

Online Learning: A Minimalist Example

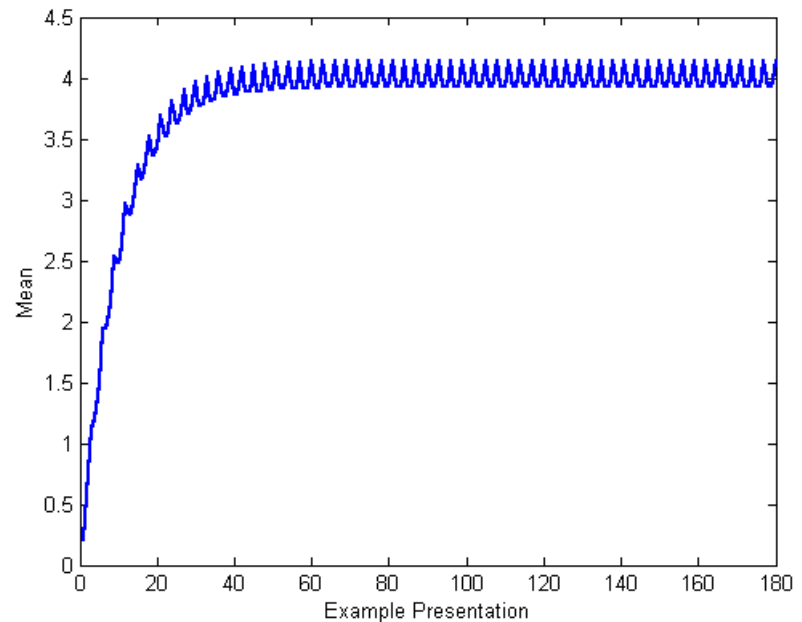
Approximate the running average of: $x_1, x_2, \dots, x_n$

$$\Delta\mu = \varepsilon (x_i - \mu)$$

Example: 2, 4, 6

$\varepsilon = 0.1$

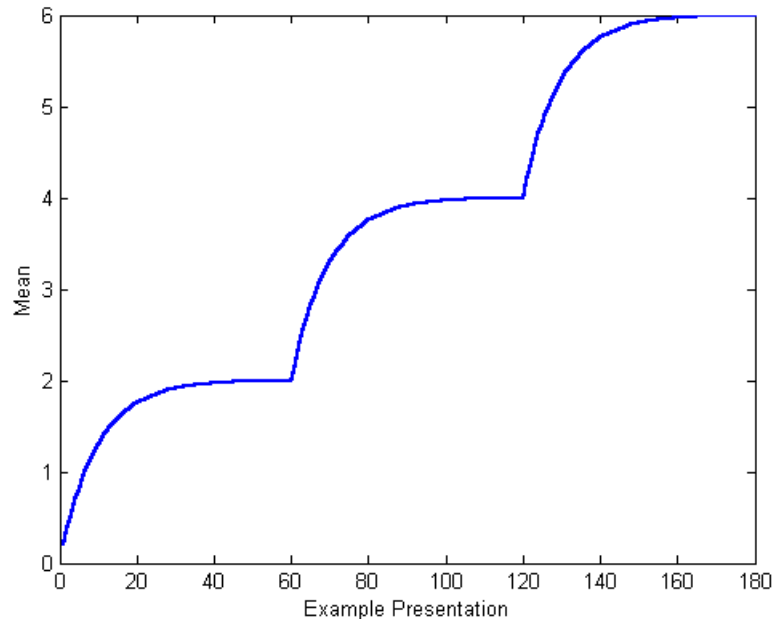
60 epochs



Catastrophic Interference I: Focused Learning

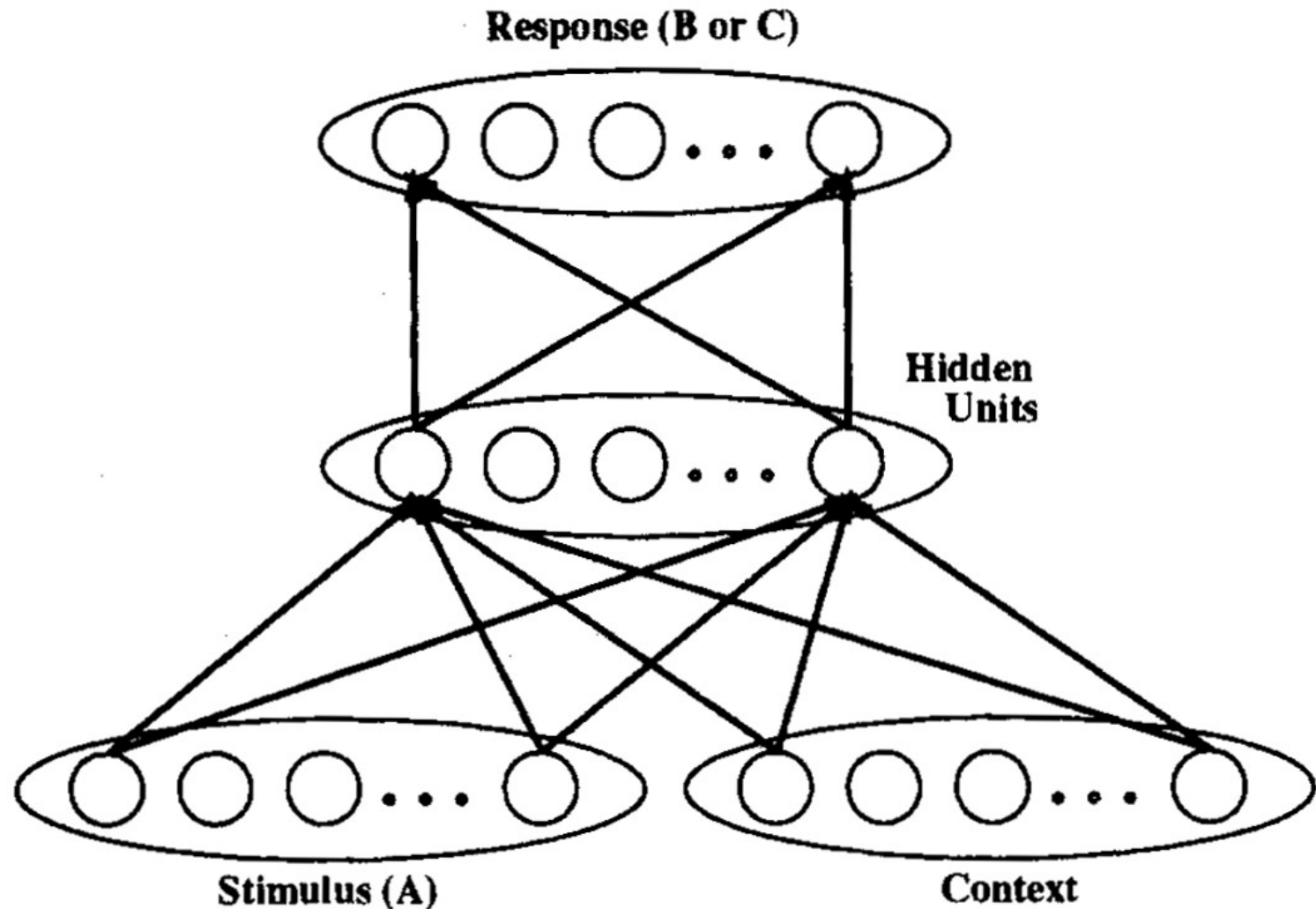
We presented 60 times {2, 4, 6} (interleaved learning).

What would happen if instead we presented:
60 times {2}, then 60 times {4}, then 60 times {6} (focused learning)?



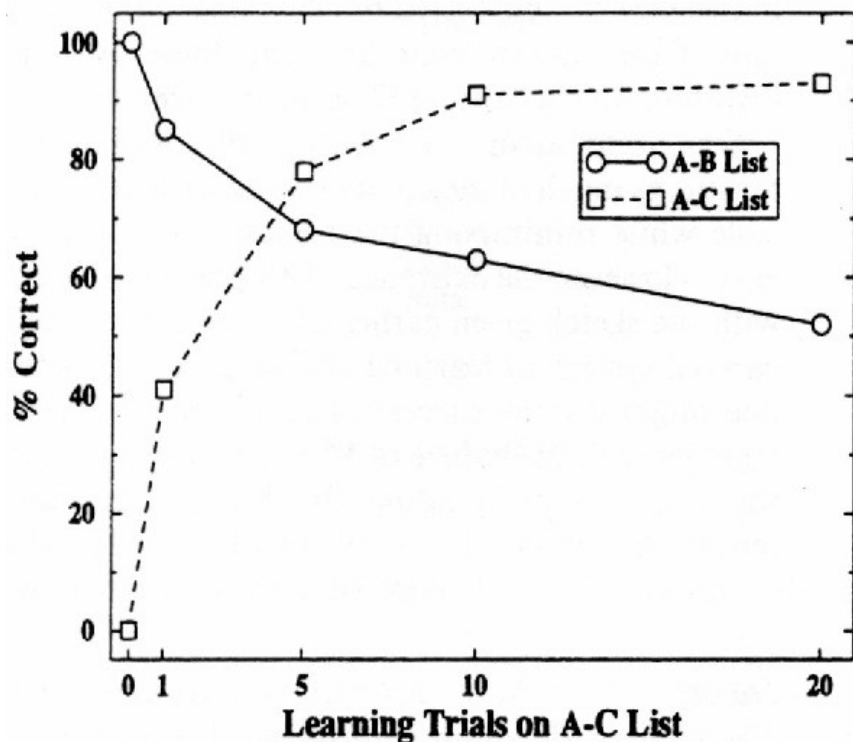
Catastrophic interference: Learning about new stuff in a focused manner makes you forget the old stuff.

