Using neural population decoding to understand high level visual processing

By

Ethan M. Meyers

B.A. Computer Science
Oberlin College 2002

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTORATE OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2011

Signature of Author……

Ethan Meyers
Department of Brain and Cognitive Sciences
October 28, 2010

Certified by……

Tomaso Poggio
Eugene McDermott Professor
Thesis Supervisor

Accepted by……..

Earl K. Miller, PhD
Picower Professor of Neuroscience
Director, Brain and Cognitive Sciences Graduate Program

Using neural population decoding to understand high level visual processing

By

Ethan M. Meyers

Submitted to the Department of Brain and Cognitive Sciences
on October 28, 2010 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Computational Neuroscience

Abstract

The field of neuroscience has the potential to address profound questions including explaining how neural activity enables complex behaviors and conscious experience. However, currently the field is a long way from understanding these issues, and progress has been slow. One of the main problems holding back the pace of discovery is that it is still unclear how to interpret neural activity once it has been recorded. This lack of understanding has led to many different data analysis methods, which makes it difficult to evaluate the validity and importance of many reported results. If a clearer understanding of how to interpret neural data existed, it should be much easier to answer other questions about how the brain functions.

In this thesis I describe how to use a data analysis method called 'neural population decoding' to analyze data in a way that is potentially more relevant for understanding neural information processing. By applying this method in novel ways to data from several vision experiments, I am able to make several new discoveries, including the fact that abstract category information is coded in the inferior temporal cortex (ITC) and prefrontal cortex (PFC) by dynamic patterns of neural activity, and that when a monkey attends to an object in a cluttered display, the pattern of ITC activity returns to a state that is similar to when the attended object is presented alone. These findings are not only interesting for insights that they give into the content and coding of information in high level visual areas, but they also demonstrate the benefits of using neural population decoding to analyze data. Thus, the methods developed in this thesis should enable more rapid progress toward an algorithmic level understanding of vision and information processing in other neural systems.

Thesis Supervisor: Tomaso Poggio
Title: Eugene McDermott Professor

$$E = 2r$$

J. Cleese, Podcast #33: The Brain Explained
(http://www.youtube.com/watch?v=FQjgsQ5G8ug)

Acknowledgements

Thesis committee

First of all I would like to thank Pawan Sinha, who got me started on vision related research. Through his creativity, and sense of humor, I learned how fun studying computer and human vision can be. Second, I would like to thank my advisor, Tommy Poggio. Tommy understands the benefit of taking many different approaches to studying the computations that underlie vision, and consequentially, he has put together an amazing group at the CBCL. By being surrounded by a broad range of research topics, through Tommy's insights, and by having the freedom to explore what has interested me the most, I have probably learned more in his lab than would be possible anywhere else. Third, I would like to thank Gabriel Kreiman, who has given me some of the best feedback on my work, and has been incredibly generous with his willingness to listen and help me with my work. Fourth, I would like to thank Emery Brown. Emery studies two of the topics that I find most interesting, statistics and consciousness, and his rigorous approach and amazing ability to do work in two fairly different fields, have been an inspirational. Finally, I would like the thank Jim DiCarlo. While Jim is not technically on my committee, I have had a great time discussing ideas with him at his lab meetings, and I have a strong appreciation of his work since he is one of the leading neuroscientists who realizes that a real understanding of the vision will come when we can build a computer system that can emulate visual processing in the brain.

Collaborators

The results in this thesis would not be possible without the hard work of the researchers who dedicated a tremendous amount of time collecting the data that I analyzed. First, I would like to thank Dave Freedman, Jefferson Roy and Earl Miller for collecting the data that is used in Chapter 3. The data Dave collected has led to the publication of four papers, which speaks volumes about the quality of the data and the experiment. Also, the task that Jefferson was able to train his monkeys on is very impressive, so I am very grateful he was willing to share this data with me. Second, I would like to that Hamdy Ebark, Winrich Freiwald, for collecting the data used in Chapter 4, and also I would like to thank Thomas Serre for his intellectual contributions to that project. While ultimately the results did not turn out exactly as we had hoped, I did learn a lot being involved in this project. Third, I would like to thank Ying Zhang, Narcisse Bichot and Bob Desimone for contributing the data and to the ideas in Chapter 5. Working with Ying was fun, and I also enjoyed the many discussions I had with Bob and Tommy as we wrote the paper together. Finally, I would like to thank Jennie Deutsch and Jim DiCalro for their help in collecting simultaneously recorded V4 data that is used in Chapter 6. Jim was incredibly generous in allowing me, Joel and Cheston to run a set of experiments for couple of months early in 2009, and I really appreciate the time Jennie spent collecting the data (and I would like to thank Joel for the time he spent spike sorting the

data!). The recordings made in that project have convinced me that using chronically implanted arrays to record data is the best way to proceed in the future.

Family and friends

Lastly, but not leastly, I would like to thank everyone who has helped keep me stay sane and motivated. First of all I would like to thank my parents who have been incredibly supportive. I know a lot of people who like their parents, but it's hard for me to imagine that I could have gotten a better pair. Second, I would like to thank Ben, Danielle and Beata for being some of my best friends outside of the lab (and also Adam and Mike for being fun roommates). I would also like to thank Beata the discussions of neuroscience and feedback on much of my work (I know it was often painful, but my work is much better because of it). Third, I would to thank all the grad students and post-docs who have been my friends over the years. In particular, I would like thank the 'karaoke crew', including Dominique, Tracy, Michelle and Barbara who were a lot of fun to hang out with, the 'computational neuroscience quals crew', John, Chris, and Viren, who I learned from (and also Srini), the dedicated poker players Simon and Max, and the Foosball players, Lorenzo, Stan, Bernd, Thomas, Barbara and Jake (your foosball and trash talking skills may still need work, but it was still fun to beat you all) . Fourth, I would like to thank all the members of the CBCL including: Joel, Charlie, Cheston, Jim, Sharat, Huei-Han, Ulf, Lorenzo, Stan, Thomas, Rif, Bernd, Tony, Gadi, Kathleen and particular Jake, who answered tons of my questions the first few years in the CBCL. Finally, I would like to thank Leo and Pearl for making me some great Caeser salads for lunch

Table of Contents

.

Chapter 1: Introduction

While many significant discoveries have been made in the field of neuroscience, we are still a long way from understanding the neural processing that enables complex behaviors. One key reason why the field does not have a better understanding of neural information processing is due to the fact that most analyses of neural data use a 'hypothesis testing' approach that is designed to demonstrate that particular effects are present in the activity of single neurons, rather than focusing on how the brain works as a computational system designed to solve particular tasks. Consequentially, the field of neuroscience seems to be overrun by facts, while lacking a real understanding of how the brain functions.

In this thesis, I discuss how to apply a data analysis method called 'population decoding' to reveal a deeper understanding of neural information processing. Population decoding works by modeling the relationship between the activity of a population of neurons and particular experimental conditions (see chapter 2 for more details). Once this relationship has been learned, predictions can be made about which experimental condition are present in new data, which allows one to assess how reliability of the relationship between neural activity and particular experimental conditions. The level of accuracy of these predictions can then be used in several ways to give insight into what computations a brain area is most involved in, and also how a given brain region is representing and processing information. For example, one can use the activity of a population of neurons to make predictions about several different types of variables that are behaviorally relevant, and then evaluate how much information there is about each of these variables in order to understand the key functions that a given brain region is involved in. Conversely, one can make predictions about one particular variable using different representations (or types) of neural activity, in order to gain insight into how information is coded in neural activity. One can also compare these predictions accuracies to the performance of humans, animals and computer algorithms on similar tasks, in order to relate the amount of information in neural activity to other systems that are trying to

solve the same task. Population decoding also has several other advantages over more conventional data analysis methods (see Chapter 6), including the ability to more clearly identify particular states the brain is in by combing the noise signals that each individual neuron has into a better picture of how a brain region is functioning as a whole. Thus, this method offers a lot of promise in terms of gaining a more computational understanding of neural processing.

The work in this thesis focuses on understanding high level visual processing by analyzing data from the inferior temporal cortex (IT) and other closely connected brain regions that are involved in the later stages of processing visual information (and possibly object recognition). Neural population decoding analyses have been applied to several other brain regions (they have been most widely used to analyze neural activity in the motor cortex of macaques, and hippocampus of rats), and decoding methods are widely used to analyze data from functional magnetic resonance imaging studies of visual processing in humans. However, apart from a few notable studies (Gochin et al., 1994; Hung et al., 2005), prior to the work in this thesis, population decoding has not been widely used to analyze data from macaque high level visual areas. Thus, applying population decoding to neural data from high level visual areas is an interesting problem because it requires new techniques to be developed, and because of the greater potential to make significant discoveries. Also, because of large amount of work being done in computer vision, findings from neuroscience in this area could lead to practical applications by giving insight into how to build better computer vision systems, and conversely, insights from the field of computer vision could help shed light into how to interpret the population decoding results, which makes it an attractive problem to research in terms of gaining a computational understanding of neural processing.

The majority of the content of this thesis concerns analyzing neural data using population decoding in order to make new discoveries. The reason why this work focuses on analyzing data rather than on developing methods, is because I felt that if population decoding did not give any additional insight compared to other methods, then proving mathematical properties about the method would not be very useful for advancing our

understanding of how the brain functions[1]. In the process of analyzing data, however, I did end up creating several methods that allow one to evaluate the significance of the results. I also spent a large amount of time making sure that the decoding results are robust. For example, before I began doing these decoding analyses, it was unclear whether the decoding results would be affected by the choice of classifier, the way the data is normalized, or a number of factors. If the decoding results varied drastically depending on these choices, then the method would not be very useful for making general statements about how the brain functioned. Consequentially, I spent a large amount of time exploring how different parameter choices could affect the decoding results, and fortunately, the empirical results have shown that decoding is remarkably robust to most parameter choices. Finally, I spent a considerable amount of time developing code to do these decoding analyses. I hope to release this code soon to allow other researchers to easily apply the decoding analyses to their data in order to help increase the pace of discovery.

[1] Chapters 3-5 clearly show that many additional insights about high level visual processing can be made using this method.

Organization of this thesis

The remainder of Chapter 1 gives background on the what is known about interior temporal cortex (which is the brain region where most of the data analyzed in this thesis comes from) and background on other data analysis methods that are relevant for studying processing of high level visual information.

Chapter 2 describes how population decoding works, and how to apply these methods to neural data in order to get useful results. The main goal of this chapter is to provide a guidance to help researchers, who have not had much experience with machine learning, successfully apply these methods to data from new experiments.

Chapters 3-5 describes findings from applying population decoding to neural data from high level visual areas. Chapter 3, shows that the both the inferior temporal and prefrontal cortex contain 'abstract' category information in the activity of a small subset of neurons that change over the course of a trial. Chapter 4 compares neural decoding results to results from computational models and psychophysics experiments, and shows that while it was possible to decode whether animal is in a complex natural scene image using neural data, this information is also present in simple low level visual features found in earlier visual areas. Finally, chapter 5 shows that attention affects neural representation of objects in IT primarily by restoring the pattern of activity to a state this is similar to when an attended object is shown in isolation.

Chapter 6 concludes the thesis by highlighting some common themes that span the chapter in the thesis, listing some of the advantages that neural population decoding has over other methods by taking examples from the results in the thesis, and discusses some interesting questions that can be addressed using this method in the future.

Background: The inferior temporal cortex

The data analyzed in this thesis comes from recordings made in the anterior inferior temporal cortex (IT) and a few connected brain regions (namely, the prefrontal cortex and V4). Below is a literature review of the inferior temporal cortex, which helps put the findings of this thesis in the context of what is currently known about this brain region.

Visual information enters the cortex of primates through a pathway that starts with the retina in the eye, passes through the lateral geniculate nucleus in the thalamus, and then enters cortex in a posterior region (occipital cortex) in a region known as primary visual cortex (V1) (Bear et al., 2006). After V1, visual information is thought to travel along two distinct pathways, known as the dorsal stream and the ventral stream (Mishkin and Ungerleider, 1982; Goodale and Milner, 1992). The dorsal stream is thought to be more involved in the processing of vision for action (it is often called the 'where' or 'how' stream), while the ventral stream is thought to be more involved visual object recognition (it is often called the 'what' stream). The feedforward connections of the ventral stream proceed from V1 to visual areas V2, V4 and then on to the inferior temporal cortex (IT), which lies in the temporal lobe, and contains the mid and late stages of the ventral visual pathway (Mishkin et al., 1983; Felleman and Van Essen, 1991). The major project sites for the outputs of IT are the medial temporal lobe which is thought to be involved in memory, and to ventral prefrontal cortex, which is thought to be involved in planning and decision making (Miyashita, 1993).

IT contains multiple subdivisions that have slightly different properties (Baylis et al., 1987; Felleman and Van Essen, 1991; Logothetis and Sheinberg, 1996), although IT is most often discussed in terms of two regions, posterior/central IT (PIT/CIT, or the closely related TEO), and anterior IT (AIT, or the closely related TE) (von Bonin and Bailey, 1948; Felleman and Van Essen, 1991). Overall neurons in IT have longer latencies (Nowak and Bullier, 1998), larger/bilateral receptive fields and respond to more complex stimuli than regions than earlier vision area (Logothetis and Sheinberg, 1996; Fujita,

2002). The complexity of features and the size of receptive fields also continues increase from PIT to AIT (Tanaka et al., 1991).

IT has traditionally been thought of as a brain region that is necessary for visual object recognition (Mishkin, 1982; Gross, 1994), although recent studies have raised questions about its role (Kirchner and Thorpe, 2006; Girard et al., 2008; Matsumoto et al., 2010). Below, I describe research in IT in terms of several different conceptual approaches that have been used to guide research questions about the function of this brain region. More detailed literature reviews of IT cortex and object recognition in primates can be found in (Miyashita, 1993; Gross, 1994; Logothetis and Sheinberg, 1996; Tanaka, 1996; Ashbridge and Perrett, 1998; Fujita, 2002).

Examining IT neurons' responses to visual features

One widely studied question in IT concerns trying to find the set of 'visual features' that individual neurons respond to. This approach has its origin in seminal early work on understanding properties of early visual stages of processing in the retina (Hartline, 1938; Lettvin et al., 1959) and primary visual cortex (Hubel and Wiesel, 1959) where simple visual features such as orientation, and spatial frequency can be used to predict a large degree of the responses of these cells. Several studies using this conceptual approach have found that many neurons in AIT and the superior temporal sulcus (STS), respond much more strongly to images of faces and other complex objects (Gross et al., 1972; Bruce et al., 1981; Perrett et al., 1982; Desimone, 1991). Numerous other studies have tried to characterize the visual response properties of non-face selective IT neurons by using parameterized stimulus sets, including Fourier descriptors (Schwartz et al., 1983), radial frequency components (Op de Beeck et al., 2001), Walsh patterns (Richmond et al., 1987), contour elements (Brincat and Connor, 2004), and other intuitive axes such as elongation and curvature (Kayaert et al., 2005). While these studies have elucidated several interesting facts about IT, such as neurons often have the highest firing rates to the parameters extreme elements in their stimulus set (Brincat and Connor, 2004) and perceptual ordering of similarity matches neural response similarity (Op de Beeck et al.,

2001), a clear set of parameters around which IT neurons are tuned has yet to be found. Another method to analyze which features IT is tuned to consists of a 'stimulus reduction' method in which a neuron is shown a series complex images (usually of isolated objects or shapes), and then the image that initially gave a large response is reduced in its complexity to try to find the simplest set of visual features that still maximally excite the neuron (Desimone et al., 1984; Tanaka et al., 1991; Kobatake and Tanaka, 1994; Tanaka, 1996, 2004). Such approaches have found most AIT neurons respond selectively to a particular feature that is usually of "moderate" complexity (Tanaka, 1996) yet again, no simple description of IT neuron tuning has emerged from these studies. While having a clear understanding of what visual features neurons in IT are tuned to would be a very significant breakthrough, it is unclear whether one should expect single neurons in IT to be tuned to along any easily interpretable dimensions (Serre et al., 2005). It is also possible that even if IT does represent visual properties along easily parameterized dimensions, that if the information is coded in a distributed manner, then examining the firing rates of individual neurons might not clearly reveal what these dimensions are.

Visual feature topology in IT

A related question to which visual features are neurons in IT tuned to, is the question of whether there is any topographic organization in IT. It is known that AIT does not have the retinotopic organization that characterizes many of the early visual areas from V1 to V4 and PIT (Tanaka, 1996, 1997). However, there is a significant amount of evidence that AIT neurons are grouped together based on their selectivity to visual image properties (even though the exact properties that these neurons respond to is unknown). A recent study has shown that faces selective neurons in IT and STS are grouped into several discrete patches that are interconnected (Moeller et al., 2008), which suggests that there are dedicated regions of cortex devoted to processing faces. Beyond face selective IT neurons, functional magnetic resonance imaging studies in macaques have shown consistent patterns of activation to images of objects with different shape properties that were not influenced by experience with the images, behavioral task, and the exact retinal

position of the objects, indicating that there is a large scale topographic map in IT based on the visual properties of images (Op de Beeck et al., 2008). On a finer scale, electrophysiological and optical imaging studies have reported "columns" in AIT in which all neurons along a vertical penetration seem to have similar selectivity, and neurons within a 400-500 micron wide region have similar selectivity, but beyond these regions selectivity of neurons changes sharply (Fujita et al., 1992; Tanaka, 1996; Wang et al., 1998; Tsunoda et al., 2001). While there has been speculation on the functional role of this topographic organization (Tanaka, 2003) the reason for this organization is still unknown.

Memory and learning in IT

Other studies have investigated visual learning and memory effects in IT (Mishkin 1982). One widely reported memory related effect, seen in a significant number of neurons (~30%), is a *decrease* in response when the same stimulus is repeatedly shown two or more times (Miller et al., 1991; Li et al., 1993; Miller and Desimone, 1994; Sawamura et al., 2006). This suppression in response occurs even if the repeated stimulus is not behaviorally relevant (Miller and Desimone, 1994), and also when the repeated stimulus is shown at a different size or position (Lueschow et al., 1994). Interestingly, when a monkey engages in a delayed match-to-sample (DMS) task, the decrease in response occurs from the "sample" to the "match" stimulus, but, there is a rebound/resetting of response on the next trial, such that the sample response is high again (i.e., significantly higher than the response to the previous match, even when the match and sample stimuli are shown in successive order) (Miller et al., 1993; Li et al., 1993). Additionally, the neurons that show a decreasing response to repetitions within a trial also show a longer lasting, more gradual decrease in response to both the sample and the match stimulus (particularly in neurons in ventral AIT and the perirhinal cortex), such that the more a particular stimulus is seen, the lower the response is to that stimulus. Thus these neurons have been suggested to be involved in representing stimulus "familiarity" (Li et al., 1993). A second memory effect seen in IT is an *enhancement* in neuronal response to the repetition of a stimulus only when the stimulus is *behaviorally relevant*, e.g, if a stimulus

is the "match" in a DMS task, and not just a repetition of a non-match stimulus (Miller and Desimone, 1994). Almost no neurons that show an enhanced response for behaviorally relevant stimuli show the decrease in response effects, which suggests that there are two separate networks in IT what underlie these different memory related responses (Miller and Desimone, 1994).

In addition to studying changes in a neuron's mean responses due to stimuli that have been presented in the past, other studies have examined short-term memory in IT in relation to a neuron's sustained temporal profile of response after the offset of a particular stimulus. Studies using a DMS task have found neurons with stimulus specific modulation (including constant high levels of sustained activity) to colors and to particular shapes over delay periods that are up to 20 seconds in length (Fuster and Jervey, 1982; Miyashita and Chang, 1988). Other DMS studies, though, have questioned whether these sustained responses are memory signals since intervening stimuli disrupt the stimulus specific modulation even though they do not interfere with the monkey's ability to complete the task (Baylis and Rolls, 1987; Miller et al., 1993). However, a recent study using a population decoding analysis has shown that the disruption caused by subsequent stimuli is not complete (Woloszyn and Sheinberg, 2009), and that residual information does persist over intervening stimuli[2].

Many studies have also examined longer lasting "learning effects" in IT in which neurons appear to change their stimulus selectivity as a result of extensive experience with a particular stimulus set or task. Studies have shown that if a monkey *passively* views a series of images that are in a fixed order numerous times, neurons in AIT will tend to respond similarly to images that are adjacent to each other in the learned sequence, even

[2] The first version of this literature review was written for my thesis proposal, and was completed before the study by (Woloszyn and Sheinberg, 2009) had been published. In that version of this literature review, I speculated that a memory signal that persisted over intervening stimuli might be found in IT if it a more sensitive analysis based population decoding analyses was conducted. The reason I thought one might find such a memory trace was based on the fact that a similar memory trace had been found in the locust olfactory bulb when a population decoding analysis was done by Broome et al., (2006). Thus I found it very interesting to see that this speculation was confirmed by the results of Woloszyn and Sheinberg, (2009).

when the images are presented later in a random order (Miyashita, 1988; Erickson and Desimone, 1999). Additional studies have also shown that many neurons become more selective for stimuli that are frequently seen (Logothetis et al., 1995; Kobatake et al., 1998; Baker et al., 2002; Freedman et al., 2006), although most studies have found that individual neurons still respond to multiple stimuli rather than being tuned exclusively to one stimulus (Kobatake et al., 1998; Baker et al., 2002; Freedman et al., 2006). The increased selectivity of neurons seen in several of these studies (Baker et al., 2002; Freedman and Assad, 2006) is due to decrease in firing to non-preferred stimuli rather than increase in response to the preferred stimuli, and thus could be related to the familiarity effects described by Li et al. (1993).

Other research has looked at changes in neuronal responses that could be a result of monkeys participating in an *active* task. Several studies had monkeys engage in a paired-associate (PA) task, where monkeys needed to learn to associate pairs of stimuli together, such that when the first cue stimulus was shown, the monkeys needed to correctly select a particular second stimulus from two alternative choices. Results from these studies found two effects: first, that many neurons respond at a higher firing rate to both images in the pair (acting as if they associated both stimuli together) (Sakai and Miyashita, 1991; Messinger et al., 2001) and second, many other neurons that responded particular strongly to a particular image, would respond with an increasing firing rate during a delay period when the paired image was used as a cue, thus acting as if the neuron was anticipating the onset of the pair image (Sakai and Miyashita, 1991). However, further work has suggested that such paired-associate effects are stronger in adjacent perirhinal cortex than in IT proper, and that paired associations in IT are large due to feedback connections from this area (Higuchi and Miyashita, 1996; Naya et al., 2001). Other studies have shown that when a monkey must discriminate between two visually distinct classes of items, more neurons are found that respond to visual features that can differentiate between the two image classes than to features that cannot differentiate between the classes (Sigala and Logothetis, 2002). Additionally, the reward structure associated with particular stimuli and behavioral responses can influence response properties of neurons in AIT (Mogami and Tanaka, 2006; De Baene et al., 2008),

providing further evidence that active engagement in a task is an important factor shaping neuronal activity.

Invariant object representations in IT

Another framework for studying IT is to examine which transformations of a given stimulus are neurons invariant/tolerant to (Ashbridge and Perrett, 1998; Rolls, 2000). The logic behind this approach is that primates need to be able to recognize objects under a variety of different viewing conditions such as changes to the object's size, position, ambient illumination and surrounding clutter, all of which give rise to very different retinal images of an object; thus, if IT is critical for robust object recognition, there should be neurons in this brain region that respond similarly despite changes in such parameters. Several single neuron analyses examining this issue have found neurons in anterior IT that respond similarly to images of particular objects even when the object is shown at different sizes and positions (Schwartz et al., 1983; Miyashita and Chang, 1988; Lueschow et al., 1994; Ito et al., 1995), although the majority of neurons in IT do seem to respond best to a particular size/position[3] (Lueschow et al., 1994; Ito et al., 1995; Ashbridge and Perrett, 1998; DiCarlo and Maunsell, 2003). Other studies that have looked at more complex transformation of stimuli, including shape defined by texture and motion (Sary et al., 1993), mirror reversals of shapes (Rollenhagen and Olson, 2000; Baylis and Driver, 2001), contrast changes/reversal (Baylis and Driver, 2001; Zoccolan et al., 2007), and rotation of familiar 3D shapes (Logothetis et al., 1995; Booth and Rolls, 1998), have also found that there are neurons in IT that respond similarly despite these changes in stimulus properties, although again, many neurons are more tuned to particular ranges of parameters and there seems to be a tradeoff between how selectively a neuron respond to particular stimuli and how tolerant it is to different transformations (Zoccolan et al., 2007)).

[3] However even when neurons do respond more to a particular size/location, the ordinal order of stimulus selective for almost all neurons seems to remain the same at the preferred and non-preferred locations – thus the changes at a preferred size/location can best be explained in terms of a change in gain in the neuron's response (Lueschow et al., 1994; Ito et al., 1995; DiCarlo and Maunsell, 2003) .

Learning mechanisms have also been proposed that can explain how neurons in IT could obtain such invariant properties, which ties together the learning effects seen in IT with IT cortex's involvement in shape discrimination. In such theories, the fact that there is often temporal contiguity to the transformations that images of objects undergo in the world is combined with an associative learning so that neurons become invariant to particular object transformations (Földiák, 1991; Wallis and Rolls, 1997; Wiskott and Sejnowski, 2002). For example, since objects are often seen at slightly different sizes in a precise temporal sequence as one approaches an object, a Hebbian learning rule could cause a downstream neuron to pool together the responses of two upstream neurons that each respond only to one image of a particular size, due to the upstream neurons firing in close temporal proximity; thus after such learning, the downstream neuron would respond similarly to a particular object regardless of the object's size. Similar mechanisms could explain the position, illumination and view tolerant neuron responses found in IT. Psychophysical evidence has shown that humans indeed experience perceptual learning for temporally contiguous images, such that images that occurred in a temporal sequence are subsequently perceived as being more similar (Wallis and Bülthoff, 2001; Cox et al., 2005) and computational models have been build around this principle (Földiák, 1991; Wiskott and Sejnowski, 2002). Also, recent neurophysiological experiments have shown that temporal binding can indeed change neuron's selectivity and create 'false invariances', which is strong support for this theory (Li and DiCarlo, 2008).

However, despite the great appeal of this theory, it is difficult to explain how all of the response properties of IT neurons could arise from such learning mechanisms. For example, a recent study has suggested that IT neuronal response properties might be organized by whether a stimulus is an image of a living or non-living item (Kiani et al. 2007), which is hard to explain in terms of associative learning rules. Thus IT is most likely involved in more than just creating invariant object representations.

Neural Coding

Apart from the issue of *what* information is in IT and other high level visual areas, is the issue of *how is information coded* in the neural activity in these areas. While it is clear that much of the information in neural spiking activity is present in the 'firing rate' of the neuron[4], many questions about whether additional information is contained in other aspects of neural activity have not been definitively answered (Dayan and Abbott, 2001). Questions related to information coding in single neurons include: 1) For a single neuron, is the information contained in each spike independent from the information contained in all other spikes, so that a inhomogeneous Poisson process is a full characterization of the information content of the neuron (an independent spike code), or is there additional information in the relationship between spike times (a temporal correlation code); 2) Is information contained in precise spike/rate *modulations* that are not purely due to the dynamics of the stimuli (temporal encoding), or does the mean number of spikes within a particular time interval contain all the relevant information (spike-count code)? and 3) over what temporal time scales do neurons carry information[5]?

Work in the primate visual system that has tried to address these questions includes a study by Victor and Purpura, (1996) who used a 'metric-space' analysis to examine data from V1, V2 and V3, and found that there indeed appears to be additional information in the precise times of spikes beyond what could be accounted for by an inhomogeneous Poisson process (with additional information ranging on time scales from 10-100ms

[4] For the purpose of this review, we use the term 'firing rate' to refer to either the 'spike-count rate' which is defined as the number of spikes within a particular time interval *within a single trial*, or the 'firing rate' which is defined as the number of spikes within a particular (perhaps smaller) time interval that is calculated *by averaging over repeated trials* of the same type. Both of these firing-rates measures are known to contain significant amounts of information about stimuli and other behaviorally relevant variables, although from a theoretical view point, they are significantly different from each other, and can potentially lead to very different view points about how the brain processes information (Dayan and Abbott, 2001).

[5] This question is often referred to as the temporal vs. rate coding debate, however as pointed out by several researchers (Theunissen and Miller, 1995; Dayan and Abbott, 2001; Gerstner and Kistler, 2002), the distinction between a rate code and a temporal code (as it is most commonly defined) is just a matter of degree rather than any concrete coding strategy difference, thus I will avoid using this terminology.

depending on the stimulus attribute). However, work by Heller et al., (1995) using an information theoretic approach based on decoding failed to find additional information beyond what was present in the mean spike counts over time windows of 25ms in V1 and 50ms in IT. Additionally, a series of studies by Optican and Richmond (Optican and Richmond, 1987; Richmond and Optican, 1987; Richmond et al., 1987) that used information theoretic approach applied to temporal principal components of binned spike trains from IT showed what appeared to be additional information in the temporal modulation of neural activity beyond modulations to the mean firing rate, suggesting that there was *temporal encoding* of information in IT. However, later work by (Tovee et al., 1993), found that the analyses of Optican and Richmond could largely be accounted for the fact that Optican and Richmond did not correct for a systematic limited sampling bias in their information estimates, and that indeed most of the information in IT could be accounted for based on a the temporal modulation of the mean firing rate and the initial onset latency of the IT responses.

Apart from questions about neural coding on the single neuron level, there are also many open questions related to how information is coded in populations of neurons, including: 1) Within a population of neurons, given a particular stimulus, is the coding of information within each neuron independent of the activity of all the other neurons (an 'independent-neuron' or 'population' code), or does the correlated activity (i.e., 'noise correlations') between neurons contain additional information (i.e, an 'ensemble code') (Latham and Nirenberg, 2005). 2) Is the *information* carried by different neurons highly redundant so that many neurons contain the same information about a particular stimulus/condition (redundant code) or does each neuron contain unique information or potentially more information than what is contained in the neurons individually (synergistic code)?[6] 3) Is the spiking activity of neurons sparse, or are many neurons

[6] As pointed out by (Latham and Nirenberg, 2005), whether neurons are statistically independent from one another given the stimuli, is a slightly different questions than whether neurons code information redundantly or synergistically. To measure statistical independence, Latham and Nirenberg, (2005), use the formula $p(r_1, r_2, ..., r_N \mid s) \neq \prod_i p(r_i \mid s)$. If the joint distribution is not equal to the product of the marginal distributions, then there is significant information in the 'noise correlations' in the data. To measure redundancy/synergism the formula of Schneidman et al., (2003) is used which is:

activity at the same time? 4) Is *information* highly distributed across many neurons or is their a small compact subset of neurons that contains all the information at any given point in time? and 5) Is the population code stationary such that one pattern of activity represents a particular stimulus at all points in time, or are there multiple patterns of neural activity that represent a given stimulus?

Research in the visual system that addresses these population coding issues includes: 1) Several studies have claimed to show that neurons largely act as independent encoders of information. For example, studies as of pairs of neurons in the retina (Nirenberg et al., 2001) and V1 (Golledge et al., 2003) showed at least 90% can be extracted by ignoring correlated activity and using only mean firing rates, while studies in IT (Aggelopoulos et al., 2005; Anderson et al., 2007) found that 94-100% of the information was in the spike count alone, and only 0-4% of information was in the correlated activity in pairs of neurons. Other researcher however have claimed that additional information can be found in the correlated activity of neurons including work by (Dan et al., 1998) who found that there is on average an additional 20% increase in information when treating correlated activity in strongly correlated pairs as an additional information channel, and work by (Pillow et al., 2008) who used recordings from parasol retinal ganglion cells to show that modeling the full coupling of all neurons in the retina leads to a 20% increase in the amount of information that can be decoded. 2) Studies by Gawne and colleagues of redundancy in the information carried by neurons within a population have shown that for both V1 and IT approximately 20% of the information carried by pairs of neurons is redundant (Gawne and Richmond, 1993; Gawne et al., 1996). 3) Work by Vinje and Gallant, (2000) has shown that showing monkeys natural images leads to sparse *activity* in V1, while work by Rolls and Tovee, (1995) has found that IT neurons respond to a large fraction of images shown, indicating that neurons in IT do not have sparse activity. 4) Work by Meyers et al., (2008) used a decoding based approach to show that in IT and

---

$\Delta_{synergy} = I(s;\mathbf{r}) - \sum_i I(s;r_i)$, where I(s; r) is the mutual information, **r** is the population response, and r_i is the response of the i^{th} neuron. $\Delta_{synergy}$ is a measure of how much more information is gained by the observing more neurons.

PFC at any point in time there is a small compact subset of neurons that contain all the information that the larger population has. 5) Several studies from the PFC, IT, V1, MT and other areas and in other animals (Nieder et al., 2002; Laurent, 2002; Baeg et al., 2003; Zaksas and Pasternak, 2006; Meyers et al., 2008) have shown that information is contained in dynamic patterns activity such that different patterns of activity are elicited by the same stimulus at different latencies relative to stimulus onset.

Background: Data analysis methods

Many different methods have been used to analyze neural data. Below I will discuss a three methods that are particularly relevant for the work discussed in this thesis. The first method is based on applying standard statistical tests to measurements derived from firing rates of neurons is probably the most widely used data method for analyzing *what* information is in particular visual areas. The second method is based on mutual information measurements, and has been widely used to examine questions of neural coding in the visual system. Finally, the third method discussed is neural population decoding analyses which have been used to widely in the study of the rat hippocampus and the primate motor system, and is used in this thesis to analyze data from the macaque visual system (more details about how to apply this method to neural data from visual experiments is discussed in chapter 2). For more general reviews of statistical methods in neuroscience see Brown et al. (2004) and Kass et al. (2005).

Hypothesis testing using standard statistical tests (and ad hoc indices)

One of the most widely used methods to determine presence of an effect in neural data is to apply conventional statistical tests (e.g., t-tests, ANOVAs, etc.) to the firing rate of a neuron in particular conditions. Some commonly used hypothesis tests in study of neural responses in primate vision include: 1) determining whether a neuron is 'visually responsive' by applying a t-test between the firing rates of the neuron in a baseline period before the onset of a stimulus and to the firing rates of the same neuron after a stimulus has been shown, and 2) determining whether a neuron is 'visually selective' by applying

a one-way ANOVA to the neural data using the different stimuli that were shown as conditions. If the p-value for these test is below a particular alpha level (usually set at .05 or .01) then the neuron is considered visually selective/responsive. Also, it is common practice to calculate ad hoc 'tuning/selectivity index values' for each neuron in a given brain region, which usually consists of subtracting firing rates from different conditions and then normalizing the results. Significant differences *between* different brain regions or conditions are often later calculated by applying another conventional statistical test to the *number* of selective neurons (or to the distributions of index values) from each brain area/condition.

While these methods can be useful, there are two potential pitfalls with conclusions drawn from these types of analyses. The first problem is that such analyses examine the properties of each neuron individually, and do not take into account information that is available in the joint activity of many neurons, which is contrary to the widely believed theory that information is coded in distributed patterns of activity across many neurons. The second problem in these analyses is that it is often assumed that all neurons from a given brain region are homogenous in nature (i.e., randomly distributed from the same underlying distribution), rather than coming from a diverse population. Consequently, these analyses assume that the *number* of selective neurons in a given brain region is a good indicator of how involved that brain region is in a particular task, which is questionable assumption to make and could potentially lead to wrong conclusions if the assumption is wrong. For example, it is possible that if the neurons in a brain region A are more specialized for particular tasks than a region B (i.e., the representation in A is less distributed and more explicit than the representation in B), then A might have a much lower percentage of neurons involved in any given task than B, however to conclude that B is more involved in each task than A would be a mistake. Furthermore, looking at the mean value of particular indexes (or calculating statistics over the whole population between A and B) could be misleading in such cases (although looking at the highly selective outliers could be informative).

Mutual information measures

Another widely used method to analyze data (particular in relation to answering questions about neural coding, is to calculate the mutual information between a stimulus, or other relevant condition, and a neuron's response. Mutual information was originally devised to address issues involved with 'channel coding' (i.e., it was used to evaluate how much of the capacity of a communication channel is used when a particular code is chosen to describe a particular set of data (Shannon, 1948), however it can be more broadly interpreted as a measure of the amount of information that can be obtained about one variable when observing the value of another variable.

There are several conceptual ways to view mutual information between two random variables X and Y. One common way is to view mutual information (denoted I(X; Y)), as the Kullback–Leibler divergence between the full joint probability distribution P(X, Y), and product of the two marginal distributions, P(X) and P(Y). The equation for mutual information in this case is:

$$I(X;Y) = D_{KL}(P(X,Y), P(X)P(Y)) = \sum_{x,y} P(x,y) \log_2 \left(\frac{P(x,y)}{P(x)P(y)}\right) \quad .$$

This formulation makes explicit the fact that the mutual information is measuring how *independent* the random variables X, and Y, and it also makes it readily apparent that mutual information is symmetric in its arguments (i.e., I(X; Y) = I(Y; X)). If X and Y are independent, then P(X, Y) = P(X) P(Y) and so log[P(X, Y)/P(X)P(Y)] = 0, and hence there is no mutual information between these variables.

Another common and mathematically equivalent way to view mutual information is as the difference in entropy between the marginal distribution P(X) and the entropy of the conditional distribution P(X|Y). The entropy of a distribution (denoted H[P(X)]) is a functional that gives a measure of the 'uncertainty' in a distribution (e.g., if the values in a probably distribution are all equally likely then there will be high uncertainty for the outcome of a particular random experiment and hence high entropy, while if all the mass in probability density function is centered on one value then the outcome is certain to be

that particular value, and hence there is no entropy). The mathematical formula for entropy is $H[P(X)] = -\sum_x P(x)\log_2 P(x)$, and the hence the formula for mutual information can be written as I(X; Y) = H([(X)] – H[P(X|Y]). The interpretation of this formulation is that if P(X) has a high value, many different signals will be present, and hence a lot of information can potentially be communicated. However, if in the process of trying to communicate this information to a receiver who receives the message Y, the information becomes highly corrupt (and hence there is much uncertainty about X after we see the value of Y), then the amount of information that can be communicated is substantially reduced. Thus this formulation stresses the difference between the possible messages X that can be sent, and how much uncertainty remains about X after the transmitted message Y = y is received (or equivalently, if we know the transmitted signal X = x, then how much uncertainty is there about what the received message will be).

To relate these two formulations to neural data, the first formulation based on the KL-divergence can be thought of as how independent is the neural response from the stimulus. If the neural response is completely independent of the stimuli (P(R, S) = P(R)(P(S)), then it contains no mutual information about the stimuli, and it can be interpreted that this neuron is not involved in the coding/processing of these stimuli (or equivalently, one cannot use the stimulus to predict what the neural response will be). Conversely, if there strong degree of dependence between the stimuli and the neural responses, then the neural response contains a large amount of information about the stimuli, and hence it is more likely that these neurons are involved in processing these particular stimuli.

The second formulation, based on the difference between the marginal and conditional entropies, can be related to neural responses by relating the neurons to communication channels. If we view the entropy of the stimulus set H[P(S)] as how likely all possible signals are to come from the world (and hence the total amount of *possible* information), then we can view then H[P(S)] minus the conditional entropy of the stimulus given the neural response (H[P(S|R)]), as how much information remains about the world after it has been transmitted through the neural response (here H[P(S|R)] can be interpreted as

the amount of uncertain about S left after the response R has been seen, and is sometimes referred to as the 'noise-entropy' since it measure how much entropy remains due to a neuron's/channel's variable responses to the same stimulus). Since mutual information is symmetric in its arguments, one can also use the equivalent formula $I(R; S) = H[P(R)] - H[P(R|S)]$. This lead to interpretation of $H[P(R)]$ telling us the total possible amount of information that a given neuron can transmit based on distribution of responses the neuron can generate, and $H[P(R|S)]$ as informing us about how uncertain we still are about what the neural response is, given we know what the stimulus is.

In practice, applying mutual information to neural data has run into difficulties due to the fact that if the data has been under sampled, then estimates of mutual information can have a systematic upward limited sample biased[7]. The reason that this limited sampling bias occurs is that with smaller sample sizes both the unconditional entropy $H[P(R)]$ and the conditional entropy $H[P(R|S)]$ are biased downward, because $P(R)$ and $P(R|S)$ do not take on their full range of values due to the limited sample. However since $H[P(R)]$ has been estimated for all the data, while $H[P(R|S)]$ is estimated separately for each stimulus, the downward bias is larger for $H[P(R|S)]$ and consequently when the conditional entropy is subtracted from the response entropy the mutual information is biased upward (Panzeri et al., 2007). (Similar results occur for $H[P(S)]$ and $H[P(S|R)]$, where $H[P(S)]$ has no bias while $H[P(S|R)]$ is biased downward leading to the same upward bias in mutual information). The bias is also systematic in the sense that the for a constant sample size of N, the larger the potential information the system has (i.e., the larger the marginal entropy $P(X)$), the larger the bias will generally (Panzeri et al., 2007). Thus if different neural signal representations are being compared on the same dataset, and bias correction is not employed, then the representation that has more potential responses will generally have a higher mutual information value due to bias, even if both representations actually carry the same amount of information. An example of this occurring in the literature can be seen in Optican and Richmond (1987), which claims that there is temporal encoding of information based on that fact that they calculate a higher amount of mutual information

[7] A 'limited sample bias' is defined as the difference between the expected value of the probability functional computed from the probability distributions estimated with N samples, and its value computed from the true probability distributions.

between a stimulus and the neural response when the neural response is characterized using more principal components; however this results is an artifact of a bias in mutual information due to the fact that there is a larger entropy in marginal distribution when a higher dimensional neural representation is used. Also it should be noted that obviously as the sample size N increases, the bias tends to decrease. Thus the crucial parameter for how much bias will be present is the ratio of the number of samples per stimulus (denoted Ns) to the number of potentially different responses (denoted N_r). If N_s / N_r is large, then the bias should be small (Panzeri et al., 2007).

Many methods have been devised to deal with this limited sample bias, which can largely be divided into two strategies. In the first strategy, researchers try to overcome the limited sampling bias by assuming particular properties of the neural response in order to effectively reduce the number of potential neural response, which directly reduces the amount of bias for a fixed sample size. An example of this method is to decode the neural response R into a predicted response for what the stimulus is (denoted S'), and then use I(S; S') as a surrogate for I(S; R). Since S' will likely have a smaller range of possible values than the number of potential neural responses R, this estimate should potentially reduce the bias (Samengo, 2002). Other methods to calculate mutual information that generally fall under this strategy are described in Borst and Theunissen, (1999) and in Victor (2006). The second strategy in dealing with the limited sampling bias is to directly estimate this bias, and then subtract it from the estimate of mutual information, or to compute mutual information using a method that attempts to directly correct for the bias . Several, often sophisticated methods have been devised that follow this strategy (Panzeri and Treves, 1996; Paninski, 2003; Nemenman et al., 2004), however in a detailed empirical analyses of these methods (Panzeri et al., 2007) found that the most effective bias correction method simply consisted of calculating the mutual information from randomly shuffling the relationship between stimuli and the responses, and then subtracting this shuffled bias estimate from the mutual information calculated from the real stimulus response data.

While mutual information can indeed offer much insight into questions of neural coding, there are a few potential issues with the method both in terms of practically applying the

procedure and in terms of the philosophical justification for using it. In terms of practically applying the method, one of the downsides is that estimating mutual information generally does not work well for analyzing large populations of neurons since the number of potential neural responses increases exponentially with the number of neurons used, making most estimation methods too biased to be useful. Additionally, the amount of information (i.e., *bits* of information) calculated using mutual information is highly dependent on the stimulus set used. While this criticism could apply to most data analyses methods, many researchers using mutual information would like to use the number of *bits* calculated to compare results between experiments; however due to the dependency on the stimulus ensemble used, such comparisons are questionable. Criticisms have also been raised on the philosophical underpinnings for justifying the use of mutual information. In particular, it is questionable whether one should view neurons as being analogous to information channels since information channels are merely passive mechanisms which data can flow through while neurons presumable are involved in neural computations. Indeed, by the information processing inequality, all the information that the whole brain contains must be present in the peripheral nervous system, and how much information is being retained as information is passed through a neuron seems to be of less interest in terms of generating a deep understanding of neural processing. Rather, what appears to be important is an understanding of how information is being lost in an intelligent manner in order to build complex and invariant representations that potentially allow an organism to thrive in the world, and mutual information analyses do not seem to readily lend themselves to assessing how invariant a neural representation is.

Neural decoding

In decoding based analyses of neural data, recordings of neural activity are used to predict the whether stimuli or other behaviorally relevant variables are present in the world. The rational behind this analysis is that it takes 'the organism's point of view' (Bialek et al., 1991; Rieke et al., 1999) in the sense that using a decoding algorithm to 'readout out' information from a brain area is similar to the task that a downstream neuron engages in when pooling information from an upstream area; thus the information

extracted from a population of neurons P by a decoding algorithm should theoretically be similar to the information available to a downstream neuron that has synaptic connections to this population P.

Many different decoding algorithms have been used for the classification and reconstruction/(regression) of stimuli (and other conditions presented to an animal). While there are several different ways to group these methods, one way they can be divided is along the lines of generative (probabilistic/Bayesian) and discriminative algorithms. Within the discriminative methods, many 'linear' algorithms have been used including: 1) linear filtering of a temporal spike train of the H1 visual neuron of the blowfly to predict a visual stimulus (Rieke et al., 1999); 2) linear combinations of cells' preferred movement directions in macaque motor cortex to create a 'population-vector' that can predict arm movements (Georgopoulos et al., 1986); 3) 'optimal linear filters' that minimize a squared loss function between any real valued stimulus and a reconstructed stimulus (Salinas and Abbott, 1994); 5) maximum correlation methods that categorize a stimulus based on the maximum correlation between a given neural response vector and the mean response vector from each condition to decode information about a rat's position from hippocampal place cells (Wilson and McNaughton, 1993); 6) nearest neighbor methods to categorize attention and intended movement directions from macaque motor cortex (Quiroga et al., 2006); and 7) linear support vector machines that minimize a hinge-loss function subject to regularization constraints that have been used to discriminate between visual images using data from IT (Hung et al., 2005). Also non-linear discriminative methods have been used including neuronal networks to decode motor movements from macaque motor cortex (Wessberg et al., 2000). A few generative/probabilistic/Bayesian methods that have been used include: 1) Possion Naïve Bayes classifiers (with and without additional continuity prior distributions) used to decoding a rat's position based on the firing rates of neurons in the hippocampus (Zhang et al., 1998) and 2) more sophisticated state space Bayesian/Kalman filtering algorithms that additionally model the dynamics of the decoding variances and include models of neuron's encoding properties to produce probability estimates of a rat's position or the movement of an monkeys arm (Brown et al., 1998; Wu et al., 2006).

