Towards bridging the gap between neuroscience and artificial intelligence

Surya Ganguli

Dept. of Applied Physics,
Neurobiology,
and Electrical Engineering

Stanford University

Funding:

Bio-X Neuroventures
Burroughs Wellcome
Genentech Foundation
James S. McDonnell Foundation
McKnight Foundation
National Science Foundation

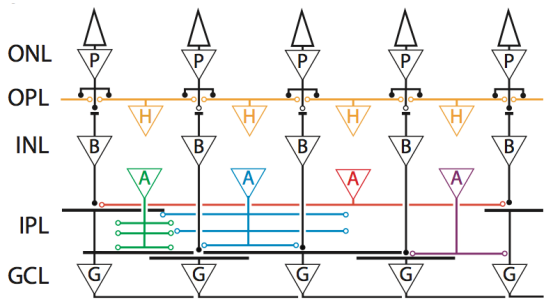
NIH
Office of Naval Research
Simons Foundation
Sloan Foundation
Swartz Foundation
Stanford Terman Award

http://ganguli-gang.stanford.edu

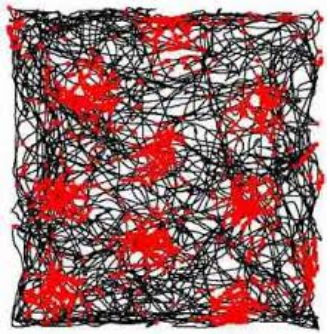
Twitter: @SuryaGanguli

Neural circuits and behavior: theory, computation and experiment

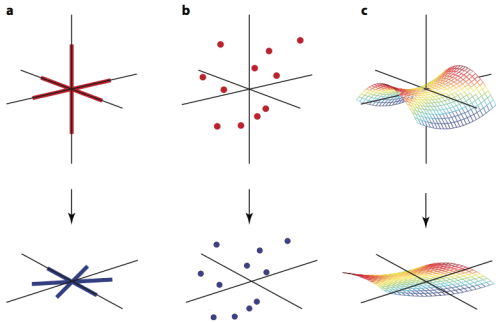
with Baccus lab: inferring hidden circuits in the retina
w/ Niru Maheswaranathan and Lane McIntosh



with Clandinin lab: unraveling the computations underlying fly motion vision from whole brain optical imaging
w/ Jonathan Leong, Ben Poole and Jennifer Esch

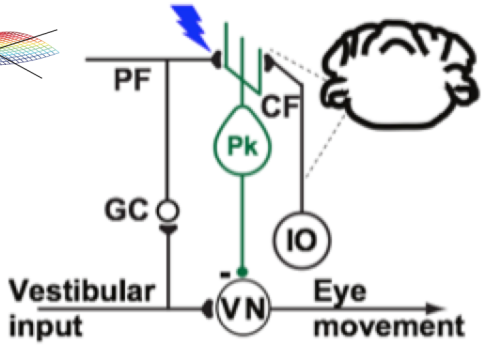


with the Giocomo lab: understanding the internal representations of space in the mouse entorhinal cortex
w/ Kiah Hardcastle and Sam Ocko



with the Shenoy lab: a theory of neural dimensionality, dynamics and measureme
w/ Peiran Gao, Eric Trautmann

with the Raymond lab: theories of how enhanced plasticity can either enhance or impair learning depending on experience
w/ Subhaniel Lahiri, Barbara Vu, Grace Zhao



Motivations for an alliance between theoretical neuroscience and theoretical machine learning

- What does it mean to understand the brain (or a neural circuit?)
- We understand how the connectivity and dynamics of a neural circuit gives rise to behavior.
- And also how neural activity and synaptic learning rules conspire to self-organize useful connectivity that subserves behavior.
- The field of machine learning has generated a plethora of learned neural networks that accomplish interesting functions.
- We know their connectivity, dynamics, learning rule, and developmental experience, *yet*, we do not have a meaningful understanding of how they learn and work!

On simplicity and complexity in the brave new world of large scale neuroscience, Peiran Gao and S. Ganguli, Curr. Op. in Neurobiology, 2015.

Towards a unification of disparate fields

i.e. understanding the trainability, expressivity, and generalizability of neural networks, and machine learning algorithms, especially in high dimensions

Engineering intuitions and theorems as at to what works

Machine Learning and Computer Science

Neuroscience

Theoretical physics and applied mathematics

Proofs of principle; clues from psychology and biophysics

Complex systems analysis

i.e. the neuron principle, convolutional architectures, substrates for reinforcement learning, memory, attention, multiple timescales

i.e. statistical physics of quenched disorder, spin glasses, dynamical criticality, chaos theory

References

- Random projections of random manifolds, S. Lahiri, P. Gao, S. Ganguli, http://arxiv.org/abs/1607.04331, under review at JMLR.
- M. Advani and S. Ganguli, An equivalence between high dimensional Bayes optimal inference and M-estimation, NIPS 2016.
- M. Advani and S. Ganguli, Statistical mechanics of optimal convex inference in high dimensions, Physical Review X, 6, 031034, 2016.
- A. Saxe, J. McClelland, S. Ganguli, Learning hierarchical category structure in deep neural networks, Proc. of the 35th Cognitive Science Society, pp. 1271-1276, 2013.
- A. Saxe, J. McClelland, S. Ganguli, Exact solutions to the nonlinear dynamics of learning in deep neural networks, ICLR 2014.
- Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio, Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, NIPS 2014.
- B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, NIPS 2016.
- S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, Deep information propagation, ICLR 2017.
- S. Lahiri, J. Sohl-Dickstein and S. Ganguli, A universal tradeoff between energy speed and accuracy in physical communication. https://arxiv.org/abs/1603.07758
- A memory frontier for complex synapses, S. Lahiri and S. Ganguli, NIPS 2013.
- Modelling arbitrary probability distributions using non-equilibrium thermodynamics, J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, ICML 2015.
- Deep Knowledge Tracing, C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas, J. Sohl-Dickstein, NIPS 2015.
- Deep learning models of the retinal response to natural scenes, L. McIntosh, N. Maheswaranathan, S. Ganguli, S. Baccus, NIPS 2016.
- Improved multitask learning through synaptic intelligence, Friedemann Zenke, Ben Poole, Surya Ganguli, arxiv.org/abs/1703.04200, under review.
- A. Nayebi and S. Ganguli, Biologically inspired protection of deep networks from adversarial attacks, arxiv.org/abs/1703.09202, under review.
- http://ganguli-gang.stanford.edu

Talk Outline

Improved multitask learning through synaptic intelligence.

Synaptic complexity -> multitask learning

Protecting deep networks from adversarial examples.

Dendritic biophysics + Kurtotic weights -> adversarial robustness

Deep learning models of the retinal response to natural scenes.

Computational reconstruction of the interior of the retina.

Improved multitask learning through synaptic intelligence

Ben Poole
poole@cs.stanford.edu

Friedemann Zenke
fzenke@stanford.edu

Subhaneil Lahiri
sulahiri@stanford.edu

Improved multitask learning through synaptic intelligence, Friedemann Zenke, Ben Poole, Surya Ganguli, arxiv.org/abs/1703.04200, under review.

A memory frontier for complex synapses, S. Lahiri and S. Ganguli, NIPS 2014. (outstanding paper award)

Improved multitask learning through synaptic intelligence

Machine Learning

Training phase

↓

Adjust synaptic weights

↓

Freeze weights

↓

Testing phase

If data distribution changes,
must retrain entire network

Biological Learning

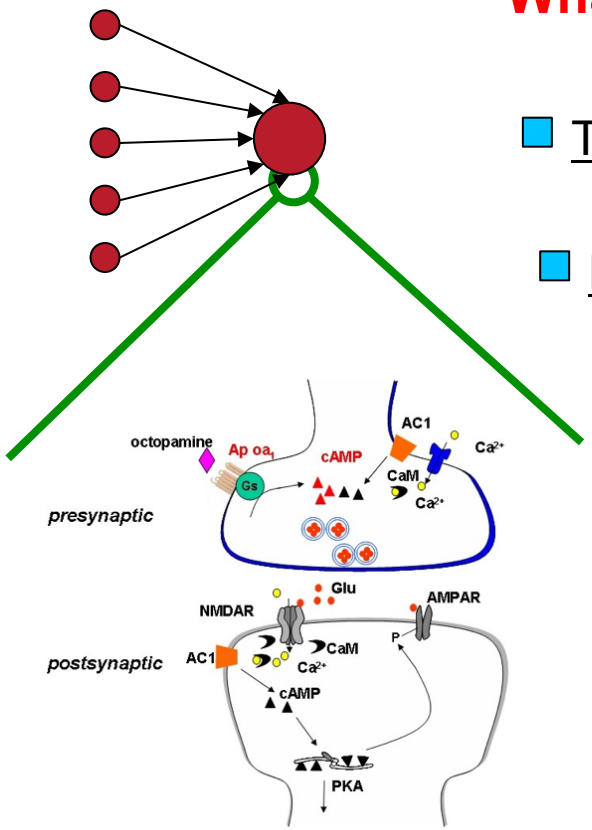
Lifelong learning

Continuous adaptation to changing domains with evolving data distributions

Can solve new tasks without forgetting old tasks

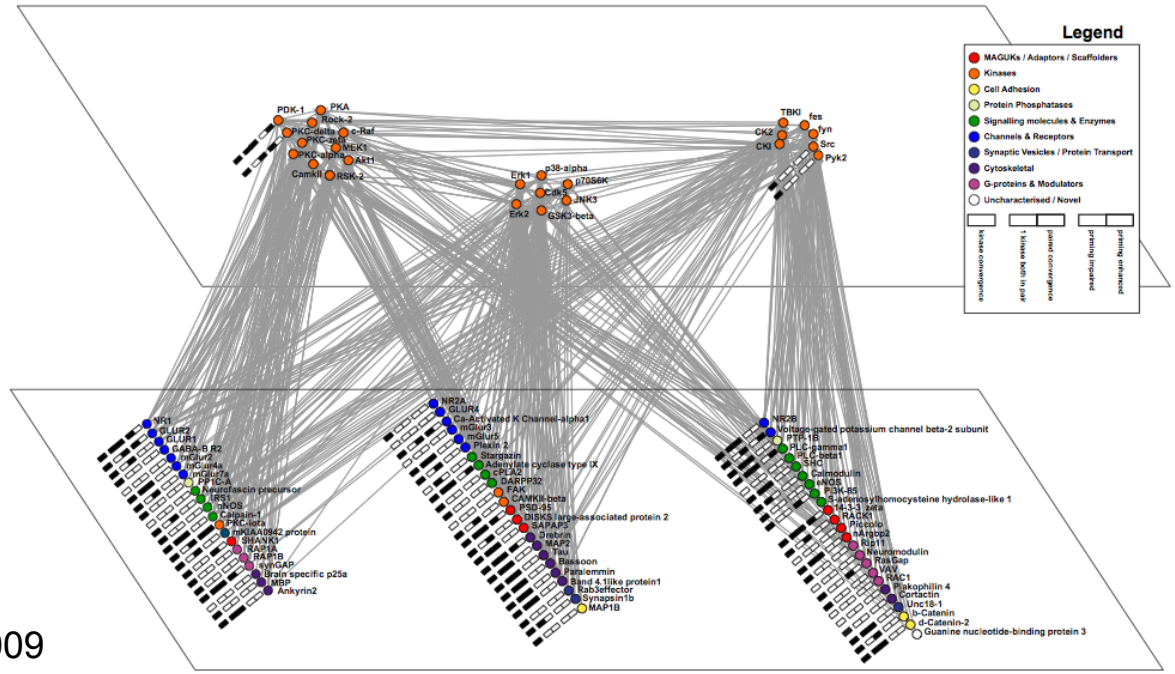
Question: How can take steps toward multitask learning in artificial neural networks?

A gulf between theory and experiment



What is a synapse from neuron j to neuron i?

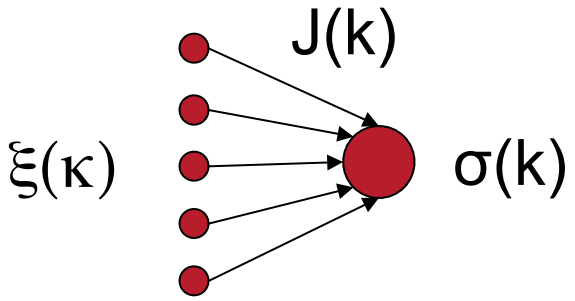
- Theorist: W_{ij} or J_{ij} ~ size of postsynaptic potential
- Experimentalist: AMPA, NMDA, CAMKII, MAPK, CREB, MHC-I, second messengers, membrane protein regulation, intracellular trafficking, new protein synthesis …….



Coba et. al.
Science Signalling 2009

Memory capacity with scalar analog synapses

Consider the number of associations a neuron with N afferent synapses can store.



$$\sigma(k) = \mathrm{sgn}\,(J \cdot \xi(\kappa) - \theta)$$

An online learning rule to store the desired association:

$$J(k+1) = e^{-1/\tau} J(k) + \sigma(k)\,\xi(\kappa)$$

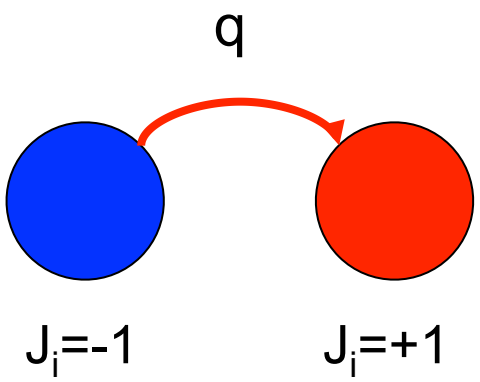
i.e. 1) Allows analog weights to decay slightly (forget the past inputs)
2) Add in the new association to the weight (learn a new input).

Memory capacity: How far back into the past can synapses reliably recall previously stored associations?

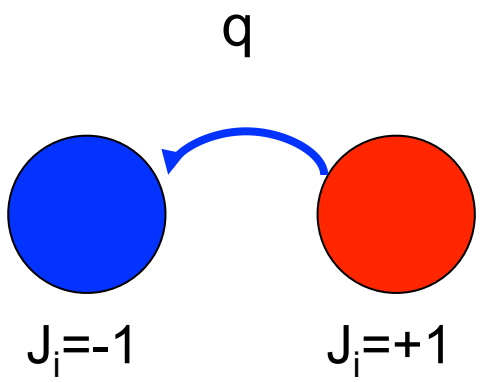
Answer: If τ is O(N) then the past O(N) associations can be recalled.

