# Gradual progression from sensory to task-related processing in cerebral cortex

Scott L. Brincat[1*], Markus Siegel[1,2*], Constantin von Nicolai[2], and Earl K. Miller[1]

[1]The Picower Institute for Learning and Memory, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139.
[2] Centre for Integrative Neuroscience & MEG Center, University of Tübingen, 72076 Tübingen, Germany

* - equal contribution

## Author contributions

M.S. and E.K.M. designed the experiments. M.S. performed the experiments and recorded the data. C.v.N. and S.L.B. spike-sorted the data. S.L.B. curated the data, and conceived and performed the analyses. S.L.B and E.K.M. wrote the manuscript.

## Classification

BIOLOGICAL SCIENCES : Neuroscience

## Abstract

Somewhere along the cortical hierarchy, behaviorally relevant information is distilled from raw sensory inputs. We examined how this transformation progresses along multiple levels of the hierarchy by comparing neural representations in visual, temporal, parietal, and frontal cortices in monkeys categorizing across three visual domains (shape, motion direction, color). Representations in visual areas MT and V4 were tightly linked to external sensory inputs. In contrast, prefrontal cortex (PFC) largely represented the abstracted behavioral relevance of stimuli (task rule, motion category, color category). Intermediate-level areas—posterior inferotemporal (PIT), lateral intraparietal (LIP), and frontal eye fields (FEF)—exhibited mixed representations. While the distribution of sensory information across areas aligned well with classical functional divisions—MT carried stronger motion information, V4 and PIT carried stronger color and shape information—categorical abstraction did not, suggesting these areas may participate in different networks for stimulus-driven and cognitive functions. Paralleling these representational differences, the dimensionality of neural population activity decreased progressively from sensory to intermediate to frontal cortex. This shows how raw sensory representations are transformed into behaviorally relevant abstractions and suggests that the dimensionality of neural activity in higher cortical regions may be specific to their current task.

## Significance statement

The earliest stages of processing in cerebral cortex reflect a relatively faithful copy of sensory inputs, but intelligent behavior requires abstracting out behaviorally relevant concepts and categories. We examined how this transformation progresses through multiple levels of the cortical hierarchy by comparing neural representations in six cortical areas in monkeys categorizing across three visual domains. We found that categorical abstraction occurs in a gradual fashion across the cortical hierarchy and reaches its apex in prefrontal cortex, that categorical neural coding does not respect classical models of large-scale cortical organization, and that the dimensionality of neural population activity is reduced in parallel with these representational changes. Our results shed light on how raw sensory inputs are transformed into behaviorally relevant abstractions.

## Introduction

Neural representations at the earliest stages of cortical processing reflect a relatively faithful copy of sensory inputs, but intelligent behavior requires abstracting the behaviorally relevant elements from sensory inputs. The sensory continuum often needs to be parsed into categories, such as dividing continuous color variations of fruits into “ripe” and “unripe”. Arbitrary sensory stimuli can also be functionally associated to acquire the same meaning (e.g. the diverse stimuli grouped into the category “food”). Impaired or atypical categorization is a hallmark of disorders such as autism (1) and schizophrenia (2). Understanding its neural basis could provide pathways to early diagnosis and treatment.

Abstract categorical representations can be found in areas at or near the top of the cortical hierarchy, such as lateral prefrontal cortex (3–7), posterior parietal cortex (7–9), and the medial temporal lobe (10). Less well understood are the processing steps that transform bottom-up sensory inputs into these task-related, and thus top-down, representations. We therefore recorded from multiple regions along the cortical hierarchy in macaque monkeys performing a multidimensional categorization task.

It required binary categorization of stimuli continuously varying along two distinct sensory domains—motion direction and color—and arbitrary grouping of a set of shape cues that signaled which feature (motion or color) should be categorized on each trial. We recorded simultaneously from six cortical areas (Fig. 1D) from both the dorsal and ventral visual processing streams, including frontal (prefrontal cortex, PFC; frontal eye fields, FEF), parietal (lateral intraparietal area, LIP), temporal (posterior inferior temporal cortex, PIT) and visual (areas V4 and MT) cortices. Our results suggest that categorization occurs in a gradual fashion across the cortical hierarchy, reaching its apex in prefrontal cortex; that categorical coding does not always correlate with classical functional divisions; and that the dimensionality of cortical activity decreases in parallel with the reduction of continuous sensory stimuli to categorical groupings.

## Results

Our main interest was to track the transformation of visual inputs from more sensory (bottom-up) representations to task-related (top-down) representations. On each trial of our multidimensional categorization task (Fig. 1A), a visual shape cue instructed the monkey whether to categorize a subsequently presented colored, moving random-dot stimulus based on its color (“greenish” vs. “reddish”) or direction of motion (upward vs. downward). Therefore, it probed three different types of sensory inputs: shape, motion, and color. Two of the shapes (arbitrarily chosen) cued motion categorization of the forthcoming dots, while the other two cued color categorization (Fig. 1B,C). Thus, four different shapes were arbitrarily grouped into pairs by virtue of them cueing the same task rule, and continuously varying colors and directions of motion were arbitrarily divided by a sharp boundary (Fig. 1C).

### Shape information

We first examined representations of the four shape cues instructing the task to be performed on each trial. Since two cues mapped to each task rule, we could dissociate cue-related

(sensory) and task-related neural activity. Thus, if information about the cues was purely sensory, neurons should be arbitrarily selective for them in a manner unrelated to their task-related groupings. Alternatively, if neural selectivity reflected their task-relevance, it should group the cues together by task.

To quantify neural information related to the cues, we computed the percent of data variance explained by cue identity (PEV) in each neuron's spike rates, at each time point, within a stimulus-based regression model that allowed distinct rates for all four cues (i.e., measuring bottom-up sensory selectivity). All sampled cortical areas—MT, V4, PIT, LIP, FEF, and PFC—conveyed significant sensory information about the task cues (Fig. 2A,B; $p < 0.01$, one-sample bootstrap test). The strongest cue (shape) information was in areas PIT and V4 (Fig. 2C, $p < 0.01$ for all comparisons with other areas, two-sample bootstrap test).

To measure task-related (top-down) information about the task rule instructed by the cues, we fitted categorical regression models that constrained spike rates to be equal for both cues instructing the same task rule but allowed them to vary between task rules (Fig. 2D, colored lines). The stimulus-based model used above provides an upper bound for the PEV in this categorical model (Fig. 2D, upper gray curves in each plot) that can only be attained by neural activity that differentiates between categories (i.e., tasks), but not at all between conditions within the same category (i.e., different cues instructing the same task). By contrast, neural activity that differentiates indiscriminately between visual cues—that is, reflects no task information—would be expected to have PEV equal to 1/3 of the upper bound, as the true task-relevant pairing of cues is one of three possible pairwise groupings of the set of cues (Fig. 2D, lower gray curves; see Methods for details). The degree of task-related information for each area's neural population was summarized by a task category information index (see Methods) that quantified where the actual observed population responses fell between these two bounds (ranging from 1 for pure task information to 0 for pure sensory information; negative values are possible if selectivity is only between cues that signal the same task). As an additional comparison of cue and task effects, we plotted population mean spike rate time series averaged across each neuron's preferred and non-preferred tasks, and preferred and non-preferred cues within each task (Fig. 2G).

Areas MT, V4, PIT all exhibited information that hewed closely to values expected from a pure bottom-up, sensory representation of shape (Fig. 2D). Their task information indices (Fig. 2E) did not differ significantly from zero ($p > 0.01$) and average spike rate differences between task rules were similar to the differences between cues instructing the same task (Fig. 2G). This was true for both V4 and PIT, areas where we found strong sensory shape information, as well as for MT, where there was weaker shape information. By contrast, PFC, FEF, and LIP all conveyed task information well above that predicted by chance from bottom-up sensory signals (Fig. 2D). They had significant task information indices (Fig. 2E; $p \leq 1\times10^{-5}$ for all three areas) and larger rate differences between tasks than between cues signaling the same task (Fig. 2G). Of these areas, PFC exhibited the most categorical task information. It was significantly greater than all other areas (Fig. 2F; $p < 0.01$) except FEF ($p = 0.05$). FEF and LIP fell in between PFC and the group of sensory areas (MT, V4, and PIT). Thus, visual areas MT, V4, and PIT appear to contain a primarily bottom-up sensory representation of the shape cues while areas LIP, FEF,

and particularly PFC represented their top-down meaning. Unlike the case with sensory information, where results were predictable from traditional areal divisions (shape-coding ventral stream areas PIT and V4 showed the strongest information), top-down task coding was observed in traditionally non-shape-coding dorsal stream areas LIP and FEF, but not in V4 or PIT.