Several papers compare the performance of different classifiers (Salinas and Abbott, 1994; Zhang et al., 1998; Brown et al., 1998; Schwartz et al., 2001), and point out reasons to prefer one type of decoding algorithm over another. Some advantages of using many of the discriminative methods are that the algorithms are generally simpler which often means that they run faster and that it is easier to apply them to many different types of data without having to substantially modify the algorithm. Also, it is often easier to interpret the results of a classifier in terms of how it could potentially relate to the functioning of downstream biological neurons (Additional supplemental material 2.1). Some advantages of using generative/Bayesian algorithms are that it is possible to incorporate prior information into the decoding procedure which often leads to higher decoding accuracies[8]. Also, many generative methods have a stronger mathematical foundation which makes it easier to assess how errors are related to the quality of the fit of a model, and it also often allows for computation of confidence intervals and other measures to assess the certainty one should have in the decoding results.

Overall, regardless of the decoding algorithm used, neural decoding (and neural population decoding in particular) have several advantages as a general method of analyzing neural data[9]. Some of these advantages are: 1) decoding allows a potentially biologically plausible way to evaluate the amount of information in a population (Zhang et al., 1998), as opposed to conventional statistics that often treat neurons as being identical samples from an underlying probability distribution; 2) population decoding can examine all the information in a population simultaneously often sidestepping some selection biases and combinatorial explosion effects that influence conventional statistics and mutual information measure respectively; 3) by training and testing a decoding algorithm under different conditions, decoding methods allow one to assess how abstract or invariant a neural representation is to particular changes in the stimulus or other

[8] While incorporating prior information is very useful for practical decoding tasks (particular when used in a brain machine interface that controls prosthetic devices based on neural activity), one needs to be careful that one is not relying too heavily on such models when inferring the function of a brain area, since the prior information in purely an invention of the creator of the decoding algorithm and thus the algorithm is not really assessing what information is directly available from the neural activity.

[9] Also see chapter 6 which illustrates some of these advantages using examples taken from this thesis.

behavioral conditions; 4) by exploring different representations of neural data (such as using different bin sizes or different neural signals), one can use decoding to assess how much information different types of signals contain and thus get an idea of how neurons code information; and 5) neural decoding methods can also be used to test for the reactivation of neural patterns that were evoked when particular stimuli were present, which could give insight into memory or other stimulus related processing that is occurring even when the stimuli are not present, such as during sleeping states (Wilson and McNaughton, 1993). Some disadvantages of decoding include: 1) the fact that the results could depend on the specific decoding algorithm, loss function, or data representation used, and that using a different decoding algorithms, loss functions or data representation could potentially yield different results[10] (Schneidman et al., 2003); and 2) the information that is being decoding might be used in a different way, or not at all, by the animal, and so it is possible that decoding results could be misleading.

References

Aggelopoulos N, Franco L, Rolls E. Object perception in natural scenes: Encoding by inferior temporal cortex simultaneously recorded neurons. J of Neurophysiol 93: 1342-1357, 2005.

Anderson B, Sanderson M, Sheinberg D. Joint decoding of visual stimuli by IT neurons' spike counts is not improved by simultaneous recording. Experimental Brain Research 176: 1-11, 2007.

Ashbridge E, Perrett D. Generalizing across object orientation and size. In: Perceptual constancy: why things look the way they do. New York: Cambridge University Press, 1998, p. 192-209.

Baeg E, Kim Y, Huh K, Mook-Jung I, Kim H, Jung M. Dynamics of population code for working memory in the prefrontal cortex. Neuron 40: 177-188, 2003.

Baker CI, Behrmann M, Olson CR. Impact of learning on representation of parts and

[10] Empirically, we have found the decoding results to be very robust to these choices (see chapter 6 for a discussion of this issue).

wholes in monkey inferotemporal cortex. Nat Neurosci 5: 1210-1216, 2002.

Baylis GC, Rolls ET, Leonard CM. Functional subdivisions of the temporal lobe neocortex. J. Neurosci 7: 330-342, 1987.

Baylis G, Rolls E. Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. Exp Brain Res 65, 1987.

Baylis GC, Driver J. Shape-coding in IT cells generalizes over contrast and mirror reversal, but not figure-ground reversal. Nat Neurosci 4: 937-942, 2001.

Bear M, Connors B, Paradiso M. Neuroscience: Exploring the Brain (Third Edition). Lippincott Williams & Wilkins.

Bialek W, Rieke F, de Ruyter van Steveninck RR, Warland D. Reading a neural code. Science 252: 1854-7, 1991.

von Bonin G, Bailey P. The Neocortex of Macaca mulatta. J Am Med Assoc 137: 1093-e, 1948.

Booth MC, Rolls ET. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cerebral Cortex 8: 510 -523, 1998.

Borst A, Theunissen FE. Information theory and neural coding. Nat Neurosci 2: 947-957, 1999.

Brincat SL, Connor CE. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nat Neurosci 7: 880-886, 2004.

Broome BM, Jayaraman V, Laurent G. Encoding and Decoding of Overlapping Odor Sequences. Neuron 51: 467-482, 2006.

Brown EN, Frank LM, Tang D, Quirk MC, Wilson MA. A Statistical Paradigm for Neural Spike Train Decoding Applied to Position Prediction from Ensemble Firing Patterns of Rat Hippocampal Place Cells. J. Neurosci. 18: 7411-7425, 1998.

Bruce C, Desimone R, Gross CG. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J. Neurophysiol 46: 369-384, 1981.

Cox DD, Meier P, Oertelt N, DiCarlo JJ. 'Breaking' position-invariant object recognition. Nat Neurosci 8: 1145-1147, 2005.

Dan Y, Alonso J, Usrey WM, Reid RC. Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus. Nat Neurosci 1: 501-507, 1998.

Dayan P, Abbott L. Theoretical neuroscience : computational and mathematical modeling

of neural systems. Massachusetts Institute of Technology Press.

De Baene W, Ons B, Wagemans J, Vogels R. Effects of Category Learning on the Stimulus Selectivity of Macaque Inferior Temporal Neurons. Learning & Memory 15: 717-727, 2008.

Desimone R, Albright T, Gross C, Bruce C. Stimulus-selective properties of inferior temporal neurons in the macaque. J. Neurosci. 4: 2051-2062, 1984.

Desimone R. Face-Selective Cells in the Temporal Cortex of Monkeys. Journal of Cognitive Neuroscience 3: 1-8, 1991.

DiCarlo JJ, Maunsell JHR. Anterior Inferotemporal Neurons of Monkeys Engaged in Object Recognition Can be Highly Sensitive to Object Retinal Position. J Neurophysiol 89: 3264-3278, 2003.

Erickson CA, Desimone R. Responses of Macaque Perirhinal Neurons during and after Visual Stimulus Association Learning. J. Neurosci. 19: 10404-10416, 1999.

Felleman D, Van Essen D. Distributed Hierarchical Processing in the Primate Cerebral Cortex. Cereb. Cortex 1: 1-a-47, 1991.

Földiák P. Learning invariance from transformation sequences. Neural Comput. 3: 194-200, 1991.

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

Freedman DJ, Assad JA. Experience-dependent representation of visual categories in parietal cortex. Nature 443: 85-88, 2006.

Fujita I, Tanaka K, Ito M, Cheng K. Columns for visual features of objects in monkey inferotemporal cortex. Nature 360: 343-346, 1992.

Fujita I. The inferior temporal cortex: Architecture, computation, and representation. Journal of Neurocytology 31: 359-371, 2002.

Fuster J, Jervey J. Neuronal firing in the inferotemporal cortex of the monkey in a visual memory task. J. Neurosci. 2: 361-375, 1982.

Gawne TJ, Kjaer TW, Hertz JA, Richmond BJ. Adjacent visual cortical complex cells share about 20% of their stimulus-related information. Cereb Cortex 6: 482-9, 1996.

Gawne T, Richmond B. How independent are the messages carried by adjacent inferior temporal cortical neurons? J. Neurosci. 13: 2758-2771, 1993.

Georgopoulos A, Schwartz A, Kettner R. Neuronal population coding of movement direction. Science 233: 1416-1419, 1986.

Gerstner S, Kistler W. Spiking neuron models: single neurons, populations, plasticity. Cambridge, U.K. ; New York:: Cambridge University Press, 2002.

Girard P, Jouffrais C, Kirchner C. Ultra-rapid categorisation in non-human primates. Animal Cognition 11: 727, 2008.

Gochin P, Colombo M, Dorfman G, Gerstein G, Gross C. Neural Ensemble Coding in Inferior Temporal Cortex. J of Neurophysiol 71: 2325-2337, 1994.

Golledge HDR, Panzeri S, Zheng F, Pola G, Scannell JW, Giannikopoulos DV, Mason RJ, Tovée MJ, Young MP. Correlations, feature-binding and population coding in primary visual cortex. Neuroreport 14: 1045-50, 2003.

Goodale MA, Milner AD. Separate visual pathways for perception and action. Trends Neurosci 15: 20-25, 1992.

Gross CG, Rocha-Miranda CE, Bender DB. Visual properties of neurons in inferotemporal cortex of the Macaque. J. Neurophysiol 35: 96-111, 1972.

Gross CG. How Inferior Temporal Cortex Became a Visual Area. Cerebral Cortex 4: 455 -469, 1994.

Hartline HK. The response of single optic nerve fibers of the vertebrate eye to illumination
of the retina. American Journal of Physiology 121: 400-415, 1938.

Heller J, Hertz JA, Kjær TW, Richmond BJ. Information flow and temporal coding in primate pattern vision. Journal of Computational Neuroscience 2: 175-193, 1995.

Higuchi S, Miyashita Y. Formation of mnemonic neuronal responses to visual paired associates in inferotemporal cortex is impaired by perirhinal and entorhinal lesions. Proceedings of the National Academy of Sciences of the United States of America 93: 739 -743, 1996.

Hubel DH, Wiesel TN. Receptive fields of single neurones in the cat's striate cortex. J. Physiol. (Lond.) 148: 574-591, 1959.

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

Ito M, Tamura H, Fujita I, Tanaka K. Size and position invariance of neuronal responses in monkey inferotemporal cortex. J Neurophysiol 73: 218-226, 1995.

Kayaert G, Biederman I, Vogels R. Representation of Regular and Irregular Shapes in Macaque Inferotemporal Cortex. Cerebral Cortex 15: 1308 -1321, 2005.

Kirchner H, Thorpe SJ. Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. Vision Res 46: 1762-1776, 2006.

Kobatake E, Tanaka K. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J Neurophysiol 71: 856-867, 1994.

Kobatake E, Wang G, Tanaka K. Effects of Shape-Discrimination Training on the Selectivity of Inferotemporal Cells in Adult Monkeys. J Neurophysiol 80: 324-330, 1998.

Latham PE, Nirenberg S. Synergy, Redundancy, and Independence in Population Codes, Revisited. J. Neurosci. 25: 5195-5206, 2005.

Laurent G. Olfactory network dynamics and the coding of multidimensional signals. Nature Reviews Neuroscience 3: 884-895, 2002.

Lettvin J, Maturana H, Mcculloch W, Pitts W. What the Frog's Eye Tells the Frog's Brain. Proceedings of the IRE 47: 1940-1951, 1959.

Li L, Miller EK, Desimone R. The representation of stimulus familiarity in anterior inferior temporal cortex. J Neurophysiol 69: 1918-1929, 1993.

Li N, DiCarlo JJ. Unsupervised Natural Experience Rapidly Alters Invariant Object Representation in Visual Cortex. Science 321: 1502-1507, 2008.

Logothetis NK, Pauls J, Poggio T. Shape representation in the inferior temporal cortex of monkeys. Curr. Biol 5: 552-563, 1995.

Logothetis NK, Sheinberg DL. Visual object recognition. Annu. Rev. Neurosci 19: 577-621, 1996.

Lueschow A, Miller EK, Desimone R. Inferior Temporal Mechanisms for Invariant Object Recognition. Cerebral Cortex 4: 523 -531, 1994.

Matsumoto N, Saunders R, Gothard B, Richmond B. Retention of perceptual categorization following bilater removal of area TE in rhesus maonkeys. Cosyne. Salt Lake City, UT, USA: 2010.

Messinger A, Squire LR, Zola SM, Albright TD. Neuronal representations of stimulus associations develop in the temporal lobe during learning. Proceedings of the National Academy of Sciences of the United States of America 98: 12239 -12244, 2001.

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: 1407-19, 2008.

Miller EK, Li L, Desimone R. A neural mechanism for working and recognition memory in inferior temporal cortex. Science 254: 1377-1379, 1991.

Miller E, Desimone R. Parallel neuronal mechanisms for short-term memory. Science 263: 520-522, 1994.

Miller E, Li L, Desimone R. Activity of neurons in anterior inferior temporal cortex during a short- term memory task. J. Neurosci. 13: 1460-1478, 1993.

Mishkin M. A memory system in the monkey. Philos. Trans. R. Soc. Lond., B, Biol. Sci 298: 83-95, 1982.

Mishkin M, Ungerleider LG. Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. Behav. Brain Res 6: 57-77, 1982.

Mishkin M, Ungerleider LG, Macko KA. Object vision and spatial vision: two cortical pathways. Trends in Neurosciences 6: 414-417, 1983.

Miyashita Y. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature 335: 817-820, 1988.

Miyashita Y. Inferior temporal cortex: where visual perception meets memory. Annu. Rev. Neurosci 16: 245-263, 1993.

Miyashita Y, Chang HS. Neuronal correlate of pictorial short-term memory in the primate temporal cortex. Nature 331: 68-70, 1988.

Moeller S, Freiwald WA, Tsao DY. Patches with Links: A Unified System for Processing Faces in the Macaque Temporal Lobe. Science 320: 1355-1359, 2008.

Mogami T, Tanaka K. Reward association affects neuronal responses to visual stimuli in macaque TE and perirhinal cortices. Journal of Neuroscience 26: 6761-6770, 2006.

Naya Y, Yoshida M, Miyashita Y. Backward Spreading of Memory-Retrieval Signal in the Primate Temporal Cortex. Science 291: 661-664, 2001.

Nemenman I, Bialek W, de Ruyter van Steveninck R. Entropy and information in neural spike trains: Progress on the sampling problem. Phys. Rev. E 69: 056111, 2004.

Nieder A, Freedman D, Miller E. Representation of the quantity of visual items in the primate prefrontal cortex. Science 297: 1708-1711, 2002.

Nirenberg S, Carcieri SM, Jacobs AL, Latham PE. Retinal ganglion cells act largely as independent encoders. Nature 411: 698-701, 2001.

Nowak L, Bullier J. The timing of information transfer in the visual system. In: Cerebral Cortex, edited by Kaas JH, Rockland K, Peters A. New York: Plenum, 1998, p. 205-241.

Op de Beeck H, Wagemans J, Vogels R. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. Nat Neurosci 4: 1244-1252, 2001.

Op de Beeck HP, Deutsch JA, Vanduffel W, Kanwisher NG, DiCarlo JJ. A Stable Topography of Selectivity for Unfamiliar Shape Classes in Monkey Inferior Temporal Cortex. Cerebral Cortex 18: 1676 -1694, 2008.

Optican LM, Richmond BJ. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. J Neurophysiol 57: 162-178, 1987.

Paninski L. Estimation of entropy and mutual information. Neural Computation 15: 1191-1253, 2003.

Panzeri S, Senatore R, Montemurro MA, Petersen RS. Correcting for the sampling bias problem in spike train information measures. J Neurophysiol (July 5, 2007). doi: 10.1152/jn.00559.2007.

Panzeri S, Treves A. Analytical estimates of limited sampling biases in different information measures. Network: Computation in Neural Systems 7: 107, 87, 1996.

Perrett D, Rolls E, Caan W. Visual neurones responsive to faces in the monkey temporal cortex. Exp Brain Res 47, 1982.

Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, Chichilnisky EJ, Simoncelli EP. Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature 454: 995-999, 2008.

Quiroga R, Snyder L, Batista A, Cui H, Andersen R. Movement intention is better predicted than attention in the posterior parietal cortex. Journal of Neuroscience 26: 3615-3620, 2006.

Richmond BJ, Optican LM. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. II. Quantification of response waveform. J Neurophysiol 57: 147-161, 1987.

Richmond BJ, Optican LM, Podell M, Spitzer H. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. I. Response characteristics. J Neurophysiol 57: 132-146, 1987.

Rieke F, Warland D, Steveninck RDRV, Bialek W. Spikes: exploring the neural code. MIT Press.

Rollenhagen JE, Olson CR. Mirror-Image Confusion in Single Neurons of the Macaque Inferotemporal Cortex. Science 287: 1506-1508, 2000.

Rolls ET. Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. Neuron 27: 205-218, 2000.

Rolls E, Tovee M. Sparseness of the Neuronal Representation of Stimuli in the Primate Temporal Visual-Cortex. Journal of Neurophysiology 73: 713-726, 1995.

Sakai K, Miyashita Y. Neural organization for the long-term memory of paired associates. Nature 354: 152-155, 1991.

Salinas E, Abbott LF. Vector reconstruction from firing rates. J Comput Neurosci 1: 89-107, 1994.

Samengo I. Information loss in an optimal maximum likelihood decoding. Neural Computation 14: 771-779, 2002.

Sary G, Vogels R, Orban G. Cue-invariant shape selectivity of macaque inferior temporal neurons. Science 260: 995-997, 1993.

Sawamura H, Orban GA, Vogels R. Selectivity of neuronal adaptation does not match response selectivity: a single-cell study of the FMRI adaptation paradigm. Neuron 49: 307-318, 2006.

Schneidman E, Bialek W, Berry MJ. Synergy, Redundancy, and Independence in Population Codes. J. Neurosci. 23: 11539-11553, 2003.

Schwartz AB, Taylor DM, Tillery SI. Extraction algorithms for cortical control of arm prosthetics. Curr. Opin. Neurobiol 11: 701-707, 2001.

Schwartz EL, Desimone R, Albright TD, Gross CG. Shape recognition and inferior temporal neurons. Proceedings of the National Academy of Sciences of the United States of America 80: 5776 -5778, 1983.

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.

Shannon C. A mathematical theory of communication. Bell system technical journal 27, 1948.

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

Tanaka J. Object categorization, expertise and neural plasticity. In: The New Cognitive Neurosciences. Cambridge, MA: MIT Press, 2004, p. 876-888.

Tanaka K. Inferotemporal cortex and object vision. Annu. Rev. Neurosci 19: 109-139, 1996.

Tanaka K, Saito H, Fukada Y, Moriya M. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. J Neurophysiol 66: 170-189, 1991.

Tanaka K. Mechanisms of visual object recognition: monkey and human studies. Current Opinion in Neurobiology 7: 523-529, 1997.

Tanaka K. Columns for Complex Visual Object Features in the Inferotemporal Cortex: Clustering of Cells with Similar but Slightly Different Stimulus Selectivities. Cerebral Cortex 13: 90 -99, 2003.

Theunissen F, Miller JP. Temporal encoding in nervous systems: A rigorous definition. Journal of Computational Neuroscience 2: 149-162, 1995.

Tovee MJ, Rolls ET, Treves A, Bellis RP. Information encoding and the responses of single neurons in the primate temporal visual cortex. J Neurophysiol 70: 640-654, 1993.

Tsunoda K, Yamane Y, Nishizaki M, Tanifuji M. Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. Nat Neurosci 4: 832-838, 2001.

Victor JD, Purpura KP. Nature and precision of temporal coding in visual cortex: a metric-space analysis. J Neurophysiol 76: 1310-1326, 1996.

Victor JD. Approaches to Information-Theoretic Analysis of Neural Activity. Biological Theory 1: 302-316, 2006.

Vinje WE, Gallant JL. Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. Science 287: 1273-1276, 2000.

Wallis G, Rolls ET. Invariant face and object recognition in the visual system. Prog. Neurobiol 51: 167-194, 1997.

Wallis G, Bülthoff HH. Effects of temporal association on recognition memory. Proceedings of the National Academy of Sciences of the United States of America 98: 4800 -4804, 2001.

Wang G, Tanifuji M, Tanaka K. Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. Neuroscience Research 32: 33-46, 1998.

Wessberg J, Stambaugh CR, Kralik JD, Beck PD, Laubach M, Chapin JK, Kim J, Biggs SJ, Srinivasan MA, Nicolelis MAL. Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. Nature 408: 361-365, 2000.

Wilson M, McNaughton B. Dynamics of the hippocampal ensemble code for space. Science 261: 1055-1058, 1993.

Wiskott L, Sejnowski TJ. Slow feature analysis: unsupervised learning of invariances. Neural Comput 14: 715-770, 2002.

Woloszyn L, Sheinberg DL. Neural Dynamics in Inferior Temporal Cortex during a Visual Working Memory Task. J. Neurosci. 29: 5494-5507, 2009.

Wu W, Gao Y, Bienenstock E, Donoghue JP, Black MJ. Bayesian Population Decoding of Motor Cortical Activity Using a Kalman Filter. Neural Computation 18: 80-118, 2006.

Zaksas D, Pasternak T. Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. Journal of Neuroscience 26: 11726-11742, 2006.

Zhang K, Ginzburg I, McNaughton BL, Sejnowski TJ. Interpreting Neuronal Population Activity by Reconstruction: Unified Framework With Application to Hippocampal Place Cells. J Neurophysiol 79: 1017-1044, 1998.

Zoccolan D, Kouh M, Poggio T, DiCarlo JJ. Trade-Off between Object Selectivity and Tolerance in Monkey Inferotemporal Cortex. J. Neurosci. 27: 12292-12307, 2007.

Chapter 2: Decoding neural data from high-level vision experiments: From Experimental Design to Interpreting Decoding Results

The following material is currently being published as a book chapter:

Meyers, E., and Kreiman, G. "From Neural Recordings to Interpreting Decoding Accuracy". In: Understanding visual population codes. Kreigeskorte, N., and Kreiman, G. (eds.), 2010, MIT Press.

Abstract

In this chapter we outline a procedure to decode information from multivariate neural data. We assume that neural recordings have been made from a number of trials in which different conditions were present, and our procedure produces and estimate of how accurately we can predict the labels of these conditions in a new set of data. We call this estimate of future prediction the 'decoding/readout accuracy,' and based on this measure we can make inferences about what information is present in the population of neurons and also on how this information is coded. The steps we cover to obtain a measure of decoding accuracy include: 1) designing an experiment, 2) formatting the neural data, 3) selecting a classifier to use, 4) applying cross-validation to random splits of the data, 5) evaluating decoding performance through different measures, and 6) testing the integrity of the decoding procedure and significance of the results. We also discuss additional topics including how to examine questions about neural coding and how to evaluate whether the population is representing stimuli in an invariant/abstract way.

Introduction

In this chapter we describe a procedure to decode information from multivariate neural data. The procedure is derived from cross-validation methods that are commonly used by researchers in machine learning (ML) to compare the performance of different classification algorithms. However, instead of comparing different ML algorithms, here we assess how accurately a particular algorithm can extract information about different experimental conditions in order to better understand how the brain processes information. These procedures and algorithms are extensively used to quantitatively examine the responses of populations of neurons at the neurophysiological level.

Our motivation for using the procedure described in this chapter is based on a simple intuition for what we believe is an important computational function that the brain must perform – namely, to *reliably* distinguish between different behaviorally relevant conditions that are present in the world in single trials. Cross-validation is an excellent measure for assessing such reliability. If we can build a model (classifier) for how neurons can distinguish between different conditions using only part of the data, and show that the same model works for distinguishing between these same conditions in a new set of data, then this gives us a significant degree of confidence that the current neural activity can reliably distinguish between these conditions, and that our model is capturing the reliability in the data. Additionally, we can compare different models, and if one model is able to extract a more reliable signal from the neural data than another model, this can give us insight into how information is coded in the data. Finally, by building a model to distinguish between one set of conditions and then seeing that the same model can generalize to a different but related set of conditions, we can infer that the brain contains information in a way that is invariant to the exact conditions that were used to build the model. Since all information entering the brain is already present in the sensory nerves and early processing areas, assessing how the brain selectively loses information in order to create behaviorally-relevant invariant representations is important for understanding the functional role of higher level brain regions.

To put things in the terminology used by the machine learning and computational neuroscience communities, we call the processes of building a model on a subset of data 'training the classifier' or 'learning', and we call the process of assessing if the model (classifier) still works on a new set of data 'testing the classifier'. The 'decoding accuracy' (also referred to as 'classification accuracy' or 'readout accuracy') is a measure of how well the classifier performs on the new 'test set' of data used to test the classifier's performance. As mentioned above, a high degree of decoding accuracy indicates that the model is capturing reliable differences between different conditions.

The following chapter is a nuts-and-bolts description of how to implement a cross-validation classification scheme that we have found works well for the analysis of multivariate neural data. The methods have been developed by analyzing real neural data and assessing what empirically works the best. While we have had experience analyzing several different datasets, there is still much more work to be done to fully characterize the best methods to use. Thus the chapter below constitutes work in progress explaining the best methods we have found so far.

Experimental design

Our discussion centers on a hypothetical experiment where a subject (human or animal) is presented with different images while the investigators record the activity of multiple neurons from implanted electrodes (e.g., see Figure 2.1A). The images belong to different "conditions". These conditions could refer to different object identities, different object categories, different object positions or viewpoints, same objects under different experimental manipulations (e.g. attention / no attention), etc. In order for population decoding methods to work properly, it is important that the experimental design follows a few basic guidelines. First, multiple trials of each condition type must be presented to the subject. For example, if the investigator is interested in decoding which particular stimulus was shown to the subject, then each stimulus must be presented multiple times. While in general the more data the better, there are often experimental restrictions (e.g. it

may be difficult to hold a stable recording for prolonged periods of time). We have found that in certain cases as few as five repetitions of each experimental condition are enough to give interpretable results (Meyers et al., 2008), although higher decoding accuracies are usually obtained with more repetitions.

Second, it is important that the stimuli are presented in random order. If the stimuli are not presented in random order (e.g., if all trials of condition 1 are presented before all trials of condition 2, etc.), then even if there is no reliable information about the stimuli in the data, above chance decoding accuracies could still be spuriously obtained due to nonstationarities in the recording or experimental procedure (e.g. due to electrode drift, varying attentional engagement in the task, adaptation, etc.; see the section on "testing the integrity of decoding" below for more details).

Finally, we note that it is not strictly necessary that the recordings from the population are made simultaneously. If the same experiment is repeated multiple times with single neurons being recorded each time, a 'pseudo-population' of responses can be constructed from piecing together the same trial type from multiple sessions (see the section on "formatting the neural data" for information on how to create pseudo-populations). Due to the experimental challenges in simultaneously recording from multiple electrodes, this approach is common in the neurophysiology community. The pseudo-population approach, by construction, assumes that the activity of the different neurons is independent, that is, time-varying correlations among neurons are ignored. While results from such pseudo-populations could potentially distort the estimate of amount of information decoding from the population (Averbeck et al., 2006; Averbeck and Lee, 2006), we have seen that much insight can still be gained from this type of analysis (for example see Hung et al., 2005; Meyers et al., 2008). Additionally, using pseudo-populations allows for population decoding to be applied to many experiments where it is currently not easy to record from populations of neurons (such as from deep brain structures like ventral macaque IT), and it allows for a population decoding reanalysis of older experiments in which simultaneous recordings were not made but for which the

same experiment was run for each neuron that was recorded (e.g., see Meyers et al., 2008).

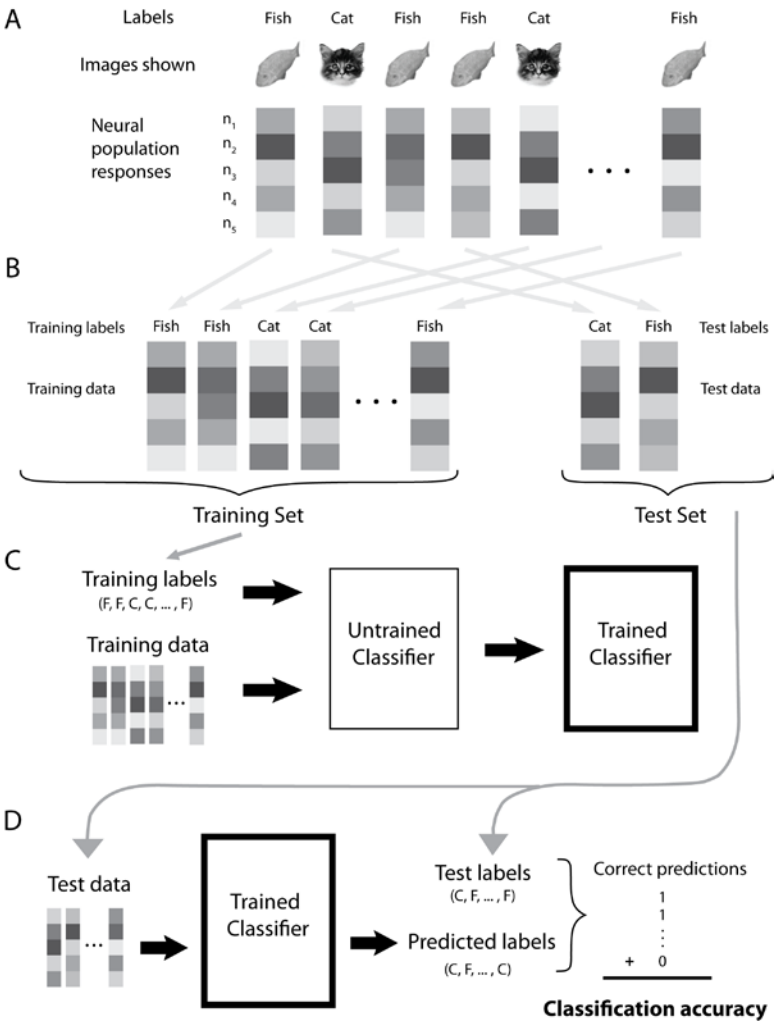


Figure 2.1 Basic steps involved in training and testing a classifier. A: An illustration of an experiment in which an image of a cat and an image of a fish were shown in random order to a subject while simultaneous recordings were made from five neurons/channels. The grayscale level denotes the activity of each neuron/channel. B: Data points and the corresponding labels are randomly selected to be in either the training set or in the test set. C: The training data points and the training labels are passed to an untrained classifier that 'learns' which neural activity is useful at predicting which image was shown – thus becoming a 'trained' classifier. D: The test data are passed to the trained classifier which produces predictions of which labels correspond to each unlabled test data point. These predicted labels are then compared to the real test labels (i.e., the real labels that were presented when the test data were recorded) and the percent of correct predictions is calculated to give the total classification accuracy.

Formatting neural data

Analyzing neural spiking data

The first step in applying population decoding to neural spiking data is to make single unit (SU) or multi-unit (MU) extra-cellular recordings. In some cases, investigators record multi-unit activity and they are interested in considering the single units that constitute those MU. There are several spike sorting algorithms for this purpose (e.g. (Fee et al., 1996; Lewicki, 1998; Wehr et al., 1999; Harris et al., 2000; Quiroga et al., 2004). Here we assume that spike extraction (and spike sorting) have already been performed in the data and we consider a binary sequence of 0's and 1's, with the 1's indicating the occurrence of spikes. The algorithms apply equally to a series of spikes from SU or MU.

It is useful to look at the average firing rate on each trial as a function of trial number for each neuron/site separately. If different conditions have been presented randomly to the animal, then there should not be any obvious temporal trend in firing rate as a function of trial number. However, there are many types of non-stationarities that could lead to trends over time (including changes in the quality of the recordings, subject fatigue, attentional changes over times, neuronal adaptation or plasticity over the course of the recordings, etc). These time-dependent trends could subsequently be confounded with the questions of interest in the absence of good trial randomization. Eliminating neurons that appear to have non-stationary responses can lead to improvements in decoding accuracy (although in practice so far we have found the improvements due to eliminating neurons with trends to be small). An automatic method that we have used to eliminate neurons that have temporal trends is to compute the average variance found in sliding blocks of 20 trials, and compare it to the overall variance among all trials. We typically eliminate all neurons for which the variance over all trials is twice as large as the average variance in 20-trial blocks.

Once neurons with trends have been removed, the next step we usually take is to bin the data. While decoding algorithms exist that use exact spike times without binning the data

(Truccolo et al., 2005), most of the common machine learning algorithms we use achieve a higher decoding accuracy when using firing rates computed over time intervals of tens to hundreds of milliseconds. The best bin size to use depends on several factors related to the types of questions of interest. For example, the degree of temporal precision in the neural code can be quantitatively evaluated by using small bins that obviously give more precise temporal accuracy, at the potential cost of having more noisy results. Conversely, if the condition that one is trying to decode seems weak, then we have found binning over larger intervals often reduces noise and leads to more robust results (see Meyers et al 2008, Hung et al 2005).

Apart from bin size, it is also of interest to consider the type of filter used to bin the data. In our work we typically have used square (boxcar) filters. The advantage of using these filters is that they provide exact boundaries in terms of the latencies of spikes that contribute to the results, and thus which time bin results are independent from other time bins[11]. Other researchers (Nikolic et al., 2007) have used exponential filters with short (20ms) time constants, in order to mimic what is believed to be the synaptic integration time of neurons, thus creating a potentially more biologically realistic model of the information available to downstream neurons.

Pseudo-populations

In many situations it is not currently practical or possible to record simultaneously from many neurons (for example, it is currently difficult to implant multi-electrode arrays in deep brain structures such as macaque inferior temporal cortex). Additionally, one might want to reanalyze older data that were not recorded simultaneously using population decoding, without having to redo the entire experiment using simultaneous recordings. In such cases, applying population decoding to 'pseudo-populations' of neurons can give some insight into population coding questions.

[11] We found this to be particularly useful when exploring how the neural code changes with time (i.e.,(Meyers et al., 2008), since there it was important to know which time periods were independent from each other.

We define a 'pseudo-population' of neurons as a population of neurons that was not recorded simultaneously but is treated as if it were[12]. To create pseudo-populations, one concatenates together responses from different neurons that were recorded when the same condition (stimulus) was presented into a 'population' response – although in fact these neurons were recorded from different experimental sessions (see Figure 2.2). We usually create these pseudo-population response vectors inside of a cross-validation procedure, and recalculate them each time we divide the cross-validation data into blocks (see section on cross-validation for more details). It should be noted that when creating pseudo-populations, all 'noise-correlations' within the data are destroyed, and the overall estimate of the amount of information in a population could be over or under estimated (Averbeck et al., 2006; Averbeck and Lee, 2006). However, at the moment it remains unclear whether such noise-correlations are important for information transmission[13] (and there is evidence that in many cases they do not matter, e.g., (Panzeri et al., 2003; Averbeck and Lee, 2004; Aggelopoulos et al., 2005; Anderson et al., 2007). Additionally, at least in principle, we might expect that in many circumstances this bias should affect all conditions equally, which would leave most conclusions drawn from experiments on pseudo-populations unchanged. Still, until more evidence is accumulated about the influence of noise-correlations, it is important to keep in mind that it is possible that population decoding results based on pseudo-populations could differ from results obtained using simultaneously recorded neurons.

---

[12] 'Pseudo-populations' have been used by several different researchers to analyze their data including Georgopoulos et al., 1983; Gochin et al., 1994; Rolls et al., 1997; Hung et al., 2005; Meyers et al., 2008 and these 'populations' are often referred to by different names including 'pseudoresponse vectors' (Gochin et al., 1994), and 'pseudosimultaneous population response vectors' (Rolls et al., 1997). Additionally, the process of recording over separate sessions to create pseudo-populations has been referred to as 'the sequential method' and the process of recording many neurons at once for the purposes of population decoding has been called the 'simultaenous method' (Tanila and Shaprio, 1998).

[13] If noise-correlations do not matter (i.e., if the activity of each neuron is statistically independent of the activity of other neurons given the current stimulus or behavioral event being represented), then a brain region is said to use a 'population code' (see Chapter 3). If interactions between neurons do code additional information then a brain region is said to use an 'ensemble code' (Hatsopoulos et al., 1998). Whether population codes or ensemble codes are used by the brain still remains and open question in neuroscience.

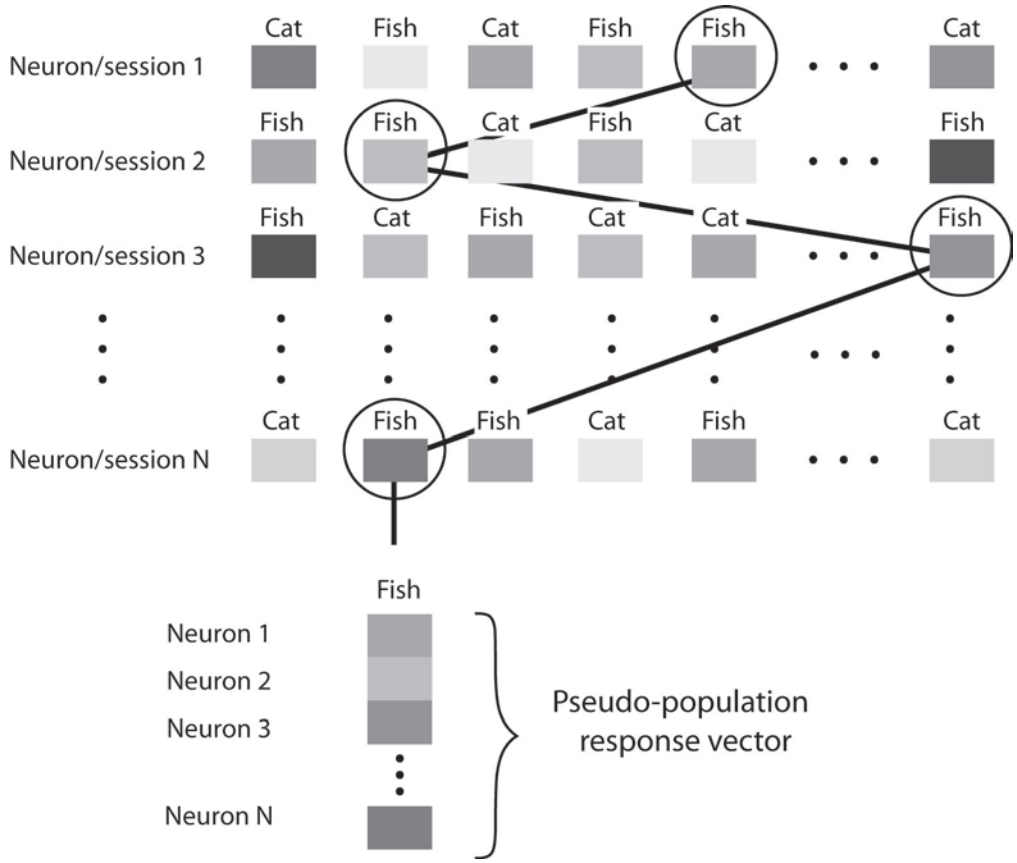


Figure 2.2 Creating pseudo-populations from data that was not recorded simultaneously. The above figure illustrates a set of experiments in which in each session the responses from a single neuron were recorded (the responses of each neuron is in each row with the darkness indicating the firing rate on particular trials). After recordings have been made from N neurons, a pseudo-population vector can be created by randomly choosing neuron responses from trials in which the same image was shown (circled neuron responses) and concatenating them together to create vector. This pseudo-population vector will be treated as if all the responses had been recorded simultaneously in subsequent population decoding analyses.

Selecting a classifier

A classifier is an algorithm that 'learns' a model from a 'training set' of data (that consists of a list of population neural responses and the experimental conditions that elicited these responses) and then makes predictions as to what experimental conditions a

new 'test' data was recorded under based on the model that was learned.[14] For example, the training data could represent neural populations responses to different images and a list of the image names that elicited these population responses, and the test data could consist of population responses to a different set of trials in which the same images were shown. The classifier would then have to predict the names of the images for each of the neural population responses in the test set. The model that was learned in this example could be a list of which neurons had high firing rates to particular images, and the classifier would make its predictions by combining the information in this list with the actual firing rates observed in the test set (see Figure 2.1).

Many different classifiers exist, although we have found that unlike in the analysis of fMRI data where using regularized classifiers greatly improves decoding performance, decoding results based on neural spiking data seems to be less sensitive to which exact classifier one uses. Empirically, we have found that we almost always achieve approximately the same level of performance using linear and non-linear support vector machines (SVMs), linear and non-linear regularized least squares (RLS), Possion Naïve Bayes classifiers (PNB), Gaussian Naïve Bayes classifiers (GNB), and a simple classifier based on taking the maximum correlation between the mean of training points for each class (MCC), (see Figure 2.3). The only classifier that consistently yielded worse results was the Nearest Neighbor classifier (NN). Since the MCC classifier has the fastest run time, and is the simplest to implement and understand, we recommend using this classifier when initially running experiments. However, since we do not have a deep

[14] To be slightly more formal, a training set consists of a pair of values (X, y), where X is an ordered set of neural population response vectors, and y is an ordered set of labels indicating the conditions that the neural responses were recorded under (with X_i being the neural population response to the *ith* training trial, and y_i indicating which conditions/stimulus was shown on that trial). 'Learning' consists of applying a function $f(X, y) \rightarrow M$ that takes the training neural data and the training labels and returns a set of model parameters M. This model can then be used by another 'inference' function $g(\hat{X}, M) \rightarrow \tilde{y}$, that takes a new set of test data $\hat{X}$, and produces a prediction $\tilde{y}$ of which labels/conditions correspond to each test point $\hat{X}_i$. The predicted $\tilde{y}$ can be compared to the real test labels $\hat{y}$ to evaluate decoding accuracy. Typically, the function g is called the 'classifier', although the learning algorithm f could also be considered part of the classifier as well. Also, it is common to write the learning function f as returning the inference algorithm g (that is, $f(X, y) \rightarrow g$.

theoretical reason why all these classifiers seem to be working equally well[15], we also recommend testing on a few different classifiers, since it is possible that better performance could be achieved on certain datasets, particularly if there are many training examples available.

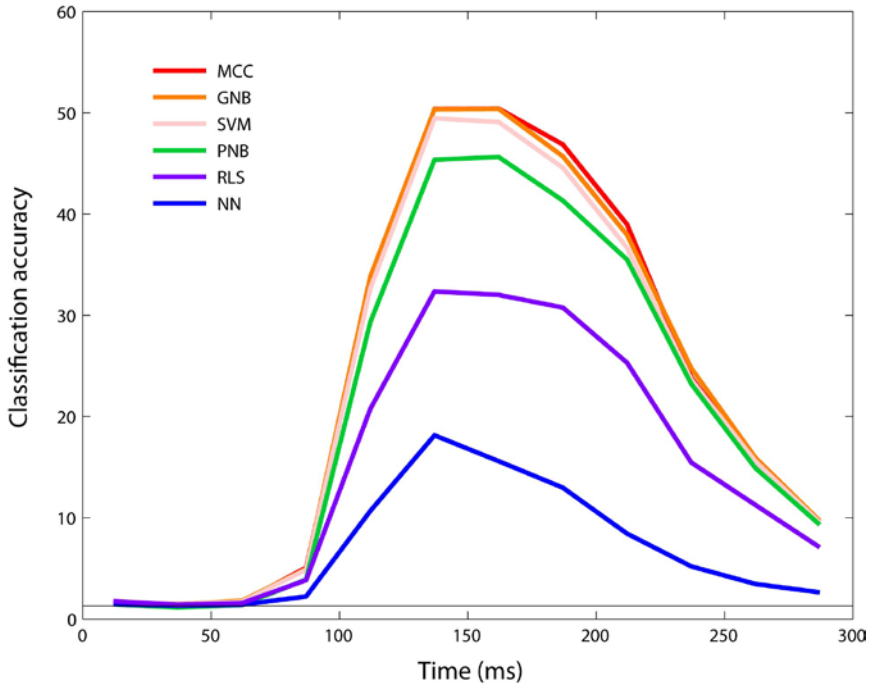


Figure 2.3 A comparison of different classifiers. The classifiers used are: a maximum correlation coefficient classifier (MCC), a Gaussian Naïve Bayes classifier (GNB), a linear support vector machines (SVM), a Possion Naïve Bayes classifiers (PNB), a linear regularized least squares (RLS), and a Nearest Neighbor classifier (NN). While the best results here were achieved with the MCC, GNB, and SVM, the over ordinal increases and decreases in decoding accuracy is similar across classifiers – thus similar conclusions would be drawn regardless of which classifier was used (although the power to distinguish between subtle differences in conditions is enhanced when better classifiers are used). The results in this figure are based on decoding which of 77 objects was shown to a macaque monkey using mean firing rates in 25ms successive bins (see Hung et al., 2005).

[15] It should be noted that in general on most classification tasks (such as on fMRI data, and on computer vision features), more complex classifiers such as SVMs and RLS tend to work better than simple ones such as the MCC. The fact that such simple classifiers work well suggests that there is something particular about neural spiking data that is well fit by this simple model.

Cross-validation

Cross-validation is the process of selecting a subset of data to train a classifier on and then using a different subset of data to test the classifier, and it forms one of the most significant components of offline neural population decoding schemes. Typically, cross-validation involves splitting the data into *k* parts, with each part consisting of *j* data points. Training of the classifier is done on *k - 1* sections of the data, and testing is done on the remaining section. The process is usually repeated *k* times, each time leaving out a different subset of data and testing on the remaining pieces. Classification accuracy is typically reported as the average percent correct over all *k* splits of the data.

When implementing a cross-validation scheme, it is critically important that *there is no overlapping data between the training set and the test set*, and that *the condition labels that belong to the test set are only used to verify the decoding performance, and that they are not used at any other point in the data processing stream*. Any violation of these conditions can lead to spurious results. Thus we recommend doing several sanity checks to insure that the cross-validation scheme has been implemented correctly (see the section on testing the integrity of the decoding procedure for more details).

