Journal of Mathematical Psychology 59 (2014) 18–29

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Mathematical learning theory through time

Marc W. Howard

Department of Psychology and Center for Memory and Brain, Boston University, United States

HIGHLIGHTS

- Traces themes established by stimulus sampling theory through subsequent memory models.
- Categorizes memory models according to the properties of the memory representations they generate.
- Describes recent neuroscientific results that place constraints on memory models.

ARTICLE INFO

Article history:
Available online 17 October 2013

Keywords:
Memory models
Cognitive neuroscience

ABSTRACT

Stimulus sampling theory (SST: Estes, 1950, 1955a,b, 1959) was the first rigorous mathematical model of learning that posited a central role for an abstract cognitive representation distinct from the stimulus or the response. SST posited that (a) conditioning takes place not on the nominal stimulus presented to the learner, but on a cognitive representation caused by the nominal stimulus, and (b) the cognitive representation caused by a nominal stimulus changes gradually across presentations of that stimulus. Retrieved temporal context models assume that (a) a distributed representation of temporal context changes gradually over time in response to the studied stimuli, and (b) repeating a stimulus can recover a prior state of temporal context. We trace the evolution of these ideas from the early work on SST, and argue that recent neuroscientific evidence provides a physical basis for the abstract models that Estes envisioned more than a half-century ago.

1. Introduction

Scientists working in mathematical learning theory wrote down equations implementing elementary psychological mechanisms. These mechanisms were then treated analytically to generate precise behavioral predictions for a variety of experimental settings. Critically, the equations were not an exercise in simple curve-fitting of behavioral data, but a concrete hypothesis about how the mind learns. In retrospect, given what was known about systems neurobiology in the 1950s, this was an audacious research program. The brain has, in principle, a huge number of degrees of freedom at its disposal to generate behavior. Writing down correct expressions for the actual physical process supporting memory, given only constraints from behavioral data, seems impossible. In this paper, we follow the implications of two key insights introduced and formalized in stimulus sampling theory (SST) through decades of subsequent memory modeling to contemporary findings from cognitive neuroscience. Even though it must have seemed impossible in the 1950s, we argue that the research program of mathematical learning theory has been largely successful in describing essential features of neural data. Moreover, the key insights of SST were essential in setting the agenda for these developments.

One key insight of SST is that the nominal stimulus – the light or tone presented to the subject – is not isomorphic to the functional stimulus. In Estes (1950), the nominal stimulus evokes a set of "conditioning elements" that can be conditioned to a particular response. In contemporary terms, we might say that the current set of active conditioning elements is the state of a "memory representation" at the time of the presentation of the nominal stimulus. At each moment, the currently active memory representation is conditioned to a response. At later times, the degree to which a particular response will be evoked is determined by the overlap between the currently active memory representation and the stored memory representation in which the response was learned.

The second key insight of SST is the concept that the memory representation following one presentation of a stimulus changes across different presentations of the stimulus. In much the same way that one cannot step into the same river twice, in SST the functional stimulus caused by different presentations of the same nominal stimulus need not be identical. Moreover, in SST, the functional stimulus caused by a particular nominal stimulus changes gradually across multiple presentations of the nominal stimulus (Estes, 1955a,b). This property enabled a treatment of a variety of phenomena that involve sensitivity to temporal variables, such as forgetting, spontaneous recovery, and the spacing effect.

E-mail address: MarcWHoward777@gmail.com.
URL: http://people.bu.edu/marc777/.

0022-2496/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmp.2013.09.003

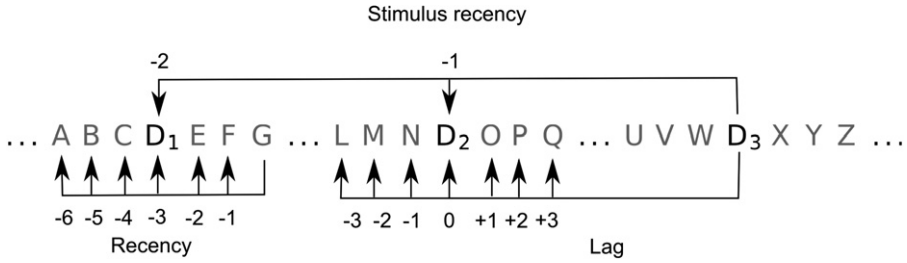

Fig. 1. Schematic for illustrating temporal structure. Models of memory can be distinguished by the similarity of the memory representations across three variables. First, do the states after different presentations of D change over presentations such that the state at D_3 is more similar to the state at D_2 than to D_1? We define the stimulus recency between D_3 and D_2 to be -1; the stimulus recency between D_3 and D_1 is -2. Second, how does the state of memory vary across time around the presentation of a stimulus. That is, is the set of cells active after presentation of G more similar to the representation after presentation of F than it is to the representation after E? We define recency as the difference in serial position between two events. The recency between G and F is -1; the recency between G and E is -2. Third, how does repeating a stimulus affect the relationships in the memory representation? This can be assessed by comparing the memory representation after D_3 to the neighbors of a prior presentation of D, here D_2. We refer to this variable as lag. The lag between D_3 and D_2 is defined to be 0. The lag between D_3 and P is $+2$; the lag between D_3 and N is -1.

In the simple conditioning experiments primarily considered by SST, it was only necessary to consider one nominal stimulus. In subsequent years, memory researchers considered more elaborate verbal learning experiments in which many stimuli are experienced and the categorical distinction between stimulus and response is blurred. For instance, in a free recall experiment, the subject might be presented with a list of 20 words presented one at a time. After a delay, the subject's task is to recall the words from the list in the order they come to mind. The nominal stimuli in this experiment are the sequence of words. But the concept of the response is more ambiguous. Associations between strings of recalls (see, e.g., Bousfield, 1953; Kahana, 1996; Pollio, Kasschau, & DeNise, 1968) suggest that memory must include a network of evolving associations between many stimuli that double as their own response. These associations could be mediated by the functional stimuli caused by each nominal stimulus.

SST specified how the memory representation following a stimulus changes over time, but it did not specify how the relationships between memory representations following different stimuli change as a function of the structure of experience. We will see that subsequent mathematical models of memory distinguish themselves from each other largely by how they respond to this structure. We will review these models in Section 2, making explicit their concrete hypotheses about how memory representations change over time. If we could directly measure the similarity between memory representations at various times, these hypotheses could be directly evaluated.

It is now possible to directly measure the similarity between brain states at different times using a variety of methods. We will discuss three such techniques. Functional magnetic resonance imaging (fMRI) provides an estimate of the oxygenation of blood, believed to be a correlate of neural function, at the spatial scale of millimeters. The pattern of activation across many individual voxels at different points in time can be compared to one another. In human epilepsy patients, electrodes are often placed below the skull for clinical reasons. In many cases these electrodes are too large to record the activity of single neurons, but they can nonetheless record meaningful signals believed to be associated with aspects of cognition. When individual neurons cannot be resolved, oscillatory fluctuations in voltage can be recorded at different anatomical locations. Finally, it is possible to record from many individual neurons using extracellular recording techniques. While it is relatively rare to record at the level of resolution necessary to identify individual neurons in humans, these methods are routinely applied in animal preparations. Extracellular recording can be used to generate a vector of firing rate across neurons, either simultaneously measured or inferred from many single neurons recorded in identical experimental preparations. In each case these methods give rise to a distributed pattern of activity across voxels, or electrodes, or neurons. Each pattern of activity can be compared to the pattern of activity at another point in time; one can construct a scalar measure to characterize the similarity between states. The similarity can be aggregated as a function of behaviorally relevant variables and compared to predictions from mathematical models describing cognition (see, e.g., Kriegeskorte, Mur, & Bandettini, 2008). In Section 3, we review recent neuroscientific work that attempts to address empirical questions about the nature of memory representations raised by SST.

2. Dynamic memory representations in mathematical memory models

Prior to SST, many models of memory simply described the strength of direct atomic associations between stimuli and responses. Modern memory models construct a description of a memory representation that changes dynamically in response to stimuli. This representation can be quite abstract (as in the SIMPLE model Brown, Steyvers, & Hemmer, 2007) or considerably more concrete (as, for instance, in TODAM2 Murdock, 1997). In this section, we describe how the memory representations developed by various mathematical memory models evolve over experience with different stimuli and how these choices endow the models with power to explain various behavioral phenomena. Although these models are in all cases quantitatively implemented, we will not focus on their precise mathematical form, focusing instead on the qualitative changes in the memory representation caused by different kinds of experience. So, for instance, we will not focus on the difference between the context representation in the Mensink and Raaijmakers (1988) model of interference and the context representation in the Murdock (1997) TODAM2 model. Although these representations change over time according to different equations, they share the property that they change gradually over time and are independent of the stimuli presented.

