MECHANISMS OF HIPPOCAMPAL RELATIONAL BINDING

BY
PATRICK D. K. WATSON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Neuroscience
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral Committee:

Professor Neal J. Cohen, Chair, Director of Research
Professor Howard B. Eichenbaum, Boston University
Professor William F. Brewer
Professor John E. Hummel
Assistant Professor Brian D. Gonsalves

Abstract

This work is about the mental representations that underlie memory for the complex compositions of people, places, things, and events that comprise everyday mnemonic experience, the mechanisms that bind, encode, and reconstruct these representations, and the mathematical frameworks that describe these mechanisms. It approaches this topic with a combination of computation modeling, human neuropsychological empirical research, and scholarly theory building. The critical components are 1) a model and discussion of memory consolidation, 2) reconstructive memory experiments, both in patients with damage to the hippocampus and in college-aged participants, 3) a pair of computational models of relational memory binding, encoding, and reconstruction. These experiments all touch on a larger debate about memory representations that dates back at least to Bartlett (1932), and touches on questions such as: "What different types of memories are there?" "Are these memories more akin to reconstructions or recordings?" "How does time and experience change these representations" and "What are the mechanisms and brain structures responsible for encoding, updating, and reconstructing these representations?"

Part 1: Memory Consolidation

The first critical component gives an overview of memory consolidation research and argues that the literature has two divergent definitions for memory consolidation: a narrow definition focused solely on how memories are protected from amnestic insult, and a broader definition that considers all kinds of representational change over time as examples of memory consolidation. Two consolidation models, one aimed at each of these definitions are presented. The first model explains the narrow definition of memory consolidation for memories of any type (episodic, semantic, procedural etc.) and of any scale (molecular, synaptic, or systems-

level) by demonstrating that in any system that has 1) multiple loci for storing information, and 2) mechanisms that transfer or copy information between these loci, information will tend to move from local to global representations. This means that any amnestic disruption, whether it attacks molecular pathways (e.g., neurotransmitters, protein synthesis), individual neurons or synapses, entire brain structures, or even networks of structures, will become less effective with time as the information it seeks to disrupt becomes more globally represented. Since the notion of multiple interacting memory systems in the brain is well established, and the model can be tuned with very minimal parameters to fit any arbitrary retrograde amnesia gradient there is no need to explain memory consolidation-as-protection-from-disruption with reference to any specific brain process, as it is an obligatory result of simply possessing a brain with multiple, interacting components. The second, broader, consolidation definition however, requires a broader explanation that varies depending upon the type of memory representation in question. The second model presents an example of memory change over time based on previous explorations of memory consolidation and interference in the hippocampus and necortex (McCloskey & Cohen 1989, McClelland, McNaughton & O'Reilly 1995, Ans et al. 2002). Unlike some of these previous works, it concludes that because the neocortical system is only learning the statistical structure of its inputs, hippocampal interleaved learning can only prevent interference in the neocortex to the degree that it creates artificial structure via temporal ordering and repetitive exposure. In a modeling context, it is of course always possible to tune the degree of ordering and repetition to match the observed result, but this does not seem very informative as to the real mechanism involved. The hippocampus cannot know what the neocortex "ought" to remember a priori, before the neocortex attempts to recall the salient fact. Thus the chapter concludes that the hippocampal memory mechanism must be more elegant than a simple copy-theory that exposes the neocortex to repeated instances of previous

experience, because such a system would still require a homunculus that decides precisely how often previous experiences should be repeated.

Part 2: Reconstruction Experiments

The second critical component is comprised of two empirical, cognitive psychology studies of hippocampal relational memory binding using novel reconstruction paradigms. Memory for complex, compositions of items and relations is most often measured using manipulated images or configurations of items (e.g., Ryan et al. 2000, Hannula, Tranel, & Cohen 2006). This allows the item information to remain constant, while the specific composition of items is manipulated. Participants typically demonstrate memory in such experiments by detecting manipulations. Yet while these experiments are effective at detecting disruptions to relational representations, they do not communicate what the change to the underlying representation is that causes it to diverge from the originally studied configuration. To allow participants to report what their mental representation looks like requires a reconstructive memory paradigm (c.f. Bartlett 1932), the results of which are often difficult to quantify. The second component presents two studies that attempt to find a middle ground between open-ended reconstruction and controlled quantifiability in hopes of developing richer relational memory datasets. The first of these experiments involved a simple spatial reconstruction paradigm. Patients with hippocampal damage at the University of Iowa or age and education matched controls studied an array of 2-5 everyday objects placed at random locations on a 1x1m table and then tried to reposition the objects in their studied locations after a brief (4s) eyes-closed delay. Previous experiments measured performance in this task exclusively with an item misplacement measure (how many cm away from their studied locations items were placed at reconstruction c.f. Smith & Milner 1981), however we found that swapping the

locations of pairs of items was far more indicative of hippocampal damage, with patients making numerous errors of this type while controls made only a single such error in the entire course of testing. Patients made swaps error even on two object arrays, and the prevalence of swapping could not be explained simply by poor performance on other metrics, though it contributed heavily to poor performance on all of the other metrics we examined. These findings suggested that the primary deficit in patients with hippocampal damage in a spatial reconstruction task was an inability to correctly bind individual item identities to their relative locations, and not a more general failure of spatial or declarative memory. The fact that these deficits are observable even at short lags traditionally associated with working memory and even with item sets as small as two additionally argues that hippocampal damage is not simply a disruption of transfer from working to long term memory. Finally, while the rate of swapping increased as the number of items increased, it did not increase faster than the number of pairwise relations present in the stimuli, suggesting that this error is directly tied to the relations between elements.

The second experiment of the second component was designed to more thoroughly explore these swap errors. Building on the first spatial reconstruction experiment, this second experiment required college-aged participants to reconstruct a short movie composed of a set of six face-background pairings, placing each face and background in their studied location, and in the correct temporal sequence. Unlike the previous experiment where participants could position objects at any location within a 1m square, thereby producing different kinds of spatial errors, this "event reconstruction" paradigm had a finite and clearly delineated number of slots for each element to be bound to, allowing only for swap errors, and making possible robust similarity analysis to determine how many adjustments to the participants' reconstructed configuration would be required to convert the reconstruction to the originally studied

configuration. In examining participants reconstruction accuracy (that is, how many elements of their reconstructed configuration were correctly bound), our central finding was that performance was tightly linked to relational complexity (i.e., bindings between sets with large numbers of elements were more difficult than those between sets with small numbers of elements), and arbitrariness (i.e., patterns of binding that were consistent across trials produced better performance than inconsistent patterns), and that both of these effects were closely related to the degree to which participants' performance could be predicted from their prior patterns of reconstructions (i.e., reconstruction "semantics"). However, once these two factors were controlled for we did not find a strong effect of the type of binding (e.g., spatial-spatial, v. item-item). Additionally, using similarity analysis we were able to demonstrate that while participants' reconstructions were dramatically better than random performance, they were only slightly more similar to the studied configuration than they were to reconstructions created before the participant saw the studied configuration. We were additionally able to demonstrate that the general "semantic" tendencies of the participant enabled them to encode approximately 12 bindings more than would be expected by chance, while the "episodic" information encoded on each trial amounted to approximately 3 additional bindings. We also demonstrated that this additional information was not simply present in the accuracy of the initial configuration but that the "errors" participants made were non-random, and contained informational structure similar to that of the original studied configuration. This study highlighted the dynamic and synergistic way in which new and previously learned information interact to provide a useful set of constraints capable of (re)creating a complex configuration of items, locations, and times, and reaffirmed the importance of examining not just how memory drives correct performance, but also how memory contributes to non-random errors.

Part 3: Computational Models of Hippocampal Function

The first of two computational models presented is the Memory and Reasoning (M&R) computational model produced in collaboration with investigators at Sandia National Laboratories. This system was meant to simulate the processing of visual sensory streams and the hippocampus. It was composed of two cortical components, both of which were composed of hierarchically arranged adaptive resonance theory (ART) networks that used an unsupervised learning algorithm to capture the statistical structure of complex visual inputs. One of these components acted on high-resolution pictures of faces (meant to be analogous to objects in the fovea), while the other acted on low-resolution pictures of scenes (meant to be analogous to the lower resolution visual surround). Given a face-scene pairing, the first component simulated the processing of the ventral stream, parsing the complex images into simple visual features, then higher-order structural features, and finally into objects corresponding to the individual faces. The second component simulated the processing of the dorsal visual stream, parsing its inputs into spatial features (e.g., "objects of any kind on the left"). Both of these streams were equipped with "recall" capabilities, if a single unit corresponding to a particular input or input category was activated, the recurrent connections within the ART network that represented that category would produce the "prototypical" input that would elicit the activation of that category. This input would in turn activate ART networks lower down in the hierarchy in the same fashion until the network would print out at the sensory camera (its "mind's eye") a visual configuration that corresponded to the originally activated component's input category. In this way the system could be a "pictures in/pictures out" searchable database for visual images. Augmenting this function was the hippocampal component that bound together information from both the "dorsal" and "ventral" components. It did this by passing the high-level activation of both components into a high-dimensional space (meant to simulate the dentate gyrus) to obtain a unique "pattern separated" key corresponding to the particular input conjunction. It

then washed this key through a heavily locally recurrent component (meant to simulate CA3) to fill in any gaps, and then mapped this output back to the two cortical components via a third hippocampal component (meant to simulate CA1). In this way faces could be "shown" to the model to elicit the "scene" they were studied with and vice versa, allowing the model to complete source memory style tasks. By tuning the time at which CA1 performed its reactivation of the cortical components it was also possible to recover additional face-scene pairings that were studied in close together in time to the original cue, and even entire sequences of studied face-scene pairings. This model was therefore able to recreate much of the performance, and even subjective experience, of relational memory tasks, but it lacked much of the flexibility of relational memory. It could not create face-face bindings, or perform transitive inference, or create novel bindings. In essence, the hippocampal component was performing the same type of category learning as the two cortical components, but it was learning highly specific, cross-domain categories.

The limits of the M&R model motivated the final model of the document the relational memory binding, encoding, and reactivation (RMBER) model. Produced in collaboration with the FRAMES team of the IARPA ICArUS project, this model was meant to perform flexible relational binding of complex compositions of stimuli. Structurally, it closely resembled the M&R model, with a cortical-inspired input/output region (the entorhinal cortex, EC layers 2 and 5 respectively), a dentate gyrus (DG) with a large number of units relative to its inputs, a highly recurrent CA3 region, and a CA1 region that performed mappings between CA3 and the EC. However, unlike the previous model the RMBER model used Mihalas-Niebur spiking neurons (Mihalas & Niebur 2009), to model the actual oscillatory dynamics of neurons. These neurons used spike-timing dependent plasticity that tuned the strength of neural connections but also the degree to which neurons were coupled with inhibitory interneurons. In this way, input

could modulate both the rate that a unit fires, and the delay at which it fires in response to the input. In addition, all units in the model were subject to an extrinsically generated theta rhythm, and a locally generated gamma rhythm. This allowed the model to do more than simply generate conjunctions corresponding to previous inputs. First, the model treats dynamics of the entire entorhinal cortex as a superposition of the activity of the entorhinal cortical units. It maps these dynamics into the high dimensional space of the dentate gyrus where complex dynamics of the entorhinal cortex are decomposed across a large number of cells, resulting in dentate cells which respond only to particular sub-frequencies of the entorhinal dynamics, and only when presented at certain phase delays (relative to the beginning of a theta or gamma cycle). These dentate dynamics are collapsed back into the CA3 region, that by summing across large numbers of dentate cells' activity and by reconstructing their signals within its highly-recurrent local network, recreates the superposition of single unit dynamics present in the EC2. These dynamics are mapped back to EC5 via CA1. Processing within each sub-region requires one gamma cycle ensuring that the output from CA1 arrives back at EC5 at the beginning of the next theta cycle, thus providing the hippocampal network's predicted dynamics for the next theta cycle. This process shares much in common with the discrete Fourier transform, decomposing a complex signal into its phase and frequency subcomponents, discarding the higher order frequencies and recomposing the original dynamics from a compressed code relying upon the stored coefficients. However, while the hippocampal model initially fills in gaps in the signal with Gaussian random noise, it learns to fill in gaps with previously observed phase and frequency sub-components. Since these sub-components are derived from the observed activity in the EC they reflect real relationships present in the activity of EC neurons. Since the model stores both phase and frequency of previous dynamics, it creates a truly compositional, concatenative code that reproduces the appropriate level of activity at the appropriate EC units

in the same order as previously observed. Thus, unlike the rigid configural categories of the previous M&R model, the RMBER model can flexibly add and remove sub-components from the entorhinal dynamics while maintaining the correct relative order of activations. We show that the model is capable of learning simple rules, and complex patterns of geometric relations such as path integration. We generate a "virtual rat" and have it run in a virtual circular enclosure according to a random path, and then allow the hippocampal model to reconstruct this path from a partial cue. We also demonstrate that the model develops many of the same cell types observed in single-cell recording studies (e.g., "place" and "time" cells). This model provides a novel way of understanding the hippocampus's relational binding function by relating its intrinsic dynamics to the discrete Fourier transform.

Together, these papers outline the need for a specialized memory system devoted to binding compositions of independent elements, experimental evidence for the existence of such a system, and computational mechanisms by which such a system might act.

Acknowledgements

This work would not have been possible with the support of a tremendous number of collaborators, friends, and supporters. My advisor Neal Cohen, and my defense committee, Howard Eichenbaum, Bill Brewer, Brian Gonsalves, and John Hummel have all provided important guidance over the course of the research outlined here. My other faculty mentors, both inside and outside the neuroscience program, Tom Anastasio, Joel Voss, Gene Robinson, Sam Beshers, and Steve Levinson, have also been instrumental to my research and my development as a scientist. In addition, this research could not have been conducted without my collaborators in Albuquerque (Michael Bernard, Tom Caudell, Shawn Taylor, Steve Verzi, and Craig Vinyard), Iowa (Dave Warren and Dan Tranel), the ICARUS project (Kenny Sharma, David Rosenbluth, and David Morgenthaler), and all of my lab mates in the memory systems laboratory at the University of Illinois. I have had the good fortune to be supported by the Memory Analogies fellowship, Sandia National Laboratories, the University of Illinois, and NIMH. All of these people and institutions, and others, were vital to my success, helping me with the what, where, when, and how of research in ways too numerous to mention, but finally, I'd like thank my son Asher, who more than any other on the list, provided me with the why.

TABLE OF CONTENTS

CONSOLIDATING CONSOLIDATION: HOW REPRESENTATIONAL CHANGE CAUSES RESISTANCE TO DISRUPTION

Contributors
Watson, P., Anastasio, T., Cohen, N.

This paper was prepared in response to data produced during the 2006-2007 Memory Analogies Project at the University of Illinois at Urbana-Champaign

Some of the concepts appear in Individual and Collective Memory Consolidation (Anastasio et al. 2013), but all of the modeling data is original to this work

Abstract

We present two models of memory consolidation. The first explains why memories become more resistant to disruption over time: memory involves multiple, dynamic, interacting processors. New information is initially represented in those processors best able to capture it, but learning and offline processing cause representations to become more distributed and global over time. Since amnestic disruptions target only a specific sub-set of these systems, and cannot "follow" the information as representations move from local-to-global, their effectiveness will decrease with time. This simple "diffusion" model of consolidation can explain any retrograde amnesia gradient or consolidation time-course at any level of analysis (i.e., synaptic, systems-level).

However, specific memory systems' mechanisms produce changes other than durability. We present a second consolidation model focused on a pair of complementary systems, the neocortex and the hippocampus (c.f.McClelland, McNaughton, & O'Reilly, 1995). We tested this model with different input corpora and training regimes. We found that while the neocortical perceptron-based model was capable of learning associative rules and arbitrary pairings, its performance on arbitrary pairings was entirely due to the artificial, repetitive, rule-based structure introduced by the interleaved learning paradigm. Further, performance on arbitrary pairings was far worse, required far more training, and was extremely vulnerable to interference. We conclude that this is due to the nature of the neocortical models learning algorithm, it is well suited toward discovering the statistical structure of its inputs, but arbitrary pairings have no such structure to discover unless it is artificially imposed via interleaved learning.

Introduction

Memory consolidation has two competing definitions. The first is over a century old: "a time and experience dependent process by which newly-acquired memories become stabilized so that they are relatively immune from disruption by behavioral interference and by amnestic agents such as lesions or drugs (c.f. Muller & Pilzecker, 1900; McGaugh, 1966; Lechner et al., 1999; Nadel, Samsonovich, Ryan, & Moscovitch, 2000)." The second is more recent: "Memory

consolidation is any change in a memory representation that takes place between encoding and recall," (Squire, Cohen, & Nadel, 1984; Winocur, Moscovitch, & Bontempi, 2010; McKenzie & Eichenbaum, 2012). There has been debate over which of these is the best framework for study (Nadel, Winocur, Ryan, & Moscovitch, 2007), yet these definitions are entirely orthogonal, refer to different data sets, and make non-comparable predictions.

The former definition is a *description* of a phenomenon: over time, memories become resistant to amnestic insult. It *describes* the temporal gradient in retrograde amnesia, but does not *explain* it. As an explanation, it would beg the question: "why do memories become more durable?" "There's a process which makes them more durable." The second definition however is an *explanation*: 1) there are different types of memory 2) over time, memories change type, 3) each amnestic intervention only disrupts one type of memory, thus 4) over time amnestic interventions become less effective. This explanation concerns itself with transforms between different types of memory representation; "durability" is a side effect of those transformations.

There are many numerous current research threads on memory consolidation. Plasticity studies have elucidated the mechanics of memory consolidation at the synapse (Shimizu, Tang, Rampon, & Tsien, 2000; Bramham & Messaoudi, 2005; Izquierdo et al., 2006). Reconsolidation (Misanin, Miller, & Lewis, 1968; Nader, Schafe, & Le Doux, 2000; Suzuki et al., 2004) experiments demonstrate that the process of memory consolidation is not a simple, unidirectional, process of solidification. Studies of consolidation during sleep have drawn attention to the effect of offline reactivation and rehearsal (Wilson & McNaughton, 1994; Stickgold, 1998; Buzsáki, 1998; Paller & Voss, 2004; Walker & Stickgold, 2004). Animal studies demonstrated that mental representations could affect the time-course of consolidation (Tse et al., 2007). But it is often unclear whether these different approaches attempt to better describe memory solidification or to provide an explanation for different types of memory transformation.

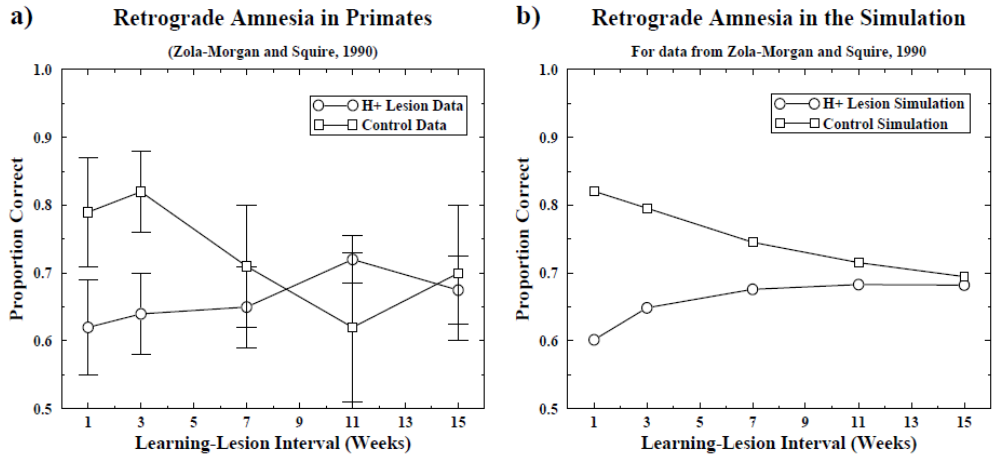
This work presents two models of memory consolidation. The first captures the "memory solidification" data and argues that memory solidification is a necessary consequence of changing mental representations-even if those representations are not converted to a more durable format. The second model highlights a particular kind of transformative consolidation: the shift from sensorial episodic representations containing information about the specific

bindings of objects, places, and sequences, to context-independent semantic representations (Nadel et al., 2000). We explain this shift via a model adapted from the complementary learning systems framework (McClelland et al., 1995; Norman & O'Reilly, 2003), that demonstrates a trade-off between representing complex, arbitrary relational information, and rule-like semantic relations. Based on this result we argue that the shift from episodic to semantic memory is due to episodic and semantic memories being different kinds of information represented in different brain systems, and that semantic memories take more time to construct not due to "consolidation" per se, but simply because semantic memories require multiple examples.

Retrograde amnesia gradients-a consequence of multiple memory systems.

The first definition of consolidation: "a time and experience dependent process by which newly-acquired memories become stabilized so that they are relatively immune from disruption by behavioral interference and by amnestic agents such as lesions or drugs," invokes a *quantitative* framework for doing memory research. In fact, it prescribes a specific experimental technique: 1) An experimental subject will learn some corpus of information either via training or incidentally. 2) At some subsequent time, the experimental subject receives an amnestic insult either via intervention or accident. 3) The learned corpus of information is tested. 4) A comparison sample is gathered from non-amnesic subjects and 5) the information retained by the amnesic subjects is plotted relative to the information of comparisons undergoing normal forgetting. The independent variable is the nature of the amnestic intervention, and the dependent variable is the time it takes for the two curves to meet. A prototypical plot of this relationship (Squire, 1992; Squire & Alvarez, 1995; McClellan et al., 1995) is below:

Figure 1.1



The shape of these curves was described more than a century ago. The control's memory curve is characterized by Jost's law of forgetting (Jost, 1897): the absolute amount remembered decreases with time, and also, that the rate of decrease decreases. The amnesic curve is characterized by Ribot's law of retrograde amnesia (Ribot, 1881), which stipulates that in retrograde amnesia recent memories are preferentially degraded; memories become more stable over time. Mathematically, this function is also in the exponential family, and resembles the inverse Gaussian distribution which has a positive first derivative, causing it to initially slopes up from zero, but has a negative second derivative, which causes it to eventually saturate at some non-zero asymptote and decay back toward zero.

There have been more than a century of consolidation experiments within this paradigm, and the take-home finding has been that there is no single time-course of consolidation. Rather, the time course can be affected by species, amenstic treatment, memory corpus, modulatory compounds, and the level of analysis-synaptic vs. systems level (for review c.f., McGaugh, 2000)

Thus memory consolidation is not only "a time and experience dependent process," it seems to involve many processes and have many dependencies. Muller & Pilzecker's original hypothesis that the time course of consolidation was a knowable constant has not been borne out. It is possible to say that consolidation takes two hours, but only if we additionally stipulate that this is true only for the finger opposition-sequence learning task disrupted by an interference

training sequence, and only in humans, and only if the humans do not nap between the two training sequences (Korman et al., 2007). Time is an important prerequisite for memory durability, the brain is a physical system that cannot instantly change, but time is hardly the only or primary factor.

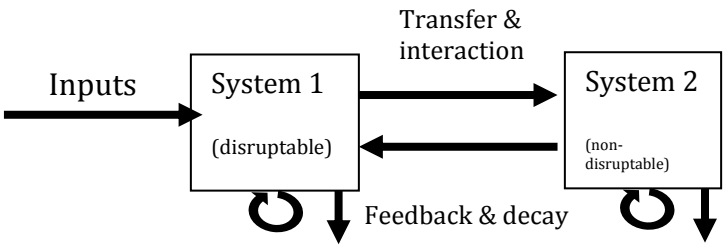
In addition to the absence of a constant time course, there is little evidence for a simple "disruptable/non-disruptable" distinction. A one-trial avoidance memory in chicks was resistant to a low dose of flurothyl after a few minutes, but not resistant to a high dose even after hours (Cherkin, 1969). There simply seems to be no simple relationship between the variables of "time" and "durability."

Yet, memories *do* become resistant to amnestic interventions over time, and this phenomenon requires an explanation. Thus we turn to the latter definition: memory change. Since at least Hebb (1949), there has been broad acceptance of multiple types of memory, implemented by different brain processes, with some mechanisms for transforming memories of one type into another. In the next section we argue that such a framework predicts a time dependant memory solidification process.

Multiple stores confer durability.

Suppose we take the minimal case: two memory systems. This could be a working memory/long term memory distinction, a reverberating circuit/long term potentiation distinction, or a hippocampus/neocortex distinction; the details of the memory system are unimportant—it could even be a pair of buckets storing water—all that is required is two stores and a means of transferring information between them (Figure 1.2).

Figure 1.2



A multi-system memory: Information is initially input to the first system. Over time, information is exchanged between the systems via reciprocal connections which mediate information transfer and interaction. Each system may also have some internal feedback processing or decay via forgetting. Each of these flows may be governed by a complex function, but from the point of view of a retrograde amnesia experiment, it is usually sufficient to consider each as a single parameter corresponding to the first derivative of information with respect to time.

When information is introduced to the first store, it slowly begins to transfer to the second store. This does not require that information be "copied" or "duplicated," though that could be the case, it is sufficient for the information to flow until it reaches an equilibrium state. In the context of memory, this flow could be the transfer from working to long-term memory, or, just as easily, the transfer of information from a neural spike train to NMDA receptors.

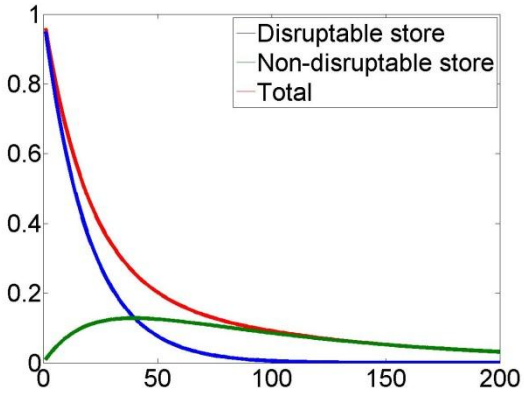
An amnestic intervention is simply removing all of the information from the first store, either by "draining" the information already there (via interference or a transient memory disruption which produces only retrograde amnesia), or by destroying the store entirely (which will produce anterograde amnesia in addition to retrograde since that store is no longer able to hold information at all). This amnestic treatment only preserves the information that has already been transferred to the second store.

Such a treatment inevitably produces temporally graded retrograde amnesia. Longer wait time will produce smaller losses up to the point where the system reaches an equilibrium state, after which any intervention will produce the smallest possible disruption given the systems' dynamics. Critically, the information has not become more durable because the second store's representations are more intrinsically durable; both stores are identical in this respect. Instead the information has become more durable because it is more distributed, and thus is partially immune to the localized amnestic disruption of the first store. Just as a drop of ink becomes increasingly hard to remove from a glass of water over time, information becomes increasingly hard to remove from the brain as it is processed by more memory systems.

This model is perfectly capable of accommodating different consolidation time courses. Instead of assuming that the two stores are identical and that at equilibrium each contains precisely half of the initial information, we need only tune the relative leakiness of the components. If the second system leaks more slowly than the first, then, in the limit, disruption of the first system

would produce no loss of memory whatsoever and we might say that this information has been "fully consolidated" into the second memory store (Figure 1.3). If the first system leaks more slowly, then we would expect that amnestic interventions ought to disrupt memory whatever the wait time. Thus some portions of memory are always dependent upon certain components (assuming our measures are sufficiently sensitive to detect precisely how much memory is disrupted). Indeed, some theories such as multiple trace theory (Nadel et al., 2000) make just such a prediction, in this case, that the hippocampus is always required for episodic memories, and as evidence points to a body of studies that found flat retrograde amnesia gradients for episodic memories (Moscovitch et al., 2005). But if we allow each of our memory components to vary in how quickly they integrate information (in other words, vary their leakiness), they can produce any pattern of consolidation.

Figure 1.3



Multisystem consolidation: Memory moves from the disruptable store to the non-disruptable store over time. The sum of the two components corresponds to intact performance and the non-disruptable alone to amnestic performance.

Further, while two components are sufficient to predict any consolidation time course (with one parameter to fit the dynamics of the disrupted memory store, and one to fit the non-disrupted store), the brain certainly contains more than two components, memory is distributed across dozens or hundreds of anatomical divisions, thousands of biochemical pathways, and billions or trillions of synapses-each with different dynamics. However, any amnestic intervention disrupts only a sub-set of these components, and performance is always measured in only two categories. Thus the experimental method itself it will necessarily divide memory into

"disrupted" and "spared," and the consolidation time-course will always be best approximated with a two-box model.

Because the particulars systems disrupted by the intervention are unknown, consolidation will have complex dynamics, individuals will vary, previously consolidated information will affect consolidation's time course, and information that was independent of a particular sub-system may become dependent upon it again, meaning that the vulnerability of a particular type of memory to a particular amnestic disruption varies with time and treatment-as observed in the reconsolidation literature. Yet if we had never observed time dependant memory consolidation or a retrograde amnesia gradient we could have predicted all of the results simply from the existence of multiple memory systems. We could even have predicted the variability of consolidation's time course from the observation that different memory systems have different properties. In principle we could predict any arbitrary consolidation time course with simple calculus if we knew all of the trillions of parameters involved (or indeed, the two parameters corresponding to the disrupted and undisrupted stores), though this prediction would have questionable practical value. "Consolidation" is an epiphenomenon produced by disrupting the normal flow of information through multiple memory systems.

However, this model tells us nothing about the actual properties of the memory systems in question because it treats memory as a unitary construct which can be "preserved" or "disrupted." It explains why memory becomes resistant to disruption by pointing out that memory is dynamic while disruptions are static. Yet this motivates a far more interesting question: "What are the dynamics of memory?" This is the heart of the second definition of memory consolidation: "any change in memory between encoding and recall," it requires research on the different mechanisms of memory systems, regardless of any change in "disruptability."[1] "Disruption" is such a grand abstraction it can be answered for all memory

[1] There is a philosophical debate to be had about whether the term "consolidation" is valuable in a framework of memory change. Although he surely did not intend it, Eric Kandel concluded his keynote address at the 2010 Society for Neuroscience with a succinct eliminativist message: "memory takes time." He did not say "consolidation takes time" or even "memory consolidates," because he did not need to. If one understands the biological and information processing that forms and changes memory, there is no additional "consolidation process" left to study.

systems in the same way, but there is no universal answer to the memory change question; it is different for each memory system. In the next section we present a discuss a particular example of memory change, the tradeoff between hippocampal and neocortical memory, and present a model similar to that of the complementary learning systems framework, (McClelland et al., 1995; Norman & O'Reilly, 2003) that helps explain the representational change between these two structures.

Multiple episodes, pattern separation, or arbitrary relations?

The hippocampus and neocortex are complementary memory systems (McClelland et al., 1995), that each provide the other with increased functionality. There is relatively broad agreement on the generalities on the role of the neocortical system: it acts as a powerful associative memory (Hebb 1949; Rosenblatt, 1961 ; Rumelhart & Ortony, 1976; McLelland, & Rumelhart 1986), building a web of semantic knowledge based on examples using some sort of multivariate statistical algorithm related to Hebbian learning. There is less consensus on the role of the hippocampal component. Some research traditions stress its role in boosting memory strength, providing additional details, performance accuracy, and memory items (Stark & Squire, 2003; Squire & Bayley, 2007), although this effect may be super-additive (Shimamura & Wickens, 2009). Other traditions stress the role of the hippocampus in pattern separation and completion, providing unique codes for individual inputs and filling in any gaps in those codes (Gilbert, Kesner, & Lee, 2001; Norman & O'Reilly, 2003). Other work is concerned primarily with the content of the hippocampal representation and its involvment in spatial maps (O'Keefe & Dostrovsky, 1971; O'Keefe & Nadel, 1978; Maguire, Burgess, & O'Keefe, 1999; Byrne, Becker, & Burgess, 2007) or temporally coded episodes (Tulving, 1984;
Endel Tulving & Markowitsch, 1998; Huxter, Burgess, & O'Keefe, 2003). Finally, one tradition stresses the role of the hippocampus encoding arbitrary relations between items (Cohen, 1995; Eichenbaum, Dudchenko, Wood, Shapiro, & Tanila, 1999; Ryan, Althoff, Whitlow, & Cohen, 2000; Konkel, Warren, Duff, Tranel, & Cohen, 2008; Komorowski, Manns, & Eichenbaum, 2009).

These accounts each provide an important perspective on hippocampal function. However, because the neocortical accounts agree so closely, there is an opportunity to test the various predictions. Since each believes that the neocortex is the seat of semantic learning, and works

by via some sort of statistical learning algorithm (like that of a multi-layer perceptron trained with back-propagation), we can ask which type of information a neocortical model has the most difficulty learning. Within the consolidation framework outlined above, we know what hippocampal representations will change into, and from that we can estimate what they originally resembled. There is no doubt that the neocortex modeled as a multi-layer perceptron trained with back-propagation can learn any of the relations in question: it can approximate any input-output mapping arbitrarily closely (Hornik, 1991). However, speed and efficacy matter: if our neocortical model can rapidly and without interference assimilate a particular class of information then a hippocampal representation of that information is unnecessary, the neocortex is well designed to capture it. However, if the neocortex struggles with a particular type of information, then that information is a good candidate for an alternate representational format.

This approach is similar to the one taken by McClellend et al. in *Why are there complementary learning systems in the hippocampus and neocortex* (1995, c.f. Norman & O'Reilly 2003), which modeled a delayed non-match to sample task in primates (Zola-Morgan & Squire, 1985). Their hippocampal model was able to precisely encode individual object-response mappings and replay these in an interleaved fashion to their neocortical model. Interleaved learning of new and old patterns protected the neocortical model from catastrophic interference, in which new patterns overwrite old. Because this account contains multiple interacting memory stores (i.e., the complementary learning systems), it produces the consolidation phenomenon--increased resilience to hippocampal damage over time. Yet, because these two systems have different properties, it also produces a shift in representations from specific, pattern separated, instances of hippocampal representation to the more generalized neocortical perceptron-like representation.

Tuning this model to address the research frameworks above requires three substantial changes. First, most of the frameworks above make claims about declarative memory, and test memory with paired associate recall rather than match/non-match. Thus we must test the model's ability to map complex inputs to complex outputs. Second, these frameworks make different claims about the informational structure of the inputs (e.g., inputs and outputs are organized like a spatial map). Thus, we must create relationships between input-output

pairings, some of which emphasize the arbitrariness of the pairings, others of which emphasize spatial relations or rules.

Finally, the training must be adjusted to reflect a critical difference between human human amnesics and hippocampally lesioned animals. The primate lesion data modeled by McClelland et al., demonstrates that the primate's training (memory for familiar objects) becomes hippocampal independent over time, but the human data upon which many of these research frameworks is based makes a slightly different claim: that older memories are less dependent on the hippocampus than more recent. If memories become more durable over time, we might assume that older episodes would be less dependent on hippocampus. However, older memories also must face longer periods of interference. Since one of the chief claims of the CLS framework is that the hippocampus protects neocortical representations from interference via interleaved learning, it may be the case that we should predict higher levels of interference post lesion, which would predominately affect older memories. This is counter to the human findings, and would provide evidence against the interleaved learning account of the hippocampus, thus we divide training inputs into three serially presented epochs (old, middle-aged, and recent memories). Additionally, within the primate data the hippocampal trace decays over roughly the same period as the neocortical trace increases. This introduces another possible confound-if the hippocampal trace is required to provide an interleaved learning signal that reduces neocortical interference; we should expect to see interference return as the hippocampal trace decays. As in the case of episodic memory, this may not result in an overall performance decrease, rather we should expect to see the impact preferentially fall on those patterns which have been lost in the hippocampal trace. Thus our model also includes a hippocampal trace decay component to measure the impact of slowly losing inputs.

Model Architecture and Training

Architecture

The model was a three-layer perceptron trained by back-propagation, there were eight input units, eight output units, and twenty hidden units. In addition there were four "noisy" inputs, assigned random states on each training trial. All weights were initially random. It was coded in MATLAB (MATLAB Version 7.0, Simulink)

Training Design

Training was accomplished by exposing the perceptron to an input-output pairing randomly drawn from the hippocampal trace and applying the back-propagation of error learning rule (Rosenblatt 1961). We repeated this procedure for a number of iterations equal to the number of patterns currently present in the hippocampal trace. After a full training cycle, the network was tested (without learning), on all of the to-be-learned input patterns, to see what proportion of them achieved a correct match. Since output was real-valued rather than binary, we applied a threshold of 0.85 to the output for the purposes of counting matches.

We used three training corpora. In the first inputs and outputs were randomly paired integers (random pairings of all 256 unique binary codes of 8 bits). In the second input corpus, these codes were ordered in a 16x16 spatial grid with each input mapped to the output above it and to the left (i.e., each input was connected to an output such that a path would be formed from that input to the upper-left corner of the grid). In the third corpus, inputs and outputs were paired based on the "minimally interesting coding problem," each binary input was paired to its equivalent Gray code output. Thus, we tested arbitrary random pairings, spatially coded pairings, and a pairings encoding simple rule.

Further, the training corpora were divided into three epochs: "old" memories, "middle aged" memories, and "recent" memories that were presented serially (i.e., instead of being trained on the entire set of possible patterns, the network was trained in three epochs, each of which was randomly assigned 1/3 of the 256 total input-output pairings). The "middle" epoch was introduced after one third of the training cycles had elapsed, and the "recent" after two thirds of the training cycle. When a new epoch of input-output pairings was added, they became part of the same general training set from which the perceptron drew its input-output pairings. Thus, new, middle, and old patterns would be interleaved in training.

Additionally, we allowed the hippocampal trace to decay by removing input-output pair examples from the training pool. This was accomplished by applying a simple exponential decay function to the input corpus in which 0.01% of input patterns were lost after each training step. This "leak" was applied to each epoch separately.

Finally, we implemented interleaved learning by allowing recurrent connections from the neocortex to the hippocampus. The neocortex reintroduced some number of successfully learned output patterns (those which passed the "correct" threshold during testing as described above) into input training cycles. This makes the network somewhat similar to an Elman network (Jordan, 1986; Elman, 1990), though without the explicit learning of inputs as outputs. It does however, allow previously learned information to be interleaved with and interfere with new learning.

Testing Design

Experiment 1

We tested three identical versions of the model, one on each of the three different input corpora (arbitrary, spatial, and rule). In each instance, the hippocampal model was the same, it provided leaky traces of three "episodes" of the inputs (old, middle, and recent). In this manipulation, we did not allow the neocortical system to train itself on previously assimilated patterns.

Experiment 2

This was identical to experiment 1, except that now the neocortical system was allowed to introduce any successfully learned input-output pairing into the training corpus. Note that this means that the model went through additional training since there were a greater total number of patterns available (hippocampal trace + neocortical trace). Any patterns that were disrupted by interference were removed from the training corpus until they were re-learned.

Experiment 3

For the arbitrary pairings alone, we explored increasing the strength of neocortical feedback and decreasing the rate of hippocampal decay. Our goal was to "rescue" the retrograde amnesia finding by tuning these two parameters until old pairs would be best retained, middle pairs would be intermediate, and new pairs would be most disrupted by hippocampal lesion. We used the same mechanism as in experiment 2; however we introduced multiple copies of each neocortically-learned pattern into the training corpus, thus allowing previously learned pairs to

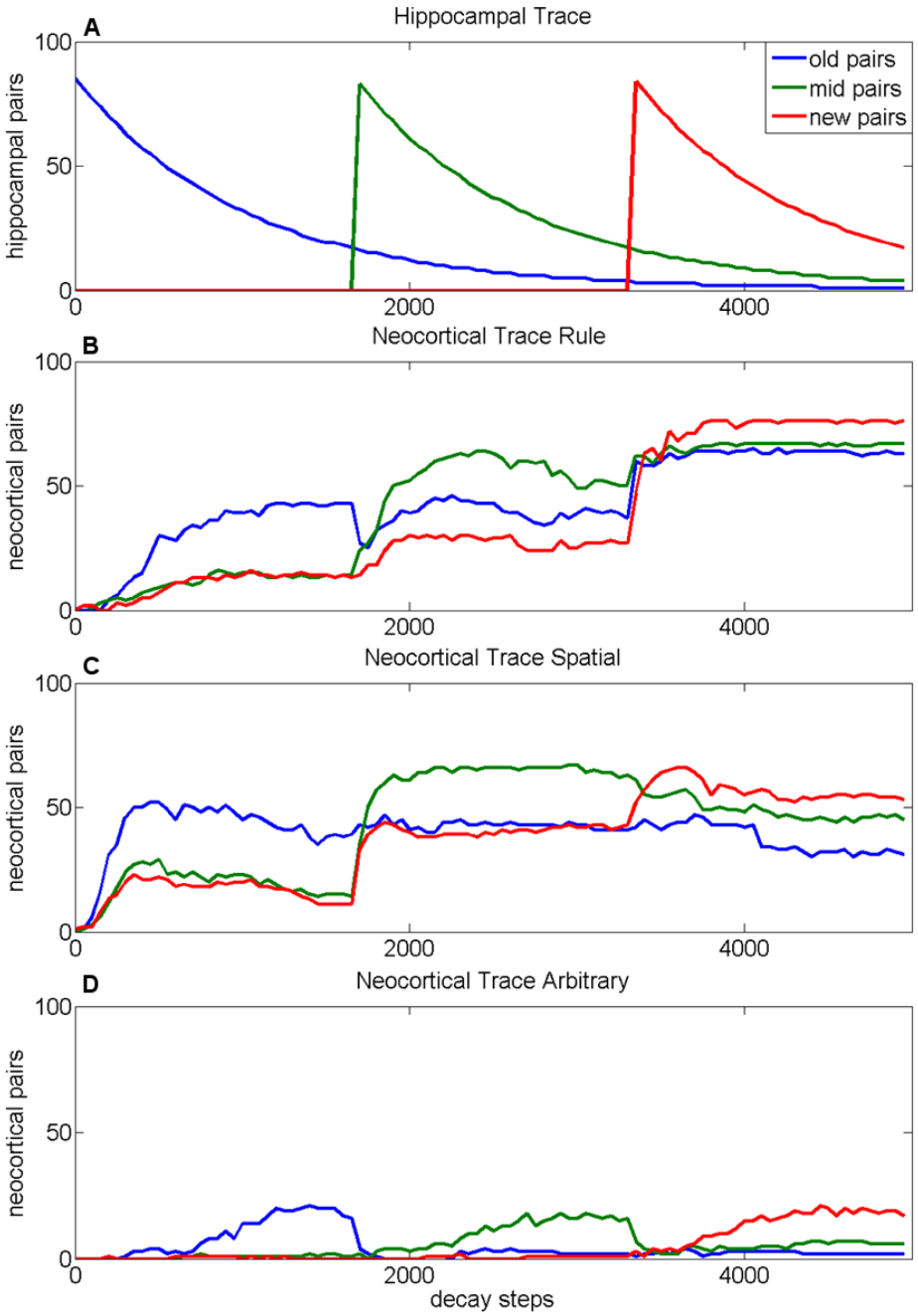
have a greater influence on model training (note that this once again results in a greater total amount of training due to the additional presentations of learned pairs).

Results

Experiment 1: Learning without neocortical feedback

In experiment 1 (Figure 1.4), at the end of the training, the neocortical model learned approximately 220 pairings from the rule based input corpus, 140 pairings from the spatial input corpus, and 28 pairings from the arbitrary corpus. For both the rule-based and spatial corpora, all three epochs (old, mid, and new) improved performance across the entire experiment, peaking in highest number of pairs learned at the end of training. For arbitrary pairings however, while the total number of patterns known generally increased throughout the training, each new epoch almost completely interfered with prior learning. At its peak, immediately before the introduction of the "mid" epoch, the "old" epoch had learned 20 patterns. This replicates the catastrophic interference finding (McCloskey & Cohen 1989, McClelland, McNaughton, & O'Reilly 1995).

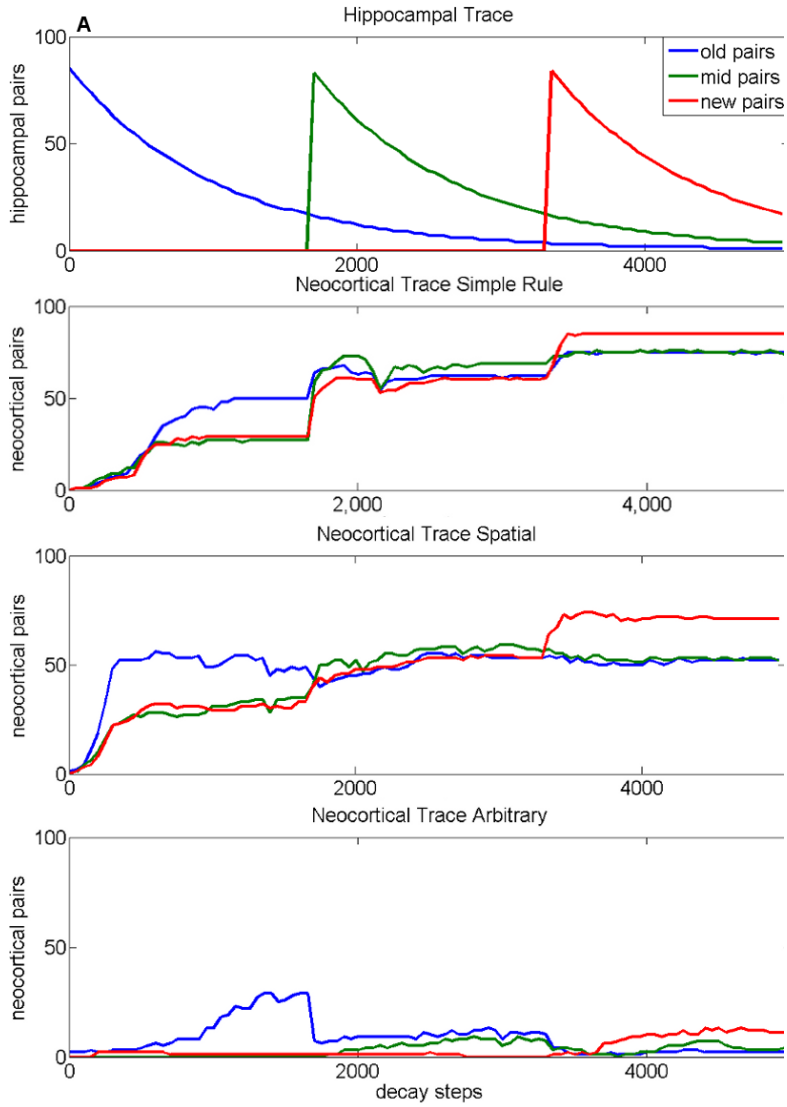
Figure 1.4


A
Hippocampal Trace
hippocampal pairs
old pairs
mid pairs
new pairs
B
Neocortical Trace Rule
neocortical pairs
C
Neocortical Trace Spatial
neocortical pairs
D
Neocortical Trace Arbitrary
neocortical pairs
0
50
100
2000
4000
decay steps

Experiment 2: Does neocortical feedback driven interleaved learning protect from catastrophic interference?

When neocortical learned patterns were returned to the training corpora, both of the structured data sets were more quickly learned, and to a higher total number of patterns. However neocortical feedback-driven interleaved learning did not substantially increase learning for arbitrary pairings (figure 1.5). However, interleaved learning did change the pattern of learning for arbitrary pairings. Upon introduction of the second epoch of patterns, the first epoch experienced interference, but not catastrophically so, and the old epoch remained the best learned during learning of the middle epoch. However, introduction of the new epoch once again caused considerable interference. This pattern of learning suggests that it may be possible to preserve the retrograde amnesia gradient by tuning the amount of hippocampal decay and neocortical feedback in a manner consistent with the model presented in the first section.

Figure 1.5

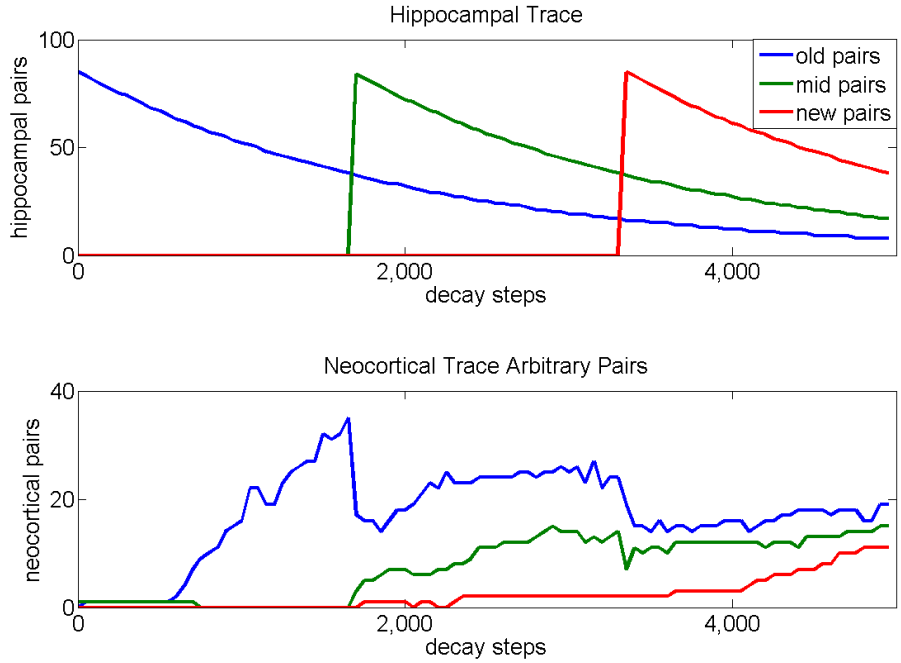


Experiment 3: Can neocortical feedback ever rescue arbitrary pairings?

With increased neocortical feedback and decreased hippocampal decay (figure 1.6) the neocortex successfully learns the arbitrary pairings without catastrophic interference and with a temporal gradient for the three epochs (i.e., old > mid > new). This replicates the experimental consolidation data from human patients (Squire, Slater, & Chance, 1975, Squire, Cohen, & Nadel 1984), and the interleaved learning finding (McClelland, McNaughton, & O'Reilly 1995, French 1999). However, it does not dramatically improve the amount of arbitrary pairings learned over

even the model without interleaved learning, despite decreases in the rate of hippocampal decay.

Figure 1.6



Discussion

In the first part of this document we argued that the *phenomenon* of memory consolidation, that memories become less disruptable over time, is best understood as a consequence of multiple, dynamic, interacting memory stores. As memory becomes more distributed, it must necessarily be more resistant to local amnestic disruptions. In the second part we examined one of many consolidation *processes*, the shift from individual memory examples to more semantic representations, using a modeling approach. This modeling approach highlighted properties of neocortical perceptron-style representations that provide some insight into what a complementary hippocampal representation ought to capture, and provides a more general to models of human memory. Further, these two models of memory consolidation are not in conflict, in fact, principles from the former model greatly informed the latter and assisted in tuning the parameters of the hippocampal and neocortical model to most accurately reflect the retrograde amnesia data.

In our model of the consolidation *process*, we explored how input-output pairings with different levels of shared informational structure might affect rates of consolidation from hippocampus to neocortex (c.f. Tse et al 2007). Additionally we explored how different tunings of hippocampal decay and neocortical feedback interact with these differences in information structure to produce different levels of preservation for memories of different ages.

In all conditions, even when there was no interleaved learning, the neocortical perceptron rapidly and easily learned both the simple rule-structured inputs, and the more complex, multi-dimensional spatially structured inputs. Yet in all conditions, even those with extremely strong interleaved learning, the neocortical perceptron struggled to learn arbitrary, random pairings, capturing fewer before hippocampal decay removed them from the training corpus, and experiencing greater levels of interference.

Why is the neocortex so poor at learning arbitrary pairings?

The neocortical perceptron works by approximating a function that maps the inputs to the outputs. For a structured relationship, such as a simple rule or a spatial manifold, this is a single function mapping all possible inputs to all possible outputs, and thus each example provides useful data for fitting all other examples. It is especially striking that in both simple rule and spatial conditions, due to the considerable intercorrelations between the pairings, model performance improved even on pairings that had *not yet been presented*. The model was not prescient; it was simply that each pair served as a sufficiently good example of the as yet unobserved pairs to drive performance.

For arbitrary pairings, however, each input-output mapping is a unique transform, and no pairing assists in fitting any other. The neocortical perceptron is certainly capable of fitting these pairings, but the function describing the mapping requires a huge number of parameters. Approximating it requires considerable bootstrapping-repetition of previous examples in an interleaved fashion. This bootstrapping induces a consistent, repetitive structure, and thus the neocortical perceptron is able to capture the information present in “random” pairings, albeit at

a much lower level than in the rule-based pairings, reaching criteria on only about a third of the total patterns.

