UNINFORMATIVE MEMORIES WILL PREVAIL

THE STORAGE OF CORRELATED REPRESENTATIONS AND ITS CONSEQUENCES

Emilio Kropff

SISSA, Trieste

Semantic memory

- Tulving, 72: "the global network codifies for a general conceptual knowledge abstracted from a large number of individual episodes or experiences".

- Nowadays the dichotomy between Episodic vs Semantic memory is under revision. Some people think that they might be different stages of the same process.

- Embeds different kinds of information: perceptual <has 4 legs>, functional <is used for hunting>, associative <likes to chase cats> and encyclopedic <may be one of many breeds>. (DOG)

Category specific deficits

- Patients were found with a significant impairment in their knowledge about living things (animals + foodstuffs) as opposed to manmade artifacts (Warrington & Shallice, 1984).

- Heterogeneous etiology: herpes encephalitis, brain abcess, anoxia, stroke, head injury and dementia of Alzheimer type (DAT). Lesions typically include inferior parts of the temporal lobe.

- Impairment for nonliving has also been reported -> double dissociation. Current ratio: 23% vs 77% (Capitani, 03)

Theoretical accounts

• sensory/functional theory (Warrington & Shallice, 84) – Representation domains depend on the type of semantic information of concepts (animals – sensory information / tools – functional properties)

• domain-specific hypothesis (Caramazza & Shelton, 98) – Evolution has created a semantic system that is specific for animals while tools have no evolutional weight and are processed by a generic separated system.

Theoretical accounts

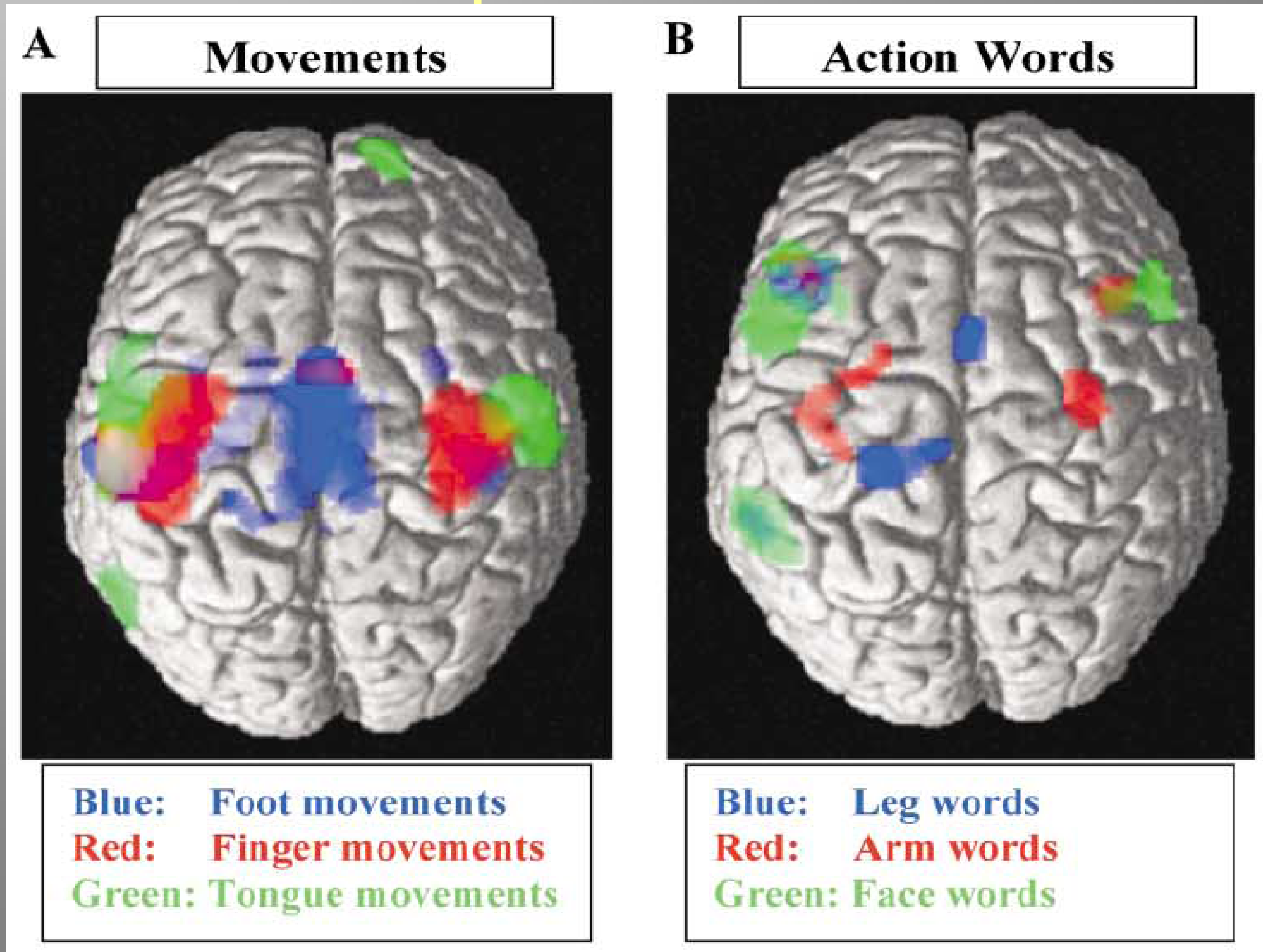
Theories concerning different measures of correlation between concepts:

• feature representation (McRee et al, 97) – concepts are represented by their features in an autoassociative memory. Problems with the storage capacity.

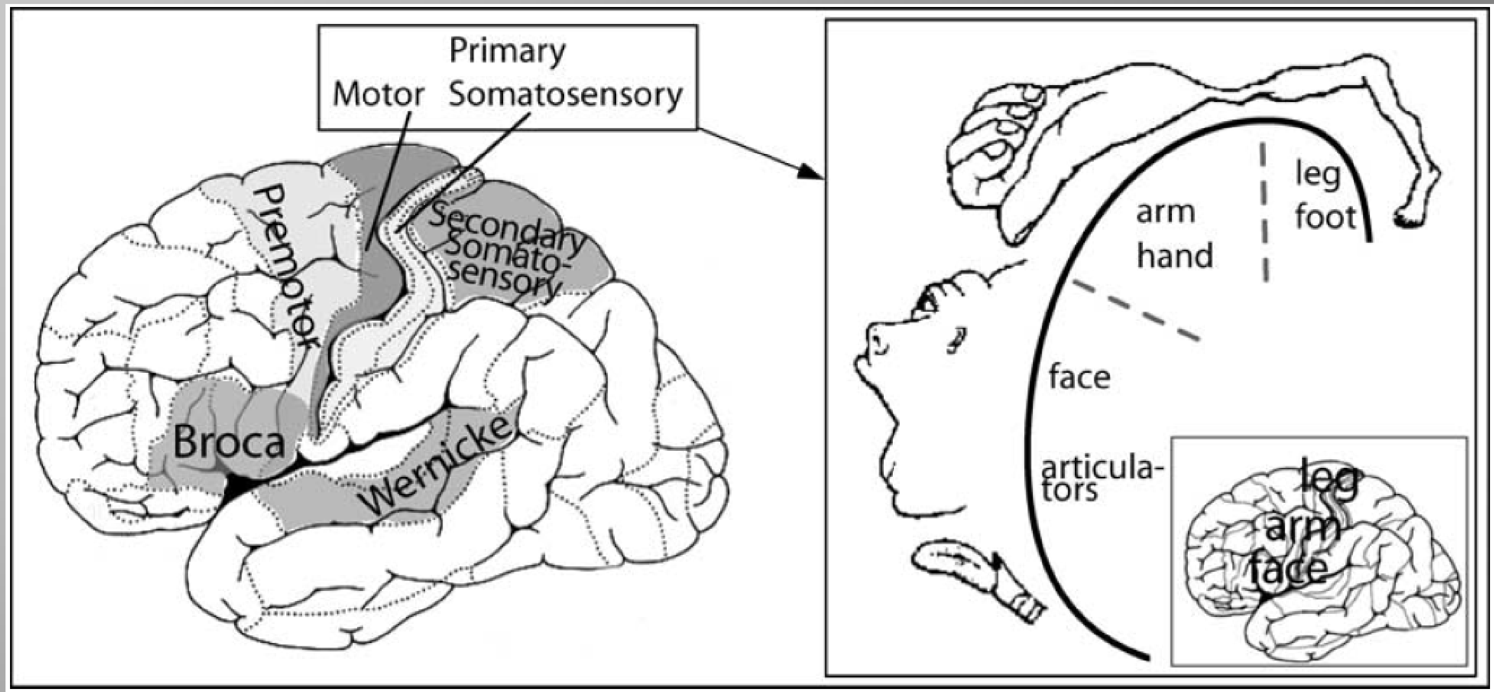
• conceptual structure account (Tyler & Moss, 01) – the structure of categories arises from: feature correlation, distinctive features and interactions between both.

• semantic relevance (Sartori & Lombardi, 04) – features have a relevance that is additive and depends on the whole structure of concepts. If a cue has a total relevance > threshold -> retrieval.

