Natural Scene Statistics and Biological Vision: from Pixels to Percepts

Bruno A. Olshausen

REDWOOD CENTER
for Theoretical Neuroscience

Helen Wills Neuroscience Institute
and School of Optometry, UC Berkeley

Tutorial outline

1. Introduction: Biological vision and theoretical neuroscience

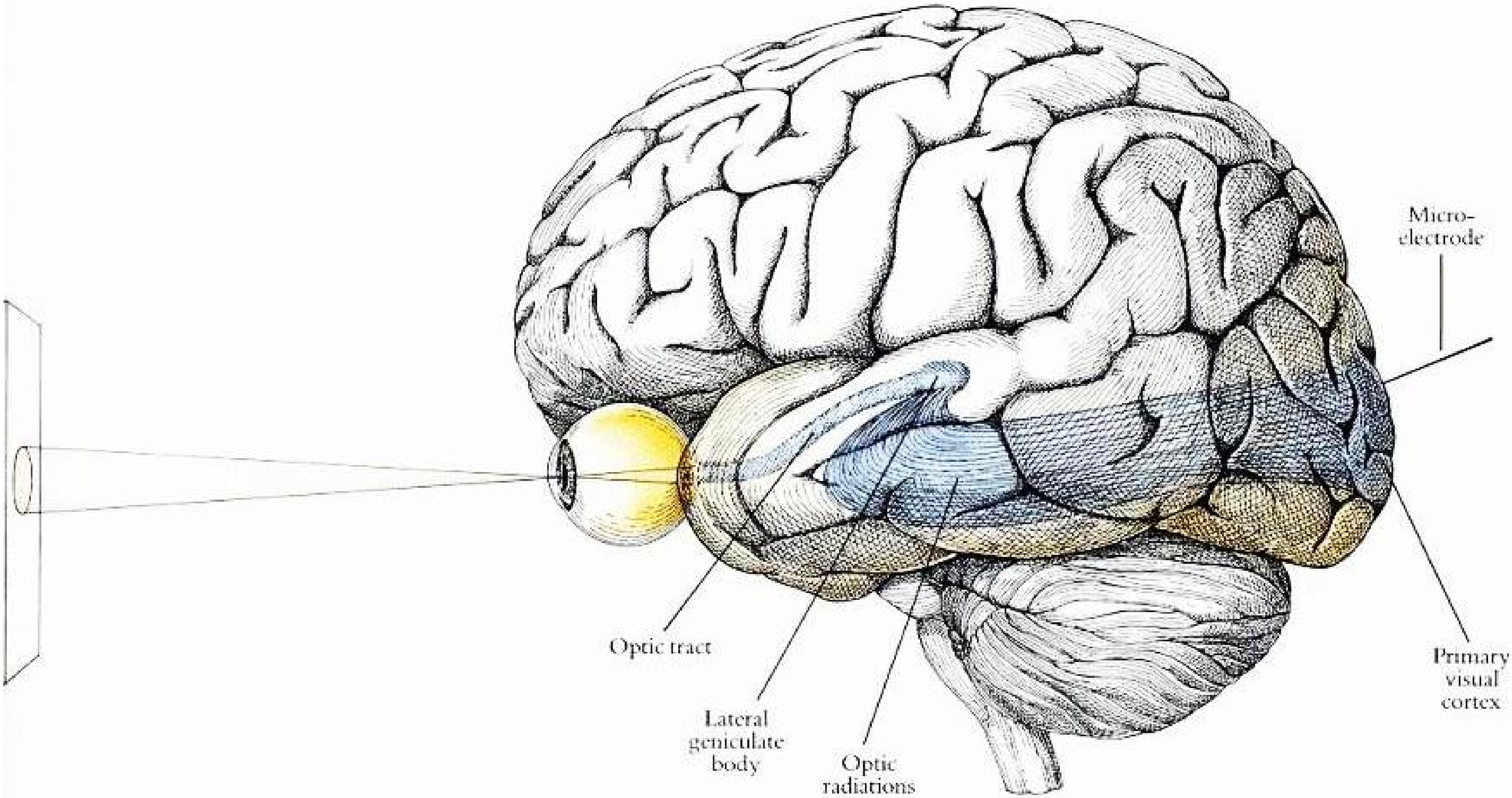
2. Natural image statistics and efficient coding

3. Vision as inference

4. Towards intermediate-level representations

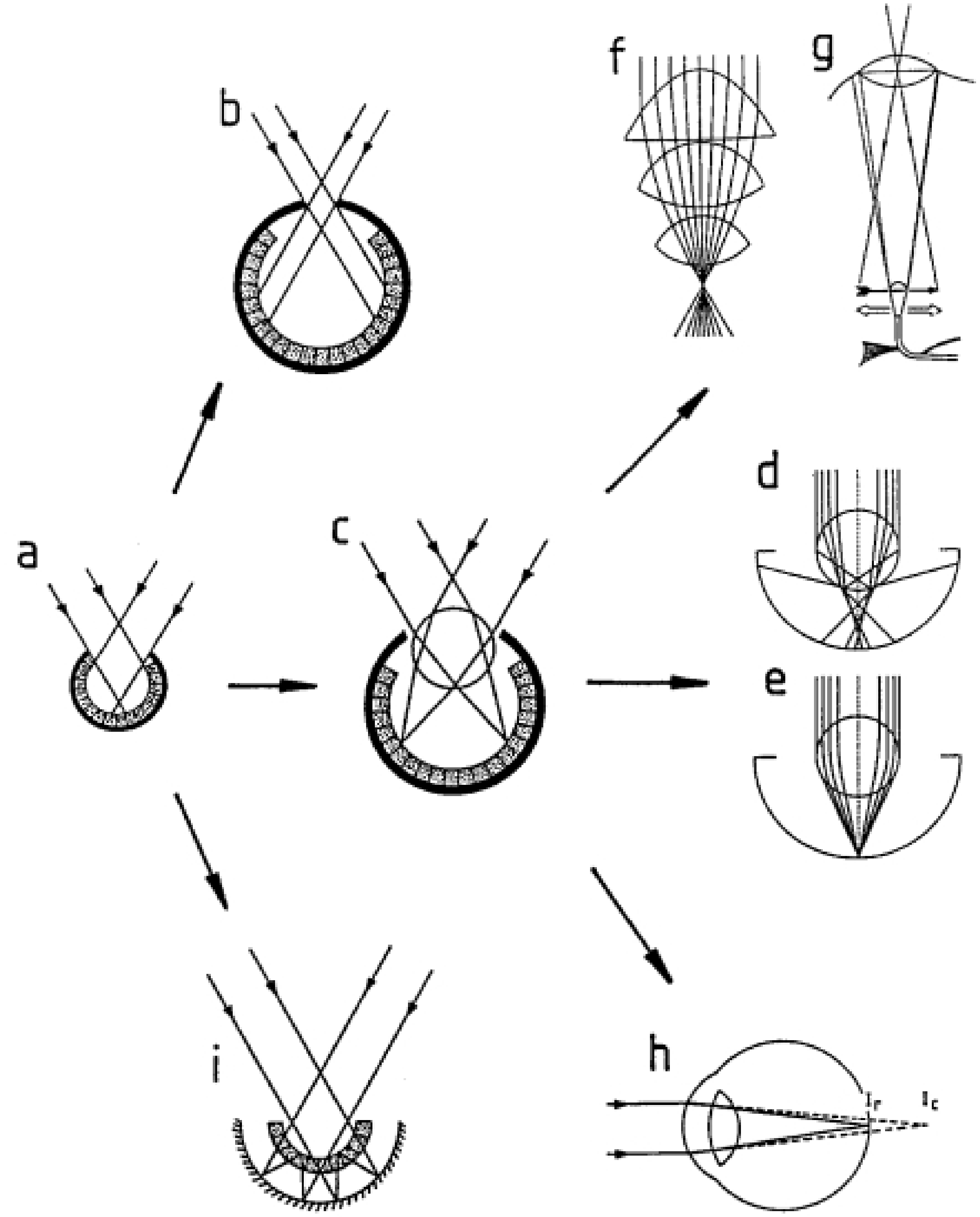
Introduction

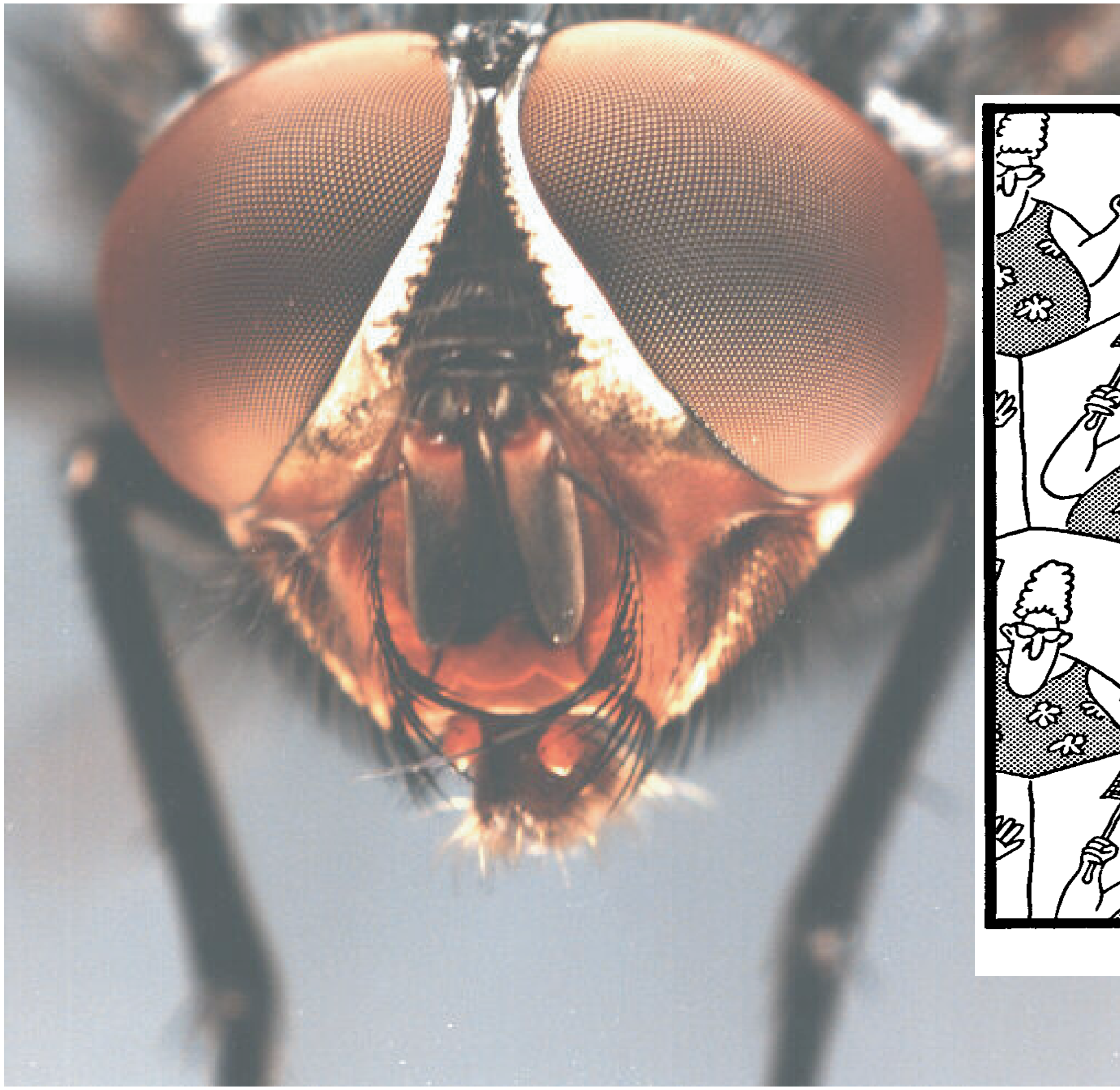
- Biological vision - what we know and and don't know.
- The efficient coding hypothesis

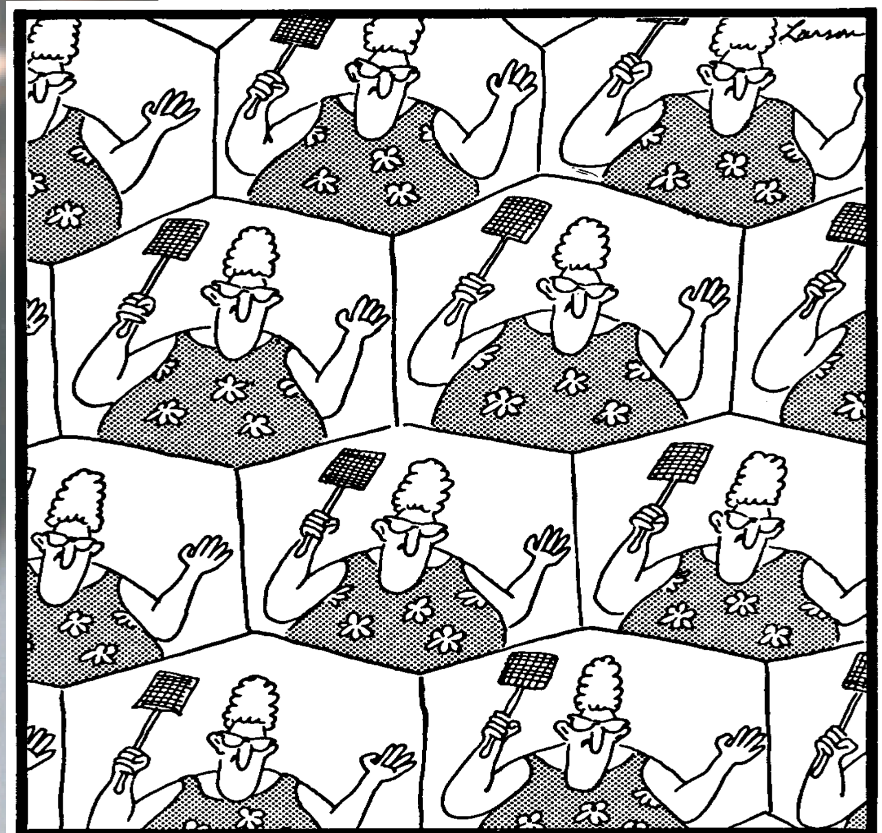

Micro-
electrode
Optic tract
Lateral
geniculate
body
Optic
radiations
Primary
visual
cortex

