Rutishauser *et al*. 10.1073/pnas.0706015105.

Supporting Information

Files in this Data Supplement:

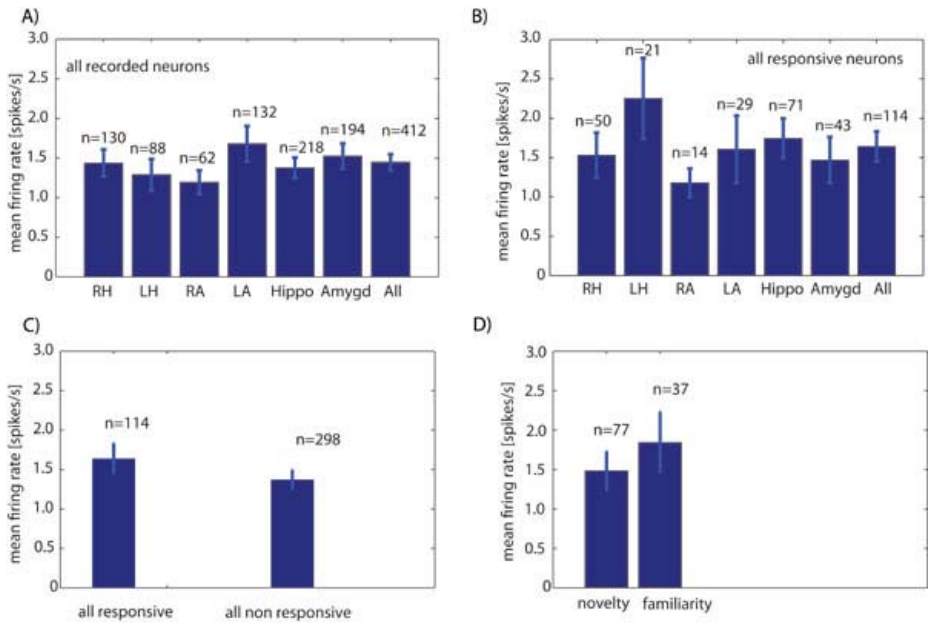

Fig. 5. Baseline firing rates calculated using the entire period recorded. Abbreviations: left hippocampus (LH), right hippocampus (RH), left amygdala (LA), right amygdala (RA). (*A*) Mean firing rate of all recorded neurons, regardless of whether they were responsive or not. There was no significant difference in baseline firing between brain areas [ANOVA for the four brain areas (LH, RH, LA, RA), $P = 0.37$]. (*B*) Same as in *A*, but only including neurons whose firing rate was significantly different for new vs. old items. There was no significant difference between brain areas (ANOVA, $P = 0.44$). (*C*) The mean firing rate of all responsive ($n = 114$) and all non-responsive ($n = 298$) neurons was not significantly different (t test, $P = 0.24$). (*D*) Mean firing rates were not different for neurons which increase firing in response to novel vs. old stimuli (t test, $P = 0.42$). All error bars are ±SEM.

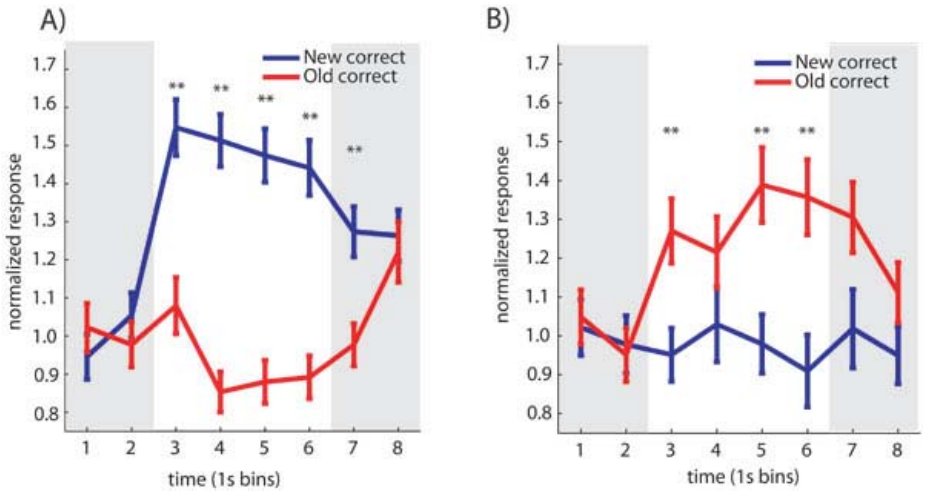

Fig. 6. Population response. (*A* and *B*) Population average of all neurons that responded significantly during the stimulus period. The stimulus was o n the screen during the 4-s period marked in white. (*A*) Average of all neurons that increased firing to correctly recognized new items ("novelty detectors") ($n = 48$). (*B*) Average of all neurons that increased firing to correctly recognized old items ("familiarity detectors") ($n = 26$). Errors are ±SEM, and ** indicates significance of a one-tailed t test at $P \leq 0.006$ ($P \leq 0.05$ Bonferroni corrected for multiple comparisons). Firing was normalized to the 2-s baseline firing before stimulus onset marked in gray. Note that this does not mean all neurons fired during the entire period, but rather represents the population average.

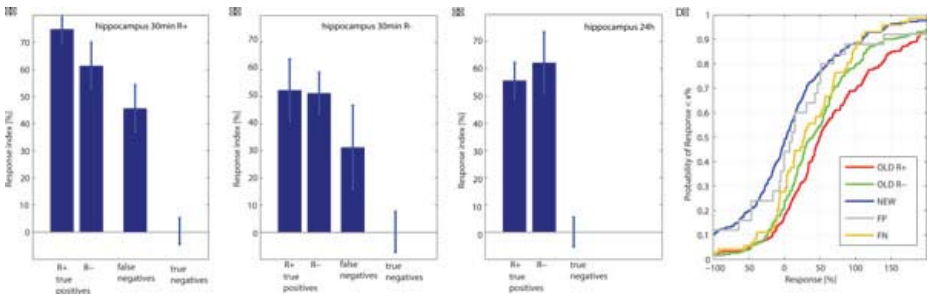

Fig. 7.
A continuous strength of memory gradient exists when the hippocampal neuronal population is considered in isola tion. In this figure, the same measures are replotted, but all units recorded from the amygdala are excluded. All findings remain valid. (*A*) Trials from the 30-min R+ sessions. There is a significant difference between R+ and R- trials ($P = 0.03$) as well as between new and false negatives ($P = 0.001$). Compare with Fig. 3*C*. (*B*) Trials from the 30-min R- session. There is no significant difference between R+ and R- trials ($P = 0.93$) but false negatives are still significantly different from new trials ($P = 0.07$). Compare with Fig. 3*F*. (*C*) Trials from the 24-h sessions. There is no significant difference between R+ and R- trials. Error trials are not shown (not enough for 24-h sessions). Compare with Fig. 3*H*. (*D*) cdf of response index of all hippocampal neurons recorded in all 30-min sessions. R+ and R- trials are signif icantly different (red vs. green, $P = 0.01$) as are new and false negatives (blue vs. yellow, $P < 0.001$). Not enough false positive trials are available to allow statistical analysis of false positives. Comp are with Fig. 4. All error bars are ±SE.

