# Learning and memory in recurrent neural networks

Nicolas Brunel

Departments of Neurobiology and Physics, Duke University

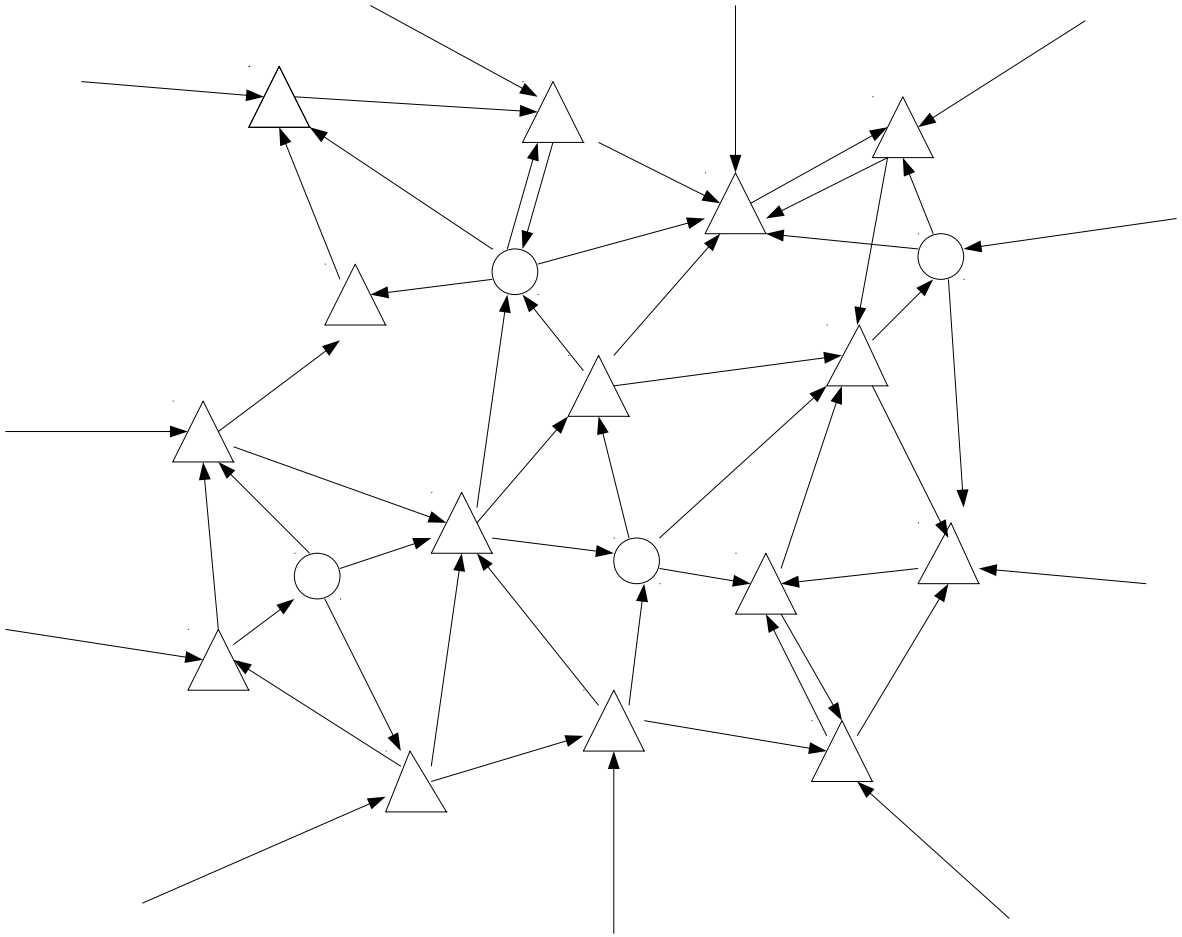
# Outline

1. Introduction: The Hebbian/Attractor Neural Network scenario
2. A brief overview of the relevant neurobiology
3. The Hopfield approach
4. The Gardner approach
5. Open questions

# Outline

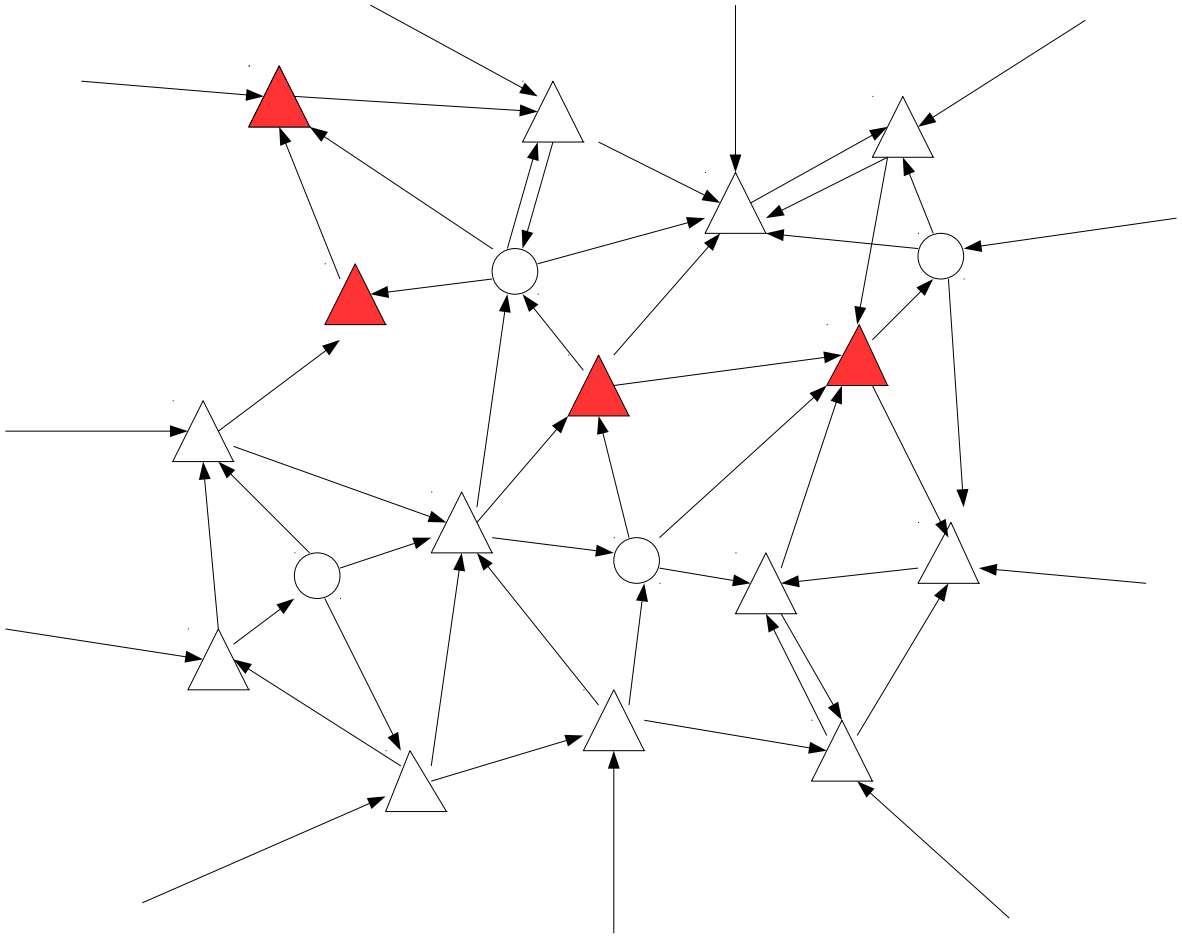
1. **Introduction: The Hebbian/Attractor Neural Network scenario**
2. A brief overview of the relevant neurobiology
3. The Hopfield approach
4. The Gardner approach
5. Open questions

# Learning and memory: The Hebbian scenario



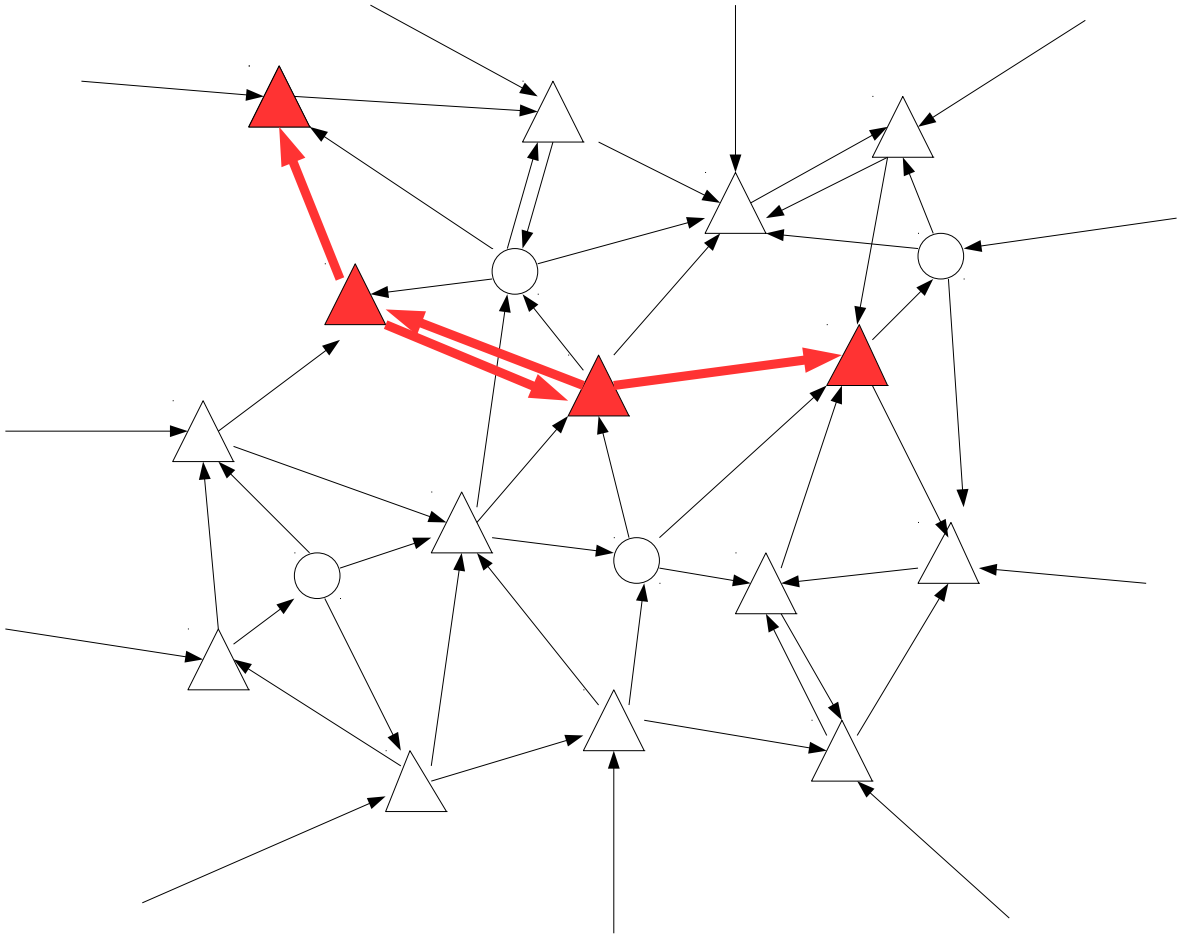
- External stimuli $\Rightarrow$ Changes in neuronal activity

Stimulus A $\Rightarrow$

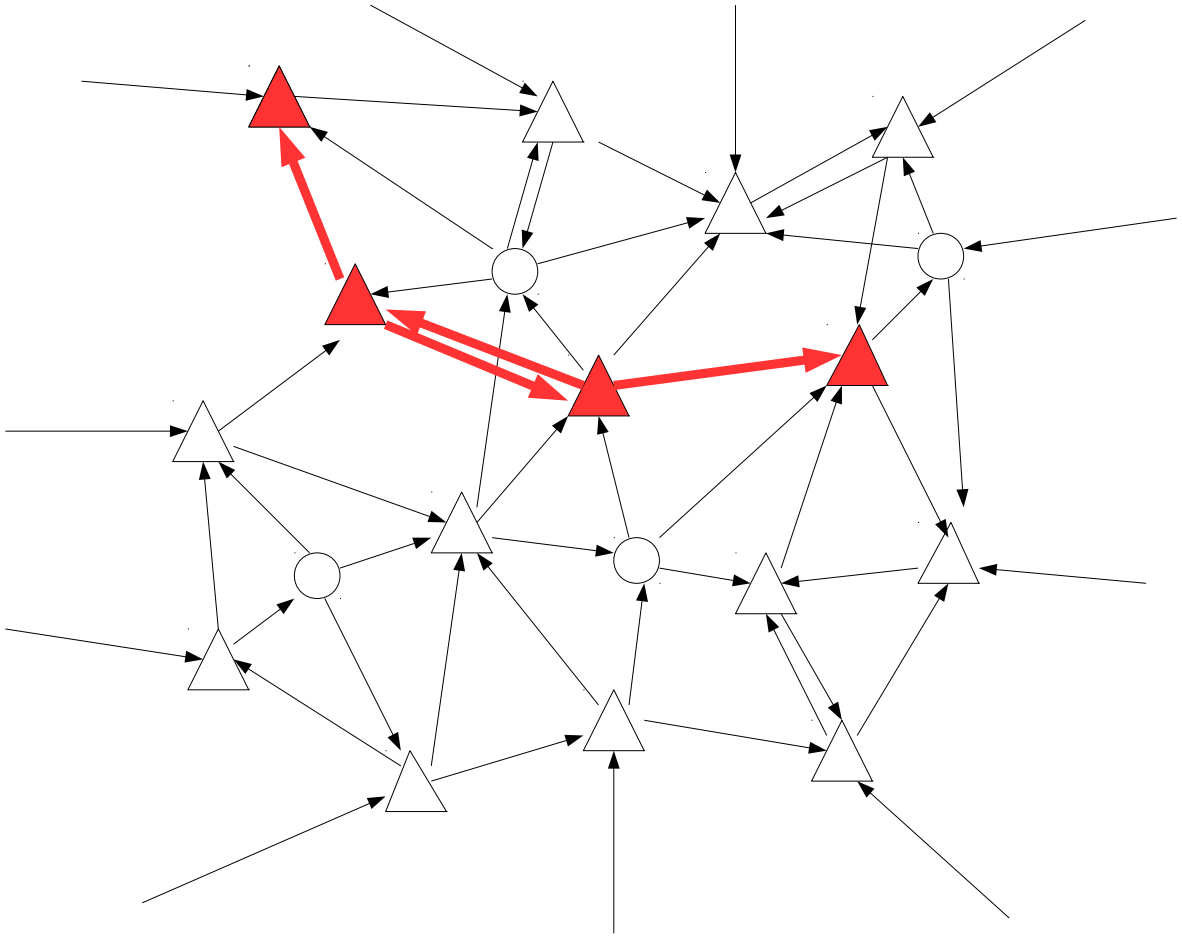


- Changes of activity $\Rightarrow$ changes in synaptic connectivity

Stimulus A $\Rightarrow$

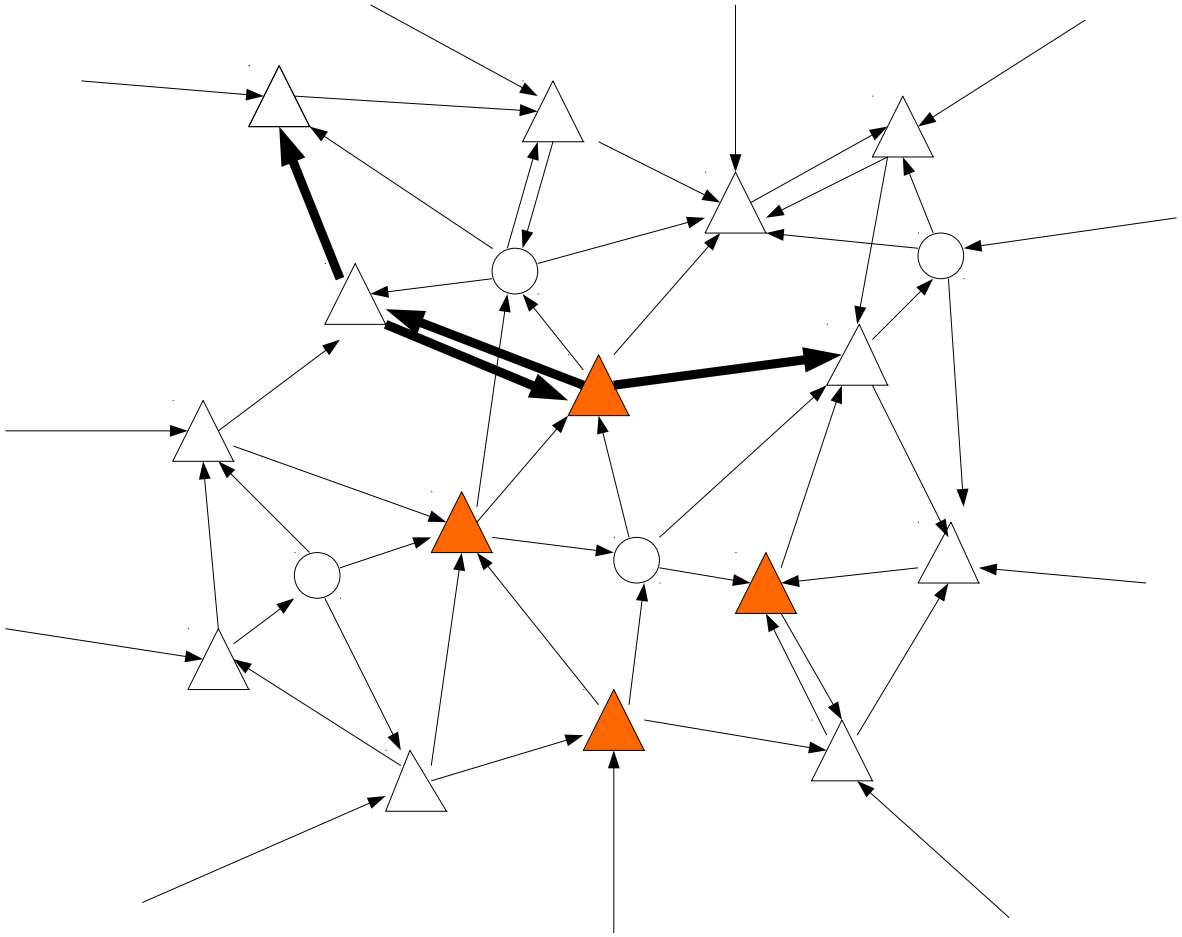


- Changes in synaptic connectivity $\Rightarrow$ changes in neuronal activity/dynamics



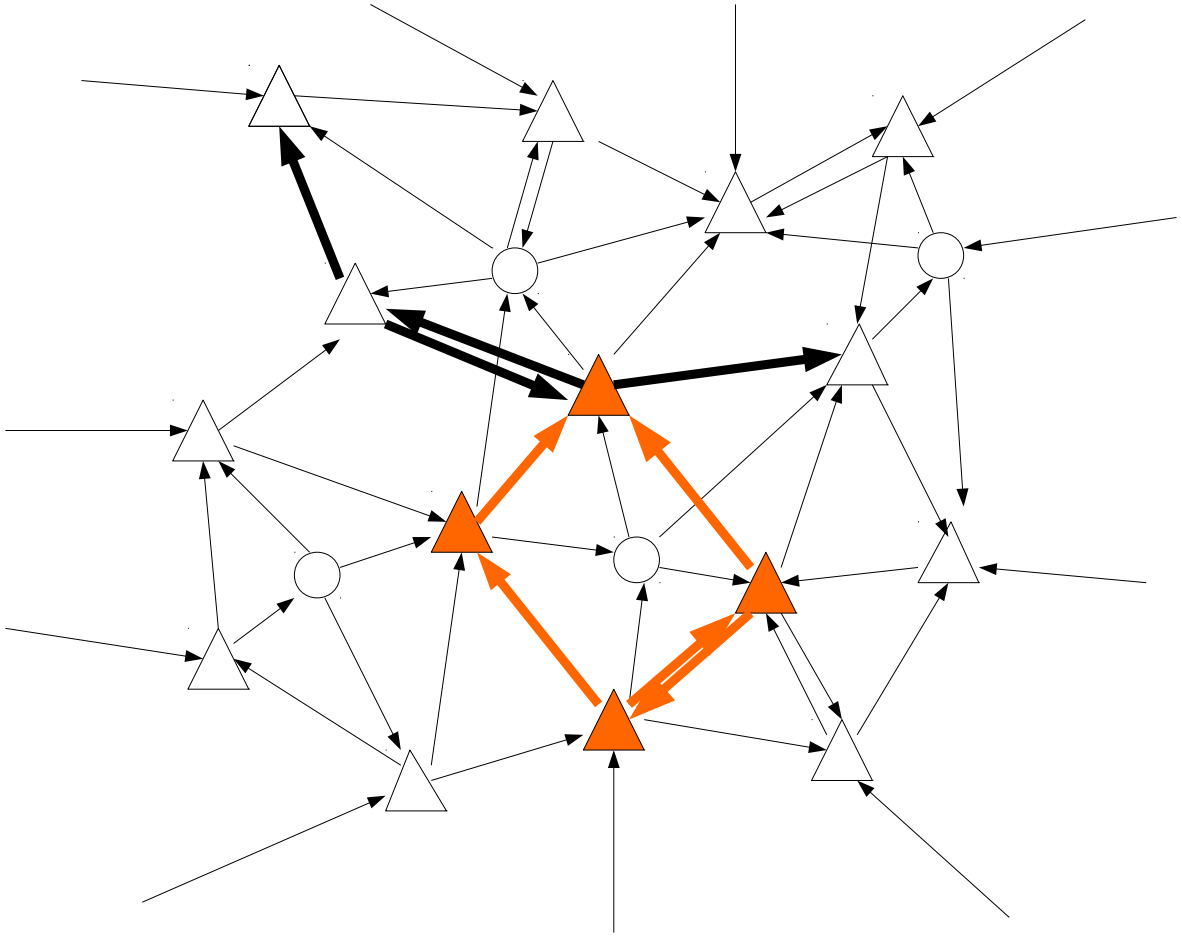
- Another stimulus triggering activity in a distinct subset of neurons...

