Donders Institute
for Brain, Cognition and Behaviour

Brain Reading (MKI43)

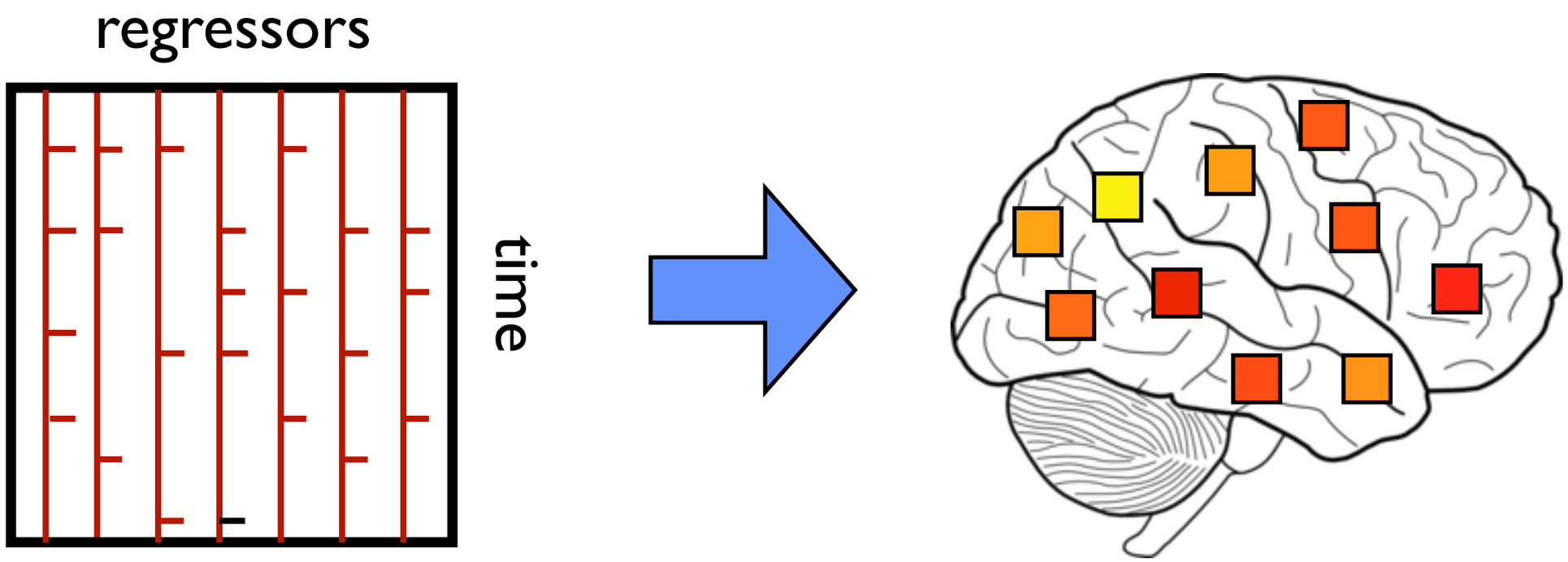
Lecture 5: Advanced generative approaches

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

Radboud University Nijmegen

Generative 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

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

Gaussian generative decoding of handwritten characters

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

We need two ingredients:

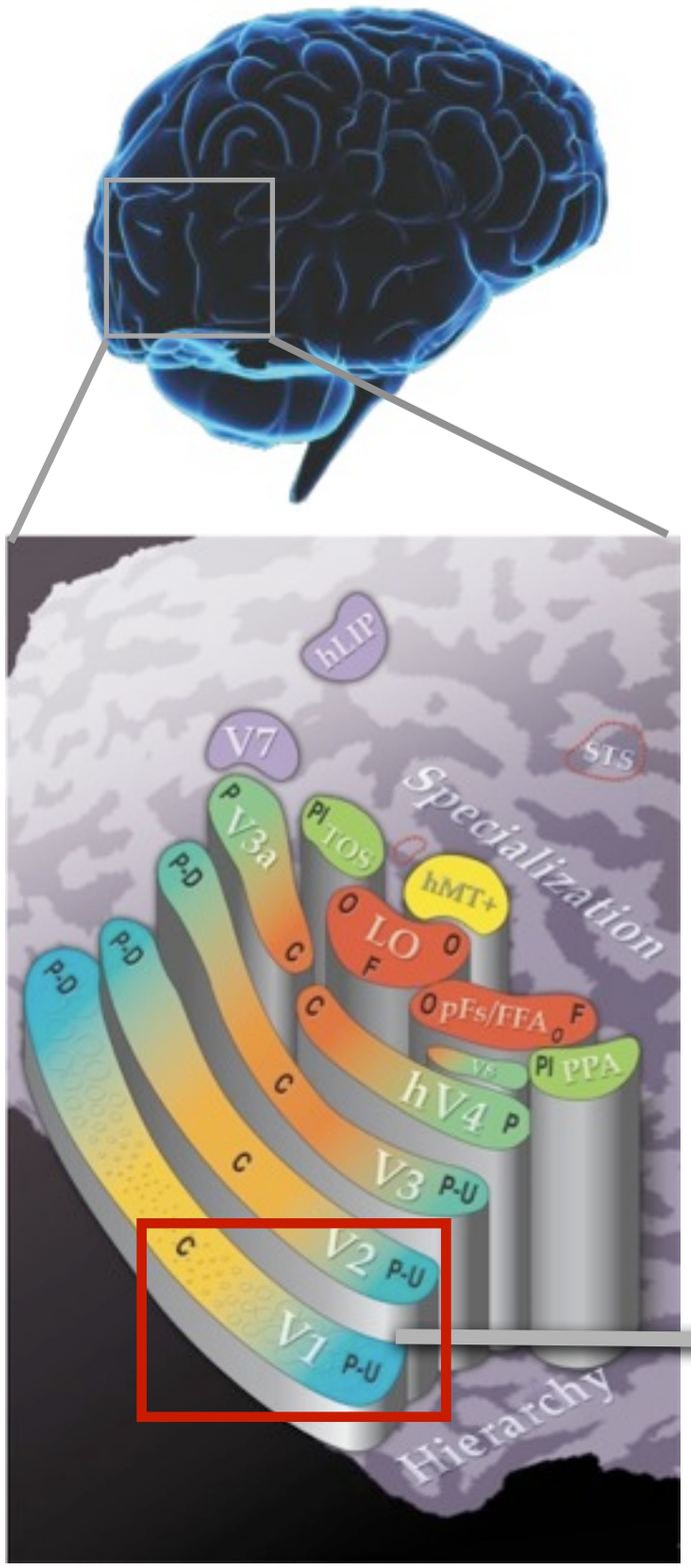
$$\text{forward model:} \quad p(\boldsymbol{y} \mid \boldsymbol{x}) \quad = \quad \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{B}^T \boldsymbol{x}, \boldsymbol{\Sigma})$$

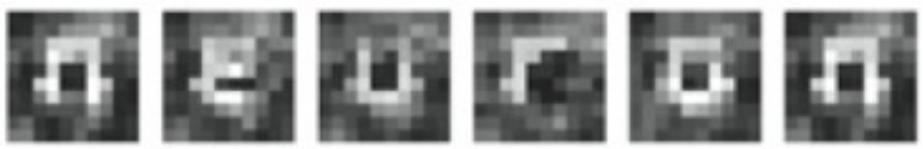
$$\text{prior:} \quad p(\boldsymbol{x}) \quad = \quad \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{0}, \boldsymbol{R})$$

Simplest option: assume everything is Gaussian!

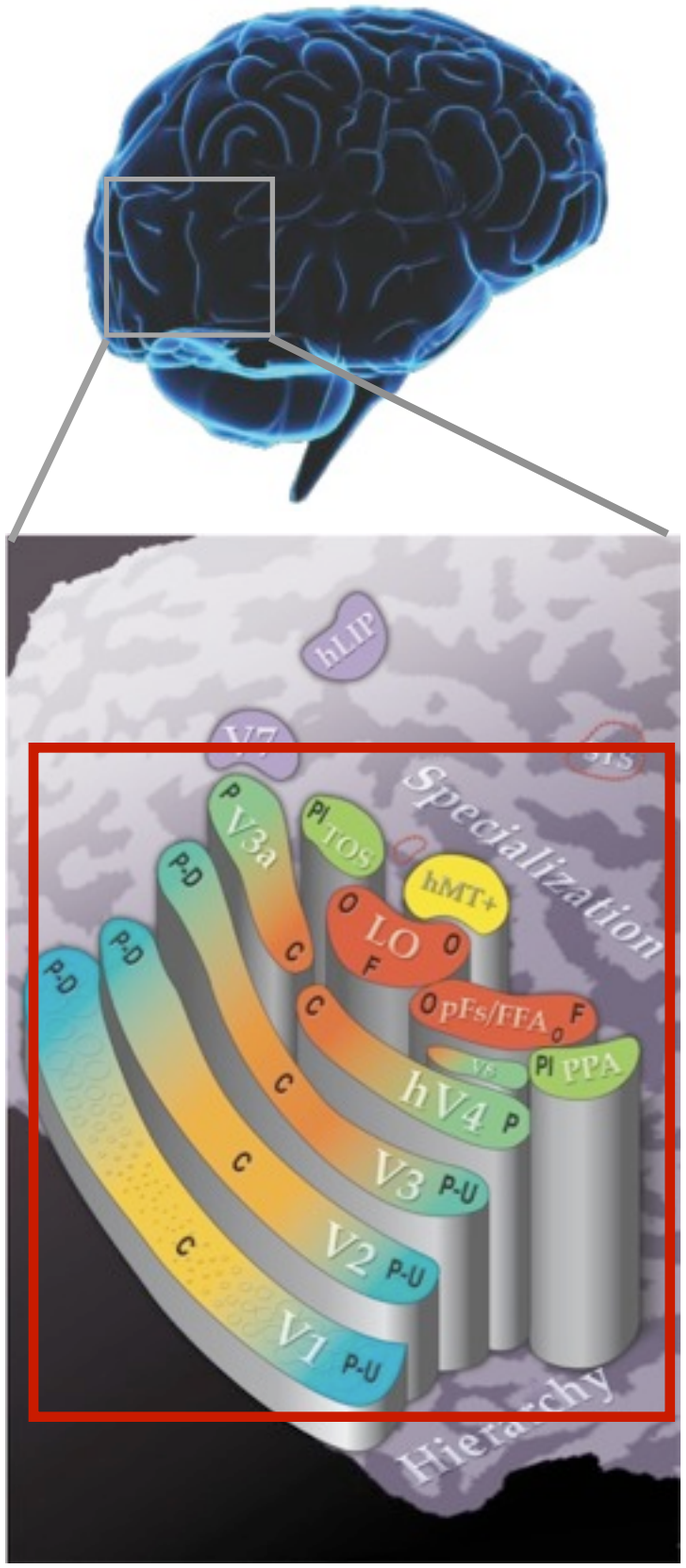
- Gaussian model is elegant but may be too simplistic
  - Can we use more realistic image priors?
  - Can we build more realistic forward models?
  - Can we be more informed by neuroscience?

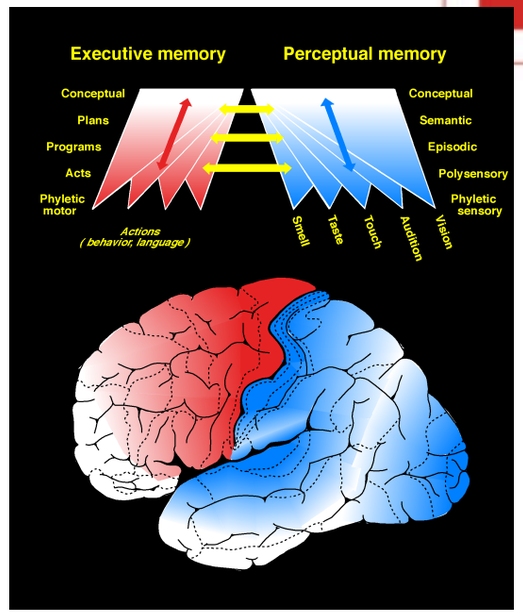
So far: flat approach to image reconstruction





Deep approach to image reconstruction





- Cortex is hierarchically organized

- Can we make use of the full cortical hierarchy?

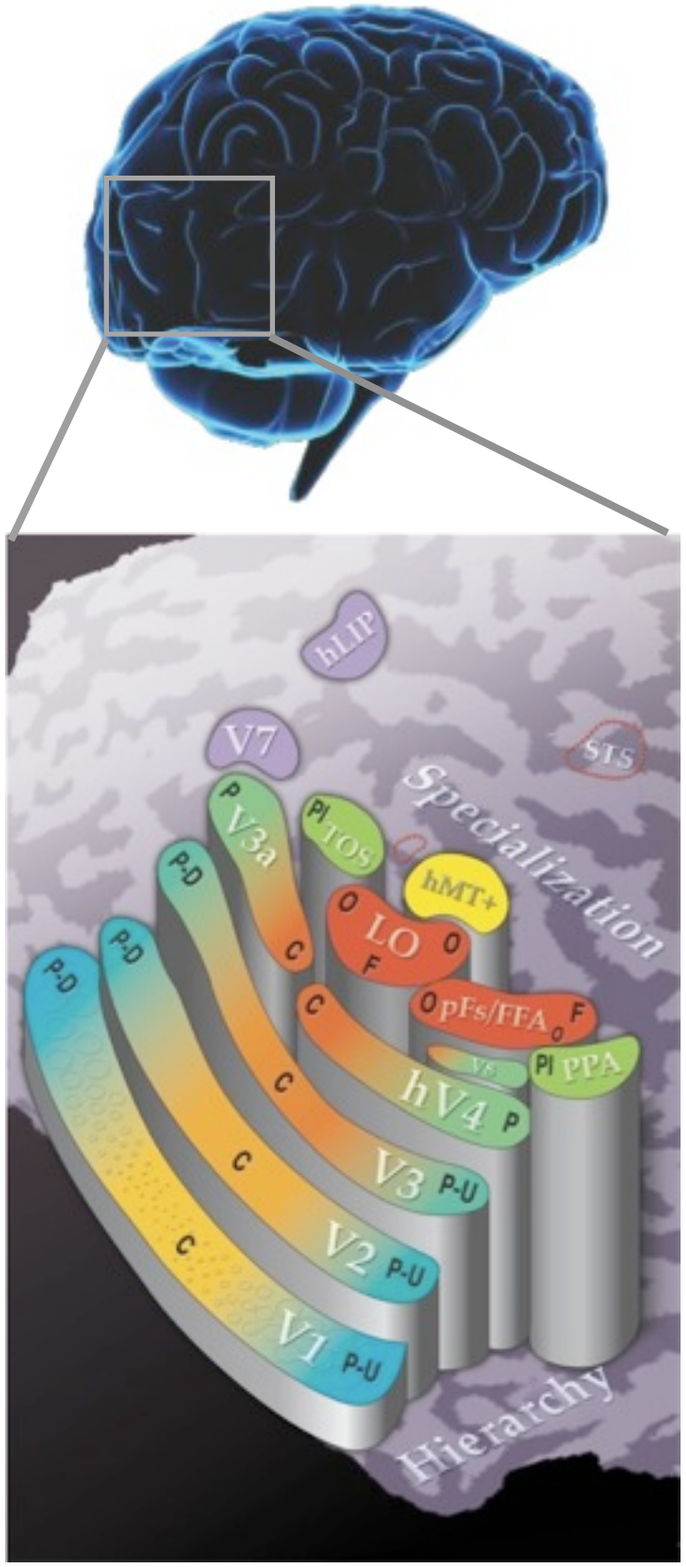
Approach 1:

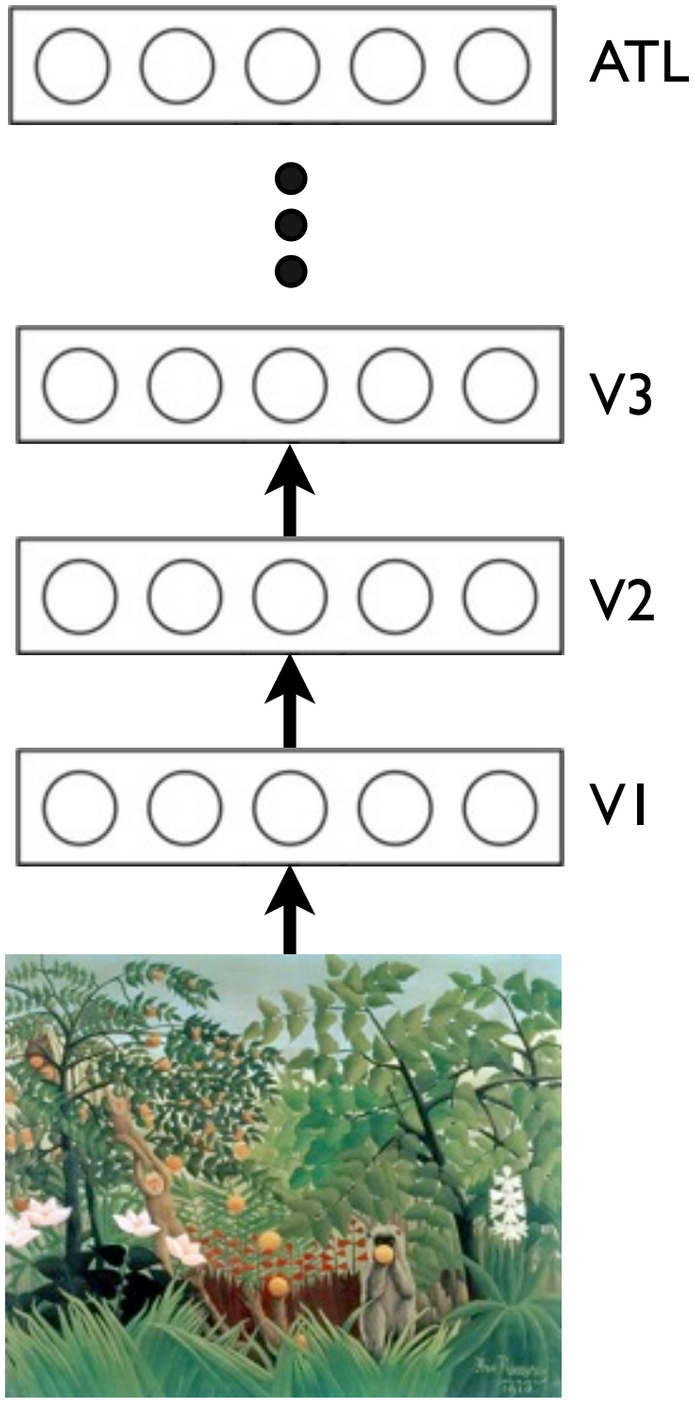
use hierarchical models that emulate the cortical hierarchy

Approach 2:

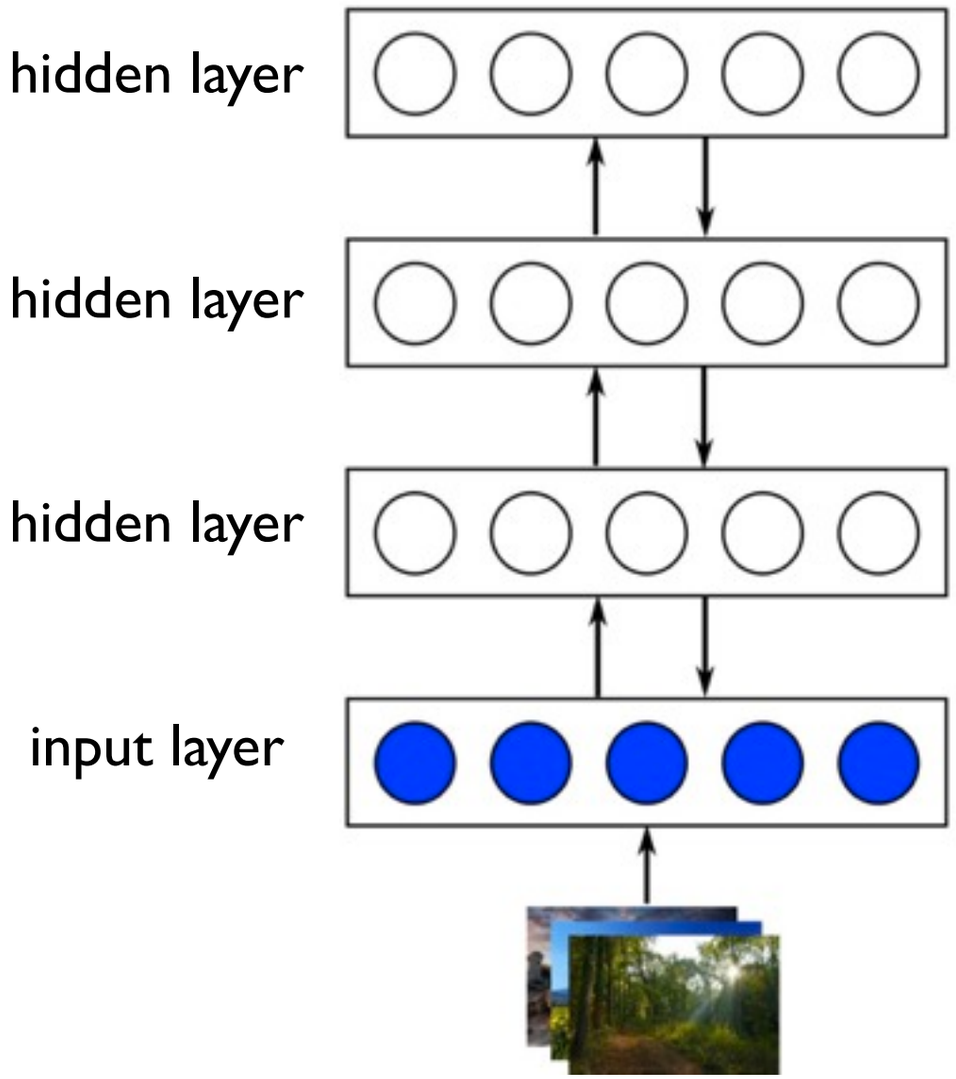
build models based on different parts of the hierarchy and combine them

Hierarchical approach



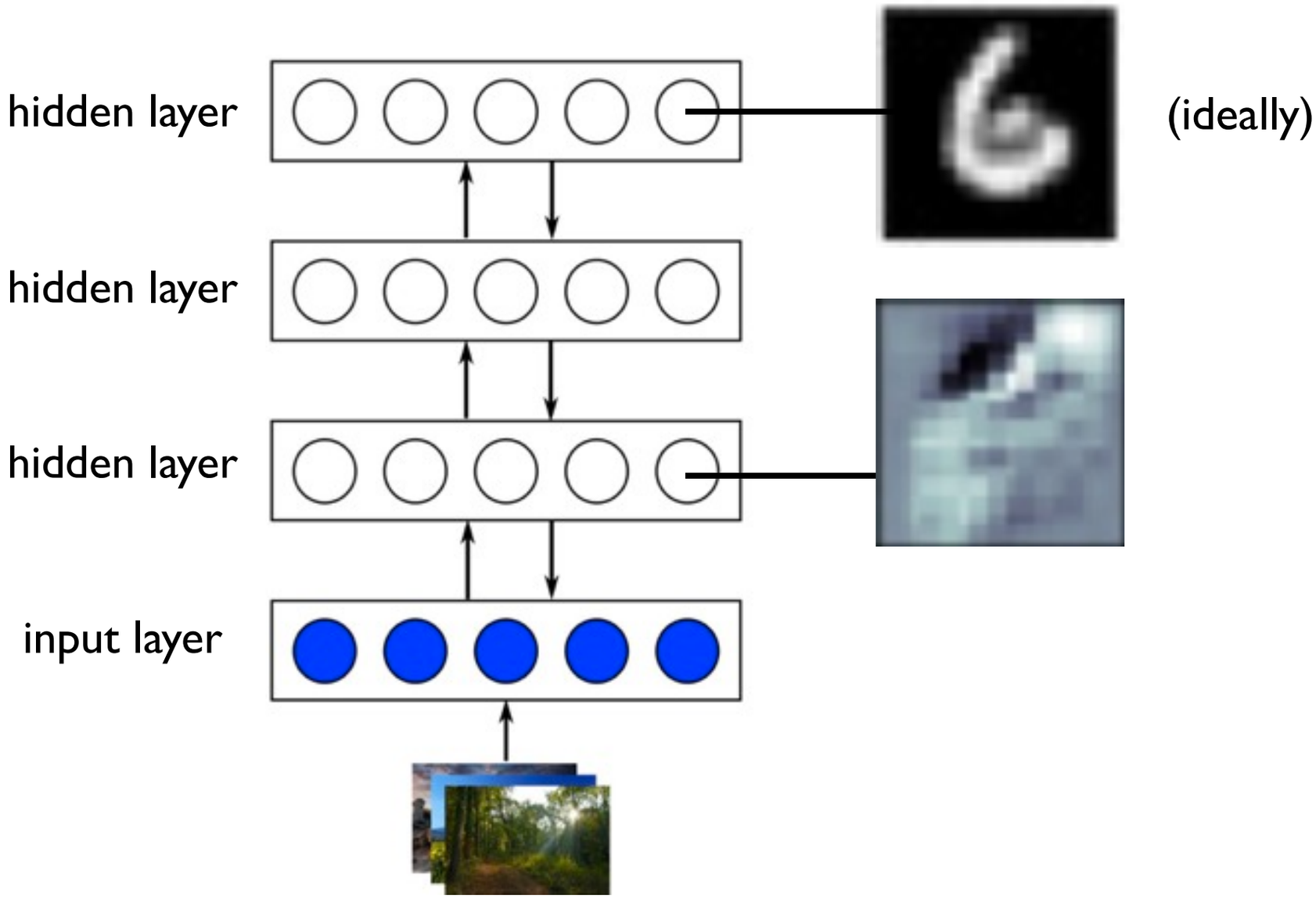


Hierarchical approach (independent of brain data!)



