J Neurophysiol 100: 1407–1419, 2008.
First published June 18, 2008; doi:10.1152/jn.90248.2008.

Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex

Ethan M. Meyers,[1,2] David J. Freedman,[3,4] Gabriel Kreiman,[2,5] Earl K. Miller,[1,3] and Tomaso Poggio[1,2]

[1]*Department of Brain and Cognitive Sciences and* [2]*The McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts,* [3]*The Picower Institute for Learning and Memory,* [4]*Department of Neurobiology, The University of Chicago, Chicago, Illinois; and* [5]*Department of Ophthalmology and Program in Neuroscience, Children's Hospital Boston, Harvard Medical School, Massachusetts*

Submitted 8 February 2008; accepted in final form 14 June 2008

Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J Neurophysiol* 100: 1407–1419, 2008. First published June 18, 2008; doi:10.1152/jn.90248.2008. Most electrophysiology studies analyze the activity of each neuron separately. While such studies have given much insight into properties of the visual system, they have also potentially overlooked important aspects of information coded in changing patterns of activity that are distributed over larger populations of neurons. In this work, we apply a population decoding method to better estimate *what* information is available in neuronal ensembles and *how* this information is coded in dynamic patterns of neural activity in data recorded from inferior temporal cortex (ITC) and prefrontal cortex (PFC) as macaque monkeys engaged in a delayed match-to-category task. Analyses of activity patterns in ITC and PFC revealed that both areas contain "abstract" category information (i.e., category information that is not directly correlated with properties of the stimuli); however, in general, PFC has more task-relevant information, and ITC has more detailed visual information. Analyses examining *how* information coded in these areas show that almost all category information is available in a small fraction of the neurons in the population. Most remarkably, our results also show that category information is coded by a nonstationary pattern of activity that changes over the course of a trial with individual neurons containing information on much shorter time scales than the population as a whole.

INTRODUCTION

The concept of population coding, in which information is represented in the brain by distributed patterns of firing rates across a large number of neurons, arguably dates back over 200 years (McIlwain 2001). Yet, despite this long conceptual history, and an extensive amount of theoretical work on the topic (Rumelhart et al. 1986; Seung and Sompolinsky 1993; Zemel et al. 1998), most electrophysiological studies still examine the coding properties of each neuron individually.

While much insight has been gained from studies analyzing the activity of individual neurons, these studies can potentially overlook or misinterpret important aspects of the information contained in the joint influence of neurons at the population level. For example, many analyses make inferences about *what* information is in a given brain region based on the *number* of responsive neurons or on the strength of index values that are *averaged* over many individual neurons. However, much theoretical and experimental work (Olshausen and Field 1997; Rolls and Tovee 1995) has indicated that information can be coded in sparse patterns of activity. Under a sparse representation, a brain region that contains fewer responsive neurons during a particular task might actually be more involved in the use of that information, and averaging over many neurons might dilute the strength of the effect, which could give rise to a misinterpretation of the data.

Another shortcoming of most single-neuron analyses is that they do not give much insight into *how* information is coded in a given brain region. Several theoretical studies have examined how information is stored in ensembles of units including attractor networks, synfire chains (Abeles 1991) and probabilistic population codes (Zemel et al. 1998) among others. However, because of the paucity of population analyses of real neural data, there is currently little empirical evidence on which to judge the relative validity of these models.

To better understand the content and nature of information coding in ensemble activity, we used population decoding tools (Duda et al. 2001; Hung et al. 2005; Quiroga et al. 2006; Stanley et al. 1999) to analyze the responses of multiple individual neurons in inferior temporal cortex (ITC) and prefrontal cortex (PFC) recorded while monkeys engaged in a delayed match-to-category task (DMC) (Freedman et al. 2003). Previous individual neuron analyses of these data had suggested that ITC is more involved in the processing of currently viewed image properties, whereas PFC is more involved in signaling the category and behavioral relevance of the stimuli and in storing such information in working memory (Freedman et al. 2003). Here, by pooling the activity from many neurons, we are able to achieve a finer temporal description of the information flow, and we can better quantify how much of the category information in these areas is due to visual properties of the stimuli versus being more abstract in nature. Additionally, by looking at the activity in a population over time, we find that the selectivity of those neurons that contain abstract category information changes rapidly. Information is being continually passed from one small subset of neurons to another subset over the course of a trial. This work not only clarifies the roles of ITC and PFC in visual categorization, but it also helps to constrain theoretical models on the nature of neural coding

Address for reprint requests and other correspondence: E. Meyers, Dept. of Brain and Cognitive Sciences, MIT, Bldg. 46-5155, 43 Vassar St., Cambridge, MA, 02139 (E-mail: emeyers@mit.edu).

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

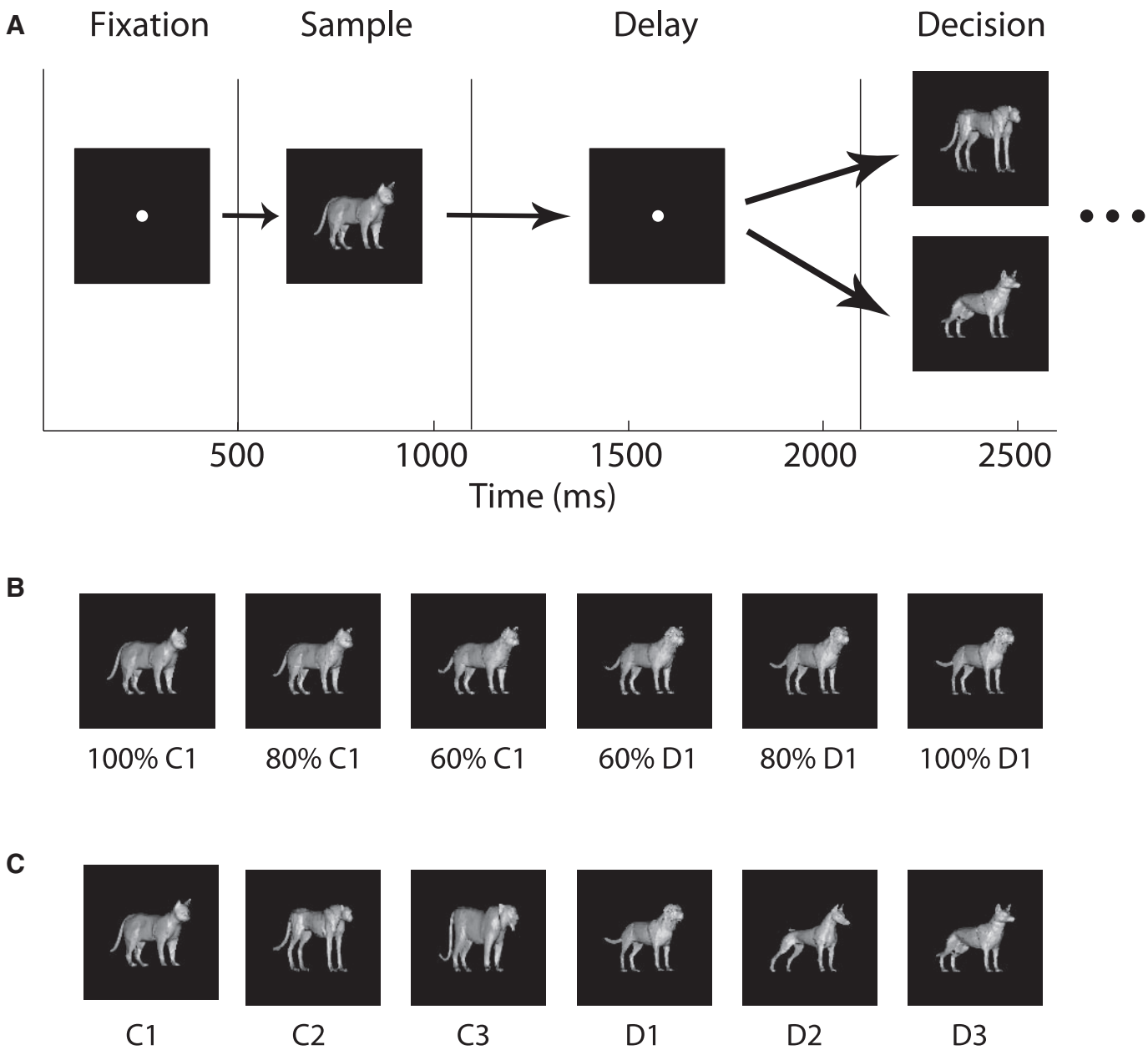

FIG. 1. Organization of the stimuli and behavioral task. *A*: time course of the delayed match to category experiment. *B*: an example of 1 of the 9 morph lines of the stimuli from the cat 1 prototype to the dog 1 prototype (the actual stimuli used in the experiment were colored orange) (see Freedman et al. 2002). *C*: the 6 prototype images used in the experiment. All the stimuli used in the experiment were either the prototype images, or morphs between the cat (*C*) and dog (*D*) prototypes.

in these structures (Riesenhuber and Poggio 2000; Serre et al. 2005).

METHODS

Behavioral task and recordings

We used the data recorded in the study of Freedman et al. (2003). Briefly, responses of 443 ITC and 525 PFC neurons were recorded from two Rhesus Macaque monkeys as the monkeys engaged in a delayed match-to-category task. Each DMC trial consisted of a sequence of four periods: a fixation period (500-ms duration), a sample period in which a stimulus was shown (600-ms duration), a delay period (1,000 ms), and a decision period in which a second stimulus was shown and the monkey needed to make a behavioral decision (Fig. 1*A*). The stimuli used in the task were morphed images generated from three prototype images of cats and three prototype images of dogs (Fig. 1, *B* and *C*). A morph stimulus was labeled a "cat" or "dog" depending on the category of the prototype that contributed >50% to its morph. During the sample period of the task, a set of 42 images (Fig. S1[1]) were used that consisted of the six prototype images and morphs that were taken at four even intervals between each dog and cat prototype. The stimuli shown in the decision period consisted of random morphs that were ≥20% away from the cat/dog category boundary, so that the category that these stimuli belonged to was unambiguous. The monkeys needed to release a lever if the sample-stimulus matched the category of the decision-stimulus to receive a

[1] The online version of this article contains supplemental data. Additional information can be found at http://cbcl.mit.edu/emeyers/jneurophys2008.

juice reward (or to continue to hold the lever and release it for a second decision-stimulus in the nonmatch trials). Performance on the task was ~90% correct. Figure 1 illustrates the time course of an experimental trial, one morph line used in the experiment, and the six prototype dog and cat images. The experimental design and recordings were previously reported by Freedman et al. (2001, 2003); and more details about the stimuli, the task, and the recordings can be found in those publications.

DATA ANALYSIS. To estimate the information conveyed by a neuronal ensemble about a particular stimulus or behavioral variable, we used a decoding-based approach (Hung et al. 2005; Quiroga et al. 2006). We trained a pattern classifier on the firing rates from a population of m neurons recorded across k trials (i.e., we have k training points in R^m, where R^m is an m-dimensional vector space). For each trial, one of c different conditions is present, and the classifier "learns" which pattern of activity across the m neurons is indicative that condition c_i was present. We assessed how much information is present in the population of neurons by using a "test data set" (firing rates from the same m neurons, but from a *different* set of h trials) and quantifying how accurately the classifier could predict which condition c_i was present in these new trials. Classifier performance was evaluated and reported throughout the text as the percentage of test trials correctly labeled. In the text we use the terms "decoding accuracy" and "information" interchangeably because there is an injective monotonic mapping between these two measures (Gochin et al. 1994; Samengo 2002). Variables (i.e., different groups of conditions) we decoded include *1*) which of the 42 stimuli was shown during the sample period ($c = 42$), *2*) the category of the stimulus shown during the sample period ($c = 2$), *3*) the category of the stimulus shown during the decision period ($c = 2$), and *4*) whether a trial was a match or nonmatch ($c = 2$). Occasionally, in the text we are informal and we say we trained a classifier on a given set of images X, by which we mean we trained the classifier on neural data that was recorded when images in set X were shown.

Because most of the neurons used in these analyses were recorded in separate sessions, it was necessary to create pseudo-populations that could substitute for simultaneous recordings. Although creating these pseudo-populations ignores correlated activity between neurons that could potentially change estimates of the absolute level of information in the population (Averbeck et al. 2006), having simultaneous recordings would most likely not change the conclusions drawn from this work because we are mainly interested in *relative* comparisons over time and between brain regions.

To create this pseudo-population for the decoding of identity information (i.e., which of the 42 stimuli were shown during the sample period) the following procedure was used. First we eliminated all neurons that had nonstationary trends (those with an average firing rate variance in 20 consecutive trials was less than half the variance over the whole session). Because the stimuli were presented in random order, the average variance in 20 trials should be roughly equivalent to the variance over the whole session (only 42 ITC and 34 PFC neurons met the trend criterion, and the decoding results were not significantly different when these neurons were included). Next we found all neurons that had recordings from at least five trials for each of the 42 stimuli shown in the sample period. This left 283 ITC and 332 PFC neurons for further analysis. From the pools of either ITC neurons or PFC neurons we applied the following procedure separately at each time period.

First, 256 neurons were randomly selected from the pool of all available neurons. This allowed a fair comparison of ITC to PFC even though there were more neurons available in the PFC pool.

Second, for each neuron, we randomly selected the firing rates from five trials for each of the 42 stimuli.

Third, the firing rates of the 256 neurons from each of the five trials were concatenated together to create 210 data points (5 repetitions × 42 stimuli) in R^{256} space.

Fourth, a cross-validation procedure was repeated five times. In each repetition, four data points from each of the 42 classes were used as training data and one data point from each class was used for testing the classifier (i.e., each data point was only used once for testing and 4 times for training). Prior to training and testing the classifier, a normalization step was applied by subtracting the mean and dividing by the SD for each neuron (the means and SD were calculated using only the data in the training set). This z-score normalization helped ensure that the decoding algorithm could be influenced by all neurons rather than only by those with high firing rates. Similar results were obtained when this normalization was omitted.

Fifth, the whole procedure from steps 1–4 was repeated 50 times to give a smoothed bootstrap-like estimate of the classification accuracy. The main statistic shown in Figs. 2–7 is the classification accuracy averaged over all the bootstrap and cross-validation trials.

A similar procedure was used to create pseudo-population vectors for decoding of sample-stimulus category, decision-stimulus category, and match-nonmatch information as shown in Fig. 2, except that 50 data points for each class were used in each of the five cross-validation splits (i.e., there were 400 training points and 100 test points), and the trial condition labels were changed to reflect the information that we were trying to decode. For the decoding of "abstract category" information in Figs. 3–7, the procedure was used exactly as described in the preceding text except that the 42 identity labels were remapped to their respective dog and cat categories, and different prototypes were used for training and testing (see section on decoding abstract category information).

Unless otherwise noted, all figures that show smooth estimates of classification accuracy as a function of time are based on using firing rates in 150-ms bins sampled at 50-ms intervals with data from each time bin being classified independently. Because the sampling interval we used is shorter than the bin size (50-ms sampling interval, 150-ms time bin), the mean firing rates of adjacent points were calculated using some of the same spikes, leading to a slight temporal smoothing of the results.

In the body of the text, we also report classification accuracy statistics. Unless otherwise stated, classification accuracy results from the sample periods are reported for bins centered at 225 ms after sample stimulus onset, results from the delay period are reported for 525 ms after sample stimulus offset, and results from the decision period are reported for 225 ms after decision stimulus offset (this corresponds to 725, 1,625, and 2,325 ms after the start of a trial, with each bin width being 150 ms). The results reported for "basic" decoding accuracies are the means ± 1 SD of the decoding accuracies over all the bootstrap trials and cross-validation splits. The results reported for decoding abstract category information are the average ±1 SD of basic decoding results taken over the nine combinations of training and test splits (see the section on decoding abstract category information for more details). Also, because there are two stimuli presented in each trial, to avoid confusion when reporting basic decoding results, we denote the first stimulus shown as the SAMPLE-STIMULUS and the second stimulus shown as the DECISION-STIMULUS with capitalized letters used to avoid confusion with the sample, delay, and decision periods (which are time periods where properties of these stimuli can be decoded). It should be noted that in this paper, we refer to the time period after the second stimulus is shown as the decision period rather than the test period as used by Freedman et al. (2003) to avoid confusion with the test set that is used to evaluate the trained classifier.

All results reported in this paper use a correlation coefficient-based classifier. Training of this classifier consists of creating c "classification vectors" (where c is the number of classes/conditions used in the analysis), and each classification vector is simply the mean of all the training data from that class (thus each classification vector is a point in R^m, where m is the number of neurons). To assess to which class a test point belongs, the Pearson's correlation coefficient is calculated

between the test point and each classification vector; a test data point is classified as belonging to the class c_i, if the correlation coefficient between the test point and the classification vector of class c_i is greater than the correlation coefficient between the test point and the classification vector of any other class. The classification accuracy reported is the percentage of correctly classified test trials.

There are several reasons why we use a correlation coefficient-based classifier. First, because this is a linear classifier, applying the classifier is analogous to the integration of presynaptic activity through synaptic weights; thus decoding accuracy can be thought of as indicative of the information available to the postsynaptic targets of the neurons being analyzed. Second, computation with this classifier is fast, and it has empirically given classification accuracies that are comparable to more sophisticated classifiers such as regularized least squares, support vector machines and Poisson naïve Bayes classifiers, which we have tested on this and other data sets (see Supplemental Fig. S2). Third, this classifier is invariant to scalar addition and multiplication of the data, which might be useful for comparing data across different time periods in which the mean firing rate of the population might have changed. And finally, this classifier has no free adjustable parameters (that are not determined by the data) which simplifies the training procedure.

For several analyses, we trained a classifier on one condition and tested the classifier on a different related condition. These analyses test how invariant the responses from a population of neurons are to certain transformations, and they help to determine whether a population of neurons contains information beyond what is directly present in the stimulus itself. We also performed analyses in which a classifier is trained with data from one time period and tested with data from a different time period; this allowed us to assess whether a pattern of activity that codes for a variable at one time period is the same pattern of activity that codes for the variable at a later time period. It is important to emphasize that for *all* analyses, training and test data come from different trials. Finally, for several analyses, we calculated the classification accuracy using only small subsets of neurons, ranked based on how category-selective these neurons were. The rank order was based on a *t*-test applied to all cat trials versus all dog trials on the training dataset, and the k neurons with the smallest P values were used for training and testing. This "greedy" method of feature selection is not guaranteed to return the smallest subset that will achieve the best performance, so the readout accuracies obtained with this feature selection method might be an underestimate of what could be obtained with an equivalent number of neurons from the same population if an ideal feature selection algorithm was applied.