Stimulus B $\Rightarrow$

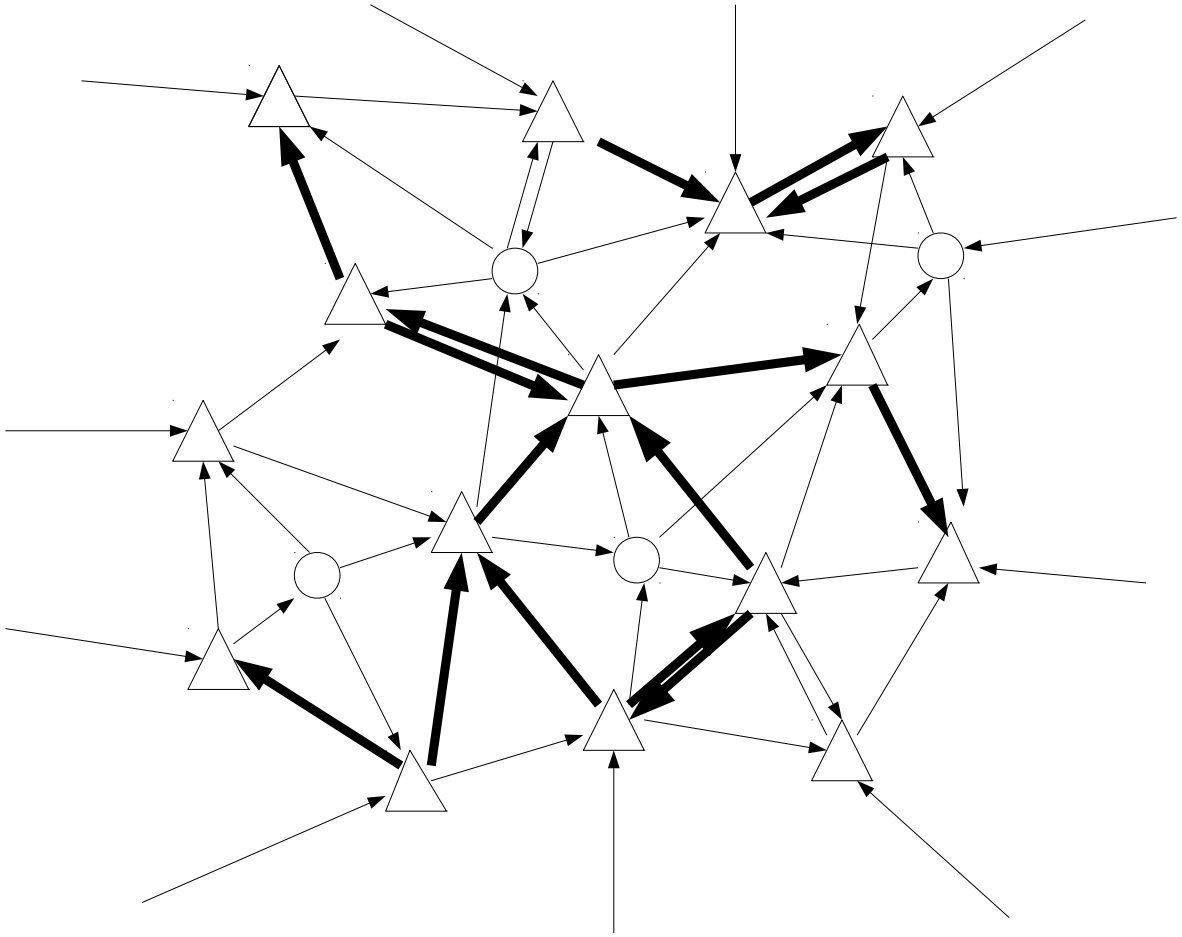


- ... will also lead to changes in connectivity

Stimulus B $\Rightarrow$



$\Rightarrow$ Synaptic connectivity $=$ superposition of traces left by external inputs



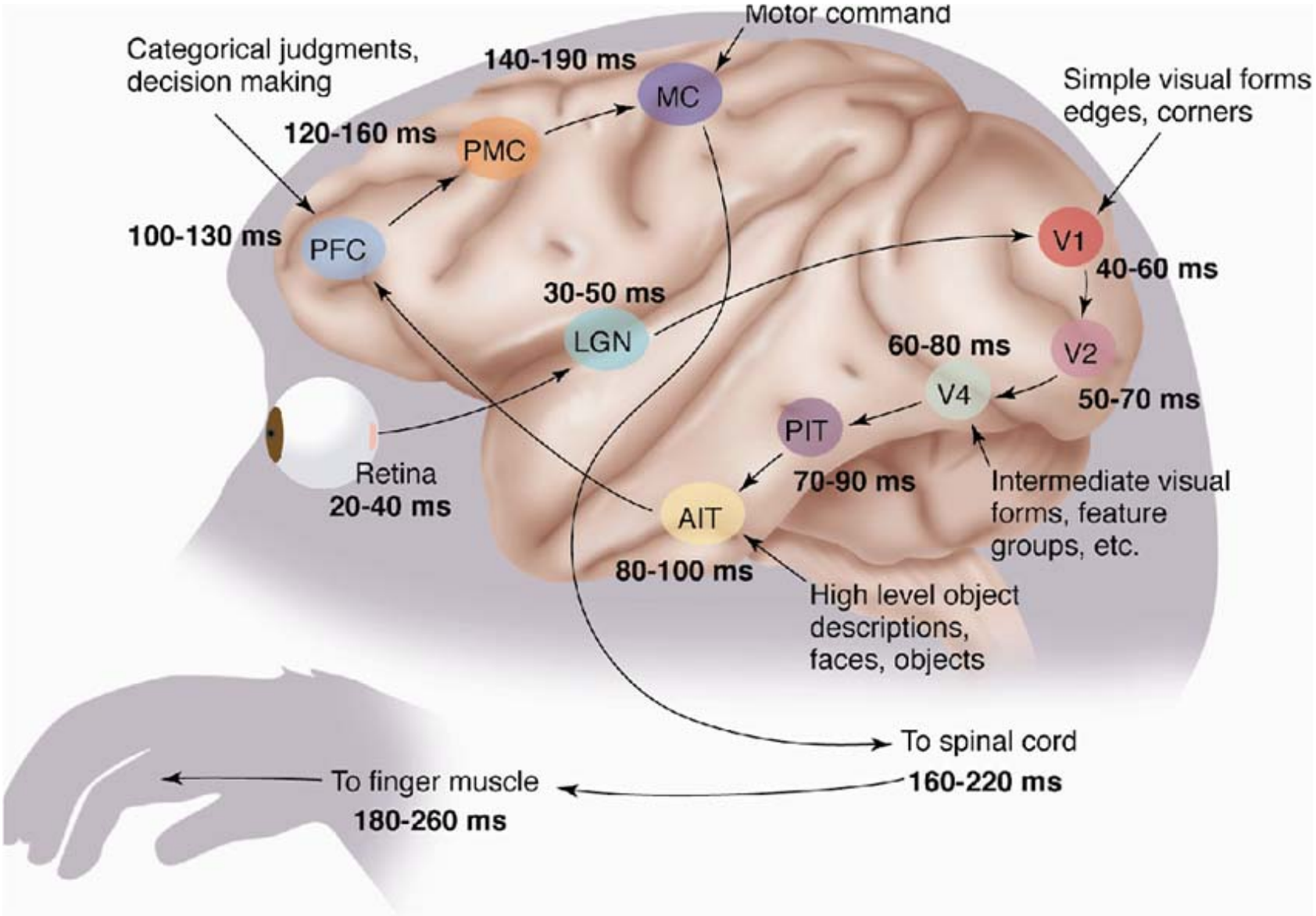
## Questions

- What is/are the synaptic plasticity (‘learning’) rule(s)?
- What are the statistics of synaptic connectivity?
- How does synaptic plasticity affect network dynamics?
- What is the storage capacity of networks?

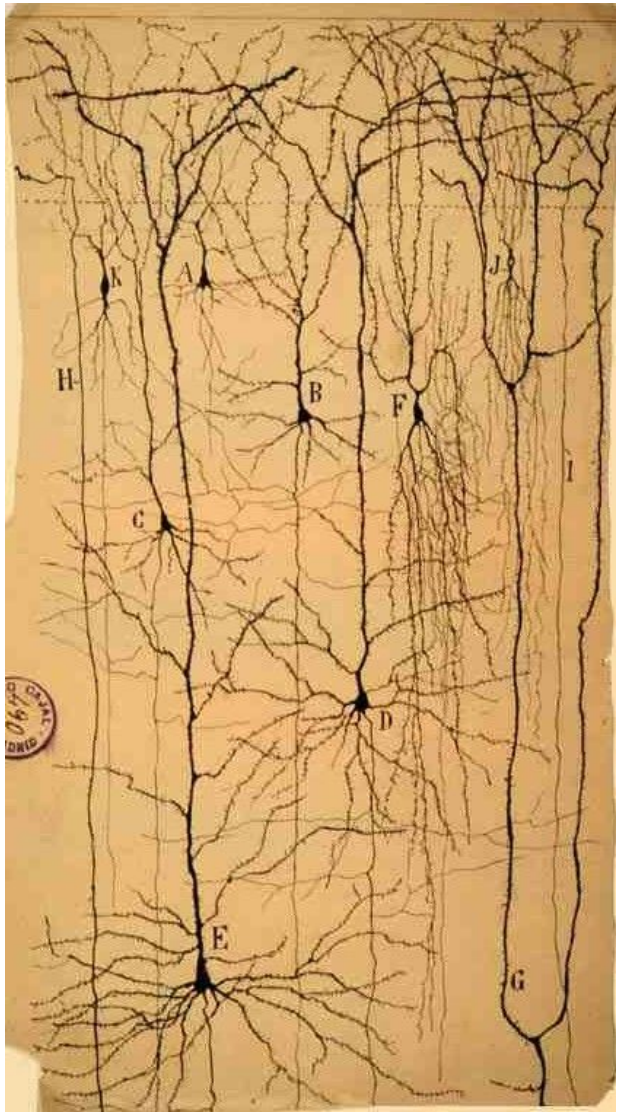
# Outline

1. Introduction: The Hebbian/Attractor Neural Network scenario
2. **A brief overview of the relevant neurobiology**
3. The Hopfield approach
4. The Gardner approach
5. Open questions

# Cerebral cortex - feedforward flow of information

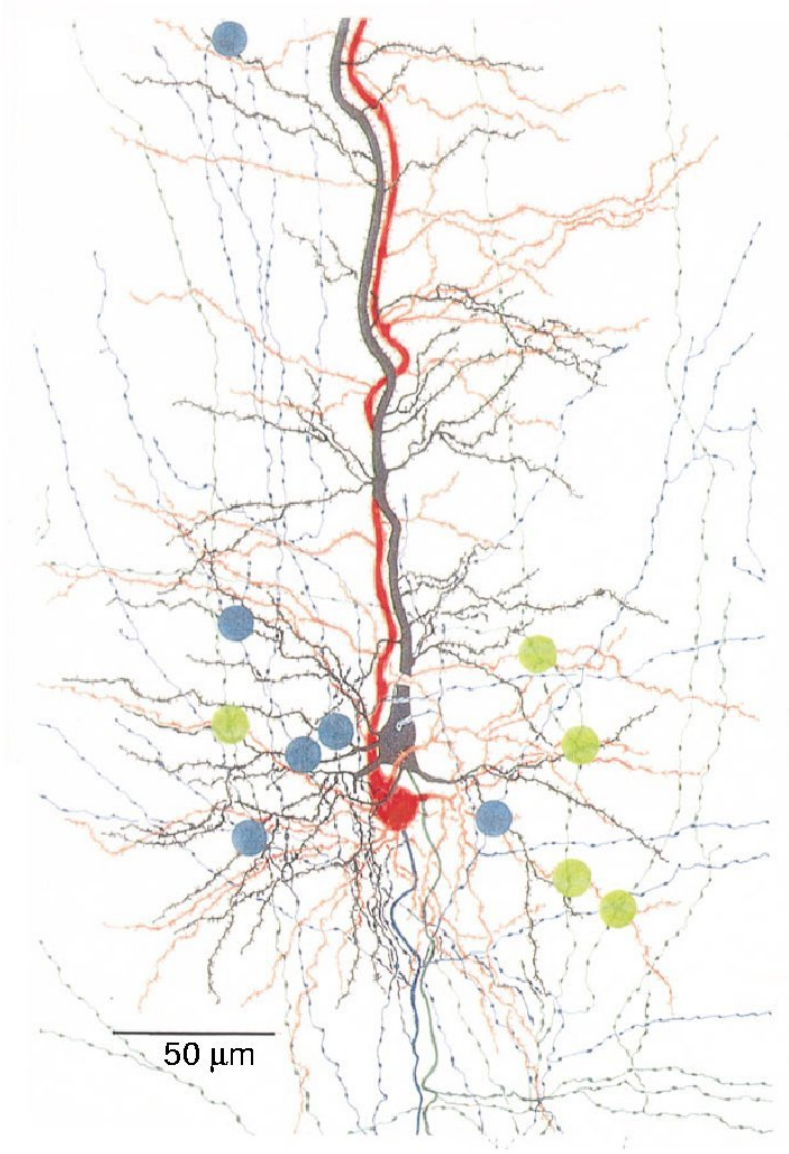


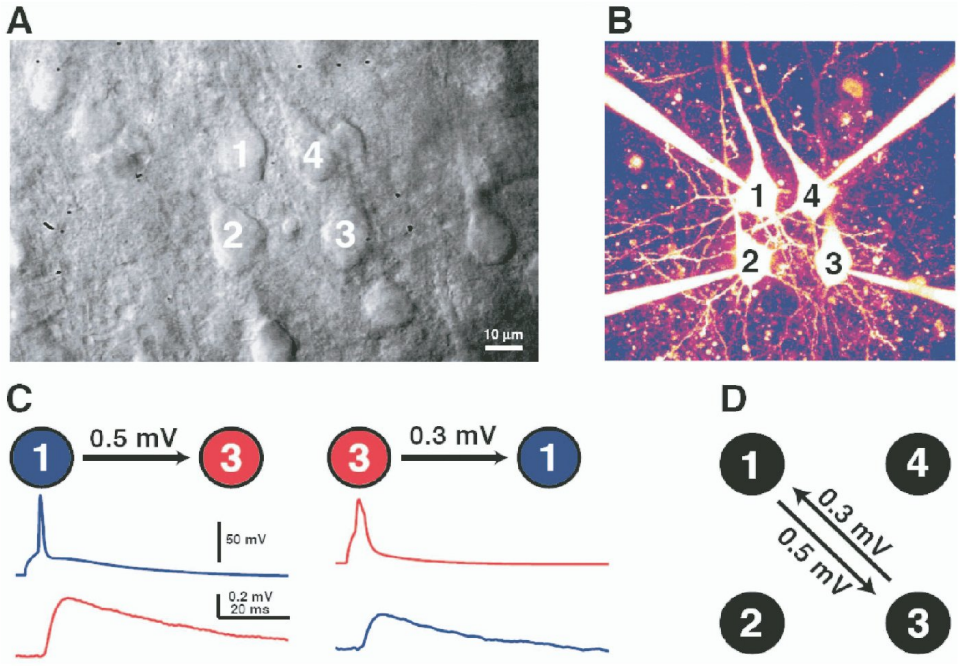
# Cerebral cortex - neurons, synapses




- Neuron density $\sim 10^5/\text{mm}^3$
- Neurons divided in two main classes:
  - Excitatory (80%) - mostly pyramidal cells
  - Interneurons (20%) - very diverse
- Synapse density $\sim 10^9/\text{mm}^3$ (10,000 synapses per neuron)
- Approximately half of the synapses are local ($<$ 1mm), other half long-range
- **Strong recurrent connectivity**

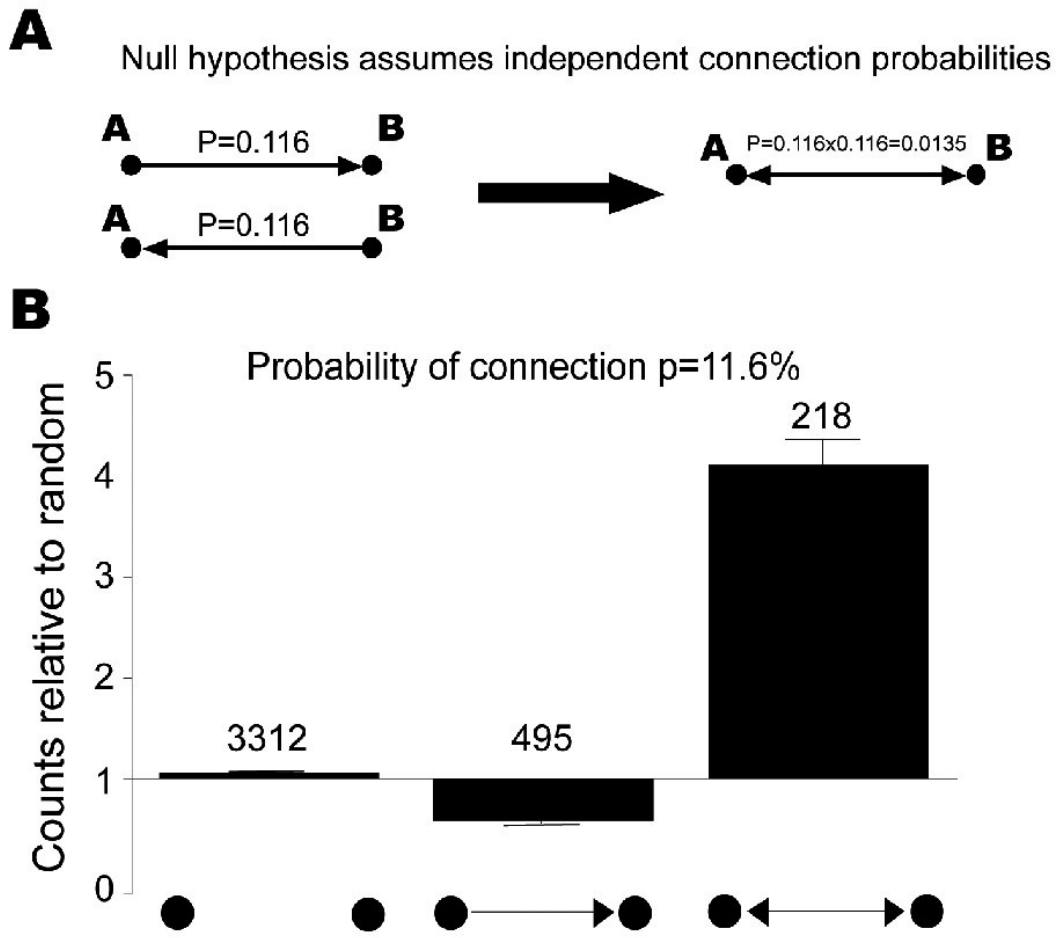
# Cortical microcircuit - connectivity



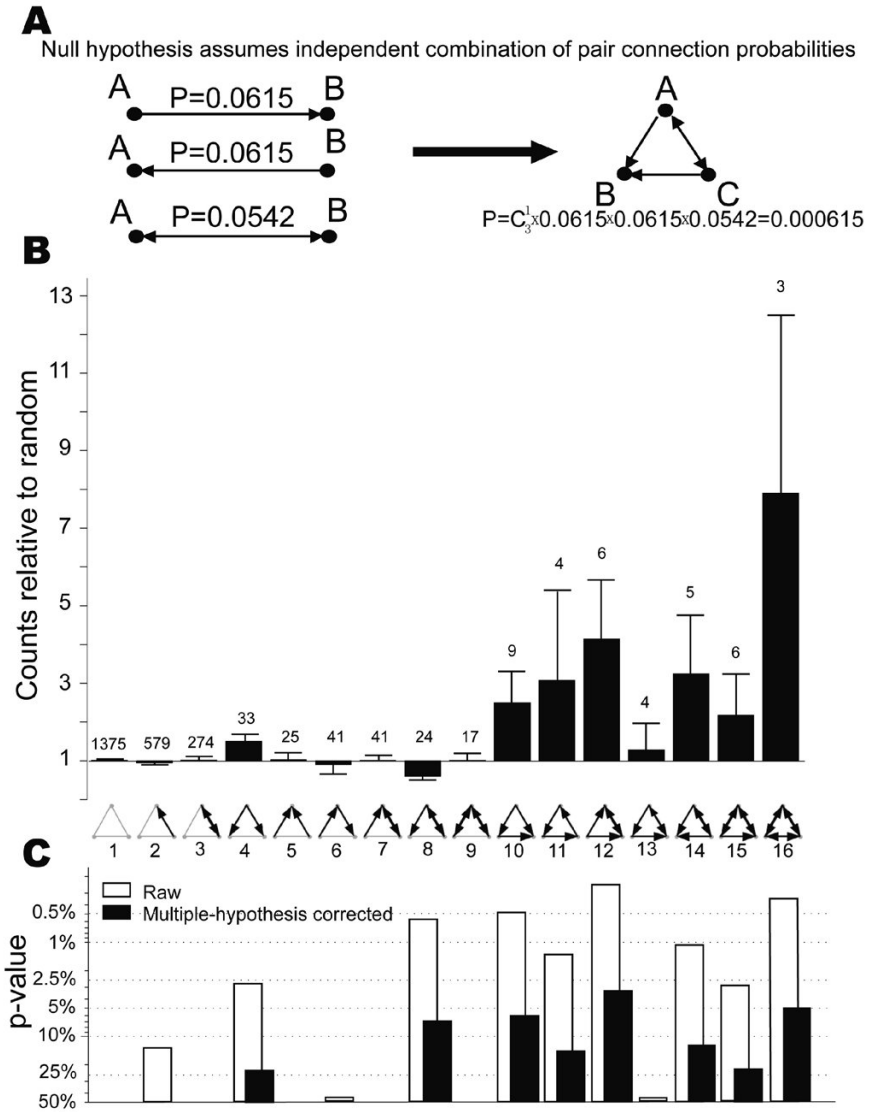

- Anatomy: Potentially fully connected network at short spatial scales (dendrite of a neuron 'touches' the axon of any other neighboring neuron with probability close to 1)
- Electrophysiology: Connection probabilities $\sim$ 0.1 (E-E), 0.5 (E-I, I-E, I-I) at short distances ($< 100 \mu$m)

# Motifs: pairs

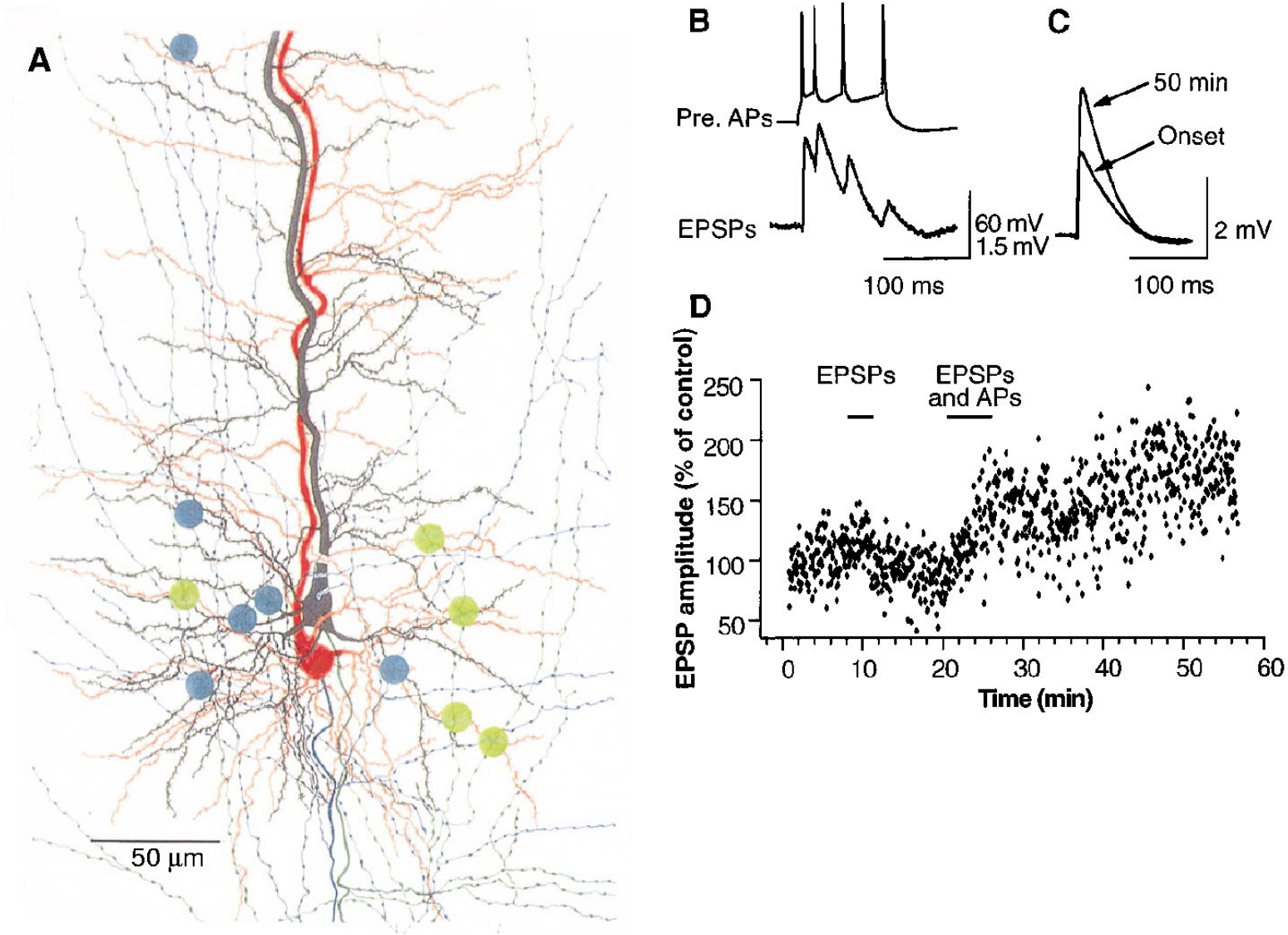



- Bidirectional connections are overrepresented
- Partially symmetric synaptic connectivity

