Sparse coding

Barlow (1972)

Perception, 1972, volume 1, pages 371-394

Single units and sensation: A neuron doctrine for perceptual psychology?

H B Barlow
Department of Physiology-Anatomy, University of California, Berkeley, California 94720
Received 6 December 1972

Abstract. The problem discussed is the relationship between the firing of single neurons in sensory pathways and subjectively experienced sensations. The conclusions are formulated as the following five dogmas:
1. To understand nervous function one needs to look at interactions at a cellular level, rather than either a more macroscopic or microscopic level, because behaviour depends upon the organized

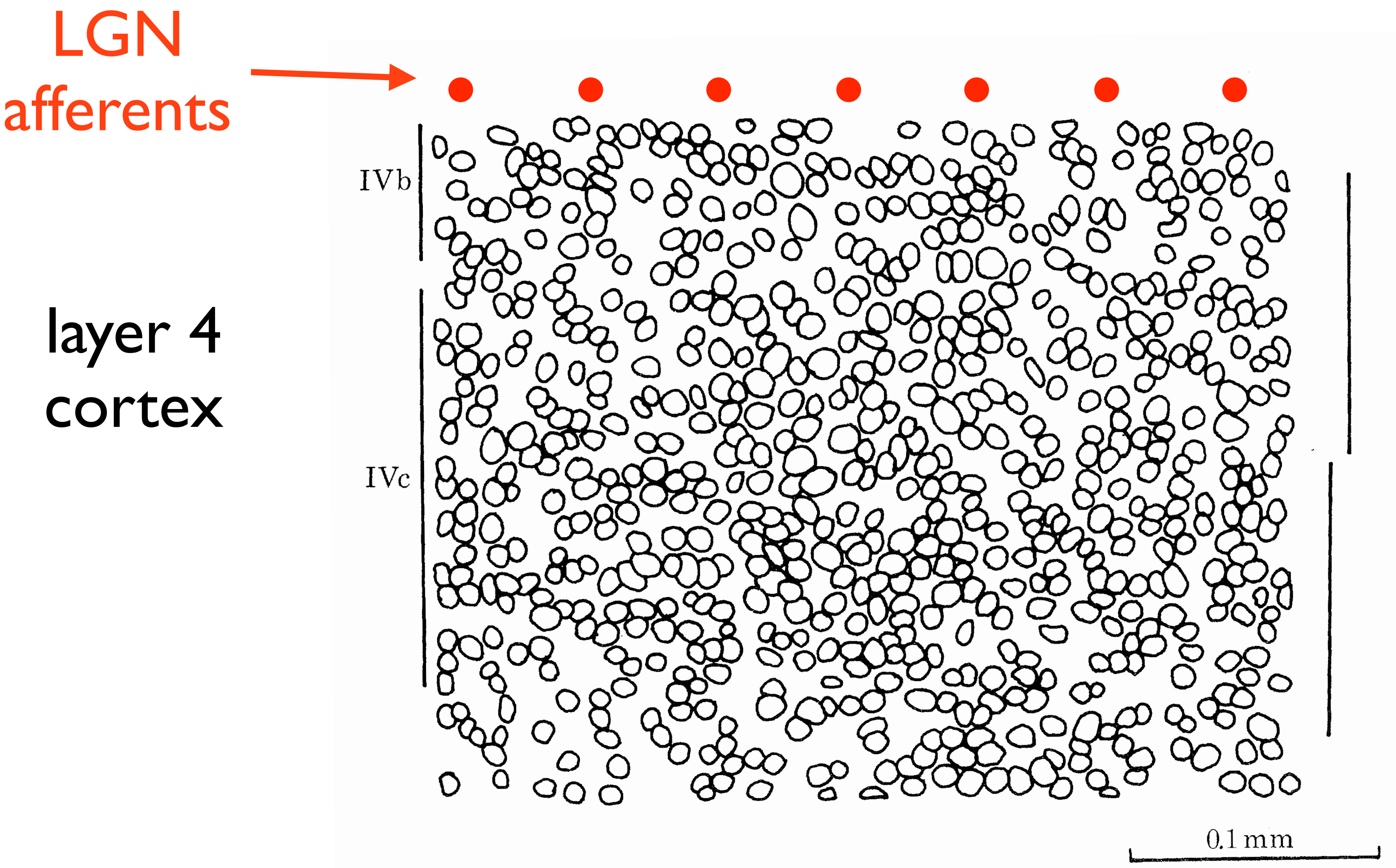
2. The sensory system is organized to achieve as complete a representation of the sensory stimulus as possible with the minimum number of active neurons.

neurons, each of which corresponds to a pattern of external events of the order of complexity of the events symbolized by a word.
5. High impulse frequency in such neurons corresponds to high certainty that the trigger feature is present.
The development of the concepts leading up to these speculative dogmas, their experimental basis, and some of their limitations are discussed.

Barlow (1972)

The second dogma goes beyond the evidence, but it attempts to make sense out of it. It asserts that the overall direction or aim of information processing in higher sensory centres is to represent the input as completely as possible by activity in as few neurons as possible (Barlow, 1961, 1969b). In other words, not only the proportion but also the actual number of active neurons, K, is reduced, while as much information as possible about the input is preserved.

V1 is highly overcomplete

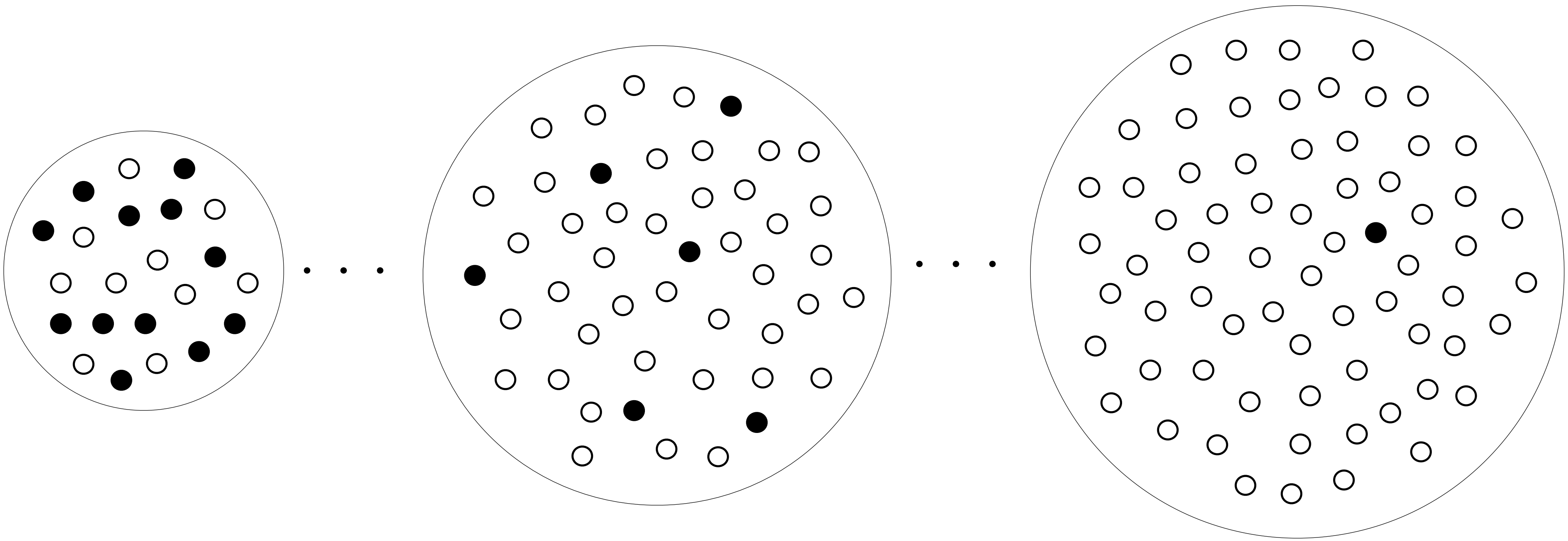


Barlow (1981)

Dense codes
(e.g., ascii)

Sparse, distributed codes

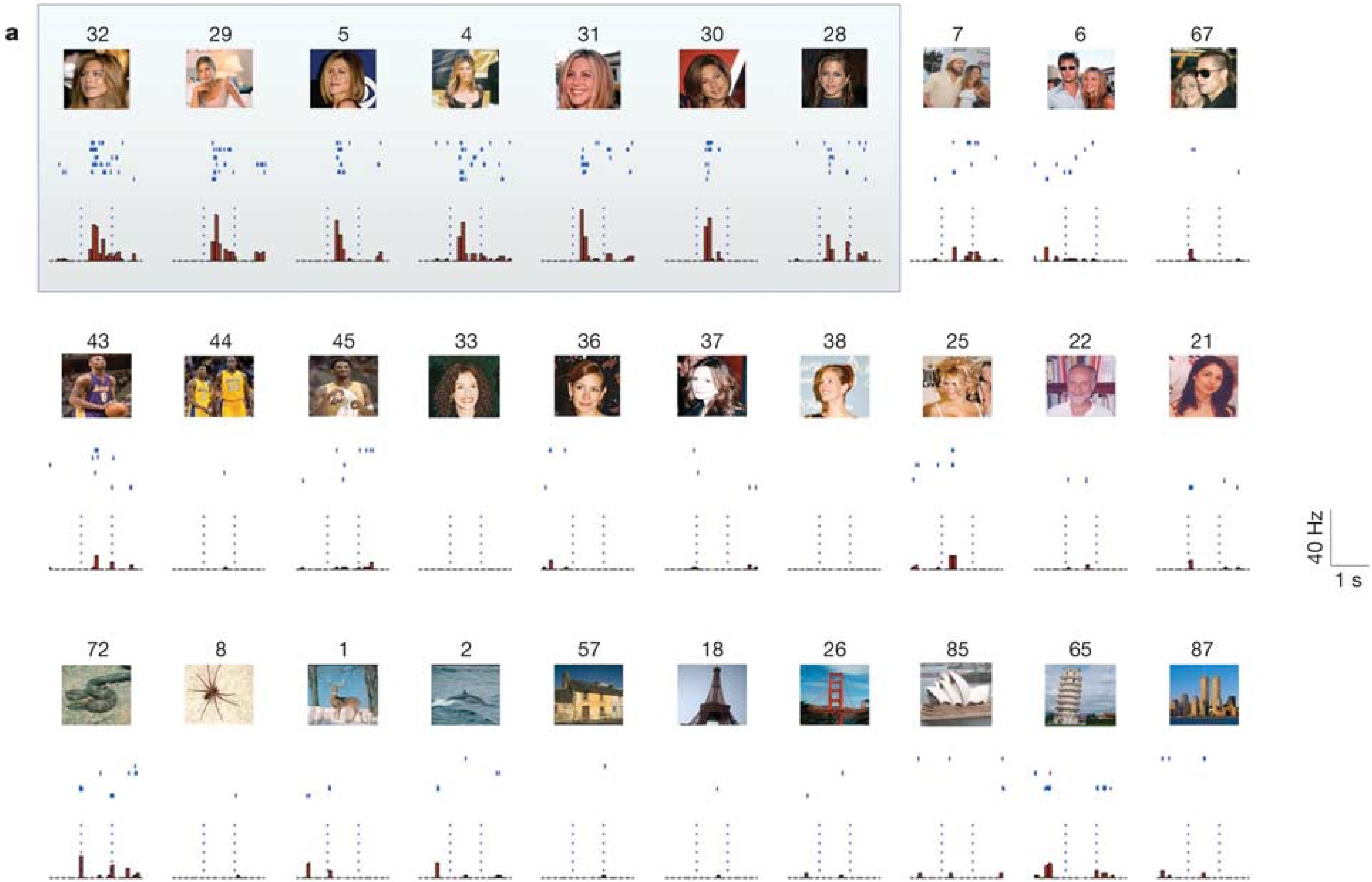
Local codes
(e.g., grandmother cells)



2^N $\binom{N}{K}$ N

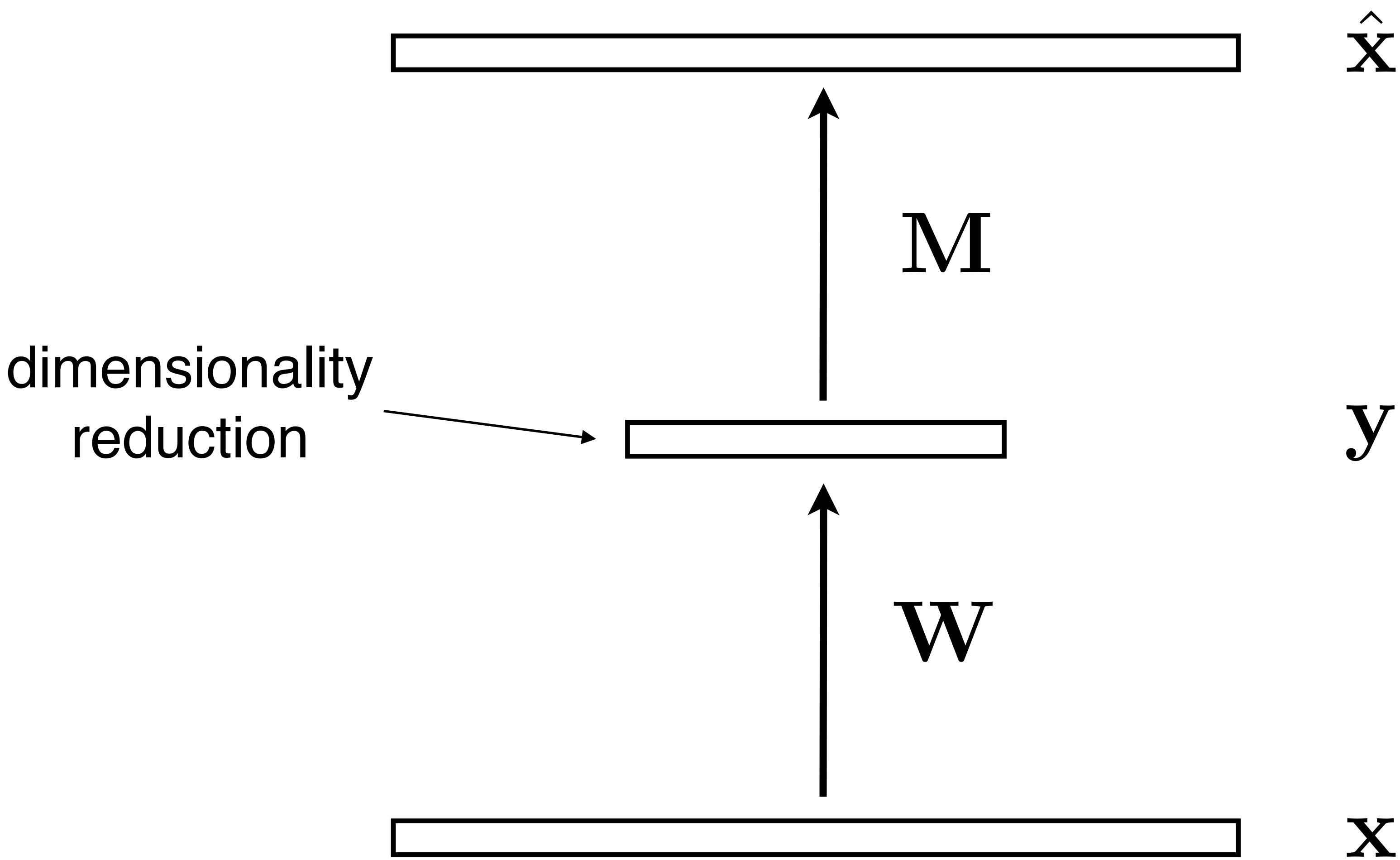
Evidence for grandmother cells?

(Quiroga, Reddy, Kreiman, Koch & Fried, *Nature* 2005)