Finally, for one set of analyses (Fig. 8), we estimated the amount of mutual information (MI) between the category of the stimuli s and individual neurons' firing rates r, using the average firing rates in 100-ms bins sampled at 10-ms intervals. To compute the mutual information, we assumed the prior probability of each stimulus category was equal, and we used the standard formula, $I = \Sigma_{s,r} P[r, s] \log_2 (P[r, s]/P[r] P[s])$ (Dayan and Abbott 2001). The joint probability distribution between stimulus and response, $P[r, s]$, was estimated from the empirical distribution using all trials. Although there exists potentially more accurate methods for estimating mutual information (Paninski 2003; Shlens et al. 2007), because our results do not depend critically on the exact MI values, we preferred the simplicity of this method.

Additional material can be found at http://cbcl.mit.edu/people/emeyers/jneurophys2008/.

RESULTS

Decoding information content in ITC and PFC

BASIC RESULTS. We used a statistical classifier to decode information from neuronal populations that were recorded as monkeys engaged in a delayed match-to-category task (Fig. 1*A*) (Freedman et al. 2003). Figure 2 shows the accuracy levels obtained when decoding four different types of information. The decoding of identity information (i.e., which of the 42 stimuli was shown during the sample period) is shown in Fig. 2*A* and provides an indication of how much detailed visual information is retained despite the variability in spike counts that occur from trial to trial. Given the high physical similarity among the images along a given morph line (Fig. 1*B*), this is a very challenging task. There was a significant amount of information only during the sample period when the stimulus was visible, and there was much more information in ITC than in PFC (17.5 ± 5.5 vs. 5.9 ± 3.5% respectively, chance = 1/42 = 2.4%). Because information about the details of the visual stimuli was not relevant for the task in which the monkey was engaged, these results are consistent with the notion that ITC is involved in the detailed analysis of the visual information that is currently visible, whereas PFC activity only contains the information necessary for completing the task (Freedman et al. 2001; Riesenhuber and Poggio 2000).

Next we examined decoding the category of the SAMPLE-STIMULUS (i.e., whether the stimulus shown at the beginning of the sample period was a cat or a dog, Fig. 2*B*). When the SAMPLE-STIMULUS was first presented, ITC had a slightly higher accuracy level than PFC (92.0 ± 2.8 vs. 81.3 ± 4.3%, at t = 225 ms, chance = 50%). However, by the middle of the sample period (t = 425 ms after stimulus onset), the information in these two areas was approximately equal (82.1 ± 4.0 vs. 82.0 ± 4.2%). During the delay and decision periods, PFC had more category information about the SAMPLE-STIMULUS than ITC [delay: 66.7 ± 4.1% (PFC) vs. 56.6 ± 4.8% (ITC); decision: 88.4 ± 4.3% (PFC) vs. 77.9 ± 4.4% (ITC), chance = 50%]. Because category information is behaviorally relevant to the monkey in this task, these results support the role of the PFC in storing task-relevant information in memory during the delay period (Miller and Cohen 2001). That ITC initially had more information about the category of the SAMPLE-STIMULUS is largely due to ITC having more information related to visual properties of the stimuli, and this visual information is being used by the classifier to decode the category of the stimuli (see section on decoding abstract category information in the following text).

Figure 2*C* shows accuracy levels from decoding the category of the DECISION-STIMULUS (i.e., the stimulus that is presented in the beginning of the decision period). ITC had slightly more information about the category of the DECISION-STIMULUS than PFC during the decision period (93.9 ± 2.7 vs. 81.1 ± 4.3%). This is probably due to the combination of visual and abstract category information by the classifier and because there is more visual information in ITC the performance level is higher there. In contrast, PFC showed higher accuracy levels when decoding whether a trial was a match or nonmatch trial during the decision period (92.3 ± 2.7 vs. 60.5 ± 4.8% Fig. 2*D*), which is again consistent with PFC containing more task-relevant information than ITC.

In addition to comparing ITC to PFC, it is also instructive to directly compare different types of information within each of these areas. Figure 2, *E* and *F*, compares the decoding accuracies for three different variables: whether a trial is a match/nonmatch trial (brown), the category of the DECISION-STIMULUS (green), and the category of the SAMPLE-STIMULUS (purple) (we start the comparison in the middle of the delay period

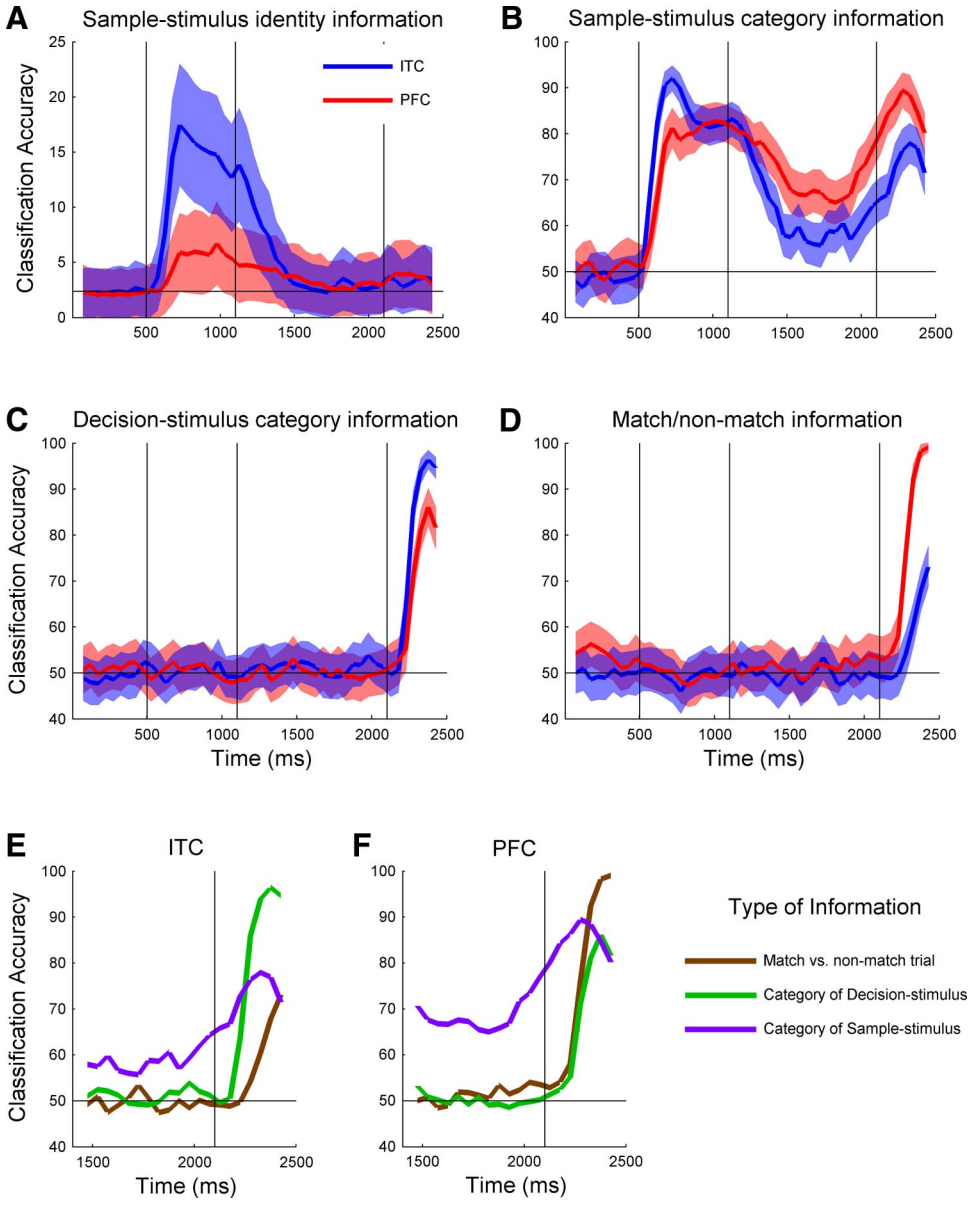

FIG. 2. Basic decoding results for 4 different types of information. *A–D*: blue lines indicates results from inferior temporal cortex (ITC) and red lines indicate results from prefrontal cortex (PFC; red, and blue shaded regions indicate one SD over the bootstrap-like trials). The 3 vertical black lines indicate SAMPLE-STIMULUS onset, SAMPLE-STIMULUS offset, and DECISION-STIMULUS onset from left to right respectively. *E* and *F*: comparison of SAMPLE-STIMULUS category decoding accuracy (purple), DECISION-STIMULUS category decoding accuracy (green), and whether a trial is a match or nonmatch trial (brown) for ITC (*E*) and PFC (*F*).

because there is no information about trial status and DECISION-STIMULUS category until the decision period). Results from ITC (Fig. 2*E*) reveal that during the decision period, there is much more information about the category of the DECISION-STIMULUS (green line) than about the category of the SAMPLE-STIMULUS (purple line) or about whether a trial is a match or nonmatch trial (brown). Also the match/nonmatch trial information showed the longest latency. This pattern shows that the variable that ITC has the most information about (of the 3 variables listed in the preceding text) is the most recently viewed visual stimulus and that there is less information about task-related variables. The pattern in PFC is quite different (Fig. 2*F*), with the most information being about task-related variables; i.e., whether a trial is a match or nonmatch trial. Also the latency of the match/nonmatch status of a trial in PFC is the same as the latency of information about the category of the DECISION-STIMULUS (and shorter than the ITC latency in the same task). It is also interesting to note that for both PFC and for ITC, the information about the category of SAMPLE-STIMULUS seems to increase just *prior* to the onset of the DECISION-STIMULUS presentation. This anticipatory increase of information might subserve the quick reaction times seen in the experiment.

ABSTRACT CATEGORY INFORMATION. From a cognitive science perspective, a category often refers to a grouping of objects based on their behavioral significance, and objects within such a group do not necessarily share any common physical characteristics (Tanaka 2004). In Fig. 2*B*, however, the decoding accuracy level for the category of the SAMPLE-STIMULUS is influenced not only by the "abstract" behaviorally relevant category of the stimulus but also by physical visual properties of the image that are also predictive of the category that the stimulus belongs to (see Supplemental Fig. S3 for more details). To better assess how much abstract category information is in ITC and PFC that is related to the behavioral grouping of the stimuli (and that not due to physical properties of the stimuli), we trained a classifier on images derived from two

dog prototypes and two cat prototypes and then tested the classifier's decoding accuracy on images derived from the remaining dog and cat prototypes (by "derived from a prototype," we mean the images that contain ⩾60% of their morph from a given prototype). The logic behind this analysis is that if the within-category prototype images were just as visually similar to each other as they are to the between-category prototype images, then using different prototypes for training and testing should eliminate the ability of visual feature information to be predictive of which class a stimulus belongs to (because there would be as many visual features shared between the training and test sets within the same category, as there are between the two different categories; see Supplemental Fig. S3). Thus obtaining above chance classification performance in this analysis would imply that a brain region had more abstract category information. While determining the visual similarity between two images is currently an ill-defined problem, we note that the prototype images used in this experiment did vary greatly in their visual appearance (Figs. 1*C* and S1). Therefore this decoding method should greatly reduce the influence of visual features (see DISCUSSION for more details on image similarity). In fact, because many of the images used to test the classifier were morphs that were blended with prototype images from the opposite category, images from opposite categories were more similar in terms of the morph coefficients than images from the same category (similar results were obtained when we did not use images that were morphs between the training and test set prototypes; see Supplemental Fig. S4*B*).

Figure 3*A*, shows the decoding results of this more abstract category information for ITC (blue) and PFC (red) averaged over all nine training/test permutations [e.g., train on (c1, c2 vs. d1, d2) test on (c3, d3); training on (c1, c2 vs. d1, d3) tested on (c3, d2) etc.]. Supplemental Fig. S4*A* shows the results for the nine individual runs for both PFC and ITC; all individual results are the average of 50 bootstrap-like trials. During the sample period when the stimuli are first shown, PFC has as much abstract category information as ITC. During the delay and decision periods, PFC has more category information than ITC. This strongly suggests that the larger amount of category information in ITC during the sample period seen in Fig. 2*B* is due to the classifier combining category information in a visually based format with information in a more abstract format.

Figure 3, *B* and *C* compare the visual plus abstract category information (purple trace) that was shown in Fig. 2*B* with the abstract category information (orange trace) that was shown in Fig. 3*A*, for ITC (*B*) and PFC (*C*). For ITC, most of the category information during the sample period is visual; however, during the delay and decision periods, almost all the category information is abstract. PFC shows a similar pattern; however, there is more abstract category information (and less visual category information) during the sample period than for ITC. Thus both ITC and PFC have category information in a visual format while the stimulus is visible, and both represent information in an abstract, task-relevant format during the delay and decision period. However, the overall ratio of abstract category information relative to total category information is greater in PFC than in ITC during the sample period.

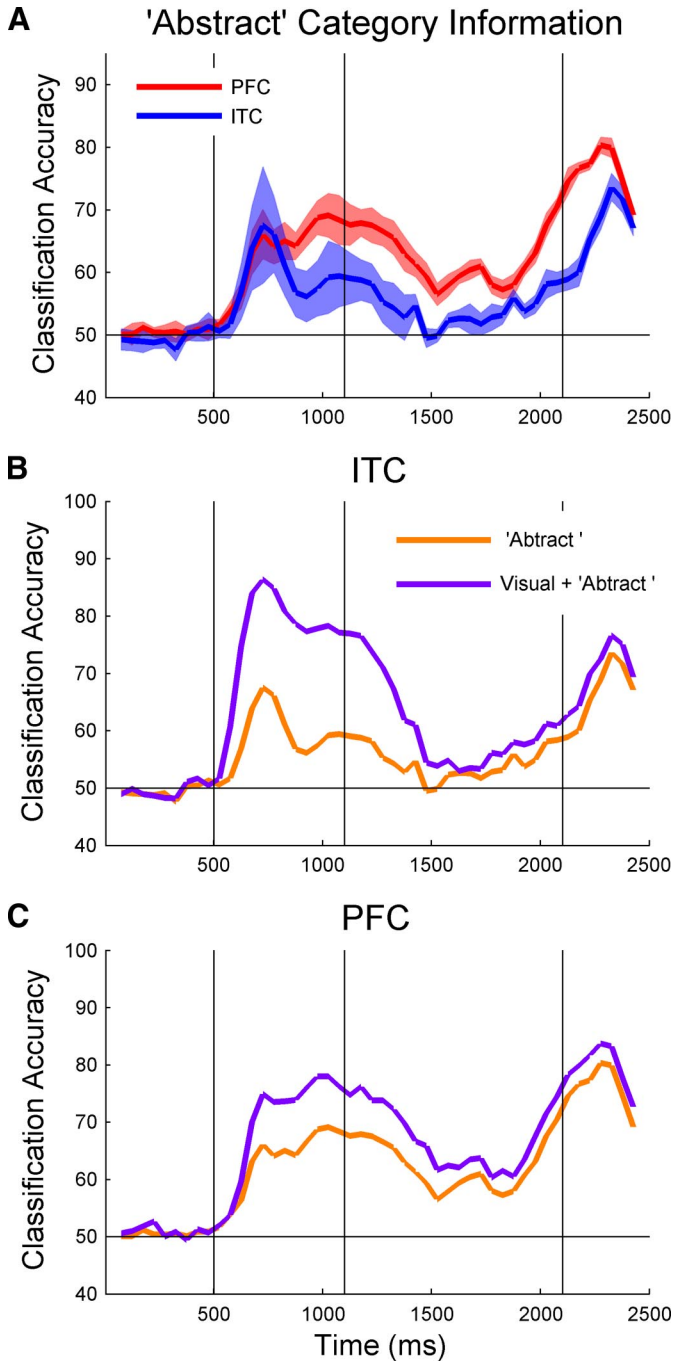

FIG. 3. Decoding task-relevant "abstract" category information. *A*: decoding accuracies for ITC (blue) and PFC (red) when training on data from 2 dog and 2 cat prototype images and testing on the remaining dog and cat prototype images. The results are the average over all 9 permutations of training/test splits and the shaded results show the SDs over the 9 permutations (the individual traces are shown in Supplementary Fig. S4*A*). *B* and *C*: comparison of visual plus category stimulus decoding accuracies (purple line) to abstract category information (orange line) for ITC (*B*) and PFC (*C*). Note that there is a larger difference between these two types of information in ITC compared with the difference between these information types seen in PFC. This is a strong indication that the high SAMPLE-STIMULUS category decoding accuracies seen in ITC in Fig. 2*B* are largely due to visual information and not abstract category information during the sample period. During the decision period, for both ITC and PFC, most of information about the category of the SAMPLE-STIMULUS is in a more abstract representation, as there is little difference between abstract category information and "basic" category information during this period.

