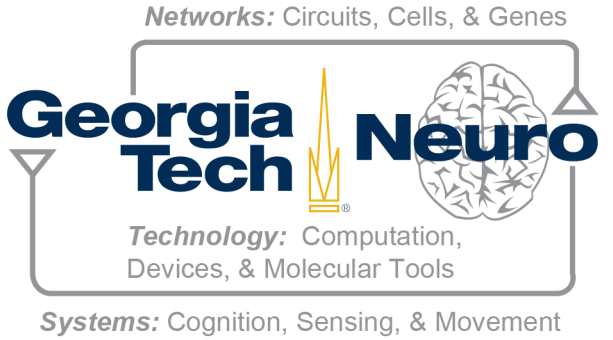
Dimensionality reduction as a model of efficient coding in sensory systems

Christopher J. Rozell

Georgia Institute of Technology



Acknowledgments

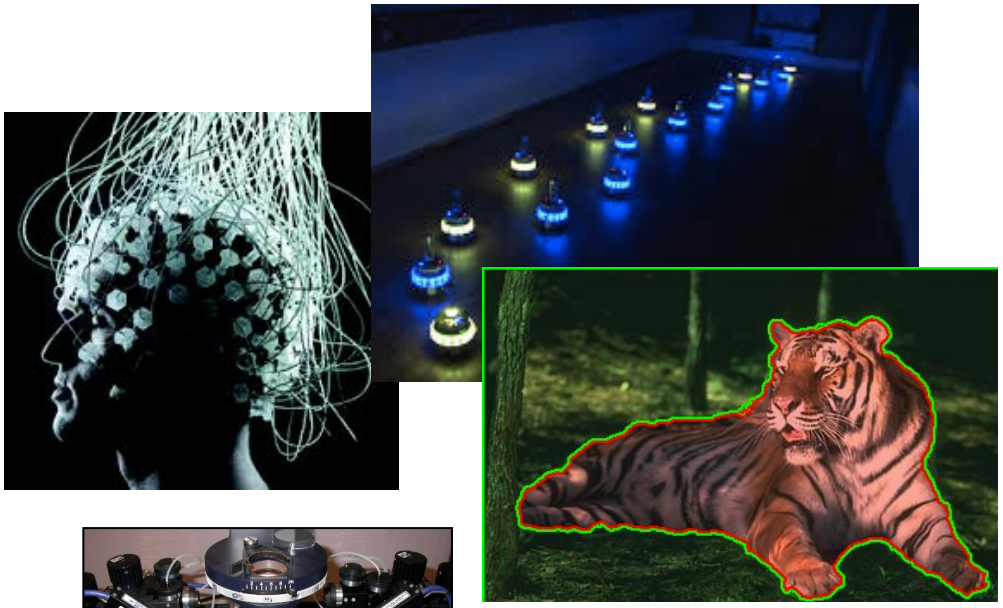
- Arish Alreja
- Aurele Balavoine
- Nick Bertrand
- Michael Bolus
- Greg Canal
- Adam Charles
- Marissa Connor
- Allison Del Giorno
- Pavel Dunn
- Stefano Fenu
- Abbie Kressner
- John Lee
- Ninghao Liu
- Siva Manavasagam
- Matt O'Shaughnessy
- Adam Willats
- Han Lun Yap
- Dong Yin
- Mengchen Zhu

"I not only use all the brains I have, but all I can borrow."

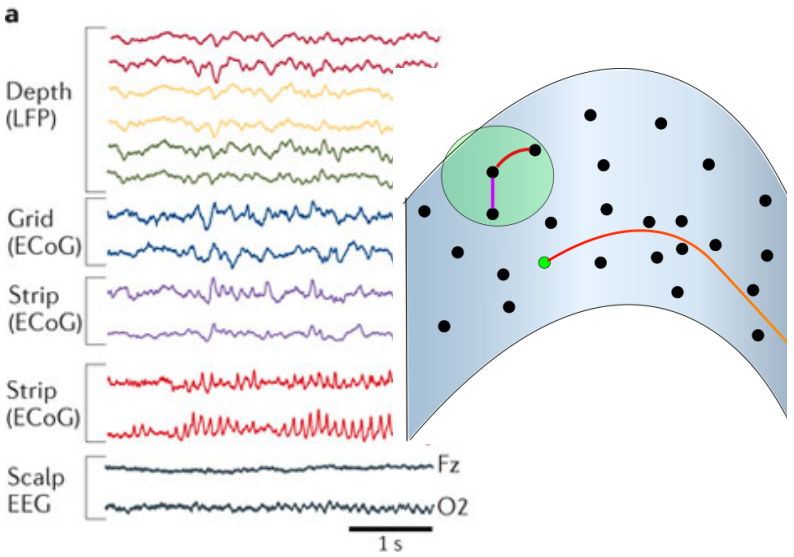
-Woodrow Wilson

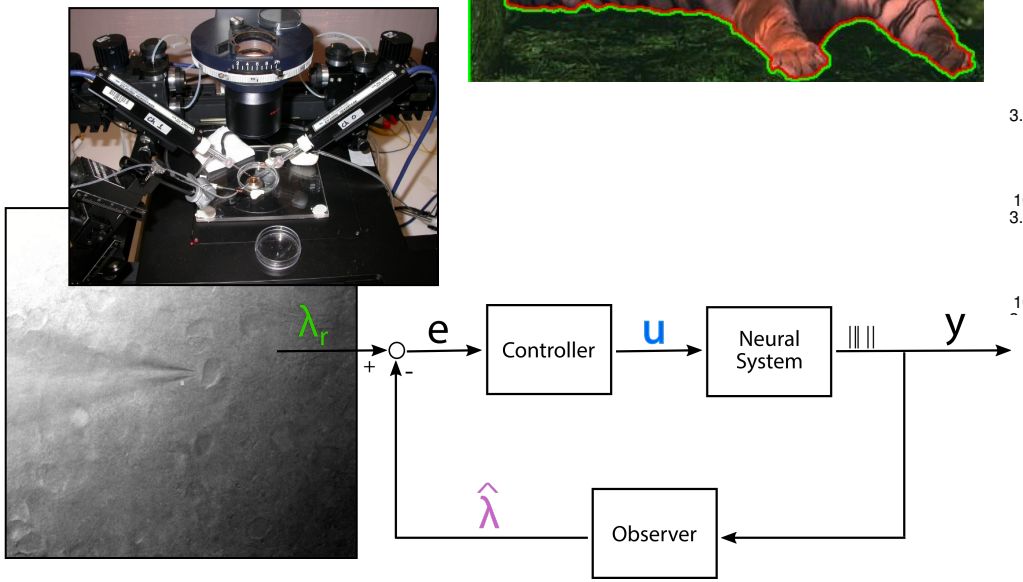
Today science…next week technology

Brain machine interfaces

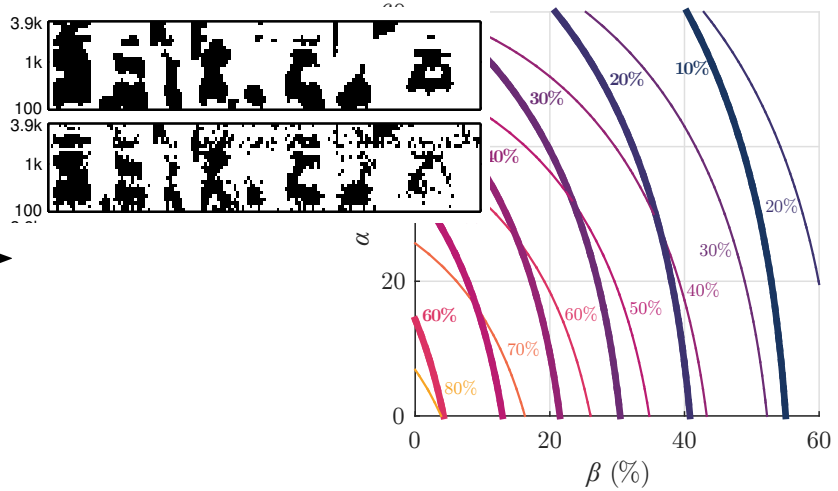


Data acquisition and analysis



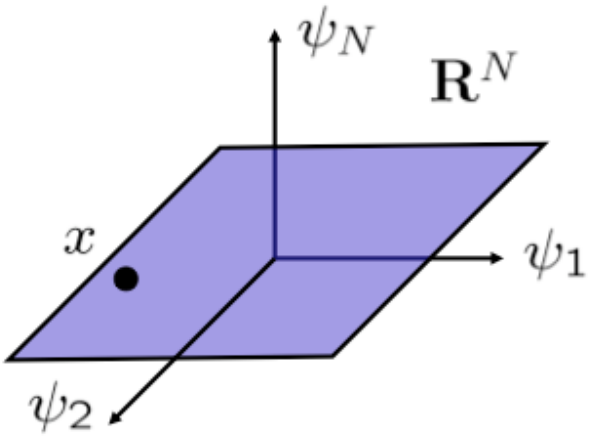


Closed-loop electrophysiology

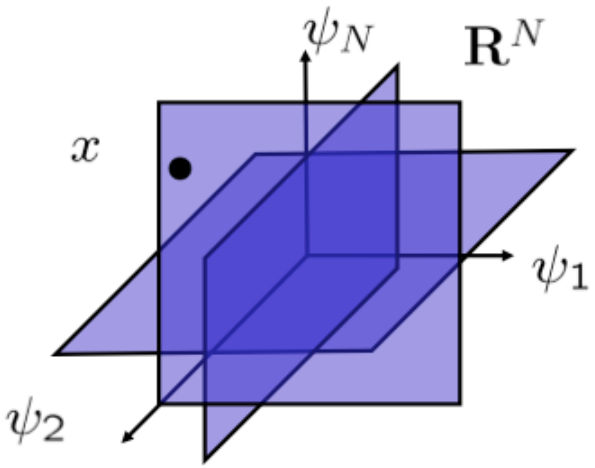


Speech intelligibility

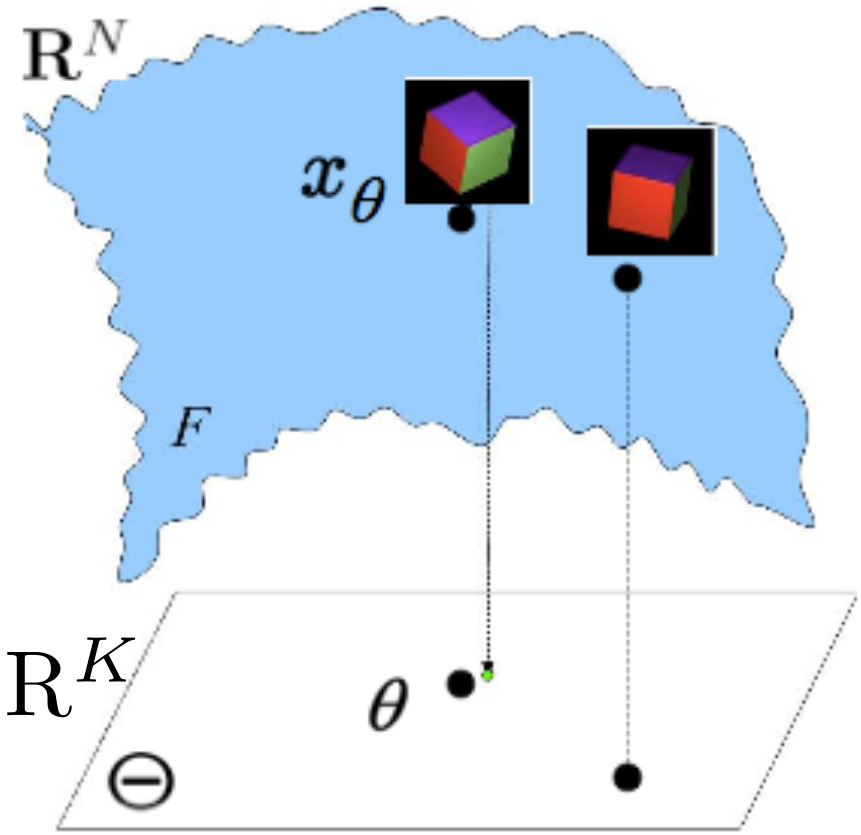
Low-dimensional geometric structure



Linear subspace

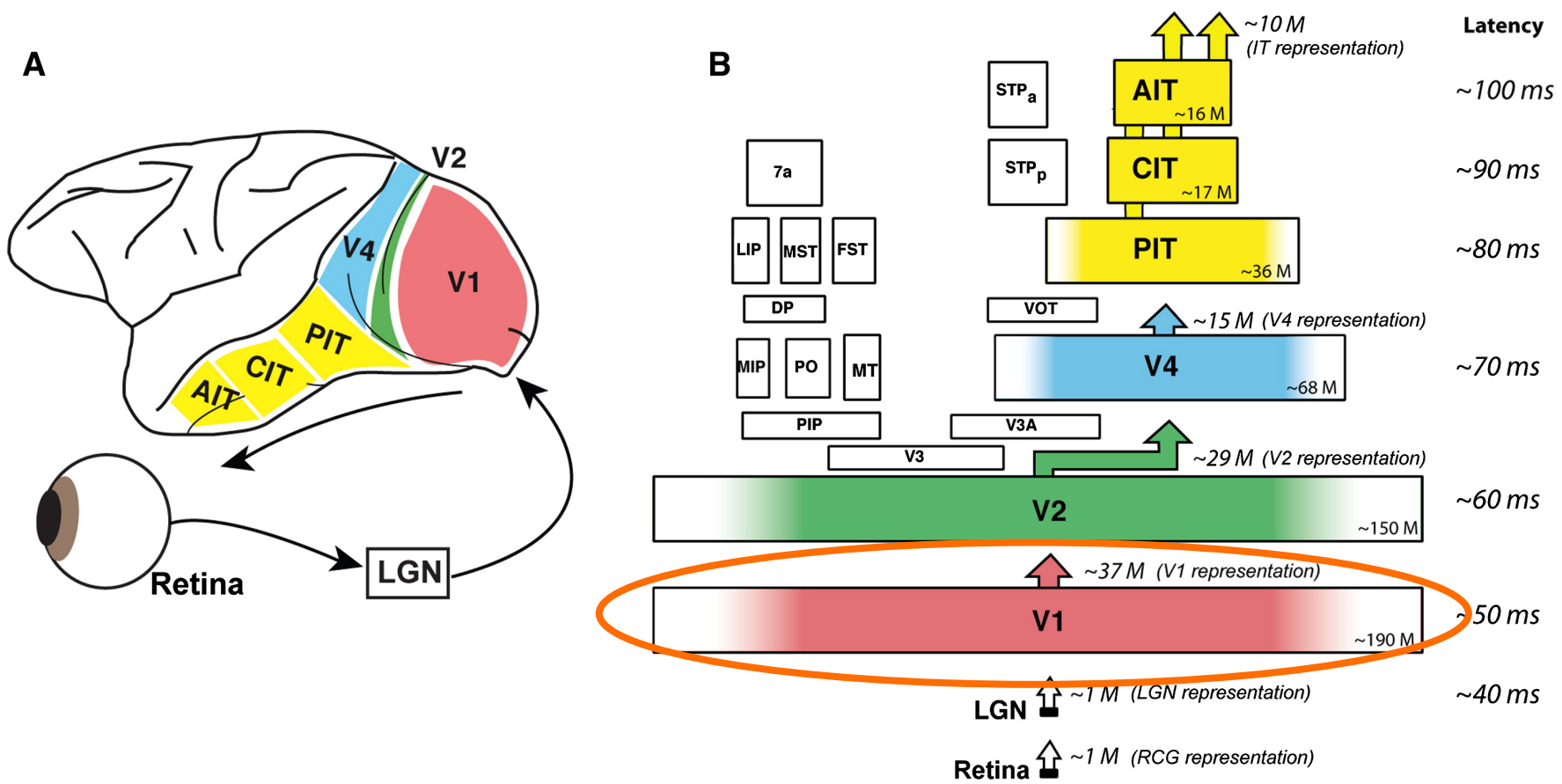


Sparse coefficients



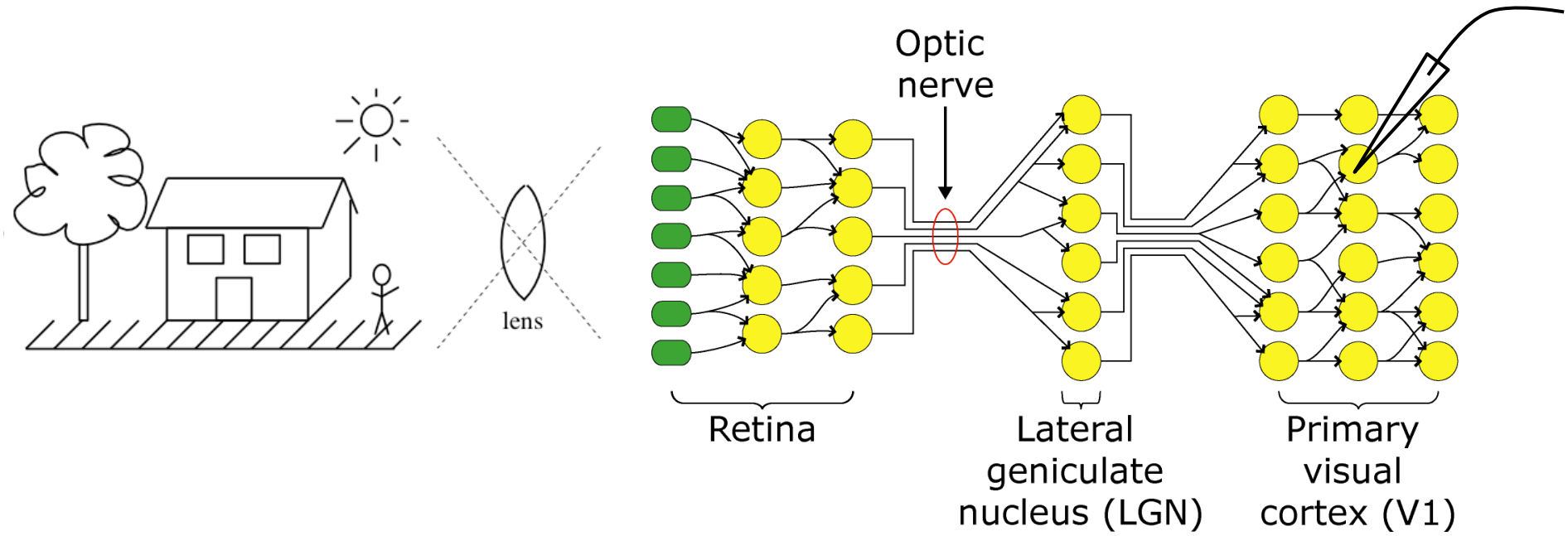
Manifold

Visual pathway



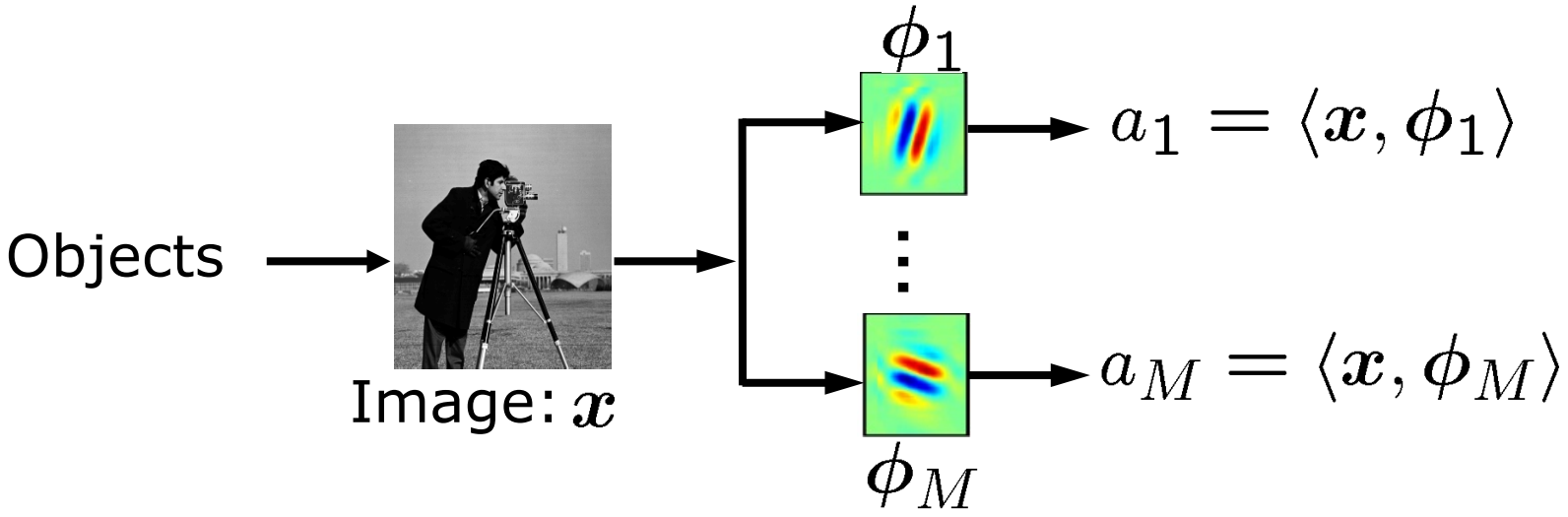
(DiCarlo, Zoccolan & Rust, 2012)

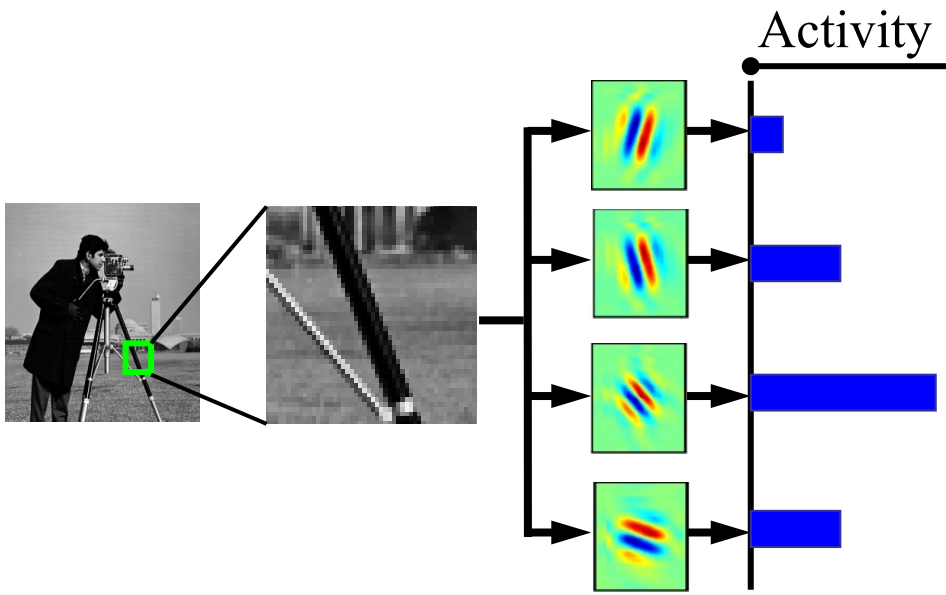
Visual receptive fields



(adapted from Hubel 1988)

