The Hippocampus and Associative Memory

Aims

•Understand how an associative memory matrix stores information through Hebbian learning and recurrent connectivity

•Explain what is meant by the terms content-addressable, pattern completion, error correction, interference, hetero-association and auto-association

•Describe how the anatomy of the hippocampal region CA3 is consistent with a role as an associative memory matrix and relate to its role in episodic memory

References

•Alvarez and Squire (1994) Memory consolidation and the medial temporal lobe: a simple network model. PNAS 91: 7041-7045.

•Burgess et al. (2001) Memory for events and their spatial context: models and experiments. Philosophical Transactions of the Royal Society London B 356 :1493-1503.

•Willshaw and Buckingham (1990) An assessment of Marr's theory of the hippocampus as a temporary memory store. Philosophical Transactions of the Royal Society London B 329: 205-215.

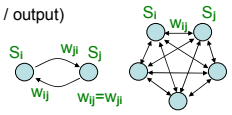
•McClelland, McNaughton and O'Reilly (1995) Why there are complementary learning systems in the hippocampus and neocortex. Psychological Review 102: 419-457.

•McNaughton and Morris (1987) Hippocampal synaptic enhancement and information storage in a distributed memory system. Trends in Neurosciences 10: 408-414.

•Wills et al. (2005) Attractor dynamics in the hippocampal representation of the local environment. Science 308: 873-876.

Hopfield (1982) Associative Memory Network

- Fully connected recurrent network (no input / output)
- Symmetric connection weights ($w_{ij} = w_{ji}$)
- Units are active ($S_i = 1$) or inactive ($S_i = -1$)



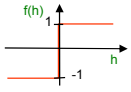
Learning: impose activation pattern and use 'Hebbian' rule to adjust weights

$$w_{ij} \rightarrow w_{ij} + \varepsilon S_i S_j$$

Δw_{ij}	S_i = -1	S_i = 1
S_j = -1	↑	↓
S_j = 1	↓	↑

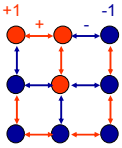
Recall: start from similar pattern of activation, change activation according to sign of input to recover original pattern

$$s_i = \text{sign}(h_i) \qquad h_i = \sum_j w_{ij} s_j$$



Hopfield (1982) Associative Memory Network

Patterns of activation are learned as 'stable states' under the rule for updating activations



Δw_{ij}	S_i = -1	S_i = +1
S_j = -1	↑	↓
S_j = +1	↓	↑

Several different patterns can be learned in the same network, but the memory capacity is limited to ~0.14N (where N is the number of neurons)

Memory is content addressable, performing 'pattern completion' of a partial cue

Spurious memories (combinations of real memories) may be formed

More plausible learning rules show similar behaviour

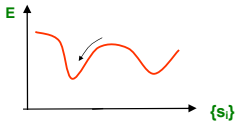
Hopfield Networks: Attractors and Stable States

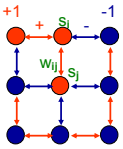
To support a pattern of activation, connections should be positive between units in the same state (i.e. 1 / 1 or -1/ -1) and negative between units in different states (1/ -1 or -1 / 1) - i.e. $s_i s_j w_{ij} > 0$

The 'frustration' or 'energy' of the system is how much this is not true – i.e. $E = -\sum_{ij} s_i s_j w_{ij}$

The update rule changes each units activation to reduce the overall frustration, until the network ends up in a stable state from which frustration cannot be further reduced

The learning rule sets the weights so that to-be-remembered patterns of activity are stable or 'attractor' states





Examples of Hopfield Networks

A 5x9 network storing 8 patterns

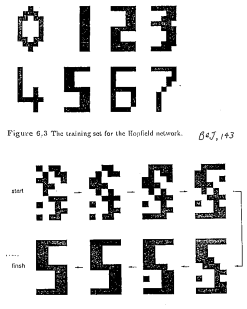
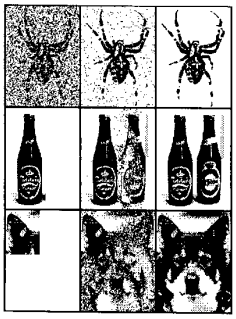


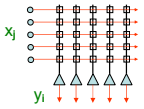
Figure 6.3 The training set for the Hopfield network.