Coding of information in ITC and PFC

COMPACT AND REDUNDANT INFORMATION. In addition to assessing *what* information is contained in ITC and PFC, the decoding analysis also allows us to examine *how* information is coded across a population of neurons. One important question of neural coding concerns whether information is contained in a widely distributed manner such that all neurons are necessary to represent a stimulus or if at a particular point in time, there is a smaller "compact" subset of neurons that contains all the information that the larger population has (Field 1994). To assess if there is a smaller compact subset of neurons in ITC and PFC conveying as much information as the larger population using population decoding, we first selected the "best" k neurons using the training data (where $k < 256$) and then trained and tested our classifier using only these neurons (Fig. 4). The best k neurons were defined as those neurons with the smallest P values based on a t-test applied to all cat-trials versus all dog-trials on the training data set (see METHODS). The selection process was done separately for each time bin. Using the 16 best neurons, we were able to extract almost all the information that was available using 256 neurons at almost all time points for both PFC and ITC. The level of compactness of information was particularly strong in PFC during the decision period

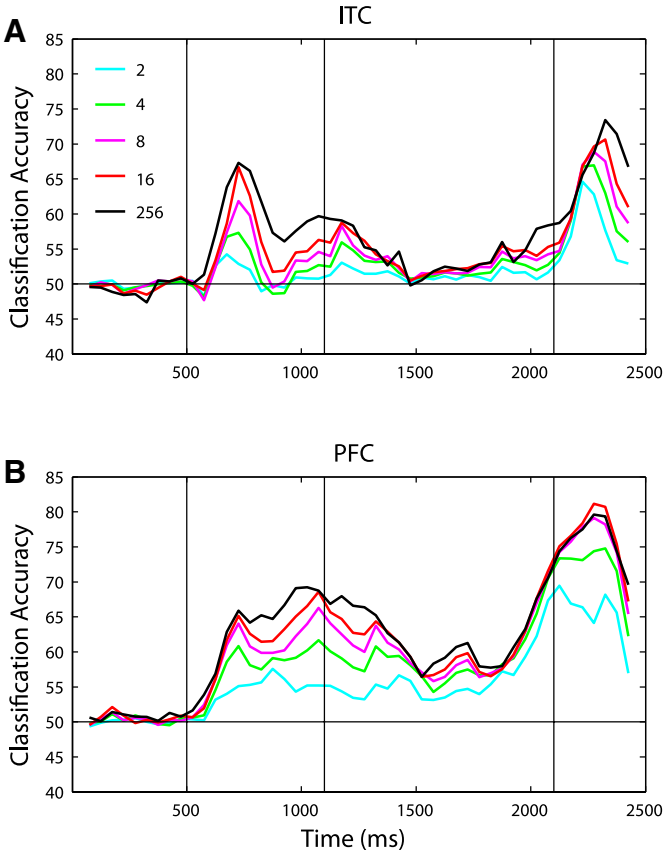

FIG. 4. Readout using the "best" 2, 4, 8, or 16 neurons, compared with readout using all 256 neurons, for ITC (*A*) and PFC (*B*). As can be seen for almost all time periods, the abstract category information available in whole population is available in only ≤16 neurons. The best neurons were determined based on t-test between cats and dogs using the training data. Because the algorithm used to select the best neurons works in a greedy manner and is not necessarily optimal, the information reported in the subsets of neurons is an underestimate of how much information would be present if the optimal k neurons were selected.

where, strikingly, eight neurons contained nearly all the information (decoding accuracy = 78.2 ± 1.2%) that was available in the whole population (79.4 ± 1.7%). It should also be noted that because our algorithm for selecting the best neurons works in a greedy fashion, the top k neurons selected might not be the best k neurons available *in combination*. Therefore all the information present in the entire population could potentially be contained in even fewer neurons. We also examined if there is a smaller subset of neurons that contains all the identity information (Supplemental Fig. S5) and found that for ITC, identity information seems to be less compact with the decoding accuracy not saturating until around 64 neurons. We speculate that this might be related to the fact that it takes more bits of information to code 42 stimuli than to code the binary category variable and also perhaps because identity information is not relevant for the task the monkey is engaged in.

Redundancy allows a system to be robust to degradation of individual neurons or synapses. This robustness constitutes a key feature of biological systems. To asses if there is redundant information present in the population of neurons, we again selected the k best neurons from the training set, but this time we excluded these neurons from training and testing and used the remaining $256 - k$ neurons for our analyses. We note that this analysis aims to assess whether there is redundant information (as opposed to estimating how much redundant information there is in the Shannon sense of redundancy). Figure 5 compares the classifier's performance using the best 64 neurons to its performance excluding the best 64 neurons. The best 64 neurons contain as much information as the whole population (magenta line). However, even when these best 64 neurons are excluded, and the remaining 192 neurons are used instead, classification performance is above chance at almost all time points (green line). Because the best 64 neurons contain as much information as the whole population, the fact the excluding these neurons does not lead to chance classification performance implies that these remaining 192 neurons contain a nonnegligible amount of redundant information with the best 64 neurons. In fact, even when half the neurons are removed, decoding accuracy is still above chance at almost all time points (Supplemental Fig. S6).

TIME-DEPENDENT CODING OF INFORMATION. Another interesting question in neural coding is whether a given variable is coded by a single pattern of neural activity in a population, as in a point attractor network (Hopfield 1982), or whether there are several patterns that each code for the same piece of information (Laurent 2002; Perez-Orive et al. 2002). To address this question, we trained a classifier with data from one time bin relative to stimulus onset and tested the classifier on data from different time bins (in all the results reported in the preceding text, training and testing were done using the same time period relative to stimulus onset). If, at all time periods, the same pattern of activity is predictive of a particular variable, then the decoding accuracy should always be highest (or at least should not decrease) when training a classifier with data from time periods that have the maximum decoding accuracy levels because the data from these time periods presumably have the least noise and would therefore lead to the creation of the best possible classifier. Alternatively if the pattern of activity that is indicative of a relevant variable changes with time (and is time-locked to the onset of a stimulus/trial), then high decoding

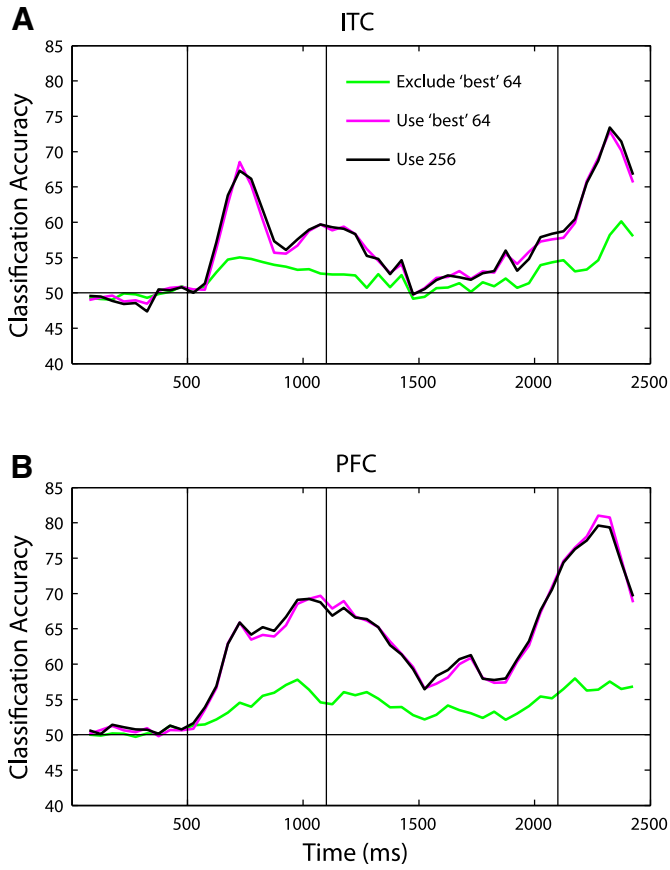

FIG. 5. Illustration of redundant information in ITC (*A*) and PFC (*B*). The magenta line indicates the readout performance when the top 64 neurons were used, and the green line indicates when the top 64 neurons were excluded and the remaining 192 neurons were used. As can be seen, the top 64 neurons achieve a performance level that is as good as using the whole population of 256 neurons. However, even when these neurons are excluded, readout is above chance, indicating that there is redundant information in these populations.

accuracies would only be achieved when using training and testing data from the same time period.

Figure 6, *A* and *B,* shows accuracy levels for decoding abstract category information when training a classifier with data from one time period (indicated by the *y* axis) and testing with data from a different time period (indicated on the *x* axis). As can be seen for both ITC and PFC, the highest decoding accuracies for each time bin occur along the diagonal of the figure, indicating that the best performance is achieved when training and testing is done using data from the same time bin relative to stimulus/trial onset. Additionally, for ITC, the decoding performance is also high when training using data from the sample period and testing using data from the decision period and vice versa, whereas for PFC, there seems to be little transfer between any different time periods. The pattern of transfer between the sample and the decision periods in ITC might indicate that there is indeed one pattern of activity in ITC that codes for the abstract category of the stimulus regardless of time; alternatively, this result might be due to visual information that is similar in the sample and decision stimuli, as the decision stimuli were created from random morphs between the prototype images. Figure 6, *C* and *D,* compares the decoding accuracies from training on three of these "fixed" time points (colored lines) to training and testing a classifier using data from the same time period (black lines) in a format that is similar to Figs. 2 and 3 (i.e., these are plots of 3 rows of Fig. 6, *A* and *B,* at time points during the sample, delay, and decision periods and compares them to the results in Fig. 3*A*). These plots again show that the highest decoding accuracy occurs when training and testing using data from the same time period, which implies that indeed the pattern of activity that codes for a particular piece of information changes with time.

Next we tested whether this changing pattern of activity was only due to neural adaptation in a fixed set of neurons or whether indeed different neurons were carrying the relevant information at different points in time. To address this question, we conducted analyses in which we eliminated the best 64 neurons (of 256 random neurons selected on each bootstrap trial) at one 150-ms time period (indicated on the *y* axis in Fig. 7) and training and test data were taken from a different 150-ms time period (indicated on the *x* axis). If the same small subset of neurons codes for abstract category information at all time periods, then eliminating these neurons from one time period should result in poor decoding accuracy at all time periods. Alternatively if different small subsets of neurons contain the abstract category information at different time periods, then there should only be a decrease in performance in the time period where the best neurons were removed. Results for both ITC and PFC show a clear pattern of lower decoding accuracies along the diagonal but largely unchanged decoding accuracies almost everywhere else, which indicates that different neurons contain the category information at different time points in a trial. Figure 7 also clearly shows that the neural code is changing faster than changes in the stimuli as illustrated by the fact that there is also a decrease only along the diagonal during the sample, delay, and decision periods even though the stimulus is not changing during these times. Additionally, Supplemental Fig. S7 shows that the neurons that code for identity information also change through the course of a trial, although the changes in code seem to be much less dramatic than is seen for the changes in code for abstract category information.

To further examine the duration of selectivity for individual neurons, we calculated an estimate of the mutual information (MI) between the category of the stimulus, and the average firing rate of neurons in 100-ms bins (see METHODS). Figure 8 shows the MI as a function of time for the four neurons that had highest MI at four different time bins. As can be seen for both PFC and ITC, individual neurons have short time windows of selectivity as expected from the results showing changing patterns of coding at the population level. It is also interesting to compare neurons 1 and 4 in Fig. 8*A*, where we can see two ITC neurons that are selective at slightly different times during the sample period even though the stimulus is constant during this time. This further supports the point that individual neuron's selectivity are occurring on a faster time scale than the changes in the stimuli.

DISCUSSION

We applied population decoding methods to neuronal spiking data recorded in PFC and ITC to gain more insight into *what* types of information are contained in these regions as well as *how* information is represented in these regions. By pooling information from hundreds of neurons, we were able to

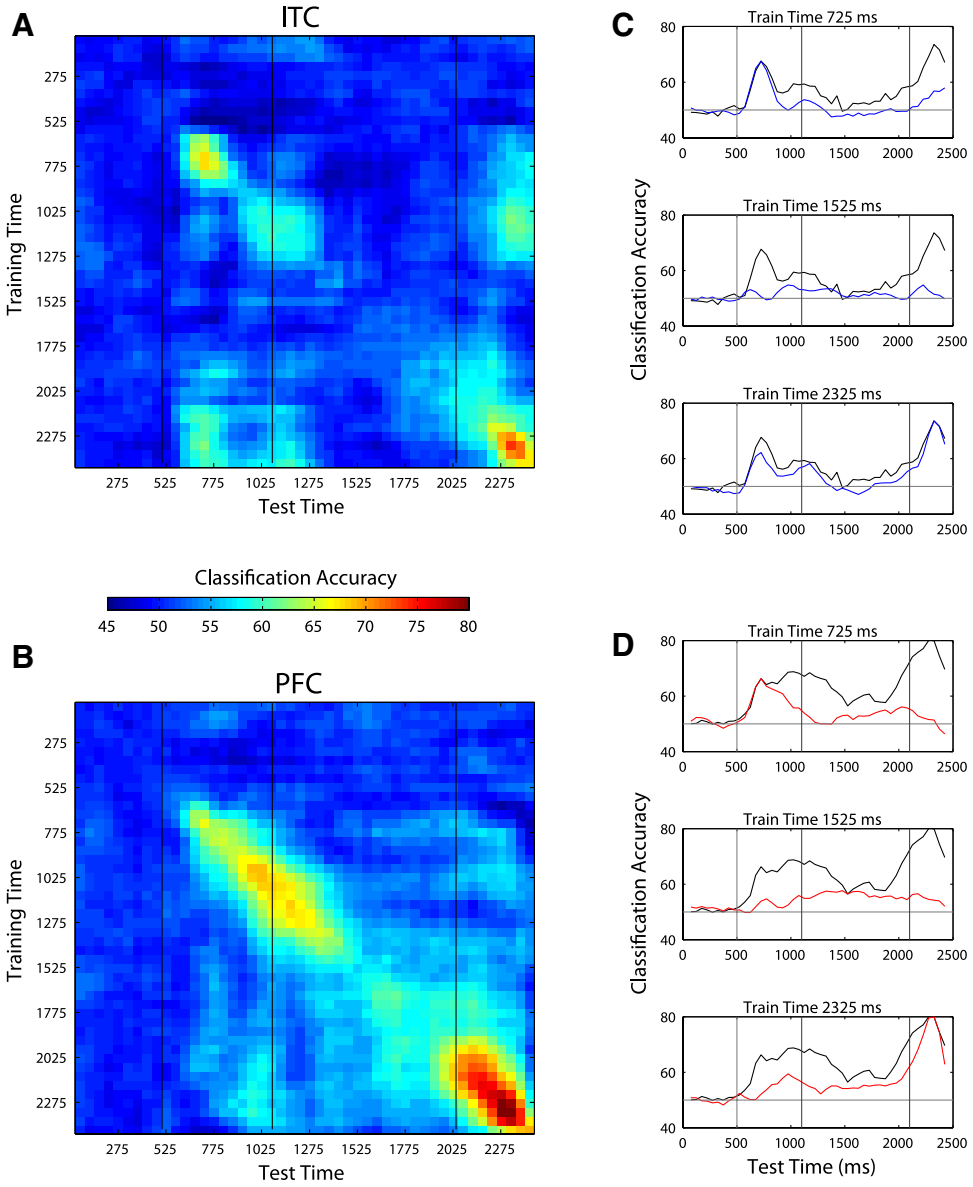

FIG. 6. Evaluating whether the same code is used at different times for abstract category information. *A*: in ITC there is some similarity in the neural code for abstract category information in the sample and the decision periods, as can be seen by the green patches near the *top right* and *bottom left* of the figure. Also there appears to be two different codes used during the sample period as can be seen by the two blob regions occurring 775–1,275 ms after the start of the trial. *B*: for PFC, the code for abstract category information seems to be constantly changing with time as indicated by the fact that the only high decoding accuracies are obtained along the diagonal of the plot. *C* and *D*: examples of decoding accuracies using 3 fixed training times from the sample, delay and decision periods (colored lines) compared decoding accuracies obtained when training and testing using the same time period (black line) for ITC (*C*) and PFC (*D*); (each of these plots corresponds to 1 row from the from *A* or *B* and the black line corresponds to the diagonal of this figure and is the same line as shown in Fig. 3*A*). These figures again illustrate that the highest performance is always obtained when training and testing is done using the same time bin relative to stimulus/trial onset, which suggests that the neural coding of abstract category information is time-locked to stimulus/trial onset.

observe the time course of the flow of information in these areas with a fine time scale. Results from basic decoding analyses (Fig. 2) showed that ITC contained more information related to the currently viewed stimulus than PFC, while PFC contained more task-relevant information than ITC, which is largely consistent with the results originally reported by Freedman et al. (2003). The finer temporal precision in our analyses also revealed an anticipatory response in both ITC and PFC, in which information about the category of the SAMPLE-STIMULUS reemerged just prior to the onset of the DECISION-STIMULUS, which seems similar to the increase in firing rate seen just prior to the onset of the decision period reported by Rainer et al. (Rainer and Miller 2002; Rainer et al. 1999) in macaque delayed match-to-sample experiments. We speculate that this anticipatory reemergence of category information might be involved in preparing the network for processing the imminent decision stimulus as soon as it is shown, which could account for the monkeys' fast reaction times.

The ability to train a pattern classifier on data of one type and test how well the classifier generalizes to data recoded under different conditions is very useful for obtaining more compelling answers to several questions. By training a classifier on data from a subset of images from one category and then testing on data recorded when a different disjoint subset of images was shown, we were able to get a better estimate of how much abstract category information is contained in both ITC and PFC (for more information about PFC's role in other categorization tasks, see Nieder et al. 2002; Shima et al. 2007). Results from our analysis of abstract category information revealed that there is initially as much abstract category information in ITC as PFC, which was not seen in the original analyses by Freedman et al. (2003) due to the long length of the time periods used in their analyses as well as potential biases introduced by only using "selective" neurons when creating category-selective indices (see INTRODUCTION).

The fact that there initially appears to be as much abstract category information in ITC as PFC (Fig. 3) raises several questions about ITC's role in categorization. One of the simplest explanations for the presence of abstract category information in ITC is that despite the morph paradigm used, the

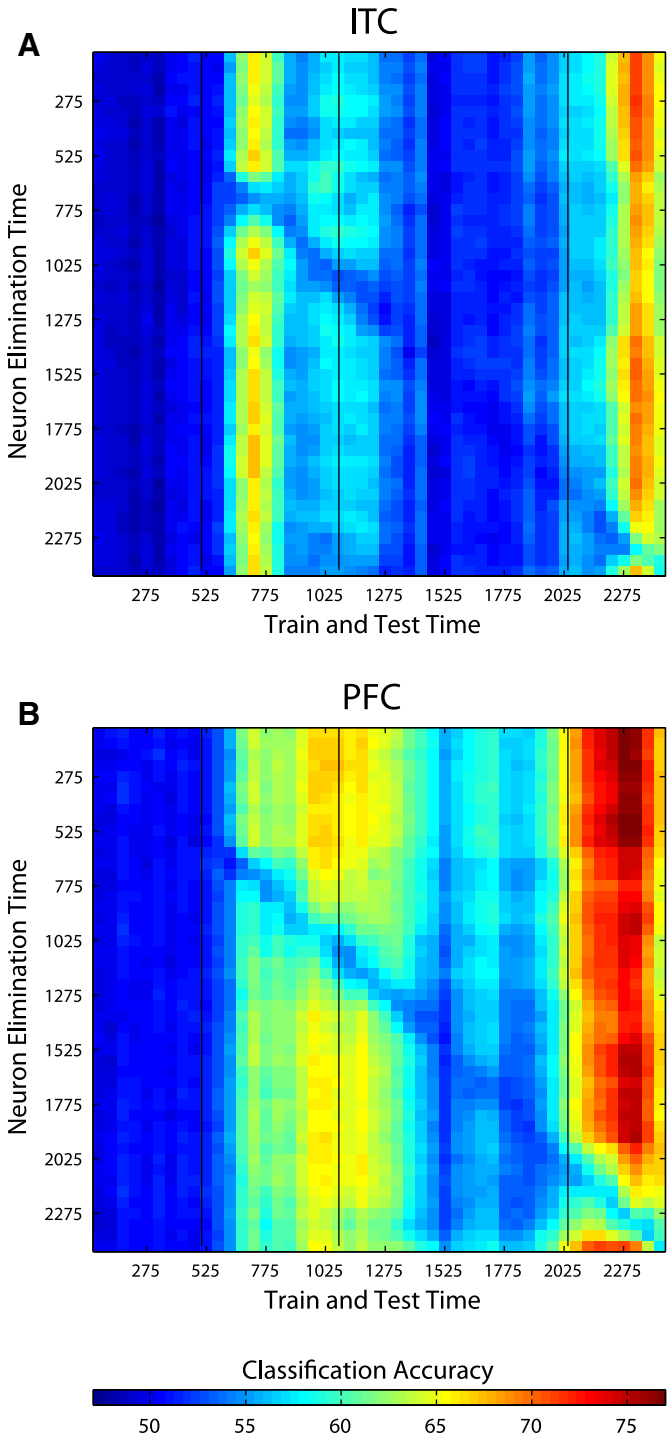

FIG. 7. Elimination of the best 64 neurons from the time period t_1 (specified on the y axis) and then training and testing with all the remaining 192 neurons at time period t_2 (as specified by the x axis) for ITC (*A*) and PFC (*B*). Eliminating the best neurons from the training set at one time period only has a large affect on decoding accuracy at that same time period and leaves other time period unaffected as can be seen by the fact that there is only lower performance long the diagonal of the figure. This indicates that the neurons in the population that carry the majority of the information change with time. Additionally, one can see a decrease only along the diagonal even during periods where the stimulus is constant (areas between the black vertical bars). This indicates that the neural code is changing at a faster rate than changes in the stimulus.

prototype images from the same category are more visually similar to each other than they are to the images from the other category (i.e., the 3 cat prototype images are more similar to each other than they are to the dog prototype images). If this was the case, then the classifier would be able to generalize across images from different prototypes from the same category based purely on visual information, which could explain the results (Sigala and Logothetis 2002). Analyses using a computational model of object recognition described in Serre et al. (2007) indeed suggest that prototype images are slightly more similar to each other than to prototypes from the opposite category. However, the level of similarity seems to be weaker than what is observed in the neural data. A direct test of whether visual image properties is giving rise to our findings could be done by running the same DMC experiment but using a different category boundary as was previously done for PFC (Freedman et al. 2001).

