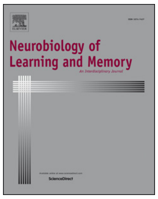
Contents lists available at ScienceDirect

Neurobiology of Learning and Memory

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



Is working memory stored along a logarithmic timeline? Converging evidence from neuroscience, behavior and models

Inder Singh[a,1], Zoran Tiganj[b,1], Marc W. Howard[b,*]

[a] Department of Psychology, Northeastern University, United States
[b] Department of Psychological and Brain Sciences, Boston University, United States

ARTICLE INFO

Keywords:
Working memory
Time cells
Judgments of recency

ABSTRACT

A growing body of evidence suggests that short-term memory does not only store the identity of recently experienced stimuli, but also information about when they were presented. This representation of 'what' happened 'when' constitutes a neural timeline of recent past. Behavioral results suggest that people can sequentially access memories for the recent past, as if they were stored along a timeline to which attention is sequentially directed. In the short-term judgment of recency (JOR) task, the time to choose between two probe items depends on the recency of the more recent probe but not on the recency of the more remote probe. This pattern of results suggests a backward self-terminating search model. We review recent neural evidence from the macaque lateral prefrontal cortex (lPFC) (Tiganj, Cromer, Roy, Miller, & Howard, in press) and behavioral evidence from human JOR task (Singh & Howard, 2017) bearing on this question. Notably, both lines of evidence suggest that the timeline is logarithmically compressed as predicted by Weber-Fechner scaling. Taken together, these findings provide an integrative perspective on temporal organization and neural underpinnings of short-term memory.

1. Introduction

Working memory is a term used to describe our ability to maintain information in an activated state. In typical working memory tasks, a relatively small amount of information is presented; after a few seconds, memory for the studied information is tested. Previous work has proposed stable persistent firing as a mechanism for maintaining memory of the stimulus identity across a temporal delay (Amit & Brunel, 1997; Chaudhuri & Fiete, 2016; Compte, Brunel, Goldman-Rakic, & Wang, 2000; Durstewitz, Seamans, & Sejnowski, 2000; Egorov, Hamam, Fransén, Hasselmo, & Alonso, 2002; Goldman-Rakic, 1995; Lundqvist, Herman, & Lansner, 2011; Mongillo, Barak, & Tsodyks, 2008; Sandberg, Tegnér, & Lansner, 2003). According to this view, the to-be-remembered information triggers a subset of neurons that remain active until the information is no longer needed. The identity of the stimulus is reflected in the subset of neurons that are activated. By examining which neurons are active at the end of the delay, one can infer what stimulus was presented at the beginning of the delay and use that information to correctly respond to the memory test.

In contrast to the classical view that information is maintained in working memory *via* a static code, a growing body of evidence suggests that working memory representations are dynamic rather than static, moving along trajectory during the delay interval (Spaak, Watanabe, Funahashi, & Stokes, 2017; Stokes, 2015). This observation is anticipated by recurrent neural network models in which an external stimulus can triggers a sequence of internal neural states (Buonomano & Maass, 2009; Maass, Natschläger, & Markram, 2002; White, Lee, & Sompolinsky, 2004). For instance, in echo state networks (Jaeger & Haas, 2004), an external stimulus provides input to a random connectivity matrix. The recurrent connectivity matrix induces a potentially complex "reservoir" of states that can be accessed some time after a stimulus. A recurrent network is a reservoir if the output, up to some tolerance, is a function of the input sequence up to some temporal window. However, the response of a particular unit triggered by a stimulus need not be unimodal in time nor a function only of one stimulus. Reservoir computing is powerful, but the complexity of the dynamics that can result from recurrent connections means that successfully decoding the sequence of past events that triggered a particular network state be challenging (Maass et al., 2002).

In this paper, we review evidence that suggests working memory maintenance could be understood as intermediate between these two approaches. Following previous theoretical (Shankar & Howard, 2012, 2013) and cognitive modeling (Howard, Shankar, Aue, & Criss, 2015) work, we consider the possibility that working memory maintenance

[*] Corresponding author.
E-mail address: marc777@bu.edu (M.W. Howard).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.nlm.2018.04.008
Received 10 December 2017; Received in revised form 31 March 2018; Accepted 6 April 2018
Available online 23 April 2018

produces a conjunctive code for what stimulus happened when in the past. Neurons participating in this representation would fire when a particular stimulus feature was experienced a certain time in the past. The "temporal receptive fields" of these predicted neurons are compact. Critically, temporal receptive fields are scale-invariant; neurons with temporal receptive fields further in the past also show an increase in their spread such that the width of their firing field goes up linearly with the time at which they peak. This property results in a logarithmic compression of the temporal dimension, enabling a natural account of behavioral effects in a range of memory paradigms (Howard et al., 2015).

Like reservoir computing approaches, this scale-invariant representation of the past gives rise to a dynamically-changing state during working memory maintenance as events fade into the past. Indeed, the mathematical implementation of this approach meets the formal definition of a liquid state machine (Shankar & Howard, 2013). However, unlike a more general reservoir computing models, this compressed representation is linear. This property enables straightforward decoding of what happened when in the past.

