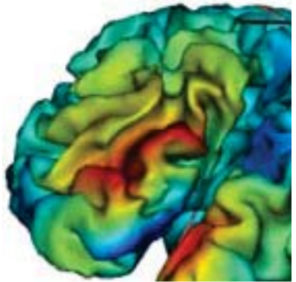
Sparse methods for functional brain imaging

Alexandre Gramfort
alexandre.gramfort@telecom-paristech.fr

Telecom ParisTech
INRIA Parietal Project Team
CEA - Neurospin, France

Workshop Sparse Models and Machine Learning
IRISA - Oct. 16 2012



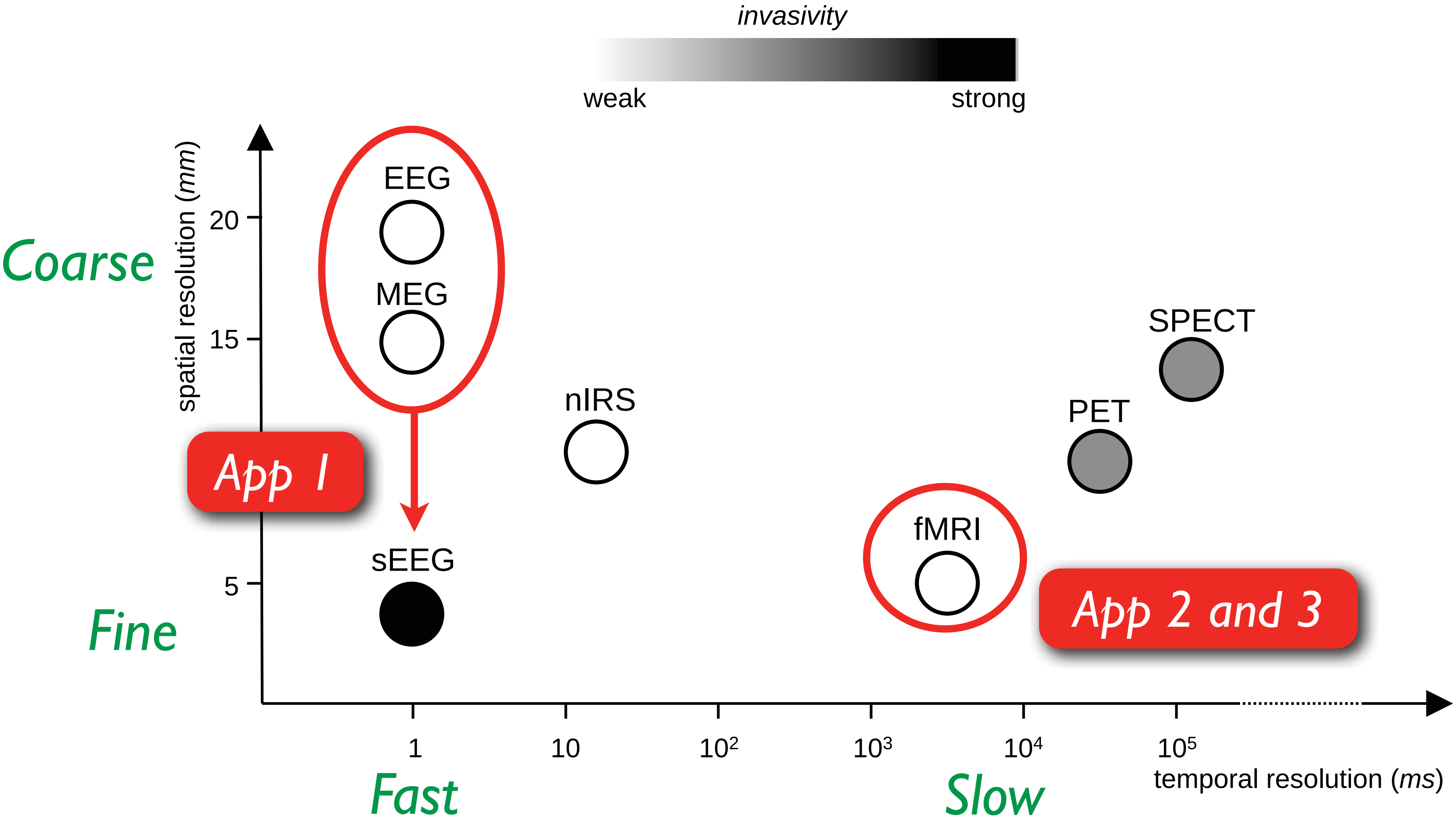
Outline: 3 good "sparse" problems

- **Brain imaging with MEG and EEG (M/EEG)**
  - Background on M/EEG (physiology and physics)
  - The inverse problem: regression with sparse structured priors using time-frequency (TF) dictionaries
- **"Brain reading" with functional MRI (fMRI)**
  - Prediction vs. recovery
  - Support recovery with correlated design?
- **Network and atlas learning with resting state fMRI**
  - Sparse covariance estimation and dictionary learning

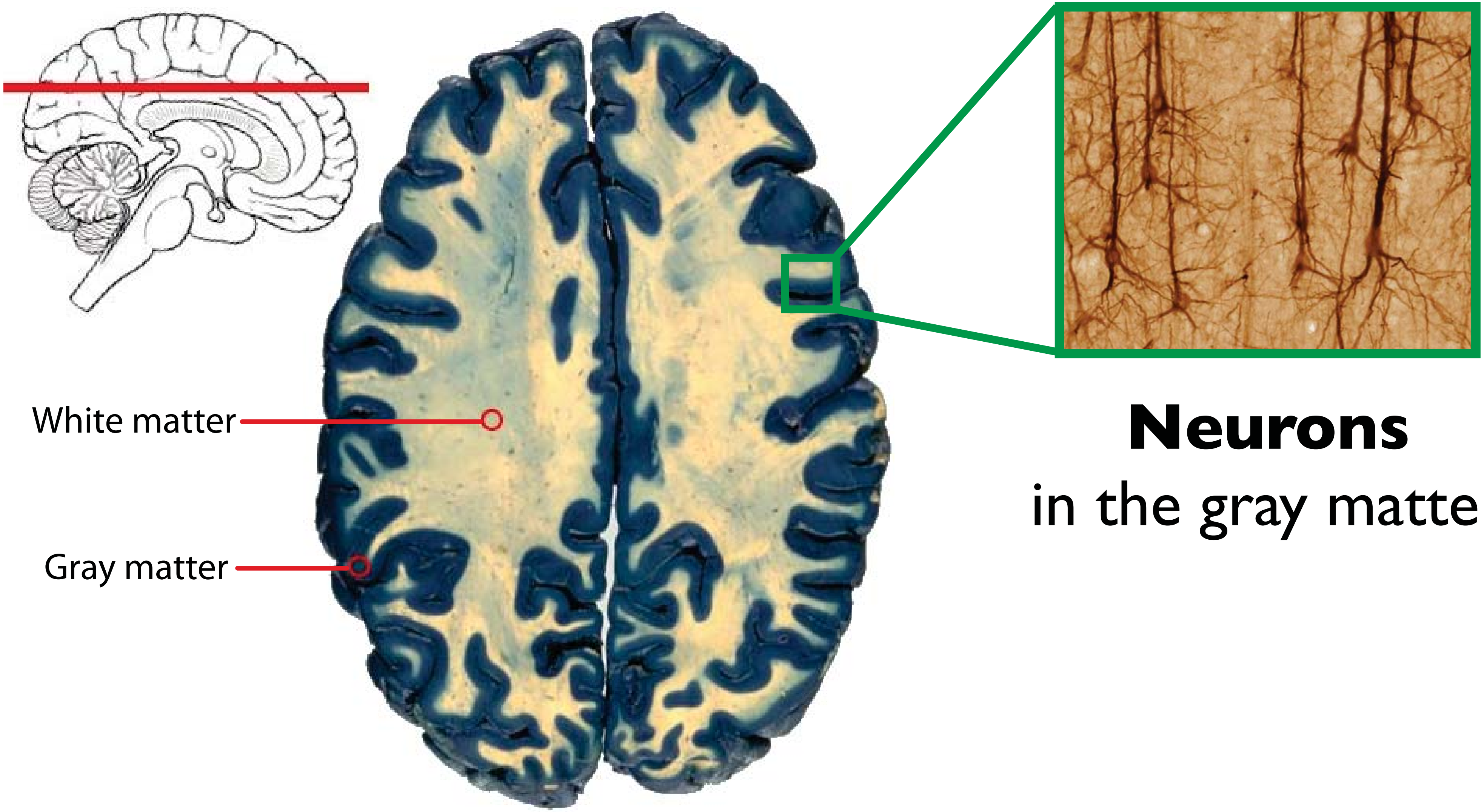
THM: Means «Take Home Message»

Background on M/EEG

Functional neuroimaging



Brain anatomy



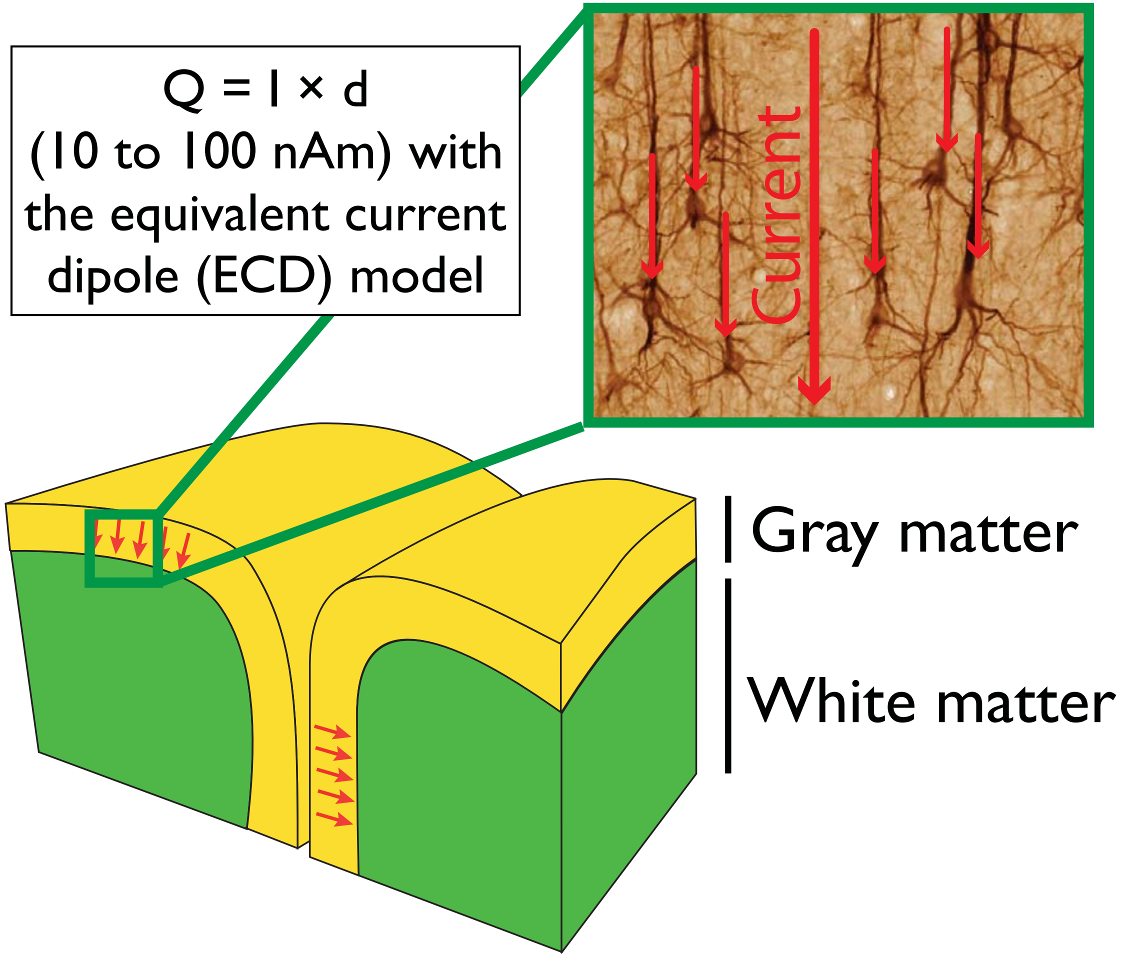
Neurons
in the gray matter

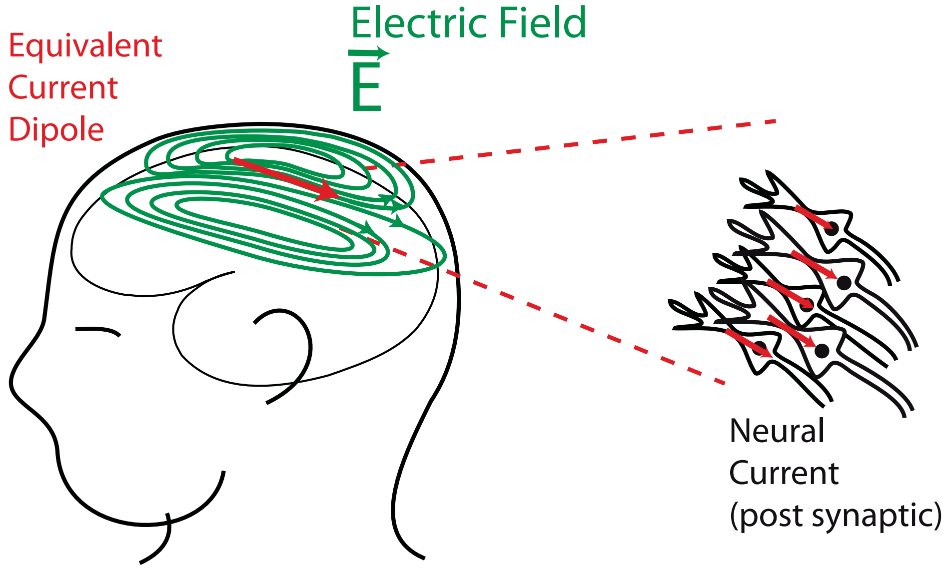
Source: dartmouth.edu

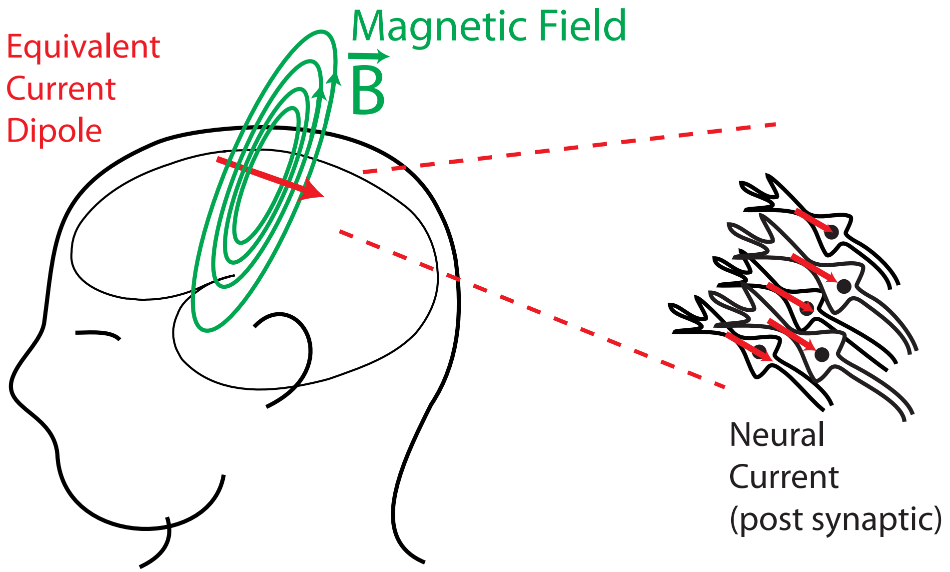
Neurons as current generators

Large cortical pyramidal cells organized in macro-assemblies with their **dendrites normally oriented to the local cortical surface**

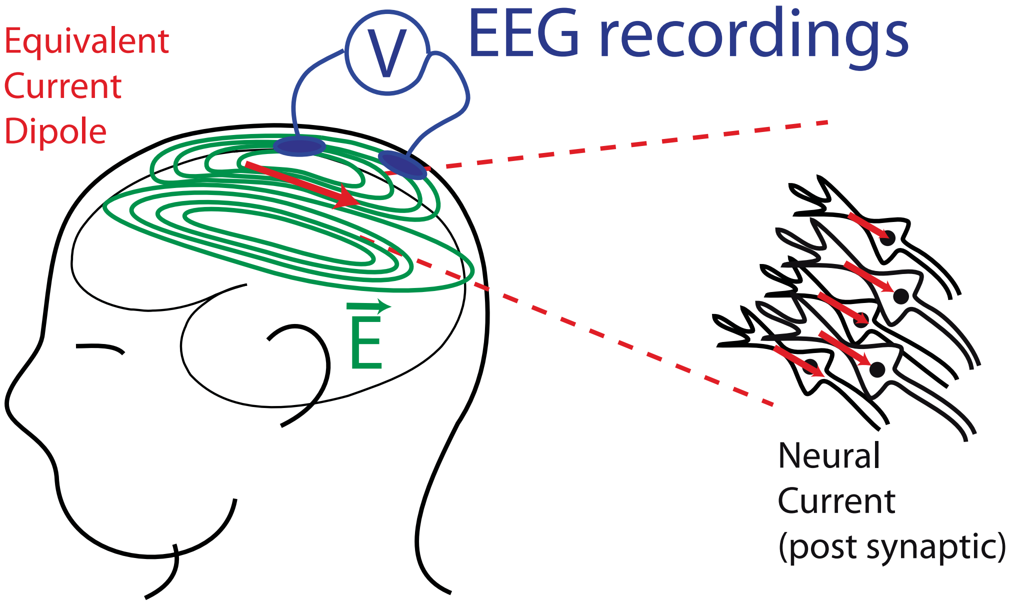
$Q = I \times d$
(10 to 100 nAm) with
the equivalent current
dipole (ECD) model

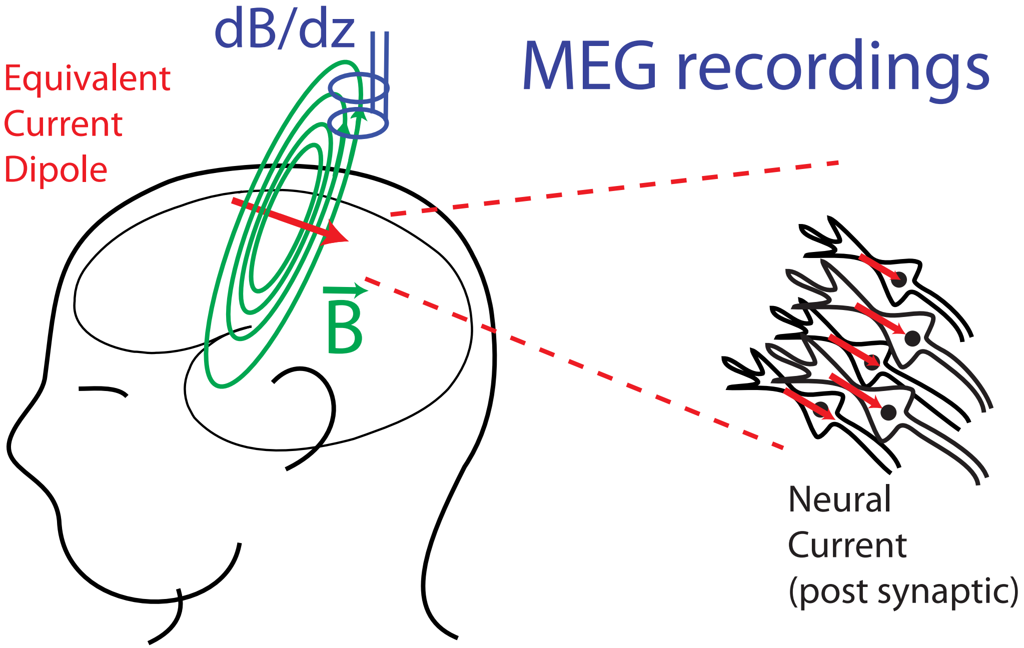




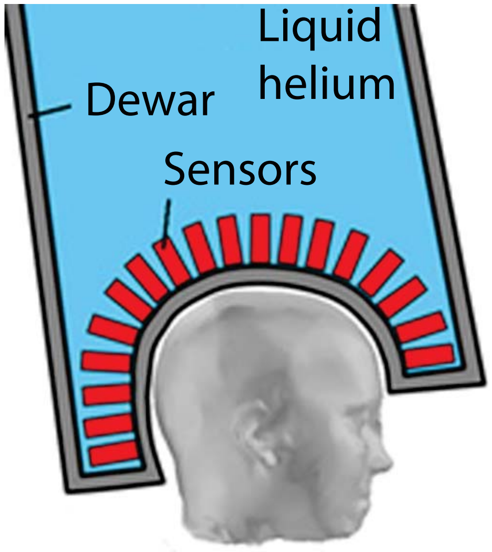


EEG & MEG systems



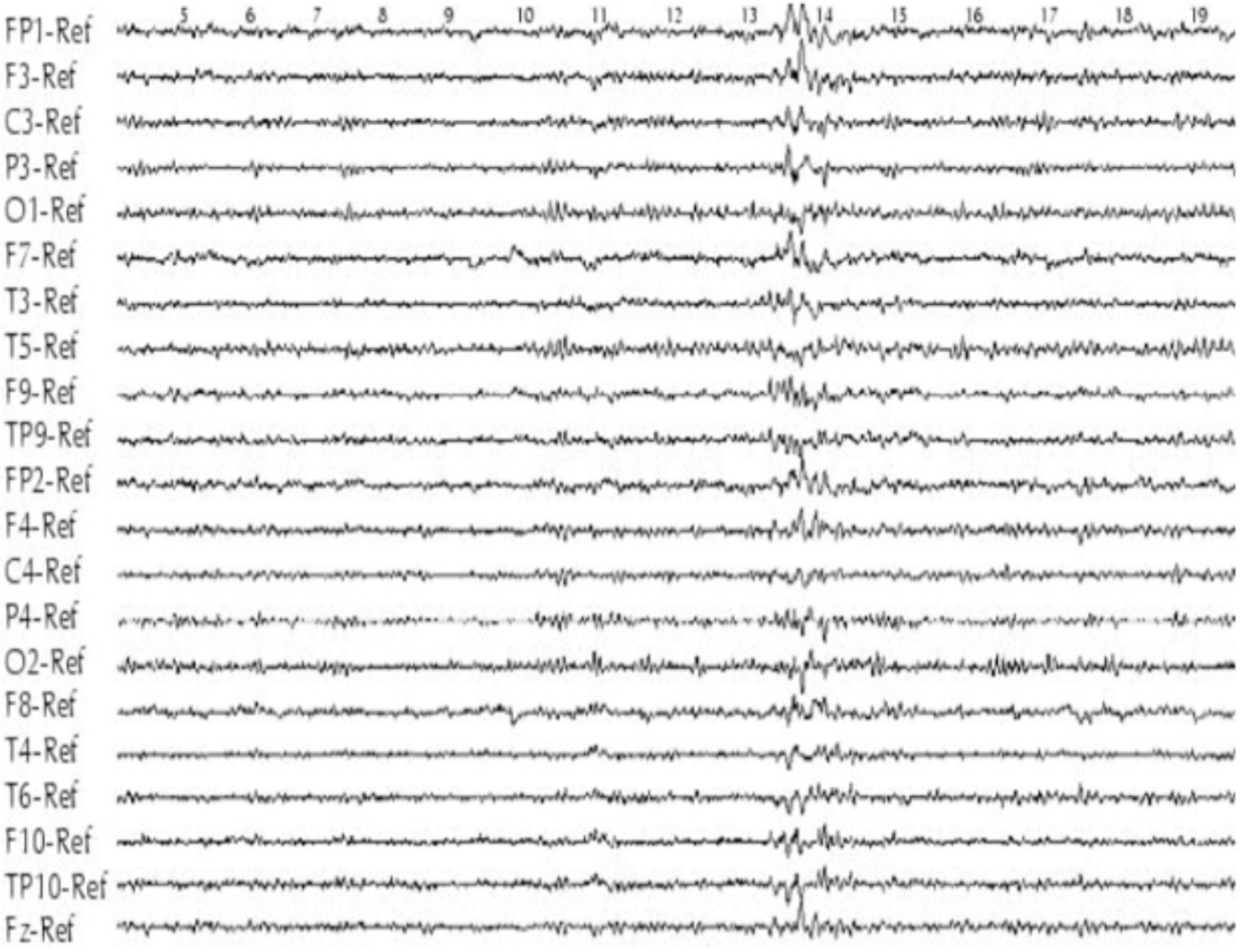


First EEG recordings in 1929 by H. Berger



Hôpital La Timone
Marseille, France

M/EEG Measurements



Sample EEG measurements

EEG :

