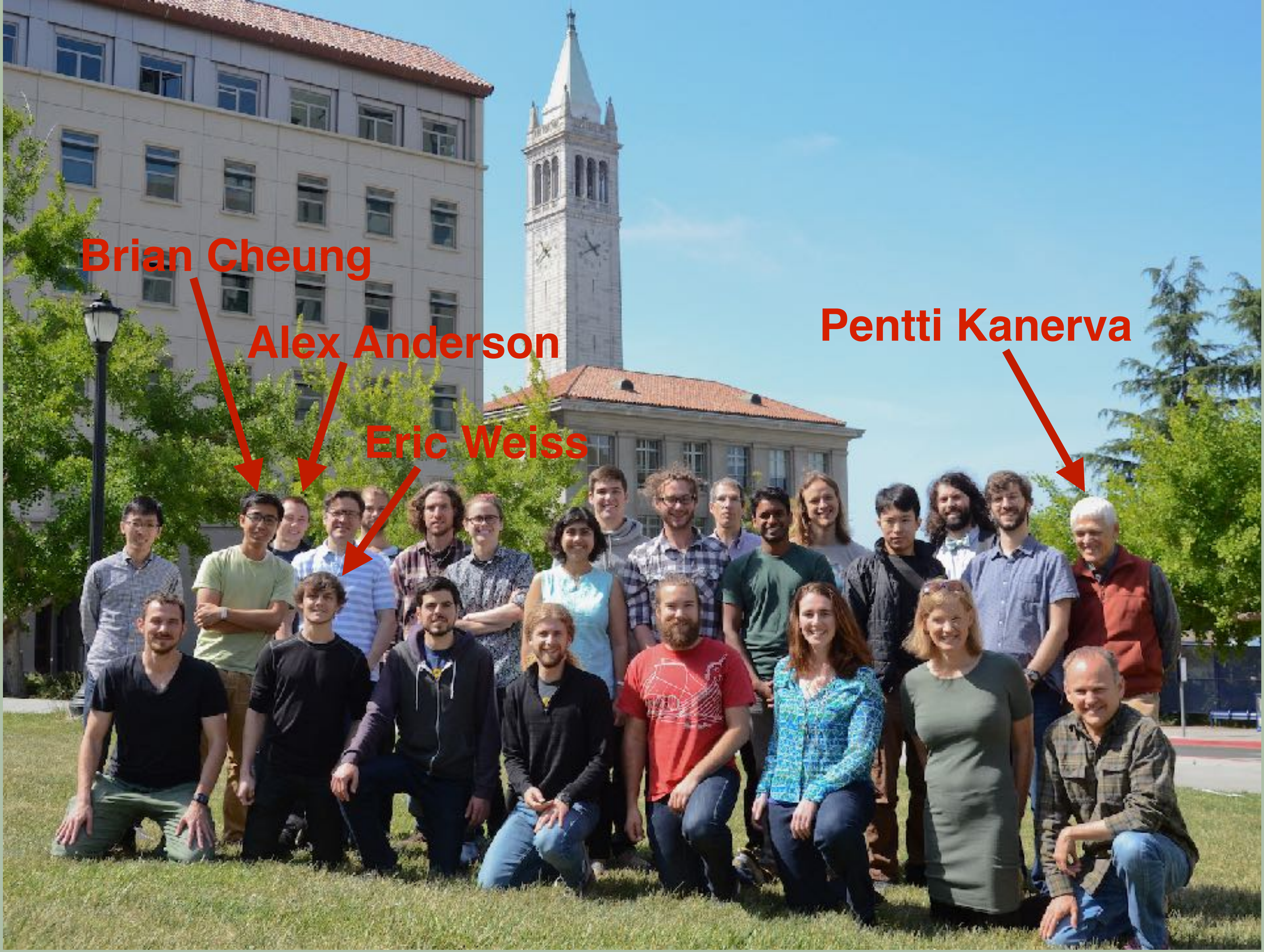
# Learning representations for active vision

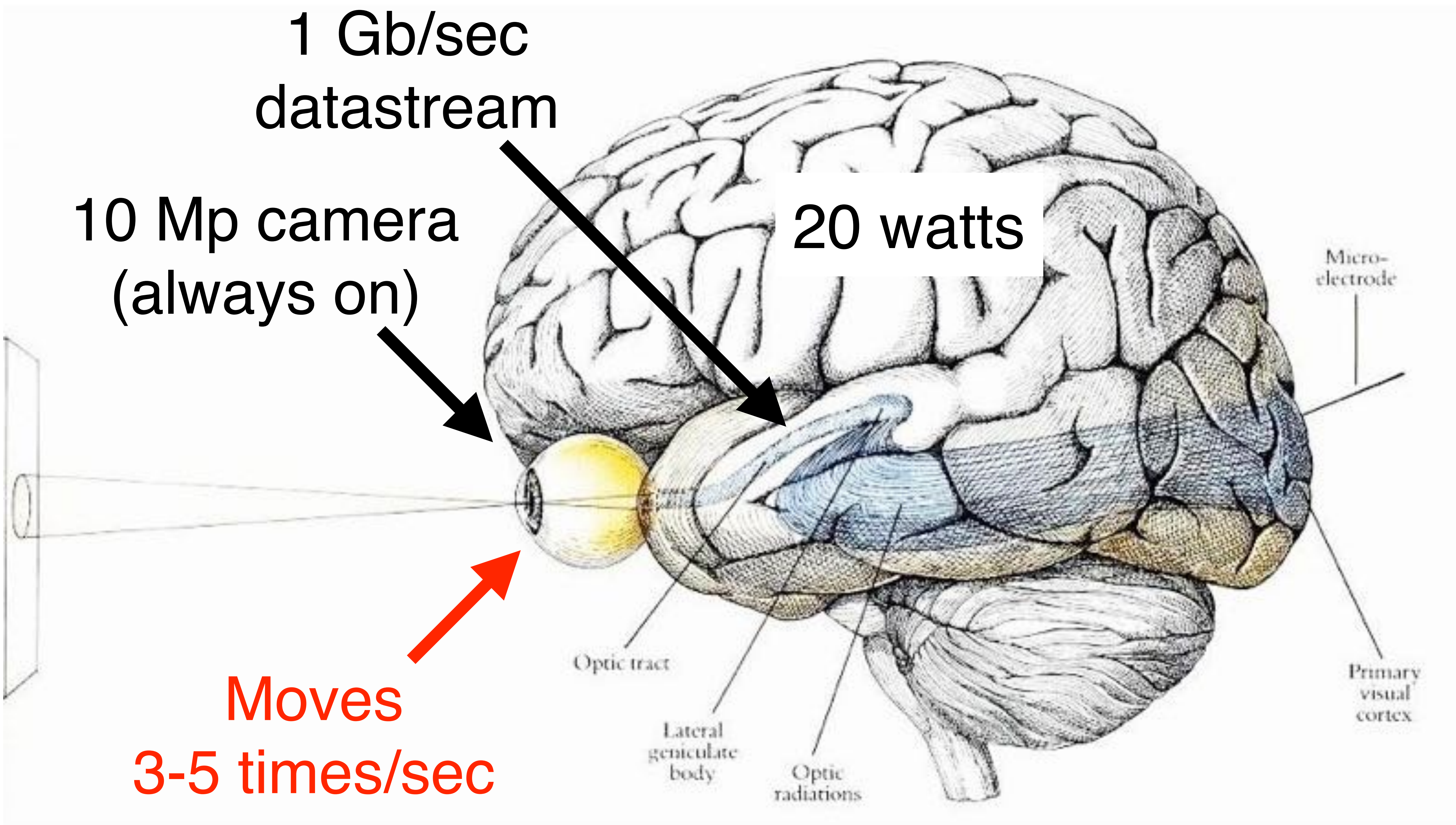
**Bruno Olshausen**
Redwood Center for Theoretical Neuroscience,
Helen Wills Neuroscience Institute, and School of Optometry
UC Berkeley

REDWOOD CENTER
for Theoretical Neuroscience

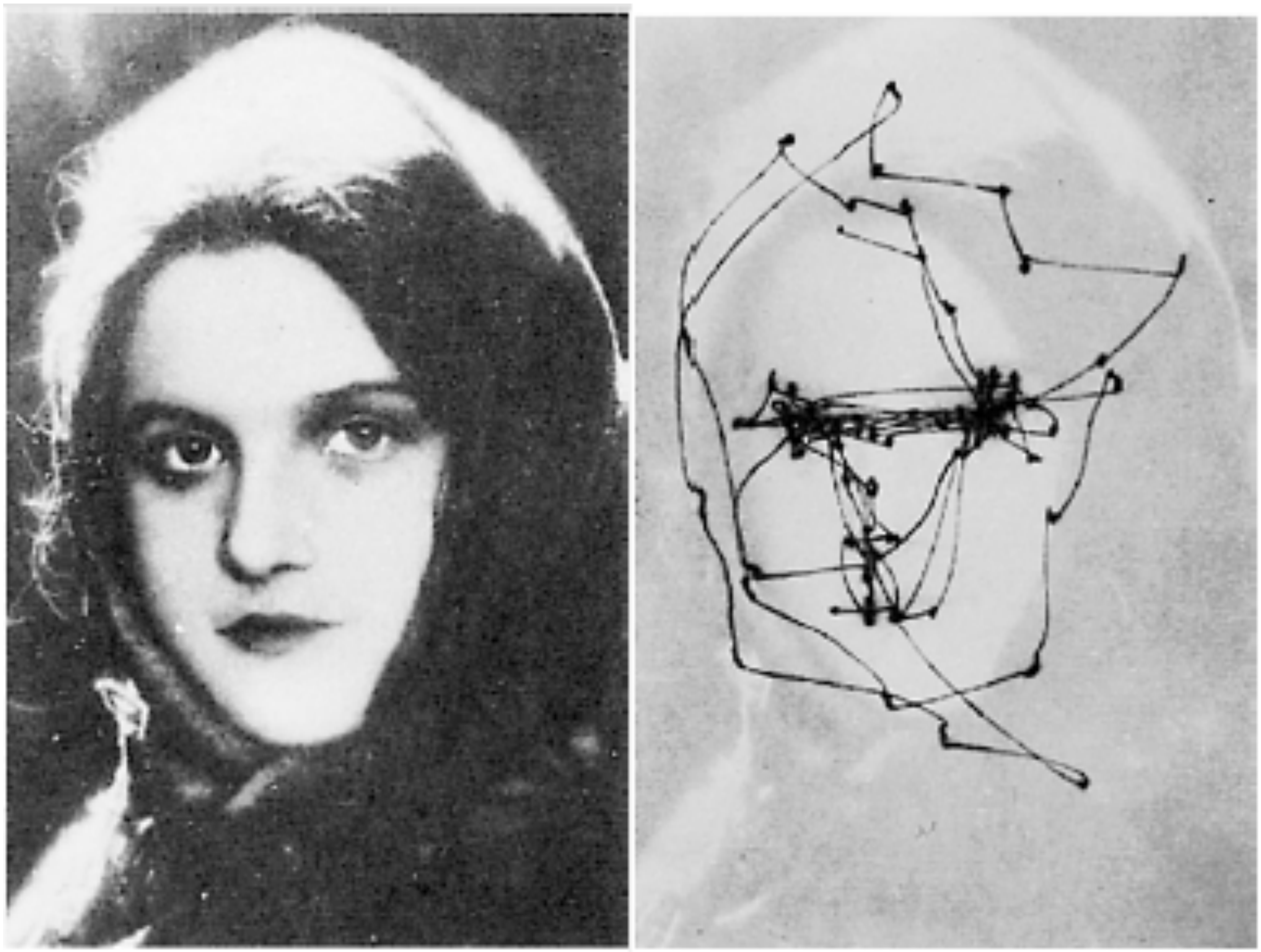



Redwood Center for Theoretical Neuroscience - April 2016

# What are the principles governing information processing in this system?

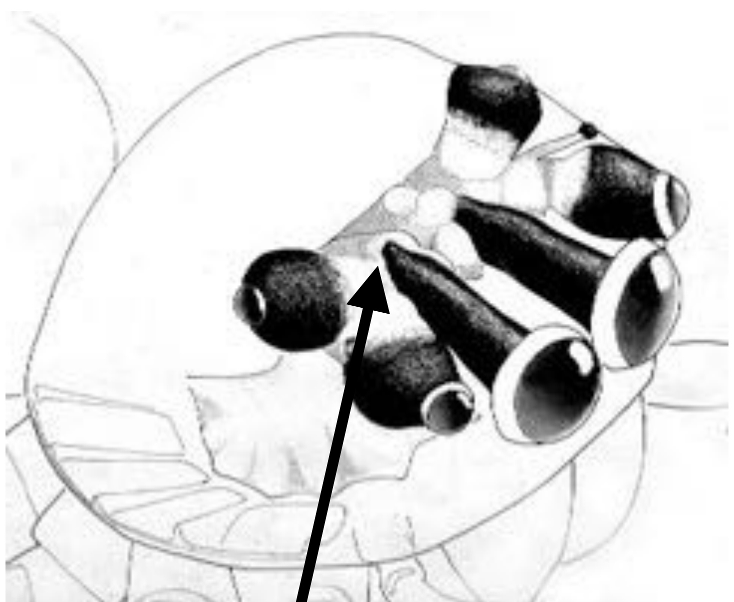


# Human eye movements during viewing of an image

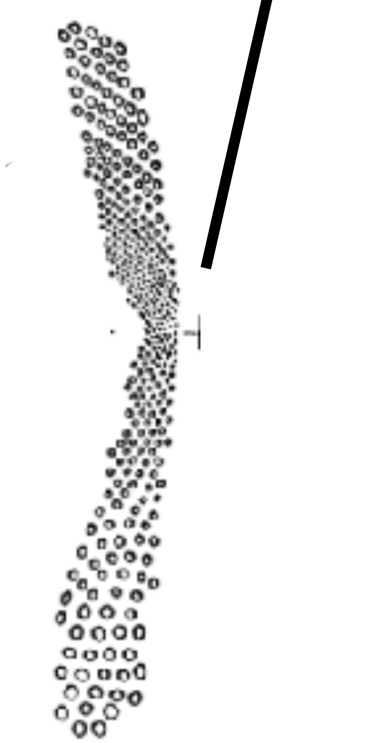


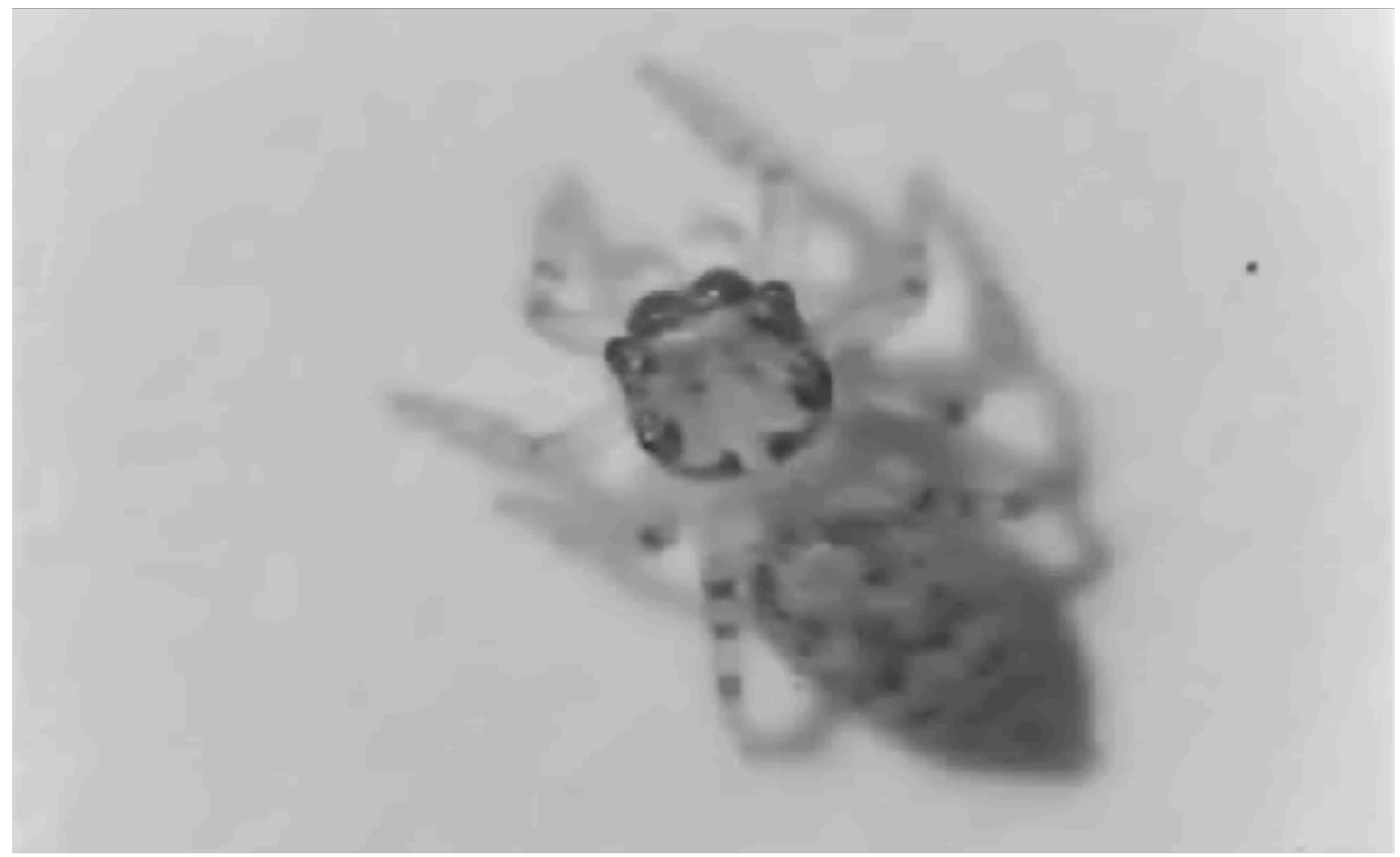
Yarbus (1967)