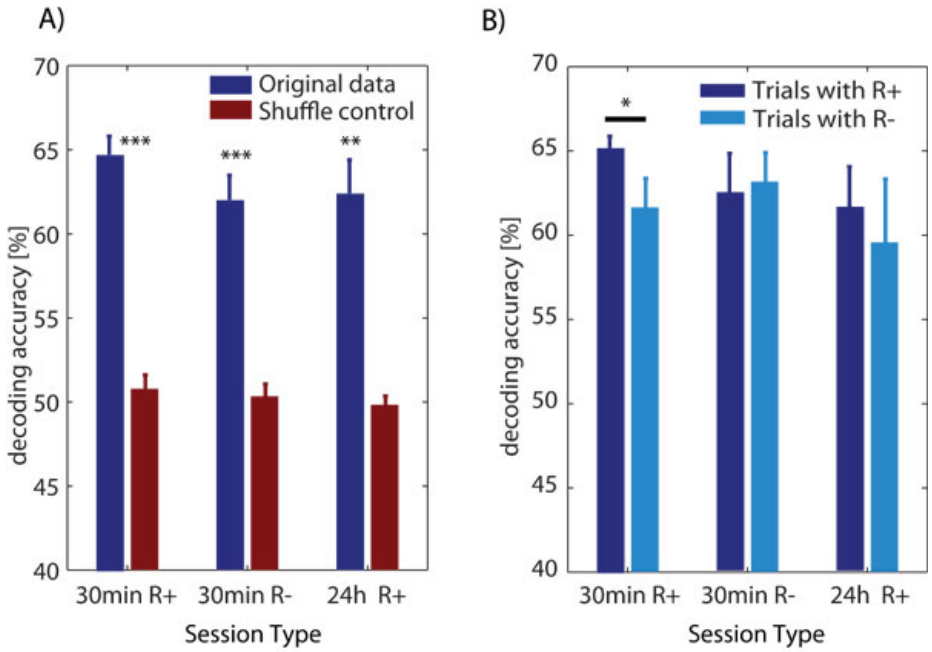

Fig. 8. Whether a stimulus is new or old can be predicted regardless of whether recall was successful or not. The decoder had access to the number of spikes fired in the three consecutive 2-s bins following stimulus onset (three numbers total). (*A*) Session-by-session differences. The performance of the decoder did not change for all three groups (ANOVA, $P = 0.35$). $n = 7,6,4$ sessions, respectively. (*B*) Trial-by-Trial differences. Here, the decoder was trained on the complete set of trials but its performance was evaluated separately either for failed (R^-) or successful (R^+) recall trials. Clearly, the familiarity of the stimulus could be decoded for trials with failed recall (R^-). In the 30min delay sessions with successful recall (30-min R^+), firing during successful recall trials contained significantly more information about the familiarity of the stimulus ($P = 0.037$, paired t test, $n = 7$ sessions). All error bars are ±SE.

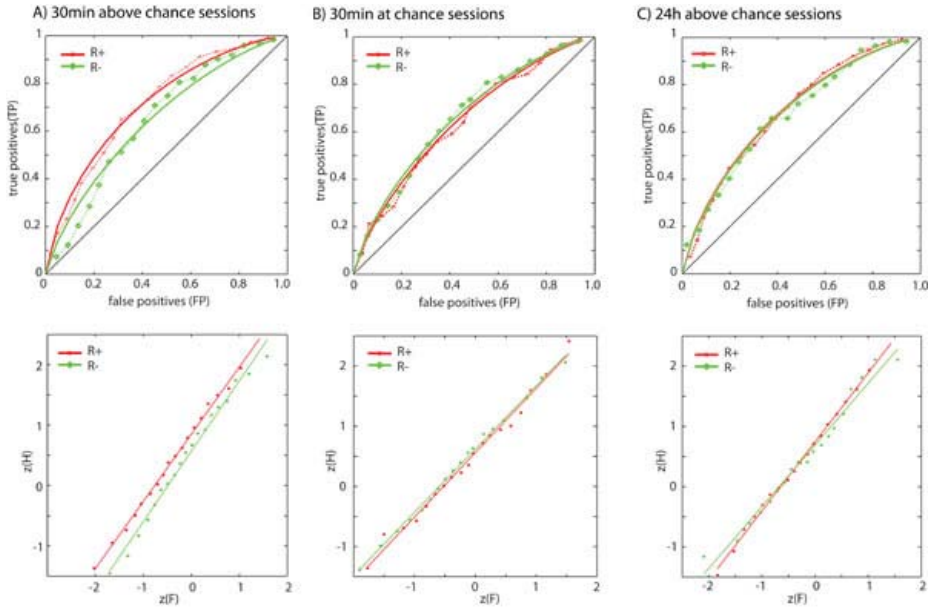

Fig. 9. ROC analysis of the neuronal data for all 3 behavioral groups (*A*, 30 min above chance; *B*, 30min at chance; *C*, 24 h above chance). The top row shows the raw datapoints as well as fits computed from d'. The bottom row shows the same but z-transformed. R^2 is >0.97 for all straight line fits. See the supplementary methods for how the ROC was computed. (*A*) d' for R^+ and R^- groups was 0.81 and 0.55, respectively. The slope (s) of the z-transformed line was 1.11 ± 0.03 and 1.16 ± 0.07, respectively. ± are 95% confidence intervals. (*B*) d' was 0.55 and 0.61 and s was 1.07 ±0.06 and 1.05 ± 0.04, respectively. (*C*) d' was 0.73 and 0.69 and s was 1.14 ± 0.04 and 1.02 ± 0.08.

