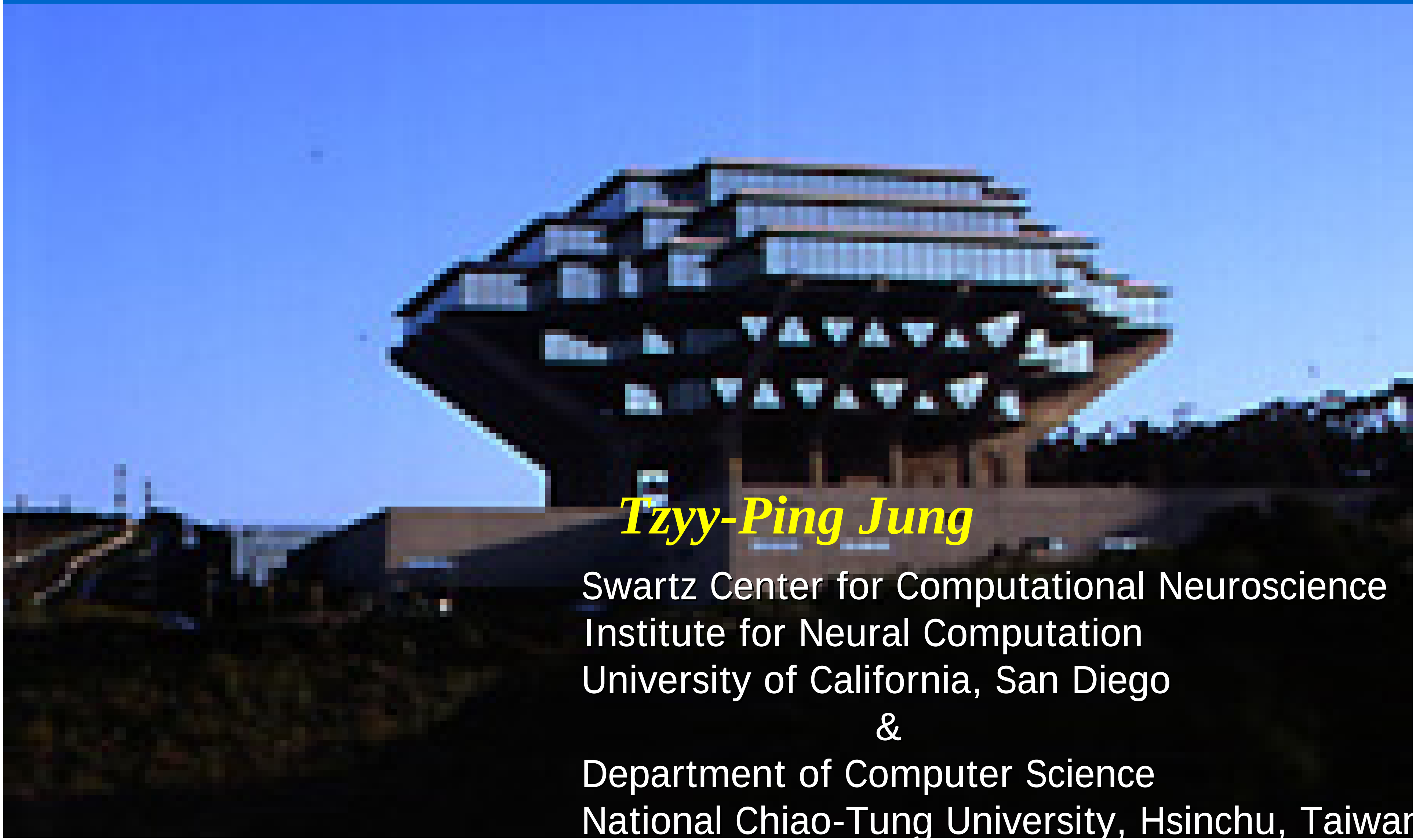
Independent Component Analysis and Its Applications



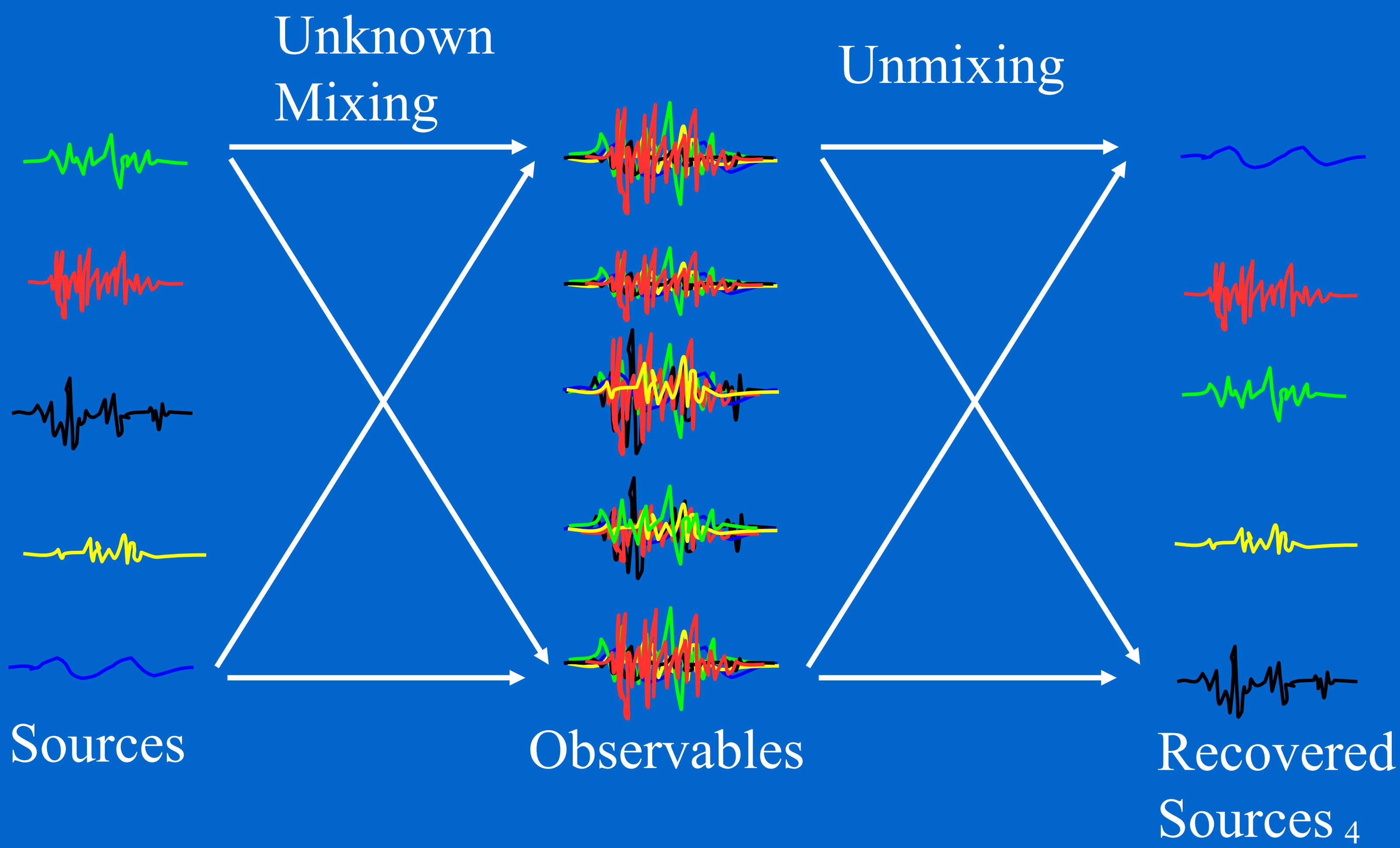
Tzyy-Ping Jung

Swartz Center for Computational Neuroscience
Institute for Neural Computation
University of California, San Diego
&
Department of Computer Science
National Chiao-Tung University, Hsinchu, Taiwan

Outline

- Blind Source Separation:
  – Solving the “cocktail party problem”
- Applications
  – Speech separation and clarity
  – Image processing
  – EEG/ERP
  – fMRI
  – other applications

Blind Source Separation



Historical Remarks

- Herault & Jutten ("Space or time adaptive signal processing by neural network models“, *Neural Nets for Computing Meeting*, Snowbird, Utah, 1986): Seminal paper, neural network
- Comon (1994): Approximation of MI by 4^{th} order statistics
- Bell & Sejnowski (1995): Information Maximization
- Amari et al. (1996): Natural Gradient Learning
- Cardoso (1996): JADE

- Applications of ICA to biomedical signals
  - EEG/ERP analysis (Makeig, Bell, Jung & Sejnowski, 1996.
  - fMRI analysis (McKeown, Jung et al. 1998)
  - ECG analysis (Cardoso 1998).

ICA Theory – Cost Functions

Family of BSS algorithms

- Information theory (Infomax)
- Bayesian probability theory (Maximum likelihood estimation)
- Negentropy maximization
- Nonlinear PCA
- Statistical signal processing (cumulant maximization, JADE)

A unifying Information-theoretic framework for ICA (Lee et al. 1999)

- Pearlmutter & Parra showed InfoMax, ML estimation are equivalent.
- Lee et al. showed negentropy has the equivalent property to InfoMax.
- Girolami & Fyfe showed nonlinear PCA can be viewed from information-theoretic principle.

Independent Component Analysis

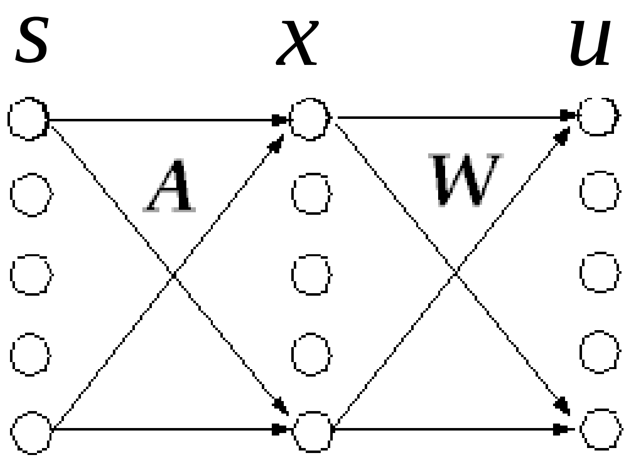
ICA is a method to recover a version, of the original sources by multiplying the data by a unmixing matrix,

$$\mathbf{u} = \mathbf{W}\mathbf{x},$$

where x is our observed signals, a linear mixtures of sources,

$$\mathbf{x} = \mathbf{A}\mathbf{s}.$$

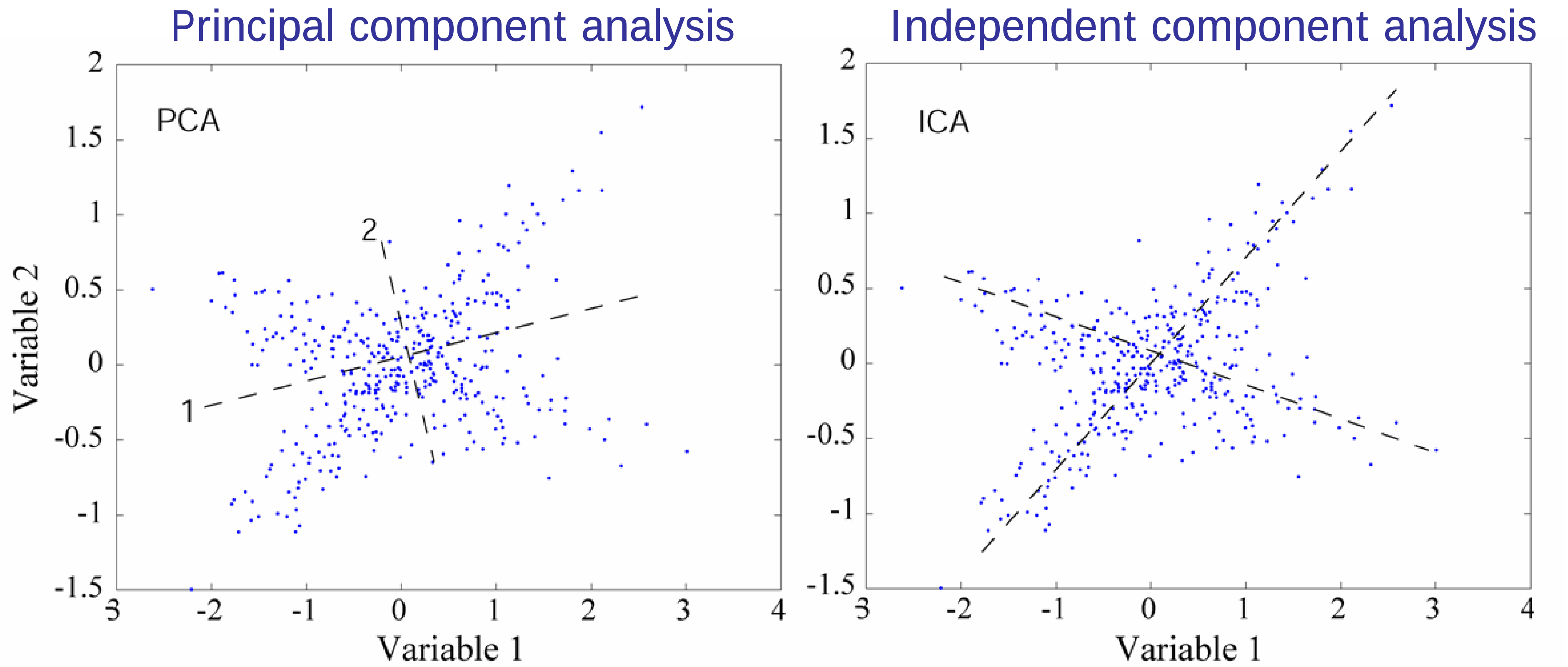
While PCA simply decorrelates the outputs (using an orthogonal matrix **W**), ICA attempts to make the outputs **statistically independent**, while placing no constraints on the matrix **W**



WA after learning:

$$\begin{bmatrix} \boxed{-4.09} & 0.13 & 0.09 & -0.07 & -0.01 \\ 0.07 & \boxed{-2.92} & 0.00 & 0.02 & -0.06 \\ 0.02 & -0.02 & -0.06 & -0.08 & \boxed{-2.20} \\ 0.02 & 0.03 & 0.00 & \boxed{1.97} & 0.02 \\ -0.07 & 0.14 & \boxed{-3.50} & -0.01 & 0.04 \end{bmatrix}$$

ICA vs PCA


Principal component analysis
Independent component analysis
PCA
ICA
1
2
Variable 1
Variable 2
2
1.5
1
0.5
0
-0.5
-1
-1.5
-3
-2
-1
0
1
2
3
4

Statistical Independence

Statistical Independence:

$$f_{\mathbf{s}}(\mathbf{s}) = \prod_{i=1}^{N} f_{s_i}(s_i)$$

Or the mutual information:

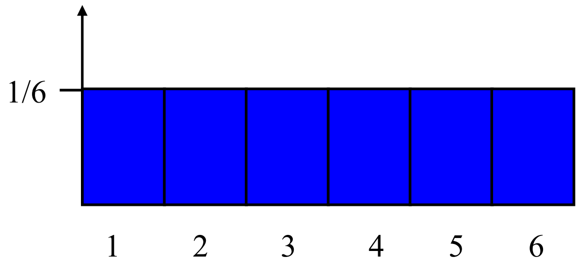
$$I(s_i, s_j) = E\left[\ln \frac{f_{\mathbf{s}}(\mathbf{s})}{\prod_{i=1}^{N} f_{s_i}(s_i)}\right] = 0, for\ \forall i \neq j$$

The problem of blind separation is to find $\mathbf{W}$ such that the linear transformation $\mathbf{u} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s}$ reestablishes the condition of statistical independence.

Entropy

$$H(X) = -\sum_{x \in X} p(x) \log(p(x))$$

Dice: 1/6



$$H = 6\left(-\frac{1}{6}\log_2\left(\frac{1}{6}\right)\right) = 2.58$$

ICA learning rule

How to make the outputs statistically independent?

Minimize their redundancy or mutual information.

Entropy: $$H(X) = -\sum_{x \in X} p(x)\log(p(x))$$

Joint entropy $$H(X,Y) = -\sum_{(x,y)\in X\times Y} p(x,y)\log(p(x,y))$$

Mutual Information $I(y_1,y_2) = H(y_1) + H(y_2) - H(y_1,y_2)$

Minimizing $I(y_1,y_2) \rightarrow$ **Maximizing $H(y_1,y_2)$**

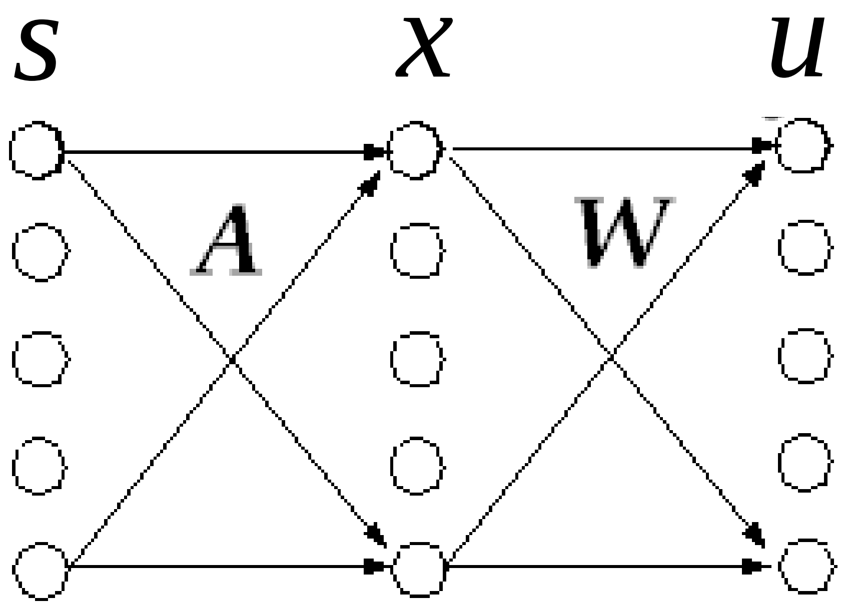
=0 if the two variables are independent

ICA learning rule

$$\Delta W = \frac{\partial H(y_1, y_2, \ldots)}{\partial W} \underbrace{W^T W}$$

Natural gradient (Amari)

Independent Component Analysis



ICA is a method to recover a version, of the original sources by multiplying the data by a unmixing matrix,

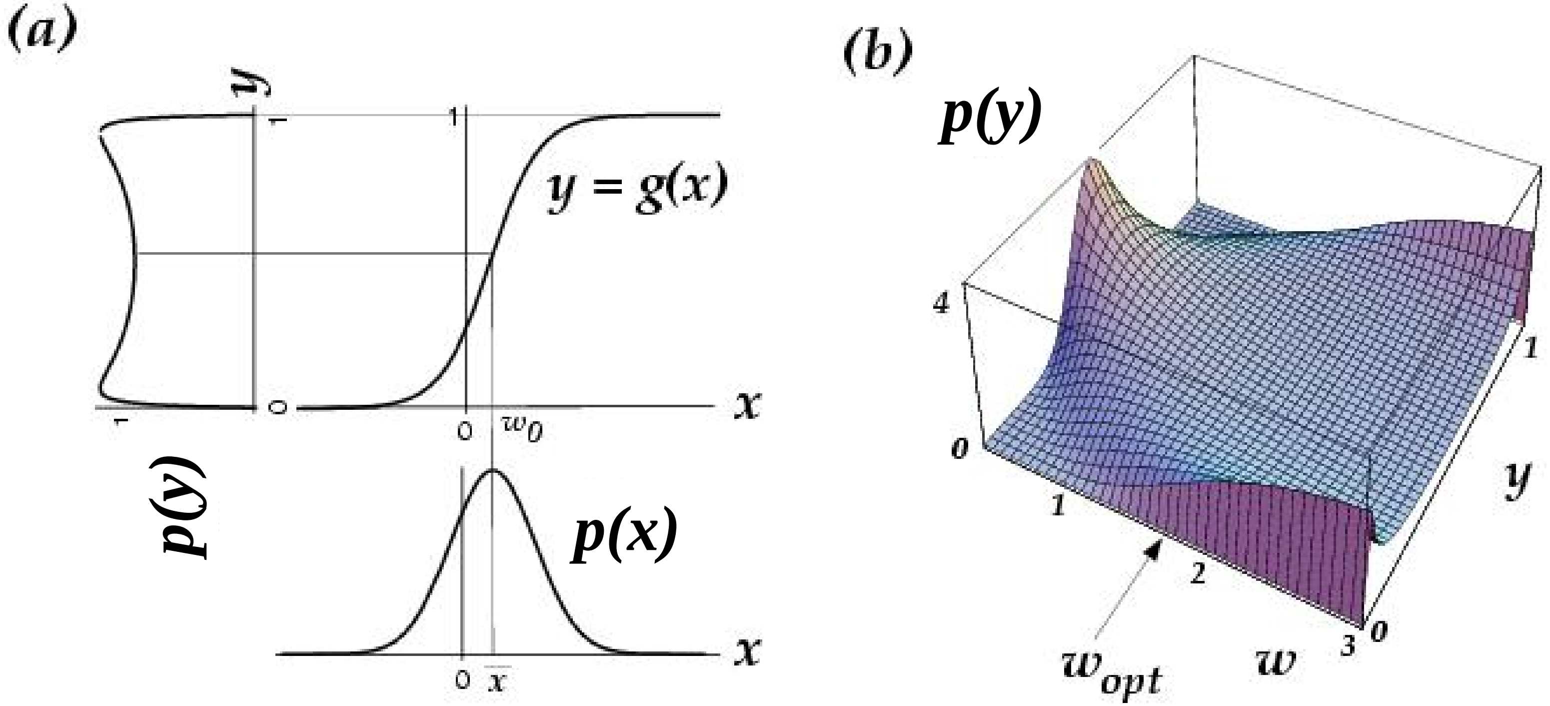
$$\mathbf{u} = \mathbf{W}\mathbf{x}$$

InfoMax (Bell & Sejnowski, 1995)

To make the u_i independent, we need to operate on non-linear transformed output variables, $\mathbf{y} = g(\mathbf{u})$, such as

$$\mathbf{y} = \frac{1}{1+e^{-\mathbf{u}}}, \quad \mathbf{u} = \mathbf{W}\mathbf{x} + \mathbf{w}_0$$

The non-linear function provides all the higher-order statistics necessary to establish independence.



From Bell & Sejnowski *Neural Compu. 1995.*

ICA learning rule

The learning rule:

$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} = \left[\mathbf{I} + \phi \mathbf{u}^T \right] \mathbf{W},$$

where $\phi_i = (\partial / \partial u_i) \ln(\partial y_i / \partial u_i)$.

