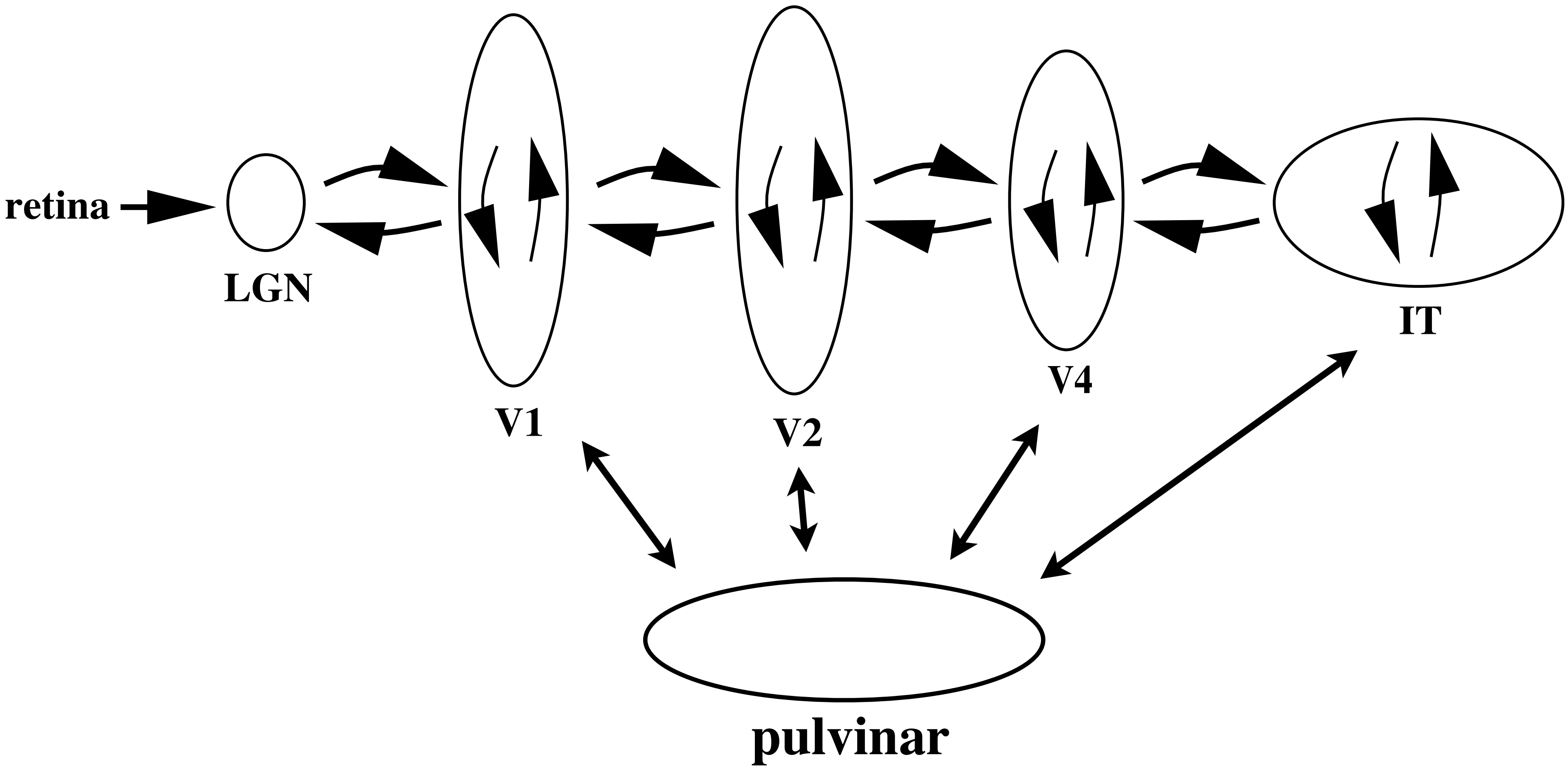
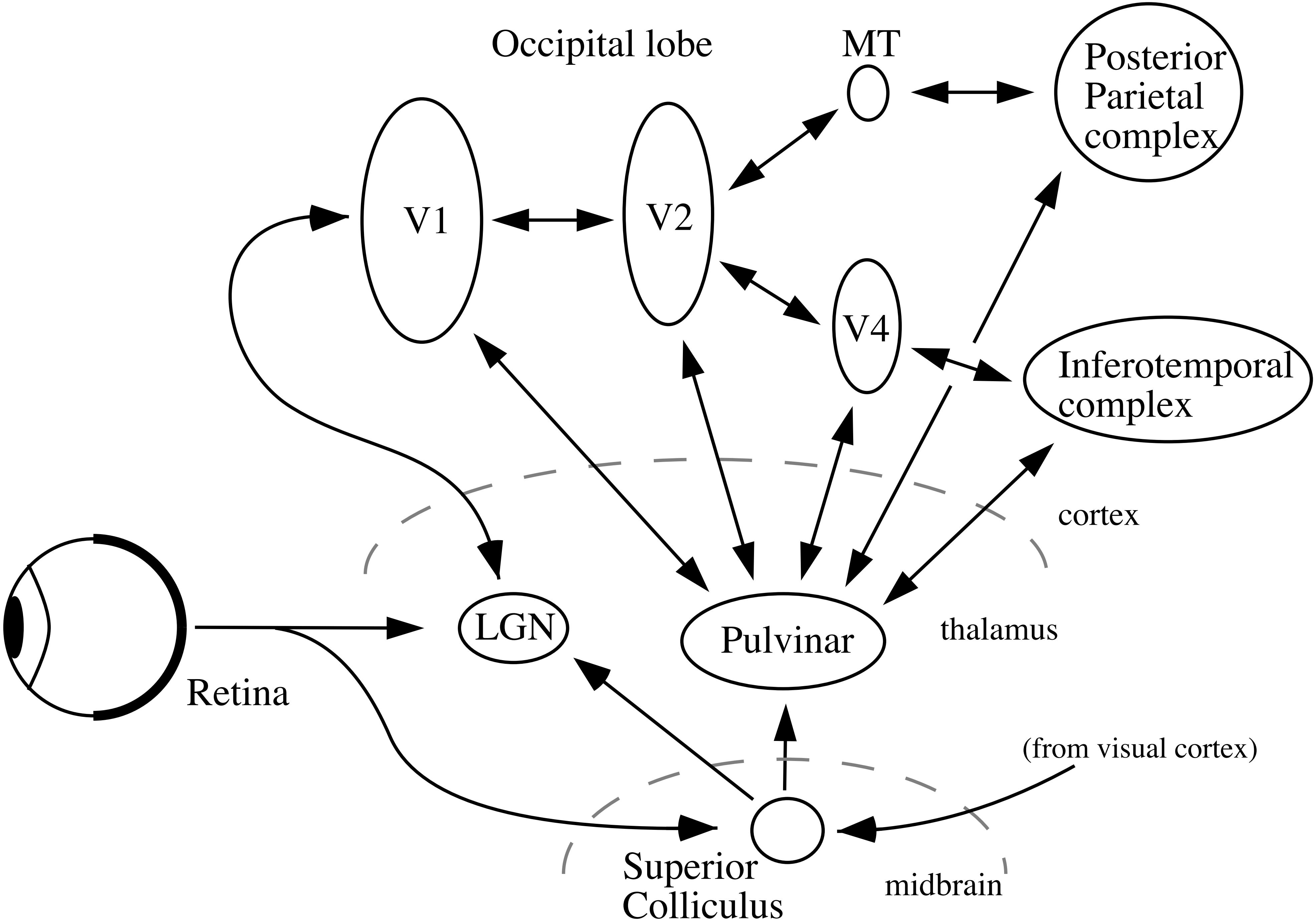
Vision as inference

- Generative models and Bayesian inference
- Sparse, overcomplete representations - a model for V1?
- Hierarchical models for capturing dependencies among sparse components
- Bilinear models and invariance (slow feature analysis)

