# The Challenge of Constructing a Robust Short-Term Memory Network

Mark Goldman
Center for Neuroscience
UC Davis

# Short vs. Long-Term Memory

## Long-term memory

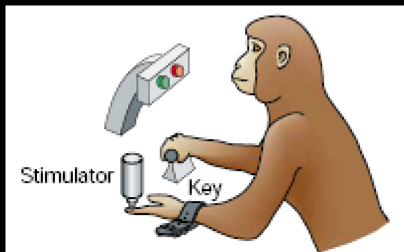
- Can last a lifetime
- Large capacity—can hold many memories
- Mechanism: physical changes in neurons & synapses

## Short-term (a.k.a. "working") memory

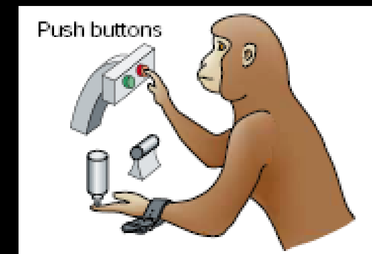
- Lasts ~1-10's of seconds
- Small capacity—only can hold a small number of memories at any time
- Mechanism: neural activity that is sustained in the absence of a stimulus

# Memory-related Neural Activity

Tactile discrimination task:




Delay period

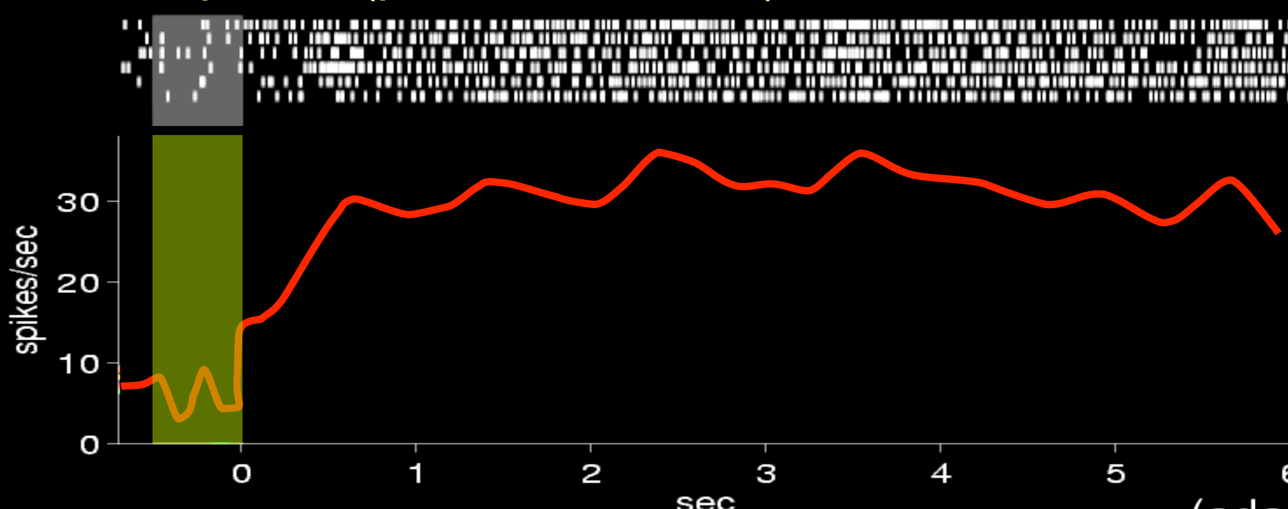



1) Vibration applied to fingertip

2) Remember stimulus frequency

3) Compare to 2nd stimulus: which is higher frequency?


Neuronal response (prefrontal cortex):

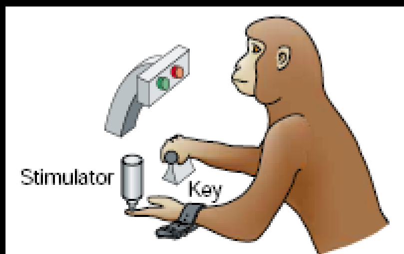



**Persistent neural activity**

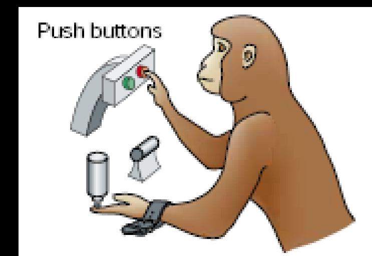
(adapted from Romo et al.)

# Memory-related Neural Activity

Tactile discrimination task:




Delay period

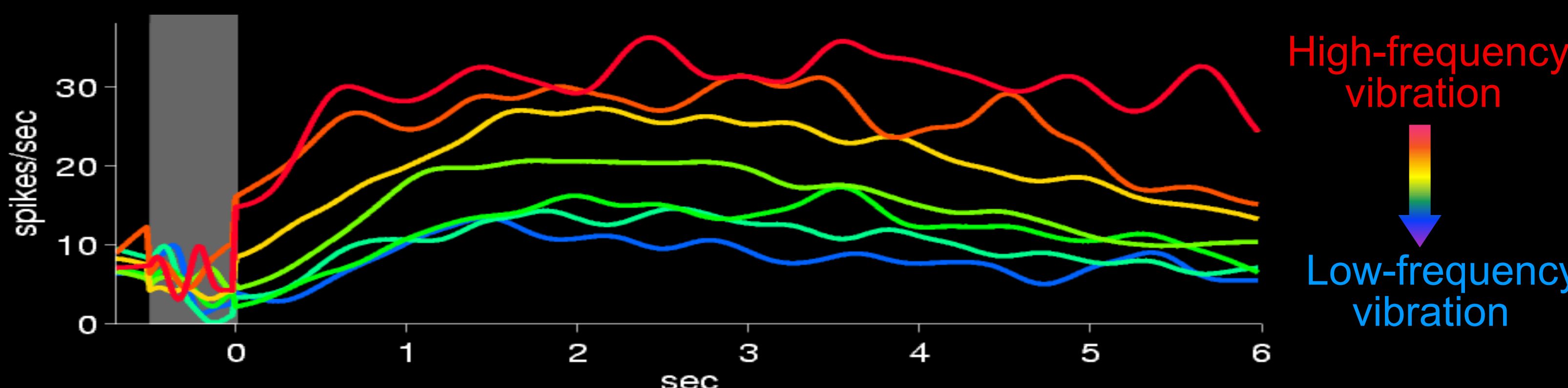



1) Vibration applied to fingertip

2) Remember stimulus frequency

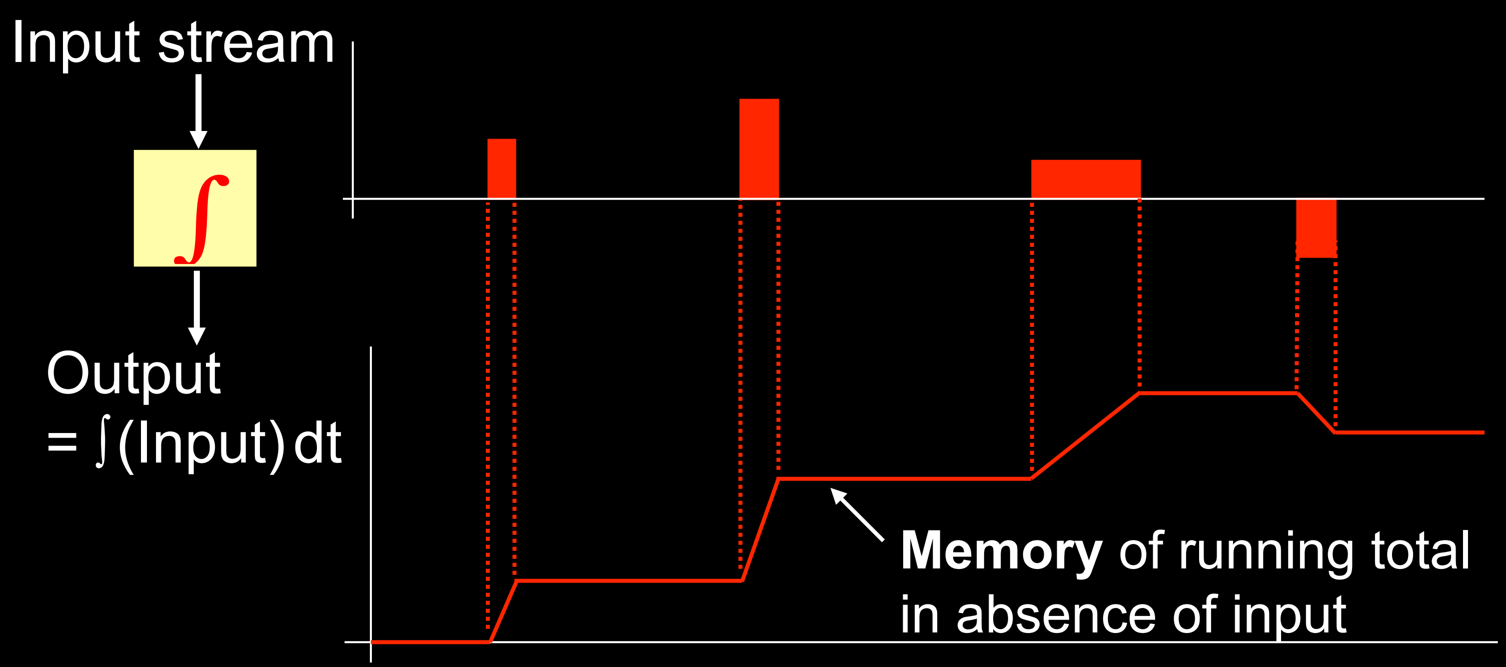
3) Compare to 2nd stimulus: which is higher frequency?


Neuronal response (prefrontal cortex):



# Another Analog Memory System

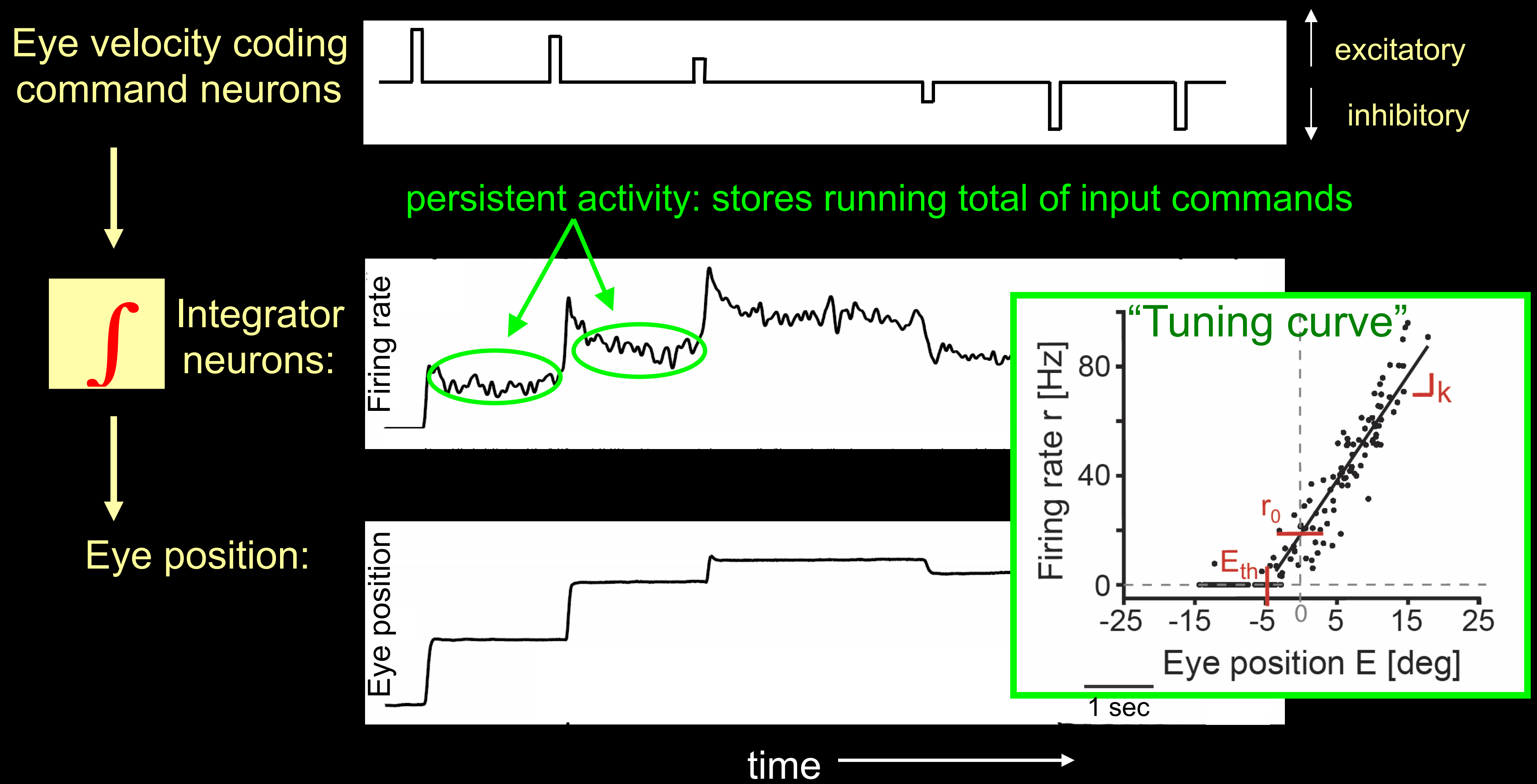
## Integrators: Store the running total of an input




Examples of integrators:

- *Decision making*: -Accumulate noisy evidence over time;
  -Make a decision when threshold is reached
- *Navigation*: Position is determined by integrating velocity signals