# Motifs: triplets

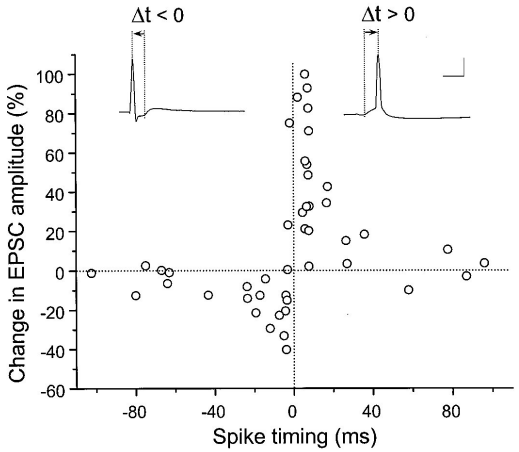


# Synaptic connectivity is plastic

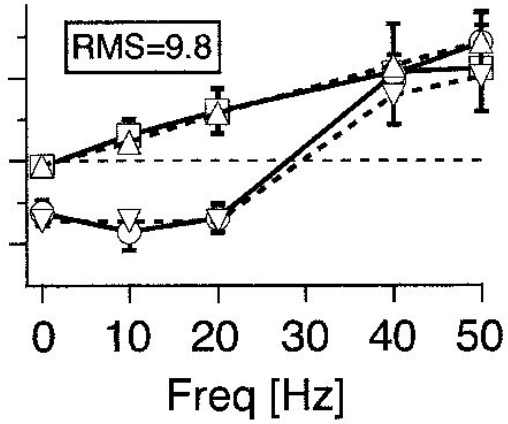


# What controls synaptic plasticity?

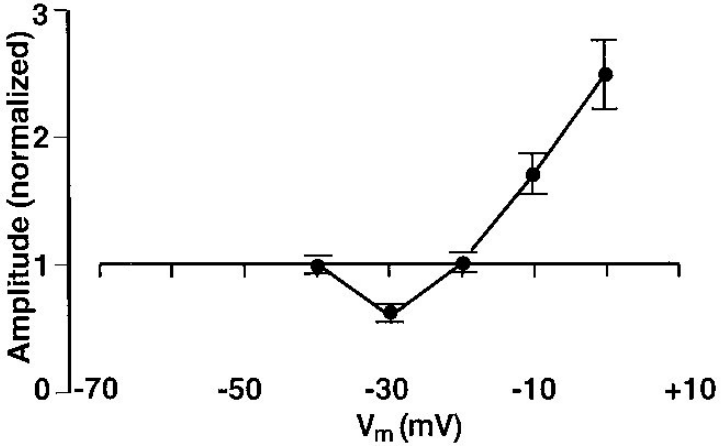
Spike timing



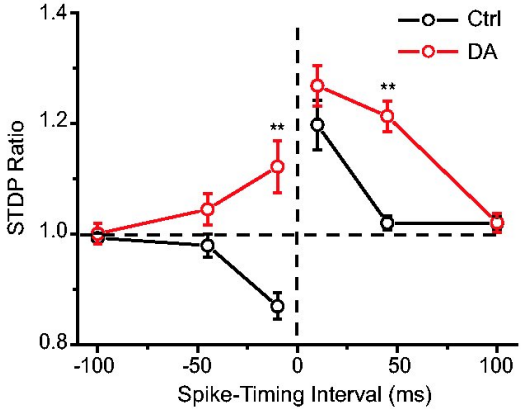

Firing rate



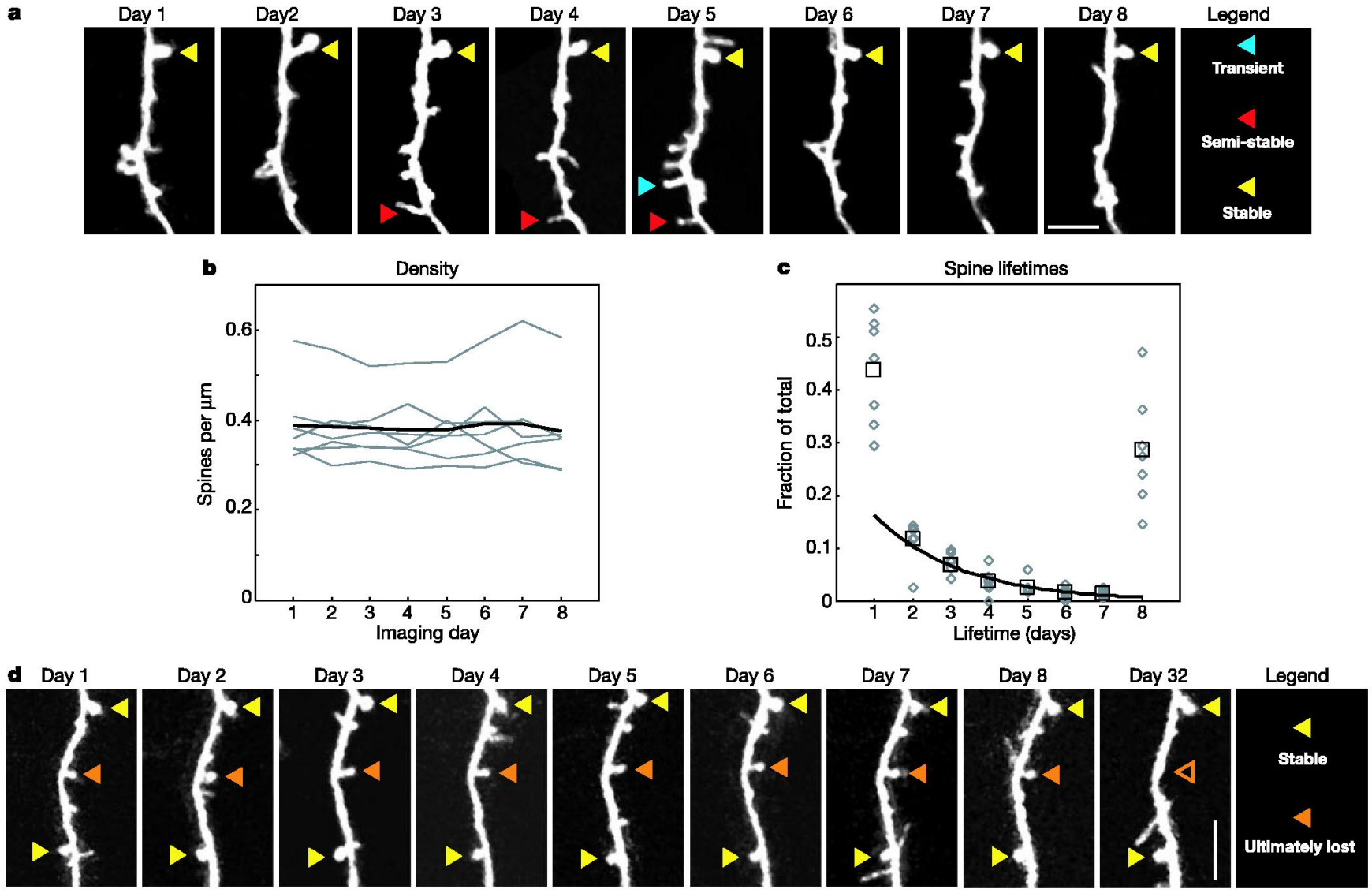

Post-synaptic membrane potential



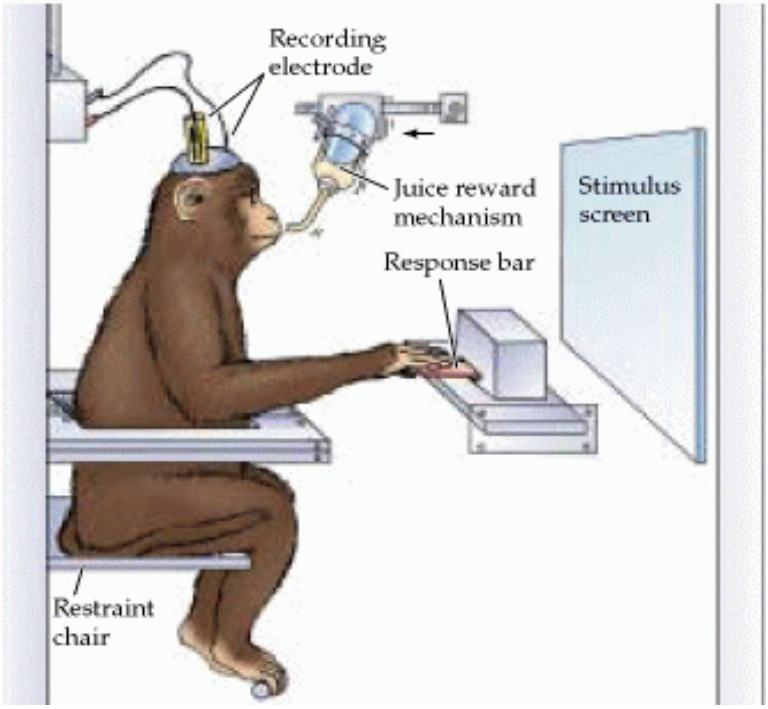

Neuromodulators (e.g. dopamine)

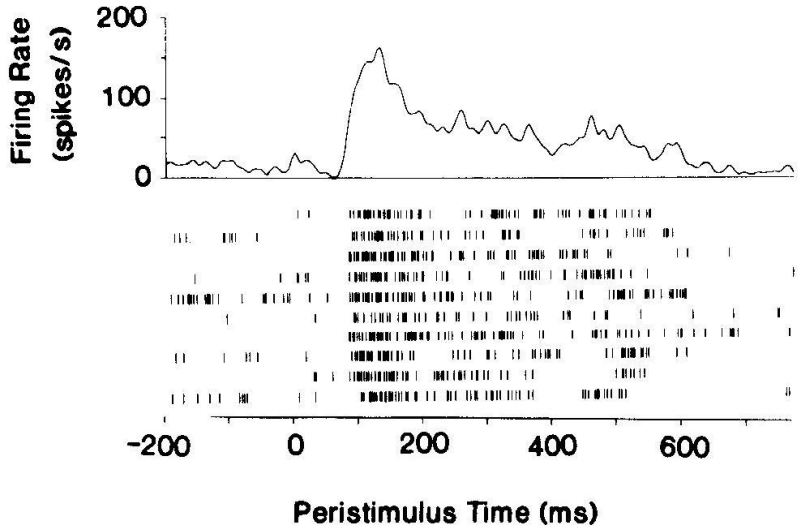


# Structural plasticity on longer time scales

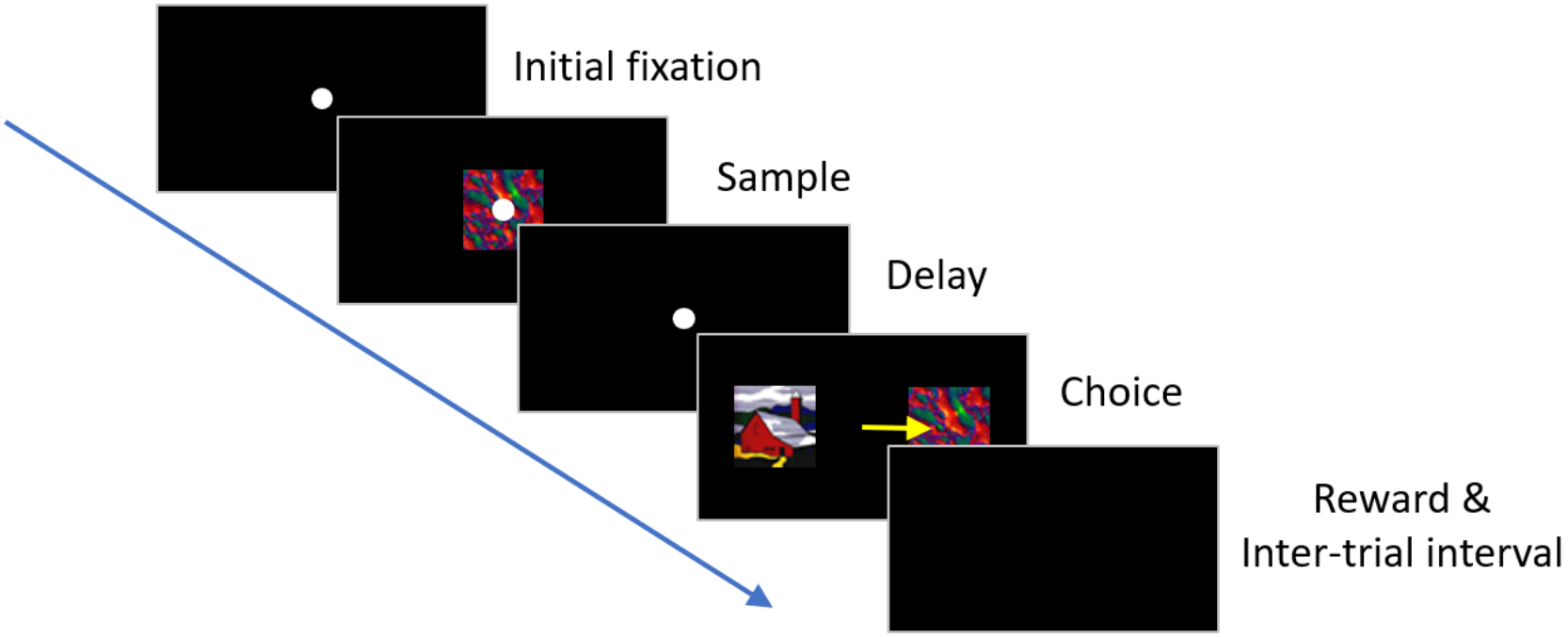


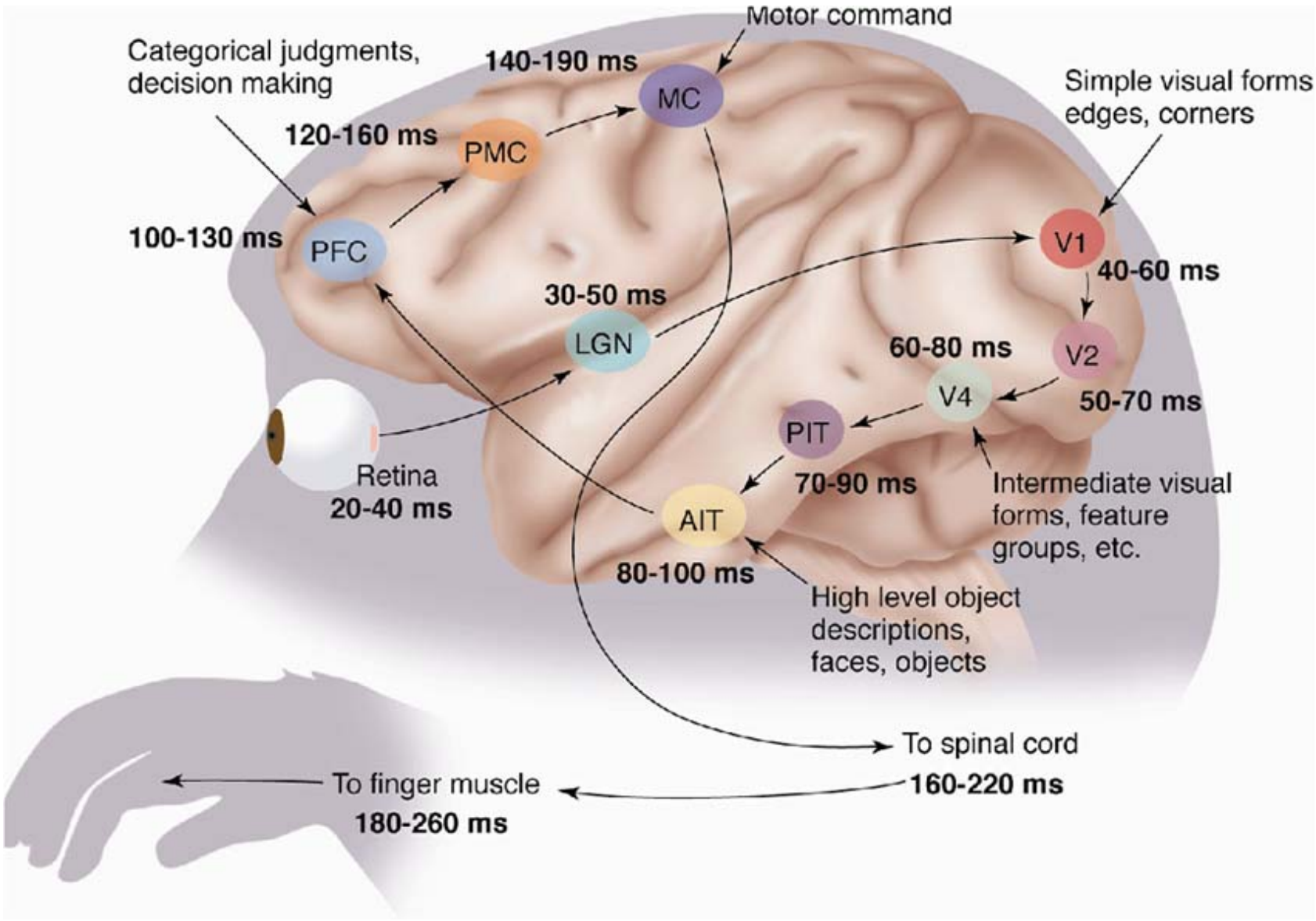
# Electrophysiological recordings in behaving animals



# Delay match to sample task

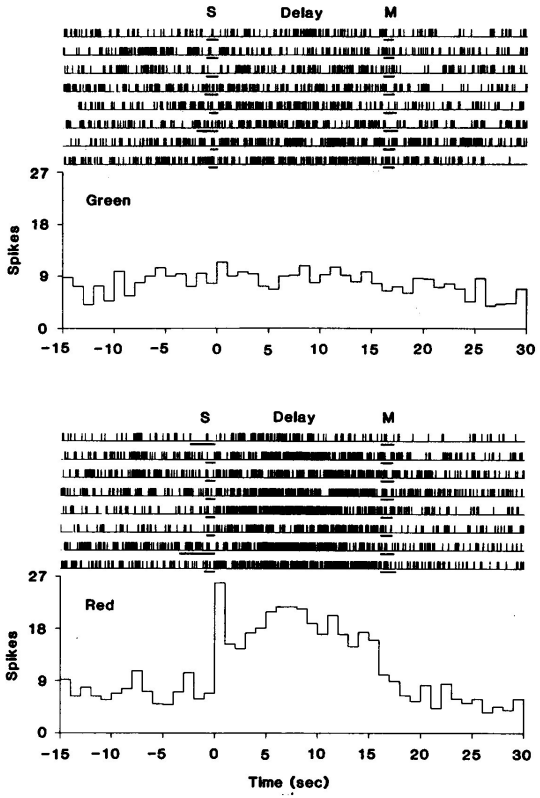



Motor command
Categorical judgments, decision making
140-190 ms
MC
Simple visual forms edges, corners
120-160 ms
PMC
100-130 ms
PFC
V1
40-60 ms
30-50 ms
LGN
60-80 ms
V2
V4
50-70 ms
PIT
Retina
20-40 ms
70-90 ms
Intermediate visual forms, feature groups, etc.
AIT
80-100 ms
High level object descriptions, faces, objects
To spinal cord
160-220 ms
To finger muscle
180-260 ms

# Selective persistent activity in DMS tasks in monkeys

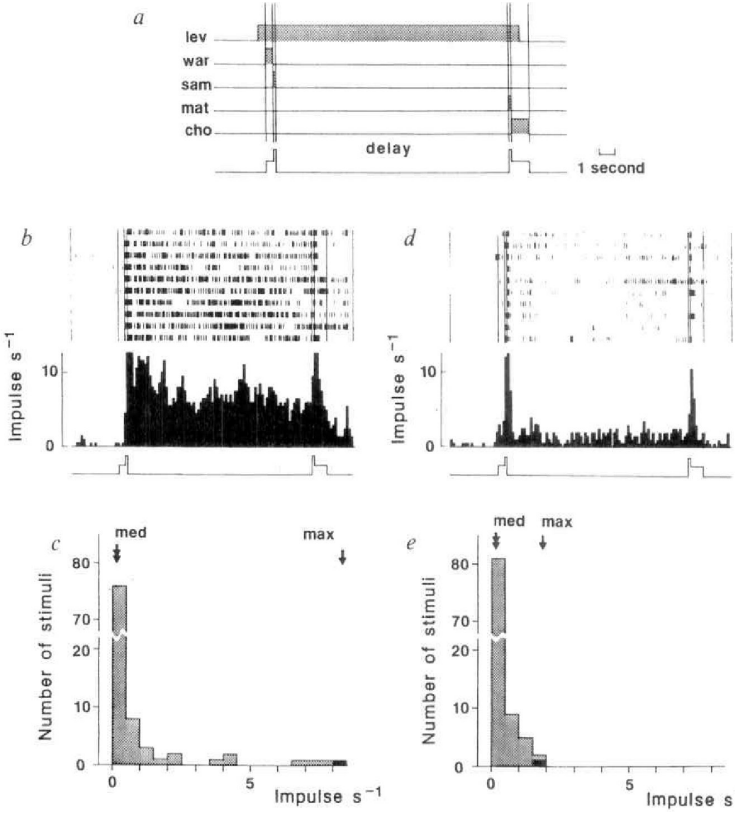
Fuster and Jervey 1981

ITC, colors



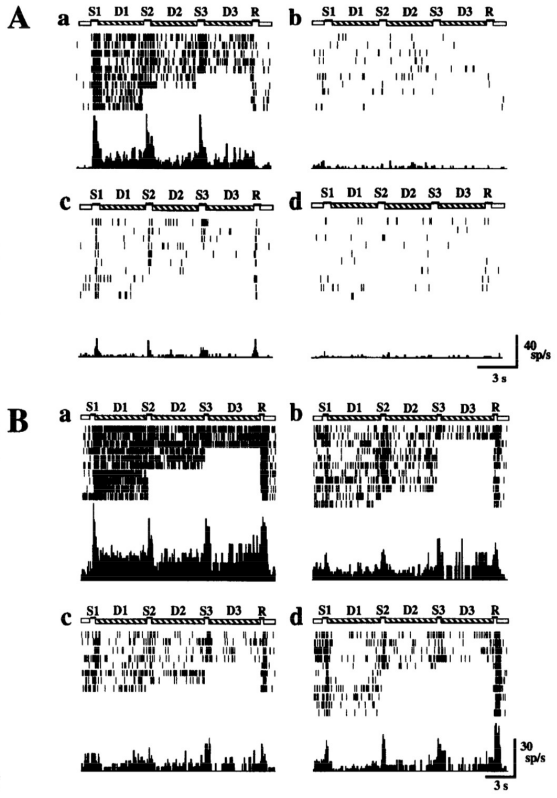

Miyashita 1988, etc...

