Models of memory

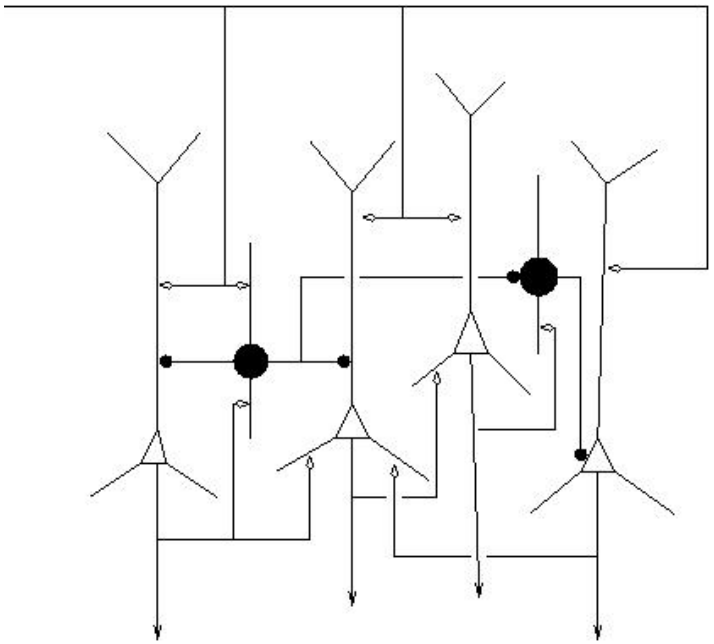
Nicolas Brunel

Laboratory
of Neurophysics
and Physiology

Laboratoire
de Neurophysique
et Physiologie

UMR 8119
CNRS - Université René Descartes
45 rue des Saints Pères
75270 Paris cedex 06, France
Tel. (33) 1 42 86 21 38
Fax (33) 1 49 27 90 62

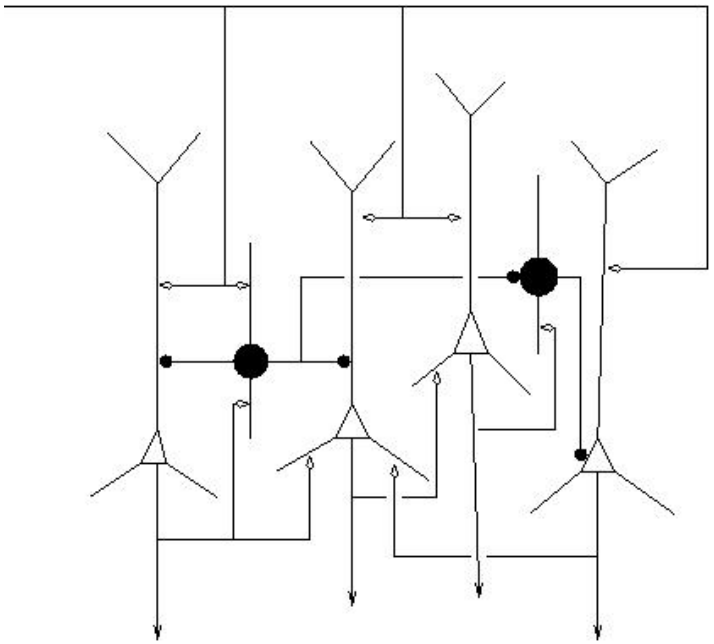
Mechanisms of memory

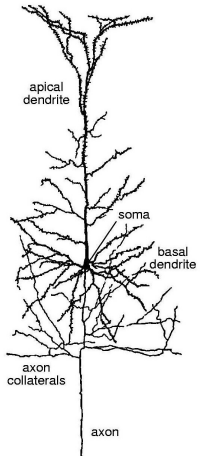


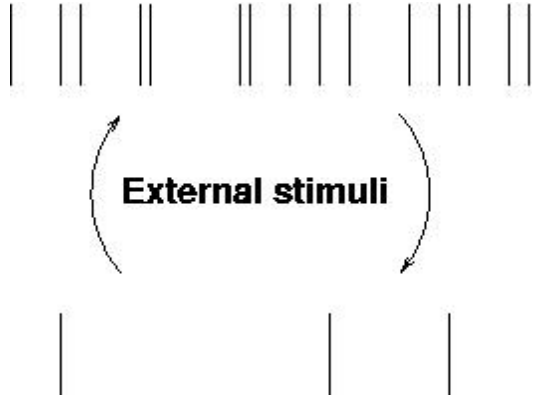
Mechanisms of memory

- Short-term (working) memory:

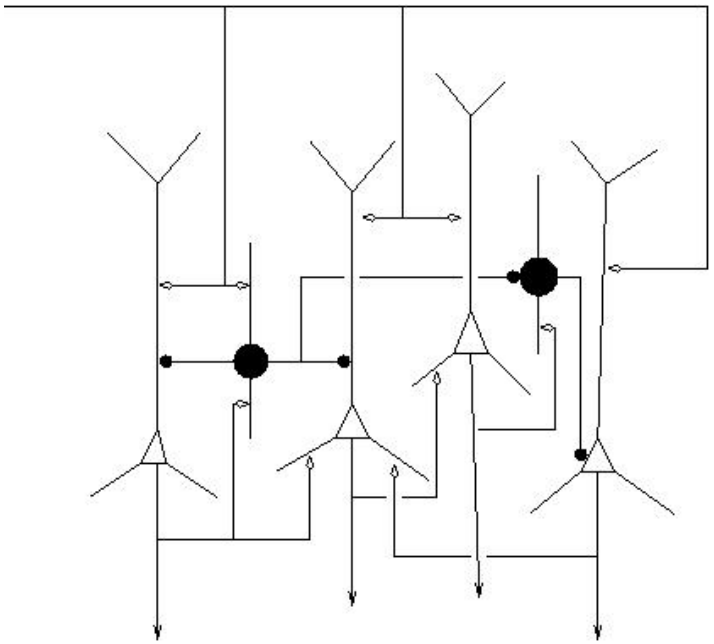
  persistent activation of neurons;





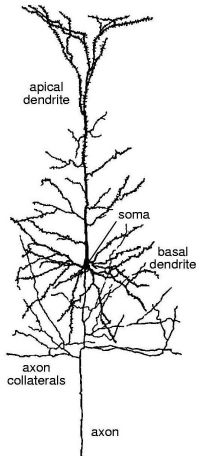


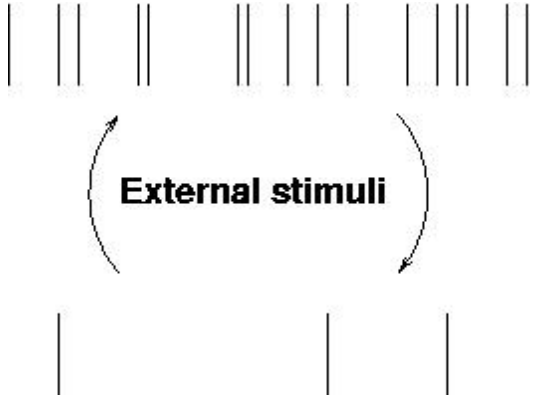
Mechanisms of memory



- Short-term (working) memory:

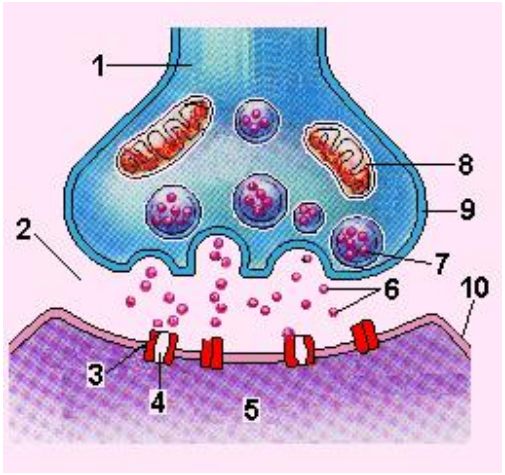
  persistent activation of neurons;

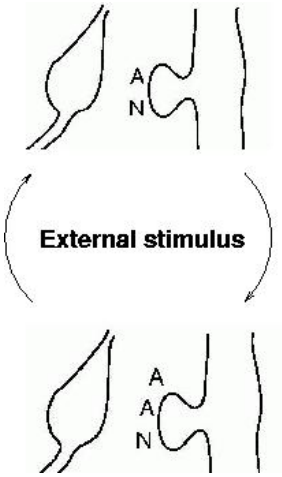




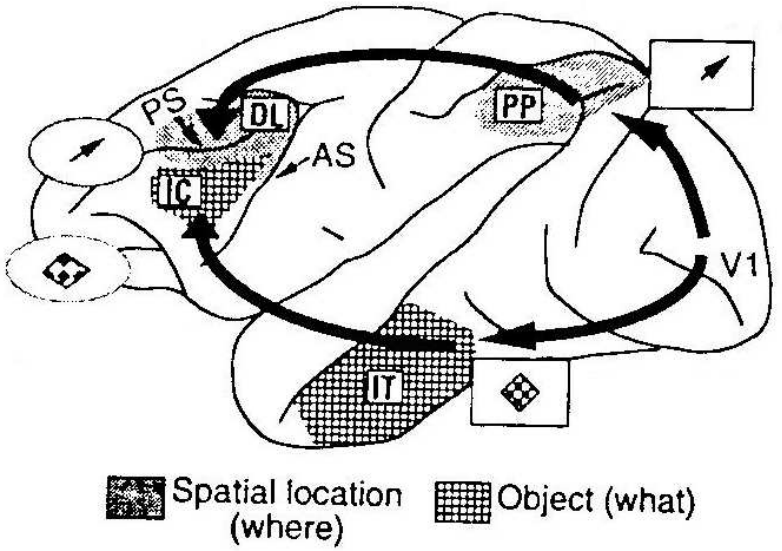
- Long-term memory:

  persistent changes of synapses.





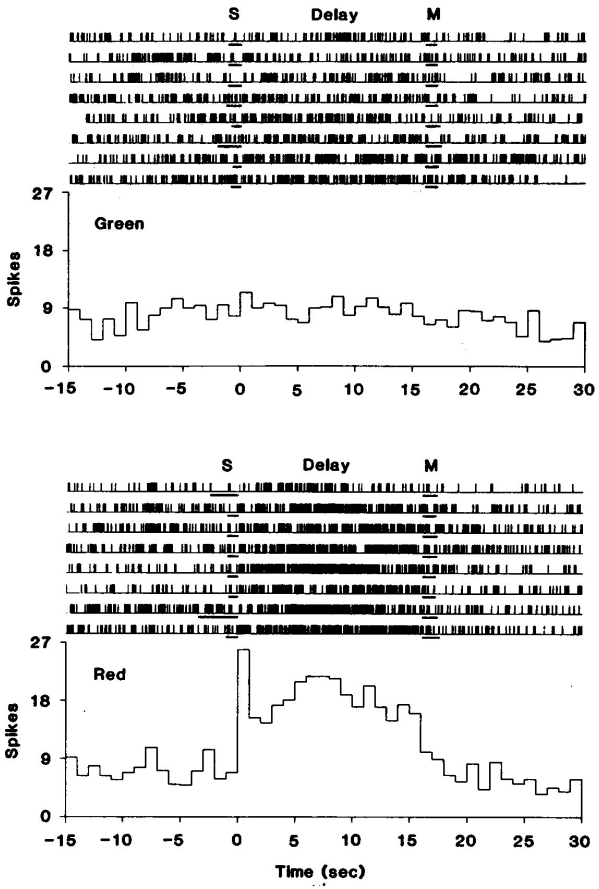
Persistent activity in delayed response tasks



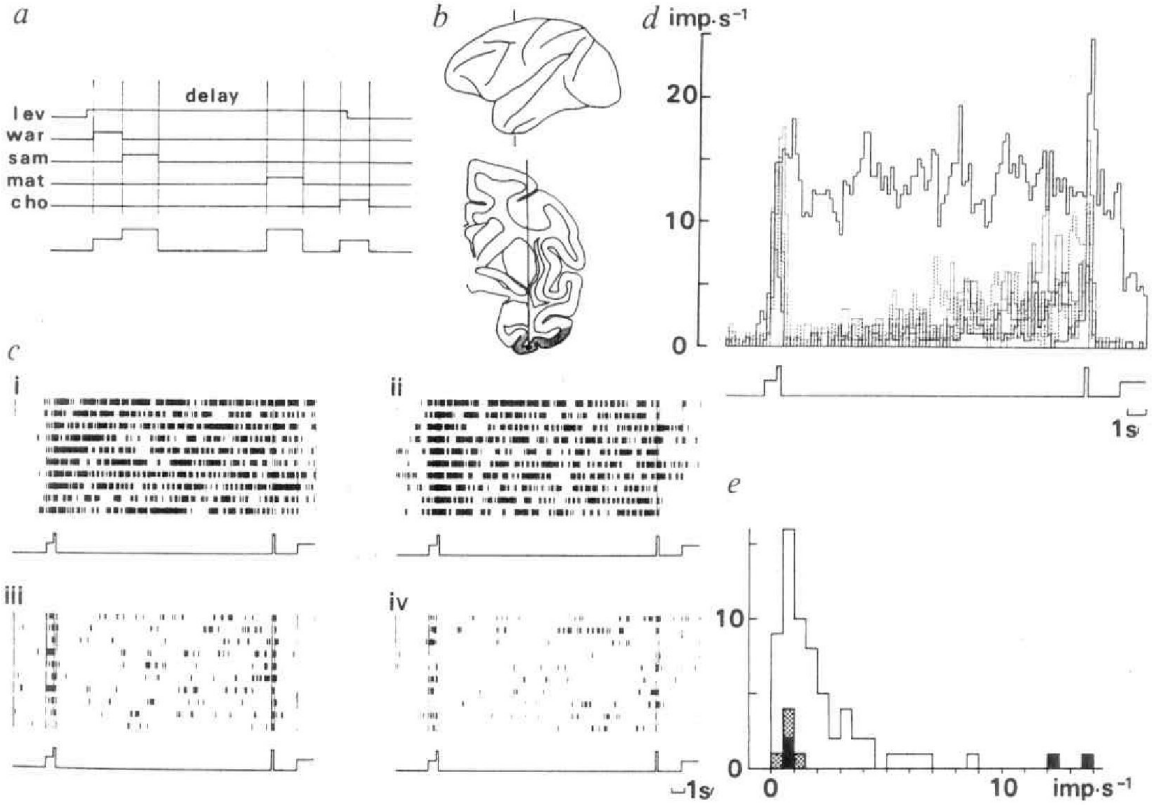
- Ventral stream and PFC:
  identity of stimuli (WHAT?)
- Dorsal stream and PFC:
  spatial location of stimuli (WHERE?)

‘Object’ working memory and persistent activity (IT)

- Fuster and Jervey 1981



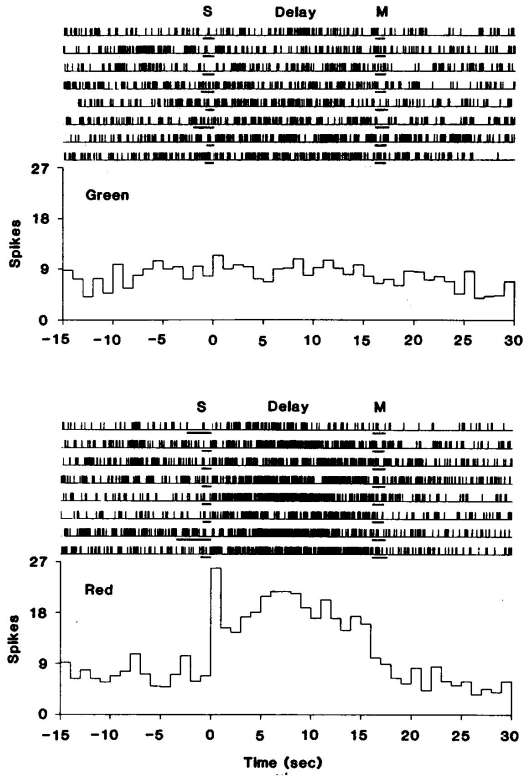
- Miyashita and Chang 1988

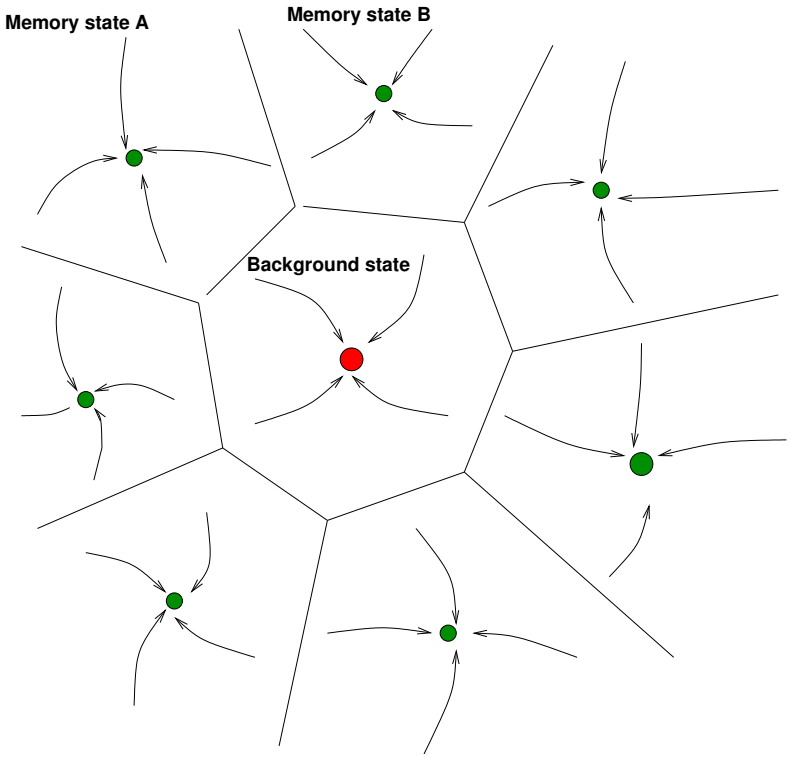


Interpretation in terms of attractor dynamics

Experimental data consistent with a system with

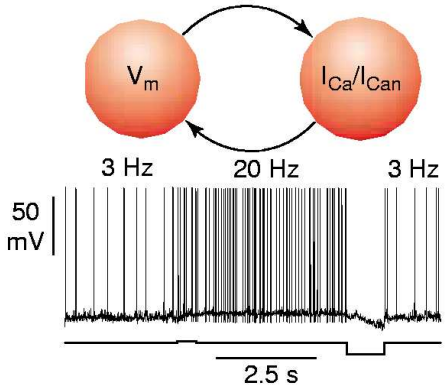
- One ‘background’ network state, with all neurons firing at low rates;
- ‘Memory’ network states, with a small fraction of neurons (specific to each memory state) active at higher rates.