When applying a cross-validation scheme to neural data, we typically use the following procedure. First, if the experimental data from different (stimulus) conditions have been repeated different numbers of times, we first calculate the number of repetitions present for the condition that has the fewest number of repeated trials;[16] for the purpose of this

[16] In most properly designed decoding experiments, different conditions are presented in a random order, and since the ability to record from a neuron often ends at a random point in time within an experimental session, it is fairly common to have a different number of stimulus presentations for different conditions (particularly when doing decoding on pseudo-populations). Since having different numbers of training examples for different conditions can bias certain types of classifiers into choosing the condition with the most training examples, we make sure that there is an equal number of training examples in each condition. Of course if there is reason to believe that in the there would be more of one condition in the world than another condition, then it could be reasonable to have this bias in the classifier (i.e., this bias could be a reasonable approximation for the a priori distribution of the conditions/stimuli in the world). Chance performance in this unbalanced training case then becomes the proportion of training points in the class with the most training points (i.e., the chance level is the expected proportion of correct responses if the classifier always selected the class with the highest a priori probability mass).

discussion, let *q* be a number that is equal to or less than the number of trial repetitions for the condition that has minimum number of repetitions, and let *k* be a number that is divides *q* evenly (i.e., $q = k * j$, where *q*, *k* and *j* are all integers). We then randomly select (pseudo-) population responses for *q* trials for each condition, and put these *q* repetitions into *k* groups, with each group having *j* population responses to each of the condition (if pseudo-populations are being used, then it is at this step that these pseudo-populations are created; see Figure 2.4A). Next we do cross-validation using a 'leave-one-group-out' paradigm, which involves training on $k - 1$ groups and testing on the last group (see Figure 2.4B). We then repeat this procedure *k* times leaving out a different group each time. Finally, we repeat the whole procedure (usually around 50 times) each time selecting a different random *q* trials for each condition, and putting these conditions together in a different random set of *k* groups. This final step of repeating the whole procedure multiple times and avering the results gives a smoother estimate of the classification accuracy and is similar to bootstrap smoothing described by (Efron and Tibshirani, 1997). See Algorithm 1 for an outline of the complete decoding procedure.

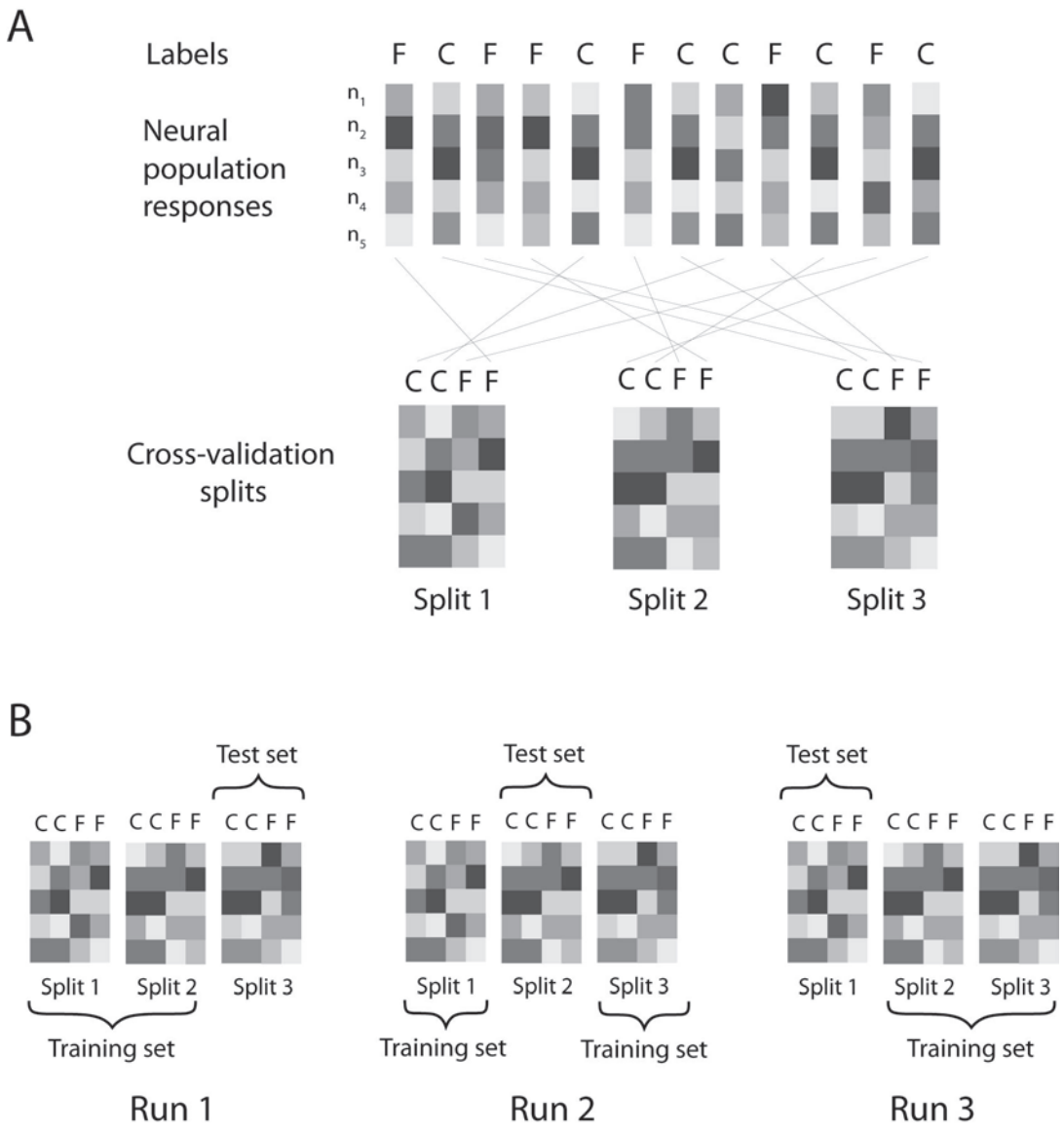


Figure 2.4 An example of cross-validation. A: An experiment in which images of a fish and an image of a cat are each shown six times in random order (q = 6). Three cross-validation splits of the data (k = 3) are created by randomly choosing data (without replacement) from two cat trials and two dog trials (j = 2) for each cross-validation split. B: A classifier is then trained on data from two of the splits and then testing on data from the third remaining split. This procedure is repeated three times (k=3), leaving out a different test split each time.

```
For 50 to 100 'bootstrap-like' trials

     Create a new set of cross-validation splits (if the data
     were not recorded simultaneously, pseudo-population
     responses are created here).

     For each cross-validation split i

          1. (Optional) Estimate feature normalization and/or
          feature selection parameters using only the training
          data. Apply these normalization and selection
          parameters to both the training and test data.

          2. Train the classifier using all data that is not
          in split i. Test the classifier using the data on
          split i. Record classification accuracy.

     end
```

Algorithm 1 The bootstrap cross-validation decoding procedure.

Feature selection and data normalization

Because different neurons often have very different ranges of firing rates, normalizing the data so that each neuron has a similar range of firing rates is often beneficial in order to ensure that all neurons are contributing to the decoding (and not just the neurons with the highest overall firing rates). Also, for some decoding analyses examining questions related to neural coding, is useful to apply feature selection methods in which only a subset of the neurons are used for training and testing the classifier (e.g., see (Meyers et al., 2008). When applying either data normalization or feature selection, it is critically important to apply these methods separately to the training and test data, since applying

any information from test set to the training set can create spurious results (see section on testing the integrity of the decoding procedure for more information). Thus, since splitting the training and test data occur within the cross-validation procedure, the normalization and feature-selection process must occur within the cross-validation procedure as well.

An example of how data normalization can be applied using a z-score normalization procedure (in which each neuron's mean firing rate is set to 0, and standard deviation is set to one) is as follows. Within each cross-validation repetition, take the $k - 1$ groups used for training, and calculate each neuron's mean firing rate and standard deviation across all the training trials, regardless of which conditions were shown. Then normalize the training data by subtracting these *training set* means from each neuron, and dividing by these *training set* standard deviations. Finally, normalize the test set data by subtracting the *training set* mean and dividing by the *training set* standard deviations for each neuron. In practice we have found that applying z-score normalization to each neuron usually marginally improves decoding accuracies, although overall we have found the results with and without such normalization to be qualitatively very similar.

A similar method can be applied when doing feature selection. In feature selection, a smaller number of neurons/features (s) that are highly selective are chosen from the larger population of all neurons. These s neurons are found *using only data from the training set*. Once the a smaller subset of neurons/features has been selected, a classifier is trained and tested using only data from these neurons/features[17]. For both the data normalization, and for the feature selection (and for all data preprocessing in general), the key notion is that the preprocessing is applied separately first to the training set without using the test set, and then it is applied to the test set separately. This insures that the test set is treated like a random sample that was selected after all parameters from training set have been fixed, and thus insures that one is rigorously testing the reliability in the data.

[17] For example, if one is trying to decode what exact images were shown to a monkey based on firing rates of individual neurons, one could use a simple feature selective method by applying a one-way ANOVA to firing rates *in the training set* (with the ANOVA groups consisting of the firing rates to particular images), and then training and testing the classifier using only neurons that had highly selective p-values..

Evaluating decoding performance

As mentioned above, the output of a classifier is usually a list of predictions for the conditions under which each test data point was recorded. The simplest way to evaluate the classification performance is to compare the predictions that the classifier has made to the actual conditions that the test data were really recorded under, and report the percent of times the classifier's predictions are correct. This method of classification evaluation is called using a 0-1 loss function, and gives reasonably interpretable results, particularly for easy classification tasks. Another method that exists for evaluating classifier performance is to use a 'rank' measure of performance (Mitchell et al., 2004). When using a rank measure of performance, the classifier must return an ordinal list that ranks how likely each test data point is to have come from each of the conditions. The rank measure then assesses how far from the bottom of the list the actual correct condition label is. The rank measure can also be normalized by the number of classes to give a 'normalized rank' measure in which a value of 1 corresponds to perfect classification, and a value of 0.5 corresponds to chance which makes the results easy to interpret. This measure also has the advantage of being more sensitive because there is not a hard limit placed on getting the actually condition exactly correct, and thus we find that this method generally works better on more difficult classification tasks.

Finally, it is also instructive to create a confusion matrix out of the classification results. If there are c conditions being decoded, a confusion matrix is a $c \times c$ sized matrix in which the columns correspond to the real condition labels of the test set, and the rows correspond to the number of times a condition labels was predicted by the classifier. The advantage of the confusion matrix is that it allows one to easily evaluate what conditions the classifier is making mistakes on, and thus what conditions elicit neural population responses that are similar. Additionally, one can convert a confusion matrix into a lower bound on the amount of mutual information (MI) between the neural population response and the conditional labels, which gives a way to compare decoding results to information theoretic measures of neural data (Samengo, 2002). Mutual information calculated from the confusion matrix can potentially be more informative than just looking at 0-1 loss

results since MI takes into account the pattern of classification errors that was made (Quian Quiroga and Panzeri, 2009). Converting a confusion matrix into a mutual information measure can be done by normalizing the confusion matrix to sum to one, and then treating the normalized matrix as a joint probability distribution between actual and predicted conditions. Applying the standard formula for mutual information[18] to this 'probability distribution' gives a lower bound estimate of mutual information.

Testing the integrity of the decoding procedure and significance of the results

Once the decoding procedure has been run, it is useful to do a few tests to ensure that any decoding accuracies that are above the expected chance level of performance are not due to artifacts in the decoding procedure. One simple test is to apply the decoding procedure to data recorded in a baseline period that occurred prior to the presentation of the condition/stimulus that has been decoded. If the decoding results are above the expected chance level during this baseline period then there is a confounding factor in the decoding procedure or in the experimental design. From our past data analyses, we have found that above-chance decoding results during baseline period are often due to changes in the average firing rate of neurons over the course of a trial combined with an experimental design or decoding procedure that is not fully randomized.

Apart from examining baseline periods, there are a few other tests that can easily be applied to check the integrity of the decoding procedure. Randomly permuting the

[18] The standard formulation being $I = \sum_{s}\sum_{s'} P(s',s)\log_2\left[P(s',s)/(P(s')P(s))\right]$ where s' are the predicted labels on the test set, s are the real labels on the test set, and P(s', s') is the joint probability distribution obtained from normalizing the confusion matrix, and the marginal distributions P(s) and P(s') can be derived from the joint distribution with the formulas $P(s') = \sum_{s} P(s',s)$ and $P(s) = \sum_{s'} P(s',s)$.

condition labels (or randomly shuffling the data itself) are other simple tests which should result in chance levels of accuracies at all time points since the relationship between the data and the condition labels is destroyed.

Randomly permuting which labels correspond to which data points also gives a way to assess when decoding accuracies are above chance. To perform this test, a null distribution is defined by the expected readout decoding accuracies if there was no relationship between the neural data and the condition labels. This null distribution can be created by permuting the relationship between the condition labels and the data, running the full decoding procedure on this label permuted data to obtain decoding results, and then repeating this permuting and decoding process multiple times. P-values can then be estimated from this null distribution by assessing how many of the values in the null distribution are less than the value obtained from decoding based on using the real labels. For example, upon performing 1000 permutations, it is possible to test if decoding accuracy with the real labels is above chance by comparing against the actual decoding accuracy with the distribution in the 1000 permutations. For an alpha level of .01, less than 10 of the 1000 decoding accuracies in the null distribution should be greater than the decoding accuracy found using the real condition label-data correspondence.

It has also been suggested that the significance of decoding results can be obtained by comparing the number of correct responses produced by a classifier to the number of correct responses one would expect by chance using a binomial distribution (Quian Quiroga and Panzeri, 2009). The method works by creating the binomial distribution $P(k) = \binom{n}{k} p^k (1-p)^{n-k}$ with n being the number of test points, and p being the proportion correct one would expect by chance (e.g., 1/(number of classes)). A p-value can then be estimated as $p - value = \sum_{k=j} P(k)$ where j is the number of correct predictions produced by the classifier. This procedure has the large advantage of being much more computationally efficient than the permutation method described above. However, there are several pitfalls of using this method that one must be aware of. In particular, for this

method to be used correctly, one should estimate the p-value for each cross-validation split separately, since using the total number of correct responses over all cross-validation splits (and/or over all 'bootstrap' repetitions) violates the assumption of data point independence that this test relies on (and hence can lead to spuriously low p-values and type 1 errors). However, estimating this p-value separately for each cross-validation split greatly reduces the sensitivity of the test leading to spuriously high p-values and potential type 2 errors. Thus unless one has a large amount of test data, it is tough to get insightful results using this method.

More advanced topics

The above sections have focused on how to run simple decoding experiments in which we are primarily interested in decoding the exact conditions/stimuli that were present when an experiment was run. However, perhaps the greatest advantage of using population decoding is that it can give insight into more complex questions about how information is coded in the brain. In the last two sections we discuss how to use neural decoding to assess how information is coded in the activity of neurons and how to assess if the information is represented in an abstract/invariant way (which is a particularly meaningful question when decoding data recorded from the highest echelons of visual cortex and pre-frontal cortex).

Examining neural coding

Despite a significant amount of research, many questions about how information is coded in the activity of neurons still have not been answered in an unambiguous way. These questions include: 1) are precise spike times important or are firing rates over longer time intervals all that matters 2) is more information present in the synchronous activity of neurons, and 3) is information at any point in time widely distributed across most of the

population of neurons, or is there a compact subset of neurons that contains all the information at any one point in time[19]. While population decoding cannot completely resolve the debate surrounding these issues, it can give some important insights into these questions. Below we describe how one can use population coding to address these issues and also some caveats one must keep in mind when interpreting the results from such analyses.

In order to address the question of how temporally precise the neural code is, it is of interest to perform population decoding using different binning schemes and to quantify how much information is lost for different representations. This can be done by simply using different bin sizes for decoding and describing which bin size gives rise to highest decoding accuracies (Meyers et al., 2009). More complex schemes can be used in which an instantaneous rate function is estimated using precise spike timing (Truccolo et al., 2005) and then this representation is used for decoding. When doing such analyses a few important caveats should be kept in mind such as the fact that the temporal precision of the recordings and a limited sampling of data could potentially influence the results.

To examine whether synchronous activity is important, or alternatively, if neurons act independently given the particular trial conditions, one can decode the activity of a population of neurons that was recorded simultaneously and compare the results to training a classifier using pseudo-populations created from the same dataset (Latham and Nirenberg, 2005). Since pseudo-populations keep the stimulus-induced aspect of the neural population code intact but destroy the correlations between neurons that occurred on any given trial (noise-correlations), this gives a measure of how much extra information is present when the exact synchronous pattern of activity on a single trial basis is preserved. Of course one must use a sufficiently powerful classifier that can exploit correlations in the data. Also, one must be careful when interpreting the results since rises or decreases in the firing rates of all neurons could potentially occur due to

[19] We use the term 'compact' subset here rather than 'sparse' subset since sparse activity usually refers to when only a few neurons are *active* at the same time, while here many neurons could be active, however only a small subset of them might *contain information* about the condition that is being decoded.

artifacts in the recording procedure. Still population decoding can begin to give an idea of how much potential additional synchronous activity could have. Additional methodological challenges when addressing these questions include the difficulties in finding neuronal combinations that could be synchronized given the large number of neurons in cortex, the potential dependencies of synchrony with distance, the potential dependencies on neuronal subtypes and others. These issues are not specific to the population decoding approach described here but they also affect other methods used to examine correlations between neurons.

Finally, feature selection can be used to examine whether information is widely distributed across most neurons or whether, at any point in time, there is a compact subset of neurons that contains all the information that the larger population has. As described in the section on cross-validation, feature selection can be used to find the most selective neurons on the training set and then use only these neurons when both training and testing the classifier. If using a reduced subset of neurons leads to decoding accuracy that is just as good as that seen in the larger population, then this indicates that indeed most of the information is contained in a small compact subset of neurons. One important caveat in this analysis is that if only one time period is examined, it is possible that some of the neurons might be non-selective due to problems with the recordings. However, if one can show that different small compact set of neurons contain the information at different points in time in the experiment (as shown in Meyers et al., 2008), this rules out problems with the recording electrode as an explanation.

Evaluating invariant/abstract representations

Since all information about the world is potentially available in early sensory organs (such as the retina for vision and the cochlea for audition), one of the more important questions about studying the brain is how information is lost in an intelligent way along the processing steps in cortex in order to create more useful invariant/abstract representations of the world. For example, many neurons in IT appear to be largely

tolerant to visual object position, since they respond similarly to particular objects regardless of the exact retinal position that an image is shown (within certain limits, see Li et al., 2009). Such an invariant representation is obviously not present in lower level areas that are retinotopic, and having this type of intelligent information loss could be behaviorally useful when an animal needs to detect the presence of an object regardless of the object's exact location on the retina.

Testing whether information is contained in an invariant/abstract way can readily be done using neural decoding. To do such a test, one can simple train a classifier on data that were recorded in one condition and then test the classifier on a different related condition. If the classifier can still perform well on the related condition then this indicates that the information is represented in an invariant or abstract way. Taking the example of position invariance again, one can train a classifier with data recorded at one retinal location and then test with data recorded at a different location, as is done in Figure 2.2 (also see Hung et al., 2005). As can be seen in Figure 2.5, neurons in IT do have information that is highly invariant/tolerant to changes in the exact retinal position. A similar type of analyses can also be done to test if different brain regions contain information in an 'abstract' format. For example, Meyers et al. (2008) used data from a task in which monkeys needed to indicate whether an image of a cat or dog was shown regardless of which exact image of a dogs and cats was shown. By training a classifier with data that was collected from a subset of images of dogs and cats and then testing the classifier when a different set of images of dogs and cats were shown, Meyers et al. (2008) could see that indeed there seemed to be information about the more abstract behaviorally relevant categories apart from the information that was due to the exact visual images of particular dogs and cats. An analogous method of training a classifier with data from one time period and testing with data from a different time period was also used by Meyers et al. (2008) to show that the neural code of an image does not appear to be stationary but instead seems to change systematically over the course of a trial – which illustrates again how training and testing with different but related data is an effective way to answer a range of different questions.

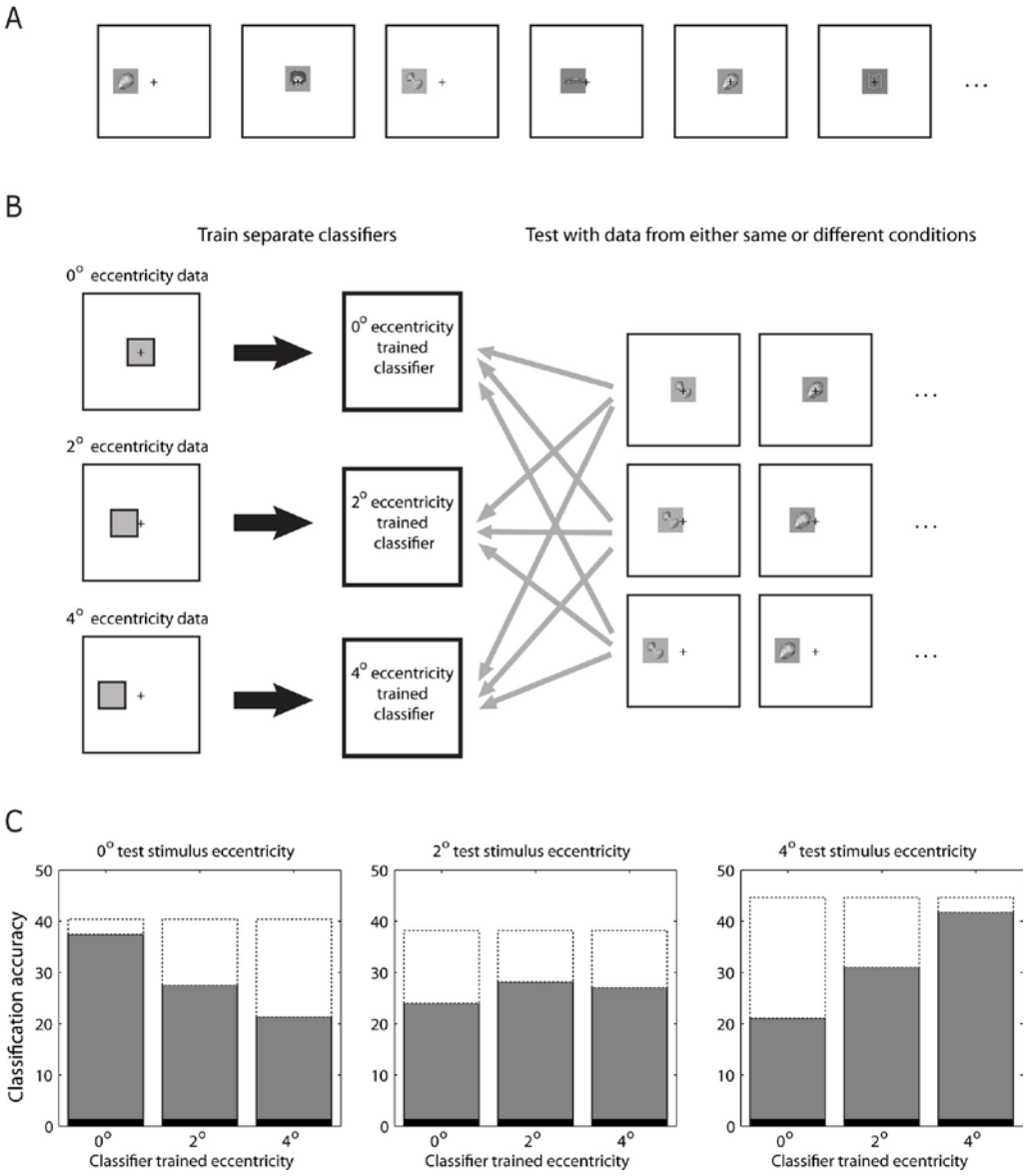


Figure 2.5 Assessing position invariance in anterior inferior temporal cortex. A: An illustration of an experiment conducted by (Hung et al., 2005) in which images of 77 objects were displayed at three different eccentricities. B: An illustration of a classifier being trained on data from different eccentricities, yielding three different models. These three models were then tested with data from either the same eccentricity that the classifier was trained on (using data from different trials), or with data at a different eccentricity. C: Results from this procedure show that the best performance was always achieved when the training and testing was done at the same eccentricity (gray bars), however performance is well above chance (black bars) at all eccentricities, indicating the population of IT is very position tolerant. Also, when the classifier is trained using data from all eccentricities (dotted bars), the results are even better than when training and testing is done at the same eccentricity, indicating that the best performance can be achieved when the classifier learns to rely mostly heavily on the neurons that have the most position invariance. Decoding results are based on multi-unit recordings from 70 neurons made by (Hung et al., 2005), using the mean firing rate in a 200ms bin that started 100ms after stimulus onset and a MCC classifier.

Conclusions

In this chapter we described how to implement a population decoding procedure, highlighted the analysis methods that we have found work best, and pointed out caveats to be aware of when interpreting results. Neural population decoding holds a great amount of potential as a method to gain deeper insight into how the brain functions, particularly with regard to answering questions related to neural coding and to how invariant and abstract representations are created in different brain regions.

Acknowledgements

We would like to thank Jim DiCarlo and Chou Hung for supplying the data that were used in this chapter. We would also like to thank Tomaso Poggio for his continual guidance. This work was supported by the American Society for Engineering Education's National Science Graduate Research Fellowship (EM) and by NSF and NIH.

References

Aggelopoulos N, Franco L, Rolls E (2005) Object perception in natural scenes: Encoding by inferior temporal cortex simultaneously recorded neurons. Journal of Neurophysiology 93:1342-1357

Anderson B, Sanderson M, Sheinberg D (2007) Joint decoding of visual stimuli by IT neurons' spike counts is not improved by simultaneous recording. Experimental Brain Research 176:1-11

Averbeck B, Latham P, Pouget A (2006) Neural correlations, population coding and computation. Nature Reviews Neuroscience 7:358-366

Averbeck B, Lee D (2006) Effects of noise correlations on information encoding and decoding. Journal of Neurophysiology 95:3633-3644

Averbeck BB, Lee D (2004) Coding and transmission of information by neural ensembles. Trends in Neurosciences 27:225-230

Efron B, Tibshirani R (1997) Improvements on Cross-Validation: The .632+ Bootstrap Method. Journal of the American Statistical Association 92:560, 548

Fee MS, Mitra PP, Kleinfeld D (1996) Automatic sorting of multiple unit neuronal signals in the presence of anisotropic and non-Gaussian variability. J. Neurosci. Methods 69:175-188

Georgopoulos AP, Caminiti R, Kalaska JF, Massey JT (1983) Spatial coding of movement: a hypothesis concerning coding of movement direction by motor cortical populations. Experimental Brain Research Supplemental 7:327-336

Gochin P, Colombo M, Dorfman G, Gerstein G, Gross C (1994) Neural Ensemble Coding in Inferior Temporal Cortex. Journal of Neurophysiology 71:2325-2337

Harris KD, Henze DA, Csicsvari J, Hirase H, Buzsaki G (2000) Accuracy of Tetrode Spike Separation as Determined by Simultaneous Intracellular and Extracellular Measurements. J Neurophysiol 84:401-414

Hatsopoulos NG, Ojakangas CL, Maynard EM, Donoghue JP (1998) Detection and identification of ensemble codes in motor cortex In H. Eichenbaum & J. Davis , eds. Neuronal ensembles: strategies for recording and decoding New York: Wiley, p. 161–175.

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

Latham PE, Nirenberg S (2005) Synergy, Redundancy, and Independence in Population Codes, Revisited. J. Neurosci. 25:5195-5206

Lewicki MS (1998) A review of methods for spike sorting: the detection and classification of neural action potentials. Network 9:R53-78

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

Meyers E, David Freedman, Gabriel Kreiman, Earl Miller, Tomaso Poggio (2009) Decoding dynamic patterns of neural activity using a 'biologically plausible' fixed set of weights In Salt Lake City, UT, USA: Frontiers in systems neuroscience. Available at: http://frontiersin.org/conferences/individual_abstract_listing.php?conferid=39&pap=1437&ind_abs=1.

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 100:1407-19

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S (2004) Learning to Decode Cognitive States from Brain Images. Mach. Learn. 57:145-175

Nikolic D, Haeusler S, Singer W, W. M (2007) Temporal dynamics of information content carried by neurons in the primary visual cortex In Advances in Neural Information Processing Systems Cambridge, MA: MIT Press, p. 1041--1048.

Panzeri S, Pola G, Petersen R (2003) Coding of sensory signals by neuronal populations: The role of correlated activity. Neuroscientist 9:175-180

Quian Quiroga R, Panzeri S (2009) Extracting information from neuronal populations: information theory and decoding approaches. Nat Rev Neurosci 10:173-185

Quiroga RQ, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput. 16:1661-1687

Rolls ET, Treves A, Tovee MJ (1997) The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. Experimental Brain Research 114:149-162

Samengo I (2002) Information loss in an optimal maximum likelihood decoding. Neural Computation 14:771-779

Tanila H, Shaprio M (1998) Ensemble Recordings and the Nature of Stimulus Representation in Hippocampal Cognitive Maps In H. B. Eichenbaum & J. L. Davis , eds. Neuronal ensembles: strategies for recording and decoding New York: Wiley, p. 177-206.

Truccolo W, Eden UT, Fellows MR, Donoghue JP, Brown EN (2005) A Point Process Framework for Relating Neural Spiking Activity to Spiking History, Neural Ensemble, and Extrinsic Covariate Effects. J Neurophysiol 93:1074-1089

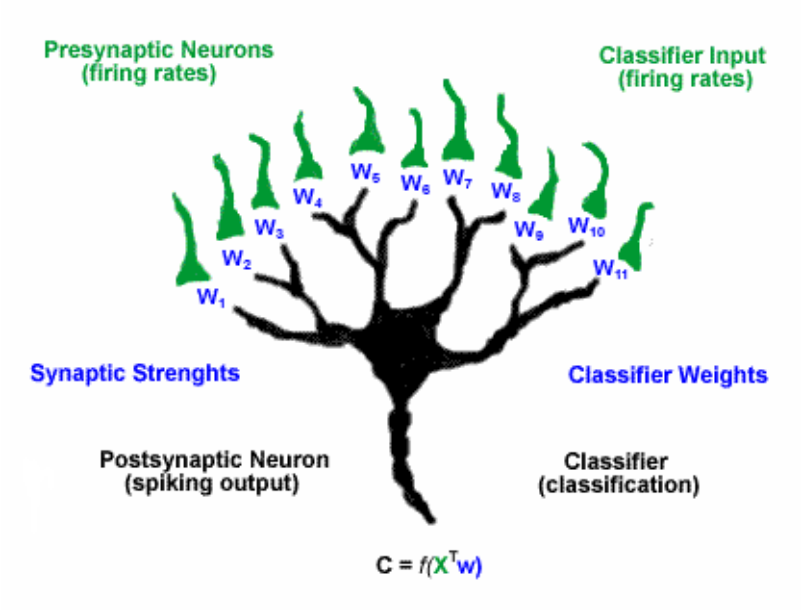
Wehr M, Pezaris J, Sahani M (1999) Simultaneous paired intracellular and tetrode recordings for evaluating the performance of spike sorting algorithms. Neurocomputing 26-27:1061-1068

Additional material: Ways to view decoding results

As described in this chapter, decoding analyses measure how well one can predict which stimulus (or other behaviorally relevant variable) was present on a given trial, based on the activity of a population of neurons. Below I briefly discuss a couple of interpretations for relating decoding results to neural processing.

As the information available to downstream neurons

One popular way to relate population decoding results to neural information processing is to propose that the decoding accuracy is measure of the amount of information that is available to a downstream neuron (Hung et al., 2005; Li et al., 2009). When a linear classifier is used, the weights of the classifier can be viewed as being analogous to synaptic strengths between the upstream population and a downstream neuron (see Additional supplemental material 2.1). While such an interpretation is appealing, one must obviously be careful not to take this interpretation too literally because: 1) it is very unlikely that there actually is a downstream neuron that has connections with a large percentage of the recorded population, and even if there was such a neuron, the chance that the weights inferred by the classifier actually reflected the real synaptic strengths seems improbable, and 2) given that one is assuming that information in the upstream area is coded by a population of neurons, inferring that all this information would be extracted into a single downstream neuron implies a strange assumption in which a distributed code is converted in a highly spare/compact code, which again seems unlikely. Thus, while one can loosely say that population decoding estimates the amount of information available to a downstream neuron, stronger interpretations should be avoided.



Additional supplemental material 2.1 Illustration of a possible correspondence between neuronal processing elements and classification algorithms.

<u>As estimating the state of a computational system</u>

A second interpretation of the decoding procedure is that it is trying to estimate the reliability that a population of neurons enters into a particular computational state. In such an interpretation, one can view the neural activity as being analogous to the activity of a set of transistors/bits in a central processing unit or in the memory of a digital computer. The decoding procedure is then estimating the state that a computational system is in, and relating this state to known stimuli, behavioral variables, or other neural states. This interpretation has the advantage of being more agnostic about the role of the decoding algorithm, and allows one to relate the decoding of neural spiking activity to decoding results from fMRI activity and recordings of other types of data.

Additional References

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

Li N, Cox DD, Zoccolan D, DiCarlo JJ. What Response Properties Do Individual Neurons Need to Underlie Position and Clutter "Invariant" Object Recognition? J Neurophysiol 102: 360-376, 2009.

Chapter 3: Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex

The material presented in this chapter was published as a paper in the Journal of Neurophysiology, and has been reprinted with permission from The American Physiological Society. The reference for the original paper is:

Meyers, E., Freedman, D., Kreiman, G., Miller, E., Poggio T. Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex. Journal of Neurophysiology, 100:1407-1419, 2008. DOI: 10.1152/jn.90248.2008

Abstract

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 (Freedman et al. 2003). 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 non-stationary 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 at least two hundred 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 coded in a given brain region based on the number of neurons that respond to particular stimuli or aspects of the task, or based on the strength of an index value 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 index values, 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 efforts 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 upon which to judge the relative validity of these models.

In order 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 pre-frontal 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 while 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 in these structures (Riesenhuber and Poggio 2000; Serre et al. 2005).

Materials and 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 (DMC). Each DMC trial consisted of a sequence of 4 periods: a fixation period (500ms duration), a sample period in which a stimulus was shown (600ms duration), a delay period (1000ms), and a decision period in which a second stimulus was shown and the monkey needed to make a behavioral decision (Figure 3.1A). The stimuli used in the task were morphed images generated from 3 prototype images of cats and 3 prototype

images of dogs (Figure 3.1B-C). A morph stimulus was labeled a 'cat' or 'dog' depending on the category of the prototype that contributed more than 50% to its morph. During the sample period of the task, a set of 42 images (Supplemental figure 3.1) were used that consisted of the 6 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 at least 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 in order to receive a juice reward (or to continue to hold the lever and release it for a second decision-stimulus in the non-match trials). Performance on the task was ~90% correct. Figure 3.1 illustrates the time course of an experimental trial, one morph line used in the experiment, and the 6 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.

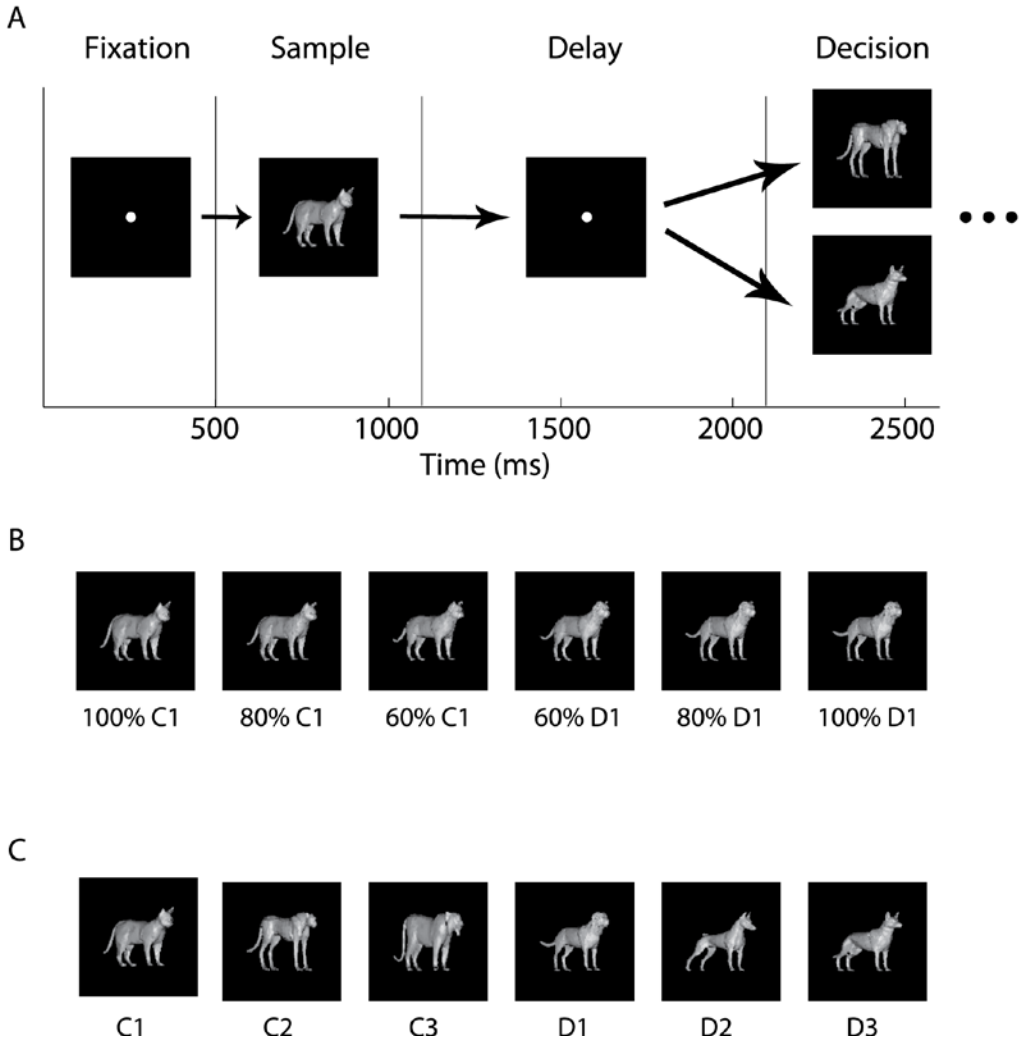


Figure 3.1 Organization of the stimuli and behavioral task. A, time course of the delayed match to category experiment. B, an example of one of the nine morph lines of the stimuli from the cat 1 prototype to the dog 1 prototype (the actual stimuli used in the experiment were colored orange, see Freedman et al. 2002). C, the six 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.

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 since 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 non-match (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 non-stationary trends (those whose average firing rate variance in 20 consecutive trials was greater than twice 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 5 trials for each of the 42 stimuli shown in the sample period. 283 ITC neurons and 332 PFC neurons were selected for further consideration after applying the constraints indicated above. From the pools of either ITC neurons or PFC neurons we applied the procedure below separately for 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 5 trials for each of the 42 stimuli. Third, The firing rates of the 256 neurons from each of the 5 trials were concatenated together to create 210 data points (5 repetitions x 42 stimuli) in R^{256} space. Fourth a cross-validation procedure was repeated 5 times. In each repetition, 4 data points from each of the 42 classes were used as training data and 1 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 standard deviation for each neuron (the mean and standard deviation 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 just 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 Figures 3.2-3.7 is the classification accuracy averaged over the 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 Figure 3.2, except that 50 data points for each class were used in each of the 5 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 Figures 3.3-3.7, the procedure was used exactly as described above except that the 42 identity labels were remapped to their respective 'dog' and 'cat' categories.

Unless otherwise noted, all figures that show smooth estimates of classification accuracy as a function of time are based on using firing rates in 150ms bins sampled at 50ms intervals with data from each time bin being classified independently. Because the sampling interval we used is shorter than the bin size (50ms sampling interval, 150ms 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 225ms after sample stimulus onset, results from the delay period are reported for 525ms after sample stimulus offset, and results from the decision period are reported for 225ms after decision stimulus offset (this corresponds to 725ms, 1625ms, and 2325ms after the start of a trial, with each bin width being 150ms). The results reported for 'basic' decoding accuracies are the mean and one standard deviation of the decoding accuracies over all the bootstrap trials and cross-validation splits (we refer to these results as 'basic decoding results'). The results reported for decoding 'abstract category' information are the average and one standard deviation of basic decoding results taken over the 9 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, in order 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), in order 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 asses 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 post-synaptic 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 figure 3.2). 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, which 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 vs. 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 under-estimate 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 (Figure 3.8), 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 100ms bins sampled at 10ms intervals. To compute the mutual information, we assumed the prior probability of each stimulus category was equal, and we used the standard formula, $I = \sum_{s,r} P[r\ s] \log_2 (P[r, s]/P[r]\ P[s])$ (Dayan and Abbott 2001). The conditional probability distribution between stimulus and response, $P[r|s]$, was estimated from the empirical distribution using all trials. While 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.

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 (Figure 3.1A) (Freedman et al. 2003). Figure 3.2 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 Figure 3.2A, 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 (Figure 3.1B), 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% versus 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, while 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, Figure 3.2B). When the SAMPLE-STIMULUS was first presented, ITC had a slightly higher accuracy level than PFC (92.0% ± 2.8% versus 81.3% ± 4.3%, at t=225ms, 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% versus 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) versus 56.6% ± 4.8%

(ITC); decision: 88.4% ± 4.3% (PFC) versus 77.9% ± 4.4% (ITC), respectively; 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 below).

Figure 3.2C 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% versus 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 non-match trial during the decision period (92.3% ± 2.7% versus 60.5% ± 4.8% Figure 3.2D), 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. Figures 2E and 2F compare the decoding accuracies for three different variables: 1) whether a trial is a match/non-match trial (brown), 2) the category of the DECISION-STIMULUS (green) 3) the category of the SAMPLE-STIMULUS (purple) (we start the comparison in the middle of the delay period because there is no information about trial status and DECISION-STIMULUS category until the decision period). Results from ITC (Figure 3.2E) 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 non-match trial (brown). Also, the match/non-match trial information showed the longest latency. This pattern shows that the variable that ITC has

the most information about (of the three variables listed above) 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. 2F), with the most information being about task-related variables; i.e., whether a trial is a match or non-match trial. Also, the latency of the match/non-match 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.

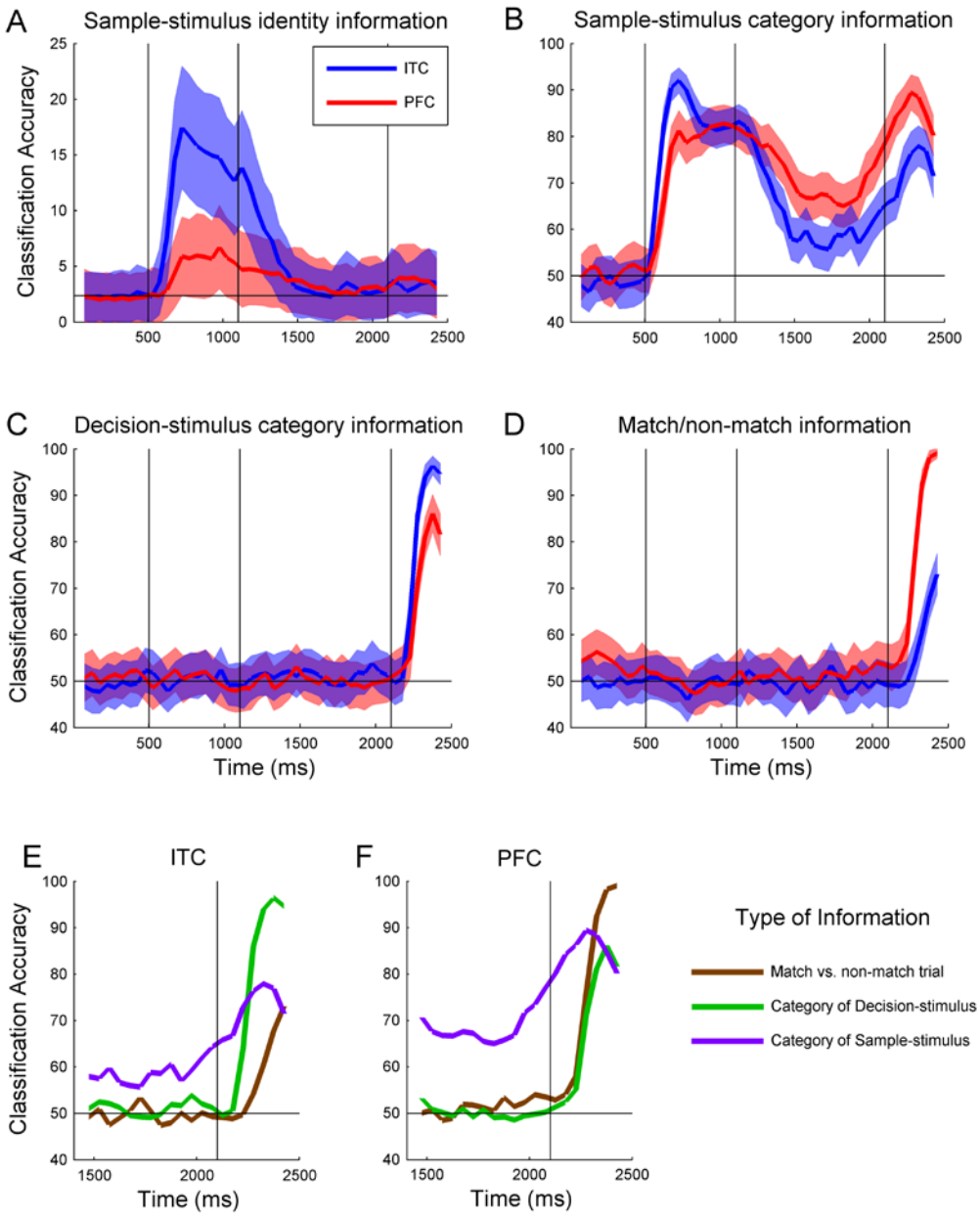


Figure 3.2 Basic decoding results for four different types of information. In figures A-D, blue lines indicates results from ITC and red lines indicate results from PFC, (red, and blue shaded regions indicate one standard deviation over the bootstrap-like trials). The three vertical black lines indicate sample stimulus onset, sample stimulus offset, and match stimulus onset from left to right respectively. E-F, comparison of sample-stimulus category decoding accuracy (purple), decision-stimulus category decoding accuracy (green) and whether a trial is a match or non-match trial (brown), for ITC (E) and PFC (F).

