Machine-Learning for Brain-Computer Interfaces

Jeremy Hill

Max Planck Institute
for Biological Cybernetics

Tübingen, Germany

MAX-PLANCK-GESELLSCHAFT

BIOLOGISCHE KYBERNETIK

BCI as a Potential Assistive Technology

- Complete paralysis (e.g. late-stage Amyotrophic Lateral Sclerosis)

BCI as a Potential Assistive Technology

- Complete paralysis (e.g. late-stage Amyotrophic Lateral Sclerosis)
  - Communication

BCI as a Potential Assistive Technology

- Complete paralysis (e.g. late-stage Amyotrophic Lateral Sclerosis)
  - Communication
- Disconnection of motor pathways (e.g. subcortical stroke, amputation)

BCI as a Potential Assistive Technology

- Complete paralysis (e.g. late-stage Amyotrophic Lateral Sclerosis)
  - Communication
- Disconnection of motor pathways (e.g. subcortical stroke, amputation)
  - Rehabilitation of movement

BCI as a Potential Assistive Technology

- Complete paralysis (e.g. late-stage Amyotrophic Lateral Sclerosis)
  - Communication
- Disconnection of motor pathways (e.g. subcortical stroke, amputation)
  - Rehabilitation of movement
  - Relief of phantom-limb pain

BCI as a Potential Assistive Technology

- Complete paralysis (e.g. late-stage Amyotrophic Lateral Sclerosis)
  - Communication
- Disconnection of motor pathways (e.g. subcortical stroke, amputation)
  - Rehabilitation of movement
  - Relief of phantom-limb pain
  - Control of prosthetics or FES

BCI as a Potential Assistive Technology

- Complete paralysis (e.g. late-stage Amyotrophic Lateral Sclerosis)
  - Communication
- Disconnection of motor pathways (e.g. subcortical stroke, amputation)
  - Rehabilitation of movement
  - Relief of phantom-limb pain
  - Control of prosthetics or FES
- Other...

Problems with Clinical Deployment

- interruptions

Problems with Clinical Deployment

- interruptions
- fatigue, pain, drugs

Problems with Clinical Deployment

- interruptions
- fatigue, pain, drugs
- noisy, non-"standard" and non-stationary EEG

Problems with Clinical Deployment

- interruptions
- fatigue, pain, drugs
- noisy, non-"standard" and non-stationary EEG
  - slower ERP responses, more low-frequency dominance

Problems with Clinical Deployment

- interruptions
- fatigue, pain, drugs
- noisy, non-"standard" and non-stationary EEG
  - slower ERP responses, more low-frequency dominance
  - blood-sugar- and fatigue-dependent changes

Problems with Clinical Deployment

- “good-day-bad-day” syndrome: any exploration of induction parameters requires an alternating or mixed design, halving the amount of data in any one experimental condition on any one day

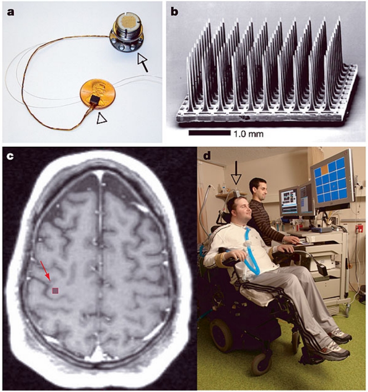
Problems with Clinical Deployment

- "good-day-bad-day" syndrome: any exploration of induction parameters requires an alternating or mixed design, halving the amount of data in any one experimental condition on any one day
- data set sizes are small to start with

Problems with Clinical Deployment

- "good-day-bad-day" syndrome: any exploration of induction parameters requires an alternating or mixed design, halving the amount of data in any one experimental condition on any one day
- data set sizes are small to start with
- more frequent session-to-session transfer problems

Measurement systems for BCI

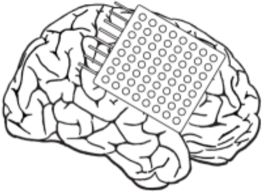


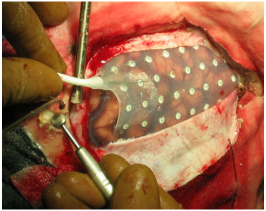
Implanted microelectrode array (Cyberkinetics, Inc)

Figure from Hochberg et al. Nature, July 2006.

Measurement systems for BCI

Department of Epileptology,
University of Bonn, 2004





Electrocorticography (ECoG)

Measurement systems for BCI

Electroencephalography (EEG)

Measurement systems for BCI

Near Infra-Red Spectrophotometry (NIRS)

Measurement systems for BCI

Magnetoencephalography (MEG)

Functional Magnetic Resonance Imaging (fMRI)

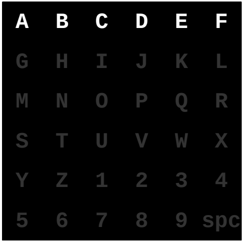
Induction

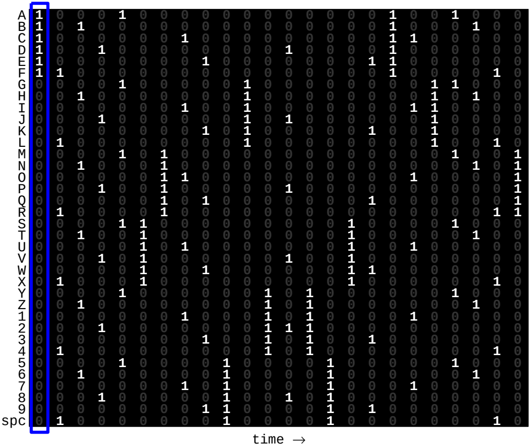
- Attention (overt and/or covert) to one of a number of stimuli

Induction

- Attention (overt and/or covert) to one of a number of stimuli
  - Most common example: visual grid speller (Farwell & Donchin 1988)