“Standard model” of V1 coding



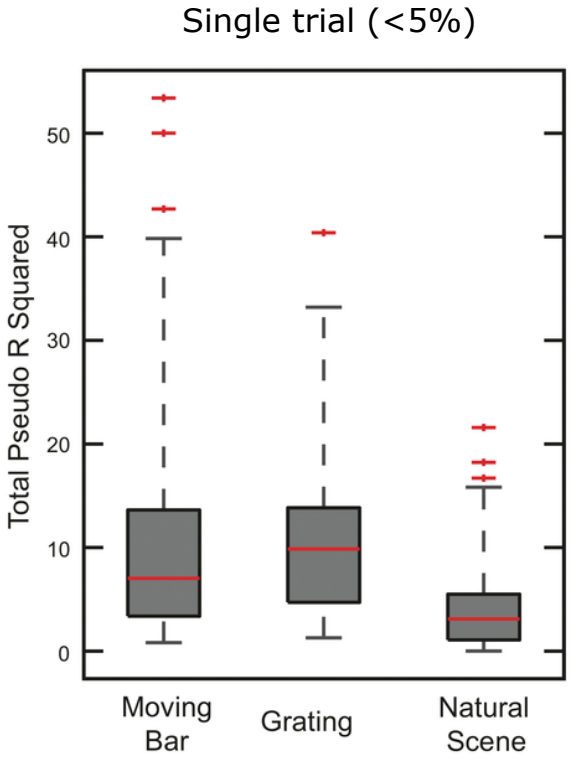


V1 receptive field
≈ Gabor wavelet

(Jones & Palmer 1987)

Betrayed by activity and anatomy

Explained variance



(Haslinger et al. 2012)

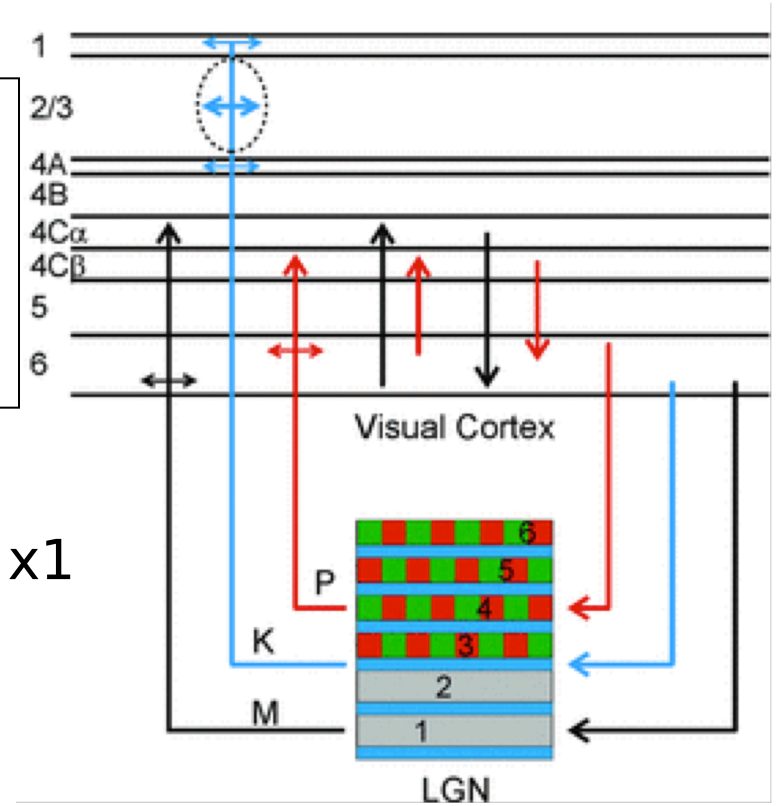
Excitatory L4 synapse origins

~5% from LGN

~28% from L4

~45% from L6

Recurrence and feedback



(Briggs & Usrey 2011; Sherman & Koch 1990; Ahmed et al. 1994; Binzegger et al. 2004; Thompson & Lamy 2007)

Betrayed by the challenge of vision…

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group — July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

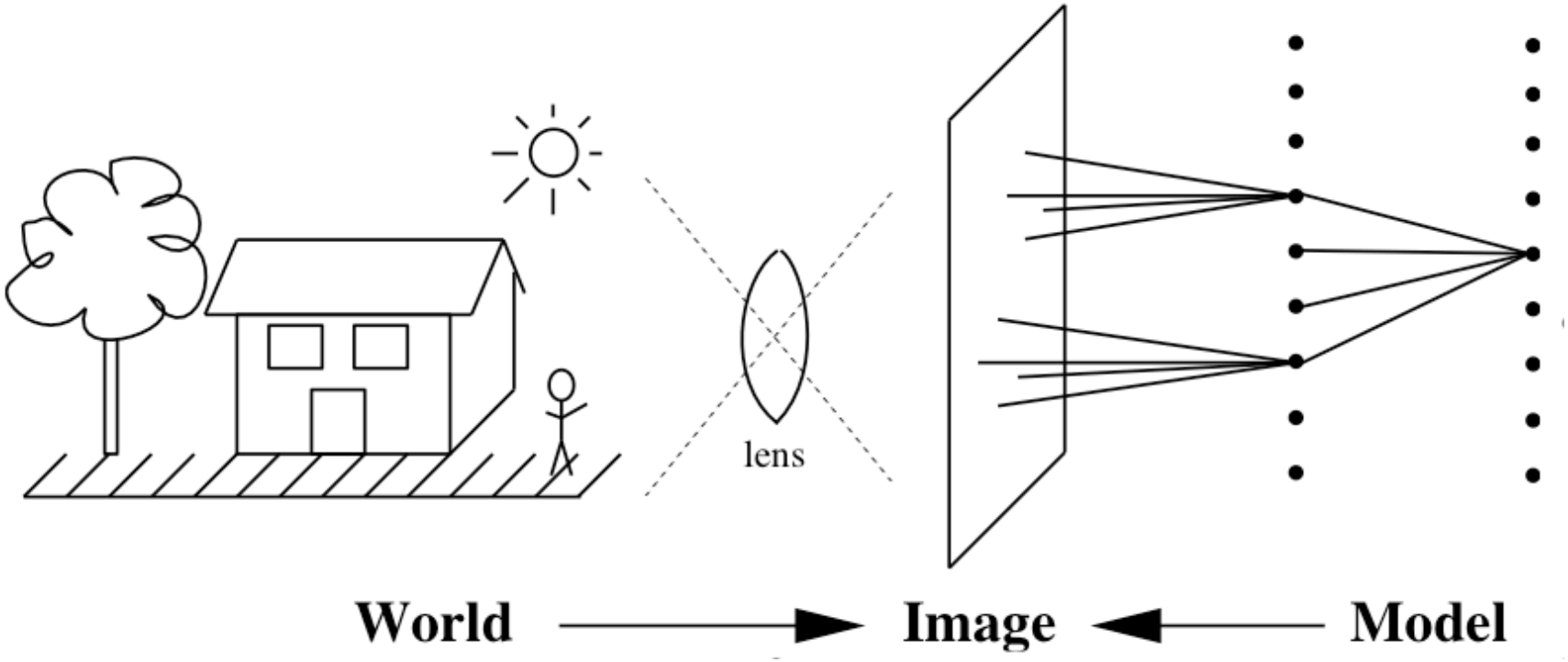
Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

(Tarr->DiCarlo->Olshausen)

Vision as inference

- Vision must infer environmental causes from scenes
  - Significant evidence for perception=probabilistic inference
  - Inherently ill-posed and ambiguous
  - Of all possible causes, which few are present now?



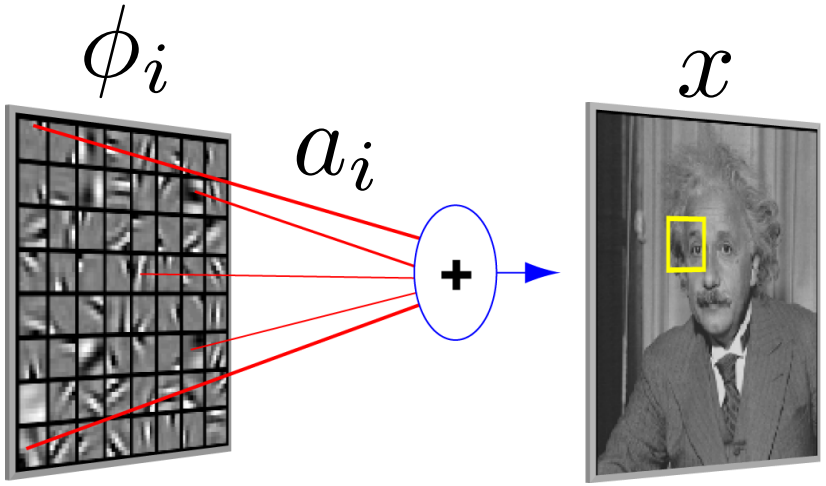
(Doya et al. 2007; Olshausen)

Modern natural image statistics

- Linear generative model:

Image Dictionary Coefficients

$$x = \sum_i \phi_i a_i + w$$

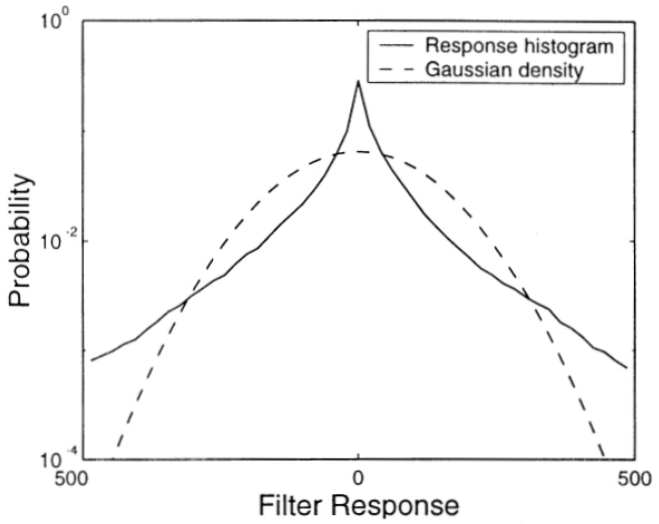


- Key notion is *sparsity*: want most a_m to be zero

1 megapixel image

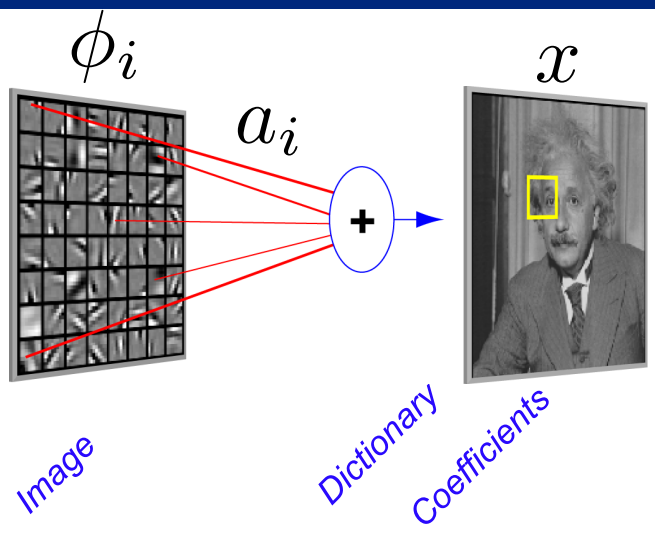
25k term approx

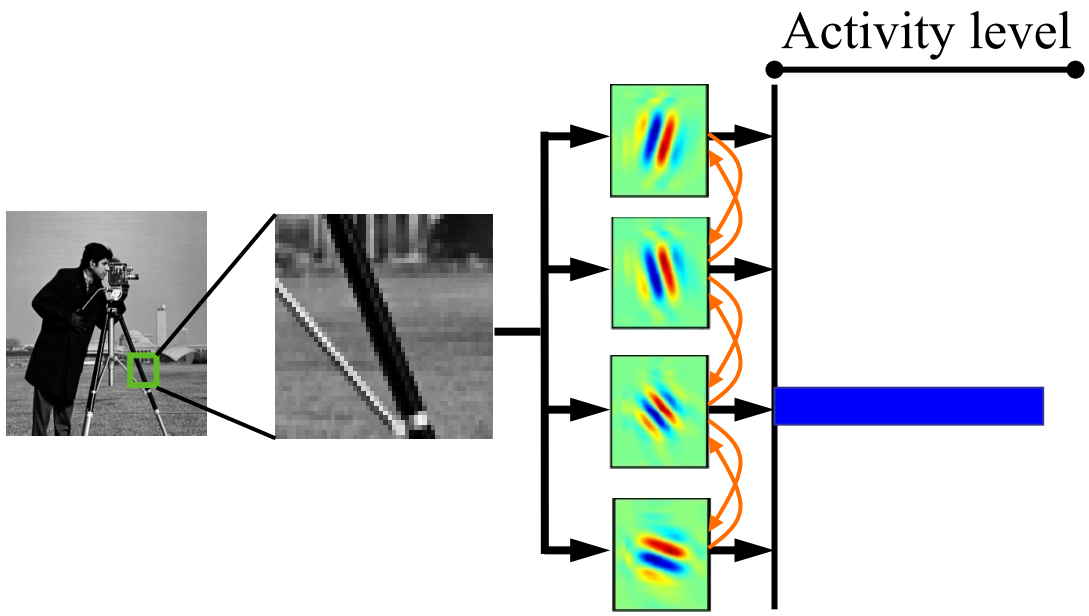
(Romberg 2008)



(adapted from Field 1994)

Hypothesis: sparse coding





$$x = \sum_i \phi_i a_i + w$$

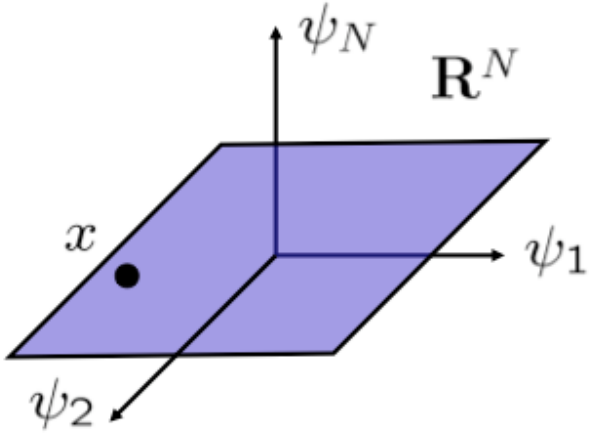
$$\{a_m\} = \arg\min_a \frac{1}{2} \left\| x - \sum_m \phi_m a_m \right\|_2^2 + \lambda \sum_m C(a_m)$$

Represent the image

Using few features

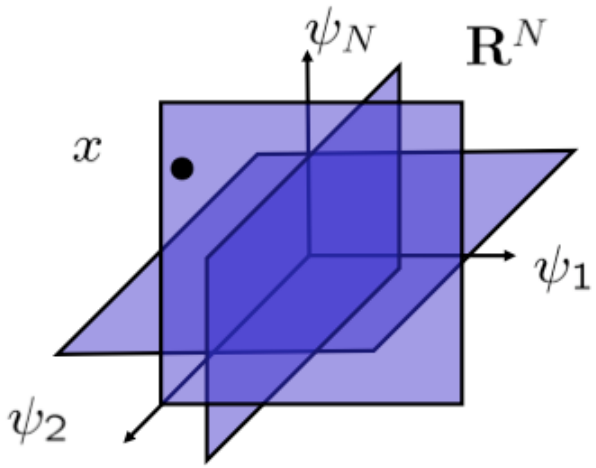
(Barlow 1961; Field 1994; Olshausen & Field 1996)

Data adaptive dimensionality reduction



Linear subspace

Find the best subspace for all data jointly (PCA)

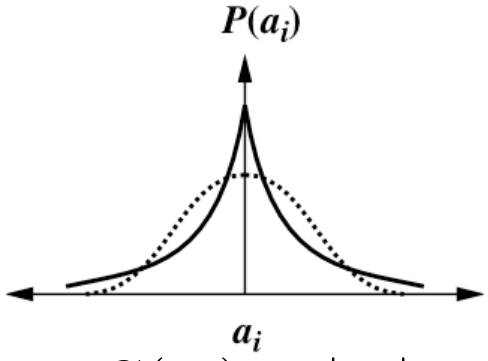


Sparse coefficients

Find subspace collection for all data (learning); decide on best subspace per datapoint (inference)

Bayesian inference

- Causes are independent: $p(\mathbf{a}) = \prod_i p(a_i)$
- Causes are sparse: $p(a_i) \propto e^{-C(a_i)}$
- Noise is Gaussian: $p(x|a) \propto e^{-\|x-\Phi a\|_2^2/2\sigma_w^2}$



$C(a_i) = |a_i|$

- Infer $\{a_i\}$ via maximum a posteriori (MAP) estimate:

$$\widehat{a} = \arg\max_a p(a|x) = \arg\max_a \left[\log p(x|a) + \log p(a)\right]$$

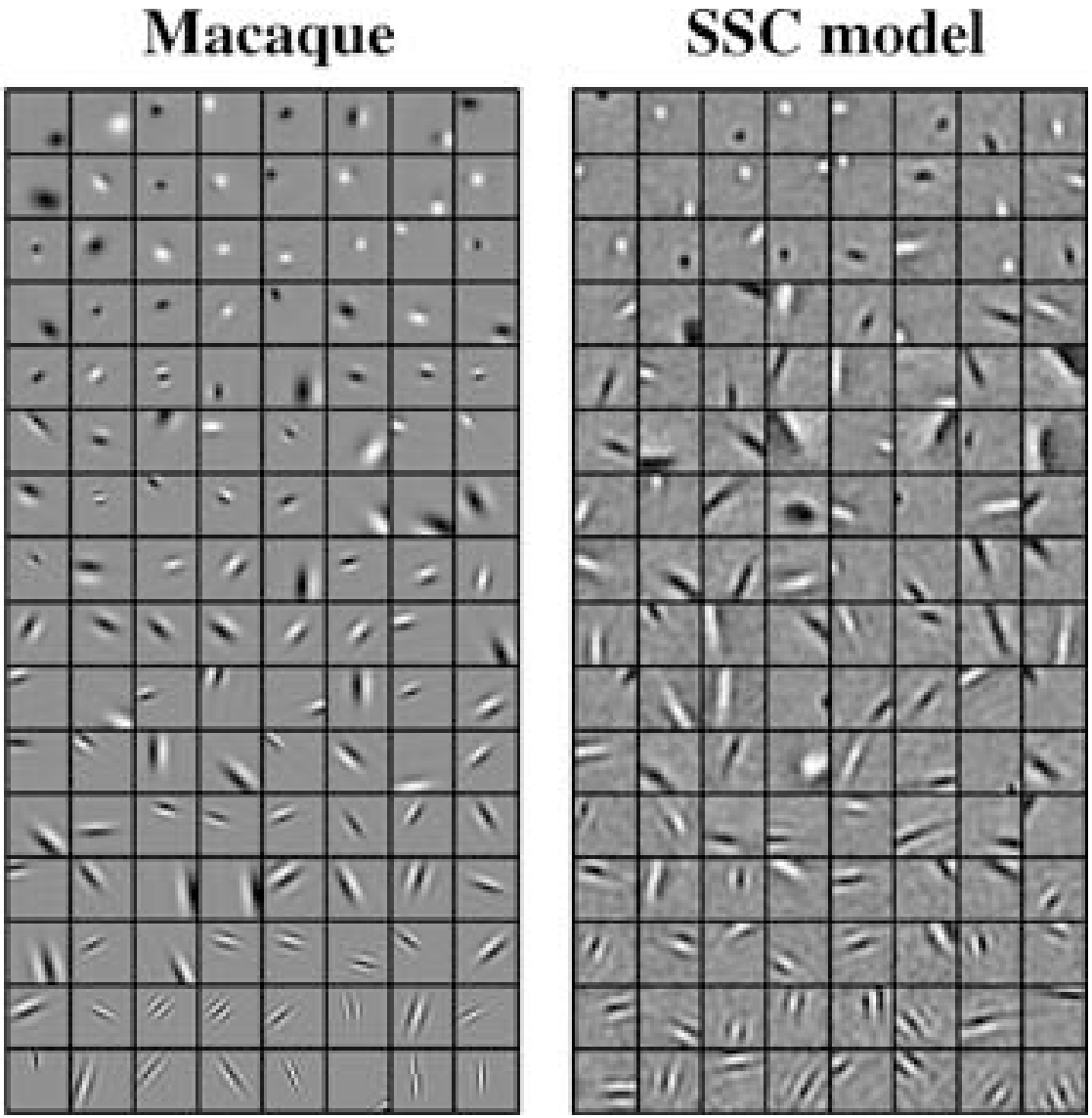
$$= \arg\min_a \left[\|x - \Phi a\|_2^2 + \lambda \sum_i C(a_i)\right]$$

Dictionary learning

- Learn dictionary Φ from image statistics
  - Start with arbitrary dictionary
  - Find sparse coefficients for image database
  - Adjust dictionary down gradient to improve performance
  - Iterate
- Very related to ICA

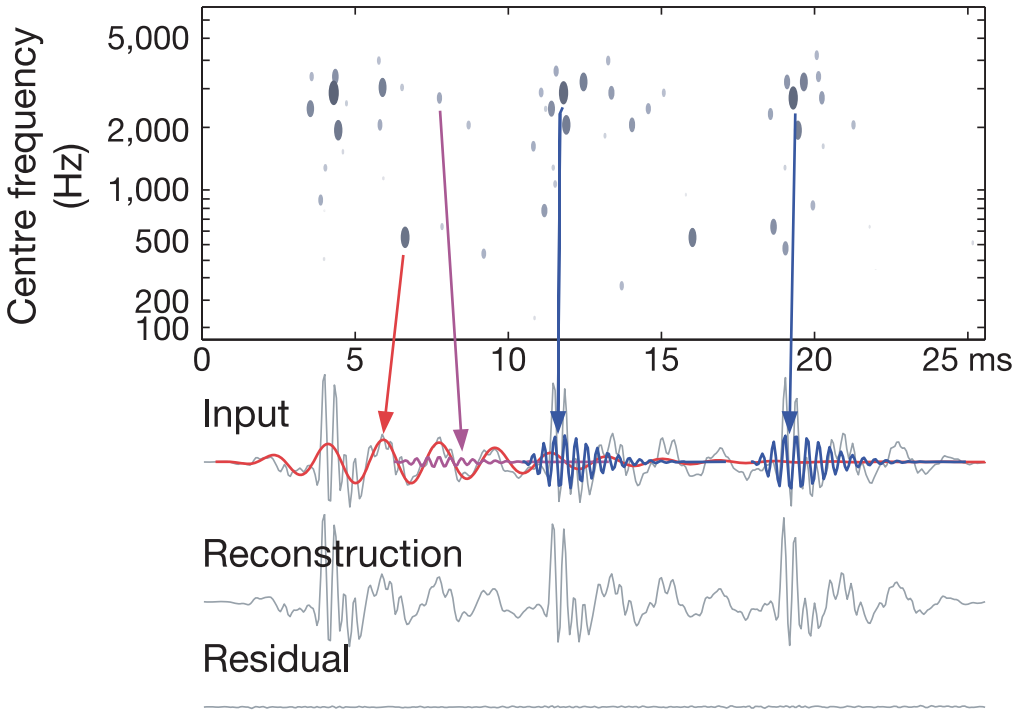
$$
\begin{aligned}
E &= \frac{1}{2}\left\| I - \sum_i a_i \phi_i \right\|_2^2 + \lambda \sum_i C(a_i) \\
\frac{\partial E}{\partial \phi_i} &= -a_i \left(I - \sum_i a_i \phi_i \right) \\
\Delta \phi_i &= \eta \left\langle a_i \left(I - \sum_i a_i \phi_i \right) \right\rangle
\end{aligned}
$$

Measured and learned RFs



(Olshausen & Field 1996; Rehn & Sommer 2007; Hunt et al. 2013)

