Investigating Mechanisms of Maintaining Multiple Items in Short-Term Memory Through Recurrent Neural Networks

Chaihyun (Catherine) Lee, Xinyu Wei, Nicolas Y. Masse
Freedman Lab, Neurobiology, University of Chicago, Chicago, IL

Key words: **short-term memory; recurrent neural network; sequential items**

Introduction

While short-term memory is incredibly important for our daily lives and behaviors, we still do not fully understand the underlying neural mechanisms. There have been recent *in vivo* studies suggesting that short-term memory is maintained by persistent neural activity when an animal is asked to remember information of a single stimulus in its short-term memory [Fig.1]. While many studies have examined how single items are encoded and maintained in working memory (WM), much less is known about how multiple items with interference are encoded largely due to the difficulties in training animals and recording neural data from these tasks. In response to these difficulties, we investigated working memory mechanisms for multiple items by training a biologically-inspired recurrent neural network on tasks that required maintenance of multiple items in memory, in order to derive putative mechanisms that our brains might be using. Furthermore, by understanding how interference between multiple items affects the networks' memory task performance, we may be able to design a novel neural network with enhanced working memory capacity.

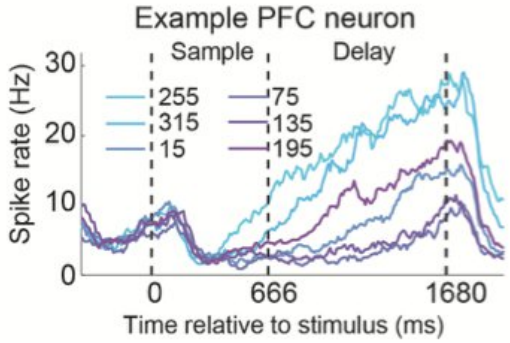

Figure 1. Persistent neural activity during delay period (from Masse et al. 2017)
This figure shows the neural activity of an example neuron recorded from prefrontal cortex (PFC) during a working memory task. The subject was trained to indicate whether a sample and test motion direction stimulus, separated by a delay period (between 666 ms and 1680 ms), matched. The eight colored curves represent the mean spike rate of the neuron for each of the eight sample motion directions (angle indicated in degrees) was presented. We observe that the identity of the sample stimulus is encoded in the neural activity of this example neuron throughout the delay period. This is consistent with previous studies that have suggested that persistent activity can maintain information in working memory.

Materials and Methods

Biologically-inspired recurrent neural network

We used a biologically-inspired recurrent neural network (RNN) to examine the putative neural mechanisms underlying working memory. In a typical neural network, neurons can exhibit positive (excitatory) or negative (inhibitory) activities without constraint. However, in a biological setting, there is a strict division between excitatory and inhibitory neurons. Therefore, we set 80% of the neurons to be excitatory and 20% to be inhibitory, consistent with what is known *in vivo*.

Furthermore, the firing of a neuron in a brain is relatively low either due to the metabolic costs (Laughlin, 1998), or possibly to facilitate information read-out from the neural activities (Hawkins, 2016 & Olshausen, 2004). Therefore, we added a spike cost to our network to encourage the network to solve problems with a low amount of neural activity.

Most importantly, synaptic connections between recurrently connected neurons were modulated by synaptic plasticity (STP) in order to mimic the ways in which brain might be encoding information for various tasks. Recent research suggests that short-term synaptic plasticity which is maintained through short-term changes in the neural network is critical for maintaining information during working memory tasks. STP in brain alters synaptic efficacy based on previous presynaptic

activity mainly through two mechanisms: presynaptic activity typically increases residual calcium concentration in the presynaptic terminal (short-term facilitation) and decreases available neurotransmitters (short-term depression) [Fig.2]. In our neural network, we incorporated two terms that represented the calcium concentration and amount of available neurotransmitters. The synaptic efficacy is proportional to the product of the two terms.