THE EVOLUTION OF EYES

Michael F. Land　　　　*Russell D. Fernald*







The last thing a fly ever sees

Jumping spiders

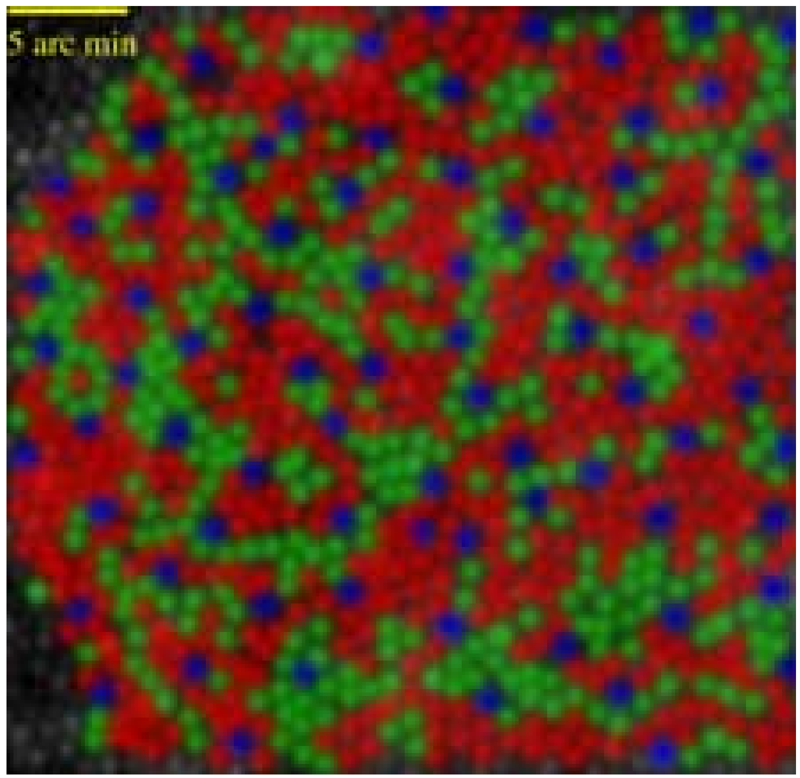
salticidae

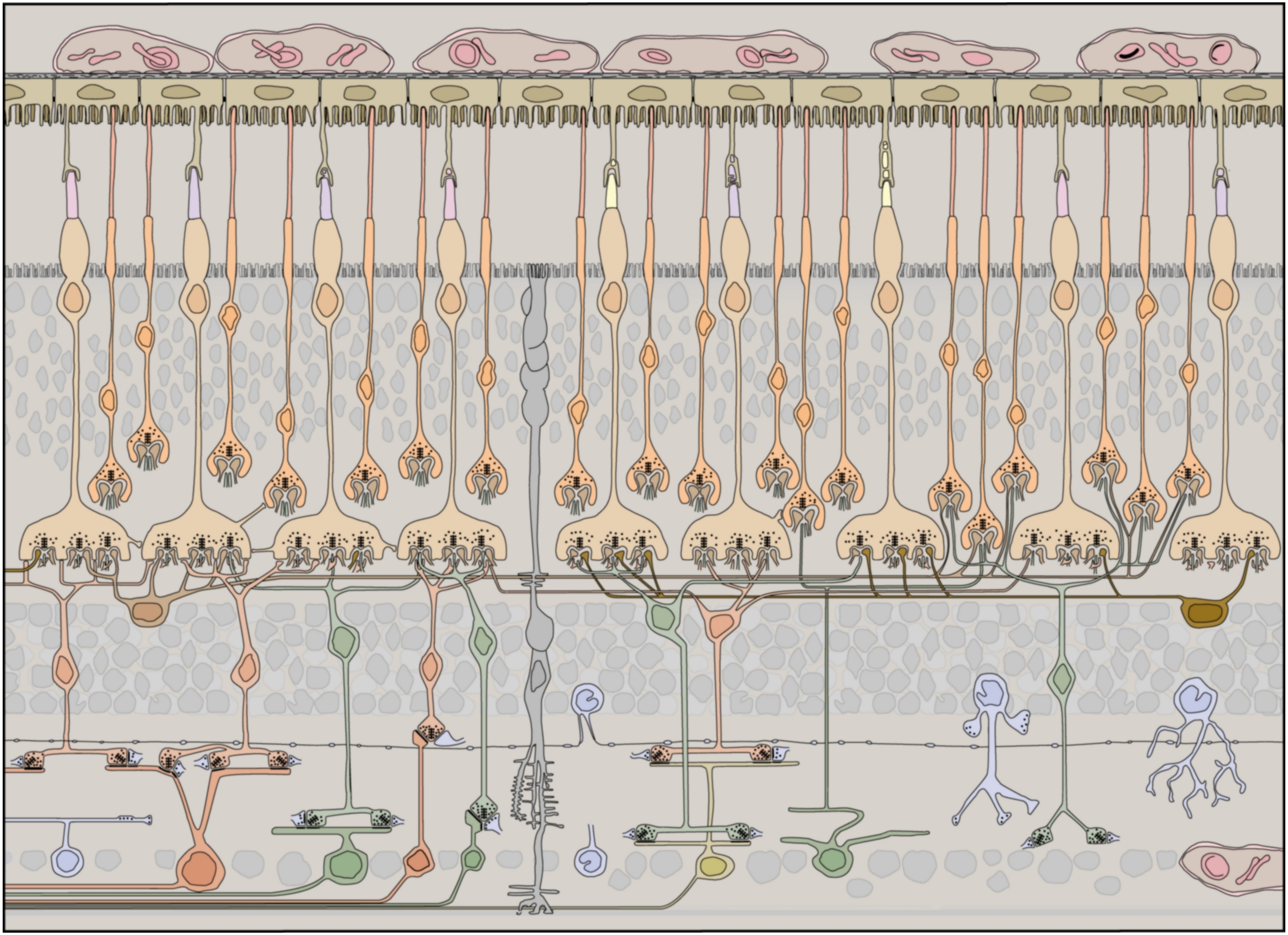
virgulatus

americanus

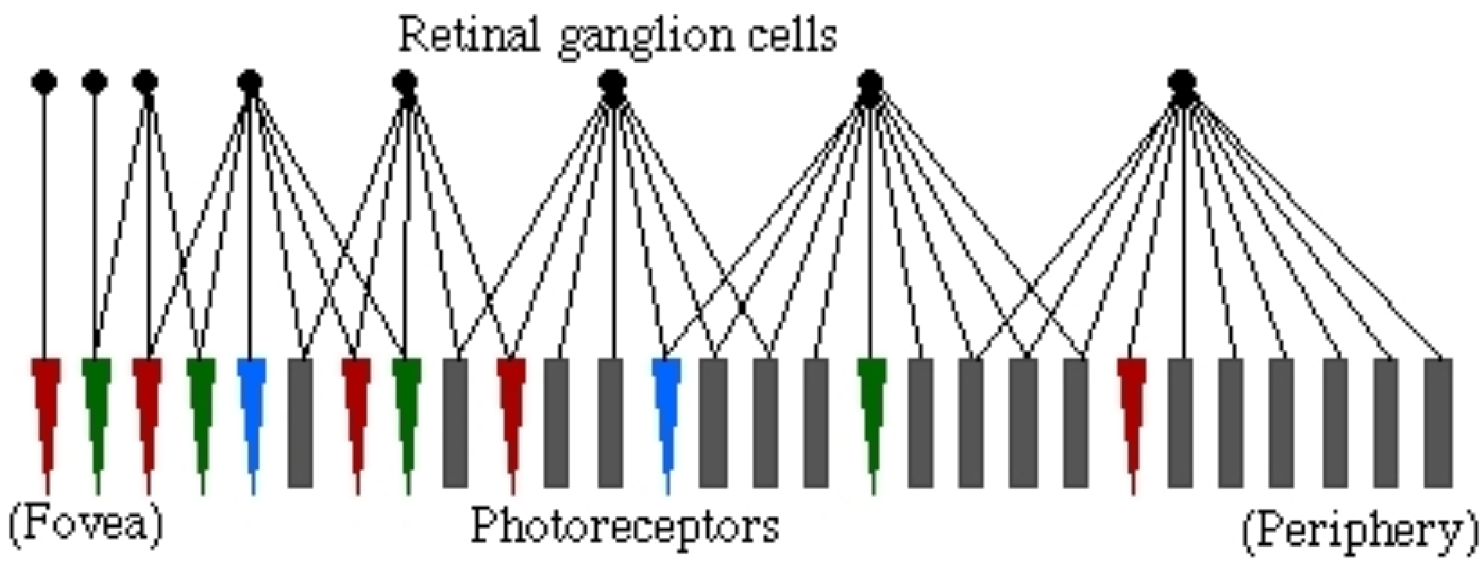
Mantis shrimp

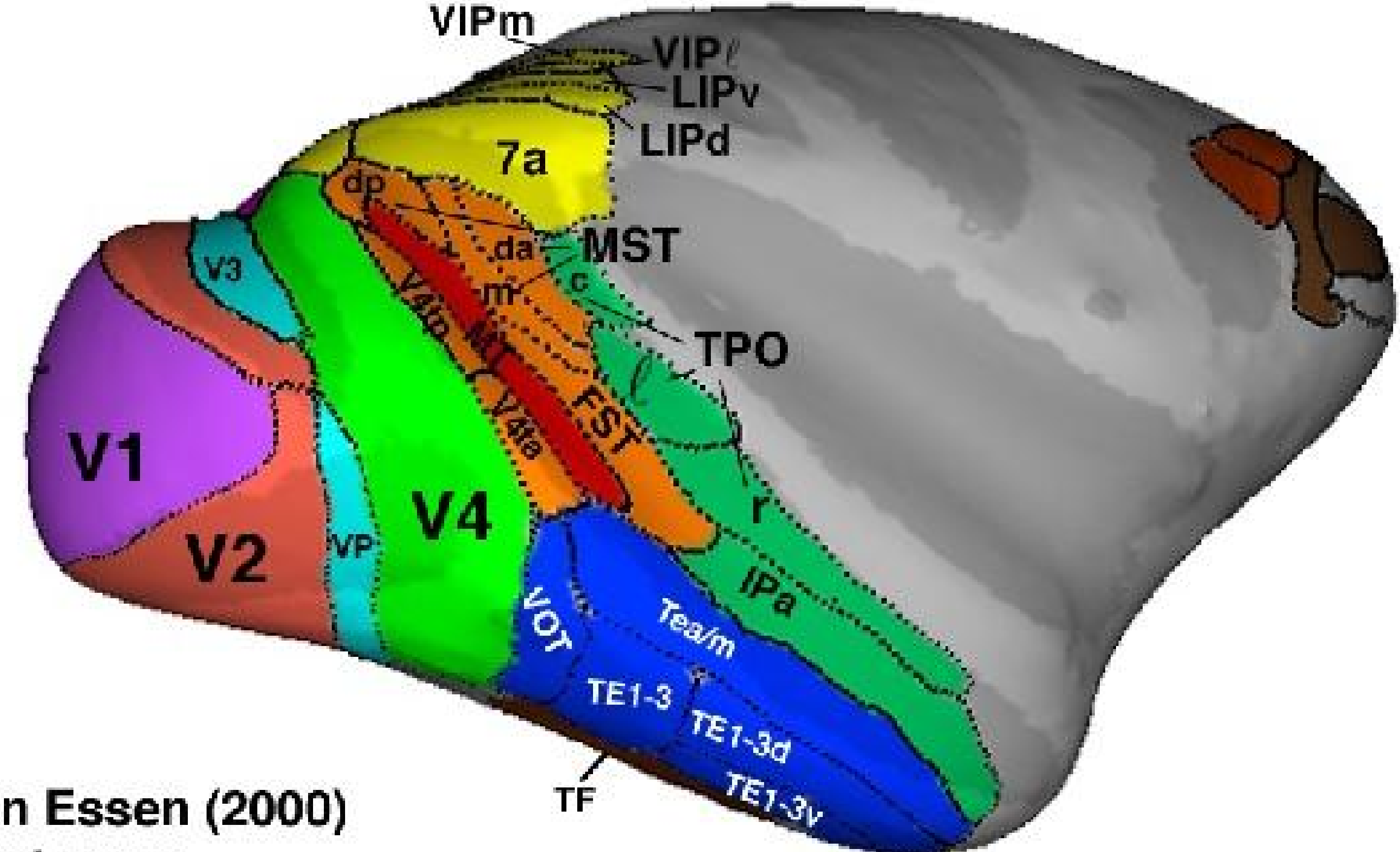
Human retina - cone mosaic





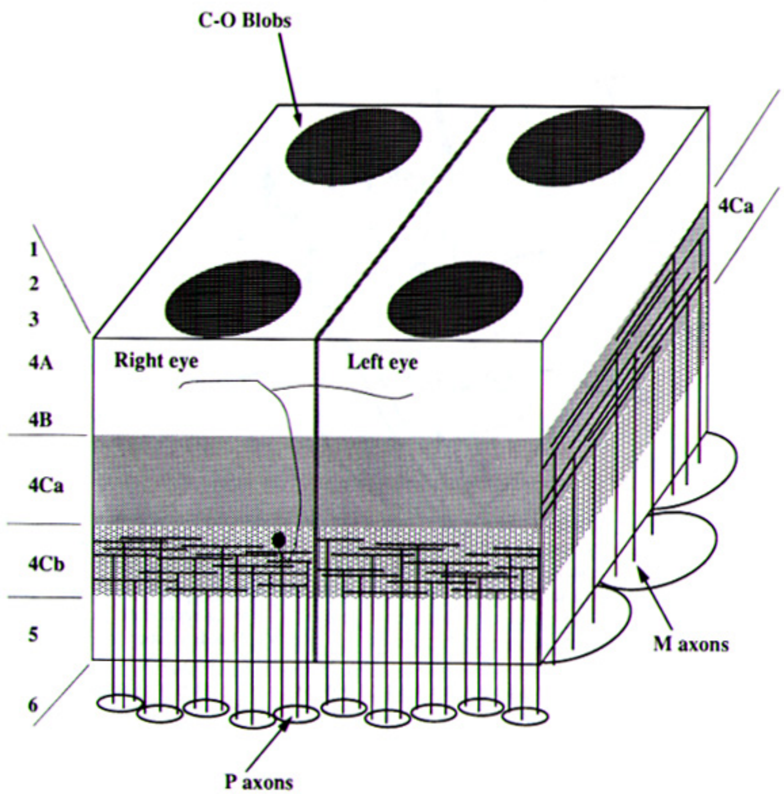
Smoothing and subsampling by retinal ganglion cells

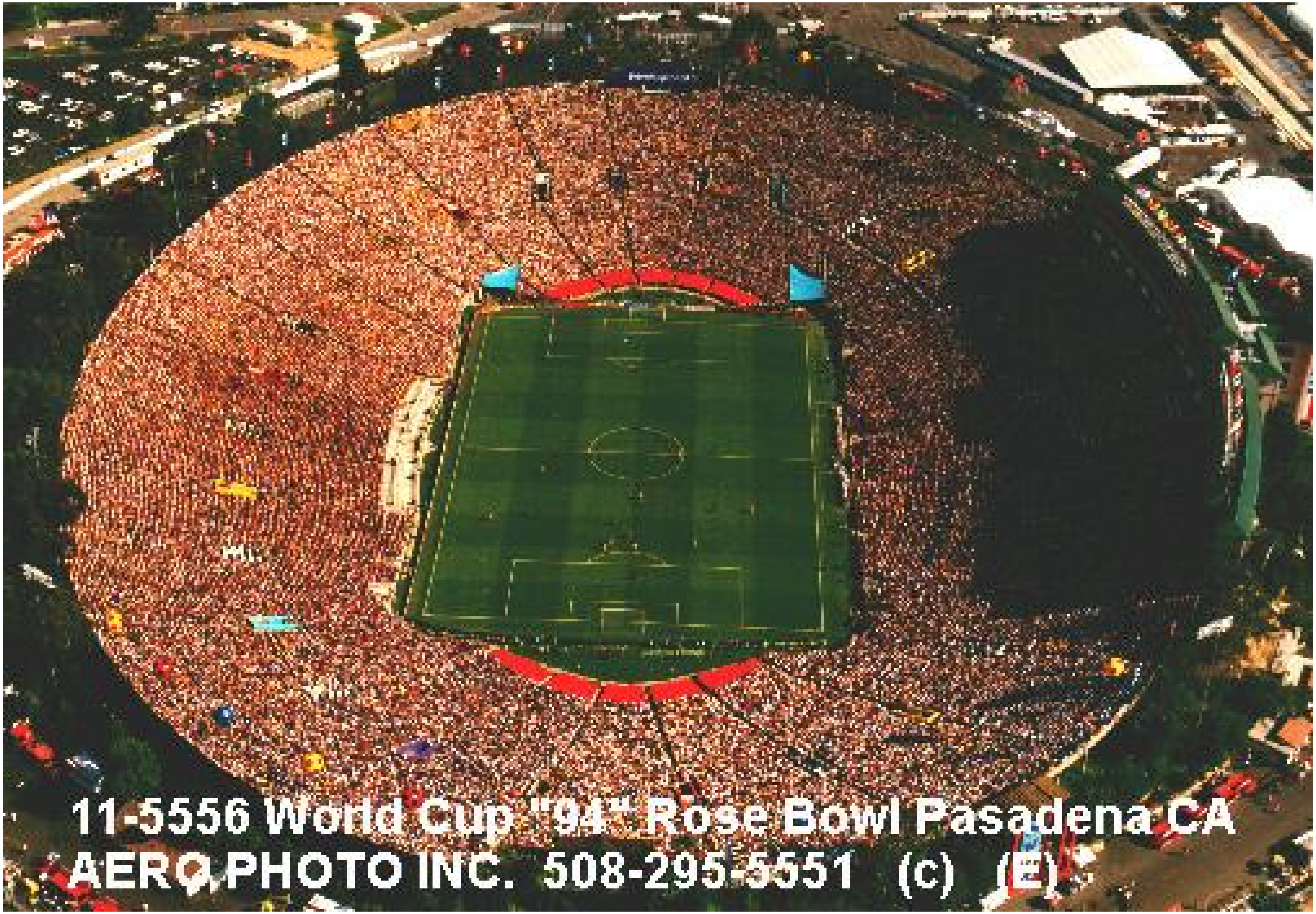




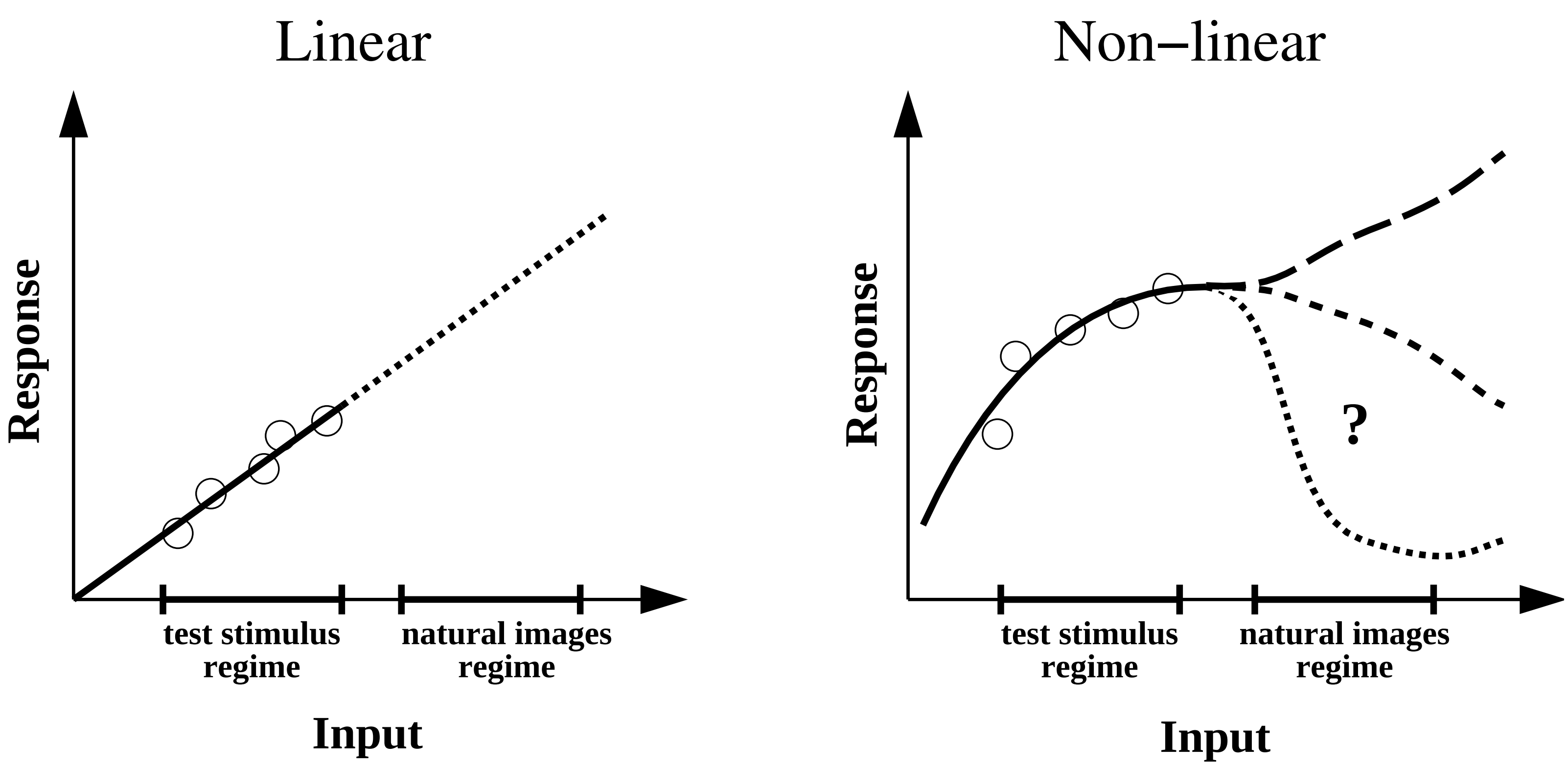
Lewis & Van Essen (2000)
Visual areas

1 mm^2 of cortex analyzes ca. 14 x 14 array of retinal sample nodes and contains 100,000 neurons




11-5556 World Cup "94" Rose Bowl Pasadena CA
AERO PHOTO INC. 508-295-5551 (c) (E)

Why natural scenes?



The efficient coding hypothesis
(Barlow 1961; Attneave 1954)

Nervous systems should exploit the statistical dependencies contained in sensory signals

Redundancy reduction (Barlow 1961)

- Natural images are *redundant* in that there exist statistical dependencies among pixel values in space and time.

- In order to make efficient use of resources, the visual system should *reduce* redundancy by removing statistical dependencies.

- This also makes it easier to store *prior probabilities*, since you can express the joint distribution of events as the product of marginal probabilities: $P(\mathbf{x}) = \Pi_i P(x_i)$.

An example - text

Redundancy Reduction as a Strategy for Unsupervised Learning

A. Norman Redlich

a l i c e w a s b e g i n n i n g t o g e t v e r y t i r e d o f s i t t i n g b y h e r s i s t e r o n
t h e b a n k a n d o f h a v i n g n o t h i n g t o d o o n c e o r t w i c e s h e h a d p e e p
e d i n t o t h e b o o k h e r s i s t e r w a s r e a d i n g b u t i t h a d n o p i c t u r e s o r
c o n v e r s a t i o n s i n i t a n d w h a t i s t h e u s e o f a b o o k t h o u g h t a l i c e w
i t h o u t p i c t u r e s o r c o n v e r s a t i o n s s o s h e w a s c o n s i d e r i n g i n h e
r o w n m i n d a s w e l l a s s h e c o u l d f o r t h e h o t d a y m a d e h e r f e e l v e r y
s l e e p y a n d s t u p i d w h e t h e r t h e p l e a s u r e o f m a k i n g a d a i s y c h a i
n w o u l d b e w o r t h t h e t r o u b l e o f g e t t i n g u p a n d p i c k i n g t h e d a i s i
e s w h e n s u d d e n l y a w h i t e r a b b i t w i t h p i n k e y e s r a n c l o s e b y h e r
t h e r e w a s n o t h i n g s o v e r y r e m a r k a b l e i n t h a t n o r d i d a l i c e t h i
n k i t s o v e r y m u c h o u t o f t h e w a y t o h e a r t h e r a b b i t s a y t o i t s e l f o
h d e a r o h d e a r

An example - text

alice was b e g in n ing t o g e t very t i r e d o f s i t t ing b y her s i s t e r o n the b a n
k and o f h a v ing n o th ing t o d o o n c e o r t w i c e she h a d p e e p e d in t o the b o
o k her s i s t e r was r e a d ing but i th a d n o p i c t u r e s o r c o n ver s a t i o n s in i
t and w h a t i s the u s e o f a b o o k th ough t alice with ou t p i c t u r e s o r c o n ver
s a t i o n s s o she was c o n s i d e r ing in her o w n m in d a s w e l l a s she c ould for
the h o t d a y m a d e her f e e l very s l e e p y and s t u p i d w he the r the p l e a s u r
e o f m a k ing a d a i s y c h a in w ould b e w o r th the t r ou b l e o f g e t t ing u p
and p i c k ing the d a i s i e s w he n s u d d e n l y a w h i t e r a b b i t with p in k e y e
s r a n c l o s e b y her the r e was n o th ing s o very r e m a r k a b l e in that n o r d i
dalic e th in k i t s o very m u c h ou t o f the w a y t o he a r the r a b b i t s a y t o i t s e
l f o h d e a r o h d e a r

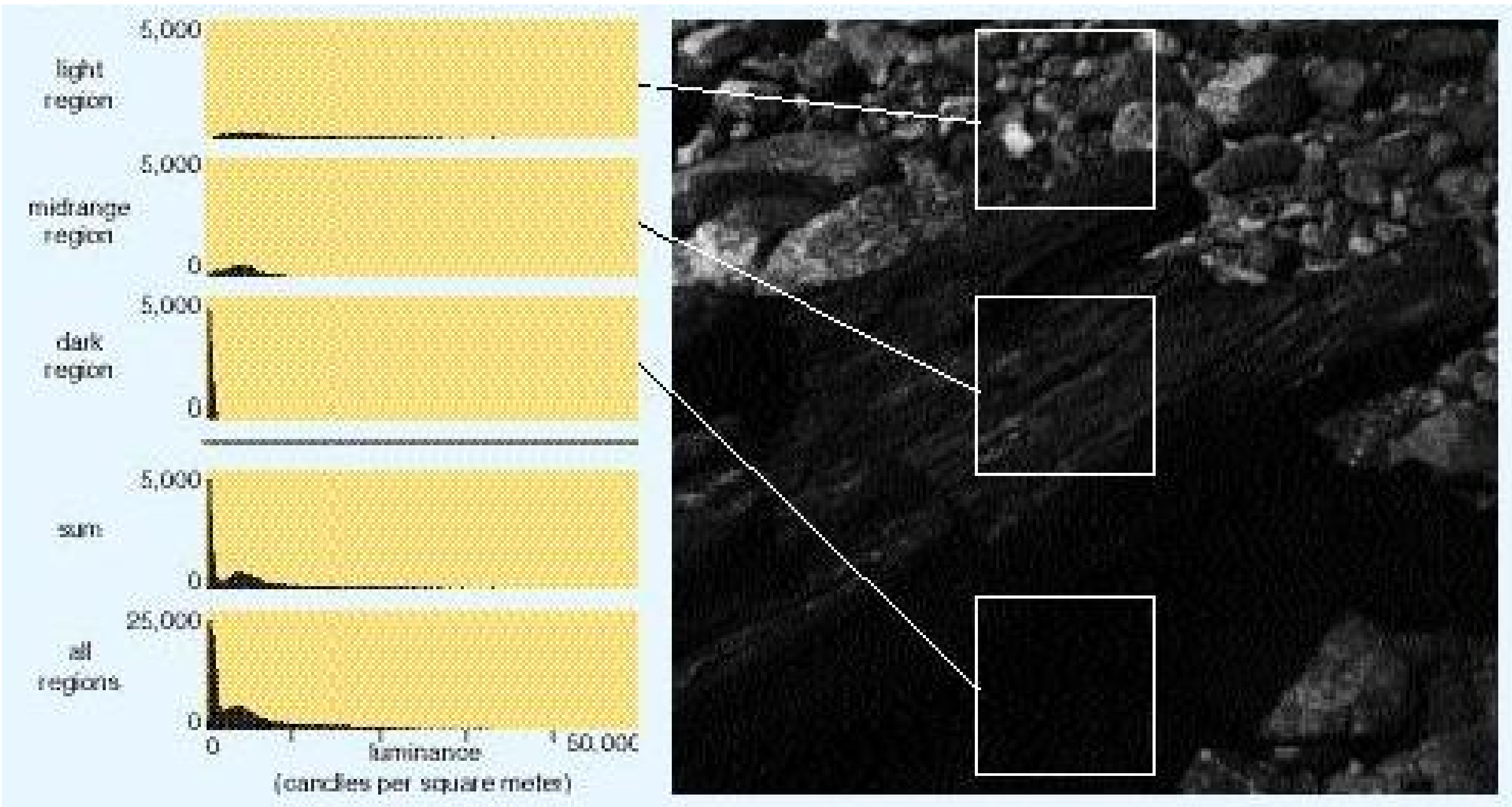
alice was beginning toget verytiredof sitting by hersister ontheban k and of having nothingtodo onceortwice shehad peeped intothe book hersister was read ing but ithad no pictures or conversation s in it and what is theuseof a book thoughtalice without pictures or conversation s so shewas consider ing in her ow n mind aswell asshecould for the hot day made her feel very sleepy and stupid whether the plea sure of making ad a is y ch a in wouldbe wor th the trouble of gettingupand p i c king the d a is ies when suddenly a whiterabbit with p in k eyes r an close by her therewas nothing so very remark able in that n or did alice think it so very much outoftheway to hear therabbit saytoitself ohdear ohdear

Natural image statistics and efficient coding

- First-order statistics
  - Intensity/contrast histograms
  - Histogram equalization
- Second-order statistics
  - Autocorrelation function $\rightarrow 1/f^2$ power spectrum
  - Decorrelation/whitening
- Higher-order statistics
  - Orientation, phase spectrum
  - Projection pursuit/sparse coding

First-order statistics (pixel histograms)

First-order statistics (pixel histograms)



Contrast: reduces dynamic range

$$C = \frac{I - \langle I \rangle}{\langle I \rangle}$$

Histogram equalization - fly LMC (Laughlin 1981)