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 Figure 3.2B, 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 figure 3.3 for more details). In order 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 greater than 60% of their morph from a given prototype). The logic beyond 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 (since 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 figure 3.3). Thus, above chance classification performance in this analysis would imply that a brain region had much 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 (Figure 3.1C and Supplemental figure 3.1). Therefore, this decoding method should greatly reduce the influence of visual features (see Discussion section 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 figure 3.4B).

Figure 3A shows the decoding results of this more 'abstract' category information for ITC (blue) and PFC (red) averaged over all 9 training/test permutations (e.g., train on [c1, c2 vs. d1 d2] test on [c3 d3]; training on [c1, c2 vs. d1, d3], etc.). Supplemental figure 3.4A shows the results for the 9 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 Figure 3.2B is due to the classifier combining category information in a visually based format, with information in a more abstract format.

Figure 3.3B compares the visual plus abstract category information (blue trace) that was shown in Figure 3.2B with the abstract category information (green trace) that was shown in Figure 3.3A, for ITC (left) and PFC (right). 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.

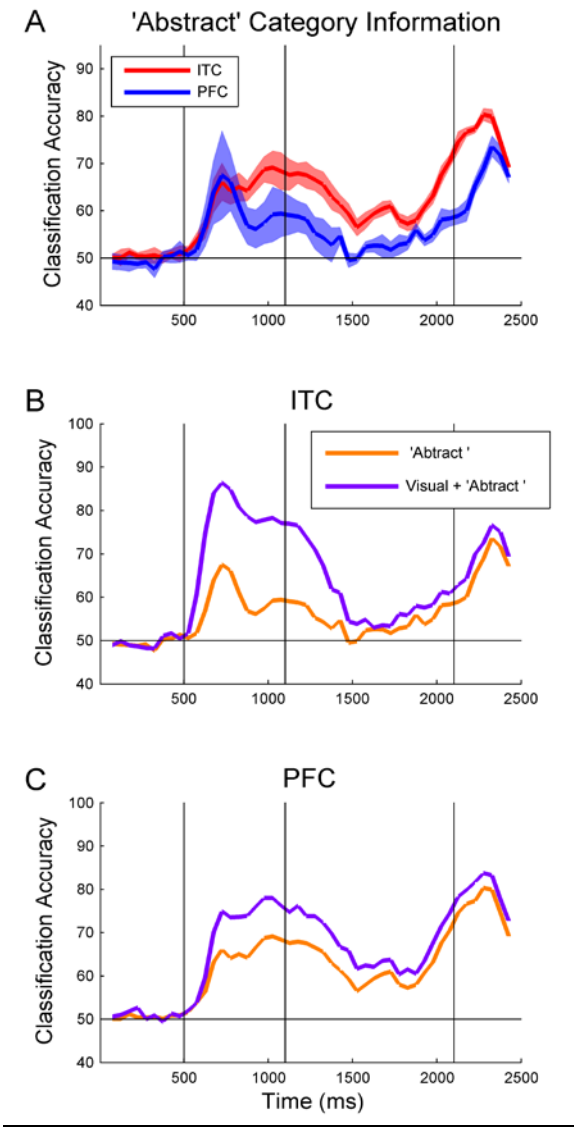


Figure 3.3 Decoding task-relevant 'abstract' category information. A, decoding accuracies for ITC (blue) and PFC (red) when training on data from two dog and two cat prototype images and testing on the remaining dog and cat prototype images. The results are the average over all 9 permutations of training/test splits and the shaded results show the standard deviations over the 9 permutations (the individual traces are shown in supplementary figure S4A). B-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 to 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 figure 2B 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). In order to asses the if there is a smaller compact subset of neurons 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 (Figure 3.4). The best k neurons were defined as those neurons with the smallest p-values based on a t-test applied to all cat trials vs. all dog-trials on the training data set (see Materials and 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 where, strikingly, 8 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 figure 3.5), 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.

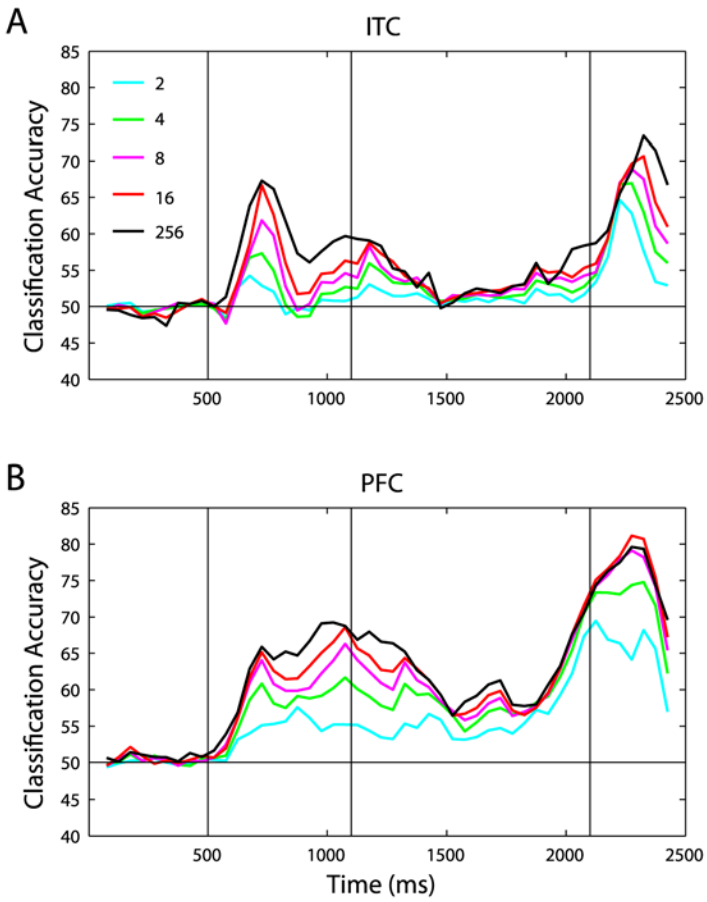


Figure 3.4 Readout using the 'best' 2, 4, 8, or 16 neurons, compared to readout using all 256 neurons, for ITC (A) and PFC (B). As can be seen, for almost all time periods, the abstract category information available in whole population is available in only 16 or fewer neurons. The 'best' neurons were 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 n neurons were selected.

Redundancy allows a system to be robust to degradation of individual neurons or synapses. This robustness constitutes a key feature of biological systems. In order 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 3.5 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). Since 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 non-negligible 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 figure 3.6).

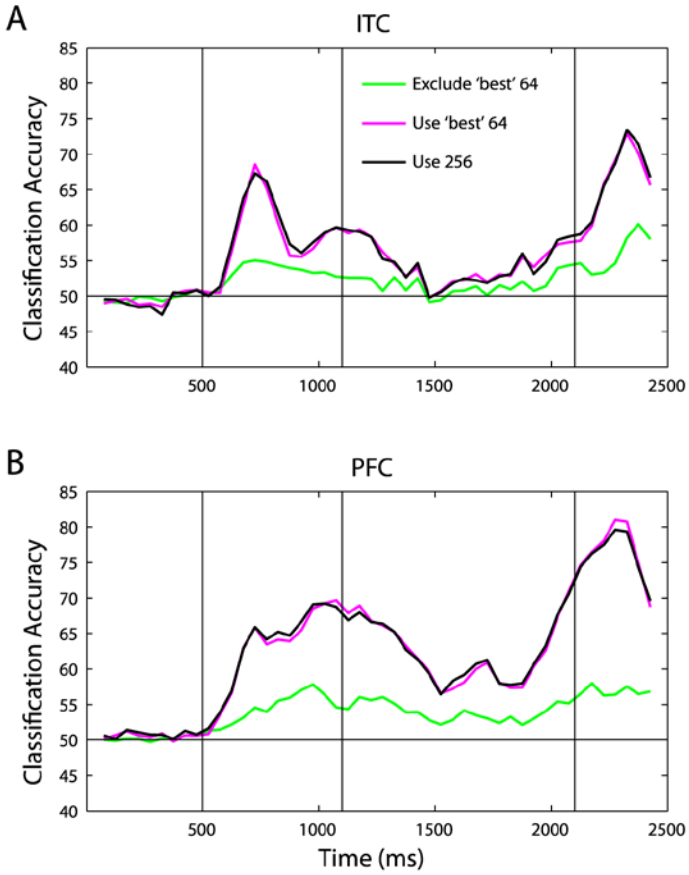


Figure 3.5 Illustration of redundant information in ITC (A), and PFC (B). The purple line indicates the readout performance when the top 64 neurons were used, and the green line indicates when the top 64 neurons were excluded and the remaining 192 neurons were used. As can be seen, the top 64 neurons 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.

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

Figure 3.6A-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 3.6C-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 Figure 3.2 and 3.3 (i.e., these are plots of three rows of Figure 3.6A and B, at time points during the sample, delay, and decision periods and compares them to the results in Figure 3.3A). 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.

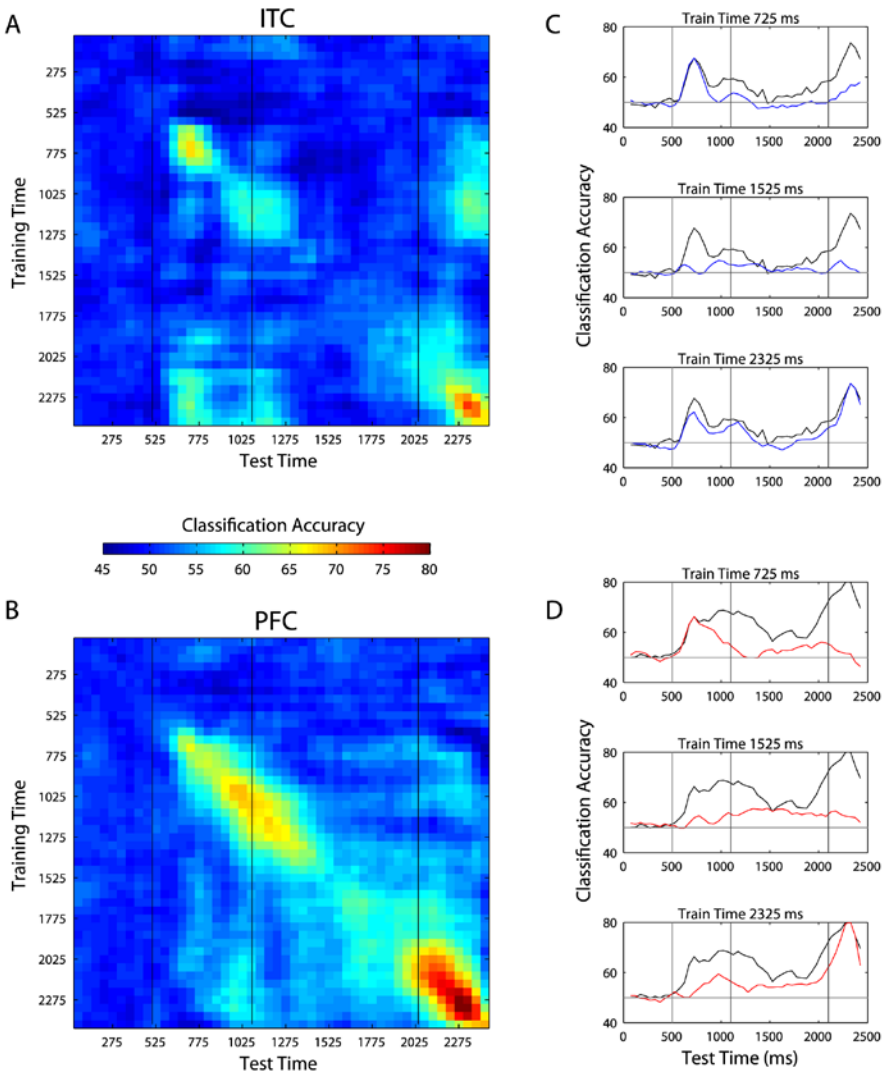


Figure 3.6 Evaluating whether the same code is used at different times for abstract category information. A, in ITC there is some similarity in the neural code for abstract category information in the sample and the match periods, as can be seen by the green patches near the upper right and lower left 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 775-1275ms 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-D, examples of decoding accuracies using three fixed training times from the sample, delay and decision periods (colored lines) compared decoding accuracies obtained when training and testing using the sample time period (black line), for ITC (C) and PFC (D); (each of these plots corresponds to one row from the from figures A or B and the black line corresponds to the diagonal of this figure, and is the same line as shown in Fig 3A). 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.

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 (out of 256 random neurons selected on each bootstrap trial) at one 150ms time period (indicated on the y-axis in Figure 3.7) and training and test data were taken from a different 150ms 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 3.7 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 figure 3.7 shows that the neurons which 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.

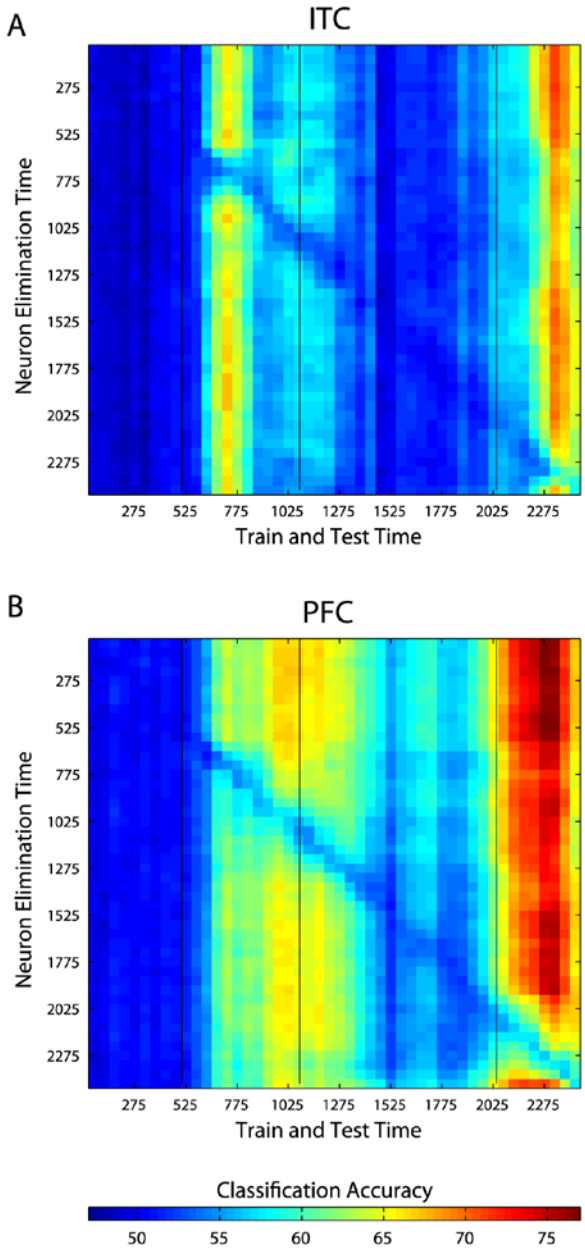


Figure 3.7 Elimination of the 'best' 64 neurons from the time period t1 (specified on the y-axis), and then training and testing with all the remaining 192 neurons at time period t2 (as specified by the x-axis), for ITC (A), and PFC (B). Eliminating the 'best' neurons from the training set at one time 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 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.

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 50ms bins (see Materials and Methods). Figure 3.8, 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 neuron 1 and neuron 4 in Figure 3.8A, 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.

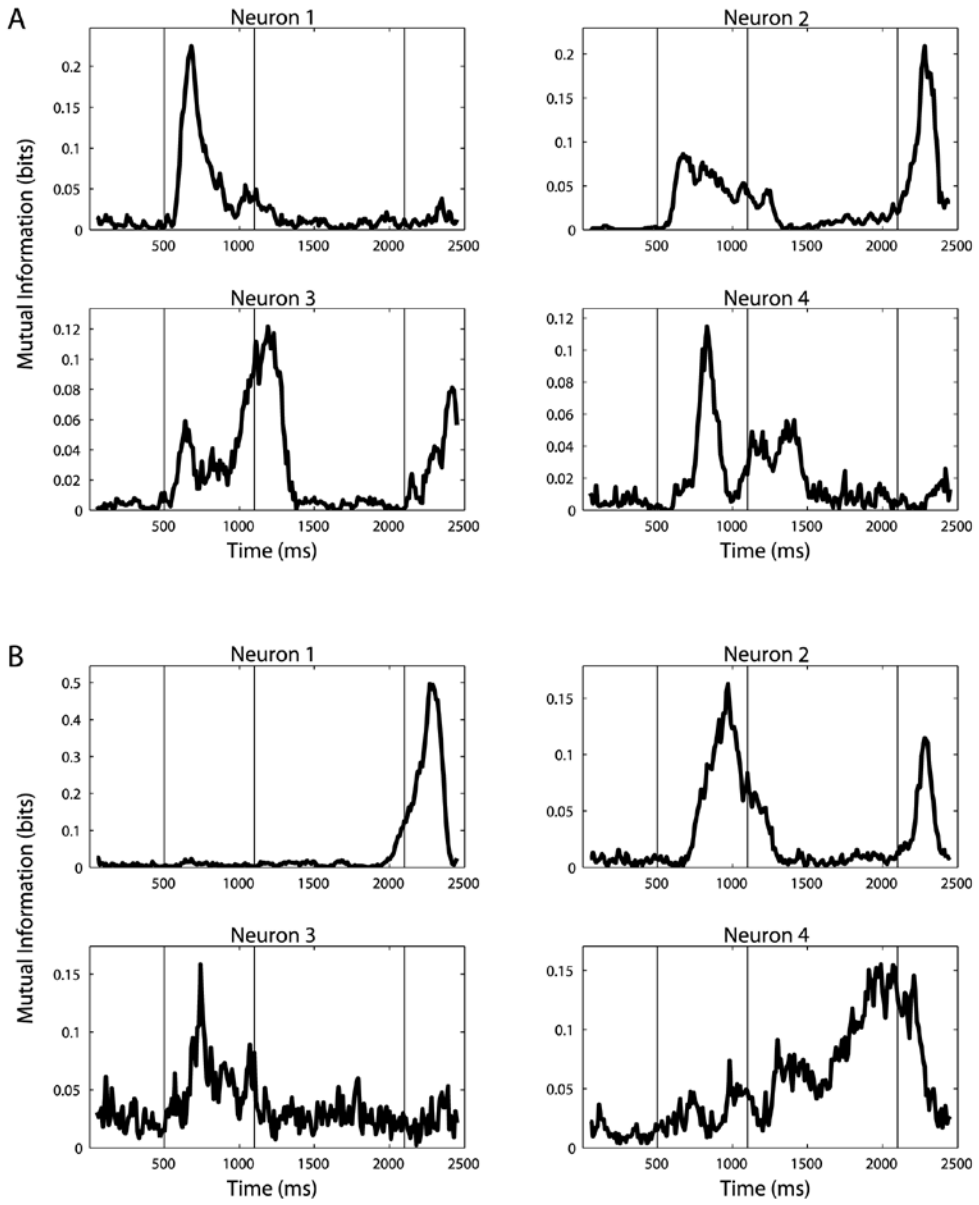


Figure 3.8 Illustration showing that many individual neurons have short periods of selectivity for ITC (A), and PFC (B). The figure plots the four neurons for ITC and PFC that had the highest the mutual information between the category of the sample-stimulus and neuron's firing rate (firing rates were calculated using 100ms bin periods sampled every 10ms). As can be seen, most neurons show high 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 neuron 1 and neuron 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.

Discussion

We applied population decoding methods to neuronal spiking data recorded in PFC and ITC in order 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 observe the time course of the flow of information in these areas with a fine timescale. Results from basic decoding analyses (Figure 3.2) 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 stimuli as soon as they are 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) and (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 (Figure 3.3) 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 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 were 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 figure 3.9). 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 (Figure 3.6). These analyses suggest that the pattern of activity coding for a particular stimulus or behaviorally relevant variable changes with time. Such results are consistent with the findings of Gochin et al. (1994), in which a paired-associate task was used to show that the pattern of activity in macaque IT 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 non-linear 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 (Figure 3.7). 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. (2002) 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 which 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 can not conclusively answer which coding scheme is correct, they do give some insight into this debate. First, because we decode mean firing rates over 150ms 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 vs. non-match 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 subset 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 correlated noise (Averbeck et al. 2006; Averbeck and Lee 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 figure 3.2, 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, since 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. 3.2-3.7) 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 would like to thank Beata Jarosiewicz and Max Riesenhuber for their helpful comments on the manuscript. This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL). This research was sponsored by grants from: National Science Foundation, and Darpa. Additional support was provided by: Children's Ophthalmology Foundation, Epilepsy Foundation, NDSEG 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 ; New York: Cambridge University Press, 1991, p. xiv, 280 p.

Aggelopoulos NC, Franco L, and Rolls ET. Object perception in natural scenes: Encoding by inferior temporal cortex simultaneously recorded neurons. Journal of Neurophysiology 93: 1342-1357, 2005.

Amit DJ. Modeling brain function : the world of attractor neural networks. Cambridge ; New York: Cambridge University Press, 1989, p. xvii, 504 p.

Averbeck BB, Latham PE, and Pouget A. Neural correlations, population coding and computation. Nature Reviews Neuroscience 7: 358-366, 2006.

Averbeck BB, and Lee D. Effects of noise correlations on information encoding and decoding. Journal of Neurophysiology 95: 3633-3644, 2006.

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

deCharms RC, and Zador A. Neural representation and the cortical code. Annual Review of Neuroscience 23: 613-+, 2000.

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

Freedman DJ, Riesenhuber M, Poggio T, and Miller EK. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. Journal of Neuroscience 23: 5235-5246, 2003.

Gochin PM, Colombo M, Dorfman GA, Gerstein GL, and Gross CG. Neural Ensemble Coding in Inferior Temporal Cortex. Journal of Neurophysiology 71: 2325-2337, 1994.

Hahnloser RHR, Kozhevnikov AA, and 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. Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences 79: 2554-2558, 1982.

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

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, and Cohen JD. An integrative theory of prefrontal cortex function. Annual Review of Neuroscience 24: 167-202, 2001.

Mogami T, and Tanaka K. Reward association affects neuronal responses to visual stimuli in macaque TE and perirhinal cortices. Journal of Neuroscience 26: 6761-6770, 2006.

Nieder A, Freedman DJ, and 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, and W. M. 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, and Hoffman T. Cambridge, MA: MIT Press, 2007, p. 1041--1048.

Olshausen BA, and Field DJ. Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Research 37: 3311-3325, 1997.

Paninski L. Estimation of entropy and mutual information. Neural Computation 15: 1191-1253, 2003.

Perez-Orive J, Mazor O, Turner GC, Cassenaer S, Wilson RI, and 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, and 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, and Andersen RA. Movement intention is better predicted than attention in the posterior parietal cortex. Journal of Neuroscience 26: 3615-3620, 2006.

Rainer G, and Miller EK. Timecourse of object-related neural activity in the primate prefrontal cortex during a short-term memory task. European Journal of Neuroscience 15: 1244-1254, 2002.

Rainer G, Rao SC, and Miller EK. Prospective coding for objects in primate prefrontal cortex. Journal of Neuroscience 19: 5493-5505, 1999.

Riesenhuber M, and Poggio T. Models of object recognition. Nature Neuroscience 3(supp): 1199-1204, 2000.

Rolls ET, and Tovee MJ. The Responses of Single Neurons in the Temporal Visual Cortical Areas of the Macaque When More Than One Stimulus Is Present in the Receptive-Field. Experimental Brain Research 103: 409-420, 1995a.

Rolls ET, and Tovee MJ. Sparseness of the Neuronal Representation of Stimuli in the Primate Temporal Visual-Cortex. Journal of Neurophysiology 73: 713-726, 1995b.

Rumelhart DE, McClelland JL, and University of California San Diego. PDP Research Group. Parallel distributed processing : explorations in the microstructure of cognition. Cambridge, Mass.: MIT Press, 1986.

Samengo I. Information loss in an optimal maximum likelihood decoding. Neural Computation 14: 771-779, 2002.

Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, and Poggio T. A Theory of Object Recognition: Computations and Circuits in the Feedforward Path of the Ventral Stream in Primate Visual Cortex 2005.

Seung HS, and Sompolinsky H. Simple-Models for Reading Neuronal Population Codes. Proceedings of the National Academy of Sciences of the United States of America 90: 10749-10753, 1993.

Shima K, Isoda M, Mushiake H, and Tanji J. Categorization of behavioural sequences in the prefrontal cortex. Nature 445: 315-318, 2007.

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

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, and Miyashita Y. Top-down signal from prefrontal cortex in executive control of memory retrieval. Nature 401: 699-703, 1999.

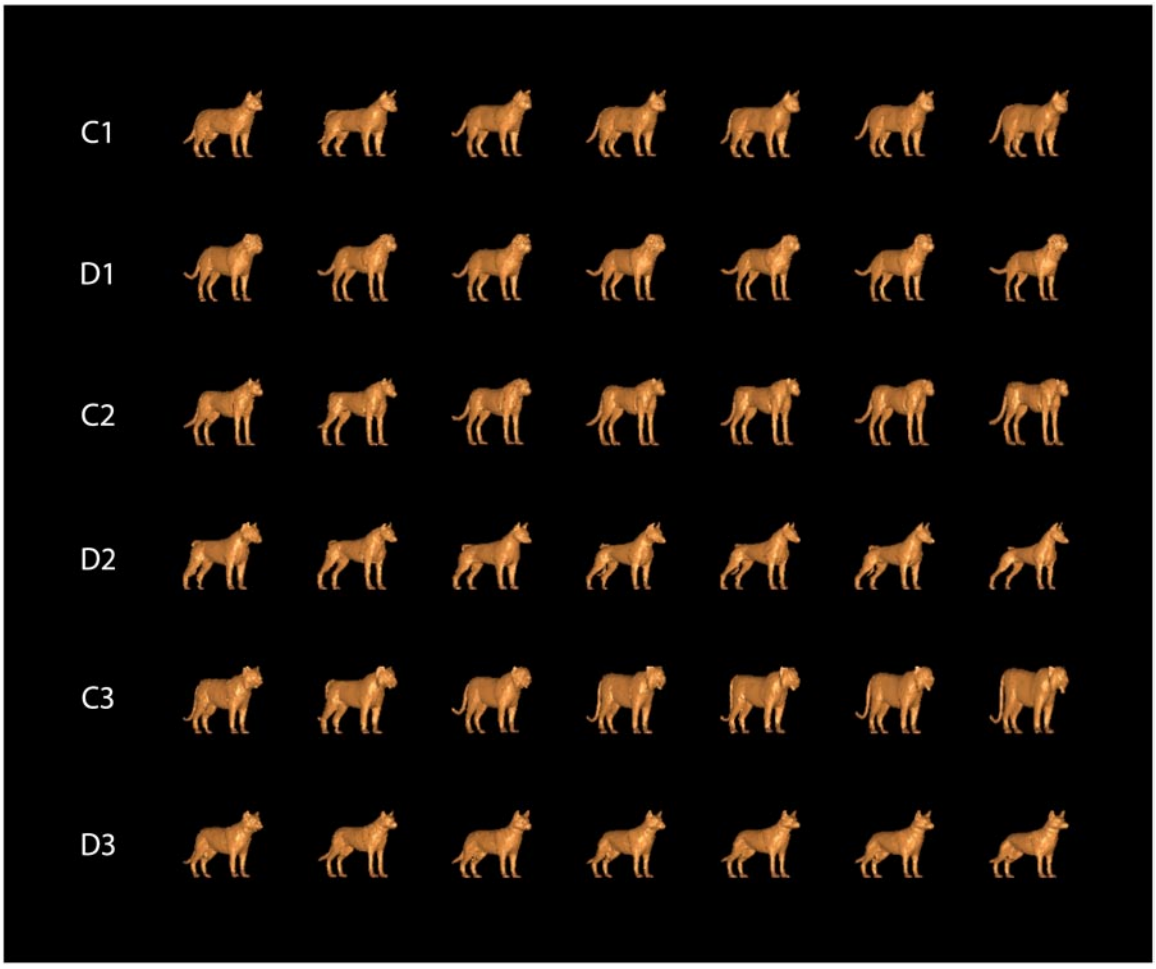
Trappenberg TP. Fundamentals of computational neuroscience. Oxford ; New York: Oxford University Press, 2002, p. xvi, 338 p.

Ungerleider LG, Gaffan D, and Pelak VS. Projections from Inferior Temporal Cortex to Prefrontal Cortex Via the Uncinate Fascicle in Rhesus-Monkeys. Experimental Brain Research 76: 473-484, 1989.

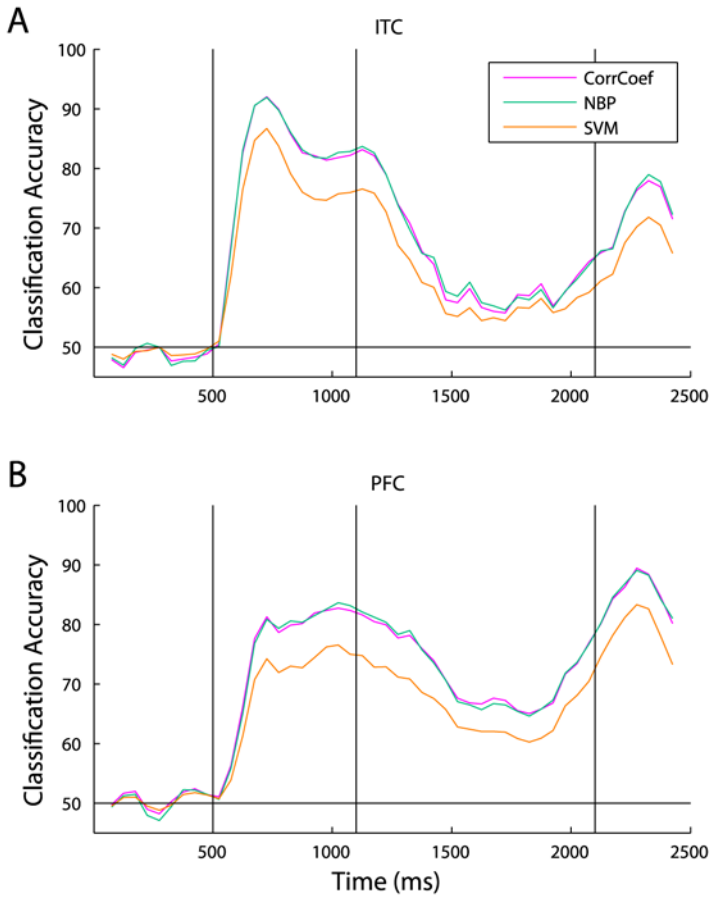
Zaksas D, and Pasternak T. Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. Journal of Neuroscience 26: 11726-11742, 2006.

Zemel RS, Dayan P, and Pouget A. Probabilistic interpretation of population codes. Neural Computation 10: 403-430, 1998.

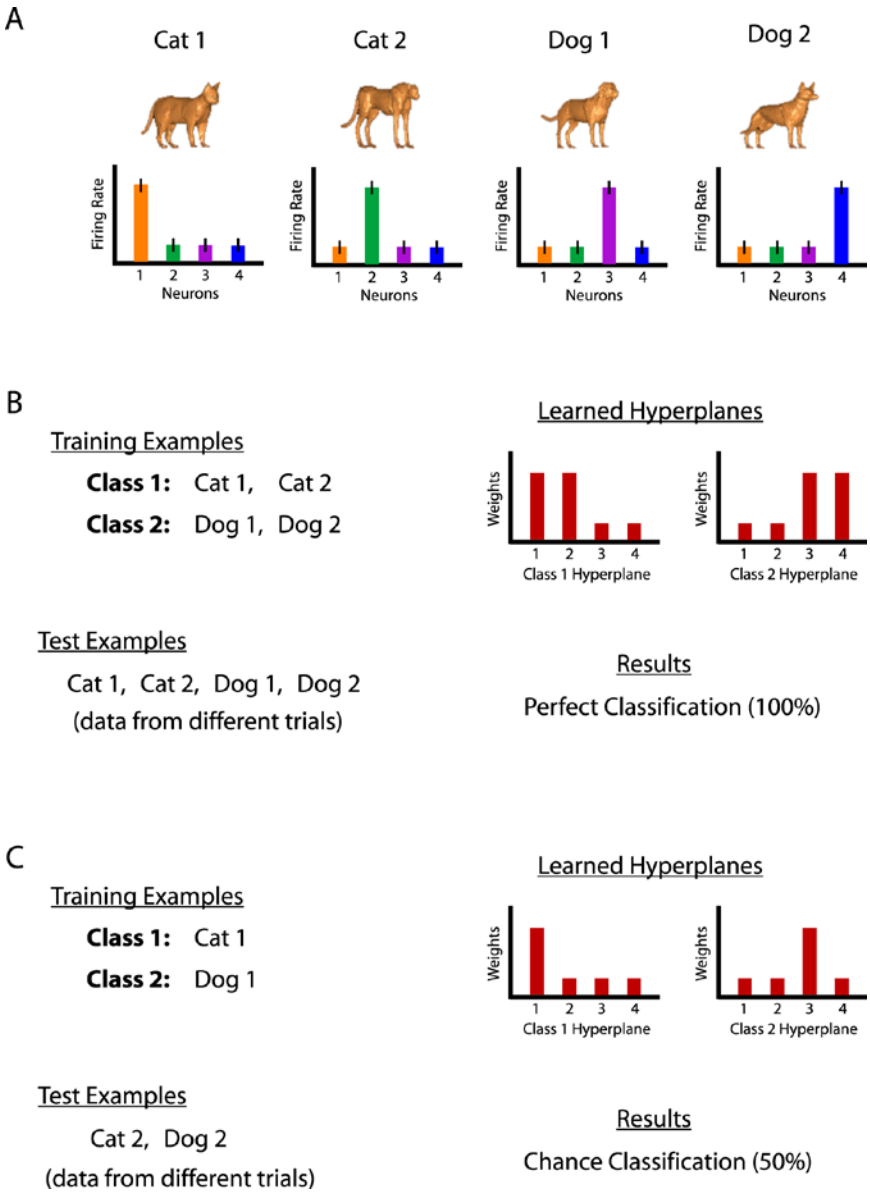
Supplemental Material



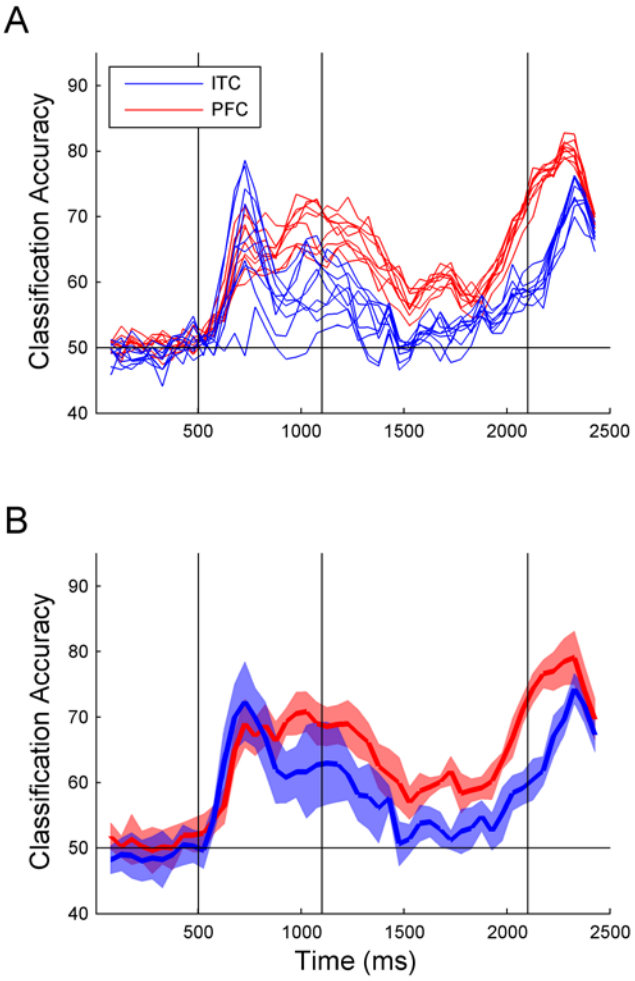
Supplemental figure 3.1 All 42 stimuli that were shown during the experiment. The images in the cat category are in the rows listed C1, C2, C3, and the images in the dog category are in the rows listed as D1, D2, D3. As can be seen, all the images look very similar, and it is not clear if the images in the cat category look more visually similar to each other than they look to images in the dog category (and vice versa for the dog category).



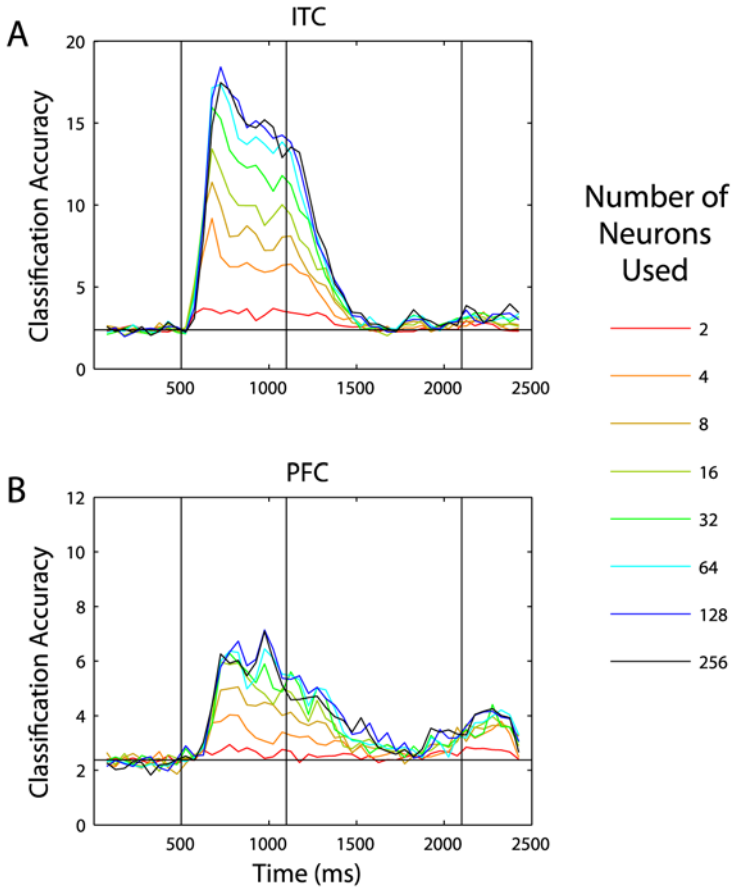
Supplemental figure 3.2 Comparison of decoding accuracy levels for three different classifiers for basic sample-stimulus category information, for ITC (A), and PFC (B). The magenta line is the classification accuracy obtained using correlation coefficient classifier, the orange line is the classification accuracy obtained using support vector machine (SVM) and the green line is the classification accuracy obtained using a Poisson Naïve Bayes classifier. As can be seen, while the mean accuracy level varies depending on which classifier is used, the trends over time remain the same, which gives us confidence that the conclusions we draw in this paper are not dependent on the classifier used since always compare results using the same classifier through the paper. It should be noted that the regularization parameter was not optimized for the SVM which could account for its overall lower accuracy level.



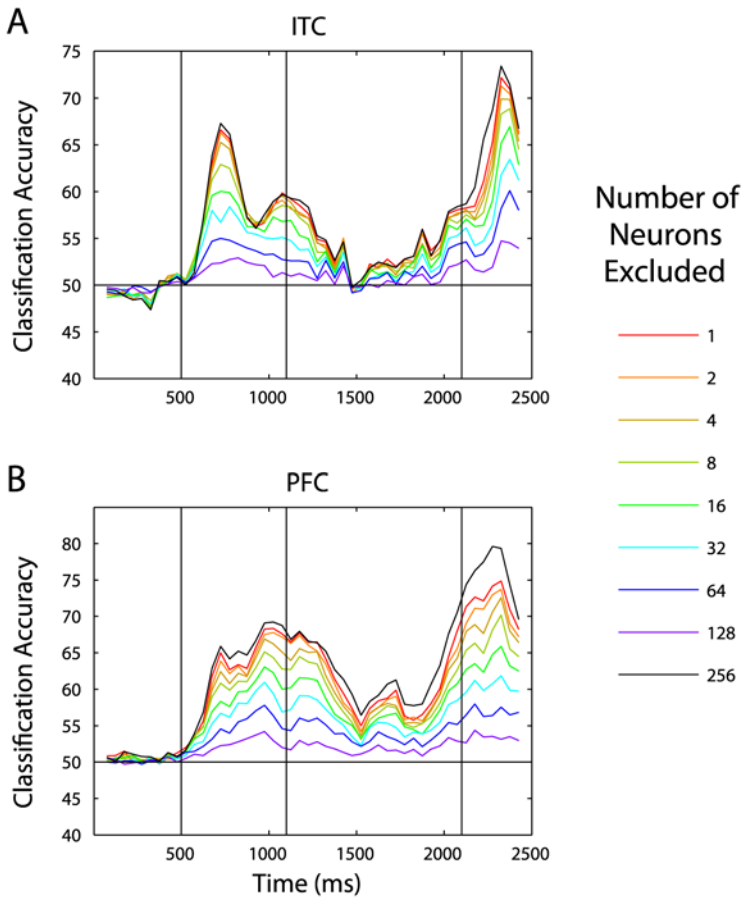
Supplemental figure 3.3 Illustration of how visual based stimulus information can lead to categorization decoding accuracy even when there is no abstract category information in the population of neurons. A, an illustration of 4 hypothetical neurons' responses to two images of dogs and two images of cats. Each neuron fires action potentials at a high rate to just one of images; thus each neuron can be thought of as being visually selective but not selective to the abstract categories. B, if training is done using trials from all when all 4 cat and dog images are shown, then one can obtain perfect cat/dog classification accuracy, even though these hypothetical neurons are only selective to visual features of the stimuli (and even though neural responses are noisy). C, if the training is done using responses from just one cat and one dog image, and the testing is done using responses to the other cat and dog images, then if the neurons are only respond to visual properties of the stimuli, classification performance will be at chance.



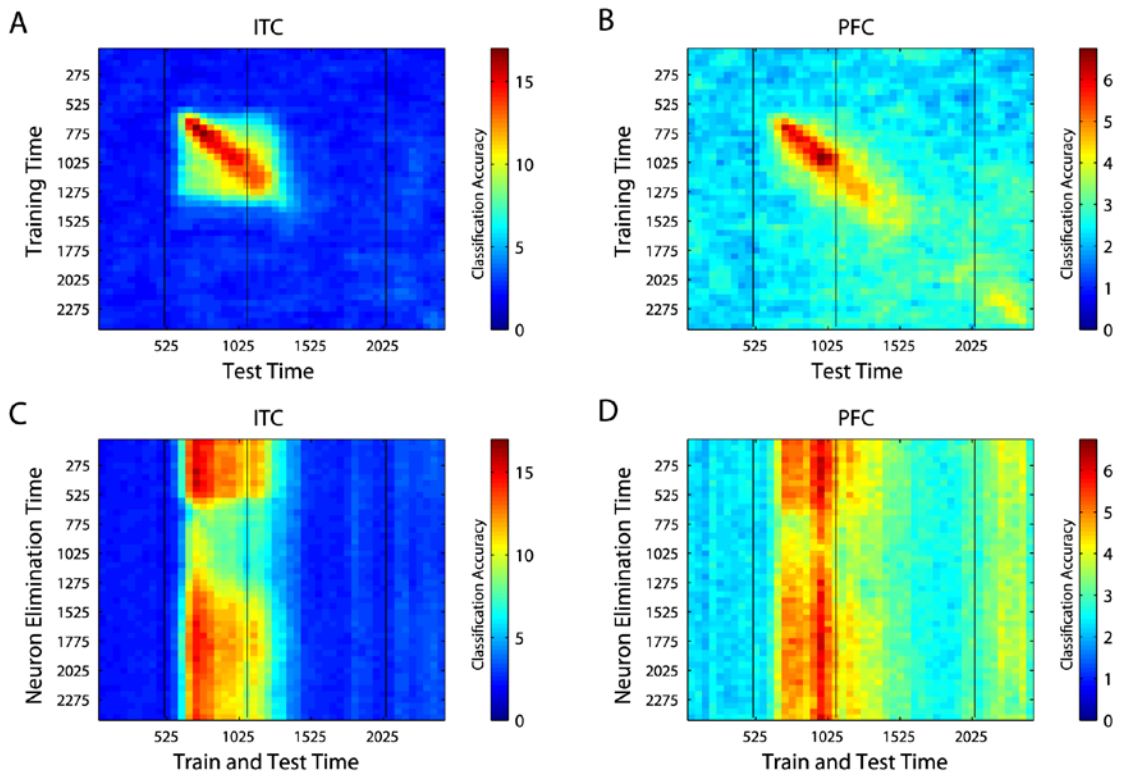
Supplemental figure 3.4 Supplementary data for the decoding of abstract category information. A, the 9 individual traces for decoding abstract category information with different permutations of training and test images; the mean of these 9 traces is what is shown in figure 3.3A. B, decoding of abstract category information excluding the morph images between the training and test prototypes. The results are very similar to those seen in figure 3.3A.