Induction





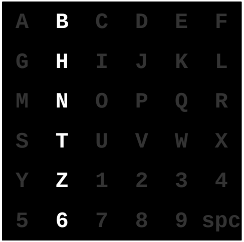
Induction

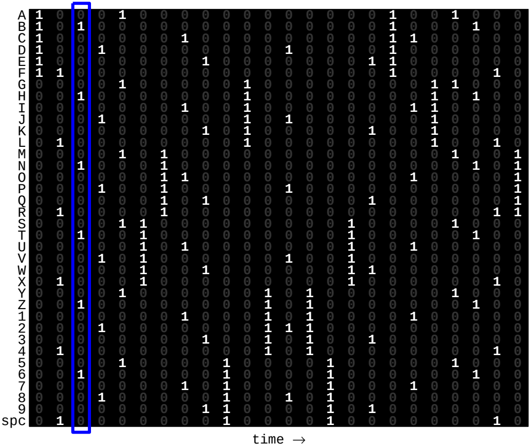
A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	spc

A	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
B	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
C	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
D	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
E	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
F	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
G	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
H	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0
I	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
J	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0
K	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0
L	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0
M	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
N	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
O	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
P	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Q	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
R	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
S	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
T	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
U	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
V	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
W	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
X	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
Y	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0
Z	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0
5	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0
6	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0
7	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0
8	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0
spc	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0

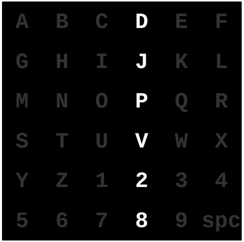
time →

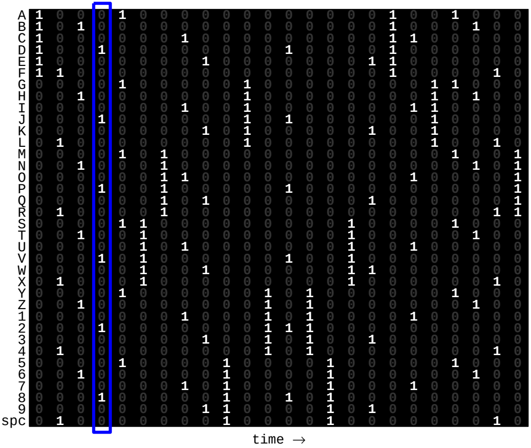
Induction


A B C D E F
G H I J K L
M N O P Q R
S T U V W X
Y Z 1 2 3 4
5 6 7 8 9 spc


A B C D E F G H I J K L M N O P Q R S T U V W X Y Z 1 2 3 4 5 6 7 8 9 spc
time →

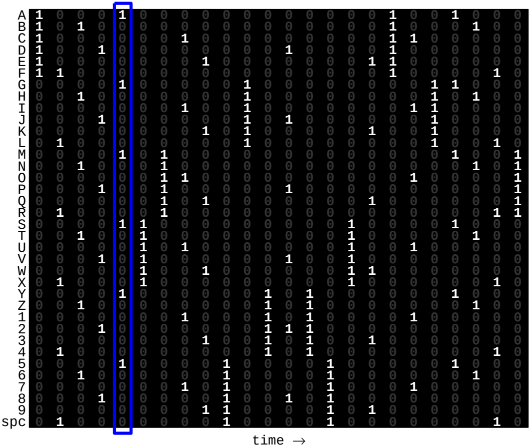
Induction


A B C D E F
G H I J K L
M N O P Q R
S T U V W X
Y Z 1 2 3 4
5 6 7 8 9 spc

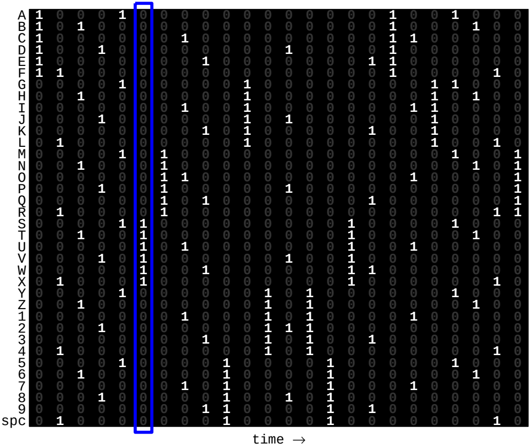

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z 1 2 3 4 5 6 7 8 9 spc
time →

Induction


A B C D E F
G H I J K L
M N O P Q R
S T U V W X
Y Z 1 2 3 4
5 6 7 8 9 spc


A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
P
Q
R
S
T
U
V
W
X
Y
Z
1
2
3
4
5
6
7
8
9
spc
time →

Induction



Induction

- Attention (overt and/or covert) to one of a number of stimuli
  - Most common example: visual grid speller (Farwell & Donchin 1988)
  - BUT: for *completely* paralysed users, vision deteriorates.

Induction

- Attention (overt and/or covert) to one of a number of stimuli
  - Most common example: visual grid speller (Farwell & Donchin 1988)
  - BUT: for *completely* paralysed users, vision deteriorates.
    $\rightsquigarrow$ incentive to design auditory-/tactile-based methods.

Induction

- Attention (overt and/or covert) to one of a number of stimuli
  - Most common example: visual grid speller (Farwell & Donchin 1988)
  - BUT: for *completely* paralysed users, vision deteriorates. $\rightsquigarrow$ incentive to design auditory-/tactile-based methods.
- "Mental tasks"

Induction

- Attention (overt and/or covert) to one of a number of stimuli
  - Most common example: visual grid speller (Farwell & Donchin 1988)
  - BUT: for *completely* paralysed users, vision deteriorates.
    $\rightsquigarrow$ incentive to design auditory-/tactile-based methods.
- "Mental tasks"
  - Most common example: imagined movement of hands or feet.