Next, we turned to the random-dot stimuli that the animals had to categorize according to their direction of motion or their color. Both motion direction and color varied along a continuum, but the animals had to group them into upward/downward or greenish/reddish. Much as with the task cues (above), we would expect bottom-up sensory signals to reflect the actual direction or color whereas task-related signals should divide them into their relevant categories. As has been previously shown (11, 12), color and motion signals were present with similar magnitude during both the color and motion tasks, and therefore we combined data across tasks. Furthermore, to isolate neural information specifically reflecting color and motion, we partitioned out from the spiking data any effects due to the monkeys' choice (left vs. right saccade) on each trial (see Methods: Data analysis).

## Motion direction information

First, we began with motion. Analogous to the task cues, there were four distinct motion directions, grouped into two categories: "upward" (90° and 30°) and "downward" (-30° and -90°) categories, with rightward motion (0°) serving as the category boundary (Fig. 1C). All areas conveyed significant sensory information about motion direction, as measured by a stimulus-based model accounting for all tested directions ($p < 0.01$). The strongest sensory motion information was found in area MT (Fig. 3A,B). It was significantly greater than all other areas (Fig. 3C; $p < 0.01$), except PIT ($p = 0.05$).

We then fit a categorical model that only captured between-category (top-down) effects. MT motion category information closely approximated the bottom-up stimulus-based prediction. The motion category information index in MT was not significantly different from zero (Fig. 3D,E; $p = 0.45$), and differences in population mean rate within and between categories were similar (Fig. 3G). Thus, MT motion information was largely bottom-up. The same was true of more weakly direction-selective areas V4 ($p \approx 1$), PIT ($p \approx 1$), and FEF ($p = 0.03$). In contrast, PFC and LIP both exhibited category direction information considerably greater than the sensory prediction (Fig. 3D), motion category information indices significantly greater than zero (Fig. 3E; PFC: $p = 0.004$; LIP: $p \leq 1\times10^{-5}$), and very little within-category difference in spike rate (Fig. 3G). As with the task instruction cues, PFC showed the most categorical motion signal overall (Fig. 3F; significantly greater than MT, V4, and PIT, all $p < 0.01$; non-significant for FEF: $p = 0.05$ and LIP: $p = 0.39$). Thus, areas MT, V4, PIT, and FEF appear to convey primarily bottom-up motion direction while areas PFC and LIP conveyed top-down motion categories. Once again, while the sensory results were consistent with traditional areal divisions (MT predictably had the strongest direction information), significant motion category information was observed in one higher-level dorsal stream area (LIP), but not in another one (FEF).

## Color information

Next, we examined neural information about the colors of the random-dot stimuli. As with motion, there were four distinct color hues grouped into two categories, “greenish” (90° and 30° hue angles) and “reddish” (-30° and -90°) with the category boundary at yellow (0°; Fig. 1C). Once again, significant sensory color information could be found in all studied areas (all $p < 0.01$). Area V4 showed the strongest sensory color information (Fig. 4A,B), significantly greater than all other studied areas (Fig. 4C; $p < 0.01$ for all pairwise comparisons). Although PIT showed the second strongest stimulus color information, it was not significantly greater than either MT or LIP (Fig. 4C; $p > 0.01$), possibly due to insufficient sampling of the relatively sparsely distributed “color patches” in and around this area (13).

Cortical representations of color were overall much more categorical than those for motion direction and task cues. All areas except MT showed significant color category information indices (Fig. 4 D,E; $p < 0.01$ for all areas). PIT, FEF, and PFC all had nearly purely categorical color representations. For each of them, categorical information nearly equaled its stimulus-based upper bound (Fig. 4D) and there was little difference in average spike rate between same-category colors (Fig. 4G). Their category indices were significantly greater than those of LIP and V4 (Fig. 4F; $p < 0.01$ for all pairwise contrasts). By comparison, strongly color-selective area V4, as well as weakly color selective areas MT and LIP, were much less categorical in their color selectivity. Thus, areas MT, V4, and LIP have a primarily bottom-up representation of color, while areas PIT, FEF, and PFC have largely categorized them into binary “greenish” and “reddish” categories. We also found that, once again, sensory color information was fairly consistent with traditional areal divisions (V4 predictably had the strongest color information) while color category coding exhibited mixed correlation with them: as expected, PIT was strongly categorical for color, but so was traditional dorsal stream area FEF.

## Changes in dimensionality across cortex

As a complementary assay of cortical coding properties, we examined the dimensionality of neural population activity using a noise-thresholded principal components analysis (PCA) method (14) previously used to measure PFC dimensionality in an object sequence memory task (15). In that context, it was found that PFC neurons contained a high-dimensional representation of task components (15). As detailed in Methods and SI Methods, we concatenated all neurons from each studied area into a “pseudo-population” and extrapolated to larger population sizes via a condition relabeling procedure. For each area and population size, we computed the mean spike rate for each of 64 task conditions (4 task cues × 4 motion directions × 4 colors) within the random-dot stimulus epoch, when all conditions were differentiated. The dimensionality of the space spanned by each resulting set of 64 neural population activity vectors was quantified as the number of principal components (eigenvalues) significantly greater than those estimated to be due solely to noise.

As expected, estimated dimensionality grew with the size of the neural population examined but generally approximated an asymptotic value (Fig. 5A) that can be taken as an estimate of the dimensionality of the underlying (much larger) neural population. Clear differences in the estimated dimensionality were observed across areas, summarized in Fig. 5B. The highest dimensionality was observed in visual areas V4, PIT, and MT. These were followed by

intermediate-level visual areas LIP and FEF. The lowest dimensionality was observed in PFC. The observed high dimensionality of area PIT was likely due in part to the inclusion of two task factors that it carried relatively strong information about: task cues (shape) and color. When the same analysis was performed in a 16-dimensional space consisting only of 4 directions × 4 colors (Fig. 5C,D), PIT dimensionality was greatly reduced and was similar to that of LIP.

Thus, the dimensionality of population activity decreased progressively up the cortical hierarchy in parallel with the gradual shift from sensory to categorical representations. Further, the estimated PFC dimensionality is close to the value that would be expected of a purely categorical representation with binary responses for each task factor (~3 for the 3-factor analysis, Fig. 5B; ~2 for the 2-factor analysis, Fig. 5D). These results suggest that high dimensionality is not an invariant property of PFC activity, but may be specific to current behavioral demands, in this case to reduce high dimensional sensory stimuli to binary categories.

## Discussion

### Abstraction of sensory inputs occurs progressively through the cortical hierarchy

Our results, summarized in Figure 6, demonstrate a gradual progression from bottom-up sensory inputs to abstracted, top-down behaviorally relevant signals as the cortical hierarchy is ascended. Across three visual domains—shape, motion direction, and color—lower-level visual areas MT and V4 conveyed strong information about sensory stimuli within their preferred domains (Fig. 6A) but showed little evidence for any abstraction beyond the raw sensory inputs (Fig. 6B). In contrast, higher-level area PFC, despite containing relatively weak information overall (Fig. 6A), showed strongly abstracted, task-relevant coding across all domains (Fig. 6B). In between, intermediate-level visual areas PIT, LIP, and FEF showed mixed representations with partial top-down coding in some domains, but not others. These results support models of cortical processing where representational transformations happen gradually across multiple cortical processing steps (16, 17), rather than in a discrete, all-or-nothing fashion.

### Graded cortical functional specialization

There is a longstanding debate about the degree to which the functions of different cortical regions are specialized or overlapping. The cortex's broad interconnections and remarkable resistance to localized damage argue for more distributed, overlapping representations (18), while many studies find evidence for seemingly circumscribed functions in at least some cortical regions (19, 20). We find evidence supporting both points of view. Information about task factors was not distributed homogeneously across cortical regions. For each factor, one or two areas clearly conveyed much stronger sensory information (Fig. 6A) than others. These results were largely predictable from the classical functions of visual cortical areas. Area MT was dominant for motion direction, consistent with its well-established role in motion processing (21). V4 was dominant for color, consistent with many reports of its robust color selectivity (22). PIT and V4 both showed strong sensory information about task cue shapes consistent with their well-established role in shape processing (23). Thus, these results support the idea of specialized cortical representations and reconfirm some of the classical functional divisions between ventral

stream and dorsal stream areas using an experimental paradigm where multiple areas from both streams were tested simultaneously with multiple stimulus domains.

On the other hand, significant information about all examined experimental factors was found in all sampled areas, supporting the idea of broadly distributed cortical representations. The fact that color and shape information can be found in dorsal stream areas MT, LIP, and FEF and motion information can be found in ventral stream areas V4 and PIT argues against any absolute functional dichotomies between cortical processing streams, consistent with previous reports (24–26). We believe this body of results supports cortical models with graded functional specialization, where cortical areas have clear innate or learned biases to represent certain attributes but retain coarse, distributed information about non-specialized attributes (27, 28).