The irony is that before any learning took place the neocortical perceptron was an excellent model of the actual function that generated the random pairings. Given a random binary digit as an input, the randomly wired perceptron produced a random paired associate. Through training the network was slowly converted from a pseudorandom number generator (i.e., an accurate model of the process that created the arbitrary pairs), to a list of the particular input-output pairings originally generated (i.e., an extremely poor random number generator). This is a simple case of over fitting; the neocortical perceptron has simply drawn a line through each studied point. This trading off of generalizability for specificity is an example of the bias/variance dilemma (Geman, Bienenstock, & Doursat, 1992).

This reluctance to learn random pairings is a feature, not a bug. The neocortical perceptron is using a gradient-descent algorithm to build a model of the statistical structure of its input-output pairings (visualize an "error" landscape with peaks and valleys, and a ball rolling down to find the lowest "error" point). Reducing the error for a one observation also reduces the error for the other observations *to the extent that these observations are correlated*. But for arbitrary pairings data, a good fit of previously observed pairings is no more likely to be a good fit of future pairings than a poor fit of previously observed pairings.

Why is there so much more interference for arbitrary pairings?

In the rule based and spatial conditions, new patterns do not cause interference, quite the opposite; they improve the fits of all three sets by providing additional inter-correlated information. However, in the arbitrary pairings condition each new corpus disrupted much of the prior learning. These disruptions can be ameliorated with the introduction of stronger neocortical training, but this change does not actually reduce interference, it simply ensures that the new patterns are more greatly interfered with than the old (a process sometimes called runaway reinforcement c.f. Meeter, 2003). This tradeoff is described in the literature as the "stability-plasticity dilemma" (Abraham & Robins, 2005; Carpenter & Grossberg, 1988;

Grossberg, 1980). Architectures can be tuned to be more sensitive to new data or old or any particular mixture thereof, but it can't be tuned to be *both* stable and plastic.

The only way to actually increase neocortical learning of arbitrary pairs is to increase the amount of hippocampal training available to the neocortex by decreasing the rate that pairings are removed from the hippocampal store (as in experiment 4), and to slow this rate as new pairings are added to allow the interleaved learning algorithm to complete reduce the comparatively greater interference between the multiple sets of pairings. The most effective tuning is to never remove pairings from the hippocampal store (as in McClelland, McNaughton, & O'Reilly 1995). This is essentially combining all of the input data into a single corpus, and re-commencing a complete search of the error space after each new example to find the global minima. This way no information can be lost.

This may seem implausible, since it predicts that all experiences are permanently stored in the hippocampus, and that after each new experience the entire neocortex (or at least the portion of it in which arbitrary associations are coded) is completely rewired to reflect the combination of new and old experience (additionally it suggests a strong need to understand what constitutes an "experience"). Yet, with our current limited knowledge of brain function perhaps this indeed what happens. In addition, some prior studies suggest that this computational problem can be considerably ameliorated by simply increasing the model's tolerance (French 1999), effectively smoothing out the error surface and allowing the model to return near matches. In this scheme the neocortex provides completely redundant representations of the hippocampal trace for arbitrary pairings, but it continues to be valuable for learning structured information, and can provide some temporally graded arbitrary pairings in the event of a hippocampal lesion. There are some theories of hippocampus-neocortex consolidation that make essentially this claim (Rosenbaum, Winocur, & Moscovitch 2001).

Why does the neocortex learn the arbitrary pairings that it does?

The neocortical perceptron is better at capturing pairings that share common structure, appear early in training or immediately accompanying new training epochs, and are frequently repeated (either via neocortical feedback or hippocampal training). These are not different properties.

"Firstness", "oldness", "repetition," etc. are all consistent structural properties of the pairings presented to the neocortical perceptron, and it learns those consistencies. In the case of arbitrary pairings, there is no structure intrinsic to the pairings that the model can leverage, but the model can learn the structure imposed by the experimenter's training algorithm. There have been many improvements to training algorithms developed since the original catastrophic interference finding (McCloskey & Cohen 1989) to leverage just such structure to minimize interference between data sets (Elman, 1990, Ans et al., 2002, Musca, Rousset, & Ans, 2009). These advances are valuable, and it is almost certain that some similar algorithms are at work in managing the interactions between hippocampal and neocortical data.

How does this speak to research traditions on hippocampal function?

This paper argues that the neocortex is capable of learning structure, but incapable of learning unstructured, random, or arbitrary associations except when such associations have some external structure imposed upon them. This suggests a role for the hippocampus: it captures unstructured, random, or arbitrary associations, and imposes some form of structured playback upon them. The neocortex can capture this structured playback using precisely the same mechanisms it uses to capture structured associations present in perception. Since everyday experience contains so many examples of unstructured, coincidental, or conjunctive data, it is no surprise that the loss of the hippocampus creates such a profound amnesia for new declarative memory. This reaffirms the ability of the hippocampus to provide a time-limited boost to neocortical function (Squire, Cohen, & Nadel 1984, Alvarez & Squire 1994), and suggests that theories of hippocampal function ought to explain the nature of hippocampally-imposed structure.

And this is precisely what these traditions do! One tradition suggest that the hippocampus is critical for learning stimuli that unfold in space (O'Keefe & Dostrovsky, 1971, O'Keefe & Nadel 1978,) and time (Tulving 2002), and that the structure it imposes upon associations between items is like the dimensional structure of space and time. A second tradition stresses the importance of the hippocampus for learning unique conjunctions of features (Norman & O'Reilly 2003), and that it imposes structure via control over the resolution of the representation (Aimone, Deng, & Gage 2011), coding precise, pattern-separated examples and filling in missing

information via pattern completion. A third tradition suggests that the hippocampus is critical for capturing arbitrary relations regardless of their resolution, or content (Eichenbaum, 2000), and that it imposes only the structure present in the bindings of those relations themselves.

The modeling results presented here challenge these research traditions to produce theories that carefully unpack how arbitrary associations are converted into structured. For example, a naïve spatial theory that claims the hippocampus is necessary for all forms of spatial representation is almost certainly wrong; the spatial path that we encoded in our spatial associations was perfectly learnable by our neocortical model. Indeed, any dataset that contains internally correlated associations, such that each example improves predictive performance for as-yet-unseen examples, does not require any additional structure to be learned. A more nuanced spatial theory, however, that argues that arbitrary, unstructured information (e.g., the order in which a set of landmarks was visited), is mapped onto a spatial representation (e.g., a 2D map), and thus can be learned by a neocortical system, is plausible. This theory may be empirically wrong, a map may not in fact be the structure taught to the neocortex, but at least it does not claim that the hippocampus works by imposing a spatial structure on spatially structured data. Such structured data could also be learned without a hippocampus.

Whatever algorithm the hippocampus uses to convert unstructured information to structure, it produces powerful and flexible recall. A hippocampal memory of a past event can answer who, what, when, where, how, and why-despite the fact that none of these questions were present at encoding. This is an interesting future direction for hippocampal research: how does the hippocampus convert unstructured information to structured at recall? What kind of representation can capture the arbitrary, incidental nature of out experiences, and yet store them in such a way that they can be easily accessed and flexibly applied in new situations? What, after all, is the similarity between events that *lack* shared structure, and how does the hippocampus bring them into alignment during recall?

References

Abraham, W. C., & Robins, A. (2005). Memory retention--the synaptic stability versus plasticity dilemma. *Trends in neurosciences*, *28*(2), 73–8. doi:10.1016/j.tins.2004.12.003

Aimone, J. B., Deng, W., & Gage, F. H. (2011). Resolving new memories: a critical look at the dentate gyrus, adult neurogenesis, and pattern separation. *Neuron*, *70*(4), 589–96. doi:10.1016/j.neuron.2011.05.010

Ans, B., Rousset, S., French, R. M., & Musca, S. (2002). Preventing Catastrophic Interference in Multiple-Sequence Learning Using Coupled Reverberating Elman Networks Dual-Network Architectures. *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, *NJ:LEA*.

Bramham, C. R., & Messaoudi, E. (2005). BDNF function in adult synaptic plasticity: the synaptic consolidation hypothesis. *Progress in neurobiology*, *76*(2), 99–125. doi:10.1016/j.pneurobio.2005.06.003

Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: a neural model of spatial memory and imagery. *… Review; Psychological Review*, *114*(2), 340–375. doi:10.1037/0033-295X.114.2.340.Remembering

Carpenter, G. a., & Grossberg, S. (1988). The ART of adaptive pattern recognition by a self-organizing neural network. *Computer*, *21*(3), 77–88. doi:10.1109/2.33

Cherkin, A. (1969). Kinetics of memory consolidation: Role of amnesic treatment parameters. *Proceedings of the National Academy of …*, 1094–1101. Retrieved from http://www.pnas.org/content/63/4/1094.short

Cohen, N. J. (1995). *Memory, amnesia, and the hippocampal system*. Retrieved from http://openurl.library.uiuc.edu/sfxlcl3?sid=google&auinit=NJ&aulast=Cohen&title=Memory, amnesia, and the hippocampal system&genre=book&isbn=0262531321&date=1995

Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M., & Tanila, H. (1999). The hippocampus, memory, and place cells: is it spatial memory or a memory space? *Neuron*, *23*(2), 209–26. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10399928

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature reviews. Neuroscience*, *1*(1), 41–50. doi:10.1038/35036213

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. doi:10.1016/0364-0213(90)90002-E

French, R. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, *3*(4), 128–135. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12079555

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/neco.1992.4.1.1

Gilbert, P. E., Kesner, R. P., & Lee, I. (2001). Dissociating hippocampal subregions: double dissociation between dentate gyrus and CA1. *Hippocampus*, *11*(6), 626–36. doi:10.1002/hipo.1077

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological review*, *87*(1), 1–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7375607

Hebb, D. (1949). The organization of behavior: A neuropsychological theory. *New York*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+Organization+of+Behavior#0

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, *4*(1989), 251–257. Retrieved from http://www.sciencedirect.com/science/article/pii/089360809190009T

Huxter, J., Burgess, N., & O'Keefe, J. (2003). Independent rate and temporal coding in hippocampal pyramidal cells. *Nature*, *425*(6960), 828–832. doi:10.1038/nature02058.Independent

Izquierdo, I., Bevilaqua, L. R. M., Rossato, J. I., Bonini, J. S., Medina, J. H., & Cammarota, M. (2006). Different molecular cascades in different sites of the brain control memory consolidation. *Trends in neurosciences*, *29*(9), 496–505. doi:10.1016/j.tins.2006.07.005

Jordan, M. (1986). Serial order: A parallel distributed processing approach. *Advances in Psychology*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0166411597801112

Jost, A. (1897). Die Assoziationsfestigkeit in ihrer Abhangigkeit von der Verteilung der Wiederholungen [The strength of associations in their dependence on the distribution of repetitions]. *Zeitschrift fur Psychologie und Physiologie der Sinnesorgane*, *16*, 436 – 472.

Komorowski, R. W., Manns, J. R., & Eichenbaum, H. (2009). Robust conjunctive item-place coding by hippocampal neurons parallels learning what happens where. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *29*(31), 9918–29. doi:10.1523/JNEUROSCI.1378-09.2009

Konkel, A., Warren, D. E., Duff, M. C., Tranel, D. N., & Cohen, N. J. (2008). Hippocampal amnesia impairs all manner of relational memory. *Frontiers in human neuroscience*, *2*(October), 15. doi:10.3389/neuro.09.015.2008

Korman, M., Doyon, J., Doljansky, J., Carrier, J., Dagan, Y., & Karni, A. (2007). Daytime sleep condenses the time course of motor memory consolidation. *Nature neuroscience*, *10*(9), 1206–13. doi:10.1038/nn1959

Lechner, H., Squire, L., & Byrne, J. (1999). 100 years of consolidation—remembering Müller and Pilzecker. *Learning & Memory*, 77–87. doi:10.1101/lm.6.2.77

Maguire, E. a, Burgess, N., & O'Keefe, J. (1999). Human spatial navigation: cognitive maps, sexual dimorphism, and neural substrates. *Current opinion in neurobiology*, *9*(2), 171–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10322179

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*(3), 419–57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7624455

McGaugh, J L. (1966). Time-dependent processes in memory storage. *Science (New York, N.Y.)*, *153*(3742), 1351–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5917768

McGaugh, J. L. (2000). Memory--a Century of Consolidation. *Science*, *287*(5451), 248–251. doi:10.1126/science.287.5451.248

McKenzie, S., & Eichenbaum, H. (2012). New approach illuminates how memory systems switch. *Trends in cognitive sciences*, *16*(2), 102–3. doi:10.1016/j.tics.2011.11.010

McLelland, J., Rumelhart, D., & Group, P. R. (1986). Parallel distributed processing: Explorations in the microstructure of cognition. *Psychological and Biological …*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Parallel+distributed+processing:+explorations+in+the+microstructure+of+cognition#7

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connections networks: The sequential learning problem. In G. H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 24, pp. 109–165). San Diego: Academic Press.

Meeter, M. (2003). Control of consolidation in neural networks: Avoiding runaway effects. *Connection Science*, *15*(1), 45–61. doi:10.1080/0954009031000149591

Misanin, J. R., Miller, R. R., & Lewis, D. J. (1968). Retrograde amnesia produced by electroconvulsive shock after reactivation of a consolidated memory trace. *Science (New York, N.Y.)*, *160*(3827), 554–5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5689415

Moscovitch, M., Rosenbaum, R. S., Gilboa, A., Addis, D. R., Westmacott, R., Grady, C., McAndrews, M. P., et al. (2005). Functional neuroanatomy of remote episodic, semantic and spatial memory: a unified account based on multiple trace theory. *Journal of anatomy*, *207*(1), 35–66. doi:10.1111/j.1469-7580.2005.00421.x

Muller, G. E., & Pilzecker, A. (1900). Experimentelle Beitrage zur Lehre vom Gedachtnis. *Zeitschrift fur Psychologie. Erganzungsband*, *1*(1), 1–300.

Musca, S. C., Rousset, S., & Ans, B. (2009). Artificial neural networks whispering to the brain: nonlinear system attractors induce familiarity with never seen items. *Connection Science*, *21*(4), 359–377. doi:10.1080/09540090903067493

Nadel, L, Samsonovich, a, Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. *Hippocampus*, *10*(4), 352–68. doi:10.1002/1098-1063(2000)10:4<352::AID-HIPO2>3.0.CO;2-D

Nadel, Lynn, Winocur, G., Ryan, L., & Moscovitch, M. (2007). Systems consolidation and hippocampus: two views. *Debates in Neuroscience*, *1*(2-4), 55–66. doi:10.1007/s11559-007-9003-9

Nader, K., Schafe, G. E., & Le Doux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, *406*(6797), 722–6. doi:10.1038/35021052

Norman, K. a, & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological review*, *110*(4), 611–46. doi:10.1037/0033-295X.110.4.611

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain research*, *34*, 171–175. Retrieved from http://psycnet.apa.org/psycinfo/1972-08318-001

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. *Why People Get Lost*. Oxford, UK: Clarendon Press. Retrieved from

http://www.ingentaconnect.com/content/oso/7347120/2010/00000001/00000001/art00006

Ribot, T. (1881). *Les maladies de la memoire*. Paris: Germer Bailliere.

Rosenbaum, R. S., Winocur, G., & Moscovitch, M. (2001). New views on old memories: re-evaluating the role of the hippocampal complex. *Behavioural brain research*, *127*(1-2), 183–97. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11718891

Rosenblatt, F. (1961). *Principles of neurodynamics; perceptrons and the theory of brain mechanisms.* Buffalo, N.Y.: Cornell Aeronautical Laboratory.

Rumelhart, D., & Ortony, A. (1976). *The representation of knowledge in memory*. Retrieved from http://www.cs.northwestern.edu/~ortony/Andrew_Ortony_files/Rumelhart and Ortony.pdf

Ryan, J. D., Althoff, R. R., Whitlow, S., & Cohen, N. J. (2000). Amnesia is a deficit in relational memory. *Psychological science*, *11*(6), 454–61. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11202489

Shimamura, A. P., & Wickens, T. D. (2009). Superadditive memory strength for item and source recognition: the role of hierarchical relational binding in the medial temporal lobe. *Psychological review*, *116*(1), 1–19. doi:10.1037/a0014500

Shimizu, E., Tang, Y. P., Rampon, C., & Tsien, J. Z. (2000). NMDA receptor-dependent synaptic reinforcement as a crucial process for memory consolidation. *Science (New York, N.Y.)*, *290*(5494), 1170–4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11073458

Squire, L R. (1992). Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological review*, *99*(2), 195–231. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1594723

Squire, L R, & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current opinion in neurobiology*, *5*(2), 169–77. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7620304

Squire, L., Cohen, N., & Nadel, L. (1984). The medial temporal region and memory consolidation: A new hypothesis. *Memory consolidation*. Retrieved from http://books.google.com/books?hl=en&lr=&id=tVwr_x2IwVoC&oi=fnd&pg=PA185&dq=The+Medial+Temporal+Region+and+Memory+Consolidation+:+A+New+Hypothesis+,+I&ots=jjMgzsRiX4&sig=iMBl9XgrKgKjODPxnsdbXhEge0U

Squire, Larry R, & Bayley, P. J. (2007). The neuroscience of remote memory. *Current opinion in neurobiology*, *17*(2), 185–96. doi:10.1016/j.conb.2007.02.006

Squire, L. R., Slater, P. C., and Chance, P. (1975). Retrograde amnesia temporal gradient in very long-term memory following electroconvulsive therapy. Science, 187, 77-79.

Stark, C., & Squire, L. (2003). Hippocampal damage equally impairs memory for single items and memory for conjunctions. *Hippocampus*, *13*(2), 281–292. doi:10.1002/hipo.10085.Hippocampal

Stickgold, R. (1998). Sleep: off-line memory reprocessing. *Trends in cognitive sciences*, *2*(12), 484–92. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21227299

Suzuki, A., Josselyn, S. a, Frankland, P. W., Masushige, S., Silva, A. J., & Kida, S. (2004). Memory reconsolidation and extinction have distinct temporal and biochemical signatures. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *24*(20), 4787–95. doi:10.1523/JNEUROSCI.5491-03.2004

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. a, Wood, E. R., Witter, M. P., et al. (2007). Schemas and memory consolidation. *Science (New York, N.Y.)*, *316*(5821), 76–82. doi:10.1126/science.1135935

Tulving, E. (1984). Precis of elements of episodic memory. *Behavioral and Brain Sciences*. Retrieved from http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=6709108

Tulving, Endel, & Markowitsch, H. (1998). Episodic and declarative memory: role of the hippocampus. *Hippocampus*, *204*, 198–204. Retrieved from

http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1098-1063(1998)8:3%3C198::AID-HIPO2%3E3.0.CO;2-G/abstract

Walker, M. P., & Stickgold, R. (2004). Sleep-dependent learning and memory consolidation. *Neuron*, *44*(1), 121–33. doi:10.1016/j.neuron.2004.08.031

Wilson, M., & McNaughton, B. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, *5*(14), 14–17. Retrieved from http://www.sciencemag.org/content/265/5172/676.short

Winocur, G., Moscovitch, M., & Bontempi, B. (2010). Memory formation and long-term retention in humans and animals: convergence towards a transformation account of hippocampal-neocortical interactions. *Neuropsychologia*, *48*(8), 2339–56. doi:10.1016/j.neuropsychologia.2010.04.016

Zola-Morgan, S., & Squire, L. (1985). Medial temporal lesions in monkeys impair memory on a variety of tasks sensitive to human amnesia. *Behavioral Neuroscience*. Retrieved from http://psycnet.apa.org/journals/bne/99/1/22/

SPATIAL RECONSTRUCTION BY PATIENTS WITH HIPPOCAMPAL DAMAGE IS DOMINATED BY RELATIONAL MEMORY ERRORS

In press in Hippocampus

Patrick D. Watson[1], Joel L. Voss[2], David E. Warren[3], Daniel Tranel[3,4] and Neal J. Cohen[1]

[1]Beckman Institute for Advanced Science and Technology,
University of Illinois Urbana-Champaign, Urbana, Illinois

[2]Department of Medical Social Sciences and Interdepartmental Neuroscience Program
Northwestern University Feinberg School of Medicine, Chicago, IL

[3]Department of Neurology,
University of Iowa, Iowa City, Iowa

[4]Department of Psychology,
University of Iowa, Iowa City, Iowa

Acknowledgements**:** Research was supported by a joint Sandia National Laboratories-Beckman Institute Fellowship to P.D.W., a US National Institutes of Health (NIH) Pathway to Independence award (K99/R00-NS069788) and a Beckman Institute Postdoctoral Fellowship Award to J.L.V., by funds from the Kiwanis Foundation to D.T., and by NIH grants R01-MH062500 to N.J.C. and P50-NS19632 to D.T.

Abstract

Hippocampal damage causes profound yet circumscribed memory impairment across diverse stimulus types and testing formats. Here, within a single test format involving a single class of stimuli, we identified different performance errors to better characterize the specifics of the underlying deficit. The task involved study and reconstruction of object arrays across brief retention intervals. The most striking feature of patients' with hippocampal damage performance was that they tended to reverse the relative positions of item pairs within arrays of any size, effectively "swapping" pairs of objects. These "swap errors" were the primary error type in amnesia, almost never occurred in healthy comparison participants, and actually contributed to poor performance on more traditional metrics (such as distance between studied and reconstructed location). Patients made swap errors even in trials involving only a single pair of objects. The selectivity and severity of this particular deficit creates serious challenges for theories of memory and hippocampus

Introduction

The precise role of the hippocampus in memory is a topic of much investigation. Observations of patients with amnesia following hippocampal damage reveal a complicated pattern of impaired and spared memory functions. The deficits include profound and pervasive impairment in learning and remembering new facts and events, preventing patients, for example, from normal learning of new routes, places, or people, and from keeping track of appointments or events of daily life. Yet other aspects of memory such as skill learning remain fully intact. Taken together with converging evidence using other neuroscience methods, the functional dissociations resulting from hippocampal damage illuminate the scope and limits of hippocampal involvement in memory (Cohen and Squire, 1980; Cohen and Eichenbaum, 1993; Schacter and Tulving, 1994; McClelland, McNaughton, and O'Reilly, 1995; Aggleton and Brown, 1999; Nadel, Samsonovitch,Ryan, and Moscovitch, 2000; Eichenbaum, and Cohen, 2001; Eichenbaum, Yonelinas, and Ranganath, 2007).

Some research has focused on the role of the hippocampus in processing spatial information and maintaining a dynamic, flexible "mental map" of space; highlighting deficits in numerous spatial tasks after hippocampal damage, along with evidence of place-sensitive cells in the

hippocampus, and correlations between hippocampal volume and spatial ability (O'Keefe, and Nadel, 1978; Hayes, Ryan, Schnyer, and Nadel, 2004; Ryan, Lin, Ketcham, and Nadel 2010). Another line of research points to the role of the hippocampus in managing declarative memory load—finding deficits following hippocampal damage when capacity limits are reached or when delays become sufficiently long (Stark and Squire 2003; Squire, Stark, and Clark 2004; Gold, Smith, Bayley, Shrager, Brewer, Stark, Hopkins, and Squire, 2006). Other research findings emphasize the nature of the representations generated by the hippocampus (Cohen and Eichenbaum, 1993; Henke, 2010). For example, one extensive body of work reports impairment following hippocampal damage for relational memory, showing deficits in representing the relationships among disparate elements of scenes or events (Eichenbaum and Cohen 2001) or in representing cross-modal bindings (Marr, 1971; Damasio, 1989; Vargha-Khadem et al., 1997; Aggleton and Brown, 1999). Such deficits are manifested for all manner of accidental or arbitrary relations (Konkel, Warren, Duff, Tranel, and Cohen, 2008; Konkel and Cohen 2009), regardless of the timescale over which the relational information must be maintained (Hannula, Tranel, and Cohen, 2006; Hannula, Ryan, Tranel, and Cohen 2006; Warren, Duff, Jensen, Tranel, and Cohen 2012).

Across all of these different lines of research, hippocampal damage is seen to produce memory impairment – and hence hippocampus is clearly engaged – in many different categories of stimuli and test formats. This highlights the broad scope and pervasiveness of hippocampal function in memory, but also makes identifying the critical factors(s) that tie these findings together challenging. What is the fundamental nature of the deficit, and hence the role of the hippocampus in memory?

Even within a single stimulus domain and test format, memory impairment following hippocampal damage can be difficult to interpret unambiguously. Deficits in learning to navigate among multiple locations in large spatial environments could be attributed to spatial, load/capacity, relational, or other demands. In the current experiment, we employed a simple memory test complemented by a set of performance analyses rich enough to identify various categories of errors arising from the different predicted deficits, permitting a more direct evaluation of various predictions within a single experimental paradigm.

Our goal was to determine whether hippocampal damage causes errors even in short-delay spatial reconstruction, and whether specific types of errors occur disproportionately, in a fashion that would be helpful in assessing the role of hippocampus in various types or aspects of memory.

Materials and Methods

Participants

Behavioral data were collected from three individuals with amnesia subsequent to hippocampal damage and from four comparison participants with no known neurological impairments. Each comparison participant was matched to an amnesic participant in age (within 2 years), educational attainment (within 1 year), sex, and handedness. Table 1 summarizes each amnesic patient's etiology along with demographic, neuropsychlogical, and hippocampal volumetric measures where available.

Amnesic etiology and neuroanatomy

All amnesic participants suffered acute episodes in adulthood that rendered them memory impaired, and previous reports have established that each amnesic participant has substantial damage to the hippocampus bilaterally. Patient 2363 became amnesic after cardiac arrest and an accompanying anoxic episode that resulted in selective regional atrophy without lesion. The bilateral volume of his hippocampus has been measured and quantified, and found to be significantly less than normal for his age and sex based on a regression model fit to hippocampal volumes of healthy comparison participants (Allen et al., 2006), with a Studentized residual value of -2.64. 2636's cerebral gray matter has been characterized as less than normative (Allen et al., 2006) with a Studentized residual value of -2.47, which was driven in large part by a normatively small amount of parietal gray matter (Studentized residual value of -2.78). Gray matter volume in the frontal and temporal lobes was less than normative but unremarkable. Patient 1846 became amnesic after a combined, hour-long episode of status epilepticus and anoxia that resulted in selective regional atrophy without lesion. Her hippocampus is atrophied bilaterally, with the atrophy being greater on the left (Warren et al., 2012). Her bilateral hippocampal volume has also been measured and quantified, and found to be significantly less than normal, with a Studentized residual value of -4.23 (Allen et al., 2006). Outside of the MTL, 1846's brain has been described as normal except for "some evidence of

cortical thinning in the paracentral lobule and precuneus" (Warren et al., 2012) that may be related to her anoxic etiology. Otherwise her brain volume (gray and white matter, both total and per-lobe) has been characterized as normative (Allen et al, 2006). Patient 2308 became amnesic after an episode of herpes simplex encephalitis (HSE) that damaged significant portions of his left and right temporal lobes. Specifically, 2308 has bilateral damage to the medial temporal lobe (including the amygdalae and the anterior hippocampi in their entirety) and medial temporal poles along with unilateral damage to left ventral and lateral temporal lobe extending to the left temporal pole (Cavaco et al., 2012). The hippocampal lesions in 2308 are so extensive that it is not possible to measure meaningfully the remaining tissue and make a quantitative comparison to normative data. Beyond the temporal lobes, 2308 has left-lateralized damage to the insular cortex, basal forebrain, and the posterior portion of orbitofrontal cortex, and right-lateralized damage to the insular cortex.

Amnesic neuropsychology

Neuropsychological examination confirmed severe declarative memory impairment in each amnesic participant, with performance on the Wechsler Memory Scale - Third edition (WMS-III, Wechsler, 1997a) at least 25 points lower than their performance on the Wechsler Adult Intelligence Scale - Third edition (WAIS-III, Wechsler, 1997b), and the average delay score on the WMS-III more than two standard deviations below the population mean. Memory impairments were selective, in that, for example, none of the amnesic participants showed any systematic impairment on a battery of neuropsychological tests of executive function, including trail making, Wisconsin card-sorting, controlled oral word associations, and the tower of London (Konkel, et al., 2008).

Experimental Paradigm

We used a spatial reconstruction task (Huttenlocher and Presson 1979; Smith and Milner, 1981; Jeneson, Mauldin, and Squire 2010). During each trial the participant studied an object array (containing between 2 and 5 objects), arranged on a 100cm-by-100cm white tabletop, and then had to reconstruct the spatial layout of the objects after a brief eyes-closed delay. During the "study" portion of the trial, the participant picked up each object with his or her dominant hand, named it, and immediately placed it back in the same location. When the participant had finished, he or she covered his or her eyes for approximately 4 s while the location of the objects was recorded in a digital photograph and the objects were cleared from the table. After the 4 s

"blind" period, the participant attempted to place the objects back into the original configuration (reconstruction). The final location of the objects was then recorded in a second digital photograph, and the next trial began after a short break. Some trials involved familiar, nameable objects (e.g., a pen, a button, a toy car, etc.) and others involved novel objects carved out of white foam blocks into various complex shapes and covered in patterns of simple lines and other shapes ("Greebles" c.f. James, Shima, Tarr, and Gauthier, 2005). All of the materials composing the novel shapes were of the same composition and color for each stimulus, such that stimuli could not be distinguished based on simple features. Blocks involving novel objects were interleaved with familiar-object blocks, with an equal number of each block type in each experimental session. During the study portion of each trial, participants picked up each object (as for the trials involving familiar objects), but counted integers aloud instead of providing names, given the obvious difficulties that would be associated with attempting to name these novel objects.

Another condition was also included that varied from the main paradigm in the instructions given to the participants. In this condition, participants were instructed to create the initial configuration of the objects (whether novel or familiar) on each trial themselves, rather than studying locations selected by the experimenter (i.e., objects were self-placed rather than experimenter-placed). This condition was administered in blocks randomly interposed with the main experimental blocks described here. Data from this condition are not reported here because amnesic participants self-positioned objects in grossly different patterns than did comparison participants, thus confounding comparisons of subsequent relational memory performance.

The digital photographs taken after the study and reconstruction portions of each trial were analyzed offline using MATLAB software (MATLAB version 7.9 Natick, Massachusetts: The MathWorks Inc., 2009). The edges of the table were identified via a semi-automated algorithm and were used to warp the coordinate space of the table into a common, Cartesian coordinate system via linear deformation. There was no more than 1cm of displacement in the position of the table edges for the reconstruction image relative to the study image for any trial, indicating that table and camera movements did not contribute significantly to measures of reconstruction

errors. The location of the center of each object was marked prior to deformation and object coordinates in the common reference frame were used for analysis.

Memory Measures

This task permitted multiple error types, and assessed reconstruction performance using five metrics (Fig. 1) capable of capturing this error heterogeneity. Because this task involved spatial reconstruction performance could be evaluated with respect to spatial theories. Because the task involved variable set sizes, including as few as two objects, performance could be evaluated for the effects of memory load. Finally, our use of unique error analyses permitted a rich evaluation of performance with respect to the different types of representation required.

Different object-configuration schemes are sensitive to different types of reconstruction errors. Spatial reconstruction experiments have historically used the *item misplacement* measure, which is simply the distance (in cm) between each item's studied location and the location where each item was placed during reconstruction (Huttenlocher and Presson, 1979, Smith and Milner, 1981, Jenesen, Mauldin, and Squire 2010). Although it is a simple, intuitively appealing analytic approach, it assumes that the underlying representation is of each item's location in a grid-like Cartesian coordinate system. While some theories of hippocampal spatial processing might endorse such a map-like representational scheme, it is not clear that all do (e.g., some spatial theories might argue that the representation is more like a path than a map). Moreover, such an item-based approach does not take into account the possibility that performance might be driven by memory representations of the configuration of the objects or the relations among the objects. Accordingly, we also measured spatial reconstruction using *edge resizing* and *edge deflection* metrics, which measure reconstructed changes in the length and direction (in cm and radians respectively) of vectors between each pair of items. These metrics assume that each items' location in the underlying representation serves as a landmark for each other item's location, with polar coordinate-like vectors between them. Moreover, we also measured memory for the overall arrangement of items with a *rearrangement* metric. This metric assumes that the underlying representation has no fine-grained representation of distance or angle, but rather reduces the configuration of studied items to simple shape via perceptual closure and measures the frequency of categorical changes in shape (e.g., a square changing to a rhombus, or a line changing its direction). This array of measures applied to amnesic

performance allowed us to better characterize the hippocampal contribution to spatial representation.

One other new metric assessed is *swaps,* which is the rate at which any pair of objects "swap" places between study and reconstruction (i.e., when the correct locations were filled but with mis-assignment of particular objects to particular locations). This metric was applied for each possible pair of objects in a given reconstruction of 2-5 objects. Because the number of possible pairwise swaps increases combinatorially while set size increases linearly, our metric here was swaps-per-pairwise-relation, thereby avoiding confounding the increase in relational complexity with the increase in the number of items. We measured such swaps by counting the frequency that the vector connecting each pair of objects reversed direction (i.e., the sign of the vector's x and y components changed simultaneously between study and reconstruction). This metric assumes an underlying representation that involves binding each trial's set of object-identities onto each trial's set of locations. The experimenters' assignment of particular objects to particular locations was random. Thus, successful performance required memory for arbitrary relations, and the incidence of swaps in patients with hippocampal damage could be used to assess the role of hippocampus in relational memory.

Patient to Comparison ratios

Patient to comparison ratios were simply calculated as patient performance over healthy comparison performance on each metric. Standard error bars were obtained via propagation of uncertainty:

For

$$f = \frac{a}{b}$$

$$(\sigma_f)^2 = \left(\frac{\sigma_a}{a}\right)^2 + \left(\frac{\sigma_b}{b}\right)^2$$

Where a corresponds to patient performance, b corresponds to healthy comparison performance, and a and b are independent and uncorrelated.

Random performance

Random performance was calculated by assuming a pair of objects was placed at two random locations (indexed by random, unique pairs of x,y coordinates drawn from our 100x100 spatial grid) during the study phase, and that these two objects were placed at two random locations during the test phase. The mean of the resulting misplacement, edge resizing, edge deflection, rearrangment, and swaps were calculated by applying our measures above. For any pair of objects placed at random, mean item misplacement ought to be 52 cm, mean edge resizing 28 cm, mean edge deflection $\pi/2$ radians, and the expected probability for rearrangements and swaps 50% and 25% respectively.

Note that this definition of random performance does not take into account biases present in our experimenter or participants. For instance, during the study phase no object was positioned less than 10cm from the outer edge of the table, meaning that the utilized area of the table was closer to 90x90. However, since making use of this information would imply some level of memory on the part of our participants, we chose to use the less constrained definition of randomness above.

Results

What kinds of errors do amnesic participants make?

We first consider reconstruction performance for displays that contained familiar, everyday objects (Fig. 2.1).

Memory load

Relative to healthy comparison participants, at all set sizes, and on all metrics, amnesic participants were impaired in reconstructing object configurations after an approximately 4s delay. ANOVAs with factors Group (amnesic/comparison) and Set Size (2, 3, 4, and 5) were conducted on each measure. While every measure yielded a main effect of group (all $p< 0.01$, see below), only *rearrangements* yielded a main effect of set size ($F(3,6)=4.11$ $p<0.01$). Thus, after accounting for relational complexity, we found no additional effect of memory load on patient performance on any measure except *rearrangements*.

Figure 2.1

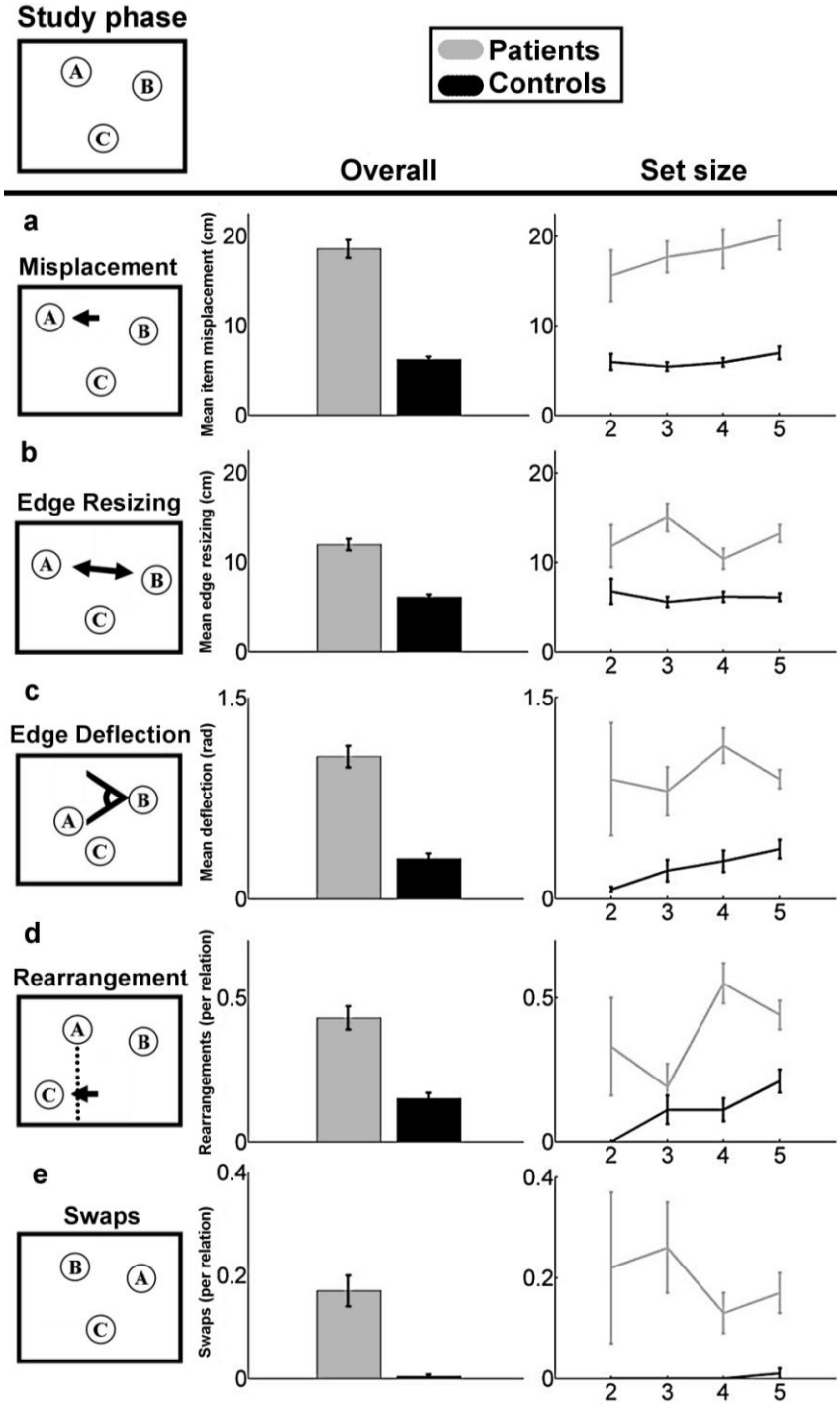


Figure 1. Patient and comparison participant performance quantified using five metrics of error quality. An example study configuration is provided at the top of the figure. (A-E) Each of the five reconstruction error types is demonstrated, with overall error (left, collapsed across the number of items in the study configuration) and error as a function of studied object set size (right) is provided for each error type. Error bars indicate SE.

Spatial Measures

We report mean performance on the *item misplacement* metric, collapsed across the number of objects that were studied (2, 3, 4, and 5), as well as the mean performance for each object set size (Fig. 2.1a). A mixed 2-by-4 ANOVA with factors of group and set size with yielded a main effect of Group ($F(1,6)=137.74$, $p<0.0001$), indicating reliably poorer placement among amnesics for all set sizes. The visual trend for more misplacement with increasing set sizes did not reach significance ($F(3,6)=1.47$, $p>0.22$), nor was there an interaction between the two factors ($F(3,6)=0.47$, $p>0.69$). On the *edge resizing* (Fig. 2.1b) and *edge deflection* (Fig. 2.1c) metrics that assess the reconstruction of the magnitude and direction of vectors between object pairs, we found similar effects. For edge resizing, a mixed 2-by-4 ANOVA with group and set size showed a significant main effect of group ($F(1,6)=48.45$, $p<0.0001$) but there was no significant effect of set size ($F(3,6)=1.49$, $p>0.21$), nor was there an interaction between the two factors ($F(3,6)=2.05$, $p>0.10$). Likewise for edge deflection, there was a significant main effect of group ($F(1,6)=36.66$, $p<0.0001$), a nonsignificant main effect of set size ($F(3,6)=1.64$, $p>0.17$), and nonsignificant interaction ($F(3,6)=1.3$, $p>0.27$). On the *rearrangements* metric (Fig. 2.1d), measuring performance at reconstructing an undistorted overall shape , showed a significant main effect of group ($F(1,6)=22.53$, $p<0.0001$), a significant main effect of set size ($F(3,6)=4.11$, $p<0.01$), and a significant interaction between the two factors ($F(3,6)=3.43$, $p<0.02$). Thus, amnesics were impaired overall, while both amnesics and comparisons made more rearrangement errors as set size increased, with this trend being disproportionately greater for amnesics.

These findings replicate previous reports of worse performance for amnesics versus comparison participants using item misplacement measures of reconstruction (Smith and Milner, 1981, Jeneson, Mauldin, and Squire, 2010). In addition, they demonstrate that amnesic participants are also impaired relative to comparisons at reconstructing the positions of objects relative to each other and are also less likely than comparisons to reconstruct an accurate version of the general shape they observed during the study phase.

Relational Measure

Errors made in reconstructing the relative positions of objects, resulting in objects "swapping" positions, are summarized (Fig. 2.1e). The *swaps* measure showed a significant main effect of

group ($F(1,6)=29.78$, $p<0.0001$), but no significant main effect of set size ($F(3,6)=0.96$, $p>0.4$) and no significant interaction between the two factors ($F(3,6)=1.01$, $p>0.3$). Thus patients were no less likely to swap a pair of items that appeared alone (set size 2) than they were to swap any pair of items that appeared in set sizes of three, four, or five. This finding shows that hippocampal amnesics are also more likely than comparisons to fail at binding item identities to locations.

Do any metrics show disproportionate impairment?

All five measures were highly inter-correlated, with R^2 between 0.95 and 0.69 (all $P<0.02$). This was expected since some reconstruction errors might lead to high values on more than one metric (e.g., a swapped item will also be misplaced relative to its original location). To identify if any of these error types disproportionately impaired to overall performance we examined the relative proportion of reconstruction errors committed by amnesic versus comparison participants. Relative to the other metrics, amnesics produced a strikingly disproportionate rate of errors for the swap metric (Fig. 2.2). For the other four metrics, amnesics performed between two and four times worse than comparisons. However, on the swap metric, amnesics were more than 40 times worse than comparisons. Amnesics made a swap error on 17% of pairwise relations (31 swaps in 182 pairs) whereas comparisons made the error on only 0.4% (1 swap in 242 pairs) of pairwise relations. There was only a single swap error made by any of the comparisons in any of the familiar-object conditions. By comparison, all three amnesic participants made numerous swap errors in the familiar object condition, with patient 2363 making an average of 0.17 swap errors per relation (10 swaps in 60 pairs), patient 1846 making 0.14 (11 swaps in 80 pairs), and patient 2308 making 0.25 (10 swaps in 40 pairs). For trials in which amnesic participants committed swap errors, 66% involved one swap error, 24% involved two errors, 5% involved three errors, and 5% involved four errors. Thus, more than one swap error in a single trial was a frequent occurrence for amnesics (34%), while never occurring in the performance of any of the comparison participants (0%). Additionally, when adjusted for number of relations amnesic participants made swap errors at an approximately equal rate for all set sizes while comparisons made the error only once and only in a 5-item set. Swap errors were thus an essentially unique identifier of amnesic participants.

Figure 2.2

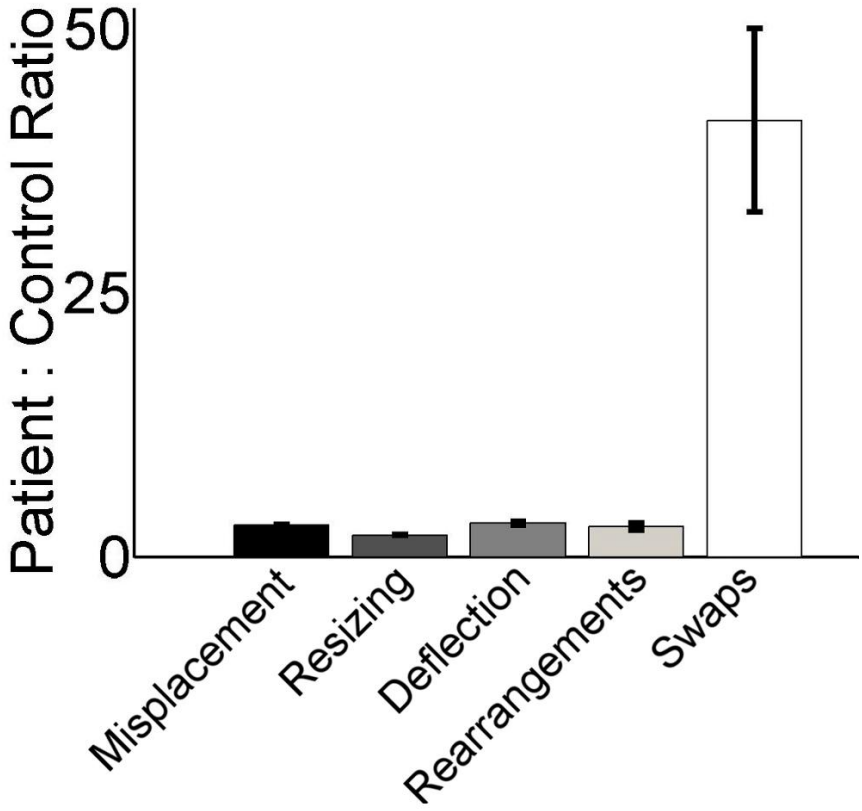


Figure 2. Disproportionately high swap errors in patients. The ratios of mean patient performance to mean comparison performance are provided for each of the five performance metrics. Error bars indicate SE, calculated by error propagation.

Are swap errors the primary deficit in amnesia?

Given the strikingly disproportionate prevalence of swap errors relative to the other error types, we next asked whether swap errors constitute the primary deficit in amnesia. In other words, what is the causal relation between swap errors and errors on the other metrics?

One possible explanation for the high incidence of swaps errors made by amnesic participants is simply that they made large misplacement errors. That is, what appeared to be swap errors actually could have resulted from misplacement errors wherein item locations were reconstructed so inaccurately that items were actually placed in another object's studied location. Thus, we examined how likely it would be for items to swap given the study-time distance between pairs of items, and the magnitude of patients' misplacement errors. We tested this with a Monte Carlo simulation that utilized the item misplacement values and the study-time inter-item distances collected from patients in our observed data. The simulation

randomly drew a pair of misplacements and an inter-item distance each iteration, controlling for set size. This had the effect of creating a new set of data in which pairs of objects moved randomly according to the distribution of the actual misplacement data. Each run of the simulation produced a number of data points equal to those observed in the real experiment, and we calculated the mean number of swaps present in the simulated data over 1,000 runs. For each run of the simulation, we performed a 1-way ANOVA on the observed versus the simulated data, and here report the mean incidence of swaps, and mean p values produced by this simulation (Fig. 2.3). The Monte Carlo simulation showed that based on item misplacements alone, patients should have made only 0.045 (SD = 0.015) swaps per pairwise relation, far less than the 0.17 swaps per relation actually observed. This meant that on average, the Monte-Carlo simulation produced approximately 8 (SD = 2.73) swaps, while the empirical data contained 31. This difference was significant: Given the observed level of misplacement error, the mean probability of observing the number of swaps in the actual data due to item misplacement alone was 0.007. Therefore, overall poor spatial positioning of items individually was not the cause of the high level of swap errors actually observed (i.e., general item misplacement was not primary to swap error).

Figure 2.3

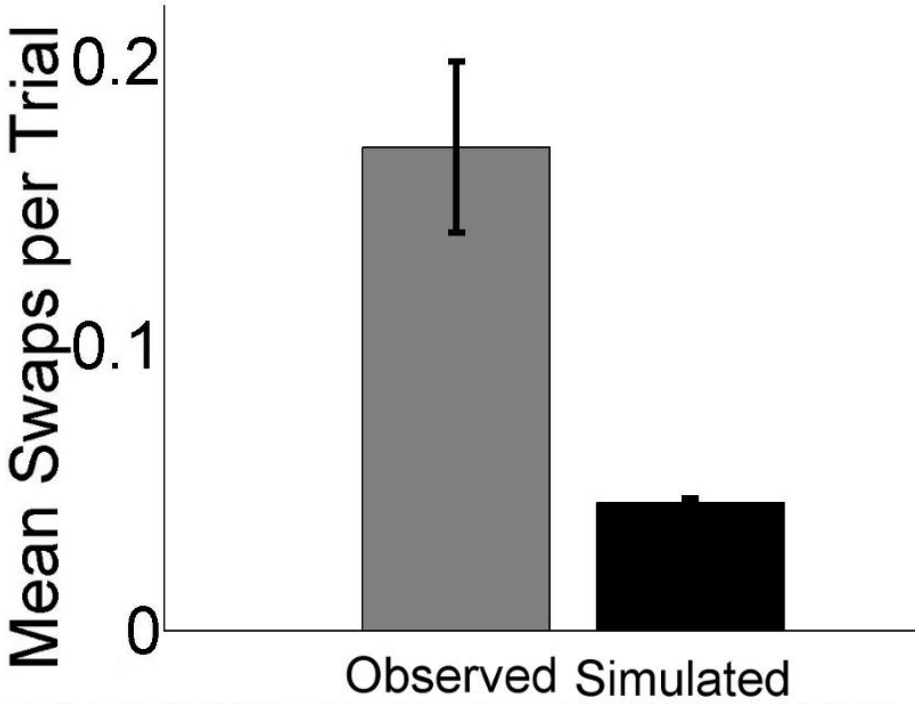


Figure 3. Swap errors in patients were more numerous than would be expected by chance. Mean actual swap errors per trial made by patients are plotted along with the number of errors expected based on chance given actual misplacement error, as determined by Monte-Carlo simulation.

Do swap errors contribute to amnesics' poor performance on item misplacement?

We next tested the opposite direction of causality (i.e., determining whether the high incidence of swap errors might be contributing to the overall poor spatial positioning performance of amnesic participants). We recalculated item misplacement after removing the error values introduced by swaps (i.e., calculating item misplacement only for un-swapped items). If swap errors were primary to item-placement errors, we would expect a significant reduction in simple spatial errors after removing the effects of swaps. This was confirmed (Fig. 2.4). Although amnesic participants still performed worse than comparisons after removal of item misplacement due to swaps (main effect of group, $F(1,3)=5.15$, $p<0.03$), removing swaps led to a significant reduction in item misplacement (one way ANOVA with swaps-present vs. swaps-removed, $F(1,3)=31.49$, $p<0.0001$). This suggests that the amnesic participants' deficits on the standard item-misplacement measure can be at least partially attributed to their poor performance on the swap metric. We also used an "unswapping" algorithm (see Appendix A) to determine whether poor misplacement performance led to swaps, the results of which converge with the above analysis showing that improvements in item misplacement are not simply due to discarding poor trials.

Figure 2.4

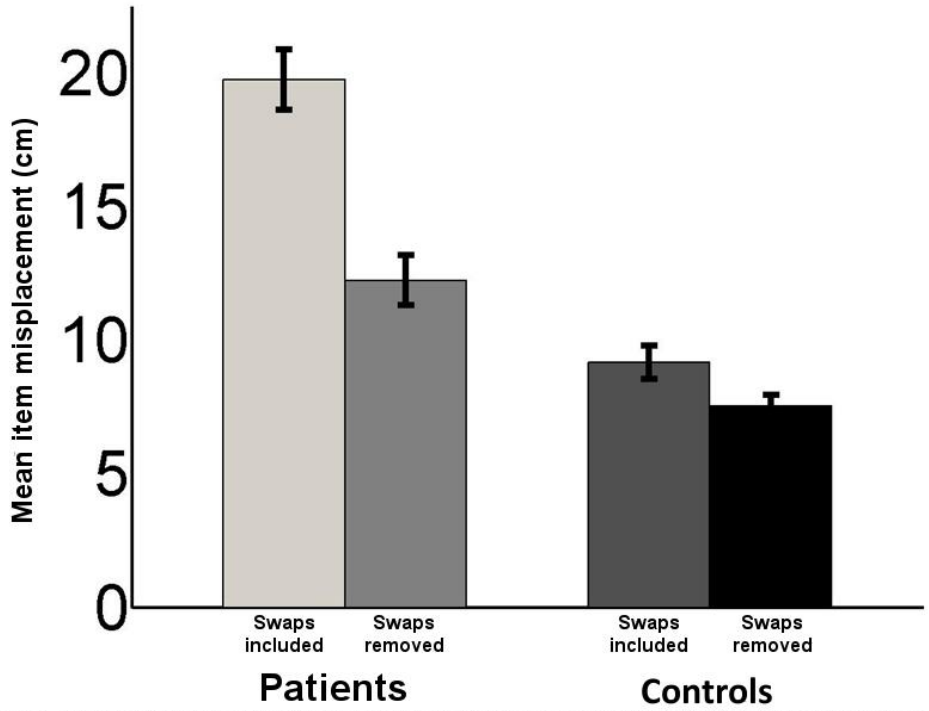


Figure 4. Swap errors were partially responsible for patient misplacement errors. Misplacement error values for patients are shown averaged for all objects in each trial as well as only for the objects for which swap errors did not occur. Removing swap errors in this manner led to a significant reduction in misplacement error for patients. Error bars indicate SE.