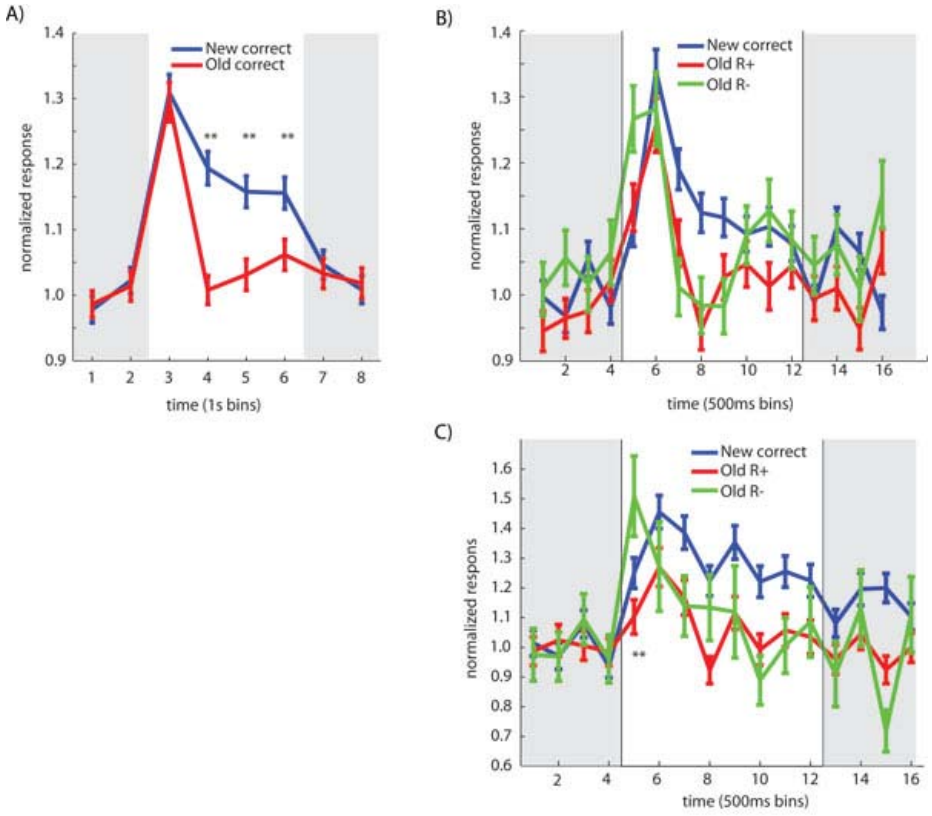

Fig. 10. (*A*) Population average of all recorded neurons that have a baseline firing rate of >0.25 Hz (n = 346). While the firing of most neurons was not significantly different between new vs. old, a significant difference between new and old stimuli could still be observed in the population average. Errors are ± SEM and ** indicates significance of a one-tailed t test at $P \leq 0.006$ ($P \leq 0.05$ Bonferonni-corrected for eight multiple comparisons). (*B*) Population average of all neurons with recollected and not recollected familiarity trials shown separately (n =). (*C*) Population average of all neurons recorded in the 30-min delay sessions with above chance recollection performance. The signal for the not recollected items peaked earlier than the signal for recollected items. ** indicates a significant difference between recollect (R^+) and not recollected (R^-) items at $P \leq 0.003$ ($P \leq 0.05$ Bonferonni-corrected for 16 multiple comparisons). The only difference was for the first time bin (0-500ms after stimulus onset). n = 134 neurons.

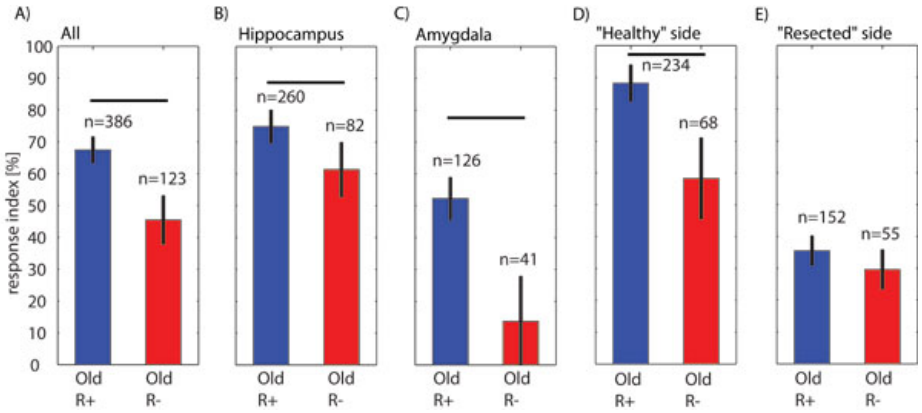

Fig. 11.
Comparison of trial-by-trial response strength for different subcategories of neurons. In this figure, only neurons from 30-min delay with successful recollection (30-min R+) are included. (*A*) All trials from all areas (same as Fig. 3*B*). (*B*) Only trials from hippocampal neurons. (*C*) Only trials from amygdala neurons. (*D*) Only trials from the "healthy" hemisphere. (*E*) Only trials from neurons in the eventually resected hemisphere. In *A-D*, the response to R+ compare to R- trials is significantly different ($P < 0.05$, two-tailed Kolmogorov-Smirnov test, compare with Fig. 3*B*). The response in *E* is not significantly different.

