Memory Models: Overview

COMS 6998 CLMM

January 2026

Memory Mechanisms in Neural Networks

(1). Common mechanisms:

A. Recurrent memory
B. Attention-based memory (standard, efficient adaptations)
C. Hybrids (state-space models)

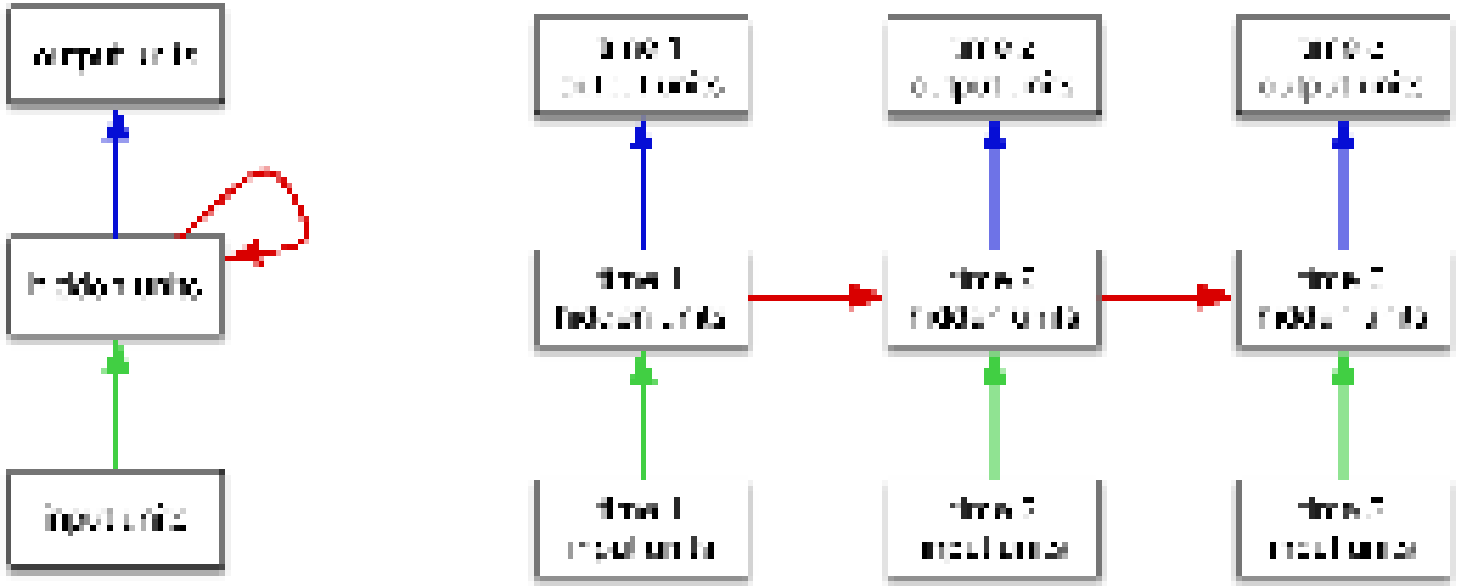
(2). Framework for multiple memory mechanisms: Complementary Learning Systems

(3). Plethora of formulations:

A. Episodic memory (replay, context manipulations)
B. Working memory (internal memory-augmented, meta-learning)
C. External memory (retrieval, differentiable memories)