For super-Gaussian,

$$\phi_i = 1 - 2y_i (for \quad logistic \quad nolinearity).$$

For sub- and/or super-Gaussian,

$$\phi_i = \begin{cases} +\tanh(u_i) - u_i & kurtosis < 0 \\ -\tanh(u_i) - u_i & kurtosis > 0 \end{cases}$$

- Remove the mean

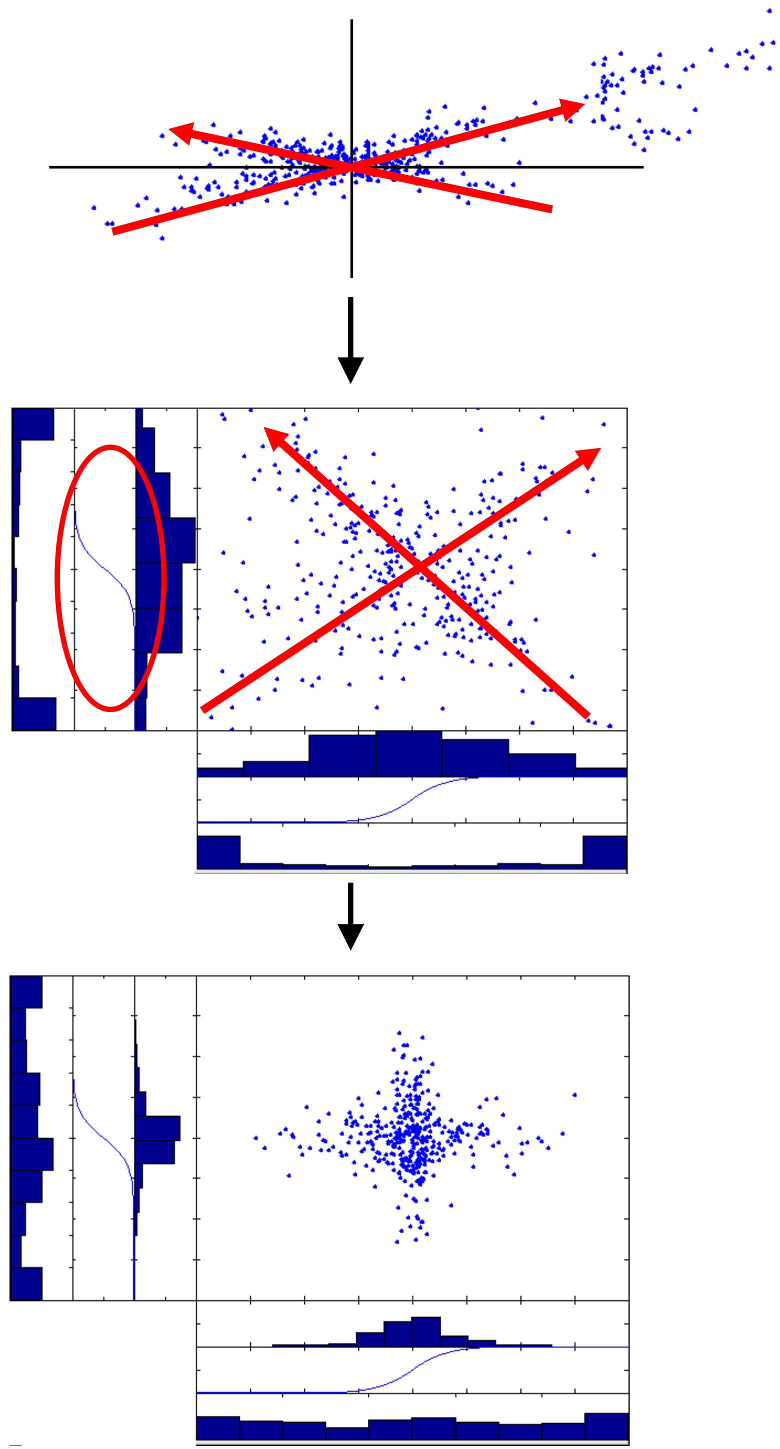
  $x = x - \langle x \rangle.$

- 'Sphere' the data by diagonalizing its covariance matrix,

  $x = 2\langle xx^T \rangle^{-1/2}(x - \langle x \rangle).$

- Update W according to

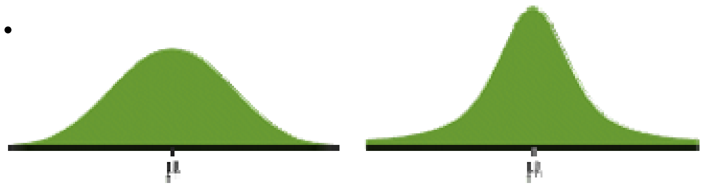
$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} = \left[\mathbf{I} + \phi \mathbf{u}^T\right] \mathbf{W}$$

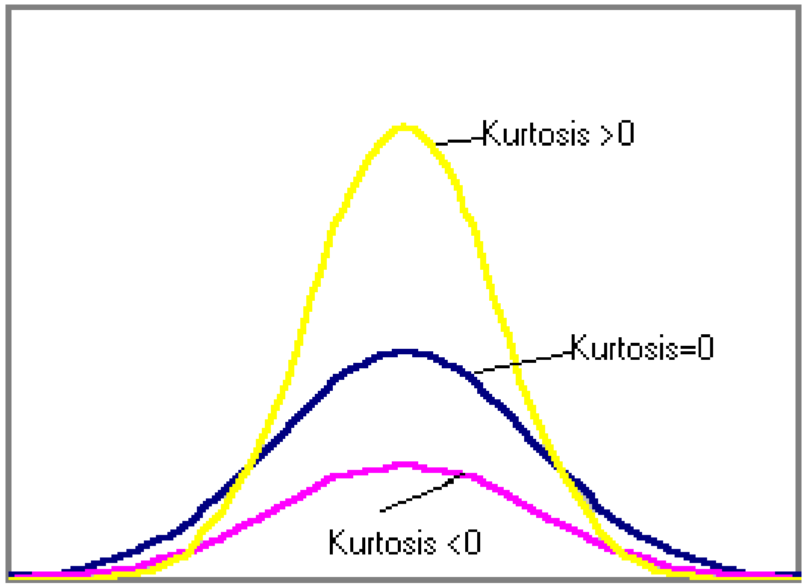


Kurtosis, Super- and Sub-Gaussian

Kurtosis: a measure of how peaked or flat of a probability distribution is.

$$kurt(X) = \frac{E\left[(X-\mu)^4\right]}{\sigma^4} - 3$$





Gaussian Dis. Kurtosis = 0

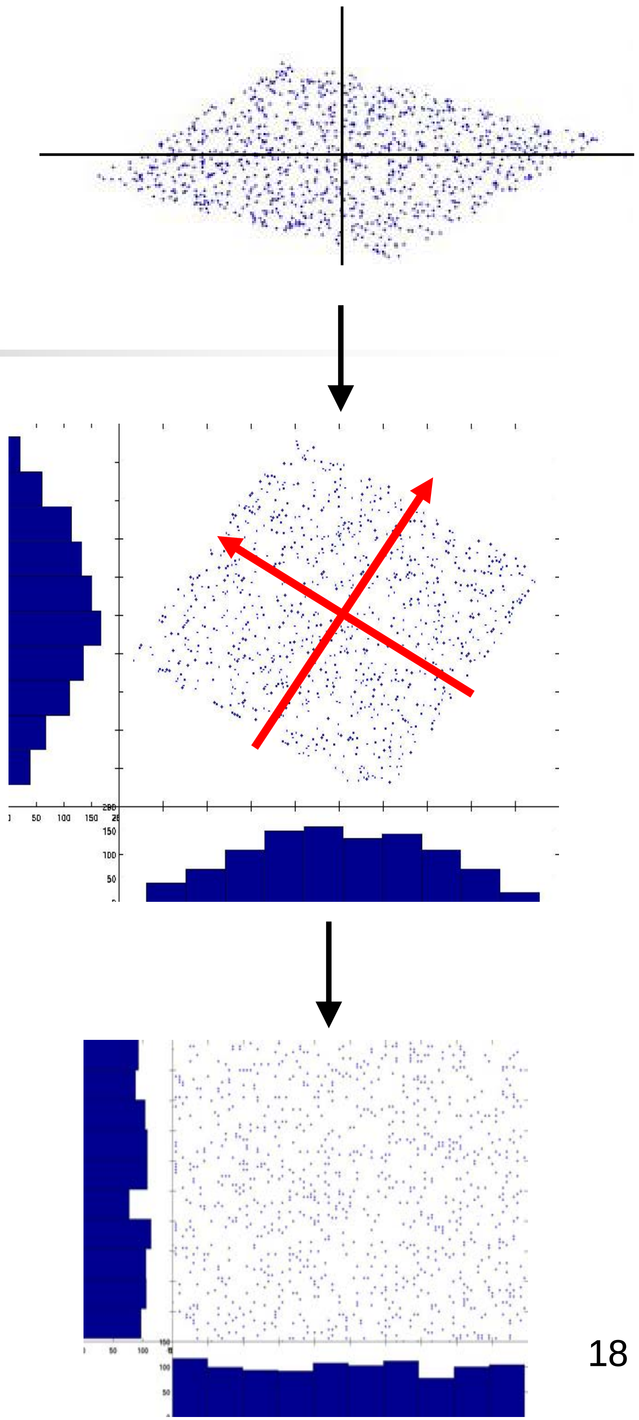
Super-Gaussian: kurtosis > 0

Sub-Gaussian: kurtosis < 0

ICA Training Process

- Remove the mean
  $x = x - \langle x \rangle$
- ‘Sphere’ the data by diagonalizing its covariance matrix,
  $x = \langle xx^T \rangle^{-1/2}(x - \langle x \rangle)$.
- Update W according to

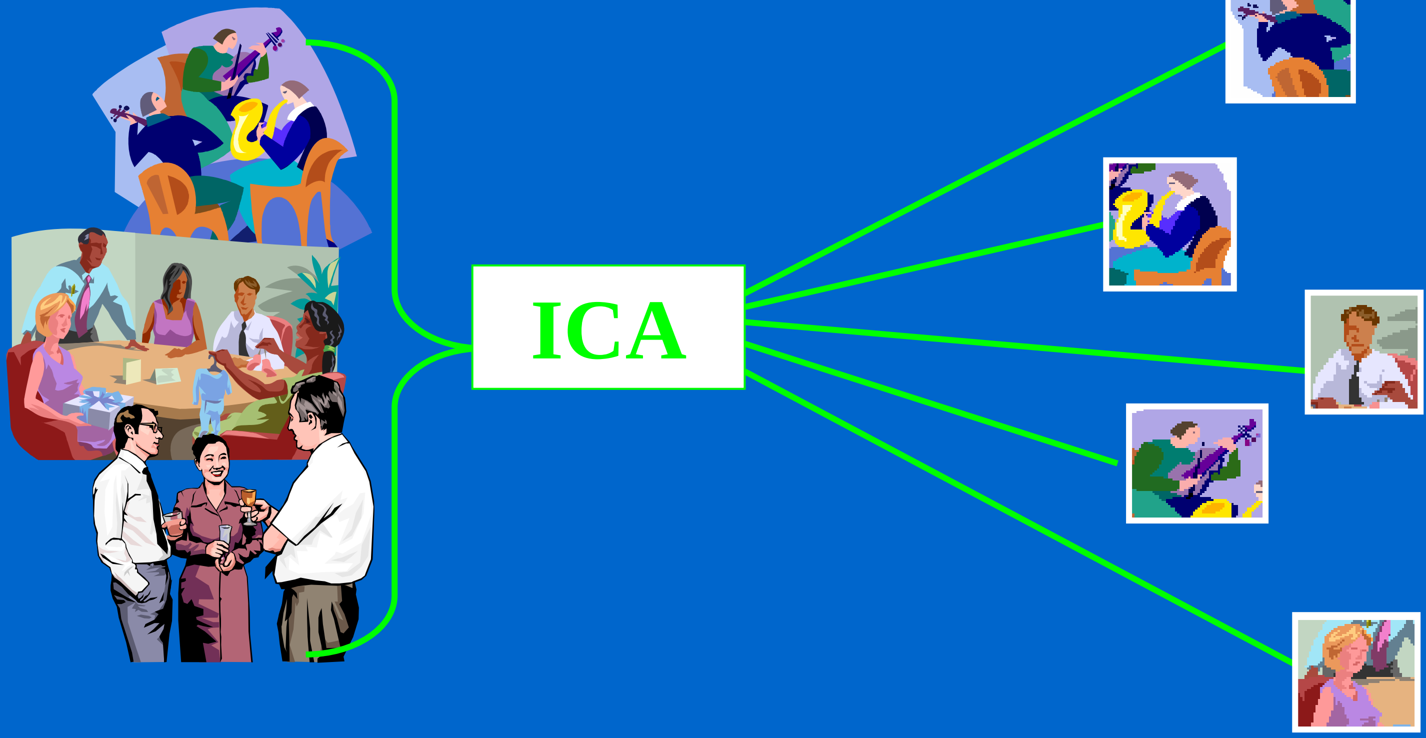
$$\Delta \mathbf{w} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{w}} \mathbf{w}^T \mathbf{w}$$



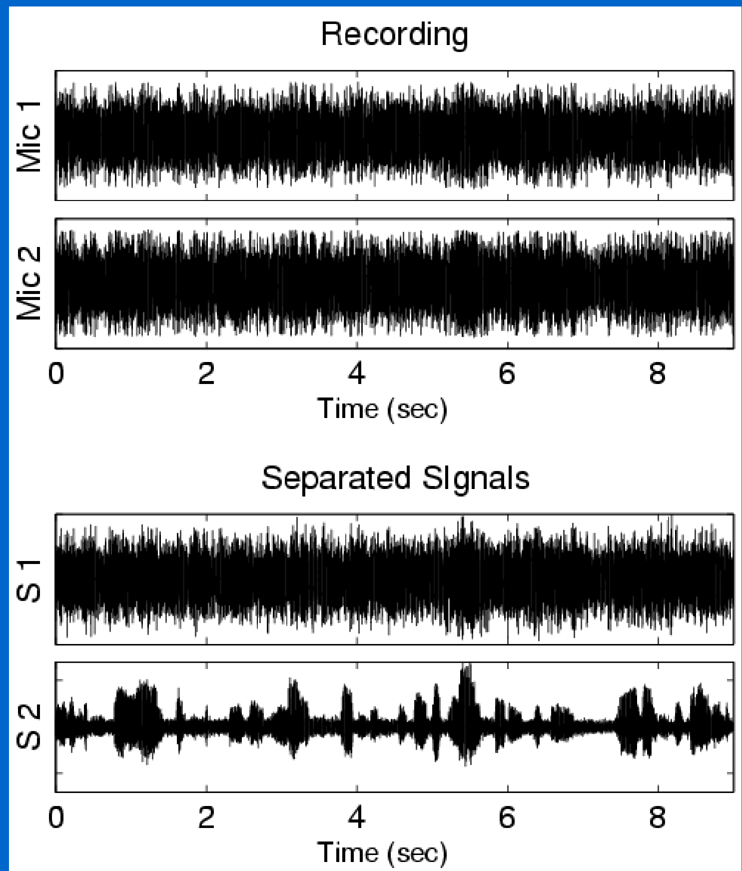
ICA Applications

- Speech enhancement (noisy speech recognition)
- Image processing
- Biomedical signal processing (**EEG**, ERP, **fMRI**, MEG)

Example: Speech Separation



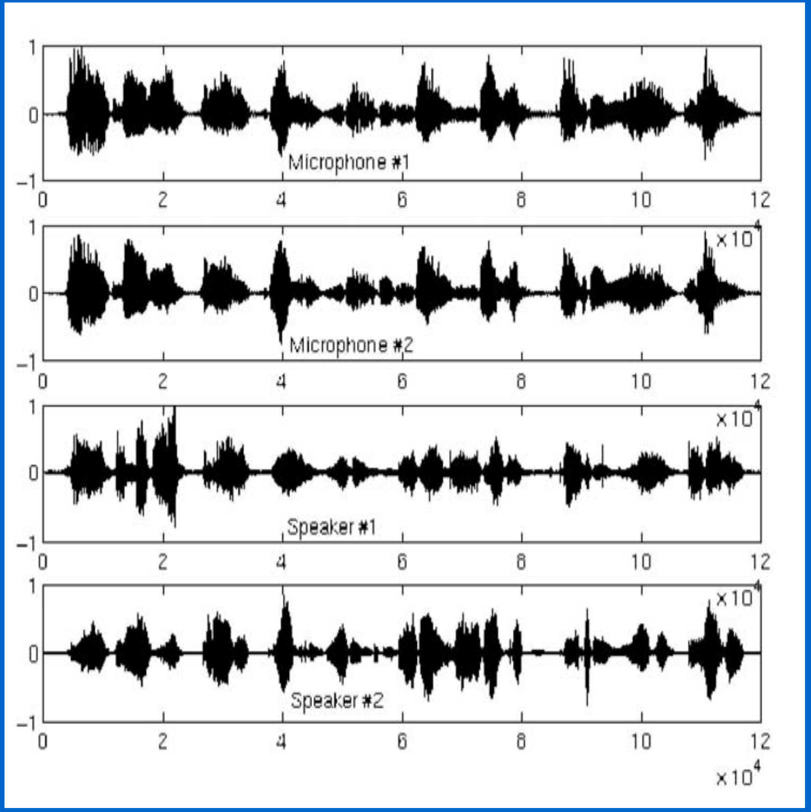
Speech Enhancement & Recognition



Speech Enhancement & Recognition

Separation of Two Speech Signals

Improves speech recognition rate after separation
Algorithm works for various sounds in different environments.



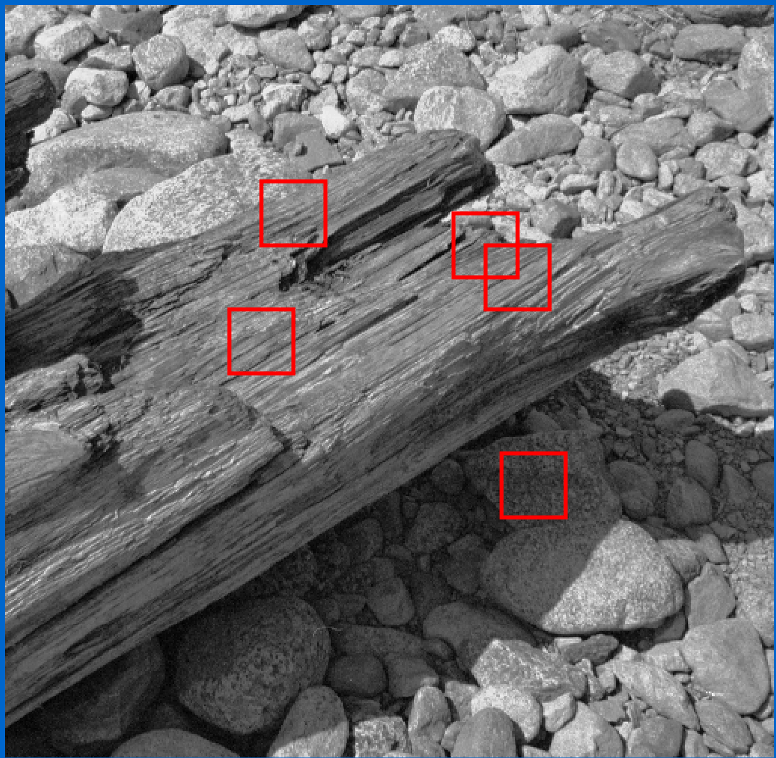
Park and Lee (1999):

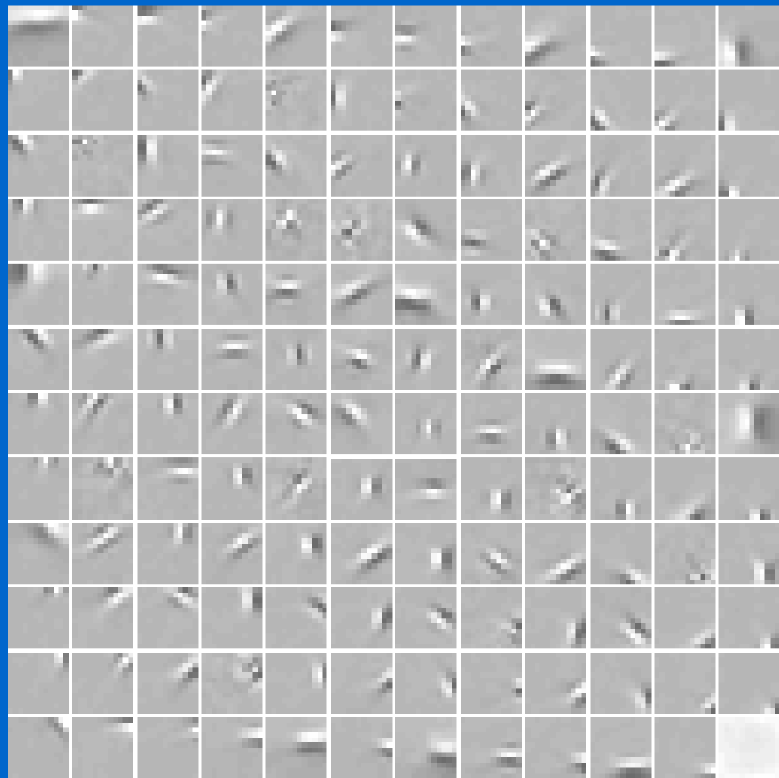
SNR [dB]	*W/o sep.*	*With sep.*
15 dB	87.8%	90.8%
10 dB	68.9%	87.9%
5 dB	37.0%	79.9%

ICA Applications

- Speech enhancement (noisy speech recognition)
- Image processing
- Biomedical signal processing (**EEG**, ERP, **fMRI**, MEG)

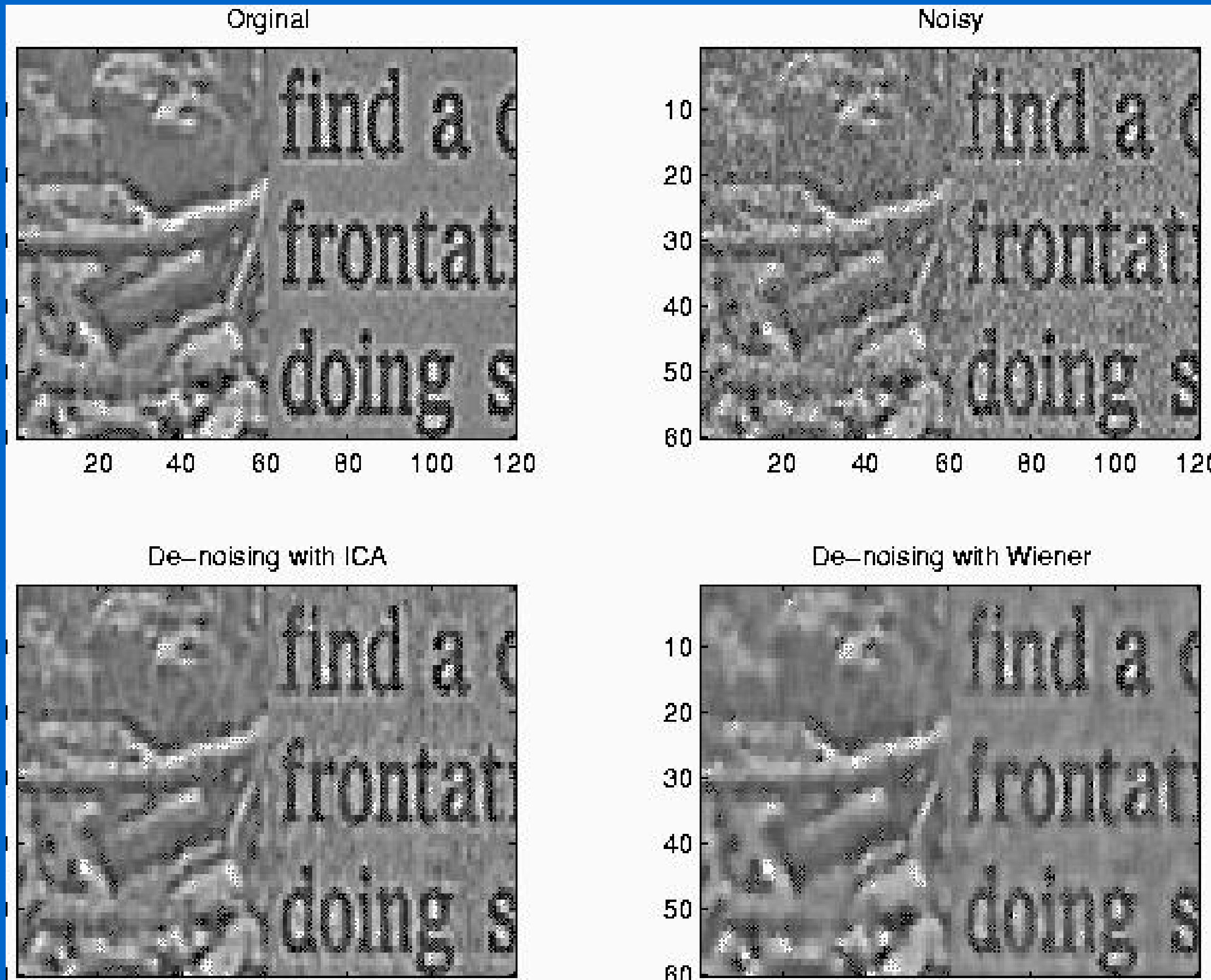
Image Processing – Finding Basis Functions of Images