In this paper we review two threads of evidence that provide support for this hypothesis. First, we review recent evidence from working memory tasks with non-human primates (Tiganj, Cromer, Roy, Miller, & Howard, in press). This evidence demonstrates that neurons in lateral prefrontal cortex (lPFC) show conjunctive receptive fields for what happened when in a working memory maintenance task. As predicted by this approach, the neurons in this task have simple temporal receptive fields that systematically spread out as the delay unfolds. The form of the spread is consistent with logarithmic compression of the temporal dimension. Second, we review recent behavioral evidence from the short-term judgment of recency (JOR) task in humans (Singh & Howard, 2017). After rapid presentation of a list of stimuli, participants can determine which of the probes was experienced more recently. It is difficult to account for this ability if participants needed to learn a new decoder for every possible probe at every possible recency. Moreover, a careful examination of the amount of time to make a successful judgment suggests that participants scan along their memory, terminating the search when a probe is identified (Hacker, 1980; Hockley, 1984; McElree & Dosher, 1993; Muter, 1979). Recent evidence shows that the time to scan for a probe goes up sublinearly, approximately with the log of the probe's recency, as predicted by this approach (Singh & Howard, 2017).

2. Neurophysiological evidence for time as a supported, compressed dimension

Models with recurrent neural networks can maintain information about preceding stimuli (Buonomano & Maass, 2009; Buonomano & Merzenich, 1995; Maass et al., 2002; White et al., 2004). The recurrent dynamics and nonlinearities in the activation function can give rise to neurons with a variety of complex responses. Such responses include stable persistent firing and temporally modulated transient activity of various forms including decaying, growing, single- and multi-peak responses. In addition, the general form of dynamics in reservoir computing can produce a variety of responses that mix the stimulus identity and the elapsed time in a highly nonlinear fashion. This type of activity is referred to as switching selectivity, and includes neurons that switch between preferred stimuli during the delay interval (Chaisangmongkon, Swaminathan, Freedman, & Wang, 2017). Because of the complexity of the internal dynamics of the recurrent neural network, information about the elapsed time is not directly readable from the firing rate. Rather it must be decoded, which can be potentially challenging.

It has been long argued that brain represents sensory and motor continuous variables with a population code dominated by neurons that have unimodal tuning curves (Dayan & Abbott, 2001; Pouget, Dayan, & Zemel, 2000). These variables include for instance visual orientation (Hubel & Wiesel, 1968), sound frequency (Goldstein & Abeles, 1975) and direction of motion (Georgopoulos, Kalaska, Caminiti, & Massey, 1982). With this type of coding different sensory and motor variables are represented as supported dimension. Elapsed time could be represented in an analogous way with neurons that have unimodal receptive fields tuned to a particular time in the past. A sequence of such neurons with receptive fields distributed along the temporal axis would constitute a representation of elapsed time that can be decoded using the same mechanisms that can be applied to decode sensory variables.

A number of studies have reported *time cells* that activate sequentially, each for a circumscribed period of time (MacDonald, Lepage, Eden, & Eichenbaum, 2011; Pastalkova, Itskov, Amarasingham, & Buzsaki, 2008). It has been argued that time cells could play an important role in timing and memory (Eichenbaum, 2013, 2014; Howard & Eichenbaum, 2015; Howard et al., 2014; MacDonald, Fortin, Sakata, & Meck, 2014). After being initially observed in hippocampus, time cells have subsequently been observed in entorhinal cortex (Kraus et al., 2015), medial prefrontal cortex (Bolkan et al., 2017; Tiganj, Kim, Jung, & Howard, 2016) and striatum (Akhlaghpour et al., 2016; Jin, Fujii, & Graybiel, 2009; Mello, Soares, & Paton, 2015). If this temporal code is logarithmically-compressed, complying with the Weber-Fechner law, then this predicts two properties of time cells. First, time fields later in the sequence should be more broad (i.e., less precise) than those earlier in the sequence. Second, there should be more neurons with time fields early in the delay and fewer neurons representing times further in the past. Both of these properties have been observed (e.g., Howard et al., 2014; Jin et al., 2009; Kraus et al., 2015; Mello et al., 2015). A recent study (Tiganj, Cromer, et al., in press) extends this work by confirming another property predicted for time cells—that stimulus identity is encoded conjunctively with the time elapsed since the stimulus presentation (see also MacDonald, Carrow, Place, & Eichenbaum, 2013; Terada, Sakurai, Nakahara, & Fujisawa, 2017).

2.1. *Conjunctive coding of what and when on a logarithmically-compressed temporal scale in a working memory task*

This hypothesis was recently tested (Tiganj, Cromer, et al., in press) using data from an earlier report (Cromer, Roy, & Miller, 2010). The experimental paradigm was a delayed match to category task. In this task a sample stimulus was presented for 500 ms followed by a 1500 ms delay interval and then by a test stimulus. The sample stimuli were divided into two category sets based on visual similarity, animals and cars. The animals category set consisted of two categories, dog images and cat images. The car category set consisted of sports cars and sedan cars.

Even though this task did not require animals to maintain temporal information, the neurons active during the delay fired consistently only during a circumscribed period of the delay (Fig. 1a), leading to a sequence of time cells. Even in the absence of a specific task demand, this population conveyed information about the time at which the stimulus was experienced. These stimulus-specific time cells show the same qualitative properties as the time cells recorded from other studies: the width of the temporal tuning curves increased with and the number density of time cells decreased with the passage of time (Tiganj, Cromer, et al., in press).

Critically, different kinds of sample stimulus triggered distinct but overlapping sequences of time cells (compare three columns of Fig. 1a). Time cells preferentially tuned to a particular category were more likely to fire for visually similar stimuli (those from the same category set) than to visually dissimilar stimuli (those from a different category set), Fig. 1a.

The decreasing temporal accuracy in these sequentially-activated stimulus-specific time cells is consistent with the hypothesis that the temporal axis is logarithmically-compressed. Fig. 1b shows the heatmaps plotted against the logarithm of time, the width and the density of the temporal tuning curves is roughly constant as function of position within the sequence.