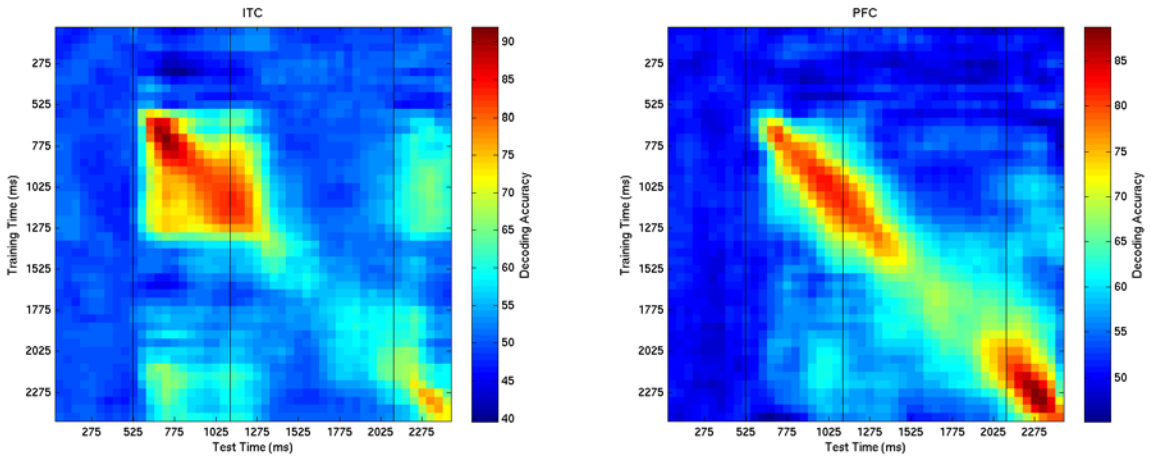
Supplemental figure 3.5 Readout of 'identity information' using the best 2, 4, 8, 16, 32, 64, or 128, compared to readout using all 256 neurons, for ITC (A) and PFC (B). As can be seen in A, identity information is less compact in ITC than abstract category information is (Figure 3.4), while for PFC the best 16 neurons seem to contain all the information in the population of 256 neurons for both abstract category information and the amount of identity information. As in Figure 3.4, the 'best' neurons were determined based on an ANOVA between cats and dogs using the training data. Due to the greedy manner the neurons were selected in, and the non-optimality of the selection method, the information represented in the subsets of neurons is an underestimate of how much information be present if the 'real' best n neurons were selected.



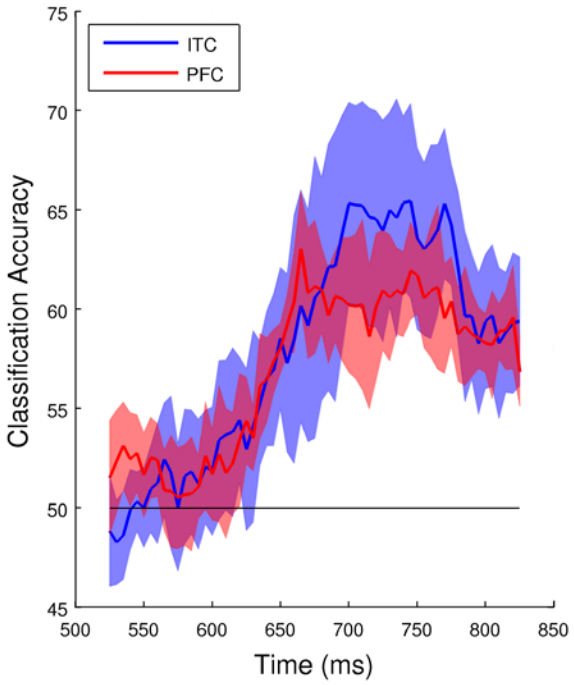
Supplemental figure 3.6 Readout results of abstract category information after excluding the "best" 1, 2, 4, 8, 16, 32, 64, and 128 neurons compared to decoding using all 256 neurons for ITC (A), and PFC (B). As can be seen, there is still information left in the population at most time periods for both IT and PFC even when the half of the best neurons have been removed.



Supplemental figure 3.7 Identity information is also coding by changing patterns of neural activity; although the code changes much less for identity information than for abstract category information (Figs. 3.4-3.5) A, B, decoding of identity information for ITC and PFC respectively, when training and testing using data from different time periods relative to stimulus onset (i.e., these plots are the same as Figure 3.6 except they show the decoding of identity information). Similar to figure 3.6, the results show that the best performance is along the diagonal, indicating a changing neural code with time. However during the sample period, the code for identity information ITC changes less than seen in the abstract category information case (Fig. 3.6A), as indicated by the green square area around the diagonal. C, D, decoding accuracies for identity information when eliminating the 'best' 64 neurons available at time period t1 (y-axis), and training and testing using all other neurons at time period t2 (x-axis), for ITC and PFC respectively (i.e., the same as Fig. 3.7, but for identity information). The 'best' 64 identity-selective neurons were determined by applying an ANOVA on the training set. As can be seen, there is some change in the 'best' identity neurons, however overall the neurons that contain identity information change much less with time than the neurons that contain the abstract category information (Figure 3.7).



Supplemental figure 3.8 Dynamic population activity for basic category information for ITC (left) and PFC (right). The same paradigm of training and testing at different times that was used in **Figure 3.6** is used here. As can be seen, there the best performance is achieved when training and testing at the same time, however for ITC, there is some transfer of performance when training during the STIMULUS-PERIOD, to testing during both the STIMULUS- and the DECISION- PERIODS.



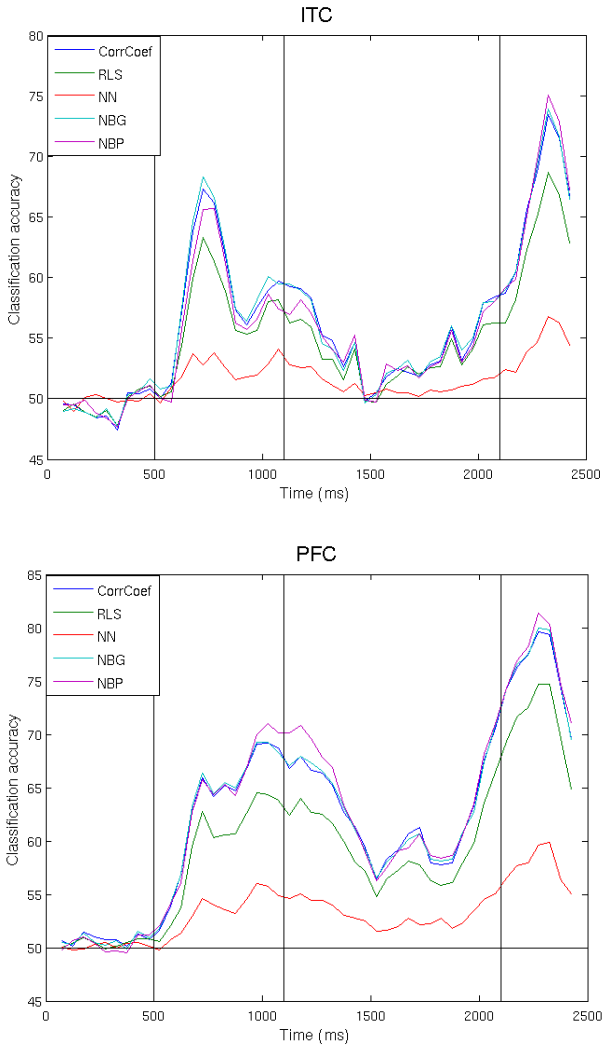
Supplemental figure 3.9 Finer time course of abstract category information in ITC (blue), and PFC (red). Results were obtain by decoding the abstract category information using a 50ms time bins, sampled at 5ms intervals, starting 25ms after sample-stimulus onset (525ms from the start of the trial). Between category morphs from the training and the test set were excluded for this analysis, because this extra visual information tended to make the results from ITC more variable (thus the results shown here are the same as the results shown in S4B, except with finer temporal resolution). As in figure 3.3 and in supplemental figure 3.4B, the results are the average over the 9 permutations of training and test sets, and the shaded regions are the standard deviations over the 9 permutations. Results from this figure show no clear latency difference between ITC and PFC for the presence of abstract category information.

Additional Supplemental Material

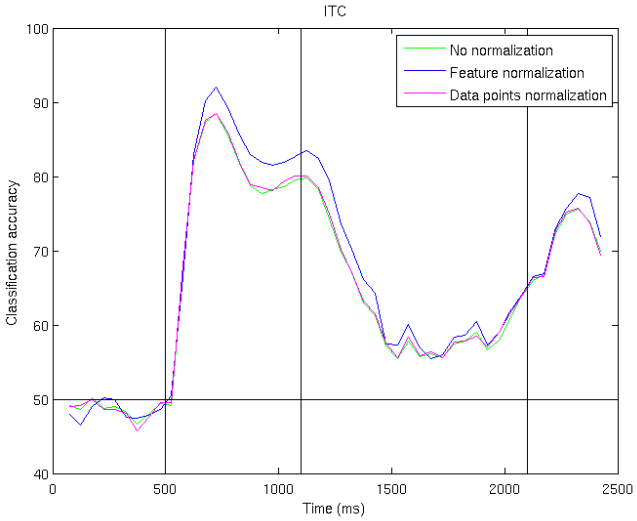
In addition to the results that was published in the 2008 Journal of Neurophysiology paper, I conducted several more decoding analyses on ITC and PFC responses during categorization tasks, that give further insight in the abstract category information, neural coding, and the robustness of population decoding methods. The results, which are shown below, are from: 1) additional web material that went with the original Journal of Neurophsyiology paper, 2) a 2009 Cosyne poster, and 3) figures based on analyzing additional data collected in the Miller lab by Jefferson Roy in a related study.

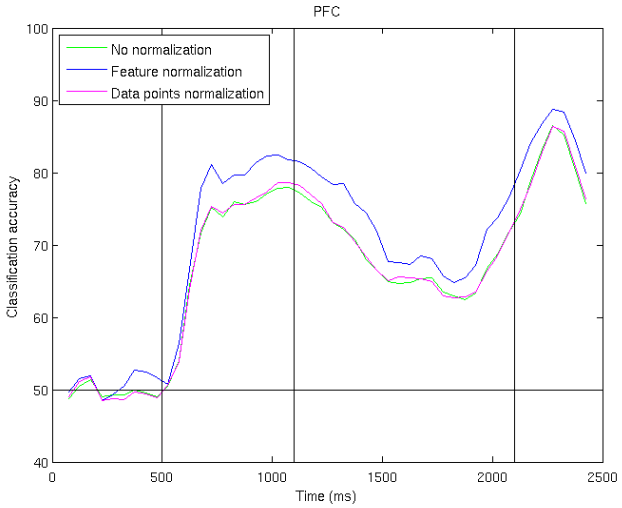
Additional 'web material' from Meyers et al., 2008, Journal of Neurophysiology paper

This additional 'web material' was put online at:
http://cbcl.mit.edu/people/emeyers/jneurophys2008/supplementary material/index.html, at the same time that the Journal of Neurophysiology paper was published. These results 1) highlight the robustness of the decoding method to choices of classifier and data normalization methods, and to drifts in firing rates, 2) give additional methods to examine abstract category information, 3) compare the abstract decoding results to computational vision features to show that the abstract category information is not likely to be due to properties of the visual stimuli, and 4) compare the decoding results to the 'category selective index' that was previously used to analyze data in Freedman et al., 2003 (which helps explan why the decoding methods found roughly the same amount of category information in ITC as in PFC early in the trial, while previous found more category information in PFC early in the trial).

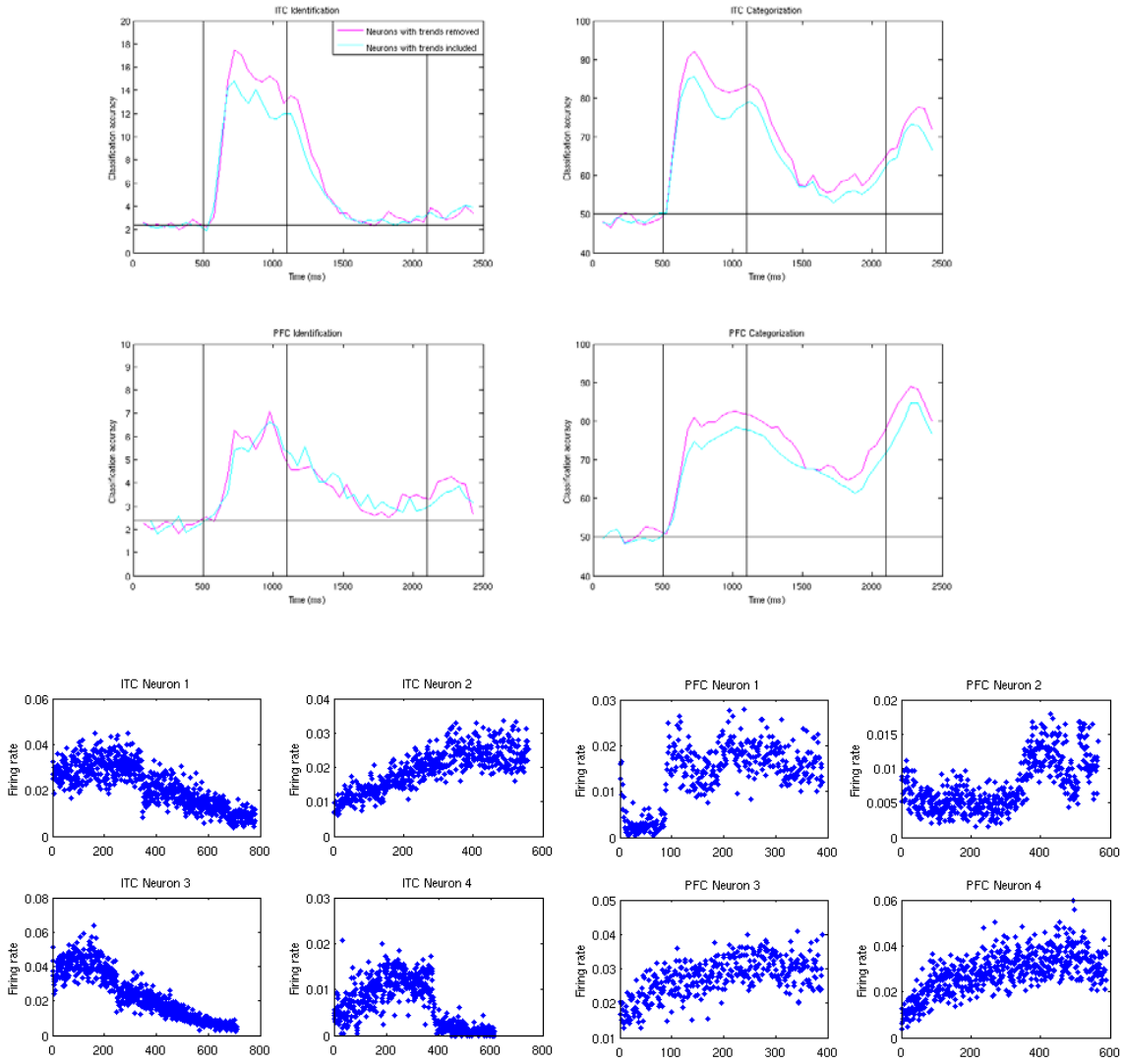


Additional supplemental material 3.1 Comparison of the decoding accuracy of 'abstract' category information using different classifiers. The results below show a comparison of decoding accuracies for five different classifier: correlation coefficient classifier (CorrCoef, blue), regularized least squares (RLS, green), nearest neighbor (NN, red), Gaussian Naive Bayes (NBG, cyan), and Poisson Naive Bayes (NBP, purple) for ITC (upper figure) and PFC (lower figure). These results are similar to those shown in Supplemental figure 3.2 except that here we are comparing more classifiers and we are decoding 'abstract' category information (as was done in Figure 3.3). As can be seen, the best performance is achieved with the CorrCoef, NBG, and NBP classifiers, RLS achieves slightly lower results and NN is by far the worst. However, for both areas and all the classifiers (apart from NN which had very poor performance overall), the general patterns of results is the same, which gives us confidence that the classifier choice is not affecting the conclusions drawn in this study.





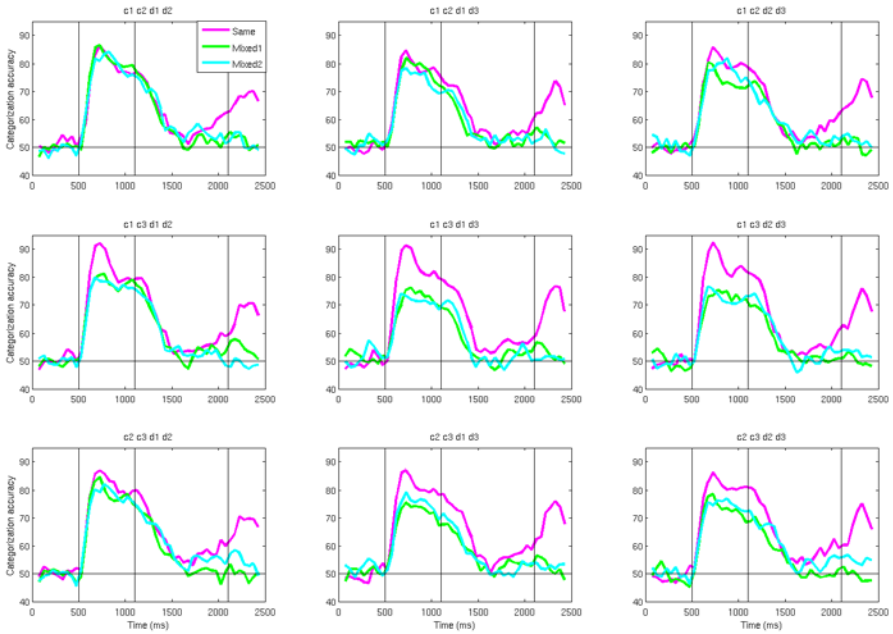
Additional supplemental material 3.2 Comparison of different data normalizations. The results below show a comparison of decoding accuracies when the data has been not been normalized (green line), when each feature has been z-score normalized (blue line), and when each data point has been z-score normalized (magenta line), for ITC (upper figure) and PFC (lower figure); by z-score normalization we mean that the data (i.e., feature or data point) has a mean of zero and a standard deviation of one. As can be seen, slightly higher results are achieved when each feature has been normalized (blue line); consequently this normalization was for all figures in the paper. The fact that z-score normalization of features increases decoding performance show that the best results are achieved when each neuron is contributing equally, since z-score normalizing of features makes all the firing rate of all neurons (averaged over all stimuli) the same; this reduces the impact of neurons that have high baseline firing rates, and increases the influence of lower firing rates neurons. All results are based on decoding basic sample-stimulus category information (the same type of information shown in Figure 3.2B). Data normalization parameters for the feature normalization (i.e,. mean and standard deviation) were gathered on the training set, and then applied to both the training and test data.



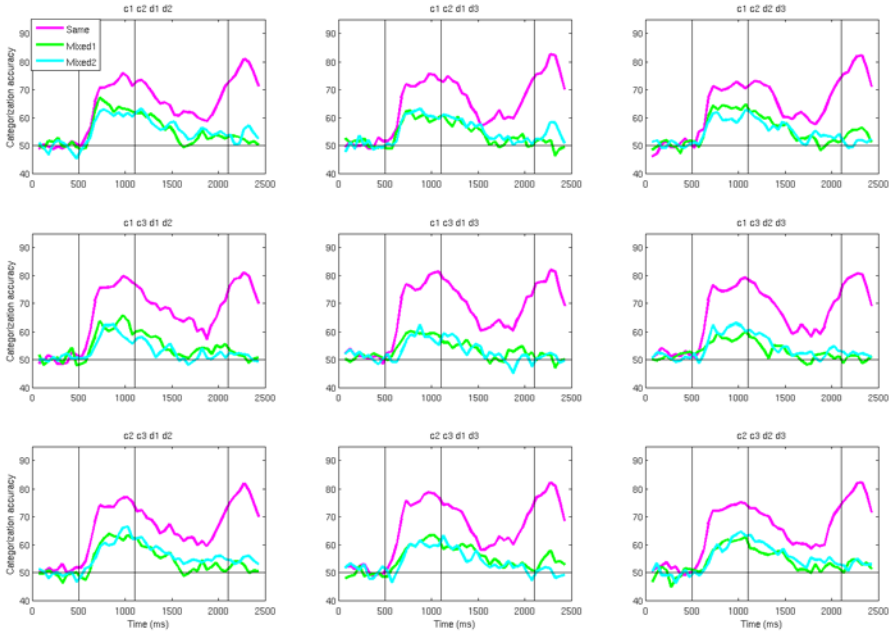
Additional supplemental material 3.3 Decoding including and excluding neurons that have temporal trends over trials. One of the preprocessing stages in the readout analysis was to remove all neurons that showed temporal trends over trials. Neurons with nonstationary trends were defined as those neurons that had an average firing rate variance in 20 consecutive trials that was less than half the trial firing rate variance over the whole session. Because the stimuli were presented in random order, the average variance in 20 tirals should be roughtly equal to the variance over the whole session (and deviations from this suggest that there could be artificats in the firing rate due such factors are movement in the recoding electrode). 42 ITC and 34 PFC neurons met this trend criterion, and were excluded from the decoding analyses in the paper. Below shows the results for decoding identity and category informatoin from ITC and PFC when these trends neurons were both included and excluded. As can be seen, the decoding accuracies are usually slighly lower when the neurons with trends are included, however overall the pattern of results is very similar. Also shown below are 4 random ITC and 4 random PFC neurons that were considered 'trend neurons' as defined by the criterion listed above (in order for the reader to get a sense of what the trends in the firing rates looked like).

Another method to examine 'Abstract' Category information

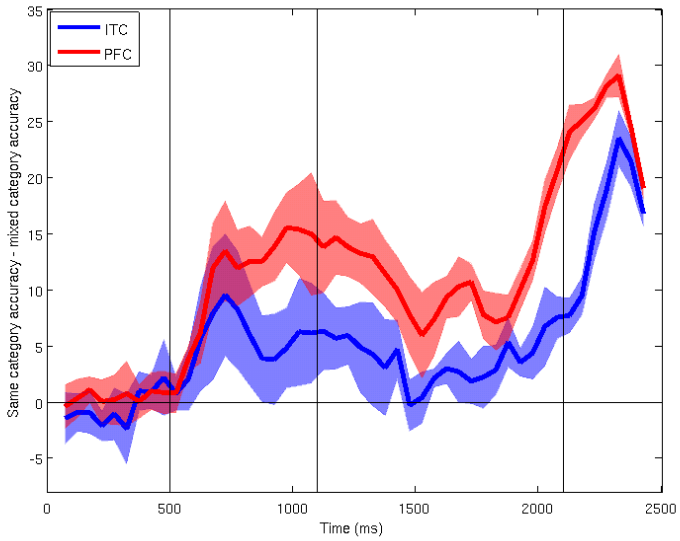
ITC results comparing within category accuracy vs. mixed category for the 9 permutations of the prototypes



PFC results comparing within category accuracy vs. mixed category for the 9 permutations of the prototypes

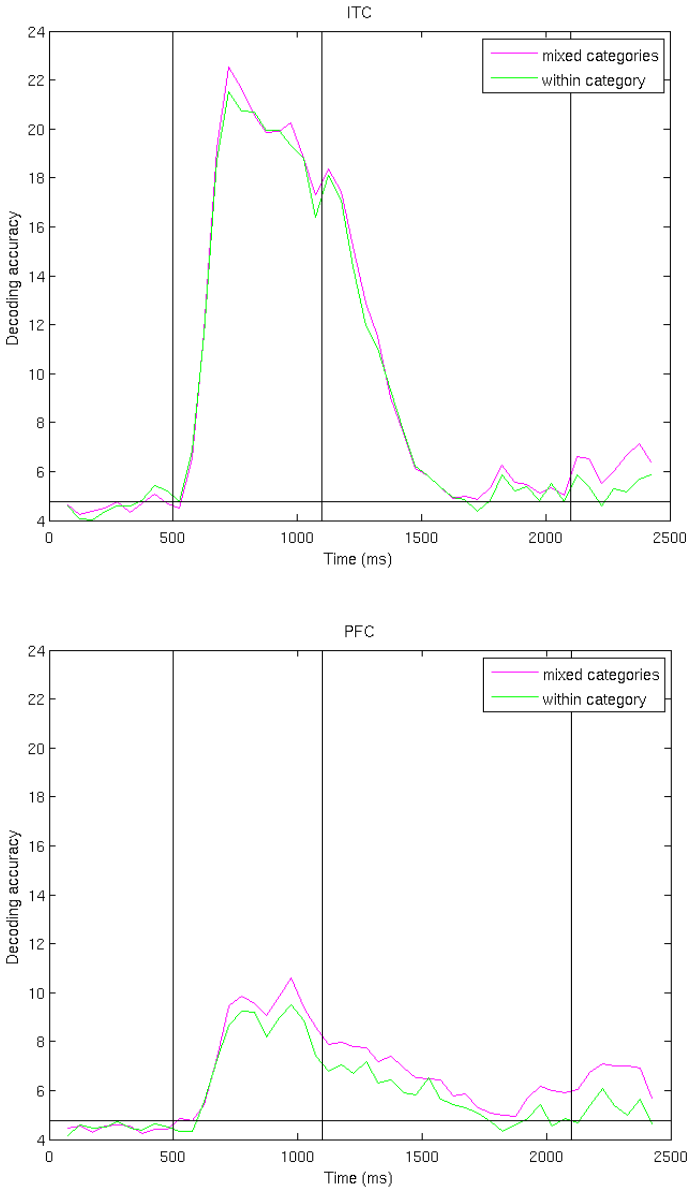


Average of within category accuracy minus the mixed category decoding accuracies

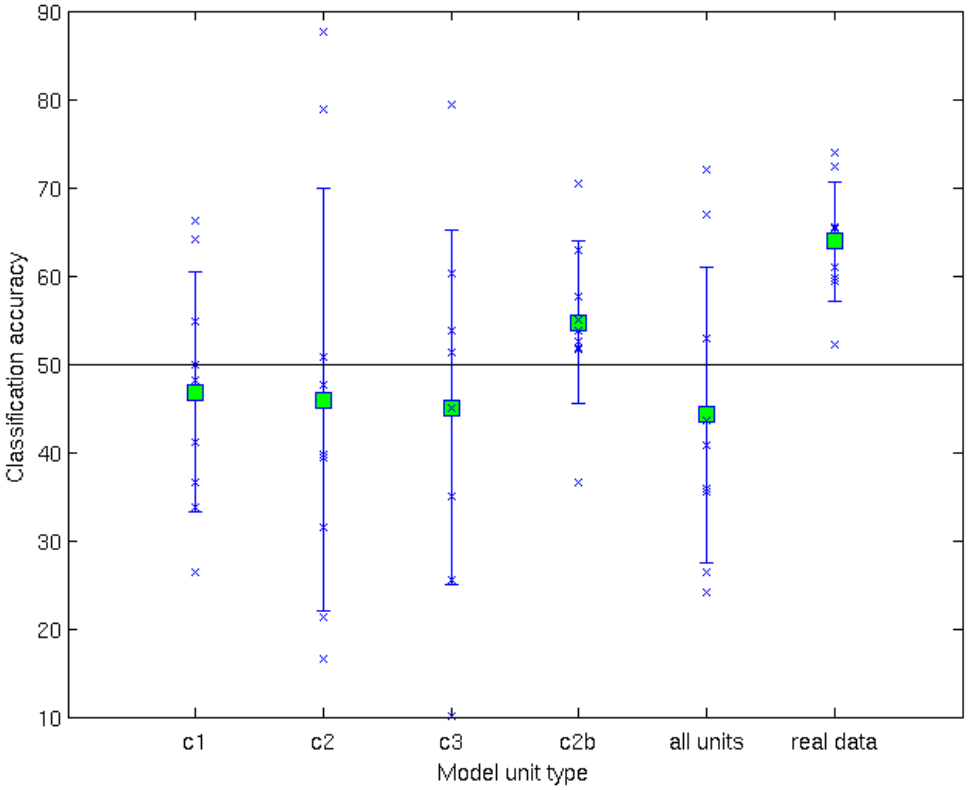


Additional supplemental material 3.4 An alternative 'abstract' category readout: within category minus mixed category accuracy. In order to further evaluate whether there is any 'abstract' category information that is possibly separate from visual properties of the stimulus, we designed an analysis in which we trained and tested a classifier using two dog and two cat prototypes, and compared the results from the within category readout (e.g., [c1 c2] vs. [d1 d2]), to the decoding accuracies obtained in the two 'mixed' category readout conditions (e.g., [c1 d1] vs. [d1 d2], and [c1 d2] vs. [c2 d1]). Results from all of the 9 permutations comparing the same vs. the two mixed conditions for ITC and PFC are shown below. For PFC, in all 9 permutations, the within category accuracy (magenta trace) is much higher than the mixed category accuracies (green and cyan traces), while in ITC the within category information is either higher or equivalent to the mixed category accuracies when the stimulus is visual, and the within category accuracy is always higher in the response period. Given the large amount of visual information in ITC, it is not surprising that for some of the conditions while the visual stimulus is being shown, the within category accuracy is equivalent to the mixed category information because for certain prototypes the stimuli in the opposite category are probably more visual similar than the stimuli within the category. However the fact that within category accuracy is never lower than the mixed category information is likely a result of additional 'abstract' category information contained in the firing rates of these neurons. For ITC during the delay period (and PFC at all time points), visual information is not present in these areas, and thus only the influence of 'abstract' category information is seen in the decoding accuracies.

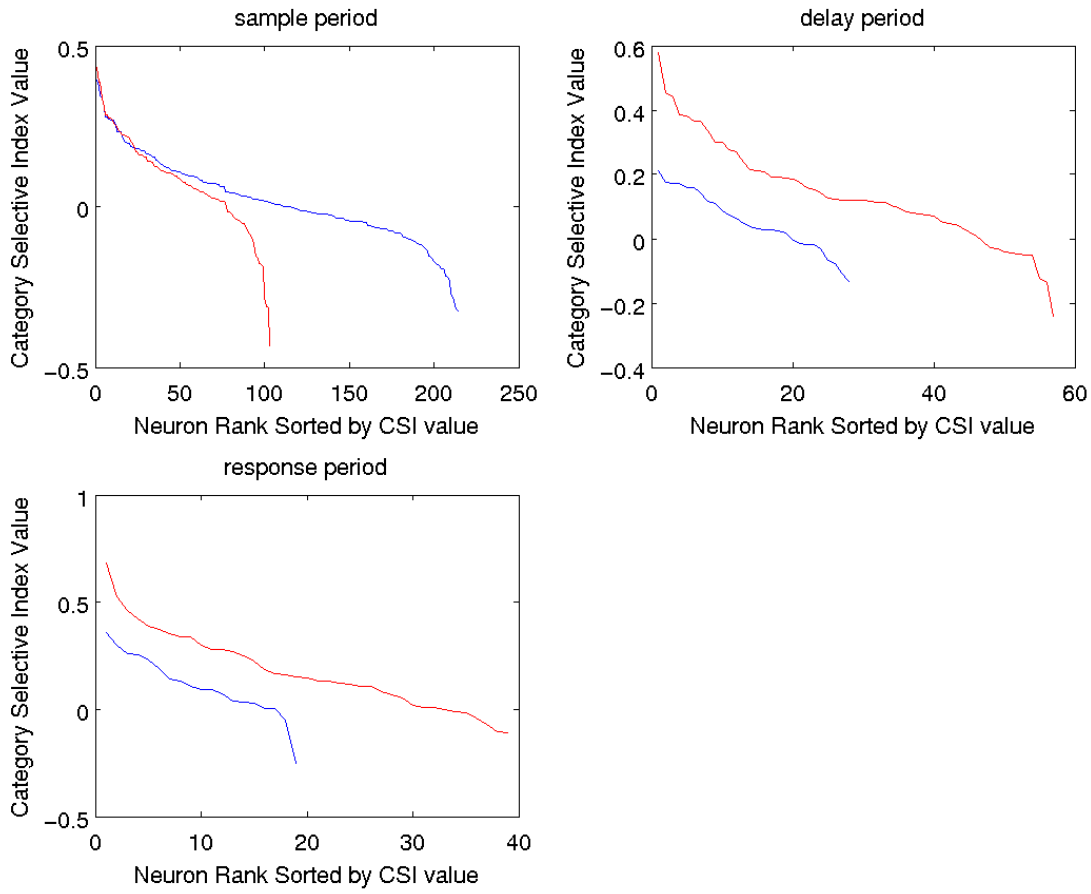
We have also created an image that summarizes the amount of 'abstract' category information in these areas using the above measure. To do this we took each of the 9 permutations of the within category decoding accuracies and subtracted from them the average of the two mixed category decoding accuracies. We then averaged over these 9 permutation and plotting the mean and standard deviation from this averaging. The results are also shown below, and examining them we see that initially ITC has about the same amount (or possibly slightly less) abstract category information as PFC, however later in the trial PFC has more abstract information than ITC, which largely confirms the results shown in Figure 3.3A of the paper.



Additional supplemental material 3.5 Identity decoding for images within the same category vs. mixed categories. The figures below show the results from decoding the identity of images using images derived from three prototypes at a time (as proposed by one of the anonymous reviewers of the paper). A comparison is made between the average of this decoding for the within category prototypes ([C1, C2, C3] and [D1, D2, D3]) (green line) to the average of the 18 other mixed category permutations ([C1 C2, D1], [C1 C2 D2], ..., [C3 D2 D3]) (magenta line). Since the readout is for the identity of the 21 images, chance is 1/21. As can be seen in the figures, for both PFC and ITC there is a significant amount of visual information present, and the potential presence of 'abstract' category information in the within-category condition does not seem to significantly degrade the ability of the classifier to decode the 'visually based' identity information.



Additional supplemental material 3.6 Decoding category information from the Model units of Serre et al., (2007). To test if the 'abstract' category information described in Figure 3.3 could be accounting for by visual image properties of the images, we applied the decoding methods to simulated neural responses created from the Model units described in Serre et al., 2007. Decoding results from training on images derived from 2 dog and 2 cat prototypes and testing on the remaking cat and dog prototype (as was done in Figure 3.3) are shown below for several different Model types. The blue x's are the results from the 9 permutations of training and test prototype splits, the green boxes are the mean from these 9 runs, and the error bars are the standard deviations. The right most column contains results from decoding the ITC neural data using one 150ms bin staring 100ms after stimulus onset (i.e., 600-750ms into the trial, where stimulus onset is at 500ms). As can be seen, the neural data achieves a higher decoding performance than all the Model units types, and perhaps more significantly, results from the decoding the neural data are always above chance for all 9 permutations of the data. These results suggest that the results reported in Figure 3.3 are due to ITC having more 'abstract' category information that is not directly inherent in the visual properties of the stimuli.



Additional supplemental material 3.7 Is there category information in ITC during the sample period? A comparison of decoding analyses and the Category Selective Index analysis of Freedman et al. 2003. In the original paper of by Freedman et al. (2003) analyses of the data suggest that there is no category information in ITC during the sample period, and that category information does not appear in ITC until the delay period. However in our new decoding based analyses, we find that there seems to be a significant amount of 'abstract' category information in ITC early in the sample period. In order to understand the different conclusion obtained by these different analyses, we reexamined the category selective index (CSI) used by Freedman et al. (2003) in more detail. The first difference between the original CSI analysis and the decoding analysis, is that in the CSI analysis, the firing rates were taken over much longer time periods; i.e., the sample period in the original analysis used firing rates averaged over 600ms starting 100ms after stimulus onset and ending 100ms after stimulus onset, compared to the decoding analysis in which firing rates in 150ms sliding bins were used. From examining the decoding results (Figure 3.3), it is clear that while early in the sample period decoding accuracies for abstract category information PFC and ITC approximately the same, later in the sample period (and for the rest of the trial), PFC has a larger amount of abstract category information than ITC. Thus by using large time windows of analysis, the fact that ITC and PFC initially have the about same amount of abstract category information could not be seen in the original CSI analysis.

A second factor that contributed to this discrepancy in results is that in the original CSI analysis an ANOVA was first run to determine which neurons were visually selective, and then the CSI values from only these visually selective neurons were used in the subsequent statistical analyses. However since ITC has many more visually selective neurons that are not category selective (particularly during the sample period when the stimuli were being shown), the results were biased by including a larger number of neurons in ITC in these analyses. Above is a plot of the CSI values for ITC and PFC using exactly the same parameters that were used in the original Freedman et al. paper (i.e., using only visually selective neurons as determined by and ANOVA, and longer time bins), but we have sorted the CSI values and then plotting them as a function of their rank order of their category selectivity as determined by the CSI value. As can be seen from the plot for the sample period, the neurons that have the highest CSI scores have approximately the same values for ITC and PFC. However for ITC, there is a long tail of neurons that have CSI values that are close to zero (which are due to the large number of visually selective neurons that are not category selective). Thus when statistics are done using all visually selective ITC neurons, the long tail of non-category selective neurons biases the results towards zero, making it seem like the population as a whole seem non-category selective. This should be contrasted with the results from the delay and response periods in which the highest CSI values are larger for PFC than for ITC, and during these time periods the decoding analysis agrees with the CSI values in stating that there is more abstract category information in PFC than ITC. It should also be noted that the decoding analyses that used feature selection (see section title 'compact and redundant information' and Figure 3.4), show that most of the abstract category information is contained in a small subset of neurons, thus calculating statistics based on larger populations can lead to incorrect conclusions. Finally, if a t-test is run between the CSI values in ITC and PFC using all the neurons (not just the ANOVA visually selective neurons) during the sample period (and even using the same larger time bins used in the original paper by Freedman et al.), the p-value is .058, which fails to meet the typical alpha level of .05 (and this value would probably be even larger only the first half of the sample period was used in which the decoding analysis indicates there is no difference in the amount of category information between ITC and PFC)

Analyses from Meyers et al., Cosyne 2009

The following results are taken from a 2009 Cosyne poster:

Meyers, E., Freedman, D., Kreiman, G., Miller, E., Poggio T. Decoding dynamic patterns of neural activity using a 'biologically plausible' fixed set of weights. Computational Systems Neuroscience, 2009

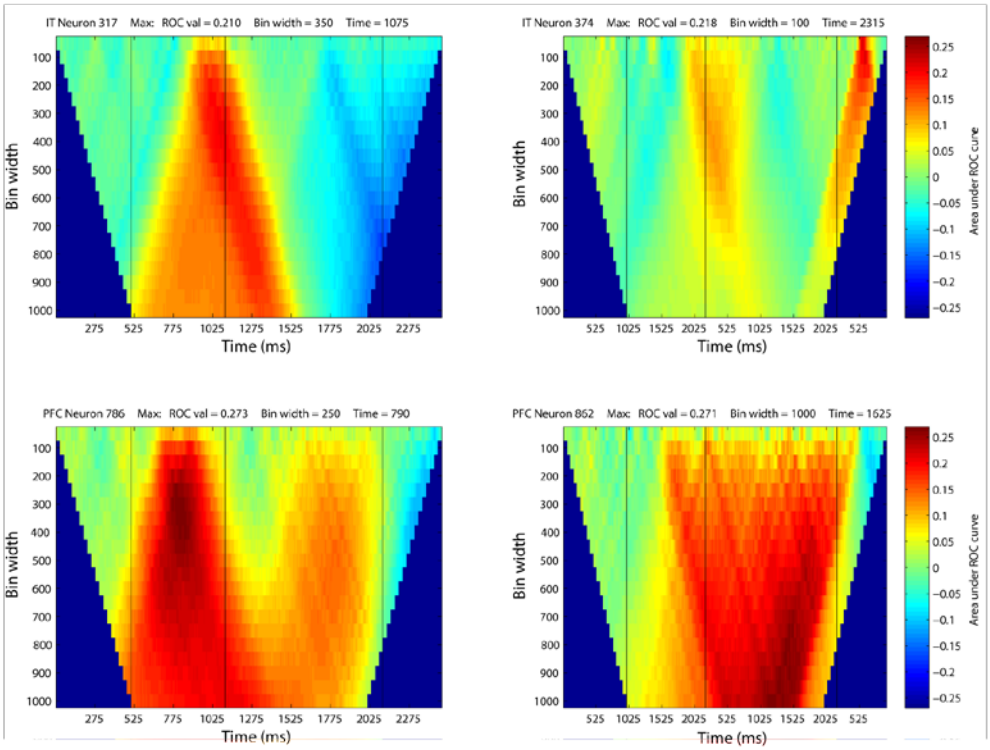
The purpose of the poster was to examine possible ways that downstream neurons could decode information from an upstream area if information in the upstream area is contained in population activity that is changing dynamically. The conclusion of the poster was that approximately 75-80% of the information that can be extracted using a more complex coding scheme that has dynamically changing weights, can be extracted with using a simplier, more 'biologically plausible', fixed set of weights (that was determined by training a classifier using the average firing rate over the whole course of a trial); thus even though the populations of ITC and PFC neurons contained abstract information in a dynamic manner, it still was possibly for a downstream area to extract the majority of information using a simple decoding scheme. In retrospect, however, we view the fact that certain types of information are coded dynamically might be an indicator that more complex processing is occurring, and to only examine how this information could be extracted is most likely glossing over some potentially interesting insights about how information is processed in the brain. Regardless of the interpretation, several of the findings related to the Cosyne poster were quite interesting, so they have been included them below.

Biologically implausibility of systematically changing synaptic weights

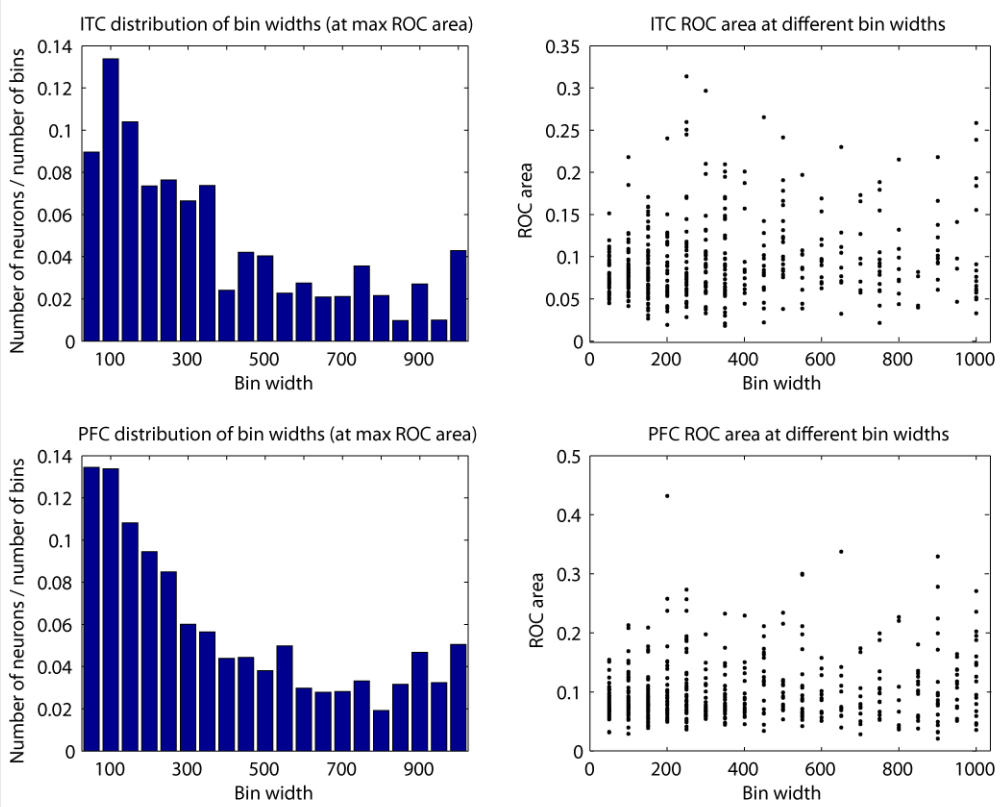
If the weights learned by a classifier are interpreted as synaptic strengths (as is commonly assumed in neural network interpretations of decoding algorithms, (see Additional supplemental material 2.1) then changing the weights of a classifier dynamically creates a biologically unlikely model whereby the synaptic strengths between neurons change within the time-course of a single trial (consistently across trials of a given type), in a way that is time locked to the onset of the stimulus. Below we examine how the brain could exploit the information that is contained in a dynamic population code.

Examining firing rate interval lengths that contain maximal information

Before assessing how effective different decoding schemes are at extracting information from dynamic populations of neurons, we examined the neurons' firing rate bin size and latency that contained the largest amount of information for discriminating between dog and cat stimuli using an ROC analysis (see Additional supplemental material 3.8). Results show that there is a large variation across the population of neurons for both the duration and the latency of selectivity (see Additional supplemental material 3.9).



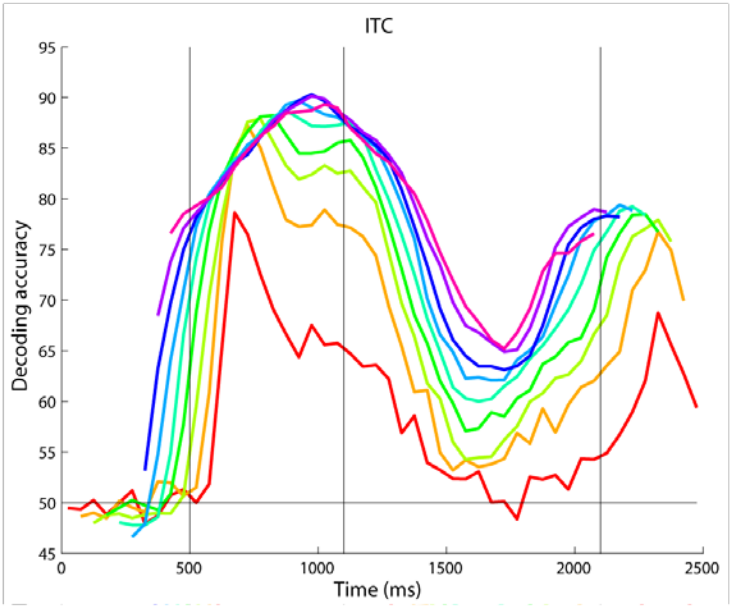
Additional supplemental material 3.8 Finding the bin size and latency that contains maximal discrimination between categories using an ROC analysis for four example neurons. The area above or below the ROC curve was calculated using firing rates calculated over different bin sizes (y-axis), and at different latencies (x-axis). For each neuron, we measured this ROC value every 5ms using bin sizes that ranged from 50ms to 1000ms. We then calculated the maximum ROC value for each neuron, the latency when this maximum ROC value occurred and the bin width at which the maximum ROC value occurred.