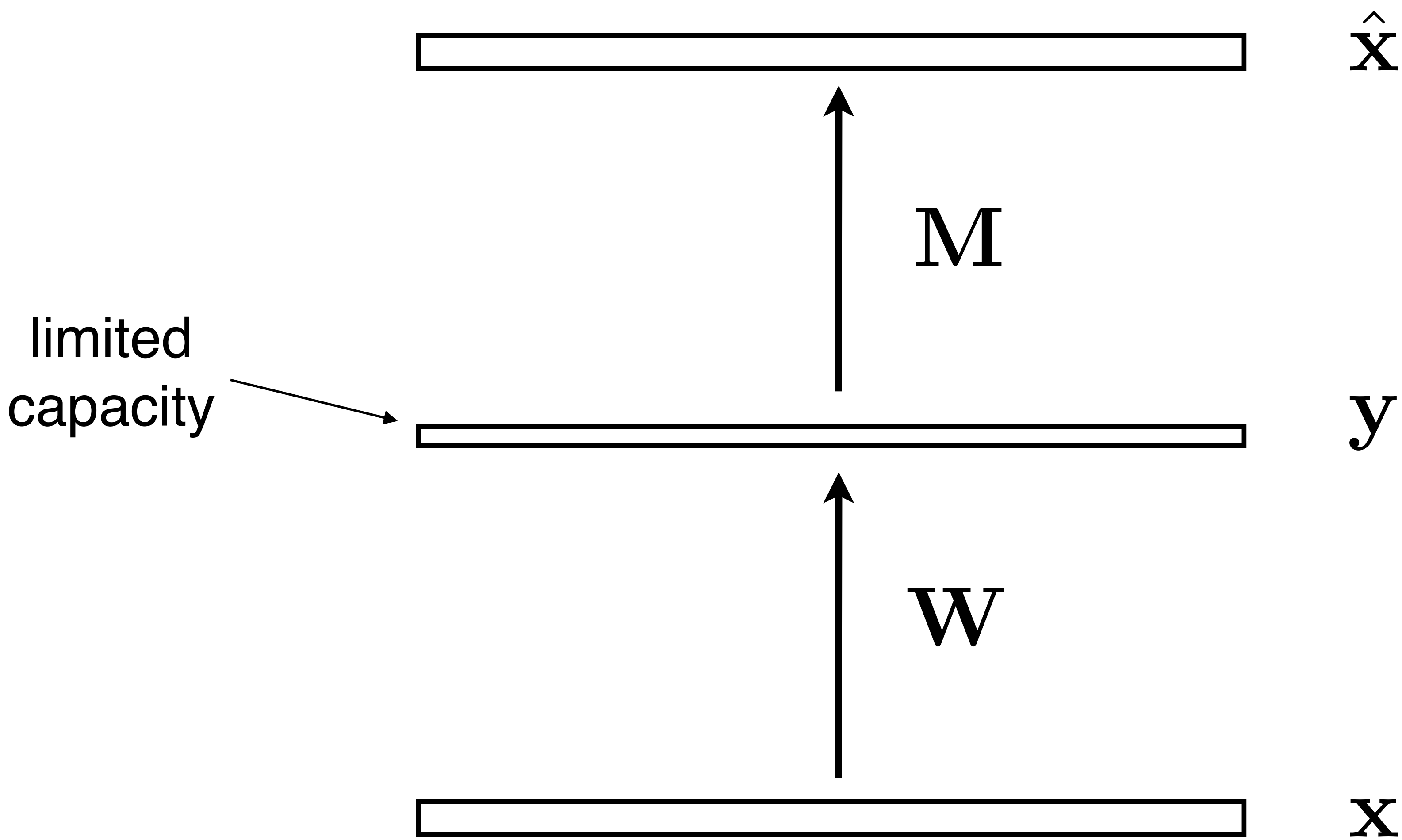
Autoencoder networks

$$\min_{\mathbf{W},\mathbf{M}} \left|\mathbf{x} - \hat{\mathbf{x}}\right|^2$$



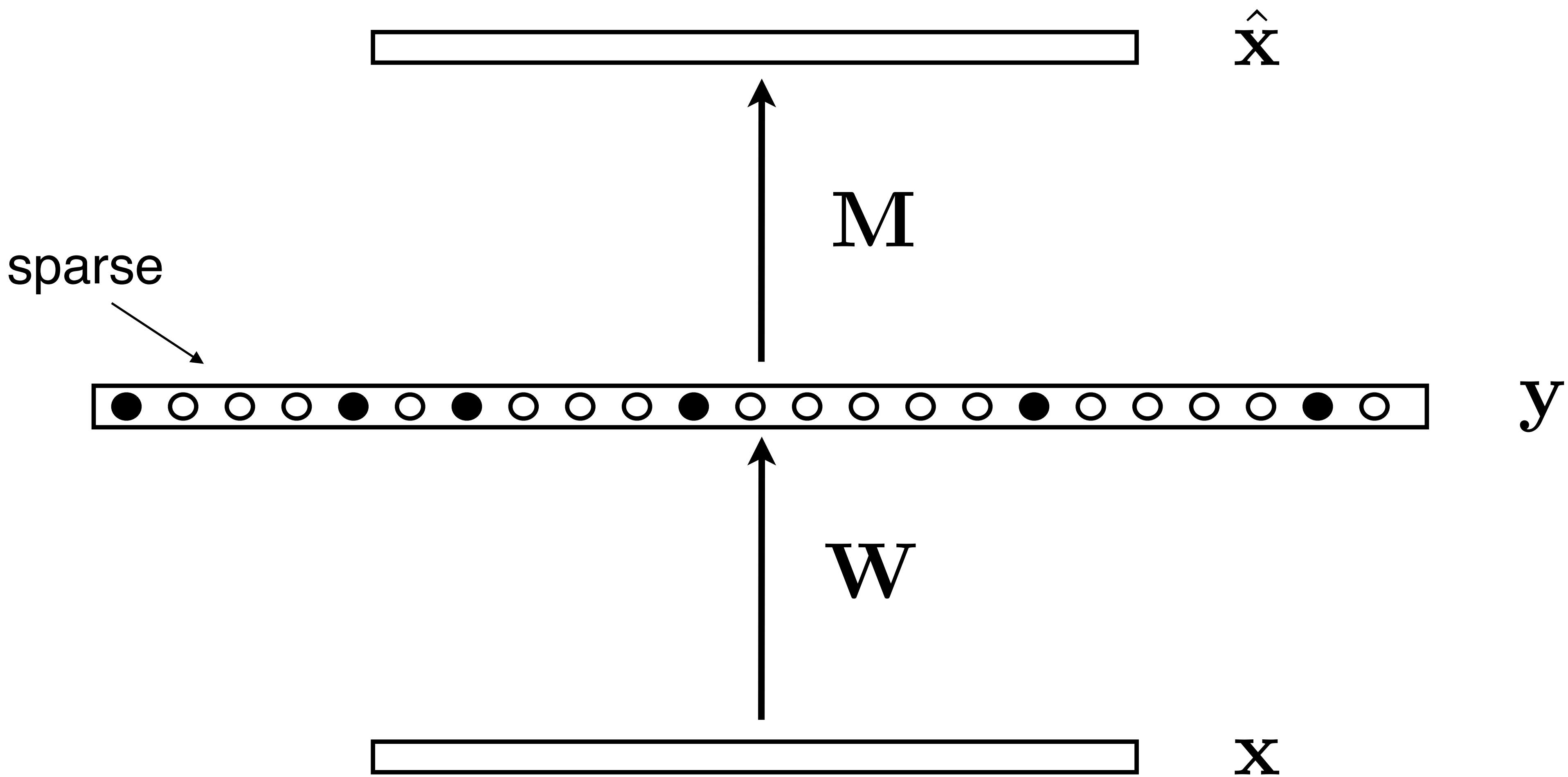
Autoencoder networks

$$\min_{\mathbf{W},\mathbf{M}} \left|\mathbf{x} - \hat{\mathbf{x}}\right|^2$$

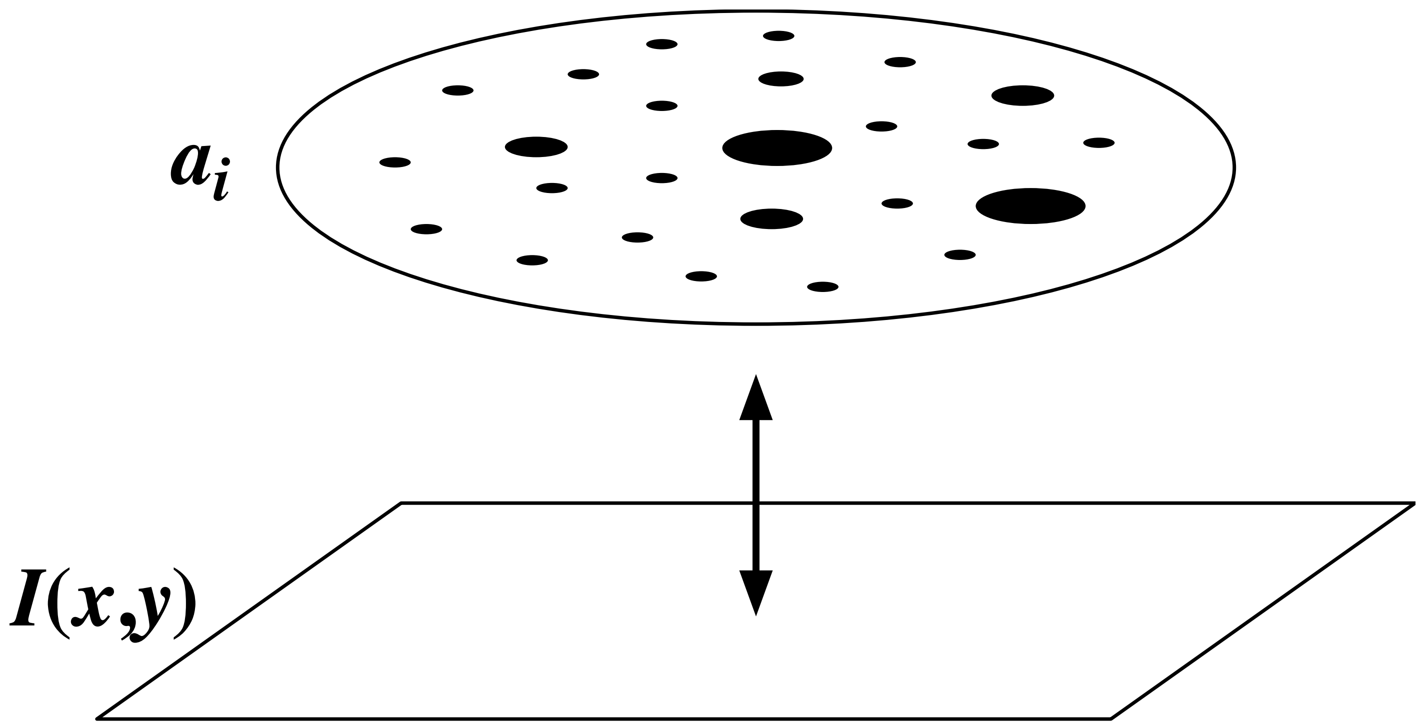


Autoencoder networks

$$\min_{\mathbf{W},\mathbf{M}} |\mathbf{x} - \hat{\mathbf{x}}|^2$$



How to learn sparse, distributed representations?



Biol. Cybern. 64, 165–170 (1990)

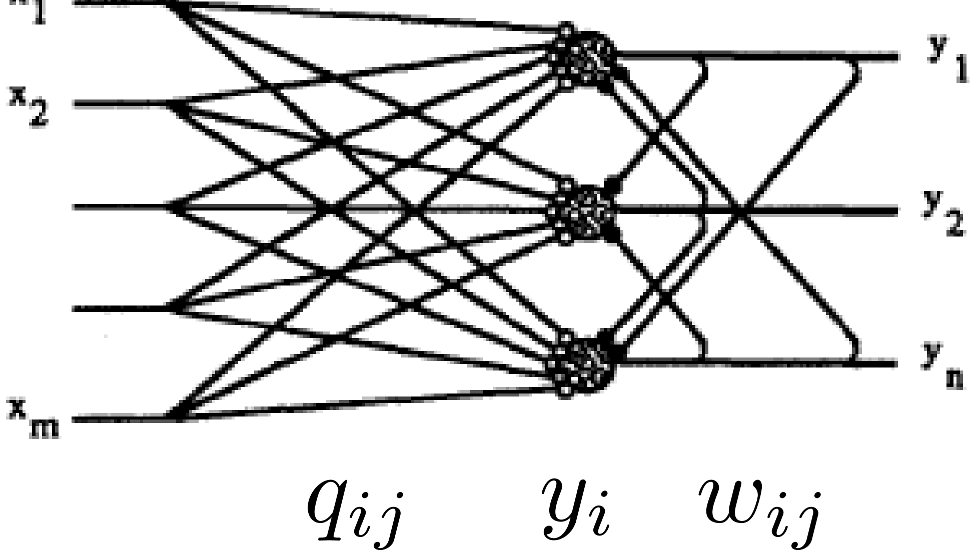
Biological Cybernetics

Forming sparse representations by local anti-Hebbian learning

P. Földiák

Physiological Laboratory, University of Cambridge, Downing Street, Cambridge CB2 3EG, United Kingdom

$$\frac{dy_i^*}{dt} = f\left(\sum_{j=1}^{m} q_{ij} x_j + \sum_{j=1}^{n} w_{ij} y_j^* - t_i\right) - y_i^*$$



anti-Hebbian rule–

$$\Delta w_{ij} = -\alpha(y_i y_j - p^2)$$

(if $i = j$ or $w_{ij} > 0$ then $w_{ij} := 0$)

Hebbian rule–

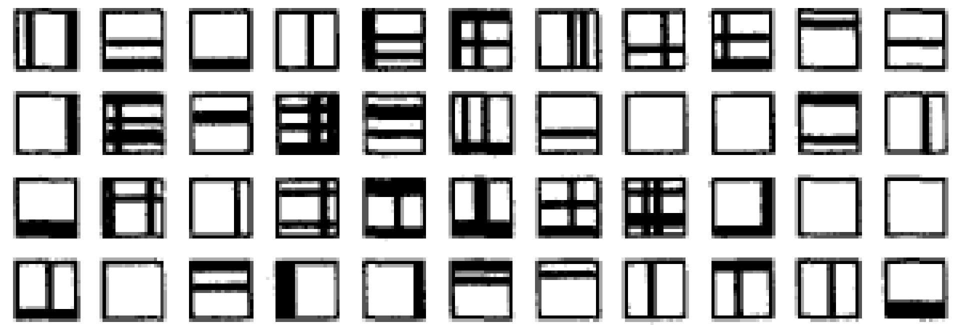
$$\Delta q_{ij} = \beta y_i (x_j - q_{ij})$$

threshold modification–

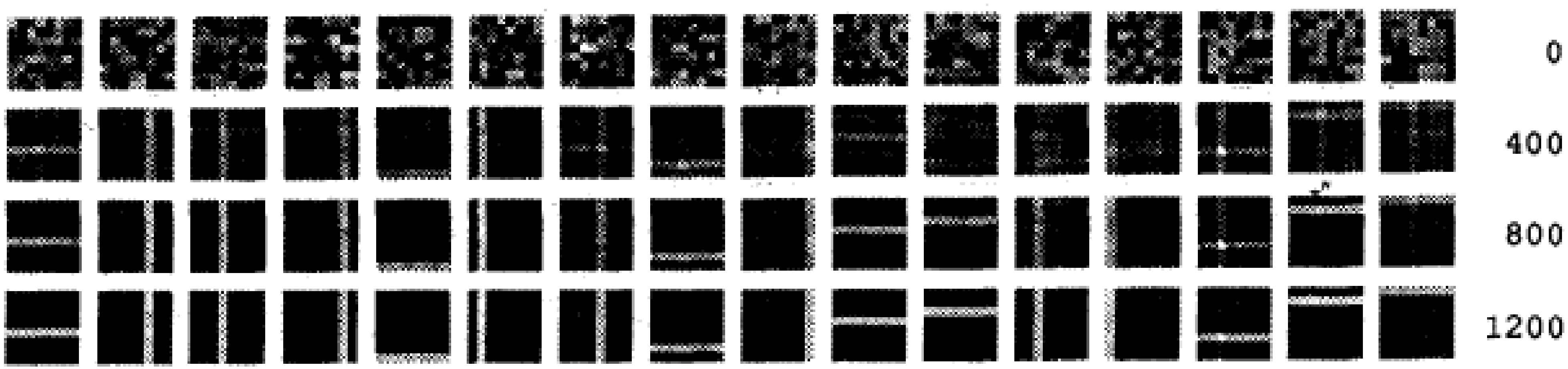
$$\Delta t_i = \gamma(y_i - p) .$$

Learning lines

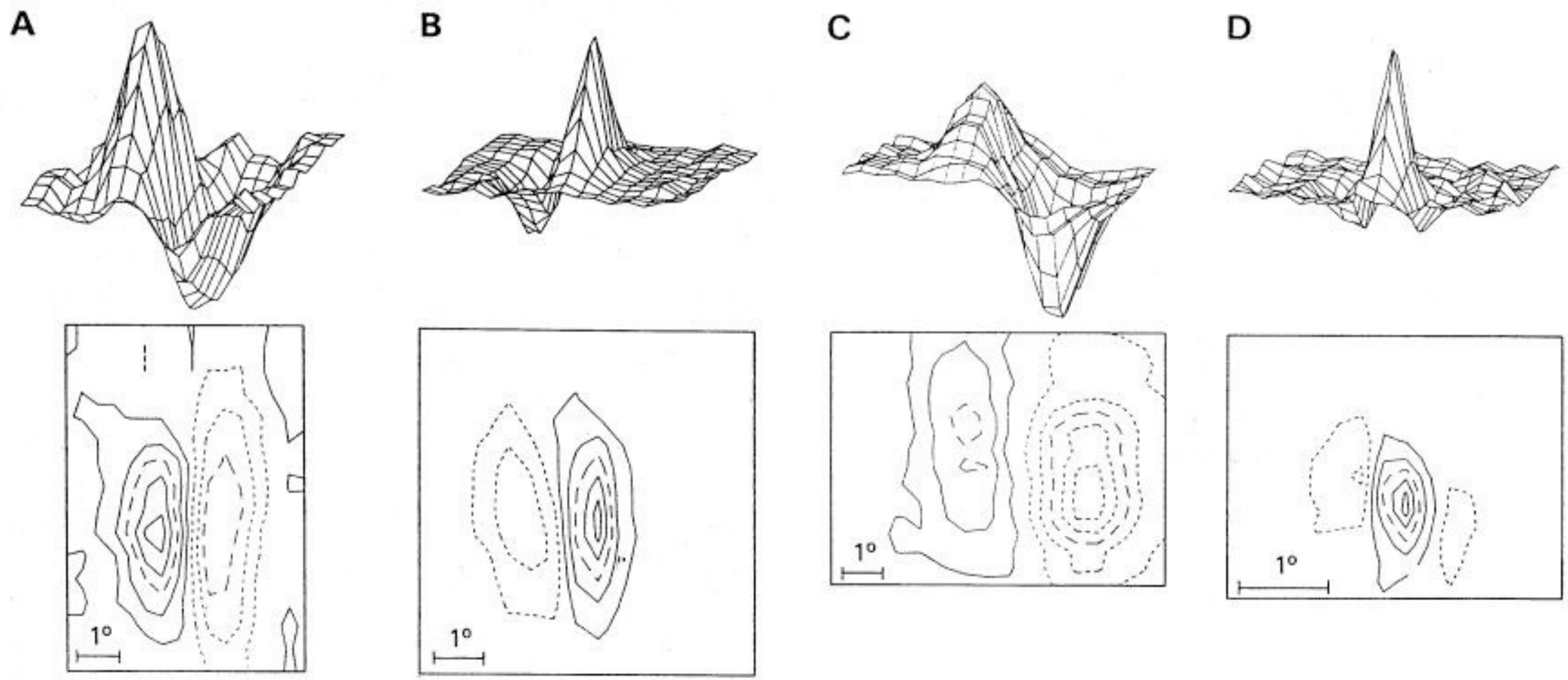
Input patterns:



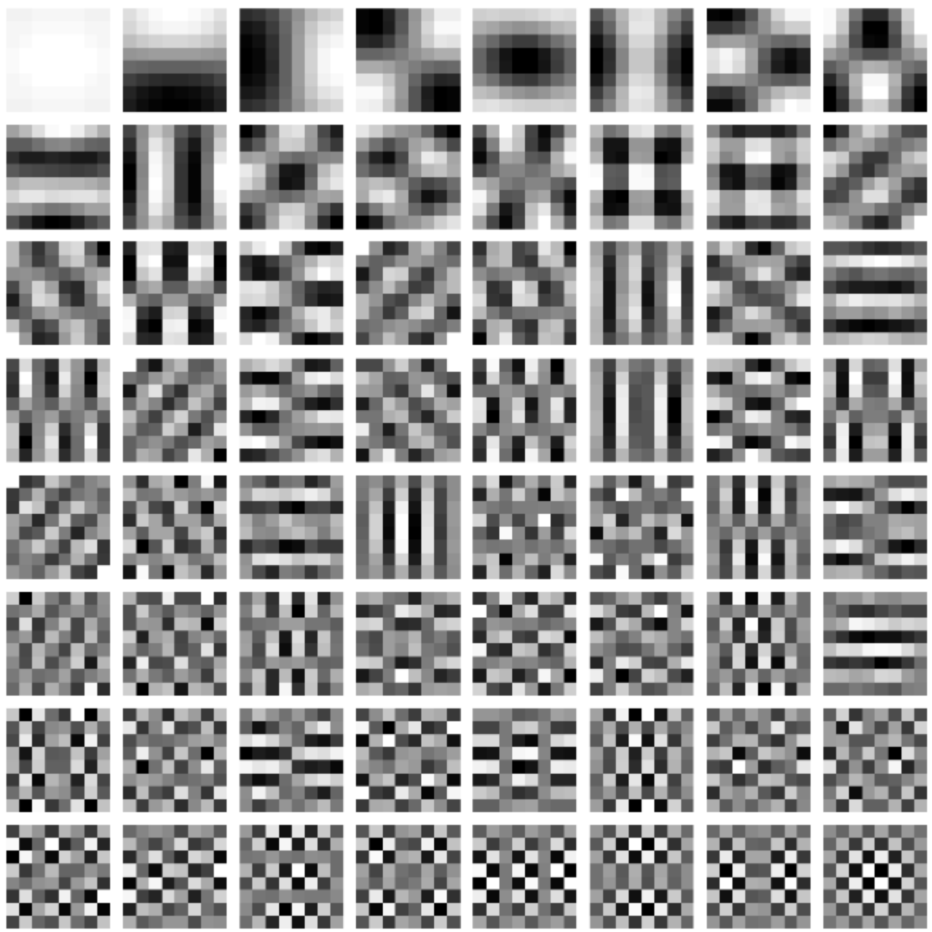
Learned weights:



V1 simple-cell receptive fields are localized, oriented, and bandpass. Why?