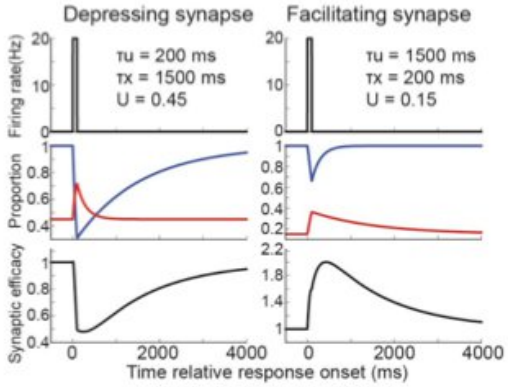

Figure 2. Short-term synaptic facilitation and depression (from Masse et al.)
This figure demonstrates how simulating residual calcium concentration and neurotransmitter availabilities can cause short-term synaptic depression and facilitation based on their time constants. The plots in the second row shows the traces of available neurotransmitters (blue) and the amount of residual calcium (red). The synaptic efficacy as shown in the third row plot is determined by the product of the available neurotransmitters and the residual calcium level. Such short-term synaptic plasticity can be used to encode information short-term.

Task

In the task, after a short initial fixation period of 200 ms, we show a sequence of motion directions pulses, each lasting for 200 ms and followed by a delay period of 200 ms in which no stimulus is present [Fig.4]. The last motion direction is followed by a long delay period of 500 ms. Until the end of the long delay period, the network should always fixate. The long delay is followed by a sequence of test periods during which the network is cued to recall the motion directions in order of their presentation. Each test period is cued by one of the eight response cues to indicate which stimuli the network should recall. The response cue for each motion direction lasts for 200 ms, and is followed by a 200 ms short delay period before being cued to recall the next motion stimuli. When the response cue is on, the network should recall by making a saccade to the direction it retrieved by generating a high activity in the output neurons that is associated with the motion direction.

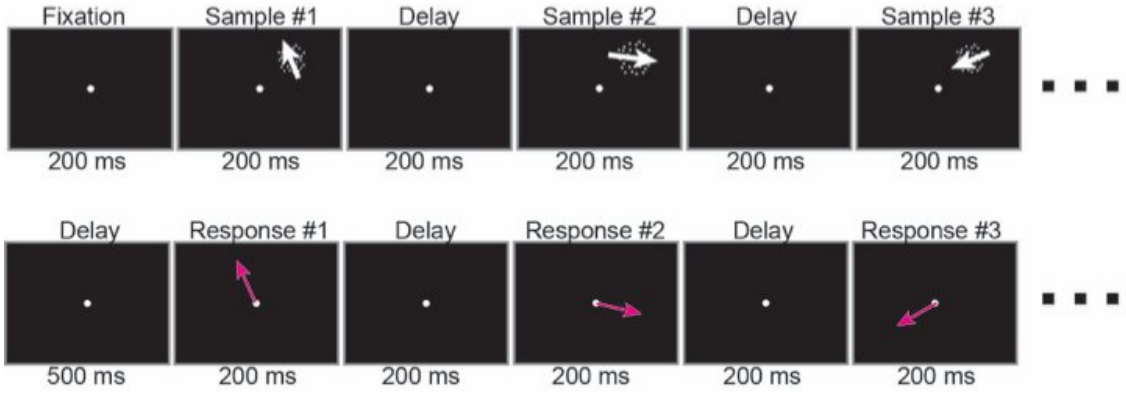

Figure 4. Task
Schematic diagram of the task in which the network is shown a sequence of stimulus chosen from the eight possible motion directions and asked to recall the directions in the same order later on.

Network configuration

For this experiment, the input layer of the model consists of 24 motion-tuned neurons for each of the four receptive fields, two fixation-tuned neurons, and one cue-tuned neuron for each pulse presented during the task [Fig.3]. The activities of the motion-tuned neurons are determined based on the motion stimuli that is being presented at each time step of the trial as well as the neurons' preferred directions. The fixation-tuned neurons fire when the network should be fixating, during the stimuli presentation period and the delay period. The cue-tuned neuron for each motion directions fire during the presentation

of that specific motion direction and when the network needs to recall that direction. The input neurons project to 100 hidden neurons, with 80 being excitatory and 20 being inhibitory. The output is a one-hot vector with a length of nine, with eight of them representing the eight motion directions respectively and one representing fixation.