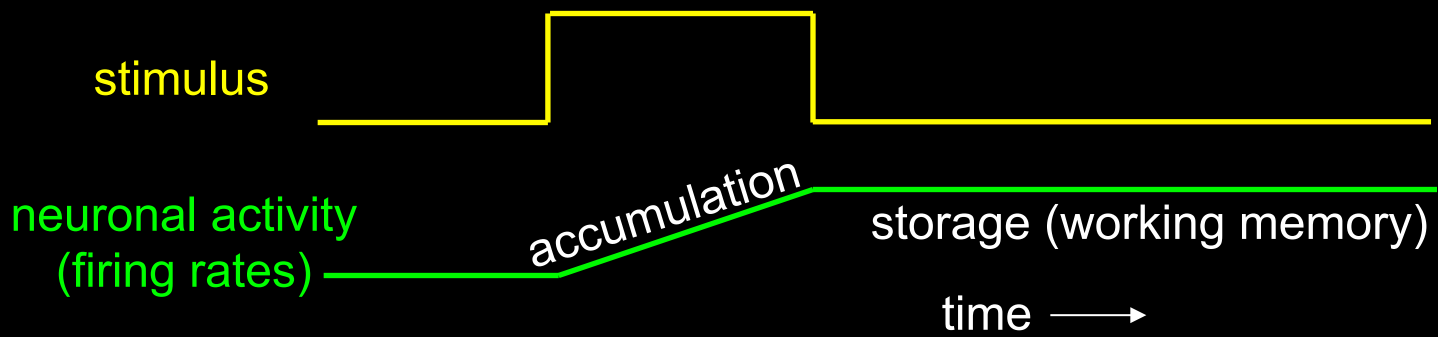
# The Oculomotor Neural Integrator:
# A Network that Stabilizes our Eye Position




(data from Aksay et al., *Nature Neuroscience*, 2001)

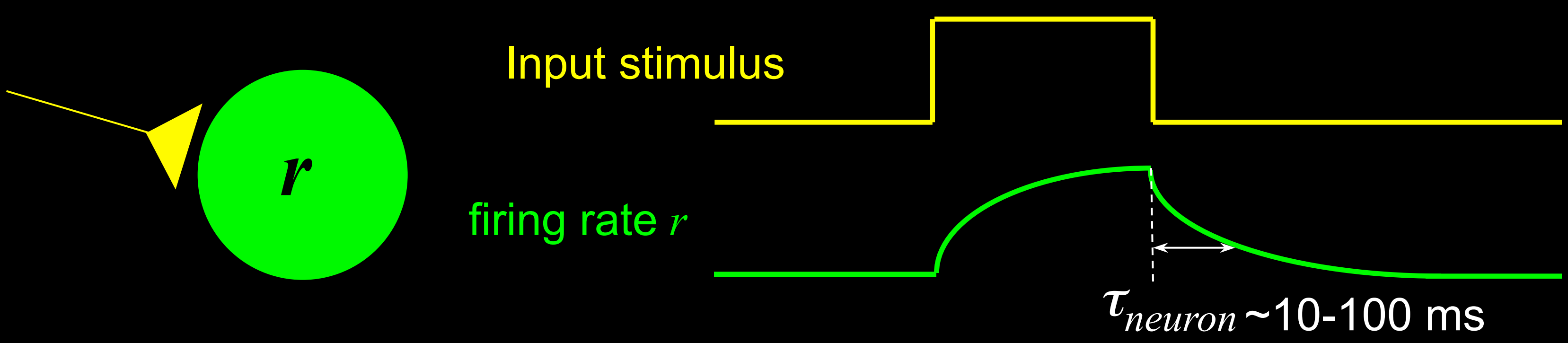
# *Issue*: How do neurons accumulate & store signals in working memory?

- In many memory & decision-making circuits, neurons accumulate and/or maintain signals for ~1-10 seconds

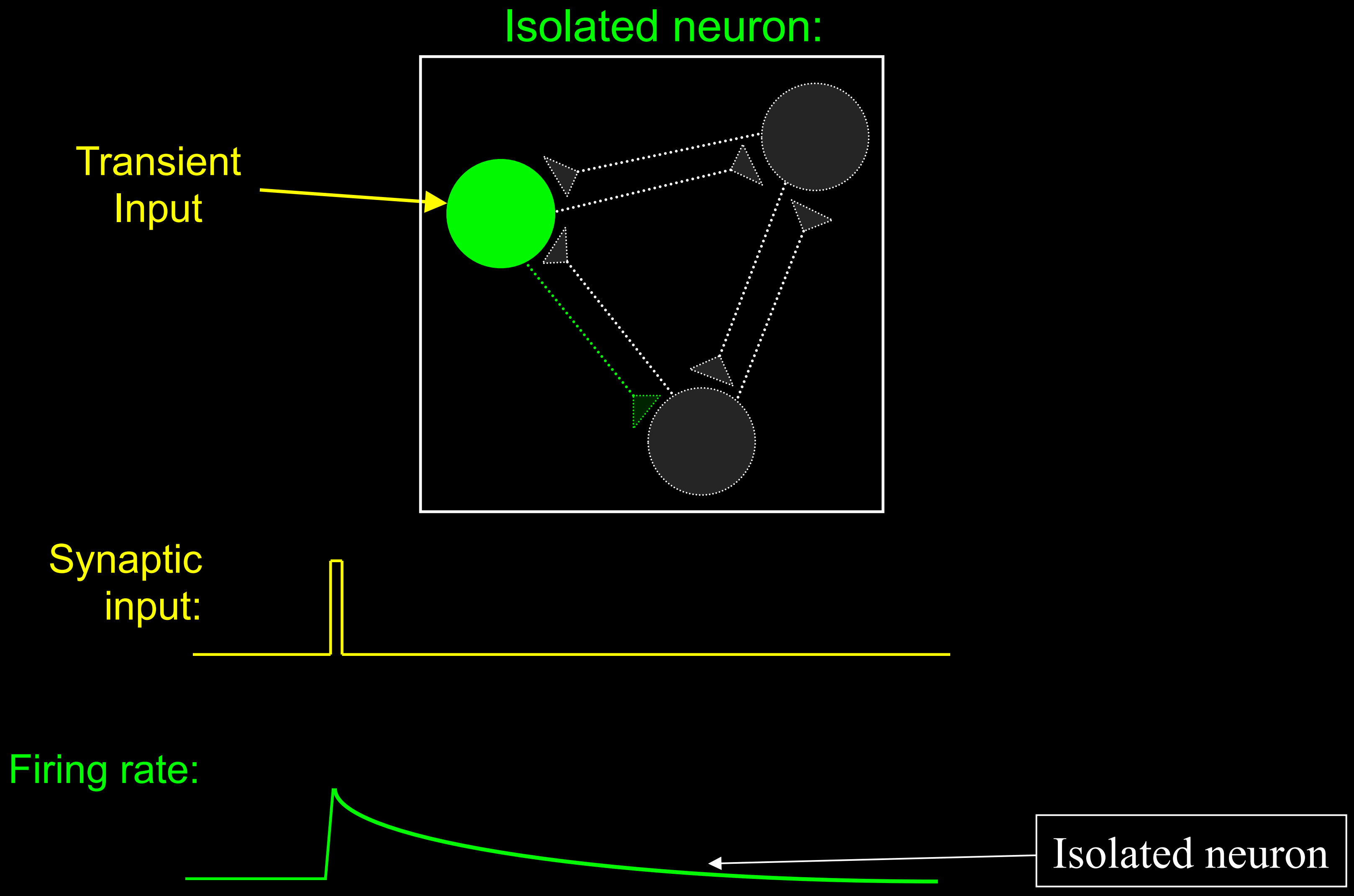



## Puzzle:

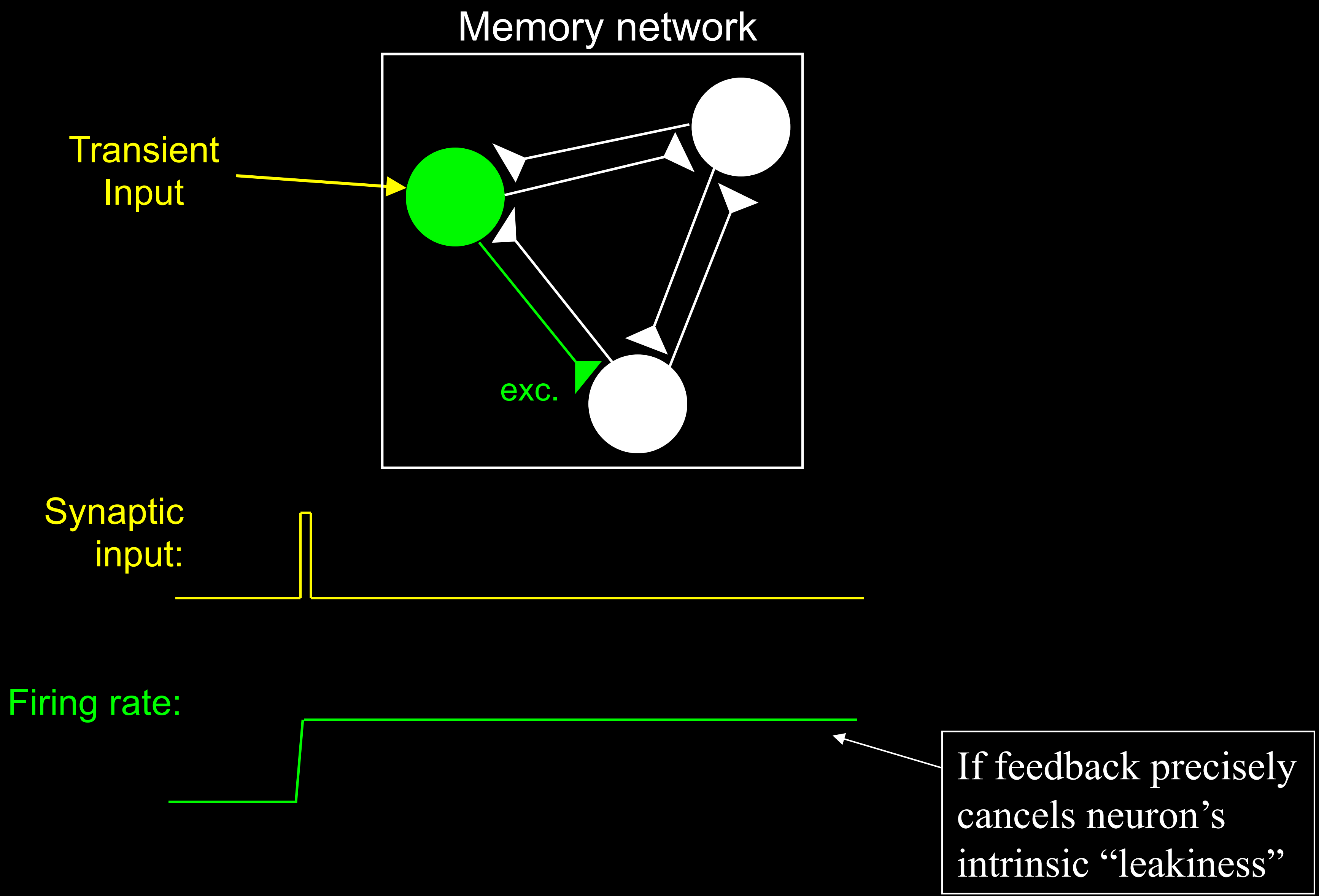
- Most neurons are intrinsically “forgetful”



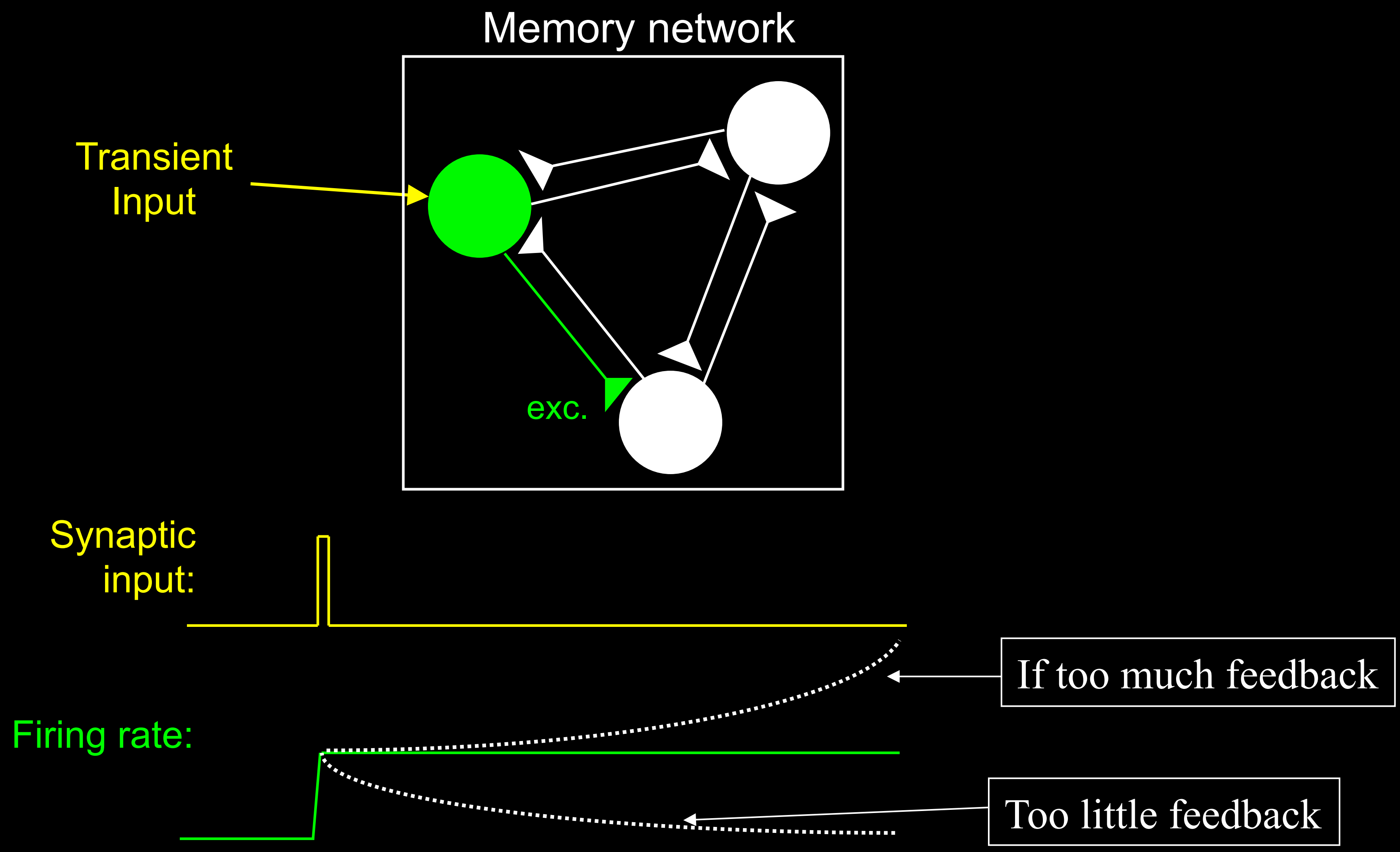
# Traditional model:  Tuned Positive Feedback