If indeed there is abstract category information in ITC that is not due to visual cues, this suggests that there is a "supervised" learning signal in ITC that is causing neurons in ITC to respond similarly to stimuli from the same category. One possible source of this supervised learning signal is that during the course of the sample presentation, PFC extracts category information from the signals arising in ITC and feeds this category information back to ITC (Tomita et al. 1999). However, with the resolution of our analyses, we could not detect any clear latency differences between the category information arising in PFC and ITC (see Supplemental Fig. S8). Given that there could be a single synapse between neurons in these two brain areas, the latency differences could be too small to detect (Ungerleider et al. 1989). Alternatively, ITC could have acquired abstract category information during the course of the monkey being trained in the task. In this scenario, which is similar to the model proposed by Risenhuber and Poggio (2000), the activity of "lower level" neurons that are selective to individual visual features present in particular stimuli are pooled together by "higher level" neurons through a supervised learning signal enabling these higher level neurons to respond similarly to all members of a given category irrespective of the visual similarity of individual members of the category. It should be noted that more recent models (e.g., Serre et al. 2007) propose a supervised learning signal is only present in PFC, while the presence of abstract category information in ITC suggests this supervised learning signal might be organizing the response properties of neurons earlier in the visual hierarchy (Mogami and Tanaka 2006); however these models could be easily modified to incorporate a supervised learning signal in stages before PFC. Because these monkeys have had an extensive amount of experience with these stimuli, it is also possible that a consolidation process has occurred when the monkey learned the task. For category grouping behavior that occurs on shorter time scales, it is possible that category signals would only be found in PFC.

By analyzing data over long time intervals, most physiological studies assume tacitly or explicitly that the neural code remains relatively static as long as the stimulus remains unchanged. We examined how stationary the neural code is by training the classifier using data from one time period and then testing with data from a different time period (Fig. 6). These analyses suggest that the pattern of activity coding for a particular stimulus or behaviorally relevant variable changes with time. Such results are

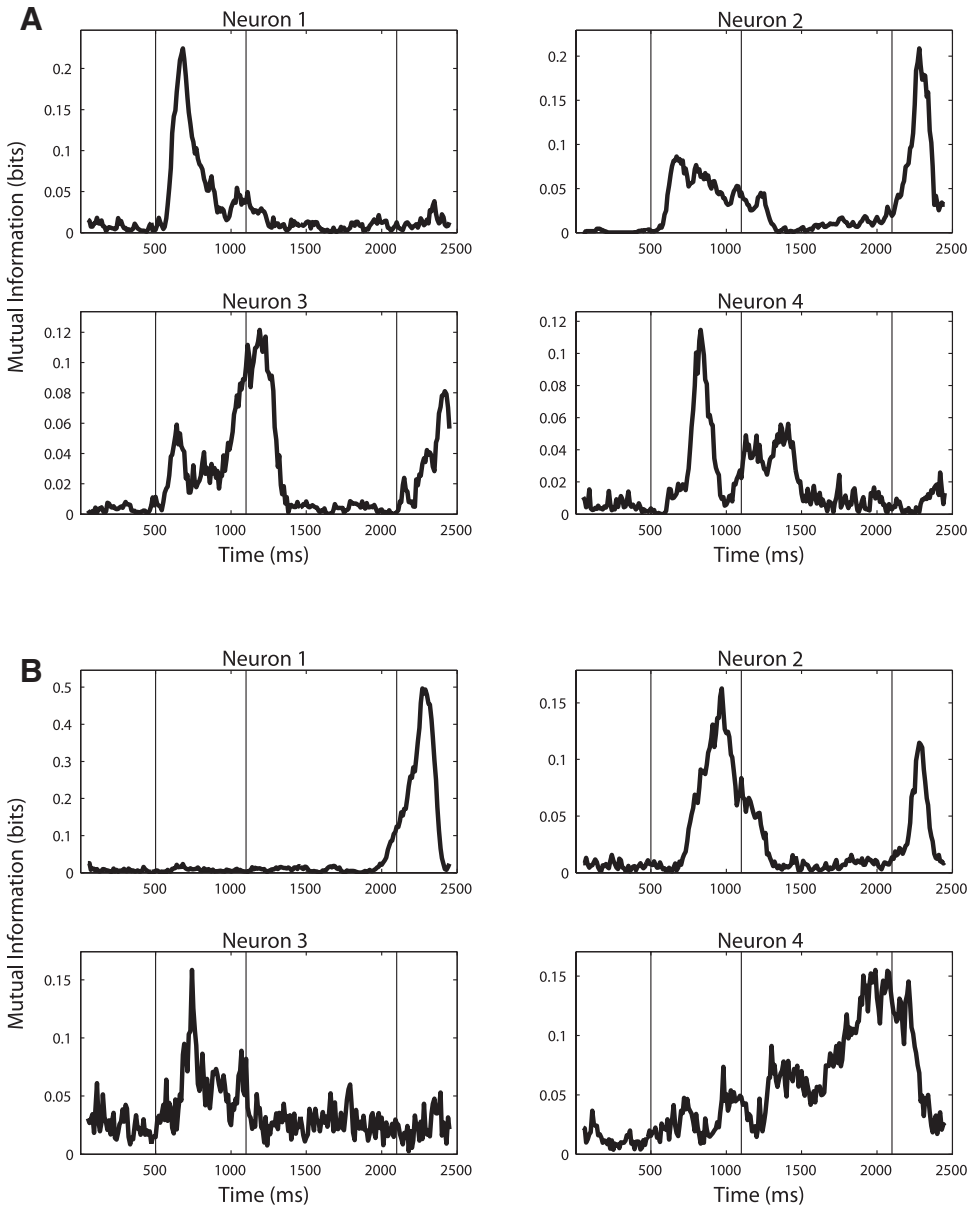

FIG. 8. Illustration showing that many individual neurons have short periods of selectivity for ITC (*A*) and PFC (*B*). The figure plots the 4 neurons for ITC and PFC that had the highest the mutual information between the category of the SAMPLE-STIMULUS and neuron's firing rate (firing rates where calculated using 100-ms bin periods sampled every 10 ms). As can be seen, most neurons show high mutual information (MI) values for only short time periods, which is what is expected for a population code that changes with time. It is also interesting to compare *neurons 1* and *4* in ITC (*A*) because it shows that individual neurons have different peak selectivity times even when the stimulus being shown is constant. Thus the changing of the neural code is not just due to changes in the stimulus.

consistent with the findings of Gochin et al. (1994), in whose study a paired-associate task was used to show that the pattern of activity in macaque ITC that is indicative of a particular stimulus during a sample period is different from the pattern of activity that is indicative of the same stimulus during a second stimulus presentation period. Also Nikolic et al. (2007) reported dynamic changes in the weights of separating hyperplanes for discriminating between visual letters using data from macaque V1. These observations suggest that the coding of particular variables through changing patterns of activity might be a general property of neural coding throughout the visual system. However, because adaptation or other nonlinear scaling of firing rates could potentially explain these results as an artifact of the decoding procedure in these studies, we further tested how stationary the neural code is by eliminating the best neurons from one time period and testing the classifier on data from another time period (Fig. 7). Results from this analysis show that there is only a temporally localized drop in classification accuracy, which indicates that different neurons carry information about the same variable at different time periods. Additionally, analyses of mutual information showed that most individual neurons are only selective for short time windows. These observations are consistent with the findings of Zaksis et al. (Zaksas and Pasternak 2006), who used an ROC analysis to show that many neurons in PFC and MT only have short time periods of selectivity. Baeg et al. (2003) also showed that past and future actions of rats can be decoded based on PFC activity during a delay period even when neurons with sustained activity are excluded from the analysis; this again agrees with our observations showing that the pattern of neural activity that codes information changes with time. While previous studies have concluded that neurons with short periods of selectivity play an important role in memory of stimuli, we also speculate that these dynamic patterns of activity might be important for the coding of a sequence of images so that the processing of new stimuli do not interfere with those just previously seen and could underlie the ability of primates to keep track of the relative timing of events.

An ongoing debate concerning the neural code is whether information is transmitted using a "rate code" in which all information is carried in the mean firing rate of a neuron within a particular time window, or whether a temporal code is used in which information is carried in by the precise timing of individual spikes (deCharms and Zador 2000). While the results in this paper cannot conclusively answer which coding scheme is correct, they do give some insight into this debate. First, because we decode mean firing rates over 150-ms bins (and shorter time bins tended to achieve lower decoding accuracies), our findings suggest that a large amount of information is still present even when the precise time of each spike is ignored (also see Hung et al. 2005). While it is possible that superior decoding performance could be achieved by using an algorithm that took exact spike times into account, considering the high performance level at certain time periods in the experiment (e.g., decoding of match versus nonmatch trial information is over 90% in PFC during the decision period, which is comparable to the 90% correct animals' performance), often there is not much more information left to extract. Second, because our results show that the pattern of neural activity that is predictive of a particular variable changes with time and that this change occurs on a faster time scale then changes in the stimulus, these findings argue against a strict rate based coding scheme in which all information about a stimulus is coded by the firing rate alone. Thus our findings suggest that neurons in ITC and PFC maintain information in their mean firing rates over time windows on the order of a few hundred milliseconds and that these periods of selectivity are time-locked to particular task events (with different neurons having different time lags), giving rise to a dynamic coding of information at the population level.

Applying feature selection methods prior to using pattern classifiers allowed us to characterize the compactness and redundancy of *information* in ITC and PFC. Results from these analyses revealed that at any one point in time, all the abstract category information available is contained in a small subset of neurons. However there still is a substantial amount of redundant information between this small highly informative subset of neurons and the rest of the more weakly selective neurons in the rest of the population. While other studies have examined sparse *spiking activity* in several different neural systems (Hahnloser et al. 2002; Perez-Orive et al. 2002; Quiroga et al. 2005; Rolls and Tovee 1995), and theoretical models have been proposed that analyze the implication of this sparse activity (Olshausen and Field 1997), our notion of compactness of *information* differs from these measures because we are not focused on whether neurons are firing, but rather we are focused on the information content that is carried by this spiking activity. It should also be noted that our notion of compactness of information differs the notion compactness described by Field (1994), because Field's notion of compactness implies that *all* neurons are involved in the coding for a stimulus, while our results suggest that only a small subset of a larger population of neurons contain the relevant information and that this subset of neurons changes in time (thus our notion of compactness could be equally well characterized as *sparseness of information,* however given the strong association in the literature between the term sparseness and firing rate, we found using this terminology to be confusing). Thus our measure adds a new and potentially useful statistic for understanding how information is coded in a given cortical region.

The neuronal responses studied here were not recorded simultaneously, and the creation of pseudo-populations can alter estimates of the *absolute* amount of information that a population contains because of noise correlations (Averbeck and Lee 2006; Averbeck et al. 2006). However, we were interested in *relative* information comparisons between different time periods or between different brain regions, so our conclusions would not be substantially altered by having data from simultaneous recordings. Furthermore, empirical evidence suggests that decoding using pseudo-populations returns roughly the same results as when using simultaneously recorded neurons (Aggelopoulos et al. 2005; Anderson et al. 2007; Baeg et al. 2003; Gochin et al. 1994; Nikolic et al. 2007; Panzeri et al. 2003). Our estimates of the absolute amount of information in the population could also be affected by the amount of data we have, the quality of the learning algorithms (however, see Supplemental Fig. S2, which suggests this is not an issue), and the features used for decoding. However, because in principle these issues affect all time points and brain areas equally, relative comparisons should be largely unaffected by them.

The ability to decode information from a population of neurons does not necessarily mean that a given brain region is using this information or that downstream neurons actually decode the information in the same way that our classifiers do. Our results using analyses in which the classifier is trained with one type of stimuli and must generalize to a different but related type of stimuli, supports the notion that the animal is using this information because such generalization implies a representation that is distinct from properties that are directly correlated with the stimuli, and having such an abstract representation coincidentally would be highly unlikely. For this reason, most of the analyses in this paper have focused on abstract category information (Figs. 3–7) because this information meets our criteria of being abstracted from the exact stimuli that are shown and hence is most likely utilized by the animal.

Using population decoding to interpret neural data is important because it examines data in a way that is more consistent with the notion that information *is actually contained* in patterns of activity across many neurons. By computing statistics on random samples of neurons, most analyses of individual neurons implicitly assume that each neuron is independent of all others and that neural populations are largely homogenous. However, such implicit assumptions are contrary to the prevailing belief that brain regions contain circuits of heterogeneous cells that have different functions and is inconsistent with empirical evidence (compact coding of information and activity) seen in this and other studies. The methods discussed in this paper can help align a distributed coding theoretical framework with analysis of actual empirical data, which should give deeper insights into the ultimate goal of understanding the algorithms and computations used by the brain that enable complex animals, such as humans and other primates, to make sense of our surroundings and to plan and execute successful goal-directed behaviors.

ACKNOWLEDGMENTS

We thank B. Jarosiewicz and M. Riesenhuber for helpful comments on the manuscript. This report describes research done at the Center for Biological and Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain and Cognitive Sciences, and which is affiliated with the Computer Sciences and Artificial Intelligence Laboratory (CSAIL).

GRANTS

This research was sponsored by grants from: National Science Foundation, National Institute of Mental Health, and Darpa. Additional support was provided by: Children's Ophthalmology Foundation, Epilepsy Foundation, National Defense Science and Engineering Graduate Research Fellowship program, Honda Research Institute USA, NEC, Sony, and the Eugene McDermott Foundation. Additional supplementary material can be found at http://cbcl.mit.edu/people/emeyers/jneurophys2008/.

REFERENCES

Abeles M. *Corticonics: Neural Circuits of the Cerebral Cortex.* Cambridge, MA: Cambridge Univ. Press, 1991.

Aggelopoulos NC, Franco L, Rolls ET. Object perception in natural scenes: encoding by inferior temporal cortex simultaneously recorded neurons. *J Neurophysiol* 93: 1342–1357, 2005.

Anderson B, Sanderson MI, Sheinberg DL. Joint decoding of visual stimuli by IT neurons' spike counts is not improved by simultaneous recording. *Exp Brain Res* 176: 1–11, 2007.

Averbeck BB, Latham PE, Pouget A. Neural correlations, population coding and computation. *Nat Rev Neurosci* 7: 358–366, 2006.

Averbeck BB, Lee D. Effects of noise correlations on information encoding and decoding. *J Neurophysiol* 95: 3633–3644, 2006.

Baeg EH, Kim YB, Huh K, Mook-Jung I, Kim HT, Jung MW. Dynamics of population code for working memory in the prefrontal cortex. *Neuron* 40: 177–188, 2003.

Dayan P, Abbott LF. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* Cambridge, MA: MIT Press, 2001.

deCharms RC, Zador A. Neural representation and the cortical code. *Annu Rev Neurosci* 23: 613–647, 2000.

Duda RO, Hart PE, Stork DG. *Pattern Classification.* New York: Wiley, 2001.

Field DJ. What is the goal of sensory coding. *Neural Comput* 6: 559–601, 1994.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291: 312–316, 2001.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK. Visual categorization and the primate prefrontal cortex: neurophysiology and behavior. *J Neurophysiol* 88: 929–941, 2002.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J Neurosci* 23: 5235–5246, 2003.

Gochin PM, Colombo M, Dorfman GA, Gerstein GL, Gross CG. Neural ensemble coding in inferior temporal cortex. *J Neurophysiol* 71: 2325–2337, 1994.

Hahnloser RHR, Kozhevnikov AA, Fee MS. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* 419: 65–70, 2002.

Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* 79: 2554–2558, 1982.

Hung CP, Kreiman G, Poggio T, DiCarlo JJ. Fast readout of object identity from macaque inferior temporal cortex. *Science* 310: 863–866, 2005.

Laurent G. Olfactory network dynamics and the coding of multidimensional signals. *Nat Rev Neurosci* 3: 884–895, 2002.

McIlwain JT. Population coding: a historical sketch. In: *Advances in Neural Population Coding*, edited by Nicolelis MAL. Amsterdam: Elsevier, 2001, p. 3–7.

Miller EK, Cohen JD. An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24: 167–202, 2001.

Mogami T, Tanaka K. Reward association affects neuronal responses to visual stimuli in macaque TE and perirhinal cortices. *J Neurosci* 26: 6761–6770, 2006.

Nieder A, Freedman DJ, Miller EK. Representation of the quantity of visual items in the primate prefrontal cortex. *Science* 297: 1708–1711, 2002.

Nikolic D, Haeusler S, Singer W, Maass W. Temporal dynamics of information content carried by neurons in the primary visual cortex. In: *Advances in Neural Information Processing Systems*, edited by Scholkopf B, Platt J, Hoffman T. Cambridge, MA: MIT Press, 2007, p. 1041–1048.

