Donders Institute
for Brain, Cognition and Behaviour

Decoding conceptual representations

Marcel van Gerven

Computational Cognitive Neuroscience Lab
(www.ccnlab.net)

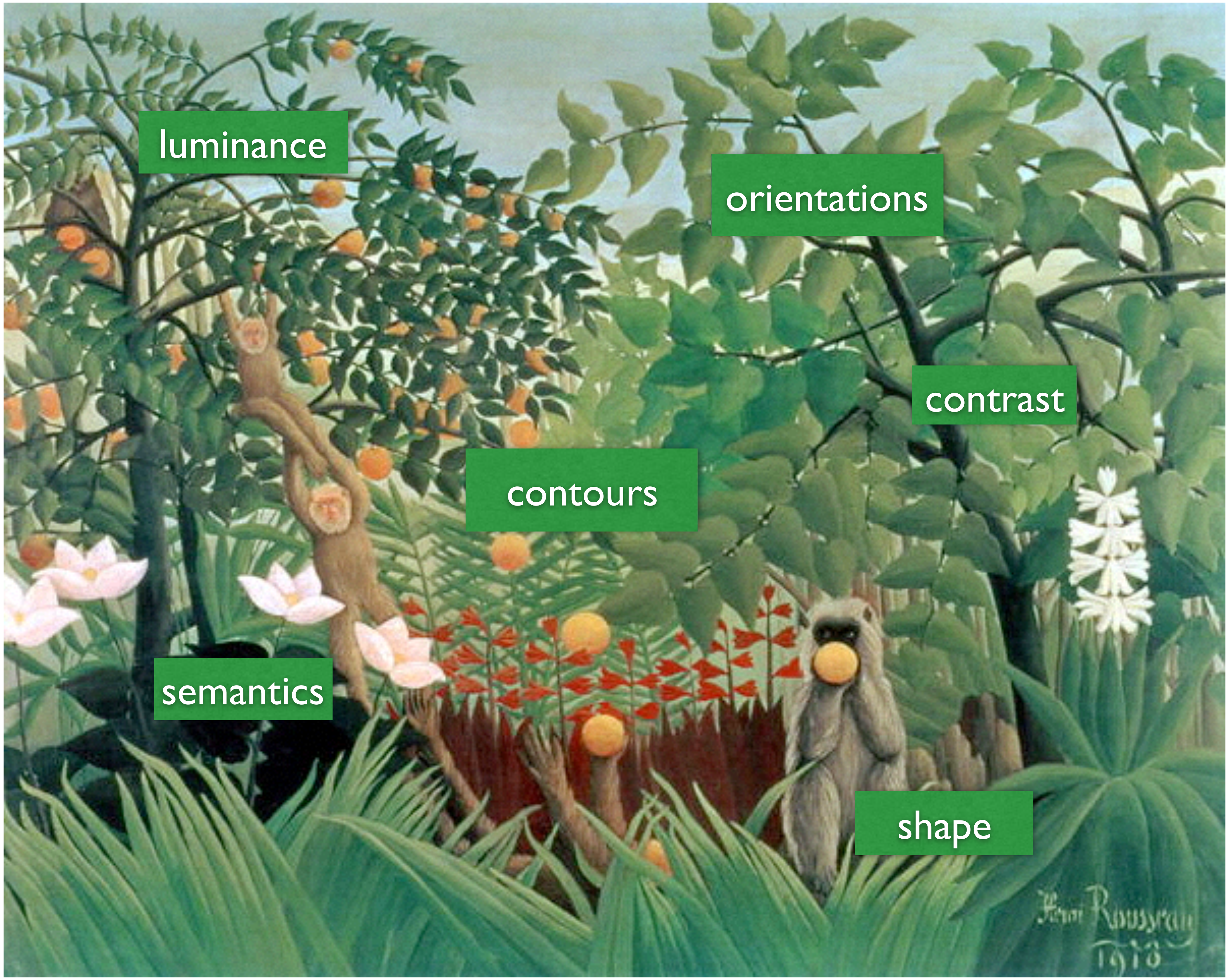
Artificial Intelligence Department

Donders Centre for Cognition

Donders Institute for Brain, Cognition and Behaviour

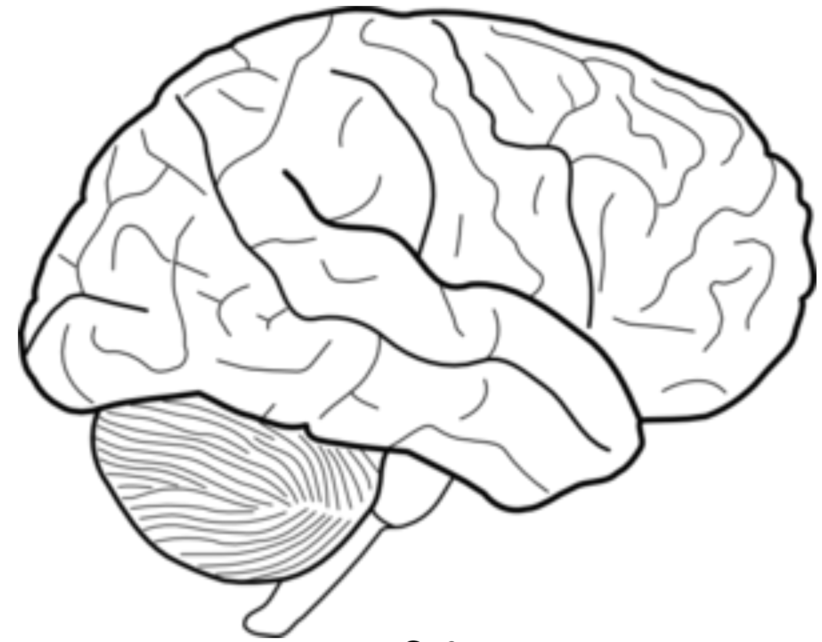
Radboud University Nijmegen

Conceptual representations

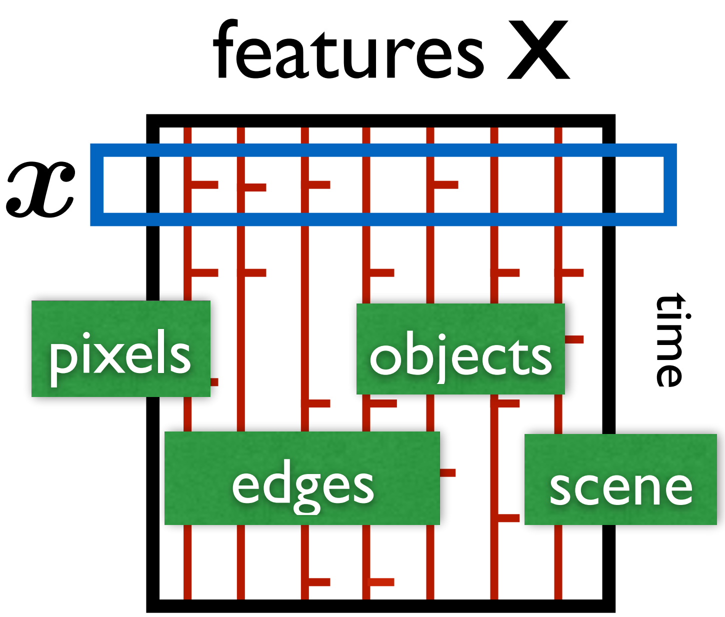


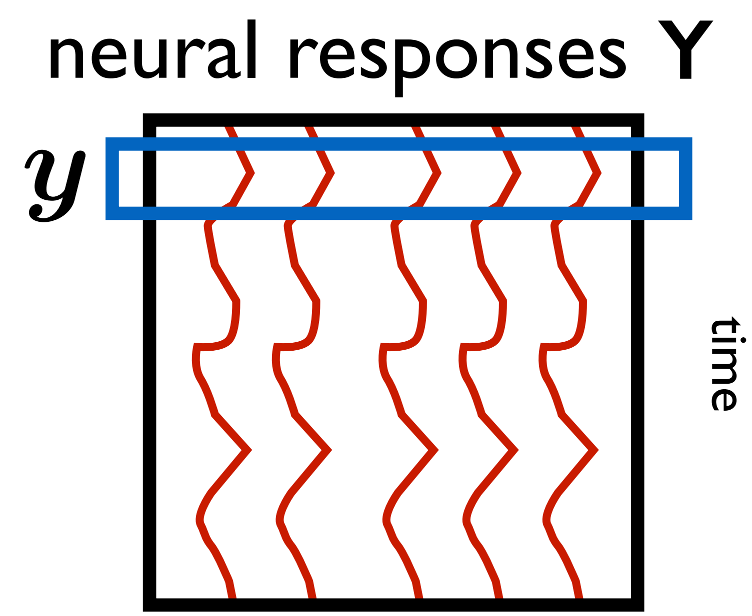
Stimulus-response relationships

x

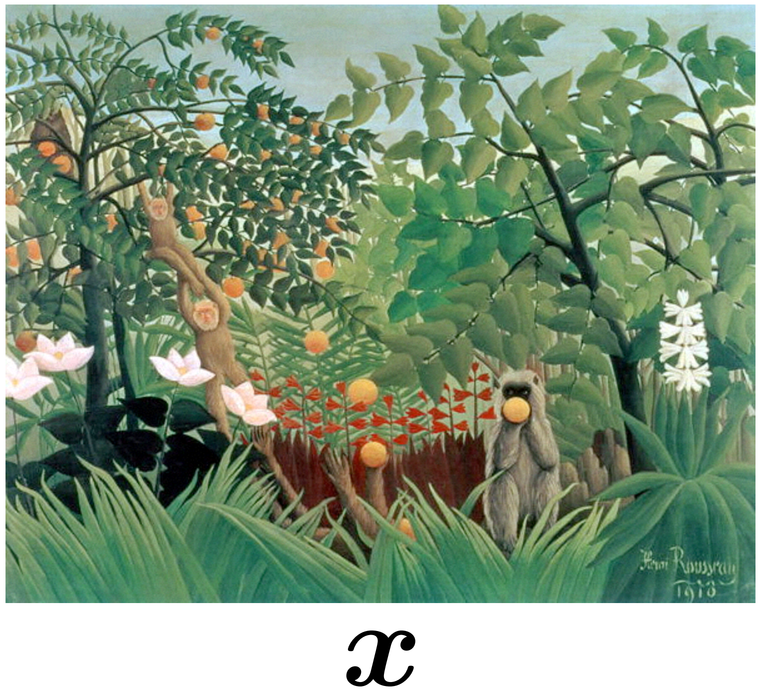


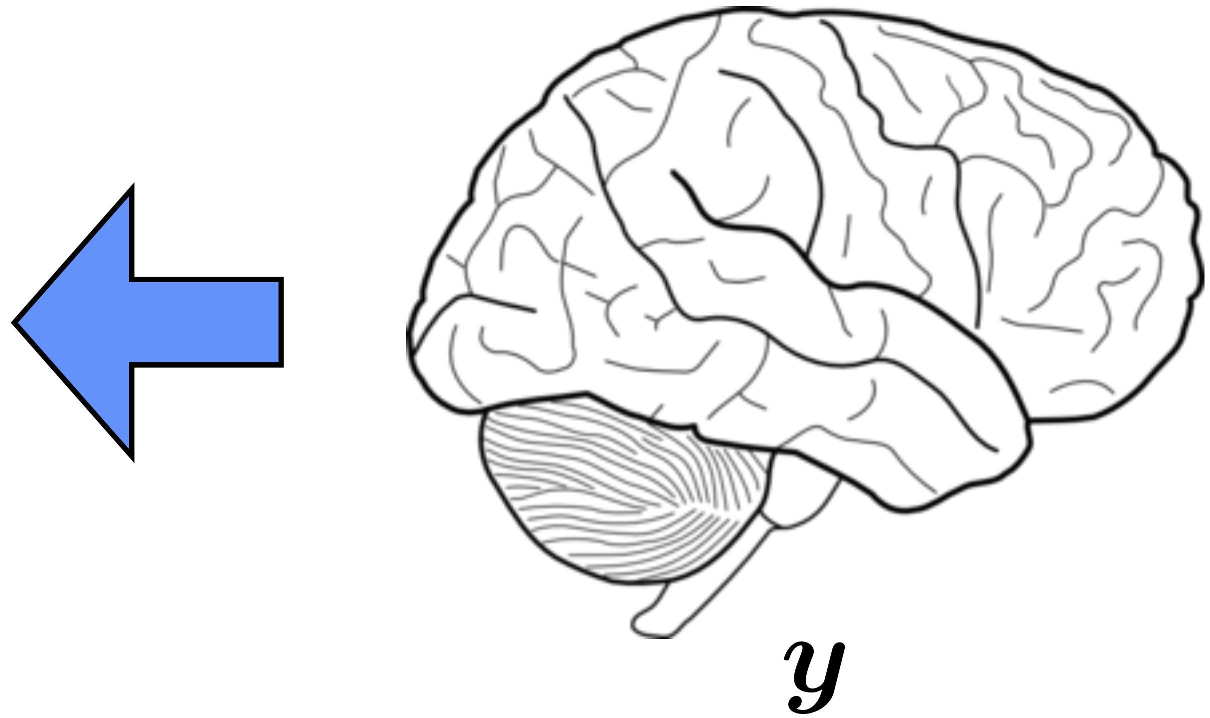
y

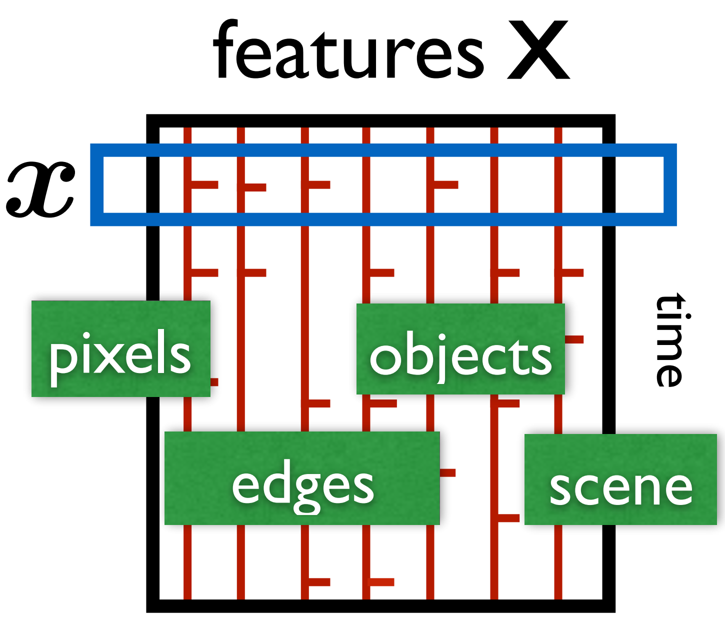


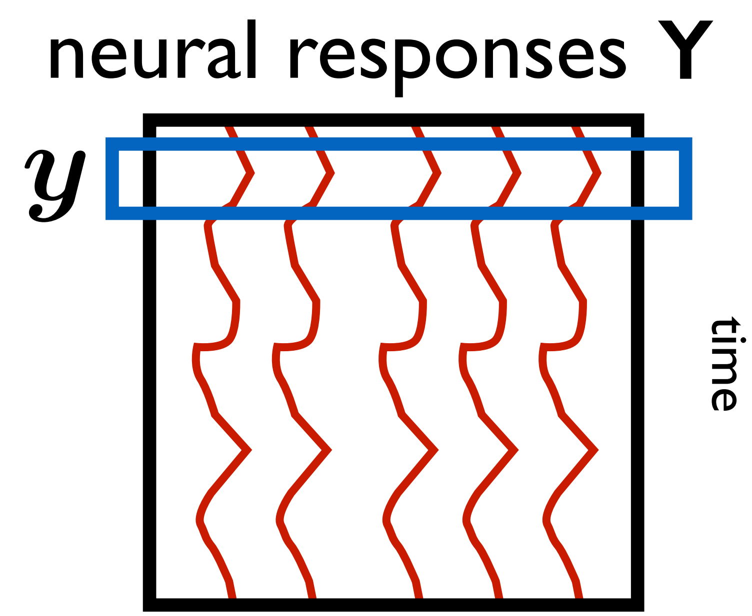


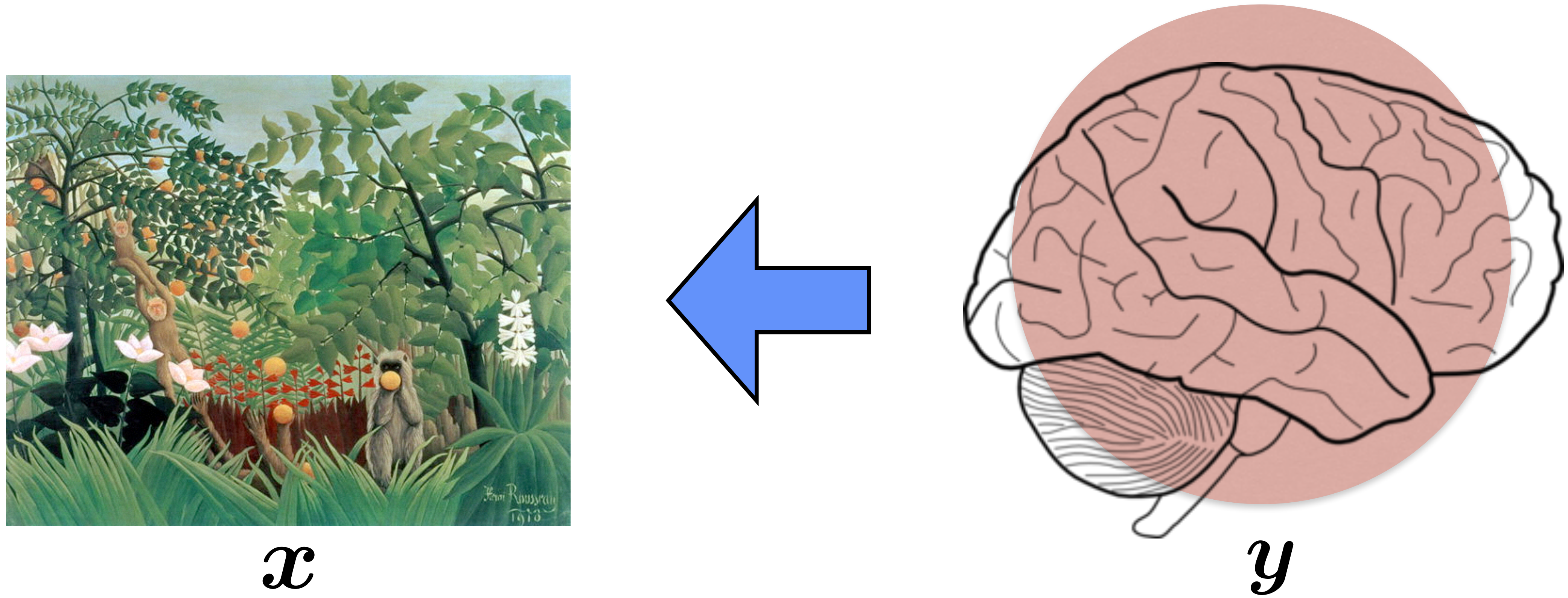
Decoding


x


y

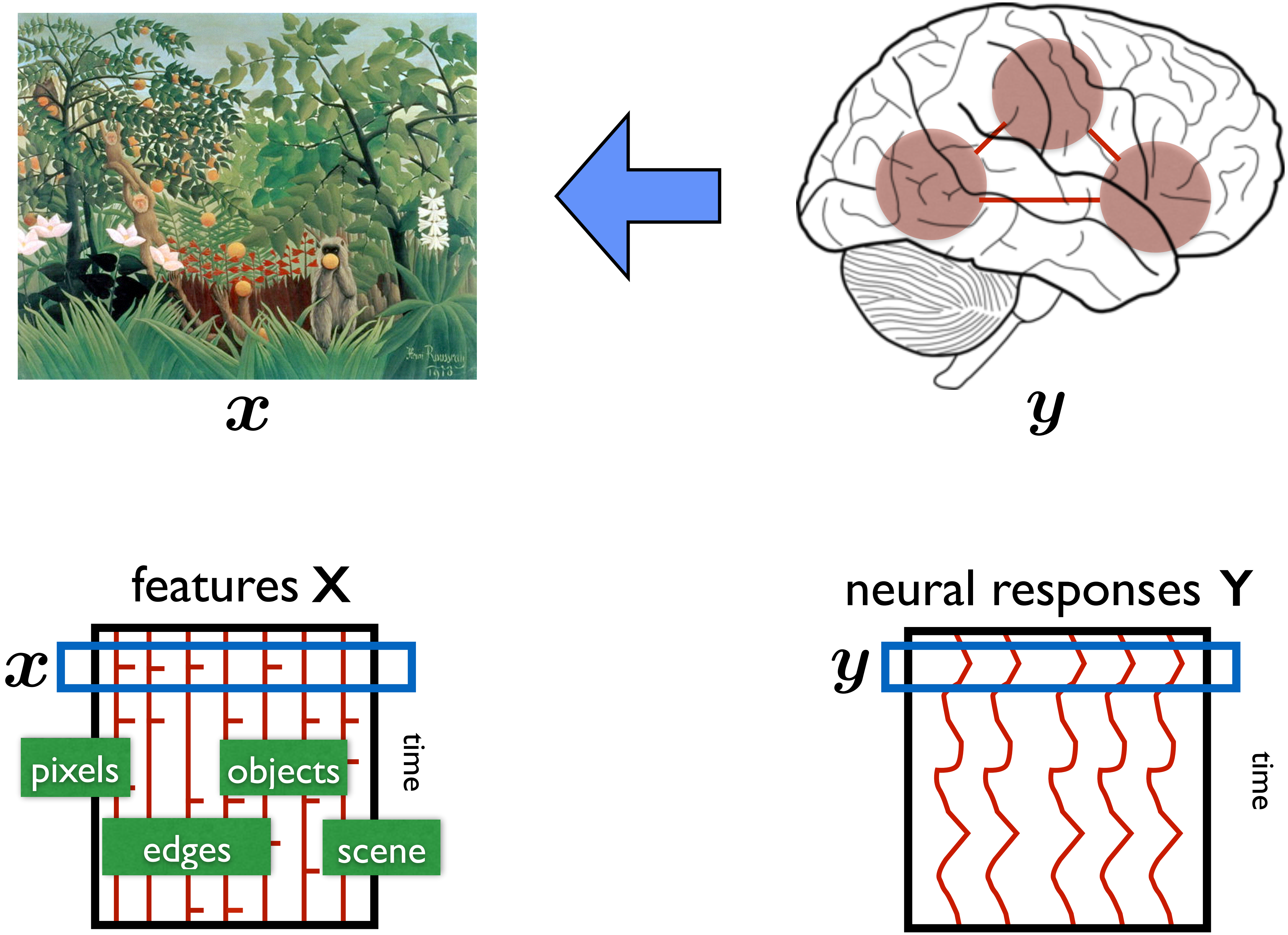

features X
x
pixels
objects
edges
scene
time


neural responses Y
y
time

