Donders Institute
for Brain, Cognition and Behaviour

Brain Reading (MKI43)

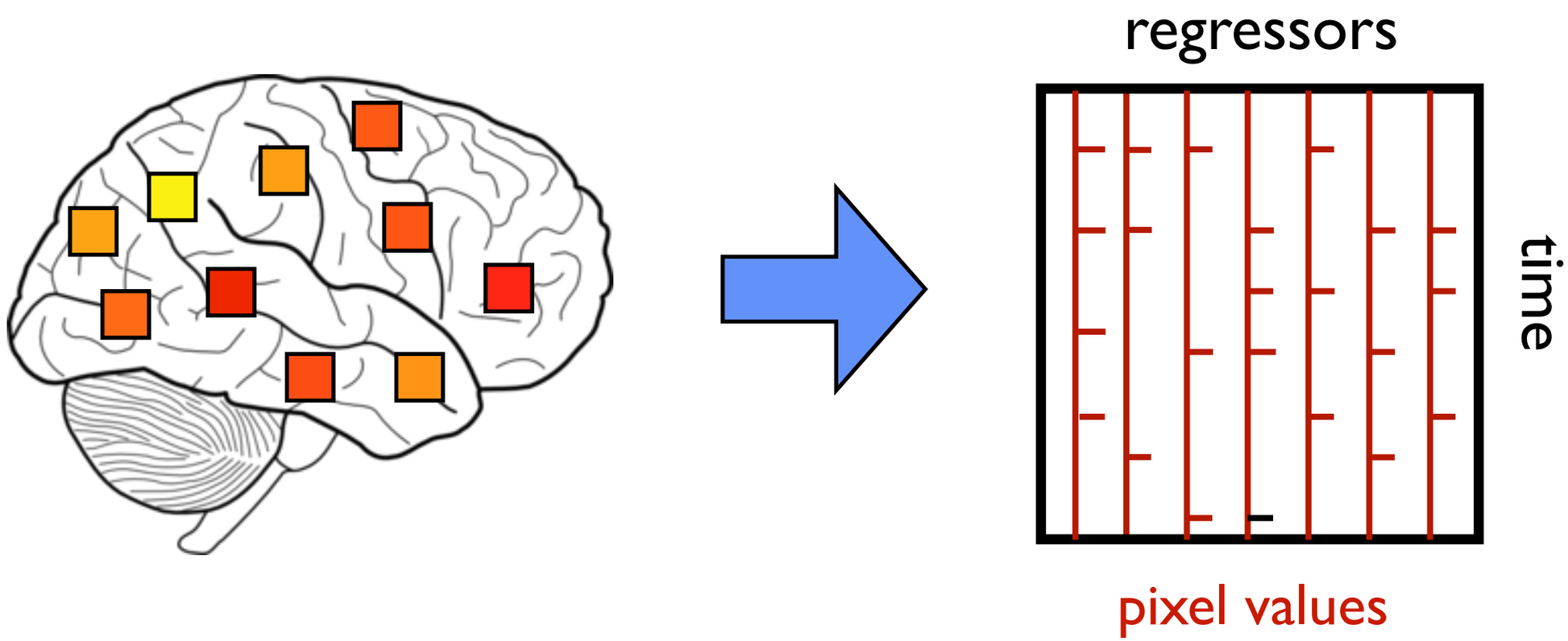
# Lecture 4: Generative decoding

Marcel van Gerven
Assistant Professor
Distributed Representations Group
Donders Centre for Cognition

Radboud University Nijmegen

# Discriminative decoding

Predict from a very high-dimensional input (fMRI voxels) to a (possibly) very high-dimensional output (image pixels).

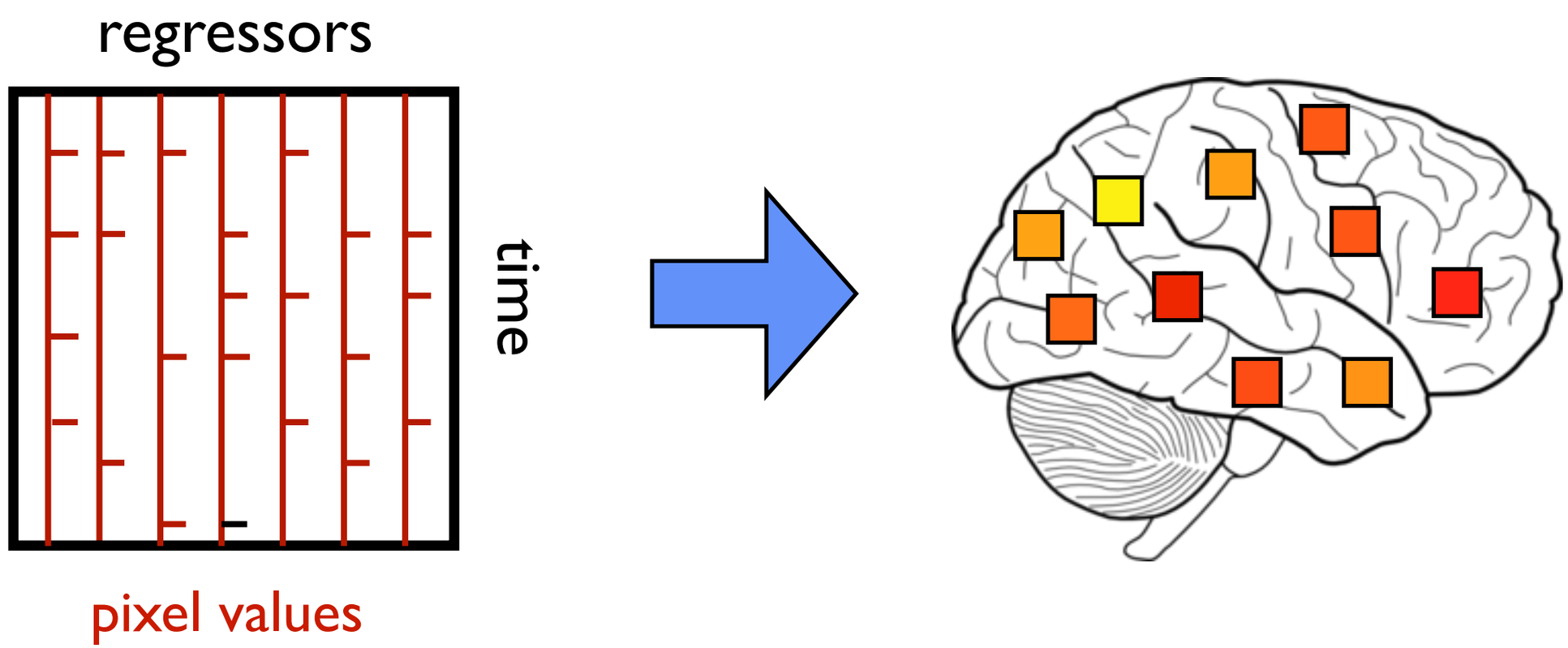



$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{y})$$

Discriminative approach

# Generative decoding

Predict from a very high-dimensional input (image pixels) to a (possibly) very high-dimensional output (voxels).

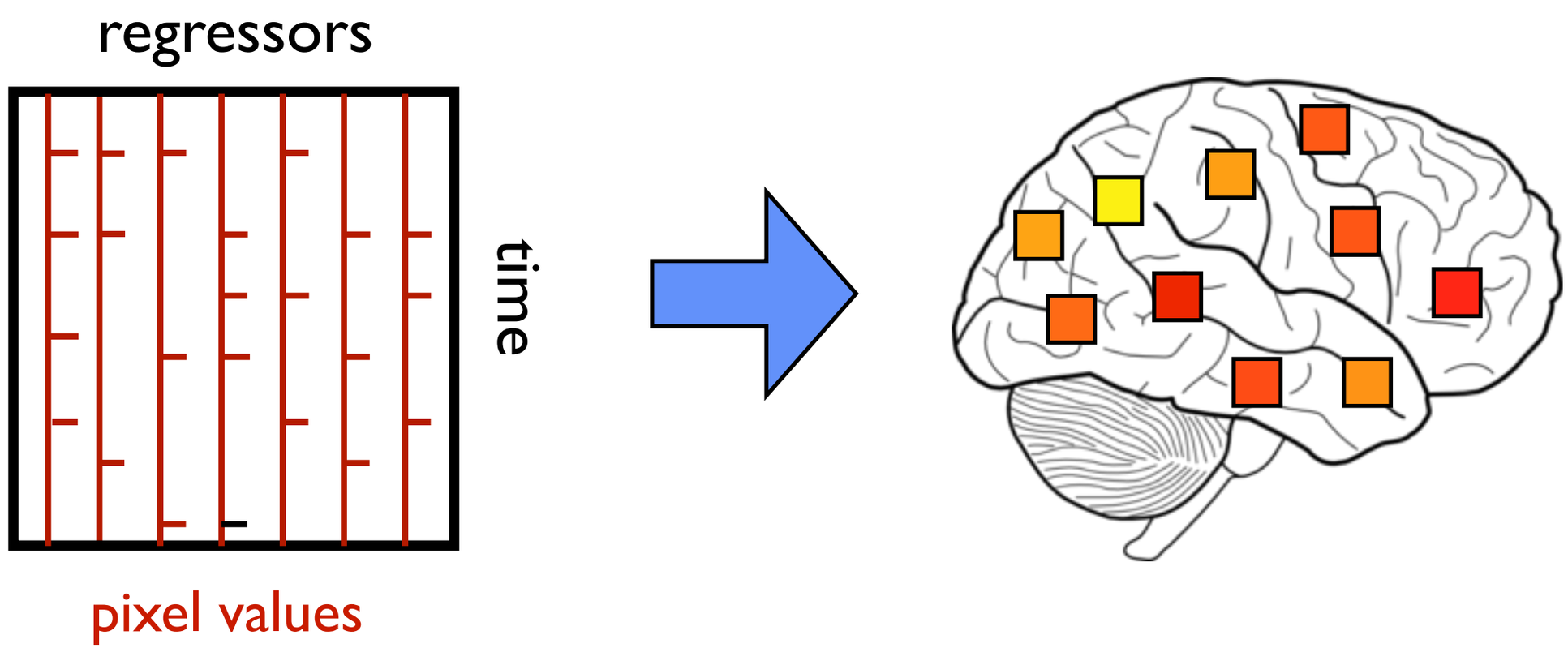



$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{y})$$ Discriminative approach

$$= \arg\max_{\boldsymbol{x}} \{p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})\}$$ Generative approach

# Maximum likelihood approach

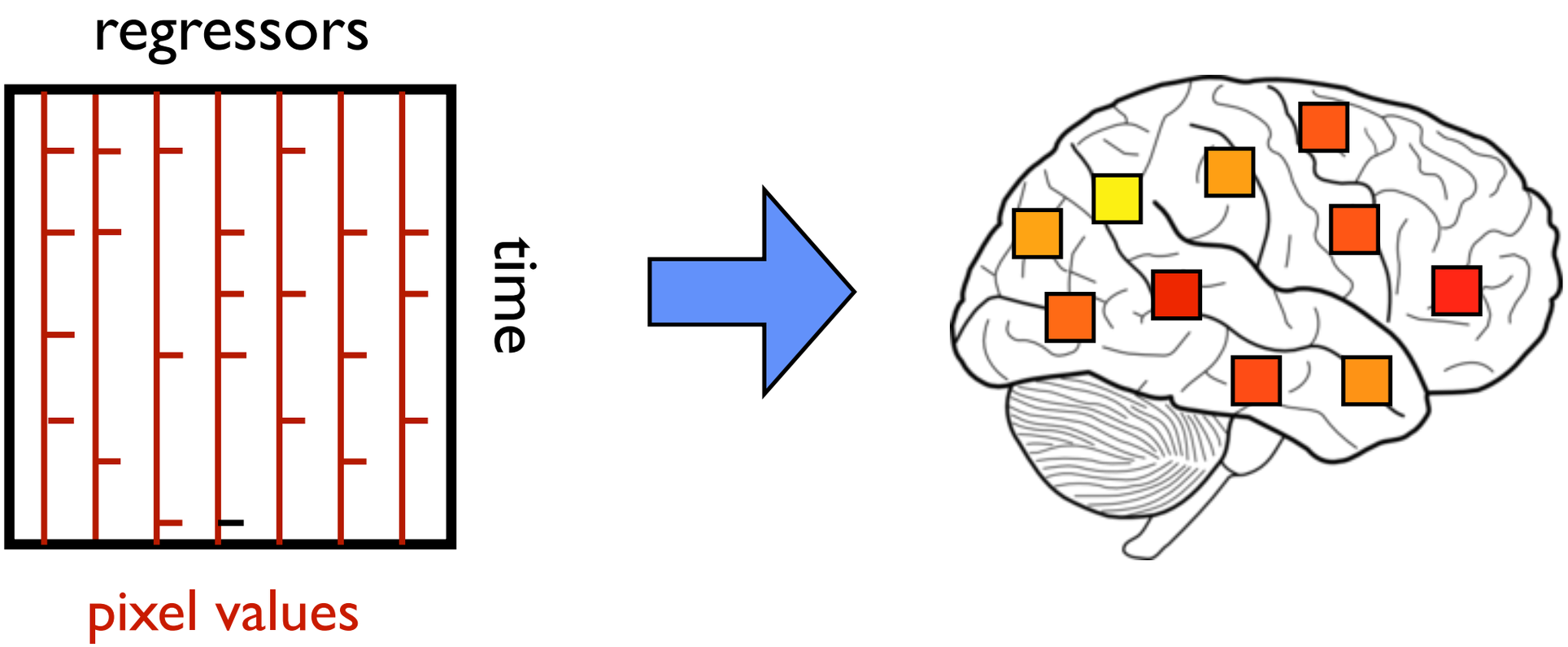
If each stimulus is equally likely, we have again the maximum likelihood solution (wrt the stimulus instead of the model parameters)...