ITC, abstract patterns



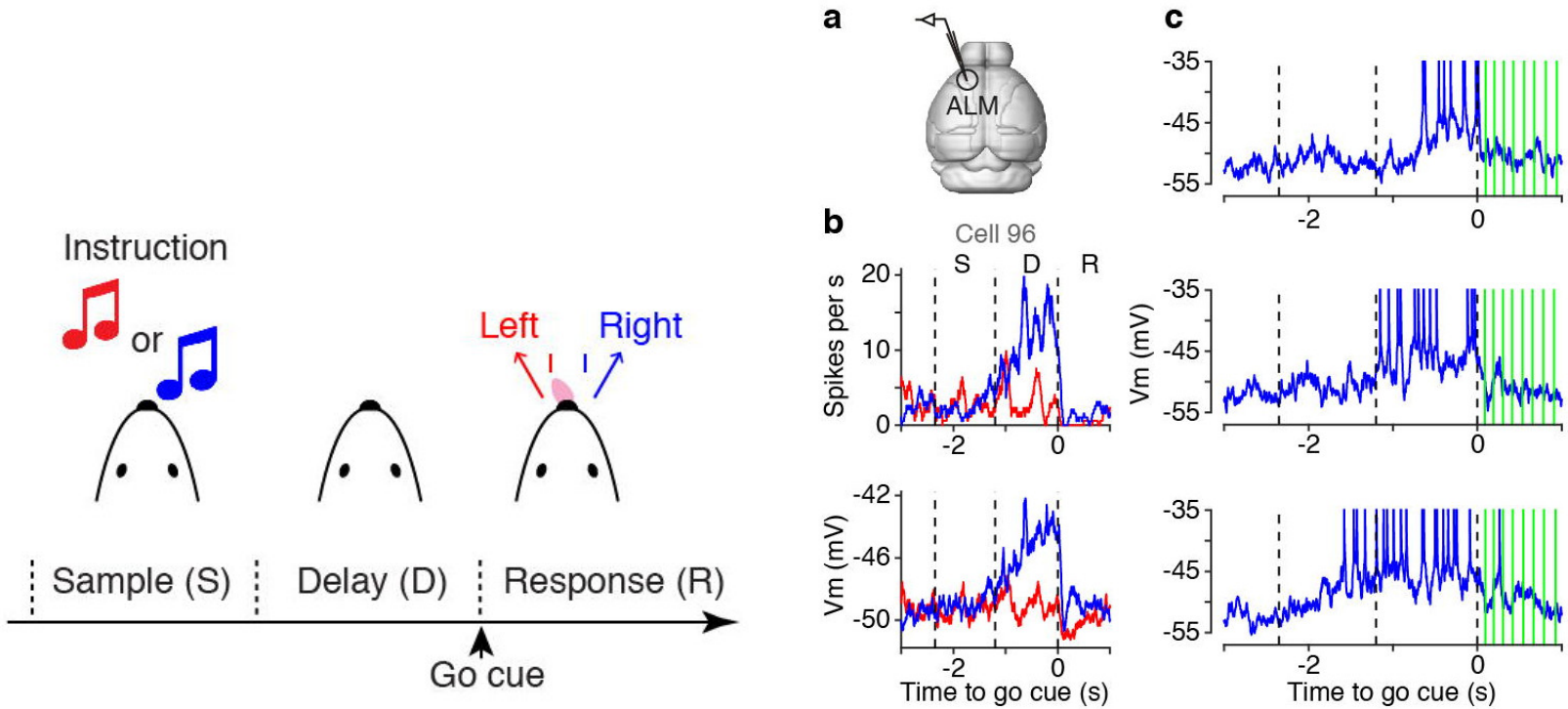

Nakamura and Kubota 1995

ITC, PRC, ERC, TPC, pictures




- **Consistent with attractor dynamics**

# Evidence for attractor dynamics in mice




- Perturbation experiments: Response consistent with attractor dynamics

# Outline

1. Introduction: The Hebbian/Attractor Neural Network scenario
2. A brief overview of the relevant neurobiology
3. **The Hopfield approach**
4. The Gardner approach
5. Open questions

# Attractor networks: the Hopfield model (1982)

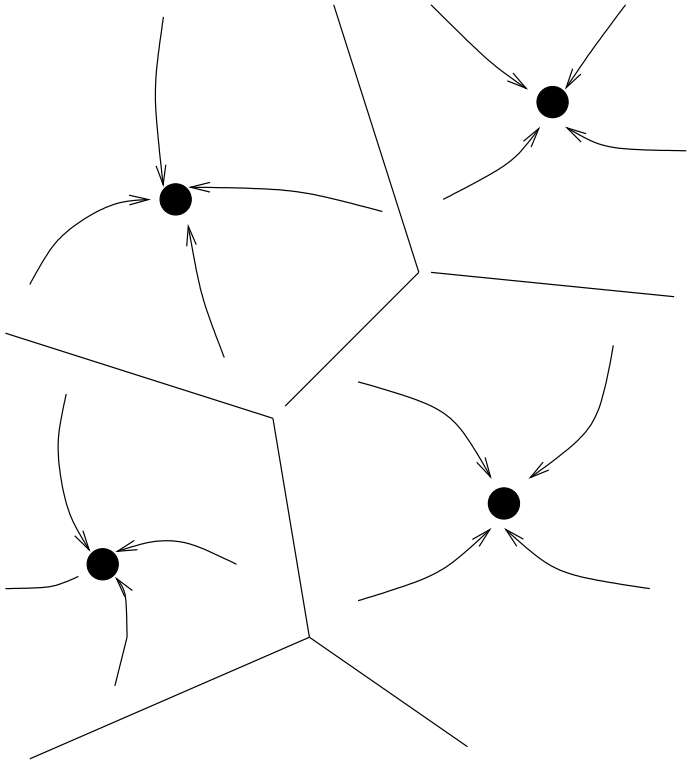
- Fully connected network of $N$ binary neurons ($S_i(t) = \pm 1$);
- Neuron dynamics (at zero temperature):

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

- How to store $p$ i.i.d. random patterns $\xi_i^\mu = \pm 1$ with prob. 0.5/0.5 as fixed point attractors?
- Use 'Hebbian' synaptic connectivity matrix

$$J_{ij} = \frac{1}{N} \sum_\mu \xi_i^\mu \xi_j^\mu$$

# Attractors in the Hopfield model



In the large $N$ limit:

- All patterns are attractors of the dynamics with prob 1 if

$$p < \frac{N}{4 \log N}$$

- A pattern is an attractor of the dynamics with prob 1 if

$$p < \frac{N}{2 \log N}$$

(Weisbuch and Fogelman-Soulie 1985, McEliece et al 1987)

- There exist stable fixed points close to the stored patterns if

$$p < \alpha_c N$$

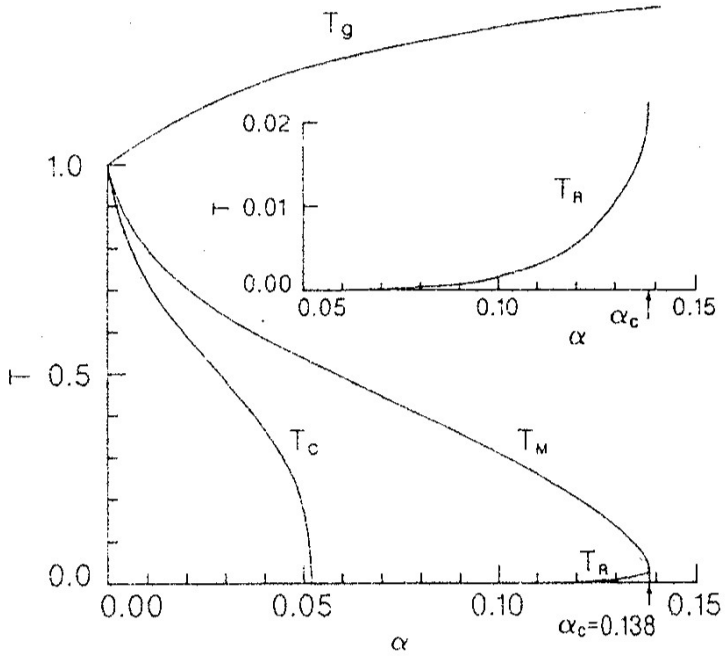
where $\alpha_c = 0.138$ (Amit, Gutfreund, Sompolinsky 1985, using the replica method)

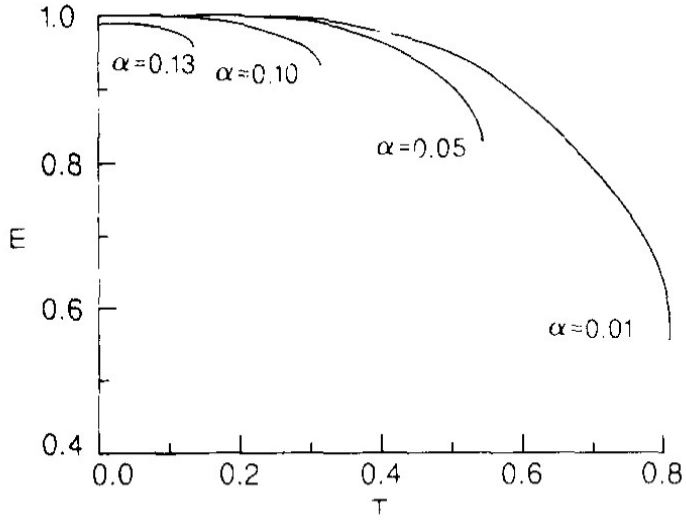
# Phase diagram

- Phase diagram can be obtained using the replica method (Amit, Gutfreund and Sompolinsky 1985)
- Order parameters quantifying quality of retrieval of memories: Overlaps with stored patterns

$$m_\mu = \frac{1}{N} \sum_i S_i \xi_i^\mu$$

- Phase diagram obtained using replica method (Amit, Gutfreund and Sompolinsky 1985)
- Basins of attraction of memories are enormous at small $\alpha$, vanish at maximal capacity



# The long road towards more realistic models

- Models with 0,1 neurons and sparse memories $\xi_i^\mu = 0, 1$ with prob $1 - f, f$ (Tsodyks and Feigel'man 1988)

$$J_{ij} = \frac{1}{f(1-f)N} \sum_\mu (\xi_i^\mu - f)(\xi_j^\mu - f)$$

  $\alpha_c \sim 1/(2f \log(1/f)$ in the sparse ($f \to 0$) coding limit
  Information stored per synapse increases with decreasing $f$ up to
  $1/(2 \log 2) \sim 0.721$ bits per synapse when $f \to 0$

- Models with highly diluted asymmetric connectivity (Derrida et al 1987)

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

  where $c_{ij} = 1, 0$ with probability $c, 1 - c$ and $c \to 0$
  $\alpha_c = p_{max}/C = 2/\pi \sim 0.64$ in the sparse connectivity limit

- 'Palimpsest' models (continuously learning new patterns, while forgetting old ones, Mézard et al 1986)

$$J_{ij} = \frac{1}{N} \sum_{\mu} \lambda^{\mu} \xi_i^{\mu} \xi_j^{\mu}$$

  Decrease in capacity (price to pay for being able to learn continuously)

- Models with discrete synapses

  Modest decrease in capacity (0.14 $\rightarrow$ 0.1 for binary synapses, Sompolinsky 1986)

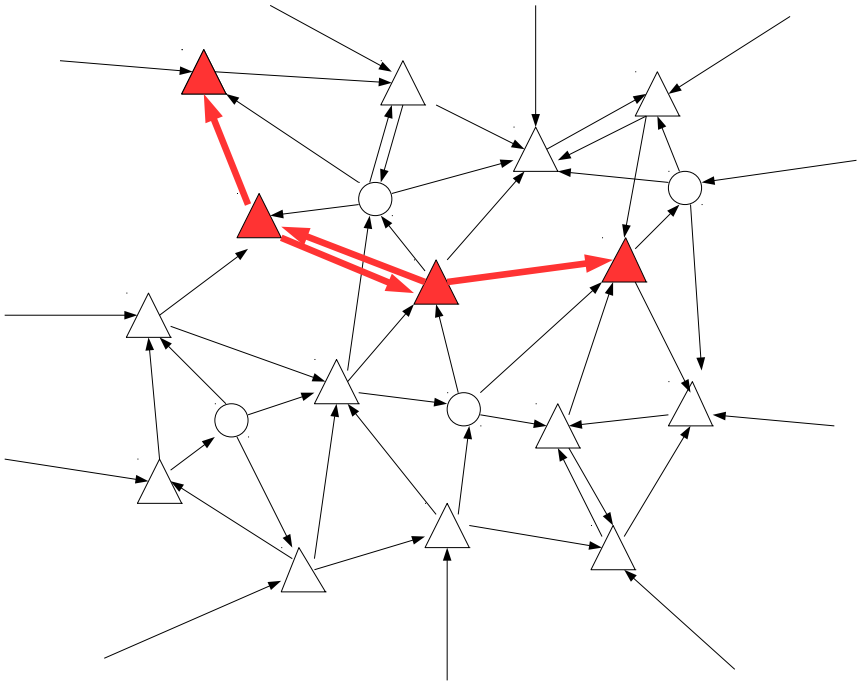
- Models with discrete synapses and one-shot learning

  $\Rightarrow$ Drastic decrease in capacity ($p \sim \sqrt{N}$, Tsodyks 1990, Amit and Fusi 1994), unless memories are sparse or synapses have a large number of states

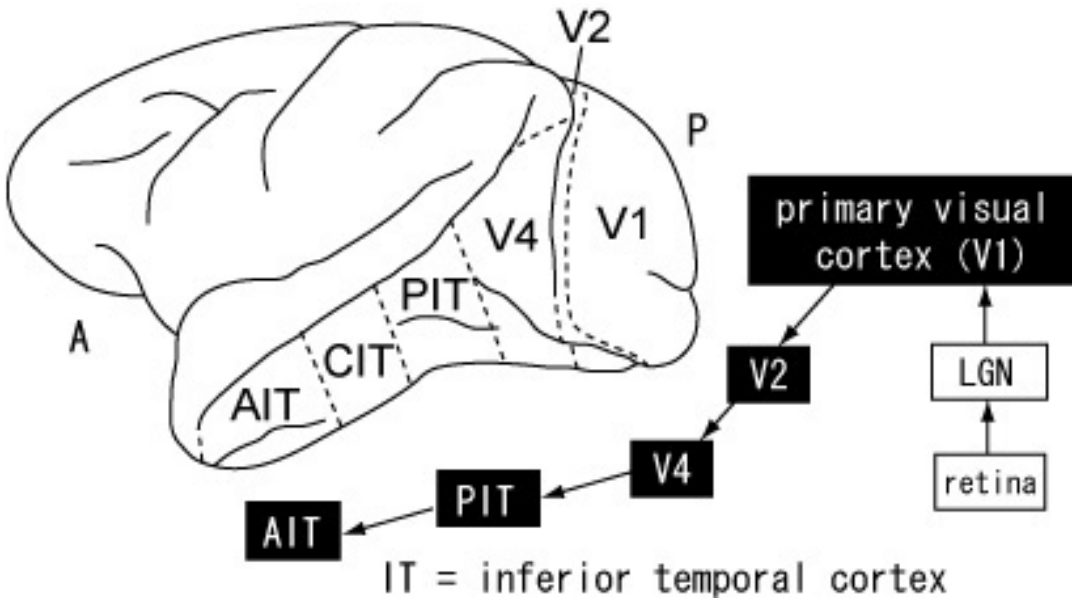
- Models with E-I separation and spiking neurons (Amit and Brunel 1997, Brunel and Wang 2001)

  Perform as ANNs provided specific conditions on connectivity are satisfied

# Inferring learning rules from in vivo data

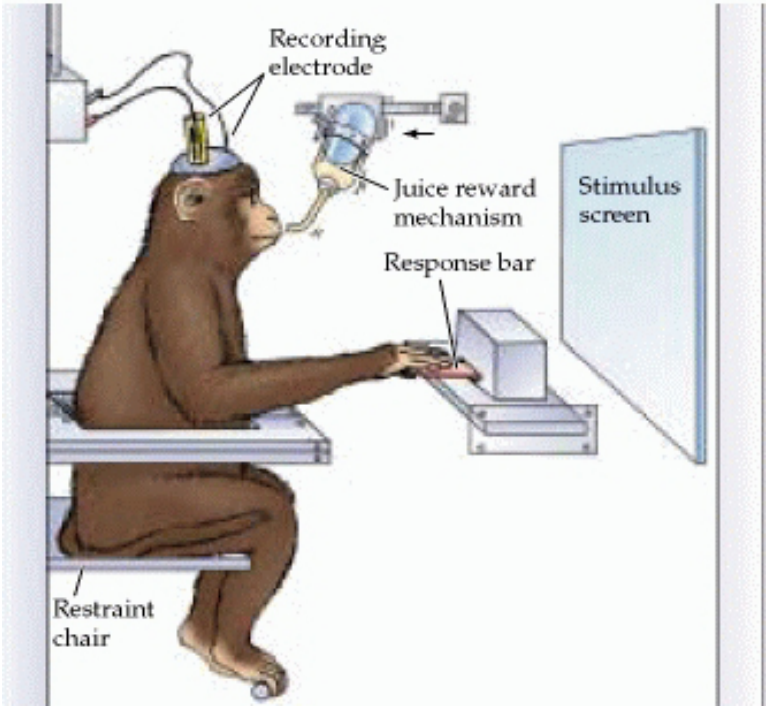


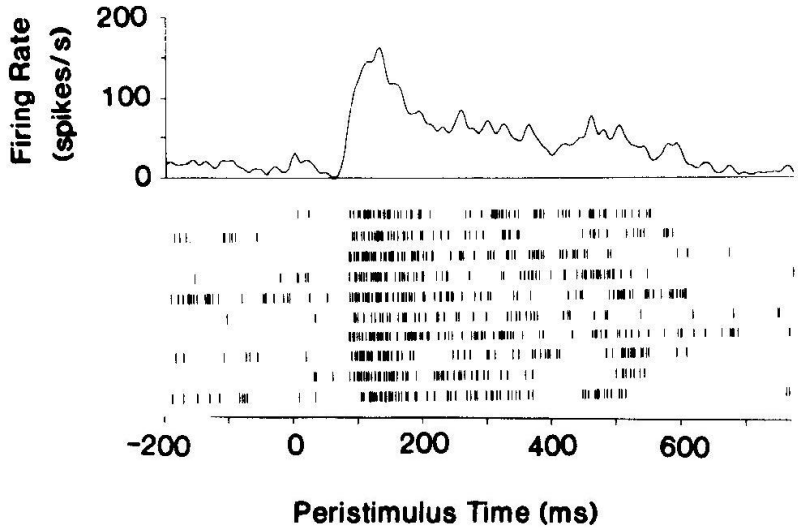
# Inferior temporal cortex (ITC)




- ITC: Where perception meets memory (Miyashita 1993)
- Last stage of purely visual processing
- Cells selective to complex visual features (faces, objects, etc)
- Stores long-term visual memories (lesion studies)
- Persistent activity has been observed in DMS tasks (Fuster, Miyashita, Desimone,...)

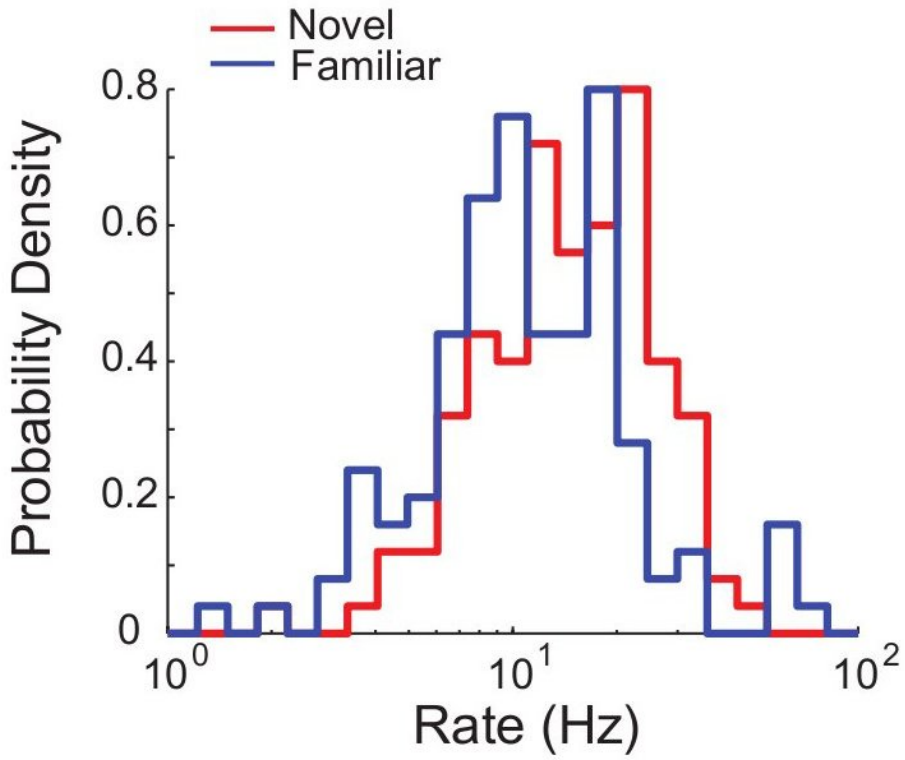
# Electrophysiological recordings in ITC of awake monkeys



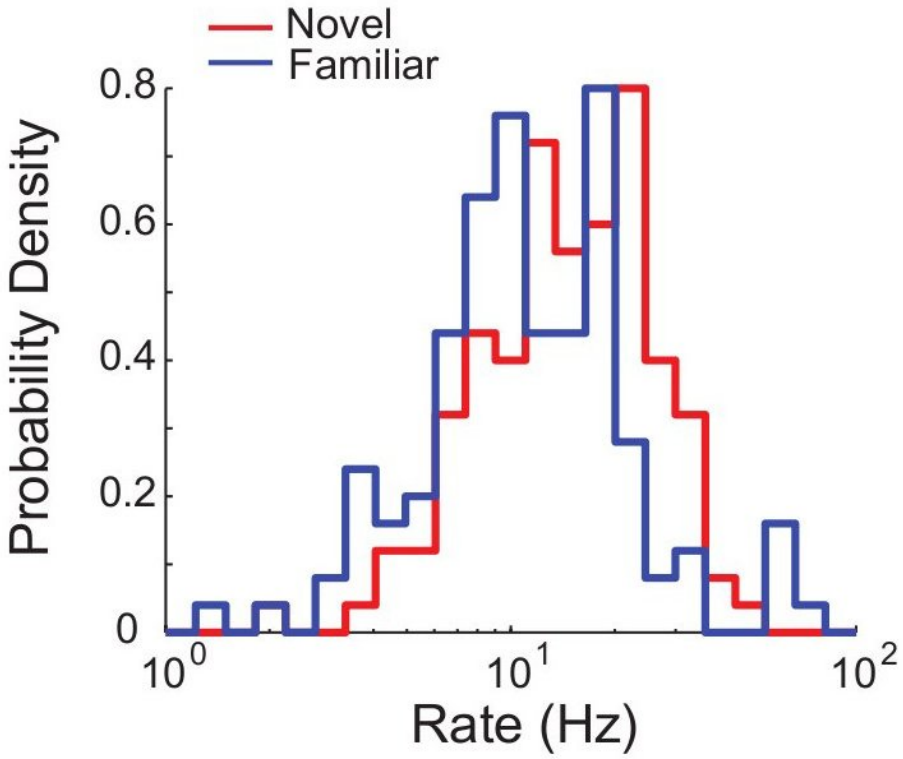

- How does neuronal activity change as an initially novel stimulus becomes progressively familiar?
- Data from Woloszyn and Sheinberg (2012): Use 125 novel/ 125 familiar images per session
- Focus on average visual response in the [75ms,200ms] interval

## Distribution of average visual responses for novel/familiar stimuli




- Mean(Familiar) $<$ Mean(Novel) for most neurons
- Best(Familiar) $>$ Best(Novel) for most 'putative' E neurons

# Questions




- Can we infer learning rule(s) from changes in firing rate distributions?
- Can learning rule(s) inferred from data generate attractors in recurrent networks?