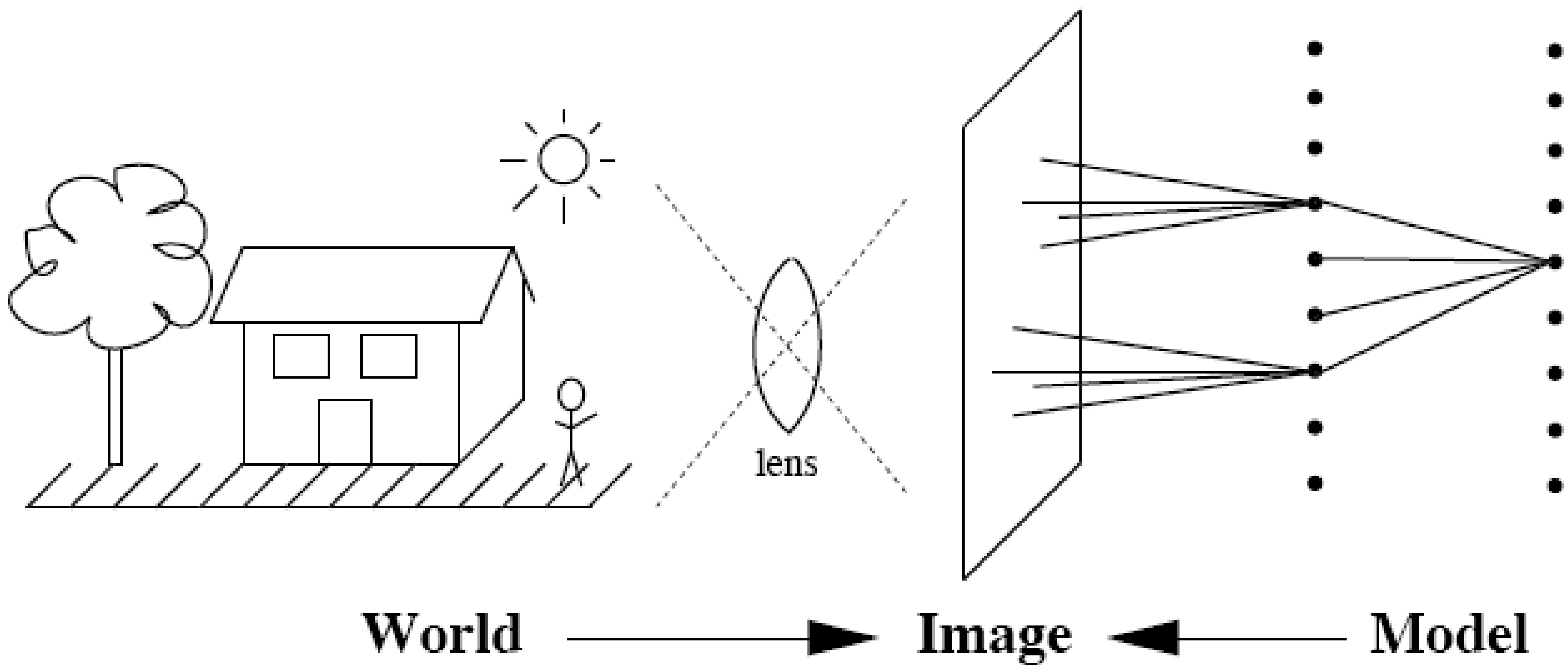
Recurrent computation is pervasive throughout cortex





Occipital lobe
MT
Posterior
Parietal
complex
V1
V2
V4
Inferotemporal
complex
cortex
LGN
Pulvinar
thalamus
Retina
(from visual cortex)
Superior
Colliculus
midbrain

Vision as inference



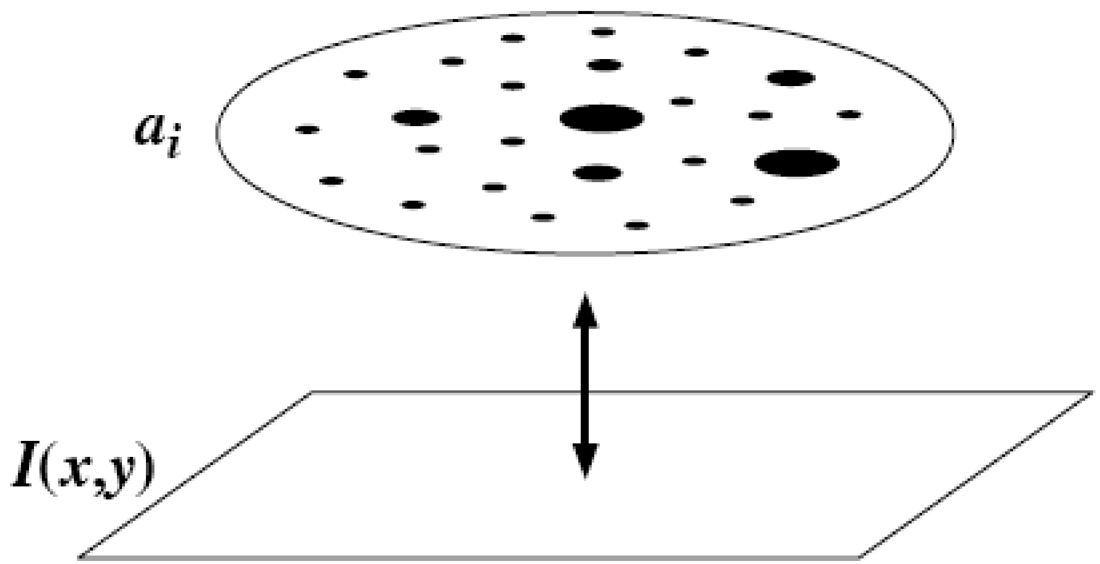
Bayes' rule

$$P(E|D) \propto \underbrace{P(D|E)}_{\substack{\text{how data is}\\ \text{generated by}\\ \text{the environment}}} \times \underbrace{P(E)}_{\substack{\text{prior beliefs}\\ \text{about the}\\ \text{environment}}}$$

$$\begin{aligned} E &= \text{the actual state of the environment} \\ D &= \text{data about the environment} \end{aligned}$$

Sparse component analysis

Olshausen & Field (1996), Bell & Sejnowski (1997)



Evidence for sparse coding

- Gilles Laurent - mushroom body, insect
- Michael Fee - HVC, zebra finch
- Tony Zador - auditory cortex, mouse
- Bill Skaggs - hippocampus, primate
- Harvey Swadlow - motor cortex, rabbit
- Michael Brecht - barrel cortex, rat
- Jack Gallant - visual cortex, macaque monkey
- Christof Koch/Itzhak Fried - inferotemportal cortex, human

See:

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

Overcomplete representations

- In oriented, multiscale pyramids, overcompleteness is necessary to ascribe meaning to coefficients (Simoncelli, Freeman, Adelson, and Heeger, 1992).

- Overcomplete time-frequency dictionaries are best able to reveal time-frequency structure embedded in signals (Chen, Donoho, Saunders, 2001).

- Area V1 is highly overcomplete, by approximately 25:1 (in cat).

Shiftable Multiscale Transforms

Eero P. Simoncelli, William T. Freeman, Edward H. Adelson, and David J. Heeger

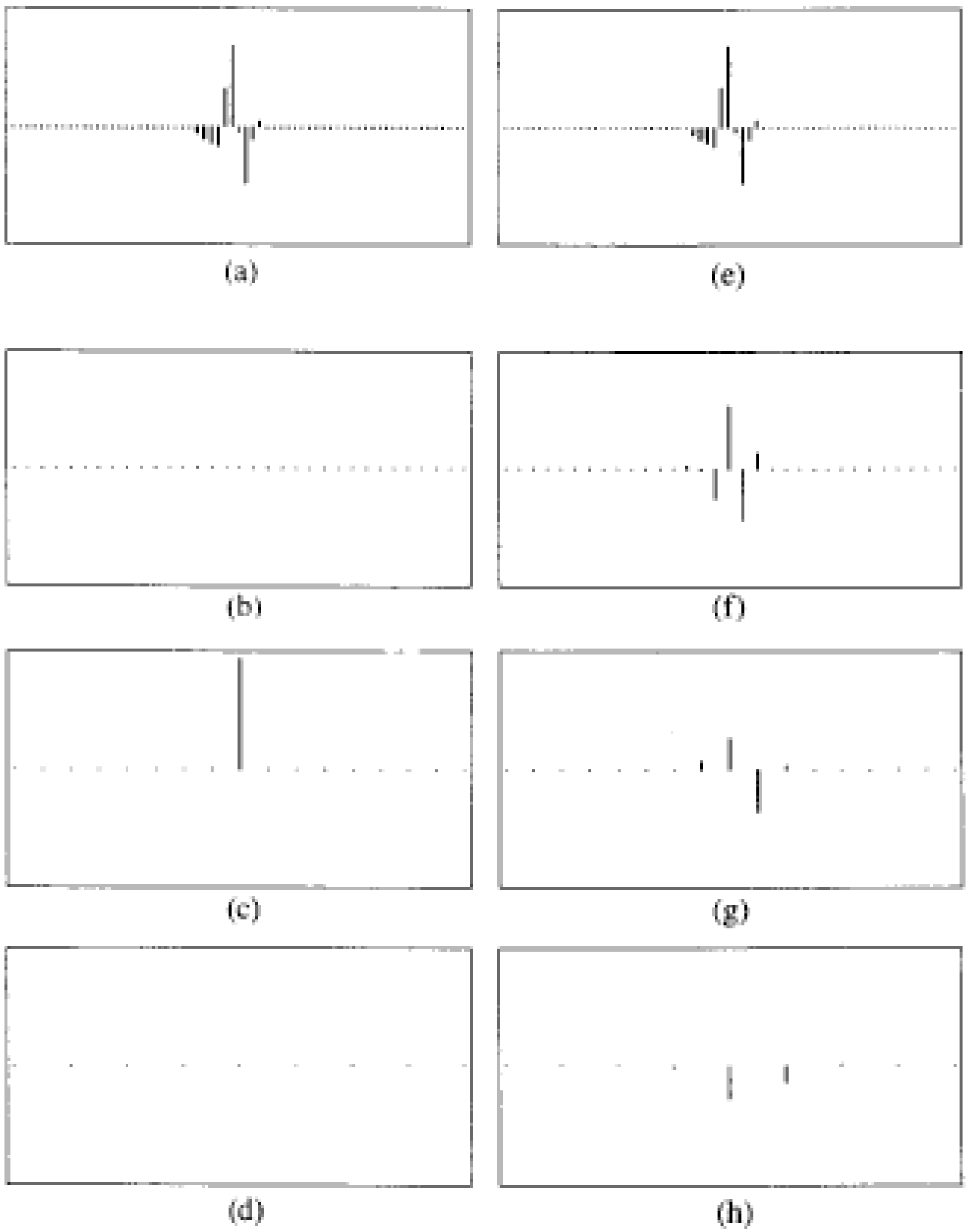


Fig. 1. Effect of translation on the wavelet representation of a signal. (a) Input signal, which is equal to one of the wavelet basis functions. (b)–(d) Decomposition of the signal into three wavelet subbands. Plotted are the coefficients of each subband. Dots correspond to zero-value coefficients. (e) Same input signal, translated one sample to to the right. (f)–(h) Decomposition of the shifted signal into three wavelet subbands. Note the drastic change in the coefficients of the transform, both within and between subbands.