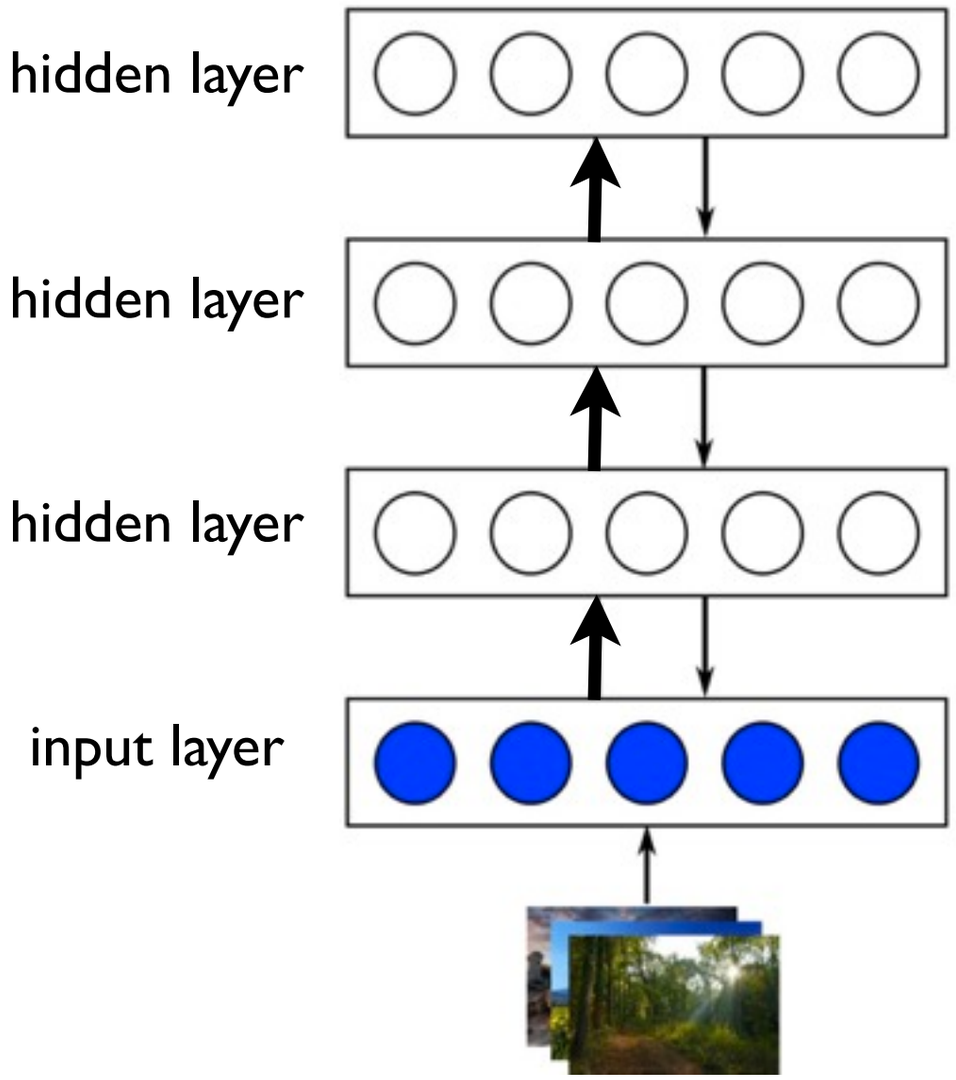
M. A. J. van Gerven, F. P. de Lange, and T. Heskes. Neural decoding with hierarchical generative models. *Neural Computation*, 2010.

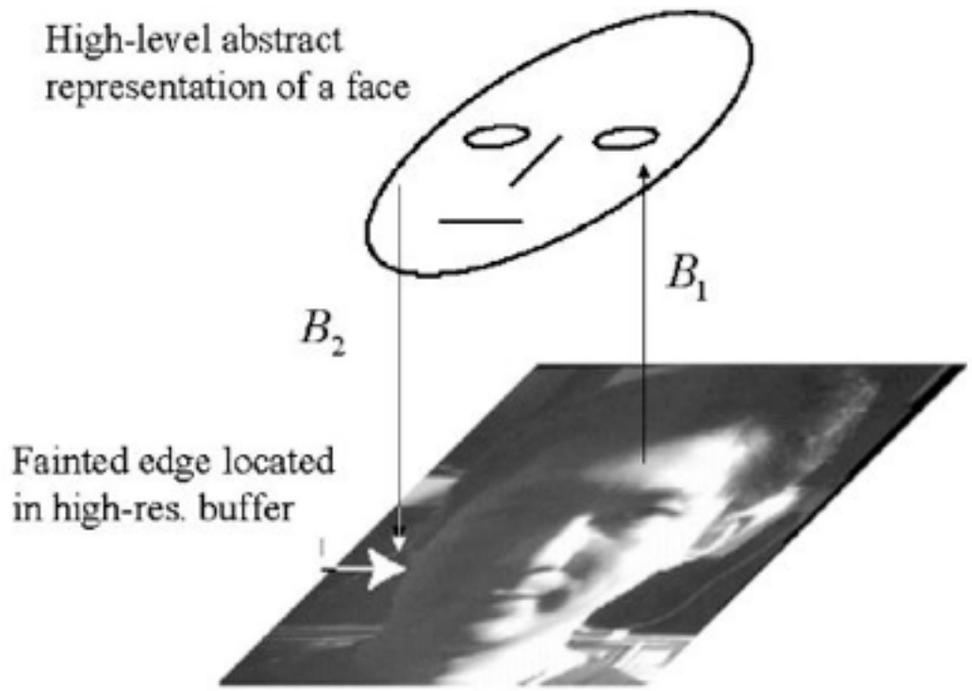
Image features



Encoding direction

How are image features activated given an input?

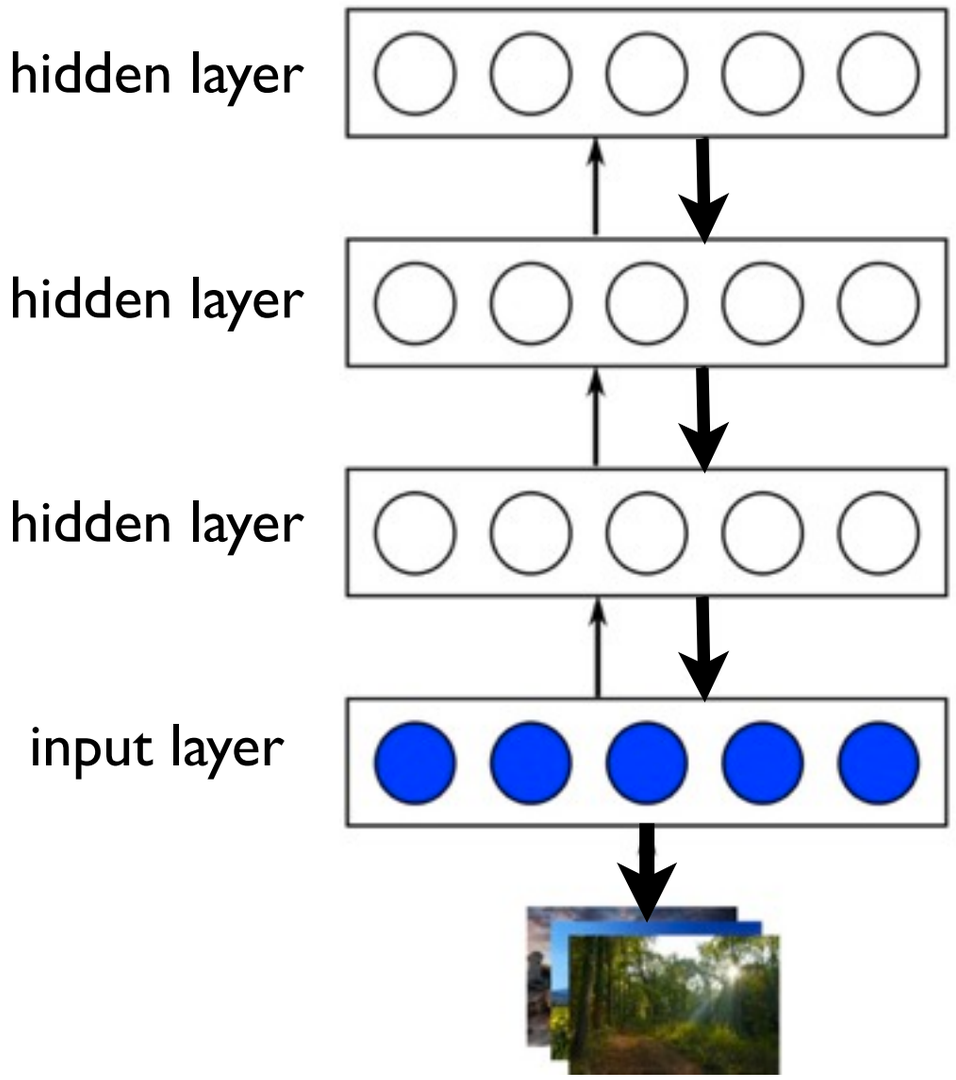


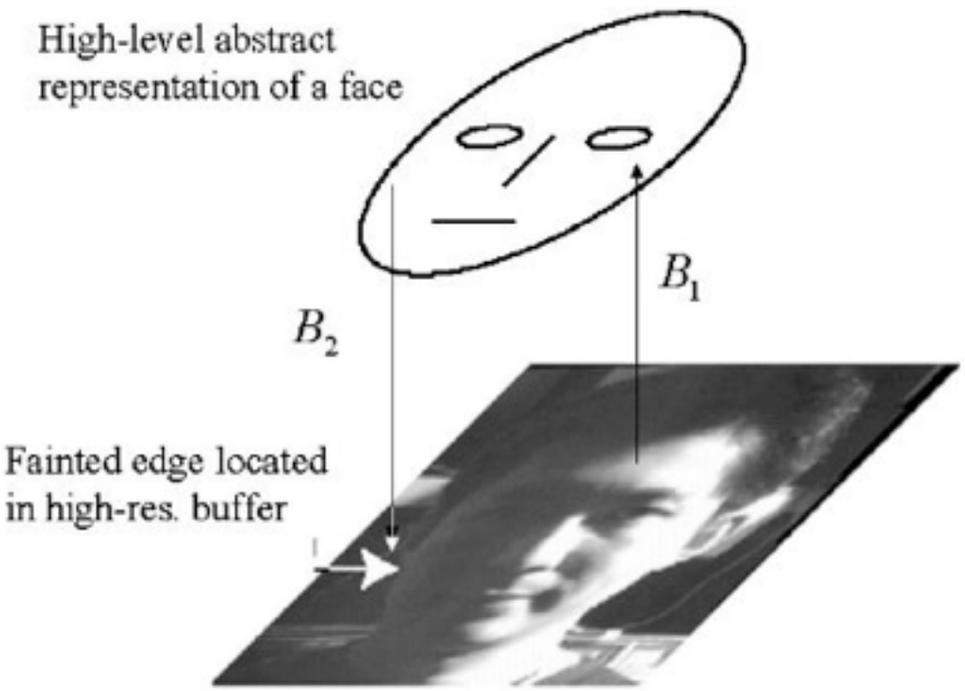


Mumford & Lee

Decoding direction

What do generated images look like?

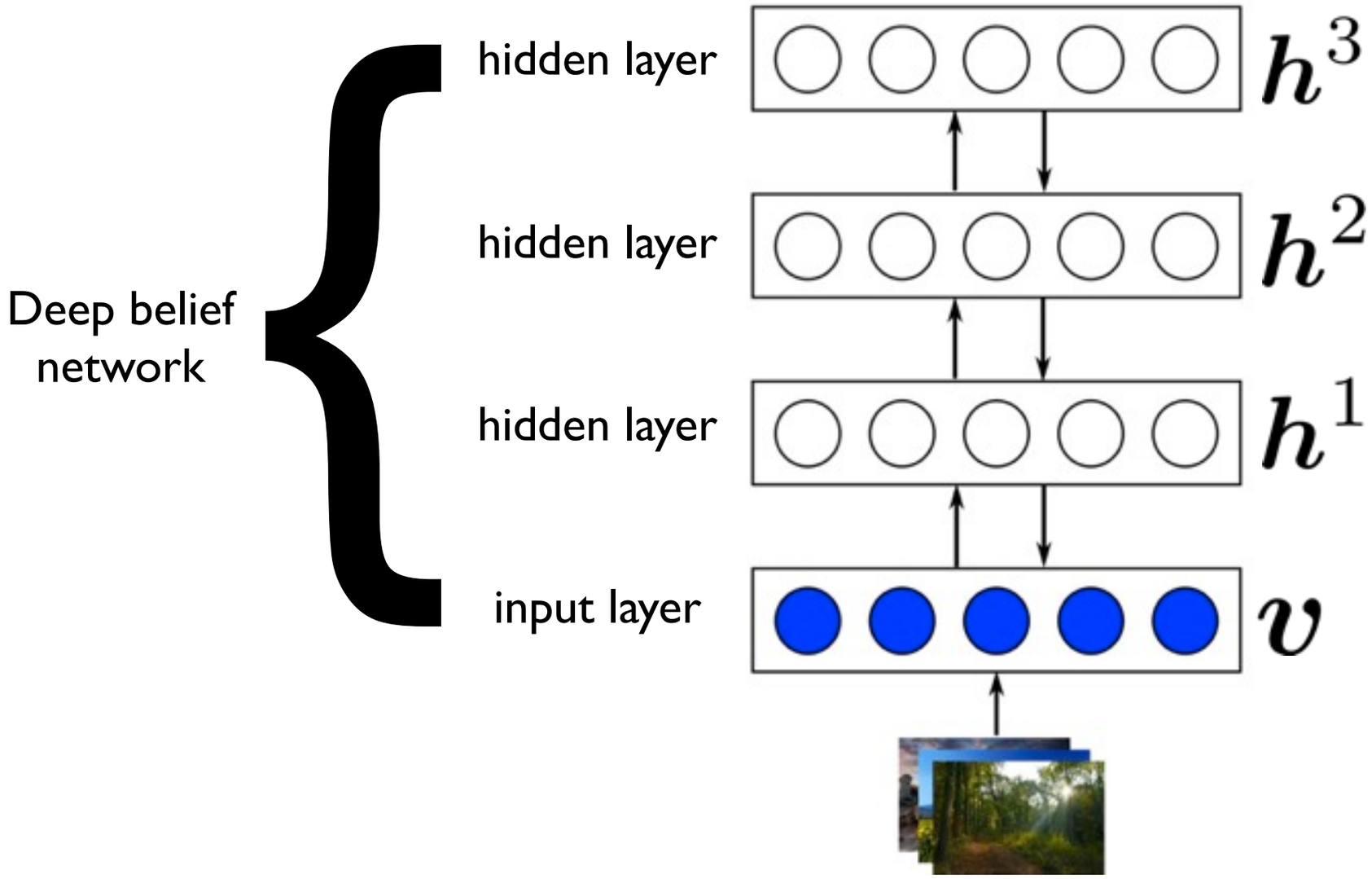




Mumford & Lee

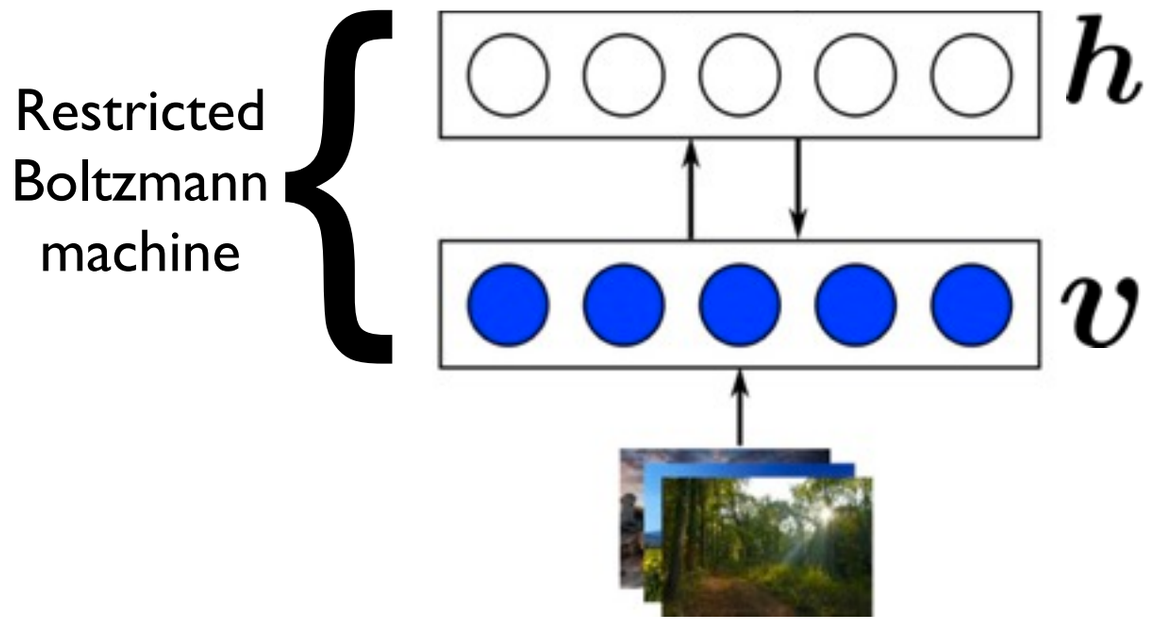
- A generative model for natural images (or text, audio, etc.)
- A very rich natural image prior P(x)

Deep belief networks (DBN)



$$P(\boldsymbol{v}, \boldsymbol{h}^1, \ldots, \boldsymbol{h}^L) = P(\boldsymbol{v} \mid \boldsymbol{h}^1) \cdots P(\boldsymbol{h}^{L-2} \mid \boldsymbol{h}^{L-1}) P(\boldsymbol{h}^{L-1}, \boldsymbol{h}^L)$$

Start with a restricted Boltzmann machine:



Restricted Boltzmann machine (RBM)

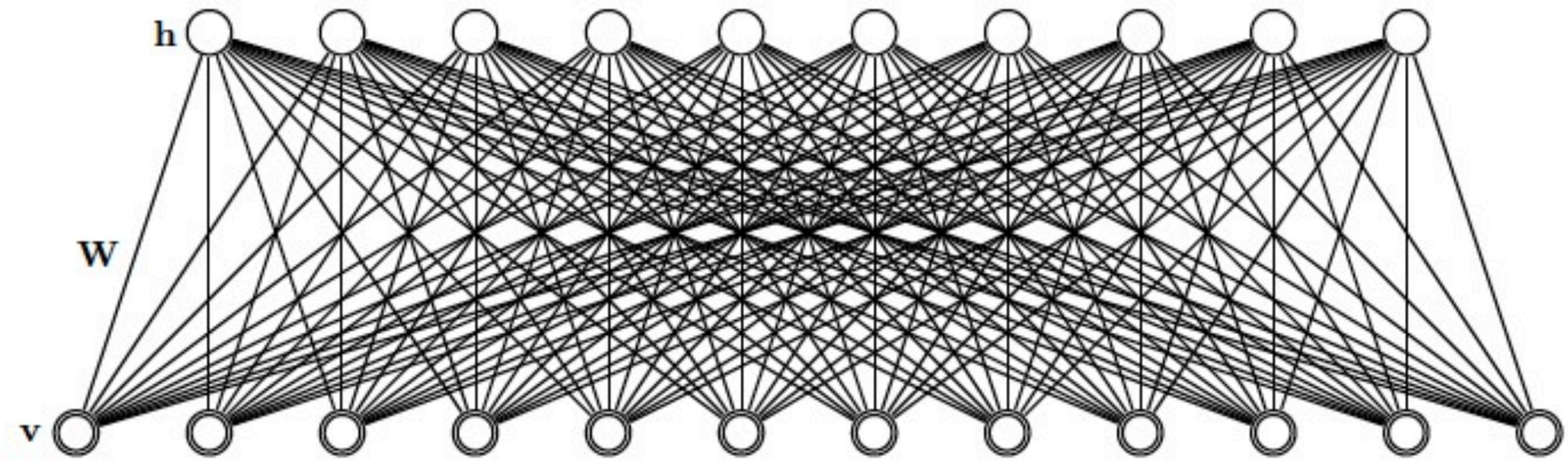


Figure 2.1: Restricted Boltzmann Machine. The top layer represents a vector of stochastic binary units **h** and the bottom layer represents a vector of stochastic binary visible variables **v**.

A bipartite Markov random field in which D visible units **v** (0 or 1) are connected to F hidden units **h** (0 or 1). A special case of the **Boltzmann machine** which is, in turn, a special kind of **energy-based model**.

PhD thesis Ruslan Salakhutdinov:
(http://www.utstat.toronto.edu/~rsalakhu/papers/Russ_thesis.pdf)

The joint distribution over **x** is defined by:

$$P(\boldsymbol{x}) = \frac{1}{Z}\exp(-E(\boldsymbol{x}))$$

where Z is the partition function: $Z = \sum_{\boldsymbol{x}} \exp(-E(\boldsymbol{x}))$

Training an EBM proceeds by maximizing log P(**x**) w.r.t the parameters by taking small steps in the direction of the gradient:

$$\frac{\partial \log P(\boldsymbol{x})}{\partial \theta} = -\frac{\partial E(\boldsymbol{x})}{\partial \theta} - \frac{\partial \log Z}{\partial \theta}$$

which can also be written as

$$\frac{\partial \log P(\boldsymbol{x})}{\partial \theta} = -\frac{\partial E(\boldsymbol{x})}{\partial \theta} + \sum_{\tilde{\boldsymbol{x}}} P(\tilde{\boldsymbol{x}}) \frac{\partial E(\tilde{\boldsymbol{x}})}{\partial \theta}$$

In many interesting cases, we have **x** = (**v**,**h**) where **v** is observed and **h** is hidden (e.g. images versus their causes).

In that case, we only care about the marginal w.r.t. **v**:

$$P(\boldsymbol{v}) = \frac{1}{Z} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}))$$

which we can map to the previous notation by writing

$$P(\boldsymbol{v}) = \frac{1}{Z} \exp(-F(\boldsymbol{v}))$$

with **free energy**

$$F(\boldsymbol{v}) = -\log \sum_{\boldsymbol{h}} \exp\left(-E(\boldsymbol{v}, \boldsymbol{h})\right)$$

By plugging this into the gradient, we obtain

$$\frac{\partial \log P(\boldsymbol{v})}{\partial \theta} = -\frac{\partial F(\boldsymbol{v})}{\partial \theta} + \sum_{\tilde{\boldsymbol{v}}} P(\tilde{\boldsymbol{v}}) \frac{\partial F(\tilde{\boldsymbol{v}})}{\partial \theta}$$

The average log-likelihood gradient over training samples is

$$\sum_{\boldsymbol{v}} \hat{P}(\boldsymbol{v}) \frac{\partial \log P(\boldsymbol{v})}{\partial \theta} = -E_{\hat{P}}\left(\frac{\partial F(\boldsymbol{v})}{\partial \theta}\right) + E_P\left(\frac{\partial F(\tilde{\boldsymbol{v}})}{\partial \theta}\right)$$

where $E_{\hat{P}}$ is the expectation under the empirical distribution $\hat{P}(\boldsymbol{v})$

and E_P is the expectation under the true distribution $P(\boldsymbol{v})$

If we can sample from P then we can get an approximation of the average log likelihood gradient and train the EBM using gradient ascent.

In an RBM, the energy of a state (**v**,**h**) is given by:

$$E(\boldsymbol{v}, \boldsymbol{h}) \quad = \quad -\boldsymbol{b}^T \boldsymbol{v} - \boldsymbol{c}^T \boldsymbol{h} - \boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v}$$

where $\theta = \{W, \mathbf{b}, \mathbf{a}\}$ are the model parameters we need to estimate.

It can be shown that the required gradient is given by

$$\frac{\partial \log P(\boldsymbol{v})}{\partial \theta} \quad = \quad -\sum_{\boldsymbol{h}} P(\boldsymbol{h} \mid \boldsymbol{v}) \frac{\partial E(\boldsymbol{v}, \boldsymbol{h})}{\partial \theta} + \sum_{\tilde{\boldsymbol{v}}, \boldsymbol{h}} P(\tilde{\boldsymbol{v}}, \boldsymbol{h}) \frac{\partial E(\tilde{\boldsymbol{v}}, \boldsymbol{h})}{\partial \theta}$$

The partial derivative w.r.t. the energy is easy to compute so we only need to be able to sample from the conditional and joint probability distributions.

Sampling P(**h**|**v**)

Sampling P(**h**|**v**) in an RBM is easy since

$$P(\boldsymbol{h} \mid \boldsymbol{v}) = \prod_i P(h_i \mid \boldsymbol{v}) = \prod_i \frac{1}{1 + \exp(-c_i + W_i \boldsymbol{v})}$$

Likewise

$$P(\boldsymbol{v} \mid \boldsymbol{h}) = \prod_i P(v_i \mid \boldsymbol{h}) = \prod_i \frac{1}{1 + \exp(-b_i + W_{.i}^T \boldsymbol{h})}$$

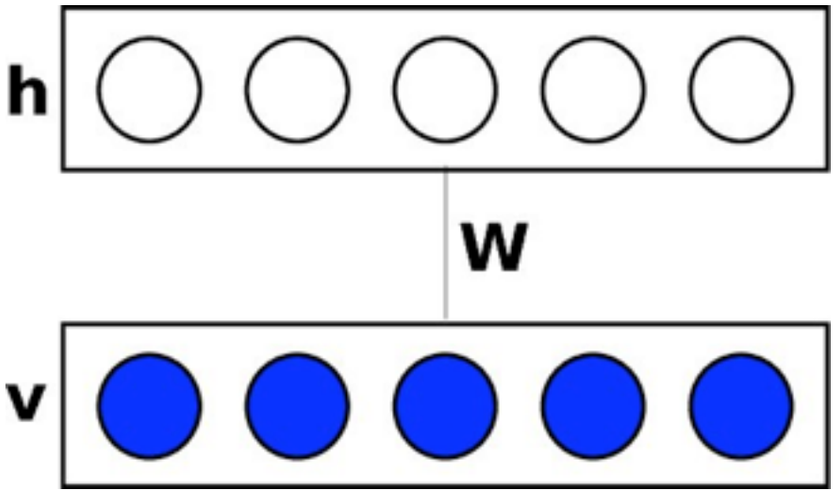
For P(**v**,**h**) we make use of **Gibbs sampling** where we alternately update

$$
\begin{aligned}
\boldsymbol{v}^1 &\sim \hat{P}(\boldsymbol{v}) \\
\boldsymbol{h}^1 &\sim P(\boldsymbol{h} \mid \boldsymbol{v}^1) \\
\boldsymbol{v}^2 &\sim P(\boldsymbol{v} \mid \boldsymbol{h}^1) \\
\boldsymbol{h}^2 &\sim P(\boldsymbol{h} \mid \boldsymbol{v}^2) \\
&\vdots \\
\boldsymbol{v}^{k+1} &\sim P(\boldsymbol{v} \mid \boldsymbol{h}^k)
\end{aligned}
$$

We make use of the **contrastive divergence** algorithm, which simply means that k is small.