# How do distributions of firing rates evolve with learning?

- Take a rate model, with $N \gg 1$ neurons described by a firing rate $r_i$;
- Total inputs to neuron $i$

$$h_i = I_{iX} + \frac{1}{N} \sum_j J_{ij} r_j$$

- Firing rate $r_i = \Phi(h_i)$
- When a novel stimulus is shown, $r_i = v_i$ where $v_i$ is drawn from $P_{nov}(v)$
- Induces changes in synaptic connectivity

$$J_{ij} \rightarrow J_{ij} + \Delta J(v_i, v_j)$$

- We assume $\Delta J(v_i, v_j) = f(v_i) g(v_j)$
- What is the new distribution of rates for the (now familiar) stimulus?

# How do distributions of rates evolve with learning?

- Firing rate of the (now familiar) stimulus $r_i = v_i + \Delta v_i$
- For small $\Delta J$, $\Delta v$, the changes in total input due to learning are

$$\begin{aligned} \Delta h_i &\approx \frac{1}{N}\sum_j \Delta J_{ij} v_j + \frac{1}{N}\sum_j J_{ij}\Delta v_j \\ &\approx f(v_i)\overline{g(v)v} + w\overline{\Delta v} \end{aligned}$$

Neuron-specific Global

- **Change in inputs $\Delta h_i$ depend on the visual response $v_i$, through $f$**

$\Rightarrow$ Change in rate of neuron $i$

$\Rightarrow$ Changes in the distribution of firing rates

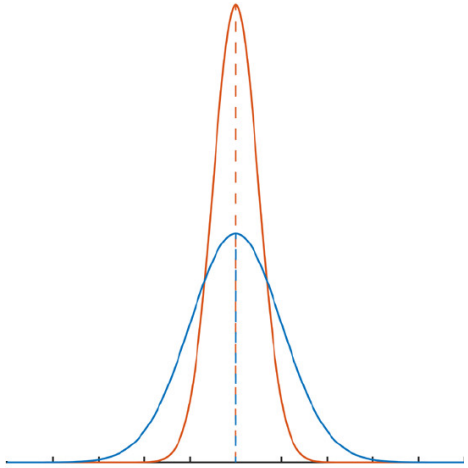
# Example: covariance learning rule

- Covariance learning rule:

$$\Delta J(v_i, v_j) = \alpha(v_i - \overline{v})(v_j - \overline{v})$$

- Changes in total inputs

$$\begin{aligned} \Delta h_i &\approx \frac{1}{N}\sum_j \Delta J_{ij} v_j + \frac{1}{N}\sum_j J_{ij}\Delta v_j \\ &\approx \alpha(v_i - \overline{v})\mathrm{Var}(v) \end{aligned}$$



- Linear f-I curve: no change in mean firing rates

$$\overline{\Delta v} = 0$$

- Individual neurons change their rates according to

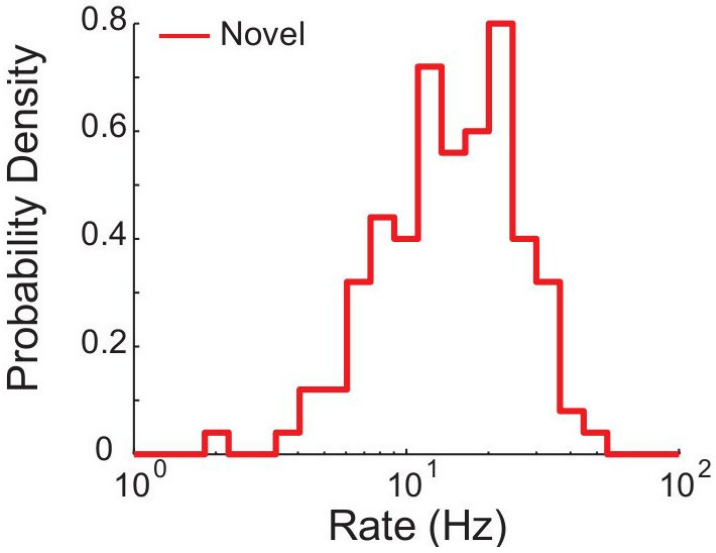
$$\Delta v_i = \alpha(v_i - \overline{v})\mathrm{Var}(v)$$

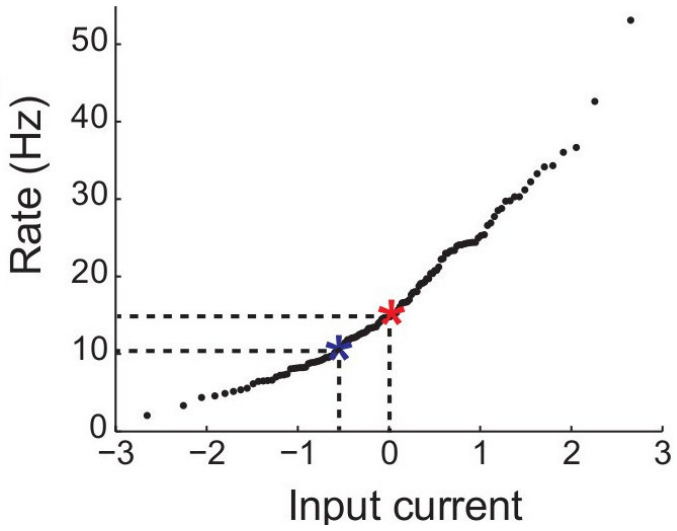
- Broadening of the distribution by a factor $1 + \alpha\mathrm{Var}(v)$
- With supra-linear f-I curves: **increase in mean firing rate**.
- **Covariance rule inconsistent with data**

# Inferring transfer function

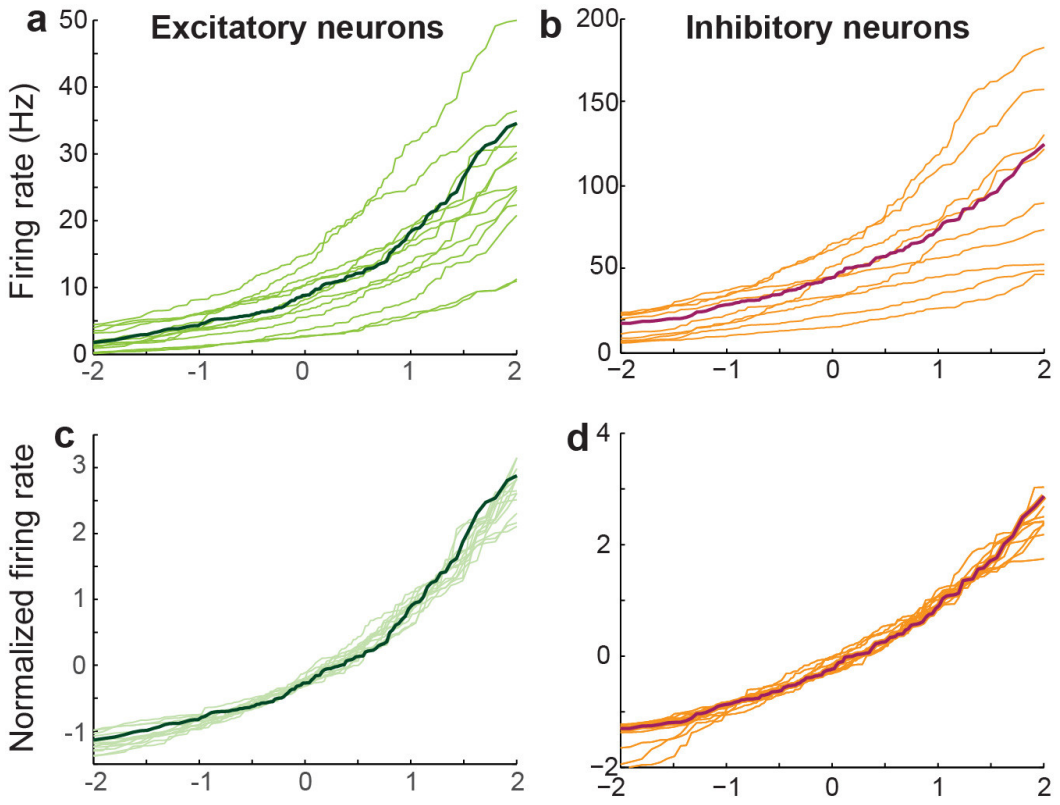
- Infer transfer function $\Phi$, from
  - Empirical distribution of rates for novel stimuli;
  - Assumed Gaussian distribution of inputs for novel stimuli

$P_{nov}(v_i)$ $\Phi(h_i)$



# Transfer functions of ITC neurons




- Supra-linear transfer functions, consistent with model and real neurons in fluctuation-driven regime

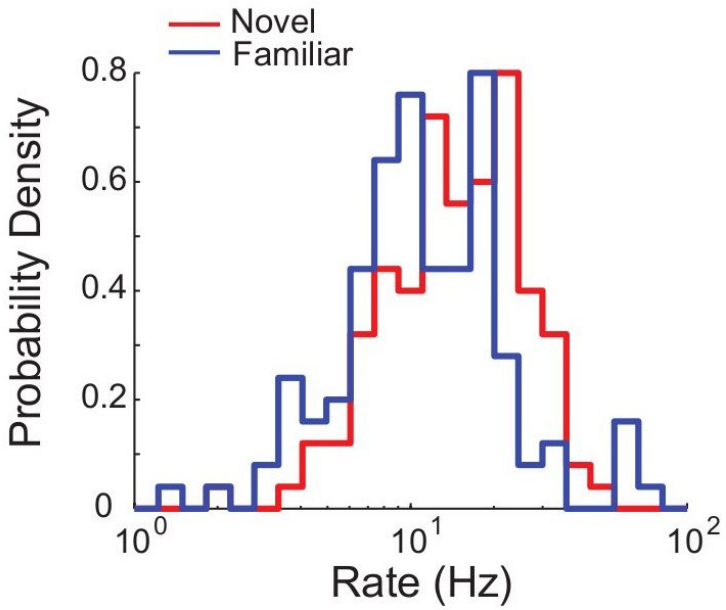
# Inferring learning rule

- Goal: Infer plasticity rule $\Delta J(v_i, v_j) = f(v_i)g(v_j)$ from $P_{nov}(v_i)$ and $P_{fam}(v_i)$?
- Assumptions:
  - Stationarity (currently familiar stimuli had, when they were novel, the same distribution as currently novel stimuli)
  - Learning rule preserves rank
- With these assumptions, it is possible to infer $f(v_i)$ - the dependence of the rule on the post-synaptic firing rate - from $P_{nov}(v_i)$ and $P_{fam}(v_i)$
- $g(v)$ undetermined, but it should be such that

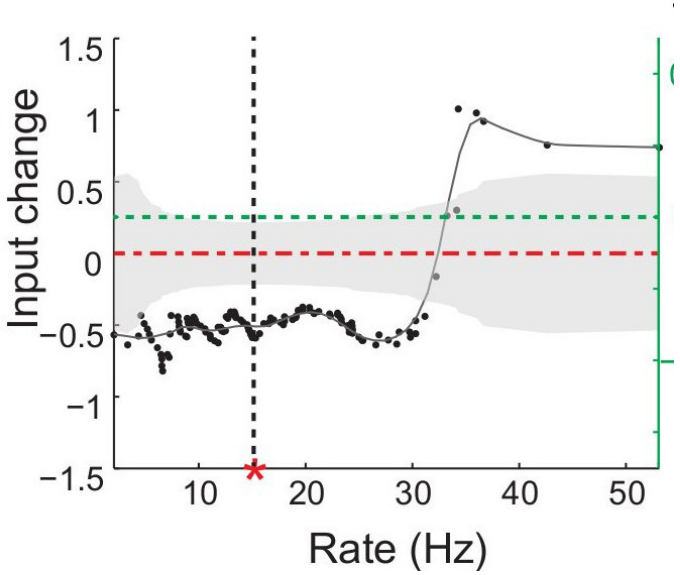
$$\int g(v)P_{nov}(v)dv = 0$$

$$\int g(v)vP_{nov}(v)dv > 0$$

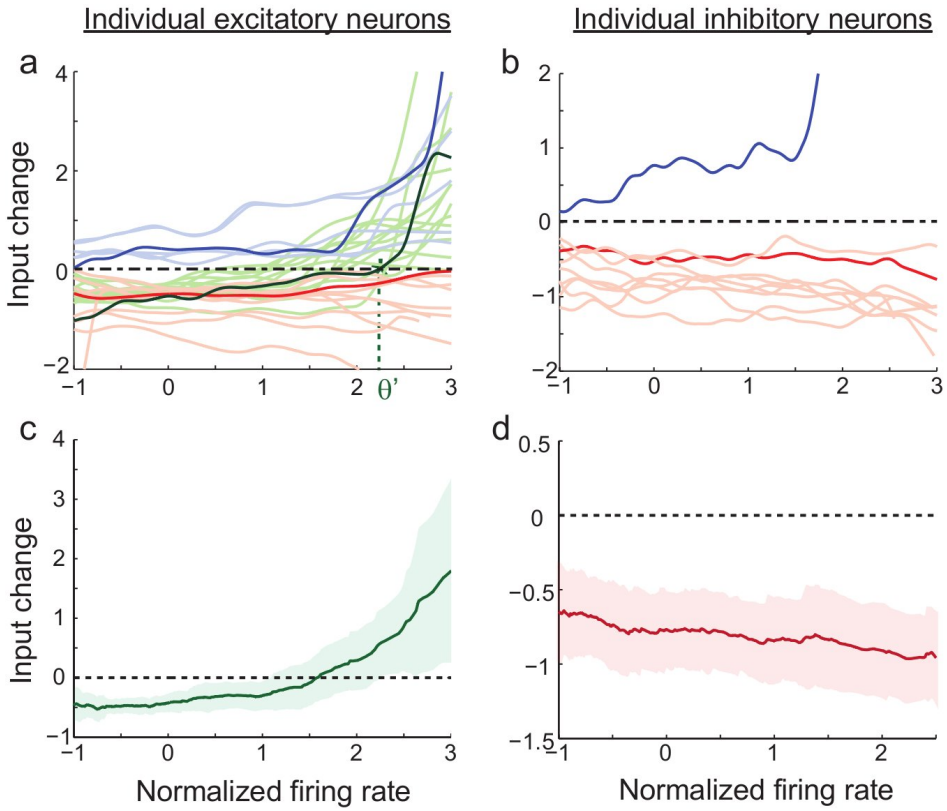
$P_{nov}(v_i), P_{fam}(v_i)$




$f(v_i)$



# Learning rules of individual ITC neurons




Simplest model that quantitatively describes data:

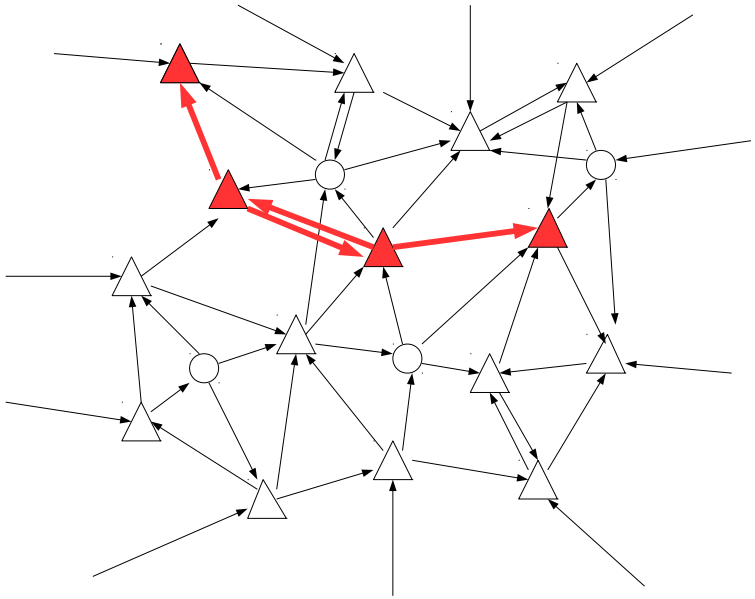
- Hebbian rule in approximately half of E$\rightarrow$ E synapses, whose dependence on post-synaptic firing rate is non-linear, and biased towards depression
- No plasticity in synapses involving I neurons

# Conclusions

- Inferred post-synaptic dependence of learning rule from in vivo data
- Data consistent with Hebbian plasticity in E neurons, no plasticity in I neurons;
- Firing rate dependence is consistent with a BCM rule
- Sparsening of representations in ITC
- Simple readout for stimulus familiarity (average network activity)

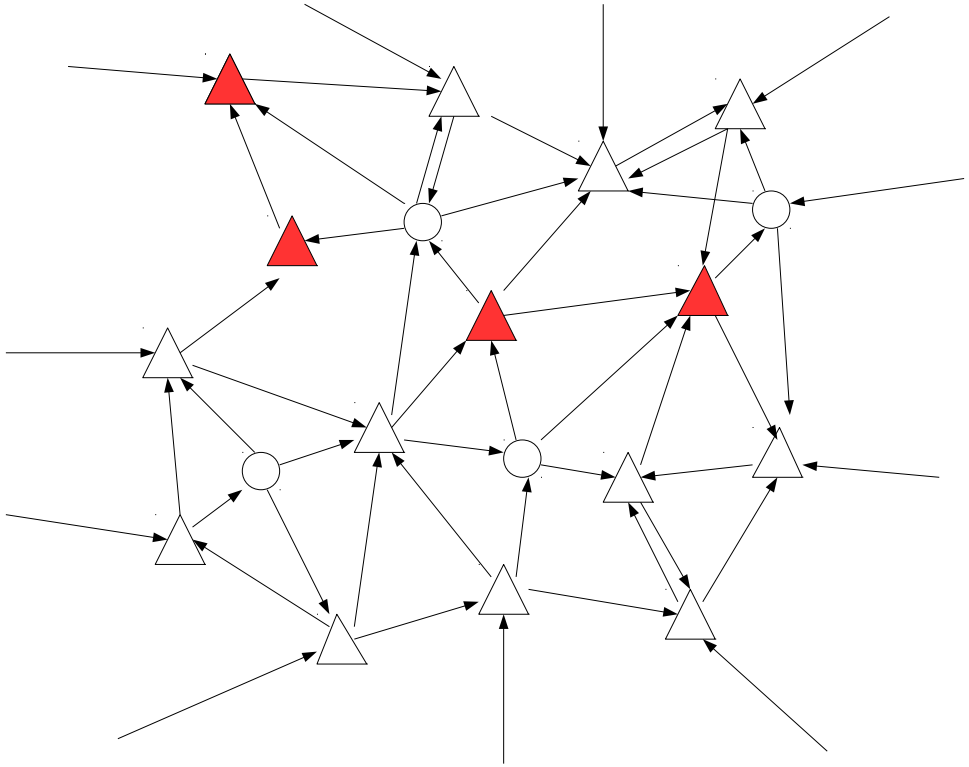
Lim, McKee, Woloszyn, Amit, Freedman, Sheinberg and Brunel (Nat. Neurosci. 2015)

# Dynamics of networks with learning rules inferred from data



- Data consistent with a non-linear Hebbian rule whose post-synaptic dependence is dominated by depression
- Does such a rule lead to attractor dynamics?
- What is the storage capacity of such a rule?