Problem: This solution relies on individual synapses to reliably maintain O(N) distinguishable analog states. Fusi and Amit 92

Memory capacity with binary synapses



Potentiation



Depression

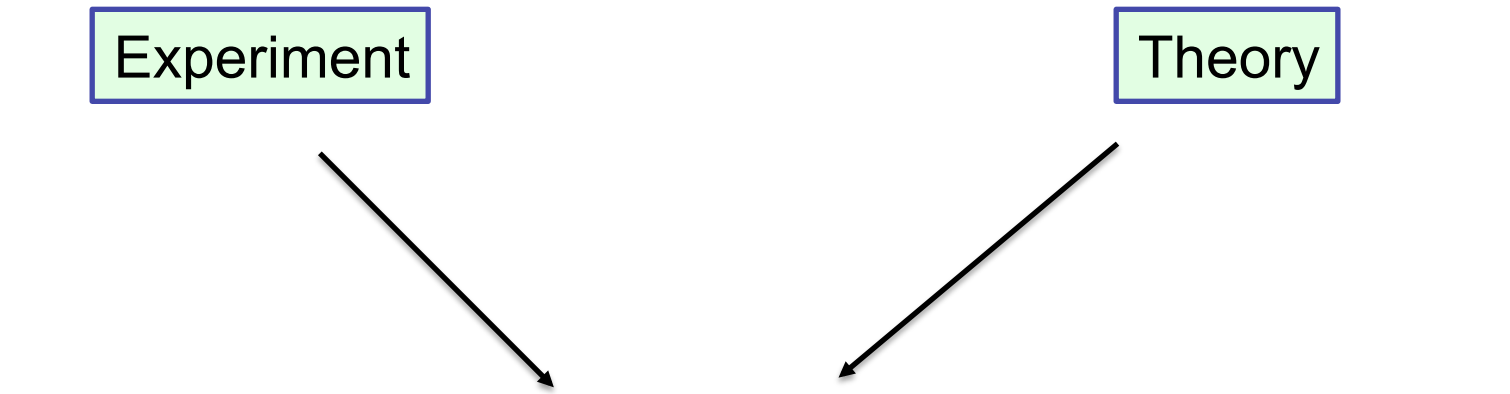
q = prob a synapse changes strength under appropriate conditions
N = number of synapses

Memory Capacity

$q = O(1)$	$\log N$	Quickly learn, quickly forget
$q = O(N^{-1/2})$	$N^{1/2}$	Sluggish to learn, slow to forget

Fusi and Amit 92

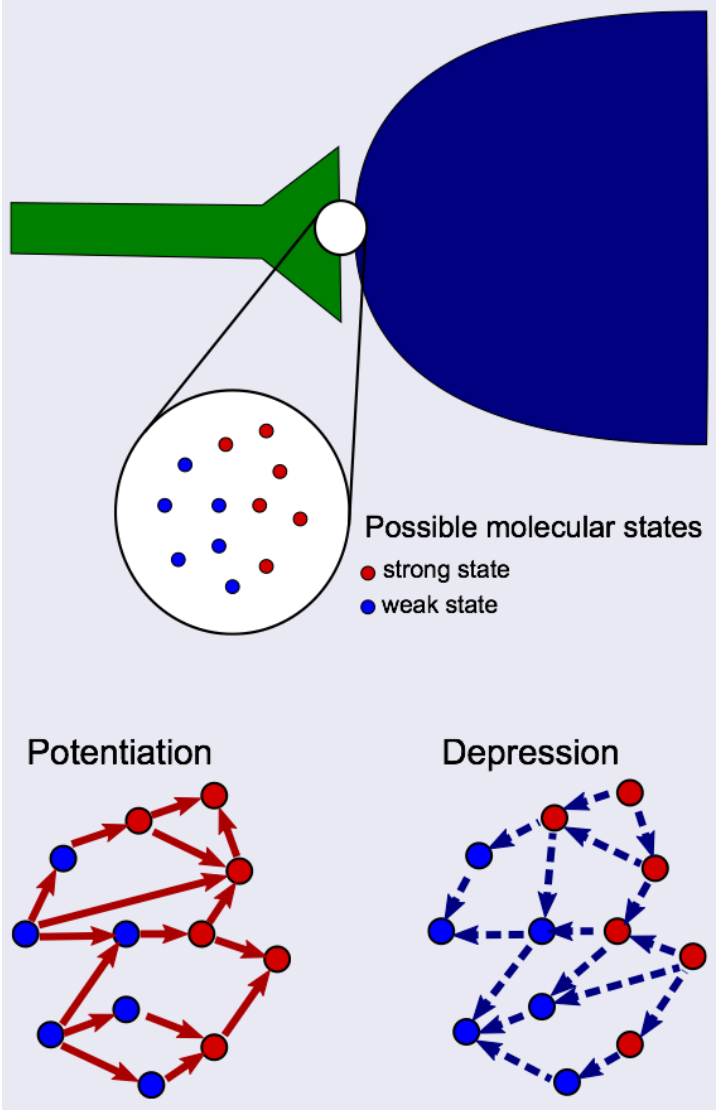
Synaptic complexity: from scalars to dynamical systems



We must expand our theoretical conception of
a synapse from that of a simple scalar value to
an entire (stochastic) dynamical system in its own right.

This yields a large universe of
synaptic models to explore and understand.

Framework for synaptic dynamical systems

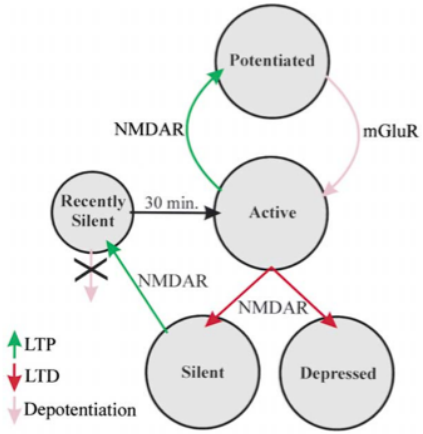


Theoretical approach:

A synapse is an arbitrary stochastic dynamical system with M internal states.

Some internal states correspond to a strong synapse, others a weak synapse.

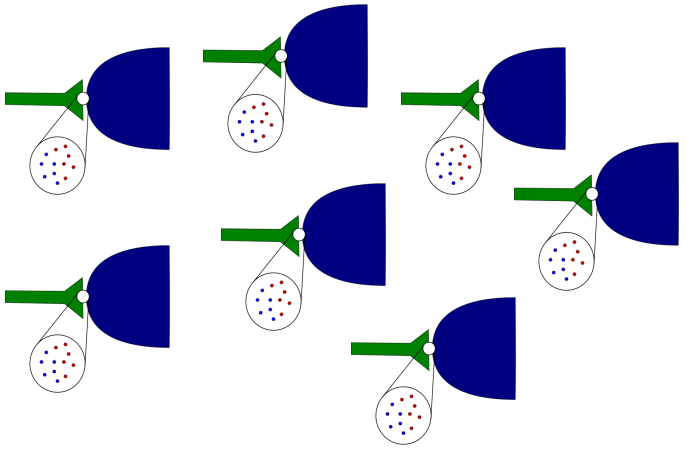
A candidate potentiation (depression) event induces an arbitrary stochastic transition between states.



Montgomery and Madison Neuron 2002

Ideal observer measure of memory capacity: SNR

A continuous stream of memories are stored (at poisson rate r) in a population of N synapses with M internal states.



The memory stored at time t=0 demands that some synapses potentiate, while others depress, yielding an ideal synaptic weight vector w_{ideal}.

The storage of future memories after t=0 changes the weight vector to w(t).

An upper bound on the quality of memory retrieval of any memory readout using neural activity is given by the SNR curve:

$$\mathrm{SNR}(t) = \frac{\langle \vec{w}_{\mathrm{ideal}} \cdot \vec{w}(t) \rangle - \langle \vec{w}_{\mathrm{ideal}} \cdot \vec{w}(\infty) \rangle}{\sqrt{\mathrm{Var}\left(\vec{w}_{\mathrm{ideal}} \cdot \vec{w}(\infty)\right)}}$$

Each choice of

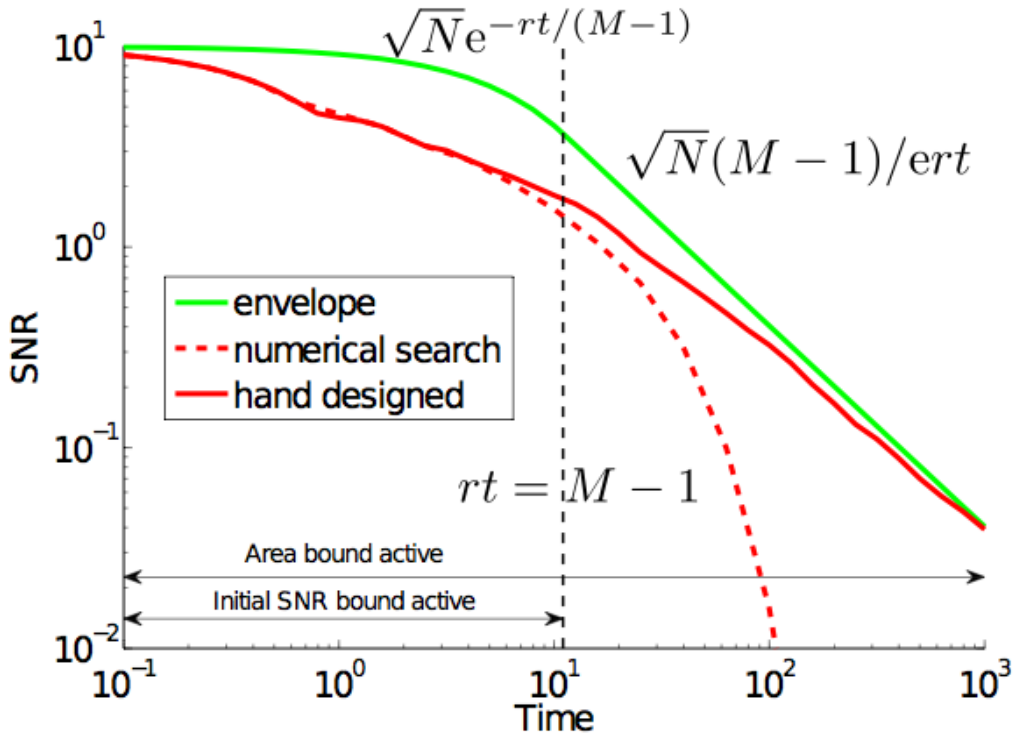
N, M, M^{pot} and M^{dep}

yields a different memory curve.

Fusi et. al. 2005, Fusi et. al. 2007, Barrett and van Rossum, 2008

A frontier beyond whose bourn no curve can cross

Area bound implies a maximal achievable memory at any finite time given N synapses with M internal states:



Chains with different transition rates come close to the frontier at late times.

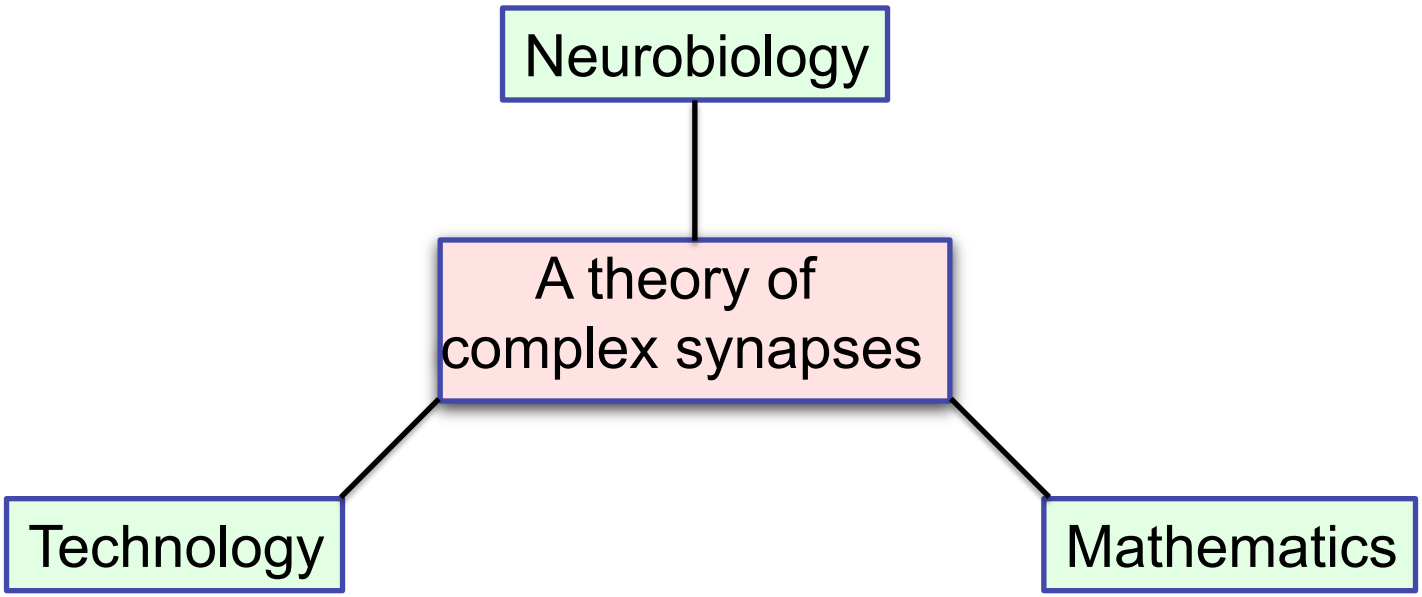
Various measures of memory (area, frontier, lifetime) grow linearly with the number of internal states M, but grow only as the square root of the number of synapses N.

Lahiri and Ganguli, NIPS 2014, outstanding paper award (3/1400)

The dividends of understanding synaptic complexity

(Cerebellar learning with complex synapses)

A framework for interpreting
molecular neurobiology data



The next generation of
artificial neural networks?

(Spatiotemporal credit assignment)
(Learning as message passing)

New theorems about
perturbations
to stochastic processes.

(Tighter bounds)

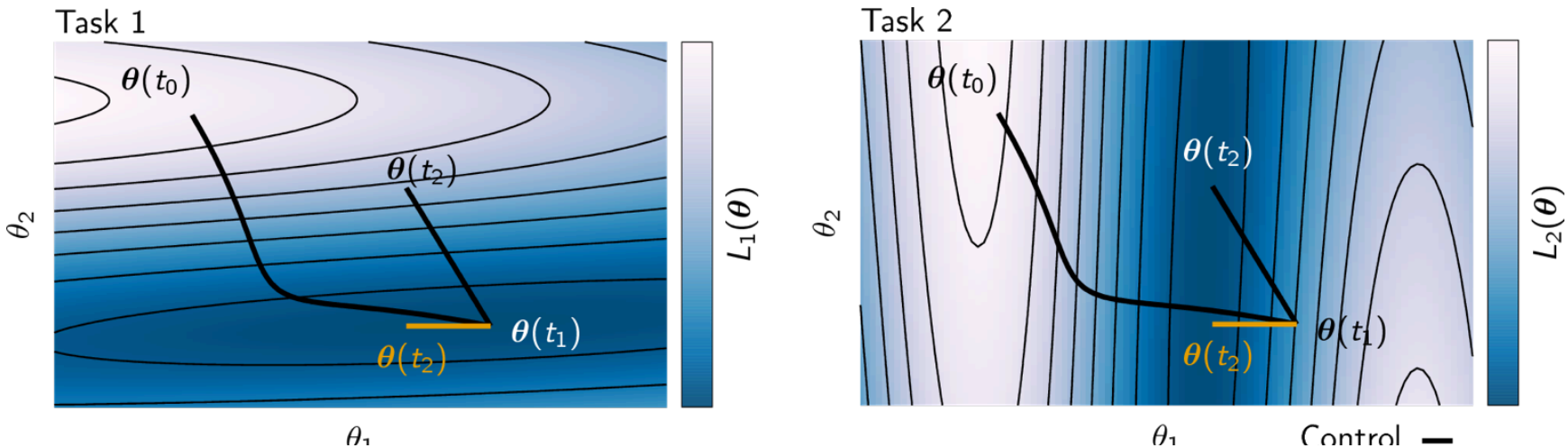
Improved multitask learning through synaptic intelligence

Task 1 constrains certain dimensions in weight space but leaves other dimensions unconstrained.

Task 2 constrains *different* dimensions in weight space, and leaves a *different* set of dimensions unconstrained.

Idea: Each synapse keeps track of its ***importance*** in solving all previous tasks and uses this importance to freeze its learning dynamics.

Theory: This local online synaptic computation can approximately compute The sum of the Hessians of all previous tasks!



Local synaptic path integral -> Global computation of Hessian

During each task, each synapse computes its contribution to the path integral that measures change in loss.