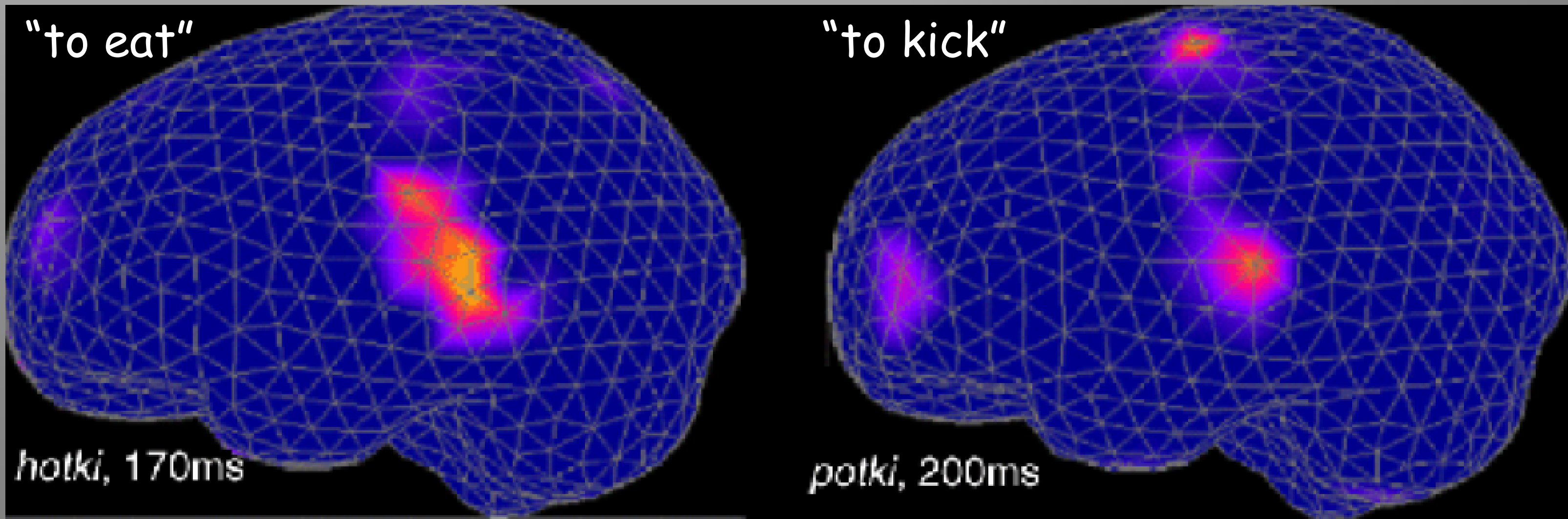
Embodied theories & Feature representation



(Pulvermuller, 04)



Early quasi-automatic word-evoked cortical activity.



Pulvermuller, 2003 – MCE on MEG recordings

The cerebral cortex

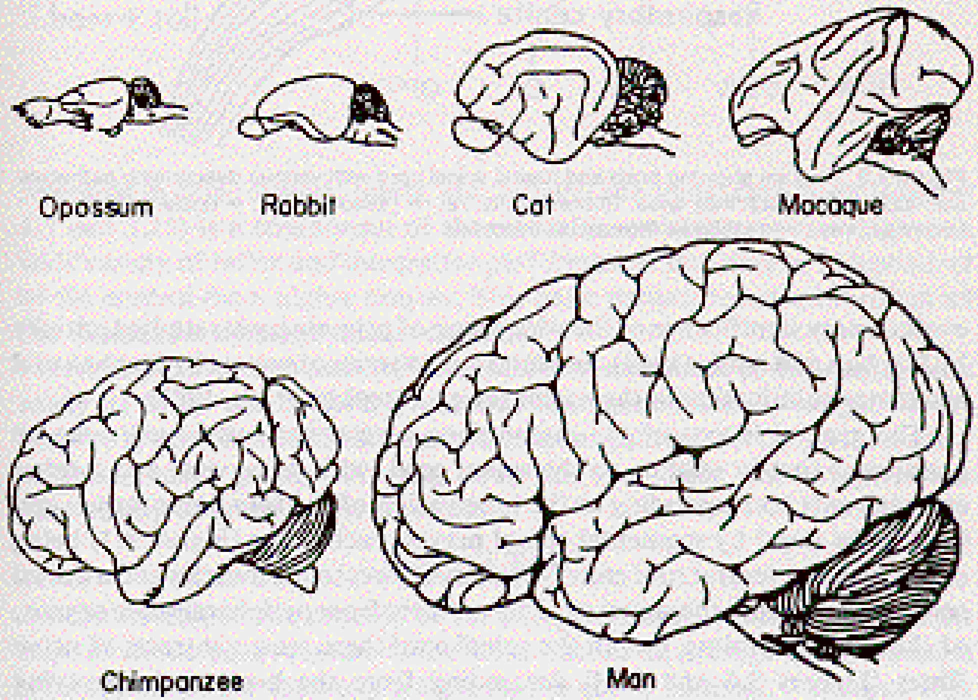


Figure 3.1 Brains of mammals drawn on the same scale (courtesy of Professor J. Jansen).

- 85% of human brain
- Processing of sensory information
- voluntary movement
- problem solving
- language