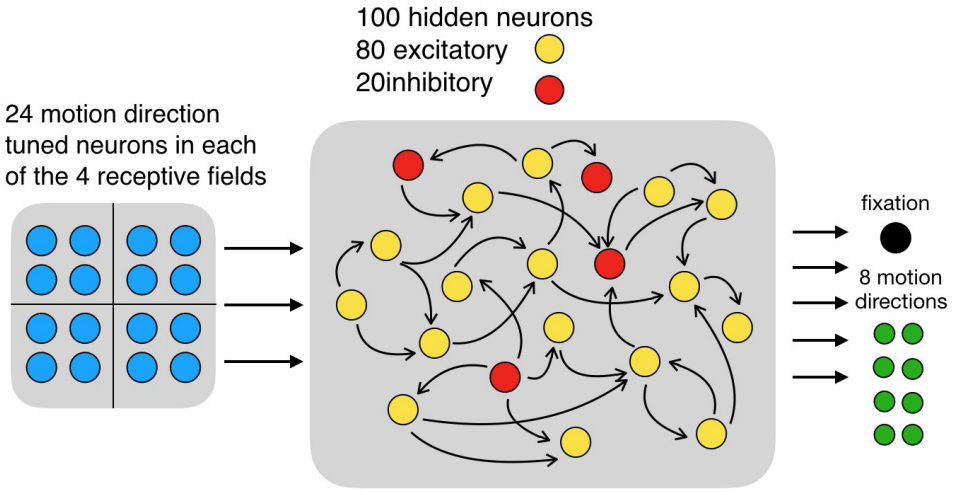

Figure 3. Overall schematic diagram of a neural network with EI network

Training method

To examine the effect of the number of items to memorize on network's behavior, we trained networks on trials with different number of motion stimuli. Connection weights and biases were trained using stochastic gradient descent to minimize a loss function comprised of two terms: 1) the cross-entropy between the expected output and actual response of the output neurons and 2) the mean spike rate of the hidden layer neurons. We started from training the network on a sequence of three motion directions, and increased the number of motion directions until the network's learning performance could not reach the accuracy of 85% any more. Throughout training, the accuracy for each motion stimulus and the neural and synaptic activities of hidden neurons were recorded to be used for analysis. In order to make the network flexible to various lengths of delay period, we make the short delay periods between stimuli and response cues vary at each iteration of the training, but no longer than 500 ms.

Results

Accuracy

Throughout the test period of the task, we determined the network's recall of the stimuli based on the activities of its output neurons. The motion direction represented by the neuron with the highest activity is the network's recall. The pulse accuracy was calculated by dividing the number of the times the network correctly recall a specific pulse in the sequence by the total number of trials. The overall task accuracy was calculated by averaging all the pulse accuracies in the sequence.

For less than six total motion stimuli, the network was able to achieve 90% accuracy in recalling all stimuli after 38,000 iterations. For tasks with more than six total motion stimuli, the network required much more number of iterations to learn.

Encoding and maintenance of information via neural activity and synaptic plasticity

To investigate how the network maintained information of multiple stimuli, we decoded the neuronal activity and synaptic efficacy of the hidden neurons. We used a linear support vector machine (SVM) to classify each sample motion directions from a population of neuronal activity or synaptic efficacy. The decoding accuracy was calculated as the percentage of the motion directions that can be correctly predicted by the SVM classifier. The decoding accuracy represented the amount of information about the stimuli that can be extracted from the hidden neurons.

For both tasks that involved only a few stimuli (3 sample stimuli) and tasks that involved more stimuli (8 sample stimuli), information about the stimuli could be extracted well from the synaptic efficacy during the delay period, in which the stimuli were no longer presented, but the network needed to maintain the information for later recall. However, the information about the stimuli was no longer stored in the neuronal activity during the delay period [Fig. 5]. This encoding and maintenance behavior aligns with previous research conducted by Masse et al., 2018 that highlighted the importance of synaptic plasticity in maintaining information saliently.