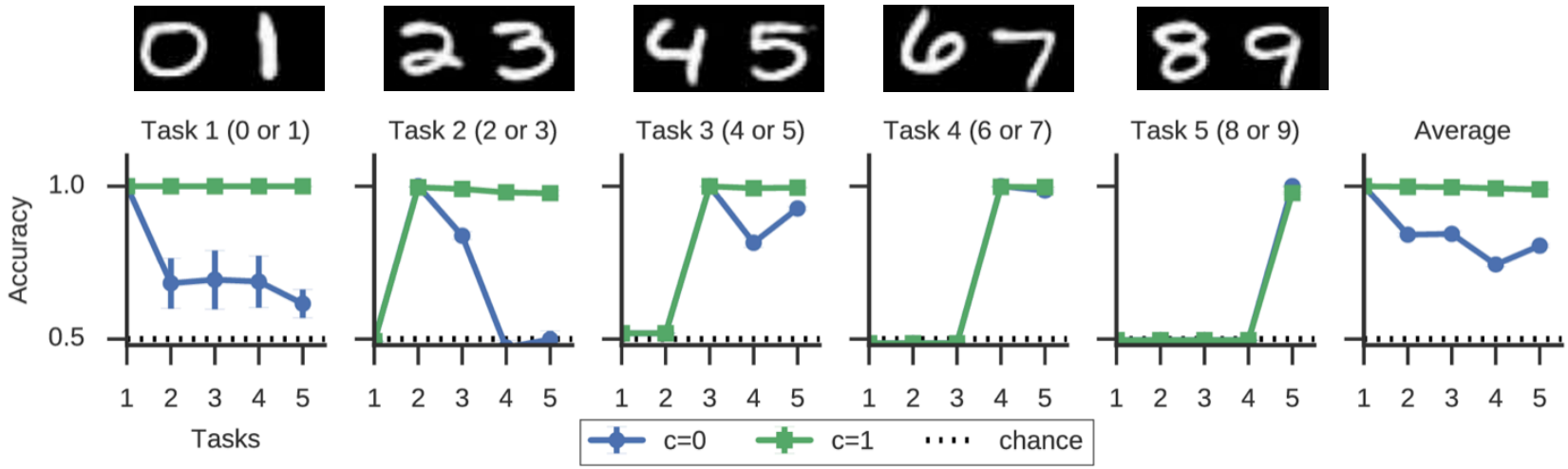
This is a local computation that involves multiplying gradient of error with change in synaptic

$$\int_{t^{\mu-1}}^{t^{\mu}} \boldsymbol{g}(\boldsymbol{\theta}(t)) \cdot \boldsymbol{\theta}'(t) dt = \sum_k \int_{t^{\mu-1}}^{t^{\mu}} g_k(\theta(t)) \theta_k'(t) dt$$
$$\equiv -\sum_k \omega_k^{\mu}. \qquad (3)$$

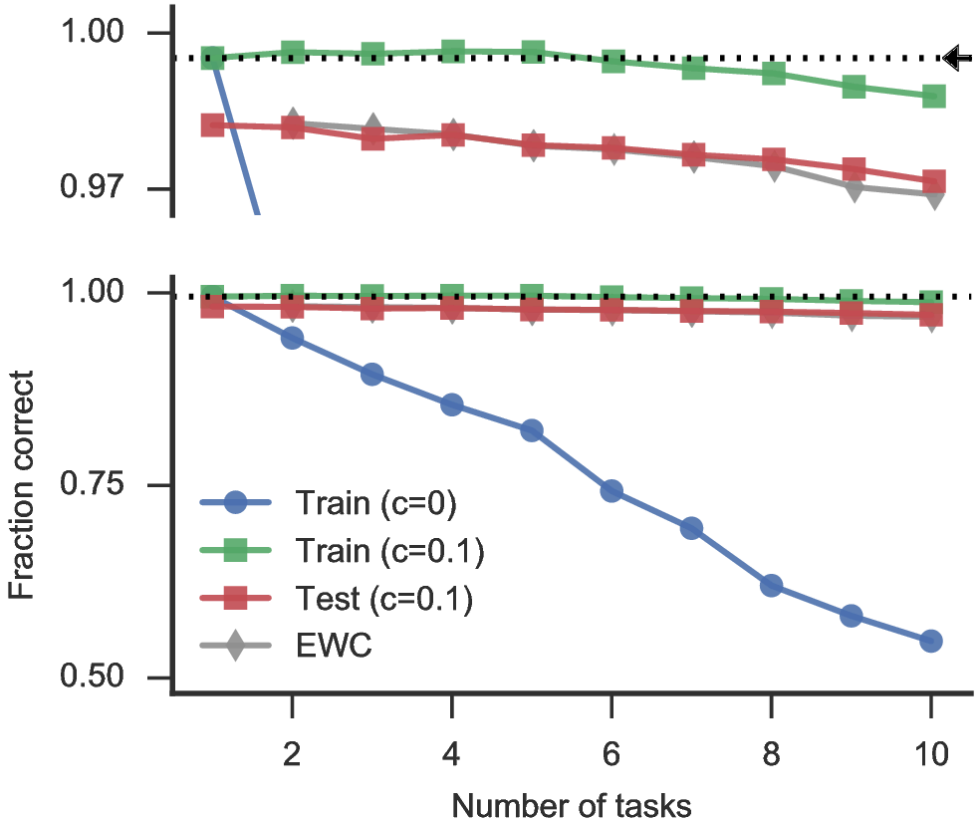
A local-to-global theorem: under gradient descent dynamics on a low rank quadratic error landscape, the importance parameters are proportional to the diagonal elements of the Hessian!!

Split MNIST

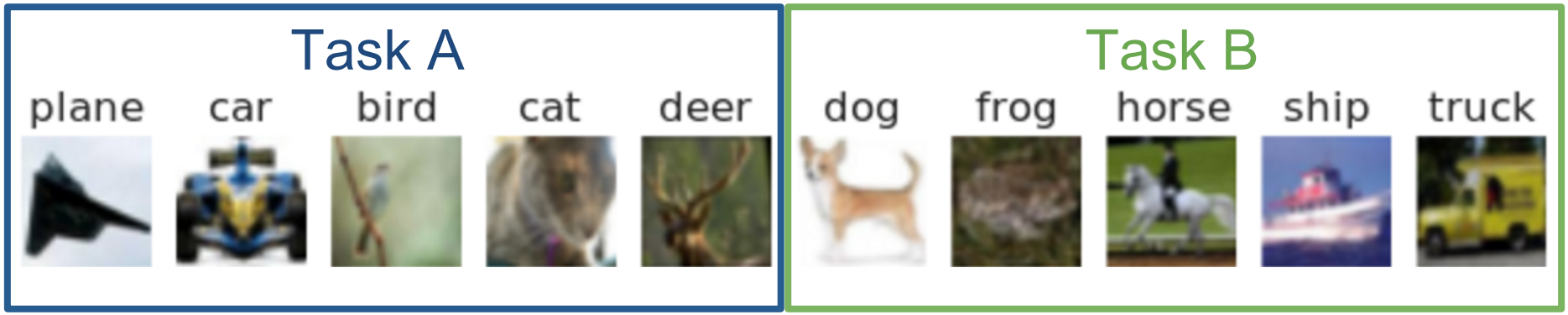
Sequentially learn to classify 0 vs. 1, then 2 vs. 3, …., 8 vs. 9



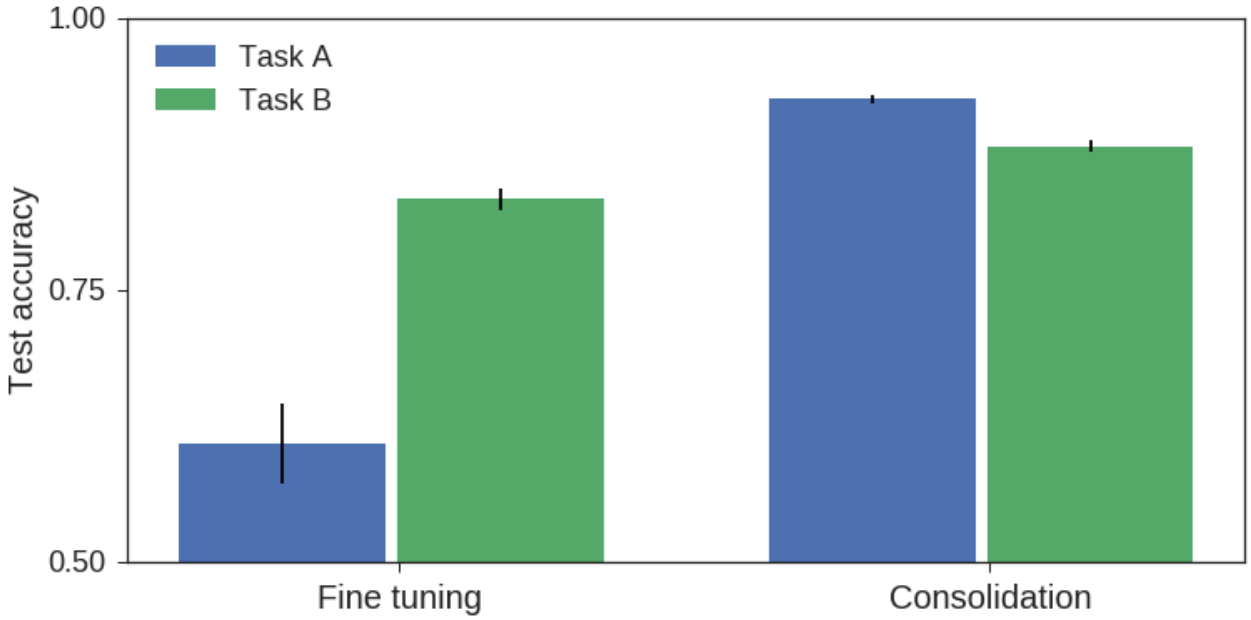
Permuted MNIST



Split CIFAR-10



Train on task A, then on task B. Evaluate test accuracy on both:



Chance accuracy is 0.2 (⅕)

Talk Outline

Improved multitask learning through synaptic intelligence.

Synaptic complexity -> multitask learning

Protecting deep networks from adversarial examples.

Dendritic biophysics + Kurtotic weights -> adversarial robustness

Deep learning models of the retinal response to natural scenes.

Computational reconstruction of the interior of the retina.

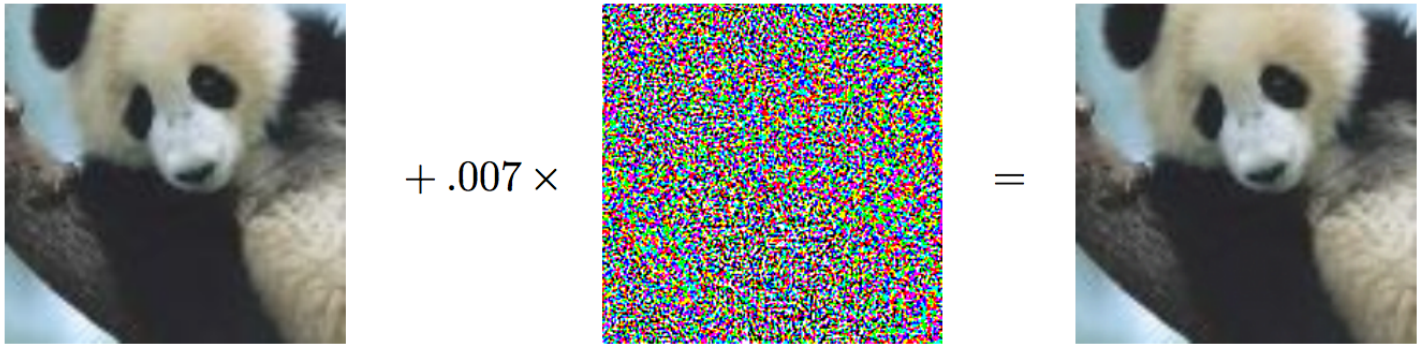
Protecting deep networks from adversarial examples

Aran Nayebi

anayebi@stanford.edu

Nayebi and S. Ganguli, Biologically inspired protection of deep networks from adversarial attacks, arxiv.org/abs/1703.09202, under review.

The problem of adversarial examples



$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

Goodfellow, Shlens, Szegedy, ICLR 2015

These neural networks do not fundamentally understand the task in the same way we do.

Technological concern: adversarial examples generalize across different Architectures and subsets of training data -> black box attacks.

Scientific concern: as foundational models in neuroscience, they get fooled by examples that do not fool us -> our models are missing biological ingredients

A linear explanation of adversarial examples

Idea behind adversarial example generation:

x = test example
x + **dx** = perturbed example

Choose **dx** so as to maximize error E subject to an L_∞ bound on **dx**.

This optimization is difficult, so linearize the error as a function over input space to get a gradient vector **g**, and optimize:

$$\mathbf{dx}^* = \arg\max_{\mathbf{dx}} \quad \mathbf{g} \cdot \mathbf{dx}, \quad \text{such that} \quad ||\mathbf{dx}||_\infty \leq \epsilon$$

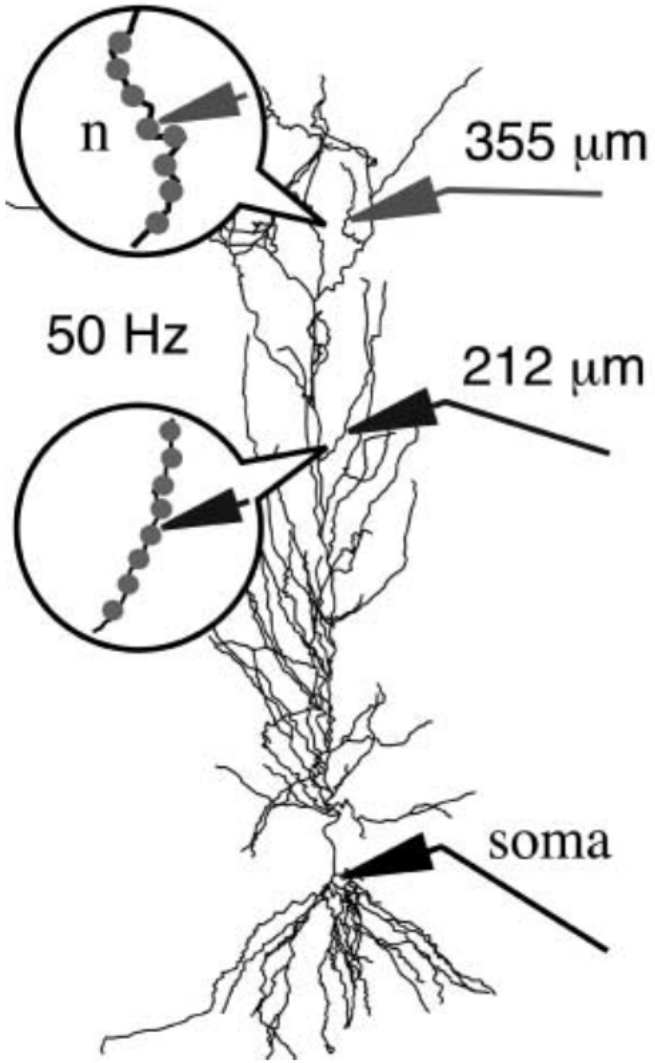
$$\mathbf{dx}^* = \epsilon \, \text{sign}(\mathbf{g})$$ The adversarial perturbation.

$$dE = \epsilon \, ||\mathbf{g}||_1$$ The increase in error.

Intuition: due to linear summation of many variables, one can change each individual variable by a *small* amount, while moving the sum by a *lot.*

Goodfellow, Shlens, Szegedy, ICLR 2015

The brain likely never linearly sums many variables



The biophysical basis for linear summation: linear superposition of trans-membrane voltages under passive cable theory.

However, active ionic conductances can destroy linear superposition by introducing nonlinear saturating thresholds.

Individual dendritic branches can be in highly saturated states or be far below threshold.

Either way, small input perturbations cannot easily propagate through such a nonlinear system.

synapses on a branch that can linearly sum is O(10) to O(100).

Poirazi, Brannon, Mel, Neuron 2015

Avoiding adversarial examples by exploring the saturated regime

Overall idea: penalize neurons according to how far their activation lies from the saturated regime.

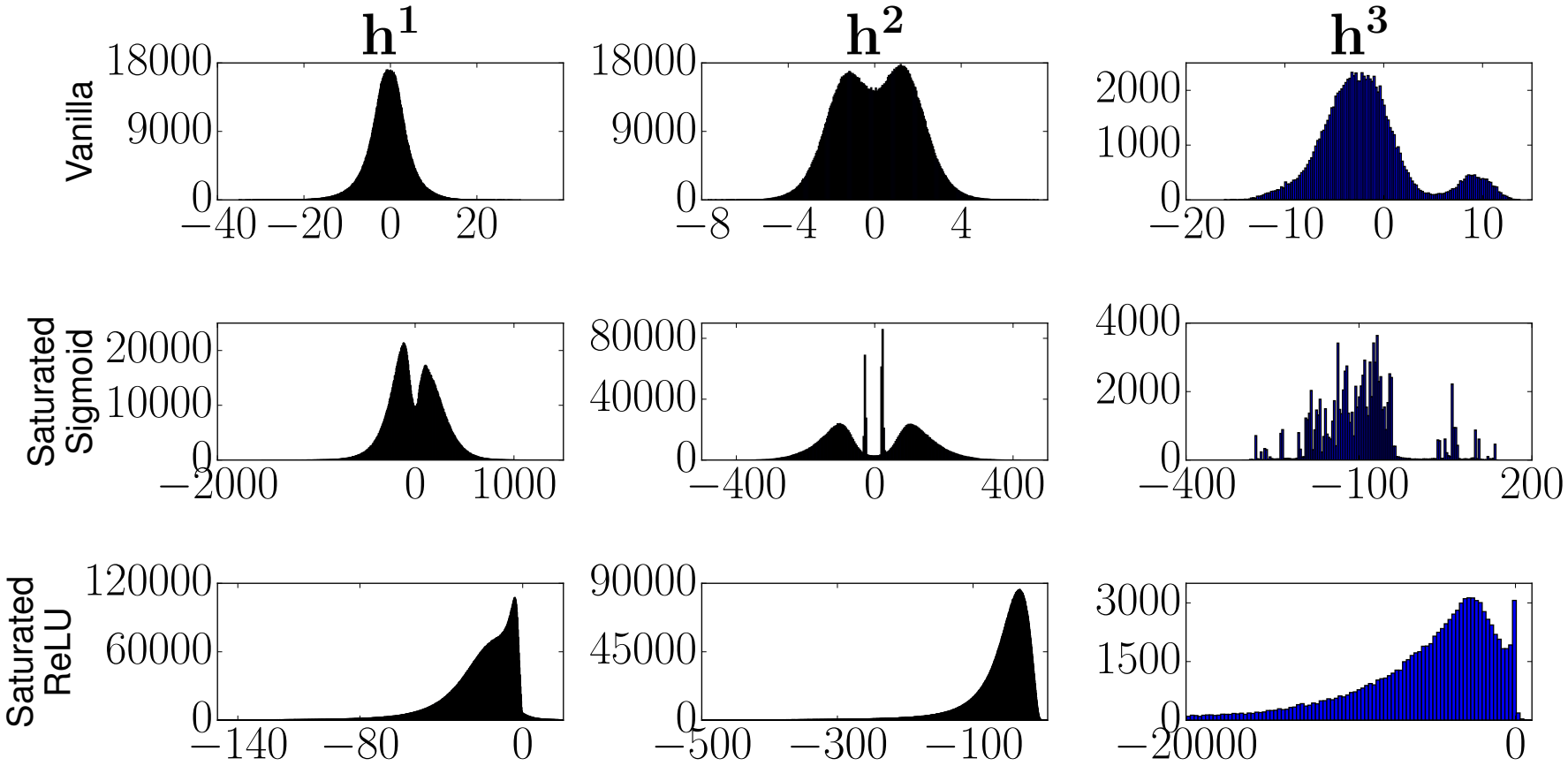
Problem: If many neurons are saturated it can be difficult to learn due to vanishing gradient problem.