$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} p(\boldsymbol{y} \mid \boldsymbol{x})$$

# Identification

Suppose we just make one prediction of the response $\boldsymbol{y} = f(\boldsymbol{x})$




Suppose we have a set of *N* stimuli. Then, the best reconstruction is given by

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}_n} g(\boldsymbol{y}_n, f(\boldsymbol{x}_n))$$

for some performance measure *g*. e.g., residual variance

# Identification

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}_n} g(\boldsymbol{y}_n, f(\boldsymbol{x}_n))$$

1. Build rich representation of input stimulus
2. Learn how the representation maps to neural response
3. Predict the response for each of a set of candidate stimuli using 1 and 2
4. Select the stimulus whose predicted response best matches the observed response according to *g*

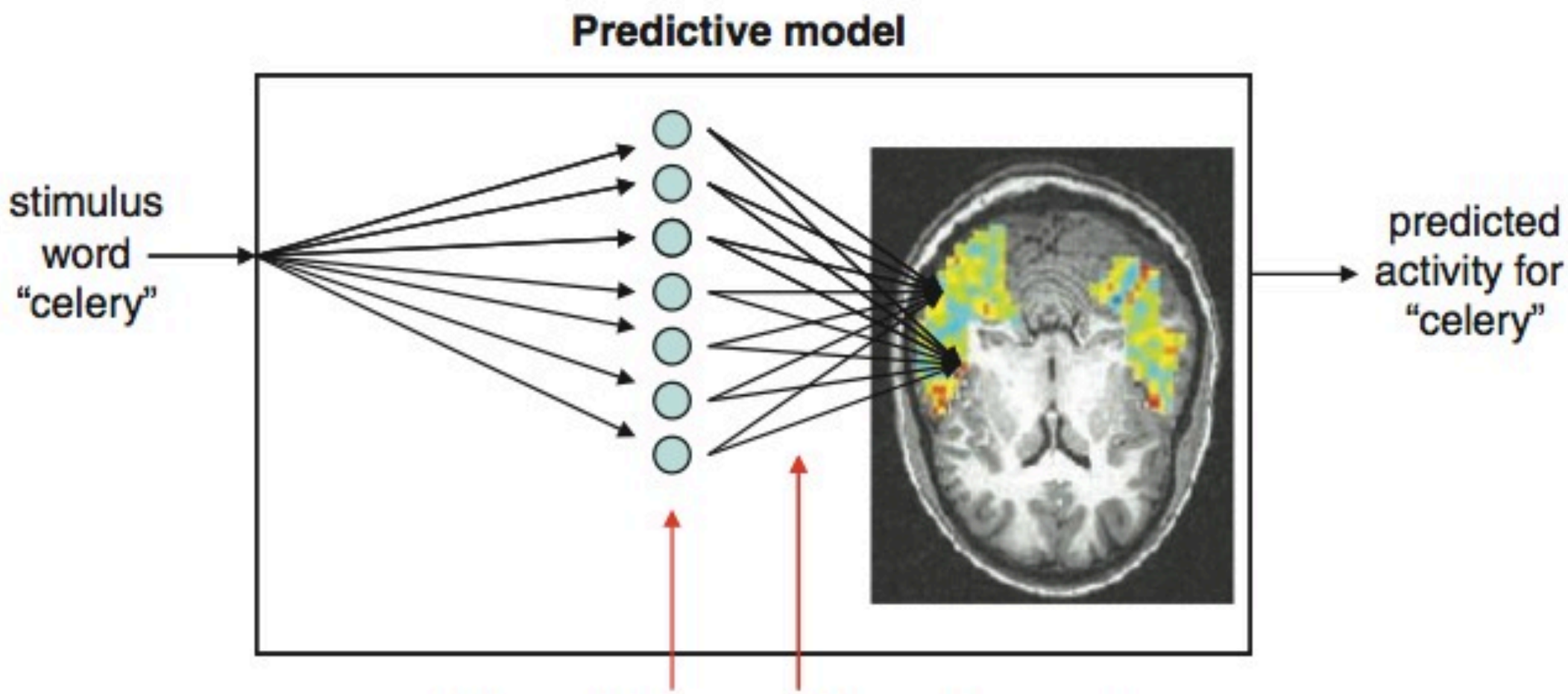
# Predicting Human Brain Activity Associated with the Meanings of Nouns

Tom M. Mitchell,[1*] Svetlana V. Shinkareva,[2] Andrew Carlson,[1] Kai-Min Chang,[3,4] Vicente L. Malave,[5] Robert A. Mason,[3] Marcel Adam Just[3]

The question of how the human brain represents conceptual knowledge has been debated in many scientific fields. Brain imaging studies have shown that different spatial patterns of neural activation are associated with thinking about different semantic categories of pictures and words (for example, tools, buildings, and animals). We present a computational model that predicts the functional magnetic resonance imaging (fMRI) neural activation associated with words for which fMRI data are not yet available. This model is trained with a combination of data from a trillion-word text corpus and observed fMRI data associated with viewing several dozen concrete nouns. Once trained, the model predicts fMRI activation for thousands of other concrete nouns in the text corpus, with highly significant accuracies over the 60 nouns for which we currently have fMRI data.

Mitchell et al. Science, 2008

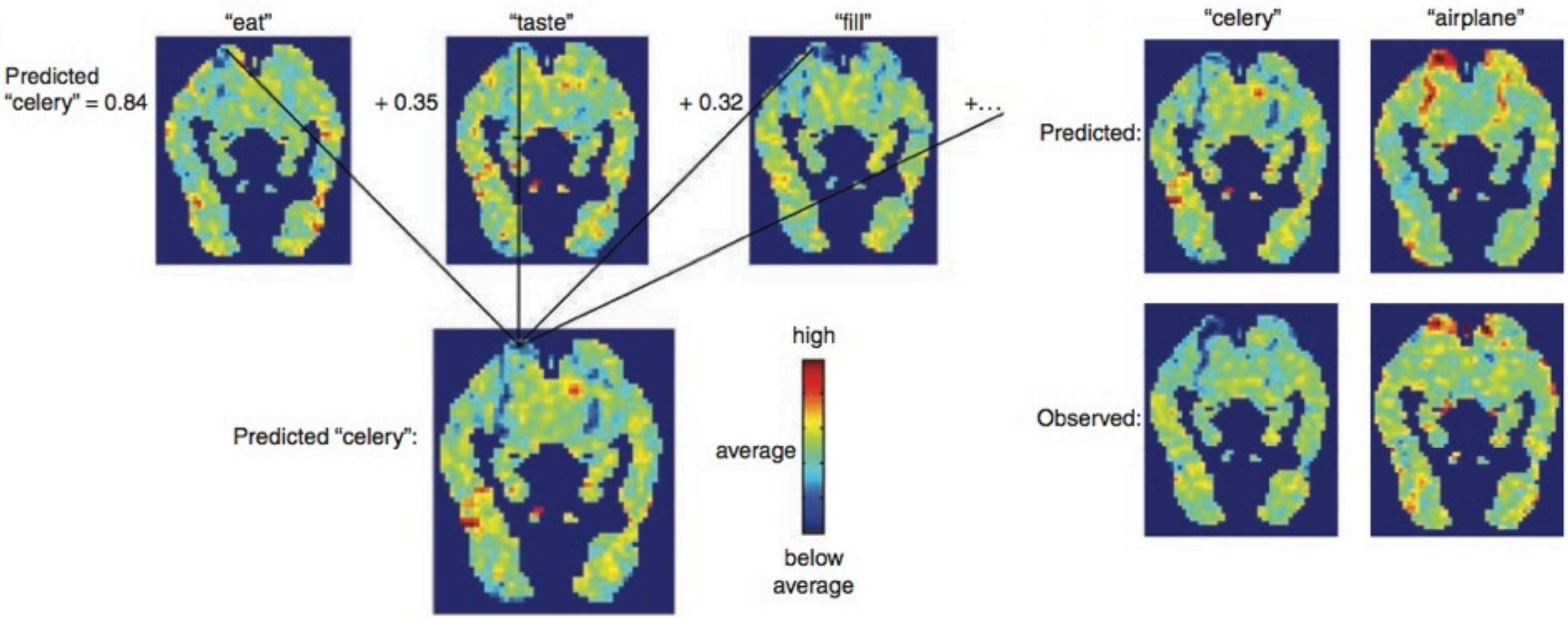
# Example: Concept decoding



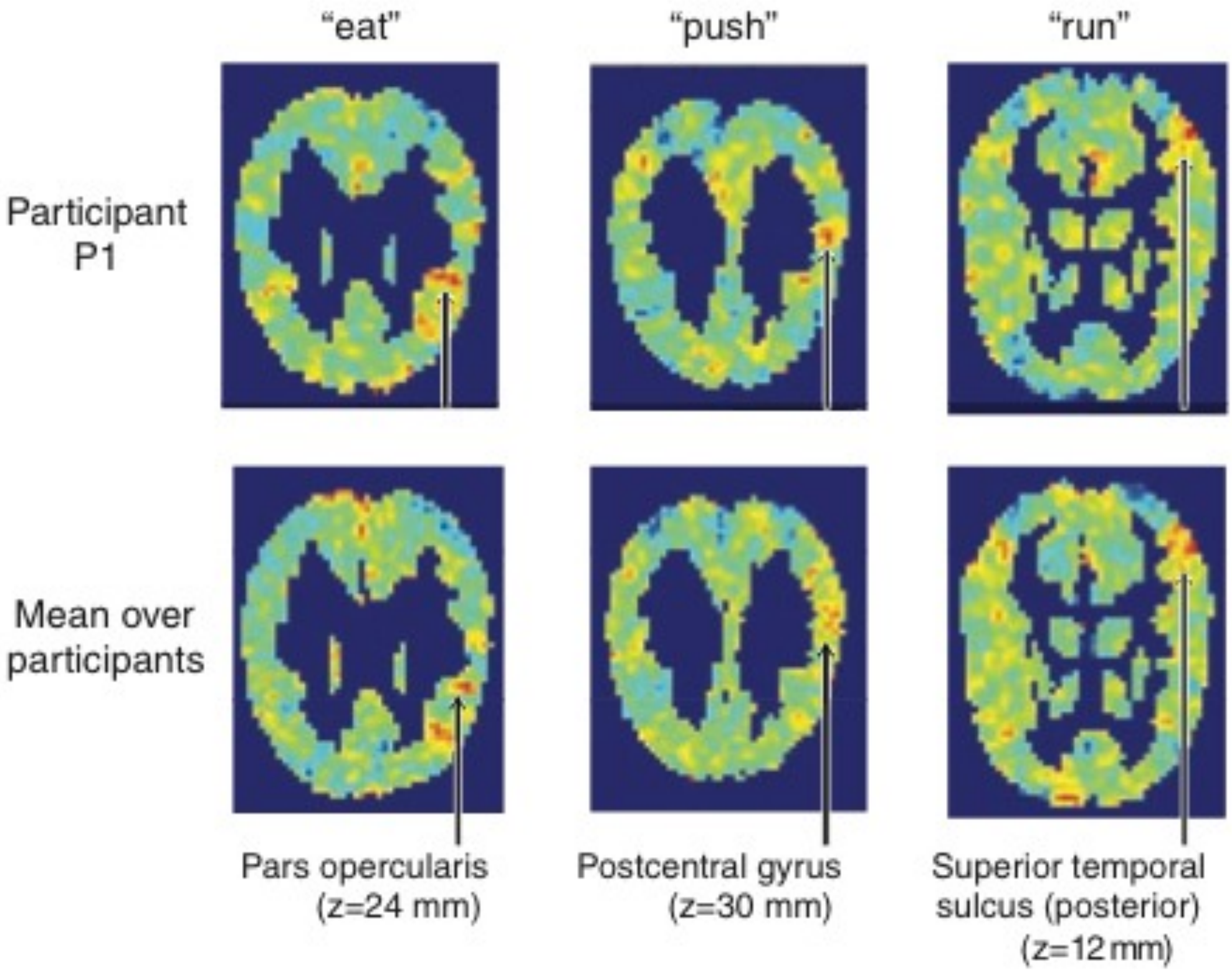

$$y_v = \sum_{i=1}^{n} c_{vi} f_i(w)$$

Mitchell et al. Science, 2008

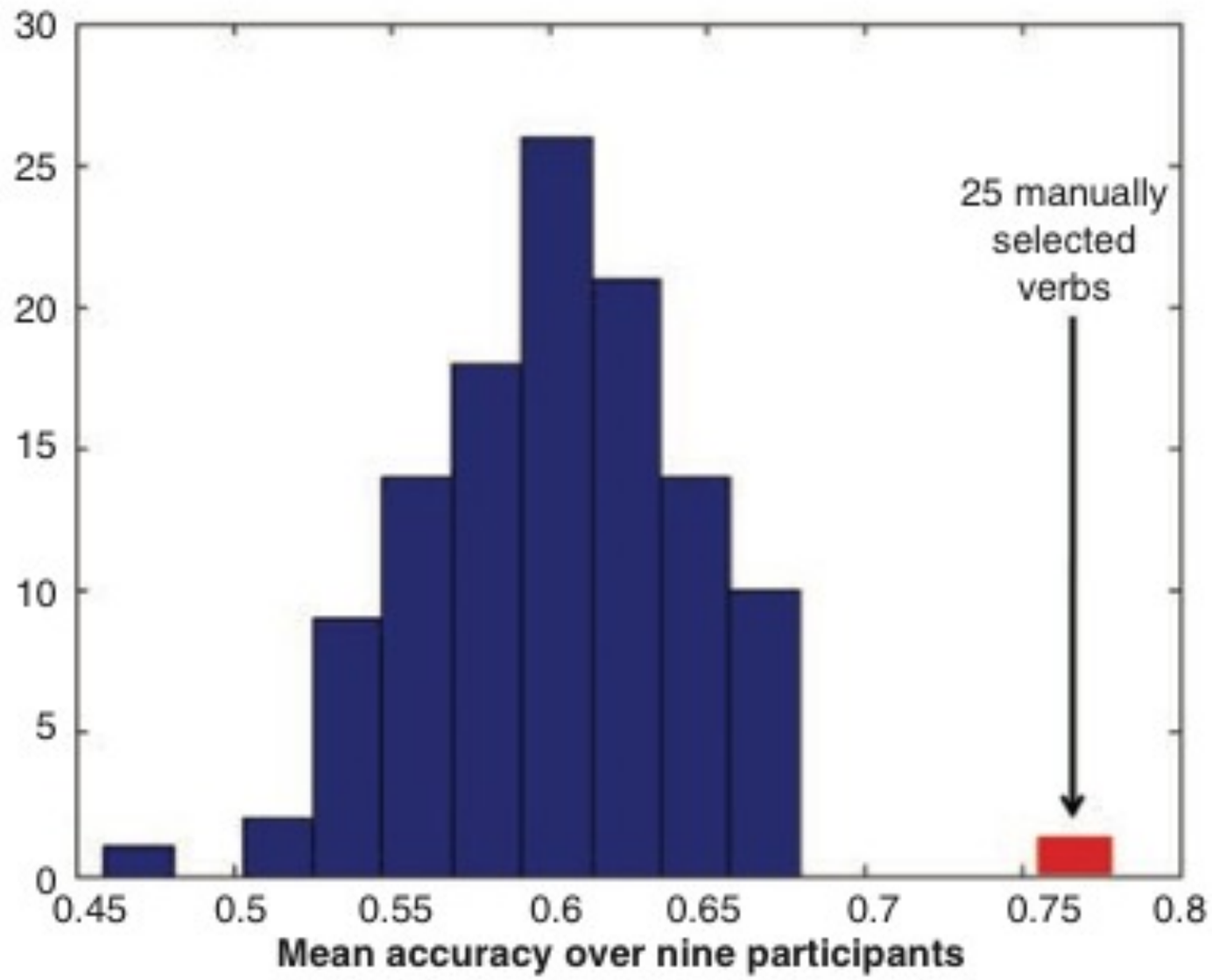
# Example: Concept decoding



# Localization



# Feature set


30
25
20
15
10
5
0
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
Mean accuracy over nine participants
25 manually selected verbs

2012 Second International Workshop on Pattern Recognition in NeuroImaging

# A Linear Gaussian Framework for Decoding of Perceived Images

Marcel A. J. van Gerven
*Radboud University Nijmegen*
*Donders Institute for Brain, Cognition and Behaviour*
*Nijmegen, The Netherlands*
*Email: m.vangerven@donders.ru.nl*

Tom Heskes
*Radboud University Nijmegen*
*Institute for Computing and Information Sciences*
*Nijmegen, The Netherlands*
*Email: tomh@cs.ru.nl*

***Abstract*—With the advent of sophisticated acquisition and analysis techniques, decoding the contents of someone's experience has become a reality. We propose a simple linear Gaussian framework where decoding relies on the inversion of properly regularized encoding models. We show that this approach yields state-of-the-art decoding performance on an fMRI dataset.**

***Keywords*-Bayesian decoding, perception, fMRI analysis**

assigns a uniform probability to images in a predefinec and zero probability to all other images. This essent allows the decoding to be performed by the forward m only, without the need for inverse inference.

In this paper we present a framework for decoding expands on the ideas put forward in the aforementic papers. Specifically, similar to [11], we assume that

Develop a simple probabilistic generative framework for percept reconstruction

forward model: how does a stimulus cause a response?

$$x^* = \arg\max \{p(y \mid x)p(x)\}$$

image prior: what do stimuli typically look like?

# Stimulus set

handwritten characters

Miyawaki et al., Neuron, 2008

Kay et al, Nature, 2008

# Gaussian generative decoding of handwritten characters

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} \{p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})\}$$

We need two ingredients:

likelihood model: $p(\boldsymbol{y} \mid \boldsymbol{x})$

prior: $p(\boldsymbol{x})$

# Gaussian generative decoding of handwritten characters

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} \{p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})\}$$

We need two ingredients:

likelihood model: $p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{B}^T\boldsymbol{x}, \boldsymbol{\Sigma})$

prior: $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{0}, \boldsymbol{R})$

Simplest option: assume everything is Gaussian!