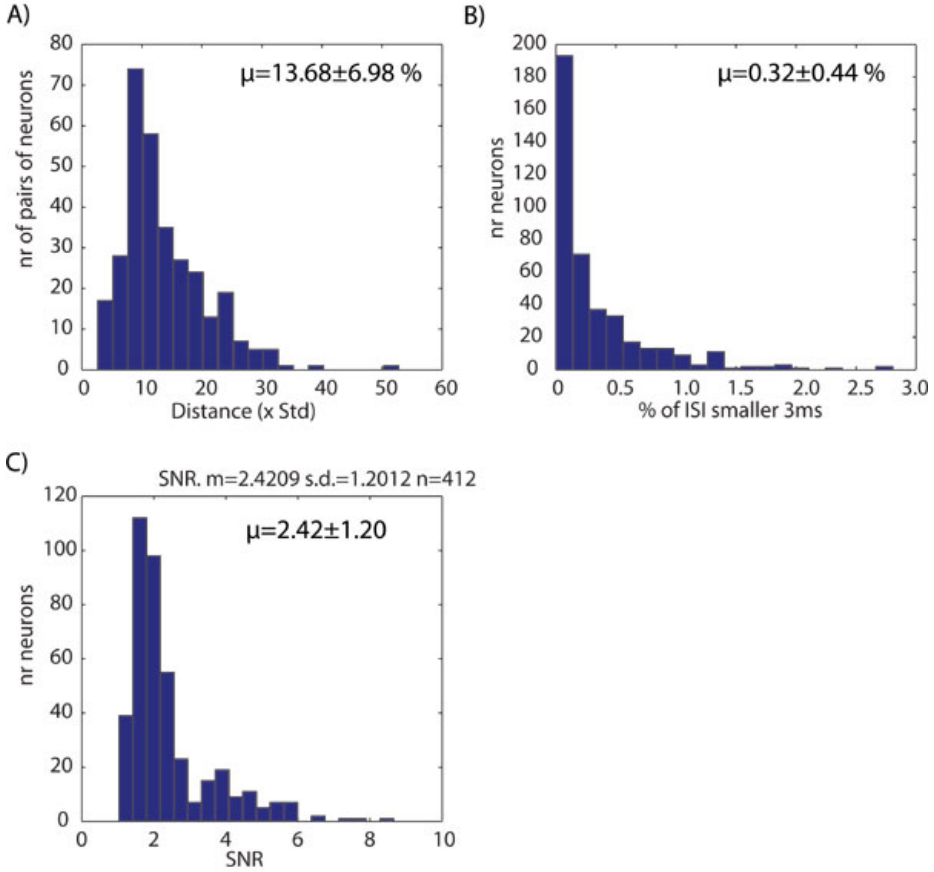

Fig. 12. Sorting quality for the 412 recorded units. (*A*) Histogram of the distance, in standard deviations, between all pairs of clusters. Only channels on which more than one unit was detected are included (315 pairs from 130 channels). The mean distance was 13.68 ±6.98 (±SD) standard deviations. (*B*) Histogram of the percentage of interspike intervals (ISI) that were <3 ms. On average 0.32 ± 0.44% of all ISIs were <3 ms (n = 412). (*C*) Histogram of the SNR of all 412 units.

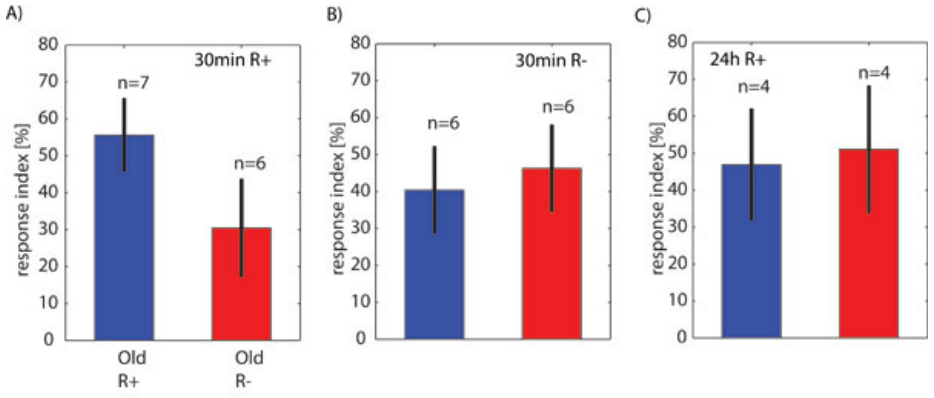

Fig. 13.
Comparison of response strength across different recording sessions (days). The difference is only significant for the 30-min R+ sessions. The data displayed here is the same as detailed in Fig. 3. However, here the mean response index for R+ and R- trials is compared between recording sessions. (*A*) The response index for all recording sessions that had above chance recollection. The difference approaches significance (P = 0.07). Number of sessions is seven and six, respectively (from four patients; one session had no R- trials). (*B*) Same as *A* but for all recording sessions with at chance recollection. Number of sessions is six for both groups (from five patients). There was no significant difference (P = 0.63). (*C*) Same as *A* but for all recording sessions with 24-h delay and above chance recollection. Number of sessions is four from three patients. There was no significant difference (P = 0.57). Error bars are ±SEM with n as specified. P values are from a t test.

Table 1. Neuropsychological evaluation of patients

				WAIS-III	WMS-R

Patient	Age	Sex	Diagnosis	PIQ	VIQ	FSIQ	VerbalMem	Mentalcontrol	VPA 2	LM 2	Vis Rep 1	Vis Rep 2
1	28	m	left temporal	125	98	110	114	6	4	24	37	39
2	41	f	left temporal	92	91	91	91	5	8	18	37	29
3	20	f	left temporal	92	93	93	83	6	8	16	34	28
4	58	f	left temporal	85	83	83	83	6	4	10	22	7
5	23	m	left temporal & frontal pole	144	111	126	122	6	8	26	39	39
6	44	m	right temporal	76	92	84	83	6	5	10	29	14
7	51	f	left temporal	90	95	93	89	6	4	23	34	34
8	16	m	right lateral frontal	84	91	88	n/a	n/a	8	n/a	31	29
av	*35.1*	-	-	*98.5*	*94.3*	*96.0*	*95.0*	*5.9*	*6.1*	*18.1*	*32.9*	*27.5*
mean raw								5.0±1.2	7.6±0.7	21.9±9.2	32.5±5.3	29.5±7.1

Intelligence was measured using the Wechsler Intelligence Scale (WAIS-III) measures of performance IQ (PIQ), verbal IQ (VIQ) and full scale IQ (FSIQ). All IQ scores have an average of 100 (by design). Memory measures are from the Wechsler Memory Scale Revised (WMS-R). Verbal memory is an WMS-R index score with a mean of 100 of the normal population (by definition). The remaining WMS-R scores are raw (unnormalized) scores. For the raw scores, the mean and standard deviation of the normal population (from WMS-R) is shown in the last row for the average age of our population. Abbreviations: Verbal paired associates 2 (VPA 2), Logical Memory 2 (LM 2), Visual Reproduction 1 (Vis Rep 1), Visual Reproduction 2 (Vis Rep 2).

Table 2. Number of neurons recorded in each area (first row) and number of neurons that responded in each behavioral group(2^{nd}, 3^{rd}, 4^{th} row)

	Group	Hippocampus			Amygdala			All		
Recorded	*30min R+*	77			103			180		
	30min R-	96			47			143		
	24h R+	45			44			89		
	all	218			194			412		
		Nov	Fam	All	Nov	Fam	All	Nov	Fam	All
New v. old	*30min R+*	25	7	32	10	5	15	35	12	47
	30min R-	11	11	22	13	3	16	24	14	38
	24h R+	11	6	17	7	5	12	18	11	29
	all			71			43	77	37	114
New v. old	*30min R+*	14	5	19	6	3	9	20	8	28
	30min R-	5	6	11	6	1	7	11	7	18
	24h R+	5	4	9	5	2	7	10	6	16
	all			39 (55%)			23 (53%)			62 (54%)

& baseline 1										
New v. old & baseline 2	*30min R+*	22	7	29	10	5	15	32	12	44
	30min R-	10	10	20	11	3	14	21	13	34
	24h R+	9	6	15	7	5	12	16	11	27
	all			64 (90%)			41 (95%)			105 (92%)

The second row shows the number of neurons which had a significantly different firing rate for old v. new trials during the post-stimulus period (6s). The last two rows show the number of neurons which are, in addition, also significantly different for two different baseline comparisons (1 and 2). The two baseline comparisons are: i) the trials associated with the type of unit are significant from baseline. (That is, if the neuron is classified as a familiarity neuron, the old trials were significantly different from baseline. The same applies for the novelty neurons, but for the new trials). ii) either the new or the old trials are significantly different from baseline. Note that the first (i) baseline condition is the most restrictive: for example, a familiarity unit that decreases firing to novel items but remains at baseline for familiar items would not pass this test. For the second baseline condition, 92% of units (105 of 114) remain significant. Thus, almost all units fired significantly different from baseline for either the new or old condition. Note that some of the n's reported in the main analysis are slightly lower than the numbers reported in this table. This is because additional constraints were applied (for example, at least one R+ and one R- trial for each included unit).