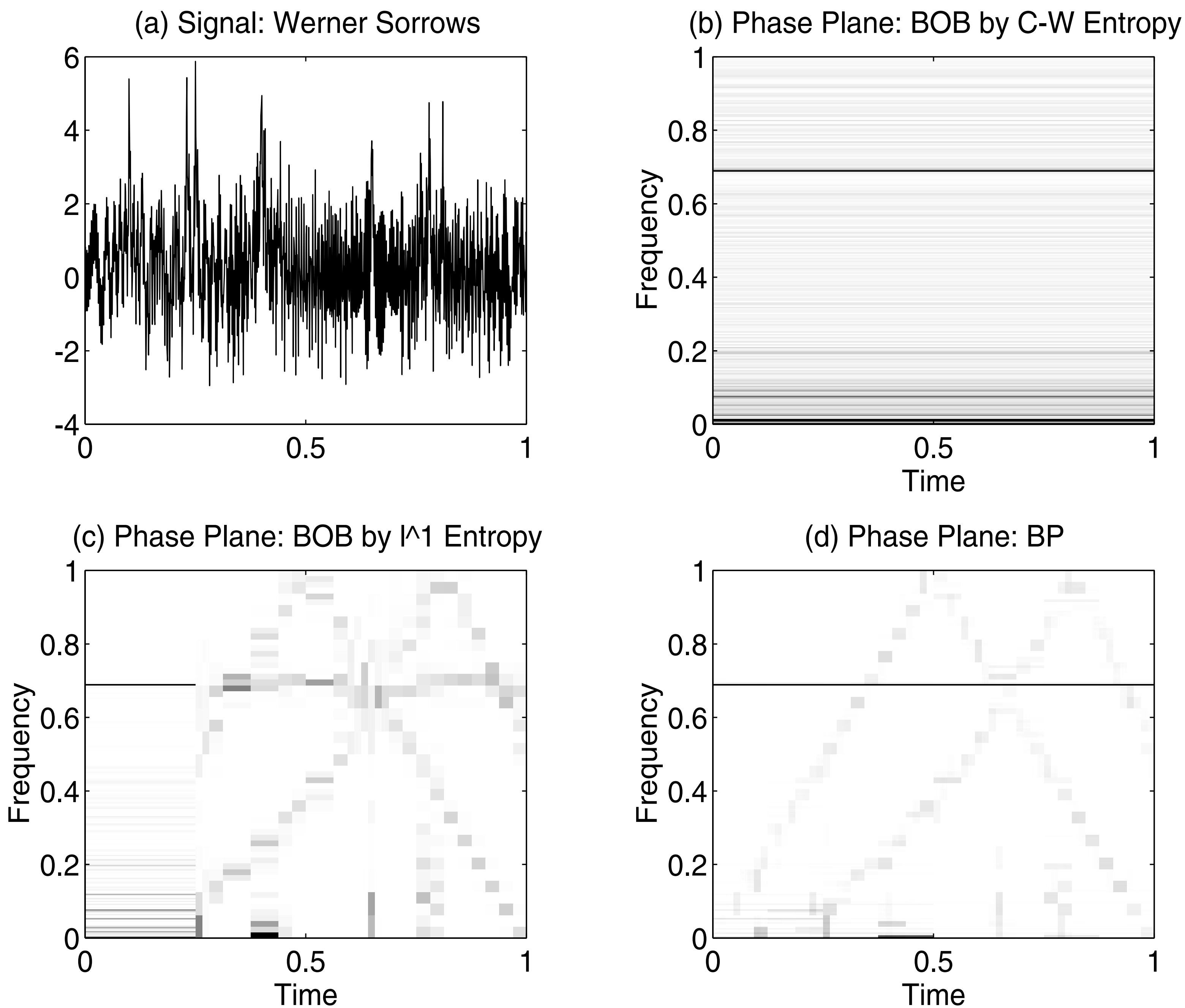

(a) Signal: Werner Sorrows
6
4
2
0
-2
-4
0
0.5
1
(b) Phase Plane: BOB by C-W Entropy
1
0.8
0.6
0.4
0.2
0
Frequency
0
0.5
1
Time
(c) Phase Plane: BOB by l^1 Entropy
1
0.8
0.6
0.4
0.2
0
Frequency
0
0.5
1
Time
(d) Phase Plane: BP
1
0.8
0.6
0.4
0.2
0
Frequency
0
0.5
1
Time

Image model

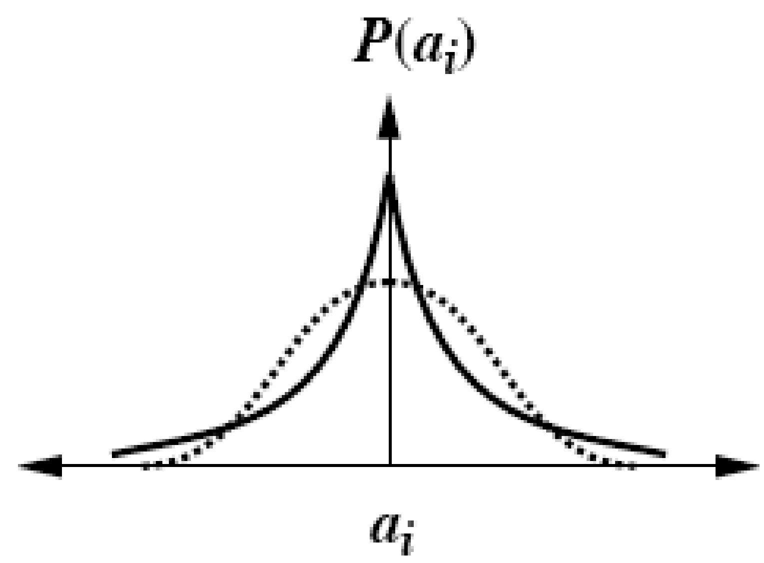
$$I(x, y) = \sum_i a_i \, \phi_i(x, y) + \nu(x, y) \; .$$

Goal: Find a set of basis functions $\{\phi_i\}$ for representing natural images such that the coefficients a_i are as sparse and statistically independent as possible.

Prior

- Factorial: $P(\mathbf{a}|\theta) = \prod_i P(a_i|\theta)$

- Sparse: $P(a_i|\theta) = \frac{1}{Z_S} e^{-S(a_i)}$



Inference (perception)

MAP estimate:

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} P(\mathbf{a}|\mathbf{I},\theta)$$

$$P(\mathbf{a}|\mathbf{I},\theta) \propto P(\mathbf{I}|\mathbf{a},\theta)\, P(\mathbf{a}|\theta)$$

Energy function:

$$\begin{aligned} E(\mathbf{I},\mathbf{a}) &= -\log P(\mathbf{a}|\mathbf{I},\theta) \\ &= \frac{\lambda_N}{2}|\mathbf{I} - \Phi\mathbf{a}|^2 + \sum_i S(a_i) + \text{const.} \end{aligned}$$

Dynamics:

$$\begin{aligned} \dot{\mathbf{a}} &\propto -\frac{\partial E}{\partial \mathbf{a}} \\ &= \lambda_N\, \Phi^T\, \mathbf{I} - \lambda_N \Phi^T \Phi\, \mathbf{a} - S'(\mathbf{a}) \end{aligned}$$

Learning

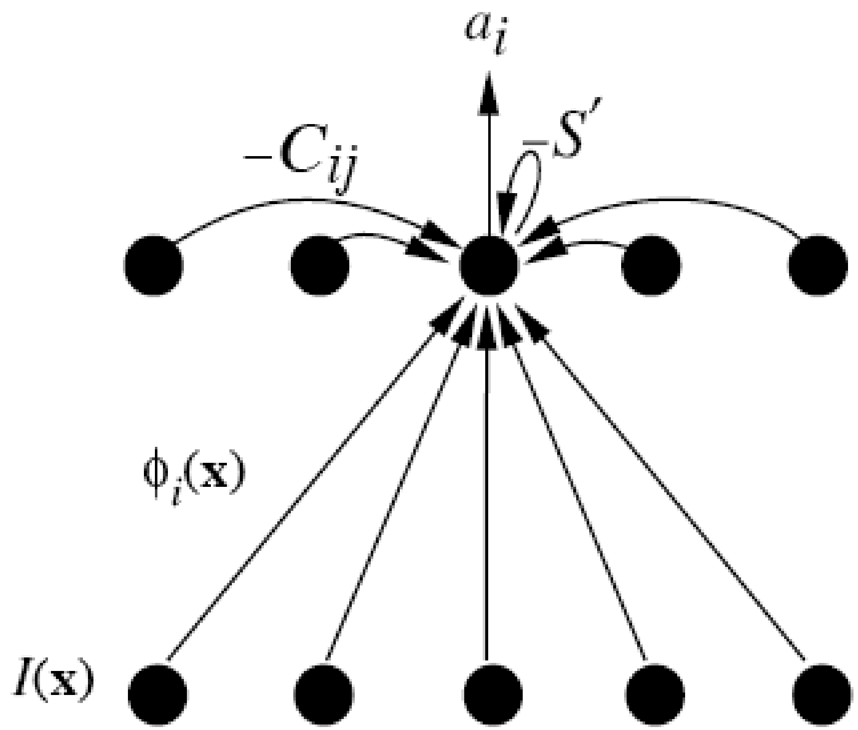
Objective function:

$$\begin{aligned}\mathcal{L} &= \langle \log P(\mathbf{I}|\theta) \rangle \\ P(\mathbf{I}|\theta) &= \int P(\mathbf{I}|\mathbf{a},\theta)\, P(\mathbf{a}|\theta)\, d\mathbf{a}\end{aligned}$$

