bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Single-exposure visual memory judgments are reflected in IT cortex

Travis Meyer and Nicole C. Rust
Department of Psychology, University of Pennsylvania

Abstract

Our visual memory percepts of whether we have encountered specific objects or scenes before are hypothesized to manifest as decrements in neural responses in inferotemporal cortex (IT) with stimulus repetition. To evaluate this proposal, we recorded IT neural responses as two monkeys performed a single-exposure visual memory task designed to measure the rates of forgetting with time. We found that a strict interpretation of the repetition suppression hypothesis could not account for the monkeys' behavior, however, a weighted linear read-out of the IT population response accurately predicted forgetting rates, reaction time patterns, and individual differences in task performance. Additionally, the linear weights were largely all the same-sign and consistent with repetition suppression. These results suggest that a subpopulation of IT neurons reflect behaviorally-relevant memory information via changes in overall population response magnitudes, in contrast to visual identify information, which is largely reflected by changes in IT population response patterns.

Introduction

The everyday act of viewing the things around us leaves us with memories of the things that we have encountered. Under the right conditions, this type of "visual recognition memory" can be quite remarkable. For example, after viewing thousands of images, each only once and only for a few seconds, we can determine with high accuracy the specific images that we have viewed (Brady et al., 2008; Standing, 1973). Additionally, we can remember not just the objects that we've seen, but also the specific configurations and contexts we saw them in (Brady et al., 2008), suggesting that our brains store these memories with considerable visual detail. Where and how are visual memories stored and where and how is the percept of visual memory signaled?

One candidate mechanism for signaling visual memory percepts is the adaptation-like response reduction that occurs in high-level visual brain areas with stimulus repetition, known as "repetition suppression" (Fahy et al., 1993; Li et al., 1993; Miller and Desimone, 1994; Riches et al., 1991; Xiang and Brown, 1998). Consistent with that proposal, individual viewings of a novel image produce response reductions in inferotemporal cortex (IT) that can last tens of minutes to hours (Fahy et al., 1993; Xiang and Brown, 1998). Signaling visual memories in this way is attractive from a computational perspective, as it could explain how IT supports visual identity and visual memory representations within the same network. That is, visual representations of different images are thought to be reflected as distinct patterns of spikes across the IT population (Fig1; DiCarlo et al., 2012; Hung et al., 2005), which translates into a population representation in which visual information is reflected by the population vector angle (Fig 1). If it were the case that visual recognition memories were reflected by changes in the total numbers of spikes or equivalently population response vector length, this would minimize interference when superimposing visual memories and visual identity representations within the same network (Fig 1).

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

While attractive, there are also reasons to question whether visual memory percepts manifest purely as repetition suppression in IT cortex. For example, following many repeated image exposures (e.g. hundreds to thousands), IT neurons exhibit tuning sharpening (Anderson et al., 2008; Freedman et al., 2006), and a subset of neurons reflect tuning peak enhancement (Lim et al., 2015; Woloszyn and Sheinberg, 2012), and these changes could happen during single-exposure memory as well. Similarly, in the case of highly familiar images, neurons in a brain area that lie beyond IT, perirhinal cortex, are reported to signal familiarity with increases (as opposed to decreases) in firing rate (Tamura et al., 2017) and highly familiar faces produce larger perirhinal fMRI BOLD responses as compared to faces that are unfamiliar (Landi and Freiwald, 2017). In humans, tests of the hypothesis that limited-exposure visual memory percepts are supported by repetition suppression signals have produced mixed results, with some studies providing support (Gonsalves et al., 2005; Turk-Browne et al., 2006) and others refuting the hypothesis (Ward et al., 2013; Xue et al., 2011). Additionally, studies have implicated factors beyond overall response strength in limited-exposure familiarity, including the repeatability of human fMRI response patterns across exposures (LaRocque et al., 2013; Xue et al., 2010) and synchronization between gamma band oscillations and spikes in monkey hippocampus (Jutras et al., 2013). Notably, while a number of studies have investigated limited-exposure repetition suppression effects in IT at the resolution of individual-units (De Baene and Vogels, 2010; Li et al., 1993; McMahon and Olson, 2007; Ringo, 1996; Sawamura et al., 2006; Verhoef et al., 2008; Xiang and Brown, 1998), no study to date has attempted to determine whether these putative visual memory signals can in fact account for visual memory behaviors.

To evaluate the hypothesis that repetition suppression in IT accounts for familiarity judgments during a visual memory task, we trained two monkeys to view images and report whether they were novel (had never been seen before) or were familiar (had been seen exactly once), across a range of delays between novel and familiar presentations ('n-back'). To explore the IT representation of visual memory on both correct and error trials, we parameterized the task such that visual memories were remembered over a timescale of minutes within experimental sessions that lasted approximately one hour. We found that while both monkeys displayed characteristic forgetting functions and reaction time patterns, these behavioral patterns were not well-predicted by a strict interpretation of the repetition suppression hypothesis. However, these behavioral patterns were well-accounted for by a cross-validated linear read-out that weighted each neuron proportional to the amount of visual memory information it carried.

Results

The single-exposure visual memory task:

While compelling, the robustness with which visual memories are stored also presents a challenge to investigating their underlying neural correlates. Ideally, investigations of the neural signals supporting a behavior are made in a context where a task is parametrically varied from easy-to-challenging, and one can evaluate the degree to which behavioral sensitivities and behavioral confusions are reflected in neural responses (Parker and Newsome, 1998). Following on visual recognition memory studies that demonstrate a relationship between the time that images are viewed and how well they are remembered (Brady et al., 2009; Potter and Levy, 1969), we increased task difficulty by reducing image viewing time from the 2-3 seconds used in the classic human visual recognition memory studies to 400 ms. To titrate task difficulty within this regime, we explored a range of delays between novel and repeated presentations.

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In these experiments, two monkeys performed a task in which they viewed images and indicated whether they were novel or familiar with an eye movement response. Monkeys initiated each trial by fixating a point at the center of the screen, and this was followed by a brief delay and then the presentation of an image (Fig 2a). After 400 ms of fixating the image, a go cue appeared, indicating that the monkeys were free to make their selection via a saccade to one of two response targets (Fig 2a). Correct responses were rewarded with juice. While the first image presented in each session was always novel, the probability of subsequent images being novel versus familiar quickly converged to 50%. Novel images were defined as those that the monkeys had never viewed before (in the entire history of training and testing) whereas familiar images were those that had been presented only once, and earlier in the same session. Delays between novel and familiar presentations (Fig 2b) were pseudorandomly selected from a uniform distribution, in powers of two (n-back = 1, 2, 4, 8, 16, 32 and 64 trials corresponding to mean delays of 4.5s, 9s, 18s, 36s, 1.2 min, 2.4 min, and 4.8 min, respectively).

The monkeys' performance on this task was systematic, as illustrated by smoothly declining "forgetting functions", plotted as the proportion of trials that images were reported familiar as a function of n-back (i.e. the number of trials between novel and familiar presentations; Fig 3a,c). When familiar images were immediately repeated (n-back = 1), both monkeys most often called them familiar (proportion chose familiar = 0.98 and 0.94; Fig 3a,c). Similarly, when images were novel, monkeys were unlikely to call them familiar (proportion chose familiar = 0.13 and 0.07; Fig 3a,c). Between these two extremes, the proportion of familiar reports systematically decreased as a function of n-back (Fig 3a,c). In monkey 1, performance at 32 and 64 back fell below chance (32-back = 0.46, 64-back = 0.27, chance = 0.50), indicating that this animal most often reported that familiar images repeated after these longer delays were novel (Fig 3a). In monkey 2, performance at 32 and 64 back remained above chance (32-back = 0.76, 64-back = 0.54), indicating higher performance in this animal as compared to monkey 1 (Fig 3c).

We also analyzed reaction times for novel and familiar trials, parsed by correct and error trial outcomes. We found that mean reaction times on correctly reported familiar trials systematically increased as a function of n-back, or equivalently, reaction times on correct trials increased with increasing difficulty (Fig 3b,d red). Conversely, reaction times on error trials decreased as a function of n-back, or equivalently, reaction times on error trials decreased with increasing difficulty (Fig 3b,d, cyan). In both animals, this led to an x-shaped pattern in the mean reaction times on familiar trials when plotted as a function of n-back. On novel trials, reaction times mimicked the pattern observed for the low n-back familiar cases in that reaction times were faster on correct as compared to error trials (Fig 3b,d).