While the overall strength of task-related information accorded well with classical divisions, the degree of top-down categorical abstraction painted a somewhat different picture. Dorsal stream area FEF exhibited a strongly categorical representation of color and task rule (derived from cue shape) but a non-categorical, sensory representation of motion direction. LIP was predictably categorical for motion but also showed a moderately categorical representation for task rule (shape) and color. While areas V4 and PIT were somewhat more predictable—they were relatively categorical for color, but not at all for motion direction—they unexpectedly exhibited no categorical coding for the cue-shape-derived task rule.

We quantified these observations by explaining the summary data in Figure 6 with two predictors related to large-scale cortical organization: 1) the anatomically-derived hierarchical level of each area (29), and 2) the expected functional congruence of each combination of task factor and area—positive for those consistent with classical dorsal/ventral processing stream divisions (e.g. MT and motion, V4 and color), and negative for inconsistent combinations (e.g. MT and color, V4 and motion; see Methods for details). We found that both the decrease in sensory information and the increase in categorical coding across cortical areas were well explained by their anatomical hierarchical level ($p \leq 1\times10^{-5}$ for both), with only a marginally significant difference between them ($p = 0.04$). In contrast, only sensory information was also significantly explained by classical processing stream divisions ($p \leq 1\times10^{-5}$), while category index values were not ($p = 0.11$), with a significant difference between them ($p = 0.003$). Thus, while our sensory information results confirm classical areal divisions, the degree of categorical coding is not well explained by them. These results suggest cortical regions may form different functional networks for bottom-up vs. top-down functions, putatively reflecting primarily feedforward and feedback/recurrent circuits, respectively.

## Categorization reached its apex in prefrontal cortex

Many studies have now reported neural correlates of categories in PFC (3, 5–7), and in area LIP and related areas of posterior parietal cortex (6, 7, 9). A recent study suggested that LIP might contain a stronger representation of categories that could drive categorical processing in PFC (6). We found that, for all examined domains, PFC exhibited a degree of categorical abstraction either greater than all other studied areas (task rule and motion) or not significantly different from the other most categorical area (color). For all domains, the prefrontal representation was more categorical than LIP, though this difference was significant only for

task rule and color, not motion direction. On the other hand, despite being less categorical than PFC, LIP did also exhibit a significantly categorical representation for all tested domains, which was not the case for any other studied area besides PFC. We interpret these results to mean that PFC does play an important—perhaps the most important—role in categorization (see also (7)). However, LIP clearly also plays a central role, and categorization likely involves reciprocal interactions between these areas, as well as others (30, 31).

### Cortical dimensionality may be task-specific

At first blush, our population dimensionality results might seem at odds with a recent report showing prefrontal population activity is high dimensional (15). That study found that PFC neurons tend to exhibit "nonlinear mixed selectivity" for specific conjunctions of task variables and consequently PFC population activity had a dimensionality near the theoretical maximum (24 dimensions) for the studied task. However, that study employed a task involving encoding and maintenance in working memory of a sequence of visual objects and responding either via a recollection or recall probe (32). Thus, correct performance required remembering which of 12 different sequences was shown and which of two modes of behavioral output was mandated. By contrast, the task used here emphasized dimensionality *reduction*. First, four visual cues were grouped into either of two task instructions. Next, 16 random-dot stimuli (4 colors × 4 directions) was mapped onto binary color or motion categories, depending on the currently instructed task rule. Finally, the deduced category was translated into a binary response. Thus, this task, unlike the previous one, emphasized reduction of high-dimensional sensory inputs to lower-dimensional abstractions. Our results therefore suggest the possibility that prefrontal dimensionality may flexibly reflect current cognitive demands (33). Inputs are expanded to higher dimensions when decisions depend on multiple variables but reduced to lower dimensionality when categorical abstraction is required. Thus, PFC dimensionality, like other PFC coding properties (34), appears to flexibly adapt to behavioral needs.

## Methods

The current work reanalyzes previously published data from our group (11). Experimental methods are briefly reviewed here, but further details can be found in SI Methods, as well as in our prior publication. All procedures followed the guidelines of the Massachusetts Institute of Technology Committee on Animal Care and the National Institutes of Health.

### Electrophysiological data collection

Neuronal activity was recorded simultaneously from up to 108 electrodes acutely implanted in each of 47 experimental sessions into up to six cortical regions (Fig. 1D): MT (60 neurons total), V4 (121), PIT (36), LIP (516), FEF (612), and PFC (1069), and sorted offline into isolated neurons. To minimize any sampling bias of neural activity, we did not prescreen activity for responsiveness or task selectivity.

### Behavioral paradigm

Two adult rhesus macaques (*Macaca mulatta*) were trained to perform a multidimensional categorization task. On each trial (Fig. 1A), a visual cue instructed the monkey to perform one of two tasks: color categorization ("greenish" vs. "reddish") or motion categorization ("upward" vs.

“downward”) of a subsequently presented colored, moving random-dot stimulus. They responded via a saccade towards a target to the left (greenish/upward) or right (reddish/downward).

Each task could be instructed by either of two distinct cues (Fig. 1B), distinguished by their global shapes, in order to dissociate cue-related (sensory) and task-related activity. Each color category comprised two distinct hues, and each motion category comprised two distinct motion directions (Fig. 1C). Thus, the behavioral paradigm was composed of three orthogonal factors, each of which can be characterized at either of two hierarchical levels of description: the raw sensory stimuli (visual cues, colors, and directions) or the corresponding abstracted meaning of those stimuli (task rule, color category, and motion category).

## Data analysis

**General.** For most analyses, spike trains were converted into smoothed rates with a Hann function (half-width 125 ms). To summarize results, we pooled rates or other derived metrics within empirically-defined epochs of interest for each factor and area (see SI Methods:General). Only correctly performed trials were included in analysis. All hypothesis tests used distribution-free bootstrap methods, with false discovery rate correction for multiple comparisons where appropriate.

**Neural information analysis.** Our primary interest was to characterize the degree to which neurons in each cortical region reflected the raw sensory stimuli or their abstracted meaning (task rule or category). We therefore fit two contrasting regression models to spike rate data—a stimulus-based one and a categorical one—separately for each neuron and time point. The stimulus-based model accounted for distinct rates for each cue, color, and motion direction, and therefore captured all data variance linearly related to individual cues, colors, and motion directions. The categorical model instead only accounted for distinct rates between tasks or categories—assuming identical rates within each task or category—and therefore only captured data variance across tasks, color categories, and motion categories. The information conveyed by neural activity about each model factor was quantified by its bias-corrected percentage of variance explained (PEV).

Because the categorical model is nested within the stimulus-based model, the stimulus-based PEV is an upper bound for the categorical PEV, which can only be attained for neural activity with purely categorical effects. On the other hand, a neural population only reflecting stimulus attributes, without respect to their categorical meaning, has an expected categorical PEV of one third of the stimulus-based PEV (see SI Methods: Neural information analysis for details). To determine the degree to which each area’s representation was stimulus-driven or categorical, we computed a “category information index” reflecting where its population mean categorical PEV lies between these two bounds. Separate indexes were computed to contrast task cues vs. tasks, motion directions vs. motion categories, and colors vs. color categories. Each index ranges from 1 for a purely categorical representation to 0 for a stimulus-driven representation (values less than 0 are possible for populations that specifically encode within-category differences). Because overall neural information is normalized out of this index, it reflects a pure measure of the categorical quality of a neural representation, similar to previous measures of

category selectivity (3), but taking the reliability of neural coding into account, as it's based on neural information (PEV), rather than raw spike rates.

**Preferred condition analysis.** To visualize within- and between-category effects, we also plotted population average spike rates for preferred vs. non-preferred tasks/categories and preferred vs. non-preferred constituent stimuli, selecting preferred conditions via two-fold cross-validation to avoid circularity.

**Population dimensionality analysis.** To measure the dimensionality of population activity, we estimated the number of principal components required to describe the space spanned by condition-mean neural population activity vectors (14, 15). We characterized neural activity within the full space of 64 task conditions (4 task cues × 4 motion directions × 4 colors), during the random-dot stimulus time epoch, when all conditions were differentiated. Epoch spike rates were computed for each trial and neuron, averaged across all trials of each condition, and concatenated across all neurons and sessions into a set of neural "pseudo-population" vectors for each studied area. Results were extrapolated to larger populations via a condition relabeling procedure (15). Dimensionality was computed as the number of principal components (eigenvalues) of each resulting [conditions × neurons] matrix significantly greater ($p < 0.05$) than the estimated distribution of principal components due to noise. Noise was estimated by computing the difference between randomly sampled pairs of trials within each condition, weighted by its expected contribution to the estimated condition mean, and submitting the resulting matrix of 64 noise vectors to PCA (14).