SI Text

SI Results

Behavior Quantified with d'. d' was 3.11 ± 0.08, 2.40 ± 0.28 and 2.67 ± 0.68 for the 30 min R^+, 30 min R^- and 24 h groups, respectively. Pairwise tests revealed a significant difference between the 30 min R^+ and R^- group (*t* test, $P \leq 0.05$). Thus, in terms of d', patients that exhibited no recollection had significantly lower recognition performance.

Neuronal ROCs.
Based on the response values as summarized in Fig. 3 we constructed two neuronal ROCs (1): one for trials with spatial recollection and one without (SI Fig. 9). The z-transformed ROC was fit well by a straight line (R = 0.997 and R = 0.988 for R+ and R-, respectively). The slope for both curves was significantly different from 1, indicating that the variance of the targets and distractors was different (for a 95% confidence interval the slope was 1.11 ±0.03 and 1.16 ± 0.07, respectively). The d' for recognized and recollected targets was 0.81 and for targets that were only recognized it was 0.55. Thus, the d' was increased by the addition of recollective information. This is in analogy to the behavioral recognition performance, which was also increased (Fig. 1*E*, see above).

Interestingly, the slopes of the neuronal z-ROCs are >1 (see above). This indicates greater variability for distractors (here new items) compared to familiar items. z-ROC slopes derived from behavioral data are found to be <1 (2). This has been used as evidence that the target distribution has higher variance compared to the distractor distribution. Intriguingly, we found that the slopes of our z-ROCs are >1. This further indicates that the neuronal signals in the medial temporal lobe (which we analyze here) represents a memory signal that should be regarded as the input to the decision process, not its output. What is measured behaviorally is the decision itself and it is thus conceivable that the decision process adds sufficient variance to change the slope of the z-ROC.

Responses of Novelty and Familiarity Neurons Compared with Baseline.
The neurons used for our analysis were selected based on a significant difference in firing in response to new vs. old stimuli. This is the most sensitive test because it detects many different patterns in which activity could differ. Example patterns that are detected by this way of classifying units are: (*i*) increase of firing only for one category (new or old) whereas the other remains at baseline; (*ii*) decrease of firing only for one category, with the other remaining at baseline; and (*iii*) a bimodal response with an increase to one category and a decrease to the other category. One concern with this analysis is that the response itself might not be significantly different from baseline. This would primarily be the case if the response is bimodal, i.e., a slight increase to one category and a slight decrease to the other. To investigate this possibility we performed additional analysis by comparing the activity of neurons that are classified as novelty or familiarity detecting units against baseline (SI Table 2). We used two different methods: the first ("method 1") tests whether the unit increases its firing rate significantly for either the old (familiarity neurons) or the new trials (novelty neurons). However, there are several classes of units that this method misses. For example, a unit that remains at baseline for old trials and reduces its firing rate for new trials would be classified as a familiarity unit. However, it would not pass the baseline test since the response for old trials remains at baseline. To include such units we used a second method ("method 2"): for a unit to be considered responsive, the activity of either the new or the old trials needs to be significantly different from baseline. The unit in the above example would pass this test.

Using method 2, we found that 92% of all units that were classified as signaling a difference between new vs. old were in addition also firing significantly different relative to baseline (see Table S2 for details). Using method 1, 54% of all units pass this additional test. Thus, ≈40% of the units signal information by a decrease in firing rate rather then an increase.

Population Activity.
So far we have analyzed the spiking of single neurons that fired significantly different for new vs. old stimuli. However, the majority of neurons (72% of neurons; 298 of 412) did not pass this test and thus were not considered in our first set of analyses. Was there a difference in mean firing between new and old stimuli if neurons were not preselected? To address this, we calculated a mean normalized activity for all recorded neurons in all sessions, separately for new and old trials (SI Fig. 10*A*). This signal reflects the overall mean spiking activity of all neurons and is thus similar to what might be measured by the fMRI signal (see *Discussion*). Only trials where the stimulus was correctly recognized were included. The mean firing activity of the entire population was significantly different in the time period from 2 to 4s relative to stimulus onset ($P \leq 0.05$, *t* test, Bonferroni-corrected for $n = 8$ comparisons). Thus, a difference in overall mean activity for novel vs. familiar stimuli can be observed even without preselecting neurons. However, the initial response (first 1 s, SI Fig. 10*A*) did not differentiate between the two types of stimuli. Rather, a sharp onset in the response could be observed for both classes of stimuli. Did the population only differentiate because the novelty and familiarity detectors were included in the average? We also calculated the population average (as in SI Fig. 10*A*) using only the units, which were not classified as either novelty or familiarity detectors. The average population activity still exhibited a sharp peak for both types of stimuli after stimulus onset and significantly differentiated between novel and familiar items in subsequent time bins ($P \leq 0.05$, *t* test, Bonferroni corrected for $n = 8$ comparisons).

Is the population response different for stimuli that are recollected compared to stimuli that are only recognized? The previous average included all old trials, regardless of whether the stimulus was recollected or not. Next, we averaged all trials from all neurons recorded for the 30min delay sessions with good recollection performance (30 min R^+). We found a similar pattern of population activity (SI Fig. 10*B*). Crucially, however, the neuronal activity in response to familiar stimuli that were later not recollected peaked earlier. Measured in time bins of 500 ms, the only significant difference between familiar stimuli that were recollected or not was in the first 500 ms after stimulus onset ($P \leq 0.05$, *t* test, Bonferroni-corrected for $n = 16$ comparisons). Thus, the population activity peaks first for stimuli that are not recollected, followed by novel and recollected stimuli.

Decoding of Recognition Memory.
Is the ability to determine whether a stimulus is old influenced by whether the stimulus was recollected or not? In the main text we have shown that the responses to recollected stimuli are stronger compared to items that are not recollected. Here, we investigate whether this increased response leads to an improvement in the ability to determine (based on the neuronal firing only) whether a stimulus is new or old. If the two types of information (familiarity and recollection) interact, one would expect that the ability to recollect would increase the ability to determine whether a stimulus has been seen before. Alternatively, recollection could be a process that is only triggered after the familiarity is already determined and these two types of

information would thus be independent. Thus, one would expect no difference in the ability to determine the familiarity from the spiking of single neurons in cases of successful vs. failed spatial recollection. To answer this question, we used a simple decoder. It used the weighted linear sum of the number of spikes fired after the onset of the stimulus. The weights were determined using regularized least squares, a method very similar to multiple linear regression (see methods). The decoder had access to the number of spikes in the 3 consecutive 2s bins after stimulus onset (3 numbers per trial).