Olshausen BA, Field DJ. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res* 37: 3311–3325, 1997.

Paninski L. Estimation of entropy and mutual information. *Neural Comput* 15: 1191–1253, 2003.

Panzeri S, Pola G, Petersen RS. Coding of sensory signals by neuronal populations: the role of correlated activity. *Neuroscientist* 9: 175–180, 2003.

Perez-Orive J, Mazor O, Turner GC, Cassenaer S, Wilson RI, Laurent G. Oscillations and sparsening of odor representations in the mushroom body. *Science* 297: 359–365, 2002.

Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I. Invariant visual representation by single neurons in the human brain. *Nature* 435: 1102–1107, 2005.

Quiroga RQ, Snyder LH, Batista AP, Cui H, Andersen RA. Movement intention is better predicted than attention in the posterior parietal cortex. *J Neurosci* 26: 3615–3620, 2006.

Rainer G, Miller EK. Timecourse of object-related neural activity in the primate prefrontal cortex during a short-term memory task. *Eur J Neurosci* 15: 1244–1254, 2002.

Rainer G, Rao SC, Miller EK. Prospective coding for objects in primate prefrontal cortex. *J Neurosci* 19: 5493–5505, 1999.

Riesenhuber M, Poggio T. Models of object recognition. *Nat Neurosci* 3 (*Suppl*):1199–1204, 2000.

Rolls ET, Tovee MJ. Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J Neurophysiol* 73: 713–726, 1995.

Rumelhart DE, McClelland JL, University of California San Diego. PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Cambridge, MA: MIT Press, 1986.

Samengo I. Information loss in an optimal maximum likelihood decoding. *Neural Comput* 14: 771–779, 2002.

Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, Poggio T. *A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex.* CBCL Paper 259/AI Memo 2005-036, Massachusetts Institute of Technology, Cambridge, MA, 2005.

Seung HS, Sompolinsky H. Simple models for reading neuronal population codes. *Proc Natl Acad Sci USA* 90: 10749–10753, 1993.

Shima K, Isoda M, Mushiake H, Tanji J. Categorization of behavioral sequences in the prefrontal cortex. *Nature* 445: 315–318, 2007.

Shlens J, Kennel MB, Abarbanel HDI, Chichilnisky EJ. Estimating information rates with confidence intervals in neural spike trains. *Neural Comput* 19: 1683–1719, 2007.

Sigala N, Logothetis NK. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* 415: 318–320, 2002.

Stanley GB, Li FF, Dan Y. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *J Neurosci* 19: 8036–8042, 1999.

Tanaka JW. Object categorization, expertise and neural plasticity. In: *The New Cognitive Neurosciences*, edited by Gazzaniga M. Cambridge, MA: MIT Press, 2004, p. 876–888.

Tomita H, Ohbayashi M, Nakahara K, Hasegawa I, Miyashita Y. Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature* 401: 699–703, 1999.

Ungerleider LG, Gaffan D, Pelak VS. Projections from inferior temporal cortex to prefrontal cortex via the uncinate fascicle in rhesus monkeys. *Exp Brain Res* 76: 473–484, 1989.

Zaksas D, Pasternak T. Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *J Neurosci* 26: 11726–11742, 2006.

Zemel RS, Dayan P, Pouget A. Probabilistic interpretation of population codes. *Neural Comput* 10: 403–430, 1998.

Object decoding with attention in inferior temporal cortex

Ying Zhang[a,1], Ethan M. Meyers[a,1,2], Narcisse P. Bichot[a], Thomas Serre[a,b], Tomaso A. Poggio[a], and Robert Desimone[a]

[a]Department of Brain and Cognitive Sciences, McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI 02912

Edited by Charles G. Gross, Princeton University, Princeton, NJ, and approved April 11, 2011 (received for review January 20, 2011)

Recognizing objects in cluttered scenes requires attentional mechanisms to filter out distracting information. Previous studies have found several physiological correlates of attention in visual cortex, including larger responses for attended objects. However, it has been unclear whether these attention-related changes have a large impact on information about objects at the neural population level. To address this question, we trained monkeys to covertly deploy their visual attention from a central fixation point to one of three objects displayed in the periphery, and we decoded information about the identity and position of the objects from populations of ∼200 neurons from the inferior temporal cortex using a pattern classifier. The results show that before attention was deployed, information about the identity and position of each object was greatly reduced relative to when these objects were shown in isolation. However, when a monkey attended to an object, the pattern of neural activity, represented as a vector with dimensionality equal to the size of the neural population, was restored toward the vector representing the isolated object. Despite this nearly exclusive representation of the attended object, an increase in the salience of nonattended objects caused "bottom-up" mechanisms to override these "top-down" attentional enhancements. The method described here can be used to assess which attention-related physiological changes are directly related to object recognition, and should be helpful in assessing the role of additional physiological changes in the future.

macaque | vision | readout | population coding | neural coding

Previous work examining how attention influences the ventral visual pathway has shown that attending to a stimulus in the receptive field (RF) of a neuron is correlated with increases in firing rates or effective contrast, increases in gamma synchronization, and decreases in the Fano factor and noise correlation, compared with when attention is directed outside the RF (1–8). However, because these effects are often relatively modest, it has been unclear whether these effects would have a large impact on information contained at the population level when any arbitrary stimulus needs to be represented. Indeed, recent work has suggested that high-level brain areas can represent multiple objects with the same accuracy as single objects even when attention is not directed to a specific object (9), which raises questions about the importance of the attention-related effects that have been reported in previous studies.

Another feature of the previous neurophysiology work on attention has been that it has primarily focused on the neural mechanisms that underlie attention (i.e., what neural circuits/processing underlie the changes seen with attention). This approach, which is related to David Marr's implementational level of analysis (10), has been fruitful, as evidenced by the fact that several mechanistic models have been created that can account for a variety of firing-rate changes seen in a number of studies (11–20). Less work, however, has focused on Marr's "algorithmic/representational level," which in this context would address how particular physiological changes enable improvements in neural representations that are useful in solving specific "computational-level" tasks (such as recognizing objects).

To assess the significance of particular physiological changes associated with changes in attentional state, and to gain a deeper algorithmic/representational-level understanding of how attention impacts visual object recognition, we use a neural population decoding approach (21, 22) to analyze electrophysiological data. Our approach is based on the hypothesis that visual objects are represented by patterns of activity across populations of neurons. Thus, we assess how these representations change when the visual objects are displayed in clutter and when spatial attention is deployed. Our results show that even limited clutter decreases information about particular objects in inferior temporal cortex (IT), and that attention-related firing-rate changes significantly increase the amount of information about behaviorally relevant objects in IT. Additionally, by focusing on how information is represented by populations of neurons, we find that "competitive" effects that occur when two stimuli are presented within a neuron's RF, and global "gain-like" effects that occur when a single stimulus is presented within a neuron's RF, can both be viewed as restoring patterns of neural activity for object identity and position information, respectively. Future work using this approach should help assess whether other physiological changes apart from firing-rate changes have an important impact on information content of IT, and should further help illuminate the computations that underlie object recognition.

Results

We recorded the responses of IT neurons to either one or three extrafoveal stimuli in the contralateral visual field while monkeys fixated a spot at the center of a display (Fig. 1*A* and Fig. S1). The three stimuli were positioned so that each was likely to be contained within a different RF of cells in V4 and lower-order areas but within the same large RFs of IT cells. When one stimulus appeared in isolation, it was always the task-relevant target, but when three stimuli appeared, one was the target while the other two stimuli were distractors on a given trial. Approximately 525 ms after the stimuli onset, a directional cue (line segment) appeared that "pointed" to the target stimulus to attend. The monkey was rewarded for making a saccade to the target stimulus when it changed slightly in color, which occurred randomly from 518 to 1,260 ms after cue onset. On half of the trials, one of the distractor stimuli changed color before the target change (foils), but the monkey was required to withhold a saccade to it. Of trials that the monkeys fixated until the time of cue onset, correct saccades to the target color change occurred on ∼72% of trials, and incorrect saccades to

Author contributions: Y.Z., E.M.M., N.P.B., T.S., T.A.P., and R.D. designed research; Y.Z. and N.P.B. performed research; E.M.M. contributed new reagents/analytic tools; Y.Z. and E.M.M. analyzed data; and E.M.M., T.A.P., and R.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]Y.Z. and E.M.M. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: emeyers@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1100999108/-/DCSupplemental.

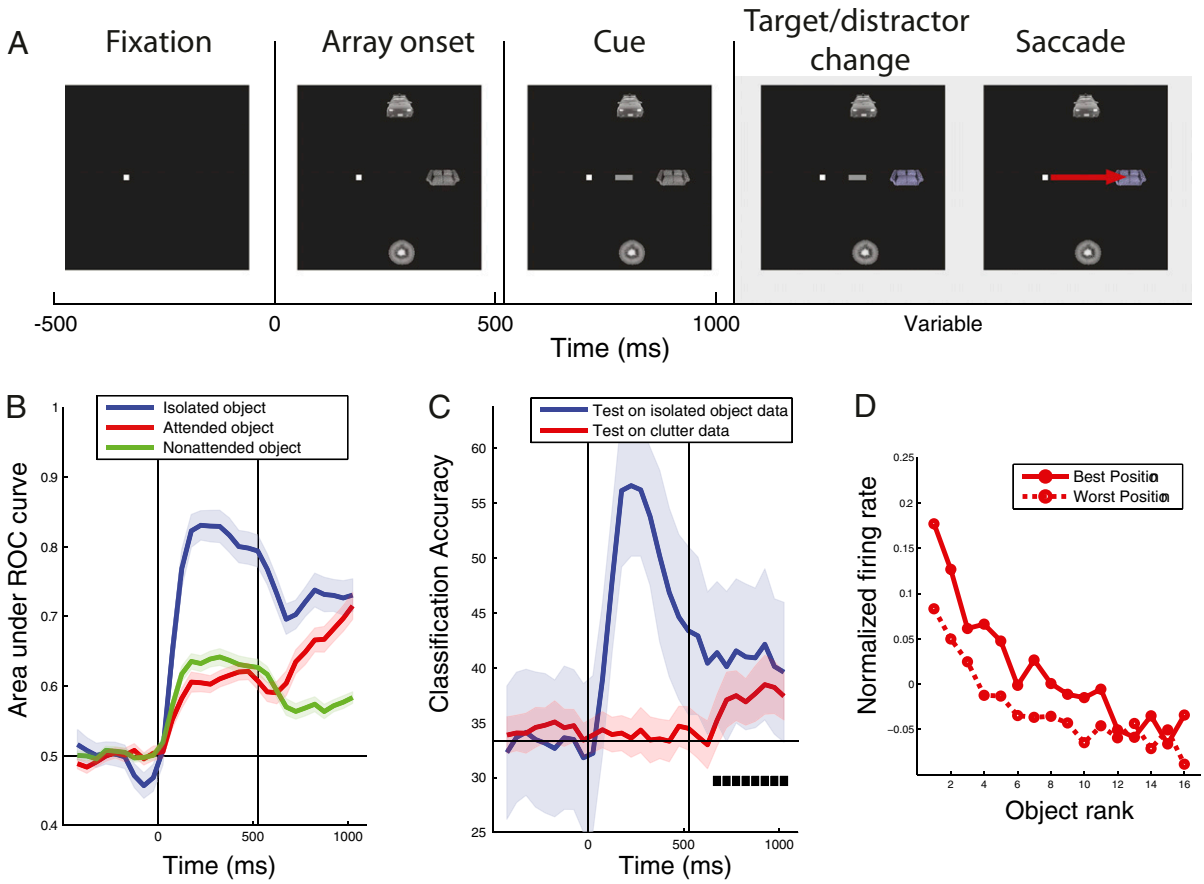

Fig. 1. Effects of attention on decoding accuracy. (*A*) Timeline for three-object trials. Single-object trials had the same timeline, except only one object was displayed. It should be noted that the attentional cue was shown for both isolated- and three-object trials, and once the cue was displayed it remained on the screen for the remainder of the trial (which could lead to potential visual–visual interactions). (*B*) Decoding accuracies for which object was shown on isolated-object trials (blue traces), and the attended object (red trace) and nonattended objects (green trace) in the three-object displays. Vertical lines indicate the times of stimulus onset, and cue onset. Colored shaded regions indicate ±1 SE of the decoding results (*Methods*). (*C*) Decoding accuracies for the position of the isolated stimulus (blue trace) and the attended stimulus (red trace). Black square boxes indicate times when the decoding accuracy for the position of the attended object was above what would be expected by chance (chance performance is 33%). (*D*) Z-score–normalized population firing rates to cluttered-display images ranked based on their isolated-object preferences. The data from isolated-object trials were first used to calculate each neuron's best and worst position and the ranking of its best to worst stimuli. The firing rates to these stimuli on cluttered trials were then calculated and averaged over all neurons, and are plotted separately for attention to the best versus worst position. Attending to the neuron's preferred position led to a relatively constant offset in the neuron's object tuning profile.

a distractor color change were made on only ~1% of trials. On the other 27% error trials, 36% of those were due to early saccades to the target location before the color change, and the remaining errors were simply random breaks in fixation.

To understand how information about objects is represented by populations of IT neurons, we applied population decoding methods (21, 22) to the firing rates of pseudopopulations of 187 neurons from two monkeys on a first stimulus set (similar results were obtained from each monkey, so the data were combined; Fig. S2) and on a second stimulus set shown to monkey 2 (Fig. S3). (By "pseudopopulation" response, we mean the response of a population of neurons that were recorded under the same stimulus conditions but the recordings were made in separate sessions, i.e., the neurons were not recorded simultaneously but treated as though they had been.) We trained a pattern classifier on data from isolated-object trials and then made predictions about which objects were shown on either different isolated-object trials or on trials in which three objects had been shown (*Methods*). Fig. 1*B* shows that information about the identity of isolated objects (blue trace) rose rapidly after stimulus onset, reaching a peak value for the area under the receiver operating characteristic (AUROC) curve of 0.83 ± 0.022 at 225 ms after stimulus onset, whereas information about the objects in the multiple-object displays also rose after the onset of the stimuli (red and green traces) but only reached a peak value of 0.62 ± 0.014 before the onset of the attentional cue. An AUROC of 0.5 represents chance performance. Thus, 75 ± 75 ms after the onset of the stimulus array, the amount of information about the objects in the three-object displays was greatly reduced compared with when these objects were shown in isolation ($P < 0.01$, permutation test; see *SI Text* for more details), showing that clutter has a significant impact on the amount of information about specific objects in IT (also see Fig. S2).

Approximately 150 ± 75 ms after the attentional cue was displayed, information about the attended object (red trace) rose significantly above the amount of information seen in the non-attended object ($P < 0.01$, permutation test). By 400 ms after cue onset, information about the attended object had reached an AUROC value of 0.64 ± 0.017, which was similar to the value of 0.68 ± 0.024 for decoding isolated-object trials during the same trial period. At the same time, information about the non-attended stimuli (green trace) decreased to a value of 0.56 ± 0.010. Thus, location-directed attention can have a significant impact on the amount of information about specific objects in IT.

NEUROSCIENCE

These attention-related changes can also be observed in the firing rate of the population of neurons to preferred and nonpreferred stimuli (Fig. S4).

In addition to identity information, position information was also enhanced (Fig. 1*C*). When this position enhancement was examined using more conventional analyses that create tuning curves for each neuron (by ordering each neuron's responses from the best to the worst stimulus) and then plotting the population average tuning curves separately for the best versus worst location (Fig. 1*D*), this position enhancement with attention appeared as an upward shift in the population tuning curves, as has previously been reported (1). However, this upward shift is a consequence of aligning all neurons' responses to their preferred and nonpreferred locations, rather than a result of all neurons increasing their firing rates with attention. When attention is allocated to a particular location, increases and decreases in activity occur in different neurons (depending on a neuron's RF structure), which creates a distributed pattern of activity that contains information about the location of where the monkey is attending.

In the cluttered decoding results described above, we trained the classifier with data from isolated-object trials and then tested the classifier with data from three-object trials. This allowed us to test whether one of the effects of attention was to restore the pattern of neural activity to a state that was similar to when an object was shown in isolation. However, it is possible that attention could have additional effects on neural representations that modify the representation of each object to make them more distinct from one another (and thus increase the amount of information about the objects), but in a way that is not related to the neural representations that are present when the objects are shown in isolation. To test this possibility, we compared the decoding accuracies when training the classifier on cluttered-display data (Fig. 2, dashed traces) to the decoding accuracy when training with isolated-object data (Fig. 2, solid traces). If attention added additional information, then there should be higher accuracy when training with cluttered data from the cue period (Fig. 2 *Right*) than when training with isolated-object data. The results show that training on isolated-object trial data was better than training on cluttered trial data during the array period before the attentional cue (Fig. 2 *Left*). This result is consistent with the idea that clutter decreases information about specific objects. Most importantly, training the classifier with cluttered data in the array period (right subplots) did not lead to better performance than training the classifier with isolated-object data (in particular, the dashed line is not higher than the solid line in the right plot of Fig. 2*B* at any point in time). Thus, it appears that attention's main effect was to restore the pop-

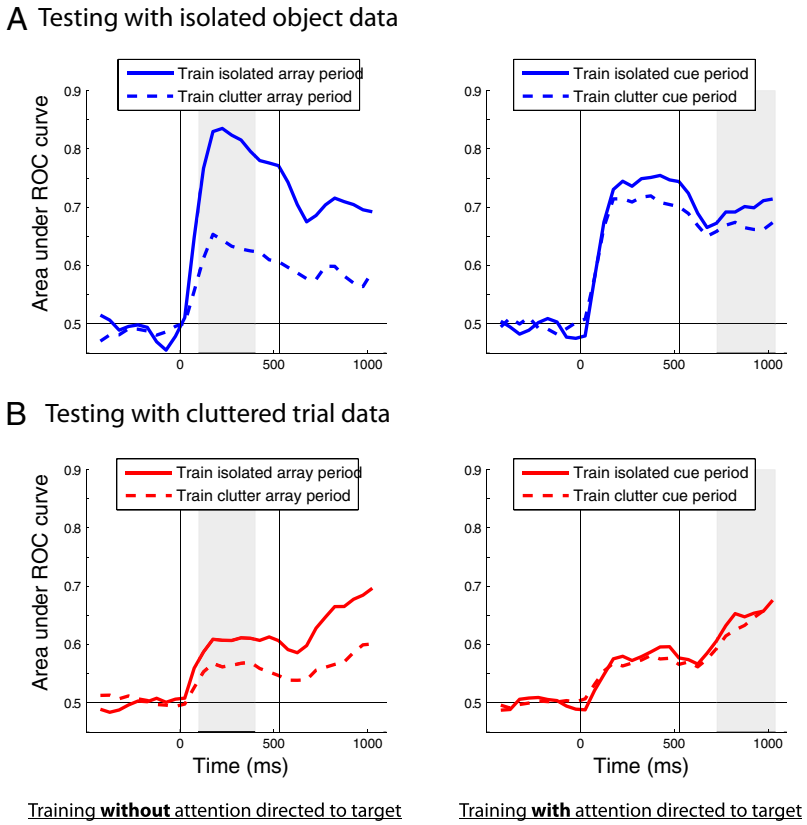

Fig. 2. Attention restores neural activity to a state that is similar to when the attended object is presented alone. Results are based on training a classifier with either 11 examples of isolated objects (solid lines) or with 11 examples of the same (attended) objects in cluttered displays (dashed line). The classifier was trained with either 300 ms of data from the array period (*Left*, gray shaded region) before the attentional cue or 310 ms of data from the period following the onset of the attentional cue (*Right*, gray shaded region). The classifier's performance was tested with either isolated-object trials (*A*) or with cluttered array data (*B*), using firing rates from 150 ms of data sampled in 50-ms sliding intervals. The results show that when training with data from the array period, better performance is achieved when training with isolated-object data compared with clutter data (left two plots), confirming that clutter reduces information about particular objects. Approximately equal levels of performance were obtained when training with either isolated or cluttered data from the period following the attentional cue (right two plots), indicating that attention caused the neural representation to enter a state that was similar to when objects were presented in isolation. It should be noted that for the isolated-object trials there is slightly more information about the stimuli when they are first shown, and hence the decoding accuracies are higher when training with data from the array period compared with training with data from the cue period (i.e., the solid blue trace in *A Left* is slightly higher than the solid blue trace in *A Right*).

ulation of neural activity to a state that was similar to that when the attended object was shown in isolation.