**Cortical organization analysis.** To relate our results to classical models of cortical organization, we fit each of the population summary matrices of Figure 6 with a two-predictor regression model: (1) the Felleman & Van Essen (29) hierarchical level of each area, and (2) the expected functional congruence of each combination of task factor and area based on classical functional divisions—positive for consistent combinations (e.g. MT and motion), and negative for inconsistent ones (e.g. MT and color).

## Acknowledgements

We thank Andre Bastos, Mikael Lundqvist, Jefferson Roy, and Andreas Wutz for helpful discussions, and Mattia Rigotti for providing code and advice on the dimensionality analysis. This work was supported by grant NIMH 5R37MH087027 (E.K.M), grant ERC StG335880 (M.S.), and the Centre for Integrative Neuroscience (DFG, EXC 307) (M.S.).

## References

1. Klinger LG, Dawson G (1995) A fresh look at categorization abilities in persons with autism. *Learning and Cognition in Autism* (Springer), pp 119–136.

2. Kéri S, et al. (1999) Category learning and perceptual categorization in schizophrenia. *Schizophr Bull* 25(3):593.

3. Freedman DJ (2001) Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex. *Science* 291(5502):312–316.

4. Wallis JD, Anderson KC, Miller EK (2001) Single neurons in prefrontal cortex encode abstract rules. *Nature* 411:953–956.

5. Roy JE, Riesenhuber M, Poggio T, Miller EK (2010) Prefrontal Cortex Activity during Flexible Categorization. *J Neurosci* 30(25):8519–8528.

6. Swaminathan SK, Freedman DJ (2012) Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. *Nat Neurosci* 15(2):315–320.

7. Goodwin SJ, Blackman RK, Sakellaridi S, Chafee MV (2012) Executive Control Over Cognition: Stronger and Earlier Rule-Based Modulation of Spatial Category Signals in Prefrontal Cortex Relative to Parietal Cortex. *J Neurosci* 32(10):3499–3515.

8. Stoet G, Snyder LH (2004) Single neurons in posterior parietal cortex of monkeys encode cognitive set. *Neuron* 42(6):1003–1012.

9. Freedman DJ, Assad JA (2006) Experience-dependent representation of visual categories in parietal cortex. *Nature* 443(7107):85–88.

10. Kreiman G, Koch C, Fried I (2000) Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat Neurosci* 3(9):946.

11. Siegel M, Buschman TJ, Miller EK (2015) Cortical information flow during flexible sensorimotor decisions. *Science* 348(6241):1352–1355.

12. Mante V, Sussillo D, Shenoy KV, Newsome WT (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503(7474):78–84.

13. Conway BR, Moeller S, Tsao DY (2007) Specialized Color Modules in Macaque Extrastriate Cortex. *Neuron* 56(3):560–573.

14. Machens CK, Romo R, Brody CD (2010) Functional, But Not Anatomical, Separation of “What” and “When” in Prefrontal Cortex. *J Neurosci* 30(1):350–360.

15. Rigotti M, et al. (2013) The importance of mixed selectivity in complex cognitive tasks. *Nature* 497(7451):585–590.

16. Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. *Nat Neurosci* 2(11).

17. DiCarlo JJ, Zoccolan D, Rust NC (2012) How Does the Brain Solve Visual Object Recognition? *Neuron* 73(3):415–434.

18. Lashley KS (1963) *Brain mechanisms and intelligence: A quantitative study of injuries to the brain* (Dover Publications, Inc., New York).

19. Fodor JA (1983) *The modularity of mind: An essay on faculty psychology* (MIT press).

20. Kanwisher N (2010) Functional specificity in the human brain: A window into the functional architecture of the mind. *Proc Natl Acad Sci* 107(25):11163–11170.

21. Born RT, Bradley DC (2005) Structure and function of visual area MT. *Annu Rev Neurosci* 28:157–189.

22. Conway BR (2009) Color Vision, Cones, and Color-Coding in the Cortex. *The Neuroscientist* 15(3):274–290.

23. Connor CE, Brincat SL, Pasupathy A (2007) Transformation of shape information in the ventral pathway. *Curr Opin Neurobiol* 17(2):140–147.

24. Desimone R, Schein SJ (1987) Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. *J Neurophysiol* 57(3):835–868.

25. Sereno AB, Maunsell JHR (1998) Shape selectivity in primate lateral intraparietal cortex. *Nature* 395(6701):500–503.

26. Peng X, Sereno ME, Silva AK, Lehky SR, Sereno AB (2008) Shape Selectivity in Primate Frontal Eye Field. *J Neurophysiol* 100(2):796–814.

27. Schiller PH (1996) On the specificity of neurons and visual areas. *Behav Brain Res* 76:21–35.

28. Singer W (2013) Cortical dynamics revisited. *Trends Cogn Sci* 17(12):616–626.

29. Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1(1):1–47.

30. Crowe DA, et al. (2013) Prefrontal neurons transmit signals to parietal neurons that reflect executive control of cognition. *Nat Neurosci* 16(10):1484–1491.

31. Antzoulatos EG, Miller EK (2016) Synchronous beta rhythms of frontoparietal networks support only behaviorally relevant representations. *eLife* 5:e17822.

32. Warden MR, Miller EK (2010) Task-Dependent Changes in Short-Term Memory in the Prefrontal Cortex. *J Neurosci* 30(47):15801–15810.

33. Fusi S, Miller EK, Rigotti M (2016) Why neurons mix: high dimensionality for higher cognition. *Curr Opin Neurobiol* 37:66–74.

34. Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24(1):167–202.

35. Asaad WF, Santhanam N, McClellan S, Freedman DJ (2013) High-performance execution of psychophysical tasks with complex visual stimuli in MATLAB. *J Neurophysiol* 109(1):249–260.

36. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol*:289–300.

37. Olejnik S, Algina J (2003) Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychol Methods* 8(4):434–447.

38. Rao SC, Rainer G, Miller EK (1997) Integration of what and where in prefrontal cortex. *Science* 276(5313):821–824.

## Figures & Figure legends

### Figure 1 – Experimental design

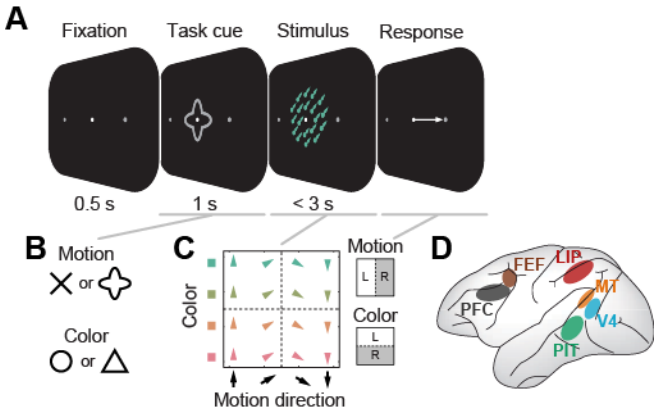
(A) Trial sequence for multidimensional visual categorization task. On each trial, monkeys categorized either the motion direction or color of a random-dot stimulus. This stimulus was immediately preceded by a symbolic cue that instructed which feature (motion or color) to categorize for that trial. The monkey responded with a leftward or rightward saccade at any time during the 3 s stimulus.

(B) Either of two different cue shapes was used to instruct each task, in order to dissociate cue- and task-related activity.

