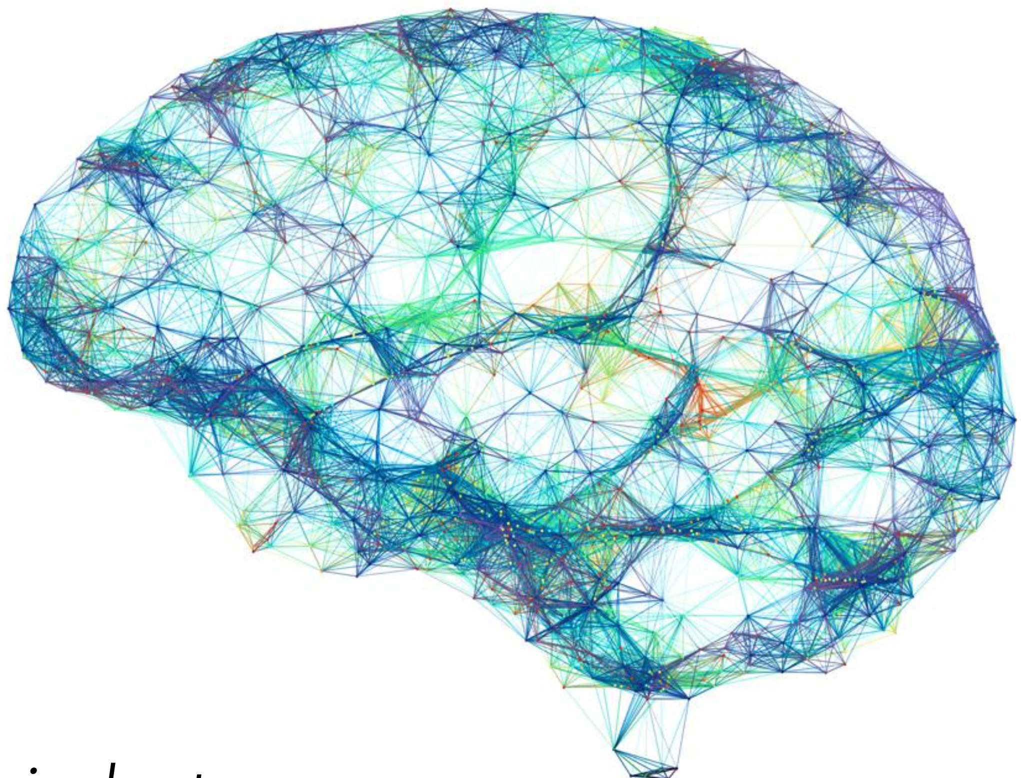
Connecting Variational Autoencoders *Back to the Brain*

arXiv:2011.07464

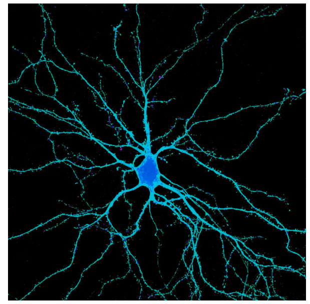
Joseph Marino
California Institute of Technology



what is the relationship between deep learning & neuroscience?

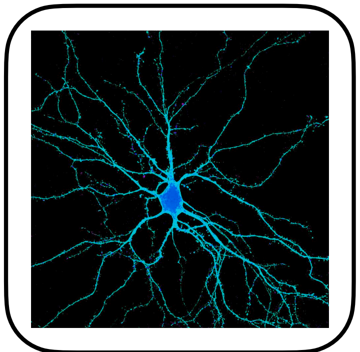
TALK OUTLINE

I. background

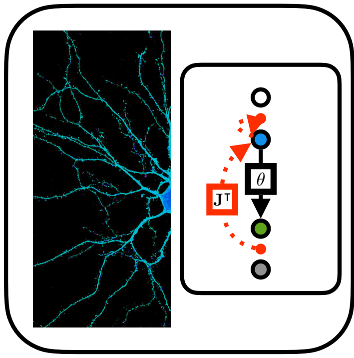


TALK OUTLINE

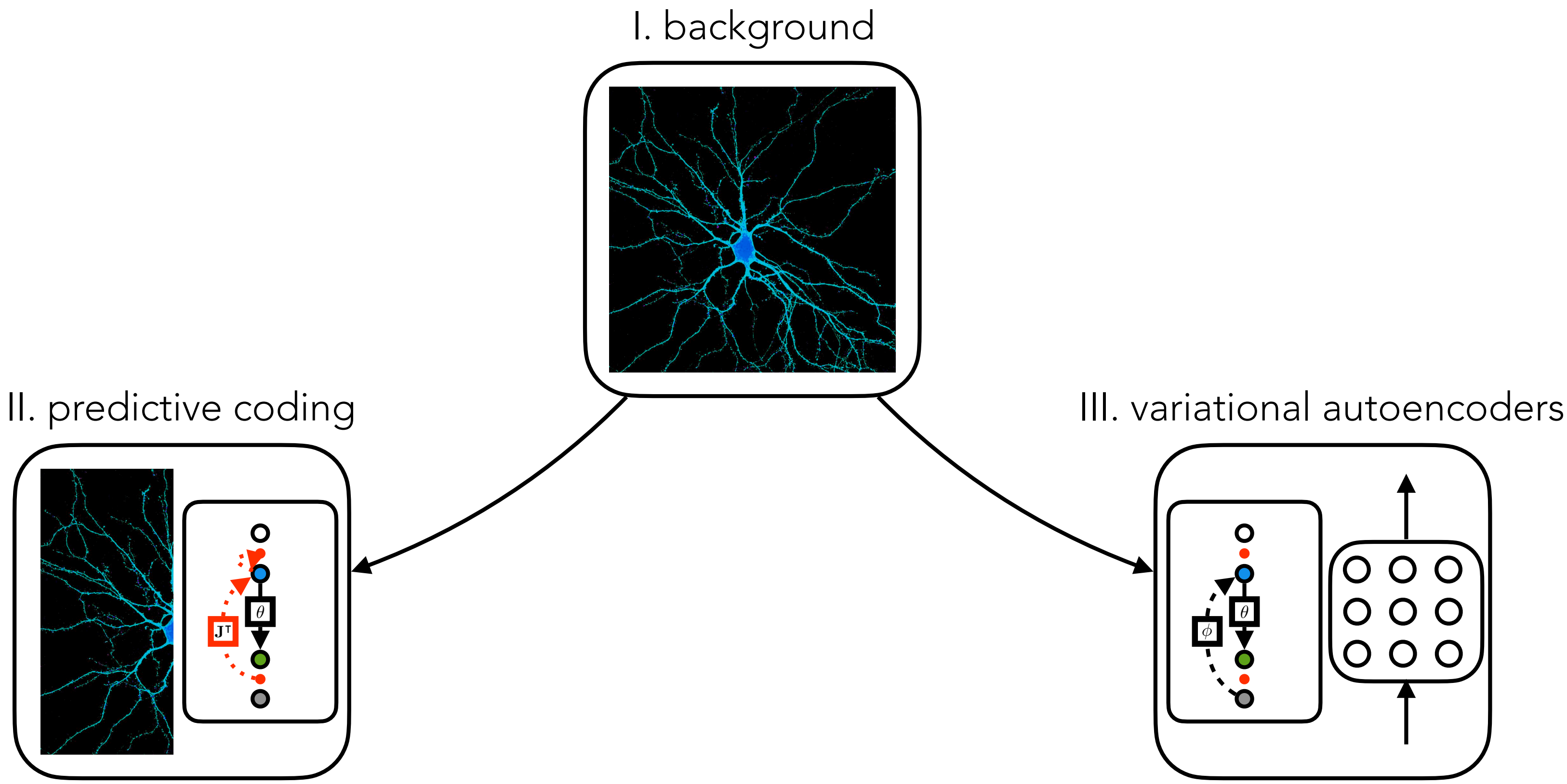
I. background



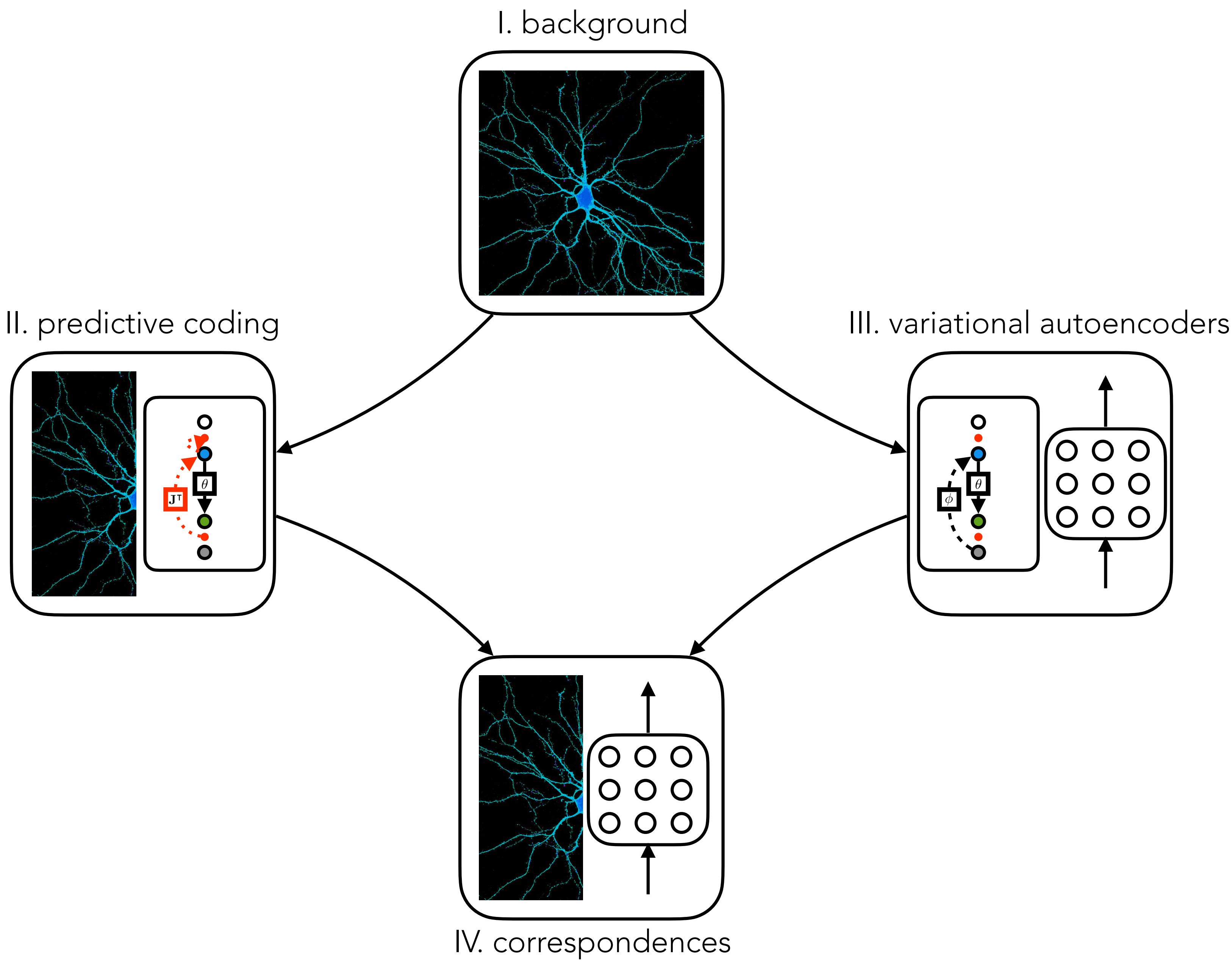
II. predictive coding



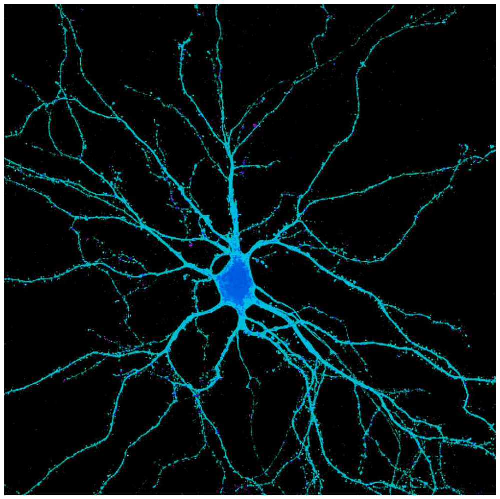
TALK OUTLINE

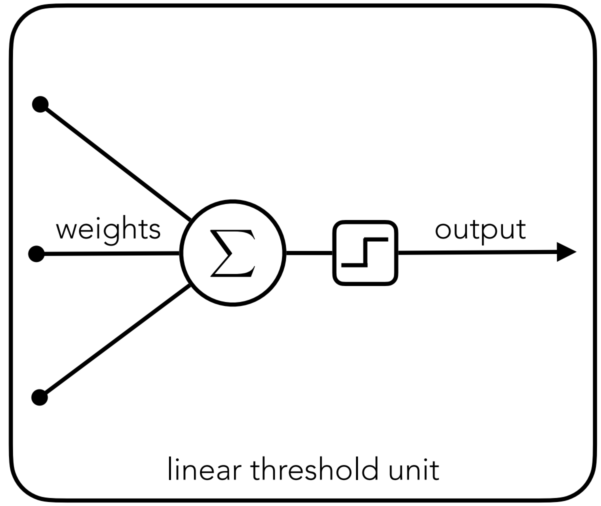


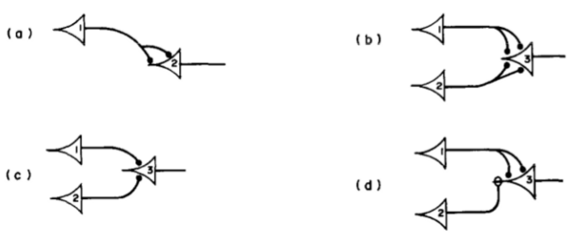
TALK OUTLINE



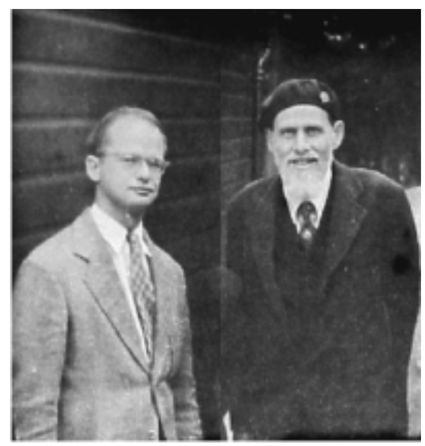
ARTIFICIAL NEURAL NETWORKS







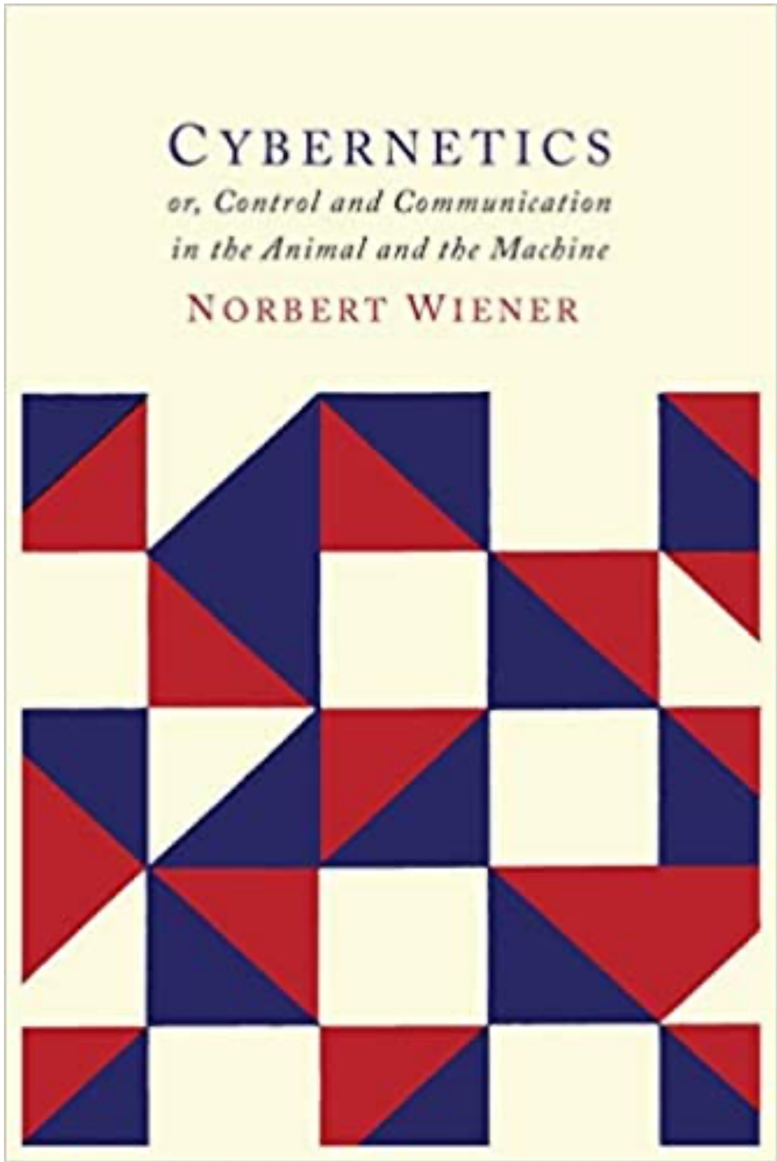
Logical Calculus for Nervous Activity
McCulloch & Pitts, 1943

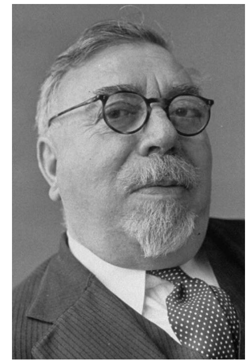


Walter Pitts   Warren McCulloch

cybernetics

CYBERNETICS





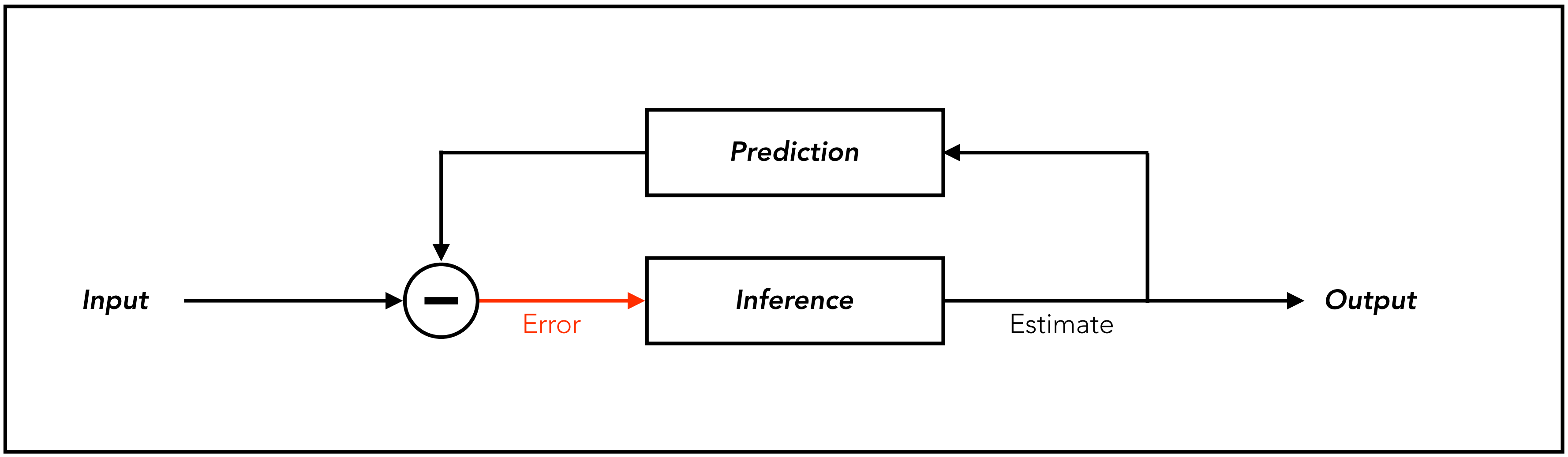
Norbert Wiener

- defined information
- probabilistic models
- variational optimization
- control & perception
- connections to neuroscience