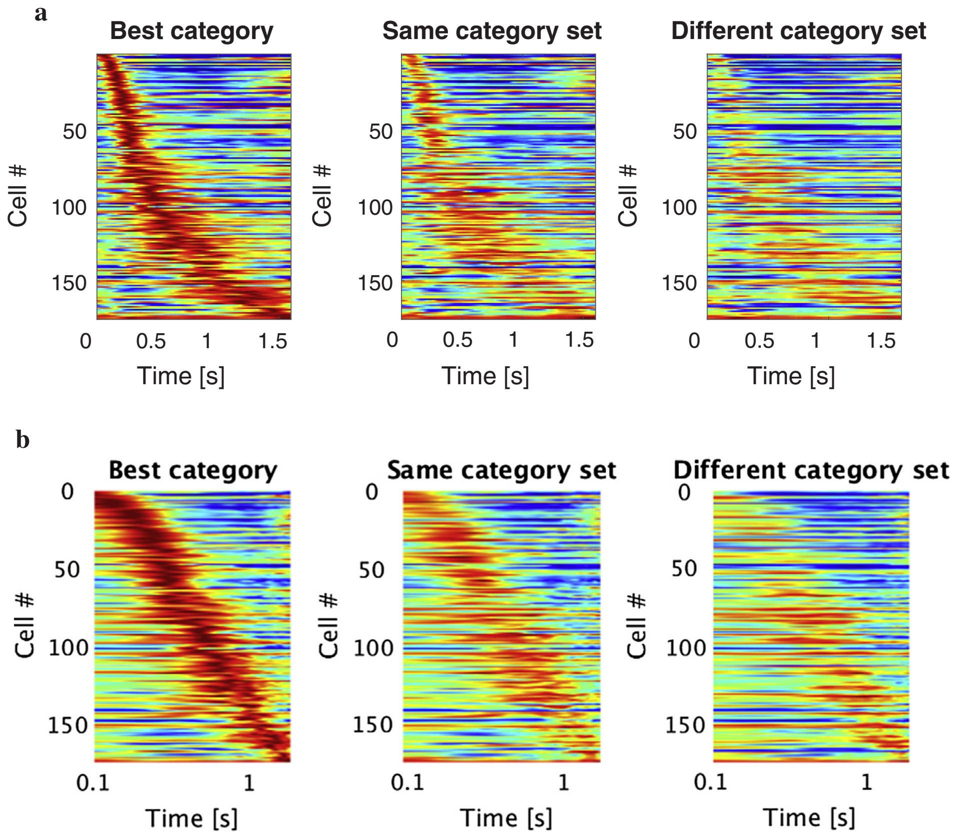


Fig. 1. (a) Sequentially activated time cells in lPFC encode time conjunctively with stimulus identity. The three heatmaps each show the response of every unit classified as a time cell. The heatmap on the left ("Best category") shows the response of each unit to the category that caused the highest response for that unit, sorted according to the units estimated time of peak activity. The second column ("Same category set") shows the heatmap for the same units, but for the other category from the same category set as that unit's "Best category." For instance, if a unit responded the most on trials in which the sample stimulus was chosen from the CAT category, then that unit's response to CAT trials would go in the first column and its response to DOG trials would go in the second column. The third column shows the response of each unit to trials on which the sample stimulus was from the other category set. Continuing with our example, a unit whose best category was CAT would have its response to CAR trials in the third column. The scale of the colormap is the same for all three plots and it is normalized for each unit such that red represents the unit's highest average firing rate and blue represents its lowest average firing rate across time bins. (b) Compression of the time axis is approximately logarithmic. Same data as in (a), but with time shown on a logarithmic-scale (note that the axes are trimmed to avoid edge effects). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Although these results are consistent with the predictions of a logarithmically-compressed representation of what happened when, they rule out many forms of a more general dynamic working memory. For instance, a general reservoir computing model could have easily generated much more complex receptive fields, with neurons showing complex receptive fields in time or responding to different stimuli at different times. These were not observed (Tiganj, Cromer, et al., in press). Moreover, there is nothing in the specification of a reservoir computing model that requires the temporal compression to be logarithmic. The results from (Tiganj, Cromer, et al., in press) suggest that the receptive fields were compact in the 2D space spanned with time and stimulus identity. Thus, time and stimulus identity were represented as continuous variables through a conjunctive (mixed selective) neural code. This is a very powerful representation because simple linear associations are sufficient to learn specific temporal relationships (Fusi, Miller, & Rigotti, 2016; Rigotti et al., 2013).

3. Behavioral evidence for a supported timeline

In the preceding section we saw that even in the absence of an explicit task demand to encode time, neurons in the macaque lPFC were sequentially activated, enabling reconstruction of temporal information. Notably, with the passage of time the temporal resolution of the representation became less accurate. This parallels the behavioral *recency effect* which is manifest as a reduction in the accuracy and an increase in response times for events that are further in the past. The recency effect is observed in all of the major memory paradigms and has similar properties over a range of time scales from a few hundred milliseconds up to at least tens of minutes (Glenberg et al., 1980; Monsell, 1978; Moreton & Ward, 2010; Neath, 1993; Shepard & Chang, 1963; Standing, 1973). The existence of a recency effect and its persistence over a range of time scales follow naturally if behavioral memory performance is extracted from a scale-invariant temporal representation of the past.

Cognitive psychologists considering how memory is accessed have proposed *scanning models* to describe the cognitive processes supporting a range of memory tasks. In visual scanning, people direct their gaze along a display to find a particular piece of information (e.g., Treisman & Gelade, 1980). Scanning models assume that an analogous process operates in memory (Hacker, 1980; Sternberg, 1966). Continuing the metaphor to vision, memory contains a store of information about many events that have been experienced in the past. However, to access the information from this memory store in enough detail, attention must be focused on a subset of the information in this memory store at a given time. Many scanning models assume that the information in the memory store is organized. For instance, in many scanning models remembered items are stored along a sequentially organized timeline. If memory is organized, then scanning models imply that the time to access a particular memory can reveal the organization of the memory store.