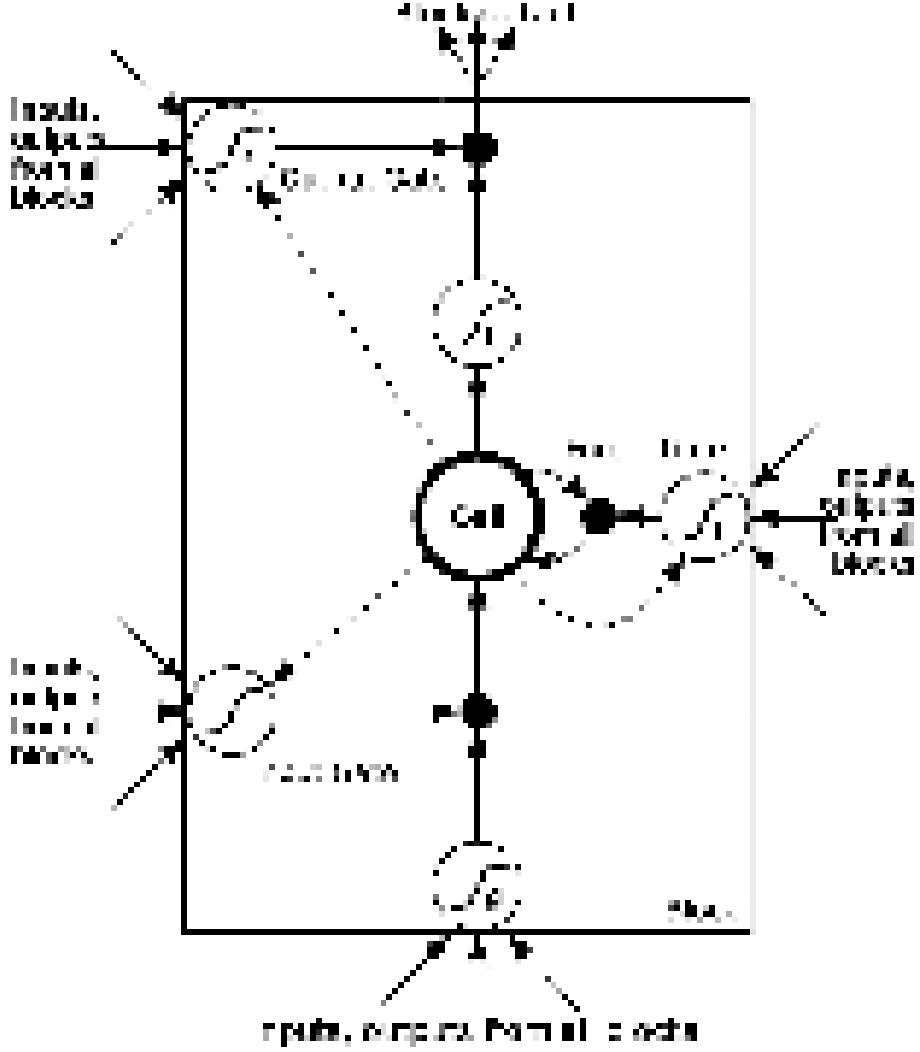
1A. Recurrent memory: RNN

- We can think of an RNN as a dynamical system with one set of hidden units which feed into themselves. The network's graph would then have self-loops.
- We can unroll the RNN's graph by explicitly representing the units at all time steps. The weights and biases are shared between all time steps
  - Except there is typically a separate set of biases for the first time step.