Principal components of natural image patches (8 x 8 pixels)

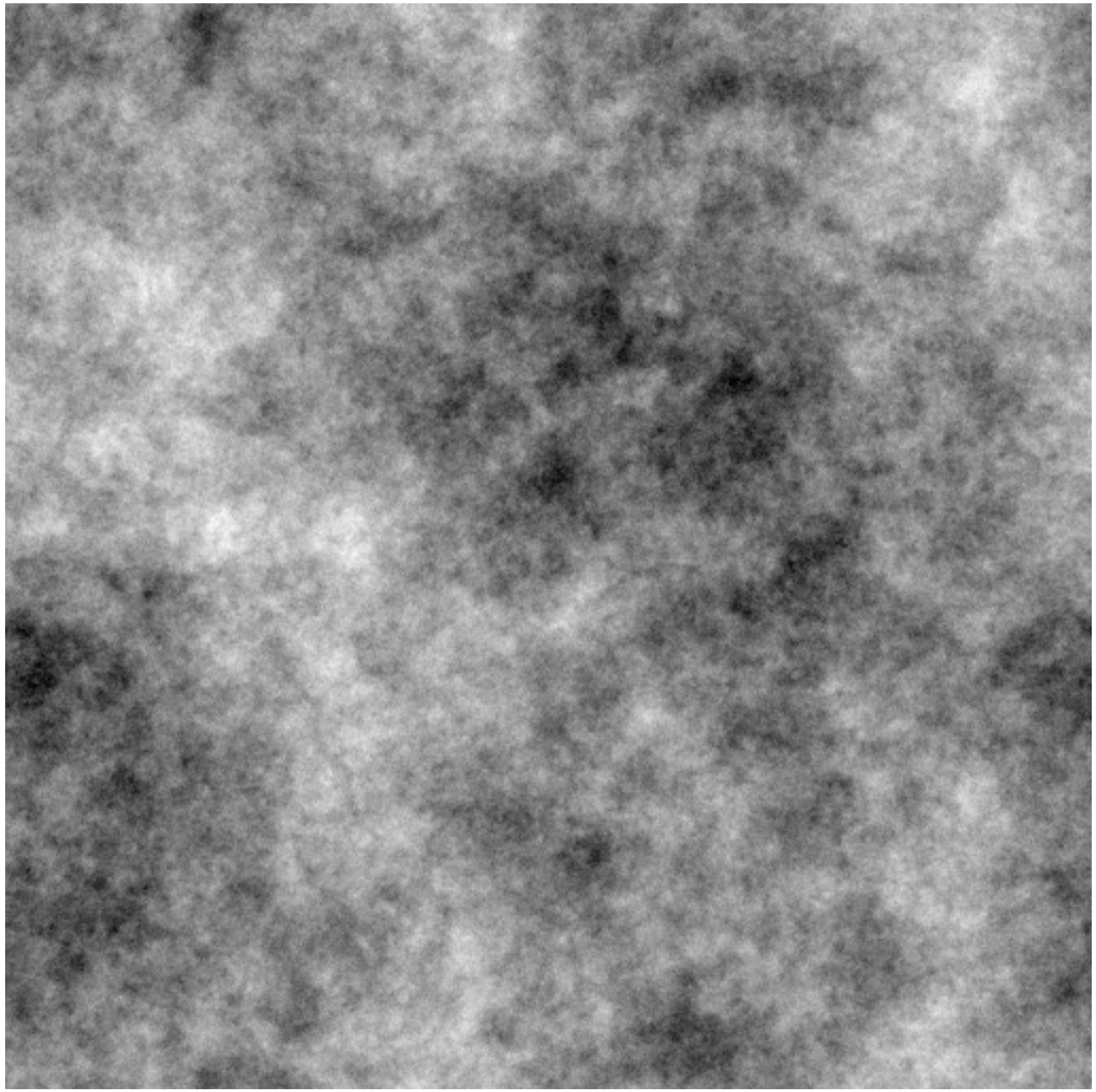


- Not localized
- Not oriented

PCA is incapable of learning about localized, oriented structure in images.

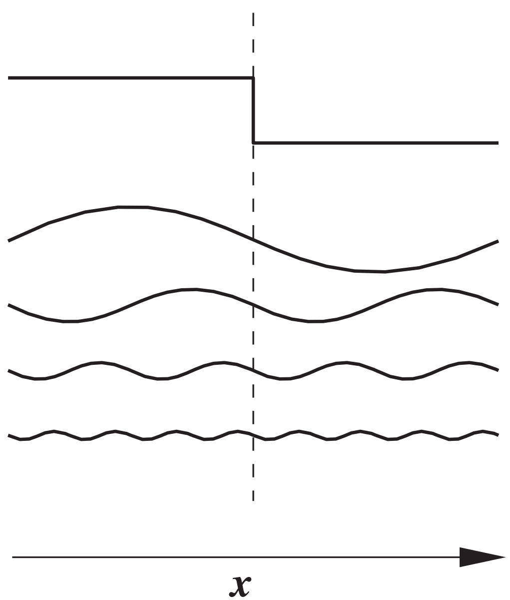
1/f noise

(what the world looks like if all you care about are pairwise correlations)

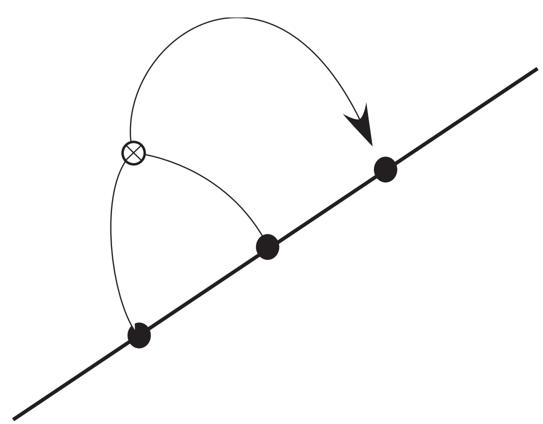


Higher-order image statistics

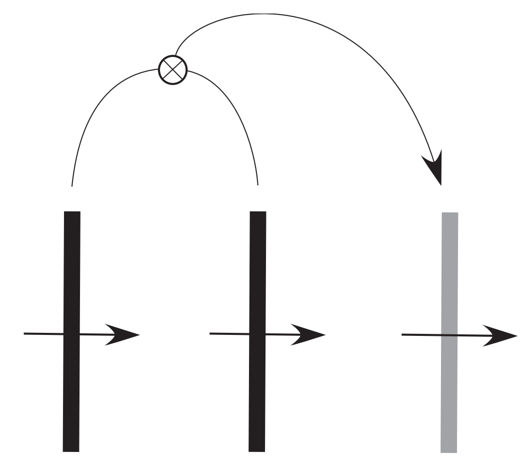
phase alignment



orientation

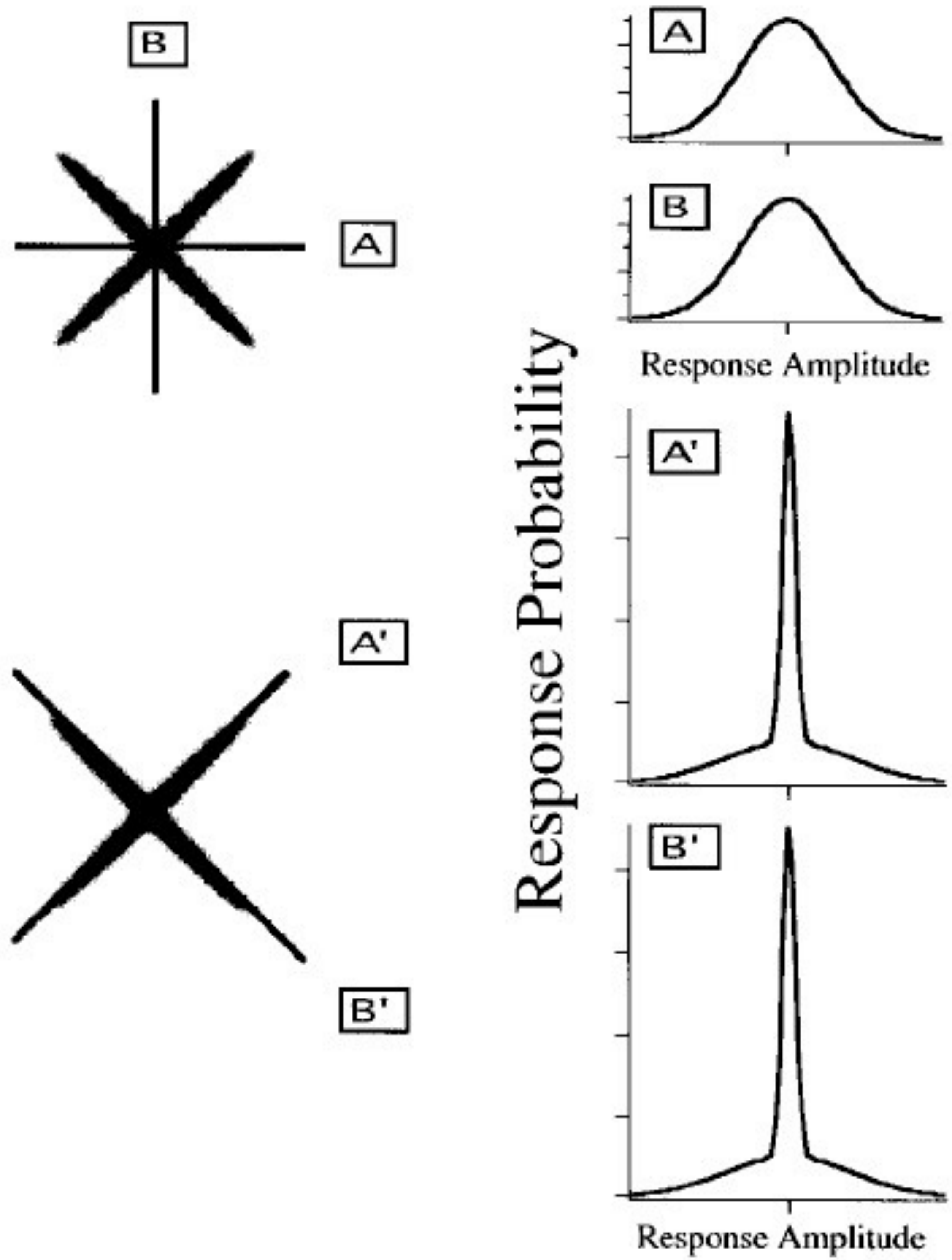


motion

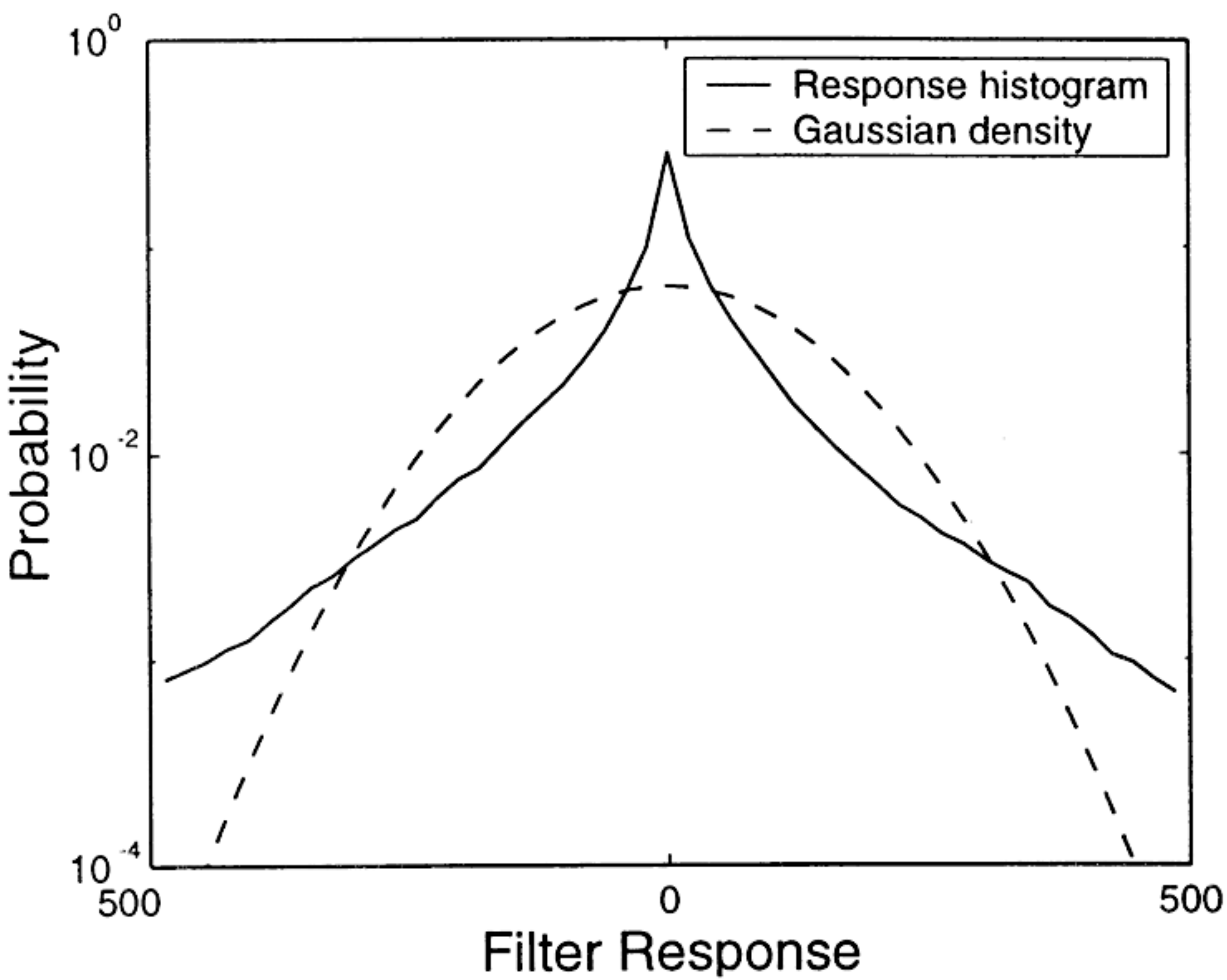


Projection pursuit
(from Field 1994)

Find higher-order structure by maximizing non-Gaussianity of projections

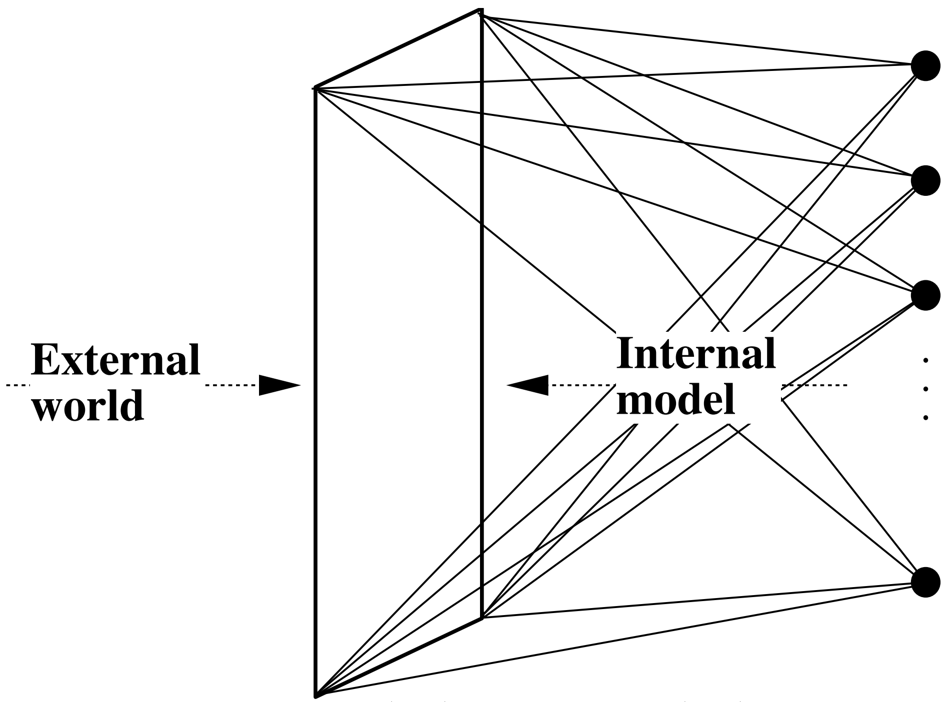


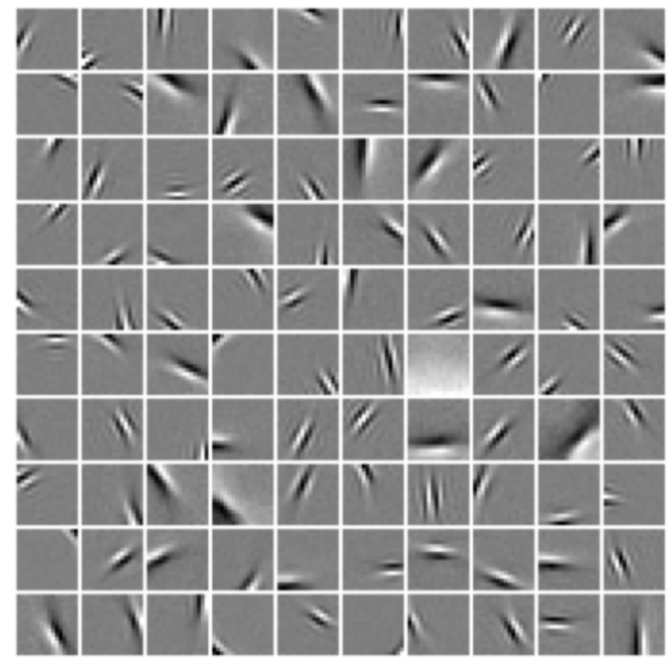
Gabor-filter response histograms are highly non-Gaussian



Sparse coding model of V1

(Olshausen & Field, 1996)