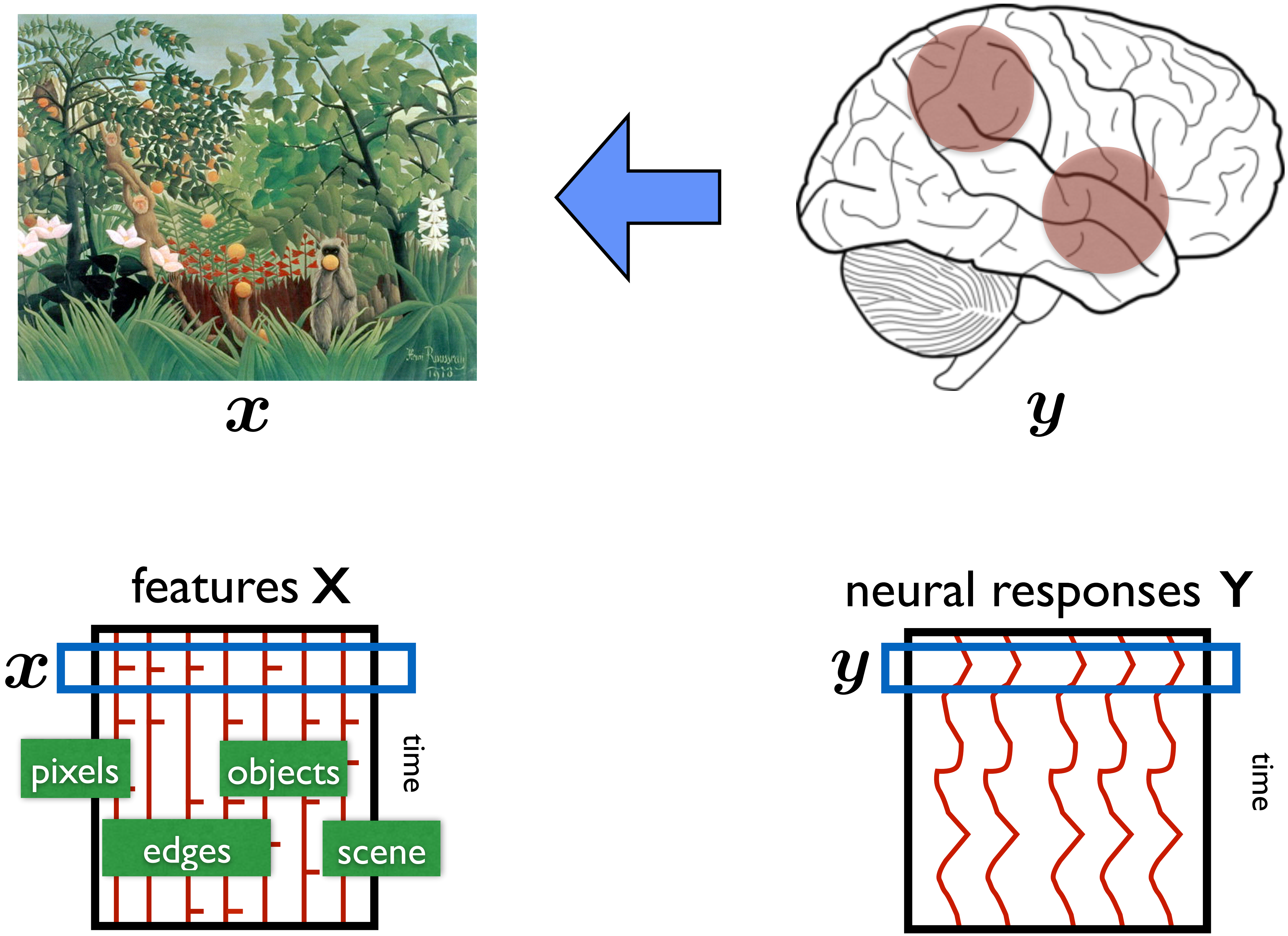

x
y


features X
x
pixels
objects
edges
scene
time
neural responses Y
y
time

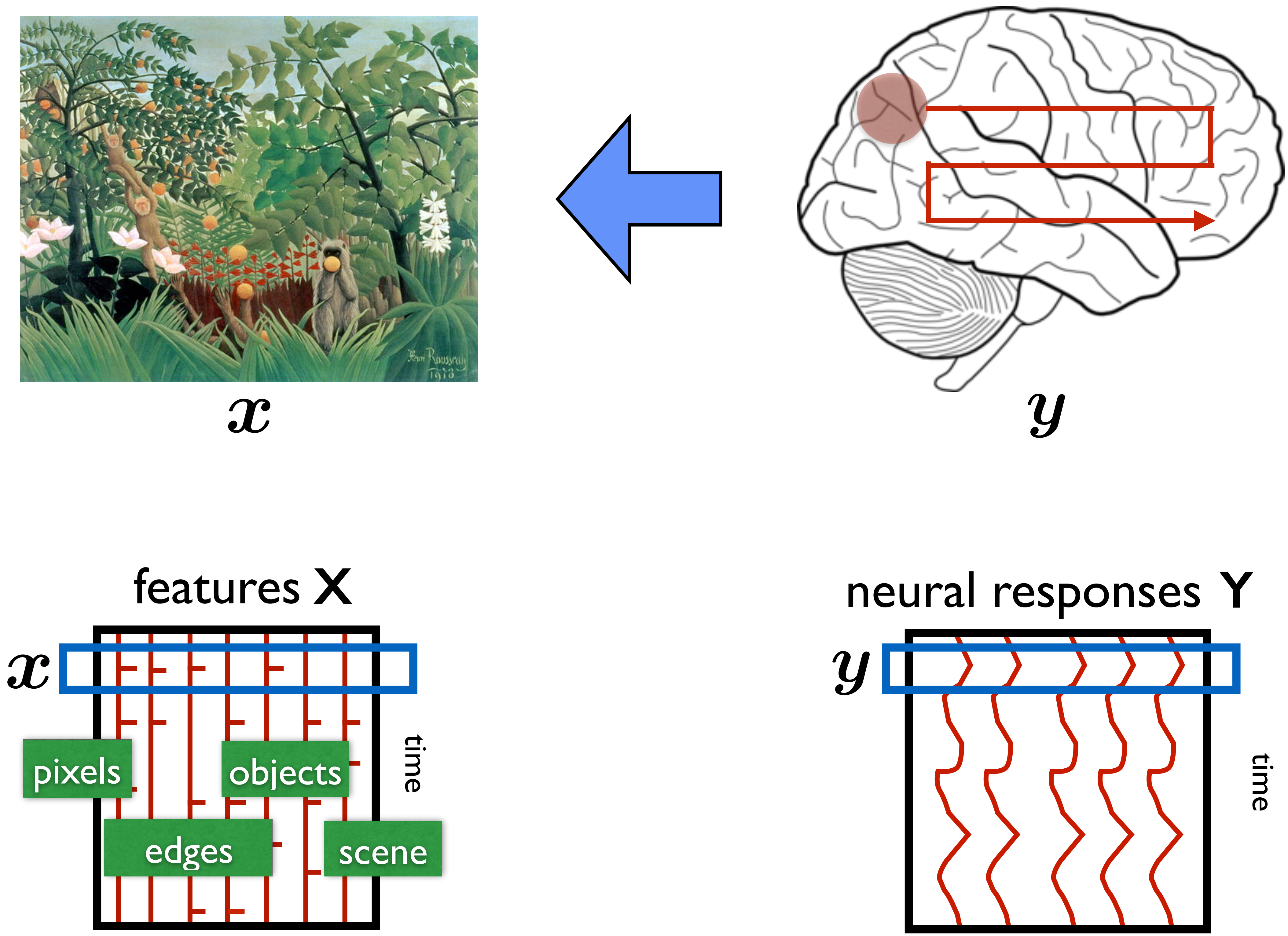
Connectivity-based approach



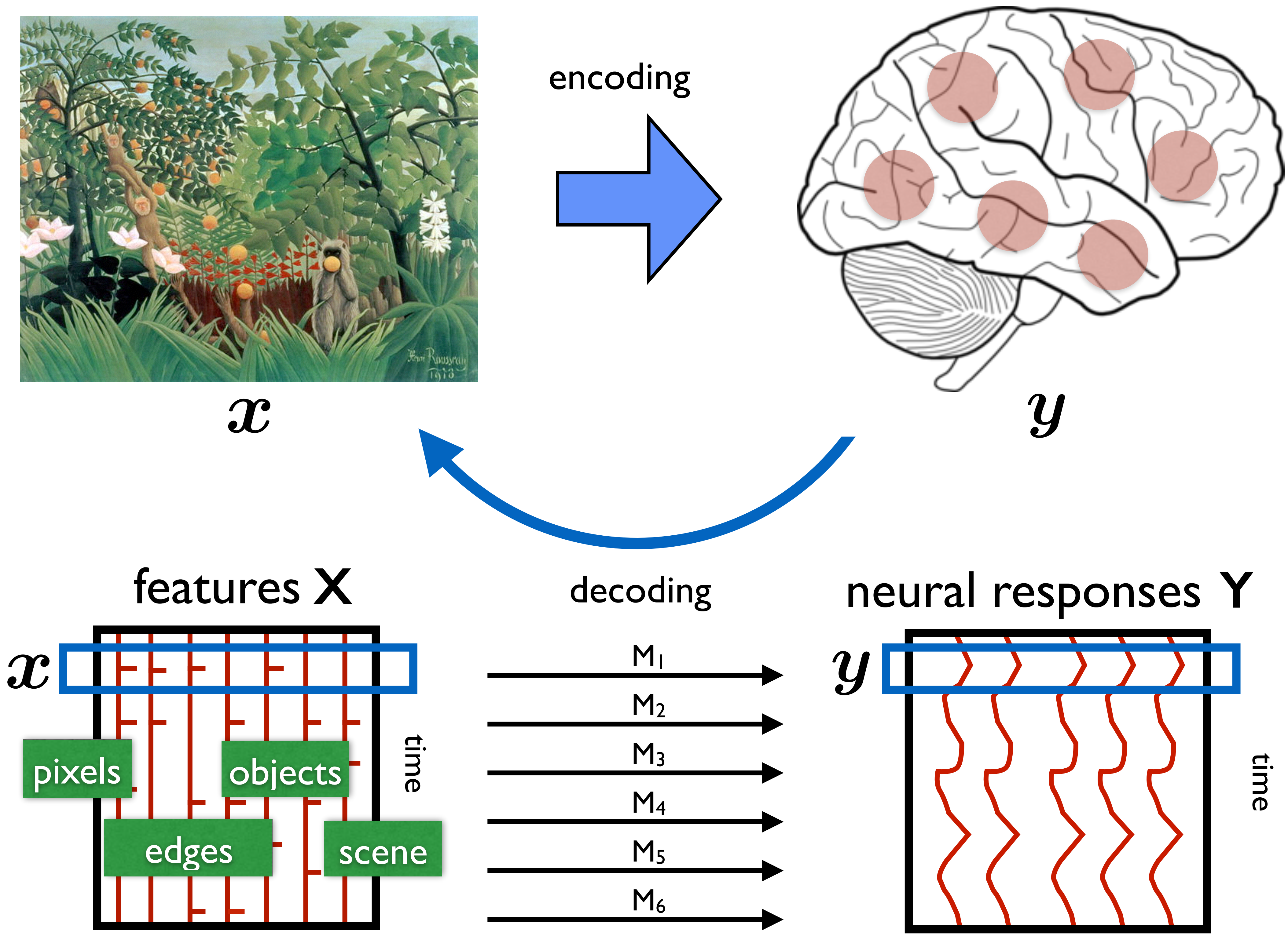
Region-of-interest approach



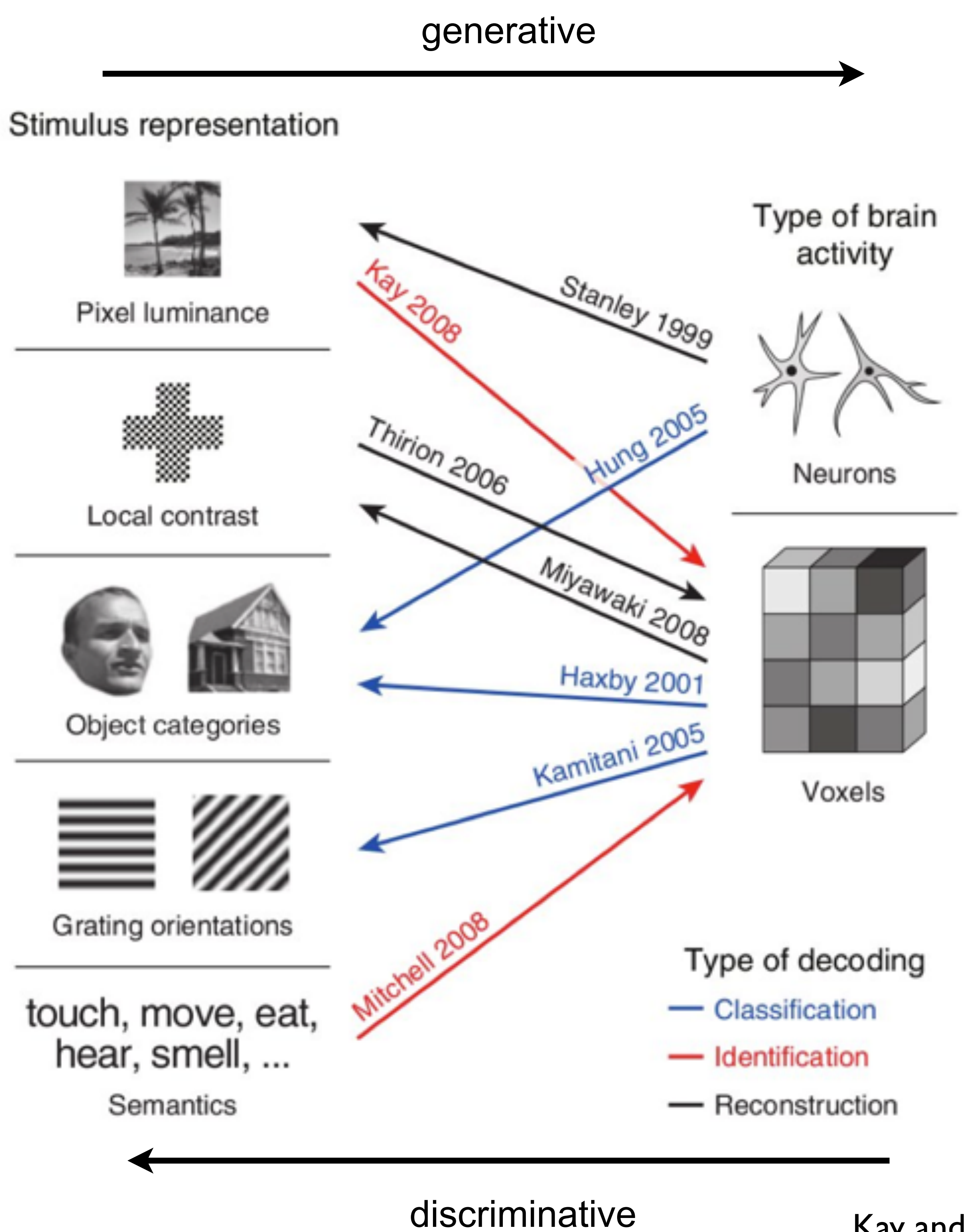
Searchlight approach



Generative approach



Decoding flavors



Kay and Gallant, Nature, 2009

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

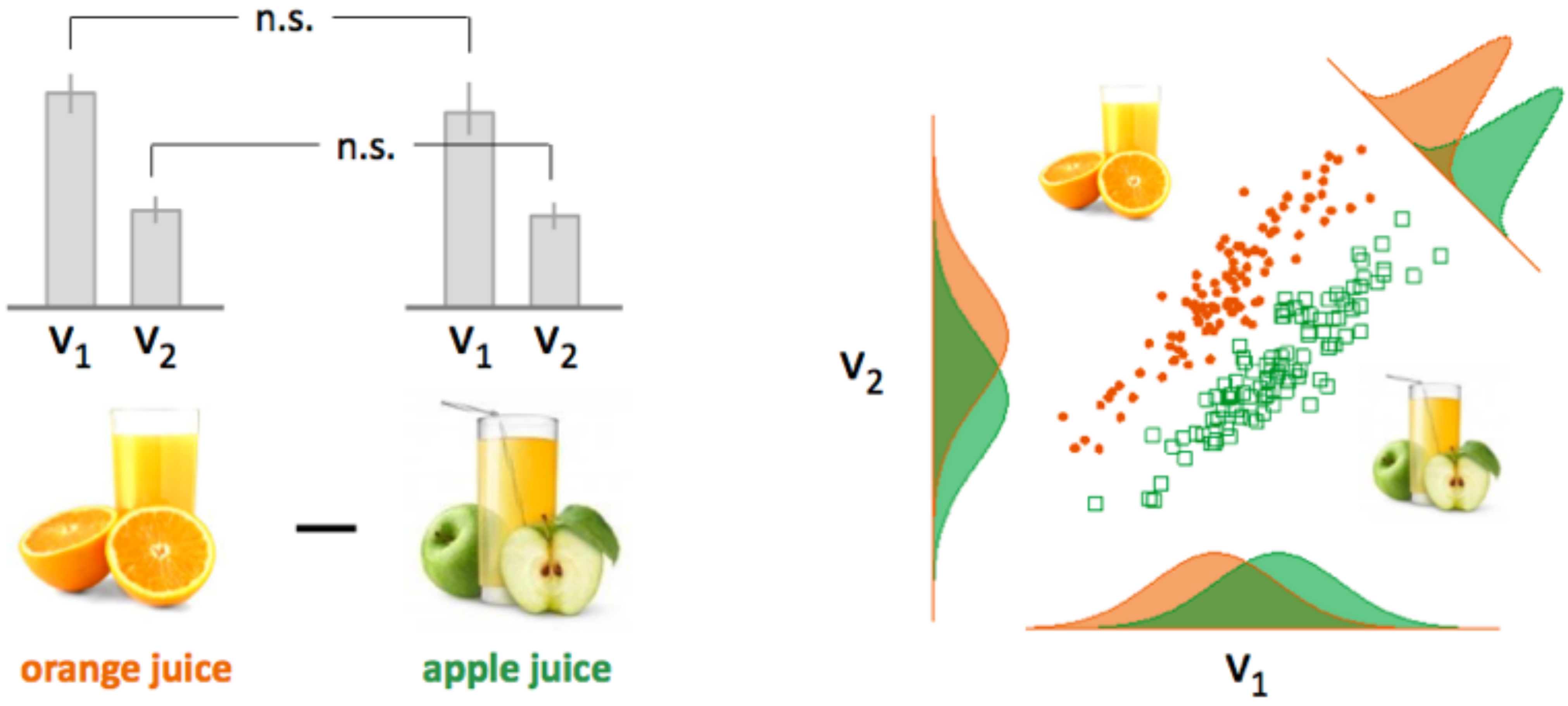
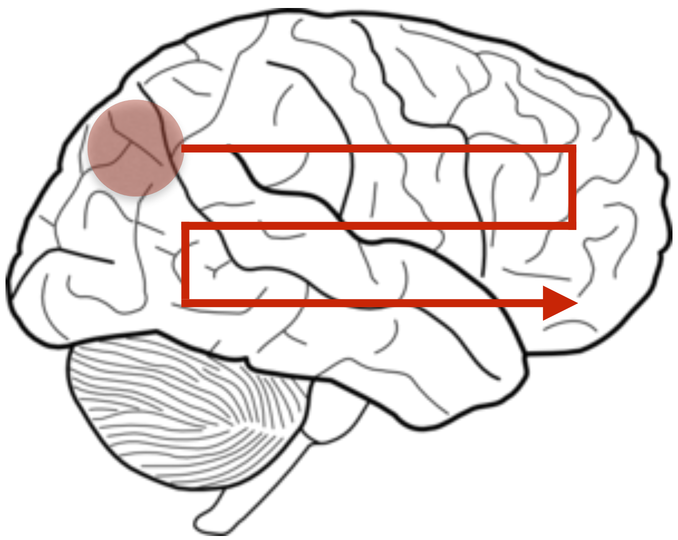


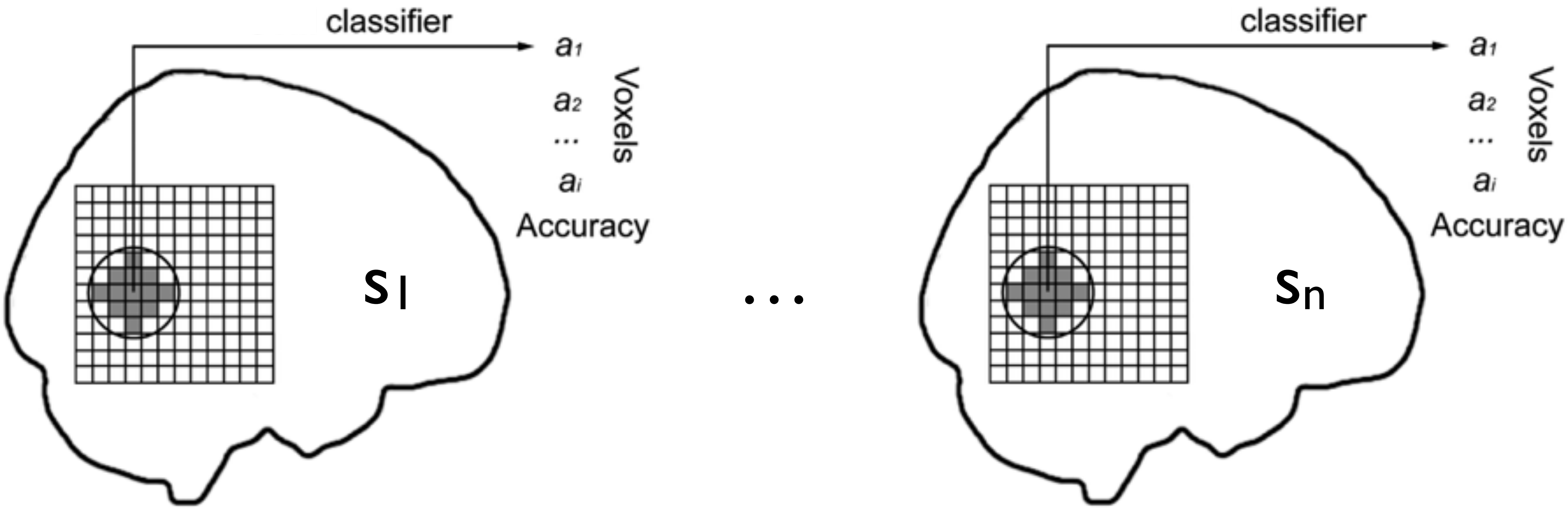
Figure courtesy Kai Brodersen

Example: Decoding amodal conceptual representations



- Animal and tool stimuli presented in four different modalities:
  - pictures
  - sounds
  - spoken words
  - written words
- What brain regions respond to concept information independent of modality?
- Can we decode conceptual representations from these areas during free recall?