1A. Recurrent memory: LSTM

- Replace each single unit in an RNN by a memory block -



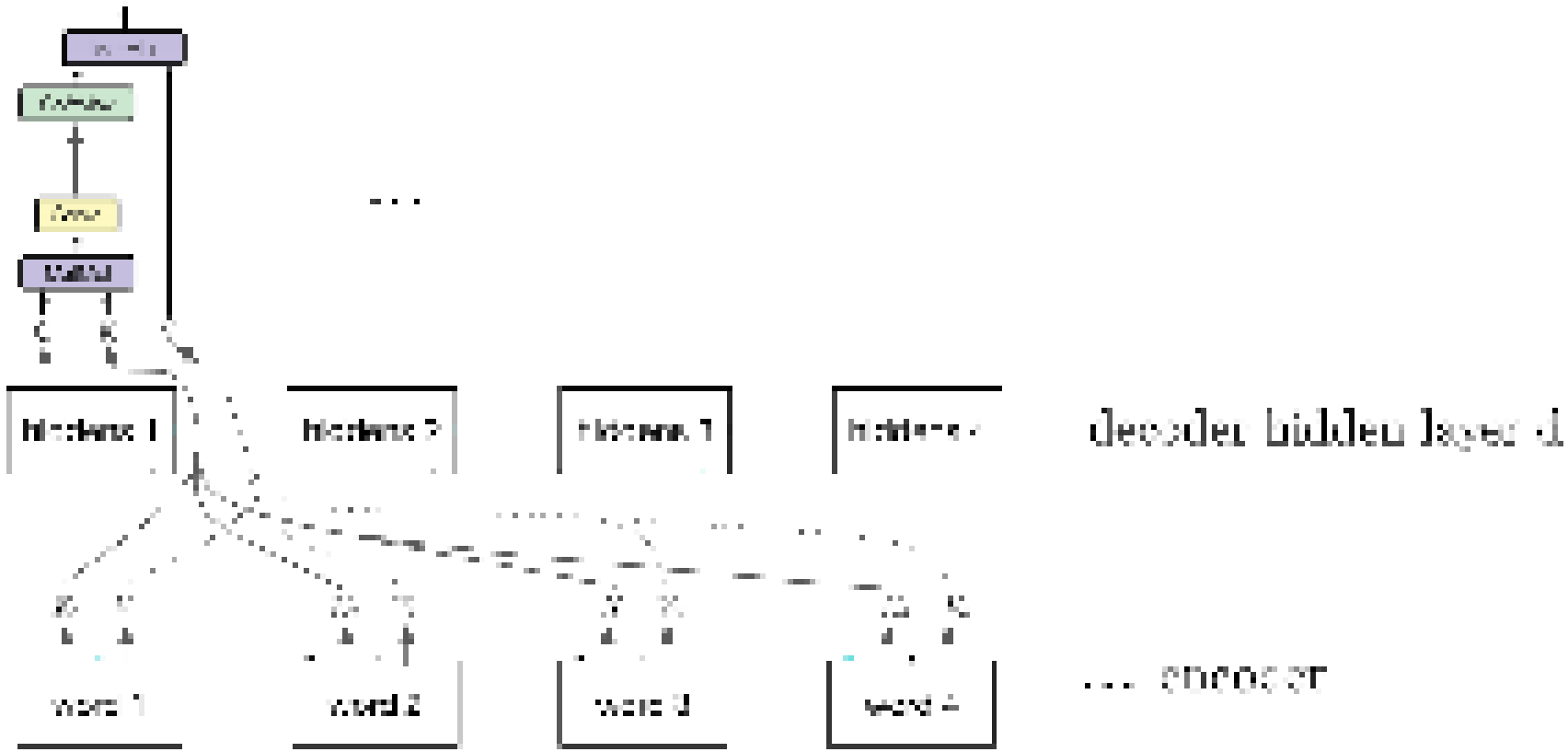
$c_{t+1} = c_t \cdot \text{forget gate} + \text{new input} \cdot \text{input gate}$

- $i = 0, f = 1 \Rightarrow$ remember the previous value
- $i = 1, f = 1 \Rightarrow$ add to the previous value
- $i = 0, f = 0 \Rightarrow$ erase the value
- $i = 1, f = 0 \Rightarrow$ overwrite the value

Setting $i = 0, f = 1$ gives the reasonable "default" behavior of just remembering things.