# Traditional model:  Tuned Positive Feedback



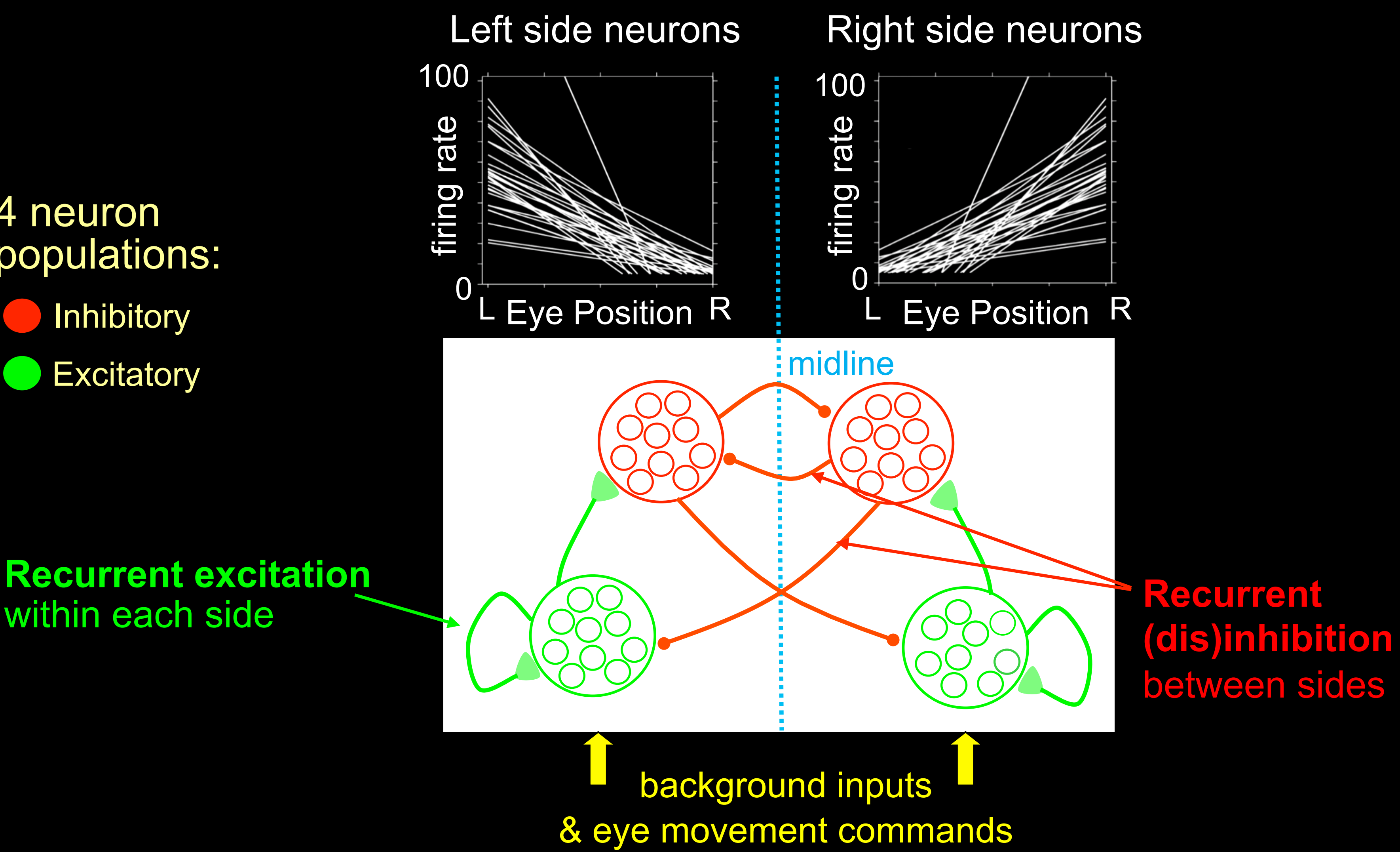
# Traditional model: Tuned Positive Feedback



# Oculomotor Integrator Network

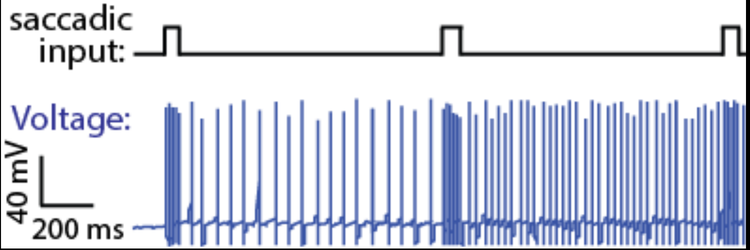
4 neuron populations:

- Inhibitory
- Excitatory

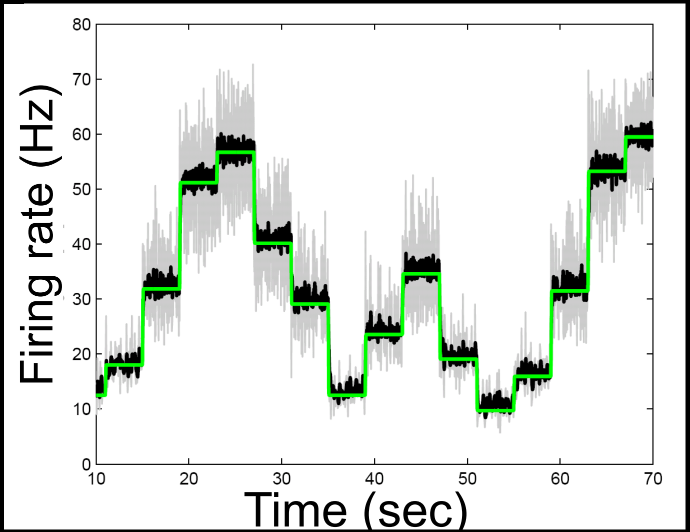


# Model can be Tuned to Integrate its Inputs and Reproduce the Tuning Curves of Every Neuron

*Example model neuron voltage trace:*

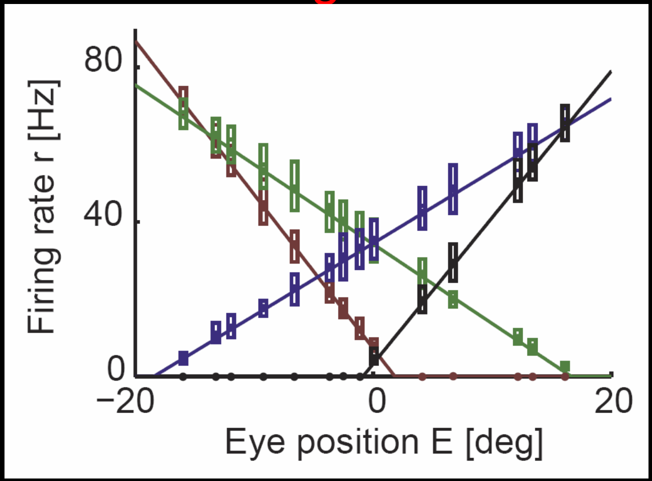



Network integrates its inputs




gray: raw firing rate
(black: smoothed rate)
green: perfect integral

…and all neurons precisely match tuning curve data



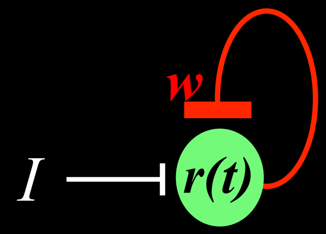

solid lines: experimental tuning curves

boxes: model rates (& variability)

# Back of the Envelope Calculation: Robustness of Analog Memory Network

Integrator equation:

$$\tau_{bio}\frac{dr}{dt} = -1r + wr + I$$




$$\tau_{network} = \frac{\tau_{bio}}{|1-\mathrm{w}|}$$

Experimental values:

Single isolated neuron: $\tau_{bio} \sim 100$ ms

Integrator circuit: $\tau_{network} \sim 30$ sec

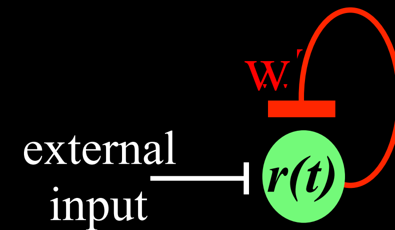
Synaptic feedback w must be tuned to accuracy of:

$$|1-\mathrm{w}| = \frac{\tau_{bio}}{\tau_{network}} \sim 0.3\%$$

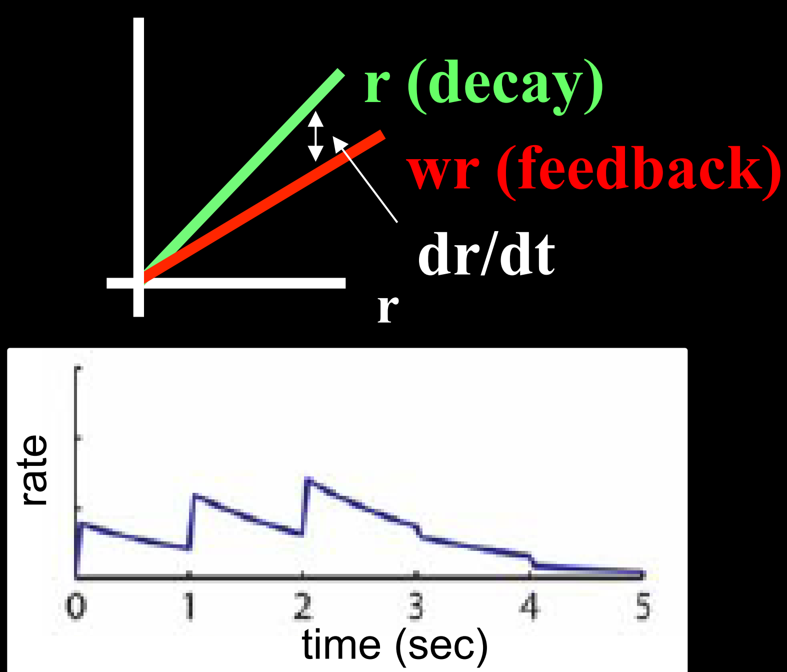
# Robustness Problem in Positive Feedback Memory Models

Fine-tuned model:

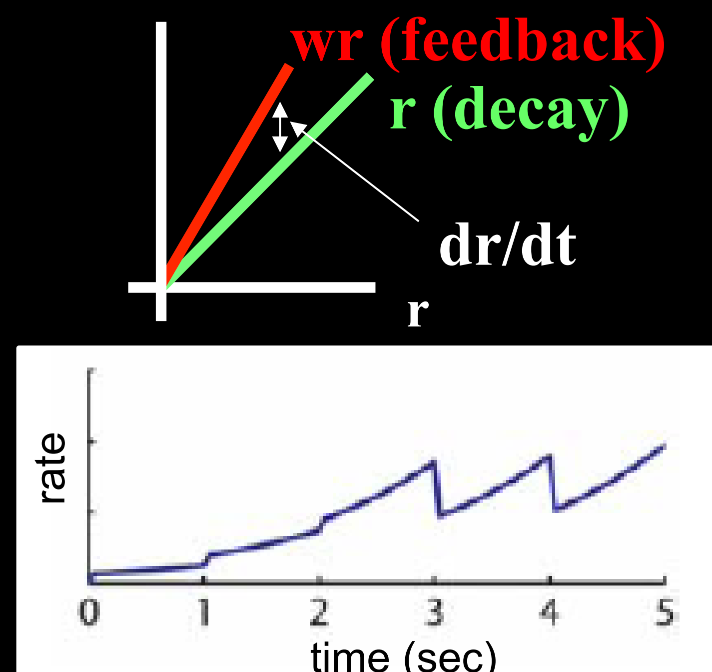
$$\tau_{neuron}\frac{dr}{dt} = \underbrace{-r}_{\text{decay}} + \underbrace{wr}_{\text{feedback}} + \textit{external input}$$