Solution: Anneal the penalty that promotes saturation over time in the training process.

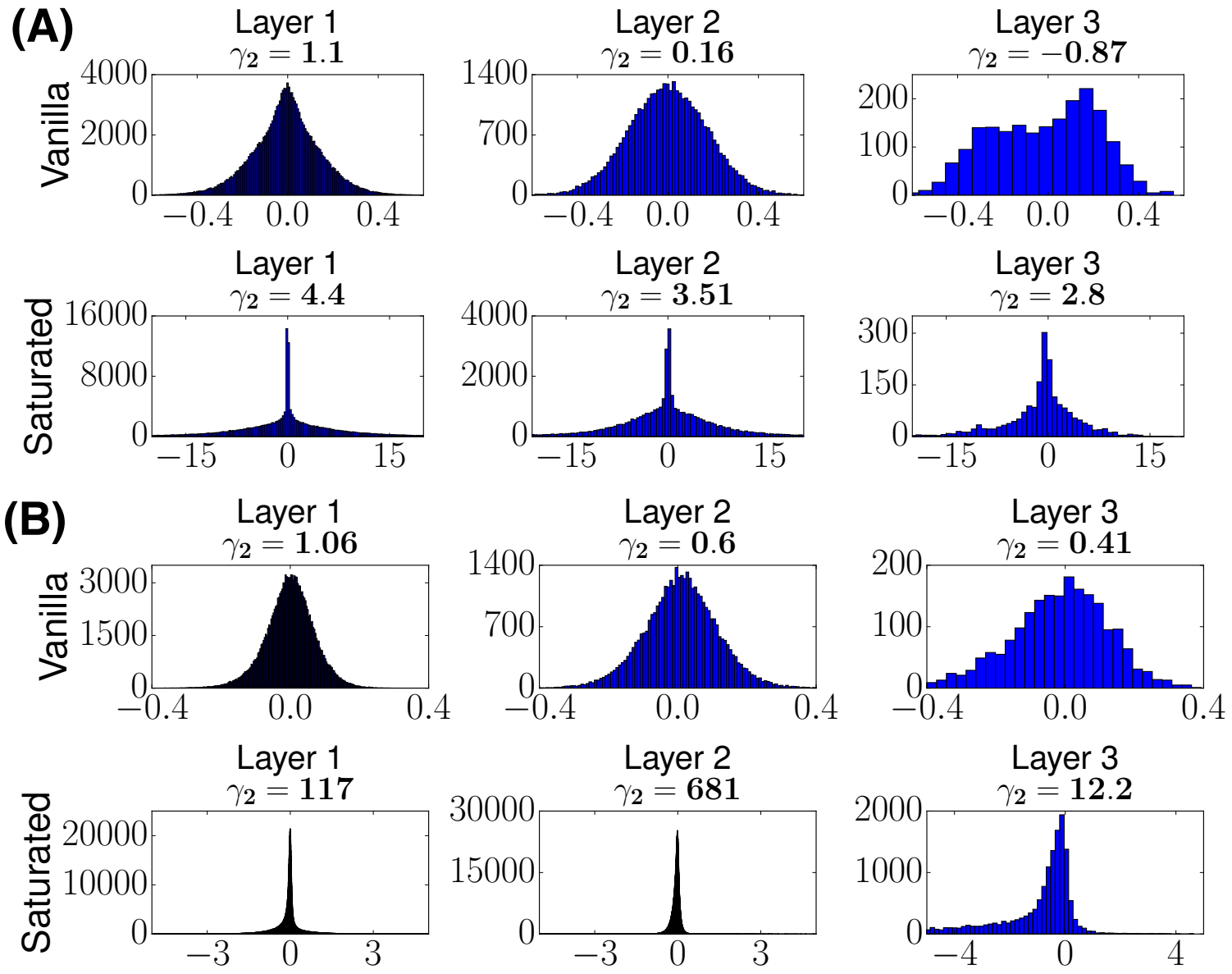
Results on MNIST:

Training	Sigmoid MLP	ReLU MLP	CNN
Vanilla	97.6% 0%	98.1% 0.41%	99.35% 5.62%
Adversarial	92.27% 81.71%	92.29% 91.04%	99.32% 83.83%
Saturated	97.01% **94.43%**	95.24% **94.59%**	99.33% **98.45%**

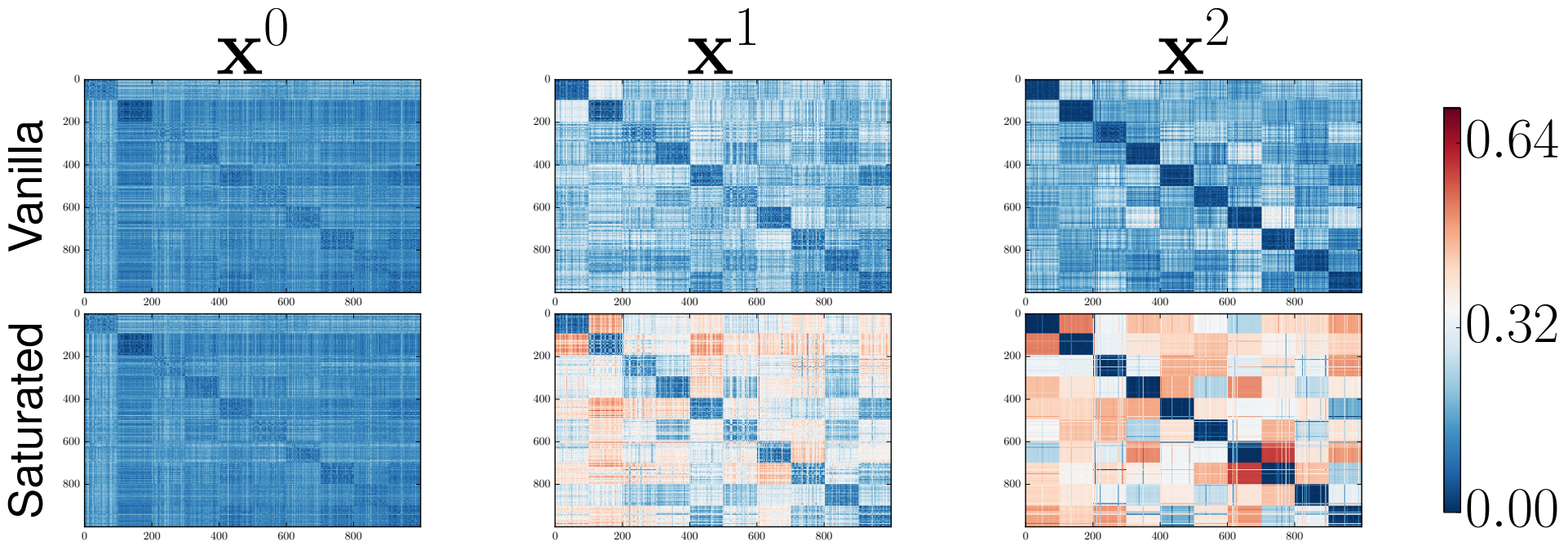
Confirmation that adversarially robust networks are operating in the saturated regime



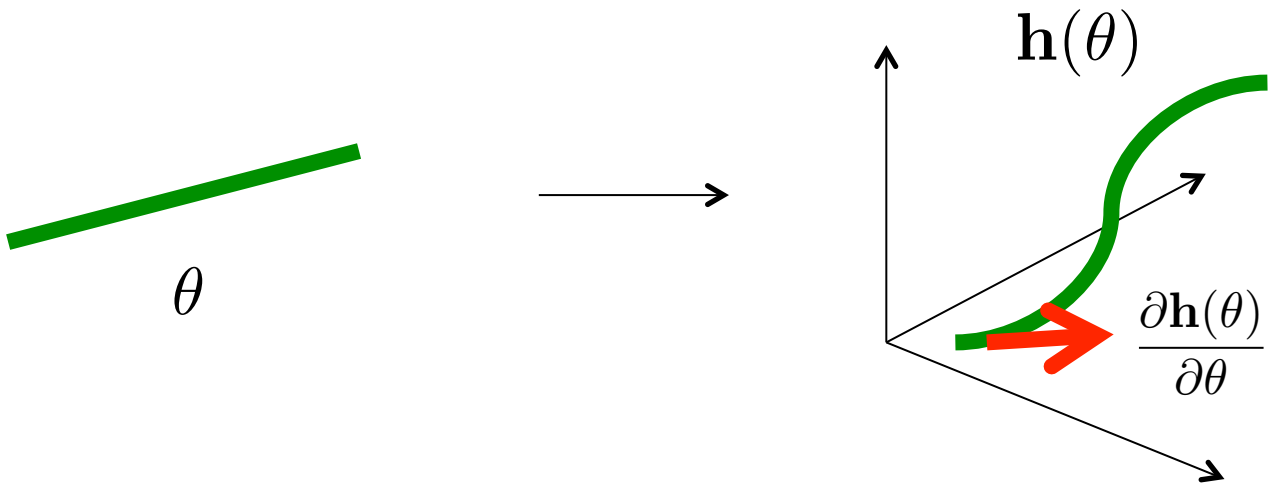
Adversarially robust networks learn highly kurtotic weight distributions, as in the brain



RDM analysis reveals highly cluster separation and tightness in adversarially robust networks



Riemannian geometry of input-output map

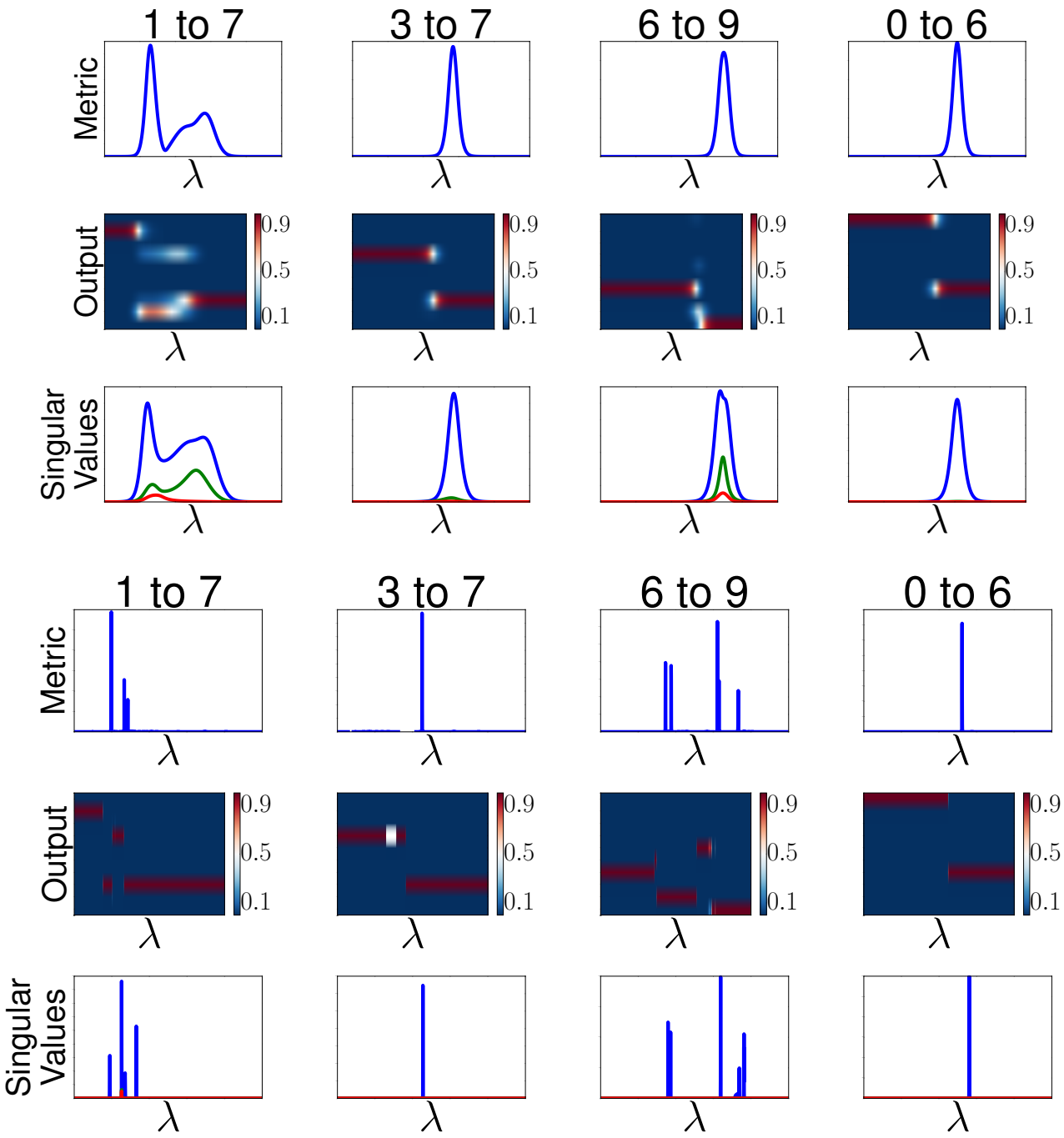


$$g^E(\theta) = \left(\frac{\partial \mathbf{h}(\theta)}{\partial \theta}\right)^T g^F(\mathbf{h}) \frac{\partial \mathbf{h}(\theta)}{\partial \theta}$$

Metric on manifold coordinate θ induced by metric in internal representation space **h**.

$$d\mathcal{L}^E = \sqrt{g^E(\theta)}d\theta$$

Length element: if one moves from Θ to $\Theta + d\Theta$ along the manifold, then one moves a distance dL^E in internal representation space


1 to 7
3 to 7
6 to 9
0 to 6
Metric
Output
Singular Values
λ
0.9
0.5
0.1

Outcome of geometric analysis

Adversarially robust networks form highly flat input-output maps.

One must move large distances in input space to move a given distance in output space.

These maps compress at center of decision volumes and expand at boundaries.

Even at expansion points, networks are sensitive to only one dimension of input perturbations.

This greatly reduces the number dimensions an adversary can exploit to fool the network.

Comparison to a random network: input-output chaos as a source of expressivity

Stanford

Google

Ben Poole

Subhaneil Lahiri

Maithra Raghu

Jascha Sohl-Dickstein

Expressivity: what kinds of functions can a deep network express that shallow networks cannot?