Induction

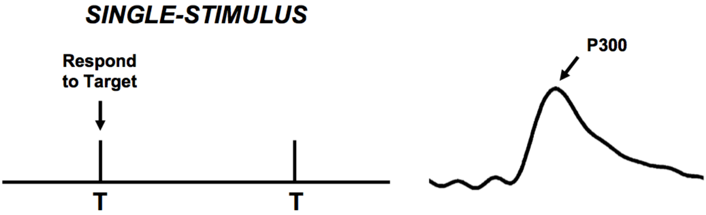
- Attention (overt and/or covert) to one of a number of stimuli
  - Most common example: visual grid speller (Farwell & Donchin 1988)
  - BUT: for *completely* paralysed users, vision deteriorates. $\rightsquigarrow$ incentive to design auditory-/tactile-based methods.
- "Mental tasks"
  - Most common example: imagined movement of hands or feet.
  - BUT: for users with motor-neuron disease, will the motor system continue functioning well enough long-term?

Induction

- Attention (overt and/or covert) to one of a number of stimuli
  - Most common example: visual grid speller (Farwell & Donchin 1988)
  - BUT: for *completely* paralysed users, vision deteriorates.
    $\rightsquigarrow$ incentive to design auditory-/tactile-based methods.
- "Mental tasks"
  - Most common example: imagined movement of hands or feet.
  - BUT: for users with motor-neuron disease, will the motor system continue functioning well enough long-term?
    $\rightsquigarrow$ incentive to explore non-motor mental tasks.

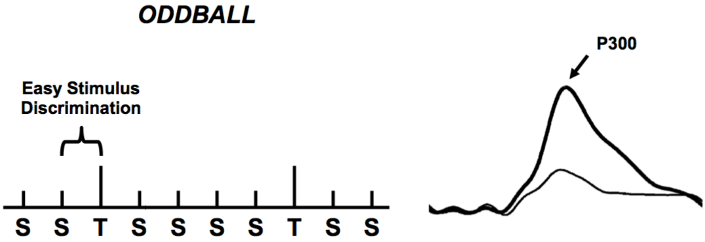
Event-Related Potentials

figures from Polich (2007)
Clinical Neurophysiology



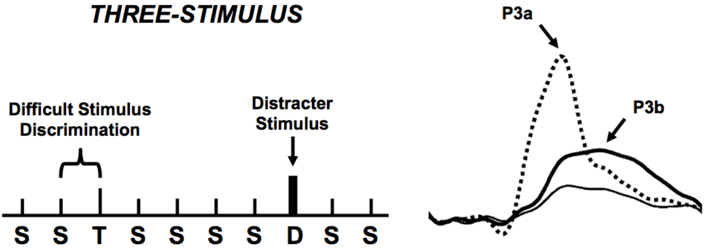
Event-Related Potentials

figures from Polich (2007)
Clinical Neurophysiology



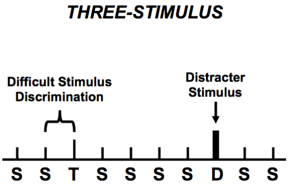
Event-Related Potentials

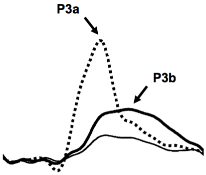
figures from Polich (2007)
Clinical Neurophysiology

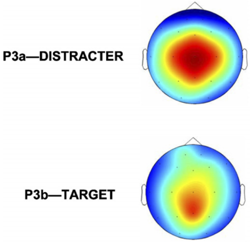


Event-Related Potentials

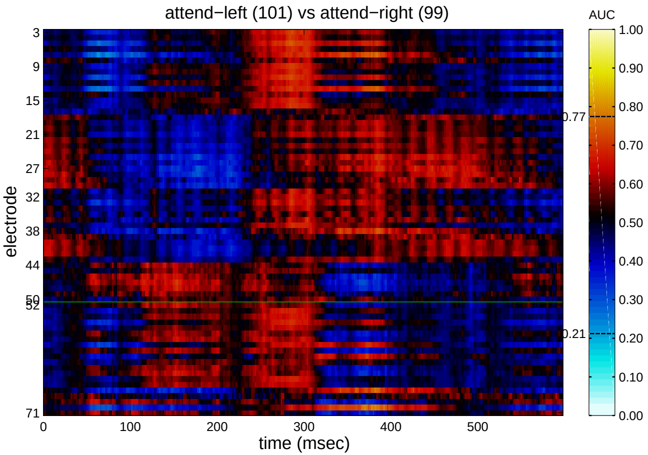
figures from Polich (2007)
Clinical Neurophysiology