$\phi_i(x,y)$

$$I(x,y) = \sum_i a_i\, \phi_i(x,y) + \epsilon(x,y)$$

Energy function

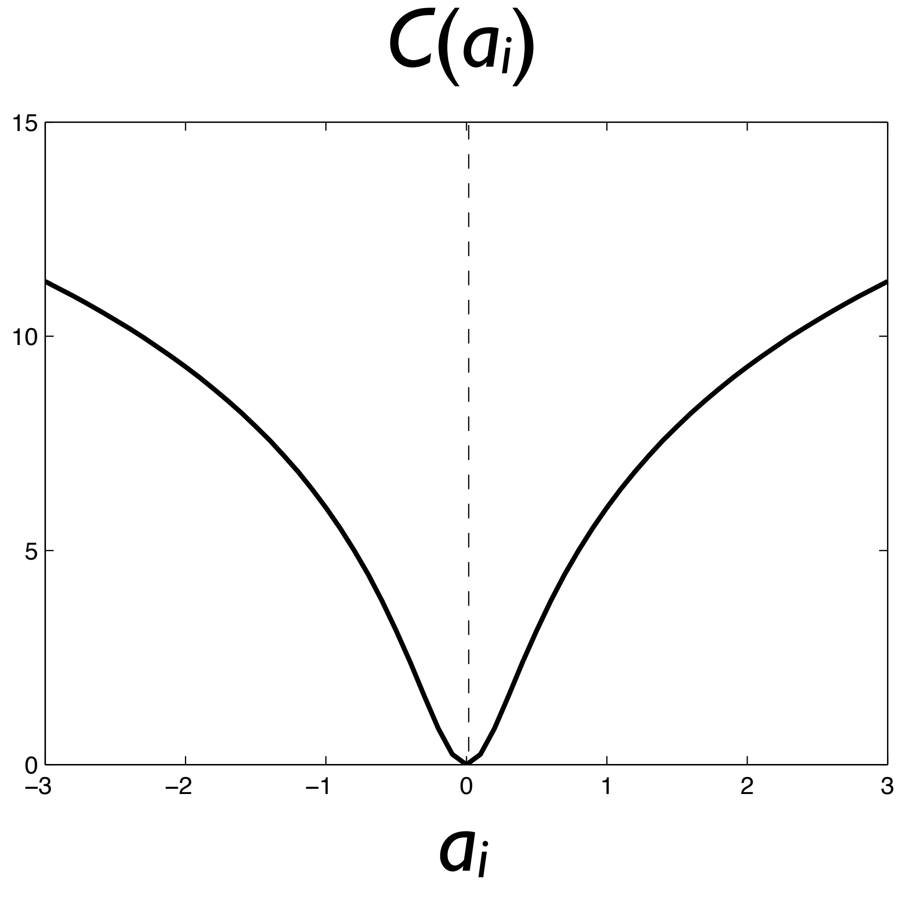
$$E = \frac{1}{2}|\mathbf{I} - \Phi\,\mathbf{a}|^2 + \lambda \sum_i C(a_i)$$

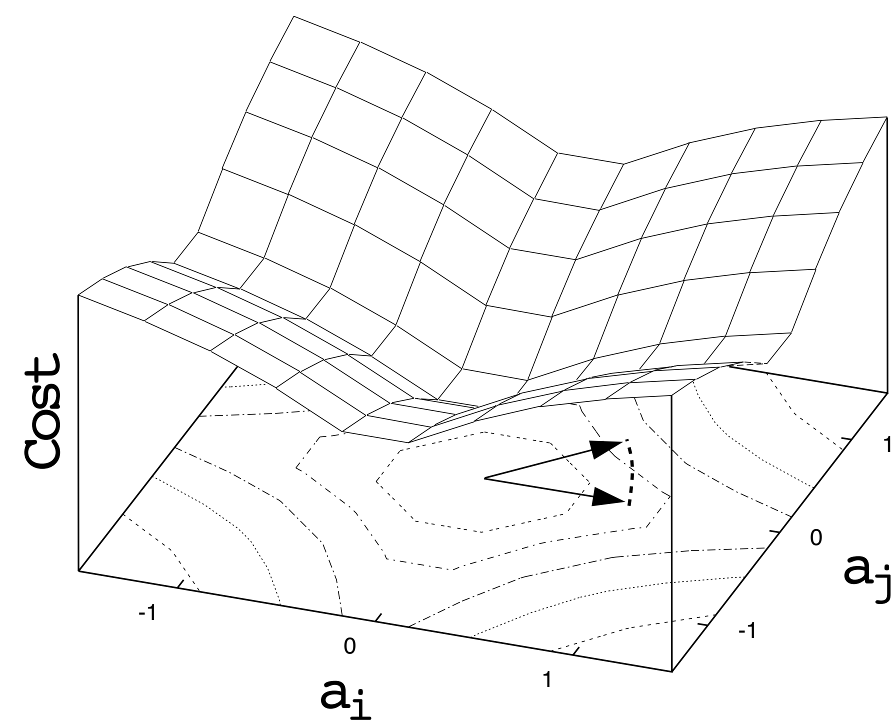
preserve information

be sparse

Cost function

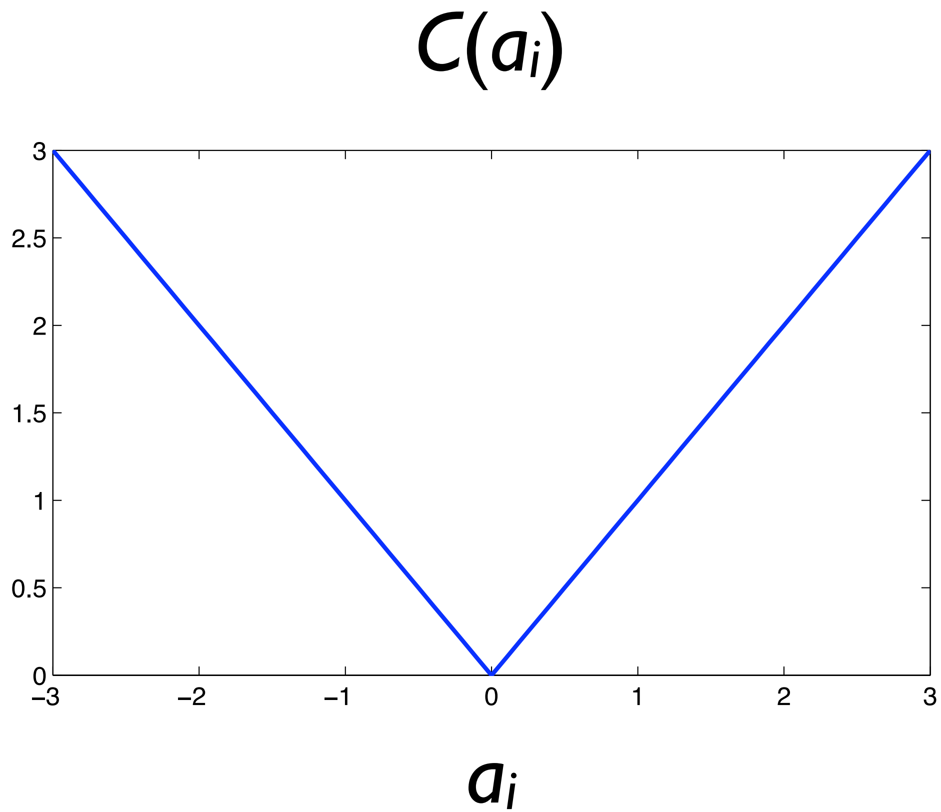
$$C(a_i) = \log(1 + a_i^2)$$

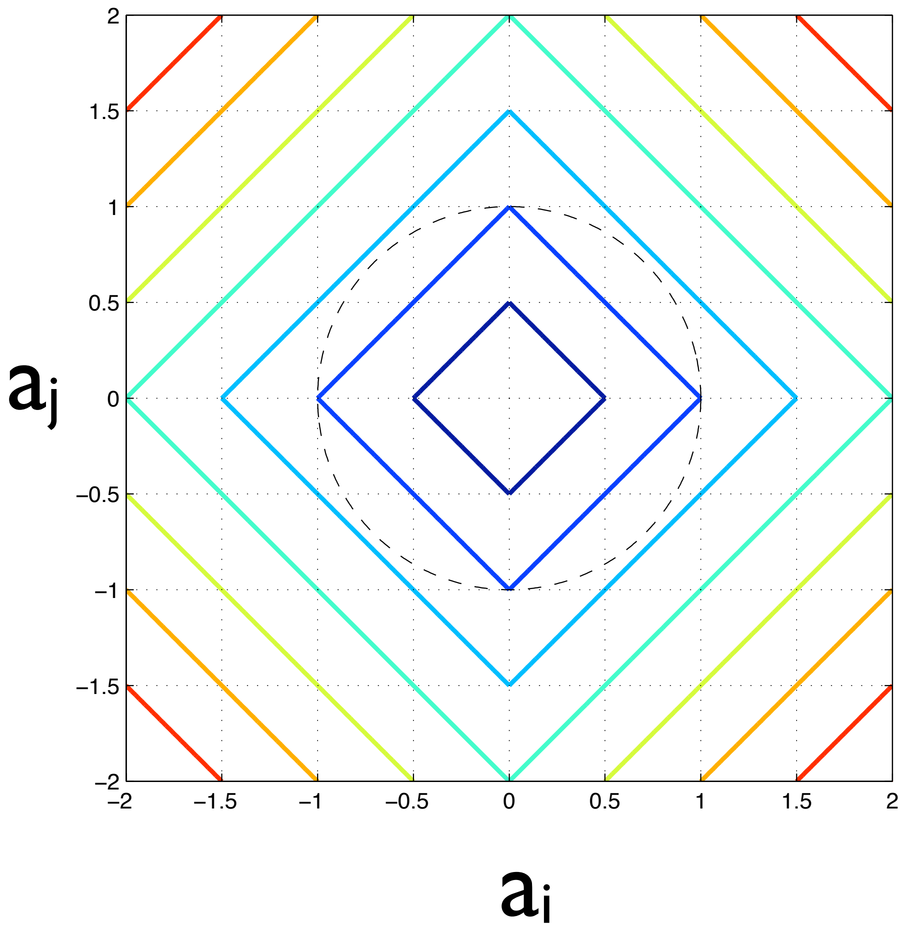




Cost function

$$C(a_i) = |a_i|$$





Inference

$$\hat{\mathbf{a}} = \arg\min_a \left[\frac{1}{2} |\mathbf{I} - \mathbf{\Phi}\,\mathbf{a}|^2 + \lambda \sum_i C(a_i) \right]$$

Compute coefficients via gradient descent

$$\begin{aligned}\tau\,\dot{a}_i &= -\frac{dE}{da_i} \\ &= b_i - \sum_{j\neq i} G_{ij}\,a_j - f_\lambda(a_i)\end{aligned}$$

Where

$$b_i = \sum_{x,y} \phi_i(x,y)\,I(x,y)$$

$$G_{ij} = \sum_{x,y} \phi_i(x,y)\,\phi_j(x,y)$$

$$f_\lambda(a_i) = a_i + \lambda\,C'(a_i)$$

Alternative formulation (the Hopfield trick)

Let

$$u_i = f_\lambda(a_i), \;\; \text{or} \;\; a_i = f_\lambda^{-1}(u_i) \equiv \; g(u_i)$$

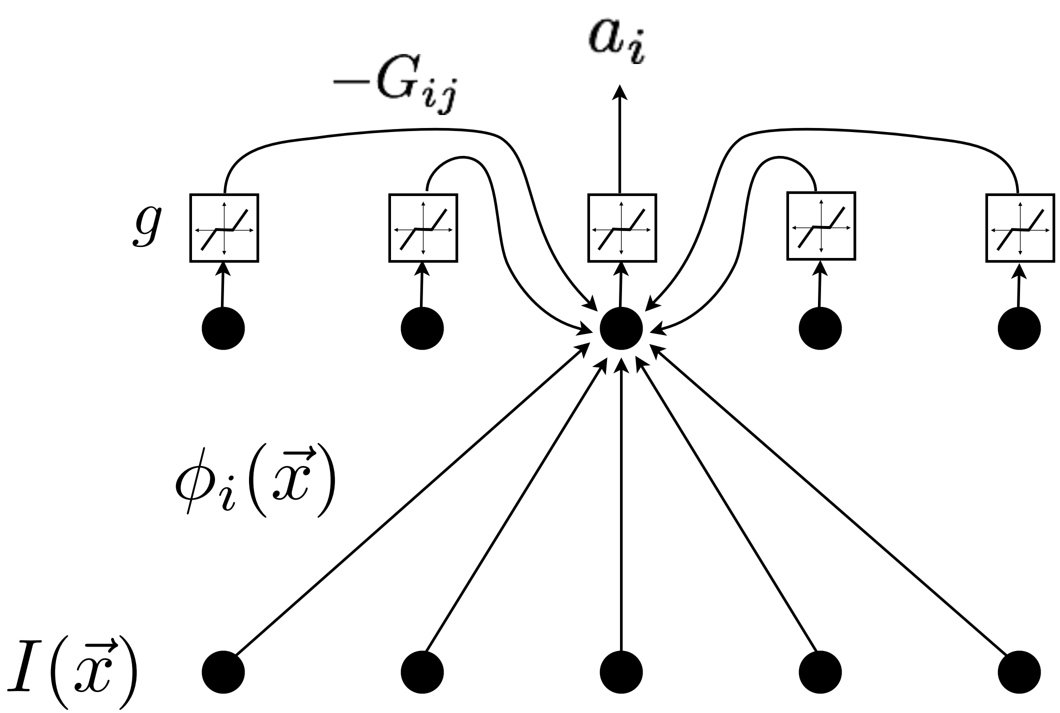
$$\begin{aligned} \tau\, \dot{u}_i &= -\frac{dE}{da_i} \\ &= b_i - \sum_{j \neq i} G_{ij}\, a_j - u_i \end{aligned}$$

Thus

$$\boxed{\begin{aligned} \tau\, \dot{u}_i + u_i &= b_i - \sum_{j \neq i} G_{ij}\, a_j \\ a_i &= g(u_i) \end{aligned}}$$

Neural circuit for computing sparse codes

(Rozell, Johnson, Baraniuk & Olshausen, 2008)



Solves

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} |\mathbf{I} - \Phi\,\mathbf{a}|^2 + \lambda \sum_i C(a_i)$$

$$\begin{aligned} \tau\,\dot{u}_i + u_i &= b_i - \sum_{j \neq i} G_{ij}\, a_j \\ a_i &= g(u_i) \end{aligned}$$

$$b_i = \sum_{\vec{x}} \phi_i(\vec{x})\, I(\vec{x})$$

$$G_{ij} = \sum_{\vec{x}} \phi_i(\vec{x})\, \phi_j(\vec{x})$$

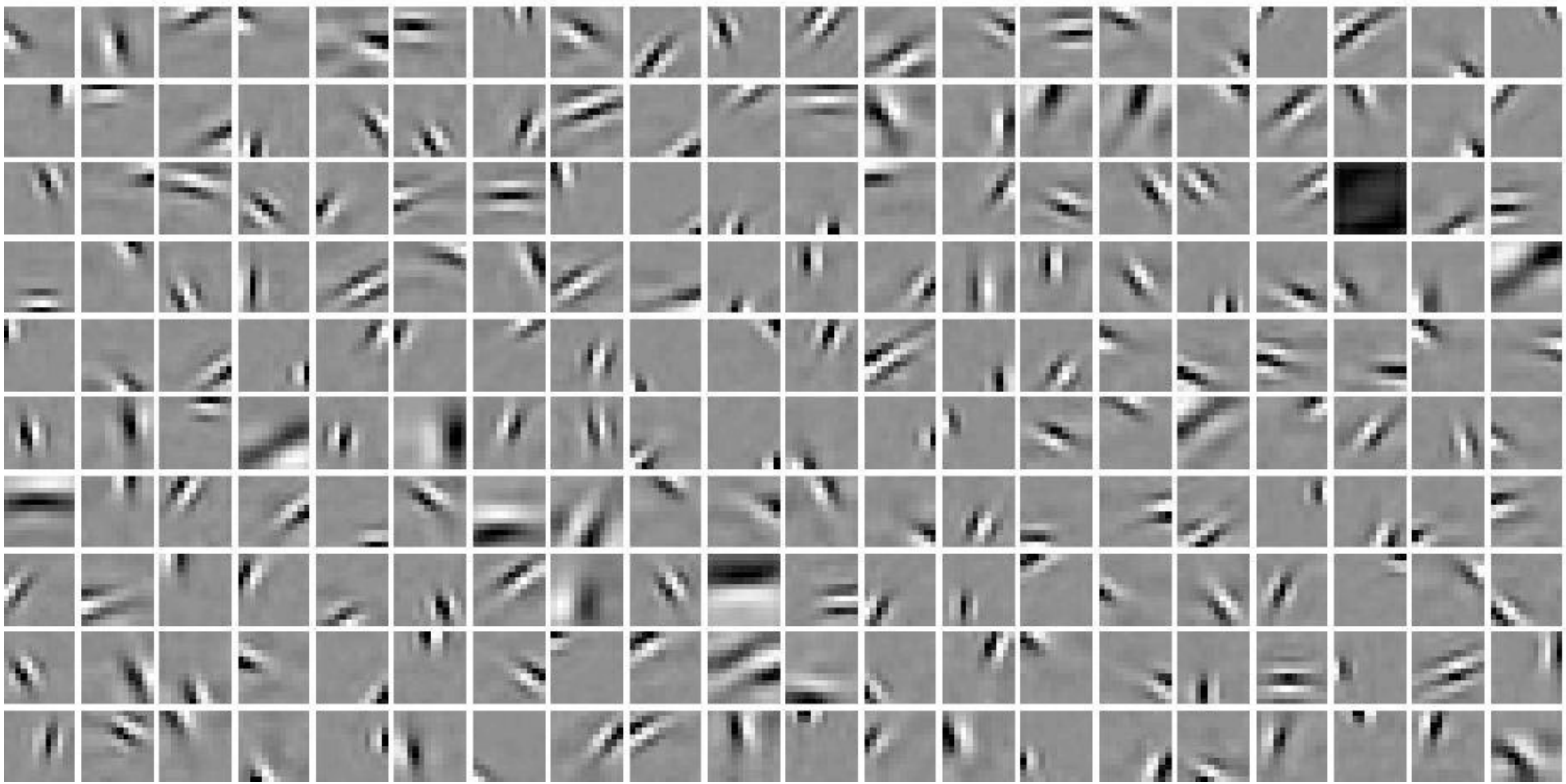
Learning

$$\hat{\mathbf{\Phi}} = \arg\min_{\Phi} \frac{1}{2} |\mathbf{I} - \mathbf{\Phi}\,\hat{\mathbf{a}}|^2$$