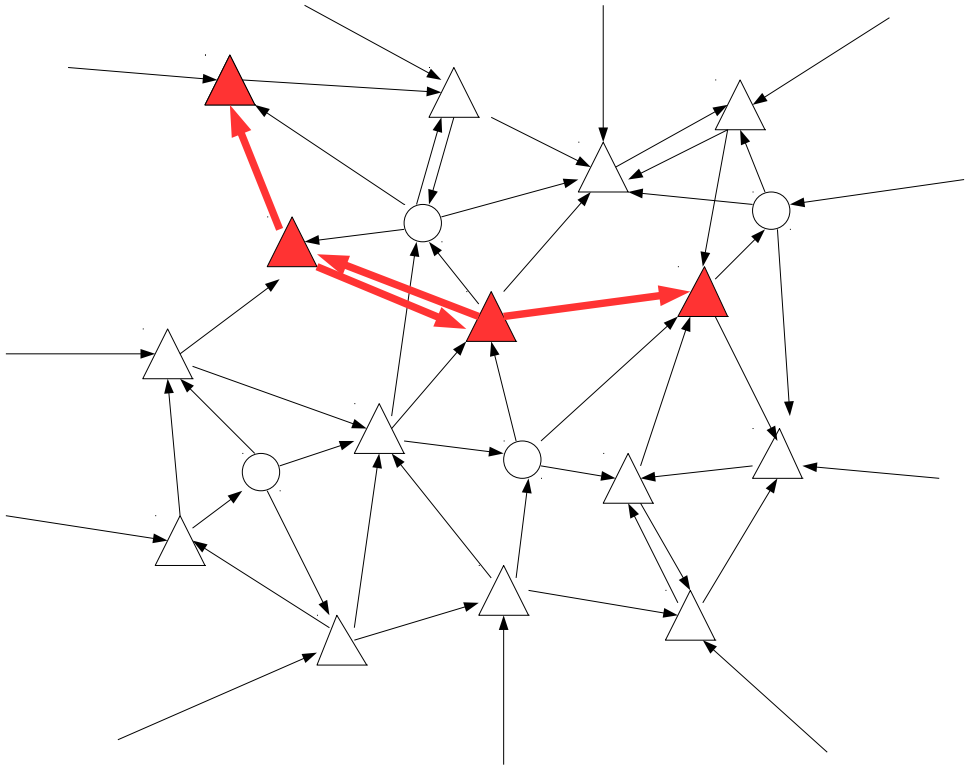
# Associative memory model constrained by data



- Generate $p$ i.i.d. Gaussian input patterns $\xi_i^\mu$

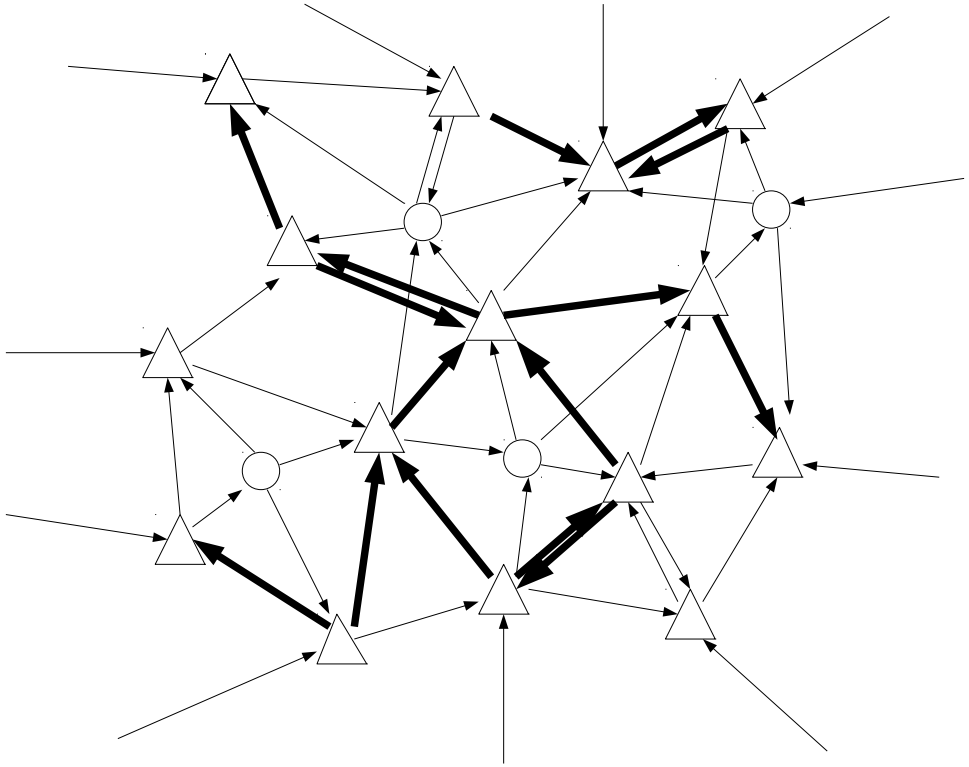
  $\Rightarrow$ With appropriate transfer function $\Phi$, distribution of firing rates automatically matches the data;

# Associative memory model constrained by data



- Learning rule $\Delta J_{ij} = f(r_i^\mu) g(r_j^\mu)$ inferred from data;

# Associative memory model constrained by data



- Final connectivity matrix

$$J_{ij} = \frac{c_{ij}}{cN} \sum_{\mu=1}^{p} f\left(r_i^{\mu}\right) g\left(r_j^{\mu}\right)$$

## Associative memory model with analog patterns and non-linear Hebbian rule

- $N$ neurons, whose firing rate obey

$$\tau \frac{dr_i}{dt} = -r_i + \Phi\left(I_i + \sum_{i \neq j}^{N} J_{ij} r_j\right)$$

- $p$ random uncorrelated Gaussian input patterns $\xi_i^\mu \sim \mathcal{N}(0,1)$
- Connectivity matrix

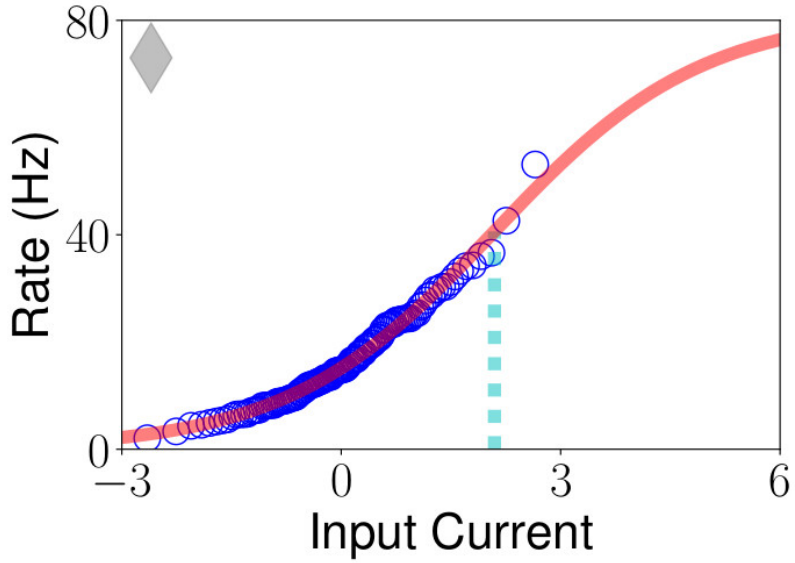
$$J_{ij} = \frac{c_{ij}}{cN} \sum_{\mu=1}^{p} f\left(\Phi(\xi_i^\mu)\right) g\left(\Phi(\xi_j^\mu)\right)$$

where

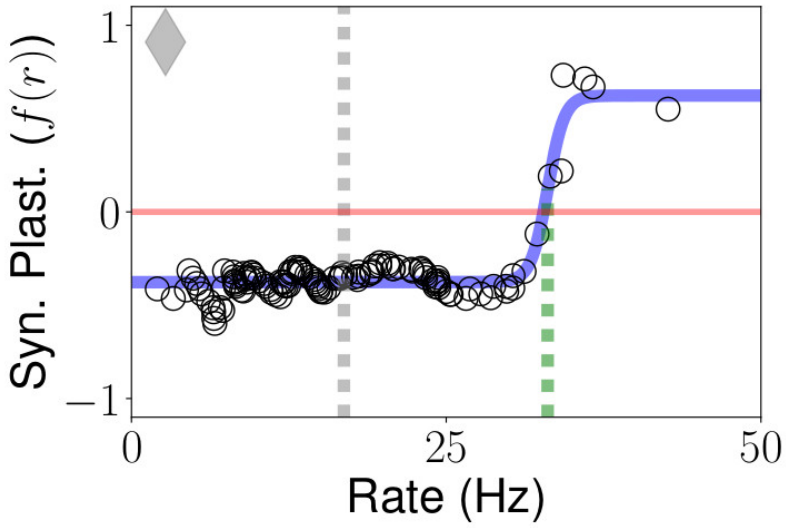
$c_{ij} =$ ER 'structural' connectivity matrix ($c_{ij} = 1$ with prob. $c \ll 1$), $g$ should be such that $\int Dx g(\Phi(x)) = 0$, $\int Dx g(\Phi(x))\Phi(x)dx > 0$.

# Transfer functions and learning rules inferred from data

**Transfer function $\Phi$**




**Post dependence of learning rule $f$**




- Fit both $\Phi$ and $f$ by sigmoidal functions, for all neurons with significant 'Hebbian' plasticity rules;
- Take $g$ as a sigmoidal function, with threshold and gain identical to $f$, and offset set such that $\int Dx g(\Phi(x)) = 0$
- Simulate and analyze the dynamics of a network with median parameters

# Mean-field theory

- Can the network retrieve a stored pattern (i.e. converge to an attractor that is correlated with the pattern)?
- Define order parameters

$$m = \langle \frac{1}{N} \sum_i \tilde{g}(\xi_i^1) r_i \rangle \qquad \text{(Overlap with pattern)}$$

$$\sigma^2 = \langle \frac{1}{N^2} \sum_{\mu>1,j} \tilde{f}^2(\xi_i^\mu) \tilde{g}^2(\xi_j^\mu) r_j^2 \rangle \qquad \text{(Quenched noise due to other patterns)}$$

- In the limits $N \to \infty$, $N \gg cN \gg 1$, $p \sim cN$, order parameters given by MF equations

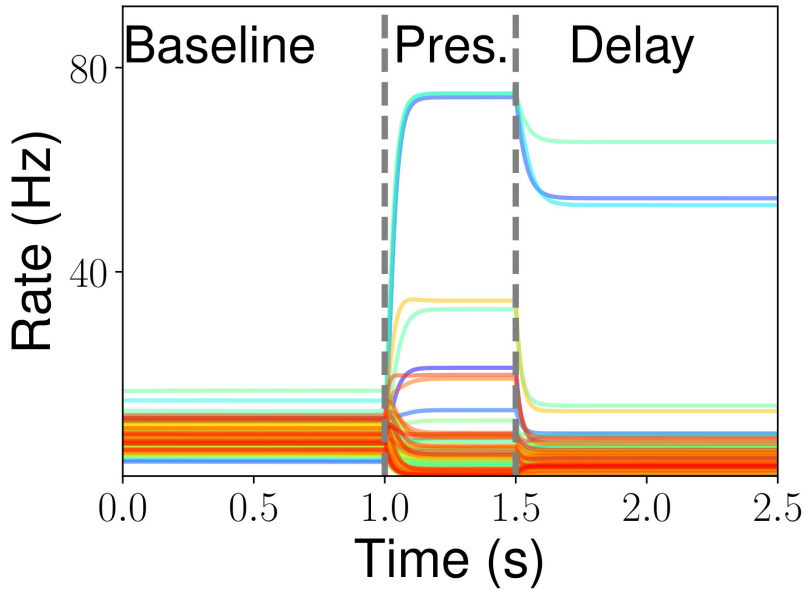
$$m = \int D\xi Dz \tilde{g}(\xi) \Phi(\tilde{f}(\xi) m + \sigma z)$$

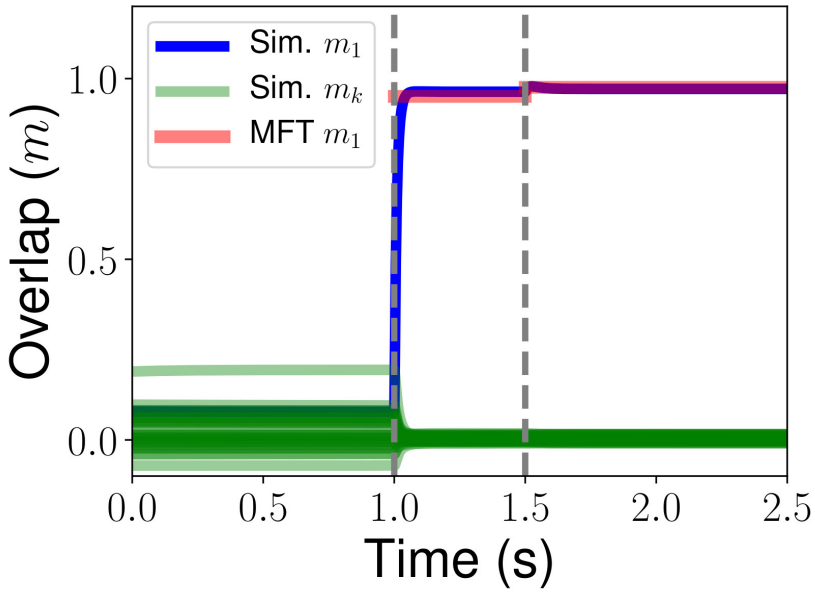
$$\sigma^2 = \alpha \int D\xi \tilde{f}^2(\xi) \int D\xi \tilde{g}^2(\xi) \int D\xi Dz \Phi^2(\tilde{f}(\xi) m + \sigma z)$$

  where $\alpha = p/(cN)$.

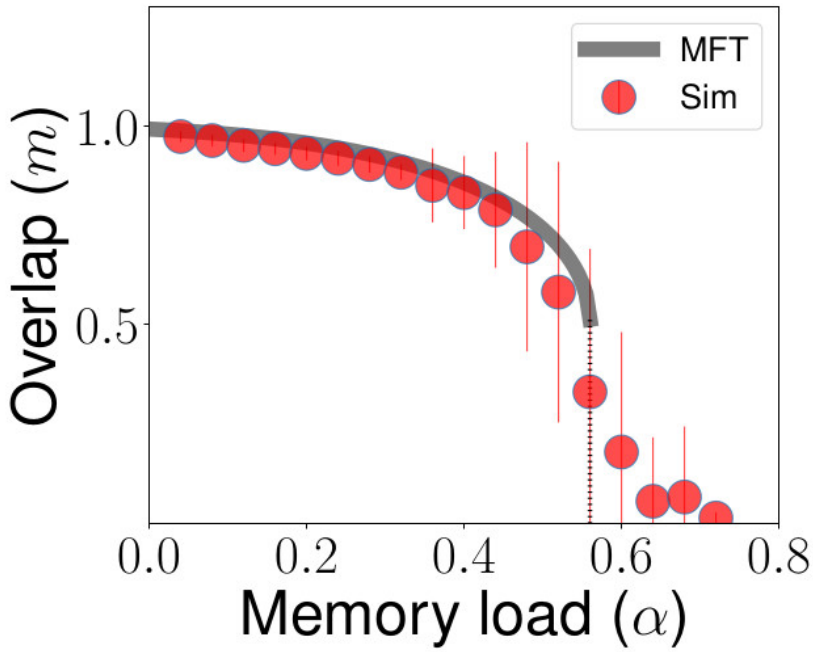
- Retrieval states: Solutions such that $m > 0$;
- Storage capacity: largest $\alpha$ for which retrieval states exist.

# Learning rules inferred from data lead to attractor dynamics and delay period activity



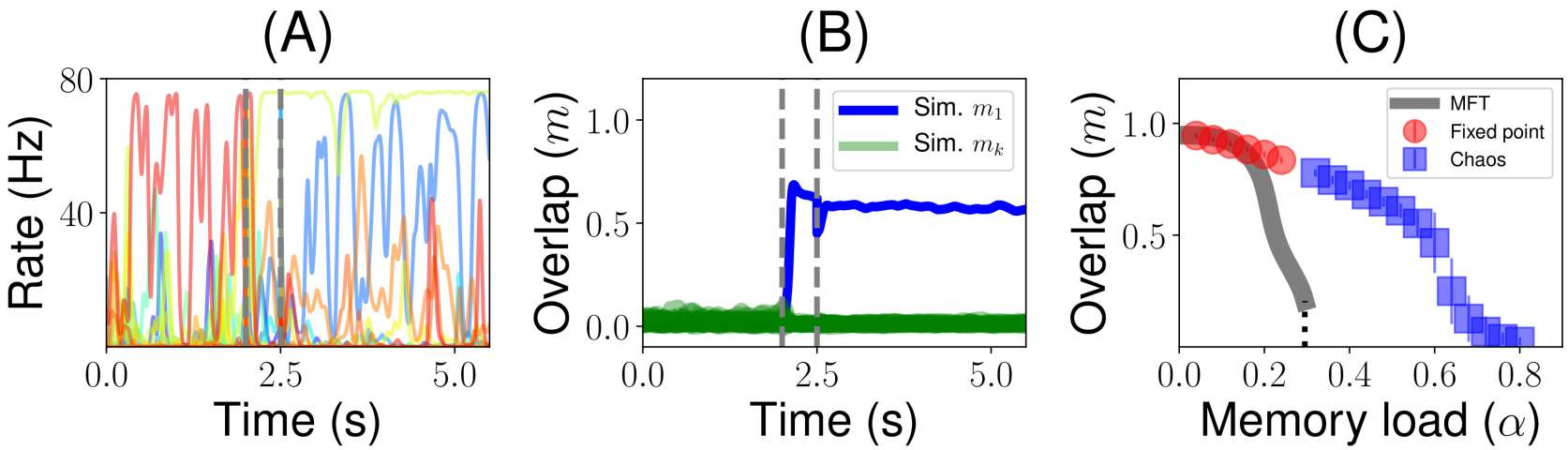



# Storage capacity




- Storage capacity for median parameters close to 0.6;
- Close to optimal capacity ($\alpha_{max} \sim 0.8$), in the space of sigmoidal functions $f$ and $g$.
- Optimal learning rule in such a space: Both $f$ and $g$ are step functions with high thresholds $\Rightarrow$ Tsodyks-Feigel'man model

# Transition to chaos at strong coupling




- Increasing coupling strength leads to chaotic retrieval states
- Similar to chaotic states in simpler asymmetric rate models (Sompolinsky et al 1988, Tirozzi and Tsodyks 1991)
- Reproduces strong irregularity and diversity of temporal profiles of activity seen in delay periods in PFC

# Conclusions

- Network model with distribution of patterns and learning rule inferred from data exhibits attractor dynamics
- Learning rule inferred from data close to optimal in terms of storage capacity (in the space of Hebbian learning rules with sigmoidal dependence on pre and post rates)
- Transition to chaos at sufficiently strong coupling - leads to strong irregularity and diversity of temporal profiles of activity in the delay period, similar to observations in PFC

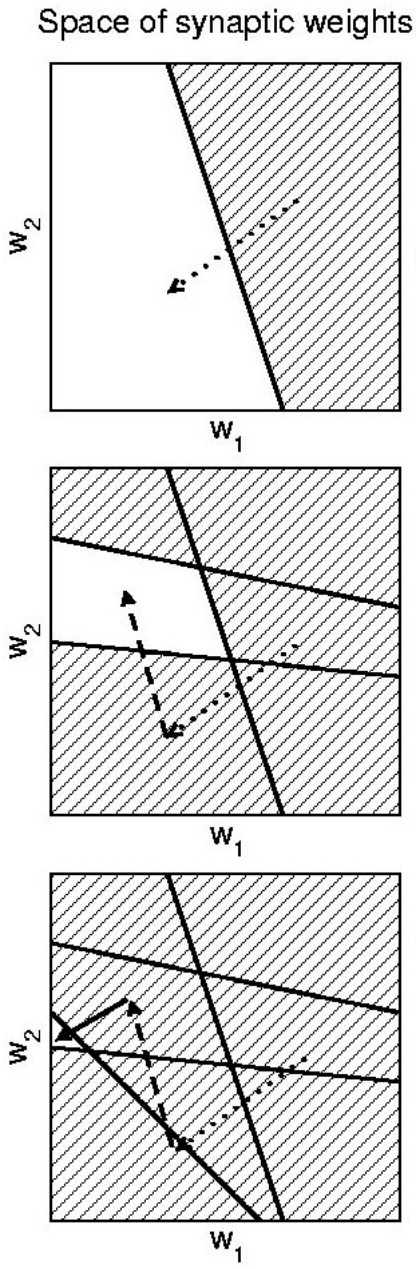
Pereira and Brunel (Neuron 2018)

# Outline

1. Introduction: The Hebbian/Attractor Neural Network scenario
2. A brief overview of the relevant neurobiology
3. The Hopfield approach
4. **The Gardner approach**
5. Open questions

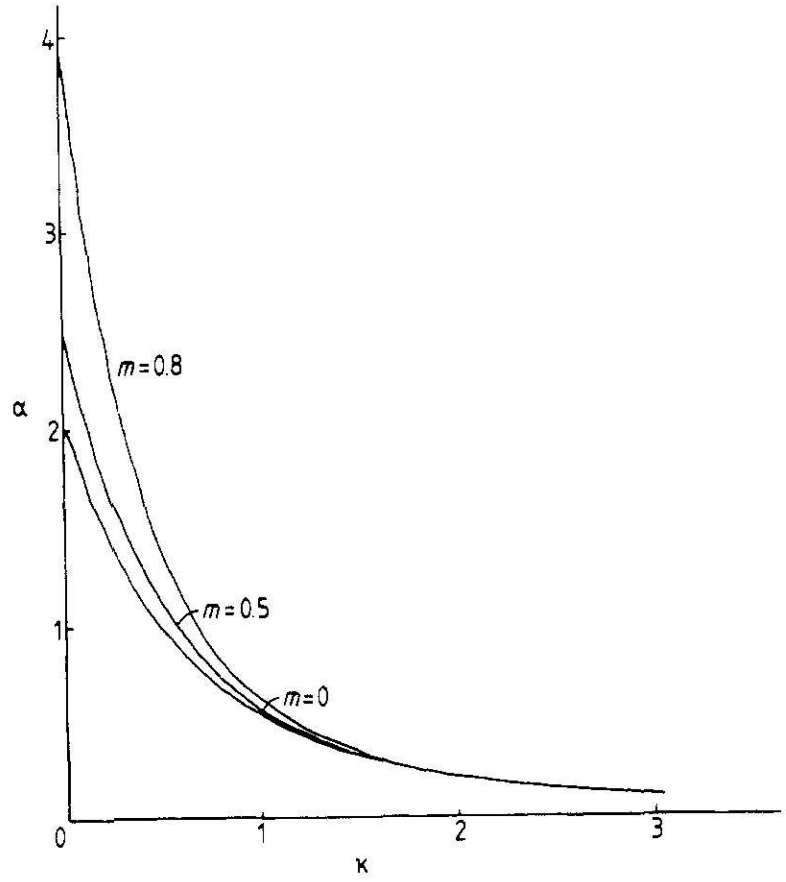
# Gardner approach (1988)

- Instead of focusing on specific learning rules, study space of all possible connectivity matrices that store a given set of fixed point attractors, with a given robustness level
- In the case of networks of binary neurons, each neuron has to solve its own perceptron problem - find a hyperplane separating two sets of patterns, those in which it should be active/inactive
- Compute the typical volume of the subspace of solutions using replica method (Gardner 1988);
- Storage capacity obtained when volume goes to zero;
- This storage capacity is an upper bound valid for all possible learning rules.
- Alternative derivation using the cavity method (Mézard 1989)



# Storage capacity of the perceptron

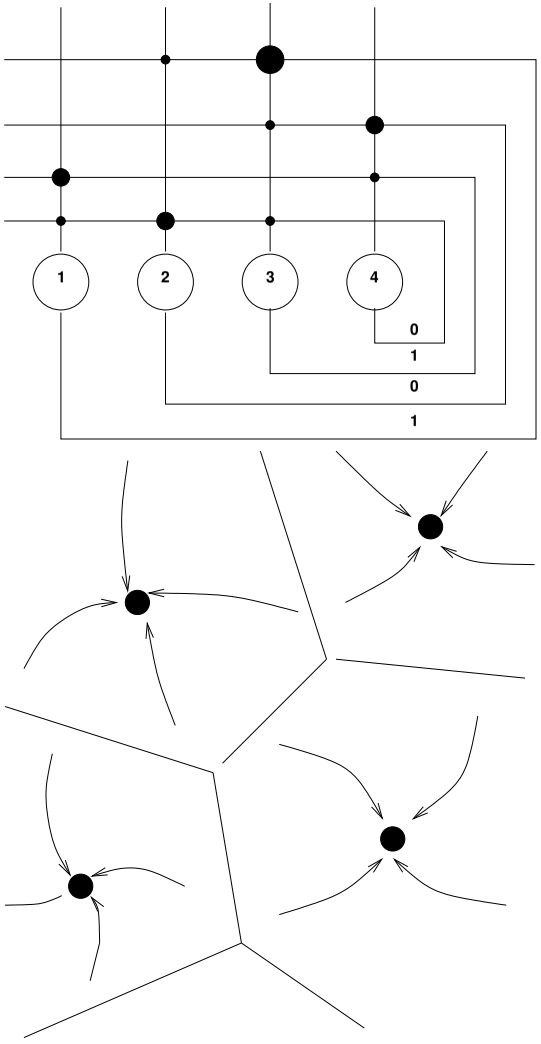
- How many random associations can a perceptron with $N$ binary inputs learn, in the large $N$ limit?
- Answer: $p = \alpha N$, where $\alpha$ stays finite in the large $N$ limit
  Unconstrained weights, $f' = 0.5$, $\kappa = 0$
  $\Rightarrow \alpha = 2$ (Cover 1965, Gardner 1988)
- Tradeoff between capacity and robustness (Gardner 1988)
- Sign-constrained synapses:
  $\Rightarrow \alpha = 1$ (Amit et al 1989)
- Capacity increases with decreasing output coding level, but information content decreases (Gardner 1988)



## Questions

- What is the distribution of synaptic weights in the volume of solutions?
- Other statistics of synaptic connectivity matrix like joint distributions of pairs of weights, motifs?
- Compare with available data

# Set-up




- Fully connected network of $N \gg 1$ binary excitatory neurons;

$$S_i(t+1) = \Theta\left(\sum_j J_{ij} S_j(t) - \theta\right)$$

  where $J_{ij} \geq 0$ for all $i, j$

- Constraints on synaptic connectivity: the network should have a large number ($p \sim O(N)$) of fixed point attractor states $S_i = \xi_i^\mu$ (stable representations of external stimuli)
- Each attractor state: random binary pattern, coding level $f$

$$P(\xi_i^\mu = 1) = f, \quad P(\xi_i^\mu = 0) = 1 - f$$

- Robustness level $\kappa$ (measures size of basin of attraction of each attractor);

# Constraints

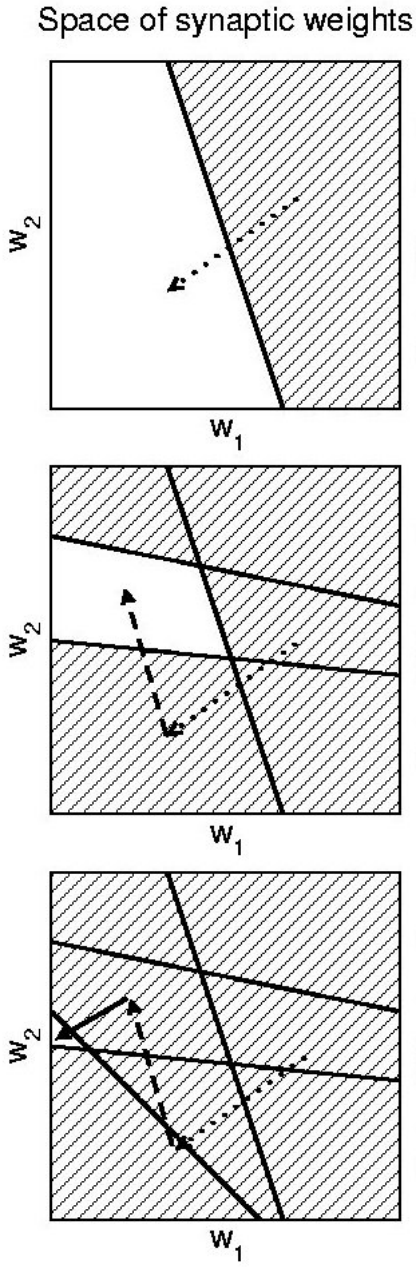
- Define 'stabilities' $\Delta_i^\mu$ as

$$\Delta_i^\mu = \frac{(2\xi_i^\mu - 1)}{\sqrt{N}} \left( \sum_j J_{ij} \xi_j^\mu - \theta \right)$$

- Subspace of solutions defined by

$$\begin{aligned} \Delta_i^\mu &\geq K \quad \text{for all } i, \mu \\ J_{ij} &\geq 0 \quad \text{for all } i \neq j \end{aligned}$$

- Goal: Compute distributions of $\Delta$s and $J$s



## Computing statistics of connectivity using the cavity method

- Follow Mézard (1989)
- Assume that we have already learned $p$ patterns, in a network of $N$ neurons;
- Add one pattern and learn it; compute the distribution of 'stabilities' of this pattern. This distribution is a function of the distribution of synaptic weights.
- Add one synaptic weight: compute the distribution of this new weight, as a function of the distribution of stabilities.
- Leads to self-consistent equations for parameters of both distributions.
- Add $n \geq 2$ synaptic weights: compute joint distributions of weights/probabilities of motifs.