Retrieval in a 130x180 network



Memory Matrices

A feed-forward single-layer neural network can be drawn as:



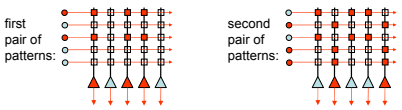
axon
input neurons x_j = 0 or 1
output neurons y_i = 0 or 1
synapse w_{ij} = 0 or 1
dendrite

Learning: $w_{ij} = \max(w_{ij}, x_j y_i)$

Output: threshold or divide by no. active inputs (N) so that:

$$y_i = \begin{cases} 1 \text{ if } \sum_j w_{ij} x_j \geq N \\ 0 \text{ otherwise} \end{cases} \qquad \text{or: } y_i = [\sum_j w_{ij} x_j / N]_{\text{floor}}$$

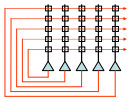
This network learns input-output associations: 'hetero-association'



Willshaw (1969); Marr (1971); McNaughton & Nadel (1990)

Memory Matrices

A recurrent (feed-back) neural network can be drawn as:



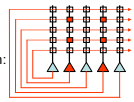
axon
neurons — x_i = 0 or 1
synapse — w_{ij} = 0 or 1
dendrite

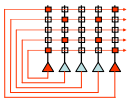
Learning: $w_{ij} = \max(w_{ij}, x_j y_i)$

Output: threshold or divide by no. active inputs (N) so that:

$$y_i = \begin{cases} 1 \text{ if } \sum_j w_{ij} x_j \geq N \\ 0 \text{ otherwise} \end{cases} \quad \text{or: } y_i = [\textstyle\sum_j w_{ij} x_j / N]_{\text{floor}}$$

This network learns to associate a pattern of activity with itself: 'auto-association'

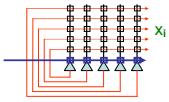
first pattern: 

second pattern: 

This represents a simpler, more realistic version of the Hopfield model

Memory Matrices

A recurrent (feed-back) neural network can be drawn as:



Need to be able to impose a new pattern of activity to be learned, while ignoring feedback from the current pattern: "detonator synapses"?

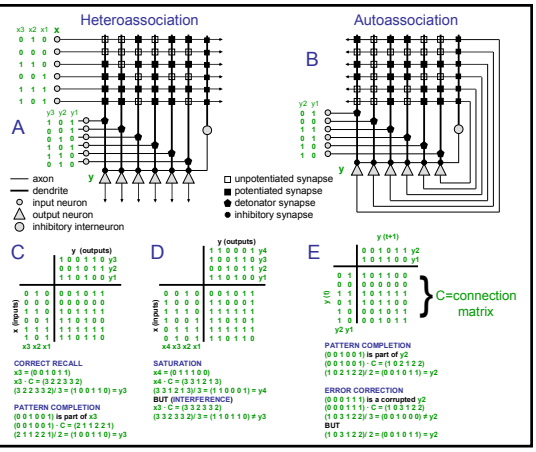
Summary of memory matrices

- Analagous to the Hopfield auto-associative network but:
  - connection weights are 0 / 1 and don't need to be symmetric
  - connection weights only increase (with pre-and post-synaptic activity)
  - neuron activation values are 0 / 1 (not -1 / 1)
- Performs pattern completion and error correction
- Prone to interference – need to use non-overlapping (e.g. sparse) codes
- Need to distinguish learning from recall

Willshaw (1969); Marr (1971); McNaughton & Nadel (1990)

'Memory Matrices': worked examples

(McNaughton and Nadel, 1990)

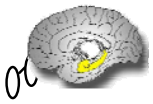


The Hippocampus as an Associative Memory Network

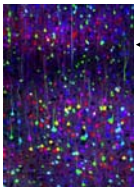
The human brain is made of very many neurons (~100bn) sending electrical signals to each other

These are particularly closely interconnected in the hippocampus

Human



The 'brainbow' mouse
(Livet et al. 2007)