In the short-term JOR task (Hacker, 1980; Muter, 1979) participants are asked to make judgments about the relative recency[2] of two probe items. In this task, the participants are rapidly presented with a list of consonants with one letter every 180 ms. At the end of the list, participants are presented with two probes from the list and asked to indicate which of the two items was presented more recently. For instance in Fig. 2a, the probes are G and T and the correct answer is G. The key finding is that the correct response time to make a correct response depends *only* on the recency of the more recent probe. That is after learning the list in Fig. 2a, correct response time would be slower if G was replaced as a probe with Q, but would not be affected if T was

[2] In this task time and recency are confounded. Prior work using behavioral tasks (Brown, Morin, & Lewandowsky, 2006; Brown, Vousden, & McCormack, 2009; Hintzman, 2004) and electrophysiology (Kraus, Robinson, White, Eichenbaum, & Hasselmo, 2013) has shown that both temporal and ordinal information is stored in the brain.

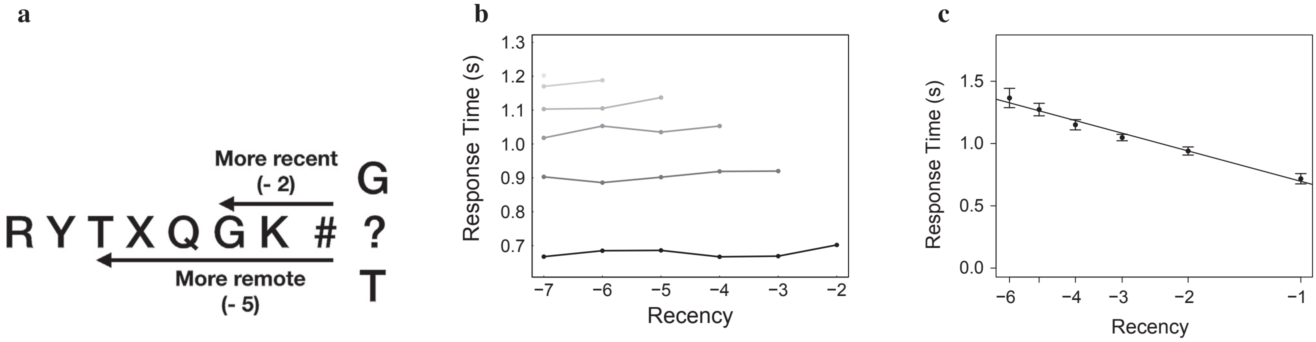


Fig. 2. Behavioral results from the short-term judgment of recency (JOR) task are consistent with backward scanning along a logarithmically-compressed timeline. (a) The participants are shown a list of letters followed by a probe containing two letters from the list. The participants are required to choose the more recent of the two probe items. (b) Empirical results in the JOR task. The response time for correct JORs is shown as a function of the more remote (*less* recent) probe. Different lines correspond to different recencies of the more recent probe. The darkest line corresponds to trials where the last item in the list was the more recent probe; successively lighter lines correspond to trials when the more recent probe was further in the past. The separation between the lines shows that correct RT depends strongly on the lag to the more recent probe (the separation between the lines), consistent with a backward scanning model. The flatness of each of the lines shows that the recency of the more remote (*less* recent) probe does not affect RT. (c) The median RT for selected probe as a function of its recency. The RT decreases sublinearly with recency (note the scaling of the x axis), as would be predicted if the timeline is compressed.

replaced as a probe with Y (Fig. 2b, Singh & Howard, 2017). This finding is as one would expect from a serial self-terminating backward scanning model.

Suppose that the participant sequentially compares the two probes to the contents of memory, stopping the search when one of the probes matches the information found in that region of the memory store. Moreover, suppose that memory is organized like a timeline and that the search begins at the present and proceeds towards the past. Because the search begins at the present and proceeds towards the past, it should take less time to find more recent probes. Because the search terminates when a match is found, the time necessary for a successful search for the more recent probe should not depend on the recency of the more remote (less recent) probe. This is just the result that is found experimentally.

If response times in the JOR task reflect the amount of "distance traversed" along the timeline, then the rate at which RT increases as the selected probe is chosen further and further into the past gives a measure of the organization of the temporal axis. The logarithmic compression in the neural data suggests that the one would expect a logarithmic compression in the reaction time data. Although it is difficult to argue specifically for a logarithmic compression, there is no question that the increase in RT is sublinear as the most recent probe recedes into the past (Fig. 2c). Prior modeling work has shown that the framework used in the proposed model can account for both accuracy and response times (Howard et al., 2015). While response times do not vary as a function of the more remote probe, accuracy shows a distance effect. In a self terminating scanning model, more remote items are missed at a higher rate than more recent items and the number of incorrect responses depends on contributions to the search from the less recent lags.

The finding of scanning along a logarithmic temporal axis in short-term JOR aligns with a number of other findings from long-term memory. For instance, in the numerical JOR task, participants report a numerical estimate of the recency of a probe stimulus. Numerical JORs are not a linear function of objective recency. Rather, they approximate a logarithmic function of actual recency (Hinrichs, 1970; Hinrichs & Buschke, 1968). Moreover, when participants are asked to judge the recency of a probe that has been presented multiple times, their judgments go up like the logarithm of the recency of the most recent presentation, but depend only weakly on the existence of an earlier presentation (Hintzman, 2010). These and other findings can be addressed with cognitive models based on a logarithmically-compressed representation of the past (Howard et al., 2015).