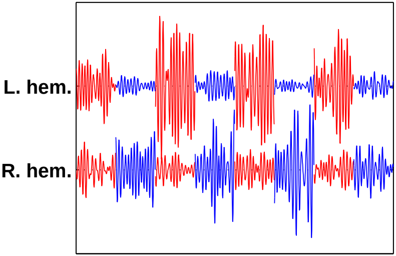


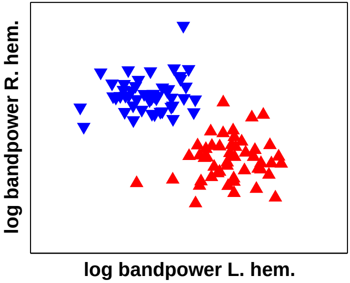
Event-Related Potentials



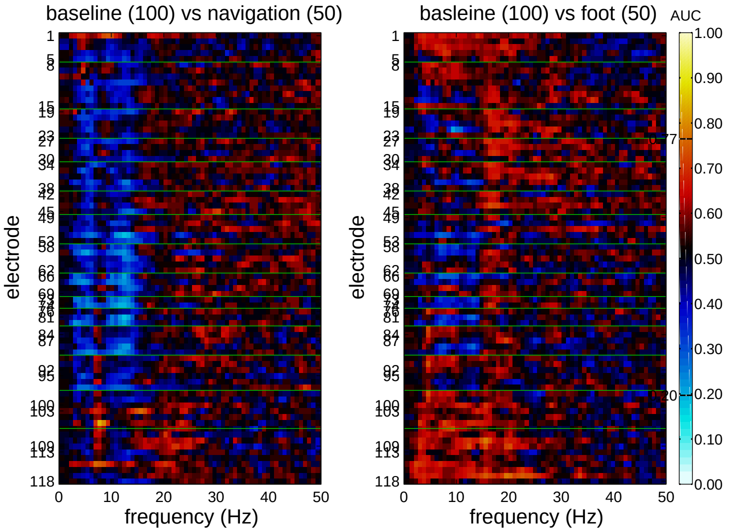
Bandpower

Event-Related Desynchronization in motor imagery: classify imagined left hand movement vs. imagined right hand movement based on power in (say) 10 Hz-band of estimated pre-motor cortex sources in the left and right hemispheres.



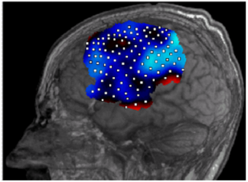


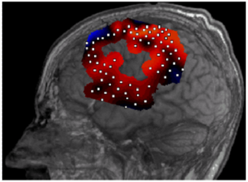
Bandpower



baseline (100) vs navigation (50)
basleine (100) vs foot (50)
AUC
electrode
frequency (Hz)
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00

Bandpower





An Overfitting Nightmare?

- High noise

An Overfitting Nightmare?

- High noise
- Small number of data exemplars

An Overfitting Nightmare?

- High noise
- Small number of data exemplars
- Very large number of features.

An Overfitting Nightmare?

- High noise
- Small number of data exemplars
- Very large number of features.
  Well actually, the features are usually *highly* correlated.

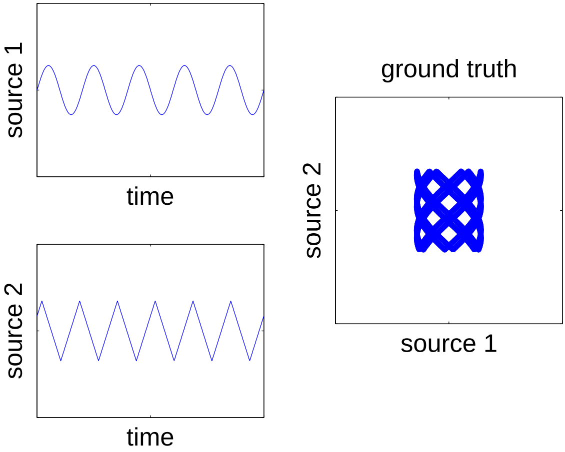
An Overfitting Nightmare?

- High noise
- Small number of data exemplars
- Very large number of features.
  Well actually, the features are usually *highly* correlated.
  - This is a good thing—we only need to worry about a low-dimensional *subspace*.

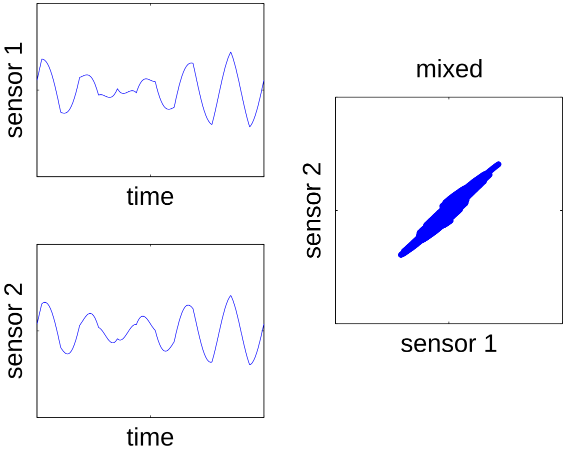
An Overfitting Nightmare?

- High noise
- Small number of data exemplars
- Very large number of features.
  Well actually, the features are usually *highly* correlated.
  - This is a good thing—we only need to worry about a low-dimensional *subspace*.
  - This is a bad thing—can lead to trying to optimize very "stiff" systems.