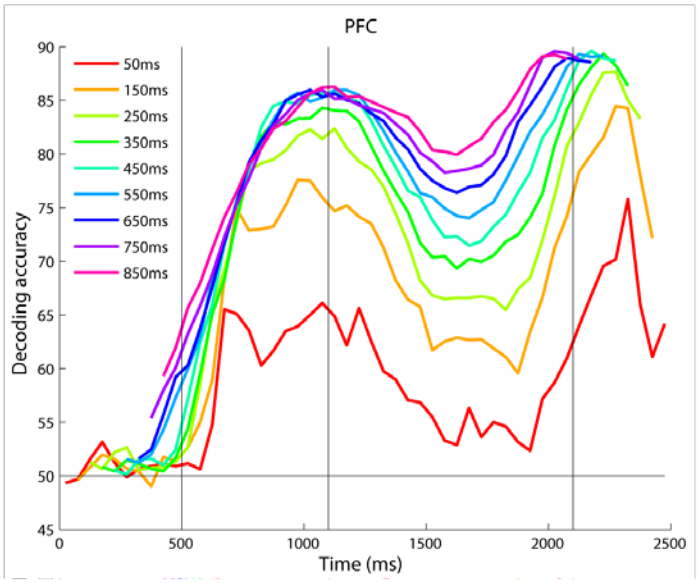


Additional supplemental material 3.9 Distribution of ROC area bin width sizes (left), and ROC areas as a function of latency (right). Population statistics derived from these ROC values show that there is a wide range of optimal bin widths for different neurons, although overall there are more neurons with small bin widths than with large bin widths (left plots). Also, there was no consistent relationship between the magnitude of the ROC values and neurons' optimal bin widths (right plots).

Population decoding using different bin sizes

Given that the optimal bin width varies widely between different neurons, we examined how bin size affects decoding accuracy when a population of neurons is used. Results from this analysis revealed that the highest decoding accuracies occurred when bin sizes of 350ms or larger are used (Additional supplemental material 3.10). The reason larger bin sizes lead to higher decoding accuracies is most likely due to more neurons with short time windows of selectivity being included within these larger integration windows.

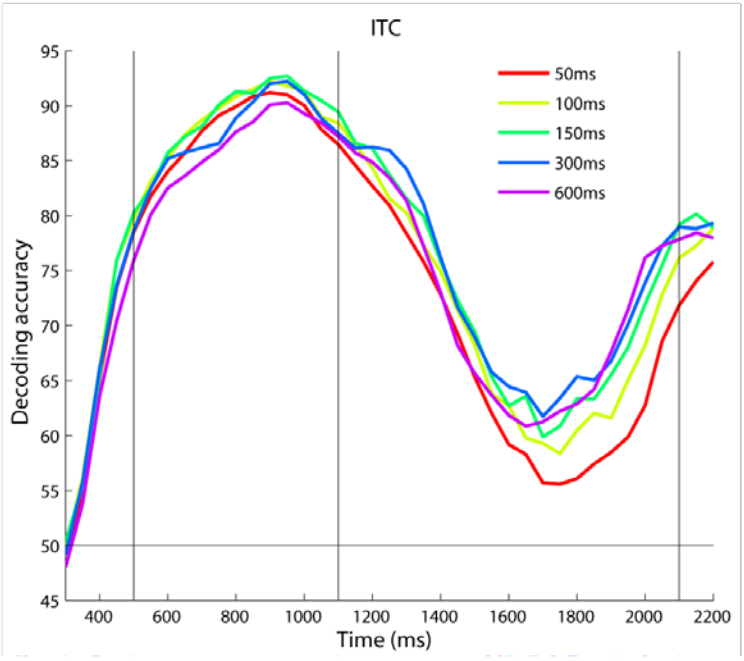




Additional supplemental material 3.10 Decoding accuracies in ITC and PFC using different bin sizes. Training and testing was done at the same time, at 50ms intervals, using the firing rates in different bin sizes centered at the time shown above in the figures. Best decoding accuracies were obtained for bin sizes greater than 350ms.

Weighting sub-bins within a 600ms sliding time window

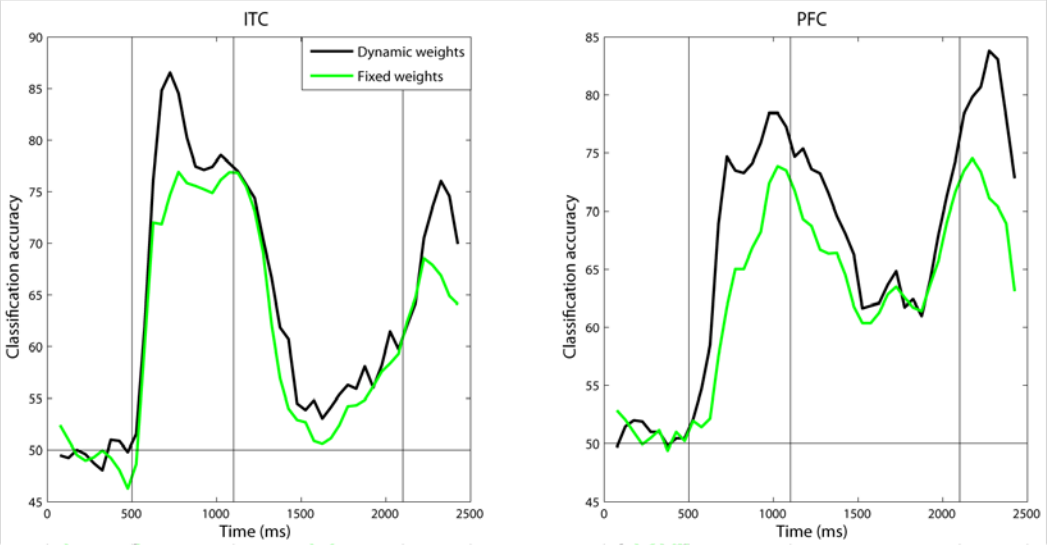
Next we examined whether a higher decoding accuracy could be achieved by having separate weights on smaller sub-bins which are then combined within a larger integration window. To do this we used a 600ms integration window, and we compared the results from using twelve 50ms bins, six 100ms bins, four 150ms bins, two 300ms bins or one 600ms bin. Overall the results showed that using separate weights for smaller sub-bins did not lead to a large increase in decoding accuracy (see Additional supplemental material 3.11). We believe the reason that having more weights on smaller sub-bins did not lead to a large improvement in accuracy is due to the fact that spiking activity outside neurons' windows of selectivity does not drastically interfere with the selectivity established within neurons' windows of selectivity.



Additional supplemental material 3.11 Decoding accuracies using different sub-bin sizes within a 600ms integration window. Results show that allowing different weights for sub-bins within a larger 600ms window did not improve decoding accuracy, despite the fact that the population code was dynamically changing in time. We interpret this finding as due to the fact that activity outside a neuron's maximal discrimination window did not dramatically interfere with the selectivity established within a neuron's maximal discrimination window (it should also be noted that having more training data could potentially change these results, although given there were 50 points per class, this does not seem too likely).

Decoding dynamic patterns using a fixed set of weights

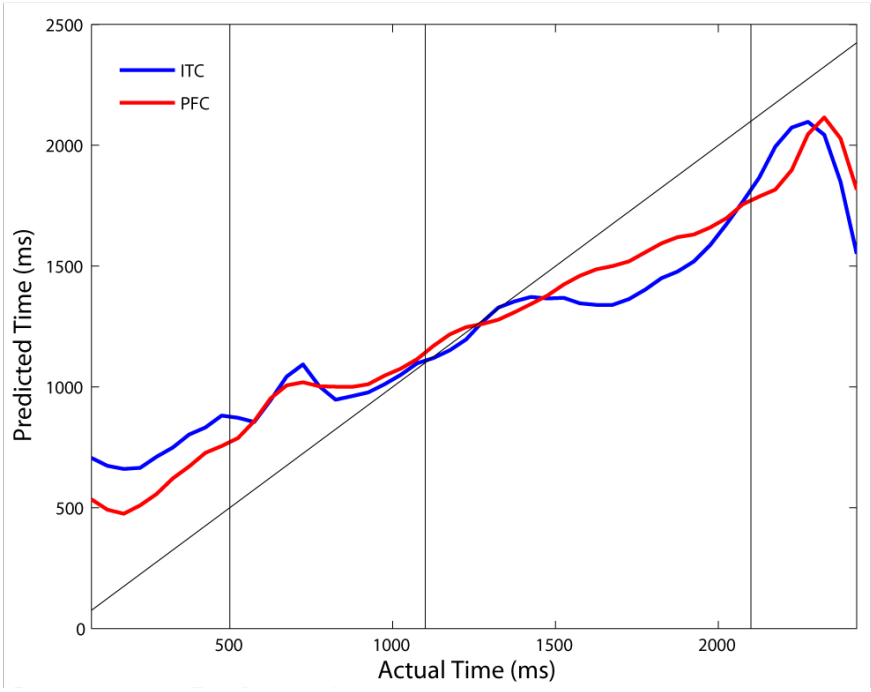
We also compared the decoding accuracy using dynamic weights based on 150ms sliding bins, to using a fixed more 'biologically plausible' set of weights based on training the classifier using a 2000ms bin starting at stimulus onset (see Additional supplemental material 3.12). As can be seen, the fixed set of weights led to a decrease in performance at most time periods. However, overall from stimulus onset to the end of the trial, the classifier based on a fixed set of weights still did 80% as well the dynamic weight classifier in ITC, and 76% as well as the dynamic weight classifier in PFC. Thus while clearly dynamic weights are capable of extracting more information, the fixed set of weights is still capable of achieving a high level of decoding accuracy.



Additional supplemental material 3.12 Comparison of training a classifier using dynamic weights (black trace), to training using one fixed set of weights for all time points (green trace) for category information decoding. The results from the dynamic weights are based on a 150ms sliding bin training paradigm (as as was used for Figure 3.2). The results for the fixed set of weights are based on training a classifier on 2000ms of data that start at the time of stimulus onset. The results show that decoding accuracy is higher when dynamically changing weights are used, but that good performance is still achieved using a fixed set of weights.

Decoding the time since stimulus onset

Finally, we examined if we could decode the time since the start of a trial using a regularized least squares regression algorithm. The results show that it is indeed possible to decoded the latency from stimulus onset in ITC and PFC at above change levels of accuracy (Additional supplemental material 3.13).



Additional supplemental material 3.13 Decoding the time since the start of a trial using data from ITC (blue trace) and PFC (red trace). Results are based on using a regularized least squares regression algorithm to predict the time since the start of a trial. The fact that both the curves for ITC and PFC slope upward indicate that it was possible to decode the time since the start of a trial at level above chance (regardless of which type of stimulus was shown).

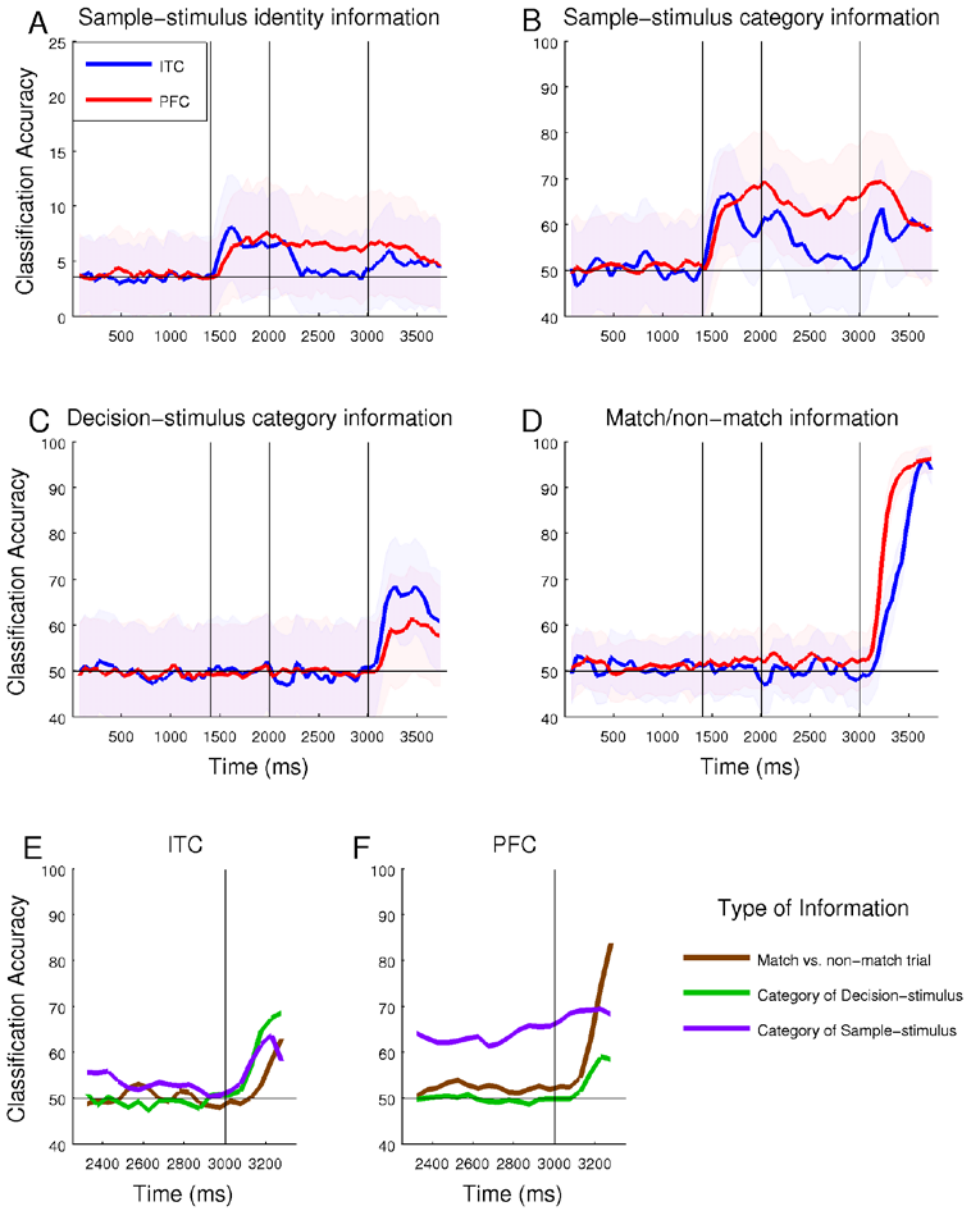
Additional analyses of abstract category information using data from Roy et al., 2010

One of the main concerns about the analysis in Figure 3.3, is that the results are not due to populations of ITC and PFC neuron's grouping the stimuli into abstract categories, but rather it is an artifact of visual properties of the stimuli. In particular, if the three cat prototypes are more visually similar to each other than they are to the three dog prototypes (and conversely, if the three dog prototypes are more visually similar to each other than they are to any of the cat prototype images), then the decoding analysis would indicate that there is indeed abstract category information, even if there was not any present. While the analyses examining computational model units suggests this is not the case (see Additional supplemental material 3.6), having further verification using actual neural data would lead to much more confidence in the interpretation of these findings. Fortunately, data collected in recent study by Roy et al. (2010), has the potential to address this issue.

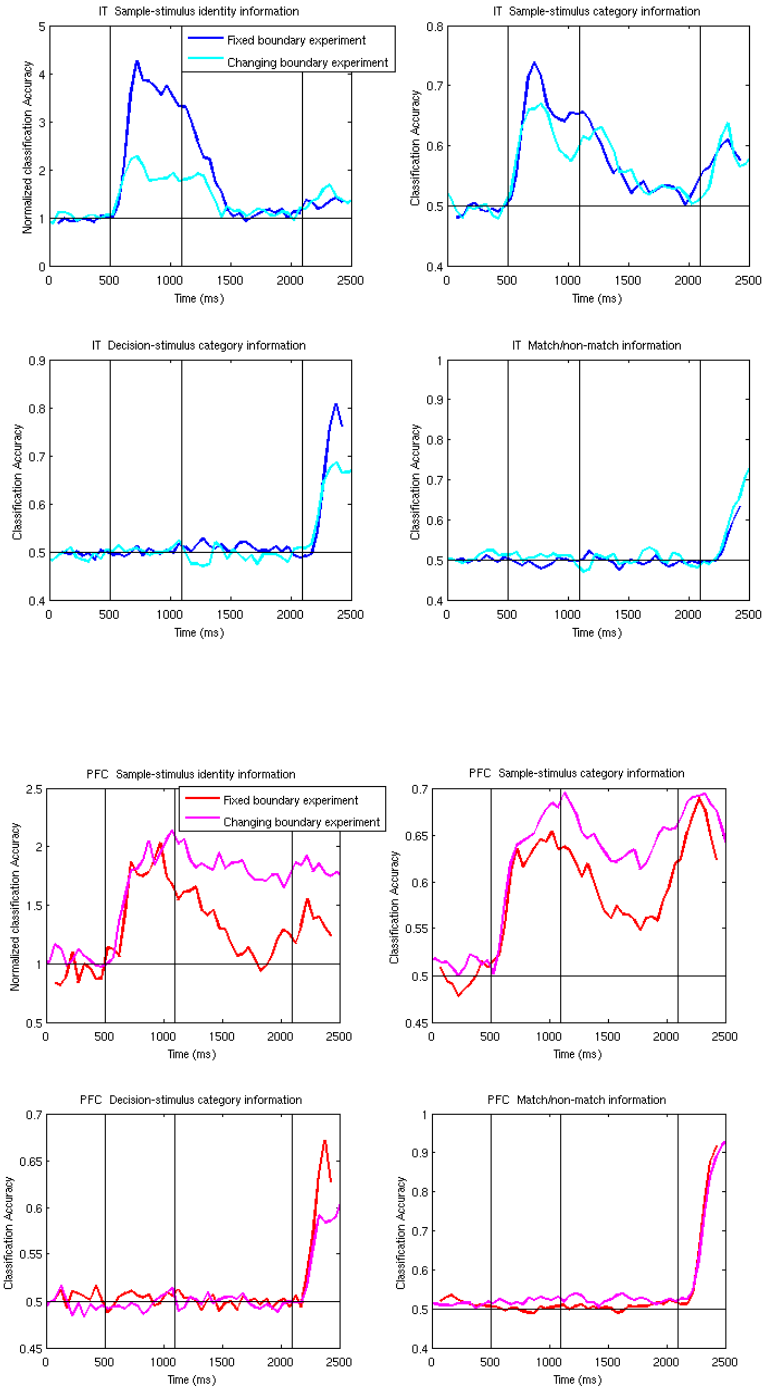
The study of Roy et al. (2010) was very similar to the Freedman et al. (2003) study (both studies were conducted in Earl Miller's lab at MIT). Monkeys were trained to discriminate between morphs from 2 cat and 2 dog prototypes, using an almost identical experimental design as Freedman et al. (2003). The one critical difference, however, was that in Roy et al. (2010), the category boundary that indicated which stimuli should be grouped together changed from trial to trial, based on a cue that was given at the start of the trial (i.e., one type of cue indicated that cat1 prototype derived images should be grouped with cat2 prototype derived images, while the other cue indicated that cat1 prototype derived images should be grouped with dog1 prototype derived images). Because of this changing category boundary, it was now possible to redo the decoding analyses for abstract category information separately for each category boundary and compare the results. If higher decoding accuracies are obtained for the task-relevant category grouping for both category boundaries, then this would indicate that indeed ITC and PFC have abstract category information (since if it were the case that a higher measure of abstract category information was obtained due to the visual similarity in the

prototype images, this would lead to a lower classification accuracy for the other category boundary).

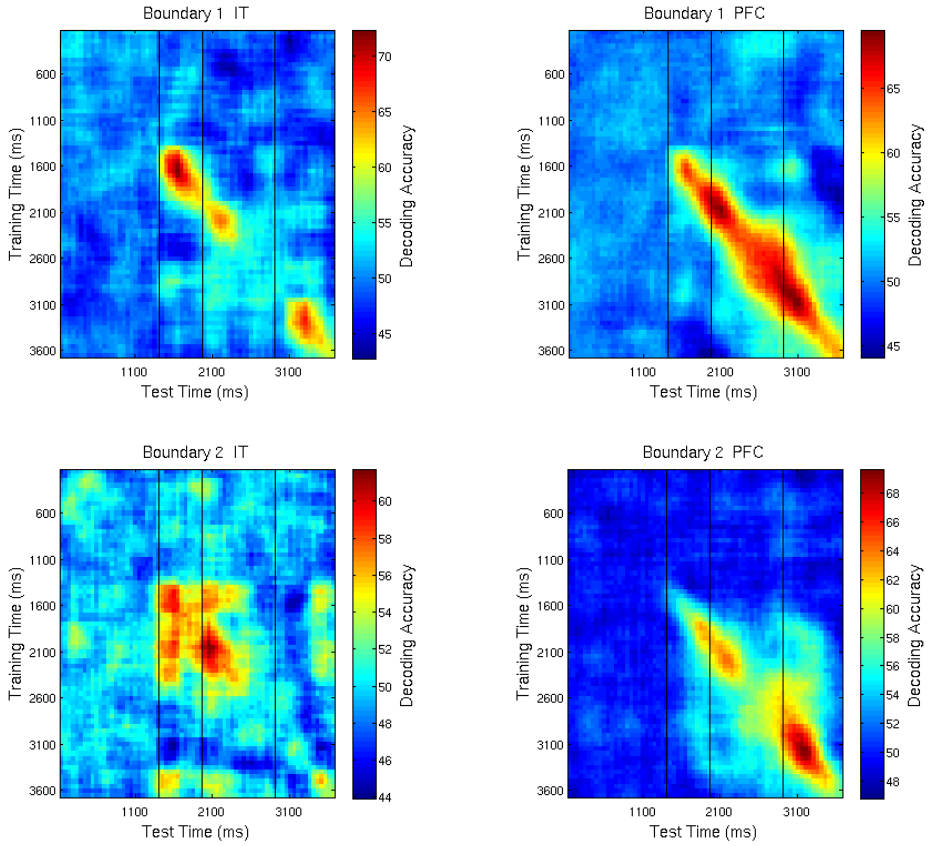
Before conducting the abstract category analyses on this data, we first replicated some of the other findings of Meyers et al. (2008). Additional supplemental material 3.14 shows a replication of Figure 3.2 using the data from Roy et al. (2010). As can be seen the results between the two studies look very similar (although the decoding accuracies from Roy et al. (2010) as slightly lower due to the fact that fewer neurons were used in these decoding analyses). We also directly compared the decoding accuracies from the two studies on the same plot using the same number of training and test points in both studies (see Additional supplemental material 3.15), which again show that the results are very similar. Finally, Additional supplemental material 3.16 replicates the dynamic population coding results of Figure 3.6 by training and testing the classifier at different type periods (Additional supplemental material 3.16 shows the dynamics for basic category information rather than abstract category information).



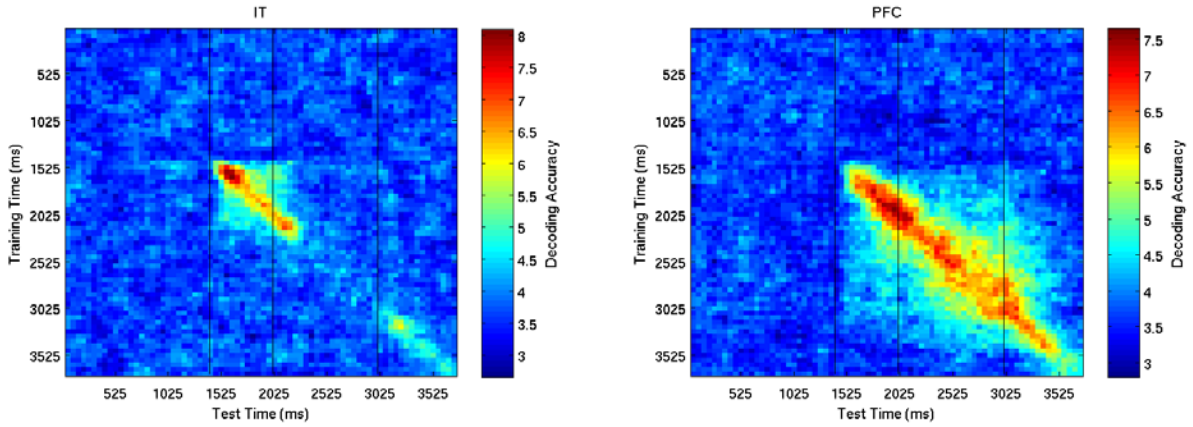
Additional supplemental material 3.14 Replication of Figure 3.2 using data from Roy et al. (2010). A, Decoding identity information. 100 neurons, and a 4-fold cross-validation paradigm was used (this study had 28 unique stimuli, leading to 3 x 28 = 84 training points, and 28 test points per CV split). B, Decoding category of the SAMPLE-STIMULUS. 128 neurons were used in a 5-fold cross-validation paradigm that had 10 repetitions of trials in each category in each cross-validation split (thus a total of 4 x 10 x 2 = 80 training points, and 20 test points were used per CV split). This analysis was run separately and then averaged over data from the two different category boundaries. C, Decoding the category of the DECISION-STIMULUS. The same parameters/paradigm that were used in subplot B were used here. D, Decoding the match/non-match status of a trial. 128 neurons and a 5-fold cross-validation paradigm was used with 30 repetiation of each trial type (240 training points, 60 test points, per CV split). E-F comparing different types of information in for ITC and PFC respectively.



Additional supplemental material 3.15 Comparison of the different types of information using the fixed boundary data of Freedman et al. 2003 (blue and red traces) to the changing category boundary data of Roy et al. 2010 (cyan and magenta traces), for ITC (upper blue and cyan plots) and PFC (lower red and magenta plots). The same parameters were used for both analyses in order to make a fair comparison.



Additional supplemental material 3.16 Dynamic coding of basic category information in ITC and PFC plotted separately for data from the two category boundaries from the Roy et al . (2010) experiment. The classifier was trained using data from the time indicated on the y-axis, and was tested using data from the time indicated on the x-axis (300ms bins were used sliding at 50ms intervals). Upper plots show results for boundary 1, and lower plots show the results for boundary 2. Data on the left is from ITC while data on the right is from PFC. 50 bootstrap neurons were used on each iteration. The same dynmanic population code is seen here as was seen for the Freedman et al. (2003) data (the results are slightly weaker for ITC on boundary 2, although they still seem present). Results are based on a 5-fold cross-validation paradigm (one example of each class was used in each cross-validation split).



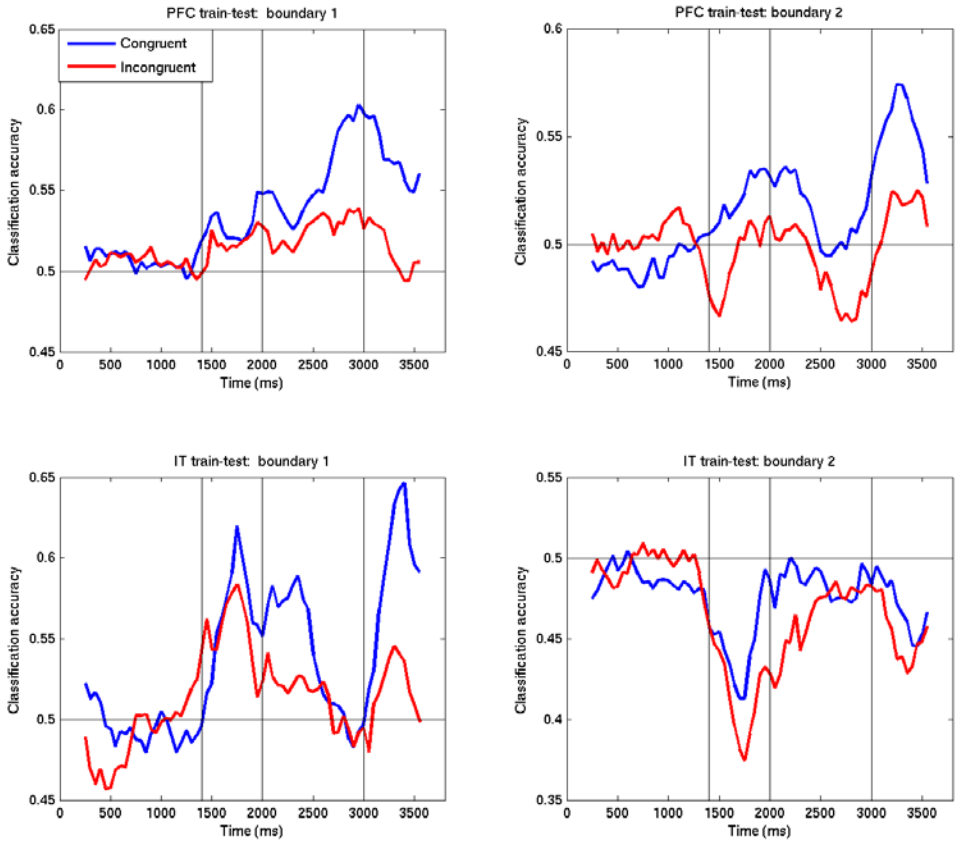
Additional supplemental material 3.17 Training at time 1 and testing at time 2 for decoding the identity of the 28 stimuli, for ITC (left) and PFC (right). The results show that there seems to be higher decoding accuracies when training and testing at the same time, but that there is some spread to other times periods as well. Given that the decoding accuracy is low, these results should be interpreted cautiously since category information could be contributing to the decoding accuracy of this identity information.

Since it was possible to replicate the basic results, we went on and applied decoding analysis to evaluate how much abstract category information was in ITC and PFC. To do this analysis, the classifier was trained to discriminate between data from two of the prototypes, and then tested on data from the other two prototypes (similar to the analysis in Figure 3.3), and the analysis was run separately for training to the classifier to group the stimuli along boundary 1, and training to group the stimuli along boundary 2. The key comparison was to evaluate the decoding accuracy in the 'congruent' case, where the category boundary the category boundary used by the classifier matched the category boundary used by the monkey, to the 'incongruent' case, where the category used by the classifier did not match the category boundary used by the monkey. For example, we could train the classifier to discriminate between c1 and c2, and then test the classifier on d1 and d2 data (i.e. train the classifier along category boundary) and then compare the congruent trial decoding, where the category boundary the monkey was using grouped c1 and d1 together, to the incongruent case, where the category boundary the monkey was

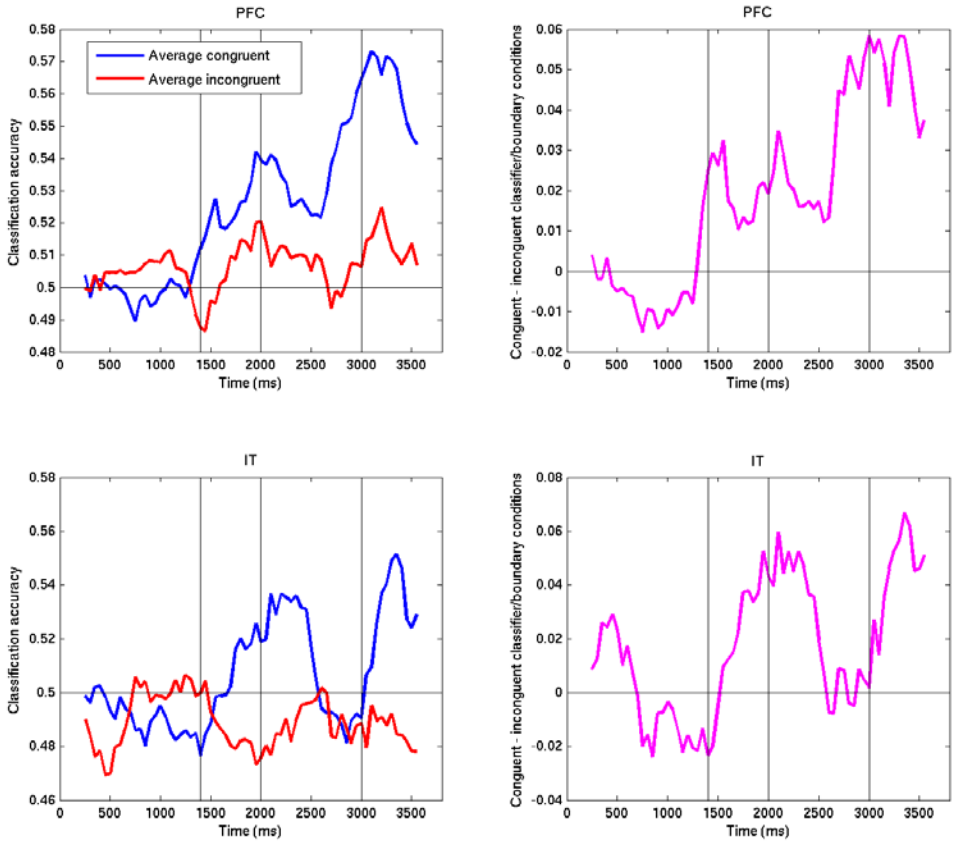
using grouped c1 and c2 together[20]. If abstract category information was really present, then there should be a higher decoding accuracy for the congruent trials than for the incongruent trials, indicating that the similarity structure in the neural data matched the similarity structure used by the monkey. It should be noted that whether the classifier was above or below 'chance' level is based on the similarity of the stimuli, so we should expect that for one of the boundaries classification decoding accuracy might be above chance, while for the other boundary it might be below chance. However what is important is whether the accuracy for the congruent decoding case is higher than the decoding accuracy for the incongruent case, which indicates a warping of the visual similarities based on the behavioral significance of the stimuli.

The results from this analysis are shown in Additional supplemental material 3.18. As can be seen, after the onset of the stimuli (first vertical line before 1500ms), the decoding accuracy for the congruent grouping was higher than the decoding accuracy for the incongruent grouping for both training boundaries, and for both ITC and PFC, indicating that the behaviorally relevant grouping of stimuli was influencing the neural similarity of the stimuli in these areas (unfortunately the results are a bit noisy due to the fact that the analysis is based on only 36 bootstrap neurons, since there were only 40 neurons in ITC that had enough repetitions of all the stimuli). Additional supplemental material 3.19 left figures, show the results averaged over the two training-testing category boundaries for the congruent and incongruent cases, and Additional supplemental material 3.19 right figures, show the average incongruent results subtracted from the congruent results, both of which indicate that there is abstract category information in both ITC and PFC.

[20] It should be noted that all four permutations of training and testing of the prototypes for a given training-testing boundary were run, and the results were averaged over these permutations; e.g., for training on c1-d1 vs. c2-d2 boundary, the classifier was run on: train on c1 vs. c2, test d1 vs. d2; trained on c1 vs. d2, test c2 vs. d2, etc..



Additional supplemental material 3.18 Plots showing that there is abstract category information in ITC and PFC in the changing category boundary experiment of Roy et al., (2010). For this analysis the classifier was training to discriminate between category boundary 1 (left plots) or category boundary 2 (right plots), and the results were compared when the category boundary the monkey was were congruent with the category boundary used to train the classifier (blue traces) or were incongruent with the category boundary used to train the classifier (red traces). As can be seen, after the stimulus onset (black vertical line around 1500ms) the decoding accuracy is generally higher for the congruent trials compared to the incongruent trials for both ITC and PFC, indicating that there is abstract category information in both these areas. The results in this figure are based on using 36 bootstrap neurons, and a 500ms sliding bin sampled at 50ms intervals.



Additional supplemental material 3.19 Abstract category decoding results for the changing boundary experiment combined over the two category boundaries. The figures on the left show the results from Additional supplemental material 3.18 averaged over the two training category boundaries. As can be seen, the fact that the congruent category decoding accuracy (blue trace) is higher than the incongruent decoding accuracy (red trace) gives support to the idea that there is abstract category information in both ITC and PFC. The figures on the left show the results of subtracting the incongruent category accuracy from the congruent category decoding accuracy (i.e., subtracting the red trace from the blue traces on the left plots). The above chance decoding accuracies again indicate that there is abstract category in ITC and PFC.

Chapter 4: Examining high level neural representations of cluttered scenes

The material in this chapter has been published as a CBCL/AI memo:

Meyers, E., Embark, H., Freiwald, W., Serre, T., Kreiman, G., and Poggio T. Examining high level neural representations of cluttered scenes. MIT-CSAIL-TR-2010-034/CBCL289, Massachusetts Institute of Technology, Cambridge, MA, July 29, 2010

Abstract

Humans and other primates can rapidly categorize objects even when they are embedded in complex visual scenes (Thorpe et al., 1996; Fabre-Thorpe et al., 1998). Studies by Serre et al., 2007 have shown that the ability of humans to detect animals in brief presentations of natural images decreases as the size of the target animal decreases and the amount of clutter increases, and additionally, that a feedforward computational model of the ventral visual system, originally developed to account for physiological properties of neurons, shows a similar pattern of performance. Motivated by these studies, we recorded single and multi unit neural spiking activity from macaque superior temporal sulcus (STS) and anterior inferior temporal cortex (AIT), as a monkey passively viewed images of natural scenes. The stimuli consisted of 600 images of animals in natural scenes, and 600 images of natural scenes without animals in them, captured at four different viewing distances, and were the same images used by Serre et al. to allow for a direct comparison between human psychophysics, computational models, and neural data. To analyze the data, we applied population 'readout' techniques (Hung et al., 2005; Meyers et al., 2008) to decode from the neural activity whether an image contained an animal or not. The decoding results showed a similar pattern of degraded decoding performance with increasing clutter as was seen in the human psychophysics and computational model results. However, overall the decoding accuracies from the neural data lower were than that seen in the computational model, and the latencies of information in IT were long (~125ms) relative to behavioral measures obtained from primates in other studies. Additional tests also showed that the responses of the model units were not capturing several properties of the neural responses, and that detecting animals in cluttered scenes using simple model units based on V1 cells worked almost as well as using more complex model units that were designed to model the responses of IT neurons. While these results suggest AIT might not be the primary brain region involved in this form of rapid categorization, additional studies are needed before drawing strong conclusions.

Introduction

Human and other non-human primates are able to rapidly extract information from complex visual scenes. Psychophysics studies have shown that humans can make reliable manual responses indicating whether an animal is present in a visual scene as early as 220ms after stimulus onset (Thorpe et al., 1996; Rousselet et al., 2002; Delorme et al., 2004), and when shown an animal and a non-animal image simultaneously (one in the left and right visual field), humans can reliably initiate saccades to the animal image with latencies as fast as 120m*s* after stimulus onset (Kirchner and Thorpe, 2006). Additional studies in humans have also shown that this rapid categorization behavior can occur in the absence of attention (Li et al., 2002), that performance is just as accurate when engaging in the task simultaneously in both left and right visual fields (Rousselet et al., 2002), and that categorization accuracy decreases as the amount of clutter in an image increases (and the size of the target decreases) (Serre et al., 2007). Similar studies using macaques have shown similar results although monkeys have even faster reactions times, with manual reaction times as quick as 180-230ms and saccade reaction times as fast as 100ms after stimulus onset (Fabre-Thorpe et al., 1998; Delorme et al., 2000; Macé et al., 2005; Girard et al., 2008). Thus humans and macaques have the ability to rapidly categorize complex and diverse images, potentially in parallel, and seemingly without the need to deploy attention.

A few studies have also examined the neural mechanisms that underlie this rapid categorization behavior. Electroencephalography (EEG) studies in humans on animal detection tasks have shown differences in event-related potentials (ERPs) around 150-170ms after stimulus onset between target present and target absent trials (Thorpe et al., 1996; Rousselet et al., 2002; Delorme et al., 2004). Functional magnetic resonance imaging (fMRI) studies in humans have also shown that when subjects need to detect a particular category of object in a scene, patterns of activity BOLD activity in lateral occipital complex (LOC) are similar to the patterns seen when an isolated image of an object from the same category is shown (Peelen et al., 2009).

Electrophysiological studies in macaques have also examined the effects that cluttered images have on neural responses and have shown that neurons' selectivity is not changed when a monkey fixates (and notices) a preferred object in the context of a cluttered scene (Sheinberg and Logothetis, 2001; Rolls et al., 2003). Additionally, studies on the neural basis of categorization have shown a diverse set of areas including the inferior temporal cortex (IT) (Sigala and Logothetis, 2002; Freedman et al., 2003; Kiani et al., 2007; Meyers et al., 2008), the prefrontal cortex (PFC) (Freedman et al., 2000, 2001; Shima et al., 2007), and lateral intraparietal cortex (LIP) (Freedman and Assad, 2006) are involved in different types of categorization behavior. However, these studies have generally used simpler stimuli of isolated objects, and a direct examination of the neural processing that underlies the *rapid* categorization behavior *in complex cluttered images* has not been undertaken.

In this study we begin to examine the neural activity that could be directly relevant for rapid categorization in macaques. In particular, we are interested in relating neural activity to a class of hierarchical feed-forward computational models of the ventral visual pathway (Serre et al., 2005, 2007) in order to assess whether this class of models is a good description of the neural processing underlying rapid categorization. Recent work Serre et al., (2007), showed that such computational models could match several aspects of human performance on rapid categorization tasks. In the study of Serre et al., (2007), a stimulus set was used that consisted of images of animals and natural scenes that were taken at four different distances from a camera (see Figure 4.1A). These images were then used in a psychophysics task in which each image was briefly presented to human subjects who had to press one button if an animal was in the image and a different button if the image did not contain an animal. Results from this study showed that humans achieved the highest accuracy when the full body of the animal was in an image, and that detection accuracy was lower for close-up images of animals' heads, and also for images in which the animal appeared further from the camera (see Figure 4.1B). A similar pattern of detection accuracy was also seen when using the output of model computational units to classify the presence/absence of an animal was in the same

images. Additionally, there was a high correlation between the mistakes that humans made and the mistakes that were made by the model ($r = \sim 0.70$), which suggests that the model was using similar visual information as humans. While these correlations suggest that humans and the computational model might be processing information in a similar way, directly testing whether neural responses match the outputs of the computational model would give much stronger evidence as to whether the computational model is a good description of the visual processing involved in this rapid categorization task. Thus, the purpose of this study was to test the plausibility of the computational model more directly by comparing the computational model output to the responses of neurons in areas that have thought to be involved in rapid categorization tasks.

In order to compare the computational model to neural data, we recorded from neurons in the ventral visual pathway as macaque monkeys viewed the same images used by Serre et al., (2007). We then analyzed the data using a decoding procedure (Hung et al., 2005; Meyers et al., 2008) that tried to predict whether an image contained an animal based on using either the neural recordings or the computational model output. Results from our analyses show that indeed several aspects of the neural activity matched both the computational model and the psychophysics including the relationship between classification accuracy and the size of the animal in the scene. However surprisingly, overall the classification accuracy from using the neural data was lower than the accuracy seen than when using computational model units and the correlation between the pattern of mistakes made by the classifier using the neural data and the computational model units, while highly significant, was still much lower than the correlation of mistakes previously seen between humans and the computational model. Additionally, the latency of information in inferior temporal cortex (IT) was relatively long (100-150ms) relative to the fastest saccade times previously reported (which were on the order of 100ms), which suggest that perhaps IT is not the critical area involved in rapid categorization when saccades are involved. In the discussion section of the paper we review several factors could have contributed to these discrepancies in the results between the model and the neural data that could potentially explain our results, however further electrophysiological studies are needed to make more conclusive statements.

Methods

Subjects and surgery

Two male adult rhesus macaque monkeys (referred to as Monkey A and Monkey B) were used in this study. All procedures conformed to local and NIH guidelines, including the NIH Guide for Care and Use of Laboratory Animals as well as regulations for the welfare of experimental animals issued by the German Federal Government. Prior to recording, the monkeys were implanted with ultem headposts (for further details see Wegener et al., 2004) and trained via standard operant conditioning techniques to maintain fixation on a small spot for a juice reward.

Recordings and eye-position monitoring

Single-unit recording & Eye-position monitoring. We recorded extracellularly with electropolished Tungsten electrodes coated with vinyl lacquer (FHC, Bowdoinham, ME). Extracellular signals were amplified, bandpass filtered (500Hz-2 kHz), fed into a dual-window discriminator (Plexon, Dallas TX) and sorted online. Spike trains were recorded at 1 ms resolution. Quality of unit isolation was monitored by separation of spike waveforms and inter spike interval histograms (ISHs). A total of 116 well isolated single units were recorded from dorsal anterior inferior temporal cortex (AITd) from monkey A, and 256 well isolated single units were recorded from AITd from monkey B. Additionally for monkey A, 444 well isolated units were recorded from dorsal posterior inferior temporal cortex (PITd), and 99 well isolated units were recorded from ventral posterior inferior temporal cortex (PITv). Eye position was monitored with an infrared eye tracking system (ISCAN, Burlington MA) at 60 Hz with an angular resolution of 0.25°, calibrated before and after each recording session by having the monkey fixate dots at the centre and four corners of the monitor.

Ophthalmic examination

Monkey A's eyes were inspected by one of the experimenters and a trained ophthalmologist with two different ophthalmoscopes. These measurements, performed in the awake and the ketamine anesthetized monkey, revealed myopia on the left (-3 dioptries) and right (-9 dioptries) eyes. In addition signs of astigmatism and possible retinal deficiancies were observed.

Stimuli and task

Two sets of stimuli were used in two different experiments. In the 'animal-scenes' experiment, the stimuli consisted of 600 images of animals in natural scenes, and 600 images of scenes without animals (see Figure 4.1A for examples of these stimuli). The animal and scene images were captured at four different distances from a camera, which we will refer to as 'head', 'close-body', 'medium-body' and 'far-body' images, which describes how the animals appeared in the different types of images, as determined by a set of human ratings (see Serre et al., 2007 for details). The images used in our experiments are same images as Serre et al., (2007) which allows us to directly compare results from the neural data to previous human psychophysics and computational modeling results. In the second 'isolated objects' experiment, the stimuli consisted of 77 images of objects from 8 different categories (food, toys, cars/airplanes, human faces, monkey faces, cats/dogs, boxes, and hands). These stimuli were previously used in a study by Hung et al., (2005), and allowed us to compare our neural data to previous recordings made from anterior IT (see Figure 4.1C for examples of these stimuli). More details about the stimulus sets can be found in Serre et al., (2007) and in Hung et al., (2005).