Wiener, 1948

CYBERNETICS

comparator circuit



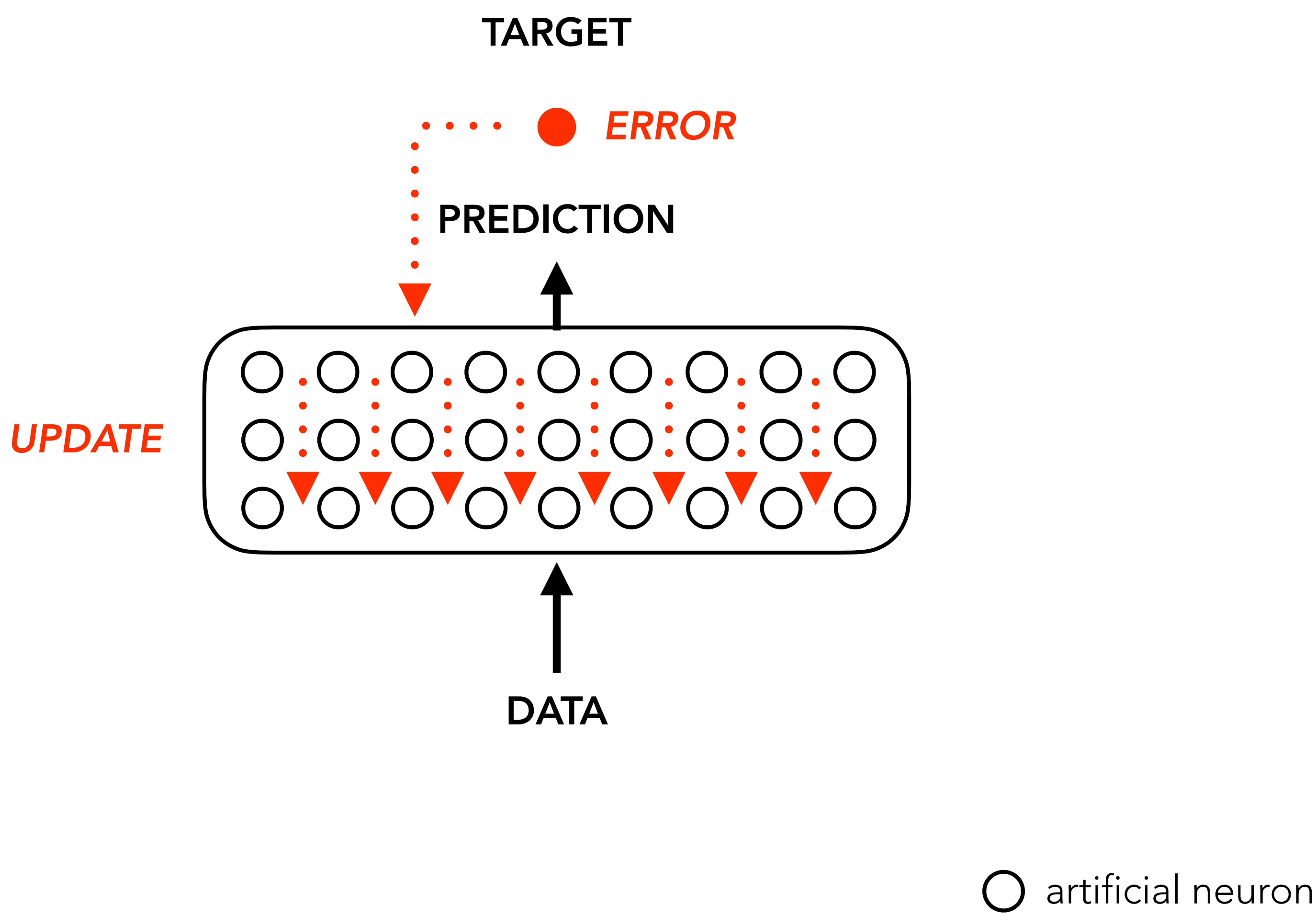
negative feedback: *use <u>errors</u> to correct estimates,*

e.g., Kalman filter, linear quadratic regulator, PID

Wiener, 1948

DEEP LEARNING

artificial neural networks + negative feedback

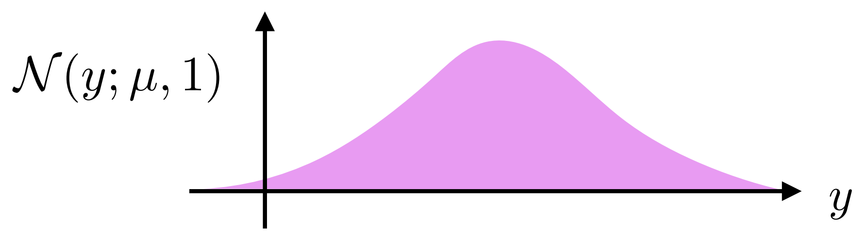


PROBABILISTIC MODELS

why error?

gradient of log-probability for exponential family distributions

example: Gaussian density



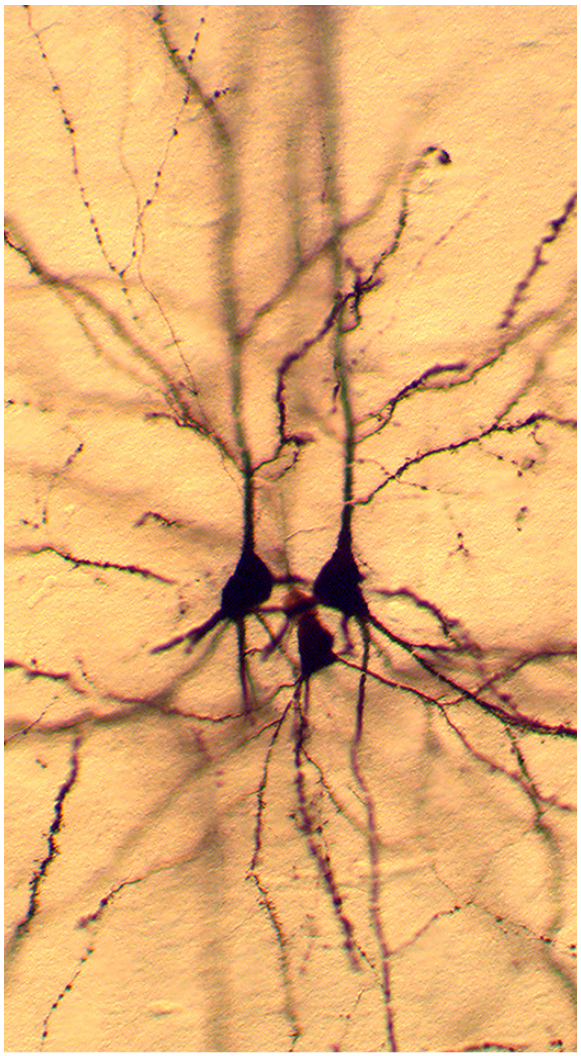
$$\log \mathcal{N}(y;\mu,1) = \frac{1}{2}(\mu - y)^2 + \text{const.}$$

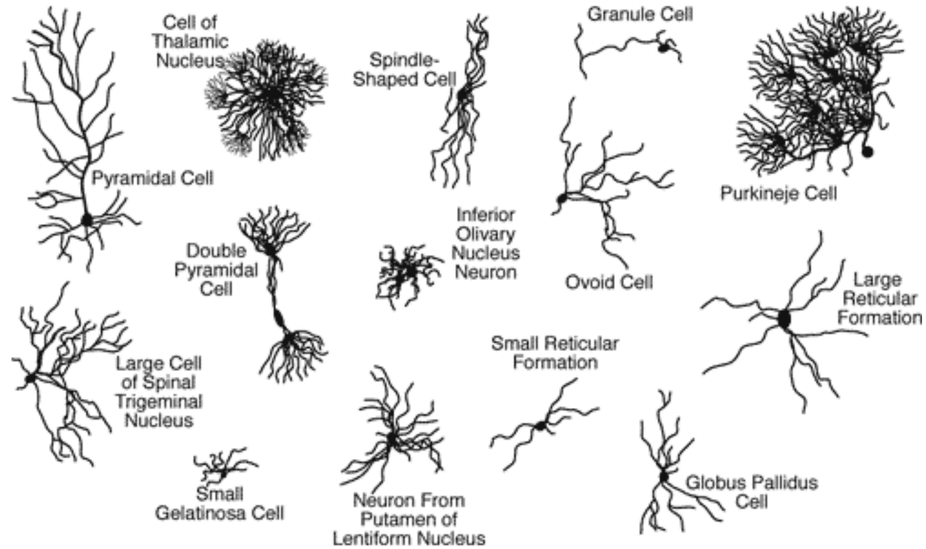
$$\frac{d}{d\mu} \log \mathcal{N}(y;\mu,1) = \mu - y$$

improving predictions requires evaluating and reducing errors

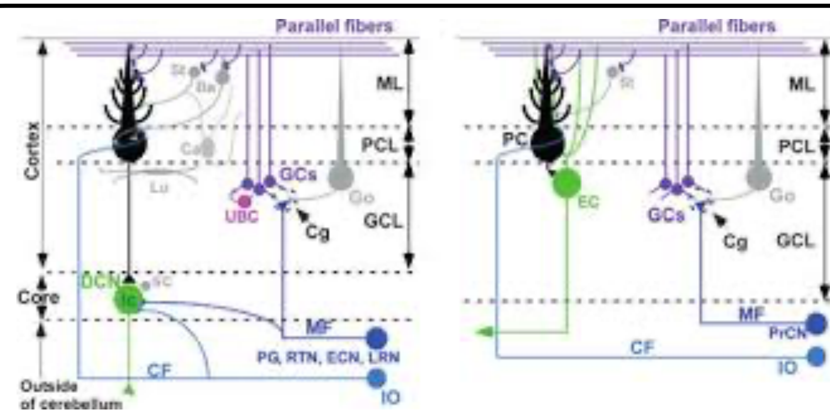
NEUROSCIENCE

neurons are **complex**



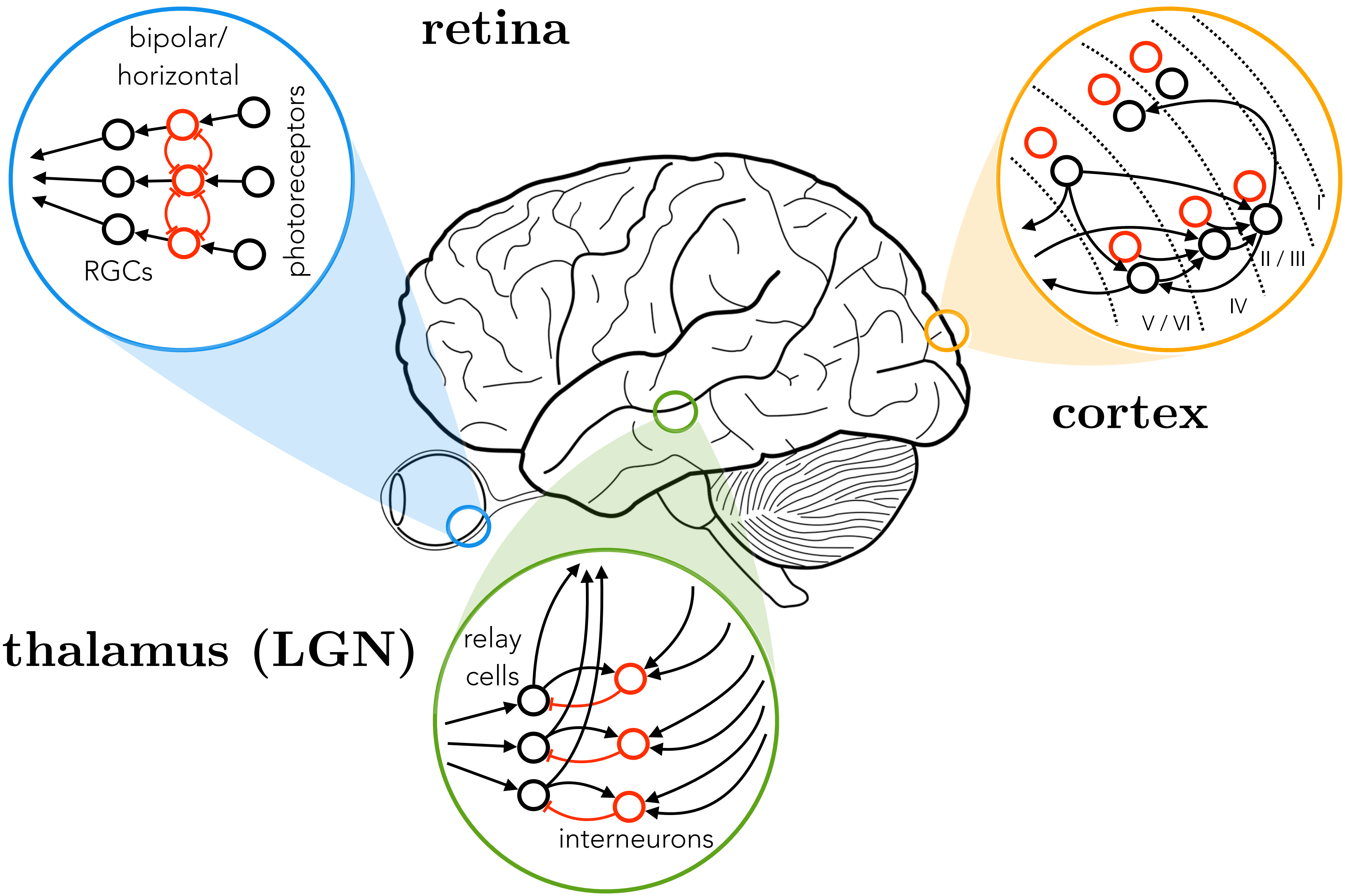


extremely **diverse**

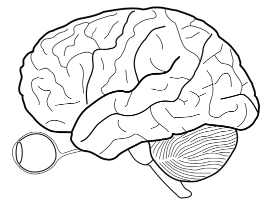


and arranged in intricate **circuits**

VISUAL PATHWAY

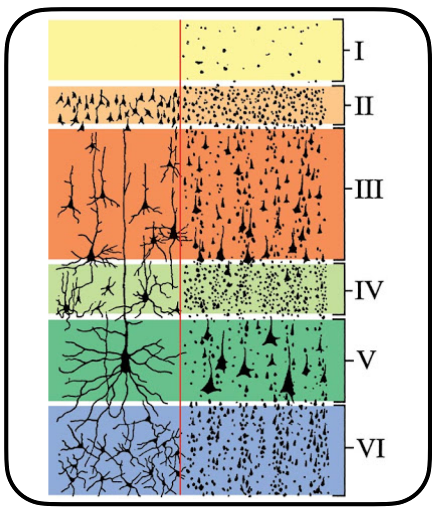


NEOCORTEX

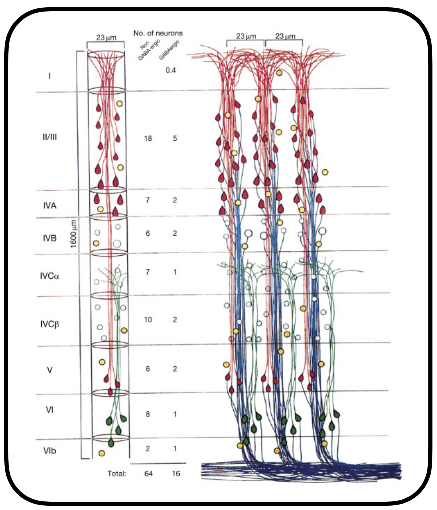


neocortex is involved in all "higher-level" sensorimotor processing

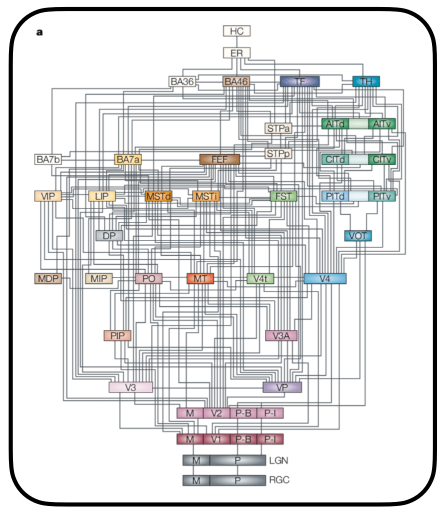
layered structure



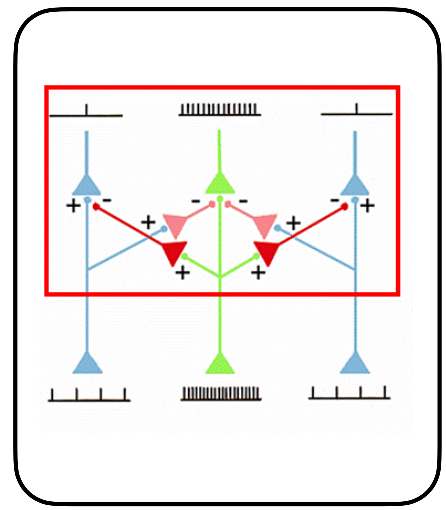
columns of neurons



hierarchical



lateral inhibition

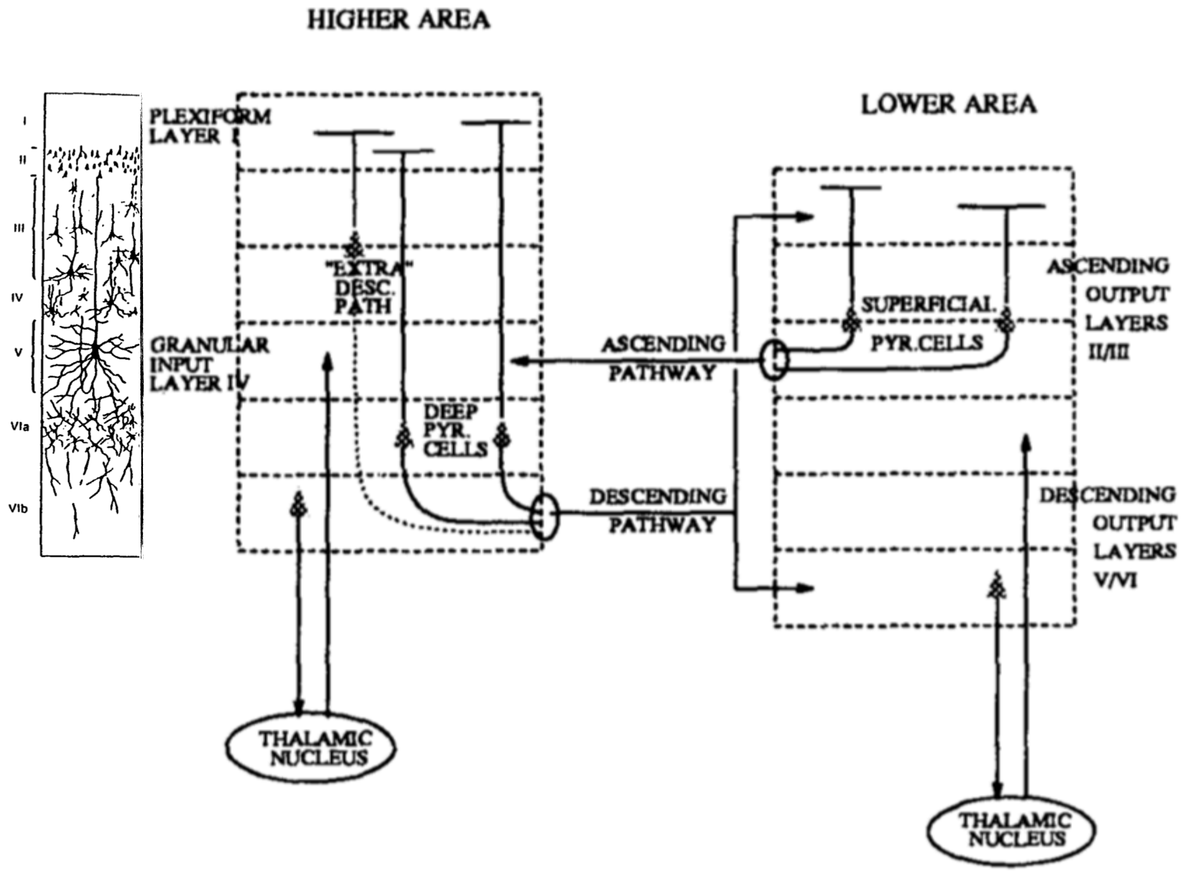


basic circuit motif **forward/backward** **normalization**

backward (top-down) connections outnumber forward (bottom-up) connections!