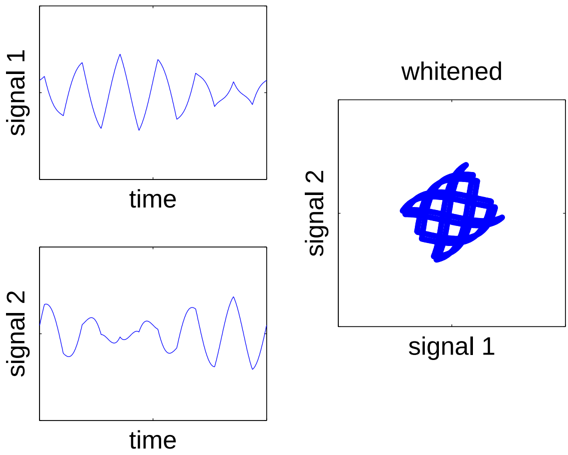
Source Separation



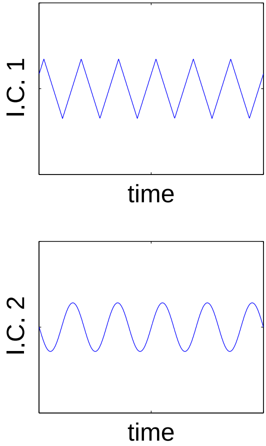
Source Separation

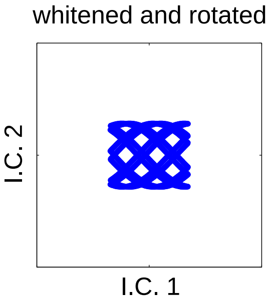


Source Separation

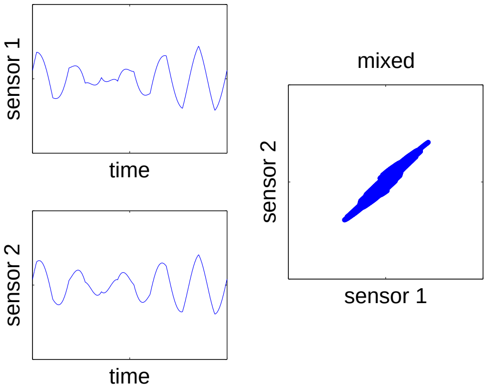


Source Separation

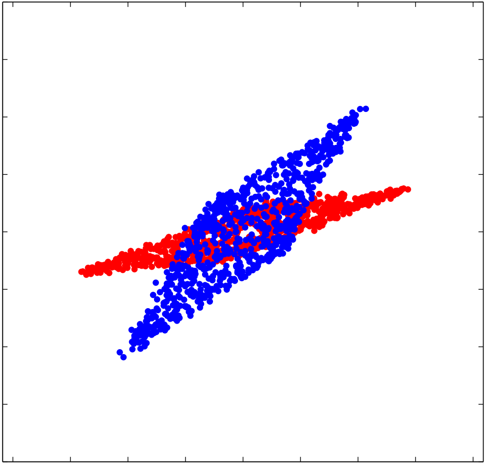




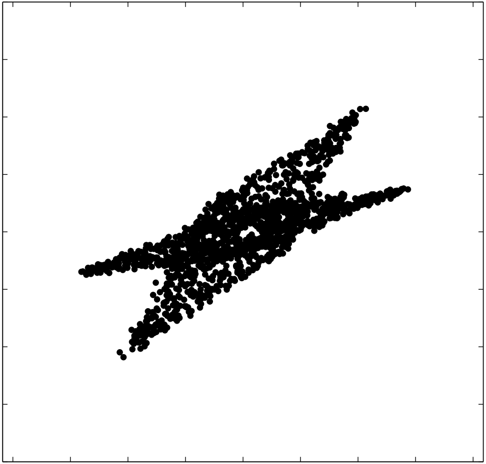
Cheap supervised rotation with CSP



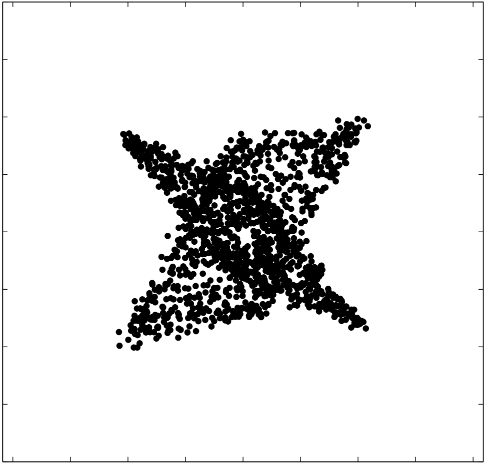
Cheap supervised rotation with CSP



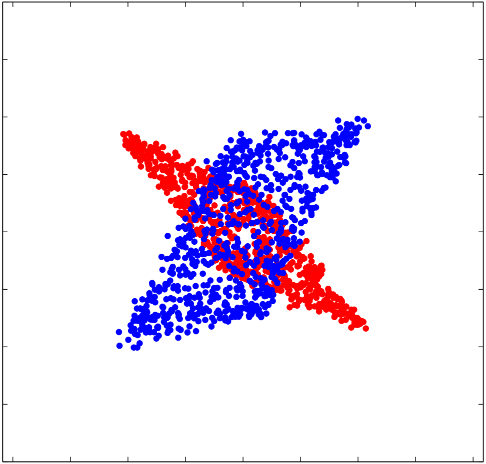
Cheap supervised rotation with CSP



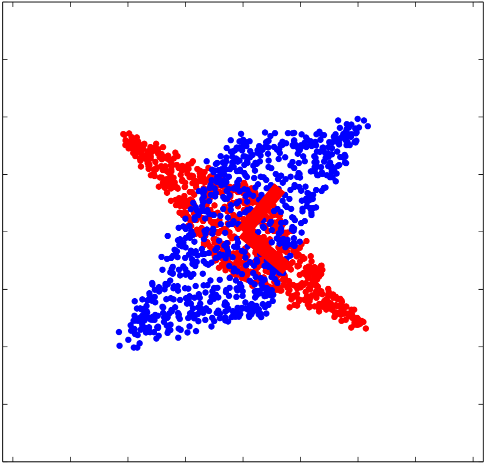
Cheap supervised rotation with CSP



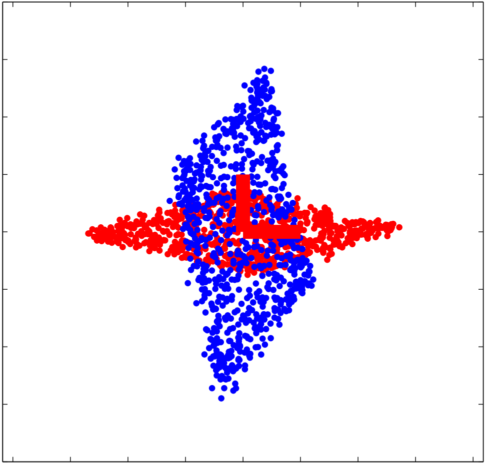
Cheap supervised rotation with CSP



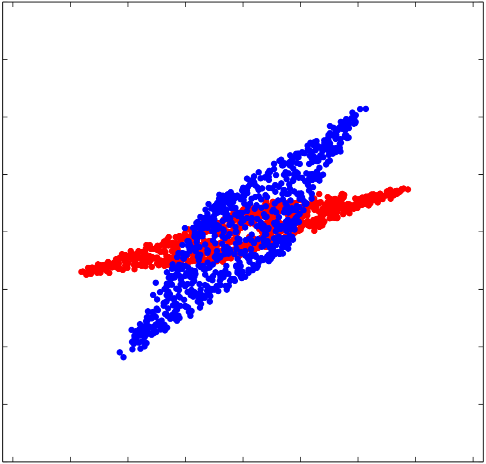
Cheap supervised rotation with CSP



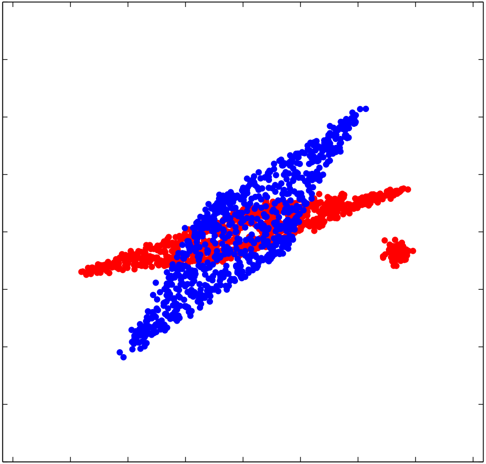
Cheap supervised rotation with CSP



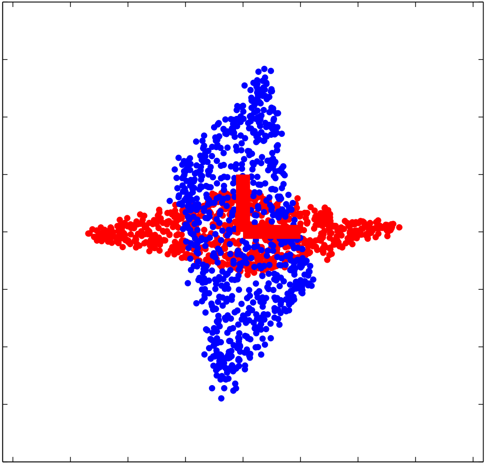
CSP: outlier- (artifact-) sensitivity



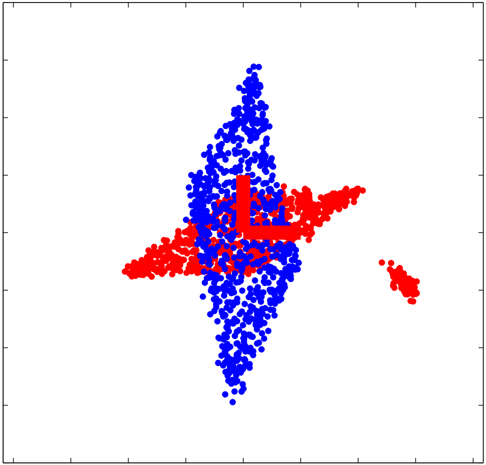
CSP: outlier- (artifact-) sensitivity



CSP: outlier- (artifact-) sensitivity



CSP: outlier- (artifact-) sensitivity



Slightly deeper learning?

From Collobert & Weston's NIPS 2009 tutorial:

Engineering: complex features, simple algorithm.

vs

Machine-Learning: simple input, implicitly learn the features.

Slightly deeper learning?

From Collobert & Weston's NIPS 2009 tutorial:

Engineering: complex features, simple algorithm.
Preprocessing (spatial subspace, spectral filtering...) then *classification*

vs

Machine-Learning: simple input, implicitly learn the features.

Slightly deeper learning?

From Collobert & Weston's NIPS 2009 tutorial:

Engineering: complex features, simple algorithm.

Preprocessing (spatial subspace, spectral filtering...) then *classification*

vs

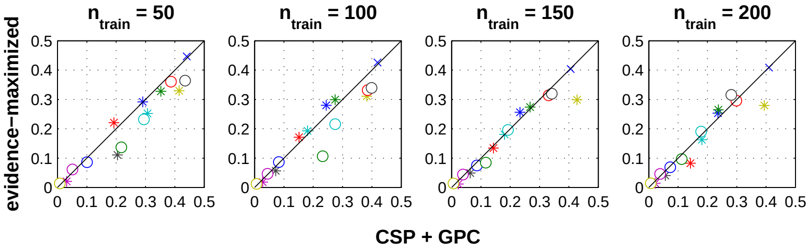
Machine-Learning: simple input, implicitly learn the features.

Idea: instead of performing CSP's least-square criterion to estimate discriminative sources

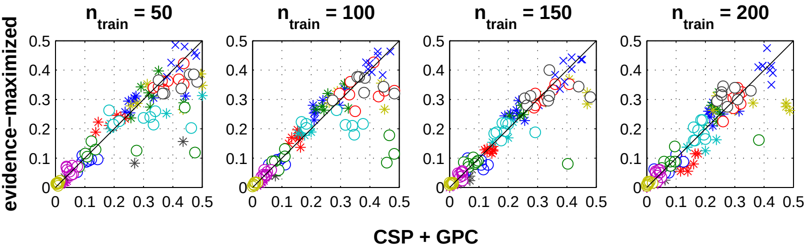
$$\mathrm{S} = \mathrm{W}\mathrm{X}$$

then classifying the resulting bandpower features diag $(\mathrm{S}\mathrm{S}^{\top})$ according to some *other* loss function, let's treat W as the hyperparameters of (e.g.) a Gaussian Process classifier and optimize them according to the marginal-likelihood...

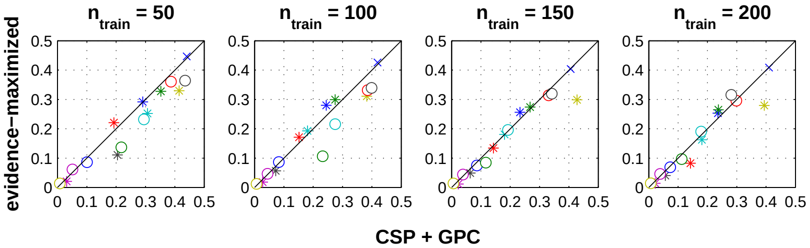
Slightly deeper learning?