Cerebral cortex – Braitenberg & Schüz, 1991

of neurons >> # of input fibers

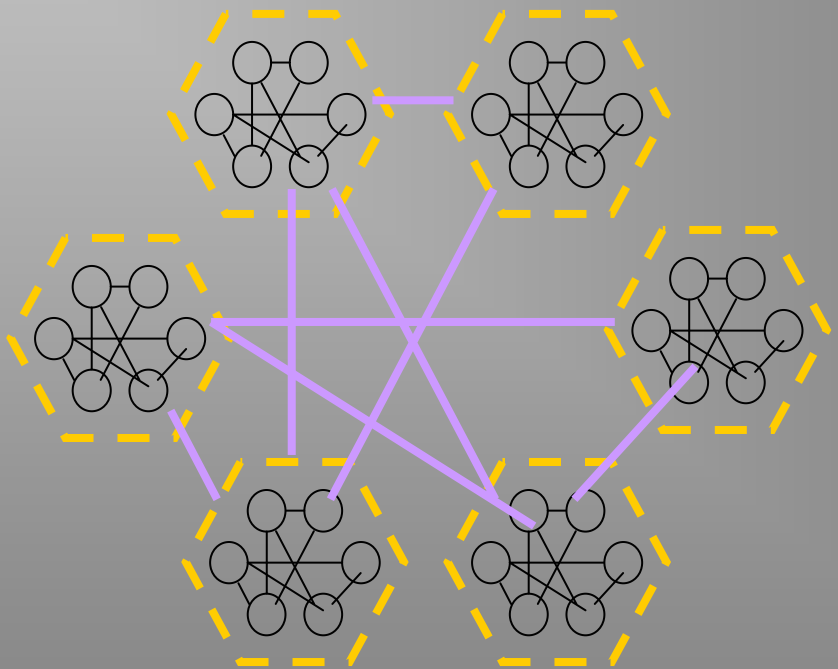
Modifiable synapses

No prefered direction in the connections

Mostly excitatory synapses

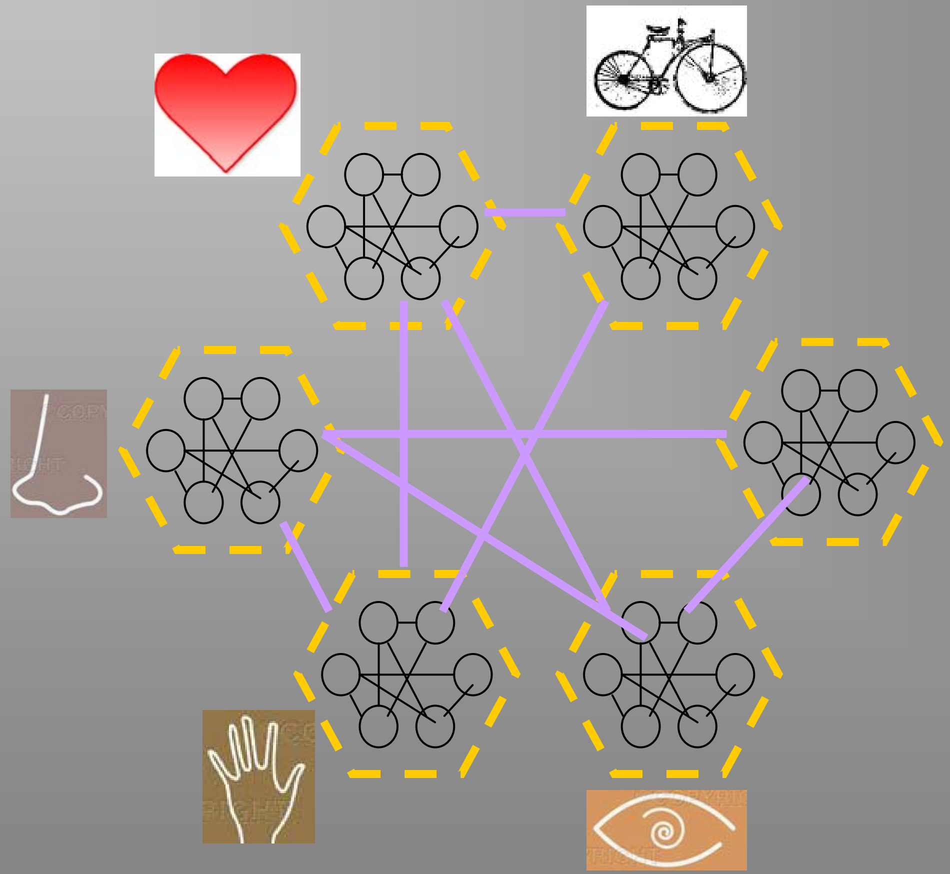
Great convergence & divergence

Connections are very weak

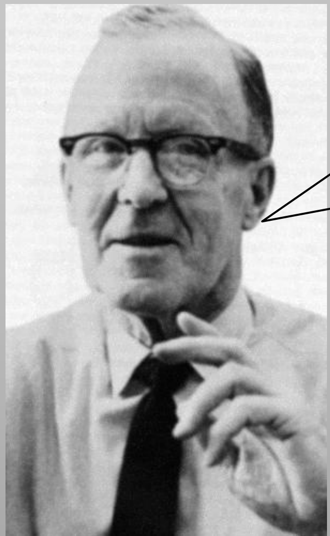


Two-level associative memory with formation of cell assemblies

Cerebral cortex – Braitenberg & Schüz, 1991

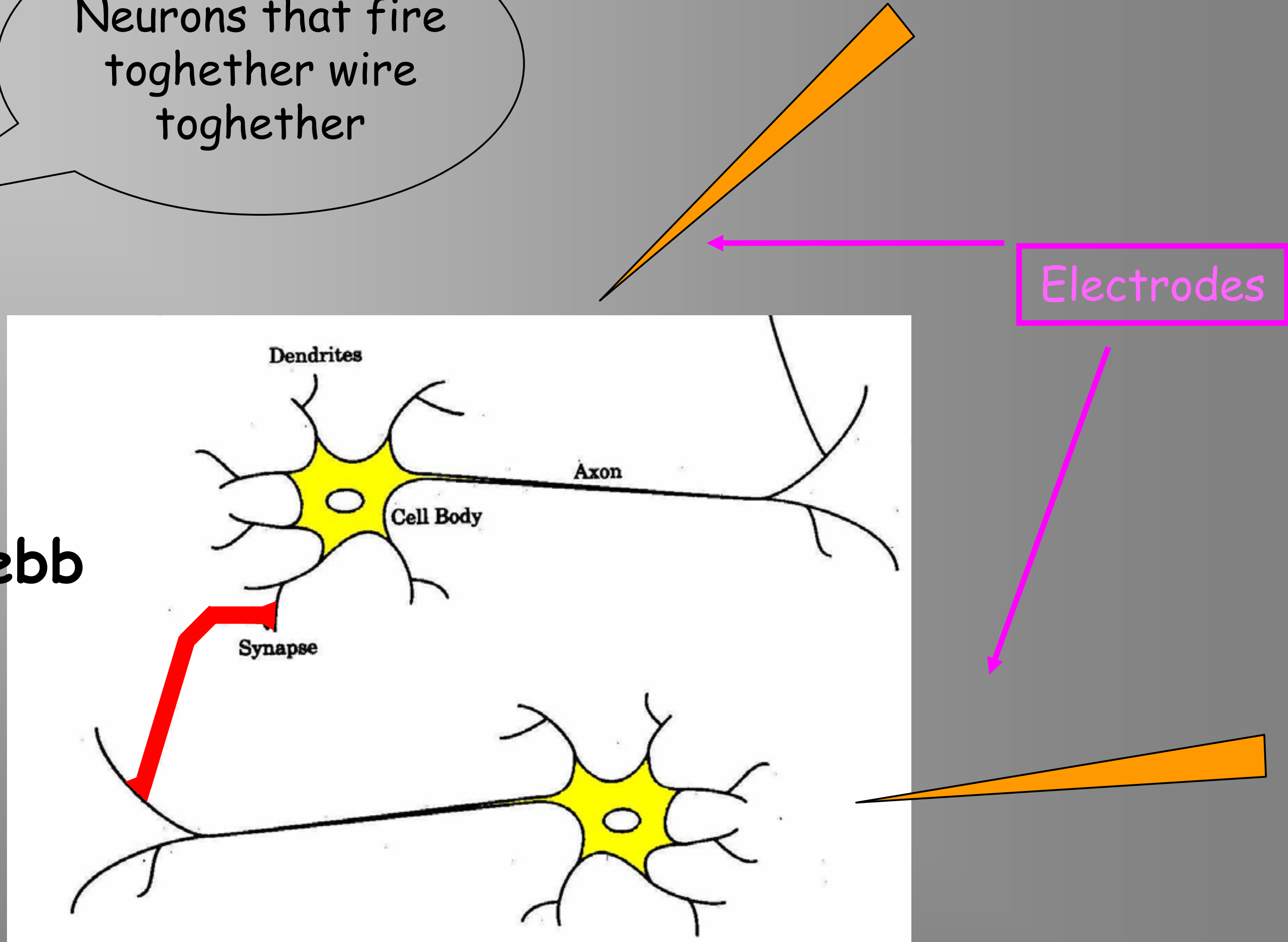


Two-level associative memory with formation of cell assemblies

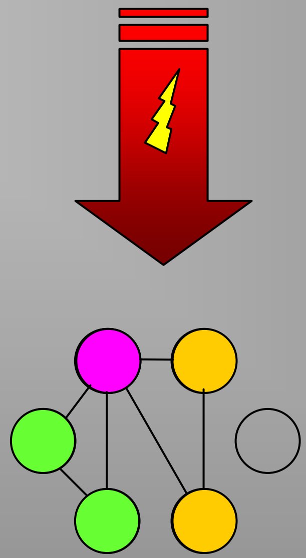


Neurons that fire toghether wire toghether

Donald O. Hebb
1904 - 1985



Auto-associative memories

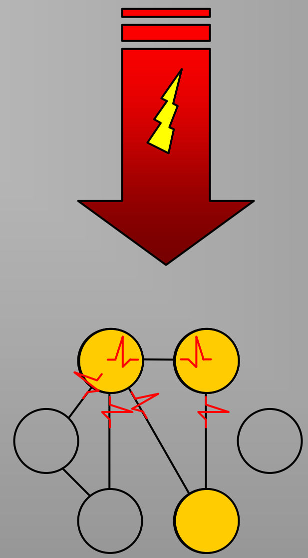


- No activity
- Pattern #1 active
- Pattern #2 active
- Pattern #3 active

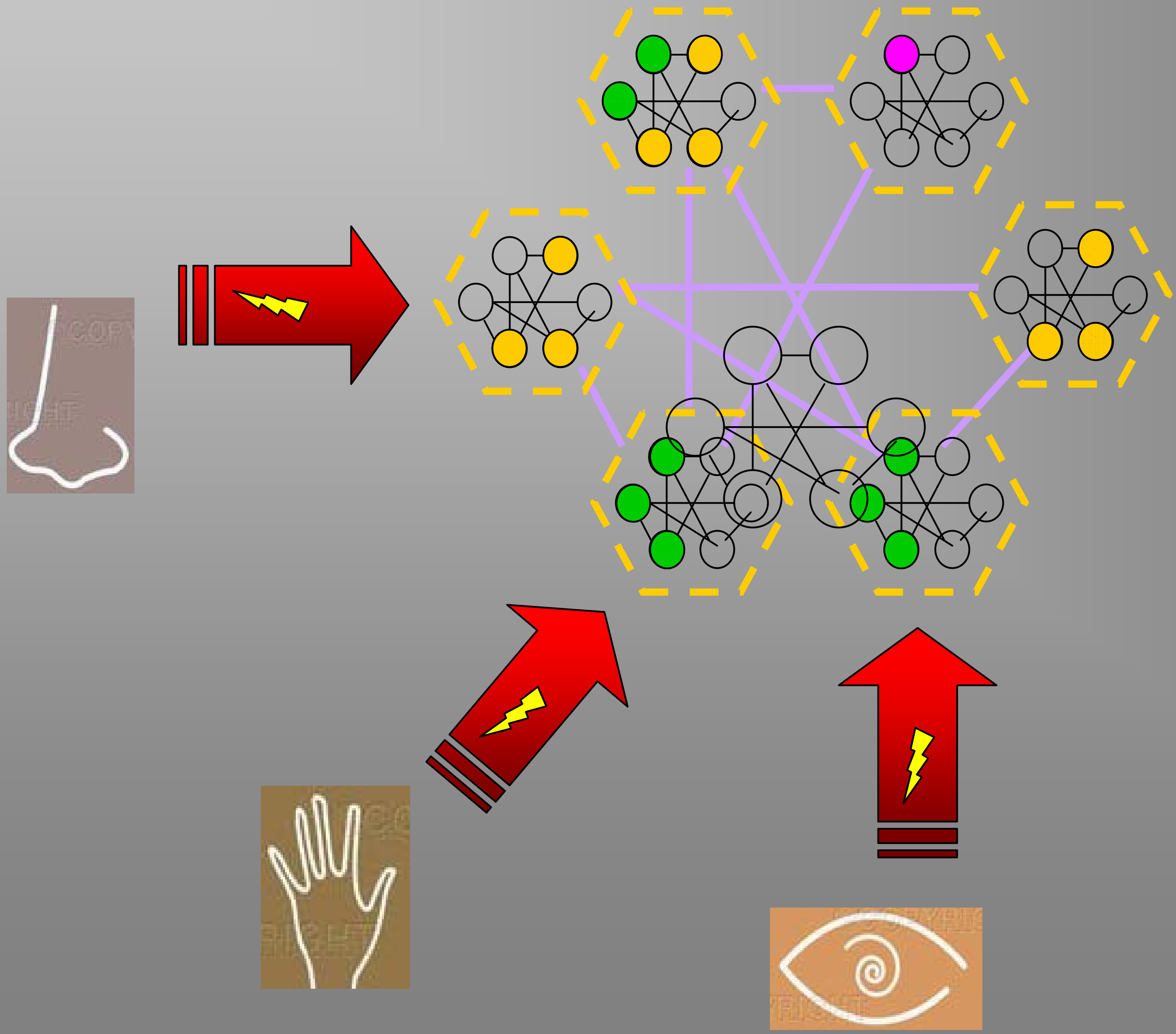
Learning !!

Testing the memory

- Pattern #2 active



Cerebral cortex – Braitenberg & Schüz, 1991



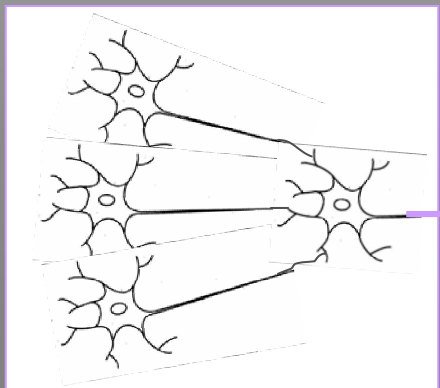
Two-level associative memory with formation of cell assemblies

Hopfield memories

- The network stores **p** patterns, each one characterized by a vector ξ in **N** dimensions, with components following:

$$P(\xi_i^{\mu}) = (1 - a)\,\delta(\xi_i^{\mu}) + a\,\delta(\xi_i^{\mu} - 1) \qquad [i = 1 \ldots N,\ \mu = 1 \ldots p]$$

were a *is the sparseness, the fraction of active neurons when the network is in an attractor state.*



$$h_i = \sum_{j=0}^{N} J_{ij}\,\sigma_j - U$$

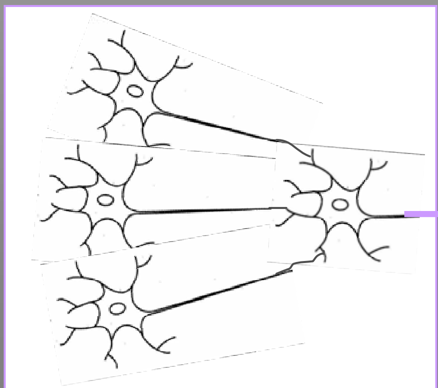
$$\sigma_i = \frac{1}{1 + e^{-\beta h_i}}$$

Hopfield memories

- U is a threshold of order 1, necessary to mantain the activity low, avoiding storage capacity colapse (Tsodyks, 89).
- β is an inverse temperature
- J_{ij} are the weights following the hebbian rule:

$$J_{ij} = \frac{C_{ij}}{C.a.(1-a)} \sum_{\mu=1}^{p} (\xi_i^{\mu} - a)(\xi_j^{\mu} - a)$$

$$\sum_{j=1}^{N} C_{ij} = C$$



$$h_i = \sum_{j=0}^{N} J_{ij}\, \sigma_j - U$$

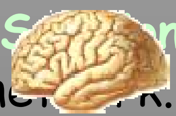
$$\sigma_i = \frac{1}{1 + e^{-\beta h_i}}$$

Hopfield memories

• If patterns are randomly correlated (Tsodyks,89),

$$p_{max} \sim \frac{C}{a \ln\left(\frac{1}{a}\right)}$$

• However, if patterns have a non-trivial structure of correlations, the storage capacity colapses.