From what underlying process might these x-shaped reaction time patterns arise? As is the case for nearly any task, behavioral performance can be thought of as the outcome of passing a signal (in this case a memory signal) through a decision process. The x-shaped patterns that we observe differ from the patterns reported for tasks that are well-accounted for by standard drift diffusion models of decision making, such as the dot-motion-direction task (Gold and Shadlen, 2007), as these models predict reaction times that are matched on correct and error trials and that both increase as task performance decreases. While it is the case that extensions to this framework can predict reaction time asymmetries (Ratcliff and McKoon, 2008), additional parameters are required for it to do so, and these additions make it less well-suited for the purposes of this study (focused on evaluating the plausibility that IT visual memory signals can quantitatively account for visual memory behavior). We have, however, determined that these

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

x-shaped reaction time patterns can be captured by a very simple, low-parameter extension to the signal detection theory framework, as proposed by "strength theory" (Murdock, 1985; Norman and Wickelgren, 1969). Like signal detection theory, strength theory proposes that a noisy internal variable ("memory strength") is compared to a criterion to determine whether an image is novel or familiar (Fig 4, left). Strength theory also predicts that during a visual memory task, reaction times will be inversely related to the distance of this variable from the criterion, loosely analogous to a process in which increased certainty produces faster responses (Fig 4, middle). This leads to the qualitative prediction that when images are repeated with short n-back, memories are strong, and this will produce reaction times that are faster on correct as compared to error trials (Fig 4, red vs. blue). In contrast, at long n-back, memories are weak, and this can produce reaction times that are slower on correct as compared to error trials (Fig 4, green vs. purple). The combined consequence of strong and weak memories is an x-shaped pattern.

In sum, the reproducible patterns reflected in both monkeys' forgetting functions, along with their reaction time patterns, place non-trivial constraints on the candidate neural signals that account for single-exposure visual memory behavior. The x-shaped patterns of reaction times that we observe cannot be account for by a standard drift diffusion process, but they can, in principle, be captured by the predictions of strength theory. However, a successful description of the neural signals supporting single-exposure visual memory behavior requires identifying a neural signal whose sensitivity to the elapsed time between initial and repeated presentations of an image matches the sensitivity reflected in the monkeys' behavior.

Single-exposure visual memory signals in IT cortex:

As monkeys performed this task, we recorded neural responses from IT using multi-channel probes acutely lowered before the beginning of each session. For quality control, recording sessions were screened based on their neural recording stability across the session, their numbers of visually responsive units, and the numbers of behavioral trials completed (see Methods). The resulting data set included 15 sessions for monkey 1 (n=403 units), and 12 sessions for monkey 2 (n= 396 units). Both monkeys performed many hundreds of trials during each session (~600-1000, corresponding to ~300-500 images each repeated twice). The data reported here correspond to the subset of images for which the monkeys' behavioral reports were recorded for both novel and familiar presentations (e.g. trials in which the monkeys did not prematurely breaking fixation during either the novel or the familiar presentation of an image).

We began by considering the proposal that the signals differentiating novel versus familiar presentations of images were systematically reflected as response decrements with stimulus repetition (i.e. "repetition suppression"). As a first, simple illustration of the strength of these putative single-exposure memory signals, shown in Figure 5 is a plot of the grand mean firing rate response of all 799 units parsed by n-back, plotted as a function of time relative to stimulus onset (Fig 5a). This plot reveals a fairly systematic decrement in the response with repetition that diminishes with time since the novel presentation. We quantified the magnitude of suppression as the decrement in the area under each n-back trace relative to the novel trace, computed 150 – 400 ms after stimulus onset (Fig 5b). Immediate stimulus repetition resulted in a decrement in the response of ~11% and suppression magnitudes decreased systematically with n-back.

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

As described above, we were interested in evaluating the plausibility that IT repetition suppression signals could account for the monkeys' behavioral reports, including both the sensitivity of their forgetting functions as well as their reaction time patterns. A comparison of this measure of suppression with n-back (Fig 5b) with the monkeys' behavioral forgetting functions (Fig 2a,c, left) reveals that these two do not in fact align. For example, a small amount of enhancement (as opposed to suppression) was observed for familiar images presented at 64 back whereas both monkeys were successful at distinguishing 64 back from novel images, on at least a fraction of the trials (Fig 2a,c left). Examination of the timecourse of the response (Fig 5a) confirms that suppression did not exist for the 64-back case during any portion of the 400 ms viewing period (Fig 5a, compare purple versus black). We also confirmed that even after applying stringent screening criteria to select units with the largest stimulus-evoked responses (relative to baseline), the discrepancy between behavioral reports and neural signals at 64-back remained, suggesting that this discrepancy is not due the noisy responses of units that are only weakly modulated or are unmodulated (not shown). These results suggest that our behavioral and neural data cannot be aligned via a population read-out that differentiates novel versus familiar responses based on repetition suppression in the strictest sense (i.e. the total or average numbers of spikes across the population). However, it leaves open the possibility that the monkeys' behavioral reports might arise via a more traditional weighted linear read-out in which IT neurons are weighted proportional to the amount of task-relevant information that they carry.

A weighted linear read-out of IT accurately predicts behavior:

As described above, we were interested in determining the degree to which a weighted linear read-out of IT could accurately predict the monkeys' behavior, including the sensitivity of their forgetting functions to n-back and the patterns observed in their reaction times (Fig 3). As an overview of our approach, we computed neural predictions of behavioral forgetting functions by assessing performance of a linear decoder applied to IT pseudopopulation data, and we computed neural predictions for reaction time patterns by applying the simple extensions to these approaches proposed by strength theory (Fig 4). We begin by demonstrating our results for the data pooled across the two subjects, followed by its application to each monkey individually.

Because the neural data collected in any individual recording session had too few units to fully account for the monkeys' behavior (e.g. near 100% correct for 1-back familiar images), we concatenated units across sessions to create a larger pseudopopulation of 799 units, where responses were quantified 150-400 ms following stimulus onset. When creating this pseudopopulation, we aligned data across sessions in a manner that preserved whether the trials were presented as novel or familiar as well as their n-back separation (see Methods). Because different images were used in each session, aligning images this way implicitly assumes that the total numbers of stimulus-evoked spikes are matched across different images and that the responses of the units recorded in different sessions are uncorrelated.

To compute neural predictions for behavioral forgetting functions, we began by training a linear decoder to discriminate the same images presented as novel versus as familiar (Fig 6a) using the data corresponding to all n-backs simultaneously. The specific linear decoder we used, a Fisher Linear Discriminant (FLD), assigned a weight to each neuron proportional to the amount of linearly-separable visual memory information reflected in its responses (i.e. it's d'; see Methods), and a single criterion value to parse the combined, weighted population responses for

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

novel versus familiar predictions. A final parameter specified the size of the IT population under consideration. Shown in Fig 6b are the neural estimates of distributions of memory signal strength at each n-back, computed across many iterations of the cross-validated linear classifier training and testing procedure for the best sized population (n=799 units, described below). As expected, we found that the weighted population response strengths were largest for novel images (Fig 6b, black) and were weakest for familiar images presented as immediate repeats (Fig 6b, red). Between these two extremes, we observed a continuum of strengths loosely organized according by n-back (Fig 6b, rainbow). Finally, a neural prediction for the forgetting function was computed as the fraction of each distribution that fell on the "familiar" side of the criterion differentiating novel versus familiar predictions (Fig 6c). This analysis revealed a high degree of alignment between the neural prediction at each n-back and behavior, including high performance for familiar images presented at low n-back, performance at mid-range n-back that fell off with a similar sensitivity, and performance at the longest n-back (64) that fell below chance (Fig 6c). Similarly, neural predictions for novel images were well-aligned with the monkeys' behavioral reports (Fig 6c, "N").

To produce neural predictions for reaction times, we turned to strength theory (Fig 4). Shown in Fig 6d is the first step required for making those predictions: a plot of the neural predictions for the proportions of "correct" and "error" trials, plotted as a function of n-back. Note that the correct predictions simply replicate the forgetting function shown in Fig 6c, and the error predictions are simply those same values, subtracted from 1. While these plots directly follow from Fig 6c, we include them to illustrate that they qualitatively reflected an inverted version of the monkeys' behavioral reaction time plots, including an x-shaped pattern. To determine the degree to which these qualitative relationships quantitatively predict the monkeys' reaction times, we examined the relationship between the proportions plotted in Fig 6d and the monkeys' mean reaction times, and found it to be approximately linear (Fig 6e). We thus fit a two parameter linear function to convert the neural predictions of these proportions into reaction times (Fig 6e, black line). The resulting neural predictions were largely aligned with the monkeys' mean reaction times (Fig 6f), including increasing reaction times as a function of n-back on correctly reported familiar trials, decreasing reaction times as a function of n-back on familiar trials in which the monkeys' made errors, and the characteristic x-shaped pattern. Additionally, shorter mean reaction times for novel images on correct versus error trials were largely well-predicted by the neural data.