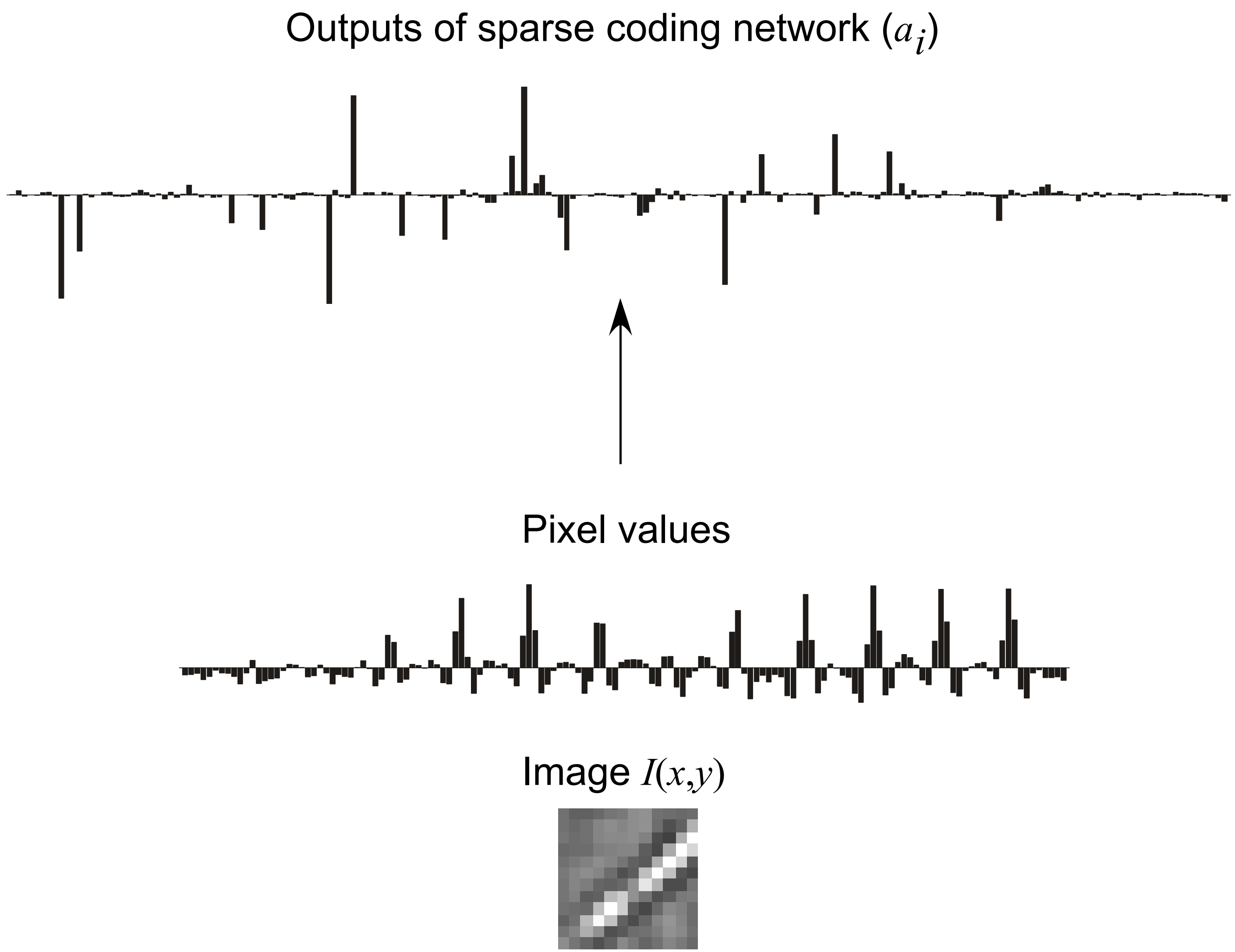
$$\begin{aligned} \Delta\phi_i &= -\eta\,\frac{\partial E}{\partial \phi_i} \\ &= [\mathbf{I} - \Phi\,\hat{\mathbf{a}}]\,\hat{a}_i \end{aligned}$$

learning rule

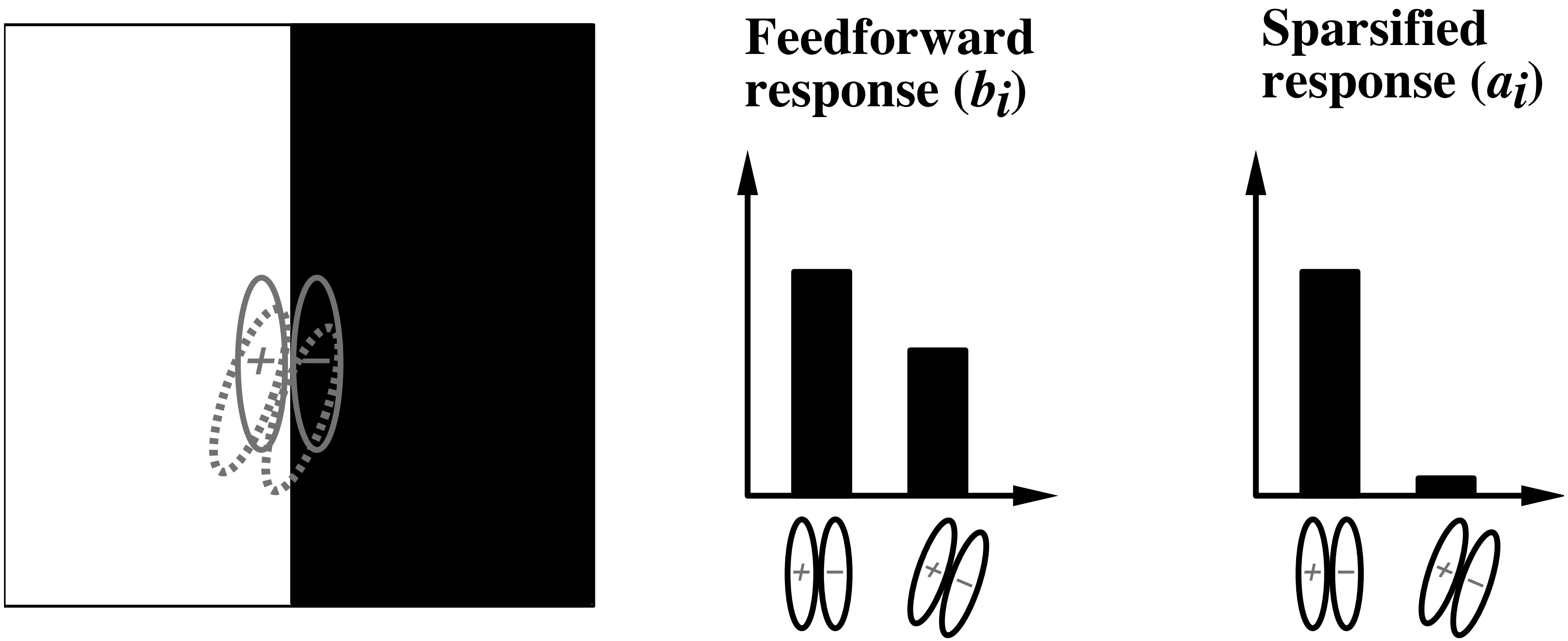
Features learned from natural images
(200, 12x12 pixels)



Sparsification

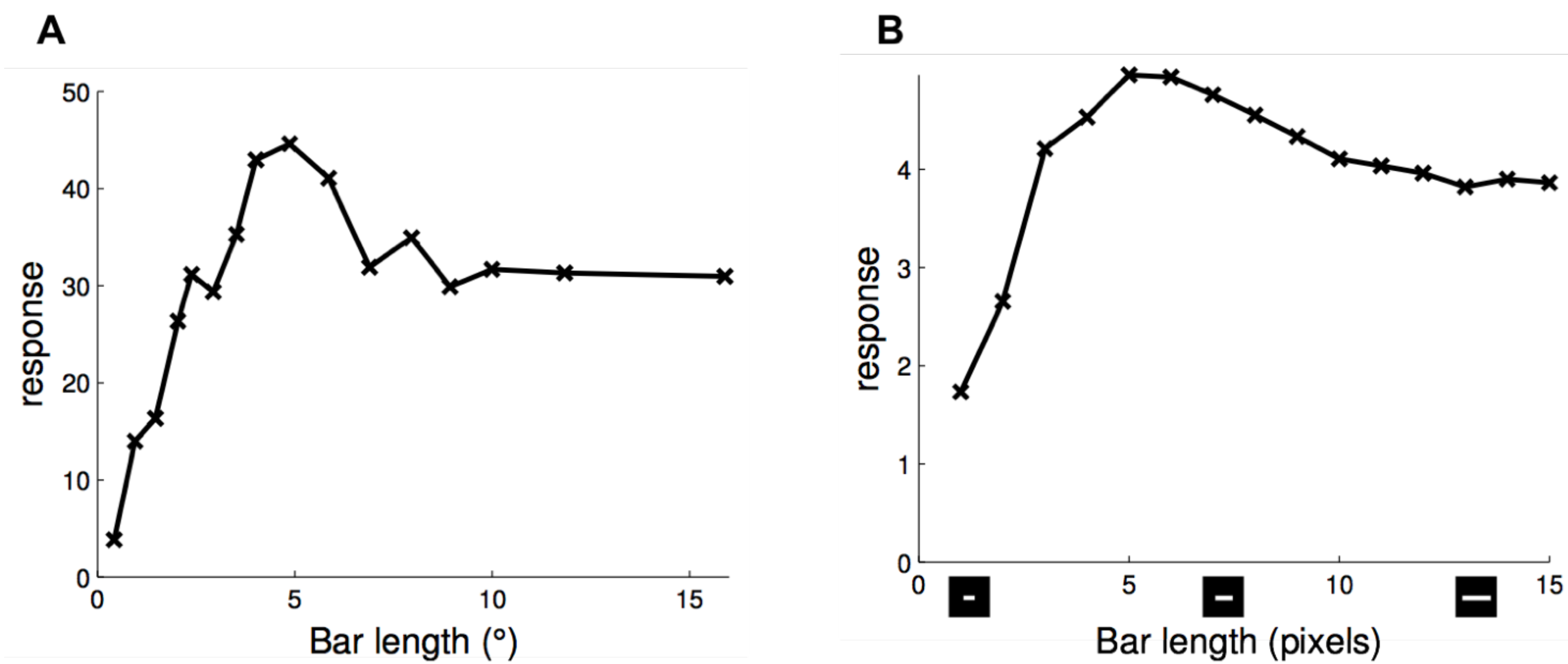


‘Explaining away’



Explaining away can account for non-classical surround effects such as end-stopping

(Lee et al., 2006; Zhu & Rozell, 2013)



Evidence for sparse coding

Mushroom body, locust (Laurent)

HVC, zebra finch (Fee)

Auditory cortex, mouse (DeWeese & Zador)

Hippocampus, rat/primate (Thompson & Best; Skaggs)

Motor cortex, rabbit (Swadlow)

Barrel cortex, rat (Brecht)

Visual cortex, monkey/cat (Vinje & Gallant)

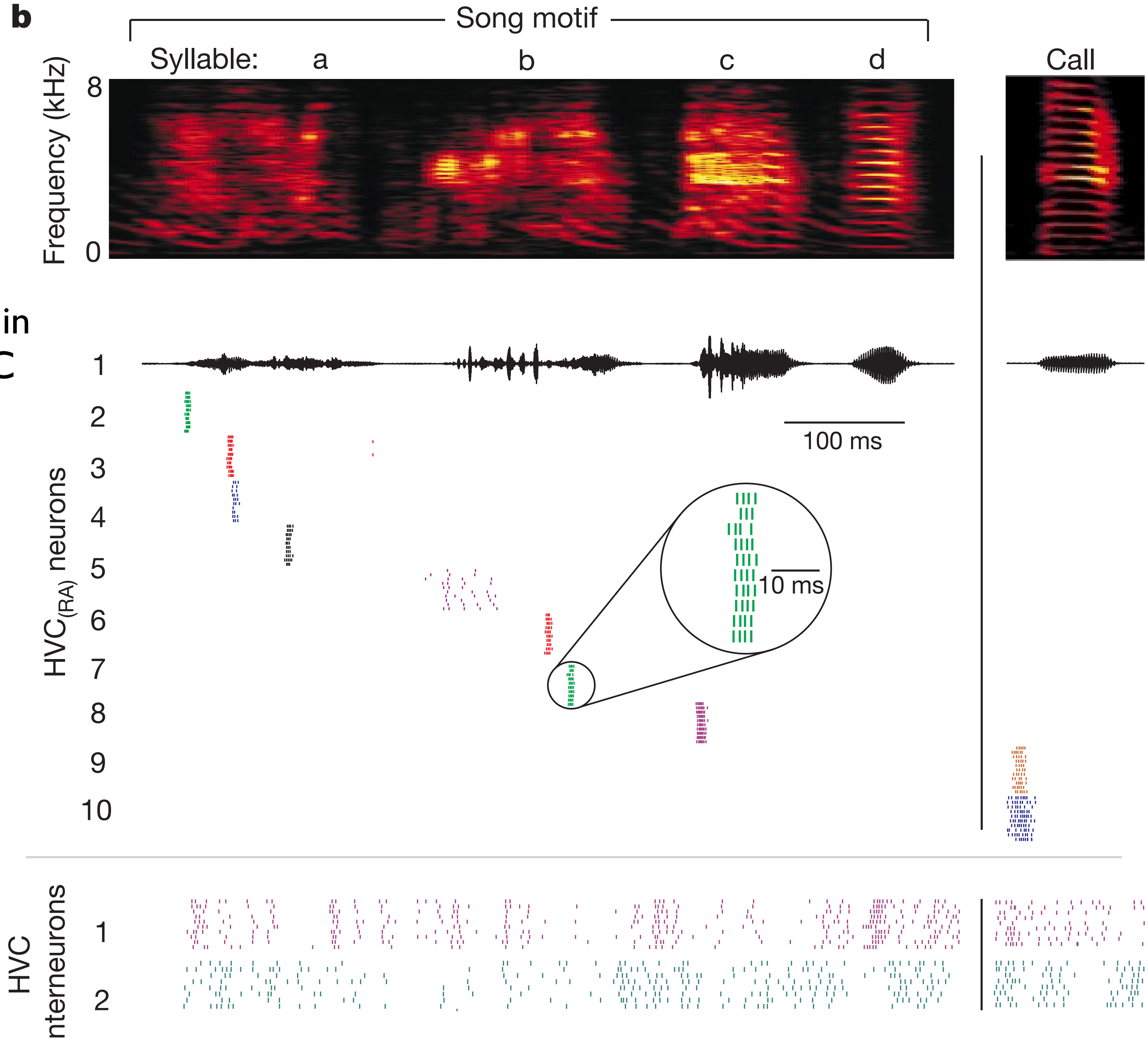
Visual cortex, cat (Gray; McCormick)

Inferotemporal cortex, human (Fried & Koch)

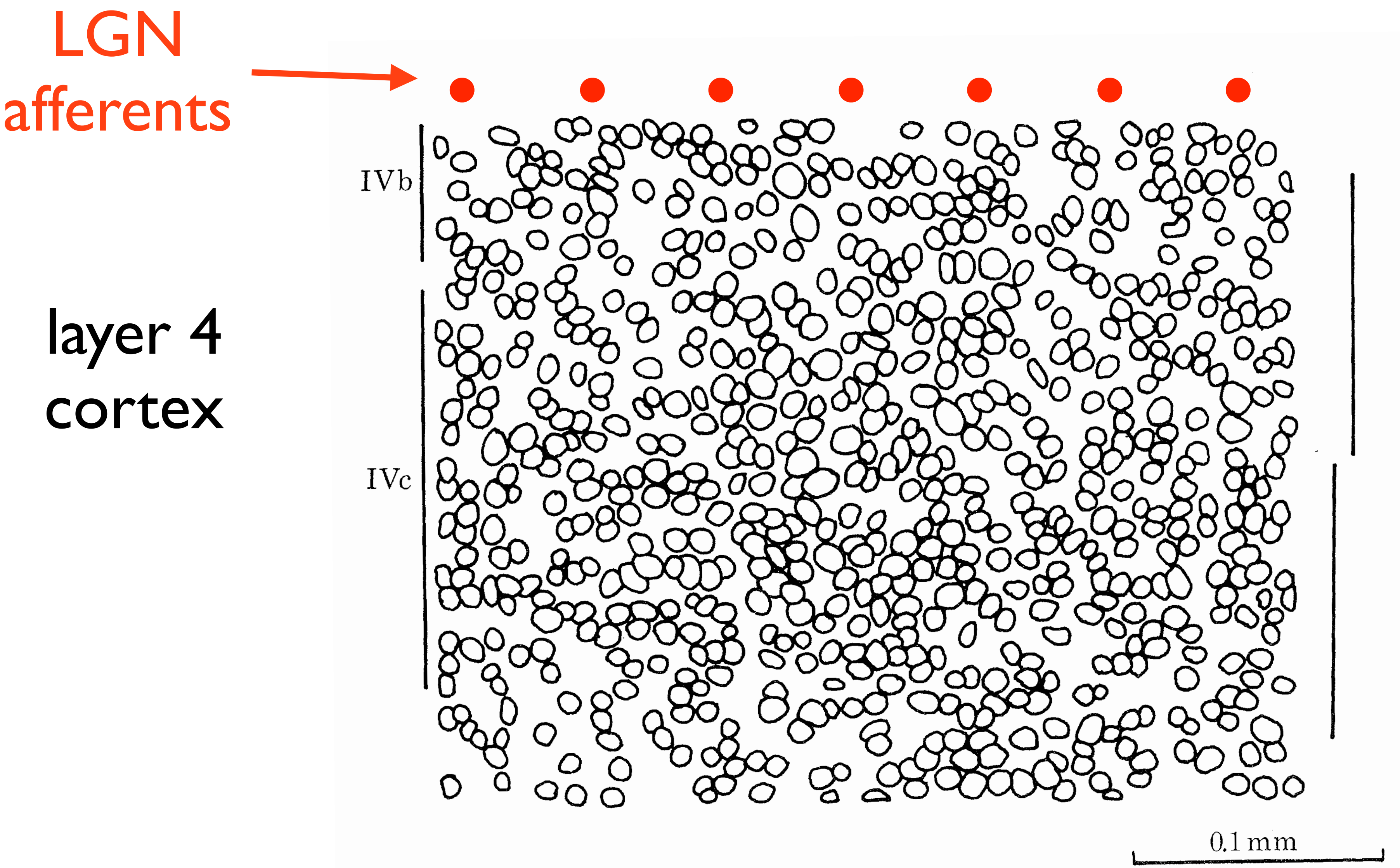
Olshausen BA, Field DJ (2004) Sparse coding of sensory inputs. *Current Opinion in Neurobiology, 14*, 481-487.