Slightly deeper learning?



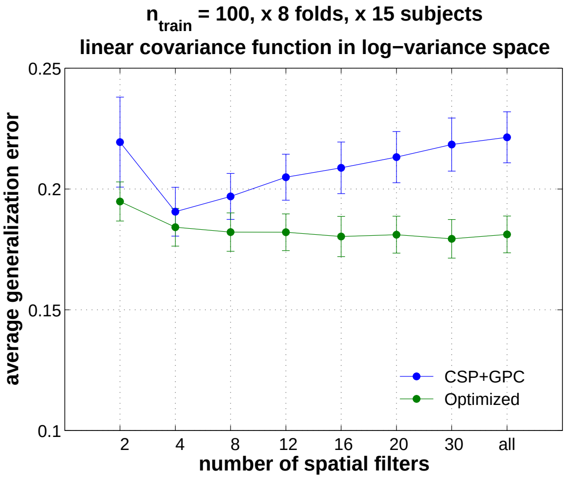
Slightly deeper learning?



Note:

- large individual variation
- particular benefits for smaller, noisier datasets.

Deeper learning ⇝ more “hands-free” operation



Deeper still?

Automatic combination of/selection between first- and second-order features

- Christoforou et al. (2008) JMLR

Deeper still?

Automatic combination of/selection between first- and second-order features

- Christoforou et al. (2008) JMLR
- Tomioka & Müller (2010) Neuroimage

Deeper still?

Automatic combination of/selection between first- and second-order features

- Christoforou et al. (2008) JMLR
- Tomioka & Müller (2010) Neuroimage

Convex optimization of spatial filters, with automatic selection/weighting between frequency bands

- Tomioka & Müller (2010) Neuroimage

Deeper still?

Automatic combination of/selection between first- and second-order features

- Christoforou et al. (2008) JMLR
- Tomioka & Müller (2010) Neuroimage

Convex optimization of spatial filters, with automatic selection/weighting between frequency bands

- Tomioka & Müller (2010) Neuroimage
- Farquhar (2009) Neural Networks

Deeper still?

Automatic combination of/selection between first- and second-order features

- Christoforou et al. (2008) JMLR
- Tomioka & Müller (2010) Neuroimage

Convex optimization of spatial filters, with automatic selection/weighting between frequency bands

- Tomioka & Müller (2010) Neuroimage
- Farquhar (2009) Neural Networks
  - extensible to arbitrary number of dimensions (time, frequency, cross-subject, cross-condition, ...)

Deeper still?

Automatic combination of/selection between first- and second-order features

- Christoforou et al. (2008) JMLR
- Tomioka & Müller (2010) Neuroimage

Convex optimization of spatial filters, with automatic selection/weighting between frequency bands

- Tomioka & Müller (2010) Neuroimage
- Farquhar (2009) Neural Networks
    - extensible to arbitrary number of dimensions (time, frequency, cross-subject, cross-condition, ...)

Pre-processing can still make a difference to performance (e.g. equalizing variance across frequency bands to compensate for $1/f$; spatial pre-whitening in both first- and second-order cases).

Deeper still?

Automatic combination of/selection between first- and second-order features

- Christoforou et al. (2008) JMLR
- Tomioka & Müller (2010) Neuroimage

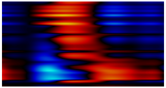
Convex optimization of spatial filters, with automatic selection/weighting between frequency bands

- Tomioka & Müller (2010) Neuroimage
- Farquhar (2009) Neural Networks
  - extensible to arbitrary number of dimensions (time, frequency, cross-subject, cross-condition, . . .)

Pre-processing can still make a difference to performance (e.g. equalizing variance across frequency bands to compensate for $1/f$; spatial pre-whitening in both first- and second-order cases).
Pre-processing the data can be seen as equivalent to changing the regularization environment. What is the "ideal" regularization strategy?

Low-rank Classification

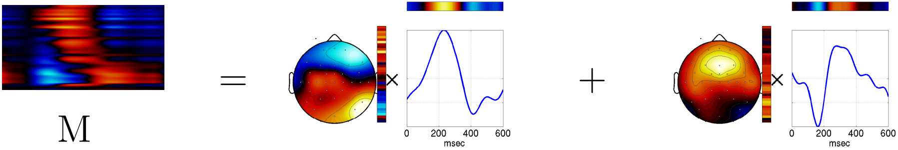
In linear ERP classification: classifier finds weights M for classifying space-$\times$-time "image" segments:



$$\mathrm{M} \quad =$$

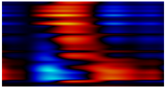
Low-rank Classification

In linear ERP classification: classifier finds weights M for classifying space-$\times$-time "image" segments:



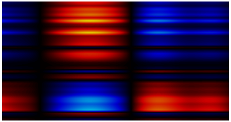
Low-rank Classification

In linear ERP classification: classifier finds weights M for classifying space-$\times$-time "image" segments:



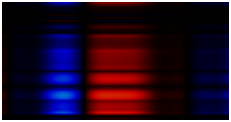
M

$=$



$\mathbf{w}_{s1}\ \mathbf{w}_{t1}^{\top}$

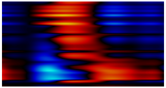
$+$



$\mathbf{w}_{s2}\ \mathbf{w}_{t2}^{\top}$

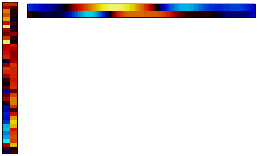
Low-rank Classification

In linear ERP classification: classifier finds weights M for classifying space-$\times$-time "image" segments:



M

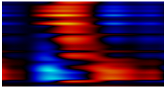
$=$



$\mathrm{W}_{\mathsf{s}} \; \mathrm{W}_{\mathsf{t}}^{\top}$

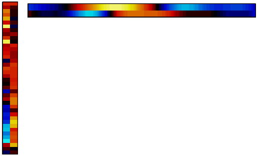
Low-rank Classification

In linear ERP classification: classifier finds weights M for classifying space-$\times$-time "image" segments:



M

$=$

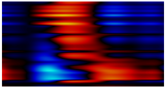


$\mathrm{W}_\mathsf{s} \;\; \mathrm{W}_\mathsf{t}^\top$

L_Σ regularization: regularize by putting an L-1 penalty on the singular values of M.