The above results show that top-down attention had a large impact on the object information represented in IT. We then asked how resistant these object representations would be to salient changes in the distractor. To test this, we aligned the data to the time when a distractor underwent a color change, and we decoded the identity of both the target and the distractor stimuli. The results, plotted in Fig. 3, show that before the distractor change there was a large improvement in decoding with attention (red trace) as seen before. However, when the distractor changed color, the dominant representation in IT switched transiently to the distractor object (light green trace), before returning to the attended-object representation (red trace). Thus, bottom-up, or stimulus, changes in the saliency of the distractor objects overrode the top-down attention-induced enhancements of particular objects. An examination of behavioral data (Fig. S5) revealed that reaction times were longer when the target changed soon after the distractor change, suggesting that the monkeys transiently switched their attention to the salient distractor, which impaired their ability to detect a target color change.

Discussion

Previous work has shown that several different physiological changes are correlated with changes in attentional state. It has been unclear, however, which of these physiological changes are important for object recognition. In this work, we show that visual clutter does indeed reduce the amount of information about specific objects in IT, and that attention-related *firing-rate changes* do indeed have a large impact on the amount of information present about particular objects. By applying these same methods to simultaneously recorded neural activity in future studies, it should be possible to assess whether changes in noise correlations or synchrony also have an impact on information at the population level.

Across the studies that have examined attention-related firing-rate changes in neurons in the ventral visual pathway (and in area V4 in particular), two seemingly distinct effects have been widely reported. The first effect occurs when a preferred stimulus and a nonpreferred stimulus are presented simultaneously in a neuron's RF and the monkey must pay attention to either the preferred or the nonpreferred stimulus depending on a cue that varies from trial to trial. (By "preferred stimulus," we mean a stimulus that elicits a high firing rate from the neuron when it is presented in isolation, and by "nonpreferred stimulus," we mean a stimulus that elicits a low firing rate when it is presented in isolation.) Results from these studies show that firing rates increase when the monkey attends to the preferred stimulus and that firing rates decrease when the monkey attends to a nonpreferred stimulus. The second attention-related firing-rate change occurs when a single stimulus is shown in a neuron's RF, and orientation tuning curves for this single stimulus are mapped out when the monkey attends either inside the neuron's RF or outside the neuron's RF. Under these conditions, the tuning curve for the neuron is scaled upward at all orientations when the monkey attends inside the neuron's RF, in a way that is consistent with the tuning curve being multiplied by a constant. Together, these effects are consistent with "biased competition" and closely related normalization models (1, 3, 11, 13, 14). By analyzing our IT data in a similar way to these previous studies, we were able to see similar attention effects on IT responses (Fig. S4 and Fig. 1*D*). However, from a population coding perspective, these effects appear to be very similar because they both create distinct patterns of neural activity that contain information about attended objects (with the patterns of activity for identity and position information overlapping one another within the same population). A consequence of this viewpoint is that the limited spatial nonuniformity/extent of a neuron's RF is not a deficit in terms of achieving complete position invariance but rather a useful property that enables more precise signaling of position information.

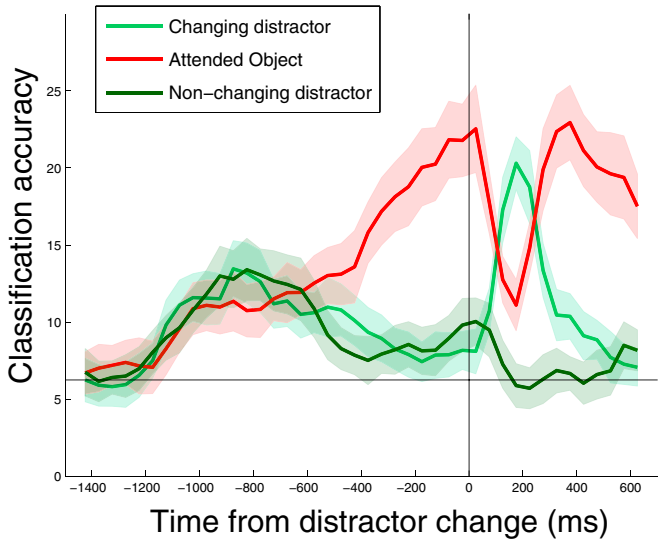

Fig. 3. Changes in the salience of distractor stimuli dominate over attention-related enhancements. A comparison of the decoding accuracies for the attended stimulus (red trace) to the distractor that underwent a color change (light green trace) and the distractor that did not undergo a color change (dark green trace). The data are aligned to the time when one of the distractors underwent a color change (black vertical bar). Chance decoding accuracy is 1/16 or 6.25%.

One discrepancy between our results and previous findings is represented by a study by Li et al. (9), which used similar population decoding methods and reported that clutter does not affect the amount of information about particular objects in IT. Although differences in stimulus parameters might be able to partially account for these effects (the stimuli used by Li et al. were smaller and presented closer to the fovea), we think that the largest factor contributing to the difference in the results was the way the classifiers were trained and tested. In particular, Li et al. trained and tested their classifier using the exact same cluttered scenes. Thus, it is possible that their classifier relied on the exact configuration in the images (by perhaps relying on visual features that spanned multiple objects) to achieve a high level of classification performance in the cluttered condition. In our study, we trained the classifier either on isolated objects (Fig. 1*B*) or using different cluttered scenes (Fig. 2) so that we would capture what is the more behaviorally relevant condition, namely being able to learn an object in isolation or on a particular background, and then being able to recognize it when seen in a different context. Indeed, when Li et al. replicated our analysis by training on isolated objects, they also found a similar decrease in classification accuracy for the cluttered conditions.

It is also important to note that the effects reported here may underestimate the impact that attention has on neural representations in IT. If the monkeys' attentional state was under stronger control by using a more difficult task (23, 24) or if the task the monkey engaged in more closely matched the information that was to be decoded (e.g., if the monkey was doing a shape discrimination task rather than a color change detection task), the effects of attention might have been even stronger. Additionally, we have seen in this study (Fig. 1*B*), and in a number of analyses of different datasets from IT, that the largest amount of information occurs when the stimuli first appear (21, 22, 25). Thus, we might also see larger attentional effects using a precuing attentional paradigm. However, we should note that even with these limitations, IT object representations with clutter

NEUROSCIENCE

were restored by attention to nearly the same level of accuracy found with isolated objects in the visual field. Finally, it should be noted that we might be able to find stronger attention-related effects by decoding data from cells that were recorded simultaneously. Indeed, a recent study by Cohen and Maunsell (4) has suggested that one of the primary ways that attention improves the signal in a population is through a decrease in noise correlations. We briefly tried to address this issue by adding noise correlations to our pseudopopulation vectors, and found that the decoding accuracies were largely unchanged (Fig. S6). However, a more detailed examination of these effects using actual simultaneously recorded data is needed before we can draw any strong conclusions.

From an algorithmic-level viewpoint, the results seen in our study are consistent with the following interpretation. Spatial attention gates signals from a retinotopic area (say V4, in which RFs are smaller than the distance between the objects in our stimuli) to IT so that the responses to clutter stimuli do not interfere with the activity elicited by the attended object in IT. Recent experimental results (26) and computational models of attention (11–20, 27, 28) are consistent with this interpretation and furthermore suggest that the clutter interference has the form of a normalization operation similar to the biased competition model. Overall, our results support the view that the main goal of attention is to suppress neural interference induced by clutter to allow higher modules to recognize an object in context after learning its appearance from presentations in isolation (or on a different background).

Methods

Experimental Procedures. Procedures were done according to National Institutes of Health guidelines and were approved by the Massachusetts Institute of Technology Animal Care and Use Committee. All unit recordings were made from anterior IT.

Visual Stimuli. The visual stimuli consisted of 16 objects from four categories (cars, faces, couches, and fruit), and are shown in Fig. S1. The stimuli were 2.3° × 2.3° in size and were shown at an eccentricity of 5.5° from fixation at angles of +60°, 0°, and −60° relative to the horizontal meridian. The stimulus sizes/locations were chosen such that there would be little overlap between the three simultaneously presented stimuli in terms of most V4 neurons' RFs (29). For the three-object displays, 864 configurations were chosen (out of the possible 3,360 permutations of three unique objects). The cluttered displays could potentially consist of either one, two, or three objects belonging to the same category. To have a variety of hard and easy displays, we selected the displays such that two-thirds of the displays (576 displays) consisted of all three objects belonging to the same category and one-third of the displays (288 displays) consisted of all three objects belonging to different categories. After analyzing the data, we did not find a large difference between these display types, and so we grouped the results from both types of categories equivalently. A second set of seven stimuli was also shown to the second monkey for additional results presented in Fig. S3 all 630 configurations of three stimuli were used for the three-object displays in this second set of experiments.

Data Selection. A total of 98 and 139 neurons were recorded from monkey 1 and monkey 2, respectively. All of the recorded neurons were used for the individual-neuron analyses (Fig. 1*D* and Figs. S3*D* and S4). For the population decoding analyses, all neurons that had at least 12 presentations of the isolated objects and 800 trials with three-object displays were included. This resulted in 75 neurons from monkey 1 and 112 neurons from monkey 2. Because the monkeys did not always complete the full experiment, not all of the neurons had recordings from all 864 three-object images. Consequently, for the decoding analyses, we only used data from the three-object images that had been shown to all of the 75/112 neurons listed above, which gave 635 three-object trials. For the data recorded on the second stimulus set, we used all neurons that had been shown 60 repetitions of the isolated-object stimuli and all 630 three-object images, which gave us 87 usable neurons of the 132 recorded.

Data Analyses. The decoding results are based on a cross-validation procedure that has previously been described (22). The decoding method works by training a pattern classifier to "learn" which patterns of neural activity are indicative that particular experimental conditions are present (e.g., which visual stimulus has been shown) using a subset of data (called the "training set"). The reliability of the relationship between these patterns of neural activity and the different conditions (stimuli) is then assessed based on how accurately the classifier can predict which conditions are present on a separate "test set" of data.

To assess how well we could decode which stimulus was shown on the isolated-object trials (Fig. 1*B*, blue trace, and Fig. 2*A*, solid blue traces), we used a cross-validation procedure that had the following steps. (*i*) For each neuron, data from 12 trials from each of the 16 stimuli were randomly selected. For each of these trials, data from all of the neurons were concatenated to create pseudopopulation response vectors (i.e., "population" responses from neurons that were recorded under the same stimulus conditions on separate trials/sessions but were treated as though they had been recorded simultaneously). Because there were 187 neurons used, this gave 12 × 16 = 192 data points in 187-dimensional space. (*ii*) These pseudopopulation vectors were grouped into 12 splits of the data, with each split containing one pseudopopulation response vector to each of the 16 stimuli. (*iii*) A pattern classifier was trained using 11 splits of the data (176 training points), and the performance of the classifier was tested using the remaining split of the data (16 test points). Before sending the data to the classifier, a preprocessing normalization method was applied that calculated the mean and SD of each feature (neuron) using data from the training set, and a z-score normalization was applied to the training data and the test data using these means and SDs. This normalization method prevented neurons with high firing rates from dominating the outcome of the classifier. (*iv*) This procedure was repeated 12 times, leaving out a different test split each time (i.e., a 12-fold leave-one-split-out cross-validation procedure was used). (*v*) The classification accuracy from these different splits was evaluated using a measure based on the area under an ROC curve (see *SI Text* for a more detailed description of this measure). Different measures of decoding accuracy gave similar results (Fig. S7). (*vi*) The whole procedure [steps (*i*)–(*v*)] was repeated 50 times (which allowed us to assess the performance for different pseudopopulations and data splits), and the final results were averaged over all 50 repetitions.

To generate the SEs of the decoding accuracy, we used a bootstrap method that applied the above decoding procedure but created pseudopopulation vectors that sampled the neurons with replacement (being careful not to include any of the same data in the training and test sets). The SE was then estimated as the standard deviation of the mean decoding accuracy over the 50 bootstrap runs, which gave an estimate of the variability that would be present if a different subset of neurons had been selected from a similar population. Unless otherwise specified below, the decoding results in this paper are based on using a correlation coefficient classifier that was trained on the mean firing rate from 500 ms of data that started 85 ms after the onset of the stimulus, and the classifier was tested using the mean firing rates in 150-ms bins that were sample at 50-ms intervals (this created smooth curves that estimated the amount of information present in the population as a function of time). In contrast to some of the results of Meyers et al., (21), we found the neural representations in this study to be largely stationary (Fig. S8), which allowed us to use data from one training time period to decoding information at all other time points.

A similar method was used for the other decoding results reported here. To obtain the decoding accuracies for the cluttered-display objects (red and green traces in Fig. 1*B*, and also solid red traces in Fig. 2*B*), the classifier was trained on isolated-object trials using 11 repetitions of each of the 16 objects exactly as described above, but the classifier was then tested using the clutter-display trials, and the accuracies for the attended and nonattended objects were measured separately [also, because the data in the test set came from a completely different set of trials there was no need to divide the data into separate splits, so all test points were evaluated in one step (i.e., step [*iv*] was omitted)]. For the dashed traces in Fig. 2, we trained the classifier using data from the cluttered trials (again using 11 trials from each of the 16 stimuli to make a fair comparison), and the classifier was then tested using either isolated-object data (blue dashed lines in Fig. 2*A*) or the remaining cluttered-display trials that had not been used to train the classifier (dashed red lines in Fig. 2*B*). The training data for Fig. 2 were from the mean firing rates of either 300 ms of data that started 100 ms after stimulus onset (left plots) or 310 ms of data that started 200 ms after cue onset (right plots).

For the decoding of position information (Fig. 1*C*), the classifier was trained using the firing rates from isolated-object trials from a 300-ms bin that started 100 ms after the stimulus onset (thus avoiding the possibility that any visual information in the cue itself could influence the results). The results were based on a threefold cross-validation scheme, where each split contained each stimulus at all three locations (i.e., 96 total training points, and

48 test points on each split). The results from decoding the location of attention in Fig. 1*C* were based on using the same isolated-object training paradigm but the classifier was tested with cluttered displays that had 12 repetitions of each of the 16 attended stimuli at all 3 locations (576 test points). The results in Fig. 3 were based on training the classifier on isolated-object trials using 500 ms of data and 11 training points (i.e., the same paradigm used for Fig. 1*B*). The classifier was tested on 288 data points (16 stimuli $\times$ 18 repetitions) using data from the cluttered trials that were aligned to the time that the distractor underwent a color change (using 150-ms bins sampled at 50-ms intervals), and the results were compiled separately for the attended stimulus (red trace), the distractor stimulus that underwent a color change (light green trace), and the other distractor stimulus that did not undergo a color change (dark green trace).

As described in more detail in the *Discussion*, we used the simpler and more commonly used zero-one loss decoding measure (21, 22) for Figs. 1*C* and 3 because we did not need to compare attended and nonattended conditions (although the results were very similar when an AUROC measure was used). The methods used to calculate the AUROC and zero-one loss values, evaluate the statistical significance of the results, and create the supplemental figures are described in *SI Text*.

Fig. 1*D* was created by using isolated-object trials to find the position that elicited the highest and lowest firing rate for each neuron and then assessing the best to worst stimulus using 500 ms of data from the array period. These tuning curves were then plotted for the attended object using 300 ms of data from cluttered trials when attention was directed to either the best or worst position. Each neuron's firing rate was z-score–normalized before being averaged together, so that neurons with higher firing rates did not dominate the population average.

ACKNOWLEDGMENTS. We thank Jim DiCarlo and Nuo Li for their comments and for conducting additional analyses on data they collected. This research was sponsored by Defense Advanced Research Planning Agency grants (Information Processing Techniques Office and Defense Sciences Office), National Science Foundation Grants NSF-0640097 and NSF-0827427, and National Eye Institute Grant R01EY017292. Additional support was provided by Adobe, Honda Research Institute USA, and a King Abdullah University Science and Technology grant to B. DeVore. E.M.M. was supported by a National Defense Science and Engineering Graduate Research Fellowship and by a Herbert Schoemaker fellowship.