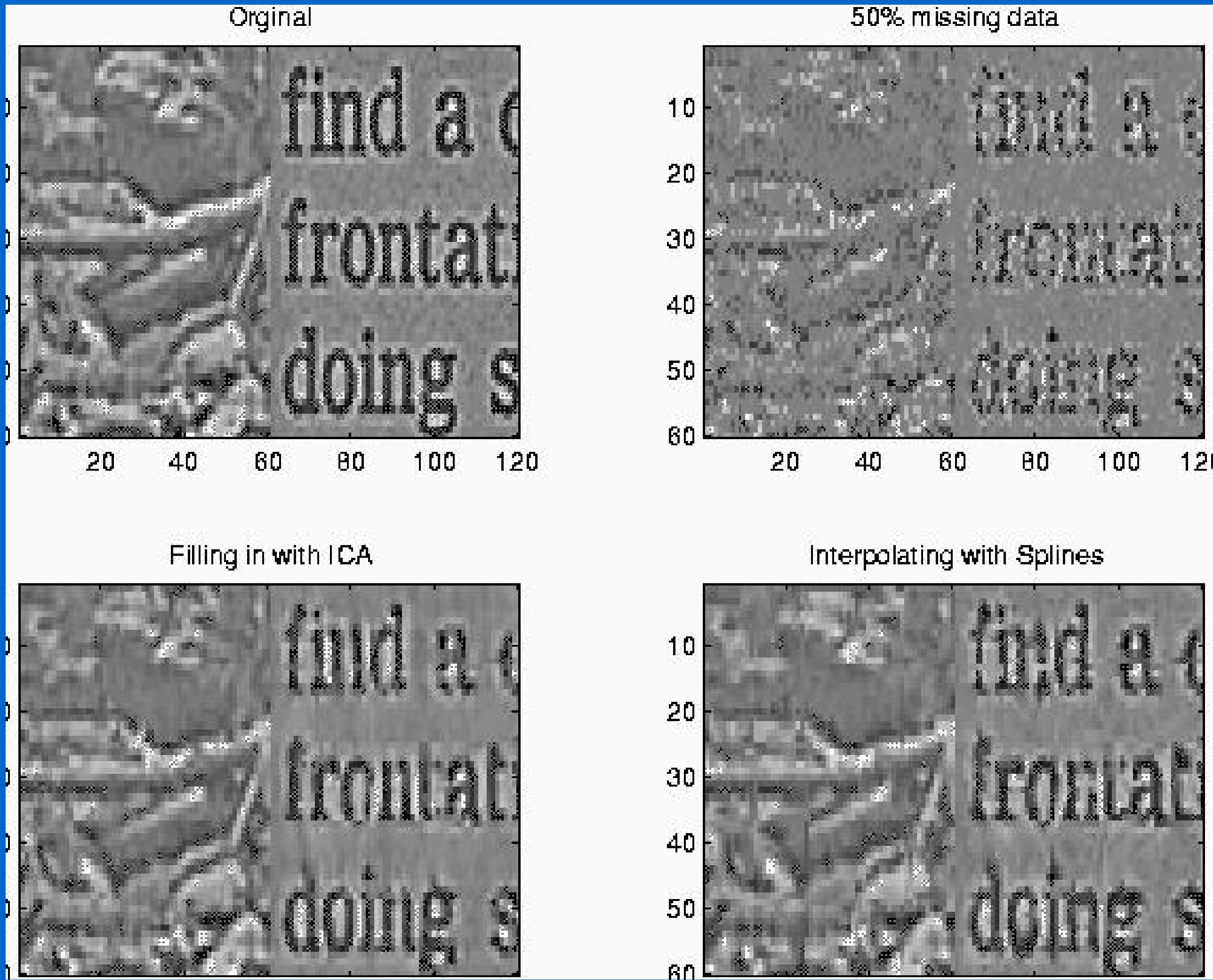


Set of 144 basis functions

Image De-noising



Filling in missing data

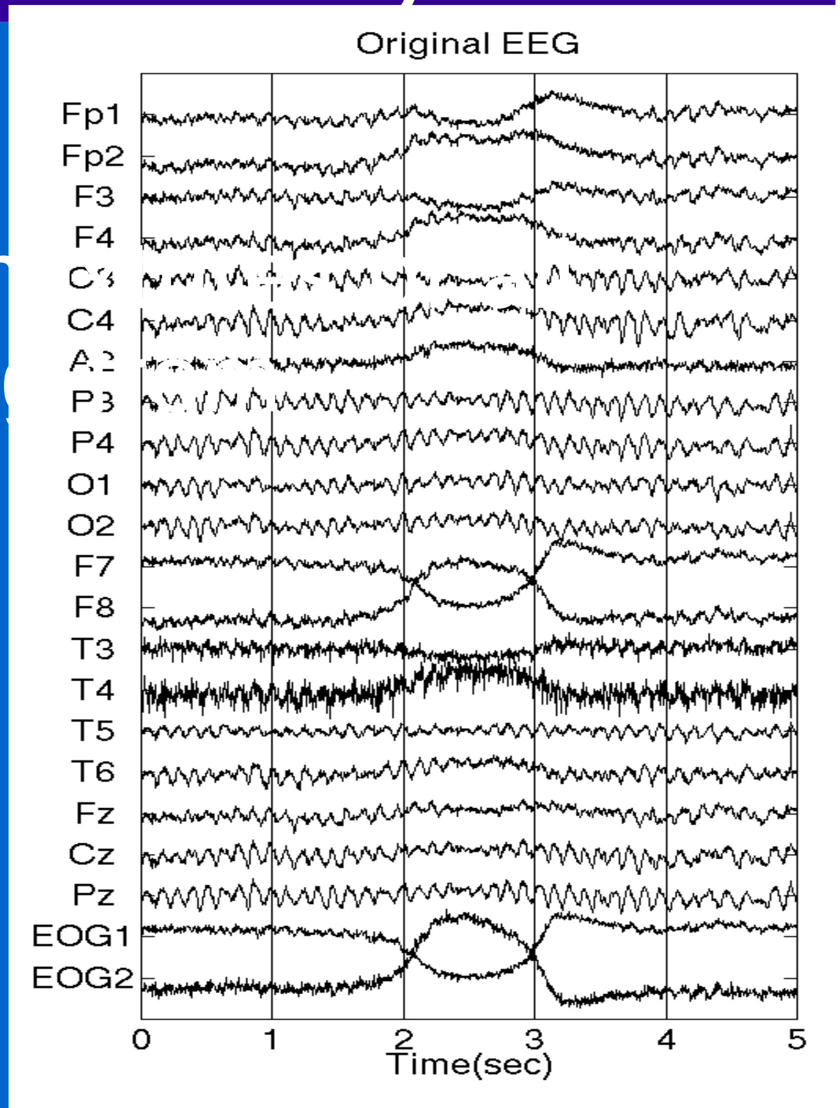


ICA Applications

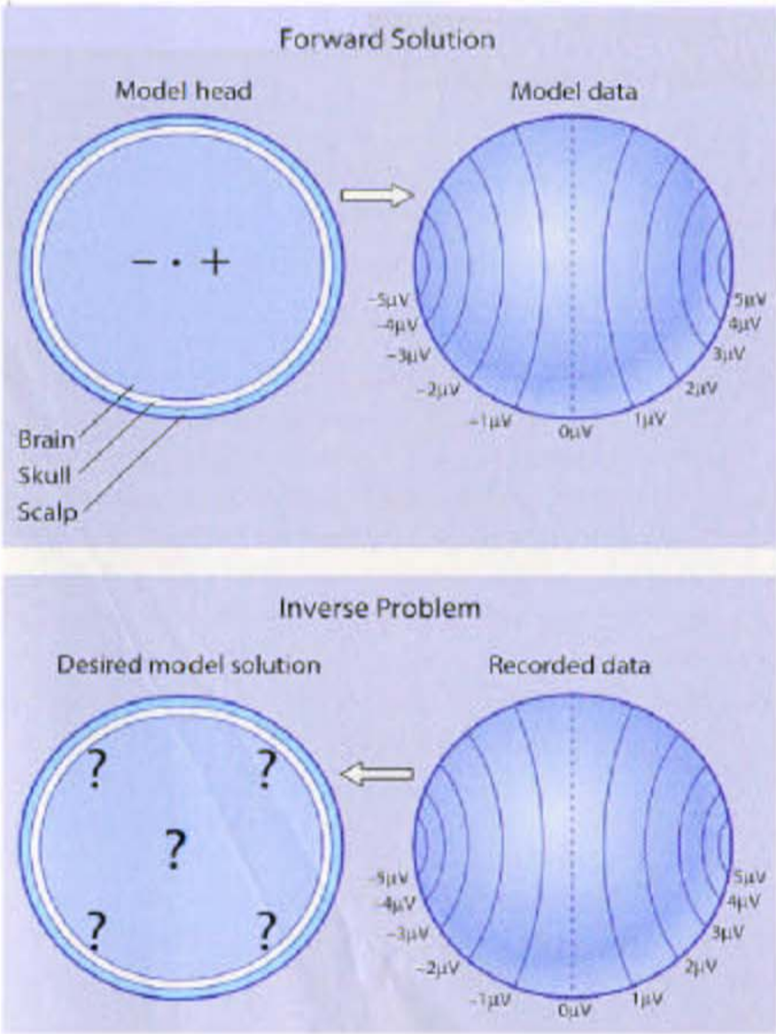
- Speech enhancement (noisy speech recognition)
- Image processing
- Biomedical signal processing (**EEG**, ERP, **fMRI**, MEG)

Challenges of EEG Analysis

- Pervasive artifacts
- EEG recordings are n
  brain activities arisin
  different networks
- Response variability
- Inverse problem
- etc



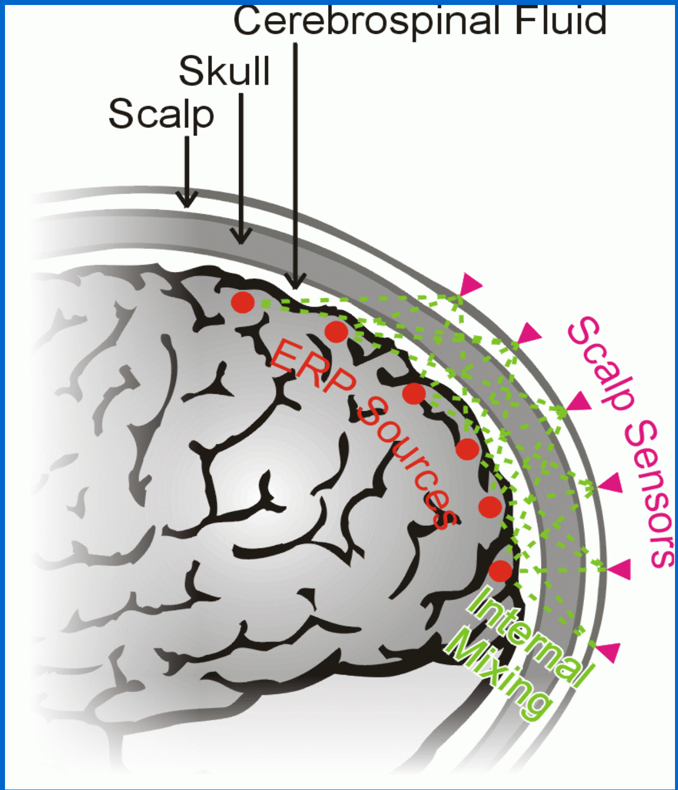
2. Inverse solution is not unique



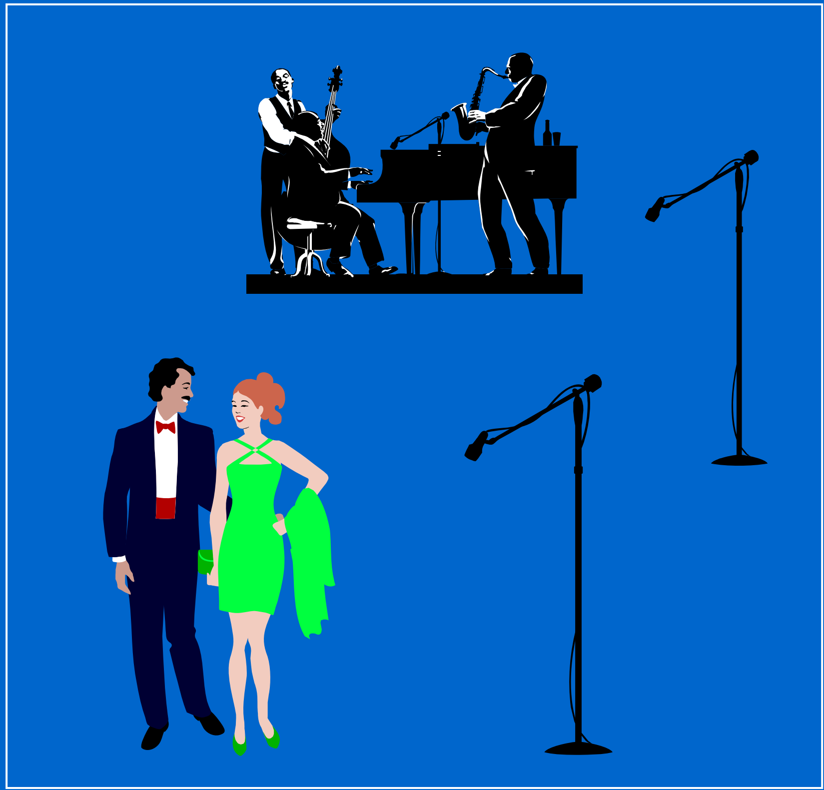
A single pattern of neural activity will produce a unique scalp map

BUT …A single scalp map could have been produced by an infinite number of patterns of neural activity

3. EEG data are mixtures of source signals

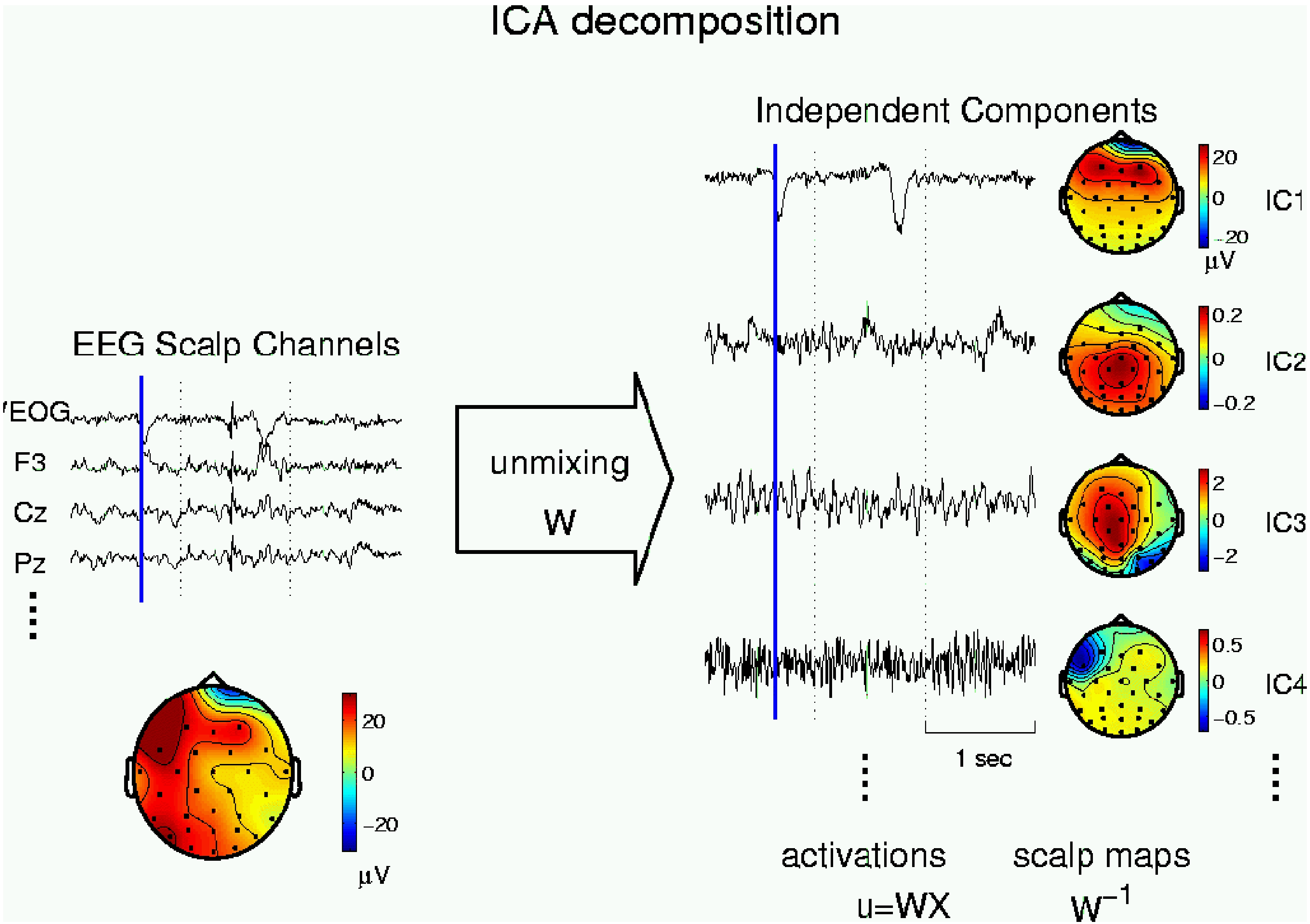


Cocktail Party



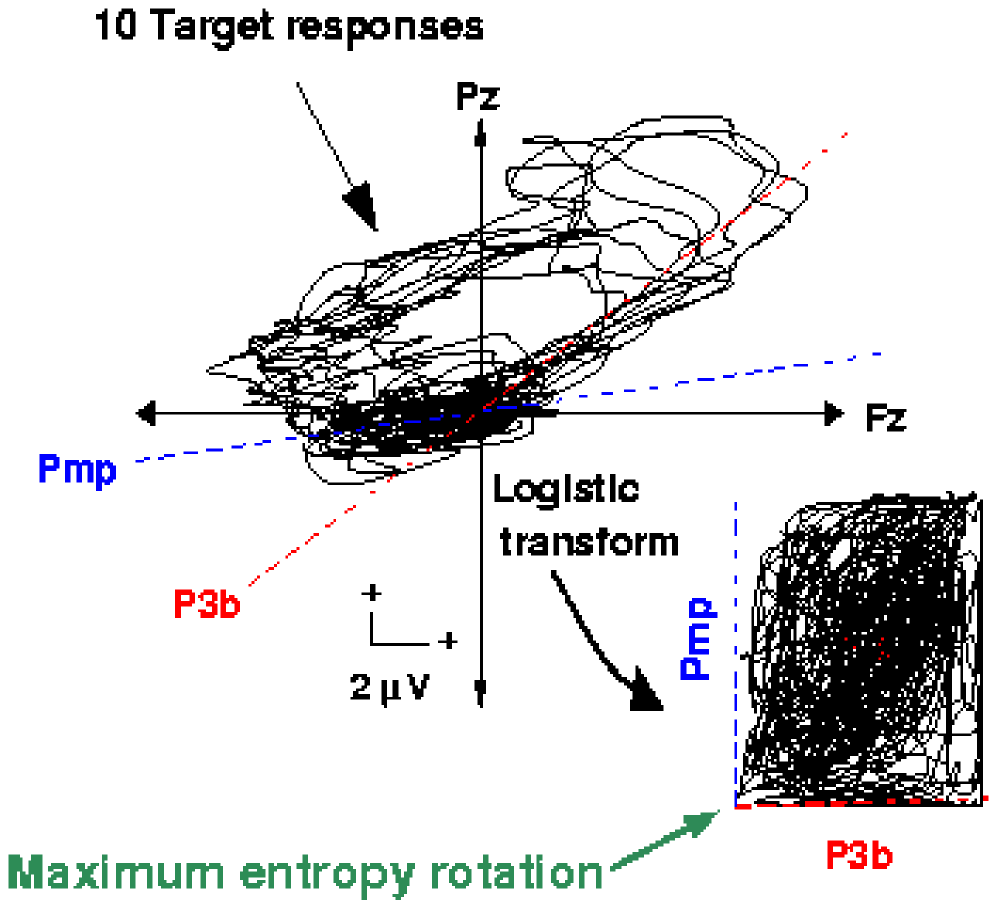
ICA/EEG Assumptions

- Mixing is linear at electrodes
- Propagation delays are negligible
- Component time courses are independent
- Number of components $\leq$ number of channels.