•S[illegible]n #1: Orthogonalize the patterns before feeding the ne[illegible]k. (i.e: Dentate Gyrus in Hippocampus)

•In semantic memory correlation between stored patterns seems to play a major role.

Solution #2 ??

$$J_{ij} = \frac{1}{C\,a} \sum_{j=1}^{N} C_{ij} \left(\xi_i^{\mu} - A_i\right)\left(\xi_j^{\mu} - B_j\right)$$

$$h_i = \sum_{j=0}^{N} J_{ij}\, \sigma_j$$

$$m_i^{\mu} = \frac{1}{C\,a} \sum_{j=1}^{N} C_{ij} \left(\xi_j^{\mu} - B_j\right) \sigma_j$$

$$h_i = \sum_{\mu=1}^{p} \xi_i^{\mu}\, m_i^{\mu} = \xi_i^{1}\, m + \sum_{\mu \neq 1} \xi_i^{\mu}\, m_i^{\mu}$$

$$A_i = B_i = a_i = \frac{1}{p} \sum_{\mu=1}^{p} \xi_i^{\mu} = popularity$$

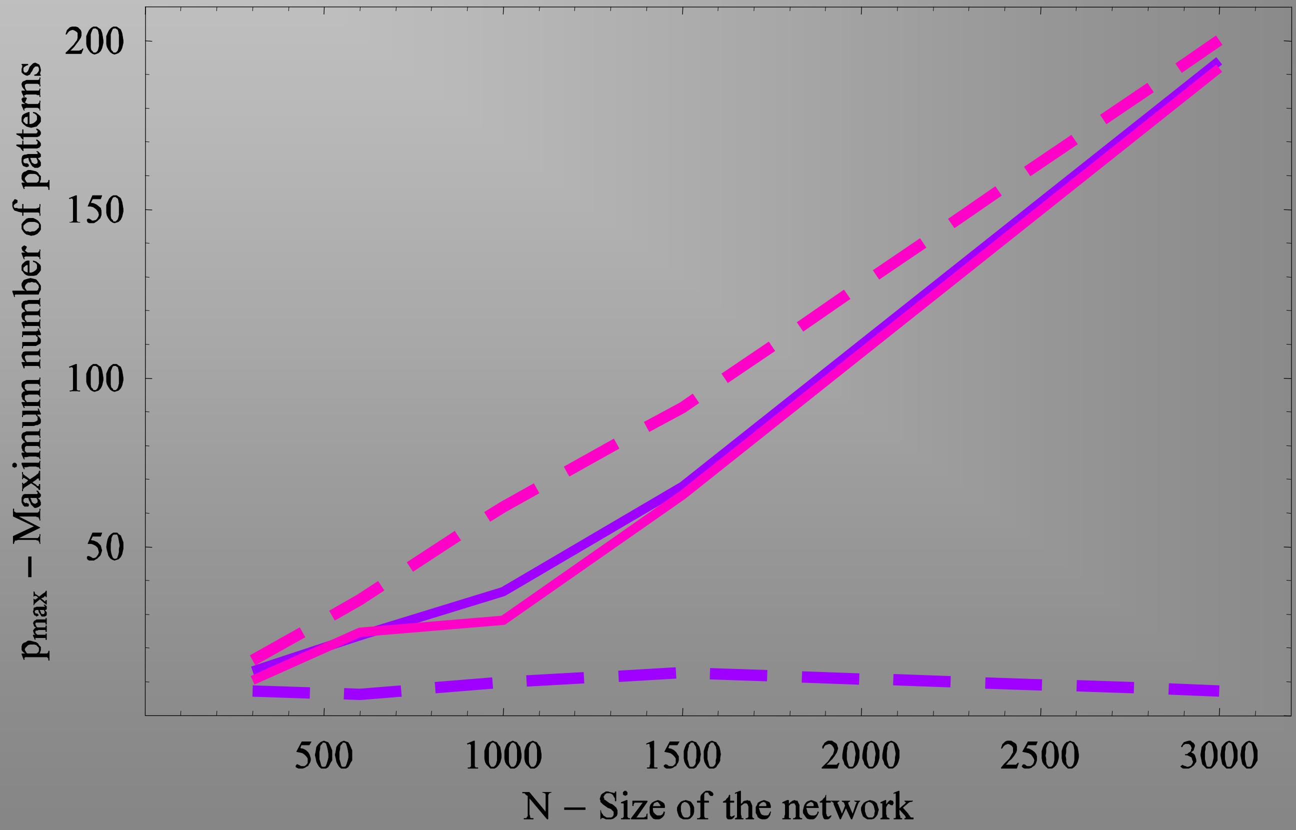
Classical result: hebbian learning supports uncorrelated memories

Classical result: catastrophe associated to correlated memories

$J_{ij} = \Sigma_\mu (\xi_i^\mu - a).(\xi_j^\mu - a)$

$J_{ij} = \Sigma_\mu (\xi_i^\mu - a_i).(\xi_j^\mu - a_j)$

popularity: $a_k = 1/p \, \Sigma_\mu \xi_k^\mu$



New result: a modification that supports correlated memories

New result: the performance is the same with uncorrelated memories

Propeties with $\alpha \approx 0$, $C \approx \ln(N)$

$$h_i = \sum_{\mu=1}^{p} \xi_i^{\mu} m_i^{\mu} = \xi_i^{1} m + \sum_{\mu \neq 1}$$

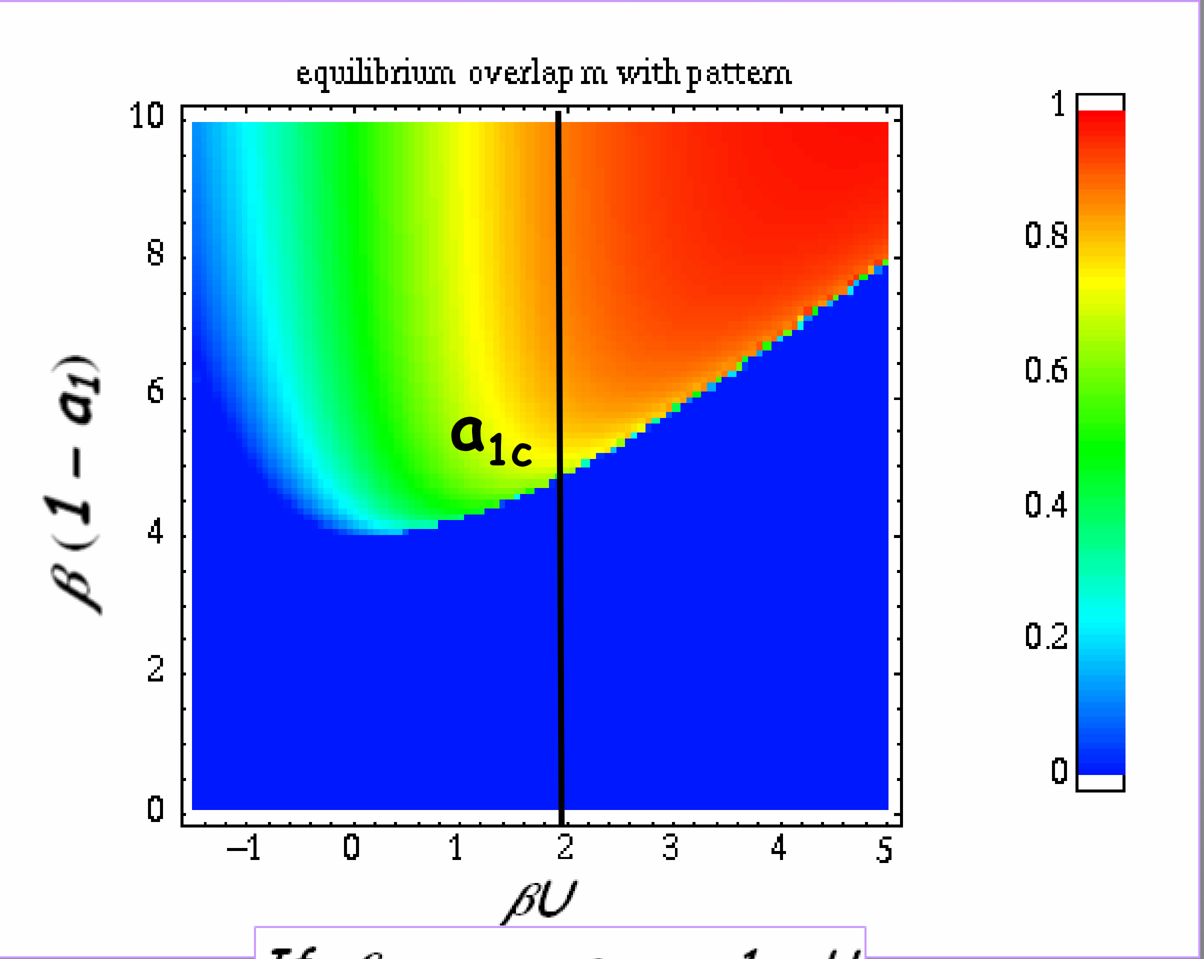
$$m = \frac{1}{N a} \sum_{j=1}^{N} \left(\xi_j^{1} - a_j\right) \sigma_j$$

$$\sigma_i = \frac{1}{1 + e^{-\beta h_i}}$$

$$m = (1 - a_1) \left\{ \frac{1}{1 + e^{\beta (U - m)}} - \frac{1}{1 + e^{\beta U}} \right\}$$

$$a_1 = \frac{1}{N . a} \sum_{\mu=1}^{p} \xi_j^{1} a_j = < a_{\xi^1} >= < \xi^{1} . \xi^{\mu} >_{\mu}$$

Propeties with $\alpha \approx 0$, $C \approx \ln(N)$



$$If\ \beta \to \infty, \quad a_{1c} = 1 - U$$

Propeties with $\alpha \approx 0$, $C \approx \ln(N)$

~ Conclusions ~

- If you want to be an attractor, you should pick at least some unpopular units.