Restricted Boltzmann machine

contrastive divergence, Hinton et al. 2006

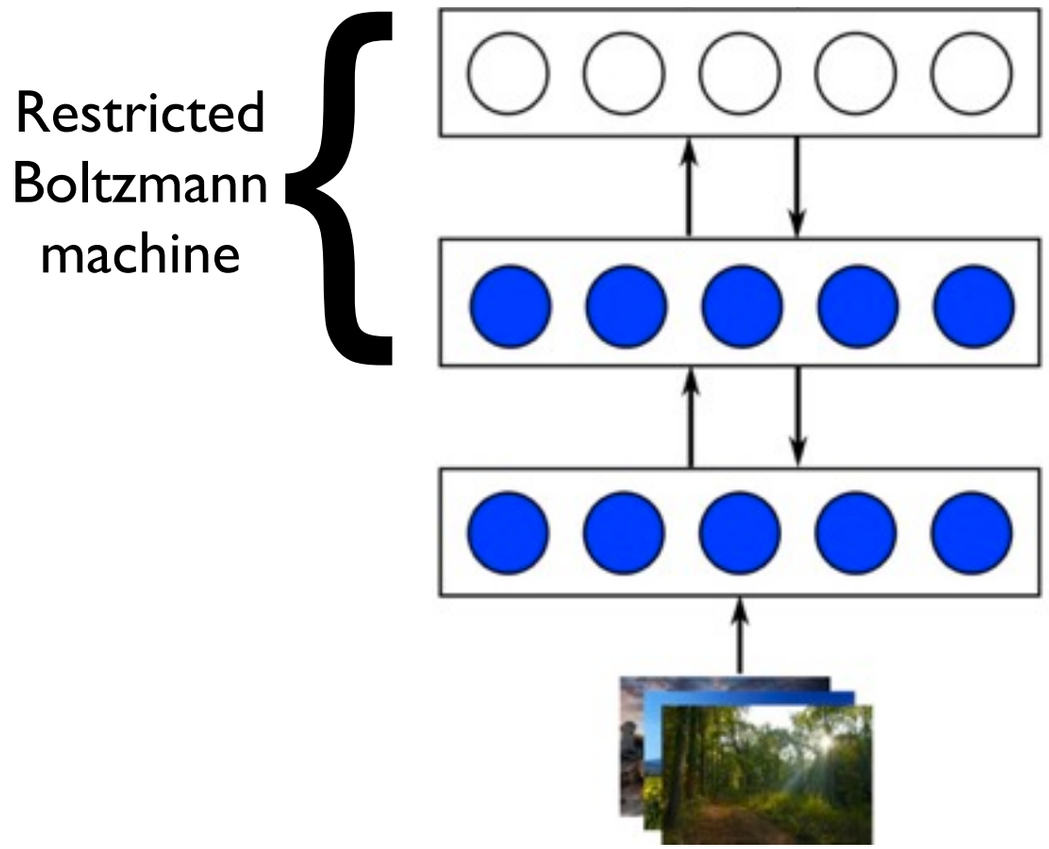


$$
\begin{aligned}
\boldsymbol{v}^1 &\sim \hat{P}(\boldsymbol{v}) \\
\boldsymbol{h}^1 &\sim P(\boldsymbol{h} \mid \boldsymbol{v}^1) \\
\boldsymbol{v}^2 &\sim P(\boldsymbol{v} \mid \boldsymbol{h}^1) \\
\boldsymbol{h}^2 &\sim P(\boldsymbol{h} \mid \boldsymbol{v}^2) \\
&\vdots \\
\boldsymbol{v}^{k+1} &\sim P(\boldsymbol{v} \mid \boldsymbol{h}^k)
\end{aligned}
$$

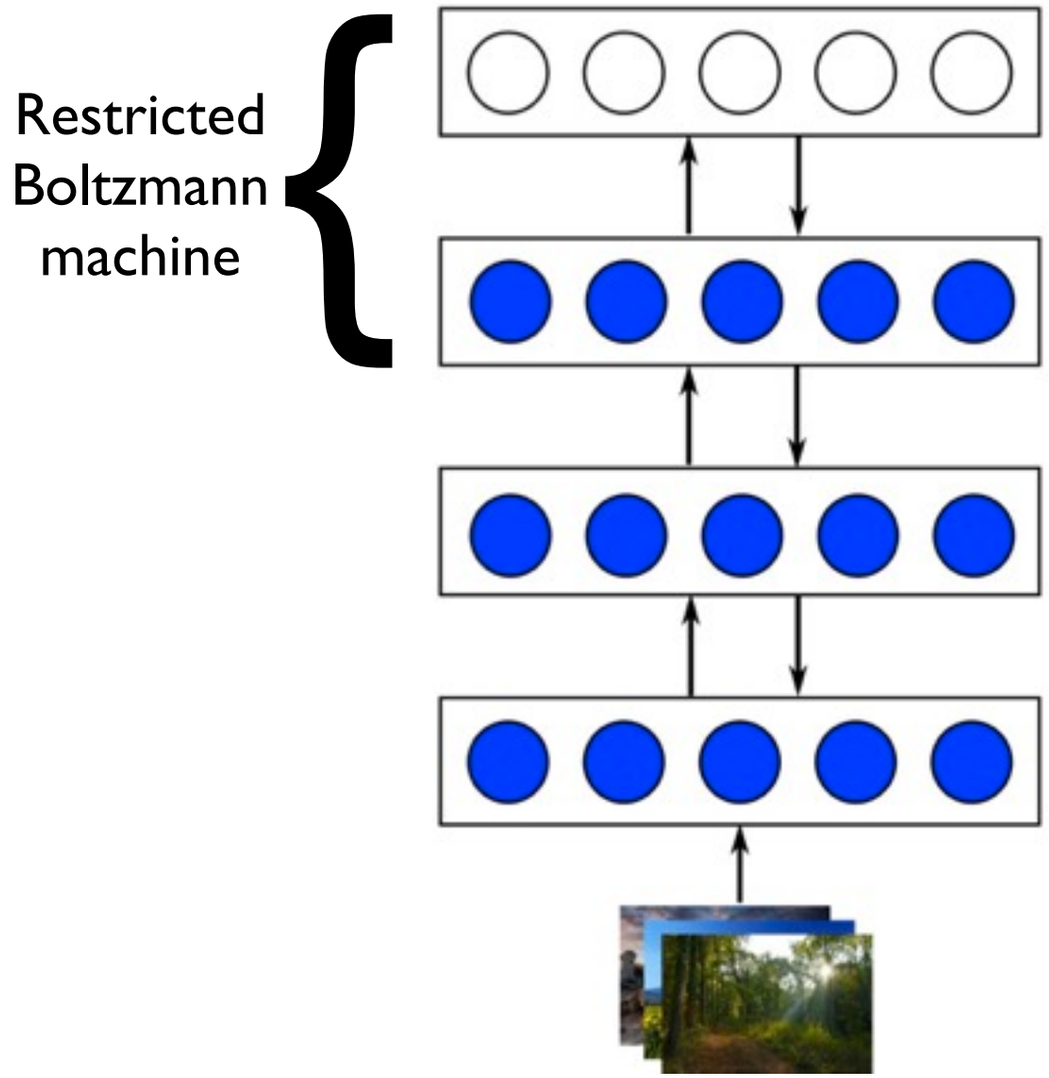
$$
\Delta\theta \propto \frac{\partial F(\mathbf{v}_{k+1})}{\partial \theta} - \frac{\partial F(\mathbf{v}_1)}{\partial \theta}
$$

“minimize surprise”

Learning deep belief networks by stacking restricted Boltzmann machines

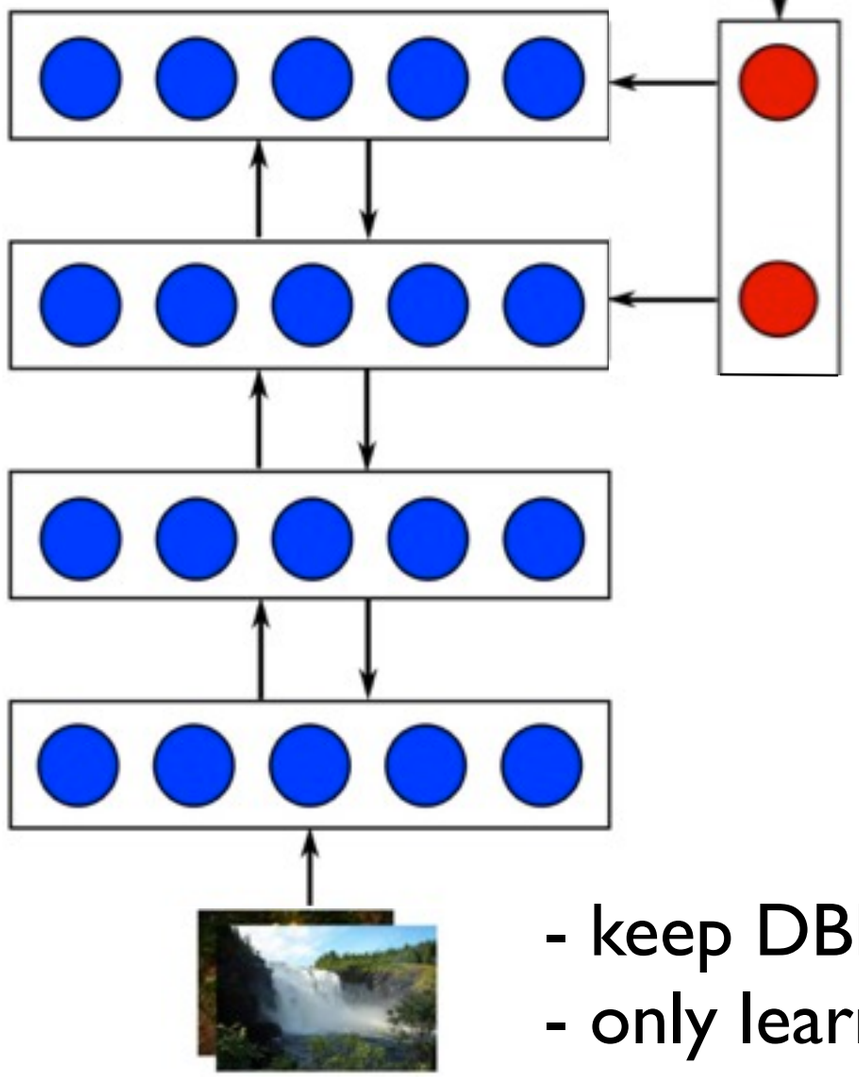


Learning deep belief networks by stacking restricted Boltzmann machines



conditional restricted Boltzmann machine:

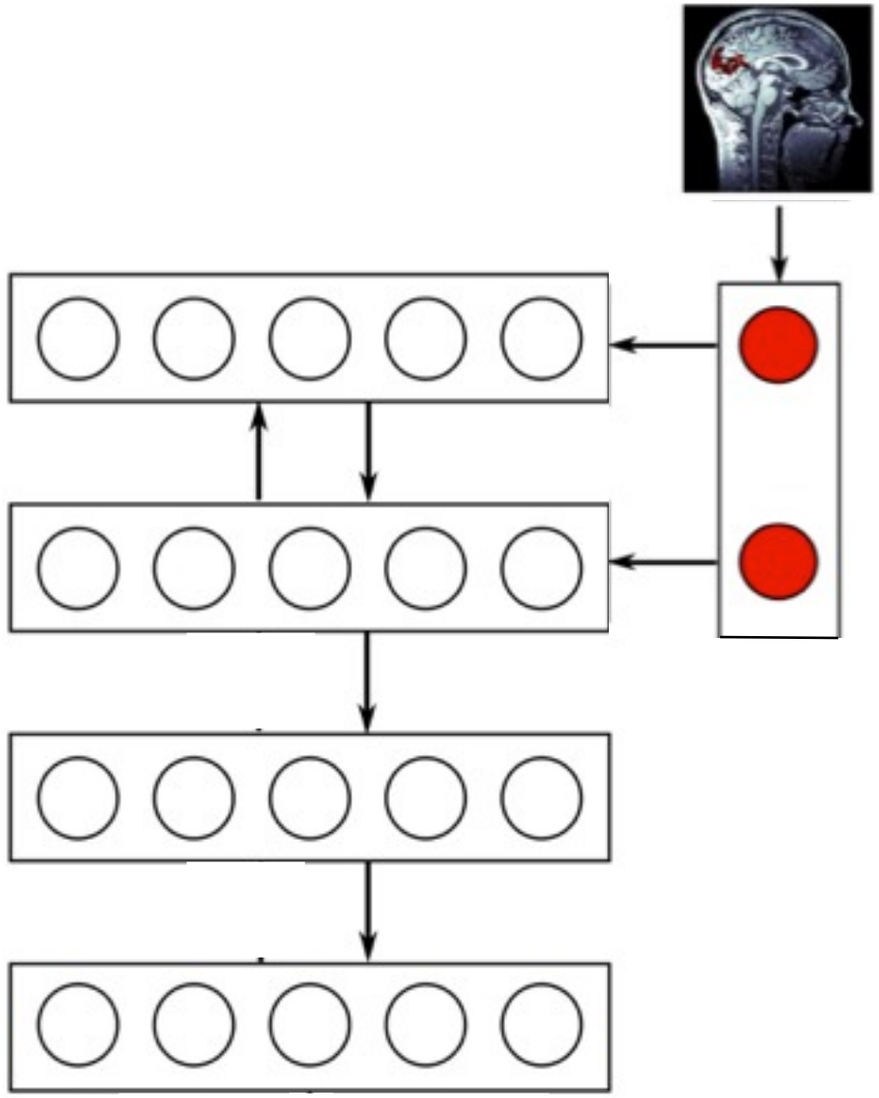
$$E(\boldsymbol{v}, \boldsymbol{h} \mid \boldsymbol{z}) = -\boldsymbol{h}^T \boldsymbol{W} \boldsymbol{v} - \boldsymbol{z}^T \boldsymbol{C} \boldsymbol{v} - \boldsymbol{z}^T \boldsymbol{B} \boldsymbol{h}$$