Do patients simply perform randomly?

It is illustrative to compare participants' performance across the measures to an objective benchmark: random performance. Supposing neither patients nor comparisons were allowed to see the study phase, but were still able to place objects in random positions. Their performance could only be randomly related to the experimenter-positioned objects (based on the premise that both the experimenter and participants are equally likely to place any object at any coordinate on a 100x100 cm grid, see Methods).

On all measures except swaps, both patients and comparisons perform far above chance across all set sizes (the 95% confidence interval does not include chance), demonstrating that, in an information processing sense, they possessed some useful information about the original configuration. Yet on swaps, comparisons' performance for familiar objects was nearly perfect: the 95% confidence interval included 0%. For amnesics, swap performance was nearly at chance: the 95% confidence interval included 25%. This quantitative difference cannot be much larger. For reconstructing the object-identity-to-vertex-position bindings, comparisons behaved as if they had nearly perfect information, while patients behaved as if they had nearly no information about such bindings.

Do novel-object displays have a disproportionate impact on patients?

The same reconstruction was also performed using a set of 14 novel objects, composed of white foam blocks carved into various complex shapes and covered in patterns of simple lines and other shapes (James, Shima, Tarr, Gauthier 2005).

We performed the same series of analyses (i.e., misplacement, edge resizing, edge deflection, rearrangements, and swaps) of reconstruction performance for arrays composed of novel objects (Fig.2.5), as well as an ANOVA for each measure with a factor of item type (familiar v. novel). Here we describe the main findings averaged across all set sizes to facilitate comparisons with the effects identified using familiar objects.

Figure 2.5

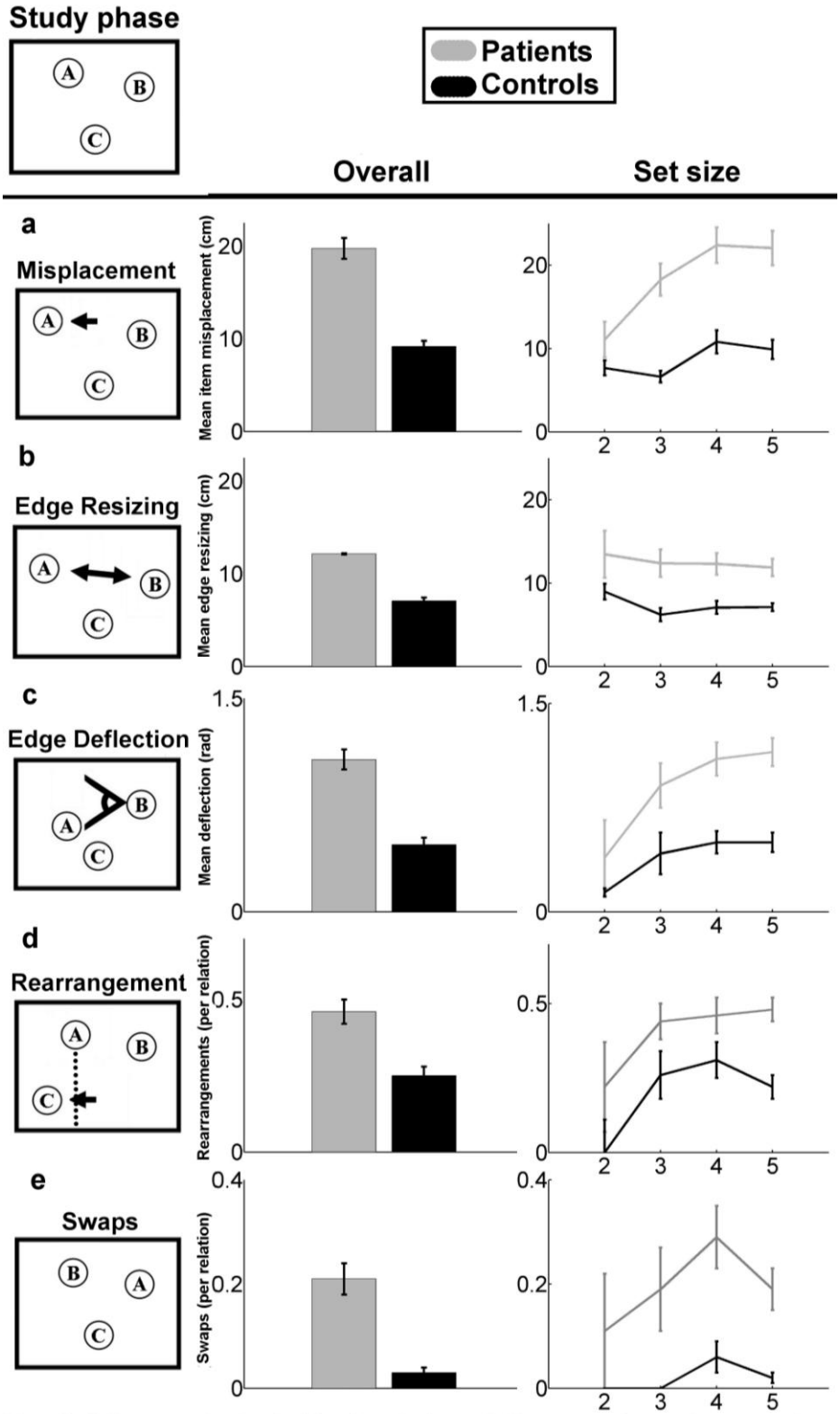


Figure 5. Patients were also impaired for all metrics in novel-object arrays. Once again, an example study configuration is provided at the top of the figure (A-E) Each of the five reconstruction error types is demonstrated, and overall error (collapsed across the number of items in the study configuration) and error as a function of studied object set size is provided for each type. Error bars indicate SE.

Performance was significantly worse for novel objects relative to familiar objects for three of the five metrics, as indicated by significant main effects of the object-type factor (for the item-misplacement, edge-deflection, and rearrangement metrics, ($F(1,6)=7.48$, 6.6, and 4.932, respectively, p's < 0.01, 0.02, and 0.04), although not for the edge-resizing ($F(1,6)=0.3$, $p=0.60$) and swap ($F(1,6)=3.25$, $p=0.72$) metrics). As was the case for familiar objects, for novel objects amnesic participants showed worse performance than comparison participants for all five metrics, as indicated by significant main effects of group (for the item-misplacement, edge-resizing, edge-deflection, rearrangement, and swap metrics, ($F(1,6)=220.2$, 136.7, 129.9, 60.3, and 84.2, respectively, all p's < 0.001). However, there was no evidence that novel objects impaired performance to a greater extent in amnesic than in comparison participants for any of the metrics; the interactions of object-type by group were all nonsignificant (p's between 0.15 and 0.65).

As for familiar objects, a striking characteristic of amnesic participants' performance with novel objects was the highly disproportionate incidence of swap errors compared to all other error types. Relative to comparisons, amnesic participants' performance was 2.2 times worse for item misplacement, 1.7 times worse for edge resizing, 2.3 times worse for edge deflection, and 1.75 times worse for rearrangement, but 7.3 times worse for swaps. Amnesics made a swap error on 21.4% of the opportunities they had to do so (39 swaps in 182 pairs), whereas comparisons made the error on only 3% of their opportunities (7 swaps in 249 pairs).All three amnesic participants made numerous swap errors, with patient 2363 making an average of 0.18 swap errors per pairwise relation (11 swaps in 60 pairs), patient 1846 making 0.28 (22 swaps in 80 pairs), and patient 2308 making 0.15 (6 swaps in 40 pairs). For trials on which any swap error occurred, amnesics frequently made multiple swaps: 54% involved one swap error, but 19% involved two errors, 11% involved three errors, 5% involved four errors, and 11% involved five errors. By contrast, multiple swaps were much less frequent in comparison participants: 85% involved only one swap error and the other 15% involved only two. As was the case for familiar objects, amnesics made swap errors for all set sizes whereas comparisons made these errors only for 4- and 5-item sets.

In addition, since comparison participants were no longer at "ceiling" on the swap metric, we once again examined patient to comparison relative performance across the five metrics (Fig. 2.6). Once again, the swap measure showed the most disproportionate deficit, with patients

performing 7.2 times worse than comparisons, with relative performance on the other four measures between 1.7 and 2.3.

Figure 2.6

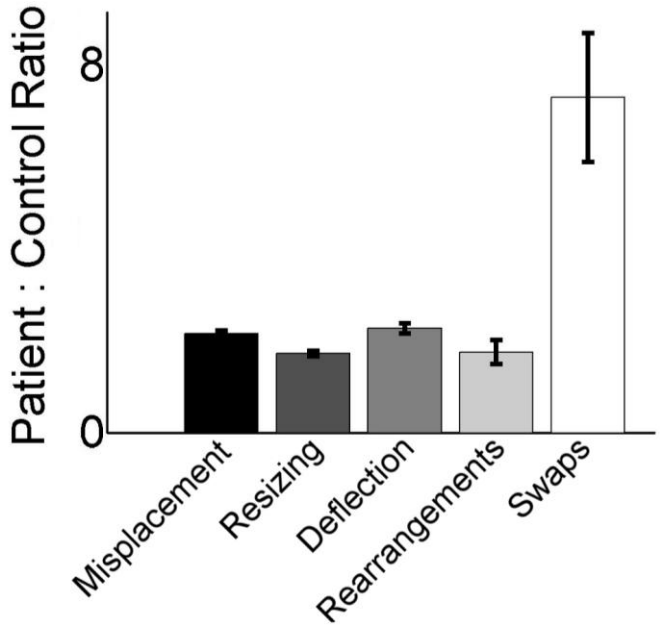


Figure 6. Disproportinately high swaps in patents in the novel object condition despite non-ceiling comparison performance. The ratios of mean patient performance to mean comparison performance are provided for each of the five performance metrics. Error bars indicate SE calculated by error propagation.

Finally, Monte Carlo simulations showed the same direction of causality for the various error types for novel objects as was observed with familiar objects. The prevalence of swap errors was significantly higher than would have been expected by chance if they were due entirely to pure item-misplacement error (Mean Simulated Swaps per relation=0.05, mean $p< 0.01$). Furthermore, removal of all item-misplacement error due to swap error led to a significant reduction in item-misplacement error for both groups ($F(1,6)=216.31$, $P<0.001$), and a disproportionately greater reduction in error for amnesics. Thus, for novel objects as for familiar ones, swap errors cause an over-estimate of the memory deficits suggested by item-misplacement errors.

Discussion

Individuals with hippocampal amnesia displayed impaired performance in reconstruction of spatial locations of small arrays of objects over a short delay interval. Impairments were present both in the standard measure in such paradigms, involving the degree of item misplacements (mean distance between objects' position at study versus reconstructed position as placed by the participant at test), and for all other metrics we used to examine aspects of the memory

representations needed to support reconstruction of object locations. Strikingly, the observed deficit was markedly disproportionate for errors involving object-for-object swapping, which evaluated object-identity-to-relative-location bindings. For arrays of familiar objects, amnesic participants committed swap errors at a rate more than 40 times that of comparison participants, who almost never committed this kind of error; making swaps nearly a unique identifier of amnesia in our sample.

The high incidence of swap errors in amnesia was shown via simulation analysis not to arise from larger-than-normal item misplacement. Instead, the causal relationship between these error types was in the other direction. Findings showed that removal of swap errors from the analysis led to a significant reduction in the estimates of item misplacement error, suggesting that a significant proportion of the overall poor performance resulted from the inability to track object-identity-to-relative-location bindings. Notably, the prevalence of swap errors in the performance of participants with hippocampal amnesia was seen for two independent stimulus categories (familiar objects and novel objects), and held across all set sizes.

One possible explanation of the pattern of performance across the various metrics is that both patients and comparisons were able to represent the object arrays as simple "shapes" formed by perceptual closure (with each object corresponding to a vertex of the shape c.f., Uttal, and Chiong 2004), and/or as a motoric sequence indexing each location within the array (as in the Corsi block tapping task, which is partially spared in patients with hippocampal damage c.f. Corsi 1972, Kessels et al. 2000). However, whatever spared representation underlies patient performance; it seems to lack the cross-domain binding information about which items occupy which spatial indices. Thus, comparison participants were highly successful both at reconstructing the array outline and at placing each object at the specific vertex position at which it was studied, as demonstrated by their relatively successful reconstruction performance measured using all metrics. Amnesic participants were somewhat less successful at reconstructing the geometry of object arrays (as indicated by their deficits in our first four edge metrics), but showed strikingly disproportionate impairment in representing the arbitrary object-to-vertex mappings required to replace the correct objects in their specific vertex positions, instead swapping object-to-vertex relations. These swap errors were nearly

diagnostic of hippocampal amnesia; in our sample if a swap error was observed on a familiar object trial, it was 97% likely that a patient was positioning the objects.

This is best illustrated by one especially striking feature of patient performance: the presence of swap errors by patients on two-object trials. Even in our task's simplest condition, with only single binding between a single pair of familiar objects, requiring maintenance for a few seconds, patients still reversed the positions of the two objects approximately once every five opportunities—a rate of swapping similar to that which would be produced without any knowledge of the proper arrangement of objects. Intuitively, and as observed in the performance of healthy comparisons, errors of this kind should be vanishingly rare in neurologically intact participants.

The deficits observed here were for brief retention intervals and short lags traditionally associated with working memory. This is consistent with other findings of relational memory deficits in amnesia at short retention intervals (Ryan and Cohen, 2004, Hannula, Tranel, and Cohen, 2006, Hannula, Ryan, Tranel, and Cohen, 2007). and also with recent evidence that the human hippocampus is essential for the expression of memory even with no interposed study-test delay (e.g., when memory is used online to guide exploration behavior, as in Voss et al., 20011a, Voss, et al., 20011b, or to assemble and maintain complex representation as in Warren et al., 2012). It also converges with fMRI findings of hippocampal activation for relational memory over the same very short timescale (Mitchell, Johnson, Raye, and D'Esposito, 2000, Piekema et al., 2006, Hannula and Ranganath, 2009) as well as with imaging data implicating the hippocampus more generally on the short-term/working memory timescale (Ranganath and D'Esposito, 2001, Stern, Sherman, Kirchoff, and Hasselmo, 2001, Ranganath and Blumenfeld, 2005).

One of our goals in using a simple but open-ended test complemented by a suite of performance metrics was to test different theoretical accounts of hippocampal deficits. Our analysis provides the strongest support for theories that emphasize arbitrary relational bindings as the primary hippocampal representation (indexed by our swaps metric). Because removal of swap errors did not entirely ameliorate patients' reconstruction deficits, our analysis also provides partial support for theories that emphasize geometric, spatial, hippocampal

representations (at least where these spatial representations correspond to simple Cartesian coordinate maps, the vectors and landmarks of a polar coordinate representation, or unitized, shapes formed by perceptual/motor processes), as indexed by our misplacement, edge resizing, edge deflection, and rearrangement metrics. However, we were able to explain the preponderance of these spatial deficits as the secondary consequences of swap errors. While the remaining deficit does partly support spatial theories of hippocampal function, it may also be a consequence of our pair-wise swap measure which counts only two dimensional rotations of inter-item vectors in a continuous, Cartesian space. Swap measures that take into account multi-item swaps (e.g., rotation of trios of objects), categorical swaps (e.g., swapping the left and right halves of a figure), or non-spatial swaps (e.g., perseveration of a previously reconstructed shape) could perhaps account for some additional portion of the deficit and are an intriguing avenue of future study. Finally, we found little evidence that hippocampal amnesia was best explained by a deficit in transferring information from a limited capacity working-memory system to long-term memory (indeed patients made swap errors over short delays even in the two object condition, and the error rate did not increase for larger set sizes after accounting for the increase in relational complexity). We would explain findings that amnesics make disproportionally greater misplacement errors on arrays with large numbers of items (Jeneson, Mauldin, and Squire, 2010) as a natural consequence of linear growth in item counts producing combinatorially more opportunities to commit swap errors.

However, our goal was not to adjudicate competing theories, but to identify the primary memory deficit resulting from hippocampal damage. Since the experiment examines only spatial reconstruction, it is impossible to infer if swap errors are a consequence of a deficit specific to item-identity-to-location bindings (c.f., Lee et al. 2005, Hartley et al. 2006), or if they arise from a deficit in a more domain-general binding system. Our findings are compatible with any theory that proposes that the hippocampal is critical to performance that relies upon flexible, reconfigurable bindings that index locations, and that it is the disruption of such bindings that causes generally poor spatial performance. Our assertion that these findings more strongly support a representation scheme based upon arbitrary relational binding than a scheme that emphasizes spatial relations arises from the fact that spatial representations often carry connotations of geometric properties such as coordinates, distances, angles, and shapes which, according to the measures reported here, were less disrupted by hippocampal than item-

identity-to-location-bindings. Four of the metrics (edge resizing, edge deflection, rearrangement, and swaps) are spatial relational measures, but only the relations supporting performance on the swap metric are disproportionately impaired by hippocampal damage. The spatial reconstruction paradigm we used provides rich behavioral records, and avoids many confounds between relational and item memory (e.g., test format, test difficulty, retention interval, etc.). We showed that while the hippocampus is certainly involved in spatial representations of all types, there is one kind of representation (which binds item identities to their relative locations) for which an intact hippocampus is the difference between chance level performance and perfect performance. Furthermore, this binding deficit is primary to more traditionally measured spatial reconstruction impairments. The precision of our result supports a similarly precise theoretical account that emphasizes that general memory impairments result from specific binding deficits. Rather than simply highlighting tasks which are impaired by hippocampal damage, theoretical accounts should be able to explain why a measure uniquely sensitive to binding errors (e.g., our swap metric) is most indicative of hippocampal impairment, and why representations of arbitrarily assigned, reconfigurable, item-identity-to-relative-location-bindings so critically depend on hippocampal function.

References

Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *The Behavioral and brain sciences*, *22*(3), 425–44; discussion 444–89.

Allen JS, Tranel D, Bruss J, Damasio H (2006) Correlations between regional brain volumes and memory performance in anoxia. J Clin Exp Neuropsychol 28:457-476.

Cavaco, S., Feinstein, J. S., Van Twillert, H., & Tranel, D. (2012). Musical memory in a patient with severe anterograde amnesia. *Journal of clinical and experimental neuropsychology*, *34*(10), 1089–100. doi:10.1080/13803395.2012.728568

Cohen N, Eichenbaum H (1993) Memory, amnesia, and the hippocampal system. Cambridge: MIT Press.

Cohen, N., & Squire, L. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science*.

Corsi, P. (1972). Human memory and the medial temporal region of the brain.

Damasio, A. R. (1989). The Brain Binds Entities and Events by Multiregional Activation from Convergence Zones. *Neural Computation*, *1*(1), 123–132.

Eichenbaum H, Cohen N (2001) From conditioning to conscious recollection: Memory systems of the brain. New York: Oxford University Press.

Eichenbaum, H., Yonelinas, a P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual review of neuroscience*, *30*, 123–52.

Gold JJ, Smith CN, Bayley PJ, Shrager Y, Brewer JB, Stark CE, Hopkins RO, Squire LR (2006) Item mem ory, source memory, and the medial temporal lobe: concordant findings from fMRI and memory-impaired patients. Proc Natl Acad Sci U S A 103:9351-9356.

Hannula DE, Ryan JD, Tranel D, Cohen NJ (2007) Rapid onset relational memory effects are evident in eye movement behavior, but not in hippocampal amnesia. J Cogn Neurosci 19:1690-1705.

Hannula DE, Tranel D, Cohen NJ (2006) The long and the short of it: Relational memory impairments in amnesia, even at short lags. J Neurosci 26:8352-8359.

Hannula, D.E., & Ranganath, C. (2009). The eyes have it: hippocampal activity predicts expression of memory in eye movements. *Neuron*, *63*(5), 592–599.

Hartley, T., Bird, C. M., Chan, D., Cipolotti, L., Husain, M., Vargha-Khadem, F. and Burgess, N. (2007), The hippocampus is required for short-term topographical memory in humans. Hippocampus, 17: 34–48. doi: 10.1002/hipo.20240

Hayes, S. M., Ryan, L., Schnyer, D. M., & Nadel, L. (2004). An fMRI study of episodic memory: retrieval of object, spatial, and temporal information. *Behavioral neuroscience*, *118*(5), 885-96.

Henke, K. (2010). A model for memory systems based on processing modes rather than consciousness. *Nature reviews. Neuroscience*, *11*(7), 523-32.

Huttenlocher J, Presson C (1979) The coding and transformation of spatial information. Cogn Psychol 11:375-394.

James TW, Shima DW, Tarr MJ, Gauthier I (2005) Generating complex three-dimensional stimuli (Greebles) for haptic expertise training. Behavior Research Methods, Instruments, and Computers, 37(2):353-8

Jeneson A, Mauldin KN, Squire LR (2010) Intact working memory for relational information after medial temporal lobe damage. J Neurosci 30:13624-13629.

Kessels, R. P., Van Zandvoort, M. J., Postma, a, Kappelle, L. J., & De Haan, E. H. (2000). The Corsi Block-Tapping Task: standardization and normative data. *Applied neuropsychology*, *7*(4), 252–8.

Konkel A, Cohen NJ (2009) Relational memory and the hippocampus: Representations and methods. Front Neurosci 3:166-174.

Konkel A, Warren DE, Duff MC, Tranel DN, Cohen NJ (2008) Hippocampal amnesia impairs all manner of relational memory. Front Hum Neurosci 2:15.

Konkel A, Warren DE, Duff MC, Tranel DN, Cohen NJ (2008) Hippocampal amnesia impairs all manner of relational memory. Front Hum Neurosci 2:15.

Lee, A. C.H., Buckley, M. J., Pegman, S. J., Spiers, H., Scahill, V. L., Gaffan, D., Bussey, T. J., Davies, R. R., Kapur, N., Hodges, J. R. and Graham, K. S. (2005), Specialization in the medial temporal lobe for processing of objects and scenes. Hippocampus, 15: 782–797. doi: 10.1002/hipo.20101

Marr, D. (1971). Simple Memory: A Theory for Archicortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *262*(841), 23–81.

McClelland, JL, McNaughton, BL, & O'Reilly, RC (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*(3), 419-57.

Mitchell KJ, Johnson MK, Raye CL, D'Esposito M (2000) fMRI evidence of age-related hippocampal dysfunction in feature binding in working memory. Brain Res Cogn Brain Res 10:197-206.

Nadel, L., Samsonovich, a, Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. *Hippocampus*, *10*(4), 352-68.

O'Keefe, J, Nadel L, (1978) The hippocampus as a cognitive map. Oxford University Press.

Piekema C, Kessels RP, Mars RB, Petersson KM, Fernandez G (2006) The right hippocampus participates in short-term memory maintenance of object-location associations. Neuroimage 33:374-382.

Ranganath C, Blumenfeld RS (2005) Doubts about double dissociations between short- and long-term memory. Trends Cogn Sci 9:374-380.

Ranganath C, D'Esposito M (2001) Medial temporal lobe activity associated with active maintenance of novel information. Neuron 31:865-873.

Ryan JD, Cohen NJ (2004) Processing and short-term retention of relational information in amnesia. Neuropsychologia 42:497-511.

Ryan, L., Lin, C.-Y., Ketcham, K., & Nadel, L. (2010). The role of medial temporal lobe in retrieving spatial and nonspatial relations from episodic and semantic memory. *Hippocampus*, *20*(1), 11-8.

Schacter, D. L.; Tulving, E. (1994) Memory Systems. First Edition, MIT Press.

Smith ML, Milner B (1981) The role of the right hippocampus in the recall of spatial location. Neuropsychologia 19:781-793.

Squire LR, Stark CE, Clark RE (2004) The medial temporal lobe. Annu Rev Neurosci 27:279-306.

Stark CE, Squire LR (2003) Hippocampal damage equally impairs memory for single items and memory for conjunctions. Hippocampus 13:281-292.

Stern CE, Sherman SJ, Kirchhoff BA, Hasselmo ME (2001) Medial temporal and prefrontal contributions to working memory tasks with novel and familiar stimuli. Hippocampus 11:337-346.

Uttal DH, Chiong C (2004) Seeing space in more than one way: Children's use of higher order patterns in spatial memory and cognition. In: Human spatial memory: Remembering where (Allen GL, ed), pp 125-142. Mahwah: Lawrence Erlbaum.

Vargha-Khadem, F., Gadian, D., Watkins, K., Connelly, A., Van Paesschen, W., & Mishkin, M. (1997). Differential Effects of Early Hippocampal Pathology on Episodic and Semantic Memory. *Science*, *277*(5324), 376–380.

Voss JL, Gonsalves BD, Federmeier KD, Tranel D, Cohen NJ (2011a) Hippocampal brain-network coordination during volitional exploratory behavior enhances learning. Nat Neurosci 14:115-120.

Voss JL, Warren DE, Gonsalves BD, Federmeier KD, Tranel D, Cohen NJ (2011b) Spontaneous revisitation during visual exploration as a link among strategic behavior, learning, and the hippocampus. Proc Natl Acad Sci U S A.

Warren, D. E., Duff, M. C., Jensen, U., Tranel, D., & Cohen, N. J. (2012). Hiding in plain view: Lesions of the medial temporal lobe impair online representation. *Hippocampus*. doi:10.1002/hipo.21000

Wechsler, D (1997a) Weschler adult intelligence Scale--Third edition, New York, New York: The Psychological Corporation.

Wechsler, D (1997a) Weschler memory Scale--Third edition, New York, New York: The Psychological Corporation.

EVENT RECONSTRUCTION REVEALS THE INTERDEPENDENCE OF EPISODIC AND SEMANTIC MEMORIES

Contributors
Watson, P., Cohen, N.

This paper was prepared for submission to Psychological Review

Abstract

Binding together complex configurations of stimulus relations supports all kinds of mental construction, including remembering, imagining, reasoning, and prediction (Hassabis & Maguire, 2009a). However, the diverse properties of the stimuli used in studies of mental (re)construction often produces asymmetries in difficulty across stimulus categories, for example, pitting memory for small numbers of spatial locations against memory for large numbers of objects that occupy those locations (Kumaran & Maguire, 2005; Ryan, Lin, Ketcham, & Nadel, 2010). This mismatch in memory requirements is a confound for memory theories suggesting that increased hippocampal involvement in domain-specific relations may be due to the increased relational binding demands of more complex or arbitrary tasks, rather than domain-specific hippocampal processing preferences. In this paper we present the Event Reconstruction Technique (ERT), a method for comparing arbitrary relational binding of stimuli across domains and complexity levels. Among healthy college-aged participants, we found 1) that behavioral performance was best predicted by the complexity of the binding at hand; 2) once this variation was accounted there was no significant effect of domain (spatial or temporal) on performance. When different types of choice complexity were constrained by non-arbitrary rules participants were able to leverage these rules to dramatically increase their performance above chance. Finally, participants made errors that revealed a hierarchically organized memory for the stimuli. For example if participants switched the locations of pair of scenes, they would often "import" faces which were paired with those scenes during study to their new, incorrectly reconstructed location.

Introduction

"Going to the feelies this evening Henry? Enquired the Assistant Predestinator. "I hear the new one at the Alhambra is first-rate. There's a love scene on a bearskin rug; they say it's marvelous. Every hair of the bear reproduced. The most amazing tactile effects."
-Brave New World (3.42), Aldous Huxley

Our everyday experience of episodic remembering is a rich reconstructive process that brings together past visual, auditory, tactile, olfactory, and gustatory experiences to produce vivid internal scenes not unlike those of Huxley's "feelies." For particularly detailed or salient memories, this reconstruction can feel almost as if we had returned to the moment, via "mental time travel" (Tulving, 2002). This mental construction system has recently garnered increased research interest to help understand how perceptual, memory, and cognitive systems interact to produce rich, internal, representations (Hassabis, Kumaran, Vann, & Maguire, 2007; Hassabis & Maguire, 2009)

However, this involvement of multiple brain and memory systems makes mental reconstruction particularly difficult to study. Reconstruction requires configuring many different dimensions of information (items, space, time), which may come from different sensory modalities. This process is dynamic and interactive (Voss, Gonsalves, Federmeier, Tranel, & Cohen, 2011),with constant updates and edits to the reconstruction. Of particular difficultly, is finding a measure that can compare the accuracy of a mental reconstruction to the original event (Koriat, Goldsmith, & Pansky, 2000), since the reconstruction often involves qualitatively different kinds of information it is unclear how to compare performance, especially if some domains are more complex than others.

For example, in Watson et al. (2013), patients with hippocampal damage were far more likely to commit "high level" reconstruction errors (e.g., switching the relative positions of object pairs), than "low level" errors (e.g., shifting an object relative to its studied position) as compared to control participants. However, because this spatial reconstruction was analogue, it is difficult to compare the relative difficulty of these different aspects of the task. It is clear that neither patients nor controls had a representation fine-grained enough to replace each item at the exact 1cm by 1cm point where it was originally studied, but it is unclear at what level participants

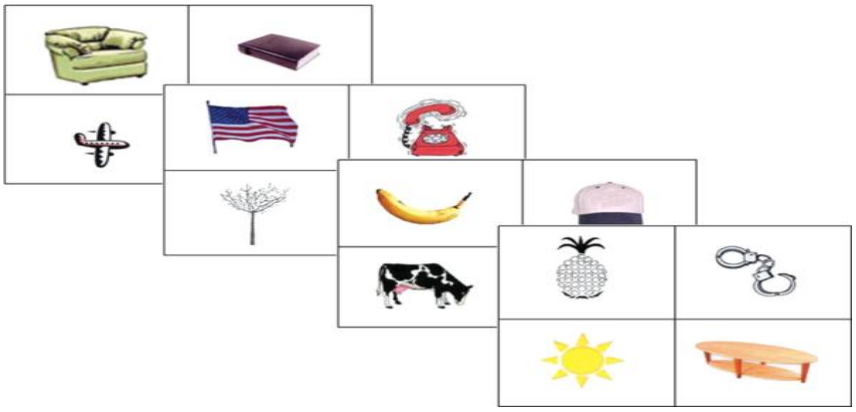
were “chunking” space. Without knowing how many elements were in play, it’s impossible to tell how *complex* (that is, how alternative configurations it is possible to construct) the bindings required to reconstruct the configuration were. Since the task requires choosing the correct reconstruction from a vast family of well-formed but incorrect alternatives, higher levels of complexity are simply more difficult, and thus require more reconstructive memory.

In addition, even when the level of complexity is controlled for, different aspects of a task may vary in how *arbitrarily* the different components of the stimulus are related, that is, given partial information about the stimulus to what degree could one predict the missing information of the stimulus. For example, Konkel et al., (2008), observed a decreasing gradient of hippocampal dependence for item, spatial, temporal, and re-pair conditions despite each of these conditions involving the same number of elements and relations at encoding time. However, in the “item” condition, two familiar objects were paired with a novel object. Correct rejection could be accomplished by noticing the novel object or by correctly recalling the “missing” familiar object via its relationship with one of two familiar objects. Since memory constrains the different conditions in different ways, each will require a different amount of reconstruction, and thus require different contributions of hippocampal processing.

However, many experiments assume the null hypothesis: that different levels of complexity and arbitrariness at reconstruction ought to result in the *same* level of reconstructive difficulty and hippocampal involvement. A more constrained recognition memory study (Ryan et. al 2010), compared spatial, temporal, object, and semantic performance within a single common experimental framework. Participants studied object arrays and later answered questions about the locations, order, and properties of items in the arrays. However, while items were trial-unique, the same four spatial locations, were used in each block, and there were six sequential arrays (six “moments”). Thus, while each item uniquely indexed a particular location in space and moment in time, each location was mapped to six different objects, and each moment to four (Figure 3.1). This confounds memory strength with complexity and arbitrariness. Since there are six paired associates with the upper-left hand location, participants have a better chance of selecting an item that appeared in the upper-left than of picking an item that appeared in the first array (1/4 v. 1/6), even if they remember nothing about the arrays. But for precisely the same reason, if participants have a poorer chance of identifying “the item that

appeared in the upper-left first" (drawing the correct item out of a set of six), than they do of identifying the "upper-left-most-item in the first array" (drawing the correct item out of a set of four). It is difficult to say then, if a variation in performance or the involvement of different brain regions in such an experiment is due to an associative strength or interference.

Figure 3.1: Mismatches in the complexity of item, spatial, and temporal information



Different numbers of object, spatial, and temporal elements argue for different levels of performance, and differential involvement of different brain systems. Without controlling for these differences in complexity, any observed differences might simply be due to predictably different baselines in different conditions (Adapted from Ryan et. al 2010).

This article presents an original, interactive, reconstructive experiment that attempts to account for the richness of memory reconstructions, while at the same time controlling for different levels of associative strength and confusability to allow an apples-to-apples comparison of performance across different levels of multi-modal stimulus complexity. Our goal was to determine which type(s) of information participants used to guide memory reconstruction, and how that information was organized.

Participants viewed a series of short movie clips and reconstructed the relations between the people, places, locations, and time present therein. By analyzing the number of free parameters (i.e., the number of ways it was possible to configure each set of relations), we were able to obtain expected performance due to chance in each category of relations, and by comparing these expected levels of performance to those actually produced by our participants we were thus able to determine what information participants were using to guide their reconstruction. By analyzing systematic patterns of errors, we are able to construct a set of dependencies that permitted us to determine which types of information were the primary organizers of reconstruction, and which types were subvenient classes.

Methods

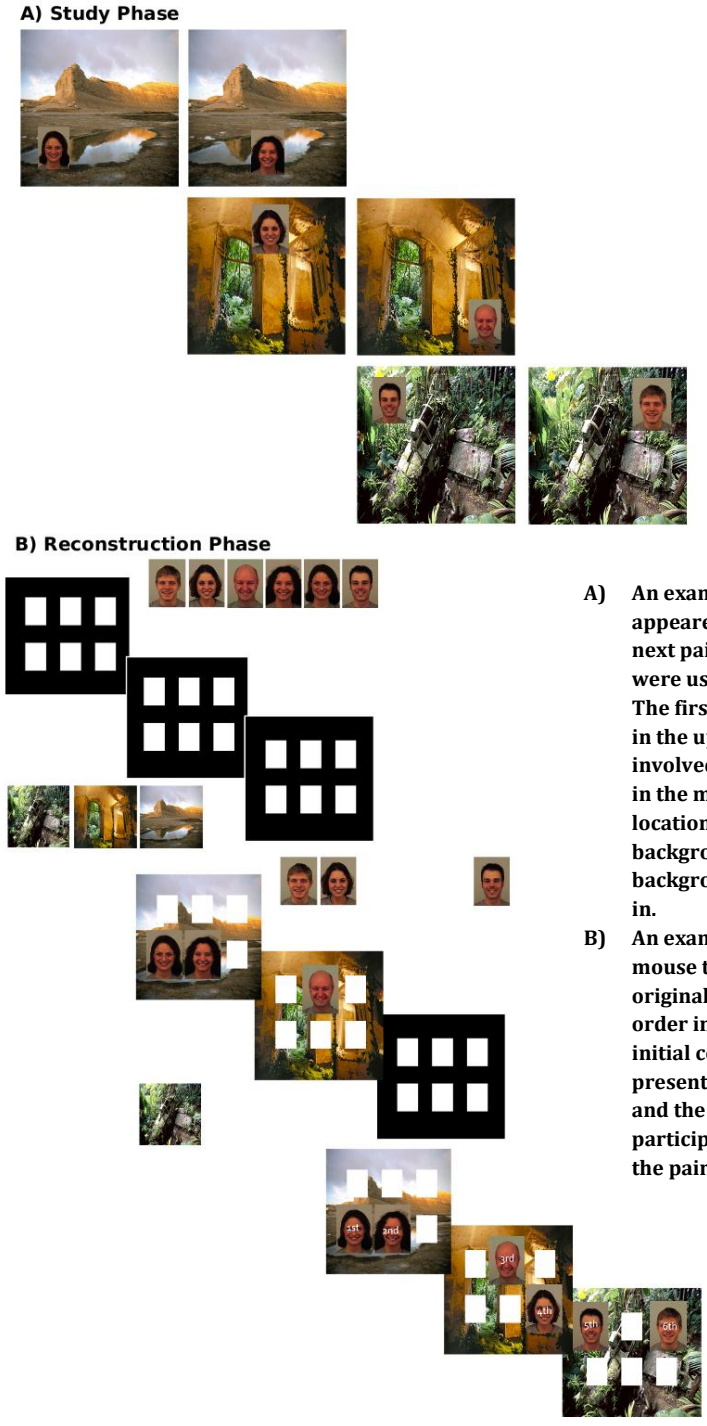
Participants

Behavioral data were collected from 26 college aged participants (14 females and 12 males). 16 of these participants were paid, while 10 received class credit. One participant was excluded for failing to complete the experiment.

Experimental paradigm

Participants performed a complex reconstruction task (Figure 3.2). During the "study" phase of the task participants viewed a brief (6s) movie clip consisting of a set of 6 still pictures, each involving a face superimposed on a scene. There were six unique faces and three unique scenes. Each face appeared at one of six possible "sockets" within a scene and each scene appeared at one of three possible "venues" on the screen. While the faces could appear at any of the sockets at any time, the scenes appeared sequentially with the first one always appearing in the upper left hand venue, the second always in the middle venue, and the last always in the lower right hand venue. During the "test" phase of the task participants reconstructed the configuration of stimuli they had originally seen by dragging each face and each scene to the remembered "sockets" and "venues" respectively. Once all of the stimuli had been placed spatially, participants labeled each face with the time it appeared (1^{st}, 2^{nd} etc.). Participants were not timed, and were free to re-order the faces, scenes, and times to their satisfaction before moving on to the next block. There were a total of 50 study/test blocks, interleaved with pleasant images (pictures of kittens) as a brief break between blocks. Before beginning participants were extensively briefed on the form and nature of the experiment, and given the opportunity to attempt two un-scored practice trials.

Figure 3.2: An example trial



A) **An example study phase. Each face-background pair appeared for 2s and was immediately followed by the next pair. The same faces, backgrounds, and locations were used in all trials, only the configuration varied. The first two face-background pairs always appeared in the upper left quadrant of the screen, and always involved the same background. The second appeared in the middle and the last in the lower right. The locations and order of the faces within the backgrounds varied, as did the mapping of each background to the "venue" on the screen it appeared in.**

B) **An example reconstruction. Participants used the mouse to drag the faces and backgrounds to their original locations. And then labeled them with the order in which they appeared. The figure shows the initial configuration of faces and backgrounds presented to the participants, a partial reconstruction, and the completed reconstruction. Note the error the participant made swapping the order and locations of the pair of faces in the second background.**

Stimuli

Stimuli were divided into scenes and faces. There were three unique scenes, and six unique faces which were reused in all 50 blocks of the experiment, differing only in their relational configuration. The scenes (Brand X Photography) were visually complex, and included two outdoor and one indoor scene. The faces included three male faces and three female faces all white and all in neutral or smiling expressions (Althoff & Cohen, 1999).

There were three unique "venues" into which scene could be placed, which were occupied sequentially from upper left to lower right. Any scene could appear in any venue on a given trial. There were six unique sockets into which faces could be placed, each of these locations was associated with each of the scene locations meaning there were a total of 18 possible sockets into which faces could be placed. However, faces were constrained to appear only once per socket, and only two locations would be occupied by faces per venue. The face/scene distinction mirrored the socket/venue distinction: 6 small objects (faces) were placed in six small locations (venues), while three large objects (scenes) were placed in three large locations (venues). Participants were made aware of these constraints in their initial briefing.

In total, this made for 24 unique stimulus components, 3 scenes, 3 venues, 6 faces, 6 unique face sockets, and 6 unique "time-slots" for faces. The combinatorics of this set of objects allows for 3,110,400 (3!*6!*6!) possible unique study trials. Each of these combinations of components involves a set of bindings between the different components, 3 scene-venue mappings, 6 scene-face mappings, 6 scene-socket mappings, 6 scene-time mappings, 6 venue -face mappings, 6 venue -socket mappings, 6 venue –time mappings, 6 face-socket mappings, 6 face-time mappings, and 6 socket -time mappings for a total of 57 unique relations present in each study and reconstruction trial.

The set of study trials was generated by randomly assigning a complete set of relations (57 in total) to the 24 components in each of 50 blocks. The set was checked to ensure there was no exactly repeated configuration. This corpus was divided into four counter balanced versions and each participant viewed one of these. Thus each participant viewed the same set of study trials, though not necessarily in the same order to avoid any incidental ordering effects.

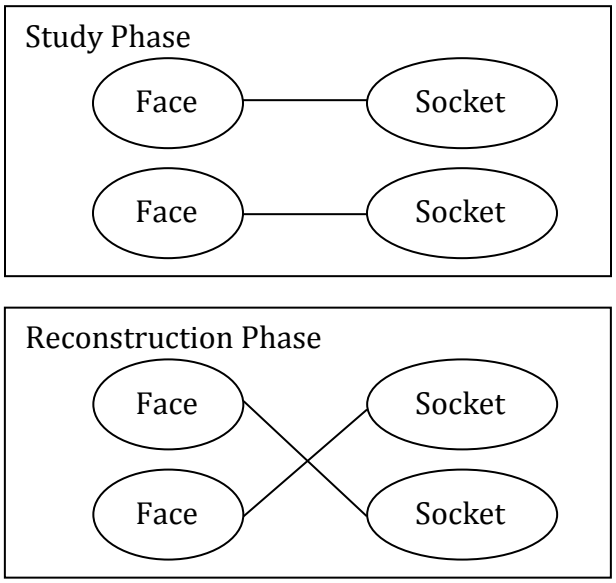
Analysis

Relational Performance via graph theoretic analysis

A natural framework for analyzing these data is graph theory (for an introduction, c.f. Diestel, 2010). Each of the 24 unique items is mapped to a "node" in a graph, and each of the 57 unique relations is an "edge" connecting a pair of item nodes. By constructing a graph composed of the configuration items and relations in the study phase, and comparing it to the graph constructed from the reconstruction configuration we can straightforwardly measure the relations that were reconstructed correctly and those that were misassigned (c.f. Figure 3.3). This data can be analyzed to see if performance varies across different types of relations (e.g., are participants more successful at putting faces in the correct sockets or at putting backgrounds in the correct venues?)

Additionally, this method helps disambiguate which portion of performance is due to memory for particular configuraitons and performance which might be predicted based soley on the general rules that generate each configuraiton. That is to say, since all of the studied configurations were generated by a single psuedorandom algorithm, participants could achive above-zero reconstruction performance by simply following all of the rules of that algorithm *even if they never saw the particular configuration they were meant to reconstruct!* Thus, the lower bound on performance (i.e. "chance level" performance), depends upon the participants' model of the algorithm that generated the inputs (and upon random effects such as idiosynchratic strategies, fluctuations in attention etc.). This is roughly equivalent to saying that the participants' performance is a combination of "episodic" (i.e., information tied to a particular reconstruction), and "semantic" (i.e., information common to all reconstructions), components. The elegant way to decompose the "episodic" and "semantic" components is to compute the degree of overlap between a studied-reconstruction pair, and then subtract from that the degree of overlap between that reconstruction and the reconstruction immediately prior. Since the two reconstructions can only resemble each other due to information which is shared across trials, their overlap must correspond to the "semantic" component and any additional information provided by the study trial (i.e., the "episodic" component) is then equivalent to increase in similarity between studied and reconstructed configurations over the semantic component.

Figure 3.3



An example of a study-reconstruction pair of graphs. In this case, the reconstruction has an error, the Face-socket relations for Face 1 and 2 have been swapped.

Using this framework we were able to examine participants performance by measuring how many relations are correctly reconstructed. Additionally, we are able to compare different types of relations to see if some are better remembered than others.

Domains, Complexity, Arbitrariness

There are ten different types of relations reconstructed by participants (generated by the ten ways it is possible to bind together a pair of elements, e.g., a face and a socket). We examined three critical comparisons:

1. Object-Object relations v. Spatial-Spatial relations

   Face-background relations and socket-venue relations are matched in the degree of complexity (3 elements mapped to 6 elements), and arbitrariness (2-to-1 mappings), but differ in the domain they concern (faces and backgrounds are both "objects dropped into slots" while sockets and venues are both "slots that accept objects"). Thus if performance differs it should be due to the underlying effect of the domain.
2. Simple mappings v. complex mappings

The background-to-venue binding maps 3 elements to 3 elements, using a 1 to 1 and onto rule. The face-to-time binding maps 6 elements to 6 elements, also using a 1 to 1 and onto rule. Thus differences in performance between these two bindings are likely due to differences in their complexity.

3. Strongly rule-bound mappings v. arbitrary mappings.
   The most strongly constrained binding is venue-to-time which is completely deterministic. We compared it to the equally complex, but less strongly constrained background-to-time binding (since backgrounds appear at arbitrary times).

Structural measures via realignment

Looking at reconstruction performance only in terms of "errors" is somewhat misleading. When a participant binds an element incorrectly, they are guaranteed to make a second error since by assigning relations incorrectly excludes the nodes to which those relations were assigned from being bound correctly. Thus, no error can be made independently of any other error. Worse still, some errors introduce unresolvable conflicts in subsequent reconstruction. For example, if the participant transposes a pair of background-venue relations (i.e., swapping the locations of a pair of scenes), it is impossible to correctly assign both face-background and face-venue bindings (since the scenes are in the wrong venues, the participant must choose to place the appropriate faces in the correct background or the correct venue, but cannot do both).

To capture how well participants are able to preserve the general structure of relations in the presence of different error types we created a measure that can quantify the distance between study and test trials. This measure is aimed at capturing quantitatively what is often obvious to memory researchers qualitatively-that a participant confused a pair of items or a block of trials-but made an excellent reconstruction of the configuration given that initial mistake.

This measure was constructed by describing each configuration with a sentence of 36 words, one for each of the elements that appeared at reconstruction time (i.e., 3 venues, 3 backgrounds, 6 faces, 6 times, and 18 sockets). The order of the symbols corresponded to the order in which they appeared reading from left to right placing venues before backgrounds, backgrounds before sockets, sockets before faces, and faces before times (i.e., the symbols were ordered according to a hierarchical, right-linear grammar). For example, the sentences

describing the study and reconstruction (with the mismatched sub-strings underlined) trials depicted in Figure 3.2 are:

Study

V1, B1, S1, S2, S3, S4, F1, T1, S5, F2, T2, S6,
V2, B2, S7, S8, F3, T3, S9, S10, S11, S12, F4, T4,
V3, B3, S13, F5, T5, S14, S15, F6, T6, S16, S17, S18.

Reconstruction

V1, B1, S1, S2, S3, S4, F1, T1, S5, F2, T2, S6,
V2, B2, S7, S8, F4, T4, S9, S10, S11, S12, F3, T3,
V3, B3, S13, F5, T5, S14, S15, F6, T6, S16, S17, S18.

The distance between sentences can be measured by calculating the number of operations required to convert one such sentence to another. There are several possible operation measures (e.g., weighted Levenstein which measures substitutions, deletions, and insertions), however, since all of the symbols were present in the sentence, we defined operation cost in this case in terms of the number of *rotations* (i.e., moving all the symbols one step to the right or left) or *reflections* (reversing the order of the symbols) of the non-matching sub-region of the sentence pair. To count the operations, the re-alignment algorithm identifies the mismatched region, and applies the rotation and reflection transforms to the string of mismatched elements. It then finds the transform that results in the greatest increase in match, and then recurs, focusing on the new, smaller error region. For example, the above study/reconstruction sentences are realigned via the following steps:

Mismatched region
F4, T4, F3, T3
Rotation+1
T3, F4, T4, F3
Rotation +2
F3, T3, F4, T4
Target
F3, T3, F4, T4

These two sequences are thus described as being 2 operations (2 rotations + 0 recursions), from each other. In addition to the operations required to convert one sentence to the other, we can count the number of symbols which are simultaneously placed into the correct configuration. If multiple elements "snap" into place simultaneously more often than might be expected due to chance, it suggests that these elements were bound together in an associated "chunk." Just as with cost, a difference in chunk size suggests different strategies for different conditions. For details on the realignment algorithm and specifics on the structural similarity measures see Appendix B.

More fundamentally, this approach can help to elucidate the representational format of reconstructive memory by telling us about the kind of information that participants store and how they organize it.

For additional information about how metrics are calculated, see Appendix C. We examine two measures produced by this algorithm: total cost (the mean of operations + recursions, for the example trial, the total cost would be 2), and chunk size (the mean of the size of the steps in the cost function, this is equivalent to the number of elements that snap into place simultaneously when the two sentences are realigned, suggesting that these elements were associated in a single chunk, for the example trial, the chunk size would be 4).

This metric corresponds very closely to the qualitative descriptions of configurations and the modifications necessary to convert the reconstruction to the studied configuration. We might describe the two configurations by saying that face 3 and face 4 had been swapped (in both space and time) by the participant, between study and reconstruction, and that to fix this mismatch we ought to swap them back.

The "Semantic" and "Episodic" components

In all of the above measures, there is no guarantee that participants' performance, whether measured by accuracy or similarity, is due to exposure to the stimulus configuration presented at study time. Even if participants did not observe the study trial, so long as they created *some* reconstruction, it would resemble the study trial at some chance level. Some of this baseline performance is due to the constraints of the physical universe (e.g., each item may only occupy a single place and time, and no two items may occupy the same place and time), some of the baseline performance is due to constraints of the experimental apparatus (e.g., it was

impossible to place a background into a socket, only faces could be placed there), some of baseline is due to explicit experimental instructions (e.g., use each socket only once), and some is due to implicit instructions (e.g., place exactly two faces per background), and so on. The experimenter cannot know exactly what mixture of these constraints a participant chooses to use. Further, participants are free to use their own strategies and biases during reconstruction, including being willfully perverse and reconstructing configurations that systematically differ from the studied configuration.

To disambiguate the degree to which participants relied upon information present in studied configurations (e.g., the bald man appeared in the upper left hand corner of the desert), from participants general reconstruction tendencies regardless of the particular studied configuration we performed a Monte-Carlo style control, using participants' own performance as a model for their own general tendencies.

We first considered how accurately or similarly a participant had reconstructed a studied configuration. We then considered how accurately or similarly a participant had reconstructed their immediately prior reconstruction. Since the each of the two reconstructions were constructed by the same participant, but did not share the information contained in the intervening study trial (i.e., a prior reconstruction could only contain information gleaned from the current study trial if the participants were precognitive), their performance can only correspond to whatever set of rules or strategies the participant chose to employ, and this performance corresponds to how well *any* of the participants reconstructions would serve as an example of the studied configuration-even if they had not yet seen the studied configuration. Any additional performance above this baseline measure could only result from information present in the studied configuration. The logic of this comparison is very similar to that of word-stem completion priming tasks that control for the chance a participant would produce the prime in response to the stem if they had not seen the prime (Warrington & Weiskrantz, 1970), with the additional benefit that each participant serves as their own control. For an extensive discussion of why this approach is preferable to more traditional measures of chance, see Appendix B.

Because performance on reconstruction-reconstruction pairs corresponds to information present *across trials*, coded by general constraints, rules, strategies, or biases, we refer to this component of performance as the "semantic" component, taking this to mean "the portion of performance that could have been predicted without any knowledge of the particular studied configuration." The component of performance above this baseline that is present only *within trials*, and is tied to an individual studied configuration we refer to as the "episodic" component, taking this to mean "performance that could not have been predicted without information of the particular configuration being reconstructed." It is important to note that these terms, are not always used in precisely this way.

Results

How much did participants' reconstructions depend upon episodic v. semantic information? Accuracy

Participants' mean reconstruction accuracy was 64.0% with a standard error of 1.6%. Of this performance approximately 55.1% was due to semantic knowledge about the experiment (as measured by overlap between reconstructions), and 45.0% was due to information tied to the particular study-reconstruction episode (Figure 3.4). This is equivalent to saying that of the original 64% of performance, knowledge of the general rules was sufficient to raise performance from 0 to 35.3% (s.e. 0.2%) , and information tied to a particular trial was sufficient to add an additional 28.6% (s.e., 1.5%) to bring the total to 64% (s.e. 1.6%).