First, we used the decoder to determine for how many trials we could correctly predict whether the stimulus was new or old, based only on the firing of a single neuron. For all sessions (n = 17), the decoder was able to predict the correct identity for 63 ± 1% of all trials. We repeated this analysis for each of the 3 behavioral groups (R^+ 30 min, R^- 30min and R^+ 24 h). We found (SI Fig. 8*A*) that the recognition decoding accuracy (chance 50%) did not depend on whether the subject was able to recollect the source of the stimulus or not (1-way ANOVA, P = 0.35). Thus, decoding of familiarity is equally effective, even in the group where patients were not able to recollect at all (SI Fig. 8*A*, 30 min R- sessions).

Was there a difference in decoding performance in the same day group where subjects had good recollection performance? We selectively evaluated the performance of the decoder for two groups of trials: trials with correct recollection and trials with failed recollection. We find that firing during trials with failed recollection does carry information about the familiarity of the stimulus (SI Fig. 8*B*, R-). The ability to predict the familiarity of the stimulus was slightly improved for the behavioral group with good recollection performance on the first day (SI Fig. 8*B*, right. P = 0.03, paired t test).

SI Discussion

Differences Between Amygdala and Hippocampal Neurons.
So far, we have analyzed neurons recorded from the Amygdala and the Hippocampus as a single group. We pooled the responses from both groups because we previously found that both structures contain units that respond to novel and familiar items in a very similar fashion (3). Nevertheless we also analyzed the activity separately for both brain structures. We find that the previous finding still holds - while the response magnitude differs, the overal response pattern is very similar. In particular all primary findings of our article hold independently for the hippocampus and the amygdala (see below).

We found that the increased response to old stimuli that are recollected (R+) compared to stimuli that are not recollected (R-) is present in both hippocampal and amygdala neurons (SI Fig. 11; 74.8 ±5.3% vs. 61.3 ± 8.6% for the hippocampus and 52.2 ± 6.8% vs. 13.7 ± 14.2% for the Amygdala). The response magnitude (comparing all old trials, regardless of whether they are R+ or R-), however, is larger in the hippocampus (71.6 ±4.5% vs. 42.8 ± 6.3%, P < 0.001). While the amplitude of the response is different there is nevertheless a significant difference between R+ and R- trials in both areas.

This is further illustrated in SI Fig. 7, where we replotted the response to old R+, old R-, new and false negatives (forgotten items) for all 3 behavioral groups only considering hippocampal units (SI Fig. 7 *A-C*). The relevant differences (R+ vs. R-, New vs. false negative) are the same as for the pooled responses (see SI Fig. 7 legend for statistics). Similarly, the responses during the error trials (false negatives and false positives) are the same (compare SI Fig. 7*D* with Fig. 4*B*).

We also repeated the within-group ANOVA for only the hippocampal units of the 30min R+ session. The ANOVA was significant for novelty (P = 4.1e-6) and familiarity (P = 1.3e-19) units. The planned contrasts of R- vs. New and R+ vs. R- revealed a robust difference for novelty (P = 5.1e-5 and P = 0.04, respectively) units. For familiarity units, the R- vs. New contrast was significant (P = 0.002) whereas the R+ vs. R- contrast was only approaching significance (P = 0.17). This is because there were only seven familiarity units that contribute to this comparison. Repeating the same comparisons while excluding all units that do not fire significantly different from baseline (see SI Table 2) reveals a similar pattern: the ANOVA for familiarity units remains unchanged (all units different from baseline) whereas the novelty units ANOVA still shows a significant difference between R- vs. New (P = 2.7e-5) and R+ vs. R- (P = 0.016).

Differences Between Epileptic and Nonepileptic Tissue.
Was the neuronal response reported here influenced by changes induced by disease? All subjects for this study have been diagnosed with epilepsy and as such some of the effects may not extend to the normal population. Behaviorally, our subjects were comparable to the normal population (see Table 1). Also, we separately analyzed a subset of neurons that were in a non-epileptic region of the subject's brain. We found a comparable (but stronger) response to old stimuli in this "healthy" neuron population (SI Fig. 11*D*). Similarly, we find that neurons from the "to be resected" tissue still exhibited a response to old stimuli (SI Fig. 11*E*). This response was, however, weaker and there was no significant difference between recollected and not recollected stimuli. Thus, it is possible that the average difference between recollected and not recollected items in normal subjects will be larger than that observed in the epileptic patients in our study.

Relationship to Previous Single-Cell Studies.
A previous human single-cell study (4) concluded that the neuronal activity observed during retrieval is due to recollection. The task used was the repeated presentation of word pairs with later free recall and thus included no recognition component. Due to the choice of words and the repeated presentation of the same word pairs, the novelty/familiarity of the stimuli was not controlled for. It is thus not clear whether the activity observed was related to recollection or to the recognition of the familiarity of the stimuli. Here, we combine both components in the same task and thus demonstrate that the same neurons represent information about both aspects of memory simultaneously. Similar paired associates tasks have been used with monkeys (5, 6). Changes in neuronal firing were, however, only observed after many learning trials (>10). A neuronal correlate of episodic memory requires changes after a single learning trial. It thus seems possible that this study documented the gradual acquisition of well learned associations rather than episodic memories.

Relationship to Evoked Potentials.
Both surface and intracranial evoked potentials show prominent peaks in response to new stimuli. Scalp EEG recordings during recognition of previously seen items show an early frontal potential (≈300 ms), which distinguishes old from new items and a late potential (≈500-600 ms) that is thought to reflect the recollective aspect of retrieval (7). However, the signal origin of these scalp recordings is not known. These differences between evoked potentials in response to new and old items are reduced or absent in patients with hippocampal sclerosis (8). Intracranial EEG recordings from within the hippocampus and the amygdala show prominent differences between new and old items (≈400-800 ms) (8-10), further suggesting the MTL as a potential source for the scalp signal. The latencies and nature of these potentials are also in agreement with the average population activity that we have analyzed (SI Fig. 11). We find that the peak activity is within the 500-1000 ms timeframe (SI Fig. 10*B*). Remarkably, the activity peaks first (within the first 500 ms) if recollection fails. If recollection is successful, the peak is in the second bin (500-1000 ms). This suggests that a recognition judgment based purely on familiarity occurs quicker. In addition, it is worth noting that the average population activity we recorded is compatible with the previous intracranial EEG findings but conflicts with BOLD signals obtained by others (11, 12) .