Fig. 1 provides a schematic that enables us to illustrate three distinguishable types of temporal relationship. Let us denote the state of the memory representation when, say, stimulus A is presented as $\mathbf{s}_A$. First, we can consider how the state changes across different presentations of a particular nominal stimulus. Consider the three occurrences of D in Fig. 1. Indexing the three presentations by a subscript, we can ask whether these representations change gradually over time, or if they are independent of one another. That is, if the memory states are independent, then $\mathbf{s}_{D_1} \cdot \mathbf{s}_{D_2} = \mathbf{s}_{D_1} \cdot \mathbf{s}_{D_3}$. In contrast, if the memory representation after presentation of the nominal stimulus D changes gradually over time, then we would expect that $\mathbf{s}_{D_1} \cdot \mathbf{s}_{D_2} > \mathbf{s}_{D_1} \cdot \mathbf{s}_{D_3}$. We refer to the variable describing the number of presentations of the same stimulus as *stimulus recency* (Fig. 1). For instance, the stimulus recency

between D_3 and D_2 is -1, whereas the stimulus recency between D_3 and D_1 is -2. A decrease in the similarity of the memory representation as a function of stimulus recency is a key feature of SST (Estes, 1955a,b).

Second, we can ask what is the moment-to-moment change in the stimulus representation as different stimuli are presented. Rather than asking about different presentations of the same stimulus, we can compare, for instance, $\mathbf{s}_A$ to $\mathbf{s}_B$. We refer to the difference between serial positions of different stimulus presentations as *recency* (Fig. 1). Referring to the list described in Fig. 1, the recency of E to D_1 is -1, whereas the recency between E and C is -2. Note that it is logically possible that there could be an effect of stimulus recency on the similarity of the states in a memory representation but no effect of recency on memory states following the presentation of different stimuli. We will see that a decrease in the similarity of the memory representation as a function of recency can be used to account for behavioral recency effects is predicted by all three successors to SST that we consider.

There are many distinct mechanisms that predict an effect of stimulus recency on similarity of the memory representation. We need a third variable to distinguish among these mechanisms. Thus far, we have described two variables. Stimulus recency can be used to describe the similarity of the memory representation following successive presentations of the same nominal stimulus. Recency can be used to describe the similarity of the memory representation following successive presentations of different nominal stimuli. It remains to describe the relationship between a repetition of a nominal stimulus and the nominal stimuli surrounding its prior presentation. We define a third variable, *lag*, to describe the relationship between a repetition of a nominal stimulus and the neighbors of the previous presentation of that stimulus (Fig. 1). Referring to Fig. 1, the lag between D_3 and O is $+1$; the lag between D_3 and M is -2. We will see that predictions about the effect of lag on the similarity of the memory representation distinguish successors of SST from one another.

2.1. Stimulus sampling theory

In SST, a nominal stimulus evokes a set of conditioning elements each time the nominal stimulus is presented. The idea is that there is a large set of internal states that might follow presentation of, say, a 500 Hz tone. These internal states are composed of different combinations of "conditioning elements", which we can think of as analogous to neurons. Across different presentations of the tone, different states consisting of different combinations of stimulus elements are activated. Critically, like neurons obeying the Hebb rule, only the elements that are actually activated during a particular presentation of the stimulus can be conditioned to a response. In a testing situation, the nominal stimulus activates some set of conditioning elements—only the elements that are both activated at test and previously conditioned to the response can cause a behavioral response. As a consequence, the correlation structure among the different presentations of a nominal stimulus have a tremendous effect on the behavioral predictions that the model makes.

In the earliest versions of SST (exemplified by Estes, 1950), the set of elements available after each stimulus presentation was chosen independently of one another. That is, referring to the currently active set of conditioning elements after presentation of the ith presentation of nominal stimulus D as $\mathbf{s}_{D_i}$, the assumption is that $E\left[\mathbf{s}_{D_i}\cdot\mathbf{s}_{D_j}\right]$ is independent of the relationship between j and i for $j\neq i$. As the nominal stimulus is repeated, the probability that a randomly chosen set of elements is conditioned to the response increases, leading to a gradual (exponential) increase in the proportion of active elements that have been conditioned to the response. This gradual increase in the number of conditioned elements specifies a learning curve. Similarly, this simple model can account for the gradualness of extinction.

In later versions of SST, the memory representation following a nominal stimulus changed slowly across presentations of that stimulus. Estes (1955a,b) assumed that the population of active elements available to be conditioned fluctuated over time, with a probability that an active element would become inactive and a probability that an inactive element would become active. Because the active elements are no longer chosen independently, this means that $E\left[\mathbf{s}_{D_i}\cdot\mathbf{s}_{D_j}\right]$ is a decreasing function of stimulus recency $-|i-j|$.

The simple assumption of autocorrelation in the active stimulus elements leads directly to a number of remarkable behavioral predictions. First, it provides an account of forgetting over time (Estes, 1955b). Imagine that an association has been rapidly learned such that all of the active elements are conditioned to a response but the inactive elements are not. Over time, the conditioned elements will tend to fluctuate out of the active state, and unconditioned elements will tend to fluctuate into the active state. As a consequence, the more time that has elapsed since learning an association, the smaller the number of active conditioned elements, and thus the lower the probability of a behavioral response. Using similar reasoning, one can also account for spontaneous recovery. Suppose that after extensive learning the set of all active and inactive elements had been conditioned. Now, the active elements are unlearned by extinguishing the response. Immediately after extinguishing, all of the active elements are unconditioned; over time, inactive elements that are still conditioned fluctuate into the active state, causing a recovery of the response. The simple assumption of stimulus fluctuation also provides a natural account of the spacing effect (Estes, 1955a), the very general finding that memory at long time delays is better if repeated practice is spaced out in time (see, e.g., Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006).

The Estes (1950) formulation of SST differs from the Estes (1955a,b) formulation in the autocorrelation of the active stimulus elements as a function of stimulus recency. If the Estes (1950) formulation is correct, there would be no effect of stimulus recency, because the activated set is chosen independently each time the nominal stimulus is shown. In contrast, the Estes (1955a,b) version predicts decreasing similarity as a function of stimulus recency. The schematic pattern of results for stimulus recency for the Estes (1955a,b) variant is summarized in the leftmost column of Fig. 2 (top row).

Note that, while the Estes (1955a,b) variant of SST specifies the similarity of the memory representation when the nominal stimulus is presented, it does not specify how the representation changes between presentations of the nominal stimulus. On the one hand, we might assume that the representation changes gradually during that interval, which would lead to a decrease in similarity as a function of recency as well as of stimulus recency. This would mean that stimulus elements after presentation of D would remain active even long after the nominal stimulus presented. But this seems to contradict the observation that the response is presumably not made during the time between presentations of the nominal stimulus. SST as such is ambiguous as to the state of the memory representation during the times when the nominal stimulus is not presented.

This ambiguity in SST is reflected in Fig. 2. Because SST does not specify how the set of activated stimulus elements will change as a function of recency across presentations of different stimuli, we have left the middle column blank. Similarly, the similarity as a function of lag predicted by SST is ambiguous. While SST tells us that there will be a measure of similarity between the memory representation after the repeated stimulus and the representation after the original presentation of that stimulus, reflected in the