- ≈ 32 to 100 sensors

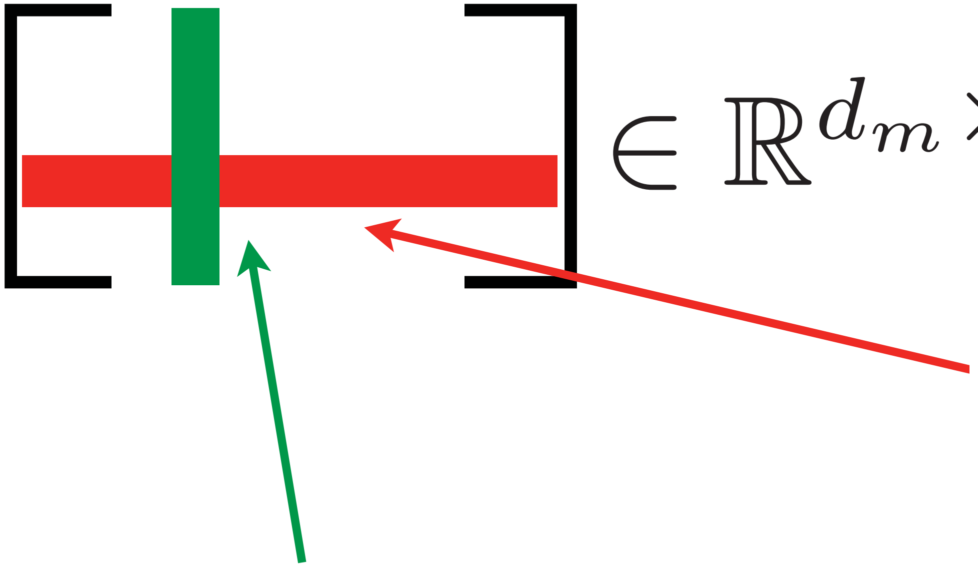
MEG :

- ≈ 150 to 300 sensors

Sampling between 250 and1000 Hz

High temporal resolution

M/EEG Measurements: Notation



M =

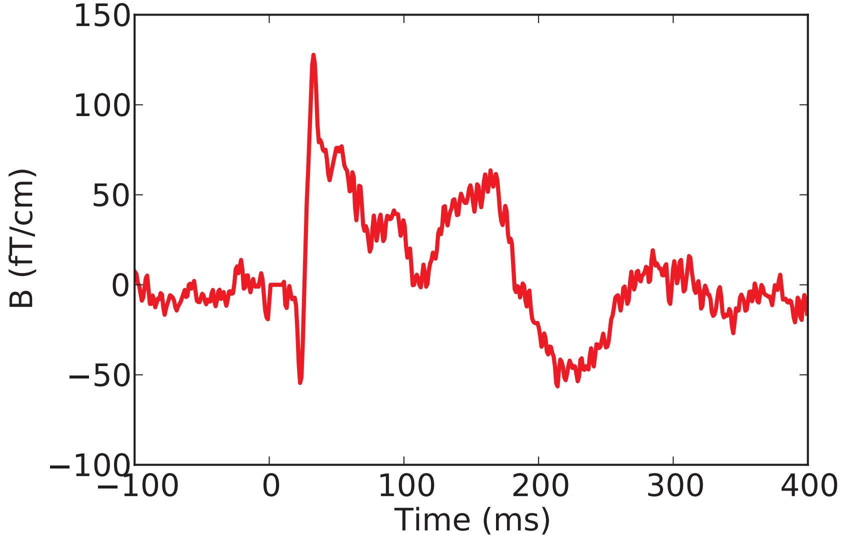
MEG
and/or
EEG

$$\in \mathbb{R}^{d_m \times d_t}$$

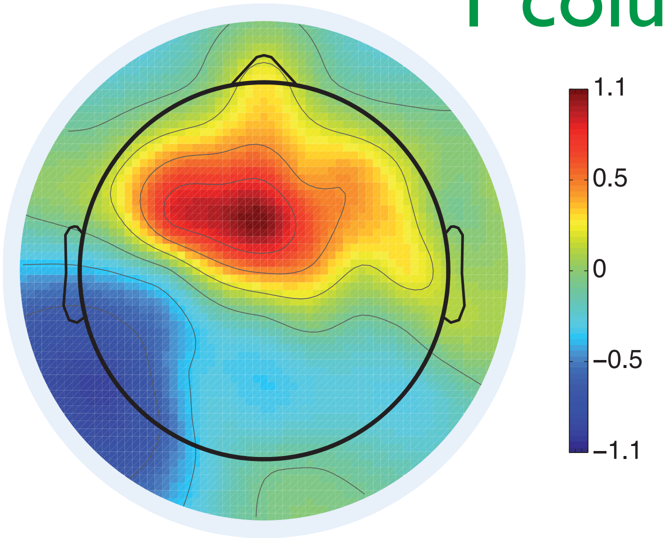
d_m : Nb of sensors

d_t : Nb of time points

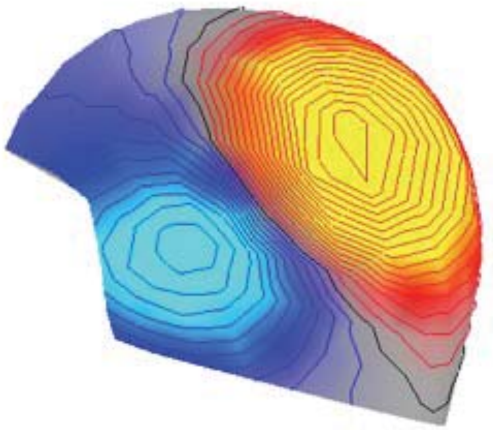
1 column



1 row = 1 time series
on 1 sensor



2D topography



3D topography

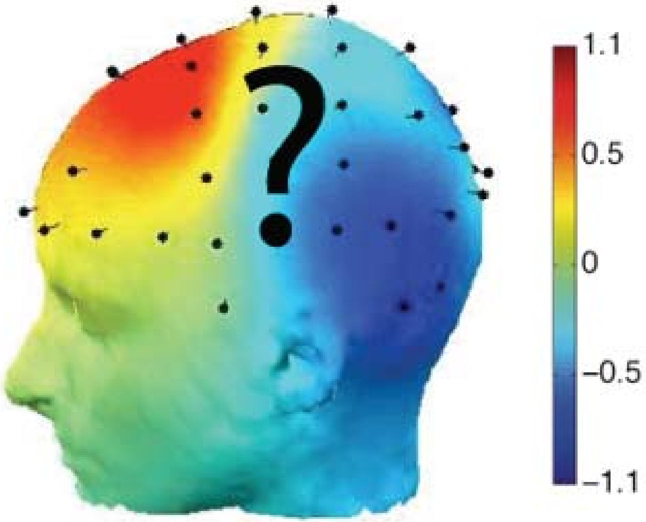
The M/EEG inverse problem with structured sparse priors and time-frequency dictionaries

[Gramfort et al., Physics in Medicine and Biology 2012]
[Gramfort et al., IPMI 2011]
[Gramfort et al., submitted]

collaboration with Strohmeier D., Haueisen J., Hämäläinen M., Kowalski M.

Inverse problem: Objective

Find the current generators that produced the M/EEG measurements



Linear forward problem: Maxwell

Maxwell Equations
with **quasi-static**
approximation

$$\begin{cases} \nabla \times \vec{E} = & 0 \\ \nabla \cdot \vec{B} = & 0 \\ \nabla \times \vec{B} = & \mu_0 \vec{J} \\ \nabla \cdot \vec{E} = & \dfrac{\rho}{\epsilon_0} \end{cases}$$

Remark: quasi-static implies no temporal derivatives and no propagation delay

Total currents: $\vec{J} = \vec{J}_p + \vec{J}_c$

Primary currents

Conduction currents

Ohm's law: $\vec{J}_c = -\sigma \nabla V$

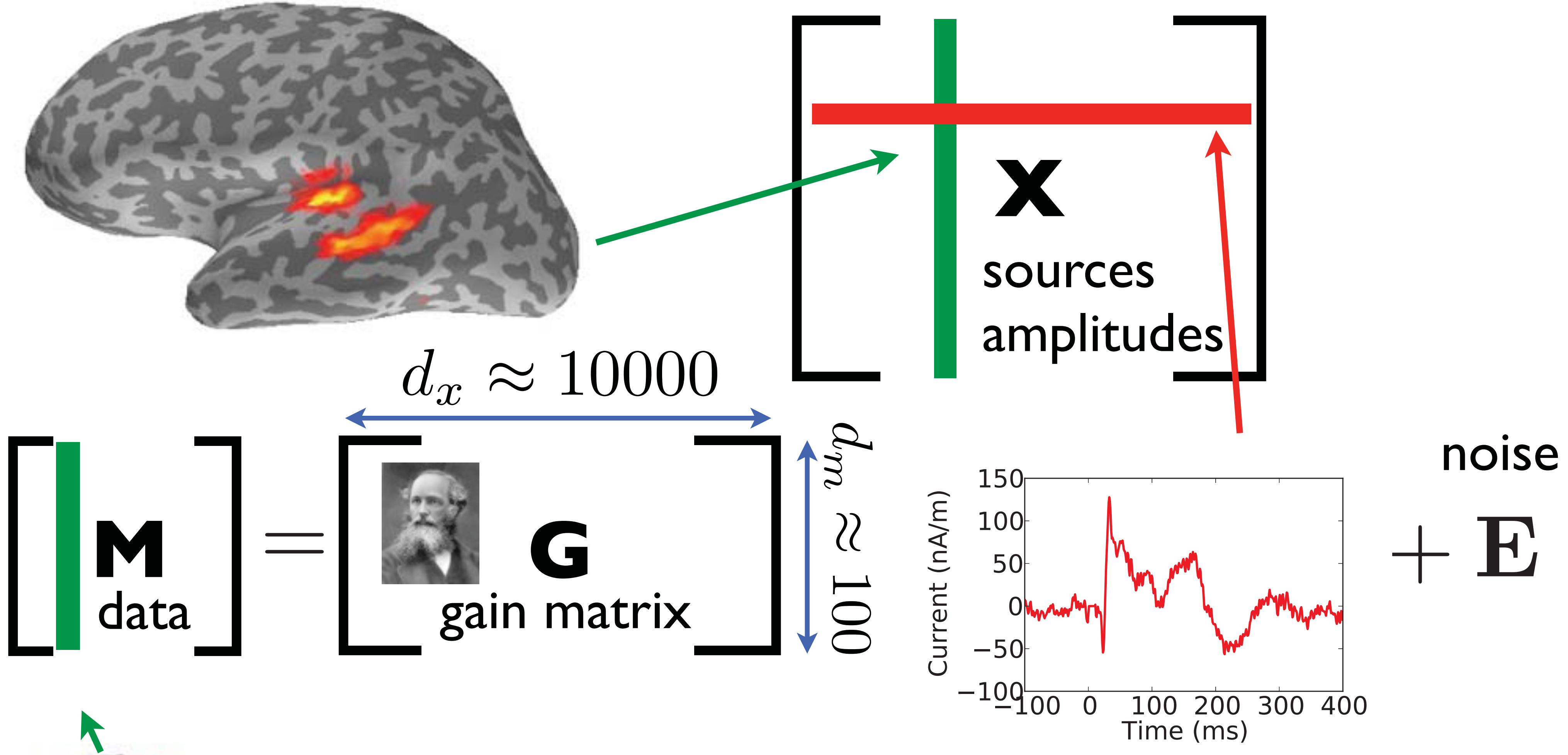
V Electric potential
σ Tissue conductivity

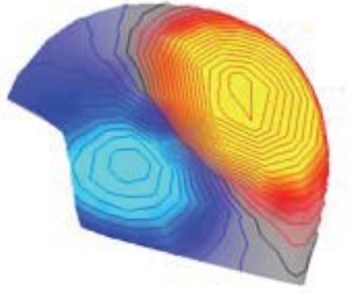
Potential equation:
(relation btw. the potential and the sources)

$$\nabla \cdot \nabla \times \vec{B} = 0 \Rightarrow \nabla \cdot (\vec{J}_s + \vec{J}_c) = 0$$

$$\Rightarrow \nabla \cdot \vec{J}_p = \nabla \cdot (\sigma \nabla V)$$

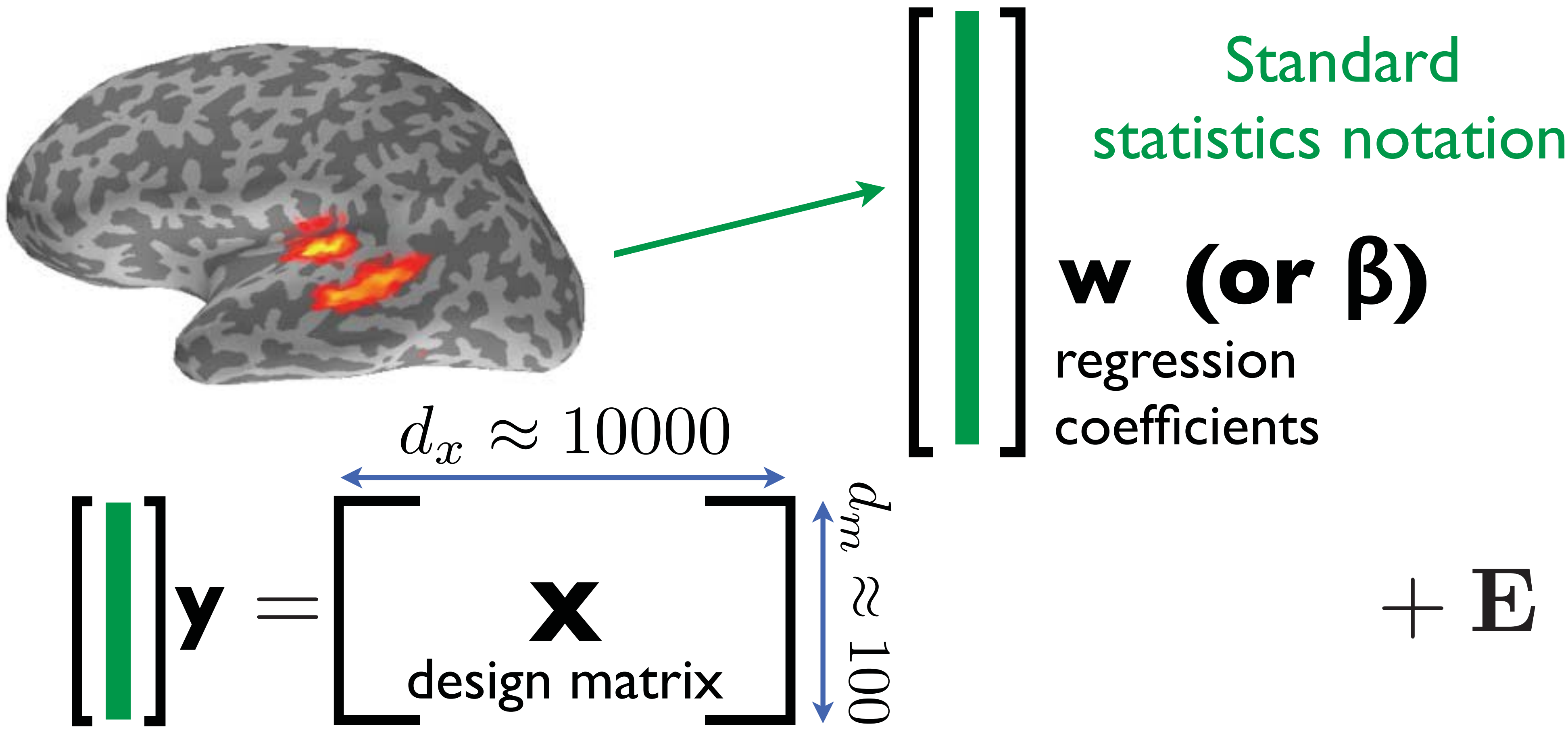
M=GX+E : An ill-posed problem





THM: Following **Maxwell's equations** each source adds its contribution **linearly**

y = Xw+E : An ill-posed problem



Standard statistics notations

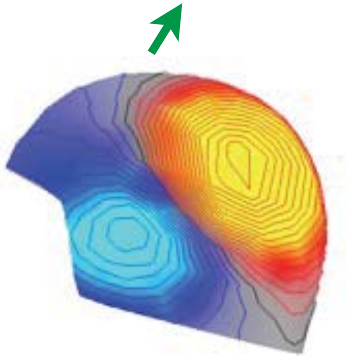
w (or β)
regression coefficients

$d_x \approx 10000$

$d_m \approx 100$

$$\mathbf{y} = \mathbf{X} \; + \mathbf{E}$$

design matrix



THM: At **each time instant** the M/EEG inverse problem **IS** a **regression** with more variables than observations

Inverse problem framework

Penalized (variational) formulation (with whitened data):

$$\mathbf{X}^* = \underbrace{\arg\min_{\mathbf{X}} \|\mathbf{M} - \mathbf{G}\mathbf{X}\|_F^2}_{\text{Data fit}} + \underbrace{\lambda\phi(\mathbf{X})}_{\text{Prior}}, \lambda > 0$$

λ : Trade-off between the data fit and the prior

where $\|\mathbf{A}\|_F^2 = \mathbf{tr}(\mathbf{A}^T\mathbf{A})$

$\phi(\mathbf{X})$ is the prior.

Examples for $\phi(\mathbf{X})$: ℓ_1, ℓ_2, Total-Variation ...

THM: when SNR goes UP λ goes DOWN.

L2 a.k.a. Minimum Norm Estimates (MNE)

$$\phi(\mathbf{X}) = \|\mathbf{W}\mathbf{X}\|_F^2 = \sum_{i,j} w_i^2 x_{ij}^2 = \|\mathbf{X}\|_{\mathbf{\Sigma},2}^2$$

$$\mathbf{W}^2 = \mathbf{\Sigma}$$ *source covariance*

Leads to a **closed form solution** (matrix multiplication):

$$\mathbf{X}^* = \mathbf{\Sigma}^{-1}\mathbf{G}^T(\mathbf{G}\mathbf{\Sigma}^{-1}\mathbf{G}^T + \lambda \mathbf{Id})^{-1}\mathbf{M}$$

[Tikhonov et al. 77, Wang et al. 92, Hämäläinen et al. 94]

Remarks:

- **MNE** is known as **Ridge regression** in statistics.
- **Really fast** to compute (SVD of **G**), hence very much used in the field.
- In practice, it's **much more complicated** (whitening data, correcting artifacts, channels with different SNRs, setting λ based on SNR, loose orientation, ...)

THM: A lot of domain knowledge to make it work

Mixed-Norm Estimates (MxNE) & sparse priors

Why sparse priors?

- **M/EEG data** are commonly assumed to be **produced** by a **few brain regions** (justifies the use of multi-dipole fits)
- **Activations** have **small spatial extents** w.r.t. meas. distance

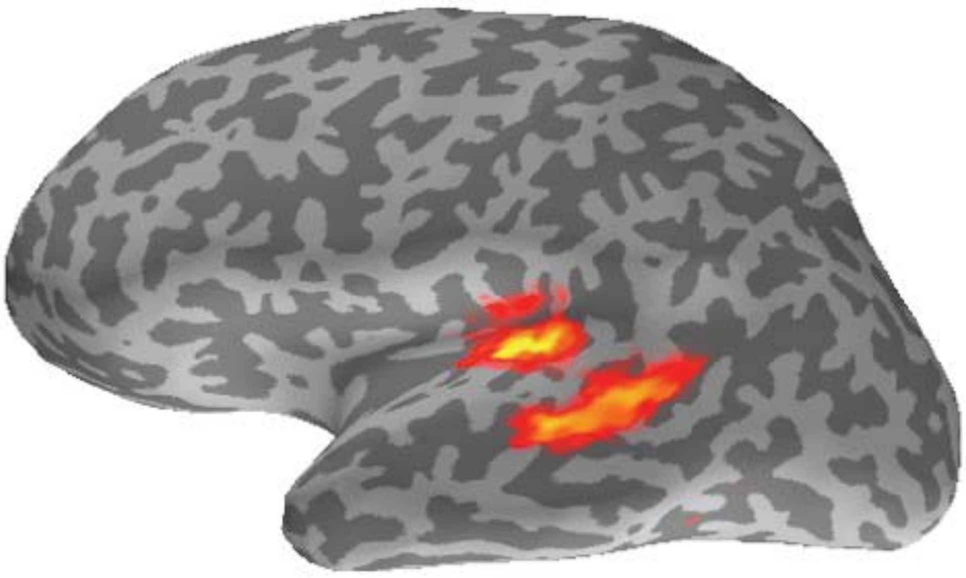
Brief history of contributions up to now:

- [MCE 95, Focuss 95] : single instant solvers (not adapted)
- [Nummenmaa 2007, Wipf 2009, Friston (MSP) 2009] : Bayesian methods based on automatic relevance determination (ARD)
- [Haufe 2008, Ou 2009] convex mixed-norm prior but uses a very slow SOCP solver (*sedumi*).

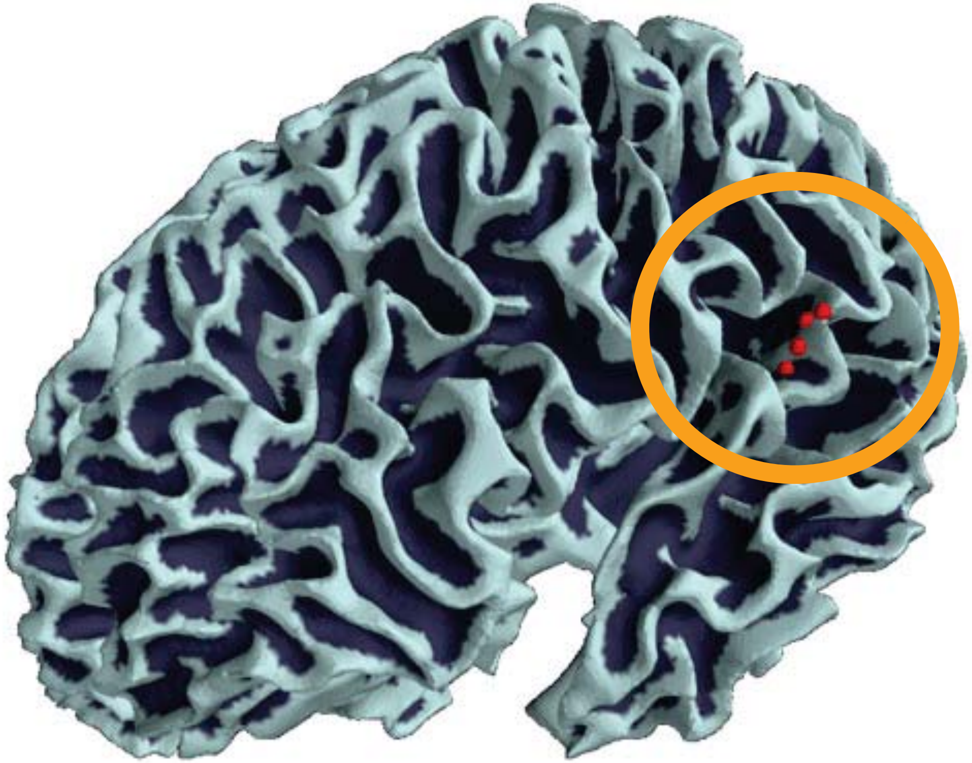
L1 vs L2 norms on combined M/EEG data

Activation in left-auditory cortex

L2 result



L1 result



Why does not everybody use sparse priors?