II. PREDICTIVE CODING

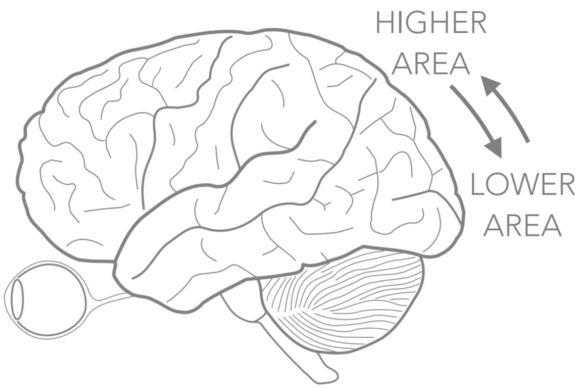
HIERARCHICAL PREDICTIVE CODING

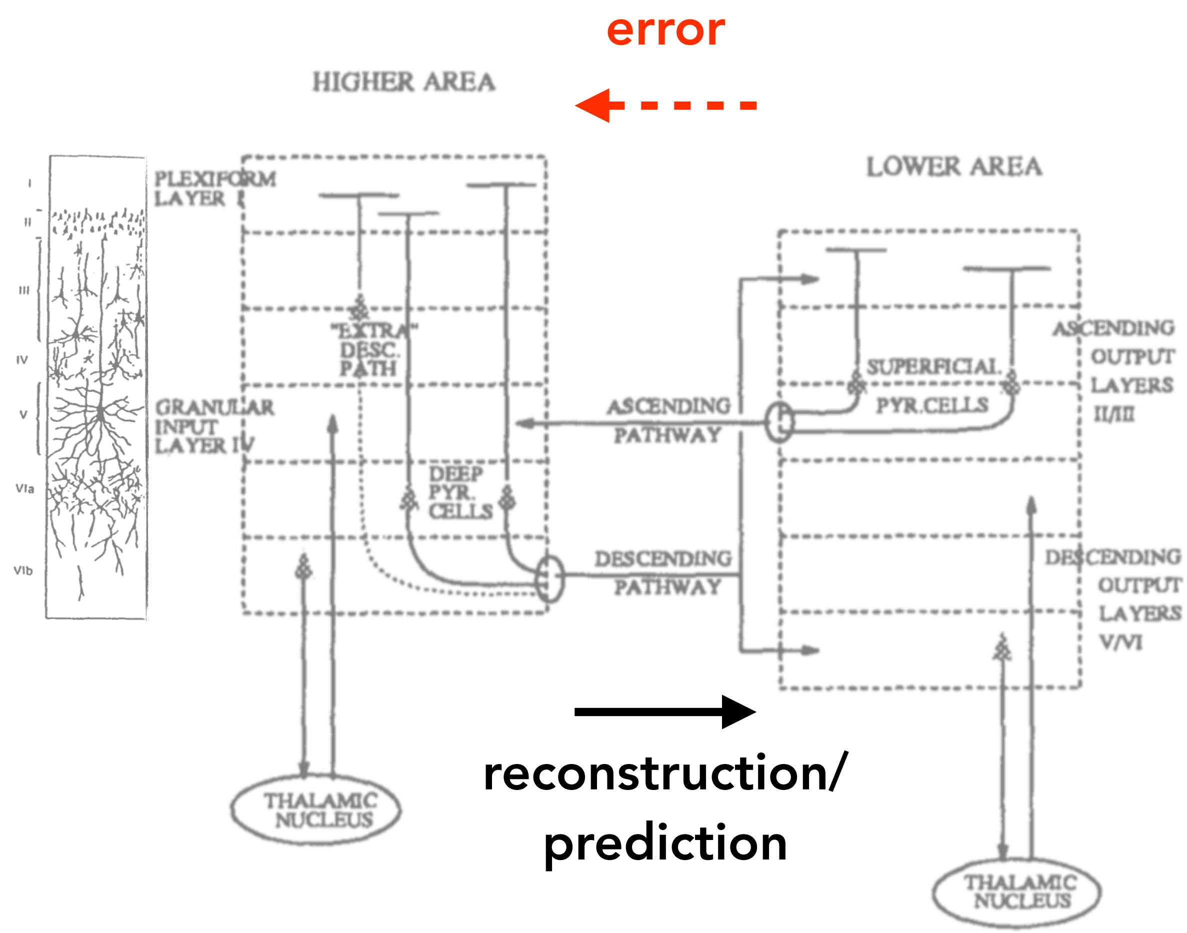


Thalamus "plays the role of an 'active blackboard' on which the current best ***reconstruction*** *of some aspect of the world is always displayed."*

Mumford, *Biological Cybernetics* (1991)

HIERARCHICAL PREDICTIVE CODING



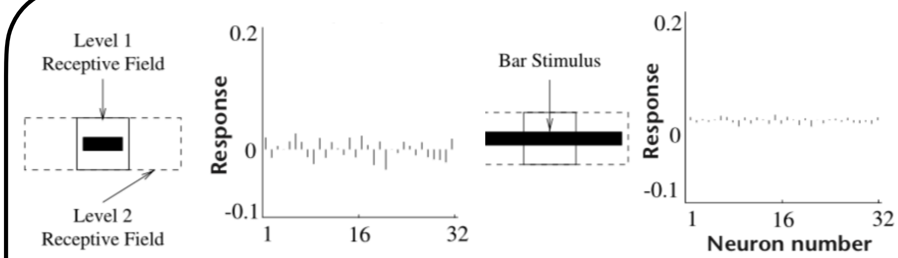


*Thalamus "plays the role of an 'active blackboard' on which the current best **reconstruction** of some aspect of the world is always displayed."*

Mumford, *Biological Cybernetics* (1991)

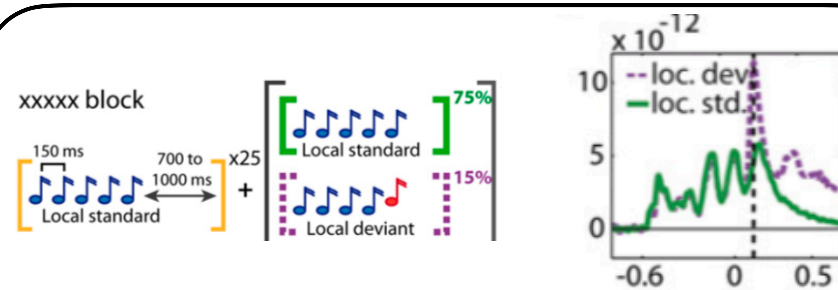
HIERARCHICAL PREDICTIVE CODING

some evidence in support of predictive coding



top-down effects in visual processing, e.g. "end-stopping"

Rao & Ballard, 1999



auditory sequence prediction errors

Wacongne et al., 2011

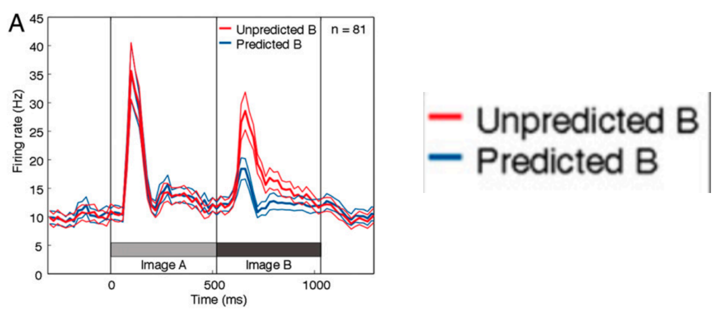
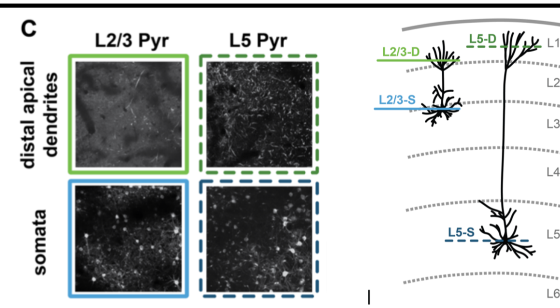


image sequence prediction errors

Meyer & Olson, 2011

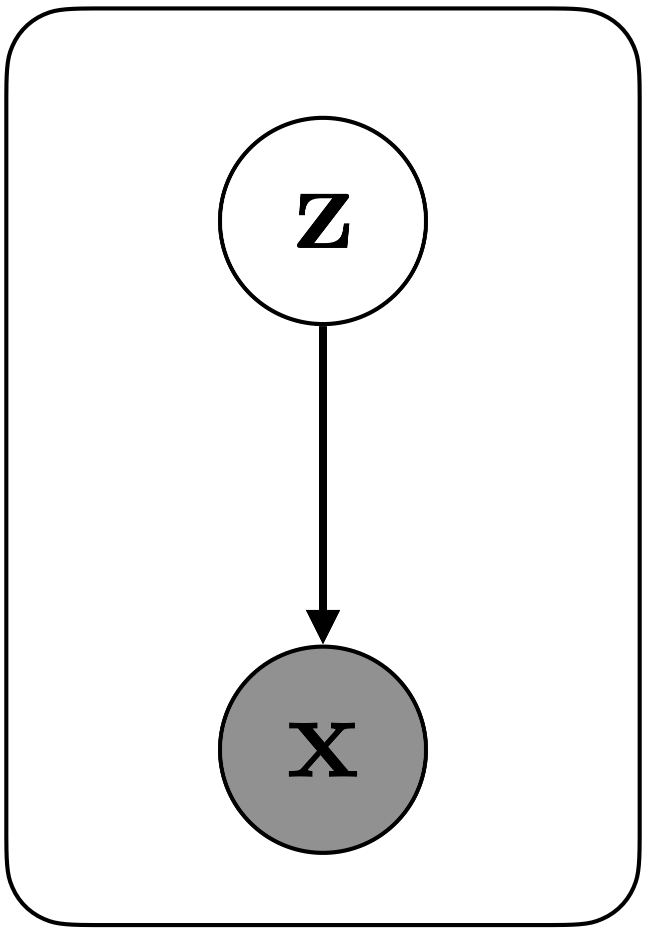


targeted top-down predictions

Gillon et al, 2021

HIERARCHICAL PREDICTIVE CODING

we can formalize this process as *probabilistic modeling & inference*



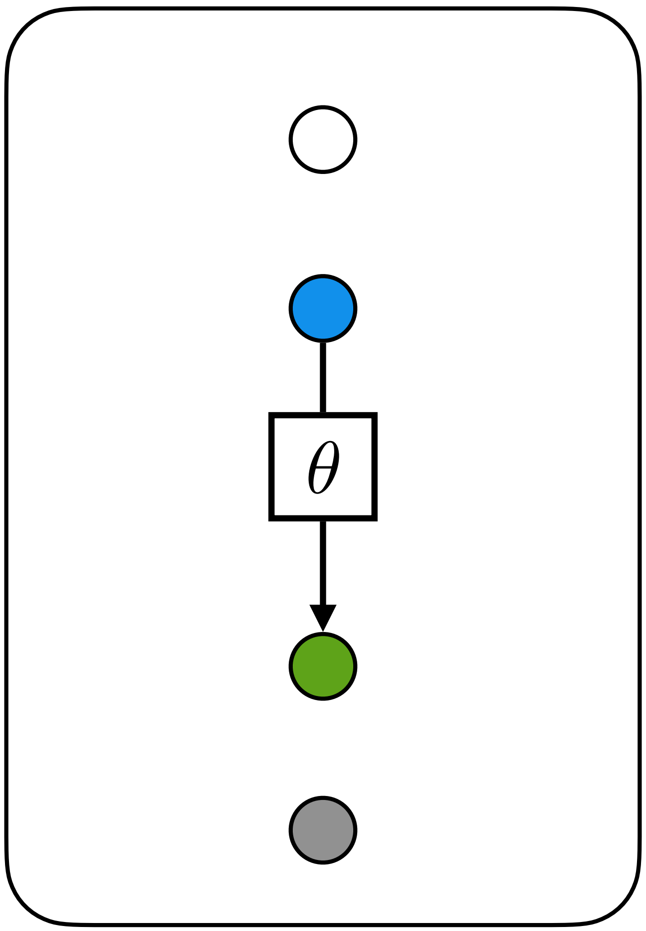
$$p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$$

latent variable model

Rao & Ballard, 1999

HIERARCHICAL PREDICTIVE CODING

we can formalize this process as *probabilistic modeling & inference*



$p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\mathbf{z}, \mathrm{diag}(\boldsymbol{\sigma}^2_\mathbf{z}))$

$q(\mathbf{z}|\mathbf{x}) = \delta(\hat{\mathbf{z}})$

prediction/reconstruction

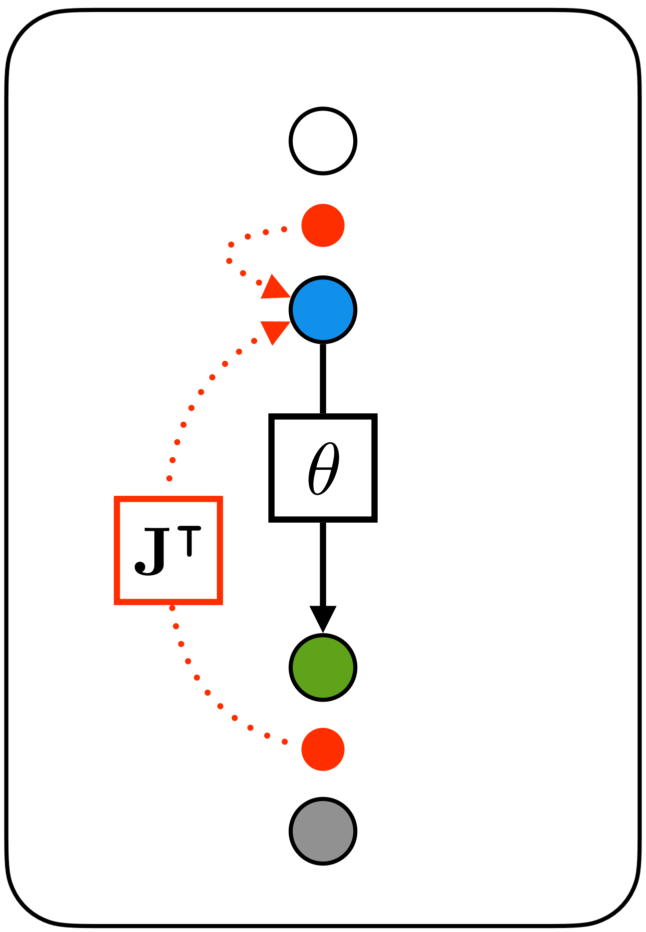
$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\mathbf{x}(\mathbf{z}), \mathrm{diag}(\boldsymbol{\sigma}^2_\mathbf{x}))$

$\mathbf{x} \sim p_{\mathrm{data}}(\mathbf{x})$

Rao & Ballard, 1999

HIERARCHICAL PREDICTIVE CODING

we can formalize this process as *probabilistic modeling & inference*



$$p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_\mathbf{z}, \mathrm{diag}(\boldsymbol{\sigma}^2_\mathbf{z}))$$

$$q(\mathbf{z}|\mathbf{x}) = \delta(\hat{\mathbf{z}})$$

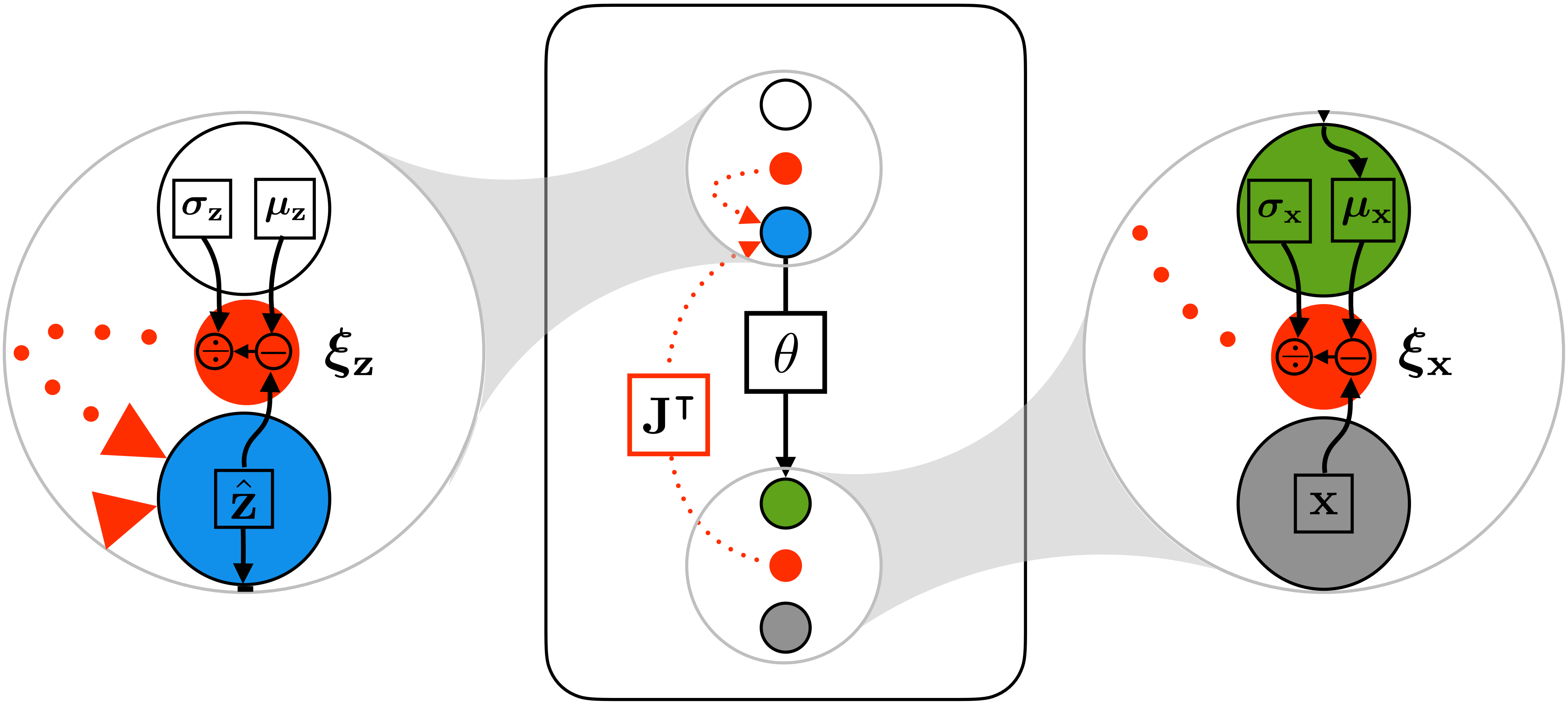
prediction/reconstruction

$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_\mathbf{x}(\mathbf{z}), \mathrm{diag}(\boldsymbol{\sigma}^2_\mathbf{x}))$$

$$\mathbf{x} \sim p_{\mathrm{data}}(\mathbf{x})$$

inference: maximize a (variational) objective $\mathcal{L}$ w.r.t.