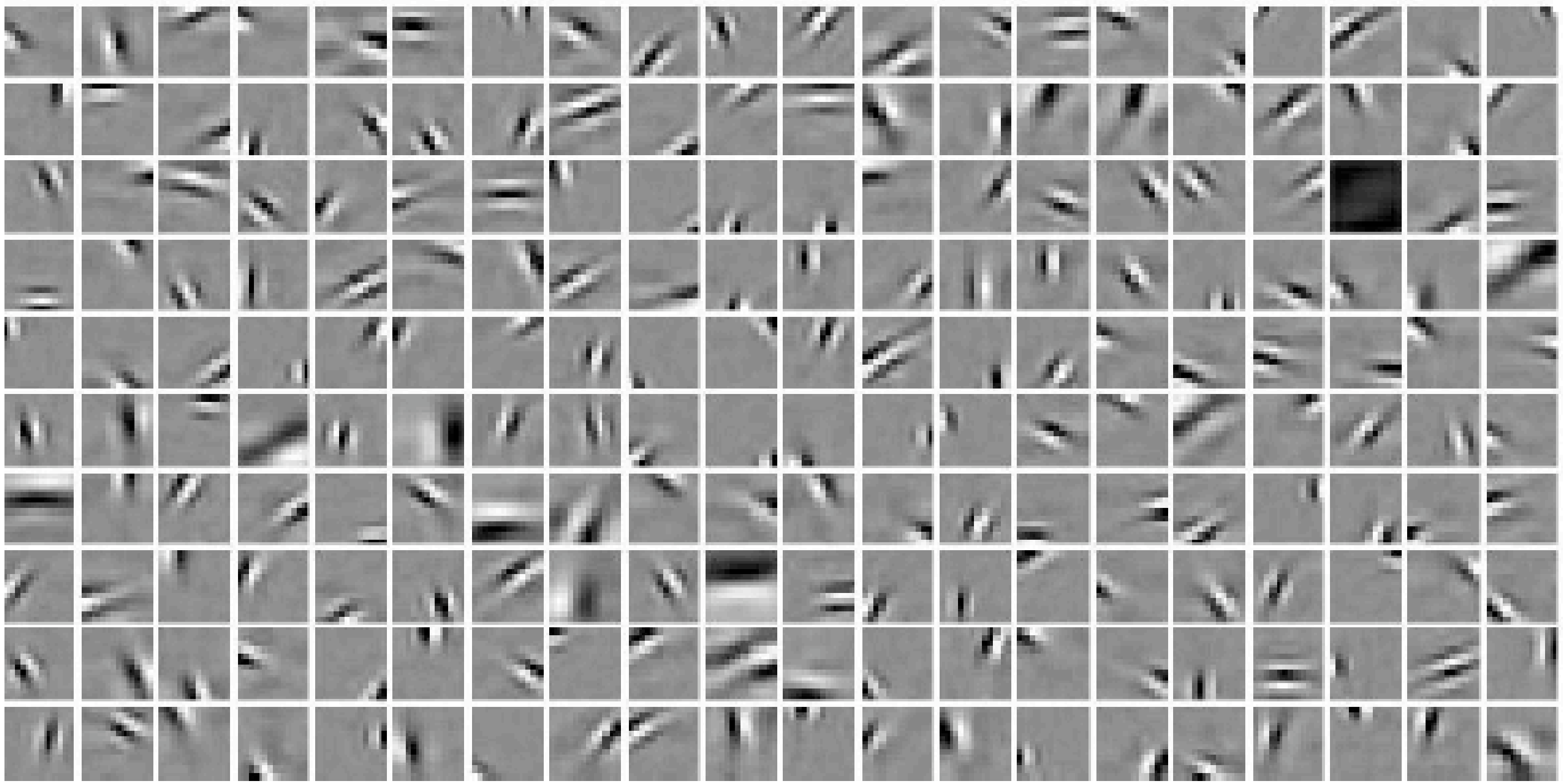
Learning rule:

$$\begin{aligned}\Delta\Phi &\propto \frac{\partial \mathcal{L}}{\partial \Phi} \\ &= \lambda_N \int [I - \Phi\,\mathbf{a}]\; P(\mathbf{a}|\mathbf{I},\theta)\, d\mathbf{a}\end{aligned}$$