4. The challenge of decoding what and when in working memory

The previous section showed behavioral evidence that short-term human JOR performance relies on backward scanning of a logarithmically-compressed timeline. Earlier we saw that neural representations in a macaque working memory task appeared to construct a logarithmically-compressed timeline. It is of course possible that one has nothing to do with the other. Perhaps the behavioral evidence is actually generated by a different cognitive model. Perhaps the dislocation between species and/or the methodological differences between the behavioral tasks have conspired to create an illusion of a connection where none exists. Here we argue that taking the cognitive model for backward scanning seriously requires a neural representation very much like that observed in the macaque working memory task and argues against many possible representations of what and when information that would be subsumed under the more general framework of reservoir computing.

Consider the computational challenge of implementing a backward self-terminating search model neurally. The backward scanning model requires that we can query the content available at different times. That is, one must be able to specify a when and retrieve information about the what. This places a strong constraint on the organization of the memory store. It is not sufficient that the store contains information about what happened when, but also that the information about different times can be separately queried. Moreover, because at most a few seconds intervene between presentation of a novel list and the JOR test, it is difficult to reconcile successful performance on this task with models that require extensive training to learn a decoder. It is known that humans can perform the JOR task with unfamiliar pictures (Hintzman, 2005). If a decoder for what happened when must be learned, this immediately raises the question of how the training signal should be generated. It is circular to assume that the training signal contains information about what happened when, i.e. in order to learn what happened when one starts with information about what happened when. Moreover, because techniques for learning via gradient descent are typically slow, requiring many trials to successfully learn, there is the additional technical challenge of generating a decoder that can be used for unfamiliar pictures.

Fig. 3 provides a schematic depiction of the properties that would be necessary to account for this set of findings. The ability to separately decode both identity and temporal information follows from a linear system in which each possible stimulus triggers a sequence of activity that is not affected by subsequent stimulus presentations. In this way each unit is identified with a single time point in the past and a

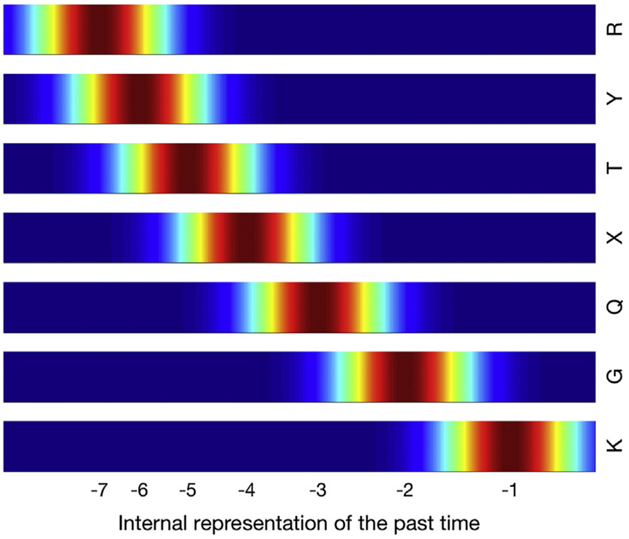


Fig. 3. Conjunctive representation of what happened when. Each horizontal strip shows the activation of a different set of units triggered by each of several possible stimuli. The pattern of activation across the units is logarithmically compressed such that units peaking further in the past (on the left of the figure) have wider temporal fields (as in Fig. 1). This figure shows the state of the representation after presenting the list from Fig. 2a at a constant rate. Note that each stimulus shows the same width across units. The temporal compression can be seen by noting that the location of the peaks across cells overlap more for stimuli further in the past. If scanning proceeds at a constant rate in cell space, the time to find a target depends on the logarithm of its recency as found in the behavioral data.

projection from the stimulus space.[3] However, in order to rapidly decode the recency of arbitrary stimuli appearing in arbitrary sequences it is necessary that the response of the neurons coding for the history is available in a form that does not require learning a different decoder for each time point. In the case of the proposed model, the information is encoded through a set of leaky integrators and decoded through a linear transformation that gives rise to a logarithmically-compressed sequential activation constituting a timeline. Because there is no mixing of stimulus dimensions through the stages of the network and the same form of stimulus coding is respected at each point of the timeline, there is no need to learn a sequence-specific decoder.

5. Discussion

The results reviewed in this paper are perfectly consistent with a dynamic view of working memory (Spaak et al., 2017; Stokes, 2015) and a subset of reservoir computing models. However, they imply a coding scheme more specific than general reservoir computing or recurrent network models.

First, the results here suggest that time in the working memory representation is logarithmically compressed. Logarithmic compression provides a natural implementation of the Weber-Fechner law and is optimal in the sense that it enables comparable amount of information to be extracted from the past at different scales of resolution (Howard & Shankar, in press). Logarithmic compression requires a system that is scale-invariant. Scale-invariance is very difficult to implement in a linear chain of neurons (Goldman, 2009; Liu, Tiganj, Hasselmo, & Howard, in preparation). In the context of reservoir computing, scale-invariance implies a broad and specific spectrum of eigenvalues of the dynamics of the system. Logarithmic compression implies that the spectrum of eigenvalues gives a distribution of time constants τ that goes down like τ^{-1}. The long tail of this power law distribution requires that the system has some very long time constants. It is possible that these time constants are not the consequence of recurrent connections but that they result from very slow intrinsic properties of individual neurons (Egorov et al., 2002; Fransén, Tahvildari, Egorov, Hasselmo, & Alonso, 2006; Tiganj, Hasselmo, & Howard, 2015).

Second, in order for the time decoder to be extensible to novel stimuli and novel lists, the dynamics of the system must be linear (or nearly so). In a linear system, the state of the network can be expressed as a sum of the previously-presented stimuli. This means that the information about whether or not a particular stimulus was presented at a particular time can be decoupled from the information carried about other stimuli. This property is extremely useful in developing models of working memory in which arbitrary information can be queried from novel temporal sequences.