Figure 3.4: Semantic and episodic accuracy

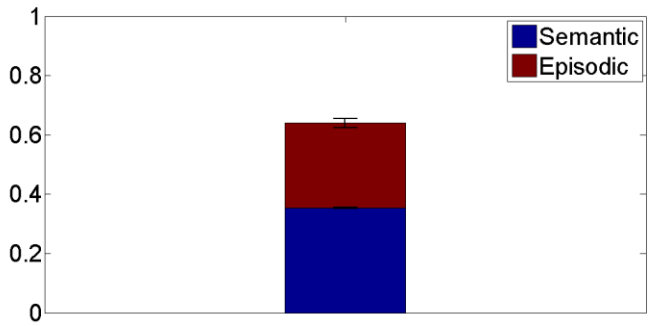


Figure 3.4 shows the relative proportion of performance attributable to cross-trial "semantic" information, and within trial "episodic" information.

Critical comparisons of performance

We made three critical comparisons between different types of bindings. First we compared bindings within two different domains: items and space. Item-item binding performance compared to location-location bindings (i.e., Were faces more strongly associated with backgrounds, regardless of location, or were sockets more strongly associated with venues, regardless of what item occupied them). Second we compared a simple 3-to-3 mapping to a more complex 6-to-6 mapping. Finally we compared a completely constrained deterministic binding to a more flexibly remappable binding of equal complexity (Figure 3.5). For an overview of all ten possible pair-wise bindings, see Appendix D.

Figure 3.5: Critical comparisons

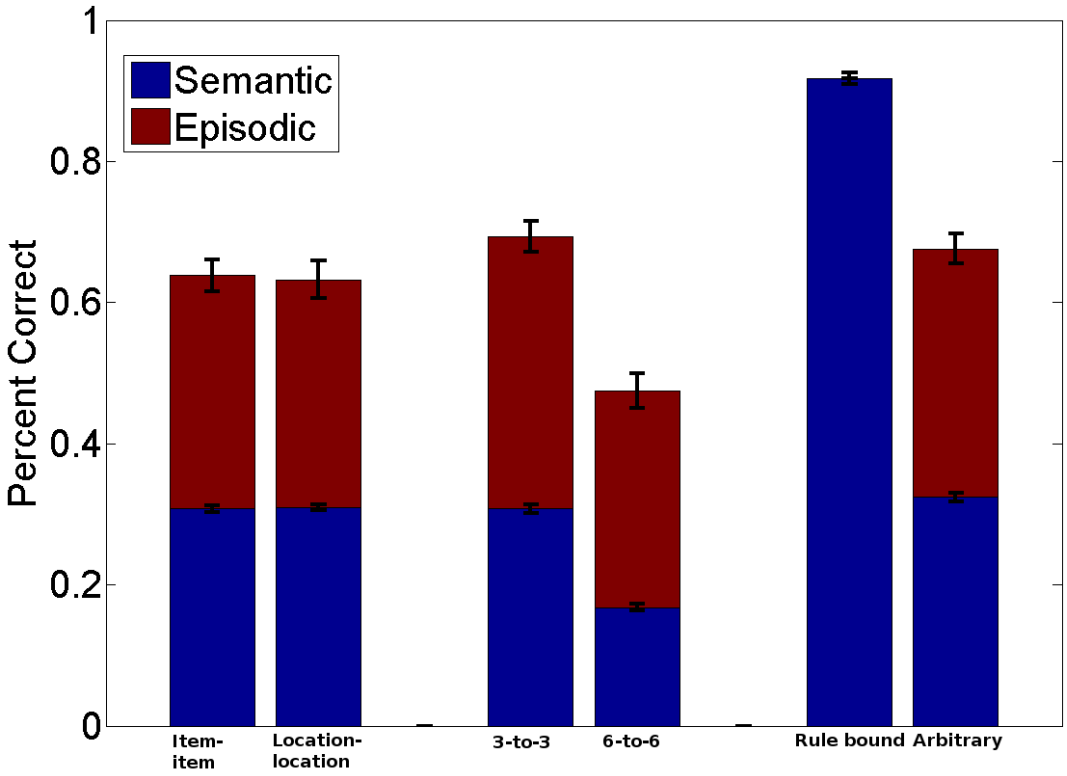


Figure 3.5 shows the episodic and semantic components of the three critical comparisons.

In the three critical comparisons, there was no effect of domain (item v. location) after controlling for complexity and arbitrariness, but there was an effect of complexity and

arbitrariness, with more complex and more arbitrary patterns of bindings yielding lower performance.

Additionally, a mixed ANOVA with factors of domain (item-item, item-space, item-time, space-space, space-time), complexity (3 to 3, 3 to 6, and 6 to 6), arbitrariness (deterministic, flexible), and memory type (semantic, episodic), found significant effects of complexity (d.f., 1, $F=20.17$, $P<0.0001$), arbitrariness (d.f., 1, $F=14.4$, $P<0.0002$), and memory type (d.f. 1, $F=12.37$, $P<0.0005$), but no significant effect of domain (d.f., 4, $F=0.44$, $P>0.783$).

Structural similarity

Comparing structural similarity of configurations produced a slightly different story of the relative value of semantic and episodic information. With no constraints (i.e., with a random string of symbols) it took an average of 63.8 steps (s.e., 0.3) to convert one string to another (i.e., any pair of configurations would be approximately 63.8 operations apart in the experiment's "configuration space"). However, it took only 14.1 steps (s.e., 0.2) to convert a reconstruction to the immediately preceding reconstruction, suggesting that the participants' grasp of the semantic rules considerably constrained and improved their performance. However, realigning a reconstruction with the appropriate study trial was even better, taking only 12.2 steps (s.e., 0.6), providing a marginal improvement upon the semantic performance (Figure 3.6). This relationship held for both the complexity of the operations required to translate one configuration to another (Random: mean = 51.8, s.e. = 0.3; Semantic: mean = 9.1, s.e. = 0.2; Episodic: mean = 7.8, s.e.= 0.4) and for the number of recursions required by the algorithm (Random: mean = 12.0, s.e. = 0.03; Semantic: mean = 5.0, s.e. = 0.06; Episodic: mean = 4.4, s.e. = 0.18), suggesting that knowledge of the general semantics, and of specific information tied to a particular episode decreased both the breadth and depth of the search required to realign the two configurations.

In addition, the mean number of elements that simultaneously "snapped" into place as a single chunk (ignoring those initially correctly ordered) varied across these three conditions. For randomly arranged sentences, the average chunk size was 2.6 (s.e., 0.01), for semantically related reconstructions 3.6 (s.e., 0.03), and for episodically related study-reconstruction pairs,

3.0 (s.e., 0.06), suggesting that part of the increase in similarity over randomness was due to using larger "chunks."

Figure 3.6: Mean number of operations to restore studied target

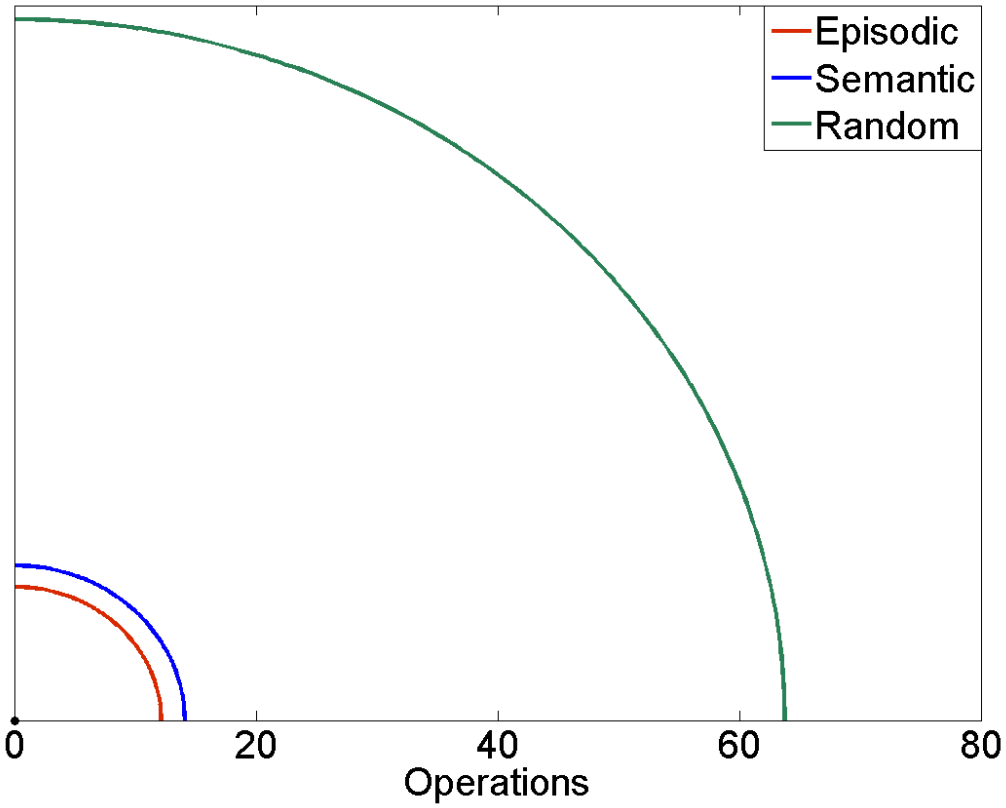


Figure 3.6 shows the mean number of operations required to convert a configuration to the studied target (represented by the point at the origin). The green circle represents the number of steps from a randomly ordered sentence to the target, the blue circle the number of steps from a previous reconstruction to the target, and the red circle the number of steps from the participants' current reconstruction to the target.

Is structural similarity due to chunking?

There two possibilities that might explain the pattern of similarity in Figure 6. First, study-reconstruction, or reconstruction-reconstruction pairs might simply have better initial performance: more items are correctly bound initially, but incorrectly bound items are bound randomly. Second, the incorrectly positioned items share useful structural relationships, that is, they are associated in a chunk, but the chunk as a whole is bound incorrectly to the other items.

This latter possibility cannot be the case in the randomly bound case, since there are by definition, no meaningful structural relationships or associations between individual elements.

Since each step in the realignment algorithm assumes that all of the incorrect items belong to the same chunk and attempts to find the transform that maximizes the number of correct bindings while preserving the structure of the chunk, we can evaluate these two hypotheses by examining how much similarity improves over the course of realignment for the "episodic" and "semantic" representations relative to randomly ordered elements (Figure 3.7).

Figure 3.7: Similarity improves over the course of realignment

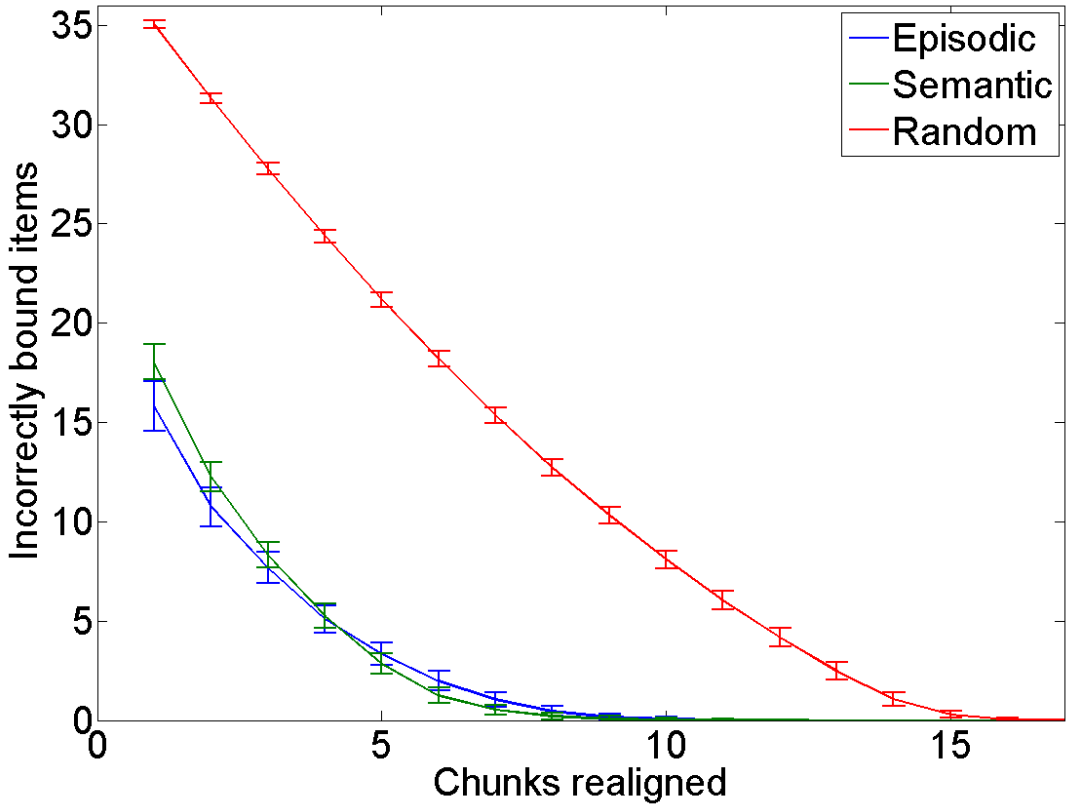


Figure 3.7 shows the course of realignment for episodic, semantic, and random comparisons. While much of the benefit of episodic and semantic representations over random comes from their lower initial number of incorrectly bound items, both slopes are also more curvilinear, suggesting that the incorrectly bound items within the episodic and semantic representations are realigned in larger "chunks" than the randomly ordered items.

There is a large difference in the initial performance between the episodic and semantic realignments and the random comparison, the random comparison begins with a mean of 35.0

(s.e., 0.2) incorrectly bound items, compared to just 18.0 (s.e., 0.9) incorrectly bound items for the semantic (reconstruction-reconstruction) realignments, and only 15.8 (s.e., 1.2) for episodic (study-reconstruction) realignments.

However, both episodic and semantic reconstructions had additional structure present in their incorrectly aligned elements. The first "chunk" aligned by the algorithm for random data contained a mean of 3.7 elements (s.e., 0.01), while the first chunk realigned in the semantic condition contained 5.8 (s.e., 0.2) elements, and the episodic condition contained 5.0 (s.e., 0.3) elements. However, this is not an apples-to-apples comparison, since the "first" chunk of the random data involved 35 elements of the 36 total elements, compared to just 18 for the semantic and 15.8 for the episodic conditions. When the realignment algorithm acted on a random data chunk of size 18, it was only able to realign 2.8 (s.e., 0.1) items. When it acted on a random chunk of size 16, it was only able to realign 2.7 elements (s.e. 0.1). Thus, the chunks successfully realigned in the semantic and episodic conditions were nearly twice as large as those aligned at similar chunk sizes in randomly bound elements.

To control for this, we measured the difference between both episodic and semantic reconstructions and random performance controlling for chunk size by finding a random chunk of the same size as the one present in the episodic and semantic reconstructions. The mean number of bound elements per-reconstruction in excess of what would be predicted from realigning random data was 12.4 (s.e., 0.4) for semantic reconstructions, and 15.4 (s.e., 0.8), for episodic reconstructions.

Discussion

We used an event-reconstruction paradigm to analyze reconstructive memory for previously viewed events assessing both accuracy and general structural similarity. In addition, we separated performance tied to the general cross-trial "semantics" from performance tied to a specific single-trial "episode," and showed that these two components made roughly equivalent contributions to performance. Using a set of critical comparisons we showed that participants' reconstruction accuracy was strongly related to the level of complexity of the relations they reconstructed, and to the arbitrariness of the rules for binding, but we found little to differentiate performance on different stimulus categories.

This pattern of performance suggests that much of the data explained by theories of mental processing can be equally well explained with reference to intrinsic differences in the complexity of the data, without the need to propose any special mechanism. Theories which emphasize the spatial or temporal nature of cognitive processing for example, cannot rest solely on differences in performance on spatial or temporal tasks, because spatial and temporal tasks must necessarily contain types of relational complexity (e.g., ordering, dimensionality, an ambiguous number of "locations"), absent in simpler item based tasks.

At a structural level, both the episodic and semantic reconstructions performed dramatically better than randomly bound configurations, both in their initial accuracy, and in the structure contained in the initially incorrectly bound items. These initially incorrectly bound items were shown to contain "chunks" of associated elements that could be simultaneously correctly bound with the same operation. Semantic similarity between reconstructions accounted for a preponderance of this chunked information, providing slightly more than 12 additional bindings per trial, while the episodic component provided approximately three additional bindings per trial (similar to what has been observed in other working memory tasks c.f. Cowan 2001). These findings suggest that traditional quantitative performance measures (such as "number of items correctly recalled"), tend to overestimate the degree to which memory performance is tied to particular episodes. Our measure of memory accuracy estimated that approximately 45% of participants' performance was explained by their episodic memory contribution, while the realignment measure suggested viewing the study trial only increased the number of known bindings by 25% and overall similarity between the reconstruction and the study trial by 14%.

To help explain the seeming tension between the episodic component's large contribution to accuracy and small contribution to similarity consider the following thought experiment: A participant is asked to guess a number between 1 and 5. The experimenter, however, never chooses 4 or 5. Once the participant learns this, their performance increases from 20% to 33%, this semantic property of the experiment is therefore, worth a 13% increase in their success rate. However, on one trial, the participant learns that the experimenter has not selected the number 3. On this trial, therefore, they can achieve 50% performance, a boost of 17%! In one sense, they have half as much episodic information as semantic, on the other hand, because the

episodic information about a *difference* between this trial and prior trials, it is synergistic with the previously known semantic information provides a greater benefit than all of the semantic information put together. Since episodic information is a marginal increase on semantic, any linear increase in episodic information will yield exponential gains in performance by leveraging the existing semantic information and further, the more semantic information, the greater the leverage.

This phenomenon was born out anecdotally in participants' post-experiment interviews. Nearly all participants found the experiment extremely difficult, and estimated that they had done very poorly, yet performance was relatively high: 64% of the bindings in each reconstruction were correctly reconstructed. Participants accurately reported their difficulty in encoding the large amount of rapidly presented information, capturing only about 15 of the total 57 unique bindings present in episode, but by leveraging the rules and the small amount of episodic information they were able to encode they were able to achieve high performance.

This dynamic interactivity of episodic and semantic knowledge is a constant feature of every day experience, and has strong basis in past literature. In animal literature using radial arm mazes, rats must remember both which arms tend to be baited, and which arms were baited, but now have been consumed (Davis et al. 1986). This distinction (often referred to as the difference between "reference" and "working" memory), maps very closely onto the distinction made here between "semantic" and "episodic." The rats "episodic" knowledge that a particular arm is no longer baited is far more valuable to performance than one would expect in the absence of the "semantic" knowledge of which of the arms tend to be baited.

However, this is not to say that the interplay of episodic and semantic always redounds to our benefit. In the famous DRM paradigm (Deese 1959, Roediger & McDermott 1995), participants see a list of words all of which are related to a single target word that is not presented (e.g., bed, snooze, awake, alarm, dreams, blanket, pillow. Unpresented target word: sleep), and often are more likely to produce the unpresented target word during tests of free recall than they are to produce any word actually on the list. This recalled, yet unpresented, word is often taken as an example of a "false" memory, in which the general "semantic" information present across the list has repeatedly primed the missing word causing it to be recalled. However, this entirely

frames the memory question in terms of inhibiting the target word, and entirely neglects how valuable the target word is for generating the words that *were* studied. Knowing the fact that everything on the list was sleep related helps generate a large number of sleep related words that might plausibly have been on the list at the minor cost of accidentally adding the target word to the list. The participants' "episodic" memory, the ability to recall presented words, depends not just on which individual words they saw, but upon the conceptual features that all of the words share. It is impossible to know if a word like "bed" was generated because of some specific knowledge, or simply because it was related to sleep.

These constant interactions between general knowledge and specific experience help to explain the richness of human memory, and how such memory can be applied to construct novel imaginings, plans, or scenarios. Participants' ability to create meaningful "chunks" of information, each of which itself contained meaningful internal structure dramatically improved their performance on the task, because even when they erred, there error was due to a misalignment of already meaningful units, and therefore the studied structure was preserved except for at the point of the swap. Participants' performance could be understood both in terms of its similarity to their target, and similarity to their prior performance. Since prior performance closely matched the general statistical tendencies of the data being presented, whenever a participant's knowledge of the particular target configuration failed, they could fall back on a pattern of behavior that general worked well, providing a robust level of performance even in the absence of information tied to a particular target. An interesting future avenue of study is to delve into the rules that organize these chunks. What are the criteria for binding together associates, and upon what brain systems and mechanisms does it depend?

References

Althoff, R. R., & Cohen, N. J. (1999). Eye-movement-based memory effect: a reprocessing effect in face perception. *Journal of experimental psychology. Learning, memory, and cognition*, *25*(4), 997–1010. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10439505

Burgess, N., Maguire, E. a, & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron*, *35*(4), 625–41. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12194864

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and brain sciences*, *24*(1), 87–114; discussion 114–85. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11515286

Davis,H., Tribuna,J., Pulsinelli,W., Volpe,B.; Reference and working memory of rats following hippocampal damage induced by transient forebrain ischemia, Physiology & Behavior, Volume 37, Issue 3, 1986, Pages 387-392, ISSN 0031-9384, 10.1016/0031-9384(86)90195-2.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. Journal of Experimental Psychology, 58, 17-22.

Diestel, R. (2010). *Graph Theory*. *Oberwolfach Reports* (fourth., pp. 1–451). Springer. doi:10.4171/OWR/2010/11

Dusek, J. a, & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(13), 7109–14. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=21293&tool=pmcentrez&rendertype=abstract

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature reviews. Neuroscience*, *1*(1), 41–50. doi:10.1038/35036213

Eichenbaum, Howard, & Cohen, N. (2001). *From conditioning to conscious recollection*.

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. a. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(5), 1726–31. doi:10.1073/pnas.0610561104

Hassabis, D., & Maguire, E. a. (2009). The construction system of the brain. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *364*(1521), 1263–71. doi:10.1098/rstb.2008.0296

Konkel, A., Warren, D. E., Duff, M. C., Tranel, D. N., & Cohen, N. J. (2008). Hippocampal amnesia impairs all manner of relational memory. *Frontiers in human neuroscience*, *2*(October), 15. doi:10.3389/neuro.09.015.2008

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual review of psychology*, 481–537. Retrieved from http://www.annualreviews.org/doi/pdf/10.1146/annurev.psych.51.1.481

Kumaran, D., & Maguire, E. a. (2005). The human hippocampus: cognitive maps or relational memory? *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *25*(31), 7254–9. doi:10.1523/JNEUROSCI.1103-05.2005

Maguire, E. a, Burgess, N., & O'Keefe, J. (1999). Human spatial navigation: cognitive maps, sexual dimorphism, and neural substrates. *Current opinion in neurobiology*, *9*(2), 171–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10322179

McKenzie, S., & Eichenbaum, H. (2012). New approach illuminates how memory systems switch. *Trends in cognitive sciences*, *16*(2), 102–3. doi:10.1016/j.tics.2011.11.010

Nadel, L., Samsonovich, a, Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. *Hippocampus*, *10*(4), 352–68. doi:10.1002/1098-1063(2000)10:4<352::AID-HIPO2>3.0.CO;2-D

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. *Why People Get Lost*. Oxford, UK: Clarendon Press. Retrieved from

http://www.ingentaconnect.com/content/oso/7347120/2010/00000001/00000001/art00006

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 803-814

Rosenbaum, R. S., Winocur, G., & Moscovitch, M. (2001). New views on old memories: re-evaluating the role of the hippocampal complex. *Behavioural brain research*, *127*(1-2), 183–97. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11718891

Ryan, L., Lin, C.-Y., Ketcham, K., & Nadel, L. (2010). The role of medial temporal lobe in retrieving spatial and nonspatial relations from episodic and semantic memory. *Hippocampus*, *20*(1), 11–8. doi:10.1002/hipo.20607

Sutherland, R. J., Weisend, M. P., Mumby, D., Astur, R. S., Hanlon, F. M., Koerner, a, Thomas, M. J., et al. (2001). Retrograde amnesia after hippocampal damage: recent vs. remote memories in two tasks. *Hippocampus*, *11*(1), 27–42. doi:10.1002/1098-1063(2001)11:1<27::AID-HIPO1017>3.0.CO;2-4

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual review of psychology*, *53*, 1–25. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev.psych.53.100901.135114

Tulving, E., & Markowitsch, H. (1998). Episodic and declarative memory: role of the hippocampus. *Hippocampus*, *204*, 198–204. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1098-1063(1998)8:3%3C198::AID-HIPO2%3E3.0.CO;2-G/abstract

Voss, J. L., Gonsalves, B. D., Federmeier, K. D., Tranel, D., & Cohen, N. J. (2011). Hippocampal brain-network coordination during volitional exploratory behavior enhances learning. *Nature neuroscience*, *14*(1), 115–20. doi:10.1038/nn.2693

Warrington, E. K., & Weiskrantz L. (1970) Amnesic syndrome: Consolidation or retrieval?. *Nature, 228*, 628-630.

Watson, P. D., Voss, J. L., Warren, D. E., Tranel, D. N., & Cohen, N. (in press). Hippocampal amnesia is dominated by relational errors. Hippocampus

Winocur, G., Moscovitch, M., & Bontempi, B. (2010). Memory formation and long-term retention in humans and animals: convergence towards a transformation account of hippocampal-neocortical interactions. *Neuropsychologia*, *48*(8), 2339–56. doi:10.1016/j.neuropsychologia.2010.04.016

MODELING ASPECTS OF HUMAN MEMORY FOR SCIENTIFIC STUDY

Authors
Michael Bernard, J. Dan Morrow, Shawn Taylor, Stephen Verzi, Craig Vineyard, Thomas Caudell, Neal Cohen, Howard Eichenbaum, Mark McDaniel, & Patrick Watson

This document was an interdisciplinary modeling effort involving collaborators at the University of Illinois at Urbana-Champaign, Sandia National Laboratories, Boston University, and the University of New Mexico.

It was published as a technical research report by Sandia National Laboratories in 2009. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

1. INTRODUCTION

Cognitive neuroscience research has found that the hippocampus plays a central role in forming and temporarily storing representations of personal experiences. These representations are later migrated to widespread areas of the cerebral cortex, which are then permanently stored. The focus of the Laboratory Directed Research and Development (LDRD) work was to produce a computational model that: (1) represents the fundamental features of hippocampus-dependent relational processing and (2) tests this representation against human memory by comparing the performance of normal humans subjects and people lacking normal hippocampal function (amnesic subjects) with the performance of the full model and the model without a functional "hippocampus" on the same memory tasks. Success of the model was measured in terms of similar performance compared to that of humans with and without the contribution of hippocampal processing, as a function of experimental parameters and controls. The intent was to develop a falsifiable model (i.e., one where we can find and understand its limitations and characterize them properly), where the trend of results from human experimentation match the trends of simulation. A second focus, modeling aspects of human reasoning, was dropped after the first fiscal year due to cutbacks in the overall funding of this project. As such, this technical report will only discuss the modeling of declarative memory.

1.1. Overview of the Problem and Idea

Memory is usually thought of as a passive record of past events and acquired factual knowledge. But our adaptive application of memory is to make plans for our future actions. Therefore, our conscious lives are dominated by interactions between retrospective memory, the capacity for recollection of general knowledge and one's personal history of previous actions and their outcomes, and prospective memory, our intentional application of knowledge, and history in directing ongoing decisions and behavior. This project is currently modeling of how the brain accomplishes retrospective recollection and memory. Our capacity for recollection is known to be supported by a system composed of several cortical association areas interacting with structures in the medial temporal lobe, and in particular, the hippocampus. There is a general consensus that the cortex is the repository of detailed representations of perceptions and thoughts and that the hippocampus supports the ability to bind together cortical representations and, when cued by part of a previous representation, to reactivate the full set

of cortical representations that compose a recollective, declarative (explicit) memory.

To date, computational models have not neurocognitively represented episodic recollection memory within an embodied, simulation environment. This creates several limitations regarding the plausibility of current models. First, current approaches do not dynamically collect "what," "where," and "when" perceptual information to produce an episodic memory trace. Second, current approaches typically create a false distinction between semantic and event-based, episodic memory. While semantic memory has a different phenomenology than episodic memory, there is strong evidence they are part of the same system (McKoon et al., 1986).

Research has found that the hippocampus plays a central role in forming and temporarily storing representations of personal experiences. These representations are later migrated to widespread areas of the cerebral cortex, which are then permanently stored.

To address the need for more plausibility model Sandia National Laboratories (SNL) has (1) produced a computational model that represents the fundamental features of hippocampus-dependent relational processing and (2) tested this representation against human memory by comparing the performance of normal humans subjects and people lacking normal hippocampal function with the performance of the full model and the model without a functional "hippocampus" on the same memory tasks. Success of the model was measured in terms of similar performance compared to that of humans with and without the contribution of hippocampal processing, as a function of experimental parameters and controls.

This effort is extending the current Sandia Cognitive Framework by incorporating a representation of memory processing, focusing on hippocampus and neocortical systems described in current complimentary learning systems theory (i.e., cortical-hippocampal theory of declarative memory, Eichenbaum, 2007). The model also specifies how hippocampal and cortical representations interact at multiple levels of abstraction to support the interleaving of new information within the cerebral cortex. For example, we integrated the perceptual features of relational memory processing into our computational model. This project produced two main products: (1) a neuro-cognitive computational architecture that represents episodic memory and (2) major review paper(s) submitted for publication. This work extended current computational models (for example, McClelland et al.; Psych Rev. 1995) wherein a pre-existing knowledge structure in cortical areas is challenged to incorporate new information within an

existing network. To accomplish this goal we collaborated with leading experts from academia. Specifically, the external research team consisted of: (1) Howard Eichenbaum, professor of psychology and neuroscience and Director of the Center for Memory and Brain at Boston University, (2) Neal Cohen, professor of psychology and neuroscience at the Beckman Institute of the University of Illinois, (3) Thomas Caudell, professor and Director of the Center for High Performance Computing at University of New Mexico, and (3) Mark McDaniel, professor of psychology at Washington University.

2. THE DECLARATIVE MEMORY SYSTEM

Like the Roman god Janus, memory looks both into the past and the future. Memory is usually thought of as a passive record of past events and acquired factual knowledge. But our adaptive application of memory is to make plans for our future actions. Therefore, our conscious lives are dominated by interactions between *retrospective memory,* the capacity for recollection of general knowledge and one's personal history of previous actions and their outcomes, and *prospective memory,* our intentional application of knowledge and history in directing ongoing decisions and behavior. The discussion will begin by outlining the experimental evidence on the cognitive and neural mechanisms of recollection, and then consider retrospective memory from experimental studies in cognitive science. The paper will then outline a formal model and its implementation in software.

2.1 What is recollection?

We have all been in the situation where we meet someone who seems highly familiar but we cannot recall who they are or why we know them. Sometimes, we just give up and say, "Don't I know you?" Alternatively, when a clue or sufficient mental searching helps us retrieve a wealth of information all at once, including the name, where we met before, and the circumstances of the meeting. Considerable current research on recollection has focused the distinction between a vivid recollection the lesser condition of a sense of familiarity with a particular person or object. Familiarity comes rapidly and reflects the strength match between a cue and a stored memory template. It is an isolated ability to identify a person or object as previously experienced. Recollection is typically slower and measured by the number of qualitative associations retrieved and the organization of the memory retrieved. Thus, recollections typically include not only the item sought in memory but also three other kinds of additional

information:

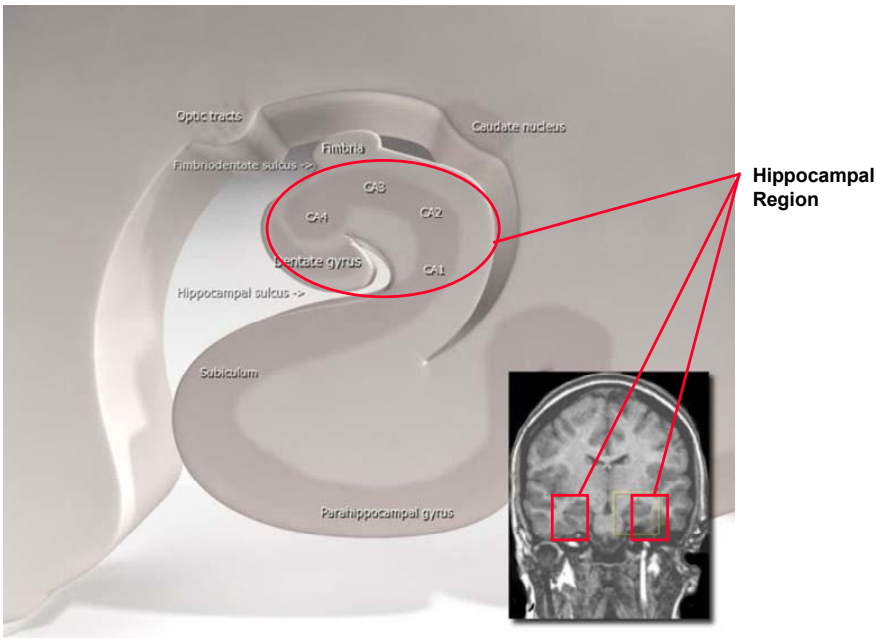
(1) a spatial and temporal context of the experience in which the item was previously encountered

(2) a replay of the sequence of events that compose an entire episode with that item

(3) and remembering additional related experiences with the item.

Furthermore, one brain area, the hippocampus, is critically involved in each of these aspects of recollection. Yonelinas et al. (2002) ROC analysis on recognition memory performance to show that mild hypoxia that causes damage largely confined to the hippocampus resulted in a severe deficit in recollection but normal familiarity. A similar pattern of deficient recollection and preserved familiarity was reported in a patient with relatively selective hippocampal atrophy related to meningitis (Aggleton et al., 2005). Further consideration of the three properties of introduced above provides insights into both the fundamental elements of recollection and the role of the hippocampus in memory processing (see Figure 4.1).

Figure 4.1. The hippocampal and parahippocampal regions within the brain



Events are represented as items in the context in which they were experienced. A fundamental

feature of recollection is memory for the spatial, temporal, and associational context in which experiences occur. Functional imaging studies support the notion that the hippocampus is activated during the encoding or retrieval of associations among many elements of a memory, a characteristic of context-rich episodic memories (for review see Cohen et al., 1999; Eldridge et al., 2000; Maguire, 2001; Addis et al., 2004). For example, Henke et al., (1997) observed greater hippocampal activation when subjects associated a person with a house, as compared to making independent judgments about the person and house and others have found selective hippocampal activation during recollection of the context of learning in formal tests of memory (e.g. Davachi et al., 2003; Ranganath et al., 2003). The coding of associations extends beyond item and context associations such that the hippocampus is also selectively activated during the encoding or retrieval of verbal (Davachi & Wagner, 2002; Giovanello et al., 2003a) and face-name associations (Small et al., 2001; Zeineh et al., 2003; Sperling et al., 2003). Correspondingly, recent neuropsychological studies have found that recognition of associations is impaired even when recognition for single items is spared in amnesic patients (Giovanello et al., 2003, Turriziani et al., 2004). These studies reported impairment in recognition memory for associations between words or between faces or face-occupations pairs, as compared to normal performance in recognition of single items. At the same time, other functional imaging studies and characterizations of amnesia have suggested that the hippocampus is sometimes involved in both associative and single item recognition, highlighting the need to clarify the nature of associative information that composes an "event" (Squire et al., 2004). Nevertheless, these findings are generally consistent with the notion that the hippocampus plays a distinct role in recollection associated with binding features of items and their context to represent salient events (Eichenbaum et al., 2007).

Studies that employ animal models can provide compelling evidence on the effects of selective hippocampal damage. Several studies have shown that damage limited to the hippocampus results in deficits in forming a memory for the context or location where items were once experienced (reviewed in Mumby, 2001). In one recent study, rats were initially exposed to two objects in particular places in one of two environmental chambers (Mumby et al., 2002). In subsequent recognition testing, the place of the object or the context was changed. Normal rats increased their exploration of objects that were moved to new places or put in novel contexts. By contrast, rats with hippocampal damage failed to recognize objects when either the place or

context was changed (see also Eacott & Norman, 2004).

Several investigators have argued that animals are indeed capable of remembering the temporal as well as spatial context in which they experienced specific stimuli (Clayton et al., 2003; Day et al., 2003). To further explore these aspects of episodic memory, Eichenbaum (a team member of this LDRD) developed a task that assesses memory for events from a series of events that each involve the combination of an odor ("what"), the place in which it was experienced ("where"), and the order in which the presentations occurred ("when"; Ergorul & Eichenbaum, 2004). On each of a series of events, rats sampled an odor in a unique place along the periphery of a large open field. Then, memory for the when those events occurred was tested by presenting a choice between an arbitrarily selected pair of the odor cups in their original locations. Normal rats initially employed their memory of the places of presented cups and approached the location of the earlier experience. Then they confirmed the presence of the correct odor in that location. Animals with selective hippocampal damage fail on both aspects of this task even though their memory for independent features of location and odor items was intact. These findings indicate that the hippocampus is critical for effectively combining the "what", "when", and "where" qualities of each experience to compose the retrieved memory.

Studies on the firing properties of single neurons in animals provide insights into the nature of neural population representations in the hippocampus. There is a large body of evidence that hippocampal neurons encode associations among stimuli, actions, and places that compose discrete events. Many studies have shown that hippocampal neurons encode an animal's location within its environment, and some view this as the principle function of hippocampal populations (Muller et al., 1999; Best et al., 2001). However, many other studies have shown that hippocampal neurons also fire associated with the ongoing behavior and the context of events as well as the animal's location (Eichenbaum et al., 1999). In the most direct examination of this issue, Wood et al (1999) directly compared spatial and non-spatial coding by hippocampal neurons by training animals to perform the same memory judgments at many locations in the environment. A large subset of hippocampal neurons fired only associated with a particular combination of the odor, the place where it was sampled, and the match-non-match status of the odor. In a similar study on the coding properties of hippocampal neurons in humans, Ekstrom et al. (2003) recorded in subjects as they played a taxi driver game, searching for passengers picked up and dropped off at various locations in a virtual reality town. They

observed that many of these cells fired selectively associated with specific combinations of a place and the view of a particular scene or a particular goal. These and other studies indicate that, in rats, monkey, and humans, a prevalent property of hippocampal firing patterns involves the representation of unique associations of stimuli, their significance, specific behaviors, and the places where these events occur (see Eichenbaum et al., 2004).

Episodes are represented as sequences of events. We live our lives through personal experience, and our initial construction of reality within consciousness is a form of episodic buffer that contains a representation of the stream of events as they just occurred (Baddeley, 2000). In an early characterization of episodic recollection, Tulving (1983) distinguished episodic memory as organized in the temporal dimension, and contrasted this scheme with a conceptual organization of semantic memory. Tulving (1983) argued that the central organizing feature of episodic memory is that "one event precedes, co- occurs, or follows another." This is reminiscent of Aristotle's (350BC) characterization of vivid remembering: "Acts of recollection, as they occur in experience, are due to the fact that one thought has by nature another that succeeds it in regular order." These characterizations emphasize the temporal organization of episodic memories.

The order of events within human memory depends on hippocampal function. In a study using a design similar to that described above, Hopkins et al. (1995) found that patients with hypoxic brain injury involving in shrinkage of the hippocampus are impaired in memory for the order of a series of 6 words, pictures, or spatial locations. These patients were, however, also impaired in recognition of the items, undermining an unambiguous interpretation of a deficit in the order of the events independent of memory for the events. More recently, Spiers et al. (2001) reported a selective deficit in order memory independent of item memory in a patient with selective hippocampal damage due to perinatal transient anoxia (Vargha-Khadem et al., 1997). In this study the patient explored a virtual reality town in which he received objects from virtual characters. His recognition of the familiar objects was intact, but he was severely impaired in memory for the order in which he received objects, as well as for where he received them. Also, Downes et al. (2002) reported that patients with medial temporal lobe damage that included bilateral hippocampal damage were impaired in memory for the order of presentation of words for which recognition of the items was equivalent. Also, evidence from the deferred imitation task, where subjects are required to remember an action sequence, indicate a critical role for

the hippocampus (McDonough et al., 1995; Adlam et al., 2005). Thus, humans with hippocampal damage are impaired in memory for the order of events in unique episodes even in cases where recognition memory is intact.

Studies on animals also show that the representation of memories by the hippocampus incorporates not only items that must be remembered, but also the events that precede and follow. For example, Honey et al. (1998) provided a simple demonstration of the importance of temporal order in hippocampal processing, reporting that hippocampal lesions disrupted animals' normal orienting response when a pair of stimuli are presented in the opposite order of previous exposures. The specific role of the hippocampus in remembering the order of a series of events in unique experiences has been explored using a behavioral protocol that assesses memory for episodes composed of a unique sequence of olfactory stimuli (Fortin et al., 2002; see also Kesner et al., 2002). Memory for the sequential order of odor events was directly compared with recognition of the odors in the list independent of memory for their order. On each trial rats were presented with a series of five odors, selected randomly from a large pool of common household scents. Memory for each series was subsequently probed using a choice test where the animal was reinforced for selecting the earlier of two of the odors that had appeared in the series. In later sessions we also tested whether the rats could identify the odors in the list independent of their order, by was rewarding the selection of a novel odor against one that had appeared in the series. Normal rats performed both tasks well. Rats with hippocampal lesions could recognize items that had appeared in the series but were severely impaired in judging their sequential order.

How do hippocampal neuronal populations represent the sequences of events that compose distinct episodes? A common observation across many different behavioral protocols is that different hippocampal neurons become activated during every event that composes each experience, including during simple behaviors such as foraging for food (e.g., Muller et al., 1987) as well as learning related behaviors directed at relevant stimuli that have to be remembered in studies that involve classical conditioning, discrimination learning, and non-matching or matching to sample tasks to tests and a variety of maze tasks (e.g. Hampson et al., 1993; for review, see Eichenbaum et al, 1999). In each of these paradigms, animals are repeatedly presented with specific stimuli and rewards, and execute appropriate cognitive judgments and conditioned behaviors. Corresponding to each of these regular events, many hippocampal cells

show time-locked activations associated with each sequential event. Also, as described above, many of these cells show striking specificities corresponding to particular combinations of stimuli, behaviors, and the spatial location of the event. Thus, hippocampal population activity can be characterized as a sequence of firings representing the step-by-step events in each behavioral episode.

Furthermore, these sequential codings can be envisioned to represent the series of events and their places that compose a meaningful episode, and the information contained in these representations distinguishes related episodes that share common events and therefore could be confused. Recent studies on the spatial firing patterns of hippocampal neurons as animals traverse different routes that share overlapping locations provide compelling data consistent with this characterization. In one study, rats were trained on the classic spatial alternation task in a modified T-maze (Wood et al., 2000; see also Frank et al., 2000; Ferbinteanu and Shapiro (2003). Performance on this task requires that the animal distinguish left-turn and right-turn episodes that overlap for a common segment of the maze and requires the animal to remember the immediately preceding episode to guide the choice on the current trial, and in that way, the task is similar in demands to those of episodic memory. If hippocampal neurons encode each sequential behavioral event and its locus within one type of episode, then most cells should fire only when the rat is performing within either the left-turn or the right-turn type of episode. This should be particularly evident when the rat is on the “stem” of the maze, when the rat traverses the same set of locations on both types of trials. Indeed, a large proportion of cells that fired when the rat was on the maze stem fired differentially on left-turn versus right-turn trials. The majority of cells showed strong selectivity, some firing at over ten times the rate on one trial type, suggesting they were part of the representations of only one type of episode. Other cells fired substantially on both trial types, potentially providing a link between left-turn and right-turn representations by the common places traversed on both trial types.

Functional imaging studies in humans have also revealed hippocampal involvement in both spatial and non-spatial sequence representation. Several studies have shown that the hippocampus is active when people recall routes between specific start points and goals, but not when subjects merely follow a set of cues through space (Hartley et al.2003). In addition, the hippocampus is selectively activated when people learn sequences of pictures (Kumaran & Maguire, 2006). Even greater hippocampal activation is observed when subjects must

disambiguate picture sequences that overlap, parallel to the findings on hippocampal cells that disambiguate spatial sequences (Wood et al., 2000).

Memories are networked to support inferential memory expression. Further consideration of the cognitive properties of episodic memory suggest that related episodic representations might be integrated with one another to support semantic memory and the ability to generalize and make inferences from memories. Referring to how related memories are integrated with one another, William James (1890) emphasized that "...in mental terms, the more other facts a fact is associated with in the mind, the better possession of it our memory retains. Each of its associates becomes a hook to which it hangs, a means by which to fish it up by when sunk beneath the surface. Together they form a network of attachments by which it is woven into the entire tissue of our thought." James envisioned memory as a systematic organization of information wherein the usefulness of memories was determined by how well they are linked together.

There are two main outcomes of the linking of representations of specific experiences. One is a common base of associations that are not dependent on the episodic context in which the information was acquired. Thus when several experiences share considerable common information, the overlapping elements and common links among them will be reinforced, such that those items and associations become general regularities. The representation of these general regularities constitutes semantic "knowledge" that is not bound to the particular episode or context in which the information was encoded. The networking of episodic memories by common elements provides a mechanism for the commonly (albeit not universally, see Tulving, 2002) held view that semantic knowledge is derived from information repeated within and abstracted from episodic memories.

There is considerable evidence that hippocampal neurons indeed extract the common features among related episodes. In all the studies described above, a subset of hippocampal neurons encode features that are common among different experiences – these representations could provide links between distinct memories. For example, in the Wood et al. (1999) study on odor recognition memory, whereas some cells showed striking associative coding of odors, their match/non-match status, and places, other cells fired associated with one of those features across different trials. Some cells fired during a particular phase of the approach towards any

stimulus cup. Others fired differentially as the rat sampled a particular odor, regardless of its location or match-non-match status. Other cells fired only when the rat sampled the odor at a particular place, regardless of the odor or its status. Yet other cells fired differentially associated with the match and nonmatch status of the odor, regardless of the odor or where it was sampled. Similarly, in Ekstrom and colleagues' (2003) study on humans performing a virtual navigation task, whereas some hippocampal neurons fired associated with combinations of views, goals, and places, other cells fired when subjects viewed particular scenes, occupied particular locations, or had particular goals in findings passengers or locations for drop off. In studies that have recorded hippocampal neuronal activity as rats perform alternation tasks in a T-maze (Wood et al., 2000; Frank et al., 2000; Ferbintineau & Shapiro, 2003), whereas many cells distinguish overalapping actions and locations on the maze, some cells capture the common places and events between the different types of episodes.

The notion that hippocampal cells might reflect the linking of important features across experiences and the abstraction of common information was also highlighted in recent studies on monkeys and humans. Hampson et al. (2004) trained monkeys on matching to sample problems, then probed the nature of the representation of stimuli by recording from hippocampal cells when the animals were shown novel stimuli that shared features with the trained cues. They found many hippocampal neurons that encoded meaningful categories of stimulus features and appeared to employ these representations to recognize the same features across many situations. Kreiman et al., (2000a) characterized hippocampal firing patterns in humans during presentations of a variety of visual stimuli. They reported a substantial number of hippocampal neurons that fired when the subject viewed specific categories of material, e.g., faces, famous people, animals, scenes, houses, across many exemplars of each. A subsequent study showed that these neurons are activated when a subject simply imagines its optimal stimulus, supporting a role for hippocampal networks in recollection of specific memories (Krieman et al., 2000b). A subsequent study showed that some hippocampal neurons are activated a subject views any of a variety of different images of a particular person, suggesting these cells could link the recollection of many specific memories related to that person (Quiroga et al., 2005). This combination findings across species provides compelling evidence for the notion that some hippocampal cells represent common features among the various episodes that could serve to link memories obtained in separate experiences.

The second outcome from a network of linked memories is a capacity to use the common elements to retrieve multiple memories that include that element. Furthermore, hippocampal representations could support a capacity to "surf" the network of linked memories and identify relationships and associations among items that were experienced in distinct memories and therefore are only indirectly related. A single cue could generate the retrieval of multiple episodic and semantic memories, and cortical areas can access these multiple memories to analyze the consequential, logical, spatial, and other abstract relationships among items that appeared separately in distinct memories. These logical operations on indirectly related memories can support inferences from memory. The activity of searching and surfing networks of memories, and then comparing and contrasting memories could underlie our awareness of memories and the experience of conscious recollection. The organization of linked experience-specific and experience- general memories with the capacity for association and inference among memories is called a "relational memory network."

In a series of studies, Eichenbaum has used a model system of rodent olfactory memory to explore the importance of the hippocampus in the linking memories and using the resulting relational networks to make associational and logical inferences from memory. One study examined the role of the hippocampus in making indirect associations between stimuli that were each directly associated with a common stimulus. Initially, control rats and rats with hippocampal lesions were trained on a series of overlapping "paired associates" (Bunsey & Eichenbaum, 1996). On each trial, the rat was initially presented with one of two initial items in a pairing, and then had to select the arbitrarily assigned associate. For example, for training on the pairs A-B and X-Y, if A was the initial item, then the rat had to select B and not Y; conversely, if X was the initial item the rat had to select Y and not B. Then the rats were trained on a second paired associated list where the initial items were the second items in the first list and new items were the associates (B-C and Y-Z). Thus, when B was presented initially, the rat was required to select C and not Z; when Y was presented initially, the rats was then required to select Z and not C. After training on all four paired associates, the rats were tested on their knowledge of the indirect relations among the pairings. These tests involved presentations of an initial item from the first learned paired associates (A or X) followed by a choice between the second items of the later learned associates (C versus Z). Normal rats demonstrated their ability to express these indirect relations by selecting C when A was presented and Z when X was

presented, whereas rats with selective hippocampal damage showed no capacity for this inference from memory. These findings, combined with observations on another transitive inference task (Dusek & Eichenbaum, 1997), indicate that the hippocampus is critical to binding distinct memories into a relational network that supports flexible memory expression.

In another experiment, rats learned a hierarchical series of overlapping odor choice judgments (e.g., A > B, B > C, C > D, D > E), then were probed on the relationship between indirectly related items (B > D ?). Normal rats learned the series and showed robust transitive inference on the probe tests. Rats with hippocampal damage also learned each of the initial premises but failed to show transitivity (Dusek & Eichenbaum, 1997). The combined findings from these studies show that rats with hippocampal damage can learn even complex associations, such as those embodied in the odor paired- associates and conditional discriminations. But, without a hippocampus, they do not interleave the distinct experiences according to their overlapping elements to form a relational network that supports inferential and flexible expression of their memories (see also Buckmaster et al., 2004).

Complementary evidence on the role of the hippocampus in networking of memories comes from two recent studies indicating that the hippocampus is selectively activated when humans make inferential memory judgments. In one study, subjects initially learned to associate each of two faces with a house and, separately, learned to associate pairs of faces (Preston & Gabrieli, 2004). Then, during brain scanning, the subjects were tested on their ability to judge whether two faces who were each associated with the same house were therefore indirectly associated with each other, and on whether they could remember trained face pairs. The hippocampus was selectively activated during performance of the inferential judgment about indirectly related faces as compared to during memory for trained face-house or face-face pairings. In the other study, subjects learned a series of choice judgments between pairs of visual patterns that contained overlapping elements, just as in the studies on rats and monkeys, and as a control they also learned a set of non-overlapping choice judgments (Heckers et al., 2004). The hippocampus was selectively activated during transitive judgments as compared to novel non-transitive judgments.

These findings indicate that the hippocampal relational network mediates the linking of distinct episodes that may contain items that have not been experienced in the same episode or in the

same context. In doing so, the hippocampus plays a role in more than simply binding items within memories, but also mediates associations between distinct memories. During recollection, the hippocampus supports a capacity to generate multiple memories that share a common element, and the information contained within these memories can be used by many brain systems to make judgments about causal, logical, temporal, and spatial relations among the items in those memories (Cohen & Eichenbaum, 1993). Iterations of association, retrieval, and re-coding memories according to deduced relationships among the items would lead to the development of a systematic organization of items and episodes in memory wherein facts and events are linked to one another by a broad range of causal, logical, temporal, spatial, and other relevant relationships among the items. And this organization supports flexibility in the expression that is characteristic or recollective memory, specifically involving inferences between items that are only indirectly related.