HIERARCHICAL PREDICTIVE CODING



inference involves a sum of (prediction) errors

$$\nabla_{\hat{\mathbf{z}}}\mathcal{L} = \mathbf{J}^{\intercal}\boldsymbol{\xi}_{\mathbf{x}} - \boldsymbol{\xi}_{\mathbf{z}}$$

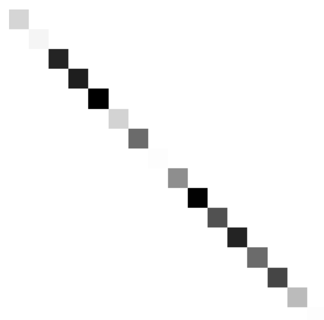
Rao & Ballard, 1999

SPATIOTEMPORAL PREDICTIVE CODING

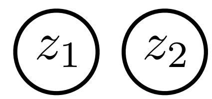
lateral inhibition implements a covariance matrix/normalization

independent dimensions: $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$

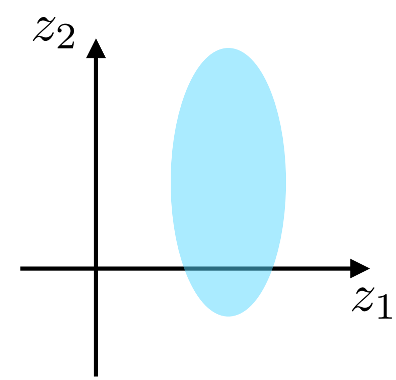
$\mathrm{diag}(\boldsymbol{\sigma}^2) =$



in 2D:

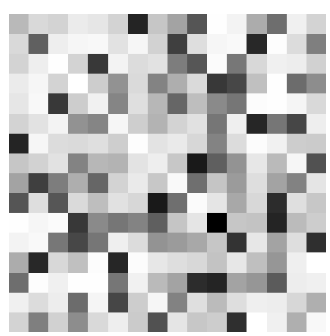


independent sampling

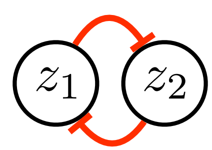


linearly-dependent dimensions: $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$

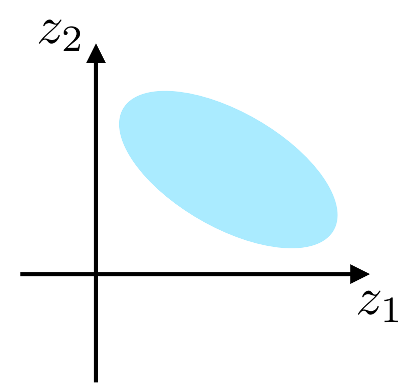
$\boldsymbol{\Sigma} =$



in 2D:

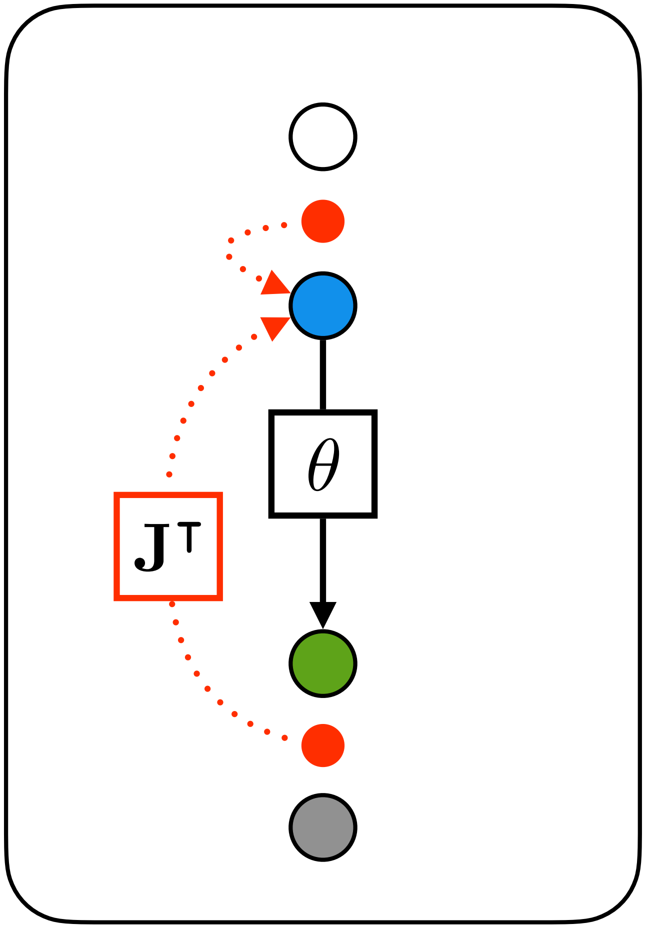


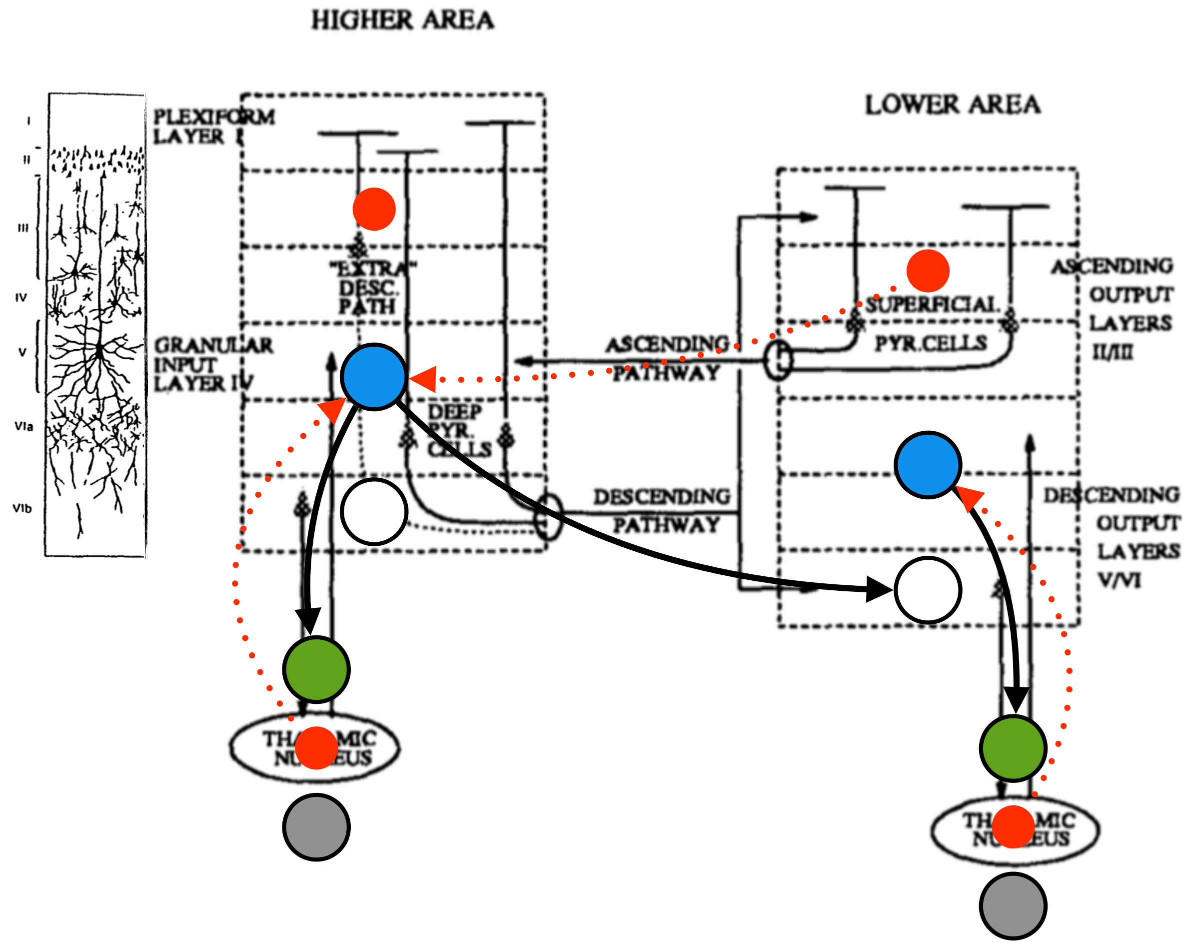
linearly-dependent sampling



PREDICTIVE CODING

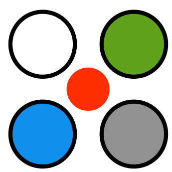
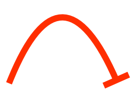
(proposed) biological correspondences





PREDICTIVE CODING

(proposed) biological correspondences

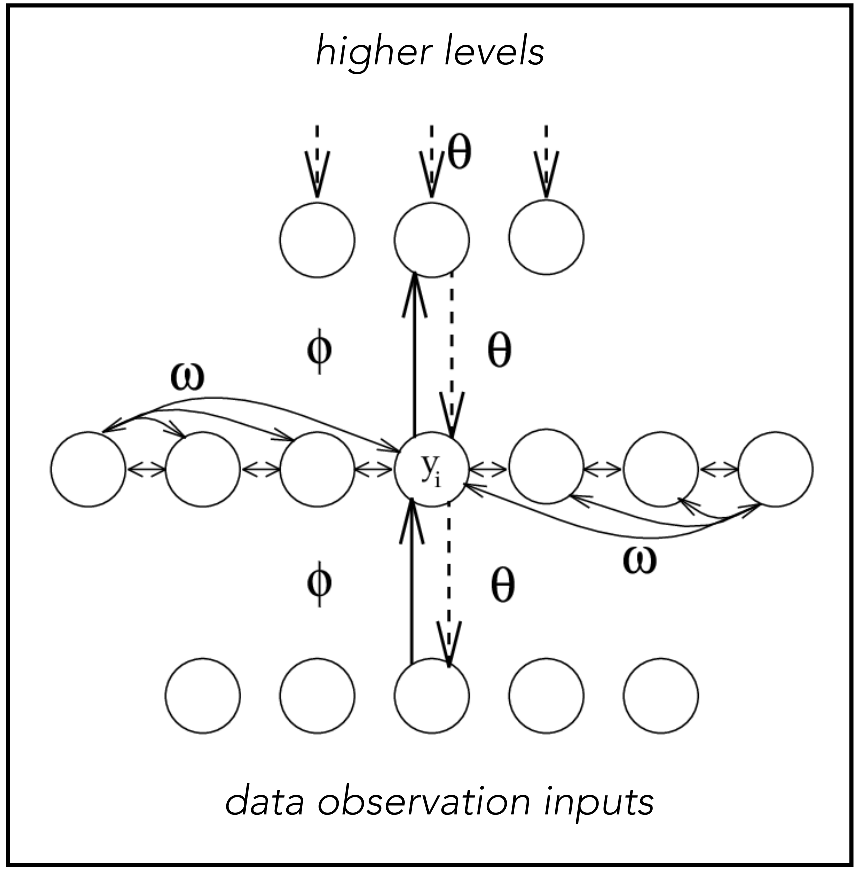
	MODEL		NEUROSCIENCE
	distributions & errors	→	pyramidal neuron firing rates
$\boxed{\theta}$	model weights	→	*descending* pyramidal neuron synapse weights
$\boxed{\mathbf{J}^{\top}}$	inference Jacobian	→	*ascending* pyramidal neuron synapse weights
	covariance matrix	→	lateral inhibition

III. VARIATIONAL AUTOENCODERS

DEEP LATENT VARIABLE MODELS

The Helmholtz Machine: *learn a separate inference model*

Inference



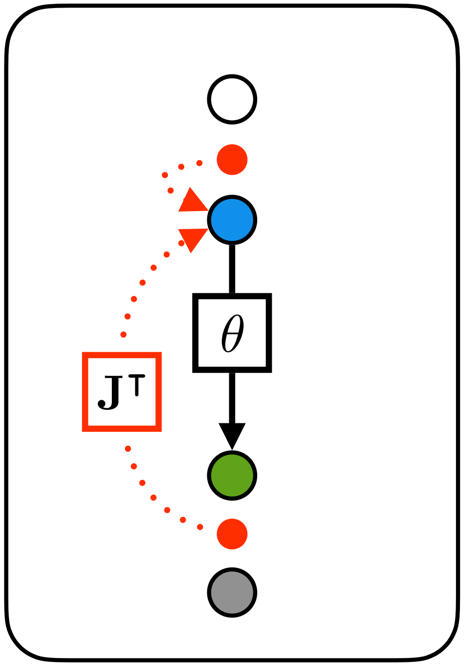
Prediction

Dayan et al., 1995
Dayan & Hinton, 1996

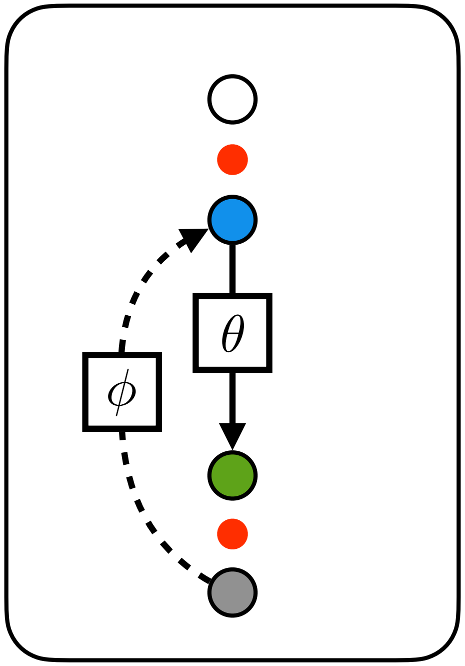
AMORTIZED INFERENCE

amortized inference:

spread out inference costs by learning a separate model (learning to infer)



gradient-based inference



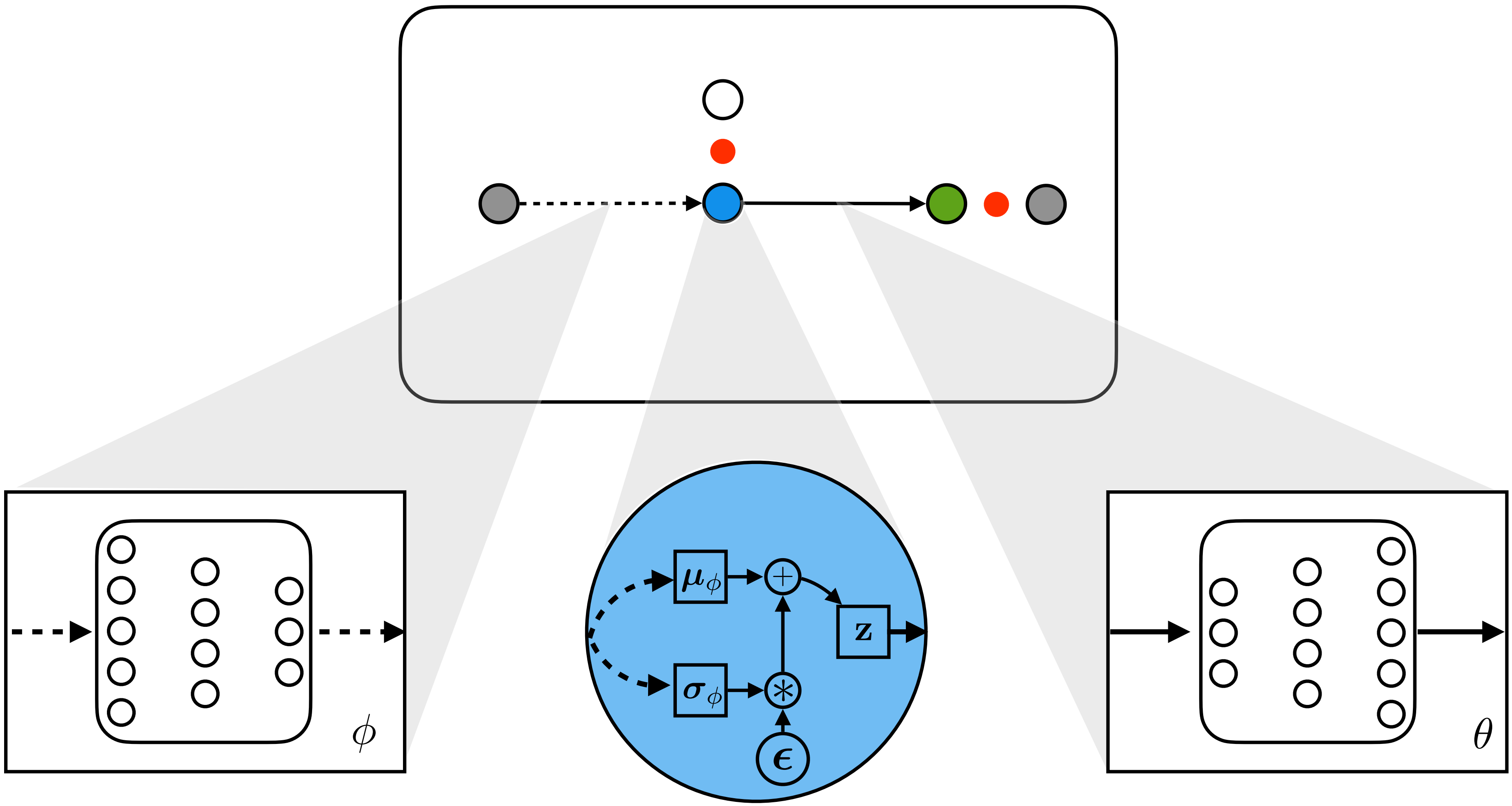
amortized inference

<u>substantially</u> more efficient!

VARIATIONAL AUTOENCODERS

Variational Autoencoder (VAE):

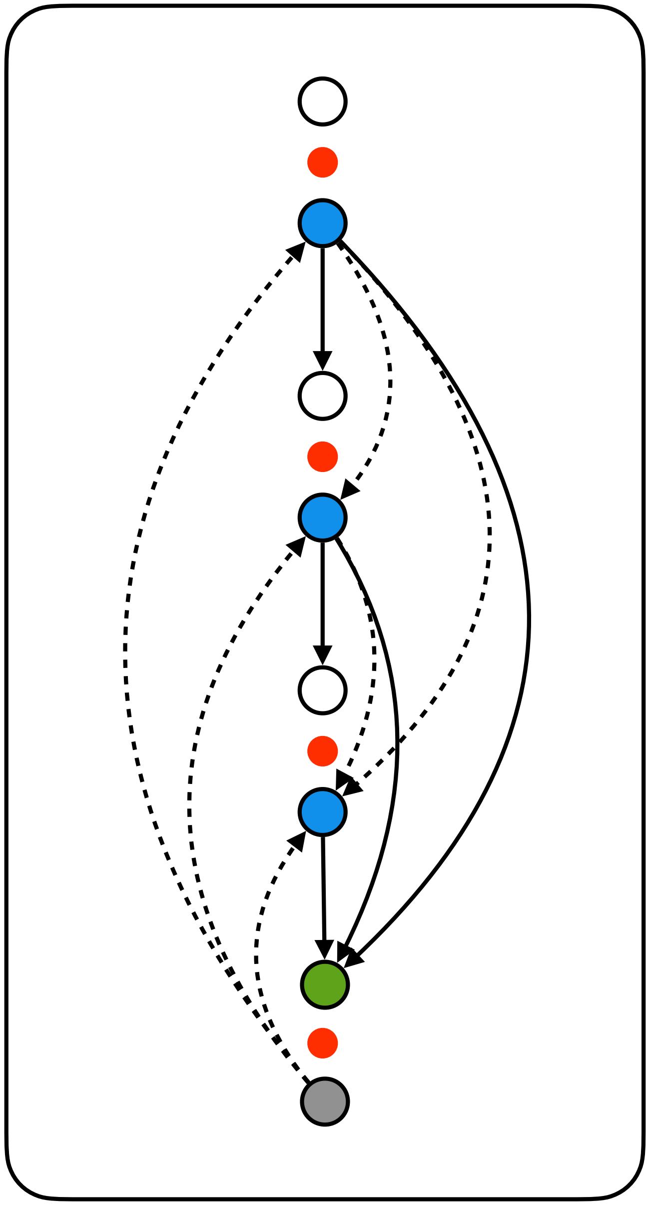
deep latent variable model + variational inference + direct encoder + *reparameterized* Gaussian



Kingma & Welling, 2014

Rezende et al., 2014

HIERARCHICAL VAES



256 x 256 samples

Child, 2020

up the hierarchy

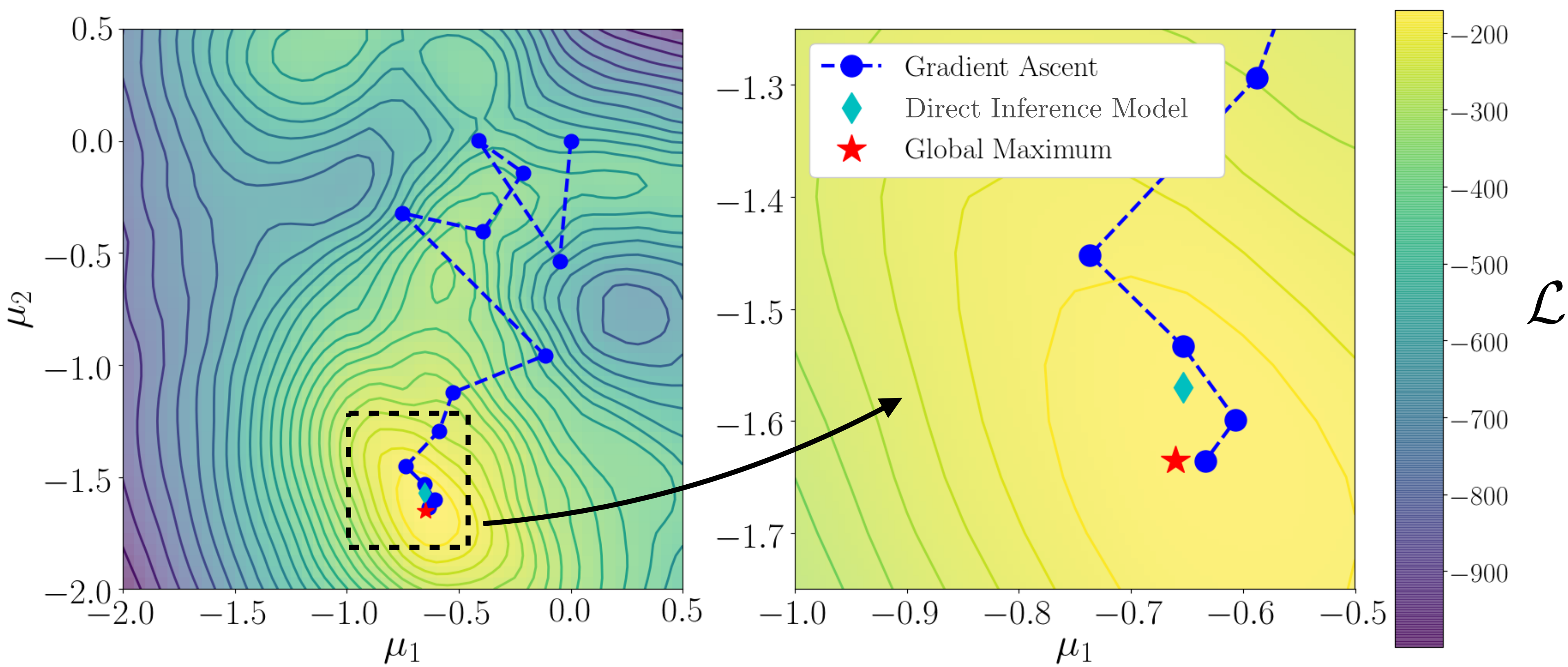
increasingly abstract

Sønderby et al., 2016

INFERENCE SUBOPTIMALITY

direct inference models provide suboptimal estimates

"amortization gap"



see also Cremer et al., 2018

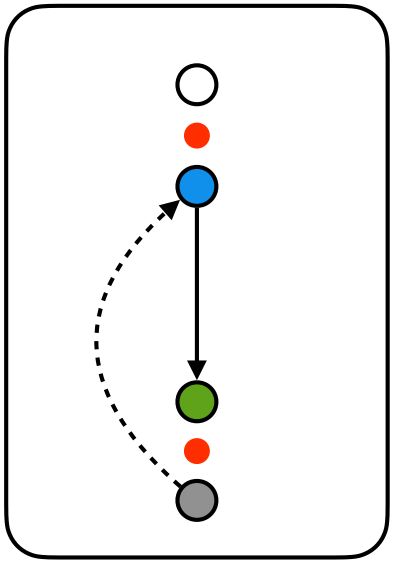
Marino et al., 2018

ITERATIVE AMORTIZED INFERENCE

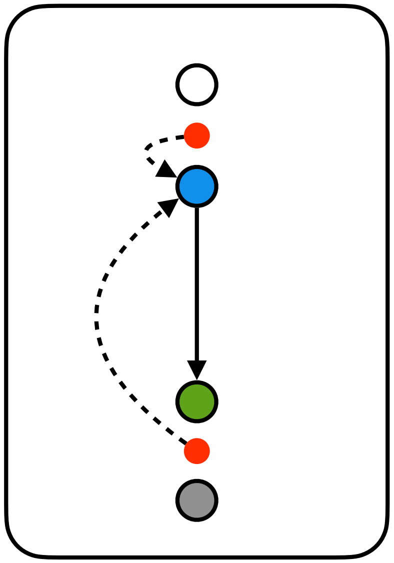
perform inference via a learned *iterative* procedure

encode inference gradient or errors

$\nabla_{\hat{\mathbf{z}}}\mathcal{L}$ $\quad$ $\boldsymbol{\xi}_{\mathbf{z}}$ $\boldsymbol{\xi}_{\mathbf{x}}$