# Active vision in jumping spiders



(Wayne Maddison)





(Bair & Olshausen, 1991)

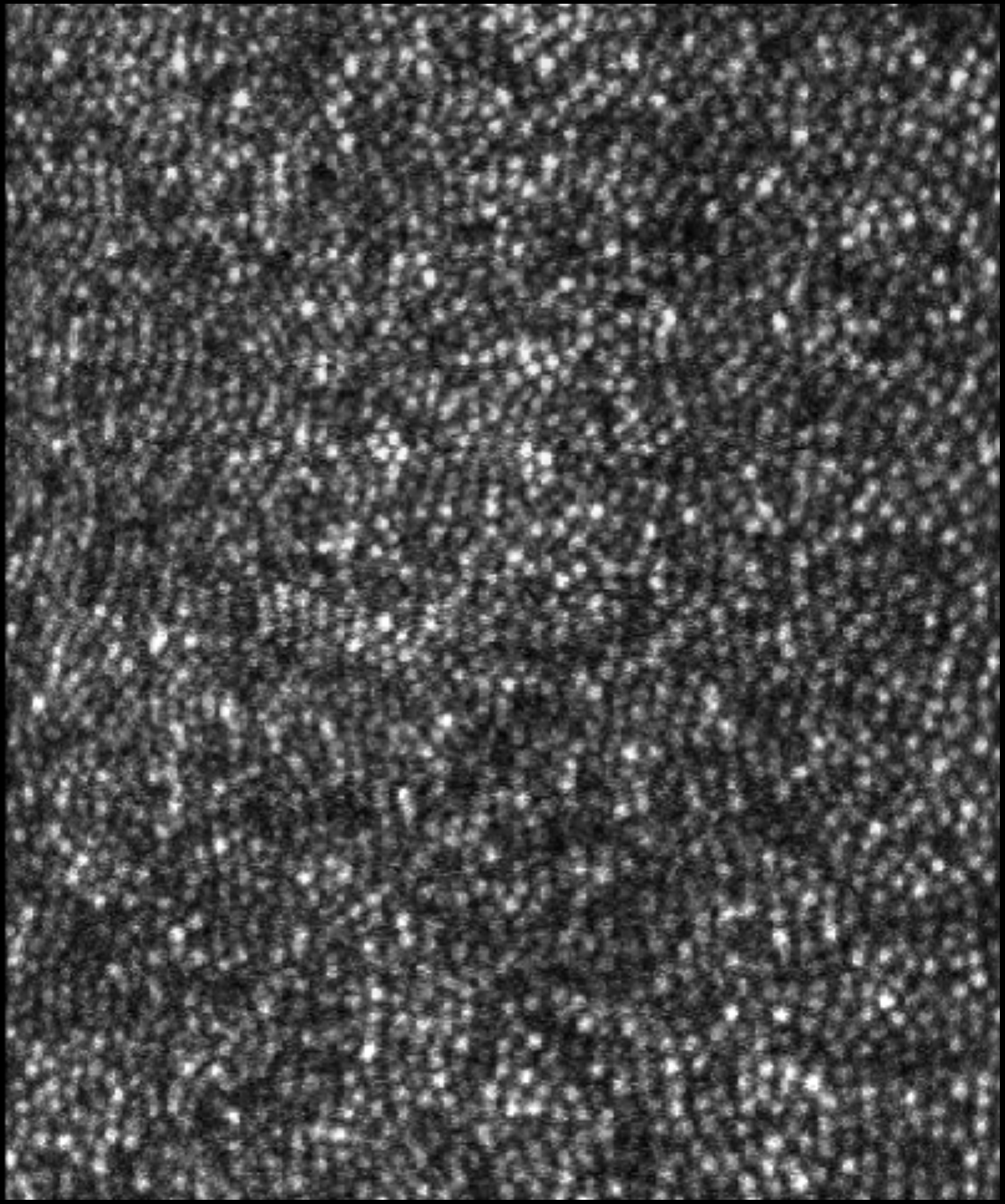
# Three questions

1. How do we see in the presence of fixational eye movements?

2. What is the optimal spatial layout of the image sampling array?

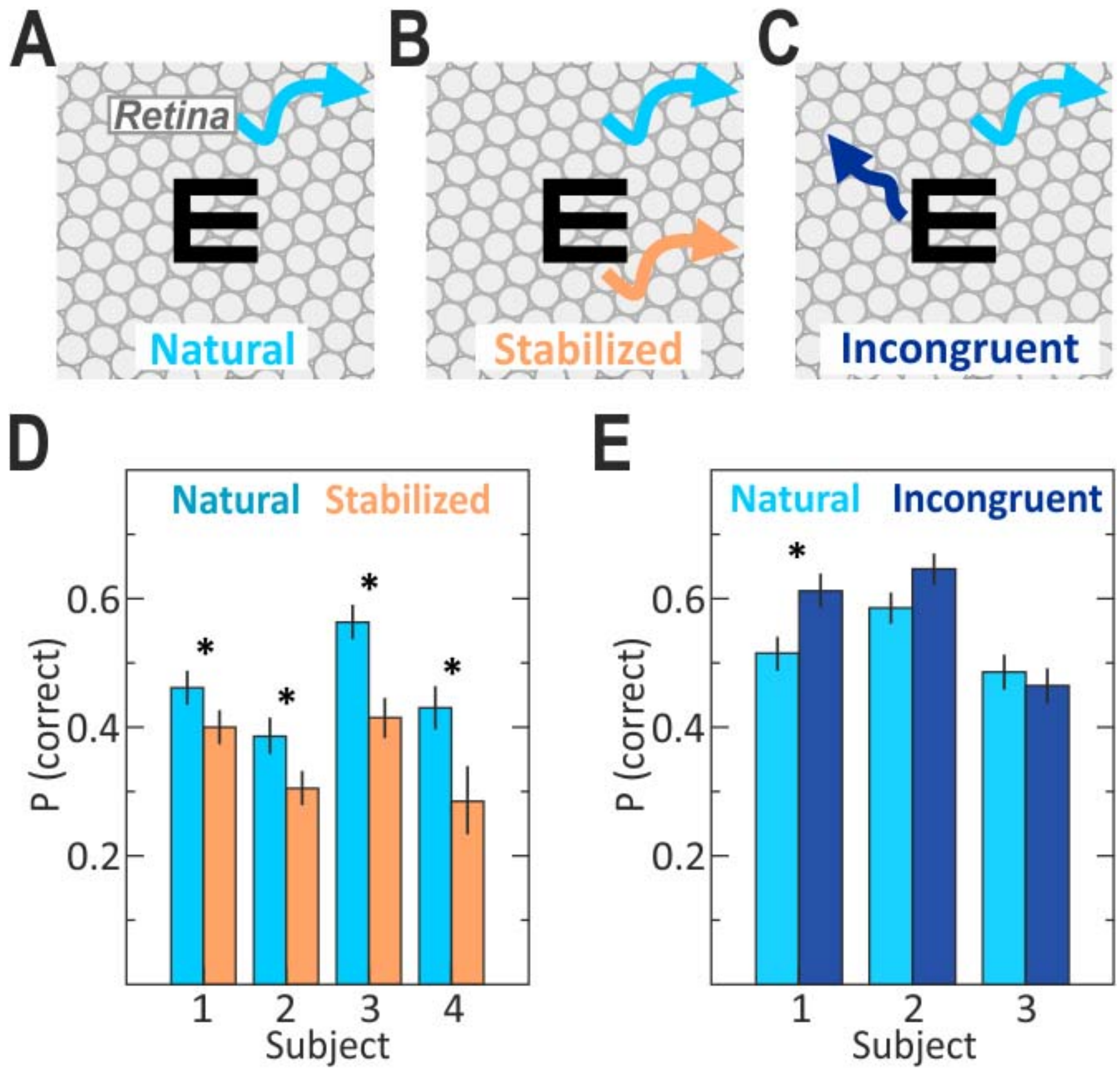
3. How is information integrated across multiple fixations?

# Fixational eye movements (drift)



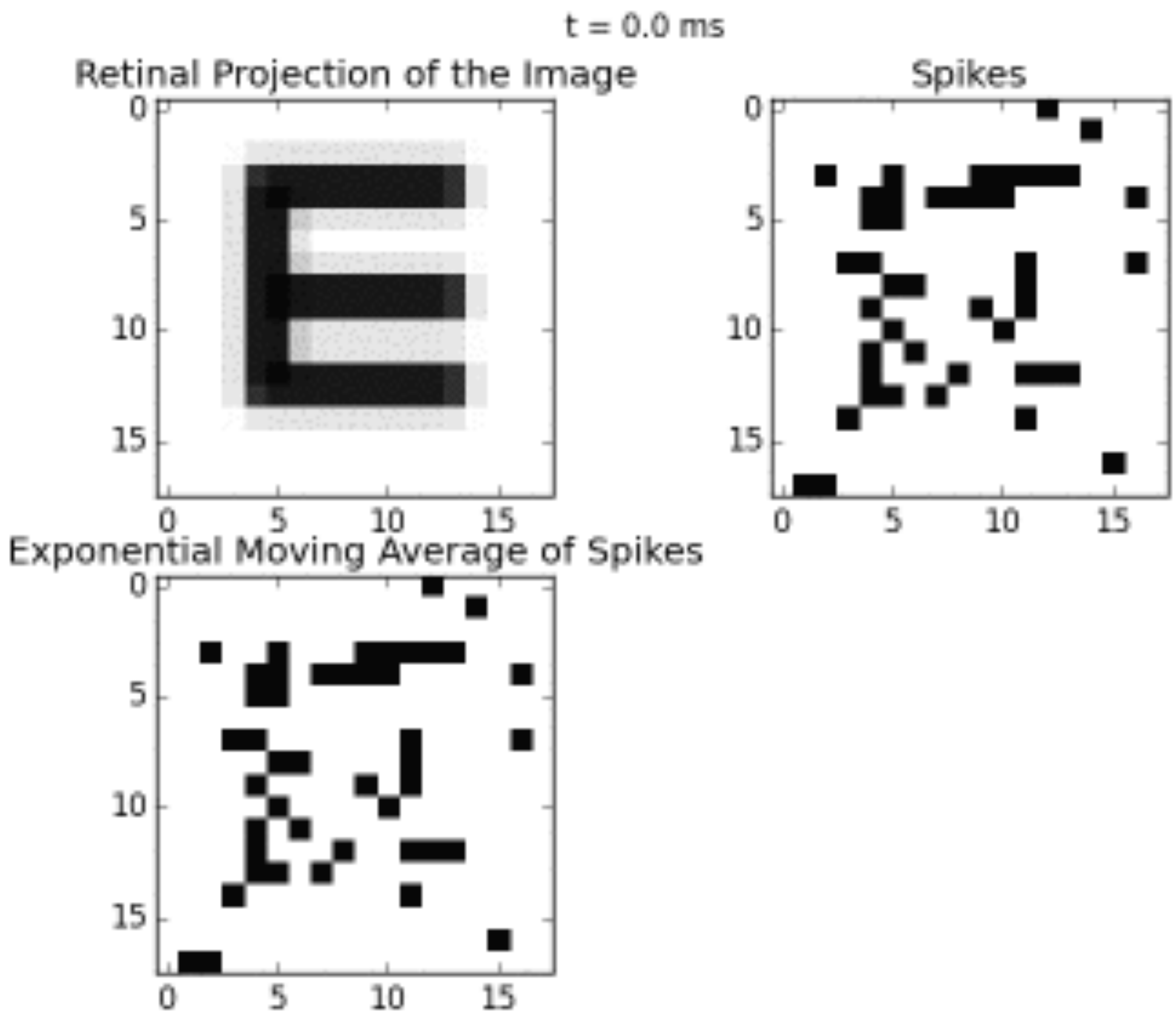
(from Austin Roorda, UC Berkeley)

# Retinal image motion helps pattern discrimination




Ratnam, K., Domdei, N., Harmening, W. M., & Roorda, A. (2017). Benefits of retinal image motion at the limits of spatial vision. *Journal of Vision, 17*, 1–11.

# Simple averaging is not sufficient



# The problem

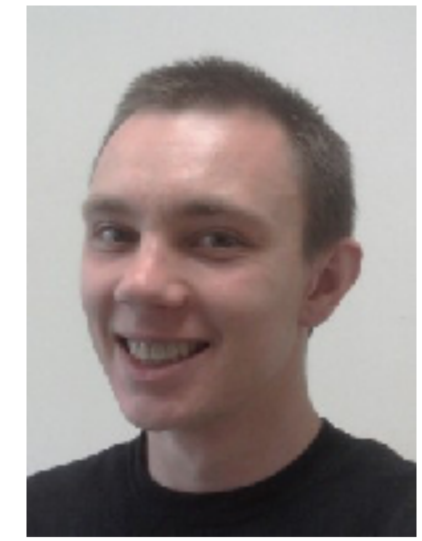
$$I(\vec{x}, t) = S(\vec{x} - \Delta\vec{x}(t)) + \epsilon(\vec{x}, t)$$

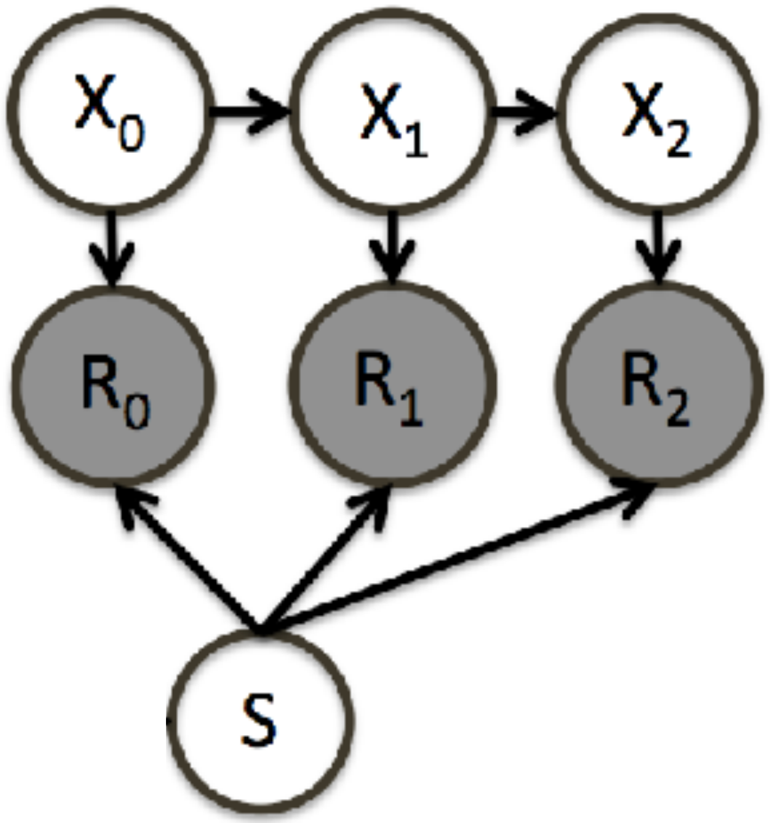
$$\hat{\Delta}\vec{x}(t) = \arg\min_{\Delta\vec{x}(t)} |I(\vec{x}, t) - S(\vec{x} - \Delta\vec{x}(t))|^2$$

$$\hat{S}(\vec{x}) = \int I(\vec{x} + \Delta\vec{x}(t))\, dt$$

# Graphical model for separating form and motion

(Alex Anderson, Ph.D. thesis)






**Eye position**

**Spikes**

(from LGN afferents)