## Likelihood model

$$
\begin{aligned}
p(\boldsymbol{y} \mid \boldsymbol{x}) &= \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{B}^T\boldsymbol{x}, \boldsymbol{\Sigma}) \\
&\propto \exp\left( (\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\boldsymbol{y})^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}^T\boldsymbol{x} \right)
\end{aligned}
$$

canonical form

where we assume that the predicted voxel responses are independent

Can be written equivalently as

$$
\mathbf{y} = \mathbf{B}^\top\mathbf{x} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}; \boldsymbol{\Sigma})
$$

with $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_q)$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2)$.

# Prior

$$
\begin{aligned}
p(\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{y} \mid \mathbf{0}, \boldsymbol{R}) \\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{R}^{-1} \boldsymbol{x}\right)
\end{aligned}
$$

Models the statistical properties of the outside world

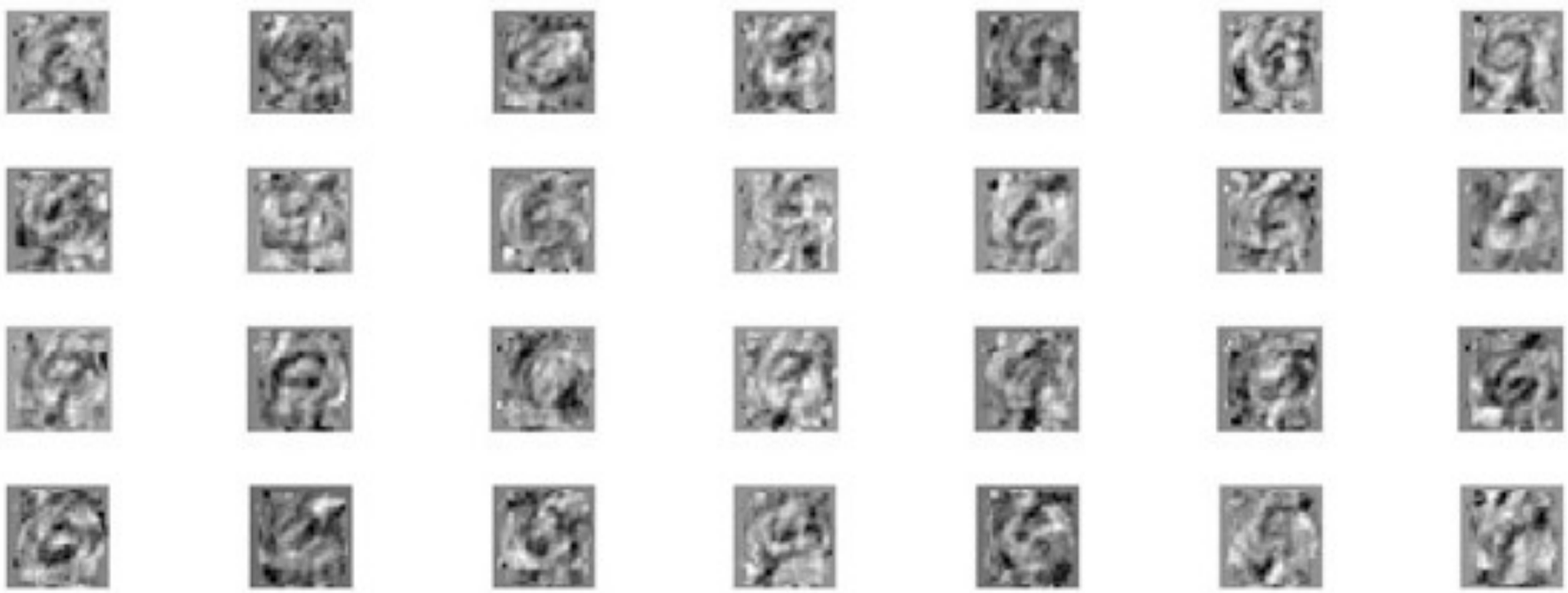
ideal samples from p(**x**):

# Prior

$$
\begin{aligned}
p(\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{y} \mid \mathbf{0}, \boldsymbol{R}) \\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{R}^{-1} \boldsymbol{x}\right)
\end{aligned}
$$

Models the statistical properties of the outside world

actual samples from p(**x**):



Gaussian priors oversmooth and do not accurately model image statistics...

# Putting it all together

$$p(\boldsymbol{y} \mid \boldsymbol{x}) \quad \propto \quad \exp\left( (\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\boldsymbol{y})^T \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T \boldsymbol{B}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}^T \boldsymbol{x} \right)$$

$$p(\boldsymbol{x}) \quad \propto \quad \exp\left( -\frac{1}{2}\boldsymbol{x}^T \boldsymbol{R}^{-1}\boldsymbol{x} \right)$$

Such that

$$p(\mathbf{x} \mid \mathbf{y}) \propto \exp\left( \left(\mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{y}\right)^\top \mathbf{x} - \frac{1}{2}\mathbf{x}^\top \left(\mathbf{R}^{-1} + \mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{B}^\top\right)\mathbf{x} \right)$$

giving the MAP estimate

$$\hat{\mathbf{x}} = \mathbf{m} = \left(\mathbf{R}^{-1} + \mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{B}^\top\right)^{-1} \mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

# Some tricks

MAP estimate

$$\hat{\mathbf{x}} = \mathbf{m} = \left(\mathbf{R}^{-1} + \mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{\top}\right)^{-1}\mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

Matrix inversion lemma:

$$\hat{\mathbf{x}} = \left(\mathbf{R} - \mathbf{R}\mathbf{B}\left(\boldsymbol{\Sigma} + \mathbf{B}^{\top}\mathbf{R}\mathbf{B}\right)^{-1}\mathbf{B}^{\top}\mathbf{R}\right)\mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

Learning on linear filters

$$\mathbf{x}^{*} = \mathbf{m} = \left(\tilde{\mathbf{R}}^{-1} + \tilde{\mathbf{B}}\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{B}}^{\top}\right)^{-1}\tilde{\mathbf{B}}\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

where

$$\tilde{\mathbf{B}} = \mathbf{U}\mathbf{B}$$

$$\tilde{\mathbf{R}}^{-1} = \mathbf{V}\mathbf{R}^{-1}\mathbf{V}^{\top}$$

Estimate for each voxel:

$$(\hat{\mathbf{b}}, \hat{\sigma}) \quad = \quad \arg\min_{\mathbf{b},\sigma^2} \left\{ \frac{||\mathbf{y} - \mathbf{X}^\top \mathbf{b}||_2^2}{2N\sigma^2} + R_{\lambda,\mathbf{G}}(\mathbf{b}) \right\}$$

**DATA FIT** **CONSTRAINTS**

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \left\{ \frac{1}{2N} ||\mathbf{y} - \mathbf{X}^\top \mathbf{b}||_2^2 + R_{\lambda,\mathbf{G}}(\mathbf{b}) \right\}$$

$\hat{\sigma}^2$ estimated from hold-out data

# Constraints

Graph-constrained elastic net regularizer:

$$R_{\lambda,G}(\mathbf{b}) = \lambda||\mathbf{b}||_1 + \frac{1}{2}\mathbf{b}^\top \mathbf{G}\mathbf{b}$$

The pixel-by-pixel matrix **G** expresses constraints on **b**:

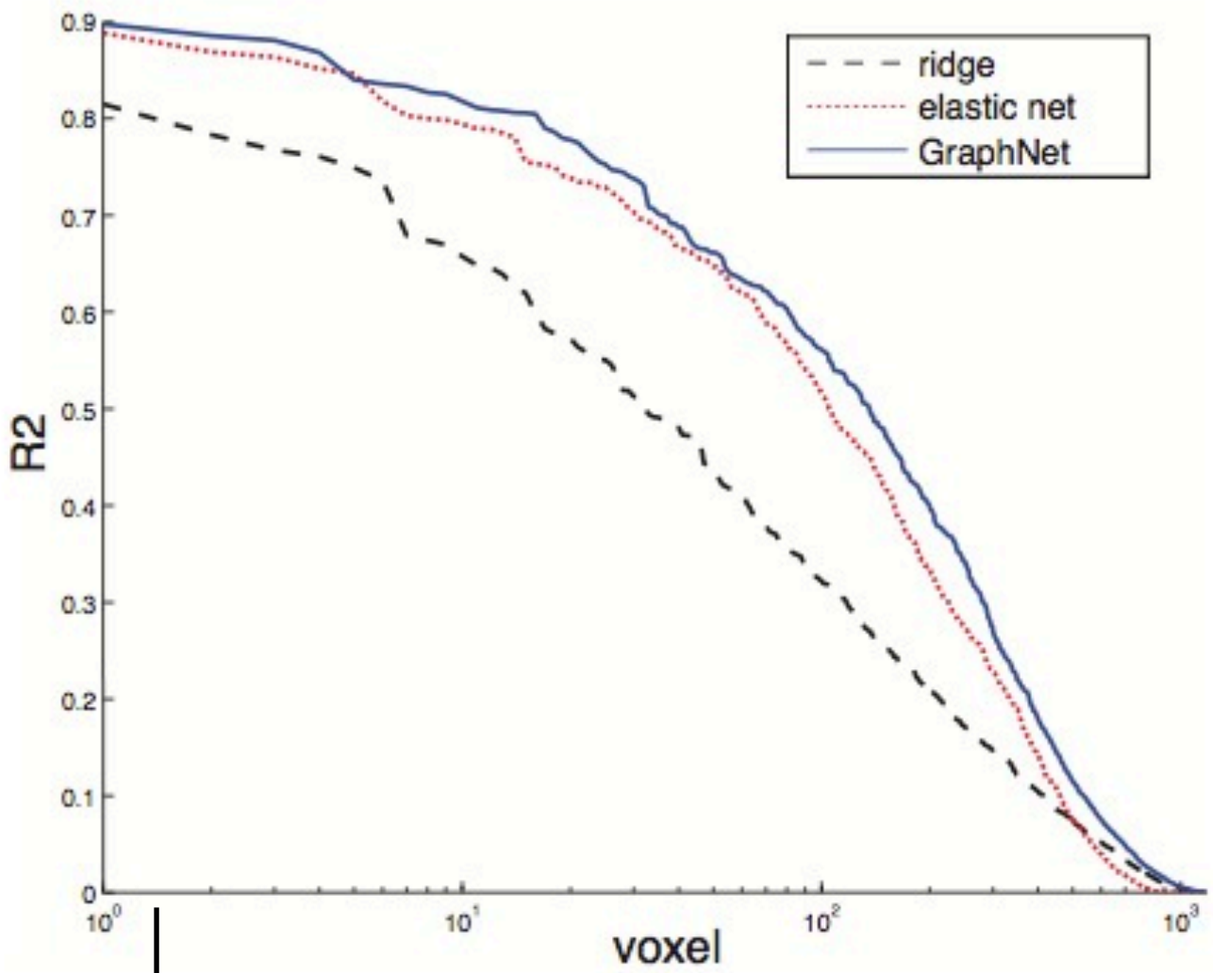
The choice **G** = **D** - **A** with degree matrix **D** and adjacency matrix **A** encourages neighbouring pixels to have the same weight (see Lecture notes)

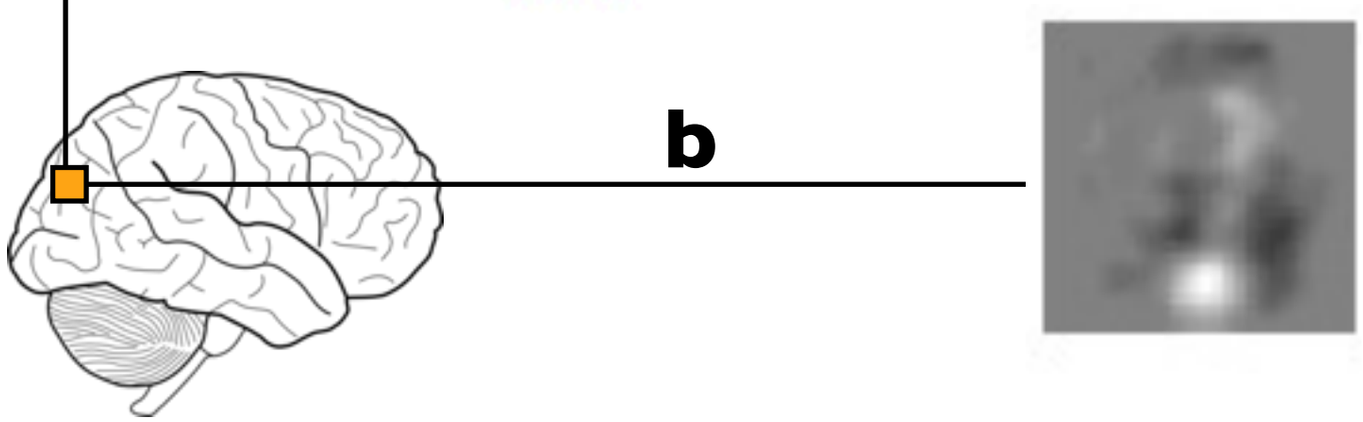
For a 2 x 2 image:

$$\boldsymbol{D} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix} \quad \boldsymbol{A} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \quad \boldsymbol{G} = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}$$

# Encoding performance

Explained variance per voxel for the three encoding models

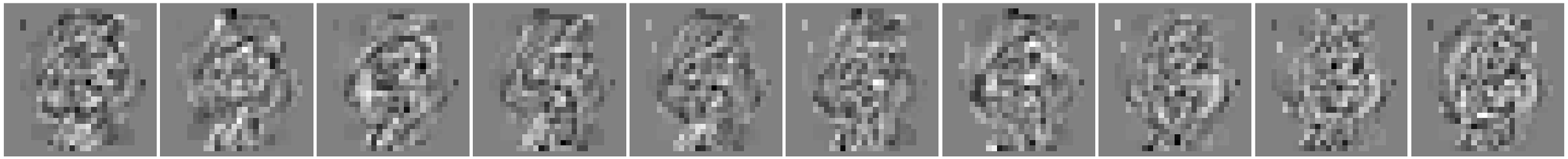


# Learned filters

Learned filters for the ten best voxels:

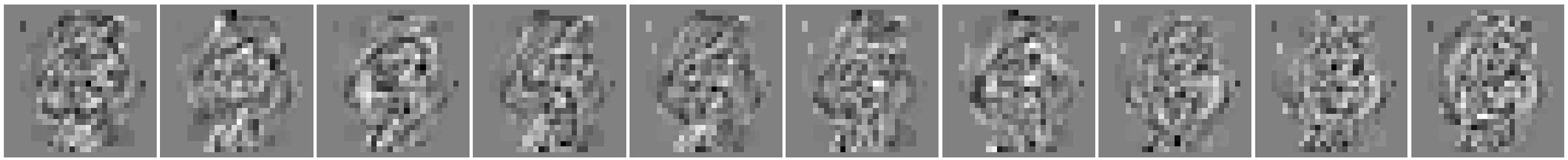
ridge



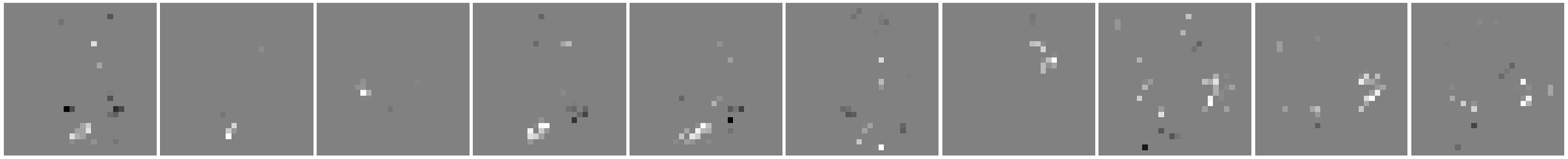
# Learned filters

Learned filters for the ten best voxels:

ridge



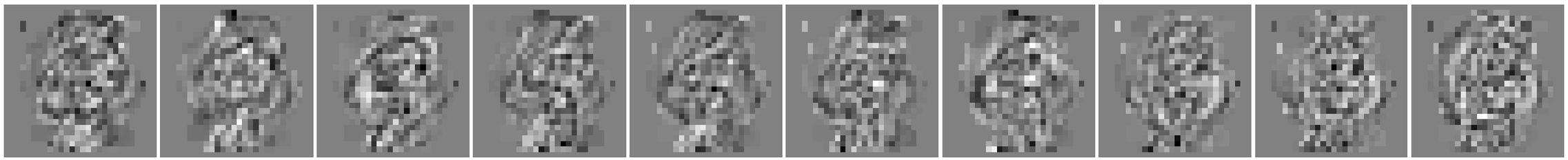
elastic net



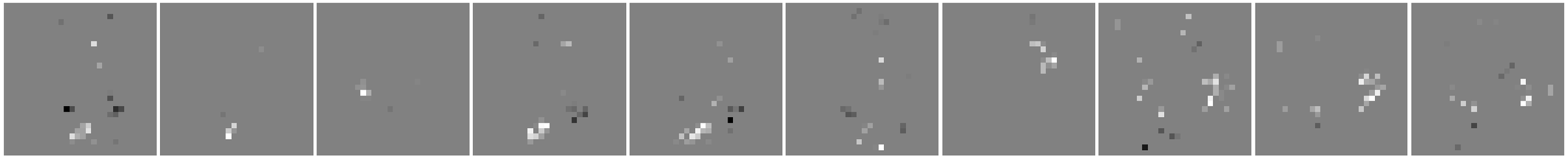
# Learned filters

Learned filters for the ten best voxels:

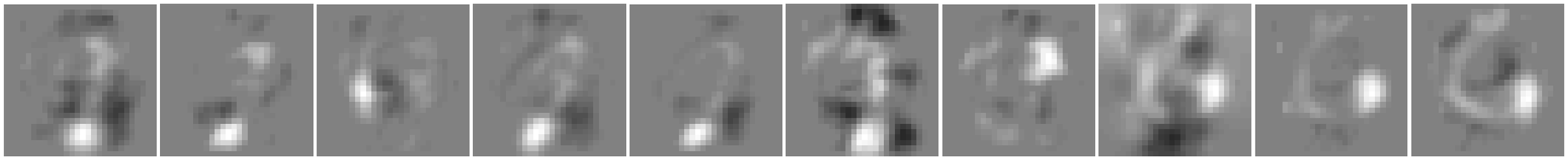
ridge



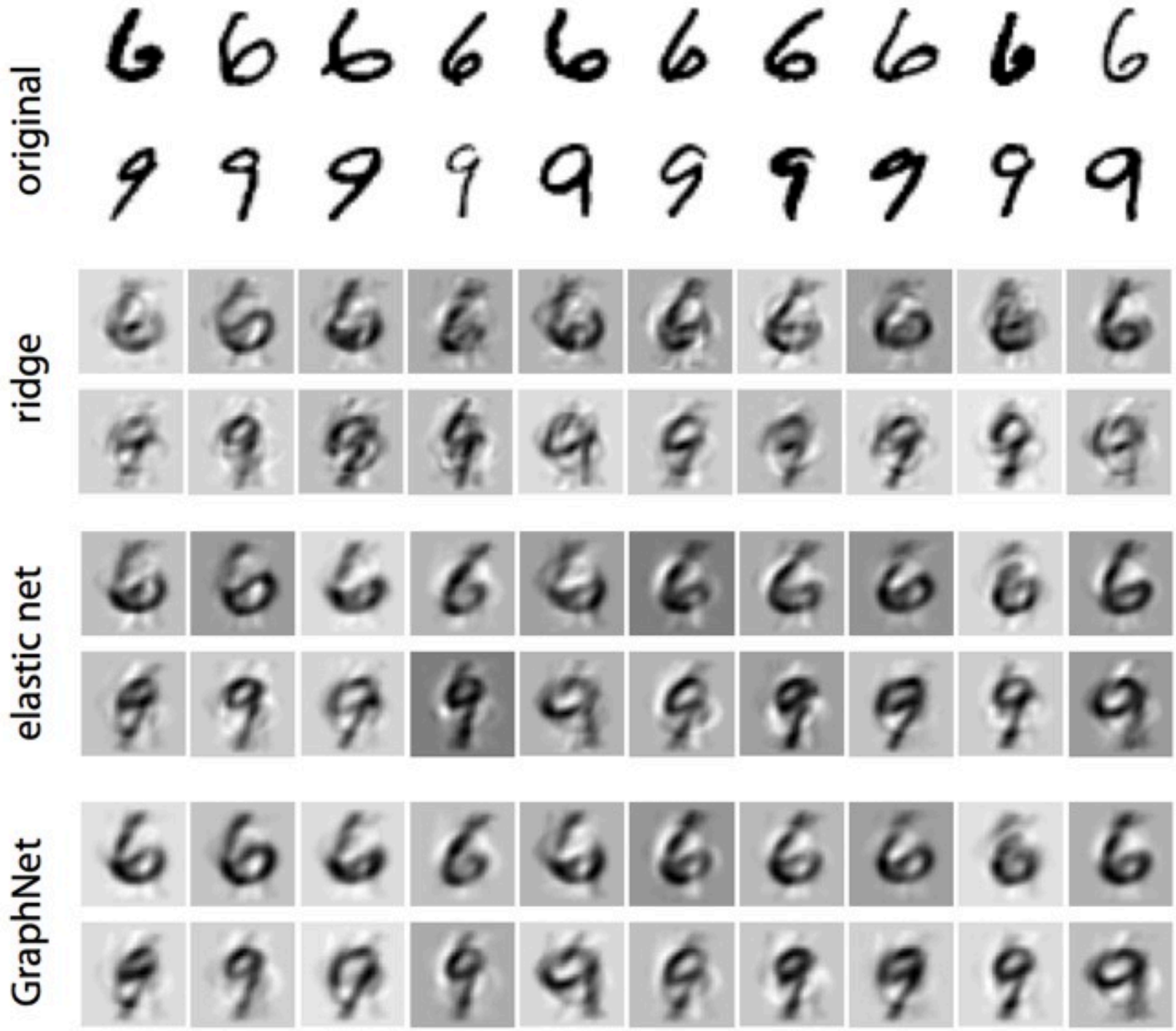
elastic net



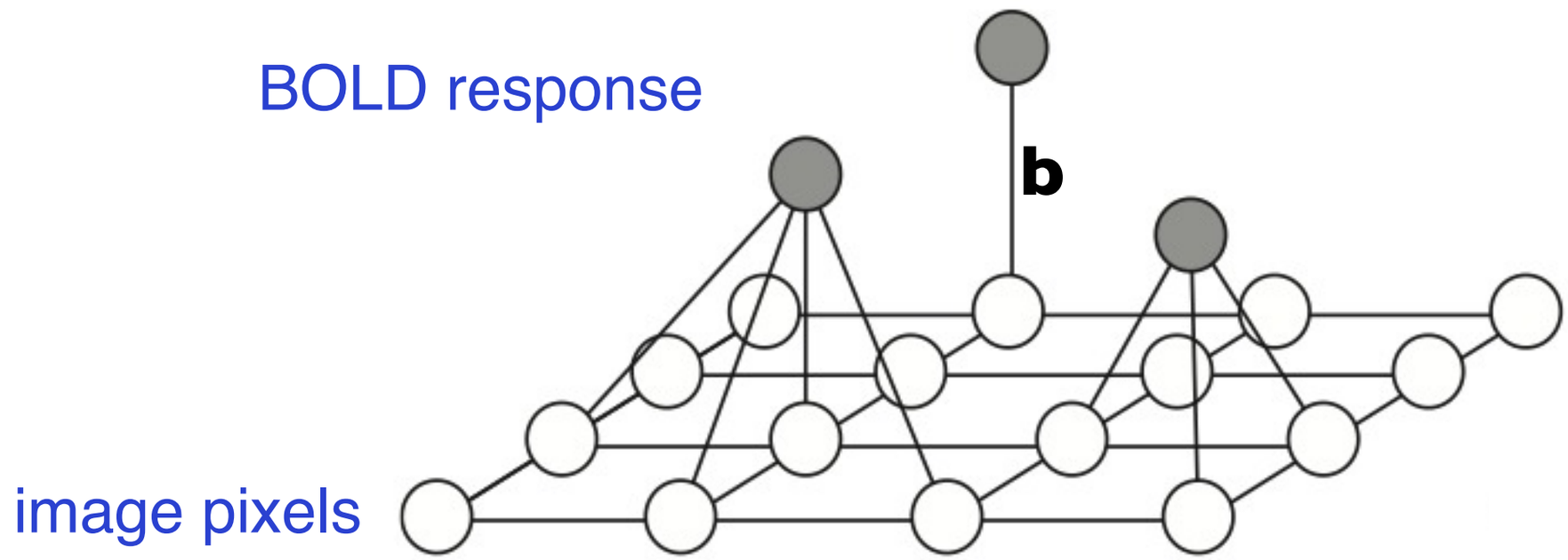
GraphNet



# Reconstructions



# A bit of interpretation

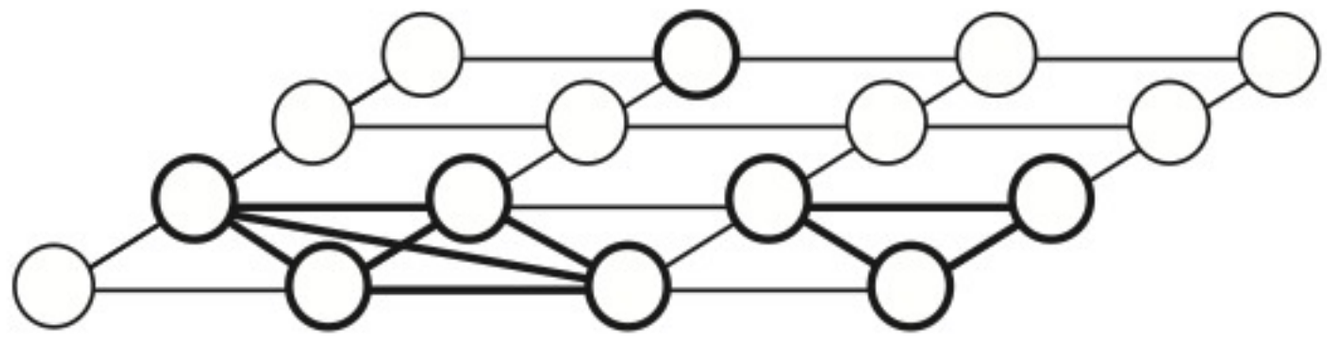



Probabilistic form of the Graphnet regularizer:

$$p(\mathbf{b}) \quad \propto \quad \prod_i \exp\left(-\lambda |b_i|\right) \prod_j \exp\left(-\frac{1}{2}\sum_{i\sim j} b_i G_{ij} b_j\right)$$

# A bit of interpretation

image pixels

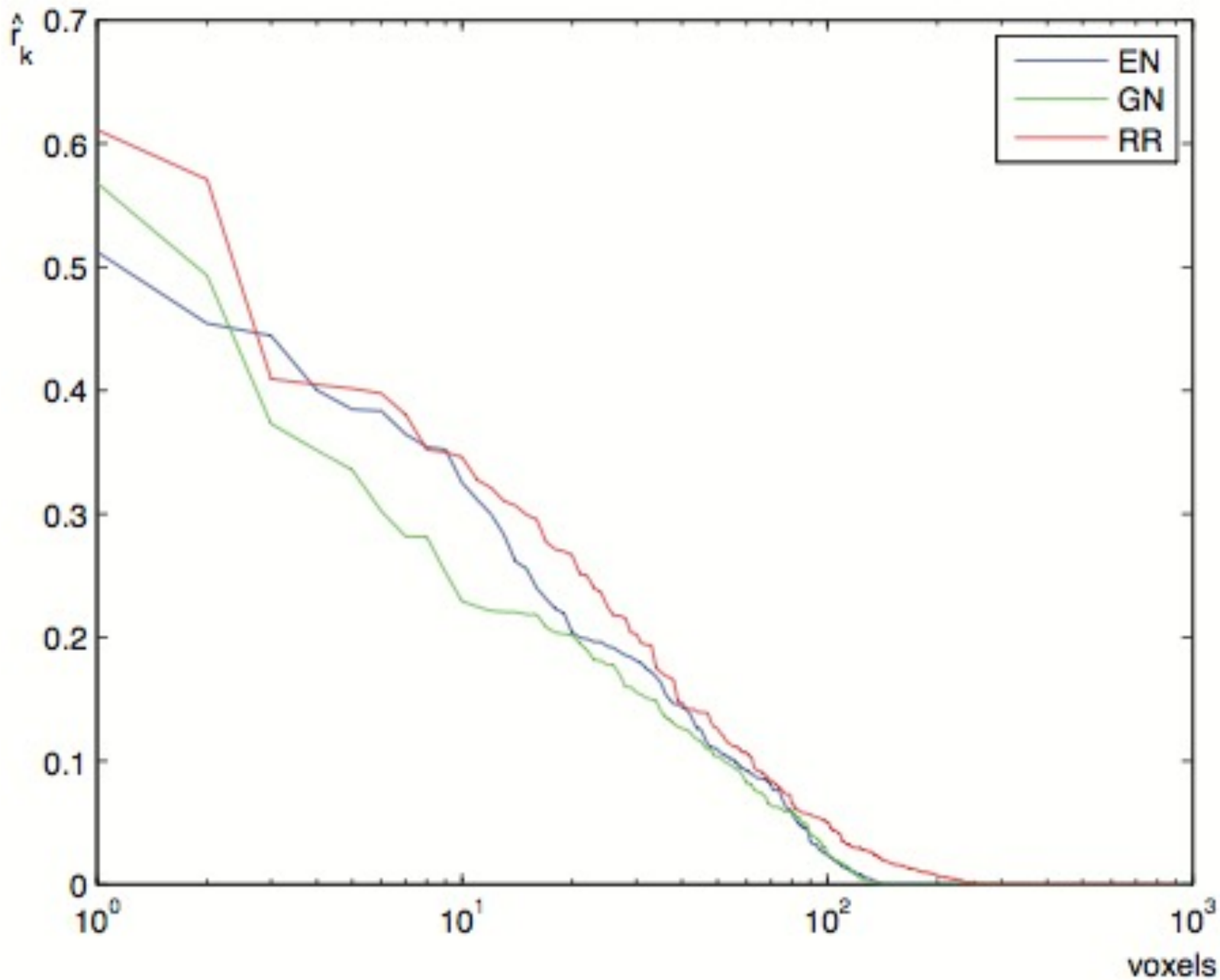


- correlations between pixels change as a function of **b**
- mean of the updated Gaussian is the most likely input