Catastrophic Interference and Focused Learning: A More Interesting Example

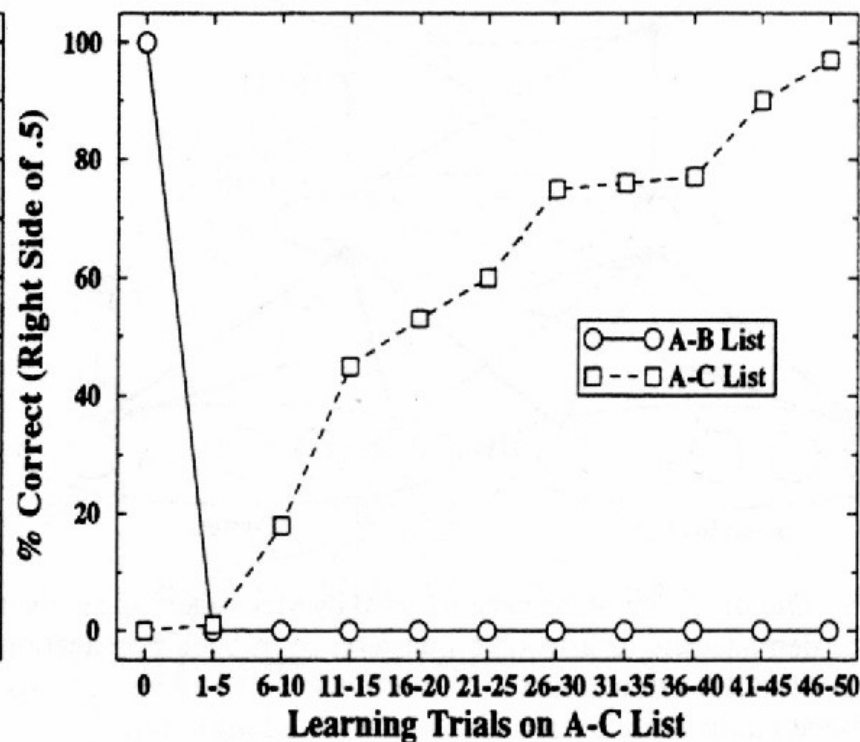


Catastrophic Interference and Focused Learning: A More Interesting Example

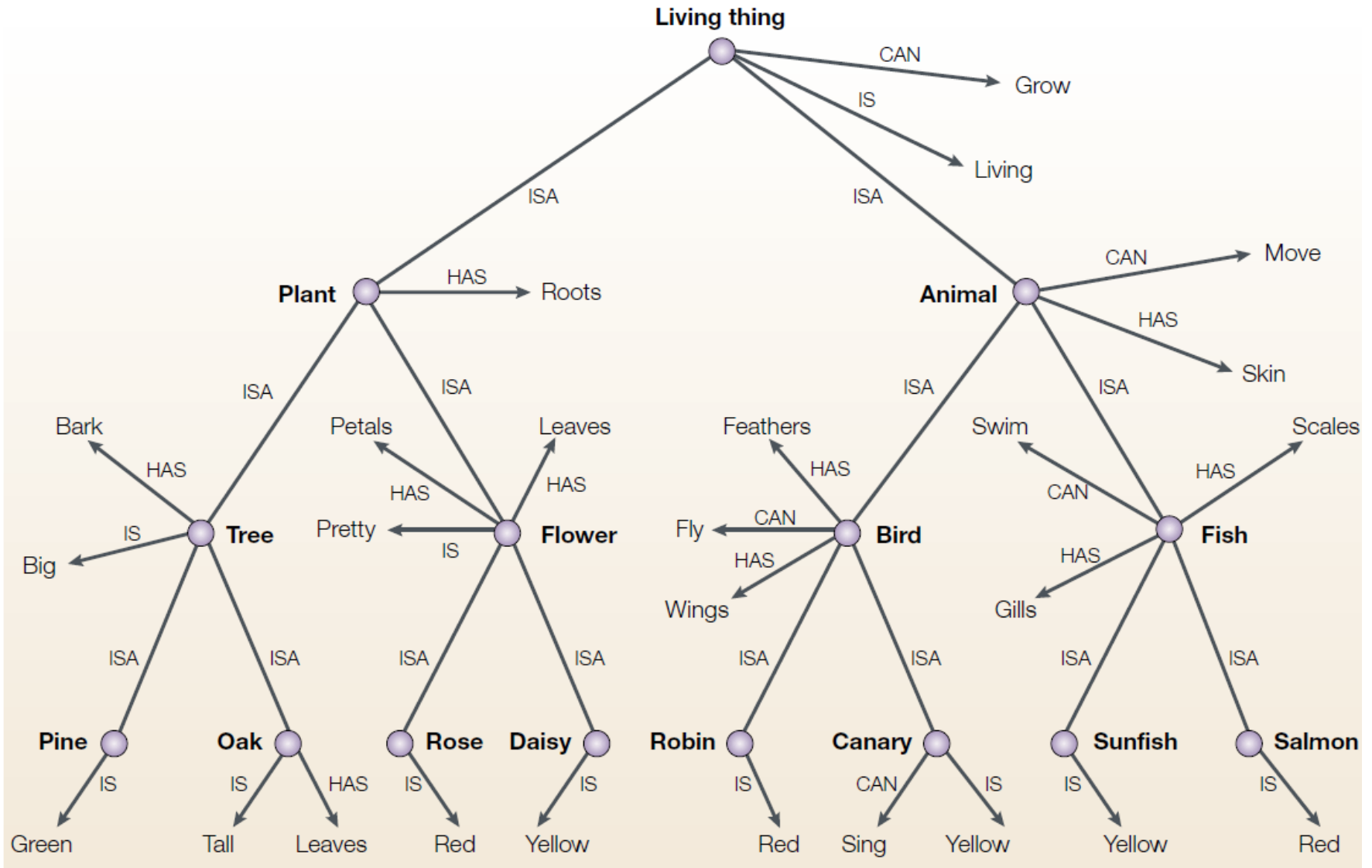
a) AB-AC List Learning in Humans



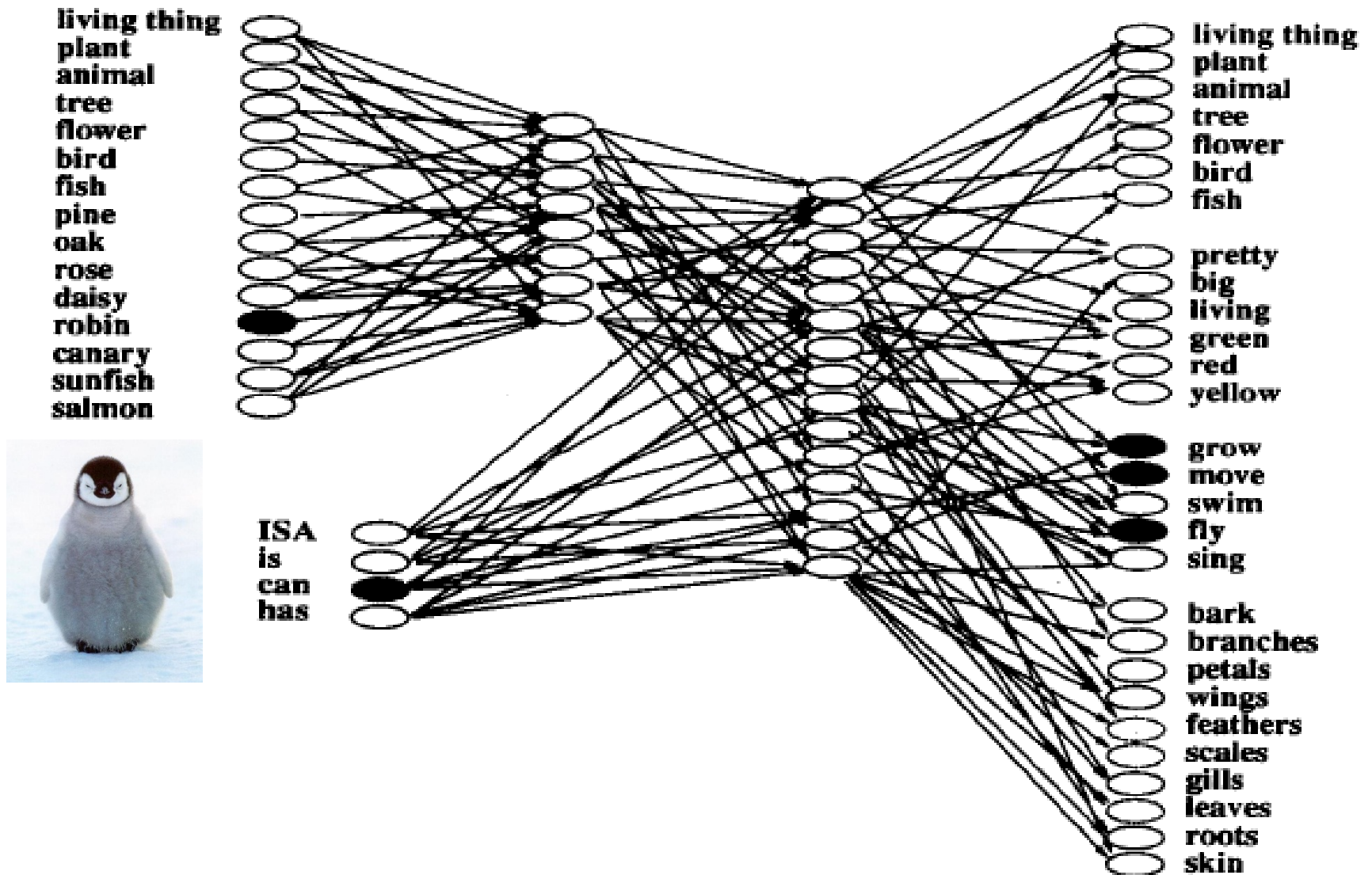
b) AB-AC List Learning in Model



Rumelhart and Todd (1993)

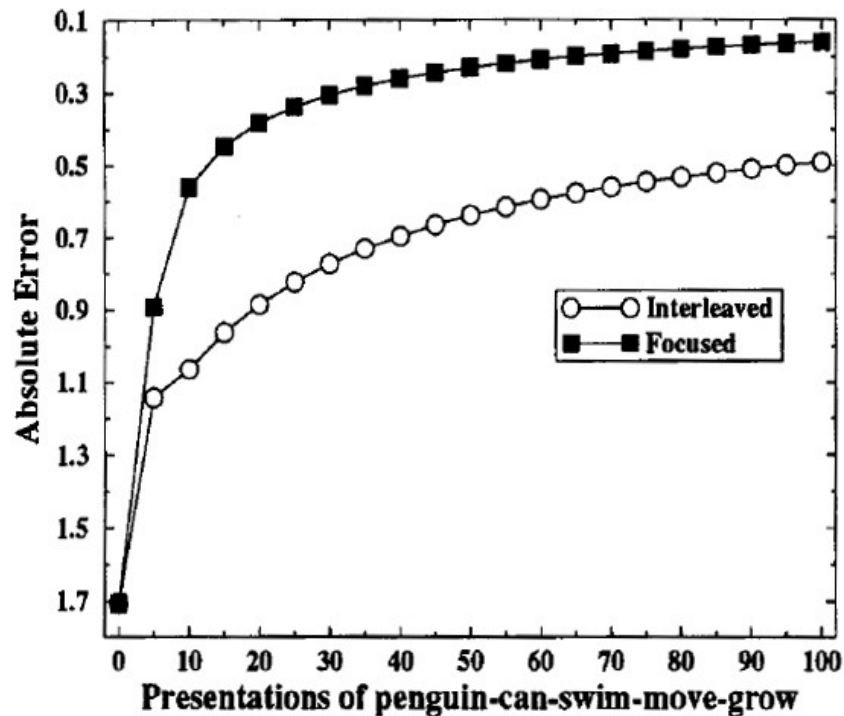


Rumelhart and Todd (1993)