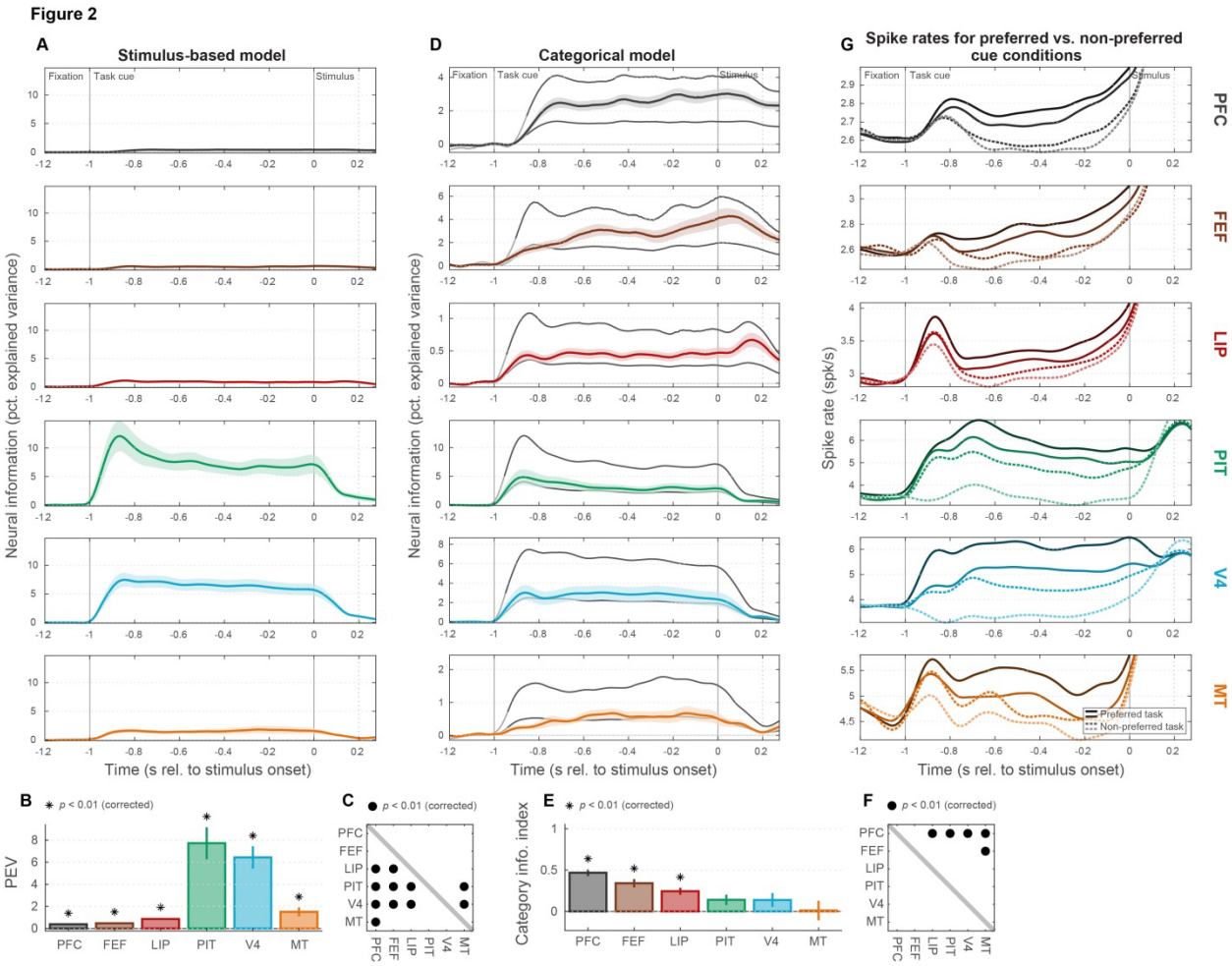
(C) Stimuli systematically sampled motion direction (upward to downward) and color (green to red). Each color category comprised two distinct colors and each motion category comprised two distinct motion directions (additional ambiguous stimuli on the category boundaries were included in the paradigm, but not analyzed here). Dashed lines indicate category boundaries. For each task, the two categories had a fixed mapping to a leftward (L) vs. rightward (R) saccadic response.

(D) Illustration of sampled brain regions: lateral prefrontal cortex (PFC), frontal eye fields (FEF), lateral intraparietal (LIP), posterior inferotemporal (PIT), V4, and MT.

**Figure 1**



## Figure 2 – Cue/task representation




(A) Mean (± SEM) neural information (percent of explained variance; PEV) about task cues in each studied area as a function of time within the trial (referenced to the onset of the random-dot stimulus). Shaded areas indicate SEM.

(B) Summary (across-time mean ± SEM) of cue stimulus information for each area. All areas contain significant information ($p < 0.01$, false discovery rate corrected for multiple comparisons; asterisks).

(C) Cross-area comparison matrix indicating which regions (rows) had significantly greater cue information than others (columns). Dots indicate area pairs that attained significance ($p < 0.01$, corrected for multiple comparisons). PIT and V4 contain significantly greater cue information than all other areas.

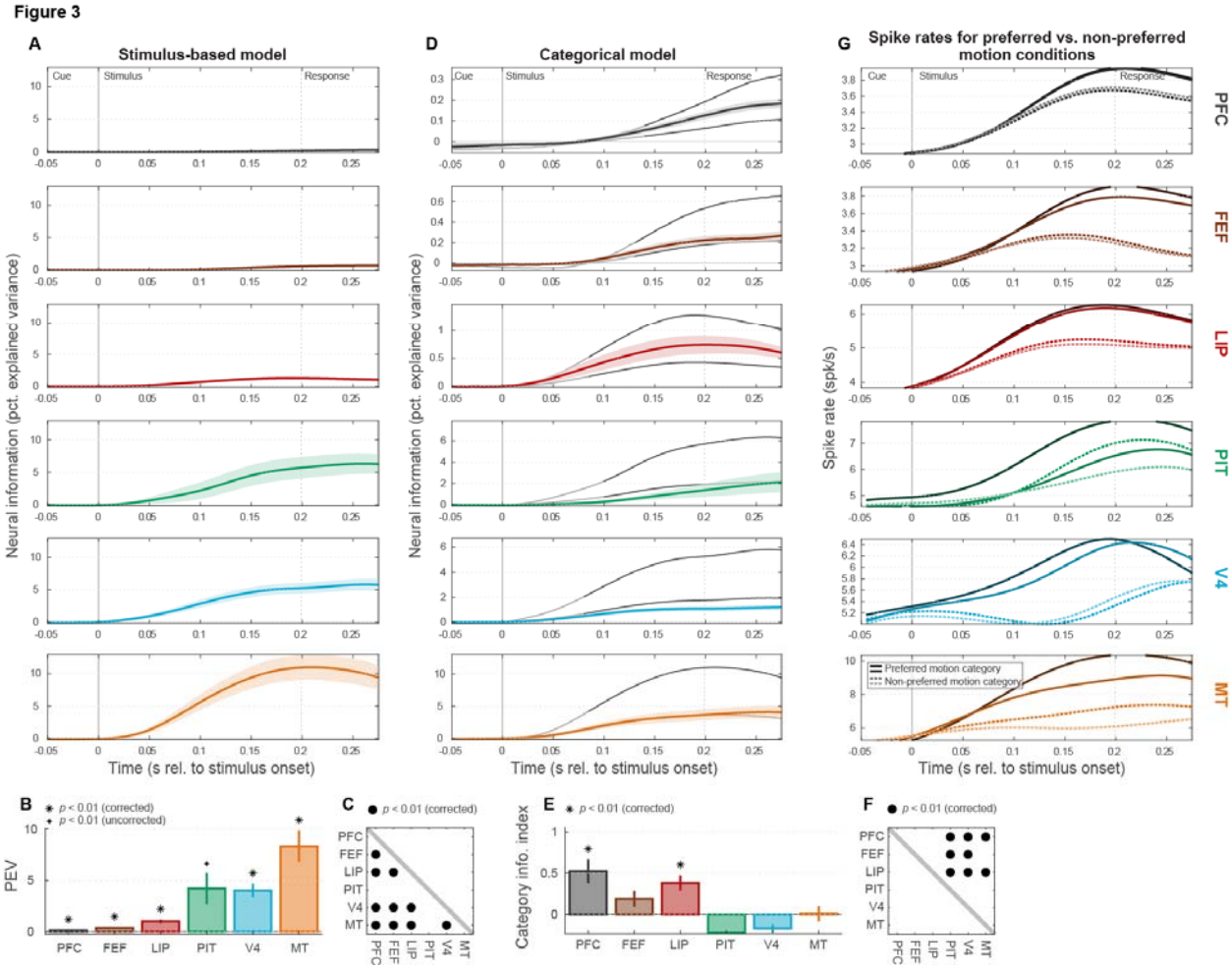
(D) Mean (± SEM) neural information (PEV) about the task performed on each trial, as a function of time. Note differences in *y*-axis ranges from panel A. Gray curves indicate expected values of this statistic corresponding to a purely task-driven representation (upper line), and to a purely sensory representation of task cues (lower line). The transitions from light to dark gray in these curves indicate the estimated onset latency of overall cue information (see Methods), which was used as the start of the summary epoch for each area.

(E) Task category information index (± SEM) for each area, reflecting where its average task information falls between the expected values for pure sensory (0) and task (1) representations. Only PFC, FEF, and LIP are significantly different from zero ($p < 0.01$; asterisks).

(F) Indicates which regions (rows) had significantly greater task information indices than others (columns). PFC was significantly greater ($p < 0.01$; dots) than all others, except FEF.

(G) Mean spike rates for preferred (solid lines) and non-preferred tasks (dotted lines), and for preferred (darker) and non-preferred (lighter) task cues within each task. Note that rates for some areas exceed the plotting range outside the time epoch of interest here (the cue period). Consistent with the information analyses, PFC rate differences for cues instructing the same task were small compared to differences between tasks, whereas for visual areas MT, V4, and PIT these differences were of comparable size.

## Figure 3 – Motion direction representation




(A) Mean (± SEM) neural information about random-dot stimulus motion directions in each studied area as a function of time (note difference in time axis from Fig. 2).