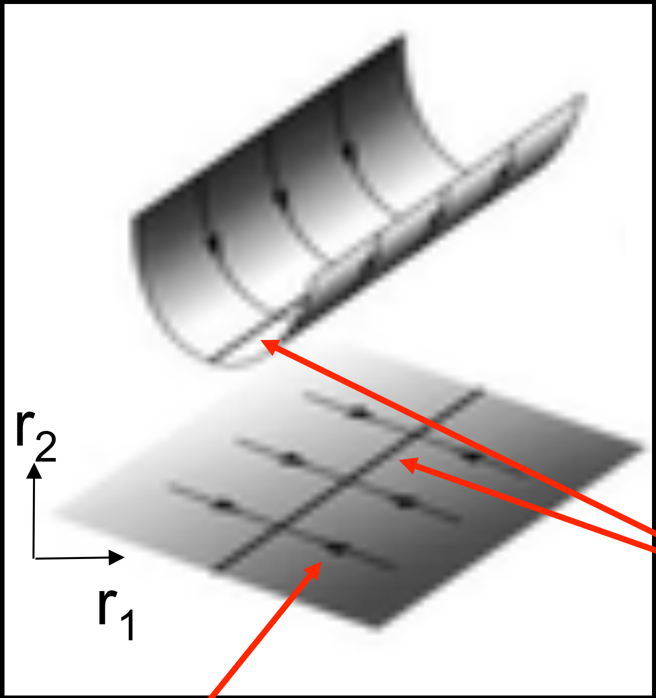

## Leaky behavior




## Unstable behavior



# Geometrical ("Line attractor") Picture of Analog Memory Storage & the Robustness Problem

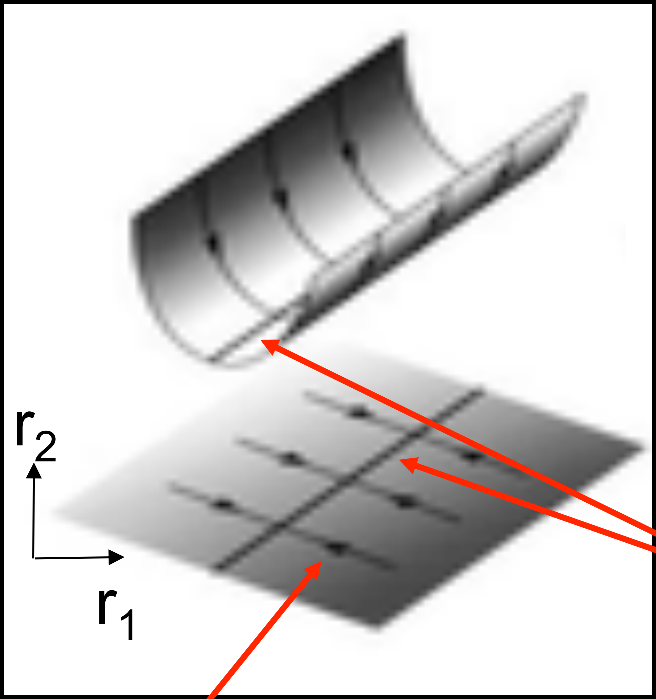


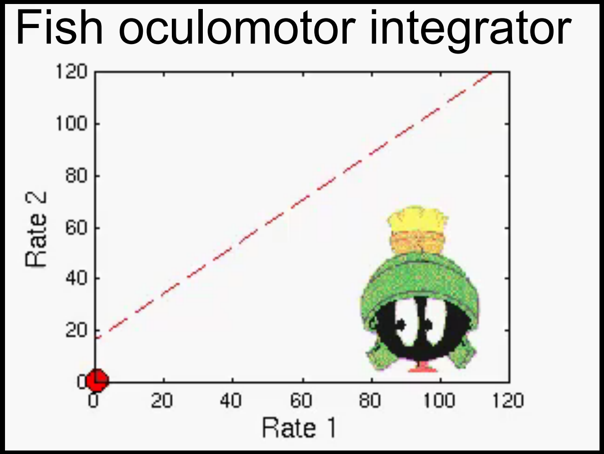

Network state maintained stably at any point along trough of "energy" surface

Activity decays along other directions

⇨ "Line attractor", or "Line of fixed points"

# Geometrical ("Line attractor") Picture of Analog Memory Storage & the Robustness Problem




(H.S. Seung, D. Lee)

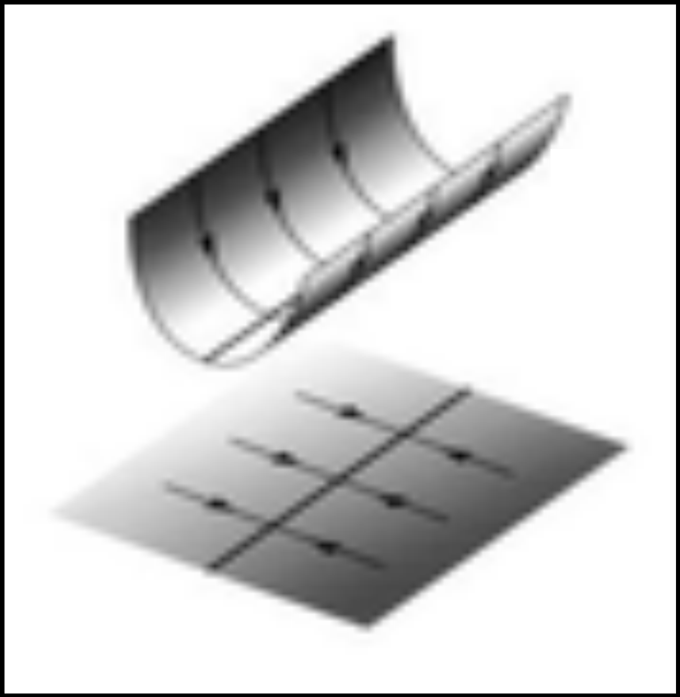
Network state maintained stably at any point along trough of "energy" surface

Activity decays along other directions

⇒ "Line attractor", or "Line of fixed points"

Problem: 1) Noise → diffusion of memory representation
2) If surface isn't flat (network isn't tuned perfectly), network activity state slips!

# Geometry of Robustness
# & Hypotheses for Robustness on Faster Time Scales

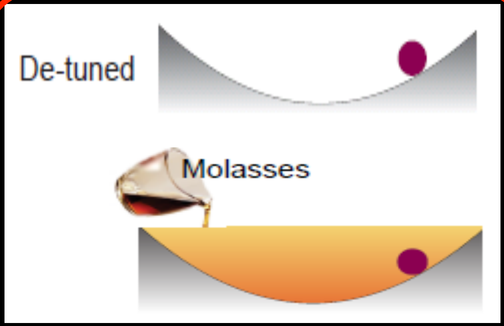


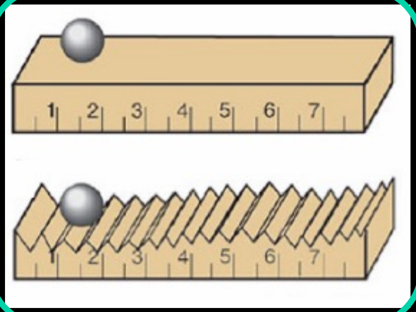
1) Plasticity on slow time scales:
Reshapes the trough to make it flat

2) To control on faster time scales:
Add ridges to surface to add "friction"-like slowing of drift

-OR-

Fill attractor with viscous fluid to slow drift

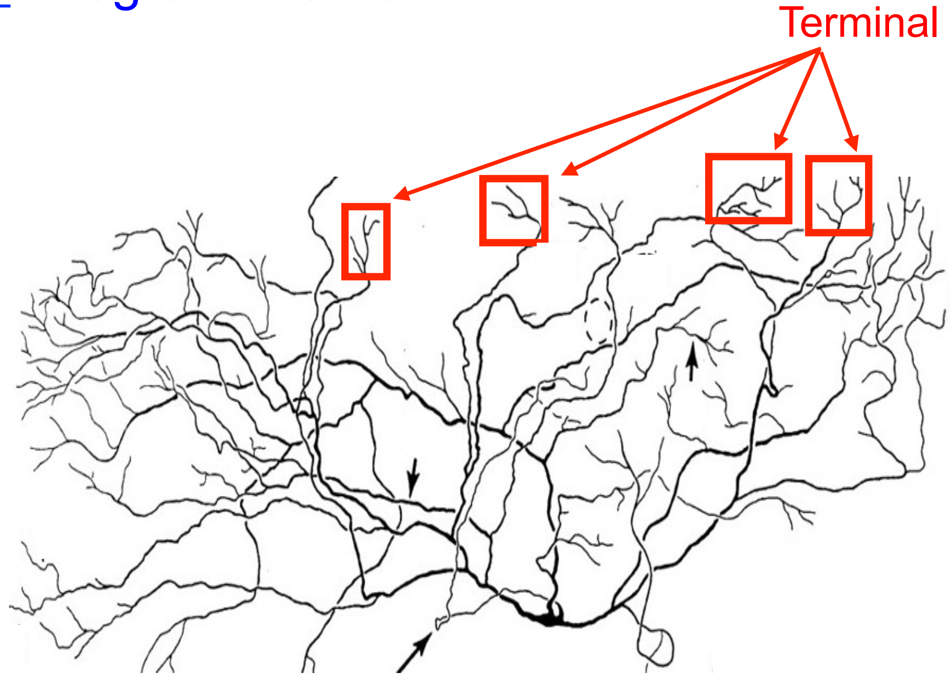


# **Idea 1:** Neurons may have intrinsic properties that help to maintain the persistent neural activity

❖ Concept: Dendritic branchlets may act as <u>bistable, digital elements</u> (i.e. <u>flip-flops</u>) that add robustness to the circuit

**<u>Real</u>** integrator neuron:




=

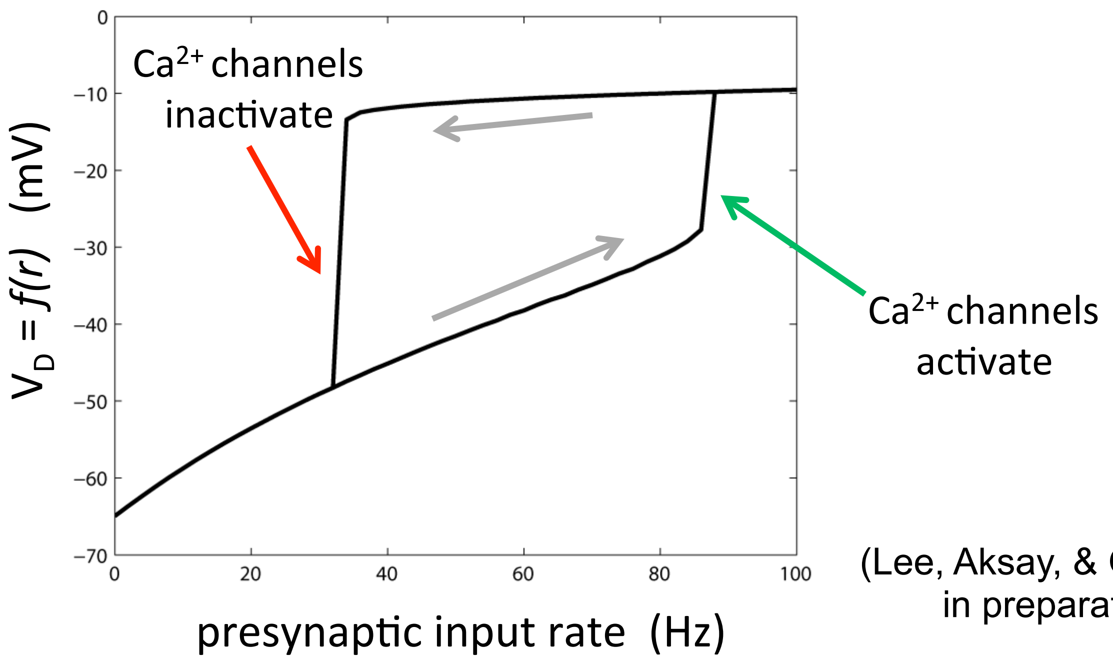
Ratchet:
resists slippage

# Evidence for dendritic bistability & independence

1) Dendritic bistability has been observed experimentally

-Due to the *self-sustaining* properties of, e.g., NMDA, NaP, or $Ca^{++}$ channels

Steady-state voltage-firing rate relation

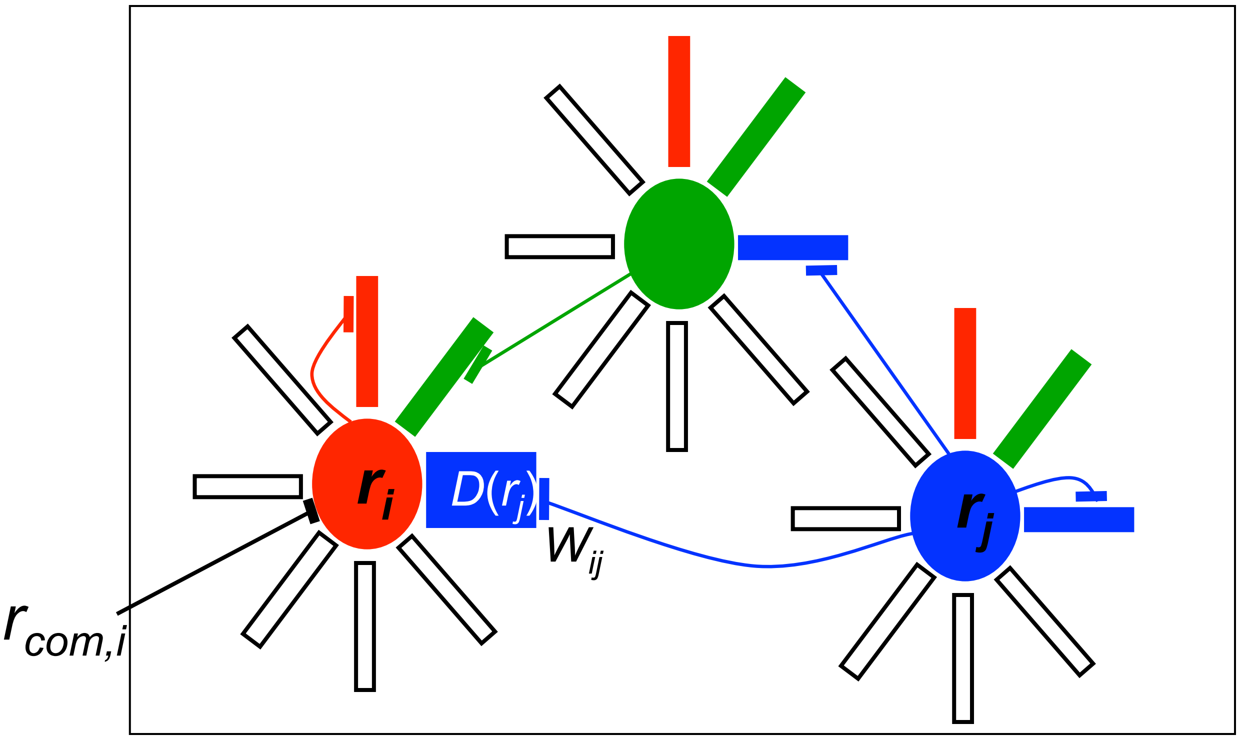



(Lee, Aksay, & Goldman, in preparation)

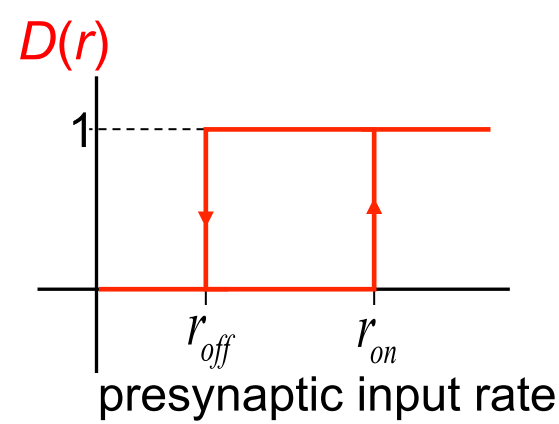
2) Anatomically realistic models suggest that different dendritic branches may behave approximately independently (Koch et al., 1983; Poirazi et al., 2003)