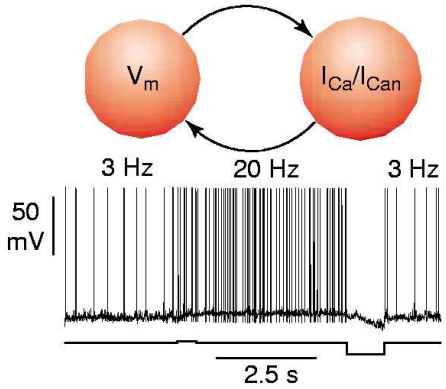
Mechanisms of persistent activity?

1. **Single cell**: persistent activity due to non-linear dynamics of voltage-dependent channels

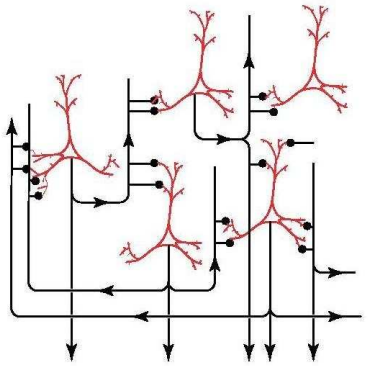


Mechanisms of persistent activity?

1. **Single cell**: persistent activity due to non-linear dynamics of voltage-dependent channels

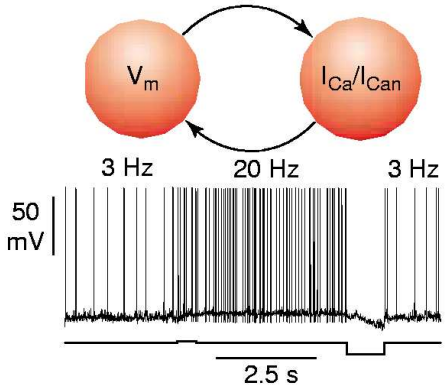


2. **Local network**: persistent activity due to local excitatory connectivity

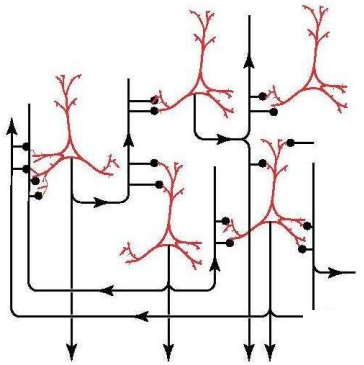


Mechanisms of persistent activity?

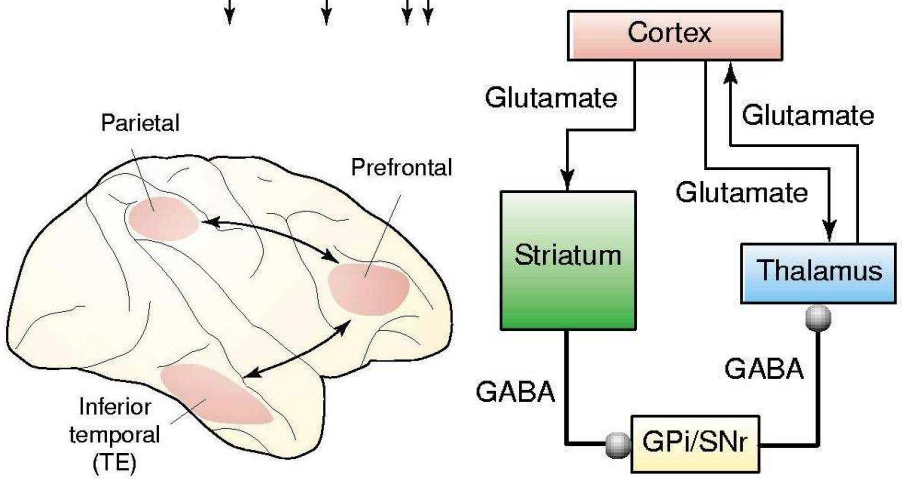
1. **Single cell**: persistent activity due to non-linear dynamics of voltage-dependent channels



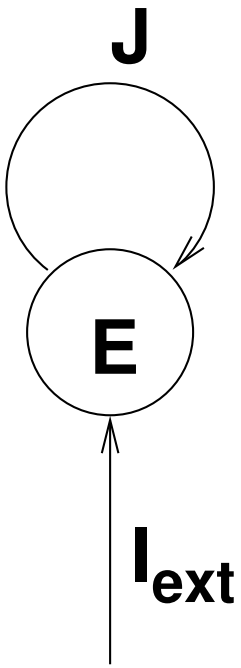
2. **Local network**: persistent activity due to local excitatory connectivity



3. **Systems**: persistent activity due to long-range connections between cortical areas or between cortical and subcortical areas



Local network: Minimal model



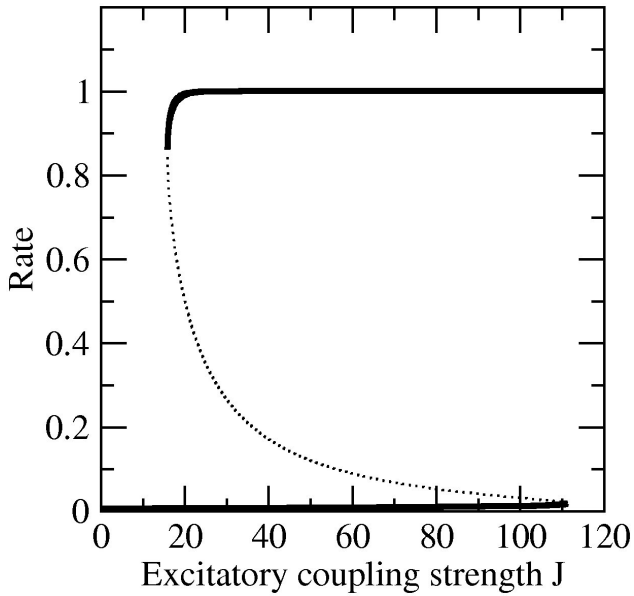
Recurrent excitation $\longrightarrow$ Bistability

$$\tau \frac{dr}{dt} = -r + \Phi(I_{ext} + Jr)$$

where I_{ext} is an external input, and Φ is the transfer function (e.g.sigmoidal).

Provided I_{ext} is low enough:

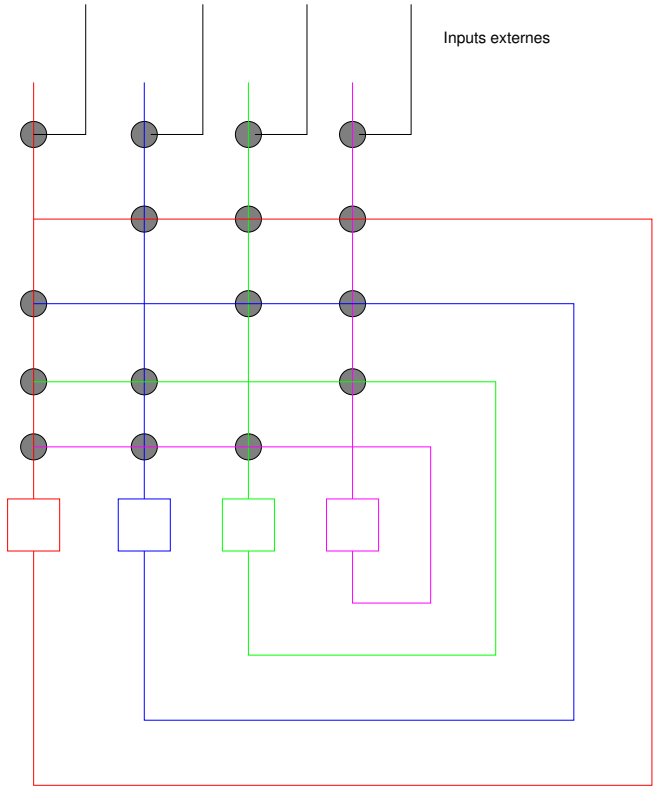
- $J < J_1$: one low activity state;
- $J_1 < J < J_2$: bistability
- $J > J_2$: one high activity state



Networks with many attractors: the Hopfield model

- N binary neurons ($S_i(t) = \pm 1$);
- Neuron dynamics:

$$S_i(t+1) = \text{sign}\left(\sum_j J_{ij} S_j(t)\right)$$



- p 'memory states' ξ_i^μ
- 'Hebbian' synaptic matrix storing memories

$$J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$$

- Energy function

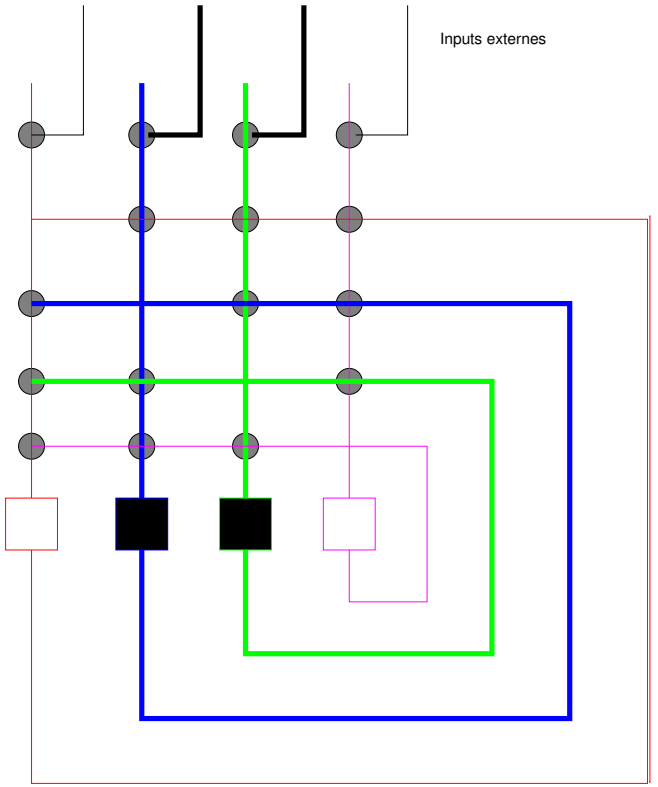
$$E(S) = -\frac{1}{2} \sum_{i,j=1}^{N} J_{ij} S_i S_j$$

- Tools of statistical mechanics apply
- Attractor states close to stored memories if $p < p_{max} \sim N$

Networks with many attractors: the Hopfield model

- N binary neurons ($S_i(t) = \pm 1$);
- Neuron dynamics:

$$S_i(t+1) = \text{sign}\left(\sum_j J_{ij} S_j(t)\right)$$



- p 'memory states' ξ_i^μ
- 'Hebbian' synaptic matrix storing memories

$$J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$$

- Energy function

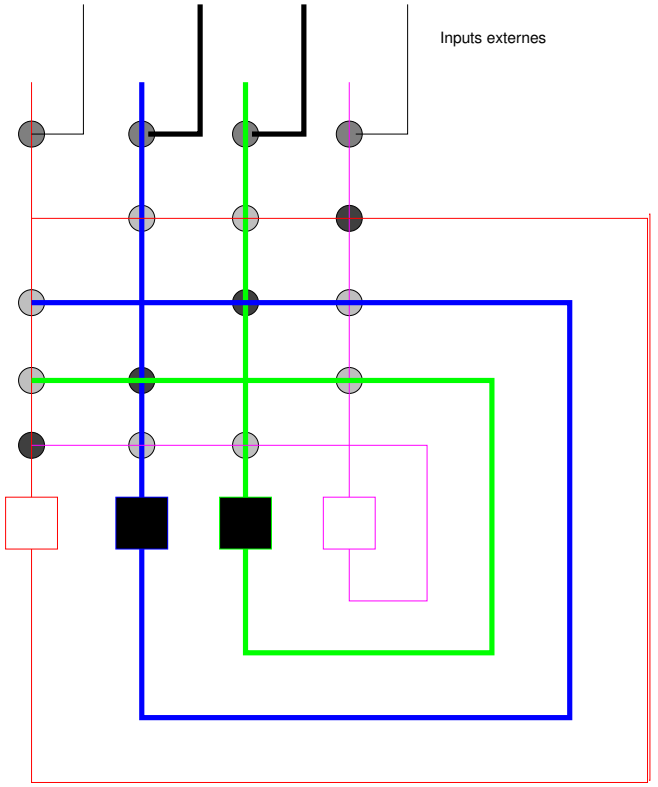
$$E(S) = -\frac{1}{2} \sum_{i,j=1}^{N} J_{ij} S_i S_j$$

- Tools of statistical mechanics apply
- Attractor states close to stored memories if $p < p_{max} \sim N$

Networks with many attractors: the Hopfield model

- N binary neurons ($S_i(t) = \pm 1$);
- Neuron dynamics:

$$S_i(t+1) = \text{sign}\left(\sum_j J_{ij} S_j(t)\right)$$



- p 'memory states' ξ_i^μ
- 'Hebbian' synaptic matrix storing memories

$$J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$$

- Energy function

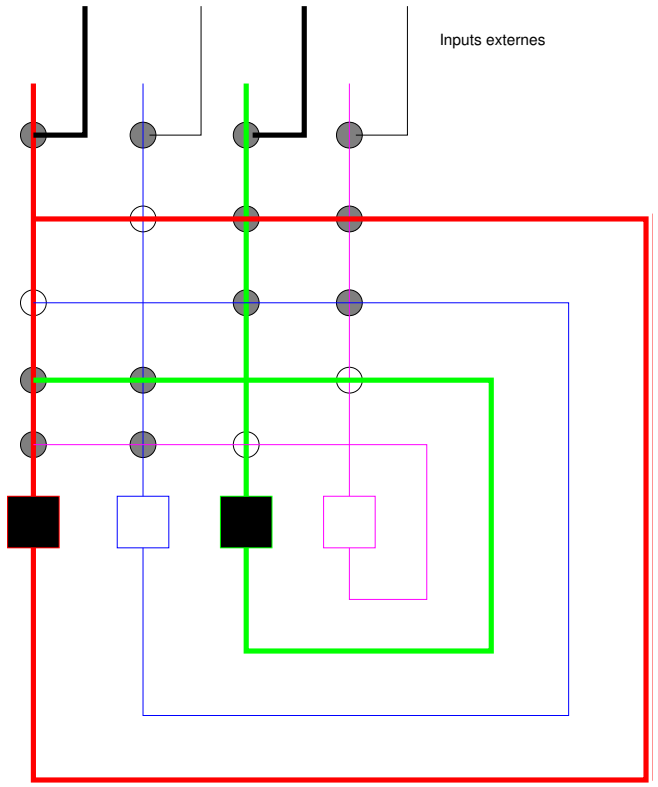
$$E(S) = -\frac{1}{2} \sum_{i,j=1}^{N} J_{ij} S_i S_j$$

- Tools of statistical mechanics apply
- Attractor states close to stored memories if $p < p_{max} \sim N$

Networks with many attractors: the Hopfield model

- N binary neurons ($S_i(t) = \pm 1$);
- Neuron dynamics:

$$S_i(t+1) = \text{sign}\left(\sum_j J_{ij} S_j(t)\right)$$



- p 'memory states' ξ_i^μ
- 'Hebbian' synaptic matrix storing memories

$$J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$$

- Energy function

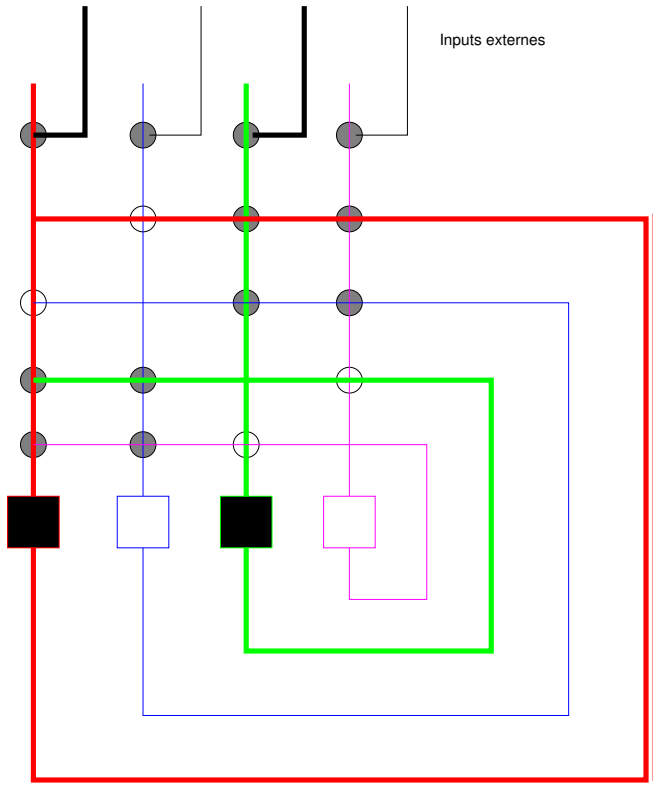
$$E(S) = -\frac{1}{2} \sum_{i,j=1}^{N} J_{ij} S_i S_j$$