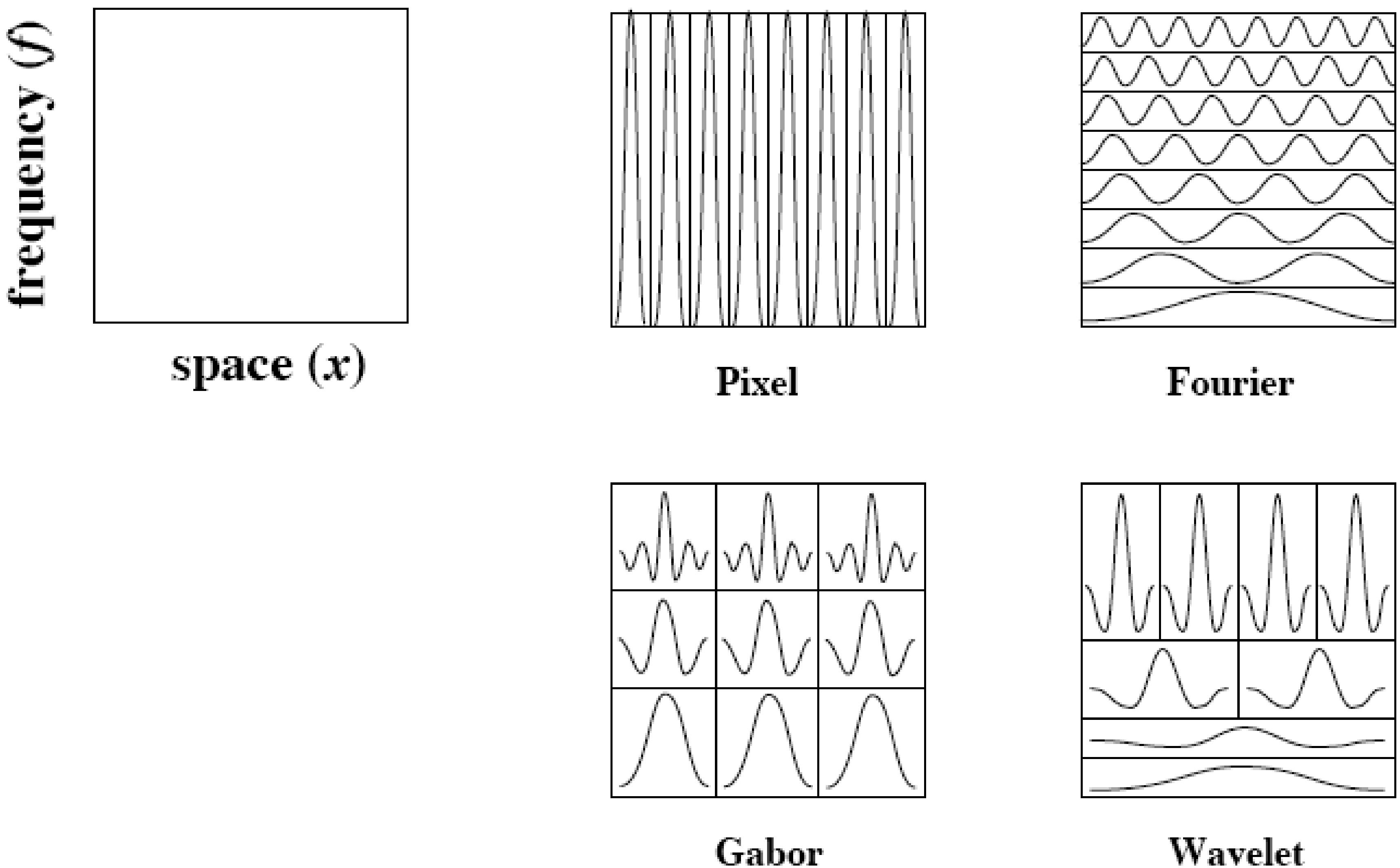
Network implementation



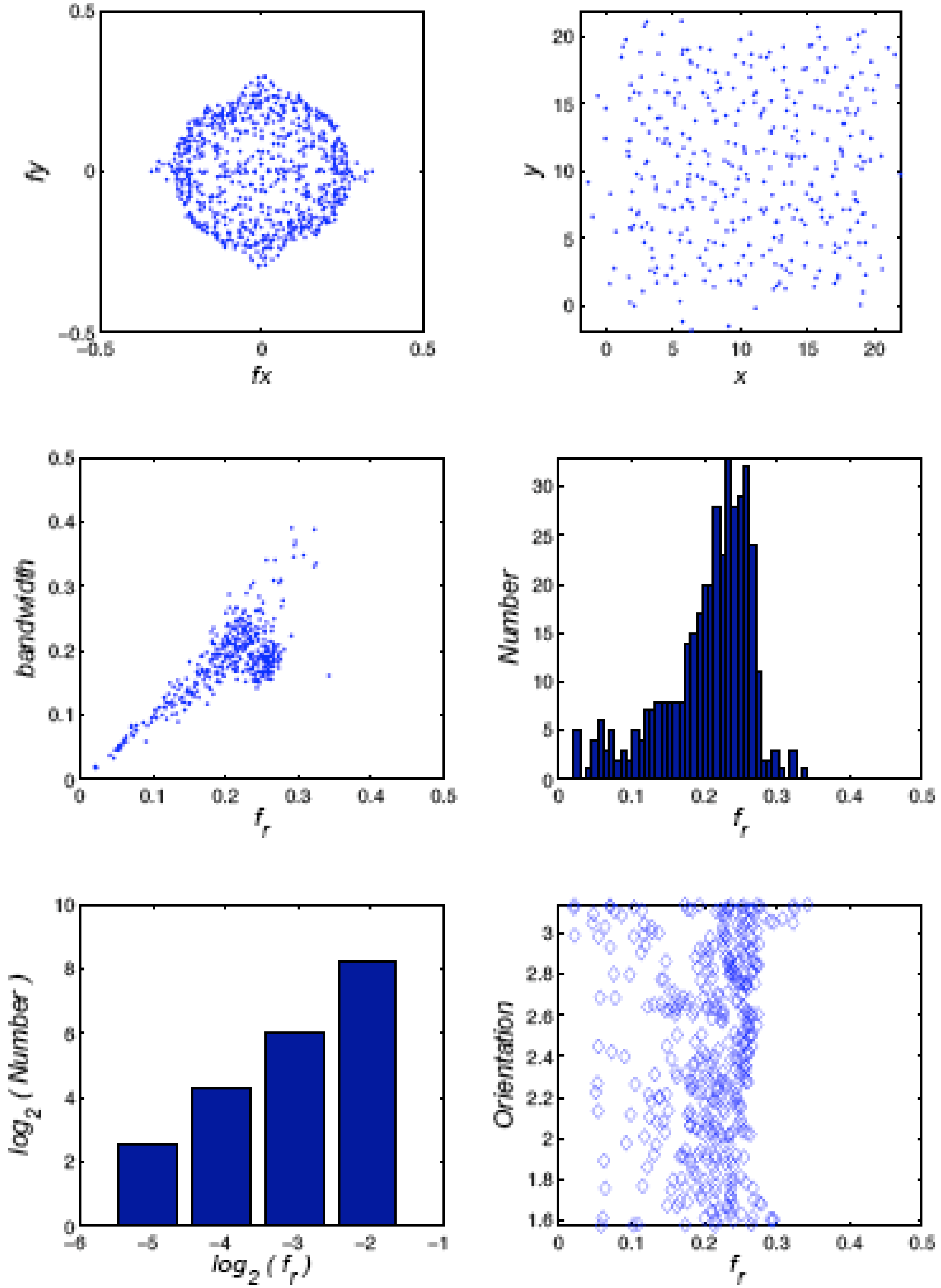
Learned basis functions (200, 12x12)



Scale space (or “phase space”)

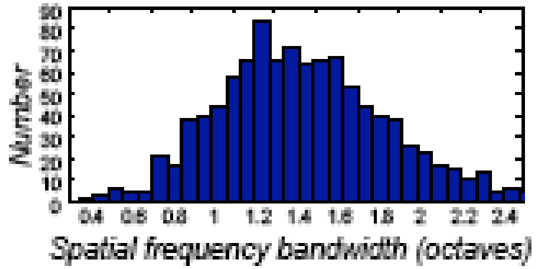


Tiling properties

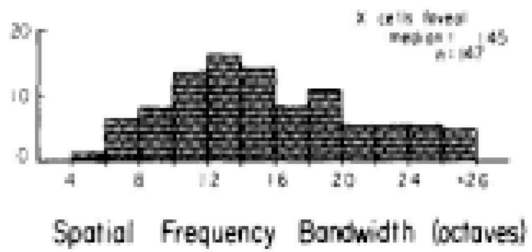


Spatial-frequency bandwidth

Model:

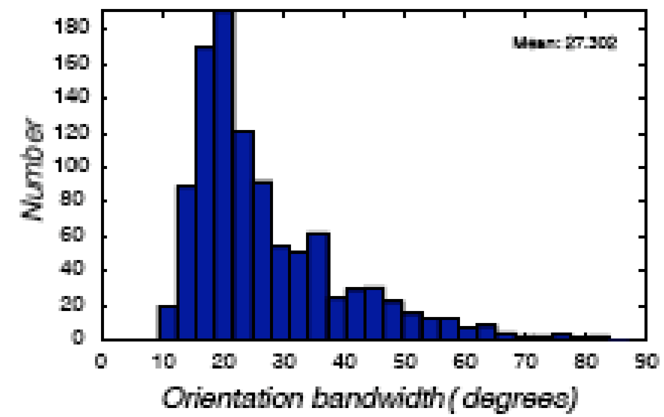


Physiology (DeValois lab):

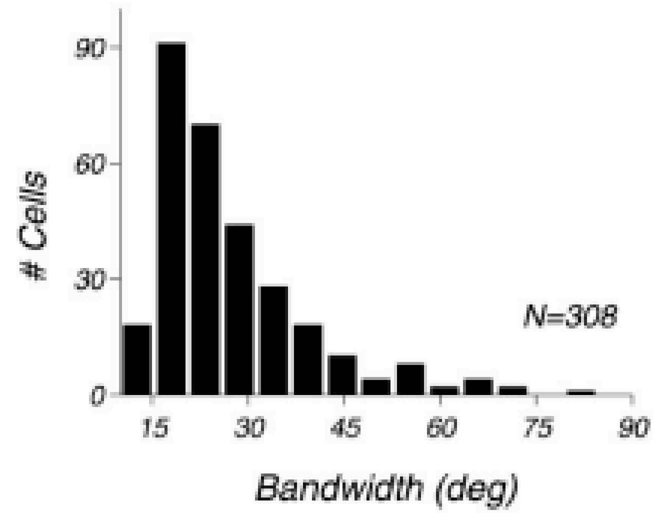


Orientation bandwidth

Model:

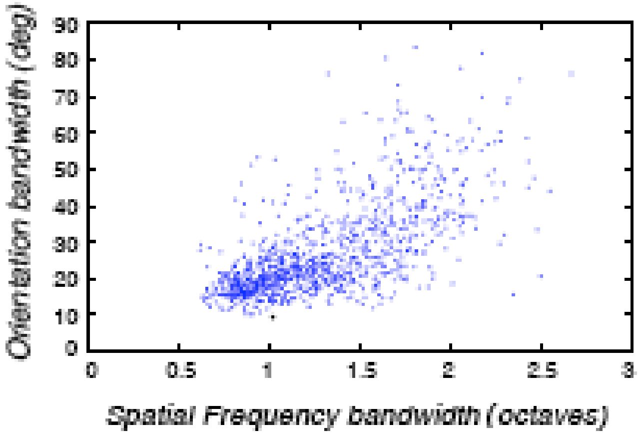


Physiology (Shapley lab):