One important step in the procedure, not described above, involved determining the appropriate IT population size for making neural and behavioral comparisons. Because there really wasn't a way to do this *a priori*, we applied a fitting approach in which we computed the mean squared error between the actual forgetting functions and their neural predictions at a range of population sizes, including simulated extensions of our population up to sizes 50% larger than the maximal size we recorded (see Methods; Fig 6g). When too few units were included in the population, the neural prediction of the forgetting function was too flat and underestimated high performance at low n-back (Fig 6g, left inset). When too many units were included in the population, the neural prediction of the forgetting function was over-saturated and fell off with n-back with a slope that was too steep (Fig 6g, right inset). Of interest was the question of whether a global alignment of behavioral and neural sensitivity produced accurate neural predictions of the shape for forgetting function with n-back. The best population size fell near the maximal size of the total number of units that we recorded (Fig 6g, red dot).

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

As a final analysis for the pooled data, we were interested in better understanding whether the signs of the linear weights were consistently positive across units or whether IT reflected mixtures of positive and negative weights. The former scenario would be indicative of an IT population in which visual memories are reflected by spike numbers or equivalently population vector length (albeit only in a subpopulation of units), whereas the latter scenario would be more indicative of an IT memory representation that is reflected by population spike patterns, or equivalently vector angle (Fig 1). As a proxy for these weights, Fig 6h shows a histogram of d' values across the units in our recorded population. Unit d' was clearly skewed toward positive values, suggestive of a noisy IT visual memory representation that is largely reflected by the numbers of spikes evoked in a memory-modulated IT subpopulation.

IT single-exposure visual memory signals predict individual differences:

While the behavioral patterns were qualitatively similar for the two monkeys that participated in these experiments, quantitatively, monkey 2 performed better at the task than monkey 1 (Fig 3). Concomitantly, the strength of repetition suppression, computed as described in Fig 5b but separately for the units recorded in each monkey, was larger in the monkey that was better at the task (Fig 7a). We thus sought to determine whether a linear read-out of the IT data could accurately predict each monkey's individual behavioral response patterns.

Fig 7b (black) illustrates how the error between the neural predictions and forgetting functions changes as a function of population size for each animal. These plots demonstrate that a population of ~800 units in monkey 1 and ~500 units in monkey 2 most accurately accounted for each monkey's behavioral reports. This difference in population size is interesting in light of the fact that the monkey that required more units to account for behavioral performance was also the one that was worse at the task. To assess the degree to which our analysis had the sensitivity to distinguish individual differences, we also attempted to predict each monkey's behavioral forgetting functions from the other monkey's neural data (Fig 7b, gray). For both monkeys, the minimal error (as a function of population size) was lower when behavioral and neural data came from the same monkey as compared to when they were mixed between monkeys. These results support the hypothesis that non-trivial individual differences in behavioral performance are predicted by individual differences in IT visual memory signals.

Finally, Fig 7c-d compares each monkey's neural predictions for the best-sized population with their behavioral responses. In the case of both monkeys, these were generally well-aligned, including higher predicted performance for monkey 2 as compared to monkey 1 (e.g. at 64-back). Together, these results suggest that in addition to robust predictions of subject averaged behavior, individual differences in the ability to remember images are reflected by the sensitivities of memory signals in IT cortex.

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Discussion

This study was designed to test the hypothesis that the signals supporting single-exposure visual recognition memories, or equivalently answers to the question, "Have I seen that image before?", are reflected as decrements in the responses of neurons in IT with stimulus repetition (Fahy et al., 1993; Li et al., 1993; Miller and Desimone, 1994; Riches et al., 1991; Xiang and Brown, 1998). Prior to this study, this hypothesis had received mixed support from human fMRI studies (Gonsalves et al., 2005; Turk-Browne et al., 2006; Ward et al., 2013; Xue et al., 2011) and was largely untested at the resolution of individual neurons. We found that a strict interpretation of the repetition suppression hypothesis in the form of counting the total numbers of spikes across the IT population could not account for single-exposure visual memory behavior, however, a weighted linear read-out of IT responses accurately predicted 1) the rates of forgetting as a function of time elapsed between novel and repeated image presentations (Fig 6c), 2) mean reaction time patterns parsed by correct versus error trials (Fig 6f), and 3) individual differences in task performance (Fig 7). Additionally, linear weights were largely the same-sign and positive (Fig 6h), suggesting that behaviorally-relevant single-exposure visual memory information is in fact signaled by repetition suppression, but only by a subset of IT neurons.

In the neural coding literature, sensory variables are typically regarded as being one of two types: those that produce bell-shaped tuning curves (e.g. the orientation of a grating or the frequency of a sound) versus those that produce monotonically-increasing tuning curves that do not decrease again (e.g. contrast or luminance). This distinction is important, as this qualitative difference in how stimulus variables are encoded has important implications for how they must be decoded downstream: bell-shaped representations must be read-out by considering population spike patterns, whereas monotonically-increasing representations must be read-out by considering population spike numbers. In IT, representations of visual image and object identity are thought to largely arise from bell-shaped tuning curves, and this translates into a population representation of identity that is reflected by population vector angle (Fig 1). Our study sought to test the hypothesis that IT also reflects information about visual memory, but via monotonically-increasing tuning curves, or equivalently a population representation of memory that is reflected by population vector length (Fig 1). We found that while the simplest form of this hypothesis (where total spike numbers in IT are used to decode memory) could not account for the monkeys' behavior, a largely positively-weighted linear read-out of the IT population could. These results are thus consistent with an IT representation of visual memory that is largely reflected by population vector length, but within a subpopulation of neurons or equivalently, a linear subspace.

While changes in the IT representations of images after many image exposures (e.g. hundreds-to-thousands) has been explored previously at the resolution of individual units (Anderson et al., 2008; Baker et al., 2002; Freedman et al., 2006; Lim et al., 2015; Meyer et al., 2014; Woloszyn and Sheinberg, 2012), changes in IT after a limited number of exposures have received less focus (Fahy et al., 1993; Li et al., 1993; Riches et al., 1991; Xiang and Brown, 1998).This is presumably because traditional neurophysiology approaches rely on collecting the repeated responses to the same condition across many repeated trials whereas an image can only be "novel" on a single trial. In this study, we employed a population-based, decoding approach that formulated a single-exposure visual familiarity task as a two-way classification between novel versus familiar choices. This allowed us to pool neural responses across many units and many

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

images. We found that accounting for the monkeys' behavioral reports required a large number of units (500-800), presumably as a consequence of the small-sized signals that follow the single exposure to an image that was presented minutes ago. To create such a large population, we concatenated responses across different sessions into a pseudopopulation, and in that process implicitly assumed that 1) every image evoked the same total numbers of spikes in IT and 2) the responses of units recorded in different sessions were uncorrelated. Future work will be targeted at testing the validity of these assumptions.

In this study, we adjusted the task parameters such that images were forgotten over minutes within sessions that lasted approximately one hour. This included reducing the viewing time from the longer durations used in previous human behavioral experiments (2-3 seconds) to 400 ms. Our results suggest that forgetting rates are well-aligned between behavioral reports and IT neural signals within this regime. Will longer timescale memories be reflected by signals in IT as well? That remains to be seen. It could be the case that IT reflects single-exposure visual memories across all behaviorally-relevant timescales, alternatively, it could be the case that the signals reflecting single-exposure memories across longer timescales (e.g. hours and days) are only reflected in higher brain areas such as perirhinal cortex and/or the hippocampus. A highly related issue is the question of where and how single-exposure visual memories are stored in the brain. Crucially, it is important to recognize that it does not necessary follow from the fact that a particular brain area reflects a memory signal, that it must be the locus at which storage occurs. It is likely the case that the visual memory signals that we observe are at least partially the consequence of the cumulative adaptation-like processes that happen within IT and within brain areas preceding IT. What is less clear is whether these signals also reflect contributions from higher brain areas as well.