- Tools of statistical mechanics apply
- Attractor states close to stored memories if $p < p_{max} \sim N$

Networks with many attractors: the Hopfield model

- N binary neurons ($S_i(t) = \pm 1$);
- Neuron dynamics:

$$S_i(t+1) = \text{sign}\left(\sum_j J_{ij} S_j(t)\right)$$



- p 'memory states' ξ_i^μ
- 'Hebbian' synaptic matrix storing memories

$$J_{ij} = \sum_\mu \xi_i^\mu \xi_j^\mu$$

- Energy function

$$E(S) = -\frac{1}{2} \sum_{i,j=1}^{N} J_{ij} S_i S_j$$

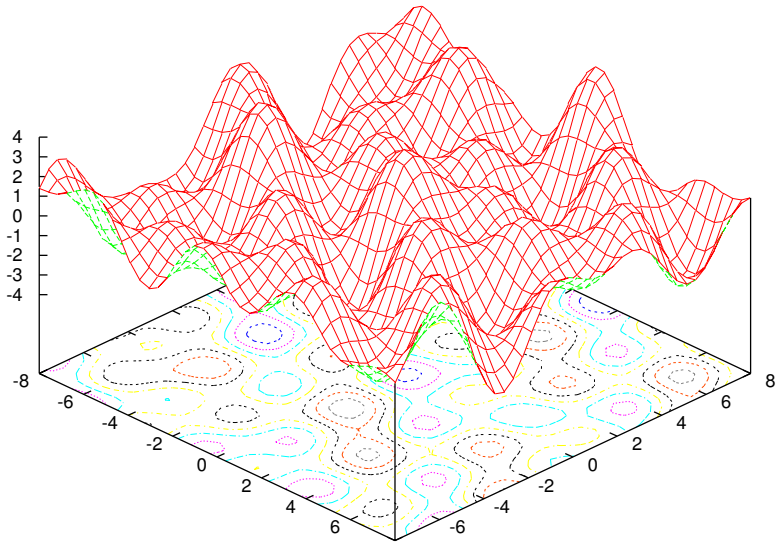
- Tools of statistical mechanics apply
- Attractor states close to stored memories if $p < p_{max} \sim N$

Energy landscape and memories

- Each memory is a network configuration that is an attractor of network dynamics

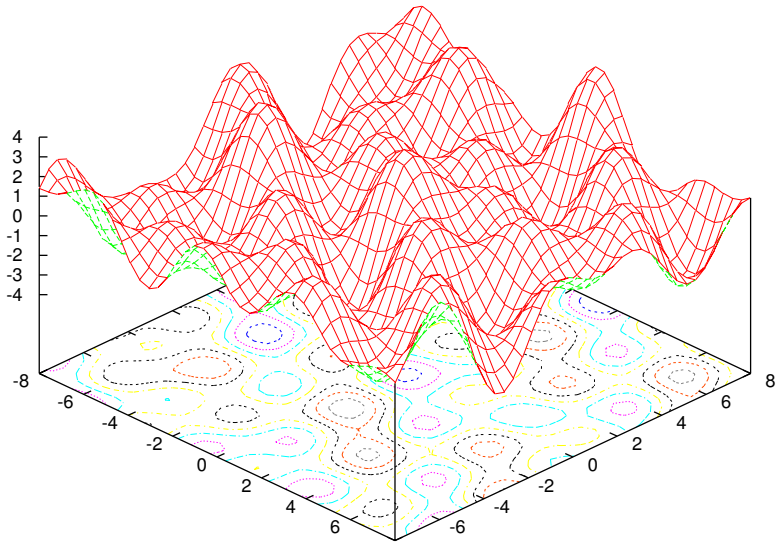
Energy landscape and memories

- Each memory is a network configuration that is an attractor of network dynamics



Energy landscape and memories

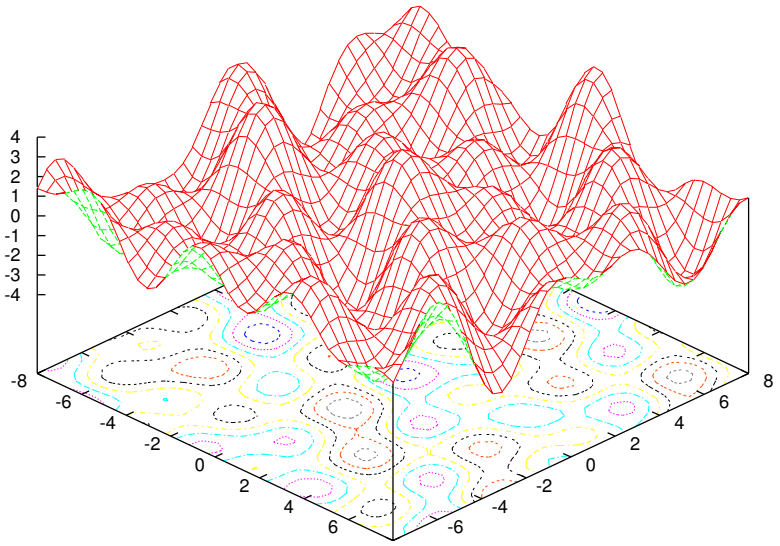
- Each memory is a network configuration that is an attractor of network dynamics



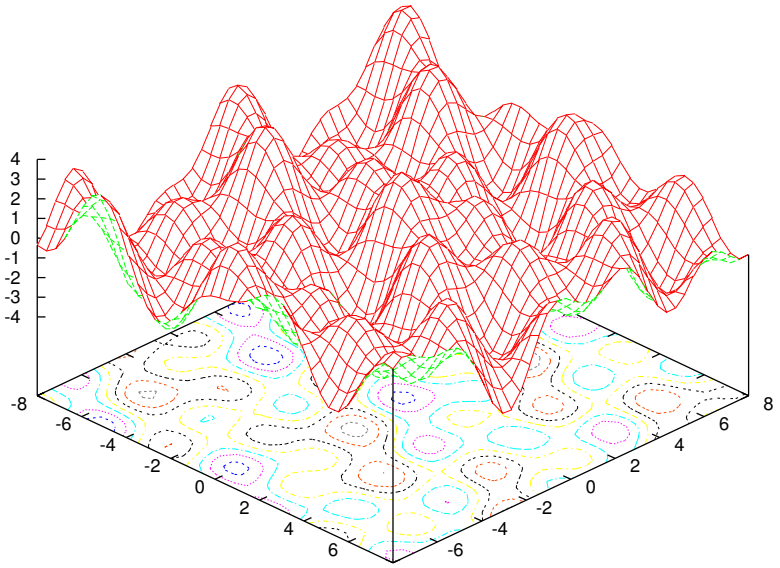
- Changes in synaptic efficacies due to learning lead to modifications of these attractors (creation, movement, destruction)

Energy landscape and memories

- Each memory is a network configuration that is an attractor of network dynamics

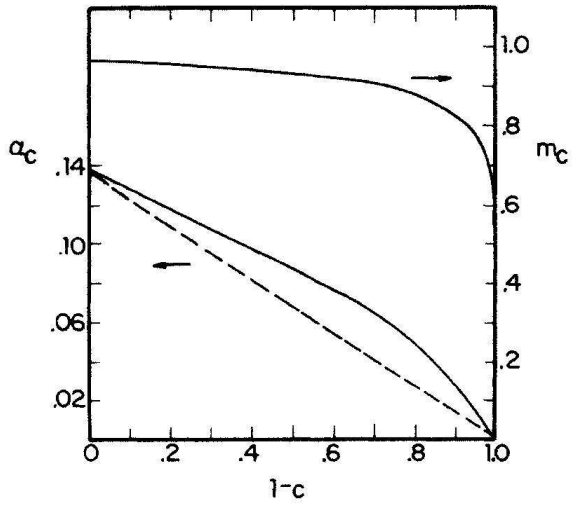


- Changes in synaptic efficacies due to learning lead to modifications of these attractors (creation, movement, destruction)



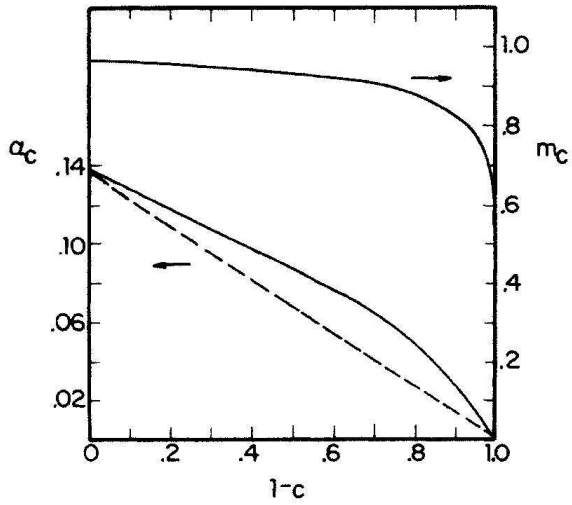
Summary of properties of Hopfield and related models

- Hopfield model (1982 - $\pm$ 1 neurons, dense coding, analog synapses)
  - Capacity (max number of memories) $\sim 0.14N$
    (Amit et al 1985)
  - Trade-off between number of attractors and size of attractor basins
  - Very robust to random dilution
    (capacity of order $C =$ number of synapses per neuron)
    (Sompolinsky 1986, Derrida et al 1987)



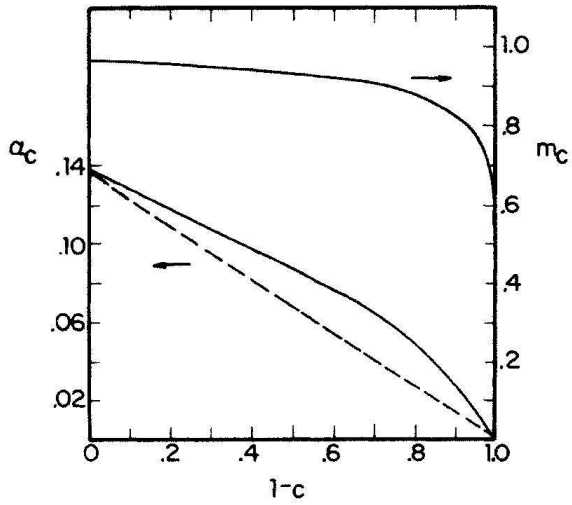
Summary of properties of Hopfield and related models

- Hopfield model (1982 - ± 1 neurons, dense coding, analog synapses)
  - Capacity (max number of memories) $\sim 0.14N$
    (Amit et al 1985)
  - Trade-off between number of attractors and size of attractor basins
  - Very robust to random dilution
    (capacity of order $C =$ number of synapses per neuron)
    (Sompolinsky 1986, Derrida et al 1987)
- Tsodyks-Feigelman model (1988 - 0,1 neurons, arbitrary coding level $f =$ Prob$(\xi_i^\mu = 1)$, analog synapses)
  - Capacity (max number of memories) $\sim C/f\ln(f)$;
  - Quiescent state ('no recognition') stable.



Summary of properties of Hopfield and related models

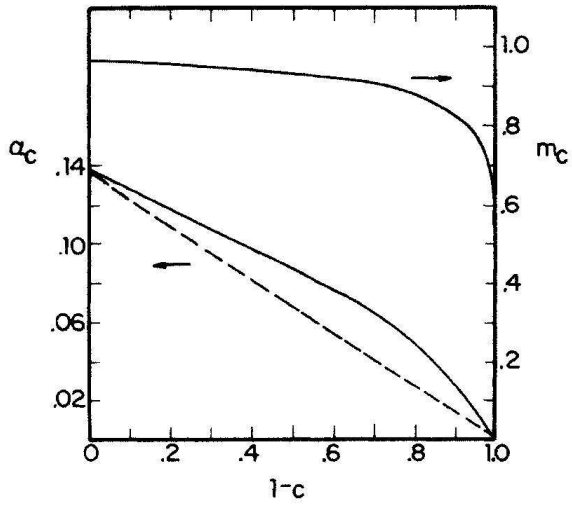
- Hopfield model (1982 - $\pm$ 1 neurons, dense coding, analog synapses)
  - Capacity (max number of memories) $\sim 0.14N$
    (Amit et al 1985)
  - Trade-off between number of attractors and size of attractor basins
  - Very robust to random dilution
    (capacity of order $C =$ number of synapses per neuron)
    (Sompolinsky 1986, Derrida et al 1987)



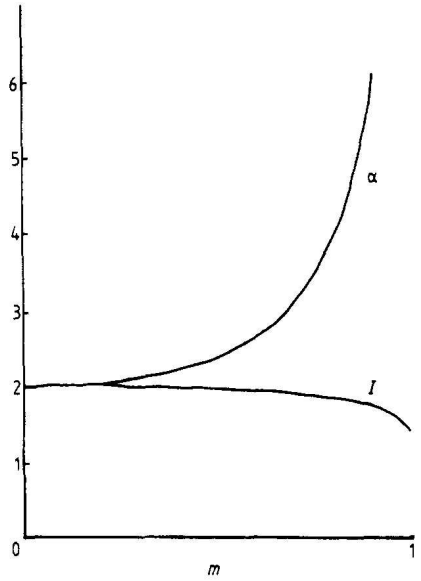
- Tsodyks-Feigelman model (1988 - 0,1 neurons, arbitrary coding level $f =$ Prob$(\xi_i^\mu = 1)$, analog synapses)
  - Capacity (max number of memories) $\sim C/f\ln(f)$;
  - Quiescent state ('no recognition') stable.
- Willshaw model (1969 - 0,1 neurons, sparse coding, discrete synapses)
  - Works well only for sparse coding, $f \sim \ln N/N$ where capacity is close to optimal!

Summary of properties of Hopfield and related models

- Hopfield model (1982 - $\pm$ 1 neurons, dense coding, analog synapses)
  - Capacity (max number of memories) $\sim 0.14N$
    (Amit et al 1985)
  - Trade-off between number of attractors and size of attractor basins
  - Very robust to random dilution
    (capacity of order $C =$ number of synapses per neuron)
    (Sompolinsky 1986, Derrida et al 1987)



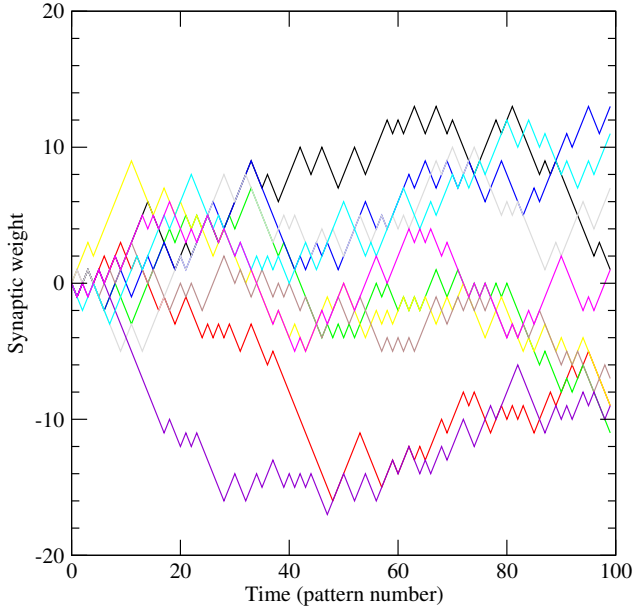
- Tsodyks-Feigelman model (1988 - 0,1 neurons, arbitrary coding level $f =$ Prob$(\xi_i^\mu = 1)$, analog synapses)
  - Capacity (max number of memories) $\sim C/f\ln(f)$;
  - Quiescent state ('no recognition') stable.



- Willshaw model (1969 - 0,1 neurons, sparse coding, discrete synapses)
  - Works well only for sparse coding, $f \sim \ln N/N$ where capacity is close to optimal!
- Theoretical capacity limit (max over all possible matrices J_{ij}):
  $2C$ memories (dense coding), $\sim C/f\ln(f)$ memories (sparse coding)
  (Gardner 1988)

Learning

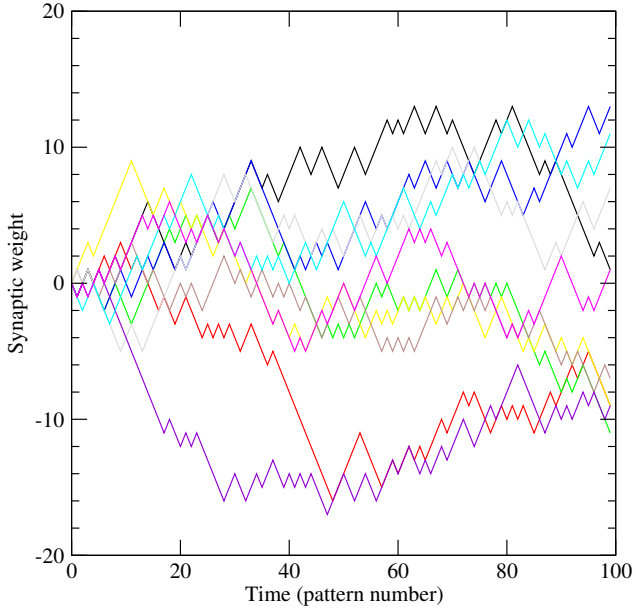
- In the presence of a continuous stream of incoming stimuli: problem of memory black-out in Hopfield-type models



Learning

- In the presence of a continuous stream of incoming stimuli: problem of memory black-out in Hopfield-type models

- ‘Palimpsest’ models: old patterns are progressively erased by more recently seen patterns.