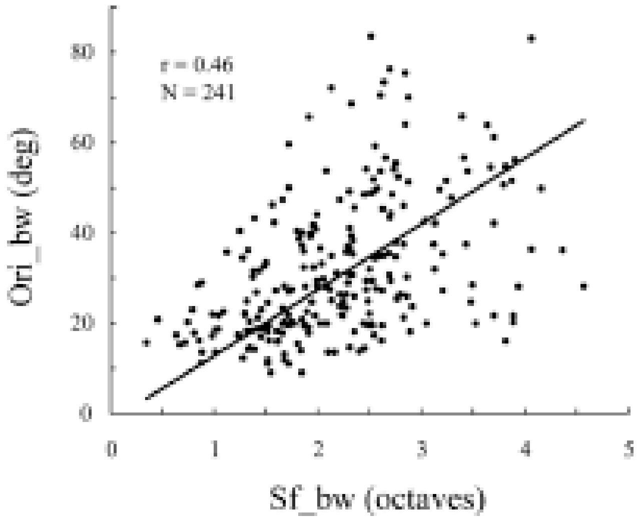


Orientation bandwidth vs. spatial-frequency bandwidth

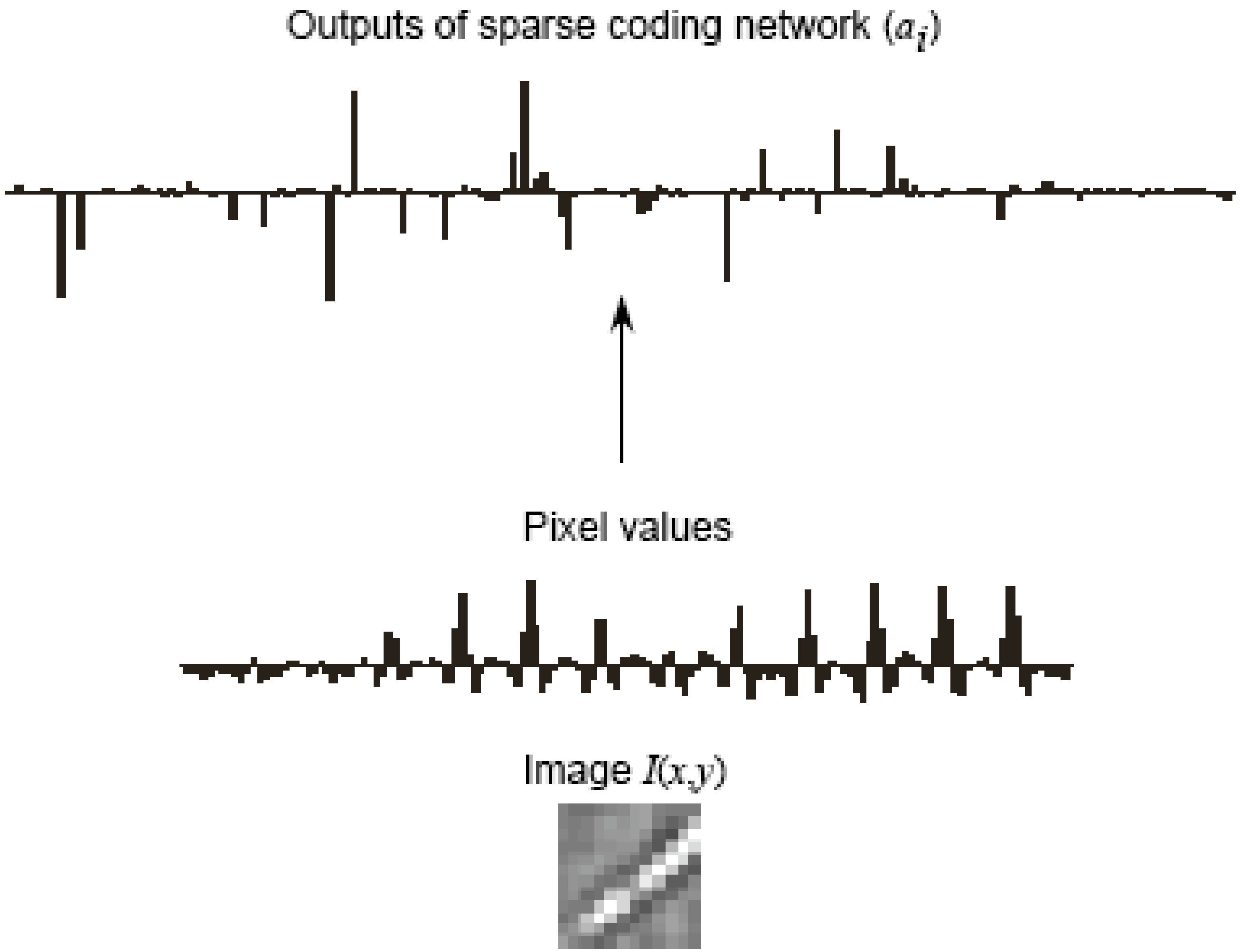
Model:



Physiology (Shapley lab):