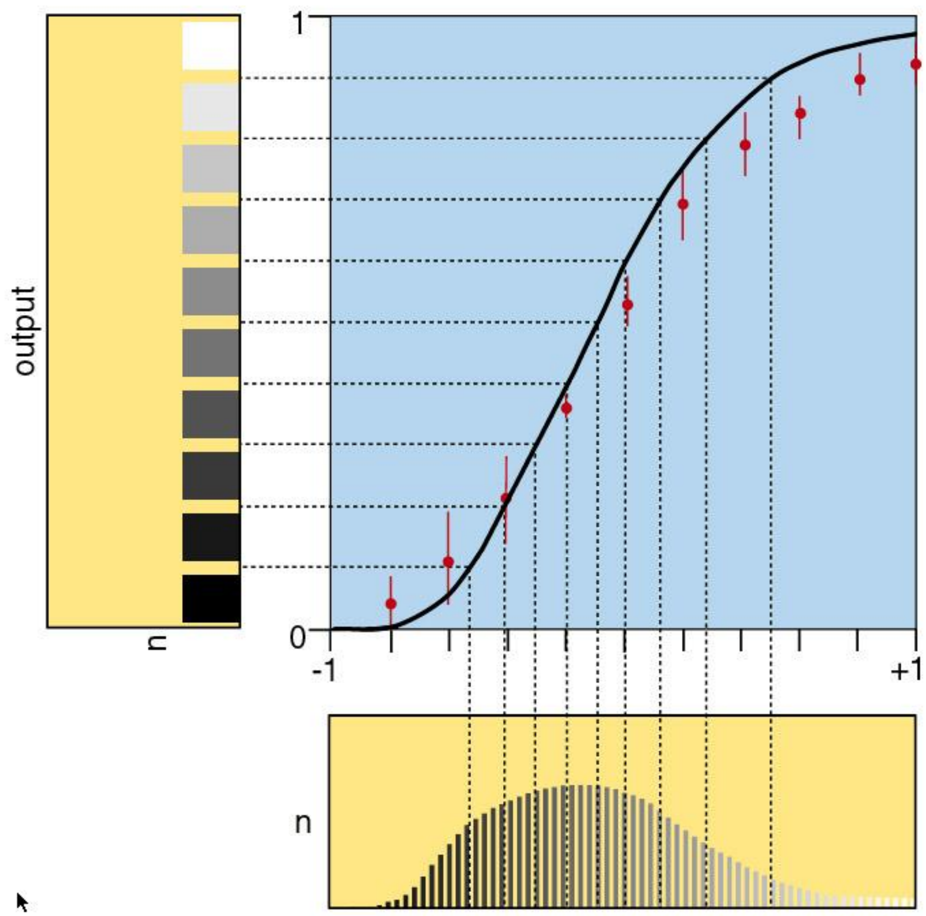
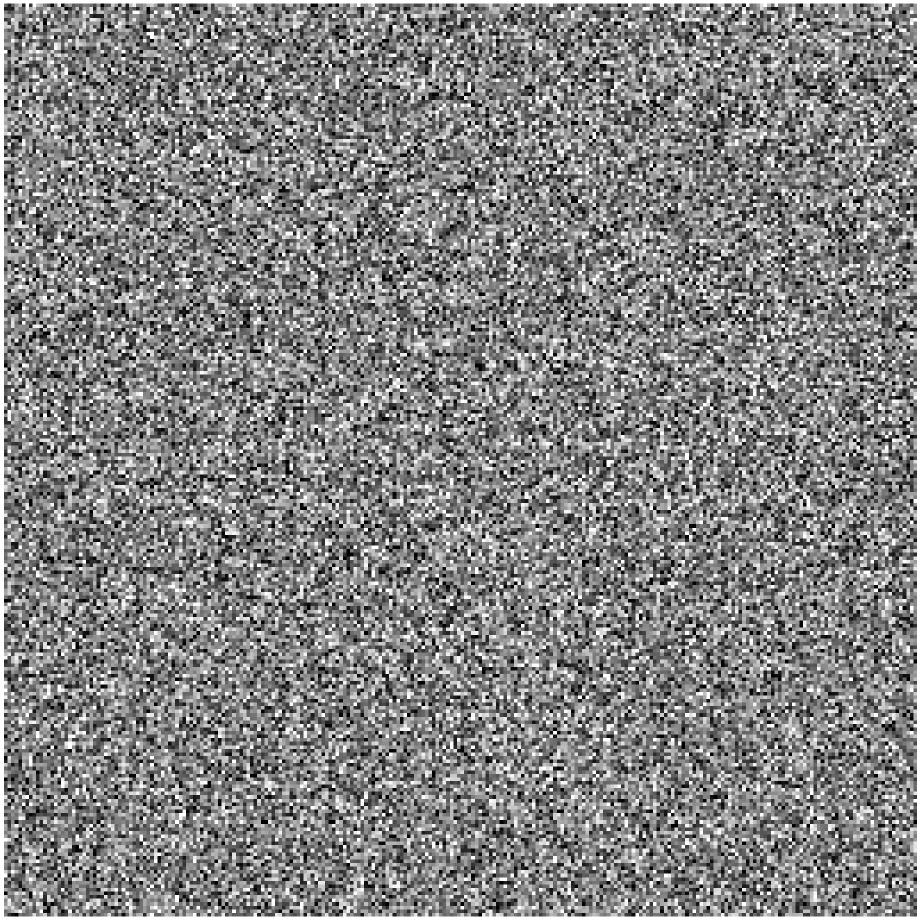
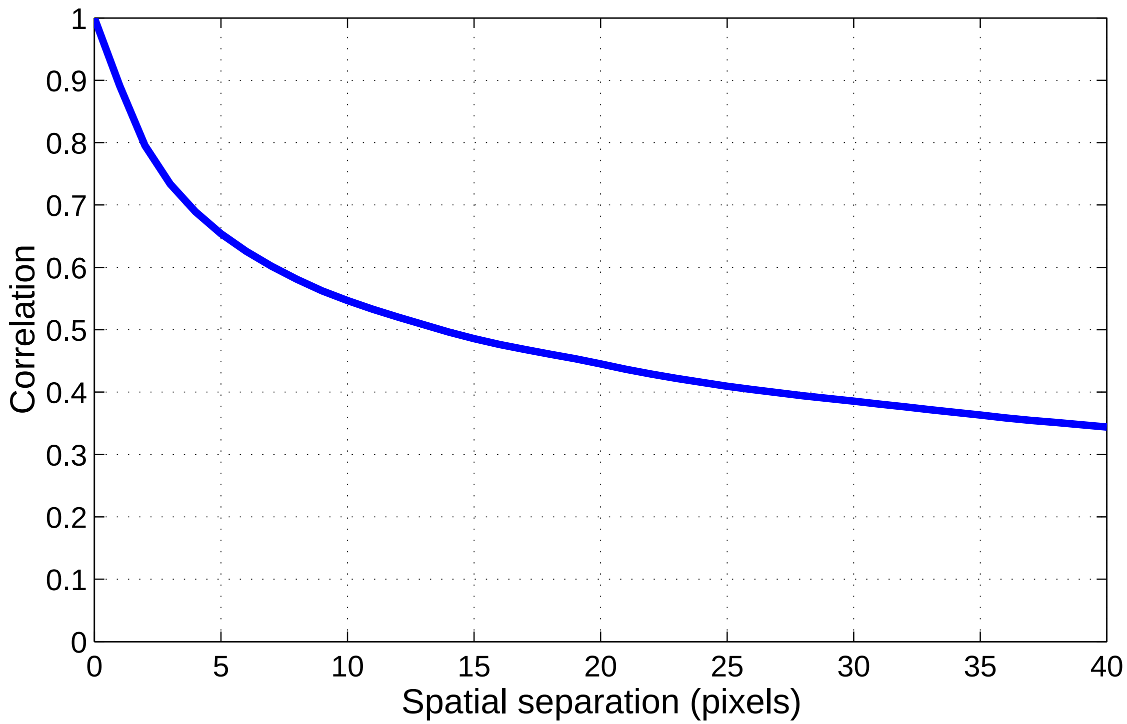


Image synthesis - first-order statistics

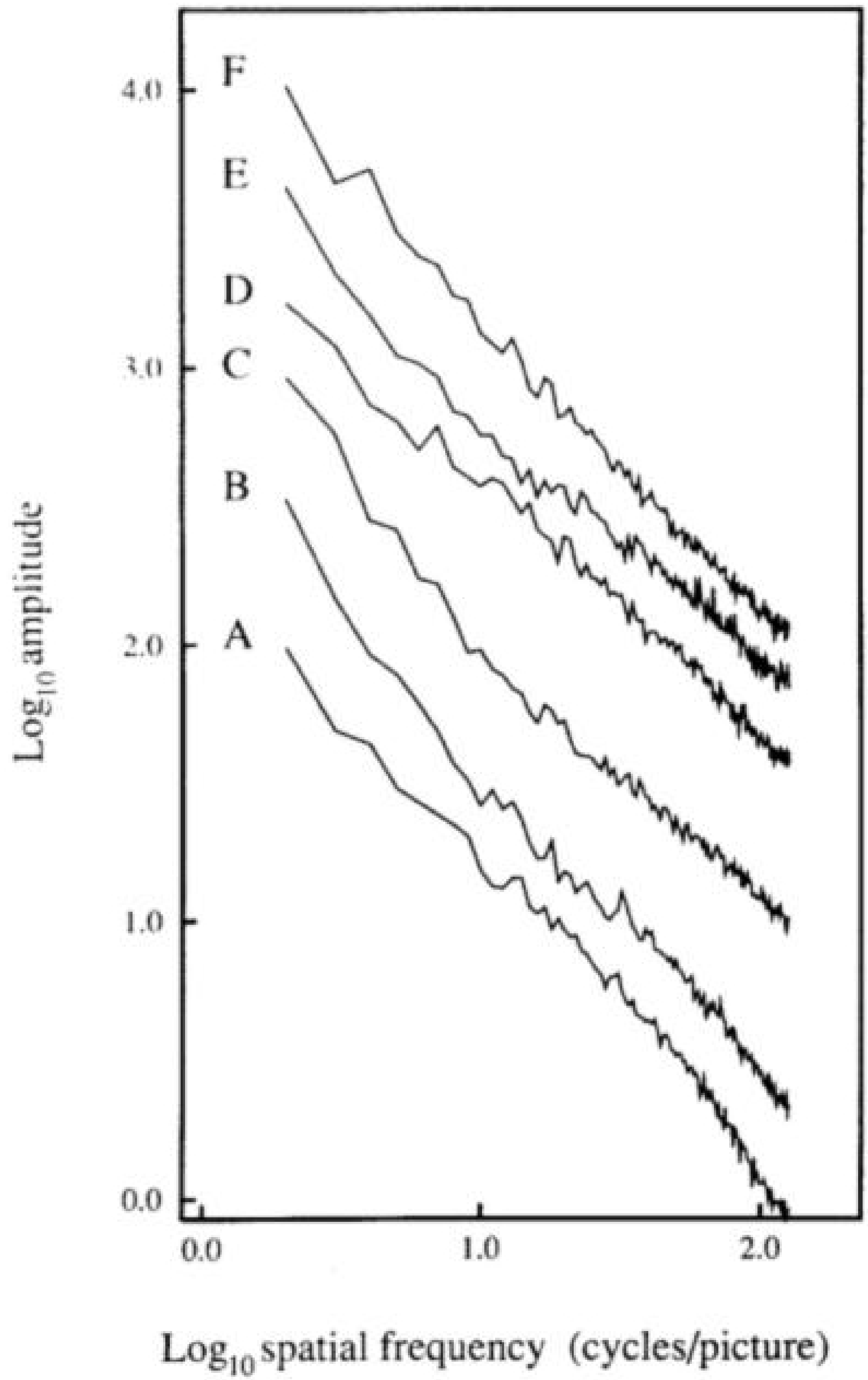


Second-order statistics (auto-correlation function)

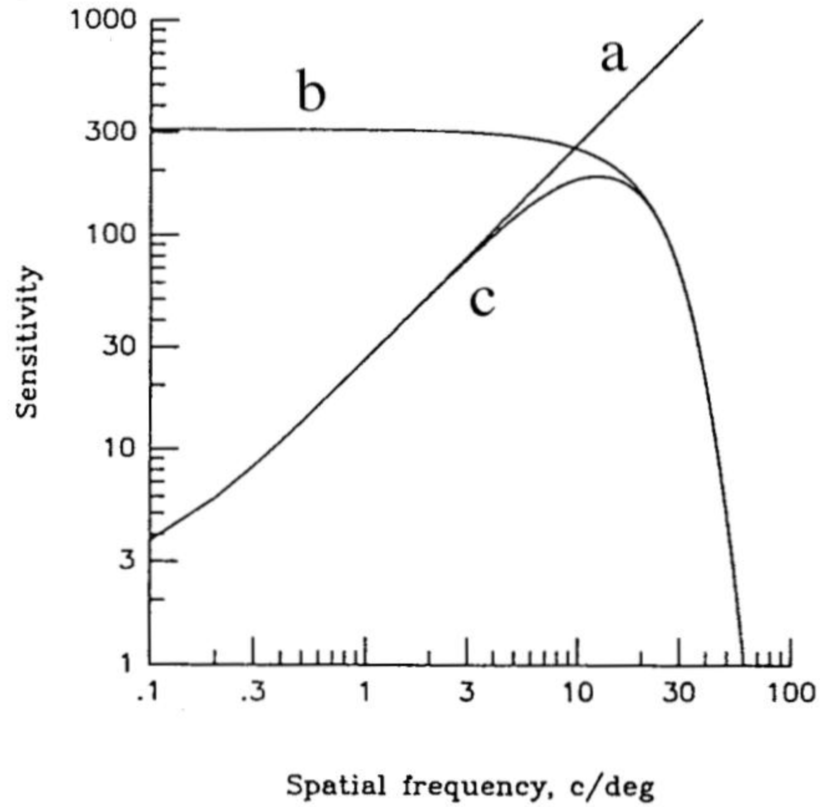
$$C(\Delta x) = \langle I(x)\, I(x + \Delta x) \rangle_x$$



Power spectrum of natural images (Field, 1987)



Whitening (Atick & Redlich, 1992)



Whitening removes second-order correlations

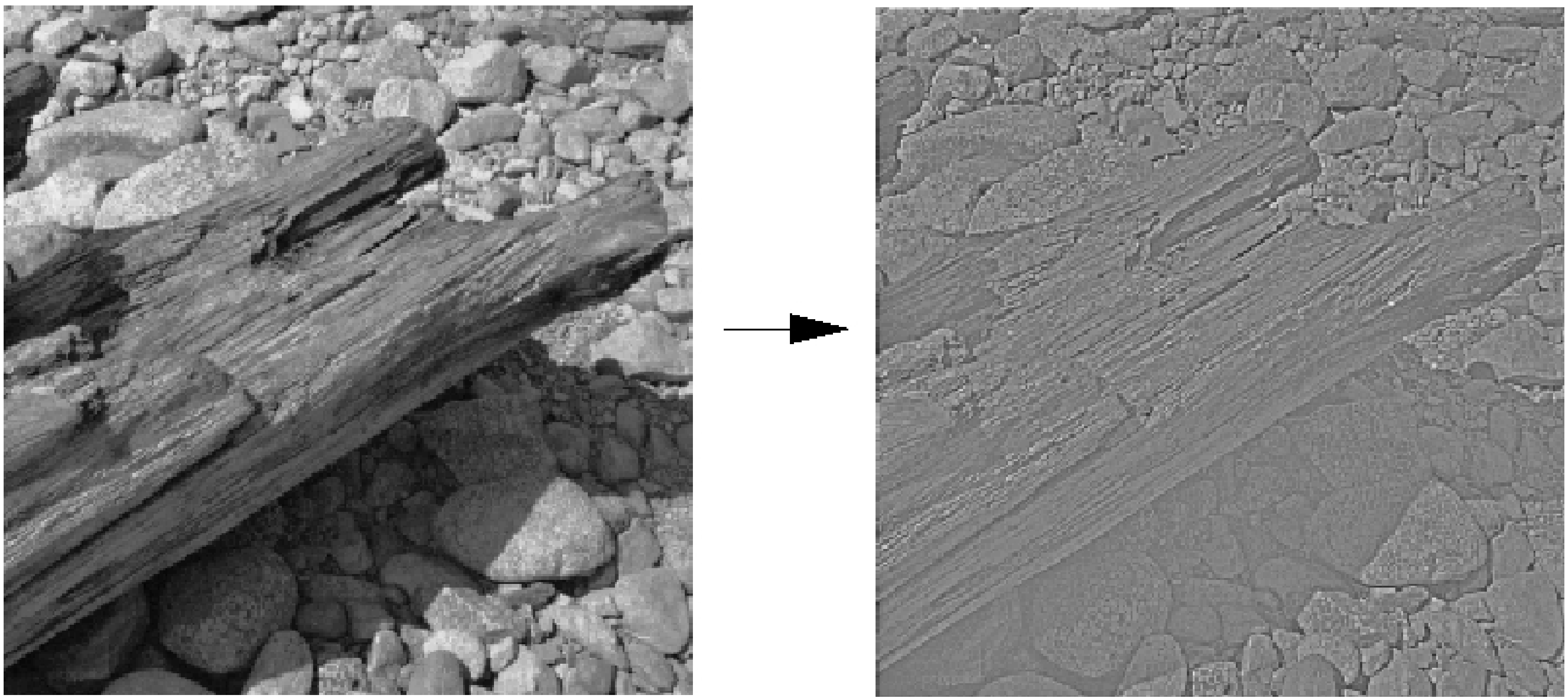
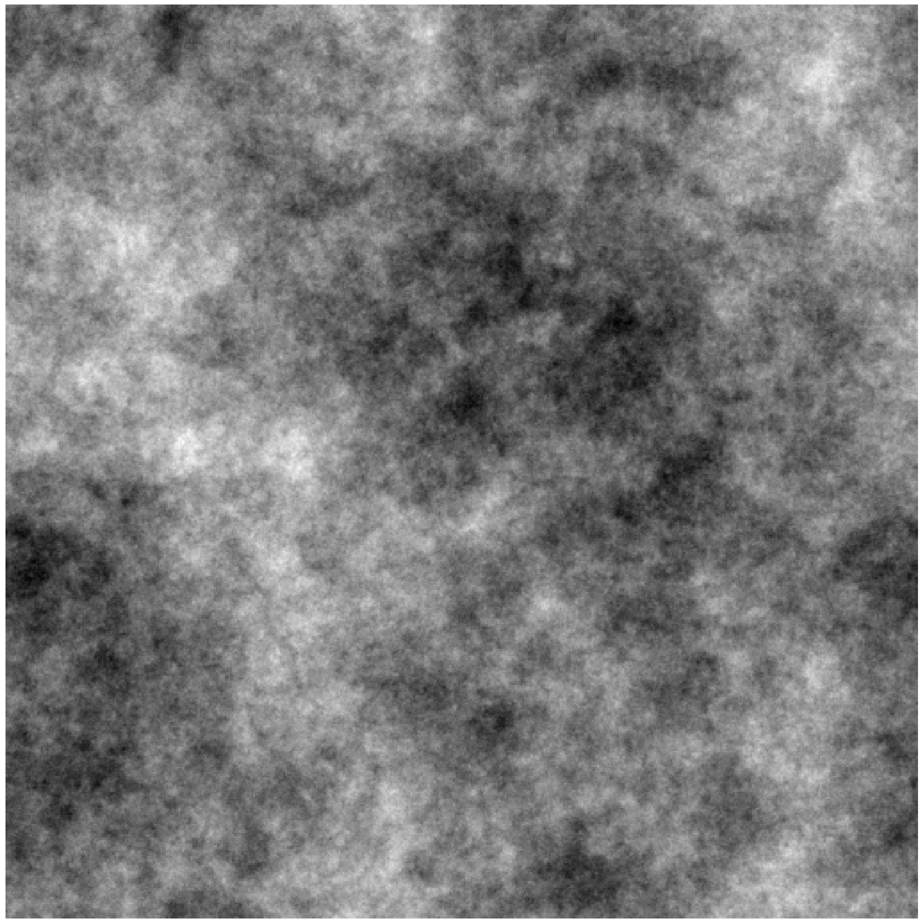
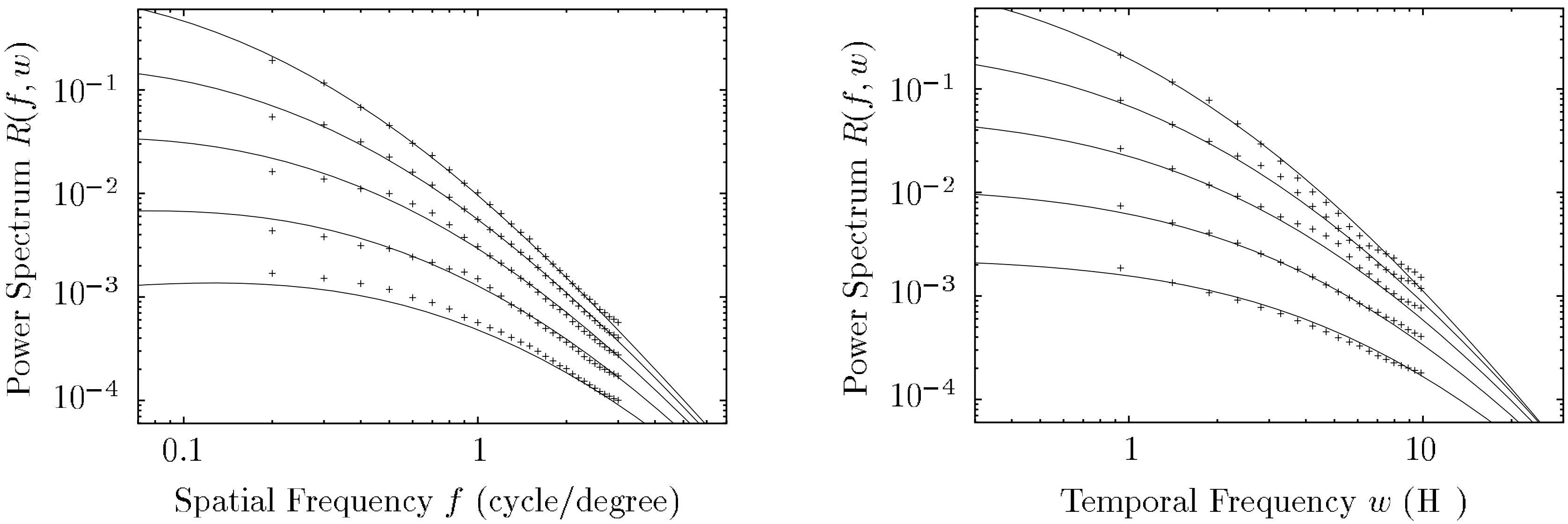


Image synthesis - second-order statistics



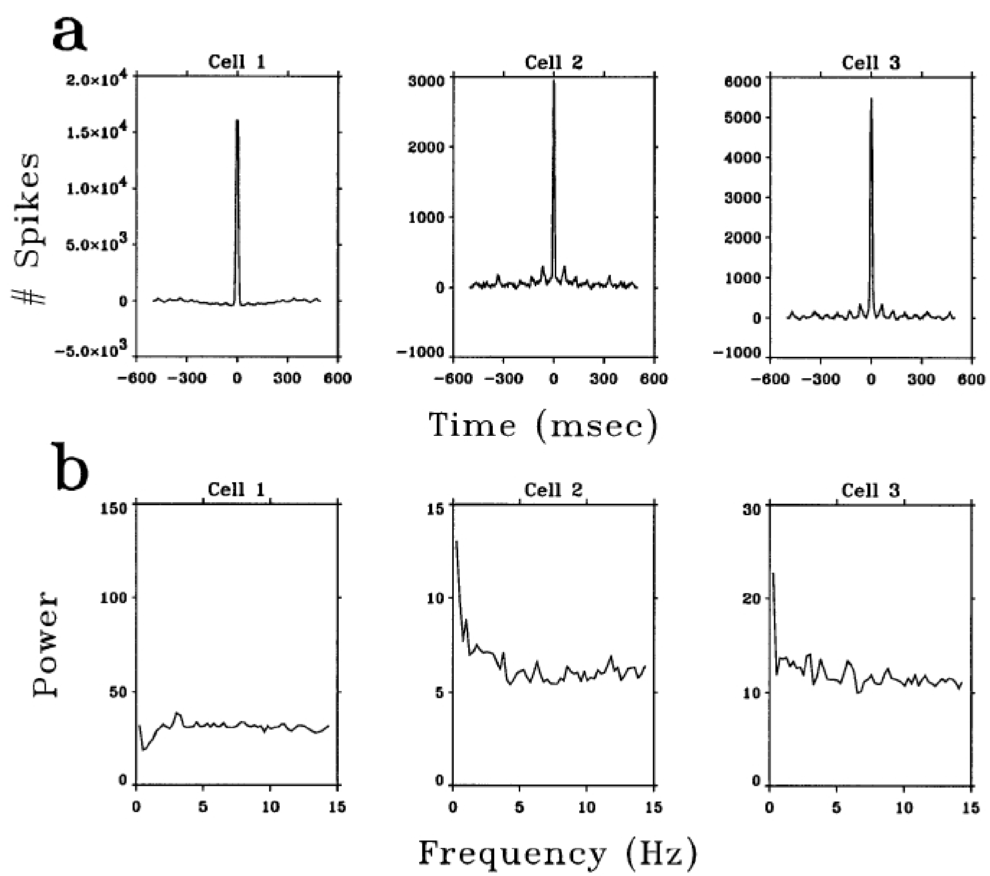
Spatiotemporal power spectrum of natural scenes