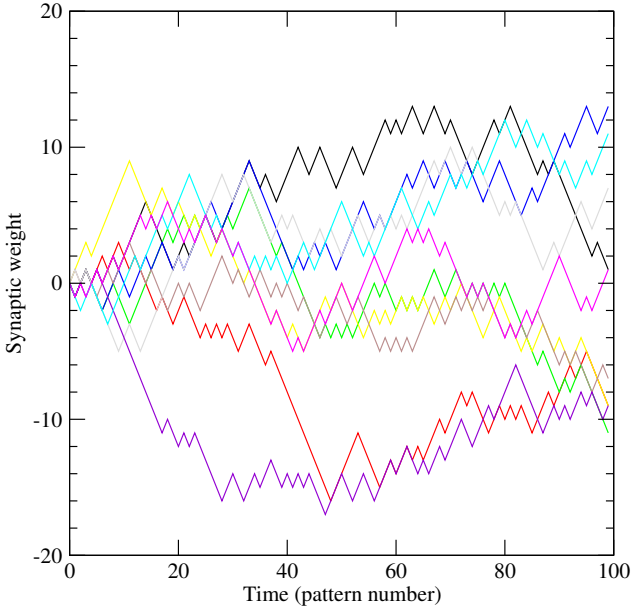


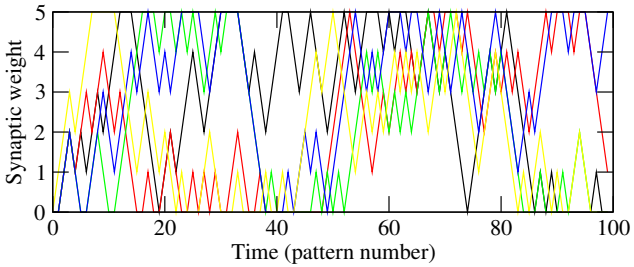
Learning

- In the presence of a continuous stream of incoming stimuli: problem of memory black-out in Hopfield-type models

- ‘Palimpsest’ models: old patterns are progressively erased by more recently seen patterns.

- Models with analog synapses
  - Add bounds to synaptic weights (Parisi 1986)
  - Exponential decay of old memories (Mézard et al 1986)





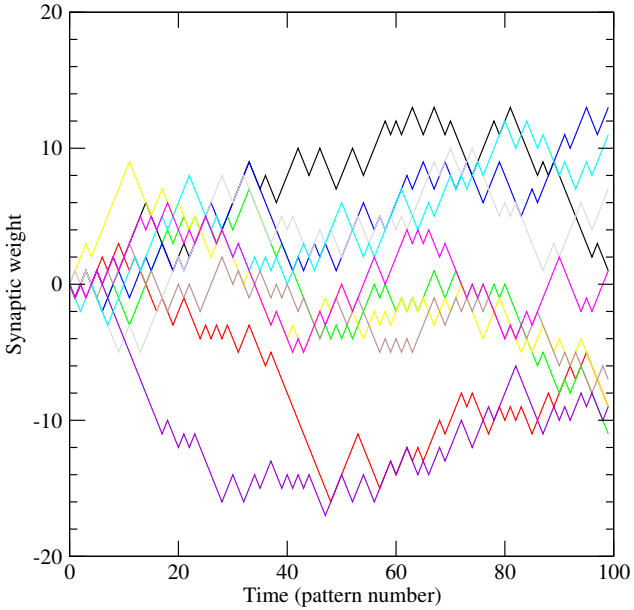
Learning

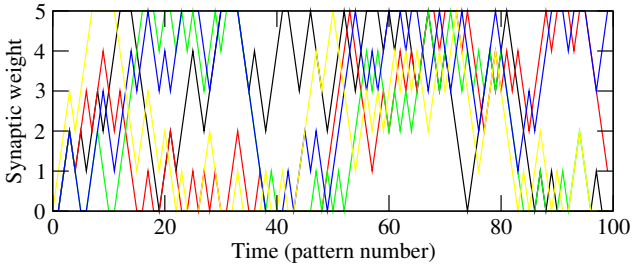
- In the presence of a continuous stream of incoming stimuli: problem of memory black-out in Hopfield-type models

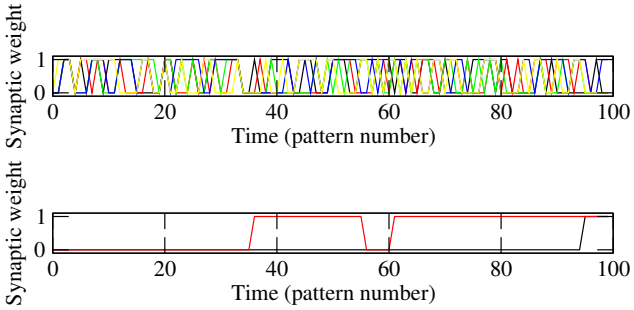
- ‘Palimpsest’ models: old patterns are progressively erased by more recently seen patterns.

- Models with analog synapses
  - Add bounds to synaptic weights (Parisi 1986)
  - Exponential decay of old memories (Mézard et al 1986)
- Models with binary synapses (low/high efficacy states), and stochastic transitions between states (Amit and Fusi 1994, Fusi et al 2005).

  Very poor performance, unless:
  - Balance between LTP and LTD-like transitions, AND sparse coding;
  - Hidden states are added (e.g. cascade model)



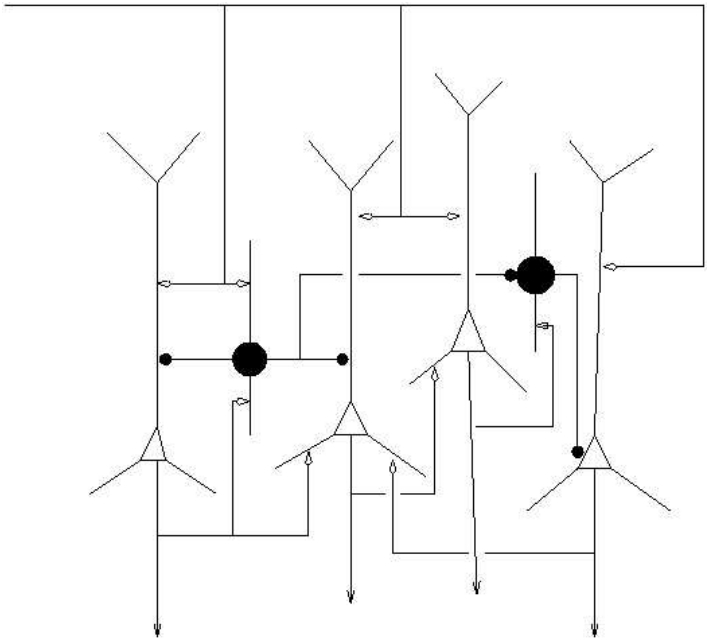




What do we learn from networks of binary neurons?

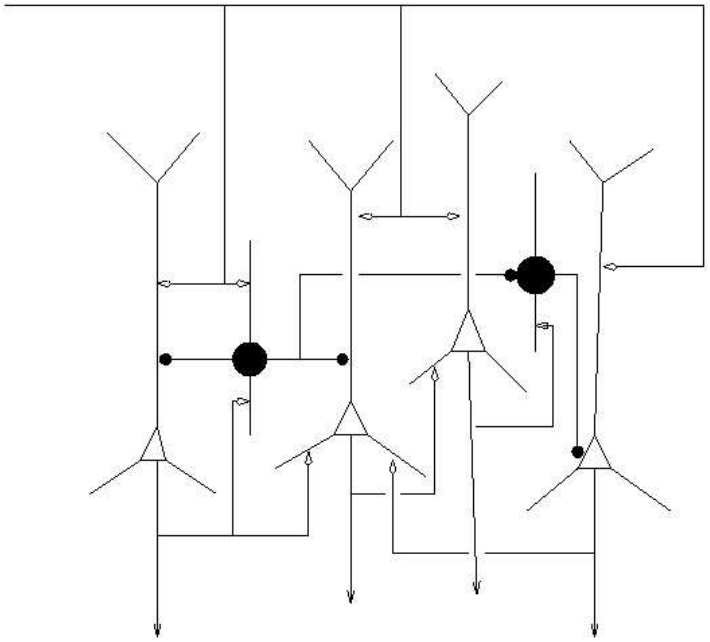
- A network can work as an associative memory in a robust way and quasi-optimal manner (stored information of order 1 bit per synapse)
  - with a diluted binary synaptic matrix;
  - and stochastic learning;
- But only when some conditions are fulfilled
  - sparse coding;
  - balance between ‘LTP’ and ‘LTD’
- These models are too simple to be compared with experiments;
- $\Rightarrow$ More realistic networks (network of spiking neurons)

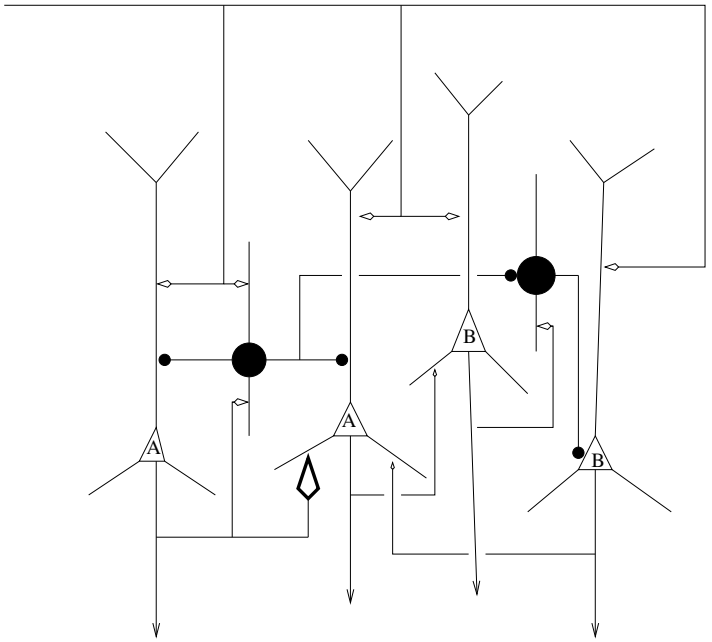
Local cortical network model



- Local network ($\sim$ 1mm^3, 10^5neurons), 80% exc, 20% inh;
- Connection probability $\sim$ 10%;
- Neurons: integrate-and-fire neurons;

Local cortical network model

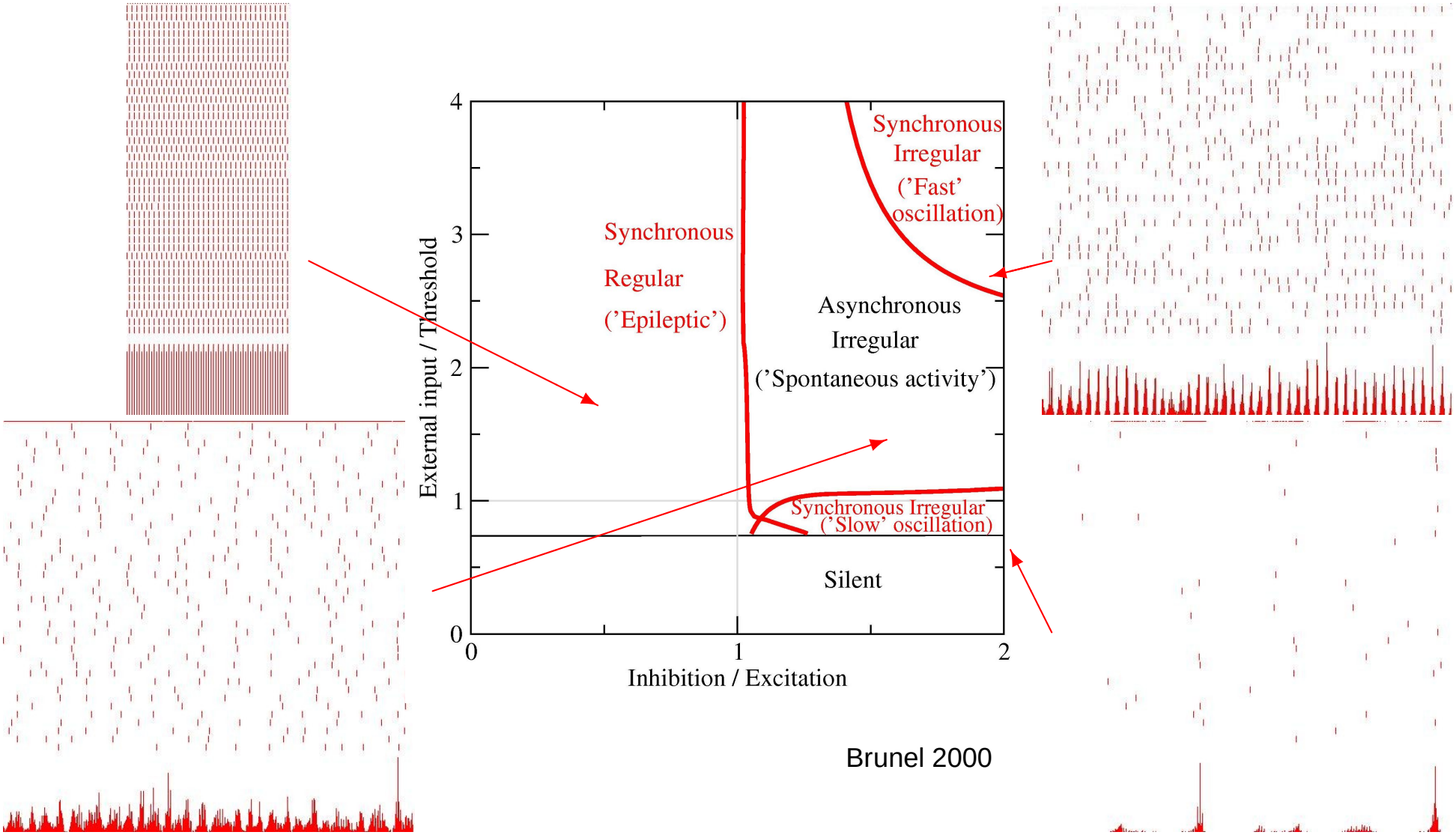




- Local network ($\sim$ 1mm^3, 10^5neurons), 80% exc, 20% inh;
- Connection probability $\sim$ 10%;
- Neurons: integrate-and-fire neurons;
- Each stimulus activates a small fraction of cells ($\sim$ 1%)
- Both potentiation and depression of synapses by Hebbian mechanisms (by a factor $\sim$ 2)

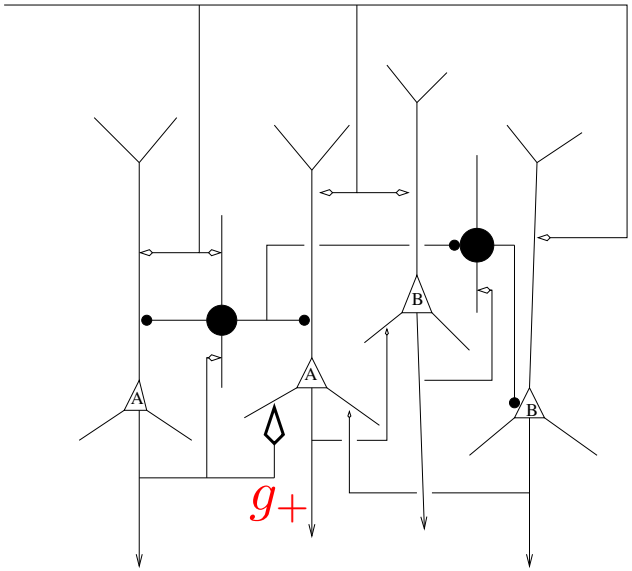
Amit and Brunel 1997

Phase diagram of unstructured network

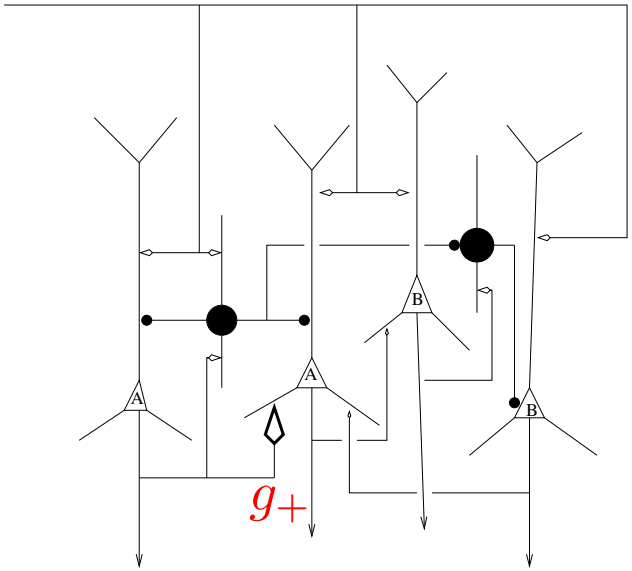


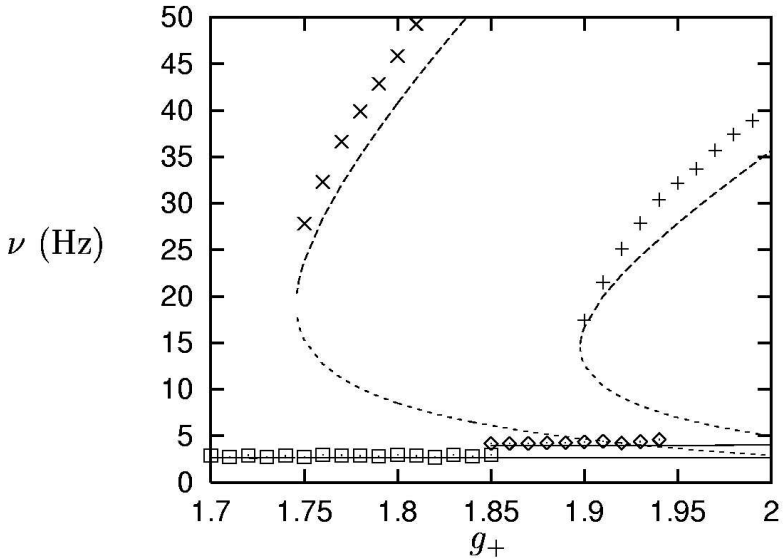
Brunel 2000

Emergence of persistent activity following learning

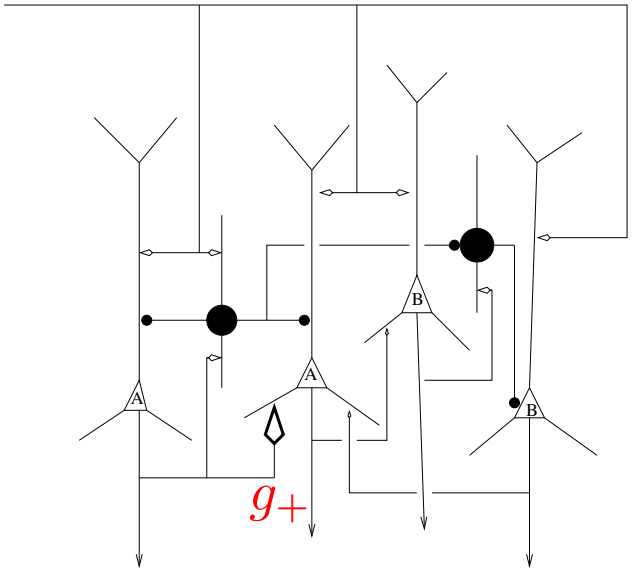


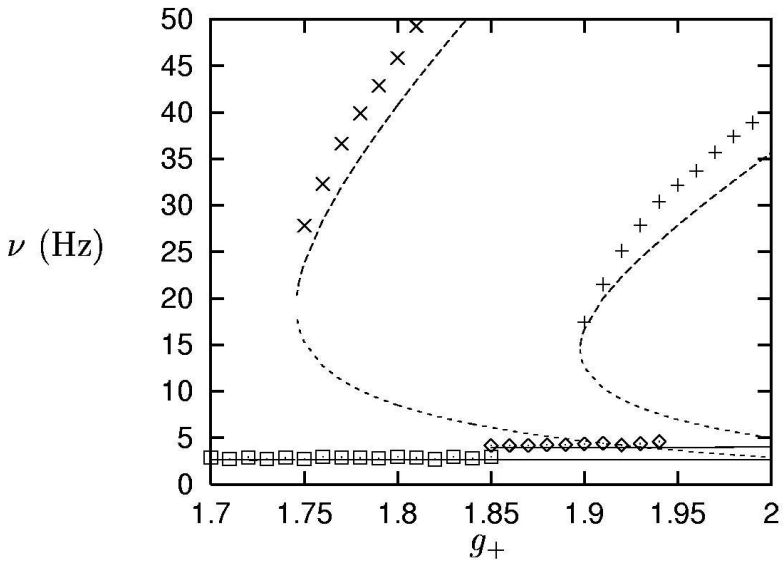
Emergence of persistent activity following learning