Sparse coding in songbird HVC

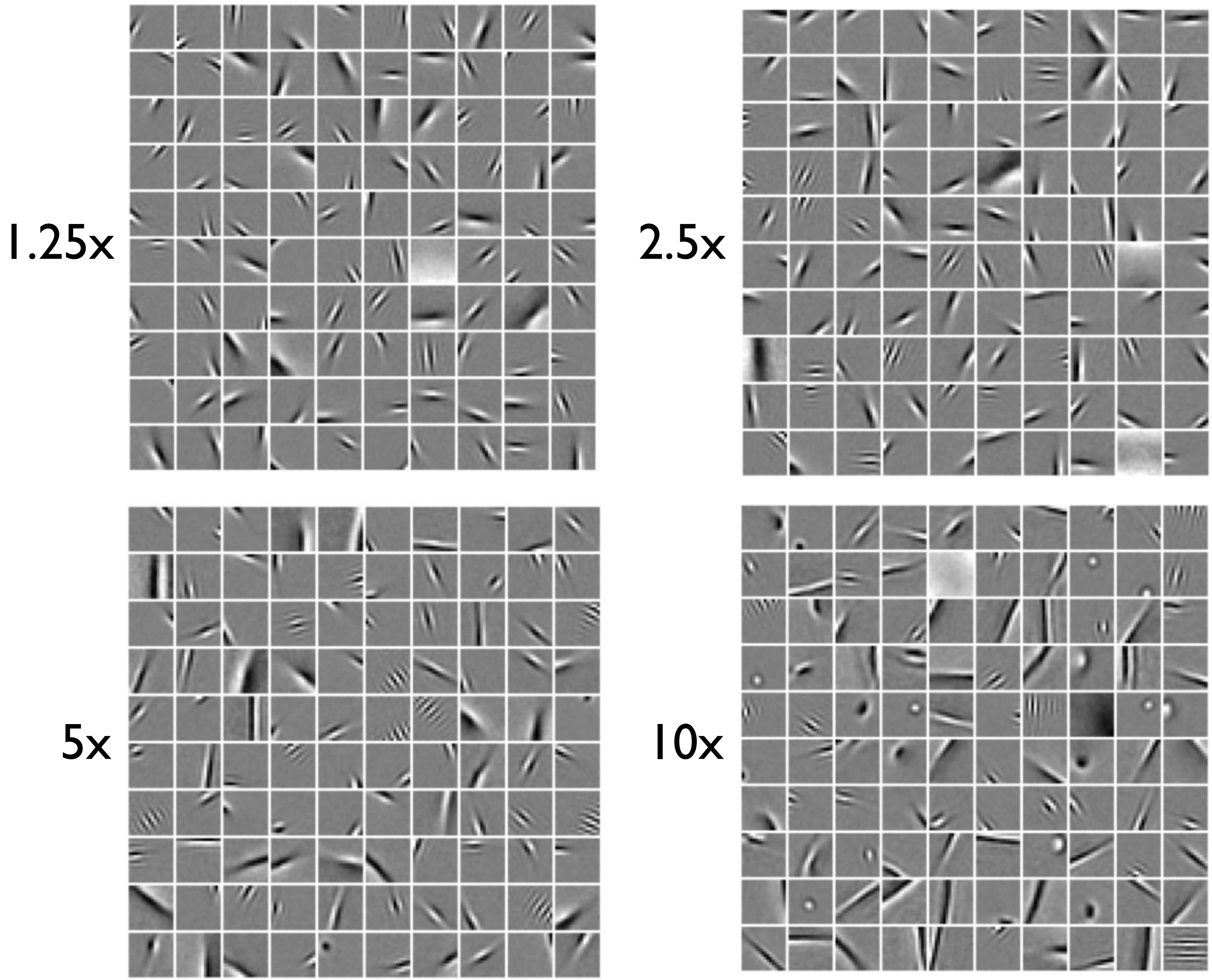
Hahnloser, Kozhevnikov & Fee (2002)



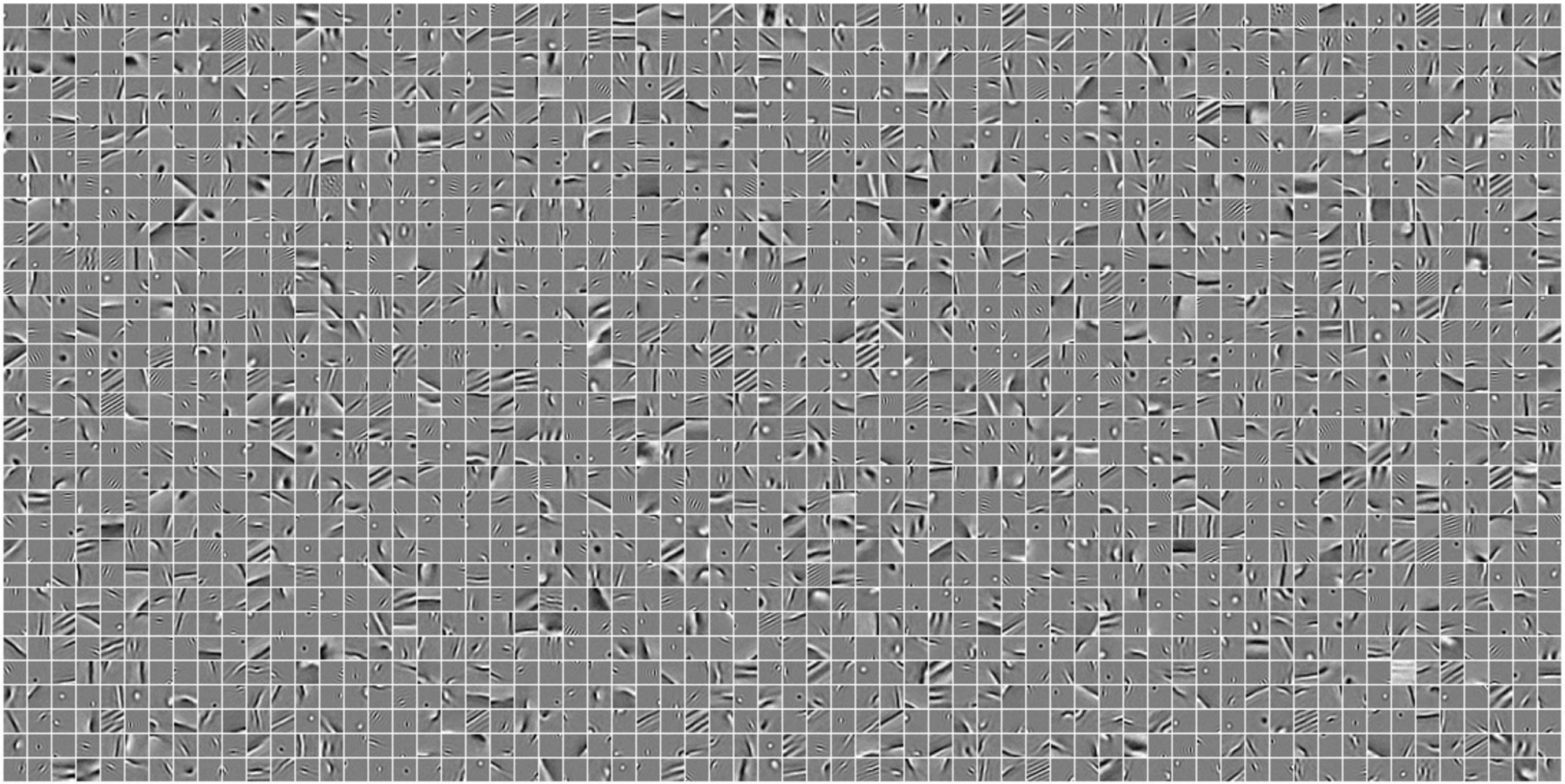
V1 is highly overcomplete

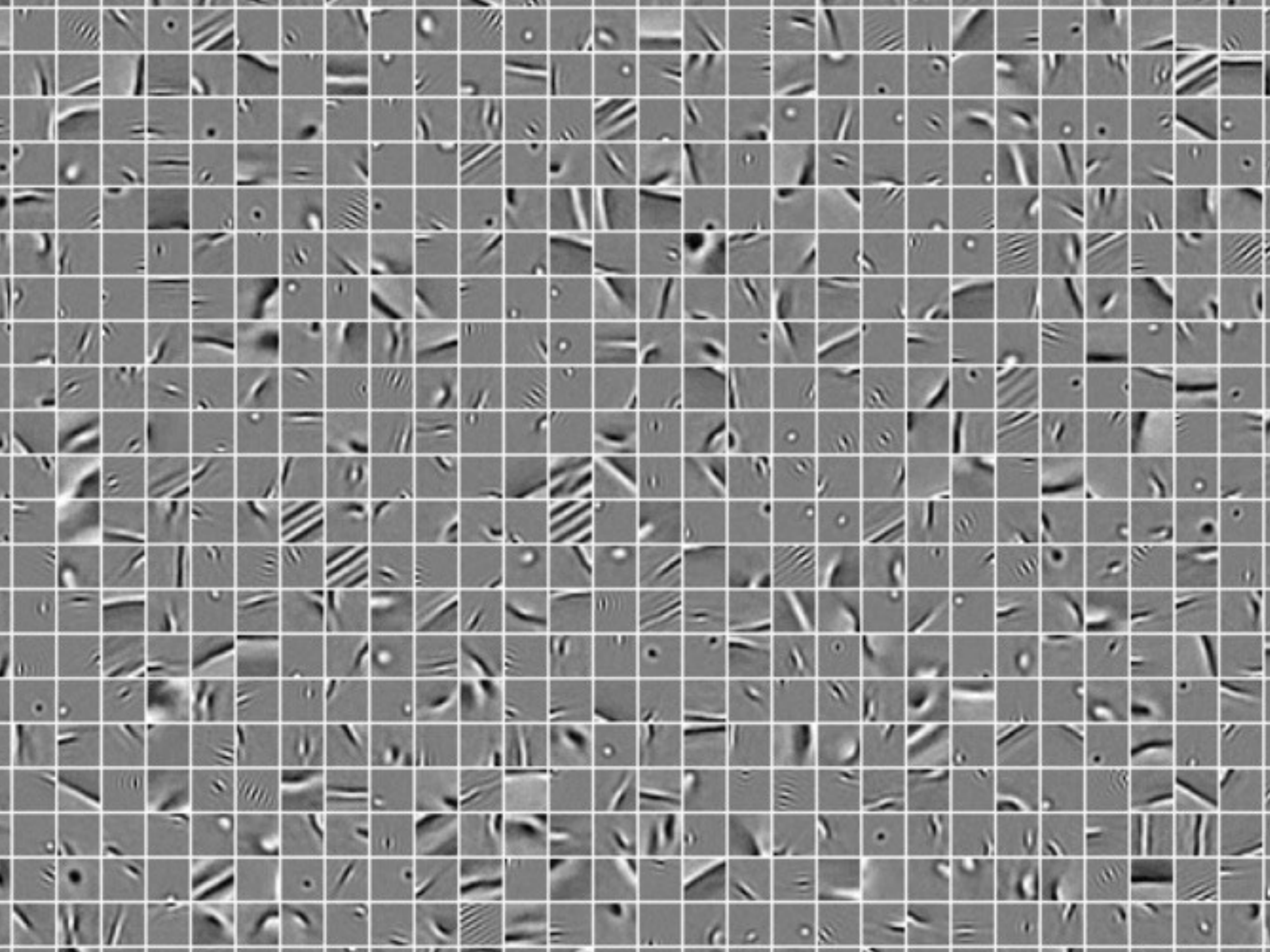


Barlow (1981)