Direct
Amortization



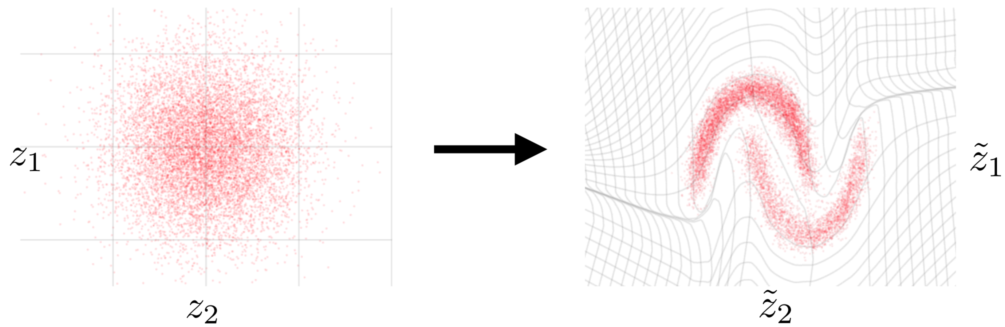
Iterative
Amortization

+ more accurate
+ more general

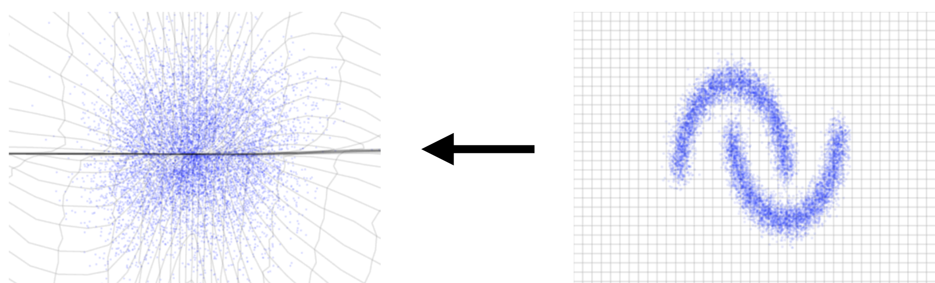
NORMALIZING FLOWS

normalizing flows add/remove dependencies

un-normalize



independent $\qquad z_1 \quad z_2 \quad \underset{f_\theta^{-1}}{\overset{f_\theta}{\rightleftarrows}} \quad \tilde{z}_1 \quad \tilde{z}_2 \qquad$ dependent



normalize

Dinh *et al.*, 2015; 2017

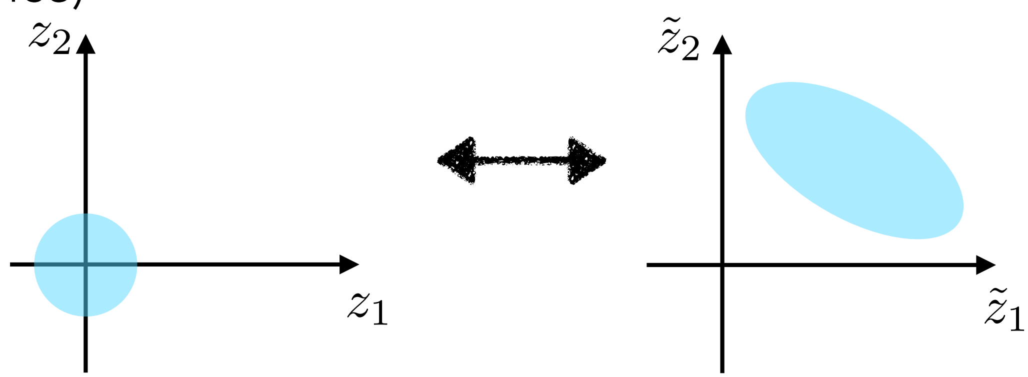
Rezende & Mohamed, 2015

NORMALIZING FLOWS

a covariance matrix is an affine normalizing flow with *linear* dependencies

LINEAR (covariance)

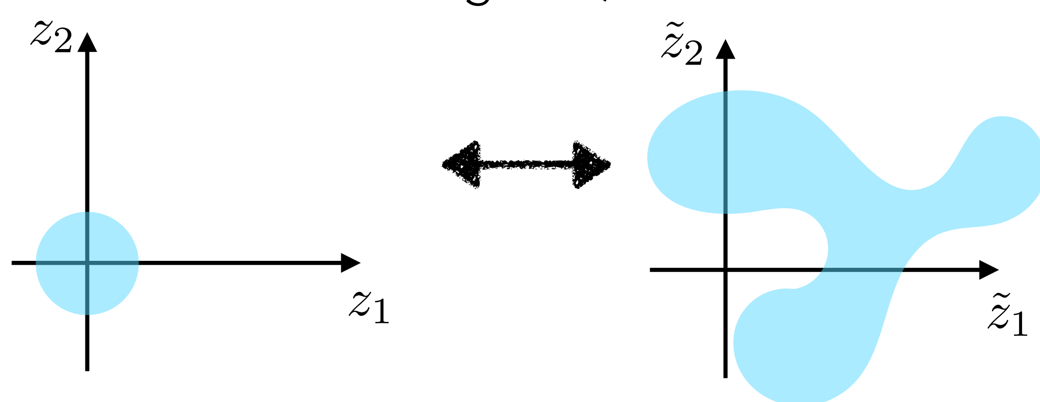
$$\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$$



$$\mathcal{N}(\tilde{\mathbf{z}}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

NON-LINEAR (non-linear normalizing flow)

$$\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$$



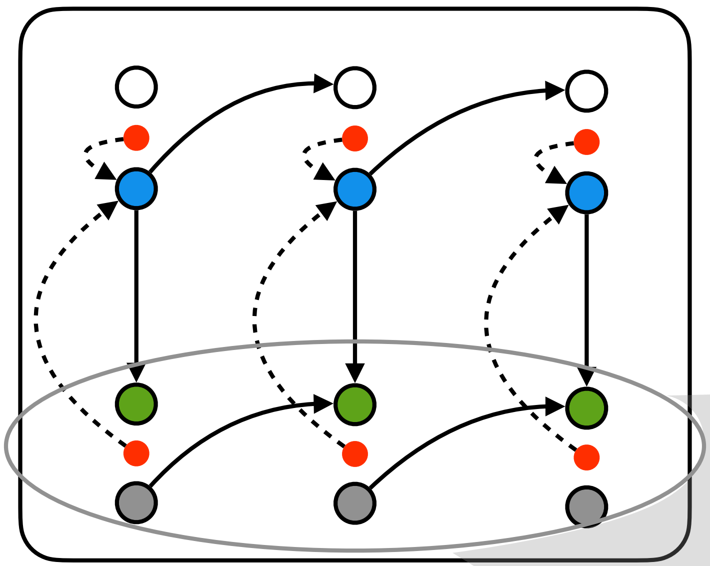
$$p(\tilde{\mathbf{z}})$$

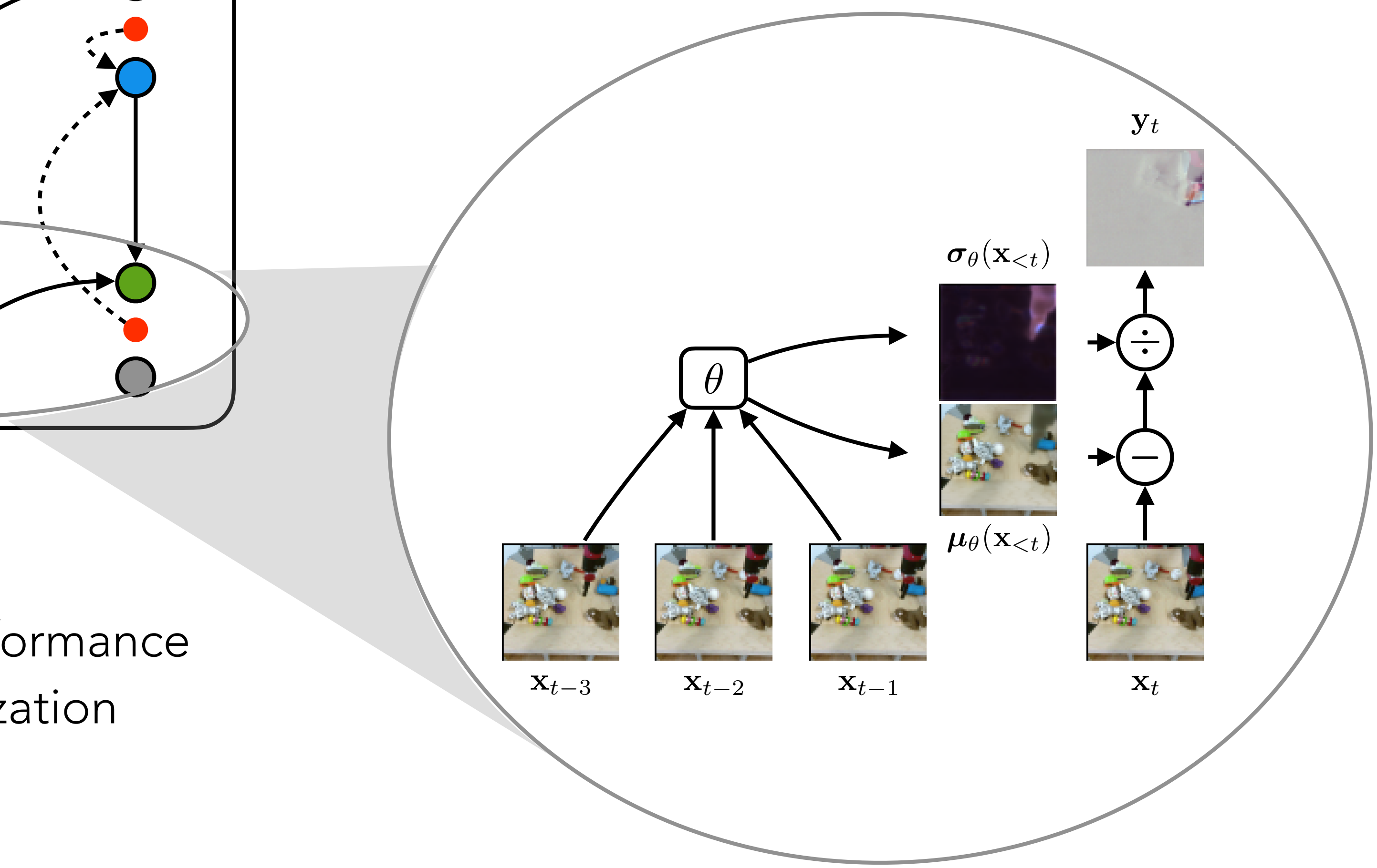
NORMALIZING FLOWS

normalizing flows can be applied to any parametric distribution

sequential autoregressive flows

remove dependencies across time

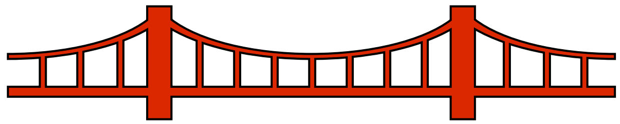


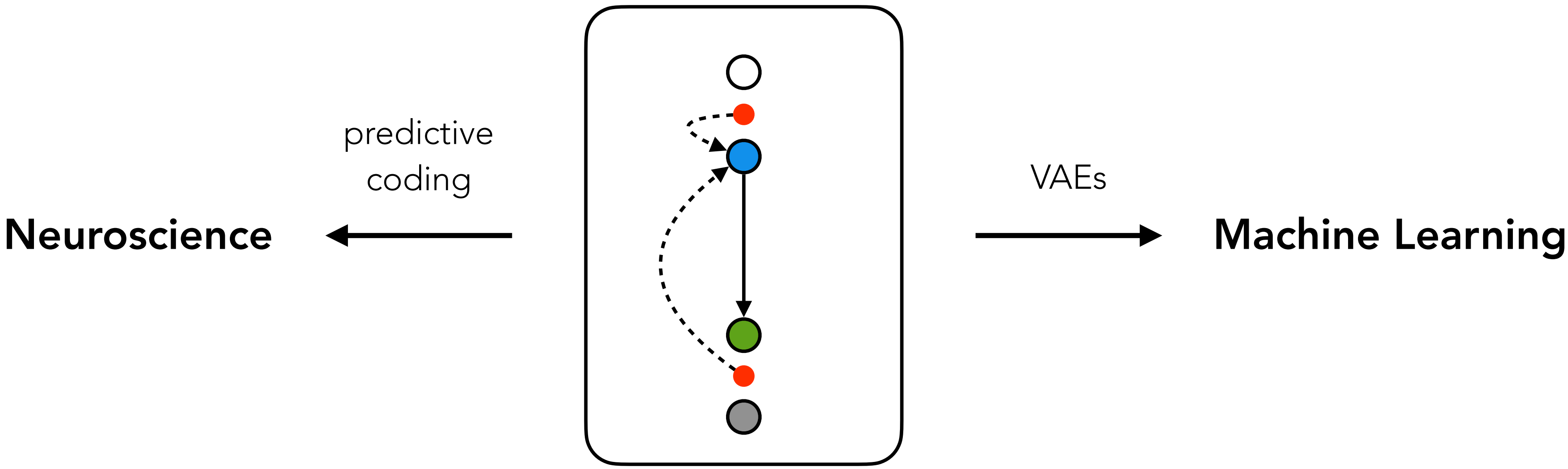


improved performance
& generalization

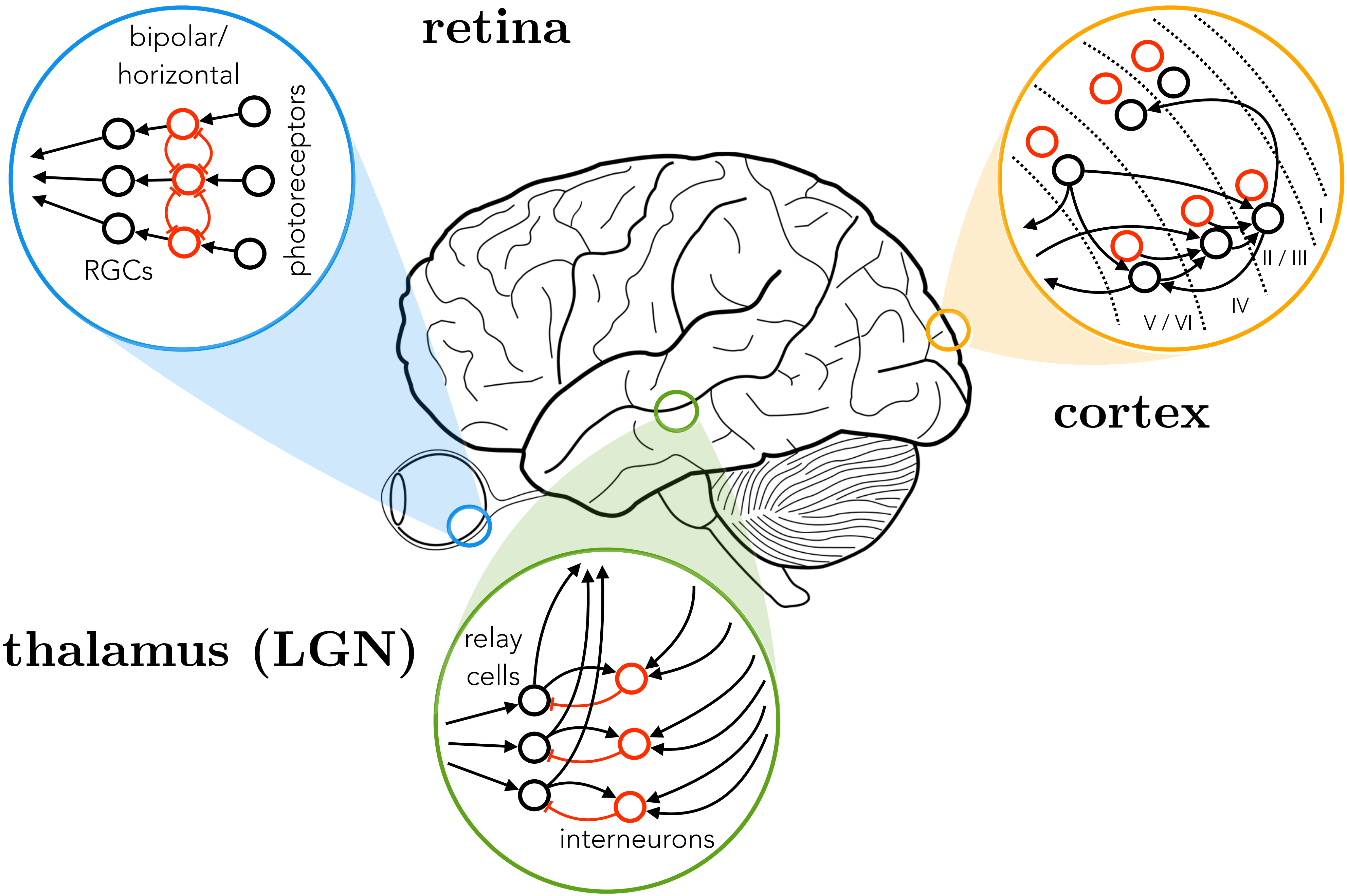
IV. CORRESPONDENCES

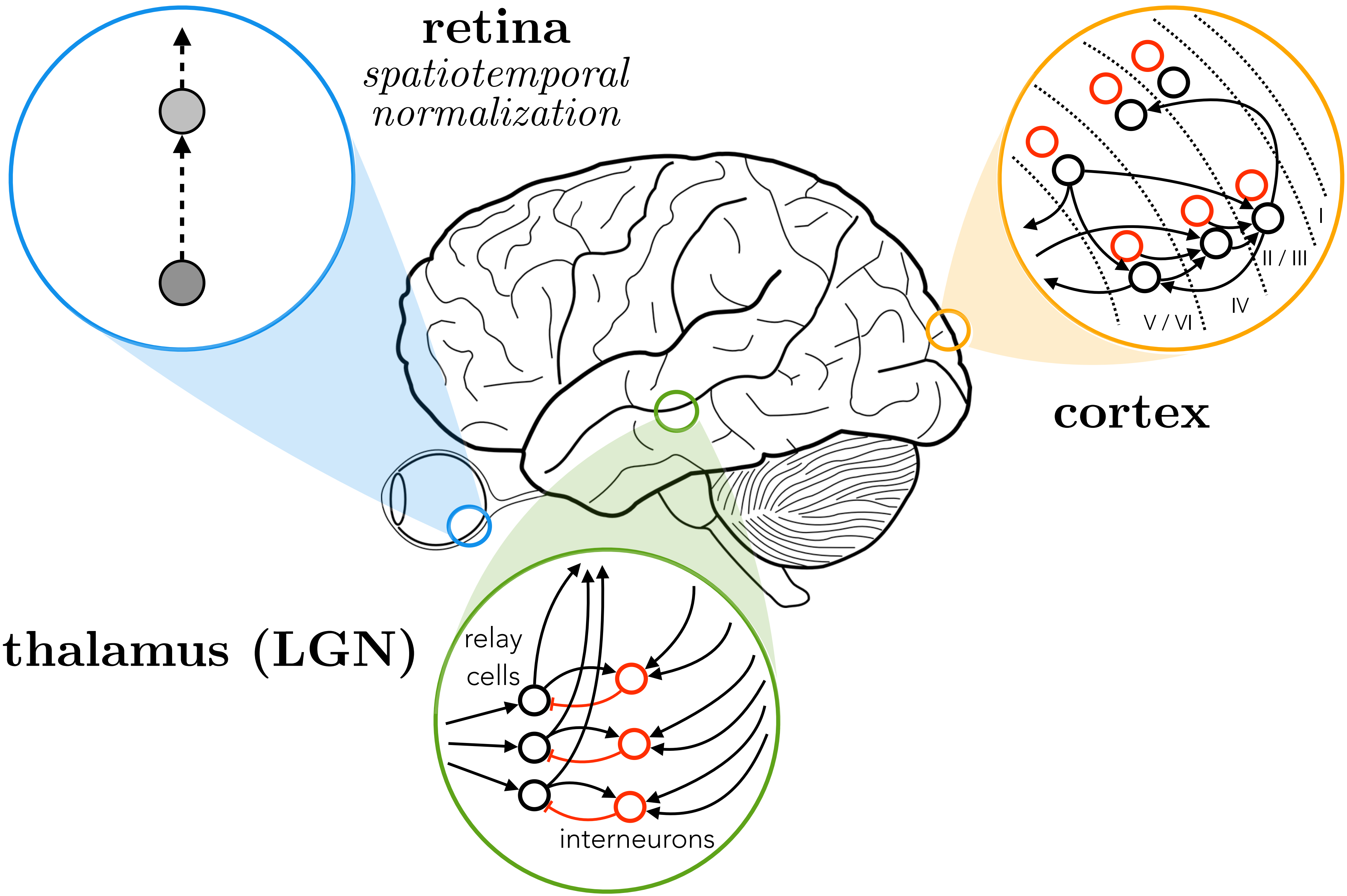
a *bridge* between neuroscience and machine learning