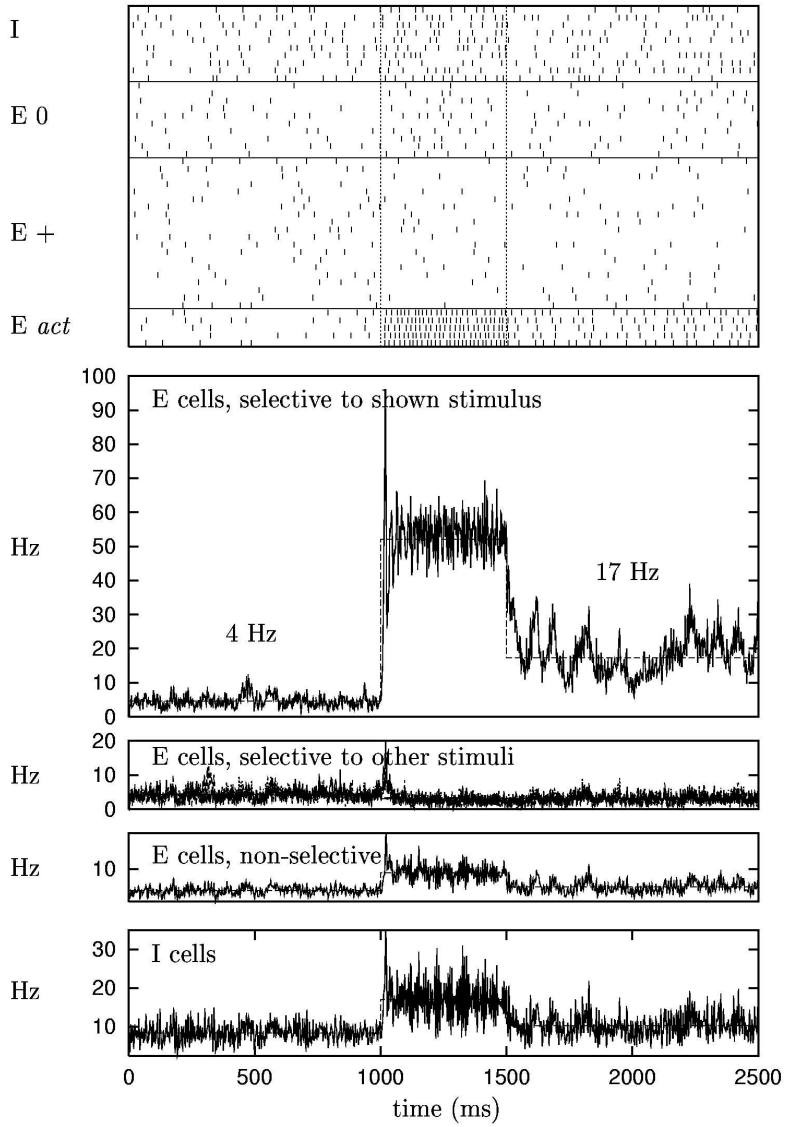




Emergence of persistent activity following learning

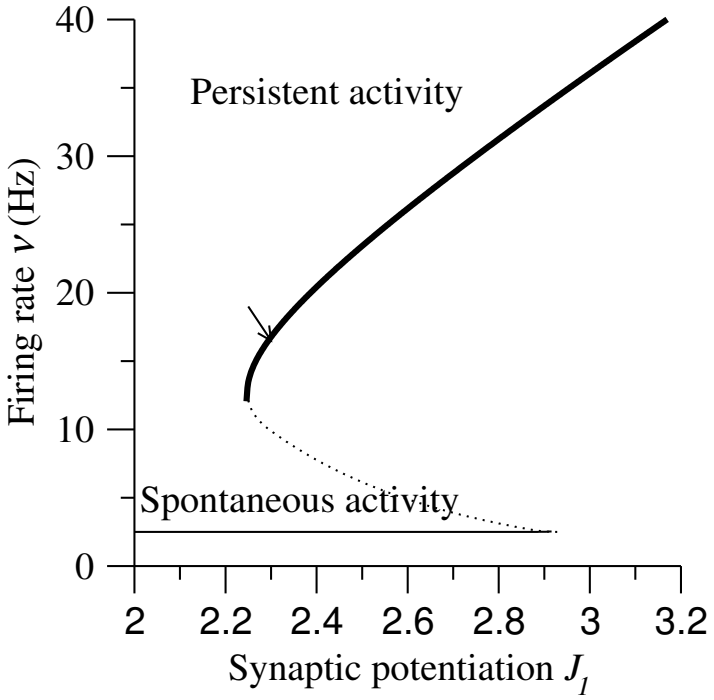




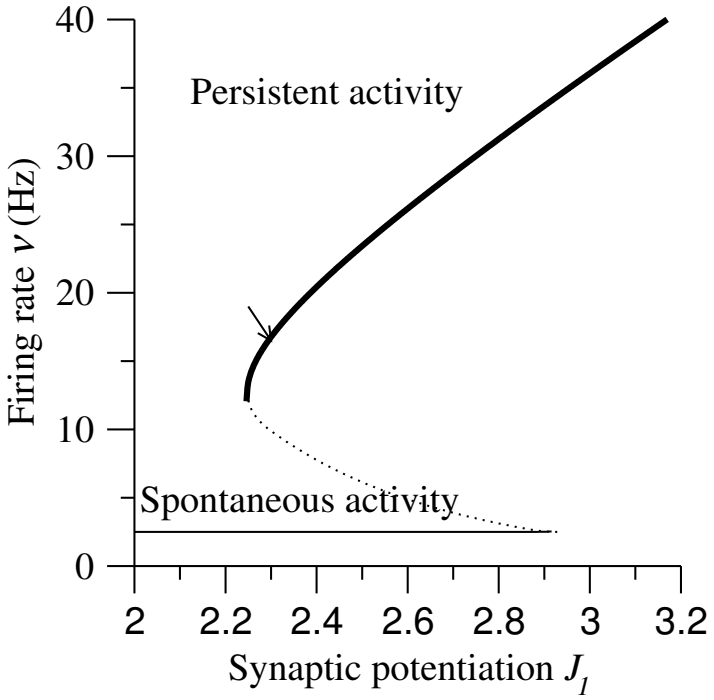


Brunel 2000

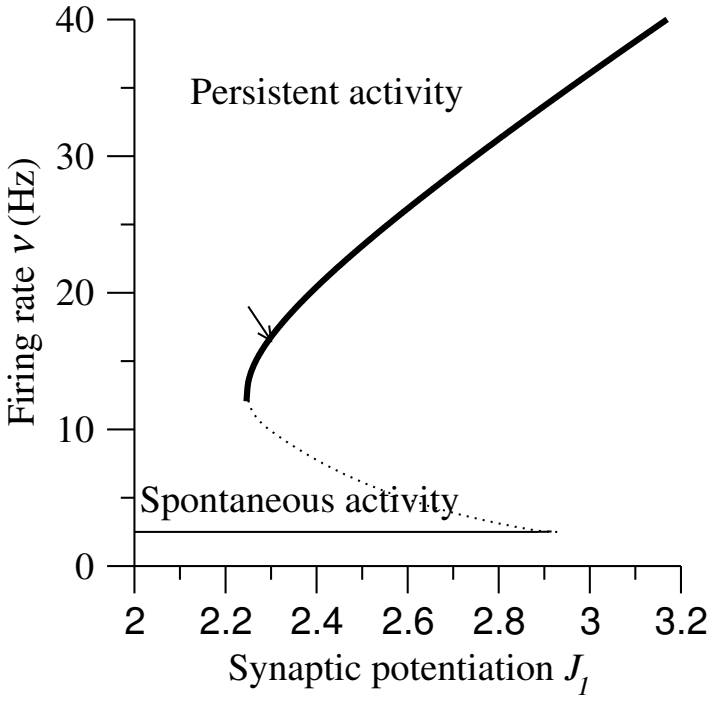
Switching the network back to the spontaneous state