- keep DBN fixed
- only learn adjustments of bias

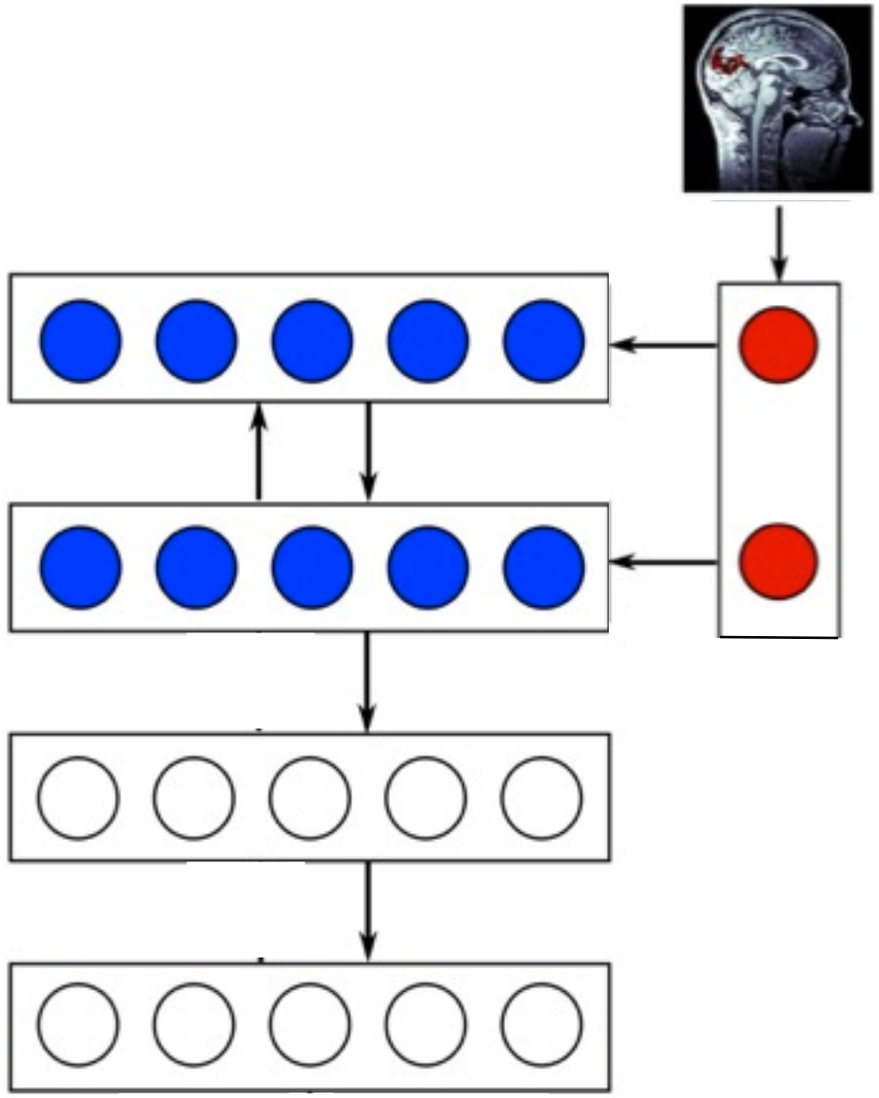
Reconstruction

Reconstruction phase



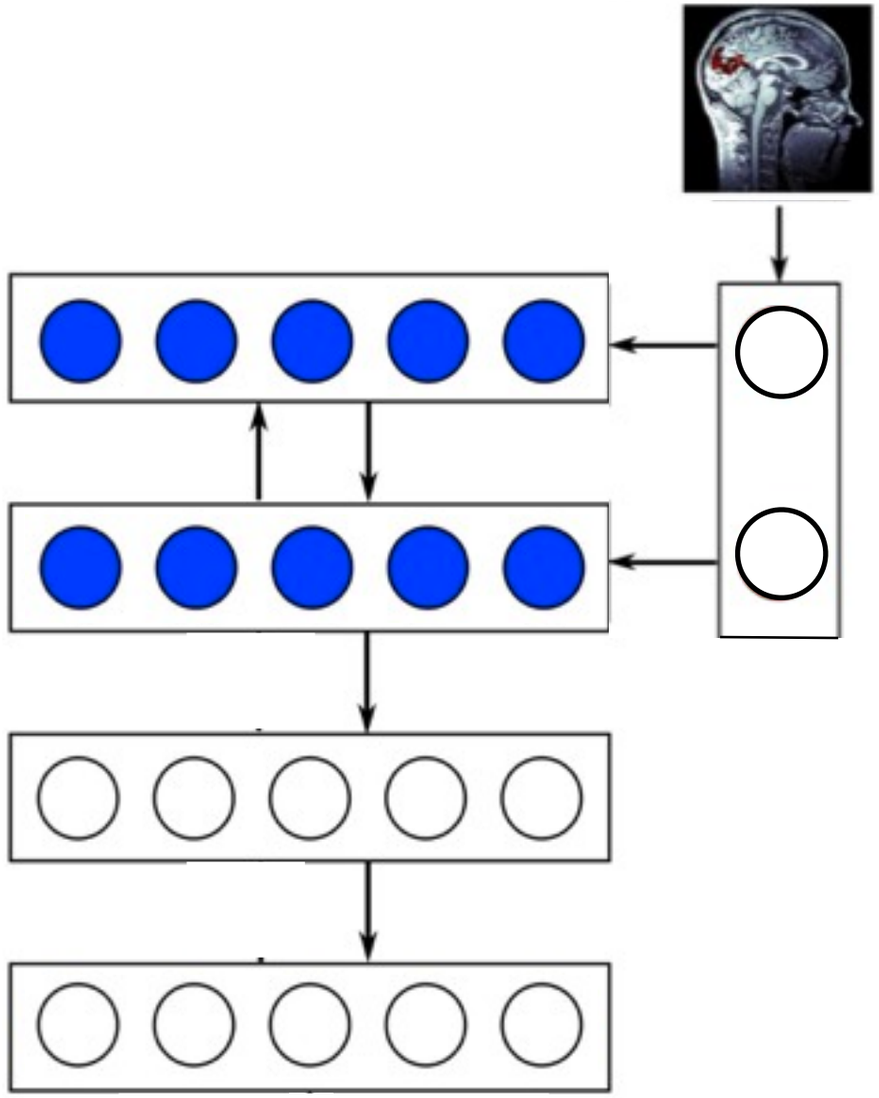
Reconstruction

Reconstruction phase



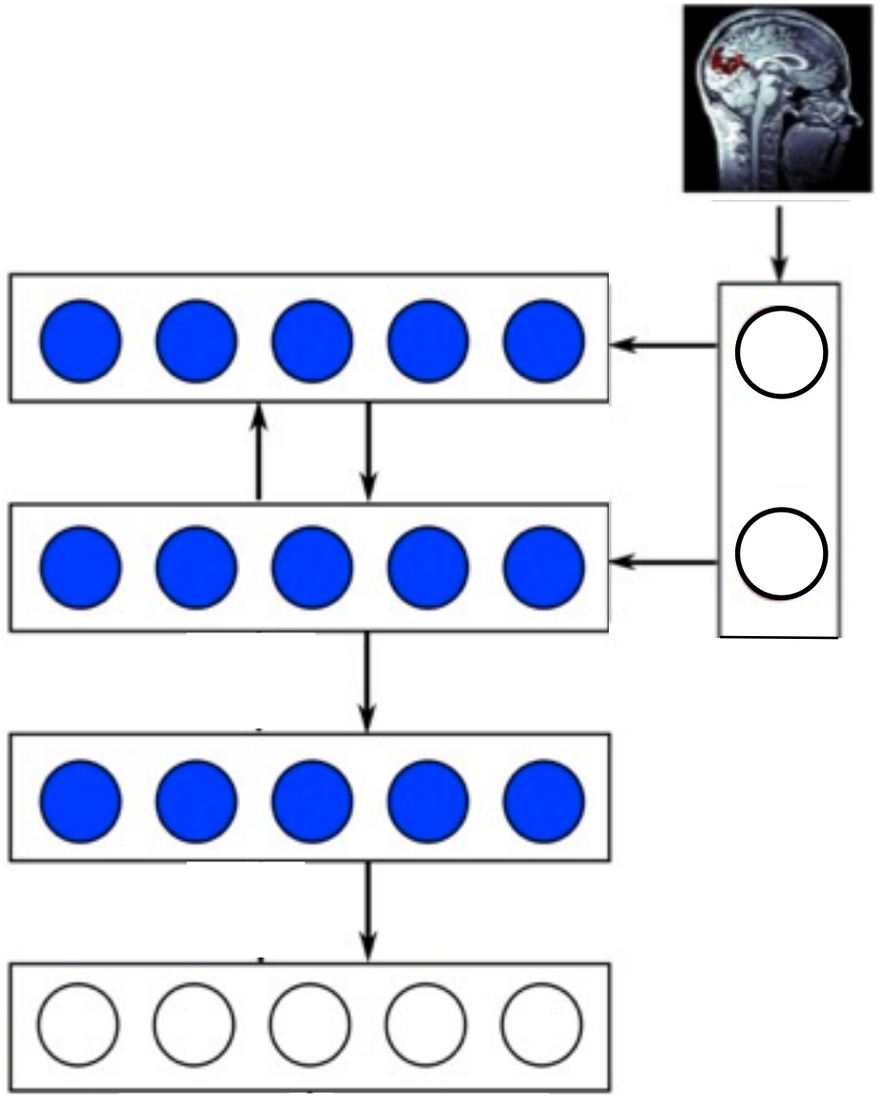
Reconstruction

Reconstruction phase



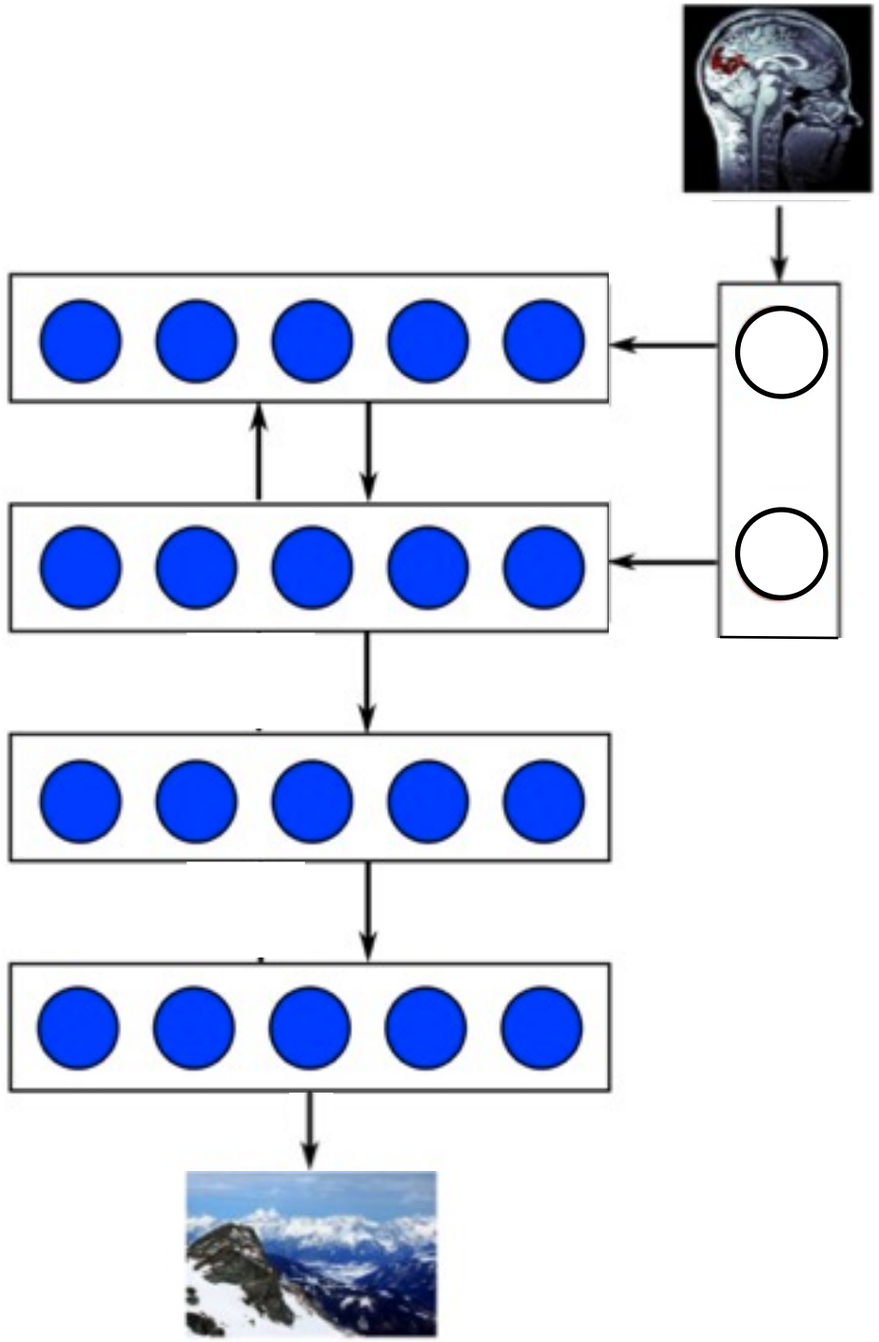
Reconstruction

Reconstruction phase



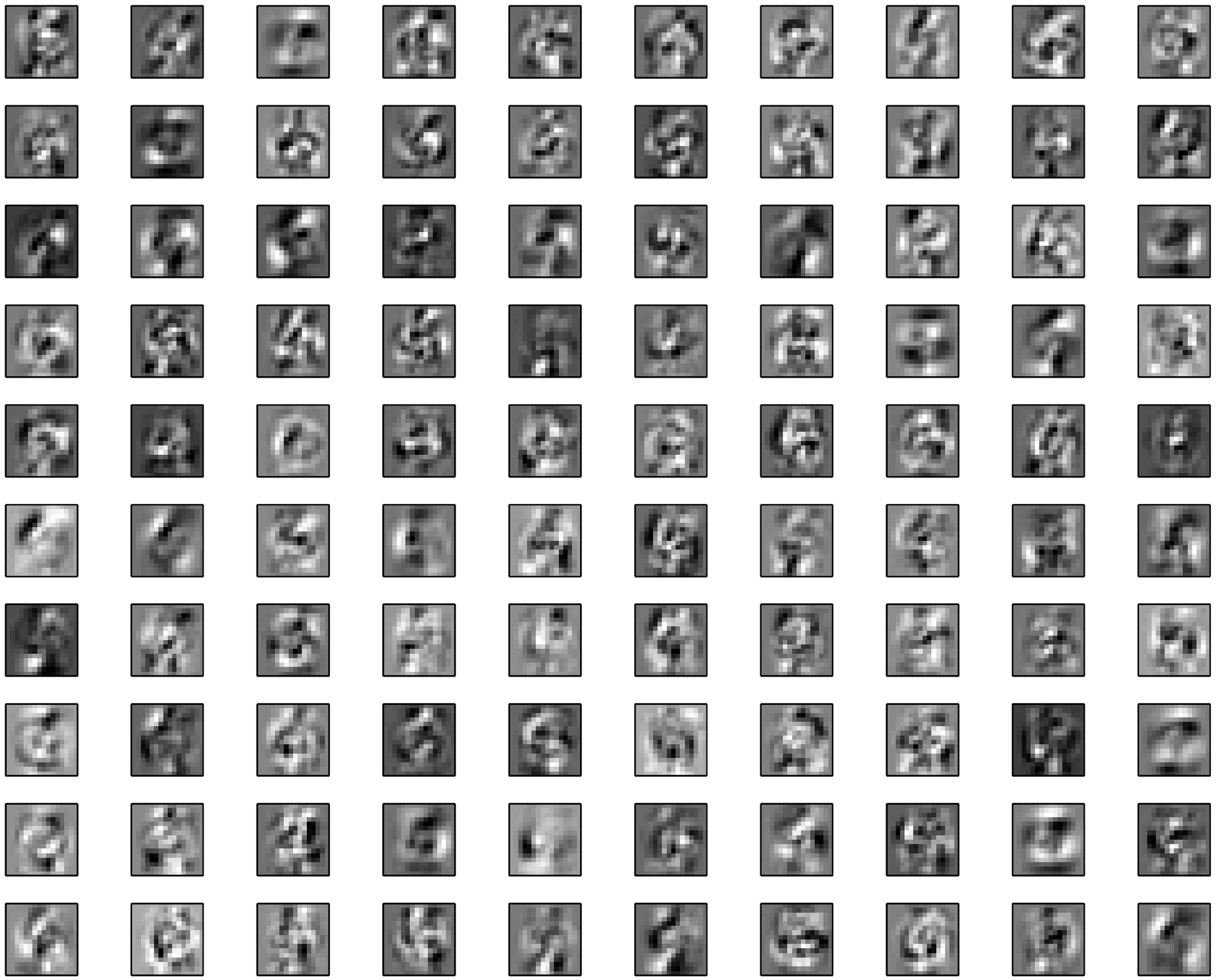
Reconstruction

Reconstruction phase

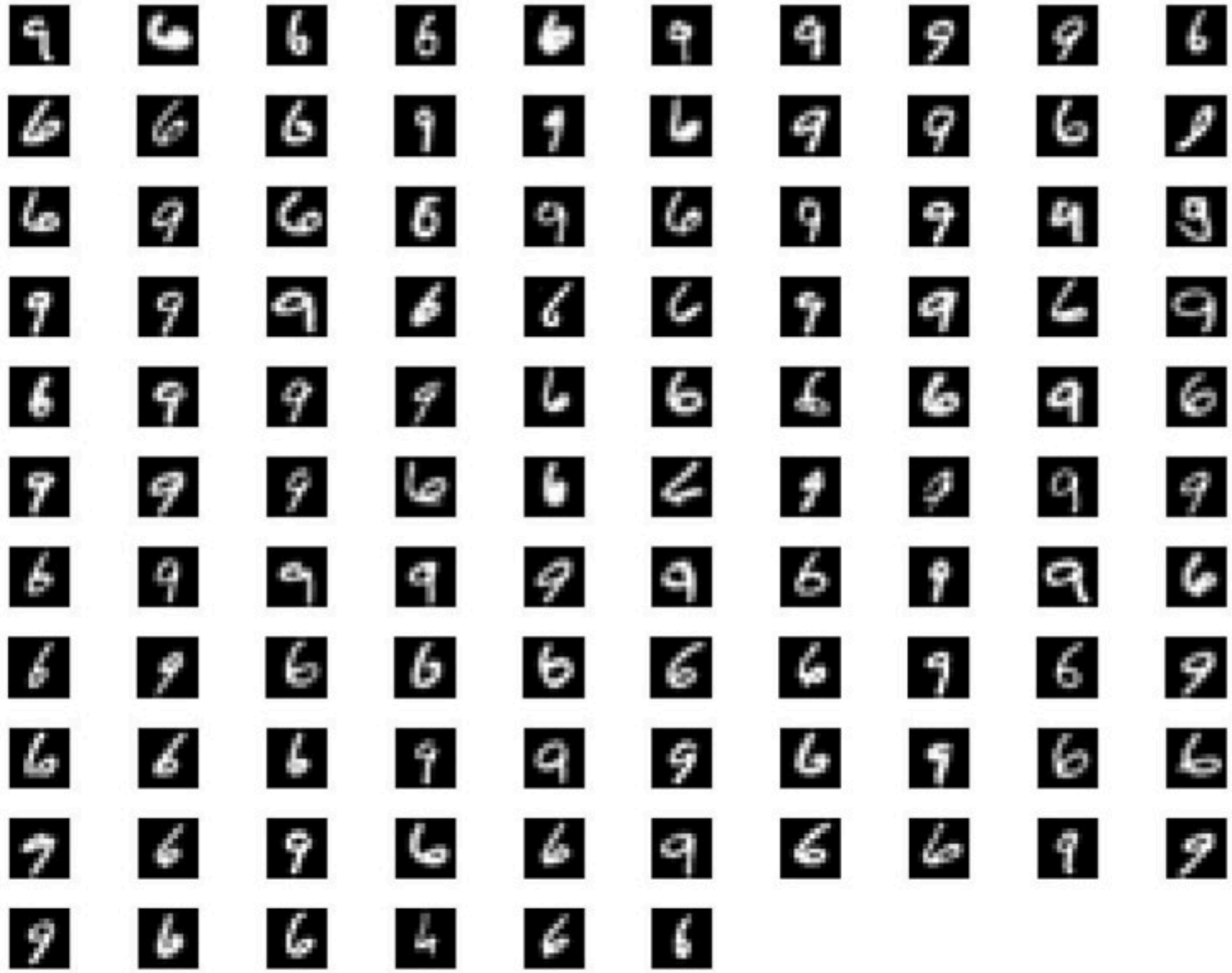


First layer filters

Second layer filters

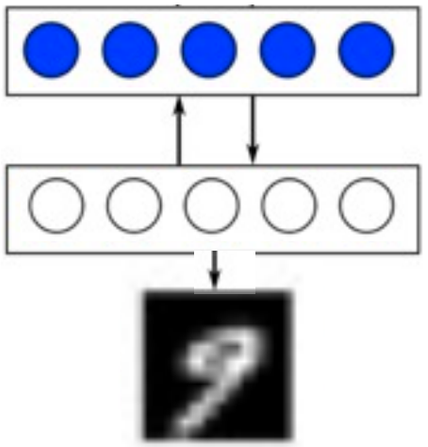


Stimuli

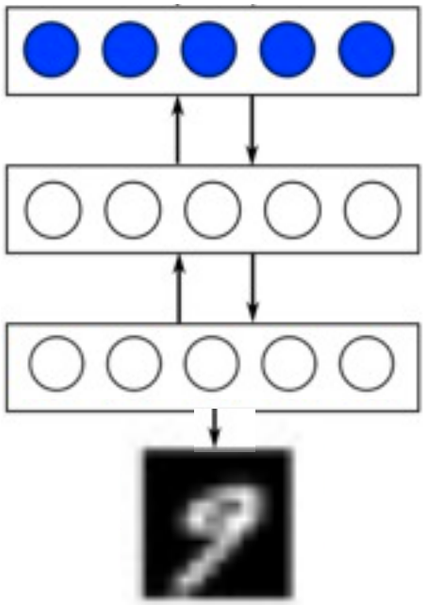


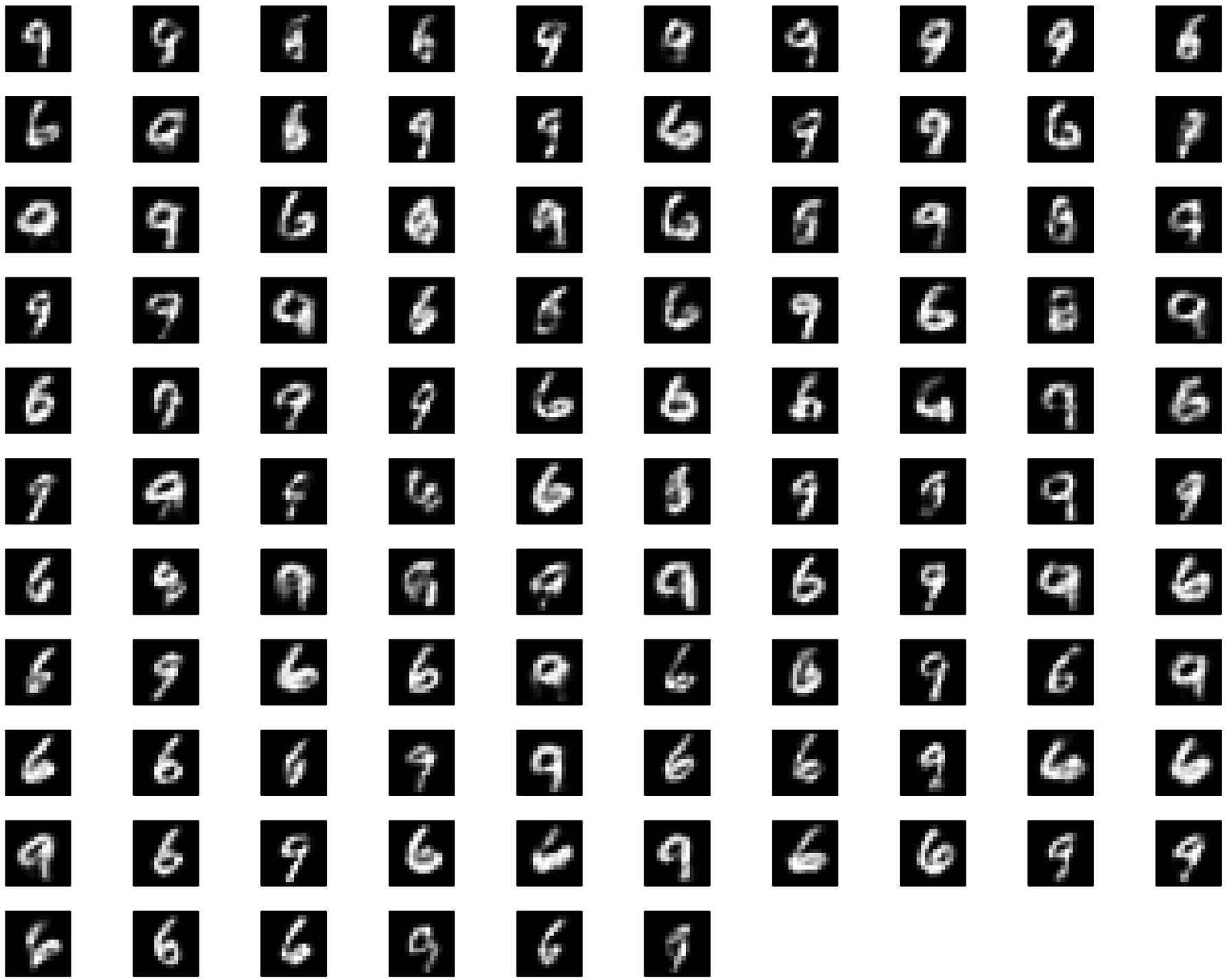
Zero layer reconstructions

First layer reconstructions

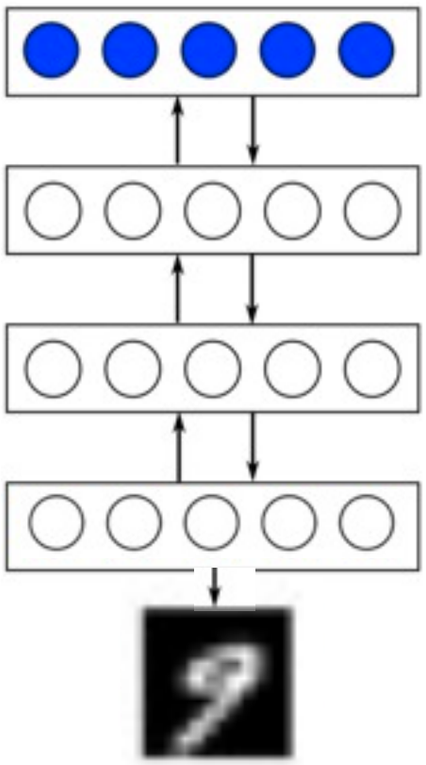


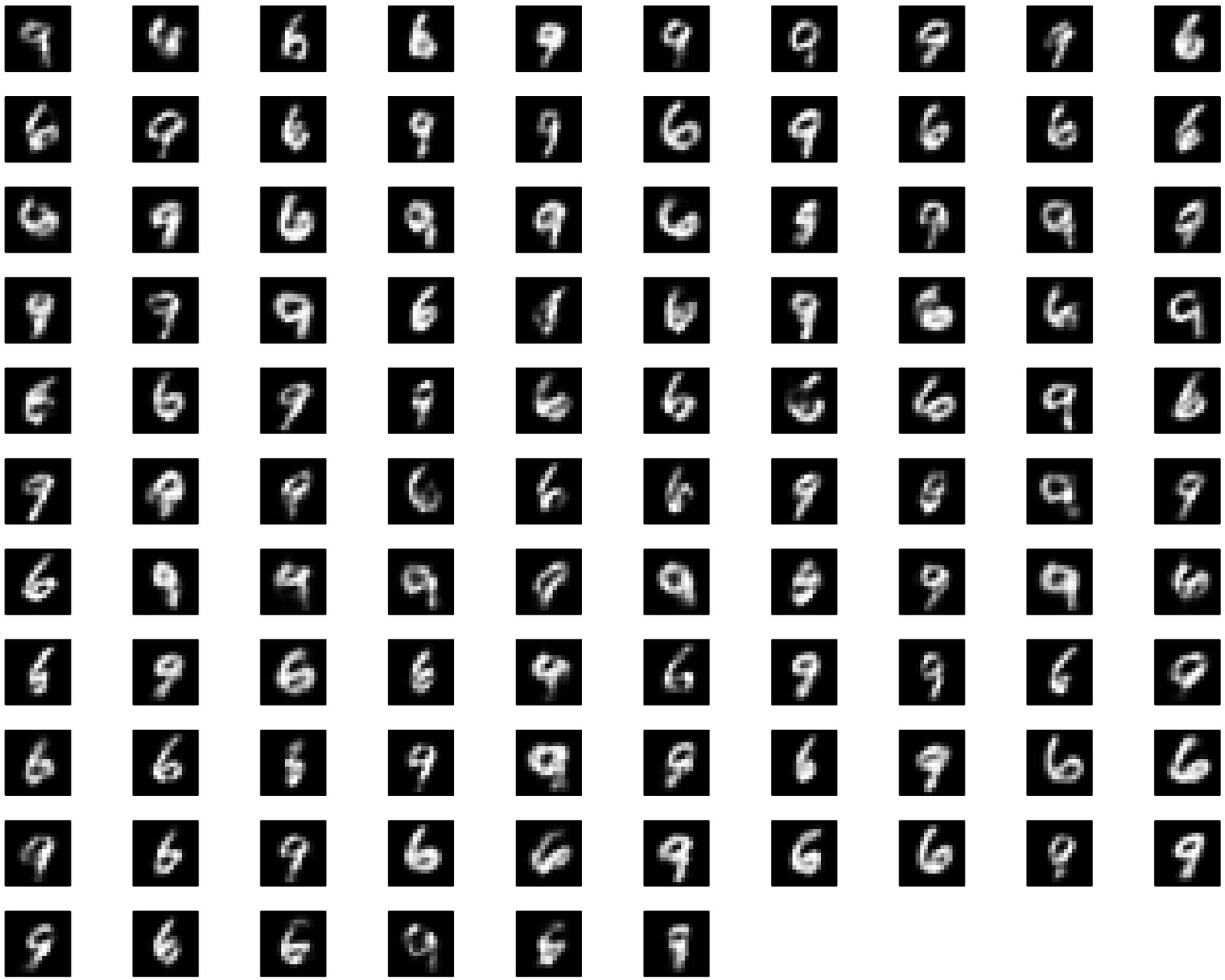
Second layer reconstructions



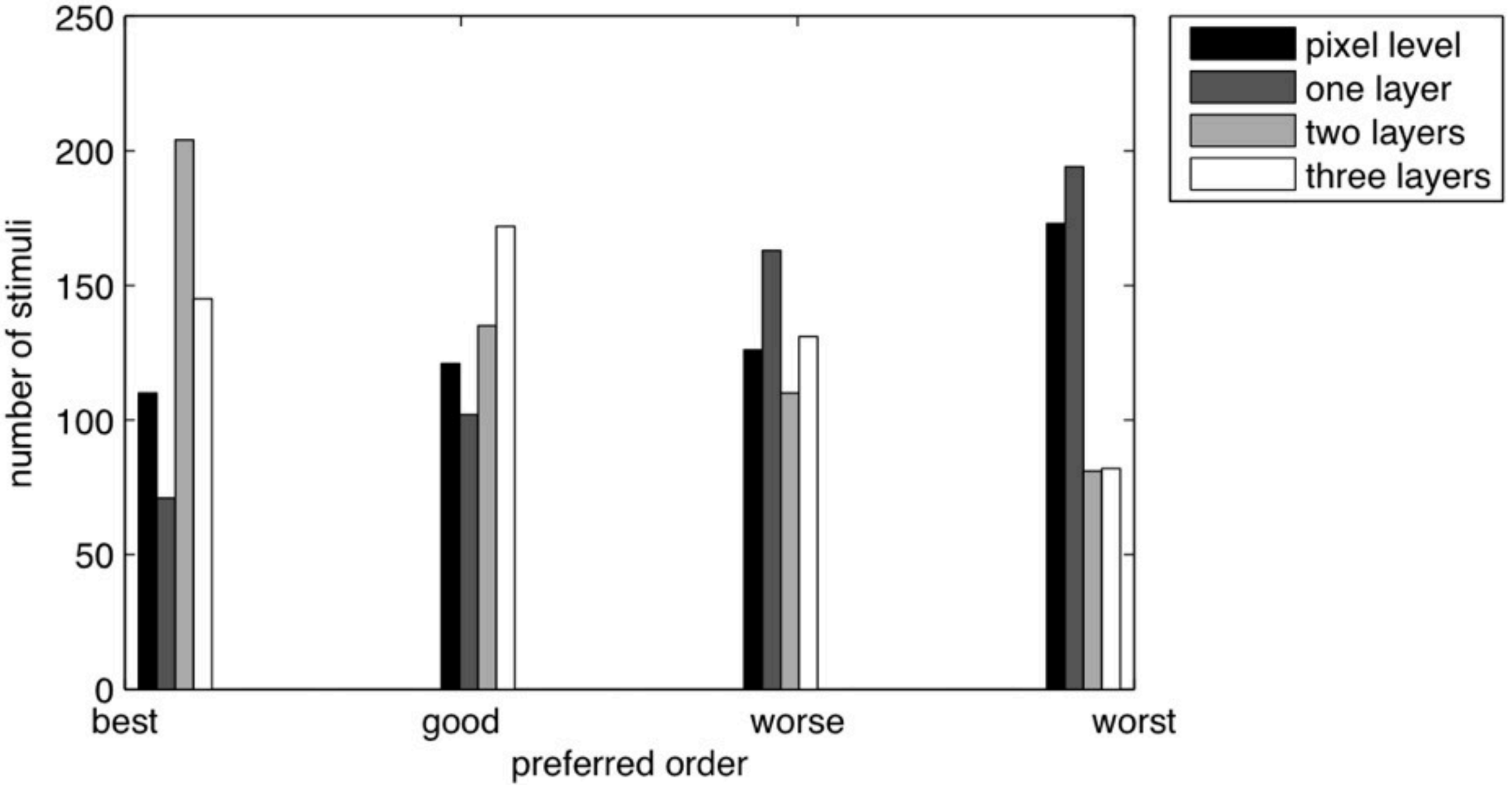


Third layer reconstructions

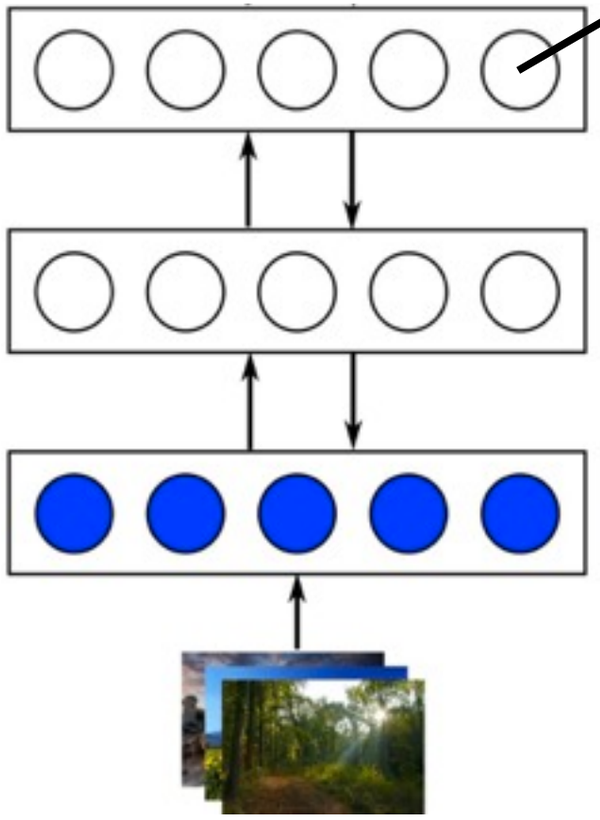


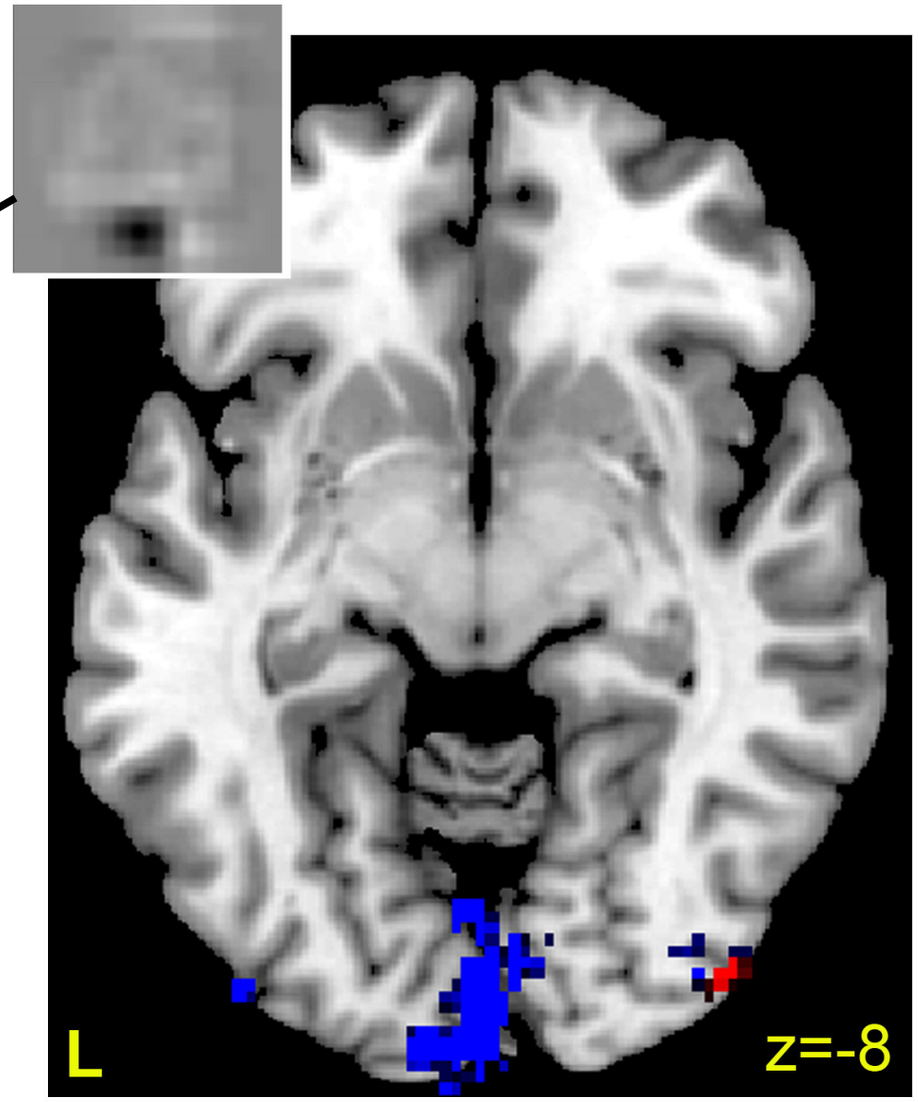


Decoding performance



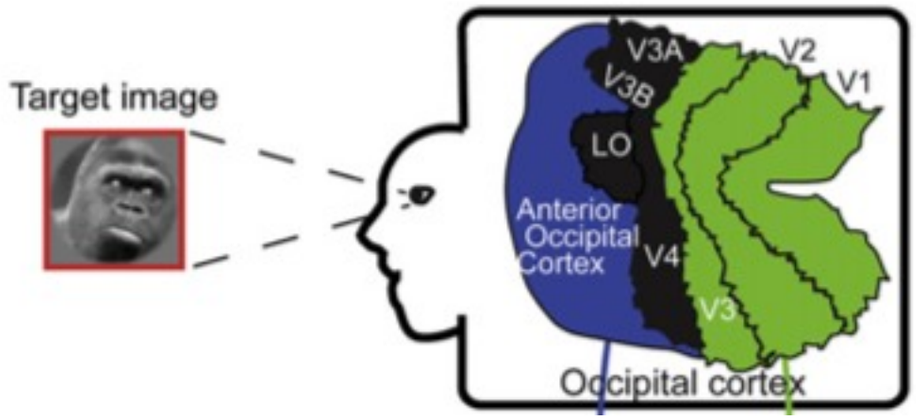
Interpretation





Reconstruction: combining sources of information

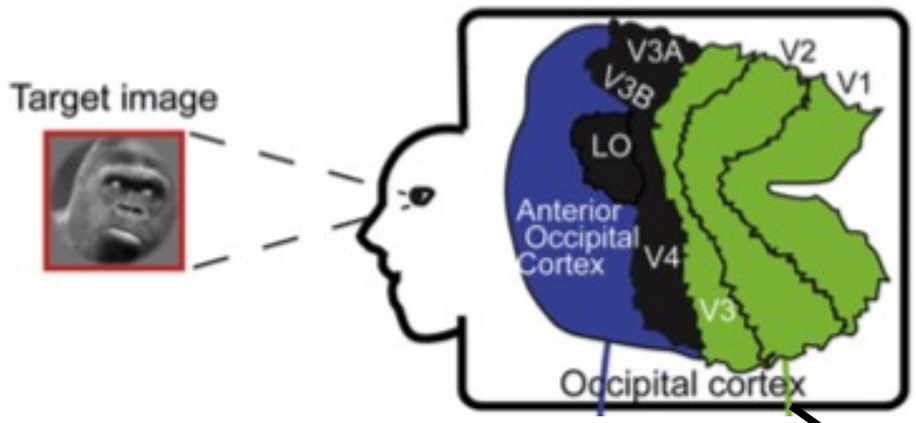
generative approach



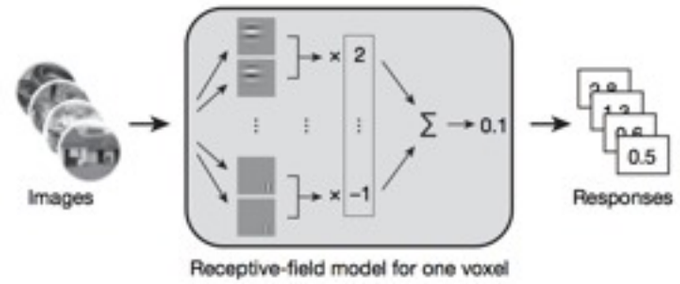
T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