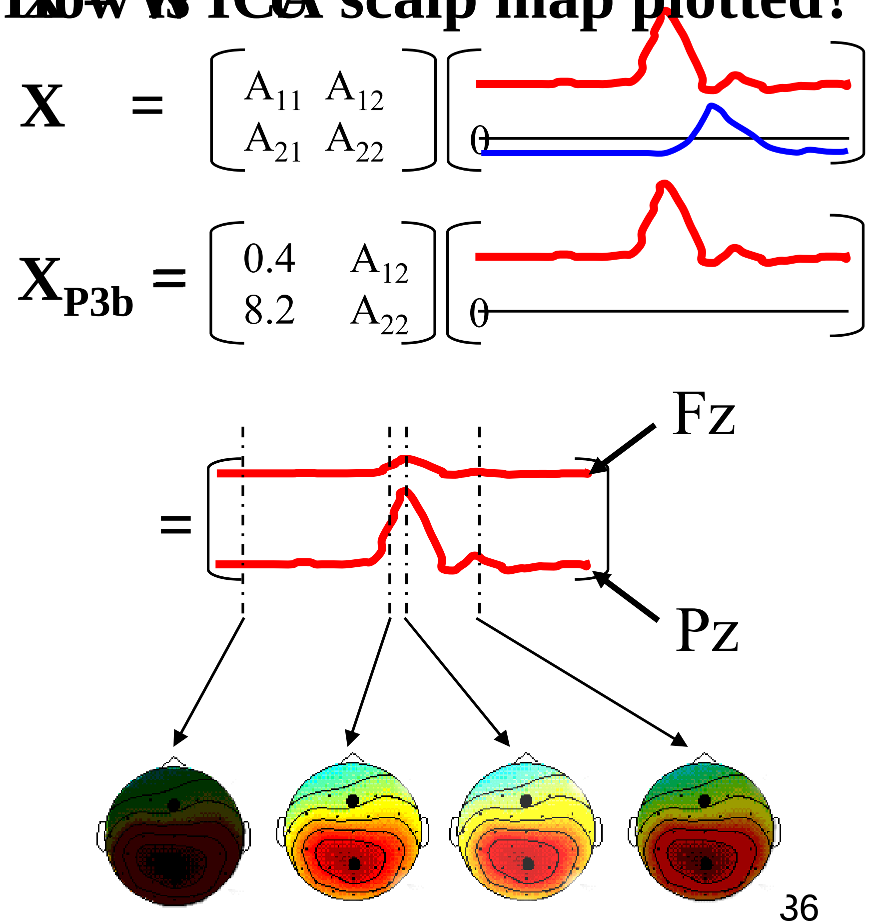
From Jung et al., *Clinical Neurophysiology*, 2000.

Independent components of EEG/ERP



From Makeig et al., *JNS,* 1999.

How is ICA scalp map plotted?



Frequently Asked Questions

- What is temporal and spatial ICA?

For EEG, we are looking at temporally independent brain activities arising from different brain networks.

For fMRI, the independence is considered over voxels because of brain modularity. i.e., Simplistically, "Different places do different things."

Frequently Asked Questions (cont.)

- How much data is enough data?

  There is no fixed limit to the number of points needed for a "good" ICA solution - and in fact no fixed way to judge whether an ICA solution is "good" or not.

Frequently Asked Questions (cont.)

- How should the activations be scaled?

  $$U = WX, \quad X = W^{-1} * U$$

  The strength of source activity is distributed between the columns of W^{-1} and the rows of U.

- Can ICA separate ‘correlated’ source activities?

Practical Issues with ICA of EEG/ERP

1. Apply ICA to averaged ERPs
   - How many time points are needed for training?

     Suggestion: At least several times number of variables in the unmixing matrix.
   - Which EEG processes may express their independence in the ERP training data?

     Suggestion: Decompose the concatenated collection of ERP averages in respond to the experimental stimulus and task conditions.
   - ICA decomposition of averaged ERPs must be interpreted with caution.

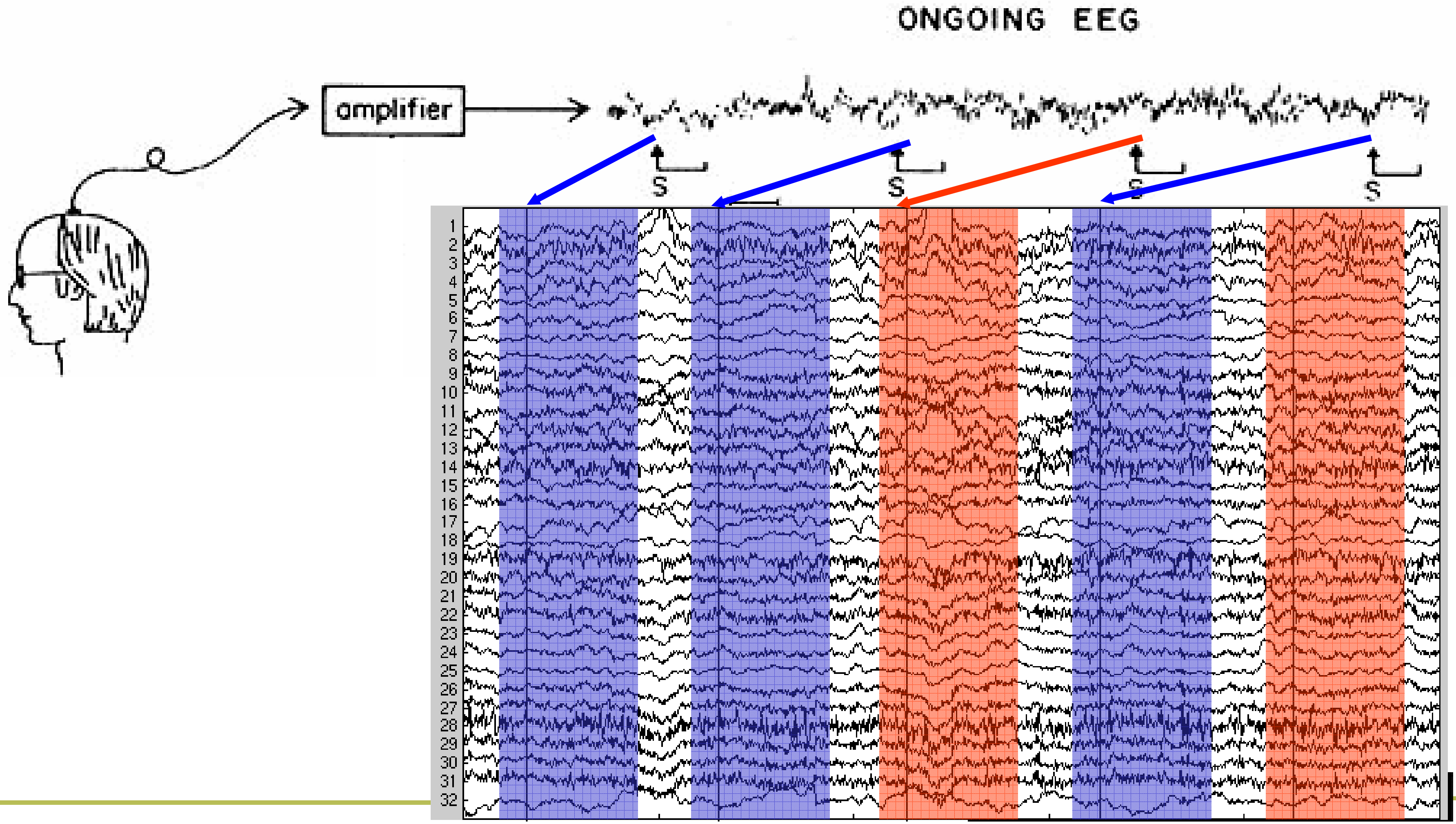
Practical Issues with ICA of EEG/ERP

2. Apply ICA to continuous EEG data
   - Are components spatially stationary through time?

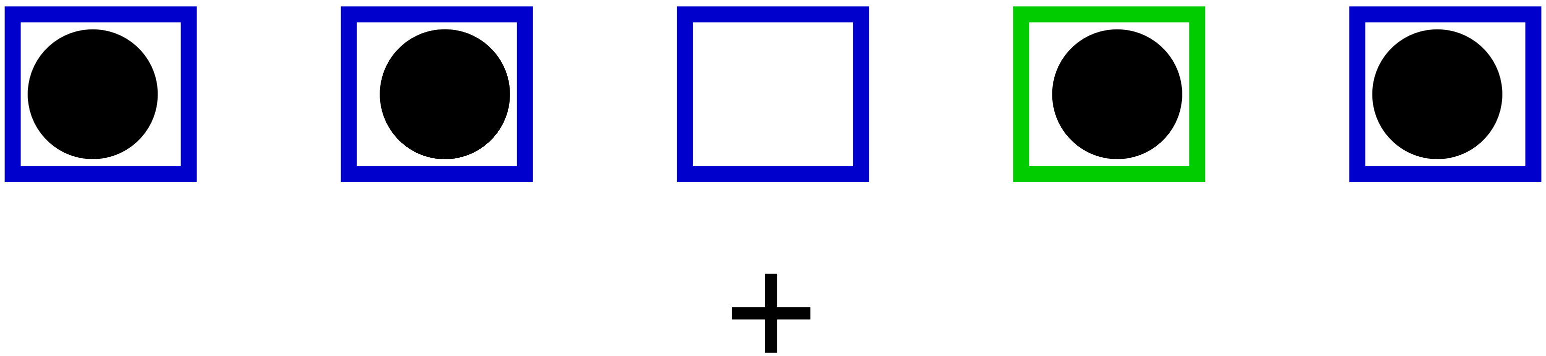
     Suggestion: Perform separate decompositions of subsets of the recorded data, each consisting of periods during which the sources may be stationary.

Practical Issues with ICA of EEG/ERP

3. Apply ICA to unaveraged event-related EEG



Experiment

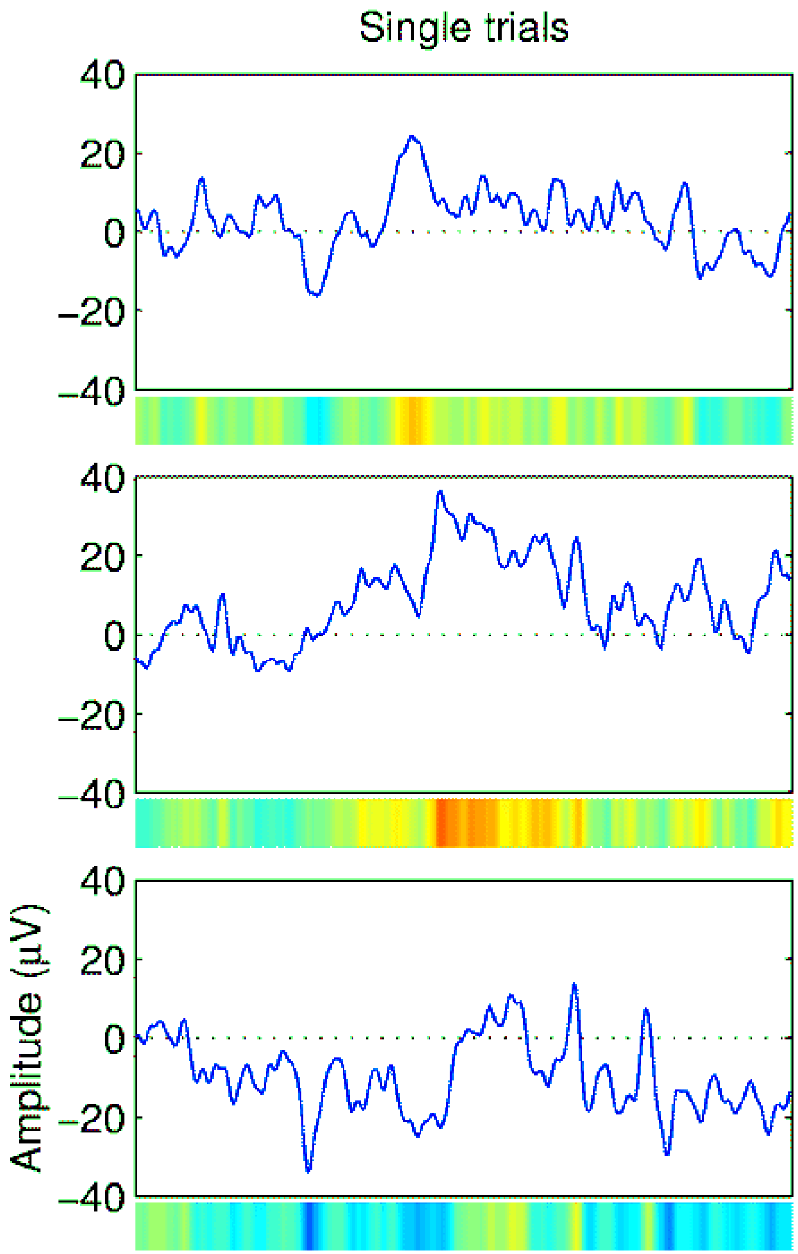


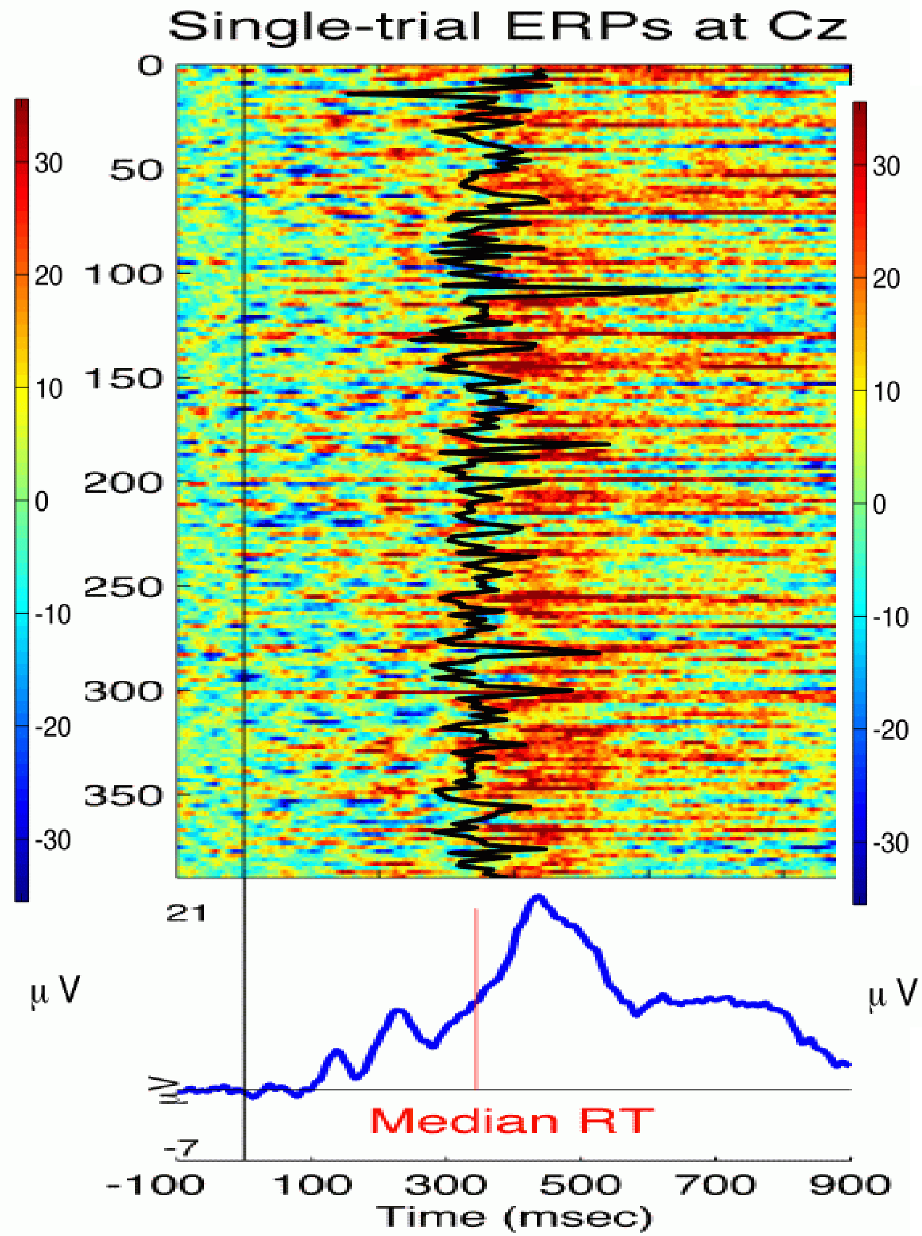
Task: Fixate **cross** while covertly attending to **green box**. Press button when **circle** is flashed in green box.

Subject: 28 normal control, 14 autistic and 8 cerebellar lesion subjects.

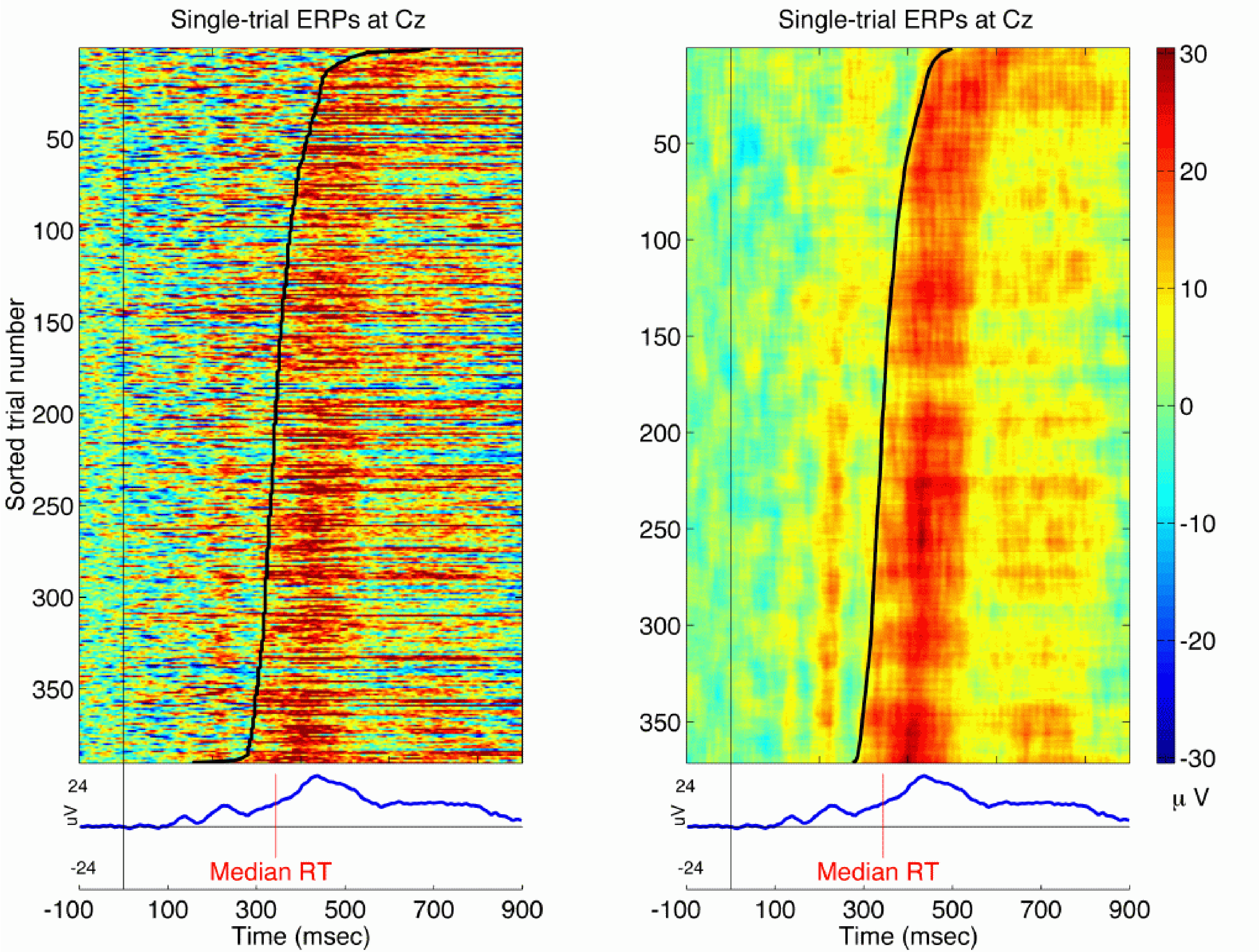
Session: 30 72-s task blocks, including 120 **targets** and 480 nontargets in each of the 5 locations.

ERP Image





From Jung et al., *NIPS*, 1999.



From Jung et al., *NIPS*, 1999.