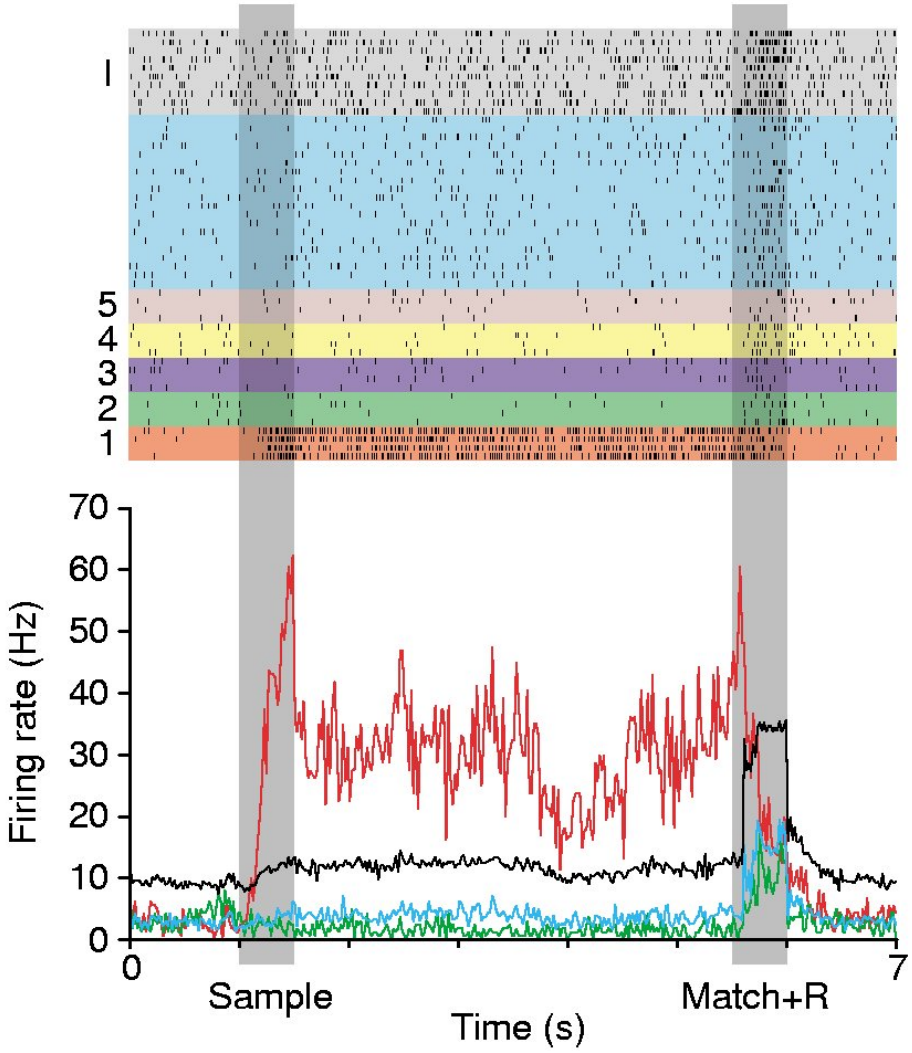


Switching the network back to the spontaneous state

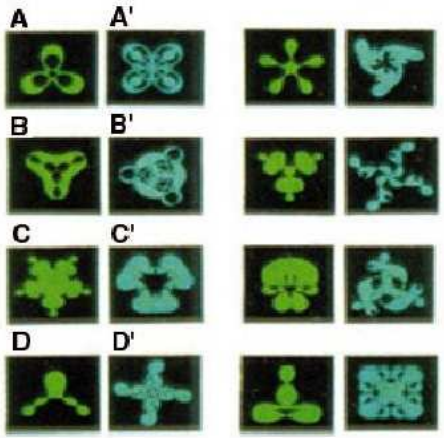


Switching the network back to the spontaneous state

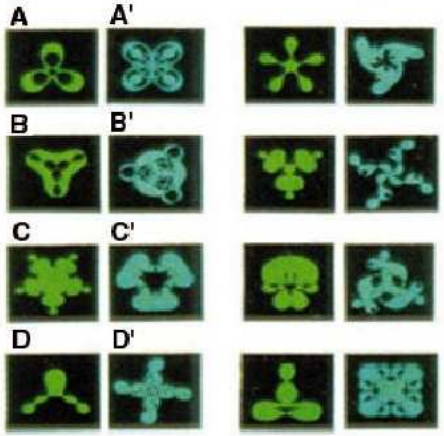


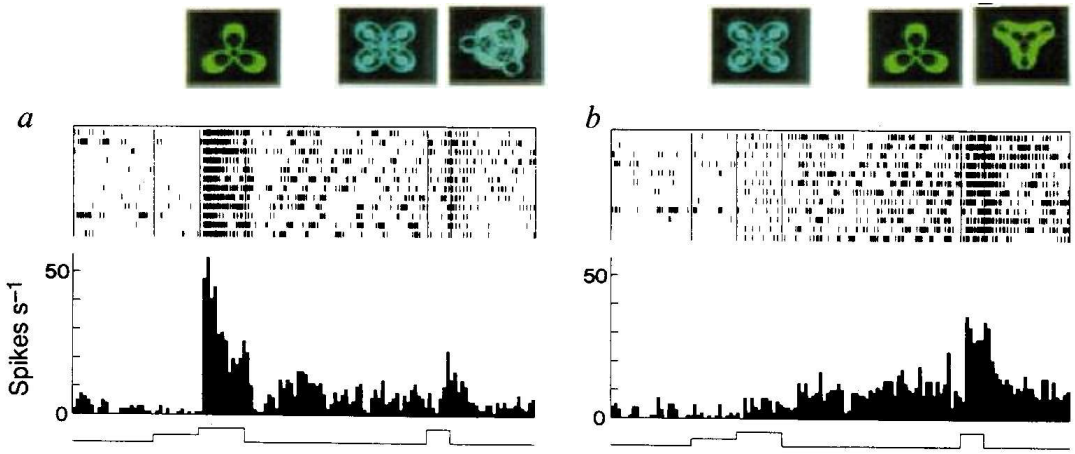


Pair-association experiments and prospective activity

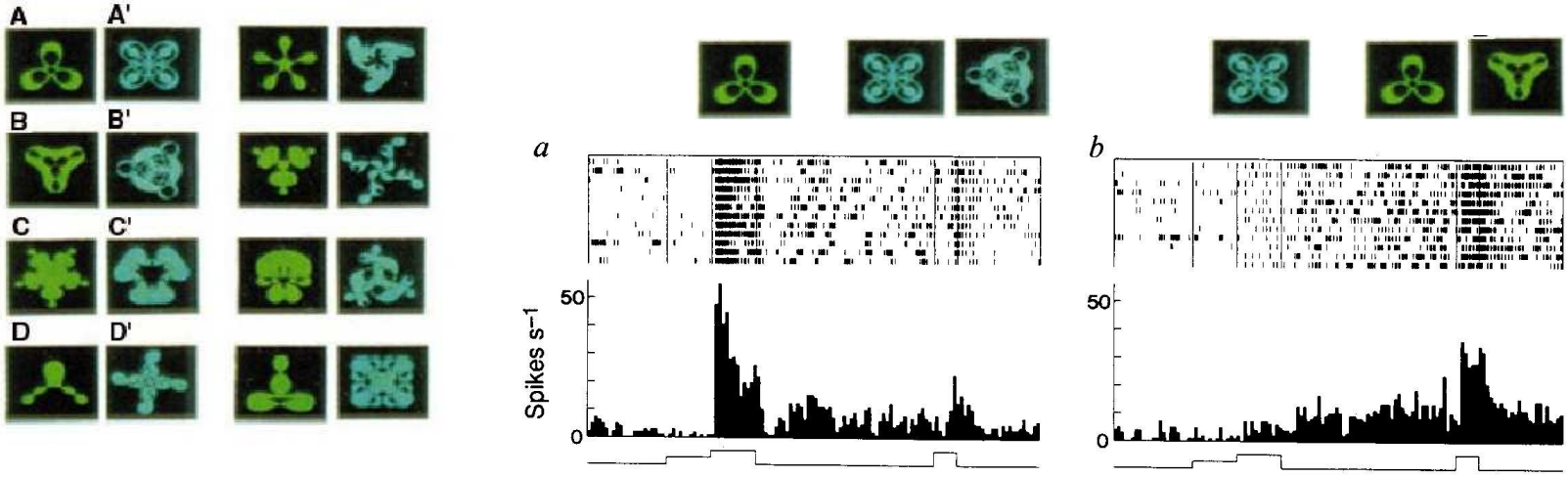


Pair-association experiments and prospective activity





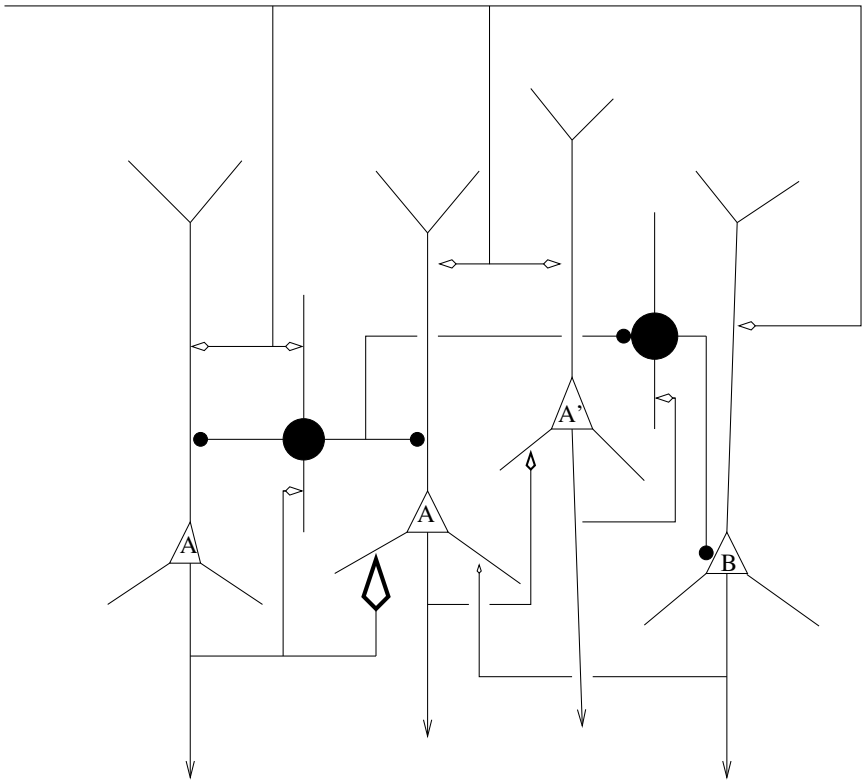
Pair-association experiments and prospective activity



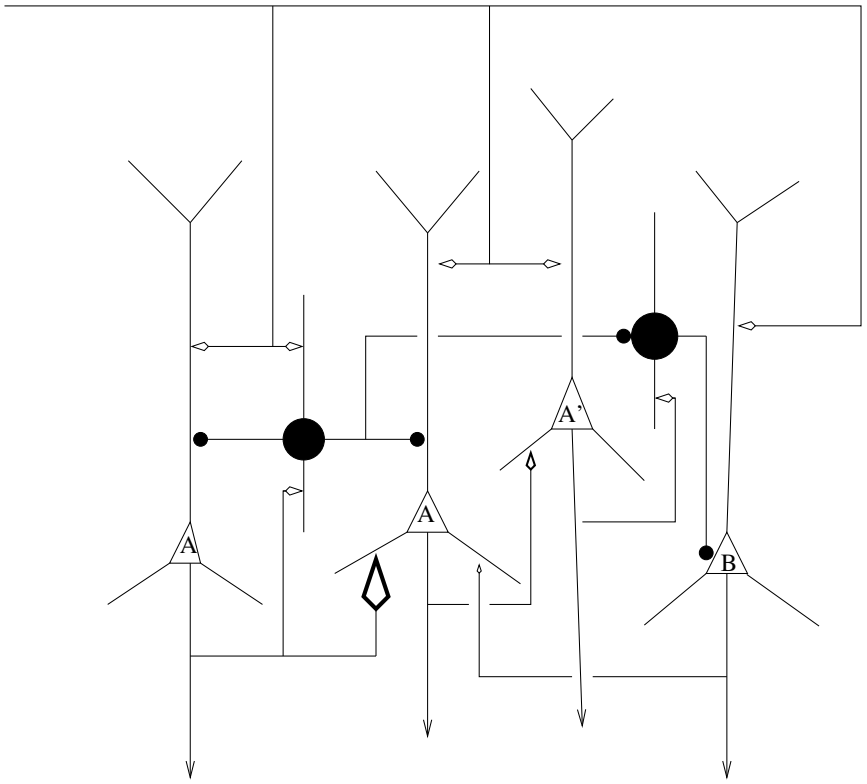
Sakai and Miyashita 1991; Naya et al 1996, 2001, 2003

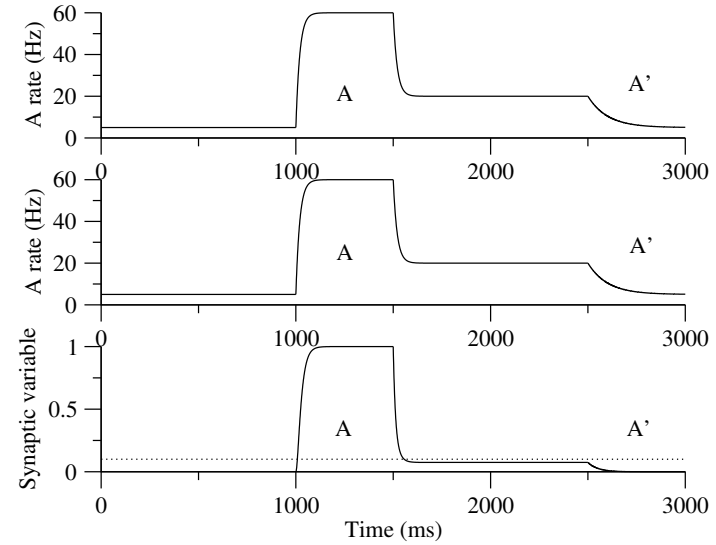
Erickson and Desimone 1999 (perirhinal cortex); Rainer and Miller (prefrontal cortex).

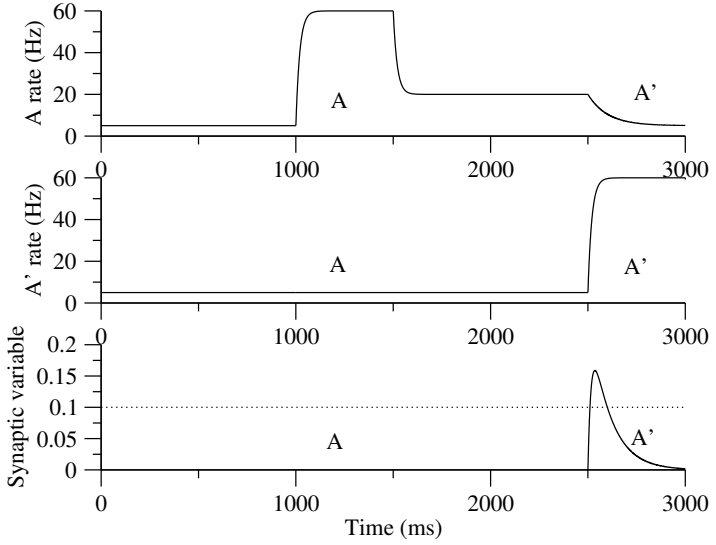
How synaptic matrix is structured during the pair-association task



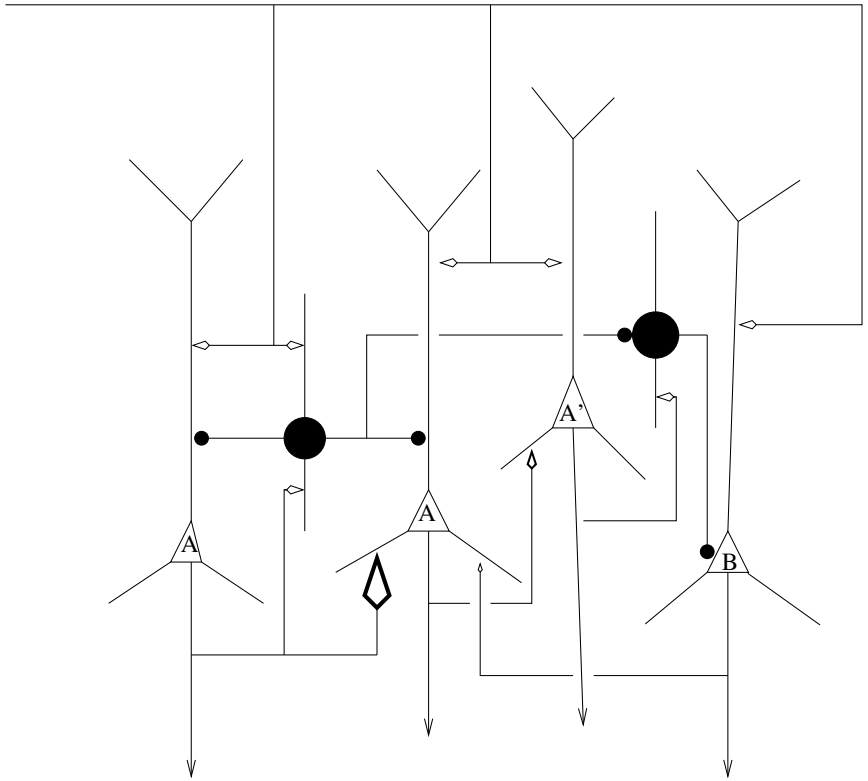
How synaptic matrix is structured during the pair-association task



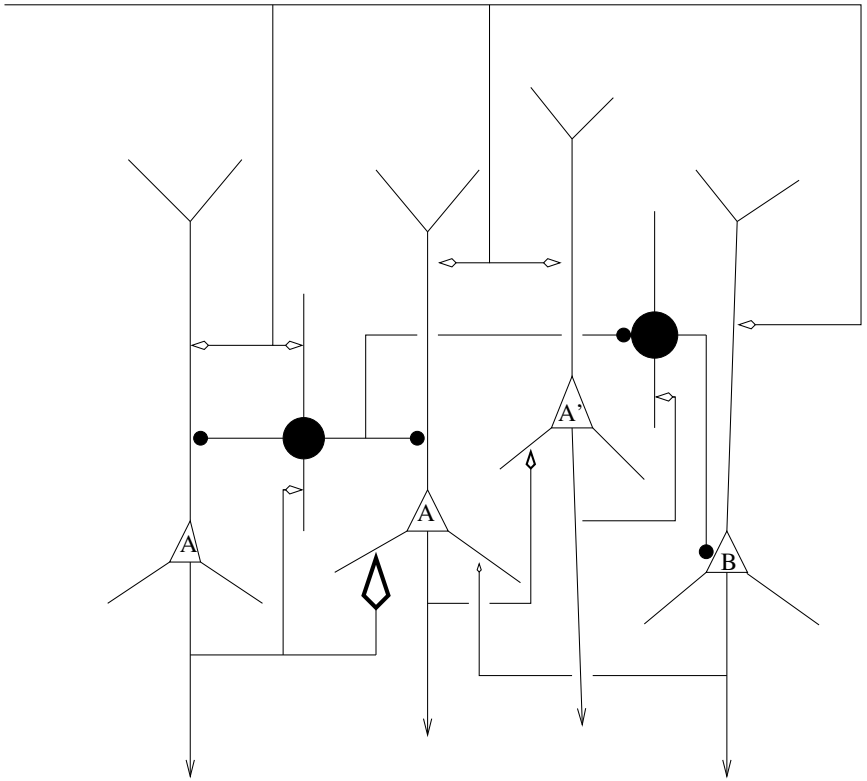


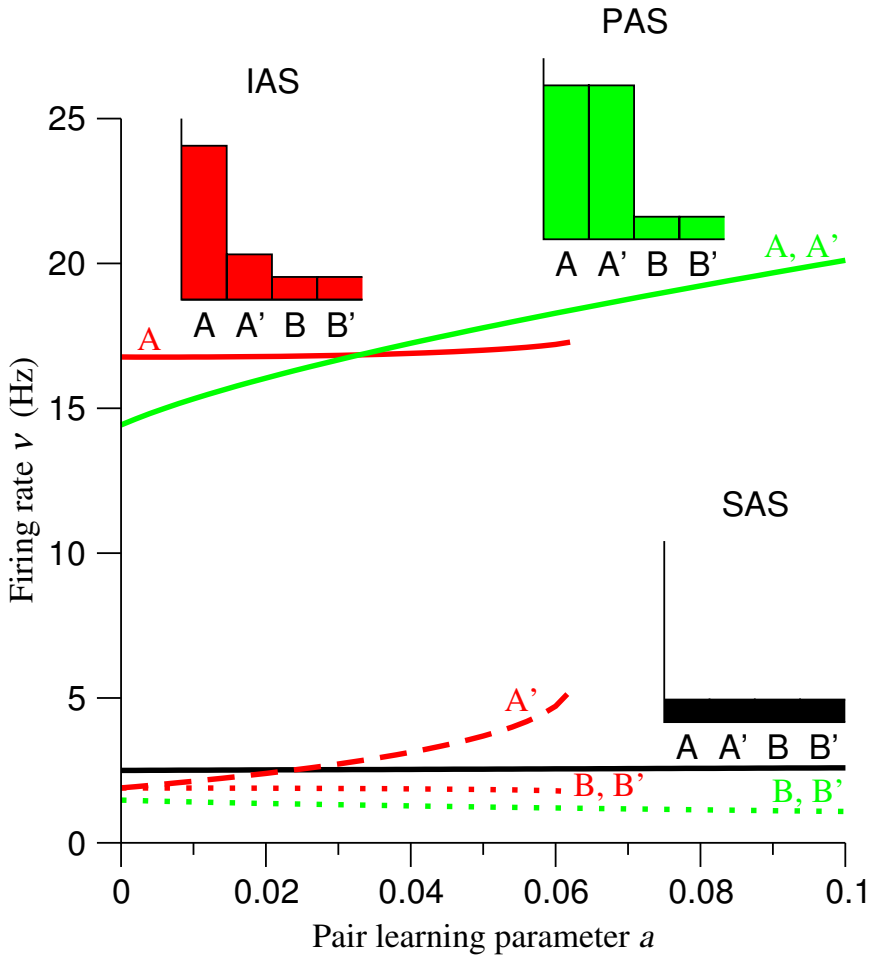


Network states after pair-association learning



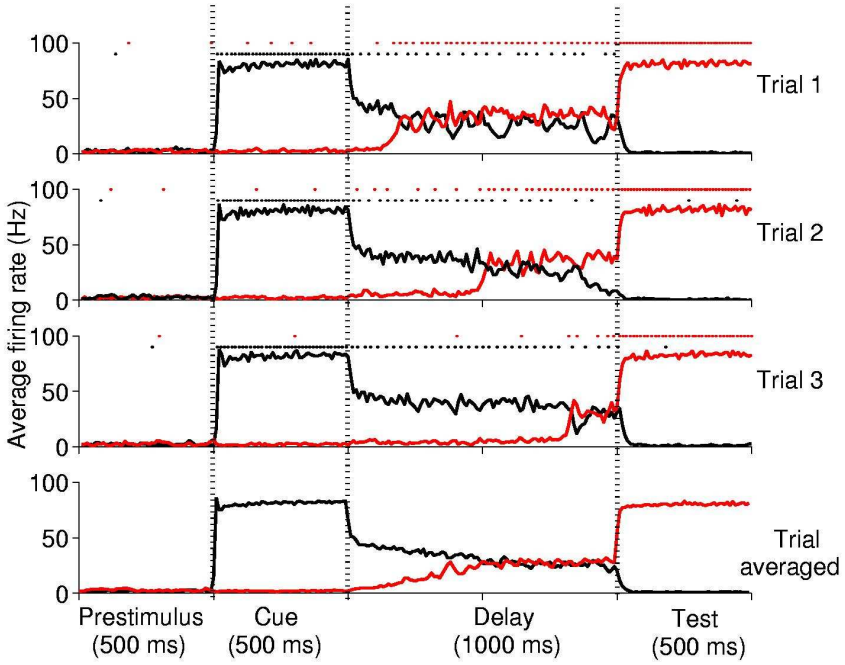
Network states after pair-association learning



