug. 2007.
- M. Grosse-Wentrup, C. Liefhold, K. Gramann, and M. Buss, "Beamforming in noninvasive brain-computer interfaces", IEEE Trans. Biomed. Eng., vol. 56, no. 4, pp. 1209–1219, Apr. 2009.

BCI Surveys

- A. Bashashati, M. Fatourechi, R. K. Ward, and G. E. Birch, "A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals", J. Neural Eng., vol. 4, no. 2, pp. R32–R57, Jun. 2007.
- F. Lotte, M. Congedo, A. Lecuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces", J. Neural Eng., vol. 4, no. 2, pp. R1–R13, Jun. 2007.
- S. Makeig, C. Kothe, T. Mullen, N. Bigdely-Shamlo, Z. Zhang, K. Kreutz-Delgado, "Evolving Signal Processing for Brain–Computer Interfaces", Proc. IEEE, vol. 100, pp. 1567-1584, 2012.

Interesting Technical Papers

- D.P. Wipf and S. Nagarajan, “A Unified Bayesian Framework for MEG/EEG Source Imaging,” NeuroImage, vol. 44, no. 3, February 2009.
- S. Haufe, R. Tomioka, and G. Nolte, “Modeling sparse connectivity between underlying brain sources for EEG/MEG,” Biomedical Engineering, no. c, pp. 1-10, 2010.
- S. Boyd, N. Parikh, E. Chu, and J. Eckstein, “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers,” Information Systems Journal, vol. 3, no. 1, pp. 1-122, 2010.
- P. Zhao and B. Yu, “On Model Selection Consistency of Lasso,” Journal of Machine Learning Research, vol. 7 pp. 2541-2563, 2006.

Technical Papers, ct'd

- J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal Deep Learning," in Proceedings of the 28th International Conference on Machine Learning, 2011.
- K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," Nature, vol. 452, no. 7185, pp. 352-355, Mar. 2008.
- O. Jensen et al., "Using brain-computer interfaces and brain-state dependent stimulation as tools in cognitive neuroscience," Frontiers in Psychology, vol. 2, p. 100, 2011.
- D.-H. Kim, N. Lu, R. Ma,. Y.-S. Kim, R.-H. Kim, S. Wang, J. Wu, S. M. Won, H. Tao, A. Islam, K. J. Yu, T.-I. Kim, R. Chowdhury, M. Ying, L. Xu, M. Li, H.-J. Cung, H. Keum, M. McCormick, P. Liu, Y.-W. Zhang, F. G. Omenetto, Y Huang, T. Coleman, J. A. Rogers, "Epidermal electronics," Science vol. 333, no. 6044, 838-843, 2011.

Researchers to Watch

- Klaus-Robert Mueller et al. (TU Berlin) – one of the leading BCI groups http://www.bbci.de/publications.html
- Marcel van Gerven et al. (Donders) – BCI and Neuroscience with a Bayesian approach https://sites.google.com/a/distrep.org/distrep/publications
- Ryota Tomioka (U Tokyo) – known for some technical achievements http://www.ibis.t.u-tokyo.ac.jp/RyotaTomioka
- Karl Friston et al. (UC London) – working on relevant underpinnings for neuroimaging (outside BCI) http://www.fil.ion.ucl.ac.uk/Research/publications.html
- Leading Statisticians and Machine Learners: Michael I. Jordan, Andrew Ng, Lawrence Carin, Zoubin Ghahramani, Francis Bach, Geoffrey Hinton, Ruslan Salakhutdinov, Yeh Whye Teh, David Blei, …

Extended version of this lecture:

See Additional Materials

Thanks!

Questions?