How to solve the reconstruction problem?

generative approach

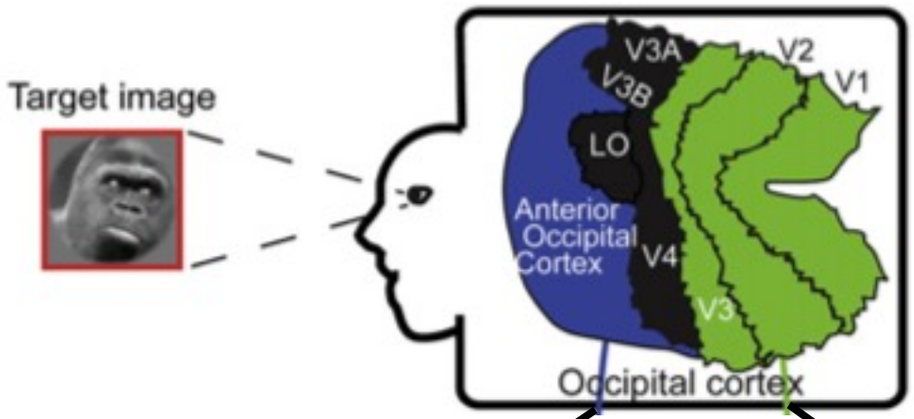


structural encoding model

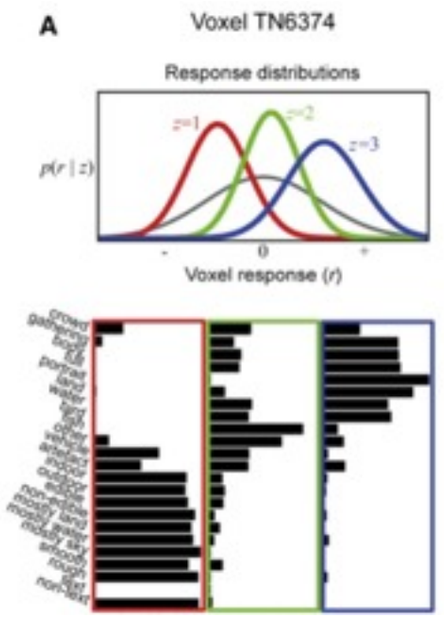


How to solve the reconstruction problem?

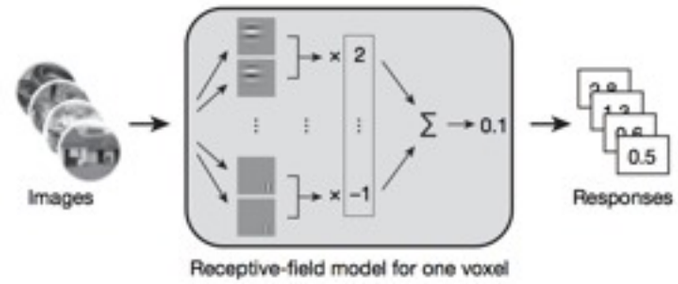
generative approach



semantic encoding model

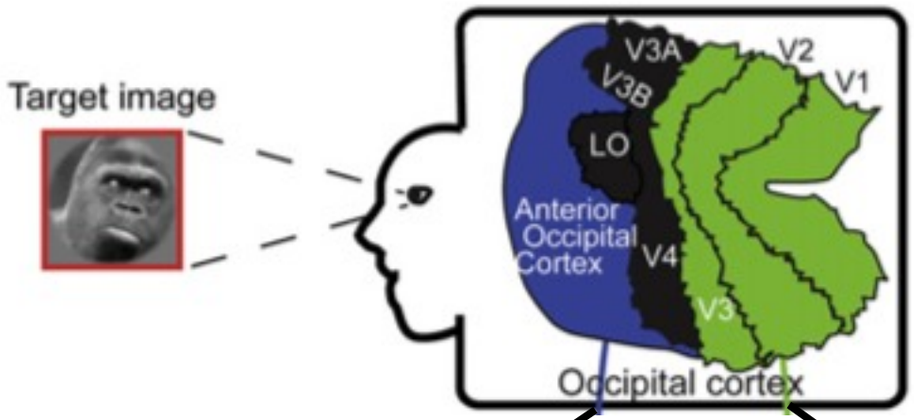


structural encoding model

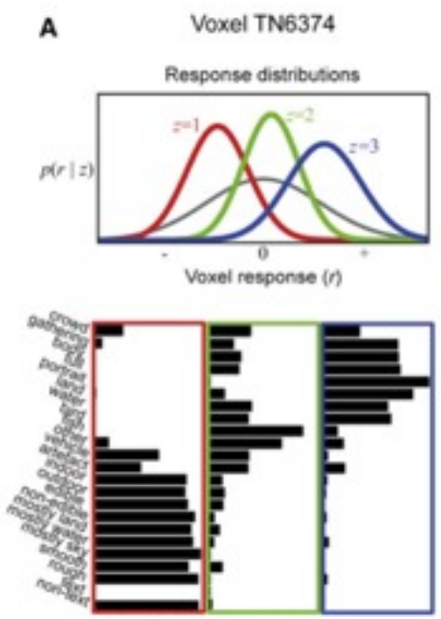


How to solve the reconstruction problem?

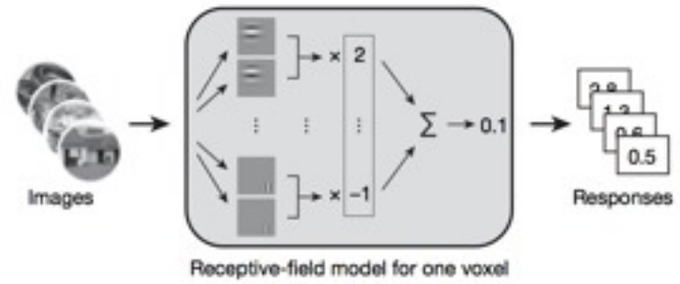
generative approach



semantic encoding model



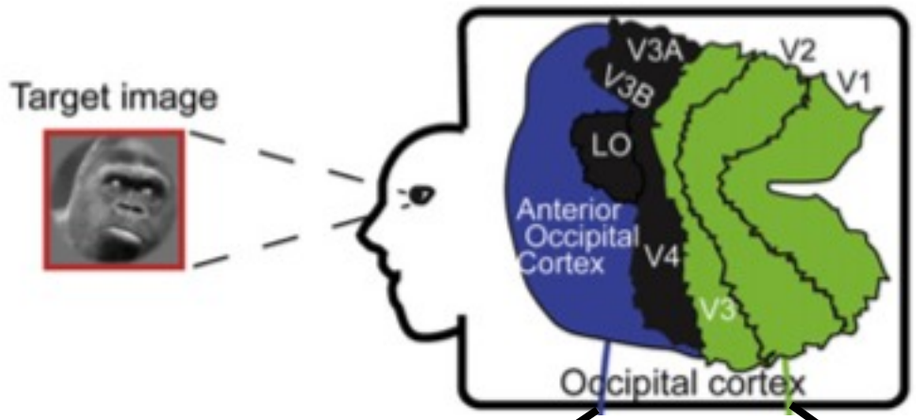
structural encoding model



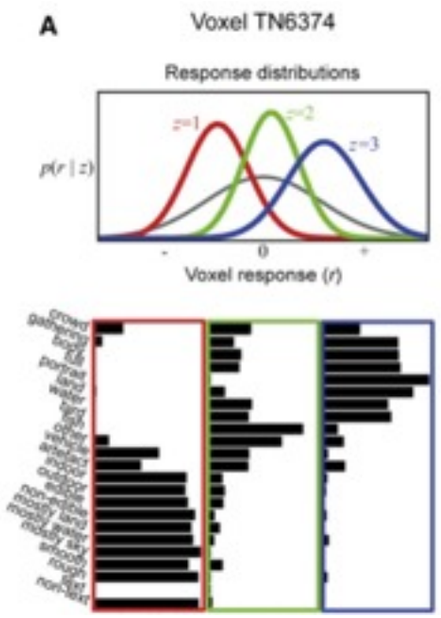
natural image prior

How to solve the reconstruction problem?

generative approach



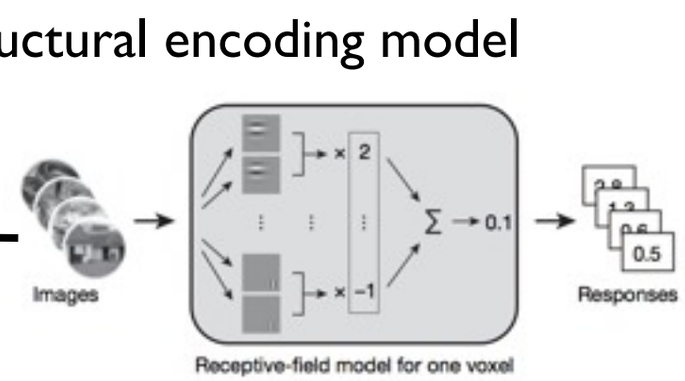
semantic encoding model



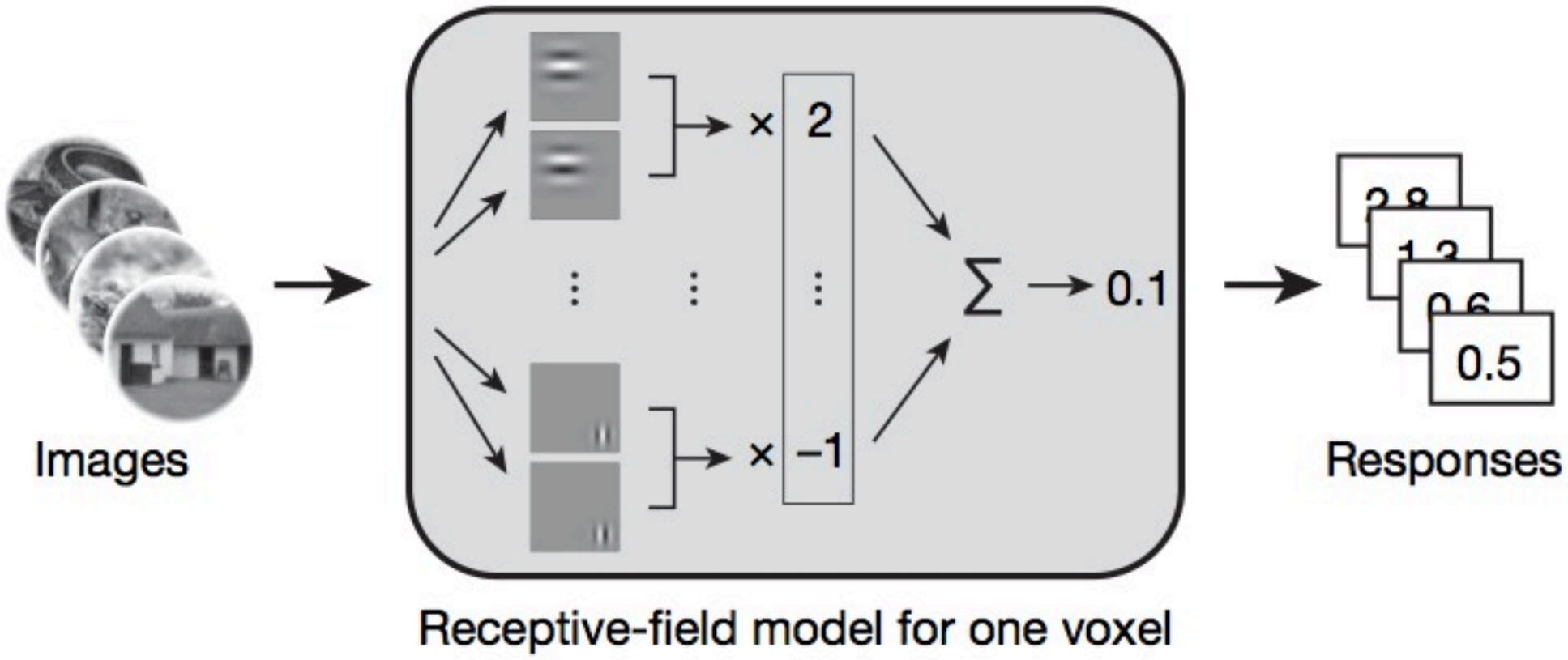
$$p(\mathbf{s}|\mathbf{r}) \propto p(\mathbf{s}) \prod_i p_i(\mathbf{r}_i|\mathbf{s})$$

natural image prior

structural encoding model



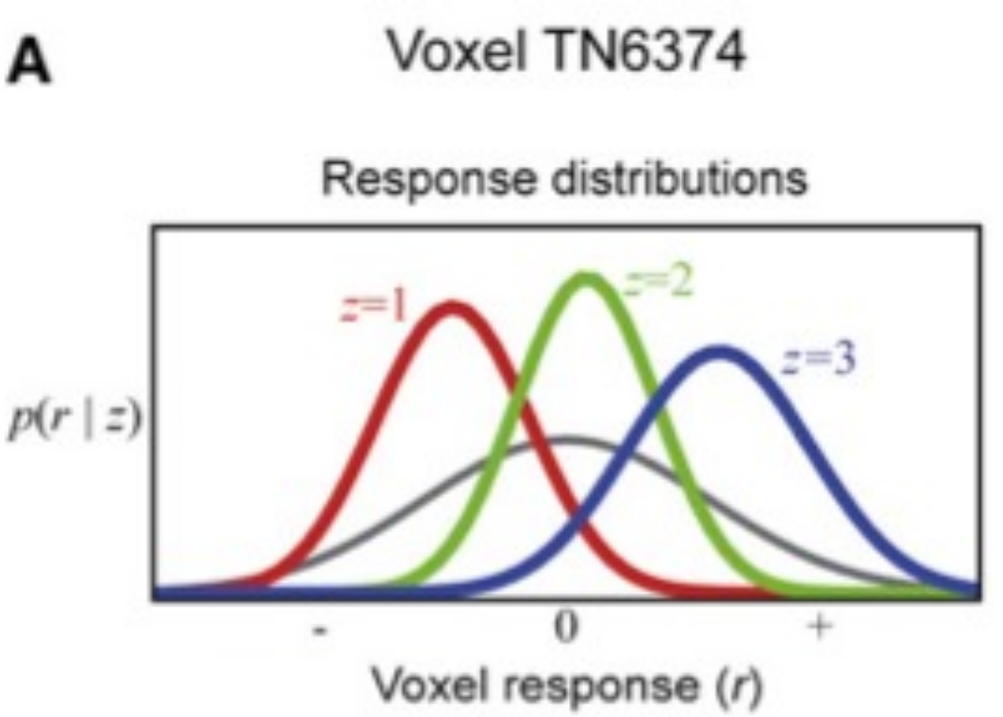
Structural encoding model



$$p(r \mid \boldsymbol{s}) \propto \exp\left(-\frac{r - \boldsymbol{h}^T \boldsymbol{f}(\boldsymbol{s}))^2}{2\sigma^2}\right)$$

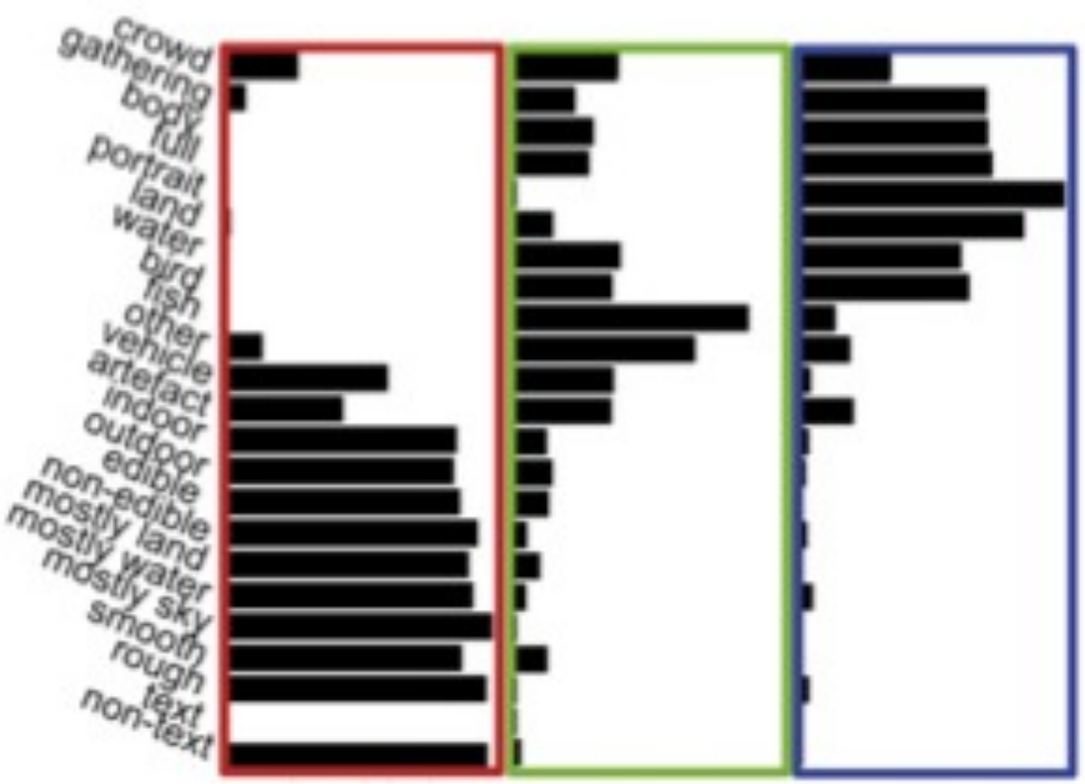
where $\boldsymbol{f}(\boldsymbol{s}) = \log(|W^T \boldsymbol{s}| + 1)$