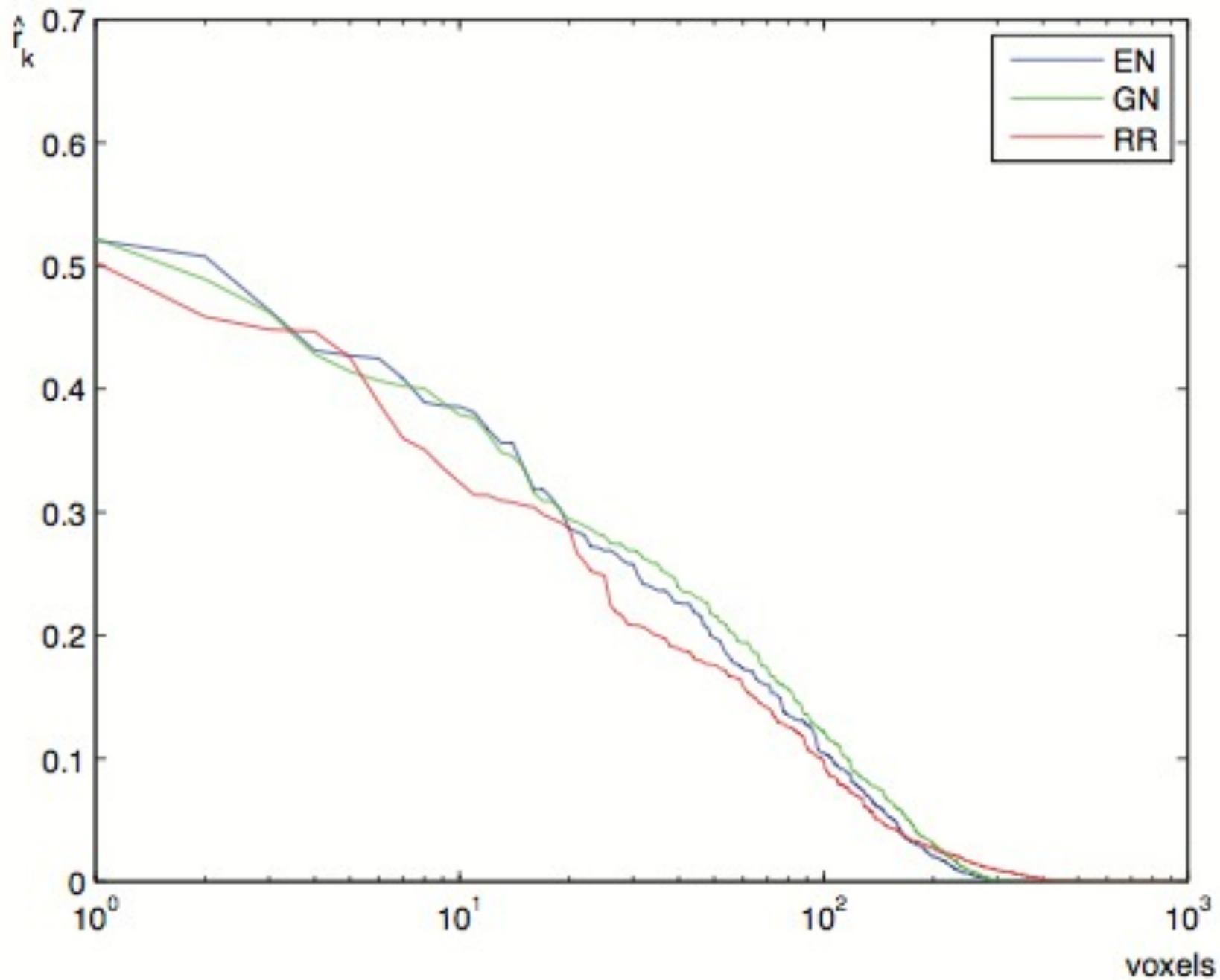
# Some new results

-Test approach with large stimulus set
and rapid event related design:

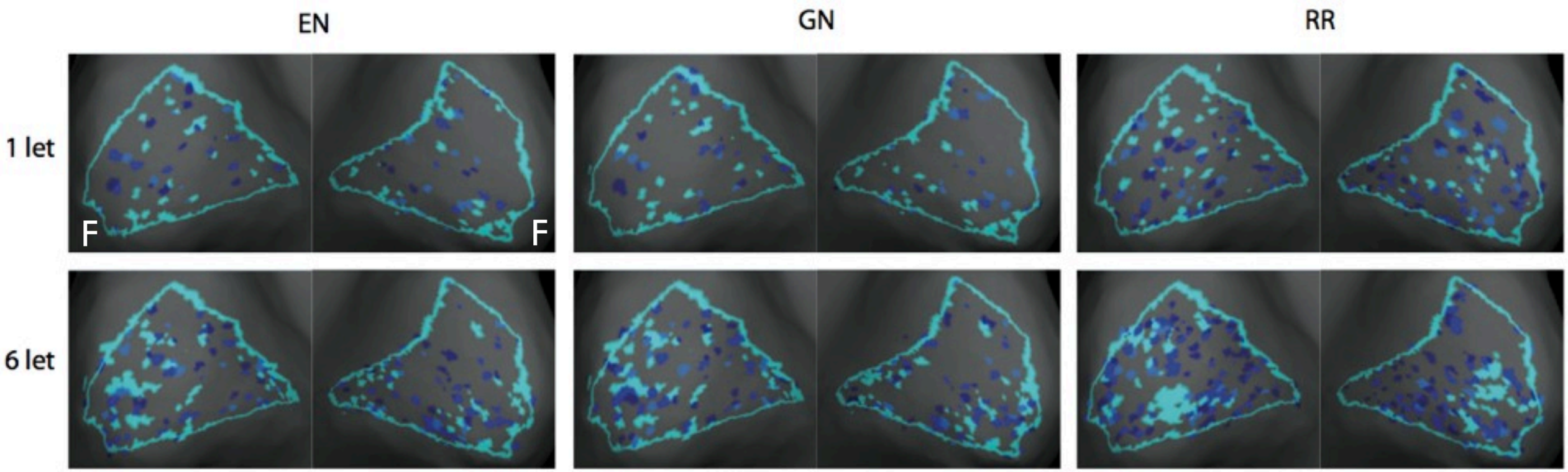
# Encoding performance; trained on ‘B’ only



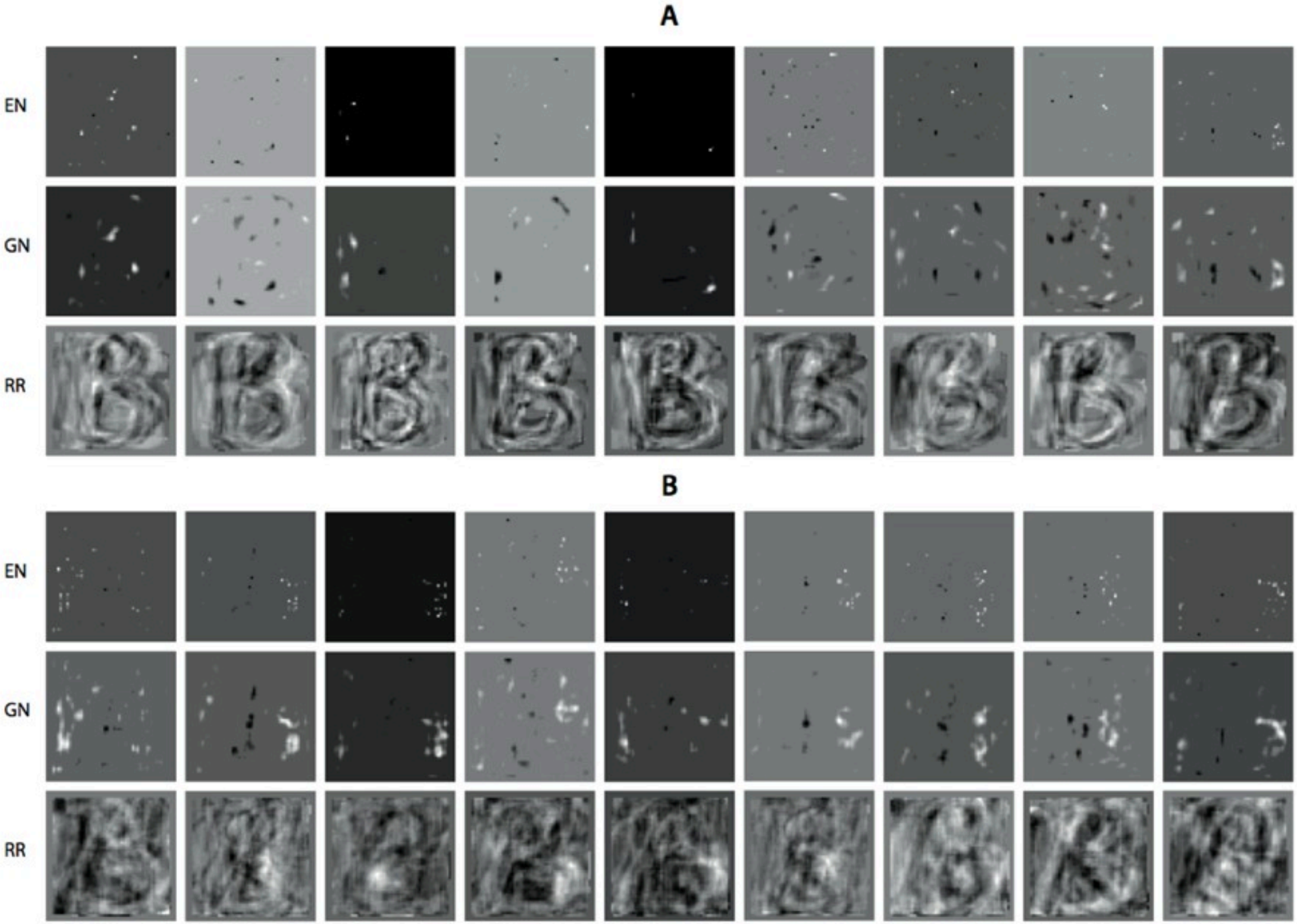
# Encoding performance; trained on all letters