# Network with Bistable Dendrites

Network of *N* neurons, each with *N* identical dendrites:



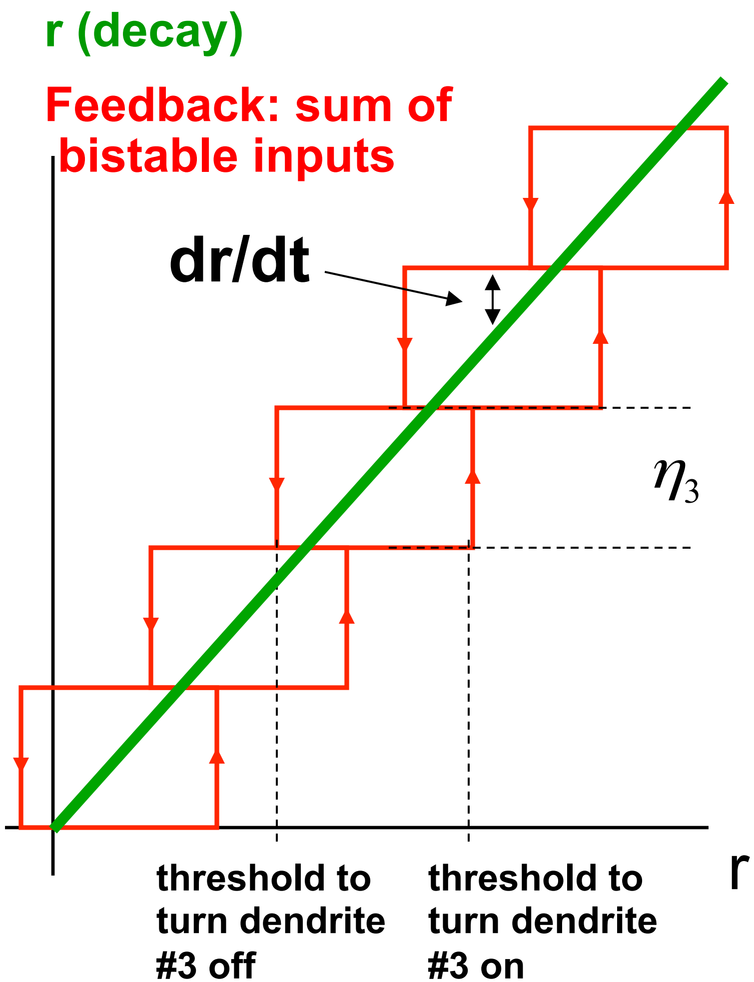

Bistable dendritic response



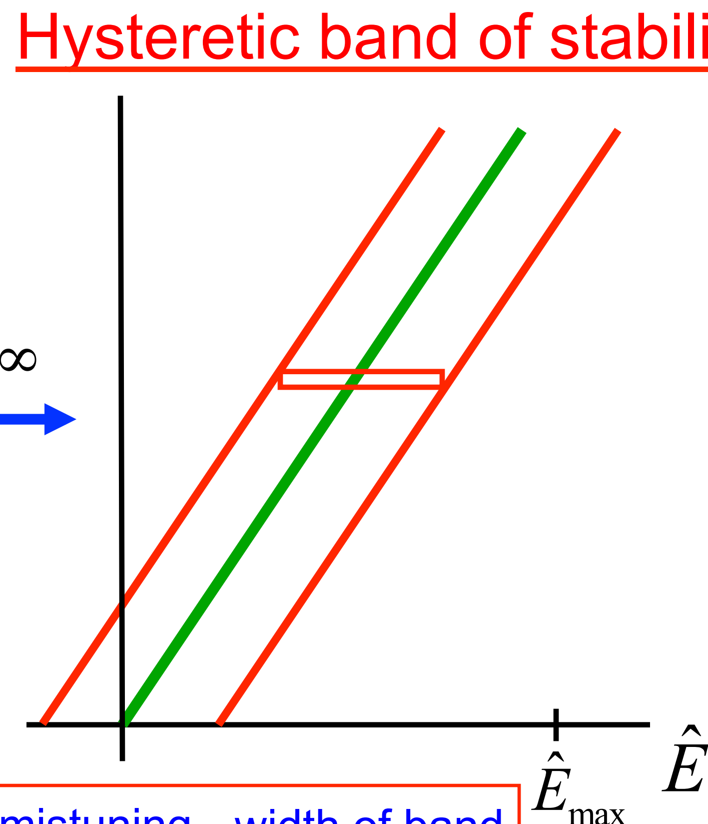

$$\tau \frac{dr_i}{dt} = \underbrace{-r_i}_{\text{decay}} + \underbrace{\sum_{j=1}^{N} W_{ij} D(r_j)}_{\text{recurrent input}} + \underbrace{r_{com,i}}_{\text{eye movement commands}}$$

# Graphical Solution of Balanced Leak & Feedback

During fixations: $$\tau \frac{dr_i}{dt} = -r_i + \sum_{j=1}^{N} W_{ij} D(r_j)$$




$N \rightarrow \infty$

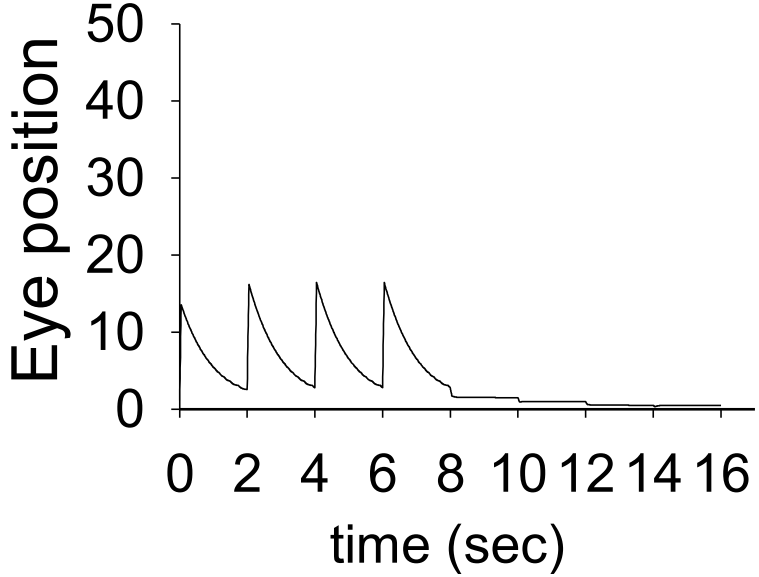



% mistuning tolerated ~ $\dfrac{\text{width of band}}{\hat{E}_{max}}$

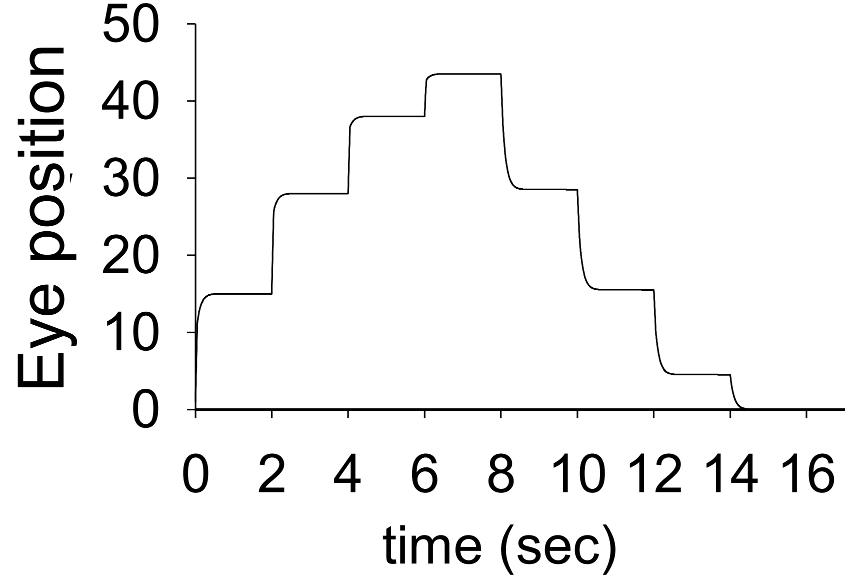
(Goldman et al., 2003; see also Koulakov et al., 2002)

# Comparison of Robustness With & Without Bistability

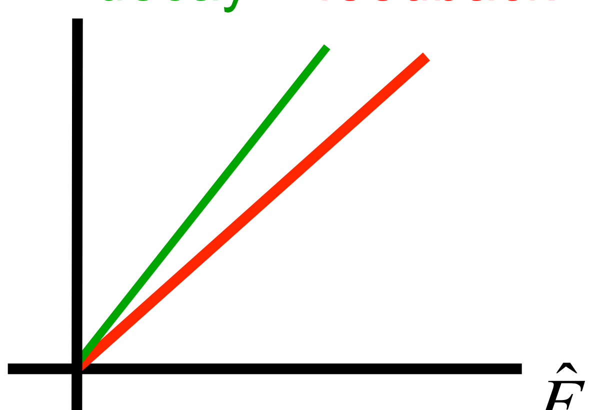
**10% mistuning, no bistability**



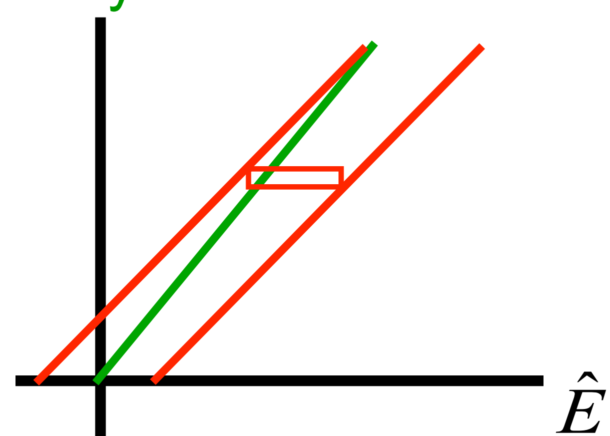

**10% mistuning, w/ bistability**




decay > feedback



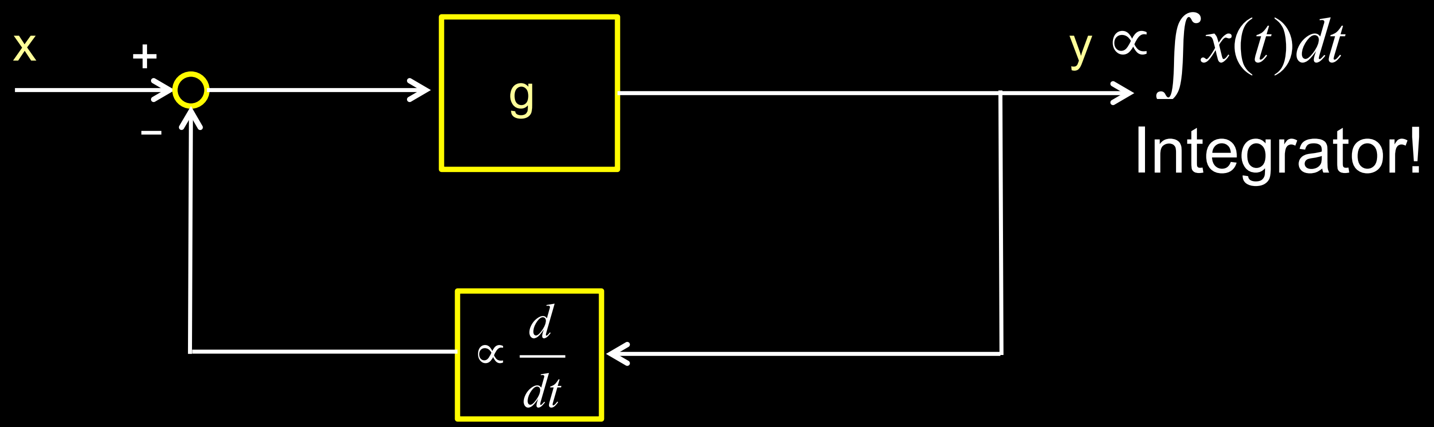

decay within feedback band



# **Idea 2**: Designing Networks to be Robust to Common Perturbations

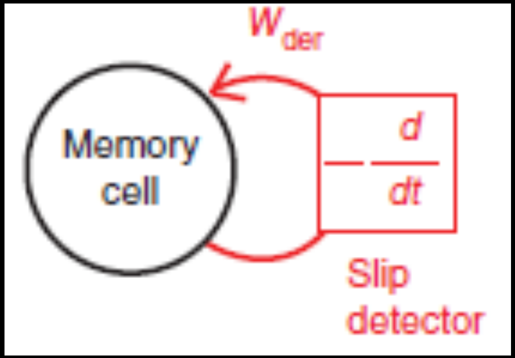
Fundamental control theory result:

*Strong negative feedback* of a signal produces an output equal to the inverse of the negative feedback signal



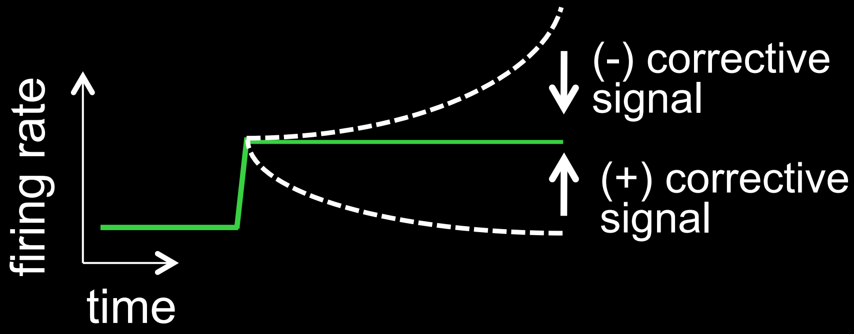
# Persistent Activity from Negative-Derivative Feedback