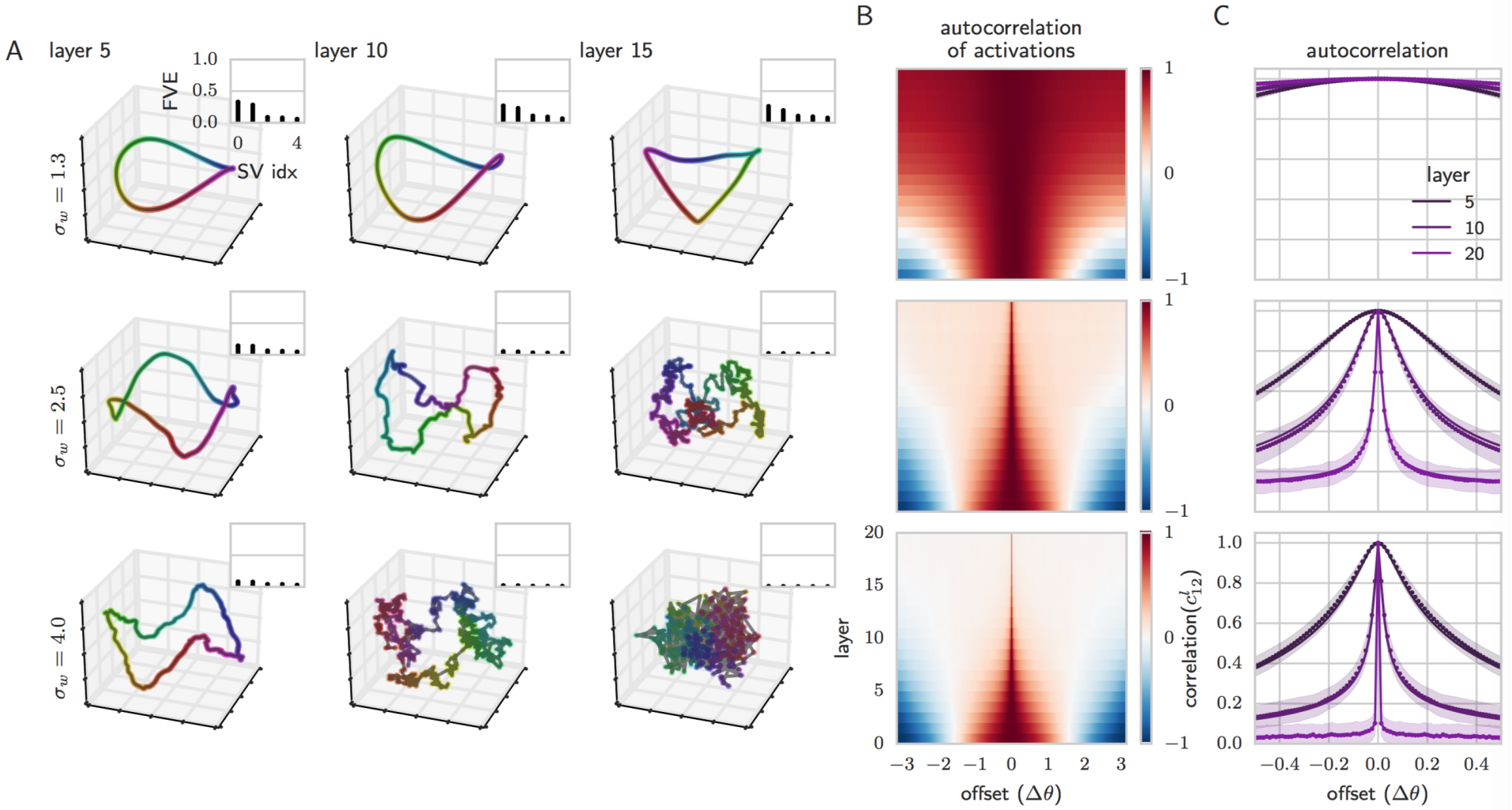
Exponential expressivity in deep neural networks through transient chaos, B. Poole, S. Lahiri,M. Raghu, J. Sohl-Dickstein, S. Ganguli, NIPS 2016.

On the expressive power of deep neural networks, M.Raghu, B. Poole,J. Kleinberg, J. Sohl-Dickstein, S. Ganguli, under review, arxiv/1606.05336

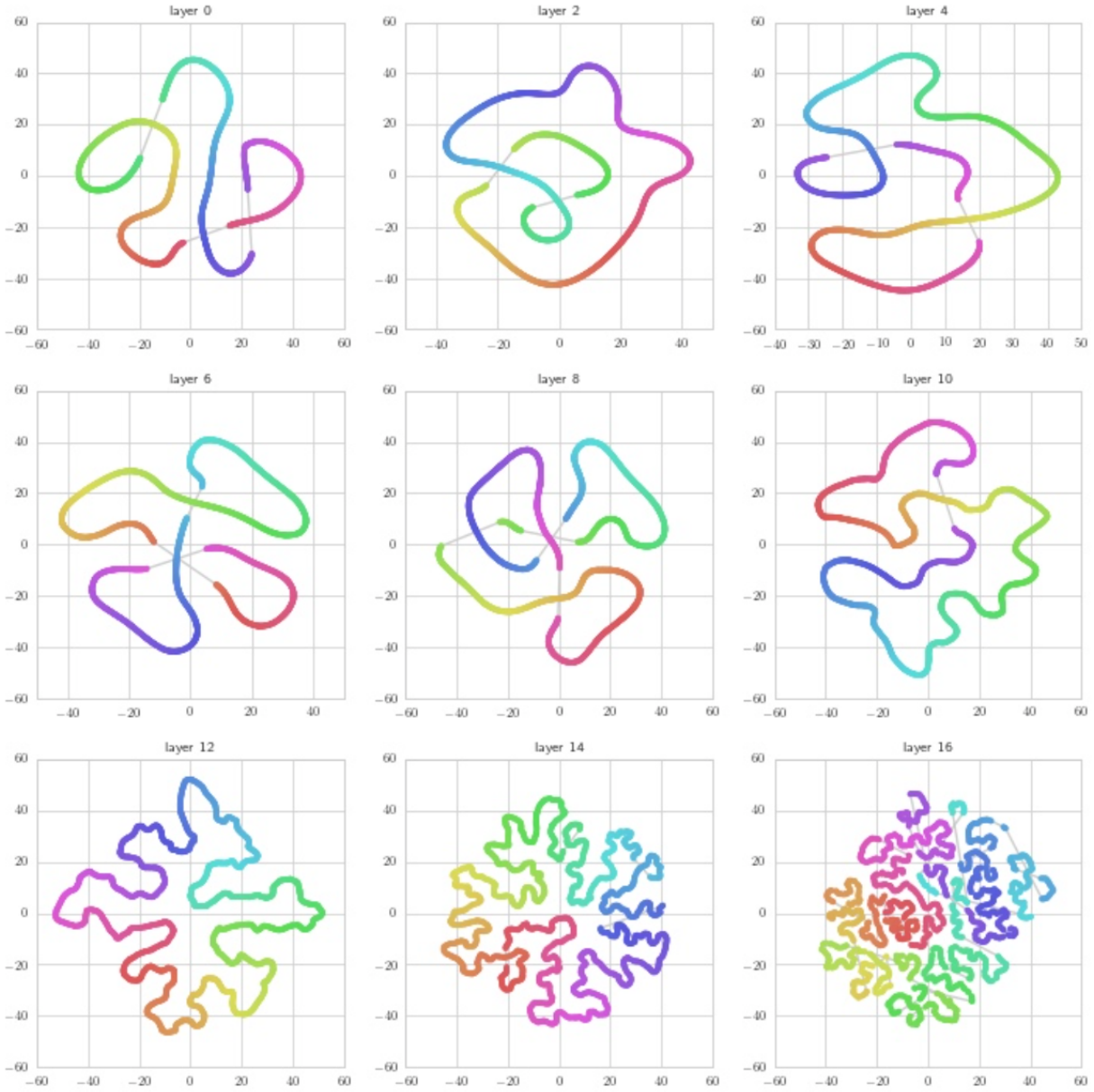
Propagation of a manifold through a deep network

$$\mathbf{h}^1(\theta) = \sqrt{N_1 q^*}\left[\mathbf{u}^0\cos(\theta) + \mathbf{u}^1\sin(\theta)\right]$$

A great circle input manifold



Propagation of a manifold through a deep network

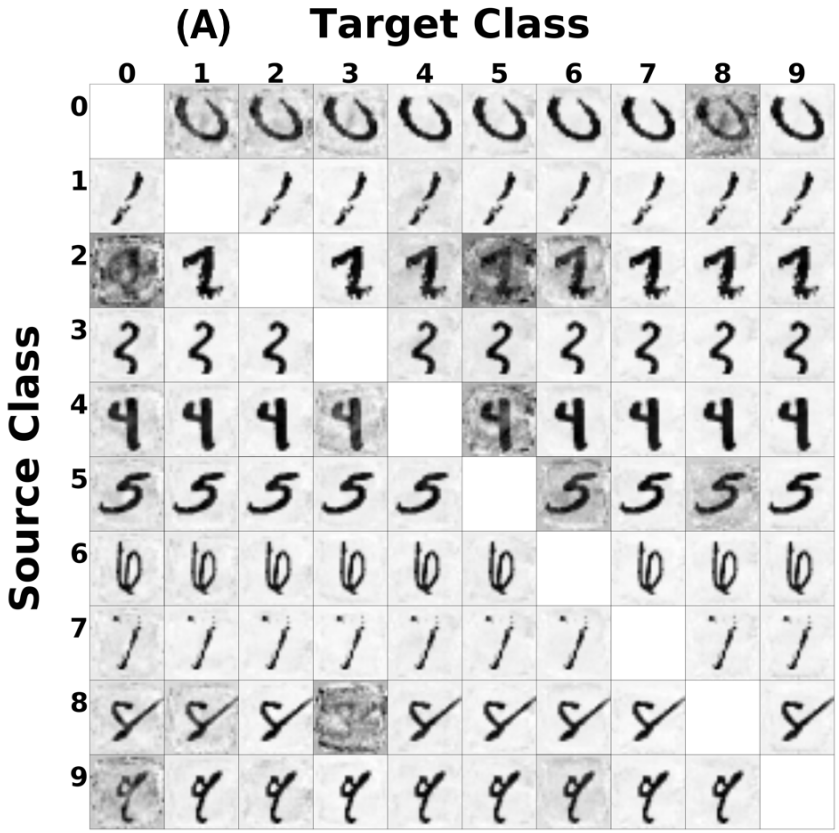


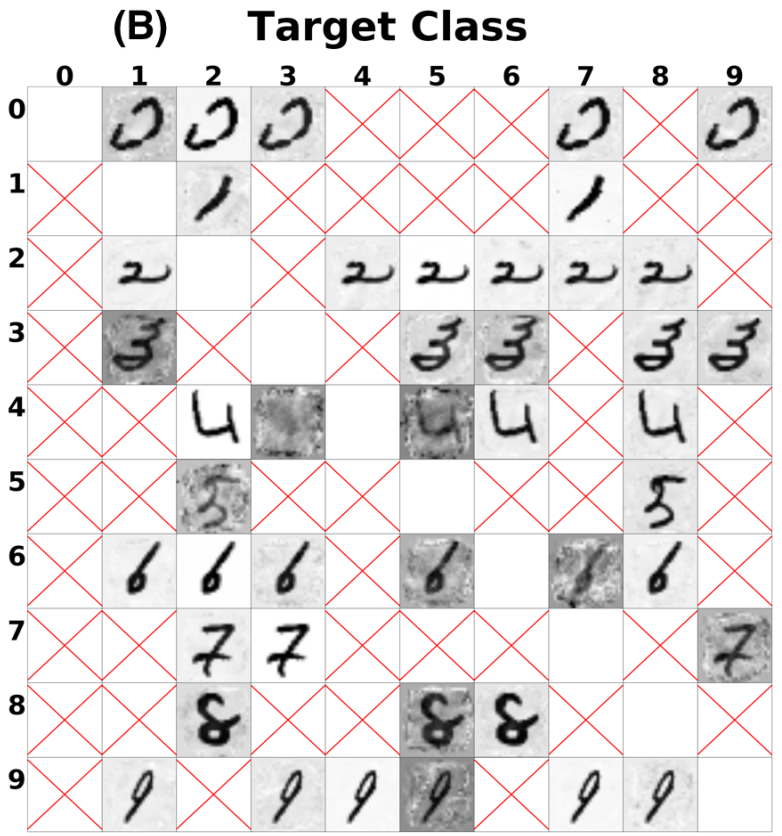
More powerful, iterative second order adversaries

Pick the least confident image from a source class.

For each target class: decrease source prob and increase target prob

Use 1000 iterations of LBFGS.



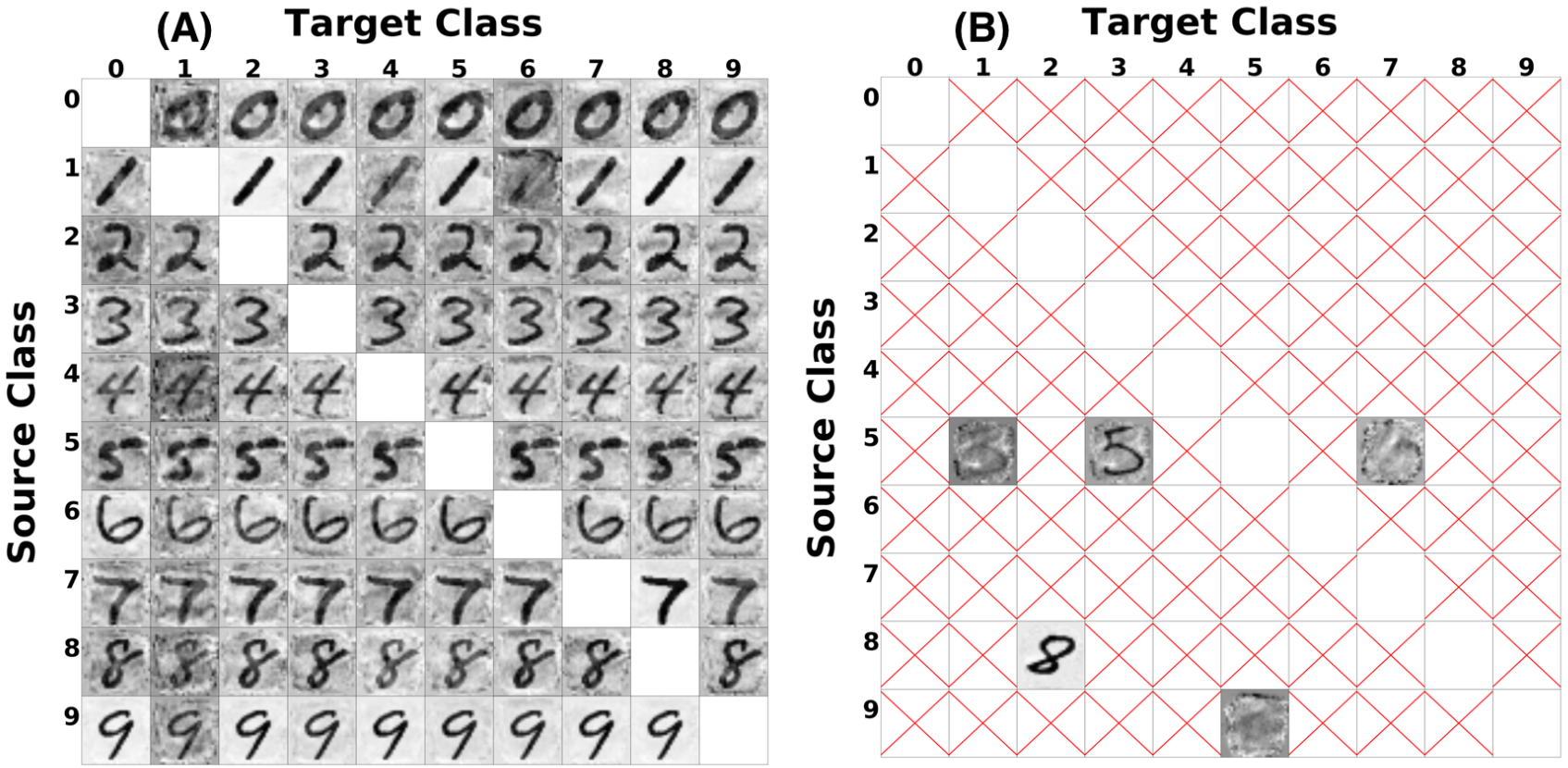


More powerful, iterative second order adversaries

Pick the most confident image from a source class.

For each target class: decrease source prob and increase target prob

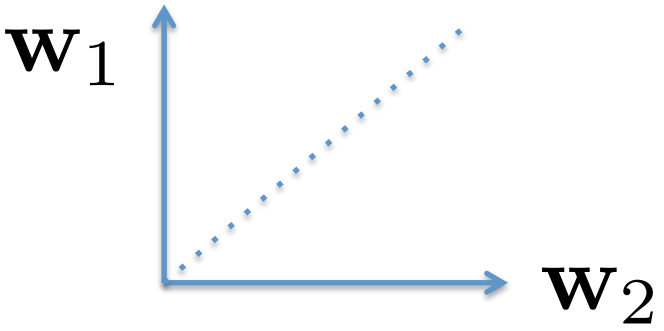
Use 1000 iterations of LBFGS.



Role of weight kurtosis: a linear protective mechanism

Consider linear classification with two cluster prototype vectors $\mathbf{w}_1$ and $\mathbf{w}_2$
A test example **x** is class 1 iff $\mathbf{w}_1 \cdot \mathbf{x} > \mathbf{w}_2 \cdot \mathbf{x}$

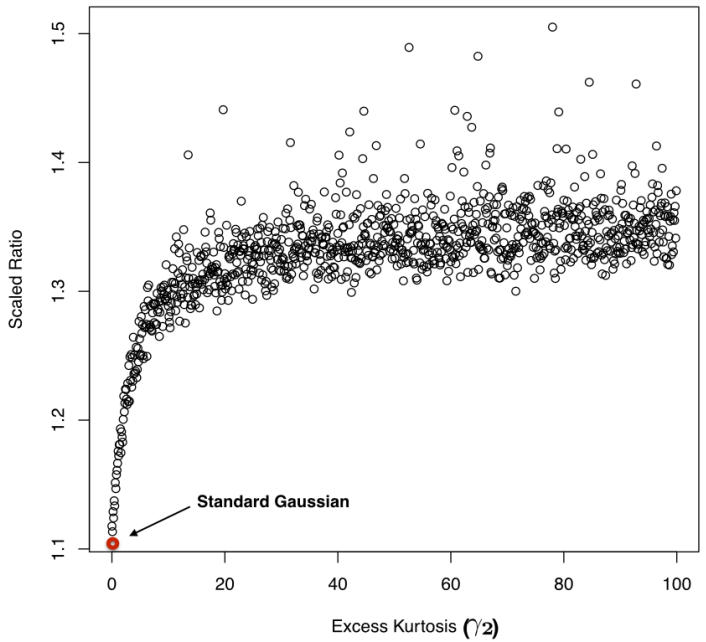
x = $\mathbf{w}_1$ be a test example
x + **dx** = perturbed example



A bounded l_∞ norm perturbation will fool the system iff:

$$\epsilon > \epsilon_{\min} \equiv \frac{||\mathbf{w}_1||_2^2}{||\mathbf{w}_2 - \mathbf{w}_1||_1}$$

For unit l_2 norm **w**: $1 \leq ||\mathbf{w}||_1 \leq \sqrt{N}$



Talk Outline

Improved multitask learning through synaptic intelligence.