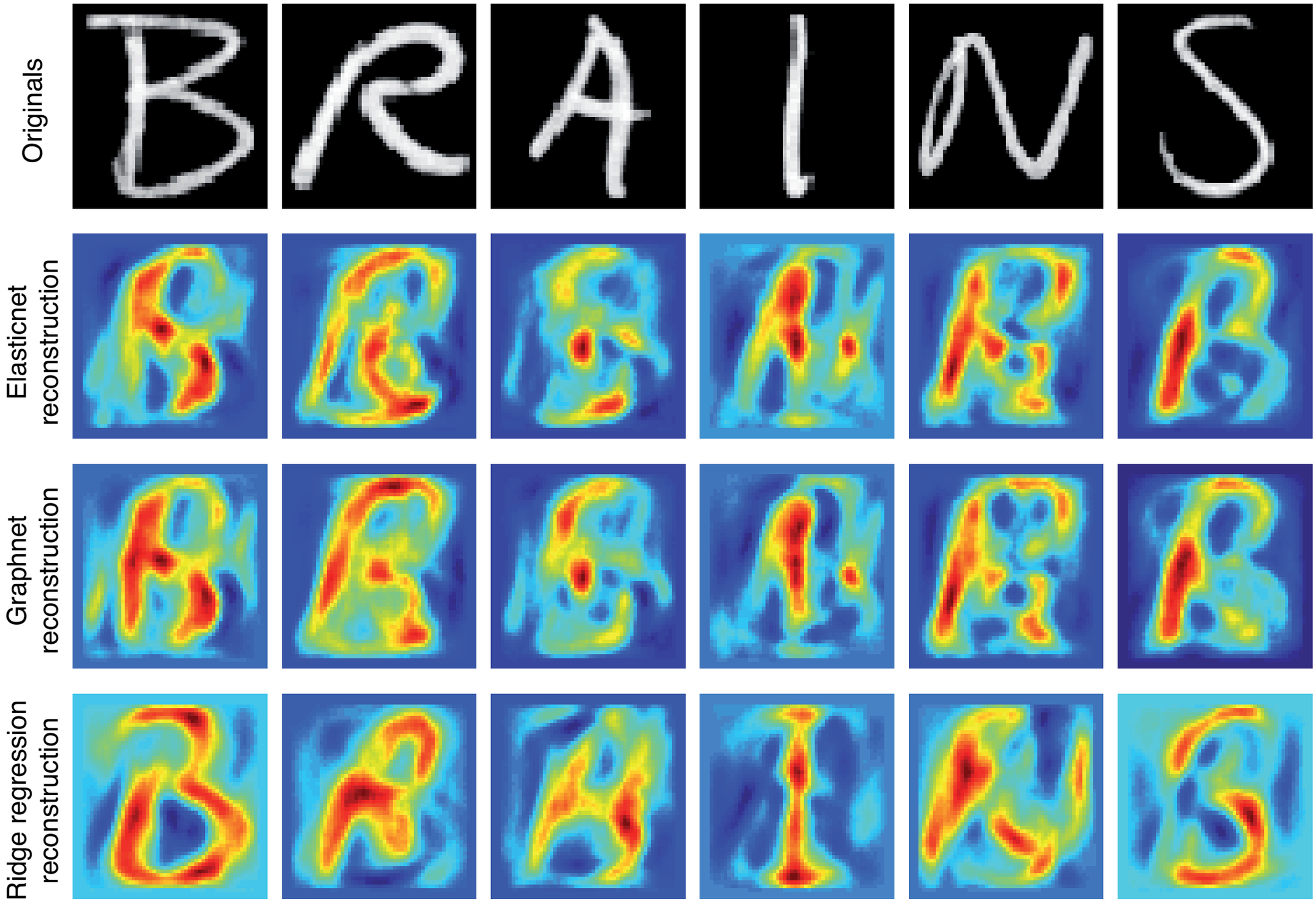
# Localization in V1



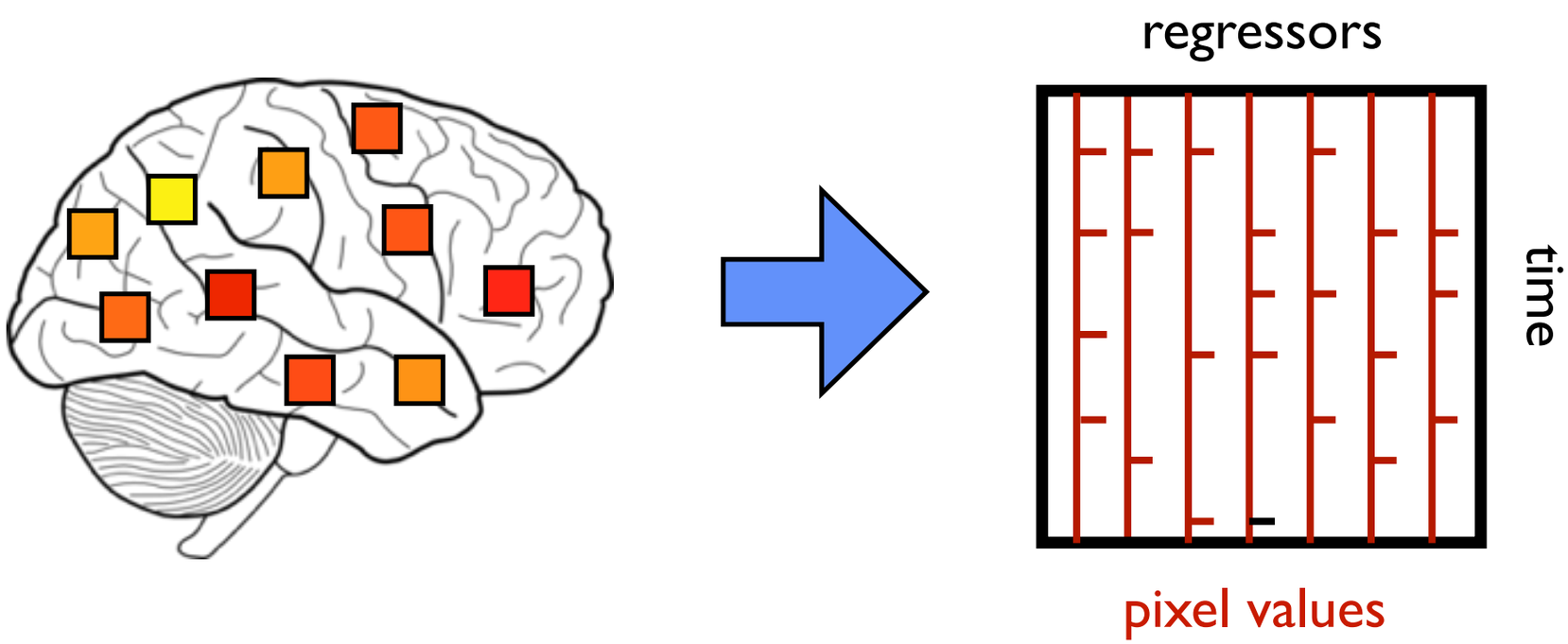
# Learned filters

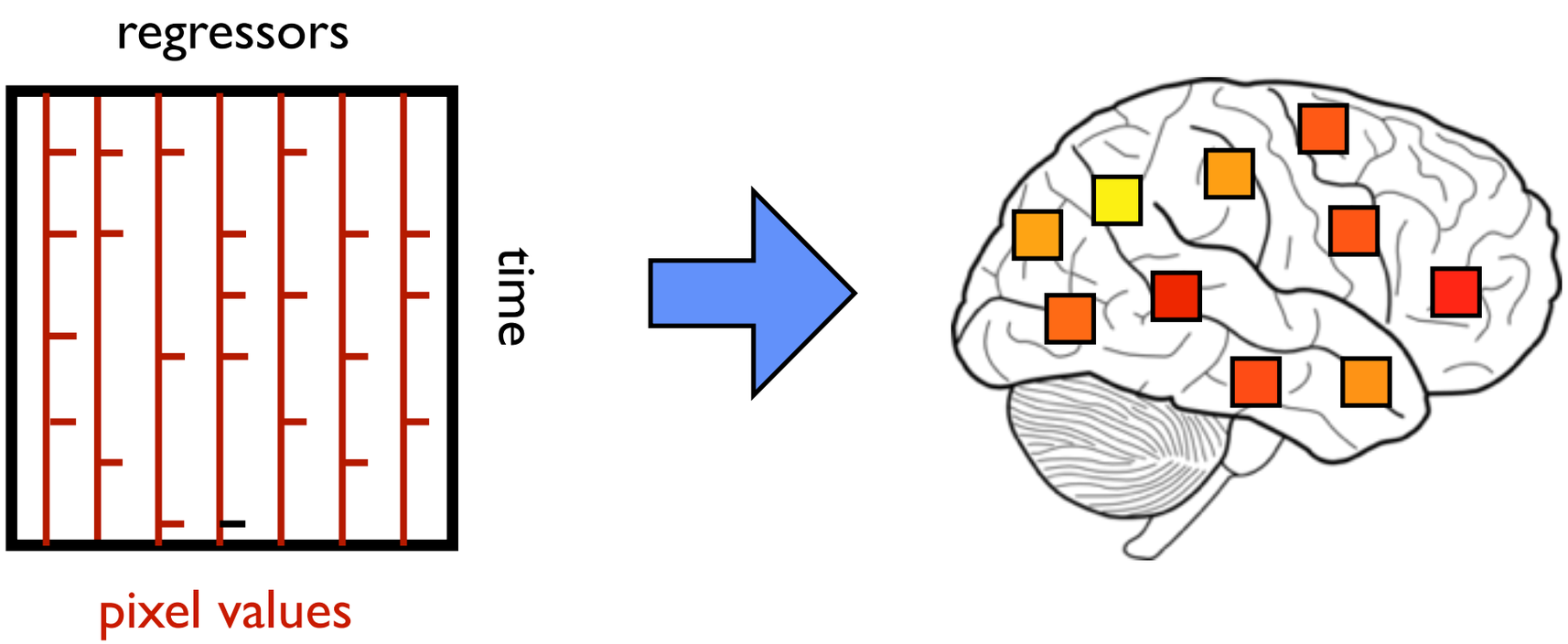


# Reconstructions



# Discussion: discriminative versus generative models



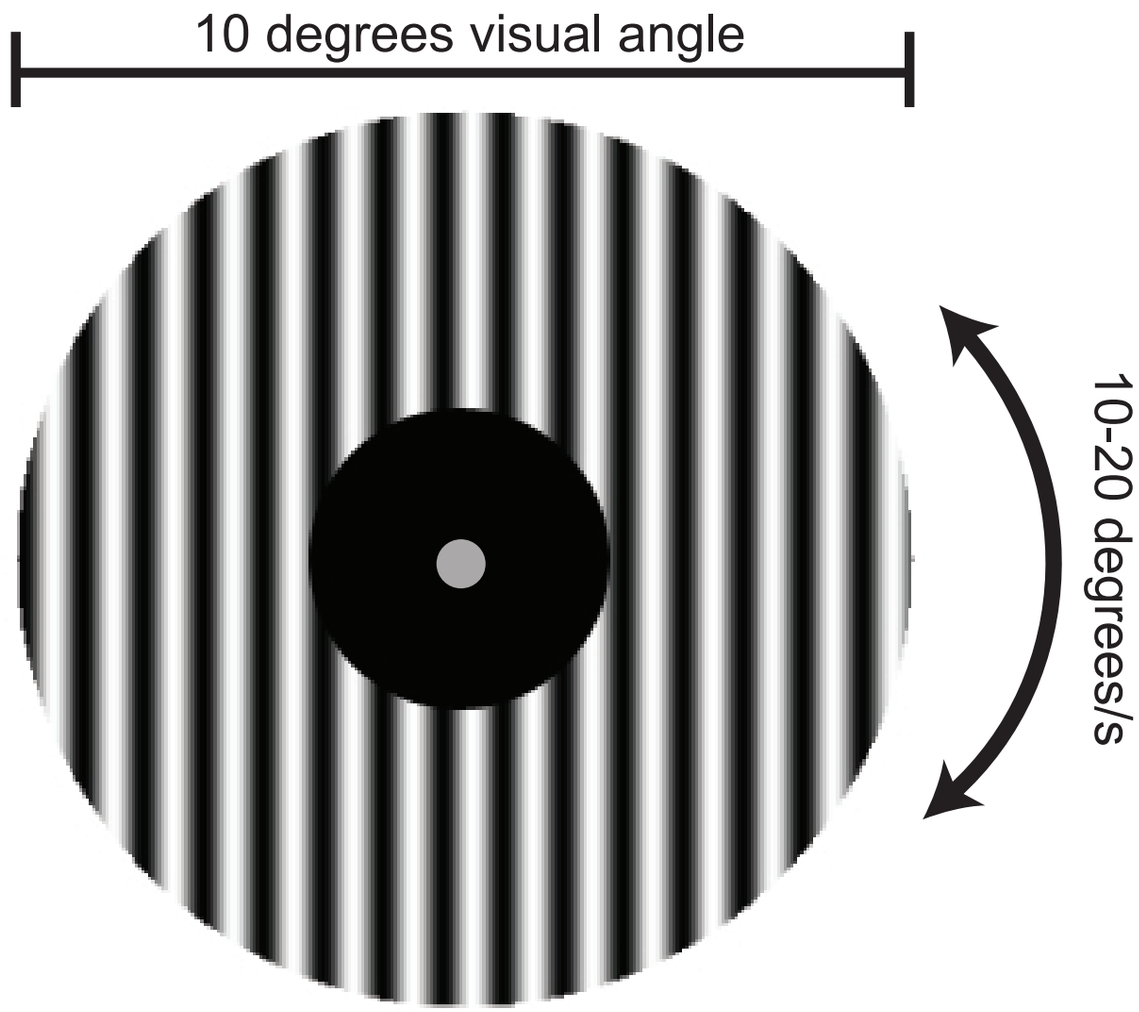
# Towards dynamic decoding

- most decoding studies are trial-based
- mental state is an ongoing process
- can we decode the dynamics of mental state?

Applications:

- ongoing perception
- reactivations in memory retrieval
- bistable illusions
- asynchronous BCIs

# Dynamic decoding of grating orientation

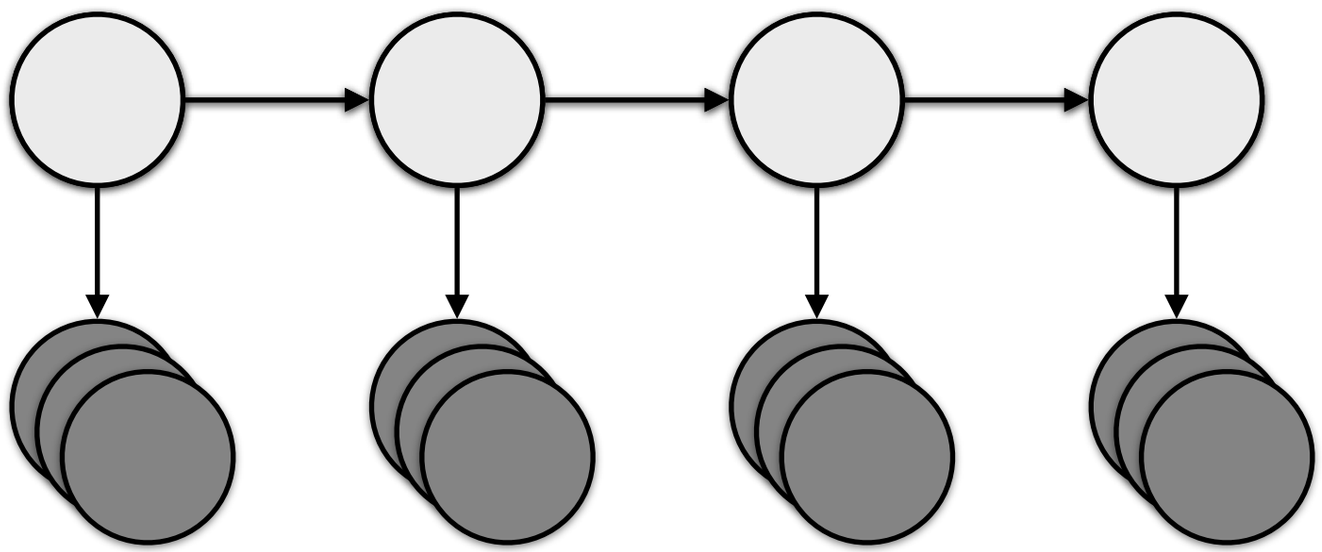



Continuously rotating grating
random changes in angular velocity and rotation direction

van Gerven MAJ, Kok P, de Lange FP, Heskes T. Dynamic decoding of ongoing perception. Neuroimage. 2011;57:950–957.

# sine/cosine angle, angular velocity, rotation direction

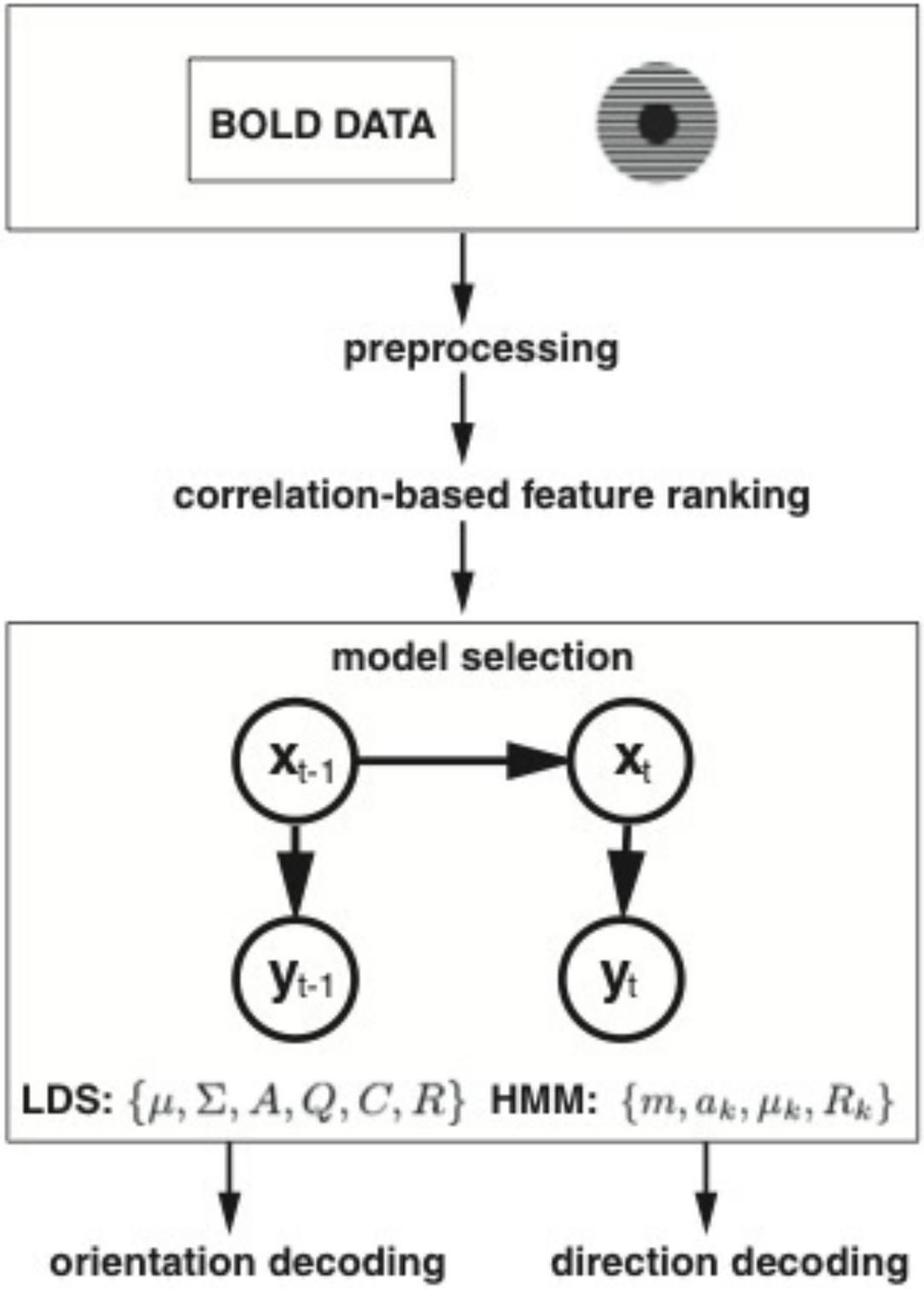
BOLD



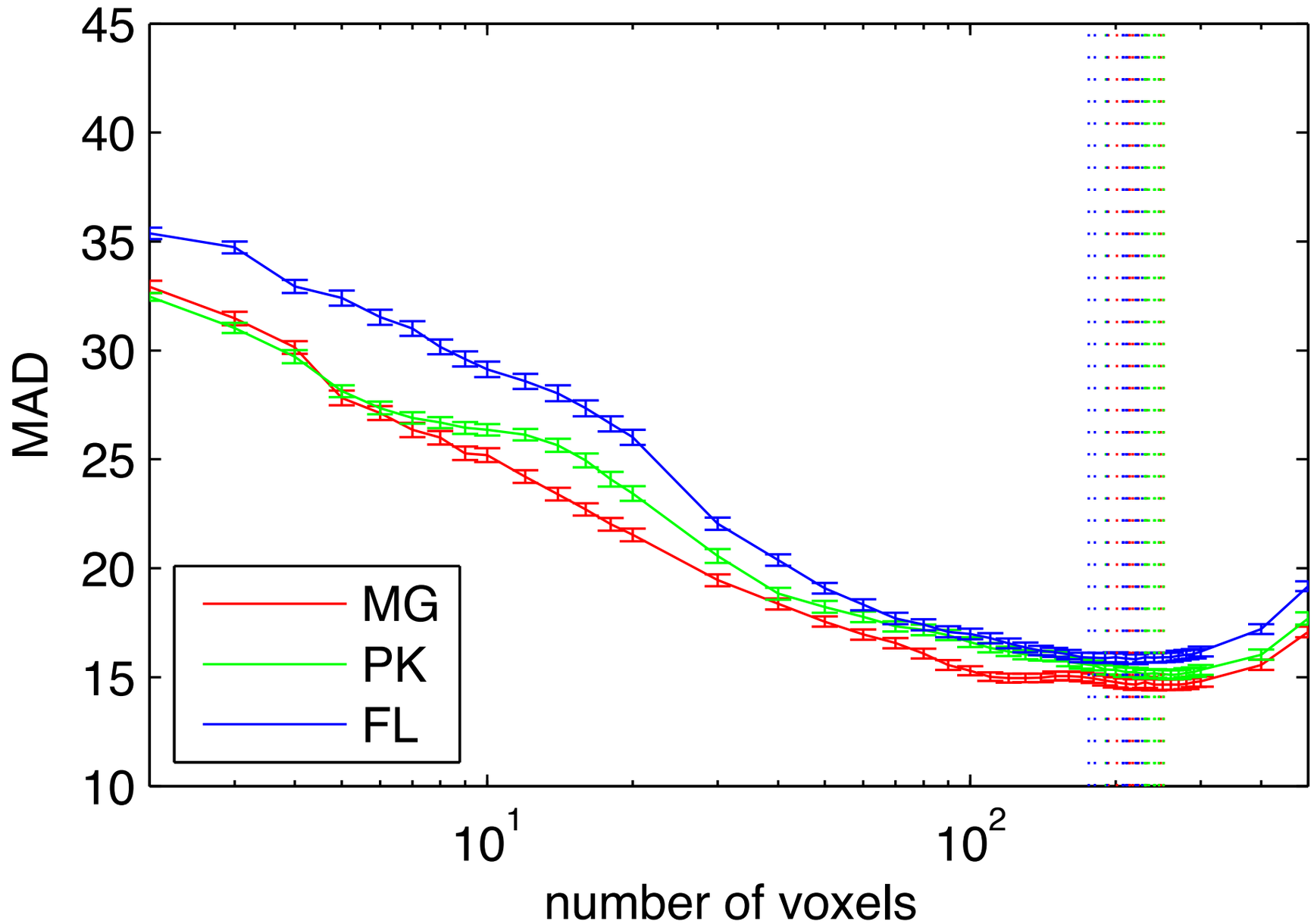
$$p(x,y) = p(x_1) \prod_{t=2}^{T} p(x_t|x_{t-1}) \prod_{t=1}^{T} p(y_t|x_t)$$