Dictionary learning in auditory periphery

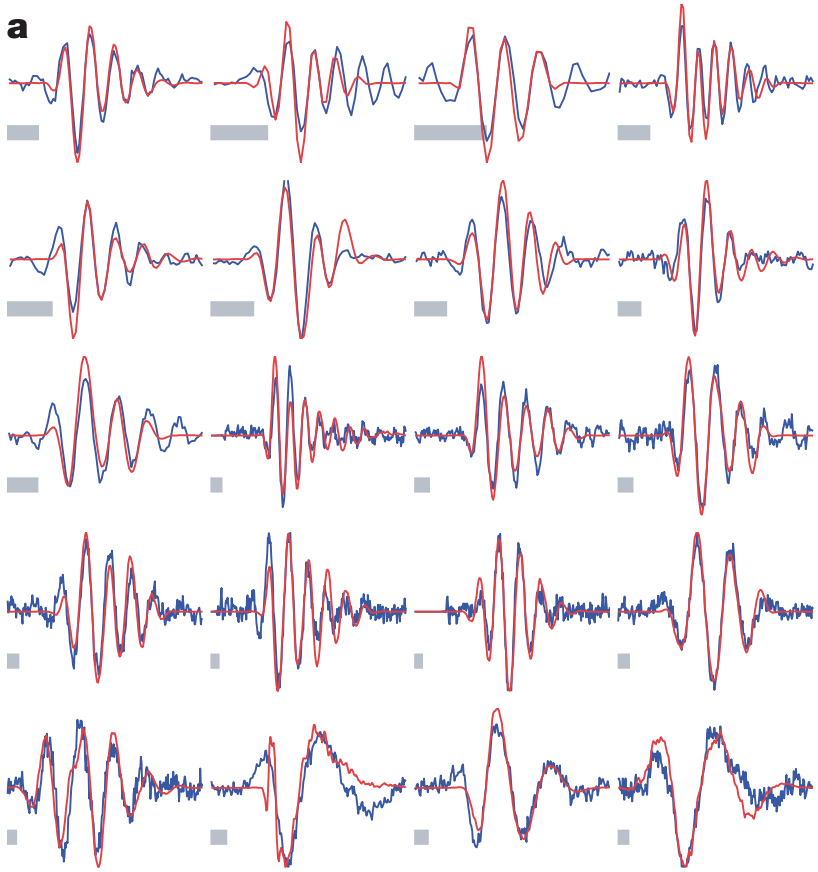


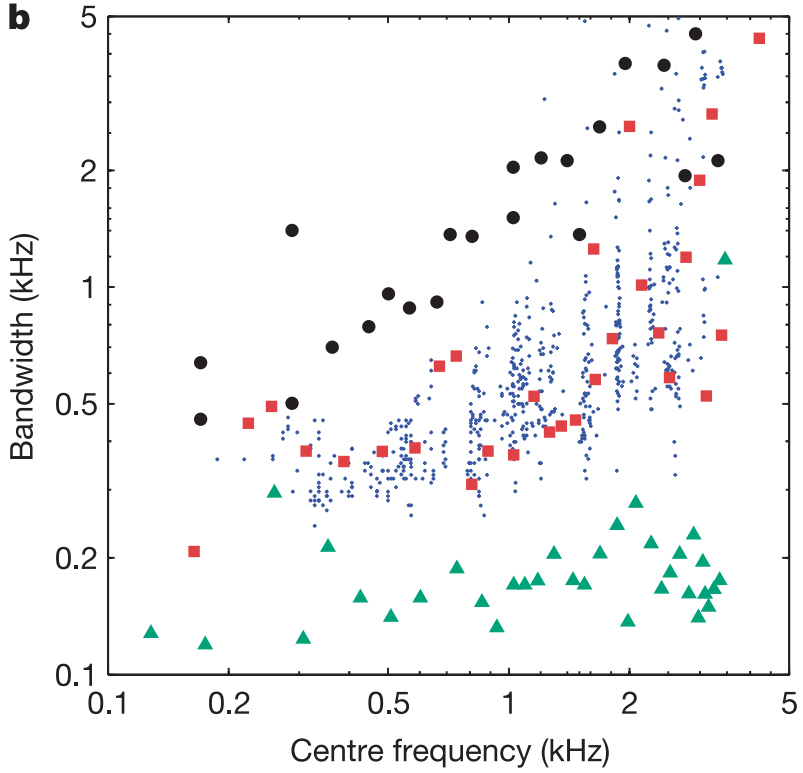
(Lewicki 2002;
Smith & Lewicki 2006)

$$x(t) = \sum\sum s_i^m \phi_m \left(t - \tau_i^m\right) + \varepsilon(t)$$

- Convolutional adaptation of sparse coding model
- Trained on a mixture of mammalian vocalizations and environmental sounds (transient and ambient)

Results match revcor AN filters

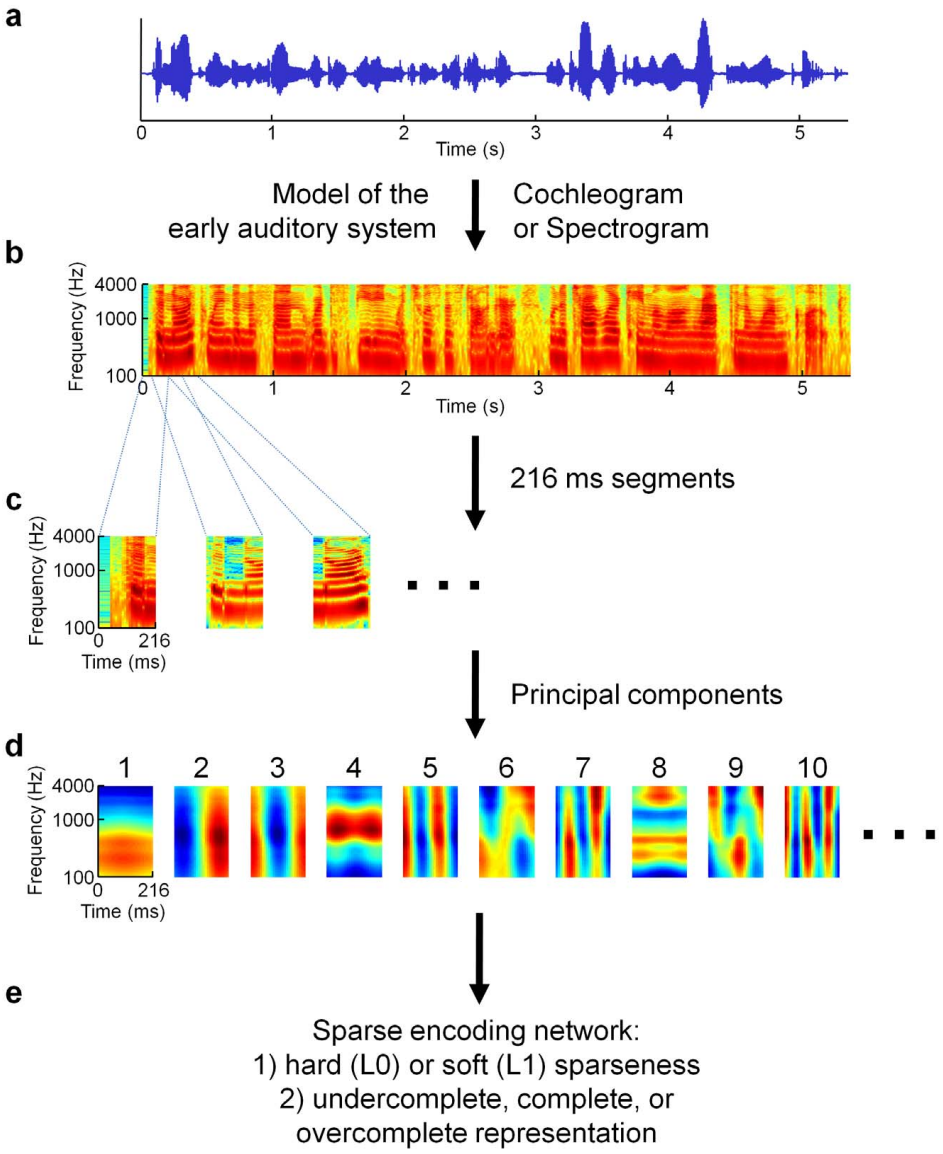




Blue: revcor AN filters
Red: learned from mixture (speech similar)
Black: learned from environmental
Green: learned from vocalizations

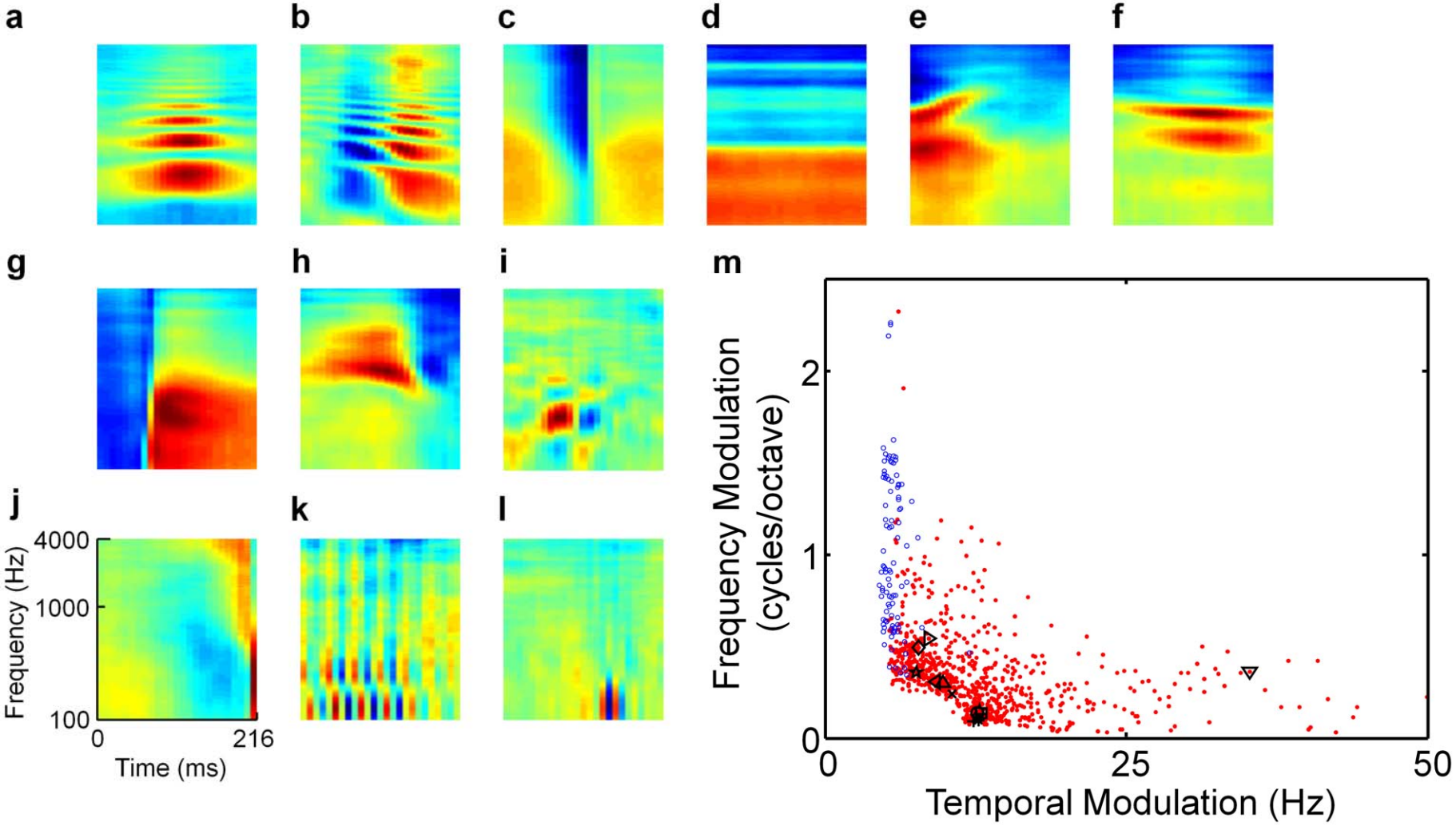
(Smith & Lewicki 2006)

Dictionary learning in higher representations



(Carlson et al., 2012;
see also Klein et al. 2003)

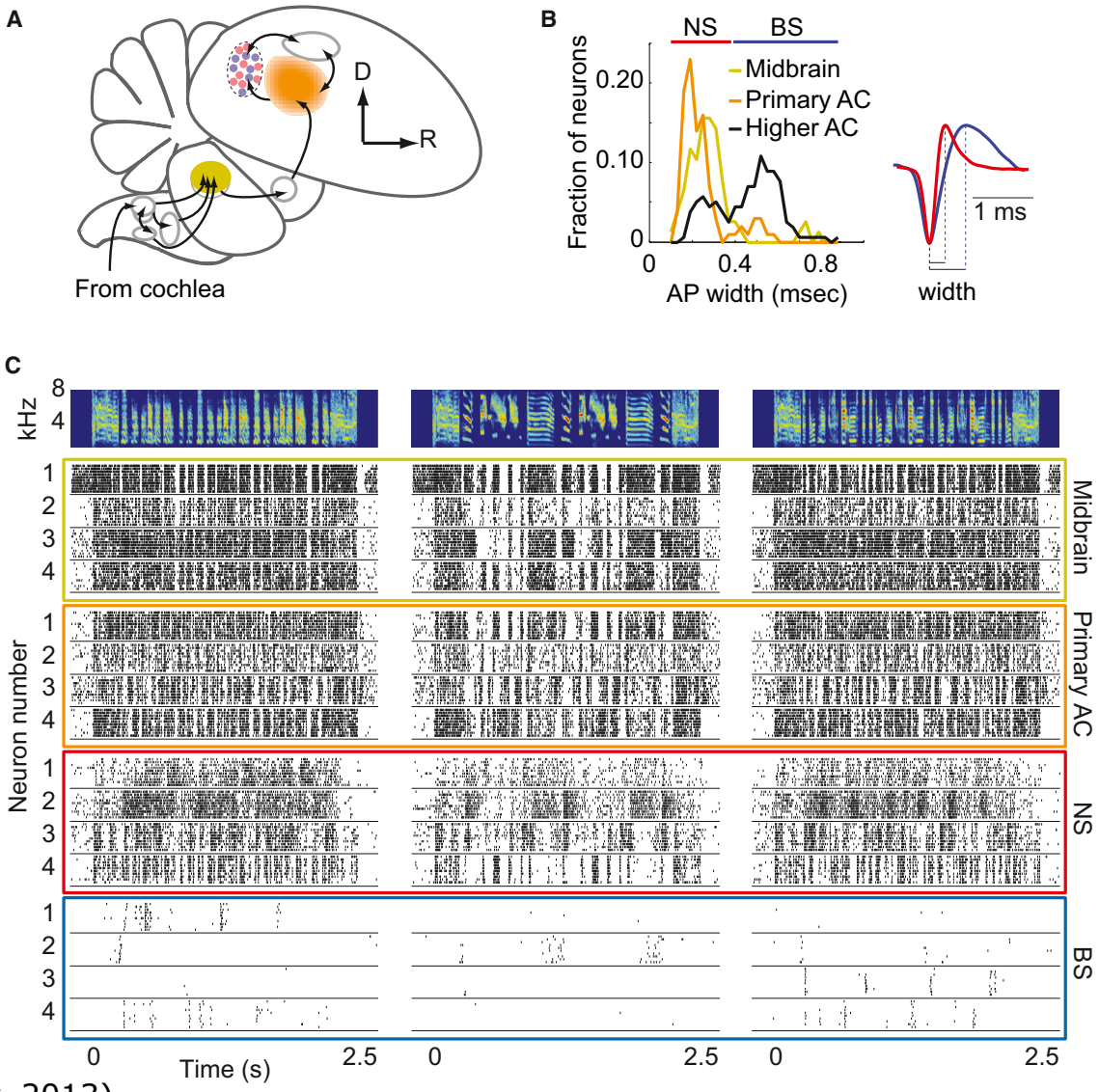
STRF properties similar to IC/thalamus/A1



(Carlson et al., 2012)

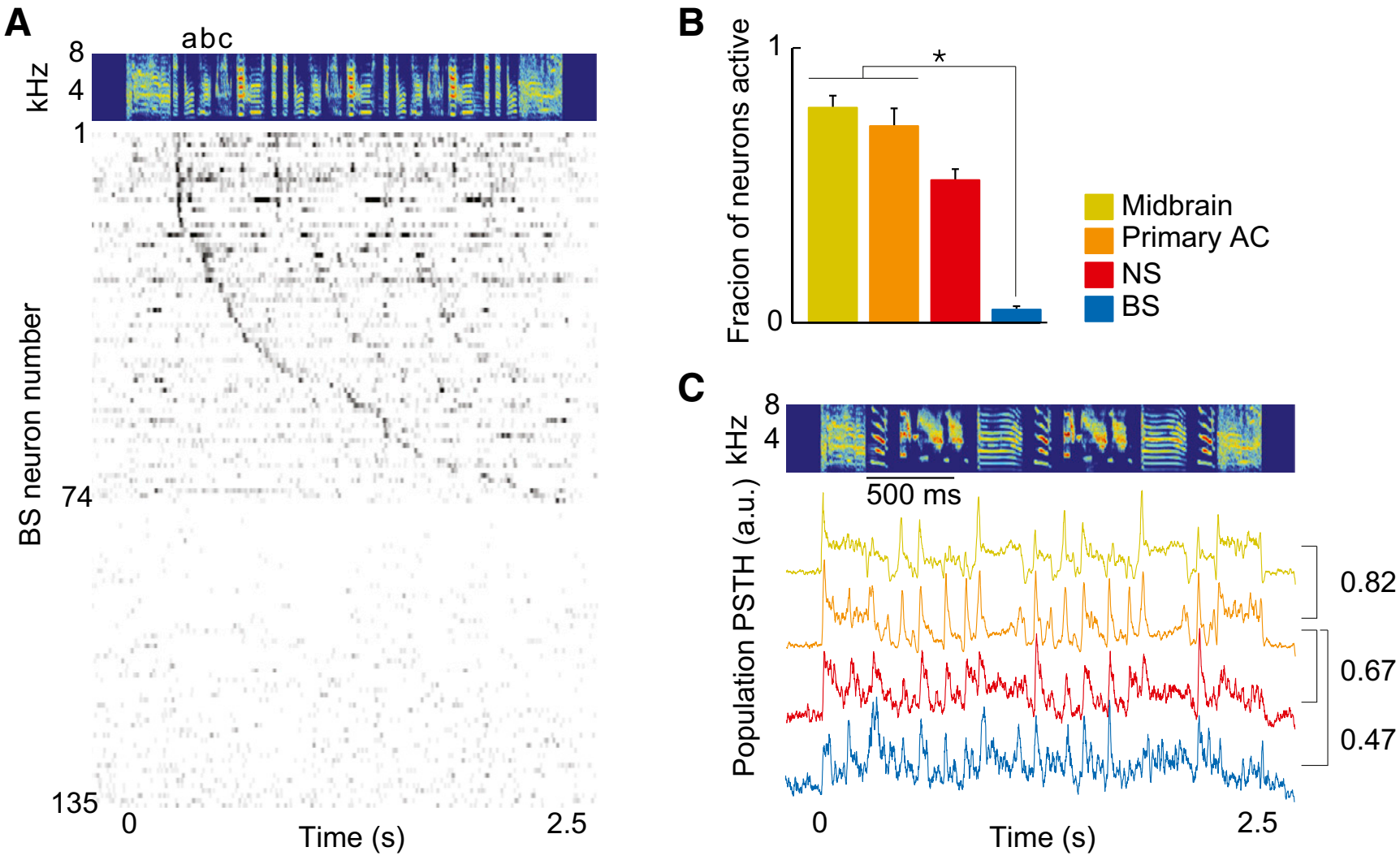
Red: spectrogram; 4x overcomplete
Blue: cochleaogram; 0.5x overcomplete

Sparsity in songbird AC broad spiking cells



(Schneider & Woolley, 2013)

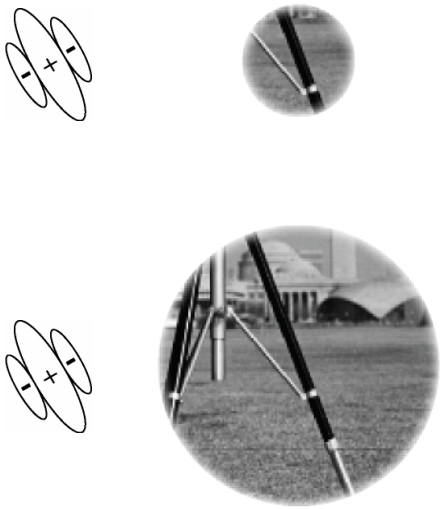
High reliability and increasing sparsity

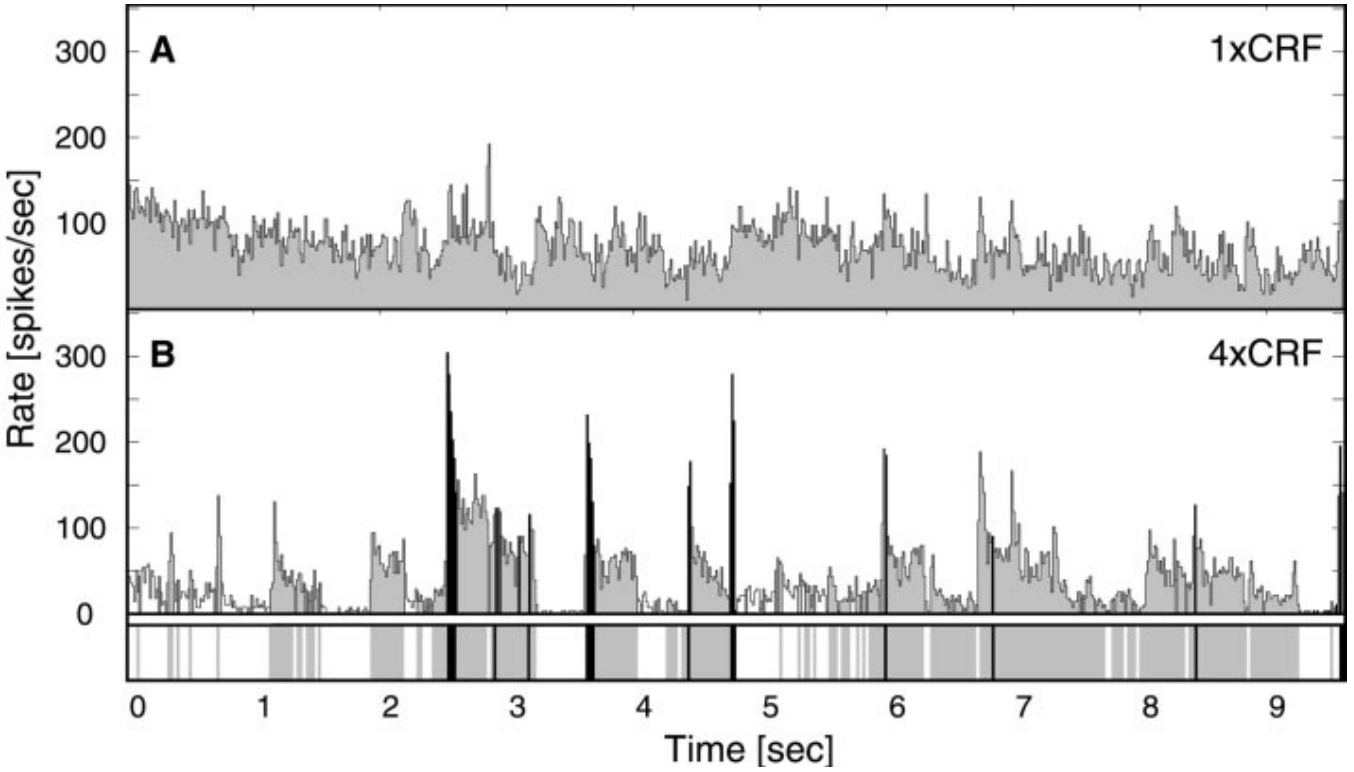


(Schneider & Woolley, 2013)

Sparsity in primate V1

(Vinje & Gallant 2002)

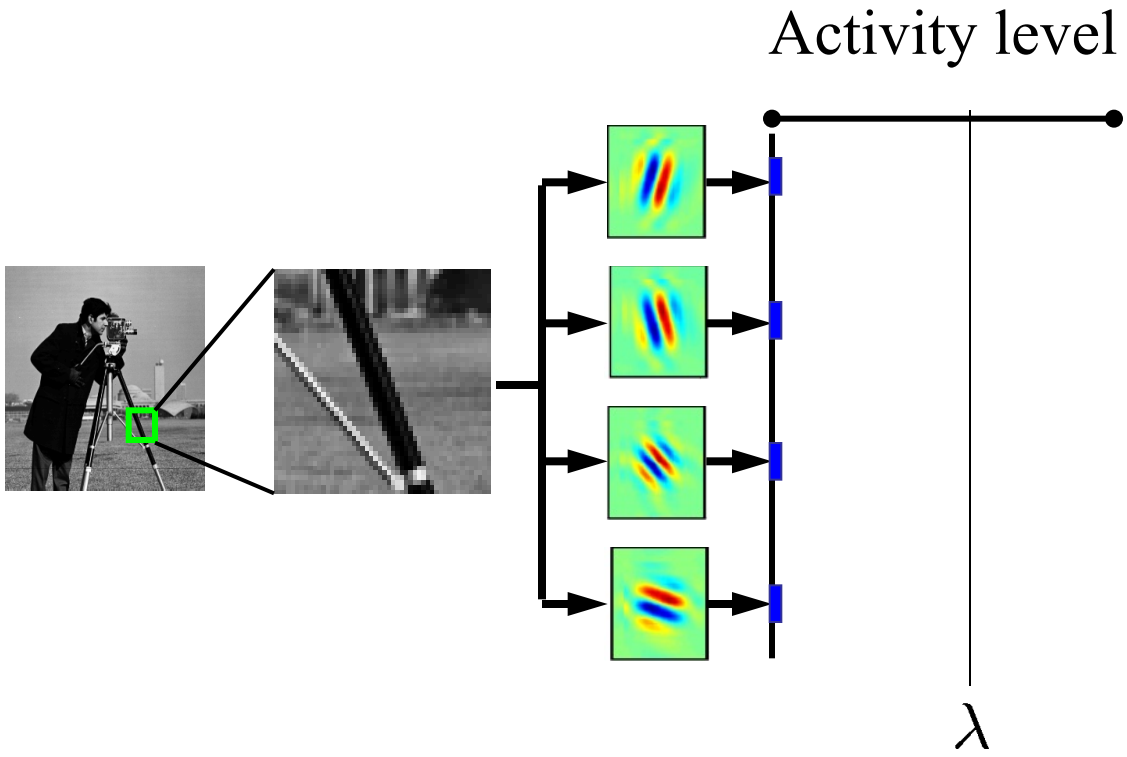




(see also Haider et al. 2010)

Can sparse coding account for response properties?