- Sparse priors lead to **harder optimization problems** (non-differentiable with **no closed form solution**).
- Solvers are iterative and **slower than L2.**

Contribution:

- Provide relevant sparse priors and fast algorithm:
  - Definition of good **convex priors** (beyond simple L1)
  - Come up with **fast algorithms** exploiting sparsity of the solution
  - Handle **specificities of M/EEG**: depth bias, loose/free orientation, whitening etc.

Inverse problem

Optimization problem:

$$\mathbf{X}^* = \arg\min_{\mathbf{X}} \|\mathbf{M} - \mathbf{G}\mathbf{X}\|_F^2 + \lambda\phi(\mathbf{X}), \lambda > 0$$

Data fit Prior

convex + convex

=

convex

- Data fit is **quadratic** hence **convex**
- If $\phi(\mathbf{X})$ is **convex**, then it is a **convex optimization problem**

L1 in the MEG world

L1 priors a.k.a. Minimum current estimate (MCE) :

$$\phi(\mathbf{X}) = \|\mathbf{X}\|_1 = \sum_i |x_i| \quad \text{with} \quad d_t = 1$$

[Matsuura et al. 95]

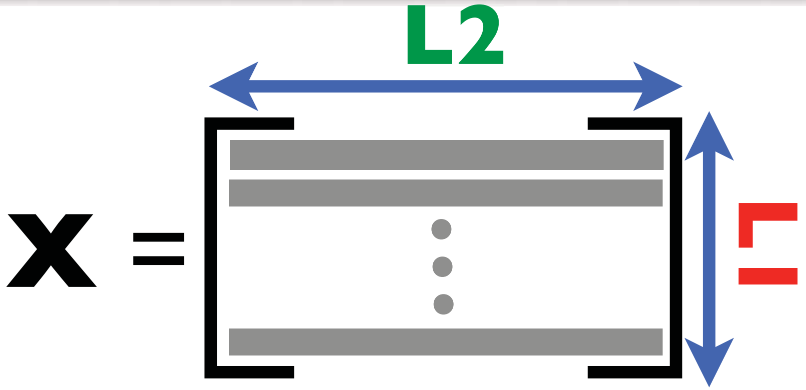
$\phi(\mathbf{X})$ is **convex**, **non differentiable** and has **no closed form solution**.

Remarks:

- It’s known as **LASSO** in machine learning / stats [Tibshirani 96], **basis pursuit denoising** (BPDN) in signal processing [Chen Donoho Saunders 99] and **MCE** [Matsuura 95, Uutela 99] in M/EEG
- **Not good enough** for M/EEG

$\phi(\mathbf{X})$ with M/EEG data: L21

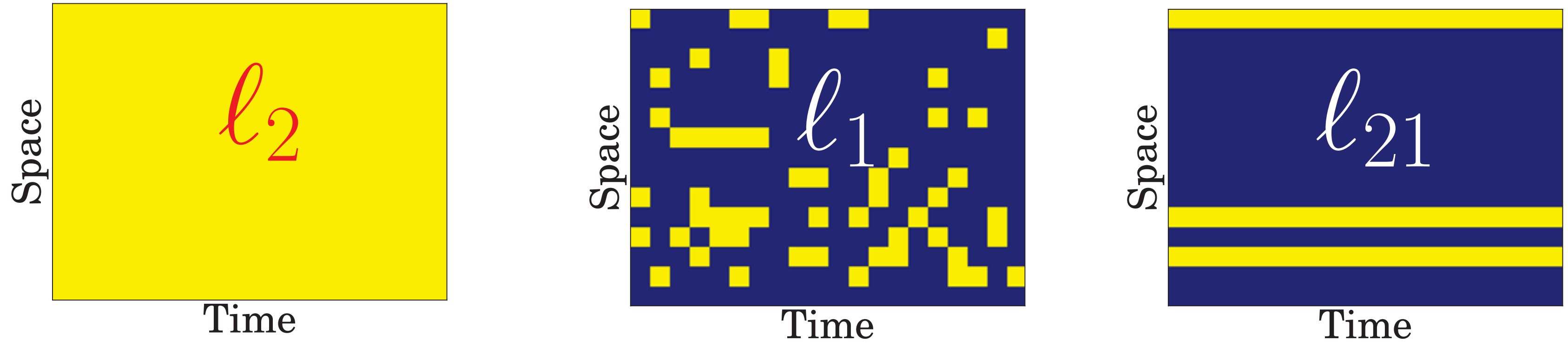
$$\phi(\mathbf{X}) = \|\mathbf{X}\|_{21} = \sum_i \sqrt{\sum_t |x_{i,t}|^2}$$



2-level mixed-norm

[Ou et al. Neuroimage 2009]

- It introduces **temporal structure** in the prior
- It guarantees that the active sources are the same over time

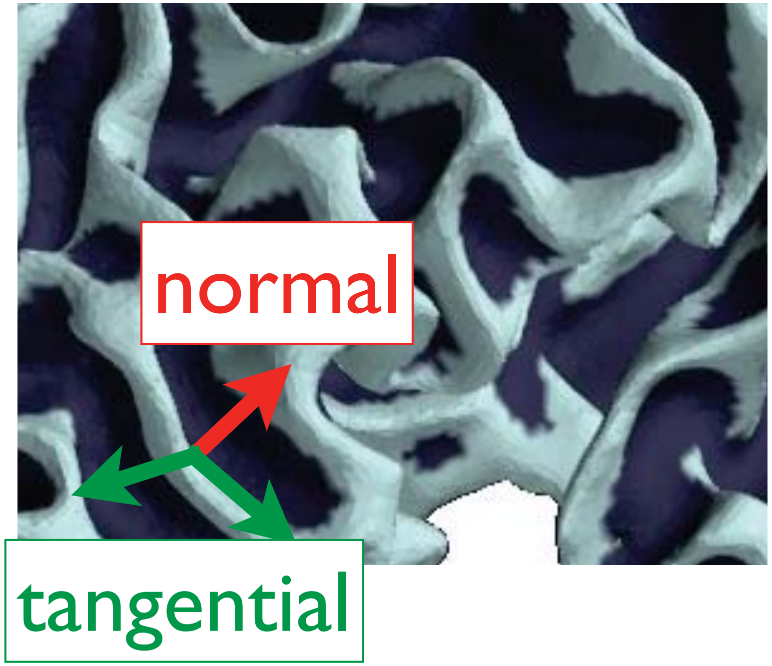


Remark : It is known as Group Lasso in Machine Learning & «joint feature selection»

[Yuan et al. 2006, Obozinski 2009 ...]

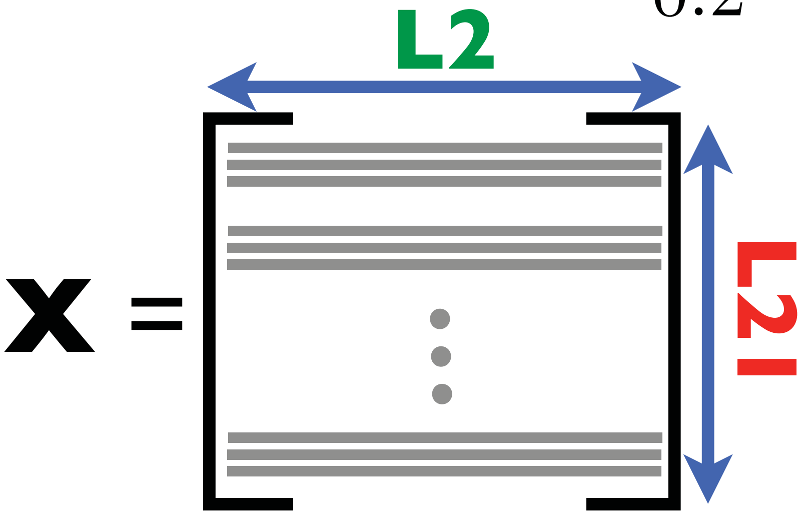
L21 with loose orientation

$$\phi(\mathbf{X}) = \|\mathbf{X}\|_{21} = \sum_i \sqrt{\sum_t |x_{i,t}^{normal}|^2 + \rho|x_{i,t}^{tang1}|^2 + \rho|x_{i,t}^{tang2}|^2}$$



custom but still a 2-level mixed-norm

with for example $\rho = \frac{1}{0.2}$



THM: you need custom sparse solvers adapted to M/EEG

Proximal iterations

- Very **generic** method (works for L1, L2, L21, etc.)
- **Iterative** method
- **First order method** (only requires to compute gradients)
- Algorithms **scalable** with **highly sampled source spaces**
- Can be much **faster** when combined with an **active-set strategy** that exploits the known sparsity of the solution

[Gramfort et al., *Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods*, PMB 2012]

[Kowalski et al., NIPS Optim. Workshop 2011]

Proximal iterations

Definition:

The **proximal operator** associated to $\lambda\phi$ is given by

$$\mathrm{prox}_{\lambda\phi}(\mathbf{Y}) = \arg\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\|_2^2 + \lambda\phi(\mathbf{X})$$

[Moreau 65]

Remark: It's the inverse problem with no **G** ie. no smoothing kernel

Forward-Backward iterations

Pb:

$$\mathbf{X}^* = \arg\min_{\mathbf{X}} \|\mathbf{M} - \mathbf{G}\mathbf{X}\|_F^2 + \lambda\phi(\mathbf{X}), \lambda > 0$$

Algorithm:

- *Initialize*: **Choose** $\mathbf{x}^{(0)} \in \mathbb{R}^{d_x}$ **(for example** $\mathbf{0}$**).**
- *Iterate*:

$$\mathbf{x}^{(k+1)} = \text{prox}_{\underline{\mu\lambda\phi}}\left(\mathbf{x}^{(k)} + \mu\underbrace{\mathbf{G}^T(\mathbf{m} - \mathbf{G}\mathbf{x}^{(k)})}_{\text{gradient of data fit}}\right)$$

where $0 < \mu < 2|||\mathbf{G}^T\mathbf{G}|||^{-1}$.

[Daubechies et al. 2004, Combettes et al. 2005]

Remarks:

- a.k.a. Iterative soft thresholding (ISTA)
- Convergence rate proportional to 1/k

Some proximal operators: L1

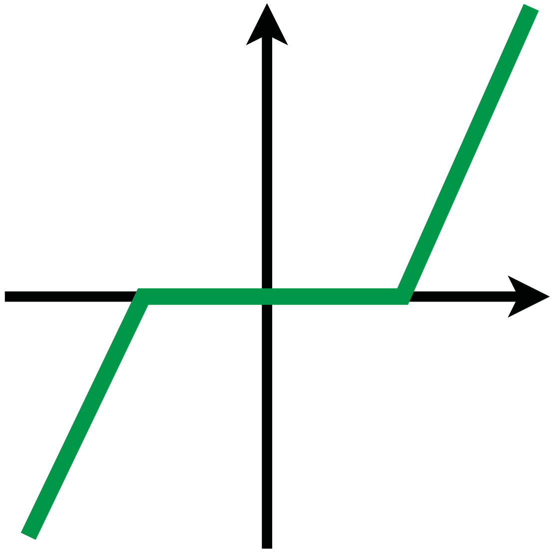
$$\phi(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_i |x_i|$$

Proximal operator:

$$\mathrm{prox}_{\lambda\|\ \|_1}(\mathbf{y}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

Solution:

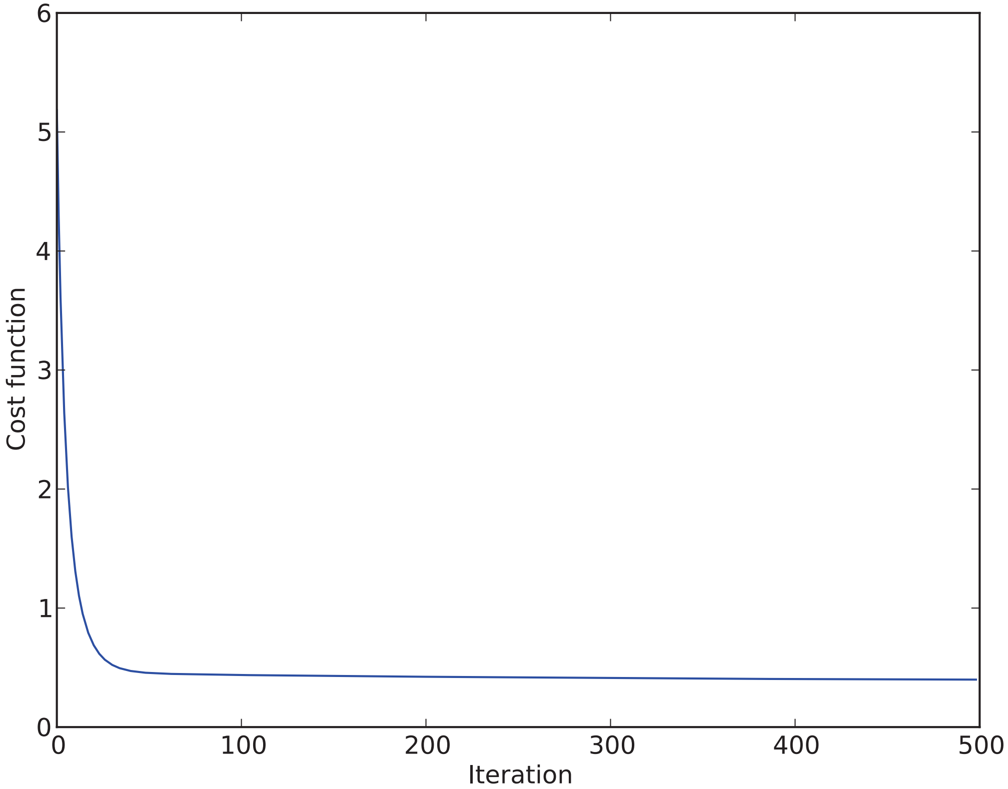
$$x_i^* = y_i \left(1 - \frac{\lambda}{|y_i|}\right)^+$$



Remark: It is referred to as **Soft Thresholding**

Lasso/MCE PythonISTA

```
alpha = 0.1 # Lambda parameter
L = 1.05 * linalg.norm(G)**2
for i in xrange(maxit):
    X += (1 / L) * np.dot(G.T, M - np.dot(G, X))
    X = np.sign(X) * np.maximum(np.abs(X) - (alpha / L), 0)
```



Ok but how many iterations?

Optimality conditions & Duality gaps

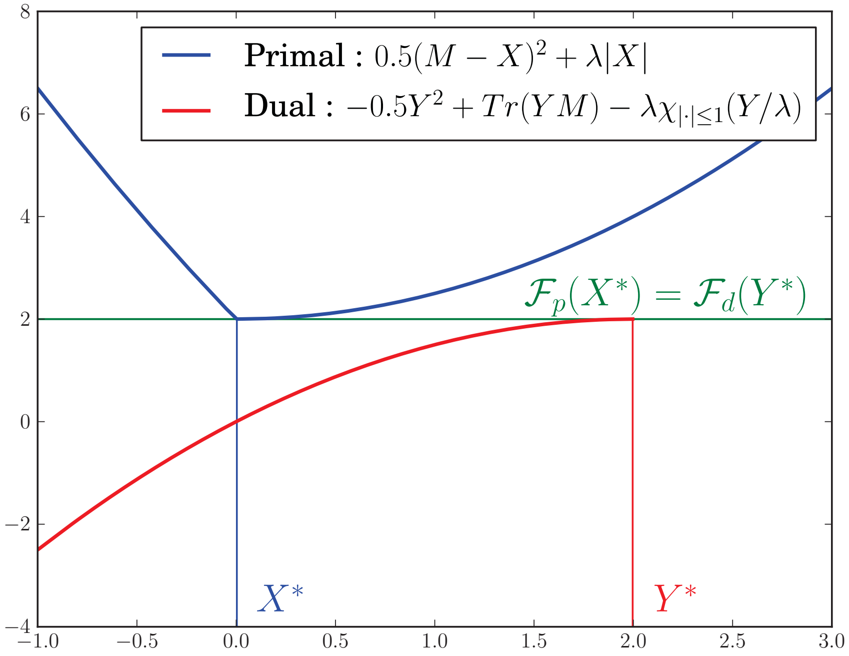
Primal problem $\min_X \frac{1}{2}\|M - GX\|_2^2 + \lambda\phi(X) = \min_X \mathcal{F}_p(M)$

Dual problem $\max_Y -\frac{1}{2}\|Y\|_2^2 + \mathrm{Tr}(Y^T M) - \lambda\phi^*(G^T Y/\lambda) = \max_Y \mathcal{F}_d(Y)$

Gap $\eta(X, Y) = \mathcal{F}_p(X) - \mathcal{F}_d(Y) \;\geq 0$

Slater's conditions «say» : $\eta = 0$ at optimum (strong duality)

Example with Lasso :



THM:
A principled way to test the optimality of a solution for a non-smooth problem

Active set methods (L1 & L21 priors)

- You know 2 things:
  - only a few sources will be active
  - how to test the optimality of a solution

The idea :

1. Start with a small problem (only a few sources)
2. Test optimality assuming all left out sources have 0 activation
3. If not good enough

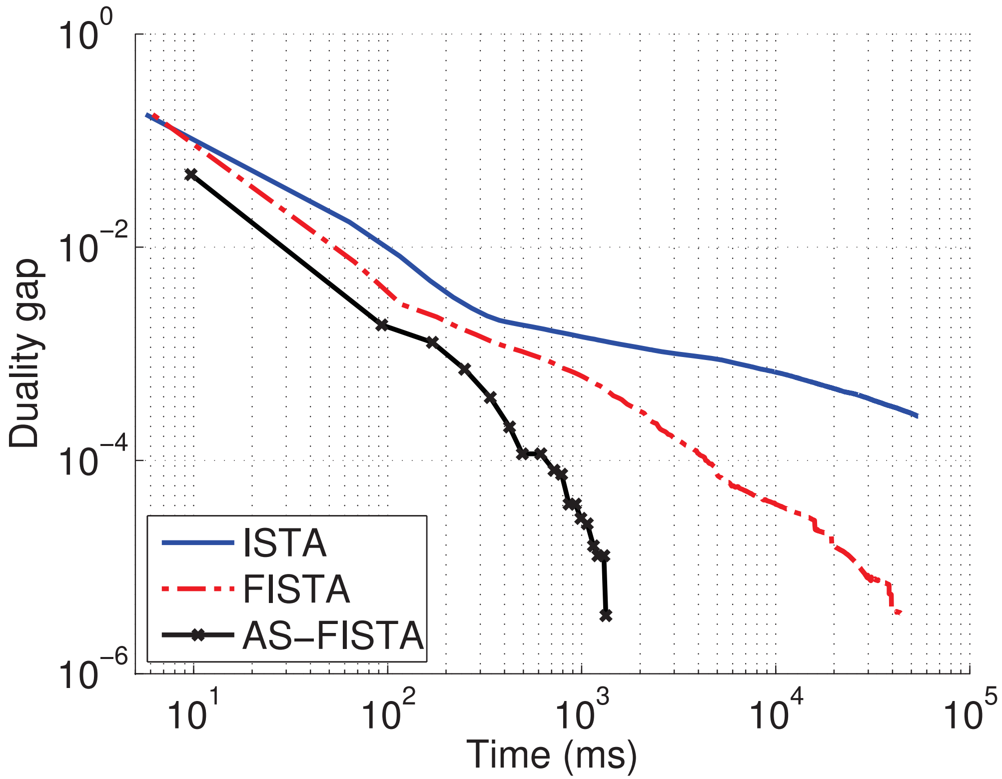
    add new sources to the problem and goto 1

  else

    stop !

[Markowitz 1952, Osborne «Homotopy methods» 2001, Efron «Lars» 2004,
Roth «active-set for the group-lasso» ICML' 08,
Kowalski et al., NIPS Optim. Workshop 2011]

ISTA vs. FISTA vs. Active Set



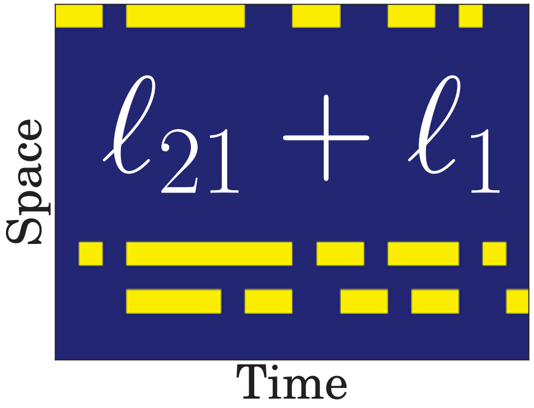
The M/EEG inverse pb can be solved with non-l2 priors also in a few seconds !

- It is possible to reach an $1/k^2$ using multi-steps methods e.g. FISTA (Fast - ISTA) [Nesterov 2007, Beck et al. 2009]
- It is possible to be even faster for certain problems using an «active set» strategy.