2.2 The Anatomy of Memory

How do the above described memory functions emerge from the circuitry of the hippocampus? The brain system that mediates retrospective and prospective memory is composed of several cortical association areas interacting with structures in the medial temporal lobe (MTL), and in particular, the hippocampus. There is a general consensus that areas of the cerebral cortex are specialized for distinct aspects of cognitive and perceptual processing that are essential to memory, and that the cortex is the repository of detailed representations of perceptions and thoughts. The MTL is the recipient of inputs from widespread areas of the cortex and supports the ability to bind together cortical representations such that, when cued by part of a previous representation, the MTL reactivates the full set of cortical representations that compose a retrospective memory. Areas of the cortex both direct the storage of memories in the MTL and interpret the reconstructed memories generated by the MTL to support prospective memory. This simple, anatomically based scheme provides the framework on which our model is built. In the following sections, we will describe in greater detail the functional components of this system and the pathways by which information flows among them, and a qualitative model of how they interact to support retrospective and prospective memory.

The anatomy of the brain system that supports memory is remarkably conserved across mammalian species (Manns & Eichenbuam, 2007). Information processing in this system occurs

in three main stages. The first stage involves virtually every neocortical association area (Burwell et al., 1995; Suzuki, 1996). Each of these neocortical areas projects to one or more subdivisions of the parahippocampal region, which includes the perirhinal cortex, the parahippocampal cortex, and the entorhinal cortex. The subdivisions of the parahippocampal region are interconnected and send major efferents to multiple subdivisions of the hippocampus itself. Thus, the parahippocampal region serves as a convergence site for cortical input and mediates the distribution of cortical afferents to the hippocampus. Within the hippocampus, there are broadly divergent and convergent connections that could mediate a large network of associations (Amaral & Witter, 1989), and these connections support plasticity mechanisms that could participate in the rapid coding of novel conjunctions of information (Bliss & Collingridge, 1993).

The outcomes of hippocampal processing are directed back to the parahippocampal region, and the outputs of that region are directed in turn back to the same areas of the cerebral cortex that were the source of inputs to the MTL.

Only highly pre-processed sensory information reaches the MTL, but these inputs come from virtually all higher-order cortical processing areas. Perhaps the most thoroughly studied cortical area afferent to the hippocampus is the inferotemporal (IT) cortex, the highest-order visual object processor in primates. Ablation (removal of material from the surface of an object by vaporization or other erosive processes) of the inferotemporal cortex results in a visual-guided learning and deficits without impairment in visual fields, acuity, or threshold. The behavioral physiology of inferotemporal cortex is consistent with the data from ablation studies, showing that IT neurons are maximally driven by complex visual patterns, and the response properties of these cells are dependent on attentional mechanisms and reward association. Many IT neurons are preferentially responsive to a particular pattern, often one that is of obvious significance to the animal, including cells that respond selectively to faces. IT neurons respond differently to the same stimuli when they appear as stimuli to-be-remembered, or when they were novel versus familiar, and some cells maintain firing during the memory delay periods during performance of short term memory tasks. In humans, distinct ventral temporal areas that include and surround IT are activated by presentation of different categories of visual cues, including faces, tools, and animate objects (Martin, 2007; Kanwisher, 2007).

Other major inputs to the MTL arise from the posterior parietal area. Damage to this cortical area results in impairment in neglect of contralateral sensory stimulation across sensory modalities (Mountcastle et al., 1975; Andersen, 1989). One area within parietal cortex that has received particular interest is area 7a where most cells are visually driven. These cells have very large receptive fields and neuronal responsiveness is highly dependent on attentional factors. These cells respond best when the stimulus is the target of an eye or hand movement and they prefer moving stimuli but show little preference for stimulus form or color. These and other data indicate that the posterior parietal area is specialized for attention and egocentric spatial analyses including localization and visual and manual acquisition of targets in space. Also, areas of the parietal and temporal cortex are involved in complex perceptual processing essential to configuration of the conceptual contents of information that is the subject of recollection (e.g., Uncapher et al., 2006).

Additional major inputs to the MTL arise from several areas within the prefrontal cortex, a sensory-motor-limbic integration area involved in the highest-order cognitive functions including motor programming, vicarious trial and error, and memory (Fuster, 1995). In humans components of the prefrontal cortex mediate working memory, effortful retrieval, source monitoring, and other processing currently being specified that contribute critically to cognitive functions essential to recollection (Dobbins et al., 2002). In addition, midline structures within the prefrontal and cingulate cortical areas have been identified as activated during processing of self-referential information that may be strongly related to autobiographical memory (Northoff & Bermpohl, 2004; Fink et al., 1996; Cabeza & St Jacques, 2007).

The nature of cortical inputs to the MTL differs considerably across mammalian species (Manns & Eichenbaum, 2006). The proportion of inputs derived from different sensory modalities also varies substantially between species, such that olfaction (e.g., rats), vision (e.g., primates), audition (e.g., bats), or somatosensation (e.g., moles) have become disproportionately represented in the brain in different animals (Krubitzer and Kaas, 2005). Nevertheless, the sources of information derived from prefrontal and midline cortical areas, as well as posterior sensory areas, are remarkably consistent across species.

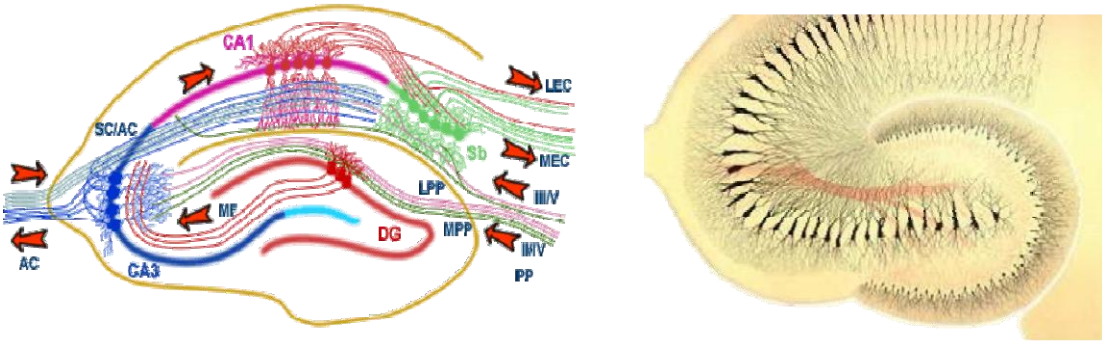
Despite major species differences in the neocortex, the organization of cortical inputs to the hippocampus is remarkably similar in rodents and primates. Across species, most of the

neocortical input to the perirhinal cortex comes from association areas that process unimodal sensory information about qualities of objects (i.e., “what” information), whereas most of the neocortical input to the parahippocampal cortex comes from areas that process polymodal spatial (“where”) information (Suzuki & Amaral, 1994; Burwell et al., 1995. There are connections between the perirhinal cortex and parahippocampal cortex, but the “what” and “where” streams of processing remain largely segregated as the perirhinal cortex projects primarily to the lateral entorhinal area whereas the parahippocampal cortex projects mainly to the medial entorhinal area. Similarly, there are some connections between the entorhinal areas, but the “what” and “where” information streams mainly converge within the hippocampus. The cortical outputs of hippocampal processing involve feedback connections from the hippocampus successively back to the entorhinal cortex, then perirhinal and parahippocampal cortex, and finally, neocortical areas from which the inputs to the MTL originated.

2.3 Towards a functional organization of a cortical-hippocampal memory system

The anatomical evidence reviewed above suggests the following hypothesis about how information is encoded and retrieved during memory processing. During encoding, representations of distinct items (e.g., people, objects, events) are formed in the perirhinal cortex and lateral entorhinal area. These representations along with back projections to the “what” pathways of the neocortex can then support subsequent judgments of familiarity. In addition, during encoding, item information is combined with contextual (“where”) representations that are formed in the parahippocampal cortex and medial entorhinal area, and the hippocampus associates items and their context. When an item is subsequently presented as a memory cue, the hippocampus completes the full pattern and mediates a recovery of the contextual representation in the parahippocampal cortex and medial entorhinal area. Hippocampal processing may also recover specific item associates of the cue and reactivate those representations in the perirhinal cortex and lateral entorhinal area. The recovery of context and item associations constitutes the experience of retrospective recollection.

Figure 4.2. The Hippocampal Structures of the Brain



Perirhinal cortex and lateral entorhinal area. Substantial evidence indicates that neurons in the perirhinal cortex and lateral entorhinal cortex are involved in the representation of individual perceptual stimuli. Electrophysiological studies on monkeys and rats performing simple recognition tasks have identified three general types of responses (Brown & Xiang, 1998; Suzuki & Eichenbaum, 2000). First, many cells in these areas exhibit selective tuning to memory cues such as odors or visual stimuli. Second, some cells maintain firing in a stimulus-specific fashion during a memory delay, indicating the persistence of a stimulus representation. Third, many cells have enhanced or suppressed responses to stimuli when they re-appear in a recognition test, indicating involvement in the recognition judgment. Similarly, in humans, among all areas within the medial temporal lobe, the perirhinal area selectively shows suppressed responses to familiar stimuli (Henson et al., 2003). Complementary studies in animals with damage to the perirhinal cortex indicate that this area may be critical to memory for individual stimuli in the delayed non-matching to sample task in rats (Mumby & Pinel, 1994; Otto & Eichenbaum, 1992) and monkeys (Suzuki et al., 1993). These and other data have led several investigators to the view that the perirhinal cortex is specialized for identifying the memory strength of individual stimuli (e.g. Brown & Aggleton, 2001; Henson et al., 2003; Aggleton et al., 2004).

Parahippocampal cortex and medial entorhinal area. The parahippocampal cortex and medial entorhinal area may be specialized for processing spatial context. Whereas perirhinal and lateral entorhinal neurons have poor spatial coding properties, parahippocampal and medial entorhinal neurons show strong spatial coding (Burwell & Hafeman, 2003; Hargreaves et al., 2005). Further, the immediate early gene *fos* is activated in perirhinal cortex by novel visual cues, but *fos* is activated in the postrhinal cortex by a spatial re-arrangement of the cues (Wan et al., 1999). In

addition, whereas object recognition is impaired following perirhinal damage, object-location recognition is deficient following parahippocampal cortex damage in rats (Gaffan et al., 2004) and monkeys (Alvarado & Bachevalier, 2005). Similarly, perirhinal cortex damage results in greater impairment in memory for object pairings whereas parahippocampal cortex lesions results in greater impairment in memory for the context in which an object was presented (Norman & Eacott, 2005). Parallel findings from functional imaging studies in humans have dissociated object processing in perirhinal cortex from spatial processing in the parahippocampal cortex (Pihlajamaki et al., 2004). Furthermore, whereas perirhinal cortex is activated in association with the memory strength of specific stimuli (Henson et al., 2003), the parahippocampal cortex is activated during recall of spatial and non-spatial context (Ranganath et al., 2003; Bar and Aminoff, 2003).

Hippocampus. Compelling in support for differentiation of functions associated with recollection come from within-study dissociations that reveal activation of the perirhinal cortex selectively associated with familiarity and activity in the hippocampus as well as parahippocampal cortex was selectively associated with recollection (Deselaar et al., 2006; Davachi & Wagner, 2002; Davachi et al., 2003; Ranganath et al., 2003). These and many other results summarized in a recent review suggest a functional dissociation between the perirhinal cortex, where activation changes are consistently associated with familiarity, and the hippocampus and parahippocampal cortex, where activation changes are consistently associated with recollection (Eichenbaum et al., 2007). An outstanding question in these studies is whether the parahippocampal cortex and hippocampus play different roles in recollection. In particular, the above described findings on parahippocampal activation associated with viewing of spatial scenes suggests the possibility that this area is activated during recollection because recall involves retrieval of spatial contextual information. By contrast, the hippocampus may be activated associated with the combination of item and context information.

CA1 versus CA3. Several recent studies have suggested that subregions of the hippocampus may play distinct roles in memory. A particularly striking contrast comes from a comparison between two studies by Kesner and colleagues (Gilbert and Kesner, 2003; Kesner et al., 2005). In one experiment, normal rats learned associations between a particular object or odor and their locations in specific places in an open field. On each trial, one of two objects (differentiated by visual or olfactory cues) was placed at one of two locations on a large open field. If object A was

in place 1, a reward could be found underneath. Similarly, if object B was in place two a reward could be obtained by displacing the object. However, no reward was available if either object was presented in the alternate location. Normal animals improved in performance across days, as reflected in differentiating their latencies to approach object in rewarded vs non-rewarded locations. Selective lesions of CA3 completely blocked acquisition of object-place associations, whereas CA1 lesions had no effect. In contrast, the opposite pattern of results was found in another study were rats were taught associations between an object and an odor that were separated by a short delay. The animals learned that if object A was presented before the delay, then a cup of sand would contain a food reward if it was scented with odor 1 (but not with odor 2). Conversely, if object B was presented first, then a cup of sand would contain a food reward if it was scented with odor 2 (but not odor 1). Memory was measured by a briefer latency to approach the scented cup on rewarded pairings (A-1 and B-2) than on non-rewarded pairings (A-2 and B-1). In normal rats, the latency to approach rewarded cups gradually decreased over daily training sessions, at about the same rate as observed in the previous object-place association study. In contrast, rats with selective CA1 lesions showed no sign of acquiring the associations between temporally separated objects, whereas rats with CA3 lesions acquired the task just as rapidly as normal animals. These results are consistent with the idea CA1 is specialized for representation of the order of events that are separated in time (Manns & Eichenbaum, 2006).

3. CURRENT COMPUTATIONAL REPRESENTATION

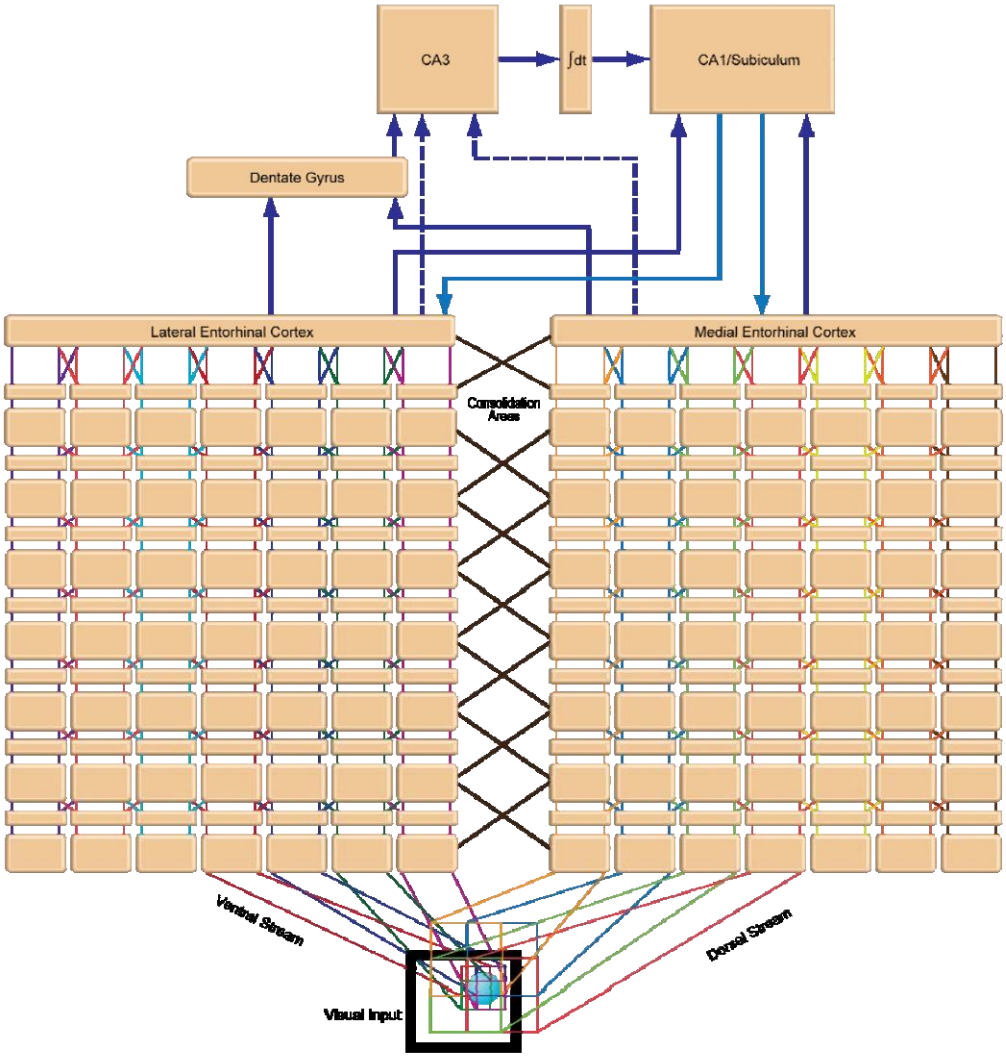
3.2. General Description of Architecture

For this LDRD project we focused on higher-level processing, occurring after eye foviation and movements such as saccades. The model starts at the point where the visual input images have been separated into two components. The first sub-image corresponds to the area seen by the focal area of the eye, and the second sub-image contains the entire field-of-view for the eye. This division models the higher resolution present in the fovea, as well as the way focus and context information are treated separately through some parts of biological cortex. We will begin the description of the computational system by describing the bottom-up behavior involved in encoding episodic memories into the representation of the system. As illustrated in Figure 4.3, the first sub-image (the blue sphere at bottom of image) is directed to the ventral

stream of neural processing where object detection and categorization is handled (i.e., “what” information). The second sub- image (the upper-right portion of the visual input) is directed to the dorsal stream where the spatial context of what the eye is seeing is determined and categorized (i.e., “where” information). This stream consists of a lower resolution (zoomed out) view of the entire field of view including the focal area itself. Note that both sub-images are further segmented into overlapping sub-sections for even greater specificity in category formation of our episodic memory.

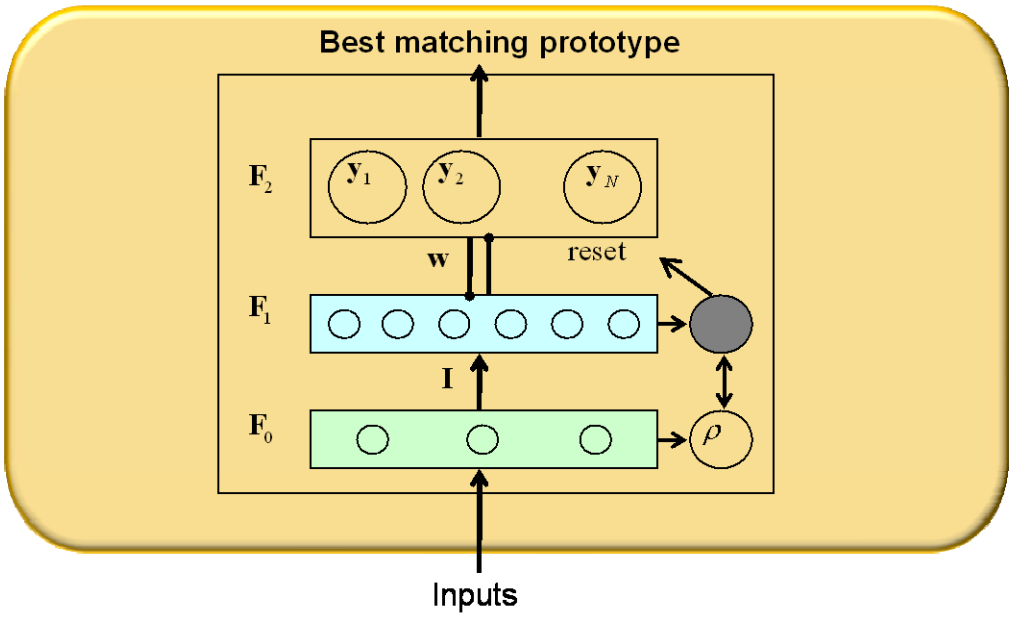
Our cortical model is comprised of stacked layers of fuzzy Adaptive Resonance Theory (ART) networks. ART is a well established self-organizing neural technique for classifying input activations. The interested reader can find a wealth of literature on ARTs details, performance, and stability (see Vila, 1994; Walczak, 2005). While ART is a good choice for the classification modules, other unsupervised learning techniques could be expected to render similar results, though probably with subtle and interesting differences. Between each pair of ART modules there is a layer of Temporal Integrators (Taylor et al., 2009), with an adjustable time constant of integration for each layer. This TIART network is meant to model a biological cortical column, which receives afferent connections from either a particular subsection of the input field or a particular subsection of the previous layer. These subsections overlap in a manner inspired by the biological cortex (Kingsley, 2000). The system as a whole is a simple but powerful cortical classifier. Progressively higher layers encode progressively more abstract objects or spatial locations. Low levels correspond to simple perceptual primitives (edges etc.), high levels might correspond to whole objects or other semantic concepts. However, the cortical ART networks have an interesting modification. Each ART network has a "top down" recall mode driven by input from higher-level, more abstract layers as well as the traditional "bottom up" mode driven by stimulus input. While in "top down" mode, an activated F2 node in the ART network reinstates the prototypical input pattern which it encodes in synaptic weights between layers F1 and F2 (see Figure 4.4). So, for example if a particular F2 node encoded the concept of "dog" it could read out the features ("fur," "tail," "barks," etc.) in the F1 layer. These features would in turn activate the F2 layer of the next lowest ART module and so on. In this way the network learns new inputs from the bottom up, and can then recall and reconstruct these features from the top down.

Figure 4.3. A conceptual view of the computational architecture



In our version of the cortex, one half of the cortical input columns receive high resolution input from the center of the visual field. This causes the network to develop templates which corresponds to "objects" and is meant to simulate the fovea near the sensory level and the ventral visual stream near the associative levels. The other half of the cortical columns receive low resolution input from the periphery which leads the development of spatial representations. This is meant to correspond to the off-center visual field at low levels and the dorsal visual stream at associative levels.

Figure 4.4. An example of an Adaptive Resonance Theory (ART) module



For each cortical modality (i.e., focus and context) there are three levels of cortex. For each level of cortex, we have implemented a grid of fuzzy ART modules (the current version has 6 layers at 7 X 7 modules; see Appendix 1). There exist temporal integrators between the fuzzy ART levels, but for these experiments the time constant of integration is dialed down to where they are inconsequential from the point of view of pasting together temporal events. The temporal integrators are dialed down because the experimental tasks are static in nature rather that temporal. A future extension of this architecture would be a single mechanism that dynamically adapts both static and temporal tasks.

3.1.1 The Hippocampus

The hippocampal system makes use of several ART variations, so it is related to the cortical system at a single unit level, but has a very different architecture and accomplishes fundamentally different information processing. While the cortex attempts to represent the conceptual structure of its inputs, the hippocampus attempts to quickly bind snapshots of high level cortical activity. Behaviorally, this gives us an episodic memory mechanism where concepts originating in multimodal sensory input are bound together. By way of this binding, the

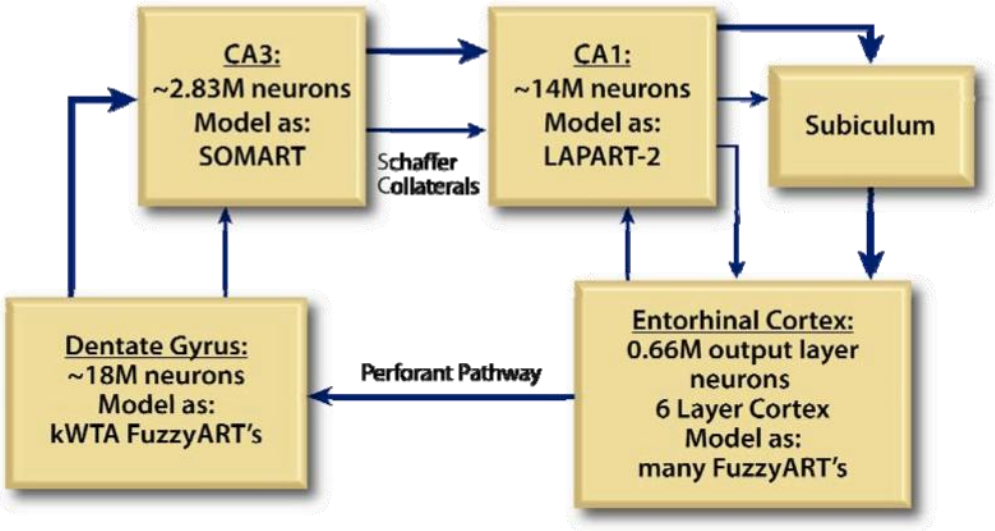
hippocampal representation can also be used to recover neocortical representations from partial activations.

Hippocampus is modeled as a loop of neural modules starting at entorhinal cortex, proceeding to dentate gyrus, continuing to CA3, then returning to entorhinal cortex through CA1, where some of the function of subiculum is implicitly captured in CA1. Entorhinal Cortex (EC) is the last level of cortex on the way from sensory input to the hippocampus. The EC is where all information that will be encoded in episodic memory

must converge. The Dentate Gyrus (DG) provides a pattern separation function for the information received from entorhinal cortex. The CA3 provides pattern completion and semanto-spatial association. CA1 closes the loop and provides temporal association.

The EC is an area of multi-modal convergence, where several data streams from different senses come together (Anderson, et al., 2007). The EC's cytoarchitechture resembles that of the cortex, so it here is modeled as is the rest of cortex. However, the ART networks which make up the EC have two sets of connections to the hippocampus, a feed forward connection to the DG and CA3 meant to simulate the perforant path, and bidirectional connections to the CA1/Subiculum component of the hippocampal model. The forward connections provide inputs to the hippocampal module. The back connections use LAPART rules to learn associative links between activity in CA1 and EC, thereby closing the autoassociative hippocampal loop. When a CA1 representation is activated, these back connections can drive top-down cortical recall (Figure 4.5).

Figure 4.5. The modeled representation of the hippocampal system



In our implementation, the pre-Medial Temporal Lobe (MTL) sensory cortex and EC are represented by layers of fuzzy-ART modules which are modified to encode temporal semantic data. Individually, these temporally integrated adaptive resonance theory (TIART) modules are capable of encoding categorical representations of their given input vectors over time (Taylor, et al., 2009). By combining layers of TIART modules, our EC creates categories of categories to represent larger semantic concepts and combine the "dorsal stream" containing contextual information and the "ventral stream" of focal information before these streams enter the hippocampus. Within the hippocampal representation in our model, each of the primary regions is represented by a different ART variant selected to achieve the particular functionality of the individual region. The relative size of each module is scaled in accordance with approximate human neuroanatomy.

The DG has peculiar anatomical properties. It has a large number of neurons with relatively low activity and it is one of the few places in the brain in which new neurons are generated in the adult brain. These properties have lead to the suggestion that the DG creates sparse, non-overlapping codes for unique events via pattern separation (Leutgeb, et al., 2007).

The DG in our model receives the conjoined multimodal sensory signals from EC. It performs pattern separation on this abundance of sensory information to produce sparse output activation, which ensures different semantic concepts are given unique encoding (Rolls &

Kesner, 2006). Computationally, a series of k-winner-take-all (k-WTA) fuzzy- ART modules constitute the DG module of our model. A WTA module is a competitive network in which a single concept beats out competing concepts to represent the input vector. Effectively, a sparse encoding is created as each of the k WTA modules yields a single output. Similar input vectors will be represented by the same single winning output, and dissimilar inputs will be represented by a differing winning output, yielding pattern separated outputs. These outputs serve as the input for CA3.

Anatomical studies of the hippocampus proper reveal cytoarchitechture which differs radically from that of the cortex (Anderson, et al., 2007). While both CA1 and CA3 both contain pyramidal cells like the cortex, existence of extensive recurrent connections in CA3 and the presence of inhibitory and excitatory interneurons have led some investigators to suggest that CA3 may be involved in pattern completion (O'Reilly & Rudy, 2001).

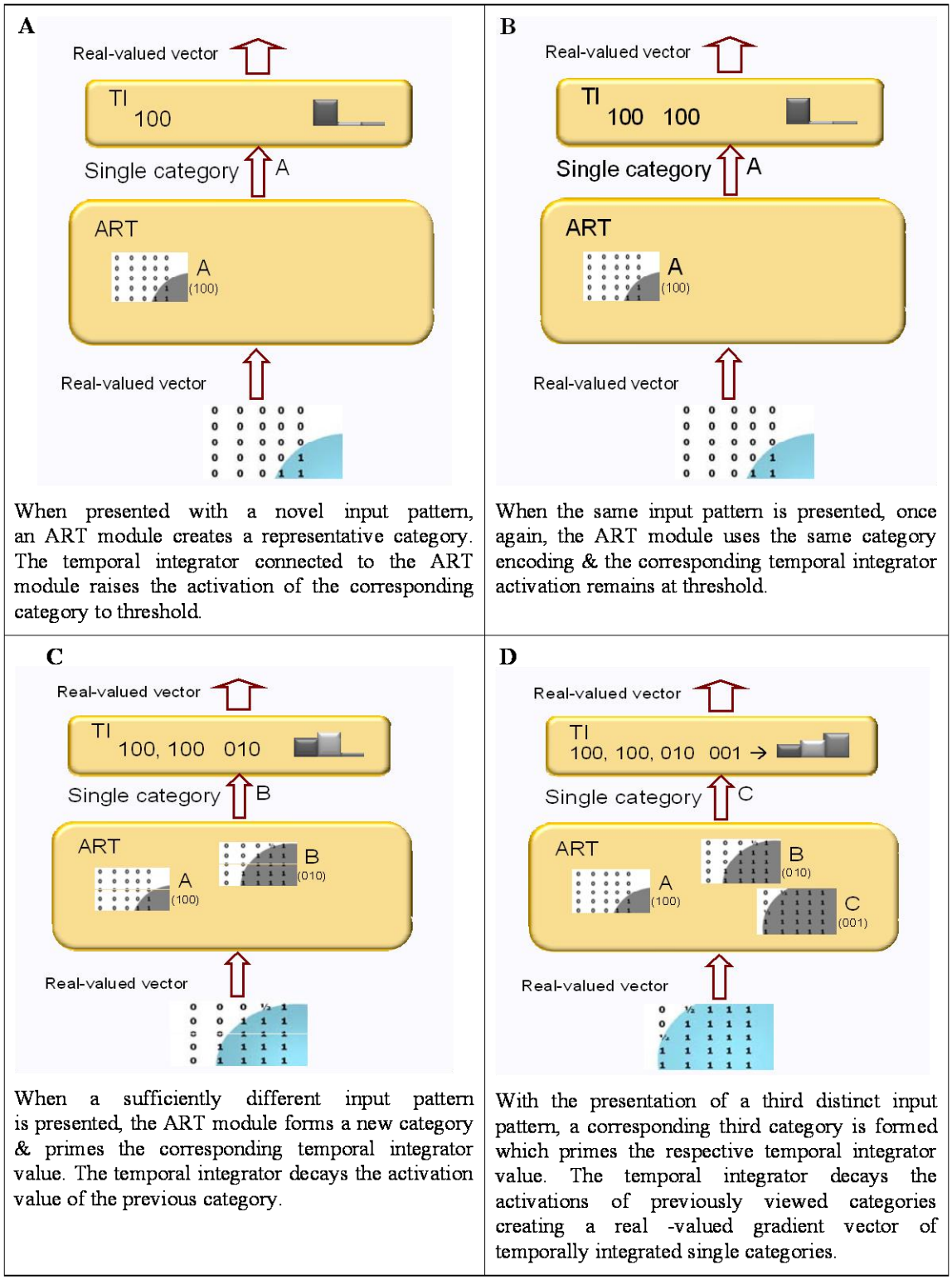
In this implementation the sparse output pattern from DG serves as input to CA3. Functionally, CA3 assists with episodic binding through auto-association. In our model, this functionality is represented by a Self-Organizing Map (SOM) structure. A standard SOM transforms a given input vector into a distinct topological region without supervision guiding the classification (Haykin, 1999). Incorporating the neighborhood updating capabilities of a SOM within a fuzzy-ART module, we have created a SOMART module to represent CA3. This module is capable of mapping semantically similar inputs to proximate topological regions. Thus the learning algorithm creates "islands" of activity which respond to similar input sets, but avoids a global topology. In effect, related concepts are clustered together to help associate episodic memories and these "islands" of relational bindings form the inputs to CA1.

Anatomically, the output of CA3 proceeds to CA1 and then to the subiculum as the major output region of the hippocampus. However, the exact functionality of the subiculum is largely unknown, so we have merged the capabilities of CA1 and subiculum in our model. CA1 has been implicated in learning relational information for temporal sequences and connecting these episodic encodings back to the original sensory inputs from EC. This ability to link sequences allows for temporal packaging of episodes. Since our CA3 can only encode momentary conjunctions, we need a mechanism which can capture sequences of changing relations. Thus, CA1 contains a unit which temporally integrates CA3 outputs using a set of leaky integrators.

This provides a temporal gradient of input conjunctions coming from CA3, the oldest bindings will have the weakest signal in the temporal integrator, while the most recent bindings will be most strongly represented. This temporally coded sequence of CA3 activity is used by CA1 to create a topology of temporal sequence.

Once this temporal topology has been established, activity in CA1 is associated with activity in the EC via a Laterally Primed Adaptive Resonance Theory (LAPART) partially-supervised learning paradigm. Local CA1 learning is supervised in that a certain sequence of CA3 activations corresponds with certain EC activation. CA3 sequence *A,* where that sequence is translated to an instantaneous representation through the temporal integrator, is bound through learning to EC activation *B.* LAPART uses two ART modules connected by a lateral activation field, so the activations on each side are generalized via the ART classification mechanism. Through experience, a connection weight is learned to bind the node that corresponds to each classified CA3 sequence to a node that corresponds to some EC activation. This mapping of sequences onto the high- level cortical representations closes the hippocampal loop, and allows activations in CA1 to cue top-down recall in the cortex and unspool the temporal representations it has created. This entire process is graphically described in Figure 4.6 (A-K).

Figure 4.6. Step-by-step process of episodic/semantic activation



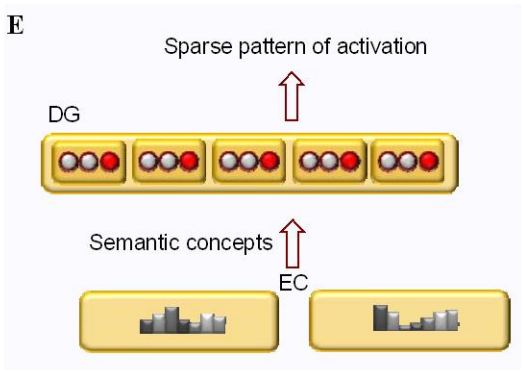
When presented with a novel input pattern, an ART module creates a representative category. The temporal integrator connected to the ART module raises the activation of the corresponding category to threshold.

When the same input pattern is presented, once again, the ART module uses the same category encoding & the corresponding temporal integrator activation remains at threshold.

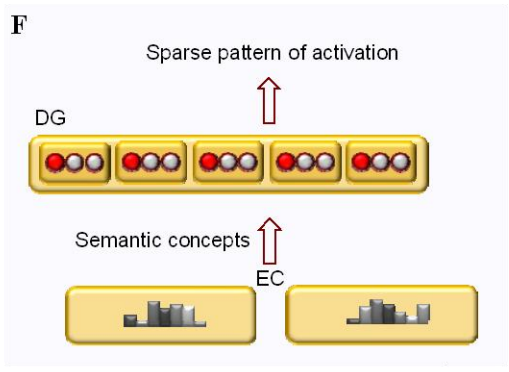
When a sufficiently different input pattern is presented, the ART module forms a new category & primes the corresponding temporal integrator value. The temporal integrator decays the activation value of the previous category.

With the presentation of a third distinct input pattern, a corresponding third category is formed which primes the respective temporal integrator value. The temporal integrator decays the activations of previously viewed categories creating a real -valued gradient vector of temporally integrated single categories.

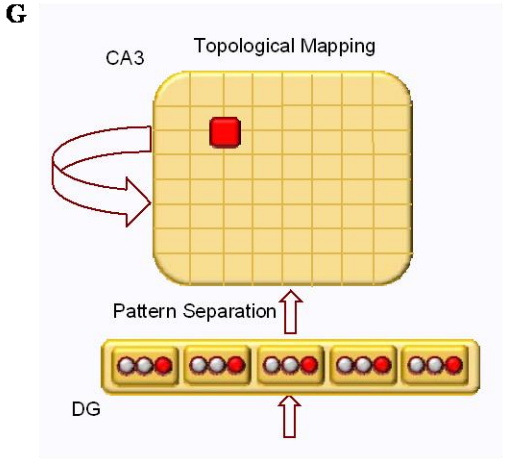
Figure 4.6 Cont.



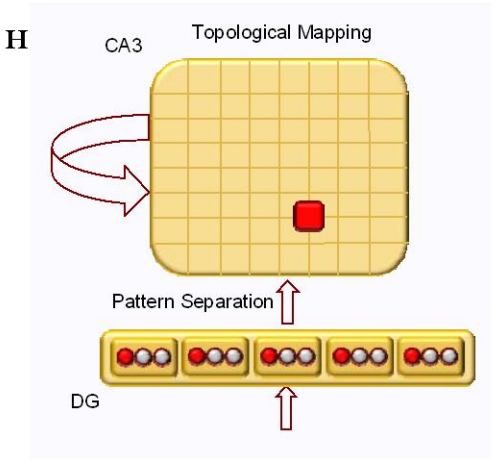
Layers of sequential ART modules and temporal integrators allow for the creation of categories of categories, which further allow for the formation of semantic concepts in the EC as the convergence of sensor modality input patterns. Multi-modality semantic concepts from EC are passed as inputs into the DG which yields a sparse pattern of output activation.



Unique DG inputs create distinctly different sparse activation patterns in effect performing pattern separation.

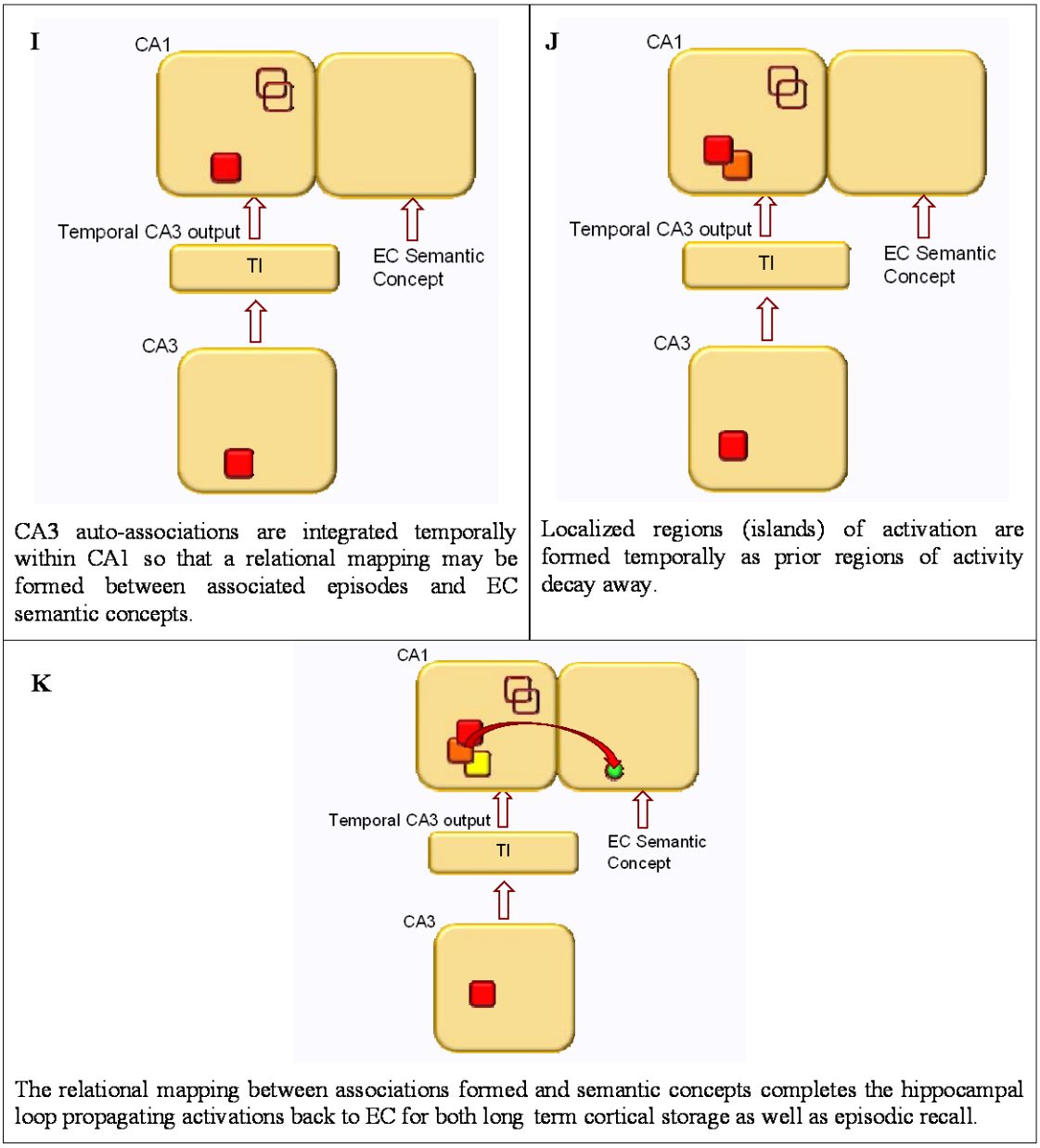


The separated sparse activations serve as inputs to the highly recurrent CA3 hippocampal region which processes these inputs performing an auto-associative mapping.



Unrelated sparse activations are mapped to distinctly different CA3 regions, and related patterns are mapped in close proximity to each other facilitating association of related concepts.

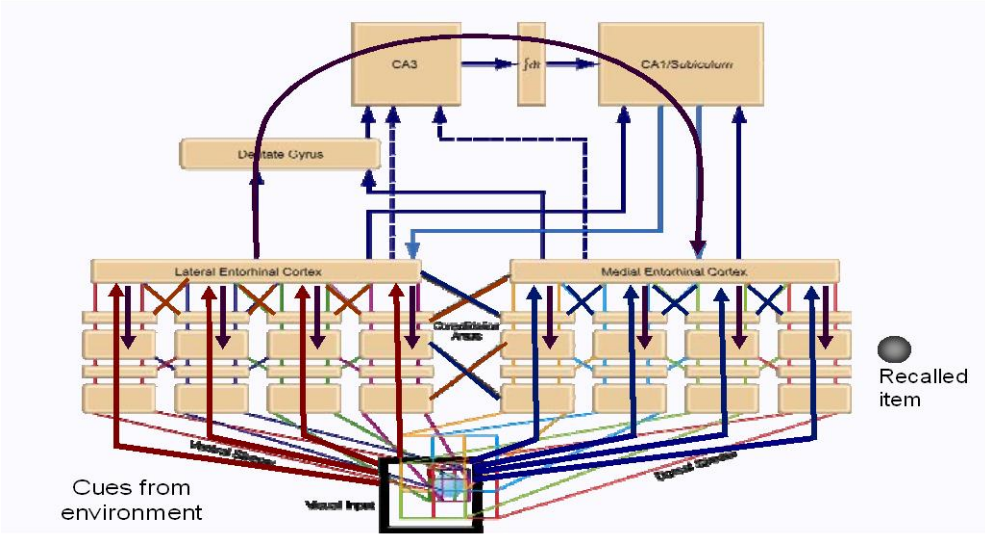
Figure 4.6 Cont.



In our computational system, when an episodic memory is recalled, the hippocampus activates one of the categories in CA1, and it begins the top-down recall of the episodic trace which it encodes. In the recall, the active category in the CA1 ART unit encodes a temporal sequence of conjunctive associations which were formed in CA3 during the episodic memory formation. During recall of a specific conjunctive association in CA3, each stream (ventral and dorsal) that makes up part of the conjunctive association is activated in top-down fashion. The top-down activation of each stream is similar, thus the top-down description of a single stream will be

given in the interest of brevity. In the top- down activation of a cortical stream (either “what” or “where”), the top-level ART category is activated (in the EC or DG). This top-level category consists of a concatenation of ART categories from each column in the cortical stream. Top-down recall of this concatenation of categories consists of a simultaneous recall of each column starting at the ART unit which is at the top of each column. It is this ART unit which provides its category for the concatenation during the formation of the episodic memory. The top-down recall for each column is similar, thus the description of a single column will be given in the interest of brevity. During recall of a cortical column, the category in the top-most ART unit is activated in top-down fashion. This category contains a temporal sequence of categories from the next lower ART unit as integrated through the integration unit between them. The recall continues from ART unit to ART unit downward through the connecting temporal integration units until the bottom ART unit is reached. In top-down activation, each temporal integration unit contains a temporal sequence of ART category activations which were fed to it as input during episodic memory formation. During recall of the temporal sequence of ART categories in a temporal integration unit, each category in the sequence is re-activated in top-down fashion in the same temporal order as was originally experienced and encoded in the episodic memory trace. When the recall reaches the lowest level ART unit, it is ready for “replay.” During replay, the memory is re-activated in forward or bottom-up fashion in the same temporal sequence it was originally experienced. In this system, temporal information is stored in the activation potentials of temporal nodes. Local semantic information is stored in the synaptic weights of the ART modules. Long-term, memory can occur through Hebbian-like adaptation of synaptic connection weights between local cortical areas. In our model, a local cortical area at a given level is comprised of a collection of nodes that all influence the activation of the same ART output node.

Figure 4.7. The bottom-up/top-down flow of information that both consolidates memory into LTM and produces recall when prompted.



4. ASSESSMENT OF THE MODEL

Assessing a computational model for the degree of neuro-cognitive plausibility is a significant challenge. No one qualitative or quantitative method is sufficient to adequately evaluate the level of agreement between a computational model and the analogous brain system it seeks to represent. Thus, we employed several evaluative methods that were based on empirical, human studies as a means to quantitative compare the model. These comparisons helped to qualitatively assess the accuracy of the model, whereby the more qualitative comparisons, the greater potential for an accurate assessment of the model. Each comparison described below was meant to address a key aspect or function of episodic memory.

4.1 Temporal and Sequential Memory and Recall of Objects

This section details our approach to assessing the model's learning of temporal semantics. First, the method of assessment is described and then a step-by-step example of its operation is given. Finally, the experimental method of assessment is demonstrated with a discussion of the results.

This explanation and the first experimental assessment have been presented in Taylor et al.,

2009. We encode temporal semantic data as a recency gradient of generalized classifications. Fuzzy ART is used as the classifier that creates an active output on a certain *F2* node for any given input vector. We implemented an ART network with a fixed number of *F2* nodes available for recruitment, where each available *F2* node (whether it had yet been recruited or not) is connected to a leaky temporal integrator. This implements a static architecture, which is convenient for explanation. The method might be extended to dynamic architecture creation (allowing ART to recruit new *F2* nodes indefinitely) in modeling neuro-development. The integrators need be leaky, otherwise the output would continually increase over time (assuming continual input greater than zero), eventually saturating. A general leaky integrator is modeled as (Carpenter, Grossberg, & Rosen, 1991). Where *y* is the integrator output, *x* the input, and *a* an integration constant.

$$\mathrm{dy/dt} = -ay + x \qquad (4.2)$$

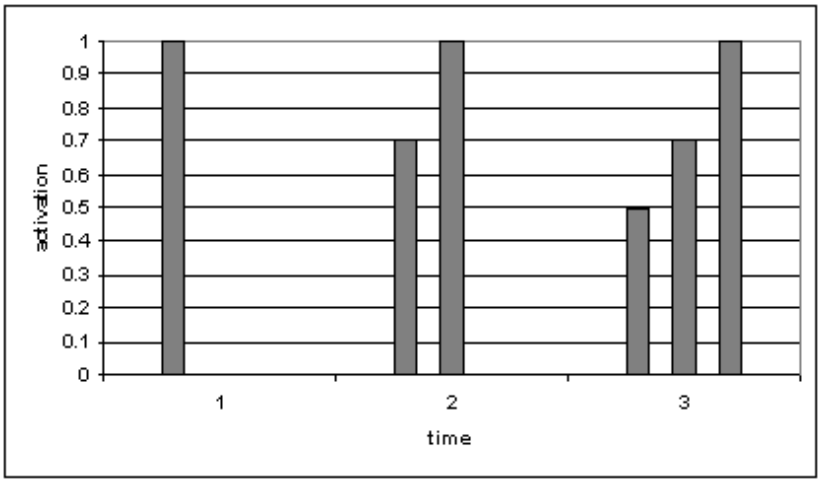
$$y(n+1) = (1-a)y(n) + x(n) \qquad (4.3)$$

We implement a discretized (using Euler's method with an arbitrary sample period of one time unit) version of the leaky integrator (4.2) as formula (4.3). Where 1−a is a decay constant. This method is simple enough in concept that we can provide a comprehensive example of its operation. The following example assumes three arbitrary sensory input vectors. We use sensory inputs to create grounded, stand-alone, examples. However, if the sequence of inputs over time were internal cortical activation patterns, the example could be describing an additional functional level of cortex, above that which created the inputs.

Let there be three distinct sensory input activation vectors *A, B, C* that form a temporal semantic sequence that we wish to encode. Inputs feed into an ART module (see Figure 9). Each input results in a different active node output on the ART. By placing a leaky temporal integrator on each ART output node, we encode a temporal sequence of inputs as a single real valued vector. The integrated vector is a recency gradient, where the order of element amplitudes (from low to high) represents the order of occurrence of the input vectors (from oldest to most recent). As mentioned, the value of a given integrator output node will decrease over time. As a result, the farther in the past a given input was observed, the smaller a value the corresponding integrator output will have (until at some small activation level, the integrator output is lost in the noise of the system). A more biologically faithful sequence recall scheme (Sun & Giles, 2001), could

involve an extra step where by leaky integrator activations provide an inhibitory signal to other nodes, such that highest activations could be ordered first in temporal sequence.

Figure 4.8. Time series of temporal integrator outputs given example input: "A", "B", "C"



If the first system input is a semantic concept *A* (the specific vector of input data is not important, the simplest assumption would be that the semantic concept *A* is represented by an array of pixel values that visually correspond to the letter *A),* then let the output of the ART be [1 0 0], considering only the first three outputs for simplicity. Likewise, let the ART outputs corresponding to inputs *B* and *C* be [0 1 0] and [0 0 1] respectively. The temporal integration array initializes to [0 0 0]. Figure 8 illustrates the temporal sequence of integrator node outputs that result from presenting *A* at timestep 1, *B* at timestep 2, and *C* at timestep 3.

[0.5 0.707 1.0] (4.4)

Finally then, the temporal input sequence *A, B, C* is encoded at a single point in time as vector (4.4). Ascending values in the vector indicated encoded temporal order, while connectivity to the rest of the architecture encodes semantic content. In the context in which it was formed, and to the level of detail that the ART categories have formed generalized templates encodes both the temporal and semantic information of the input sequence (Sun & Giles, 2001).

Figure 4.9. ART with temporal integrator

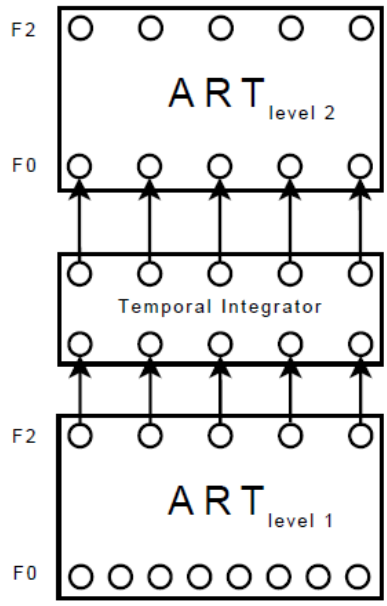
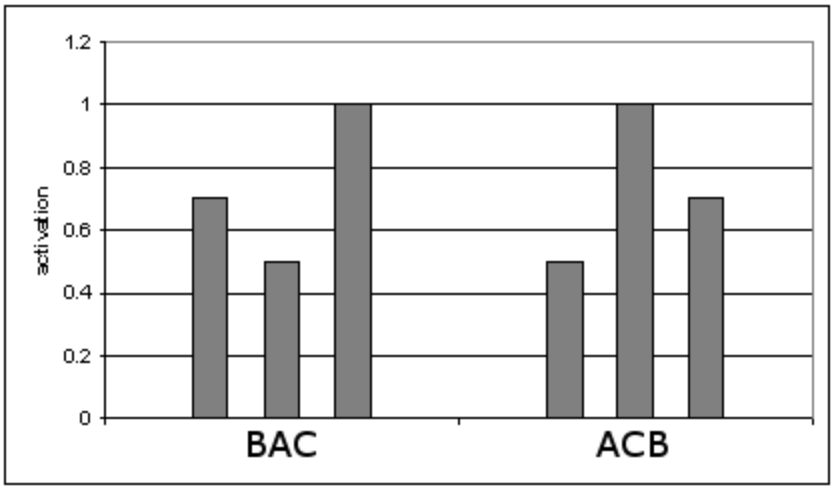


Figure 4.10. Example temporal integrator outputs



Vector (4.4) can now be encoded by a level 2 ART (see Figure 9) to uniquely represent the sequence *A, B, C.* Figure 10 illustrates some other possible temporal integrator outputs, given the indicated input sequence. Level 2 ART could encode these other vectors as representations of the corresponding temporal semantic sequence. Because the level 2 ART representations are unique (to the level of precision determined by the ART operational parameters), a top-down recall operation can recover the original sequence *A, B, C* from the level 2 ART encoding of (4.4). For example, let the temporal integrator output (4.4) result in activation of the first F2 output on

level 2 ART. Then, given an augmented ART with top-down recall ability, as well as a playback mechanism for the temporal integrator, top level stimulation of the first F2 node on level 2 ART will result in the sequence *A, B, C* being played back at the system input level.