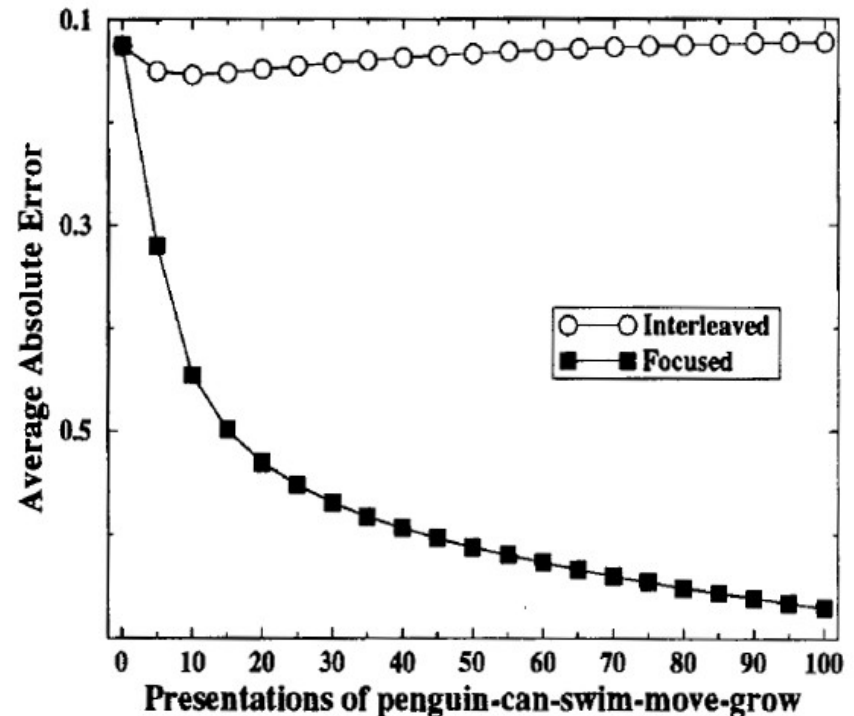


The Benefits of Interleaved Learning

a) Aquisition of New Information



b) Interference with Existing Memories

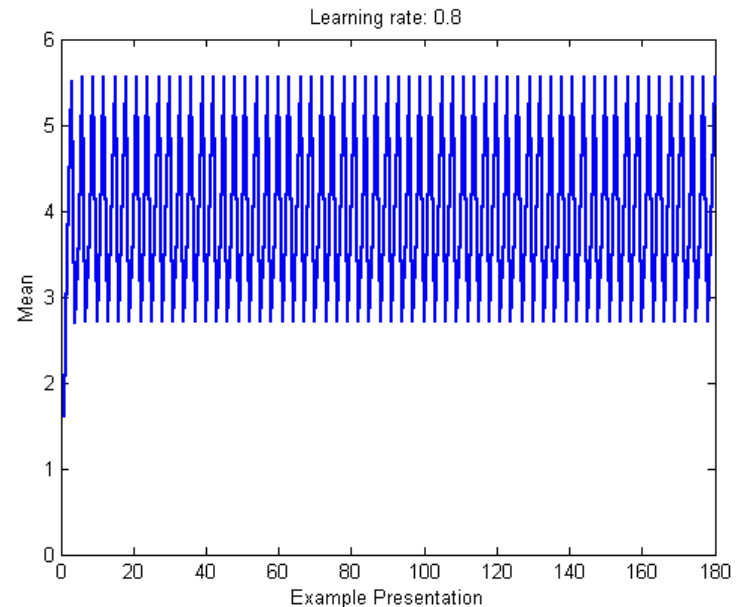
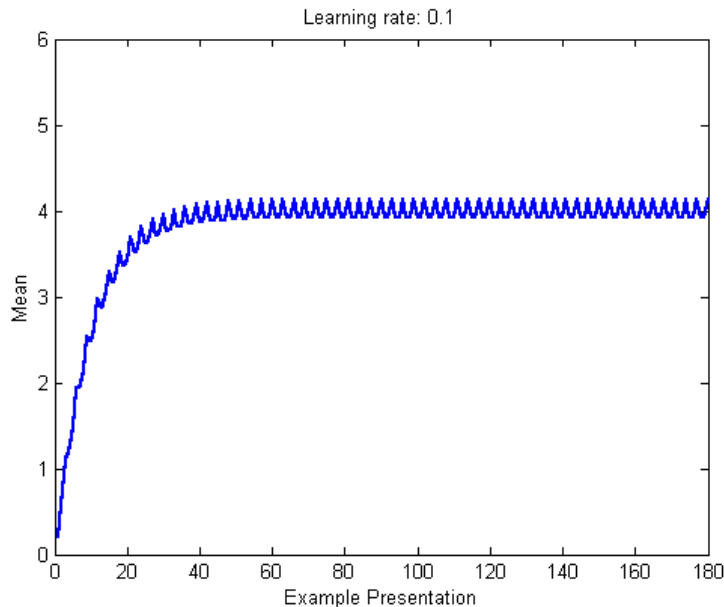


Catastrophic Interference II: Role of Learning Rate

Back to our toy example: $\Delta\mu = \varepsilon (x_i - \mu)$

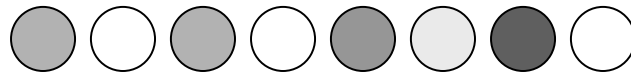
Assuming interleaved learning, should ε be large or small?