Synaptic complexity -> multitask learning

Protecting deep networks from adversarial examples.

Dendritic biophysics + Kurtotic weights -> adversarial robustness

Deep learning models of the retinal response to natural scenes.

Computational reconstruction of the interior of the retina.

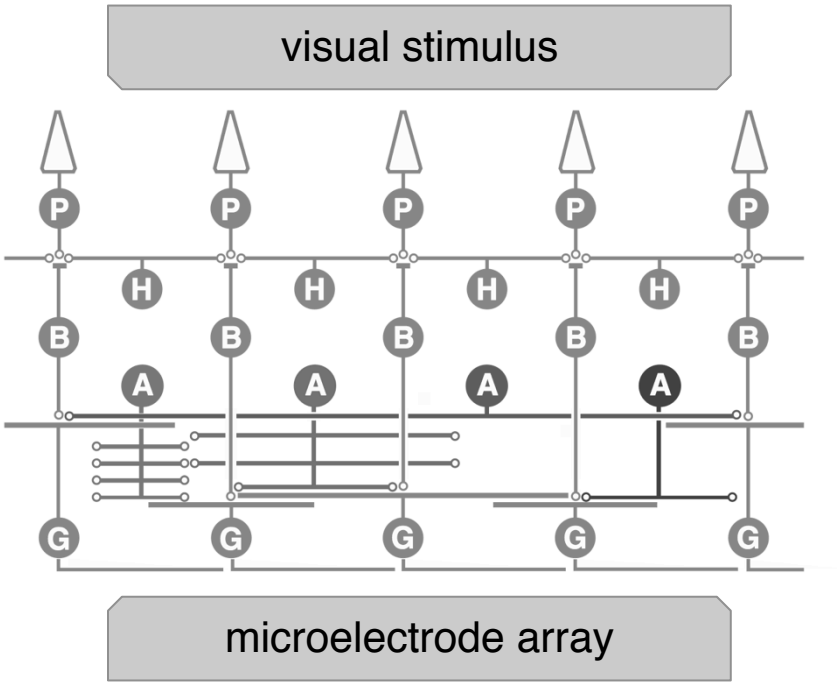
Deep neural network models of the retinal response to natural scenes

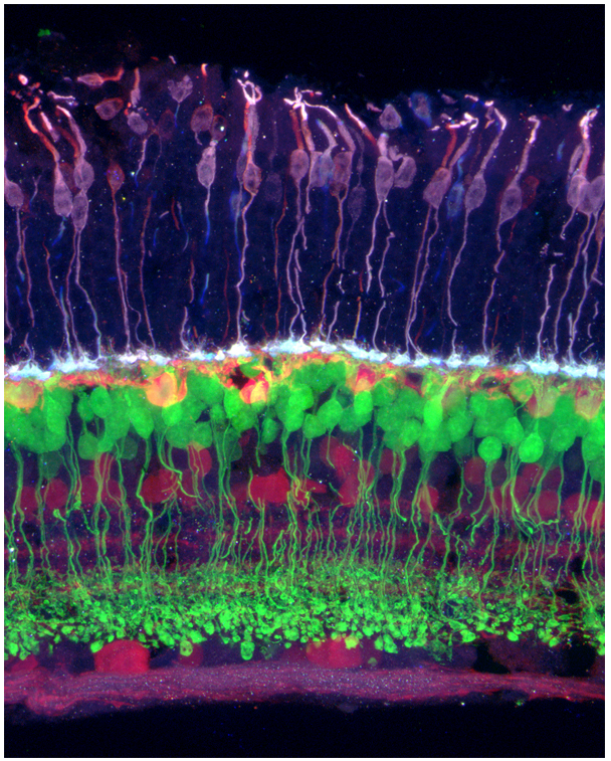


Lane McIntosh and Niru Maheswaranathan, Aran Nayebi,
Surya Ganguli and Stephen Baccus

McIntosh, L.*, Maheswaranathan, N.*, Nayebi, A., Ganguli, S., Baccus, S.A. *Deep Learning Models of the Retinal Response to Natural Scenes.* NIPS 2016.

A brief tour of the retina

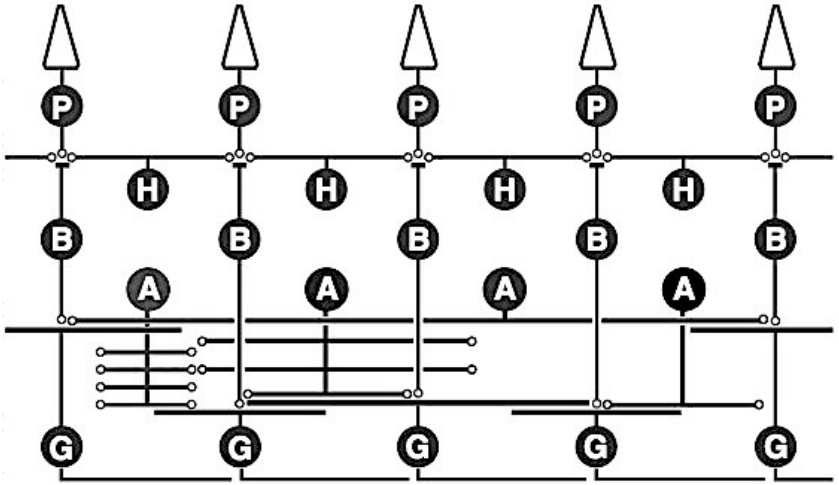




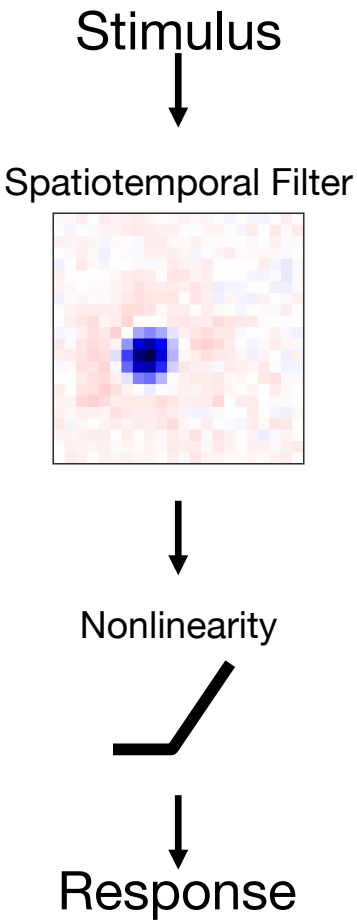
From Rachel Wong's Lab

Linear-Nonlinear models

Visual Stimulus

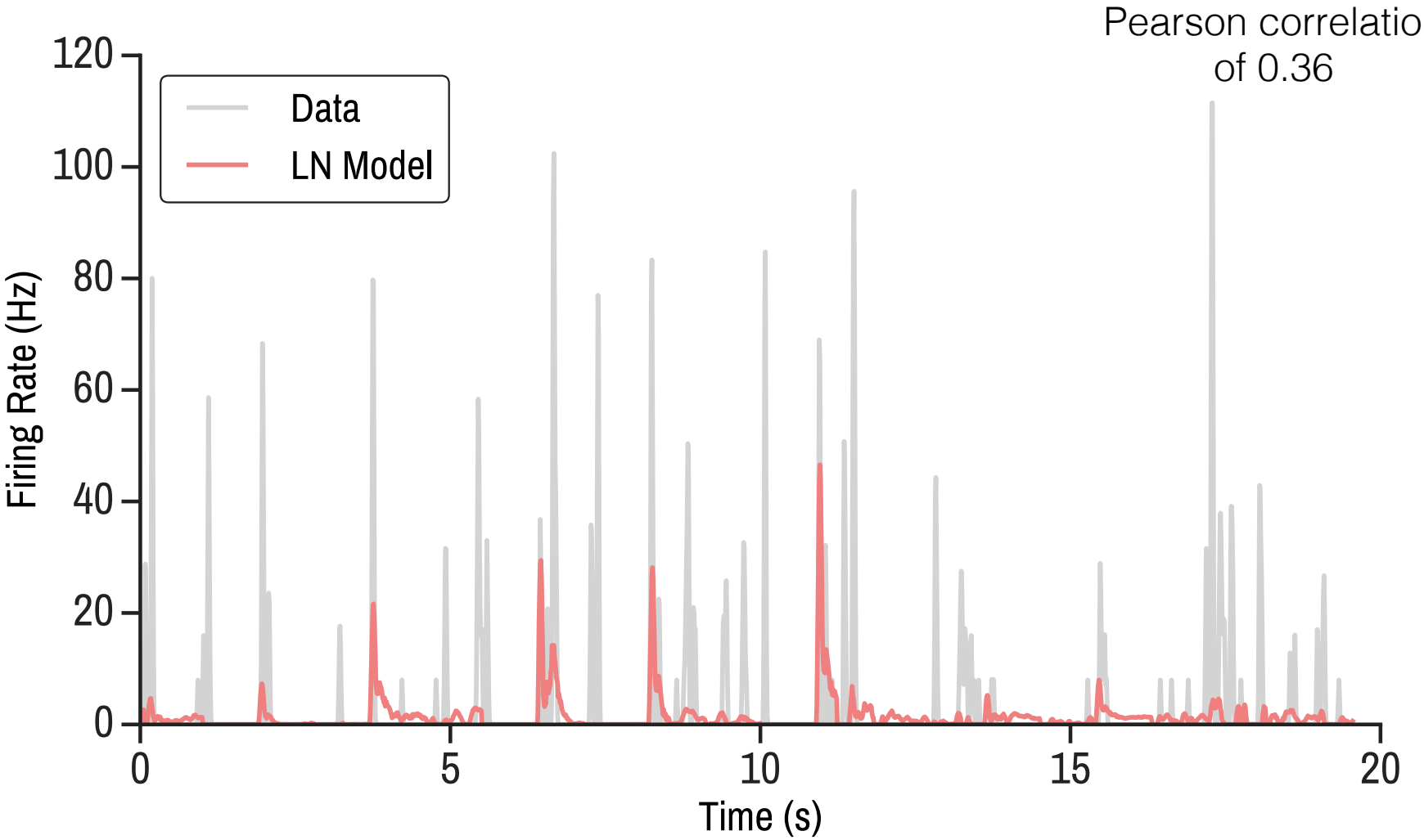


Multielectrode array (MEA)



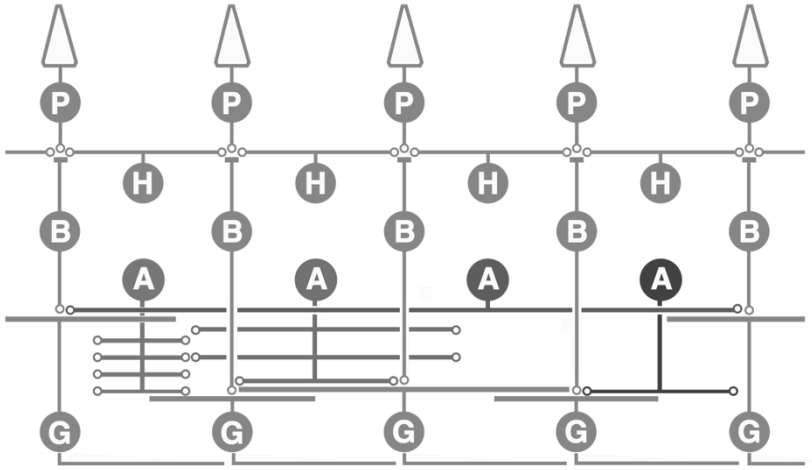
Chichilnisky 2001
Baccus and Meister 2002
Pillow et al 2005, 2008

How well do linear-nonlinear models explain the retina in natural vision?



see also
Heitman et al., 2014

Modeling ganglion cells with convolutional neural networks (CNNs)

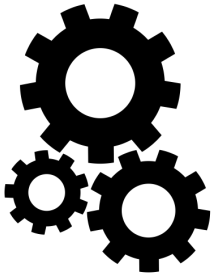


Train the model to minimize the error between predictions and recorded data

Modeling ganglion cells with convolutional neural networks (CNNs)

Challenges

trainability

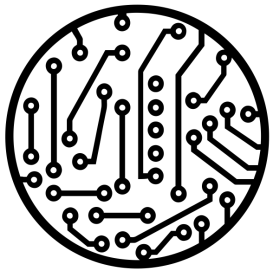


Models are complex, can easily over-fit training data

Modeling ganglion cells with convolutional neural networks (CNNs)

Challenges

neural structure



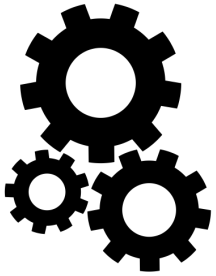
No reason why the structure or features of learned CNNs would be similar to the retina

Modeling ganglion cells with convolutional neural networks (CNNs)

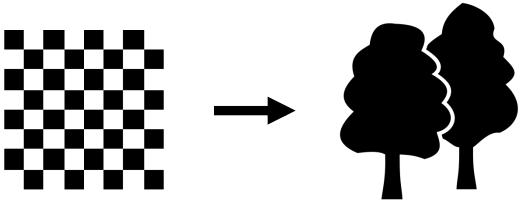
Challenges

neural function

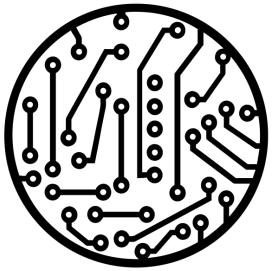
Algorithms identified by the model may not be the same as those used by the retina



CNNs capture substantially more retinal responses than previous models



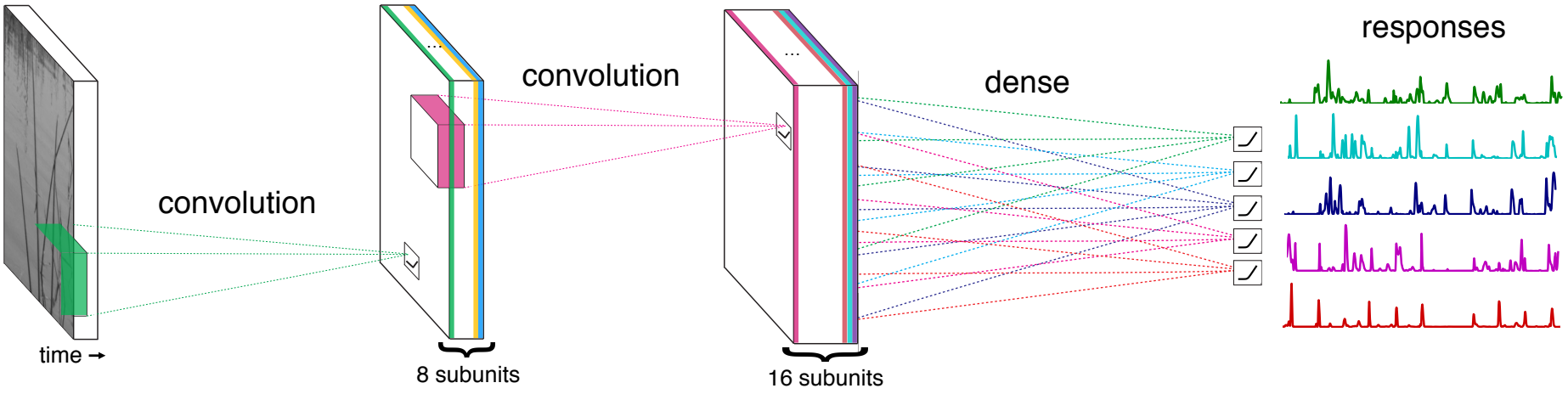
CNNs generalize better than simpler models