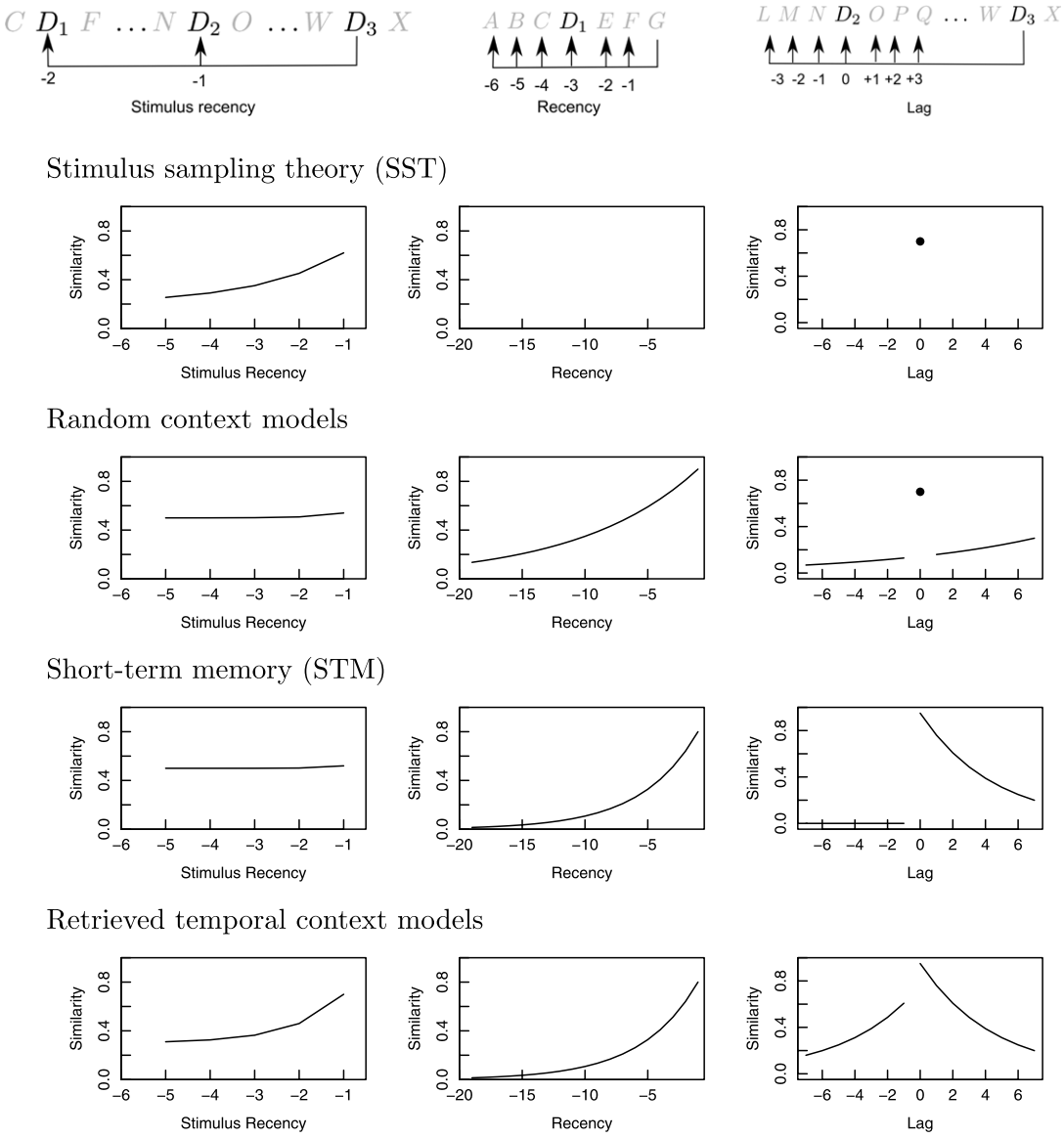

Fig. 2. Schematic of similarity relationships among memory representations induced by various models of memory. Each row shows a model described in the text. Each column shows a different relationship. If a panel is blank, this indicates that the model does not specify that relationship. The three columns show the similarity of the memory representation as a function of the three relationships depicted Fig. 1. The top row provides a schematic for the three relationships. See the text for explanations of the models and why they predict these relationships.

point at lag zero in the right column of Fig. 2, SST does not specify the pattern as a function of lag for non-zero values of lag.

2.2. Successors to SST

The subsequent decades saw numerous mathematical models of memory that can be seen as successors to SST. Like SST, these models depend critically on the properties of an internal representation that is distinct from the nominal stimuli and exploit gradual changes in that representation over time to describe canonical memory effects. We will focus on a subset of these models to highlight the impact of different choices to fill in the ambiguities left by SST with respect to the causes of autocorrelation in the memory representation. For clarity of exposition, we will go out of historical order. We will first discuss random context models (Mensink & Raaijmakers, 1988; Murdock, 1997; Sirotin, Kimball, & Kahana, 2005). Next, we will discuss models of short-term memory (STM, Atkinson & Shiffrin, 1968; Grossberg & Pearson, 2008; Raaijmakers & Shiffrin, 1980).

2.2.1. Random context models

Mensink and Raaijmakers (1988) introduced an extension of the search of associative memory (SAM, Raaijmakers & Shiffrin, 1980) model to account for interference effects in paired associate learning (see, e.g., Barnes & Underwood, 1959; Melton & Irwin, 1940; Postman & Underwood, 1973). Their model assumes that associations in paired associate learning are not simply formed by strengthening of direct item–item bonds, but are also mediated by a gradually changing context representation. The context representation followed precisely the equations governing the change in the active state described by Estes (1955b). By including the context representation, the model was able to account for a variety of findings in which the results of various memory tests after paired associate learning change with the amount of time after learning (see, e.g., Briggs, 1954; Koppenaal, 1978). Although the equations governing the context representation in Mensink and Raaijmakers (1988) are identical to those in SST, their conception makes several choices to disambiguate the gaps in SST. First, the context representation changes gradually over time independently of the

pairs that are presented. Second, conceptually the representation can change continuously across items within a list.[1]

Subsequent to the Mensink and Raaijmakers (1988) model, other authors constructed models of various memory tasks that maintained a categorical distinction between item and context representations (Murdock, 1997; Murdock, Smith, & Bai, 2001; Sirotin et al., 2005). In these models, the context representation did not change according to the same equations as Estes (1955b), but were conceptually quite similar to the Mensink and Raaijmakers (1988) formulation. In the Murdock (1997) model, the state of context at time step i evolved according to

$$\mathbf{c}_i = \rho \mathbf{c}_{i-1} + \sqrt{1-\rho^2}\boldsymbol{\eta}_i, \tag{1}$$

where $0 < \rho < 1$ and the vector of noise at each time step $\boldsymbol{\eta}_i$ is chosen independently. From Eq. (1), it is easy to see that the state of context is autocorrelated such that $E\left[\mathbf{c}_i \cdot \mathbf{c}_j\right]$ falls off with $|i-j|$.

In these "random context" models, the memory representation includes the context representation and the item representation. The item representation is assumed to be caused by the nominal stimulus that is currently presented, but otherwise does not change over time. Let us refer to the representation of the item presented at time step i as $\mathbf{f}_i$. In this case, we would expect $\mathbf{f}_i \cdot \mathbf{f}_j$ to be large when the nominal stimulus presented at time step i is the same as the nominal stimulus presented at time step j. The entire state of the memory representation in a random context model at time step i is just $\mathbf{f}_i \oplus \mathbf{c}_i$. The similarity of this representation at one time step with that at another time step is just the sum of the two similarities, $\mathbf{f}_i \cdot \mathbf{f}_j + \mathbf{c}_i \cdot \mathbf{c}_j$.

Fig. 2 summarizes the temporal relationships among the states of the memory representation composed of a fixed item representation and a randomly varying context representation as a function of the three variables we have described. In the left column we see that, unlike the Estes (1955b) variant of SST, the state $\mathbf{f}_i \oplus \mathbf{c}_i$ after presentation of a particular nominal stimulus does respond robustly to stimulus recency. The similarity of the item representation with itself is constant across presentations of the nominal stimulus. The similarity of the random context representation with itself changes gradually as a function of recency independently of the nominal stimuli presented. However, although recency is correlated with stimulus recency, if one chooses the time between repetitions of the nominal stimulus to be sufficiently large, then there could be a vanishingly small effect of stimulus recency on similarity. Unlike SST, random context models also specify that the representation should change gradually within a series of nominal stimuli (middle column). The memory representation after presentation of a nominal stimulus after presentation of a series of different nominal stimuli should fall off over time. This autocorrelation is due to the properties of random context (see Eq. (1)).

The rightmost column of Fig. 2 shows the similarity of the memory representation following repeated presentation of a nominal stimulus with neighbors of the original presentation of that stimulus. Here, the item representation and the random context representation contribute separately. First, the item representations following different presentations of a nominal stimulus D should be similar to one another. There is, however, no reason to expect that the item representation of D should be systematically related to the item representations of the neighbors of D (B, C, etc.) in a randomly assembled list. If the delay between the presentations of D is sufficiently large, there will be no detectable change due to the context representation either. However, in the right column, lag is confounded with recency. That is, E is closer to D_1 than F is to D_1. However, E is also closer to D_2 than F is to D_2. The increase in the function for non-zero lags from left to right is intended to communicate the possibility of a residual recency effect due to the changing context representation.

2.2.2. Short-term memory (STM)

STM models also make predictions about the way the memory representation – here the current contents of STM – changes over different experiences. STM is understood as a mechanism by which an item representation remains in an activated state after the nominal stimulus that caused it is no longer available. Typically, a small number of item representations can be simultaneously activated in STM. We can consider both discrete models of STM, in which a stimulus is either in STM or not (see, e.g., Atkinson & Shiffrin, 1968; Raaijmakers & Shiffrin, 1980), or continuous models, in which items gradually decay from STM (see, e.g., Grossberg & Pearson, 2008). Models of STM have been used to describe a wide range of temporal effects in immediate and delayed free recall (for a recent treatment, see Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005) and paired associate learning (see, e.g., Atkinson & Shiffrin, 1968).

Consider the set of activated stimuli in STM as a memory representation. The similarity of the representation at two times in a randomly assembled word list is a function of the degree to which the same items are active in STM at those two times. In particular, if STM contains item representations, then the input to STM caused by presentation of a nominal stimulus should be consistent from one presentation of that stimulus to another. If the number of other stimuli intervening between the two presentations of a nominal stimulus is large relative to the capacity of STM, then we would expect a flat function relating representational similarity to stimulus recency (Fig. 2). A decreasing function could result if the number of stimuli between presentations of a nominal stimulus is not large relative to the capacity of STM.[2] Like random context models, STM can account for an effect of stimulus recency only insofar as it is confounded with recency.

Like randomly changing context, the contents of STM also change gradually over time. Unlike in random context models, the change in STM depends critically on the identity of the nominal stimuli presented. For concreteness, let us consider the Atkinson and Shiffrin (1968) model with a buffer capacity of r items and a random drop-out rule. In this case, after capacity has been reached (i.e., after more than r items have been presented), the probability of any item present in the buffer falling out is just $\frac{1}{r}$. If each stimulus enters with certainty, then the probability that the buffer contains stimulus i at a later time step j is just $\left(1-\frac{1}{r}\right)^{j-i}$. Assuming that each item is presented only once and that the list is randomly assembled, this immediately gives rise to a recency effect, as the similarity of one state of STM to another is solely determined by the degree to which they contain the same item representations (middle column, Fig. 2). Like random context models, STM models also predict a representation that falls off with recency. Unlike in random context models, the gradually changing representation is sensitive to the nominal stimuli being presented, which leads to divergent predictions for lag.