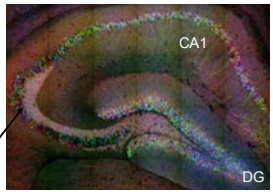


Neocortex

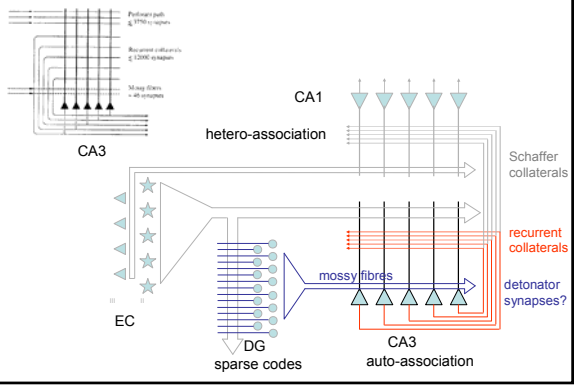
Mouse

CA3 recurrent connections

The hippocampus

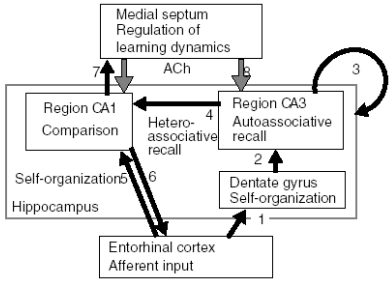


The Hippocampus as an Associative Memory Network



The Hippocampus as an Associative Memory Network

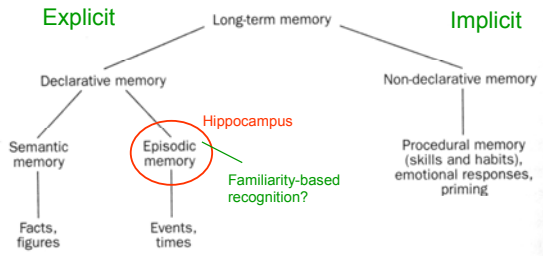
(b) Computational model inc. novelty/ACh switching between learning and recall



Hasselmo et al. (1995)

The Hippocampus as an Associative Memory Network

Categories of human memory



The Patient HM

Severe epilepsy, treated with surgery to bilaterally remove medial temporal lobes

Bilateral removal of temporal pole, amygdala, entorhinal cortex and anterior hippocampus. Posterior hippocampus is sclerotic

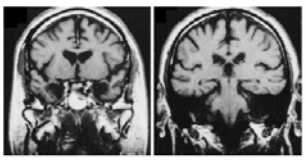
Operation: September 1953 (27 years old)

Memory tested: April 1955, 29 years old)

No new memories formed since operation. Reported date as March 1953.

IQ better than before the operation (112) and fewer seizures

Can't find new home (after 10 months) remember new people, names etc.



Scoville and Milner (1957); Corkin et al. (1997)

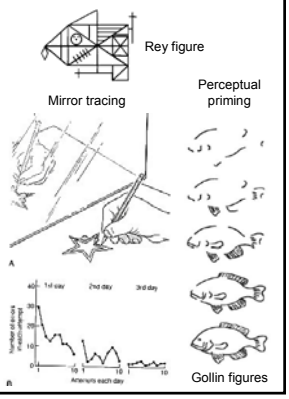
The Patient HM

Impaired explicit / declarative memory

- Episodic memory of events / people since operation; location of new home; Rey figure: can copy but not recall
- Semantic memory of new vocabulary, speech is frozen in 1950's (Gabrieli et al. 1988) with the exceptions: "ayatollah", "rock 'n' roll"

Preserved implicit / procedural memory

- Perceptual priming with Gollin figures (Milner, 1986); Mirror tracing task (Milner 1962, 1965); Pursuit rotor tracing (Corkin, 1968)



The Patient Jon

Jon has developmental amnesia

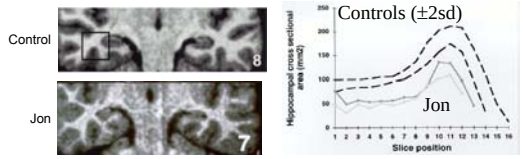
Peri-natal anoxia caused bilateral hippocampal damage, 50% volume reduction

Jon was premature, born at 26 weeks, on artificial ventilation for two months

Memory impairments first noticed at 5½ years of age

Jon has impaired episodic memory but preserved semantic and recognition memory, a normal vocabulary, and obtained one GCSE (in History)

Highlights a distinction between episodic memory and semantic memory / familiarity-based recognition?



Varga-Khadem et al. (1997)

The Patient Jon

This distinction can be demonstrated using the 'doors and people' difficulty-matched test of verbal / visual recognition / recollection

Four names for verbal recall:

Four shapes for visual recall:

Twelve names for verbal recognition:

JILL ASHLEY

JILL ASHDON
JILL ASHMAN
JILL ASHLEY
JILL ASHTON

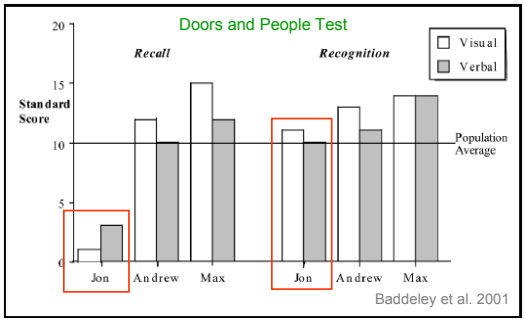
Twelve doors for visual recognition:

Baddeley et al. (1994)