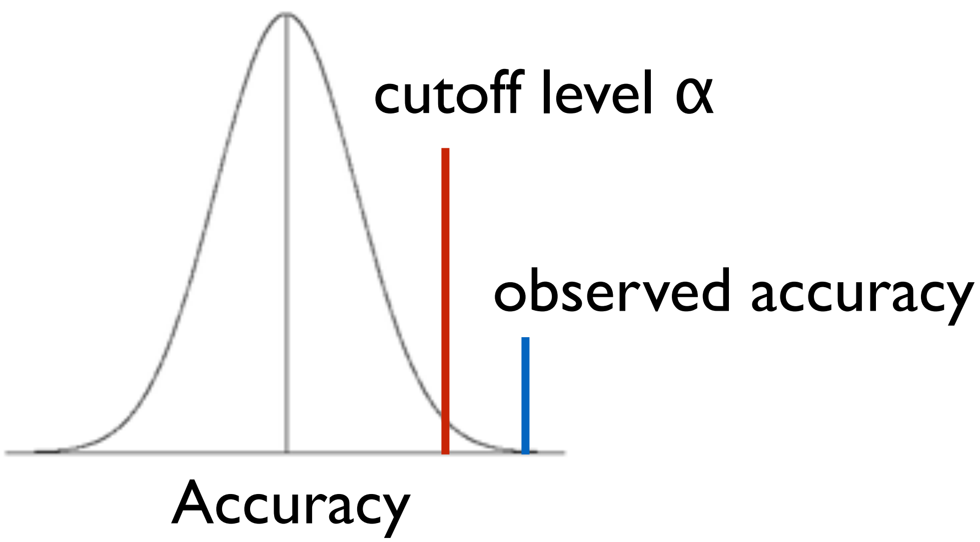
Searchlight approach



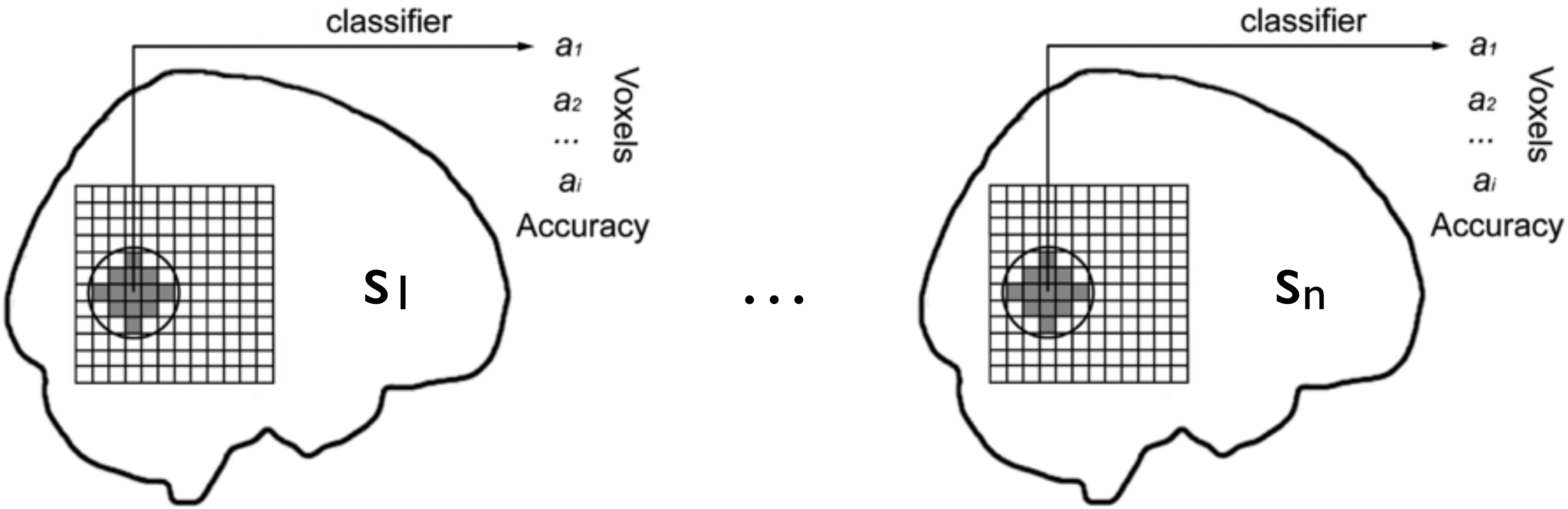
Group-level significance maps:

1. non-parametric permutation test; for each sphere in each subject:

- randomly relabel trials
- run classifier (cross-validated)
- record accuracy
- repeat hundreds of times

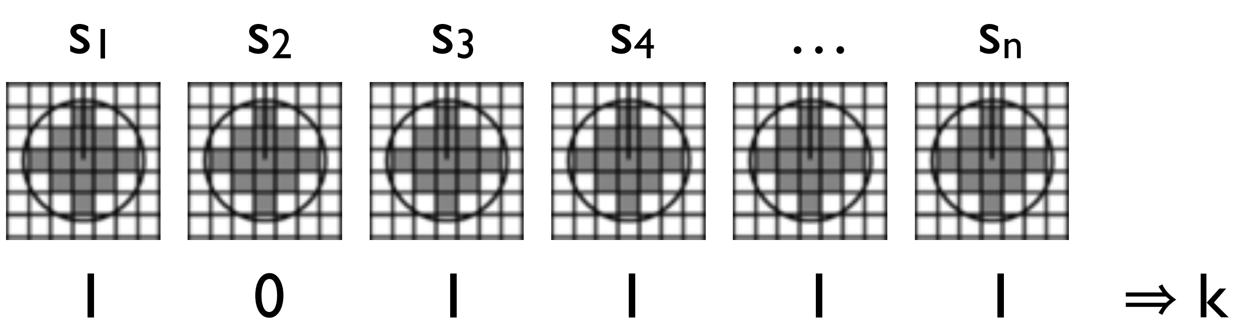


Searchlight approach



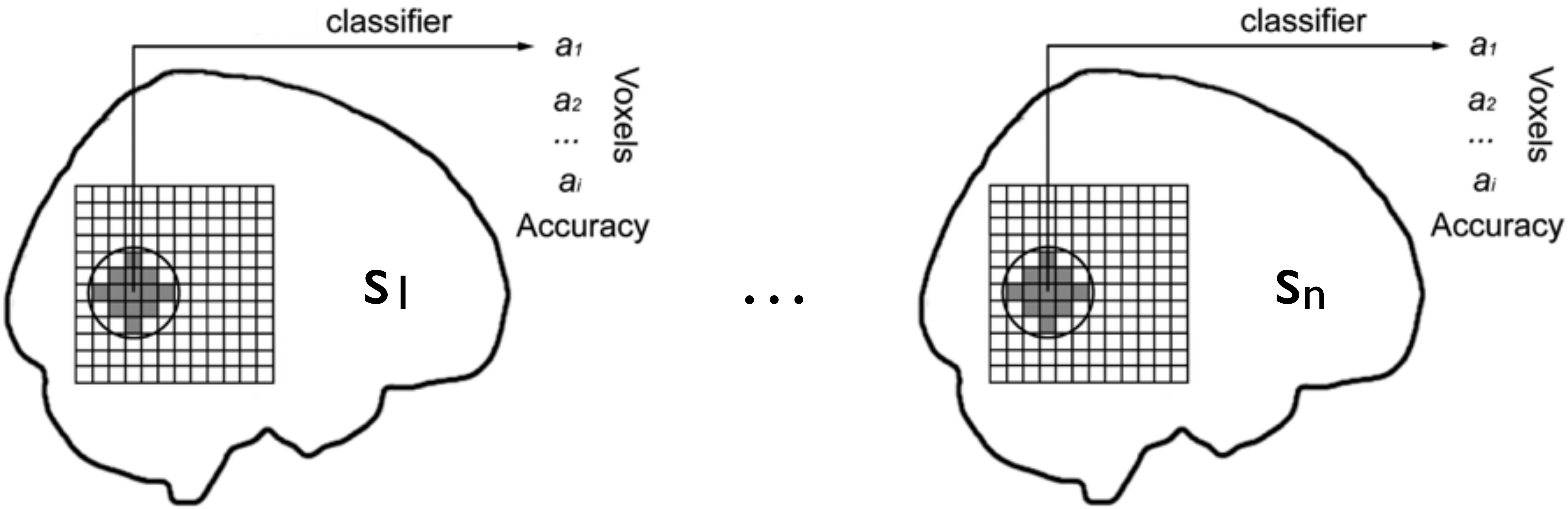
Group-level significance maps:

2. binomial test over subjects:



$$p = 1 - \mathrm{binocdf}(k-1, n, \alpha)$$

Searchlight approach

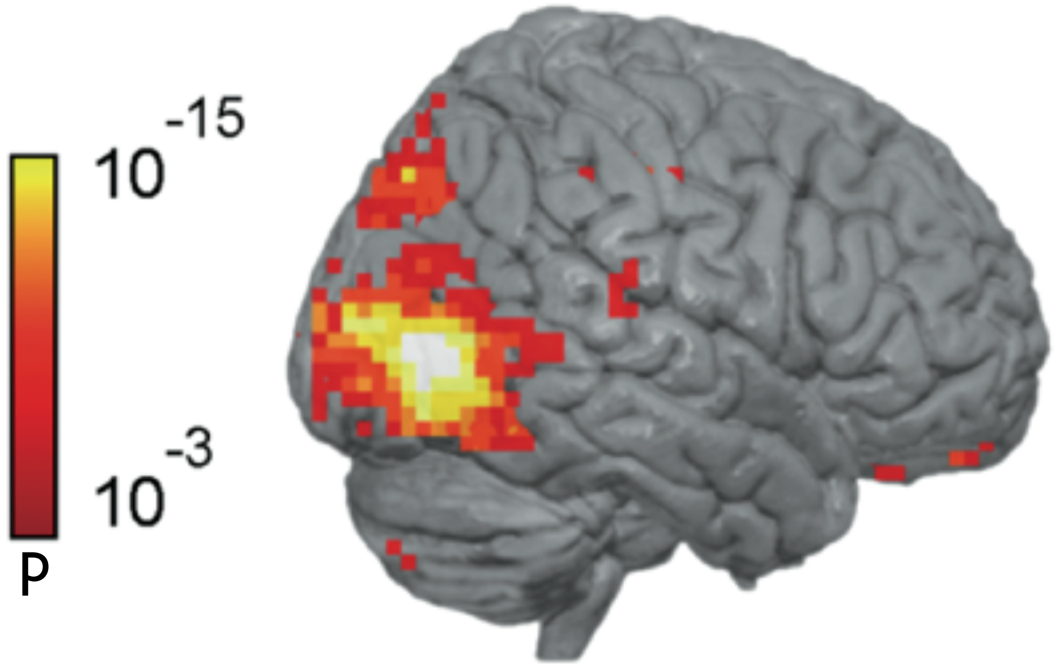


Group-level significance maps:

3. Selection of significant spheres using p value FDR-corrected for nr of spheres

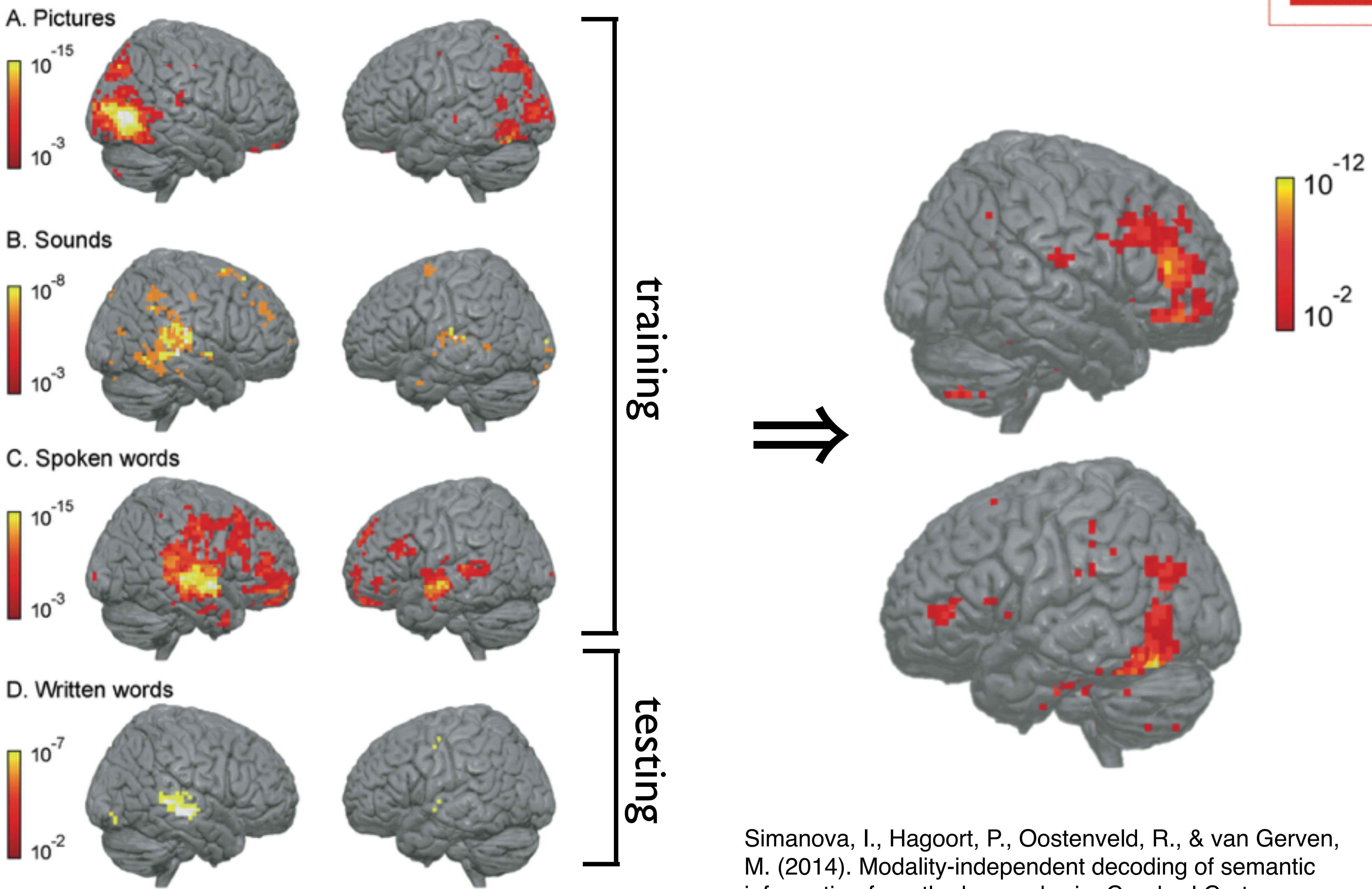
- sort group-level spheres according to p-values: $(p_1,\ldots,p_M)$
- find largest m such that $p_m \leq \alpha m/M$
- keep spheres 1,..,m