One consequence of our findings is that care needs to be taken when interpreting measures of single-exposure visual memory signals that pool across large numbers of neurons, such as the BOLD signal measured by human fMRI. Our results suggest that near-single unit resolution is required to determine how to appropriately weight IT units to account for single-exposure visual memory behaviors. In contrast, measures that average the responses across large numbers of neurons result in an information loss that cannot be recovered (e.g. via a multi-voxel pattern analysis). This difference in resolution could contribute, in part, to the discrepancy between our study and human fMRI studies that fail to find a relationship between repetition suppression magnitudes in high-level visual brain areas and explicit visual memory reports (Ward et al., 2013; Xue et al., 2011). Another factor that may contribute to differences between our results and those studies is a distinct difference in experimental design: our study correlates repetition suppression and behavioral reports on the same trial, whereas these studies correlate repetition suppression to a second viewing of an image with the behavioral report about remembering during a third viewing. The rationale behind the fMRI design is a desire to dissociate memory processes from the processes involved in decision making and response execution. In our study, we were focused on evaluating the plausibility that the signal supporting visual memory behavioral reports is reflected in IT cortex, as opposed to the plausibility that memory signals are reflected in IT in the absence of a subject being engaged in a memory task. The consistent (positive) sign of the linear weights recovered across IT units suggests that our results cannot be accounted for by motor responses, as the task required the monkeys to saccade to two different targets to report novel versus familiar predictions and a motor account would require that all the IT neurons were tuned for the same target (e.g. "upward" for monkey 1 and "downward" for monkey 2). Finally, differences between our study and those reports could also arise from differences between species, analogous to the differences reported between monkey

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

IT and human LOC for changes in the representations of highly familiar images as measured with fMRI (Op de Beeck et al., 2006; Op de Beeck et al., 2008).

Analysis of our reaction time patterns parsed by trial outcome (correct/error) revealed a characteristic x-shaped pattern (Fig 3) at odds with the predictions of standard models of decision making such as standard instantiations of the drift diffusion model. Extensions of the drift diffusion framework have been proposed in which reaction time asymmetries on correct versus error trials can be accounted for by adding per-trial noise in the decision variable drift rate or the decision variable start position (Ratcliff and McKoon, 2008). Our task was not designed to differentiate between these and other similar models, but rather to test the hypothesis that signals reflecting single-exposure visual memories are found in IT cortex. As such, we opted for the much simpler, lower-parameter description suggested by strength theory (Murdock, 1985; Norman and Wickelgren, 1969). The inverted relationship between proportion correct and reaction time captured by strength theory can loosely be thought of as a signature of confidence (e.g. when performance is higher, reaction times are faster), however, the drawback of strength theory is that it lends little biophysical insight into how this process might happen in the brain. Our study provides important constraints on models of the decision making process for single-exposure memory tasks, and should constrain future work in which this process is investigated more completely.

Acknowledgements:

We thank Alex Smolyanskaya for her contributions to early phases of this work. This work was supported by the National Eye Institute of the National Institutes of Health (award R01EY020851), the Simons Foundation (through an award from the Simons Collaboration on the Global Brain), and the McKnight Endowment for Neuroscience.

Competing Interests:

N.C.R. serves as a Board Review Editor for eLife.

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Methods:

Experiments were performed on two adult male rhesus macaque monkeys (*Macaca mulatta)* with implanted head posts and recording chambers. All procedures were performed in accordance with the guidelines of the University of Pennsylvania Institutional Animal Care and Use Committee.

The single-exposure visual memory task:

All behavioral training and testing was performed using standard operant conditioning (juice reward), head stabilization, and high-accuracy, infrared video eye tracking. Stimuli were presented on an LCD monitor with an 85 Hz refresh rate using customized software (http://mworks-project.org).

As an overview of the monkeys' task, each trial involved viewing one image for 400 ms and indicating whether it was novel, (never seen before) or familiar (seen exactly once prior) with an eye movement to one of two response targets. Images were never presented more than twice (once as novel and then as familiar) during the entire training and testing period of the experiment. Trials were initiated by the monkey fixating on a red square (0.25°) on the center of a gray screen, within a square window of ±1.5°, followed by a 200 ms delay before a 4° stimulus appeared. The monkeys had to maintain fixation of the stimulus for 400 ms, at which time the red square turned green (go cue) and the monkey made a saccade to the target indicating that the stimulus was novel or familiar. In monkey 1, response targets appear at stimulus onset; in monkey 2, response targets appeared at the time of the go cue. In both cases, targets were positioned 8° above or below the stimulus. The association between the target (up vs. down) and the report (novel vs. familiar) was swapped between the two animals.

The specific random sequence of images presented during each session was generated offline before the start of the session. The primary goal in generating the sequence was to select trial locations for novel images and their repeats with a uniform distribution of n-back (where n-back = 1, 2, 4, 8, 16, 32 and 64). This was achieved by constructing a sequence slightly longer than what was anticipated to be needed for the day, and by iteratively populating the sequence with novel images and their repeats at positions selected from all the possibilities that remained unfilled. Because the longest n-back locations (64) were the most difficult to fill, a fixed number of those were inserted first, followed by systematically working through the insertion of the same fixed number at each consecutively shorter n-back (32, 16 …). In the relatively rare cases that the algorithm could not produce that fixed number at each n-back, it was restarted. The result was a partially populated sequence in which 83% of the trials were occupied. Next, the remaining 17% of trials were examined to determine whether they could be filled with novel/familiar pairs from the list of n-back options (64, 32, 16-back …). For the very small number of trials that remained after all possibilities had been extinguished (e.g. a 3-back scenario), these were filled with 'off n-back' novel/familiar image pairs and these trials were disregarded from later analyses.

"Forgetting functions" (Fig 3a,c and Fig 6c) were computed as the mean proportion of trials each monkey selected the familiar target, across all trials and all sessions. Because behavioral outcome is a binary variable, error was estimated by computing the mean performance trace for each session, and then computing the 97.5% confidence interval as 2.2*standard error of those

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

traces. Mean reaction times (Fig 3b,d and Fig 6f) were computed as means across all trials and sessions, and 97.5% confidence intervals were computed as 2.2*standard error of those same values.

Neural recording:

The activity of neurons in IT was recorded via a single recording chamber in each monkey. Chamber placement was guided by anatomical magnetic resonance images in both monkeys. The region of IT recorded was located on the ventral surface of the brain, over an area that spanned 5 mm lateral to the anterior middle temporal sulcus and 14-17 mm anterior to the ear canals. Recording sessions began after the monkeys were fully trained on the task and after the depth and extent of IT was mapped within the recording chamber. Combined recording and behavioral training sessions happened 5 times per week across a span of 5 weeks (monkey 1) and 4 weeks (monkey 2). Neural activity was largely recorded with 24-channel U-probes (Plexon, Inc) with linearly arranged recording sites spaced with 100 μm intervals. Continuous, wideband neural signals were amplified, digitized at 40 kHz and stored using the Grapevine Data Acquisition System (Ripple, Inc.). Spike sorting was done manually offline (Plexon Offline Sorter). At least one candidate unit was identified on each recording channel, and 2-3 units were occasionally identified on the same channel. Spike sorting was performed blind to any experimental conditions to avoid bias. A multi-channel recording session was included in the analysis if: (1) the recording session was stable, quantified as the grand mean firing rate across channels changing less than 2-fold across the session; (2) over 50% of neurons were visually responsive (a loose criterion based on our previous experience in IT), assessed by a visual inspection of rasters; and (3) the number of successfully completed novel/familiar pairs of trials exceeded 100. The sample size (number of successful sessions recorded) was chosen to approximately match our previous work (Pagan et al., 2013).

Neural predictions of behavioral performance:

Because the data recorded in any individual session (on 24 channels) corresponded to a population too small to provide a full account of behavioral performance, we combined data across sessions into a larger pseudopopulation. In doing so, trials were aligned across sessions in a manner that preserved whether images were presented as novel or familiar, as well as their n-back. The resulting pseudopopulation consisted of the responses of 799 neurons to 107 images presented as both novel and familiar (i.e. 15, 15, 16, 17, 17, 15 and 12 trials at 1, 2, 4, 8, 16, 32 and 64-back, respectively).

To measure performance of the IT pseudopopulation to perform the monkeys' task (Fig 6-7), we applied a Fisher Linear Discriminant (FLD; a linear decoder). The population response $\boldsymbol{x}$ was quantified as the vector of simultaneously recorded spike counts on a given trial, computed in window positioned 150-400 ms following stimulus onset. To ensure that the decoder did not erroneously rely on visual selectivity, the decoder was trained on pairs of novel/familiar trials in which monkeys viewed the same image (regardless of behavioral outcome and for all n-back simultaneously).