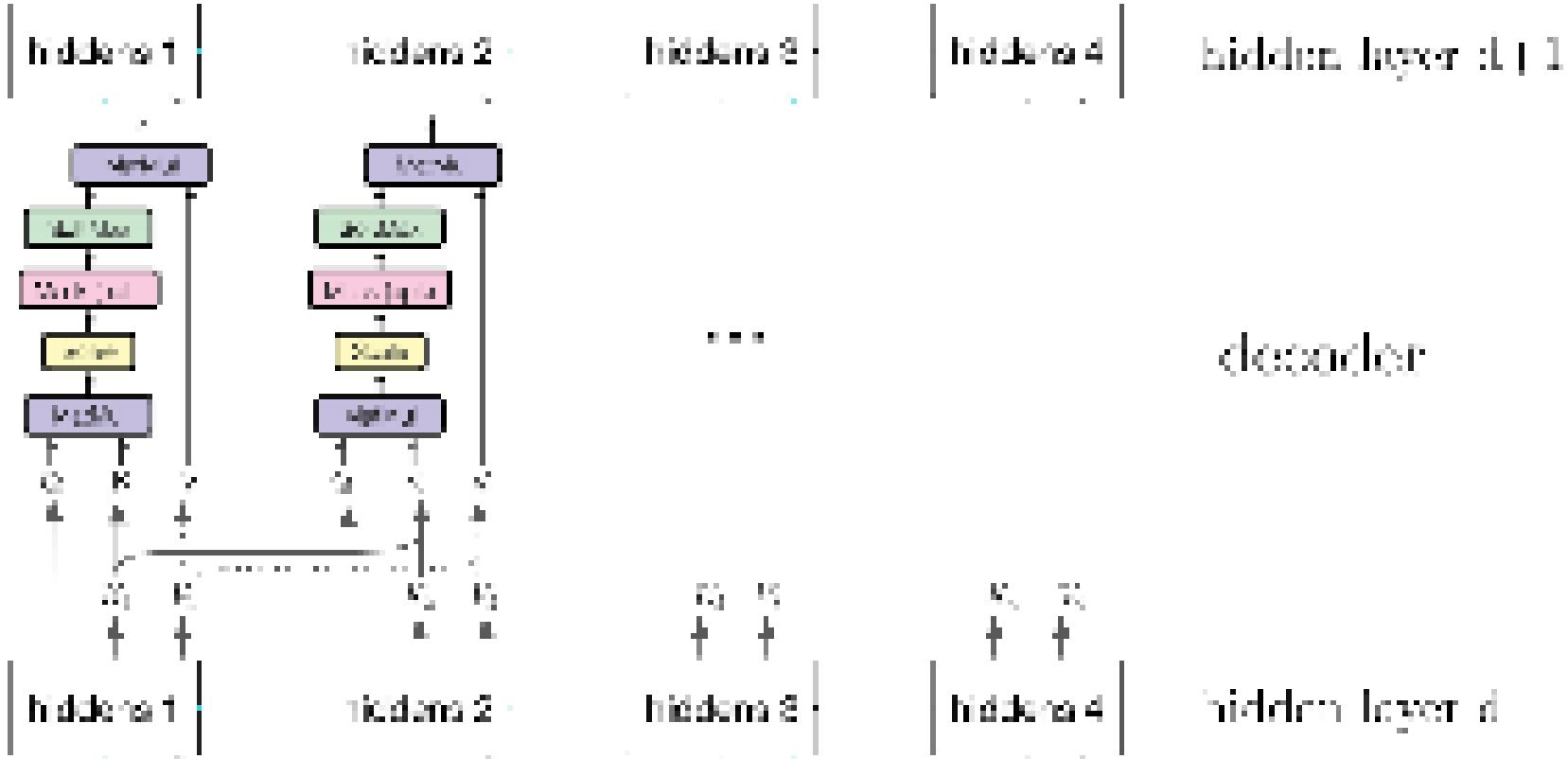
1B. Attention-based memory in transformers

- Transformer models attend to both the encoder annotations and its previous hidden layers.
- When attending to the encoder annotations, the model computes the key-value pairs using linearly transformed encoder outputs.