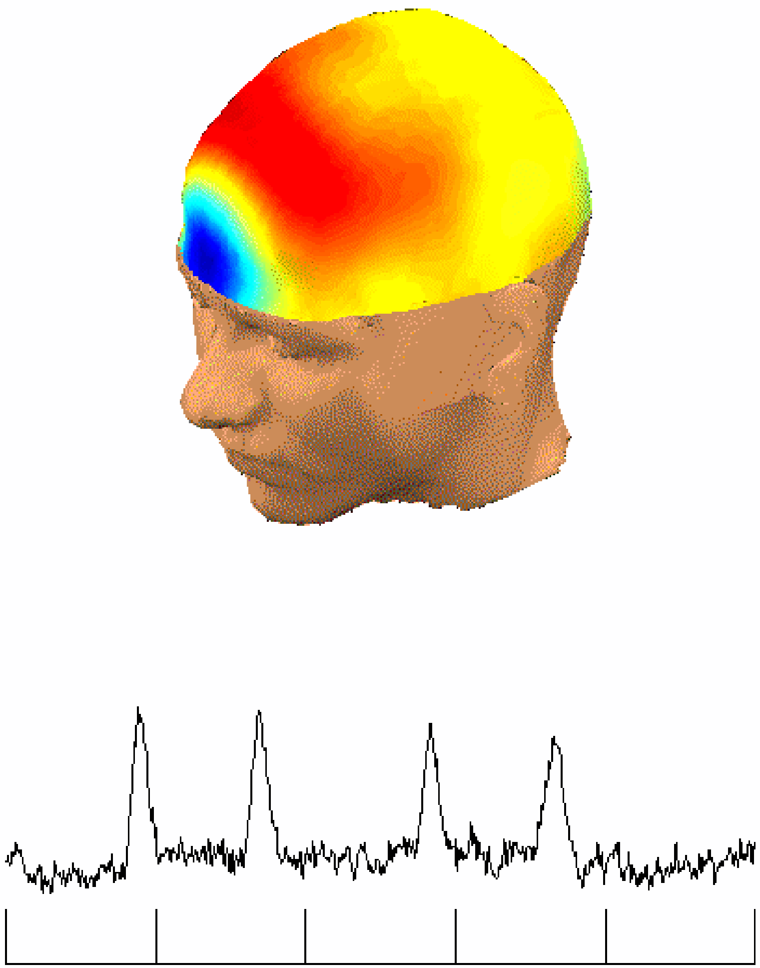
Analysis of Single-trial ERPs

ICA applied to ~600 (single-subject, 31-channel, 1-s) concatenated single-trial response epochs timelocked to detected **target stimuli**

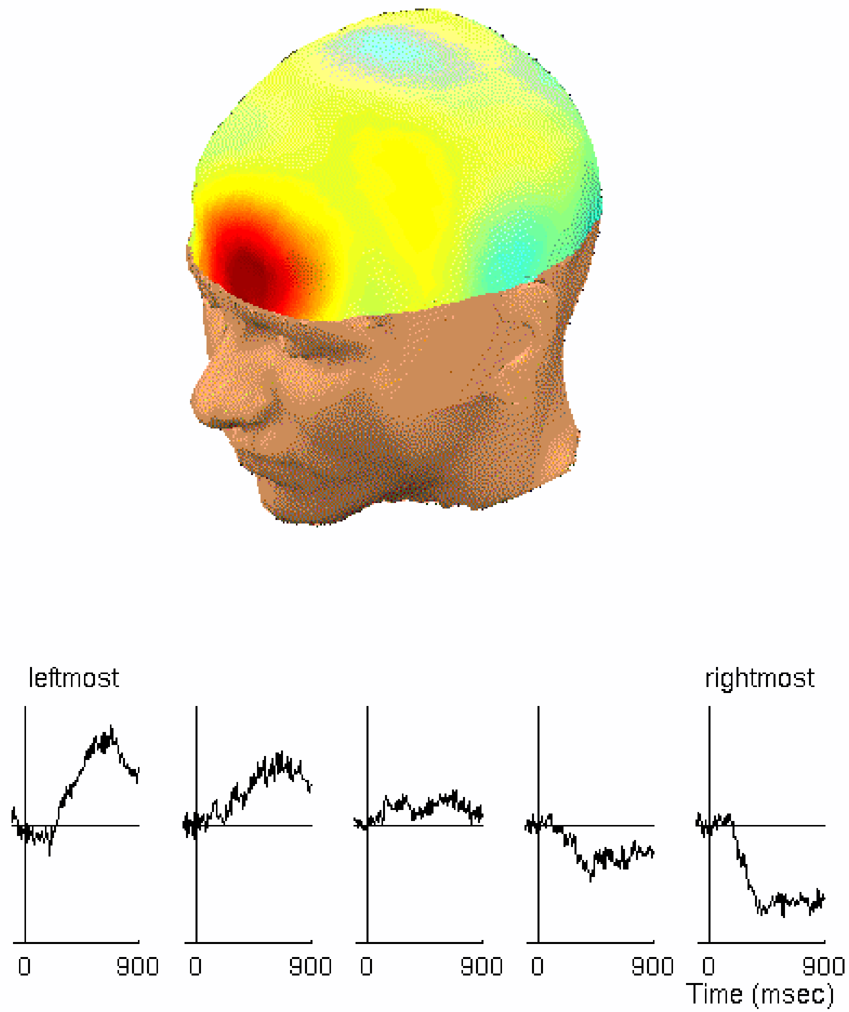
31 independent components having:

- fixed spatial projections to the scalp
- temporally independent time courses of activation

Component 1

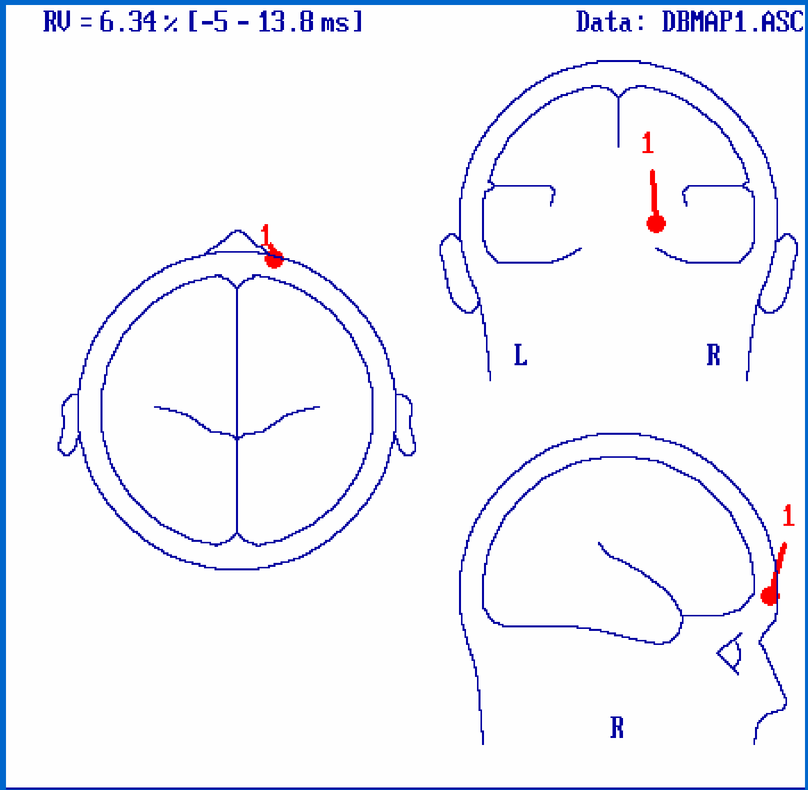


Component 2

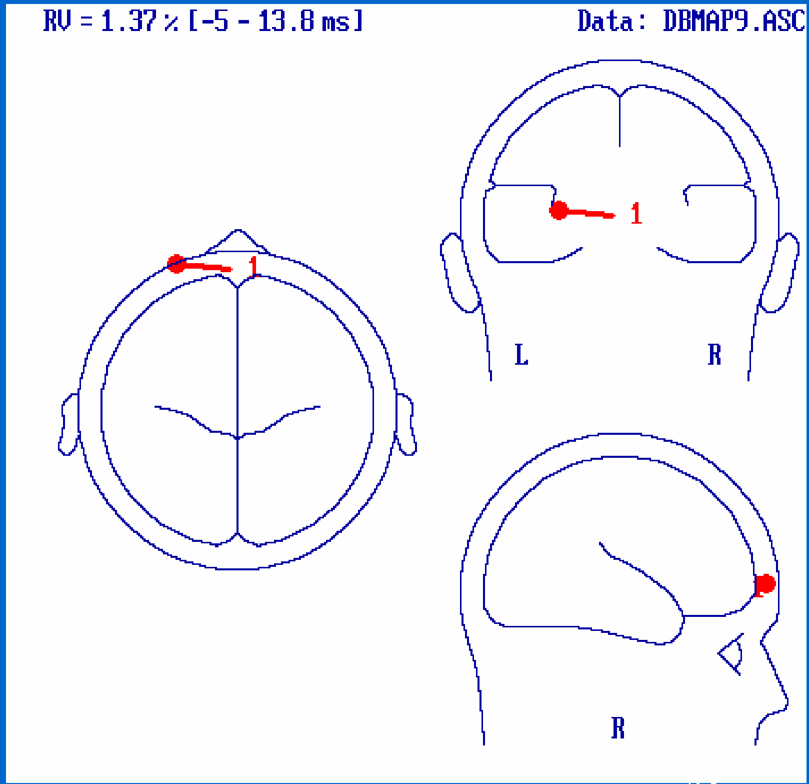


Single-dipole BESA Modeling

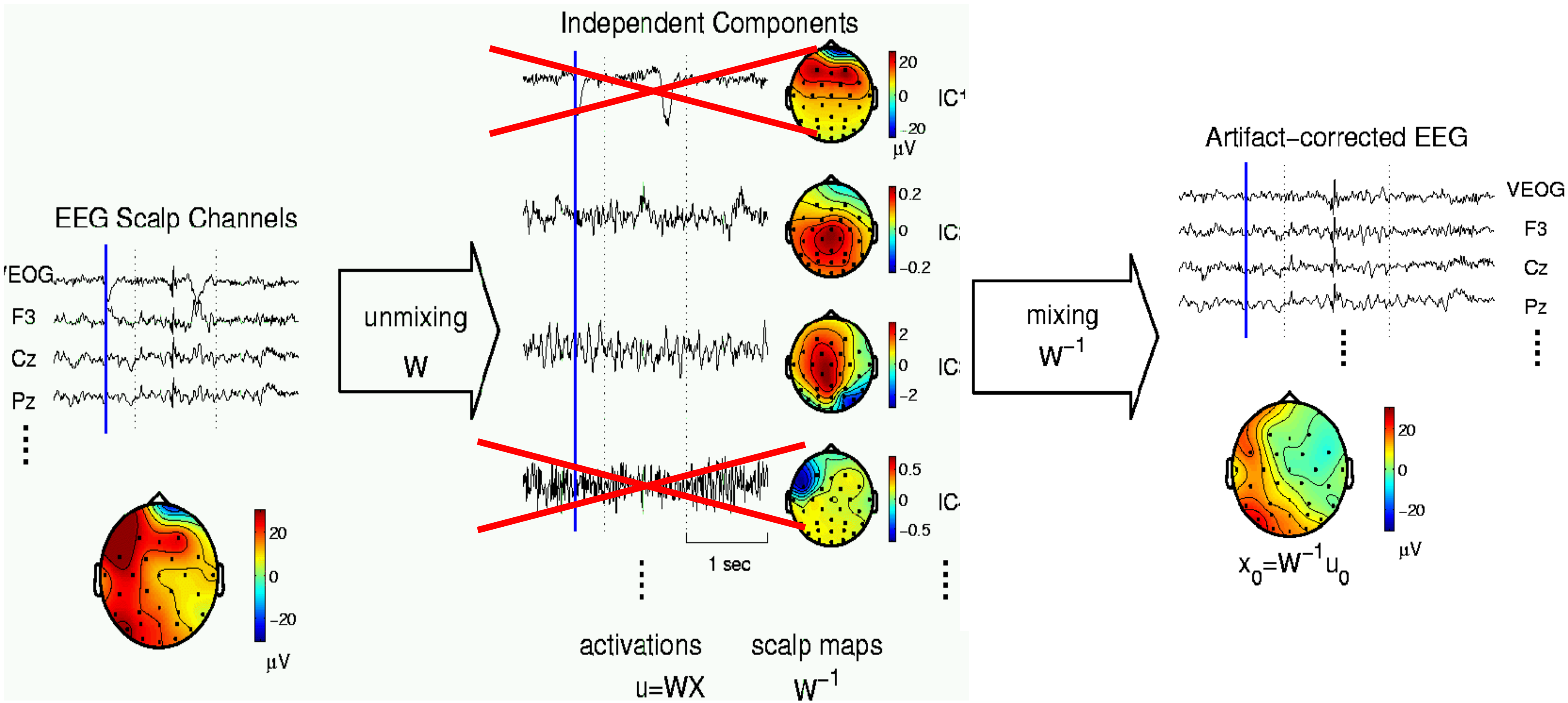
Component 1



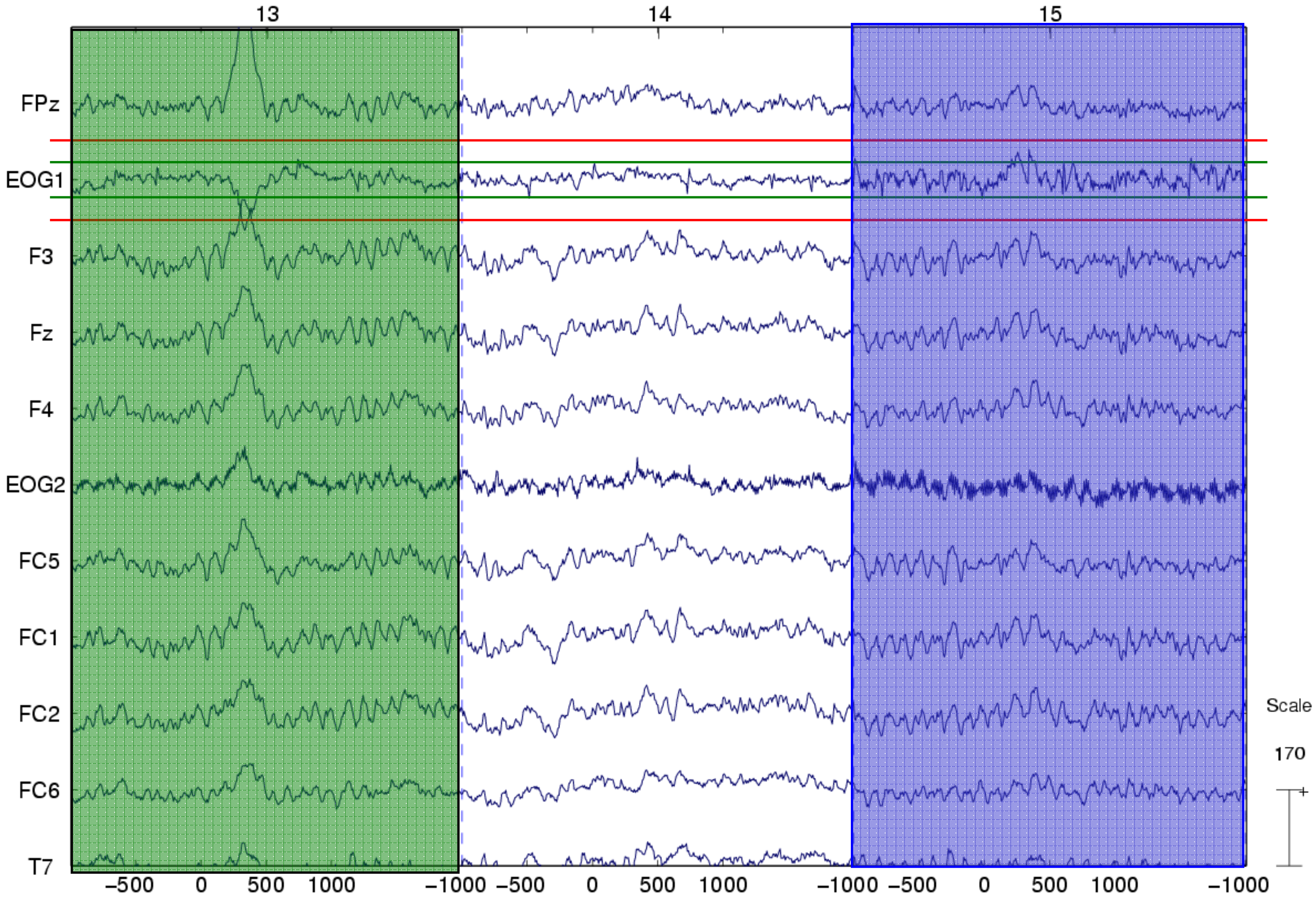
Component 2



ICA-based Artifact Correction

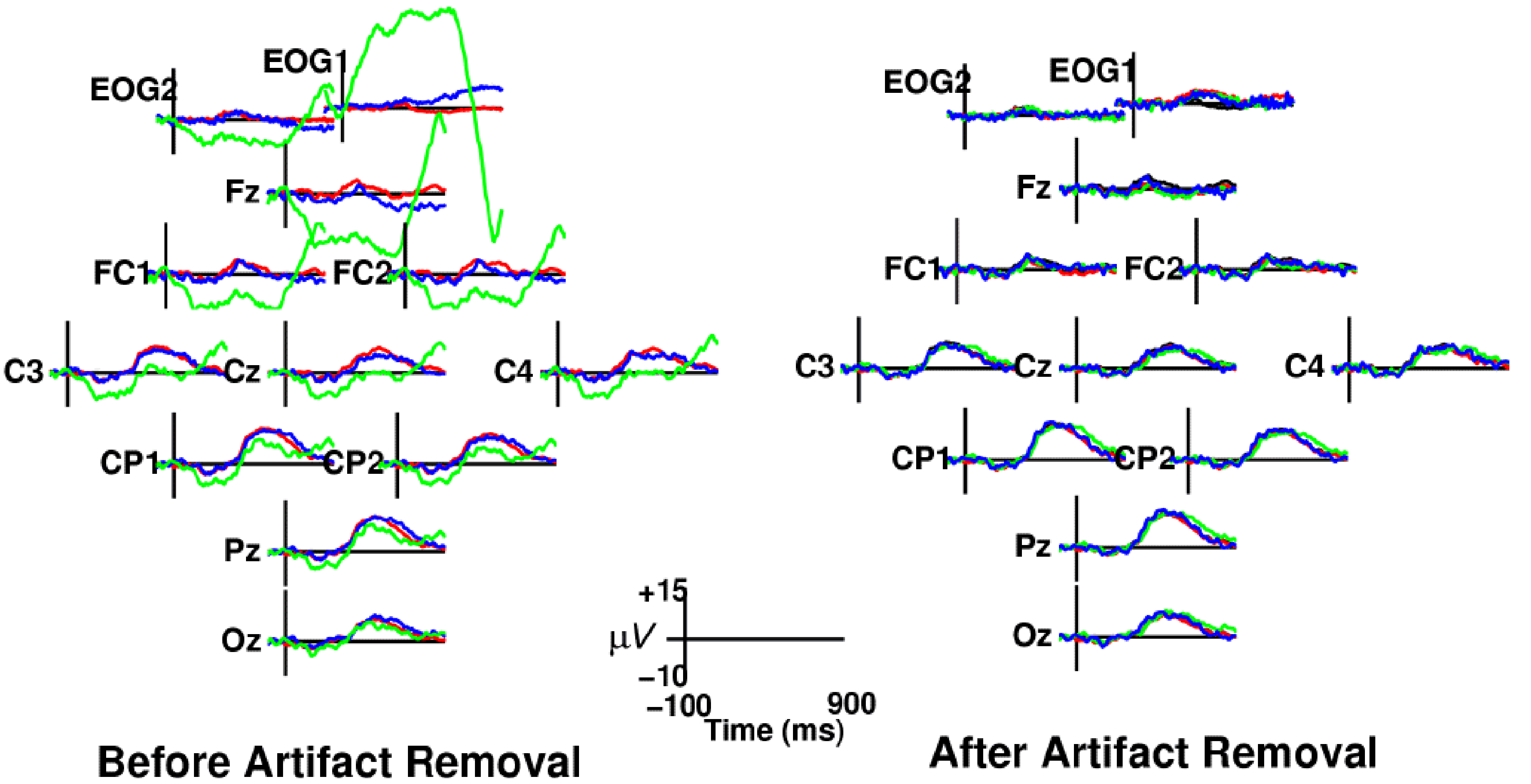


Split Single Trials based on EOG



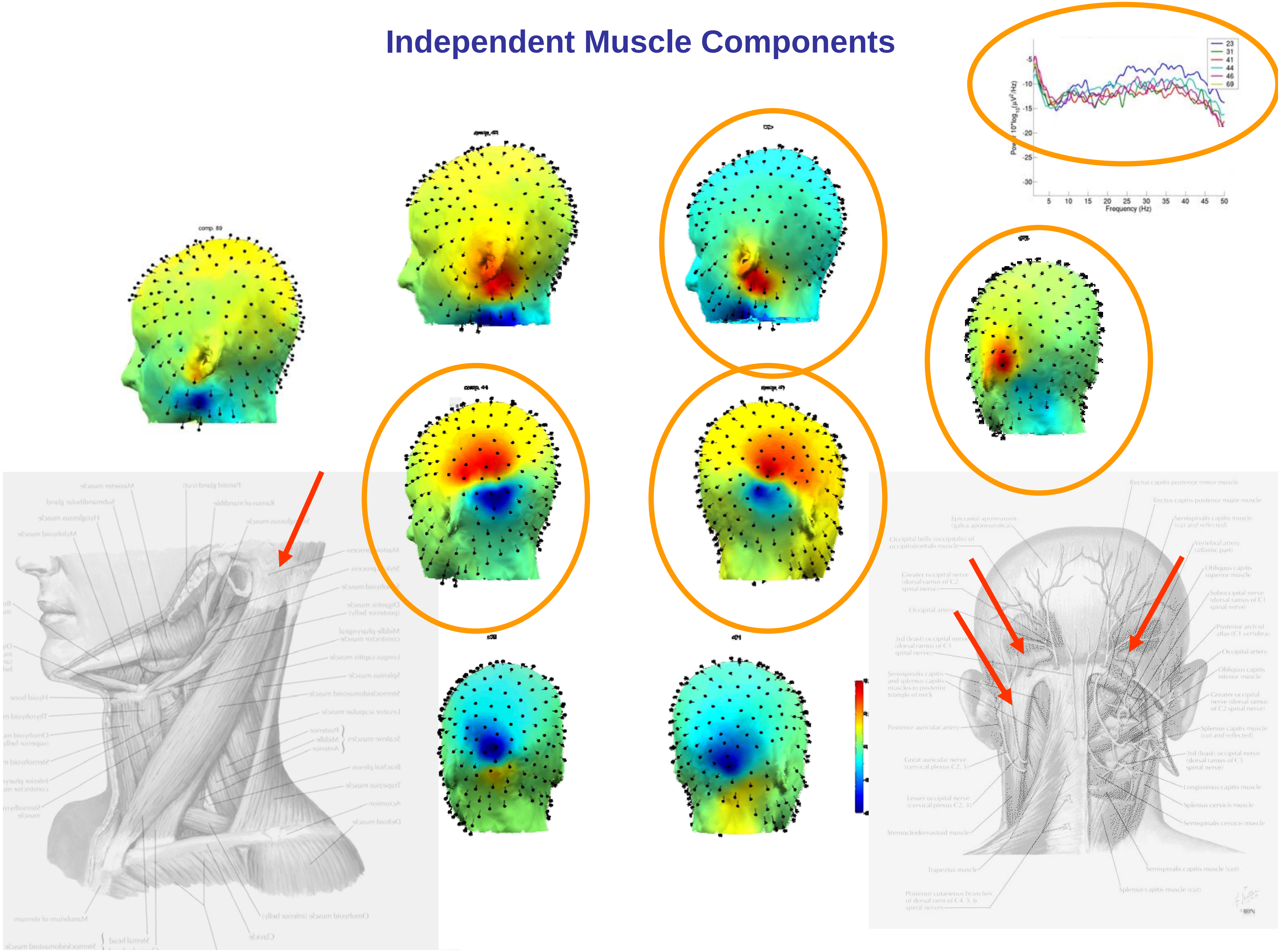
Heavily contaminated **Clean trials** **Contaminated**

Averages of Least, Moderately and Heavily Contaminated Trials

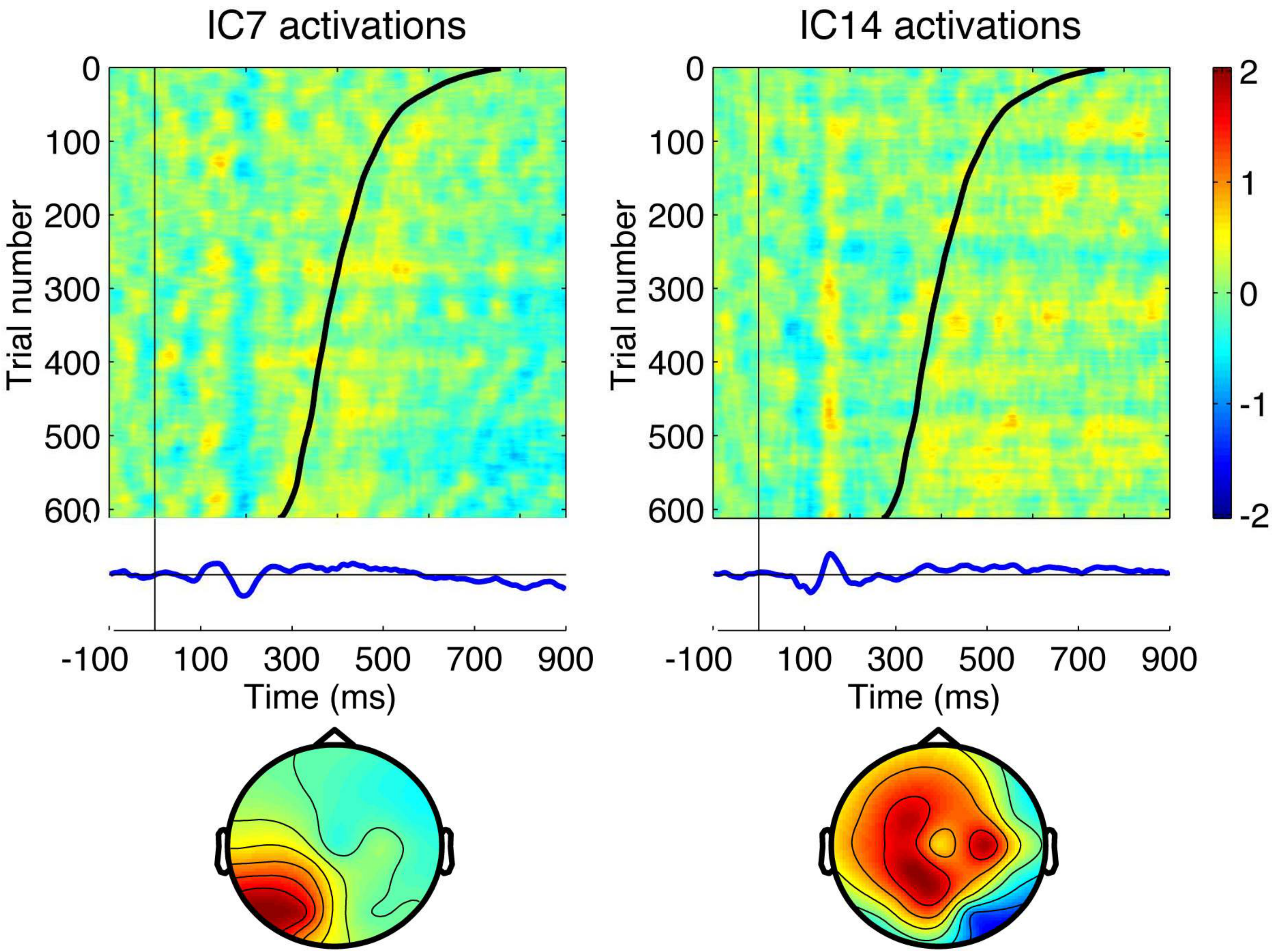


From Jung et al., *Clinical Neurophysiology*, 2000.