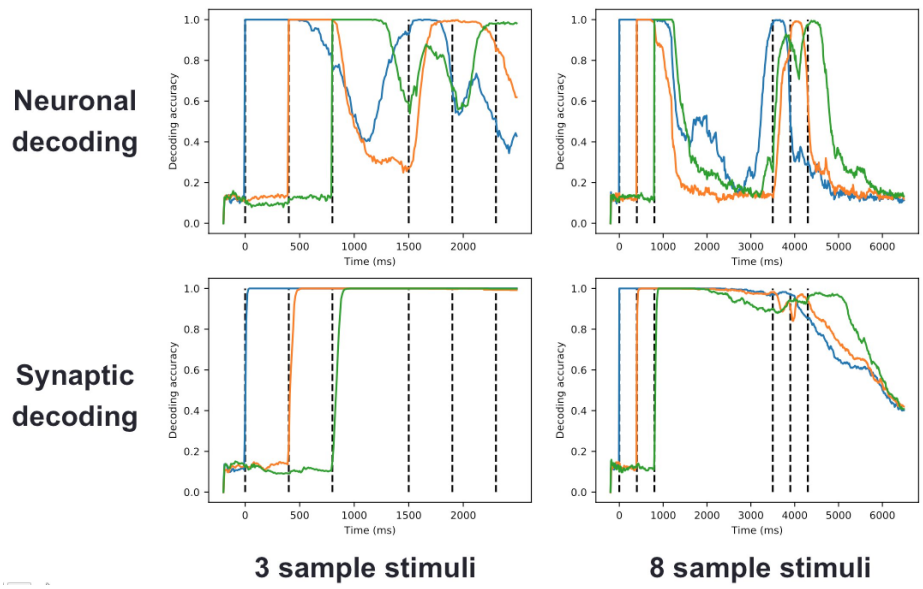

Figure 5. Decoding accuracy of neuronal and synaptic activity of the hidden neurons for the first three stimulus
Biologically-inspired RNN were trained on tasks that require the network to maintain varying number of sequentially presented stimuli. This figure shows the decoding accuracy of hidden neuronal activity and synaptic efficacy for the first three motion direction pulses for a task in which only three stimuli were presented (left panels) and another task in which eight stimuli were presented (right panels).

Sparse encoding

One possible way of maintaining multiple stimuli in short-term memory could be by using sparse encoding. By having a selective number of hidden neurons encoding and maintaining information about a specific stimulus, the network may be able to minimize the interference coming from other stimuli in the sequence and successfully remember each item.

In order to analyze the encoding specificity of the hidden neurons, we calculated the number of neurons that are robustly encoding two different motion stimuli. We calculated the proportion of explained variance (PEV) of each neuron's neuronal activity and synaptic efficacy throughout the trial. The PEV measures the amount of explained variance a linear model relates neuronal activity or synaptic efficacy to the motion direction. Thus, higher PEV means a stronger relationship between the neuronal activity or synaptic efficacy. We defined neurons with PEV values larger than 0.25 at the end of the long delay as robustly encoding the motion stimulus. Much more neurons were found encoding for a single stimulus by synaptic efficacy than by neuronal activity [Fig. 6].

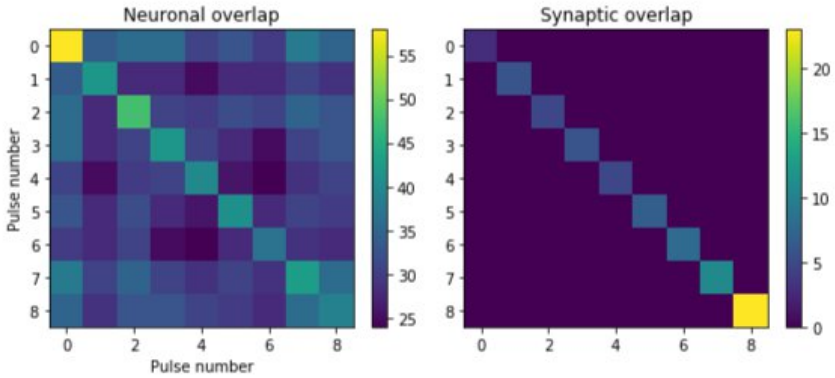

Figure 6. Specificity in neuronal and synaptic encoding
Above are heat maps illustrating the number of hidden neurons that are responsible for encoding two different motion stimuli (one indicated on the x axis and another indicated on the y axis). The two heat maps were generated by counting the number of neurons that have PEV values larger than 0.25, calculated from neuronal activity (left) and synaptic efficacy (right) respectively at the end of the long delay period.

Potential mechanism for sparse encoding

To understand the mechanisms enabling the sparse encoding of each stimulus, we selected five neurons that were most responsible for encoding each stimulus based on the hidden neurons' synaptic PEV measured at the end of the long delay period. Then we calculated the currents into these neurons following the formula: presynaptic neurons' neural activities * corresponding connection weights * synaptic efficacy. These currents were separated into activity from excitatory (EXC) neurons, inhibitory (INH) neurons, motion tuned (stimulus) neurons, and cue neurons. We were able to identify a neuronal circuitry in which the neurons responsible for encoding each stimulus had high excitatory input soon followed by a strong inhibitory current. Having a strong inhibitory current inhibits any neuronal activity in these neurons, keeping the information in short-term synaptic plasticity intact. This mechanism could explain how these neurons are minimizing the interference from subsequent or previous sample stimulus and allow them to specifically encode one stimulus [Fig.7].