## Math:

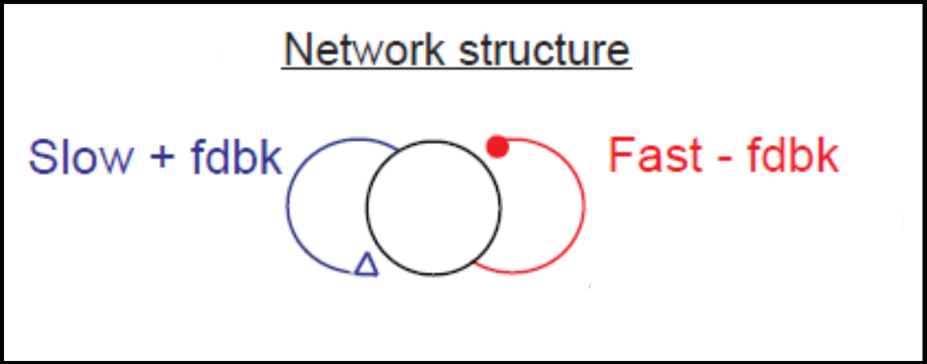



$$\tau \frac{dr}{dt} = -r - W_{der.} \frac{dr}{dt} + Input$$

## Picture:

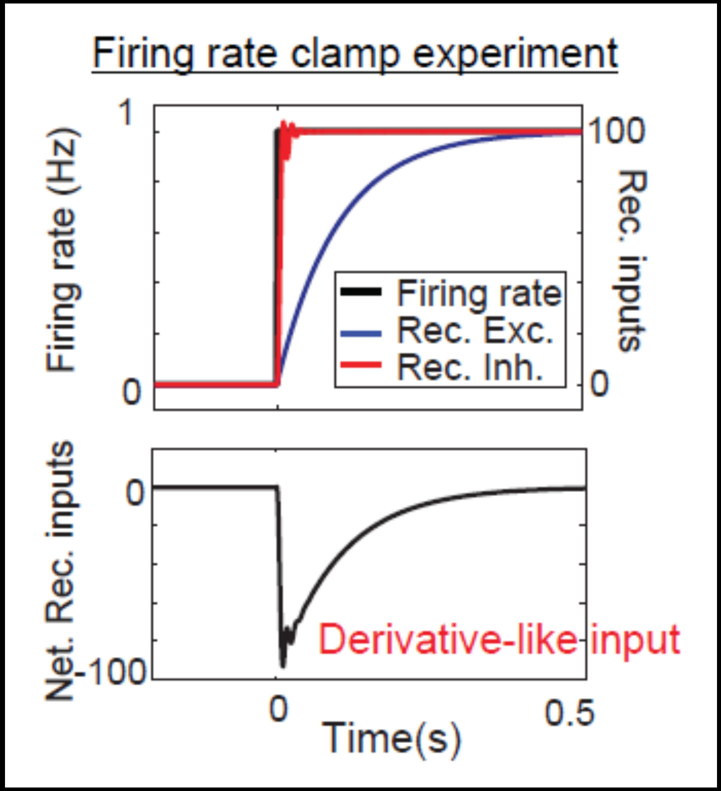



# Negative derivative feedback arises naturally in balanced cortical networks



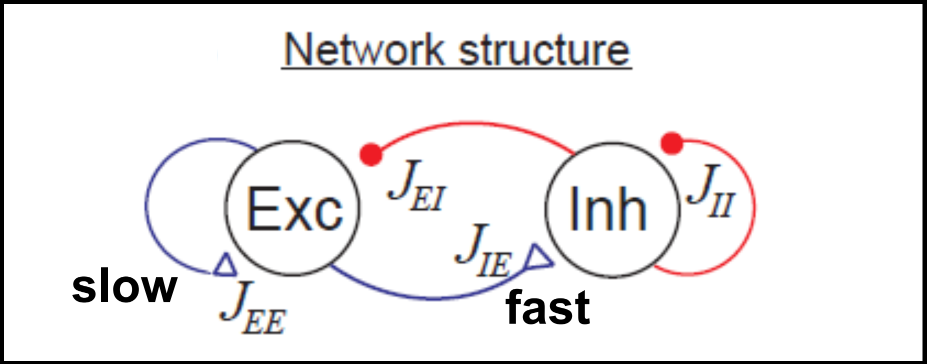

Derivative feedback arises when:

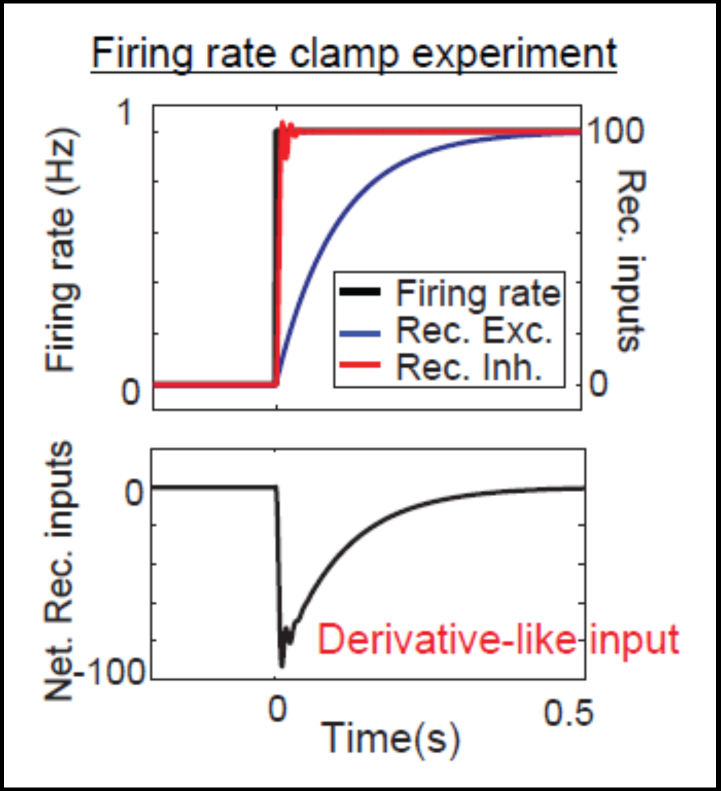
1) Positive feedback is slower than negative feedback
2) Excitation & Inhibition are balanced




Lim & Goldman, *Nature Neuroscience*, 2013

# Negative derivative feedback arises naturally in balanced cortical networks



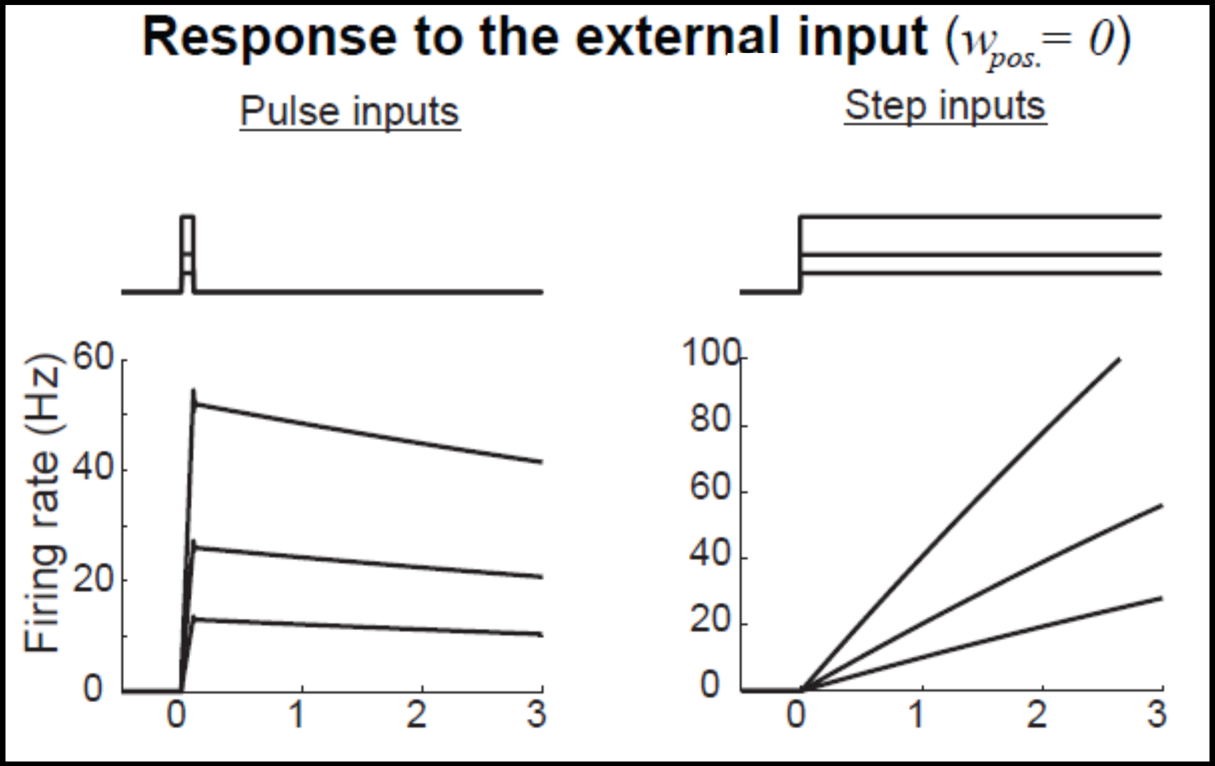




Derivative feedback arises when:

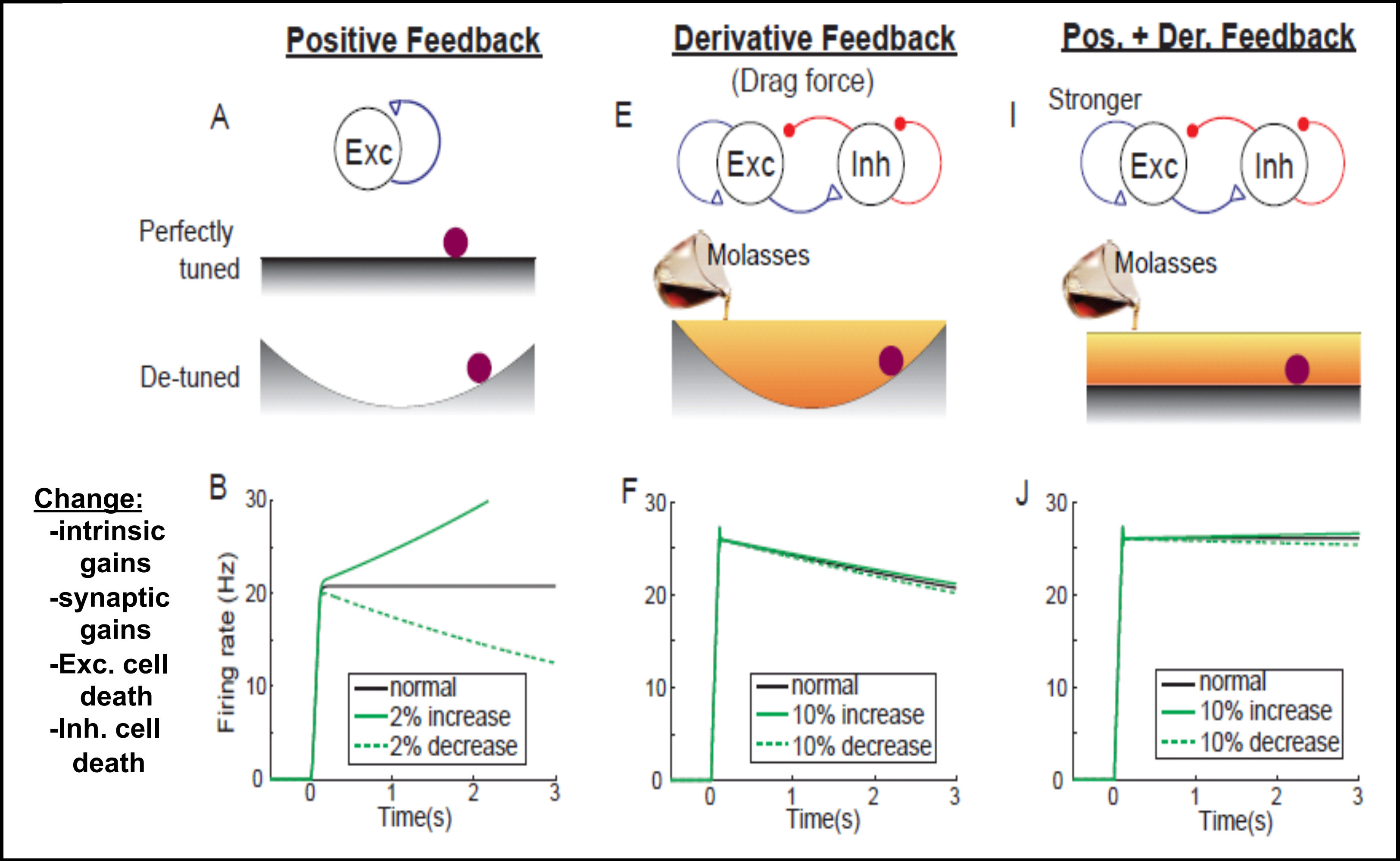
1) Positive feedback is slower than negative feedback
2) Excitation & Inhibition are balanced

Lim & Goldman, *Nature Neuroscience*, 2013

# Networks Maintain Analog Memory and Integrate their Inputs



# Robustness to Loss of Cells or Perturbations in Intrinsic or Synaptic Gains



# Summary

- Short-term memory (~10's seconds) is maintained by *persistent neural activity* following the offset of a remembered stimulus

- Classic model: Line Attractor made from Positive Feedback

  → Open Issue: Inherently requires fine-tuning/is not robust to perturbations

- Possible missing concepts (for memory & more generally…)

  → Neurons are smarter than simple linear filters plus static nonlinearites

  → Well-designed systems aren't robust to everything, but are robust to the most common perturbations they experience (…but how do we determine what these are???)