- Characterizes pairwise correlations across space and time.
- "$1/f^2$" but non-separable (Dong & Atick, 1995)

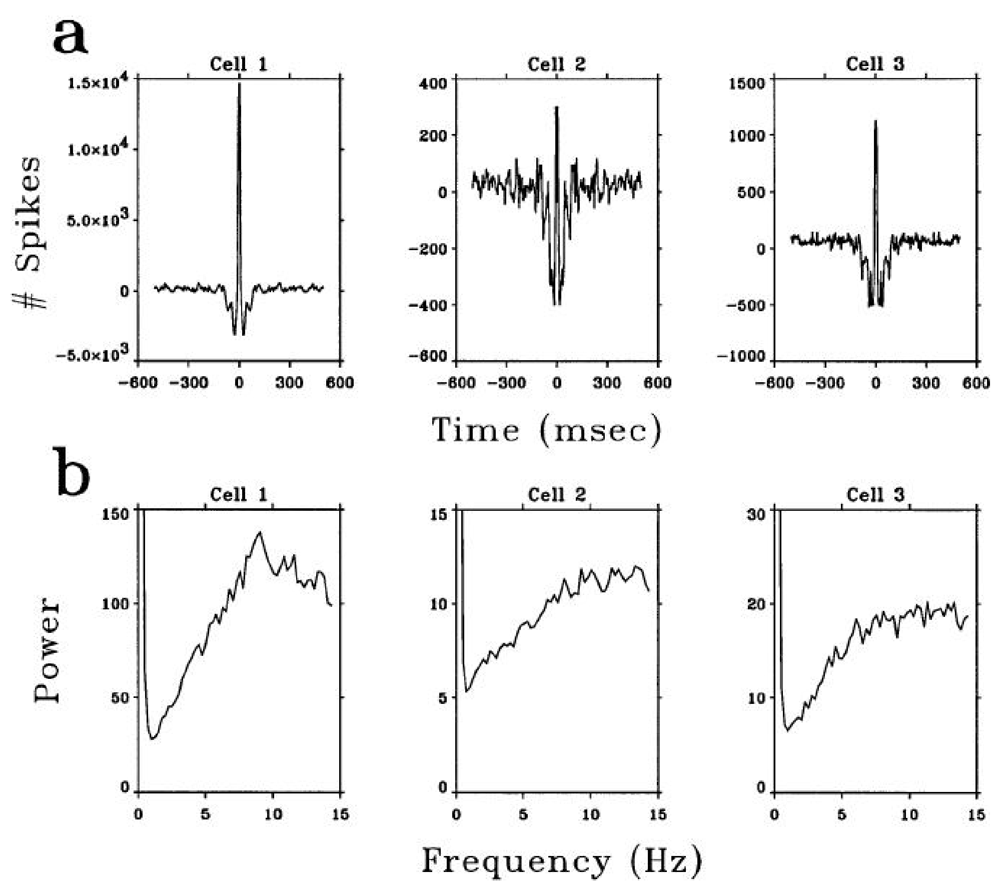


LGN neurons whiten time-varying natural images

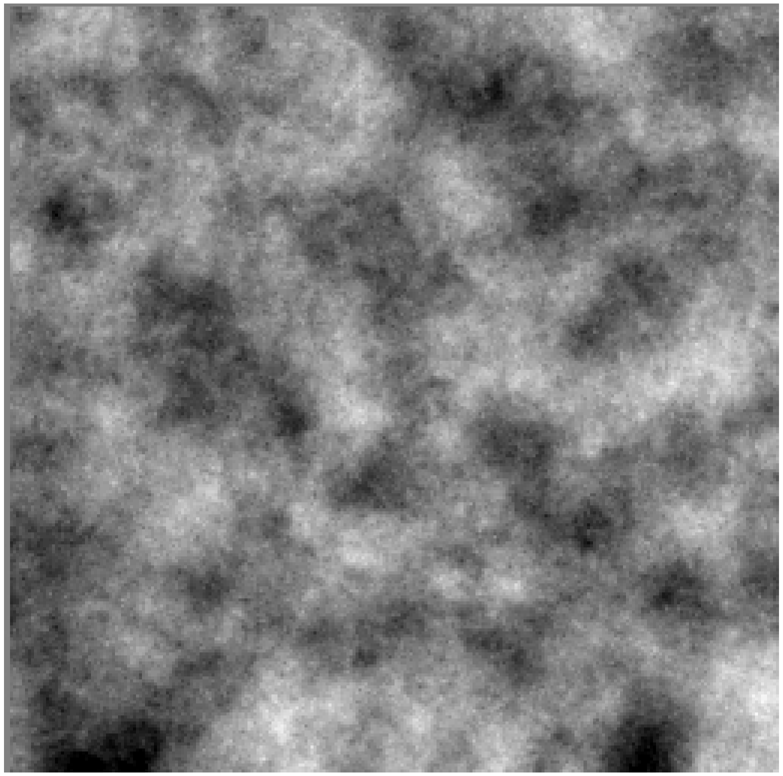
Dan et al, 1996



… but not white noise

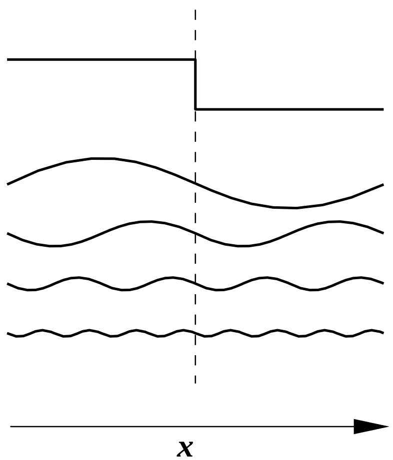


Movie synthesis - second-order, s-t statistics)

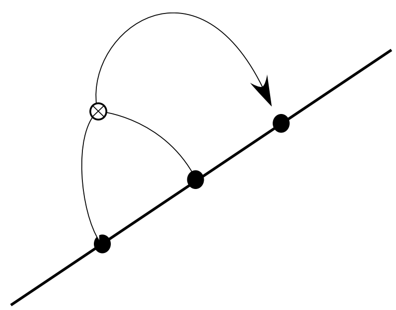


Higher-order statistics

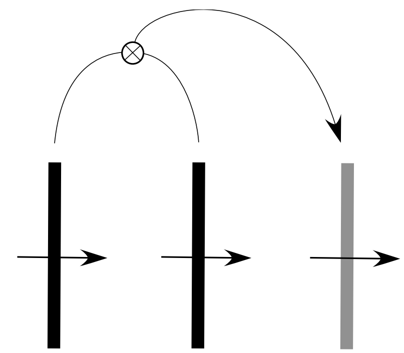
Phase alignment



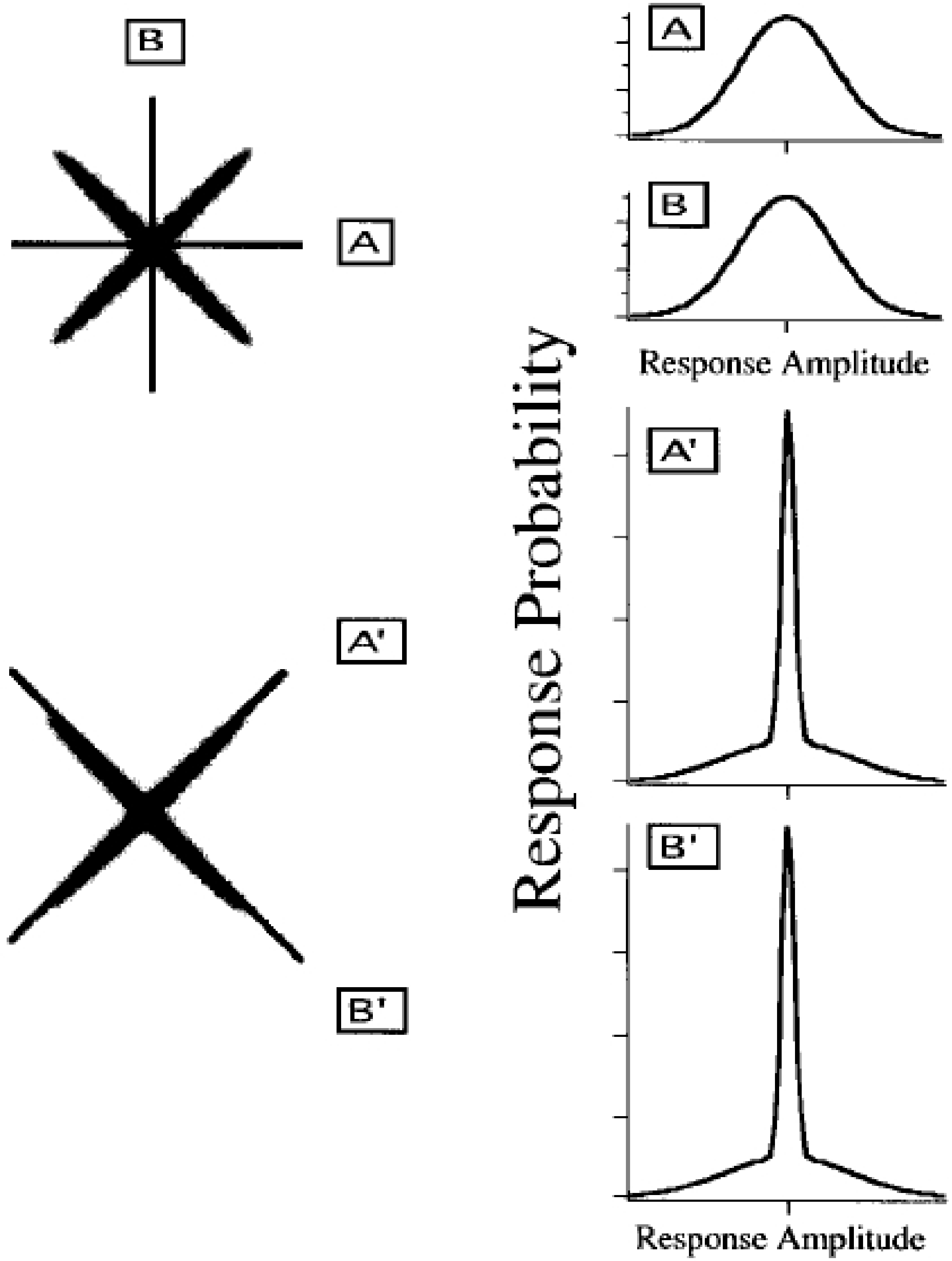
Oriented structure



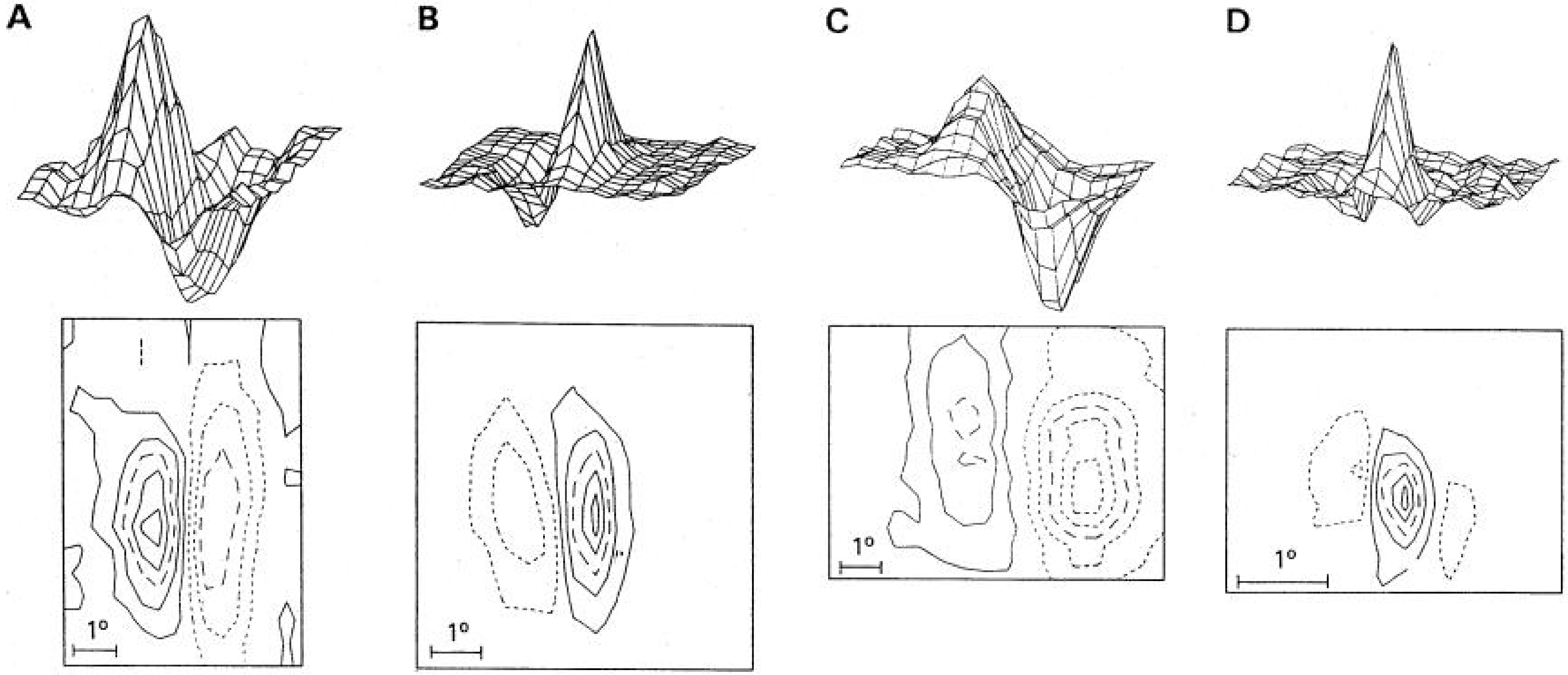
Motion



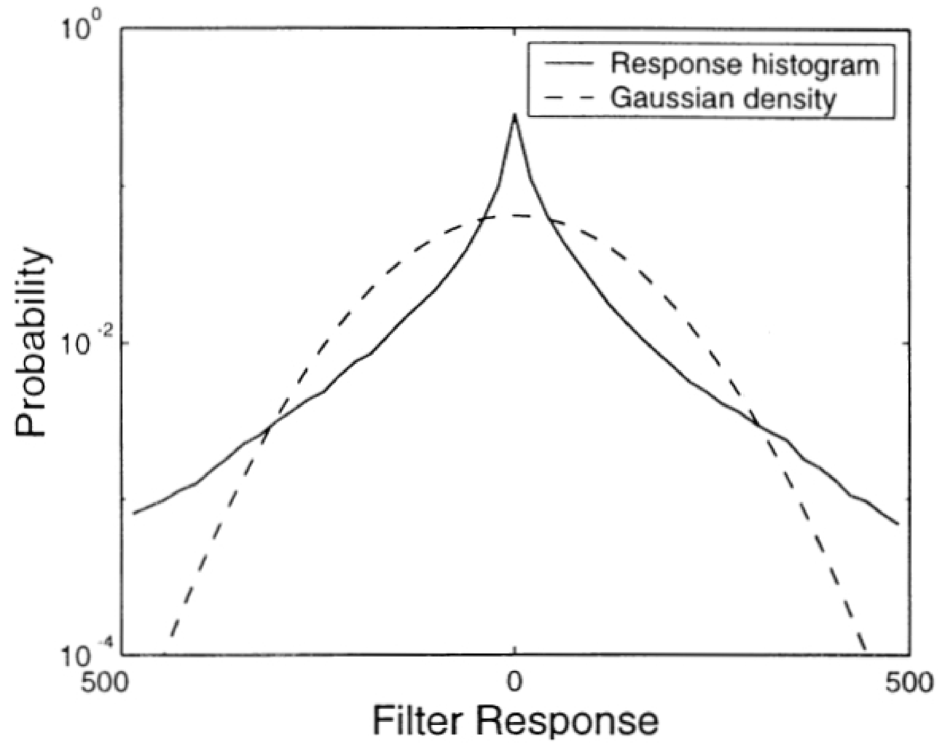
Projection Pursuit
(from Field 1994)



Simple cell receptive fields (Jones & Palmer, 1987)

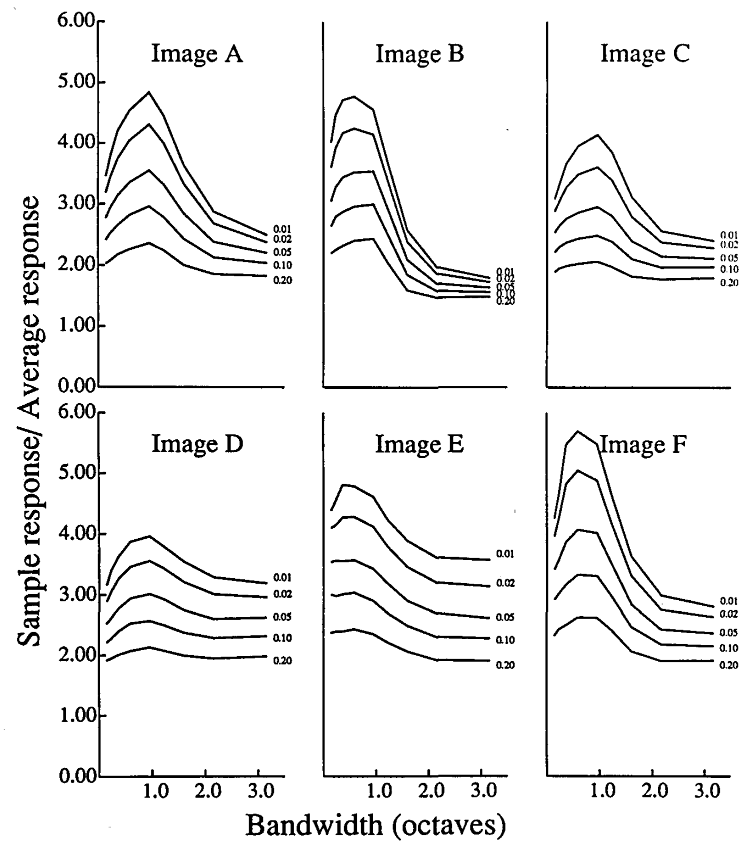


Gabor-filter histogram



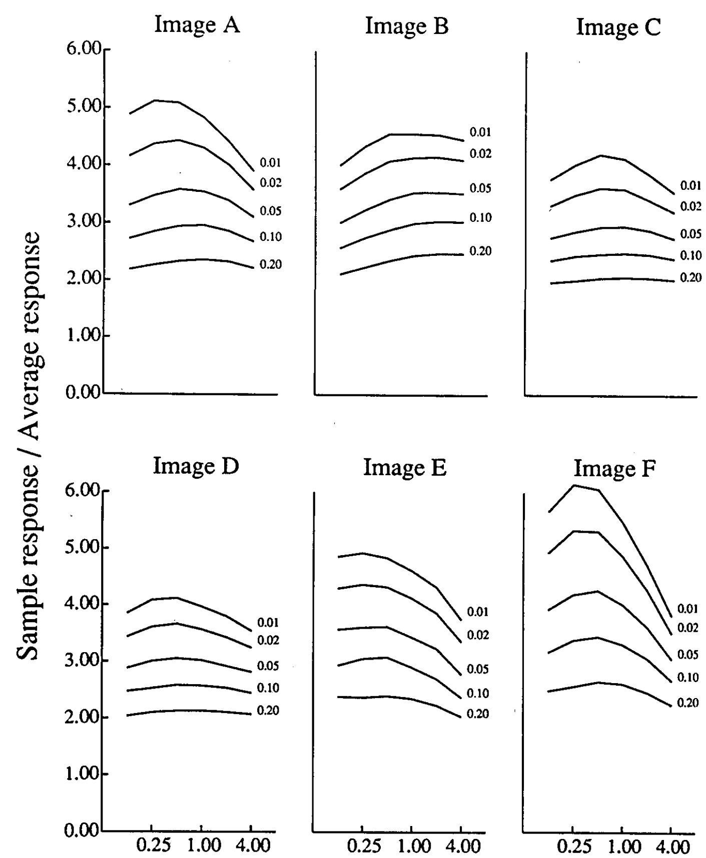
Optimal spatial-frequency bandwidth

Field (1987)



Optimal orientation bandwidth

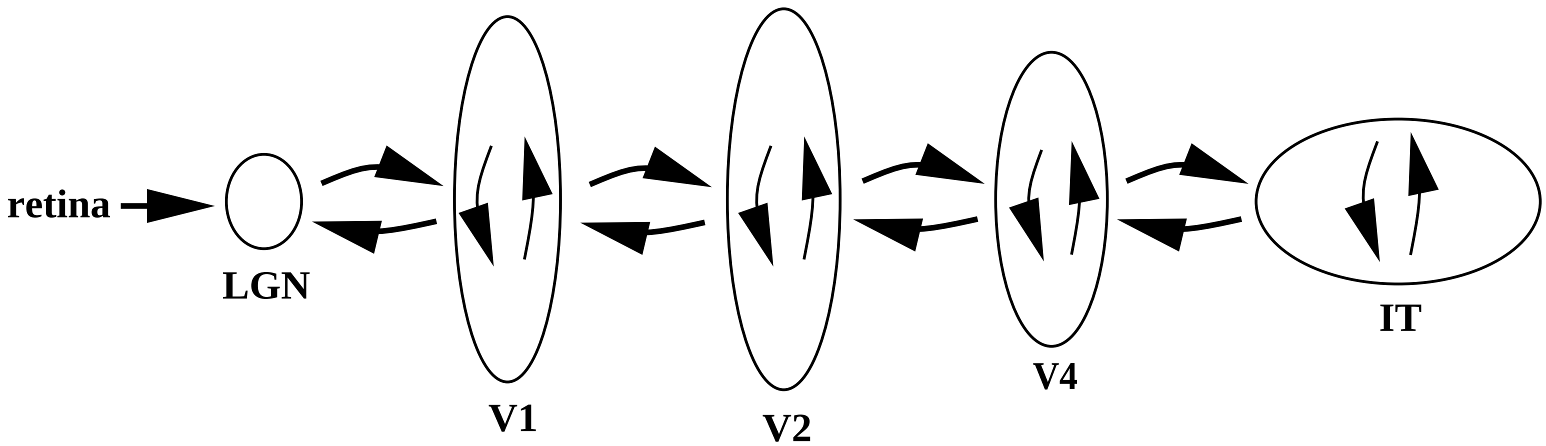
Field (1987)