As mentioned previously, for a random word list, temporal relationships in the contents of STM can only arise from the same stimulus occupying STM at two different times.[3] Let us consider the states of STM surrounding the initial presentations of a stimulus,

[1] In Mensink and Raaijmakers (1988), within-list changes were neglected under the assumption that those changes are small compared to the changes between phases of learning.

[2] In the Grossberg and Pearson (2008) model, more complex relationships are also possible.

[3] In behavioral models based on STM, such as SAM (Raaijmakers & Shiffrin, 1980; Sirotin et al., 2005), repetition of an item can cause a search and recovery process that results in a behavioral contiguity effect. If one allows for search and recovery to

$L\ M\ N\ D_2\ O\ P\ Q$, with the state after D being repeated for the second time, D_3, long after D_2. First note that the contents of STM following presentation of D_3 include a representation of D and whatever stimuli were present prior to that presentation. In the example from Fig. 1, there would be high probabilities that stimuli U, V, and W are active in STM when D_3 is presented. None of those stimuli should be in STM prior to presentation of D_2, so they cannot contribute to the similarity of the states of STM. More precisely, if the number of intervening stimuli between D_2 and D_3 is large with respect to the buffer capacity, then the similarity of the contents of STM after presentation of D_3 to the contents of STM prior to the presentation of D_2 should be flat as a function of negative values of lag. Now, when D is presented at D_2, D enters STM. Because D has also entered STM on D_3, we would expect the similarity between the states of STM at D_2 and D_3 to shoot up relative to the states at L, M, and N. As additional stimuli are presented after D_2, the stimulus D has the opportunity to fall out of STM. As a consequence, on average, the similarity of STM following D_3 decreases gradually with increasing positive values of lag (Fig. 2, right). Note that this asymmetric similarity function with respect to lag is qualitatively distinct from the pattern that results from random context models.

2.2.3. Retrieved temporal context models

The state of STM changes gradually over multiple serial positions as stimuli enter, persist in, and then leave STM. The changes in STM over time are completely dependent on the identity of the stimuli presented. The random context also changes gradually, but the changes in the context representation are completely independent of the nominal stimuli presented. Retrieved temporal context models (Howard & Kahana, 2002; Polyn, Norman, & Kahana, 2009; Sederberg, Howard, & Kahana, 2008) offer an alternative that is something of a hybrid of these two approaches. Like STM, retrieved context models hypothesize a gradually changing memory representation – referred to as the temporal context – that is caused by the nominal stimuli presented. Unlike STM, however, the input to temporal context caused by a nominal stimulus can change across presentations of that stimulus. In particular, repeated stimuli can recover previous states of temporal context in which they were experienced.

It will turn out that the predictions of retrieved temporal context models for stimulus recency are somewhat involved. For that reason, we will address stimulus recency last. Each time a nominal stimulus is presented, it provides some input to the current state of temporal context. The state changes gradually from presentation of one stimulus to the next, with information persisting for some time. Let us refer to the state of temporal context at time step i as $\mathbf{t}_i$. The nominal stimulus presented at time step i causes a particular input pattern $\mathbf{t}_i^{IN}$. Then, the state of temporal context changes from one list position to the next according to

$$\mathbf{t}_i = \rho_i \mathbf{t}_{i-1} + \beta \mathbf{t}_i^{IN}, \tag{2}$$

where $0 < \beta < 1$. In many formulations, ρ_i is chosen such that the Euclidean length of $\mathbf{t}_i$ is constant at all time steps and is typically less than 1. From Eq. (2), we can readily see that the similarity of $\mathbf{t}_i$ to $\mathbf{t}_j$ falls off with recency $|i-j|$ (Fig. 2, middle). This is the same qualitative pattern we would expect for random context models and STM.

The nature of $\mathbf{t}^{IN}$, though, has a big effect on the behavioral predictions of the model and results in a qualitative difference between the similarity of the memory representations of a retrieved temporal context model from the predictions of either random context models or models based on STM. When a nominal stimulus is repeated, the input to the temporal context vector $\mathbf{t}^{IN}$ caused by that stimulus is not the same as that for the first time that nominal stimulus was presented. The *input* that enters into the memory representation changes, including the state of temporal context prior to the original presentation of the stimulus. In particular, when a nominal stimulus is repeated, it can cause a "jump back in time" in which a prior state of context is recovered. To make this more concrete, let us suppose that the stimulus presented at time step i is later repeated at time step r. Then,

$$\mathbf{t}_r^{IN} = \gamma \mathbf{t}_{i-1} + (1-\gamma)\, \mathbf{t}_i^{IN}. \tag{3}$$

First, note that, if the delay between i and r is sufficiently long, then the contribution of $\rho_r \mathbf{t}_{r-1} \cdot \mathbf{t}_{i+j}$ for small j will be negligible compared to the contribution of $\beta \mathbf{t}_r^{IN} \cdot \mathbf{t}_{i+j}$.[4] Let us consider the contribution of each of these two terms to the similarity of $\mathbf{t}_r$ to the neighbors of $\mathbf{t}_i$. First, let us consider the similarity arising from $\mathbf{t}_i^{IN} \cdot \mathbf{t}_{i+j}$. We can see that for a random word list around i this is maximal for $j = 0$. This item-specific contribution falls off as j increases from zero. $\mathbf{t}_i^{IN}$ would not tend to resemble the states of context *prior* to time step i. Taken in isolation, the $\mathbf{t}_i^{IN}$ component of Eq. (3) has the same qualitative pattern as a function of lag as we would expect for STM, for analogous reasons.

The other component in Eq. (3), $\mathbf{t}_{i-1}$, reflects the recovery of the previous state of temporal context available when the stimulus was originally presented. This "jump back in time" results in a component that overlaps with states of temporal context both forward and backward in the sequence from serial position i. This can be seen by noting that $\mathbf{t}_i \cdot \mathbf{t}_j = \mathbf{t}_j \cdot \mathbf{t}_i$. The expectation of $\mathbf{t}_{i-1} \cdot \mathbf{t}_{i+j}$ is maximal at $j = -1$ and symmetric around this maximum. Combining these two components, the similarity as a function of lag will on average be asymmetric, falling off in both the forward direction and the backward direction (Fig. 2, right).

We are now in a position to address the predictions of stimulus recency in retrieved temporal context models. There are two reasons that retrieved temporal context models predict an effect of stimulus recency. First, if ρ is sufficiently close to 1, we may expect the effect of recency to also be manifest as a function of stimulus recency (Fig. 2, left) for just the same reasons as random context models can exhibit a decreasing function of stimulus recency. Above and beyond this effect, note that Eq. (3) describes a change in the input caused by a nominal stimulus across multiple presentations. Consider what would happen if the stimulus is again repeated. There is a strong analogy to Eq. (2). There, the temporal context changes from one moment to the next driven by the input from the current nominal stimulus. In Eq. (3), the input caused by a nominal stimulus changes gradually from one presentation of that stimulus to the next, driven by the temporal context in which it was presented. Assuming for clarity that the contexts are uncorrelated because the repetitions of the nominal stimulus are widely separated in a list with no other repetitions, this process gives rise to something like an autoregressive process. Note, however, that this autoregressive process is distinctly different from the autoregressive process in Eq. (2) in two ways. First, the process in Eq. (3) is over presentations of the nominal stimulus itself, and is thus relatively resistant to large gaps between presentations of the nominal stimulus. In this regard, the stimulus recency effect predicted by retrieved temporal context models differs from stimulus recency in SST, which was assumed to depend only on the spacing between presentations of the nominal stimulus. Second, the rate at which

take place, and if that retrieval process exhibits a behavioral contiguity effect, and if the results of the retrieval process are fed back into the contents of STM during study, then the contents of STM could exhibit a more subtle contiguity effect, like that seen for retrieved context models. In this case, however, STM can no longer be seen as a simple container that holds a set of recently presented items.

[4] To the extent that it is not negligible, the tendency is for a recency gradient like that obtained for the random context model in the right column of Fig. 2.

the input pattern changes – the rate constant in Eq. (3), if you will – is largely decoupled from the rate constant in Eq. (2).

2.3. Lag and stimulus recency analyses discriminate among models of memory

SST (Estes, 1950) introduced the idea of a memory representation separate from the nominal stimulus. Estes (1955a,b) demonstrated the crucial importance of allowing the memory representation to change gradually over time. All three of the classes of successors to SST we have examined share the property that the memory representation should change gradually as successive stimuli are presented, decreasing in similarity as a function of recency (Fig. 2, middle). However, these three classes of models make very different assumptions about the causes of the change in their gradually changing representations. These different causes result in qualitatively different predictions about how the memory representation should change as a function of lag (Fig. 2, right) and stimulus recency (Fig. 2, left).

The lag analysis clearly distinguishes the models from one another. Random context models, because the context representation is independent of the stimuli presented, predict that there should be no systematic effect of lag above and beyond any residual confound with recency. Because the gradually changing representation depends on the identity of the nominal stimulus presented, both STM and retrieved temporal context models predict systematic and robust lag effects. However, there is also a qualitative property to the shape of the lag curve that is unique to retrieved temporal context models. In particular, the recovery of a prior state of temporal context by a repeated stimulus causes a curve that falls off in both the forward direction and the backward direction around zero. Finding a decrease in the backward direction above and beyond that attributable to recency is a unique prediction of this class of models.