**Pattern**

$$\begin{aligned}\hat{S} &= \arg\max_S \log P(R|S) \\ &= \arg\max_S \log \sum_X P(R|X,S)\, P(X)\end{aligned}$$

# Alternating estimation of pattern (S) and position (X)

Given current estimate of position (X), update S

$$\hat{S}^{t+1} = \arg\max_S \sum_{t'=0}^{t} \sum_j \langle \log P(R_{j,t'} | X_{t'}, S) \rangle_{P(X_{t'} | S^t, R_{0:t})}$$

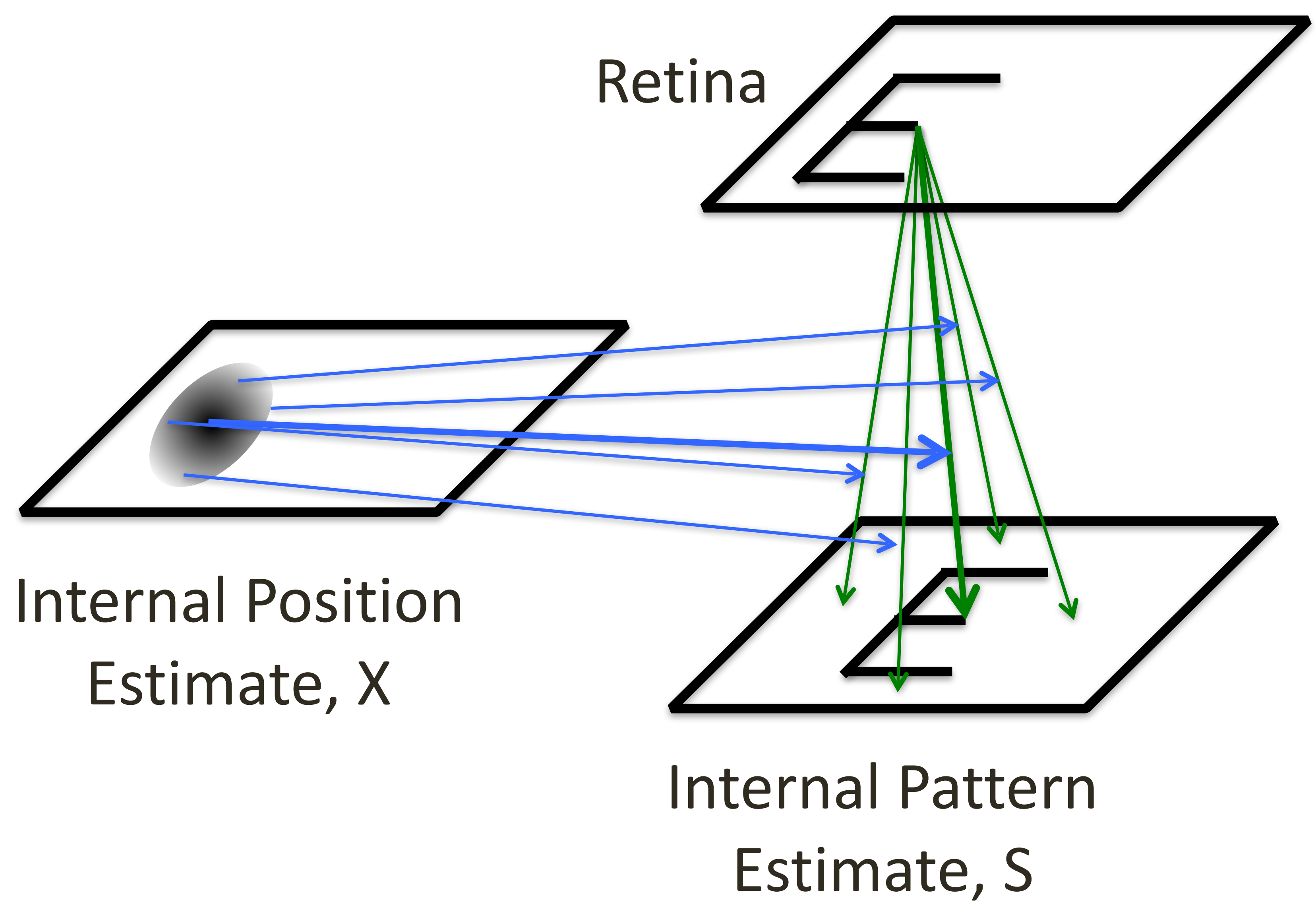
$$\log P(R_{j,t} | X_t, S) = R_{j,t} \, \log \lambda_j - \lambda_j \, dt$$

$$\log \lambda_j = \sum_{\vec{x}} g_j(\vec{x}) \, S(\vec{x} - \vec{X}_t)$$

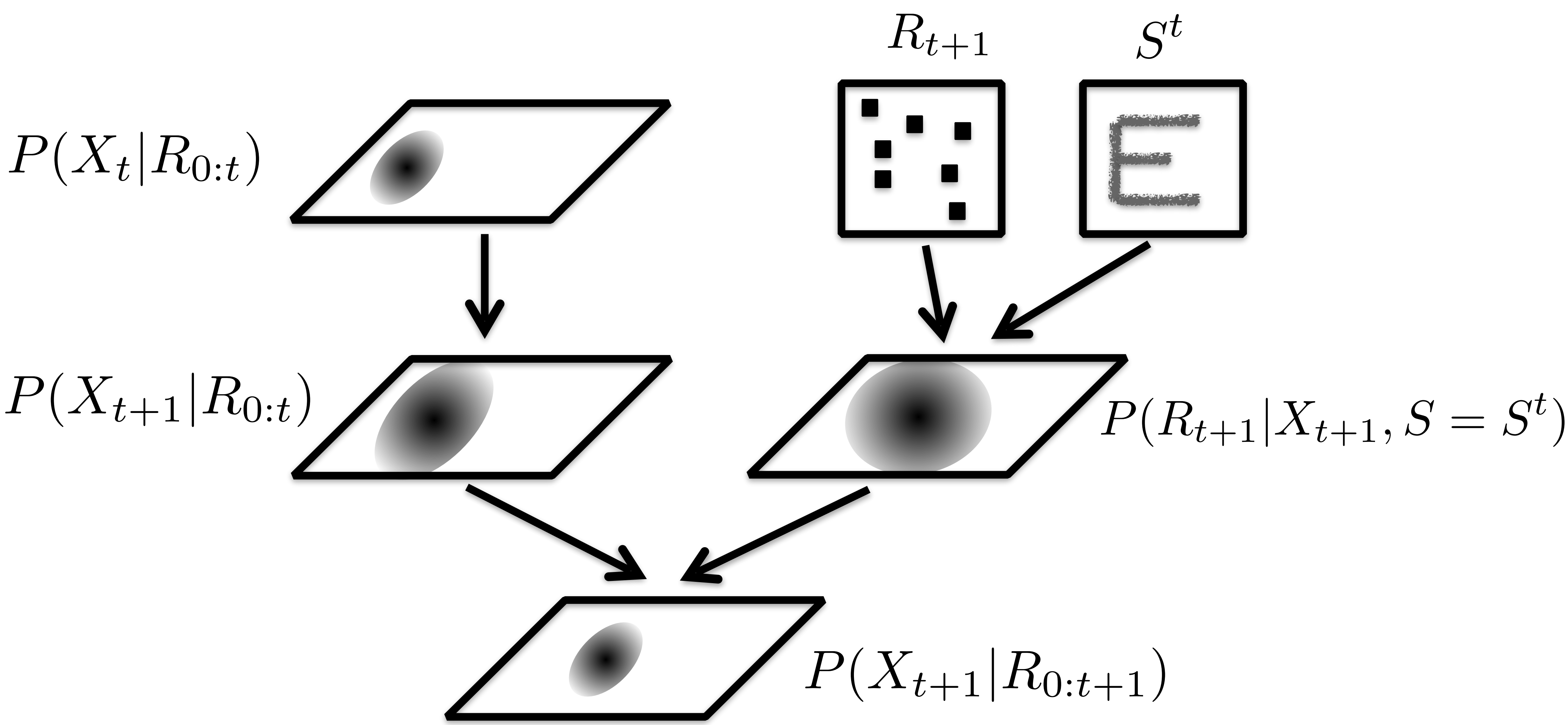
Given current estimate of pattern (S), update X

$$P(X_t | S^t, R_{0:t}) \propto P(R_t | X_t, S^t) \sum_{X_{t-1}} P(X_t | X_{t-1}) \, P(X_{t-1} | S^{t-1}, R_{0:t-1})$$

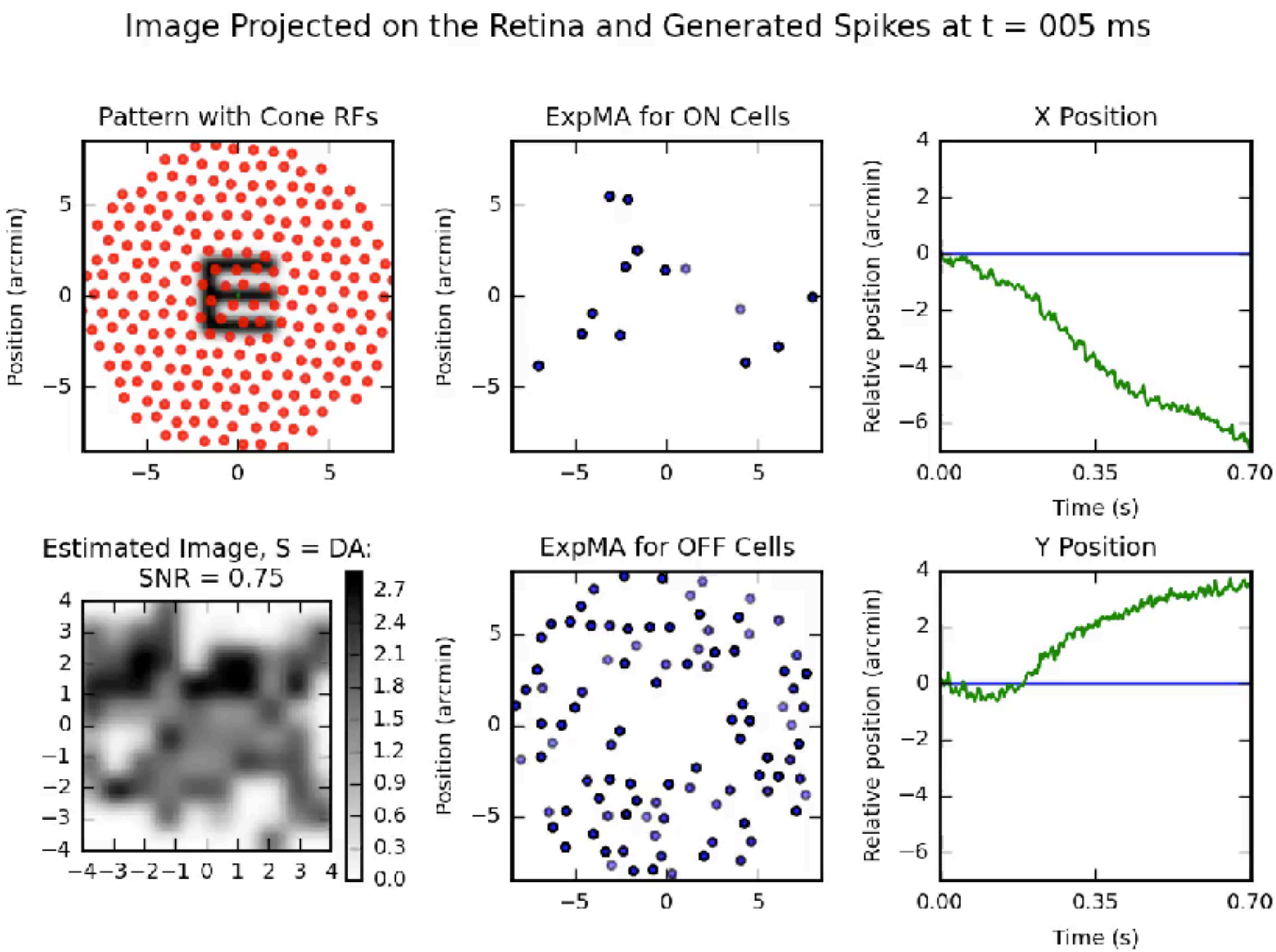
# Given current estimate of position (X), update S



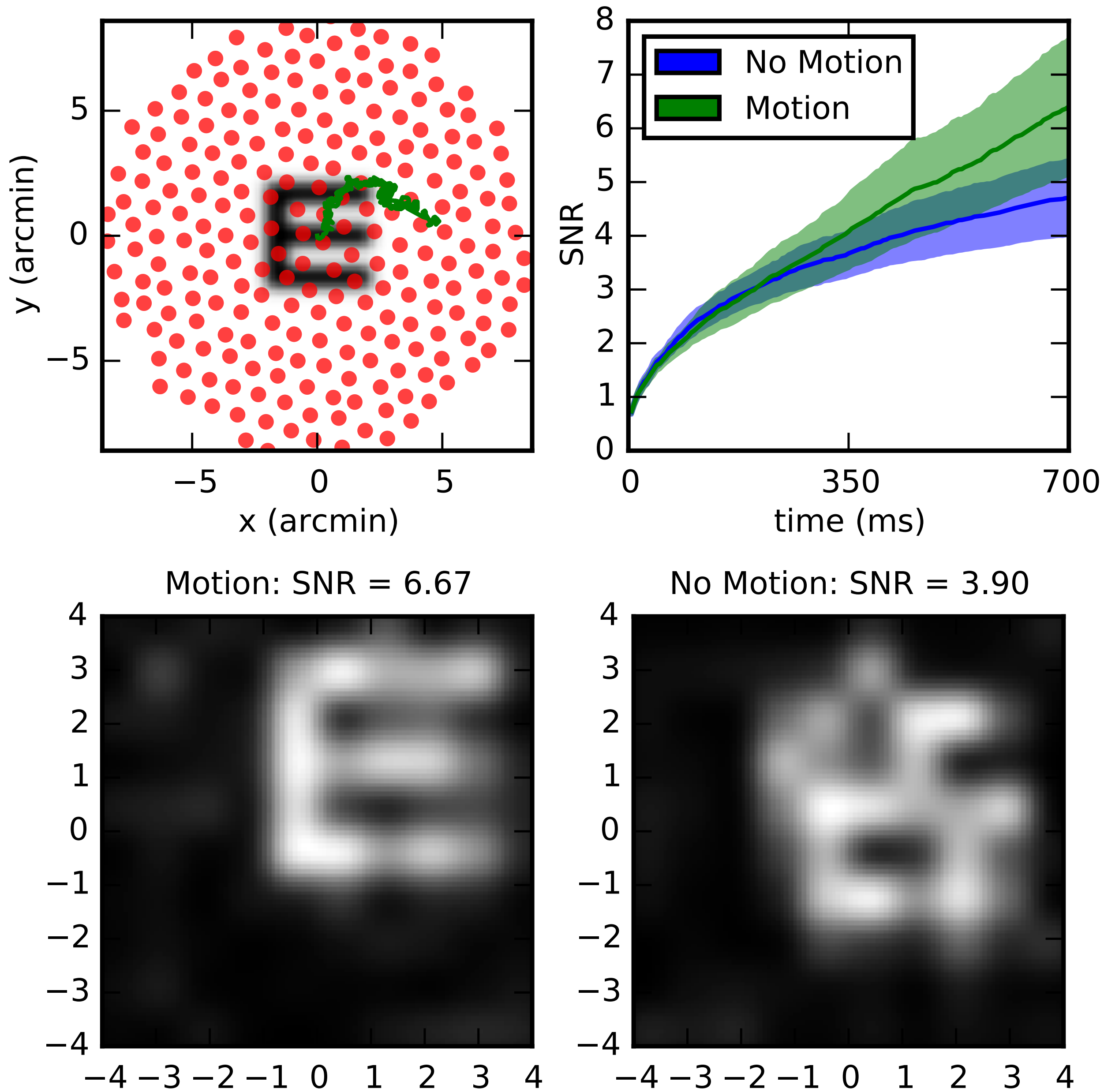
# Given current estimate of pattern (S), update X