To compute the FLD parameters, we considered the general form of a linear decoding axis:

(1) $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b,$

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

where **w** is an N-dimensional vector (and N is the number of units) containing the linear weights applied to each unit, and b is a scalar value. We fit these parameters using an FLD, where the vector of linear weights was calculated as:

(2) $\boldsymbol{w} = \Sigma^{-\mathbf{1}}(\mu_1 - \mu_2)$

and b was calculated as:

(3) $b = \boldsymbol{w} \cdot \frac{1}{2}(\mu_1 + \mu_2) = \frac{1}{2}{\mu_1}^T \Sigma^{-\mathbf{1}} \mu_1 - \frac{1}{2}{\mu_2}^T \Sigma^{-\mathbf{1}} \mu_2$

Here $\mu_1 \; and \; \mu_2$ are the means of the two classes (target matches and distractors, respectively) and the mean covariance matrix is calculated as:

(4) $\Sigma = \frac{\Sigma_1 + \Sigma_2}{2}$

where Σ_1 and Σ_2 are the covariance matrices of the two classes with the off-diagonal entries set to zero. We set these terms to zero (as opposed to regularization) because we found that the off-diagonal terms were very poorly estimated for our data set. Calculating FLD weights in this manner is thus equivalent to weighting each unit by its d' alone (while ignoring any optimization that considers correlated activity between units).

We used the FLD analysis to measure the strength of the visual memory signal at each n-back using an iterative cross-validated linear decoding procedure that randomly assigned images to training and testing groups. On each iteration of the resampling procedure, the responses for each unit were randomly shuffled within the set of matched n-back to ensure that artificial correlations (e.g. between the neurons recorded in different sessions) were removed. Each iteration also involved setting aside the responses to 2 images at each n-back (presented as both novel and familiar, for 4 trials in total) for testing classifier performance. The remaining trials were used to train an FLD linear decoder (Eqs 1-4) to distinguish novel versus familiar images, where the novel and familiar classes included the data corresponding to all n-backs and all trial outcomes. Memory strength was measured as the dot product of the test data vectors $\boldsymbol{x}$ and the weights **w**, adjusted by *b* (Eq. 1). Histograms of these distributions are shown in Fig 6b across 1000 resampling iterations. A neural prediction of the proportion of trials on which "familiar" would be reported was computed as the proportion of each distribution that took on a value less than the criterion (Fig 6c & 7c). This process was repeated for a broad range of population sizes and for each size, the mean squared error between the actual and predicted forgetting functions were computed to determine the best sized population (Fig 6g & 7b).

To compute predictions of reaction times on correct and error trials, we began by considering the proportion of the distributions shown in Fig 7b predicted to be reported "correct" versus "wrong", as a function of n-back, for both novel and familiar presentations (Fig 6d). Examination of these proportions plotted against the monkeys' reaction times that they map onto revealed a linear relationship (Fig 6e), which we fit with a line by minimizing mean squared error. The final neural predictions for reaction times were produced by passing the predicted proportions for correct and error trials through the resulting linear equation (Fig 6f).

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Estimating performance for larger-sized populations:

Accounting for the pooled performance across the two monkeys required all of the units recorded, combined across the two animals (n=799). We were unable to accurately predict performance for either animal (e.g. near 100% correct for 1-back familiar) with the neural data recorded from that animal (~400 units; Fig 8b-c). To estimate performance for larger sized populations than those we recorded, we computed performance trajectories as a function of population size and quantified how the mean and standard deviation of the distributions (depicted in Fig 6b), as well as the value of the criterion, grew as a function of population size. Our results (not shown) revealed that the trajectories of the means and the criterion were highly linear as a function of population size, and the standard deviations saturated. Our estimates of performance for larger sized populations thus began by maintaining the standard deviations computed for the full recorded population while extending the linear trajectory means and linear trajectory criterion. We then computed performance for the larger sized populations after replacing the histograms in Fig 6b with Gaussians matched for those means and standard deviations, relative to the estimated criterion. The error trajectories for these extended populations are shown in Fig 6g & 7b (dashed lines).

Unit d':

Unit d' was calculated, for each unit, as the difference in the mean responses to the set of images presented as novel versus the set presented as familiar, divided by the average standard deviation across the two sets (Fig 6h).

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References:

Anderson, B., Mruczek, R.E., Kawasaki, K., and Sheinberg, D. (2008). Effects of familiarity on neural activity in monkey inferior temporal lobe. Cereb Cortex *18*, 2540-2552.

Baker, C.I., Behrmann, M., and Olson, C.R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. Nat Neurosci *5*, 1210-1216.

Brady, T.F., Konkle, T., Alvarez, G.A., and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. Proc Natl Acad Sci U S A *105*, 14325-14329.

Brady, T.F., Konkle, T., Oliva, A., and Alvarez, G.A. (2009). Detecting changes in real-world objects: The relationship between visual long-term memory and change blindness. Communicative & integrative biology *2*, 1-3.

De Baene, W., and Vogels, R. (2010). Effects of adaptation on the stimulus selectivity of macaque inferior temporal spiking activity and local field potentials. Cereb Cortex *20*, 2145-2165.

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

Fahy, F.L., Riches, I.P., and Brown, M.W. (1993). Neuronal activity related to visual recognition memory: long-term memory and the encoding of recency and familiarity information in the primate anterior and medial inferior temporal and rhinal cortex. Exp Brain Res *96*, 457-472.

Freedman, D.J., Riesenhuber, M., Poggio, T., and Miller, E.K. (2006). Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. Cerebral Cortex *16*, 1631-1644.

Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. Annu Rev Neurosci *30*, 535-574.

Gonsalves, B.D., Kahn, I., Curran, T., Norman, K.A., and Wagner, A.D. (2005). Memory strength and repetition suppression: multimodal imaging of medial temporal cortical contributions to recognition. Neuron *47*, 751-761.

Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863-866.

Jutras, M.J., Fries, P., and Buffalo, E.A. (2013). Oscillatory activity in the monkey hippocampus during visual exploration and memory formation. Proc Natl Acad Sci U S A *110*, 13144-13149.

Landi, S.M., and Freiwald, W.A. (2017). Two areas for familiar face recognition in the primate brain. Science *357*, 591-595.

LaRocque, K.F., Smith, M.E., Carr, V.A., Witthoft, N., Grill-Spector, K., and Wagner, A.D. (2013). Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. J Neurosci *33*, 5466-5474.

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

666 Li, L., Miller, E.K., and Desimone, R. (1993). The representation of stimulus familiarity in anterior
667 inferior temporal cortex. J Neurophysiol *69*, 1918-1929.

668 Lim, S., McKee, J.L., Woloszyn, L., Amit, Y., Freedman, D.J., Sheinberg, D.L., and Brunel, N.
669 (2015). Inferring learning rules from distributions of firing rates in cortical neurons. Nat Neurosci
670 *18*, 1804-1810.

671 McMahon, D.B., and Olson, C.R. (2007). Repetition suppression in monkey inferotemporal
672 cortex: relation to behavioral priming. J Neurophysiol *97*, 3532-3543.

673 Meyer, T., Walker, C., Cho, R.Y., and Olson, C.R. (2014). Image familiarization sharpens
674 response dynamics of neurons in inferotemporal cortex. Nat Neurosci *17*, 1388-1394.

675 Miller, E.K., and Desimone, R. (1994). Parallel Neuronal Mechanisms for Short-Term-Memory.
676 Science *263*, 520-522.

677 Murdock, B.B., Jr. (1985). An analysis of the strength-latency relationship. Mem Cognit *13*, 511-
678 521.

679 Norman, D.A., and Wickelgren, W.A. (1969). Strength theory of decision rules and latency in
680 retrieval from short-term memory. J Math Psychol *6*, 192-208.

681 Op de Beeck, H.P., Baker, C.I., DiCarlo, J.J., and Kanwisher, N.G. (2006). Discrimination
682 training alters object representations in human extrastriate cortex. J Neurosci *26*, 13025-13036.

683 Op de Beeck, H.P., Deutsch, J.A., Vanduffel, W., Kanwisher, N.G., and DiCarlo, J.J. (2008). A
684 stable topography of selectivity for unfamiliar shape classes in monkey inferior temporal cortex.
685 Cereb Cortex *18*, 1676-1694.

686 Pagan, M., L.S., U., M.P., W., and Rust, N.C. (2013). Signals in inferotemporal cortex and
687 perirhinal cortex suggest an untangling of visual target information. Nature Neuroscience *16*,
688 1132-1139.

689 Parker, A.J., and Newsome, W.T. (1998). Sense and the single neuron: probing the physiology
690 of perception. Annu Rev Neurosci *21*, 227-277.

691 Potter, M.C., and Levy, E.I. (1969). Recognition memory for a rapid sequence of pictures. J Exp
692 Psychol *81*, 10-15.

693 Ratcliff, R., and McKoon, G. (2008). The diffusion decision model: theory and data for two-
694 choice decision tasks. Neural Comput *20*, 873-922.