Group-level map for animal versus tool images



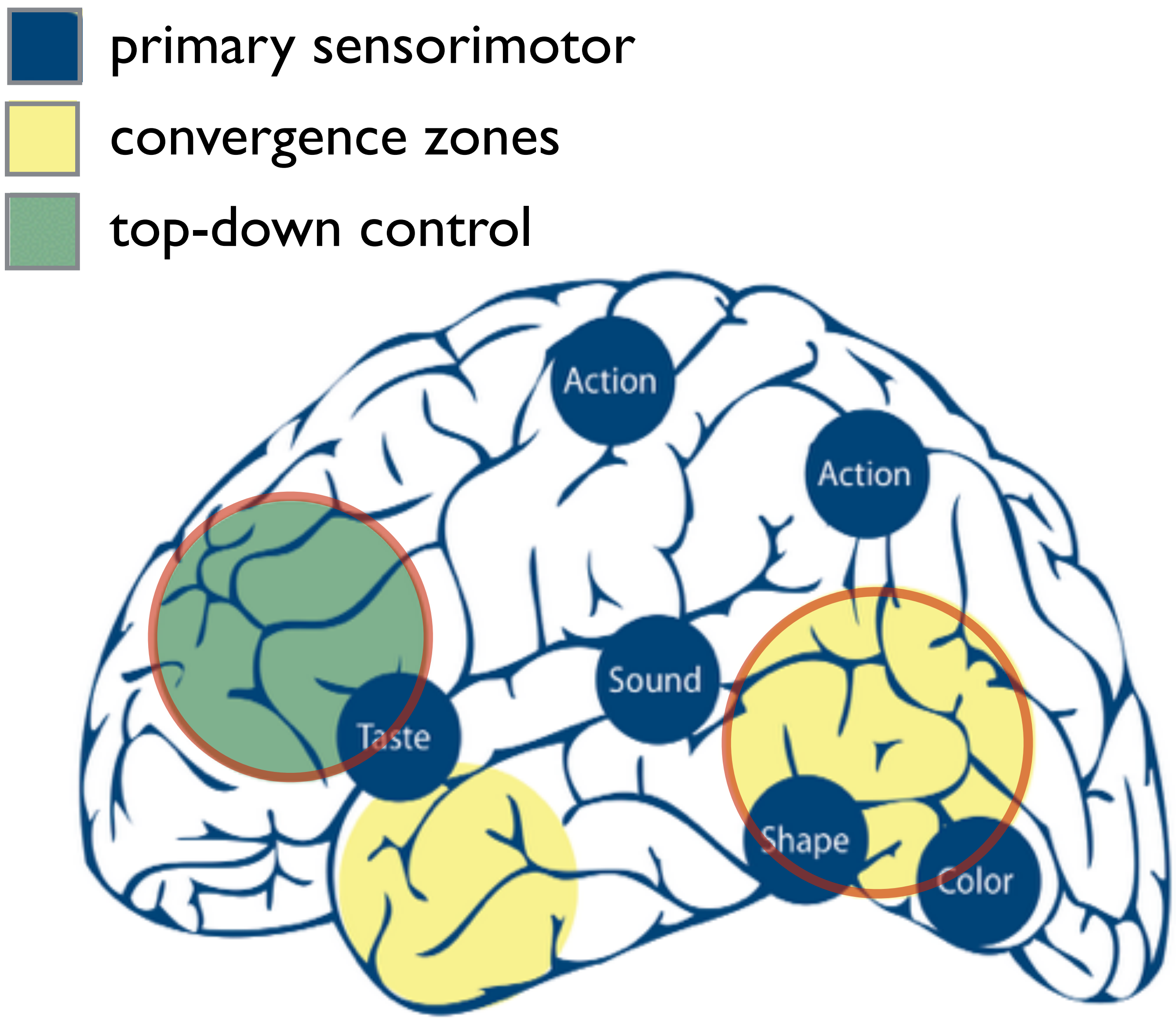
- Decoding of animals versus tools driven by early visual areas
- However, could be driven by low-level visual properties…

From single modalities to amodal representations

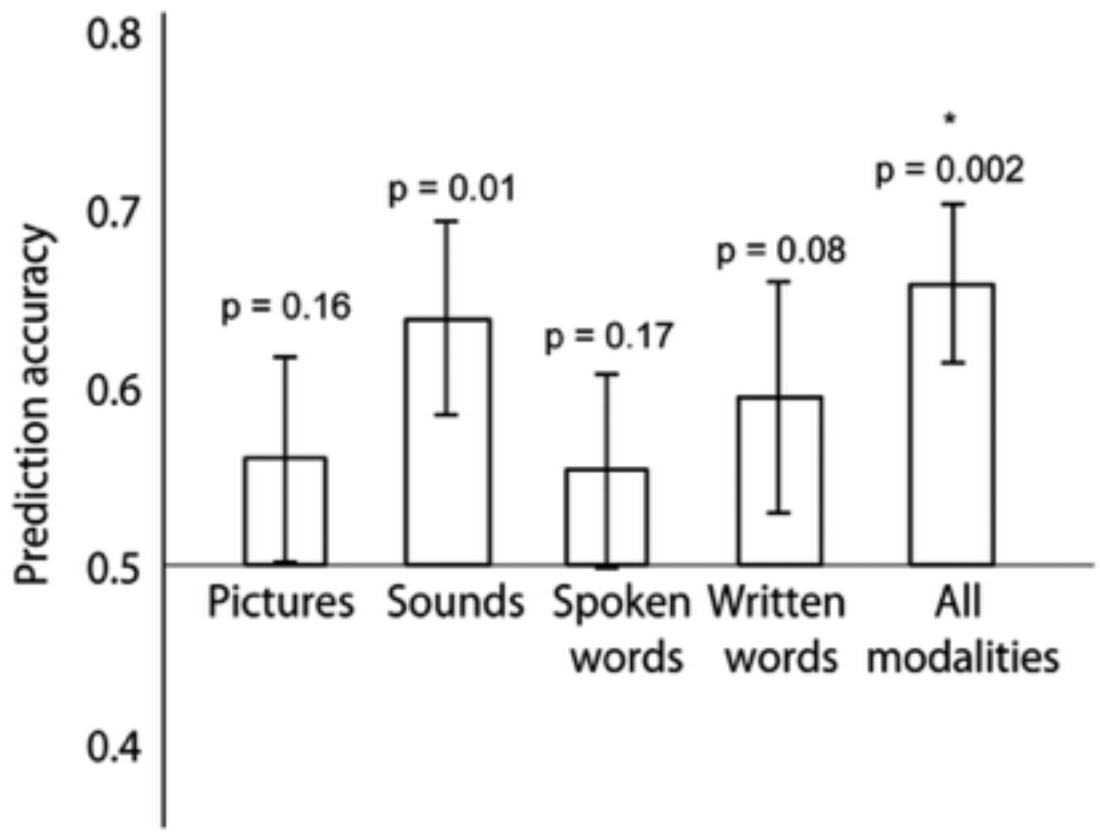


Simanova, I., Hagoort, P., Oostenveld, R., & van Gerven, M. (2014). Modality-independent decoding of semantic information from the human brain. Cerebral Cortex.

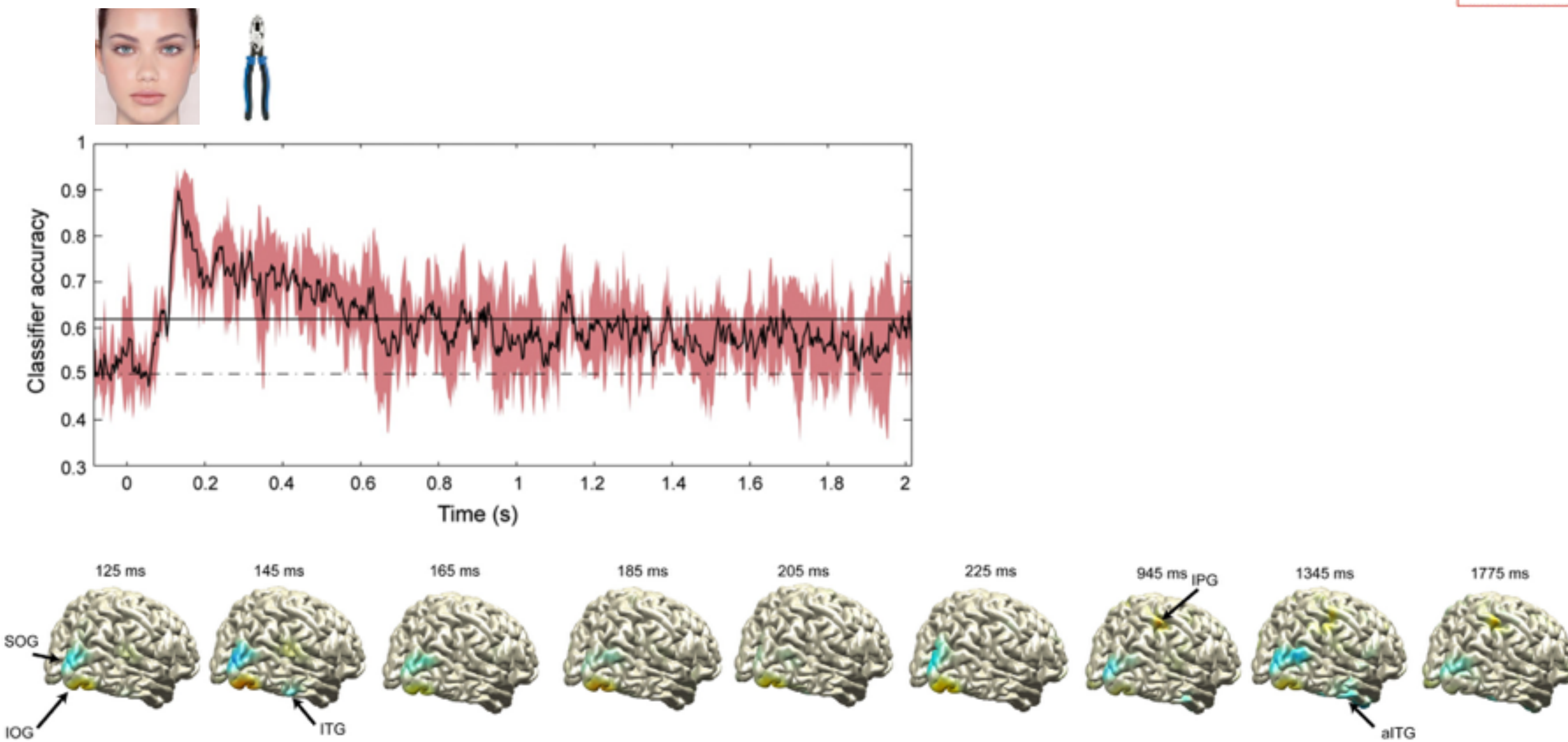
Decoding amodal conceptual representations



Free recall:



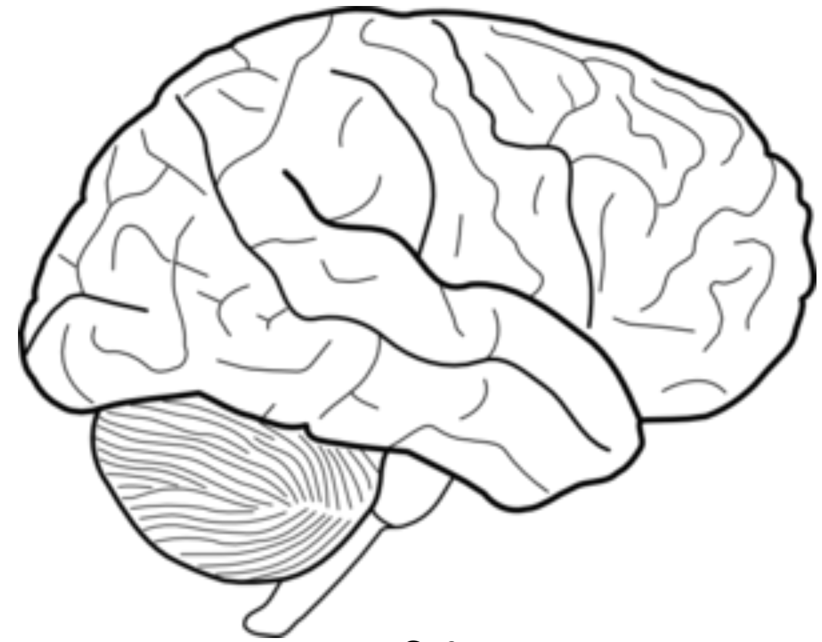
Probing spatiotemporal dynamics



Van de Nieuwenhuijzen et al. (2013). MEG-based decoding of the spatiotemporal dynamics of visual category perception. NeuroImage.

Also see e.g.:
Harrison & Tong (2009). Nature; Sudre et al. (2012). Neuroimage; Carlson et al. (2013). JoV; King & DeHaene (2014). TiCS; Albers et al. (2013). Current Biology; Isik et al. (2014). J. Neurophys.

$\boldsymbol{x}$



$\boldsymbol{y}$

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

Decoding representations (generative approach)

$\boldsymbol{x}$

encoding

decoding

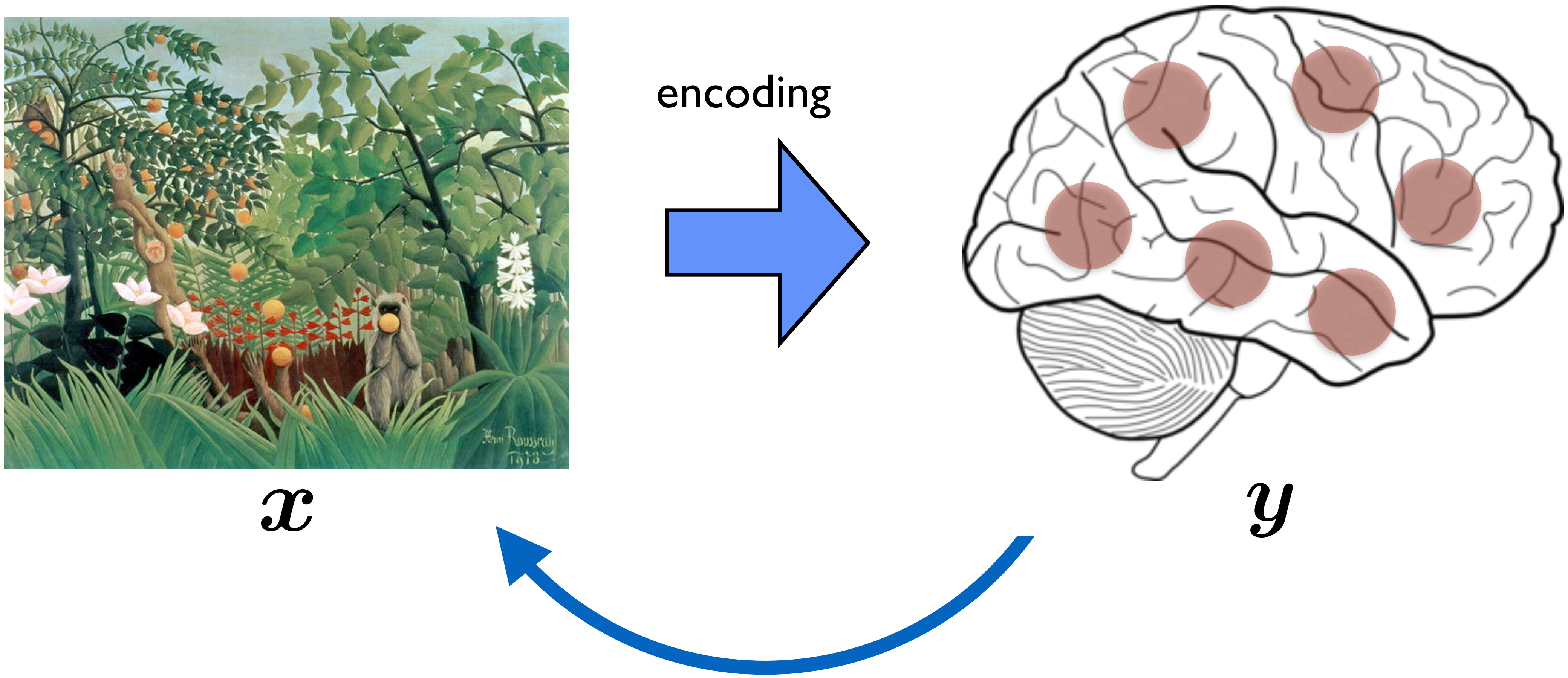
$\boldsymbol{y}$

encoding model

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

stimulus prior

Gaussian framework



$$\mathcal{N}(\boldsymbol{y}; \boldsymbol{B}^T \boldsymbol{x}, \boldsymbol{\Sigma})$$

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

$$\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \boldsymbol{R})$$

Without confounds and ignoring the HRF, we get

$$\mathbf{y}_k = \mathbf{X}\boldsymbol{\beta}_k + \boldsymbol{\epsilon}$$

This boils down to a linear Gaussian model:

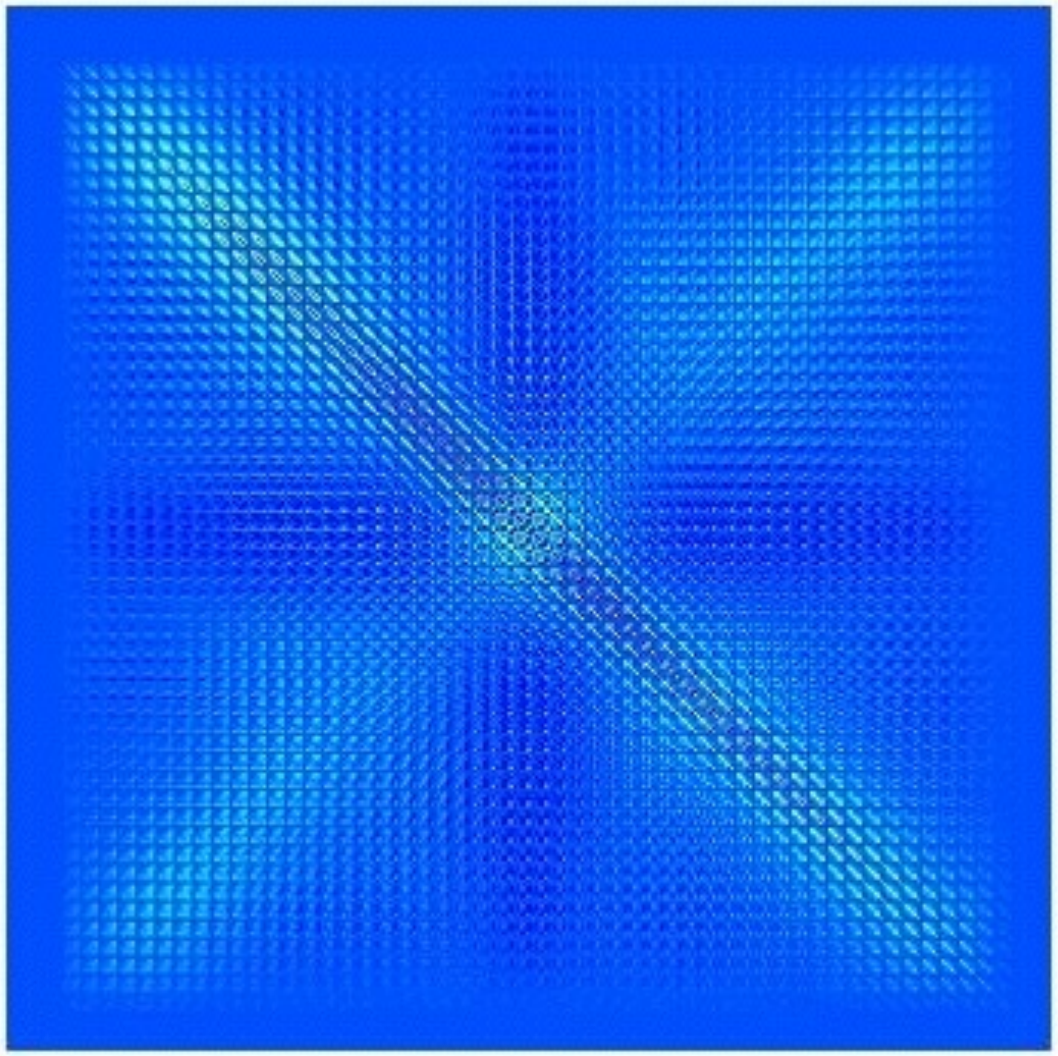
$$p(\mathbf{y}_k \mid \mathbf{X}, \boldsymbol{\beta}_k, \sigma^2) = \mathcal{N}(\mathbf{y}_k; \mathbf{X}\boldsymbol{\beta}_k, \sigma^2 \mathbf{I}_N)$$

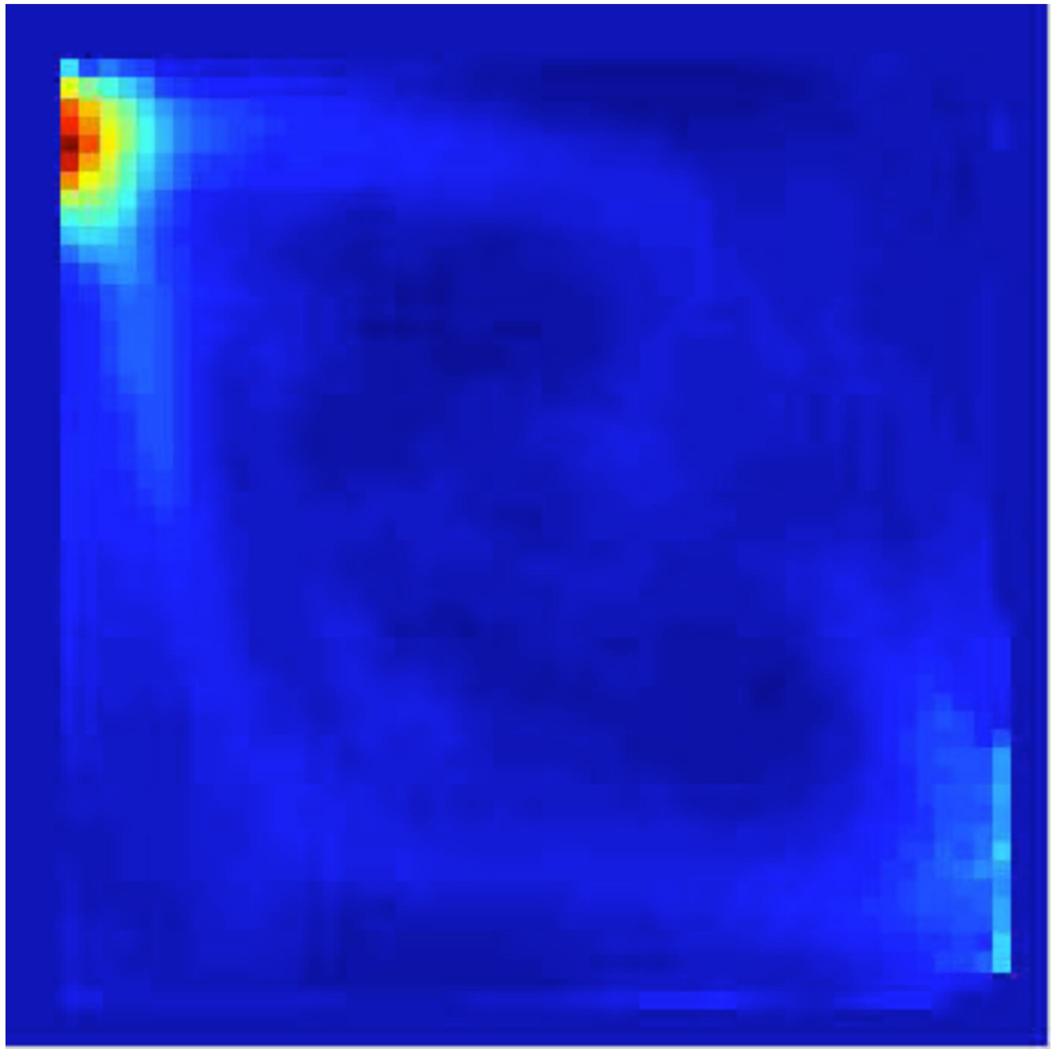
The least squares solution for $\boldsymbol{\beta}_k$ given a Gaussian prior on $\boldsymbol{\beta}_k$ is given by

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{X}\mathbf{X} + \lambda \mathbf{I}_N)^{-1} \mathbf{X}^T \mathbf{y}_k$$

The matrix **B** of regression coefficients is given by $[\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K]$.

For a Gaussian image prior $\mathcal{N}(\boldsymbol{x}; \mathbf{0}, \boldsymbol{R})$ we compute the covariance matrix $\boldsymbol{R}$ using a separate set of handwritten images





The posterior is given by

$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}(\mathbf{x}; \mathbf{m}, \mathbf{Q})$$

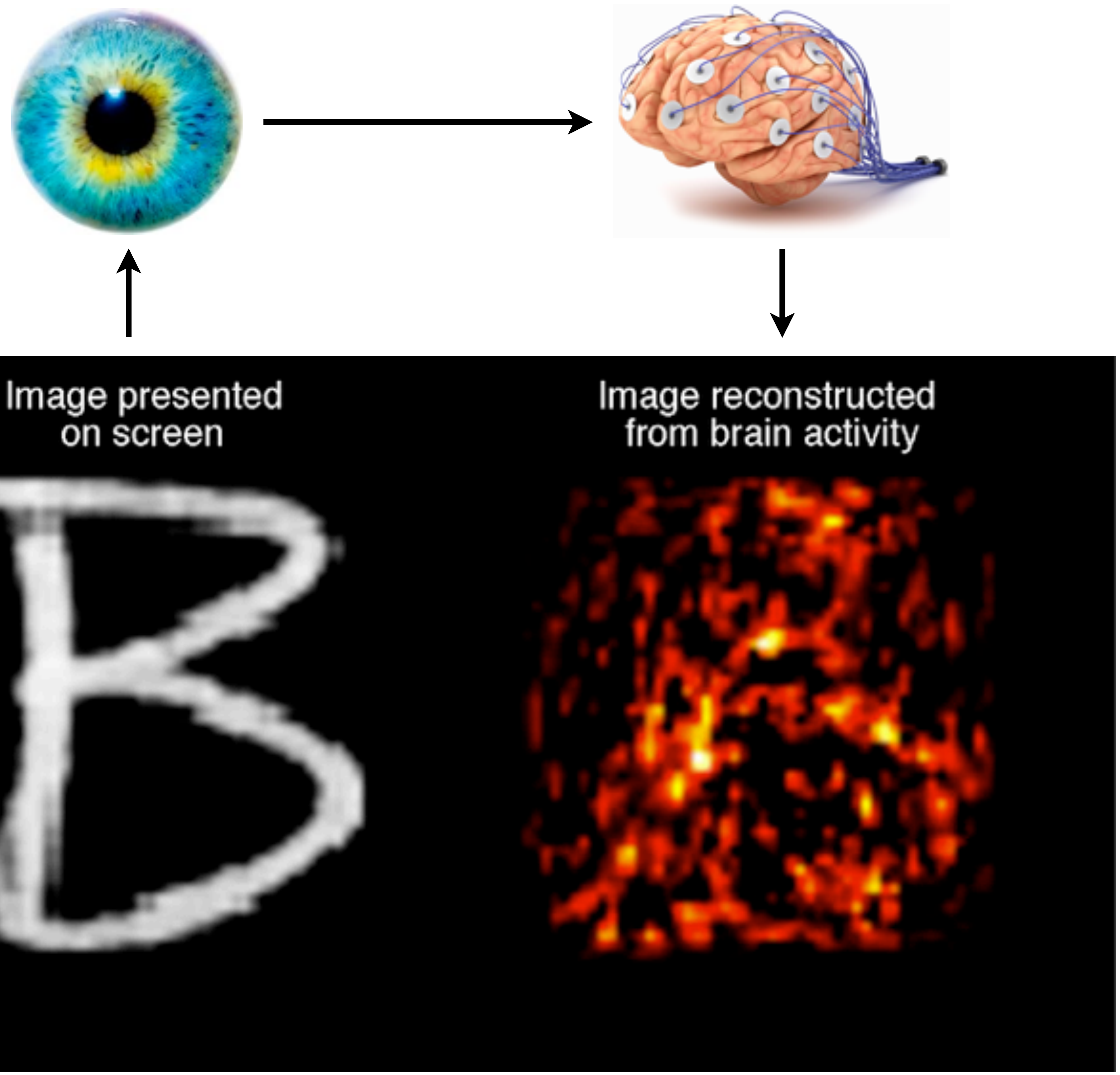
with mean $\mathbf{m} \equiv \mathbf{Q}\mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{y}$ and covariance $\mathbf{Q} = (\mathbf{R}^{-1} + \mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{B}^{\top})^{-1}$.

It immediately follows that the most probable stimulus is given by

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

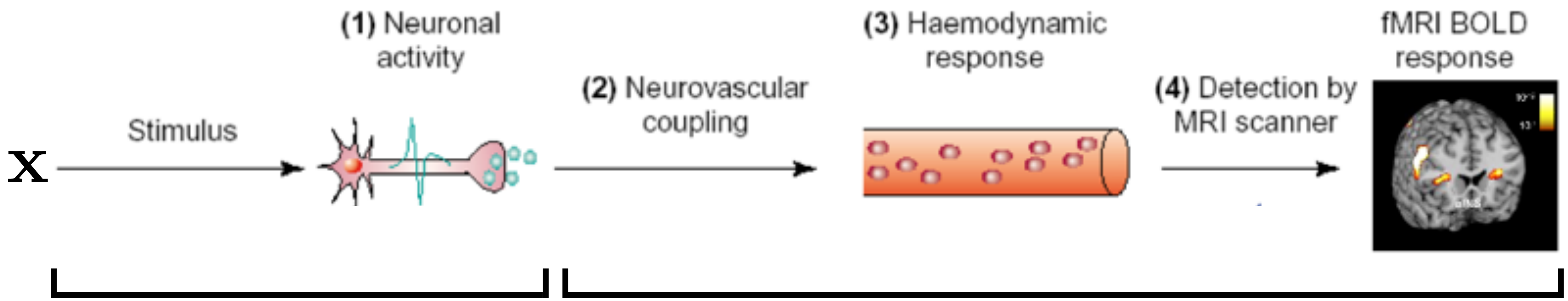
Also see Thirion et al. (2006) Neuroimage

Example: handwritten characters



Schoenmakers, S., Barth, M., Heskes, T., & van Gerven, M. A. J. (2013). Linear reconstruction of perceived images from human brain activity. NeuroImage, 83, 951–961

Building better encoding models



$\boldsymbol{\phi}(\mathbf{x})$ — (non)linear feature space

$\boldsymbol{f}(\cdot)$ — forward model

See Naselaris et al. (2011). Encoding and decoding in fMRI. NeuroImage

$$\mathbf{y} = \mathbf{f}(\phi(\mathbf{x})) + \epsilon$$

- What is the optimal $\boldsymbol{f}(\cdot)$?
- What is the optimal ε?
- What is the optimal $\boldsymbol{\phi}(\mathbf{x})$?

What is the optimal $\boldsymbol{\phi}(\mathbf{x})$?

- Conceptual representations as instantiations of $\boldsymbol{\phi}(\mathbf{x})$

$$\mathbf{x} = \text{image pixels (gaussian model)}$$

$$\phi_1(\mathbf{x}) = \text{local contrast energy}$$

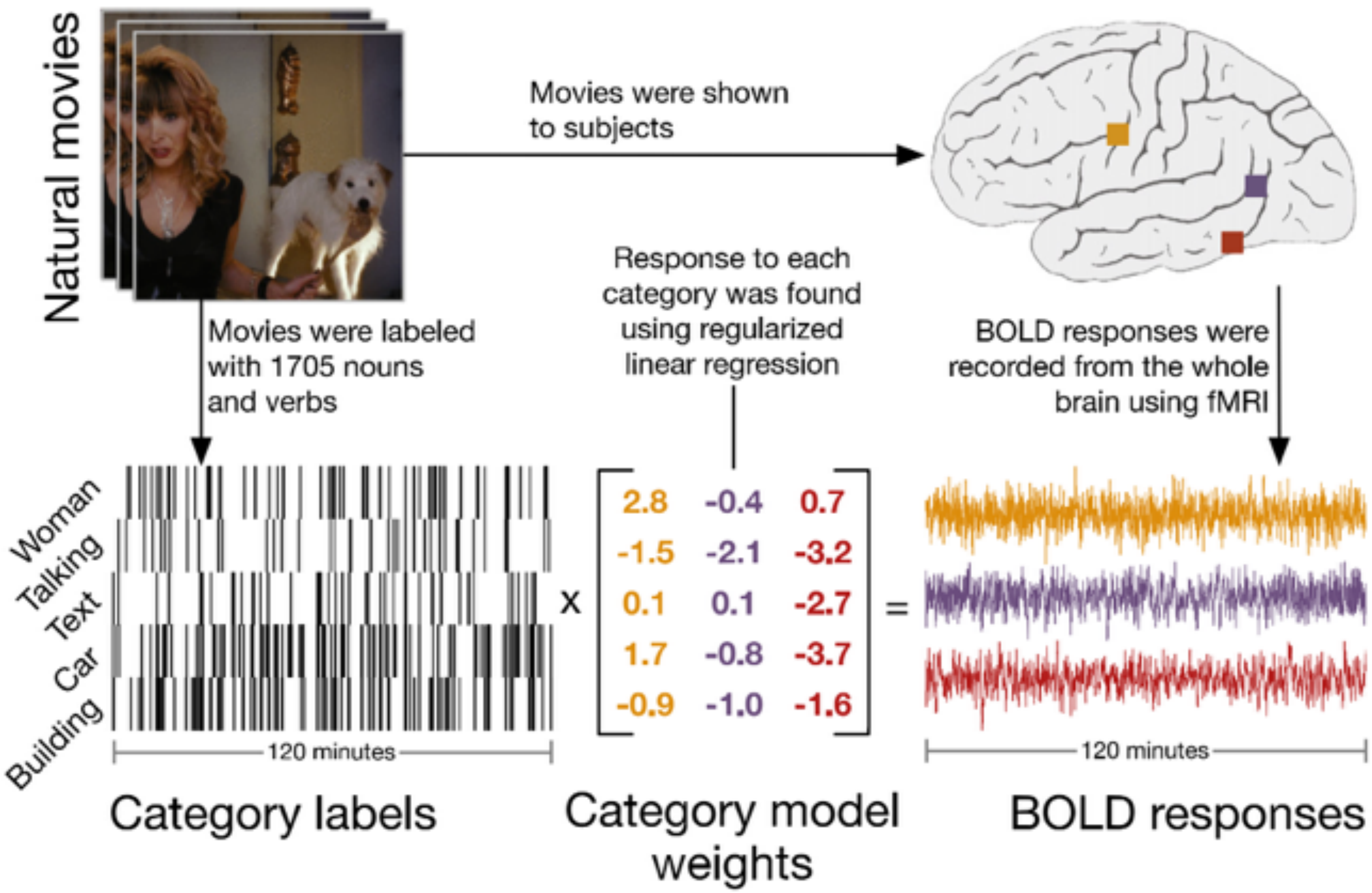
$$\phi_2(\mathbf{x}) = \text{edges and contours}$$

$$\vdots$$

$$\phi_M(\mathbf{x}) = \text{objects}$$

- How to obtain these $\boldsymbol{\phi}_i(\mathbf{x})$?

$\phi(\mathbf{x})$ as a manual relabelling

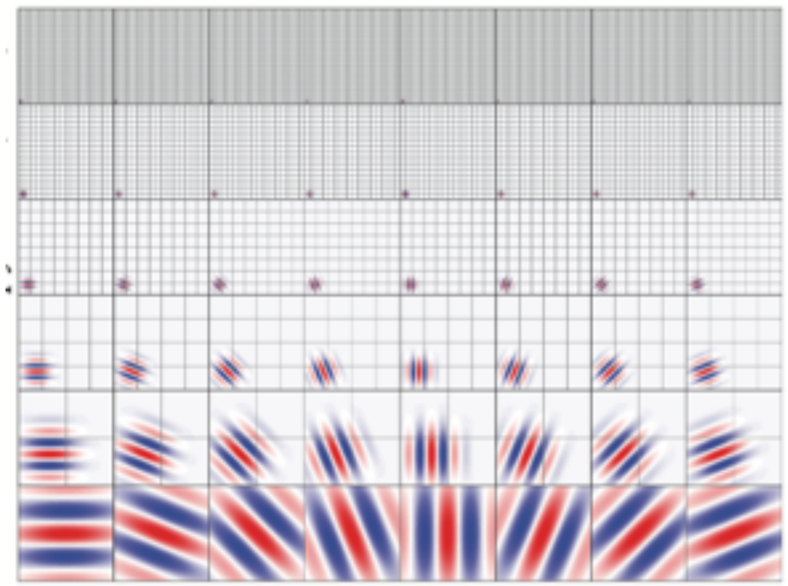


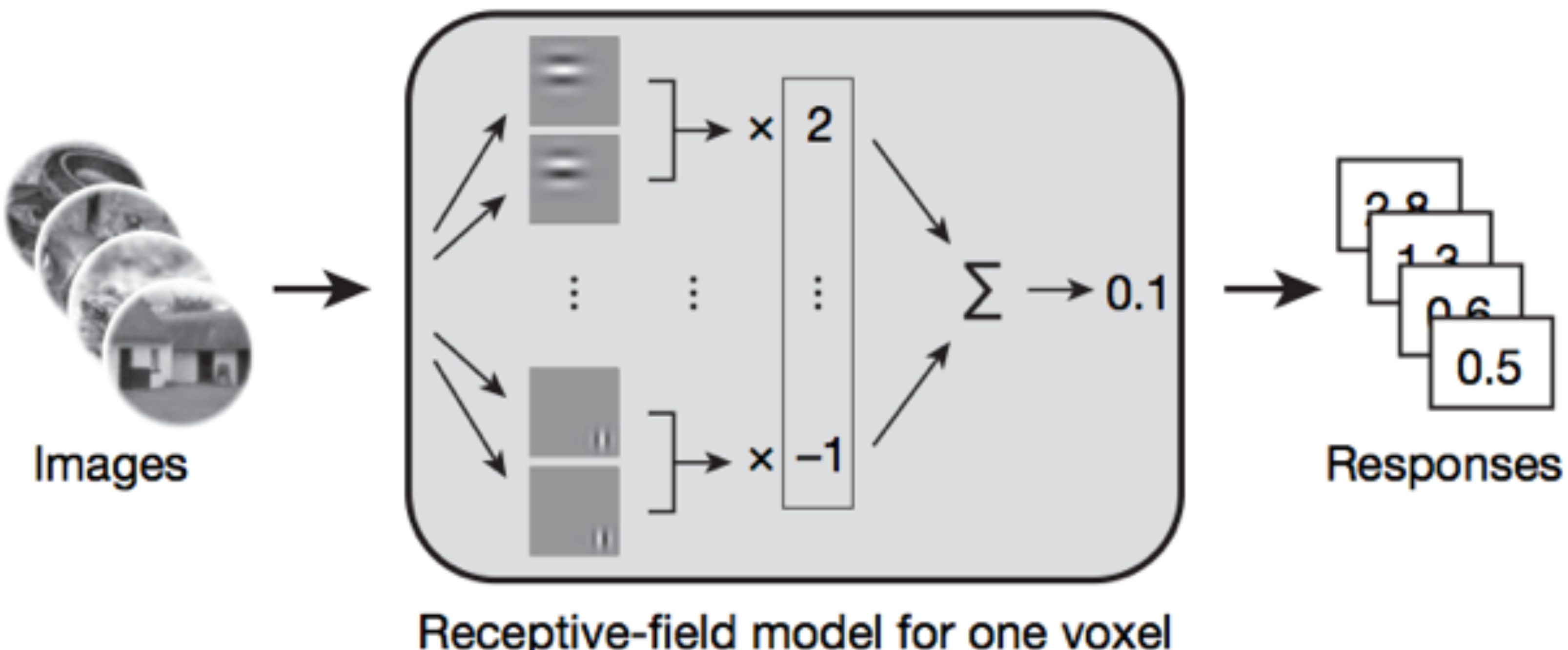
Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron.