The Estes (1955a,b) version of SST assumed that the conditioning elements available when a CS was presented change gradually across multiple presentations of that nominal stimulus. The successors of SST are also distinguished by the degree to which they incorporate this property. Random context models and STM can both describe stimulus recency only to the extent that it is confounded with recency. Suppose that there is a large gap between repetitions of a nominal stimulus. Then, for both random context models and STM, the rate at which the representation changes as a function of recency also determines the rate at which the representation changes as a function of stimulus recency. For a fixed rate of change as a function of recency, as the time between repetitions of a nominal stimulus grows larger, the effect of stimulus recency becomes vanishingly small. For instance, as long as the contents of STM from one presentation of a nominal stimulus have been emptied before the next presentation of that stimulus, there is no reason to expect any effect of stimulus recency on the memory representation. If one needed to describe an effect of stimulus recency on the representation using STM, one could assume that STM has a much larger capacity. Then, however, recency would also be affected. Retrieved temporal context models have a robust mechanism to describe a stimulus recency effect that is only weakly sensitive to the spacing between repetitions of a nominal stimulus and is decoupled from the effect of recency.

Moreover, retrieved temporal context models make a specific hypothesis about the source of the variability in the input caused by repetition of a nominal stimulus, which leads to a further prediction. In retrieved temporal context models, the input caused by a nominal stimulus does not change randomly, as in SST, but rather in response to the particular temporal contexts in which it is experienced (Eq. (3)). Stimuli that are presented in similar temporal contexts will tend to develop similar input patterns. This prediction forms a point of contact with computational models of semantic learning, which estimate the meaning of words by observing their cooccurrence statistics in natural text; see Dennis (2005), Griffiths, Steyvers, and Tenenbaum (2007), Howard, Shankar, and Jagadisan (2011), Jones and Mewhort (2007), Landauer and Dumais (1997), and Shankar, Jagadisan, and Howard (2009).

3. Neural similarity analyses

We mentioned at the outset that the research program of mathematical learning theory – to write down a set of precise equations that described mechanisms detailed enough to account for behavior – was audacious, given what was known about neurobiology in the 1950s. Constructing a correct process model of memory unconstrained by neurobiology is an almost impossible task. While it is still not an easy task to compare models to neurobiological data, recent decades have seen tremendous progress in our ability to measure brain processes in humans and in animal models. Human neuroimaging studies can resolve hemodynamic response on the scale of a few millimeters. Intracranial recordings from human epileptic patients combine spatial resolution of local field potentials on the order of about a centimeter with temporal resolution on the scale of milliseconds. Rodent electrophysiology labs routinely record extracellularly from a few hundred neurons simultaneously in parts of the brain believed to be important in learning and memory.

All of these technologies generate multivariate responses. To the extent that we have a hypothesis about how the memory representation changes across conditions, as in Fig. 2, we can evaluate whether these similarity relationships are respected in multivariate brain responses. In the case of fMRI, we can compare the distributed activity across voxels at one time to the pattern of activity across voxels at another time (Norman, Polyn, Detre, & Haxby, 2006; Polyn, Natu, Cohen, & Norman, 2005). In the case of intracranial recordings, we can construct vectors of principal components that vary over time (Manning, Polyn, Litt, Baltuch, & Kahana, 2011). When we have many neurons, we can compare population vectors to one another. One can construct a population vector over a certain interval by estimating the firing rate for each neuron. If one has recorded from n neurons, one can construct an n-dimensional population vector. The similarity of these vectors to one another can be compared using standard methods (e.g., inner product, standard correlation methods, Euclidean distance, Mahalanobis distance, etc.).

Recent years have seen the development of three themes of research that place constraints on memory representations across time and, thus, place constraints on mathematical process models of memory. One theme is the finding that memory representations change gradually over long periods of time, ranging up to at least thousands of seconds. A second theme is recent evidence suggesting that, in episodic memory tasks, remembered items can cause the signature of a "jump back in time". Finally, nominal stimuli that are experienced close together in time develop similar stimulus representations.

3.1. Memory representations change over long periods of time

Many models have suggested that the change in the memory representation across presentations of different nominal stimuli (Fig. 2, middle) could account for the behavioral recency effect. The recency effect refers to the finding that recently presented items from the end of a list are better remembered than less recent items from the middle of the list. If this is the case, then the time scale over which behavioral recency effects are observed should be reflected in neural recency effects. Although some authors have assumed that the recency effect extends only over short

time scales associated with the capacity of STM, behavioral evidence shows that in free recall the recency effect is observed over time scales ranging from a hundred milliseconds (Neath & Crowder, 1996), to seconds (see, e.g., Murdock, 1962), to tens of seconds (see, e.g., Bjork & Whitten, 1974; Glenberg et al., 1980; Howard & Kahana, 1999), to hundreds of seconds (Glenberg et al., 1980; Howard, Youker, & Venkatadass, 2008), and perhaps even to days and weeks (Baddeley & Hitch, 1977; da Costa Pinto & Baddeley, 1991; Moreton & Ward, 2010). If changes in memory representations are responsible for these recency effects, they should also change over similarly long time scales. We briefly review three studies that suggest that neural ensembles change gradually over perhaps hours.

Manns, Howard, and Eichenbaum (2007) presented rats with lists of odors for a judgment of recency task. They measured population vectors across simultaneously recorded neurons from the hippocampus for a four-second period around each odor sampling event. Restricting their attention to events within the same list, they found that population vectors from different events changed reliably as a function of the recency relating the two events. More surprisingly, they also found that population vectors changed gradually across lists. That is, the population vectors from two events from different lists became reliably less similar to one another as the number of intervening lists of odors increased. Lists were separated by about a minute, indicating that there were reliable changes continuing even after several hundred seconds.

Similar results have been found in two recent studies conducted while rats foraged in an open environment. It is well known that "place cells" in the rat hippocampus discriminate location within a small environment such as a one-meter-square enclosure (see, e.g., Wilson & McNaughton, 1993). That is, one place cell might fire when the animal's physical location is in one circumscribed part of the environment relatively independently of other time-varying variables. Another place cell would fire only when the animal is in some other circumscribed part of the environment. An ensemble of such place cells would have the ability to reconstruct the animal's location within the environment. Hyman, Ma, Balaguer-Ballester, Durstewitz, and Seamans (2012) recorded from ensembles of neurons from the rodent hippocampus and medial prefrontal cortex while rats randomly foraged in one of two environments. As the rats foraged randomly, they would revisit the same location at different times. The authors constructed population vectors for these different visits to the same location and measured similarity as a function of the time between the visits. Population vectors in both regions changed gradually over several hundreds of seconds.

Similarly, Mankin et al. (2012) recorded place cells from the rodent hippocampus while the animal explored an enclosure at various times throughout the day for consecutive days. They found that the population vectors at a particular location within the environment continued to change reliably after several hours had passed. Critically, ensemble similarity remained at a low level at a delay of 24 h, ruling out an account based on time of day.

A memory representation that changes gradually over time is a key feature that can be used to account for the behavioral recency effect. While there is undoubtedly much work that would have to be done to clarify the relationship between the behavioral recency effect and these neural ensembles, it is clear that neurons in the hippocampus change their firing gradually over periods of time sufficiently long to account for the recency effect.

3.2. Memory representations may jump back in time

The preceding discussion reveals very strong evidence for gradual changes in neural representations over time scales of at least a few hours. This effect of recency does not discriminate among models specifying the memory representation (see the middle column of Fig. 2). Comparing the memory representation after a repeated item to the neighbors of its original presentation provides a way to distinguish different models specifying different causal relationships between nominal stimuli and the changes in the memory representation (Fig. 2, right column). We review recent studies that have a bearing on the question of whether the brain manages to "jump back in time".

Two recent studies are consistent with a neural contiguity effect. First, Manning et al. (2011) recorded intracranially from epileptic patients who performed a free-recall task. They observed a reliable effect of recency on the pattern of oscillatory components present at the electrodes. Comparing the interval just before a word was free-recalled with the interval surrounding the words' presentation they found a reliable neural contiguity effect that correlated with the behavioral contiguity effect. While suggestive of a jump back in time, because the sequence of recalls was generated by the subject, it is possible that their neural contiguity effect reflected the fact that the words recalled prior to a word also came from near the about-to-be-recalled word in the list, rather than a jump back in time *per se*. Second, Zeithamova, Dominick, and Preston (2012) examined multivoxel patterns of activity in an fMRI experiment while subjects learned pairs of images. The images were chosen from categories that can be decoded using fMRI activation. Subjects studied pairs A–B and later studied B–C. Critically, during study of the B–C pair, they found that patterns in the hippocampus partially reconstructed the category of A, suggesting that this prior episode, including the content of the stimuli, was recovered. Because the pairs were presented simultaneously, this study leaves open the possibility that the brain would not reconstruct the temporal context that preceded a repeated item (see Gershman, Schapiro, Hupbach, & Norman, 2013, for a related fMRI study).