- Lowering U can make any pattern retrievable -> ATTENTION

Propeties with finite α, C ≈ ln(N)

$$h_i = \sum_{\mu=1}^{p} \xi_i^{\mu} m_i^{\mu} = \xi_i^{1} m + \sum_{\mu \neq 1} \xi_i^{\mu} m_i^{\mu}$$

GAUSSIAN noise (If there is independence between neurons i and j).

$$h_i \sim \xi_i^{1} m + \sqrt{\alpha\, q\, a_i}\; z_i$$

$$q = \frac{1}{N a^2} \sum_{j=1}^{N} a_j (1 - a_j)\, \sigma_j^{2}$$

Propeties with finite α, C ≈ ln(N)

$$m = \frac{1}{N.a} \sum_{j=1}^{N} \left(\xi_j^{\,1} - a_j\right) \int_{-\infty}^{\infty} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}\, \theta\left(z + \frac{(m-U)}{\sqrt{\alpha . q . a_j}}\right) d z$$

$$q = \frac{1}{N.a^2} \sum_{j=1}^{N} a_j \left(1 - a_j\right) \int_{-\infty}^{\infty} \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}\, \theta\left(z + \frac{(m-U)}{\sqrt{\alpha . q . a_j}}\right) d z$$

Propeties with finite α, $C \approx \ln(N)$

$$m = \int_0^1 f(x)\{(1-x)\,\phi_1 + x\,\phi_0\}\, dx - \frac{1}{a}\int_0^1 F(x)\, x\, \phi_0\, dx$$

$$q = \frac{1}{a}\int_0^1 f(x)\, x\,(1-x)\{\phi_1 - \phi_0\}\, dx + \frac{1}{a^2}\int_0^1 F(x)\, x\,(1-x)\,\phi_0\, dx$$

$$\phi_k = \frac{1}{2}\left\{1 + Erf\left(\frac{k\,m - U}{\sqrt{2\,\alpha\, q\, x}}\right)\right\}$$

a_j	*follow a distribution F(x)*
a_ξ	*follow a distribution f(x)*

Propeties with finite α, $C \approx \ln(N)$

$$I_f = \int_0^1 f(x).x.(1-x)\,dx$$

$$I_F = \int_0^1 F(x).x.(1-x)\,dx$$

- At zero order, $\alpha_c = (p/C)_c \sim 1/I_f$

- At first order, the correction depends on I_F. The faster the distribution falls, the better the storage capacity.

Propeties with finite α, $C \approx \ln(N)$

$$I_f = \int_0^1 f(x).x.(1-x)\,dx$$

$$I_F = \int_0^1 F(x).x.(1-x)\,dx$$

If F(x) decays fast enough

$$\alpha_c \propto \frac{1}{I_f \ln\left(\frac{I_F}{a I_f}\right)}$$

... $\left(\text{If } F(x) = \delta(x-a) \rightarrow I_f = I_F = a \rightarrow \alpha_c \propto \frac{1}{a \ln\left(\frac{1}{a}\right)} \right)$

Propeties with finite α, $C \approx \ln(N)$

$$I_f = \int_0^1 f(x).x.(1-x)\,dx$$

$$I_F = \int_0^1 F(x).x.(1-x)\,dx$$

If F(x) decays fast enough

If F(x) decays exponentially

If F(x) decays as a power law

$$\alpha_c \propto \frac{1}{I_f \ln\left(\frac{I_F}{a I_f}\right)}$$

$$\alpha_c \propto \frac{1}{I_f\left[\ln\left(\frac{I_F}{a I_f}\right)\right]^2}$$

$$\alpha_c \propto \frac{a}{I_f \ln\left(\frac{a^{\gamma-2}}{I_f}\right)}$$

Propeties with finite α, $C \approx \ln(N)$

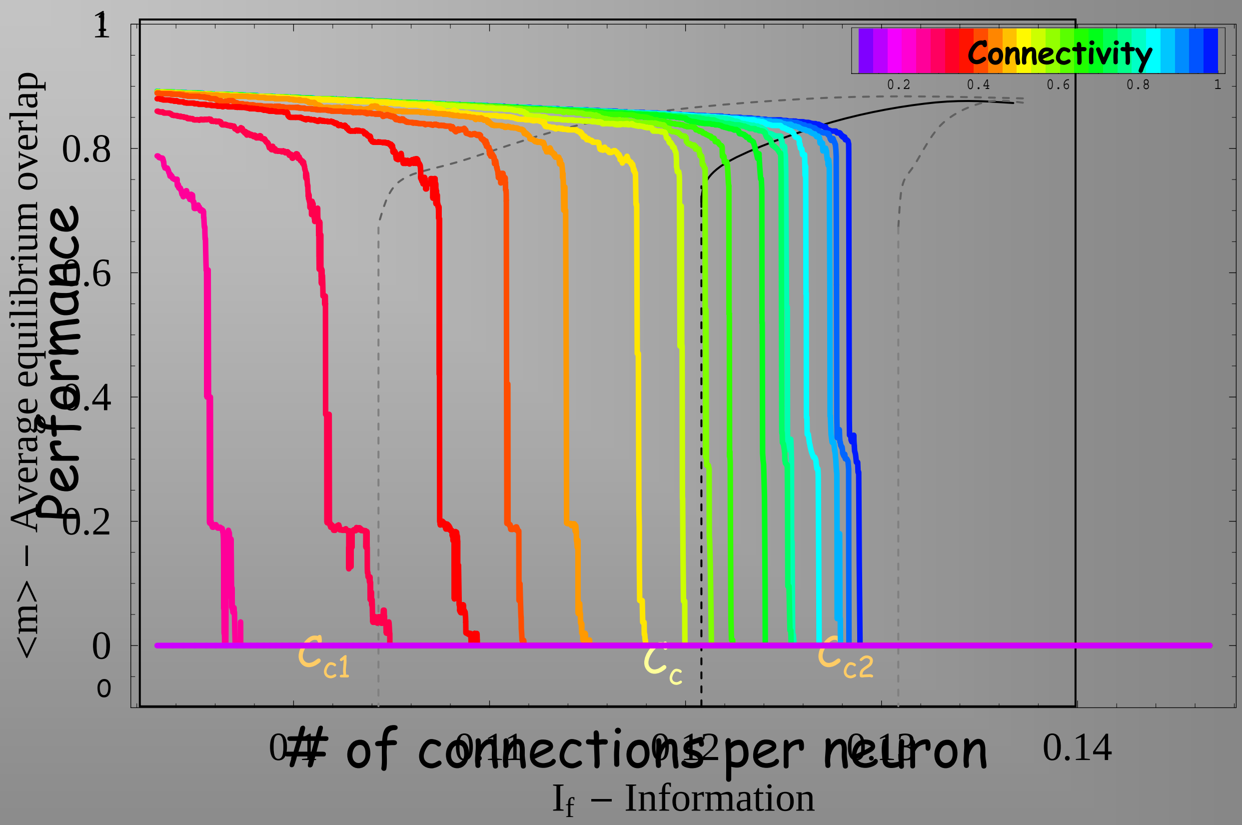
~ Conclusions ~

- α_c depends on the retrieved pattern (selective impairment).

- A pattern is more resistant to lesioning or to forgetting if it has a smaller value of:

$$I_f = \int_0^1 f(x).x.(1-x)\,dx$$

Storage capacity ~ fixed p



Information = $\Sigma_i \, a_i \, (1-a_i)$

summed over active neurons in the pattern

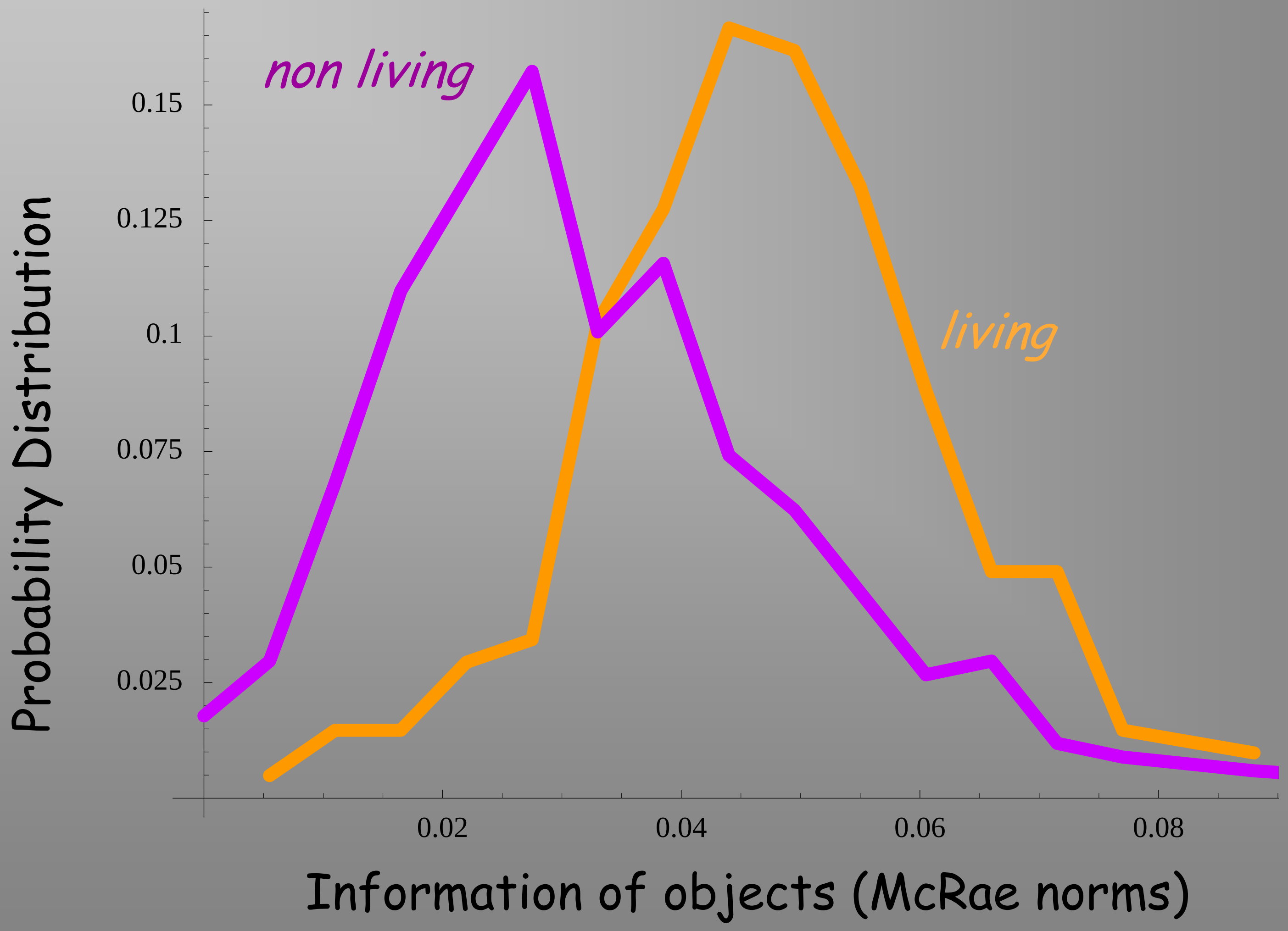
McRae's feature norms

- (McRae et al, 05) - www.psychonomic.org/archive

- 541 concepts covering a wide range of living and non-living examples used in previous studies. Participants were provided with 20 unrelated concepts and asked to list at most 10 features. Recording identifying sinonymous features, etc.

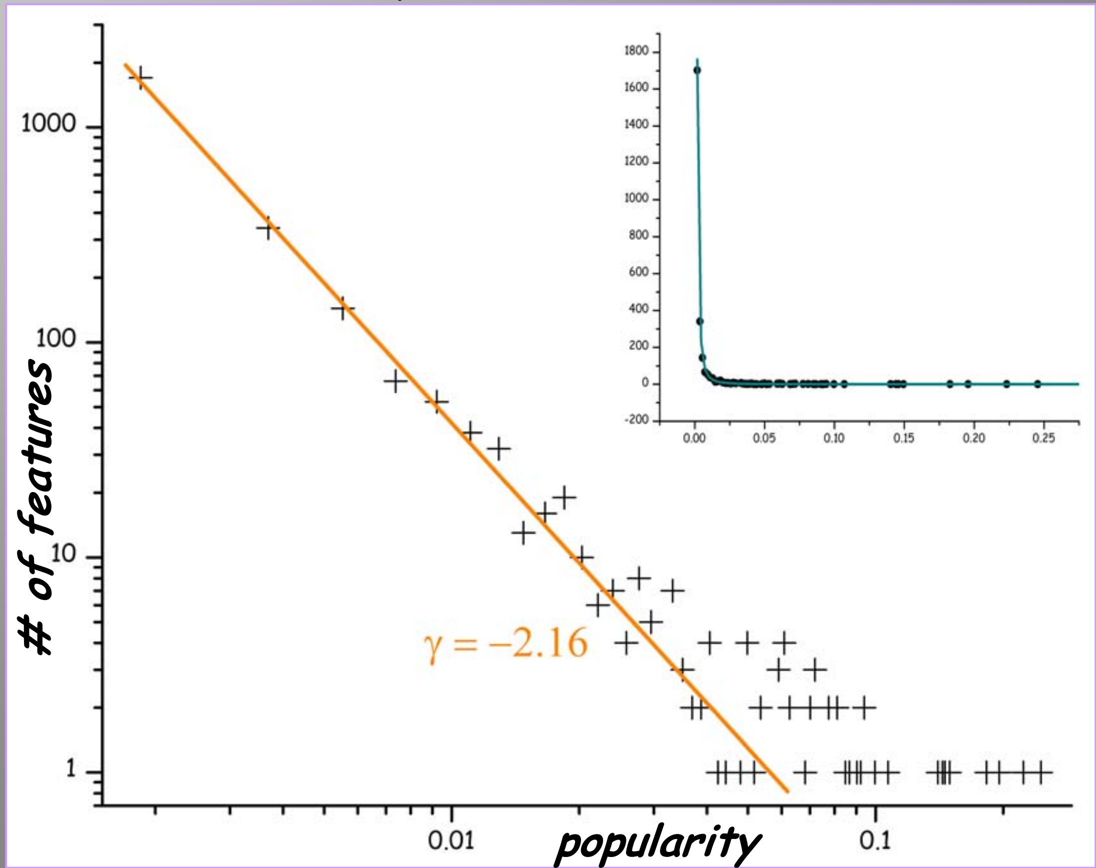
- "Feature norms are assumed to provide valid information not because they yield a literal record of semantic representations, but rather because such representations are used systematically by participants when generating features."

Category specific effects



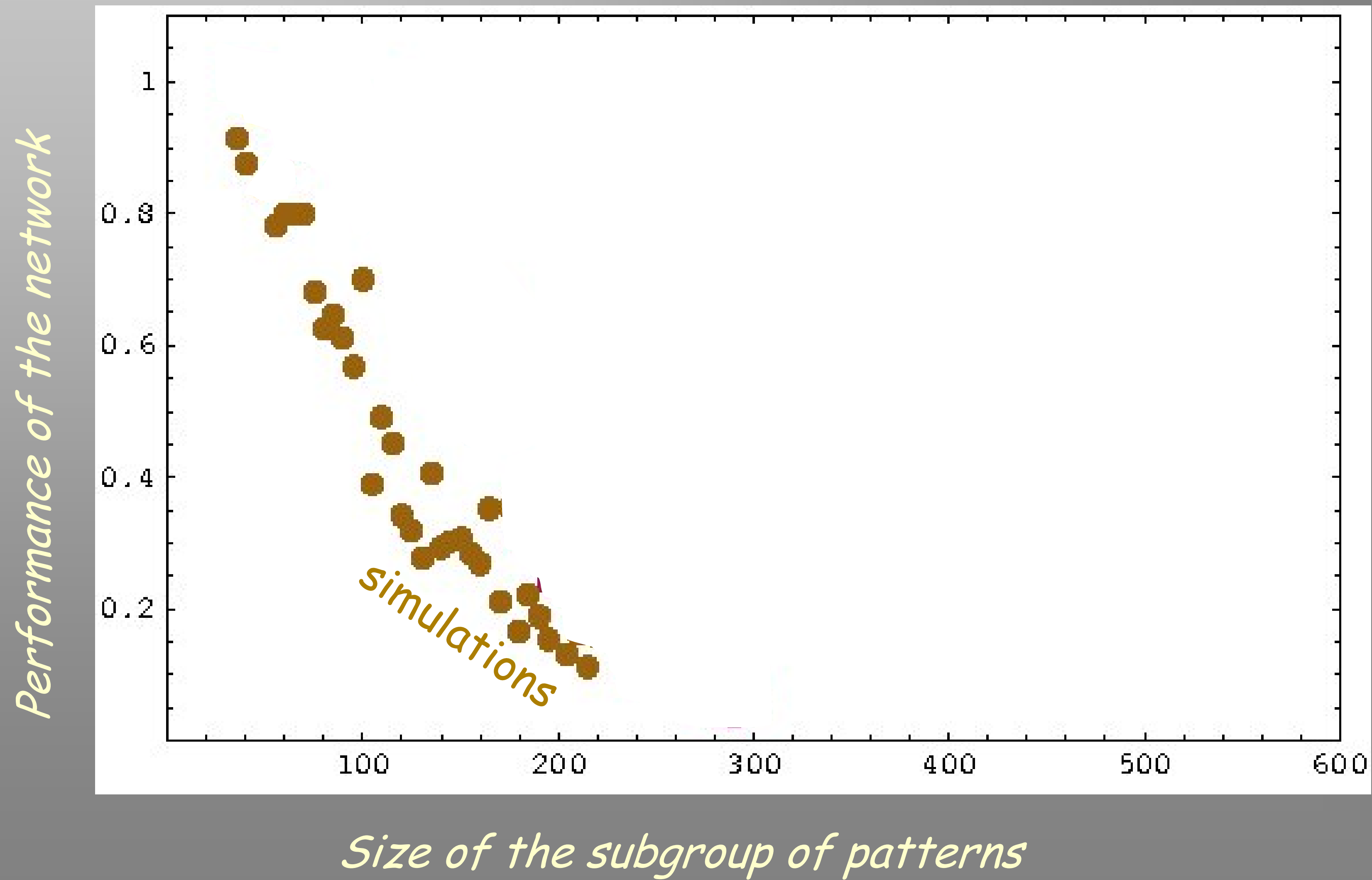
McRae's feature norms

• In the semantic memory literature, auto-associative networks are often presented as weak models. Why?



• To convince psychologists one must show an auto-associative memory that is able to store feature norms.