1.25x
2.5x
5x
10x

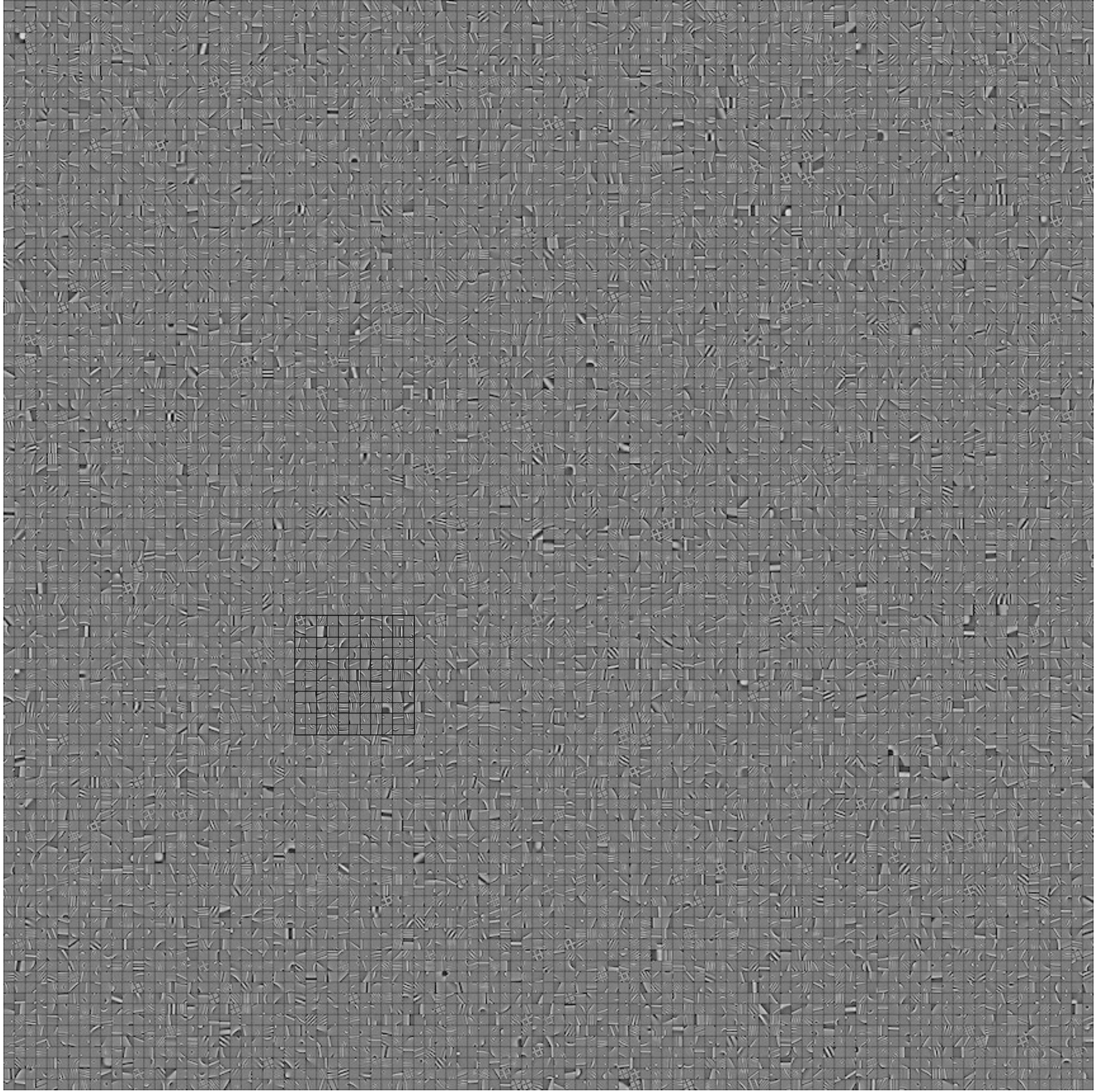
Full 10x dictionary

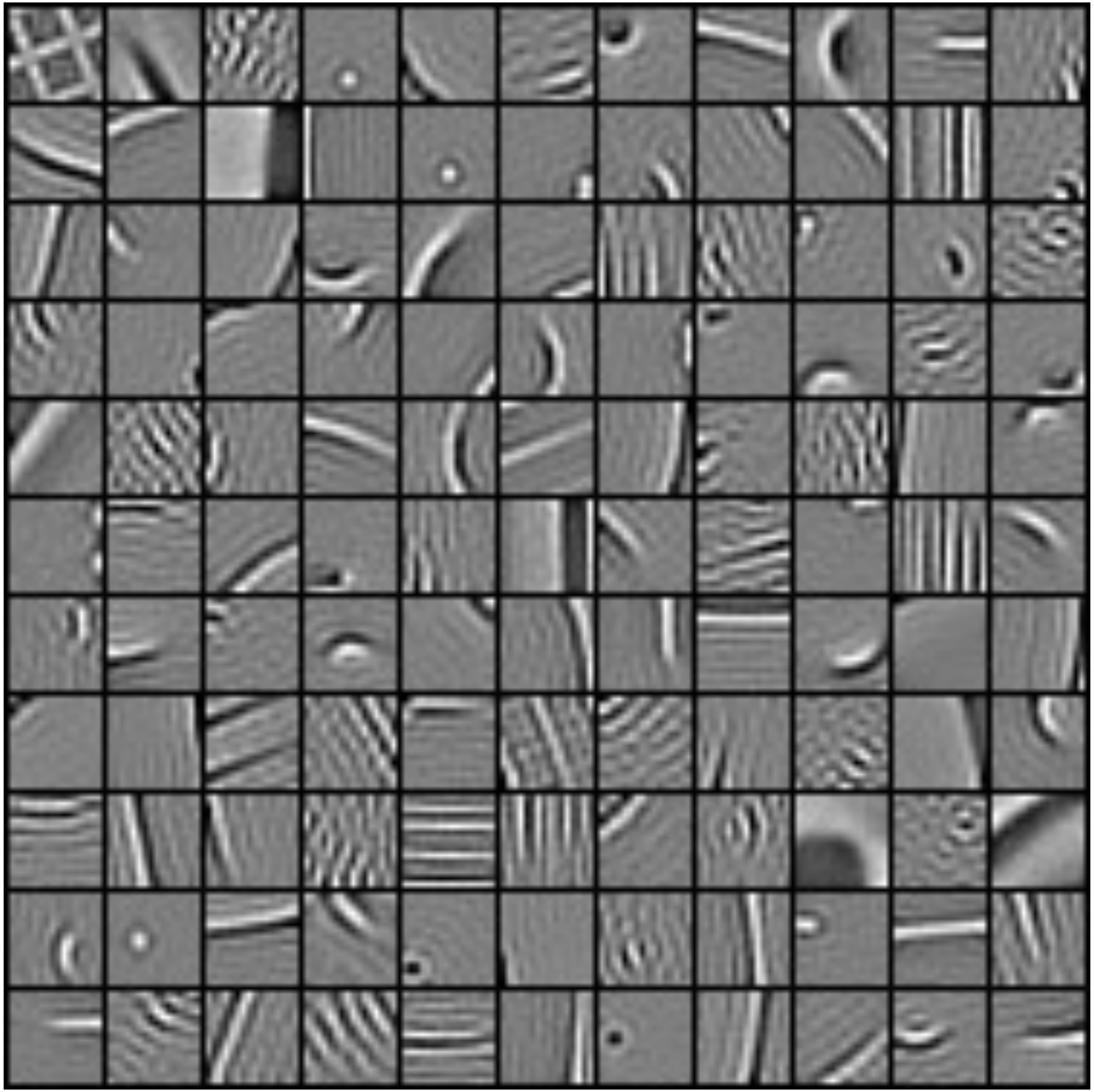




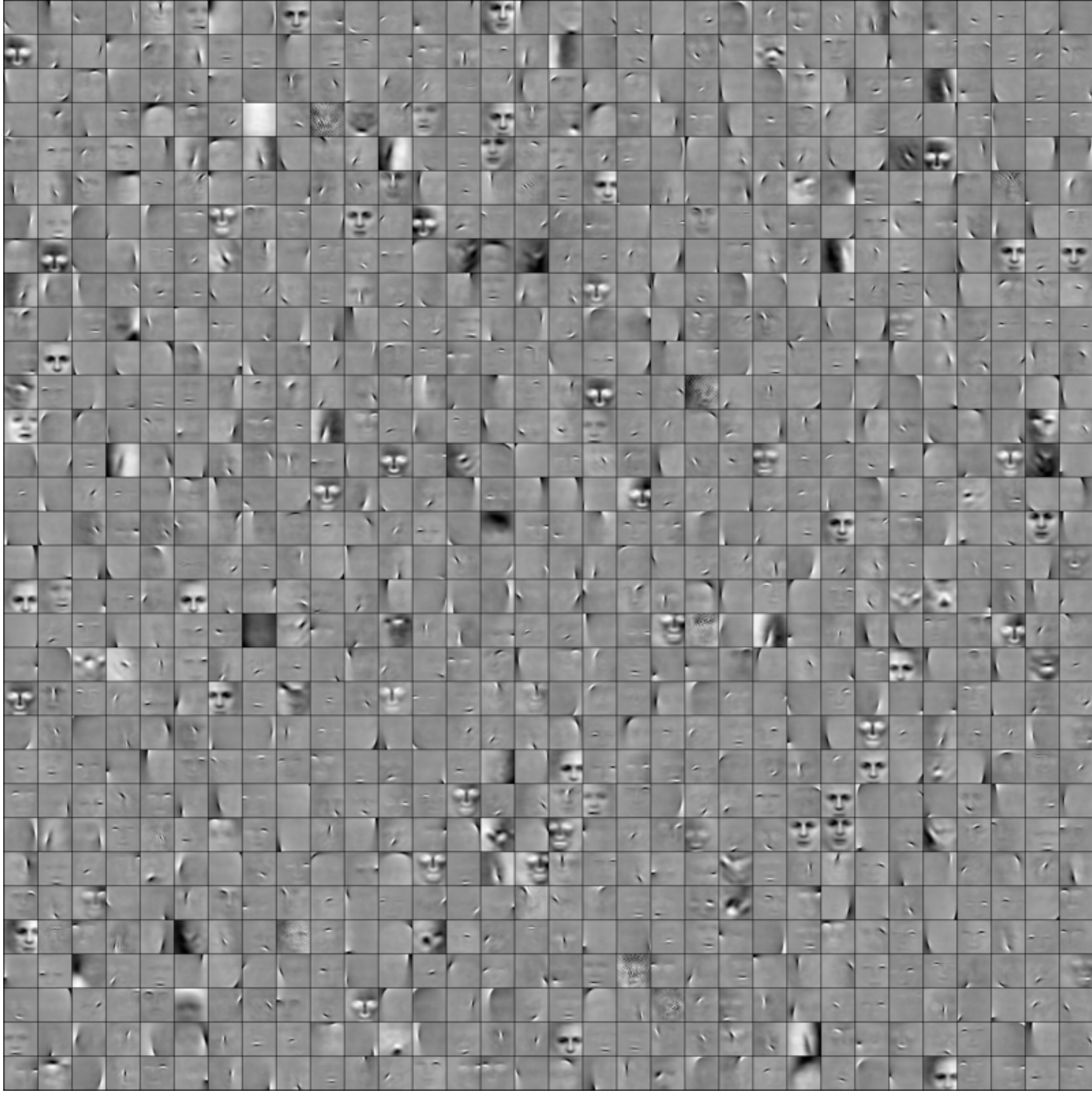
100x overcomplete learned dictionary

(obtained by Charles Cadieu after running for 8 hours on16 GPU's)



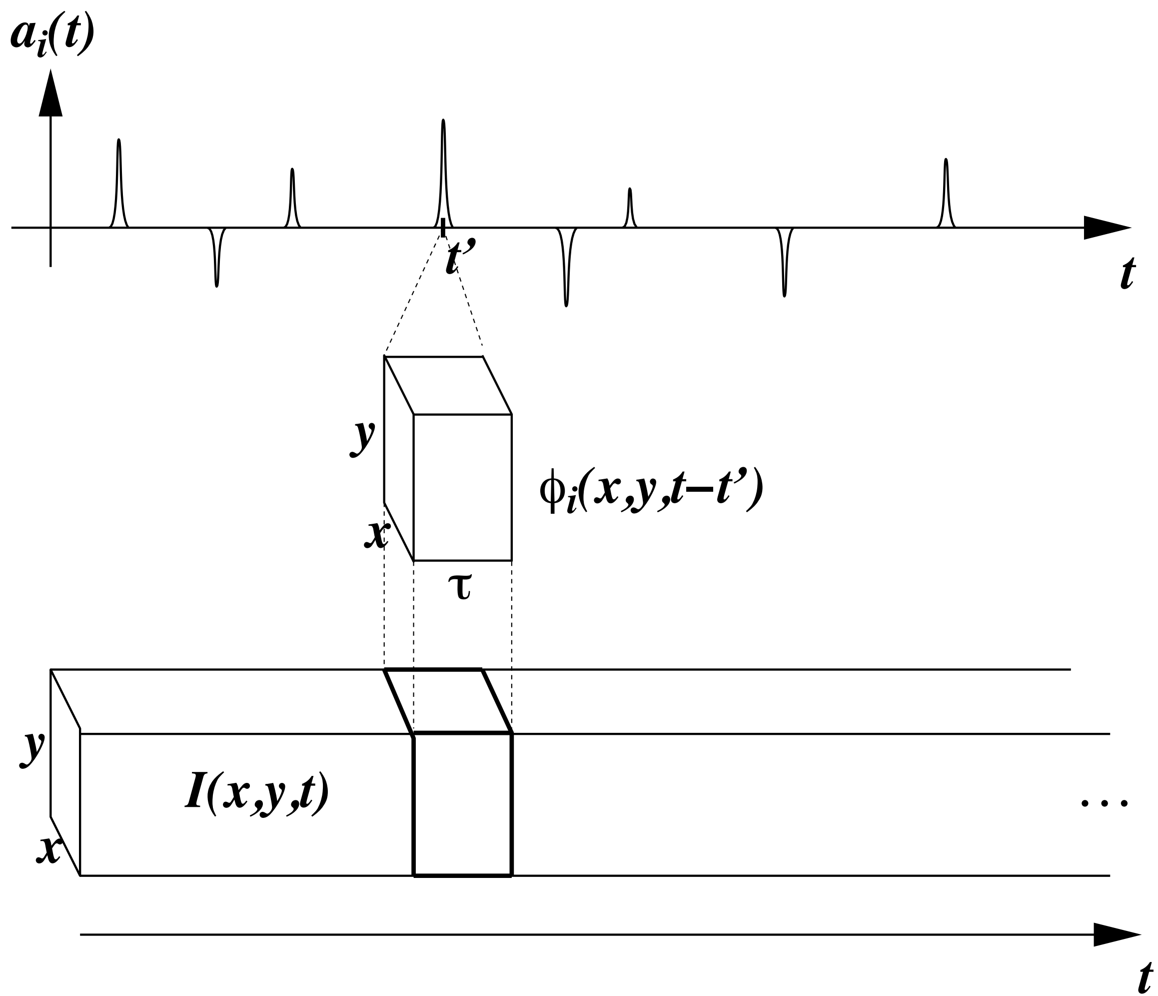


Faces (charles cadieu)



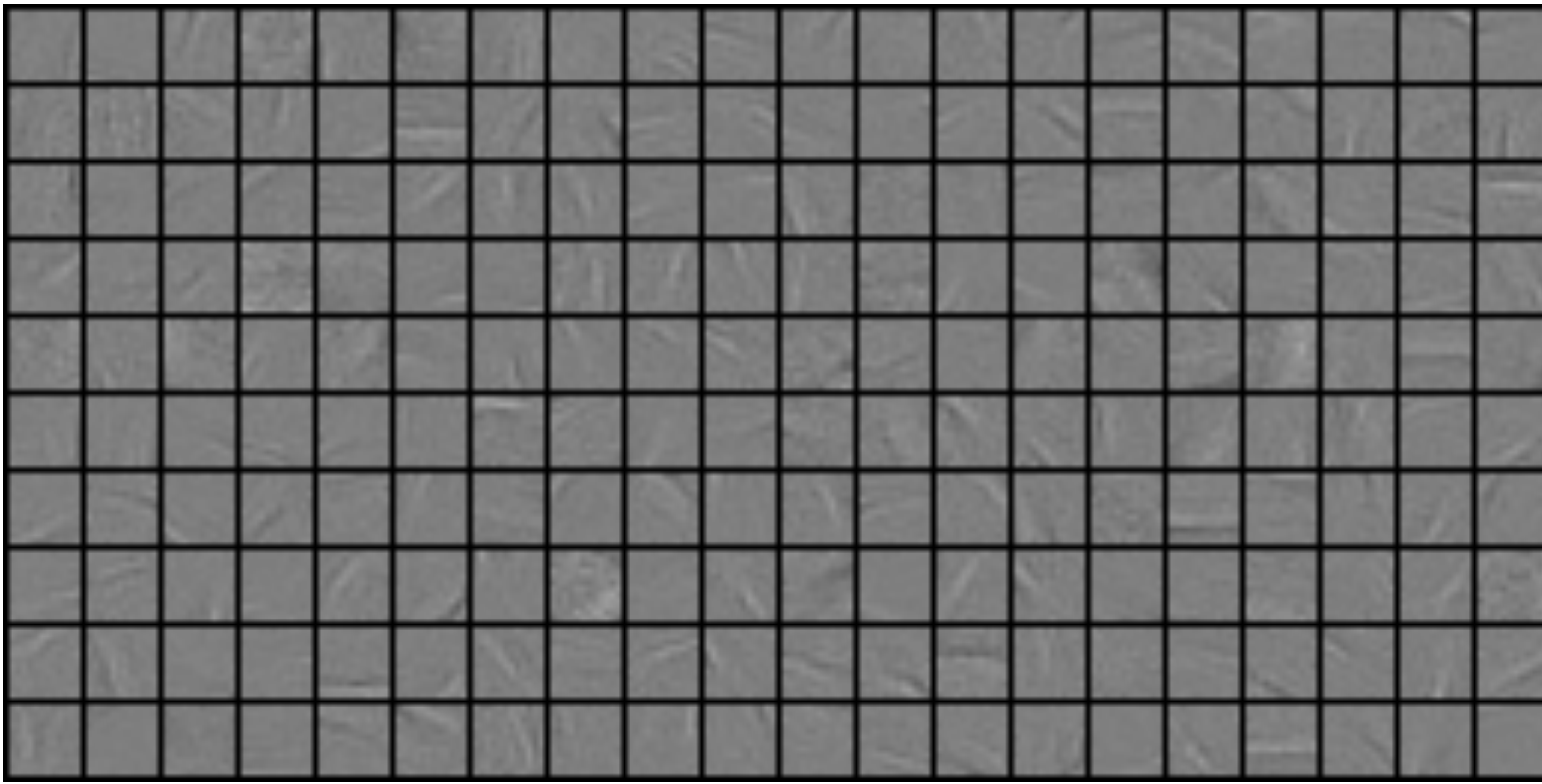
Sparse coding of time-varying images

$$I(x,y,t) = \sum_i a_i(t) * \phi_i(x,y,t) + \nu(x,y,t)$$

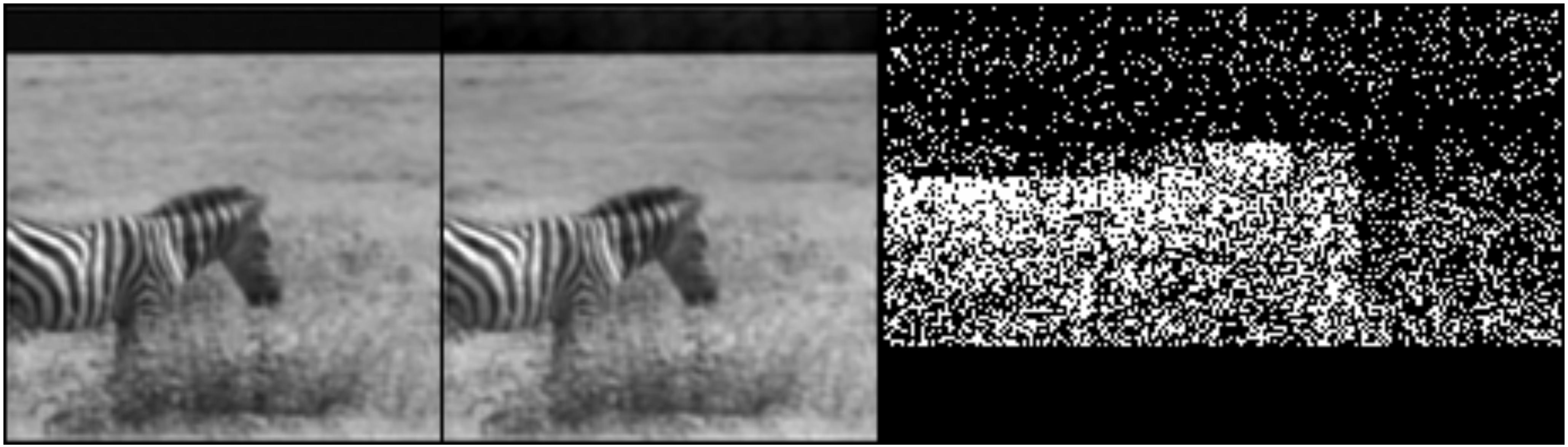


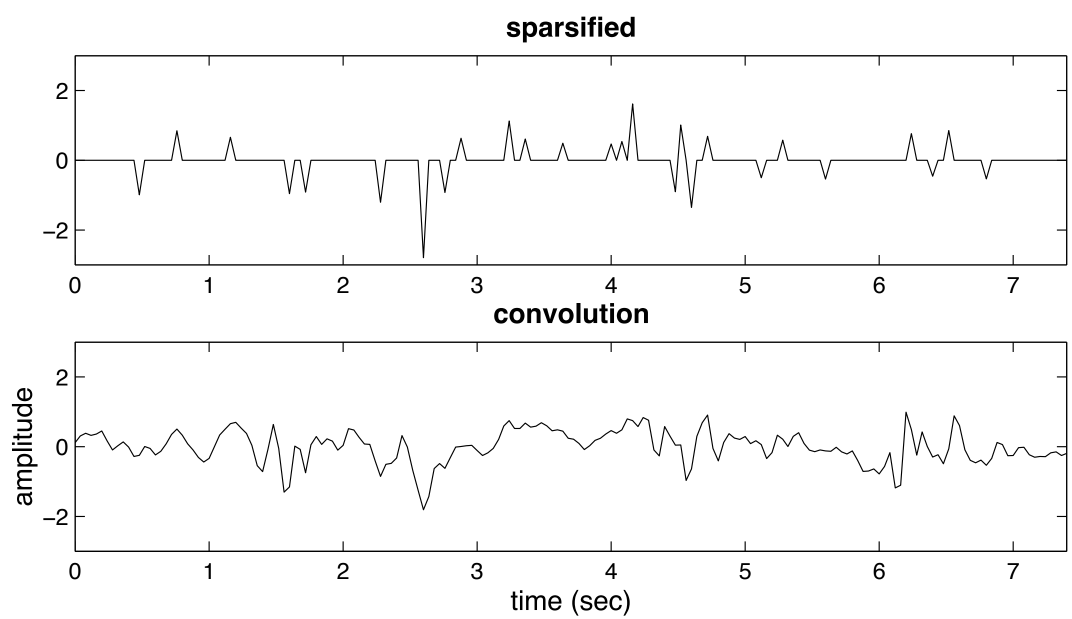
Learned basis space-time basis functions

(200 bfs, 12 x12 x 7)



Sparse coding and reconstruction

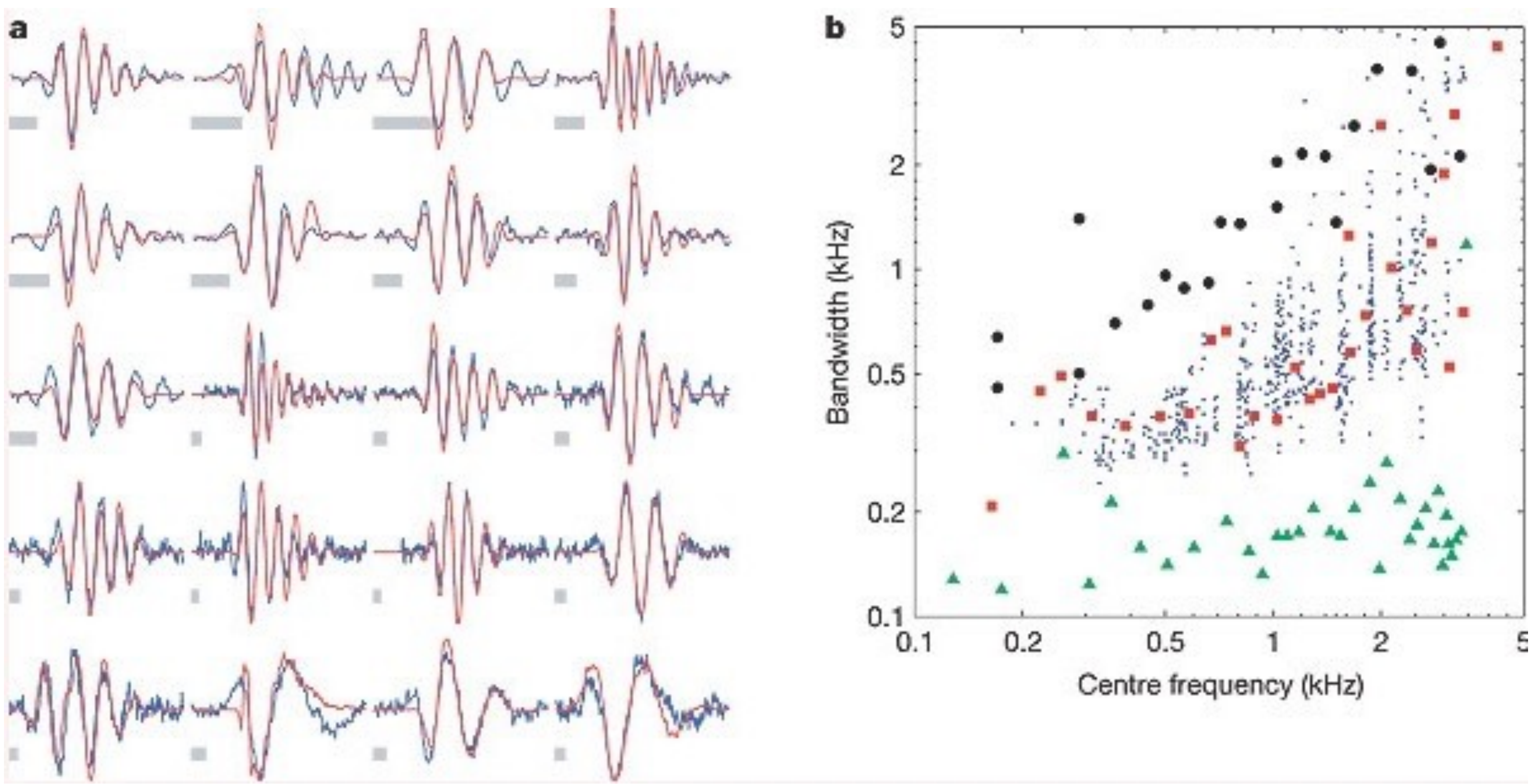




Sparse coding of natural sounds
(Smith & Lewicki 2006)