But… the brain is not stationary

L21 like any other sparse solver available today

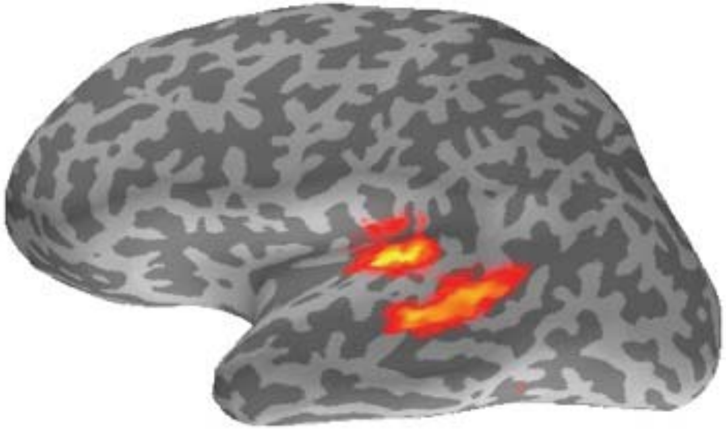
it imposes the sources to be the same over the entire time interval

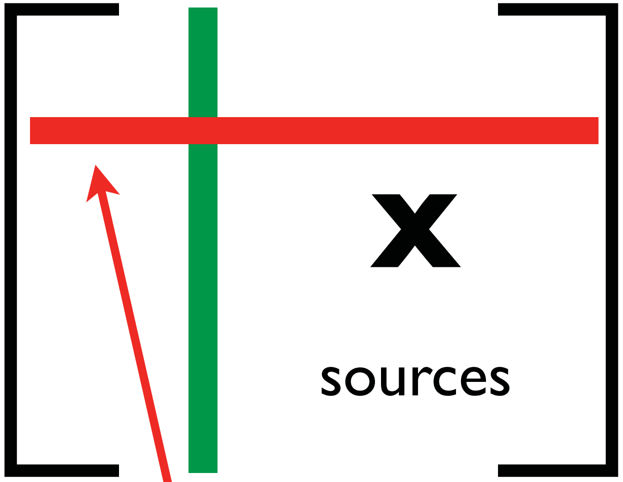


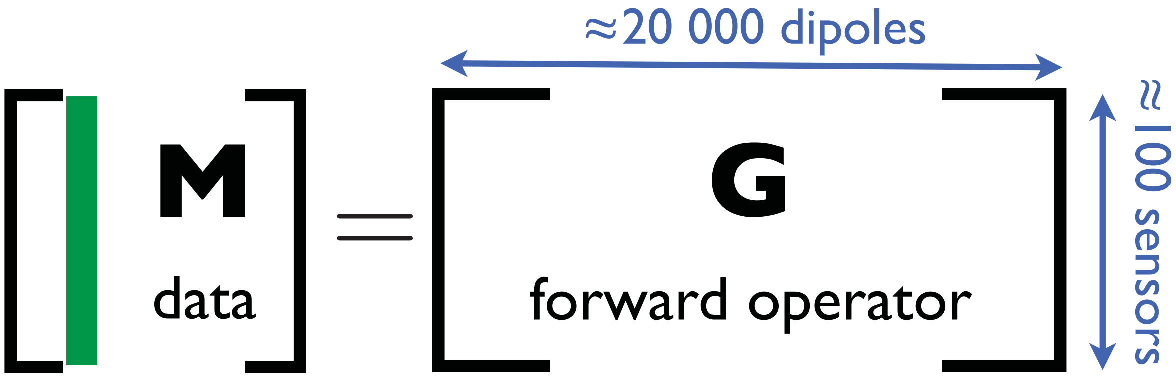
Challenge:

How do you promote sparse solutions with non-stationary sources?

back to M = G X + E

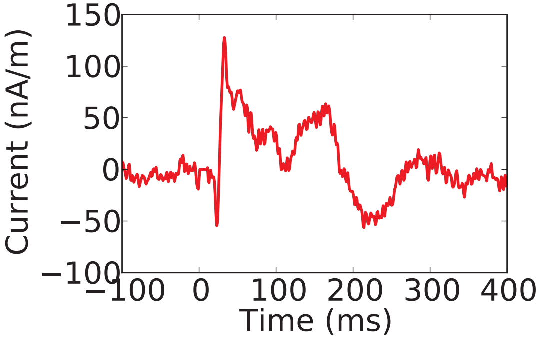


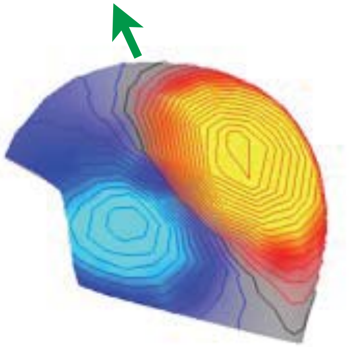




+ E

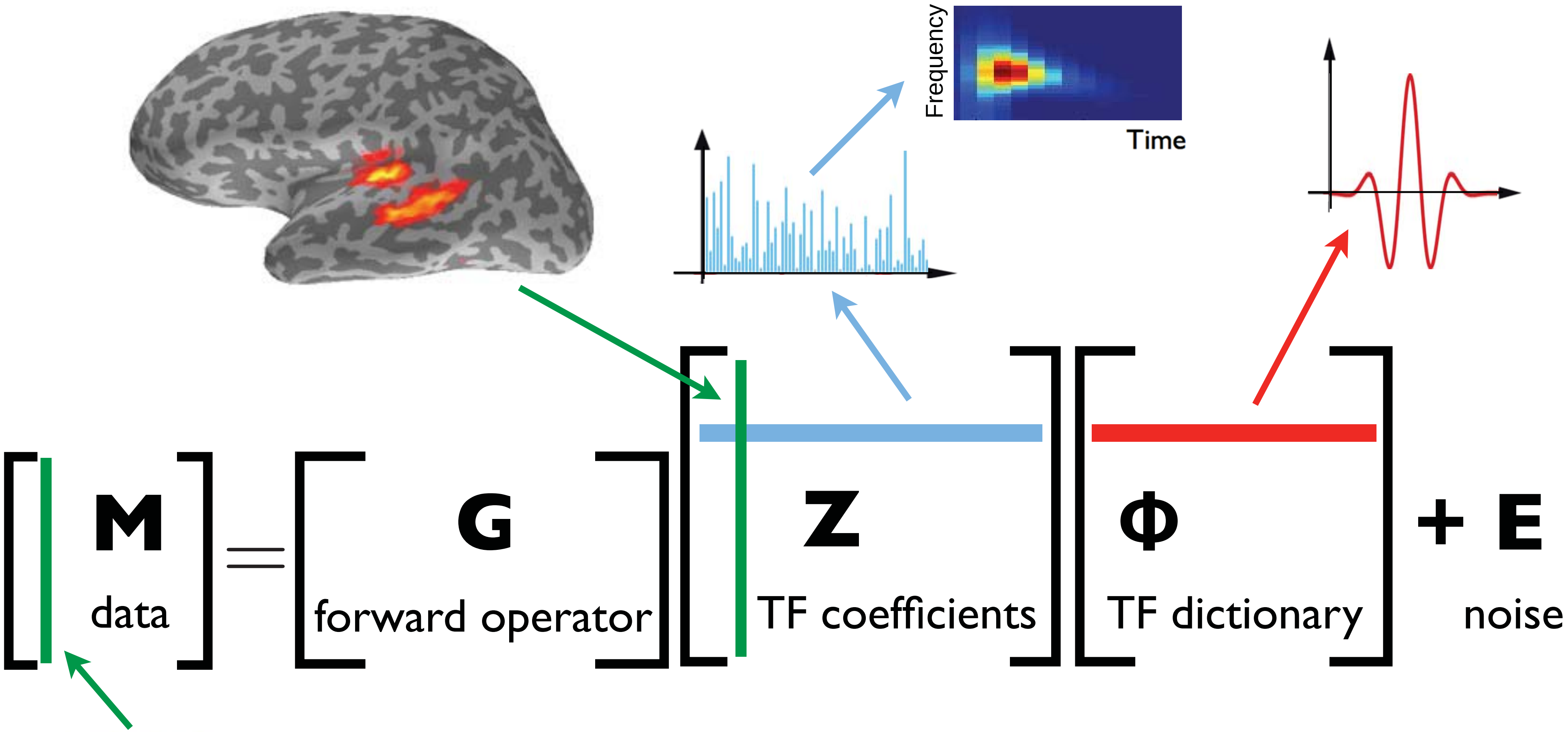
noise



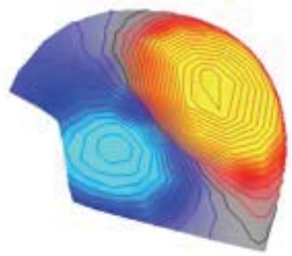


Objective: estimate X given M

$M = GZ\Phi + E$



M data = **G** forward operator **Z** TF coefficients **Φ** TF dictionary **+ E** noise



Objective: estimate Z given M

Time-frequency (TF) prior

The classical approach [MNE, dSPM, sLORETA]:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \underbrace{\|\mathbf{M} - \mathbf{G}\mathbf{X}\|_F^2}_{\text{data fit}} + \underbrace{\lambda\phi(\mathbf{X})}_{\text{prior}}, \;\; \lambda > 0$$

we propose:

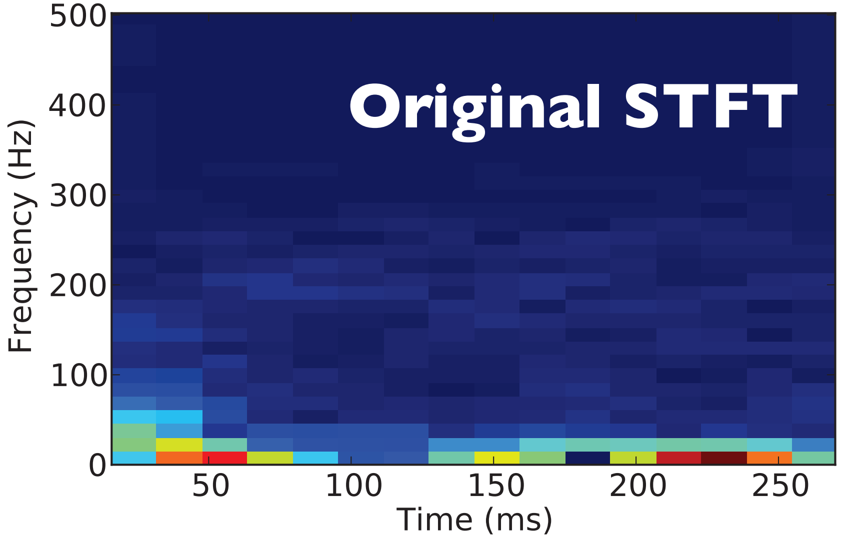
$$\hat{\mathbf{Z}} = \arg\min_{\mathbf{Z}} \|\mathbf{M} - \mathbf{G}\mathbf{Z}\mathbf{\Phi}^{\mathcal{H}}\|_F^2 + \lambda\phi(\mathbf{Z}), \;\text{then}\; \hat{\mathbf{X}} = \hat{\mathbf{Z}}\mathbf{\Phi}^{\mathcal{H}}$$

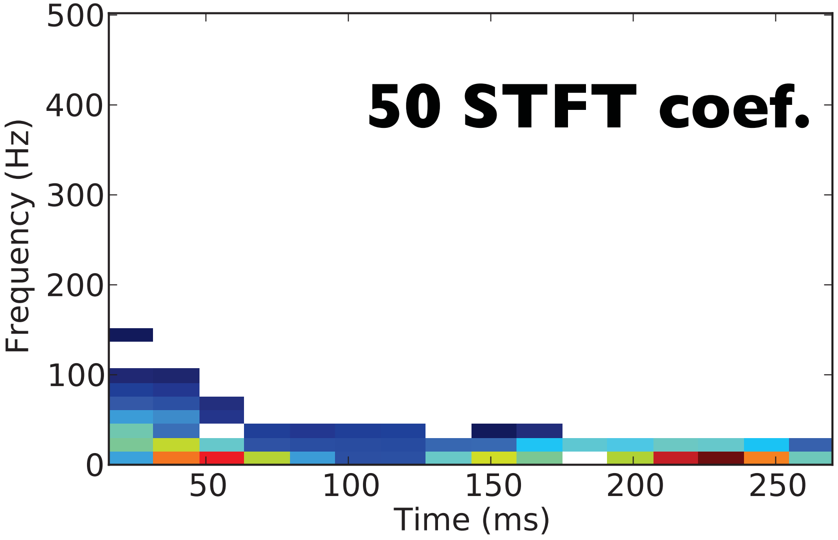
- $\mathbf{\Phi}$: is a **TF dictionary** of Gabor atoms
- Z : **coefficients** of the **TF transform** of the sources

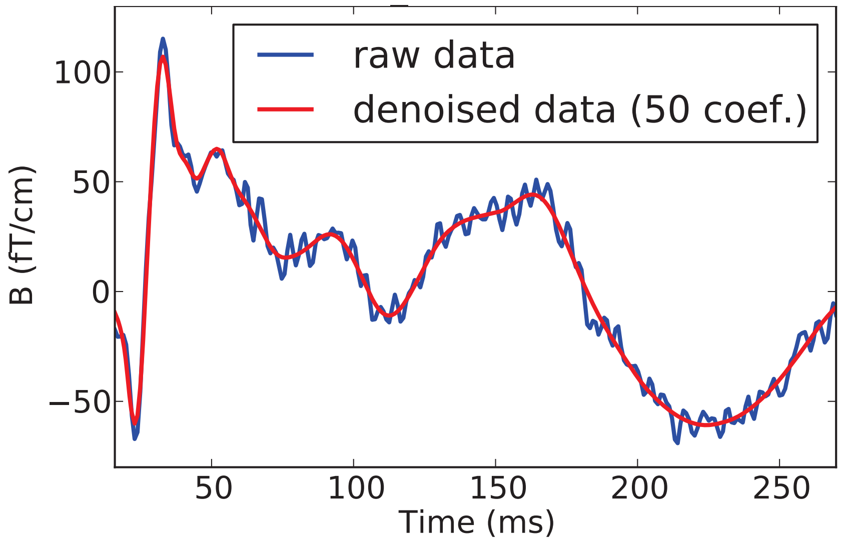
Advantage:
localization in
space, time and frequency
in one step

Why does it make sense?

and why a sparse prior shall work ?







[«Denoising by soft-thresholding» Donoho 95]

Time frequency dictionaries

discrete version of the
complex Gabor transform = short time fourier transform (STFT)

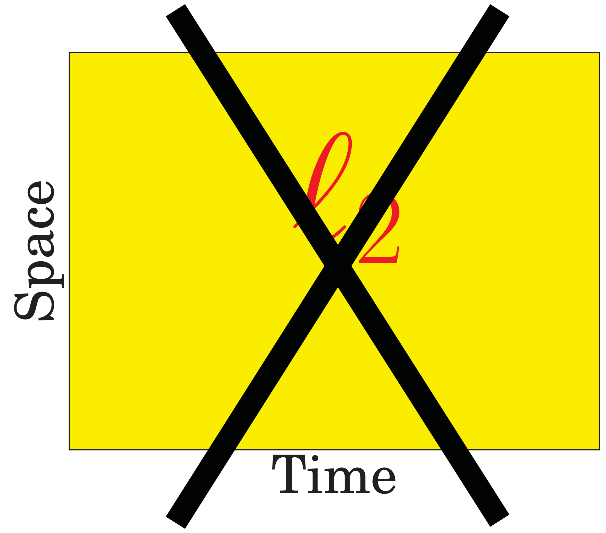
- It is **invertible**
- It is **translation invariant**

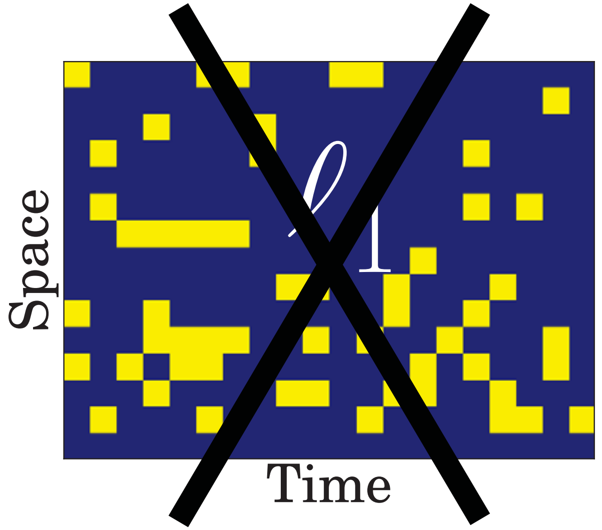
  (not like classical dyadic wavelets)
- It can capture **non-stationary signals** (not like FFT)

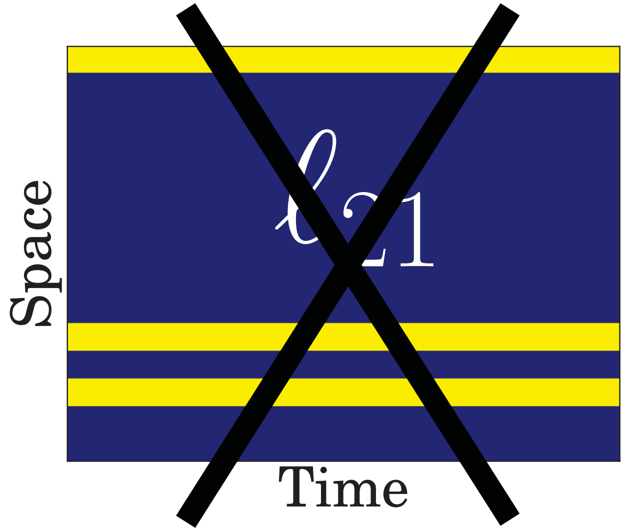
  (It is classically used in M/EEG on sensor measurements)
- It is **relatively fast** to compute

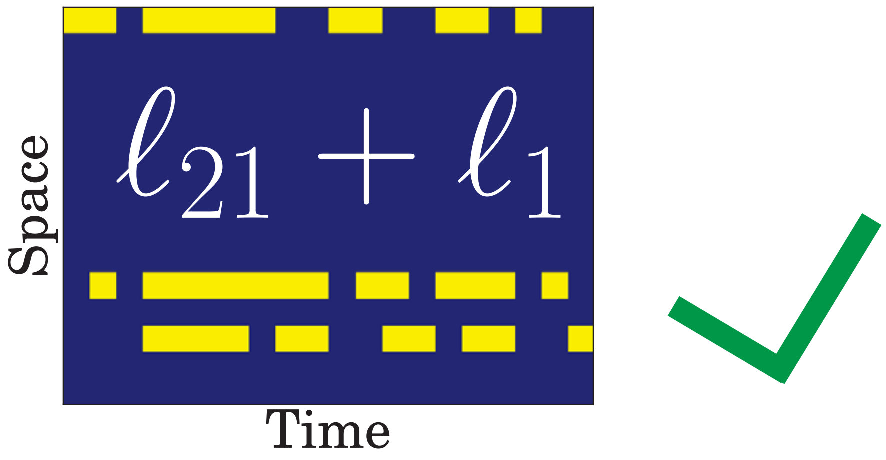
What is a good prior on Z?

What prior?









$$\phi(Z) = \lambda(\rho\|Z\|_1 + (1-\rho)\|Z\|_{21})$$

Algorithm

Definition 1 (Proximity operator). *Let* $\varphi : \mathbb{R}^M \to \mathbb{R}$ *be a proper convex function. The proximity operator associated to* φ*, denoted by* $\text{prox}_\varphi : \mathbb{R}^M \to \mathbb{R}^M$ *reads:*

$$\text{prox}_\varphi(\mathbf{Z}) = \arg\min_{\mathbf{V}\in\mathbb{R}^M} \frac{1}{2}\|\mathbf{Z}-\mathbf{V}\|_2^2 + \varphi(\mathbf{V}) \enspace .$$

Lemma 1 (Proximity operator for $\ell_{21} + \ell_1$). *Let* $\mathbf{Y} \in \mathbb{C}^{P\times K}$ *be indexed by a double index* (p,k)*.* $\mathbf{Z} = \text{prox}_{\lambda(\rho\|.\|_1+(1-\rho)\|.\|_{21})}(\mathbf{Y}) \in \mathbb{C}^{P\times K}$ *is given for each coordinates* (p,k) *by*

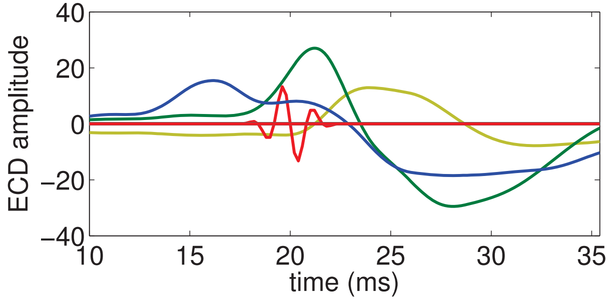
$$Z_{p,k} = \frac{Y_{p,k}}{|Y_{p,k}|}(|Y_{p,k}| - \lambda\rho)^+ \left(1 - \frac{\lambda(1-\rho)}{\sqrt{\sum_k (|Y_{p,k}| - \lambda\rho)^{+^2}}}\right)^+ \enspace .$$

where for $x \in \mathbb{R}$*,* $(x)^+ = \max(x,0)$ *, and by convention* $\frac{0}{0} = 0$ *.*

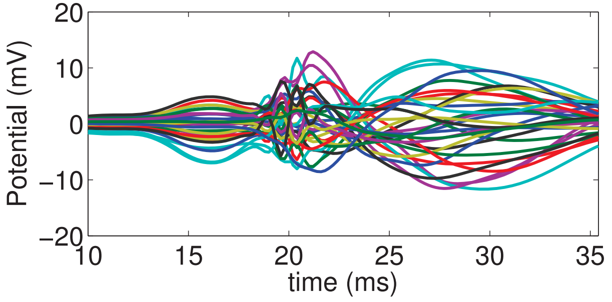
THM: It boils down to 2 successive thresholdings

[Jenatton et al. 2011, Gramfort et al. IPMI 2011]

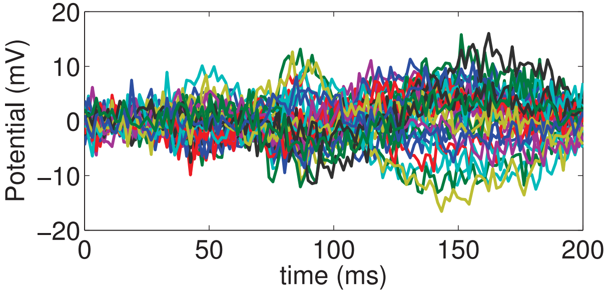
Simulation results (part 1)



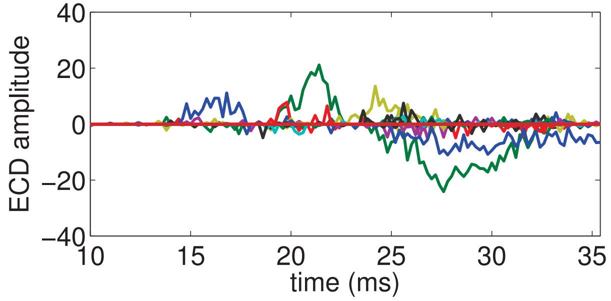
(a) $\mathbf{X}$ ground truth



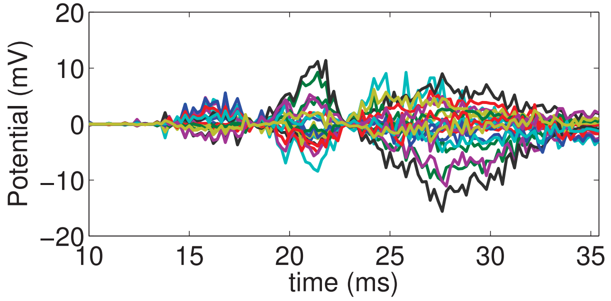
(b) $\mathbf{M}$ noiseless



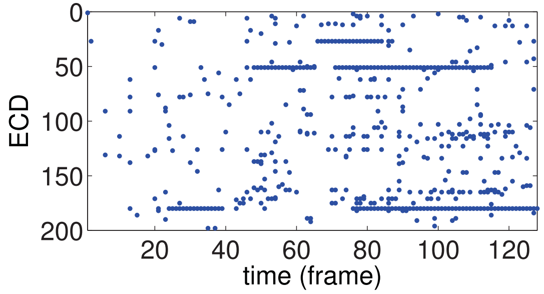
(c) $\mathbf{M}$ noisy (SNR=6dB)



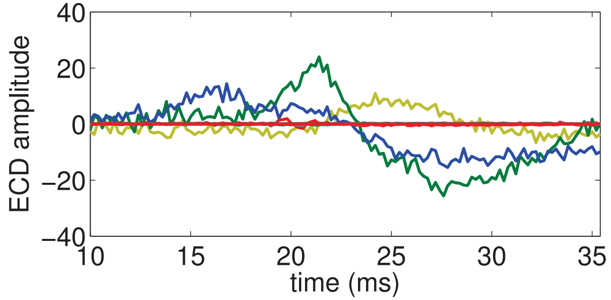
(d) $\mathbf{X}_{\ell_1}$



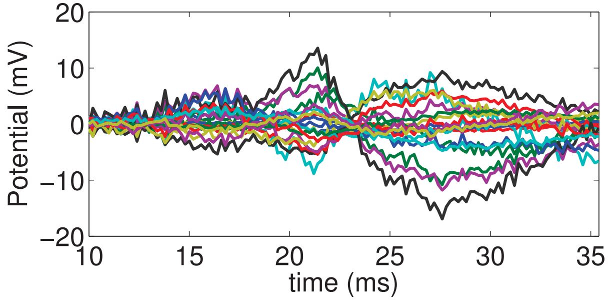
(e) $\mathbf{M}^\star_{\ell_1} = \mathbf{G}\mathbf{X}^\star_{\ell_{21}}$



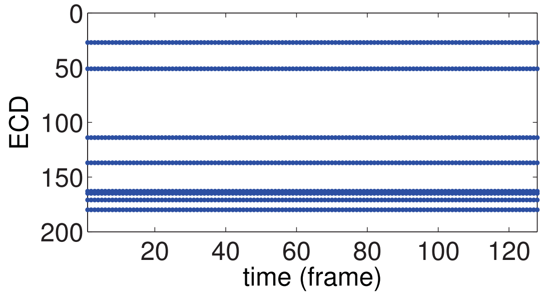
(f) Non-zeros of $\mathbf{X}^\star_{\ell_1}$