Sparsity computation in dynamical systems

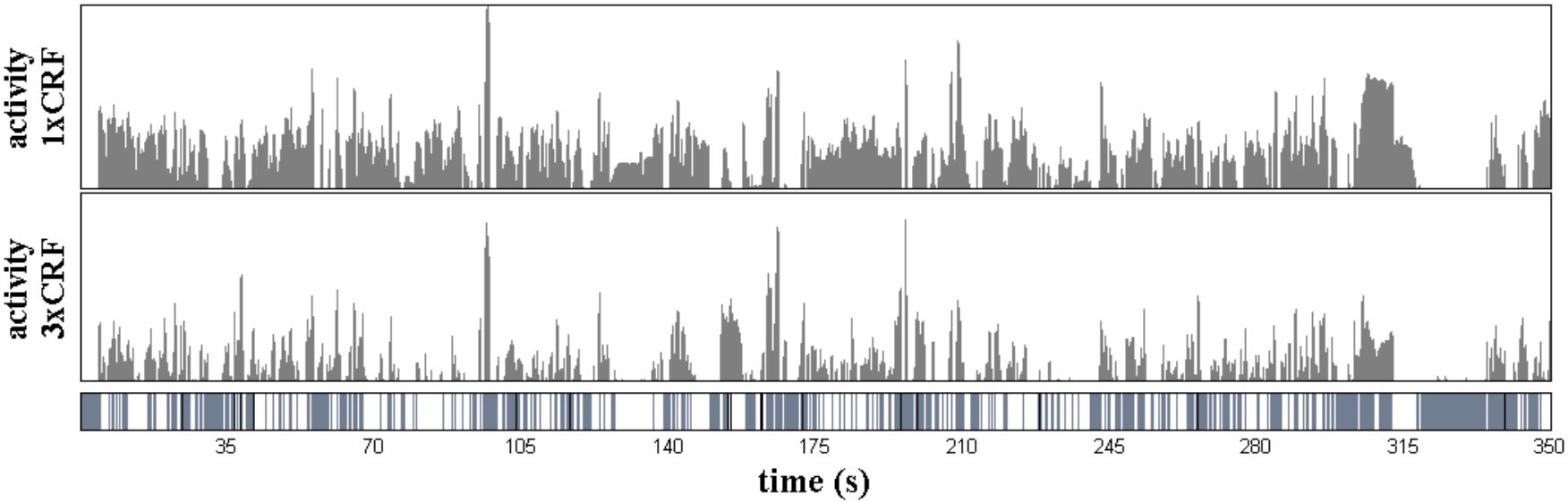


$$\dot{u}_m(t) = \frac{1}{\tau}\left[\langle \phi_m, \boldsymbol{x}(t)\rangle - u_m(t) - \sum_{n \neq m} \langle \phi_m, \phi_n \rangle \, a_n(t)\right]$$

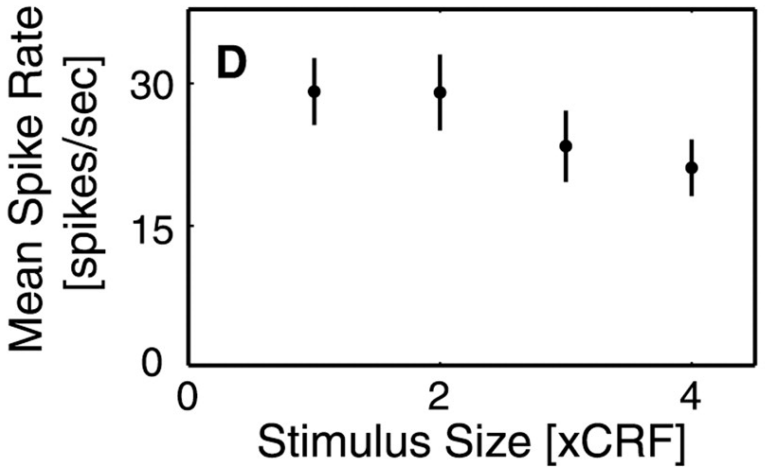
$$a_m = T_\lambda(u_m)$$

(R., Johnson, Baraniuk and Olshausen 2008; Balavoine, Romberg & R., 2012; Balavoine, R. & Romberg, 2013a,b; Charles, Garrigues & R., 2012; Shapero, Charles, R. & Hasler 2012; Shapero, R. & Hasler 2012,2013; Shapero, Zhu, Hasler & R. 2014; Balavoine, R. & Romberg, 2015; Olshausen & R., 2017)

Model response to natural stimuli

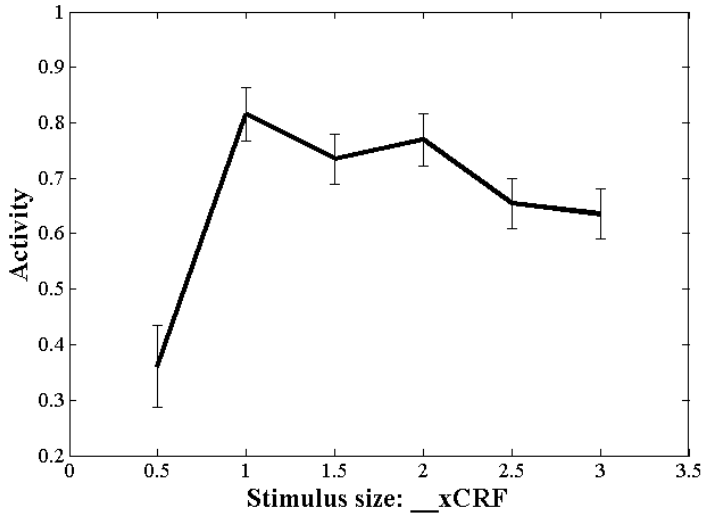


Physiology



(Vinje & Gallant 2002)

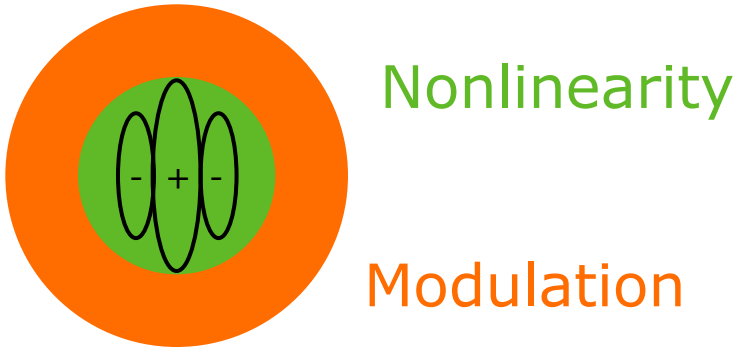
Model



(Del Giorno, Zhu & R. 2013)

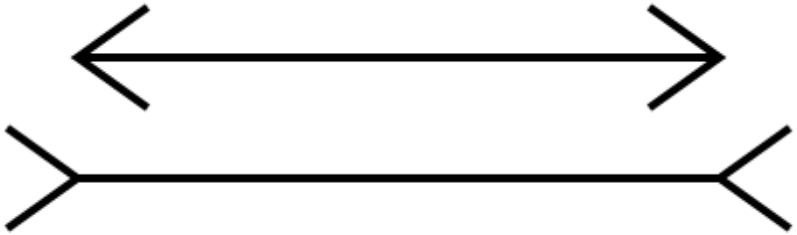
Nonclassical RF (nCRF) effects

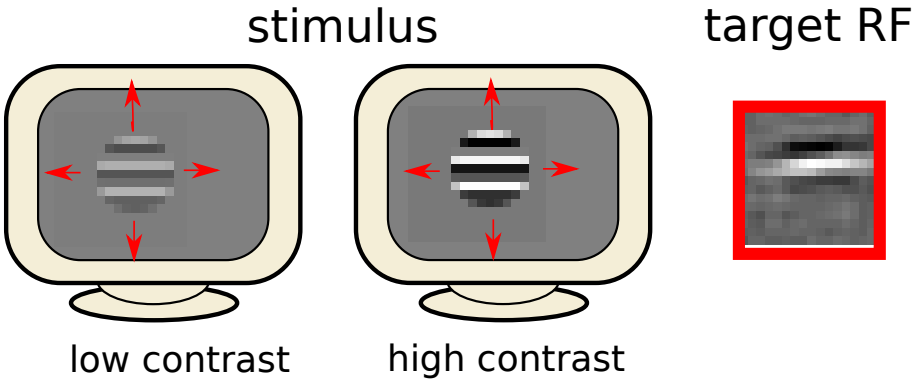
- Nonlinear response properties in classical RF
- Context modulation outside classical RF



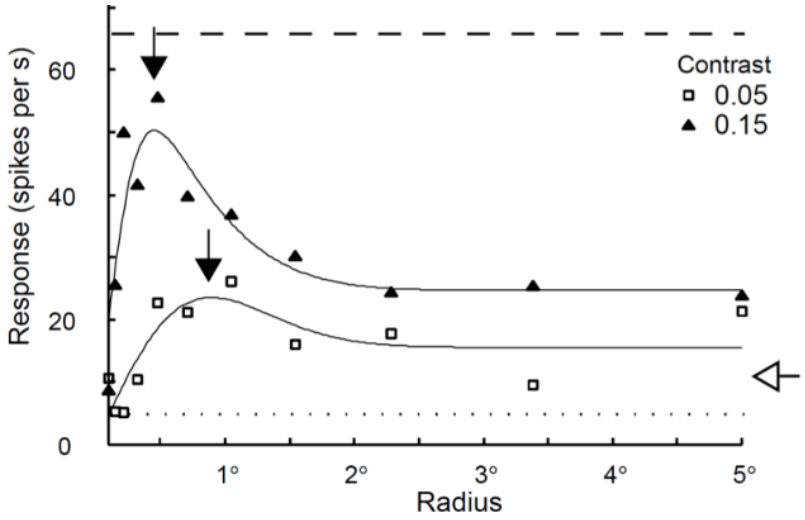
- Many models:
  - Mechanistic (Priebe & Ferster 2012)
  - Phenominological (Series et al. 2003)
  - Some previous connection to optimal coding rules

Surround suppression



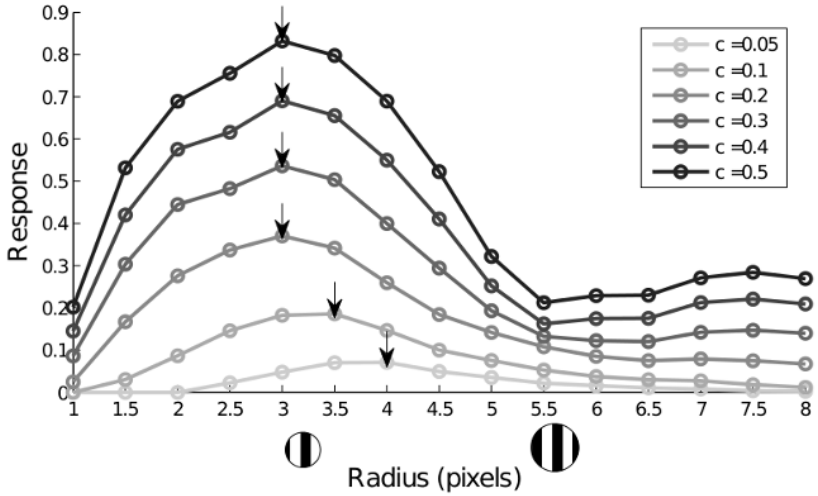


Physiology



(Sceniak et al. 1999)

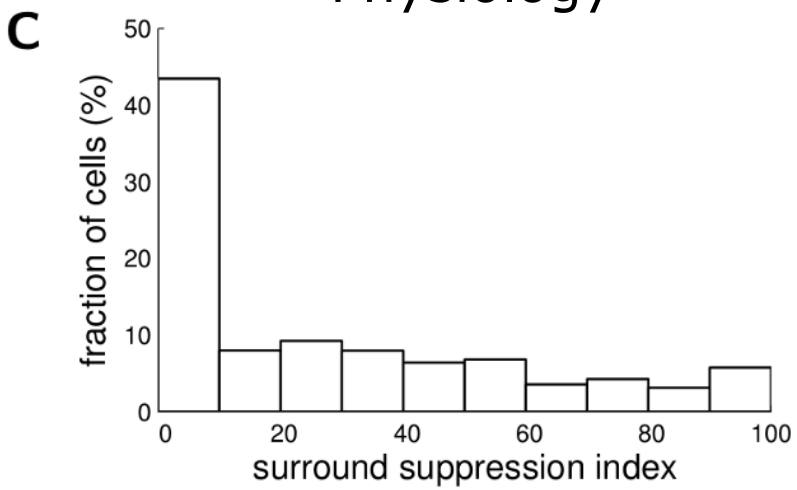
Model



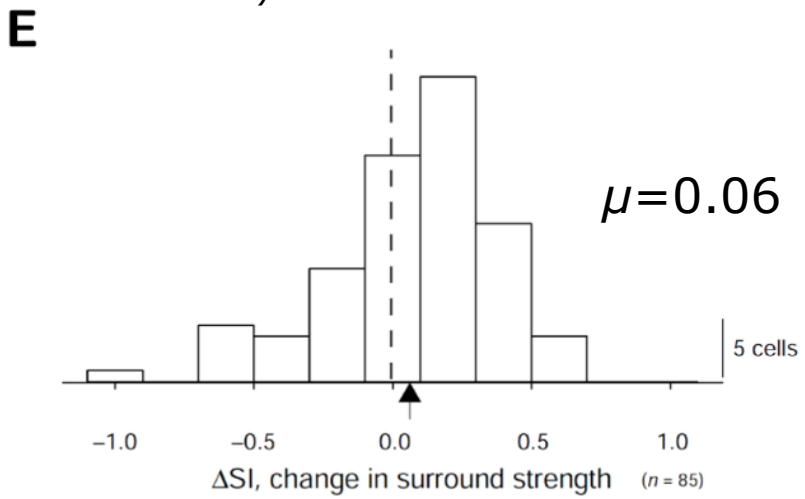
(Zhu & R. 2013)

Surround suppression index

Physiology

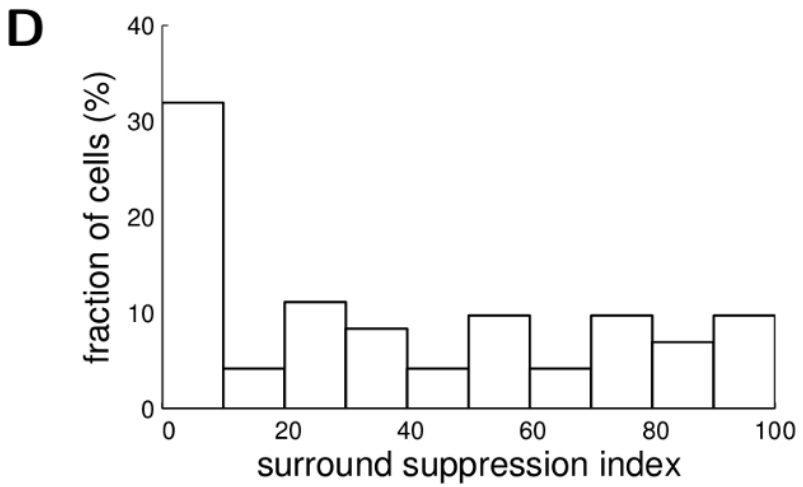


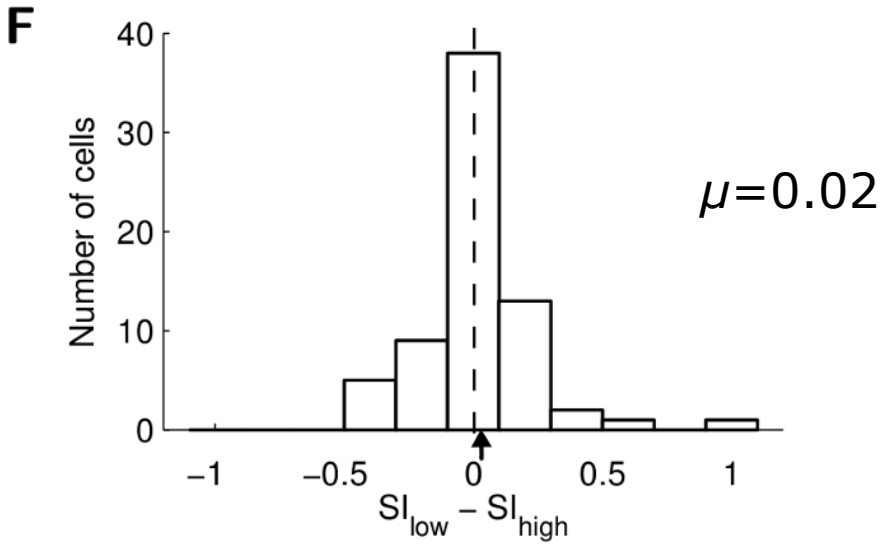
(Walker et al. 2000)



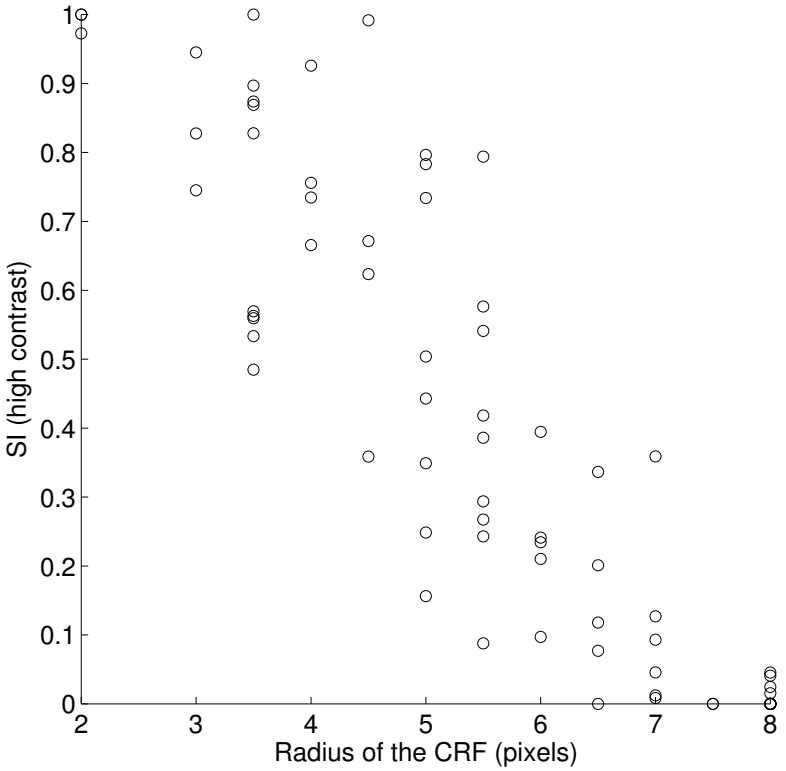
(Sceniak et al. 1999)