# Acknowledgments

Theory (Goldman lab, UCD)
Michiel Berends
Itsaso Olasagasti (USZ)
Dimitry Fisher (Brain Corp.)
Sukbin Lim (U. Chicago/
NYU-Shanghai)

Experiments
David Tank (Princeton Univ.)
Emre Aksay (Cornell Med.)
Guy Major (Cardiff Univ.)
Robert Baker (NYU Medical)

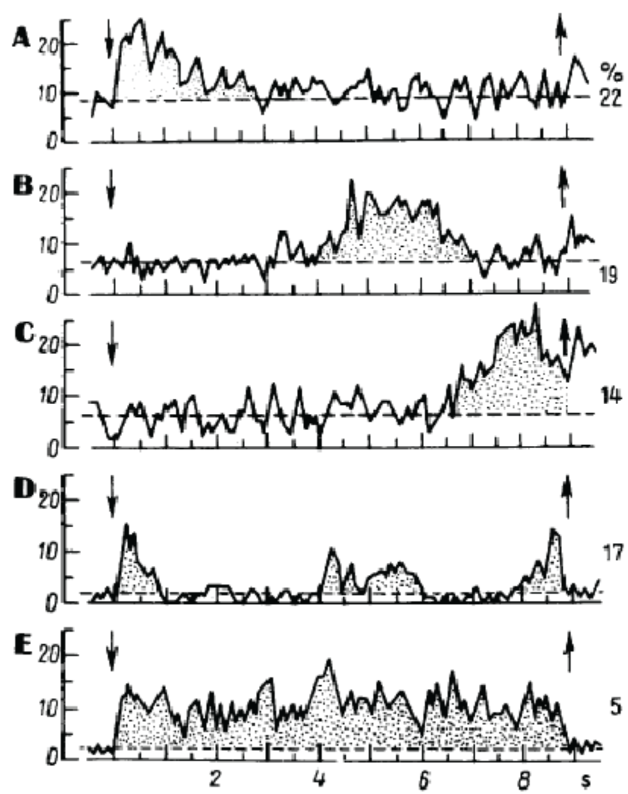
## Idea 3:

- Are attractors, created through feedback loops, even necessary?

- Could there be advantages to alternative, higher-dimensional representations?

# Working memory task not easily explained by traditional feedback models

5 neurons recorded during a PFC delay task (Batuev et al., 1979, 1994):




Early stage neuron?

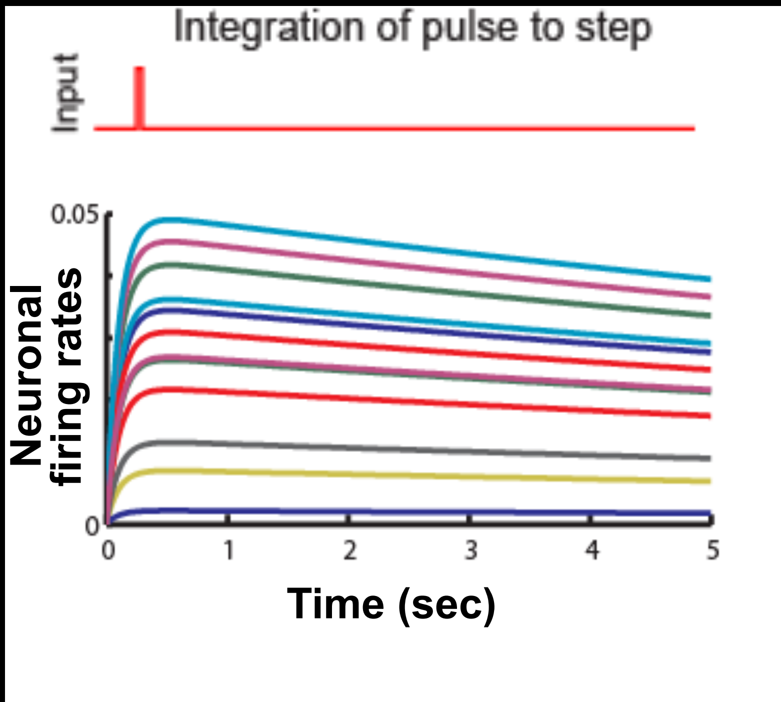
Middle stage neuron?

Late stage neuron?

Sum early, mid, late stage neurons?

Sum all stages?

# Response of Individual Neurons in Line Attractor Networks



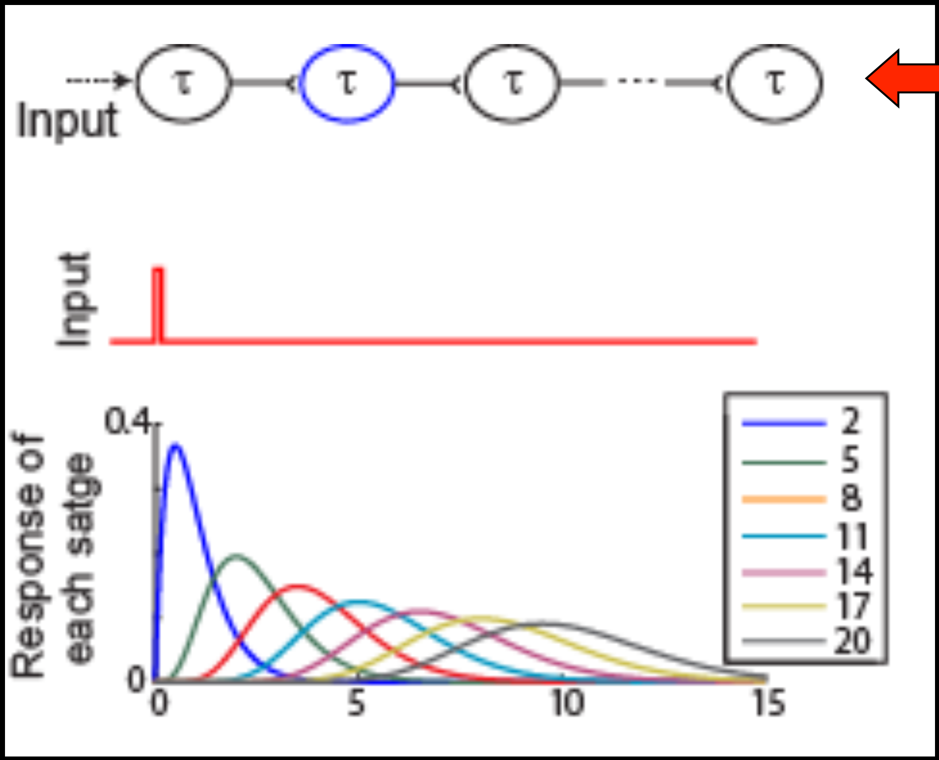

All neurons exhibit similar slow decay:
Due to strong coupling that mediates positive feedback

Problem: Does not reproduce the differences between neurons seen experimentally!

# Feedforward Networks Can Integrate!

(Goldman, *Neuron*, 2009)

Simplest example:

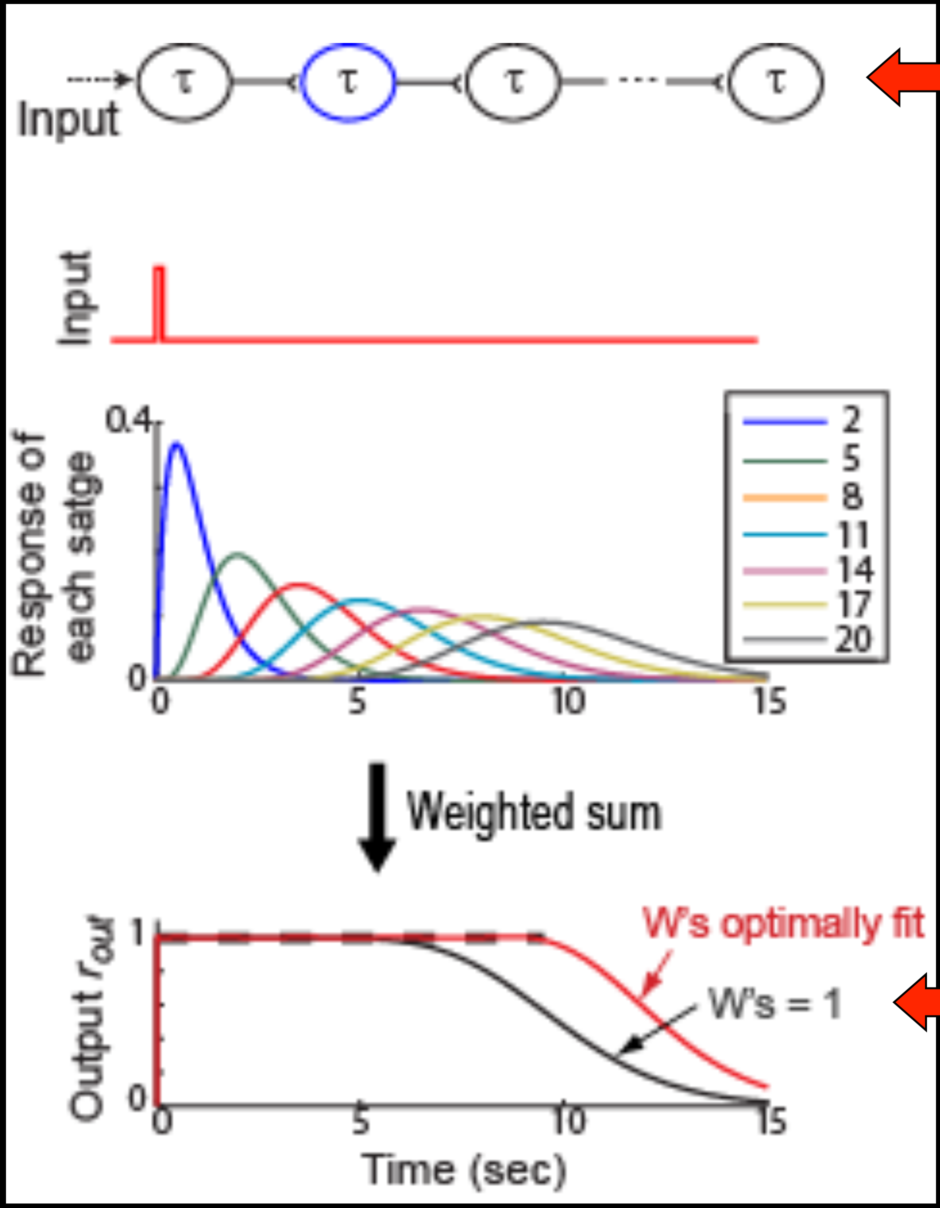



Chain of neuron clusters that successively filter an input

# Feedforward Networks Can Integrate!

(Goldman, *Neuron*, 2009)

Simplest example:



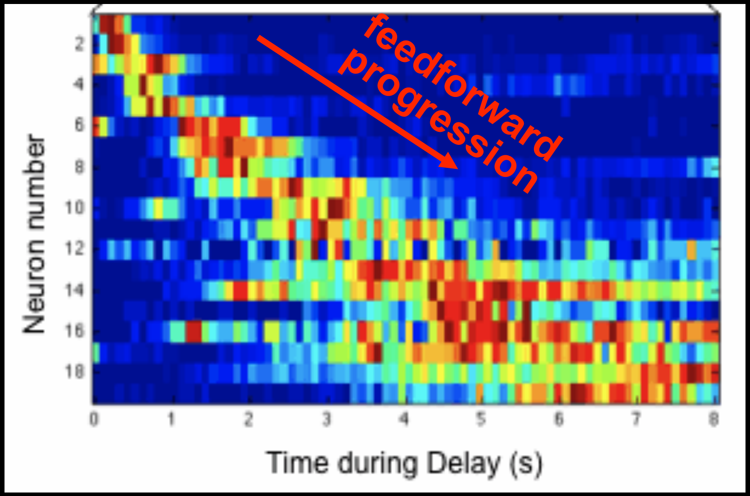

Chain of neuron clusters that successively filter an input

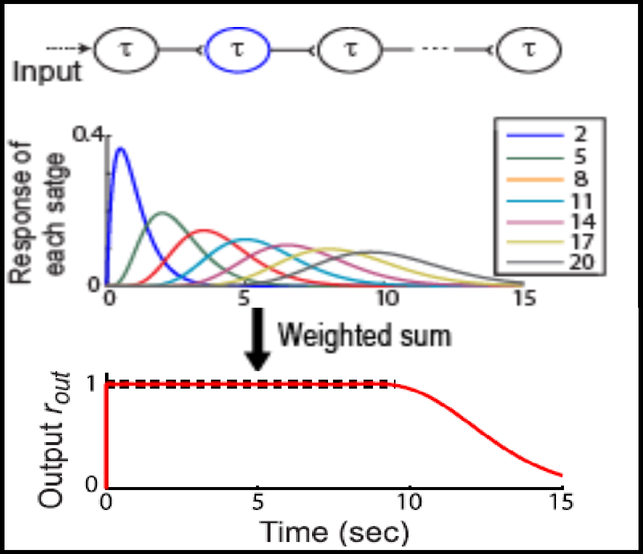
Integral of input!
(up to duration ~N$\tau$)

(can prove this works analytically)

# Recent data: “Time cells” observed in rat hippocampal recordings during delayed-comparison task

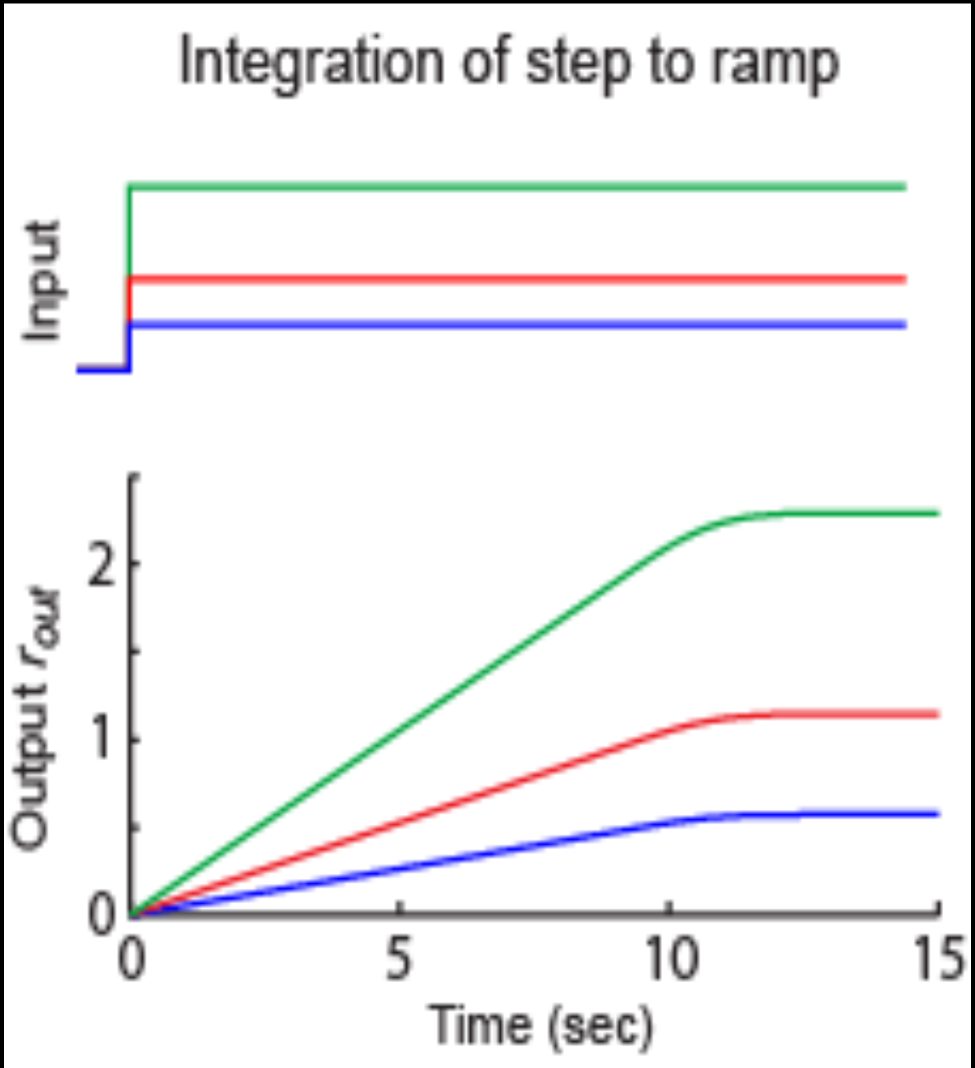
Data courtesy of H. Eichenbaum [Similar to data of Pastalkova et al., *Science*, 2008; Harvey et al., *Nature*, 2012]