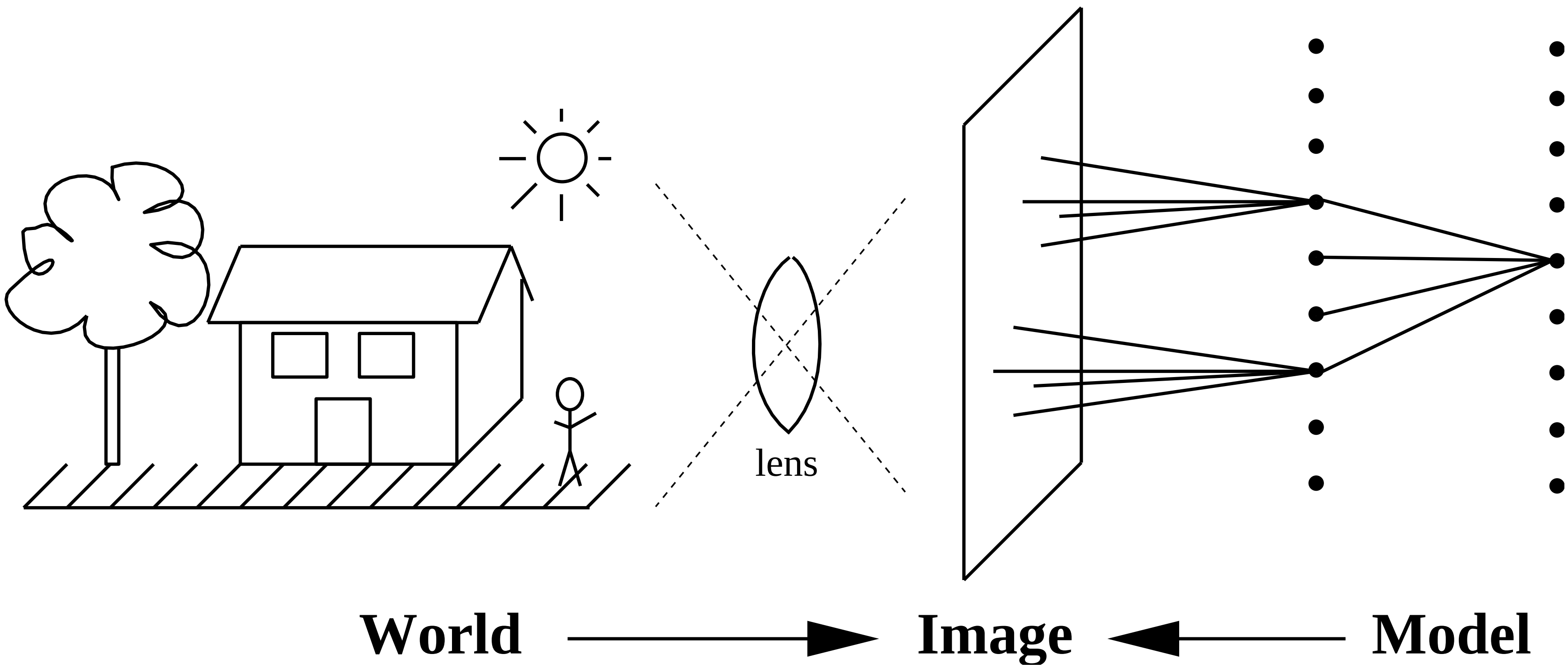
Vision as inference

- 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



Vision as inference



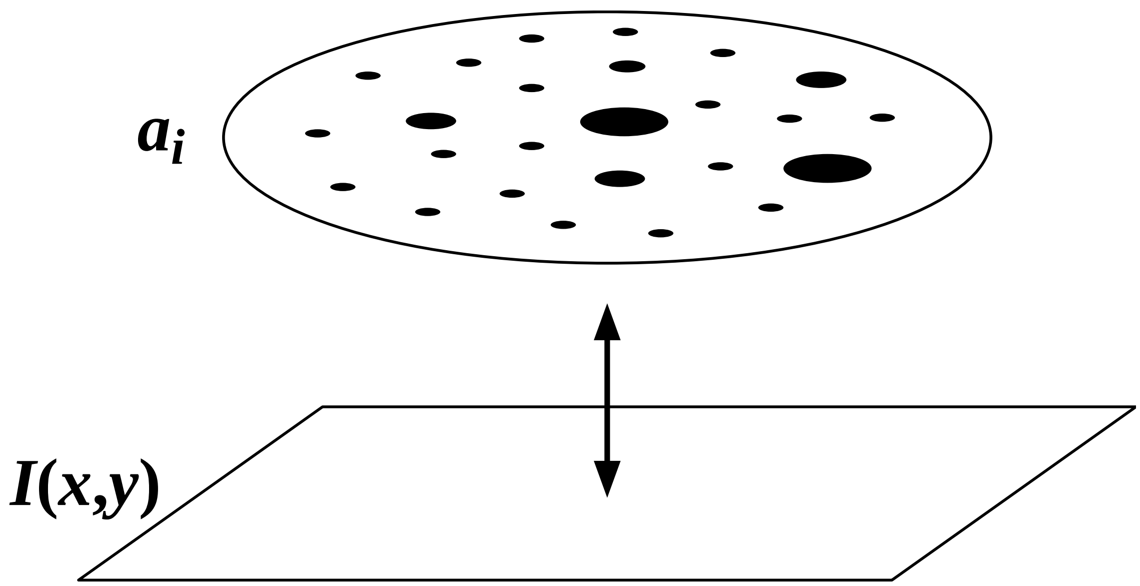
Bayes' rule

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

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

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 - inferotemporal 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).

Image model

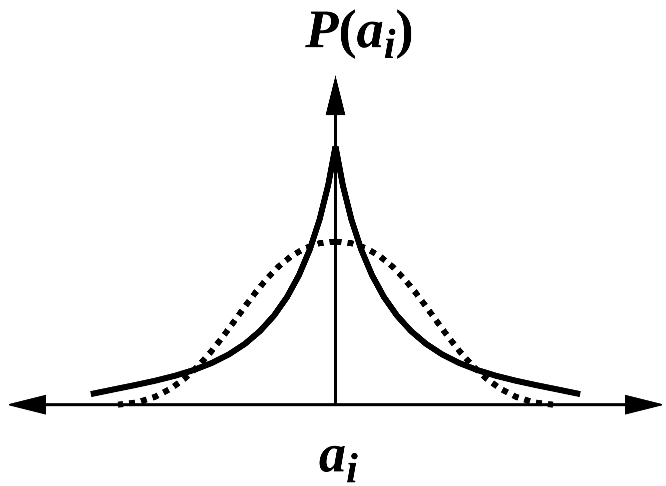
$$I(x,y) = \sum_i a_i \, \phi_i(x,y) + \nu(x,y) \ .$$

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

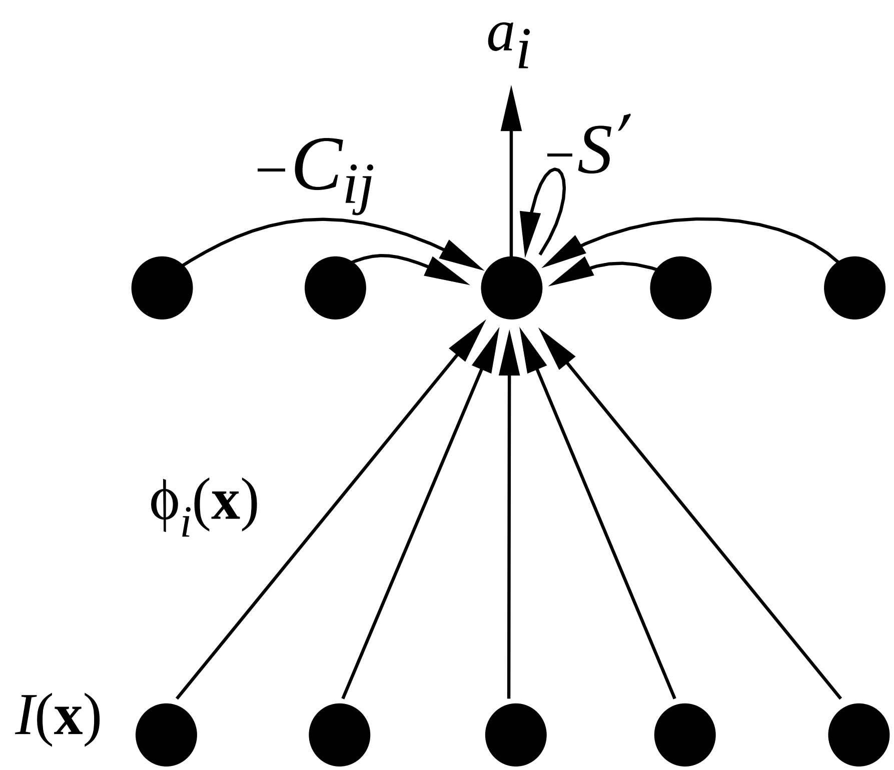
Objective function:

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

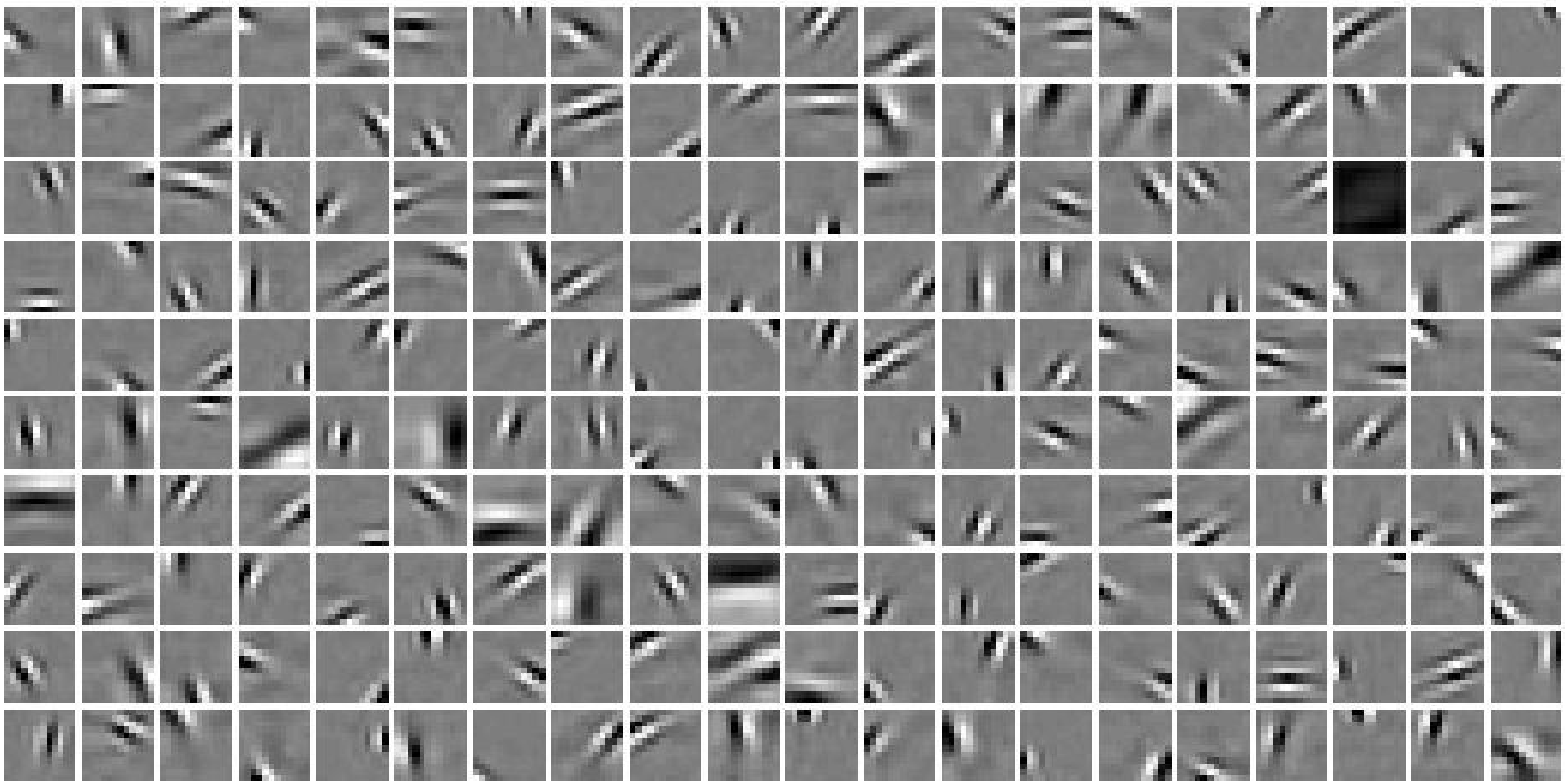
Learning rule:

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

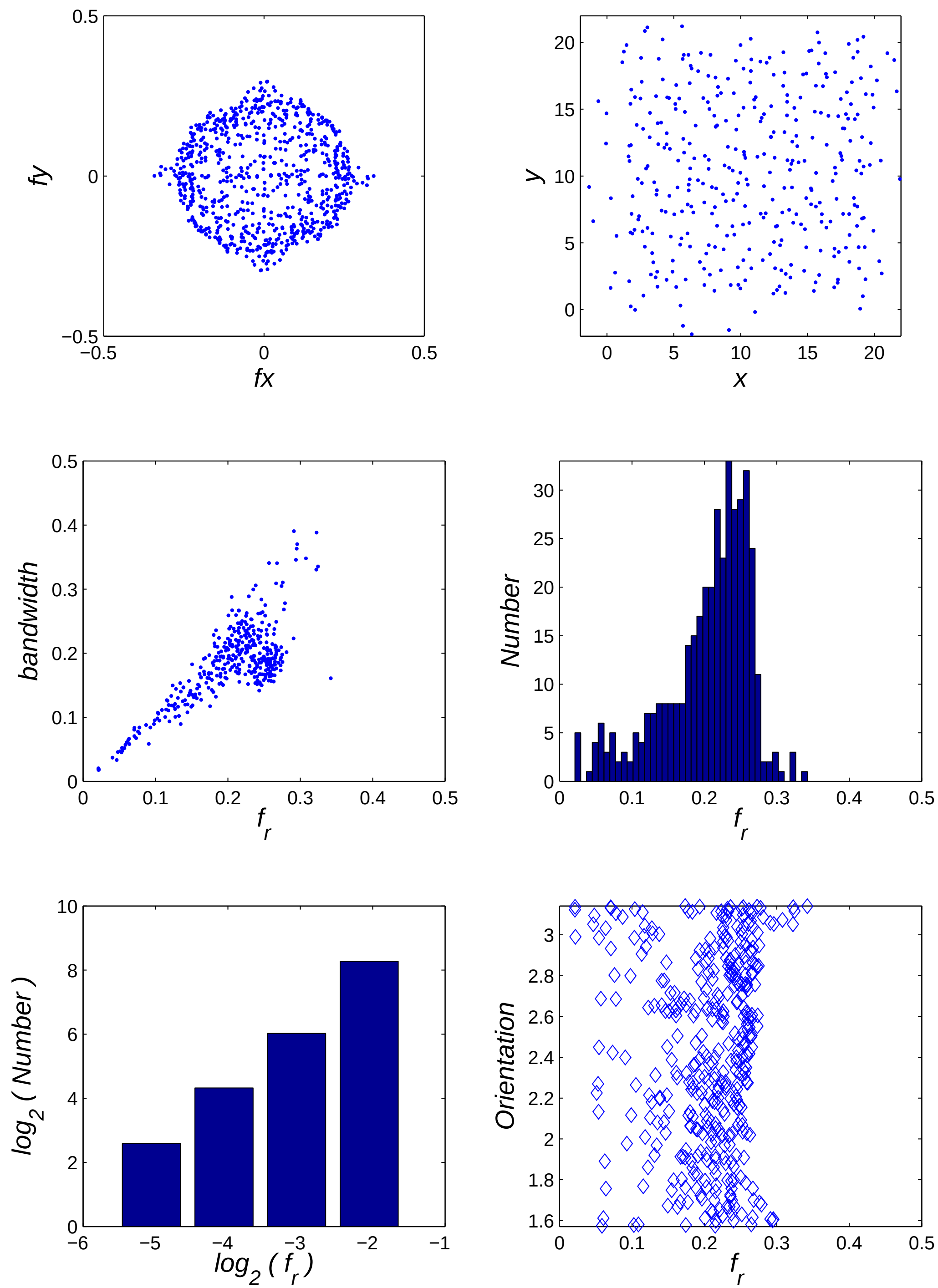
Network implementation



Learned basis functions (200, 12x12)

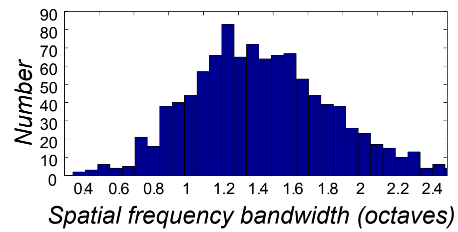


Tiling properties

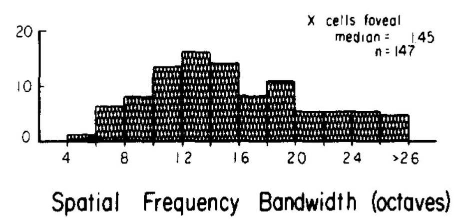


Spatial-frequency bandwidth

Model:

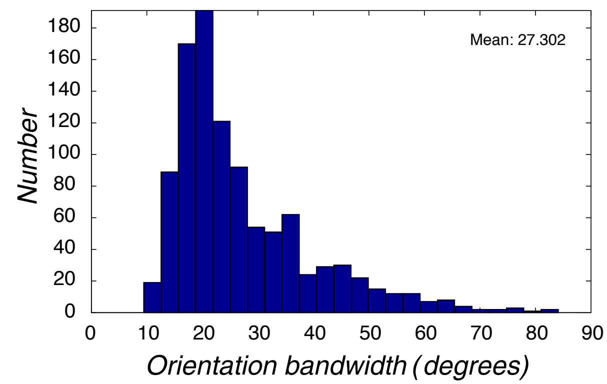


Physiology (DeValois lab):

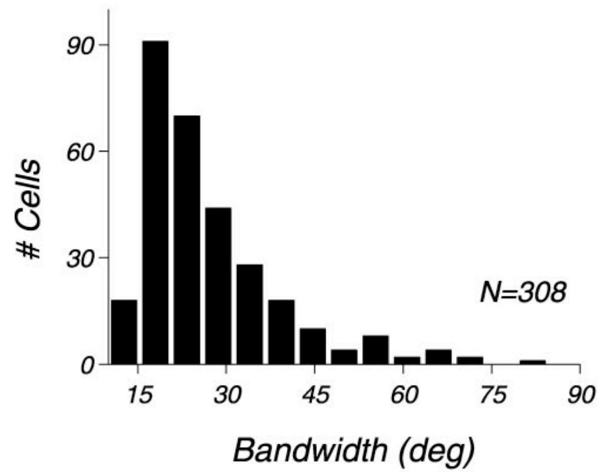


Orientation bandwidth

Model:

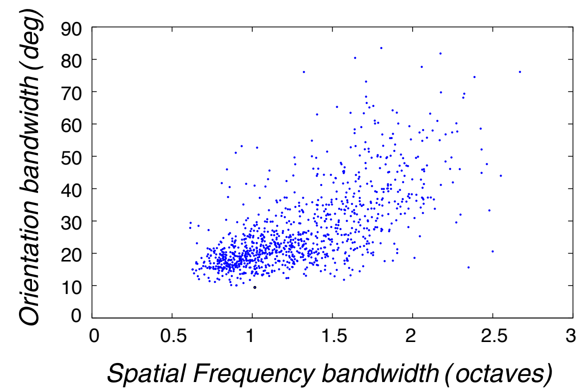


Physiology (Shapley lab):

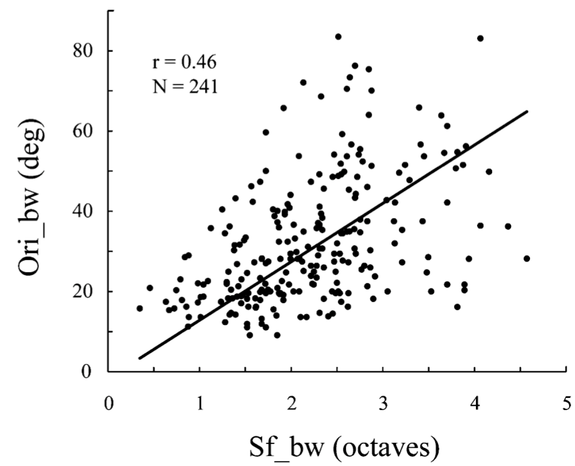


Orientation bandwidth vs. spatial-frequency bandwidth

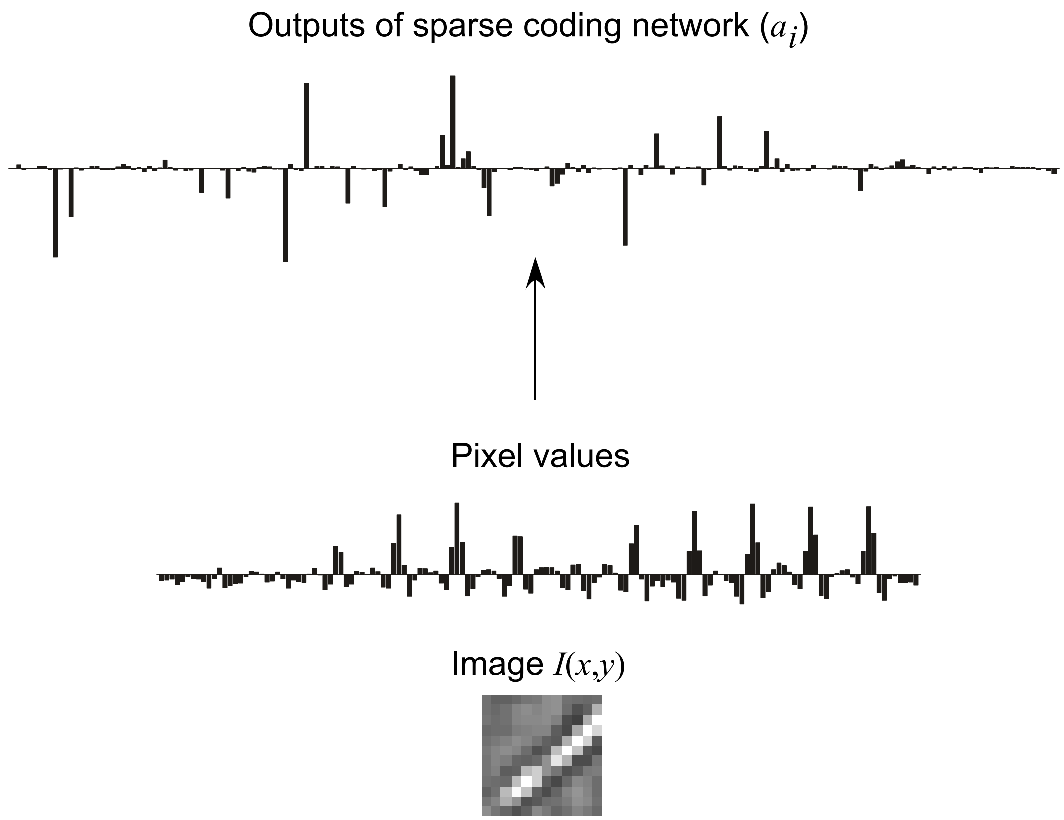
Model:



Physiology (Shapley lab):