Low-rank Classification

In linear ERP classification: classifier finds weights M for classifying space-$\times$-time "image" segments:



M

$=$

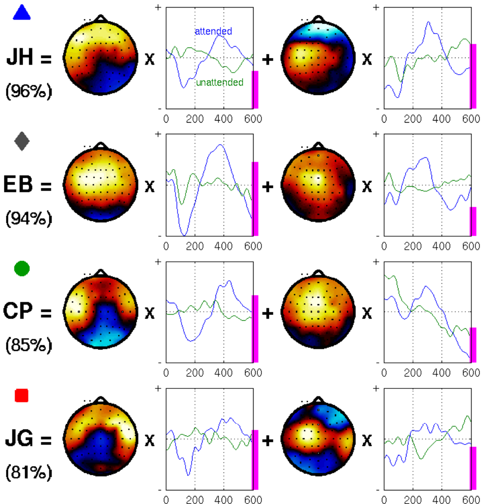
$\mathrm{W}_{\mathsf{s}}\ \mathrm{W}_{\mathsf{t}}^{\top}$

L_{Σ} regularization: regularize by putting an L-1 penalty on the singular values of M.

- Tomioka & Aihara (2007) ICML 2007.
- Tomioka & Müller (2010), Neuroimage.
- Farquhar (2009), Neural Networks.

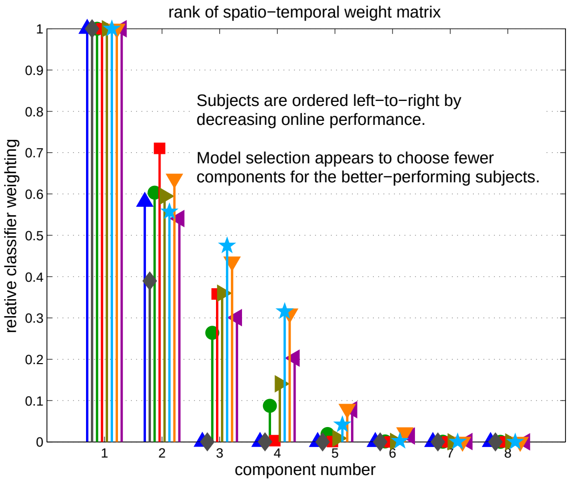
Example Sparsification Results

A BCI based on auditory stimuli (Hill et al., NIPS 2004 & BBCI Workshop 2009):



Example Sparsification Results

A BCI based on auditory stimuli (Hill et al., NIPS 2004 & BBCI Workshop 2009):



How Can Machine-Learners Help to Make BCI a Clinical Reality?

How Can Machine-Learners Help to Make BCI a Clinical Reality?

Moving towards "deeper" learning strategies

- improve performance on small/noisy datasets
- make systems run more "hands-free"

How Can Machine-Learners Help to Make BCI a Clinical Reality?

Moving towards "deeper" learning strategies

- improve performance on small/noisy datasets
- make systems run more "hands-free"

Use of L_Σ regularization (and its generalization to > 2 dimensions) to find the right subspace solutions.

How Can Machine-Learners Help to Make BCI a Clinical Reality?

Moving towards "deeper" learning strategies

- improve performance on small/noisy datasets
- make systems run more "hands-free"

Use of L_Σ regularization (and its generalization to > 2 dimensions) to find the right subspace solutions.

Incorporating prior knowledge/setting up the regularization environment in better ways.

How Can Machine-Learners Help to Make BCI a Clinical Reality?

Moving towards "deeper" learning strategies

- improve performance on small/noisy datasets
- make systems run more "hands-free"

Use of L_Σ regularization (and its generalization to > 2 dimensions) to find the right subspace solutions.

Incorporating prior knowledge/setting up the regularization environment in better ways.

Better transfer-learning and zero-training methods (e.g. see Fazli et al., this NIPS).

How Can Machine-Learners Help to Make BCI a Clinical Reality?

Moving towards "deeper" learning strategies

- improve performance on small/noisy datasets
- make systems run more "hands-free"

Use of L_{Σ} regularization (and its generalization to > 2 dimensions) to find the right subspace solutions.

Incorporating prior knowledge/setting up the regularization environment in better ways.

Better transfer-learning and zero-training methods (e.g. see Fazli et al., this NIPS).

Dealing with non-stationarities in brain data (see Klaus-Robert Müller's talk at this symposium, re SSA).

How Can Machine-Learners Help to Make BCI a Clinical Reality?

Moving towards "deeper" learning strategies

- improve performance on small/noisy datasets
- make systems run more "hands-free"

Use of L_{Σ} regularization (and its generalization to > 2 dimensions) to find the right subspace solutions.

Incorporating prior knowledge/setting up the regularization environment in better ways.

Better transfer-learning and zero-training methods (e.g. see Fazli et al., this NIPS).

Dealing with non-stationarities in brain data (see Klaus-Robert Müller's talk at this symposium, re SSA).

Finding ways of *encoding* information in more user- and brain-friendly ways (e.g. see Hill et al., last NIPS).