An additional study addresses some of these limitations. Subjects in the Howard, Viskontas, Shankar, and Fried (2012) study were epileptic patients who studied lists of pictures in a continuous recognition study. In each block, the pictures in the list were presented twice. The first time the picture was presented, the subject should respond "new"; the second time the picture is presented, the subject should respond "old". While the patients performed the task, extracellular recordings were taken from a variety of locations in the MTL, isolating clusters of spikes from both single neurons and clusters of a few neurons. Howard et al. (2012) measured a population vector across these clusters averaged over the 3 s following presentation of each stimulus. As we would expect based on the foregoing evidence for gradually changing representations, Howard et al. (2012) observed that the neural ensemble changed gradually both within and across blocks of stimuli (Fig. 3a). Within block, the effect of recency persisted over a few dozen seconds; across block, the effect of recency persisted for a few minutes.

Fig. 3b shows the results of a neural contiguity analysis on the same data set, giving neural similarity as a function of lag. The results were consistent with a "jump back in time", suggesting that the repeated item caused recovery of gradually changing information available before presentation of the repeated item. However, several limitations of this study preclude a definitive conclusion. First, the design of the continuous recognition study meant that repetition D_2 was not always at a long delay relative to the initial presentation D_1. Rather, repetitions came at a variety of recencies such that the recency effect between the repeated stimulus and its predecessors had to be statistically removed. This additional level of analysis complicates the interpretation. A study-test design in which a long delay intervened between study of the list items and the test probes would have avoided this potential problem. Second, identifying individual spikes from extracellular recordings with humans is much more error prone than from extracellular recordings from animal preparations, leading to a

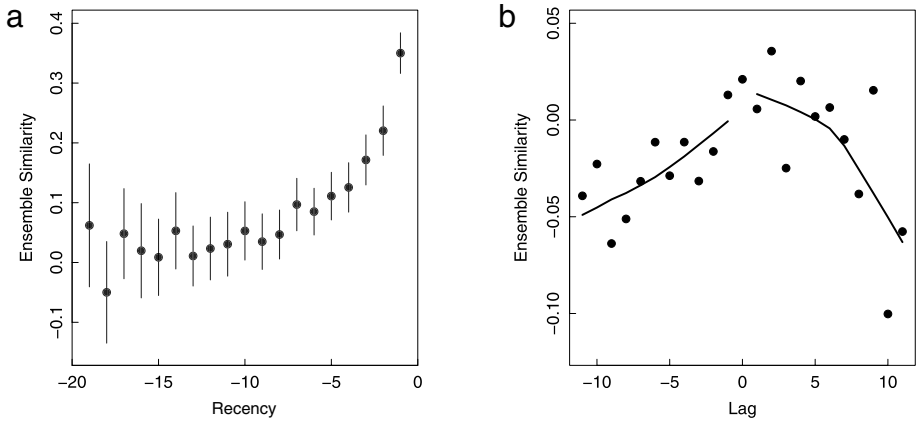

Fig. 3. Recency and contiguity in the brain. Neural recency and contiguity effects in human MTL neurons. Single units were recorded from epileptic patients while they performed a continuous recognition task. (a) The inner product between the normalized population vector at each time and at each preceding time was compared as a function of the number of stimuli between them. Each stimulus presentation takes at least 3 s, so the population changes over at least a few dozen seconds. (b) The recency-controlled comparison between the population vector caused by a repeated item and the neighbors of its original presentation. After Howard et al. (2012).

$$A\ B\ C\ D_1 E_1\ F\ \ldots\ L\ M\ N\ E_2 D_2\ O\ \ldots\ U\ V\ W\ D_3 E_3\ X$$

Fig. 4. Nominal stimuli repeated in similar temporal contexts. Here, stimuli *D* and *E* tend to cooccur with one another across temporal contexts. According to retrieved temporal context models, they will develop similar representations.

potential source of noise. Third, there was relatively little neural data available, and behavioral performance was good, so it was not possible to establish a correlation between the neural contiguity effect and either a behavioral contiguity effect (Schwartz, Howard, Jing, & Kahana, 2005) or even successful memory performance.

While none of these three studies in isolation conclusively demonstrates that a neural jump back in time supports episodic retrieval, their limitations are complementary. While no behavioral contiguity effect was observed in the Howard et al. (2012) study, the Manning et al. (2011) study was able to measure a behavioral contiguity effect and to demonstrate that it was correlated with the neural contiguity effect they observed. While one might be concerned about the statistical procedure to remove the recency effect in the Howard et al. (2012) study, no such artifact was present in the Manning et al. (2011) study, because it used delayed recall. While there is a concern about correlation with previous recalled items in the Manning et al. (2011) study, because the probes were chosen by the experimenter in the Howard et al. (2012) study, that concern does not apply there. If one is willing to make the leap that these studies reflect different neural signatures of the same phenomenon, then taken together they provide a strong case that memory depends on the maintenance and recovery of a gradually changing state of temporal context. It remains to be seen if this neural contiguity effect extends over longer time scales over which the behavioral contiguity effect is manifest (see, e.g., Howard et al., 2008).

3.3. Stimulus representations reflect the temporal context in which a nominal stimulus was experienced

In retrieved temporal context models, the input caused by a nominal stimulus changes across repetitions of that nominal stimulus. More than just random fluctuations, the changes come to reflect the temporal contexts in which the nominal stimulus is experienced. This can result in the development of functional stimulus representations that reflect the structure of learning.

A concrete example may help to illustrate this point. Consider the effect of learning on two nominal stimuli, say *D* and *E*, that initially have uncorrelated input patterns and are then presented close together in time repeatedly (Fig. 4). Initially, the input patterns $\mathbf{t}^{IN}_{D_1}$ and $\mathbf{t}^{IN}_{E_1}$ can only recover contexts from prior to learning. As a consequence, $\mathbf{t}^{IN}_{D_1} \cdot \mathbf{t}^{IN}_{E_1}$ should be no higher on average than a randomly chosen pair of stimuli. Now, after learning, $\mathbf{t}^{IN}_{D_2}$ recovers $\mathbf{t}_{D_1-1}$ and $\mathbf{t}^{IN}_{E_2}$ recovers $\mathbf{t}_{E_1-1}$. Because these contextual states are close to one another, this means that $\mathbf{t}^{IN}_{D_2} \cdot \mathbf{t}^{IN}_{E_2}$ will be much higher than the input caused by these two nominal stimuli before learning. Repeated presentations of the two stimuli close together in time cause the input patterns of the two stimuli to become more and more similar to one another. There are some subtleties to this argument that depend on the details of the learning rules employed (Howard, Jing, Rao, Provyn, & Datey, 2009; Rao & Howard, 2008; Shankar et al., 2009), but this simple illustration is sufficient for present purposes.

There is ample neural evidence that the representations caused by nominal stimuli that are experienced close together in time become similar to one another. For instance, Miyashita (1988) recorded from neurons in the inferotemporal (IT) cortex of monkeys. The stimuli in his experiment were complex visual patterns assembled randomly into a list that was repeatedly presented in a fixed order for many days of training. Neurons did not respond to any identifiable visual property of the stimuli (Miyashita & Chang, 1988). However, after training, neurons that fired persistently in response to one nominal stimulus from the list also tended to respond to other stimuli that were close in position within the list (Miyashita, 1988). This effect was notably absent for a new set of stimuli that were not trained in order. Later work showed a number of additional properties of these neurons, including that this property depends on connections from the medial temporal lobe (Naya, Yoshida, & Miyashita, 2001). The medial temporal lobe includes the hippocampus, which, as discussed earlier, contains a representation that changes gradually over long periods of time.

The connection between the hippocampus and this effect of learning temporal contexts has been made explicit by recent neuroimaging work in humans. Schapiro, Kustner, and Turk-Browne (2012) presented subjects with a series of complex images in a statistical learning paradigm. Pairs of stimuli that predicted each other with high probability were embedded in a sequence of random images. Using similarity of multivoxel patterns of activation, they found that the entire medial temporal lobe, including the hippocampus, developed representations such

that nominal stimuli that were repeatedly paired caused more similar neural representations than nominal stimuli that were not reliably paired (see also Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013). Neural firing comes to reflect temporal contiguity over shorter time scales as well (see Li and DiCarlo (2008) and DiCarlo, Zoccolan, and Rust (2012)), suggesting that temporal contiguity may be a powerful cue in parsing the visual world (see, e.g., Becker, 1999; Franzius, Wilbert, & Wiskott, 2011; Wiskott & Sejnowski, 2002).

4. Discussion

Mathematical psychology has several potential functions. One function is to develop measurement models that provide a way to summarize behavioral data in a concise and insightful way. Another function is to demonstrate that abstract heuristics, such as Bayesian inference, could support human behavior. A third function is to provide concrete mechanistic hypotheses about the computations that could support behavior. Of course, these hypotheses should generate testable predictions about behavior, but they can also, in principle at least, be tested literally using neurobiological findings.[5] SST was formulated as a concrete mechanistic hypothesis explaining the development and change in memories over time. In the ensuing decades, several distinct hypotheses have been developed that build on these ideas. It is troubling that so much time has passed with essentially contradictory models persisting in parallel. One reason for this is that the behavioral constraints from a single task, such as free recall, or paired associate learning, or item recognition, are ill suited to constrain a general model of how memory representations are updated and maintained over time. In addition, it is quite possible that behavioral results are simply insufficient to constrain a satisfactory mechanistic hypothesis. Neurobiological data can provide an additional set of constraints that speak directly to the mechanistic nature of mathematical models of memory; mathematical psychology would benefit tremendously if it can incorporate these insights into future model development.