Relationship to fMRI Studies.
This is also in apparent conflict with previous functional Magnetic Resonance Imaging (fMRI) findings (11, 12) that identified regions within the MTL that are selectively activated only for memories that are recollected. Crucially, however, these studies assumed *a priori* that model (*i*) above is correct by searching for brain regions that correlate with the components identified by that model. If model (*i*) is not correct, however, these results are subject to alternative interpretation. Also, these studies used the "remember/know" paradigm to identify memories that were recollected by the subjects. However, this paradigm requires a subjective decision (yes/no) as to whether the memory was recollected or not (as discussed above). It is thus possible that the brain areas identified using these paradigms reflect the decision taken about the memory rather than the retrieval process itself. In our study, no decision as to whether or not recollection succeeded was necessary. Also, our data analysis makes no assumptions about the validity of any particular model.

What is the appropriate baseline activity to consider in the MTL? The MTL is highly active during quiet rest. In fact it is often more active during rest than during memory retrieval (13). Imaging studies can suffer from this undefined baseline and results may vary owing to different choices of representative baseline activity (13). This may also contribute to the apparently disparate findings regarding the involvement of the MTL in recognition memory.

To further investigate the discrepancy between fMRI and single-cell studies, we averaged the neuronal activity of all neurons recorded regardless of their behavioral significance, to approximate a signal that might be similar to an fMRI signal (SI Fig. 10; see *Results*). We found that even under this condition, the overall population activity successfully distinguished between new and old items. The response to old items was not selective for recollected items and was clearly present even if the failed recollected trials were considered separately (SI Fig. 10*B*). Clearly these data differ from previously measured hippocampal BOLD signals (e.g., ref. 11).

SI Methods

Electrophysiology.

All patients were diagnosed with drug resistant temporal lobe epilepsy and implanted with intracranial depth electrodes to record intracranial EEG and single units. Electrodes were placed based on clinical criteria. Electrodes were implanted bilaterally in the amygdala and hippocampus (4 electrodes in total). Each electrode contained 8 identical microwires, one of which we used as ground. We were able to identify single-neurons in the hippocampus and/or amygdala in 9 of the 10 patients. One additional patient was excluded because he had no recognition memory (performance was at chance). Thus, this study is based on 8 patients [6 of which overlap with a previous study (3)]. We recorded a total of 21 retrieval sessions from these eight patients. Four of these sessions (from four different patients) were excluded due to insufficient recognition performance (see below). Thus, this study is based on 17 retrieval sessions from eight different patients. The 17 retrieval sessions were distributed over 16 different days (on one day, two retrieval sessions were conducted). We recorded from 24-32 channels simultaneously (three or four electrodes) and found, on average, 11.9 ±4.4 (±
SD) active microwires (counting only microwires with at least one well separated unit). The average number of identified units per wire was 2.0 ±1.0 (± SD). Inactive wires (no units identified) are excluded from this calculation (77 of 280). There were 130 wires with more than one unit (on average 2.6 ± 0.8 for all wires with >1 unit). For those wires, we quantified the goodness of separation by applying the projection test (14) for each possible pair of neurons. The projection test measures the number of standard deviations the two clusters are separated after normalizing the data such that each cluster is normally distributed with a standard deviation of 1 (see ref. 14 for details). We found that the mean separation of all possible pairs ($n = 315$) is 13.68 ±6.98 (± SD) (SI Fig. 12*A*). We identified, in total, 412 well separated single units. We quantified the quality of the unit isolation by the percentage of all interspike intervals (ISI) that are <3 ms. We found that, on average, 0.3 ± 0.4% of all ISIs were below 3 ms (SI Fig. 12*B*). The signal-to-noise ratio (SNR) of the mean waveforms of each cluster relative to the background noise was on average 2.4 ± 1.2 (SI Fig. 12*C*).

For the purpose of comparing only neurons from the "healthy" brain side (left or right), we excluded all neurons from either the left or right side of the patient if the patients diagnosis (Table S1) included temporal lobe damage (SI Fig. 11). No neurons were excluded if the diagnosis indicated that the seizure focus was outside the temporal lobe.

Behavior.
Each session consisted of a learning and retrieval block. We quantified, for each session, the recognition rate (percentage of old Stimuli correctly recognized), the false positive rate (percentage of new stimuli identified as old) and the recollection rate. The recollection rate was the percentage of stimuli identified as old for which the spatial location was correctly identified. Sessions with a recognition rate of ≤50% were excluded (three sessions). Each session was assigned to either the 24h or 30min delay group.

For each session, we estimated whether spatial recollection rate was significantly different from chance (25%). Due to the small number of trials (maximally 12), the significance was estimated using a bootstrap procedure (see below). Based on this significance value, we further divided each of these two groups into a group with good spatial recollection performance ($P \leq 0.05$, above chance, R^+) and one with poor spatial recollection performance (not significantly different from chance, $P > 0.05$, R^-). For the 24 h group there was only one session with poor recollection performance and thus this analysis was not conducted. Thus, there were 3 behavioral groups that were used for the neuronal analysis: 30 min R^+ ($n = 7$), 30 min R^- ($n = 6$) and 24 h R^+ ($n = 4$). The assignment of sessions to groups was based entirely on behavioral performance. Neuronal activity was not considered.

Data Analysis. *Behavioral*. We labeled each retrieval trial during which a correctly recognized old stimulus was presented as either correctly or incorrectly recollected. For each session we then tested (bootstrap, $P \leq 0.05$, one-tailed, B = 20,000) whether recollection performance was above chance level. We used the bootstrap test instead of the *z* test because of the small number of samples. The resulting p values were more conservative (larger) compared to the p values obtained with the *z* test. Only sessions that passed this test were considered to have "above chance" recollection performance. Trials that failed this test were considered as "at chance". This was to ensure that only neurons from patients that had a clearly demonstrated capability for source memory were included. Also, recording sessions with less than a 50% hit rate for Old stimuli were excluded to ensure that only sessions with sufficient recognition performance were included. We verified for each group of sessions (Fig. 1 *D* and *E*) whether performance was significantly above chance using a *z* test. For this, we pooled all trials of a particular group and labeled each as either correct or incorrect. Then we used one *z*
test to test whether the ratio correct:incorrect was above chance. We used this instead of individual tests for each session to avoid artificially boosting performance due to the small sample size (e.g., 4 out of 12 correct) in each particular session.

Response index. We compared, trial-by-trial, the response (quantified by the response index) to old stimuli that were successfully recollected (R^+) to old stimuli that were not recollected (R^-). For this comparison, trials with recognition errors were excluded (thus, all trials are familiar). The error trials were analyzed separately. There was one data point for every trial for every neuron (e.g., if there are 10 trials and 10 neurons, there are 100 data points). There were 1,368 old stimulus trials (12 retrieval sessions with total 114 neurons), with 1230 trials with a correct recognition response (familiar, TP) and 138 trials that were errors (misses). We analyzed the error trials separately.