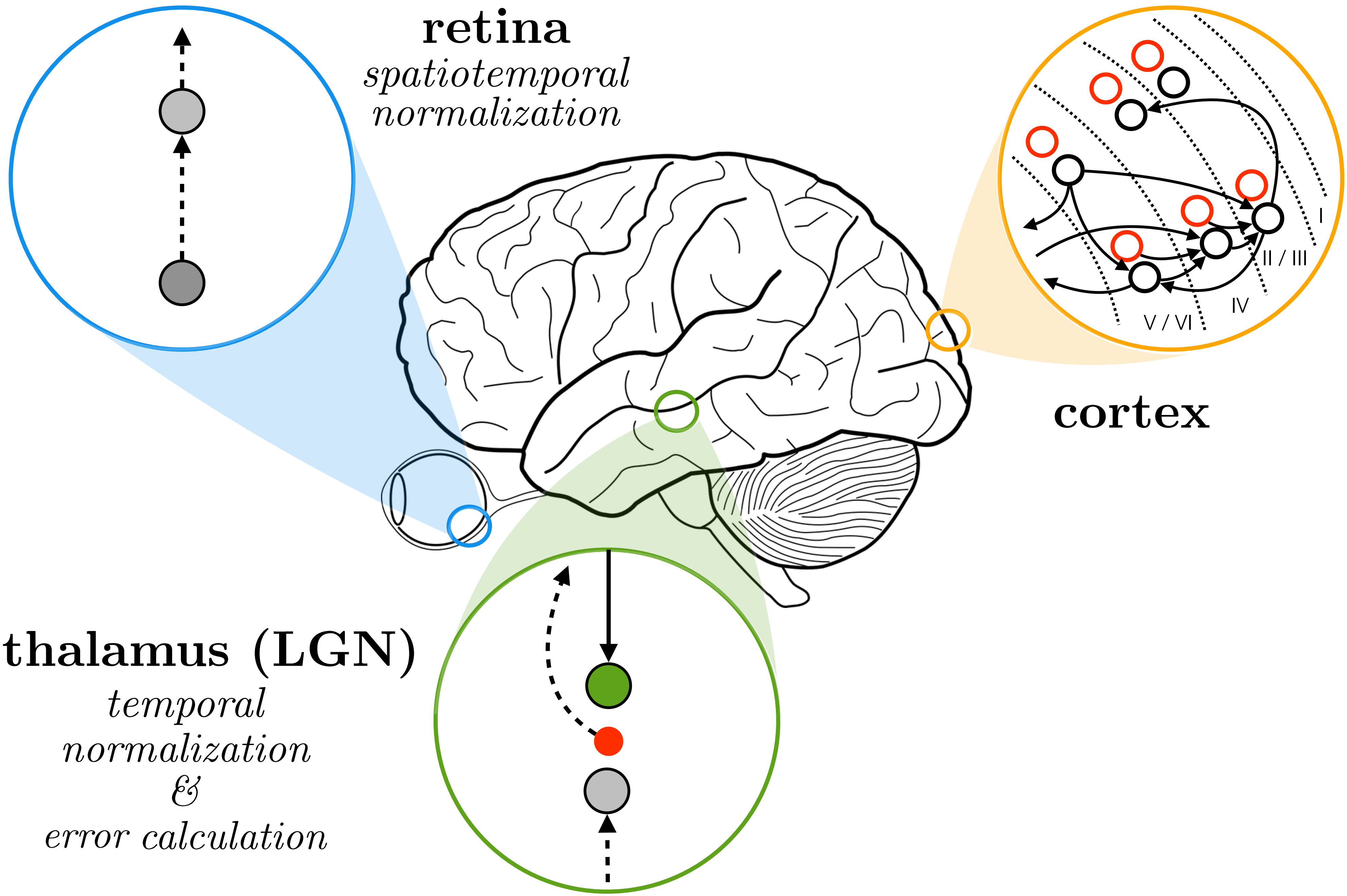




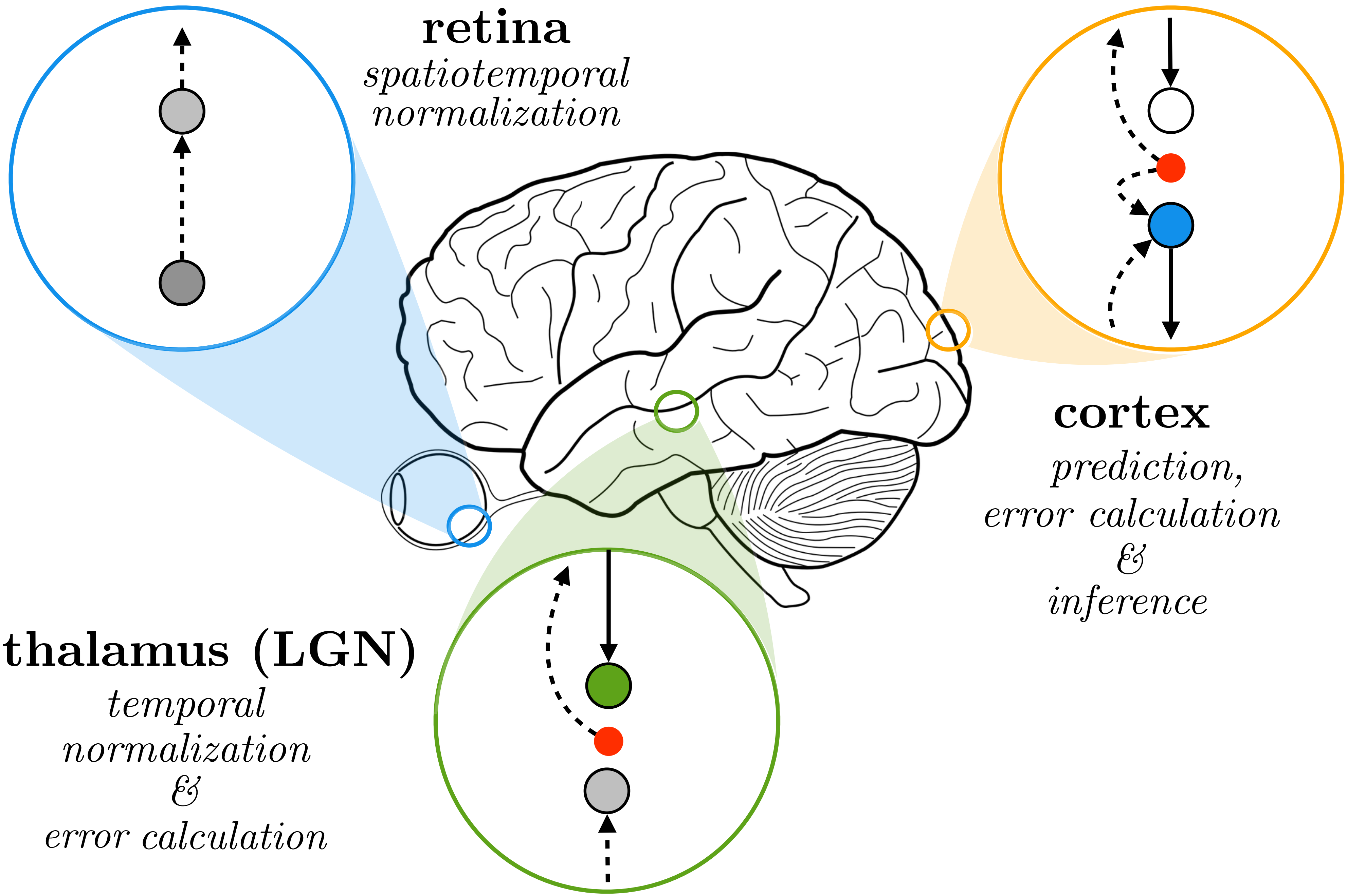
providing new correspondences…

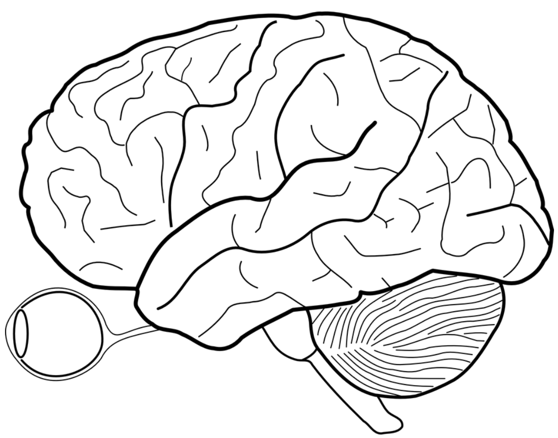

retina
bipolar/
horizontal
photoreceptors
RGCs
cortex
I
II / III
IV
V / VI
thalamus (LGN)
relay
cells
interneurons

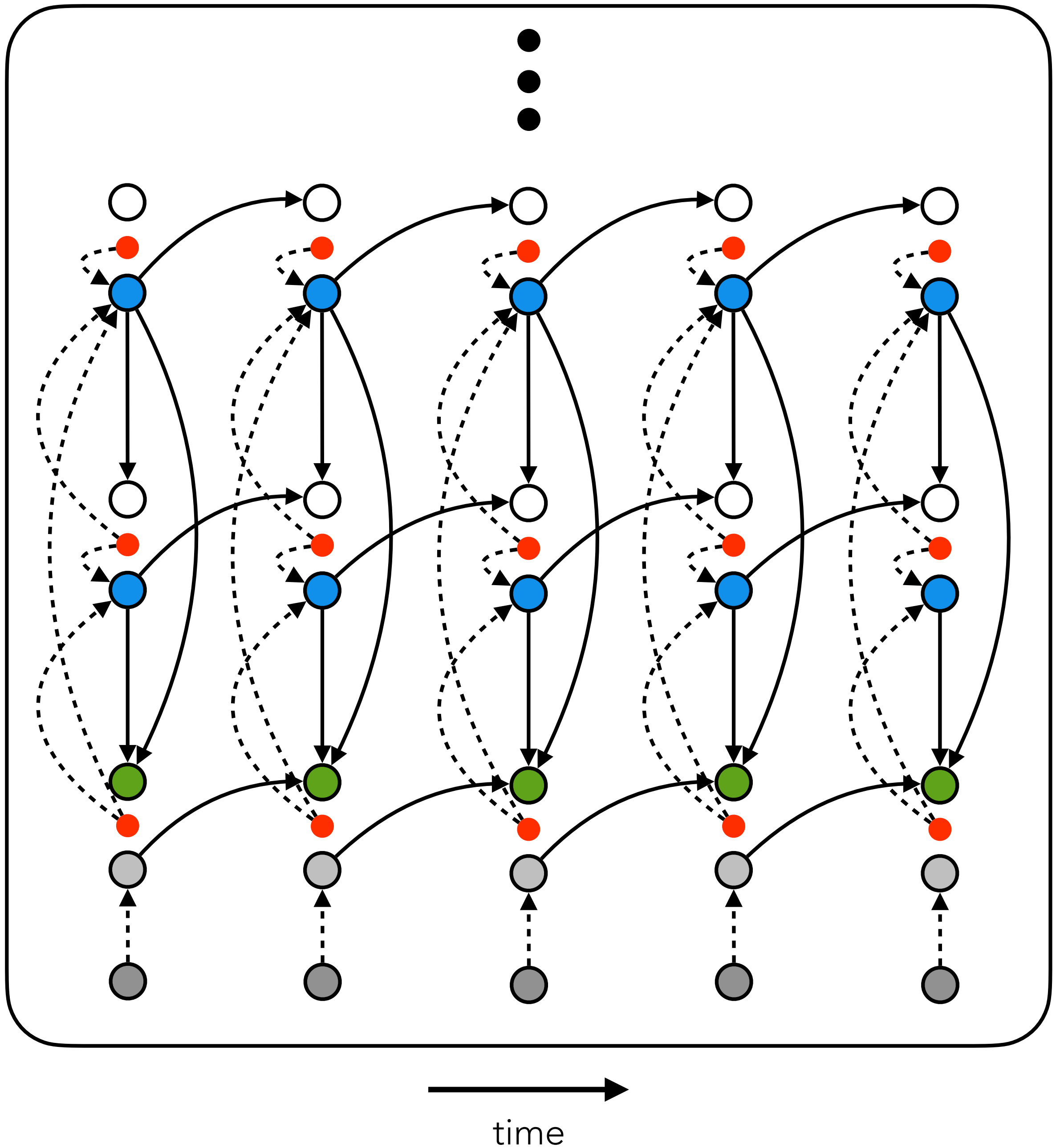

retina
spatiotemporal
normalization
cortex
I
II / III
IV
V / VI
thalamus (LGN)
relay
cells
interneurons

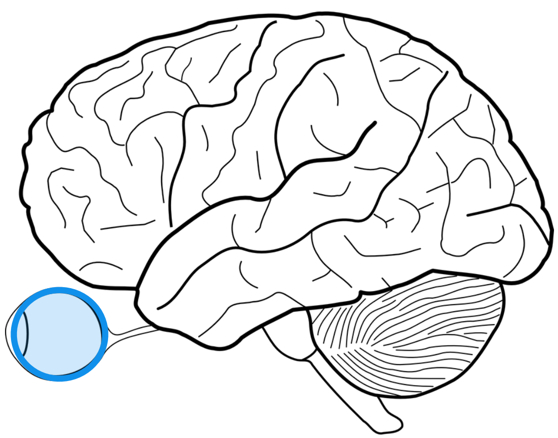


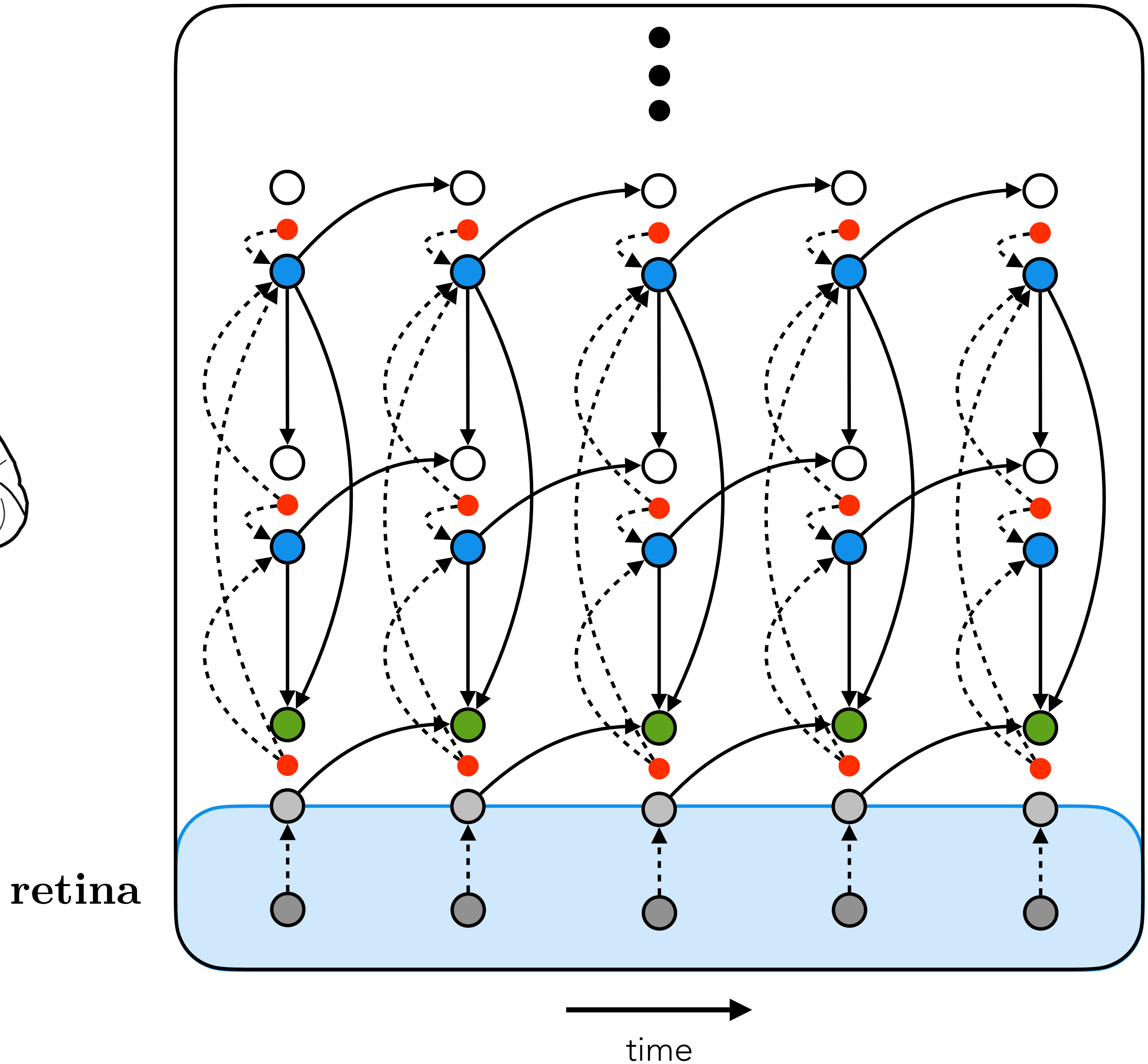
retina
spatiotemporal
normalization
cortex
I
II / III
IV
V / VI
thalamus (LGN)
temporal
normalization
&
error calculation