(B) Summary (across-time mean ± SEM) of motion direction information for each area. All areas contain significant information ($p < 0.01$, asterisks; PIT only significant when not correcting for multiple comparisons, small asterisk). MT carried the strongest motion information.

(C) Indicates which regions (rows) had significantly greater motion direction information than others (columns).

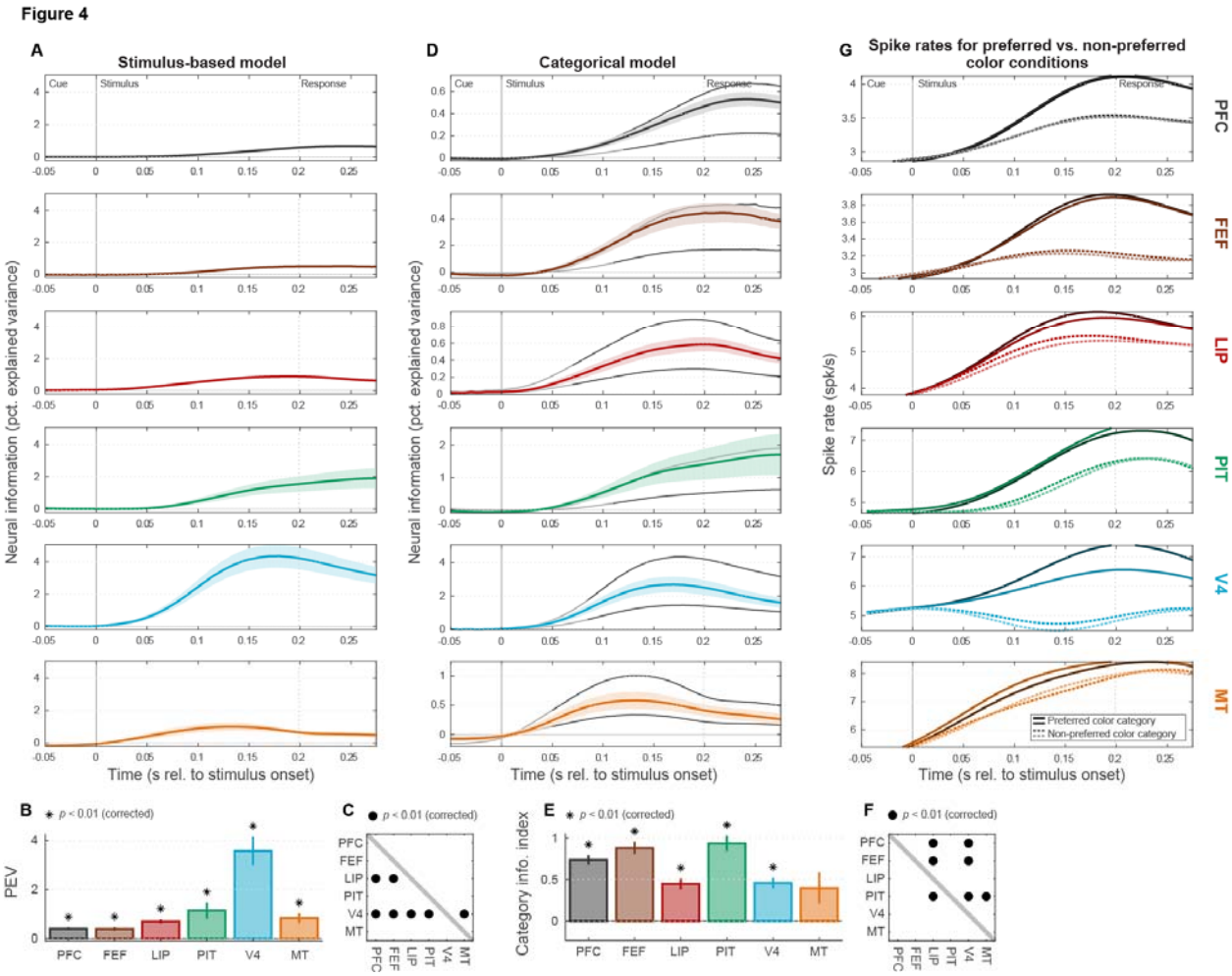
(D) Mean (± SEM) neural information (PEV) about motion categories. Gray curves indicate expected values of this statistic corresponding to a purely categorical representation (upper line), and to a purely sensory representation of motion direction (lower line).

(E) Motion category information index (± SEM) for each area, reflecting where its average motion category information falls between the expected values for pure sensory (0) and categorical (1) representations. Only PFC and LIP are significantly different from zero ($p < 0.01$, asterisks).

(F) Indicates which regions (rows) had significantly greater motion category information indices than others (columns; $p < 0.01$).

(G) Mean spike rates for preferred (solid lines) and non-preferred motion categories (dotted lines), and for preferred (darker) and non-preferred (lighter) directions within each motion category. Within-category rate differences in PFC and LIP were small compared to between-category differences, whereas for visual areas MT, V4, and PIT these differences were comparable.

## Figure 4 – Color representation




(A) Mean (± SEM) neural information about random-dot stimulus colors in each studied area.

(B) Summary (across-time mean ± SEM) of color information for each area. All areas contain significant information ($p < 0.01$, asterisks). V4 carried the strongest color information.

(C) Indicates which regions (rows) had significantly greater color information than others (columns; $p < 0.01$).

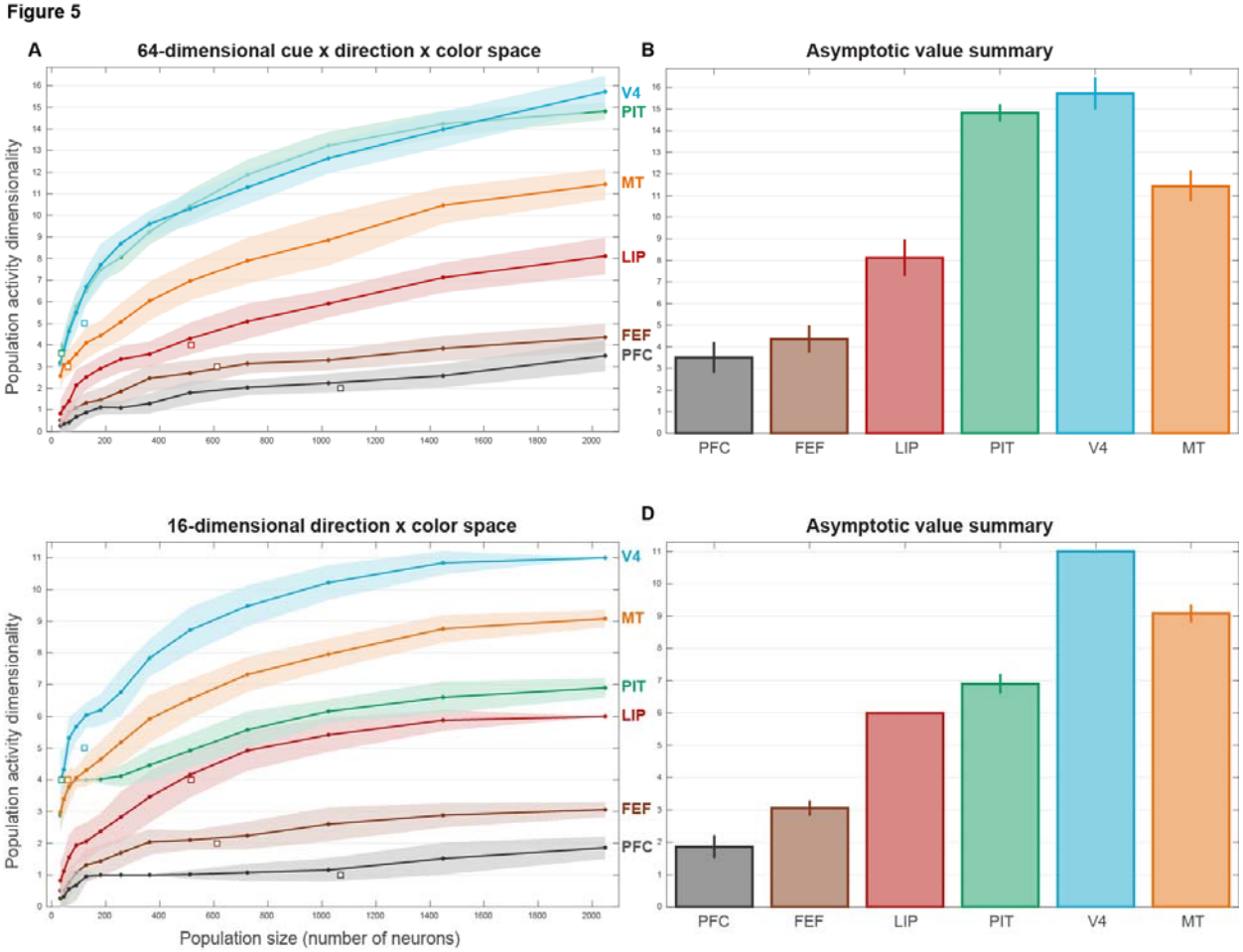
(D) Mean (± SEM) neural information (PEV) about color categories. Gray curves indicate expected values of this statistic corresponding to a purely categorical representation (upper line), and to a purely sensory representation of color (lower line).