Answer: Smaller ε will lead to more accurate (but slower) learning.



Distributed Versus Sparse Representations

Distributed representations: Each pattern is represented across most units

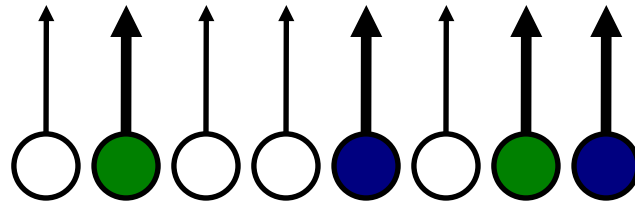


Sparse representations: Each pattern is represented in a small percentage of units



Catastrophic Interference III: Distributed Versus Sparse Representations

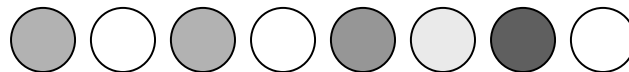
Sparse representations: Learning about one pattern doesn't affect what you know about other patterns



This is good: It prevents interference.

This is also bad: Doesn't allow generalization.

Distributed representations: Learning about one pattern affects what you know about other patterns



This is good: It allows generalization.

This is also bad: Interference.

Incompatible Goals

Quickly learn specific information (e.g., your neighbor's dog bit you)	Learn general information (e.g., what dogs are)
Requires	
Fast learning (often one example)	Slow learning (generalize over many examples)
Focused learning	Interleaved learning (don't want to forget all about cats when you learn about dogs)
Sparse representations (avoid interference)	Distributed representations (allow generalization)

Are we stuck?

Complementary Learning Systems

If we can't have it all in the same system, let's have two systems!

System 1: Remember specifics, i.e., learn quickly about new information and minimize interference

High learning rate, sparse representations

System 2: Learn slowly the overall structure of the environment

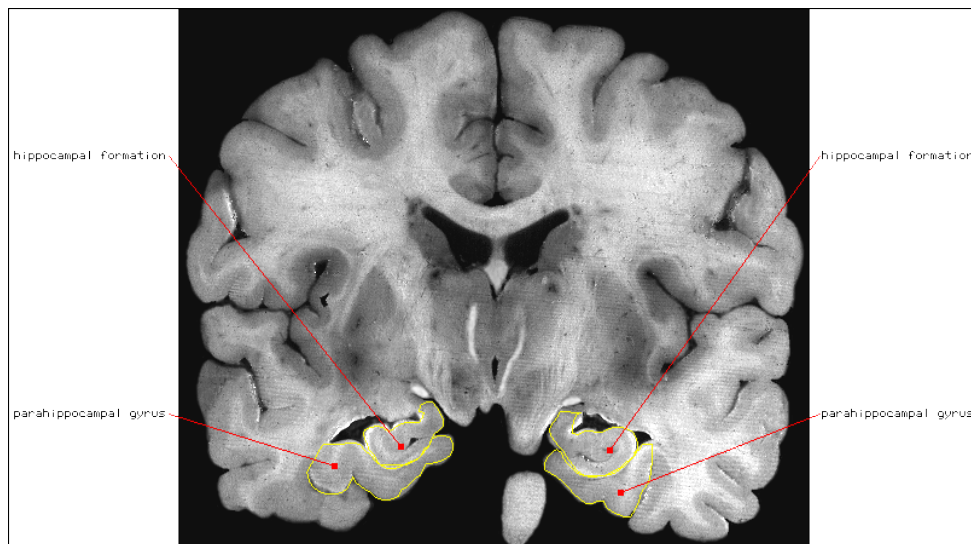
Small learning rate, distributed representations

Requires interleaved learning

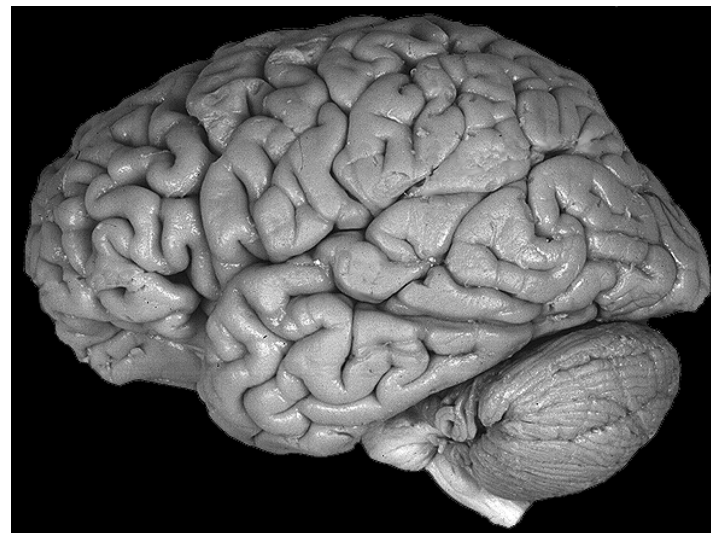
System 1 trains System 2 with many interleaved presentations of the individual examples that System 1 learns quickly!

Complementary Learning Systems in the Brain

It seems that the brain has two such systems!