Model





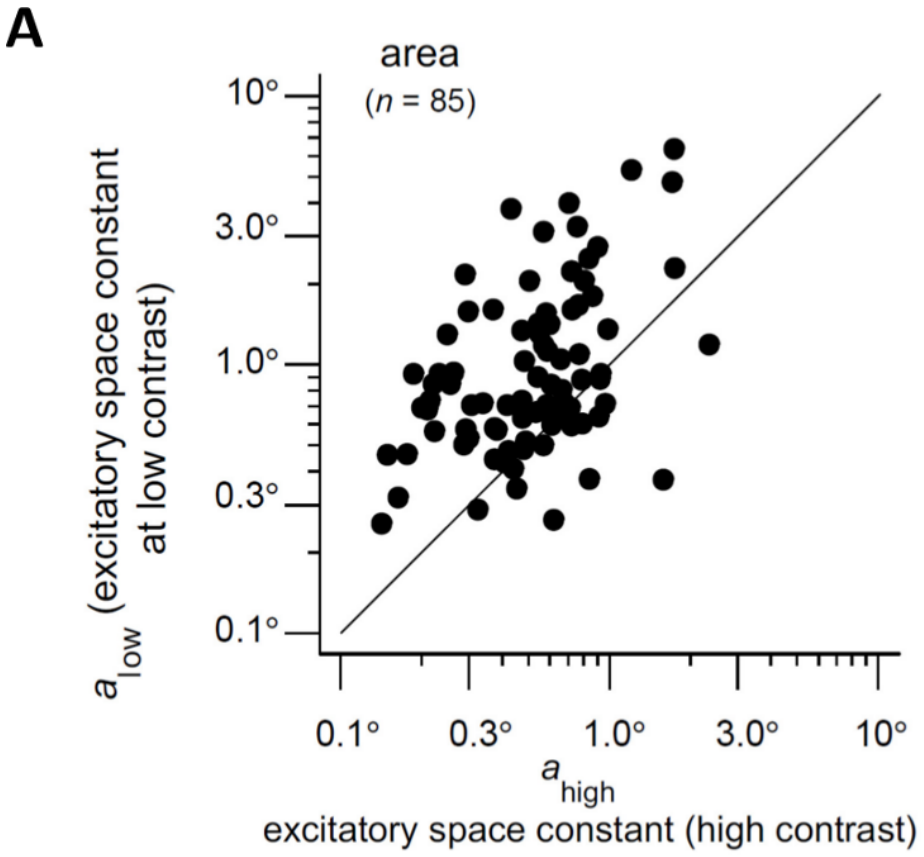
Prediction: CRF vs. SI anticorrelation



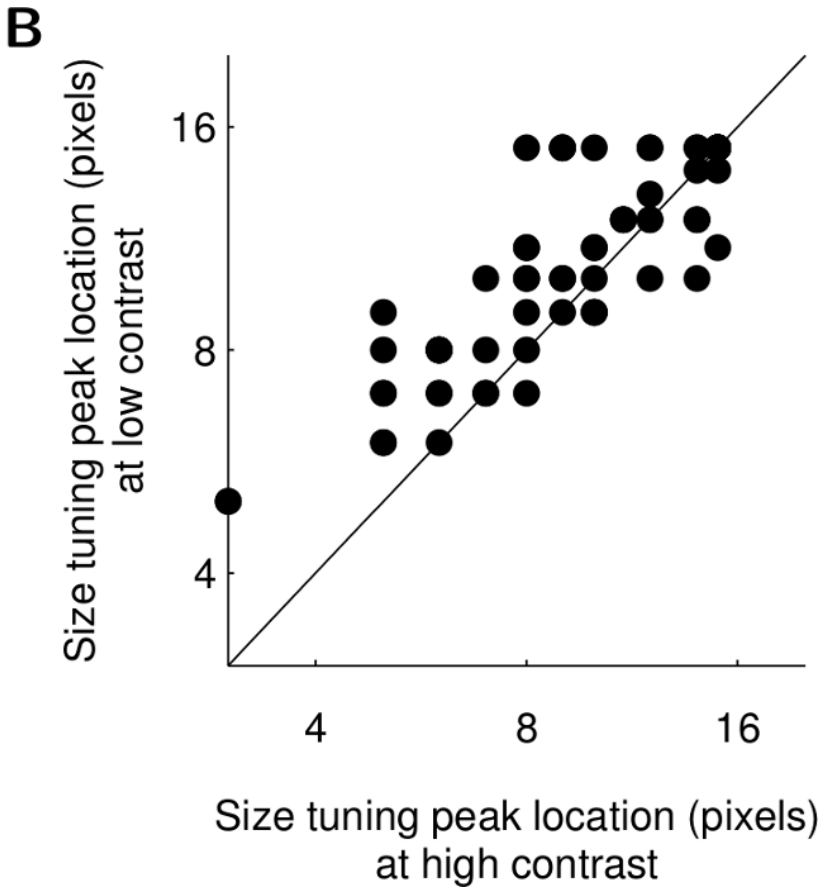
Supported! See Rose 1977 and Rose 1979

Size tuning vs. contrast

Physiology



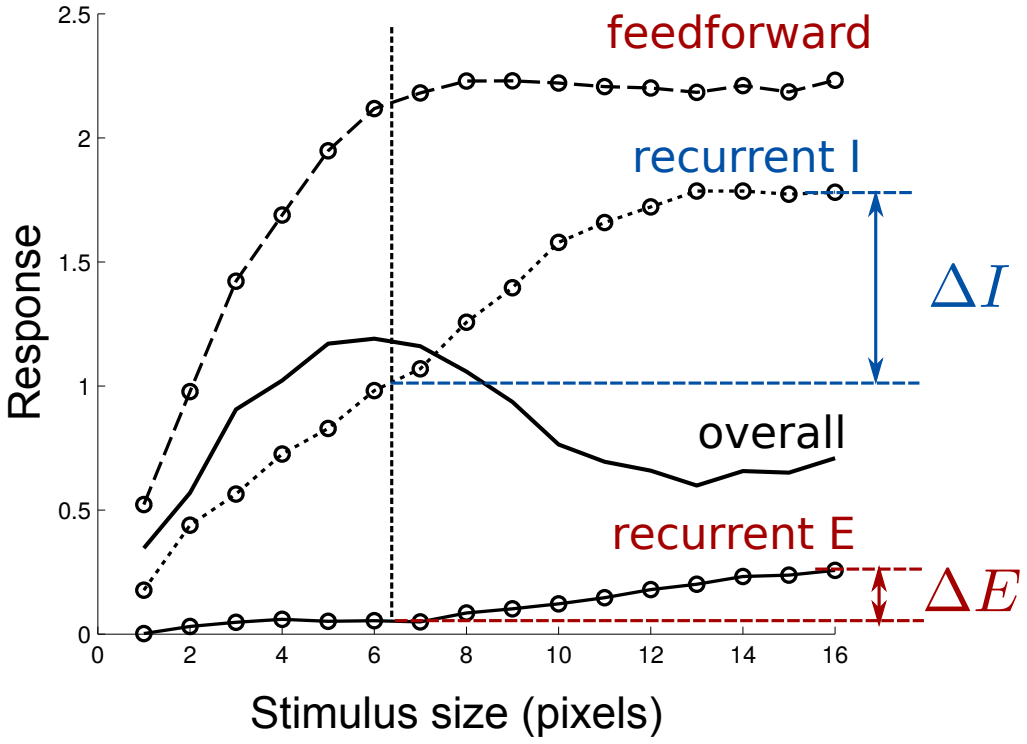
Model



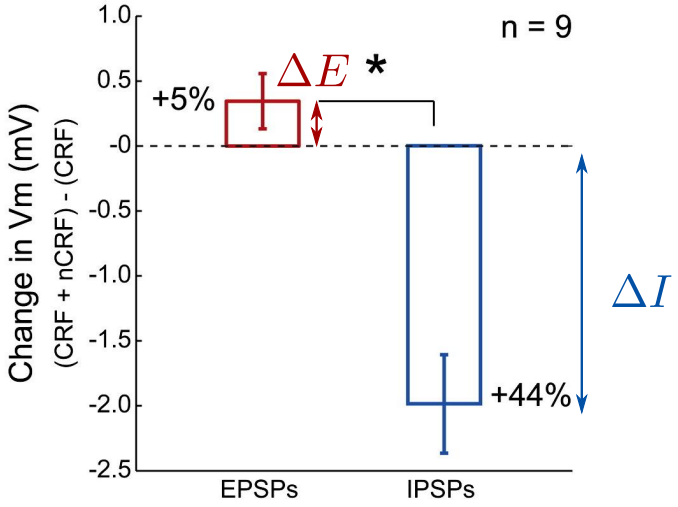
(Sceniak et al. 1999)

Mechanism: increased inhibition (~9x)

Model

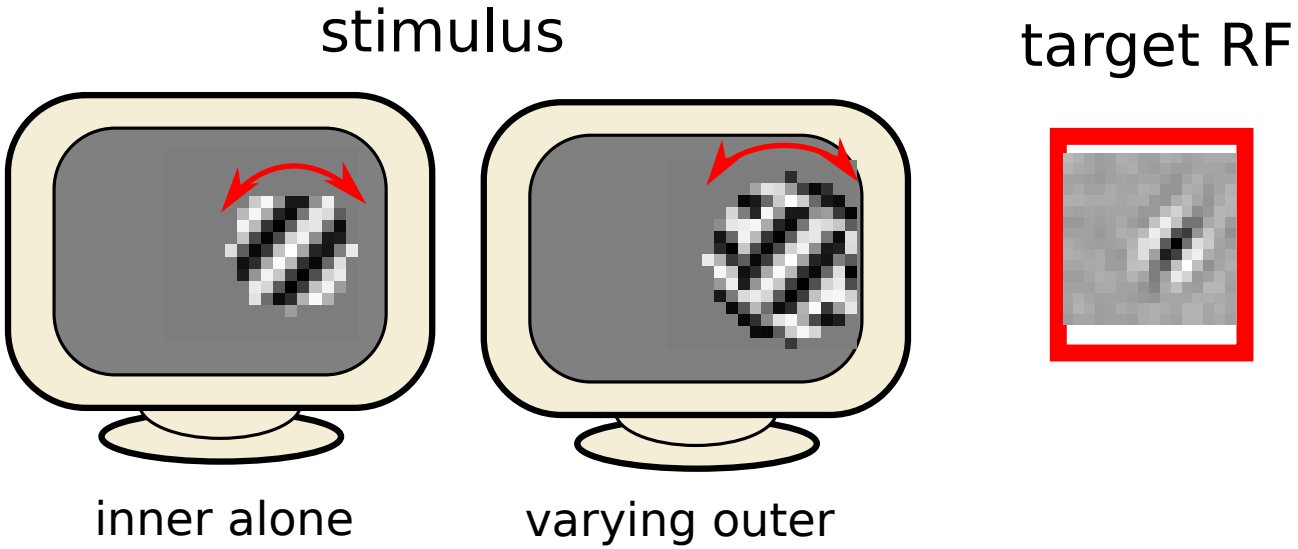


Physiology



(Haider et al. 2010)

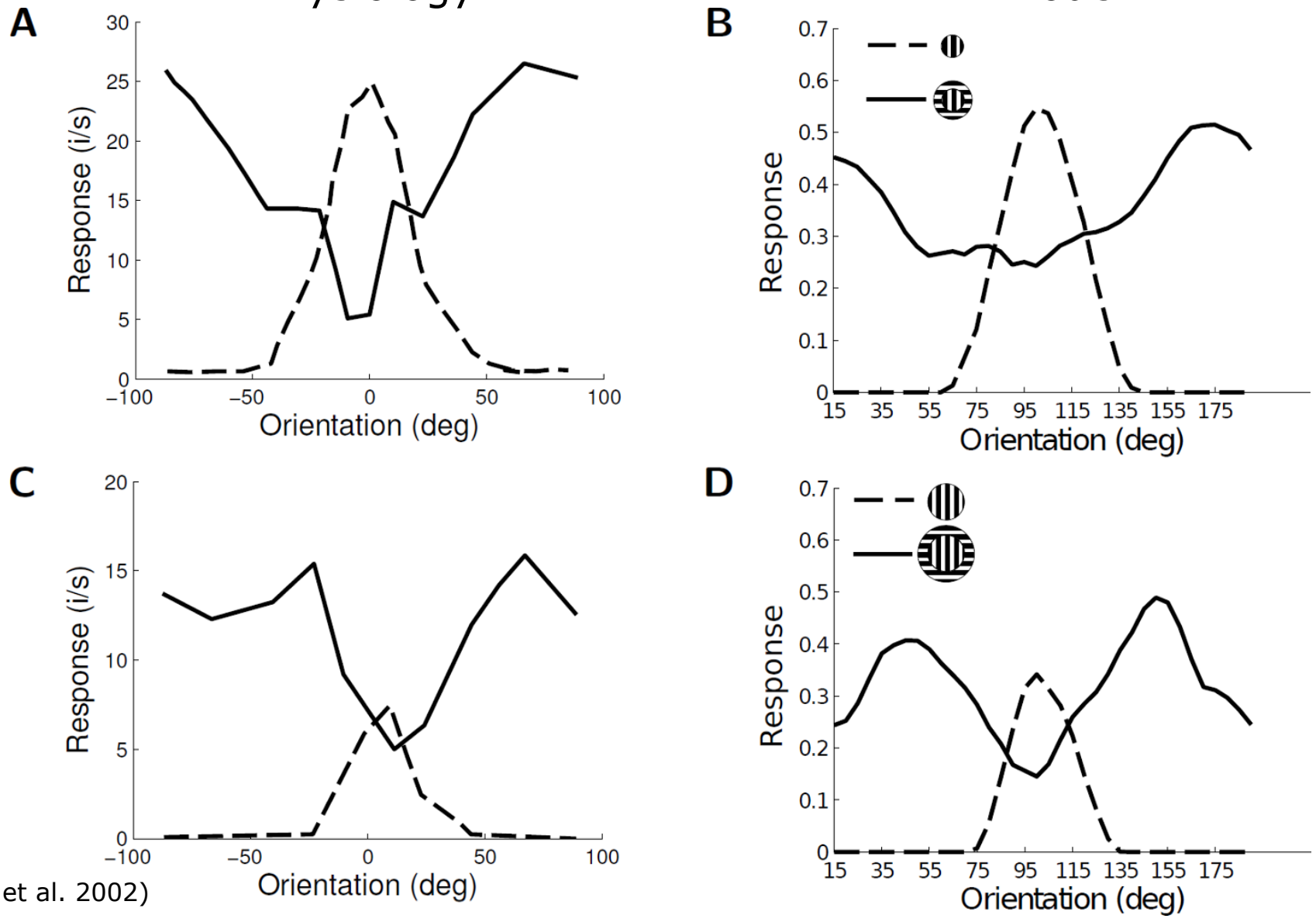
Surround orientation tuning



Surround orientation tuning

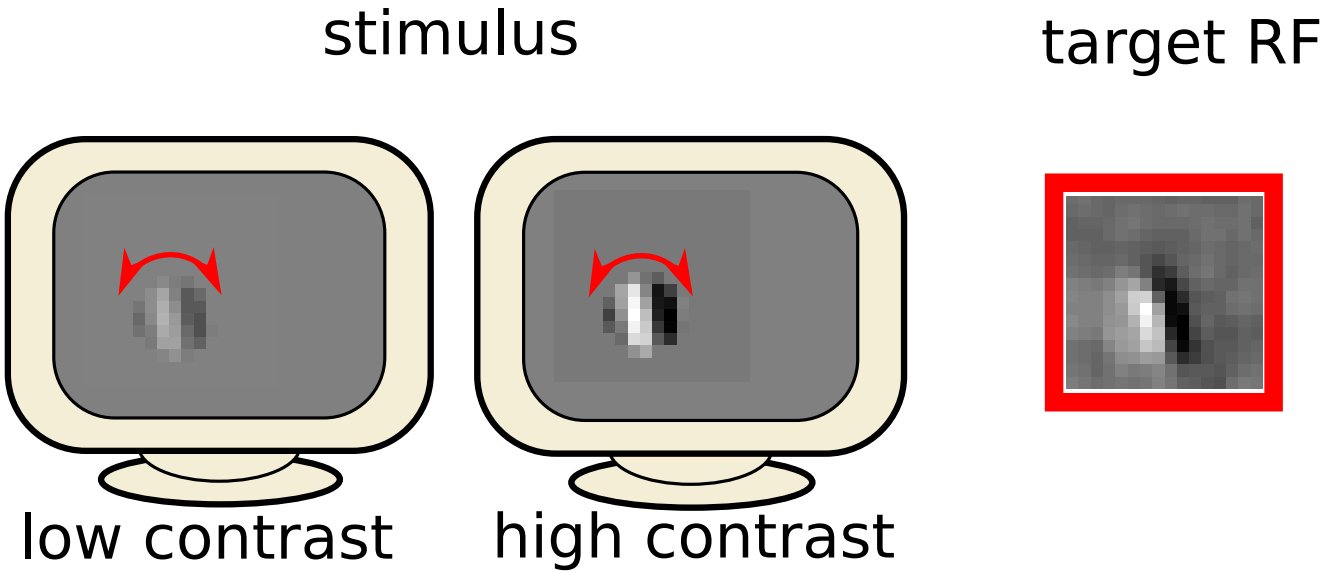
Physiology

Model



(Jones et al. 2002)

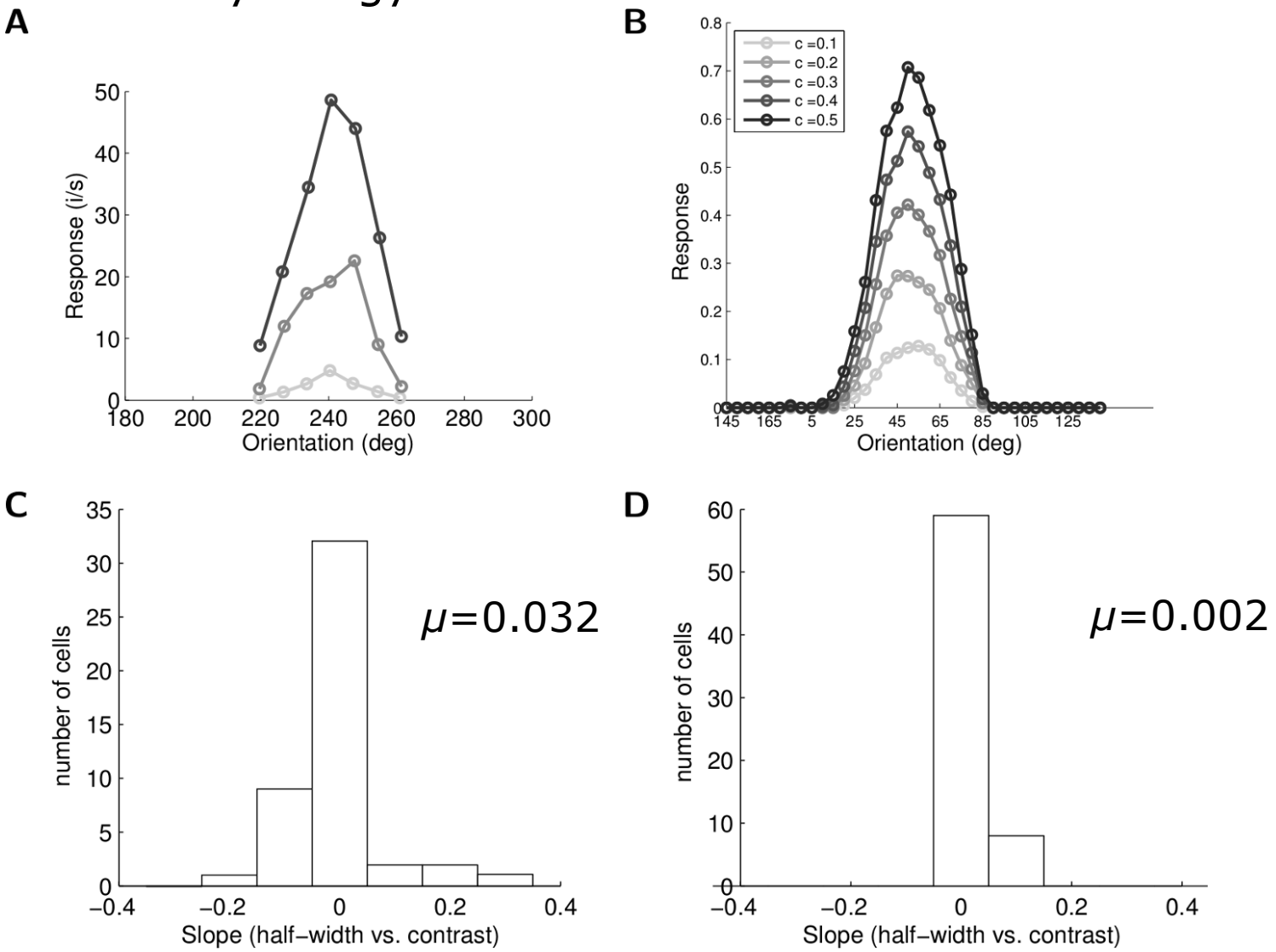
Contrast invariant orientation tuning



Contrast invariant orientation tuning

Physiology

Model



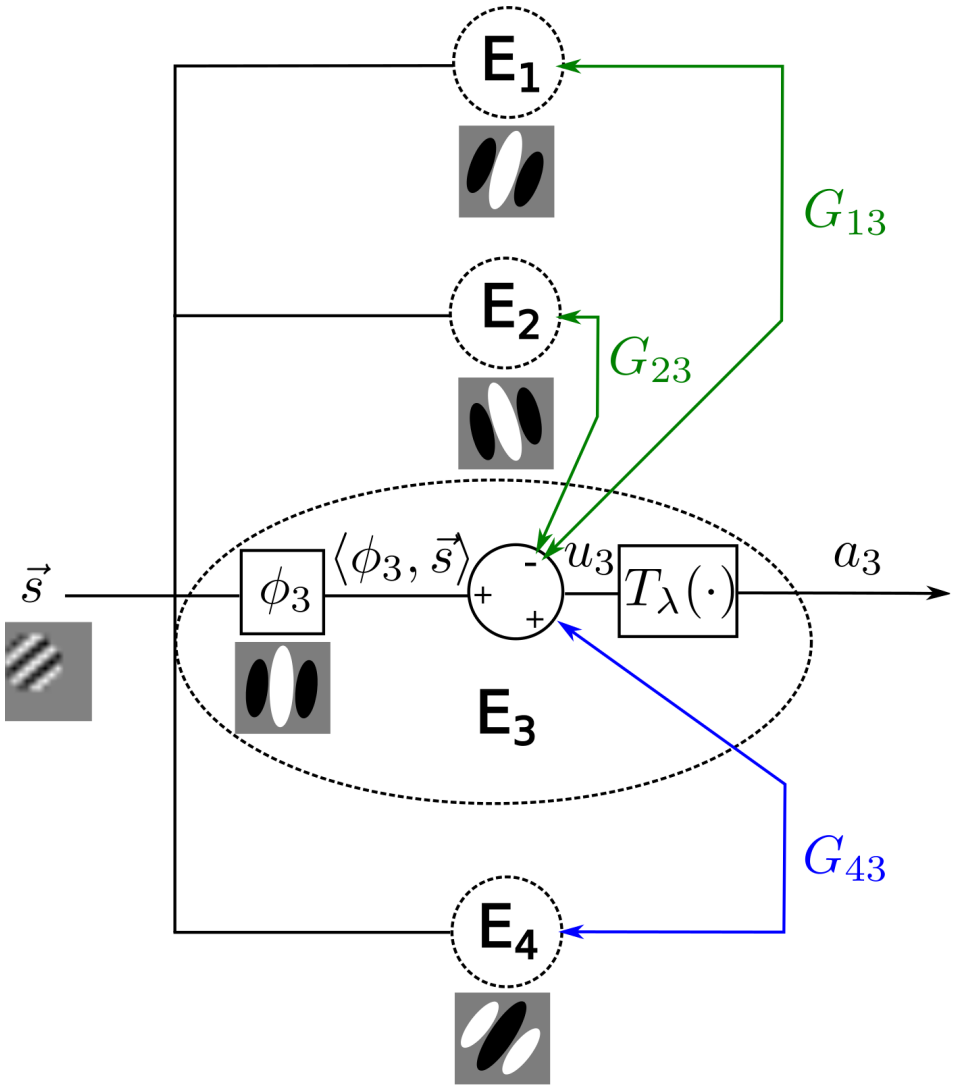
(Skottun et al. 1987; Allito & Usrey 2004)

Sparse coding response properties

- Many excitatory response properties emerge from a dynamical system implementing sparse coding

- What about inhibitory interneurons?
  - Need lateral influences via: $G = \Phi^T \Phi$
  - Dale's law -> need separate inhibitory interneurons
  - E/I ratio of ~5:1
  - Both orientation tuned and untuned cells reported