- Labelling of objects and actions in each movie frame using 1364 Wordnet terms
- Labour intensive…

$\phi(\mathbf{x})$ as a canonical transformation

Gabor wavelet pyramid (GWP)

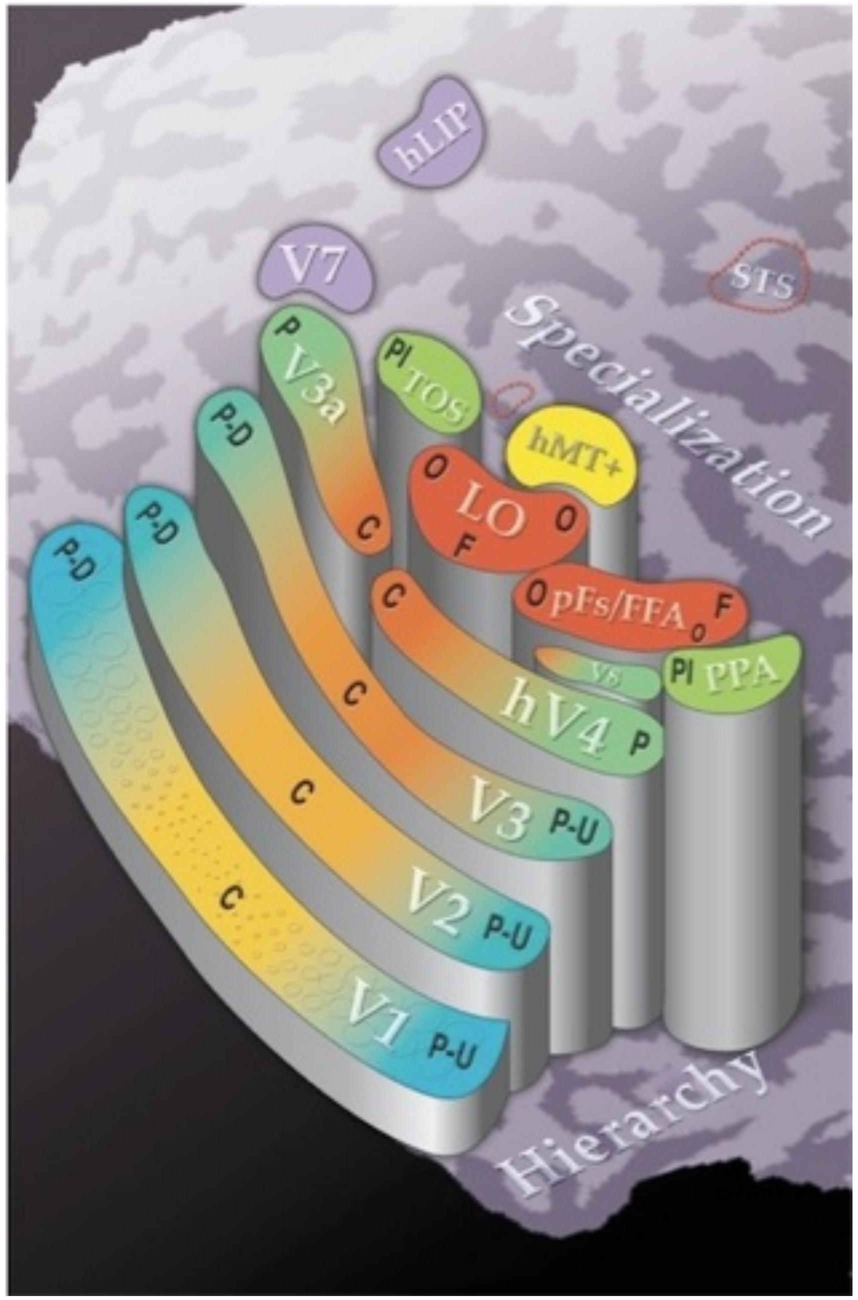




Kay et al, Nature, 2008

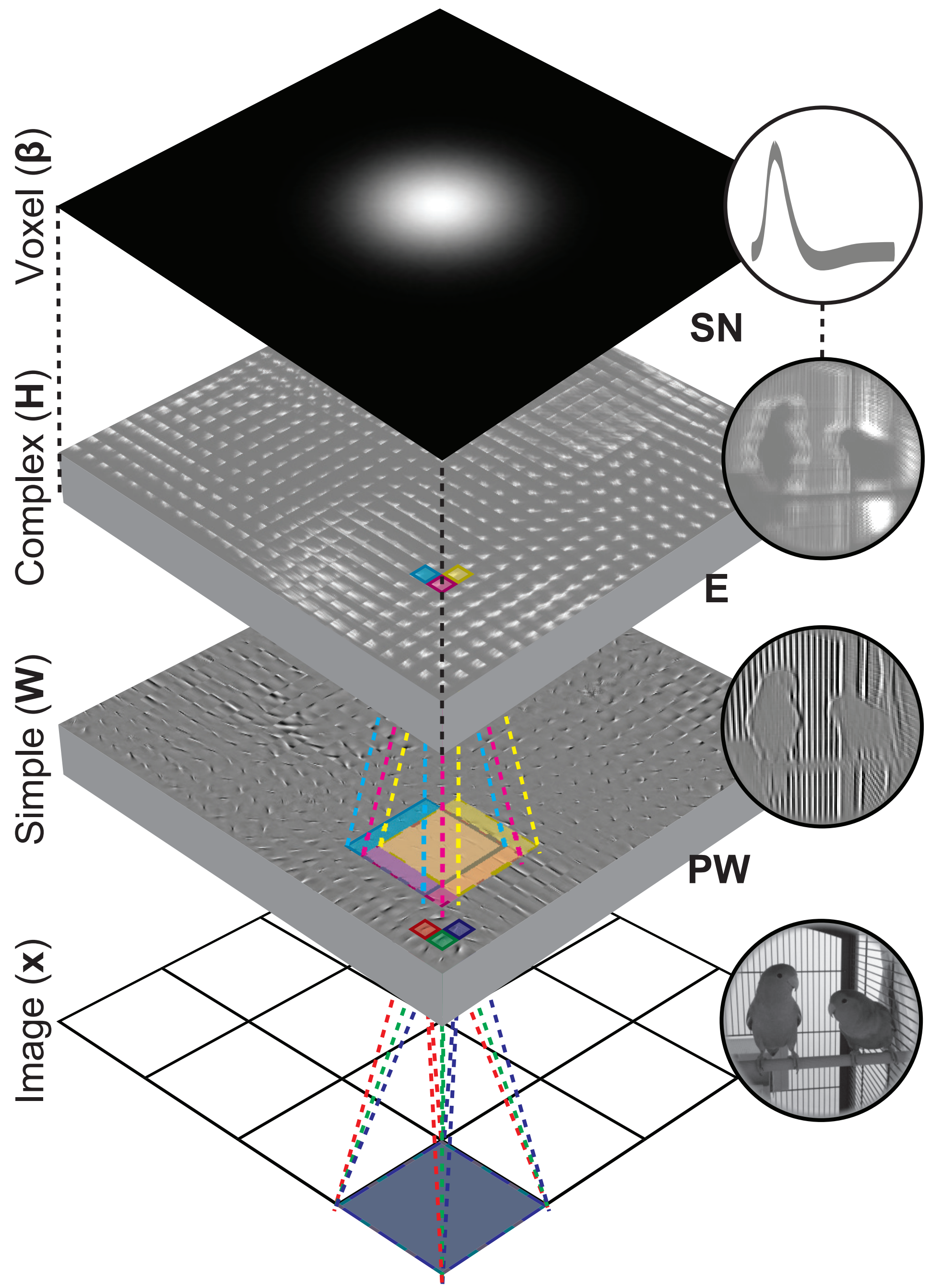
- How to extend this strategy to more complex transformations?

$\phi(\mathbf{x})$ as a result of unsupervised learning



- neurons are adapted to statistical properties of their environment
- different brain regions respond to different statistical properties
- nonlinear feature spaces improve encoding
- can we further improve results via unsupervised learning of nonlinear feature spaces?

Two-layer topographic sparse coding model



$$\mathbf{y} = \mathbf{B}^T \phi(\mathbf{x}) \text{ with } \phi(\mathbf{x}) = \phi_2(\phi_1(\mathbf{x}))$$

$$\phi_2(\mathbf{x})$$

$$\phi_1(\mathbf{x})$$

$$\mathbf{x}$$

PW = principal component analysis whitening
E = energy
SN = static nonlinearity

Güçlü, U., & van Gerven, M. A. J. (2014). Unsupervised Feature Learning Improves Prediction of Human Brain Activity in Response to Natural Images. PLoS Comp. Biol. In Press.

Simple-cell activations given by a linear transformation of whitened image patches **z**:

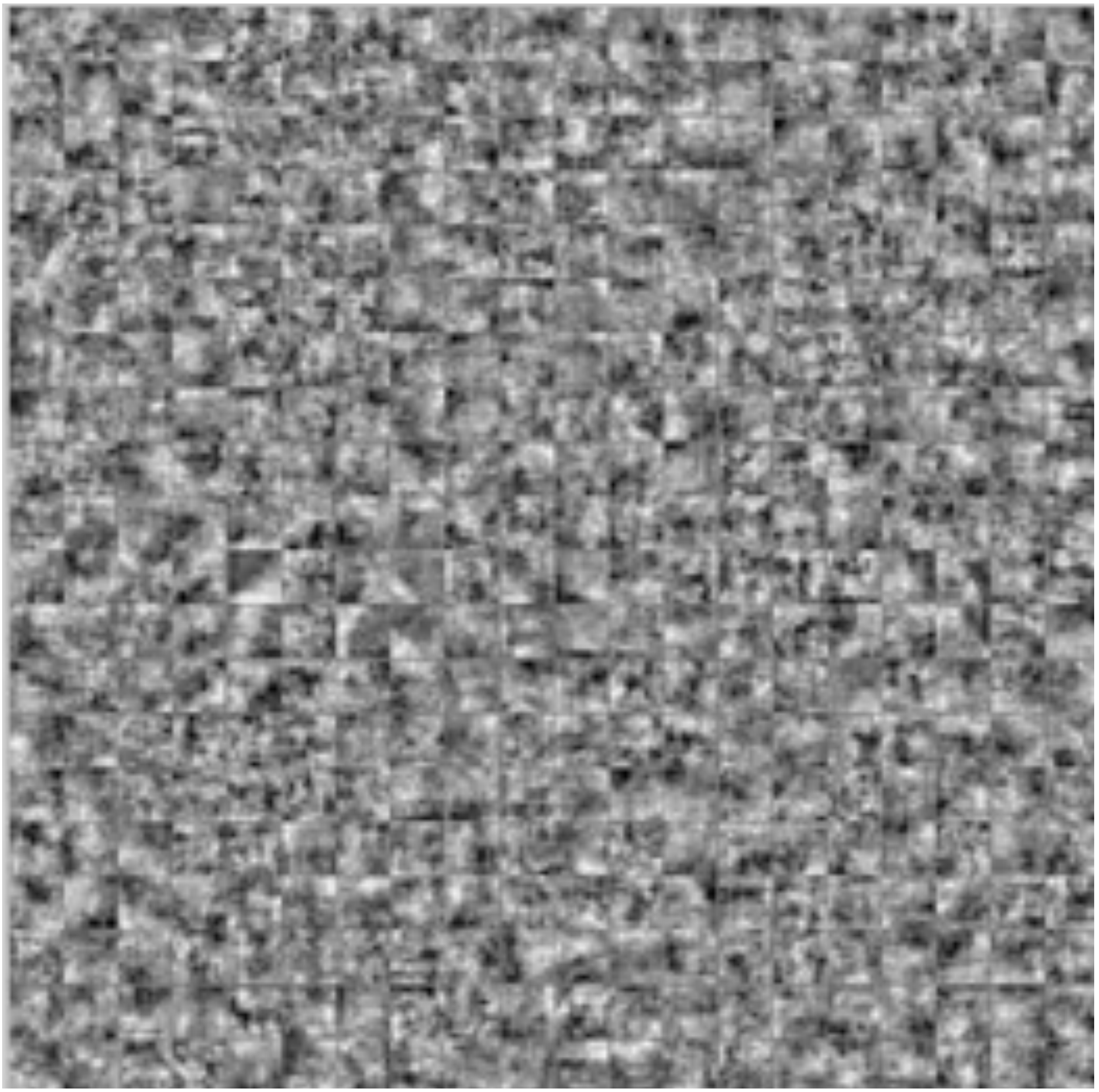
$$\phi_1(\mathbf{x}) = \mathbf{W}\mathbf{z}$$

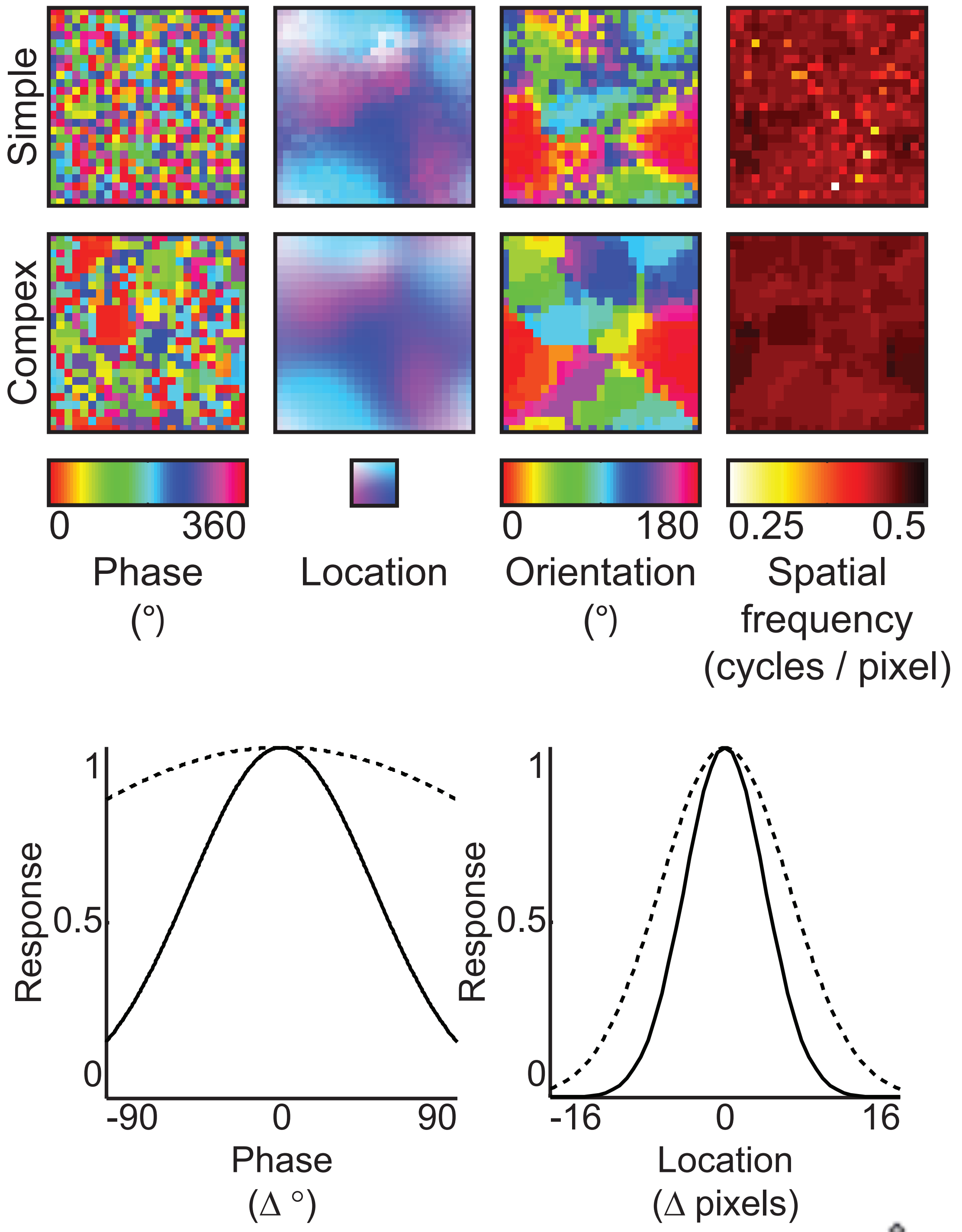
Complex-cell activations derived from the pooled energy of simple cell activations

$$\phi_2(\mathbf{s}) = \log\left(1 + \mathbf{H}\mathbf{s}^2\right)$$

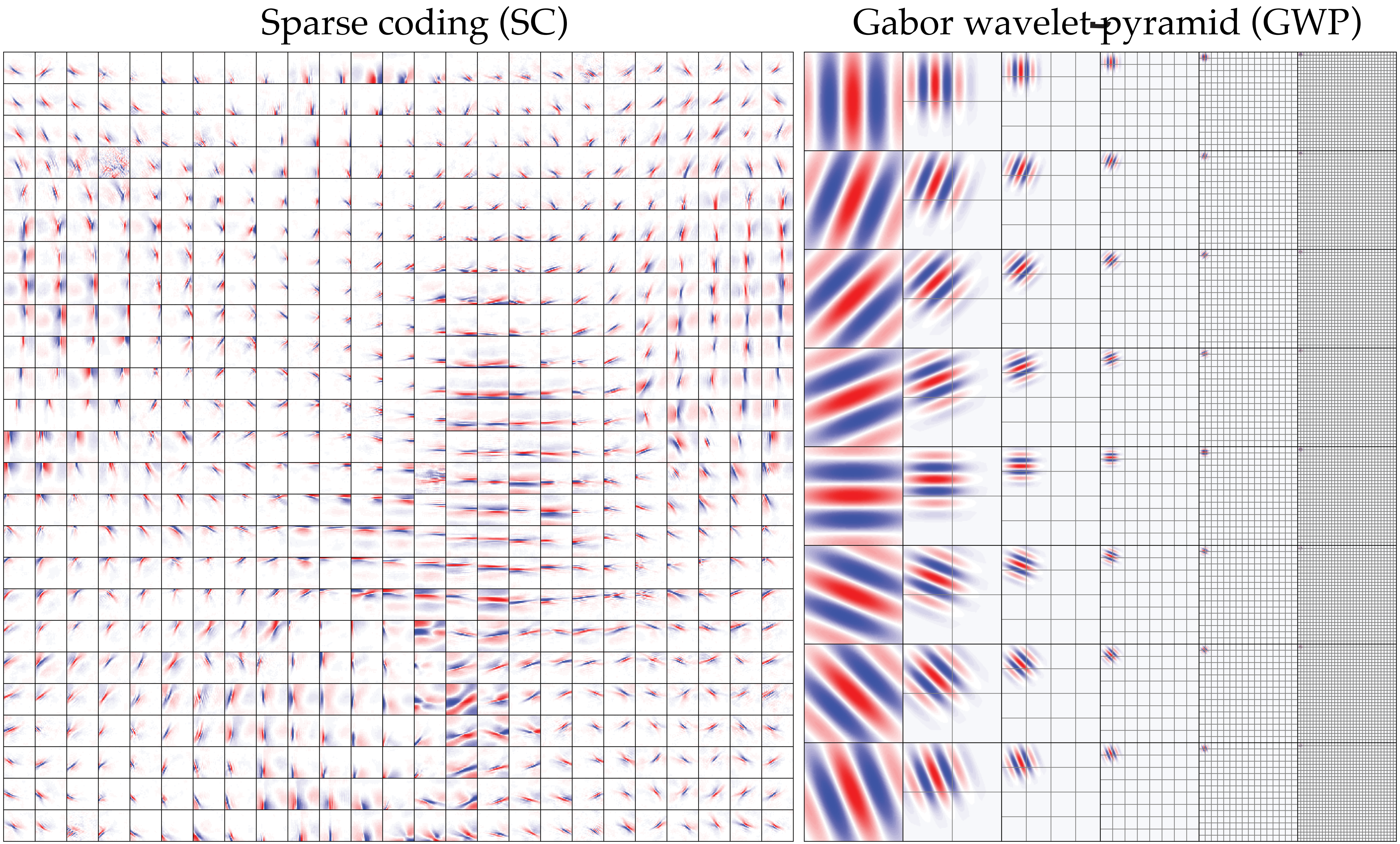
where ***H*** is a neighbourhood matrix for a square grid with circular boundary conditions.