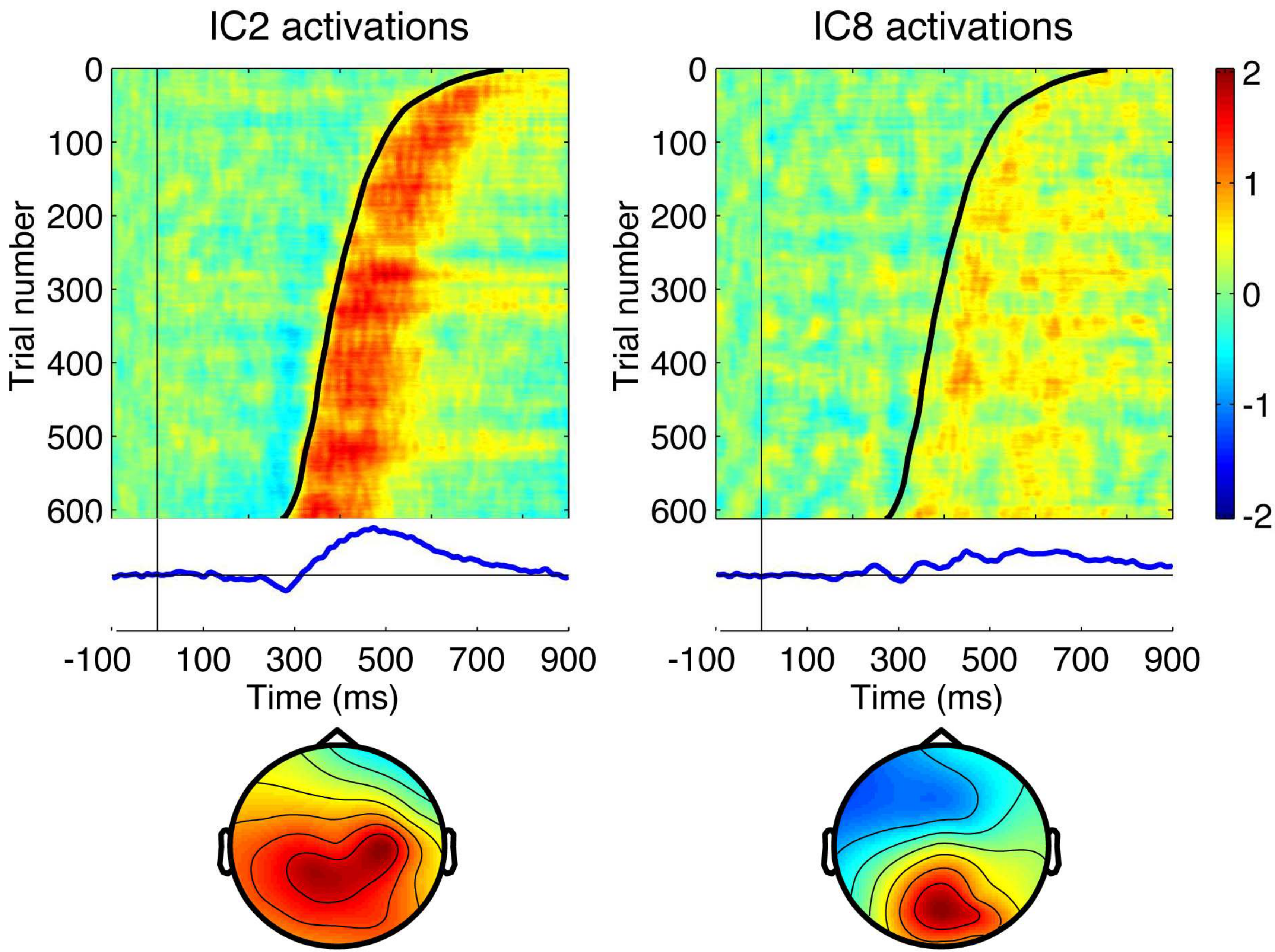
Independent Muscle Components



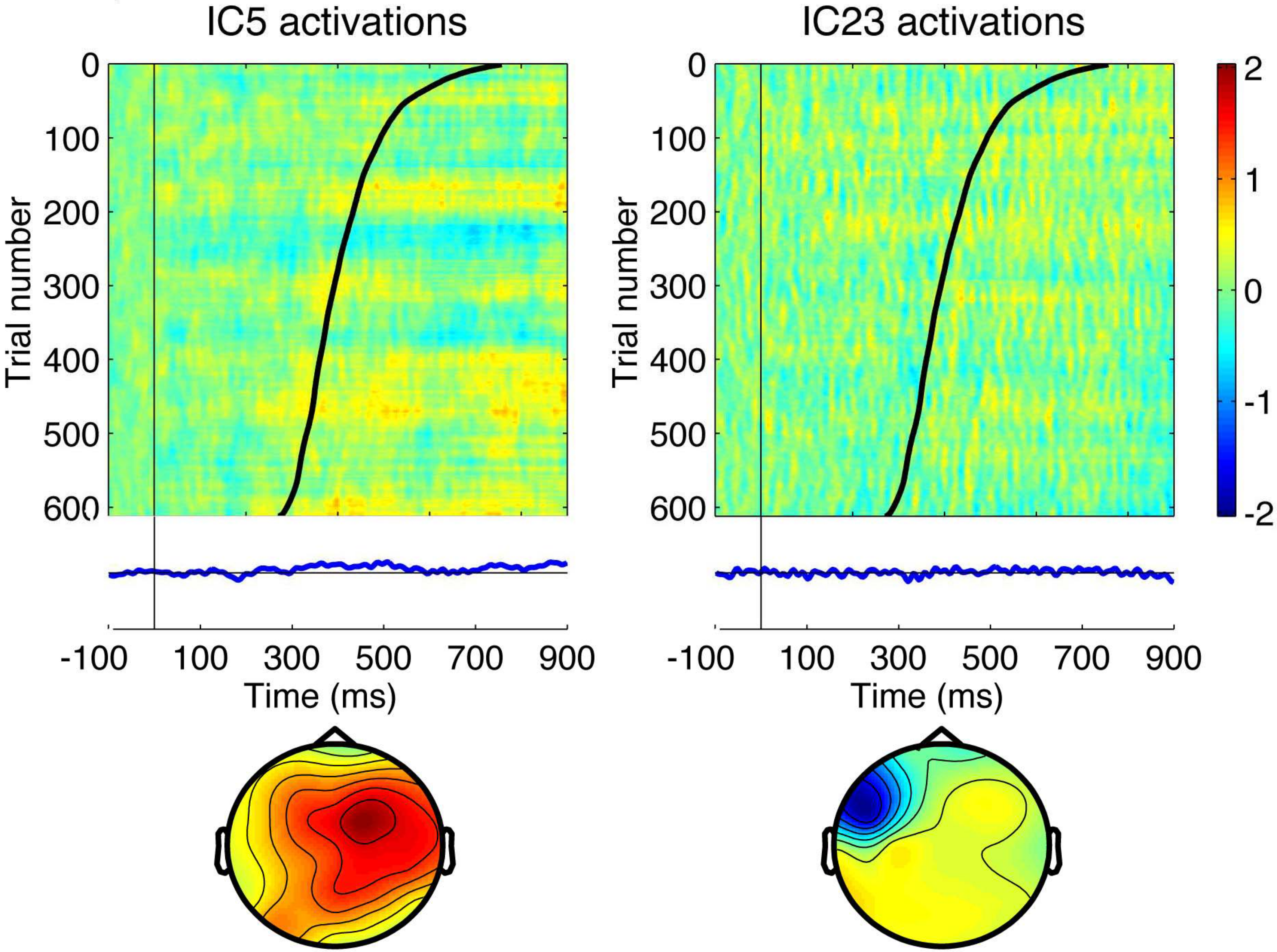
Stimulus-locked

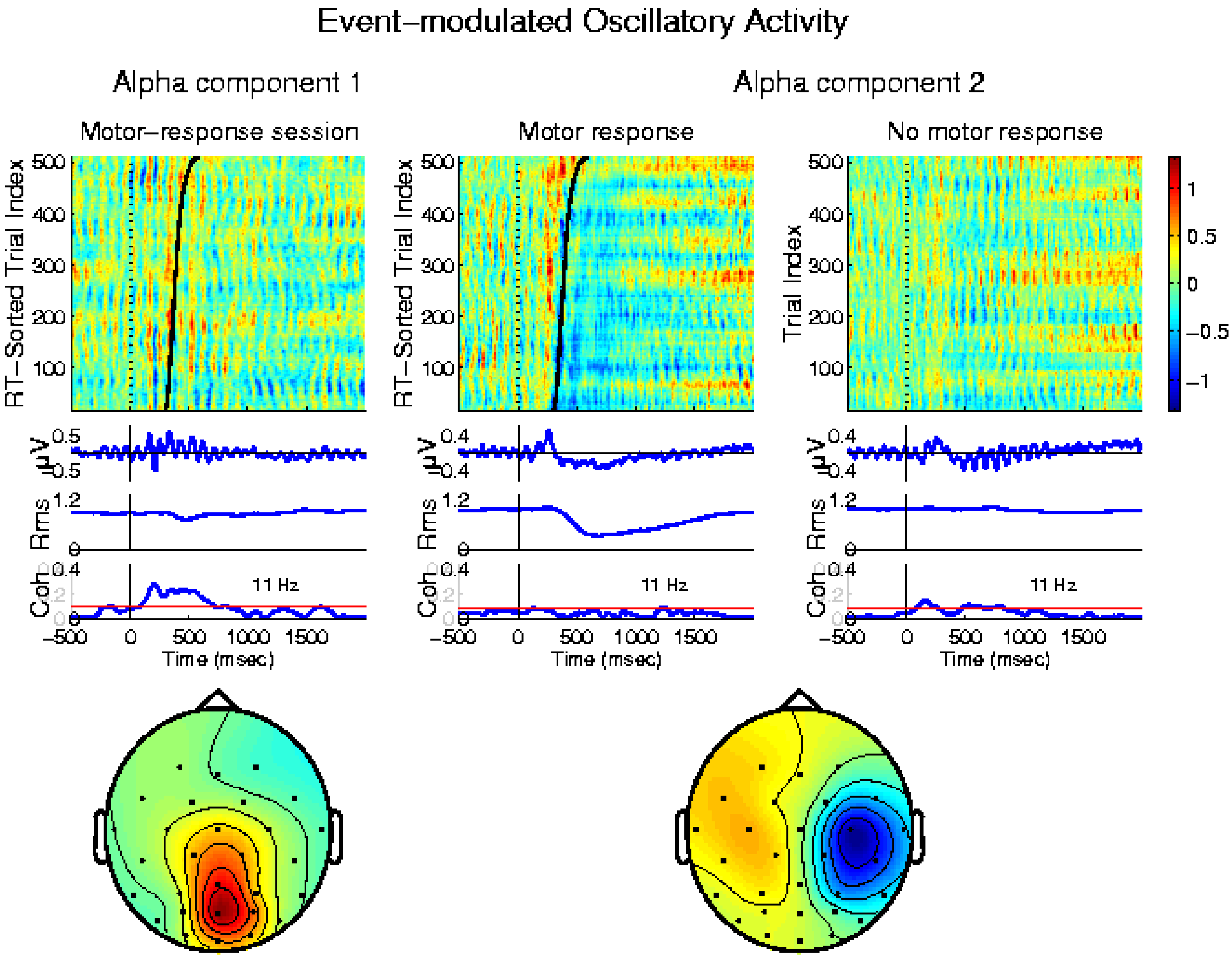


Response-locked



Non-phase locked

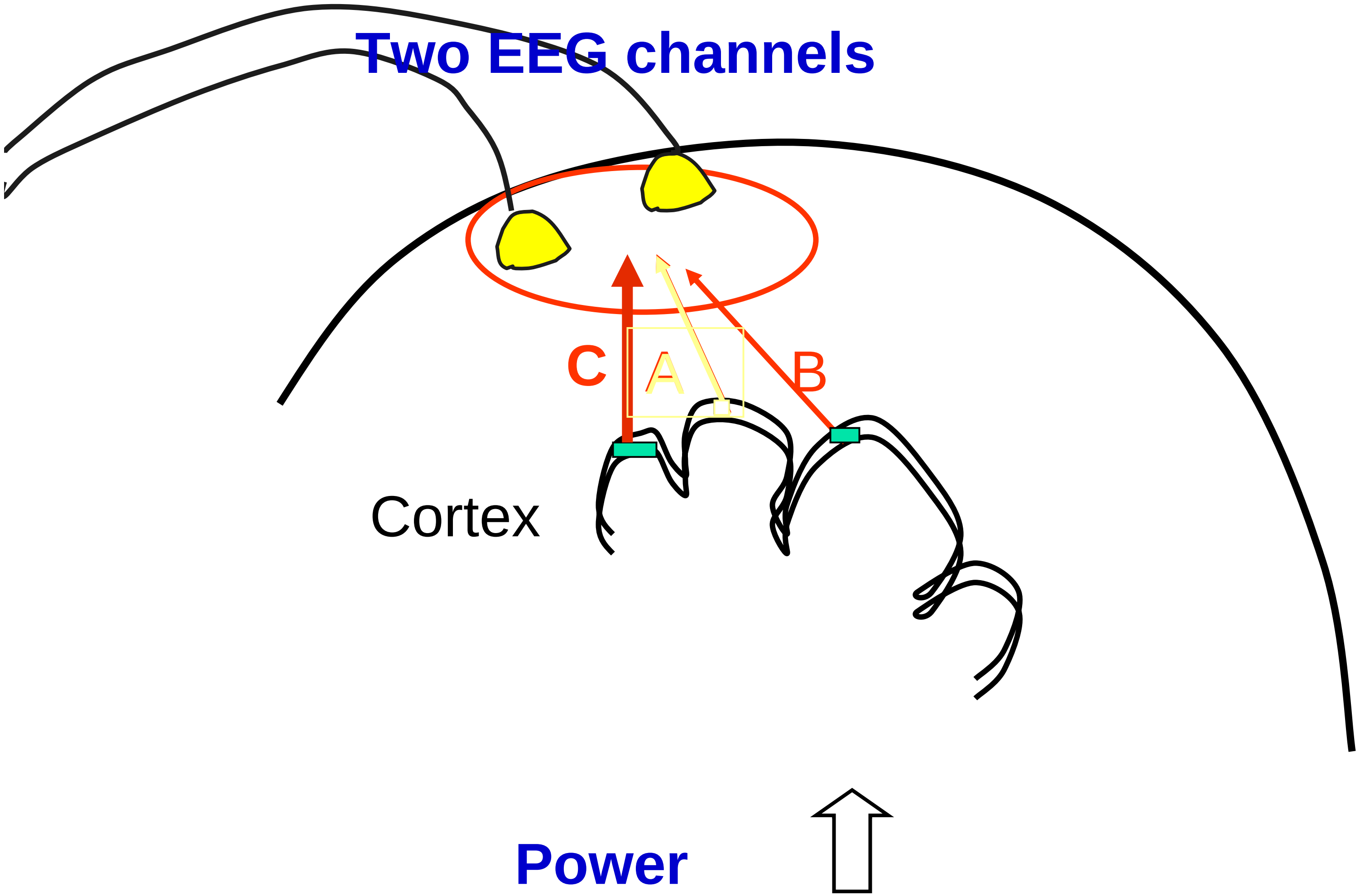



Event–modulated Oscillatory Activity
Alpha component 1
Alpha component 2
Motor–response session
Motor response
No motor response
RT–Sorted Trial Index
Trial Index
μV
Rms
Coh
11 Hz
Time (msec)

Characteristics of Independent Components

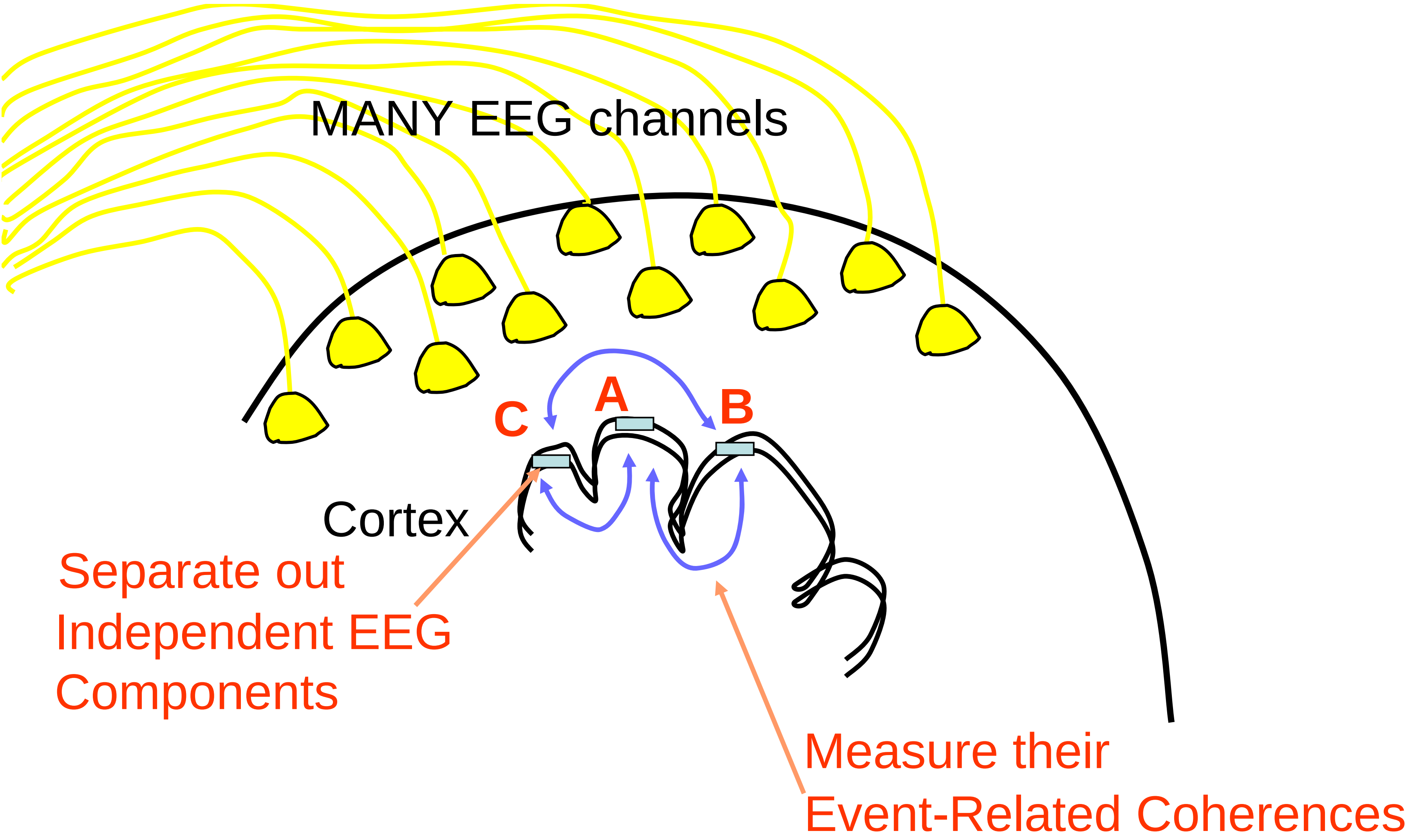
- Concurrent Activity
- Maximally Temporally Independent
- Overlapping Maps and Spectra
- Dipolar Scalp Maps
- Functionally Independent
- Between-Subject Regularity

Do the activities of maximally independent EEG domains interact ?