Decoding with the Kalman smoother (continuous) or Viterbi decoding (discrete)

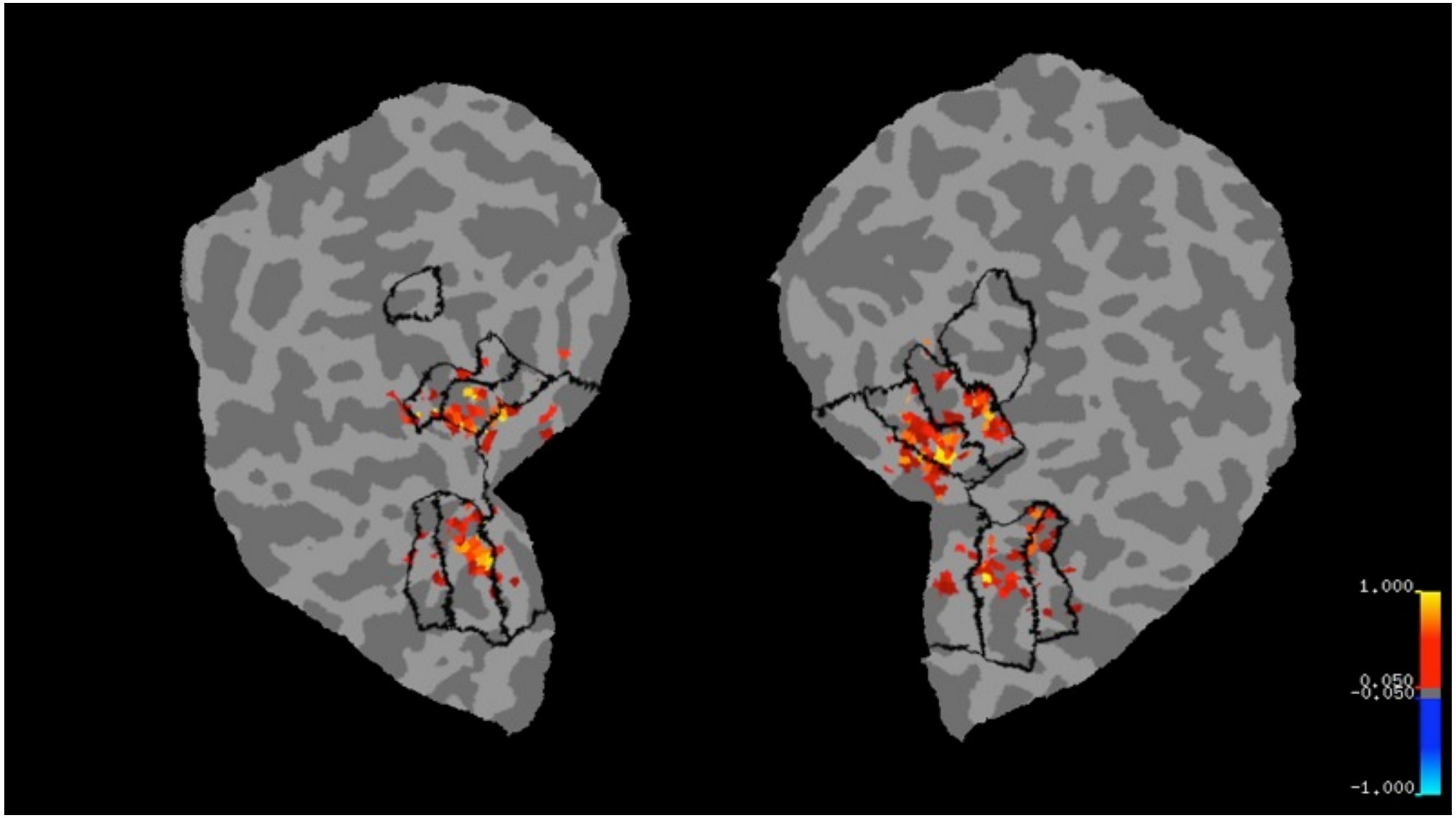
# Decoding procedure



# Angle decoding

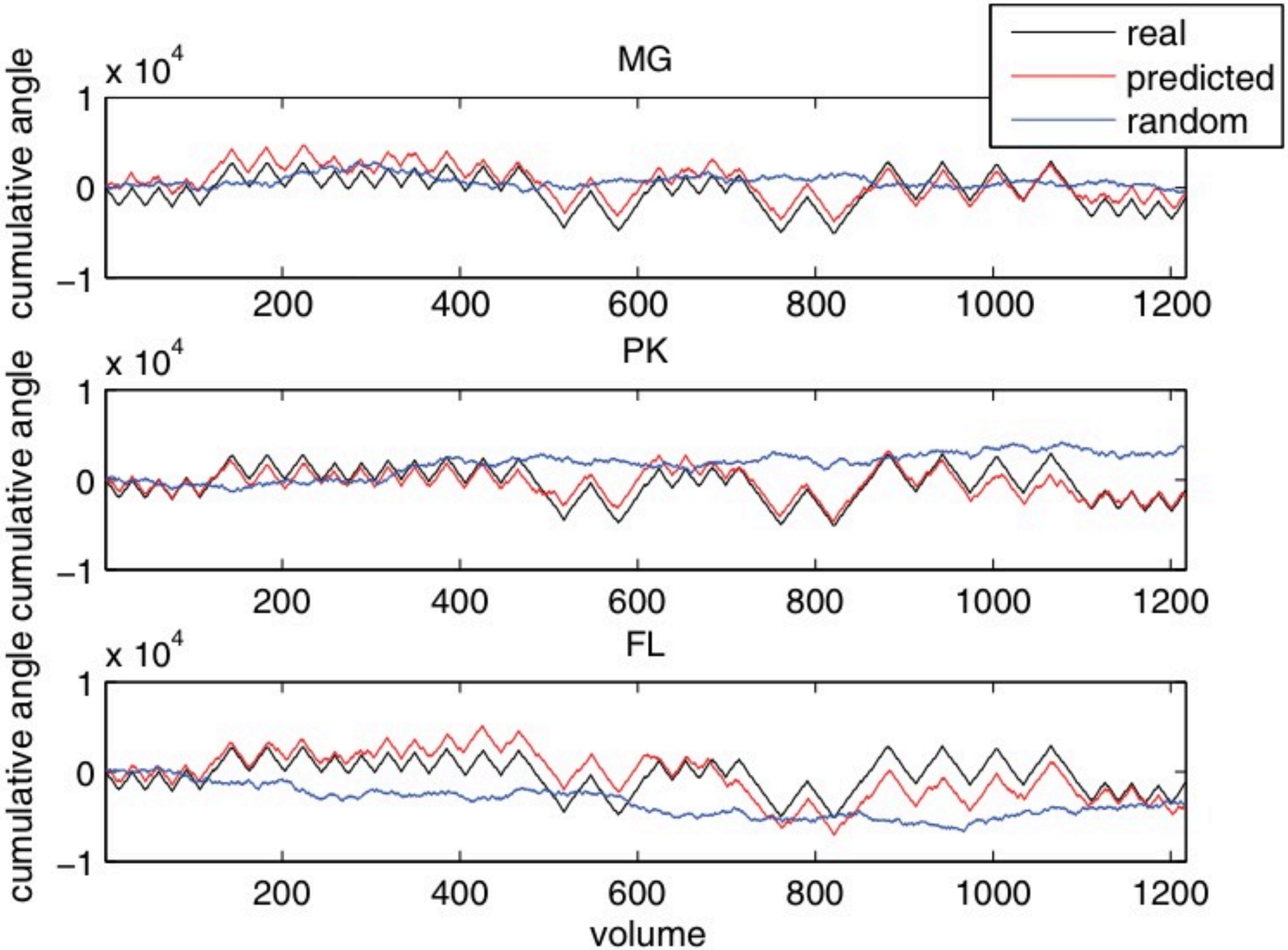


# Angle decoding

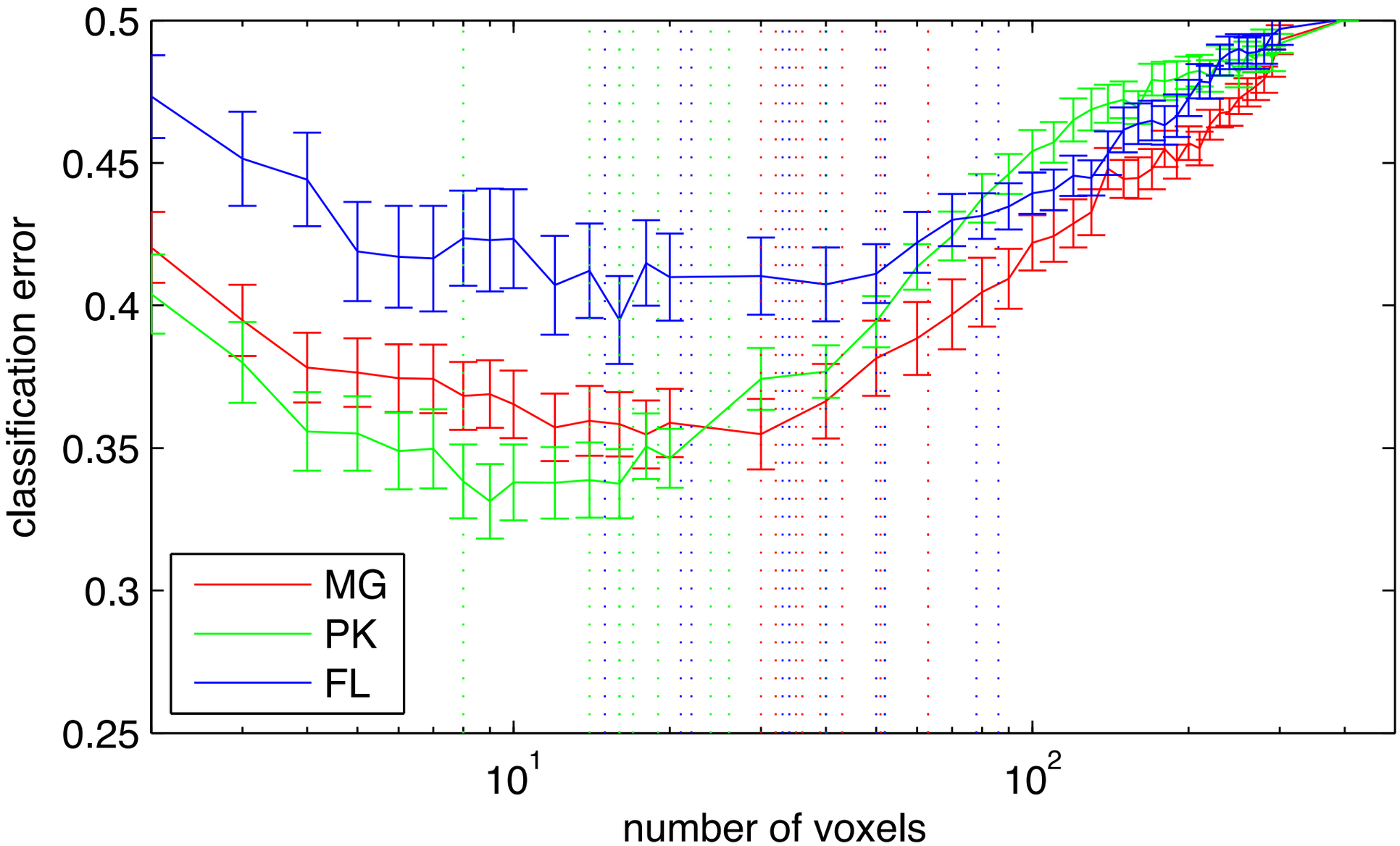



selection of voxels that correlate most strongly with the timeseries

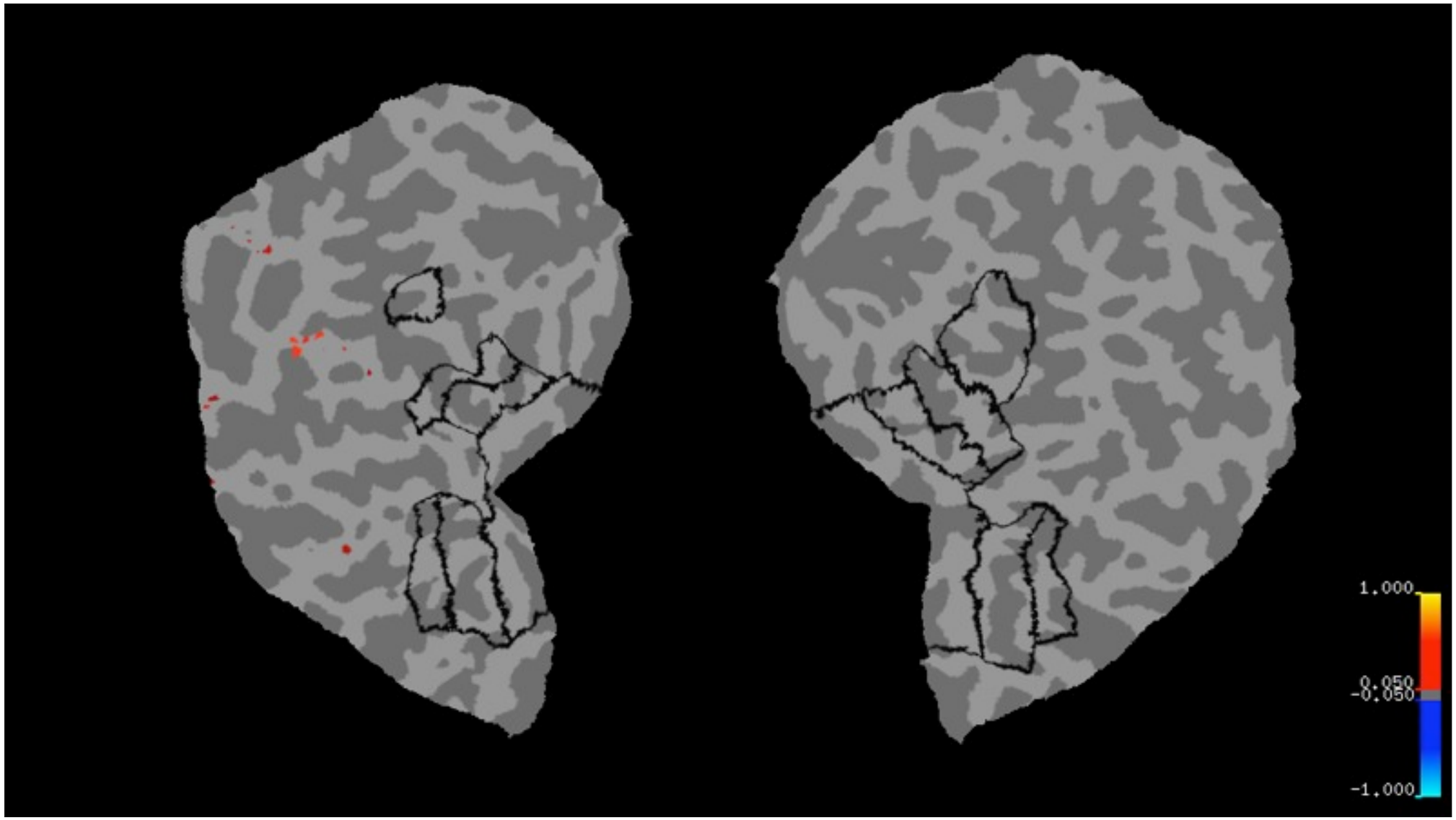
# Angle decoding



# Direction decoding

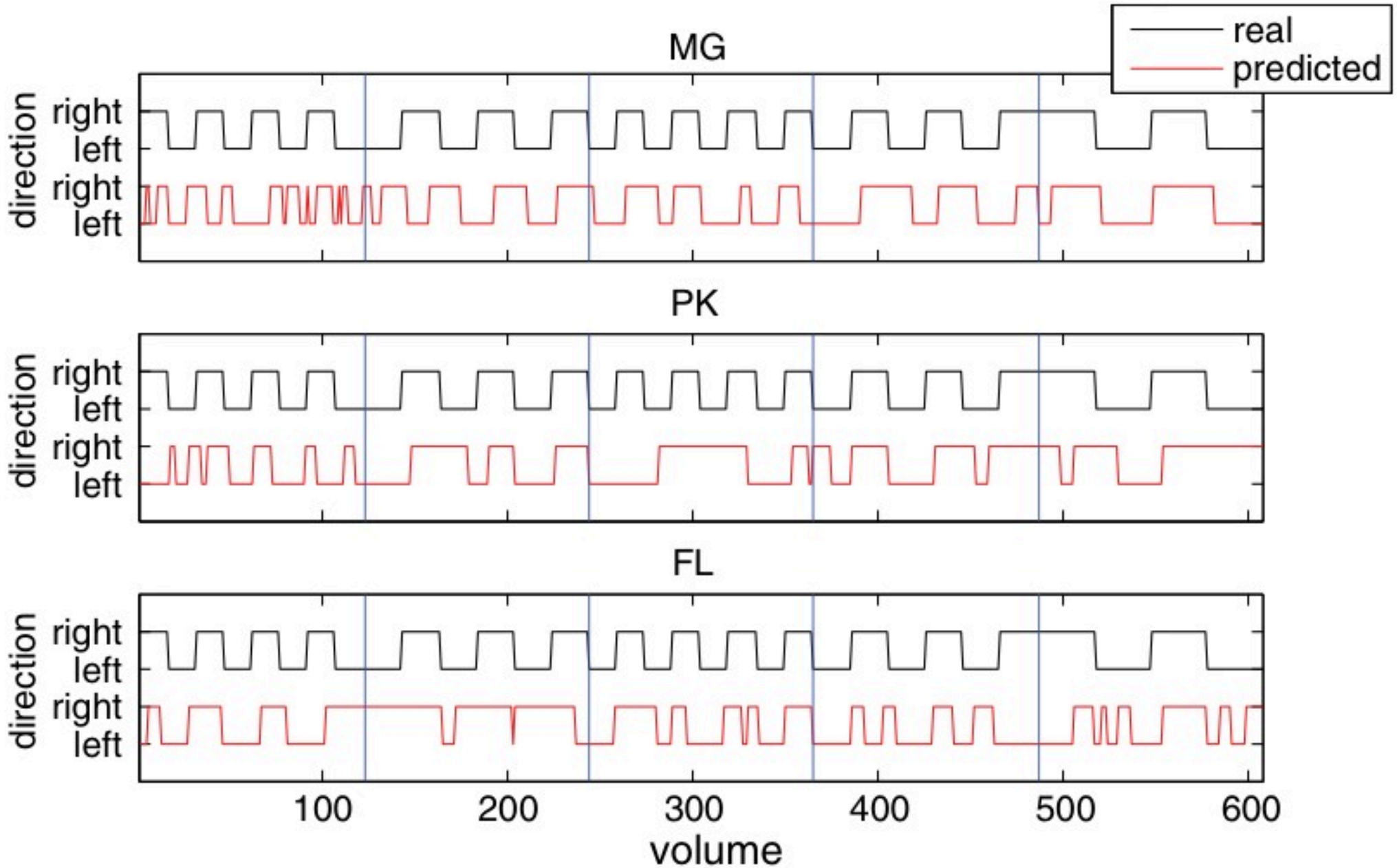


# Direction decoding

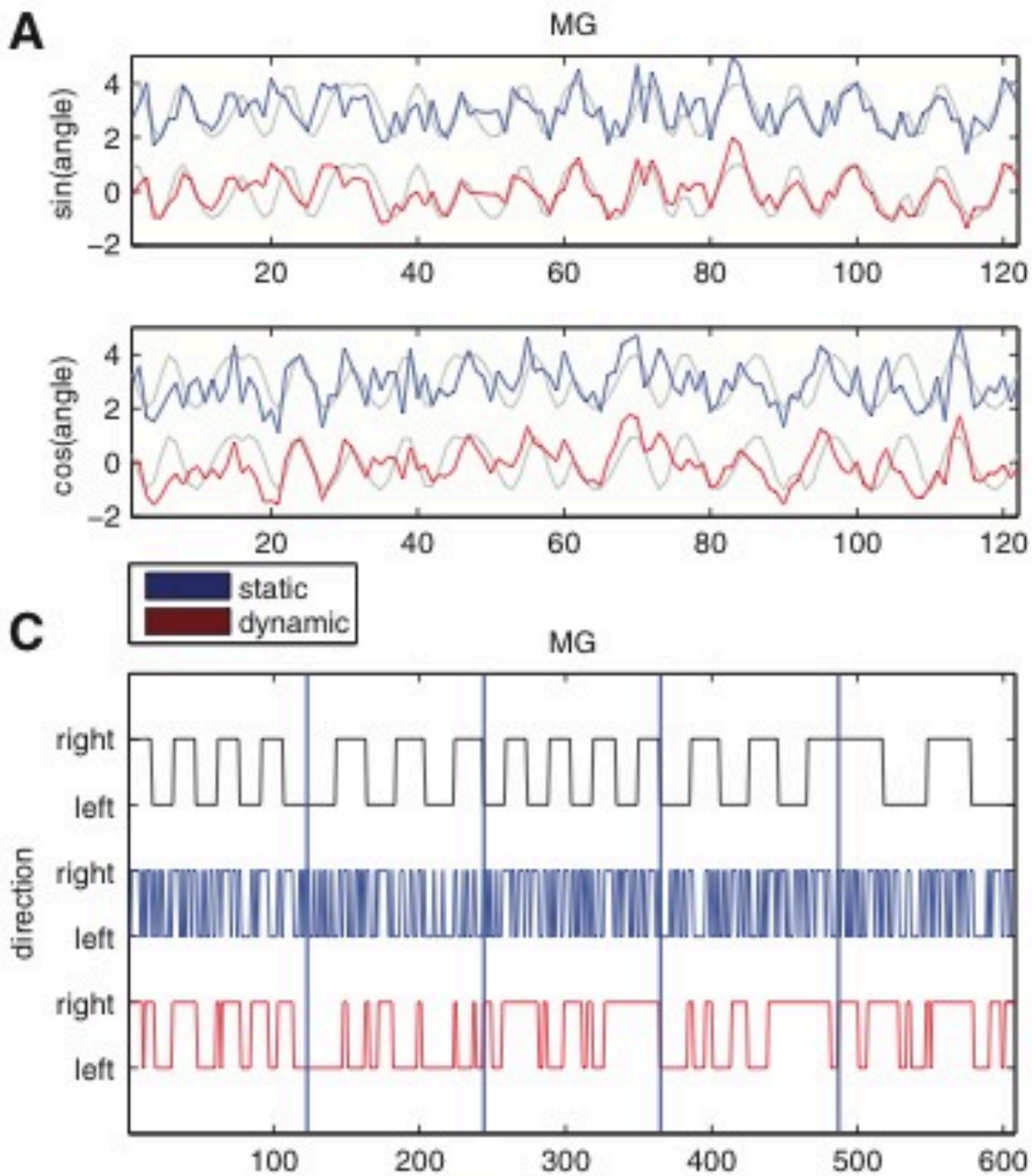


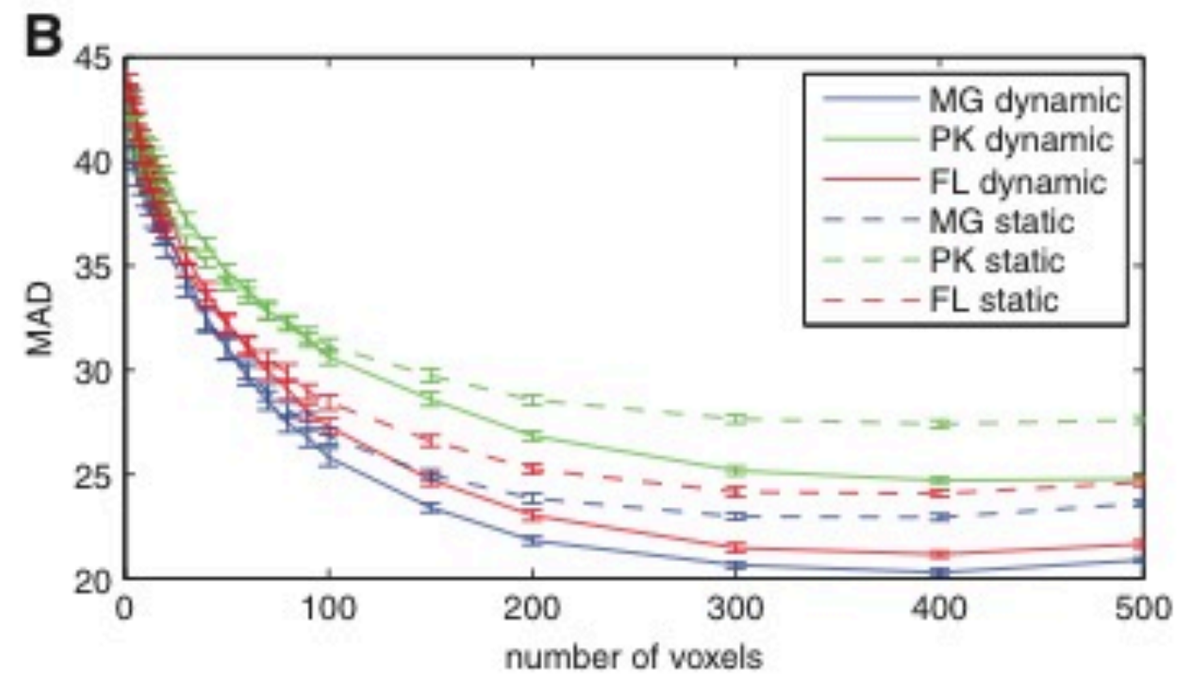

selection of voxels that correlate most strongly with the timeseries

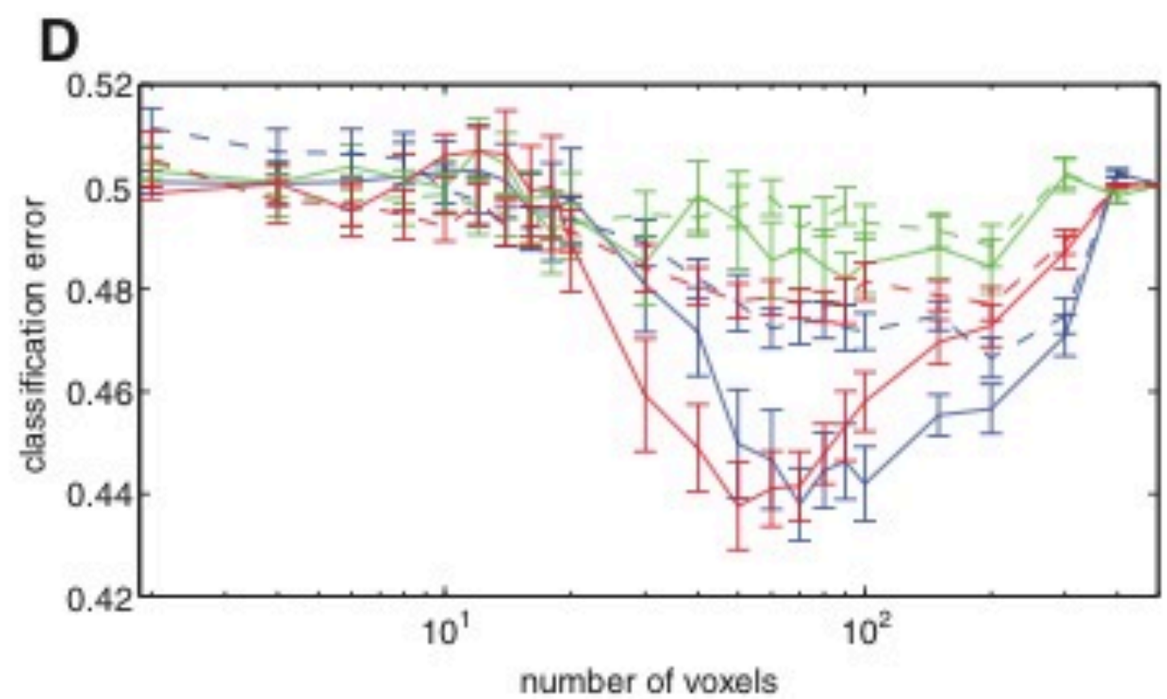
# Direction decoding



# Static versus dynamic decoding



# Required reading

- Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, Mason RA, et al. Predicting Human Brain Activity Associated with the Meanings of Nouns. Science. 2008;320(5880):1191–5.
- van Gerven MAJ, Heskes T. A Linear Gaussian Framework for Decoding of Perceived Images. 2011 International Workshop on Pattern Recognition in Neuroimaging (PRNI). 2012. p. 1–4.

## Predicting Human Brain Activity Associated with the Meanings of Nouns

Tom M. Mitchell,[1*] Svetlana V. Shinkareva,[2] Andrew Carlson,[1] Kai-Min Chang,[3,4]
Vicente L. Malave,[5] Robert A. Mason,[3] Marcel Adam Just[3]