CNN internal units correspond to interneurons in the retinal circuitry

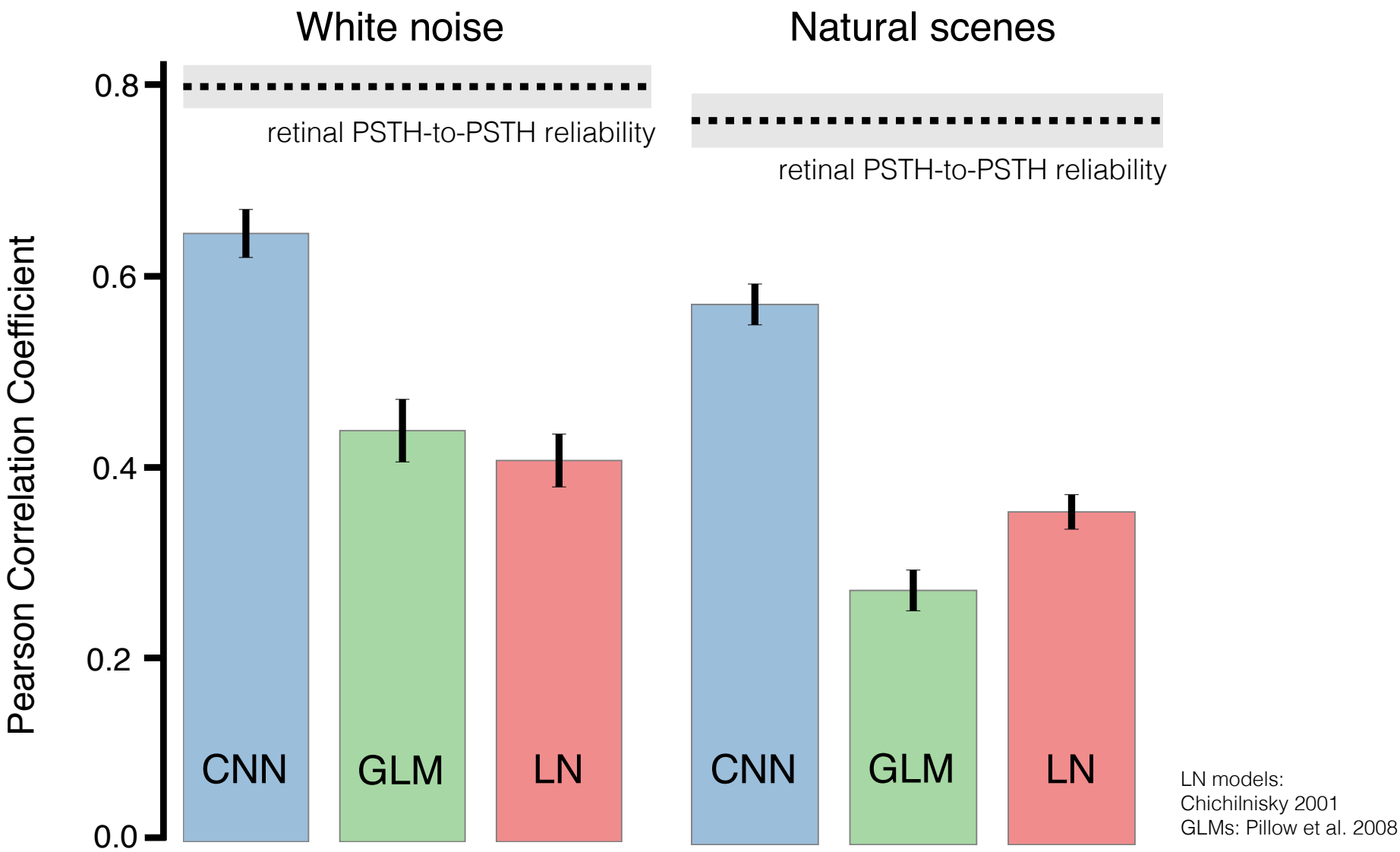
CNNs learn aspects of retinal variability, computation, and adaptation

Convolutional neural network model



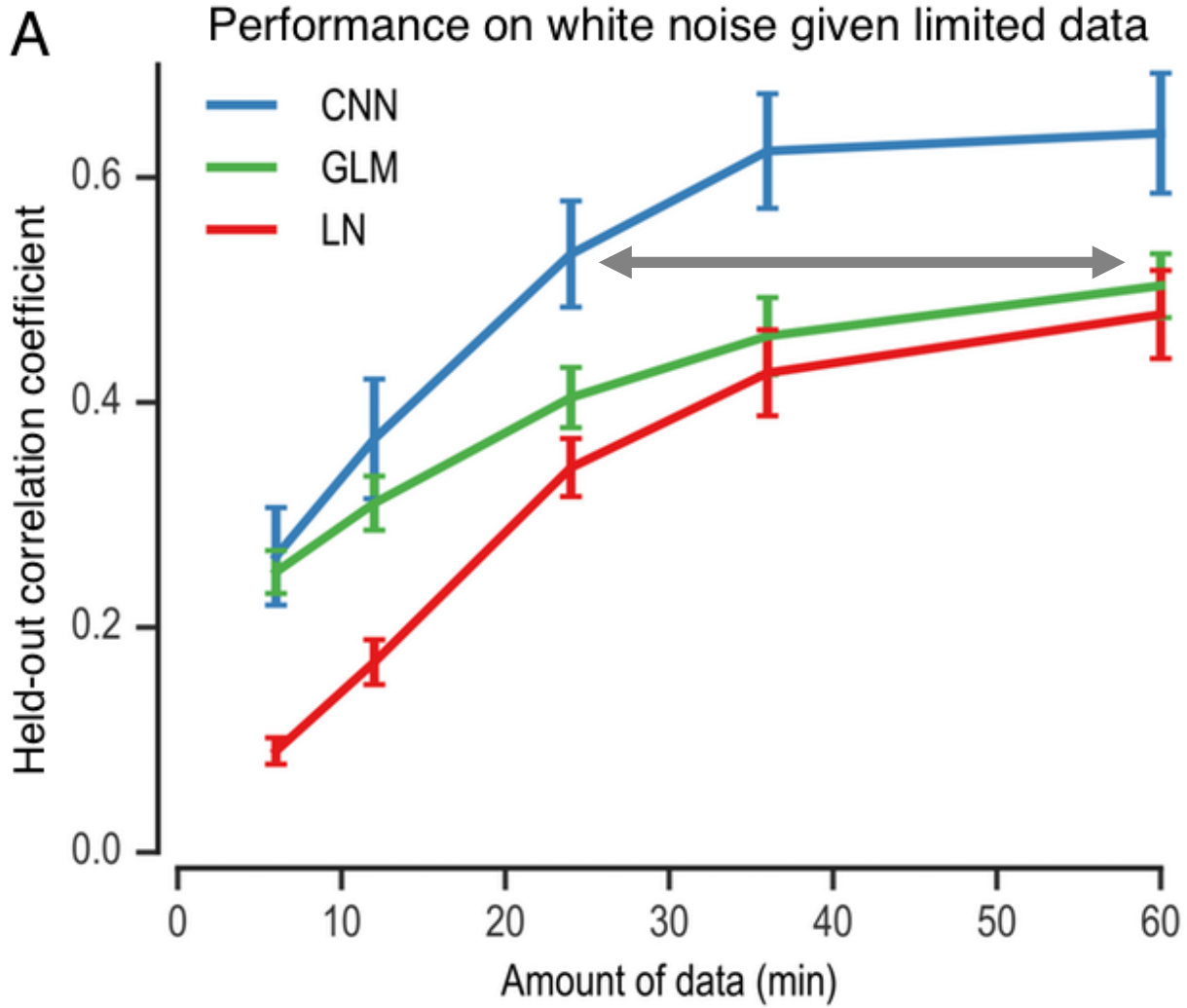
Three layers works best!

CNNs approach retinal reliability



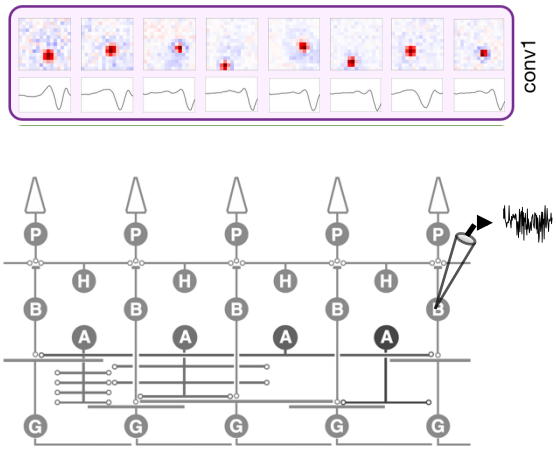
LN models:
Chichilnisky 2001
GLMs: Pillow et al. 2008