Both of these properties are straightforward to implement in a computational model based on the Laplace transform (Shankar & Howard, 2012, 2013). Intuitively, the set of cells coding the Laplace transform holds information about the past, but in a way that is distributed across all of the cells. Unlike a vector space representation, any individual cell does not carry unique identifiable information about the past. However, a subset of cells with nearby time constants uniquely codes for the history at a corresponding point in the past. By including a wide range of time constants in the set of cells coding for the Laplace transform, one can trace out the entire timeline of the past. This model meets the formal requirements for a reservoir computer and a liquid state machine. However, it is also a linear system. Moreover because the different time scales decouple from one another (unlike in a linear chain), the problem of constructing a spectrum of time constants that goes like τ^{-1} can be addressed using very general physical principles (Amir, Oreg, & Imry, 2012). This formal approach is beyond the scope of the current paper, but it has been applied to a range of problems in neuroscience (Howard & Eichenbaum, 2013; Howard et al., 2014) and cognitive psychology (Howard et al., 2015).

5.1. Open questions

The hypothesis advanced here—that working memory is constructed from a linear system with logarithmic time compression—-makes a number of testable predictions, both neurophysiologically and behaviorally.

Although there is good evidence for a logarithmic temporal scale in behavior (e.g., Hinrichs & Buschke, 1968) the quantitative evidence for *logarithmic* compression of time has not been established quantitatively with neurophysiological data. That is, although there is abundant evidence that time cells in a range of brain regions and tasks are compressed (e.g., Jin et al., 2009; Mello et al., 2015; Salz et al., 2016; Tiganj, Shankar, & Howard, 2017), it has not been quantitatively established that this compression is logarithmic.

A logarithmically compressed timeline could be an important part of neural mechanism needed for performing the JOR task. However, to fully describe the neural underpinnings of the JOR task it is necessary to explain how the compressed timeline can be sequentially scanned and how the output of that scanning can be used to accumulate evidence for each presented probe. This problem is conceptually similar to the problem of visual attention, where subjects sequentially scan the visual space (Howard, 2018). While details of such a circuit remain outside of the scope of this review, we speculate that the sequential scanning could be implemented with the same type of circuit as the compressed memory itself. If the activity of a set of neurons can be used to gate the output of the timeline, then attention to a particular point in the past amounts to setting the gate to the corresponding time cells. Sequentially scanning along the past amounts to sequentially moving the location of this gate. In order to account for response times, one would allow this gated output from the timeline to provide input to an evidence accumulating circuit (Ratcliff, 1978; Usher & McClelland, 2001).

[3] In Fig. 3 the mapping from stimulus space to the units is "localist" for clarity such that each unit responds only to a single stimulus. In general, this is not necessary.

The mathematics of the computational model can generate a scale-invariant timeline extending arbitrarily far into the past. Neural constraints would certainly limit the extent of such a timeline in practice. Behavioral evidence suggests scale-invariance in memory for at least tens of minutes (Howard, Youker, & Venkatadass, 2008). It remains unclear whether time cells can support the memory representation for that long. Although existing neural recordings have measured time cell sequences extending at least a minute (Bolkan et al., 2017; Mello et al., 2015), existing neural data do not address longer time scales. However, multiple studies have reported gradual changes in neural activity across spectrum of timescales, from minutes to days (Cai et al., 2016; Mankin, Diehl, Sparks, Leutgeb, & Leutgeb, 2015; Manns, Howard, & Eichenbaum, 2007; Mau et al., in press; Rashid et al., 2016). It is possible that those very slow changes reflect sequentially activating time cells over much longer time scales than have thus far been observed.

6. Conclusions

We reviewed recent neurophysiological and behavioral evidence that suggests that the representations supporting working memory performance have a very specific form. Our hypothesis is that sets of neurons represent what happened when in a conjunctive manner with logarithmic compression of the time axis. This hypothesis implies a specific form of a dynamic working memory representation that is a subset of the more general mathematical framework of reservoir computing. Rapid expression of arbitrary decoders requires linear dynamics. Logarithmic compression requires that the dynamics are scale-invariant. Both of these properties are satisfied by a recent proposal for constructing a scale-invariant representation of a temporal history.

Acknowledgments

The authors gratefully acknowledge support from ONR MURI N00014-16-1-2832, NIBIB R01EB022864, NIMH R01MH112169, and NSF IIS 1631460.

References

Akhlaghpour, H., Wiskerke, J., Choi, J. Y., Taliaferro, J. P., Au, J., & Witten, I. (2016). Dissociated sequential activity and stimulus encoding in the dorsomedial striatum during spatial working memory. *eLife, 5*, e19507.

Amir, A., Oreg, Y., & Imry, Y. (2012). On relaxations and aging of various glasses. *Proceedings of the National Academy of Sciences, 109*(6), 1850–1855.

Amit, D. J., & Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex, 7*(3), 237–252.

Bolkan, S. S., Stujenske, J. M., Parnaudeau, S., Spellman, T. J., Rauffenbart, C., Abbas, A. I., ... Kellendonk, C. (2017). Thalamic projections sustain prefrontal activity during working memory maintenance. *Nature Neuroscience, 20*(7), 987–996. http://dx.doi.org/10.1038/nn.4568.

Brown, G. D. A., Morin, C., & Lewandowsky, S. (2006). Evidence for time-based models of free recall. *Psychonomic Bulletin and Review, 13*(4), 717–723.

Brown, G. D. A., Vousden, J. I., & McCormack, T. (2009). Memory retrieval as temporal discrimination. *Journal of Memory and Language, 60*(1), 194–208.