# Add a new pattern, and learn it

- Add a new randomly drawn pattern $\vec{\xi}$;
- Consider neuron $i$, and define $\tilde{\xi} = (2\xi_i - 1)$
- Associated stability:

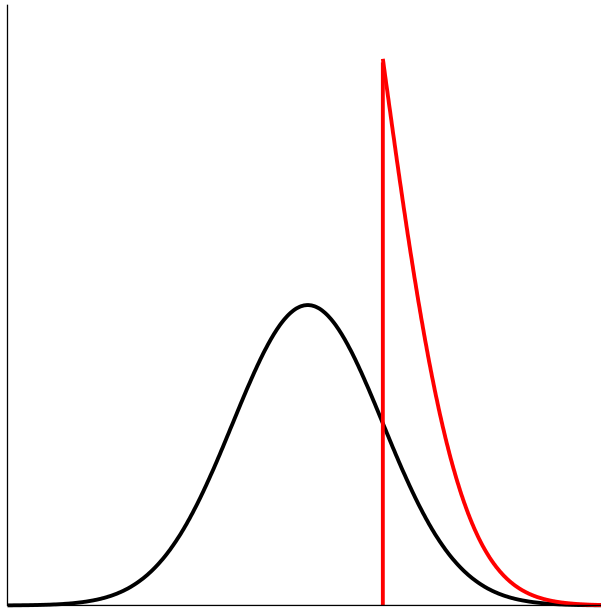
$$\Delta = \frac{\tilde{\xi}}{\sqrt{N}} \left( \sum_j J_{ij}\xi_j - N\theta \right)$$

- Distribution of $\Delta$ over the space of weights satisfying the previously learned pattern is Gaussian, with moments

$$\begin{aligned} h &= \frac{\tilde{\xi}}{\sqrt{N}} \sum_j \langle J_{ij} \rangle (\xi_j - f) + \tilde{\xi} f M \\ \sigma_h^2 &= \frac{1}{N} \sum_j \left( \left\langle J_{ij}^2 \right\rangle - \langle J_{ij} \rangle^2 \right) \left( \xi_j (1-2f) + f^2 \right)] \end{aligned}$$

- Learning the pattern $=$ removing from the space of weights those for which $\Delta < \kappa \Rightarrow$ leads to a truncated Gaussian,

$$P(\Delta, h) = \frac{1}{\sigma_h} \frac{\exp\left( -\frac{1}{2} \left( \frac{\Delta - h}{\sigma_h} \right)^2 \right)}{H\left( \frac{\kappa - h}{\sigma_h} \right)} \Theta(\Delta - \kappa)$$



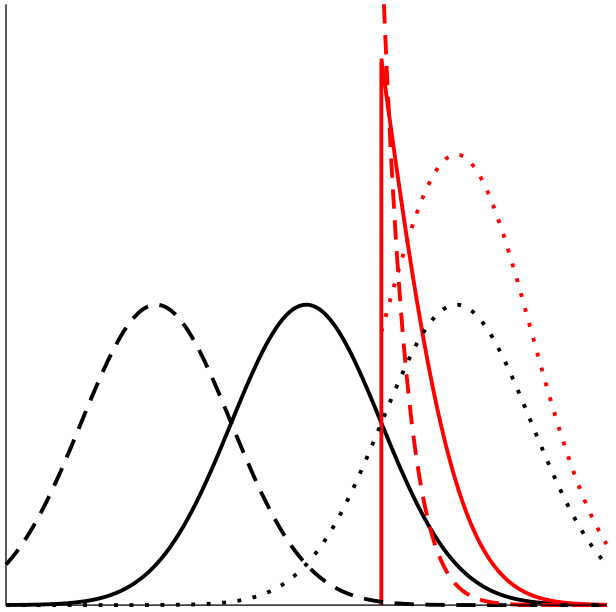
# Averaging over distribution of patterns

- Average over distribution of patterns gives:

$$\begin{aligned} \overline{h} &= \tilde{\xi} f M \\ \overline{\left(h - \tilde{\xi} f M\right)^2} &= q f (1-f) \\ \overline{\sigma_h^2} &= f(1-f)(Q-q) \end{aligned}$$

where

$$\begin{aligned} \frac{1}{N} \sum_j \overline{\langle J_{ij}^2 \rangle} &= Q \\ \frac{1}{N} \sum_j \overline{\langle J_{ij} \rangle^2} &= q \end{aligned}$$

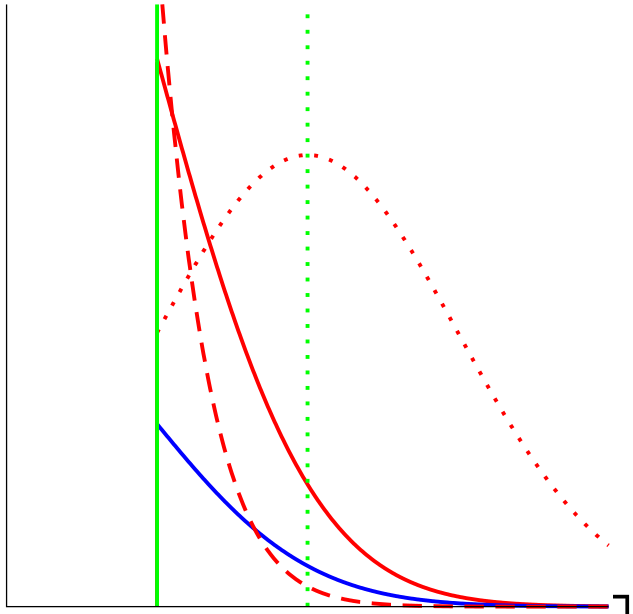


- Pattern-averaged distribution of stabilities

$$\overline{P(\Delta)} = \sum_{\tilde{\xi}=\pm 1} p_{\tilde{\xi}} \int \frac{dh}{\sqrt{2\pi q f (1-f)}} \exp\left( -\frac{1}{2} \left( \frac{h - \tilde{\xi} f M}{\sqrt{q f (1-f)}} \right)^2 \right) P(\Delta, h)$$

# Distribution of stabilities at maximal capacity

- At maximal capacity, space of weights satisfying constraints shrink to zero.
- In this limit, $\sigma_h \to 0$, $q \to Q$
- Truncated Gaussian distributions $P(\Delta, h)$ converge to delta functions, centered either on $\kappa$ (if $h < \kappa$) or $h$ (if $h > \kappa$)
- Pattern-averaged distribution becomes:



$$P(\Delta) = \sum_{\tilde{\xi}} p_{\tilde{\xi}} \left[ \frac{\exp\left( -\frac{1}{2} \left( \frac{\Delta - \tilde{\xi} f M}{\sqrt{f(1-f)Q}} \right)^2 \right)}{\sqrt{2\pi f(1-f)Q}} \Theta(\Delta - K) + H(\tau_{\tilde{\xi}}) \delta(\Delta - K) \right]$$

Kepler and Abbott 1988, Krauth et al 1988

# Adding one synapse and computing its distribution

- Add a single input, $i = 0$ to neuron $i$, with an associated weight $J_0$. Values of the patterns on this new input are $\xi_0^\mu$, $\mu = 1, \ldots, p$.
- This changes slightly the stability of all the patterns: $\Delta^\mu \to \Delta^\mu + \epsilon^\mu$ where

$$\epsilon^\mu = \frac{\tilde{\xi}^\mu}{\sqrt{N}} J_0 (\xi_0^\mu - f)$$

- The distribution of weights $J_0$ satisfying all the $p$ constraints is

$$Q(J_0) \propto \int \prod_\mu P(\Delta^\mu, h^\mu) d\Delta^\mu \Theta\left(\Delta^\mu + \epsilon^\mu - \kappa\right) \Theta(J_0)$$

- In the large $N$ limit:

$$Q(J_0) \propto \exp\left((a-c)J_0 - \frac{b}{2}J_0^2\right)\Theta(J_0)$$

with

$$\begin{aligned} a &= \frac{1}{\sqrt{N}}\sum_\mu \tilde{\xi}^\mu(\xi_0^\mu - f)P(\kappa, h^\mu) \\ b &= \frac{1}{N}\sum_\mu \left(\xi_0^\mu(1-2f)+f^2\right)\left(P(\kappa,h^\mu)^2 + \frac{\partial P}{\partial \Delta}(\kappa, h^\mu)\right) \end{aligned}$$

Again, a truncated Gaussian...

# Averaging (again) over the distribution of patterns

- Compute moments of $a$ and $b$ over the distribution of patterns,

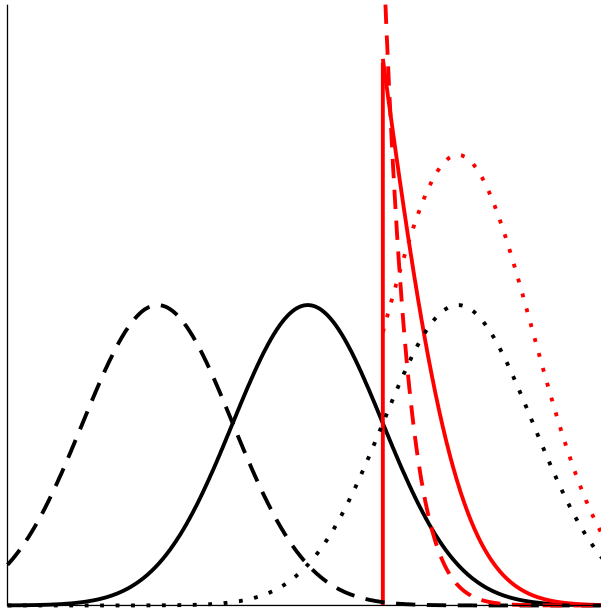
$$\sigma_a^2 = \alpha f(1-f) \sum_{\tilde{\xi}} p_{\tilde{\xi}} \int Du \frac{1}{\sigma_h^2} \frac{G\left(a_{\tilde{\xi}}(u)\right)^2}{H\left(a_{\tilde{\xi}}(u)\right)^2}$$

$$\overline{b} = \alpha f(1-f) \sum_{\tilde{\xi}} p_{\tilde{\xi}} \int Du \frac{1}{\sigma_h^2} \frac{G\left(a_{\tilde{\xi}}(u)\right)^2}{H\left(a_{\tilde{\xi}}(u)\right)^2} - \frac{a_{\tilde{\xi}}(u)}{\sigma_h^2} \frac{G\left(a_{\tilde{\xi}}(u)\right)}{H\left(a_{\tilde{\xi}}(u)\right)}$$

$$a_{\tilde{\xi}}(u) = \frac{\kappa - \tilde{\xi} f M + u\sqrt{qf(1-f)}}{\sqrt{f(1-f)(Q-q)}}$$

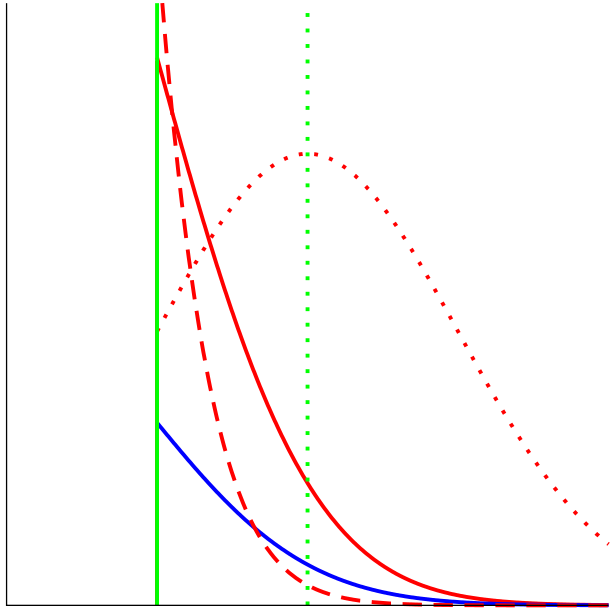
- The averaged distribution of weights is

$$\overline{Q(J_0)} = \int Du \frac{\exp\left(-\frac{\overline{b}}{2} J_0^2 + J_0\left(\overline{a} + u\sigma_a\right)\right) \Theta(J_0)}{\int_0^{+\infty} dw \exp\left(-\frac{\overline{b}}{2} w^2 + w\left(\overline{a} + u\sigma_a\right)\right)}$$



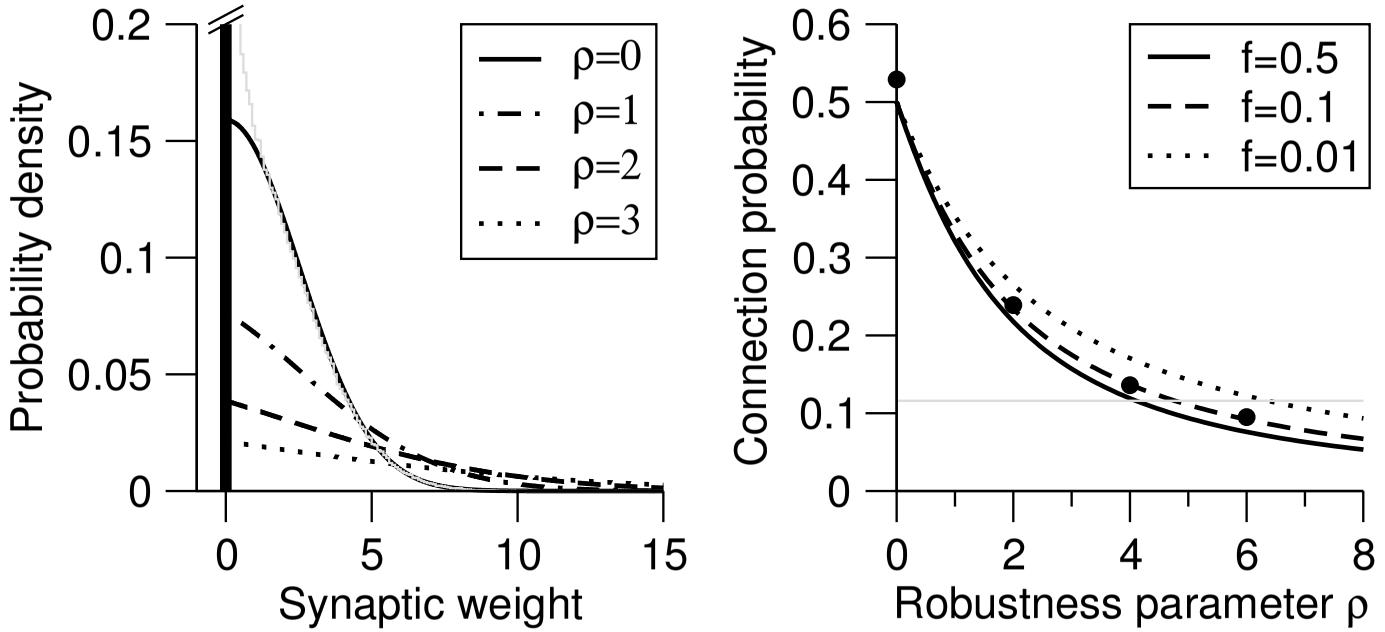
# Maximal capacity

- When $q \to Q$, both $a$ and $b$ diverge as $1/(Q-q)$;
- Again, truncated Gaussians converge to delta functions;



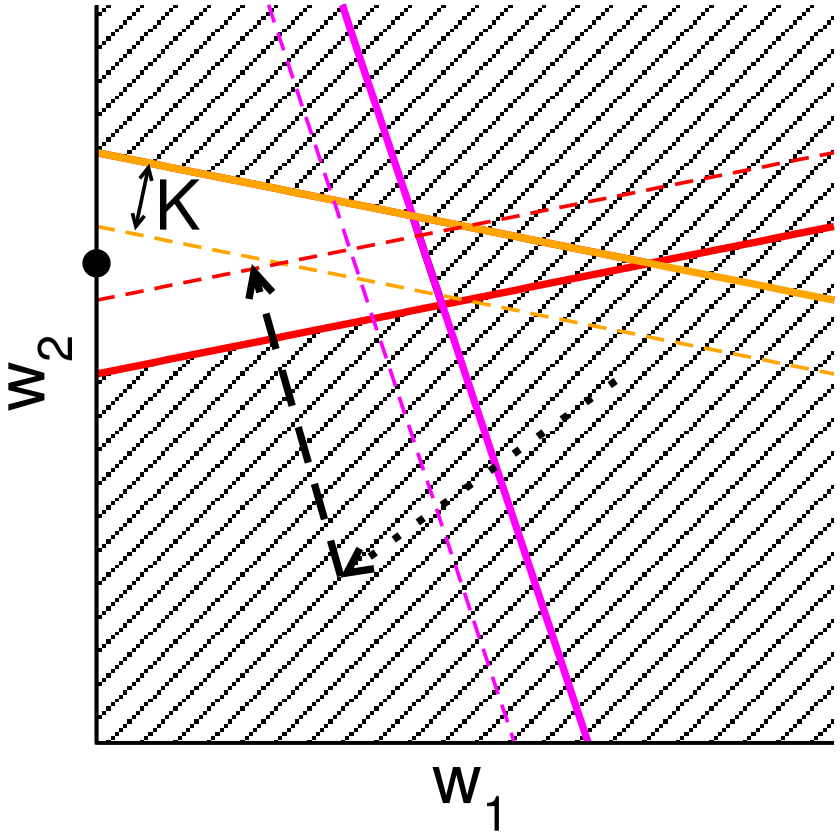
## $P(J_{ij})$ at maximal capacity

$$P(J_{ij}) = S\delta(J_{ij}) + \frac{1}{\sqrt{2\pi}\sigma_W} \exp\left[-\frac{1}{2}\left(\frac{J_{ij}}{\sigma_W} + J_0(S)\right)^2\right] \Theta(W)$$