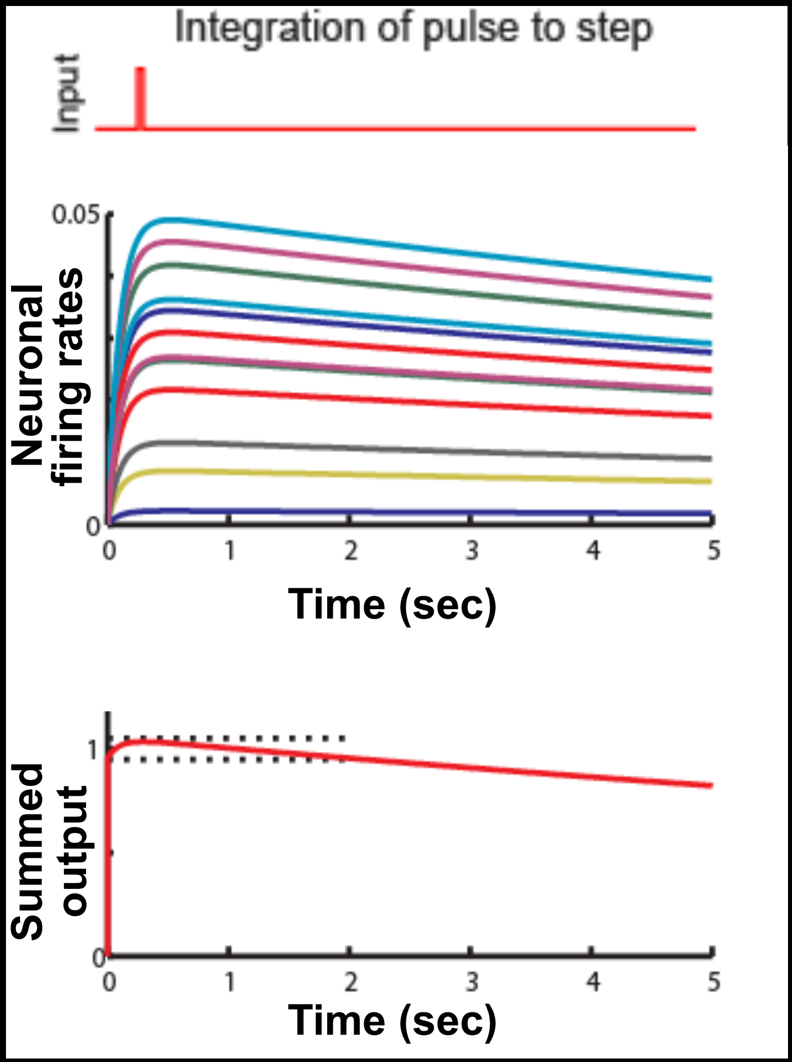

(Goldman, *Neuron*, 2009)

# Same Network Integrates Any Input for ~Nτ

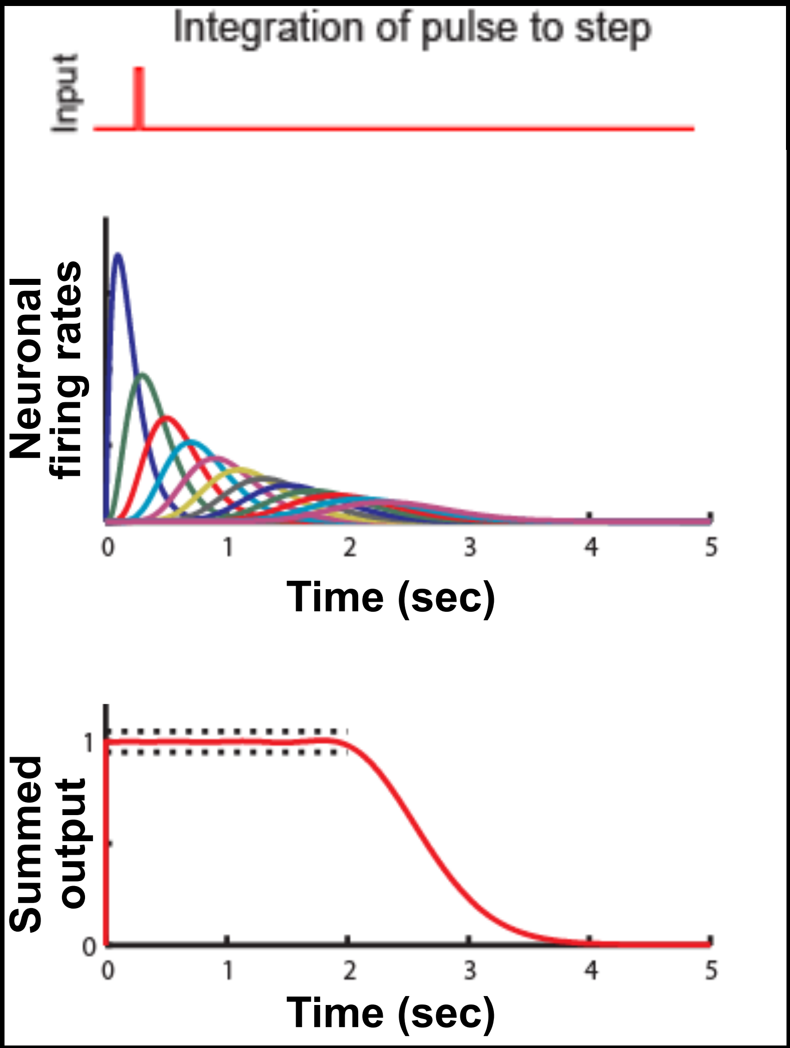


# Improvement in Required Precision of Tuning

Feedback-based Line Attractor:
10 sec decay to hold 2 sec of activity

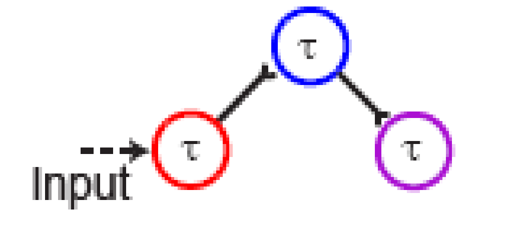



Feedforward Integrator
2 sec decay to hold 2 sec of activity

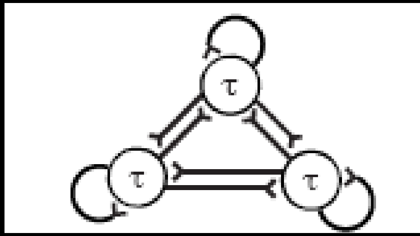


# Generalization to Coupled Networks: Feedforward transitions between *patterns* of activity

Feedforward network




Recurrent (coupled) network



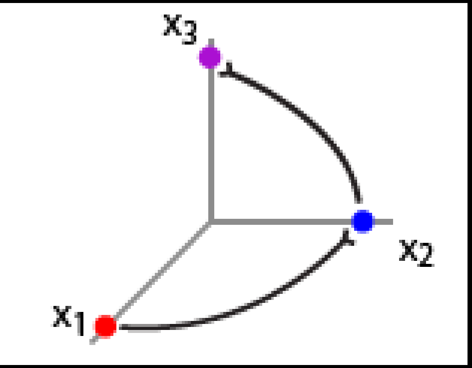

Connectivity matrix $W_{ij}$: $\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$

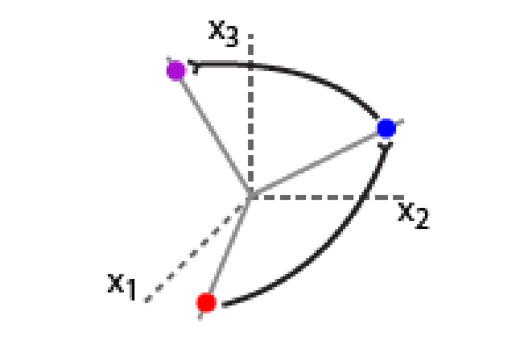
$\mathbf{W}_{recurrent} = \mathbf{RWR}^{-1}$

(Schur decomposition)

Map each neuron to a combination of neurons by applying a coordinate rotation matrix **R**

Geometric picture:

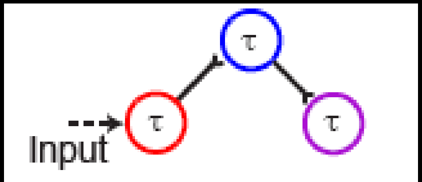


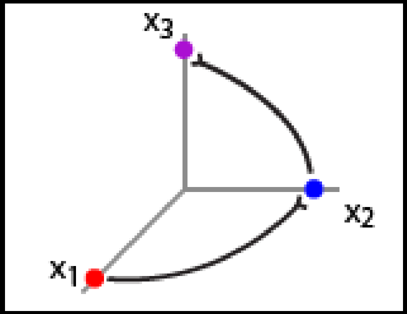




(Math of Schur: See Goldman, *Neuron*, 2009; Murphy & Miller, *Neuron*, 2009; Ganguli et al., *PNAS*, 2008)

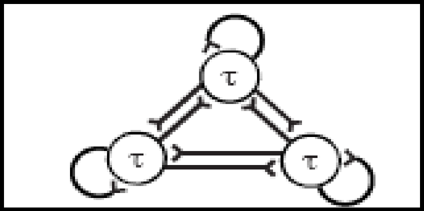
# Responses of functionally feedforward networks

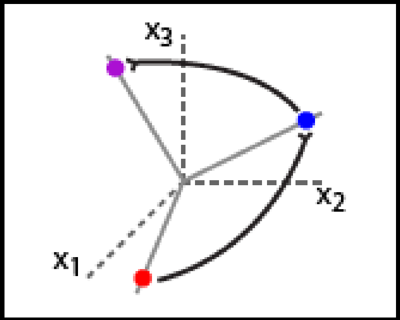
Feedforward
network activity patterns



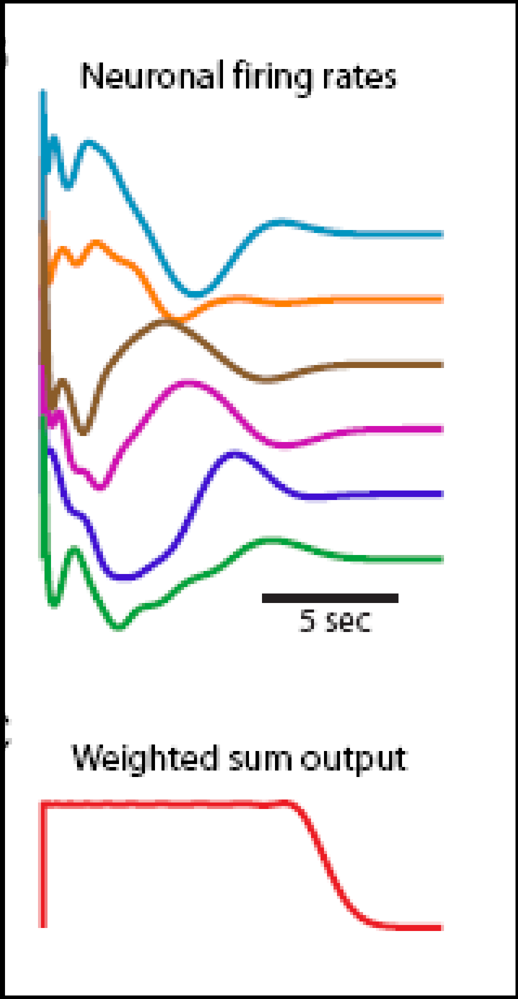




Functionally feedforward
*activity patterns…*

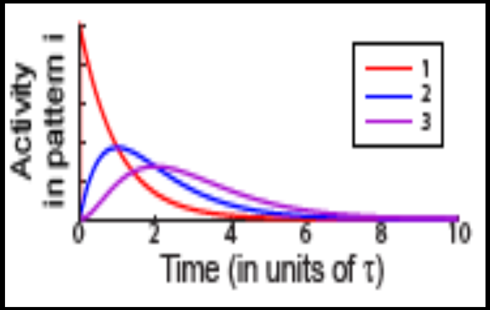


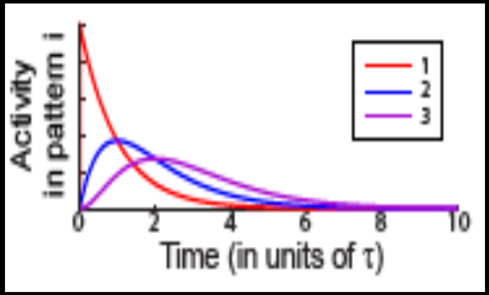




*& neuronal firing rates*




Effect of stimulating pattern 1:




Goldman, *Neuron*, 2009