Later work will delve into neural implementations of these top-down recall mechanisms. A brief overview of possible mechanisms is given now, so that their functionality can be used in the experiments that confirm temporal semantic encoding has occurred. We supplement ART by specifying a top-down behavior. When an *F2* node is stimulated from above, it plays down the associated template's activation levels to the ART input layer. Template activation level adjustment is where all memory storage in ART occurs, so playback of a given template represents recall of one memory component at the scale of that ART unit.

We also supplement the bottom-up temporal integration scheme with a top-down behavior. As the output of a temporal integrator array is a pattern encoding the order of input activation, recall of that gradient should play back the temporal integrator inputs in that order. When a pattern is placed on the temporal integrator array from above, the array will first activate the input corresponding to the lowest value in the pattern, then the next lowest, and so on. This behavior will play back the input activations in their original order. As a manner of implementation, we can imagine the top-down stimulation of the temporal integrator array as setting a threshold for each element of the array. The integrators then start integrating up from zero and fire the associated input node when their internal value reaches the threshold. The lowest threshold will be reached first, which is correct because it represents the input that occurred farthest in the past and therefore the input that should be played back first.

4.1.1 Experimental Results

This experiment demonstrated successful encoding of temporal semantic data. In this case, the semantic meaning will be visual sensory observation. The formation of the encodings themselves can be tracked by probing node activations in the architecture. However, this can at most show that some representations were formed (not that those representations are correct). We demonstrate that the representations are valid encodings of the input information by initiating a recall process that decodes the temporal semantic representations into whatever those representations encode in the context of the system. If the encoding, in the context in which it is stored and recalled, decodes to the original information (to the resolution of the

system), then we have solid evidence that the desired functionality is captured by the described method.

Figure 4.9 shows the architecture used for the recall experiment. The visual sensory input is a two dimensional array of 10 by 10 pixels. The first level ART forms templates to represent the visual input symbols. The temporal integrator forms semanto-spatial patterns to represent sequences of symbols observed by the first layer ART as passed on by the first layer ART's *F2* node activation. The second layer ART forms templates that represent the patterns output by the first layer temporal integrator array that represent temporal sequences of input symbols. In this case, the timescales are such that the second layer ART captures temporal integrator patterns corresponding to three input timesteps. Figure 4.11 shows an input sequence. Figure 4.12 shows one recall sequence, in this case recall corresponds to the middle three symbols of the input.

Figure 4.11. Memory and recall input sequence

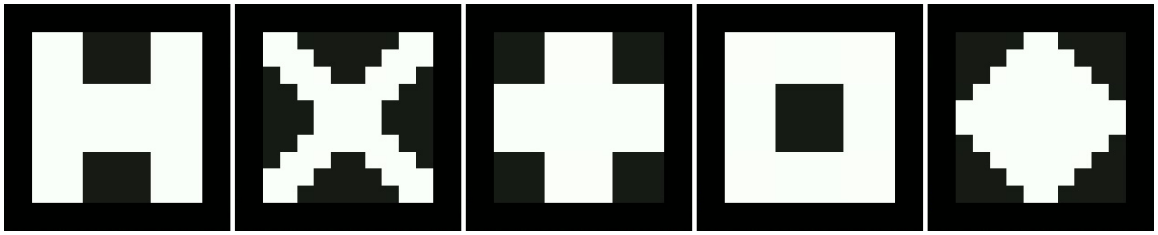
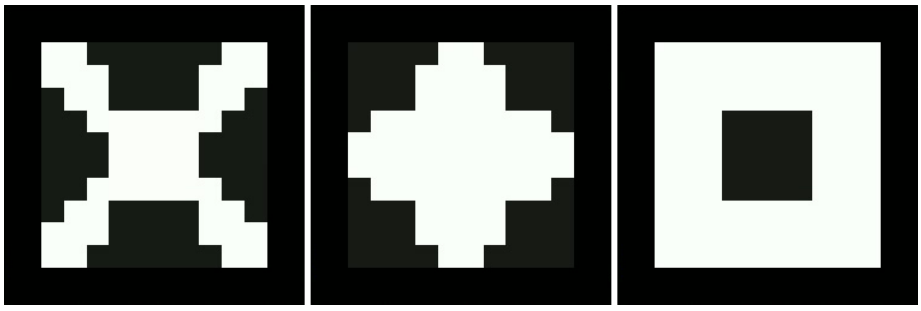


Figure 4.12. A sample memory and recall output sequence



We note that the recalled sequence is not an exact copy of the input sequence. The plus and diamond symbols have been aliased to a common symbol. This aliasing occurs because in the metric used by ART, the plus and diamond symbols are sufficiently close for classification to the same *F2* template. An easy parallel can be drawn in human memory formation. To the non-

expert, transient observation of either a grey Toyota Corolla or a grey Honda Accord is likely to form an aliased memory of "grey import sedan." In ART, the vigilance parameter determines how close two inputs must be to be classified into the same template.

The vigilance parameter was set arbitrarily at 0.8, in a range of 0 to 1, for the described experiment. In our architectures, the vigilance parameter of the ART units is an independent variable that can be used to tune the performance of the network. One of the trade offs inherent in tuning with the vigilance parameter is memory space vs. precision, greater vigilance will form a greater number of more precise templates. We can tune a part, or all, of this sort of architecture for more precise (higher resolution) temporal semantic encoding, but only at the expense of using up more memory space.

As stated earlier, the temporal semantic encoding method is capable of encoding information to the precision of the generalization used in the ART module(s). This experiment demonstrates that valid temporal semantic encodings were formed because when the encodings are decoded, the original temporal and semantic information is recovered.

In addition to the limitations of the ART parameters used, the user would want to consider supplemental mechanisms (possibly such as those found in (James, 2001)) if exact encoding of sequences is important. The anticipated use of this method is the simulation of biological cognitive processes, so some deficit in perfect memory is acceptable. Further research can characterize the consequences of potential imperfections in the system, as compared to imperfections in human cognition. Research shows better relative memory performance, as opposed to absolute memory performance in humans (Sejnowksi & Rosenberg, 1987).

Another consideration for future applications is the packaging of input episodes. The above example uses static episode size for simplicity of analysis. A more interesting system would dynamically package episodes, possibly based on rate of change and/or novelty.

$$V_m = \frac{RT}{F} \ln \frac{P_K[K^+]_o + P_{Na}[Na^+]_o + P_{Cl}[Cl^-]_i}{P_K[K^+]_i + P_{Na}[Na^+]_i + P_{Cl}[Cl^-]_o} \quad (4.5)$$

There exists a biological correlate, in the behavior of neurons, to the sort of temporal integration described here. Membrane potentials integrate down (assuming sub-threshold

stimulation to some level above resting equilibrium) over time due to leakage current through resting ion channels. The Goldman equation (Vila, 1994), describes the influence of ionic concentrations on the neuron membrane potential (Waibel, et al., 1989). An elevated membrane potential would reflect deviation from an equilibrium (where the Goldman equation expressed that equilibrium for the pertinent ions in neurons) of ionic concentrations between the inside and outside of the neuron. Ions would then flow through resting (non-gated) ion channels until equilibrium was restored. The time course of this equilibrium restoration can be described by temporal integration.

The rate of change of the membrane potential is a function of the number of resting channels and the number, connection strength, and activity of afferent neural connections. We abstract beyond gated ion channels and outgoing action potentials as we only seek to explain neural plausibility rather than a full model of neuron. With proper balance and biasing, the rate of change in membrane potential (and hence the rate of decay of the integration) could be tuned over a wide range. This tuning allows the arbitrary time scale representation that we mentioned earlier.

We illustrate the biological correlate both to show neurological plausibility of our technique, but also as evidence for pervasive temporal integration (and thereby pervasive co-encoding of both temporal and semantic information at multiple levels through cortex). No neurons function without some form of temporal integration, though a counter argument is that the time scale of ion-channel temporal integration is not relevant to information encoding.

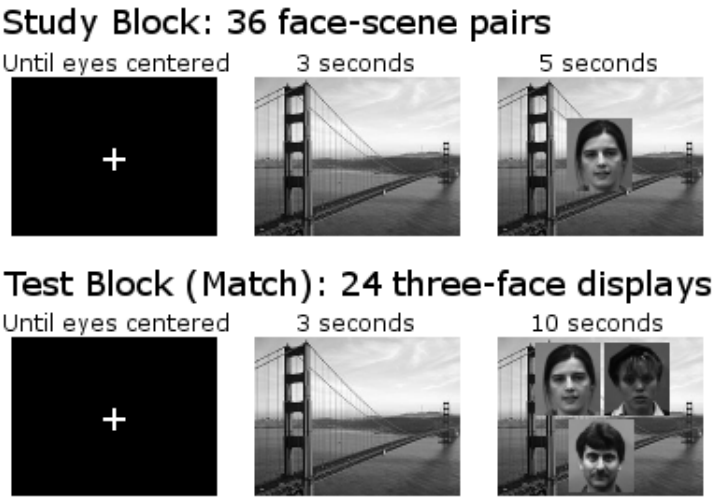
Structure is a critical characteristic of any neural system (Wan, 1994; Heathcote, 1995), no less so in the method described here. Temporal semantic information encoded as a recency gradient only has meaning within the structural context in which it was encoded. The activation of a temporal integration node, which represents the information that a square was the most recent symbol in a certain sequence, only has that specific meaning because there exists a connection between that node and a classifier node that represents square. Furthermore, that classifier node only represents square because of the particular connects between the classifier functional subsection inputs and visual sensor outputs. The guidance to be appreciated by the consideration of structure is that when building up larger architectures with these methods, one must keep in mind invariant structural mappings of the processed information, lest one lose or

corrupt the information being encoded.

4.2. Associating Object/Scenes Pairs

For the second assessment we compared the model results to study results of Hannula et al. (Haykin, 1999). The Hannula study presented human subjects with a series of face-scene pairs in a study block, and then tracked eye movements for sets of three faces, with a background scene, presented in a test block. The set of three faces can be from one of three categories: match, re- pair, or novel. The match face sets contain three known (previously seen) faces, one of which is correctly paired with the background scene. The re-pair face sets contain three known faces, but none of them are correctly paired with the background scene. The novel face sets contain three unknown faces. This task is an exercise in episodic memory for associating people and places.

Figure 4.13. Hannula image scheme



Published results show that normal subjects will spend a larger proportion of viewing time directed to a face that correctly matched the background. Subjects with hippocampal damage did not exhibit this proportional increase in dwell time on the matching face. This result indicates that hippocampus is required for the recognition of previously observed episodes.

The goal of this assessment is to show evidence that our model exhibits some of the same function as biological brain with regard to scene/object pair association. In the interest of

correlating behavior from our model to human behavior, we create a mapping of the human experimental setup to an experiment that we can run, in simulation, on our model. To mitigate visual processing effects, we map the face-scene focus-context images to simple geometric images (initially using squares and triangles, then going to orthogonal lines and dashed lines). Our input images were ten pixels by ten pixels.

We present arrangements of our focus and context images that correspond with the study and test image presentations of the original experiment. The study presentation sequence of the original experiment is fixation, scene, face. As our model lacks a mechanism for separating focus and context information in the visual field, we must simulate that separation by presenting separate images to the focus and surround modality inputs. The presentation sequence for the focus modality is fixation, scene, face. The presentation sequence for the context modality is fixation, scene, scene. These sequences reflect the fact that the scene image is a focus image during the second element of the original study presentation because the scene is the only image on the screen. The context modality only ever sees the scene because even when the face is present in the original study presentation, the scene still forms the background of the image. Part of the original experimental setup is that visual dwell time on an image is a measure of recognition of that image. As the model is lacking eyes, an alternate measure of recognition must be developed. The simulated measure of recognition is equivalent to directly probing neural activation in a human brain. Modeled neural activation can be evaluated by observing the ART classifier module output in the cortex model and the grid node outputs in the hippocampus model. A representative output report is shown in Figure 14. Model recognition scores are computed by summing contributions from each cortical classification module, and the hippocampus. The cortical classification modules can each contribute one point, and the hippocampus can contribute a point. This scoring convention was arbitrary and was sufficient for our purposes. The first ART module in the cortical focus modality contributed a point if it identified an existing template (i.e. it had previously learned a generalization) for the current input. As inputs are presented in sequences of three (fixation, scene, face), the first ART module will make its contribution based on the last element of the sequence. The second ART module in each cortical modality is located after a temporal integrator, and so it will score familiarity based on the whole sequence.

Another biological brain mechanism that our model lacks is the ability to concentrate on

different portions of an image. As such, we must simulate that ability for the purpose of the test images. Instead of a single sequence with the last image containing three faces, we present three sequences with the last image each containing one face. This way, the model does not need to consider three sub-images as the human subjects do when looking at the single test image of three faces. The experiment outputs from Figure 4.5 reflect a test sequence in the original experiment where faces 1, 2, and 3 are shown against scene 1, then faces 2, 5, and 6 are shown against scene 2. In the study portion of the experiment, face 1 was viewed with scene 1 and face 2 was viewed with scene 2.

4.2.1 Results

Figure 4.14 shows an example of the experimental results. This example reflects two test sequences from the Hannula experiment. As this example is from the match category of faces, all three faces are known. One face should have an episodic memory associating it with the tested scene.

Figure 4.14. Experiment outputs

Presentation timestep	Focus-Input Index	Context-Input Index	Total Familiarity Score
1	1	1	5
2	2	1	4
3	3	1	4
4	2	2	5
5	5	2	4
6	6	2	4

The intact model exhibits higher familiarity scores when previously studied matching focus-context (face-scene) image pairs are presented, as opposed to pairs that were not studied together. This behavior correlates to the eye dwell time of human subjects in the Hannula study. If the hippocampus section of the model is lesioned, familiarity scores are the same between matching and non-matching image pairs. This behavior also correlates with human subjects, where subjects with hippocampal damage do not preferentially dwell on particular faces during the matching and non-matching face-scene pairs. Further results show that there is no difference in familiarity scores between different focus images in the re-pair and novel tasks, with either intact or lesioned models. These results correlate with the Hannula data where

hippocampally damaged and normal subjects both view faces in the re-pair task with no preference.

4.2.2 Repeated Experiment

The aforementioned results were obtained from initial model architectures. With significant advances made to the architecture we re-ran the experiment and were subsequently able to improve upon the mapping between the procedure performed by Hannula et al. and our work. Model fidelity advances allowed us to use actual images of faces and scenes rather than representative geometric shapes.

Furthermore, advancements made to the hippocampus representation enabled us to restrict the neural activation analysis to CA3 where associations are formed. By doing so, we were subsequently able to analyze the model CA3 activations and attain the results similar to Hannula et al. These results correlate well with human performance data in which greater hippocampal activation is observed when an existent encoding may be retrieved. Likewise, partial hippocampal activity is required to explore the representation of a novel input, and little hippocampus involvement is required for re-paired episodes which do not fit prior encodings.

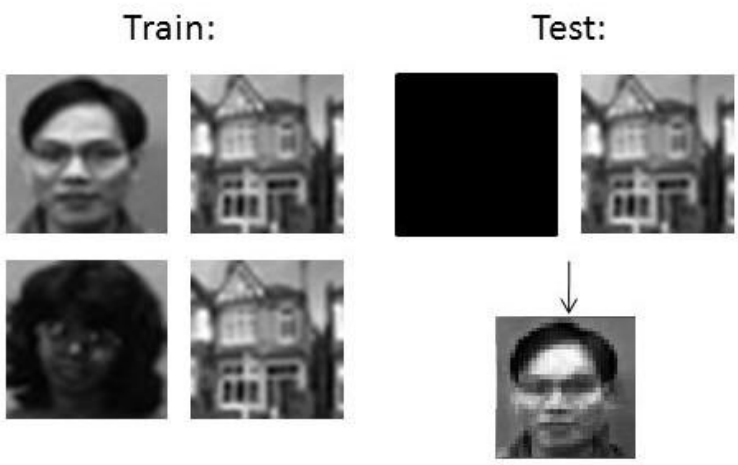
4.3 Co-occurrence of Shared Scenes with Novel Objects

For the third assessment we compared the model to a study performed by Preston et al. (Preston, 2004). In the Preston study, human subjects were trained on black and white photographs of face-house pairs and face-face pairs in three sets. The first set consists of pairs of faces and houses. The second training set introduces new faces paired with the same set of houses shown in the first set. And finally, the third training set consisted of face-face pairs which were previously unseen. During the testing phase of the Preston study, subjects performed forced- choice judgment tasks. Two of the tasks presented either a face or a house and required the subject select the corresponding house or face to complete the pair. The other two tasks focused on face-face pairs. One task was simply a test of the learned face-face pairs, whereas the other task tested subject's ability to recall related face-face pairs which shared a common house but which were never explicitly seen together.

Similarly, we trained our model using face-house pairs such that a face is processed by the

ventral stream and a house is processed by the dorsal stream. An example of the input presented to the model may be seen in the left half of Figure 4.15. See Appendix B for the full training sequence. Our model lacks the ability to perform the forced-choice judgment task. So rather we first trained the model on face-house pairs including faces with a common house. Then, we turned off learning in the model so that no new concepts could be formed, but rather only existent concepts could be used. We then presented the model an ambiguous partial input cue by inputting a blank image to the ventral stream and one of the houses previously seen during training to the dorsal stream. This partial cue presentation may be seen on the right half of Figure 4.15. Rather than selecting between possible choices as in the forced-choice judgment task, we get our model to reconstruct the image that it has stored in memory associated with the house. The resultant image is not an exact copy of the original input, but rather is an amalgamation of categorical representations distributed throughout the hierarchy of TIART modules comprising sensory cortex. A sample recalled face may be seen at the bottom right of Figure 4.15.

Figure 4.15. Face and scene pairs



4.3.1 Qualitative comparison results

In addition to the example shown in Figure 4.15, the model was successfully able to recall correct corresponding faces for each of the houses shown during testing. All of the recalled faces clearly resembled a particular face shown during the training phase; however, each was subject to slight distortions yielding an imperfect recall.

The ability to associate related face-face pairs can be observed qualitatively within the model by noting the activation regions within CA3. Portrayed in the upper left region of the graphical user interface (GUI), shown in Figure 4.16, informally one can observe whether or not the same CA3 activations are employed to encode the association of cohabitation. The visual input presented to the model can be seen in the lower left of the GUI. As displayed in Figure 16, when presented with face A and house A, a distinct region of CA3 is activated. Likewise, as shown in Figure 17, when presented with a different face B also paired with house A, an overlapping region of CA3 is indeed active indicative of the shared encoding between the related face-face pairing. On the other hand, as shown by Figure 4.17, when presented a distinctly different face C and a different house B an entirely different CA3 region is utilized for the encoding.

Figure 4.16. Person 'A' with house 'A'

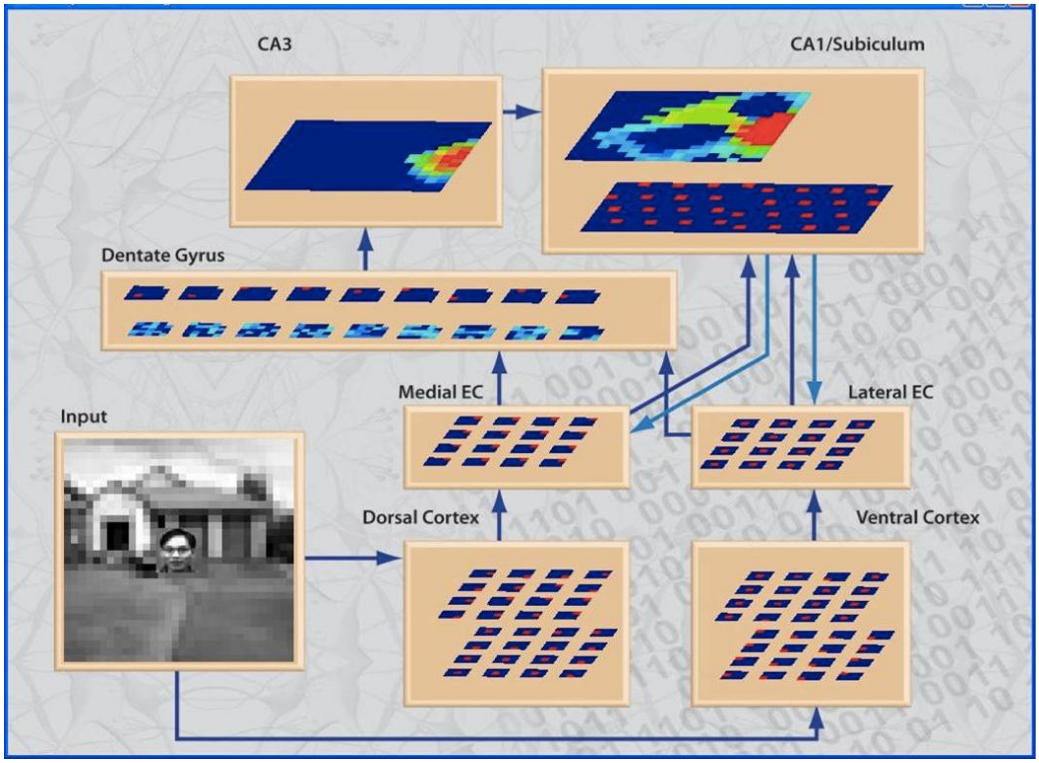


Figure 4.17. Person 'B' with house 'A'

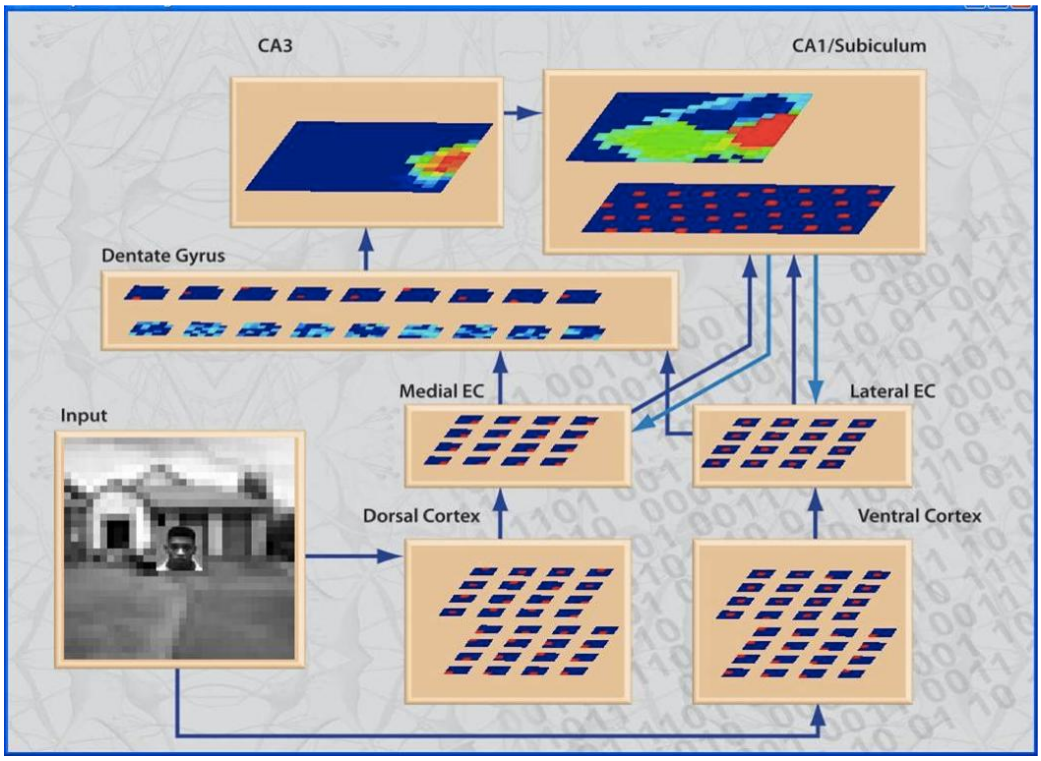
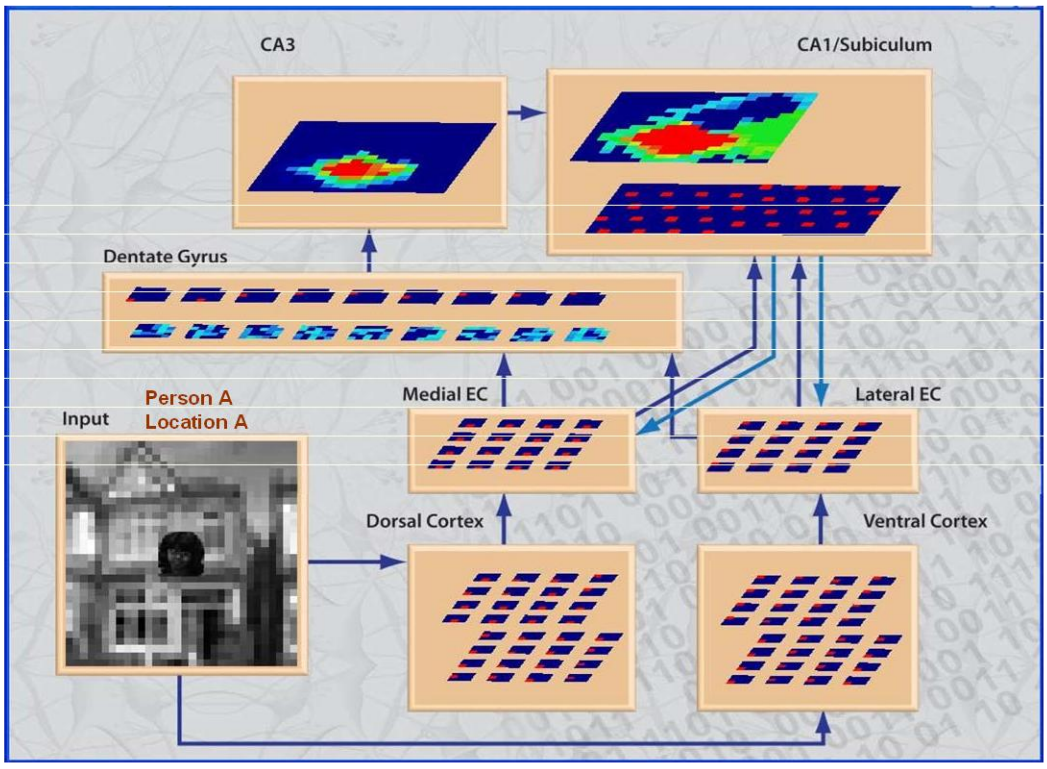


Figure 4.18. Person 'C' with house 'C' produces distinct mapping of activation



4.3.2 Quantitative comparison results

In the absence of a fully embodied model, with output modality to articulate the envisioned associations beyond cued recall, we have applied the mathematics of information theory to quantify the relationship between semantic concepts within the architectural implementation of CA3. Information theory allows for a quantitative evaluation of the information content independent of the particular computational implementation or the underlying neuroanatomical processes modeled.

More specifically, within the context of information theory, mutual information is a measure of the dependence between two random variables (Cover, 2005), and is computed by the double summation given in equation (4.6).

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \quad (4.6)$$

Treating a conjoined face-house pair semantic concept as a random variable, the various CA3 encoding regions represent the alphabet of possible states the random variable may take on. In other words, a single random variable, such as X, represents the combined concept of a particular face and a specific house. Any specific pattern of activation within the CA3 may be used to represent the random variable, and thus the entire CA3 grid is the set of all possible values the random variable may express. From this perspective, mutual information may then be used to

quantitatively evaluate whether or not our architecture recognizes and auto associates inferred relationships.

In comparison to the Preston study, this technique allows us to evaluate whether or not our model is capable of forming an association between unseen related face-face pairs. A single face is only part of a random variable, and so for two different faces to share a relationship they must have a common context. The left column of Figure 4.19 lists the mutual information for the related face-face pairs our model was trained on. For instance, the first entry is the mutual information for two random variables A and E. A different face is represented by A than that of E; however both random variables share the same house. The right column on the other hand presents an averaged mutual information value of all the unrelated faces in reference to a

particular face. As an example, the face represented by A is only related to the face represented by E. All other random variables (in this case B, C, D, F, G, and H) represent unrelated faces. Therefore, in column 2 of Figure 4.19, we represent the average mutual information values for non-matching (i.e. non-auto-associated) faces.

Furthermore, we have tested the capabilities of our model on even more complex associations than those in the Preston experiment, to demonstrate the flexibility available in forming novel arbitrary associations. As shown in Figure 4.20, we have tested our model using a vehicle context in addition to houses. In addition to contextual relationship, a more advanced partially overlapping association occurs in this more advanced example.

Figure 4.19. Model comparison to Preston study

Mutual Information of Related Face-Face Pairs	Average Mutual Information of Unrelated Pairs
I(A;E) = 0.3657	I(A;~E) = 0.0254
I(B;F) = 0.3628	I(B;~F) = 0.0303
I(C;G) = 0.3303	I(C;~G) = 0.0294
I(D;H) = 0.3570	I(D;~H) = 0.0322
I(E;A) = 0.3657	I(E;~A) = 0.0247
I(F;B) = 0.3628	I(F;~B) = 0.0293
I(G;C) = 0.3303	I(G;~C) = 0.0296
I(H;D) = 0.3507	I(H;~D) = 0.0323

Note: ~ denotes negation

Figure 4.20. Faces paired with different contexts.

The motivation behind the Preston et al. study was to investigate the role of the human hippocampus in the novel expression of declarative memories (Preston, 2004). Comparable performance by our model on an equivalent test demonstrates an apropos functional appropriation to the modules comprising our architecture.

In the Preston study, human performance was near perfect for the learned face-house pairs (Preston, 2004). Likewise, our model was successfully able to recall a correct face for each house presented such that the recalled image incurred only slight distortion. Due to algorithmic limitations, the present version of the model can only recall one of the faces associated with a given house. However, this could be corrected by allowing the model to retrieve all association pairs rather than only the single best match.

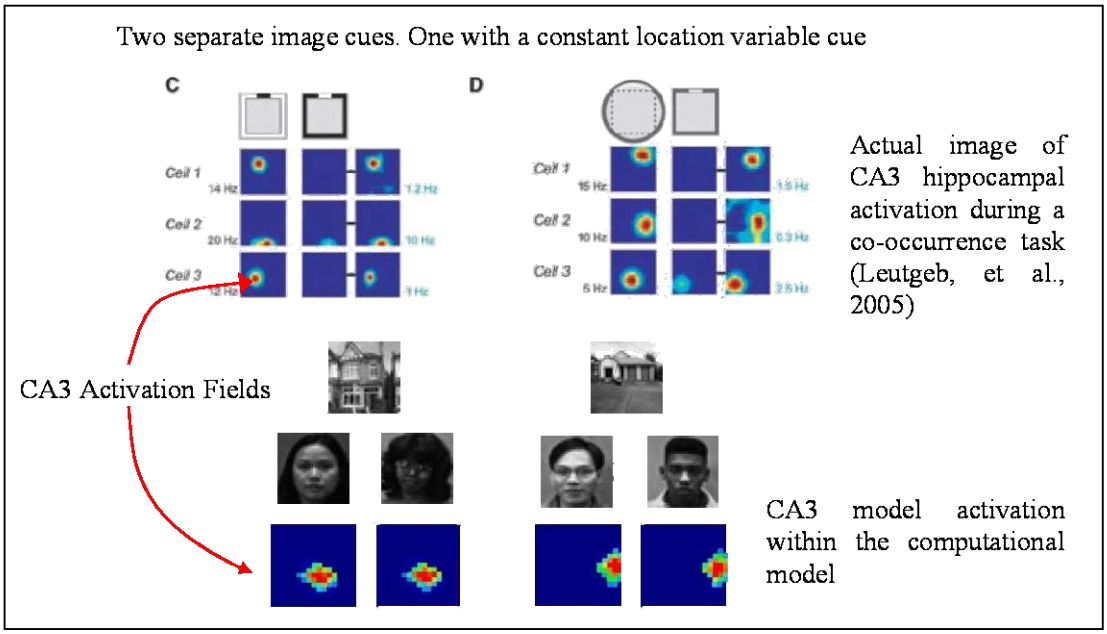
The Preston study observed increased hippocampal activation during fMRI scans when subjects were tested on related face-face pairs compared with learned face-face pairs (Preston, 2004). This observation demonstrates the important role of the hippocampus in relational tasks.

Beyond simply leveraging the hippocampus to form associations within our model, in particular mutual information has quantifiably shown our model is capable of forming associations between novel concepts. In our mutual information measure, we approximate the joint probability distributions for two semantic concepts. This approximation is calculated by computing the normalized fuzzy conjunction of the respective CA3 activations when the concepts are processed by the model individually. As can be seen by the mutual information approximation values given in Figure 20, the related face-face pairs have a significantly larger CA3 mutual information measure than that of unrelated pairs.

Furthermore, by incorporating vehicles as an additional context, we are able to demonstrate that our model is capable of processing a variety of concepts, as is true for humans, and is not only capable of processing houses. This more complex association additionally demonstrates the ability to associate multiple contexts with a single focus in addition to associating multiple foci with a single context. For example, as illustrated in Figure 8, the same person represented in concepts A and B is associated with a house in one concept and a vehicle in the next. A second person is additionally associated with the same house as shown in concept C. While both people cohabitate the same house, only the first person is associated with the vehicle. The ability to differentiate between these overlapping associations is evident by the mutual information measures. Both the mutual information value associating the first person with his house and vehicle, as well as the mutual information value associating the two people cohabitating the same house are considerably larger than the relationship between the second person and the vehicle.

Comparing human CA3 activation during a co-occurrence task (Leutgeb, et al., 2005) to the CA3 activation of the computational model during a similar co-occurrence task yielded similar results. This is shown in Figure 4.21.

Figure 4.21. Experimental and model image of CA3 hippocampal activation during a co-occurrence task



5. GENERAL DISCUSSION AND CONCLUSIONS

The progression of our computational model is driven by attempts to improve model fidelity in relation to neurobiology. Rather than striving to implement the most efficient machine learning algorithms to achieve a desired goal, our approach has been to model the neuroanatomy and processes underlying declarative memory and recall. In doing so, we have demonstrated the ability to model elements of cognitive behavior such as familiarity and recognition.

As a result of continuous improvement to the model we are also able to create automatic associations of various semantic concepts. Additionally we have presented mutual information as a means of quantitatively analyzing the dependence between semantic concepts within the CA3 region of the model. Overall, information theoretic analysis provides a mathematically rigorous means of analyzing the information storage and propagation capabilities of a model in an implementation dependent manner. In general, the artificial neural network computation model we have presented processes sensory inputs and in effect is capable of exhibiting qualitative memory phenomena such as auto-association of episodic memory concepts.

We have made both a neurophysiological and a psychological behavioral case for our model. We

assert that this approach is of great potential benefit to the field because it puts computational modelers and neuropsycholgical investigators into interdisciplinary communication. By engineering a structural neuro-cognitive model, we have highlighted areas where neuroscience could most profitably shine the light of discovery to push our understanding further forward. For instance, all information that traverses through our model goes through the entorhinal cortex to dentate gyrus connections. That connectivity scheme was modeled based on our best anatomical understanding, but what is the merit of bringing together all of the modalities before they are hippocampally bound? Why is this evolutionarily more valuable than retaining the higher information content possible with separate modalities? This is an area of in need of a neuroscientific theoretical approach and an answer could in turn, help us to construct a more veridical, powerful and explanatory model. We believe that our model provides the experimentalist with a useful tool to explore cognitive processes. The behavioral effects we suggest should be confirmed in human subjects, but the model can be used to run exhaustive trials that would not be plausible for human studies. As this model continues to be developed, the computational-to-human study paradigm will only become more attractive and the potential for interdisciplinary collaboration more alluring. This is exemplified in the statement by Neal Cohen, professor and Director of the Head, Brain & Cognition division at the Department of Psychology, & Beckman Institute for Advanced Science and Technology, & Neuroscience Program:

This model supports the ability to do classification/categorization of a range of visual inputs, to remember the prior occurrence of each of those inputs individually, to do pattern completion permitting recovery of those items based on partial or incomplete cues, to represent different locations in the visual environment, to remember which individual items occurred in which locations, and to bind together in memory representations of any arbitrary collection of items with one another and with their spatial or other contexts. And all of these capabilities are implemented in a model with biological realism greater than in any previously implemented model. Finally, it is done in a way that permits us to test the contributions of each of the individual components of the model and to compare that with what is seen in humans and animals.

We believe this work will also benefit the Science and Defense national security mission of the DOE and other federal agencies by increasing the understanding of key aspects of cognition as

well as creating a higher fidelity human modeling architecture. This will enable the DOE to better understand the thought processes underling human behavior, as well as enhance human modeling in areas such as action/counter-action predictive simulations, training, and assistive decision making.

5.1 Model Limitations

Although our model is built upon understood neuroanatomical hippocampal function using biologically plausible computational mechanisms, it is not an identical reproduction of neural anatomy and function. Our model is not an exact neuron for neuron replica of the HC. Indeed, even if we had the computational resources to implement it, a reference map of every neuron and synapse in biological sensory cortex and hippocampus does not exist. Our work provides evidence for some specific connection schemes that we consolidated from the best existing literature. Future models can iterate and improve upon our assumptions. While not implementing the absolute volume of neural nodes in modeled biological structures, the model does take into consideration the neuron density and type within distinct regions, and attempts to preserve the same ratios in allocating computational resources.

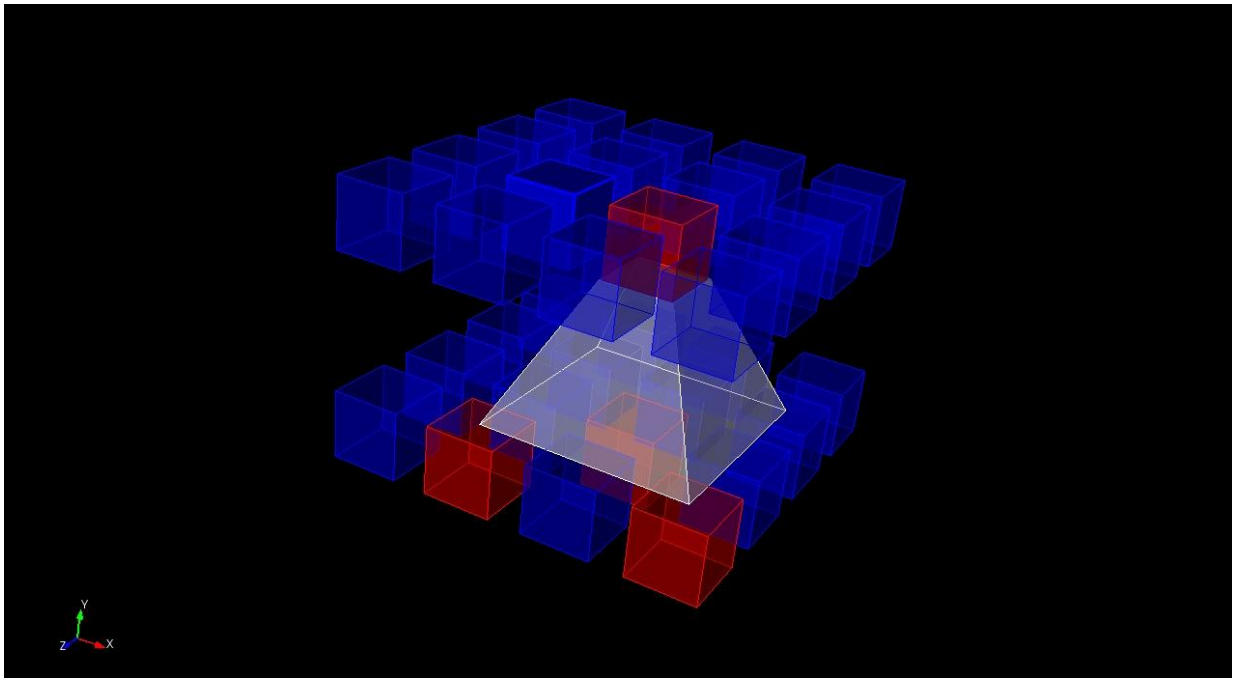
In terms of scope, the model is constrained to sensory cortex, parahippocampus and HC. This partial neural representation does not include an output modality, consequently constraining the means by which we may query and test the model. As addressed in the relevant sections describing the means by which we have tested the model, we have accounted for this limitation by constraining the means in which we extract information from the model. Rather than simply making inferences regarding model performance or knowledge based upon the underlying computational implementation, we have restricted our analysis to mechanisms such as neural activation which is somewhat analogous to brain imaging approaches.

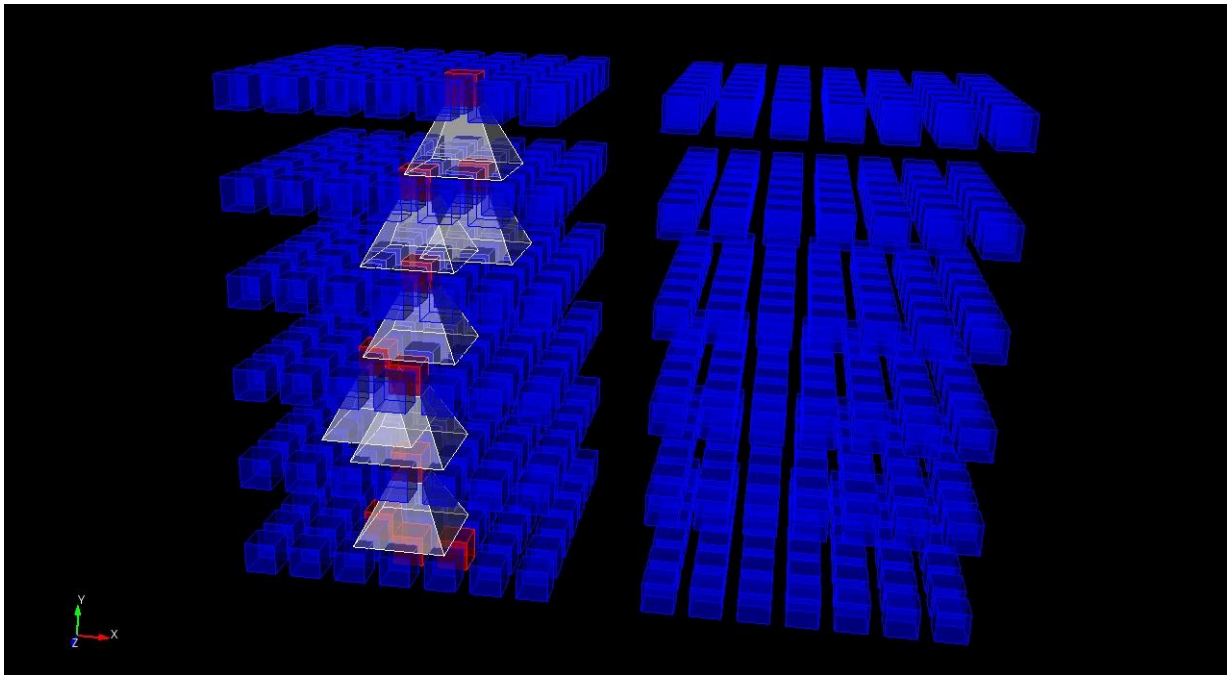
As the fidelity of the model is not at the neuron level, likewise it does not operate via action potentials. Rather, our model requires a clocking system regulating the flow of information through the model. This seemed a reasonable abstraction as we are running the model on digital computers anyway. The temporal integrators through the system do provide a means of buffering up a sequence of inputs, but include a design tradeoff impairing the ability to encode a sequence containing a repeated input separated by a different input. The temporal integrator functionality, as described formerly, decays the activation value of a category representation

over time. However, the activation is replenished upon subsequent presentations of the same input. For example, while an input sequence of ABC could correspond to an integrated vector output <0.5, 0.75, 1>, a sequence which repeats such as ABA is indistinguishable from BA (and various other possibilities). Computationally there are several simple means of compensating in a non- biological manner, but that would contradict the design intentions of this project. We see two possibilities for reconciling this approach with the biology. First, the biology may implement a more complex temporal sequence encoding scheme. We implemented a fairly simple scheme in part through a desire not to make any unreasonable demands upon what biological neural networks might be capable of. Second, the inability to distinguish ABA from AB may not be an issue at sufficiently abstract conceptual levels. It seems a reasonable claim that you can never have exactly the same experience twice, therefore the brain will never see exactly the same pattern of activations twice, hence “A” will never repeat as in ABA.

APPENDIX A: LAYOUT AND CONNECTION STRUCTURE OF THE MODEL

Modules from higher columns receive inputs from multiple modules of lower columns





6 layers at 7X7 modules

APPENDIX B: THE INPUT SET PRESENTED TO THE MODEL IN COMPARISON WITH THE PRESTON STUDY

REFERENCES

1. Addis, D.R., Moscovitch, M., Crawley, A.P., and McAndrews, M.P. (2004) Recollective qualities modulate hippocampal activation during autobiographical memory retrieval. Hippocampus 14: 752-762.

2. Adlam, A.R., Vargha-Khadem, F., Mishkin, M., and de Haan, M. (2005) Deferred imitation of action dequences in developmental amnesia. Journal of Cognitive Neuroscience 17: 240-248.

3. Aggleton, J.P., Kyd, R.J., & Bilkey, D.K. (2004). When is the perirhinal cortex necessary for the performance of spatial memory tasks? *Neuroscience & Biobehavioral Reviews, 28,* 611-24.

4. Aggleton J.P., Vann S.D., Denby C., Dix S., Mayes A.R., et al. (2005) Sparing of the familiarity component of recognition memory in a patient with hippocampal pathology. *Neuropsychologia* 43(12):1810-1823.

5. Alvarado, M.C., & Bachevalier, J. (2005). Comparison of the effects of damage to the perirhinal and parahippocampal cortex on transverse patterning and location memory in rhesus macaques. *Journal of Neuroscience, 25,* 1599-1609.

6. Amaral, DG, Witter, MP. 1989. The three-dimensional organization of the hippocampal formation: A review of anatomical data. *Neuroscience* 31: 571-591

7. Amaral, D. G., & Witter, M. P. (1995). Hippocampal formation. In G. Pacinos (Ed.), *The Rat Nervous System,* (2nd ed., pp. 443-493). San Diego, CA: Academic Press.

8. Andersen, R.A. (1989) Visual and eye movement functions of the posterior parietal cortex. Annual Review of Neuroscience 12:377-403.

9. Anderson, J. R. (1993). Rules of the Mind. Hillsdale, NJ: Erlbaum

10. Annon, A. (2009). Episodic memory modeled by an integrated cortical-hippocampal neural architecture, *Human Behavior-Computational Modeling and Interoperability,* Oak Ridge, TN.

11. Annon, A. (2009). Temporal semantics: An adaptive resonance theory approach, *Proceedings of the 2009 International Joint Conference on Neural Networks.* Atlanta, GA.

12. Aristotle (350 BC). On Memory and Reminiscence. (J. I. Beare, Trans.).

13. Baddeley, A. (2000). The episodic buffer: A new component of working memory? Trends in Cognitive Science, 4, 417-423.

14. Bar,M.&Aminoff,E.(2003).Corticalanalysisofvisualcontext.Neuron,38,347-58. 53

15. Barnes, J. M., & Underwood, B. J. (1959). Fate of first-list associations in transfer theory. *Journal of Experimental Psychology, 58,* 97–105.

16. Best, P.J, White, A.M., and Minai, A. (2001) Spatial processing in the brain: the activity of hippocampal place cells. *Annual Review of Neuroscience 24,* 459-86.

17. Bliss, T. V. P., Collinridge, G. L. (1993). A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature, 361,* 31-39.

18. Brown, M.W. & Xiang, J.Z. (1998). Recognition memory: Neuronal substrates of the judgment of prior occurrence. *Progress in Neurobiology, 55,* 149-89.

19. Brown, M.W. & Aggleton, J.P. (2001). Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience, 2,* 51-61.

20. Buckmaster, C.A., Eichenbaum, H., Amaral, D.G., Suzuki, W.A. and Rapp, P. (2004) Enothrinal cortex lesions disrupt the relational organization of memory in monkeys. *Journal of Neuroscience 24:* 9811-9825.

21. Bunsey, M., & Eichenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature, 379,* 255-257.

22. Burwell, R.D., Witter, M.P. & Amaral, D.G. (1995). Perirhinal and postrhinal cortices of the rat: a review of the neuroanatomical literature and comparison with findings from the monkey brain. *Hippocampus, 5(5),* 390-408.

23. Burwell, R.D., & Hafemanm, D.M. (2003). Positional firing properties of postrhinal cortex neurons. *Neuroscience, 119,* 577-88.

24. Cabeza, R. and St. Jaques, P. (2007) Functional neuroimaging of autobriographical memory.

Trends in Cognitive Sciences (in press).

25. Carpenter G. A., &. Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine, *Computer Vision, Graphics, and Image Processing, 37,* 54–115.

26. Carpenter, G. A., Grossberg, S. & Rosen. D. B. (1991). FuzzyART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural Networks, 4:759–771.*

27. Clayton, N. S., Bussey, T. J., & Dickinson, A. (2003). Can animals recall the past and plan for the future? *Nature Reviews Neuroscience, 4,* 685-691.

28. Cohen, N. J., Ryan, J., Hunt, C., Romine, L., Wszalek, T., Nash, C.(1999). Hippocampal system and declarative (relational) memory: Summarizing the data from functional neuroimaging studies. *Hippocampus, 9:* 83-98.

29. Cohen, N. J. & Eichenbaum, H. (1993). *Memory, Amnesia, and the Hippocampal System.* Cambridge, MA: M.I.T. Press.

30. Cohen, N.J. & Squire, L.R. (1980). Preserved learning of pattern-analyzing skill in amnesia: Dissociation of "knowing how" and "knowing that." *Science, 210,* 207-210

31. Cover, T. M. *Elements of information theory.* Hoboken, N.J: J. Wiley, 2005.

32. Daselaar S.M., Fleck M..S, & Cabeza R. (2006) Triple dissociation in the medial temporal lobes: recollection, familiarity, and novelty. *Journal of Neurophysiology* 96:1902- 1911.

33. Davachi, L. and Wagner, A. G. (2002). Hippocampal contributions to episodic encoding, Insights from relational and item-based learning. *Journal of Neurophysiology, 88,* 982-990.

34. Davachi, L., Mitchell, J. P. and Wagner, A. D. (2003). Multiple routes to memory, Distinct medial temporal lobe processes build item and source memories. *Proceedings of the National Academy of Sciences. 100,* 2157-2162.

35. Day, M., Langston, R., & Morris, R. G. M. (2003). Glutamate-receptor-mediated encoding and

retrieval of paired-associate learning. *Nature, 424,* 205-209.

36. Dobbins, I. G., Foley, H., Schacter, D. L., & Wagner, A. D. (2002). Executive control during episodic retrieval: Multiple prefrontal processes subserve source memory. *Neuron, 35,* 989-996.

37. Downes, J. J., Mayes, A. R., MacDonald, C., & Humkin, N. M. (2002). Temporal order memory in patients with Korsakoff's syndrome and medial temporal amnesia. *Neuropsychologia, 40,* 853-861.

38. Dusek, J. A., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Science, U.S.A., 94,* 7109-7114.

39. Eichenbaum, H. & Cohen, N.J. (2001). From Conditioning to Conscious Recollection: Memory Systems of the Brain. Oxford: Oxford University Press.

40. Eichenbaum, H. (2007). Declarative memory: Insights from cognitive neurobiology. *Annual. Review of Psychology, 48,* 547–572.

41. Eichenbaum, H., P. Dudchencko, E. Wood, M. Shapiro and H. Tanila (1999) The hippocampus, memory, and place cells: Is it spatial memory or a memory space? *Neuron* 23: 209-226

42. Eichenbaum H. 2004 Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron* 44:109-20.

43. Eichenbaum, H., Yonelinas A.R., and Ranganath, C. (2007) The medial temporal lobe and recognition memory. Annual Review of Neuroscience (In press).

44. Ekstrom, A.D., Kahana, M.J., Caplan, J.B., Fields, T.A., Isham, E.A., Newman, E.L., & Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature, 425,* 184- 87.