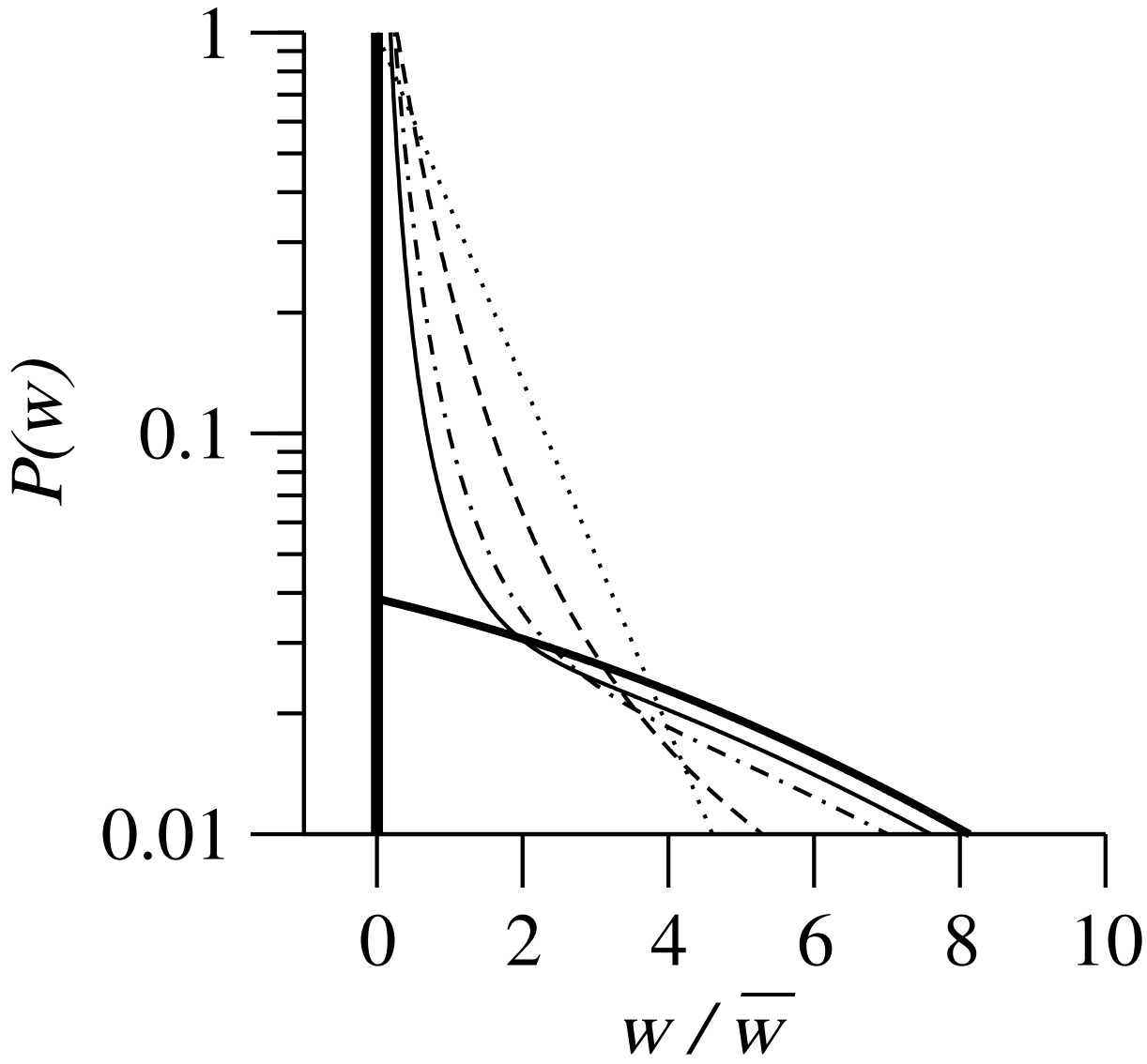


# Robust storage implies sparse connectivity

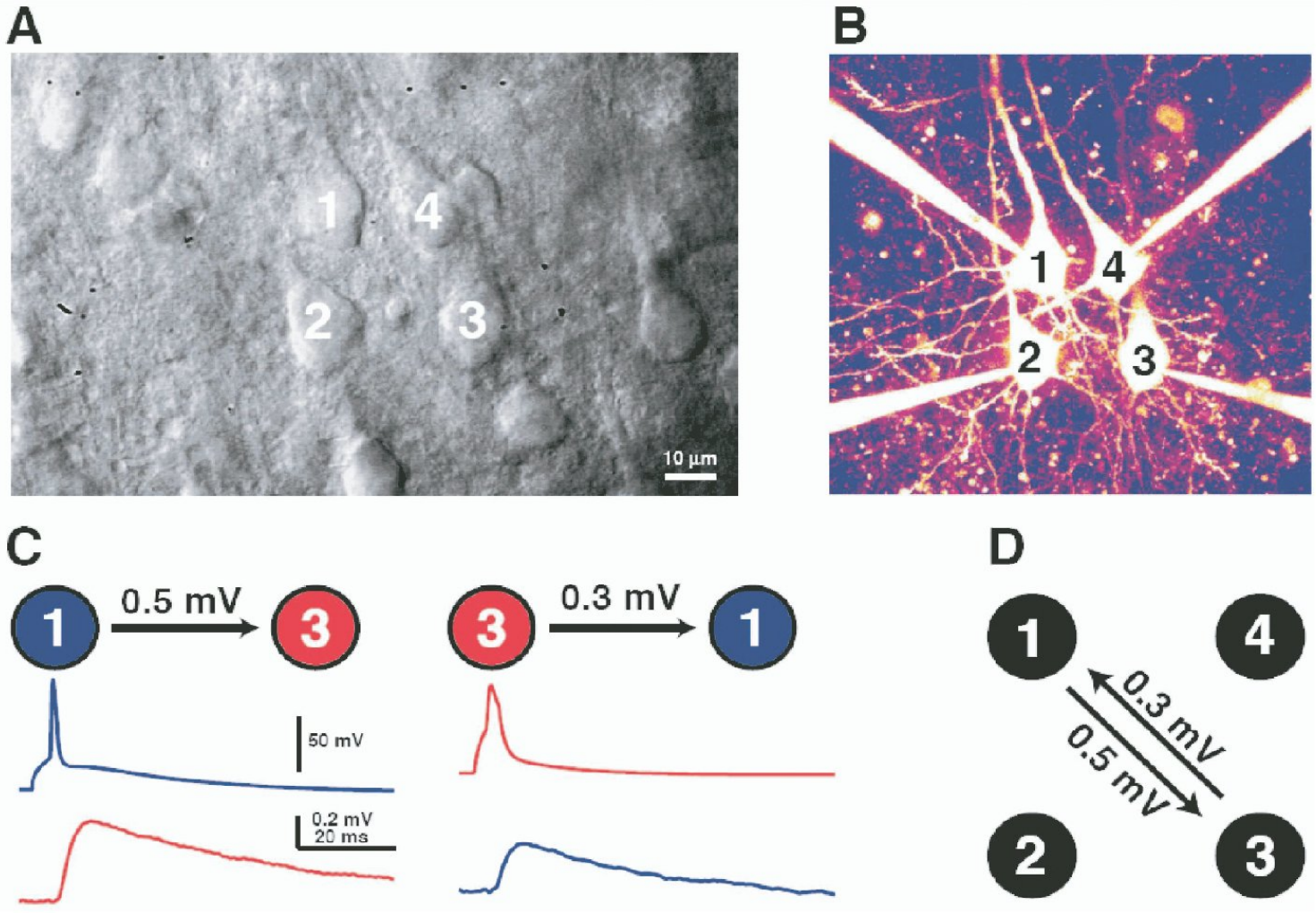
- Increasing robustness means increasing the distance to the pattern-associated hyperplanes;
- But not from the $J_{ij} = 0$ hyperplanes
- As a result, robust solutions are concentrated on a large fraction of the $J_i = 0$ hyperplanes



## Distribution of weights below capacity



# Data: intracellular recordings in cortical slices

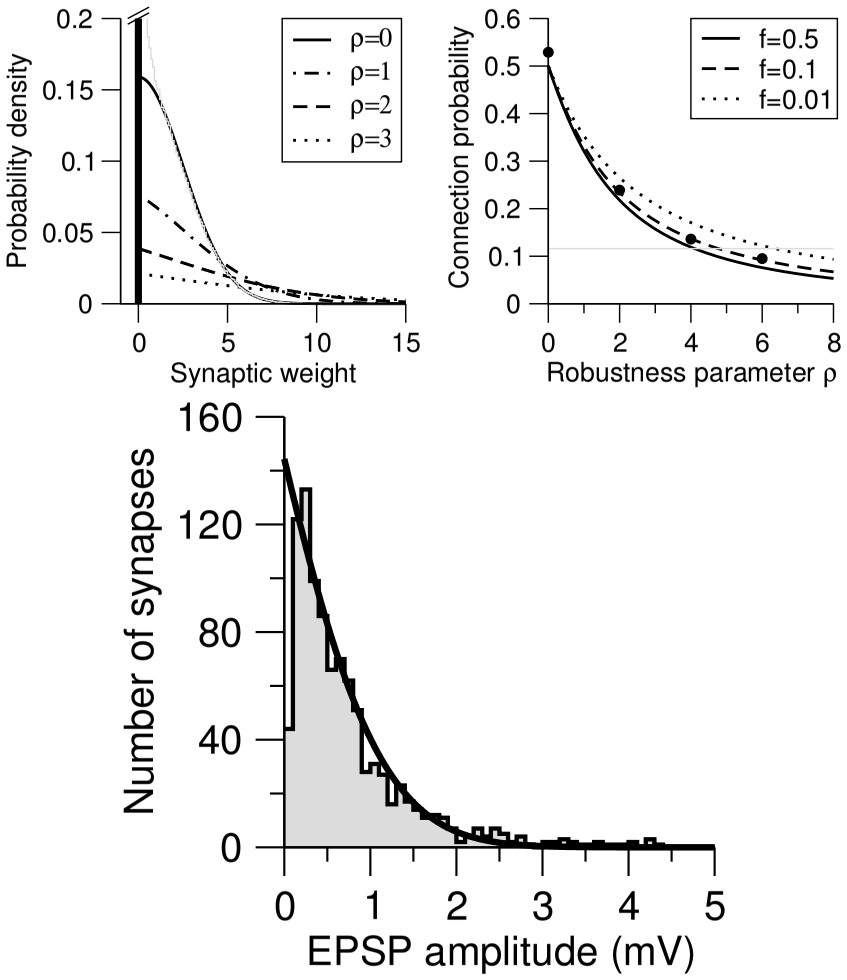



Recordings by Jesper Sjostrom (Sjostrom et al 2001, Song et al 2005)

# Distribution of synaptic weights: experiment vs theory

- Large fraction of zero weight synapses is consistent with data:
  - **Anatomy:** nearby pyramidal cells are potentially fully connected (Kalisman et al 2005)
  - **Electrophysiology:** nearby pyramidal cells have connection probability of $\sim 10\%$ (Mason et al 1991, Markram et al 1997, Sjostrom et al 2001, Holmgren et al 2003)

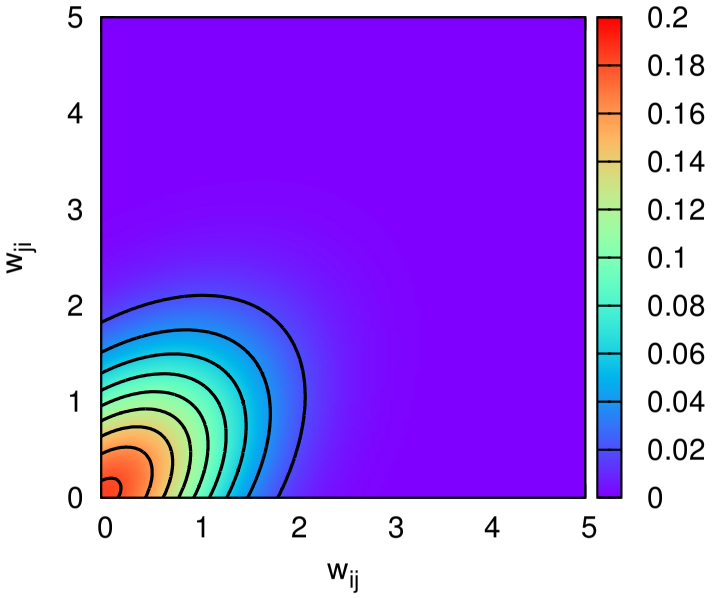
  $\Rightarrow$ **Large fraction of zero-weight ('potential' or 'silent') synapses.**

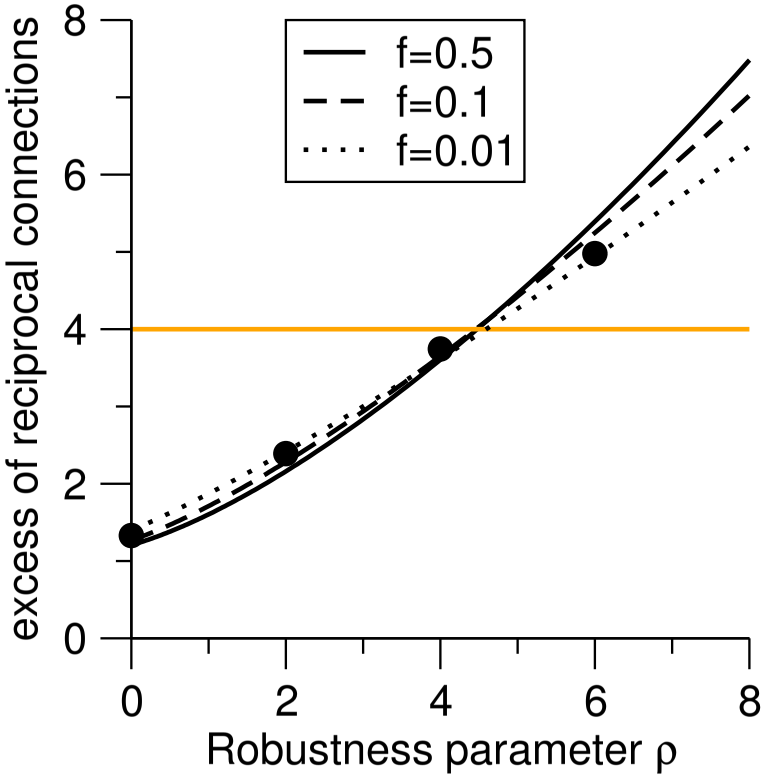



Sjostrom et al 2001; Song et al 2005

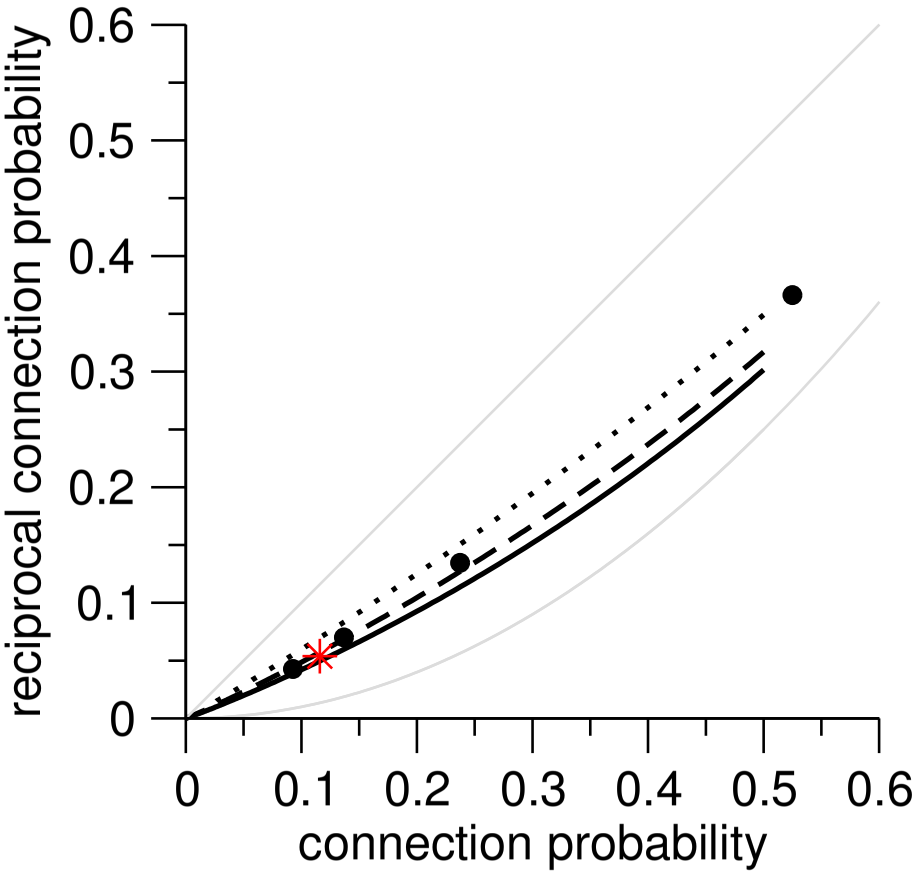
# Two-neuron connectivity

- Calculation of the joint distribution $P(J_{ij}, J_{ji})$ using cavity method;
- Leads to truncated correlated 2-D Gaussian, with delta functions on $J_{ij} = 0$ and $J_{ji} = 0$ axis (and product of two delta functions at $J_{ij} = J_{ji} = 0$).
- Over-representation of bidirectionnally connected pairs of neurons, compared to random uncorrelated network with same connection probability



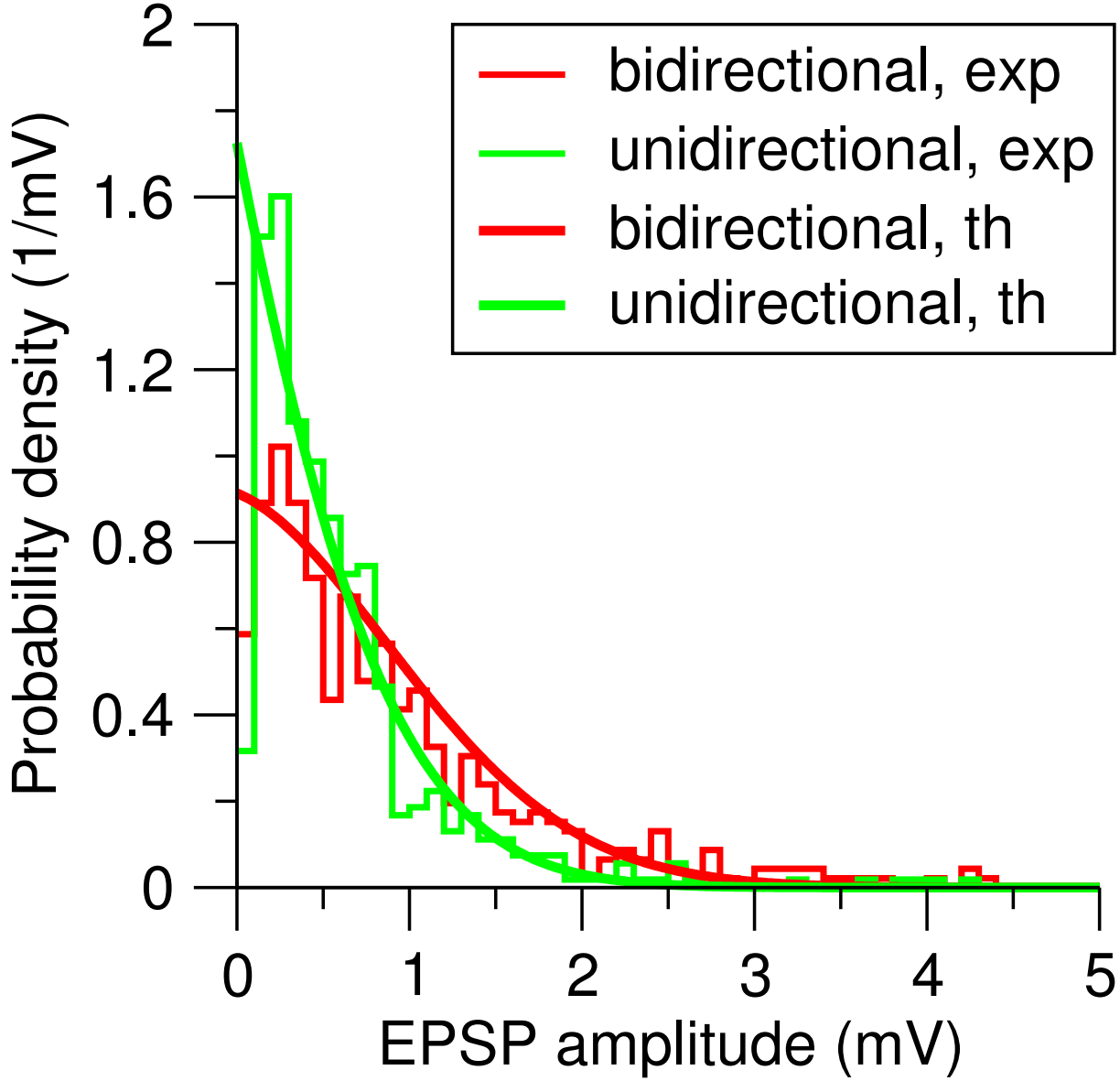



# Two-neuron connectivity: experiment vs theory




- Markram et al (1997) rat L5 somatosensory cortex: 3 x random;
- Song et al (2005) rat L5 visual cortex: 4 x random
- Wang et al (2006)
  - rat PFC: 4 x random;
  - rat visual cortex: 2 x random;
- Lefort et al (2009) mouse barrel cortex: $\sim$ random

**Distributions of weights: bidirectional vs unidirectional**



# Conclusions

- Maximizing information storage drives a large fraction ($> 0.5$) of synapses to zero
- Increasing robustness increases sparseness
- Networks storing fixed point attractors have an intermediate degree of symmetry
- Networks storing sequences are asymmetric (no overrepresentation of bidirectional connections)
- Theoretical distributions of synaptic weights match data recorded in cortex
- Consistent with optimality of information storage in cortex

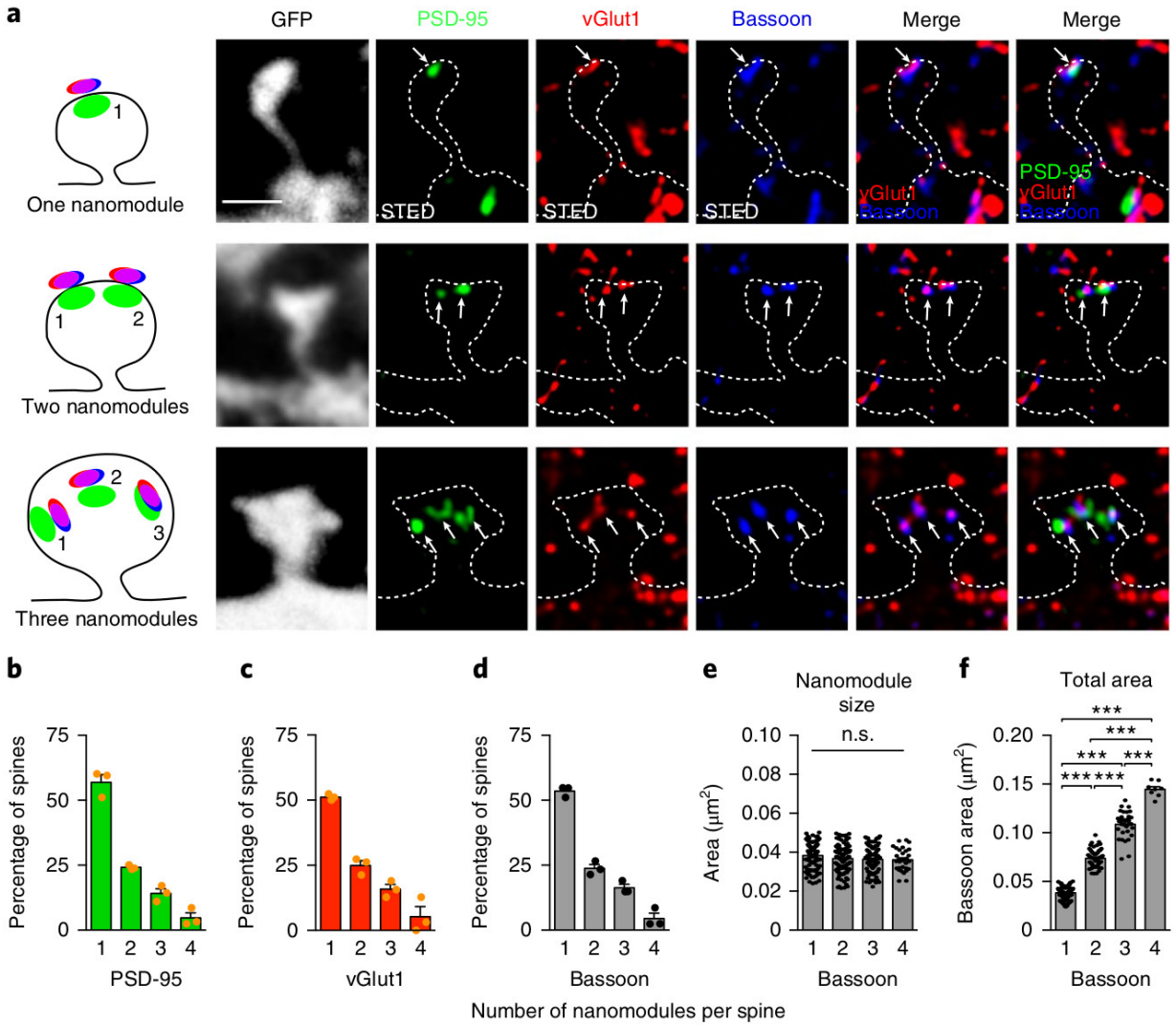
# Outline

1. Introduction: The Hebbian/Attractor Neural Network scenario
2. A brief overview of the relevant neurobiology
3. The Hopfield approach
4. The Gardner approach
5. **Open questions**

## Open questions

- Apply Gardner approach to more realistic networks
- Generalize Gardner approach: attractor states can be at a finite distance from patterns (not identical to them)
- How close to Gardner bound can we get with purely unsupervised learning?
- Information storage at larger scales?
- Storage of time-varying patterns?
- Mechanisms for storing multiple items in working memory?

- Networks with discrete synapses and on-line learning: Which learning algorithms maximize capacity?



- Suppose we can measure the full matrix $J_{ij}$ ('connectome'). Can we infer stored memories from the connectome?