(E) Color category information index (± SEM) for each area.

(F) Indicates which regions (rows) had significantly greater color category information indices than others (columns; $p < 0.01$).

(G) Mean spike rates for preferred (solid lines) and non-preferred color categories (dotted lines), and for preferred (darker) and non-preferred (lighter) colors within each color category.

## Figure 5 – Population activity dimensionality



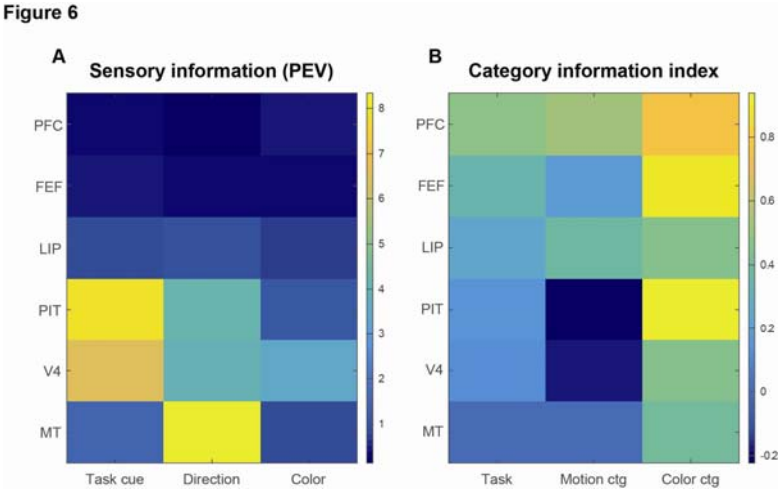

(A) Dimensionality (mean ± SEM) of neural population activity as a function of extrapolated population size, for each studied area. Dimensionality was estimated by noise-thresholded PCA within a 64-dimensional task cue × motion direction × color space. Values for the actual recorded neural populations (white squares ± SEM) were largely consistent those from the extrapolated populations.

(B) Summary of asymptotic dimensionality values (± SEM) in 64-d space.

(C) Dimensionality (mean ± SEM) of population activity as a function of population size, for each studied area. Dimensionality was estimated within a reduced 16-dimensional motion direction × color space

(D) Summary of asymptotic dimensionality values (± SEM) in 16-d space. V4 and MT have the highest dimensionality, followed by PIT and LIP, and then by FEF and PFC.

## Figure 6 – Summary of results




(A) Mean stimulus-based neural information (percent explained variance) in each studied area about task cues, motions directions, and colors.

(B) Task/category information indices for each studied area for task, motion categories, and color categories.

# Supplementary Information

## SI Methods

### Electrophysiological data collection

Neuronal activity was recorded from parylene-coated tungsten electrodes, referenced to animal ground, amplified by a high-impedance headstage, filtered to extract spiking activity, digitized, and streamed to disk by an integrated electrophysiological system (Recorder or Omniplex, Plexon). The filtered signal was threshold-triggered to separate neuronal spikes from background noise, and individual spike waveforms were manually sorted off-line into isolated neurons (Offline Sorter 3, Plexon). Unsorted multi-unit data, as used in our previous publication (11), produced very similar results. Neurons were included in analyses only for the duration of time in which they were well isolated from background noise and other neurons.

### Behavioral paradigm

Two adult rhesus macaques (*Macaca mulatta*), one male and one female, were trained to perform a cued multidimensional categorization task. All stimuli were displayed on a CRT monitor at 100 Hz vertical refresh rate. An infrared-based eye-tracking system (Eyelink II, SR Research) continuously monitored eye position at 240 Hz. Behavioral control was handled by the MonkeyLogic program (www.monkeylogic.net) (35).

Each trial (Fig. 1A) was initiated when the animal fixated on a central dot (± 1.2° visual angle). Following a 500 ms fixation period, a centrally-presented visual cue (1000 ms) instructed the monkey to perform one of two tasks: color categorization ("greenish" vs. "reddish") or motion categorization ("upward" vs. "downward") of a subsequent centrally-presented colored, moving random-dot stimulus. Within each stimulus, all dots had the same color and moved in the same direction (100% coherence). Monkeys reported the stimulus category via a saccade towards a target to the left or right. The stimulus-response mapping for each task was fixed (color task: greenish→left, reddish→right; motion task: upward→left, downward→right). Monkeys were free to respond at any point (up to 3 s) after the random-dot stimulus onset.

The cues were four different gray shapes; two of them instructed the color task, while the other two instructed the motion task (Fig. 1B). Using two distinct cues for each task allowed dissociation of neural activity related to the visual cue shapes and the task they instructed. Each motion category consisted of two directions (Fig. 1C; upward: 90°, 30°; downward: -30°, -90°), and each color category consisted of two hues (greenish: 90°, 30°; reddish: -30°, -90° hue angle). Additional tested values at the category boundaries were not included in the present analyses, as we focus here on unambiguously classifiable stimuli.

The monkeys performed both categorization tasks with high accuracy: 94% correct on average for motion, 89% for color. Across all trials and sessions in the current dataset, the median reaction time was 274 ms, with only 1% of all responses occurring earlier than 200 ms.

## Data analysis

**General.** To summarize results, we pooled rates or other derived metrics within empirically-defined epochs of interest for each factor and area. The start of each epoch was set by the onset latency of stimulus information for a given factor and area (see Neural information analysis below), as estimated by the time point where population mean information first attained 25% of its maximal value across all time points. To capture the full temporal extent of relevant neural activity, without introducing confounding factors, the ends of cue/task factor epochs for all areas were set at the onset of the random-dot stimulus, and the ends of motion- and color-related epochs were set at the earliest 1 percentile of the distribution of all behavioral reaction times (200 ms). All analyses were performed using custom Matlab code.

**Hypothesis testing.** All hypothesis tests used non-parametric bootstrap methods that do not rely on specific assumptions about the distributions of data values. One-sample tests (Fig. 2B,E; 3B,E; 4B,E) resampled the statistic of interest with replacement 100,000 times from the neuronal population. The resulting distribution, offset by the actual observed statistic value, is an empirical estimate of the distribution of the statistic under the null hypothesis, and the one-tailed significance level was computed as the proportion of resampled values greater than the actual observed value. Two-sample tests (Fig. 2C,F; 3 C,F; 4 C,F) were computed similarly, but with separate bootstrap resampling from each compared population, evaluating the difference in the statistic of interest between populations, and computing the two-tailed significance level as the proportion of absolute resampled values exceeding the absolute observed value. All tests were evaluated at the $p < 0.01$ level, using the Benjamini-Hochberg false discovery rate procedure (36) to account for multiple comparisons where appropriate. All presented standard errors were computed as the standard deviation across bootstrap samples.

**Neural information analysis.** To characterize the degree to which neurons in each cortical region reflected the raw sensory stimuli or their abstracted meaning, we fitted two contrasting linear multiple regression models to spike rate data—a stimulus-based one and a categorical one—separately for each neuron and time point. The stimulus-based model allowed for distinct rates for each cue, color, and motion direction:

$$\mathbf{rate}_{j,t} = \mathbf{X}_{cue}\boldsymbol{\beta}_{cue,j,t} + \mathbf{X}_{direction}\boldsymbol{\beta}_{direction,j,t} + \mathbf{X}_{color}\boldsymbol{\beta}_{color,j,t} + \mathbf{x}_{choice}\beta_{choice,j,t} + \beta_{0,j,t} + \varepsilon_{j,t}$$

where $\mathbf{rate}_{j,t}$ is a [trials × 1] vector of per-trial spike rates for a given neuron $j$ and time point $t$; the first three $\mathbf{X}$'s are [trials × 3] matrices of binary indicator vectors labelling which cue, color, or direction was present on each trial (the fourth value is implicitly coded by zeros on all three indicators), which are the same for all neurons and time points in a given session; the $\boldsymbol{\beta}$'s are the corresponding [3 × 1] vectors of coefficients, fitted independently for each neuron and time point; $\beta_0$ is the intercept, and $\varepsilon$ is the residual error. An additional parameter, $\beta_{choice}$, accounted for any effects of the monkey's eventual behavioral choice—reported via a left or right saccade—on each trial. Choice effects were analyzed in detail in our previous publication on this dataset (11), and were therefore not examined further in this work. However, including this term in the model allowed us to partition out any data variance related to behavioral choice and subsequent motor preparation processes, isolating variance specifically related to the variables of interest here. The full model contained 11 fitted parameters. This model captured all data

variance linearly related to individual task cues, colors, and motion directions, as well as the behavioral choice.