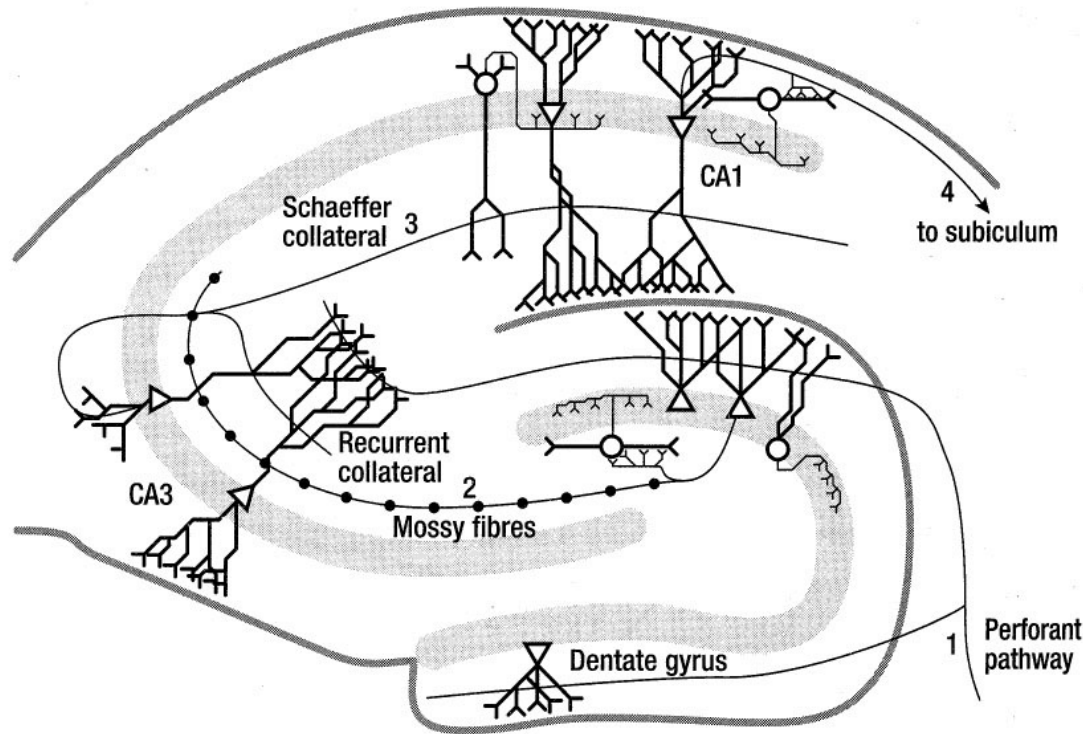


Fast learning of specific
information:
Hippocampal system



Slow learning of general
information:
Neocortex

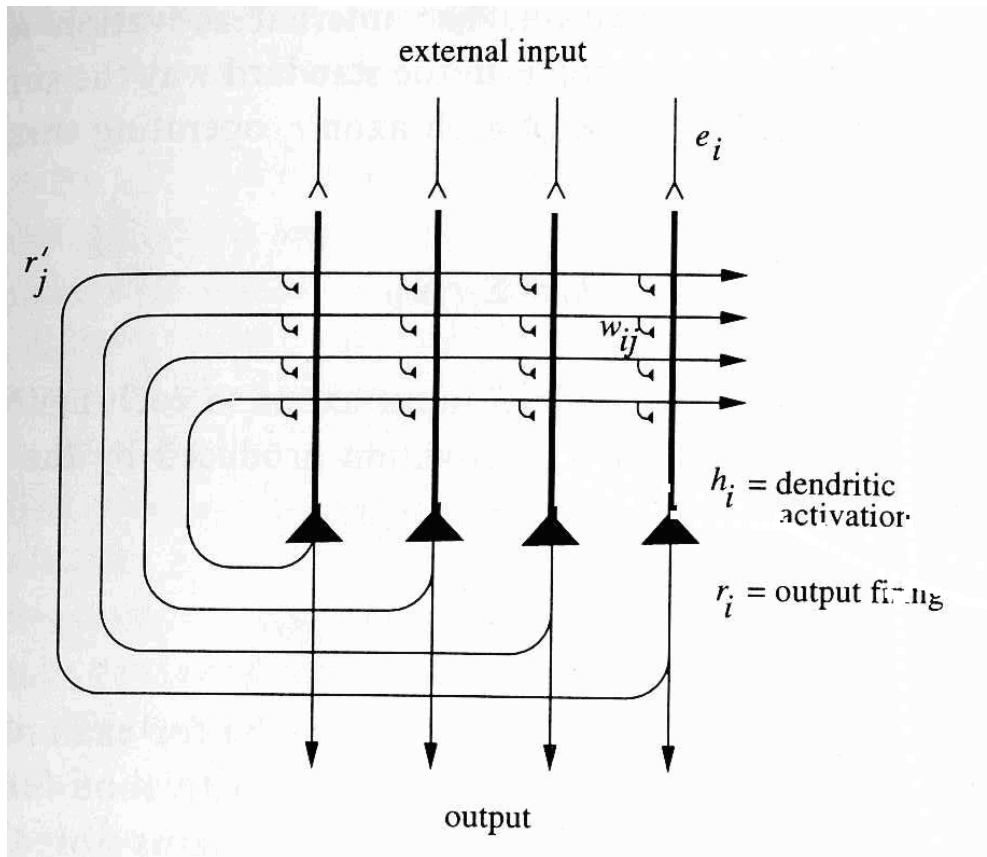
An Autoassociative Network in Region CA3 of the Hippocampus?



Some people have suggested that the extensive recurrent collaterals in CA3 form an autoassociative memory (e.g., Rolls, 1998)

The other areas may be involved in compressing/decompressing information

The Hippocampus as an Autoassociator



An autoassociative network

Hebbian learning: $\Delta w_{ij} = \epsilon a_i a_j$

Features of
autoassociative networks

Pattern completion: Give the network a bit of a stored pattern (i.e., a cue) and it can reproduce (i.e., recall) the rest

Learning is fast if ϵ is high (approx. 1)

Interference is minimized if patterns don't overlap

Complementary Learning Systems in the Brain

Hippocampal system: Remember specifics, i.e., learn new information quickly and minimize interference

High learning rate, sparse representations

Neocortex: Learn slowly the overall structure of the environment

Small learning rate, distributed representations

Requires interleaved learning

Hippocampus trains neocortex over time with many interleaved presentations of the specific memories the hippocampus has stored

Anterograde and Retrograde Amnesia: Humans

After lesions to the hippocampal system, humans:

Cannot form new explicit memories (e.g., episodic memories): **anterograde amnesia**

Lose most such memories that were laid out (relatively) shortly before the lesion, *but can remember older ones*:
temporally graded retrograde amnesia

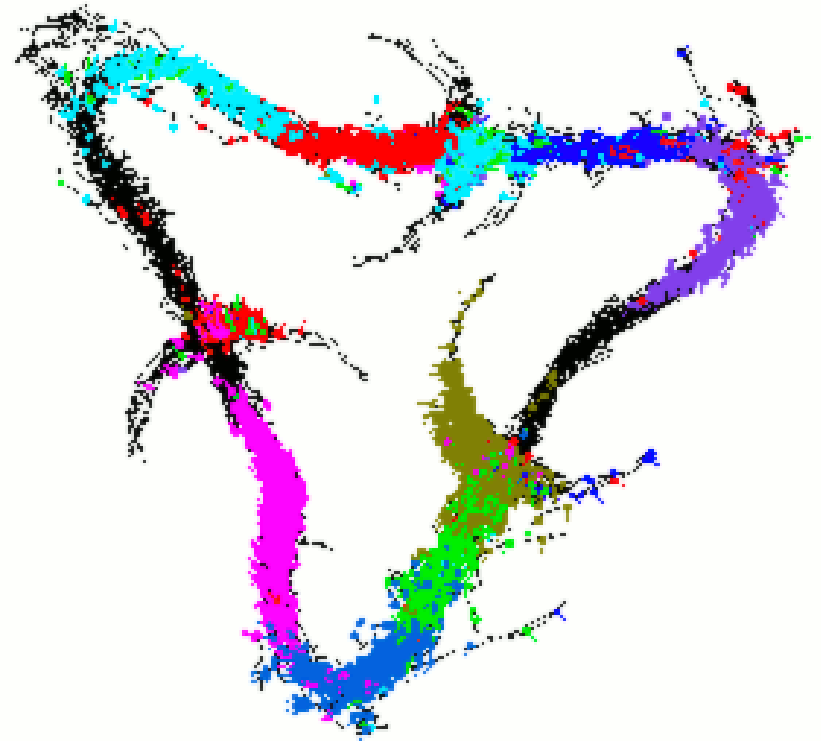
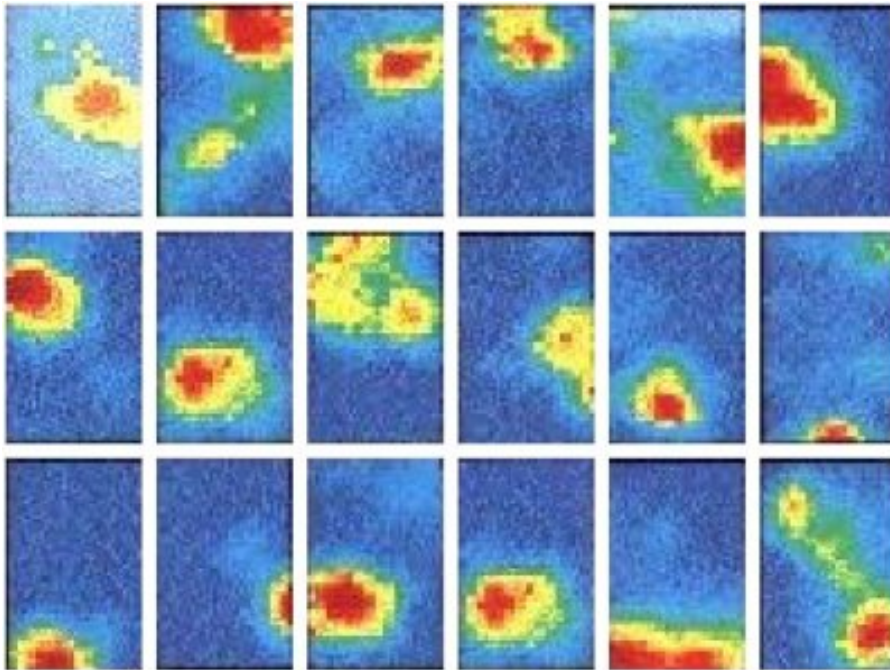
Anterograde and Retrograde Amnesia: Animals

After lesions to the hippocampus, animals show anterograde and temporally graded retrograde amnesia for memories that involve **cue configurations**

Examples:

- Negative patterning
tone → reward; light → reward; tone & light → no reward
- Spatial learning

Place Cells in the Rat Hippocampus



Role of the Hippocampus

Overall, the hippocampus seems to be important for fast learning of arbitrary conjunctive representations