Naïve implementation



- G is **low-rank** due to dictionary and scene structure
- Use PCA to find an implementation that uses fewest inhibitory interneurons

$$G = U\Sigma V^T$$

V : presynaptic weights

U : postsynaptic weights

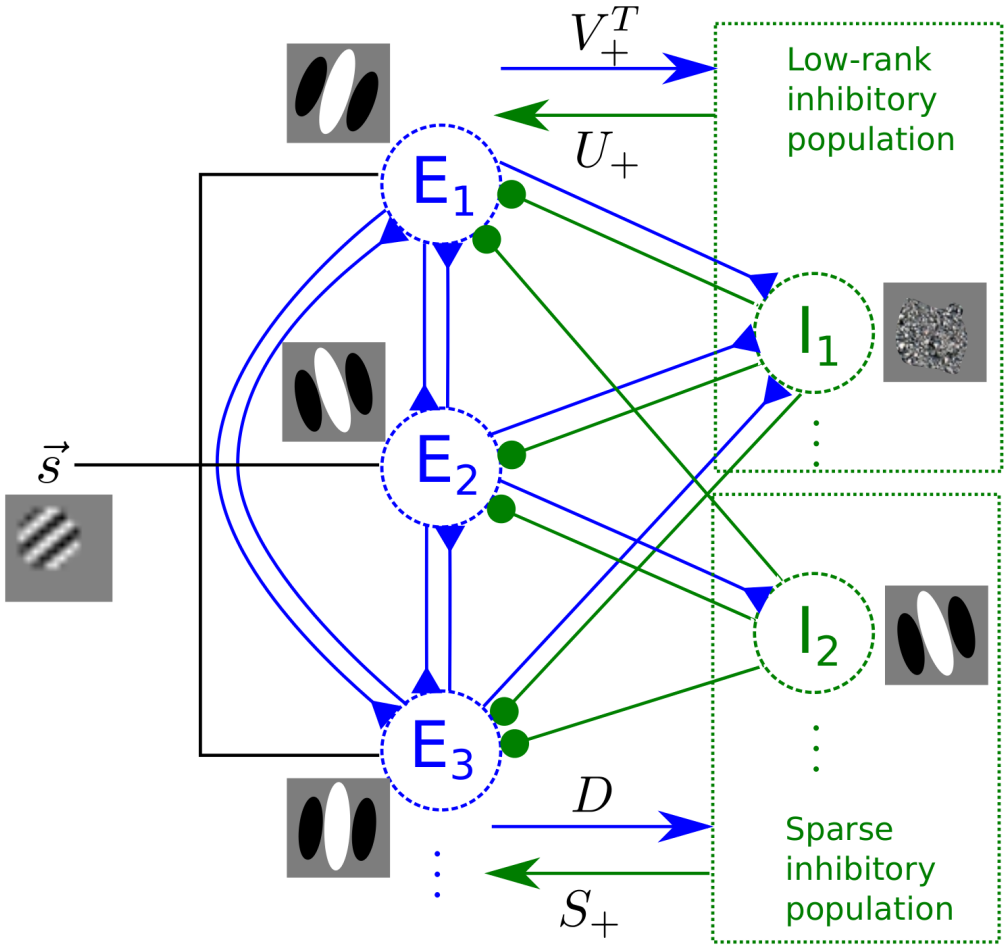
Σ : scaling/gain

- Leverage recent work on low-rank matrix factorization (robust PCA)

(Candes et al. 2011)

A tale of two inhibitions

(Zhu & R. 2015)



Encouraged by nuclear norm to min number of cells

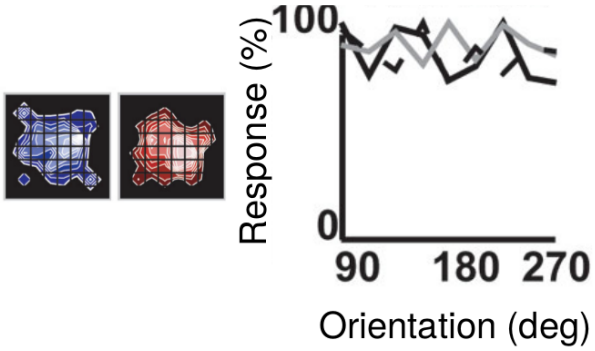
Encouraged by L1 norm to min direct inhibition

$$L, S = \arg\min_{L,S} \|L\|_* + \|\Lambda S\|_1 \quad \text{s.t. } G = L + S$$

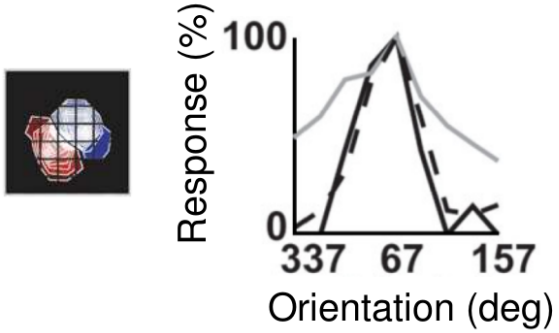
$$G = L + S = U\Sigma V^T + S = (U_+ + U_-)\,\Sigma\left(U_+^T + V_-^T\right) + (S_+ + S_-)$$

Untuned vs. tuned cells

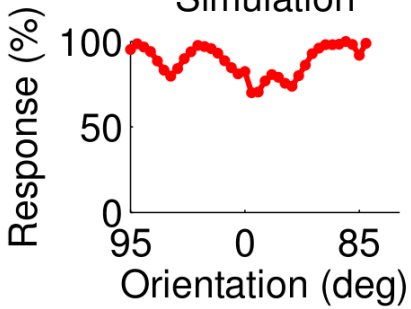
Physiology
(Hirsch et al. 2003)



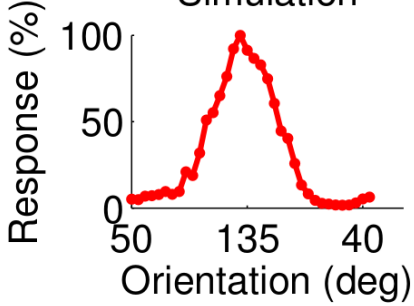
Physiology
(Hirsch et al. 2003)

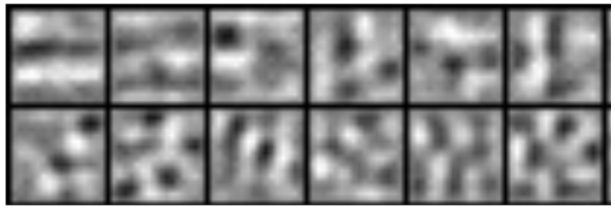


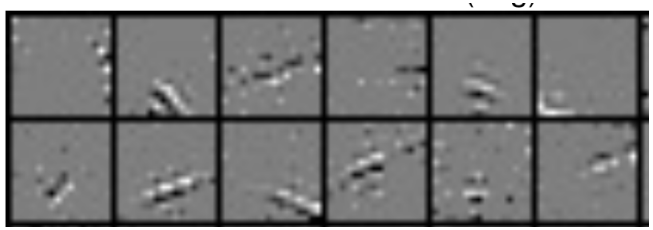
Simulation



Simulation







Optimal E:I ratio

- Ratio of cortical E:I cell types consistent in relatively narrow range

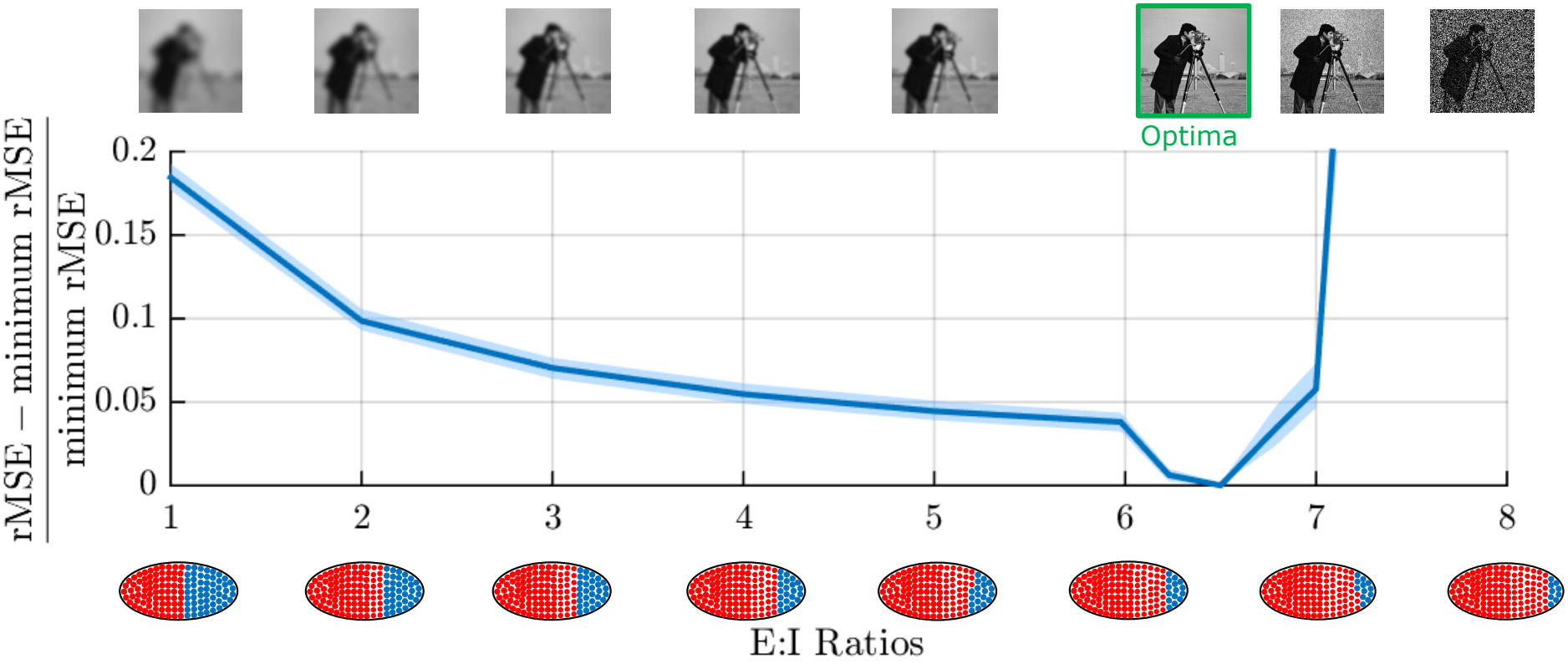
Species	A1	V1	S1
Rodent	9:1	9:1	8.5:1
Cat	3:1	4:1	2.35-3.42:1
Primate	3:1	3:1	3:1

- Why not 1:1 or 20:1? What computational principles govern cortical E:I ratios?
- Our approach: volume constraint=fixed population size
- E:I low
  - High accuracy in desired computation
  - Network sacrifices representational resolution
- E:I high
  - High capacity to represent input stimuli
  - Network sacrifices computational resolution

(Alreja, Nemenman & R. in prep.)

Optimal E:I ratios

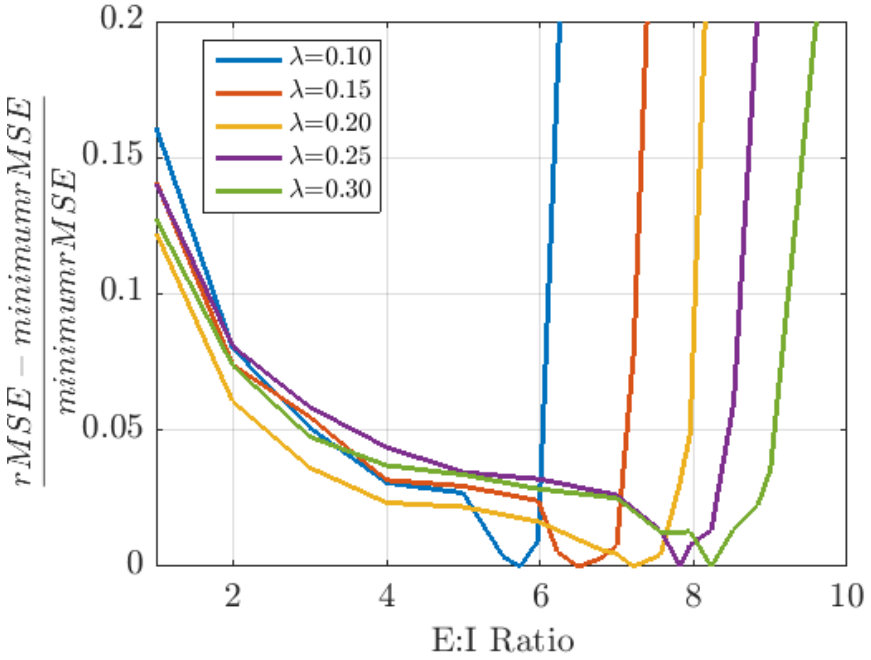
- Find approximation to sparse coding network for I size set by fixed population size and variable E:I ratio

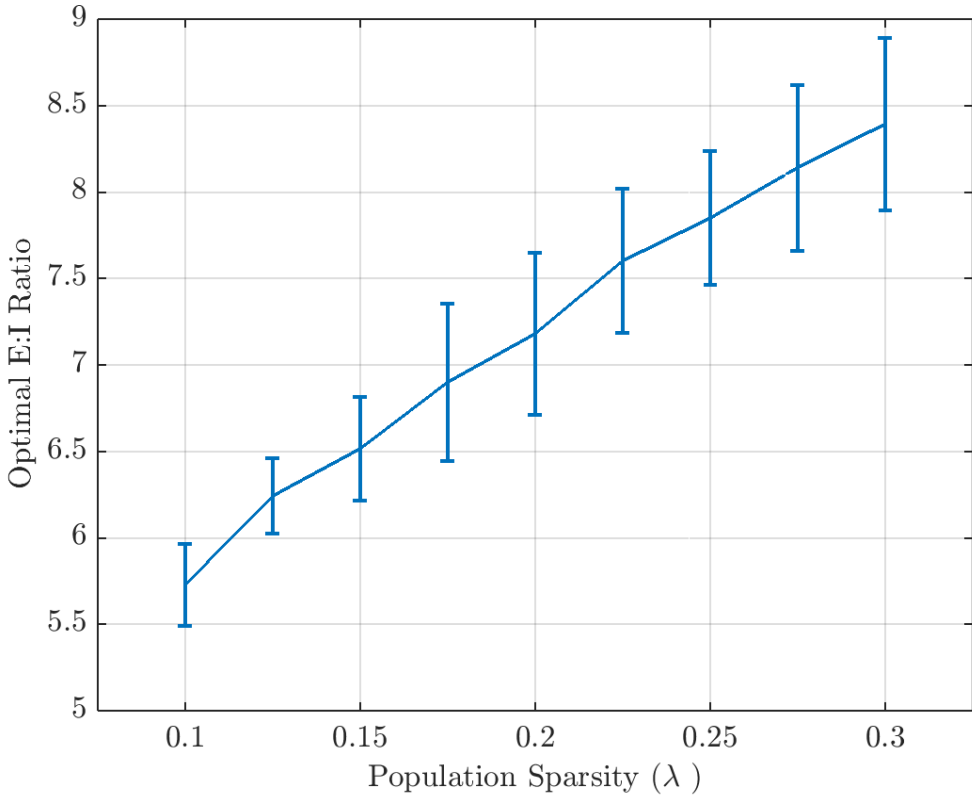


(Alreja, Nemenman & R. in prep.)

Optimal E:I ratios

- Same optima for total sparsity (Treeves-Rolls metric) and metabolic cost of the excitatory cells
- More sparse -> higher optimal E:I (some support)



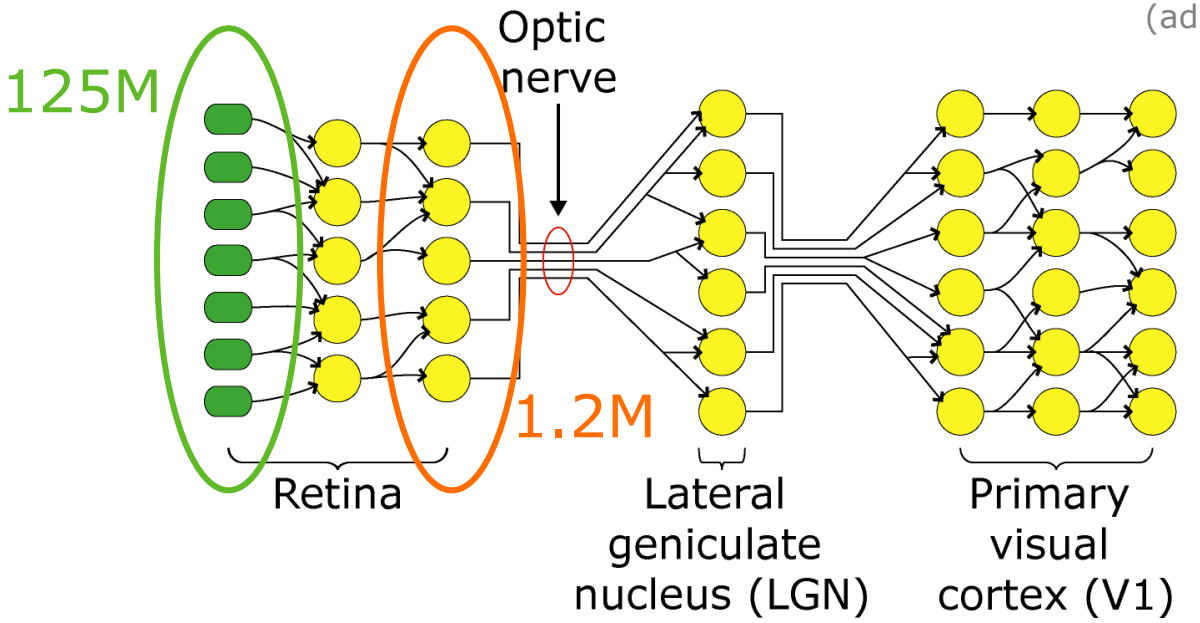


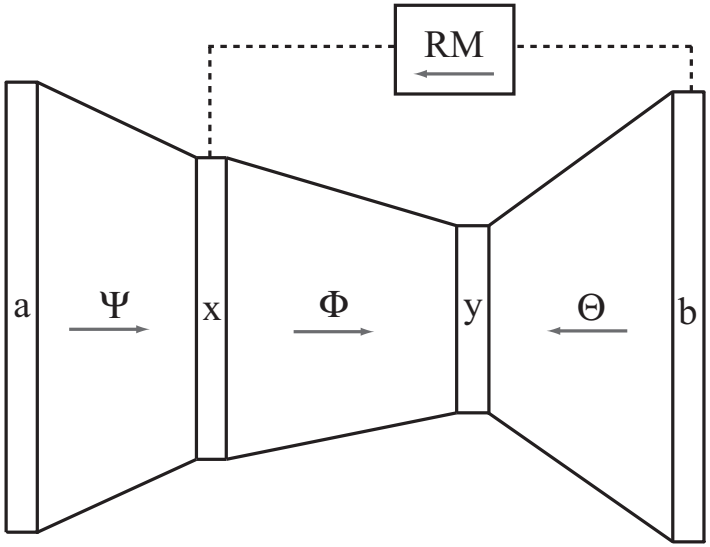
More sparsity →

(Alreja, Nemenman & R. in prep.)

Dimensionality reduction in learning

(adapted from Hubel 1988)



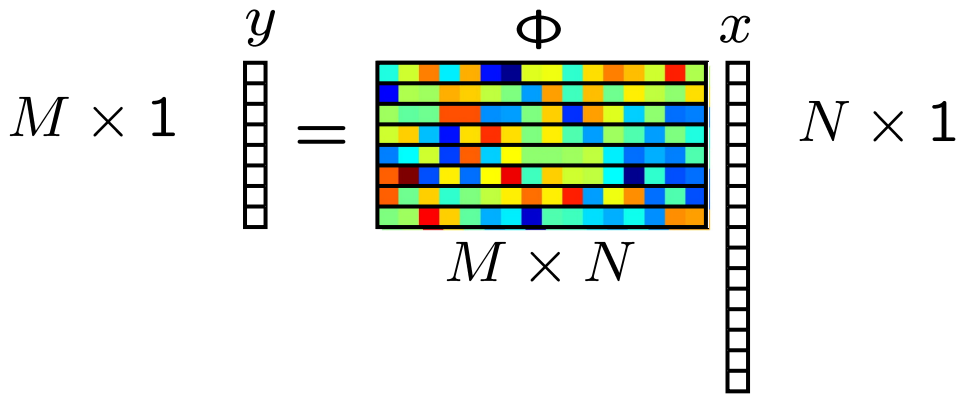


a: coefficients
x: image
y: reduced image
Θ: learned dictionary

(Isley et al. 2009)

Compressed Sensing

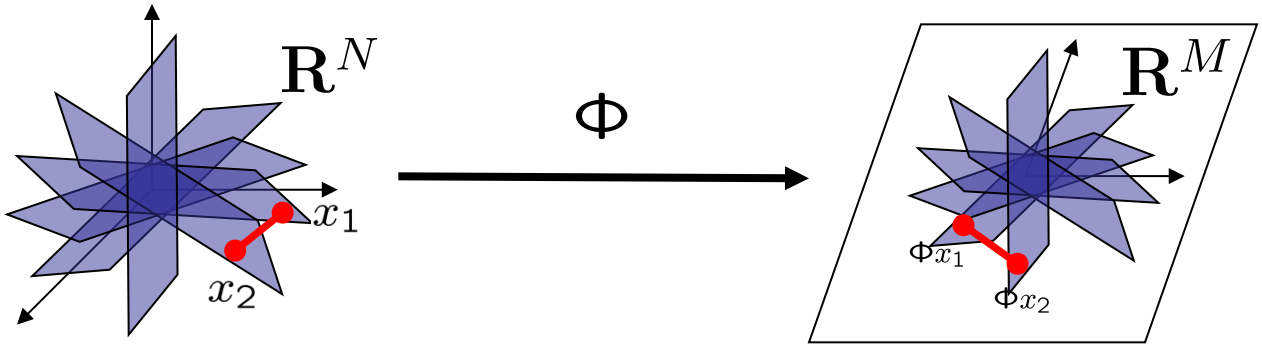
- Randomized dimensionality reduction: $y = \Phi x + w$



- Stable embedding over a signal class: $x_1, x_2 \in \mathcal{S}$

$$(1-\delta) \leq \|\Phi x_1 - \Phi x_2\|_2^2 \,/\, \|x_1 - x_2\|_2^2 \leq (1+\delta)$$

- Ex: Restricted Isometry Property (RIP) for *S*-sparse signals



Gaussian random matrix

$$M \geq O\left(\frac{S}{\delta^2}\log(N/S)\right)$$

(Candès 2006)

Compare dimensionality reduction

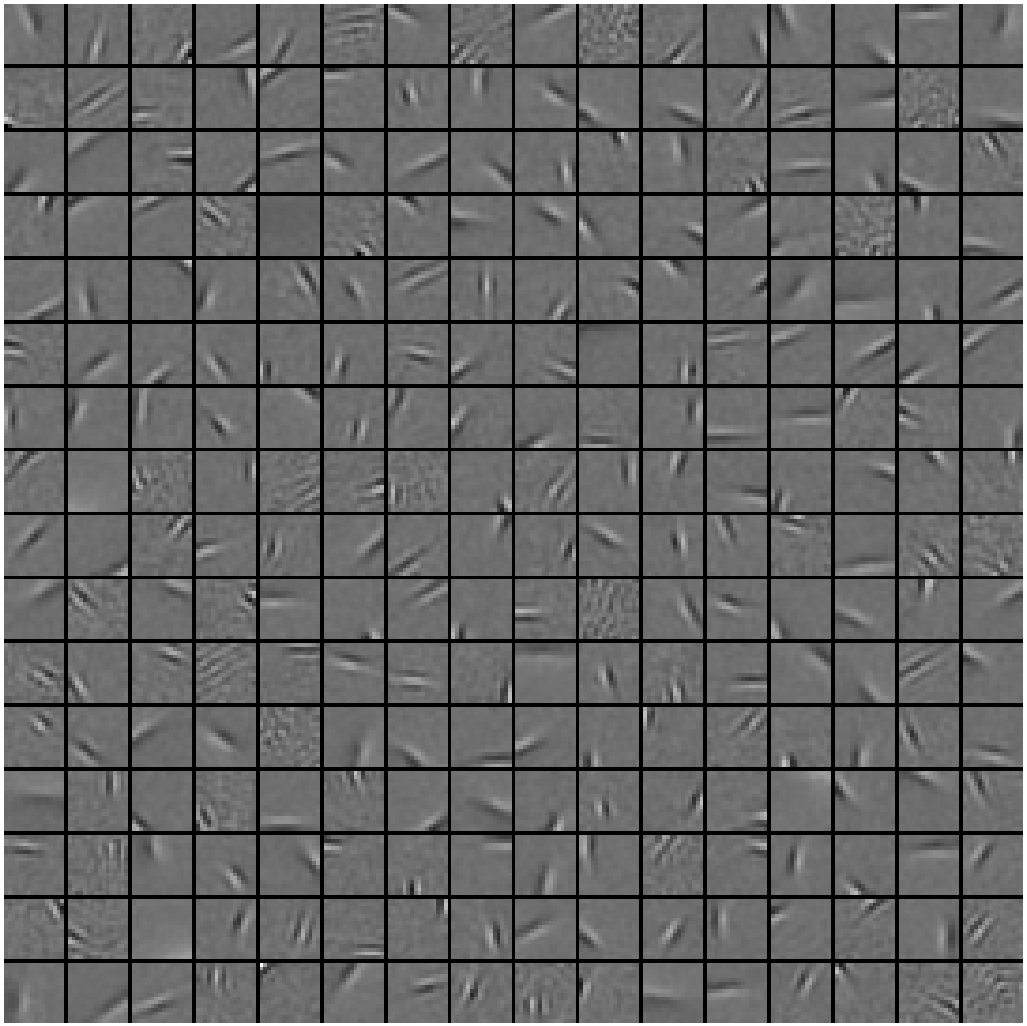
- Principal component analysis (PCA)
  - Data adaptive (expensive)
  - Linear encoding and decoding (cheap)
  - No signal model (decorrelates -> ideal for Gaussian)
  - Maximize variance; minimize approximation error

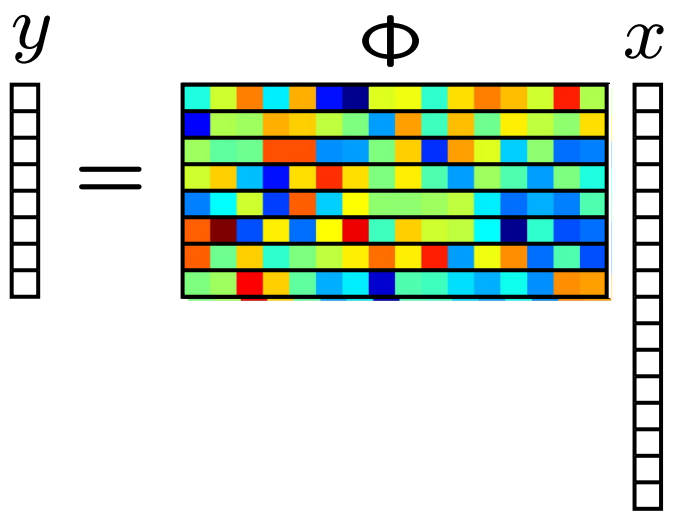
- Compressed sensing (CS)
  - Data non-adapative (cheap)
  - Linear encoding (cheap)
  - Nonlinear decoding (expensive)
  - Low-dimensional signal model (sparsity, manifold, etc.)
  - Retain geometry of signal class

- Conceptualized as useful for expensive data acquisition

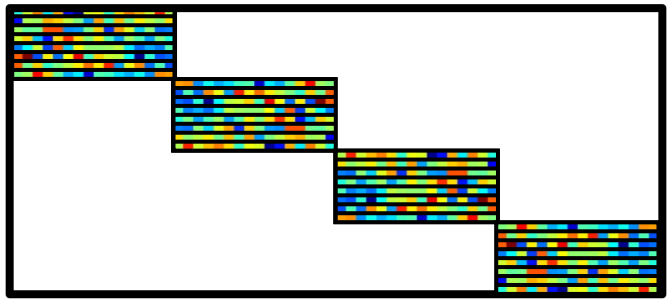
Learning in compressed space

- Model bottlenecks as random dimensionality reduction (Isley, et al. 2009)
- N=16x16=256 and M=128
- Plot recovered RFs





Requires global wiring!