Buonomano, D. V., & Maass, W. (2009). State-dependent computations: Spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience, 10*(2), 113–125. http://dx.doi.org/10.1038/nrn2558.

Buonomano, D. V., & Merzenich, M. M. (1995). Temporal information transformed into a spatial code by a neural network with realistic properties. *Science, 267*(5200), 1028.

Cai, D. J., Aharoni, D., Shuman, T., Shobe, J., Biane, J., Song, W., ... Silva, A. (2016). A shared neural ensemble links distinct contextual memories encoded close in time. *Nature, 534*(7605), 115–118.

Chaisangmongkon, W., Swaminathan, S. K., Freedman, D. J., & Wang, X. J. (2017). Computing by robust transience: How the fronto-parietal network performs sequential, category-based decisions. *Neuron, 93*(6), 1504–1517.

Chaudhuri, R., & Fiete, I. (2016). Computational principles of memory. *Nature Neuroscience, 19*(3), 394–403.

Compte, A., Brunel, N., Goldman-Rakic, P. S., & Wang, X. J. (2000). Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex, 10*(9), 910–923.

Cromer, J. A., Roy, J. E., & Miller, E. K. (2010). Representation of multiple, independent categories in the primate prefrontal cortex. *Neuron, 66*(5), 796–807. http://dx.doi.org/10.1016/j.neuron.2010.05.005.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems.* Cambridge, MA: MIT Press.

Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience, 3*, 1184–1191.

Egorov, A. V., Hamam, B. N., Fransén, E., Hasselmo, M. E., & Alonso, A. A. (2002). Graded persistent activity in entorhinal cortex neurons. *Nature, 420*(6912), 173–178.

Eichenbaum, H. (2013). Memory on time. *Trends in Cognitive Sciences, 17*(2), 81–88. http://dx.doi.org/10.1016/j.tics.2012.12.007.

Eichenbaum, H. (2014). Time cells in the hippocampus: A new dimension for mapping memories. *Nature Reviews Neuroscience, 15*(11), 732–744. http://dx.doi.org/10.1038/nrn3827.

Fransén, E., Tahvildari, B., Egorov, A. V., Hasselmo, M. E., & Alonso, A. A. (2006). Mechanism of graded persistent cellular activity of entorhinal cortex layer V neurons. *Neuron, 49*(5), 735–746.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Ppinion in Neurobiology, 37*, 66–74.

Georgopoulos, A. P., Kalaska, J. F., Caminiti, R., & Massey, J. T. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience, 2*(11), 1527–1537.

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.

Goldman, M. S. (2009). Memory without feedback in a neural network. *Neuron, 61*(4), 621–634.

Goldman-Rakic, P. (1995). Cellular basis of working memory. *Neuron, 14*, 477–485.

Goldstein, M. H., Jr., & Abeles, M. (1975). Single unit activity of the auditory cortex. *Auditory system (199–218)*. Springer.

Hacker, M. J. (1980). Speed and accuracy of recency judgments for events in short-term memory. *Journal of Experimental Psychology: Human Learning and Memory, 15*, 846–858.

Hinrichs, J. V. (1970). A two-process memory-strength theory for judgment of recency. *Psychological Review, 77*(3), 223–233.

Hinrichs, J. V., & Buschke, H. (1968). Judgment of recency under steady-state conditions. *Journal of Experimental Psychology, 78*(4), 574–579.

Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition, 32*(2), 336–350.

Hintzman, D. L. (2005). Memory strength and recency judgments. *Psychonomic Bulletin & Review, 12*(5), 858–864.

Hintzman, D. L. (2010). How does repetition affect memory? Evidence from judgments of recency. *Memory & Cognition, 38*(1), 102–115.

Hockley, W. E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(4), 598–615.

Howard, M. W. (2018). Memory as perception of the past: Compressed time in mind and brain. *Trends in Cognitive Sciences, 22*, 124–126.

Howard, M. W., & Eichenbaum, H. (2013). The hippocampus, time, and memory across scales. *Journal of Experimental Psychology: General, 142*(4), 1211–1230. http://dx.doi.org/10.1037/a0033621.

Howard, M. W., & Eichenbaum, H. (2015). Time and space in the hippocampus. *Brain Research, 1621*, 345–354.

Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., & Eichenbaum, H. (2014). A unified mathematical framework for coding time, space, and sequences in the hippocampal region. *Journal of Neuroscience, 34*(13), 4692–4707. http://dx.doi.org/10.1523/JNEUROSCI.5808-12.2014.

Howard, M. W., & Shankar, K. H. (2018). Neural scaling laws for an uncertain world (in press). Available from arXiv:1607.04886.

Howard, M. W., Shankar, K. H., Aue, W., & Criss, A. H. (2015). A distributed representation of internal time. *Psychological Review, 122*(1), 24–53.

Howard, M. W., Youker, T. E., & Venkatadass, V. (2008). The persistence of memory: Contiguity effects across several minutes. *Psychonomic Bulletin & Review, 15*(PMC2493616), 58–63.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195*(1), 215–243.

Jaeger, H., & Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science, 304*(5667), 78–80. http://dx.doi.org/10.1126/science.1091277.

Jin, D. Z., Fujii, N., & Graybiel, A. M. (2009). Neural representation of time in cortico-basal ganglia circuits. *Proceedings of the National Academy of Sciences, 106*(45), 19156–19161.

Kraus, B. J., Brandon, M. P., Robinson, R. J., Connerney, M. A., Hasselmo, M. E., & Eichenbaum, H. (2015). During running in place, grid cells integrate elapsed time and distance run. *Neuron, 88*(3), 578–589.