Scalp channel power changes/coherence

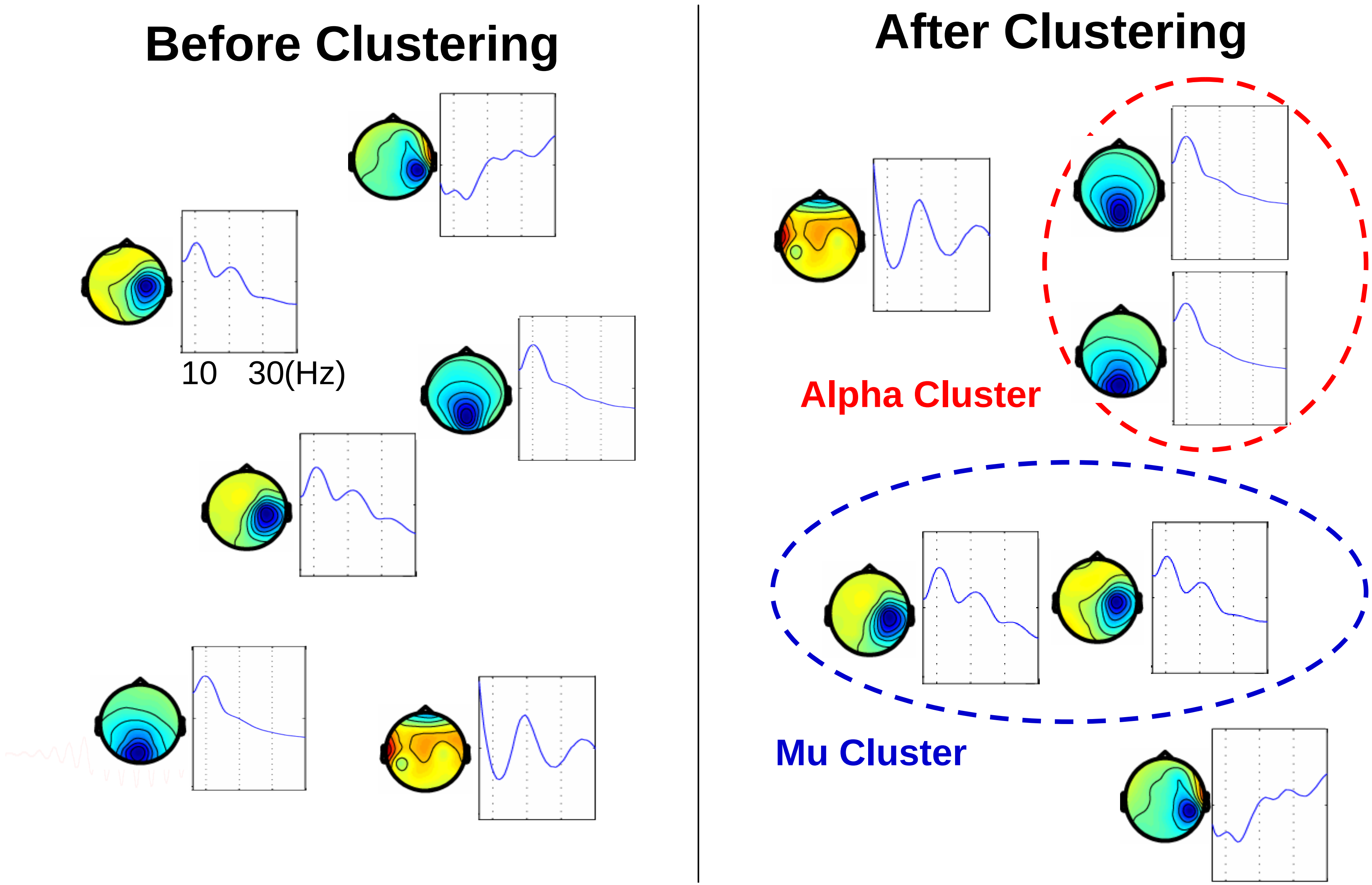
→ source confounds!



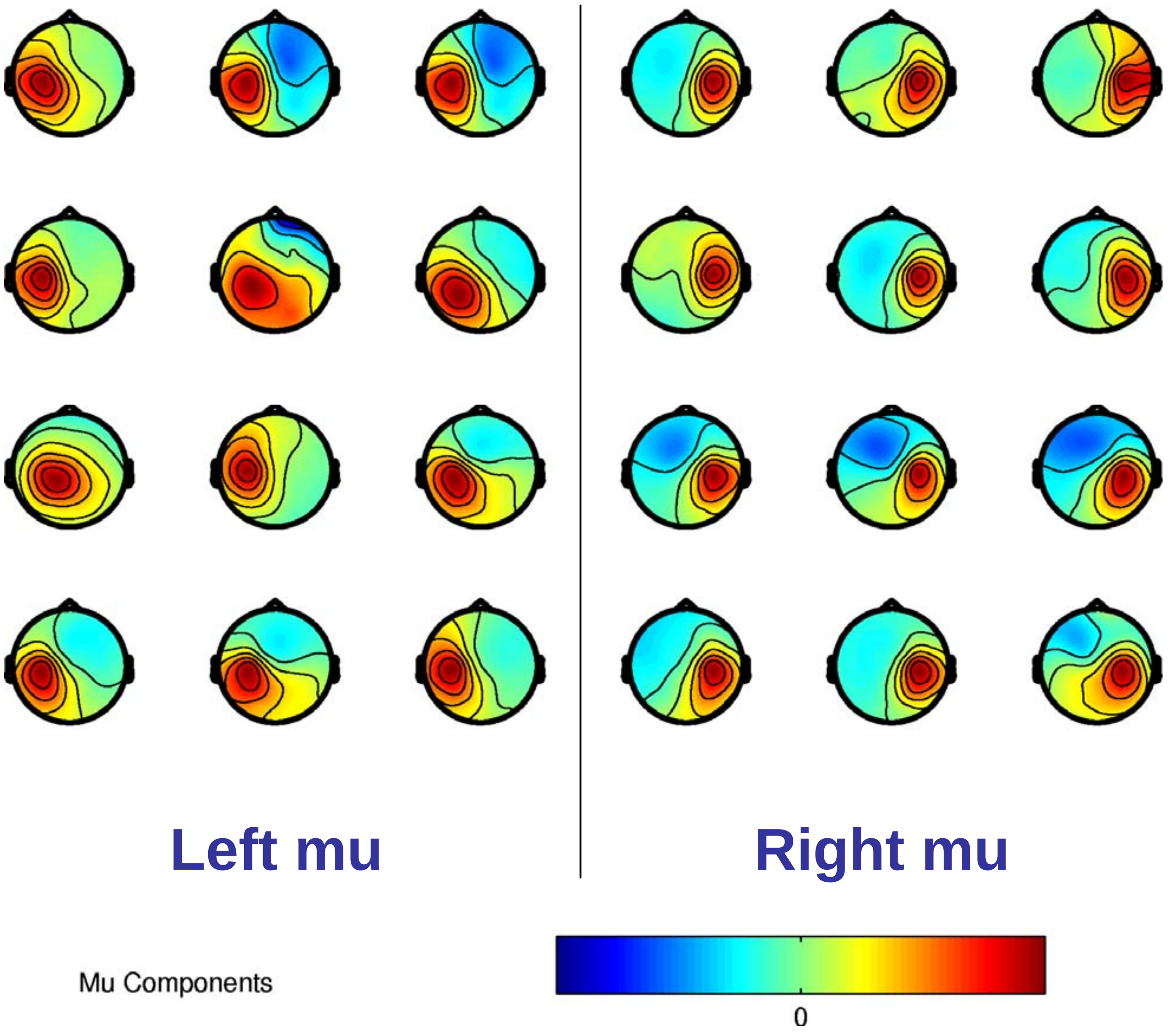
ICA Component coherence → source dynamics!

Does every subject have the same or comparable components while they perform the same tasks?

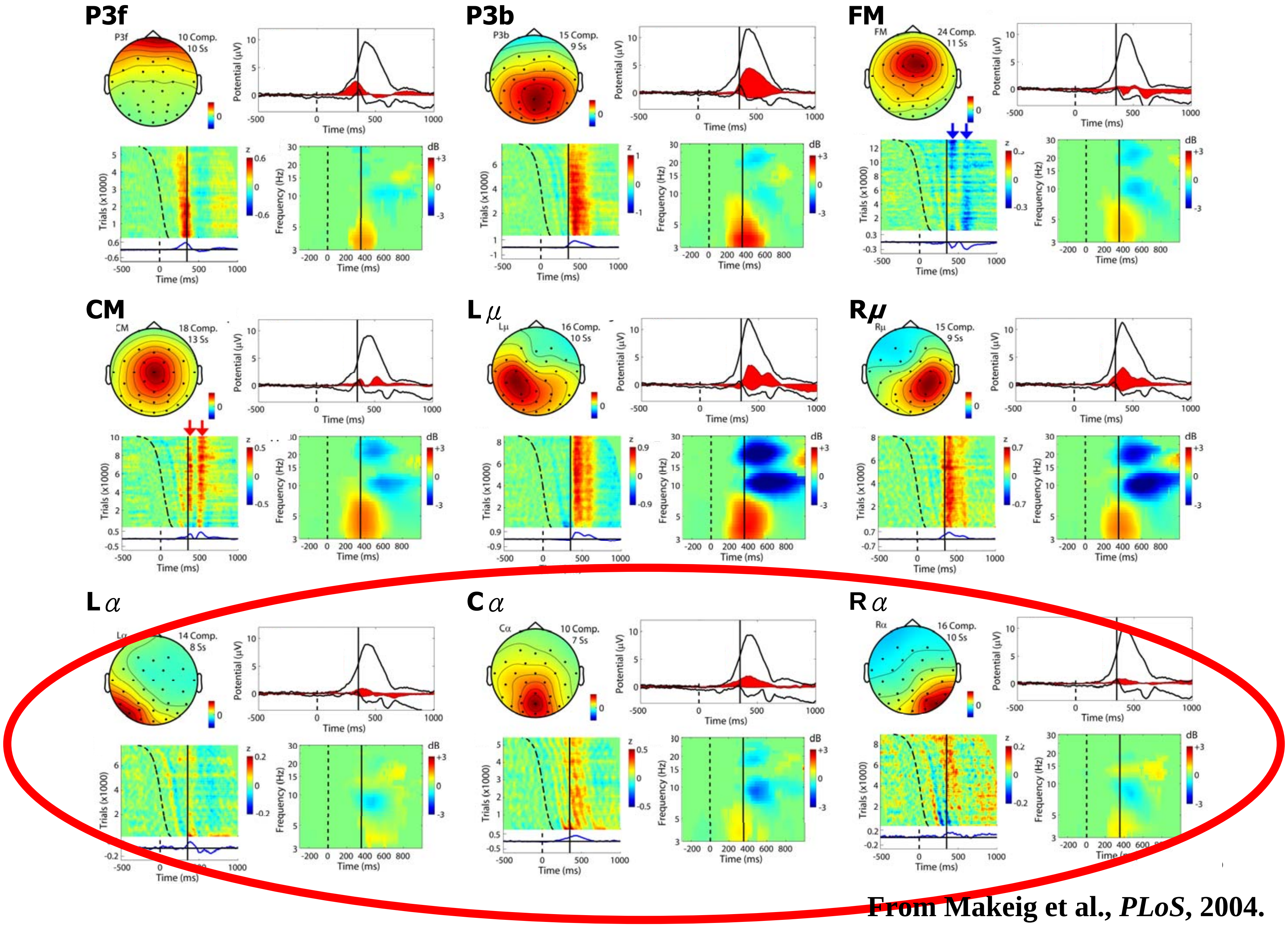
Component Stability: Cross-subject clustering analysis of ICA components



Between-Subject Regularity



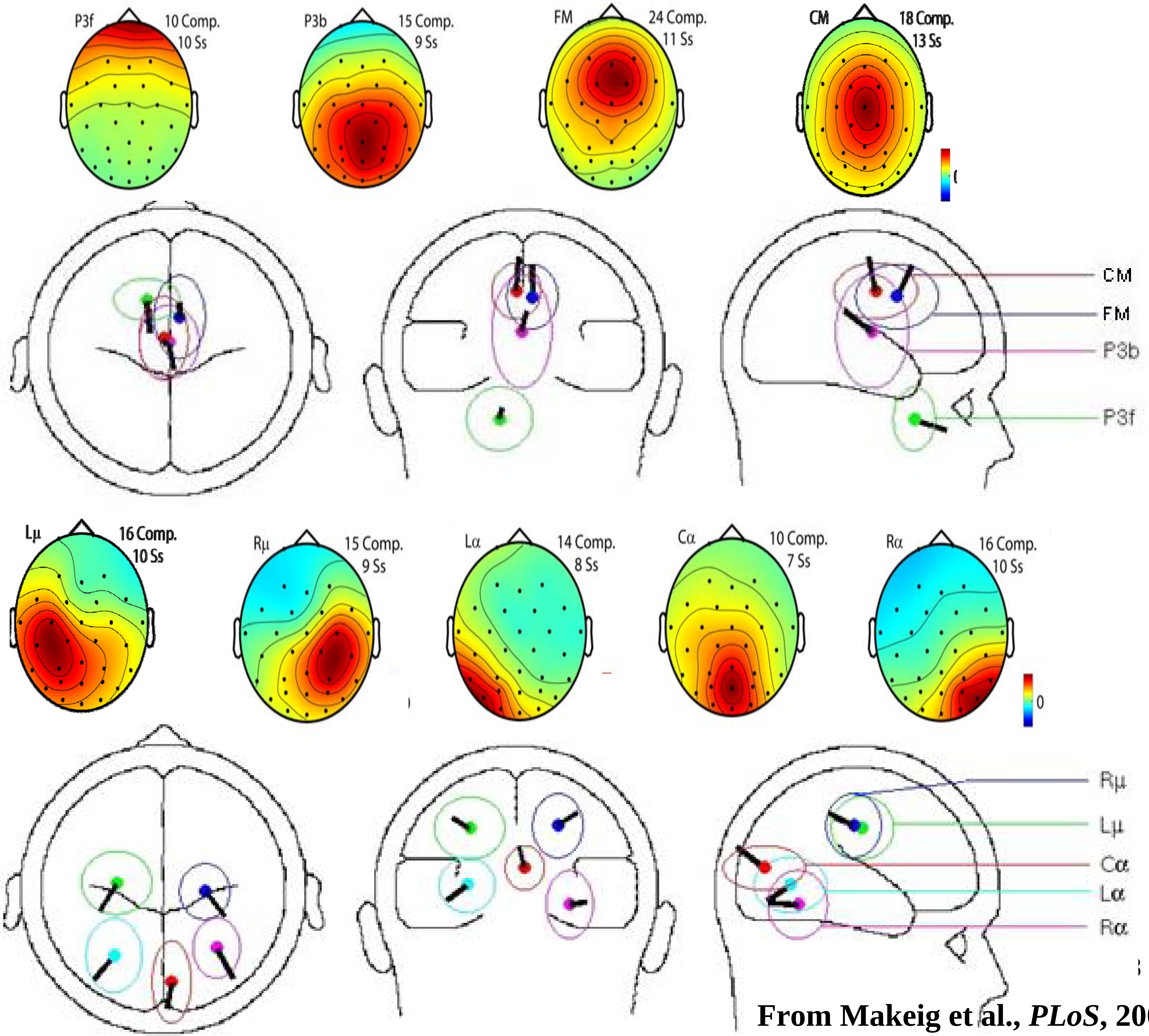
From Jung et al., *Human Brain Mapping*, 2001.



From Makeig et al., *PLoS*, 2004.

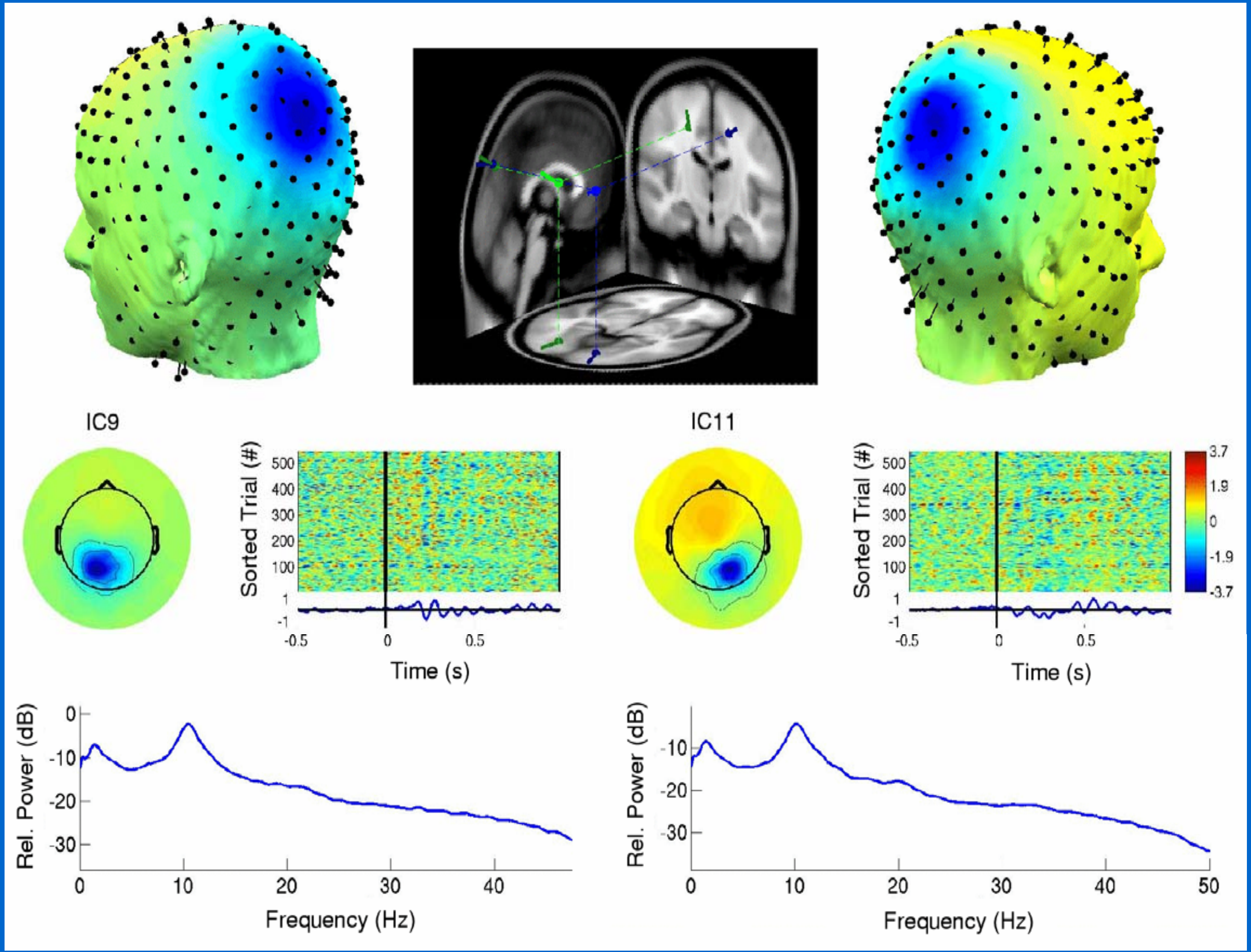
Source Localization

- EEG data collected from any point on the scalp typically includes activity projected by volume conduction from multiple EEG processes in different cortical regions. This has made it difficult to localize the sources of the EEG signals.
- By separating the data into maximally independent *domains of partial synchrony*, ICA identifies scalp maps associated with synchronous field activity in compact domains, separating the question of *source identification* from that of *source localization*.



From Makeig et al., *PLoS*, 2004.

Source Localization



From Makeig et al., *TINS*, 2005.

Balancing Caution with Enthusiasm

Although results of applying ICA to EEG and ERP data have shown great promise and given new insights into event-related brain dynamics, the analysis method is still in its infancy.

The plausibility and reliability of its results should in each case be validated using convergent evidence, typically behavioral and/or other physiological measurements, before interpreting its functional significance.

Summary

- ICA separates high-density EEG (or MEG) data into sources of distinct information in the multidimensional signals.
- ICA reveals WHAT EEG (and artifact) processes are active in the data, building spatial filters that allow:

    (1) their separate activities to be assessed and monitored,

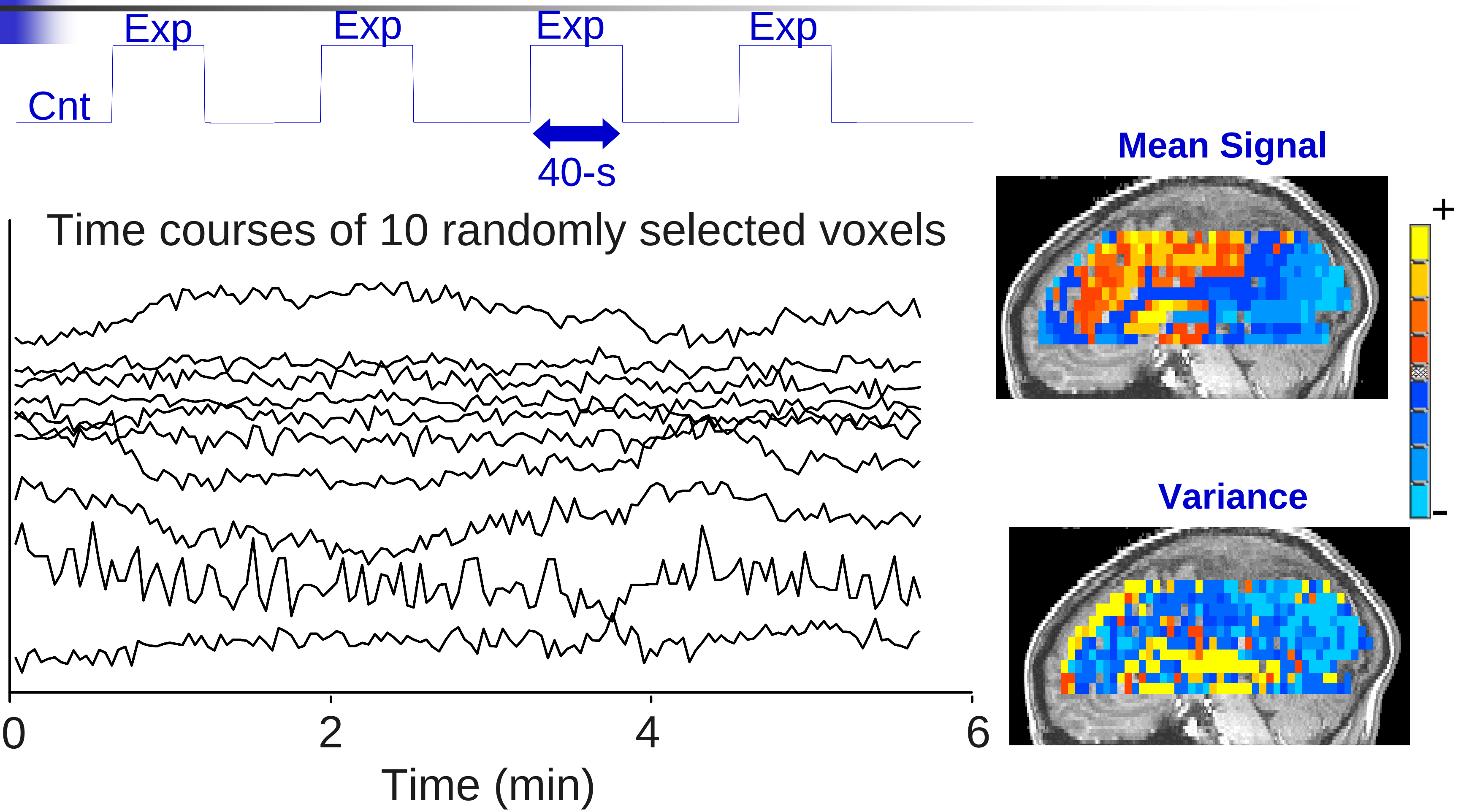
    (2) their separate projections to the scalp sensors to be mapped and inverted with little or no interference from other sources.

    (3) the interactions between multiple brain networks to be investigated.