CNNs trained on less data outperform simpler models on more data

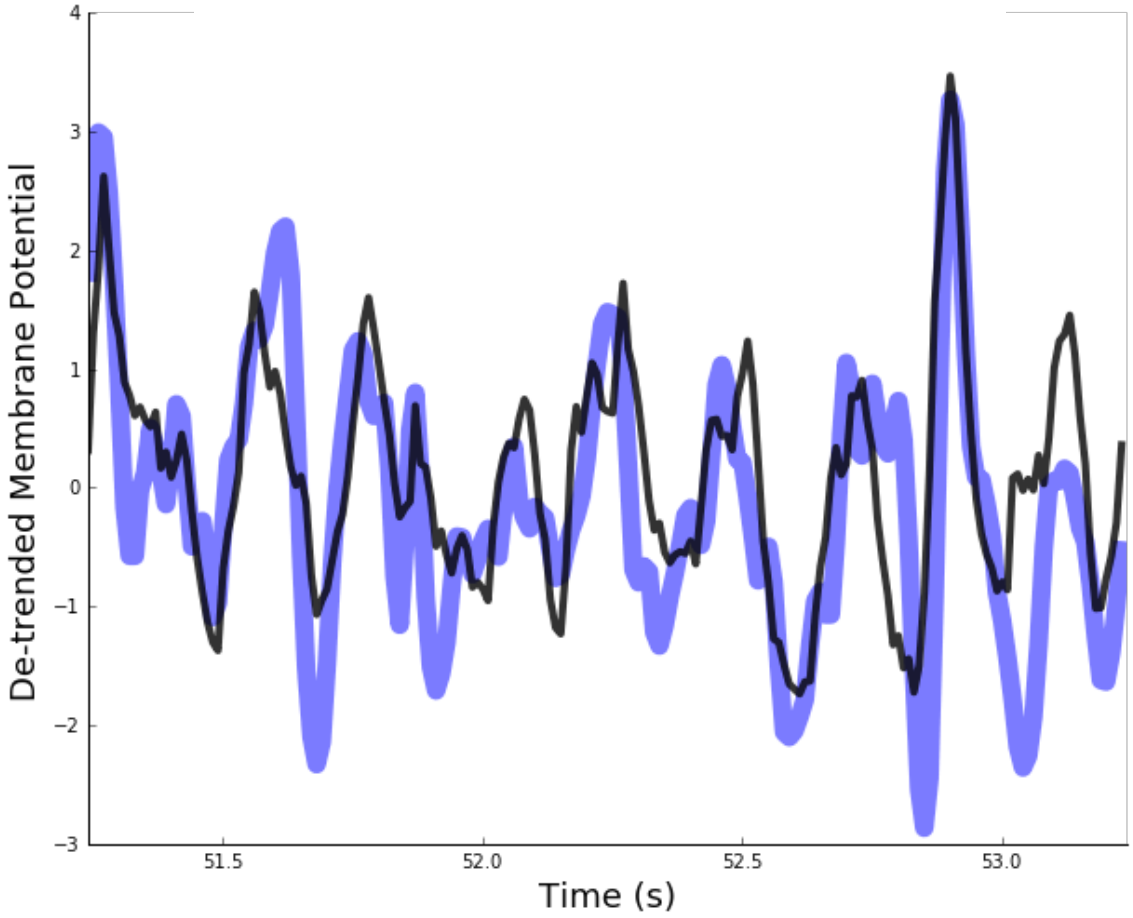


Features bear striking resemblance to internal structure in retina

CNN filters trained on white noise

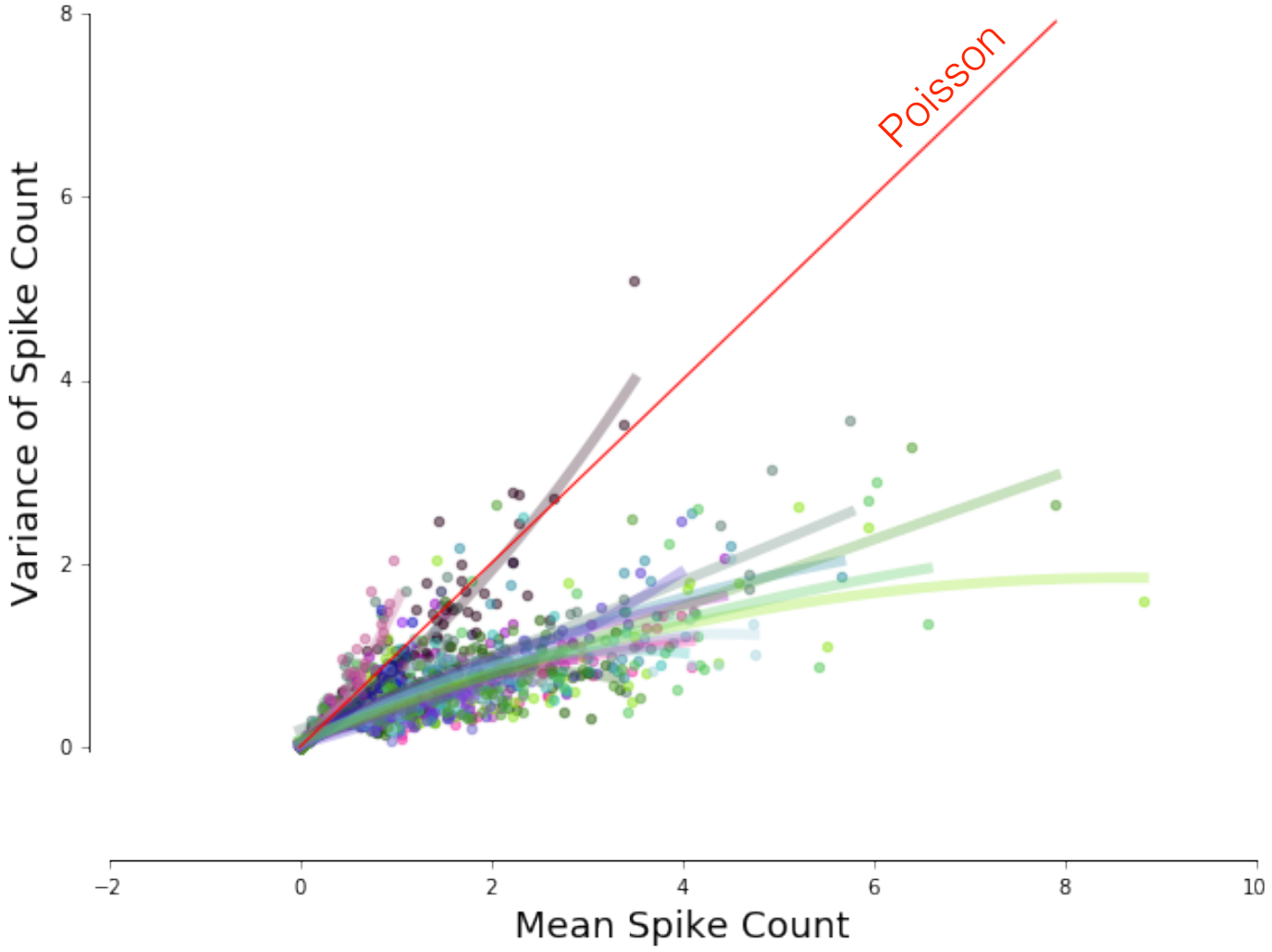


CNN first layer pre-ReLU activity
Bipolar cell membrane potential

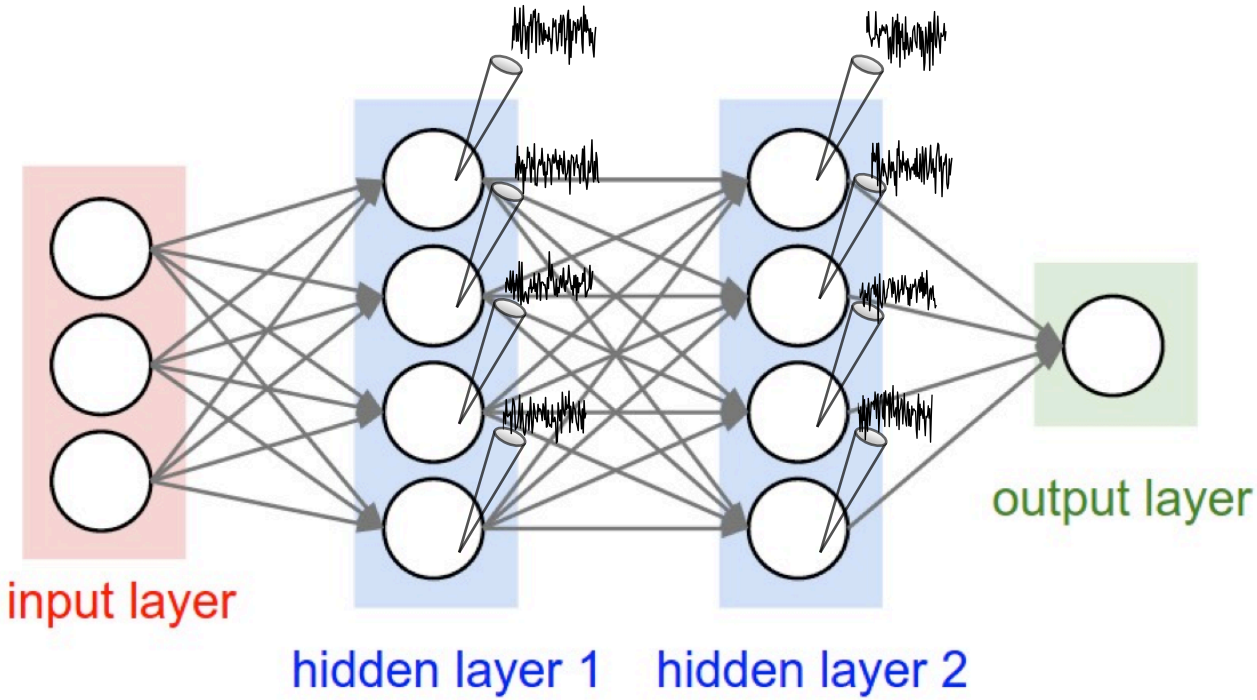


Intracellular data courtesy of
Pablo Jadzinsky and David Kastner

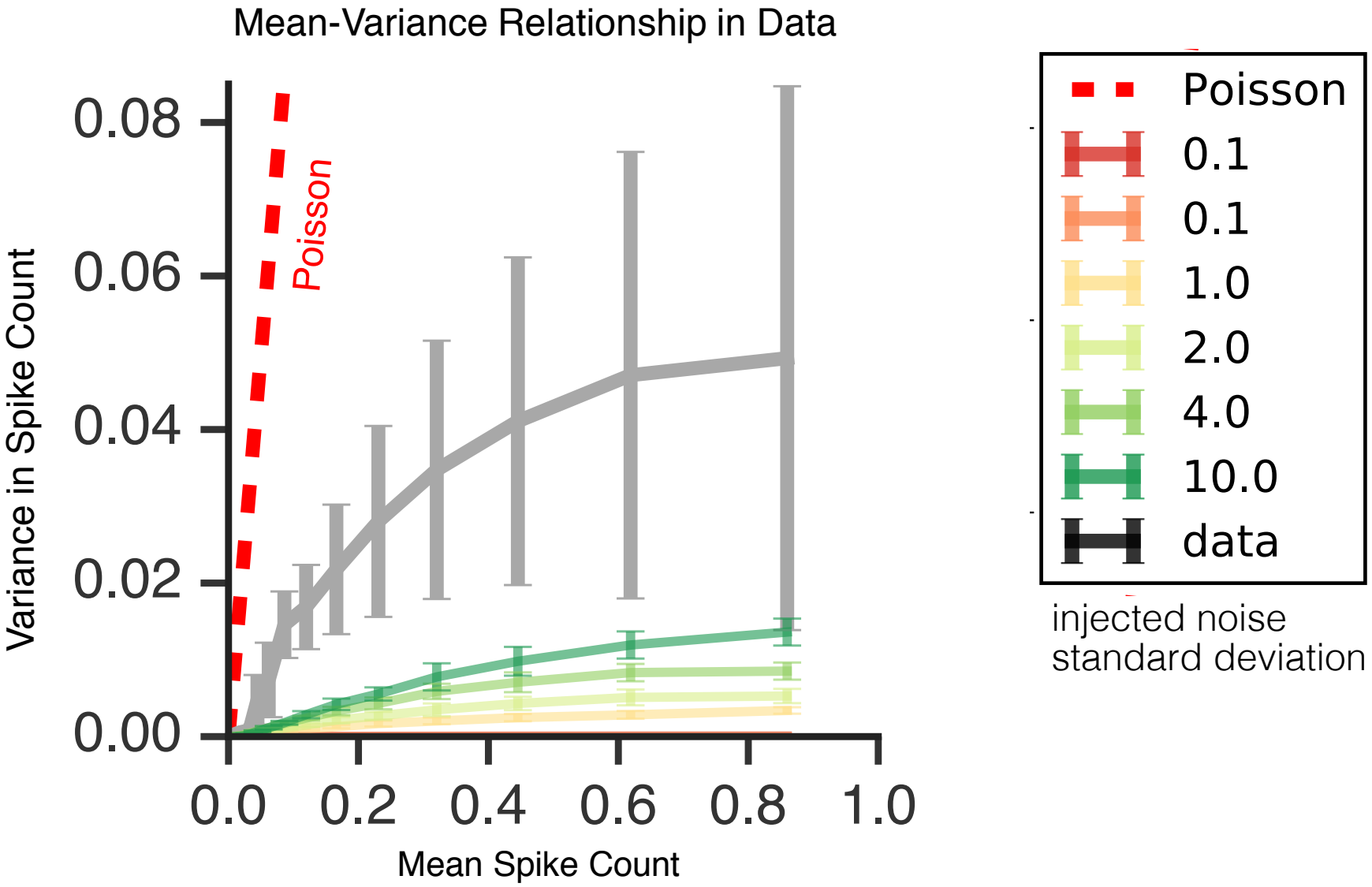
Most retinal neurons have sub-Poisson variability (while LNP models are Poisson)



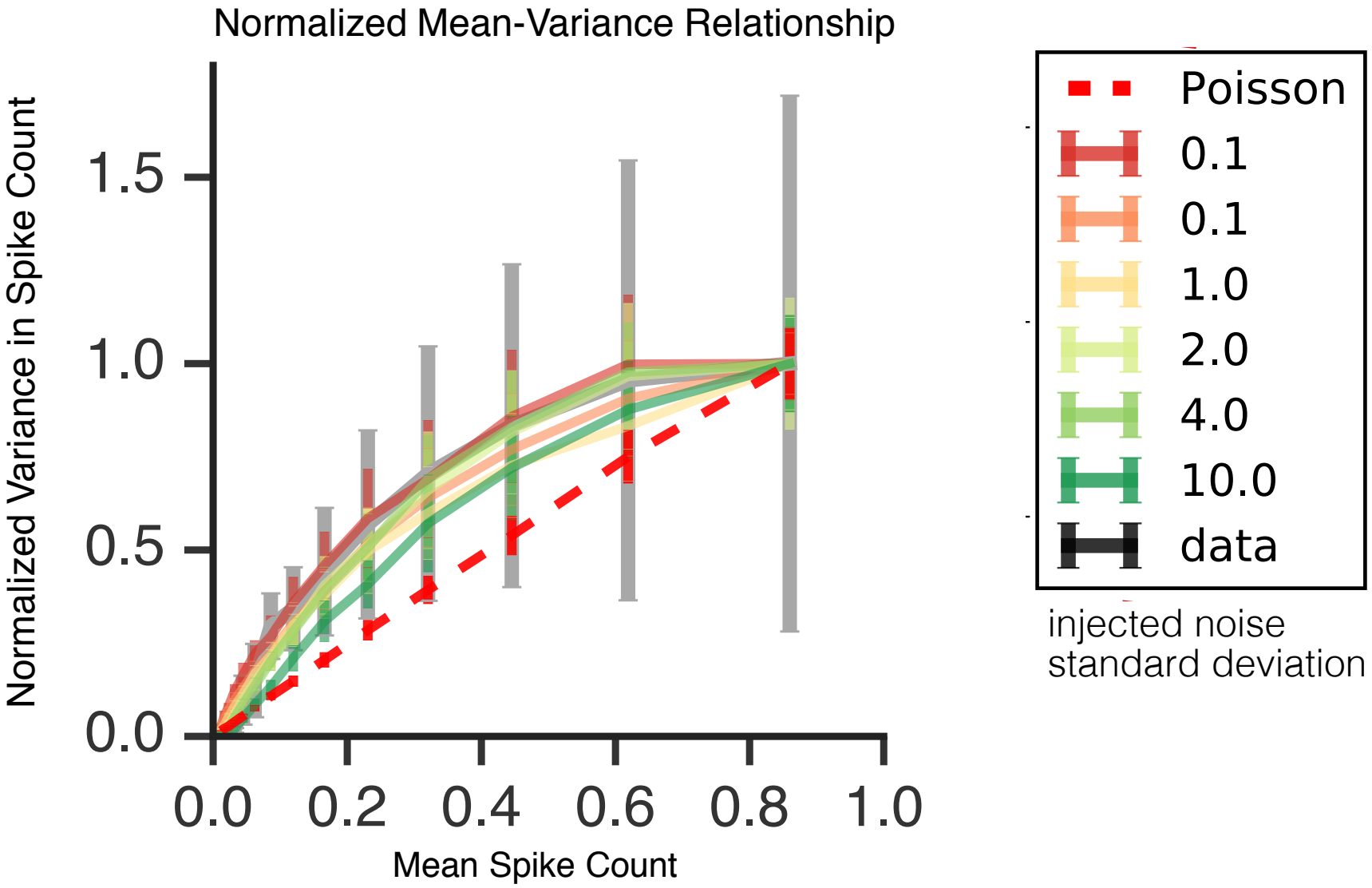
We can inject Gaussian noise into each hidden unit of our CNN model



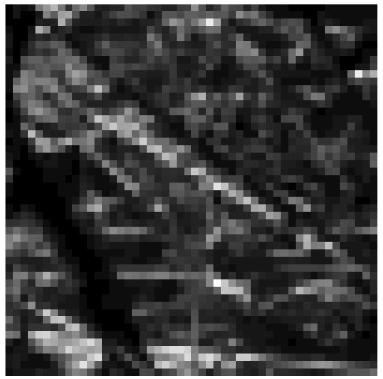
Model has lower variance than data

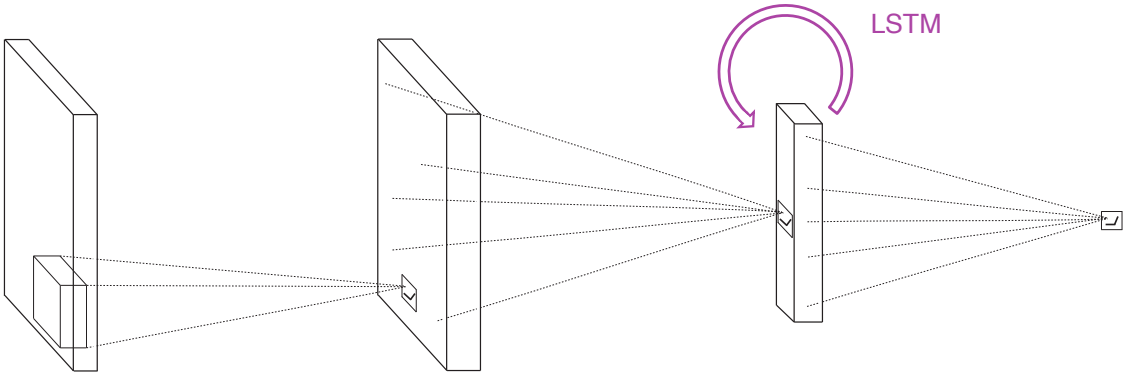


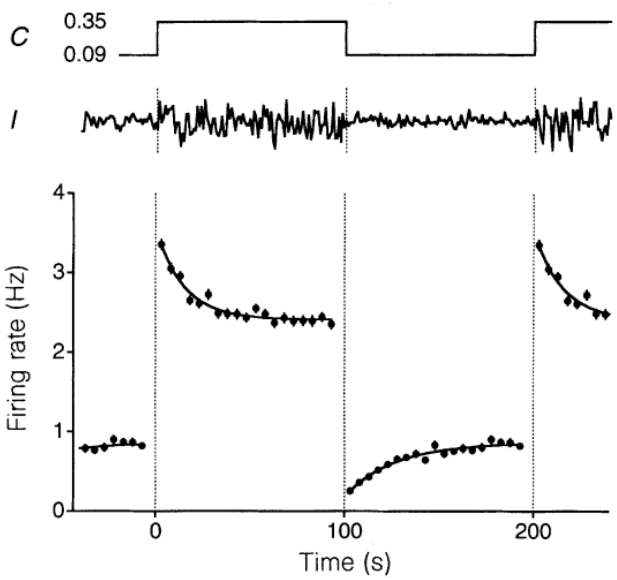
However model uncertainty has same scaling relationship as the retina



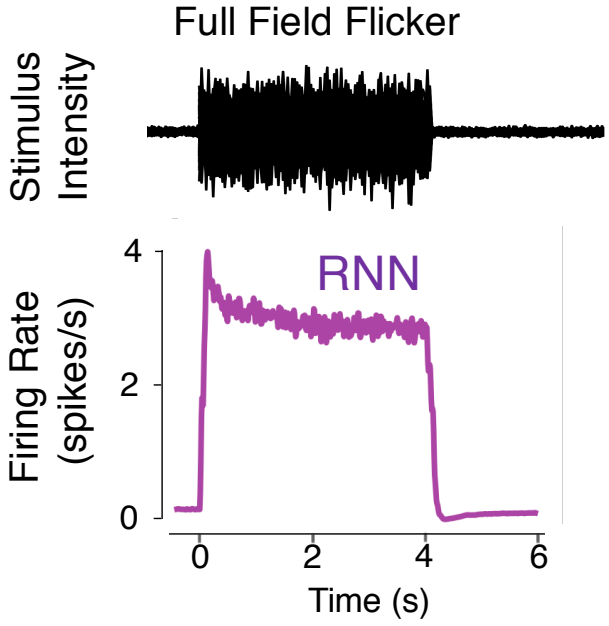
Capturing contrast adaptation from retinal responses to natural scenes





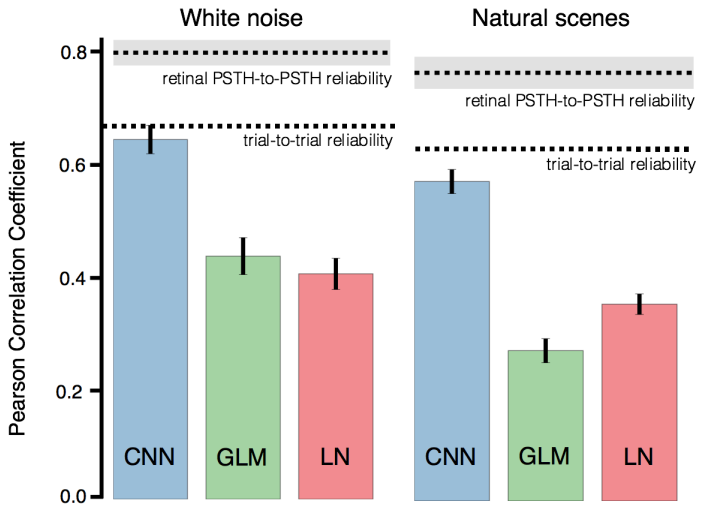


Smirnakis et al., 1997

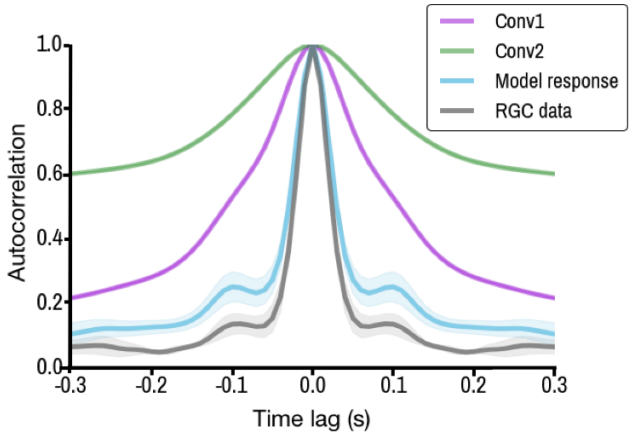


Aran Nayebi

Summary

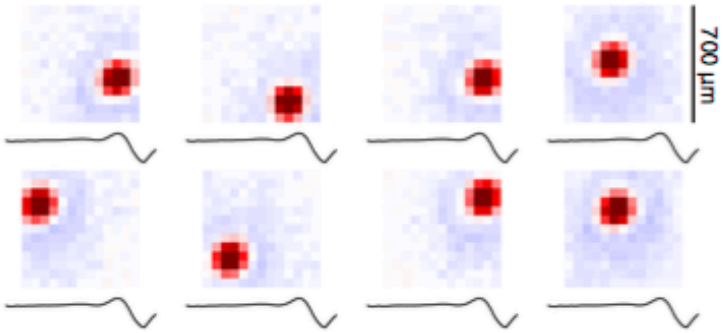


CNNs learn the internal, nonlinear structure of the retina



CNNs capture substantially more retinal responses than previous models.

CNNs also generalize better to different stimuli classes.



We can capture not only the mean response, but also how variability scales with the mean

Our CNN models reproduce principles of signal processing inside retina without having direct access to it!