4.1. Can we construct a physical theory of memory?

Stimulus sampling theory was an attempt to write down a set of equations that would provide a veridical description of the cognitive subprocesses underlying memory. While much work still remains to be done, we have learned a great deal about the ways in which ensembles of neurons in the brain, and in particular the MTL, change over repeated presentations of a nominal stimulus. For the moment let us make the leap that neural ensembles in the medial temporal lobe are the memory representations central to human behavioral performance in standard memory experiments. Let us further ignore the caveats and limitations of the foregoing neuroscientific studies, and consider the constraints these findings would have on models of human behavioral memory performance. With these assumptions, the neuroscientific findings described above imply that we must construct a model subject to the following constraints.

1. Memory representations change gradually over periods of time much longer than a single list.
2. Persistent changes in memory representations can be caused by nominal stimuli.
3. Nominal stimuli can cause previous states of the memory representation to be recovered.
4. Stimulus representations for stimuli presented close together in time come to resemble one another.

All of the successors we have considered here are consistent with Constraint 1. Constraints 2–4 argue strongly against random context models as the sole source of temporal variability in neural representations. Moreover, constraints 3–4 are inconsistent with a conception of STM as a container of unchanging item representations. To the extent that the empirical neural data described in constraints 1–4 is correct, it is unclear how work on behavioral models of memory that violate one or more of those constraints could lead to a satisfactory mechanistic model of memory in the brain.

While all four constraints are at least roughly consistent with the predictions of retrieved temporal context models, this does not in any way imply that those models are correct or constitute a satisfactory mathematical description of the memory representation underlying performance in any particular memory task. These neural constraints are much stronger than the constraints available in the 1950s, but there is still a huge space of models consistent with those constraints. It is almost inconceivable that the retrieved temporal context models published thus far are satisfactory in any meaningful sense. These constraints should serve to focus attention on specific directions that are likely to yield progress towards a satisfactory physical theory of memory.

Any process that enables recovery of a memory representation is consistent with constraint 3; constraint 3 raises a number of fundamental psychological questions. For instance, is the "jump back in time" discrete or continuous? Does the retrieval event have characteristic retrieval dynamics? What factors (at both encoding and retrieval) cause this jump back in time to succeed or fail? Even granting that a "jump back in time" is a general principle of memory, we thus far do not even have strong hypotheses that address these basic questions.

Constraint 4 is consistent with a wide variety of learning rules; the choice of rule should be constrained by broad considerations and would benefit from behavioral empirical data collected under controlled conditions. There is good reason to suspect that the development of neural representations that reflect temporal statistics is an essential operation in semantic learning and learning of perceptual representations more generally, so it is extremely pressing to connect models of memory to learning over scales longer than a few minutes (Nelson & Shiffrin, 2013).

Any set of equations that describes a vector state that changes gradually over time and that is causally affected by the nominal stimuli presented ought to be consistent with constraints 1 and 2. As such, constraints 1 and 2 together create a relatively weak physical constraint; additional theoretical concerns are necessary to move forward. Why does the memory representation in the MTL change gradually over long periods of time? One particularly intriguing possibility with potentially far-reaching implications is that the memory representation codes for the history of stimuli leading up to the present moment in a scale-invariant manner. After presentation of *ABC*, the memory representation includes information about *C* being one time step in the past, *B* two steps in the past, and *A* three steps in the past. After presentation of another stimulus, the representations of *A*, *B*, and *C* are each pushed back in time, and the new stimulus is placed in the slot corresponding to one time step in the past. In this way, the memory representation would maintain information about which nominal stimuli were presented while preserving information about their time of presentation.

Cognitive neuroscientists studying the hippocampus, a region within the MTL, have long suggested that it plays a special role in encoding sequences of stimuli, retaining information about what stimulus was presented and the order of their presentation

[5] Of course, models of behavior may serve all three functions. For instance, in the context of drift diffusion models (Ratcliff, 1978), drift rate and bias are useful summaries of behavior, and drift diffusion in two-choice decisions can be understood as computation of log-likelihood using sequential sampling. Neurobiologically, one can directly evaluate the hypothesis that neurons obey a random walk towards a threshold (Smith & Ratcliff, 2004).

(see, e.g., Eichenbaum, 1999; Hasselmo, 2009, 2012; Jensen & Lisman, 1996; Levy, 1996; Lisman, 1999; Sohal & Hasselmo, 1998). MacDonald, Lepage, Eden, and Eichenbaum (2011), recording from the hippocampus of rodents during the delay of a memory task, found neurons that fired at a particular time during the delay. These *time cells* contained information about how long in the past the beginning of the delay period was. Critically, a subset of time cells was selective for the identity of a sample stimulus that was encoded at the beginning of the delay, meaning that a population of such neurons would carry information about what nominal stimulus was presented how far in the past (see also Gill, Mizumori, & Smith, 2011; Kraus, Robinson, White, Eichenbaum, & Hasselmo, 2013; MacDonald, Carrow, Place, & Eichenbaum, 2013; Pastalkova, Itskov, Amarasingham, & Buzsaki, 2008).

It is interesting to consider how a representation of the sequences of events leading up to the present moment could cause the representation to change gradually over time. If separate neurons code precisely for the conjunction of time step and nominal stimulus, this memory representation would not exhibit autocorrelation. That is, if a neuron fires robustly when B was presented two time steps in the past but does not fire at all when B was presented three time steps in the past, this neuron could not be a source of correlated firing between those two occasions. In contrast, if the sequence coding were blurry, such that some neurons participate in coding for multiple time steps in the past, the representation as a whole would show an overlap across those scales. For instance, if a neuron fired when B was two steps in the past and also fired when B was three steps in the past, this would lead to an overlap in the firing pattern at those two times. Now suppose that the accuracy with which the time of presentation is recorded blurs out in a scale-invariant manner. That is, the time at which a stimulus presented 10 s in the past is stored with the same relative accuracy as that for a stimulus presented 100 s in the past, for instance 10 ± 1 s and 100 ± 10 s. A scale-invariant blur in the accuracy would make sense of the observation that neural representations change over hundreds or even thousands of seconds.

Shankar and Howard (2012) presented a mathematical model that computes a fuzzy representation of sequences based on an approximation to the inverse Laplace transform. Because the blur at each point in the history is proportional to the time in the past being represented, the model gives rise to scale-invariant autocorrelation. For instance, if a neuron coding for a stimulus 2 s in the past also fired when the stimulus was 3 s in the past, then the neuron coding for that stimulus 20 s in the past would also fire when the stimulus was 30 s in the past. A range of modeling work has shown that this representation can account for a variety of behavioral findings from episodic memory, conditioning, and working memory (Howard, Shankar, Aue, & Criss, 0000; Shankar & Howard, 2012). This form for a memory representation that changes gradually over time could be a promising starting point for the development of a satisfactory mechanistic theory of memory, an effort that continues the thread of work initiated by SST more than six decades ago.

5. Conclusion

The lasting effect of stimulus sampling theory can be seen in work being done today. The agenda set by SST – to describe the properties of a memory representation that intervenes between the nominal stimulus and the response – has driven most of the memory modeling work done in the subsequent decades. In particular, SST's conjecture about the importance of gradually changing memory representations has been particularly influential on models of memory. Recent neuroscientific findings have shown strong evidence for a gradual change in neural ensembles extending at least hundreds of seconds. Many subsequent memory models, including random context models and short-term memory models, can be seen as hypotheses about the source of the gradual change assumed by SST. Random context models assume that the change is caused by noise that is independent of the presented stimuli. This hypothesis is inconsistent with the neural data reviewed here. Short-term memory is an improvement in that changes in the representation are caused by the nominal stimuli presented. However, STM fails to account for the finding that the set of neurons caused by a nominal stimulus change to reflect the temporal structure of experience. Like STM, retrieved temporal context models postulate a gradually changing memory representation caused by the nominal stimuli presented. In addition, retrieved context models also propose that repetition of a stimulus can cause a "jump back in time", whereby a previous state of temporal context is recovered; stimuli presented in similar temporal contexts should thus cause similar patterns of activation after learning. While not yet definitive, some recent neuroscientific work is supportive of a neural jump back in time in the MTL.

Due to technological advances in neurobiology, the overarching goal of SST, namely the construction of a set of equations that are a veridical description of actual physical processes underlying behavioral memory performance, while still very far off, is nonetheless much more realistic now than it was in the 1950s. Successful work towards this goal will require that mathematical models incorporate neuroscientific evidence into the empirical data that constrain them.

Acknowledgments

The author gratefully acknowledges support from AFOSR award FA9550-10-1-0149, and NSF award BCS-1058937.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: a proposed system and its control processes. In K. W. Spence, & J. T. Spence (Eds.), *The psychology of learning and motivation*: *Vol. 2* (pp. 89–105). New York: Academic Press.