695 Riches, I.P., Wilson, F.A., and Brown, M.W. (1991). The effects of visual stimulation and
696 memory on neurons of the hippocampal formation and the neighboring parahippocampal gyrus
697 and inferior temporal cortex of the primate. J Neurosci *11*, 1763-1779.

698 Ringo, J.L. (1996). Stimulus specific adaptation in inferior temporal and medial temporal cortex
699 of the monkey. Behav Brain Res *76*, 191-197.

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

700 Sawamura, H., Orban, G.A., and Vogels, R. (2006). Selectivity of neuronal adaptation does not
701 match response selectivity: a single-cell study of the FMRI adaptation paradigm. Neuron *49*,
702 307-318.

703 Standing, L. (1973). Learning 10,000 pictures. Q J Exp Psychol *25*, 207-222.

704 Tamura, K., Takeda, M., Setsuie, R., Tsubota, T., Hirabayashi, T., Miyamoto, K., and Miyashita,
705 Y. (2017). Conversion of object identity to object-general semantic value in the primate temporal
706 cortex. Science *357*, 687-692.

707 Turk-Browne, N.B., Yi, D.J., and Chun, M.M. (2006). Linking implicit and explicit memory:
708 common encoding factors and shared representations. Neuron *49*, 917-927.

709 Verhoef, B.E., Kayaert, G., Franko, E., Vangeneugden, J., and Vogels, R. (2008). Stimulus
710 similarity-contingent neural adaptation can be time and cortical area dependent. J Neurosci *28*,
711 10631-10640.

712 Ward, E.J., Chun, M.M., and Kuhl, B.A. (2013). Repetition suppression and multi-voxel pattern
713 similarity differentially track implicit and explicit visual memory. J Neurosci *33*, 14749-14757.

714 Woloszyn, L., and Sheinberg, D.L. (2012). Effects of long-term visual experience on responses
715 of distinct classes of single units in inferior temporal cortex. Neuron *74*, 193-205.

716 Xiang, J.Z., and Brown, M.W. (1998). Differential neuronal encoding of novelty, familiarity and
717 recency in regions of the anterior temporal lobe. Neuropharmacology *37*, 657-676.

718 Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J.A., and Poldrack, R.A. (2010). Greater neural
719 pattern similarity across repetitions is associated with better memory. Science *330*, 97-101.

720 Xue, G., Mei, L., Chen, C., Lu, Z.L., Poldrack, R., and Dong, Q. (2011). Spaced learning
721 enhances subsequent recognition memory by reducing neural repetition suppression. J Cogn
722 Neurosci *23*, 1624-1633.
723

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

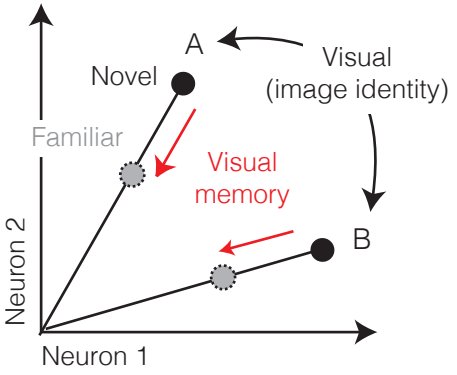

Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 1. *Multiplexing visual and visual memory representations*. Shown are the hypothetical population responses to two images (A and B), each presented as both novel and familiar, plotted as the spike count response of neuron 1 versus neuron 2. In this scenario, visual information (e.g. image or object identity) is reflected by the population response pattern, or equivalently, the angle that each population response vector points. In contrast, visual memory information is reflected by changes in population vector length (e.g. a multiplicative rescaling with stimulus repetition). Because visual memory does not impact vector angle, superimposing visual memories in this way would mitigate the impact of single-exposure plasticity on the underlying perceptual representation.

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

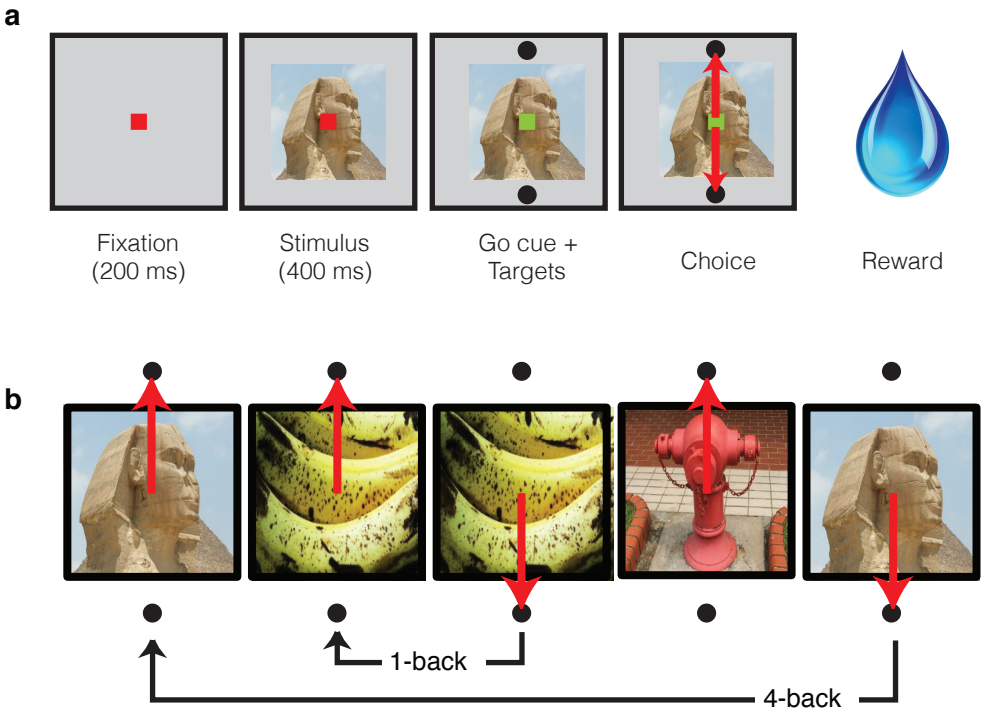

Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 2. *Single-exposure visual memory task*. In this task, monkeys viewed images and reported whether they were novel (i.e. had never been encountered before) or were familiar (had been encountered once and earlier in the same session) across a range of delays between novel and repeated presentations. **a)** Each trial began with the monkey fixating for 200 ms. A stimulus was then shown for 400 ms, followed by a go cue, reflected by a change in the color of the fixation dot. Targets, located above and below the image, were associated with novel and familiar selections, and differed for each monkey. Successful completion of the trial resulted in a juice reward. **b)** Example sequence where the upward target was associated with novel images, and the downward target with familiar images. Familiar images were presented with n-back of 1, 2, 4, 8, 16, 32, and 64 trials, corresponding to average delays of 4.5s, 9s, 18s, 36s, 1.2 min, 2.4 min, and 4.8 min, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