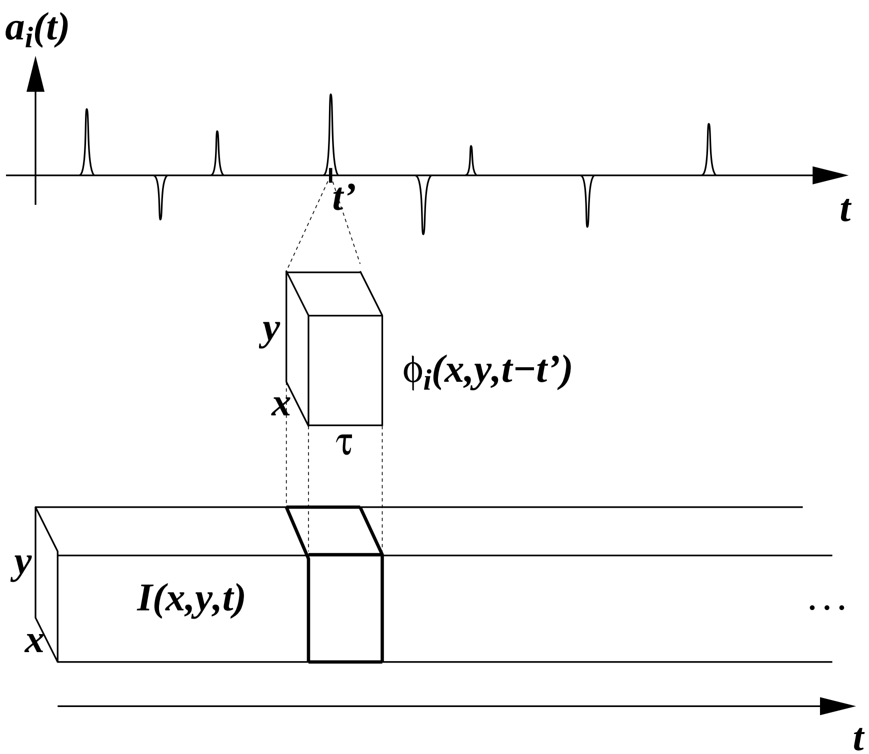


Sparsification



Space-time image model

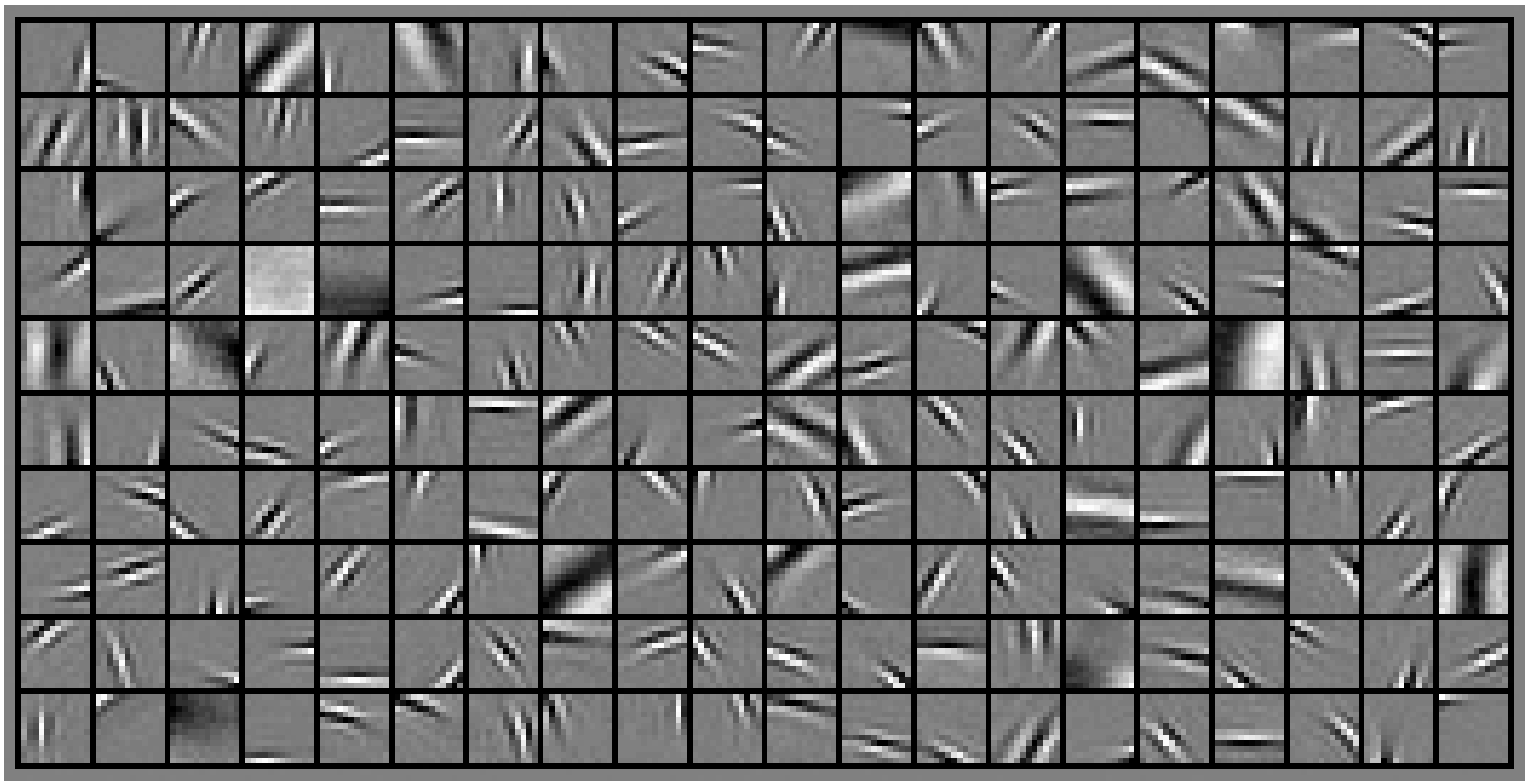
$$I(x,y,t) = \sum_i a_i(t) * \phi_i(x,y,t) + \nu(x,y,t)$$



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.*

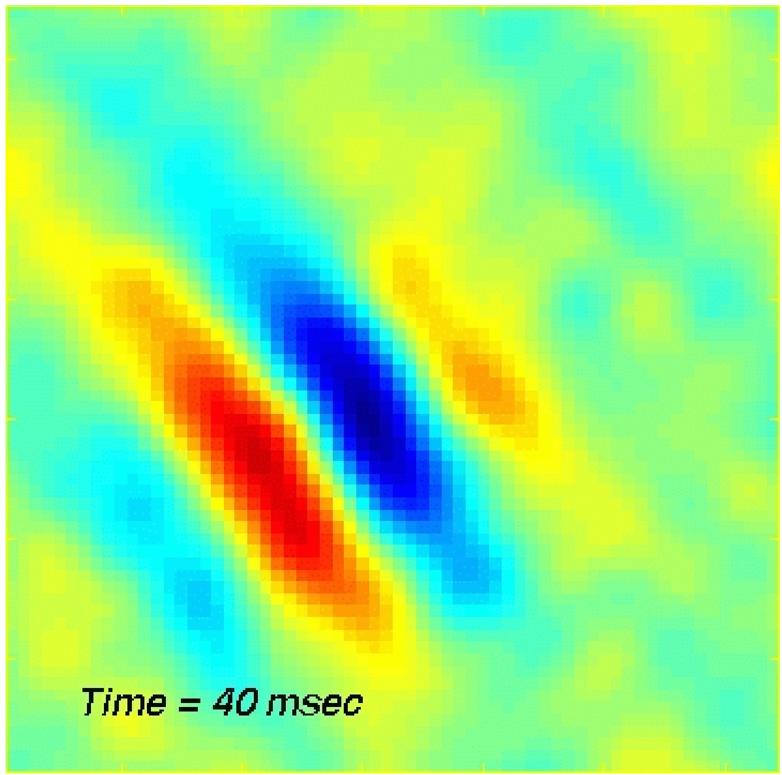
Learned space-time basis functions (200, $12 \times 12 \times 7$)

Training set: nature documentary

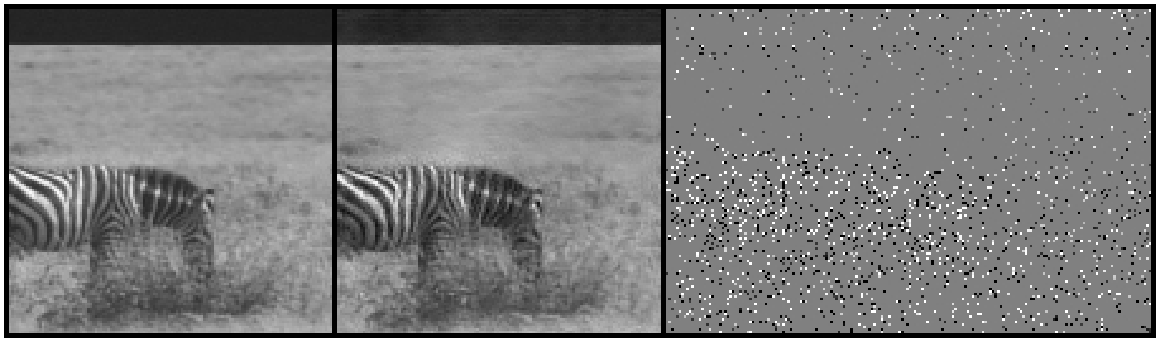


V1 space-time receptive field

(Courtesy of Dario Ringach)



Spike encoding and reconstruction



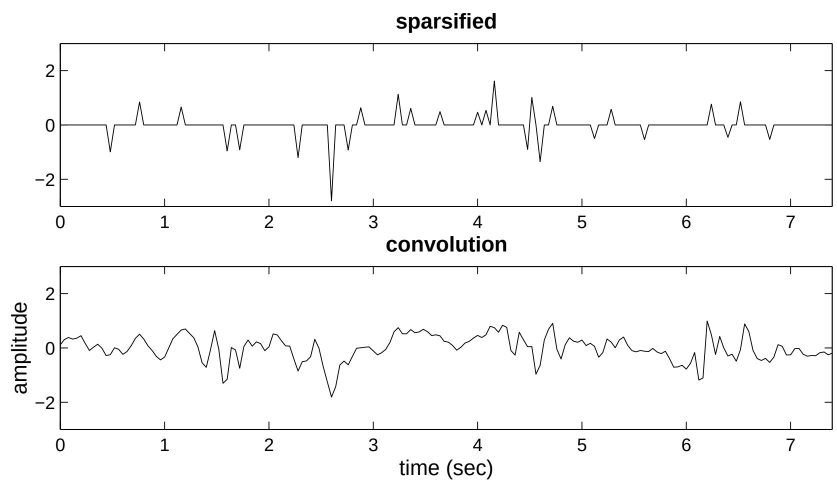
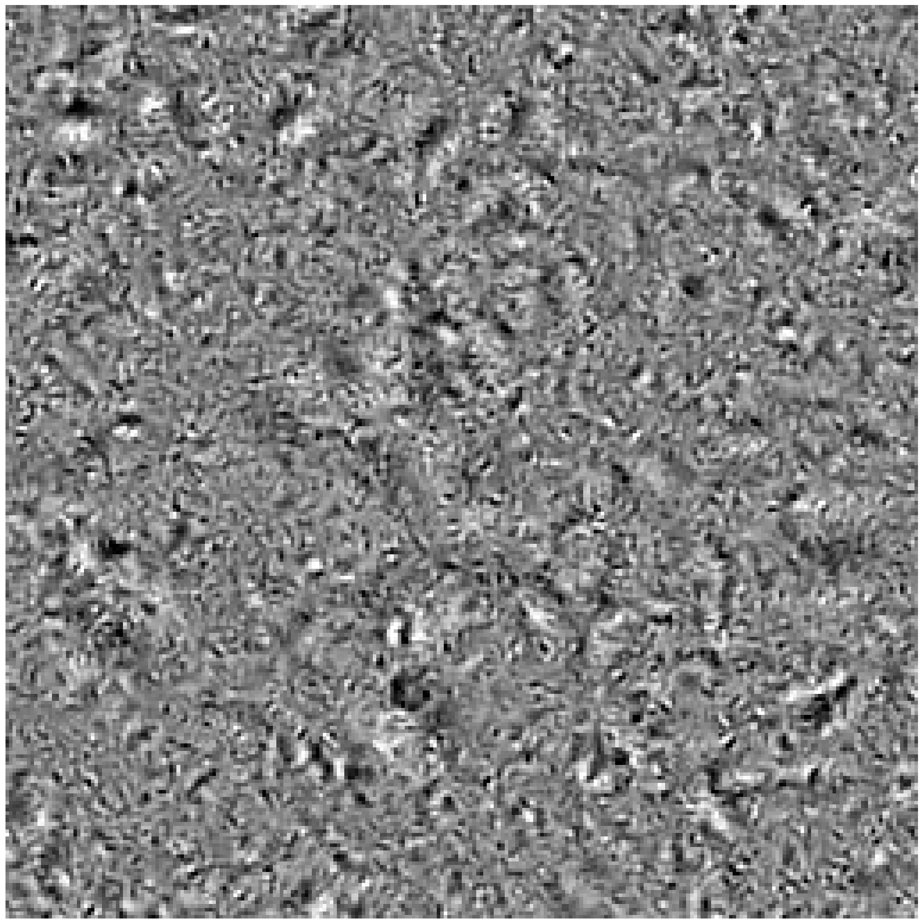
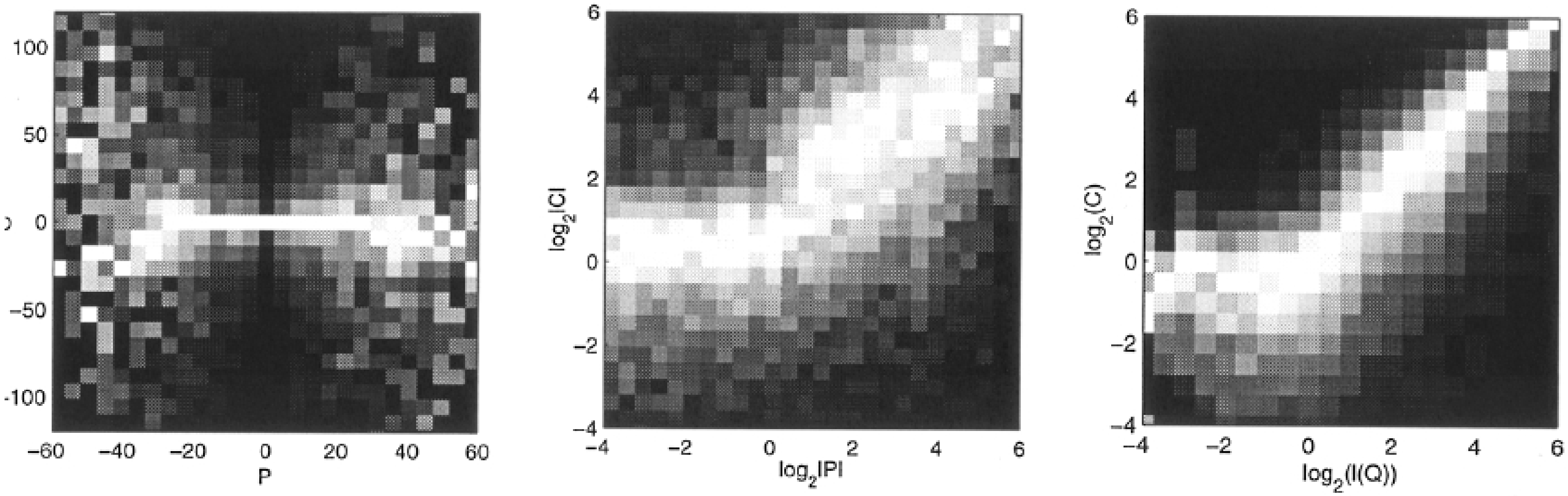


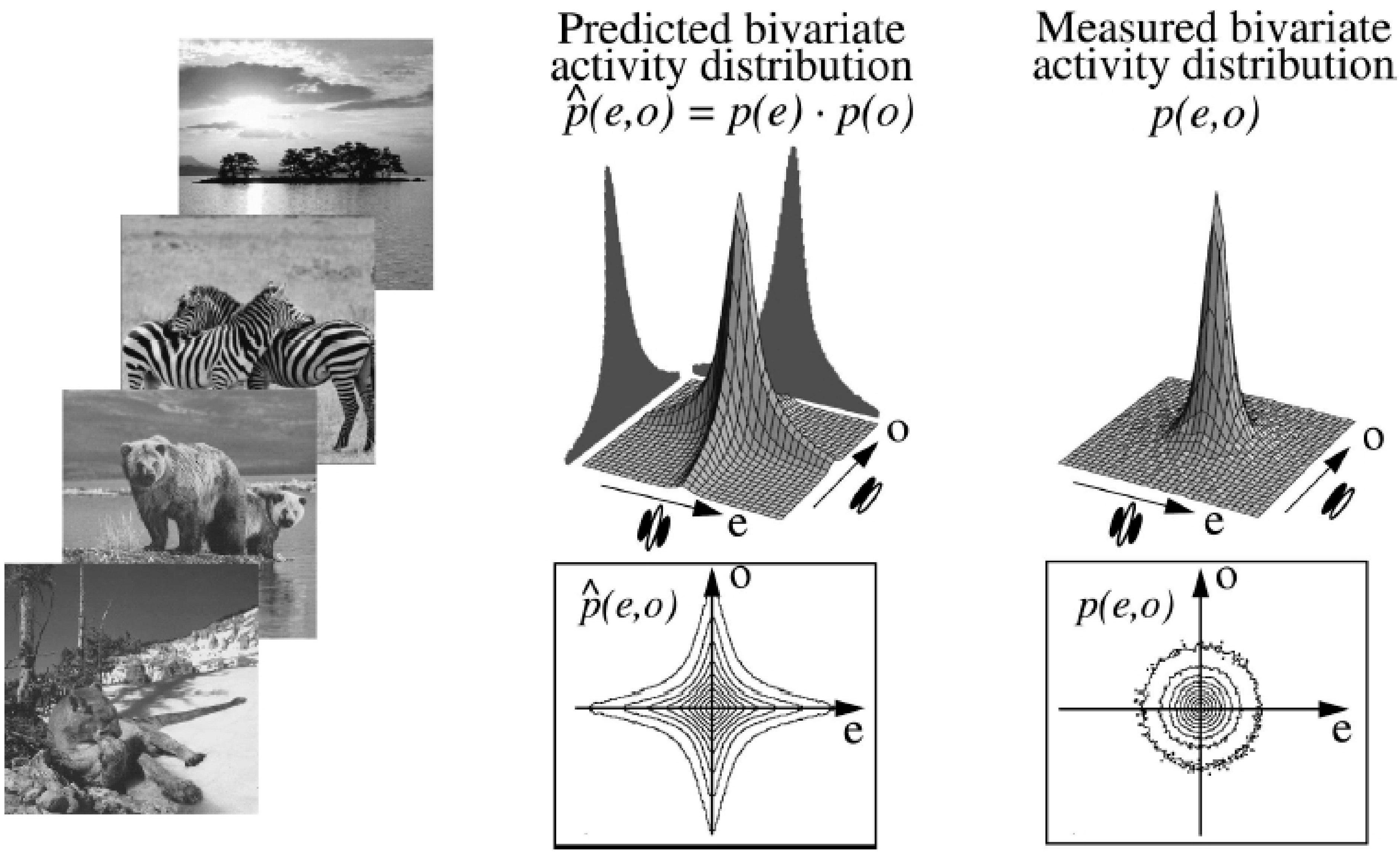
Image synthesis - higher-order statistics



Statistical dependencies among coefficients

Buccigrossi & Simoncelli (1997)



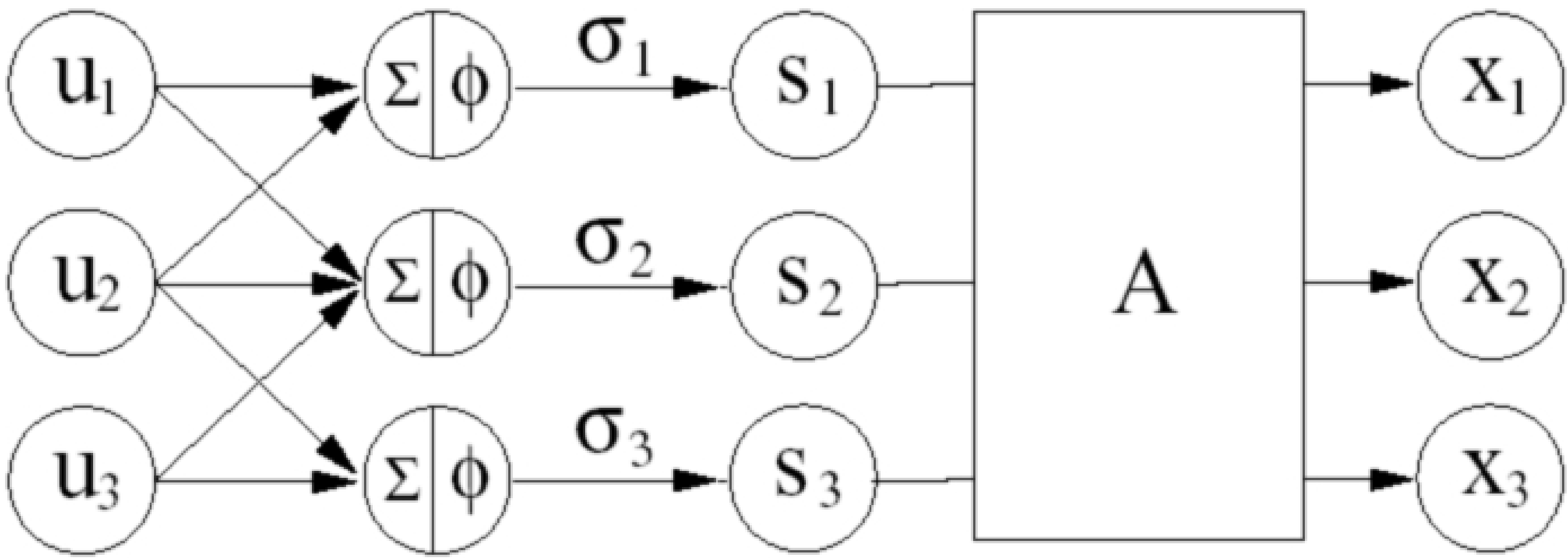

Predicted bivariate activity distribution
$\hat{p}(e,o) = p(e) \cdot p(o)$
Measured bivariate activity distribution
$p(e,o)$
o
e
o
e
$\hat{p}(e,o)$
o
e
$p(e,o)$
o
e