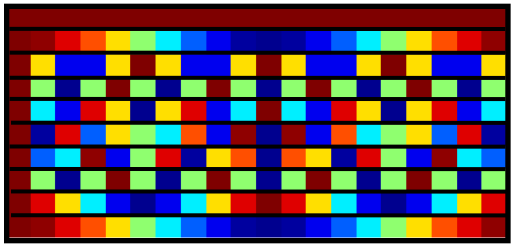


What about local wiring?

Structured Matrices in CS

- Subsampled Fourier matrices

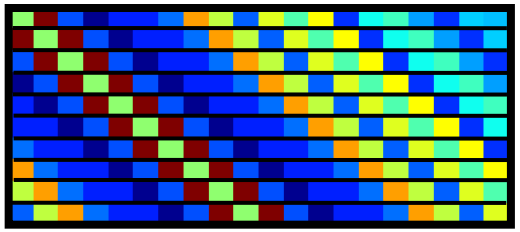
$$\mathrm{RIP}-(S,\delta) \Leftrightarrow M \geq O\left(\frac{S\log^3(S)}{\delta^2}\log N\right)$$



(Rudelson and Vershynin, 2008)

- Partial circulant matrices

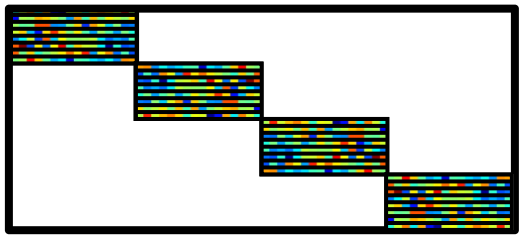
$$\mathrm{RIP}-(S,\delta) \Leftrightarrow M \geq O\left(\frac{S\log^2(S)}{\delta^2}\log^2 N\right)$$



(Krahmer et al., 2014)

- Block diagonal matrices (J blocks)

$$\mathrm{RIP}-(S,\delta) \Leftrightarrow M \geq O\left(\frac{S\log^2(S)}{\delta^2}\mu^2\log^2 N\right)$$

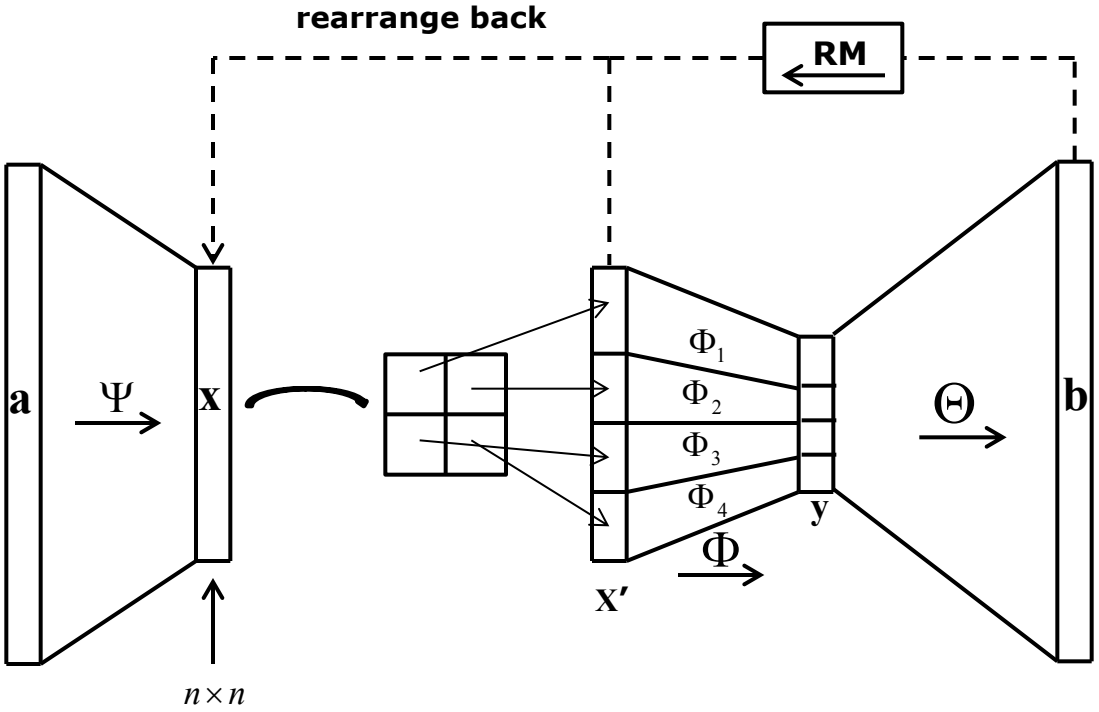


(Yap, Eftekhari, Wakin, & R., 2015)

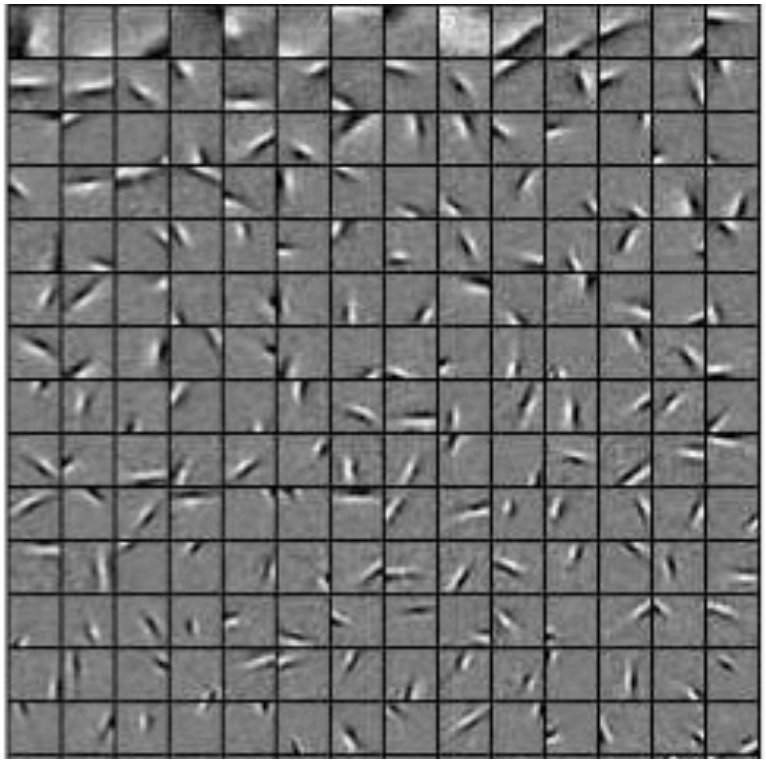
- Coherence μ measures similarity to canonical basis
  - Fourier: $\mu^2 = 1$
  - Canonical: $\mu^2 = J$

Global vs. local wiring

- Break image up into 16 equal blocks and then compress locally

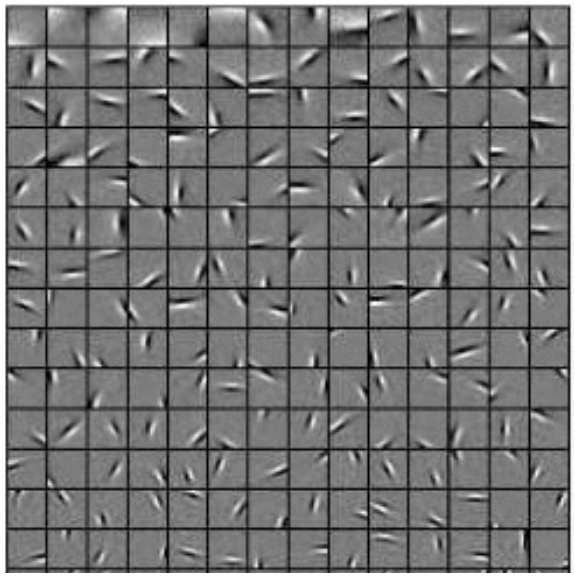


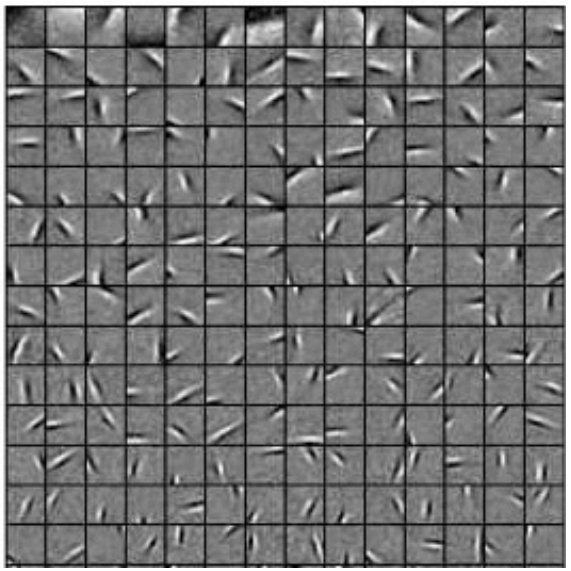
(Liu & R., in prep)

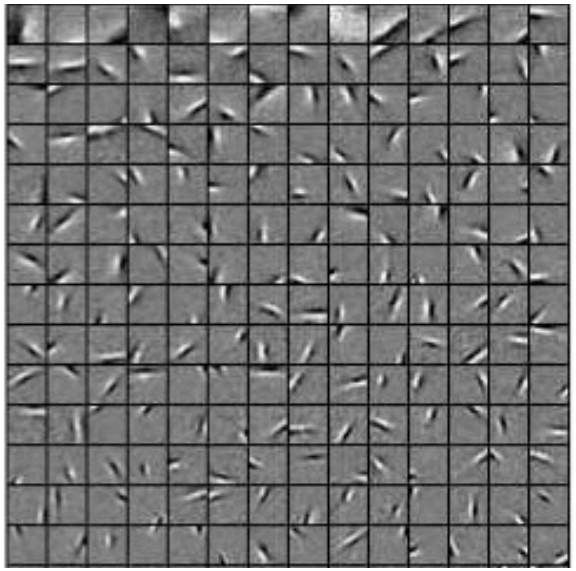


Compression does not affect RF properties

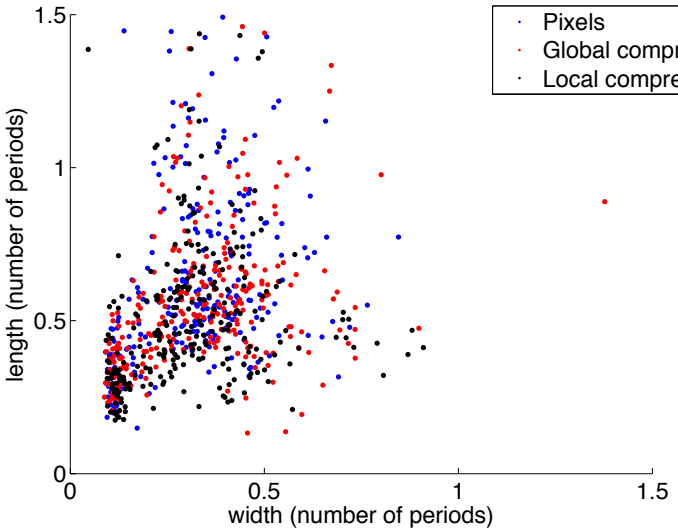
No compression | Gaussian matrix (J=1) | Local synapses (J=16)

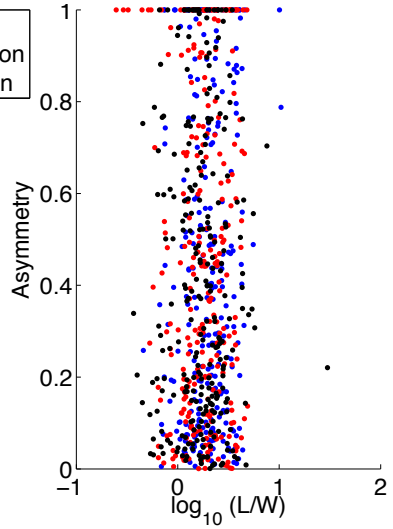


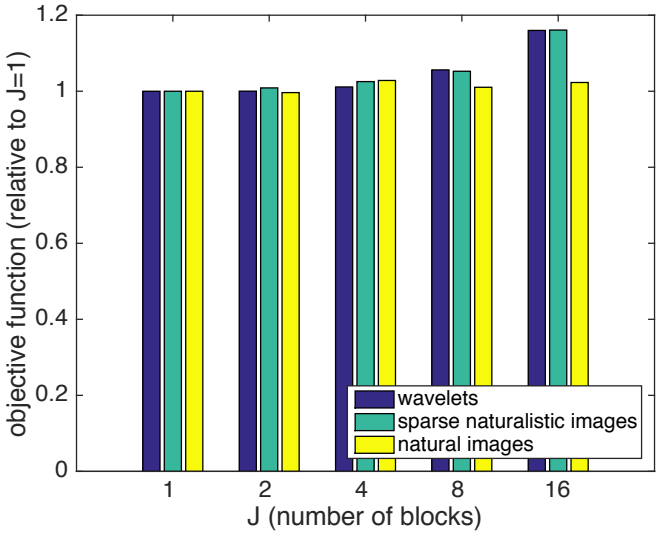




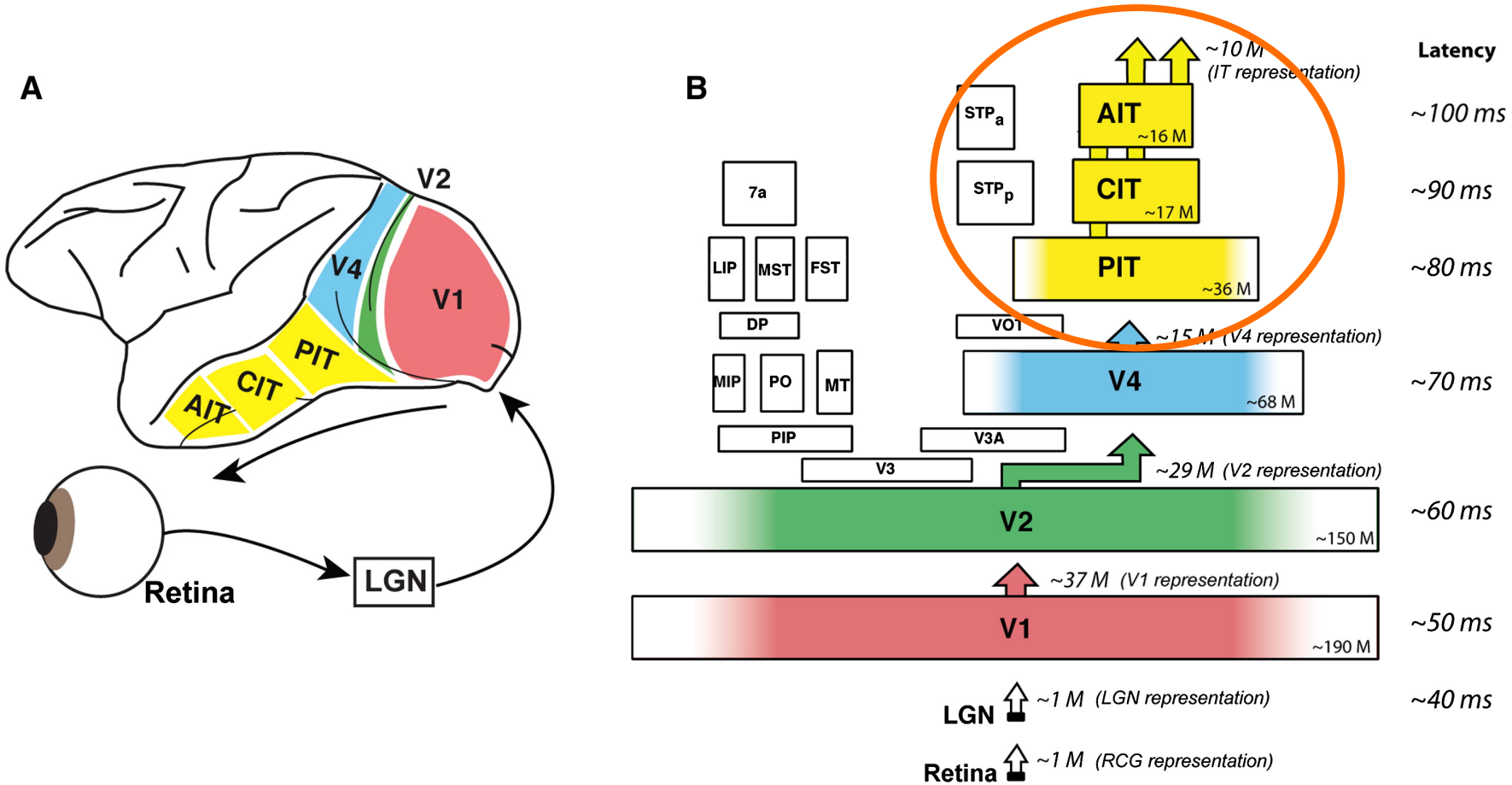
(Liu & R., in prep)





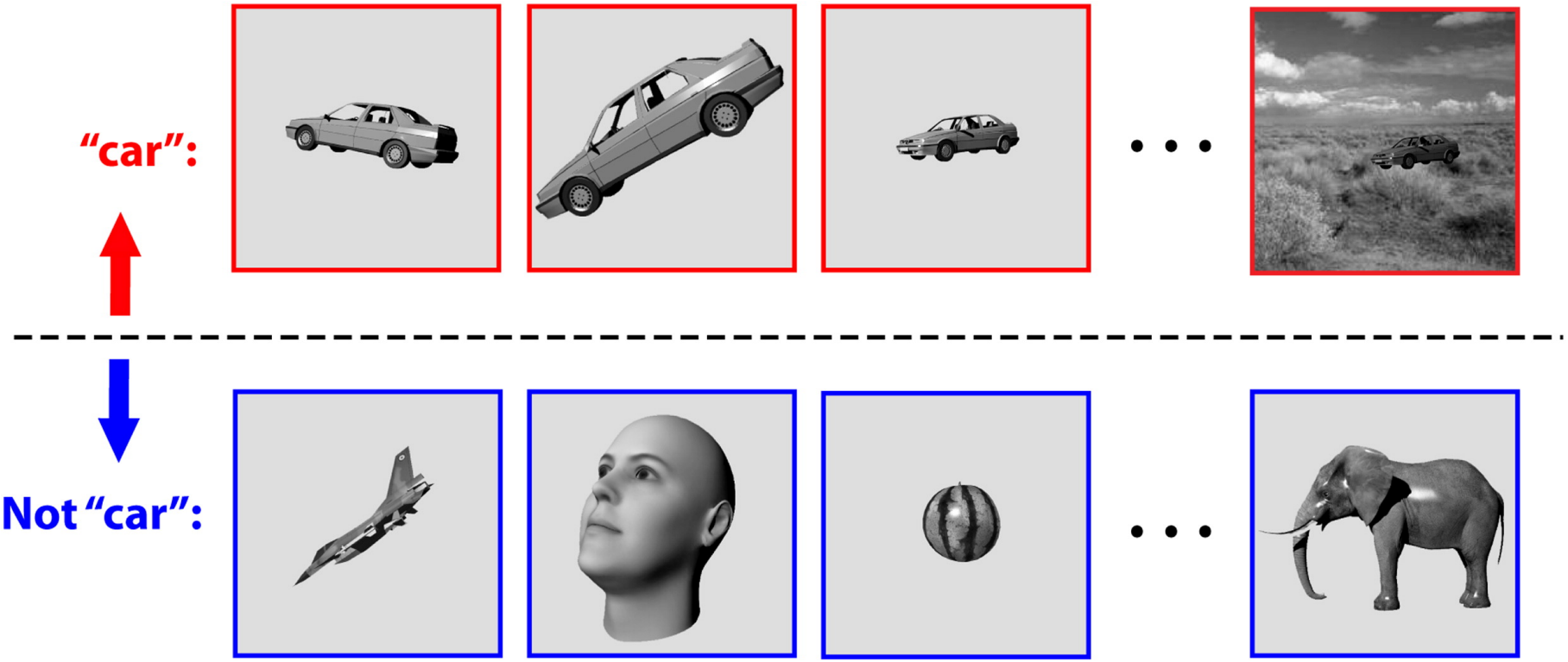


Visual pathway



(DiCarlo, Zoccolan & Rust, 2012)