(g) $\mathbf{X}^\star_{\ell_{21}}$

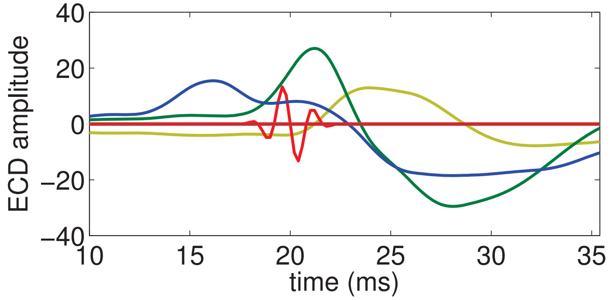


(h) $\mathbf{M}^\star_{\ell_{21}} = \mathbf{G}\mathbf{X}^\star_{\ell_{21}}$

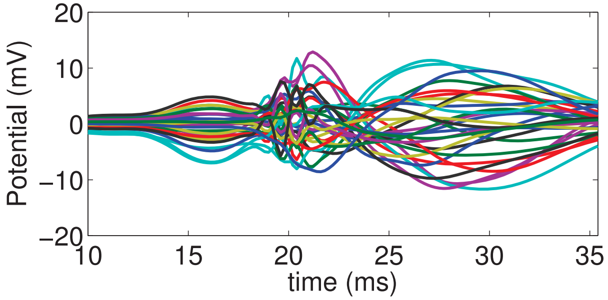


(i) Non-zeros of $\mathbf{X}^\star_{\ell_{21}}$

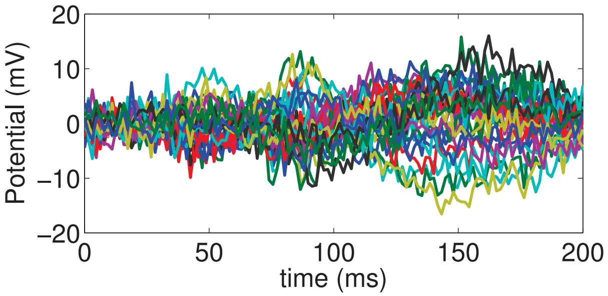
Simulation results (part 2)



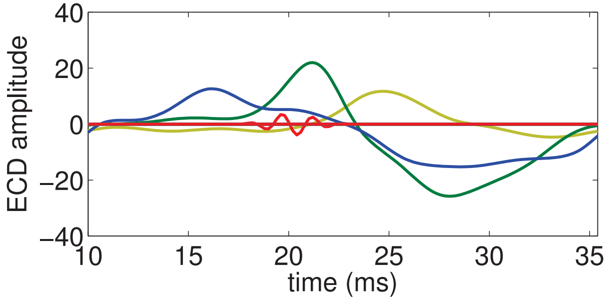
(a) $\mathbf{X}$ ground truth



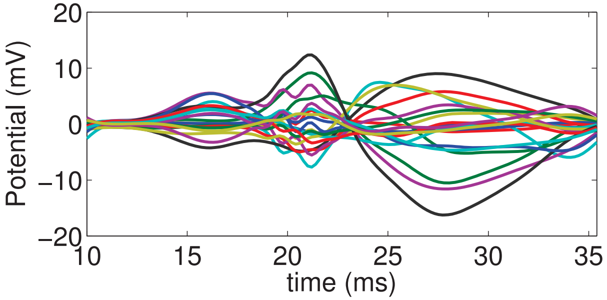
(b) $\mathbf{M}$ noiseless



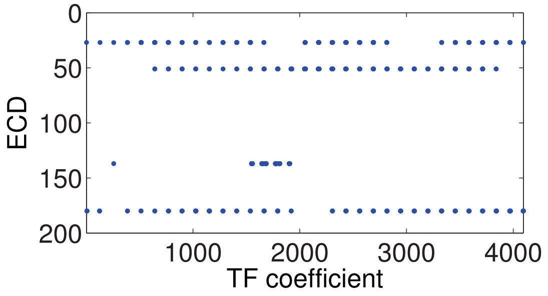
(c) $\mathbf{M}$ noisy (SNR=6dB)



(j) $\mathbf{X}^\star_{\mathrm{TF}\ \ell_{21}+\ell_1}$



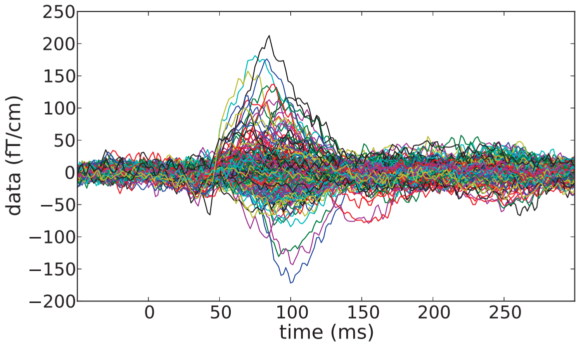
(k) $\mathbf{M}^\star_{\mathrm{TF}\ \ell_{21}+\ell_1}$



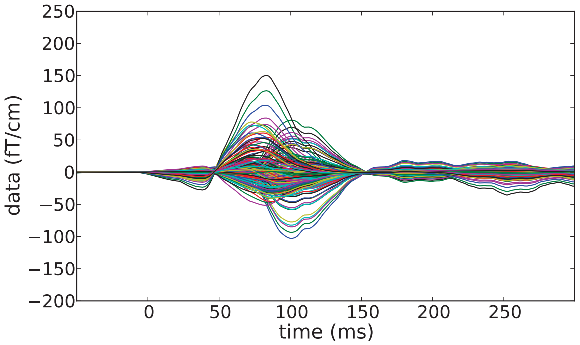
(l) Non-zeros of $\mathbf{Z}^\star_{\ell_{21}+\ell_1}$

MEG Auditory data

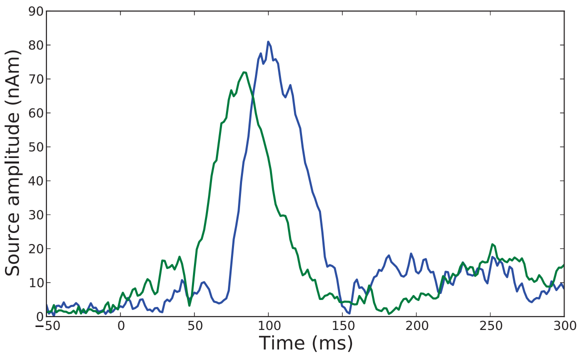
Protocol: 50 epochs of auditory tones in left ear
(305 MEG, 59 EEG channels)



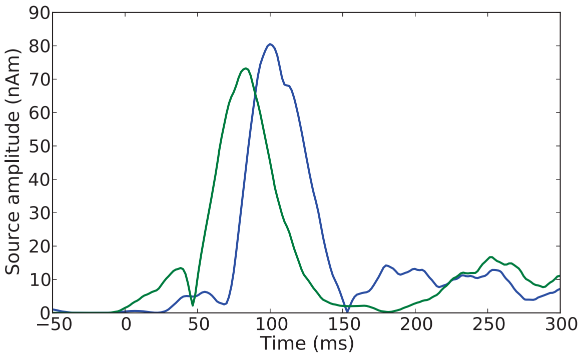
(a) MEG data (Gradiometers only)



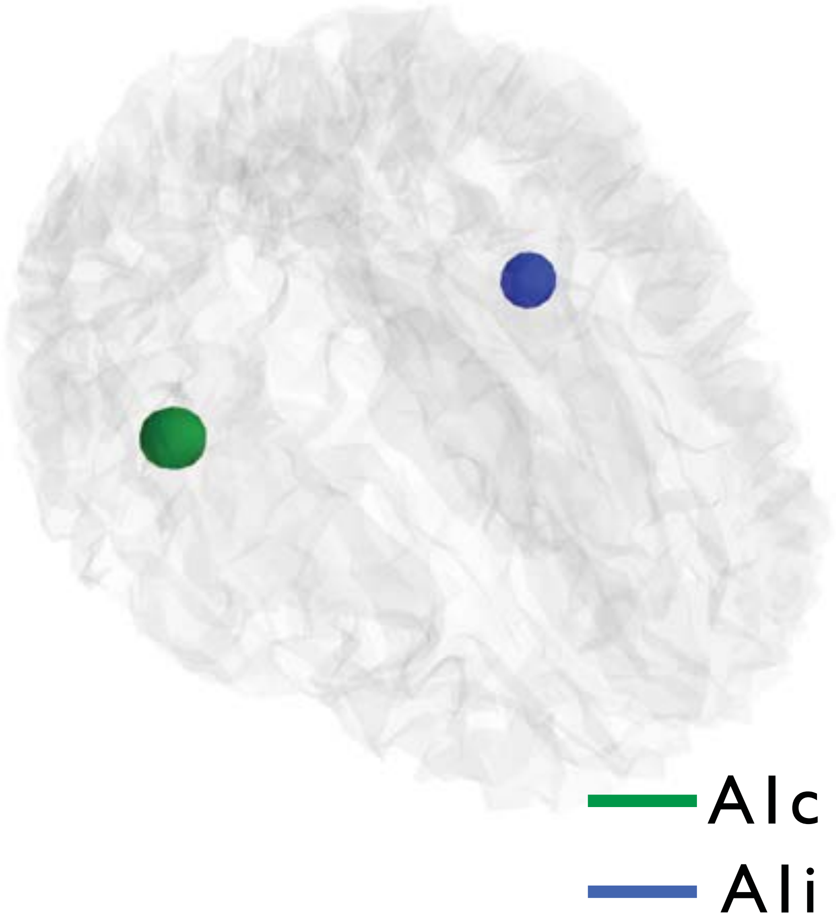
(b) $\mathbf{GX}^{\star}_{\text{TF-MxNE}}$ (explained data)



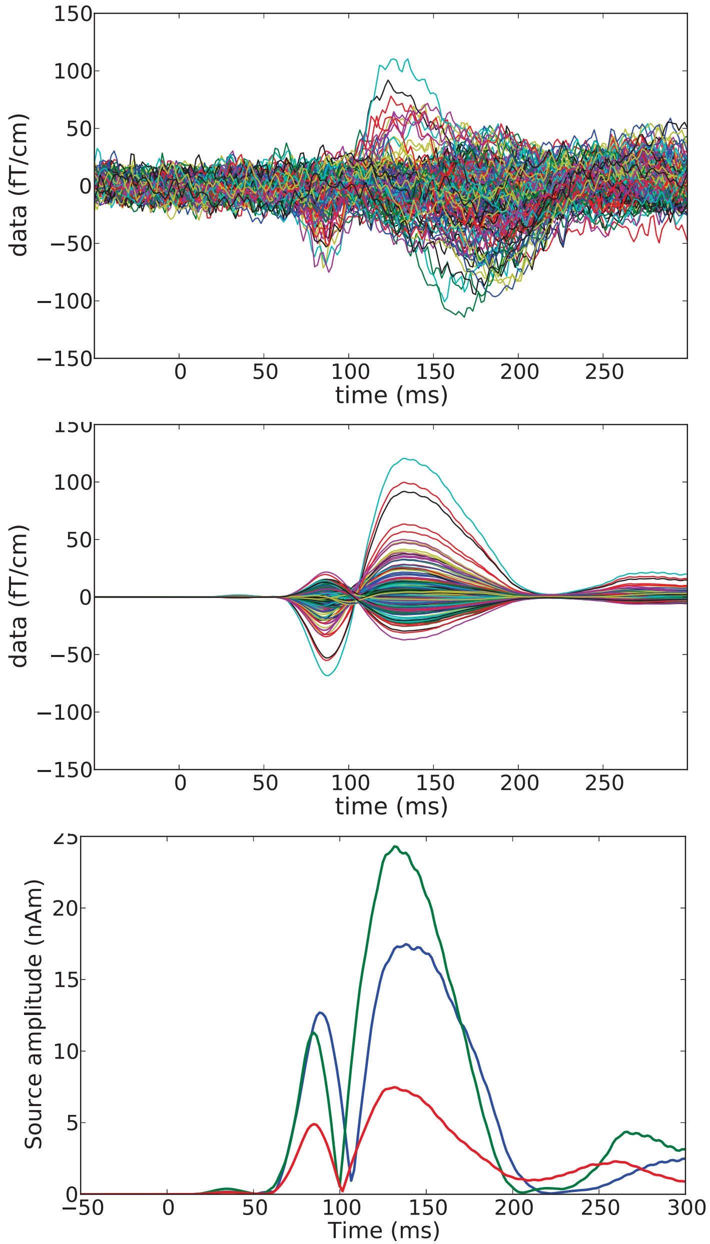
(c) $\mathbf{X}^{\star}_{\text{MxNE}}$



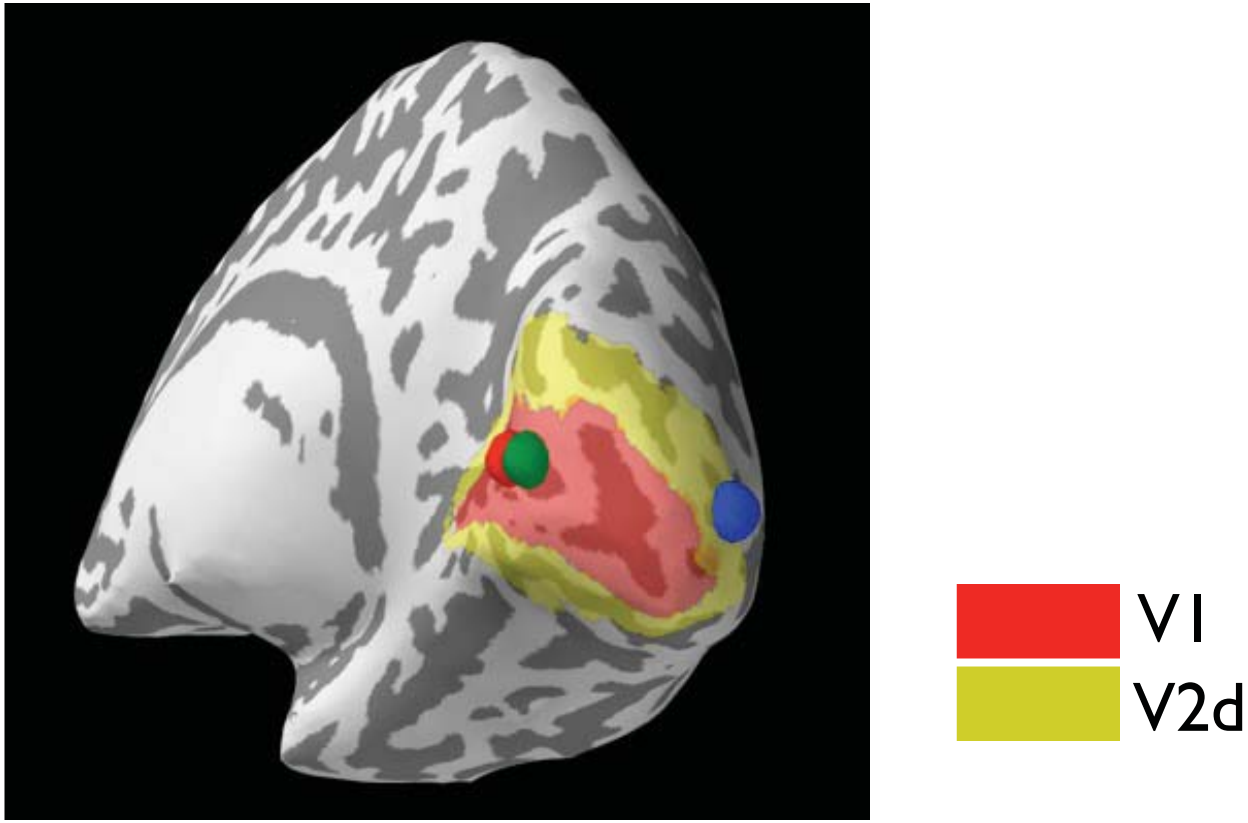
(d) $\mathbf{X}^{\star}_{\text{TF-MxNE}}$



MEG Visual data



Protocol: 50 epochs of visual flash in left hemi-field (305 MEG, 59 EEG channels)

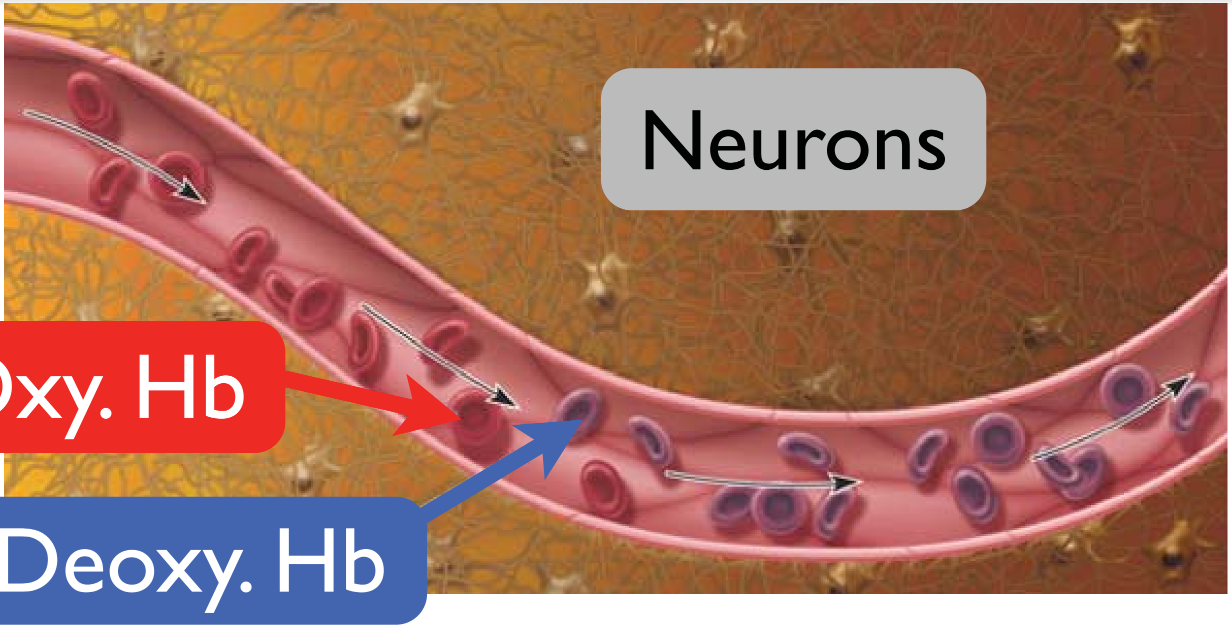


“Brain reading” with fMRI … *prediction vs. recovery*

[*Gramfort et al., Beyond brain reading: randomized sparsity and clustering to simultaneously predict and identify, NIPS Workshop 2011*]

[*Varoquaux et al., Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering, ICML 2012*]

fMRI: neurons change hemoglobin oxygenation

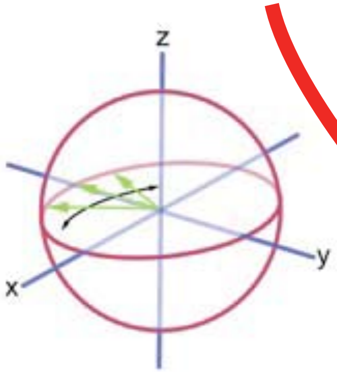


Neurons

Oxy. Hb

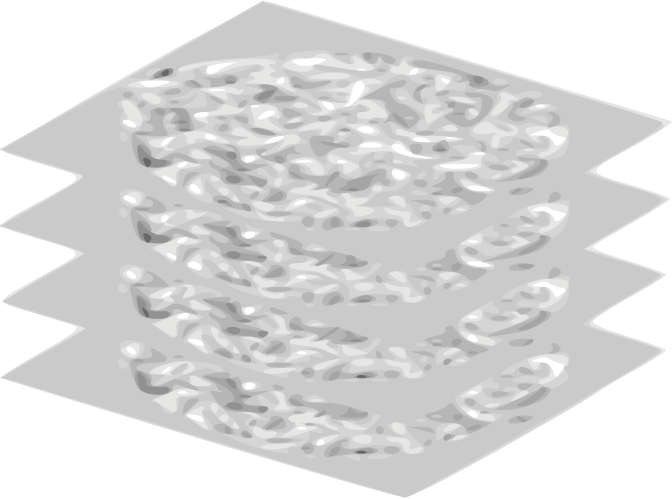
Deoxy. Hb

High spatial resolution (vox ≃ 2mm)



Nuclear
Magnetic
Resonance

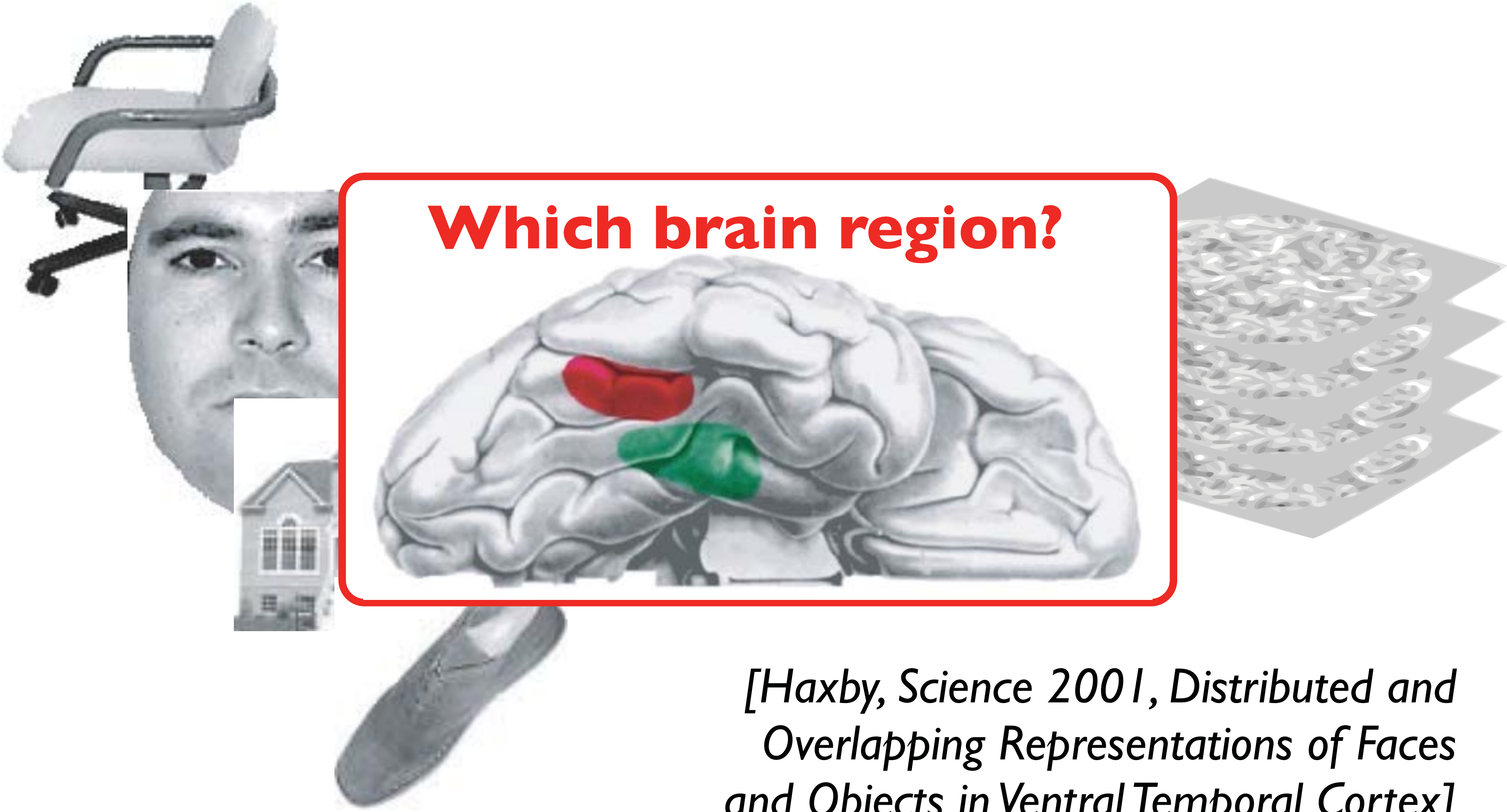
Scanner



3D volumes
(1 every 2s)

Brain mapping

Stimuli



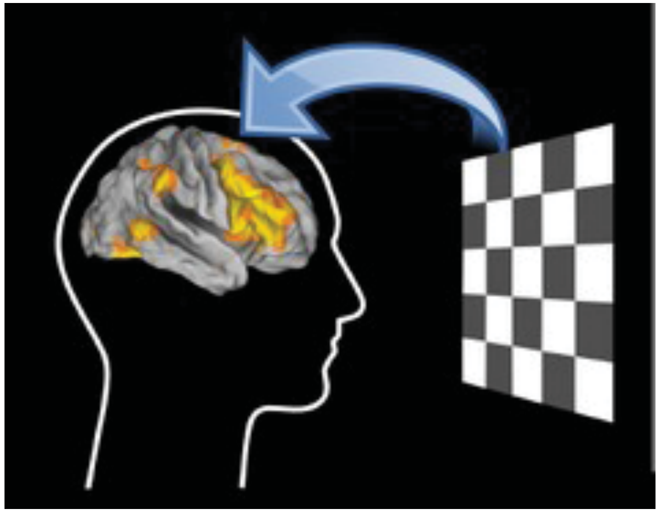
Which brain region?