Hierarchical models for capturing dependencies among sparse components

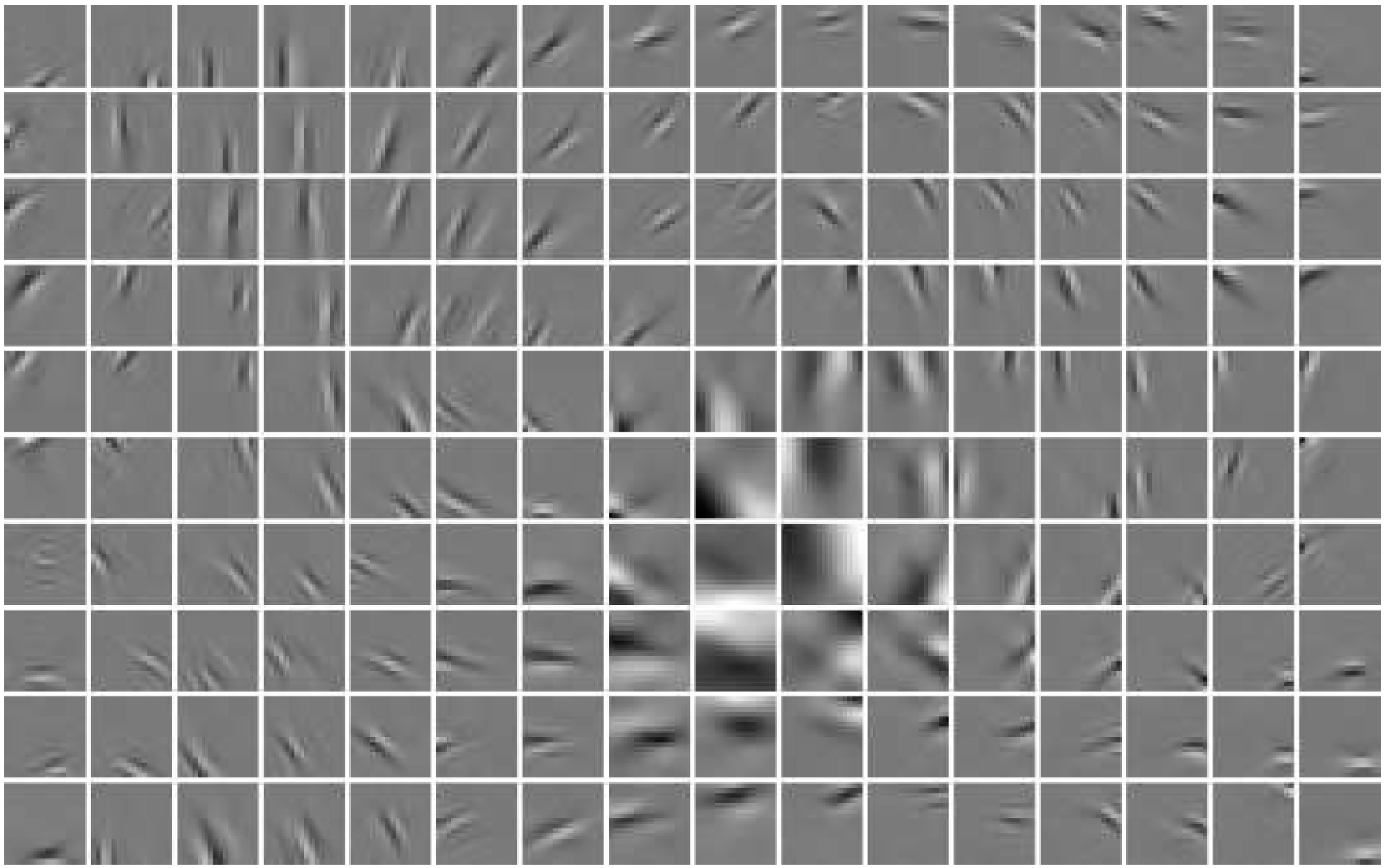
$$
\begin{aligned}
a_i &= \overbrace{\sigma_i}^{\text{power-law}} \times \overbrace{z_i}^{\text{Gaussian}} \\
\sigma_i &= f\Big(\sum_j \Psi_{ij}\, b_j\Big)
\end{aligned}
$$

Wainwright & Simoncelli (2002)
Hyvarinen & Hoyer (2002)
Karklin & Lewicki (2003)
Schwartz & Sejnowski (2004)
Osindero & Hinton (2005)

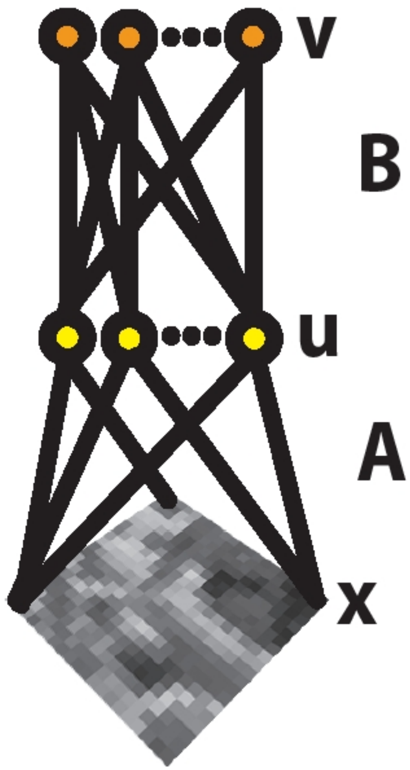
'Topographic ICA' - Hyvarinen & Hoyer (2002)



'Topographic ICA' - Hyvarinen & Hoyer (2002)



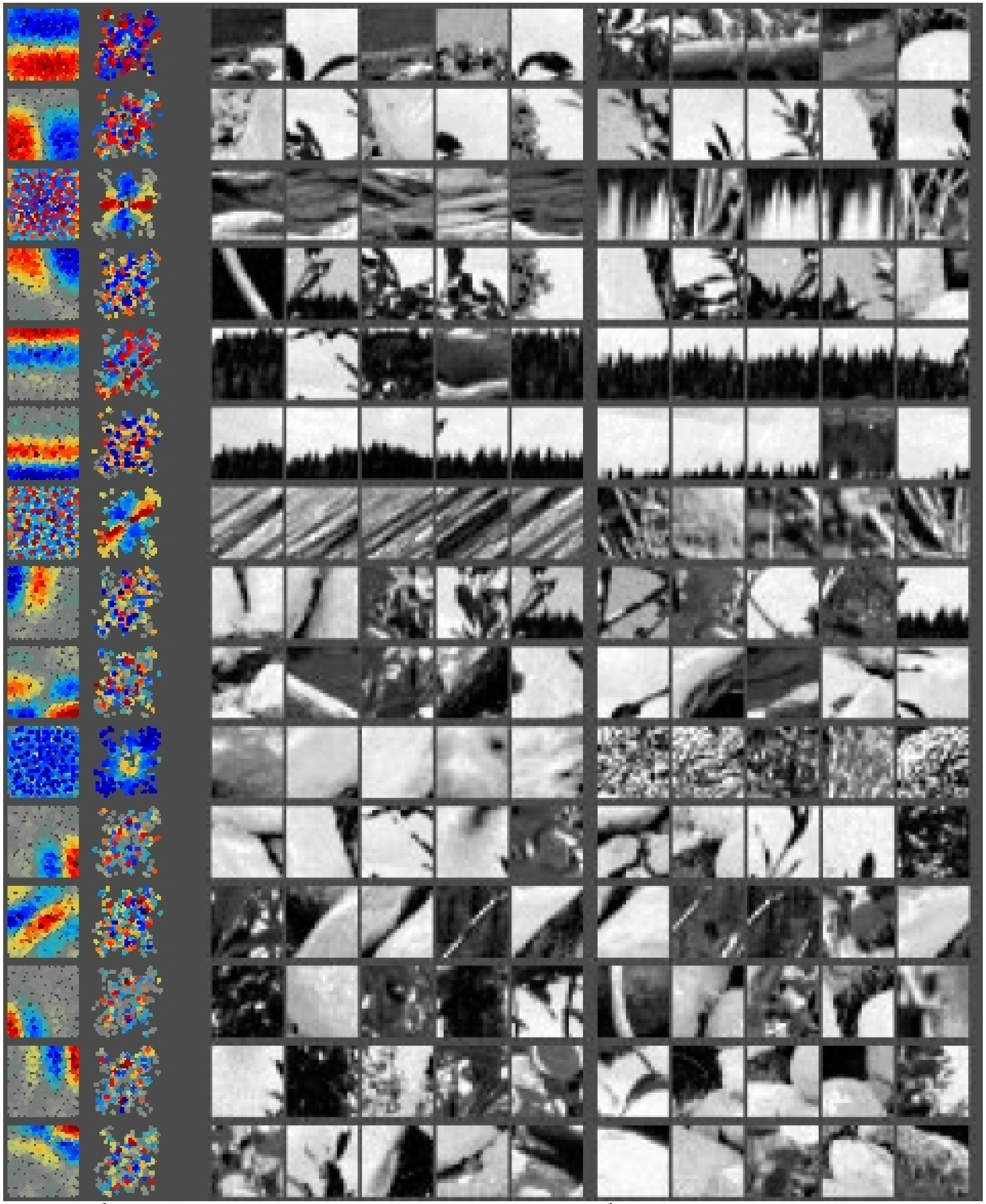
Learning the neighborhoods - Karklin & Lewicki (2003)



$$\mathbf{x} = \mathbf{A}\,\mathbf{u}$$

$$u_i = \sigma_i\, z_i$$

$$\sigma_i = e^{\sum_j B_{ij}\, v_j}$$



Bilinear models for learning invariant representations

$$\mathbf{z} = \sum_{ij} \mathbf{w}_{ij} \underbrace{x_i}_{\text{`what'}} \underbrace{y_j}_{\text{`where'}}$$

- Tenenbaum & Freeman (2000) - SVD
- Grimes & Rao (2005) - sparse coding

Generative model for transformation and shape

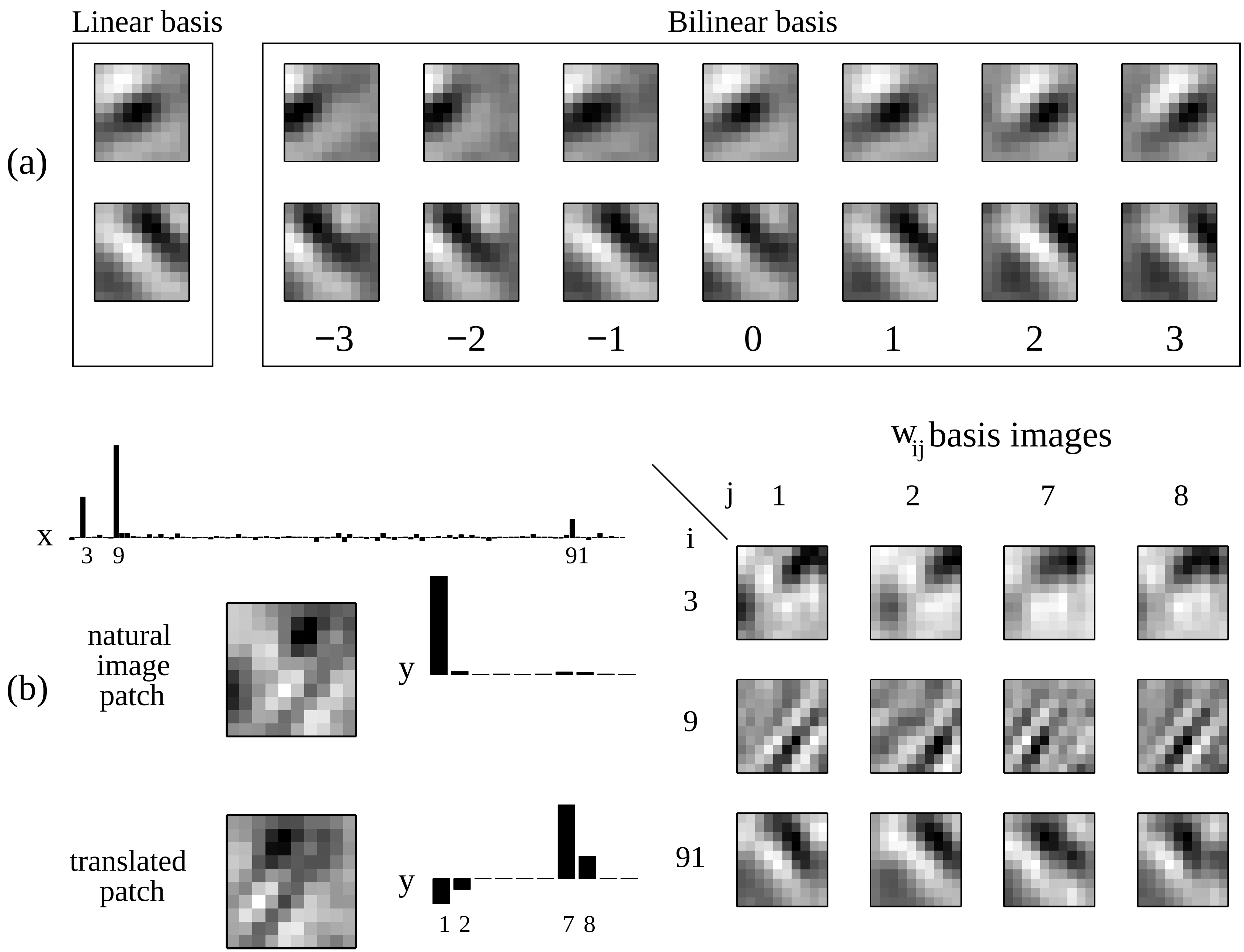
$$
\begin{aligned}
I^0(x) &= \sum_{x'} T(x,x')\, I^1(x') \\
T(x,x') &= \sum_{j} c_j\, \Psi_j(x,x') \\
I^1(x) &= \sum_{i} a_i\, \phi_i(x)
\end{aligned}
$$

Define:

$$\Gamma_{ij}(x) = \sum_{x'} \Psi_j(x,x')\, \phi_i(x')$$

Then:

$$I^0(x) = \sum_{ij} \Gamma_{ij}(x)\, a_i\, c_j$$



Linear basis
Bilinear basis
(a)
−3
−2
−1
0
1
2
3
x
3
9
91
w_{ij} basis images
j
1
2
7
8
i
3
9
91
natural image patch
y
(b)
translated patch
y
1 2
7 8

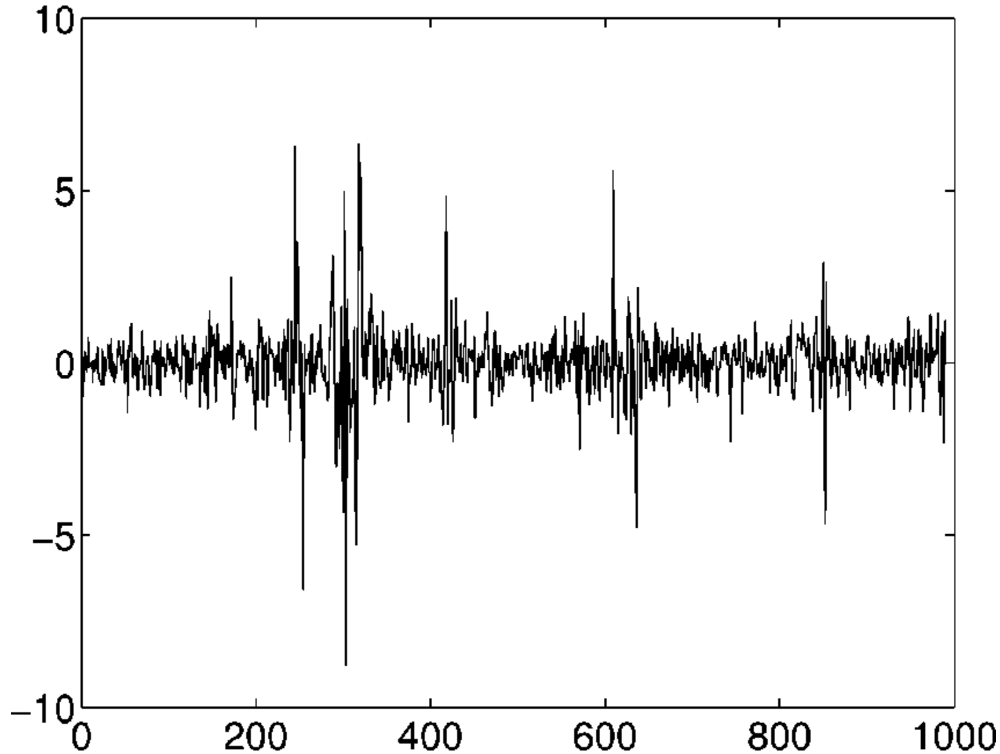
Slow feature analysis

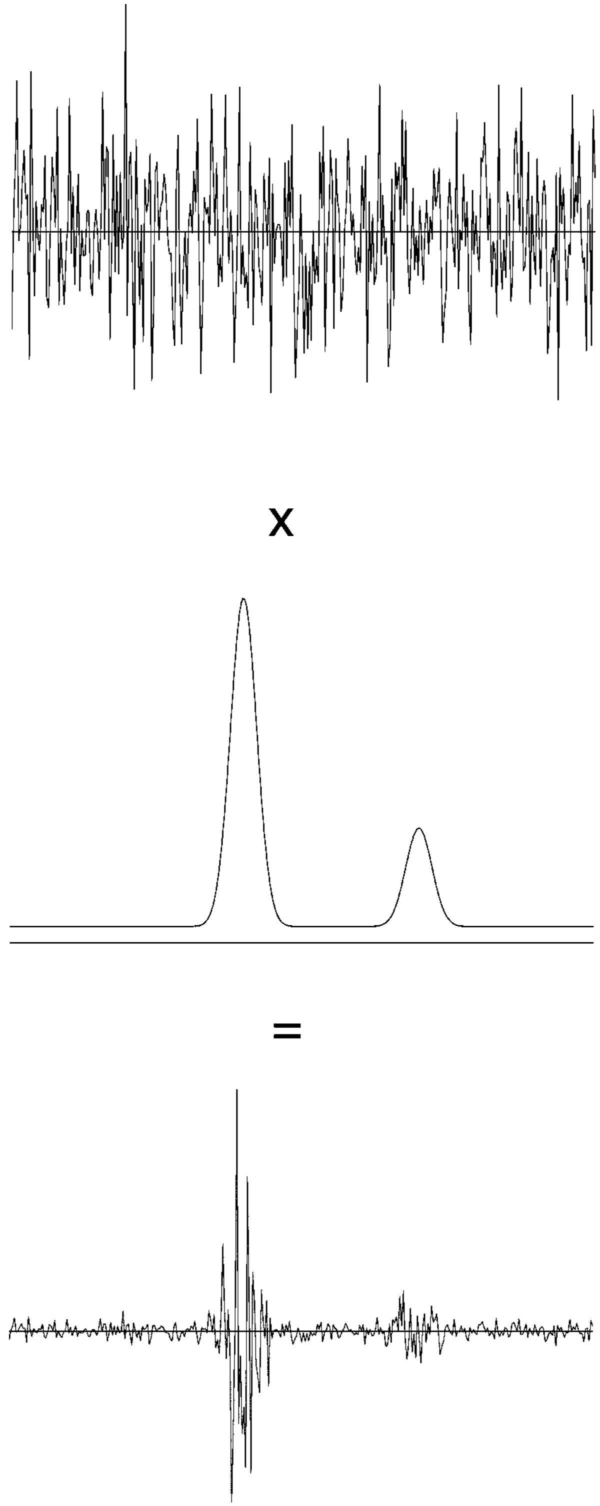
Foldiak (1991), Wiskott (2002)

Learn invariant causes by imposing slowness over time.

Bubbles

Hyvarinen et al. (2003) JOSA *20*





Generative model

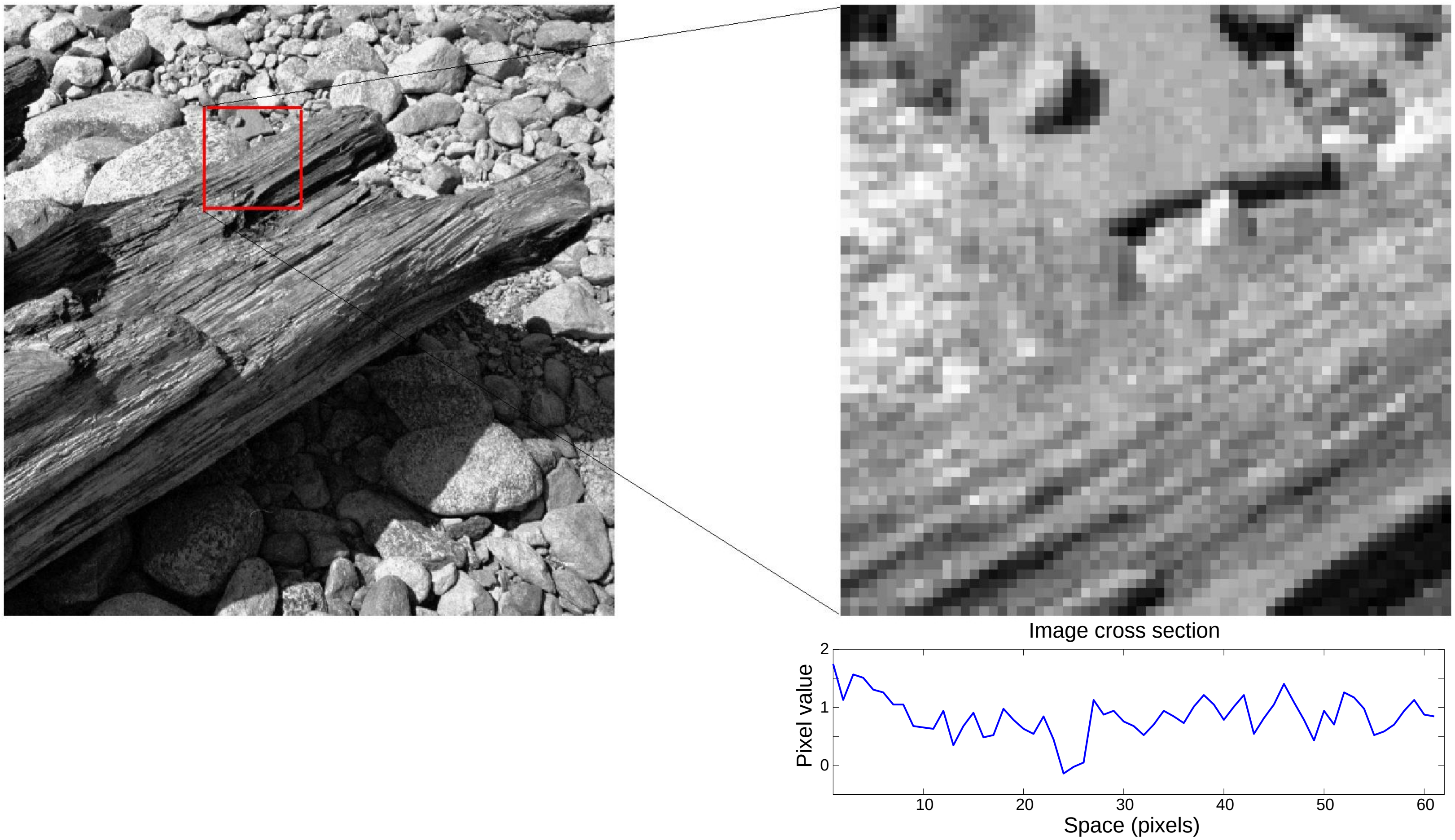
$$I(x,y,t) = \sum_i a_i(t)\,\phi_i(x,y)$$

$$a_i(t) = \underbrace{\sigma_i(t)}_{\text{slow,sparse}} \; \underbrace{z_i(t)}_{\text{fast}}$$

Towards intermediate-level representations

- The problem of scene analysis
- Insights from psychophysics
  - Occlusion and figure-ground representation (Nakayama & Shimojo)
  - Adaptation (Webster/Leopold)

The problem of scene analysis



How do you interpret an edge?