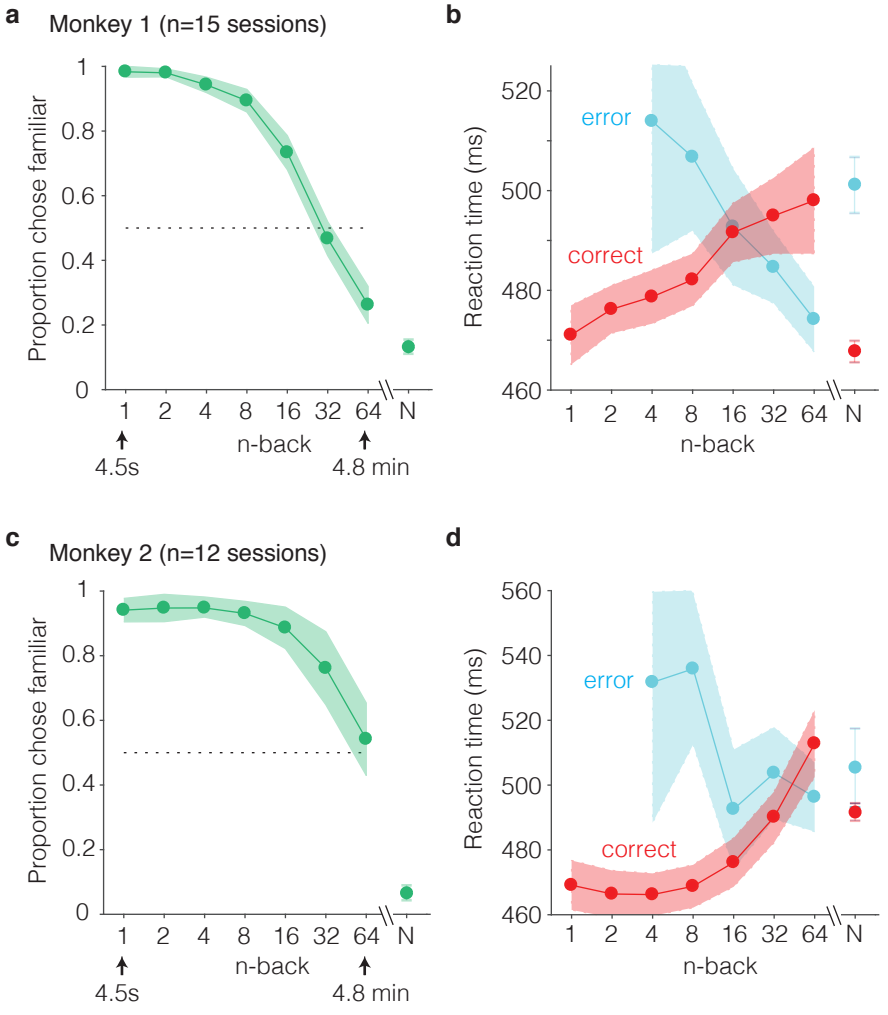

Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 3. *Behavioral performance of two monkeys on the single-trial visual recognition memory task.* **a,c)** "Forgetting functions", plotted as the proportion of trials that each monkey reported images as familiar as a function of the number of trials between novel and repeated presentations (n-back). Novel trials are indicated by "N" and a break in the x-axis. The dotted line indicates chance performance on this task, 50%. Error bars depict 97.5% confidence intervals of the per-session means. **b,d)** Mean reaction times, parsed according to trials in which the monkeys answered correctly versus made errors. Error bars depict 97.5% confidence intervals computed across all trials.

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

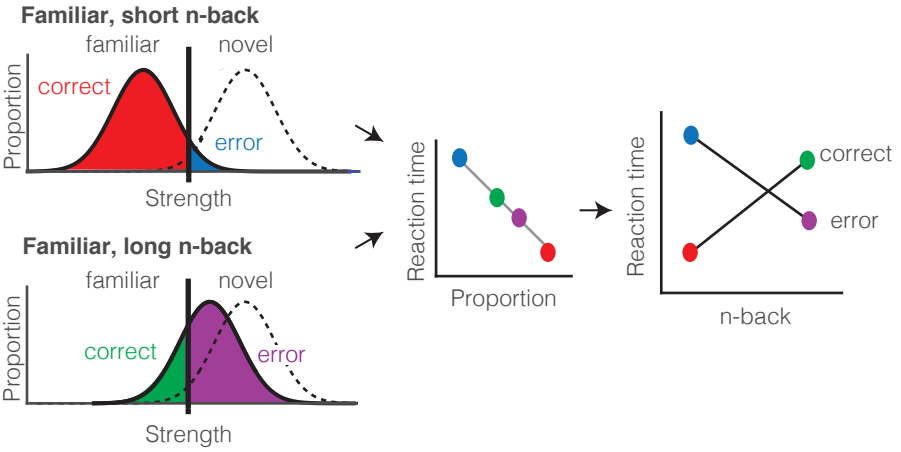

Figure 4

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 4. *Strength theory qualitatively predicts x-shaped reaction time patterns.* Like signal detection theory, strength theory proposes that the value of a noisy internal variable, memory strength, is compared to a criterion to differentiate novel versus familiar predictions. *Left:* shown are the hypothetical distributions of memory strengths across a set of images presented as novel (dashed lines) and as familiar (black), repeated after a short (top) versus long (bottom) delay. The colored portions of each familiar distribution indicate the proportion of trials corresponding to correct reports and errors, based on the position of the distribution relative to the criterion. In the case of short n-back, memory strength is high, the proportion correct is high, and the proportion wrong is low. In the case of long n-back, memory strength is low, the proportion correct is low and the proportion wrong is high. *Middle:* strength theory proposes an inverted relationship between proportion and mean reaction times, depicted here as linear. *Right:* passing the distributions on the left through the linear function in the middle produces an x-shaped reaction time pattern.

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

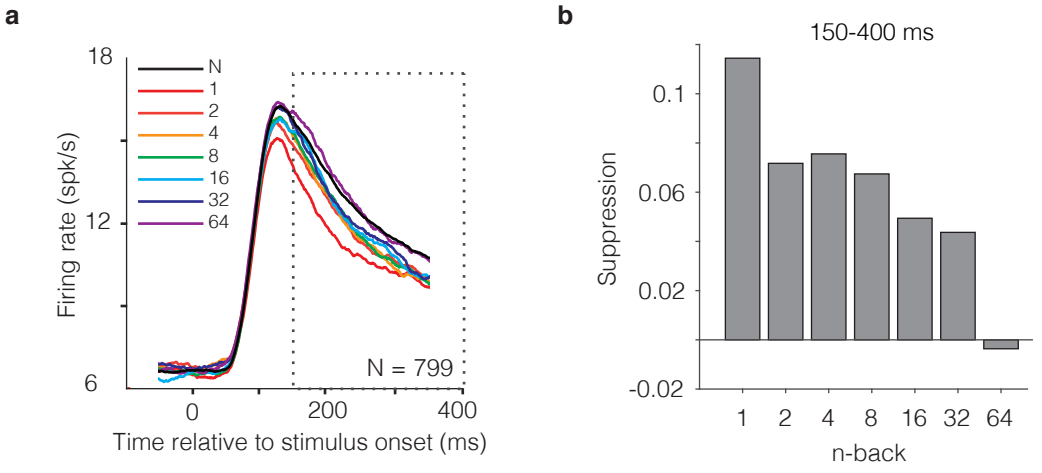

Figure 5

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 5. *IT repetition suppression cannot account for single-exposure visual memory behavior.* **a)** Grand mean firing rates for all units, plotted as a function of time aligned to stimulus onset, parsed by images presented as novel (black) versus familiar at different n-back (rainbow). Traces were computed in 1 ms bins and smoothed by averaging across 50 ms. The dotted box indicates the spike count window corresponding to the analysis presented in panel b. **b)** The calculation of suppression magnitude at each n-back began by quantifying the grand mean firing rate response to novel and familiar images within a window positioned 150 ms to 400 ms after stimulus onset. Suppression magnitude was calculated separately for each n-back as (novel – familiar)/novel. The qualitative disagreement between the monkeys partial success at reporting 64-back as familiar (Fig 3a,c) and the absence of a suppressive signal at 64-back suggests that a strict interpretation of the repetition suppression hypothesis, where total numbers of spikes in IT are used to classify images as novel versus as familiar, cannot account for the monkeys' behavior.