- Episodic memories

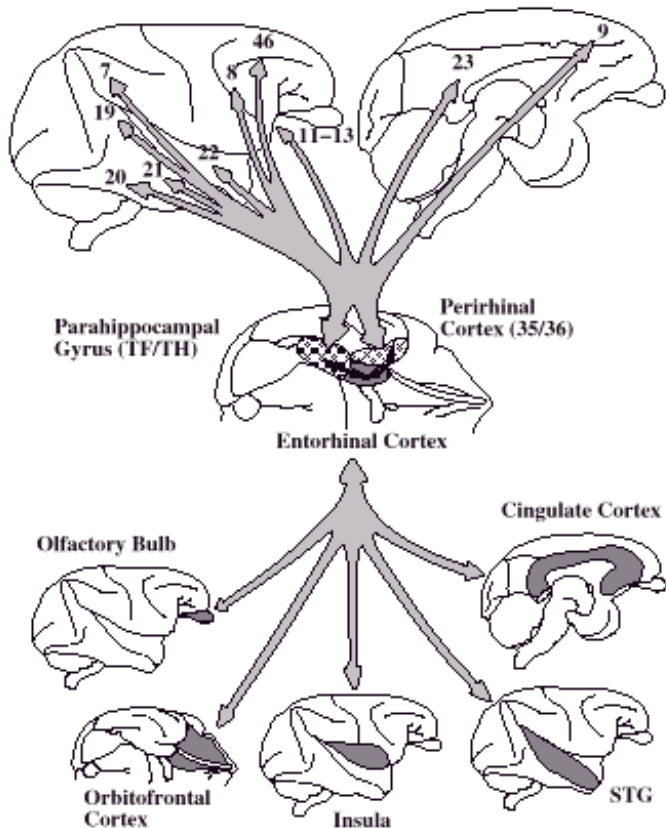
- Spatial learning

- Contextual fear conditioning

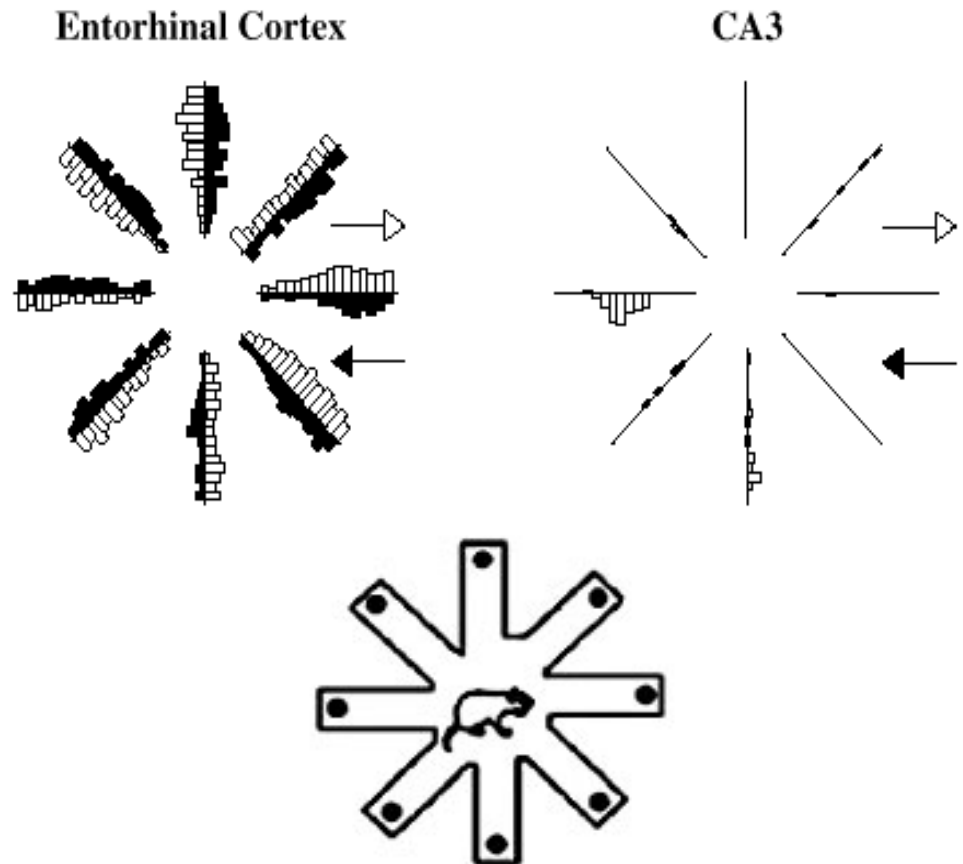
Properties of the Hippocampus

For fast learning of arbitrary conjunctive representations, one needs:

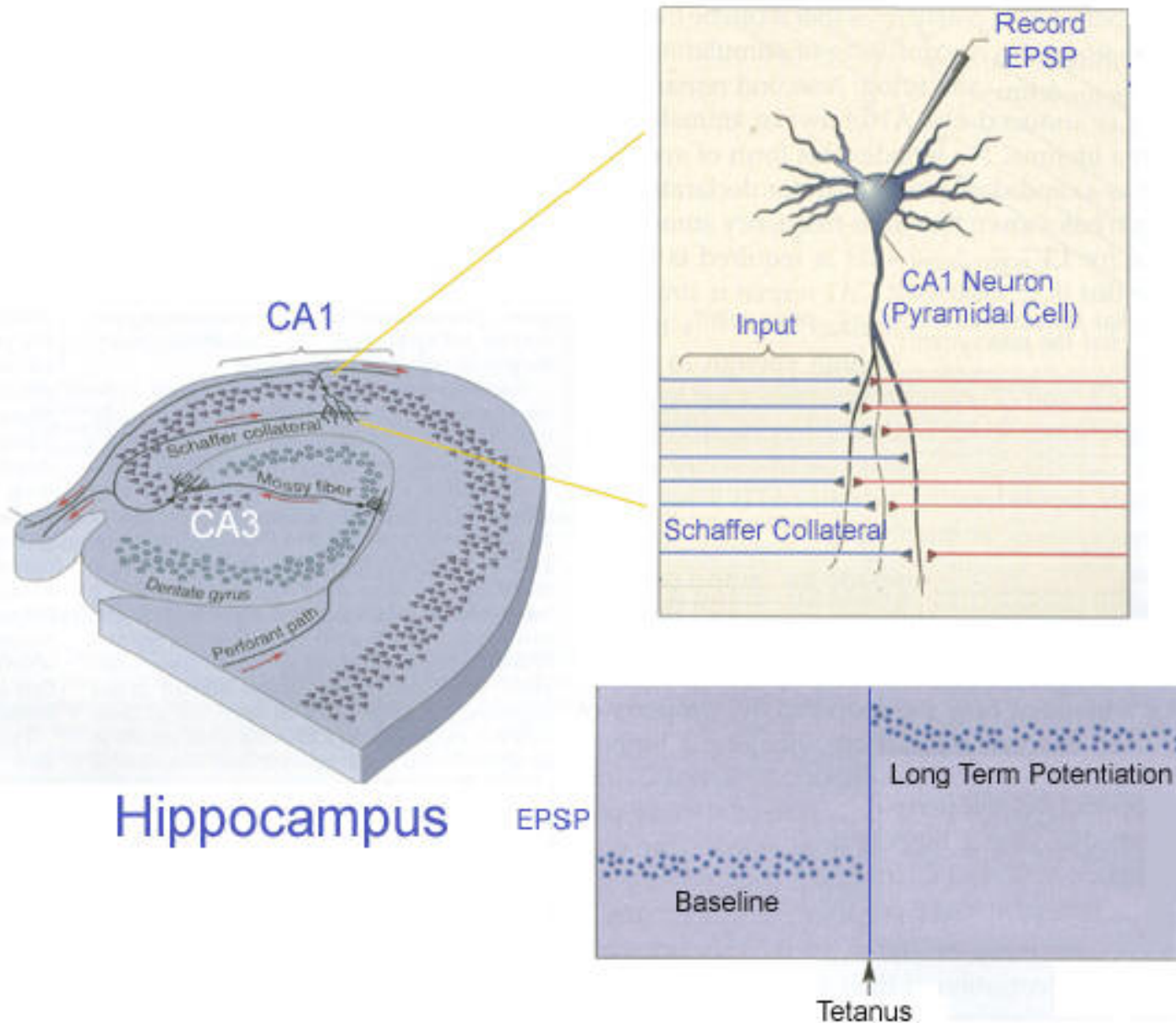
Widespread connectivity
to many brain areas



Sparse representations
(to minimize interference)



Fast learning: Long-term potentiation (LTP)



Preserved Learning and Memory After Hippocampal Lesions

Skills (e.g., tracing a figure viewed in a mirror, reading mirror-reversed print, learning how to ride a bicycle)

Suggests a spared system for slow, gradual learning: Neocortex

Repetition priming (e.g., reading aloud a pronounceable non-word)

Some forms of **classical conditioning**, such as simple associations between CS and US (e.g., tone → shock)

Likely dependent on other systems (e.g. amygdala):
Multiple memory systems (e.g., Rolls, 2000)

Where Are We?