Object recognition depends on context

Torralba & Sinha (2001)

Object recognition depends on context

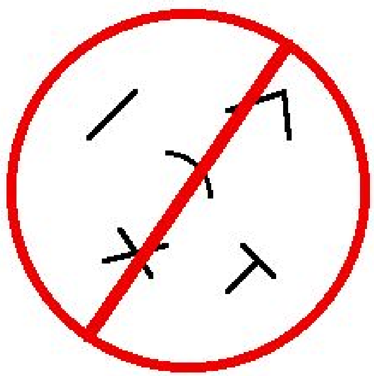
Torralba & Sinha (2001)

Object recognition depends of context

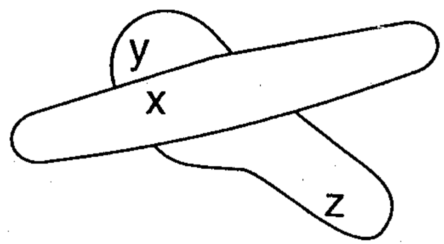
Torralba & Sinha (2001)

Visual representations are 3D, not 2D

Nakayama K, He ZJ, and Shimojo S. (1995) **Visual surface representation: a critical link between lower-level and higher level vision.** In: S.M. Kosslyn and D.N. Osherson, Eds, *An Invitation to Cognitive Science*. MIT Press, pp. 1-70.



Rules of occlusion

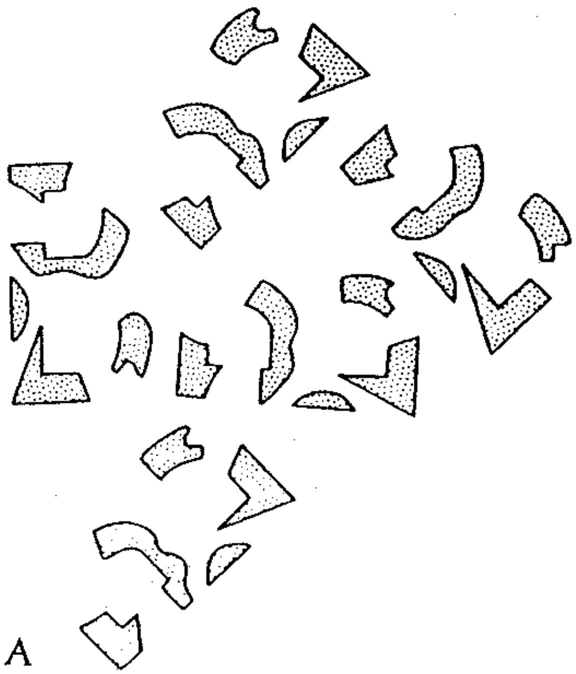


1. When image regions corresponding to different surfaces meet, only one region can "own" the border between them.

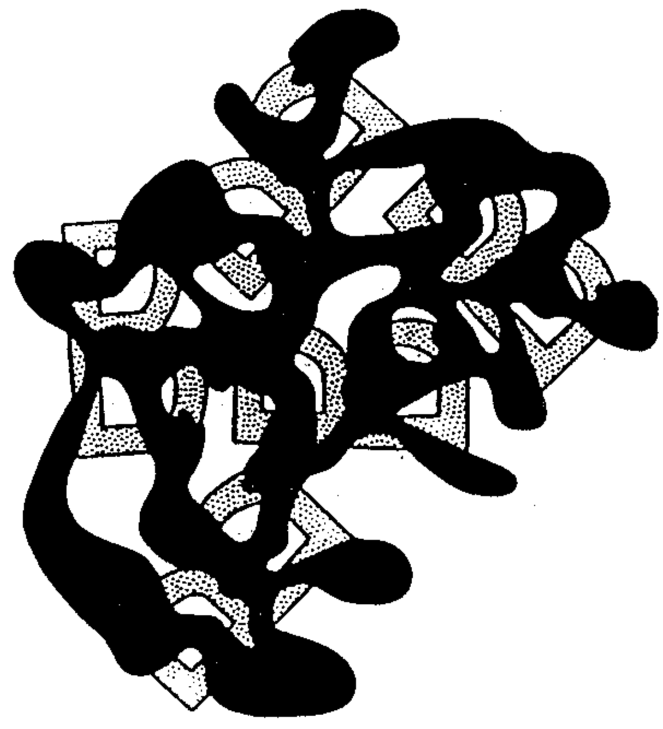
2. Under conditions of surface opacity, a border is owned by the region that is coded as being in front.

3. A region that does not own a border is effectively unbounded. Unbounded regions can connect to other unbounded regions to form larger surfaces completing behind.

Bregman B's

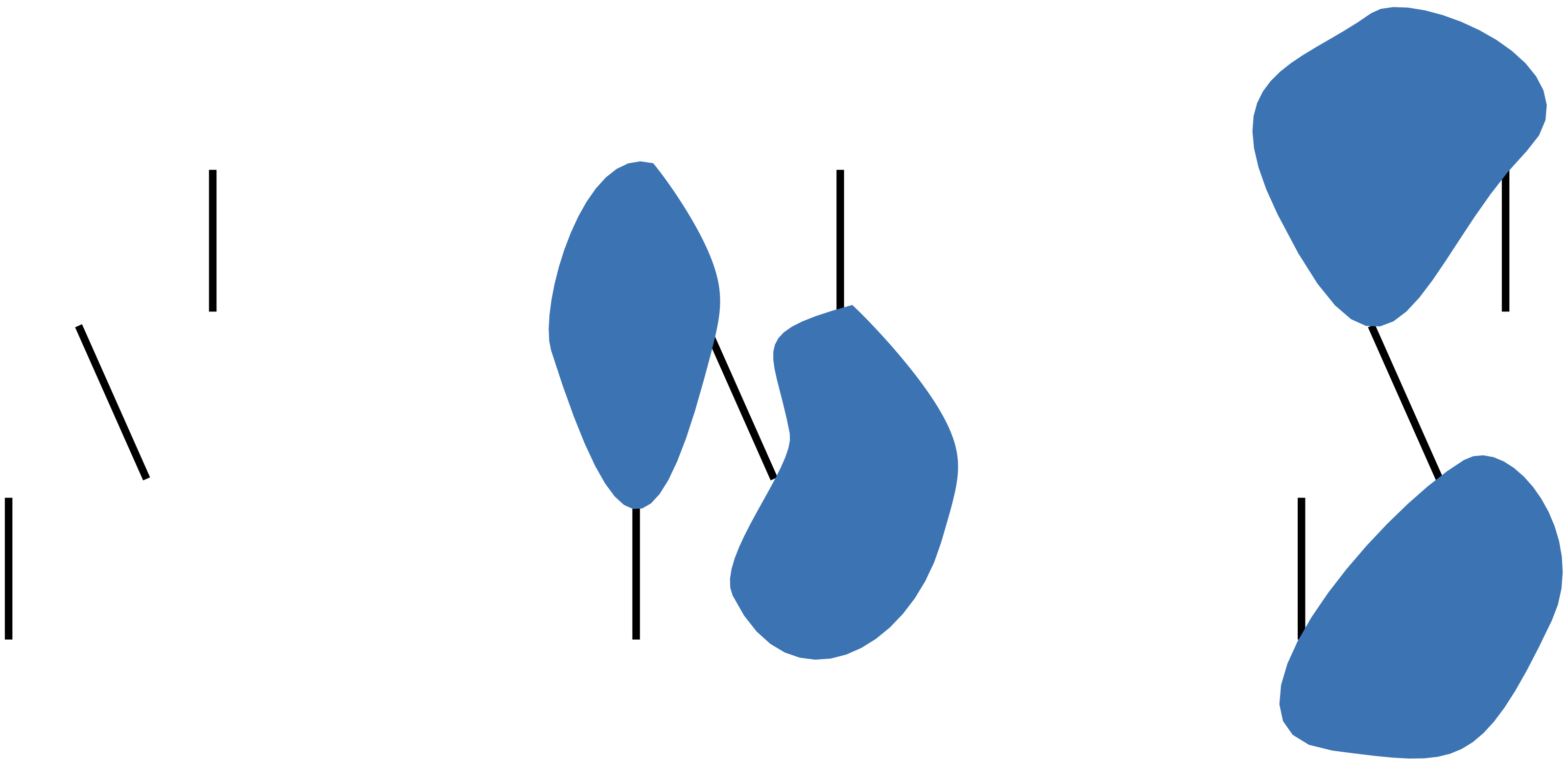


A

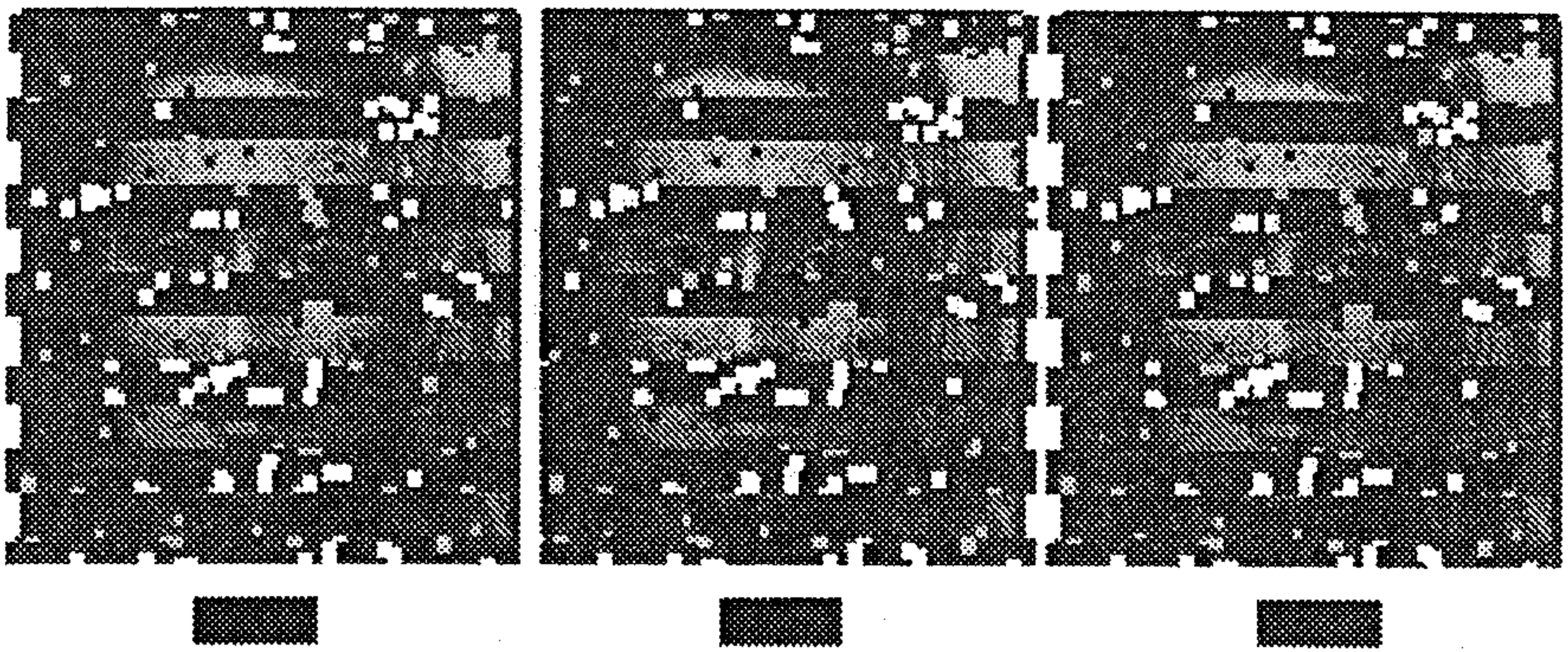


B

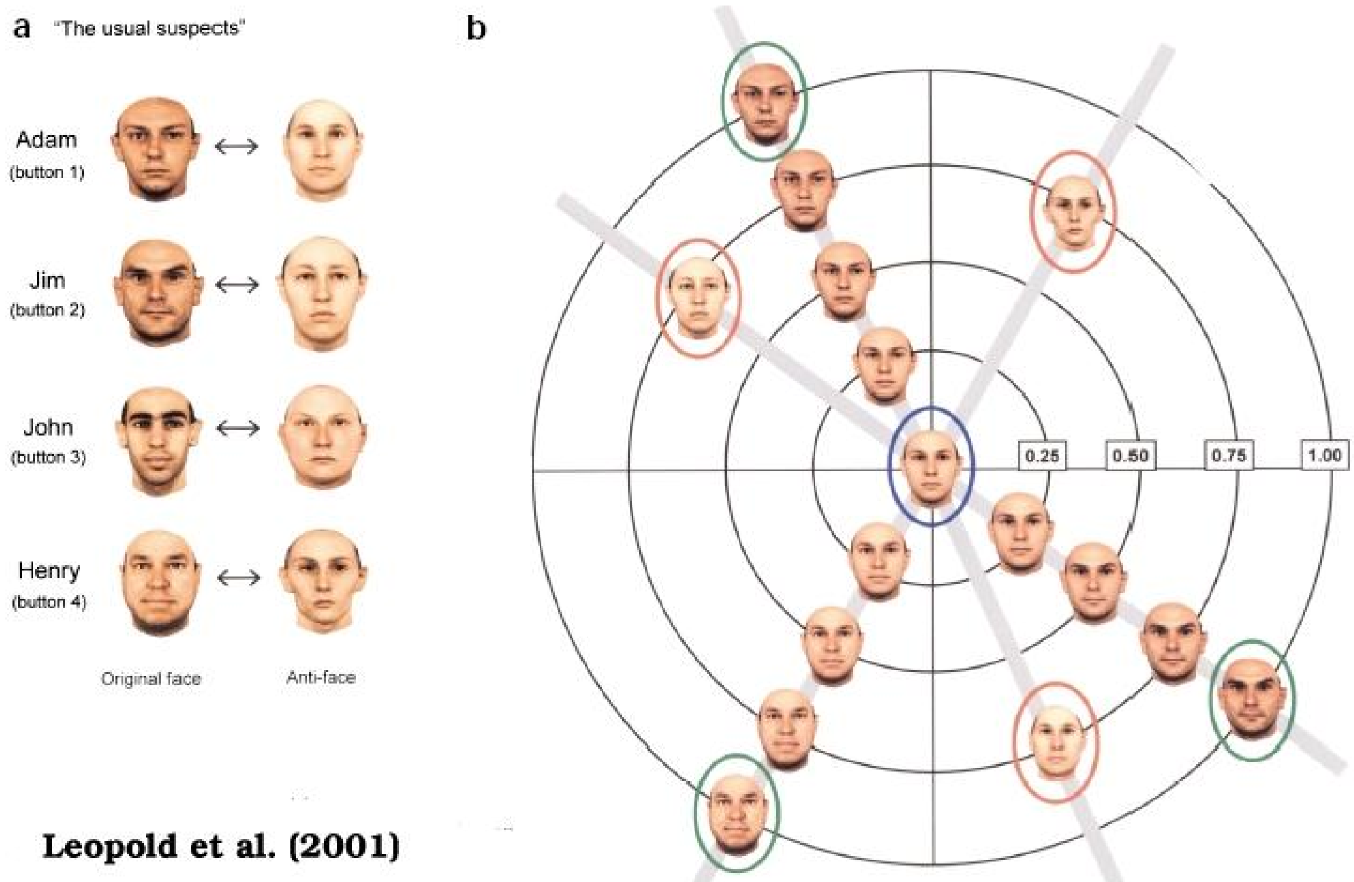
Occluders determine object completion



Amodal completion depends on depth assignment



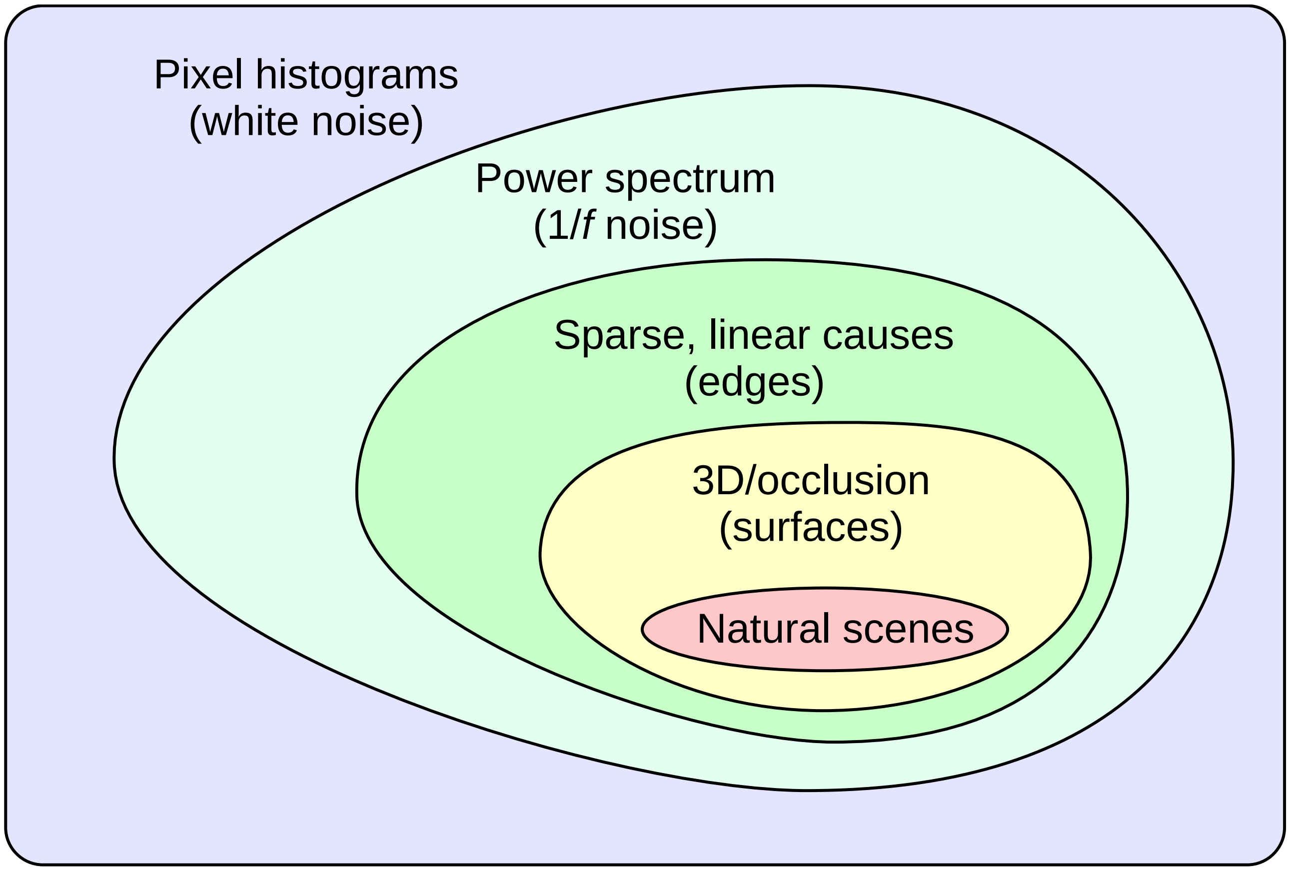
Adaptation reveals internal axes of representation



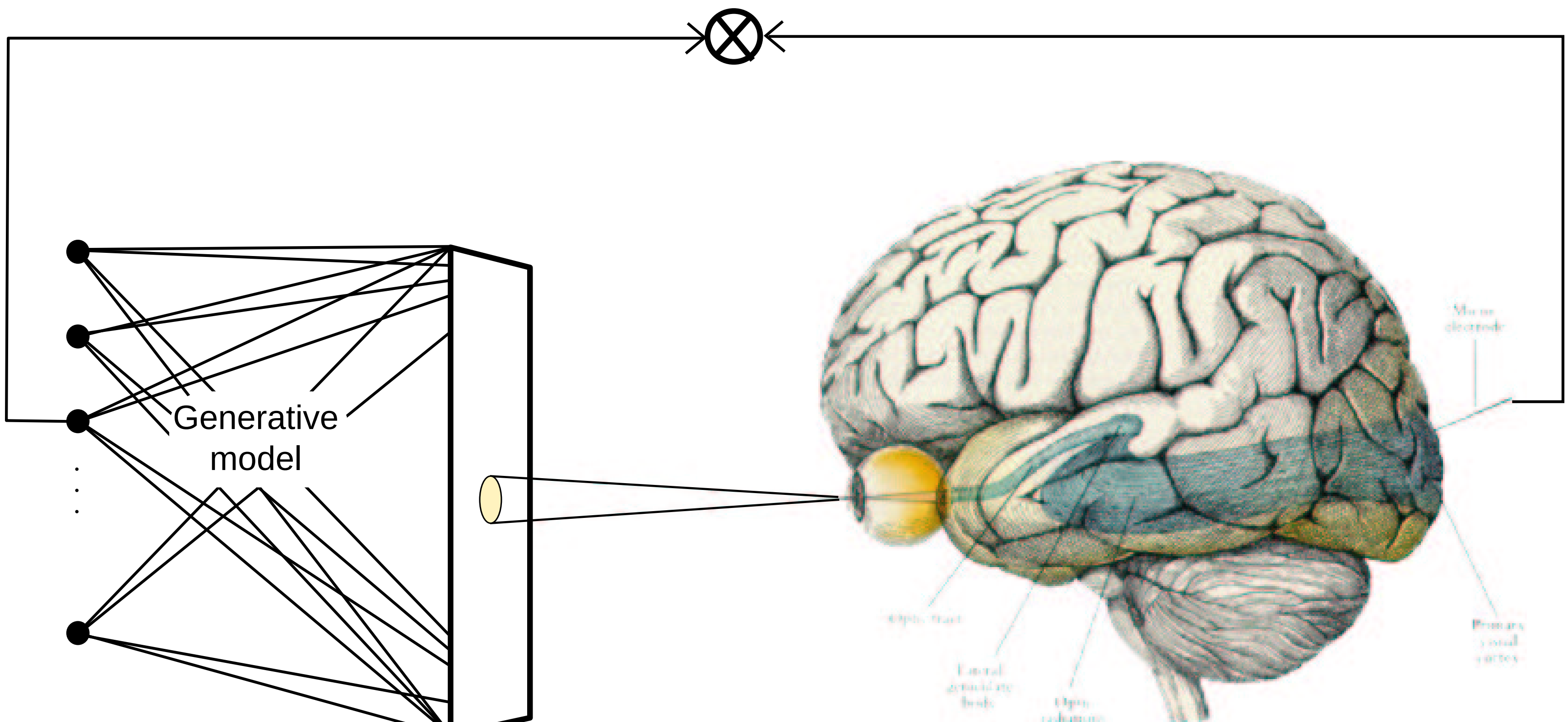
Leopold et al. (2001)

Summary

Image models



Using generative models as experimental tools



Main points

- Image statistics and neural coding
  - contrast distribution $\rightarrow$ histogram equalization
  - $1/f^2$ power spectrum $\rightarrow$ whitening
  - higher-order statistics $\rightarrow$ V1 simple cells (Gabor functions)
- Cortex as generative model
  - Neurons represent causes of natural images.
- Future challenges
  - intermediate-level representations
  - surfaces/occlusion

Further information and details

baolshausen@berkeley.edu
http://redwood.ucdavis.edu/bruno