1. McAdams CJ, Maunsell JH (1999) Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J Neurosci* 19:431–441.
2. Fries P, Reynolds JH, Rorie AE, Desimone R (2001) Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* 291:1560–1563.
3. Moran J, Desimone R (1985) Selective attention gates visual processing in the extrastriate cortex. *Science* 229:782–784.
4. Cohen MR, Maunsell JHR (2009) Attention improves performance primarily by reducing interneuronal correlations. *Nat Neurosci* 12:1594–1600.
5. Motter BC (1993) Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *J Neurophysiol* 70: 909–919.
6. Reynolds JH, Chelazzi L, Desimone R (1999) Competitive mechanisms subserve attention in macaque areas V2 and V4. *J Neurosci* 19:1736–1753.
7. Mitchell JF, Sundberg KA, Reynolds JH (2009) Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. *Neuron* 63:879–888.
8. Mitchell JF, Sundberg KA, Reynolds JH (2007) Differential attention-dependent response modulation across cell classes in macaque visual area V4. *Neuron* 55:131–141.
9. Li N, Cox DD, Zoccolan D, DiCarlo JJ (2009) What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? *J Neurophysiol* 102:360–376.
10. Marr D (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (Henry Holt, New York).
11. Lee J, Maunsell JHR (2009) A normalization model of attentional modulation of single unit responses. *PLoS One* 4:e4651.
12. Chikkerur S, Serre T, Tan C, Poggio T (2010) What and where: A Bayesian inference theory of attention. *Vision Res* 50:2233–2247.
13. Reynolds JH, Heeger DJ (2009) The normalization model of attention. *Neuron* 61:168–185.
14. Reynolds JH, Desimone R (1999) The role of neural mechanisms of attention in solving the binding problem. *Neuron* 24:19–29.
15. Lee DK, Itti L, Koch C, Braun J (1999) Attention activates winner-take-all competition among visual filters. *Nat Neurosci* 2:375–381.
16. Hamker FH (2005) The reentry hypothesis: The putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. *Cereb Cortex* 15:431–447.
17. Ardid S, Wang XJ, Compte A (2007) An integrated microcircuit model of attentional processing in the neocortex. *J Neurosci* 27:8486–8495.
18. Börgers C, Epstein S, Kopell NJ (2008) Gamma oscillations mediate stimulus competition and attentional selection in a cortical network model. *Proc Natl Acad Sci USA* 105:18023–18028.
19. Tiesinga PH, Fellous JM, Salinas E, José JV, Sejnowski TJ (2004) Inhibitory synchrony as a mechanism for attentional gain modulation. *J Physiol Paris* 98:296–314.
20. Tsotsos JK (1988) in *Computational Processes in Human Vision: An Interdisciplinary Perspective* (Ablex, Norwood, NJ), pp 286–338.
21. Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. *Science* 310:863–866.
22. Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T (2008) Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J Neurophysiol* 100:1407–1419.
23. Boudreau CE, Williford TH, Maunsell JHR (2006) Effects of task difficulty and target likelihood in area V4 of macaque monkeys. *J Neurophysiol* 96:2377–2387.
24. Spitzer H, Desimone R, Moran J (1988) Increased attention enhances both behavioral and neuronal performance. *Science* 240:338–340.
25. Gochin PM, Colombo M, Dorfman GA, Gerstein GL, Gross CG (1994) Neural ensemble coding in inferior temporal cortex. *J Neurophysiol* 71:2325–2337.
26. Buffalo EA, Bertini G, Ungerleider LG, Desimone R (2005) Impaired filtering of distracter stimuli by TE neurons following V4 and TEO lesions in macaques. *Cereb Cortex* 15:141–151.
27. Deco G, Rolls ET (2004) A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Res* 44:621–642.
28. Spratling MW (2008) Predictive coding as a model of biased competition in visual attention. *Vision Res* 48:1391–1408.
29. Gattass R, Sousa AP, Gross CG (1988) Visuotopic organization and extent of V3 and V4 of the macaque. *J Neurosci* 8:1831–1845.

Incorporation of new information into prefrontal cortical activity after learning working memory tasks

Ethan M. Meyers[a,1], Xue-Lian Qi[b], and Christos Constantinidis[b]

[a]Department of Brain and Cognitive Sciences, McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Department of Neurobiology and Anatomy, Wake Forest University School of Medicine, Winston-Salem, NC 27157

Edited by Ranulfo Romo, Universidad Nacional Autonoma de Mexico, Mexico City, D.F., Mexico, and approved February 7, 2012 (received for review January 19, 2012)

The ability to learn new tasks requires that new information is integrated into neural systems that already support other behaviors. To study how new information is incorporated into neural representations, we analyzed single-unit recordings from the prefrontal cortex (PFC), a brain region important for task acquisition and working memory, before and after monkeys learned to perform two behavioral tasks. A population-decoding analysis revealed a large increase in task-relevant information, and smaller changes in stimulus-related information, after training. This new information was contained in dynamic patterns of neural activity, with many individual neurons containing the new task-relevant information for only relatively short periods of time in the midst of other large firing rate modulations. Additionally, we found that stimulus information could be decoded with high accuracy only from dorsal PFC, whereas task-relevant information was distributed throughout both dorsal and ventral PFC. These findings help resolve a controversy about whether PFC is innately specialized to process particular types of information or whether its responses are completely determined by task demands by showing there is both regional specialization within PFC that was present before training, as well as more widespread task-relevant information that is a direct result of learning. The results also show that information is incorporated into PFC through the emergence of a small population of highly selective neurons that overlay new signals on top of patterns of activity that contain information about previously encoded variables, which gives insight into how information is coded in neural activity.

neural coding | task learning | Macaque | vision | principal sulcus

The prefrontal cortex (PFC) is a brain region involved in planning, decision making, working memory, and learning new context dependent behaviors (1–3). Although many studies have found task-related activity in the PFC for a variety of different behaviors (4–6), it is often unclear whether this task-related information had always existed in PFC or whether it emerged as a result of learning. Furthermore, in studies in which it seems likely that new information arises as a result of training (7, 8), how this new information interacts with preexisting information is not well understood. Given that we must continually learn to perform new tasks, while simultaneously maintaining the ability to perform previously learned behaviors, understanding how new information is integrated into existing neural processing is fundamental to understanding how the brain enables complex and adaptive human behaviors.

To gain insight into how learning a new task affects processing in PFC, we analyzed single-unit activity from neurons before and after two monkeys were trained to perform two distinct delayed match-to-sample tasks (9, 10) (Fig. 1 and Fig. S1). Using a neural-population decoding analysis, we were able to directly assess what information was represented in the population before training and what new information arose because of learning a new task. Our analyses sought to examine several questions concerning the content and coding of information including: (*i*) Does learning a new task change the amount of information about basic stimulus features or does it only change the amount of information about more complex task-related variables? (*ii*) Does the new information arise because of the emergence of a few highly selective neurons or is information evenly distributed across the population? (*iii*) Do neurons become specialized to process only one type of information, as suggested by some studies (11), or can the same neuron carry multiple types of information as other studies suggest (12)? (*iv*) Is the new information contained in a dynamic population code (13–15), or is there one stationary pattern of neural activity that contains the new information? (*v*) Are there differences in the information content between dorsal and ventral PFC, and does learning affect these two brain regions equally (16–20)? Thus, this work gives insight into how new information is incorporated in neural systems and helps clarify the key computations that are occurring in PFC (21, 22).

Results

Neural recordings were made from two monkeys while they passively viewed a sequence of two stimuli and after they were trained on two delayed-match-to-sample tasks. Before training, the monkeys fixated a central point and passively observed two stimuli that were separated by a delay (Fig. 1*C*). After training, the monkeys viewed the same sequence of stimuli and made a saccade that indicated whether the two stimuli were identical (Fig. 1*D* and Fig. S1*A*). In the "feature task," the monkeys indicated whether the symbols where the same (Fig. 1*D*); in the "spatial task," the monkeys indicated whether the square symbol appeared at the same location (Fig. S1). Neurons were sampled with an unbiased procedure, recording from all neurons that could be isolated, and a decoding procedure was used that jointly analyzed the activity of 750 neurons at a time recorded in separate sessions (*Methods*). Results from the feature task are shown in the main text of the article (*Results*), and the results from the spatial task are shown in the supplemental figures (*SI Text*). (Overall, the results were very similar between the two tasks.)

Our analysis examined the amount of information that could be decoded from the neuronal population (evaluated as the performance of a linear classifier) about stimulus identity or location (i.e., which of eight stimuli were shown in the feature and spatial tasks) and about the match/nonmatch status of a trial. We observed little difference in the amount of information about the identity of the first stimulus before training (Fig. 2*A*, blue trace) compared with after training (Fig. 2*A*, red trace). In contrast, there was a massive increase in information about the match/nonmatch status of the trial in PFC after training. Before

Author contributions: X.-L.Q. and C.C. designed research; E.M.M., X.-L.Q., and C.C. performed research; E.M.M. contributed new reagents/analytic tools; E.M.M. analyzed data; and E.M.M. and C.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: emeyers@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1201022109/-/DCSupplemental.

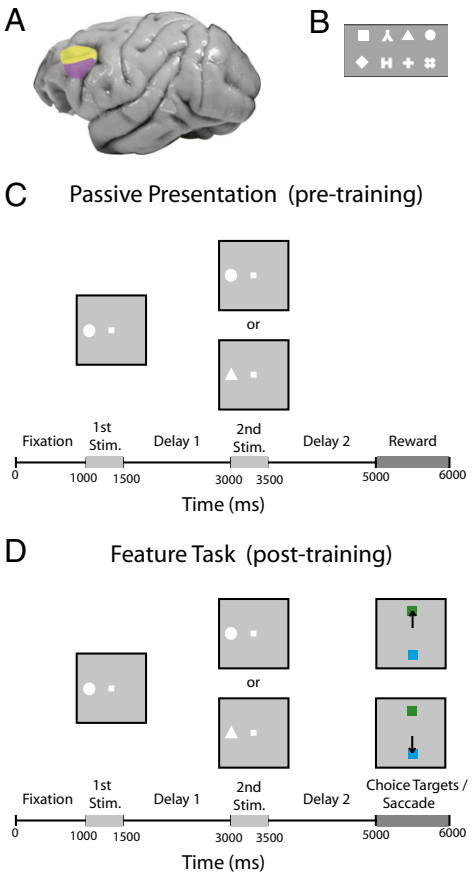

Fig. 1. Brain regions and the feature task. (*A*) Dorsal (yellow) and ventral (magenta) regions of lateral PFC where the recordings were made. (*B*) Stimuli used in the feature task. The stimuli extended 2° of visual angle. (*C*) Passive fixation task that was used before training. (*D*) Feature task. The monkeys viewed the same sequence of images as in the passive task; however, at the end of the experiment, monkeys needed to make a saccade to the green target if the stimuli matched or to the blue target if the stimuli did not match.

training, when the match/nonmatch status of a trial was irrelevant for the task, we could not decode information about whether the two stimuli matched in shape at accuracies that were above chance (Fig. 2*B*, blue trace). However, after training, when the match/nonmatch status was critical for correctly completing the task, we could decode this information at above-chance levels shortly after the onset of the second stimulus [permutation test, $P < 0.005$; see colored bars at the bottom of Fig. 2 *A* and *B*], and it was possible to decode this information with close to 100% accuracy during most of the following delay period (Fig. 2*B*, red trace). This increase in match/nonmatch information across the population was attributable to a small subset of neurons that became highly selective for match/nonmatch information after training (points above horizontal line in Fig. 2*C*), as well as a larger number of neurons that showed a small increase in their match/nonmatch selectivity. In fact, these highly selective neurons were so informative that the top eight most selective neurons contained almost all of the information that was present in the entire population (Fig. S2*A*). However, the less strongly selective neurons still contained significant amounts of redundant information, as evidenced by the fact that when we excluded the top 128 most selective neurons, we still obtained above-chance decoding accuracies (Fig. S2*B*). Similar results were observed for spatial stimuli (Fig. S3): equivalent levels of stimulus information could be decoded during the first stimulus presentation before and after training, and an increase in the match/nonmatch information was observed after training. A subtle difference between tasks was an increase in position information in anticipation of the second stimulus presentation after training in the spatial task.

Recent work has shown that neural activity in PFC shows complex temporal dynamics, with individual neurons changing the way they code information over the course of a trial (13–15). To test whether the task-relevant match/nonmatch information that emerged after training was also encoded by dynamic population activity, we applied a decoding analysis in which we trained the classifier using data from one time period (as indicated by the *y* axis on Fig. 3*A* and Fig. S4*A*) and tested the classifier using data from a different time period (as indicated by the *x* axis on Fig. 3*A* and Fig. S4*A*). If the information is represented by a stationary pattern of activity (i.e., if patterns of neural activity that encode the match/nonmatch trial status are the same at all time points), then training the classifier using one time point with high information should lead to high classification accuracy at all other time points where the information is present. Conversely, if information is represented by dynamic patterns of neural activity, then training the classifier at any one time point with high information should lead to high classification accuracy only at that time point. Fig. 3*A* (*Right*) clearly shows that high decoding accuracy is only obtained when the classifier is trained and tested on data from the same time point relative to stimulus onset. Over the second delay period, the decoding accuracy dropped from 98% correct to 68% correct when the classifier was trained on data taken 500 ms before the time the classifier was tested. Thus, new task-relevant information that emerges as a result of training is contained by a dynamic pattern of neural activity. The dynamic nature of the information coding was also evident in the mean firing rates of individual neurons. Examining activity of highly selective match/nonmatch neurons (Fig. 3*B* and Fig. S4*B*) revealed that the task-relevant information in several neurons was present for only short periods of time relative to the duration of the second delay period stimulus when this information was present (e.g., the middle neuron in Fig. 3*B* and the rightmost neuron in Fig. S4*B*), giving rise to the dynamic coding of match/nonmatch information seen at the population level. Additionally, firing-rate modulations that occurred throughout the trial were attributable to individual neurons carrying information about different variables at different points, as can be seen by the fact that the highly match/nonmatch selective neurons also contained large amounts of stimulus identity information (Fig. S5). Thus, task-relevant information is incorporated into PFC by interleaving/overlapping new information into ongoing dynamic activity that is carrying information about other variables and consequently the absolute firing rate level of a single neuron at a particular time point is often highly ambiguous if the context of the larger population is not taken into account.

The results presented thus far combined data from dorsal PFC (areas 46 and 8a) and ventral PFC (areas 12 and 45) (Fig. 1*A*). Previous research has led to conflicting findings about whether there are distinct functional differences between dorsal and ventral PFC (16–20), despite the clear anatomical connectivity differences between these two regions (23). To assess whether these brain areas have similar amounts of stimulus identity and match/nonmatch information, we applied the same decoding analyses separately to data from each of these areas. The results show that both dorsal and ventral PFC contained above-chance match/nonmatch information after the training (Fig. 4*A*; permutation test, $P < 0.005$). In contrast, information about the identity of the stimuli was largely confined to dorsal PFC, at least for the limited stimulus set that we tested neurons with (Fig. 4*B*; permutation test, $P < 0.005$). Similar results were seen in the spatial task (Fig. S6) (although the onset of match/nonmatch information was a bit longer in ventral PFC in the spatial task).

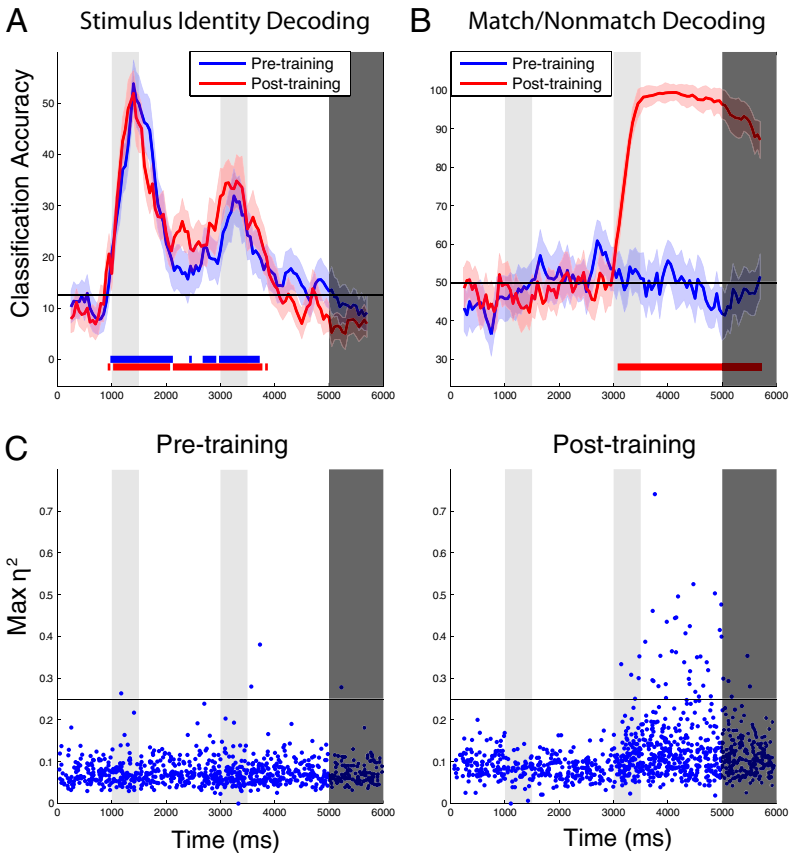

Fig. 2. Information in PFC pre- and posttraining in the feature task. (*A*) Comparison of information about the identity of the first stimulus pretraining (blue) and posttraining (red). The gray shaded regions indicate the times when the first, second, and decision stimuli were shown, the black horizontal line indicates the level of decoding expected by chance, the color shaded regions indicate 1 SE in the decoding accuracy if different neurons were used, and the red and blue bars at the bottom of the figure indicate times when the decoding accuracy was above chance (permutation test, $P < 0.005$). As can be seen, training had little effect on stimulus identity information. (*B*) Comparison of information about the match/nonmatch trial status pretraining (blue) and posttraining (red). As can be seen, there is a large increase in match/nonmatch status information after training. *C*, Match/nonmatch selectivity of individual neurons before training (*Left*) and after training (*Right*). (The black horizontal line is for visualization purposes to make the pre- and posttraining differences easier to compare.) The η^2 statistic measures the proportion of the trial-by-trial variance in firing rates explained by whether a trial is a match or a nonmatch trial. Each point corresponds to the η^2 value of a single neuron at the latency when the neuron had its maximal selectivity (see *Methods*).

Thus, there are significant differences in how basic stimulus information is processed by these brain regions, whereas the newly learned task-relevant information was more distributed. Evidence also exists for variations in anatomical structure and function within the dorsal and ventral prefrontal cortex (23, 24). In our data set, comparison of area 46 with 8a revealed no qualitative differences within the dorsal PFC (Fig. S7). The match/nonmatch information was not restricted to the prefrontal cortex: recordings made from the posterior parietal cortex (areas LIP and 7a) of one of the monkeys after training also revealed a similar level of match/nonmatch information in that brain region (Fig. S8), indicating that the task-relevant information could potentially be wide-spread across several cortical areas.