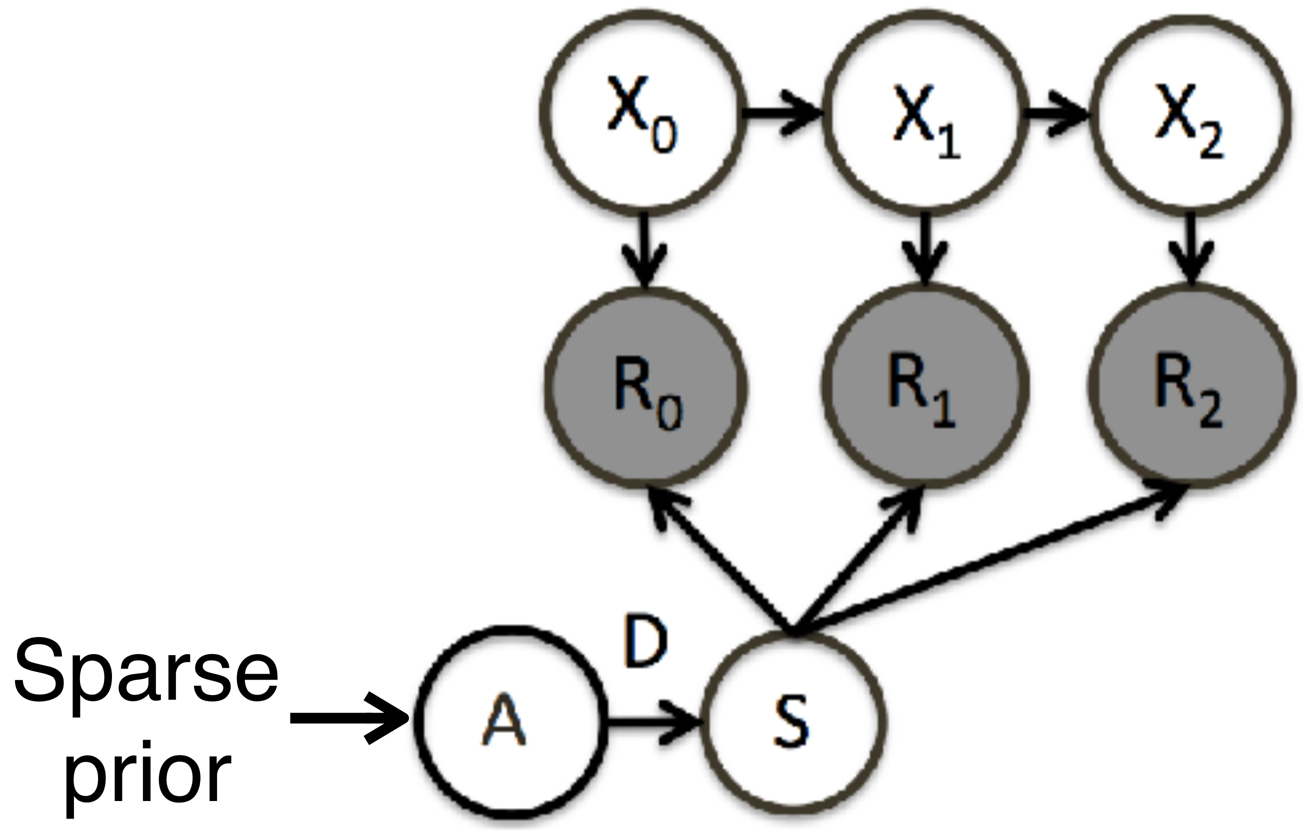
# Joint estimation of pattern (S) and position (X)



# Motion helps estimation of pattern *S*



# Including a prior over *S*




**Eye position**

**Spikes**
(from LGN afferents)

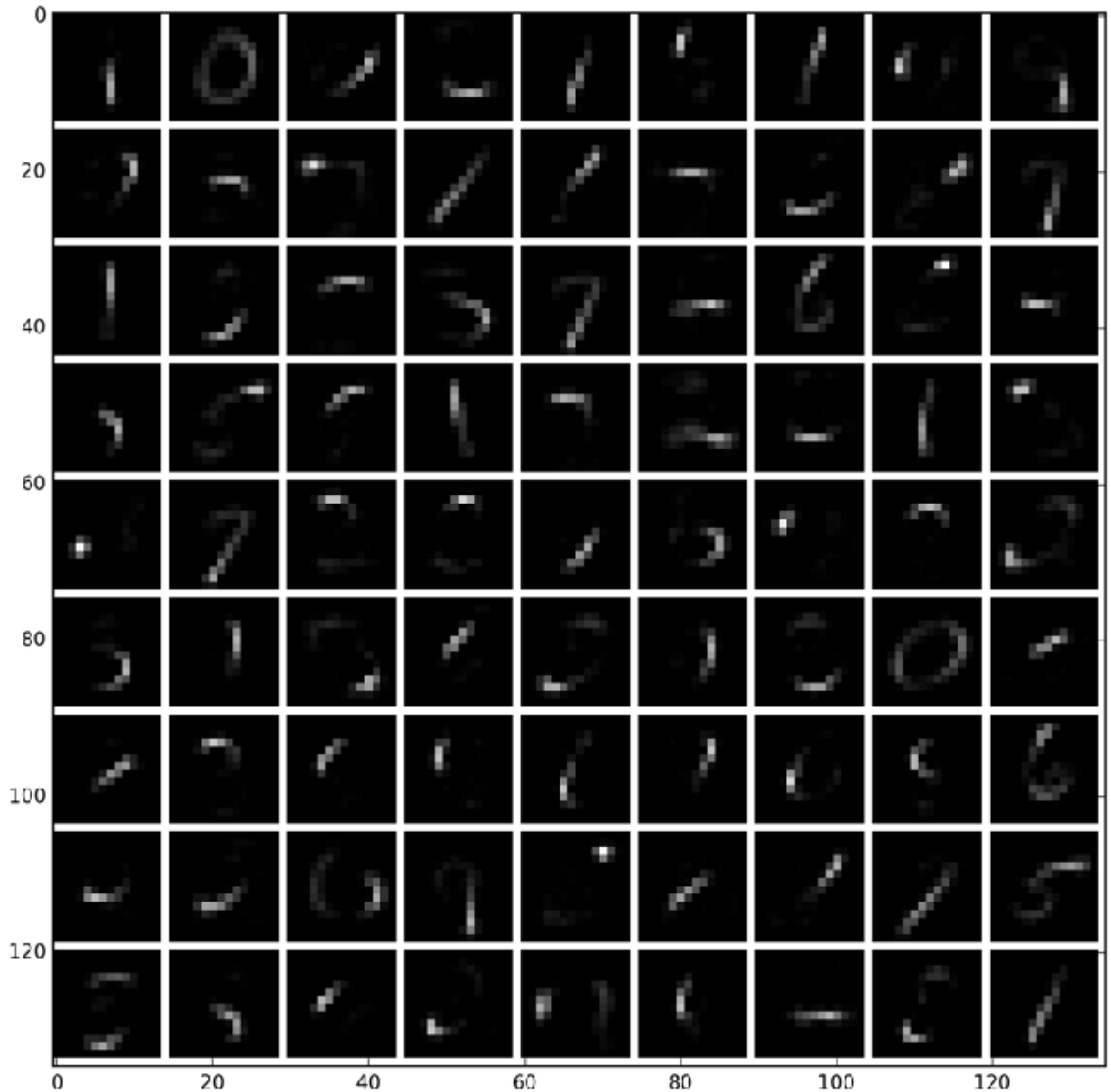
**Pattern**

$$S = D\,A$$

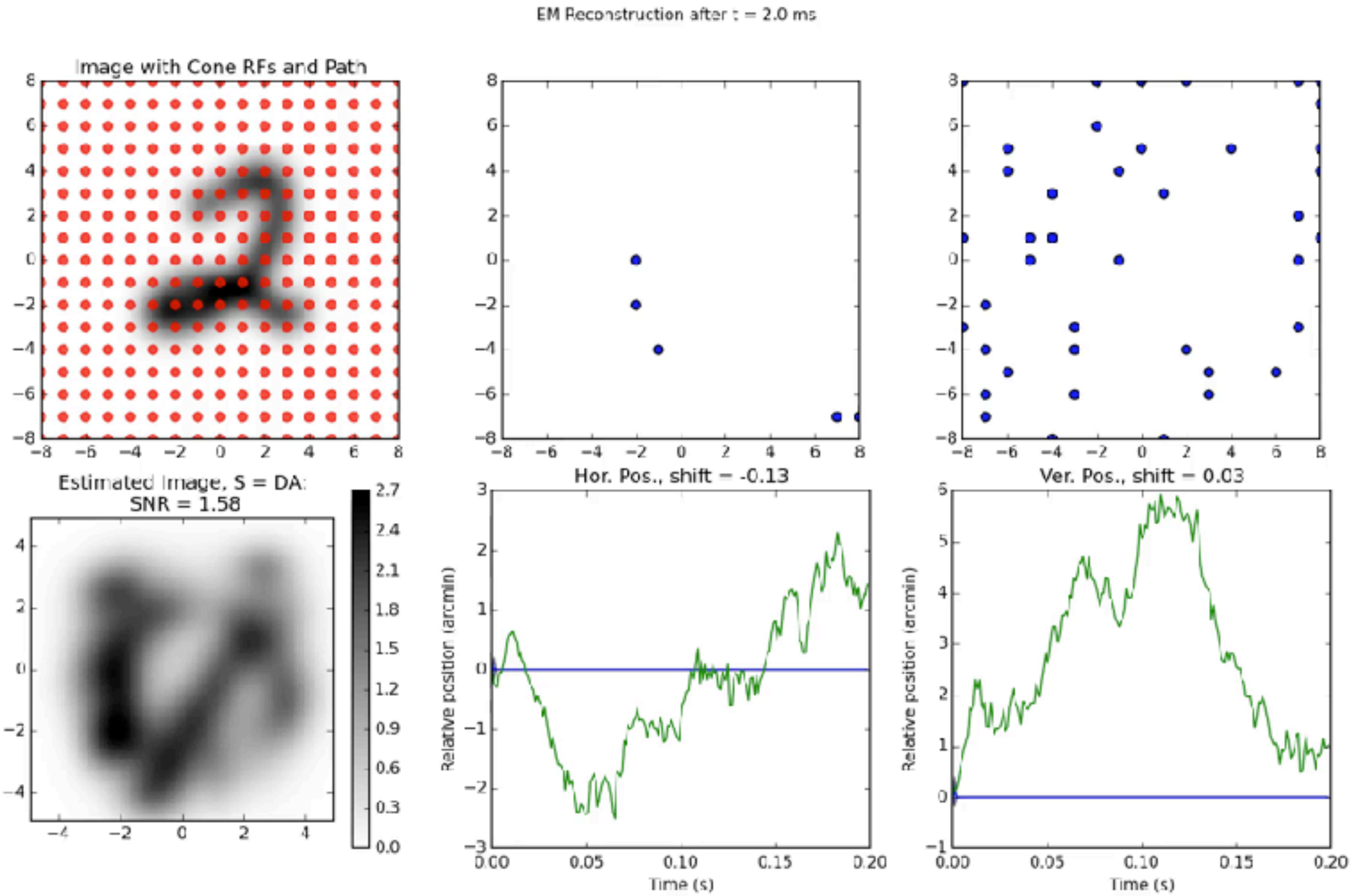
$$\hat{A} = \arg\max_{A}\ \log P(R|A) + \boxed{\log P(A)}$$