[Haxby, Science 2001, Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex]

Standard analysis vs. MVPA

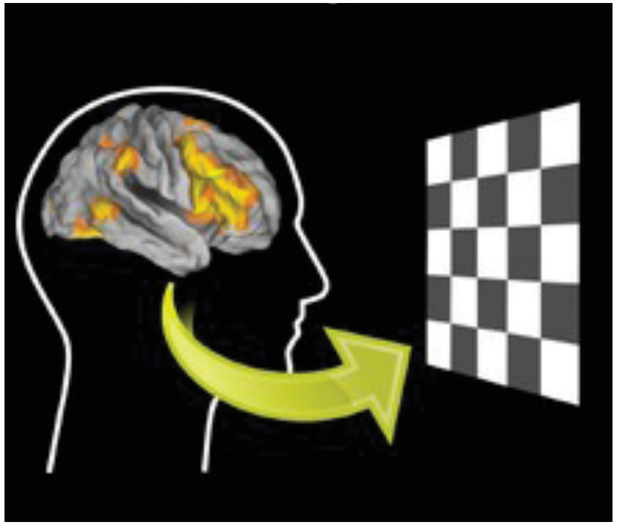
Standard analysis

- Test whether the voxel is recruited by the task
- **Many voxels** : problem of multiple comparisons
- Statistical power ∝ 1 / n_voxels



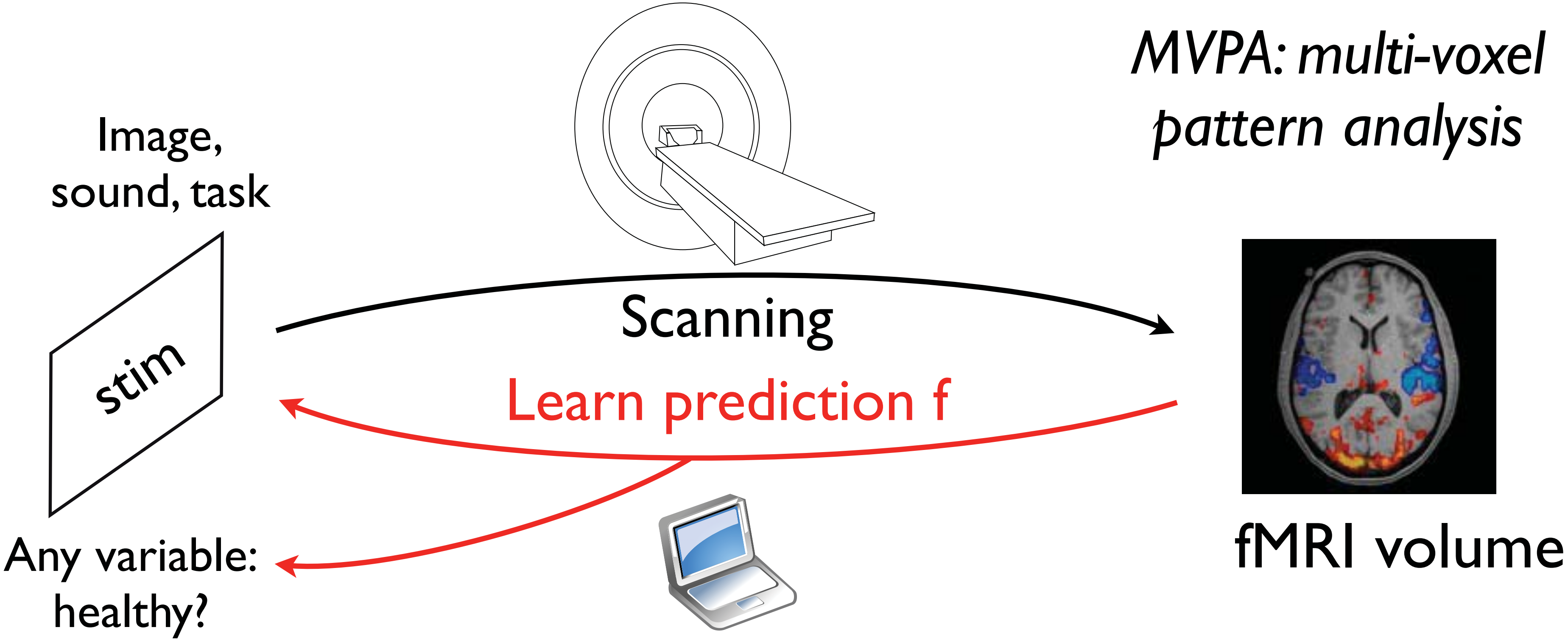
Supervised Learning

- Predictive model
- **Many voxels** : curse of dimensionality
- But can exploit the information shared between voxels: more statistical power?



Supervised learning *a.k.a. MVPA*

MVPA: multi-voxel pattern analysis

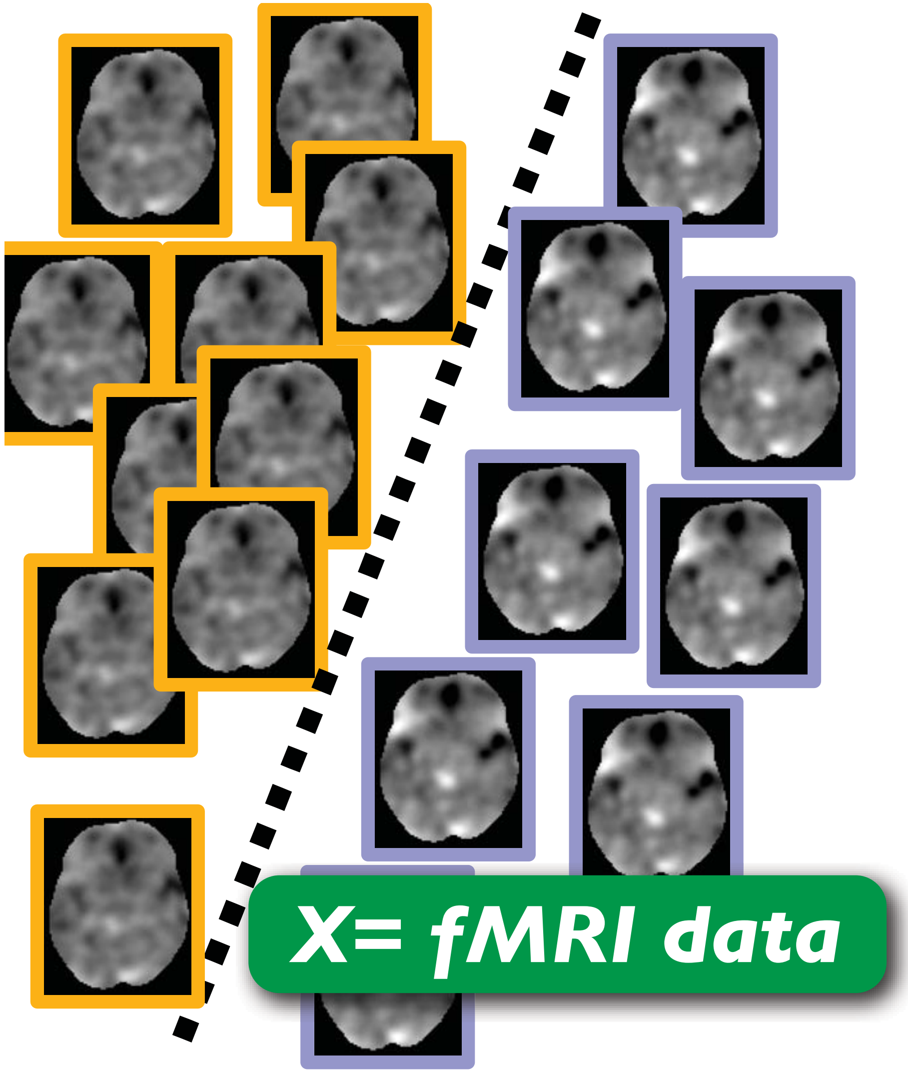


Challenge: Predict a behavioral variable from the fMRI data

Question: Is the information captured by fMRI? If so, **where**?

[Haxby et al. 01, Cox et al. 2003, Mitchell et al 04, Laconte et al 05, Kamitani et al 05, Thirion et al. 06, Haynes et al. 06, Kay et al. 08, Miyawaki et al. 08, Yamashita et al. 08, Naseralis et al. 09, Pereira et al. 09, Caroll et al. 09, Ryali et al. 2010, ...]

Classification example with fMRI



X= fMRI data

The **objective** is to be able to **predict** 🟪 or 🟧 given an fMRI activation map

🟪		🟧
Patient	vs.	Controls
Faces	vs.	Houses
...	vs.	...
1	vs.	-1

i.e. $y = \{-1, 1\}$

objective: Predict $y = \{-1, 1\}$ given $x \in \mathbb{R}^p$

fMRI // M/EEG

MNE:

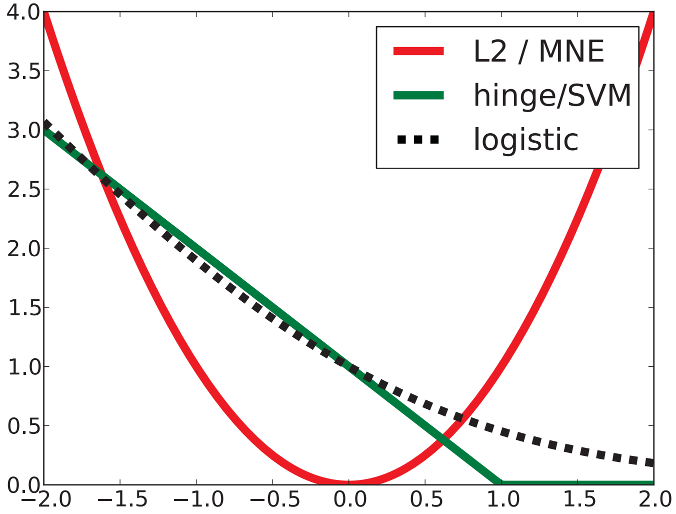
$$\min_w \frac{1}{n}\sum_{i=1}^{n}\underline{(y_i - x_i^T w)^2} + \lambda\|w\|_2^2 \qquad \|w\|_2^2 = \sum_{i=1}^{p} w_i^2$$

Linear SVM:

$$\min_w \frac{1}{n}\sum_{i=1}^{n}\underline{\mathrm{hinge}(y_i x_i^T w)} + \lambda\|w\|_2^2$$

$$y_i = \mathrm{sign}(x_i w)$$

≍ 1e2 or 1e3 observations, 1e6 voxels (variables): ill-posed

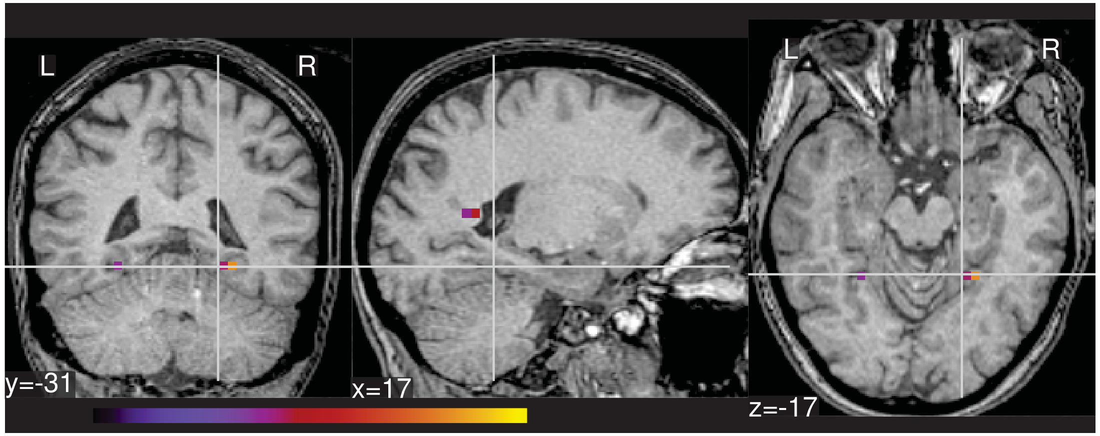


THM: Like L2 is heavily used for MEG, linear SVM is very common in fMRI

Hope and caveats

Hope: Use sparse priors to get sub-linear sample complexity ($n \propto k \log(p)$)

Problem: RIP, mutual incoherence … not valid for fMRI due to spatial redundancy: very correlated design *[Candes 06, Tropp 04, Wainright 09]*



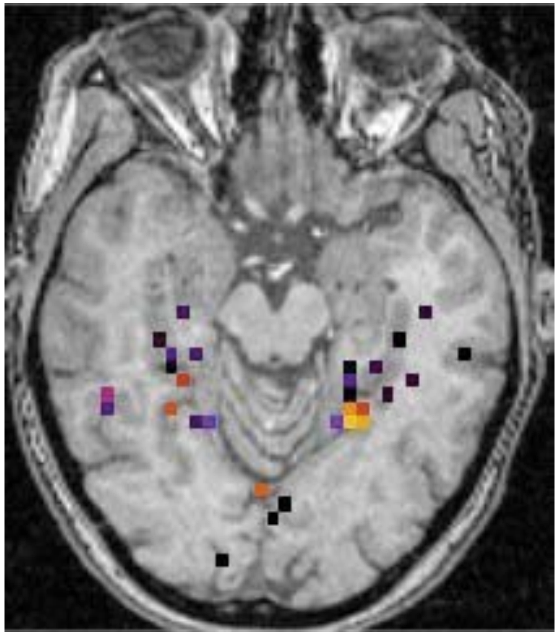
Lasso with CV : 23 Coefs

Randomized sparsity

Stability Selection:

- **Perturb design:** subsample the data (or bootstrap) & rescale features (columns)
- **Run L1 solver**
- **Keep** features that are "often" active

Good recovery without mutual incoherence property but RIP-like



Problem: Cannot recover large correlated groups of features

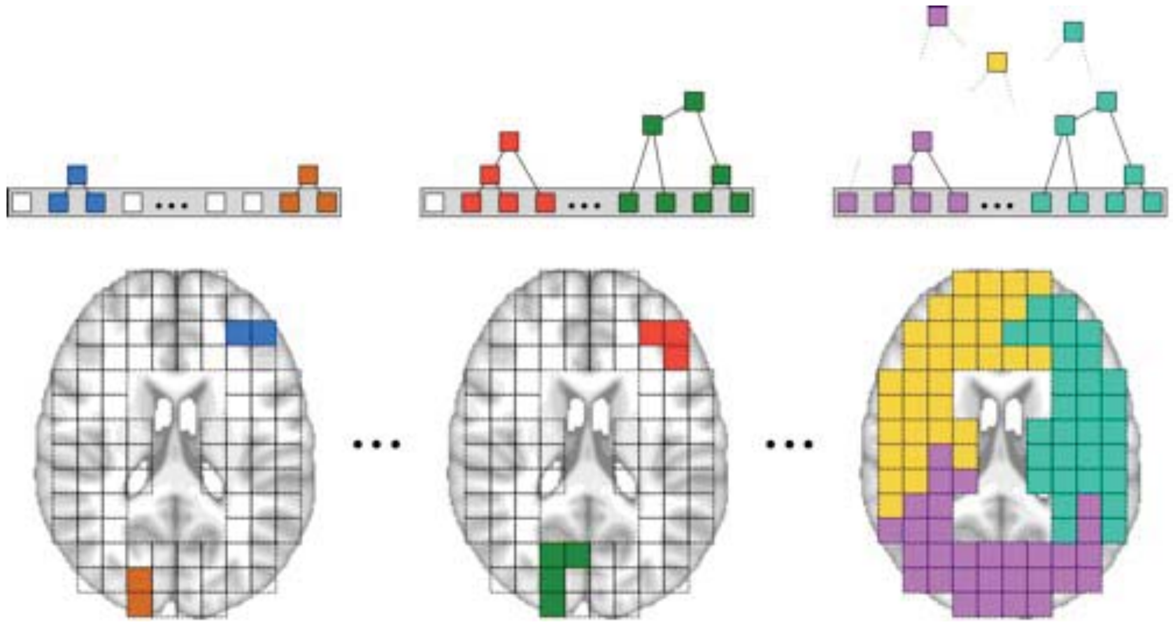
Intuition: For m correlated features, selection frequency divided by m

[Meinshausen and Buhlmann "Stability Selection" 2010, Bach "Bootstrap Lasso" 2008]

Randomized sparsity & clustering

Stability Selection:

- **Perturb design:** subsample the data (or bootstrap) & rescale features (columns)
- **Cluster features / voxels**
- **Run L1 solver**
- **Keep** features that are "often" present in an active cluster



Ward hierarchical clustering with spatial constraint

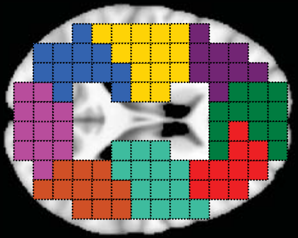
Reduces correlations: better RIP

[Michel et al. 2011]

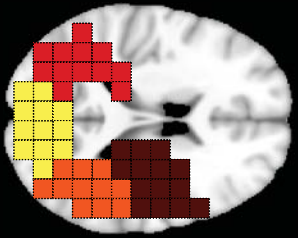
Algorithm

1 set `n_clusters` and sparsity by cross-validation

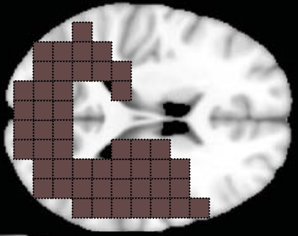
2 `loop:` perturb randomly data



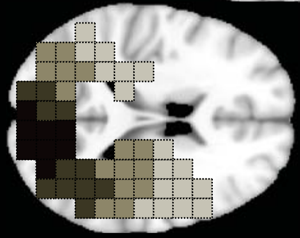
3 clustering to form reduced features



4 sparse linear model on reduced features



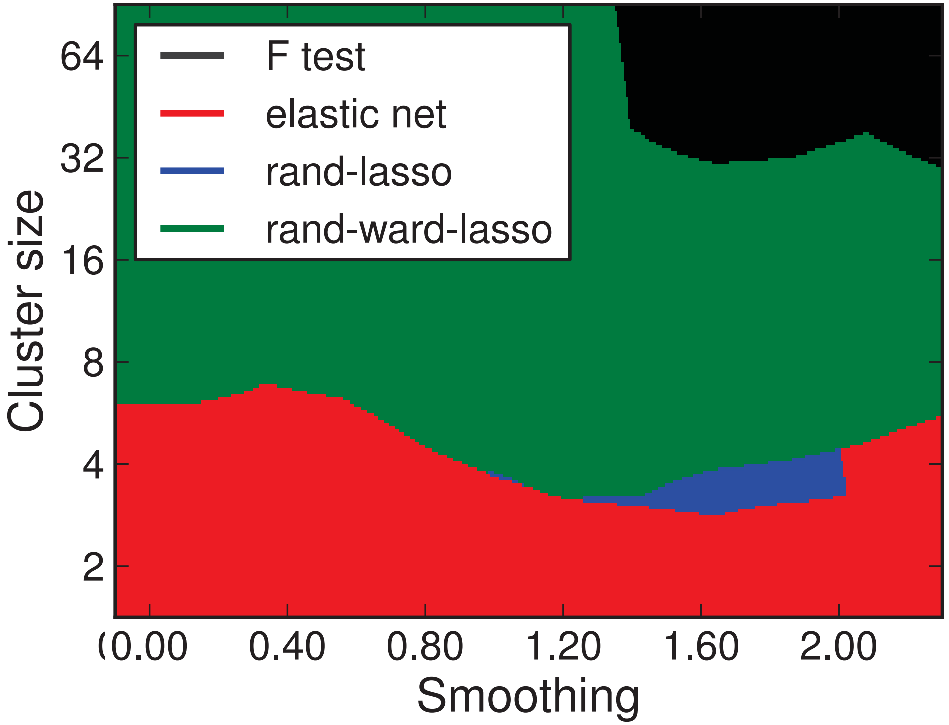
5 accumulate non-zero features



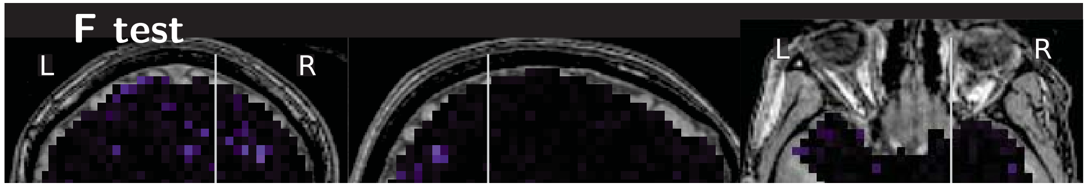
6 threshold map of apparition counts

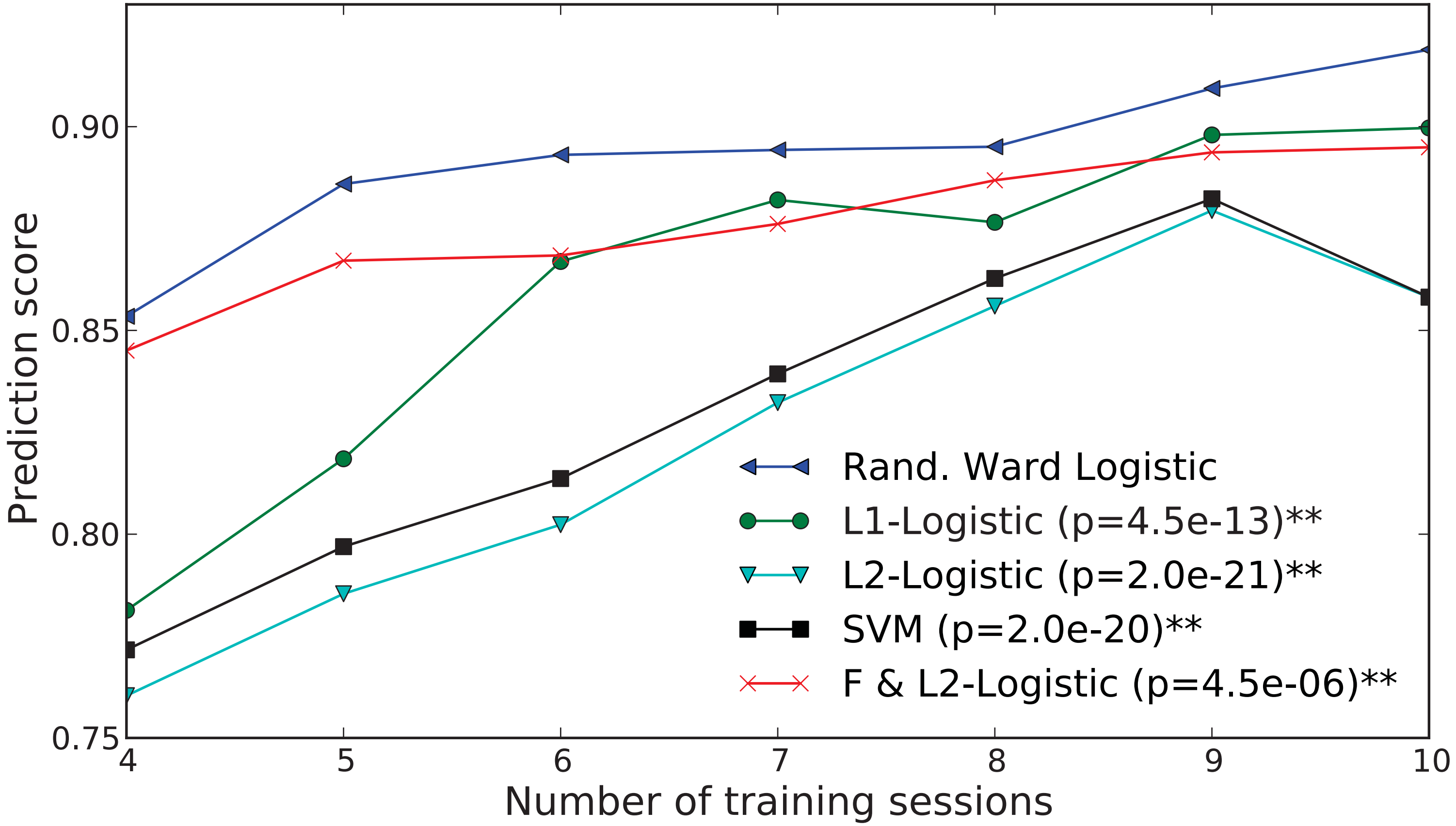
Simulations

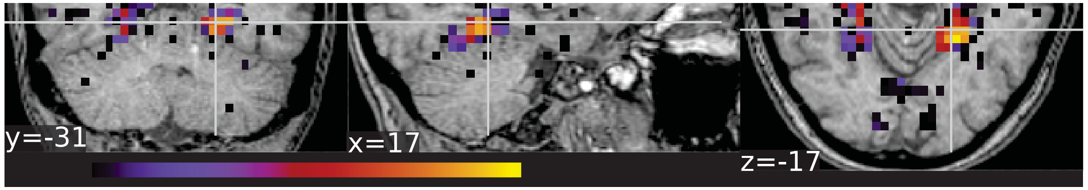
- $p = 2048$, $k = 64$, $n = 256$ ($n_{min} > 1000$)
- Weights w: patches of varying size
- Design matrix **X**: 2D Gaussian random images of varying smoothness



Results on [Haxby et al.]