Matrix **W** is learned using randomly sampled image patches

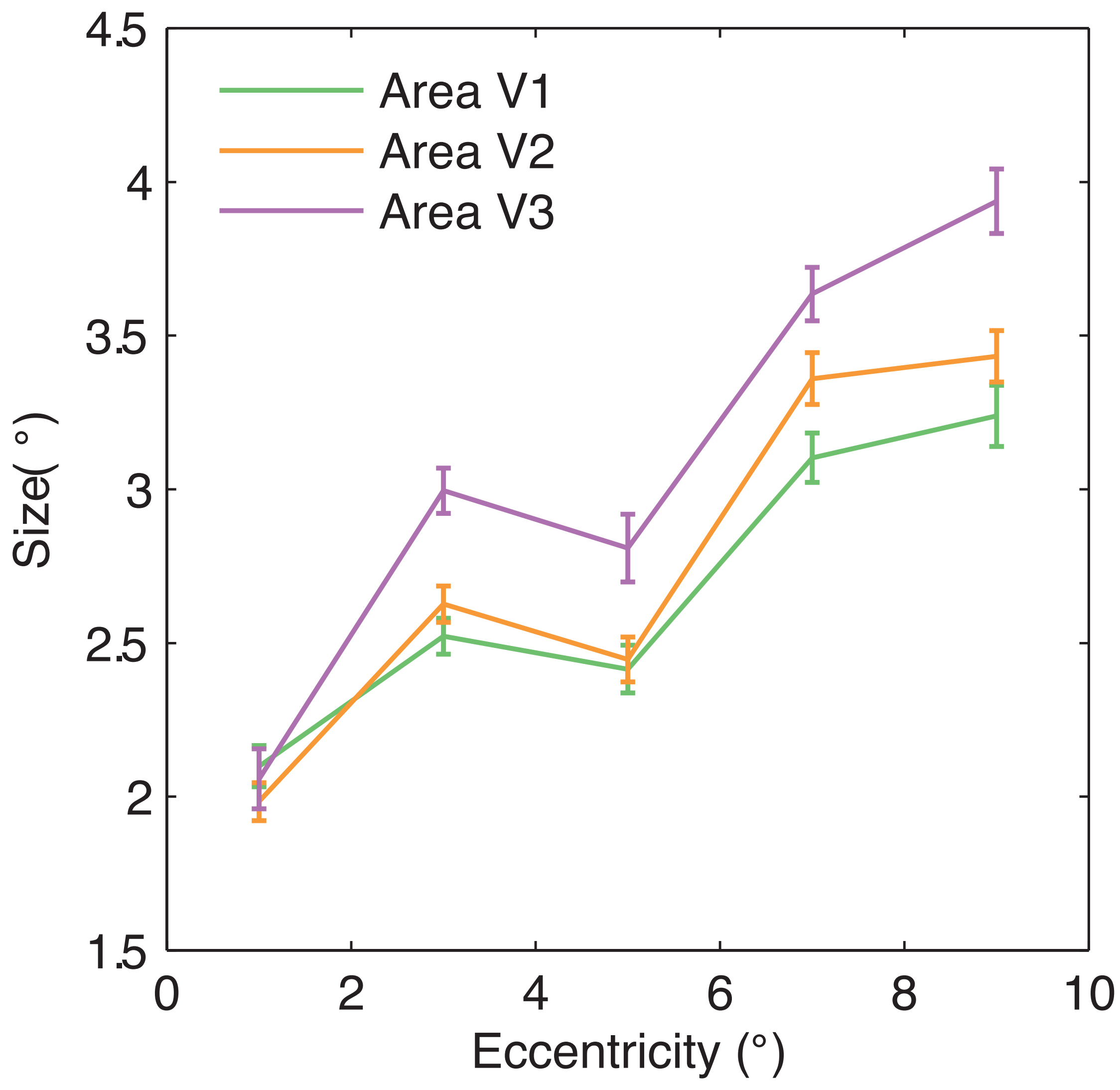


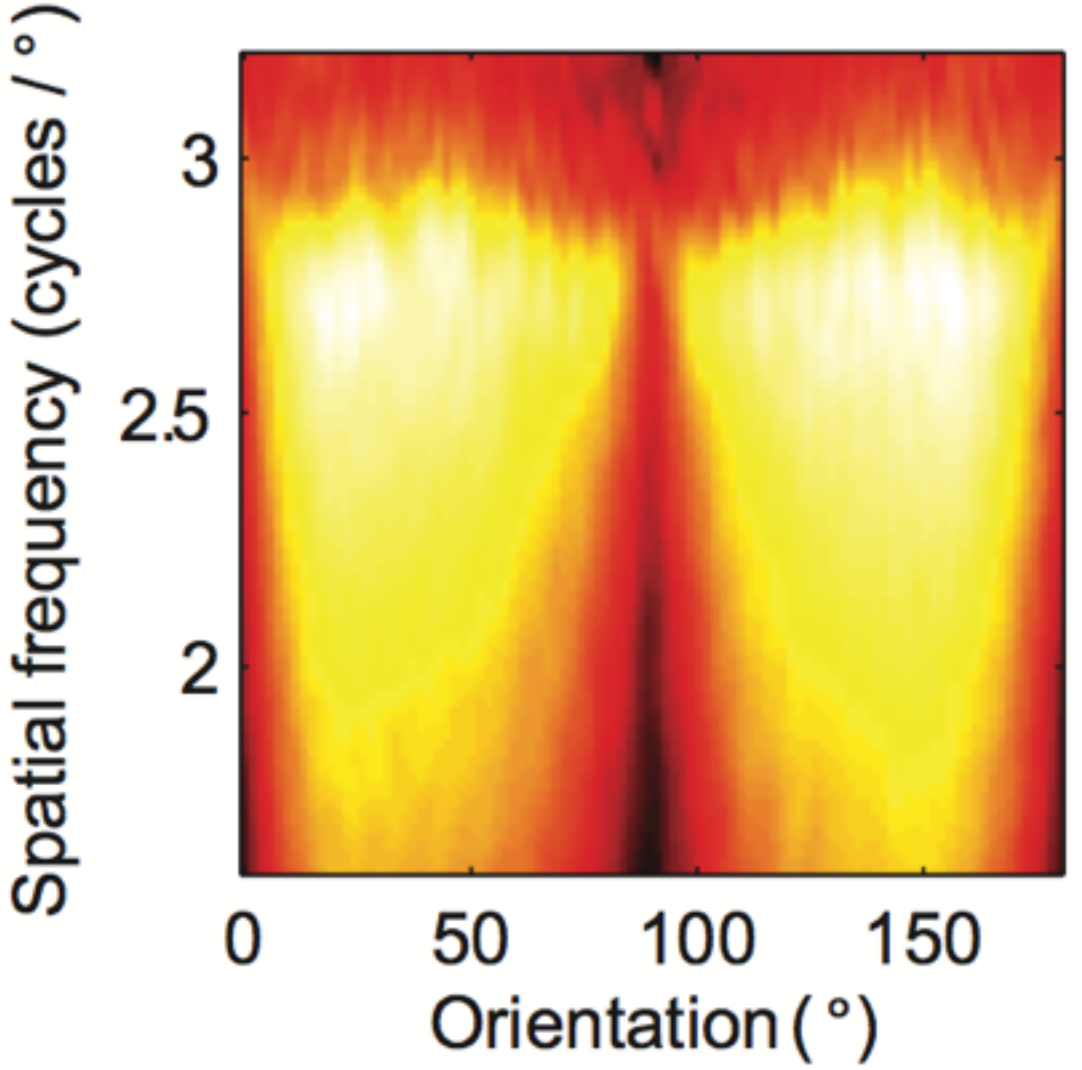


Sparse coding versus Gabor wavelet pyramid basis

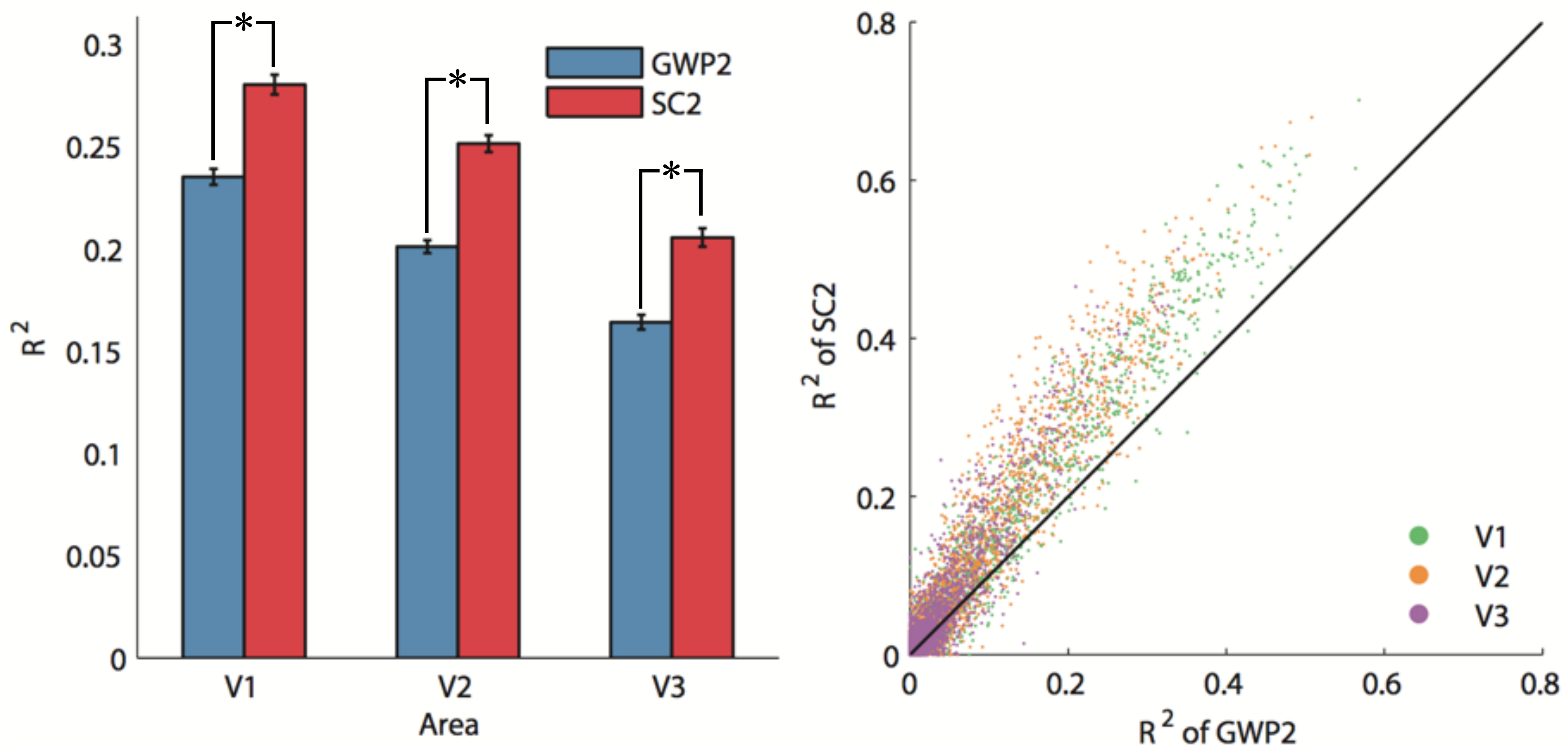


Analysis of voxel receptive fields





Encoding results



Decoding results

image identification

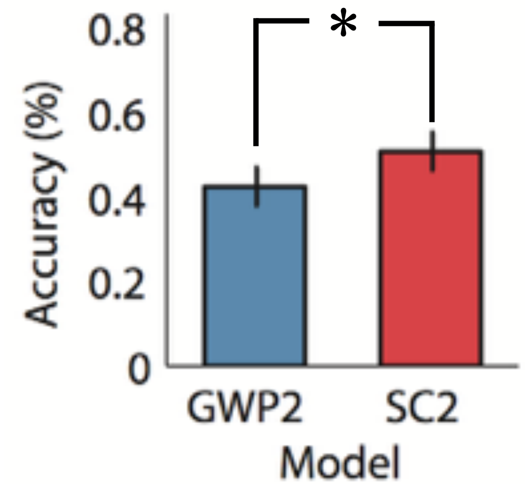
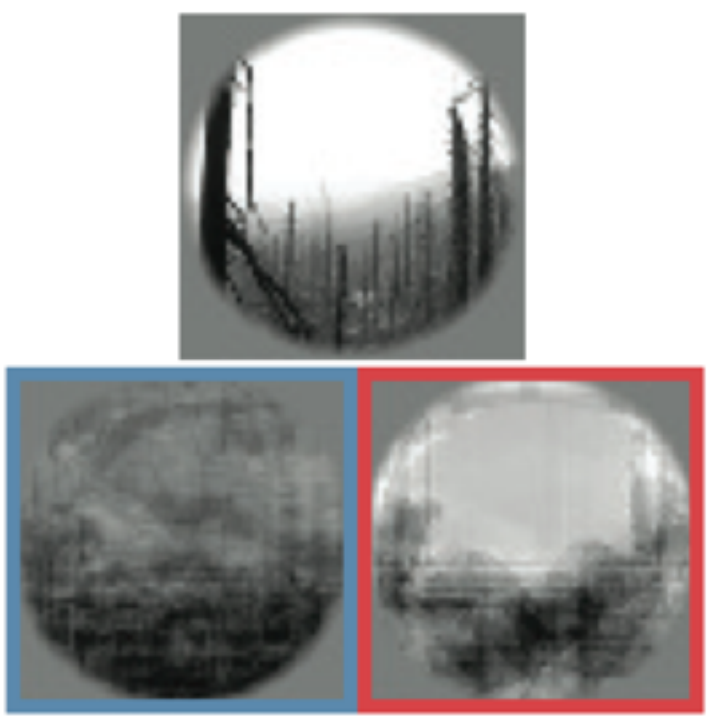
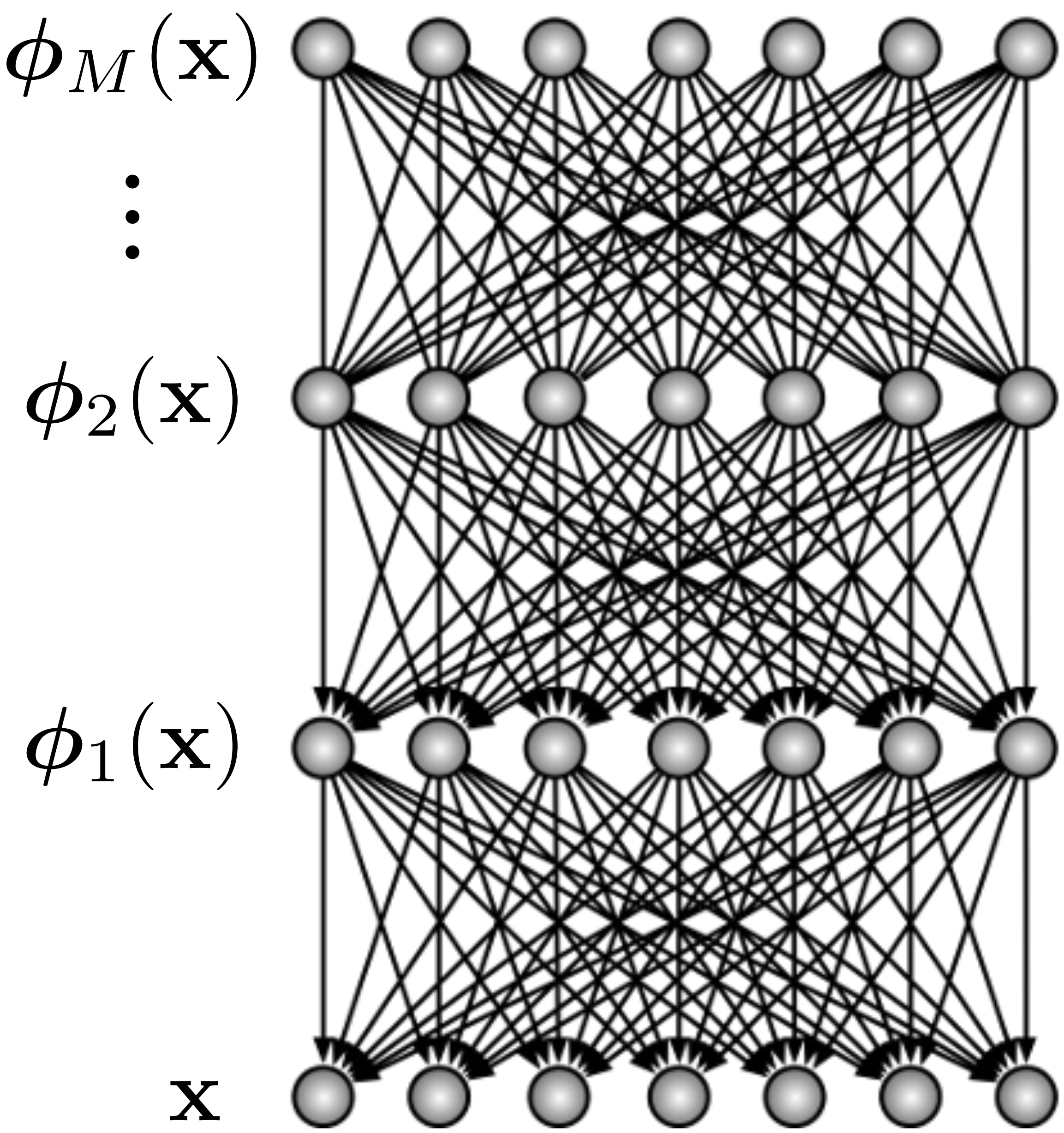


image reconstruction



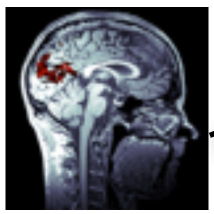


- unsupervised learning of nonlinear feature spaces provides better encoding and decoding
- deep neural networks offer a model of hierarchically structured representations in visual cortex
- deep belief networks as generative models that model complex nonlinear stimulus properties