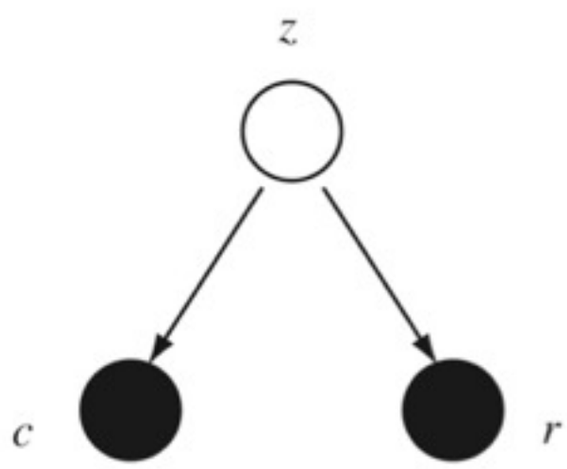
Semantic encoding model



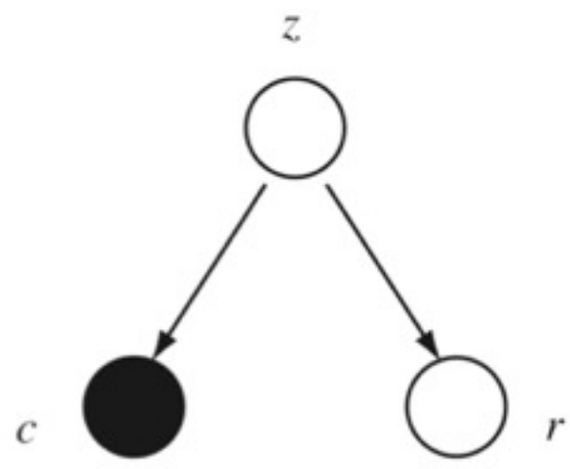
$$p(r \mid c(\boldsymbol{s})) = \sum_z p(r \mid z)p(z \mid c(\boldsymbol{s}))$$

$$p(r \mid z) \propto \exp\left(-\frac{(r-\mu_z)^2}{2\sigma^2}\right)$$



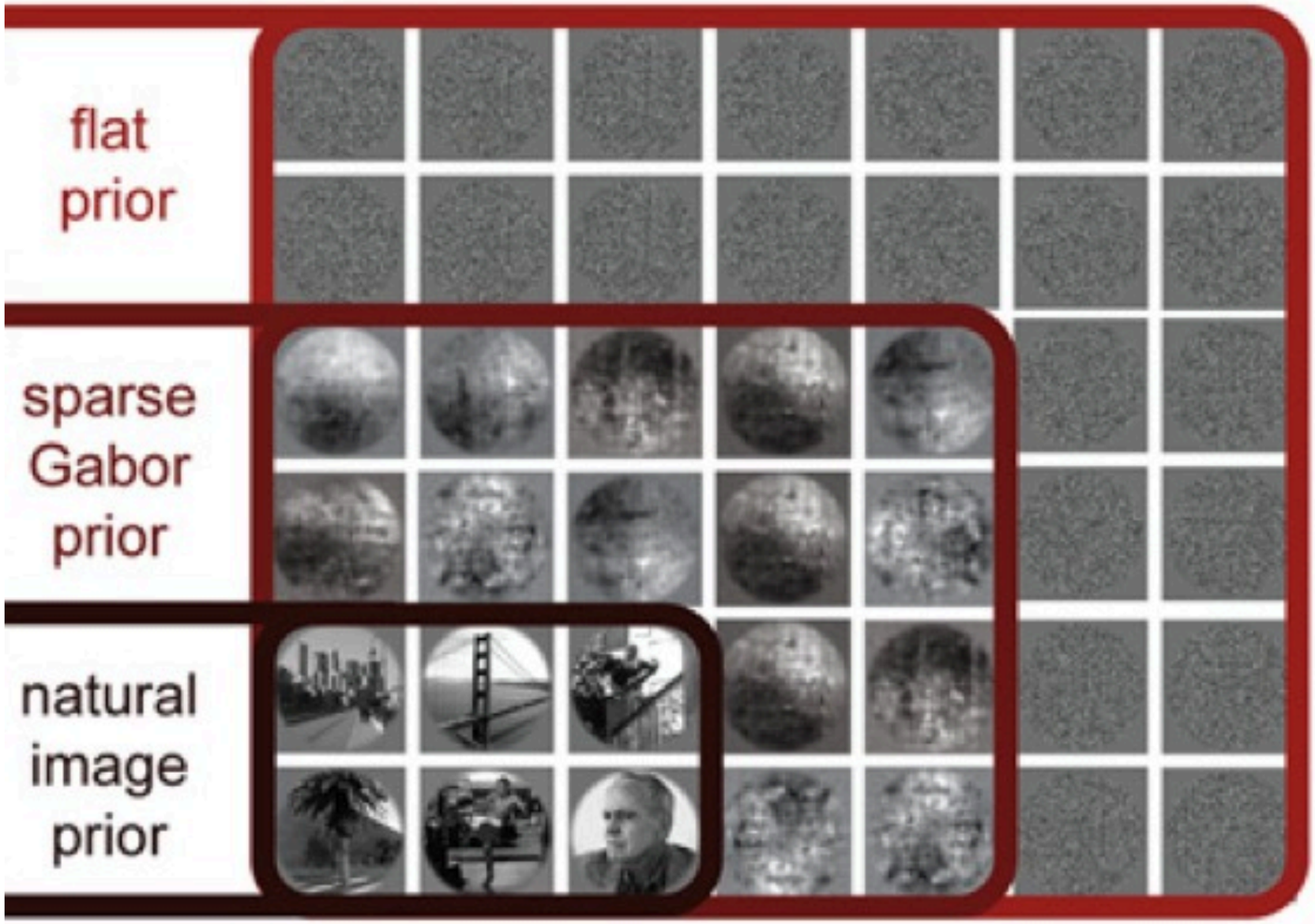


Parameter estimation



Response prediction

$p(\mathbf{s})$

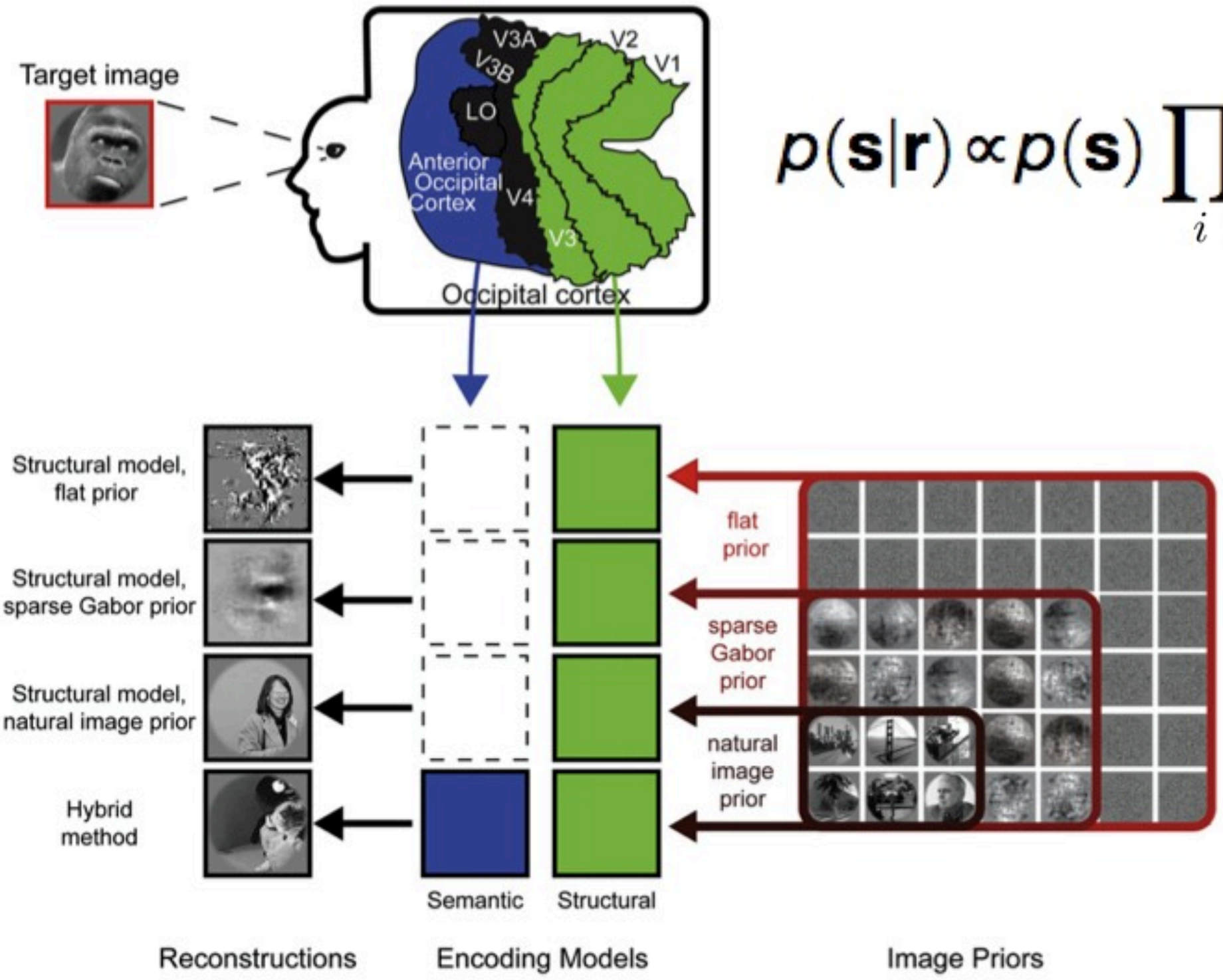


$p(\mathbf{s}) = \text{constant.}$

$$p_{SG}(\mathbf{s}) = \int_{\mathbf{a}} p(\mathbf{s}|\mathbf{a})p(\mathbf{a})d\mathbf{a}$$

$$p_{NIP}(\mathbf{s}) = \frac{1}{C}\sum_{i=1}^{C} \delta_{\mathbf{s}^{(i)}}(\mathbf{s})$$

The final algorithm



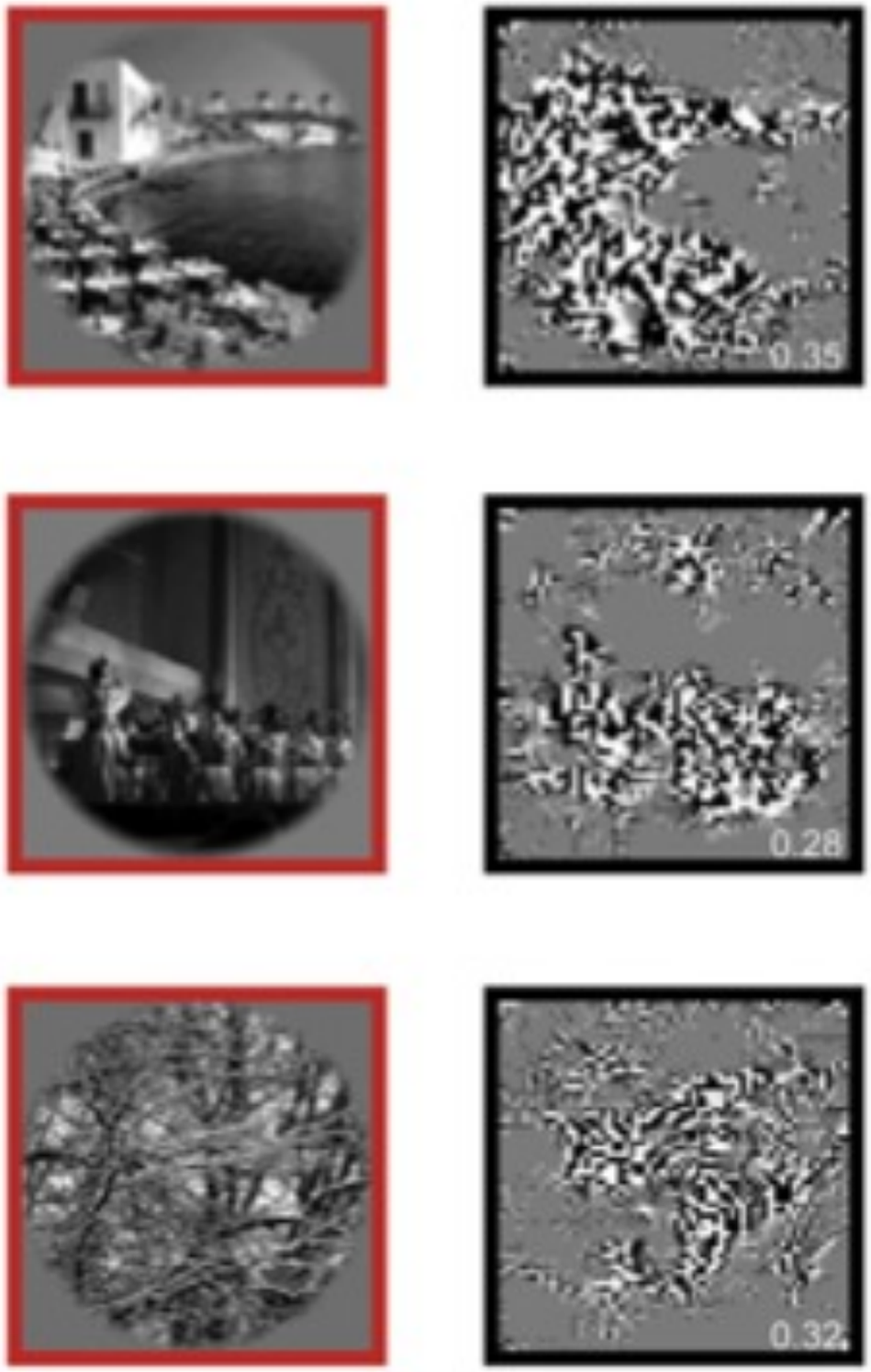
$$p(\mathbf{s}|\mathbf{r}) \propto p(\mathbf{s}) \prod_{i \in \{\text{semantic, structural}\}} p_i(\mathbf{r}_i|\mathbf{s})$$