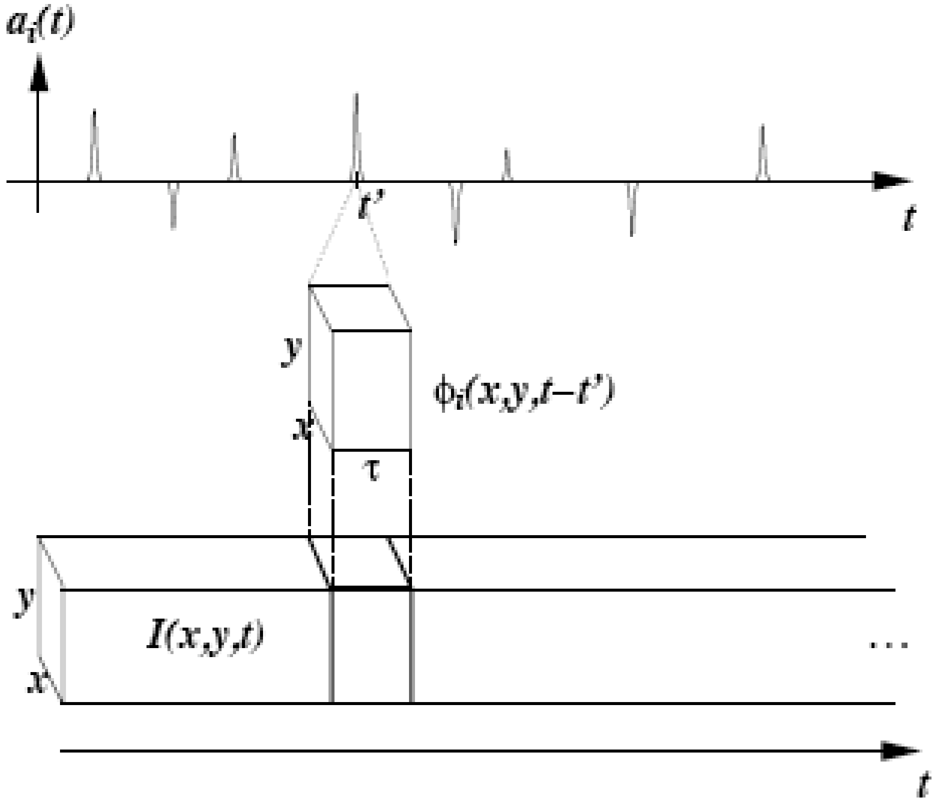


Sparsification



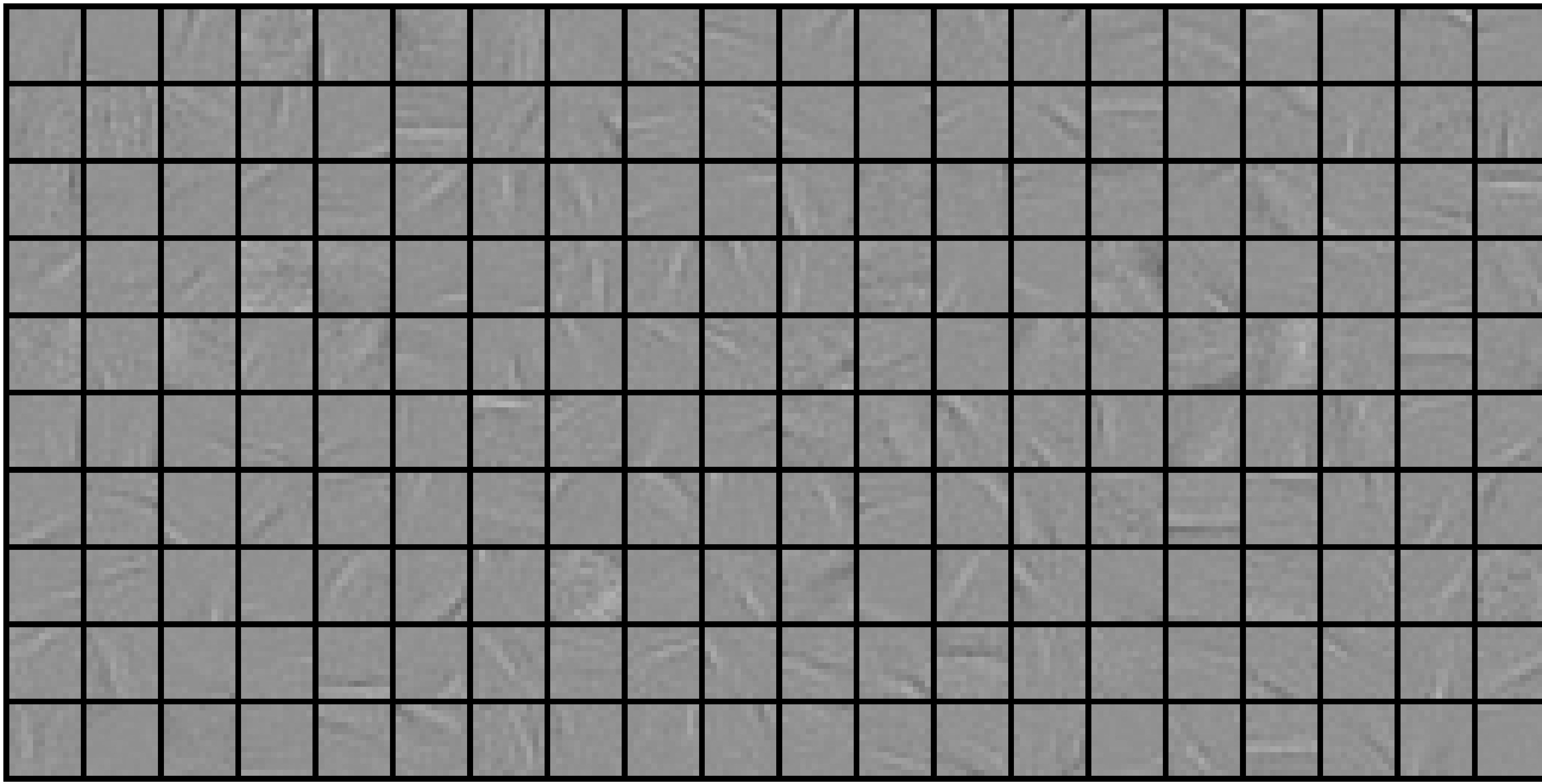
Space-time image model

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



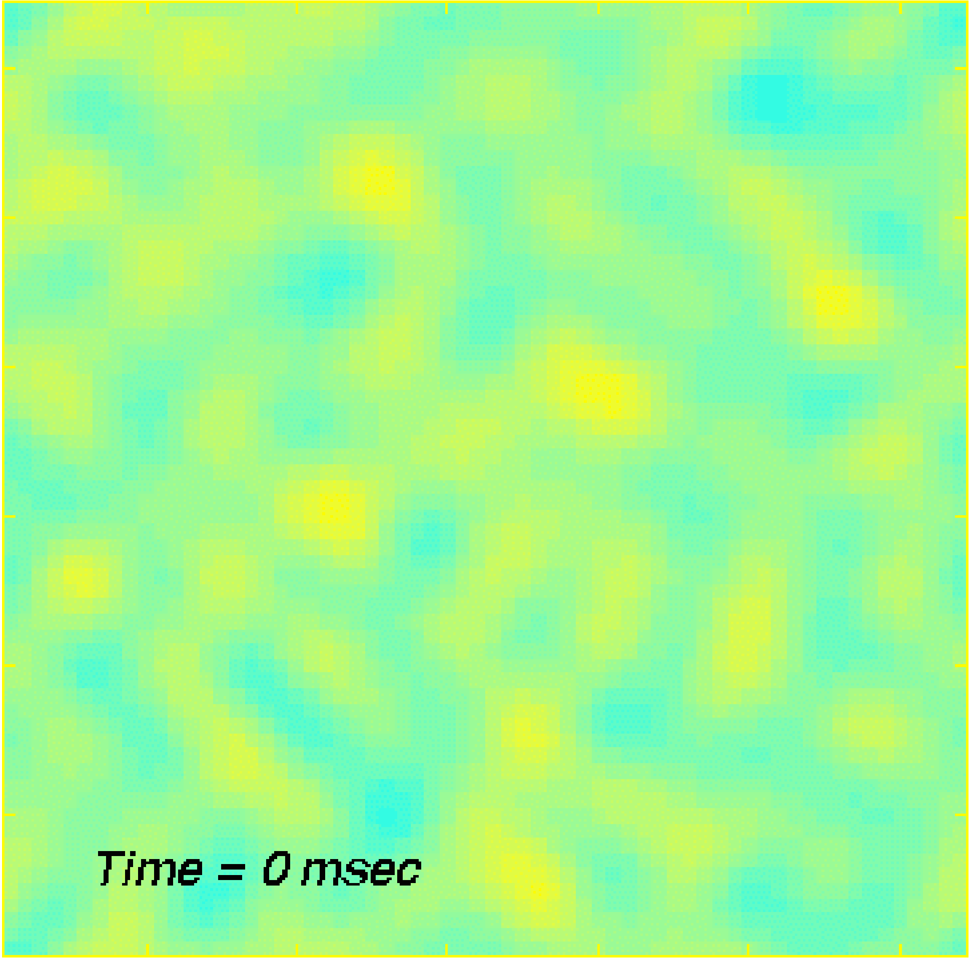
Goal: Find a set of space-time basis functions $\{\phi_i\}$ for representing natural images such that the *time-varying* coefficients $a_i(t)$ are as sparse and statistically independent as possible *over both space and time.*

Learned basis space-time basis functions
(200 bfs, 12 x12 x 7)

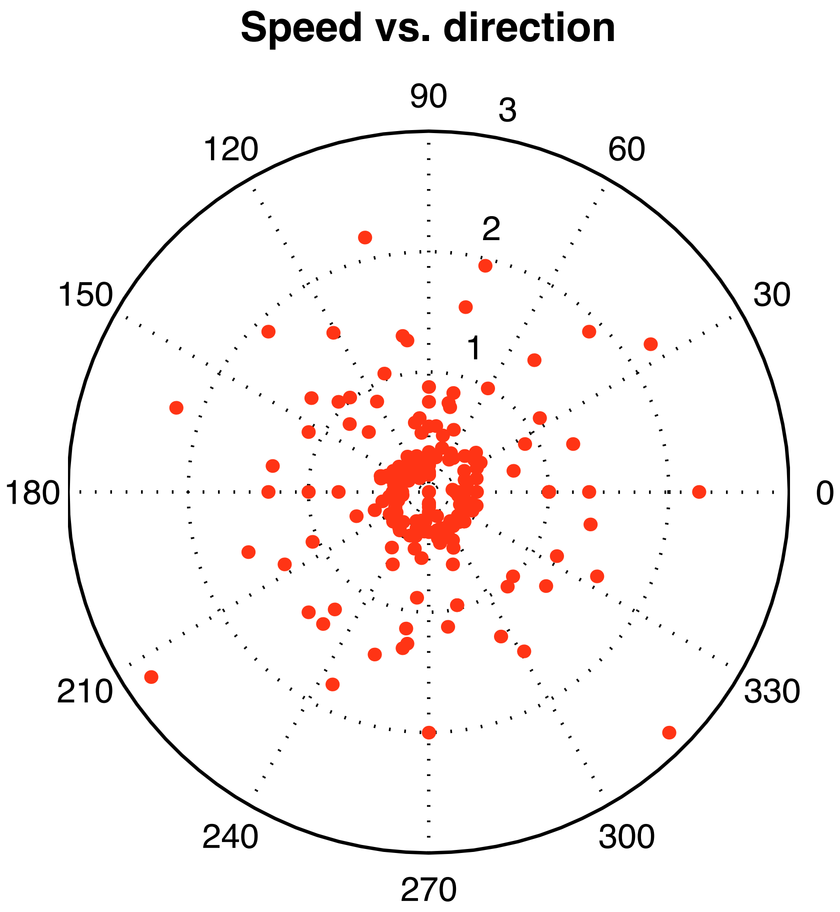


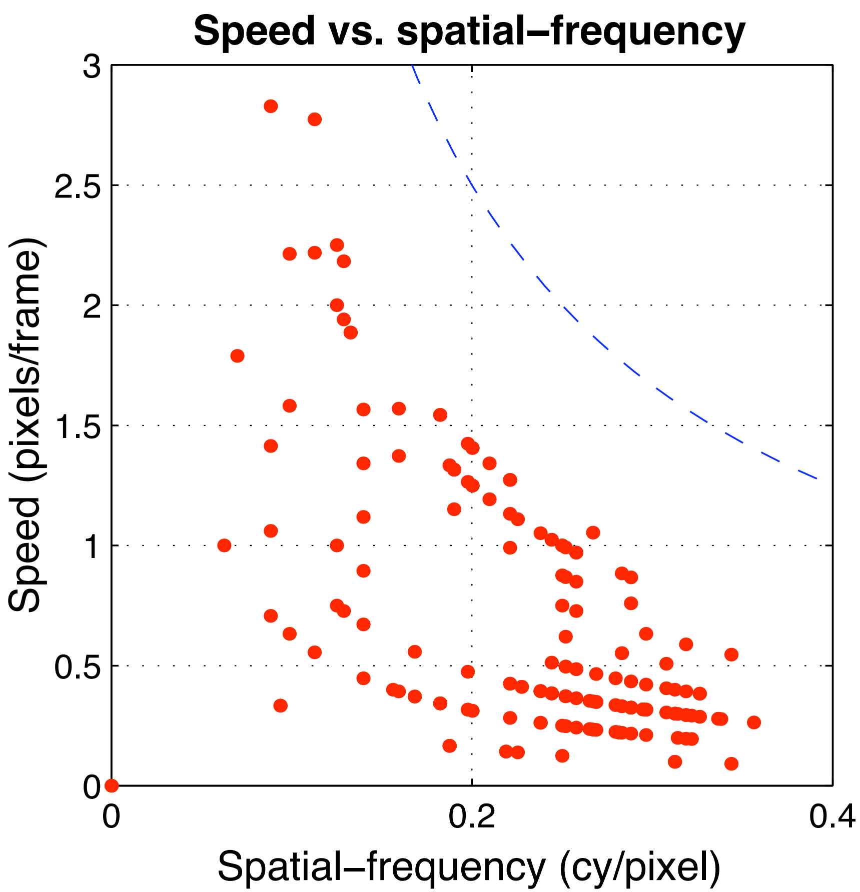
V1 space-time receptive field

(courtesy of Dario Ringach)