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

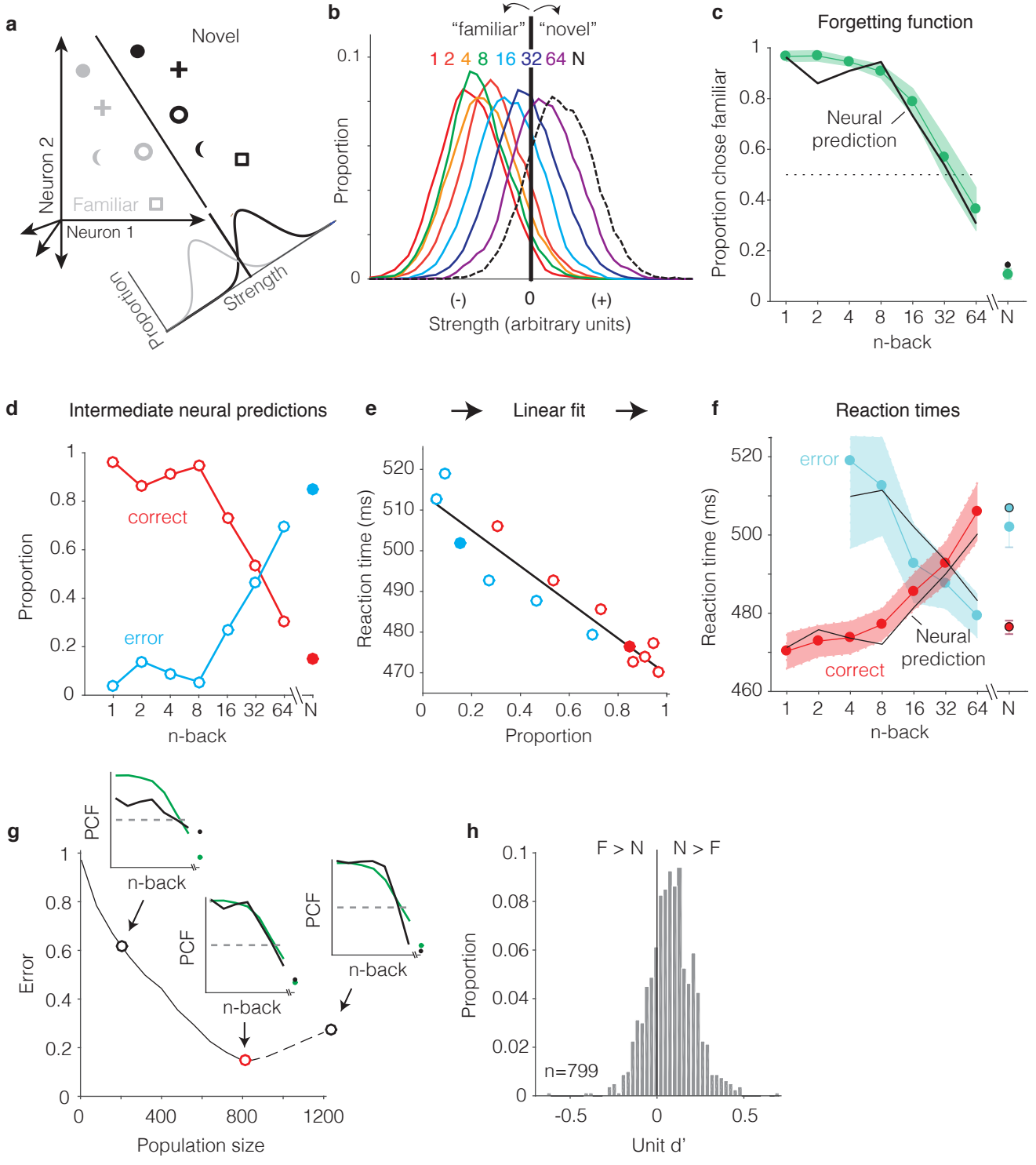

Figure 6

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 6. *Transforming IT neural data into behavioral predictions.* In all panels, behavioral and neural data correspond to the data pooled across the two monkeys. **a)** A cartoon depiction of how memory strength was measured for each n-back. Shown are the hypothetical population responses of 2 neurons to different images (represented by different shapes) shown as novel (black) versus as familiar (gray). The line depicts a linear decoder decision boundary optimized to classify images as novel versus as familiar. Distributions across images within each class are calculated by computing the linearly weighted sum of each neuron's responses and subtracting a criterion. **b)** Distributions of the linearly weighted IT population response, as a measure of memory strength, shown for novel images (black dotted) and familiar images parsed by n-back (rainbow), for a population of 799 units. To compute these distributions, a linear decoder was trained to parse novel versus familiar across all n-back via an iterative resampling procedure (see Methods). **c)** Black: the neural prediction of the behavioral forgetting function, computed as the fraction of each distribution in panel b that fell on the "familiar" (i.e. left) side of the criterion. Behavioral data are plotted with the same conventions as Fig 3a,c. **d)** The first step in the procedure for estimating reaction times, shown as a plot of the proportions of each distribution from panel b predicted to be correct versus wrong, as a function of n-back. Solid and open circles correspond to novel and familiar trials, respectively. Note that the red curve (correct trials) simply replots the predictions from panel c and the blue curve (error trials) simply depicts those same values, subtracted from 1. **e)** A plot of the proportions plotted in panel d versus the monkeys' mean reaction times for each condition, and a line fit to that data. **f)** The final neural predictions for reaction times, computed by passing the data in panel d through the linear fit depicted in panel e. Behavioral data are plotted with the same conventions as Fig 3b,d. **g)** Mean squared error between the neural predictions of the forgetting function and the actual behavioral data, plotted as a function of population size. The solid line corresponds to the analysis applied to recorded data; the dashed line corresponds to the analysis applied to simulated extensions of the actual data (see Methods). Insets indicate examples of the alignment of the forgetting function and neural prediction at three different population sizes, where green corresponds to the actual behavioral forgetting function and black corresponds to the neural prediction. PCF = proportion chose familiar. **h)** Distribution of unit d' across the 799 units. N=novel; F=familiar.

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

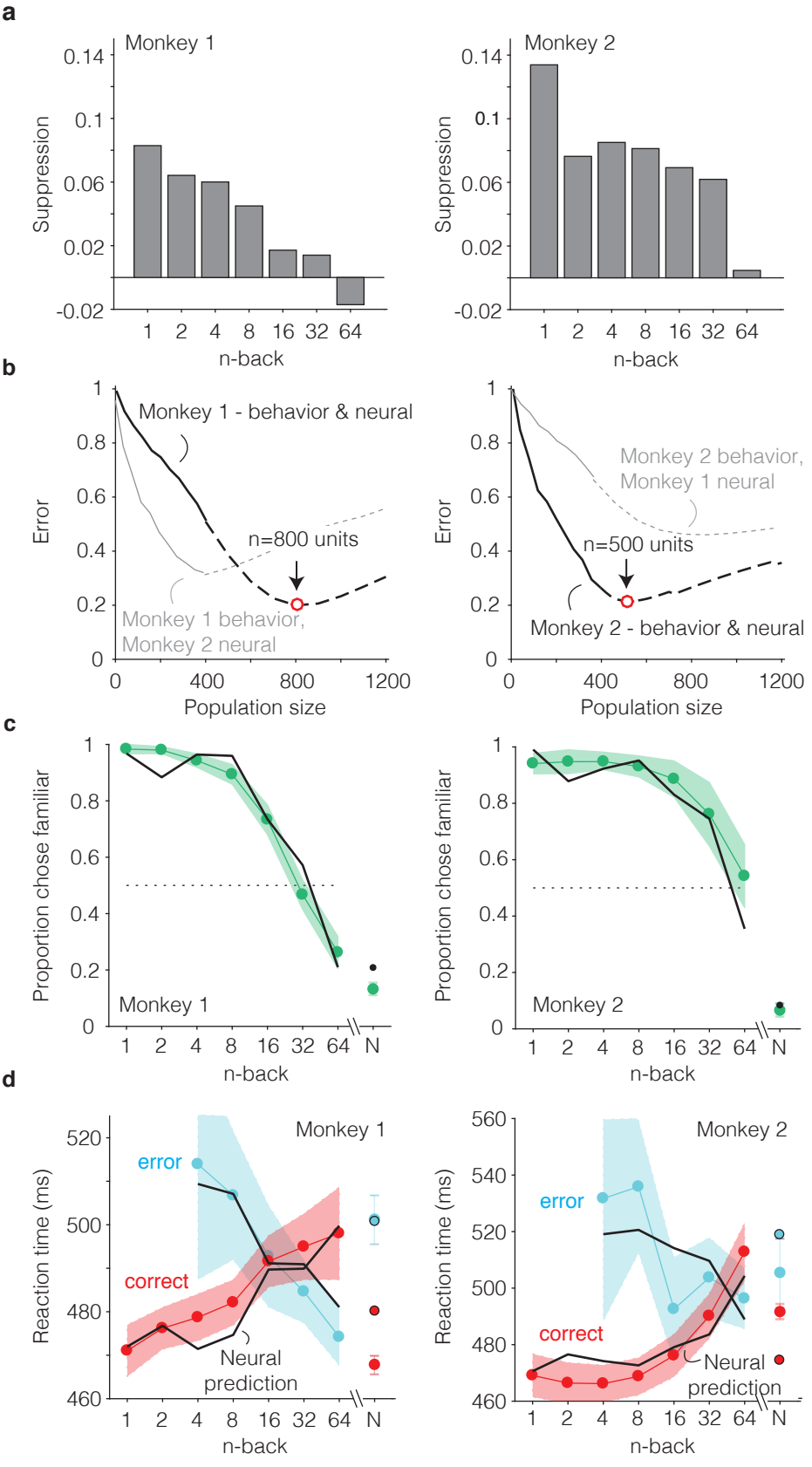

Figure 7

bioRxiv preprint doi: https://doi.org/10.1101/197764; this version posted October 3, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 7. *Individual differences.* **a)** Suppression magnitudes at each n-back, computed as described for Fig 5b but isolated to the units recorded in each monkey individually. **b)** Plot of mean squared error as a function of population size, computed as described for Fig 6g but for each monkey individually. Black: Comparisons when behavioral and neural data come from the same monkey. Gray: Comparisons when behavioral and neural data are crossed between monkeys. c-d) Alignment of neural predictions and behavior, computed as described for Fig 6c & 6f, but applied to each monkey individually, for population sizes indicated by the red dots in panel b. Behavioral data are replotted from Fig 3 and neural predictions are superimposed in black.