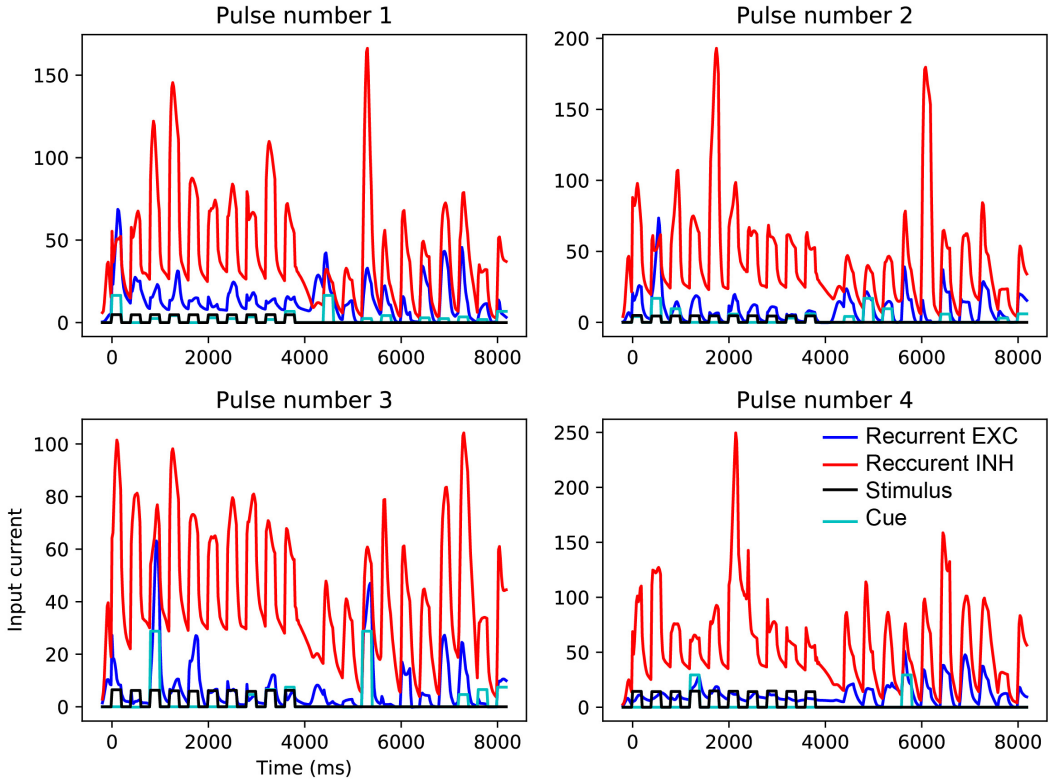

Figure 7. Current within hidden neurons for each motion stimulus
The currents flowing into the top five neurons responsible for encoding each motion stimulus were calculated. The figure shows input currents to the top five neurons of each pulse stimulus based on the source of the incoming activities: excitatory (EXC) neurons, inhibitory (INH) neurons, motion tuned (stimulus) neurons, and cue neurons.

Observation of primacy and recency effect

While training recurrent neural networks to encode multiple sequentially presented items in their memory, we observed an interesting phenomenon in which the accuracy for the first and last couple items were higher than the ones that were presented in the middle [Fig.8]. This phenomenon is analogous to the “primacy effect” (remembering items at the beginning of the list better compared to the rest) and “recency effect” (remembering items that are presented near or at the end of the list better than the ones before) which are often observed when people are asked to freely recall items from a sequence of items (Murdock, 1962 & Mayo, 1964). A potential reason for the primacy and recency effects could be that the network has a more difficult time projecting strong inhibitory signal to the hidden neurons that allow them to encode items in the middle of a sequence. With less inhibitory signal, the neurons encoding for the stimuli in the middle of the sequence would suffer from more interference from other stimuli as their neuronal and synaptic activity changes with a presentation of other stimuli.