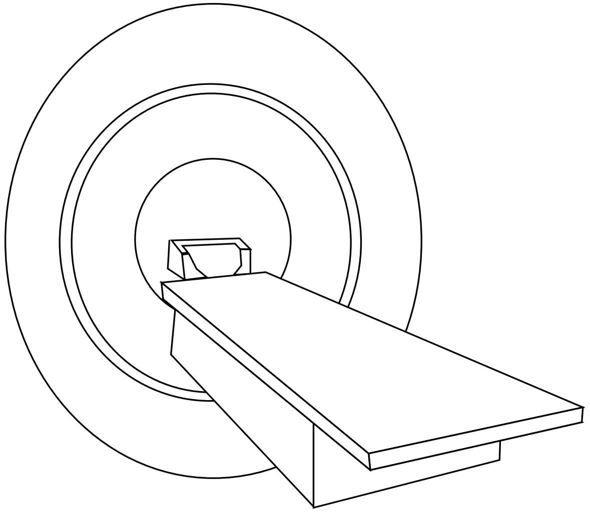


[ICML 2012]

Resting state fMRI: from networks to a population atlas

[Varoquaux, Gramfort et al. NIPS 2010
Varoquaux, Gramfort et al. IPMI 2011]

The context



fMRI resting state:

Subject with “no task” (eyes closed) for a few minutes (5 to 15 mins).

Why resting state:

- Easy to acquire
- Adapted to patients, infants

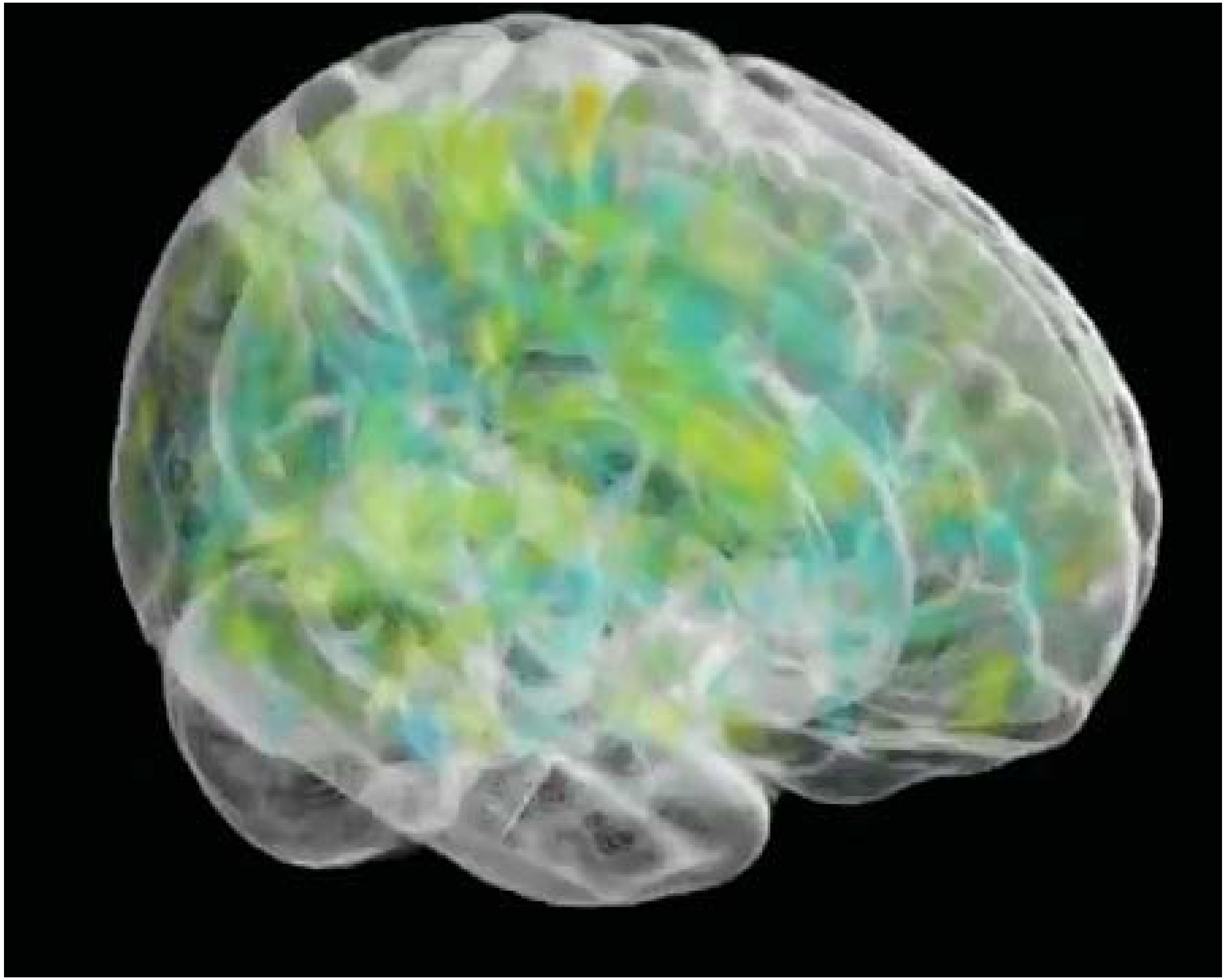
Challenge:

- Non-standard fMRI data
- Completely unsupervised
- Need new methodology

Question:

- We want to “learn” what is a “normal” resting state activity

Video of raw resting state data



courtesy of Gael Varoquaux

http://www.youtube.com/watch?v=uhCF-zlk0jY

The problem

Objective:

Estimate brain «networks» from **full brain** fMRI ongoing activity (resting state) **on a population.**

Definition [network]:

Regions that activate simultaneously, spontaneously or as an evoked response, **form an integrated network** that supports a specific cognitive function.

[Fox et al. Nat Rev Neurosci 2007, Bullmore Nat Rev Neurosci 2009, Smith PNAS 2009 ...]

The ingredients

- **Full brain**
- **Population** level **model**
- A **probabilistic model** where **likelihood of unseen data** can be tested and used for model selection with **cross-validation**
- **Gaussian graphical models** (special case of probabilistic graphical models with 2nd order statistics)
- Networks estimation using **graph partioning** with **modularity** criterion

From voxels to regions (ROIs)

Data are:

- co-registered to a template brain
- averaged within anatomically-defined regions

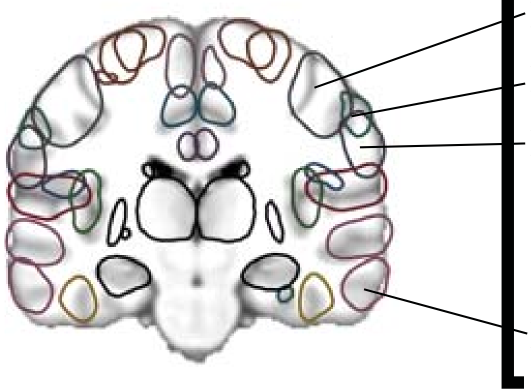
The atlas:

- 122 cortex ROIs (sulcal lines)
- 15 subcortical structures (FSL HO atlas)

THM:
A volume is *summarized* by $p=137$ values

[Perrot et al. IPMI (2009)]

Gaussian graphical model



p brain regions

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ x_p \end{bmatrix} \in \mathbb{R}^p \sim \mathcal{N}(0, \boldsymbol{\Sigma})$$

zero mean multivariate Gaussian distribution

$$p(x) = \frac{1}{(2\pi)^{p/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp(-\frac{1}{2} x^T \boldsymbol{\Sigma}^{-1} x)$$

let $\mathbf{K} = \boldsymbol{\Sigma}^{-1}$ *precision matrix*

taking the log of the likelihood gives:

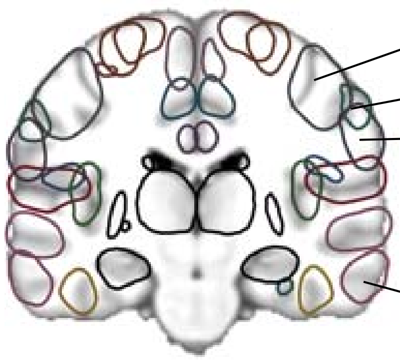
n brain volumes

$$\log(p(\mathbf{X})) = \frac{n}{2}\log(|\mathbf{K}|) - \frac{1}{2}\mathrm{tr}(\mathbf{X}^T\mathbf{K}\mathbf{X}) + cst \quad , \quad \mathbf{X} \in \mathbb{R}^{p \times n}$$

and:

$$\underbrace{\log(p(\mathbf{X})) = \frac{n}{2}\log(|\mathbf{K}|) - \frac{1}{2}\mathrm{tr}(\mathbf{K}(\overbrace{\mathbf{X}\mathbf{X}^T}^{\text{Data covariance}})) + cst}_{\text{Log-Likelihood}}$$

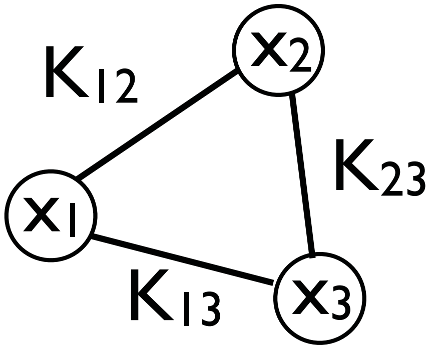
Graph and partial correlations



p brain regions

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdot \\ \cdot \\ x_p \end{bmatrix} \in \mathbb{R}^p \sim \mathcal{N}(0, \mathbf{\Sigma})$$

Let $\mathbf{K} = \mathbf{\Sigma}^{-1}$



$$p(x) = \frac{\sqrt{|\mathbf{K}|}}{(2\pi)^{p/2}} \exp(-\frac{1}{2} x^T \mathbf{K} x)$$

we have $$x^T \mathbf{K} x = \sum_{i,j} x_i \mathbf{K}_{ij} x_j$$

THM: The «connections» between x_i *and* x_j *are in* ***K***

Rq: It's the *partial correlations*

The challenges

- With 137 ROIs the covariance estimation requires to estimate (137x138)/2 =9 453 values

9 453 >> n≈250 (number of volumes for 1 subject)

THM: The estimation **problem is ill-posed**

Idea: To **increase n** take **more subjects**

Problem: Inter-subject variability

Remark: Even with **NO noise**, it is **ill-posed**

Single subject estimation

Penalized maximum likelihood:

$$\hat{\mathbf{K}}_{\ell_1} = \operatorname{argmin}_{\mathbf{K}\succ 0} \overbrace{\operatorname{tr}\left(\mathbf{K}\,\hat{\boldsymbol{\Sigma}}_{\text{sample}}\right) - \log\det\mathbf{K}}^{\text{Data fit}} + \overbrace{\lambda\|\mathbf{K}\|_1}^{\text{L1 Prior}}$$

where $\hat{\boldsymbol{\Sigma}}_{\text{sample}} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ and $\|\mathbf{K}\|_1 = \sum_{i\neq j}|\mathbf{K}_{ij}|$

Remark: It's a maximum a posteriori (MAP) estimate with i.i.d. Laplace prior on off-diagonal coefficients

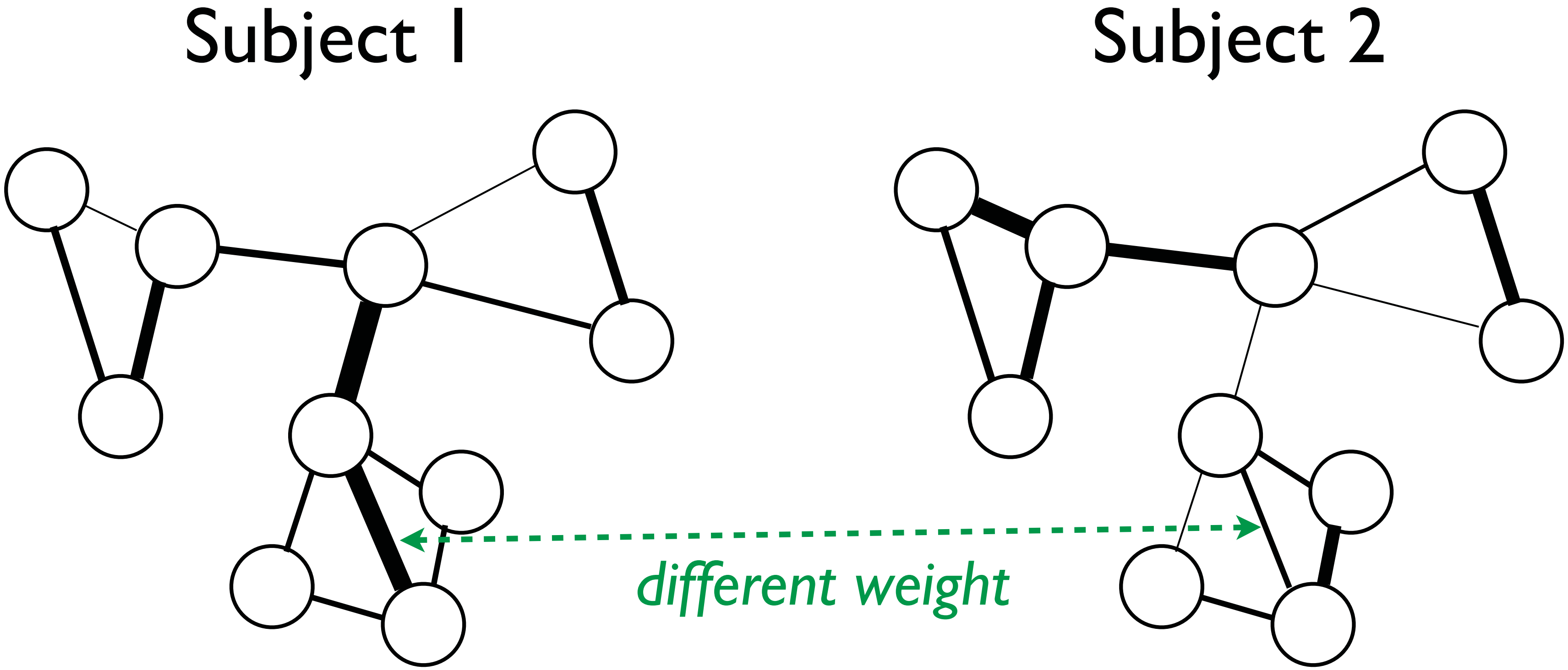
THM: L1 regularization promotes a weakly connected graph (sparse)

Optimization: Convex problem, cyclic descent

[A. Rothman, et al. : Sparse permutation invariant covariance estimation. Electron J Stat 2 (2008) 494]

Population level estimation

Idea: Promote the **same graph structure across the population** but **allow different weights** to take into account **inter-subject variability**



Population level estimation

Notations:

$\hat{\boldsymbol{\Sigma}}^{(s)}_{\text{sample}}$ is the empirical covariance for subject s

$\mathbf{K}^{(s)}$ is the precision for subject s

Optimization problem:

$$\left(\hat{\mathbf{K}}^{(s)}_{\ell_{21}}\right)_{s=1..S} = \mathrm{argmin}_{\mathbf{K}^{(s)} \succ 0} \left(\sum_{s=1}^{S} \left(\overset{\text{Data fit}}{\overline{\mathrm{tr}(\mathbf{K}^{(s)}\, \hat{\boldsymbol{\Sigma}}^{(s)}_{\text{sample}}) - \log\det \mathbf{K}^{(s)}}} \right) + \lambda \overset{\text{L1/L2 Prior}}{\overline{\sum_{i \neq j} \|\mathbf{K}^{(\cdot)}_{ij}\|_2}} \right)$$

$$\textstyle\sum_{i \neq j} \sqrt{\sum_{s=1}^{S} (\mathbf{K}^{(s)}_{ij})^2} = \sum_{i \neq j} \|\mathbf{K}^{(\cdot)}_{ij}\|_2 .$$

L1/L2 norm of off-diagonal terms

THM: The L1/L2 prior imposes the **same zeros** in **Ks** in the population (same graph edges for all subjects) but with **different weights**.

Data and preprocessing

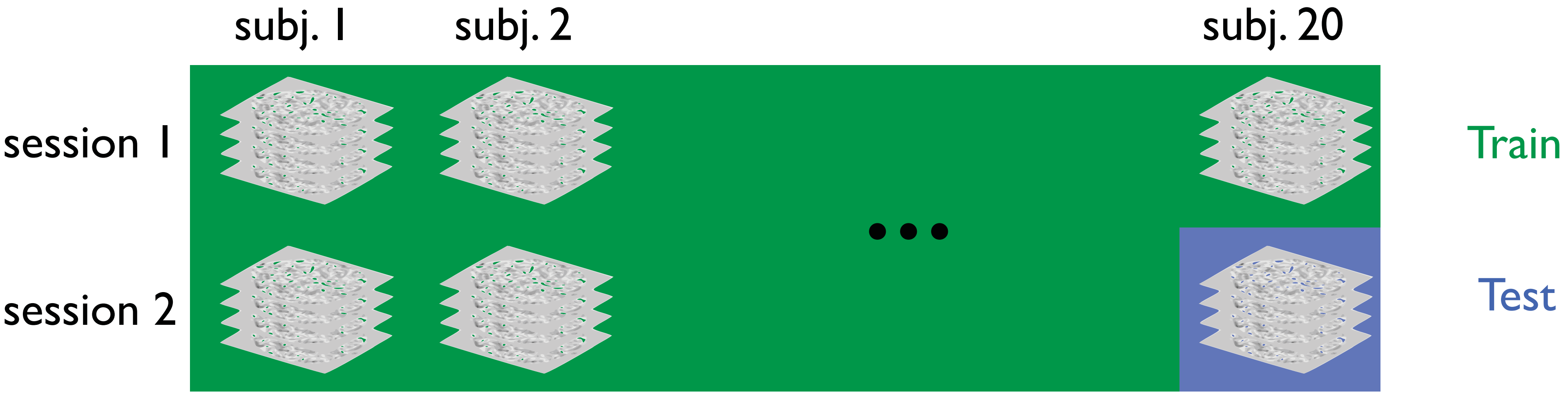
- 20 subjects
- 2 sessions with 244 volumes per session (TR 2.4s)
- Slice timing, motion correction, realignment with SPM5
- Confounds are regressed out (Ventricles, CSF, motion)
- 0.3 Hz low pass filter
- Removing of linear trend and unit variance to look at correlations

Remark: domain knowledge

Model selection

- **Leave one session out** (possibly informed by population data)
- **The likelihood of the left out session** is tested to **find the best regularization parameters**.

Example with session 2 of subj. 20 out:

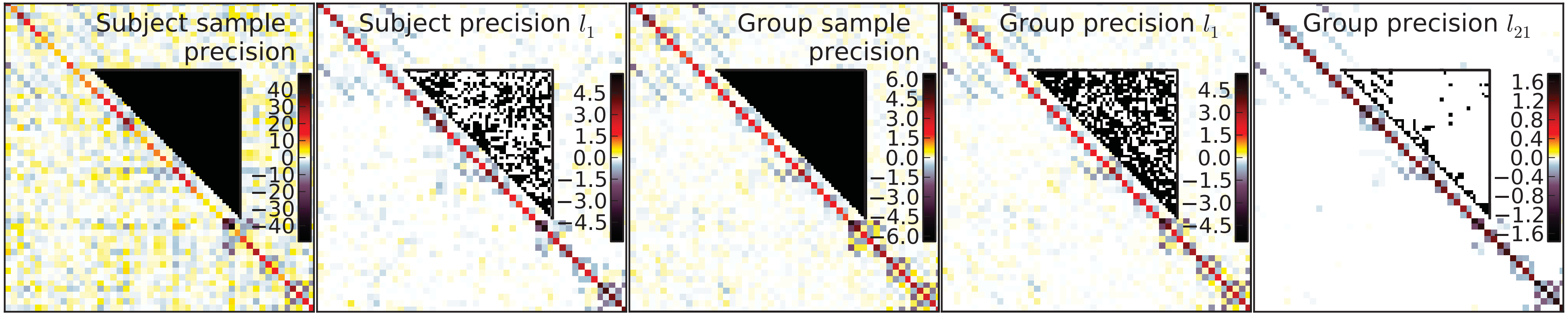


Results

Comparison between:

- MLE naive inverse
- L2 $\hat{\mathbf{K}}_{\ell_2} = (\hat{\mathbf{\Sigma}}_{\text{sample}} + \lambda \, \mathbf{I})^{-1}$
- LW *[Ledoit and Wolf 2004]*
- L1 individual subject
- L1 on concatenated data from all subjects
- L1/L2

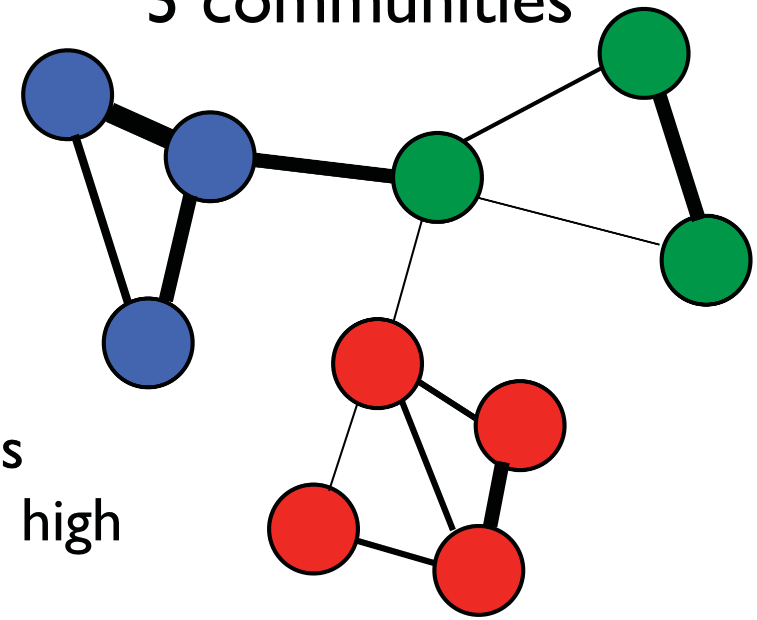
	Using subject data				Uniform group model				
	MLE	LW	ℓ_2	ℓ_1	MLE	LW	ℓ_2	ℓ_1	ℓ_{21}
Generalization likelihood	33.1	-57.1	38.8	43.0	40.6	41.5	41.6	41.8	45.6
Filling factor	100%	100%	100%	45%	100%	100%	100%	60%	8%



Communities and modularity

Now that we have the graph....