For both experiments, the stimuli were presented in a rapid sequence, with each stimulus being presented for 100ms, followed by 100ms inter-stimulus-interval in which a gray screen was shown (see Figure 4.1D). During the presentation of the stimuli, the monkey sat in a dark box with its head rigidly fixed, and was given a juice reward for keeping

fixation for 2-5 seconds within a 1.1 degree fixation box (when fixation was not kept, the image sequence during which fixation was not maintained, was repeated). Visual stimuli were presented using custom software (written in Microsoft Visual C/C++), and presented at a 60 Hz monitor refresh rate and 640 x 480 resolution on a 21" CRT monitor. The monitor was positioned 54 cm in front of the monkey's eyes, and the images subtended a 6.4° x 6.4° region of the visual field. For the isolated-objects experiment, all images were presented in random order until 10 presentations of each of the 77 objects had been shown. For the animal-scene experiment, the 1200 images were divided into blocks of 120 images, with each block consisting of 15 animal and 15 scene images from each of the four camera distances. The experiment consisted of running 5 presentations of each image within a block before going on to present the next block. For every experimental session, the blocks were presented in the same order, but the images within each block were fully randomized.

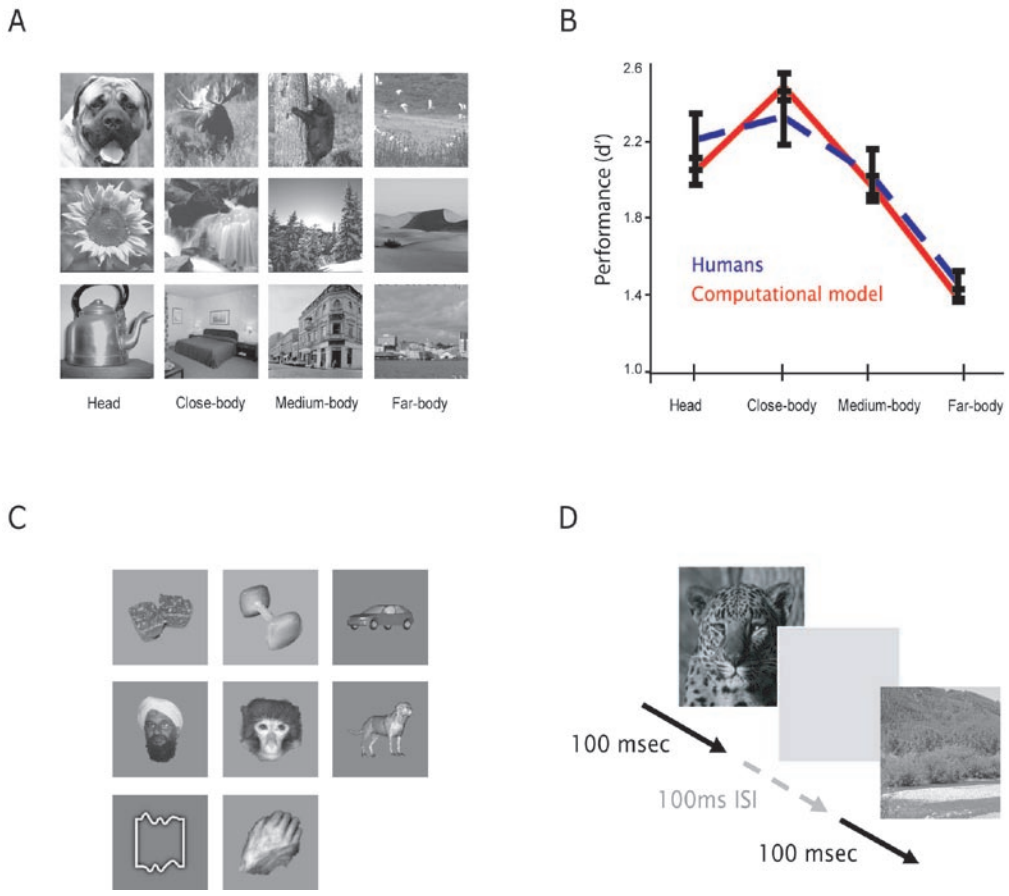


Figure 4.1 Example stimuli and the experimental design. A. Example images from the animal-scenes experiment in which 600 images of animals and 600 images of scenes that did not contain animals were shown to two monkeys. The images in the experiment come from four different camera distances (titled 'head', 'close-body', 'medium-body' and 'far-body'). The top row shows animal images examples, and the middle row shows 'natural' scene images and the bottom row shows 'artificial' scene images (the data set consisted of 300 natural and 300 artificial scene images). B. Results from (Serre et al., 2007) showing human psychophysics and computational modeling results. As can be seen, the ability of people to detect animals in these images is best for the 'close-body' condition and decreases in the medium-body and far-body conditions where the target animal becomes smaller and the amount of clutter increases (blue trace). Also, similar results were seen when training a classifier on computational model units that are based on the known properties of neurons in the ventral visual pathway (red trace). C. Example images from isolated objects experiment. D. The experimental design used for both experiments, in which images are shown for 100ms followed by a 100ms ISI in a rapid sequence.

Data analysis

Decoding analyses. The main analysis method used in this paper is neural population decoding which has previously been described in Hung et al., (2005) and Meyers et al., (2008). Briefly, this method works by creating pseudo-populations of neural data that consist of firing rates from a few hundred neurons that were recording independently but are treated as if they had been recorded simultaneously. A pattern classifier is first trained on several examples of these pseudo-population responses for each stimulus type, and then the classifier is used to make predictions about which stimulus type is present in a new set of pseudo-population responses that come from a different set of trials. The classification accuracy is calculated as the percentage of correct predictions made on the test dataset. For decoding which exact image was shown (Supplemental figure 4.1) the decoding procedure is used in a cross-validation paradigm in which the pseudo-population responses are divided into k sections; k-1 sections of data are used from training the classifier and the last section is used for testing, and the procedure is repeated k times each time using a different section of the data for testing (for supplemental figures 1, there were k=10 splits of the data, with each split consisting of pseudo-population responses from each of the 77 isolated objects). For all analyses, a bootstrap procedure is also applied in which different pseudo-populations are created from the data, and then whole cross-validation procedure is run. In this paper, the bootstrap procedure is run either 50 times for the isolated object analysis, or 250 times for the animal/non-animal analyses, and the final decoding results consist of the average decoding accuracy over all these different bootstrap-like (and cross-validation) runs. The error bars that are plotting are the standard deviation of the classification accuracy statistics calculated over all bootstrap-like runs.

Most decoding results in this paper are based on using a maximum correlation coefficient (MCC) classifier (this classifier has also been called a correlation coefficient classifier (Meyers et al., 2008), the maximum correlation classifier (Wilson and McNaughton, 1993) and the dot product algorithm (Rolls et al., 1997) and is described in those papers). We also make use of support vector machines (SVMs), and regularized least squares

(RLS) classifiers (Vapnik, 1995; Chang and Lin, 2001; Rifkin and Lippert, 2007). It should be noted that the MCC classifier does not have any 'free-parameters' that are not completely fixed by the data, while the SVM and the RLS classifiers have a single free-parameter (called the error penalty parameter and denoted by the letter C)[21] that determines the tradeoff between how well the classifier should fit the training data, versus how complex of a function should the classifier use. Larger values of the error penalty parameter C will cause the classifier to use more complex functions that better fit the data, however using a function that is too complex will often hurt the ability to correctly classify new points that are not in the training set (i.e., the classifier will overfit the training data). Conversely, if the value of C is too small, then the classifier will choose a function that is too simple that will not fit the training data very well which will also cause the classifier to generalize poorly to new data. For the RLS classifier, there is an efficient way to find the optimal value of C using only the training data which we always used (for this reason we generally prefer to use an RLS classifier over a SVM; see Rifkin and Lippert, 2007). For other analyses we are interested in comparing our work to previous work that used SVMs, thus we explicitly vary the value of C and see how changes in this parameter affect the cross-validation results (see Figure 4.6C). It should be noted that in order to make the problem of finding a good value for C computationally tractable when using an SVM, our analyses look at the *cross-validation results* from changing C rather than optimizing C using only the training data and then applying cross-validation (as is done for the RLS results); thus the classification accuracies from these analyses could be slightly biased upward due to over-fitting.

Before the data is passed to the classifier we calculate the mean and standard deviation for each neuron/model-unit using only the training data, and then we z-score normalize the training and test data using these means and standard deviations. The reason for normalizing the data is that the range of firing rates can vary drastically between different neurons, and such normalization helps prevent neurons with high firing rates from

[21] It is also common for researchers in machine learning to talk about a 'regularization constant' parameter (denoted λ) rather than the error penalty parameter C. The error penalty constant is related to the regularization constant by the formula $C = 1/(\lambda * k)$, where k is the number of training examples.

dominating the population analysis (although in practice we have found that results are largely unaffected by such normalization).

Results from these decoding analyses are reported in two different ways. The first ways is to simply report the 'classification accuracy', which is the percentage of the test data points that are correctly classified. The second way we report the results is in terms of the d-prime score. For the animal/no-animal decoding experiment, the d-prime decoding accuracy is calculated as the z-transform of proportion of animal images correctly classified as containing animals minus the z-transform of the proportion of images falsely classified as containing animals. We use this d-prime score in order to be able to easily compared our results to Serre et al., (2007) which also reported their results using this measure.

Our main decoding analyses address the question of whether we could use neural data to classify whether an image of a natural scene contains an animal (Figure 4.2). To do this analysis we use data from 50% of the images for training and data from the remaining 50% of the images for testing, making sure that the data from exactly half the images in both the training and test sets contain animals. Since each unique image was repeated 5 times when shown to the monkey, the data from all five trials for a specific stimulus went into the training set while the test set consisted of a single pseudo-population response from each image - thus the training set consisted of 3000 points and the test set consisted of 600 points[22] (using data from only a single trial of each image type for the training set did not change the results, see Supplemental figure 4.10). Additionally, for each bootstrap-like run, the images in both the training and test sets were divided evenly among the four camera distance image classes (i.e., 25% head, 25% close-body, 25% medium-body and 25% far-body in both the training and test sets), to keep all decoded conditions balanced. When training the classifier, all data from different camera distances was treated the same and a single decision boundary was learned for classifying

[22] In retrospect it might have been better to use all 5 repetitions of the test points as well, which could have possibly led to slightly smoother results and smaller errorbars, although the results overall would be very similar.

images that contained animals versus images that did not, exactly replicating the type of analysis done by Serre et al., (2007)). When testing the classifier, the decoding results for the four camera distances are typically reported separately (e.g., Figure 4.2A, Figure 4.2B etc.), and for some analyses, the results were further separated into accuracy for the animal images vs. the accuracy for the scene images (e.g., Figure 4.3). For all analysis, the decoding procedure was repeated 250 times using different images of animals/scenes in the training and test set each time.

In order to calculate the latency of information in AITd, we used a permutation test to assess when the decoding accuracies exceeded a level that would be expected by chance (Golland and Fischl, 2003). For each 25ms time bin that was used in Figure 4.2C, we randomly shuffled the image labels, and applied the full cross-validation decoding procedure using 50 bootstrap-like iterations[23]. This whole procedure was repeated 250 times to give a null distribution which indicates the range of expected decoding values obtained if there was no real relationship between the images shown and the data collected. P-values were calculated as the proportion of samples in the null distribution that were greater than or equal to the decoding accuracy from the real data-label pairing. The latency of information was then assigned to the first 25ms bin in which the p-values were below $p = 0.05$ level.

Comparison to computational model units and human psychophysics results. In some of the analyses we compare results from decoding neural activity to the results obtained from decoding the outputs of computational model units of Serre et al., (2007) that were run on the same animal/scene images. Briefly, the model of Serre et al., (2007) consists of a sequence of processing stages that alternate between template matching operations (which give rise to S units) and maximum operations (which give rise to C units), and works as follows: On the first level (the S1 level) images are convolved with a set of

[23] Ideally we would have run 250 bootstrap trials for each sample in the null-distribution to match the 250 bootstrap runs used to create the real decoding results, however this was computationally too expensive. Using only 50 bootstraps for each null sample will make each sample point in the null distribution slightly more variable, which will lead to a slightly larger standard deviation in the null distribution and consequently to more conservative p-values (i.e., more likely to make type II errors than type I error).

Gabor filters at four different orientations and 16 different spatial scales at locations distributed evenly across the image to create a larger vector of responses (these responses are analogous to the output of V1 simple cells). Next, for each S1 orientation, the maximum S1 response value within a small spatial neighborhood and over adjacent scales is taken to create a C1 vector of responses (these responses are analogous to the output of V1 complex cells). On the next level (the S2 level), for each local neighborhood, C1 unit response are compared to a number of 'templates' vectors (these template vectors were previously extracted from running the C1 model on a random subset of natural images that were not used in these experiments); the S2 response vector then consists of the correlation between each template vector and each C1 neighborhood response. For each template vector, the maximum value of S2 unit within a larger neighborhood is then taken to create the C2 responses (these C2 responses have been previously compared to the responses of V4 units by (Cadieu et al., 2007). Likewise S3 responses are created by comparing C2 responses to another set of templates, and C3 responses are created by pooling over even larger neighborhoods of S3 units. For more details on the model see Serre et al., (2007). Analysis of the outputs of the computational model units was done by applying the exact same decoding procedure that was used to decode the neural data except the neural responses were replaced by the responses of computational units. Unless otherwise specified in the text, the exact same number of computational model units and of neural responses were always compared in order to make the comparison of results are closely matched as possible.

Human psychophysics experiments were also previously run (Serre et al., 2007) using the same images that were used in the electrophysiological experiments reported in this paper. In those experiments, images were flashed for 20ms on a screen followed by a 30ms black screen inter-stimulus interval which was then followed by an 80ms mask, and humans needed to report whether an animal was present in the images. For several analyses in the paper, we compare the accuracy that humans could detect animals in specific images to the accuracy that a classifier achieved in detecting an animal in a specific image based on either neural data or data from computational model units.

Comparing the computational model units to the neural data was done in several ways. The simplest way to do this comparison was to plot the classification accuracies from the computational model units next to the classification accuracies from the neural data (Figure 4.2 and Figure 4.3). In order to do more detailed comparisons, two other methods were used. In the first method, we compared neural population activity to populations of computational model units by examining which images were consistently classified as animals (regardless of whether the classification was correct) using either neural data or computation model data as input to the classifier. To do this analysis we ran the decoding procedure 250 times, and calculated how often each of the 1200 images was classified as an animal. This yielded a 1200 dimensional 'animal-prediction' vector for both the neural and computational model data. We then correlated the animal-prediction vector derived from the neural data to the animal-prediction vector derived from computational model unit predictions to get an estimate of whether the neural data and the model units were making the same pattern of predictions (this again is similar to an analysis done by Serre et al., (2007) in order to compare human psychophysics performance on an animal detection task to the performance of computational model units). Additionally, we calculated the correlation between the animal-prediction vectors from each monkey, to get a baseline to compare the computational model animal predictions to. We also compare animal-prediction results from Serre et al., (2007) based on mistakes humans made on an animal detection psychophysics task and based on a 'full' computational model consisting of 6000 model units that was used in that work. Results are reported using both Pearson's correlation coefficient and Spearman's correlation coefficient.

In order to assess whether any of the correlations between these animal-prediction vectors could have occurred by chance, we conducted a permutation test. This test was done by randomly permuting the values each 1200 element animal-prediction vector and then calculating the correlation values in these permuted vectors. The permutation procedure was repeated 1000 times to create null-distributions for each correlation pair, and the p-value was assessed as the proportion of values in the null distribution that were greater than the correlation values from the real unperturbed animal-prediction vectors. For all

comparisons made in Table 1, all the real correlation values were greater than any value in the null distribution, indicating that each correlation was beyond what would be expected by chance. Approximate 95 percent confidence intervals were also calculated for the Pearson and Spearman correlation values on this null distribution by taking the 25th lowest and 976th highest values for all pairs of conditions that were correlated, and then choosing the pair that had the minimum value for the lower bound and the pair that had the maximum value as upper bound yielding a conservative estimate for the 95% confidence interval for all pairs (in practice the 95% confidence interval was in fact quite similar between all pairs).

We also compared the computational model units decoding results to the decoding results obtained from other simpler visual features. These features were: S1 model units (which are just Gabor filters created at four different orientations and 16 different scales), randomly chosen pixels, and the mean values of pixels in small image patches. To create the S1 units, we used the parameters previously described by (Serre et al., 2007), and then selected randomly 1600 units for each of the four orientations of Gabor filters, yielding a pool of 6400 features for each image (the same filters were chosen for all the images, thus making the decoding possible). To create the pixel representation, 1600 randomly selected pixels were chosen from each image (again, the position of each randomly selected pixel was the same in all of the images). To create the mean patch representation, we used a similar process that was used to create the S1 units, except that we convolved the image with averaging filters at 16 different patch sizes rather than Gabor functions, and there was only a total of 1600 features used, since mean filters are not oriented. When decoding whether an image contained an animal in it using these features, we applied the same decoding procedure that was used for model units and neural data; namely, on each bootstrap-like iteration, we randomly selected 100 features from the larger pool, and then repeated this bootstrap-like procedure 250 times using a different selection of 100 random features each time.

Results

Decoding whether an animal is in a natural scene image

To try to gain a better understanding of the brain regions and neural processing that underlies rapid object recognition, we used a population decoding approach to assess if we can predict whether an animal was present in a complex natural scene image that was shown to a monkey based on neural data recorded from the ventral visual stream. If it is possible to decode whether an animal is in a complex natural scene image from neural data from AIT recorded <100ms after stimulus onset, then this suggests that AIT could potentially be an important brain region in rapid categorization behavior. Alternatively, if it is not possible to decode whether an animal is in a natural scene image within a behaviorally relevant time period, this gives some support to the theory that other brain regions might be the critical areas involved in rapid categorization (Kirchner and Thorpe, 2006; Girard et al., 2008).

In order to do an animal/non-animal decoding analysis, we trained a classifier using data that was collected from half of the animal and scene images, and we tested the classifier using data that was recorded when the other half the images had been shown (i.e., the training and test sets each had data from 600 images). The training and test sets were balanced in terms of the distance that images were from the camera (i.e., balanced in terms of head, close-body, medium-body, and far-body images) and in terms of having 50% of the images containing animals in both sets. The whole decoding procedure was repeated 250 times using data from different randomly chosen image in the training and test sets in order to get a smoother estimate of the information contained in the neural data (see methods for more details).

Figure 4.2A shows the decoding results separately for the head, close-body, medium-body, and far-body conditions, based on using a MCC classifier and the neural firing rates from AITd in a 200ms time bin that started at 100ms after stimulus onset for both monkeys (red and blue lines) (Supplemental figure 4.4 plots these same results in terms

of percent correct rather than d' values). As can be seen, for both monkeys, the head, close-body, and medium-body conditions are decoded at an accuracy that occurred above chance performance, indicating that it is possible to tell whether an animal is present in a cluttered natural scene based on the neural activity from 100 neurons, provided that the image of the animal is not too small relative to the amount of surrounding clutter. For monkey A, we also recorded data from dorsal and ventral posterior inferior temporal cortex (PIT). Decoding results from these areas show an even lower accuracy than the results from AIT (Figure 4.2B), although the dorsal PIT results appear to have the same trend of decoding accuracy as a function of camera distance that is seen in AIT and the computational model units.

In order to assess the latency of information in AIT, we applied the same decoding paradigm (i.e., 50% of the images used for training and 50% of the images used for testing) to neural data using the firing rate in 25ms bins, sliding at 25ms intervals (see Figure 4.2C). The results from conducting a permutation test on this data (see methods) suggest that a latency between 125-150ms for the head, close-body, and medium-body conditions in both monkeys (the far-body condition was never consistently above chance, so it was not possible to assess the latency for this condition).

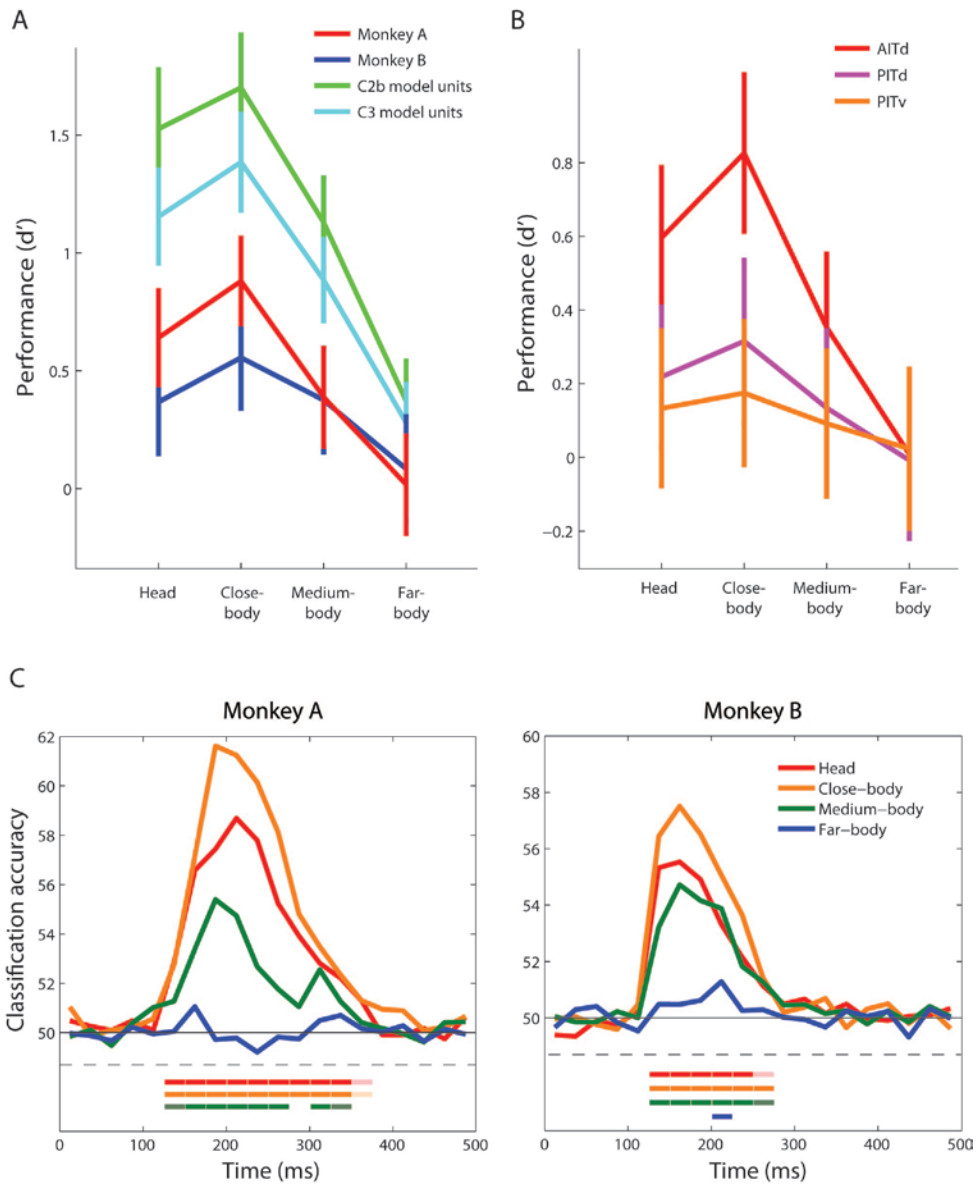


Figure 4.2. Results from decoding whether an animal is in a natural scene image using data. A. A comparison of decoding results using neural data from AITd from the two monkeys (red and blue traces), to results obtained from two different types of computational model units (cyan and green traces). As can be seen, decoding results based on the neural data and the computational model units show the same general trend as a function of camera distance. However, overall the decoding accuracies from the computational model units are better than the results from the neural data. B. A comparison of results from three different brain regions from monkey A. The results again show similar trends as a function of camera distance, but the AITd results are better than the more posterior regions. C. Decoding accuracies from both monkeys as a function of time. The colored lines below the plot show time when the results were significantly above chance ($p < 0.05$ light trances, $p < 0.01$ dark traces, permutation test). For both monkeys, for the head, close-body and medium-body distance, a significant amount of information was in AITd starting 125-150ms after stimulus onset.

Finally, we plotted the classification accuracies separately for the images that contained animals and the images that did not contain animals (Figure 4.3, upper two plots). As can be seen, decoding results based on neural data from both monkeys show a similar general trend in which the classification accuracy for images without animals *increases* from head to far-body conditions, and the classification accuracy for images with animals *decreases* from head to far-body conditions. This perhaps is not surprising since the far-body animal images consist mostly of background clutter that is perhaps be more 'similar' to the visual attributes in cluttered images that do not contain animals than to close-up images of animals heads (the fact that the far-body animal images are below chance also shows that these images were indeed seen by the classifier as more similar to images that do not contain animals than to the other animal images). When the classifier was trained and tested separately on data from the four different distances (Supplemental figure 4.6) again the decoding results showed the same pattern, but none of the results were below chance levels, confirming the fact that the far-body animal images generally were more similar in their neural responses to images from all distances that did not contain animals, than to images from all distances that did contain animals.

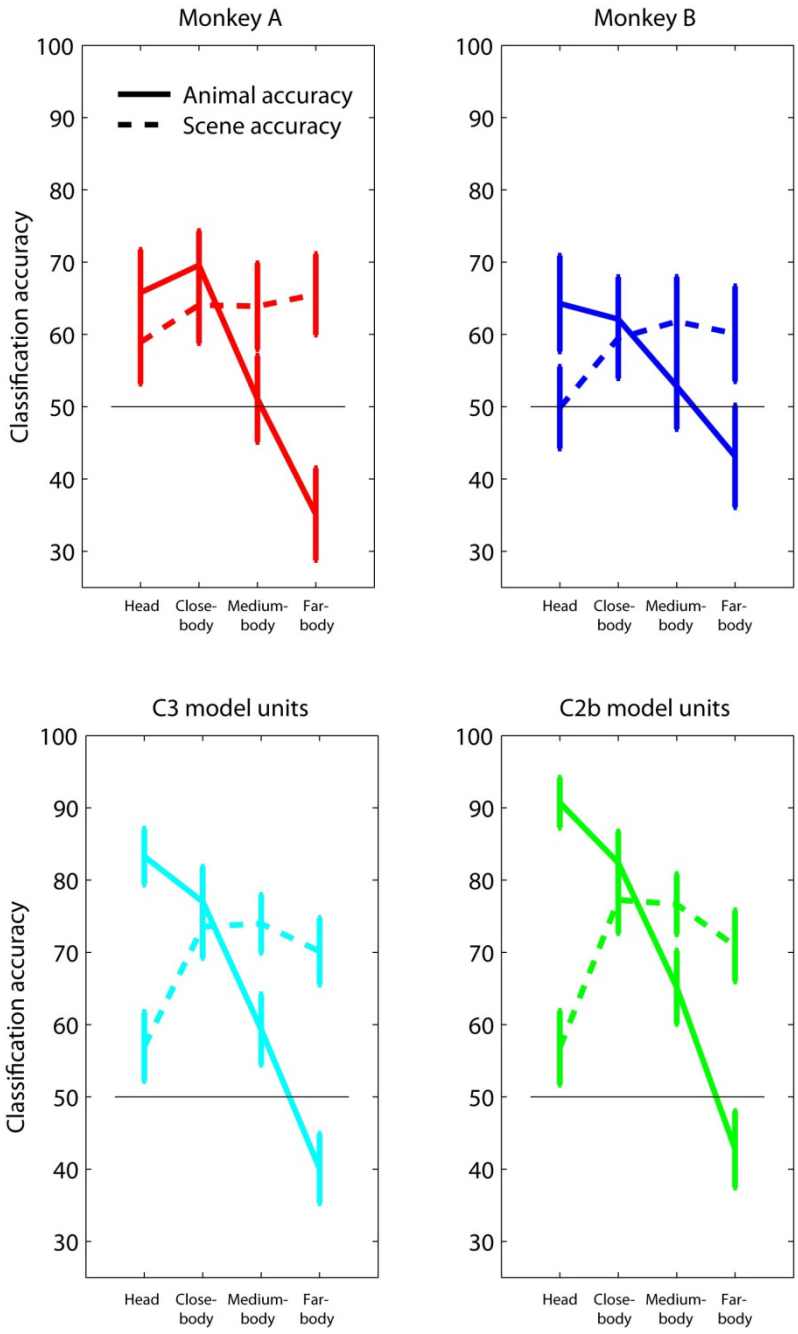


Figure 4.3 Decoding accuracies using neural and model unit data plotted separately for the animal images (solid lines) and non-animal images (dashed-lines). As can be seen, the results for the neural data from both monkeys (upper two plots) and for the C3 and C2b computational model units (lower two plots) show similar trends with an increase in decoding accuracy for the non-animal images at further camera distances, and a decrease in decoding accuracy for animal images at further camera distances. This pattern of results is due to the fact that the when the animal is far from the camera the background clutter dominates the image causing the computational and neural representations to be more similar to images that do not contain animals.

Comparing computational model units to neural AIT data

In addition to assessing whether the neural data contains sufficient information about whether an animal is present in a cluttered scene in a time frame that is fast enough to underlie behavior, we were also interested in assessing whether the computational model proposed by Serre et al., (2007) could potentially provide a satisfactory description of the neural processing that is occurring in IT. In order to investigate this question, we did several analyses to assess whether decoding results based on computational model units appeared similar to decoding results based on the neural data. For a first analysis, we applied the same decoding procedure used on the neural data in Figure 4.2A to the C2b and C3 computational model units of Serre et al., (2007), (see methods section for more details). The results are plotting on Figure 4.2A in the green and cyan lines. As can be seen in the figure, a similar trend in decoding accuracy appeared in both monkeys an in the computational model, with the best performance occurring in the close-body condition, and the results becoming worse for images in which the animal appeared further from the camera (i.e., the far-body condition), which is the same trend reported by Serre et al., (2007) in human psychophysics and computational modeling experiments. The results also clearly show that computational model units have an overall *higher* level of performance than the decoding results based on the neural data (and these differences are even larger when a regularized classifier is used, see Figure 4.5). We also compared the classification accuracies separately for the images that contained animals and the images that did not contain animals using the computational model units (Figure 4.3, lower two plots), and again observed the same pattern seen in the neural data, namely, that the classification accuracy for images without animals *increases* from head to far-body conditions, and the classification accuracy for images with animals *decreases* from head to far-body conditions.

While the fact that the computational model units and the neural data showed similar trends in performance as a function of image distance suggests that the neural data and the computational model units could be operating in a similar manner, a more detailed

analysis is needed to draw stronger conclusions. In order to better assess the similarity in performance between the model units and the neural activity we undertook two analyses, one of which focused on the information contained at the population level and the other focused more on the correspondence between individual neurons and model units.

To compare the neural population activity to populations of computational model units, we examined which images were consistently classified as animals (regardless of whether the classification was correct) using either neural data or computation model data as input to the classifier. To do this analysis we created 'animal-prediction' vectors that were based on how often each image of the 1200 images were classified as an animal (regardless of whether the image actually contained an animal) based on using either the neural data or computational model data as input to the classifier. We then correlated these animal-prediction vectors using either Pearson or Spearman's correlation coefficient. We also correlated the neural/model results to animal-prediction vectors that were based on psychophysics performance of how often humans reported an animal in an image in a rapid animal detection task, and to a previous implementation of a 'full' computational model results that was used by Serre et al., (2007) which were obtained by applying an SVM to 1500 units from C1, C2, C2b and C3 levels of the model (for a total of 6000 units).

Table 1 shows the results from this analysis using Pearson's correlation coefficient (upper triangular part of the matrix) or Spearman's correlation coefficient (lower triangular part of the matrix) (correlations with additional features are shown in supplemental table 4). Based on a permutation test (see methods section), an approximate 95% confidence interval on the Pearson's (Spearman's) correlation from the null distribution is [-0.062 0.062] ([-0.061 0.061]) for all conditions, indicating that all the correlation between animal-predictions for all conditions are well above what would be expected by chance[24].

[24] To put the values in Table 1 in perspective, we also calculated two measures of reliability of the neural data by comparing half the data from one monkey to the other half of the data from the same monkey. The first measure of within monkey reliability examined reliability across trials. To do this analysis we randomly divided the neural data from the 5 repeated trials of each stimulus into disjoint two sets, with each set having data from 2 of the trials for each stimulus. We then applied the full decoding procedure to

Thus the decoding results based on neural data, the computational model and the results of human psychophysics detections are all making similar patterns of mistakes on many images. However, the correlation level between the model units and the neural data is lower than the correlation level between the neural data from the two monkeys[25], which suggests that there is additional structure in the neural data that the computation model units are not capturing. Additionally, the correlation between the results obtained from the full computational model of Serre et al., (2007) and the results from using a subset of 100 model units from the higher levels of the computational model (C2b and C3) only have an agreement at a correlation level between .45 and .61. In the section below titled 'A closer examination of the computational model results' we examine reasons for this seemingly low correlation.

---

each set of two trials separately, and correlated the animal prediction vectors from the first set with the animal prediction vectors obtained from the second set. Finally this procedure was repeated 50 times. The average Pearson's (Spearman's) correlation value for Monkey A from this procedure was .71 (.70) and the average Pearson's (Spearman's) correlation values from Monkey B were .65 (.62). The second measure of within monkey reliability examined the reliability across neurons. For this analysis we randomly divided the neurons from one monkey into two disjoint sets of 50 neurons each, and then applied the full decoding procedure two each set separately and then correlated the animal prediction vectors. This procedure was also repeated 50 times. The average Pearson (Spearman) correlation value for this procedure for Monkey A was .66 (.62) and for Monkey B was .36 (.35). Thus relative to the comparisons between monkeys and between monkey and computational units, the within monkey reliability was typically high.

[25] A 95% confidence interval on the Pearson's correlation between the two monkeys is [.38 .47], while the 95% confidence intervals between Monkey A and the C2b and C3 units are [.21 .31] and [.15 .26] respectively, and the 95% confidence intervals between Monkey B and the C2b and C3 units are [.29 .39] and [.21 .32] respectively. Thus, the confidence intervals on the Pearson's correlation coefficient between the two monkeys only overlaps with the confidence interval between Monkey B and the C2b, which suggests that the agreement between Monkey A and the computational model units, and the agreement between Monkey B and the C3 units, are not as high as the agreement between Monkey A and Monkey B.

	Monkey A	Monkey B	C2b	C3	Human	Serre full model
Monkey A		0.43	0.26	0.21	0.38	0.36
Monkey B	0.45		0.34	0.27	0.50	0.44
C2b	0.28	0.36		0.91	0.50	0.56
C3	0.21	0.29	0.93		0.40	0.45
Human	0.37	0.48	0.50	0.42		0.72
Serre full model	0.36	0.44	0.61	0.50	0.71	

Table 4.1. Correlation coefficient values between how often each of the 1200 images were predicted as containing an animal based on using human psychophysics results, classification accuracies from neural or computational model units. Upper triangular results are based on Pearson's correlation coefficient and lower triangular results are based on Spearman's correlation coefficient. While all the correlation values are larger than would be predicted by chance, higher correlation levels occur between the two monkeys than between the monkeys and computational model units, indicating that the computational model units are not capturing all the possible variance found in the neural data.

A closer examination of the computational model results

As mentioned above, in the process of comparing the neural data to the computational model units we noticed that the correlation between results based on using a smaller subset of C2b or C3 units and the previous computational model units results using the 'full' model obtained by Serre et al., (2007) was not that high (correlation values between .45 and .63), and the overall classification accuracy using this smaller subset of computational units was lower (compare Figure 4.1B to Figure 4.2A). Since the major differences between the 'full' model of Serre et al., (2007) and the model used here were 1) the number of units used, and 2) the types of units used, and 3) the classifier used, we decided to look in more detail at how these factors influenced the decoding results.

To analyze how the type of classifier affected the results (Figure 4.4A), we recreated the analyses in Figure 4.2A, but this time we used a regularized least squares classifier (RLS) instead of a maximum correlation coefficient classifier (MCC). Regularized classifiers such as RLS and SVMs have been shown to yield very good performance in a range of machine learning problems, but there is little evidence showing that they improve the

performance when used to decode neural data see supplemental material from Meyers et al., (2008). Figure 4.4 shows that indeed the computational model unit performance greatly increases when using a regularized classifier (overall increase in d' values of 0.463, and 0.633 for C2b and C3 respectively), while the performance remained largely the same for the neural data (overall change in d' of -0.0457 and 0.001 for Monkey A, and Monkey B respectively). Since this same number of neurons and computational model units were used in this analysis, this again points to a difference in how the computational model units and real neural data are representing information about the images. Additionally, it should be noted that the correlation between the RLS C2b or C3 units and the previous computational model units results using the 'full' model obtained by Serre et al., (2007) was in the range of ~.75 to .78 (see table 2) indicating the type of classifier was a significant factor influencing the difference between our current results and the previous results of Serre et al., (2007). Finally, it should be pointed out that when using an RLS classifier, the Spearman's correlation between Monkey B and the computational model units is actually higher than the correlation between Monkey A and Monkey B, indicating that model units are capturing as much of the variation in the neural data of Monkey B as should be expected. However the results based on Pearson's correlation and the correlation between the model units and Monkey B, are still lower than the correlation between the two monkeys indicating that the model units are still not explaining all potential neural variation.

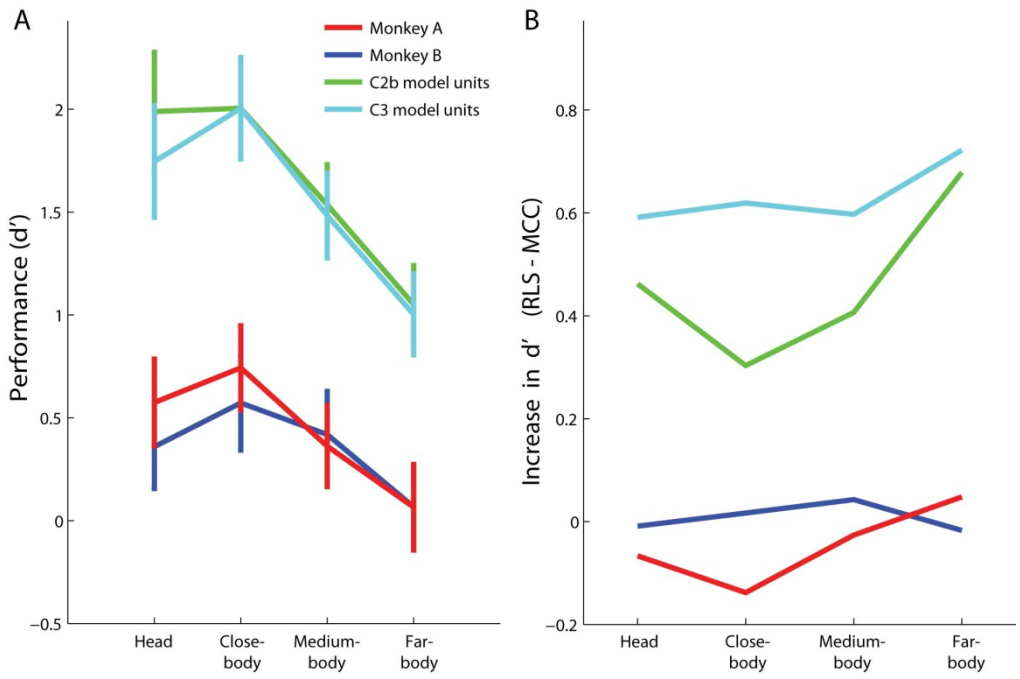


Figure 4.4 Decoding accuracies for whether an animal is in a cluttered scene image using a regularized least squares (RLS) classifier. A: Results plotted in the same format as Figure 4.2A, except that a RLS classifier is used instead of an MCC classifier. B: The change in d' classification accuracy when using a RLS classifier compared to when using an MCC classifier (values calculated by subtracting the MCC decoding accuracies in Figure 4.2A from the RLS classification accuracies shown in Figure 4.4A). As can be seen, using a regularized classifier greatly improves the classification performance of the computational model units, while leaving the neural decoding results largely unchanged.

	Monkey A	Monkey B	Model C2b	Model C3	Human	Serre full model
Monkey A		0.49	0.36	0.36	0.41	0.39
Monkey B	0.47		0.44	0.45	0.51	0.44
C2b	0.37	0.46		0.95	0.66	0.77
C3	0.36	0.47	0.93		0.66	0.75
Human	0.39	0.49	0.66	0.65		0.72
Serre full model	0.38	0.45	0.77	0.76	0.71	

Table 4.2 Correlation coefficient values between how often each of the 1200 images were predicted as containing animals based on using human psychophysics results, classification accuracies from neural or computational model units when an RLS classifier was used. Upper triangular results are based on Pearson's correlation coefficient and lower triangular results are based on Spearman's correlation coefficient. For Pearson's correlation, the agreement between the two monkeys is still higher than the agreement between the model units and data from either monkey. However, when Spearman's correlation is used, the neural decoding results from monkey B seem to be better explained by the computational model units than by matching the results to the other monkey (as can be seen by comparing the value in column 1 row 2, with the values in column 2).

To analyze how the number and type of computational model units affected the decoding accuracy, we trained a MCC and a RLS classifier on C1, C2, C2b, C3, and a random combination of all unit types, using either 100 or 1500 units. The results are shown in Figure 4.5. As can be seen again, results from the RLS classifier are significantly higher than the results from the MCC classifier. There is also an increase in decoding accuracy with more units when an RLS classifier is used, but this increase is somewhat small. More surprisingly, there does not appear to be a clear advantage to using the more sophisticated C2b and C3 features that are supposed to model the responses of IT neurons, compared to the results based on using simple C1 features which are modeled after V1 complex cells (the one exception seems to be for the 'head' condition when an MCC classifier is used, for which the C2b and the mix of all unit types tend to perform better than the C1, C2 and C3 units).

The fact that C1 units work almost as well as using a combination of all unit types differs from the findings of Serre et al., (2007) which showed that Model C1 units have a lower level of performance than the full Model (see Serre et al., (2007), supplemental table 2).

Two differences exist between the methods used here and those used by Serre et al., (2007). First, we used an RLS classifier here, while Serre et al., (2007) used an SVM. Second, Serre et al., (2007) used 1500 Model C1 units and *6000* units of all types in their 'full' model, while we used 1500 Model C1 units, and *1500* randomly chosen units of all types in our comparison. Thus either the classifier type of the number of model units used in the 'full' model should account for the difference in our findings. Below we explore these two possibilities.

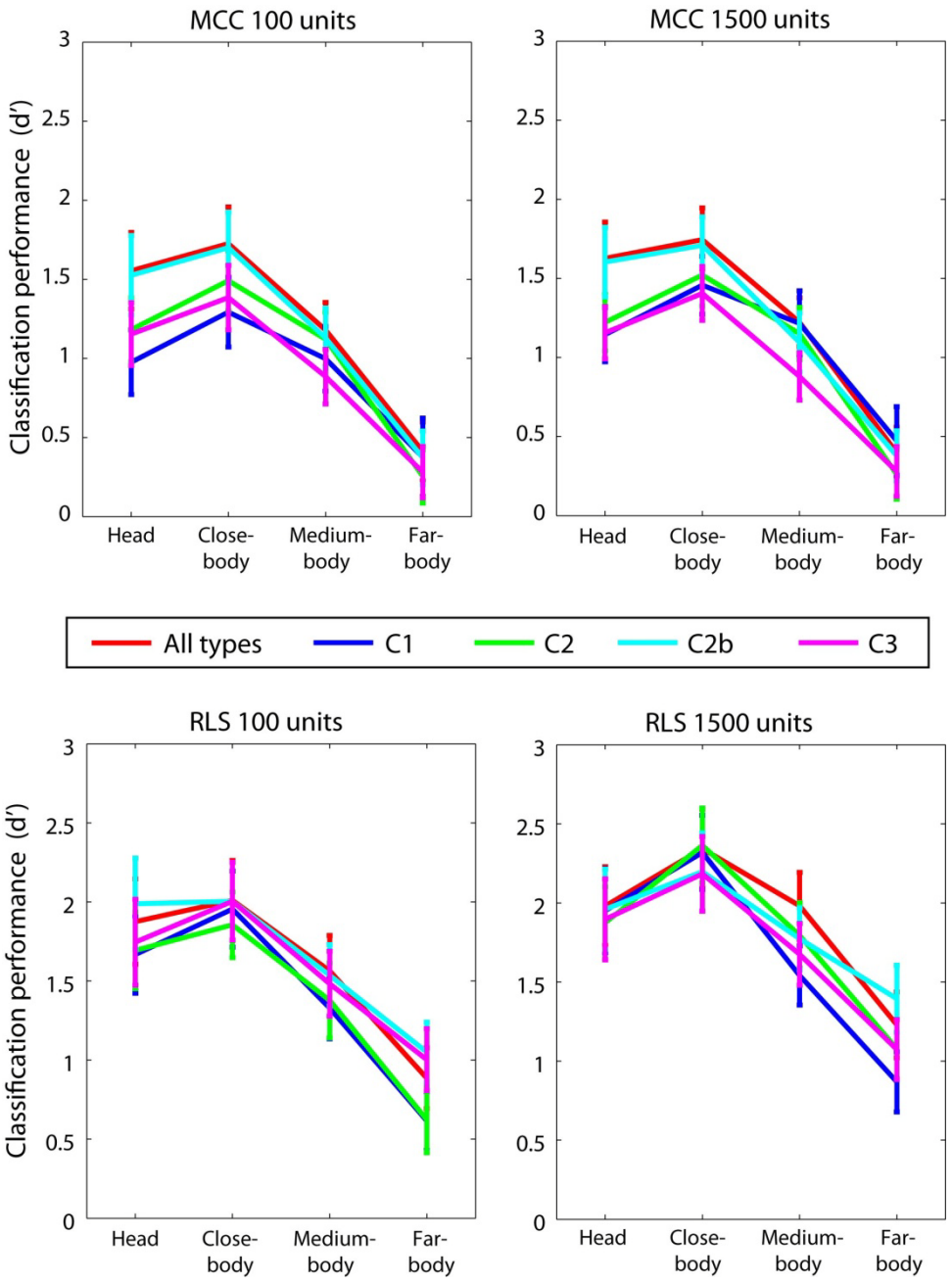


Figure 4.5 Comparing different computational model unit types using either an MCC or an RLS classifier with either 100 or 1500 model units. The results show that performance is much better when an RLS classifier is used, and that there is a slight increase in performance when more units are used. Overall in most cases, the type of computational model unit used did not have a large affect on decoding accuracy.