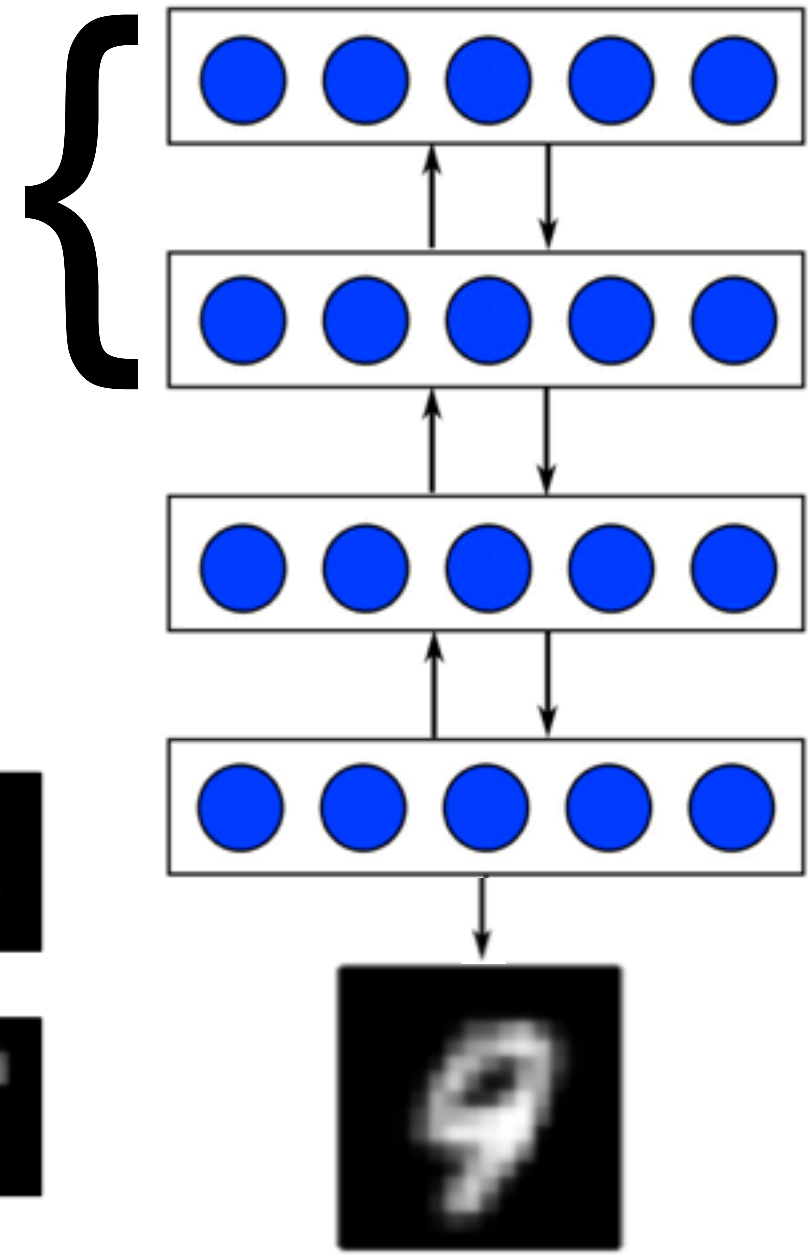
Decoding with deep belief networks

conditional restricted Boltzmann machine:

$$E(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{z}) = -\boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v} - \boldsymbol{z}^T \boldsymbol{C} \boldsymbol{v} - \boldsymbol{z}^T \boldsymbol{B} \boldsymbol{h}$$



shallow learning

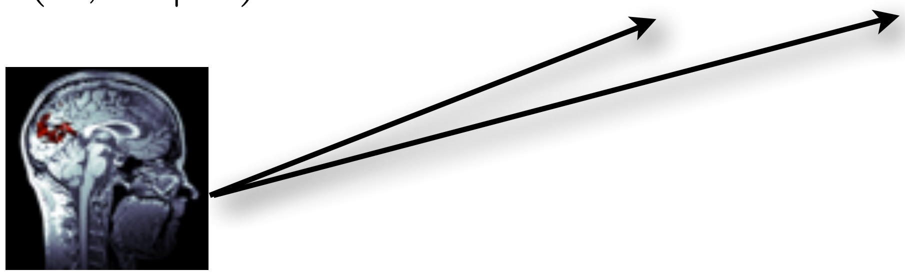


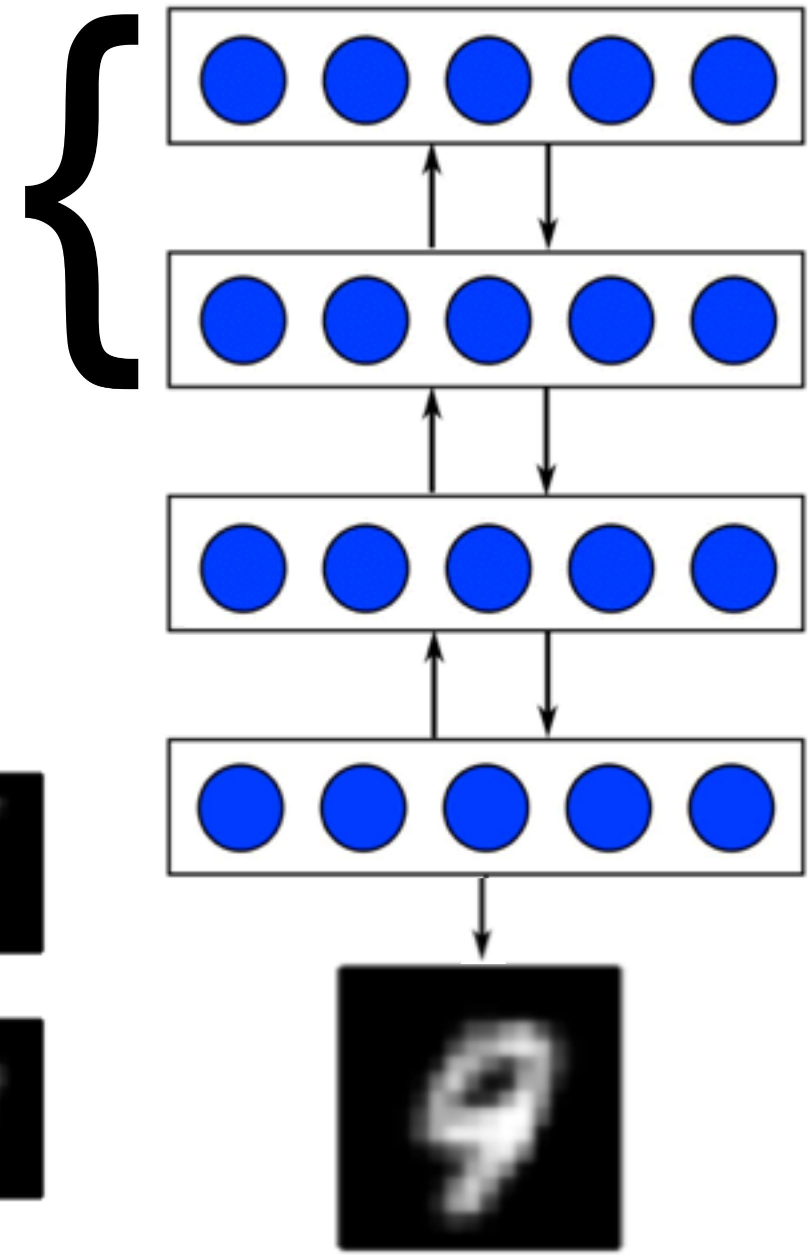
van Gerven et al. (2010). Neural decoding with hierarchical generative models. Neural Computation, 1–16

Decoding with deep belief networks

conditional restricted Boltzmann machine:

$$E(\boldsymbol{v},\boldsymbol{h} \mid \boldsymbol{z}) = -\boldsymbol{h}^T\boldsymbol{W}\boldsymbol{v} - \boldsymbol{z}^T\boldsymbol{C}\boldsymbol{v} - \boldsymbol{z}^T\boldsymbol{B}\boldsymbol{h}$$

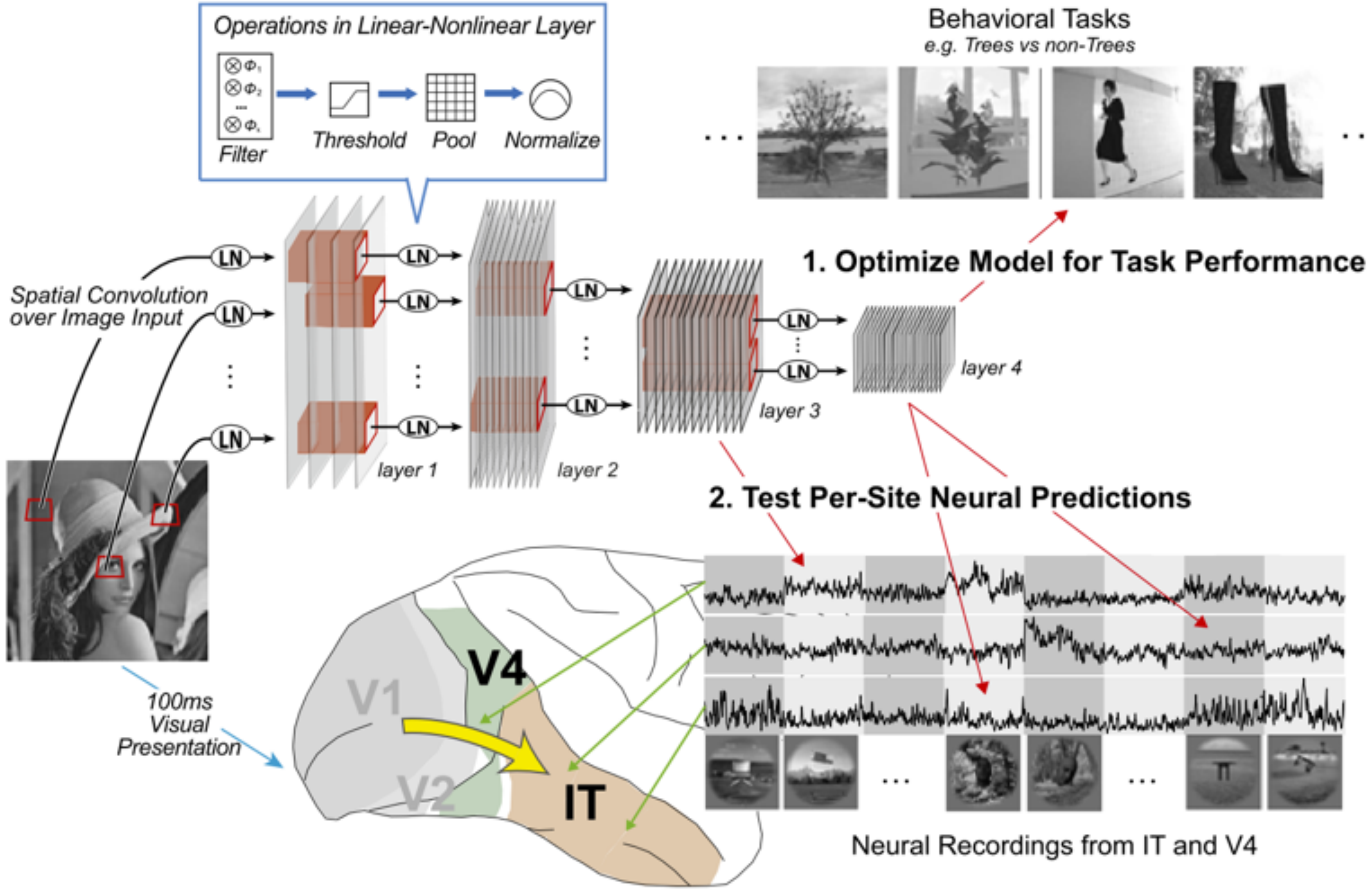




deep learning

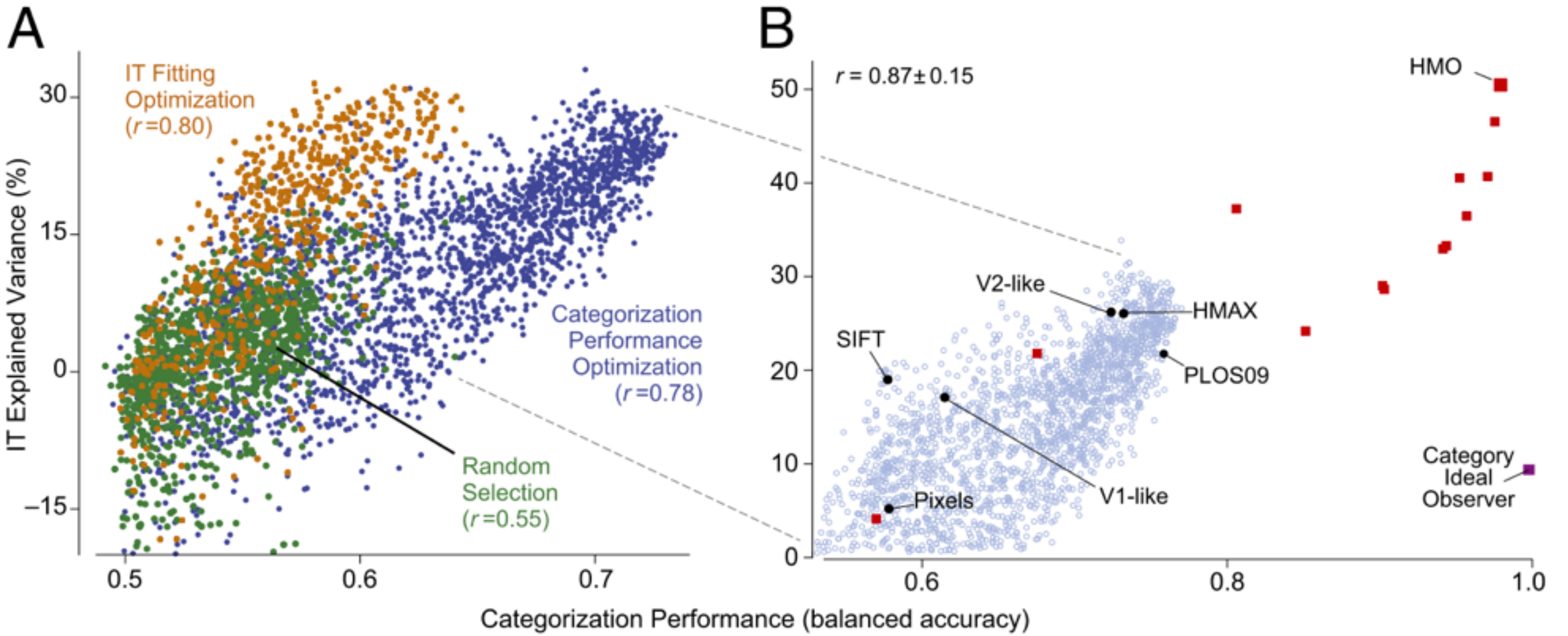
van Gerven et al. (2010). Neural decoding with hierarchical generative models. Neural Computation, 1–16

High-throughput model testing



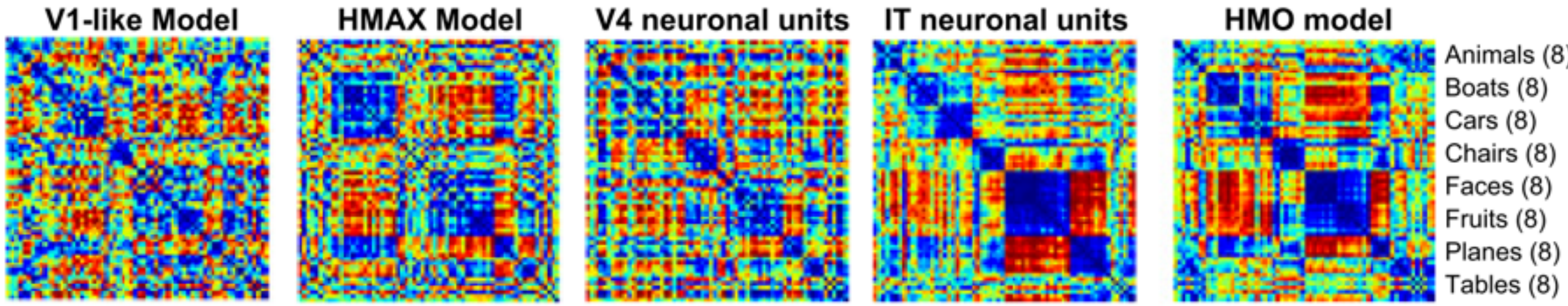
Yaminset al. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. PNAS

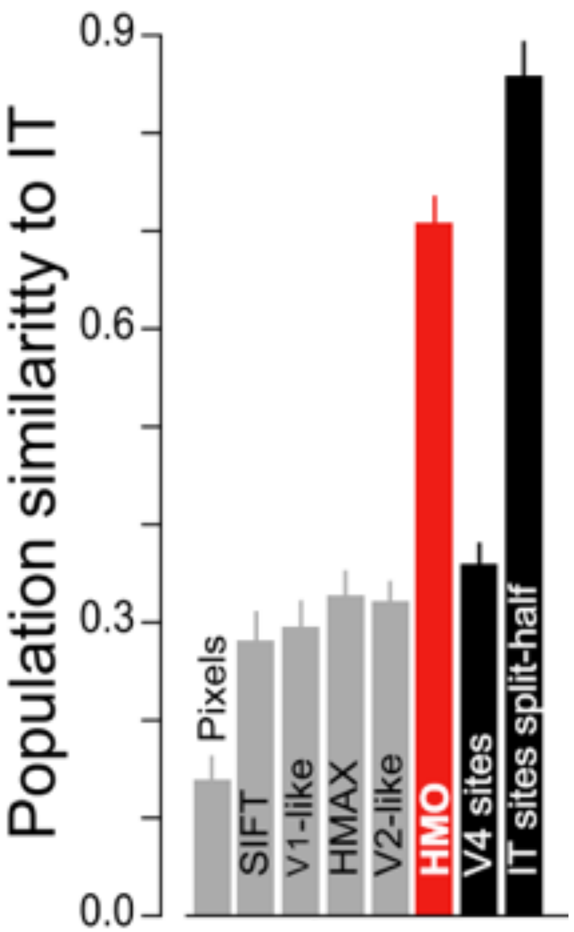
High-throughput model testing



Yaminset al. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. PNAS

Representational similarity analysis





Yaminset al. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. PNAS

Also see:
Kriegeskorte et al. (2008). Neuron; Kriegeskorte & Kievit (2013). TiCS; Pantazis et al. (2014). Nature.

Conclusions

- ➡ Discriminative approaches allow probing of representations
- ➡ Generative approaches make our assumptions explicit
- ➡ Linear Gaussian model as a baseline model for generative decoding
- ➡ Unsupervised deep learning for high-throughput analysis

Alexander Backus; Ali Bahramisharif; Markus Barth; Christian Doeller; Umut Güçlü; Peter Hagoort; Tom Heskes; Ole Jensen; Floris de Lange; Marieke van de Nieuwenhuijzen; Robert Oostenveld; Sanne Schoenmakers; Irina Simanova