45. Eldridge, L. L., Knowlton, B. J., Furmanski, C. S., Brookheimer, S. Y., & Engel, S. A. (2000). Remembering episodes: A selective role for the hippocampus during retrieval. *Nature Neuroscience, 3,* 1149-1152.

46. Elman, J. L. (1990). Finding structure in time, *Cognitive Science,* 14, 179–211, 1990. 47. Ergorul, C., & Eichenbaum, H. (2004). The hippocampus and memory for "What",

"Where", and "When". *Learning and Memory, 11,* 397-405. 48. Ferbinteanu,J.&andShapiro,M.L.(2003)Prospectiveandretrospectivememorycoding

in the hippocampus. *Neuron 40:* 1227-1239.

49. Fink, G.R., Markowitsh, H.J., Reinkemeier, M., Bruckbauer, T., Kessler, J., and Heiss, W.-D. (1996). Cerebral representation of one's own past: neural networks involved in autobiographical memory. *Journal of Neuroscience 16:* 4275-4282.

50. Fortin, N. J., Agster, K. L., & Eichenbaum, H. (2002). Critical Role of the Hippocampus in Memory for Sequences of Events. *Nature Neuroscience, 5,* 458-462.

51. Frank, L. M., Brown, E. N. & Wilson, M. (2000). Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron 27:* 169-178.

52. Fuster, J. M. (1995). *Memory in the cerebral cortex.* Cambridge, MA: M.I.T. Press.

53. Gaffan, E. A., Healey, A. N., & Eacott, M. J. (2004). Objects and positions in visual scenes: effects of perirhinal and postrhinal cortex lesions in the rat. *Behavioral Neuroscience, 118,* 992-1010.

54. Gilbert, P. E., & Kesner, R. P. (2003). Localization of function within the dorsal hippocampus: the role of the CA3 subregion in paired-associate learning. *Behavioral Neuroscience, 117,* 1385-1394.

55. Giovanello, K. S., Verfaellie, M., and Keane, M. M. (2003). Disproportionate deficit in associative recognition relative to item recognition in global amnesia. *Cognitive, Affective, and Behavioral Neuroscience 3:* 186-194.

56. Giovanello, K. S., Schnyer, D. M., and Verfaellie, M. (2003) A critical role for the anterior hippocampus in relational memory: Evidence from an fMRI study comparing associative and item recognition. *Hippocampus 14:* 5-8.

57. Hampson, R. E., Heyser, C. J., and Deadwyler, S. A. (1993) Hippocampal cell firing correlates of delayed-match-to-sample performance in the rat. *Behavioral Neuroscience, 107,* 715-739.

58. Hampson, R. E., Pons, T. P., Stanford, T. R., Deadwyler, S. A. (2004) Categorization in the monkey hippocampus: a possible mechanism for encoding information into memory.

59. Hannula, D. E., Ryan, J. D., Tranel, D. & Cohen, N. J. (2007). Rapid onset relational memory effects are evident in eye movement behavior, but not in hippocampal amnesia. *Journal of Cognitive Neuroscience, 19(10):1690–1705.*

60. Hannula, D. E.; Tranel, D.; Cohen, N. J. (2006). The long and the short of it: Relational memory impairments in amnesia, even at short lags, *Journal of Neuroscience 26,* (32), 8352-8359.

61. Hargreaves, E. L., Rao, G., Lee, I., & Knierim, J. J. (2005). Major dissociation between medial and lateral entorhinal input to dorsal hippocampus. *Science, 308,* 1792-4.

62. Hartley, T., Maguire, E. A., Spiers, H. J., & Burgess, N. (2003). The well-worn route and the path less traveled: distinct neural bases of route following and wayfinding in humans. *Neuron, 37,* 877-888.

63. Haykin, Simon S. (1999). *Neural networks a comprehensive foundation.* Upper Saddle River, N.J: Prentice Hall.

64. Heckers, S., Zalezak, M., Weiss, A. P., Ditman, T., and Titone, D. (2004) Hippocampal activation during transitive inference in humans. *Hippocampus 14,* 153-162.

65. Henson, R. N., Cansino, S., Herron, J. E., Robb, W. G., & Rugg, M. D. (2003). A familiarity signal in human anterior medial temporal cortex? *Hippocampus, 13,* 301-304.

66. Honey, R. C., Eatt, A. & Good, M. (1998). Hippocampal lesions disrupt an associative mismatch process. *Journal of Neuroscience 18* :2226-2230.

67. Hopkins, R. O. & Kesner, R. P. (1995). Item and order recognition memory in subjects with hypoxic brain injury. *Brain and Cognition, 27,* 180-201.

68. James, W. (1890). *The Principles of Psychology.* (1918 ed.). New York: Holt. 69. Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2000). *Principles of Neural Science.* New York: McGraw-Hill Co.

70. Kanwisher, N., Downing, P., Epstein, R., and Kourtzi, Z. (2007) Functional neuroimaging of visual recognition. In R. Cabeza & A. Kingstone, eds. *Handbook of Functional Neuroimaging of Cognition.* MIT Press: Cambridge. pp. 109-152.

71. Kesner, R. P., Gilbert, P. E., and Barua, L. A. (2002). The role of the hippocampus in memory for the temporal order of a sequence of odors. *Behavioral Neuroscience, 116,* 286- 290.

72. Kesner, R.P., Hunsaker, M.R., & Gilbert, P.E. (2005). The role of CA1 in the acquisition of an object-trace-odor paired associate task. *Behavioral Neuroscience, 119,* 781-786.

73. Kingsley. R (2000). *Concise Text of Neuroscience.* Lippincott, Williams & Wilkins. 74. Kohonen,T.(2001).Self-OrganizingMaps.Third,extendededition.Springer.

75. Kreiman, K., Kock, C., and Fried, I. (2000a). Category specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience, 3,* 946-953.

76. Kreiman, K., Kock, C., & Fried, I. (2000b). Imagery neurons in the human brain. *Nature 408:* 357-361.

77. Krubitzer, L., & Kaas, J. (2005). The evolution of the neocortex in mammals: how is phenotypic diversity generated? *Current Opinions in Neurobiology, 15,* 444-453.

78. Kuhn, G., Watrous, R. L. &Ladendorf, B. (1990). Connected recognition with a recurrent network, *Speech Communication, 9,* 41–48.

79. Kumaran, D. & Maguire, E. A. (2006). The dynamics of hippocampal activation during encoding of overlapping sequences. *Neuron, 49,* 617-629.

80. Lawrence, S., Giles, C. L., & Fong, I. (1996). Can recurrent neural networks learn natural language grammars, *Proceedings of the IEEE International Conference on Neural Networks,* 1853–1858.

81. Leutgeb, et al. Independent Codes for Spatial and Episodic Memory in Hippocampal Neuronal Ensembles. *Science* 309, 619 (2005)

82. Maguire E. A, (2001). Neuroimaging studies of autobiographical events memory.

Philosophical Transactions of the Royal Society of London, Series B: 356, 1441-1452.

83. Manns J. R and Eichenbaum H. (2006). Evolution of the hippocampus. In J.H. Kaas, ed. *Evolution of Nervous Systems.* Vol 3. Academic Press: Oxford pp. 465-490.

84. Martin, A. (2007). Functional neuroimaging of semantic memory. In R. Cabeza & A. Kingstone, eds. *Handbook of Functional Neuroimaging of Cognition.* MIT Press: Cambridge. pp. 153-186.

85. McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457

86. McDonough, L., Mandler, J. M., McKee, R. D., and Squire, L. R. (1995) The deferred imitation task as a nonverbal measure of declarative memory. *Proceedings of the National Academy of Sciences 92:* 7580-7584.

87. Mountcastle, V. B., J. C. Lynch, A. Georgopoulos (1975) Posterior partial association cortex of the monkey. Command functions for operations within personal space. *Journal of Neurophysiology 38:871-908.*

88. Muller, R. U., Kubie, J. L., & Ranck, J. B., Jr. (1987). Spatial firing patterns of hippocampal complex spike cells in a fixed environment. *Journal of Neuroscience 7,* 1935- 1950.

89. Muller, R. U, Poucet, B., Fenton A. A., & Cressant, A. (1999). Is the hippocampus of the rat part of a specialized navigational system? *Hippocampus. 9,* 413-22.

90. Mumby, D. G., & Pinel, P. J. (1994). Rhinal cortex lesions and object recognition in rats. *Behavoral Neuroscience, 108,* 11-18.

91. Mumby, D. G. (2001) Perspectives on object recognition memory following hippocampal damage: lessons from studies on rats. *Behavioural Brain Research 127,* 159-181.

92. Mumby, D. G., Gaskin, S., Glenn, M. J., Scharamek, T. E. and Lehmann, H. (2002) Hippocmapal damage and exploratory preferences in rats: memory for objects, place, and contexts. *Learning and Memory 9:* 49-57.

93. Norman, G., & Eacott, M. J. (2005). Dissociable effects of lesions to the perirhinal cortex and the postrhinal cortex on memory for context and objects in rats. *Behavioral Neuroscience, 119,* 557-66.

94. Northoff, G. & Bermpohl, F. (2004) Corticla midline structures and the self. *Trends in Cognitive Sciences 8:102-107.*

95. O'Reilly, R. C. (2006). Biologically based computational models of high-level cognition." *Science,* 91-94.

96. O'Reilly, R. C., Rudy, J. W. (2001). Conjunctive representations, the hippocampus, and contextual fear conditioning, *Cognitive Affective Behavior Neuroscience, 1,* 1, 66-82.

97. O'Reilly, Randall C., & Yuko Munakata. (2000). *Computational Explorations in Cognitive Neuroscience Understanding.* Mit Pr.

98. Otto, T., & Eichenbaum, H. (1992). Complementary roles of orbital prefrontal cortex and the perirhinal-entorhinal cortices in an odor-guided delayed non-matching to sample task. *Behavioral Neuroscience, 106,* 763-776.

99. Pihlajamaki, M., Tanila, H., Kononen, M., Hanninen, A., Soininen, H., & Aronen, H.J. (2004). Visual presentation of novel objects and new spatial arrangements of objects differentially activates the medial temporal lobe areas in humans. *European Journal of Neuroscience, 19,* 1939-49.

100.Preston, A. R. (2004). Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus,* 148-52.

101.Preston, A., Shrager, Y., Dudukovic, N.M. & Gabrieli, J.D.E. (2004) Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus, 14,* 148-152.

102.Quiroga R.Q., Reddy L., Kreiman G., Koch C., Fried I. (2005) Invariant visual representation by single neurons in the human brain. *Nature* 435:1102-1107.

103.Ranganath, C., Yonelinas, A.P., Cohen, M.X., Dy, C.J., Tom, S.M., and D'Esposito, M.D. (2003) Dissociable correlates of recollection and familiarity with the medial temporal lobes.

Neuropsychologia 42: 2-13.

104.Rolls, E. T., & Kesner, R. P. (2006). A computational theory of hippocampal function, and empirical tests of the theory. *Progress in Neurobiology, 79,* 1-48.

105.Rosenbloom, P. S., Laird, J. E., & Newell, A. (editors). *The Soar Papers: Research on Integrated Intelligence.* MIT Press, Cambridge, MA, 1993.

106.Scoville, W. B., & Milner, B. (2000). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neuropsychiatry Clinical Neuroscience 12,* 103-13.

107.Sejnowksi, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce english text," *Complex Systems, 1,* 145–168.

108.Shastri L. & Fontaine, T. (1995). Recognizing Hand-printed digit strings using modular spatio-temporal connectionist networks, *Connection Science,* 7, 211–235.

109.Sperling, R., Chua, E., Cocchiarella, A., Rand-Giovannetti, E., Poldrack, R., Schacter, D.L, and

110. Albert, M. (2003) Putting names to faces: Successful encoding of associative memories activates the anterior hippocampal formation. *NeuroImage 20:* 1400-1410.

111.Spiers, H. J., Burgess, N., Hartley, T., Vargha-Khadem, F., & O'Keefe, J. (2001). Bilateral hippocampal pathology impairs topographical and episodic memory but not visual pattern matching. *Hippocampus, 11,715-725.*

112.Sun, R. & Giles, C. (2001). Sequence learning: from recognition and prediction to sequential decision making, *IEEE Intelligent Systems, 16,* 67–70.

113.Suzuki, W. A., Zola-Morgan, S., Squire, L.R., & Amaral, D.G. (1993). Lesions of the perirhinal and parahippocampal cortices in the monkey produce long-lasting memory impairment in the visual and tactual modalities. *Journal of Neuroscience, 13,* 2430-51.

114.Suzuki W. A. & Amaral D. G. (1994) Perirhinal and parahippocampal cortices of the macaque monkey: cortical afferents. *Journal of Comparative Neurology 350,* 497-533.

115.Suzuki, W., & Eichenbaum, H. (2000). The neurophysiology of memory. *Annals of the NY*

Academy of Sciences, 911, 175-91.

116.Svarer, C., Hansen,L. K., &. Larsen, J. (1993). On design and evaluation of tapped-delay neural network architectures, *Neural Networks,* 46–51.

117.Suzuki W.A. & Amaral D.G. (1994) Perirhinal and parahippocampal cortices of the macaque monkey: cortical afferents. *Journal of Comparative Neurology 350,* 497-533.

118.Suzuki, W., & Eichenbaum, H. (2000). The neurophysiology of memory. *Annals of the NY Academy of Sciences, 911,* 175-91.

119.Taylor, S. E., Bernard, M. L. Caudell, T. P., Cohen, N. J. Healy, M. J., Morrow, J. D., Verzi, S. J., Vineyard, C. M., & Watson, P. (2009). Memory in silico: Building a neuromimetic episodic cognitive model. *Proceedings of the 2009 International Joint Conference on Neural Networks.* Atlanta, GA.

120.Taylor, S. E., Bernard, M. L. Caudell, T. P., Cohen, N. J. Healy, M. J., Morrow, J. D., Verzi, S. J., Vineyard, C. M., & Watson, P. (2009) Memory in silico: Building a neuromimetic episodic cognitive model. *Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering.* Los Angeles, California USA. 733-37.

121.Taylor, S. E., Healy, M. J., and Caudell, T. P. (2007). Categorical mapping from ontology to neural network: Initial studies of simple neural networks' concept capacity, *Proceedings of International Joint Conference on Neural Networks, Orlando, Florida,* 2020–2025.

122.Taylor, S. E., Bernard, M. L., Verzi, S. J., Morrow, J. S., Vineyard, C, M., Healy, M. J. & Caudell, T. P. (2009). Temporal semantics: An adaptive resonance theory approach." *Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering.* Los Angeles, California USA. 733-37.

123.Tulving, E. (1983). *Elements of episodic memory.* Oxford: Clarendon Press. 61

124.Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53,* 1-25.

125.Turriziani, P., Fadda, L., Caltagirone, C., and Carlesimo, G. A. (2004) Recognition memory for

single items and associations in amnesia patients. *Neuropsychologia 42:* 426- 433.

126.Uncapher M. R., Otten L. J., Rugg M. D. (2006) Episodic encoding is more than the sum of its parts: an FMRI investigation of multifeatural contextual encoding. *Neuron* 52:547-556.

127.Vargha-Khadem, F., Gadin, D. G., Watkins, K. E., Connelly, A., Van Paesschen, W. & Mishkin, M. (1997). Differential Effects of Early Hippocampal Pathology on Episodic and Semantic Memory. *Science, 277,* 376-80.

128.Vila, L. (1994). A survey on temporal reasoning in artificial intelligence, *AI Communications, 7,* 4–28.

129.Waibel, A., Hanazawa, T., Hinton, K., Shikano, & Lang, J. (1989). Phoneme recognition using time-delay neural networks, *Acoustics, Speech, and Signal Processing* [see also IEEE Transactions on Signal Processing], IEEE Transactions on, vol. 37, no. 3, pp. 328–339.

130.Walczak, S. (2005). Artificial neural network medical decision support tool: Predicting transfusion requirements of ER patients, Information Technology in Biomedicine, *IEEE Transactions on, 9,* 468–474

131.Wan, E. A., (1994). *Time series prediction by using a connectionist network with internal delay lines,* In Time Series Prediction. Addison-Wesley, pp. 195–217.

132.Wan, H., Aggleton, J.P., Brown, M.W. (1999). Different contributions of the hippocampus and perirhinal cortex to recognition memory. *Journal of Neuroscience, 19,* 1142-48.

133.Wood E, Dudchenko PA, Eichenbaum H. (1999). The global record of memory in hippocampal neuronal activity. *Nature 397:* 613-16.

134.Wood E, Dudchenko P, Robitsek JR, Eichenbaum H. (2000). Hippocampal neurons encode information about different types of memory episodes occurring in the same location. *Neuron, 27,* 623-33.

135.Yonelias, A. P., Kroll, N. E., Quamme, J. R., Lazzara, M. M., Sauve, M. J., Widaman, K. F., & Knight, R. T. (2002). Effects of extensive temporal lobe damage or mild hypoxia on recollection and familiarity. *Nature Neuroscience, 5,* 1236-41.

136.Zeineh, M. M., Engel, S. A., Thompson, P. M., and Brookheimer, S. Y. (2003) Dynamics of the hippocampus during encoding and retrieval of face-name pairs. *Science 299:* 577-580.

A COMPUTATIONAL MODEL OF RELATIONAL MEMORY BINDING IN THE HIPPOCAMPUS

Authors
Patrick D. Watson, Kenny Sharma, Howard B. Eichenbaum, Neal J. Cohen

This paper was prepared for submission to Neurocomputation. It was the result of an interdisciplinary effort to develop a bio-inspired system for natural understanding of complex intelligence data.

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20022. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

Abstract:

We present a non-linear dynamical spiking neuron model based upon hippocampal anatomy and electrophysiology for relational memory binding, encoding, and reconstruction (RMBER). The model decomposes the complex firing dynamics of entorhinal cortex input neurons into subcomponents that encode phase and frequency information within the high-dimensional space of the dentate gyrus. These subcomponents of the input are combined within a highly recurrent CA3 model to recover the original input dynamics. The reconstituted signal is then mapped back to the correct inputs within CA1. This decomposition and reconstruction is functionally similar to performing a discrete Fourier transform (DFT) followed by an inverse DFT on the entorhinal cortex dynamics, and allows the rapid encoding and reconstruction of any arbitrary pattern of input firing without sacrificing the fundamentally compositional nature of the input. We demonstrate how these patterns of input firing can be used to represent different relationships between items and from simple rule-based mappings to complex spatial trajectories.

Introduction

The hippocampus is required for rich, relational representations including episodic memories (Cohen & Squire, 1980; Tulving & Markowitsch, 1998; Burgess, Maguire, & O'Keefe, 2002; Konkel & Cohen, 2009); and emerging research suggests that it also contributes to online processing (Voss et al., 2011; Warren, Duff, Jensen, Tranel, & Cohen, 2012), and imagining (Hassabis, Kumaran, Vann, & Maguire, 2007). Since episodic memory, online reasoning, and imagining are all constructive processes that require binding together complex, arbitrarily or accidentally related stimuli (e.g., items, locations, and times) from different sensory modalities, while simultaneously incorporating observed, remembered, and predicted stimuli, it is no surprise that these processes share some of the same hardware. The hippocampus has long been associated with representing configurations of flexible relations, important for navigating novel paths (O'Keefe & Nadel, 1978; Samsonovich & McNaughton, 1997), learning sequences (Eichenbaum, 2000; Jensen & Lisman, 2005), and applying abstract rules such as transitive inference (Dusek & Eichenbaum, 1997; Heckers, Zalesak, Weiss, Ditman, & Titone, 2004).

Yet, mental reconstruction is a tall order, as we experience and interact with our environment, neurons in sensory and associative regions are activated when appropriate environmental patterns occupy their receptive fields (e.g., lines: Hubel & Wiesel, 1963; shapes: Murata, Gallese, Luppino, Kaseda, & Sakata, 2000; faces: Tanaka, Saito, Fukada, & Moriya, 1991). When one of these units is activated, its firing codes information about its input (Gerstner, Kreiter, Markram, & Herz, 1997), but different regions can use radically different coding schemes (Hargreaves, Rao, Lee, & Knierim, 2005). To reconstruct a rich, episode-like experience, involving a novel and arbitrary composition of multiple items and relations, requires capturing each regions' unique codes, and later rebuilding them with sufficient verisimilitude to resemble the original experience (c.f. mental time travel Tulving, 2002).

What is it about the neural architecture of the hippocampus that allows it to bind together all of this information? How are such codes represented in tissue? What mechanism encodes and reconstructs arbitrary relations within the hippocampus? Several neurocomputational models have attacked these questions.

Some focus on the hippocampus's ability to rapidly encode associations, and provide a training signal for the neocortex (McClelland, McNaughton, & O'Reilly, 1995; Norman & O'Reilly, 2003). Some focus on hippocampal spatial representations, such as the formation of place cells, grid cells, and path integration (Etienne & Jeffery, 2004; Milford, Wyeth, & Prasser, 2004; Samsonovich & McNaughton, 1997). Others focus on the electrophysiology of hippocampal theta and gamma oscillations, and map this to representations of temporal sequence (Cutsuridis, Cobb, & Graham, 2010; Mizuseki, Sirota, Pastalkova, & Buzsáki, 2009). Still others focus on the hippocampus's ability to create and distinguish between unique, non-overlapping representations (i.e., pattern separated representation), while simultaneously reconstructing partial or incomplete representations (i.e., pattern completion) (Treves & Rolls, 1994; O'Reilly & Rudy, 2000; O'Reilly & Norman, 2002; Rolls, 2010).

We present a model for relational memory binding, encoding, and reconstruction (RMBER), based on hippocampal anatomy and electrophysiology designed to encode, maintain, and reconstruct any arbitrary pattern of neural firing in its inputs in exactly the same format as it was originally presented regardless of the code used by the input neurons, or the state of affairs

that code represents. After all, from the point of view of the hippocampus, there's no outside world at all, only afferents from the entorhinal cortex and modulatory systems, the hippocampus doesn't "know" what this pattern of firing maps onto. From the hippocampus's perspective it is sufficient to:

1. Capture the identities of active entorhinal cortex neurons involved in the memory to be (re)constructed.
2. Capture the dynamics of these inputs (i.e., frequency of firing, and relative phases of each element, possibly with respect to reference rhythm such as theta).
3. Reconstruct patterns of inputs and their temporal dynamics from a partial input.

Music is a useful analogy: the entorhinal cortex has a series of instruments which can be played (i.e., cells which can fire). These instruments can produce different notes (i.e., the cells fire with some sort of complex dynamics in the frequency domain), and notes can be combined into synchronous chords, asynchronous scales, or some combination (i.e., the different firing frequencies possess phase delays relative to each other). If an input from the cortex plays a tune to the hippocampus, it need retain only the sheet music: instructions for reproducing the song (i.e., the hippocampus compresses the input), since the instruments are still present in the entorhinal cortex, the hippocampus need not store any information about how to reproduce a particular sound. So long as the instructions contain rhythms, meter, and key—the instruments will reconstruct song by overlapping these different components, producing "melodies" and "harmonies" implicitly. Given a few bars as a cue, the hippocampus should be able to fill in the rest by reactivating and playing back this compressed code using the "notes" available in the cortex and the rhythms and meters stored in the hippocampus. This code is combinatorial, the sheet music can be cut up and reassembled in different ways to produce novel tunes constrained only by the tunes that have been encoded before.

Model Architecture

In the following sections, we describe a spiking neuron based attractor neural network model of the hippocampus. The RMBER model mirrors the majority of the simplified structure, connectivity, and dynamics of the basic hippocampal-entorhinal circuit (Figure 5.1).

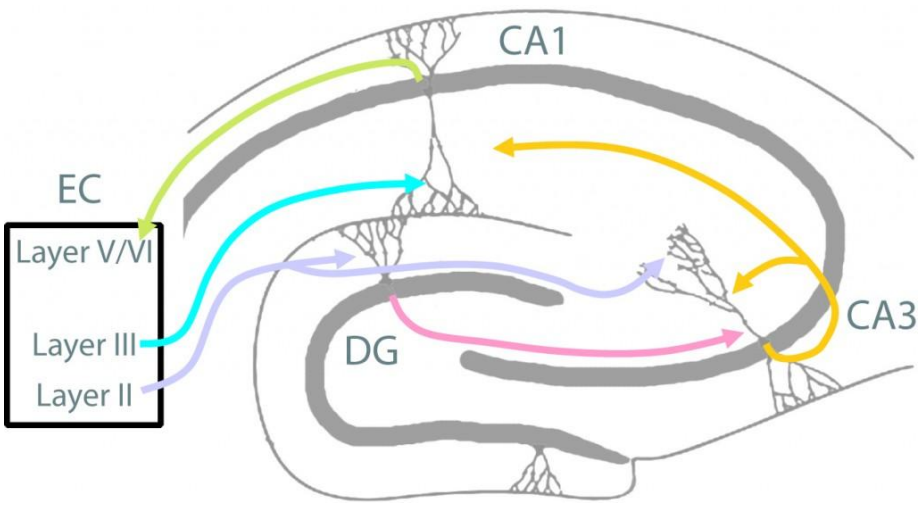


Figure 5.1 - An illustrative slice of the hippocampo-entorhinal circuit, with projections indicated by arrows.

The model is composed of four primary regions - the entorhinal cortex (which we further subdivide into input layers 2 and 3 and output layer 5 - henceforth EC2, EC3, and EC5 respectively), the dentate gyrus (DG), Cornu Ammonis region 3 (CA3), and Cornu Ammonis region 1 (CA1). Each region consists of both principle cells and inhibitory interneurons. Each region fulfills a specific functional role within the RMBER model, allowing us to assign semantic labels to describe high-level functionality. (Figure 5.2).

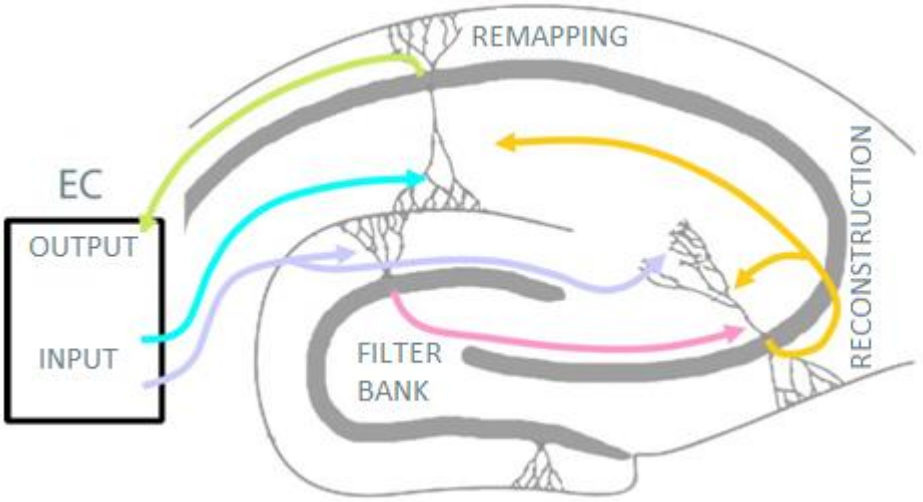


Figure 5.2 - RMBER Region Functional Roles

Synaptic connectivity also mirrors that of the hippocampus (Andersen, Morris, Amaral, Bliss, & O'Keefe, 2008). Projections between regions are meant to parallel the perforant path (linking

CA1, CA3, and the EC), and trisynaptic loop (connecting EC to DG, DG to CA3, and CA3 to CA1). Inter-region connections preserve some of the topology between regions with each unit of the input region by projecting efferents that "fan" into the most topographically proximal section of the target.

In addition, the model is designed to exhibit slow gamma phase oscillations, at approximately 40hz (Mizuseki et al., 2009). This oscillation corresponds to transient synchrony between cells within a single sub-region of the RMBER model. This means within each approximately 10hz theta cycle each successive sub-region of the model will become active once, allowing a single input pattern will propagate through all regions such that the output arrives back at the EC5 in synchrony with the peak of the subsequent theta cycle (Figure 5.3).

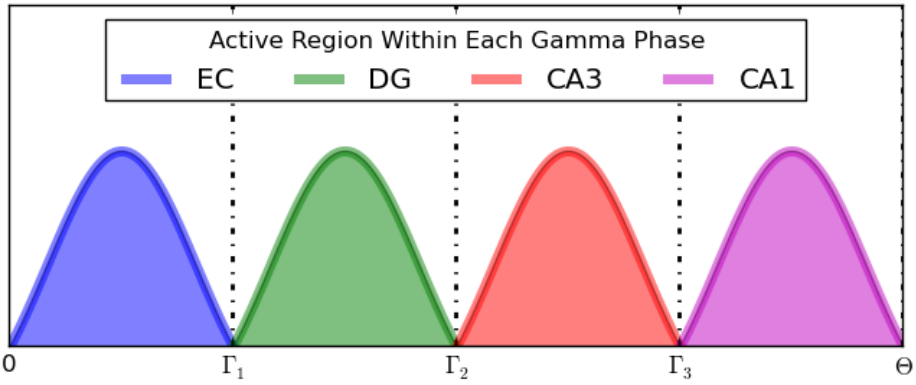


Figure 5.3 - Slow Gamma / Theta Relationship Showing Active Region

Components and Connectivity

The parameters presented below should be viewed as our "best guess" based on the anatomy of the region. We made an extensive exploration of the parameter-space, and found that the oscillatory dynamics of the model are reasonably robust to some changes in the initial weights and connectivity. Thus the general function of the model can be achieved by a wide variety of similar implementations, but the particular dynamics described in the results section will be influenced by the initial parameter assignment. Further, it is important to note that the model may not always converge or correctly encode all input terms due to the nature of the purely random connectivity and weighting parameters used.

1. Neurons and Synapses

The model is composed of Mihalas-Niebur generalized linear integrate-and-fire neurons (Mihalas & Niebur, 2009) and synapses that possess an exponential post-synaptic potential current decay rate. Synaptic plasticity plays a key role in the model's ability to learn and reconstruct patterns of similar, but disparate activity. Where applicable, synapses are updated via spike-timing dependent plasticity (STDP, Song, Miller, & Abbott, 2000) with synaptic scaling (Abbott & Nelson, 2000; Turrigiano & Nelson, 2000; Turrigiano, 2008). Initial weights and delays are drawn from Guassian distributions unless otherwise noted.

2. The entorhinal cortex input layers (EC2 and EC3)

The first layer (EC2) is meant to serve as the sole source of fully unitized input to the model. The number of neurons in this layer is solely dependent on the nature and constraints of the input source being modeled. Within our implementation, these neurons were directly connected to input stimulus neurons with a weight and delay sufficient to exactly reproduce the desired input patterns. Where different input modalities are used, they are localized to different topological regions of EC2.

Each EC2 neuron targets 5% of DG neurons with exclusively excitatory forward connections. These connections are designed to project the unitized input into a higher level space, generating a unique pattern of activity that can be used to reconstruct and disambiguate spatial and temporal relationships. Each input pattern should produce a unique and reproducible expansion that can fully express the needed level of granularity. The contribution of several closely timed input spikes is sufficient to produce this desired behavior - a one-to-many network pattern that relies on specific units and their timing.

Within our implementation, the synaptic weights were selected between 2 and 3 mV with delays between 85-115% of a single gamma cycle. The exponential decay constant was selected as one-quarter of a gamma cycle. These parameters help to add a small amount of variability to the contributions of subsets of EC2 neurons to DG neurons, such that a DG neuron may respond to a particular subset of EC2 stimulus while another will not.

The second layer (EC3) is meant to serve as an additional source of input to the CA1 region. This layer is not critical to the base functionality of the RMBER model; however, it was used to support modeling of spatial behavior.

2. The Dentate Gyrus (DG) component

The second component is meant to approximate the first region of the hippocampus proper, the Dentate Gyrus. The Dentate Gyrus contains a different cell type from the other regions (dentate granule cells), and contains many more cells than either its principle input (EC2) or its principle target (CA3). Thus we model the DG with 100 times the number of EC2 neurons. The specific scale can vary depending on the resolution of the input patterns being modeled. Simple, sequential associations between two neurons can use considerably fewer DG neurons than input patterns with complex temporal relationships. This component receives inputs exclusively from the EC2 layer.

Each DG unit has locally recurrent connections to 30% of its nearest neighbors. Again, this percentage can be modified depending on the complexity of the input pattern. This connectivity serves a critical role in being able to reconstruct a complete pattern based on partial input. The greater the number of DG neurons, the finer the resolution of reconstruction and ability to disambiguate between similar signals (c.f. Aimone, Deng, & Gage 2011). The strength of these connections is initialized at 0 mV with a delay between 0 ms and one half of a gamma cycle. STDP is configured with a coincidence window of +/- 2 ms with additive potentiation of 0.1 mV and maximum value between 0.5 and 1 mV.

The coincidence window and maximum connection weights are relatively weak since the connectivity is fairly extensive. This means that the precise contribution of a preceding pattern is required to produce features of the subsequent pattern. Highly similar patterns may contain considerable overlap; however, the number of neurons and specific weight distributions can be extended to provide further disambiguation. The exponential decay constant was selected between 1 ms and one quarter gamma phase depending again on the complexity of the input being modeled - the lower the value, the higher the resolution of the pattern reproduced.

Many DG units project to a single CA3 unit. The specific number of targets reflects the inverse of the expansion connectivity from EC2 to DG - a randomly selected 5% of DG neurons target a single CA3 neuron. Connectivity strength of these connections is initialized between 2 and 3 mV with a delay of roughly a gamma phase and a post-synaptic exponential decay constant of 1 ms. STDP is configured with a coincidence window of +/- 1 ms with additive potentiation of 0.1 mV and maximum value of 3 mV.

3. The CA3 component

The third component is meant to simulate CA3. Hippocampal anatomy suggests a rough parity between the number of CA3 neurons and EC neurons. The CA3 component receives input exclusively from DG units. This connectivity reproduces the original input signal terms from the high dimensional mapping produced by DG.

Each CA3 unit has extensive locally recurrent connections with a large neighborhood, connecting with 50% of the other cells in CA3 The strength of these connections is initialized between 0 and 0.5 mV with a delay between 0.1 and 1 ms and a post-synaptic exponential decay constant of 1 ms. STDP is configured with a coincidence window of +/- 1 ms with additive potentiation of 0.1 mV and maximum value of 0.5 mV.

CA3 sends efferents exclusively to CA1, with each CA3 cell targeting a single CA1 cell. The strength of these connections is initialized to feed the reconstructed CA3 pattern forward for reconciliation with the original EC2 input.

4. The CA1 Component.

The fourth component is meant to simulate CA1. Anatomical studies suggest it has a similar number of similar units to CA3, though its connectivity is quite different. Thus we simulate CA1 using the same number of neurons as both CA3 and the EC3. The CA1 component receives inputs exclusively from both EC2 and CA3.

Each CA1 unit has locally recurrent connections to 30% of its nearest neighbors. The strength of these connections is initialized between 0.1 and 0.5 mV with a delay between 0.1 and 1 ms.

Each EC2 neuron also targets a single CA1 neuron with exclusively inhibitory forward connections with a constant inhibitory weight sufficient to produce the "difference" between the input signal provided to EC2 and the reconstructed signal generated by CA3.

Each CA1 neuron sends output directly to a corresponding ERC5 neuron with a one-to-one mapping.

5. The ERC5 Component

The final component is meant to simulate the ERC layer 5 (ERC5), which represents the principle output of the hippocampus. We simulate ERC5 with the same number units as EC2, and it receives input solely from CA1. Activity within EC5 will occur at the beginning of a theta cycle, which corresponds to the time at which new input will be stimulated in EC2. Thus incoming and outgoing activity within EC is phase-aligned.

6. Inhibitory interneurons

All areas of the hippocampus have extensive inhibitory interneurons, and receive extensive input from modulatory regions. To simulate this, each neuron in the model receives an inhibitory signal that exhibits both tonic and phasic behavior. This function serves as a feedback control to prevent the activity of the system from growing without bound, and acts as the pattern generator that creates gamma-phase oscillations. Activity within each region progressively increases inhibitory feedback resulting in oscillations of bursting followed by silence. This oscillatory behavior was tuned to match slow gamma phase oscillations, at approximately 40hz (Mizuseki et al., 2009).

Results

Binding simple conjunctions

First we tested a simple relation represented in a pattern of asynchronous firing (Figure 5.4). This code was meant to represent a simple predicate-consequent pair (If A then B). This relationship was coded with a simple pattern of synchrony, a single lateral EC unit representing A and a single medial EC unit representing IF fired in unison to represent that A was present and bound to the predicate condition. The units representing B and THEN fire 180 degrees gamma phase advanced with respect to the A units, to represent consequent, and to represent that A

did not occupy the consequent slot (c.f. Holyoak & Hummel 1997). During training and recall, we repeated the input pattern three times, on three subsequent theta cycles. After training several intervening patterns, we test the network by presenting a partial input (only the "A + IF" portion of the conjunction) during which only four of the input neurons repeat their pattern of activation. Correct performance corresponds to the ability of the network to complete the partial pattern, reactivating the missing two neurons (B + THEN), at the correct time (180 degrees gamma phase advanced with respect to A), and to provide a code for the portion of the original signal that was "filled in" (i.e., to deconvolve the input from the filled in signal in order to distinguish that which was recalled from that which was provided by the input).

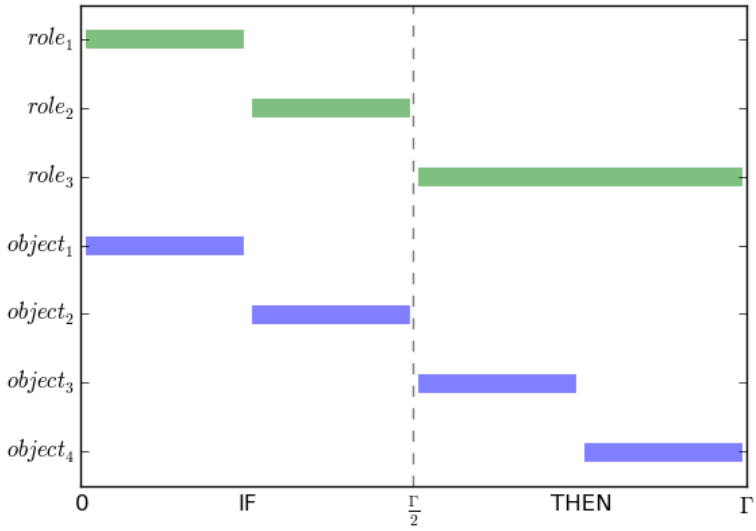


Figure 5.4 - Structure and Timing of Simple Relation

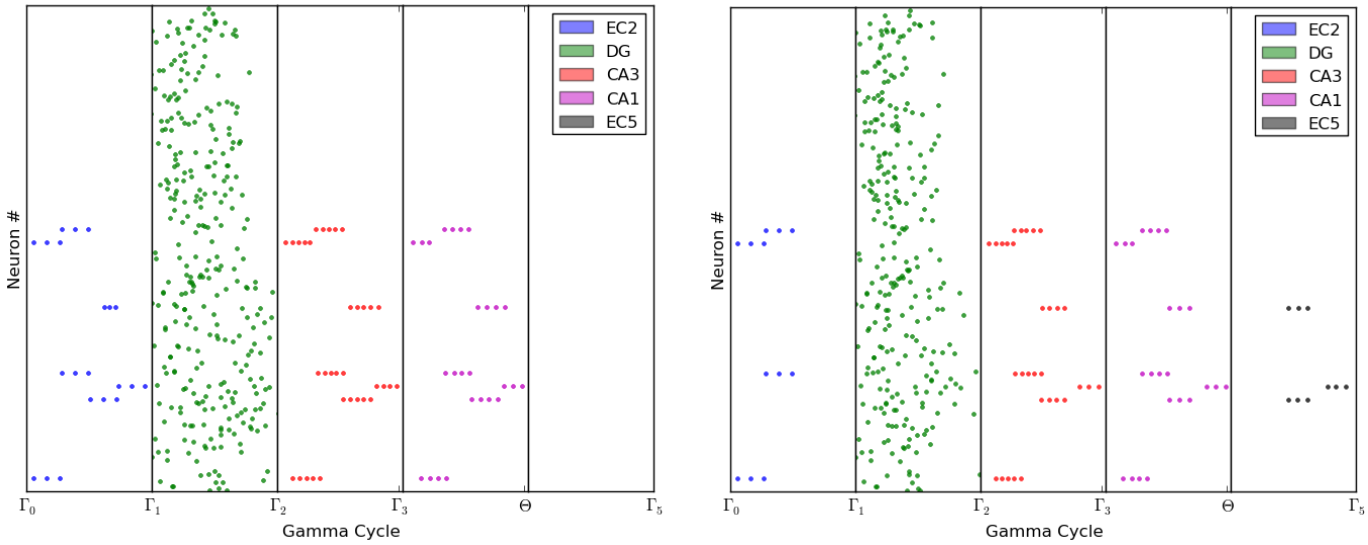


Figure 5.5 - a) Initial Training Behavior and b) Reconstruction From Partial Pattern Behavior

The model is successfully able to reconstruct the simple relation (Figure 5.5). Within each theta cycle, the first gamma cycle, activates the sub-set of neurons in EC2 that represent the simple "IF A THEN B" conjunction. During the second gamma cycle, these units activate a distributed representation in DG which uniquely codes for the phase and frequency information present in this particular conjunction of inputs. Each DG cell codes for a particular complex relationship between an input, its frequency, and its phase with the strength of the connection between the EC cell and the DG cell encoding the frequency information, and the oscillatory delay of each DG cell relative to the theta phase encoding the phase information. On the third gamma cycle, these complex conjunctive codes are summed in CA3 to produce the interference pattern of the many different phase and frequencies of the DG population. The noise produced by the initially random weights and delays of DG cells averages out, leaving only the timings of the original inputs to drive the cells in CA3. On the fourth gamma cycle, CA1 finds the mismatch between the activated input units, and CA3's recovered temporal code, reconstructing both the correct timing and input identities. Finally, on the fifth gamma cycle the mismatch between the input and the CA1 code is returned to the EC5.

Compressed codes

While the above encodes a simple rule at a single scale, the model is capable of encoding more complex configurations of input stimuli. Using a more complex pattern of inputs with multiple signals over several seconds produces a different pattern of output. In this case the

hippocampal model cortex does not reproduce exactly the pattern of units as the input signal, rather, the output is compressive-it does not reproduce any spikes that could be interpolated from adjacent spikes. In this case the output resembles a compressed or low-resolution representation of the input that eschews redundant data (Figure 5.6).

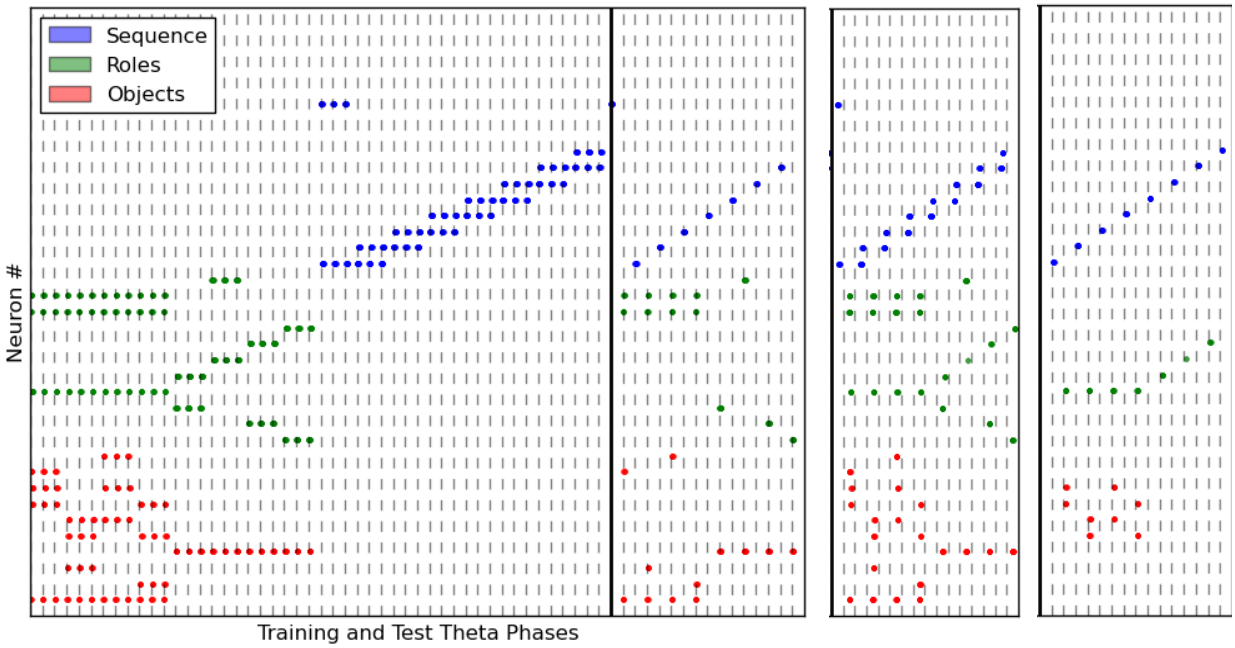


Figure 5.6 - Training, Reconstruction, and Compression of Complex, Disparate Input Stimulii A) Training and Test Stimulus in EC2 B) Reconstruction in CA3 C) Output of EC5

This compressed code closely resembles compressed path sequences observed in sleeping rats after training on a linear track task (Lee & Wilson 2002). In the case of the model, the compression arises because dentate cells which are sensitive to particular phase or frequency relationships are not sensitive to repeats. A frequency sensitive cell responds if a certain input frequency is present, not to each pulse within that input signal.

Binding biologically realistic codes

Cellular recordings of the MEC have found considerably more complex codes than the simple conjunctive binding-by-synchrony one presented above. Spatial representations, are especially well characterized; recent work has demonstrated that grid cells, as well as other classes of spatial invariant EC cells, exhibit sensitivity to periodic 2-D waves of firing oriented along a

restricted set of directions, and that the stable, spatially modulated response pattern of such cells arise from 1) the number of such wave-like components a cell is sensitive to, 2) the angle of these of those waves relative to distal landmarks and 3) the orientation of these waves relative to each other (Gustafson & Daw, 2011; Krupic, Burgess, & O'Keefe, 2012).

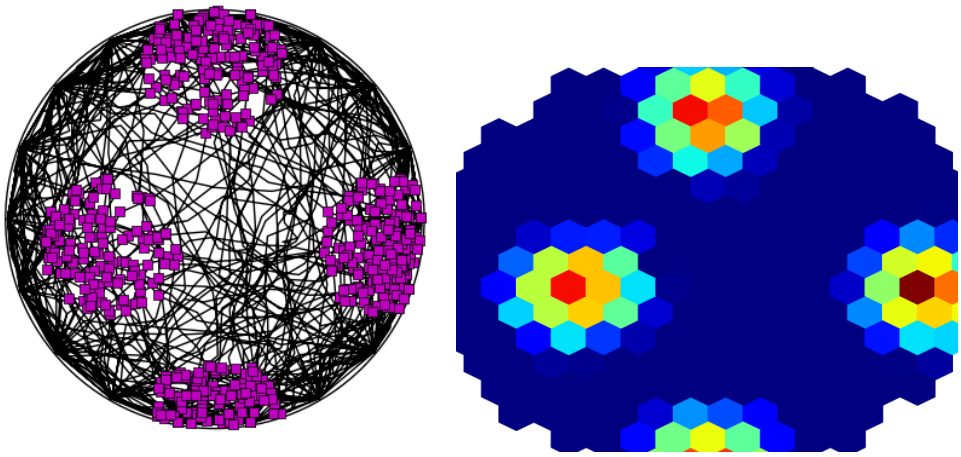


Figure 5.7 - Spatially Invariant Grid Cell Behavior Simulated in EC2 Based On Environment Position

This provides a more biologically realistic code for the model to bind. In our second test, we simulated a "rat" running a path in a simulated, 2D, 1m circular enclosure. This path was encoded into the firing of simulated grid (coding coordinate location), and head direction cells (coding angle of motion). We used six grid cells for each of three different scales receptive fields (0.5m, 0.25m, and 0.125m fields) that loosely mirror those observed in rat hippocampus (as reported in Krupic et al., 2012). We used eight head direction cells which coded for the orientation of the rat as it moved through the enclosure with receptive fields at 45 degree angles relative to each other (Figure 5.8). The rat's position and orientation (at any of our scales) could therefore be decoded from the activity of the grid and head direction cells.

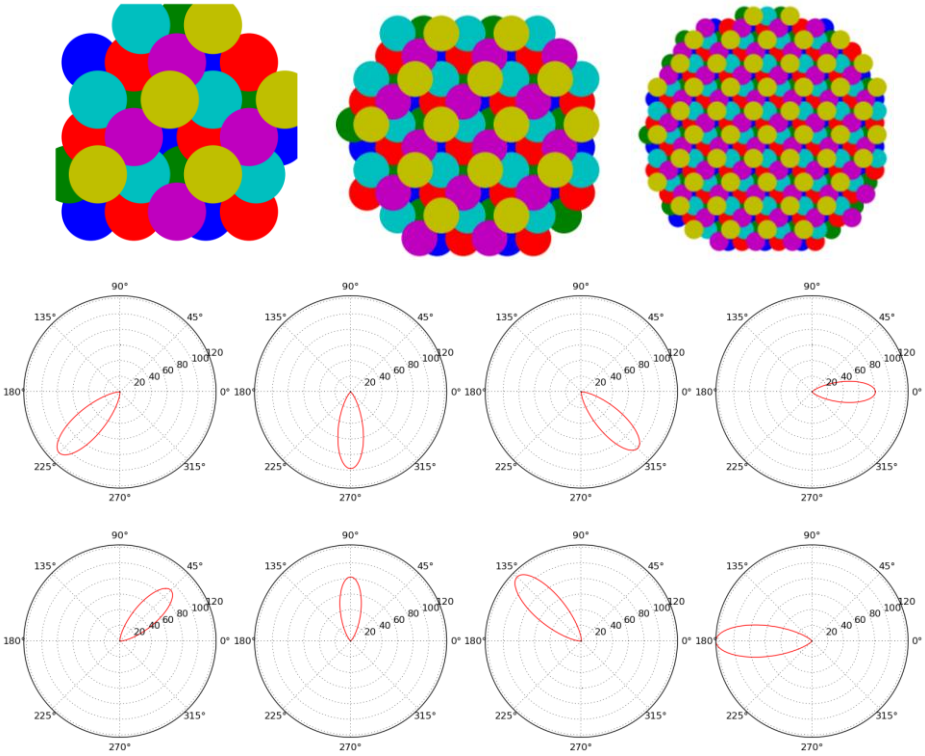


Figure 5.8 - Spatially Invariant Grid Fields at Multiple Scales and Direction Cell Sensitivity

The activity trajectories generated by the grid and head direction cells from the underlying spatial path was input as a training sequence to the hippocampal model in the same manner as the simple rule training above. Unlike the above encodings, each grid scale was trained separately to reduce the total number of combinations. In addition, EC3 was used to produce conjunctive grid-direction cells to provide additional input context to further aid in DG expansion and disambiguation of input highly similar input patterns. Timings were encoded by the rat's simulated motion in real time with roughly one transition between large-scale grid cell fields per theta cycle, and transitions between finer-scale grid cells taking place within gamma cycles. We measured output both by examining reconstructed and prospective paths and by examining hippocampal cells for "place cell" codings.