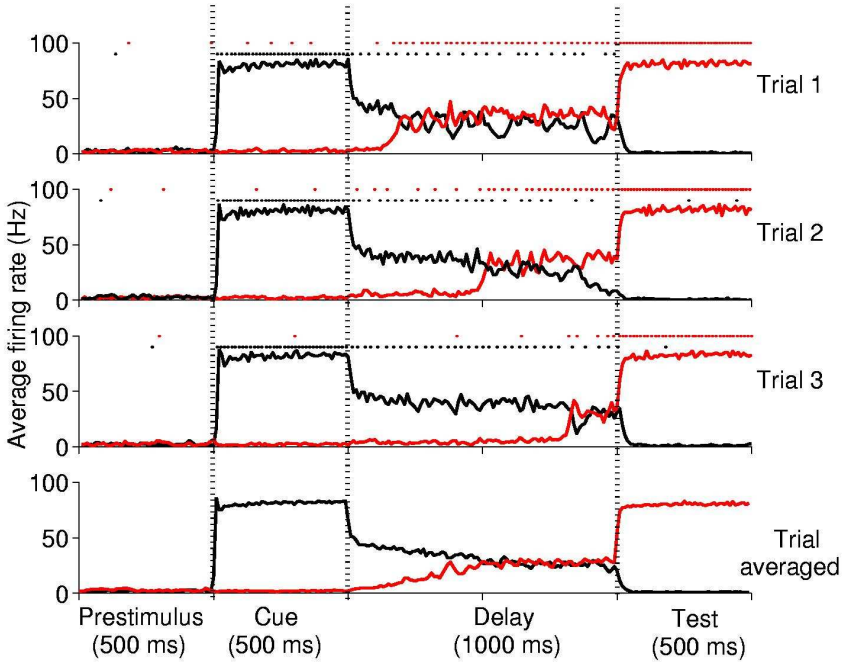


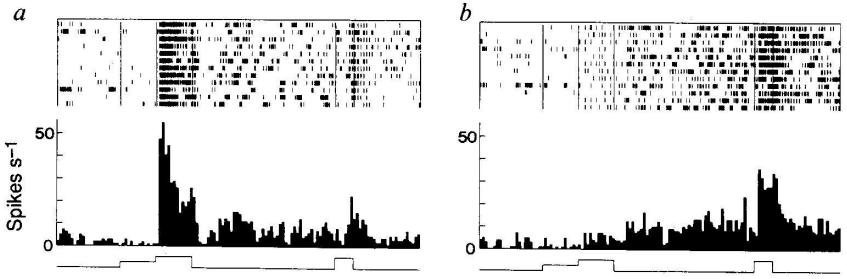
Mongillo et al 2003

Transitions between states during delay period and prospective activity



Transitions between states during delay period and prospective activity

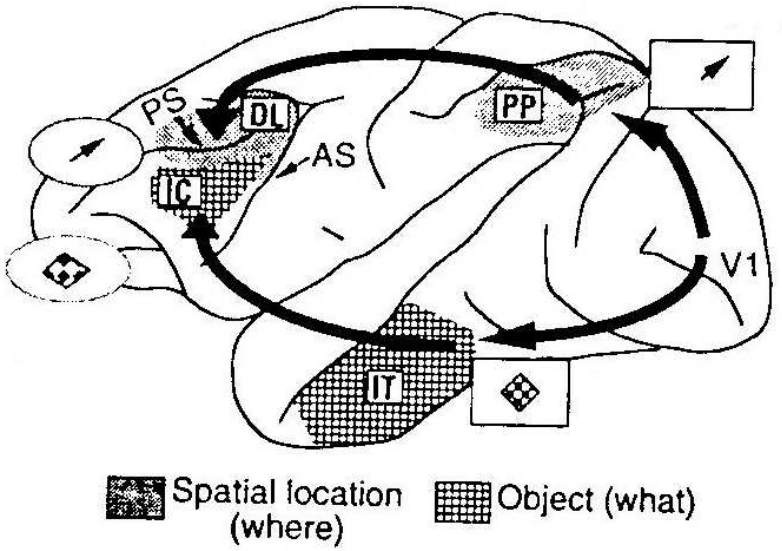




Summary - learning of associations and prospective activity

- Prospective activity due to strengthening of connections between two populations coding for associated stimuli
- Prospective activity must appear after retrospective activity (only way to link two stimuli that are separated in time) (see Erickson and Desimone, 1999)
- Similar mechanisms might underlie semantic priming phenomena

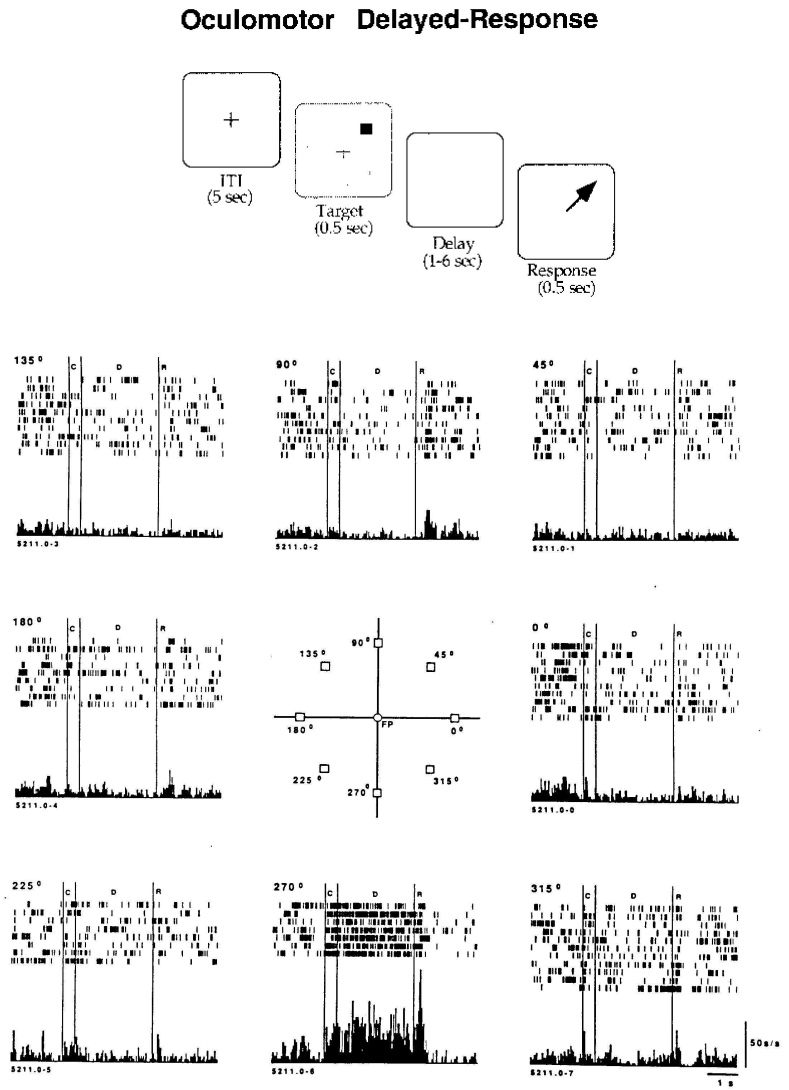
Persistent activity in delayed response tasks

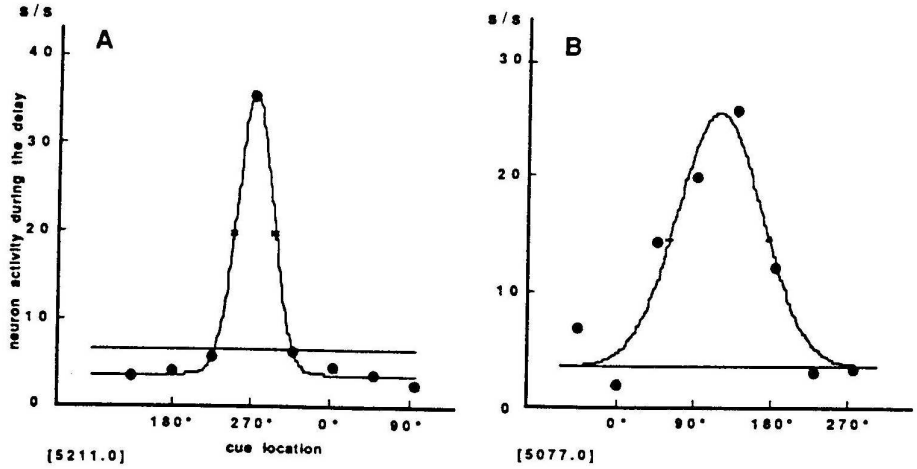


- Ventral stream and PFC:
  identity of stimuli (WHAT?)
- Dorsal stream and PFC:
  spatial location of stimuli (WHERE?)

Persistent activity in delayed oculomotor task

Persistent activity in delayed oculomotor task

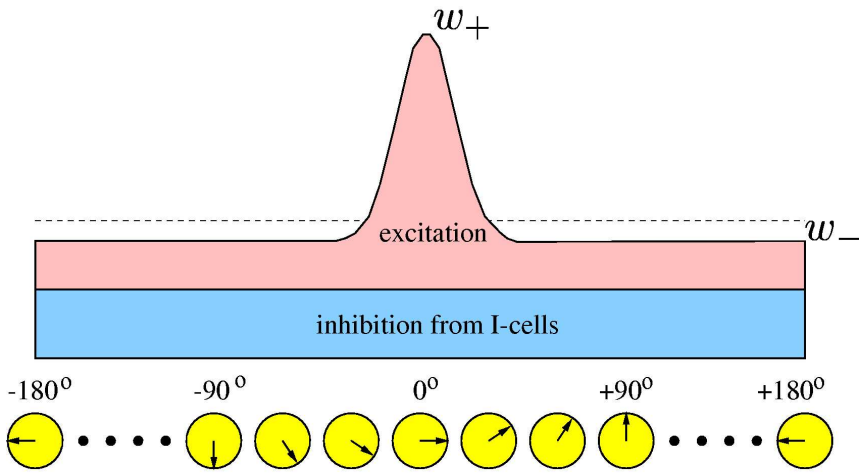




Funahashi et al 1989

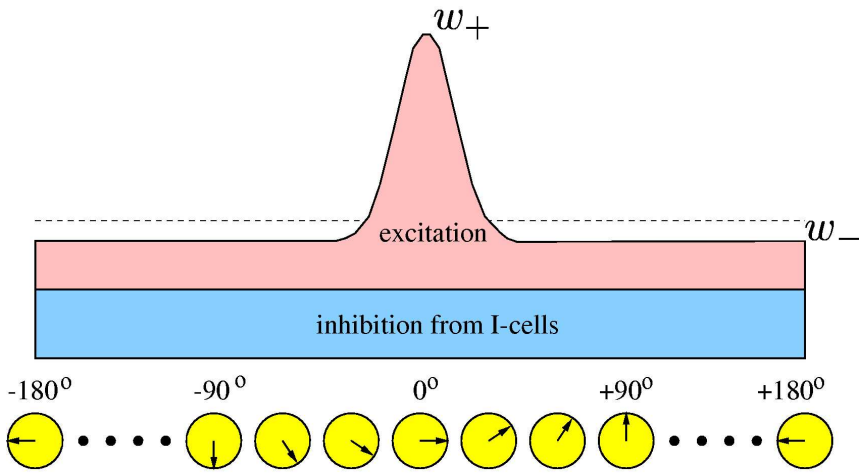
Network models for spatial working memory

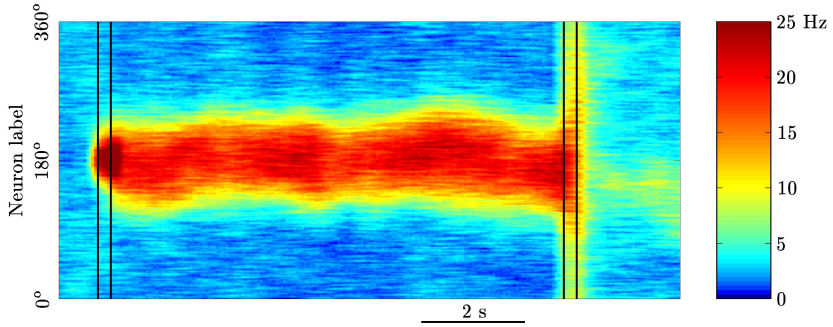
Ring model (with Gaussian footprint and integrate-and-fire neurons)



Network models for spatial working memory

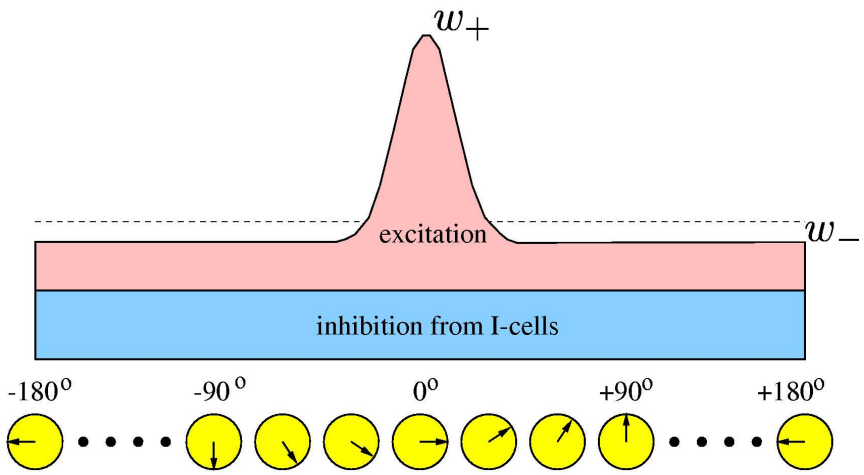
Ring model (with Gaussian footprint and integrate-and-fire neurons)

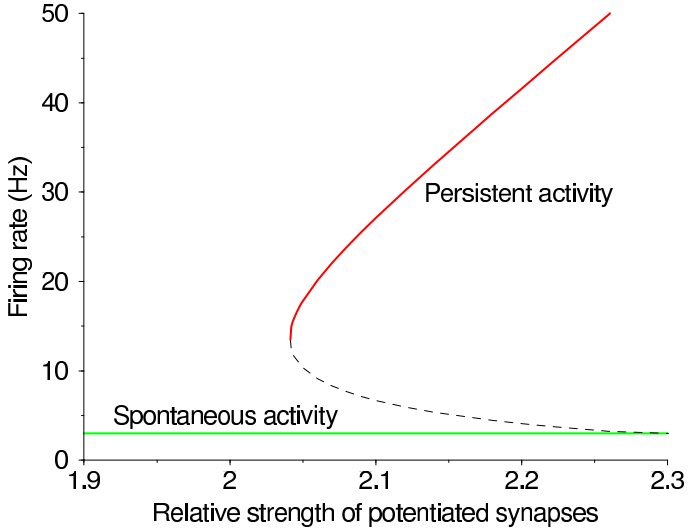


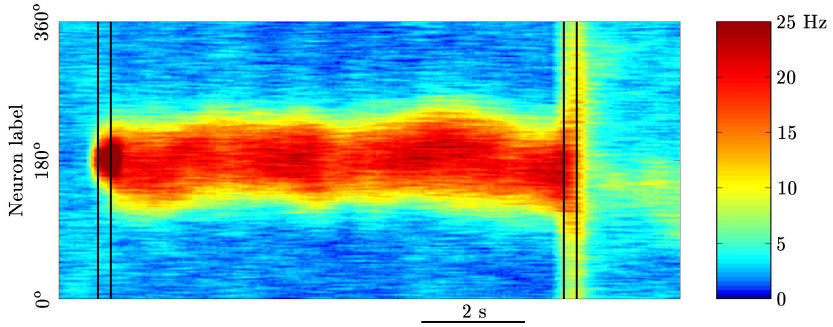


Network models for spatial working memory

Ring model (with Gaussian footprint and integrate-and-fire neurons)

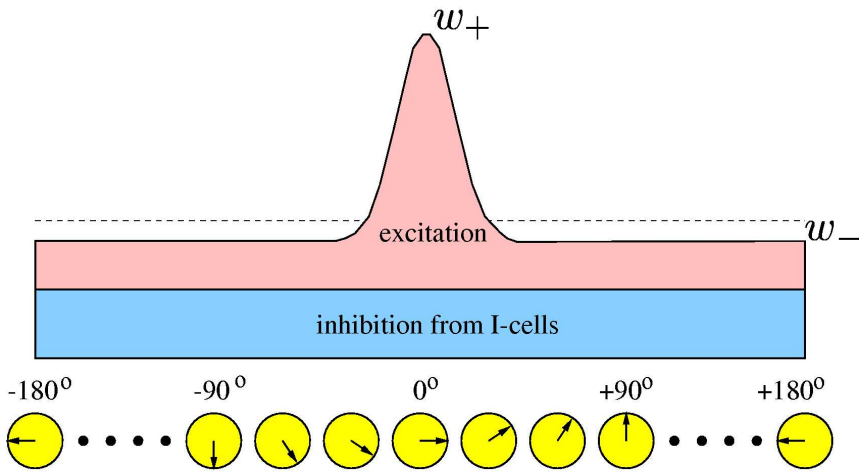


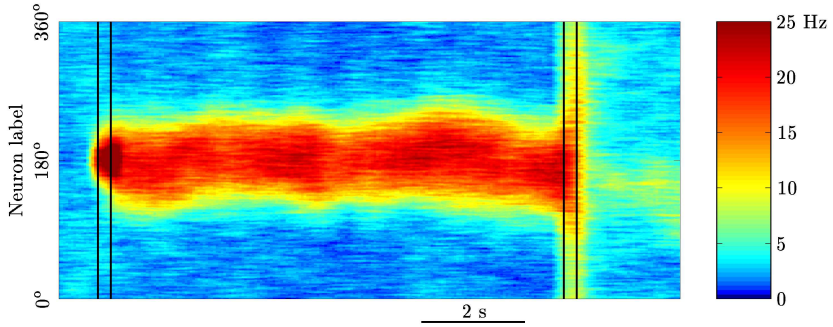


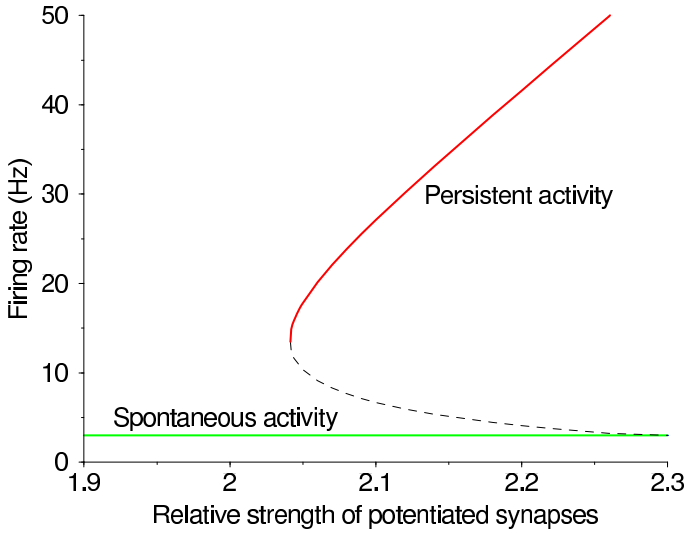


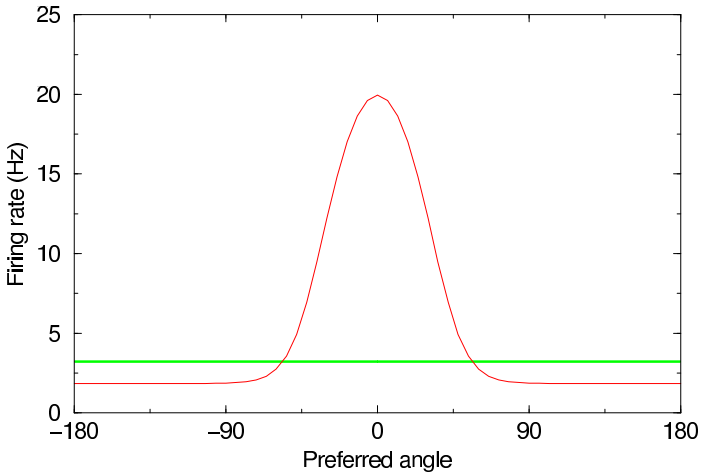
Network models for spatial working memory