= stochastic search

= identification

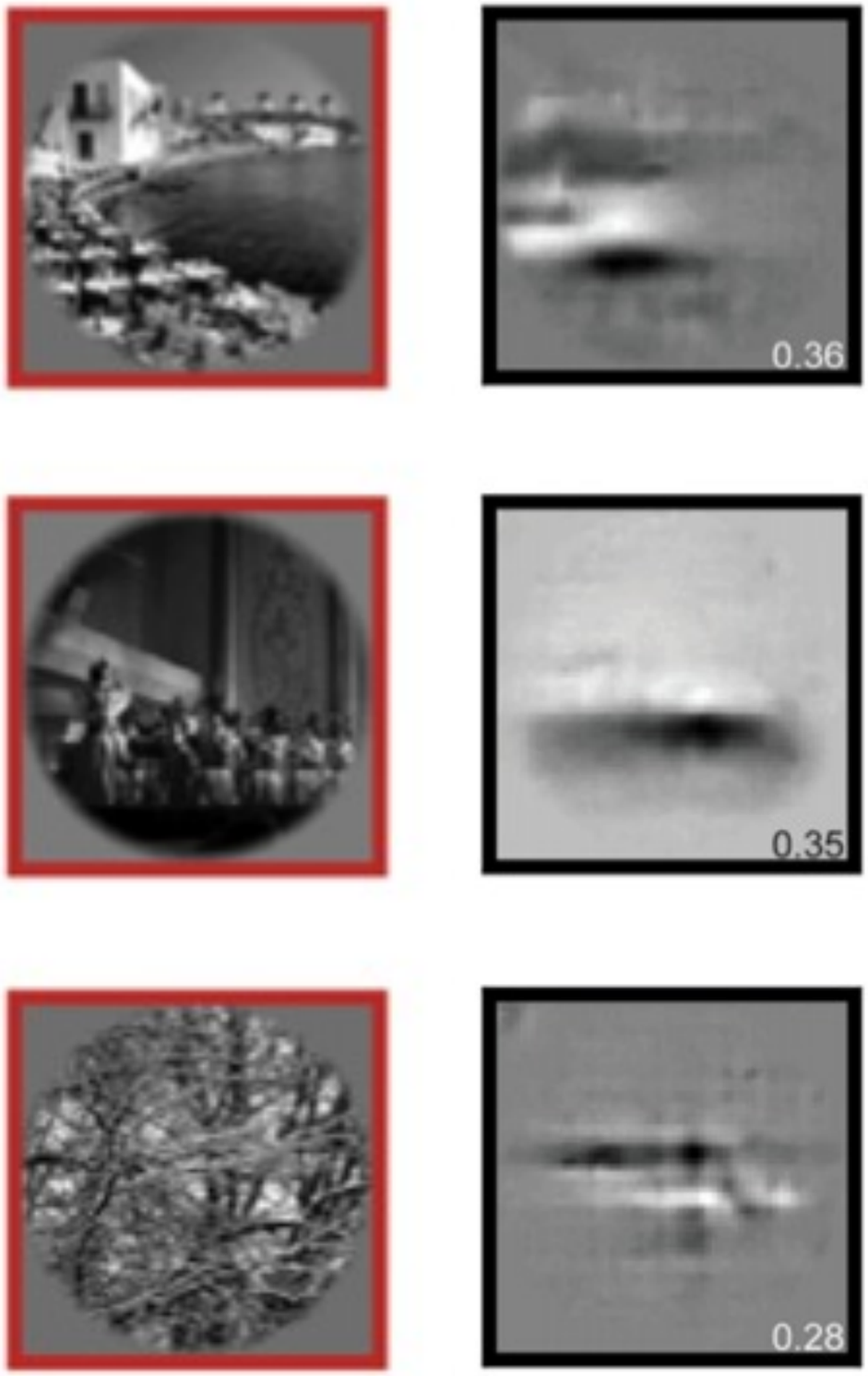
Results

structural encoding + flat prior



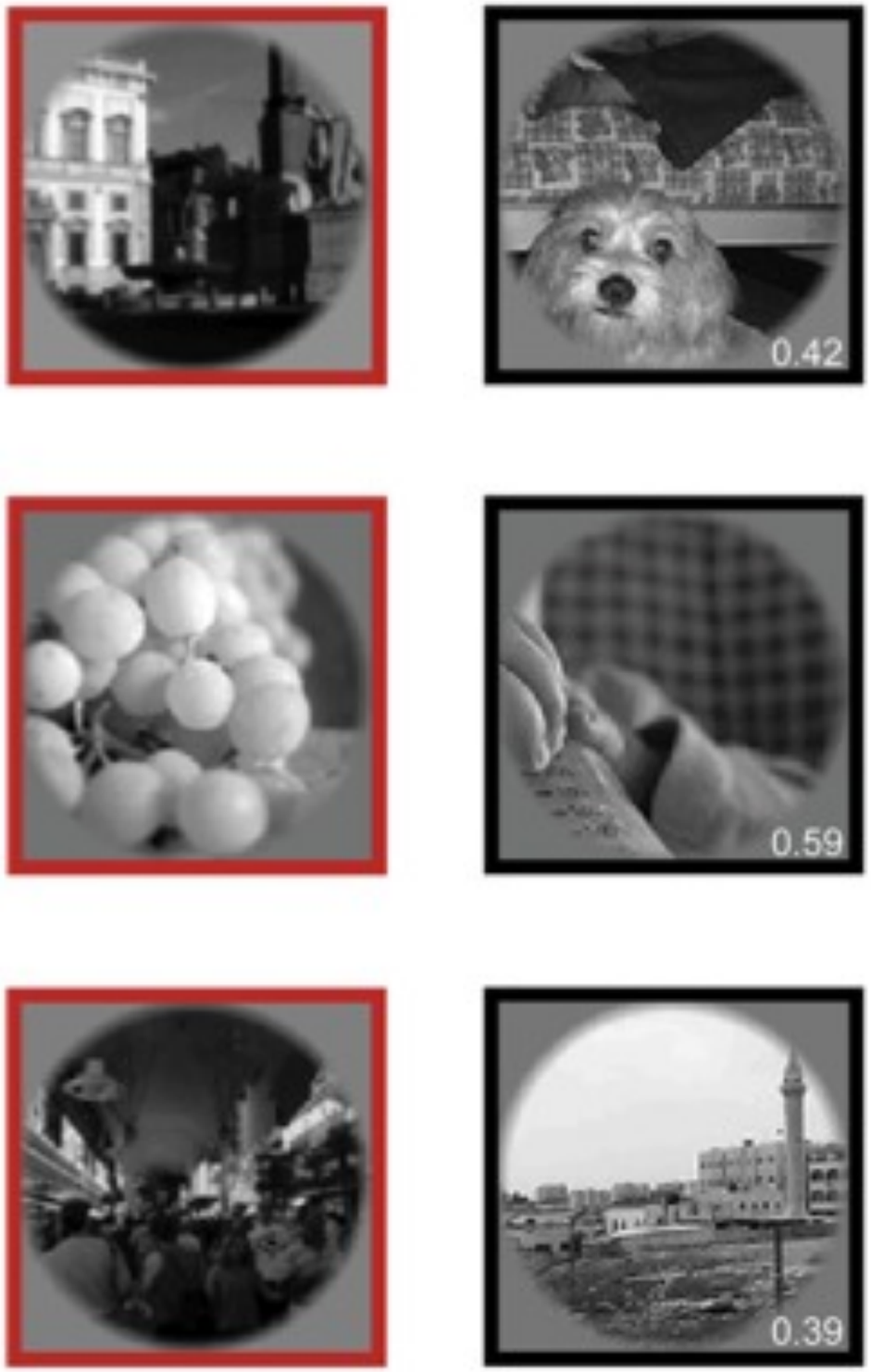
Results

structural encoding + sparse Gabor prior



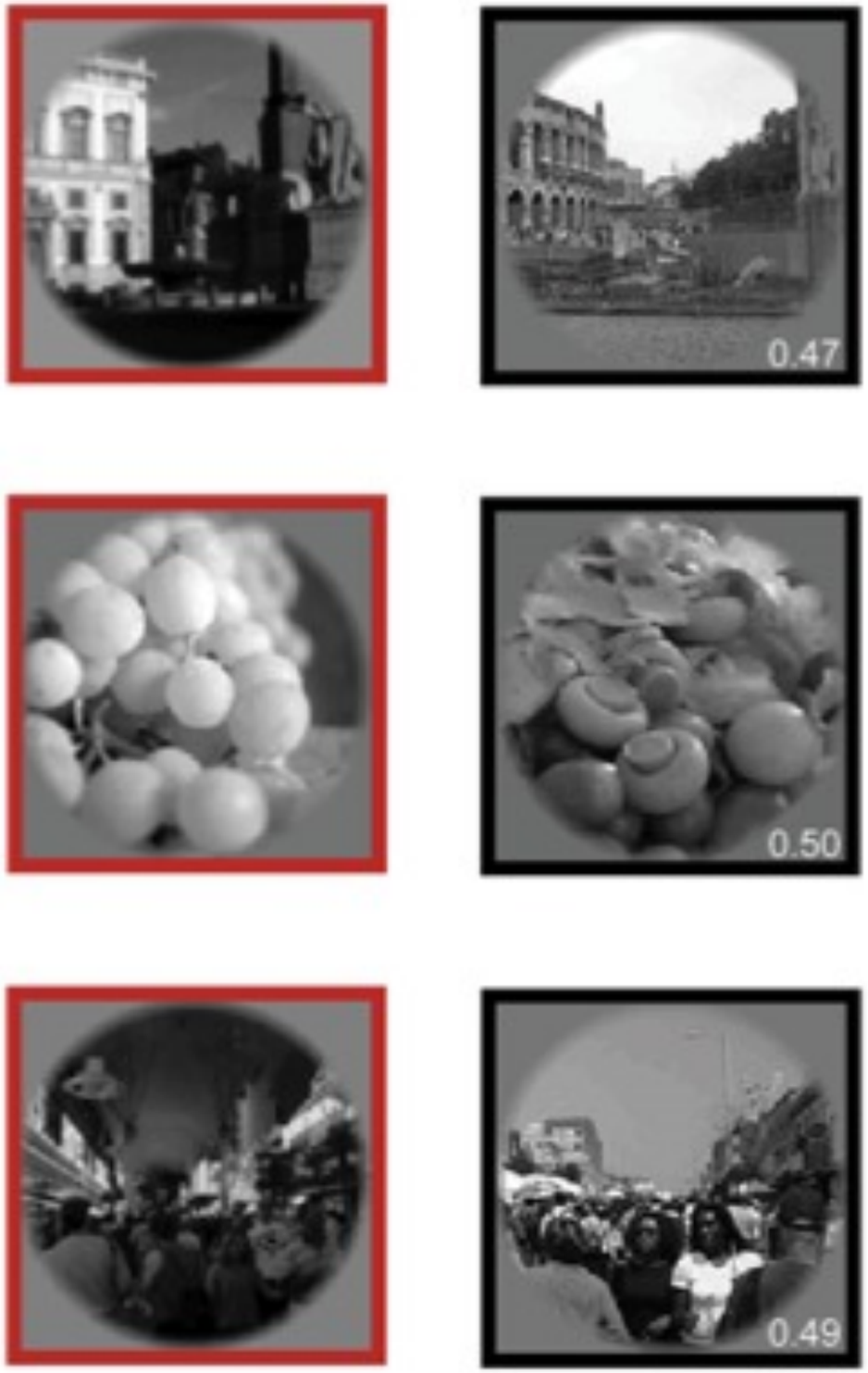
Results

structural encoding + natural image prior

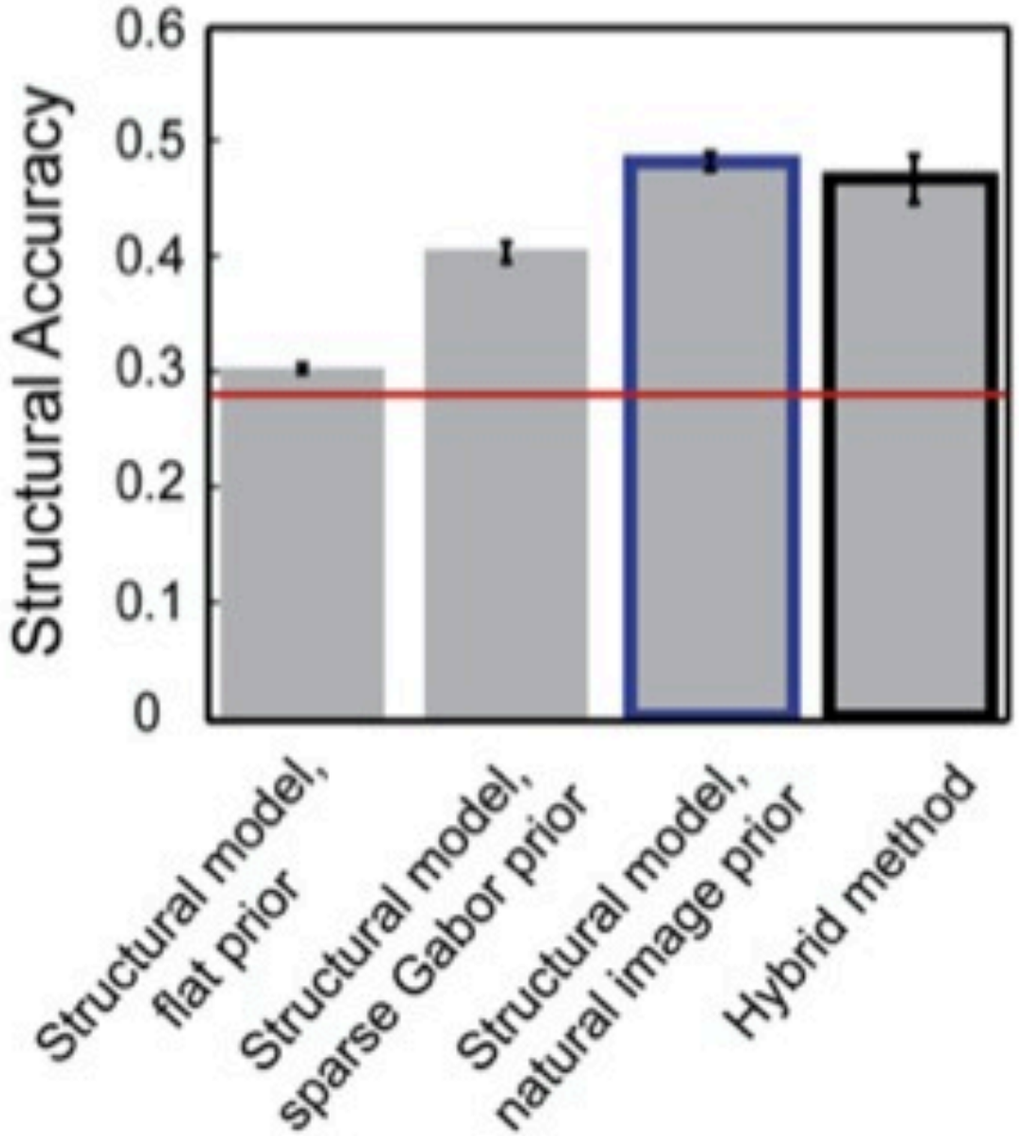


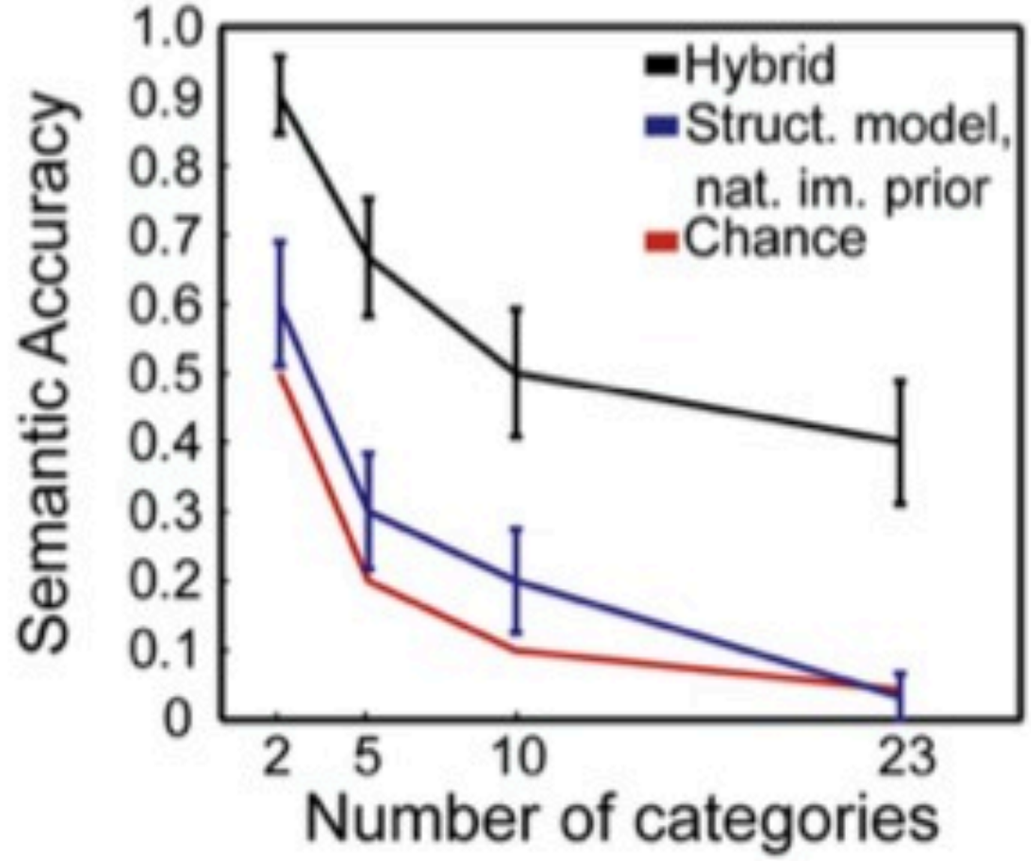
Results

structural encoding + natural image prior + semantic encoding

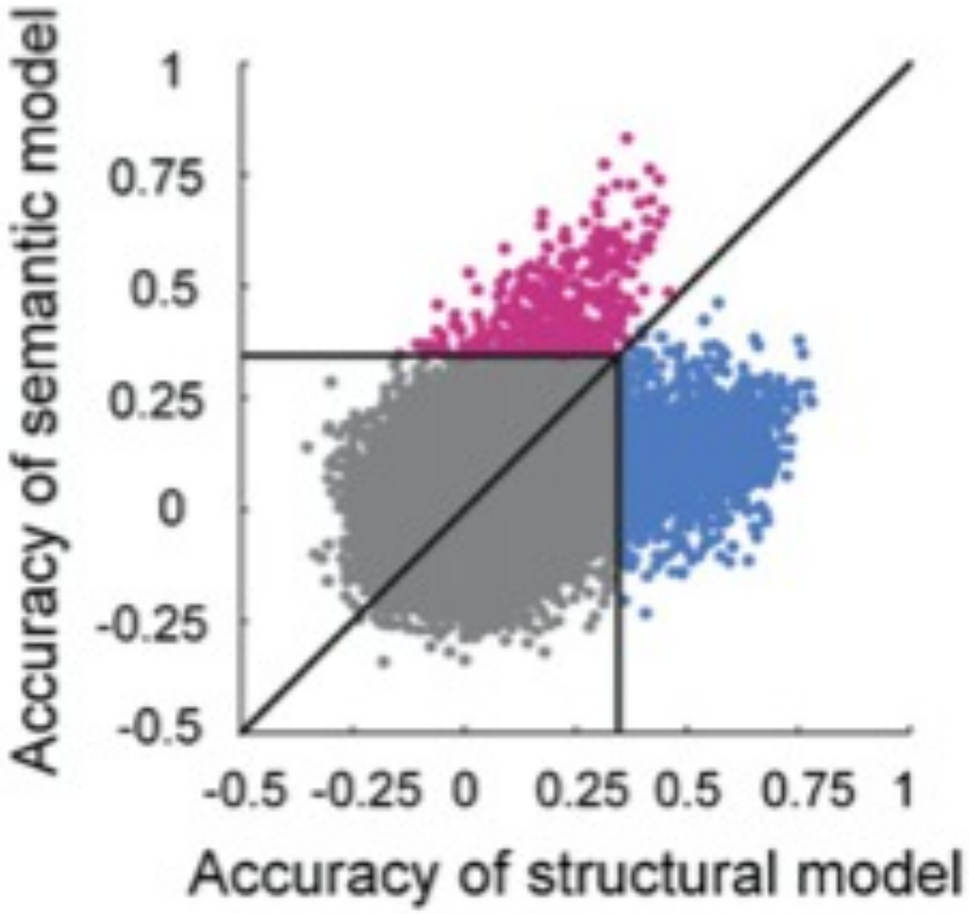


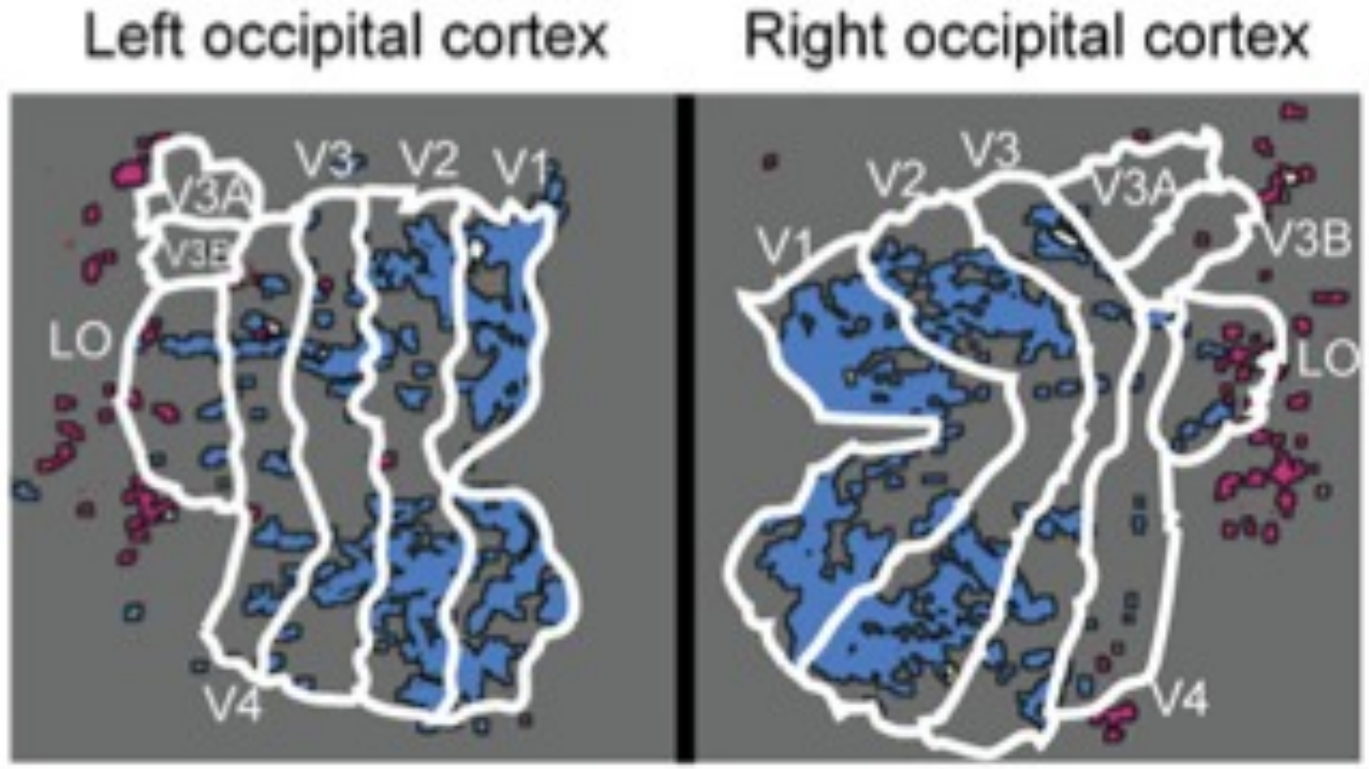
Subject TN


Structural Accuracy
0.6
0.5
0.4
0.3
0.2
0.1
0
Structural model, flat prior
Structural model, sparse Gabor prior
Structural model, natural image prior
Hybrid method

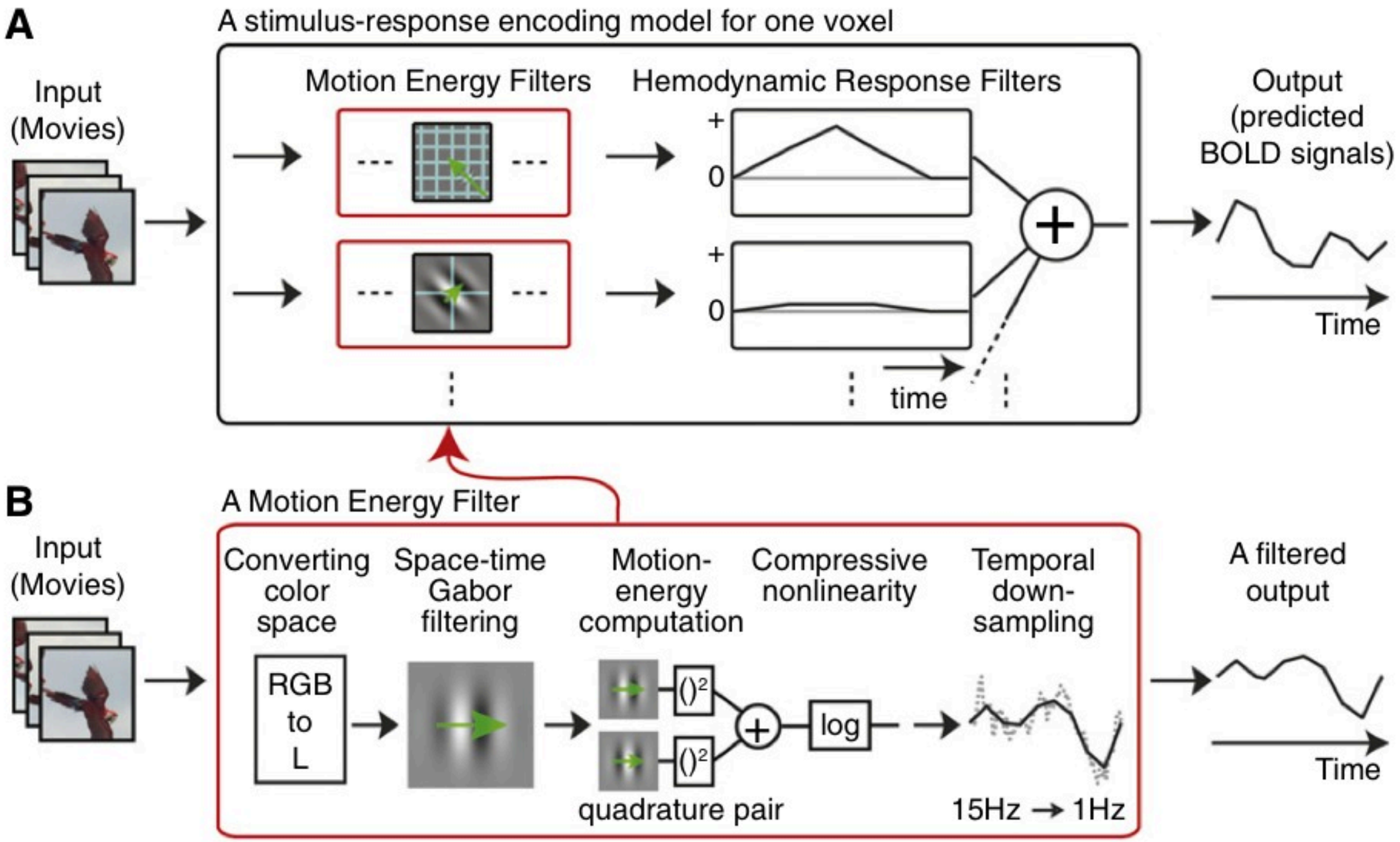

Semantic Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Hybrid
Struct. model, nat. im. prior
Chance
2
5
10
23
Number of categories

Structural versus semantic encoding voxels





Movie decoding



Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Bin Yu, Gallant JL. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. Current Biology. 2011 Sep. 21;:1–6.

Encoding model for voxel i

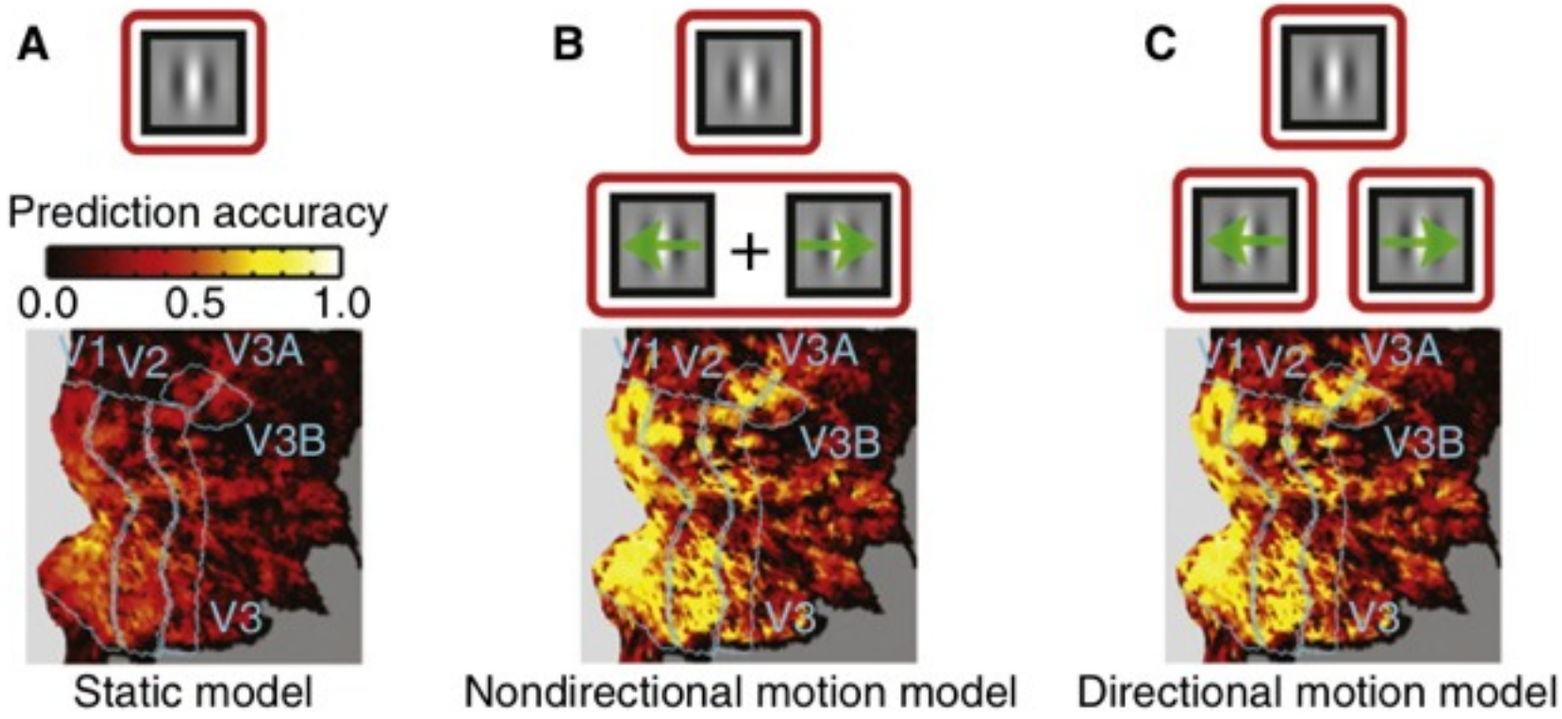
$$\hat{r}_i = s * w_i$$

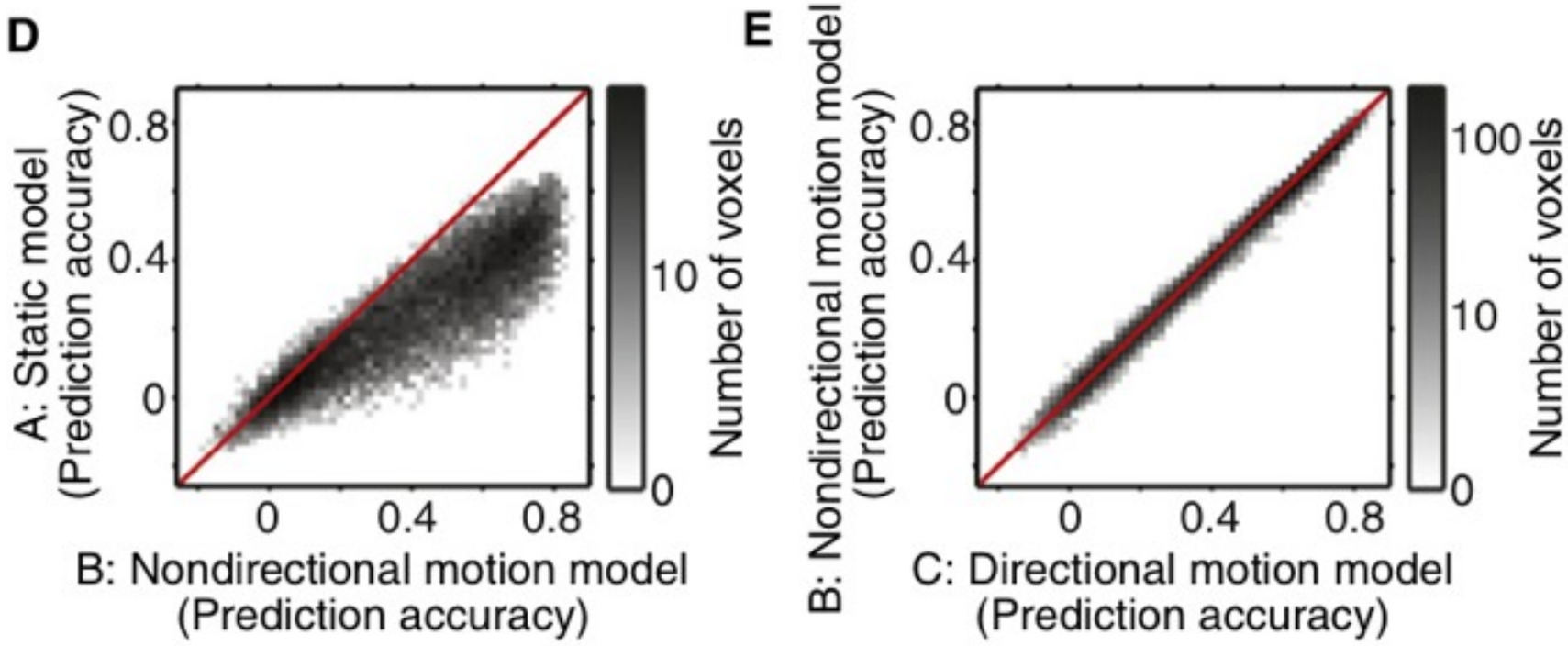
$$\hat{r}_i = \left[s_{d1} \; s_{d2} \; \cdots \; s_{dK} \right] * \begin{bmatrix} h_{i,d1} \\ h_{i,d2} \\ \vdots \\ h_{i,dK} \end{bmatrix}$$

response — motion filters at time dk — weights

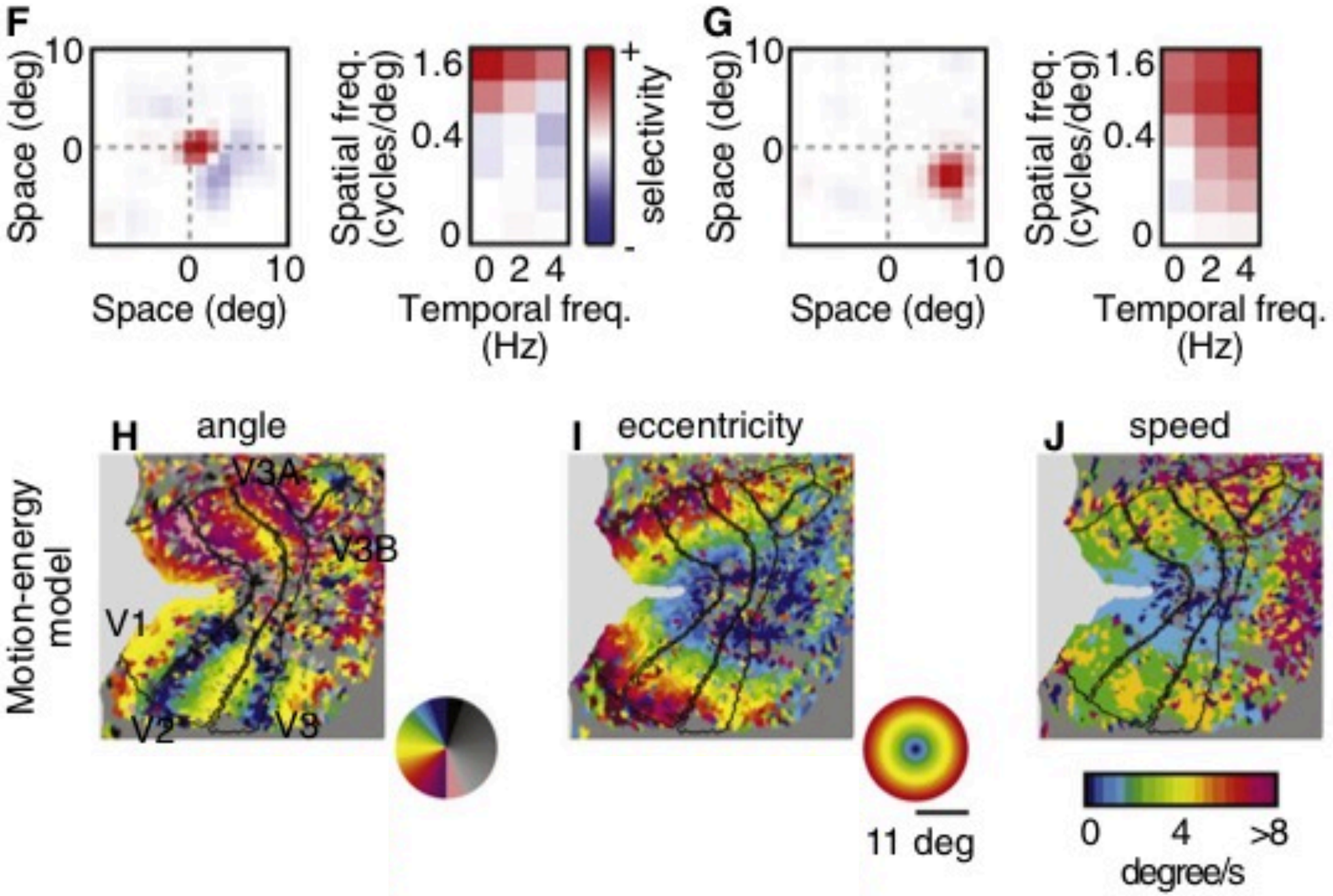
reconstruction again proceeds by identifying which movie clip in a set of 1000000 clips maximizes the likelihood.