The Patient Jon

Jon shows impaired recall but preserved unimodal recognition

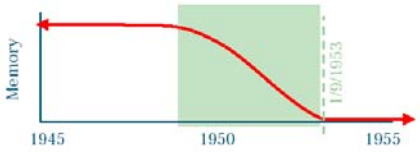
Cross-modal recognition - e.g. voice-face or object-location pairs - is impaired



Baddeley et al. 2001

Temporal Gradients in Retrograde Amnesia

HM – Old (childhood) memories undamaged
– Memories from 5-10 years before the lesion are lost
– Forgot death of favourite uncle in 1950



This might imply that hippocampal / MTL memories are 'consolidated' in the neocortex over time and therefore become hippocampal independent

Marr (1971); Alvarez and Squire (1994)

Temporal Gradients in Retrograde Amnesia

HM: photos of celebrities => more distant memories are relatively preserved
(Marslen-Wilson & Teuber, 1975)

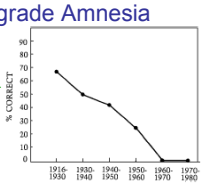
PZ: Wrote autobiography, tested on personal memories
(Butters & Cermak, 1986)

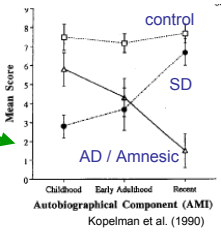
But are test items equally salient across decades?
(Warrington & Weiskrantz, 1970)

VC: Hippocampal damage and severe amnesia extending to childhood
(Cipolotti et al., 2001)

Are old memories truly episodic or do they become semantic / procedural? For example, HM uses the same words in each description
(Graham & Hodges, 1997)

Semantic Dementia => reverse gradient



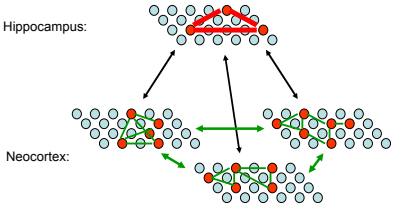


Kopelman et al. (1990)

Fast and Slow Learning in Memory Consolidation

Model
• Hippocampal recurrent collaterals for rapid ('one-shot') associative learning
• Slow learning in neocortex, trained by hippocampus

Issues
• Transfer vs training?
• Timescale? One night or twenty years → graded retrograde amnesia?
• Generalisation / abstraction of meaning → semantic memory?



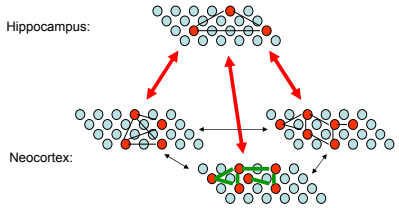
Marr (1971); Willshaw and Buckingham (1990); Alvarez and Squire (1994); McClelland et al. (1995)

Fast and Slow Learning in Memory Consolidation

Hippocampus is an index or convergence zone for disparate neocortical sites which store the memory content

This index represents cross-modal binding, and may be spatial / temporal context?

Familiarity-based unimodal recognition of content is therefore possible without the hippocampus



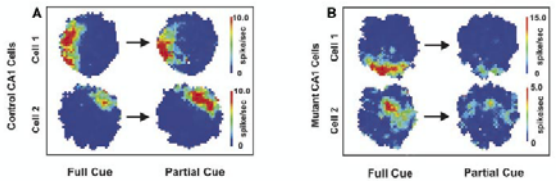
Teyler and DiScenna (1986); Damasio (1989)

Experimental Evidence for the Attractor Network Model

• Place cell firing is robust to the removal of subsets of cues

• This property is CA3 NMDA dependent

• Place cell firing is strongly affected by interchanging cues

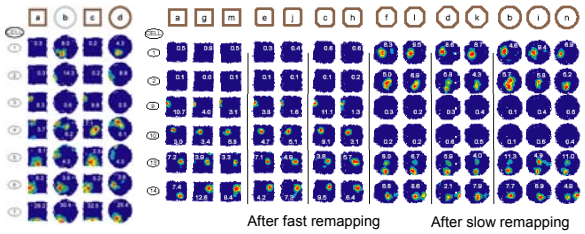


Nakazawa et al. (2002)

Experimental Evidence for the Attractor Network Model

Fast remapping…

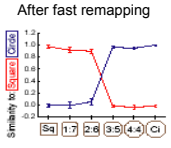
...creates attractor representations of 2 environments

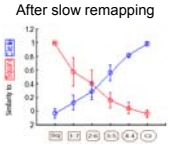


After fast remapping

After slow remapping

Memory retrieval from partial cues → 'pattern completion' in an attractor network of place cells





Wills et al. (2005)