Hippocampus seems good for fast learning of arbitrary associations

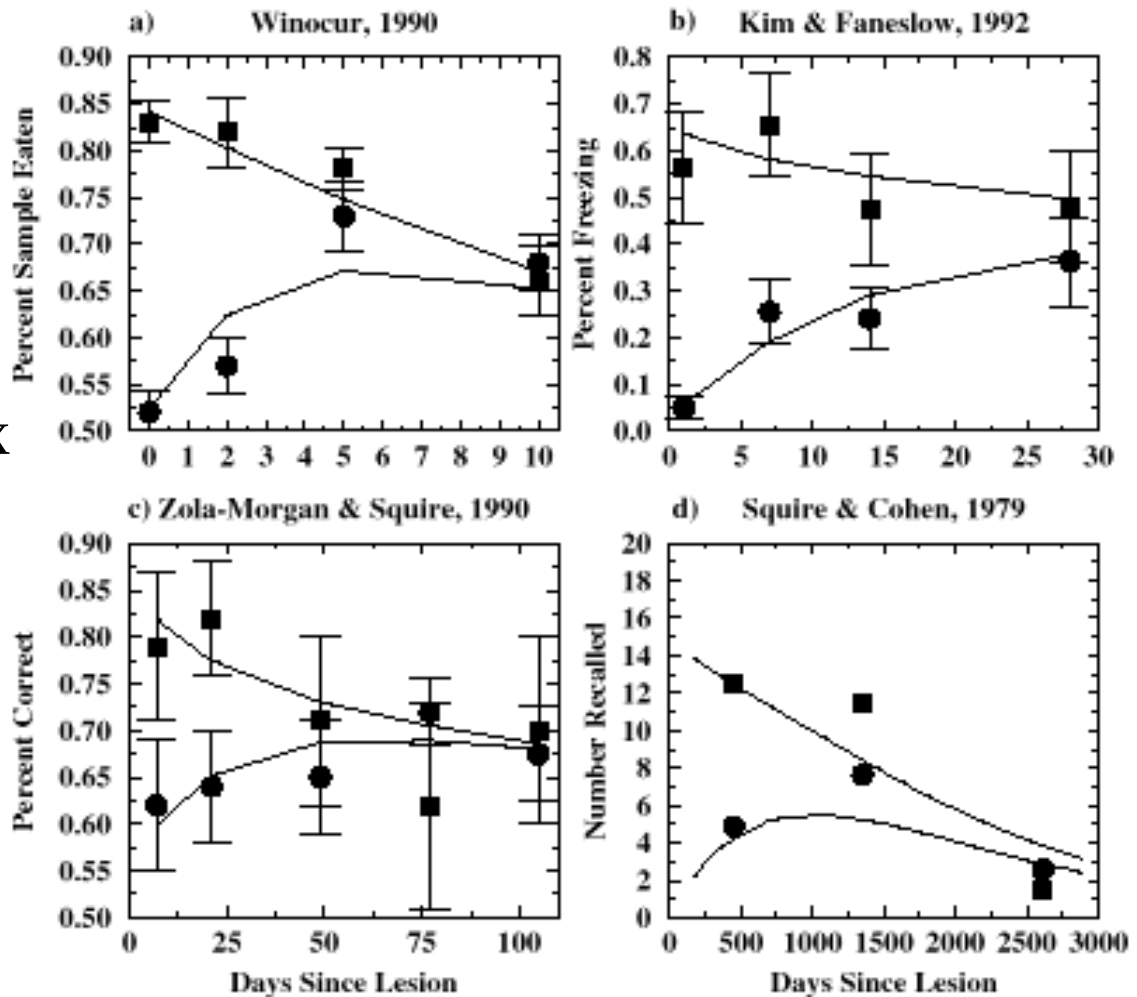
Neocortex seems good to slowly learn the general structure of a domain (e.g., skill learning)

The only thing missing for our story is to show that the hippocampus trains the neocortex with interleaved presentations of the memories it stores

Consolidation

Memories seem to be stored first in the hippocampal system and are then **consolidated**: slowly learned by (integrated into) neocortex

That's why older memories are not lost with hippocampal lesions



Why is this useful?

To allow slow, interleaved learning in the neocortex!

Training the Neocortex

Lots of evidence that during ‘off-line’ periods (e.g., slow-wave sleep or even quiet wakefulness) there is **reactivation** of memories in the hippocampus and neocortex

Examples:

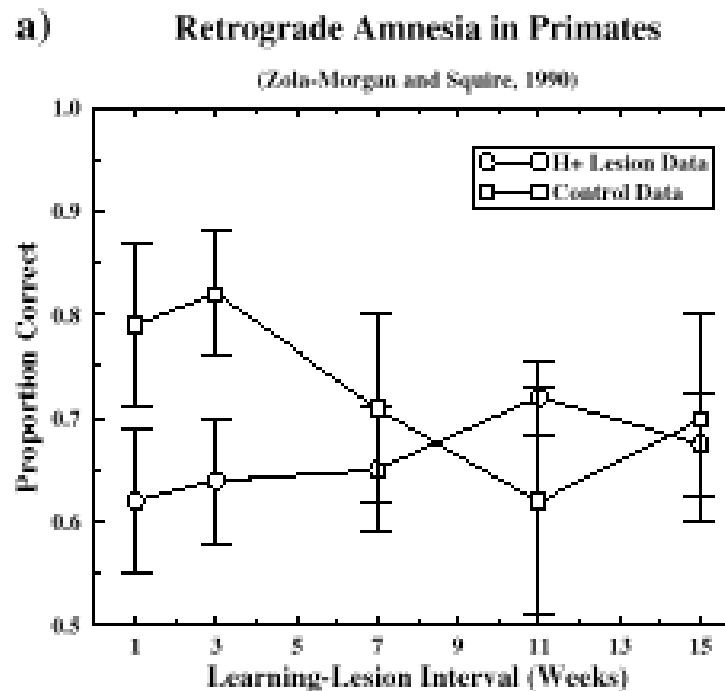
- Neurons that are active during a task tend to be more active in subsequent sleep

- Neurons whose activity is correlated during a task tend to have more correlated activity in subsequent sleep

These could be signs of the hippocampus training the cortex

A Specific Consolidation Experiment

Monkeys were trained on a set of 100 binary discriminations (one object has food, the other one doesn't)



Zola-Morgan and Squire (1990)

Model of the Zola-Morgan and Squire (1990) Experiment

Target = 1 if first object has reward
Target = 0 otherwise

Training the network: 1 day = 1 epoch

On every epoch:

- Background trials

- Learning trials from the experiment

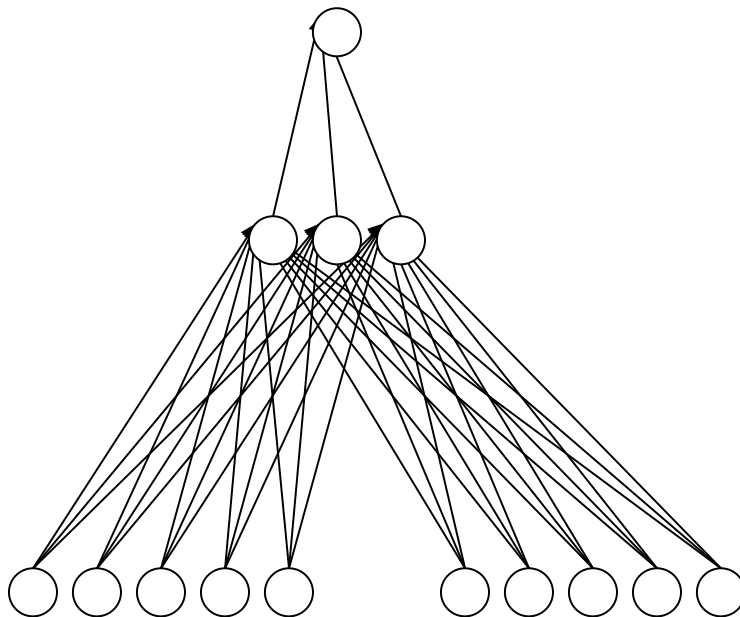
- Reinstated trials from hippocampus

‘Model’ of hippocampus:

- Each experience forms a hippocampal memory trace with strength 1.

- Every day, that strength decays by D .

- The probability of reinstatement is proportional to that strength (multiplied by a ‘reinstatement’ parameter).



Object 1

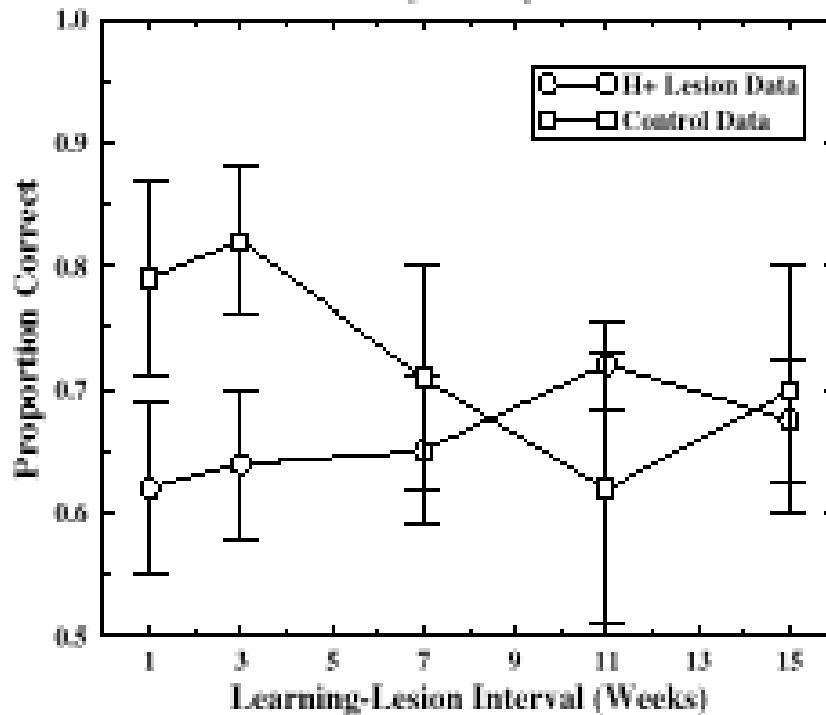
Object 2

Neocortex (hippocampus not modeled explicitly)