McRae's feature norms

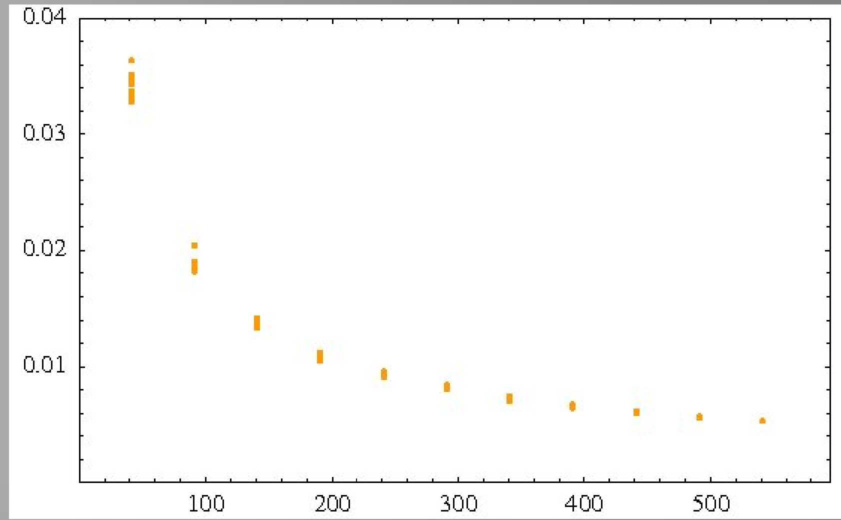


McRae's feature norms

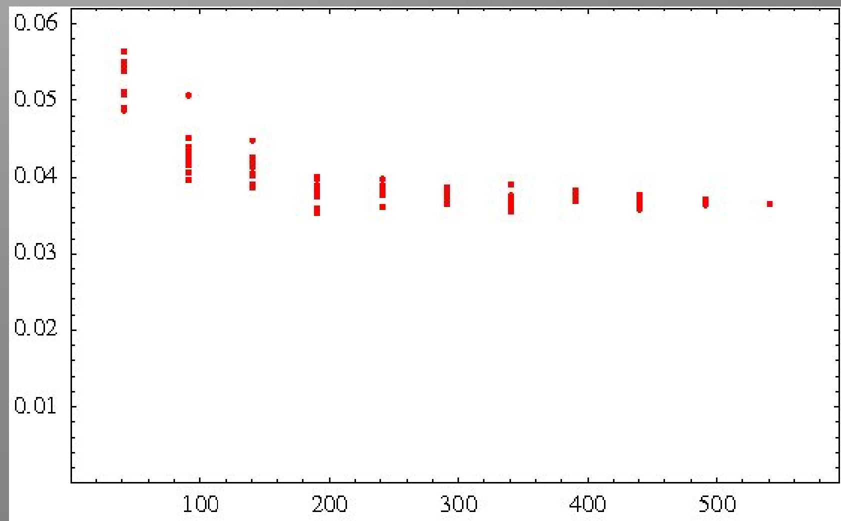
$$\alpha_c \propto \frac{a}{I_f \ln\left(\frac{a^{\gamma-2}}{I_f}\right)}$$

=

a - average sparseness

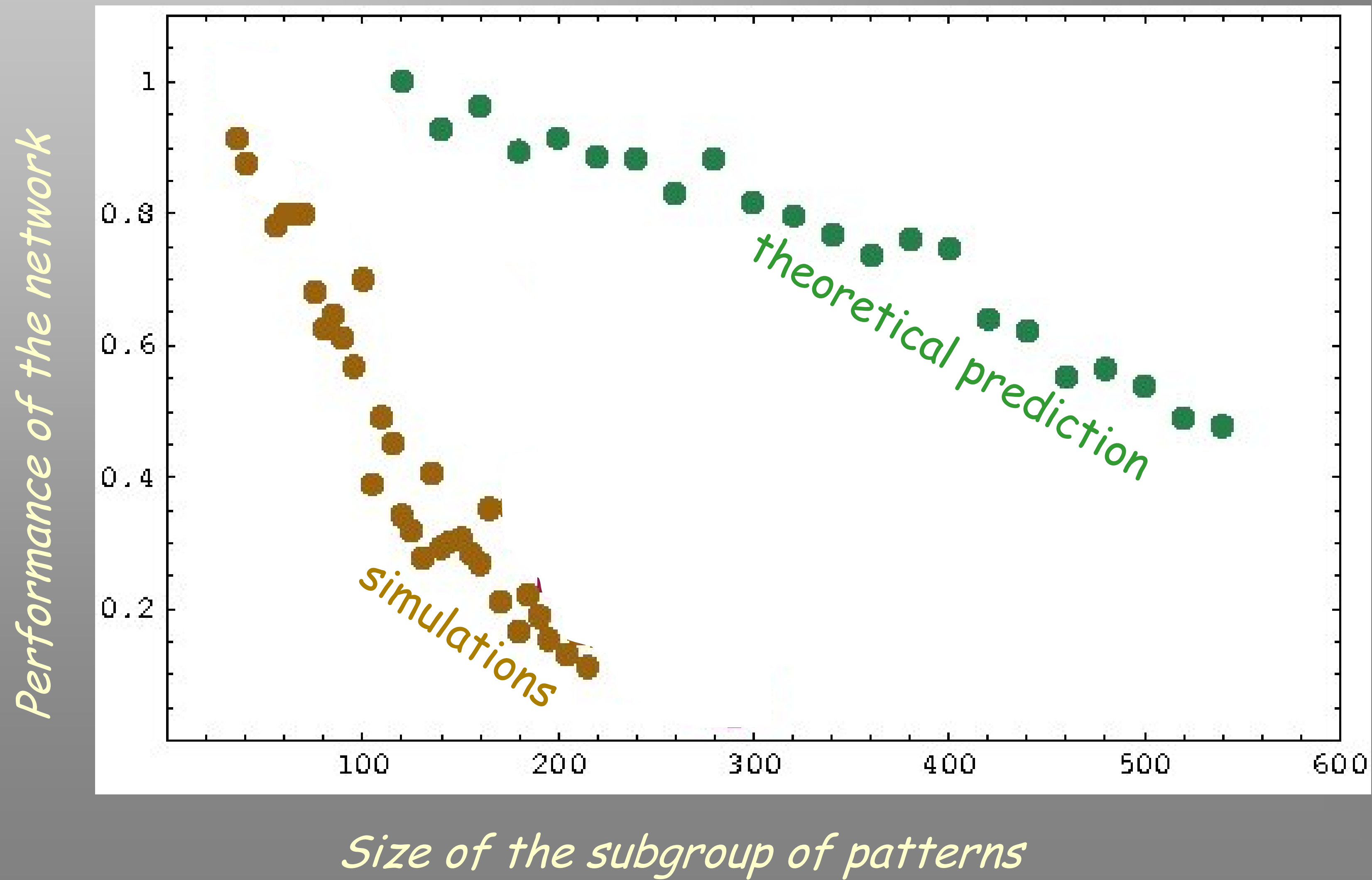


I_f - average information



Number of patterns p in the subgroup

McRae's feature norms



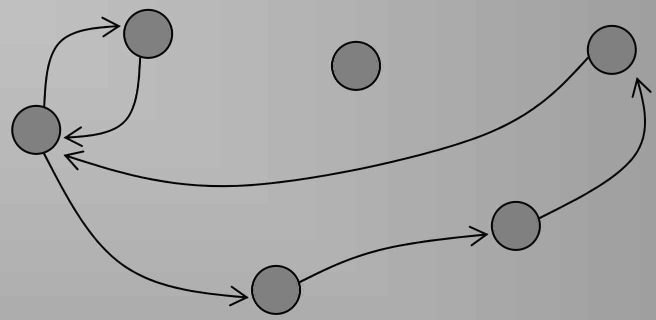
McRae's feature norms

Why the real network performs poorly?

• Independence between features is not valid (e.g: beak and wings). Is this effect strong enough? In case it is, there would be a storage capacity colapse.

• The system works but the approximation of diluted connectivity is not good.

McRae's feature norms: the full solution

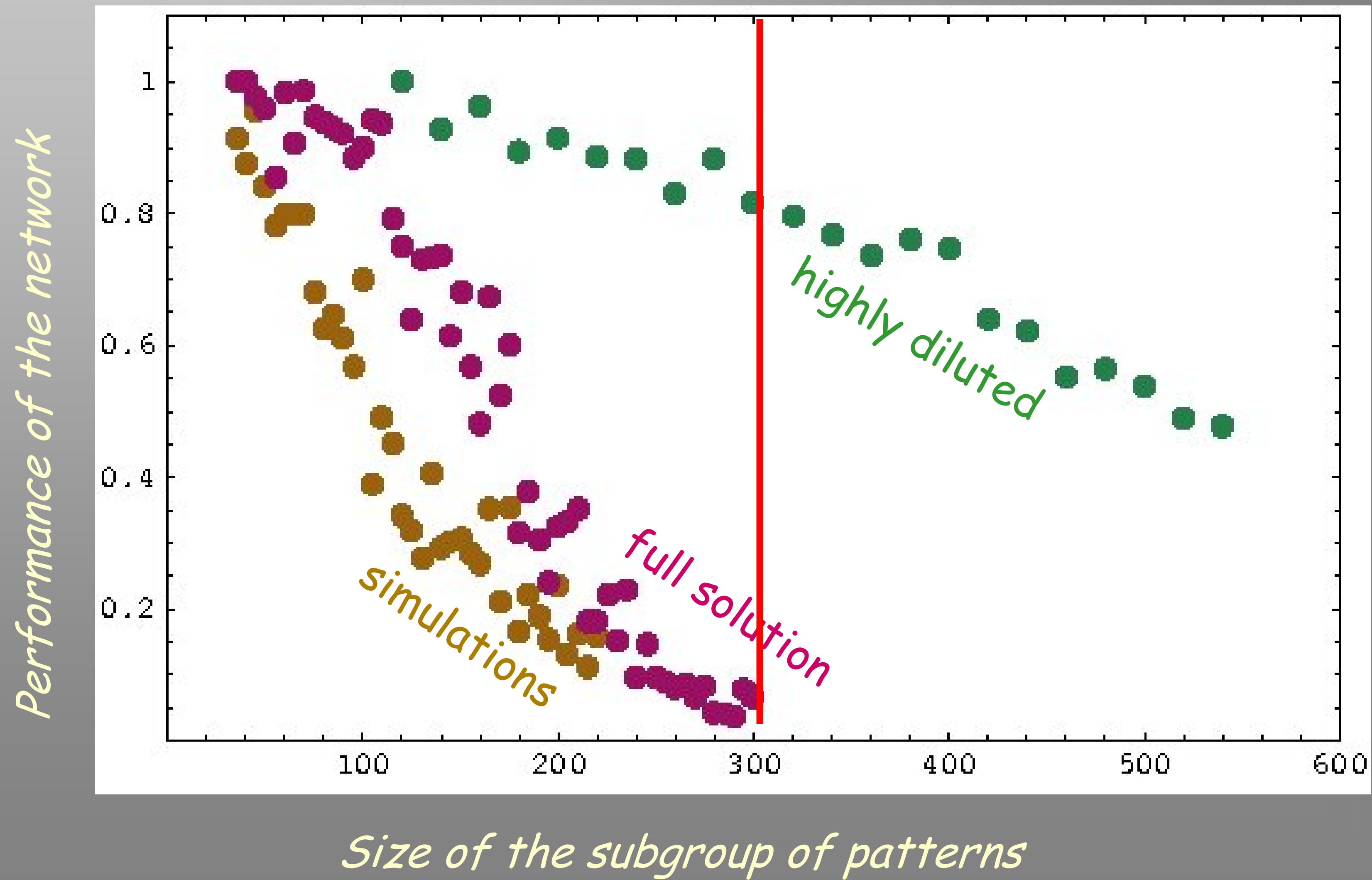


$$h_i \sim \xi_i^1 \, m + \sqrt{\alpha \, q \, a_i} \, z_i + \alpha \frac{c}{N} a_i (1 - a_i) \frac{\Omega}{1 - \Omega} \sigma_i$$