Kraus, B. J., Robinson, R. J., 2nd, White, J. A., Eichenbaum, H., & Hasselmo, M. E. (2013). Hippocampal time cells: Time versus path integration. *Neuron, 78*(6), 1090–1101. http://dx.doi.org/10.1016/j.neuron.2013.04.015.

Liu, Y., Tiganj, Z., Hasselmo, M. E., & Howard, M. W. (in preparation). Biological simulation of scale-invariant time cells.

Lundqvist, M., Herman, P., & Lansner, A. (2011). Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. *Journal of Cognitive Neuroscience, 23*(10), 3008–3020.

Maass, W., Natschläger, T., & Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation, 14*(11), 2531–2560. http://dx.doi.org/10.1162/089976602760407955.

MacDonald, C. J., Carrow, S., Place, R., & Eichenbaum, H. (2013). Distinct hippocampal time cell sequences represent odor memories immobilized rats. *Journal of Neuroscience, 33*(36), 14607–14616.

MacDonald, C. J., Fortin, N. J., Sakata, S., & Meck, W. H. (2014). Retrospective and prospective views on the role of the hippocampus in interval timing and memory for elapsed time. *Timing & Time Perception, 2*(1), 51–61.

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*(4), 737–749.

Mankin, E. A., Diehl, G. W., Sparks, F. T., Leutgeb, S., & Leutgeb, J. K. (2015). Hippocampal CA2 activity patterns change over time to a larger extent than between spatial contexts. *Neuron, 85*(1), 190–201. http://dx.doi.org/10.1016/j.neuron.2014.12.001.

Manns, J. R., Howard, M. W., & Eichenbaum, H. B. (2007). Gradual changes in hippocampal activity support remembering the order of events. *Neuron, 56*(3), 530–540.

Mau, W., Sullivan, D. W., Kinsky, N. R., Hasselmo, M. E., Howard, M. W., & Eichenbaum, H. (2018). The same hippocampal CA1 population simultaneously codes temporal information over multiple timescales. *Current Biology* (in press).

McElree, B., & Dosher, B. A. (1993). Serial recovery processes in the recovery of order information. *Journal of Experimental Psychology: General, 122*, 291–315.

Mello, G. B., Soares, S., & Paton, J. J. (2015). A scalable population code for time in the striatum. *Current Biology, 25*(9), 1113–1122.

Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science, 319*(5869), 1543–1546.

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology, 10*, 465–501.

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

Muter, P. (1979). Response latencies in discriminations of recency. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 160–169.

Neath, I. (1993). Distinctiveness and serial position effects in recognition. *Memory & Cognition, 21*, 689–698.

Pastalkova, E., Itskov, V., Amarasingham, A., & Buzsaki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science, 321*(5894), 1322–1327.

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience, 1*(2), 125–132.

Rashid, A. J., Yan, C., Mercaldo, V., Hsiang, H. L. L., Park, S., Cole, C. J., ... Josselyn, S. A. (2016). Competition between engrams influences fear memory formation and recall. *Science, 353*(6297), 383–387.

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

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature, 497*(7451), 585–590. http://dx.doi.org/10.1038/nature12160.

Salz, D. M., Tiganj, Z., Khasnabish, S., Kohley, A., Sheehan, D., Howard, M. W., & Eichenbaum, H. (2016). Time cells in hippocampal area CA3. *Journal of Neuroscience, 36*, 7476–7484.

Sandberg, A., Tegnér, J., & Lansner, A. (2003). A working memory model based on fast hebbian learning. *Network: Computation in Neural Systems, 14*(4), 789–802.

Shankar, K. H., & Howard, M. W. (2012). A scale-invariant internal representation of time. *Neural Computation, 24*(1), 134–193.

Shankar, K. H., & Howard, M. W. (2013). Optimally fuzzy temporal memory. *Journal of Machine Learning Research, 14*, 3753–3780.

Shepard, R. N., & Chang, J. J. (1963). Forced-choice tests of recognition memory under steady-state conditions. *Journal of Verbal Learning and Verbal Behavior, 2*(1), 93–101.

Singh, I., & Howard, M. W. (2017). Recency order judgments in short term memory: Replication and extension of hacker (1980). *bioRxiv*, 144733.

Spaak, E., Watanabe, K., Funahashi, S., & Stokes, M. G. (2017). Stable and dynamic coding for working memory in primate prefrontal cortex. *Journal of Neuroscience*, 3316–3364.

Standing, L. (1973). Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology, 25*(2), 207–222.

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153*, 652–654.

Stokes, M. G. (2015). "Activity-silent" working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences, 19*(7), 394–405.

Terada, S., Sakurai, Y., Nakahara, H., & Fujisawa, S. (2017). Temporal and rate coding for discrete event sequences in the hippocampus. *Neuron*.

Tiganj, Z., Cromer, J. A., Roy, J. E., Miller, E. K., & Howard, M. W. (2018). Compressed timeline of recent experience in monkey lPFC. *Journal of Cognitive Neuroscience* (in press).

Tiganj, Z., Hasselmo, M. E., & Howard, M. W. (2015). A simple biophysically plausible model for long time constants in single neurons. *Hippocampus, 25*(1), 27–37.

Tiganj, Z., Kim, J., Jung, M. W., & Howard, M. W. (2016). Sequential firing codes for time in rodent mPFC. *Cerebral Cortex, 27*(12), 5663–5671.

Tiganj, Z., Shankar, K. H., & Howard, M. W. (2017). Scale invariant value computation for reinforcement learning in continuous time. In *AAAI 2017 spring symposium series – Science of intelligence: Computational principles of natural and artificial intelligence*.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97–136.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*(3), 550–592.

White, O. L., Lee, D. D., & Sompolinsky, H. (2004). Short-term memory in orthogonal neural networks. *Physical Review Letters, 92*(14), 148102.