We compared the responses of the R^+ and R^- trials with a two-tailed *t* test and using a Kolmogorov-Smirnov test. Both were significant at $P \leq 0.05$. Paired comparisons were made with a *t* test. Normal density functions were constructed by estimating the mean and standard deviation from the data (using maximum likelihood).

Baseline comparison. To determine whether a unit was responsive relative to baseline we compared the firing during the 2s period in which the new vs. old comparison is significant to the 2s period before the stimulus onset. These comparisons were performed using a bootstrap test as described in *Methods* in the main text.

Neuronal ROCs.
Neuronal ROCs (SI Fig. 9) were constructed by considering all trials as old if the response R(i) was above a threshold T. The threshold T was varied in variable steps (see below) from the smallest to the largest value of R(i). Thresholds were varied such that each increase accounted for a 5% quantile of all available datapoints (the 0% and 100% quantile was excluded). This procedure assured that the same number of datapoints was used for the calculation of each point in the ROC. The hit/false positive rate was calculated for each threshold value. d' was calculated for each pair of hit/false positive rates and averaged. We z-transformed the ROC and fit a line through all points using linear regression to find the slope of the curve. A slope of 1.0 indicates that the two distributions (distractors and targets) are of equal variance whereas a slope of unequal 1.0 indicates a difference in variance. The z transformed ROC was fit well by a straight line for both R^+ and R^- trials (1).

Population Averages.
Population averages (SI Figs. 6 and 10) were constructed by normalizing each trial to the mean baseline firing in the 2 s before stimulus onset. The number of spikes were binned into 1s bins (non-overlapping) and averaged for all neurons. No smoothening was applied. To avoid normalization artifacts, only neurons with a baseline rate of at least 0.25 Hz were considered for the population averages (346 of 412 neurons for SI Fig. 10). Also, for SI Fig. 6 only neurons with a significant response in the stimulus period (first two of the 2-s bins) were considered (this does not apply for the trial-by-trial analysis).

Decoding. We used a linear classifier to estimate how well the firing of a single neuron during a single trial can signal the identity (new or old) of the presented stimulus. The classifier was provided with the number of spikes fired in three consecutive 2-s bins after stimulus onset (0-2 s, 2-4 s, 4-6 s). The classifier consisted of a weighted sum of these three numbers. The weights were estimated using regularized least squares (RLSC) (15, 16). This method is equal to multiple linear regression with the exception of an added regularizer term λ(see below; we used λ = 0.01 throughout). The decoding accuracy of the classifier was estimated using leave-one-out cross-validation for all training samples available. The estimated prediction error was equal to the percentage of correct leave-one-out trials. There were maximally 12 samples in each class (old or new). However, due to behavioral errors, fewer trials were sometimes available for analysis. Error rates for false positives and false negatives were approximately equal and the number of samples was thus approximately balanced in both classes. Of concern was whether a slight imbalance of the number of samples in one class could bias the results. We performed two controls to assess whether this was the case: we performed leave-one-out cross-validation with the label of the test sample randomly reassigned with 50% probability. If the classifier was biased, the resulting error would be different from 50%. We found that this was not the case (Fig. 8*A*). Also, we reran all analysis that used the decoder with a balanced number of samples (that is, equal number of samples in either class) and found no difference in the results.

The weights were determined by regularized least squares. Regularized least squares are very similar to multiple linear regression. In the following we would like to point out these differences because in a previous study we used a multiple linear regression (3).

With multiple linear regression (Eq. **S1**), the weights w are determined by multiplying the inverse of data samples Z with the trainig labels y (17).

$$w = \left[Z'Z\right]^{-1}Z'y \quad \textbf{[S1]}$$

In contrast, in regularized least squares (15, 16, 18), an additional term is added to the data samples (Eq. **S2**). Here, I is the identity matrix and λ is a scalar parameter (the regularizer).

$$w = \left[Z'Z + \lambda I\right]^{-1}Z'y \quad \textbf{[S2]}$$

The value of the regularizer is arbitrary. The bigger it is, the more constraints are placed on the solution (the less the solution is determined by the data samples). A small value of the regularizer, however, makes the solution close to the multiple linear regression solution. Importantly, however, even a small value of the regularizer punishes unrealistically large weights and also guarantees full rank of the data matrix. Regularization becomes particularly important when there are a large number of input variables relative to the number of training samples. This is the case in our study because each neuron contributed 3 variables (3 ×
2 s time periods) and the number of training samples was small (on the order of 10). Thus, regularization was necessary. We found that performance was maximal for a small (but non-zero) regularizer and used $\lambda = 0.01$ throughout.

1. Macmillan NA, Creelman CD (2005) *Detection Theory* (Lawrence Associates, Mahwah, NJ).

2. Ratcliff R, Gronlund SD, Sheu CF (1992) *Psychol Rev* 99:518-535.

3. Rutishauser U, Mamelak AN, Schuman EM (2006) *Neuron* 49:805-813.

4. Cameron KA, Yashar S, Wilson CL, Fried I (2001) *Neuron* 30:289-298.

5. Wirth S, Yanike M, Frank LM, Smith AC, Brown EN, Suzuki WA (2003) *Science* 300:1578-1581.

6. Sakai K, Miyashita Y (1991) *Nature* 354:152-155.

7. Rugg MD, Mark RE, Walla P, Schloerscheidt AM, Birch CS, Allan K (1998) *Nature* 392:595-598.

8. Grunwald T, Lehnertz K, Heinze HJ, Helmstaedter C, Elger CE (1998) *Proc Natl Acad Sci USA* 95:3193-3197.

9. Mormann F, Fell J, Axmacher N, Weber B, Lehnertz K, Elger CE, Fernandez G (2005) *Hippocampus* 15:890-900.

10. Smith ME, Stapleton JM, Halgren E (1986) *Electroencephalogr Clin Neurophysiol* 63:145-159.

11. Eldridge LL, Knowlton BJ, Furmanski CS, Bookheimer SY, Engel SA (2000) *Nat Neurosci* 3:1149-1152.

12. Yonelinas AP, Otten LJ, Shaw KN, Rugg MD (2005) *J Neurosci* 25:3002-3008.

13. Stark CEL, Squire LR (2001) *P Natl Acad Sci USA* 98:12760-12765.

14. Rutishauser U, Schuman EM, Mamelak AN (2006) *J Neurosci Methods* 154:204-224.

15. Evgeniou T, Pontil M, Poggio T (2000) *Adv Comput Math* 13:1-50.

16. Rifkin R, Yeo G, Poggio T (2003) in *Advances in Learning Theory: Methods, Model and Applications*, ed. Suykens JAK (IOS Press, Amsterdam), pp 131-146.

17. Johnson RA, D.W., W (2002) *Applied Multivariate Statistical Analysis* (Prentice Hall, New York).

18. Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) *Science* 310:863-866.