Discussion

The results presented above give insight into how new task-relevant information is incorporated into existing processing, in different regions of PFC. Our results show that whereas there was little change in the amount of basic stimulus information before training (Fig. 2*A*), more complex information about whether the stimuli matched became present throughout PFC only after this information became relevant to the monkeys' behavior (Fig. 2*B*). Additionally, our analyses revealed that the majority of basic stimulus information was restricted to dorsal PFC (Fig. 4*B*), whereas the new task-relevant information was much more widely distributed (Fig. 4*A* and Fig. S8). The fact that all regions of PFC contained the new task-relevant information is consistent with the adaptive coding model, which proposes that PFC can adapt to encode information about any property that is relevant for behavior (3). Additionally, the finding that basic visual information was restricted to dorsal PFC is consistent with domain-specific theories of PFC that claim that there are differences between different regions of lateral PFC (16, 17, 21) and raises questions about the validity of strict integrative theories, which claim there are no regional differences (18, 19). Anatomical studies have found that the cortical areas that project to dorsal PFC are different from those that project to ventral PFC and that there are extensive intra-area connections within PFC (23, 25). Based on our results and these anatomical findings, we hypothesize that the long-range anatomical connections between PFC and other cortical regions constrain the types of information that PFC can encode about basic stimulus properties and that, through learning, the task-relevant information becomes distributed more broadly via the intraarea connections within PFC.

Our results also give insight into how new task-relevant information is coded at the population and individual neuron level. At the population level, we observed that the new information was contained in a dynamic population code (13–15), with different neurons carrying information at different latencies relative to the start of the trial (Fig. 3*A*). These time-dependent representations

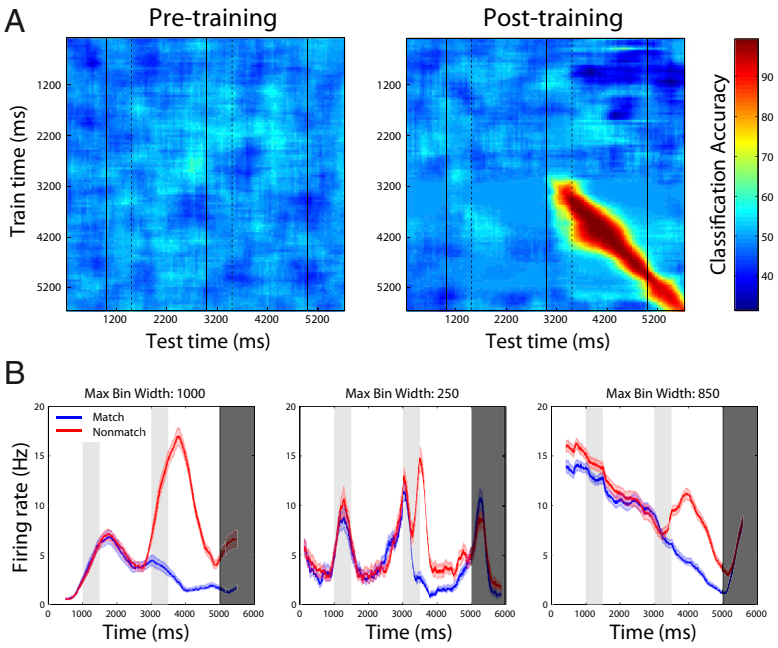

Fig. 3. Dynamic coding of task relevant information after training in the feature task. (*A*) Results from training a classifier at one time period (*y* axis) and testing the classifier at a second time period (*x* axis) for decoding the match/nonmatch trial status, either pretraining (*Left*) or posttraining (*Right*). The black solid vertical lines indicate the times when the first, second, and match/nonmatch stimuli appeared, and the black dashed lines indicate the offset times of the first and second stimuli. After the monkey was trained, high classification accuracies are seen only when the classifier is built and tested using data from around the same time periods, which shows that different patterns of neuron activity contain the task relevant information at different time points in the experiment. (*B*) Firing rates for the match trials (blue) and nonmatch trials (red) for the three most selective neurons in feature task. As can be seen, differences in firing rates between match and nonmatch trials appear to be added on top of other firing rate changes that are occurring over the course of a trial (and are carrying information about other variables). Additionally, some neurons (e.g., *Middle*) only contain large firing rate differences between match and nonmatch trials for short periods of time, which give rises to the dynamic coding of information at the population level. (These neurons are typical examples of the population of neurons that have large amounts of match/nonmatch information). Error bars indicate 1 SEM.

could enable the PFC to keep track of when particular events occurred and thus might be involved in the neural representation of time. Our population analyses also revealed that a small subset of highly match/nonmatch selective neurons emerged after training and these neurons contained almost all of the task-relevant information that was present in the larger population (Fig. 2*C* and Fig. S2*A*). These results raise the possibility that only a small percentage of the population might be critical for processing particular types of information at any one point in time and that the redundant information seen in the rest of the population (Fig. S2*B*) could help make the circuit more robust in the face of damage to these highly selective neurons. This finding also has implications for the way in which neural data are analyzed; methods that rely exclusively on average selectivity of neurons over the whole population may miss the importance of such highly selective neurons.

Finally, at the single-neuron level, we observed that the new task-relevant information was often contained in relatively short time windows that were present in the midst of other large firing-rate modulations that occurred throughout the trial (Fig. 3*B* and Fig. S4*B*). These findings suggest that, unlike the results reported in other brain regions (27), information in PFC is not coded solely by the maximum firing rate of a neuron (e.g., see the rightmost neuron in Fig. 3*B*), but, rather, that the firing rates of neurons need to be evaluated in relation to the activity of other neurons in the population. Additionally, we observe that these other firing-rate modulations carry information about other variables (Fig. S5), which shows individual neurons in PFC are multiplexing different types of information in a single spiking sequence. Such multiplexing of information could be an efficient strategy that allows the large number of time-dependent representations in PFC to be encoded by a much smaller number of neurons. Overall, these findings shed light on how novel information is incorporated into PFC activity and how neural activity codes information, which should lead to richer theories of how PFC controls behavior and how information is coded in neural activity more generally.

Methods

Recording Methods and Task. Before training, 726 and 111 neurons were recorded from the 2 rhesus monkeys (*Macaca mulatta*) while they passively viewed the stimuli in the feature task, and 810 and 210 neurons were recorded after training (for the spatial task, 595 and 113 neurons were recorded before training; 814 and 214 were recorded after training). A grid system was used for the recordings, and a map of penetrations was generated by aligning the placement of the electrodes within the grid to a magnetic resonance image of the cortical surface. Dorsal PFC in this study was defined as the area containing the two banks of the principal sulcus (≤ 2mm from the principal sulcus) and extending posterior to the arcuate sulcus, which incorporates the posterior aspect of area 46 and parts of area 8a. Ventral PFC was defined as the area in the convexity of the PFC lateral to the principal sulcus (>2 mm from the center of the principal sulcus), thus incorporating parts of areas 12 and 45. Neurons were not prescreened for stimulus selectivity; however, if a recorded neuron had a restricted receptive field, the stimuli in the feature task were typically presented in its center. The average signal-to-noise ratio of spike waveforms (in neurons that were significantly modulated by task events) was 8.0 in the pretraining population and 7.2 in the posttraining population (10). To complete the task, the monkeys needed to maintain fixation within 2° of the center of the screen. After training, the monkeys additionally needed to saccade to a green target stimulus on match trials and a blue stimulus on nonmatch trials (the location of the green and blue stimuli were randomly counterbalanced across trials, so that the decoding match/nonmatch information was not merely reflecting a planned movement direction). The stimuli were paired in each experimental session, so that, on nonmatch trials, the first stimulus was always shown with the same second nonmatching stimulus. The surgical procedures, recording methods, task details, and anatomical localization methods have been described previously (9, 10).

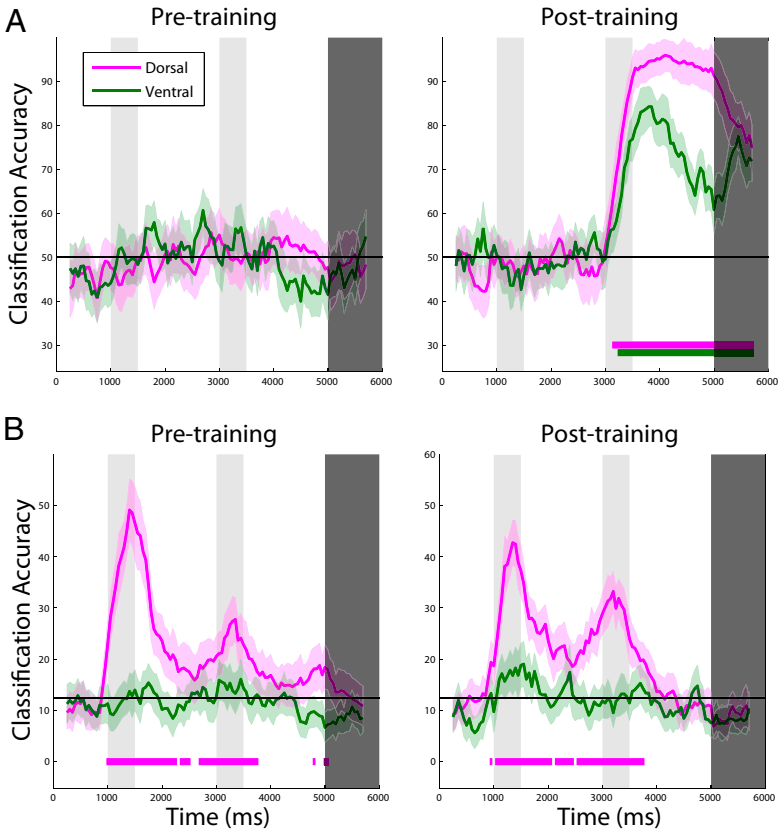

Fig. 4. Comparing information in dorsal PFC (magenta) vs. ventral PFC (green). (*A*) Match/nonmatch information pretraining (*Left*) and posttraining (*Right*) reveals that, after training, there is task relevant match/nonmatch information in both dorsal and ventral PFC. (*B*) In contrast, information about which stimulus was shown (stimulus identity information) was seen only in dorsal PFC in both the pretraining and posttraining data (left and right plots, respectively).

Data Analysis. The decoding analysis methods have been described previously (14, 28, 29). Briefly, a resample cross-validation procedure was used in which a maximum correlation coefficient classifier was trained on firing rates of pseudopopulations of neurons (i.e., populations of neurons that were recorded independently but treated as if they were recorded simultaneously). The firing rates were the average spiking activity in 500-ms bins sampled at 50-ms sliding intervals. This classifier was then used to decode the stimulus identity on the feature task and the stimulus position on the spatial task (Figs. 2*A* and 4*B* and Figs. S3*A* and S6*B*). For these tasks, 11 trials of each stimulus were used for training the classifier, and testing was done using one trial for each stimulus. Eight different stimuli were used in the feature task, and eight different locations were used in the spatial task, so chance decoding on these tasks was 1/8 = 12.5%. It should be noted that because for a given neuron the first stimulus was always paired with the same nonmatch stimulus on nonmatching trials, information about the first stimulus after the time when the second stimulus was shown (e.g., Fig. 2*A* and Figs. S3*A* and S7*B*) could be attributable to the second stimulus (30). When decoding the match/nonmatch trial status on the feature task (Fig. 2*B*, 3*A*, and 4*A*), 44 match and nonmatch trials were used for training the classifier and 4 trials from each condition were used for testing. When decoding the match/nonmatch trial status on the spatial task (Figs. S3*B*, S4*A*, and S6*A*), 48 match and nonmatch trials were used for training and 4 trials from each condition were used for testing. The chance decoding accuracy for match/nonmatch information was 1/2 = 50%. All neurons that had recordings from 12 repetitions of each stimulus on were included in the identity decoding analysis, and all neurons that have 48 (52) repetitions on the match/nonmatch feature (spatial) task were used. This led to at least 84% of the recordings being used for all analyses. To have maximum power in our results, data from both monkeys were combined in all analyses. To make a fair comparison in the decoding analyses, 750 neurons were randomly selected from the pre- and posttraining datasets and passed to the classifier (on the spatial task, only 600 neurons were used). When comparing the decoding accuracies for dorsal vs. ventral neurons, 250 neurons were used in the feature task and 200 neurons were used in the spatial task. This procedure was repeated 50 times using different neurons, creating different pseudopopulations and using different training/ test splits each time, and the results were averaged over these 50 runs. The error bars were estimates of 1 SEM decoding accuracy that would occur if different neurons had been selected and were created by sampling the 750 (or 600) neurons with replacement, and taking the SD of the decoding results over 50 bootstrap runs. To evaluate whether the decoding results were above chance, the labels from the trials were randomly shuffled and the decoding procedure was run using 10 bootstrap iterations. This procedure was repeated 200 times to get a null distribution of the decoding accuracies that would occur by chance, and significant time periods were defined as those in which the real decoding accuracy exceeded all of the values in the null distribution [i.e., $P <$ 1/200 ($P < 0.005$)].

To find the most selective neurons in Figs. S2, S3*D*, and S3*E*, an ANOVA was applied to all of the training data for each neuron, and the neurons with the smallest *P* value were classified as the most selective neurons. The effect size for the selectivity of individual neurons (Fig. 2*C* and Fig. S3*C*) was calculated using the η^2 statistic, which measures the proportion of the variance explained by the match/nomatch labels (i.e., η^2 is the between-class sum of squares divided by the total sum of squares). The neurons selected in Fig. 3*B* were chosen by calculating the ANOVA *P* values using firing rates in bin sizes from 50–1,000 ms, sampled every 5 ms, and choosing the neurons with the smallest *P* values. Because neurons have different windows of selectivity, the smoothing bin size for these neurons was based on the bin size that led to the smallest *P* value. (The bin sized used for smoothing is shown above the firing rate plots for each neuron.) To calculate the decrease in decoding accuracy when training and testing the classifier at different times, the delay period was defined as the time period 250 ms after the offset of the second stimulus to 250 ms before the onset of the decision stimuli, which corresponds to 4,250–4,750 ms after the start of the trial.

ACKNOWLEDGMENTS. We thank Tomaso Poggio for his guidance and Gabriel Kreiman, Beata Jarosiewicz, and Ram Ramachandran for their comments on the paper. This research was sponsored by the following grants: National Institutes of Health Grant EY017077, National Science Foundation Grants 0640097 and 0827427, the Defense Advanced Research Projects Agency Defense Sciences Office, and Air Force Office of Scientific Research Grants FA8650-50-C-7262 and FA9550-09-1-0606. Additional support was provided by The Tab Williams Family Endowment Fund, Adobe, the Honda Research Institute, a King Abdullah University Science and Technology grant (to B. DeVore), NEC, Sony, and especially by The Eugene McDermott Foundation.

1. Miller EK (2000) The prefrontal cortex and cognitive control. *Nat Rev Neurosci* 1: 59–65.
2. Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24:167–202.
3. Duncan J (2001) An adaptive coding model of neural function in prefrontal cortex. *Nat Rev Neurosci* 2:820–829.
4. Nieder A, Freedman DJ, Miller EK (2002) Representation of the quantity of visual items in the primate prefrontal cortex. *Science* 297:1708–1711.
5. Romo R, Brody CD, Hernández A, Lemus L (1999) Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* 399:470–473.
6. Romanski LM, Goldman-Rakic PS (2002) An auditory domain in primate prefrontal cortex. *Nat Neurosci* 5:15–16.
7. Shima K, Isoda M, Mushiake H, Tanji J (2007) Categorization of behavioural sequences in the prefrontal cortex. *Nature* 445:315–318.
8. Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291:312–316.
9. Meyer T, Qi X-L, Stanford TR, Constantinidis C (2011) Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks. *J Neurosci* 31: 6266–6276.
10. Qi X-L, Meyer T, Stanford TR, Constantinidis C (2011) Changes in prefrontal neuronal activity after learning to perform a spatial working memory task. *Cereb Cortex* 21: 2722–2732.
11. Roy JE, Riesenhuber M, Poggio T, Miller EK (2010) Prefrontal cortex activity during flexible categorization. *J Neurosci* 30:8519–8528.
12. Cromer JA, Roy JE, Miller EK (2010) Representation of multiple, independent categories in the primate prefrontal cortex. *Neuron* 66:796–807.
13. Machens CK, Romo R, Brody CD (2010) Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *J Neurosci* 30:350–360.
14. Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T (2008) Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J Neurophysiol* 100:1407–1419.
15. Warden MR, Miller EK (2007) The representation of multiple objects in prefrontal neuronal delay activity. *Cereb Cortex* 17(Suppl 1):i41–i50.
16. Goldman-Rakic PS (1996) Regional and cellular fractionation of working memory. *Proc Natl Acad Sci USA* 93:13473–13480.
17. Wilson FA, Scalaidhe SP, Goldman-Rakic PS (1993) Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science* 260:1955–1958.
18. Rao SC, Rainer G, Miller EK (1997) Integration of what and where in the primate prefrontal cortex. *Science* 276:821–824.
19. Rainer G, Asaad WF, Miller EK (1998) Memory fields of neurons in the primate prefrontal cortex. *Proc Natl Acad Sci USA* 95:15008–15013.
20. Buckley MJ, et al. (2009) Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. *Science* 325:52–58.
21. O'Reilly RC (2010) The what and how of prefrontal cortical organization. *Trends Neurosci* 33:355–361.
22. Wilson CRE, Gaffan D, Browning PGF, Baxter MG (2010) Functional localization within the prefrontal cortex: Missing the forest for the trees? *Trends Neurosci* 33:533–540.
23. Preuss TM, Goldman-Rakic PS (1991) Myelo- and cytoarchitecture of the granular frontal cortex and surrounding regions in the strepsirhine primate Galago and the anthropoid primate Macaca. *J Comp Neurol* 310:429–474.
24. Badre D, D'Esposito M (2009) Is the rostro-caudal axis of the frontal lobe hierarchical? *Nat Rev Neurosci* 10:659–669.
25. Pandya DN, Yeterian EH (1990) Prefrontal cortex in relation to other cortical areas in rhesus monkey: Architecture and connections. *Prog Brain Res* 85:63–94.
26. Tanji J, Hoshi E (2008) Role of the lateral prefrontal cortex in executive behavioral control. *Physiol Rev* 88:37–57.
27. Gold JI, Shadlen MN (2007) The neural basis of decision making. *Annu Rev Neurosci* 30:535–574.
28. Zhang Y, et al. (2011) Object decoding with attention in inferior temporal cortex. *Proc Natl Acad Sci USA* 108:8850–8855.
29. Meyers EM, Kreiman G (2001) Tutorial on pattern classification in cell recording. Visual Population Codes, eds Kriegeskorte N, Kreiman G (MIT Press, Cambridge, MA), pp 517–538.
30. Meyers et al.; The incorporation of new information into prefrontal cortical activity after learning working memory tasks. Available at http://cbcl.mit.edu/people/emeyers/pnas2012/.