Movie decoding

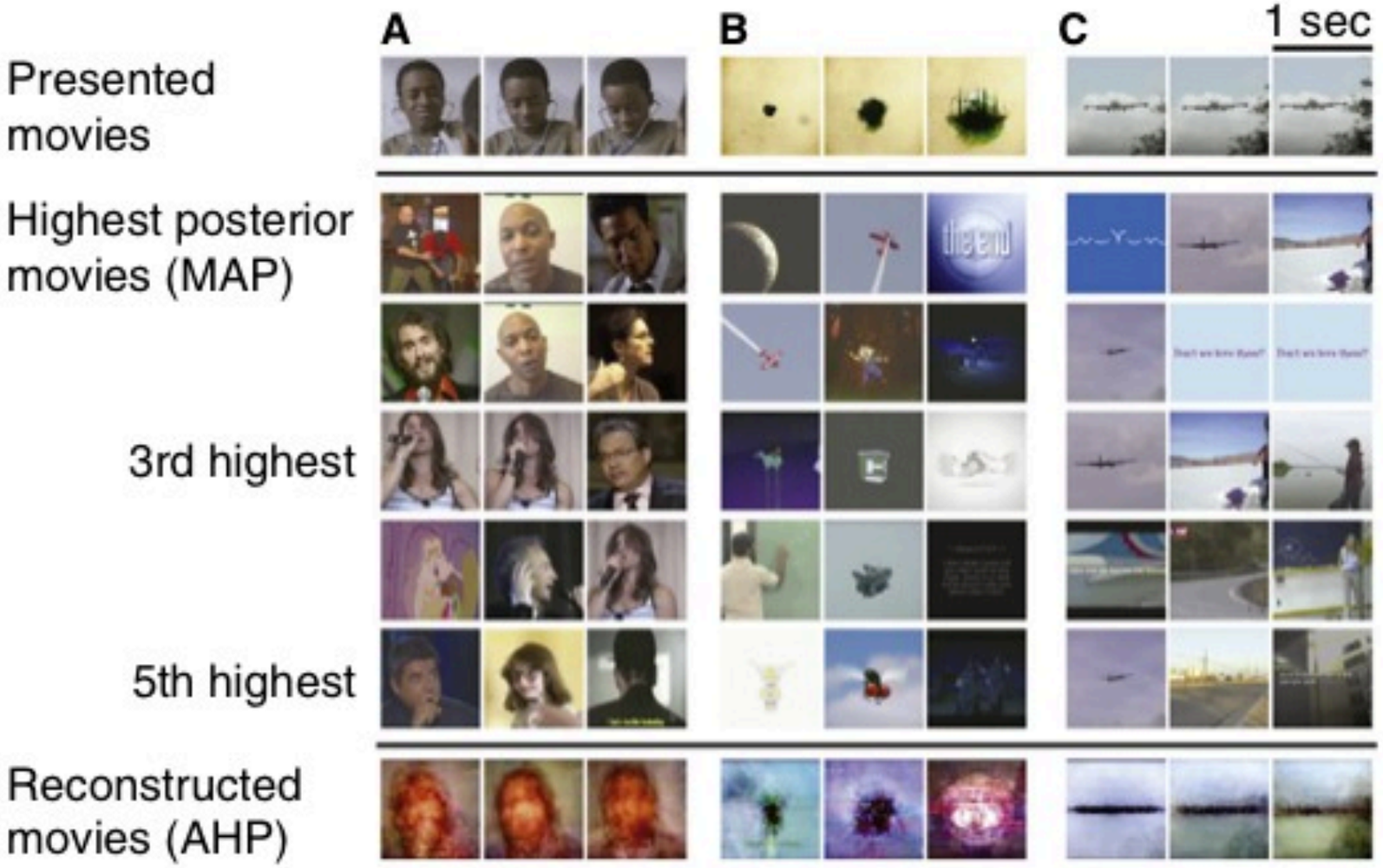


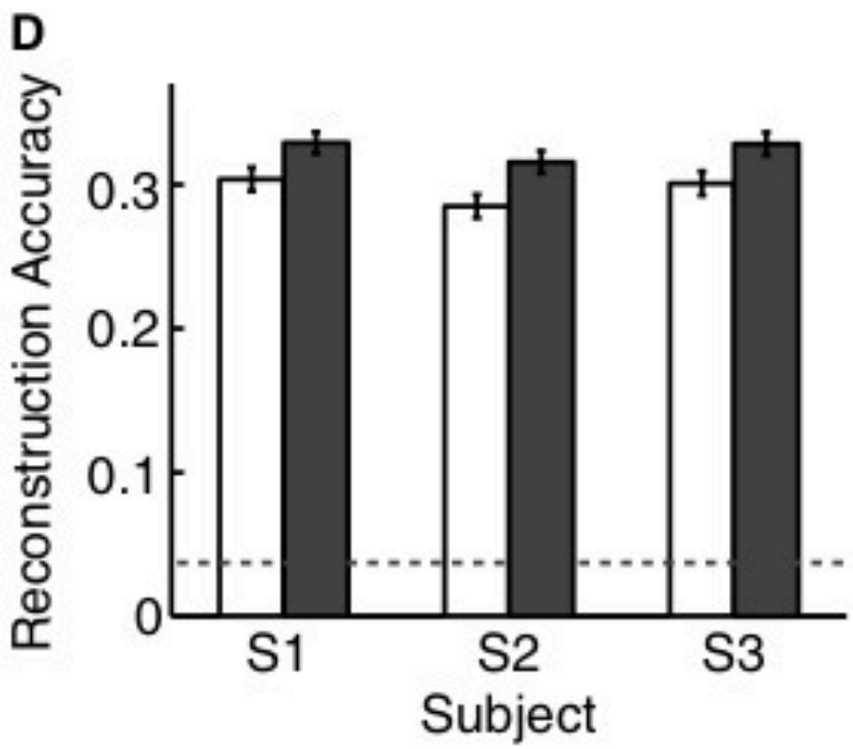


Movie decoding



Movie decoding





Presented clip

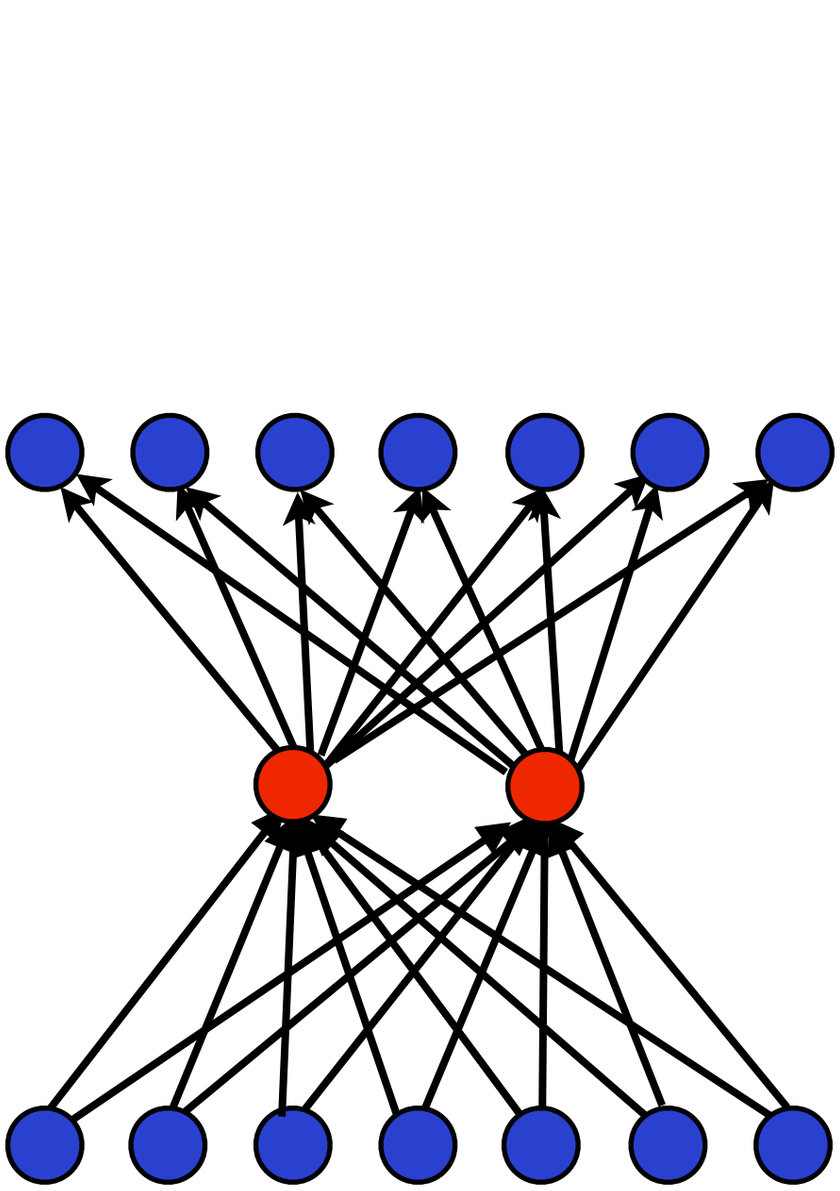
Clip reconstructed from brain activity

Reconstructing visual experiences from brain activity evoked by natural movies

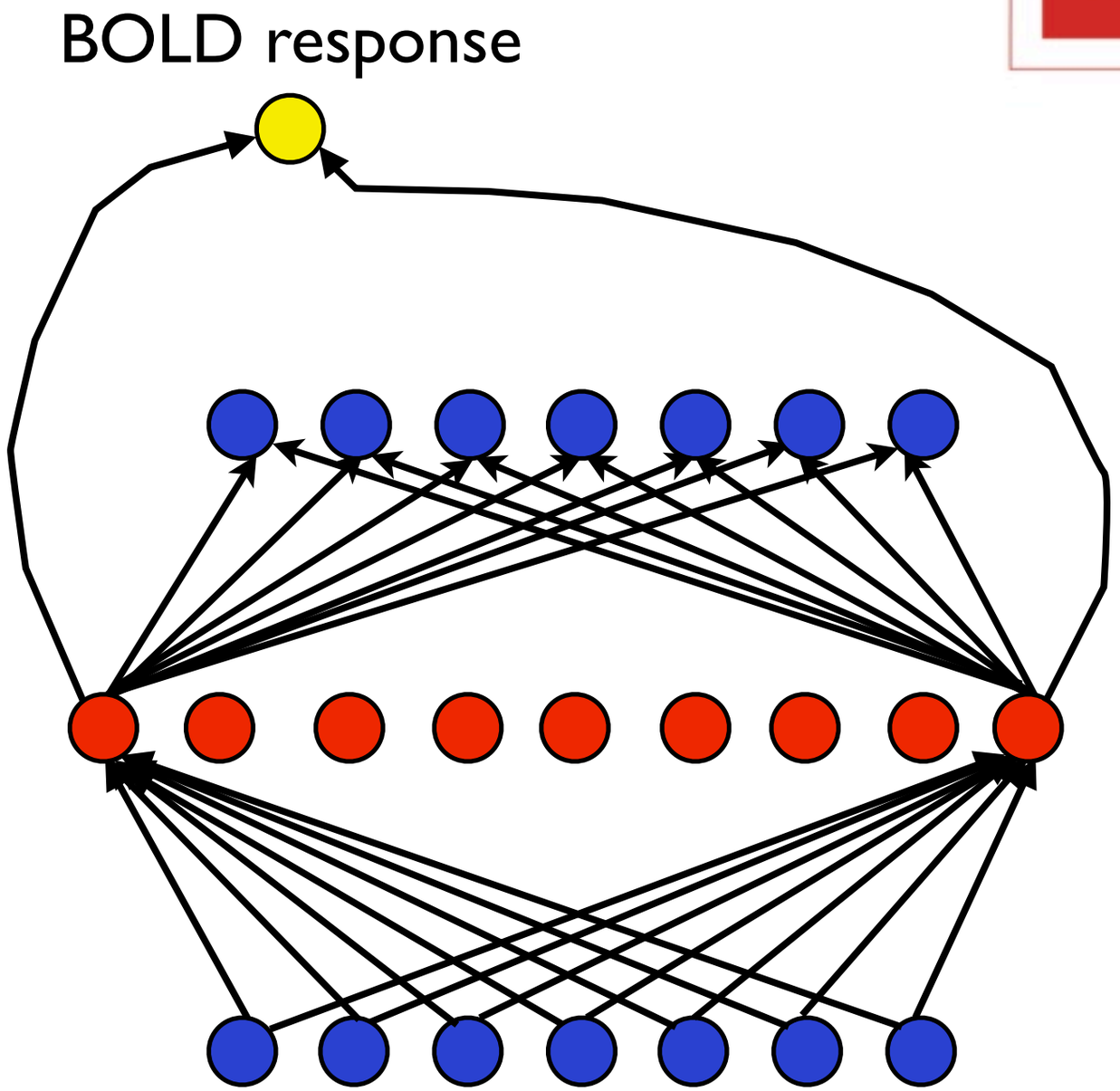
Shinji Nishimoto, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, Jack L. Gallant

Supplemental movie S1

Current work in generative modelling

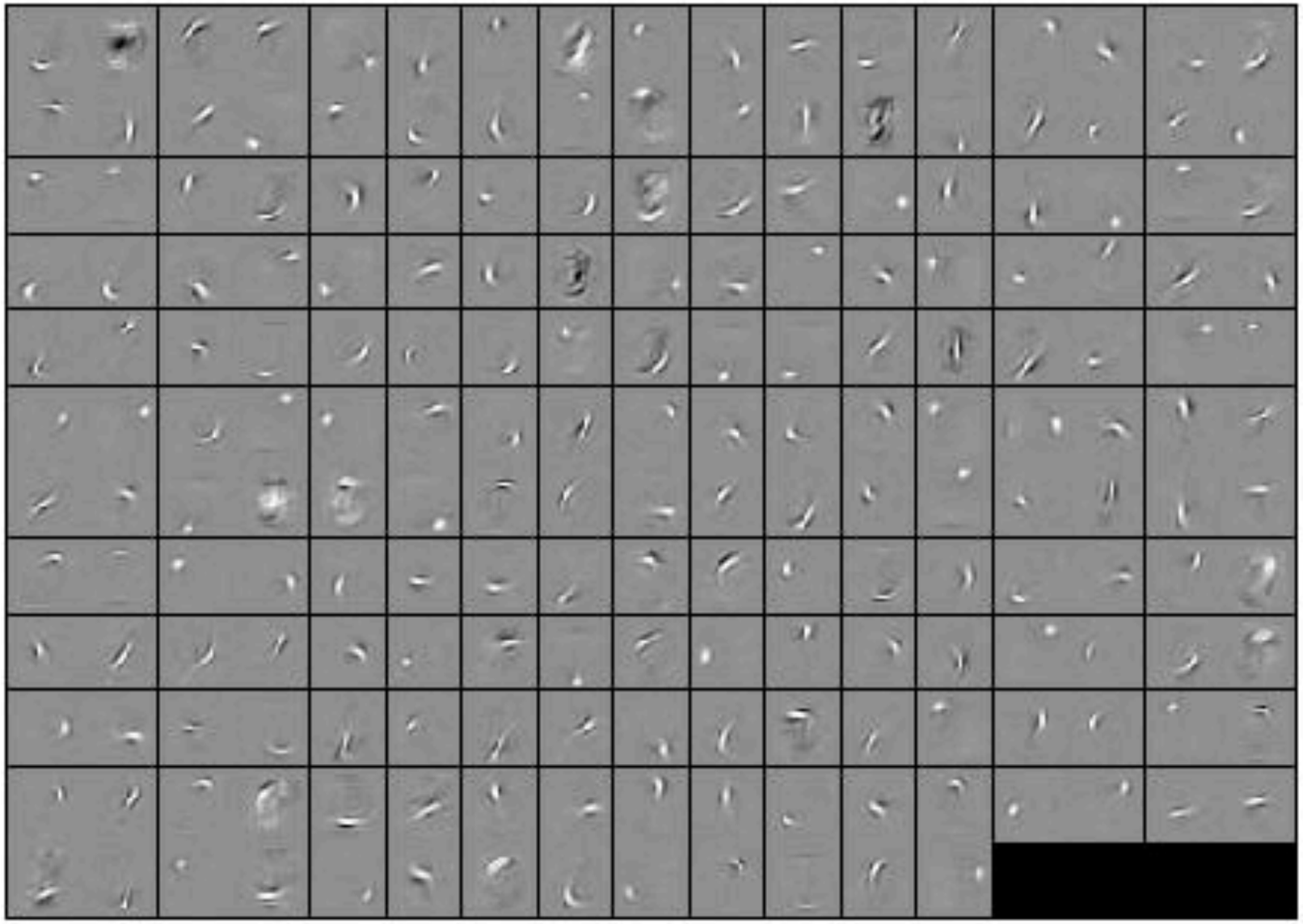


auto-encoder



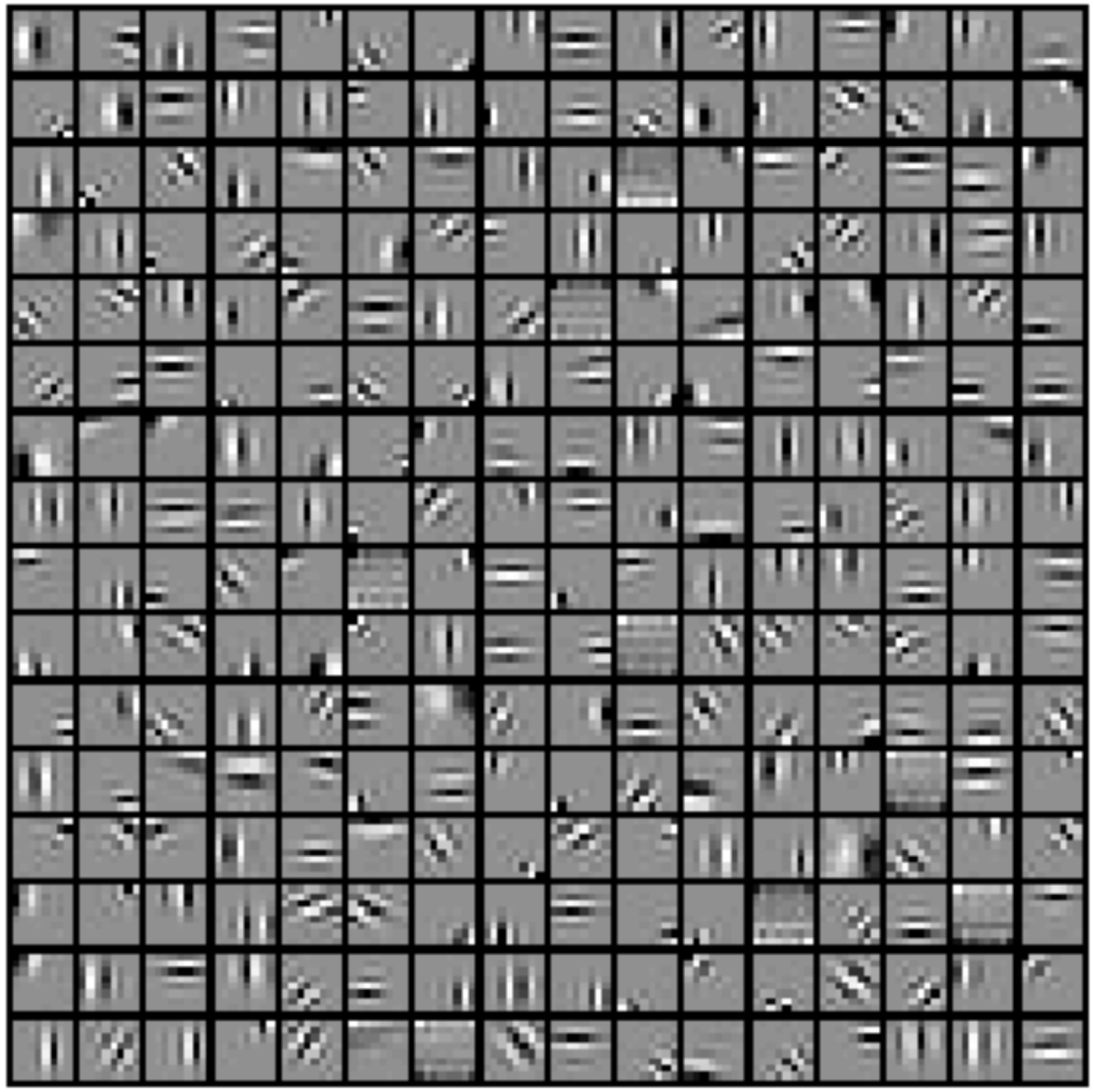
sparse auto-encoder

Current work in generative modelling



MNIST filter trained on full 28 x 28 digits (after 400 iterations)

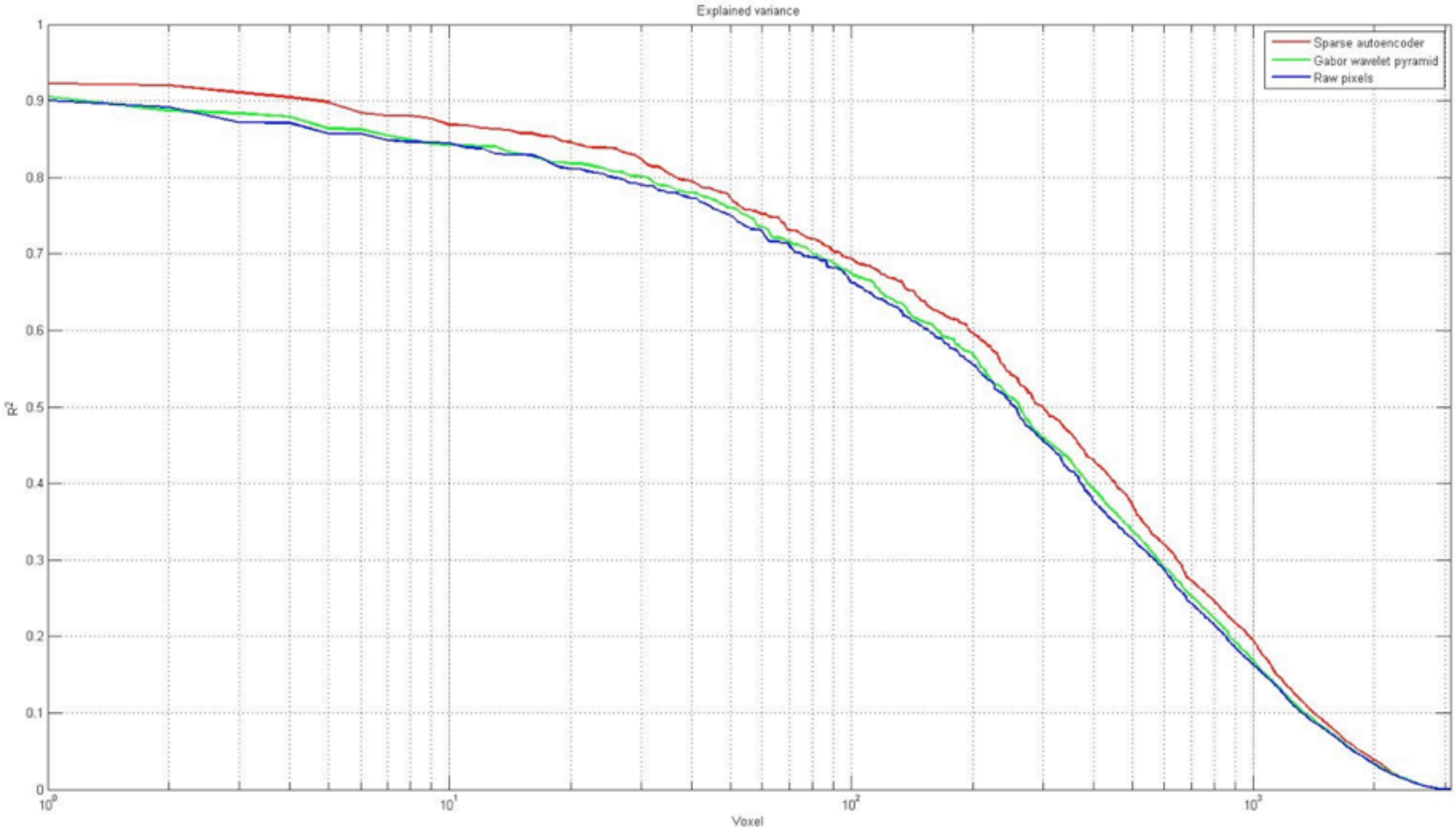
Current work in generative modelling



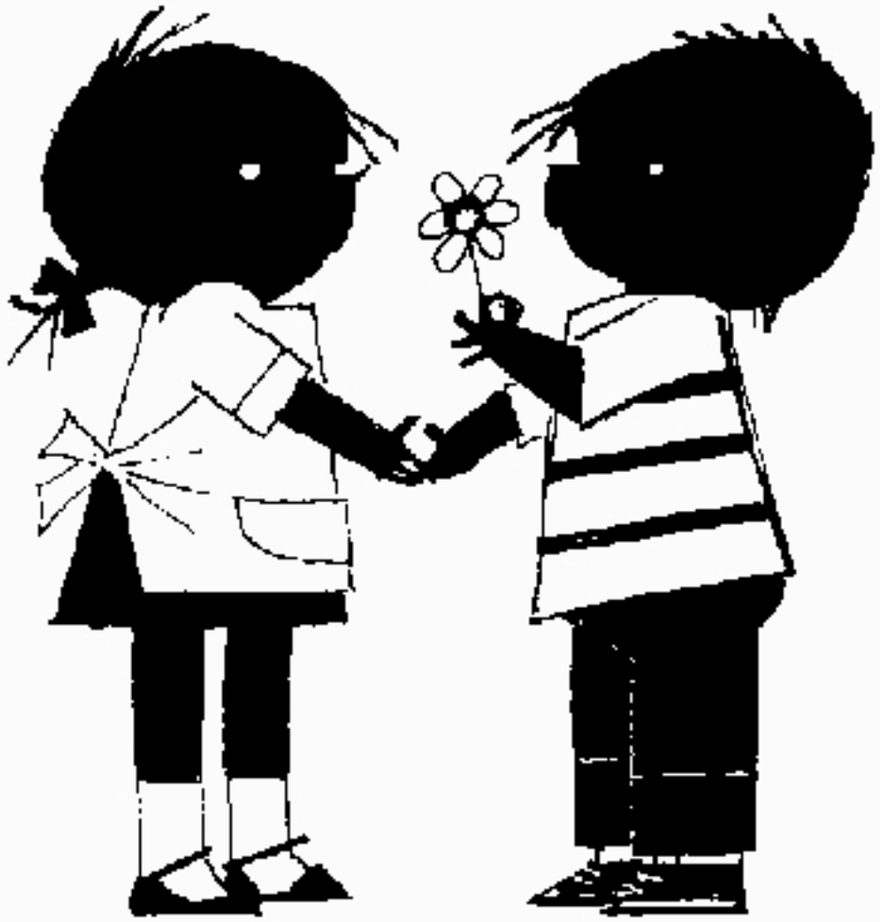
Natural image filters trained on 8 x 8 patches sampled from full 128 x 128 natural images (after 200 iterations)

Current work in generative modelling

Current work in generative modelling



Decoding of Jip en Janneke stories



Can we predict:
- the semantic content of individual stories?
- which story is presented based on functional connectivity?

- van Gerven MAJ, de Lange FP, Heskes T. Neural decoding with hierarchical generative models. Neural Comput. 2010;22(12):3127–42.

Neural Decoding with Hierarchical Generative Models

Marcel A. J. van Gerven
marcelge@cs.ru.nl
Radboud University Nijmegen, Institute for Computing and Information Sciences, 6525 AJ Nijmegen, the Netherlands, and Radboud University Nijmegen, Institute for Brain, Cognition and Behaviour, 6525 EN Nijmegen, the Netherlands

Floris P. de Lange
florisdelange@gmail.com
Radboud University Nijmegen, Institute for Brain, Cognition and Behaviour, 6525 EN Nijmegen, the Netherlands

Tom Heskes
t.heskes@science.ru.nl
Radboud University Nijmegen, Institute for Computing and Information Sciences, 6525 AJ Nijmegen, the Netherlands, and Radboud University Nijmegen, Institute for Brain, Cognition and Behaviour, 6525 EN Nijmegen, the Netherlands

Learning goals

- Hierarchical approach
- Bayesian reconstruction

Coming week

Practical

Continue practical

Lecture

Connectivity analysis