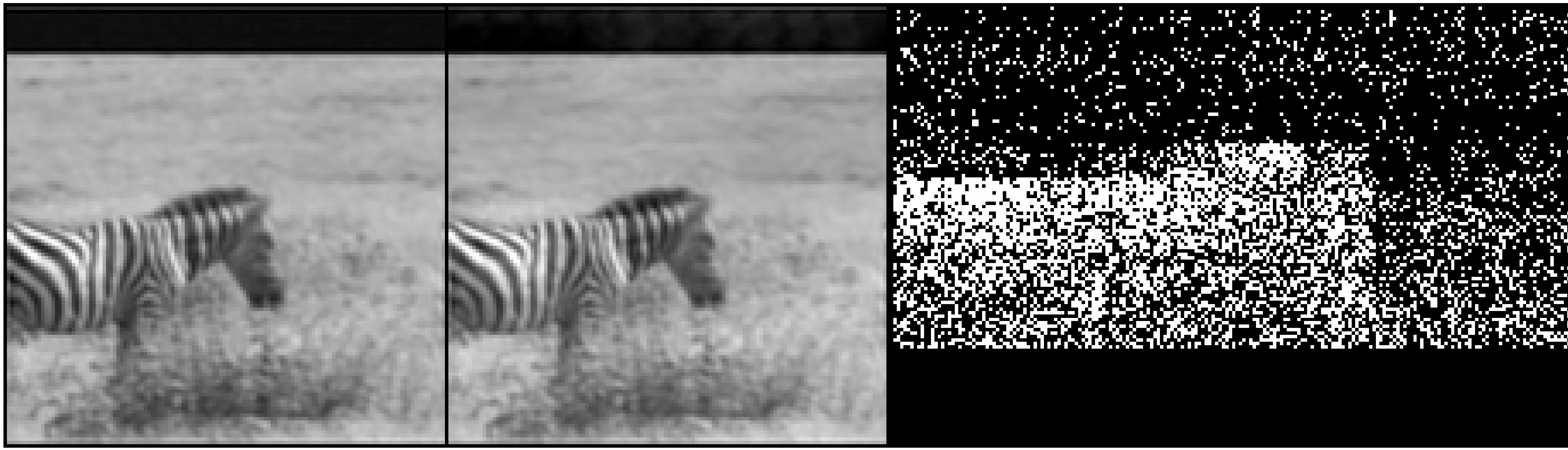


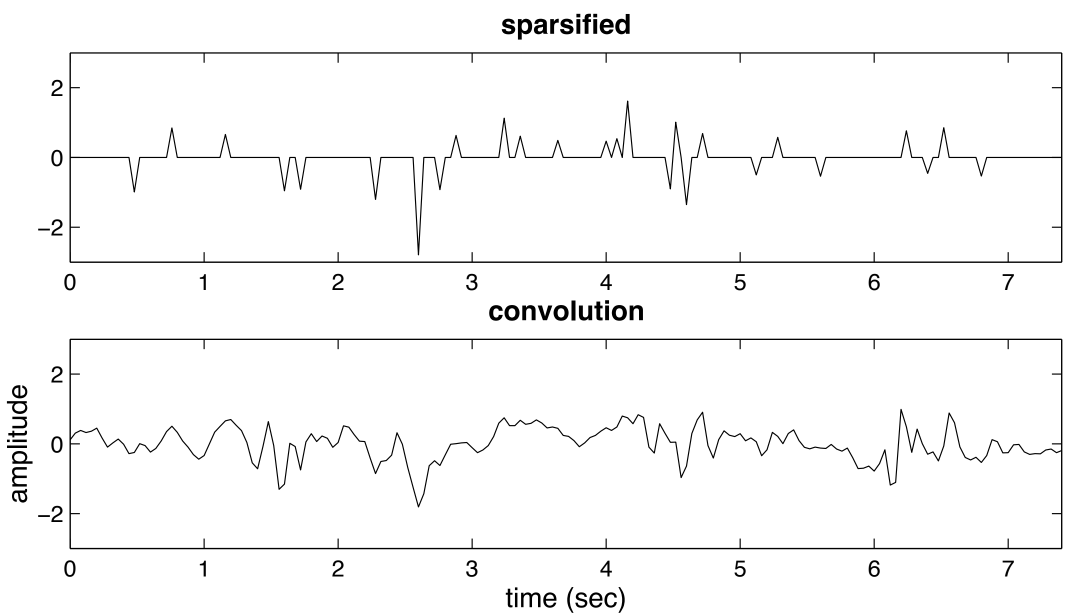
Tiling properties





Sparse coding and reconstruction





Movie synthesis

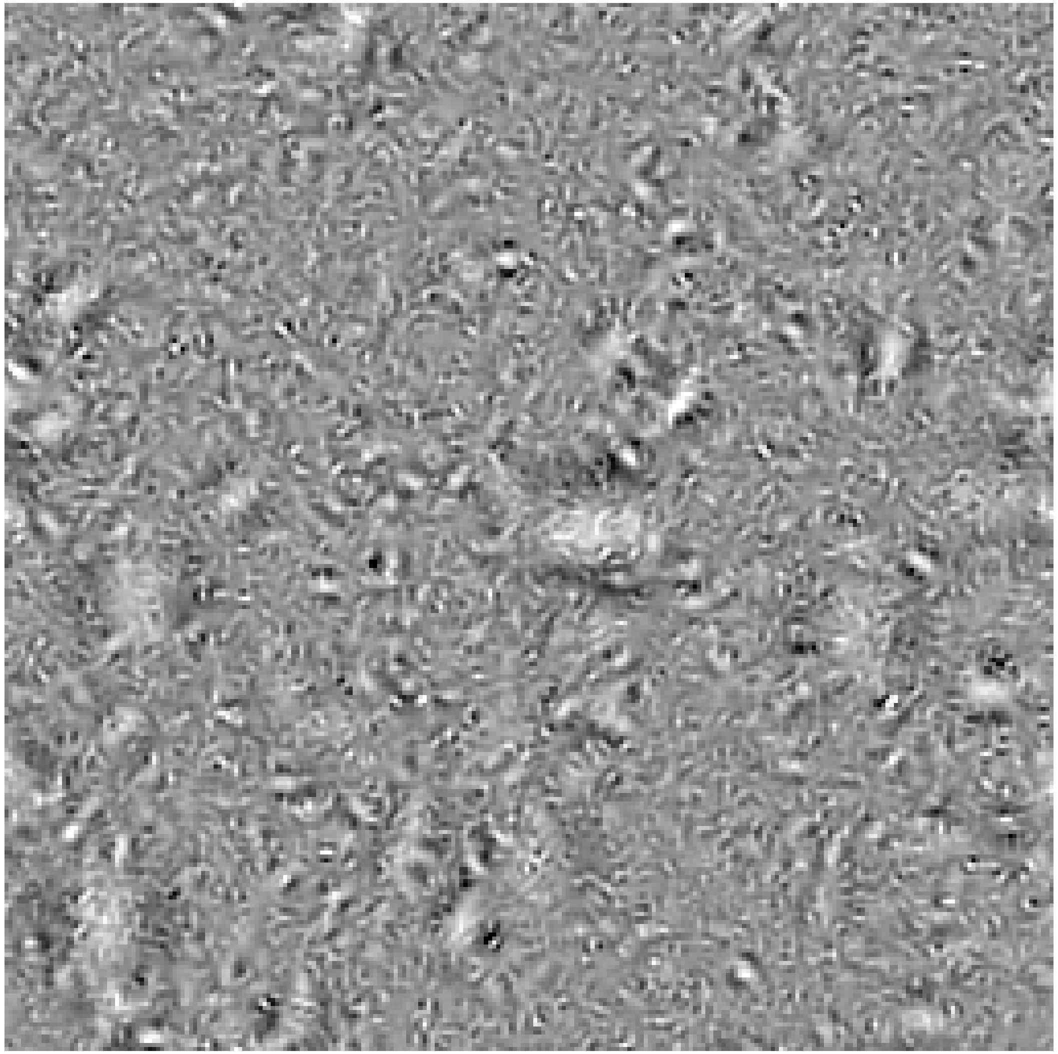
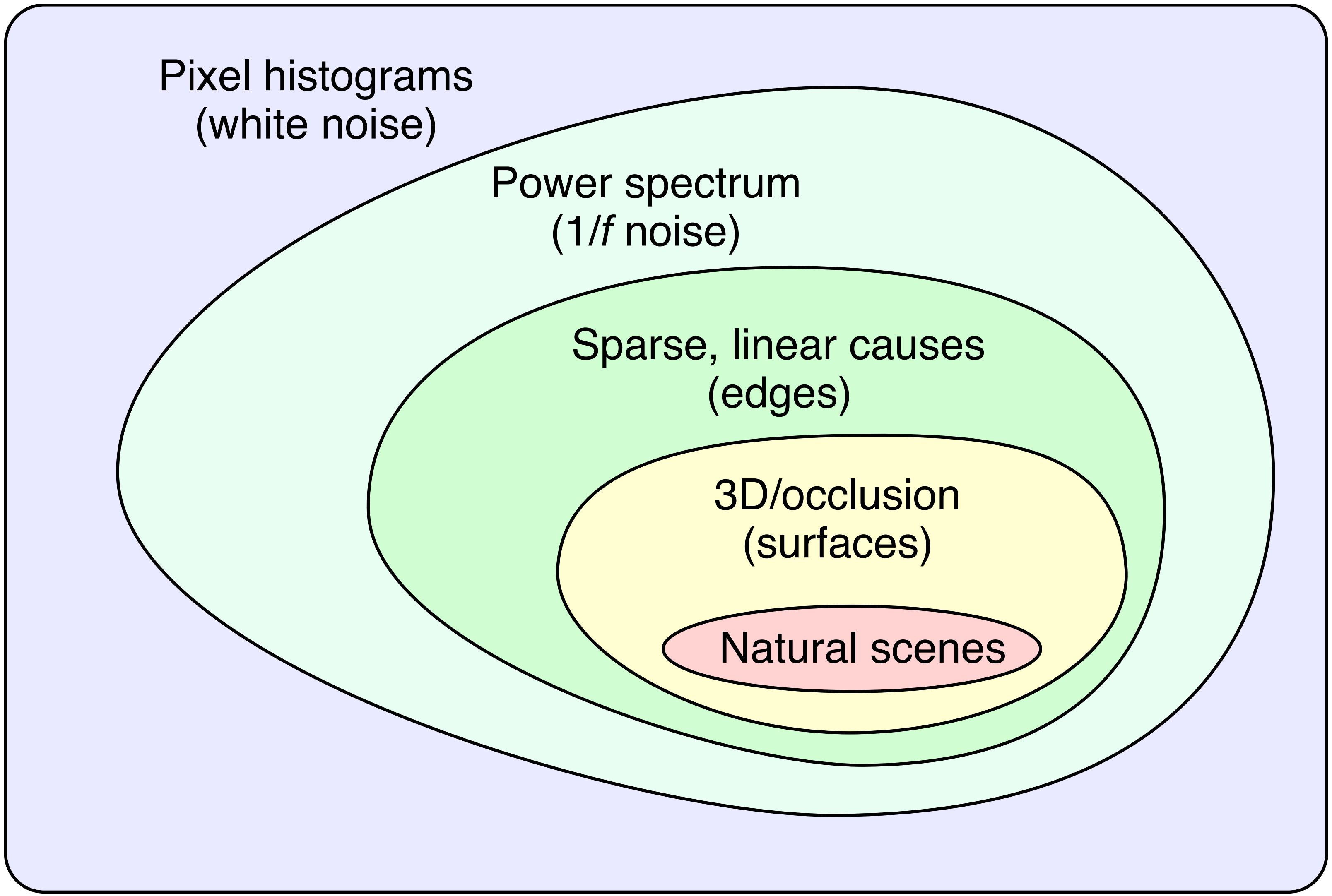


Image models



Using generative models as experimental tools