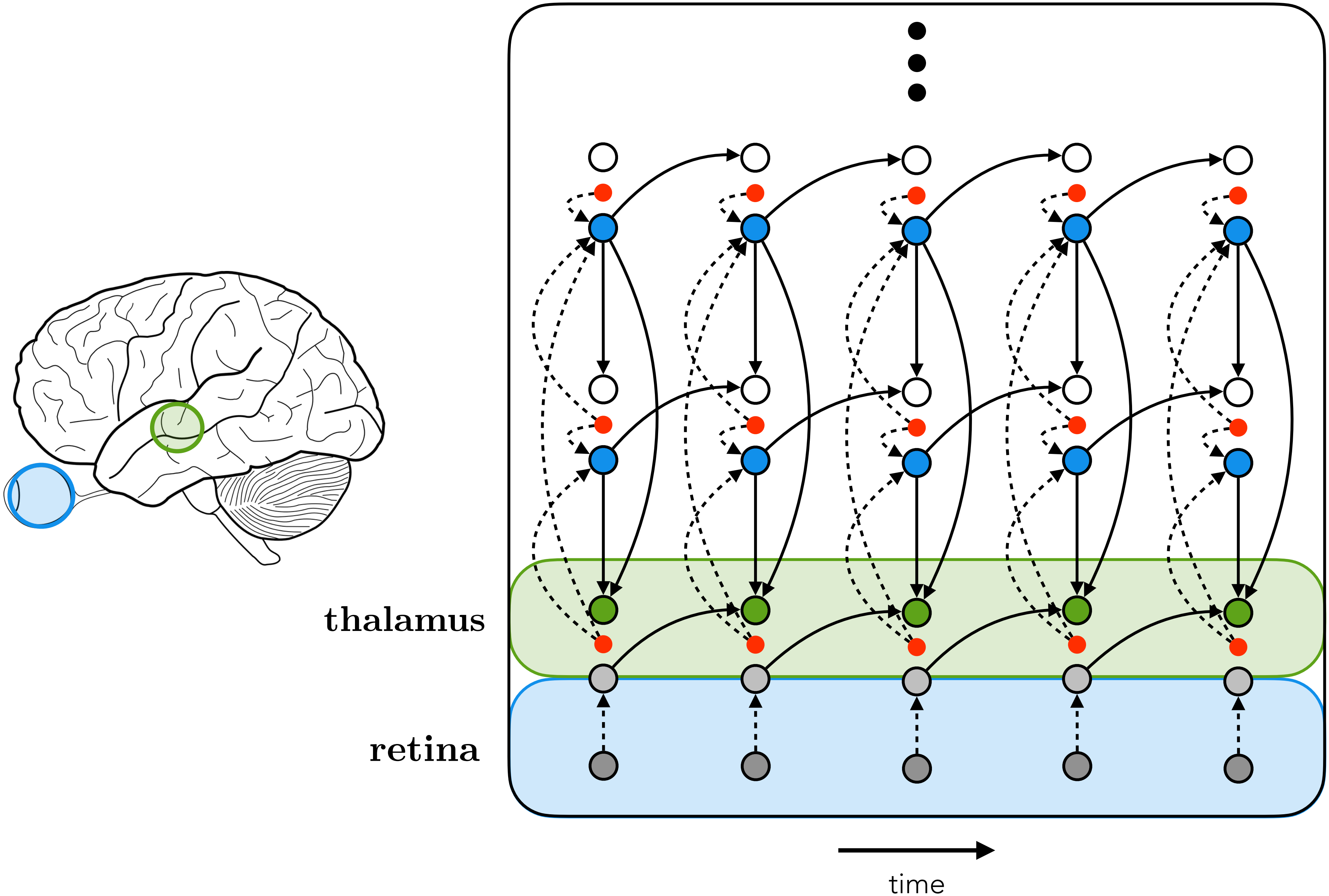

retina
spatiotemporal normalization
thalamus (LGN)
temporal normalization & error calculation
cortex
prediction, error calculation & inference

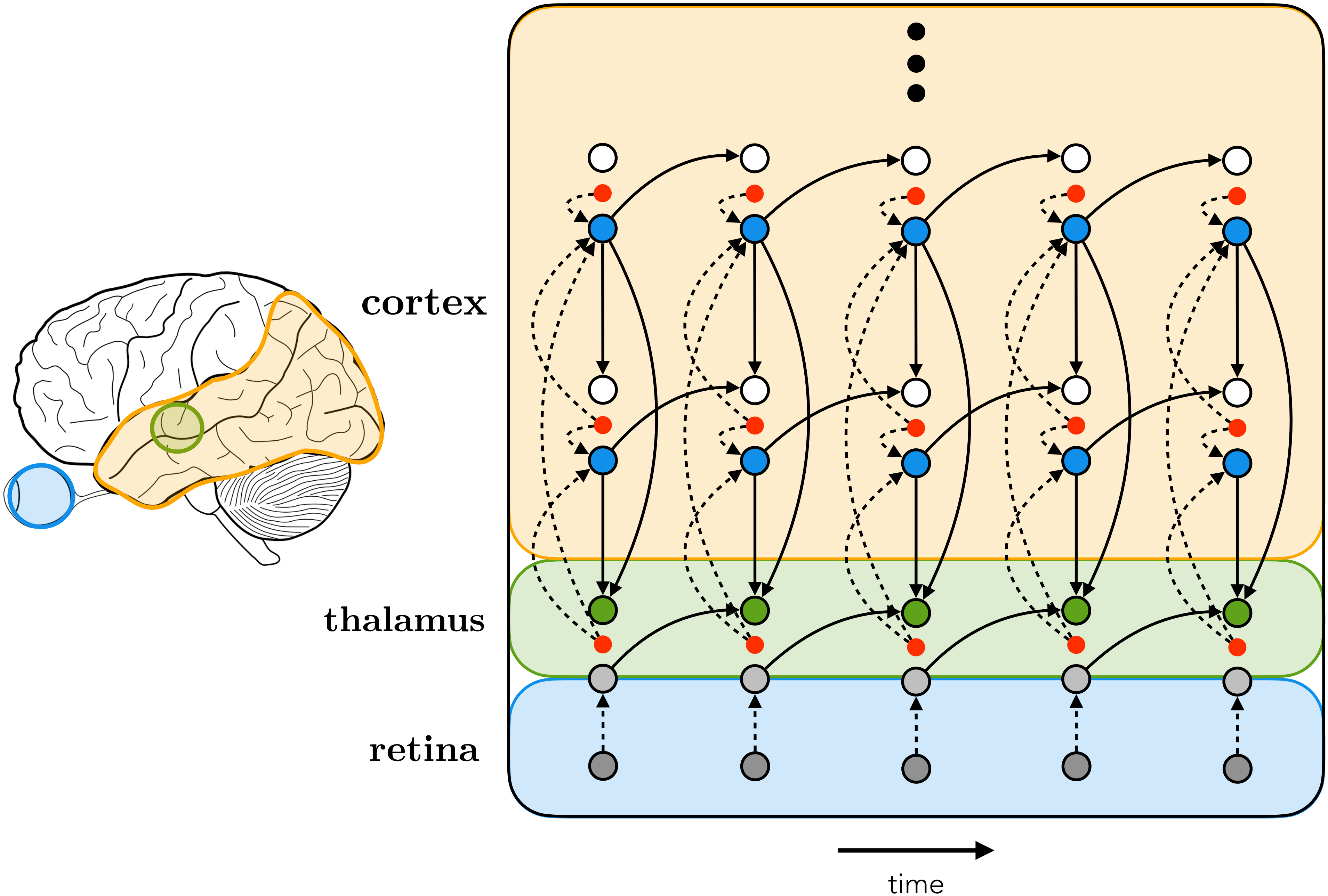



time



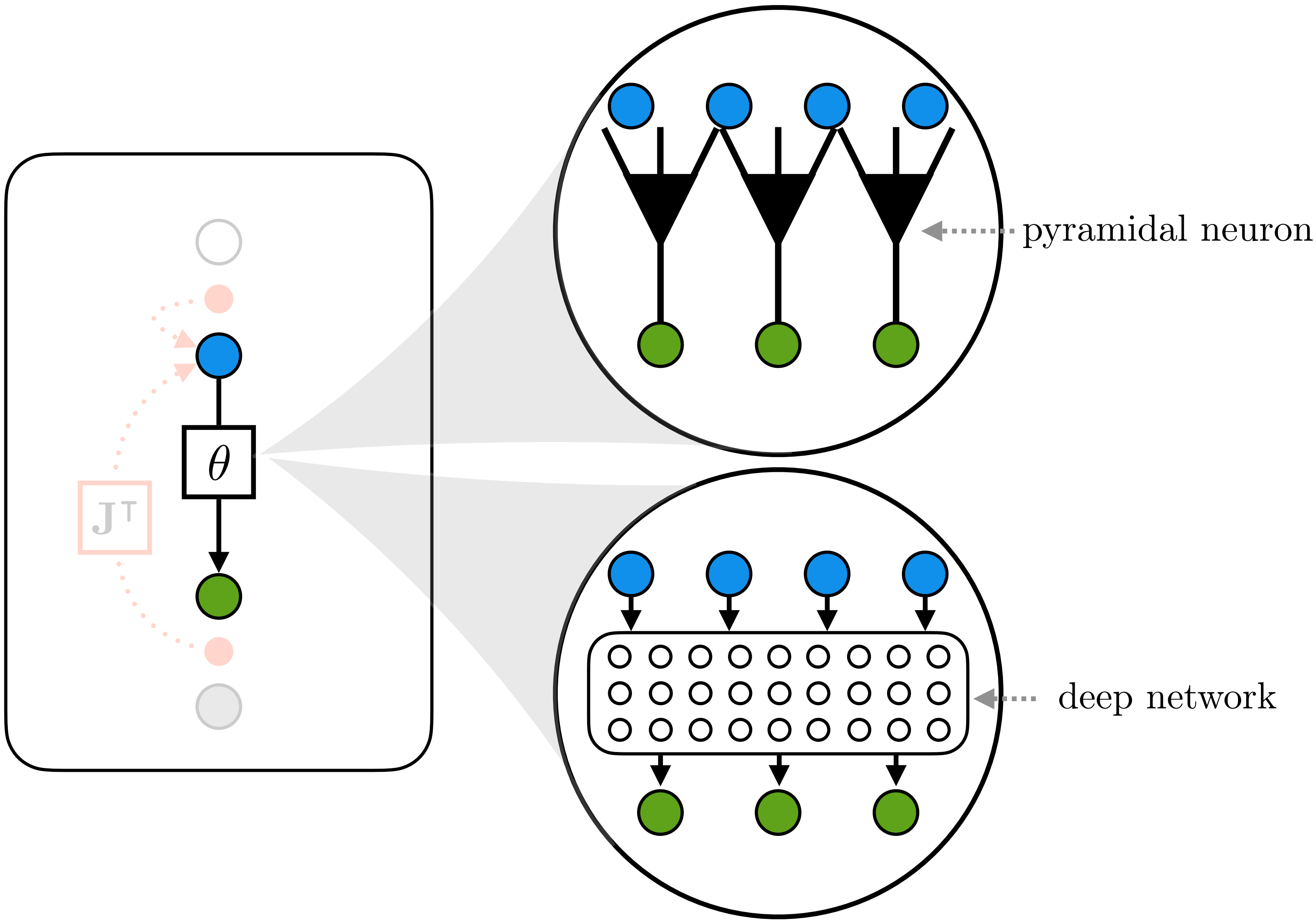

retina
time


thalamus
retina
time

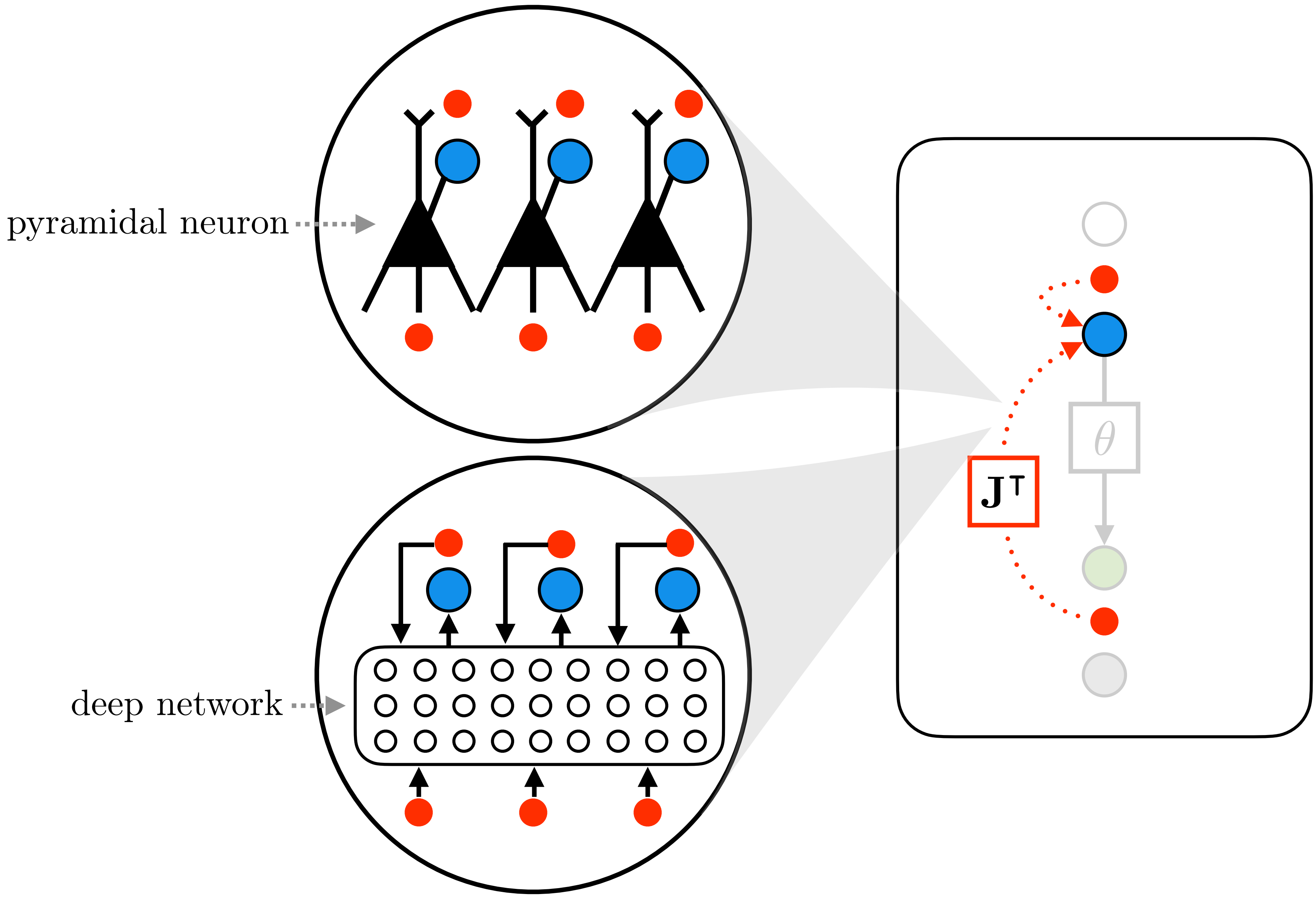


cortex
thalamus
retina
time

PYRAMIDAL NEURONS AS DEEP NETWORKS



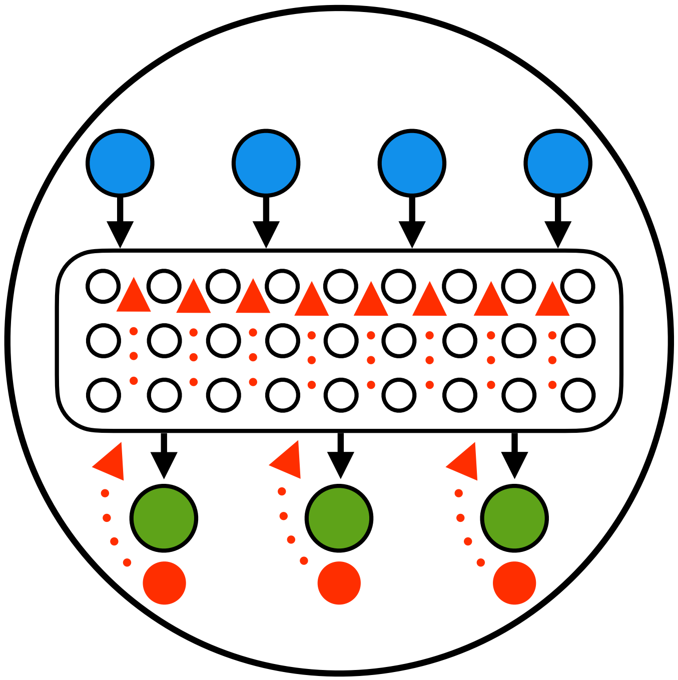
PYRAMIDAL NEURONS AS DEEP NETWORKS



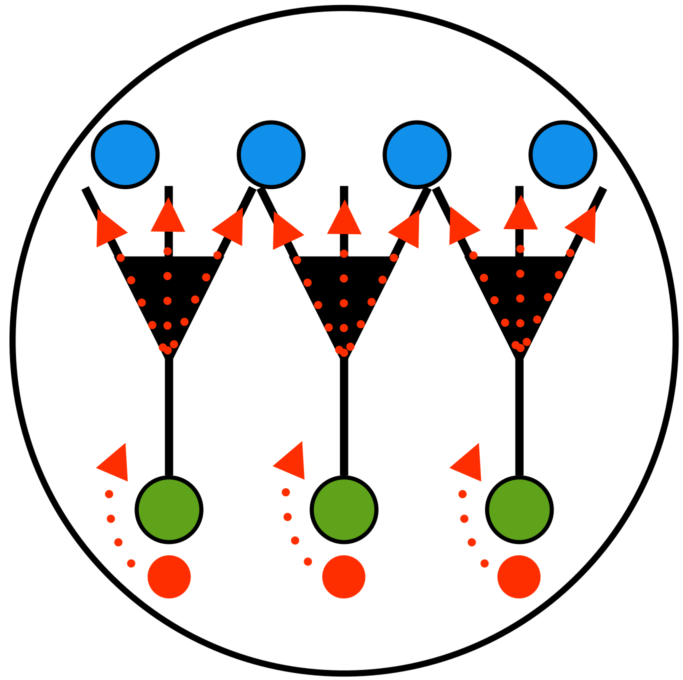
BACKPROPAGATION WITHIN NEURONS

in hierarchical latent variable models, errors provide a ***local*** training signal

see, e.g., target propagation (Bengio, 2014)