$$\phi + \phi^2 + \phi^3 + \ldots$$

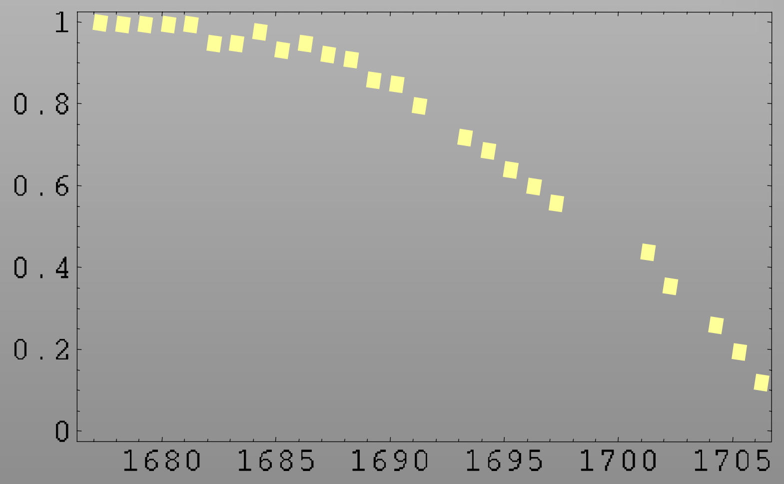
$$\Omega = \frac{1}{N} \sum_{j=1}^{N} a_j (1 - a_j) \partial(\sigma_j)$$

McRae's feature norms: the full solution



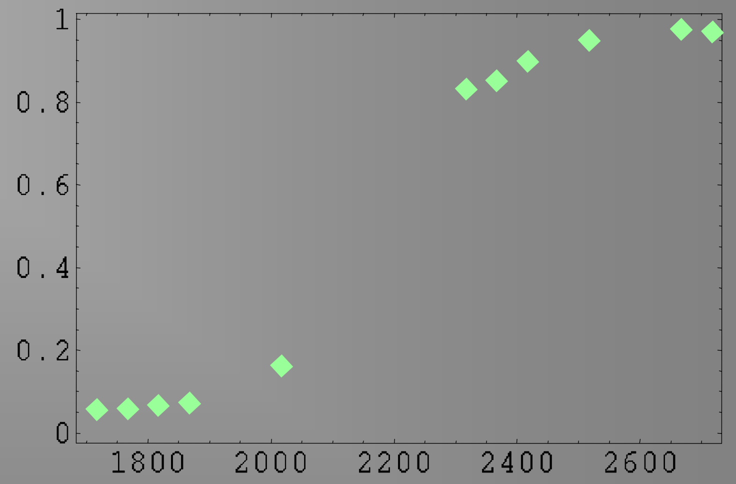
McRae's feature norms: strategies to store more patterns

1- kill popular neurons



20 most popular over 1700

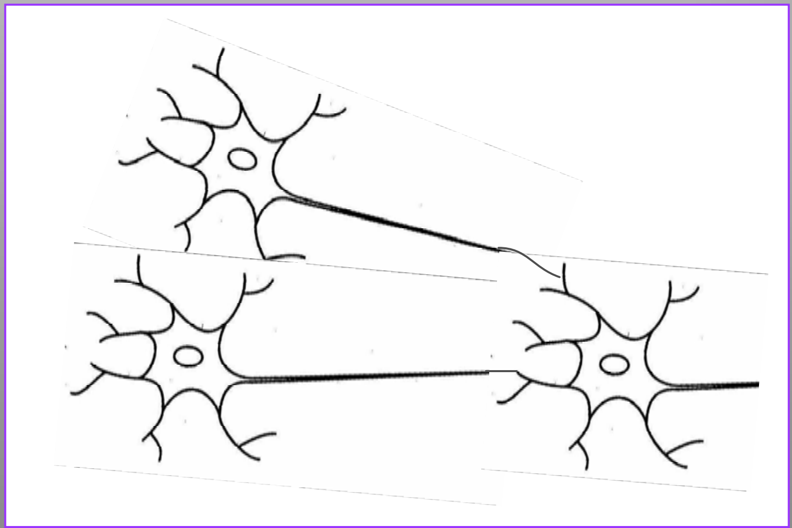
2- add unpopular neurons



800 ~ 2.7 features per pattern

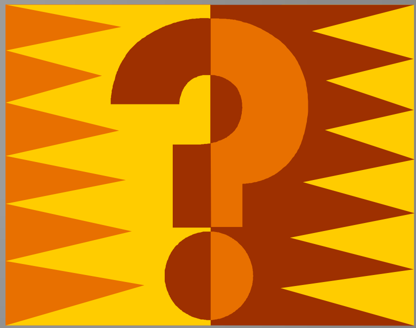
McRae's feature norms: strategies to store more patterns

3- recombination



neurons i and j have high popularity: their coincidence will be less popular. If applied massively, this principle could change the whole distribution.

4- popularity deppendent connectivity



The probability of having a connection from neuron i showld decrease with its popularity.

McRae's feature norms: plausibility of these strategies in the cortex

1- kill popular neurons

2- add unpopular neurons: thought to happen in DG to empoverish the correlation fed to the CA3 memory layer of Hippocampus.

3- recombination: found in association areas or perirhinal cortex. Could have something to do with improving storage capacity?

4- popularity deppendent connectivity

General Conclusions

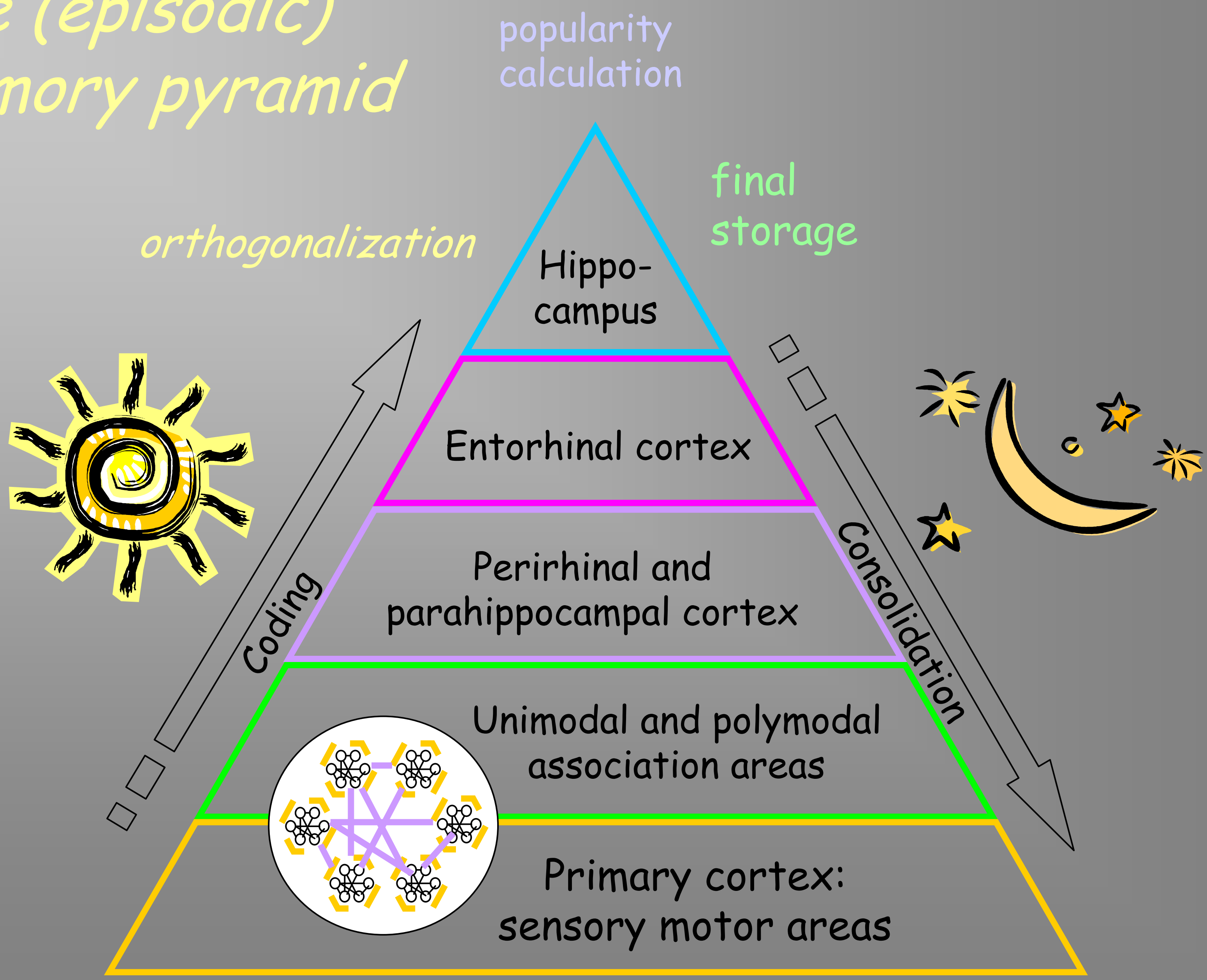
- An extension of the classical autoassociative memory model permits the storage of correlated patterns

- This storage has side-effects: memories are robust inversely to the information they carry

- The result supports accounts of category specific defficits based on correlation between patterns

- Uncorrelated memories are fast to learn while correlated memories need an intermediate step

The (episodic) memory pyramid



EBBS

Thank you