sparse

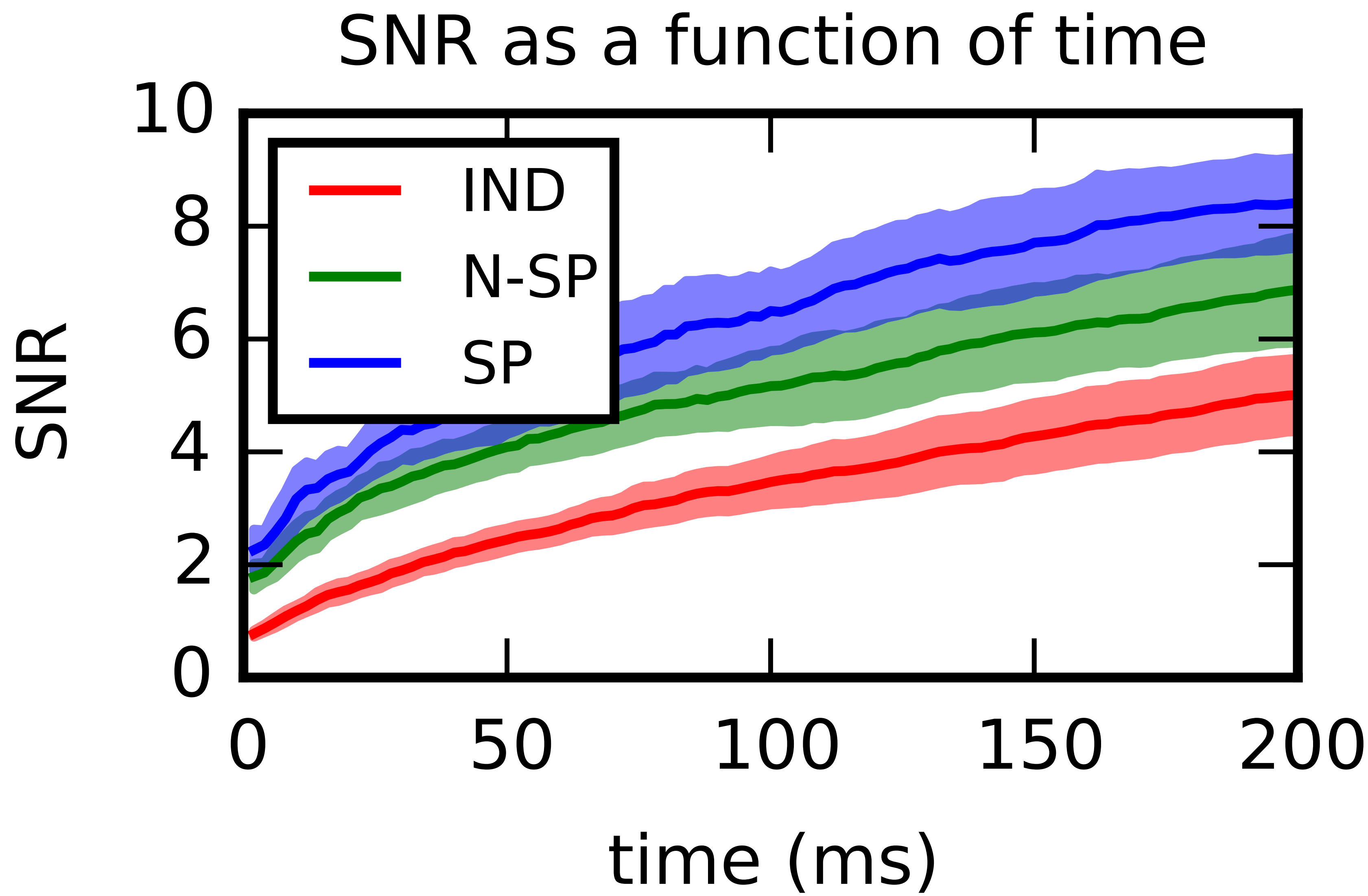
# Learned dictionary $D$



# Prior over *S* improves inference



# Prior over *S* improves inference



# Three questions

1. How do we see in the presence of fixational eye movements?

2. What is the optimal spatial layout of the image sampling array?

3. How is information integrated across multiple fixations?

# What is this?

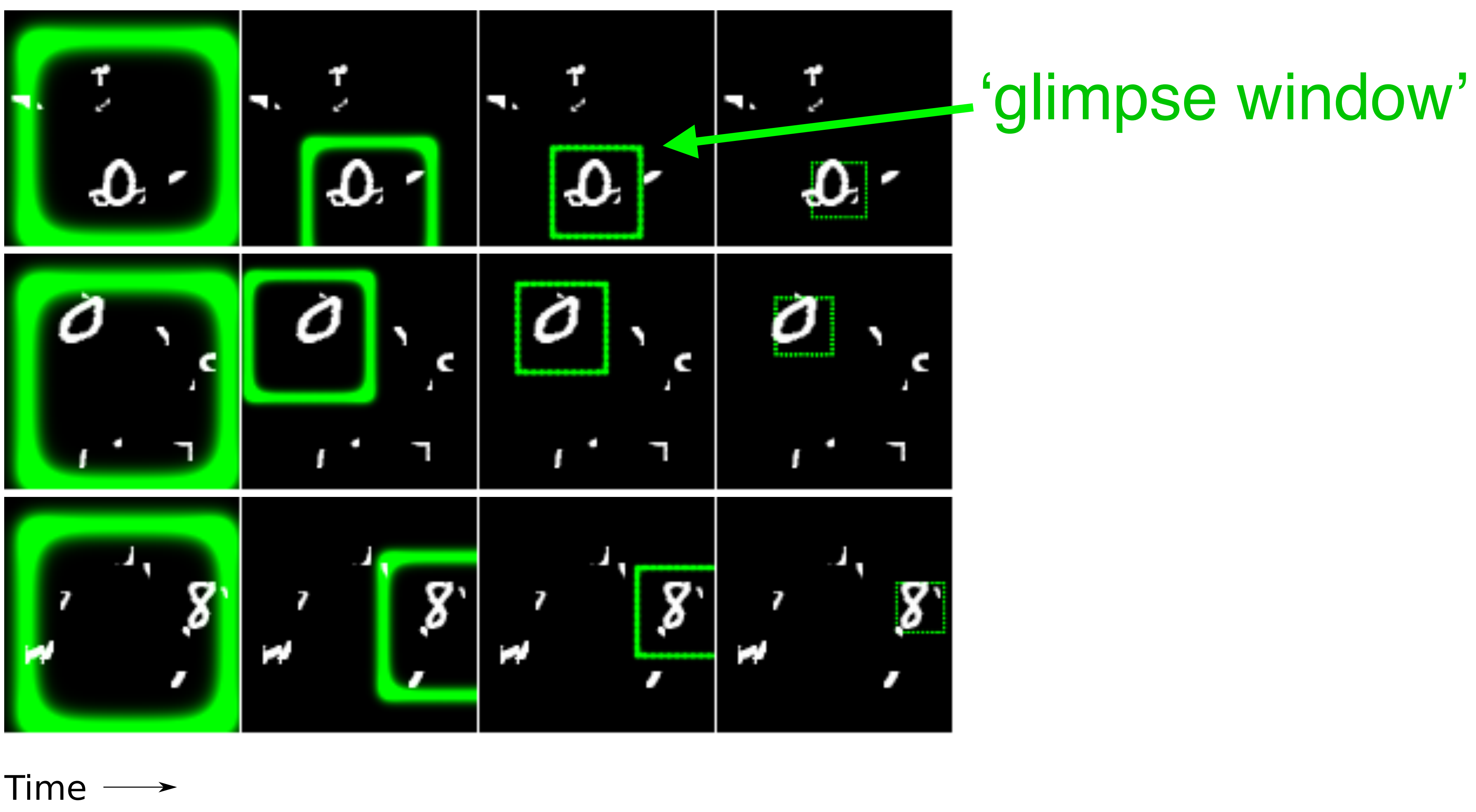
Correct label: Pomeranian

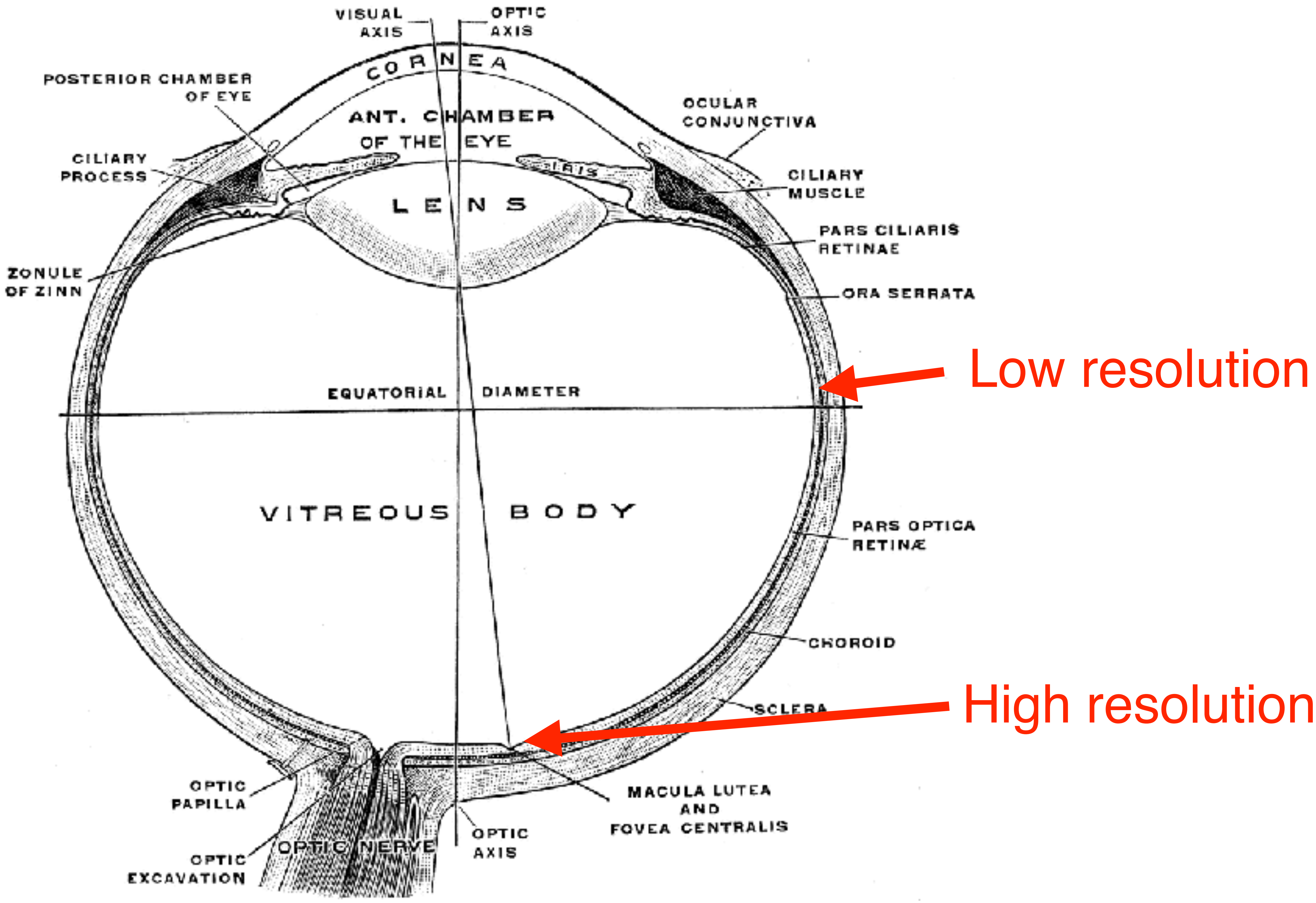
# What is this?

Correct label: Afghan hound

# DRAW: A Recurrent Neural Network For Image Generation

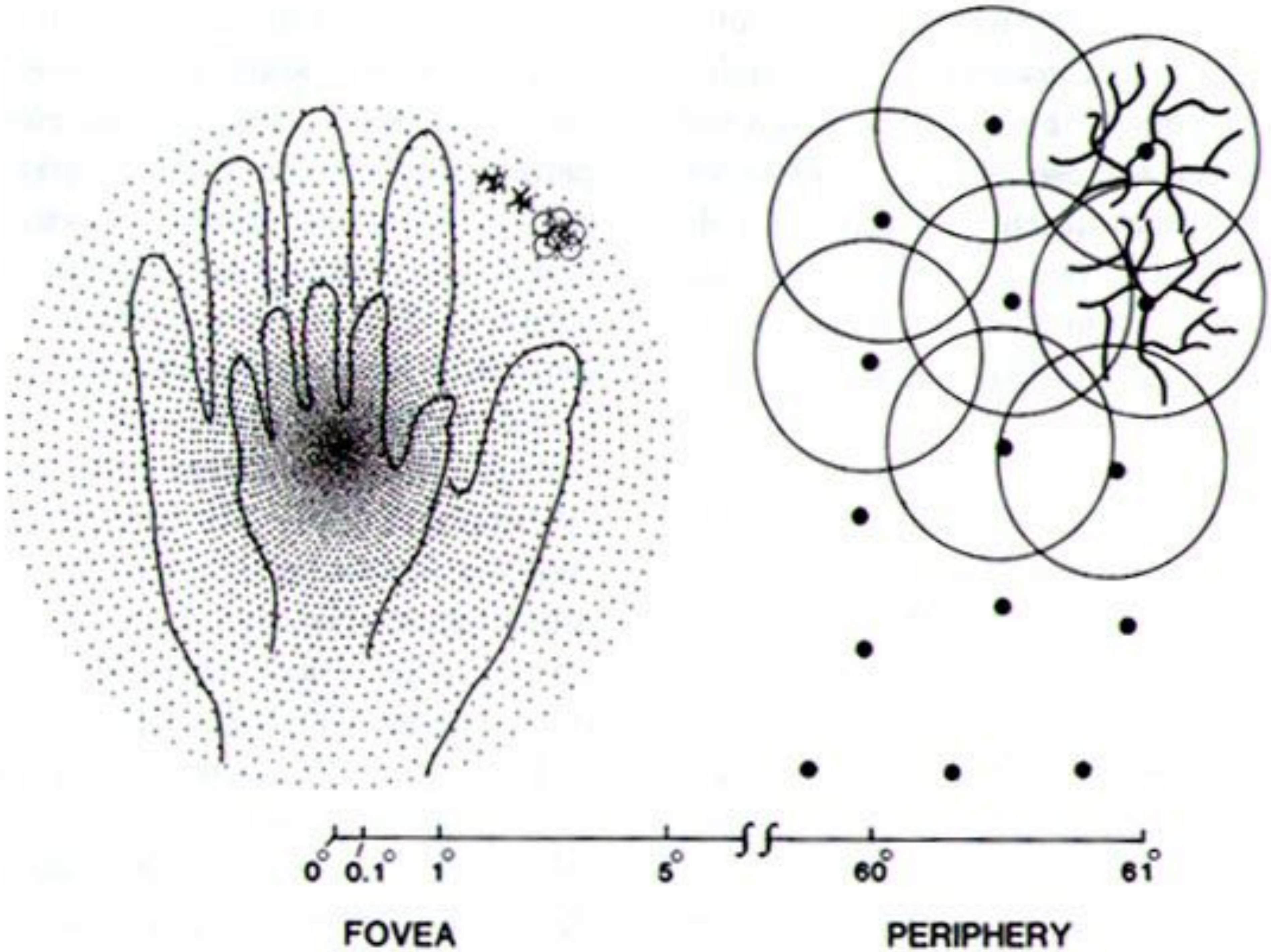
**Karol Gregor** KAROLG@GOOGLE.COM
**Ivo Danihelka** DANIHELKA@GOOGLE.COM
**Alex Graves** GRAVESA@GOOGLE.COM
**Danilo Jimenez Rezende** DANILOR@GOOGLE.COM
**Daan Wierstra** WIERSTRA@GOOGLE.COM
Google DeepMind





VISUAL AXIS
OPTIC AXIS
CORNEA
POSTERIOR CHAMBER OF EYE
ANT. CHAMBER OF THE EYE
OCULAR CONJUNCTIVA
CILIARY PROCESS
IRIS
CILIARY MUSCLE
LENS
PARS CILIARIS RETINAE
ZONULE OF ZINN
ORA SERRATA
Low resolution
EQUATORIAL DIAMETER
VITREOUS BODY
PARS OPTICA RETINÆ
CHOROID
High resolution
SCLERA
MACULA LUTEA AND FOVEA CENTRALIS
OPTIC PAPILLA
OPTIC NERVE
OPTIC AXIS
OPTIC EXCAVATION

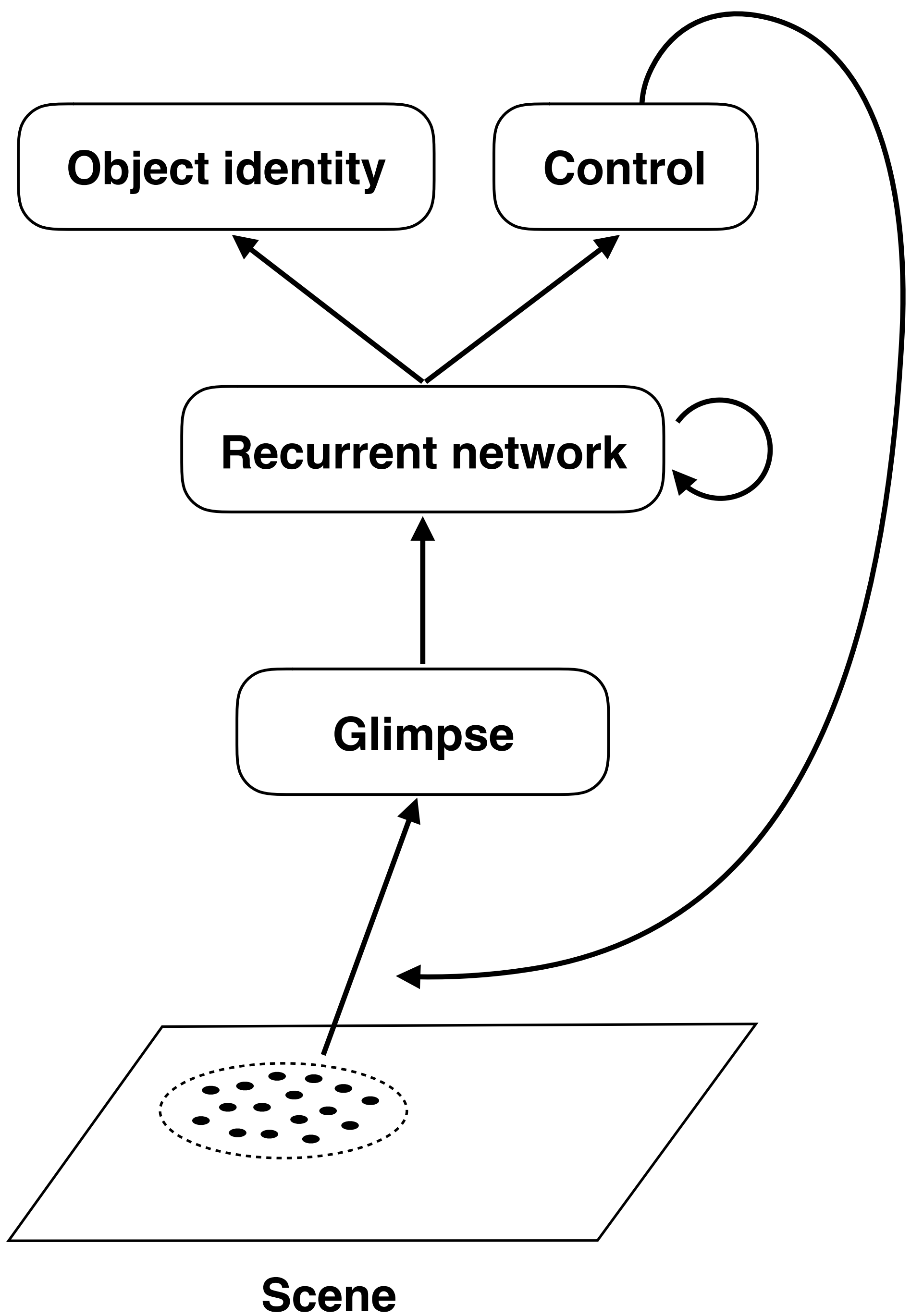
# Retinal ganglion cell sampling array
(shown at one dot for every 20 ganglion cells)



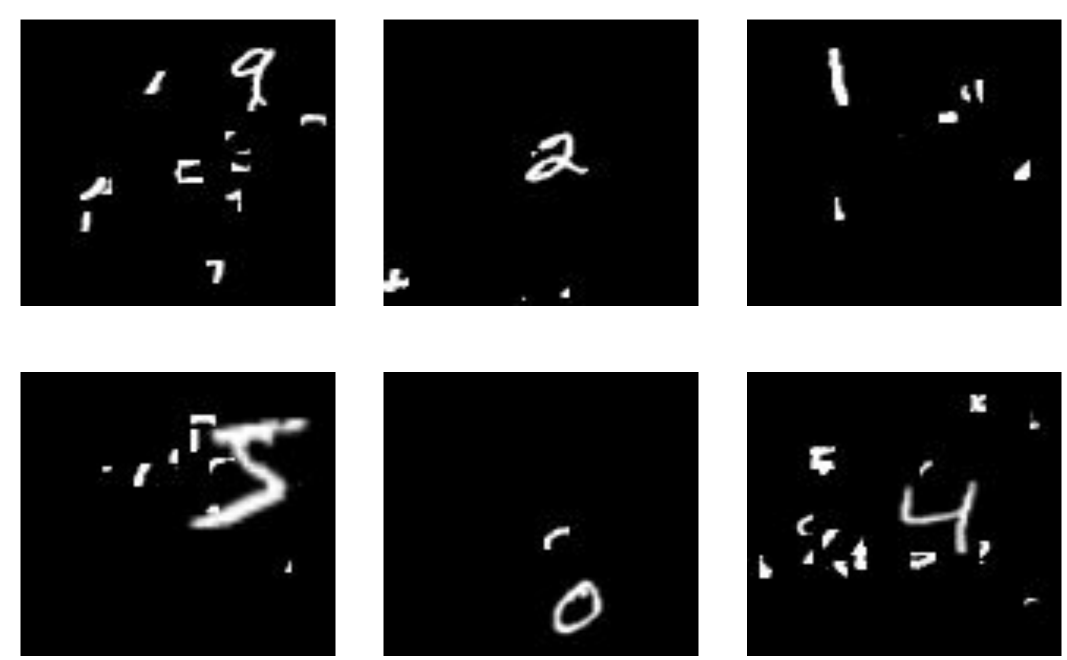

$$\Delta \approx .01(|E| + 1)$$

(from Anderson & Van Essen, 1995)

# Learning the glimpse window sampling array
(Cheung, Weiss & Olshausen, 2017)

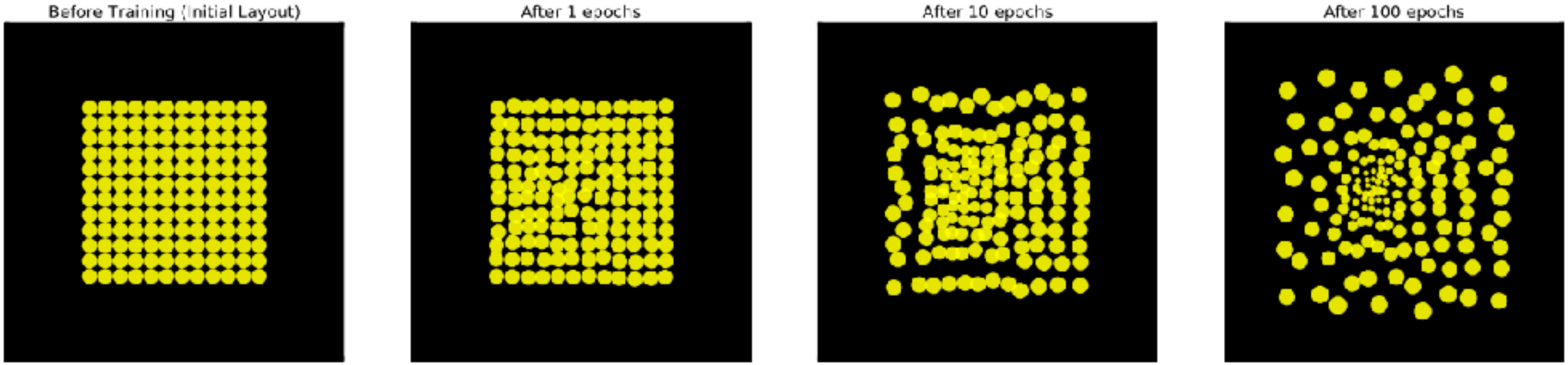



- Network is trained to correctly classify the digit in the scene.
- To do this it must find a digit and move its glimpse window to that location.