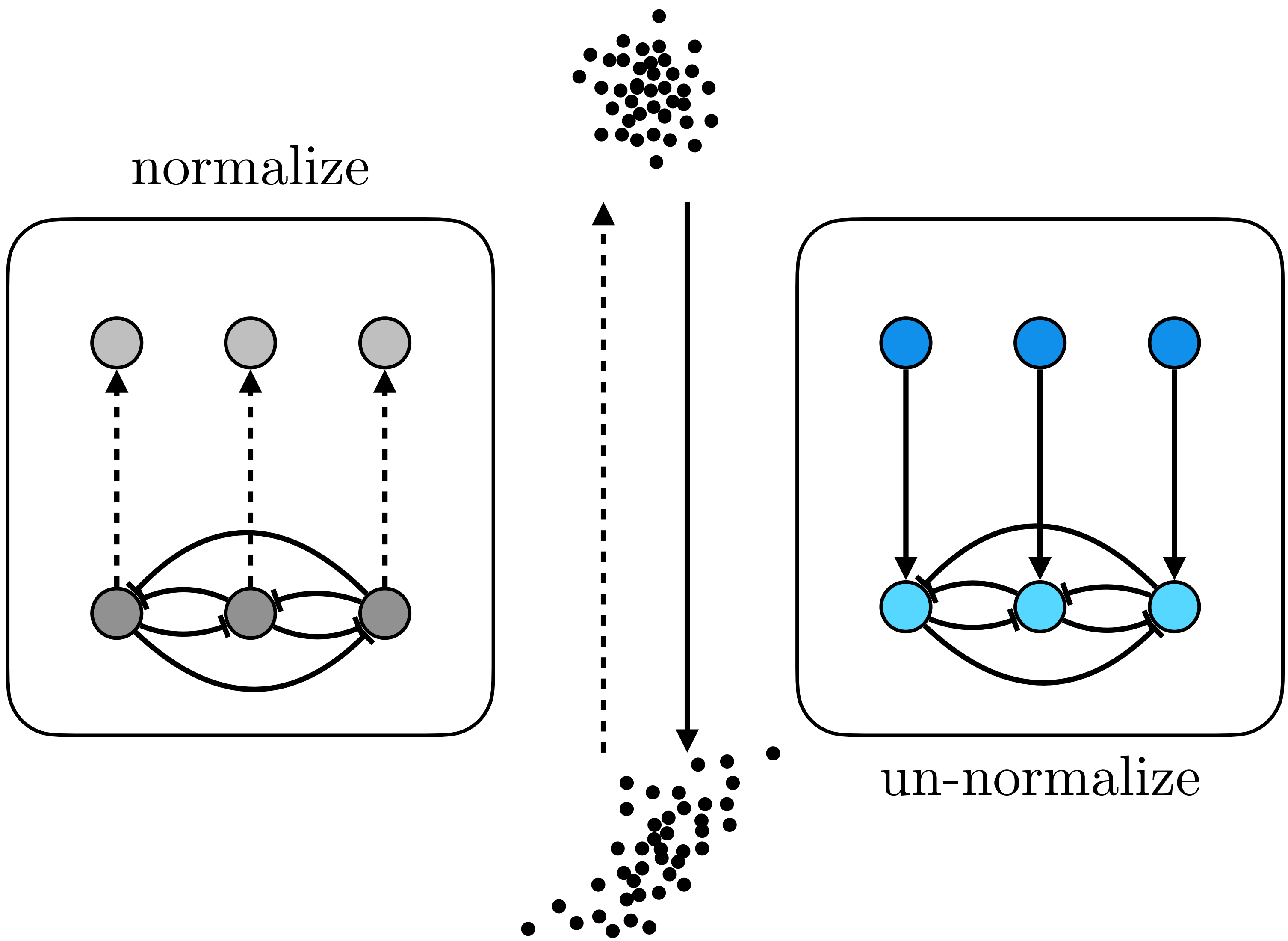


only need to
backpropagate gradients
between variables

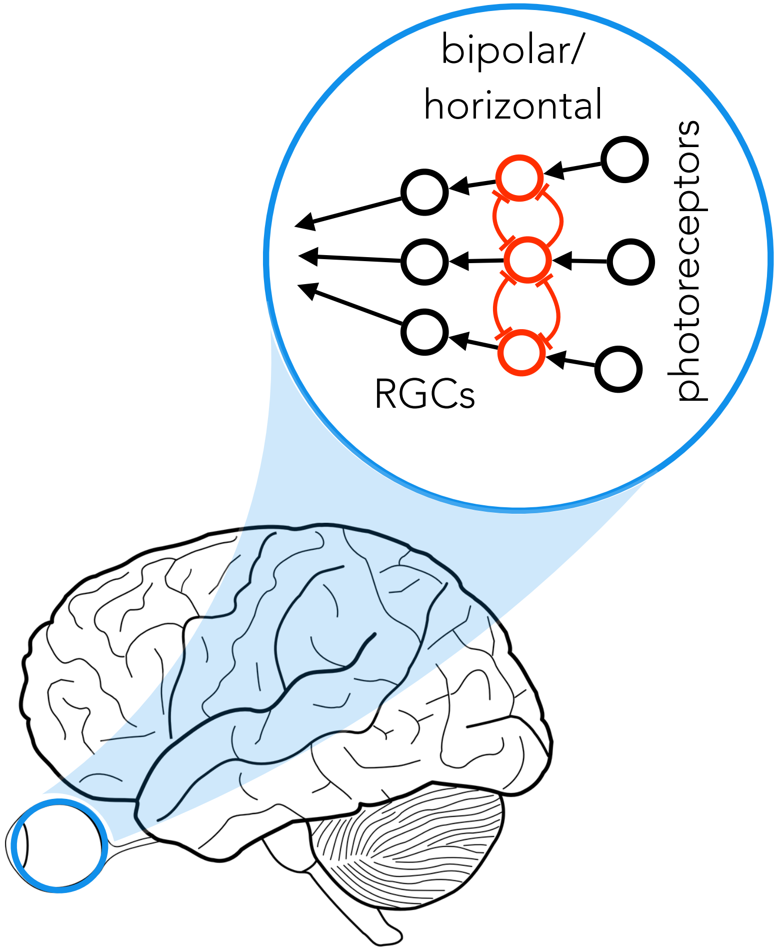


corresponding process
would occur *within neurons*

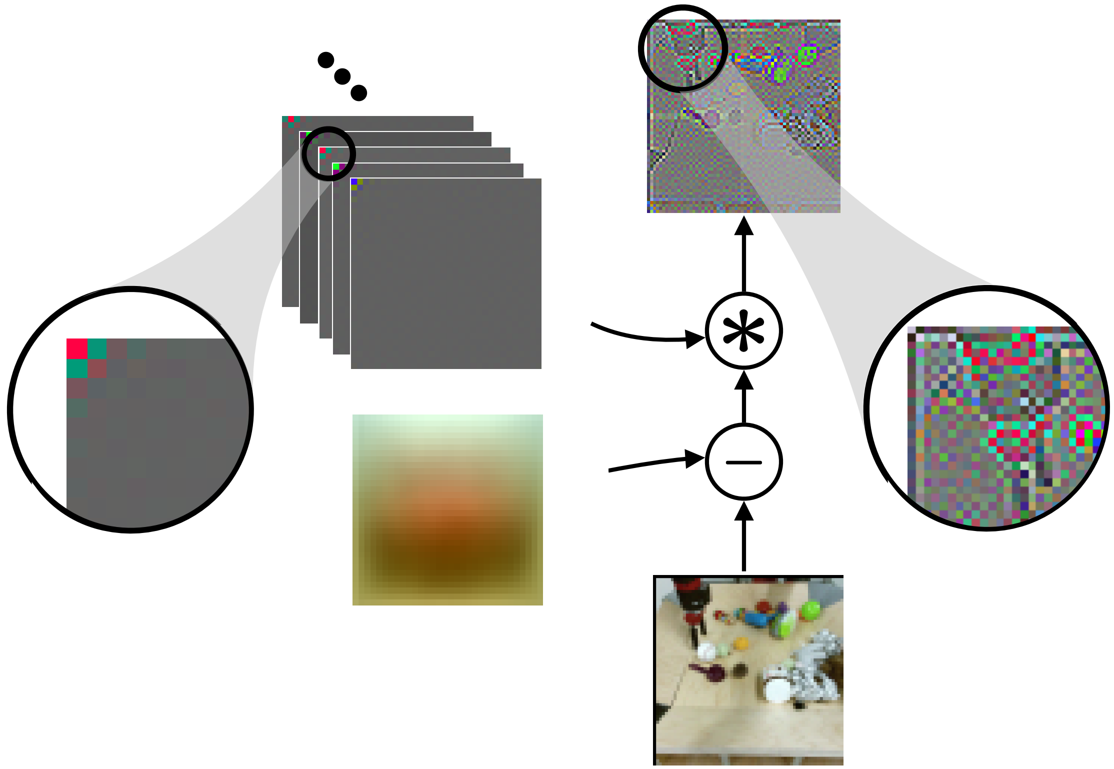
interneurons can add or remove spatiotemporal dependencies between neurons



LATERAL INHIBITION & NORMALIZING FLOWS

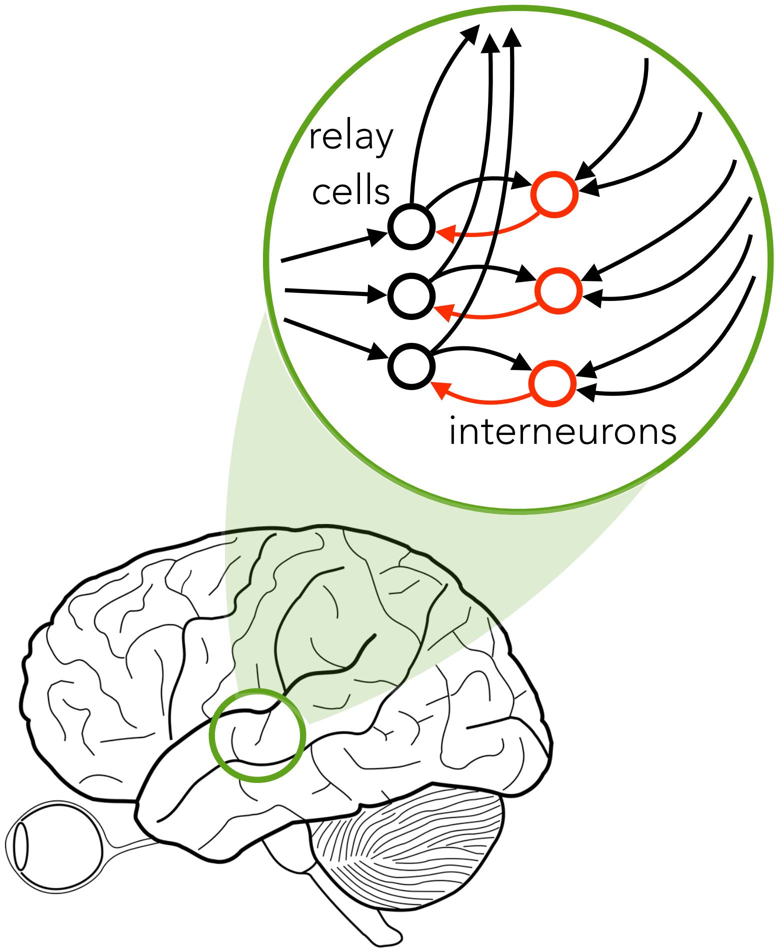


retinal interneurons
remove spatial dependencies

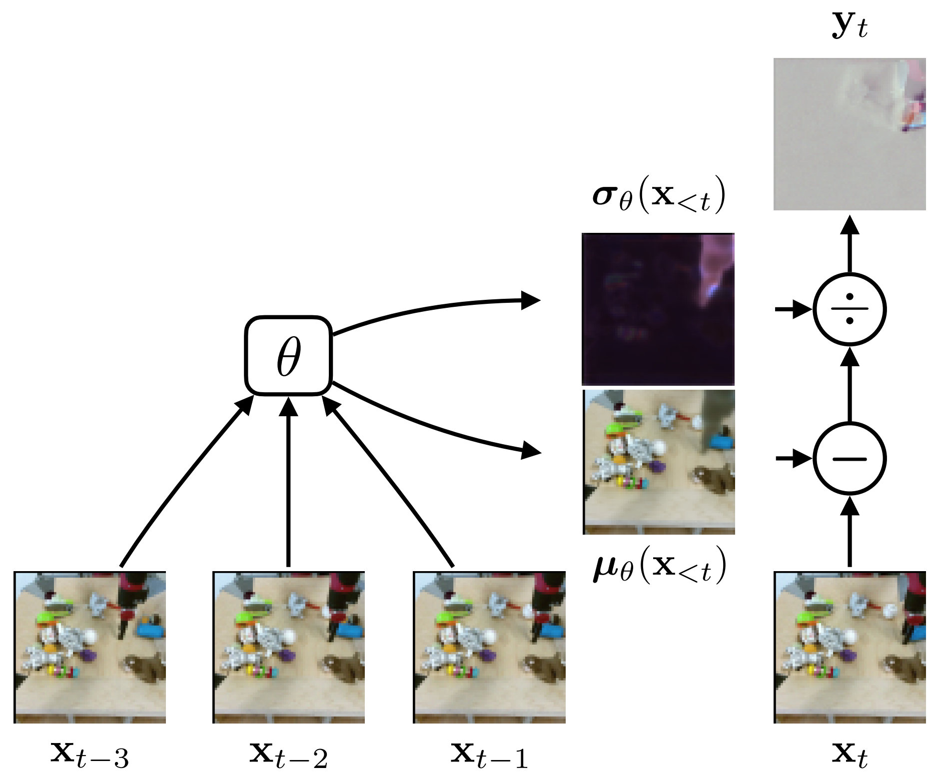


ZCA whitening:
inverting a linear affine normalizing flow

LATERAL INHIBITION & NORMALIZING FLOWS

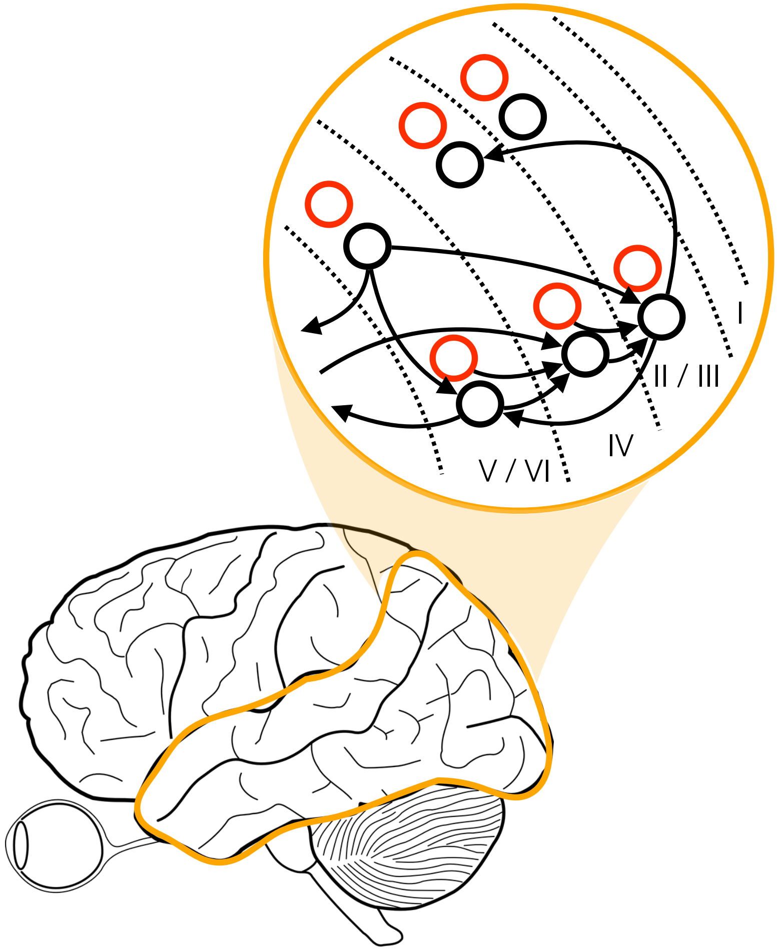


thalamic interneurons
remove low-level temporal dependencies

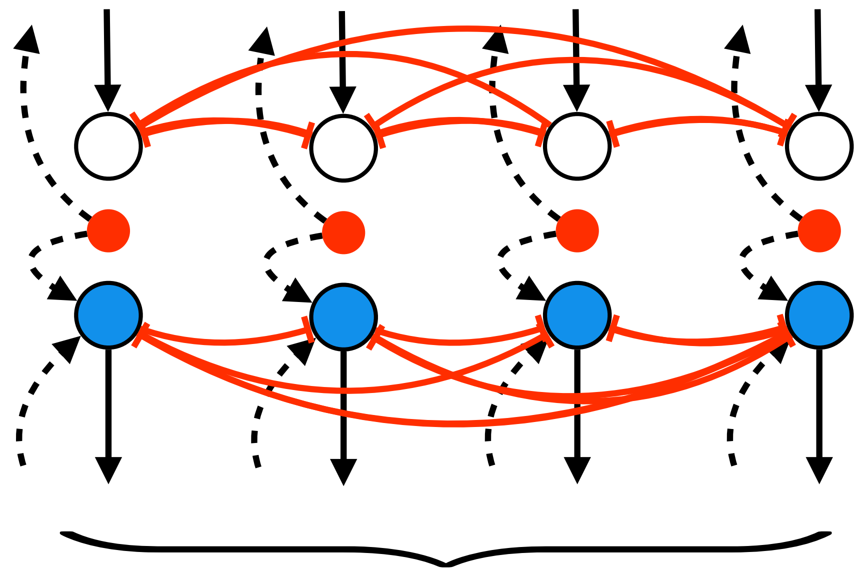


temporal whitening:
inverting an affine normalizing flow

LATERAL INHIBITION & NORMALIZING FLOWS



cortical interneurons
add/remove spatiotemporal dependencies
between columns



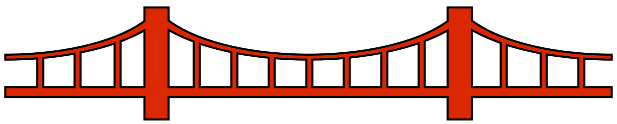
cortical columns

forward/inverse normalizing flows

IV. correspondences

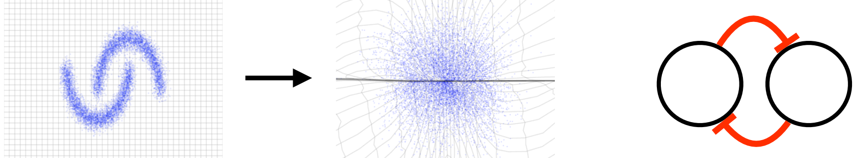
SUMMARY OF CORRESPONDENCES

by traversing the bridge between predictive coding & VAES…

MACHINE LEARNING  **NEUROSCIENCE**

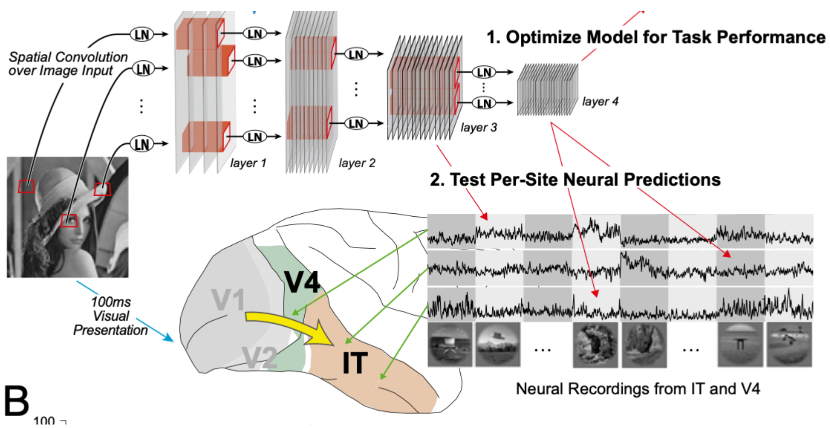
deep networks & pyramidal neurons

normalizing flows & lateral inhibition



A NEW COMPUTATIONAL MODEL FOR NEUROSCIENCE

current efforts to compare deep networks and the brain are limited

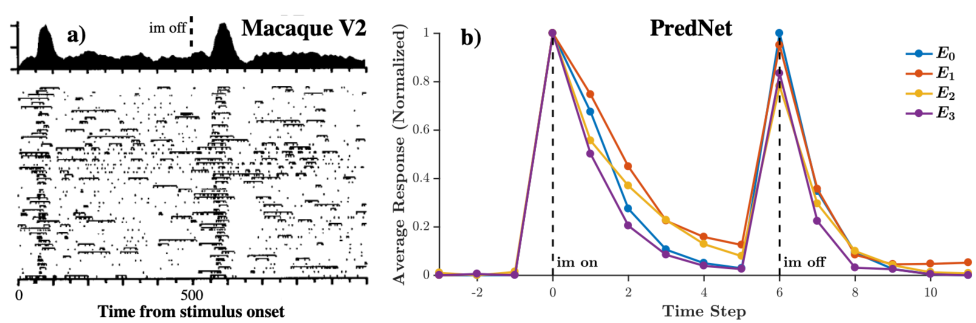


compare with supervised learning on brief image presentation

ignore feedback + dynamics

Yamins et al., 2014

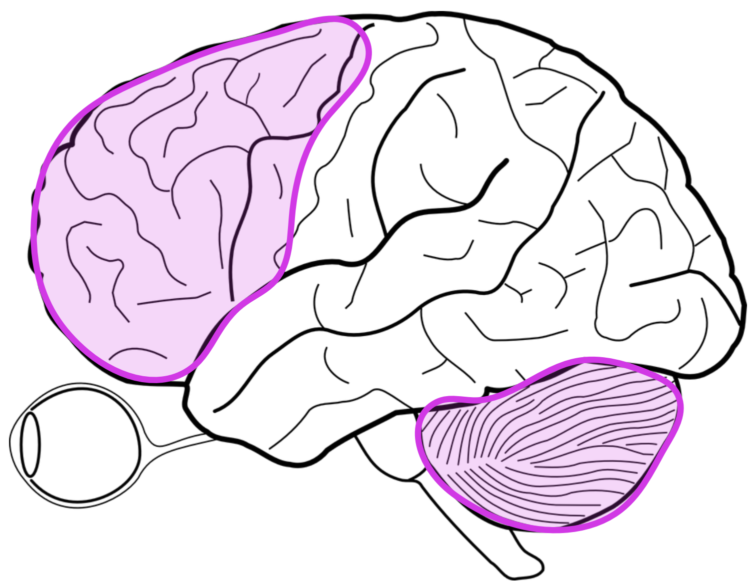
large-scale VAEs provide a route toward testing predictive coding



Lotter et al., 2018

Child, 2020

MOTOR CONTROL / EXECUTIVE FUNCTION



can apply the *same computational techniques* to control/RL

→ prediction errors drive ***feedback control***

Iterative Amortized Policy Optimization, Marino et al., 2020

`arXiv:2010.10670`

Thank You