Magnetic Resonance Imaging (MRI and/or functional MRI)

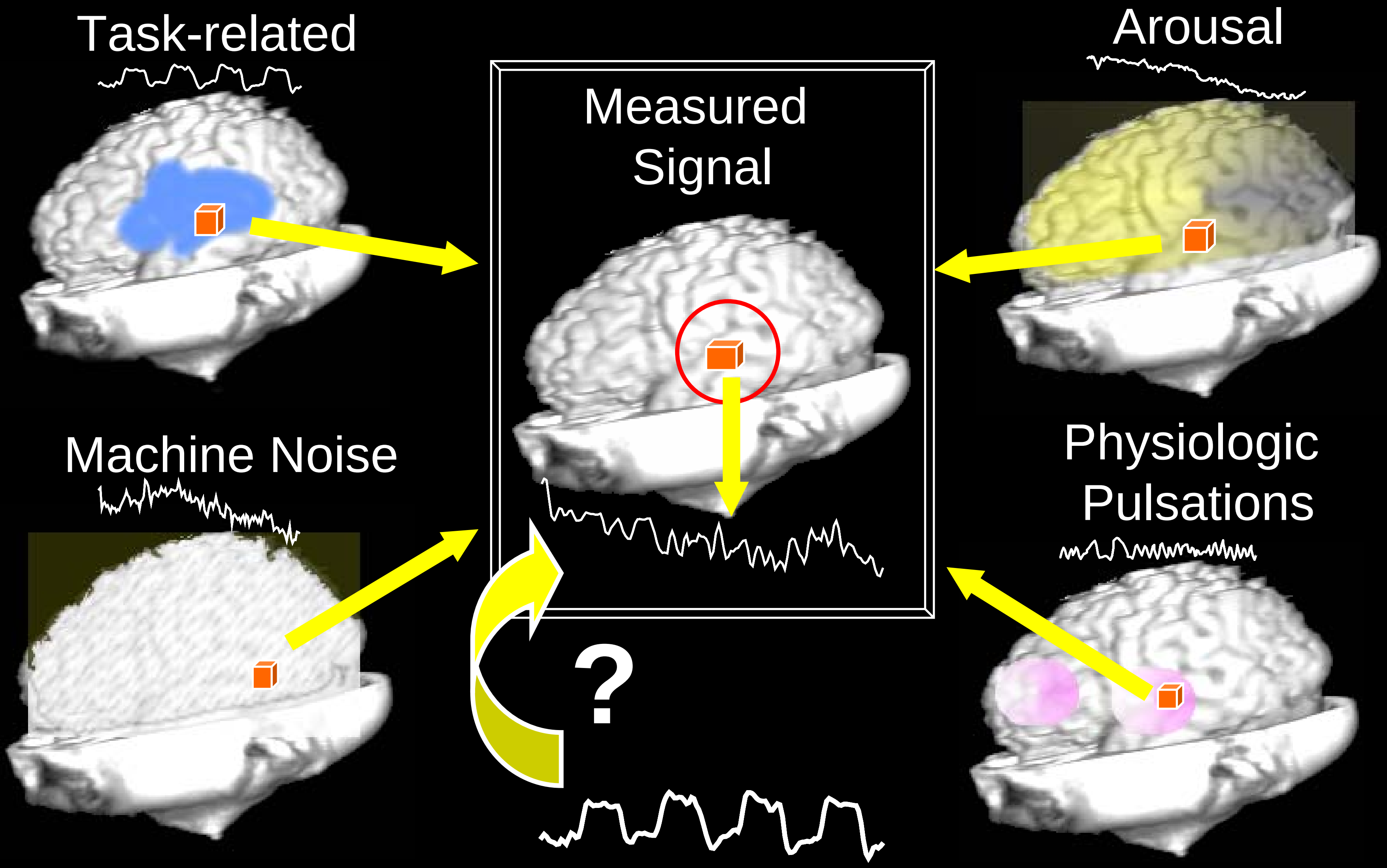
1. MRI is an imaging technique used to produce high quality images of the inside of the human body.
2. It is based on the magnetic susceptibilities of oxygenated hemoglobin (HbO2) and deoxygenated hemoglobin (HbR) to track the blood-flow changes related to neuronal activity, which is referred to as blood-oxygen-level-dependent (BOLD) contrast.

fMRI/BOLD Signal Complexity



McKeown et al., *Human Brain Mapping*, 1998

ICA Applied to fMRI Data



McKeown et al., *Human Brain Map.*, 1998

Analysis of Event-related fMRI Data

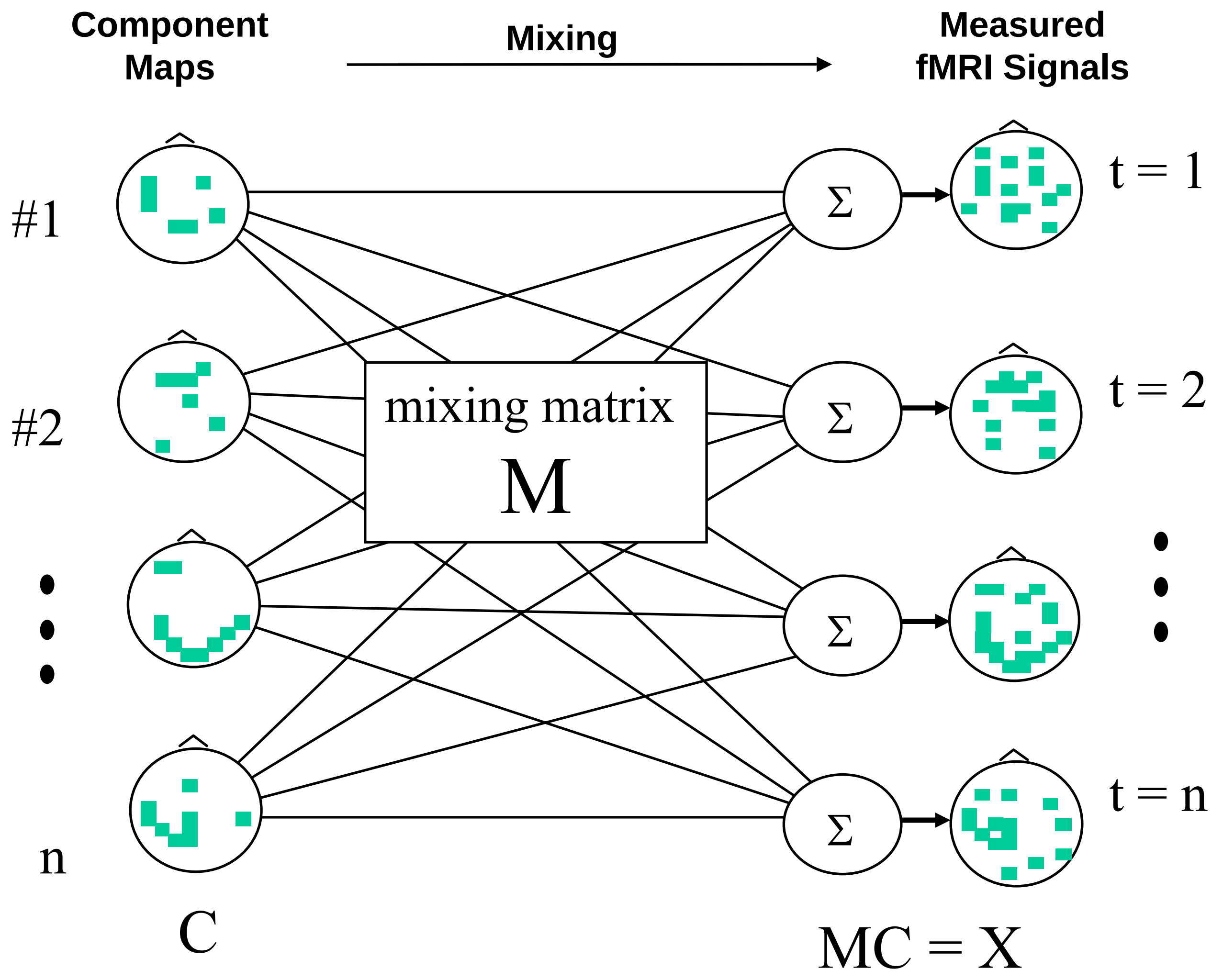
Model-based methods

- **Require** *a priori* knowledge of the time course of the hemodynamic response
- **Assume** homogeneity across different brain regions
- **Allow** tests of statistical significance within an assumed data+noise model.

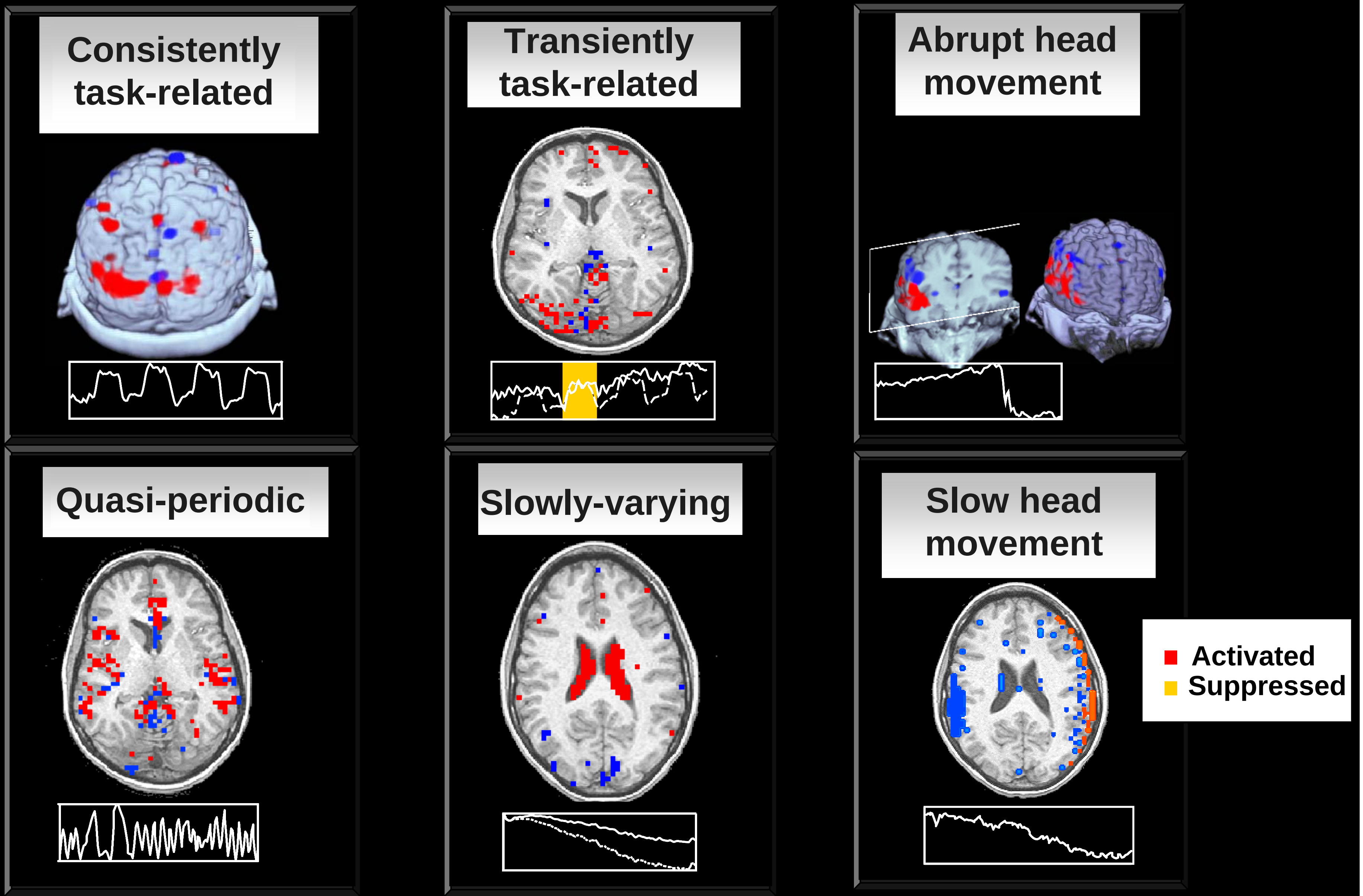
Data-driven methods

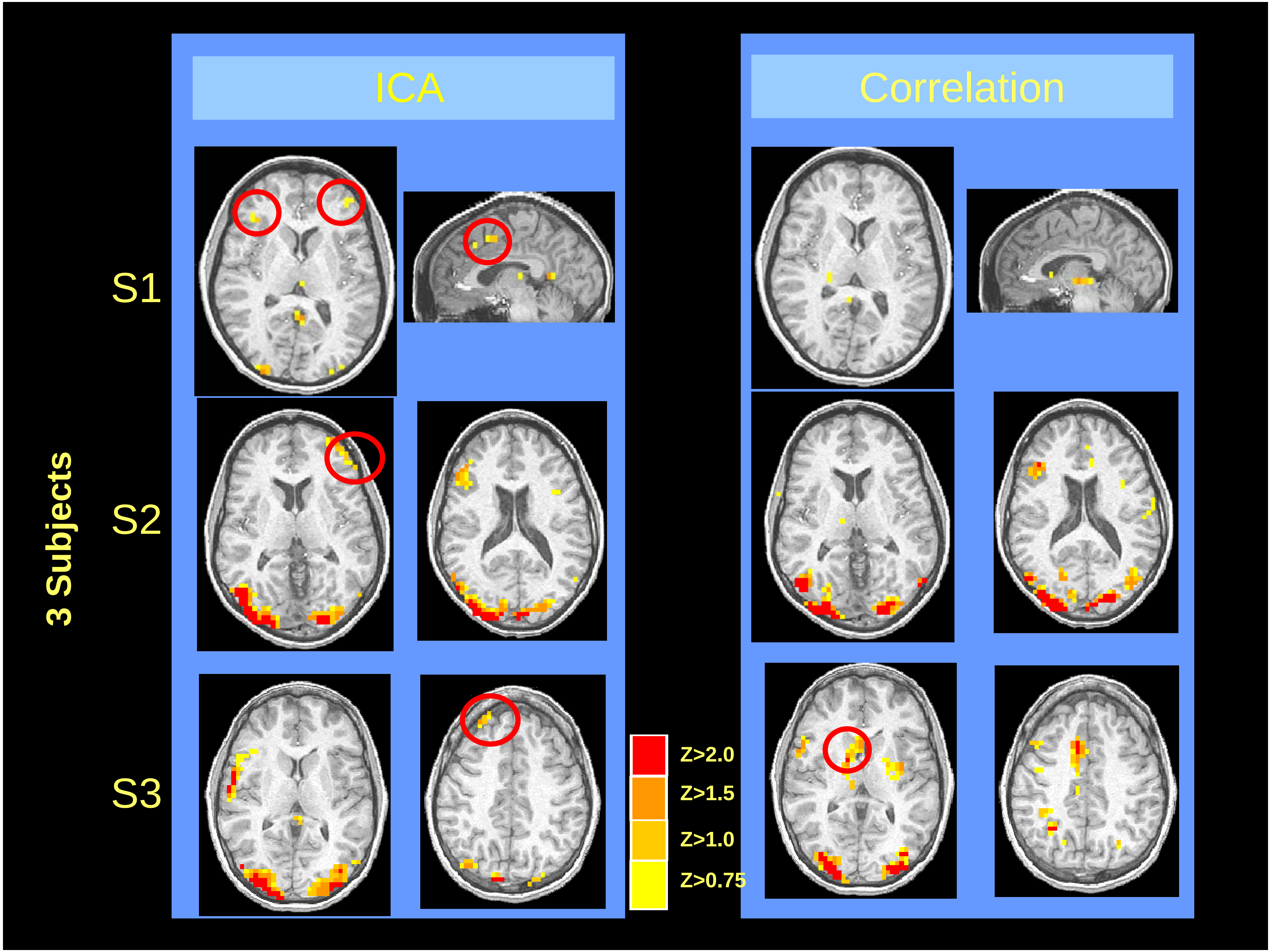
- **Require** minimal space/time assumptions
- **Explore** time courses and spatial distribution of the data
- **Reveal** unforeseen activations (time-varying, site-dependent)
- **Provide no** noise model for statistical testing.

ICA Applied to fMRI Data

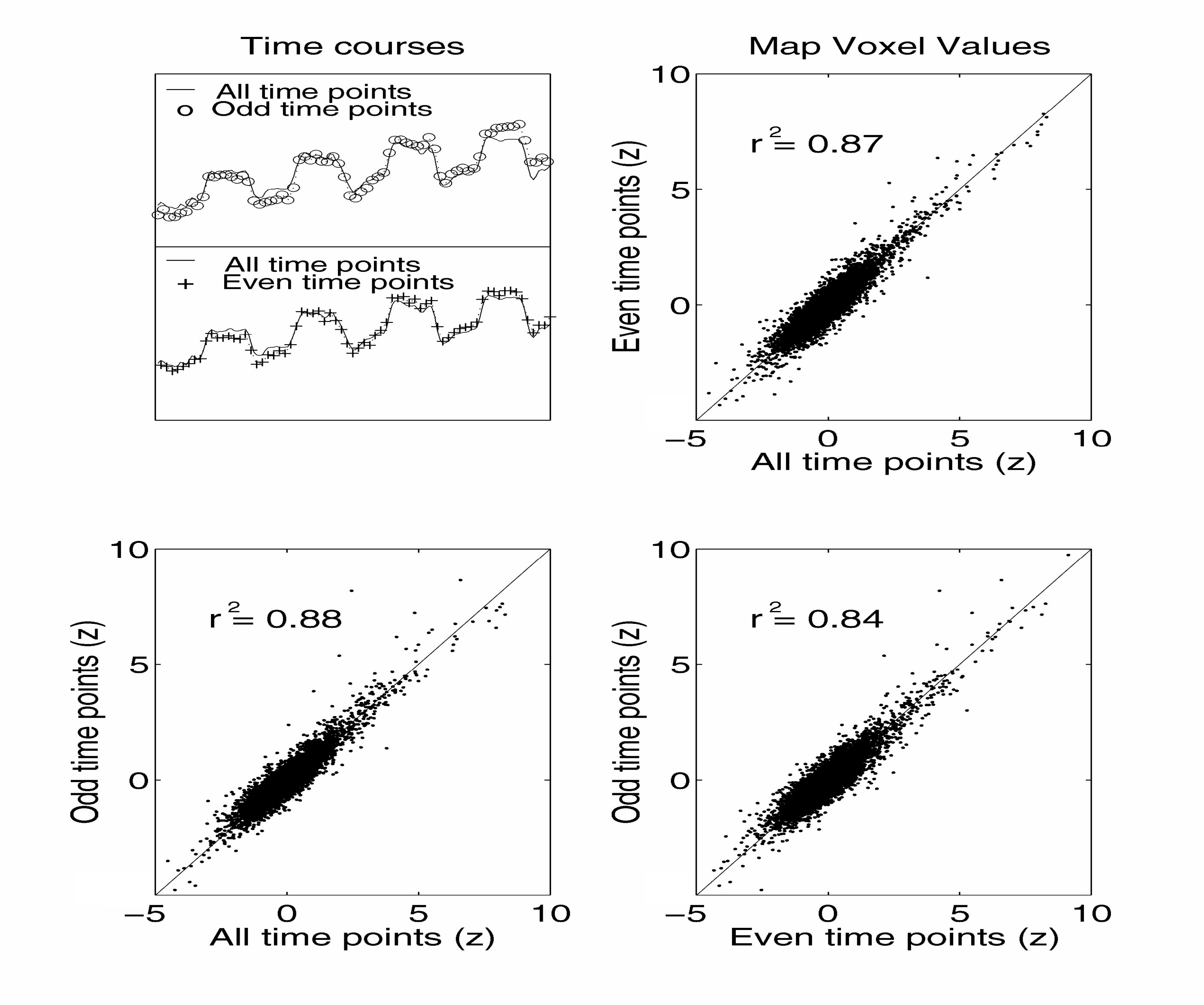


Independent fMRI Components




ICA
Correlation
3 Subjects
S1
S2
S3
Z>2.0
Z>1.5
Z>1.0
Z>0.75

Stability of ICA Component Maps



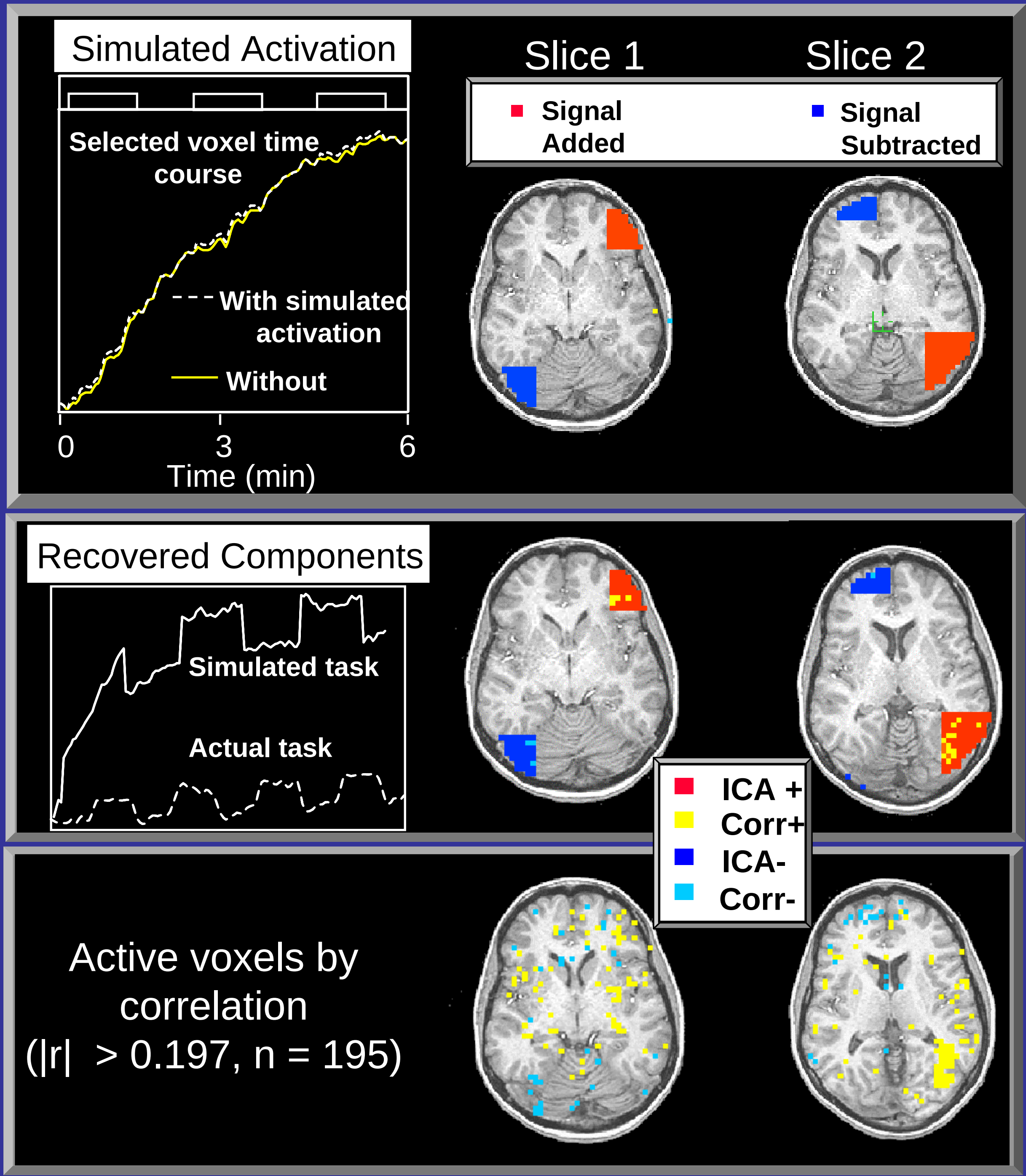
Relative Task Sensitivity

ICA

vs

Correlation

(Simulation)



Conclusions

- ICA has proven successful in many data-analysis applications.
- Great care must be taken to examine the validity of the assumptions that are used by ICA to derive a decomposition of the observed signals and/or to evaluate the reliability and functional significance of the resulting components.