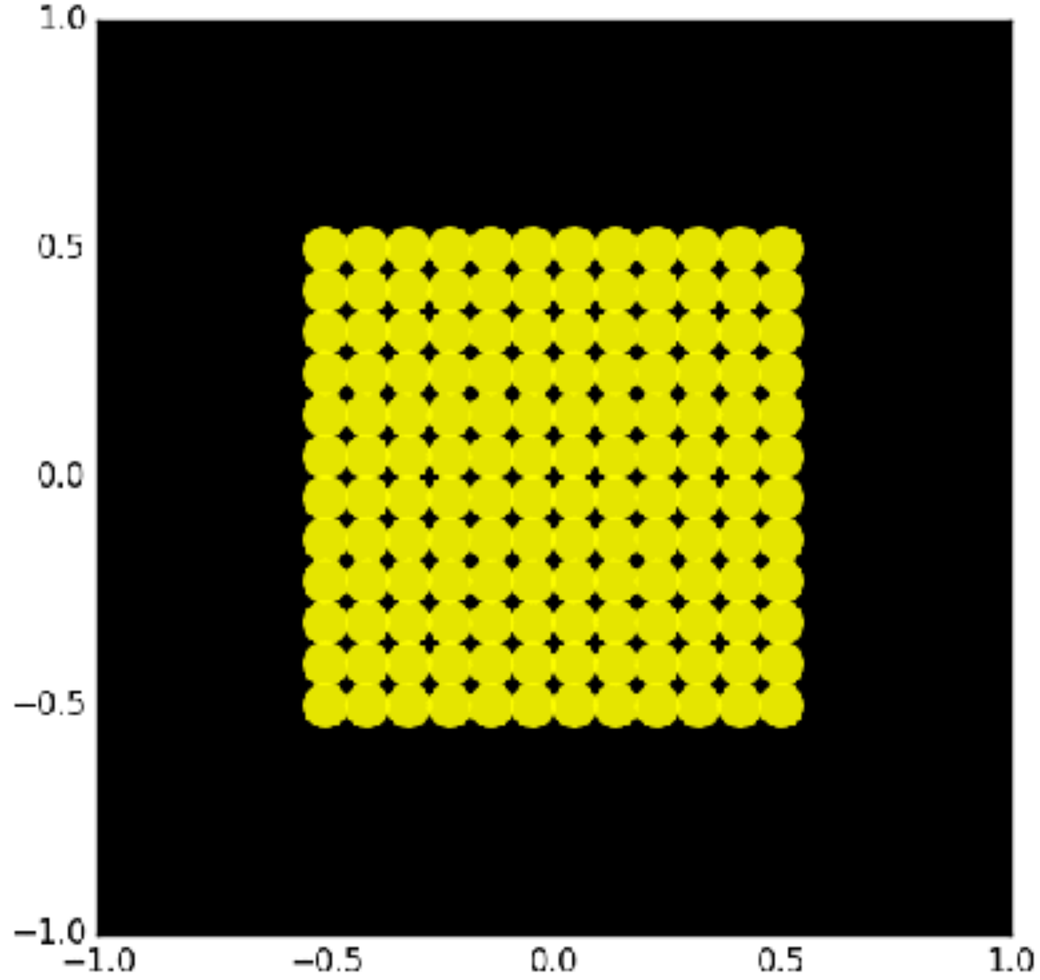


Example MNIST scenes

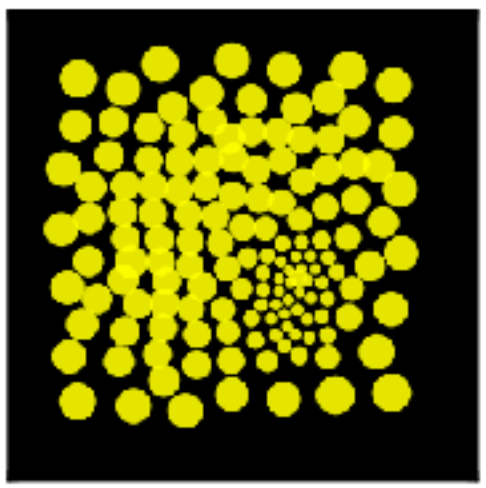
# Evolution of the sampling array during training



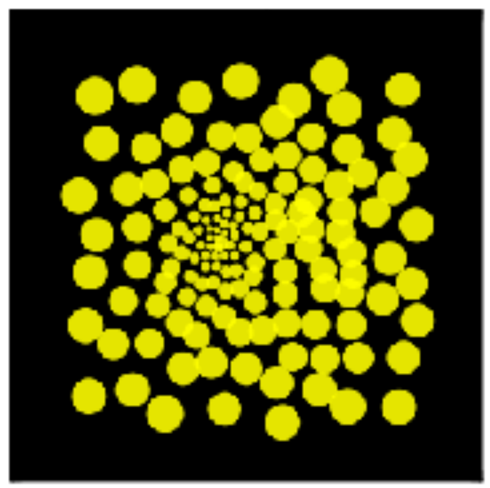
# Evolution of the sampling array during training



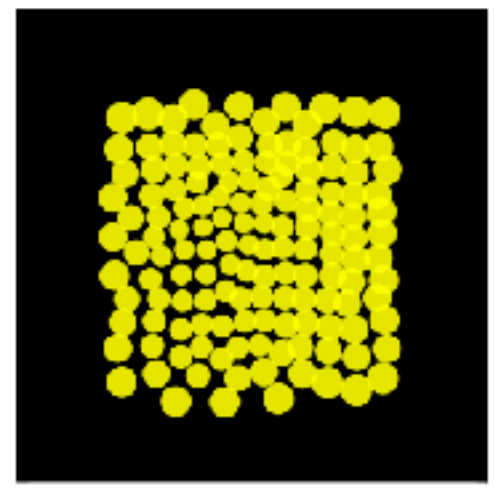
# Learned sampling arrays for different conditions



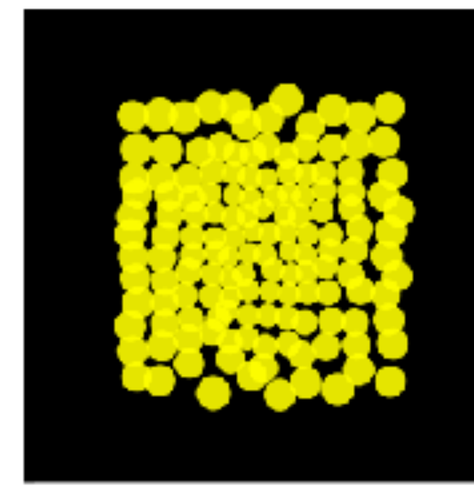
Translation only
(Dataset 1)



Translation only
(Dataset 2)



Translation & zoom
(Dataset 1)

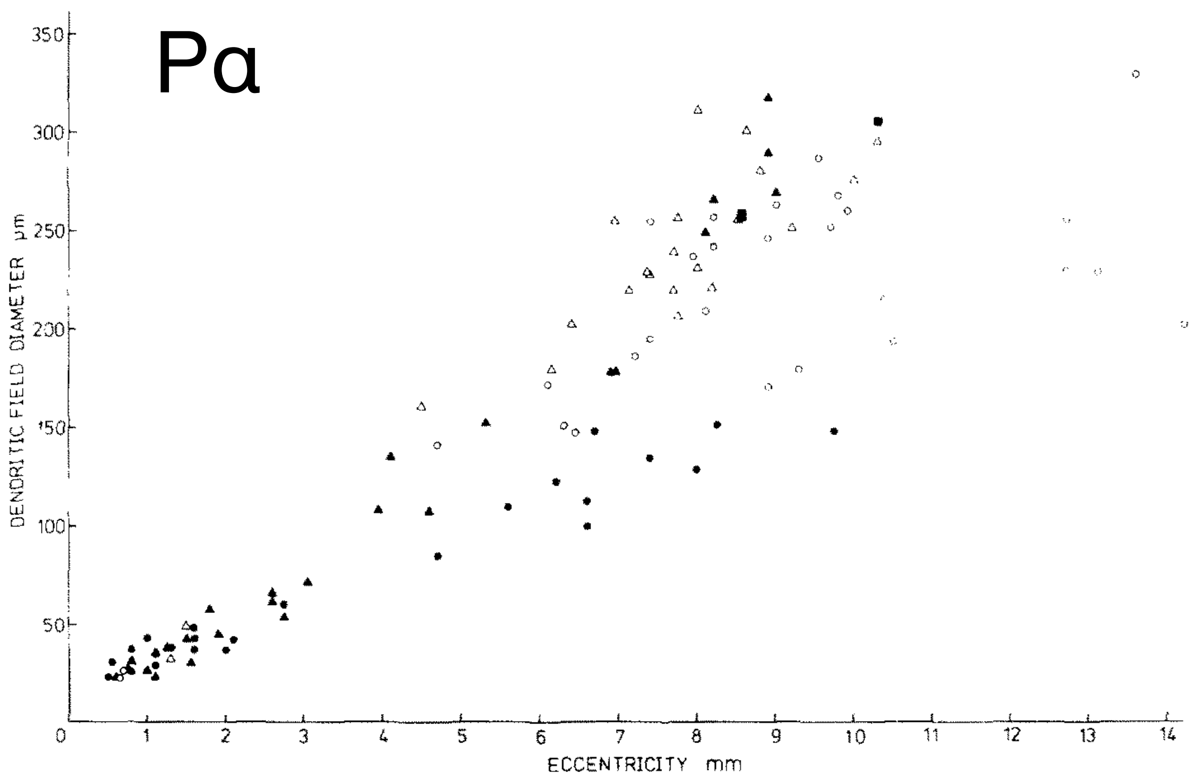


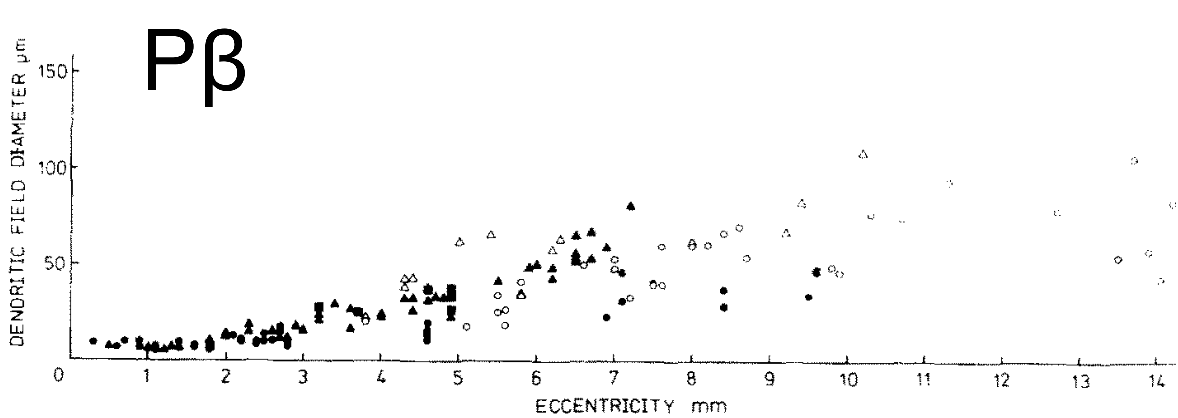
Translation & zoom
(Dataset 2)

# Comparison to primate retina

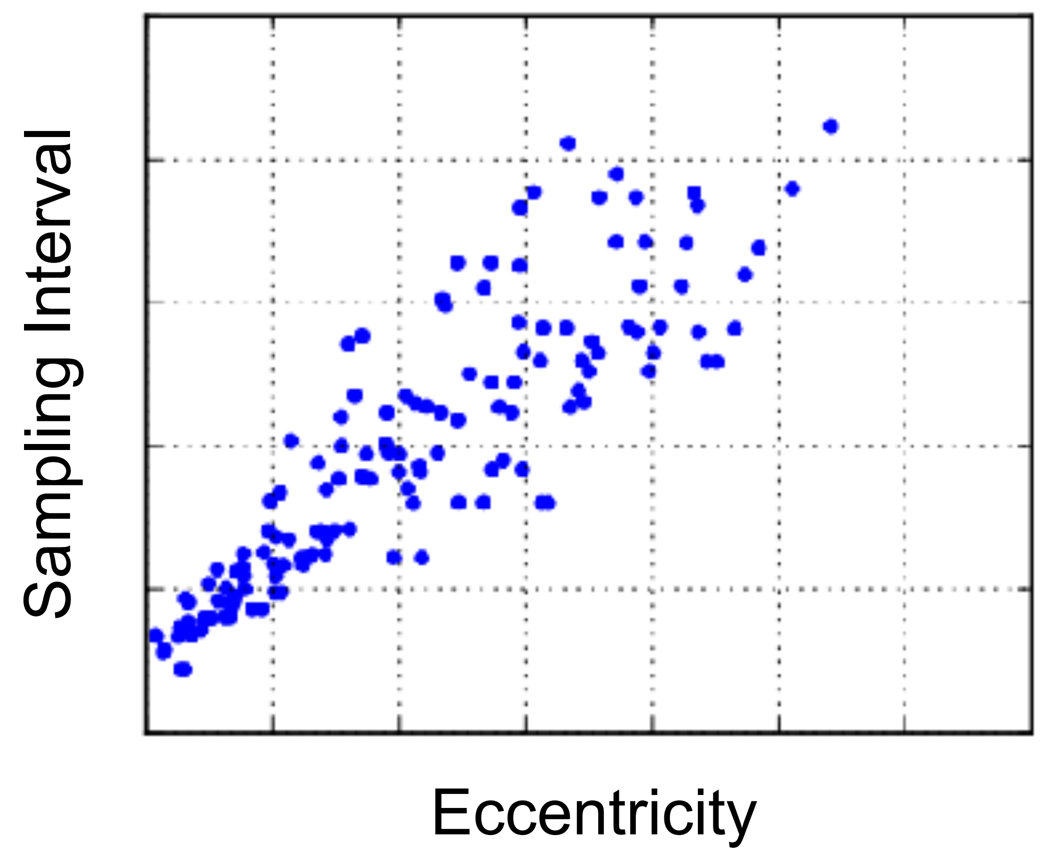
## Macaque Retina

(Perry, Oehler & Cowey, 1984)