Ring model (with Gaussian footprint and integrate-and-fire neurons)



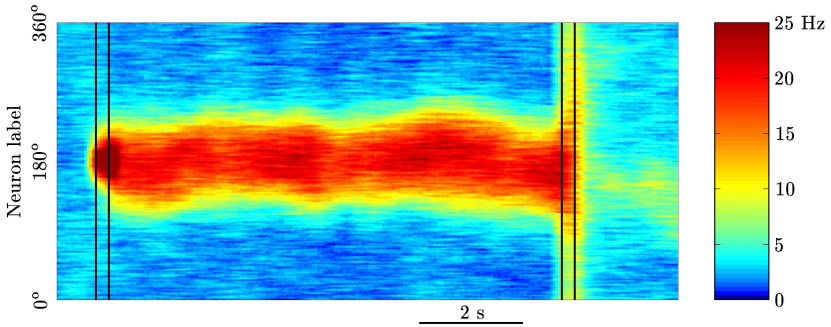




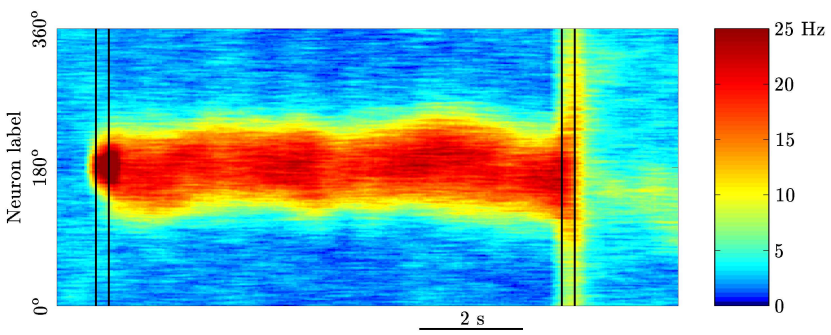


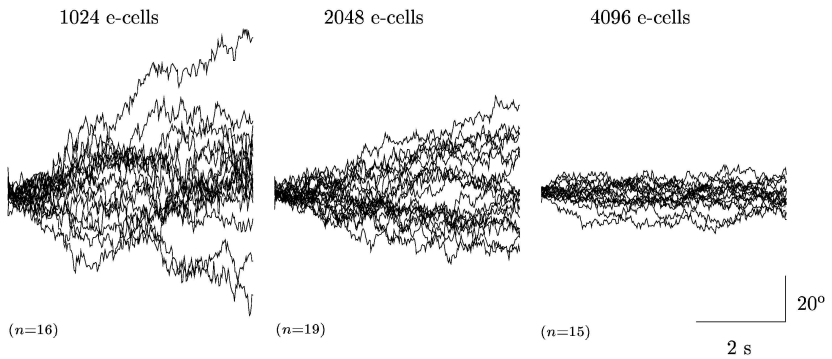
Compte et al 2000

Drift of memorized position with time

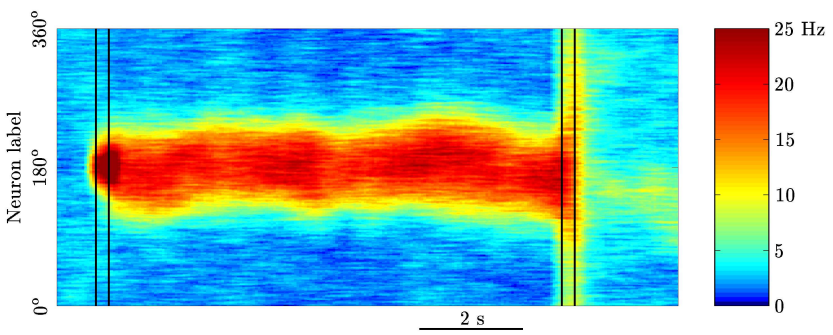


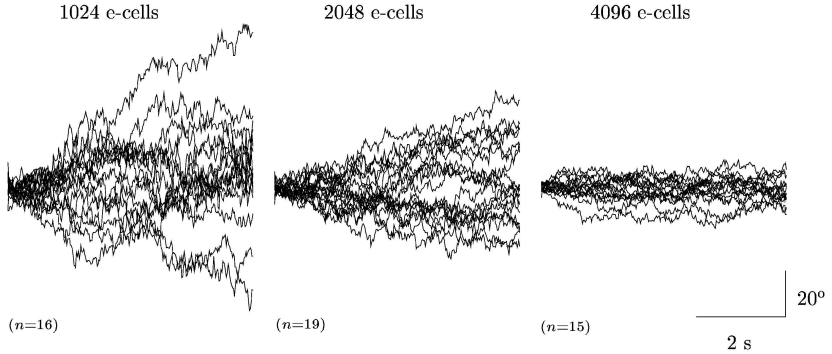
Drift of memorized position with time





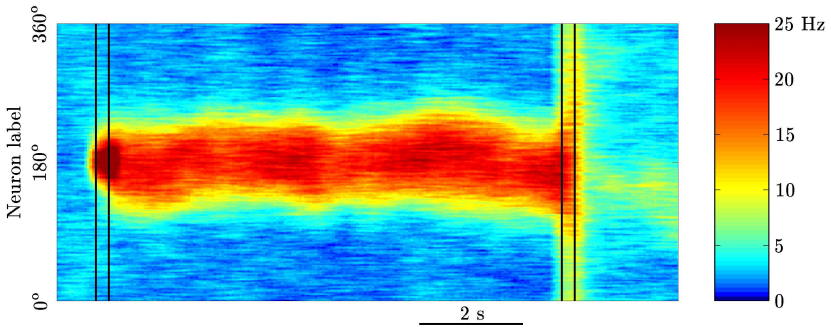
Drift of memorized position with time

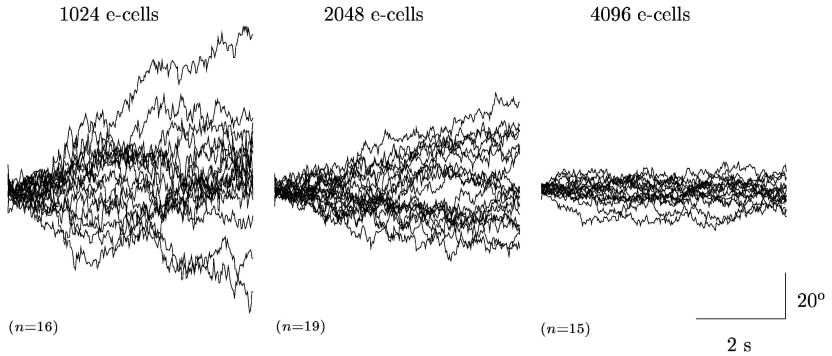




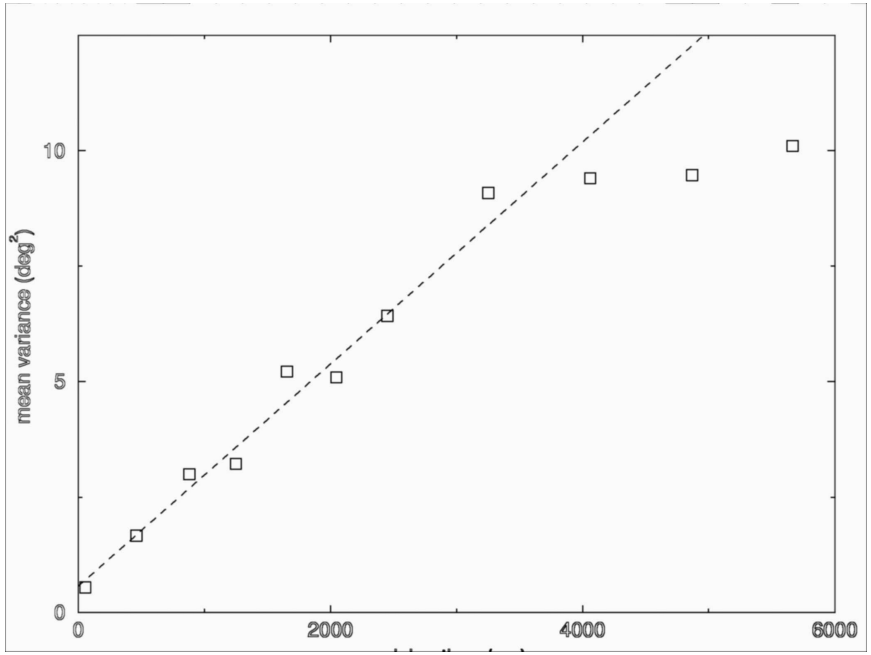
- Variance of error in memorized position increases linearly with time
- Good agreement with experimental data at short times (in both monkey and human)

Drift of memorized position with time





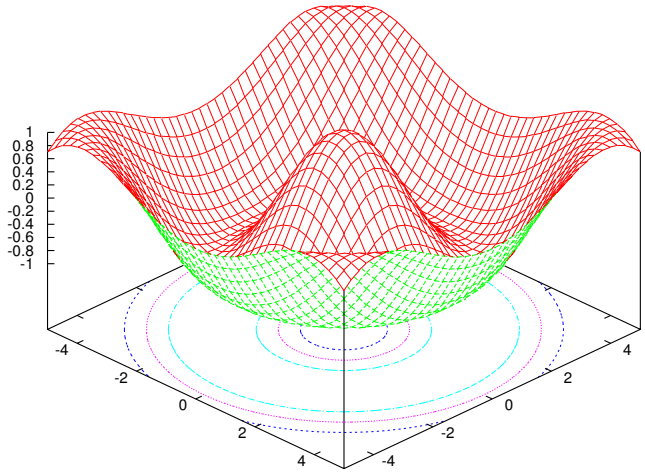
- Variance of error in memorized position increases linearly with time
- Good agreement with experimental data at short times (in both monkey and human)

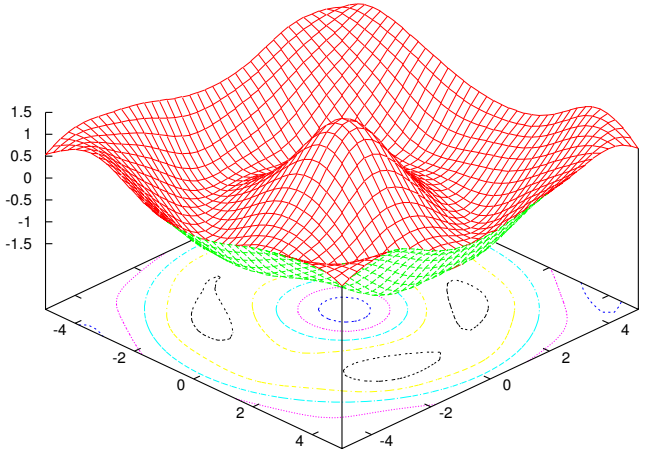


White et al 1994

The fine tuning problem and how to solve it

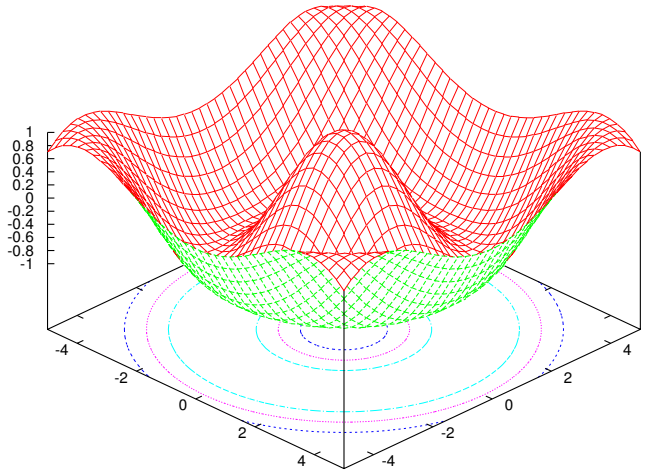
In presence of disorder/heterogeneities, a continuous attractor breaks down in a small number of discrete attractors

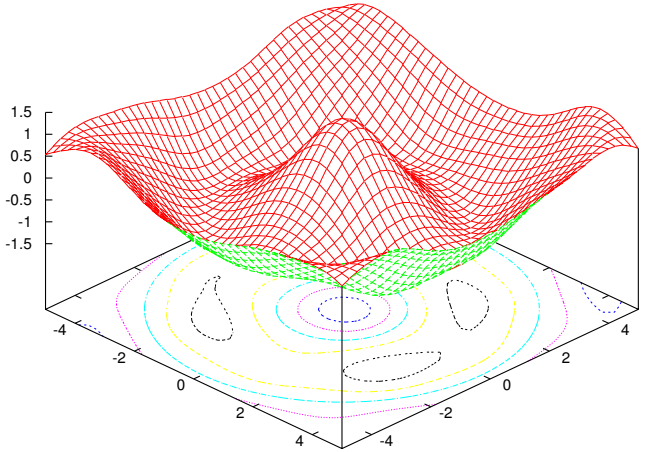




The fine tuning problem and how to solve it

In presence of disorder/heterogeneities, a continuous attractor breaks down in a small number of discrete attractors



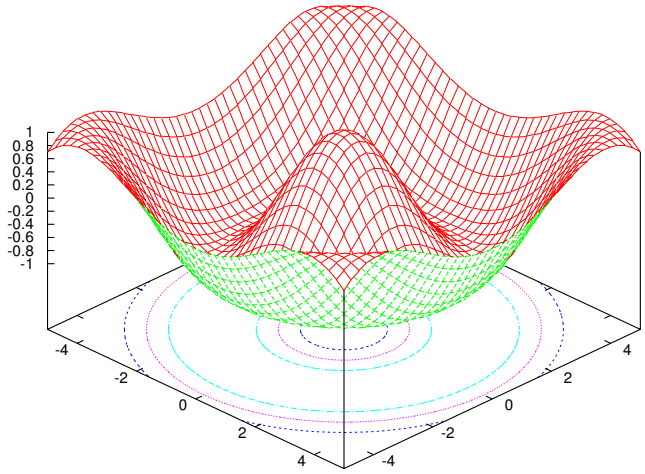


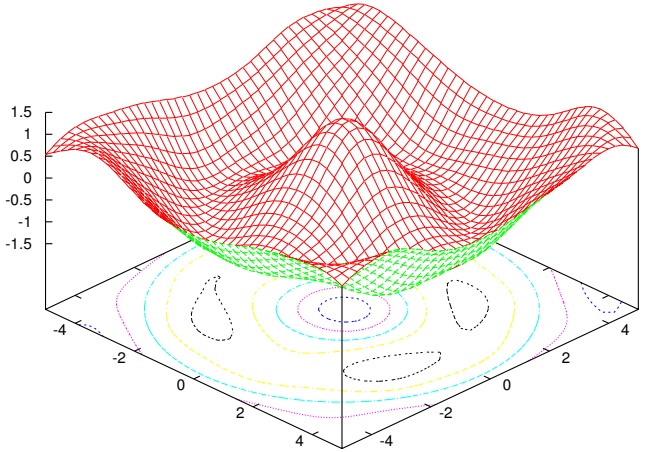
Solutions:

- Bistability at the neuronal/dendritic level (Koulakov et al 2002)

The fine tuning problem and how to solve it

In presence of disorder/heterogeneities, a continuous attractor breaks down in a small number of discrete attractors





Solutions:

- Bistability at the neuronal/dendritic level (Koulakov et al 2002)
- Homeostasis mechanisms (Renart et al 2003)

Conclusions

- Attractor network dynamics explain salient features of persistent activity in several areas of the cerebral cortex
- In this framework, learning corresponds to creation of an attractor, forgetting to the disappearance of an attractor
- Persistent activity allows to:
  - Bridge the temporal gap between stimulus and behavioral response;
  - Bridge the temporal gap between temporally separated stimuli, necessary to learn contextual information;
- Attractor networks have also been proposed to account for a variety of other neurophysiological phenomena
  - Decision-making — one attractor corresponding to each possible decision
  - Dynamics of spontaneous activity in sensory cortices — evidence for the system wandering through different attractors corresponding to representations of the external world (e.g.‘orientation states’ in V1)

A few open problems

- Relative contributions of single neuron/network mechanisms for persistent activity?
- Mechanisms/roles for temporal structure in persistent activity?
- Mechanisms for maintenance of several objects in short-term (working memory)?