Selectivity vs. generality in perception

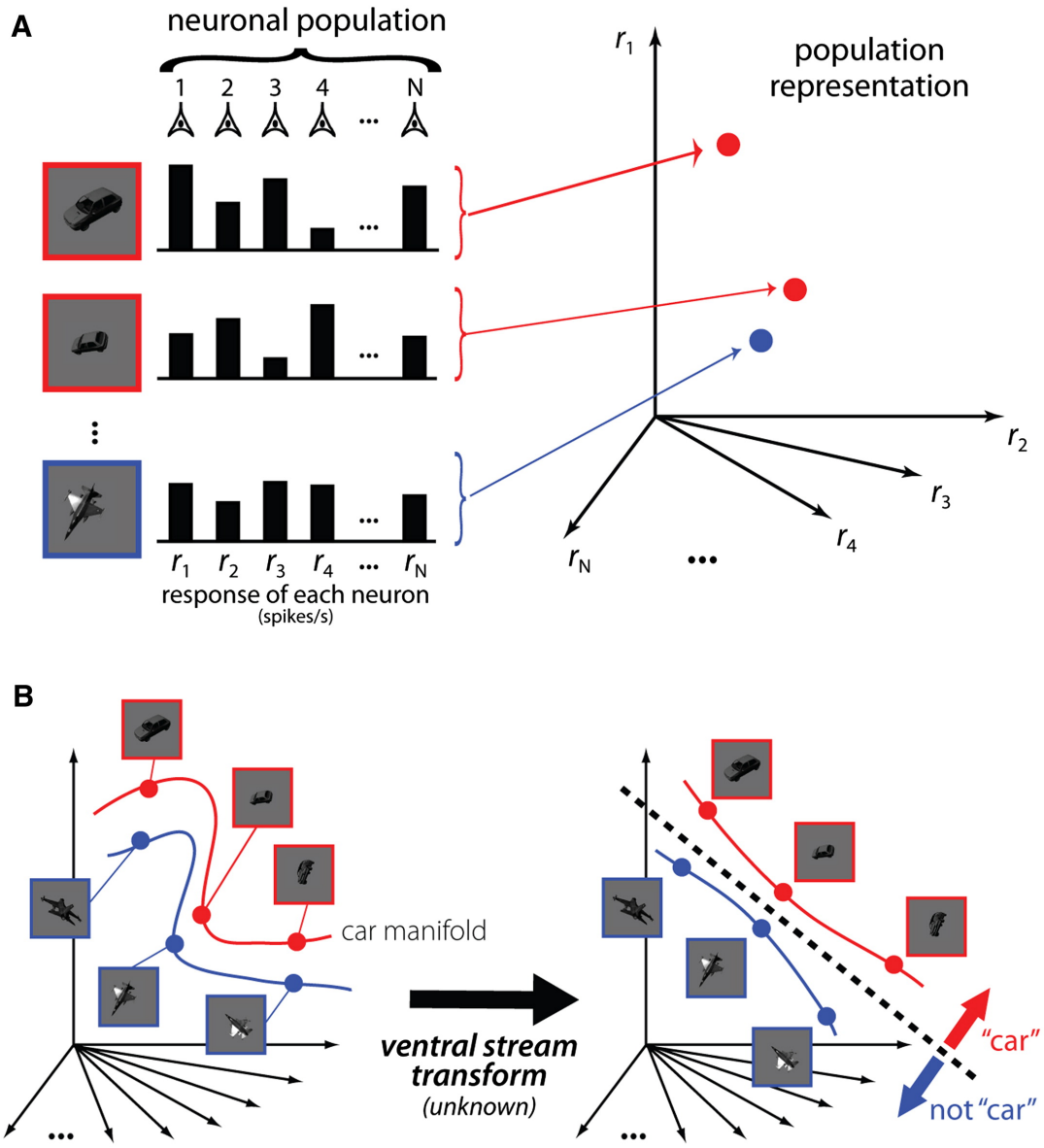


(DiCarlo Zoccolan & Rust 2012)

Scale invariances

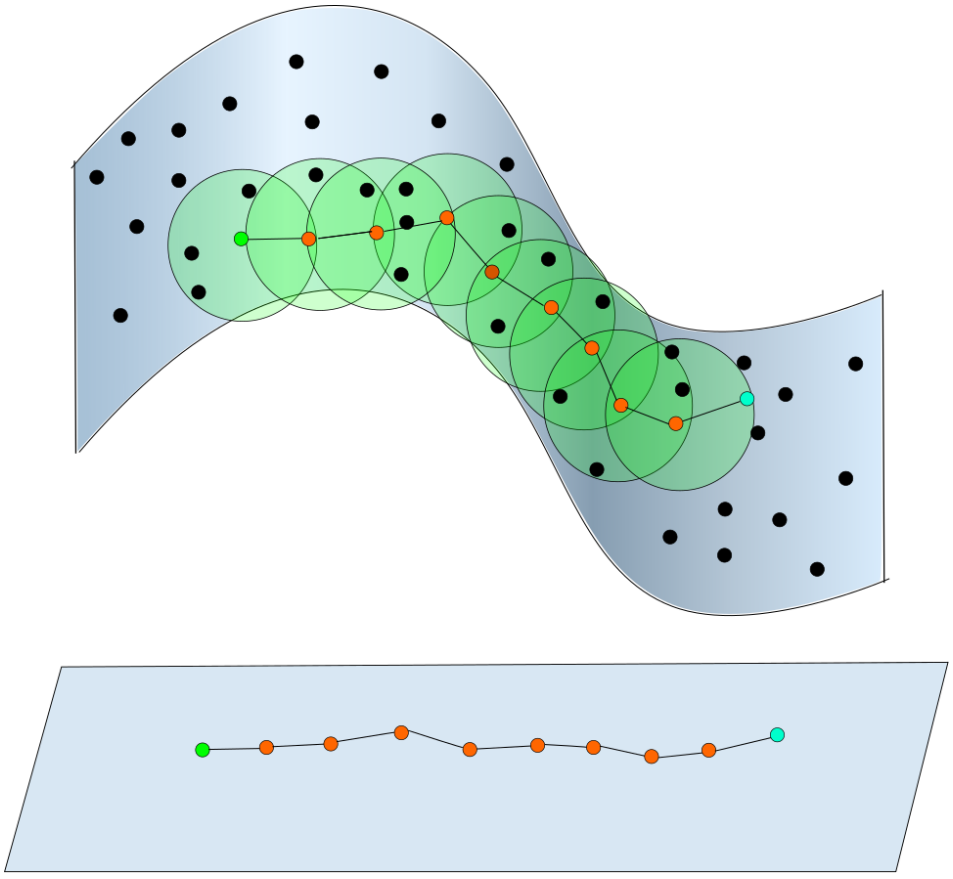
(DeLoache et al. 2004)

Manifolds of invariant representations



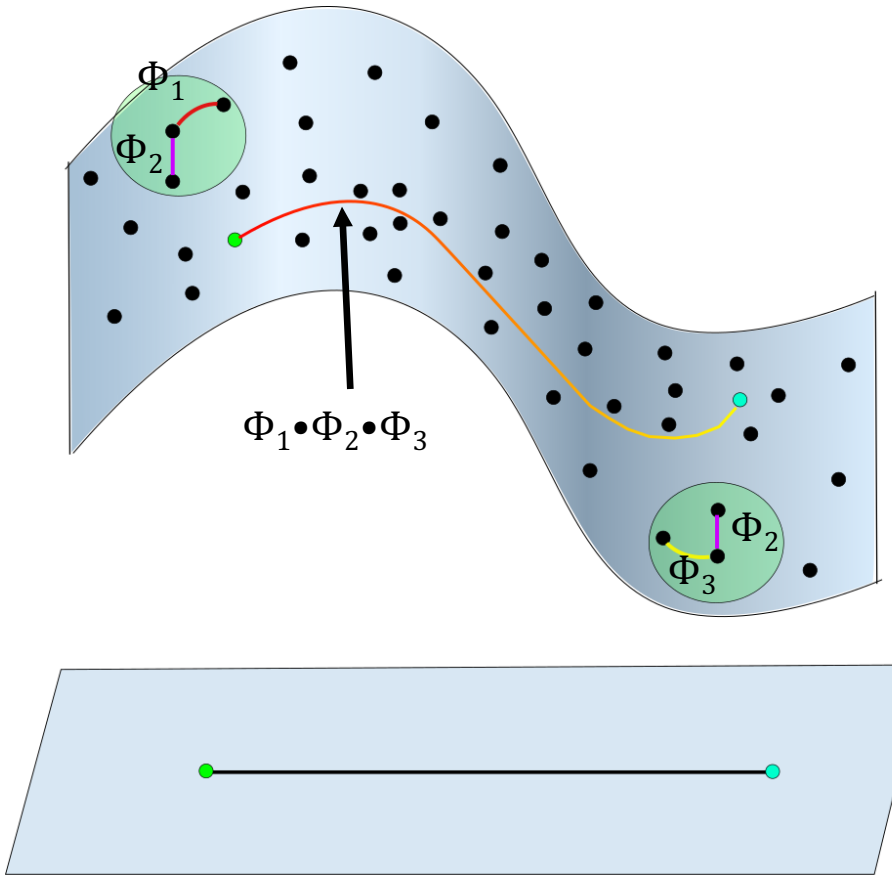
(DiCarlo Zoccolan & Rust 2012)

Manifold "learning"



Isomap

(Tenenbaum, de Silva & Langford 2000)



Manifold Transport Operator

(Culpepper & Olshausen 2009)

- Can we learn analytic operators that capture movement primitives along manifold structures?

Transform "structure from motion"

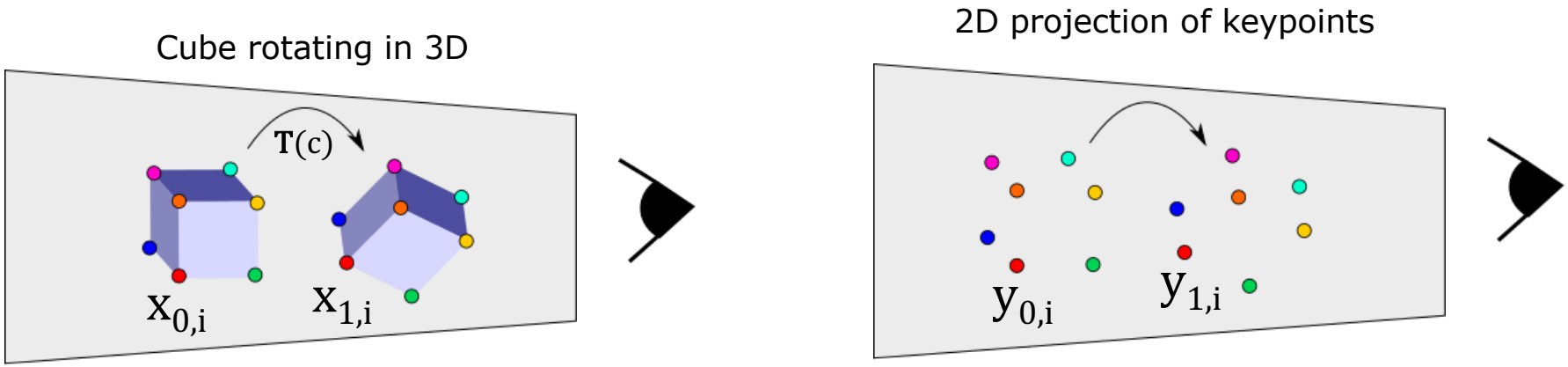


Image generation model

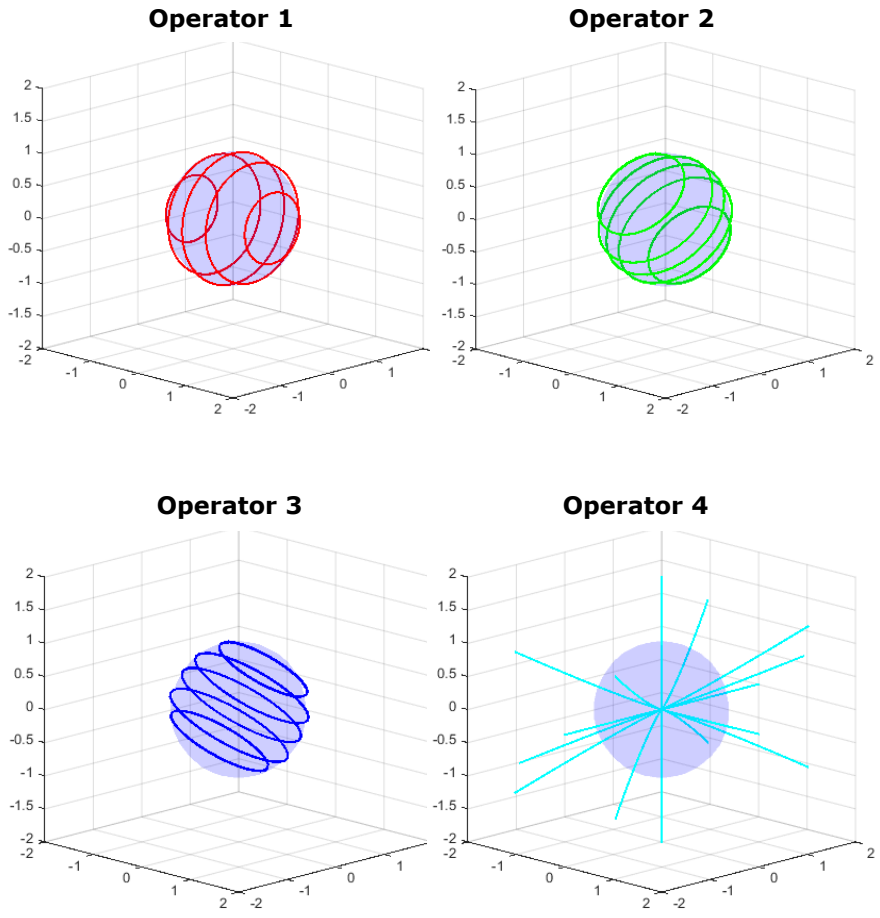
$$x(n+1) = e^{A(n)}x(n) + w$$

$$A(n) = \sum_i \phi_i c_i(n)$$

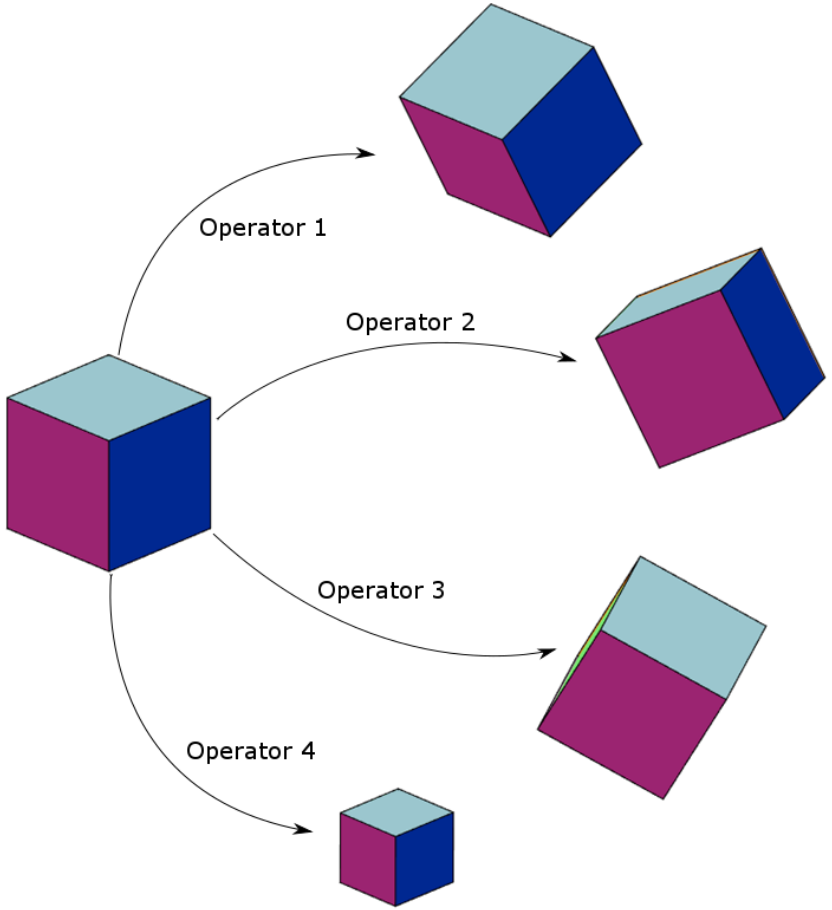
$$y(n) = Kx(n) \quad \text{(orthoprojector)}$$

- Points in 3D are constrained by 2D with unknown depth
- Infer unknowns for data samples and learn operators ϕ_i
- Can apply sparsity and/or dynamic penalty on c_i

Learned 3D transport operators

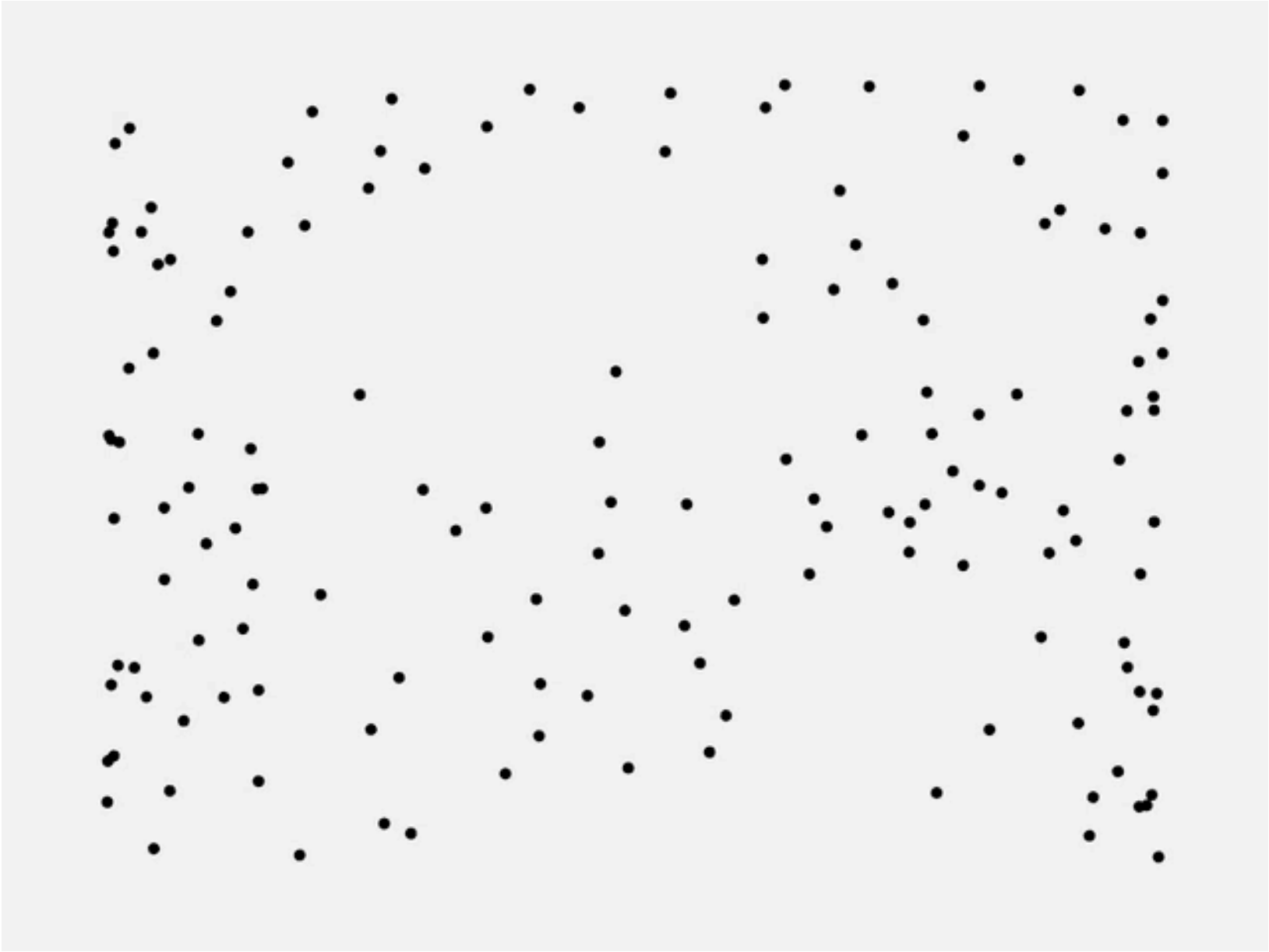


Generated transforms

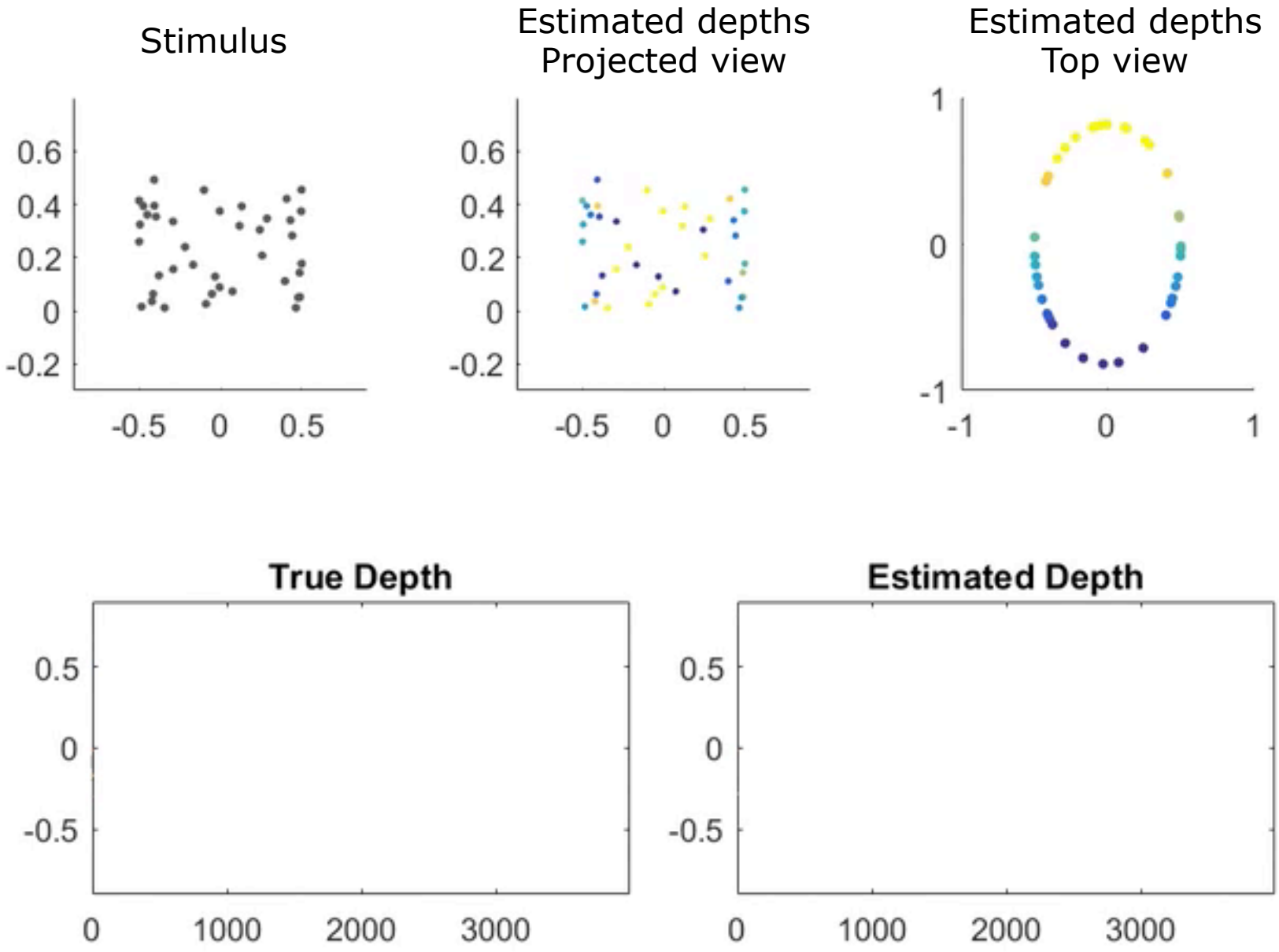


(Connor & R., in prep)

Kinetic depth effect



3D depth inference



http://siplab.gatech.edu

crozell@gatech.edu