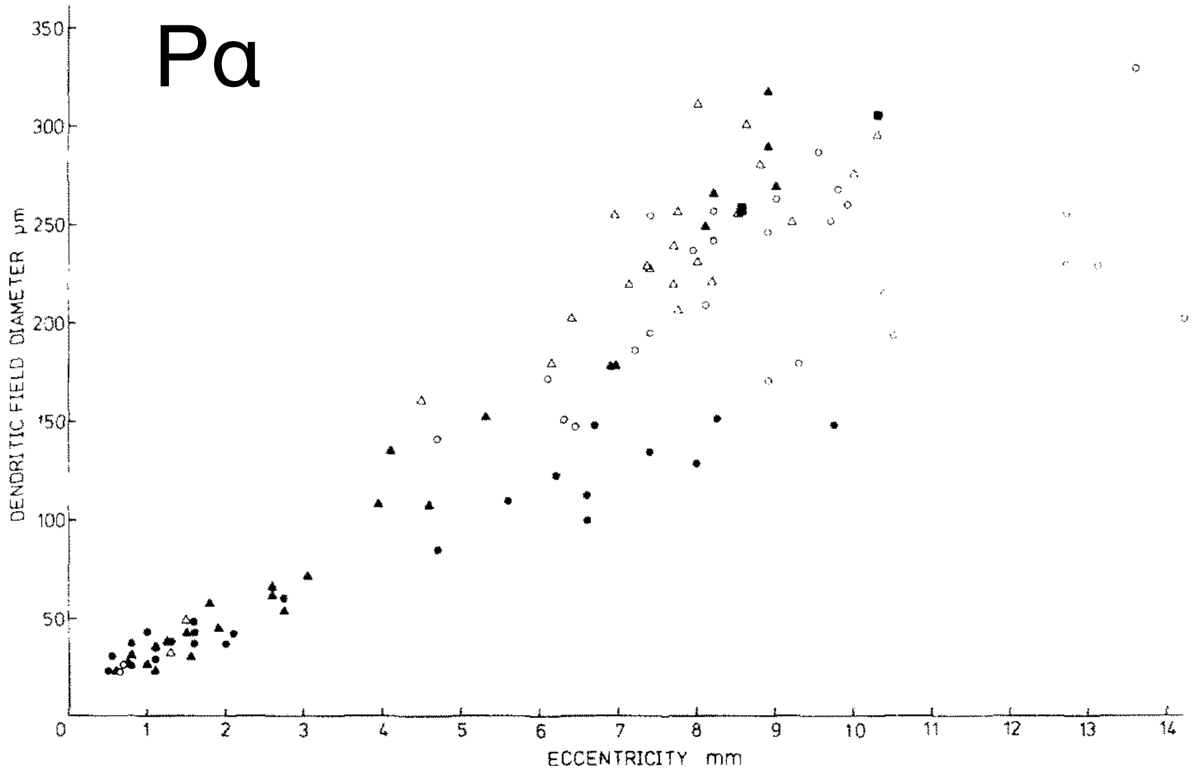

## Model

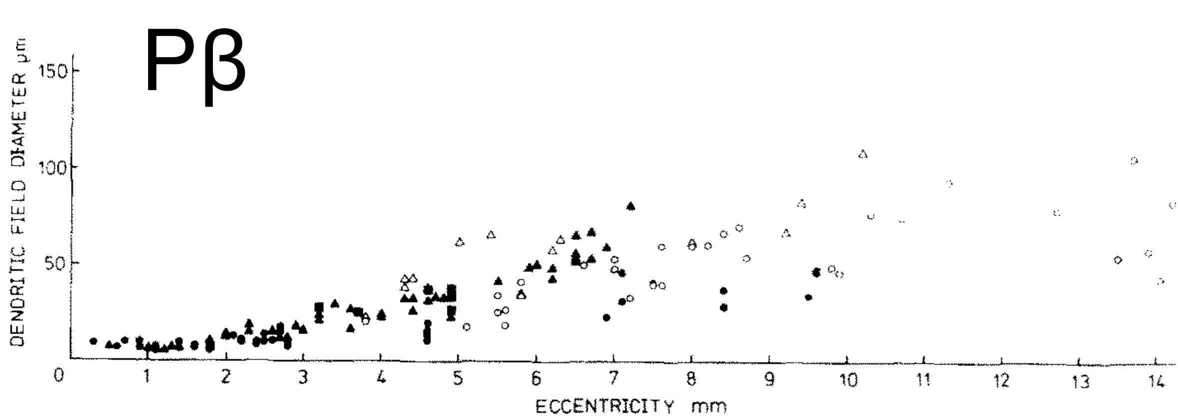


# Comparison to primate retina

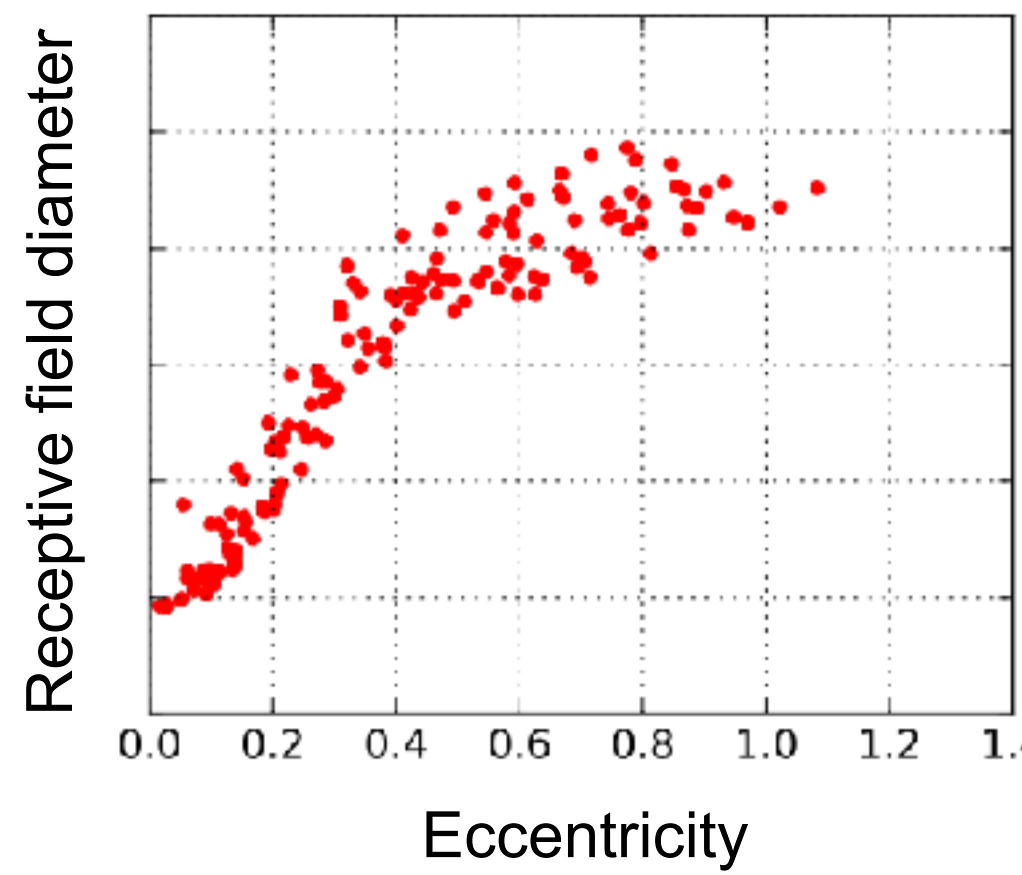
## Macaque Retina

(Perry, Oehler & Cowey, 1984)




## Model



# A FOVEATED RETINA-LIKE SENSOR USING CCD TECHNOLOGY

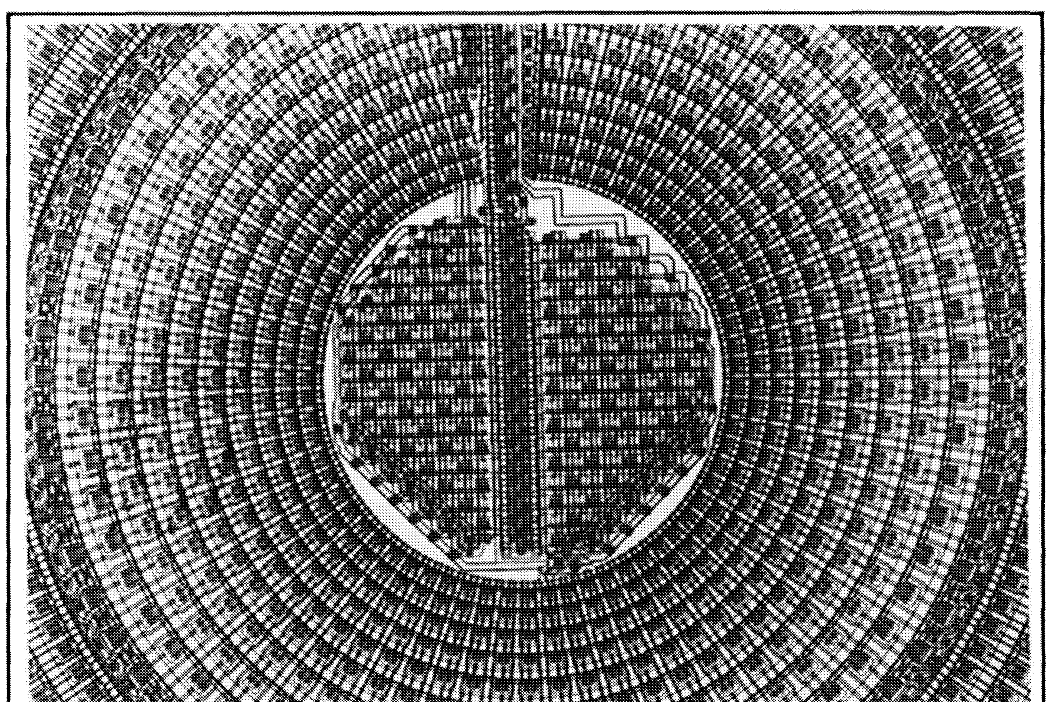
J. Van der Spiegel, G. Kreider
Univ. of Pennsylvania, Dept. of Electrical Engineering
Philadelphia, PA 19104-6390

C. Claeys, I. Debusschere
IMEC, Leuven, Belgium

G. Sandini
University of Genova, DIST, Genova, Italy

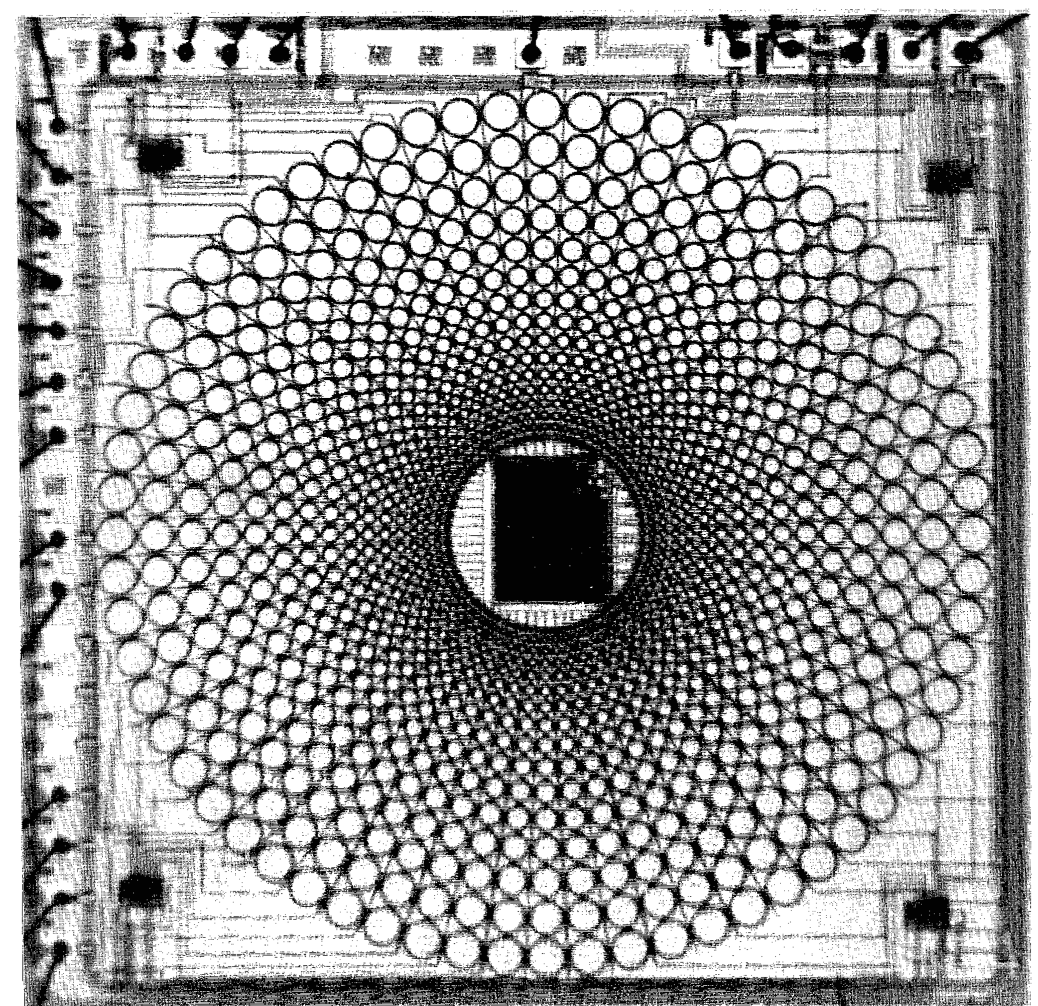
P. Dario, F. Fantini
Scuola Superiore S. Anna, Pisa, Italy

P. Bellutti, G. Soncini
IRST, Trento, Italy



# A Foveated Image Sensor in Standard CMOS Technology

Robert Wodnicki, Gordon W. Roberts, Martin D. Levine
Department of Electrical Engineering, McGill University,
Montréal, Québec, CANADA, H3A 2A7



# Three questions

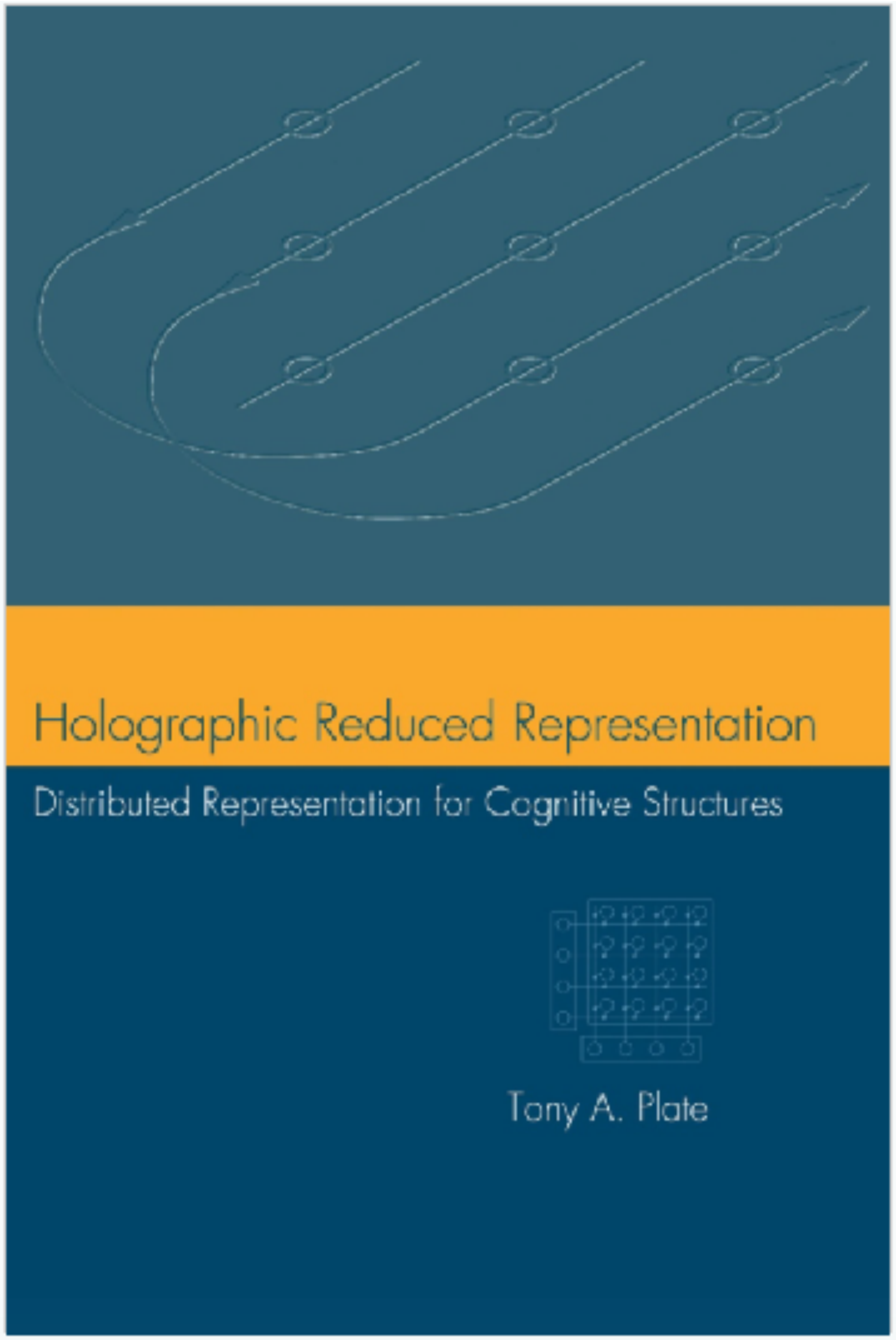
1. How do we see in the presence of fixational eye movements?
2. What is the optimal spatial layout of the image sampling array?
3. How is information integrated across multiple fixations?