3 communities



Objective [clustering]:
Graph partioning that optimizes
modularity Q

Idea: Strong edges within clusters and few edges between clusters (functional specialization with high transport properties)

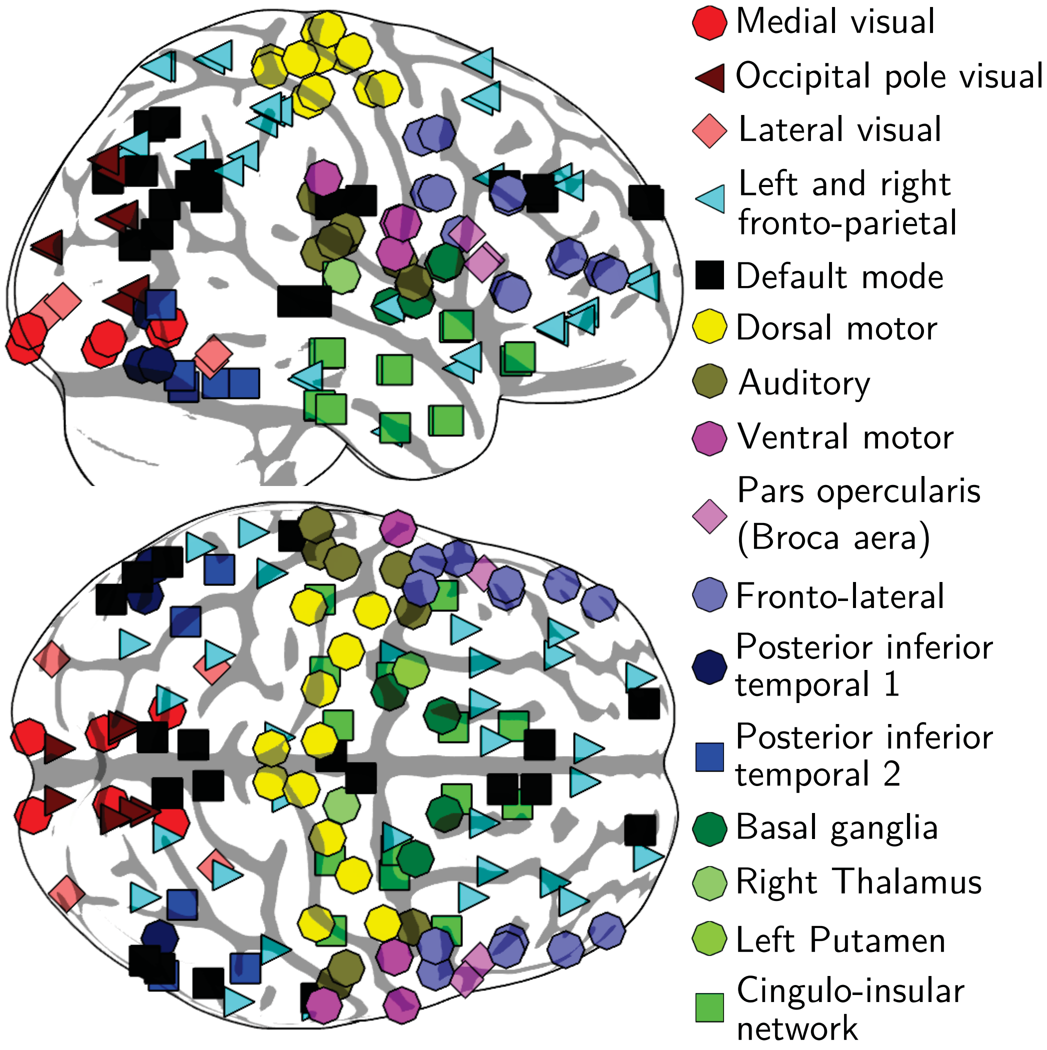
Approach:
Spectral clustering and **k-means** to maximize Q **based on the precision matrices** used as adjacency matrices

[M. Newman et al., Finding and evaluating community structure in networks. Phys rev E (2004)]

[M. Newman., Modularity and community structure in networks. PNAS (2006)]

[S. White and P. Smyth, A spectral clustering approach to finding communities in graphs. In: 5th SIAM international conference on data mining. (2005) 274]

Results



Graph is **clustered** in 16 communities **manually labelled**.

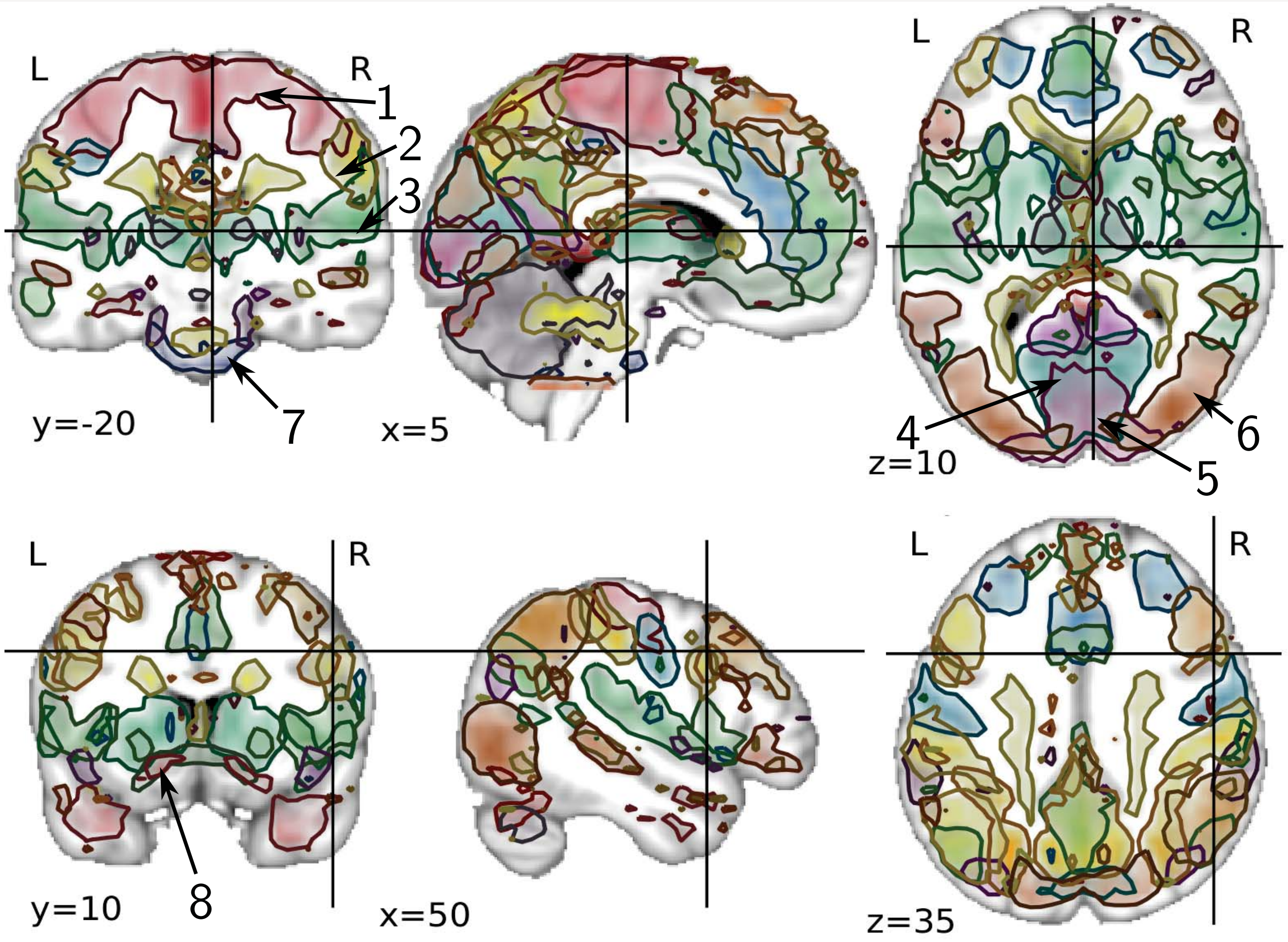
Take home messages

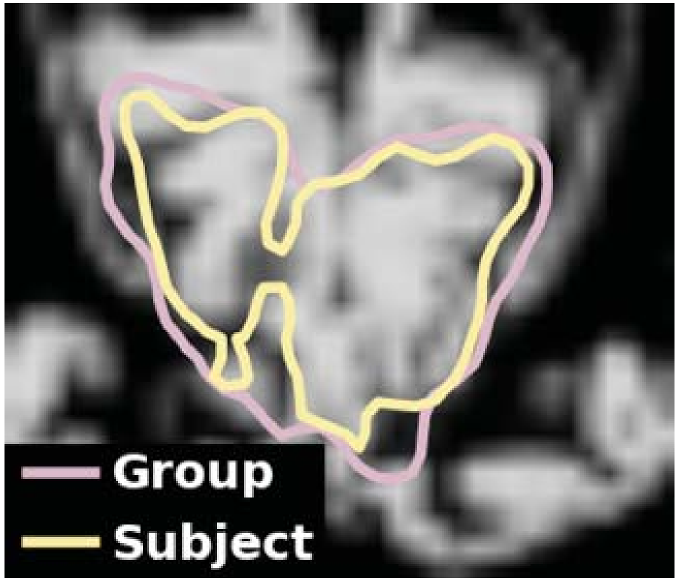
- When you have unsupervised problems, **with a likelihood** you can do proper selection by **cross-validation**
- Single-subject estimates give **poor fits due to estimation noise**
- Group-level estimates give **poor fits due to subject-variability**
- We improve on both by learning a **common structure across subjects** (shared independence structure)

But …

- **How do you learn the atlas and the ROIs** allowing inter-subject variability from the fMRI data

Dictionary learning to learn the ROIs





$$\Omega_{\mathrm{SL}}(\mathbf{v}) = \|\mathbf{v}\|_1 + \frac{1}{2}\mathbf{v}^T\mathbf{L}\mathbf{v}$$

$$\sum_{s=1}^{S} \frac{1}{2}\left(\|\mathbf{Y}^s - \mathbf{U}^s\mathbf{V}^{s\,T}\|_{\mathrm{Fro}}^2 + \mu\|\mathbf{V}^s - \mathbf{V}\|_{\mathrm{Fro}}^2 \right) + \lambda\,\Omega(\mathbf{V}) \ \ \text{s.t.}\ \|\mathbf{u}_l^s\|_2^2 \leq 1$$

[Varoquaux G., Gramfort A., J.B. Poline, B. Thirion, *IPMI 2011*]

Minimized with cyclic optimization

```
Input: ${\mathbf{Y}^s \in \mathbb{R}^{n \times p}, s = 1, \dots, S\}$, the time series for each subject; $k$, the number of maps; an initial guess for $\mathbf{V}$.
Output: $\mathbf{V} \in \mathbb{R}^{p \times k}$ the group-level spatial maps, $\{\mathbf{V}^s \in \mathbb{R}^{p \times k}\}$ the subject-specific spatial maps, $\{\mathbf{U}^s \in \mathbb{R}^{n \times k}\}$ the associated time series.
$E_0 \leftarrow \infty$, $E_1 \leftarrow \infty$, $i \leftarrow 1$ (initialize variables).
$\mathbf{V}^s \leftarrow \mathbf{V}$,   $\mathbf{U}_s \leftarrow \mathbf{Y}^s \mathbf{V}(\mathbf{V}^T \mathbf{V})^{-1}$,   for $s = 1 \dots S$
while $E_i - E_{i-1} > \varepsilon E_{i-1}$ do
  for s=1 to S do
    for l=1 to k do
      Update $\mathbf{U}^s$:   $\mathbf{u}_l^s \leftarrow \mathbf{u}_l^s + \|\mathbf{v}_l^s\|_2^{-2} (\mathbf{Y}^s (\mathbf{v}_l^s - \mathbf{U}^s \mathbf{V}^{sT} \mathbf{v}_l^s)$        Rank 1 update
      $\mathbf{u}_l^s \leftarrow \mathbf{u}_l^s / \max(\|\mathbf{u}_l^s\|_2, 1))$
    end for
    Update $\mathbf{V}^s$ (ridge regression): $\mathbf{V}^s \leftarrow \mathbf{V} + (\mathbf{Y}^s - \mathbf{U}^s \mathbf{V}^T)^T \mathbf{U}^s (\mathbf{U}^{sT} \mathbf{U}^s + \mu \mathbf{I})^{-1}$        Ridge
  end for
  Update $\mathbf{V}$ using lemma 1:   $\mathbf{V} \leftarrow \mathrm{prox}_{\lambda/S\mu\,\Omega} \left(\frac{1}{S} \sum_{s=1}^{S} \mathbf{V}^s\right)$.        Prox
  Compute value of energy:   $E_i \leftarrow \mathcal{E}(\mathbf{U}^s, \mathbf{V}^s, \mathbf{V})$
  $i \leftarrow i + 1$
end while
```

[Varoquaux G., Gramfort A., J.B. Poline, B. Thirion, *IPMI 2011*]

Conclusion

Conclusion

- **Sparse methods are great tools but there are a few caveats**:
  - Pure L1 is often not enough. You need to enforce the good structure
  - If you know you look for a sparse solution use it to be faster
  - You should promote sparsity in the right "basis" (representation)
  - Prediction (reconstruction error) is different from support recovery
- To **make something really work**:
  - a lot of domain knowledge
  - understand, adapt and improve ideas emerging in other fields (goes in both ways)
  - good software engineering: integrate your contributions/code in existing software packages to reach users.

The human inverse problem

Observations

Sparse, Convex optimization, STFT,

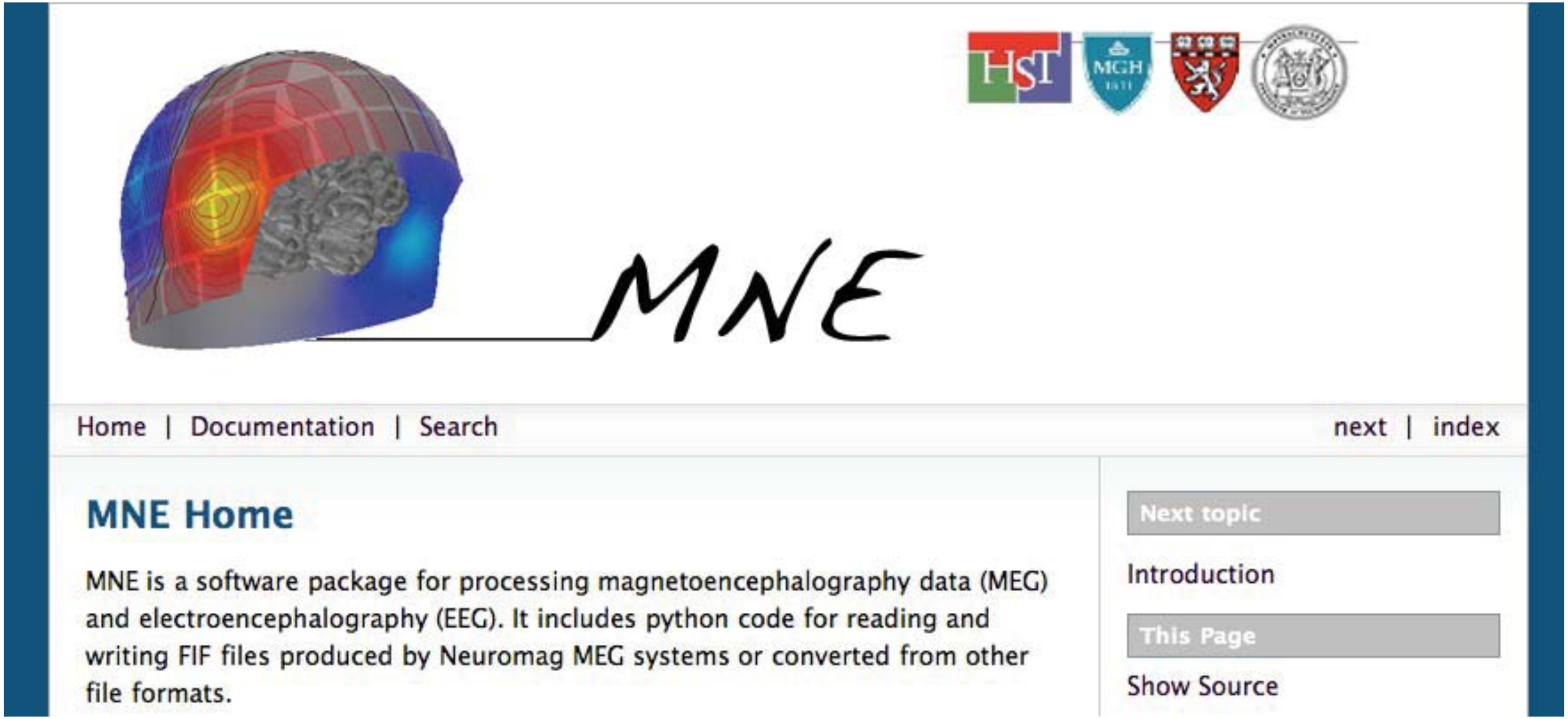
Proximal iterations,

etc...

brain
imaging
people

How do you solve this inverse problem?

For MEG



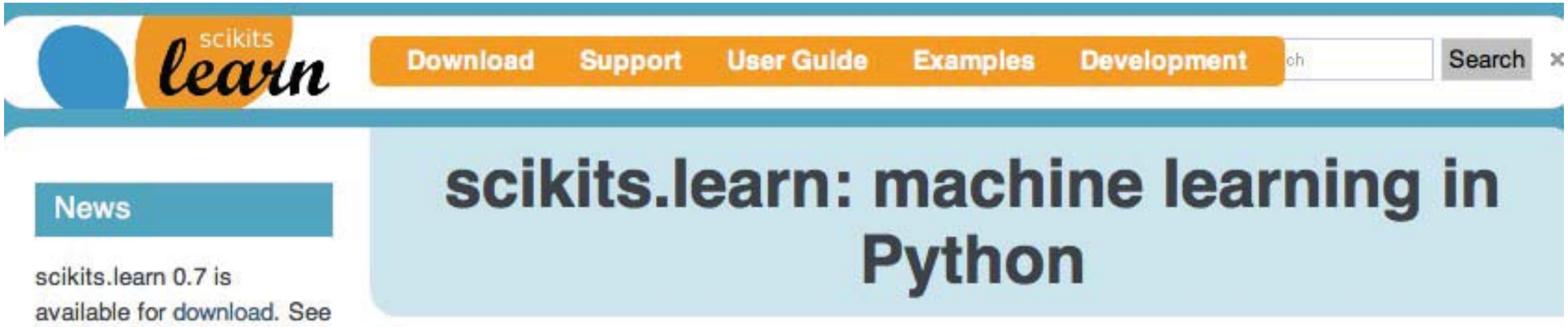
MNE Home

MNE is a software package for processing magnetoencephalography data (MEG) and electroencephalography (EEG). It includes python code for reading and writing FIF files produced by Neuromag MEG systems or converted from other file formats.

http://www.martinos.org/mne

http://www.github.com/mne-tools

Machine learning

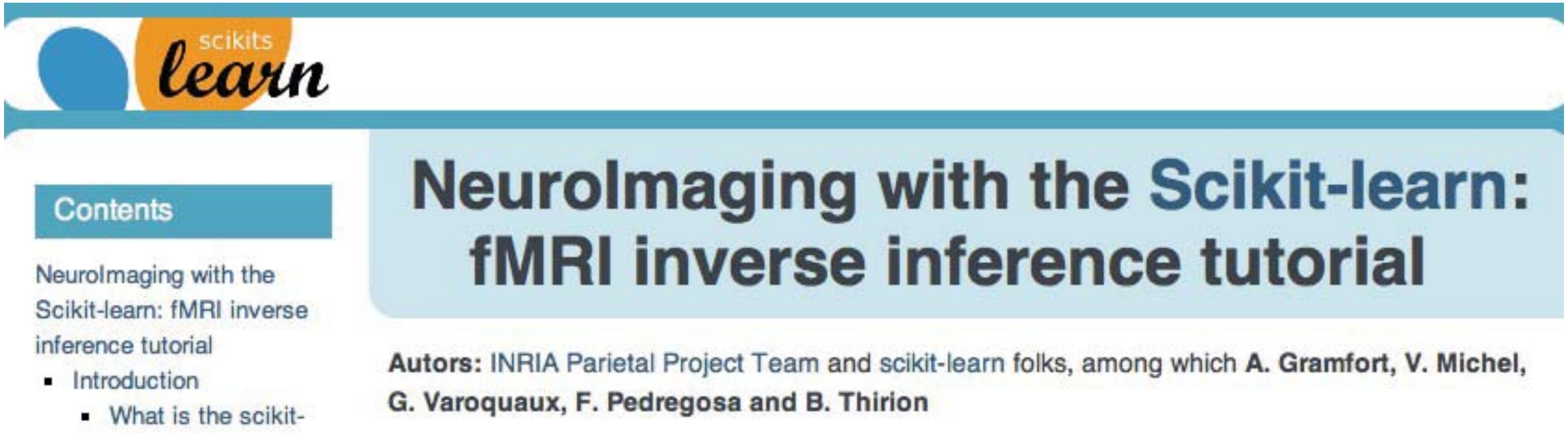


scikits.learn: machine learning in Python

News

scikits.learn 0.7 is available for download. See

http://scikit-learn.org



NeuroImaging with the Scikit-learn: fMRI inverse inference tutorial

Contents

NeuroImaging with the Scikit-learn: fMRI inverse inference tutorial

- Introduction
  - What is the scikit-

Autors: INRIA Parietal Project Team and scikit-learn folks, among which **A. Gramfort, V. Michel, G. Varoquaux, F. Pedregosa and B. Thirion**

http://nisl.github.com/

[Pedregosa et al. JMLR 2011]

Contact:

Alexandre Gramfort
alexandre.gramfort@telecom-paristech.fr

Collaborators:

@INRIA:
B. Thirion, G. Varoquaux
V. Michel, F. Pedregosa
F. Bach, R. Jenatton
G. Obozinski
M. Clerc, T. Papadopoulo

M. Hämäläinen, MGH / Harvard / MIT, Boston
M. Kowalski, L2S, Univ. Paris-sud
D. Strohmeier & J. Haueisen, TU Ilmenau, Germany

References

[Gramfort et al. IPMI 2011,
Gramfort et al. PMB 2012,
Jenatton & Gramfort et al. SIAM IS 2012,
Varoquaux & Gramfort NIPS 2010,
Varoquaux & Gramfort IPMI 2011
Gramfort et al. NIPS workshop 2011
Varoquaux & Gramfort ICML 2012]