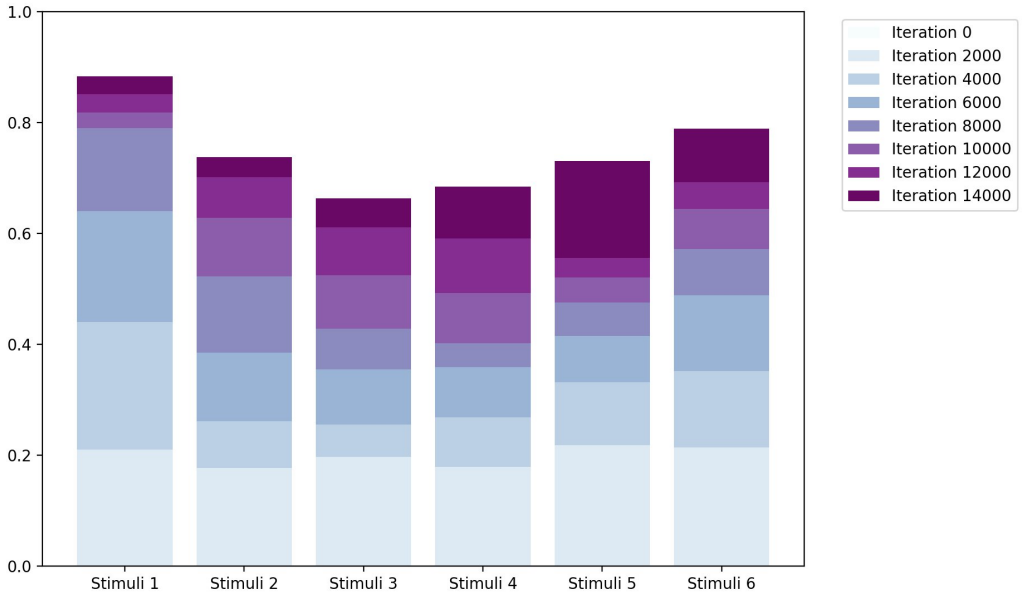

Figure 8. Recall accuracy for each stimulus showing primacy and recency effects
In a task with 6 motion stimuli, the network showed better performance for the first and last few stimuli compared to the ones in the middle of the sequence. This figure illustrates such trend over multiple training iterations.

Discussion

Our results suggest that the network may be maintaining information for multiple sequentially presented stimuli by sparsely encoding information for each stimulus in the hidden neurons' synaptic efficacy. With the understanding of how the network is using excitatory and inhibitory currents to gate what information is getting into each neuron to sparsely encode and minimize interference, we may be able to design an algorithm that further strengthens such gating mechanism to enhance the network's ability to encode multiple items for a short period of time. This could potentially alleviate the model's lack of ability to learn items presented in the middle of a sequence as well as learn to maintain more items. Such enhancements in working memory capacity would be beneficial when training a network to solve tasks that requires the network to keep track of multiple moving parts of a task, similar to what we need to do in a lot of real life situations.

References

Chafee, M. V. & Goldman-Rakic, P. S. Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memory task. J. Neurophysiol. 79, 2919–40 (1998).
Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic Coding of Visual Space in the Monkey's Dorsolateral Prefrontal Cortex. JOURNALOFNEUROPHYSIOLOGY 6, (1989).
Hawkins, J. & Ahmad, S. Why Neurons Have Thousands of Synapses, a Theory of Sequence 1936 Memory in Neocortex. Front. Neural Circuits 10, 23 (2016).
Laughlin, S. B., de Ruyter van Steveninck, R. R. & Anderson, J. C. The metabolic cost of neural 1934 information. Nat. Neurosci. 1, 36–41 (1998).
Masse, N. Y., Yang, G. R., Song, H. F., Wang, X., & Freedman, D. J. (2018). Circuit mechanisms for the maintenance and manipulation of information in working memory. doi:10.1101/305714
Mayo, C. W., & Crockett, W. H. (1964). Cognitive complexity and primacy-recency effects in impression formation. *The Journal of Abnormal and Social Psychology*, *68*(3), 335.
Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science*, *319*(5869), 1543-1546.
Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of experimental psychology*, *64*(5), 482.
Olshausen, B. & Field, D. Sparse coding of sensory inputs. Curr. Opin. Neurobiol. 14, 481– 1932 487 (2004).