Baddeley, A. D., & Hitch, G. J. (1977). Recency reexamined. In S. Dornic (Ed.), *Attention and performance*: *Vol. VI* (pp. 647–667). Hillsdale, NJ: Erlbaum.

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

Becker, S. (1999). Implicit learning in 3d object recognition: the importance of temporal context. *Neural Computation*, *11*, 347–374.

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, *6*, 173–189.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, *49*, 229–240.

Briggs, G. E. (1954). Acquisition, extinction, and recovery functions in retroactive inhibition. *Journal of Experimental Psychology*, *47*, 285–293.

Brown, S., Steyvers, M., & Hemmer, P. (2007). Modeling experimentally induced strategy shifts. *Psychological Science*, *18*, 40–45.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380.

da Costa Pinto, A., & Baddeley, A. (1991). Where did you park your car? Analysis of a naturalistic long-term recency effect. *European Journal of Cognitive Psychology*, *3*, 297–313.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychological Review*, *112*, 3–42.

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science*, *29*, 145–193.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*, 415–434.

Eichenbaum, H. (1999). The hippocampus and mechanisms of declarative memory. *Behavioural Brain Research*, *103*, 123–133.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*, 94–107.

Estes, W. K. (1955a). Statistical theory of distributional phenomena in learning. *Psychological Review*, *62*, 369–377.

Estes, W. K. (1955b). Statistical theory of spontaneous recovery and regression. *Psychological Review*, *62*, 145–154.

Estes, W. K. (1959). Component and pattern models with Markovian interpretations. In *Studies in mathematical learning theory*. Stanford, CA: Stanford University Press.

Franzius, M., Wilbert, N., & Wiskott, L. (2011). Invariant object recognition and pose estimation with slow feature analysis. *Neural Computation*, *23*, 2289–2323.

Gershman, S. J., Schapiro, A. C., Hupbach, A., & Norman, K. A. (2013). Neural context reinstatement predicts memory misattribution. *Journal of Neuroscience*, *33*, 8590–8595.

Gill, P. R., Mizumori, S. J. Y., & Smith, D. M. (2011). Hippocampal episode fields develop with learning. *Hippocampus*, *21*, 1240–1249.

Glenberg, A. M., Bradley, M. M., Stevenson, J. A., Kraus, T. A., Tkachuk, M. J., & Gretz, A. L. (1980). A two-process account of long-term serial position effects. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 355–369.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Grossberg, S., & Pearson, L. R. (2008). Laminar cortical dynamics of cognitive and motor working memory, sequence learning and performance: toward a unified theory of how the cerebral cortex works. *Psychological Review*, *115*, 677–732.

Hasselmo, M. E. (2009). A model of episodic memory: mental time travel along encoded trajectories using grid cells. *Neurobiology of Learning and Memory*, *92*, 559–573.

Hasselmo, M. E. (2012). *How we remember: brain mechanisms of episodic memory*. Cambridge, MA: MIT Press.

Howard, M. W., Jing, B., Rao, V. A., Provyn, J. P., & Datey, A. V. (2009). Bridging the gap: transitive associations between items presented in similar temporal contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 391–407.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 923–941.

Howard, M.W., Shankar, K.H., Aue, W., & Criss, A.H. (0000). Revised. A quantitative model of time in episodic memory. Psychological Review.

Howard, M. W., Shankar, K. H., & Jagadisan, U. K. K. (2011). Constructing semantic representations from a gradually-changing representation of temporal context. *Topics in Cognitive Science*, *3*, 48–73.

Howard, M. W., Viskontas, I. V., Shankar, K. H., & Fried, I. (2012). A neural signature of mental time travel in the human MTL. *Hippocampus*, *22*, 1833–1847.

Howard, M. W., Youker, T. E., & Venkatadass, V. (2008). The persistence of memory: contiguity effects across several minutes. *Psychonomic Bulletin & Review*, *15*, 58–63.

Hyman, J. M., Ma, L., Balaguer-Ballester, E., Durstewitz, D., & Seamans, J. K. (2012). Contextual encoding by ensembles of medial prefrontal cortex neurons. *Proceedings of the National Academy of Sciences USA*, *109*, 5086–5091.

Jensen, O., & Lisman, J. E. (1996). Hippocampal CA3 region predicts memory sequences: accounting for the phase precession of place cells. *Learning and Memory*, *3*, 279–287.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information composite holographic lexicon. *Psychological Review*, *114*, 1–32.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*, 103–109.

Koppenaal, R. J. (1978). Time changes in the strengths of A–B, A–C lists; spontaneous recovery? *Journal of Verbal Learning and Verbal Behavior*, *2*, 310–319.

Kraus, B. J., Robinson, R. J., White, J. A., Eichenbaum, H., & Hasselmo, M. E. (2013). Hippocampal time cells: time versus path integration. *Neuron*, http://dx.doi.org/10.1016/j.neuron.2013.04.015.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Landauer, T. K., & Dumais, S. T. (1997). Solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Levy, W. B. (1996). A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus*, *6*, 579–590.

Li, N., & DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, *321*, 1502–1507.

Lisman, J. E. (1999). Relating hippocampal circuitry to function: recall of memory sequences by reciprocal dentate-CA3 interactions. *Neuron*, *22*, 233–242.

MacDonald, C. J., Carrow, S., Place, R., & Eichenbaum, H. (2013). Distinct hippocampal time cell sequences represent odor memories immobilized rats. *Journal of Neuroscience*, *33*, 14607–14616.

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal time cells bridge the gap in memory for discontiguous events. *Neuron*, *71*, 737–749.

Mankin, E. A., Sparks, F. T., Slayyeh, B., Sutherland, R. J., Leutgeb, S., & Leutgeb, J. K. (2012). Neuronal code for extended time in the hippocampus. *Proceedings of the National Academy of Sciences*, *109*, 19462–19467.

Manning, J. R., Polyn, S. M., Litt, B., Baltuch, G., & Kahana, M. J. (2011). Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 12893–12897.

Manns, J. R., Howard, M. W., & Eichenbaum, H. B. (2007). Gradual changes in hippocampal activity support remembering the order of events. *Neuron*, *56*, 530–540.

Melton, A. W., & Irwin, J. M. (1940). The influence of degree of interpolated learning on retroactive inhibition and the overt transfer of specific responses. *American Journal of Psychology*, *53*, 173–203.

Mensink, G. J. M., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review*, *95*, 434–455.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, *335*, 817–820.

Miyashita, Y., & Chang, H. S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, *331*, 68–70.

Moreton, B. J., & Ward, G. (2010). Time scale similarity and long-term memory for autobiographical events. *Psychonomic Bulletin & Review*, *17*, 510–515.

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review*, *104*, 839–862.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*, 482–488.

Murdock, B. B., Smith, D., & Bai, J. (2001). Judgments of frequency and recency in a distributed memory model. *Journal of Mathematical Psychology*, *45*, 564–602.

Naya, Y., Yoshida, M., & Miyashita, Y. (2001). Backward spreading of memory-retrieval signal in the primate temporal cortex. *Science*, *291*, 661–664.

Neath, I., & Crowder, R. G. (1996). Distinctiveness and very short-term serial position effects. *Memory*, *4*, 225–242.

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, *120*, 356–394.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in Cognitive Sciences*, *10*, 424–430.

Pastalkova, E., Itskov, V., Amarasingham, A., & Buzsaki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science*, *321*, 1322–1327.

Pollio, H. R., Kasschau, R. A., & DeNise, H. E. (1968). Associative structure and the temporal characteristics of free recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 190–197.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*, 1963–1966.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*, 129–156.

Postman, L., & Underwood, B. (1973). Critical issues in interference theory. *Memory & Cognition*, *1*, 19–40.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: a theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory*: *Vol. 14* (pp. 207–262). New York: Academic Press.

Rao, V. A., & Howard, M. W. (2008). Retrieved context and the discovery of semantic structure. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems*: *Vol. 20* (pp. 1193–1200). Cambridge, MA: MIT Press.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology*, *22*, 1622–1627.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*, 486–492.

Schwartz, G., Howard, M. W., Jing, B., & Kahana, M. J. (2005). Shadows of the past: Temporal retrieval effects in recognition memory. *Psychological Science*, *16*, 898–904.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*, 893–912.

Shankar, K. H., & Howard, M. W. (2012). A scale-invariant representation of time. *Neural Computation*, *24*, 134–193.

Shankar, K. H., Jagadisan, U. K. K., & Howard, M. W. (2009). Sequential learning using temporal context. *Journal of Mathematical Psychology*, *53*, 474–485.

Sirotin, Y. B., Kimball, D. R., & Kahana, M. J. (2005). Going beyond a single list: Modeling the effects of prior experience on episodic free recall. *Psychonomic Bulletin & Review*, *12*, 787–805.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neurosciences*, *27*, 161–168.

Sohal, V. S., & Hasselmo, M. E. (1998). Changes in $GABA_B$ modulation during a theta cycle may be analogous to the fall of temperature during annealing. *Neural Computation*, *10*, 869–882.

Wilson, M. A., & McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, *261*, 1055–1058.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Computation*, *14*, 715–770.

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, *75*, 168–179.