The question of how the human brain represents conceptual knowledge has been debated in many scientific fields. Brain imaging studies have shown that different spatial patterns of neural activation are associated with thinking about different semantic categories of pictures and words (for example, tools, buildings, and animals). We present a computational model that predicts the functional magnetic resonance imaging (fMRI) neural activation associated with words for which fMRI data are not yet available. This model is trained with a combination of data from a trillion-word text corpus and observed fMRI data associated with viewing several dozen concrete nouns. Once trained, the model predicts fMRI activation for thousands of other concrete nouns in the text corpus, with highly significant accuracies over the 60 nouns for which we currently have fMRI data.

2012 Second International Workshop on Pattern Recognition in NeuroImaging

## A Linear Gaussian Framework for Decoding of Perceived Images

Marcel A. J. van Gerven
*Radboud University Nijmegen*
*Donders Institute for Brain, Cognition and Behaviour*
*Nijmegen, The Netherlands*
*Email: m.vangerven@donders.ru.nl*

Tom Heskes
*Radboud University Nijmegen*
*Institute for Computing and Information Sciences*
*Nijmegen, The Netherlands*
*Email: tomh@cs.ru.nl*

# Learning goals

- Discriminative vs generative decoding
- Identification
- Concept decoding
- Gaussian generative decoding
- Idea of dynamic decoding

# Coming week

**Practical**

Continue tutorial

**Lecture**

Advanced generative decoding