The categorical model instead assumed identical rates within each task or category, only modeling rate effects across tasks, color categories, and motion categories:

$$\mathbf{rate}_{j,t} = \mathbf{x}_{task}\beta_{task,j,t} + \mathbf{x}_{motionCtg}\beta_{motionCtg,j,t} + \mathbf{x}_{colorCtg}\beta_{colorCtg,j,t} + \mathbf{x}_{choice}\beta_{choice,j,t} + \beta_{0,j,t} + \varepsilon_{j,t}$$

where the overall model structure is analogous to the stimulus-based model, except here the first three $\mathbf{x}$'s are [trials × 1] vectors of binary indicators labelling which task, color category, or direction category was present on each trial, and the $\beta$'s are the corresponding scalar coefficients (five total fitted parameters). Each model was fit via ordinary least squares. To quantify the information conveyed in neural activity about each task factor of interest, we computed the percentage of variance explained (PEV) by the full set of coefficients corresponding to each factor, using the bias-corrected $\omega^2$ formulation (37).

The categorical model can be expressed as a nested subset of the stimulus-based model via a change of basis of each $\mathbf{X}$ in equation (1) to one where the first basis vector is a binary contrast between categories (or tasks), and the remaining basis vectors are binary contrasts representing the other two possible, non-task-relevant grouping of stimulus conditions into "pseudo-categories" (11). Therefore, the stimulus-based PEV provides a theoretical upper bound for the categorical PEV, which can only be attained for neural activity with purely categorical effects, and no activity differences within categories. On the other hand, a neural population only reflecting stimulus attributes, without respect to their categorical meaning, would be expected on average to exhibit equivalent effects for contrasts between true task-relevant categories and between the artificial "pseudo-categories" formed by non-task-relevant stimulus groupings. Therefore, due to the three possible (true and "pseudo") category groupings in this task design, a purely stimulus-related population would be expected to have, on average, $PEV_{categorical} = \frac{1}{3} * PEV_{stimulus}$ (note this does not represent a hard lower bound, as a population that specifically represents within-category differences can attain smaller values). These theoretical results were confirmed by extensive simulations.

To determine the degree to which population representations for each task factor, in each cortical area, are stimulus-driven or categorical, we compute a simple metric—which we refer to as a "category information index"—reflecting where its actual population mean $PEV_{categorical}$ lies between these two bounds:

$$categoryIndex = \frac{\overline{PEV}_{categorical} - lowerBound}{upperBound - lowerBound}$$

where $upperBound = \overline{PEV}_{stimulus}$, $lowerBound = \overline{PEV}_{stimulus}/3$, and the $\overline{PEV}$'s are the mean PEV values across all recorded neurons for each area and over all time points within a factor-specific epoch of interest (see above). Separate indexes were computed to contrast task cues vs. tasks, motion directions vs. motion categories, and colors vs. color categories. This index

has a value of 1 for a purely categorical representation, and a value of 0 for a stimulus-driven representation (values less than 0 are possible for populations that specifically encode within-category differences).

**Preferred condition analysis.** As an additional visualization of within- and between-category effects, we plotted population average spike rates for preferred and non-preferred categories and stimuli. For each neuron and task factor, the preferred and non-preferred category (task, motion category, or color category) and the preferred and non-preferred constituent stimuli within each category (task cue, motion direction, or color) were determined from the average spike rate within the relevant time epoch for each area and factor. Spike rate time series were averaged across the population of neurons in each area separately within each of the resulting four sorted conditions (preferred stimulus in preferred category, non-preferred stimulus in preferred category, preferred stimulus in non-preferred category, non-preferred stimulus in non-preferred category). To avoid circularity in this analysis, we used a two-fold cross-validation method, where preferred conditions were estimated from odd-numbered trials and the resulting sorting was used for population averaging of rate series across even-numbered trials. The roles of these trial subsets were then reversed, and the final presented results are the average of these two cross-validation folds.

**Population dimensionality analysis.** To measure the dimensionality of population activity, we adapted a method developed by Machens and colleagues (14), and extended by Rigotti and colleagues (15). This method estimates the dimensionality of the space spanned by population activity vectors by determining the number of principal components required to describe them. Spike rates were computed within each of the time epochs enumerated above, and the resulting data for each area was pooled across all electrodes and recording sessions into “pseudo-populations”. To extrapolate results to larger neural populations, we generated artificial neurons via specific relabeling of conditions, under the assumption that the full underlying population likely includes neurons with similar activity distributions, but slight differences in tuning for task cues, motion directions, and colors (15). Labels were swapped between the two visual cues associated with each task, between the two colors within each color category, and/or between the two directions within each motion category. Note that these label reassignments both maintain the semantic logic of the paradigm and would not alter the information carried by either stimulus-specific or categorical neurons, as estimated above. Artificial neural populations were generated by all permutations of performing each of these relabeling operations or not, resulting in a 64-fold multiplication of the actual populations ($2^2$ colors × $2^2$ directions × $2^2$ cues). Extrapolated populations of different sizes were generated by randomly subsampling from this full population of real and artificial neurons, separately for each cortical area. Dimensionality values obtained from analyzing only the actual recorded populations (Fig. 5A,C, squares) generally fell within the distribution of those obtained from the extrapolated populations, suggesting that this procedure did not substantially alter the results.

For this analysis, we characterized neural activity within the full space of 64 task conditions (4 task cues × 4 motion directions × 4 colors) and within the random-dot stimulus epoch, when all 64 conditions were differentiated (an additional analysis was performed within the reduced space of 16 motion directions × colors). To avoid PCA being dominated by a few highly active

neurons, trial spike rates for each neuron were preprocessed by z-scoring them by the mean and SD across all trials, with a regularization constant of 0.1 added to each SD (15). For each area and extrapolated population size, the vector of mean spike rates across all neurons for each condition was computed, and the dimensionality of the space spanned by each resulting set of 64 neural population activity vectors was estimated as the number of their principal components (eigenvalues) significantly greater than the estimated distribution of principal components due to finite sampling noise. Noise was estimated by computing the difference between randomly chosen trials within each condition, weighted by its expected contribution to the estimated condition mean (14), and submitting the resulting set of 64 noise vectors to PCA. This procedure was repeated 1000 times, randomly sampling within-condition trial pairs each iteration, to generate a distribution of noise-derived eigenvalues. Dimensionality was estimated as the number of data eigenvalues exceeding 95% of the distribution of first (largest) eigenvalues computed from the sampled noise (15). This entire procedure was repeated 50 times, selecting with replacement a different random subsample of area neurons in each iteration, to estimate bootstrap standard errors of the resulting dimensionality statistic. Qualitatively similar results were obtained using the method of Rigotti et al. 2013 based on the number of binary classifications that can be successfully decoded from population activity vectors.

**Cortical organization analysis.** To quantitatively relate our results to classical models of large-scale cortical organization, we separately fit the 18 population summary values (6 areas × 3 task factors) of sensory information (Fig. 6A) and category information index (Fig. 6B) with a simple two-predictor linear regression model:

$$\mathbf{summaryValue} = \mathbf{x}_{hierarchicalLevel}\beta_{hierarchicalLevel} + \mathbf{x}_{streamCongruence}\beta_{streamCongruence} + \beta_0 + \varepsilon$$

The first predictor, $\mathbf{x}_{hierarchicalLevel}$, was the Felleman & Van Essen hierarchical level of each area (MT,V4: 5; PIT,LIP: 7; FEF: 8; PFC: 10), estimated from the relative distribution of feedforward-type vs. feedback-type anatomical connections between areas (29). The second predictor, $\mathbf{x}_{streamCongruence}$, reflected the expected functional congruence of each combination of task factor and area based on the classical divisions of visual areas and functions of the ventral and dorsal processing streams. It took a value of +1 for consistent combinations (e.g. MT and motion, V4 and color), and -1 for inconsistent ones (e.g. MT and color, V4 and motion). For this purpose, task cue (shape) and color were designated as ventral stream domains, and V4 and PIT ventral stream areas; motion direction was designated a dorsal stream domain, and MT, LIP, and FEF dorsal stream areas. Since PFC is thought to integrate across both streams (29, 38), all factor/area combinations involving it were set to zero for this predictor. To equate the range of stimulus information and category index values across factors, values across all areas for each factor were normalized to range from 0 to 1. Prior to this normalization, categorical information values were rectified at zero, thus equating areas that match the sensory-based prediction with those that less than it (reflecting slight biases toward within-category selectivity). Fitting of this model to the normalized summary values was performed as described above in Neural information analysis, and evaluated by the variance explained by each of the two predictors. Errors and significance levels were estimated by bootstrapping these statistics over the contributing neurons 100,000 times.