Simulation Results

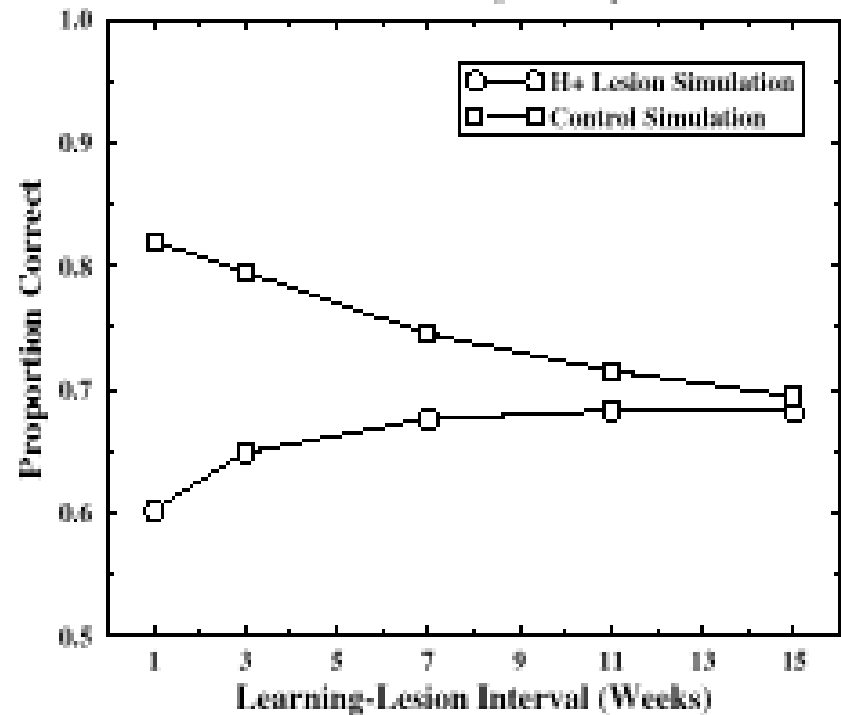
a) Retrograde Amnesia in Primates

(Zola-Morgan and Squire, 1990)

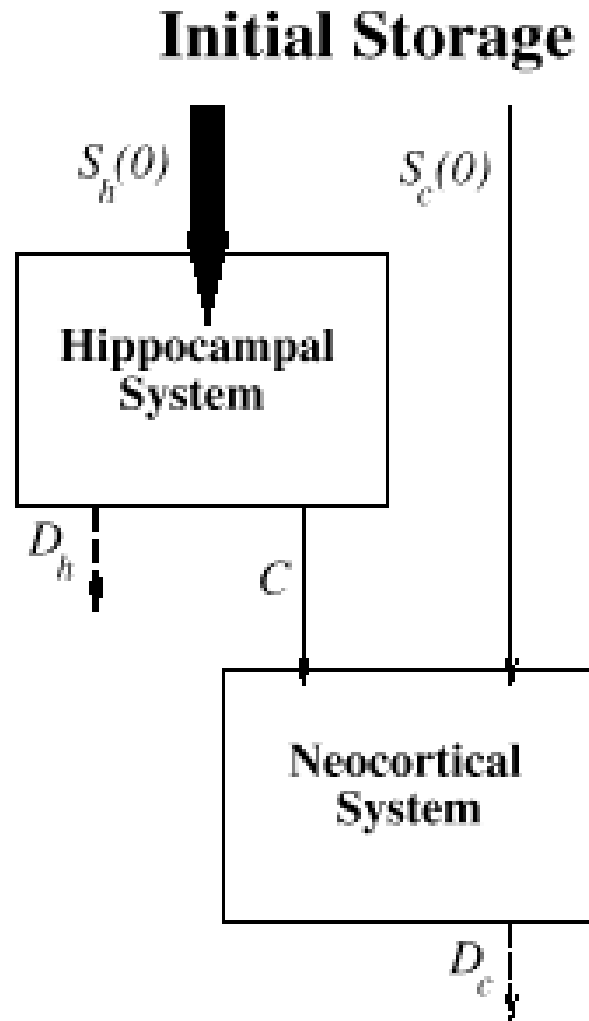


b) Retrograde Amnesia in the Simulation

For data from Zola-Morgan and Squire, 1990



A More Abstract Model



Balance between hippocampal decay, probability of reinstatement of memories, and neocortical learning rate is important

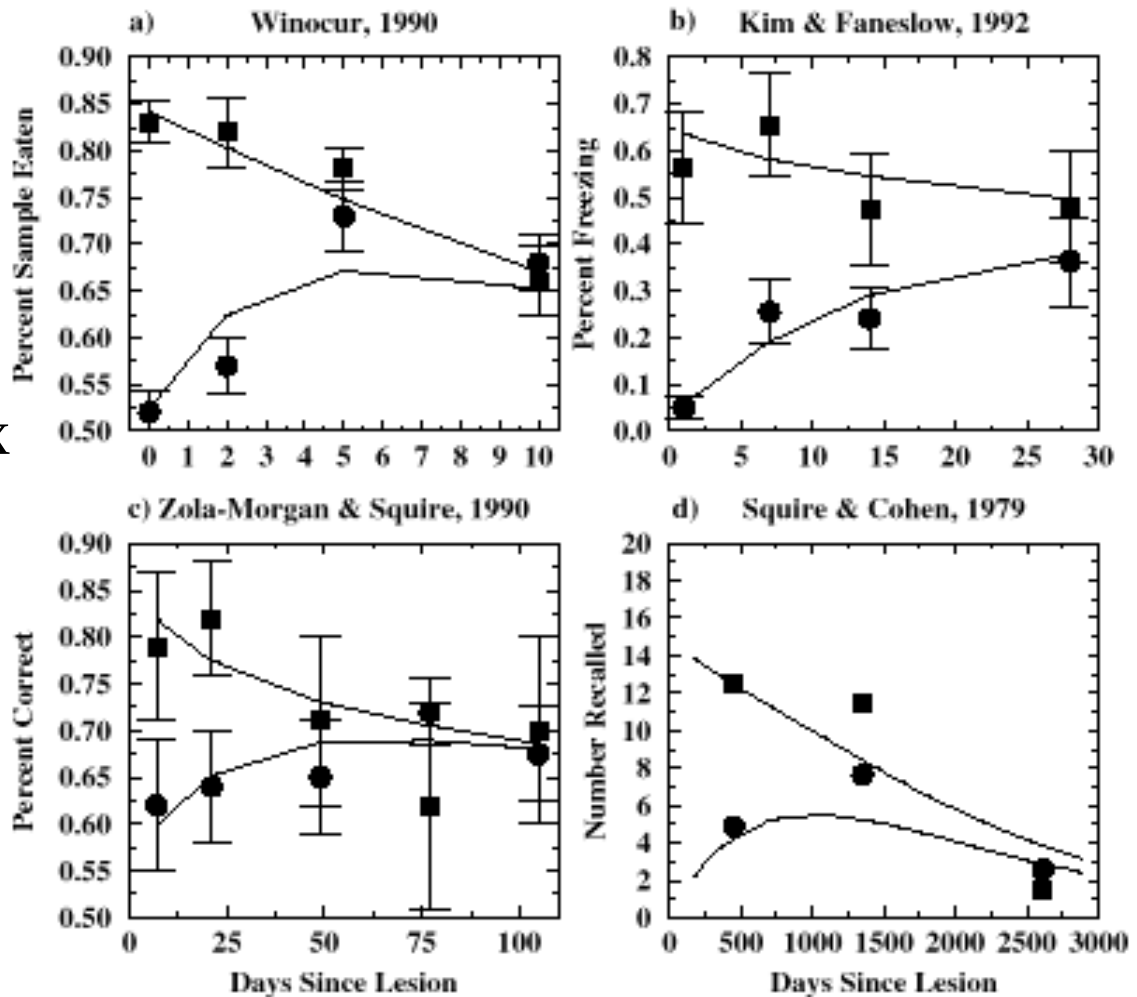
For example, if hippocampal memories decay too fast given the neocortical learning rate, there won't be enough time for consolidation

There is a neat mathematical formulation of all this (see paper)

Consolidation

Memories seem to be stored first in the hippocampal system and are then **consolidated**: slowly learned by (integrated into) neocortex

That's why older memories are not lost with hippocampal lesions



Why is this useful?

To allow slow, interleaved learning in the neocortex!

Summary

Hippocampus: Quickly learn specific information (e.g., your neighbor's dog bit you)	Neocortex: Learn general information (e.g., what dogs are)
Uses	
Fast learning (often one example)	Slow learning (generalize over many examples)
	Interleaved learning (don't want to forget all about cats when you learn about dogs)
Sparse representations (avoid interference)	Distributed representations (allow generalization)

Hippocampus trains neocortex to guarantee interleaved learning

Conclusions

We started with computational considerations:

Fast learning of specific information cannot be done in the same system as gradual learning of general information (catastrophic interference)

It seems that one needs two systems

We found that the brain probably has two such systems (!)

Several people had proposed that there might be consolidation of memory from the hippocampus to the neocortex. But McClelland, McNaughton, and O'Reilly explained **why** the brain is organized this way: it uses complementary systems that work together to solve incompatible computational goals