After training, probing the hippocampal model with a partial input sequence (c.f. Lisman & Redish 2009) produced a reconstructed (relative to previous experience) or prospective (relative

to current location) path (Figure 5.9). While reconstructed paths were generally accurate, they also contained short cuts, path inversions, and non-adjacent moves (Fig. 5.10)

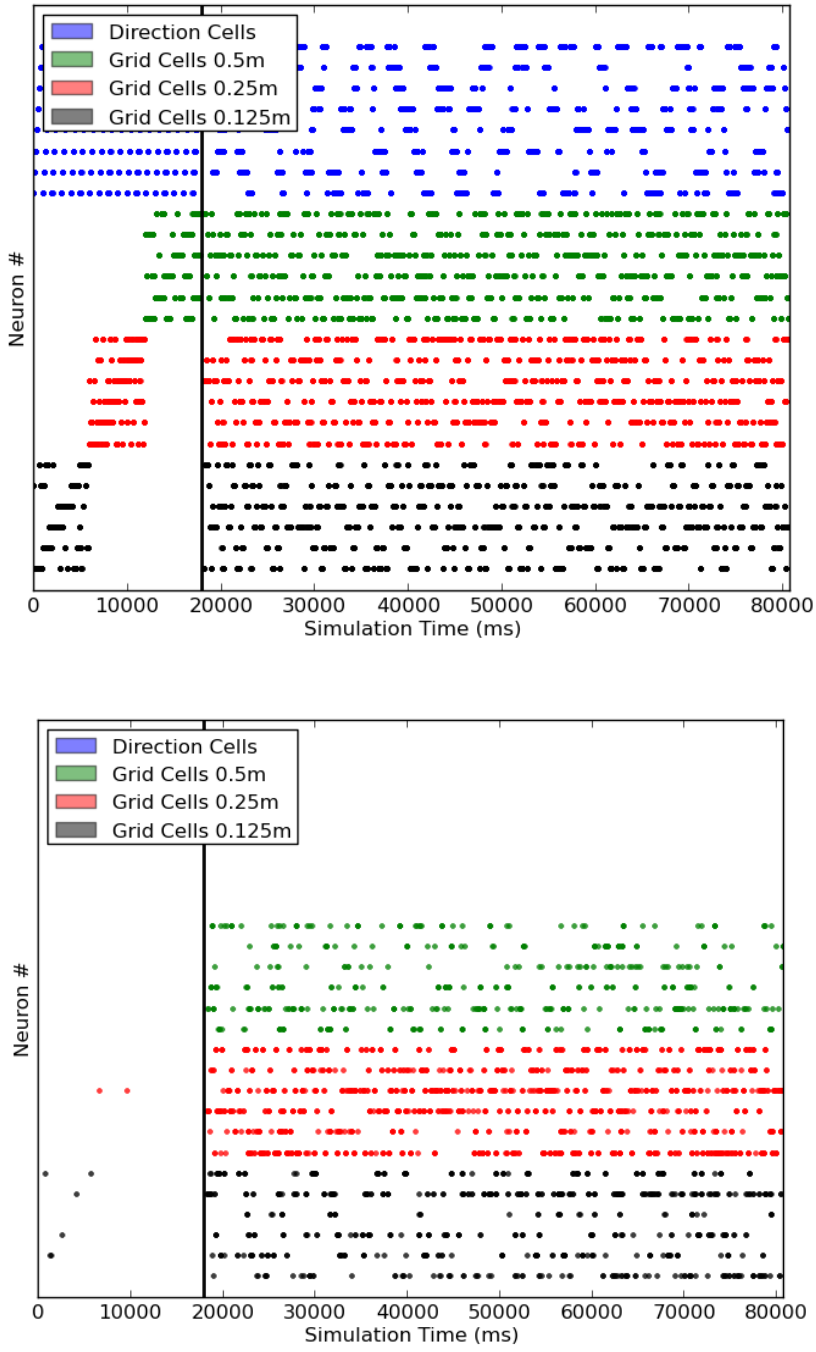


Figure 5.9 - Grid and Direction Cell Path Training and Reconstruction EC2 Input (Top) and EC5 Output (Bottom)

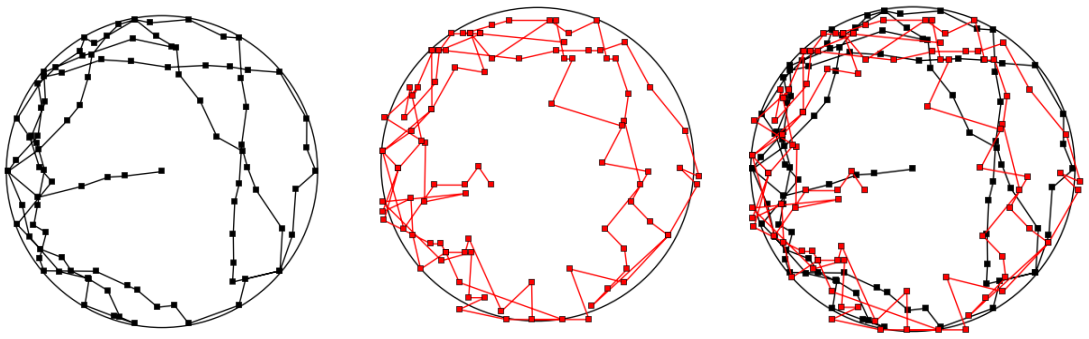


Figure 5.10 - Actual (Black) versus Reconstructed/Predicted (Red) Path

"Place" Cell Activity

The RMBER base model was augmented slightly to explore complementary "place" cell dynamics. As noted previously, forward projections were added between EC2 and EC3 to produce conjunctive grid-direction cell activity. Further, forward projections were added between EC and CA1 to implement a mechanism for conversion from spatially invariant grid cells to spatially bound place cells(Figure 5.11a) (Moser, Kropff, & Moser 2008).

A set of observed hippocampal “place” cells emerged during training (Figure 5.11b). These cells’ firing rates were modulated only when the “rat” occupied a particular point in space. Figure 5.11b further depicts the spatial variance for which each “place” cell was observed to be active – roughly fit within a bounding circle. This region of activity is directly correlated to the granularity of grid cells whose particular summation corresponds to the activation of each individual “place” cell as depicted in Figure 5.11a. A sub-set of these cells are modulated by the theta phase as well, and thus only become active when the “rat” occupies the location at a certain moment in the temporal sequence of grid cell firing (as in hippocampal “time” cells c.f. Farovik, Dupont, & Eichenbaum, 2010).

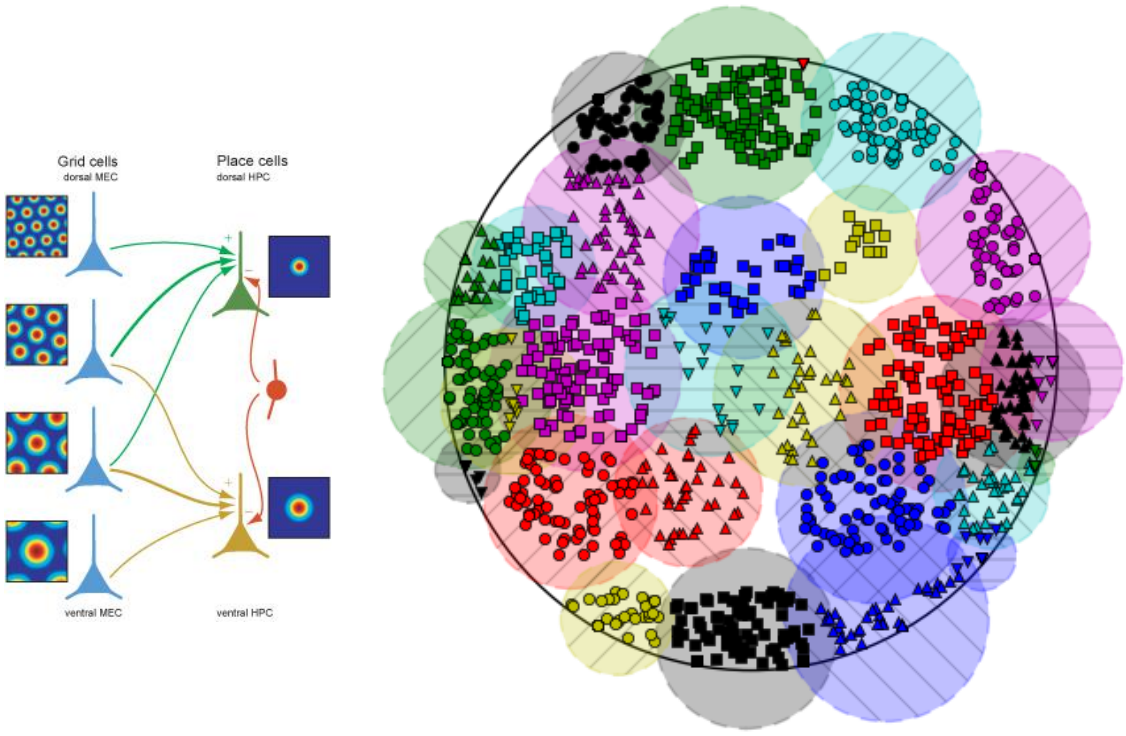


Figure 5.11 (From Moser et al. 2008- a) Spatially Invariant Grid to Spatially Bound Place Cell Conversion (b) Place Cell Activity Observed in CA1 (each marker style and color denote a different cell)

Discussion

We presented a model for binding together arbitrary relations between patterns of entorhinal cortex activity. At the neural level, our chief addition is allowing spike-timing-dependent-plasticity to tune the connectivity of inhibitory interneurons, which allows the model to learn both differences in connection strength and oscillatory delays. This delay coding allows it to learn the order of input activations to create a dynamic, concatenative code capable of capturing everything from simple, asynchronously encoded, rule-based conjunctions, to biologically realistic grid cell activity. Relations among input cells are encoded from the relative frequency and phases of EC input firings in the relatively high-dimensional space of the DG where conjunctive cells are sensitive to the co-occurance of input firing frequencies and phase delays (with respect to the theta rhythm). Relations can be (re)constructed by collapsing the firing codes of DG conjunctive cells to a smaller number of recurrent CA3 cells, before finally being mapped back into the EC via CA1. Spike-timing dependent plasticity modulates both the strength of the weights and length of the oscillatory delays to fill in gaps in input signals with

patterns that match both the items and relations present in previously encoded patterns of relations.

Encoding this oscillatory representation allows the model to be truly compositional. Neural activity corresponds to the "symbols" present in the input, and fluctuations of this activity through time (i.e., the oscillations), correspond to the dynamic binding of these symbols to particular values (e.g., activation of a time cell signifies that the rat is at a particular point in a sequence). Summing the dynamics of individual "symbols" creates an interference pattern that uniquely corresponds to the composition created by binding those symbols together. Decomposing this interference patterns in the time domain creates a novel code for the relations between symbols since it captures the order and frequency of symbols' activation. By encoding these coefficients, the hippocampus creates a database of possible bindings. Similar input patterns of bindings elicit activation of the nearest structurally similar matches, without being unduly driven by the surface features of the particular symbols involved.

In signal processing terms, this process is similar to upsampling a signal to create a high-resolution version, applying an interpolating filter to fill gaps, and then recompressing. The power of the model comes from the filter being dynamic; spike-timing dependent plasticity adjusts both what and when elements are interpolated, such that rather than filling gaps solely with noise or information from adjacent bins, gaps can be filled with information encoded during previous, similar experiences whether that similarity is due to shared items (as encoded by neural activity) or relations (as encoded by oscillatory dynamics). The large number of dentate granule cells relative to the inputs ensures that filter samples a large number of kernel sizes, improving its ability to find the optimal resolution at which to represent the input.

The model is able to capture the diverse input coding schemes used by its multi-modal inputs by transforming them into a single common code. This transformation has much in common with a discrete Fourier transforms (and wavelet functions), and indeed connectionist models with similar network architecture can perform such transforms (Silvescu, 1999; Velik, 2008). This transform creates a common language that allows diversely coded hippocampal inputs to be stored in a common compressed code, without losing the particulars of oscillatory frequency or phase that are were critical to the relational structure of the encoded input. Note however, that

this code does not *store* or *recall* the complex dynamics of the inputs per se, it stores a compressed code that contains only the information required to *adjust* and *re-align* the current input dynamics to bias them to evolve toward previously experienced dynamics.

This highlights a fundamental contrast between the RMBER model and other models of hippocampal function. Hippocampal models in the tradition of Marr (Marr 1971, McNaughton & Morris 1987, Treves and Rolls 1994, O'Reilly and McClelland 1994, Hasselmo and Wyble 1997, Rolls & Kesner 2006), assume that input information is carried solely by neural activity (i.e., firing rate) rather than by neural dynamics (i.e., phase and frequency), and that these codes all exist at the same resolution or scale. These models therefore, stress the involvement of the hippocampus in creating pattern separated codes that uniquely index particular input patterns followed by a pattern completion step to fill in missing information. Creating orthogonal codes, and filling in absent information is an important part of hippocampal function captured by these models, and the RMBER model does implement this function in the interaction between the high-dimensional DG region and the low-dimensional CA3 region.

However, pattern separation and pattern completion are simply different tunings of a classificatory process, with the former emphasizing fine-grained, specific categories that infer very little about a particular exemplar, and the latter coarse-grained, general ones with many associated features (indeed, both pattern separation and completion can be implemented by the same system, and often are in cortical models e.g., Carpenter & Grossberg 1988). Recent analyses have begun to emphasize the hippocampus's role not in pattern separation *per se* but in controlling the resolution at which memories are encoded (Aimone, Deng, & Gage, 2011). With dynamic input codes (as in RMBER), that carry information in the time domain (i.e., in oscillatory phase and frequency), and that encode information from multiple scales (e.g., cells representing entire events, and those sub-events) the criteria for resolving the category to which input samples belong must also be dynamic and multi-scale. If I remember a birthday party, and recall that one of the guests left early, I must simultaneously assign them to the "attended the birthday party" category AND the "did not attend the birthday party" category, because both assignments are true (albeit at different scales)! These dynamic, and multi-scale codes are precisely the sort of representations that require hippocampal function, but are not easily captured by the static coding scheme of models focused on pattern separation and

completion because they rely on selecting an appropriate category scale *a priori*, without knowing what information will be required by recall.

The RMBER model circumvents this by directly encoding neural dynamics at multiple scales. Instead of encoding and reconstructing patterns of activity (the party, or it's sub-elements), RMBER encodes and reconstructs both cells' activity and the relative order of that activity. This coding is not perfect, the incoming neural signal undergoes lossy compression, which tends to underweight high frequency and periodic input oscillations (these fade into the background as context). Further, representation of fine-grained oscillatory activity is limited by the rate at which gamma frequency oscillations "sample" the slower theta rhythm (i.e., the model cannot capture changes in the EC inputs that occur more frequently than gamma because these changes cannot be coordinated between the EC and the DG). In addition, larger scales (e.g., the whole birthday party) are represented by stringing together multiple theta cycles, and reconstruction at this scale may fail either due to an insufficient number of DG cells sensitive to the boundaries between theta cycles (i.e., the event is too long to be remembered as a single "chunk"), or simply due to extrinsic activity interrupting reconstruction (i.e., memory interruptions). However, these draw backs are necessary to create a truly concatenative code, RMBER reactivates neurons, at the frequency, and the phase order relative to each other as in the original event. It can "gate" in and out elements as needed. In the birthday party example, while the guest is present, the cells representing them are active, and when the guest leaves, they go silent without modifying the activity of cells representing other aspects of the party.

Another tradition of hippocampal models stresses the importance of the hippocampus to single-trial learning (Hopfield 1982, Minai & Levy, 1993, Deng, Aimone, & Gage 2010), though his tradition has recently been nuanced with the addition of models that can both rapidly encode simple associations, and generalize across larger categories (Kummeran & McClelland 2012). The goal of this rapid learning is often to train neocortical networks on infrequently experienced events (McClelland, McNaughton, & O'Reilly 1995). Since neocortical models typically rely on learning algorithms that infer the structure of their inputs via associative learning rules (i.e., they capture observed correlations) they are highly effective at learning structured semantic information about the world (e.g., dogs typically have tails), but is far less effective at capturing arbitrary, accidental, combinatorially complex, or ill-structured configurations of stimuli. One-

trial learning hippocampal model is meant to solve this by providing a learning signal that captures infrequently observed conjunctions.

However, it is worth highlighting that the neocortical systems are not "accidentally" discarding rare instances. Ignoring rare, arbitrary, or complex configurations is a feature, not a bug in associative learning systems. A single, random instance of a low-probability event (e.g., this dog is painted blue) is *not* a structured pattern of associations. It is not probabilistically likely, and an associative learning system attempting to build a statistically accurate, generative model of the world ought *not* to update its model because this peculiar instance does *not* strongly predict future events; it is possible, but not probable. Associative learning systems should either accept that this example is acceptable variance within their model ("sometimes dogs are blue") and not use it to guide future recall, or adjust the parameters to encode this single instance ("the blue dog I saw on Wednesday") and accept that this over-fitting *biases* the model (this is known as the "bias-variance" trade-off c.f. Geman, Bienenstock, & Doursat, 1992). Using a hippocampal model to train a neocortex model on individual instances simply swings the pendulum towards bias. This may be desirable, certain highly salient events may be worth over-fitting, but the hippocampal model would need to know presciently at encoding time which events would end up being important later.

The RMBER model removes the need to fit individual instances with associative parameters. Instead it creates a code that instructs the associative memory how to configure the information it already has. A good associative memory can learn that there is a cluster of features that correspond to the object we call a "chair" and another cluster of spatial features that correspond to the location "in front of the desk." However, if the associative memory system linked "chair" and "in front of desk," it would damage the independence of those categories because chairs are not locations, nor are locations chairs. Instead, the RMBER system can simply encode the transient firing *dynamics* produced by the cortex's observation of an overlap between the spatial and non-spatial category, and by reconstructing this code later can adjust cortical dynamics toward the previous state of "chair in front of the desk," without needing to create a conjunctive category tuned to the surface features of chairs in front of desks. The model is thus capable of single-trial learning, but does so via a mechanism that captures

relationships between independent categories without fitting a category to the configuration of features present in an individual instance.

This of course predicts that individual EC cells correspond to abstract, multi-modal conceptual categories (e.g., spatial locations, objects, parts of speech, emotions, etc). If the associative cortex has done a good job of clustering similar items into conceptual categories, than activation of EC components should be relatively independent (though if they are not, RMEBER is perfectly capable of binding across these proto-concepts until the clustering algorithm can sort more successfully). Therefore, the *relative* firing dynamics across these different components ought to correspond to the relations between independent concepts (although no claim is made about particulars of these dynamics, they can be arbitrary or even perverse). Remembering which concepts were bound to which relations is a matter of reconstructing their firing timings relative to each other.

This prediction speaks to the third, and most recent tradition of hippocampal models: those that take seriously information encoded in oscillatory phase and frequency. These models attempt to explain hippocampal involvement in arranging and replaying input sequences whether the dynamics of those input correspond to spatial (McNaughton et al., 2006, Fuhs & Touretzky, 2006, Burgess, Barry, & O'Keefe 2007, Giocomo et al. 2007, Kropff & Treves 2008), or temporal (Hasslemo, Bodelón, & Wyble, 2002) trajectories. This tradition of modeling has been closely tied to findings on theta phase precession. One unresolved question of these models is whether theta precession is a result of oscillatory interference between gamma and theta rhythms (O'Keefe & Recce, 1993, Bose, Booth & Recce, 1999, Lengyel, Szatmáry, & Erdi, 2003, Burgess, Barry, & O'Keefe 2007, Hasselmo 2008), or a result of sequence read out requiring ever more compressed sequences causing initially encoded elements to move retrograde relative to theta (Tsodyks et al. 1996, Jeneson & Lisman 1996, Hasselmo and Eichenbaum 2005). The RMBER model highlights possible common ground between these two perspectives. In RMBER, phase precession occurs due to interference between the oscillations of phase sensitive DG cells when their firing is collapsed into a single gamma cycle of activity in CA3. However, since DG phase sensitive cells index the position of elements in a sequence (with theta oscillations serving as a "carrier wave" for the sequence as a whole), their interference is also a required consequence sequence reconstruction. In essence, RMBER accomplishes sequence reconstruction *via* a

mechanism of oscillatory interference. Each element in the sequence is encoded within a gamma-scale oscillatory waveform, and the interference pattern of the waveforms at the theta-scale corresponds to the complete sequence.

Functionally, the RMBER model presents a slightly different interpretation for episodic memory and other forms of mental construction. Some previous theories of episodic memory (Tulving, 1984; Tulving, 2002), stressed the bottom-up, sensorial nature of episodes, and that the unique, instance-like nature of such memories came from an encoding of sensory and contextual details. This framework explains episodic memory as a recording of sensory experience, and an episode as a recalled instance as contrasted with the more general, instance-independent reconstructive process of semantic memory. The RMBER model framework, argues instead that the specific, contextually elaborate nature of episodic memories comes from the coherence of their reconstruction, and that episodes are not recorded instances, but compressed codes that do not themselves contain any sensory details, but only instructions for tuning pre-existing associative, and perceptual systems to recover the appropriate sensory details. In this formulation, the hippocampally located episodic memory system and the neocortical located semantic or gist-based memory systems do not act in parallel, but in concert. The semantic systems are not used to "fill in" missing episodic information, for the episodic information was never recorded, and does not have gaps. Rather the semantic systems are the set of brushes and paints employed by the hippocampus to compose images that look like previous experience. The parallels between the model's information processing and image compression techniques (up sampling, interpolating filters, and DFT) are not coincidental. They argue that episodic memory is better understood as a top-down reconfiguration of previously learned categorical knowledge, and that the only information encoded by the hippocampus is the minimal set of phase and frequency coefficients required to recapitulate the appropriate neocortical dynamics.

References

Abbott, L. F., & Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nature neuroscience*, *3 Suppl*(november), 1178–83. doi:10.1038/81453

Aimone, J. B., Deng, W., & Gage, F. H. (2011). Resolving new memories: a critical look at the dentate gyrus, adult neurogenesis, and pattern separation. *Neuron*, *70*(4), 589–96. doi:10.1016/j.neuron.2011.05.010

Andersen, P., Morris, R. G. M., Amaral, D., Bliss, T., & O'Keefe, J. (2008). *The Hippocampus Book*. Oxford University Press.

Bose, a, Booth, V., & Recce, M. (1999). A temporal mechanism for generating the phase precession of hippocampal place cells. *Journal of computational neuroscience*, *9*(1), 5–30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10946990

Burgess, N., Maguire, E. a, & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron*, *35*(4), 625–41. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12194864

Burgess, N., Barry, C., & O'Keefe, J. (2007). An oscillatory interference model of grid cell firing. *Hippocampus*, *812*, 801–812. doi:10.1002/hipo

Cohen, N., & Squire, L. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science*. Retrieved from http://www.sciencemag.org/content/210/4466/207.short

Cutsuridis, V., Cobb, S., & Graham, B. P. (2010). Encoding and retrieval in a model of the hippocampal CA1 microcircuit. *Hippocampus*, *20*(3), 423–46. doi:10.1002/hipo.20661

Deng, W., Aimone, J. B., & Gage, F. H. (2010). New neurons and new memories: how does adult hippocampal neurogenesis affect learning and memory? *Nature reviews. Neuroscience*, *11*(5), 339–50. doi:10.1038/nrn2822

Dusek, J. a, & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(13), 7109–14. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=21293&tool=pmcentrez&rendertype=abstract

Eichenbaum, H. (2000). A cortical-hippocampal system for declarative memory. *Nature reviews. Neuroscience*, *1*(1), 41–50. doi:10.1038/35036213

Etienne, A. S., & Jeffery, K. J. (2004). Path integration in mammals. *Hippocampus*, *14*(2), 180–92. doi:10.1002/hipo.10173

Farovik, A., Dupont, L. M., & Eichenbaum, H. (2010). Distinct roles for dorsal CA3 and CA1 in memory for sequential nonspatial events. *Learning & memory (Cold Spring Harbor, N.Y.)*, *17*(1), 12–17. doi:10.1101/lm.1616209

Fuhs, M. C., & Touretzky, D. S. (2006). A spin glass model of path integration in rat medial entorhinal cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *26*(16), 4266–76. doi:10.1523/JNEUROSCI.4353-05.2006

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/neco.1992.4.1.1

Gerstner, W., Kreiter, a K., Markram, H., & Herz, a V. (1997). Neural codes: firing rates and beyond. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(24), 12740–1. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=34168&tool=pmcentrez&rendertype=abstract

Giocomo, L. M., Zilli, E. a, Fransén, E., & Hasselmo, M. E. (2007). Temporal frequency of subthreshold oscillations scales with entorhinal grid cell field spacing. *Science (New York, N.Y.)*, *315*(5819), 1719–22. doi:10.1126/science.1139207

Gustafson, N. J., & Daw, N. D. (2011). Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLoS computational biology*, *7*(10), e1002235. doi:10.1371/journal.pcbi.1002235

Hargreaves, E. L., Rao, G., Lee, I., & Knierim, J. J. (2005). Major dissociation between medial and lateral entorhinal input to dorsal hippocampus. *Science (New York, N.Y.)*, *308*(5729), 1792–4. doi:10.1126/science.1110449

Hasselmo, M. E., & Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: simulating effects of scopolamine on human memory function. *Behavioural brain research*, *89*(1-2), 1–34. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9475612

Hasselmo, M. E., Bodelón, C., & Wyble, B. P. (2002). A proposed function for hippocampal theta rhythm: separate phases of encoding and retrieval enhance reversal of prior learning. *Neural computation*, *14*(4), 793–817. doi:10.1162/089976602317318965

Hasselmo, M. E., & Eichenbaum, H. (2005). Hippocampal mechanisms for the context-dependent retrieval of episodes. *Neural networks : the official journal of the International Neural Network Society*, *18*(9), 1172–90. doi:10.1016/j.neunet.2005.08.007

Hasselmo, M. E. (2008). Temporally structured replay of neural activity in a model of entorhinal cortex, hippocampus and postsubiculum. *The European journal of neuroscience*, *28*(7), 1301–15. doi:10.1111/j.1460-9568.2008.06437.x

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. a. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(5), 1726–31. doi:10.1073/pnas.0610561104

Heckers, S., Zalesak, M., Weiss, A. P., Ditman, T., & Titone, D. (2004). Hippocampal activation during transitive inference in humans. *Hippocampus*, *14*(2), 153–62. doi:10.1002/hipo.10189F

Hubel, D., & Wiesel, T. (1963). Shape and arrangement of columns in cat's striate cortex. *The Journal of physiology*, 559–568. Retrieved from http://jp.physoc.org/content/165/3/559.full.pdf

Jensen, O., & Lisman, J. E. (2005). Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer. *Trends in neurosciences*, *28*(2), 67–72. doi:10.1016/j.tins.2004.12.001

Konkel, A., & Cohen, N. J. (2009). Relational memory and the hippocampus: representations and methods. *Frontiers in neuroscience*, *3*(2), 166–74. doi:10.3389/neuro.01.023.2009

Kropff, E., & Treves, A. (2008). The emergence of grid cells: Intelligent design or just adaptation? *Hippocampus*, *18*(12), 1256–69. doi:10.1002/hipo.20520

Krupic, J., Burgess, N., & O'Keefe, J. (2012). Neural Representations of Location Composed of Spatially Periodic Bands. *Science*, *337*(6096), 853–857. doi:10.1126/science.1222403

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological review*, *119*(3), 573–616. doi:10.1037/a0028681

Lee, A.K., Wilson, M.A. (2002) Memory of sequential experience in the hippocampus during slow wave sleep. Neuron, 36:1183-1194.

Lengyel, M., Szatmáry, Z., & Erdi, P. (2003). Dynamically detuned oscillations account for the coupled rate and temporal code of place cell firing. *Hippocampus*, *13*(6), 700–14. doi:10.1002/hipo.10116

Lisman, J. and Jensen, O., (1997) The importance of hippocampal gamma oscillation for place cells: A model that accounts for phage precession and spatial shift. in Computational Neuroscience: Trends in Research, Plenum Publishing Corp., New York, NY, pages 683-689,

Lisman, J., & Redish, a D. (2009). Prediction, sequences and the hippocampus. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *364*(1521), 1193–201. doi:10.1098/rstb.2008.0316

Marr, D. (1971). Simple Memory: A Theory for Archicortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *262*(841), 23–81. doi:10.1098/rstb.1971.0078

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*(3), 419–57. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7624455

McNaughton, B., & Morris, R. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*. Retrieved from http://www.sciencedirect.com/science/article/pii/0166223687900117

McNaughton, B. L., Battaglia, F. P., Jensen, O., Moser, E. I., & Moser, M.-B. (2006). Path integration and the neural basis of the "cognitive map". *Nature reviews. Neuroscience*, *7*(8), 663–78. doi:10.1038/nrn1932

Mihalas, S., & Niebur, E. (2009). A Generalized Linear Integrate-and-Fire Neural Model Produces Diverse Spiking Behaviors. *Neural computation*, *718*, 704–718. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/neco.2008.12-07-680

Milford, M., Wyeth, G., & Prasser, D. (2004). RatSLAM: a hippocampal model for simultaneous localization and mapping. *Robotics and Automation, …*, (May 2004). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1307183

Minai, A., & Levy, W. (1993). Sequence learning in a single trial. *INNS World Congr. Neural Netw*. Retrieved from http://secs.ceas.uc.edu/~aminai/papers/minai_wcnn93.pdf

Mizuseki, K., Sirota, A., Pastalkova, E., & Buzsáki, G. (2009). Theta oscillations provide temporal windows for local circuit computation in the entorhinal-hippocampal loop. *Neuron*, *64*(2), 267–80. doi:10.1016/j.neuron.2009.08.037

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual review of neuroscience*, *31*, 69–89. doi:10.1146/annurev.neuro.31.061307.090723

Murata, A., Gallese, V., Luppino, G., Kaseda, M., & Sakata, H. (2000). Selectivity for the Shape , Size , and Orientation of Objects for Grasping in Neurons of Monkey Parietal Area, 2580–2601.

Norman, K. a, & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological review*, *110*(4), 611–46. doi:10.1037/0033-295X.110.4.611

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. *Why People Get Lost*. Oxford, UK: Clarendon Press. Retrieved from http://www.ingentaconnect.com/content/oso/7347120/2010/00000001/00000001/art00006

O'Keefe, J., & Recce, M. L. (1993). Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus*, *3*(3), 317–30. doi:10.1002/hipo.450030307

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus*, *4*(6), 661–82. doi:10.1002/hipo.450040605

O'Reilly, R C, & Rudy, J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus*, *10*(4), 389–97. doi:10.1002/1098-1063(2000)10:4<389::AID-HIPO5>3.0.CO;2-P

O'Reilly, R,.C., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: advances in the complementary learning systems framework. *Trends in cognitive sciences*, *6*(12), 505–510. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12475710

Rolls, E. T., & Kesner, R. P. (2006). A computational theory of hippocampal function, and empirical tests of the theory. *Progress in neurobiology*, *79*(1), 1–48. doi:10.1016/j.pneurobio.2006.04.005

Rolls, E. T. (2010). A computational theory of episodic memory formation in the hippocampus. *Behavioural brain research*, *215*(2), 180–96. doi:10.1016/j.bbr.2010.03.027

Samsonovich, a, & McNaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *17*(15), 5900–20. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9221787

Silvescu, a. (1999). Fourier neural networks. *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339)*, *1*, 488–491. doi:10.1109/IJCNN.1999.831544

Song, S., Miller, K. D., & Abbott, L. F. (2000). Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature neuroscience*, *3*(9), 919–26. doi:10.1038/78829

Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of neurophysiology*, *66*(1), 170–89. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1919665

Treves, A., & Rolls, E. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/hipo.450040319/abstract

Tsodyks, M. V, Skaggs, W. E., Sejnowski, T. J., & McNaughton, B. L. (1996). Population dynamics and theta rhythm phase precession of hippocampal place cell firing: a spiking neuron model. *Hippocampus*, *6*(3), 271–80. doi:10.1002/(SICI)1098-1063(1996)6:3<271::AID-HIPO5>3.0.CO;2-Q

Tulving, E. (1984). Precis of elements of episodic memory. *Behavioral and Brain Sciences*. Retrieved from http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=6709108

Tulving, Endel. (2002). Episodic memory: From mind to brain. *Annual review of psychology*, *53*, 1–25. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev.psych.53.100901.135114

Tulving, Endel, & Markowitsch, H. (1998). Episodic and declarative memory: role of the hippocampus. *Hippocampus*, *204*, 198–204. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1098-1063(1998)8:3%3C198::AID-HIPO2%3E3.0.CO;2-G/abstract

Turrigiano, G G, & Nelson, S. B. (2000). Hebb and homeostasis in neuronal plasticity. *Current opinion in neurobiology*, *10*(3), 358–64. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10851171

Turrigiano, Gina G. (2008). The self-tuning neuron: synaptic scaling of excitatory synapses. *Cell*, *135*(3), 422–35. doi:10.1016/j.cell.2008.10.008

Velik, R. (2008). Discrete Fourier Transform Computation Using Neural Networks. *Computational Intelligence and Security, 2008. CIS'08. International Conference on* (Vol. 1, pp. 120–123). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4724626

Voss, J. L., Warren, D. E., Gonsalves, B. D., Federmeier, K. D., Tranel, D., & Cohen, N. J. (2011). Spontaneous revisitation during visual exploration as a link among strategic behavior, learning, and the hippocampus. *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.1100225108

Warren, D. E., Duff, M. C., Jensen, U., Tranel, D., & Cohen, N. J. (2012). Hiding in plain view: Lesions of the medial temporal lobe impair online representation. *Hippocampus*, *22*(7), 1577–88. doi:10.1002/hipo.21000

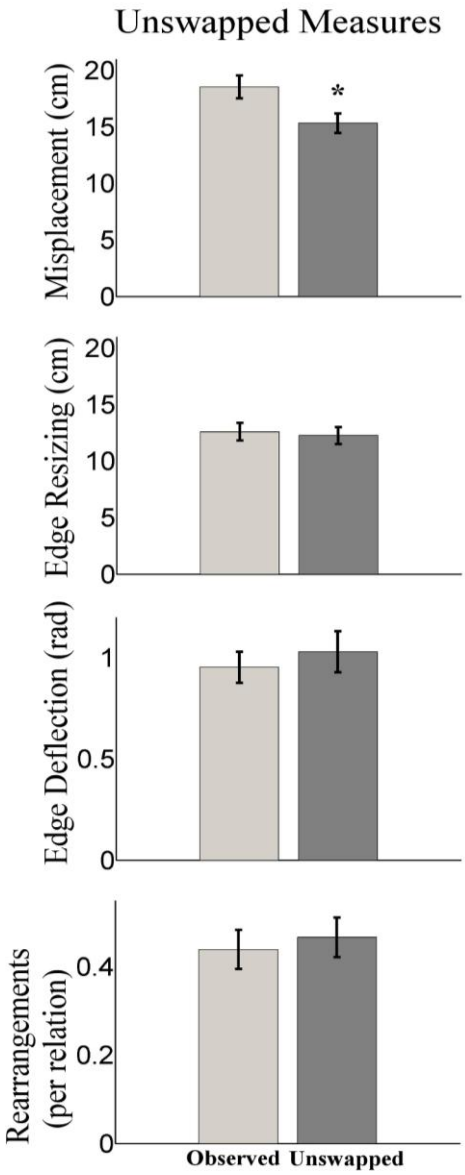
Appendix A

What impact does "unswapping" swap errors have on performance?

Though the previous analysis demonstrates that the presence of swaps is associated with poor item misplacement performance, the swap-misplacement association may be an artifact of poor performance (i.e., swaps are merely a marker of poor performance, and do not specifically cause poor item misplacement). To control for this we created an "unswapping" algorithm that substituted the reconstruction-phase coordinates of an item involved in a swap for the coordinates of the other item in the pair it was swapped with. This method retains the general spatial properties present in the reconstruction, though it does have difficulty coping with trials which contained multiple swaps. Although amnesic participants still performed worse than comparisons after "unswapping" ($F(1, 327)=127.62$, $p<0.00001$), unswapping led to a significant reduction in item misplacement (a one way ANOVA showed main effect of swaps vs. unswapped, $F(1, 327)=5.76$, $p<0.02$) (Figure A.1).

However, while *dropping* data points involved in a swap dramatically improved edge resizing ($F(1,358)=17.15$, $p<0.0001$), edge deflection ($F(1,358)=10.69$, $p<0.002$), and rearrangement ($F(1,358)=9.65$, $p<0.003$); *unswapping* did not (edge resizing $F(1,358)0.13$, $p<0.72$), edge deflection $F(1,358)=0.52$, $p<0.48$), rearrangements $F(1,358)=0.28$, $p<0.6$). This strongly suggests that the primary contribution of swapping is to item misplacement, and not to general spatial reconstruction. Since item misplacement is the overwhelmingly most common measure used in the literature, these results suggest that previous reports overestimate patients' impairment in spatial reconstruction.

Figure A.1


Unswapped Measures
Misplacement (cm)
20
15
10
5
0
*
Edge Resizing (cm)
20
15
10
5
0
Edge Deflection (rad)
1
0.5
0
Rearrangements (per relation)
0.4
0.2
0
Observed
Unswapped

Appendix B

Different patterns of binding relationships require different amounts of information *h* (i.e., *entropy*). *h* is related to the number of possible alternative binding configurations, by:

Eq. B.0

$$h = \log_2 \prod_{i=1}^{t} a_i$$

Where α is the number of possible configurations of binding in a given episode. For example, if there are two possible configurations of a binding, there is exactly 1 bit of entropy required to assign that binding. The number of configurations is determined by a covering mapping function that maps the higher-dimensional set of items onto the lower set different set according to the type of binding required by the experimental apparatus, and the experimental instructions. Let *t* be the number of different types of binding. ERT contained three distinct types of binding with the following mapping functions:

1-to-1 and onto Mappings

For mapping a set of elements one-to-one and onto a second set (e.g., mapping three backgrounds onto three venues):

Eq. B.1

$$\alpha = n!$$

Where *n* is the number of elements in each set. The number of alternatives is simply the number of ways one of the sets can be ordered since this order maps one-to-one and onto the second set.

N-to-M mappings

Mapping the element of set *A* onto the elements of set *B* without allowing for mappings other than one to one, and requiring only that all elements in the smaller set be mapped (e.g., mapping two faces onto one background). In this case:

Eq. B.2

$$\text{let } A \text{ be a set with elements } 1,2,3, \ldots n$$
$$\text{and } B \text{ be a set with elements } \{1,2,3, \ldots m\}$$
$$\text{such that } n \geq m$$

$$\alpha = nPm$$

The number of possible mappings is simply the number of permutations of *n* chosen *m* at a time.

Unconstrained bindings

These bindings allow for any number of mappings to be assigned in any way, without requiring all of the elements in either set to be used. In this case:

Eq. B.3

$$a = \sum_{k=1}^{m} nPk$$

The number of mappings is the number of permutations of *n* chosen [1,2,3,…*m*] at a time. Note that the previous two functions are special cases of this one in which *k*=*m* and *k*=*m*=*n*.

Entropy of semantic constraints

Since *h* depends upon *n, m,* and *k,* participants' performance is constrained both by the experimental apparatus (i.e., it is impossible to make certain, incorrect responses such as binding two scenes to the same venue), and by the task instructions (i.e., participants are

informed that certain assignments are incorrect, such as placing two faces in the same socket, but are not actually prevented from doing so by the reconstruction program). Note that the former set of constraints is stronger than the latter. By calculating two values of *a,* one for those constraints required by the experimental apparatus and another for those constraints suggested by the task instructions it is possible to calculate h, a measure of the amount of information provided by task instructions.

Eq. B.4

$$h = \log_2 \sum_{i=1}^{t} a_i^1 - \sum_{j=1}^{t} a_j^2$$

Where a_i^1 and a_j^2 are the number of possible configurations given the experimental apparatus and given the apparatus and instructions respectively.

Correct relational assignment by chance

Just as the number of configurations is related to which bindings are possible, so is chance performance. "Correct" configuration means matching the bindings present in the study time configuration, and the probability of a reconstructed relation matching a study time relation by chance for any *covering* (*n-to-m* and onto) mapping function is:

Eq. B.5a

$$\text{let } A \text{ be a set with elements } 1,2,3, \ldots n$$
$$\text{and } B \text{ be a set with elements } \{1,2,3, \ldots m\}$$
$$\text{such that } n \geq m$$

$$P\ chance\ = \frac{1}{m}$$

If the function is non-covering, but requires complete assignment of the smaller set:

Eq. B.5b

$$P\ chance\ = \frac{1}{n}$$

And if the function is unconstrained:

Eq. B.5c

$$P\ chance\ = \frac{1}{nm}$$

This allows us to calculate chance for each category of relation given the number of elements to be bound and the relational binding function. We do this for both the experimental apparatus alone, and the experimental apparatus plus experimental instructions:

Table B.1 Chance Level Performance

Relation	*P(chance|apparatus)*	*P(chance|apparatus+instruction)*
Scene-Venue	1/3	1/3
Scene-Face	1/6	1/3
Scene-Socket	1/18	1/3
Scene-Time	1/6	1/3
Venue-Face	1/6	1/3
Venue-Socket	1	1/3
Venue-Time	1/6	1
Face-Socket	1/18	1/6
Face-Time	1/6	1/6
Socket-Time	1/18	1/6

Unsurprisingly, our task instructions provide the most information about relations where chance performance is initially low (i.e., we explain the portions of our task which are most confusing or difficult). Note that the task instructions actually *lower* chance level performance on the Venue-Socket bindings. Without instructions, there are 18 possible sockets to which faces can be assigned, and these sockets map deterministically to their venue (i.e., are contained by their venue); however, the instructions inform participants that in fact, these 18 spatial locations actually act as references to only 6 unique spatial sockets (i.e., the top left socket in the first venue is the same as the top left socket in the second venue). This inter-socket relationship is

reinforced for participants during the experiment since one face appears in each of the 6 unique sockets. However, this method of coding space causes venue-socket bindings to become ambiguous, actually lowering expected performance. In exchange, there is considerable enhancement of other performance estimates related to sockets, since the socket-space is more greatly constrained.

Table B.2 Performance above chance

Relational category	**Chance performance**	**Observed performance**
Scene-Venue	.3333	.71
Scene-Face	.3333	.64
Scene-Socket	.3333	.54
Scene-Time	.3333	.67
Venue-Face	.3333	.79
Venue-Socket	.3333	.59
Venue-Time	1	.95
Face-Socket	.1666	.45
Face-Time	.1666	.42
Socket-Time	.1666	.40

Using the above calculations, and our empirically observed performance values, we can judge how successful participants were at encoding each type of relation by computing their performance above chance *p* according to:

Eq B.6

$$p = \frac{\%correct - chance}{1 - chance}$$

p in this case corresponds to the percentage of relations they correctly reconstructed (for each relational type), beyond those which they could be expected to get correct due to chance.

Relation	*p*
Scene-Venue	0.625
Scene-Face	0.475

Scene-Socket	0.325
Scene-Time	0.52
Venue-Face	0.7
Venue-Socket	0.4
Venue-Time	N/A*
Face-Socket	0.304
Face-Time	0.34
Socket-Time	0.04

Note: Since Venue is deterministically mapped to time, chance on this dimension if you follow the rules should be 1. This means that *p* is undefined since it is impossible to perform above chance. However, performance on this metric was not perfect, meaning either that participants' were not following the rules (and thus were acting with a chance level of 1/3) or were actually led astray by the information in their episodic memories.

Monte-Carlo sampling provides a superior chance metric

However, the existence of two different "chance" standards means that any participant performance is ambiguous: which chance standard are they using (if any!). So long as performance exceeds the lower of the two bounds, we can't know if the additional performance is due to memory or to using the other standard. In addition, real data will contain a certain amount of noise (in both the chance-level standard, and the participant's behavior), making memory performance a statistical property (i.e., are participants are above chance for mnemonic or simply random reasons?). Additionally, if participants are using particular strategies or exhibit biases, their performance might be influenced by fixed memory-independent effects, properly regarded as "semantic" effects, but captured within the "episodic" component of performance.

To avoid this confound, we can use the participants' own reconstructions as a model of their semantic memories. The reconstruction of the study configuration presented during trial *n*, and a reconstruction of the study configuration presented during trial *n-1*, differ only due to the presentation of the intervening input *n* and some pattern of random effects. Thus, if we compute the degree of relational overlap between these two reconstructions (as if one was a

reconstruction of the other), any overlap can only be due to chance constrained by those semantic rules that remained constant between trial *n-1* and trial *n*. Thus, both the mean and variance of reconstruction "performance" computed between two serially appearing reconstructions directly corresponds to the semantic component of performance. These two components become our new a_i and a_j.

By subtracting this semantic component (trial by trial) from our overall study-reconstruction performance value, we can determine the proportion of reconstruction performance that exceeds the performance we would expect due to the proper application of the semantic rules alone. Since this informational gain results exclusively from exposure to the study trial, it can be loosely thought of as the "episodic" component alone, independent of the baseline semantic performance.

This sampling method is especially useful if participants change strategy often or have flexible biases, since it adjusts trial-by-trial to match their most recent tendencies present in their reconstructions.

Episodic information encoded by relational memory

Since we know which configurations are possible given the parameters of the experiment, and how well random assignment should perform, it is possible by observing participants' performance to approximate exactly how much information about the relations they observed they have encoded in relational memory. This is done by inverting Equations 1.5a-c and solving for the number of elements n and $m$present in their mental sets. Using these values of n and m we can compute a, the number of alternative configurations of the subjects actual reconstructions as constrained by their relational memory of the study phase bindings, and h, the amount of entropy present in their reconstructions which exceeds h, the amount of information required to stay within the rules during reconstruction.

Eq. B.7

$$h = \log_2 \prod_{i=1}^{t} a_i - \prod_{j=1}^{t} a_j$$

On any given trial, n and m are always integers making calculating *a* on a trial-by-trial basis simple, however, many experiments calculate normative values of performance for convenience. It is still possible to estimate *a* from normative values of *n* and *m* by taking a weighted average:

$$a' = \frac{(a_{n+1} + a_{n-1})}{2}$$

h helps answer one thorny memory binding question: are there preferred types of memory content (e.g., space, faces, time) irrespective of the entropy captured by that specific type of memory content? For example, there is a lively debate within memory research about hippocampal memory. One theory (hereafter referred to as the "spatial theory") holds that the hippocampus is preferentially involved in spatial processing (e.g., O'Keefe & Nadel, 1978; Maguire et al., 1999; Burgess et al., 2002), because hippocampal lesions disproportionately impair performance on a wide array of spatial tasks, and FMRI studies of the hippocampus show that it is strongly activated by tasks requiring spatial processing. Another theory (the "episodic" theory), suggests that the hippocampus is important for encoding, rich, experiential, episodic memories, and for the ability to mentally "time travel" to these different episodes via a powerful reconstructive recall process (Tulving & Markowitsch, 1998; Nadel, Samsonovich, Ryan, & Moscovitch, 2000; Winocur, Moscovitch, & Bontempi, 2010). This theory is also well supported by evidence, hippocampal amnesics are unable to encode new episodic memories, and impaired at recalling old richly detailed memories (Rosenbaum, Winocur, & Moscovitch, 2001; Sutherland et al., 2001) or imagining new experiences based on old memories (Hassabis et al., 2007). A third theory (the "relational" theory), argues that the hippocampus is critical for encoding and recalling complex or arbitrary relations between items (Dusek & Eichenbaum, 1997; H Eichenbaum, 2000; Howard Eichenbaum & Cohen, 2001; Konkel, Warren, Duff, Tranel, & Cohen, 2008), and is supported by evidence that hippocampal lesions impair performance on many kinds of relational tasks (e.g., transitive inference, transverse patterning), and that hippocampal activity increases during tasks requiring relational memory.

The analyses presented here provide a method for quantitative falsification of these theories. By computing the entropy required to accurately configure a particular relational binding or pattern of bindings, it is possible to precisely quantify *complexity.* By measuring how much information is embedded in the experimental apparatus and task instructions we can compute how *semanticized* (i.e., constrained by rules which transcend a particular episode), or *arbitrary* (i.e., unconstrained by such rules), a particular binding is. By computing performance on particular trials, we can find how much *episodic* or *relational* information and about what kind of bindings (*object-space, object-time, space-time)* participants were able to encode and recall. Each theory makes strong predictions about these data.

The spatial theory predicts that hippocampal impairments will preferentially or exclusively impact spatial relations, partially sparing rule bound information, and bindings such as *object-time*, and that this impairment will hold regardless of the underlying complexity of relations (i.e., a spatial relation which requires configuring a 3-to-3 mapping will proportionally more impaired than a non-spatial relation requiring a 6 to 6 mapping).

The episodic theory predicts that hippocampal impairments will disproportionally impact non-semantic information, and may preferentially impact temporal bindings. These impairments should also hold regardless of relational complexity.

The relational theory predicts, that hippocampal impairments should primarily affect *complex* or *arbitrary* bindings regardless of content type or whether the bindings is constrained by semantic or episodic information. (i.e., 6-to-6 bindings will be harder than 3-to-3).

Further, all three theories could prove incorrect. If, despite extensive evidence to the contrary, the hippocampus was secretly an item memory region, and impairment disproportionately struck *object-space* and *object-time* bindings, it would be possible to determine this, and such a finding would argue for the rejection of all three theories.

Appendix C

Realignment algorithm

The realignment algorithm is a recursive parser. It accepts as arguments a "source" and a "target" sequence composed using the grammar outlined in the method sections. It identifies the portion of those two sequences that does not match. It assumes that these symbols constitute a "chunk," and then applies attempts every transform of the mismatched region that does not place any symbol in a location where it began (since we already know that no symbol is in the correct location), and does not make any cuts in the chunk.

If the number of symbols is odd, these transforms exactly correspond to the (*n*-1) ways it is possible to rotate the symbols through a circular buffer (i.e., moving the {1…n-1} symbols from the end of the sequence to the beginning while preserving their order). If the number of symbols is odd, there are two additional possible transformations, the reflection of the original sequence, and the reflection of the sequence rotated by exactly *n/2* steps.

Once these transforms have been applied, the algorithm selects the one that maximizes the degree of overlap between the source and target, and applies this transform to the source's string of symbols, effectively updating the source to the new, more overlapping sequence. If any symbols remain mismatched the algorithm recurs.

The algorithm produces an operation cost for realignment proportional to the number of symbols that were moved by transforms, plus the number of recursions. This cost is loosely the number of "steps" required make the source and target match if one can only move all of the mismatched elements as a single chunk.

Appendix D: Supplementary Results

Figure D.1: Reconstruction accuracy by relational type.

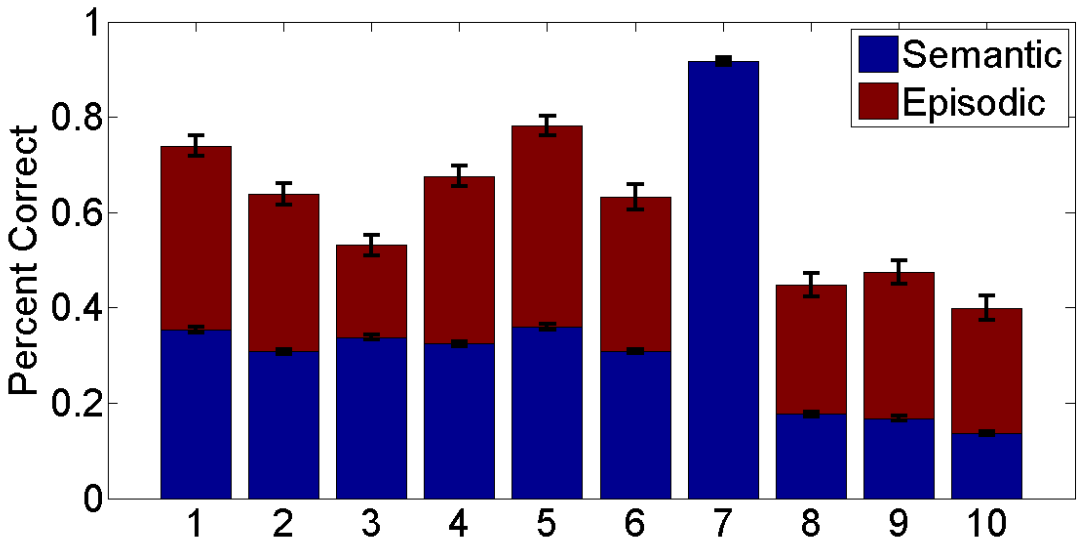


Figure D.1 shows overall reconstruction accuracy parsed into its semantic and episodic components. The ten relational types were 1) Backgrounds to Venues, 2) Backgrounds to Faces, 3) Backgrounds to Sockets, 4), Background to Times, 5) Venue to Faces, 6) Venue to Sockets, 7) Venues to Times, 8) Faces to Sockets, 9) Faces to Times, 10) Sockets to Times.

Performance varied by relational complexity and arbitrariness. Performance on simpler relations, such as Background-Venue (a 3-to-3 mapping) was generally higher than that on moderate complexity mappings such as Background-Face (3-to-6), and high complexity mappings such as Face-Socket (6-to-6). In addition, semantic performance closely tracked the level of chance predicted analytically. The first relation (Background to Venue) was one to one, but had only three elements to map to three elements, giving it an estimated chance level of 33%, very close to the semantic performance suggests. The next five relations were all two-to-one mappings with three and six elements also giving an estimated chance of 33%, again, closely approximating the semantic component. The seventh relation was completely deterministic with a 100% estimated chance level performance, again, very closely approximating the semantic component. The final three relations were one to one involving six elements in each set, and with estimated chance level of 16.7%, again very closely approximating the semantic component.

A 2-way group-level ANOVA with one factor of memory component (episodic v. semantic), and one of relational type found significant effects of both, and a significant interaction (memory component: d.f.=1, $F=104.03$, $P<0.0001$; relational type: d.f. =9, $F=94.46$, $P<0.0001$; interaction: d.f. = 9, $F=348.1$, $P<0.0001$).