$$s(t) = \sum_i a_i(t) \; * \; \phi_i(t) + \nu(t)$$

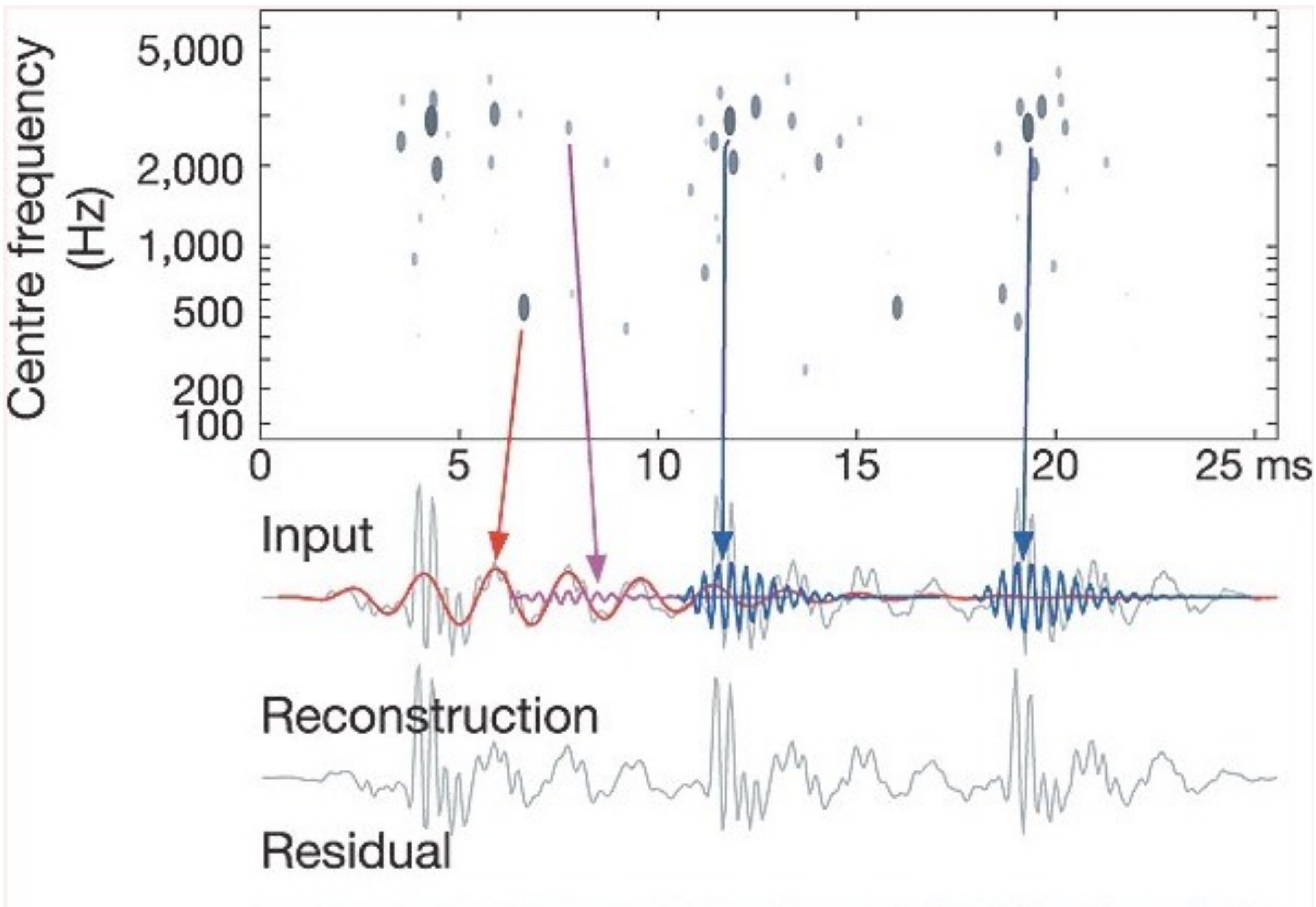
$\phi_i(t)$



Sparse coding of natural sounds
(Smith & Lewicki 2006)

$a_i(t)$

$s(t)$



Sparse coding of neural recording data

(Phil Sallee, Ph.D. thesis)

$$s_i(t) = \sum_j a_j(t) \, * \, \phi_{ij}(t) + \nu_i(t)$$

recorded voltage at electrode *i*

causes

noise at electrode *i*

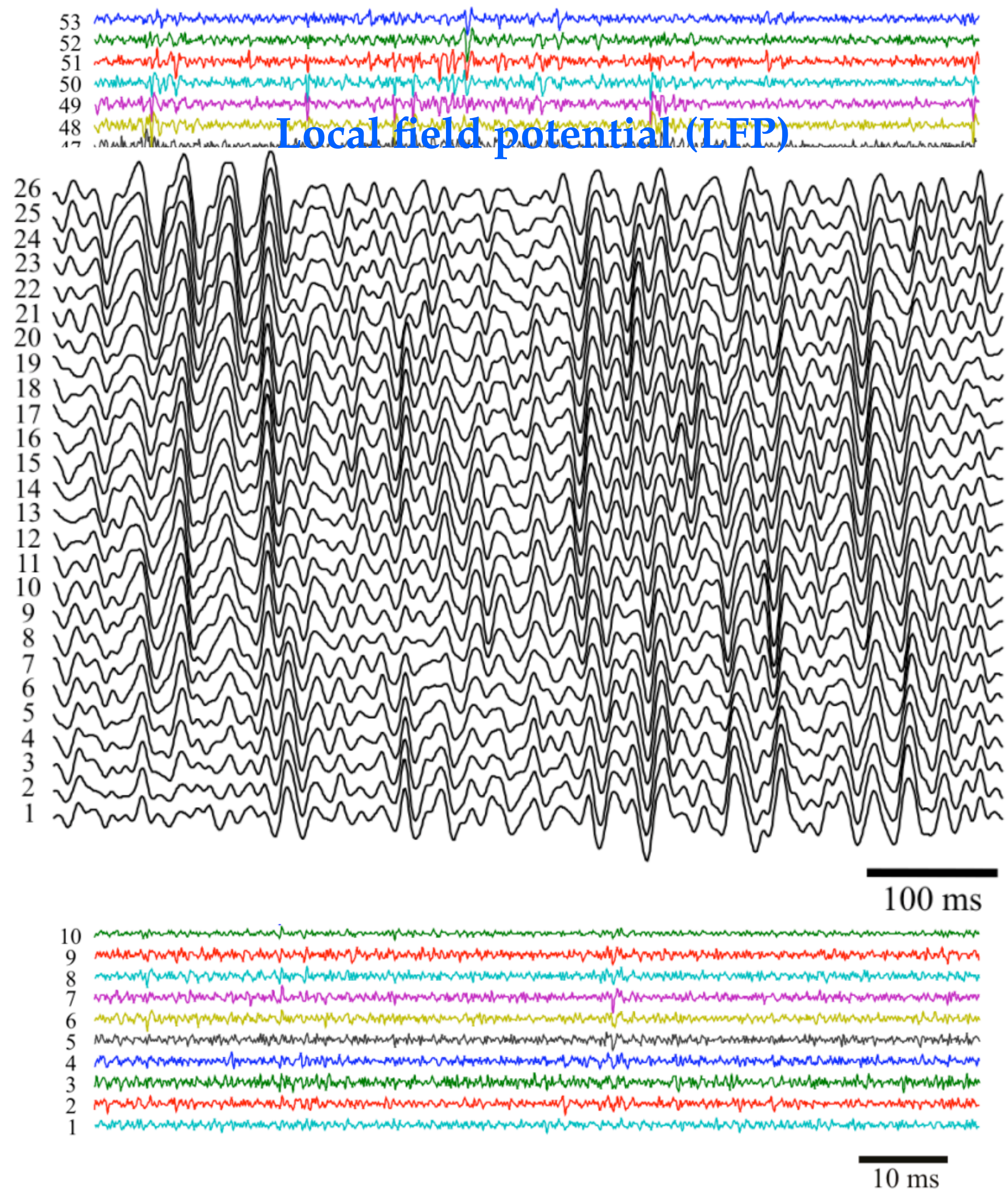
Polytrode recordings

Silicon polytrodes

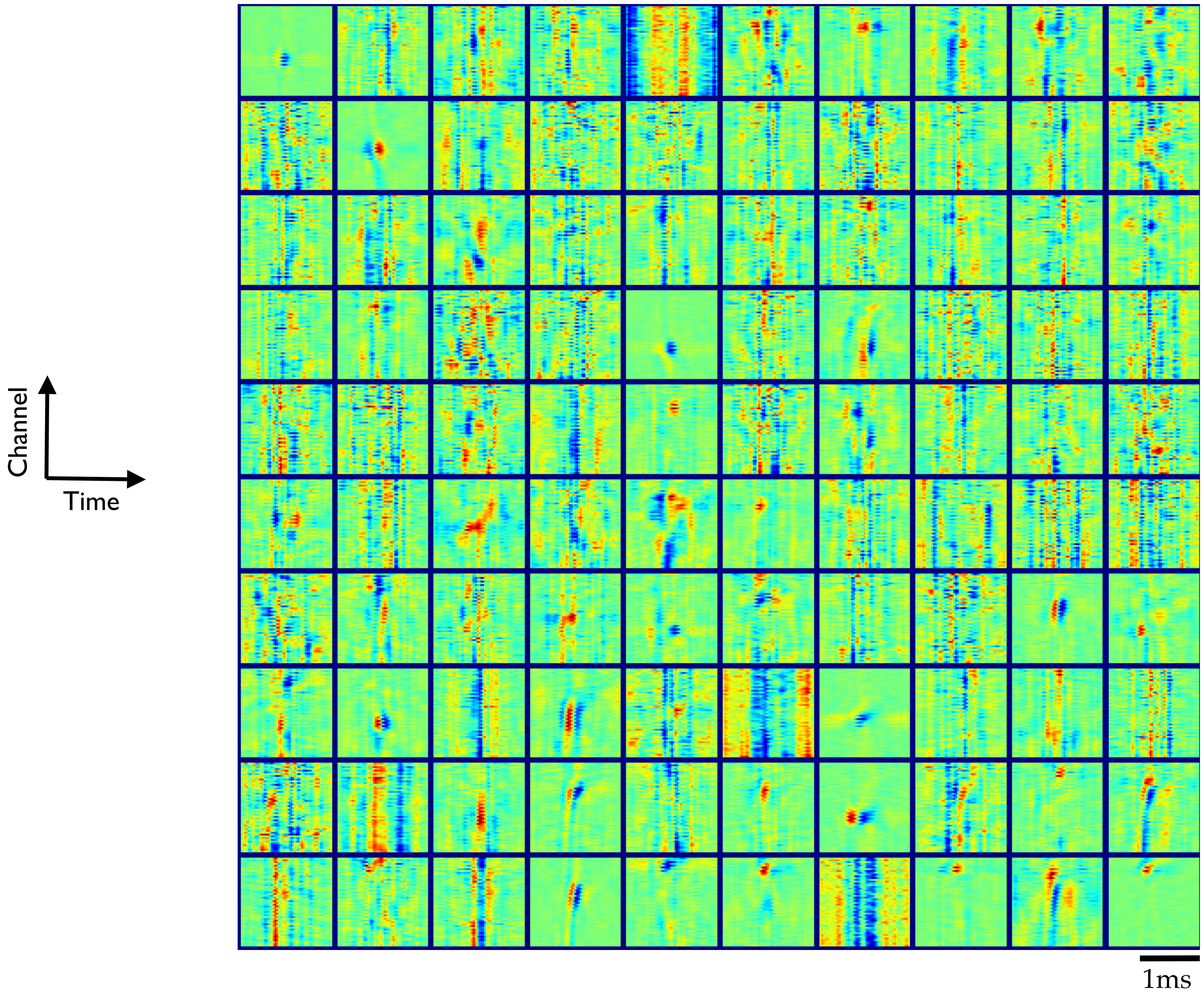
Blanche et al. (2005)

Spiking activity

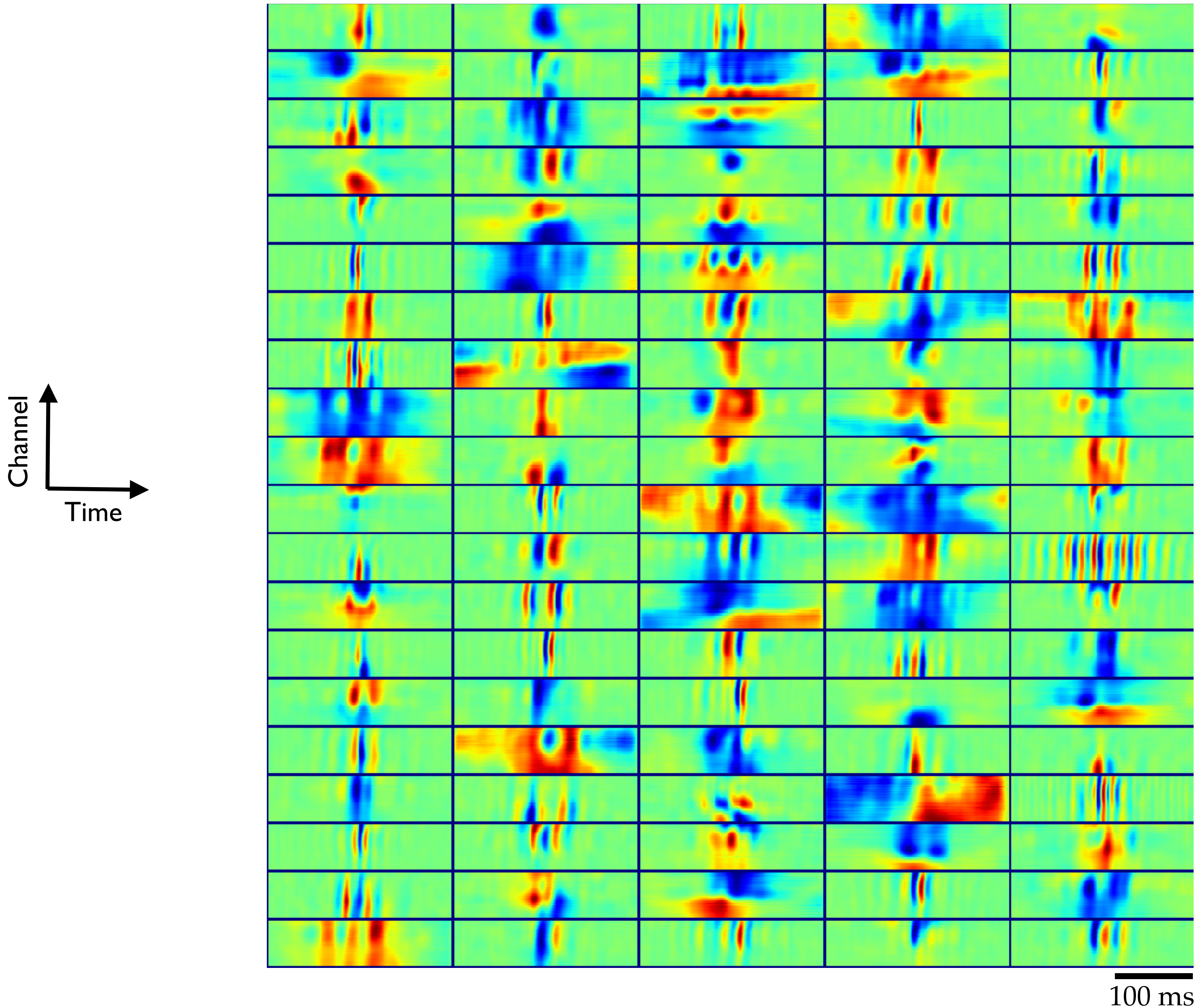
Local field potential (LFP)



Learned basis for high-pass filtered polytrode data



Learned basis for low-pass filtered polytrode data



Human MEG

(Alexandre Gramfort lab, Université Paris-Saclay)

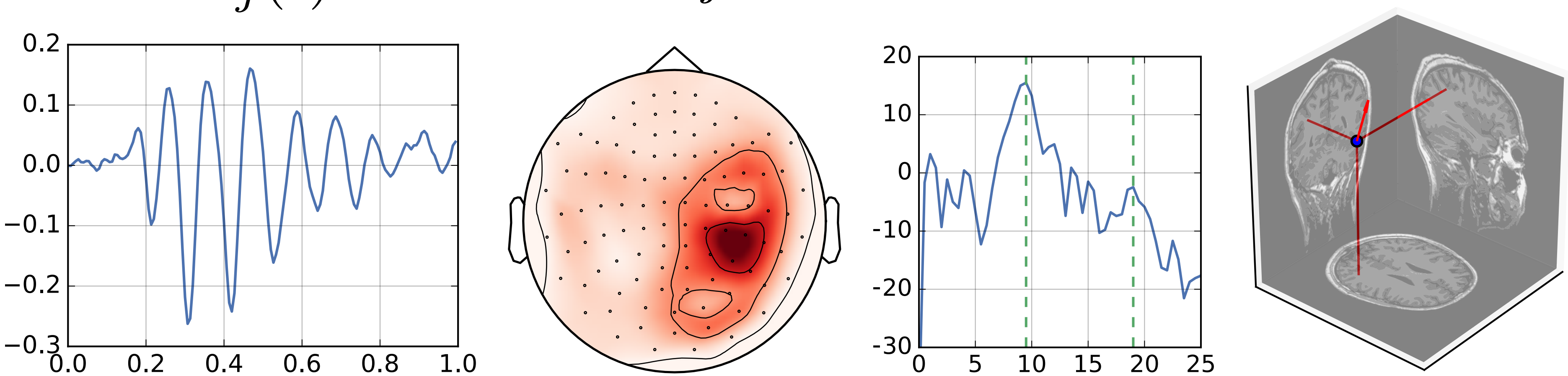
$$y_i(t) = \sum_j \phi_{ij}(t) * x_j(t) + \epsilon_i(t)$$

recorded waveform on sensor *i* — spatiotemporal features — latent cause *j* (sparse) — other stuff

$$\phi_{ij}(t) = u_{ij}\, v_j(t)$$

(assumes space-time separability)

$v_j(t)$ $\quad$ u_{ij}



(a) Temporal waveform (b) Spatial pattern (c) PSD (dB) (d) Dipole fit

Sparse coding of demodulated LFP reveals ‘place cell’ components

(Agarwal, Stevenson, Berényi, Mizuseki, Buzsáki & Sommer, 2014)