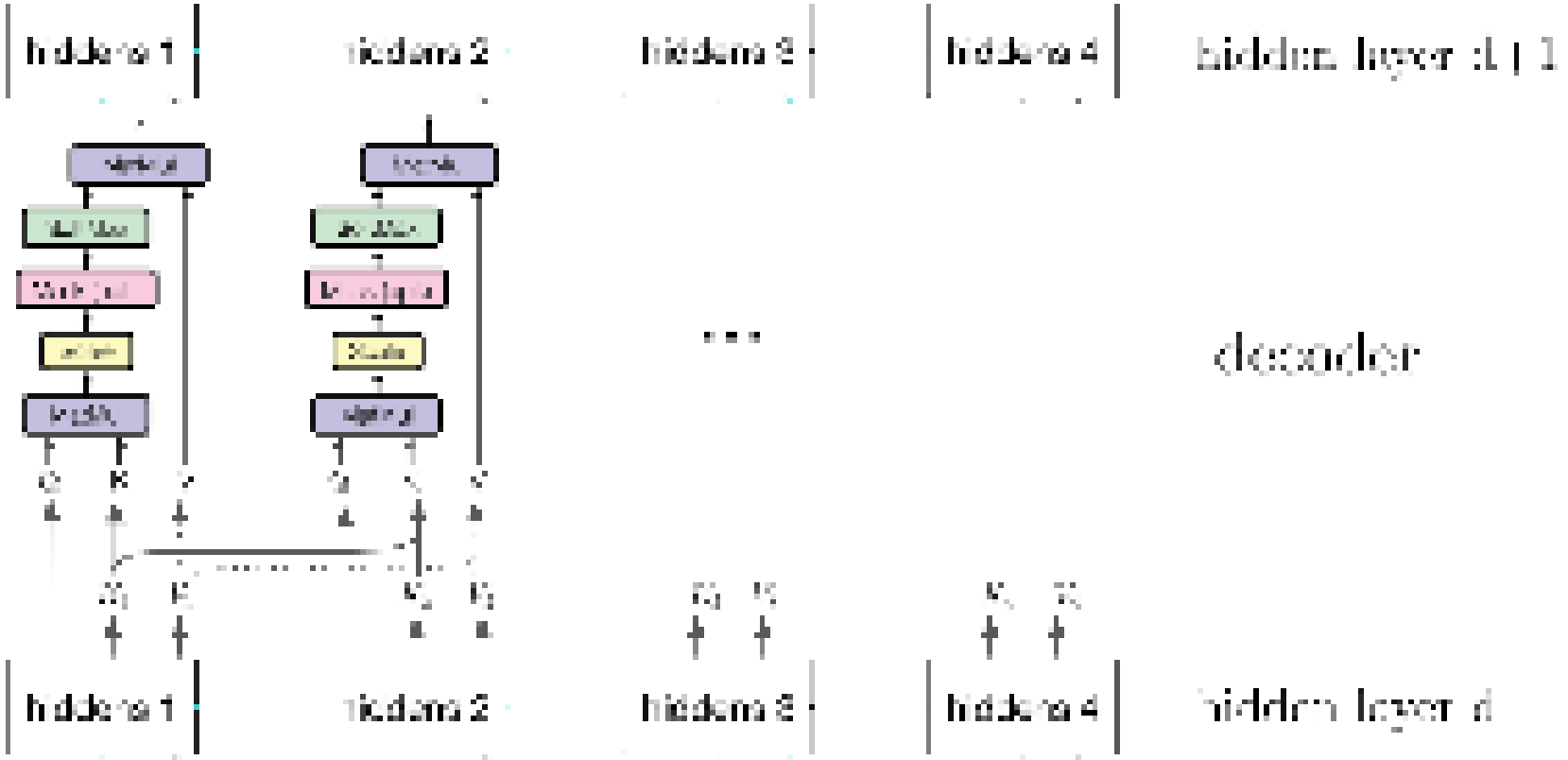
1B. Attention-based memory in transformers

- Transformer models also use "self-attention" on its previous hidden layers.
- When applying attention to the previous hidden layers, the causal structure is preserved.



1B. Attention-based memory in transformers

- Transformer models also use "self-attention" on its previous hidden layers.
- When applying attention to the previous hidden layers, the causal structure is preserved.



Memory Mechanisms in Neural Networks

(1). Common mechanisms:

A. Recurrent memory
B. Attention-based memory (standard, efficient adaptations)
C. Hybrids (state-space models)

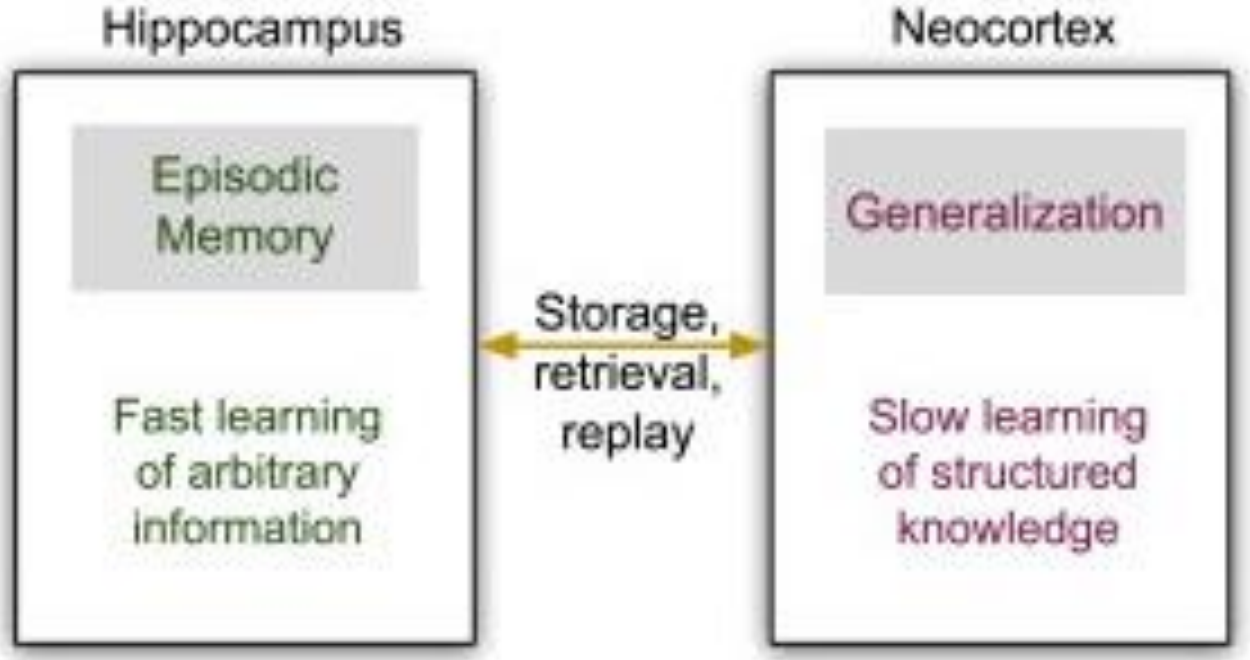
(2). Framework for multiple memory mechanisms: Complementary Learning Systems

(3). Plethora of formulations:

A. Episodic memory (replay, context manipulations)
B. Working memory (internal memory-augmented, meta-learning)
C. External memory (retrieval, differentiable memories)

2. Complementary Learning Systems

Hypothesis: brain uses **two interacting memory systems** with complementary properties to solve the stability-plasticity dilemma [McClelland et al, 1995]



- Fast learning: sparse, episodic, minimize interference, short-term
- Slow learning: distributed, semantic & procedural knowledge, supports generalization

CLS Hypothesis

Memory Consolidation:

1. Initial encoding: New experience → Hippocampus stores it quickly
2. Replay: During sleep/rest, hippocampus "replays" memories to neocortex
3. Gradual transfer: Through repeated replay, neocortex slowly learns the pattern
4. Integration: New knowledge gets integrated with existing structure in neocortex
5. Independence: Eventually, memory can be retrieved without hippocampus

Key mechanism: **Interleaved replay** - hippocampus replays both new and old memories, allowing neocortex to learn new information while rehearsing old information

Sleep:

- Hippocampus spontaneously reactivates recent memories
- Replays "teach" the neocortex gradually
- Interleaving new experiences with old prevents catastrophic forgetting
- Allows neocortex to find common structure across experiences

CLS Hypothesis: Evidence

Neuroscience:

- Patient H.M.: Hippocampal damage → can't form new memories but retains old knowledge
- Replay during sleep: Direct recordings show hippocampal replay of recent experiences
- Systems consolidation: Remote memories become hippocampus-independent over time

Behavioral:

- Sleep-dependent memory consolidation: Sleep improves memory integration
- Spacing effect: Distributed practice beats massed practice
- Interference: New learning disrupts hippocampus more than cortex

Connections to Continual Learning:

- EWC
- Generative Replay

Extended Memory Formulations