In order to integrate visual information across fixations, two things must be encoded and combined at each fixation:

1) *position* of the glimpse window
2) *contents* of the glimpse window

We need to *bind* these two things together!

A scene may then be represented as a superposition of such bindings.




Cogn Comput (2009) 1:139–159
DOI 10.1007/s12559-009-9009-8

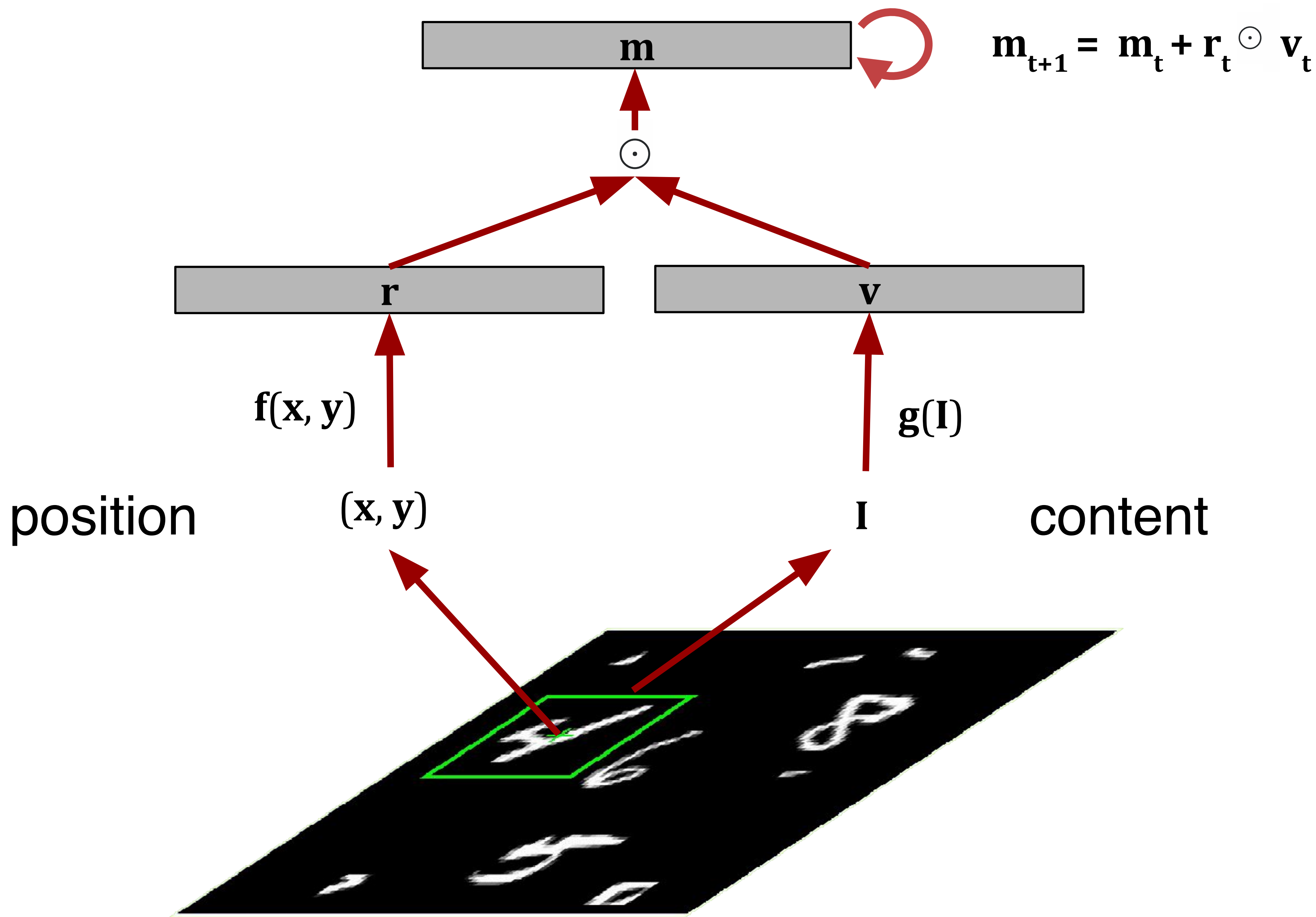
**Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors**

**Pentti Kanerva**

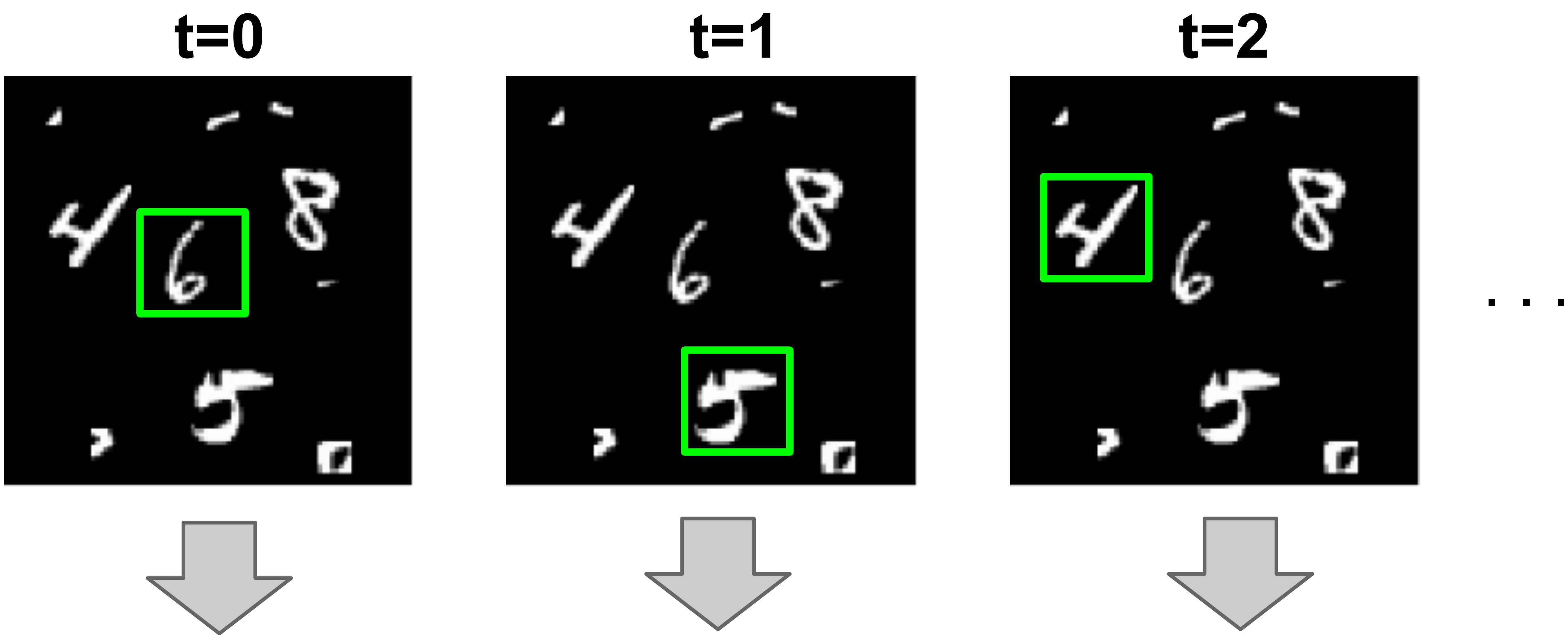
- binding without growing dimensionality
- fully distributed representation
- mathematical framework for storing and recovering information:
  - *multiplication* for binding
  - *addition* for combining
  - *operators* and *inverses*

# Network for binding and combining

(Eric Weiss, Ph.D. thesis)



# Example encoding




$$\mathbf{m} = \mathbf{v}_6 \odot \mathbf{r}_{t=0} + \mathbf{v}_5 \odot \mathbf{r}_{t=1} + \mathbf{v}_4 \odot \mathbf{r}_{t=2} + \ldots$$

# Example queries

**Where is the '5'?**

$$\text{answer} = \mathbf{v}_5^* \odot \mathbf{m}$$

$$= \mathbf{v}_5^* \odot (\mathbf{v}_6 \odot \mathbf{r}_{t=0} + \mathbf{v}_5 \odot \mathbf{r}_{t=1} + \mathbf{v}_4 \odot \mathbf{r}_{t=2} + \ldots)$$

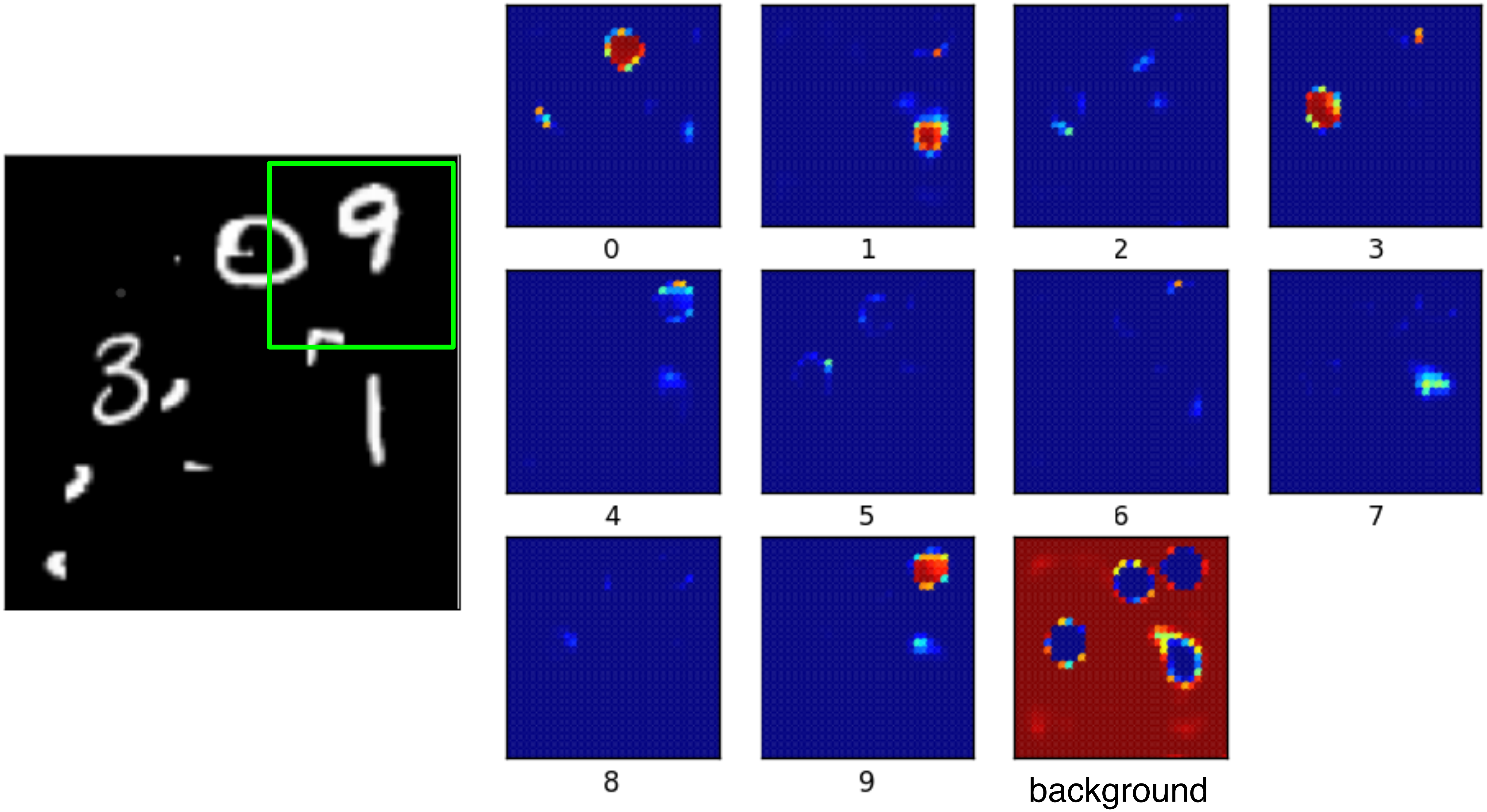
$$\approx 0 + \mathbf{r}_{t=1} + 0$$

**What object is in the center?**

$$\text{answer} = \mathbf{r}_{\text{center}}^* \odot \mathbf{m}$$

$$= \mathbf{r}_{\text{center}}^* \odot (\mathbf{v}_6 \odot \mathbf{r}_{t=0} + \mathbf{v}_5 \odot \mathbf{r}_{t=1} + \mathbf{v}_4 \odot \mathbf{r}_{t=2} + \ldots)$$

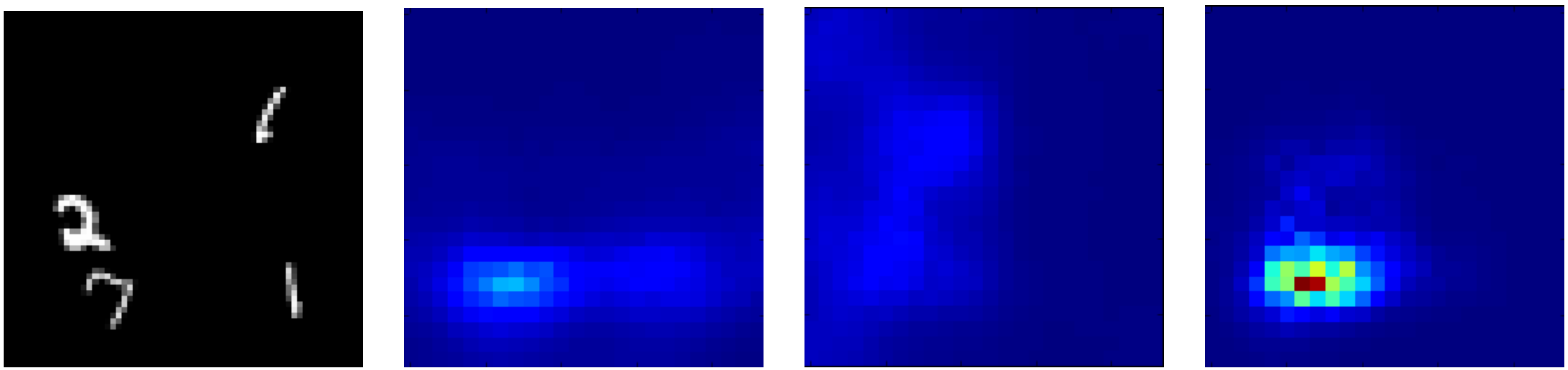
$$\approx \mathbf{v}_6 + 0 + 0$$



0
1
2
3
4
5
6
7
8
9
background

# Spatial reasoning

What is below a ‘2’ and to the left of a ‘1’?



(a) Example image (b) ”below a 2” (c) ”to the left of a 1” (d) Combined

$$\mathbf{a}_1 = f^{-1}(\mathbf{r}_{\mathbf{down}}\,(\mathbf{v}_{\mathbf{2}}^{*} \odot \mathbf{m}))$$

$$\mathbf{a}_2 = f^{-1}(\mathbf{r}_{\mathbf{left}}\,(\mathbf{v}_{\mathbf{1}}^{*} \odot \mathbf{m}))$$

$$\mathbf{a}_1 \odot \mathbf{a}_2$$

$$\text{answer} = f(\mathbf{a}_1 \odot \mathbf{a}_2) \odot \mathbf{m}$$

# Main points

- The drift movements that occur during fixation may be part of a purposeful, *active* sensing strategy to maximize the effective resolution offered by the foveal cone array.

- A *foveated* image sampling lattice similar to the primate retina emerges as the optimal solution for visual search, but only for an eye without the ability to zoom.

- Neural networks with the ability to *bind* and *combine* information across saccades are capable of building up a scene representation that supports spatial reasoning.