Figure 4.6 shows results from comparing how these two factors influence decoding accuracy for the animal-scenes dataset. As can be seen, when an SVM is used (Figure 4.6A), there is a large difference between the C1 units (blue trace) and using the full model with 6000 units (green trace), thus we were able to replicate the findings of Serre et al., (2007). Also, when 1500 units of all types are used (i.e., the 'full' model, but using just 1500 units), the decoding accuracy is between using the C1 features and using the 'full' model with 6000 units, indicating that approximately half the increase in performance when going from C1 units to the 'full' model was that there were 4 times as many features used in the full model, and half the increase was due to the diversity of features used (i.e., C1, C2, C2b and C3). When an RLS classifier was used (Figure 4.6B), the pattern of results was a bit different. First, there was almost no difference in decoding accuracy between the full model results when 1500 units are used and when 6000 units are used. Second, on the head and close-body conditions there is almost no difference between C1 units and the full model results. Finally, for the medium-body and far-body conditions, the decoding accuracy for the C1 units still appear slightly lower than the full model results. Thus when an RLS classifier is used, not only is the decoding accuracy higher using the C1 and 1500 model units than when an SVM is used, but additionally the differences between the simple C1 features and the full model are greatly reduced. This raises the question about how useful the complex and highly size and position invariant properties that are built into C2b and C3 feature responses are for animal non-animal discrimination in this dataset.

We also did an additional analysis to try to determine what was giving rise to the difference in the SVM and RLS results. There are two differences between how the SVM and RLS classifiers were used in our analyses. The first difference is that the SVM and RLS use different loss functions when the classifier is learning a separating function on the training data. The difference in these loss functions is what defines these two algorithms and thus is not a parameter that can be freely modified. The second different between the SVM and RLS algorithms is that there is an efficient way to optimize the error penalty constant on the training data for the RLS algorithm, while optimizing the

error penalty constant for an SVM is a very computationally intensive. Thus, for all the analyses above, we optimized the error penalty constant for the RLS algorithm on the training data, but for the SVM we used the default error penalty constant (which was the same method used by Serre et al., (2007)). However, since it was possible that the error penalty constant could be a large factor in contributing to the difference in results, we reran the SVM analysis several times using different error penalty constant values in order to determine whether the difference in error penalty constant values was giving rise to the difference in results.

Figure 4.6C shows the decoding accuracy for an SVM (averaged over the 4 head, close-body, medium-body and far-body conditions) as a function of the error penalty constant value *C* (Supplemental figure 4.8 shows the results separately for the 4 distances). As can be seen, the highest decoding accuracy is obtained when the error penalty constant is *C=.001* for all three model unit number/types that were tested. Additionally, having the optimal value for the error penalty constant affected the 1500 unit results more than it affected the results based on using 6000 units. When the SVM animal/non-animal results were recalculated using this optimal value of C=.001 (Figure 4.6D), the SVM results were a much closer match to the RLS results, indicating that the difference in error penalty constant values was a large factor contributing to the difference in the SVM and RLS results. More importantly, with this optimized error penalty constant value, the head and close-body conditions were no longer higher using all model unit types compared to when only using C1 features. These results indicate that for the animal-scene dataset used in this study that: 1) the model unit results (unlike the results based on neural data) are very sensitive to the exact classifier parameters used, and 2) while using a combination of more complex visual features in the higher model units as well as lower level units does lead to an improvement in discriminating between animals and natural scenes this improvement is smaller than is suggested by Serre et al. (2007) (and seems nonexistent for close-body conditions).

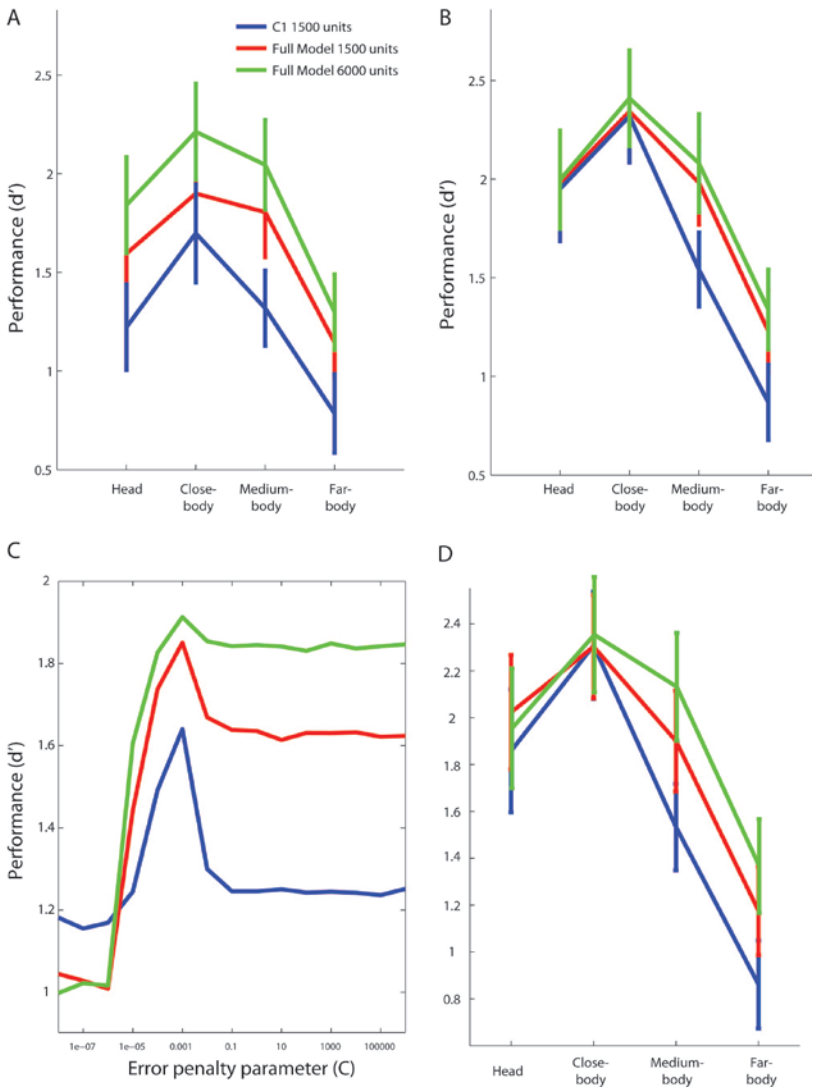


Figure 4.6 Comparison of using many model unit types, to using only C1 units, for an SVM classifier and a RLS classifier. A: results based on using an SVM (with the default error penalty constant value C =1), yields better performance when the full model using 6000 unit are used compared to using just 1500 C1 features, thus replicating the findings of Serre et al., (2007). When 1500 units of all types are used with a SVM, the results are between the C1 results and the 6000 model unit results of all types, indicating that part of the reason why the 'full' model of Serre et al., (2007) outperformed the C1 units was due to the fact that Serre's full model used four times as many units. B: For the RLS classifier, there is not much difference between using 1500 model units of all types and 6000 model units of all types. Additionally, the C1 units seem to only perform worse on the medium-body and far-body conditions. C: SVM animal/non-animal classification results (averaged over all 4 image distances), as a function of the error penalty parameter C (for all the RLS results, the optimal value of C was always determined using the training data). As can be seen the optimal value of C is .001, which yields higher performance than using the libSVM default value of C = 1. D: SVM animal/non-animal decoding results using an error penalty parameter of C = .001 (that we determined to be optimal in Figure 4.6C). With this error penalty parameter, the SVM results look much more similar to the RLS results shown in Figure 7B.

Finally, given the fact that simple C1 features did almost as well as more complex C2b and C3 features, we decided to test whether even simpler features than the C1 units could reproduce the level of performance that was seen when decoding information from the model units or the neural data. The simple features we decided to test were: randomly chosen pixels, S1 features (which are Gabor filters that match simple cell receptive fields), and the mean value of pixels in neighbors that were the size of Gabor filters used for the S1 features (see methods for more details). Results from this analysis are shown in Figure 4.7. As can be seen in most cases, the decoding accuracies based on random pixels, mean pixel intensities, and S1 features performed worse than the model unit features and the neural data, particular when an RLS classifier is used (Figure 4.7B). This matches of findings of Serre et al. (2007) who showed that many image feature types did not perform as well as the computational Model units described in this paper.

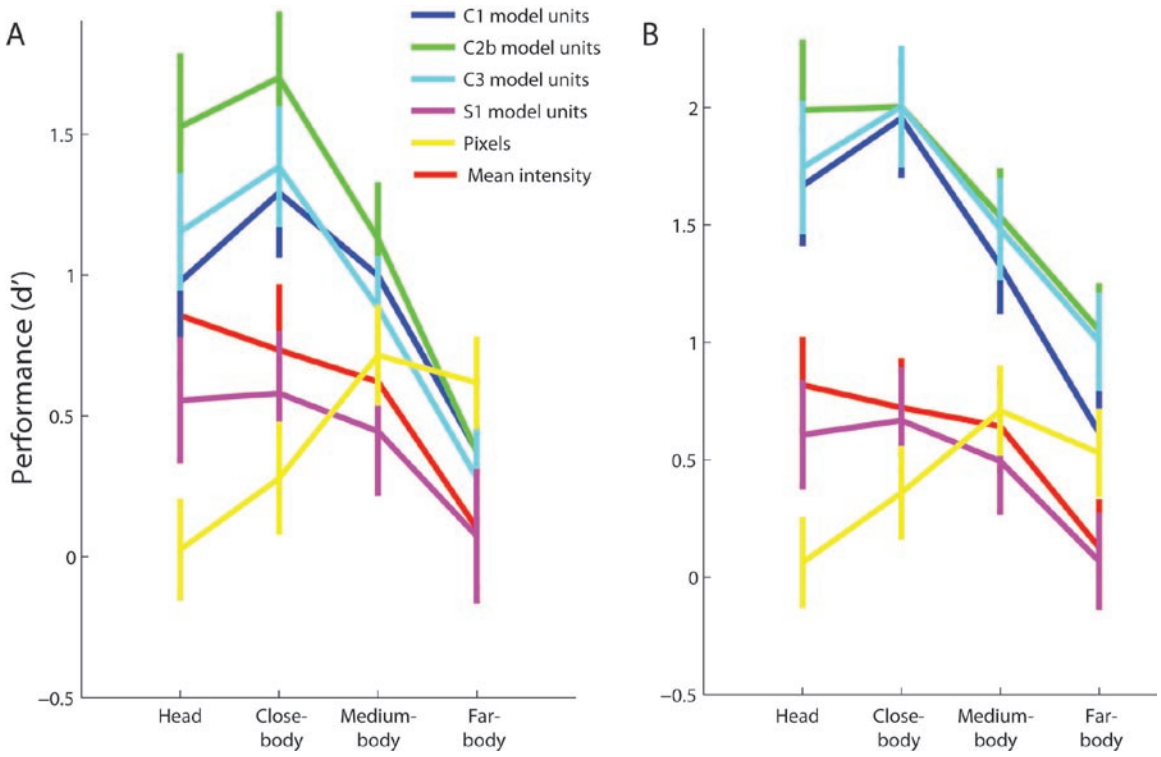


Figure 4.7 Comparing simpler visual image features based either on pixel values (yellow trace), mean pixel intensities in small image patches (red trace), or S1 model units (purple trace) to the other model units used in this paper (blue, green and cyan traces), when an MCC classifier was used (A), or a RLS classifier was used (B). As can be seen, the results of these simpler features are generally worse than the results from the model units used in this paper, particularly when an RLS classifier is used. Thus, not all image features work well for discriminating which images contain pictures of animals.

Discussion

The results of this paper show that it is possible to decode whether an animal is in a cluttered scene image using neural data from AIT and also using the computational model units of Serre et al., (2007) at levels that are well above chance. Given the diversity of visual appearance of the images used, and the fact that classification based on using simple image features perform worse, this result is not completely trivial. Additionally, the pattern of classification accuracies as a function of image distance was similar among the neural data and computational model units, suggesting that both the

neural data and computational model units could be relying on similar visual information in the images. This result is related to the findings of Serre et al., (2007) who showed the mistakes humans make in detecting animals in cluttered scenes are similar to the mistakes made by a classifier that is trained on the same computational model units, although the correspondence between the computational model units and the neural data was not as strong as that seen between the human psychophysics results and the computational model units.

One of the more surprising findings was that the decoding accuracy for the computational model units was *higher* than the decoding accuracy based on using neural data. In particular, the decoding accuracy based on simple combinations of Gabor filters (C1 features) was higher than the decoding results based on AIT neural data, which suggests that much of the information is available in simple features to detect whether an animal was in a natural scene image was not present in the neural activity. While one could easily make the decoding results from the computational model units lower by adding noise to their responses (which in a certain sense could actually make the computational model unit better match the neural data, given that unlike the neural data, the model units at the moment to not have any variation to a particular stimulus), adding such noise would not give any additional insight into what is lacking in the neural responses that is present in simple computational features. Below we speculate on a few other reasons why the decoding accuracy from the neural data was not that high.

1. <u>The monkey was not engaged in an animal detection task.</u> While several studies have shown that IT is selective for visual features even when a monkey is passively viewing images (Keysers et al., 2001; Kiani et al., 2007) it is possible that in order for the neural population to respond similarly to images that vary greatly in their visual appearance, the monkey must be actively training (or engaged) in a relevant discrimination task. Indeed, several studies have found that when monkeys are trained to discriminate between different classes of objects, neurons in IT respond more similarly to members within a category compared to members across category boundaries (Sigala and Logothetis, 2002; Meyers et al., 2008), although these effects seem to be small relative to their overall

shape turning (De Baene et al., 2008). Also, feature based attention increases the selectivity of neurons to visual stimuli (Maunsell and Treue, 2006), which could potentially increase the selectivity of neurons in IT. Thus if the monkey were engaged in an animal discrimination task, neurons in IT would most likely be more strongly tuned to complex features that discriminate between the relevant categories, which should result in a higher population decoding accuracy.

2. The brain regions we analyzed the data from might not be the areas that are critical for rapid animal/non-animal discrimination. Neurons in AIT tend to be spatially clustered next to other neurons that have similar visual response properties (Wang et al., 1998; Tsunoda et al., 2001; Op de Beeck et al., 2007). If neurons in different areas in AIT underlie the ability to discriminate between different classes of objects, it is possible that the recordings we made were not in specific regions where the neurons that are critical for analyzing information relating to animal-like shapes. In order to test this hypothesis, we reran a stimulus set that was used by of Hung et al. (2005) and compared the decoding results to the results obtained from the data from of Hung et al. (2005), (data not shown). The results indicated that indeed there was a lower decoding accuracy on the data from the monkeys used in this study compared to the data from the monkey from Hung et al. (2005) (although we also obtained slightly above chance decoding accuracy from the Hung et al. (2005) during the baseline period before the stimulus appeared on the screen, indicating that the data we had were slightly biased). Additionally, the degree of firing rate modulation in the data from this study was less than seen in the Hung et al. (2005) data, and there was more variability in the neural responses to particular stimuli, again suggesting that differences in recording site or technique could be contributing to the less selective neural responses in this study.

It is also possible, that the ventral visual pathway is not critical for rapidly detecting animals in natural scenes and that the dorsal visual pathway could be more involved in such rapid detection tasks (Kirchner and Thorpe, 2006; Girard et al., 2008). A recent study by Girard et al, (2008) has shown that macaques can reliably make saccades to animal images within 100ms of stimuli onset and given that the latency of AIT neurons is

typically reported to be around 100ms (Nowak and Bullier, 1998), there does not appear to be enough time for AIT to actually be involved in this rapid categorization behavior. Results from our analysis (Figure 4.2C) suggest that the latency of *information* about whether an animal is in an image occurs around 125-150ms after stimulus onset, which supports the view that AIT might not be critical for rapid object categorization (at least at the level that is needed to make a saccade to an animal image). However, since the monkeys in this study were engaged in a fixation task rather than a categorization task, it is possible that the relatively long latency of information was due to the fact that the monkey was in a different behavioral state than when the monkey is engaged in a categorization task, or that the rapid sequence of image presentation created forward masking effects that delayed the neural responses. Thus based on our current results it is not possible to definitively conclude that IT is not important for rapid categorization.

3. The decoding/experimental methods we used are not adequate to extract the relevant information from the AIT neural activity. In this study we used linear classifiers to decode information from populations of AIT neurons, which is a strategy that has yielded significant insight into the function of AIT in other studies (Hung et al., 2005; Meyers et al., 2008). While we have found that generally using more complex classifiers does not affect decoding performance (for example, see Figure 4.4, and Meyers et al., 2008 supplementary ma**t**erial), it is obviously not possible to test all decoding algorithms, which leaves open the possibility that a different decoding strategy might extract more information from the population of neurons and could be more biologically relevant for this animal detection task. Of more concern is the possibility that the data we used to train the classifier was not adequate to learn the relevant function necessary to discriminate between the diverse set of images used in this dataset. While in past studies we have found as few as 5 training examples was adequate to achieve seemingly high levels of classification accuracy (Meyers et al., 2008) which is much less than the 600 training images used in this study, all past decoding studies we have been involved in have used simpler stimuli such as isolated images on a gray background, and objects that were in the same class appeared to be much more visually similar than the diverse set of

animal and scene images used here. Thus it is possible that if we had much more training data that better spanned the space of visual images of animals, classification accuracy on the neural data could potentially have been as good or better than that seen based on low level model unit features.

Apart from the fact that classification accuracy was lower using neural data than we would have expected based on the computational model unit decoding results, additional differences between the computational model units and the neural data also existed. At the population level, the predictions made about whether an animal was in an image based on using model unit data generally did not match the pattern of predictions made from using neural data that well relative to the agreement based on predictions between the neural data from the two monkeys (see table 1 and 2). Thus it seems that there is potentially explainable variability in the neural responses that is not being captured by the model units.

These results prompted us to take a closer look at the computational model's performance, which lead to a number of findings. First, we observed that the decoding accuracy based on using model units increases dramatically when a regularized classifier is used compared to when using a simple MCC classifier, which again differs from the results based on using neural data which seemed to be largely insensitive to the exact classifier used (see Figure 4.4). These findings are similar to the literature in computer vision that has shown that performance can greatly improve when more complex classifiers are used, and also to vision neuroscience literature that has previously shown roughly equivalent decoding accuracies for simple and slightly more complex classifiers (Meyers et al., 2008). We speculate that this difference might be due to differences in the distributions of model unit responses and neural responses, with the neural responses having a more Gaussian like noise-structure than the computational model unit responses.

Second, we observed that decoding accuracies were not much different based on whether simpler computational model units were used (e.g., C1 units that are supposed to model complex cell responses), compared to when more complex computational model units are

used (e.g., C3 units that are supposed to model the responses of IT neurons) (see Figure 4.5). These findings differ from the results of Serre et al. (2007) in which it was suggested that a 'full' model that used all types of computational model units outperformed simple C1 features (see supplemental material Serre et al. (2007)). Further investigation showed that the discrepancy in the results can largely be explained by the fact that when Serre et al. (2007) did their comparisons they used 4 times as many units for the full-model results than for the C1 units results, and also they used a regularization constant value that generally worked better for high level units than for C1 units. Here when we corrected for these factors, we found that the higher level model units only led to a marginal improvement in this animal/non-animal classification task (see Figure 4.5 and Figure 4.6). This suggests that the database created by Serre et al., (and used in this study) contains images with position specific features that are indicative of whether an animal is present in an image. Thus the added invariance to 2D transformations of the C2b and C3 units as compared to C1 units does not add much benefit to the task on this dataset.

The finding that low level model units work about as well as higher level model units in this animal/non-animal classification task raises questions about what are the added benefits of using these more complex units for discriminating between these categories. Recent work in computer vision has also demonstrated simple Gabor-like filters can achieve state of the art performance on many popular computer vision datasets, provided that the images of the objects in the dataset do not vary too drastically in their pose (Pinto et al., 2009). Thus for object recognition tasks in which the objects do not vary greatly in size, position, and pose, units that respond to simple features might be all that is needed in order to achieve relatively high recognition rates. Similarly, behavioral work in humans and monkeys (Kirchner and Thorpe, 2006; Girard et al., 2008) has also led to the suggestion that the complex feature selectivity seen in AIT neurons might not be involved in the rapid discrimination of whether an animal is in an image, and instead that a more direct path that goes from V4 to the LIP and the FEF might underlie this rapid categorization behavior. In agreement with this theory, recent studies of LIP and FEF have shown that it is possible to discriminate between simple visual shapes based on the

neural activity from these areas (Sereno and Maunsell, 1998; Lehky and Sereno, 2007; Peng et al., 2008) (however testing whether LIP and FEF neurons can discriminate between more complex shapes is still needed).

Of course this raises the questions of what role does AIT plays in visional recognition. While we do not have a full answer, we can speculate that perhaps AIT is involved in a more detailed analysis of an image that occurs after an initial quick recognition and is perhaps useful for recognizing objects across highly different poses, positions, sizes, and other more complex images transformations (and/or AIT could be involved in processing that is involved in linking visual information to memory and decision based systems in the hippocampus and the prefrontal cortex). Indeed, visual responses of neurons in AIT do appear to generalize more across image transformations than neurons in (Janssen et al., 2008), supporting this theory. Thus, perhaps the visual system uses a two-staged processing strategy in which a fast coarser recognition is carried out first by neurons in the dorsal stream that respond to simple features, followed by a more detailed analysis that occurs in AIT. Such a system would could explain the chicken and egg like problem of being able to fixate on relevant objects of interest before knowing exactly what the object is. Additionally, such a coarse-to-detailed recognition strategy has been shown to be an extremely efficient method used in computer vision for the detection of faces (Viola and Jones, 2004), and perhaps a similar strategy would also be an effective for object recognition in general.

Acknowledgments

This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL). This research was sponsored by grants from DARPA (IPTO and DSO), National Science Foundation (NSF-0640097, NSF-0827427.). Additional support was provided by: Adobe, Honda Research Institute USA,

King Abdullah University Science and Technology grant to B. DeVore, NEC, Sony and especially by the Eugene McDermott Foundation. E.M. was also supported by the Hubert Schoemaker Graduate Student Fellowship, and the National Defense Science and Engineering Graduate Fellowship.

References

Cadieu C, Kouh M, Pasupathy A, Connor C, Riesenhuber M, Poggio T (2007) A model of V4 shape selectivity and invariance. Journal of Neurophysiology 98:1733-1750

Chang C, Lin C (2001) LIBSVM: a Library for Support Vector Machines. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.9020 [Accessed August 28, 2009].

De Baene W, Ons B, Wagemans J, Vogels R (2008) Effects of Category Learning on the Stimulus Selectivity of Macaque Inferior Temporal Neurons. Learning & Memory 15:717-727

Delorme A, Richard G, Fabre-Thorpe M (2000) Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. Vision Res 40:2187-2200

Delorme A, Rousselet G, Mace M, Fabre-Thorpe M (2004) Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. Cognitive Brain Research 19:113, 103

Fabre-Thorpe M, Richard G, Thorpe SJ (1998) Rapid categorization of natural images by rhesus monkeys. Neuroreport 9:303-308

Freedman D, Riesenhuber M, Poggio T, Miller E (2003) A comparison of primate prefrontal and inferior temporal cortices during visual categorization. Journal of Neuroscience 23:5235-5246

Freedman D, Riesenhuber M, Poggio T, Miller E (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. Science 291:312-316

Freedman D, Riesenhuber M, Shelton C, Poggio T, Miller E (2000) Categorical representation of objects in the primate prefrontal cortex. Journal of Cognitive Neuroscience:143-143

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

Girard P, Jouffrais C, Kirchner C (2008) Ultra-rapid categorisation in non-human primates. Animal Cognition 11:727

Golland P, Fischl B (2003) Permutation Tests for Classification: Towards Statistical Significance in Image-Based Studies In Information Processing in Medical Imaging , p. 341, 330.

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

Janssen P, Srivastava S, Ombelet S, Orban GA (2008) Coding of Shape and Position in Macaque Lateral Intraparietal Area. J. Neurosci. 28:6679-6690

Keysers C, Xiao DK, Földiák P, Perrett DI (2001) The speed of sight. J Cogn Neurosci 13:90-101

Kiani R, Esteky H, Mirpour K, Tanaka K (2007) Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. Journal of Neurophysiology 97:4296-4309

Kirchner H, Thorpe SJ (2006) Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. Vision Res 46:1762-1776

Lehky SR, Sereno AB (2007) Comparison of Shape Encoding in Primate Dorsal and Ventral Visual Pathways. J Neurophysiol 97:307-319

Li FF, VanRullen R, Koch C, Perona P (2002) Rapid natural scene categorization in the near absence of attention. Proc Natl Acad Sci U S A. 99:9596–9601

Macé MJ, Richard G, Delorme A, Fabre-Thorpe M (2005) Rapid categorization of natural scenes in monkeys: target predictability and processing speed. Neuroreport 16:349-354

Maunsell JHR, Treue S (2006) Feature-based attention in visual cortex. Trends Neurosci 29:317-322

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 100:1407-19

Nowak L, Bullier J (1998) The timing of information transfer in the visual system In J. H. Kaas, K. Rockland, & A. Peters, eds. Cerebral Cortex New York: Plenum, p. 205-241.

Op de Beeck HP, Deutsch JA, Vanduffel W, Kanwisher NG, DiCarlo JJ (2007) A Stable Topography of Selectivity for Unfamiliar Shape Classes in Monkey Inferior Temporal Cortex. Cereb. Cortex:bhm196

Peelen MV, Fei-Fei L, Kastner S (2009) Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature advanced online publication Available at: http://dx.doi.org/10.1038/nature08103 [Accessed June 15, 2009].

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

Pinto N, DiCarlo J, Cox D (2009) How far can you get with a modern face recognition test set using only simple features?

Rifkin R, Lippert R (2007) Notes on Regularized Least Squares. MIT. Available at: http://hdl.handle.net/1721.1/37318.

Rolls ET, Treves A, Tovee MJ (1997) The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. Experimental Brain Research 114:149-162

Rolls ET, Aggelopoulos NC, Zheng F (2003) The Receptive Fields of Inferior Temporal Cortex Neurons in Natural Scenes. J. Neurosci. 23:339-348

Rousselet GA, Fabre-Thorpe M, Thorpe SJ (2002) Parallel processing in high-level categorization of natural images. Nat Neurosci 5:629-630

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

Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, Poggio T (2005) 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

Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences of the United States of America 104:6424-6429

Sheinberg DL, Logothetis NK (2001) Noticing Familiar Objects in Real World Scenes: The Role of Temporal Cortical Neurons in Natural Vision. J. Neurosci. 21:1340-1350

Shima K, Isoda M, Mushiake H, Tanji J (2007) Categorization of behavioural sequences in the prefrontal cortex. Nature 445:315-318

Sigala N, Logothetis N (2002) Visual categorization shapes feature selectivity in the primate temporal cortex. Nature 415:318-320

Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381:520-522

Tsunoda K, Yamane Y, Nishizaki M, Tanifuji M (2001) Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. Nat Neurosci 4:832-838

Vapnik VN (1995) The nature of statistical learning theory. Springer-Verlag New York, Inc. Available at: http://portal.acm.org/citation.cfm?id=211359 [Accessed August 28, 2009].

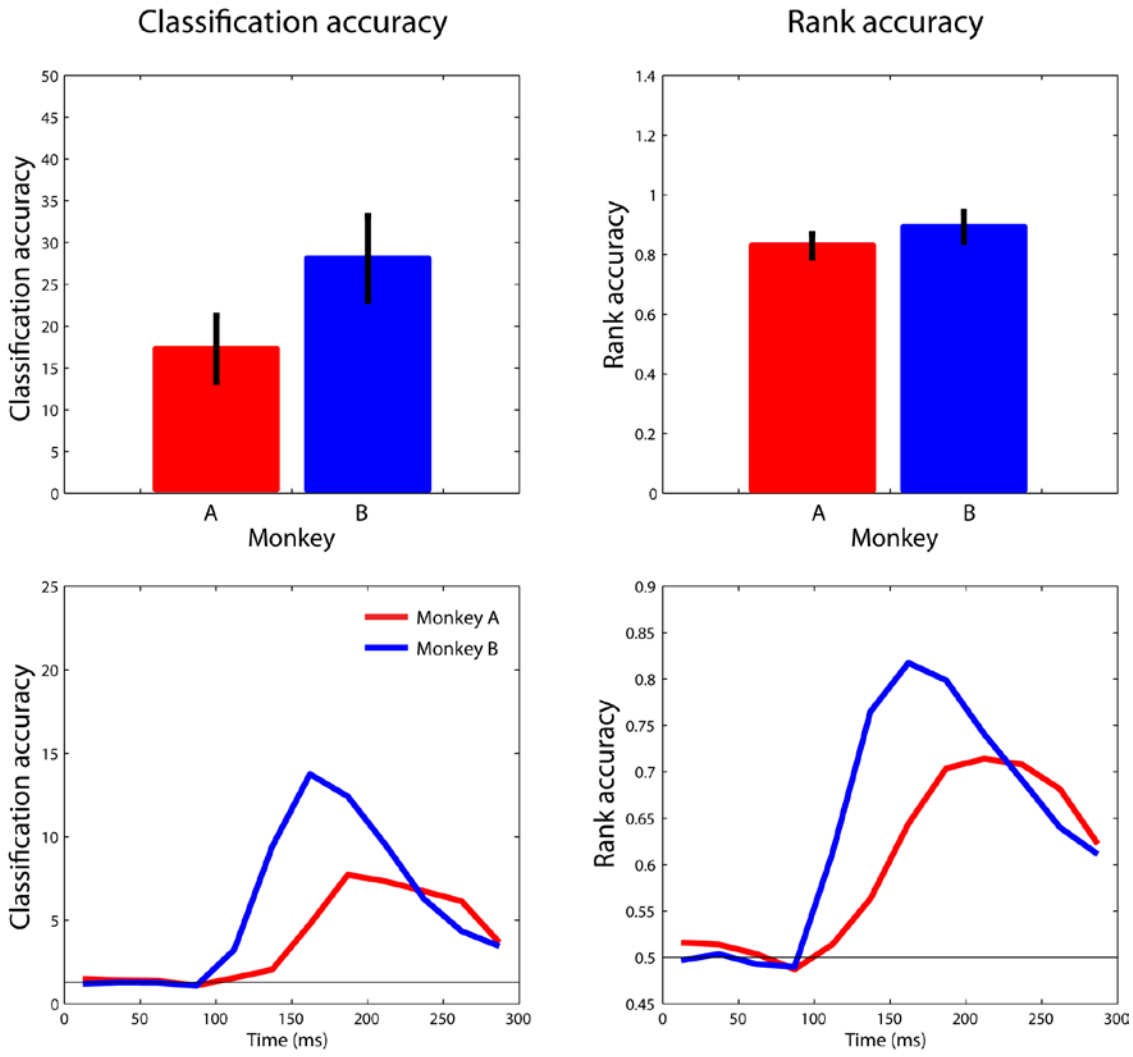
Viola P, Jones MJ (2004) Robust Real-Time Face Detection. Int. J. Comput. Vision 57:137-154

Wang G, Tanifuji M, Tanaka K (1998) Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. Neuroscience Research 32:33-46

Wegener D, Freiwald WA, Kreiter AK (2004) The influence of sustained selective attention on stimulus selectivity in macaque visual area MT. J. Neurosci 24:6106-6114

Wilson M, McNaughton B (1993) Dynamics of the hippocampal ensemble code for space. Science 261:1055-1058

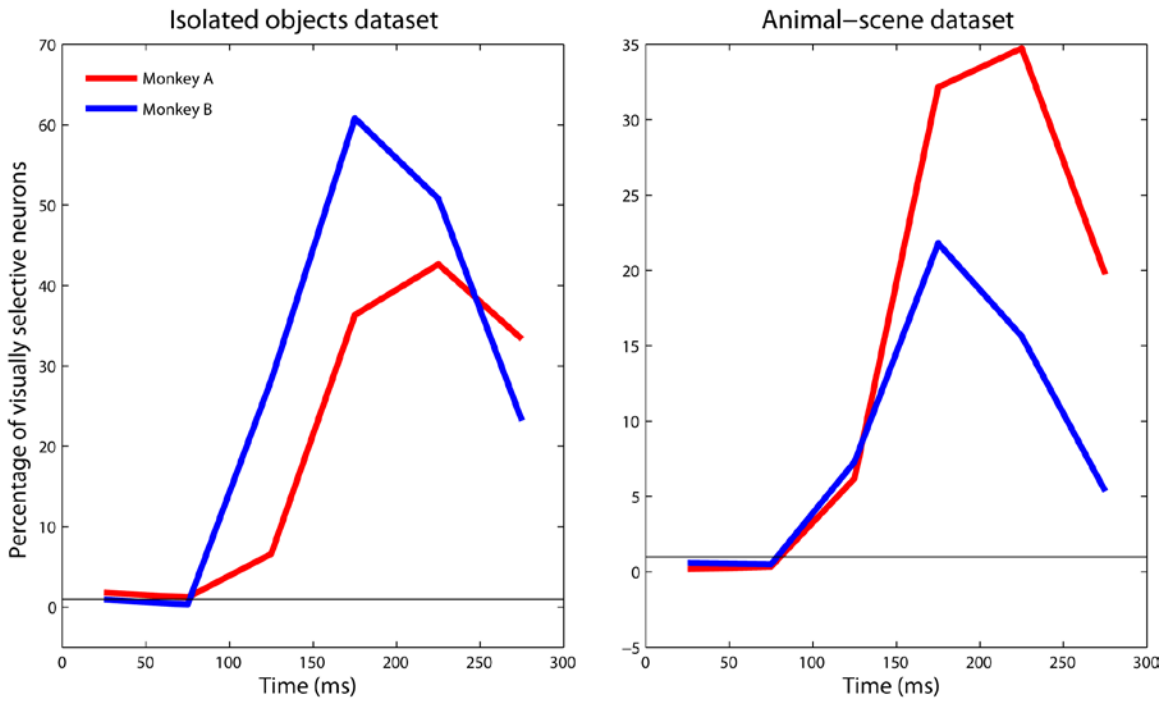
Supplementary material



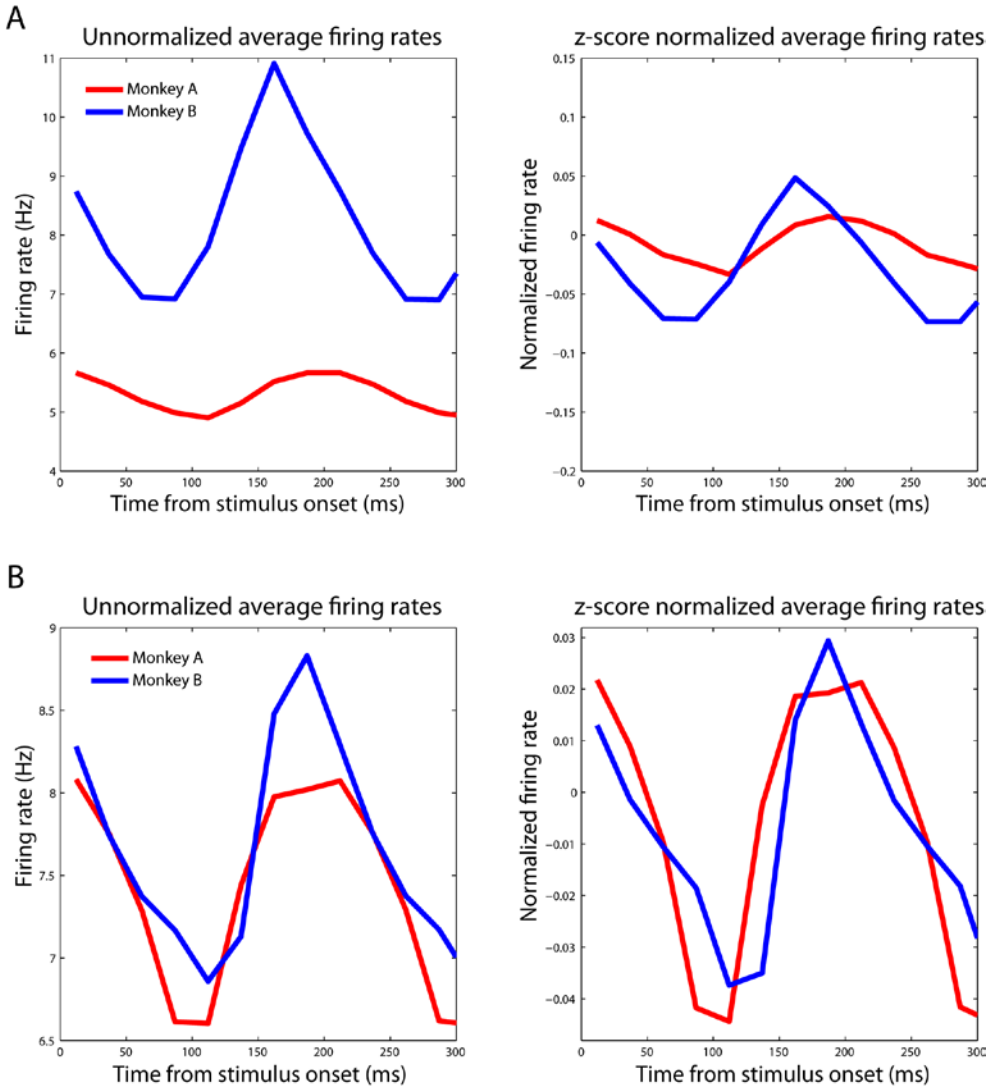
Supplemental figure 4.1 Decoding accuracy for the 77 isolated object stimuli used by Hung et al., (2005). Upper plots show decoding accuracy and rank accuracy using the average firing rates in a bin taken 100-300ms after stimulus onset, while the lower plots show decoding accuracies using 25ms sliding bins taken at 25ms intervals (i.e., a separate classifier was trained and tested using data sampled every 25ms). Plots on the left show the raw classification accuracy, where the black horizontal line represents change (1/77). Plots on the right show the 'normalized rank' accuracy which shows that when a classifier gives an ordered list of predictions, how far down the list was the label of the actual image that was shown (a normalized rank of .5 indicates chance performance). All the results shown above were lower than the accuracies obtained from the data used by Hung et al. (2005) (results now shown), although there appeared to be above chance classification accuracy during the baseline period in the Hung et al., (2005) dataset, so it is not clear if the higher accuracy on that dataset is an artifact. Also monkey A in this study was myopic which could have led to lower decoding accuracies from his data.

	Monkey A		Monkey B	
77 object	54.05	(51.65)	66.77	(64.26)
animal/non-animal	59.74	(48.71)	33.09	(23.91)

Supplemental table 4.1 Percentage of neurons that were 'visually selective' as determined by either an ANOVA or a Kruskal Wallis test (numbers in parentheses) using the mean firing rates in a 200ms bin that started 100ms after stimulus onset. As can be seen for the 77 objects, monkey B had a higher percentage of selective of selective neurons than monkey A. For the animal/non-animal data, the number of selective neurons was found using and ANOVA (or Kruskal Wallis test) separately for each 120 block of images and then averaged over the 10 blocks to counter the effects of non-stationarity in firing rate over the course of the experiment that can lead to an upward bias in the number of selective neurons. Results for the animal/non-animal data show that monkey A had a slightly higher percentage of selective neurons than monkey B. These results show a very similar pattern to the decoding results in seen in Figure 4.2 and Supplemental figure 4.1 in terms of how 'good' the neural responses were from the different monkeys.



Supplemental figure 4.2 Percentage of visually selective neurons found using a Kruskal-Wallis test (i.e., percent of neurons that had p-values less than 0.01 with the different images as conditions in the test) using 50ms sliding bins for the isolated objects data (left) or the animal-scenes data (right). The results from the percent of selective neurons from isolated objects data look very similar to decoding results from this data (Supplemental figure 4.1), with the percentage of selective neurons from the data from monkey B being higher than the percentage of selective neurons from monkey A. Since the alpha level for this test was set to 0.01, the number of selective neurons should be approximately 1% during the baseline period. For the animal/non-animal data, the number of selective neurons was found using and ANOVA (or Kruskal Wallis test) separately for each 120 block of images and then averaged over the 10 blocks to counter the effects of non-stationarity in firing rate over the course of the experiment that can lead to an upward bias in the number of selective neurons. The results show similar patterns as seen in Figure 4.2A with Monkey A higher more selective neurons than Monkey B, although the difference here appears even greater.



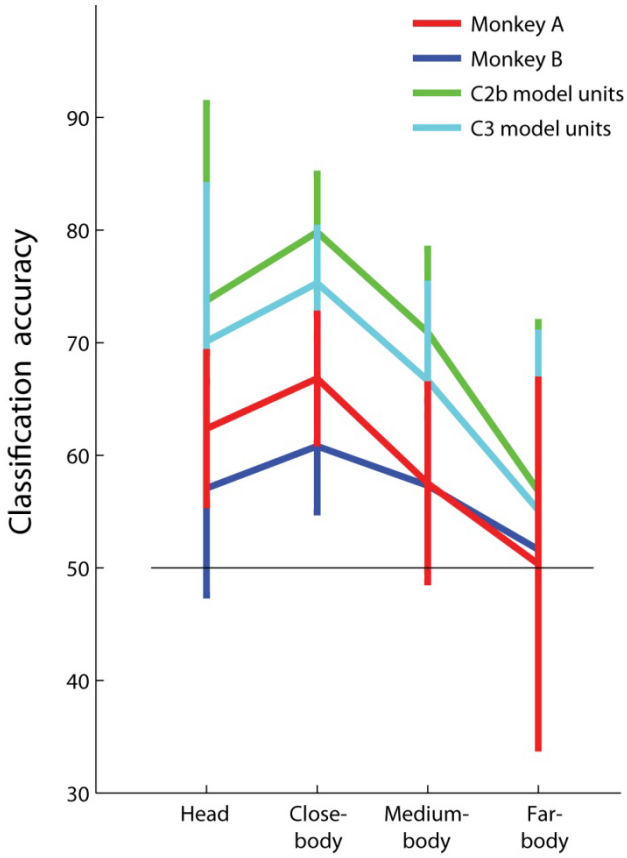
Supplemental figure 4.3 Average firing rates of neurons in the 77 isolated object experiment (A), and the average firing rates for the animal-scenes data (B). Plots on the left show the average of the raw firing rates of all neurons, while plots on the right show the average firing rate once each neuron was z-score normalized by having each neuron's firing rate have a zero mean firing rate and a standard deviation of one over the time course of a trial. The normalized firing rate give a better sense of the modulation of the population since neurons with overall higher firing rates do not dominate the average. As can be seen for the 77 isolated object experiment (A), the modulation in firing rate from the data recorded from monkey A (red trace) is less than the modulation in firing rate from the data recorded from Monkey B (blue trace). In the animal-scenes experiment, the modulation in firing rates for the two monkeys appears somewhat comparable. We are not sure why there is a different in level of neural modulation between the two experiments from Monkey A. However we do note that this difference mirrors the difference seen in the decoding accuracies in which the decoding accuracy for monkey A seems to be lower than the decoding accuracy from Monkey B in the isolated object experiment, but the decoding accuracies from both monkeys appear comparable in the animal/non-animal decoding experiments.

	77 isolated objects		Animal/non-animal	
	Median CV	Max time	Min time	Max time
Monkey A	112	162	112	187
Monkey B	87	137	112	162

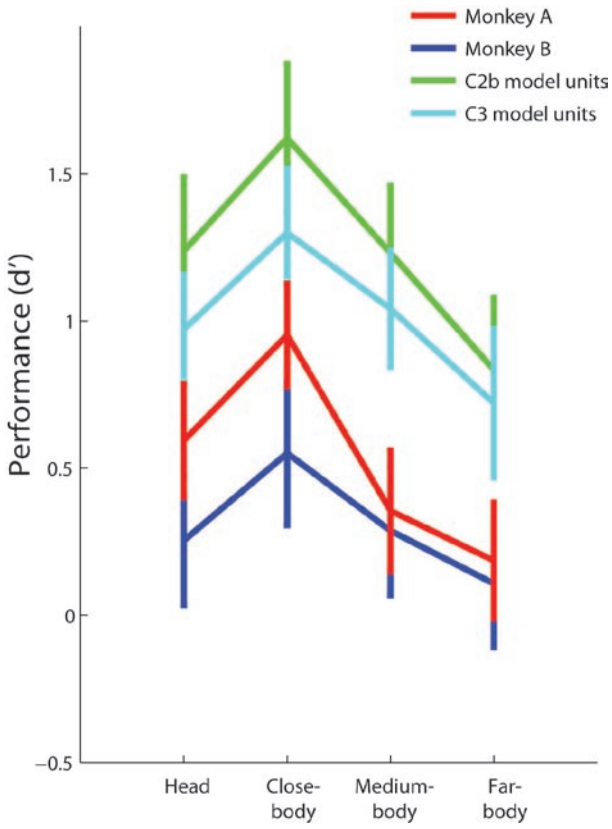
Supplemental table 4.2 The bin that had the minimum and maximum firing rate taken from Supplemental figure 4.2. Values were based on the average firing rates over all neurons using 25ms bins sampled every 25ms (values in parenthesis are given for the normalized average firing rate when they differ from the raw firing rate min or max).

	sep ID ave	p-val	over all ID	p-val
Monkey A	1.47	(A, B), $p = .13$	1.42	(A, B), $p = .23$
Monkey B	1.36	(B, H), $p < 10^{-7}$	1.36	(B, H), $p < 10^{-3}$

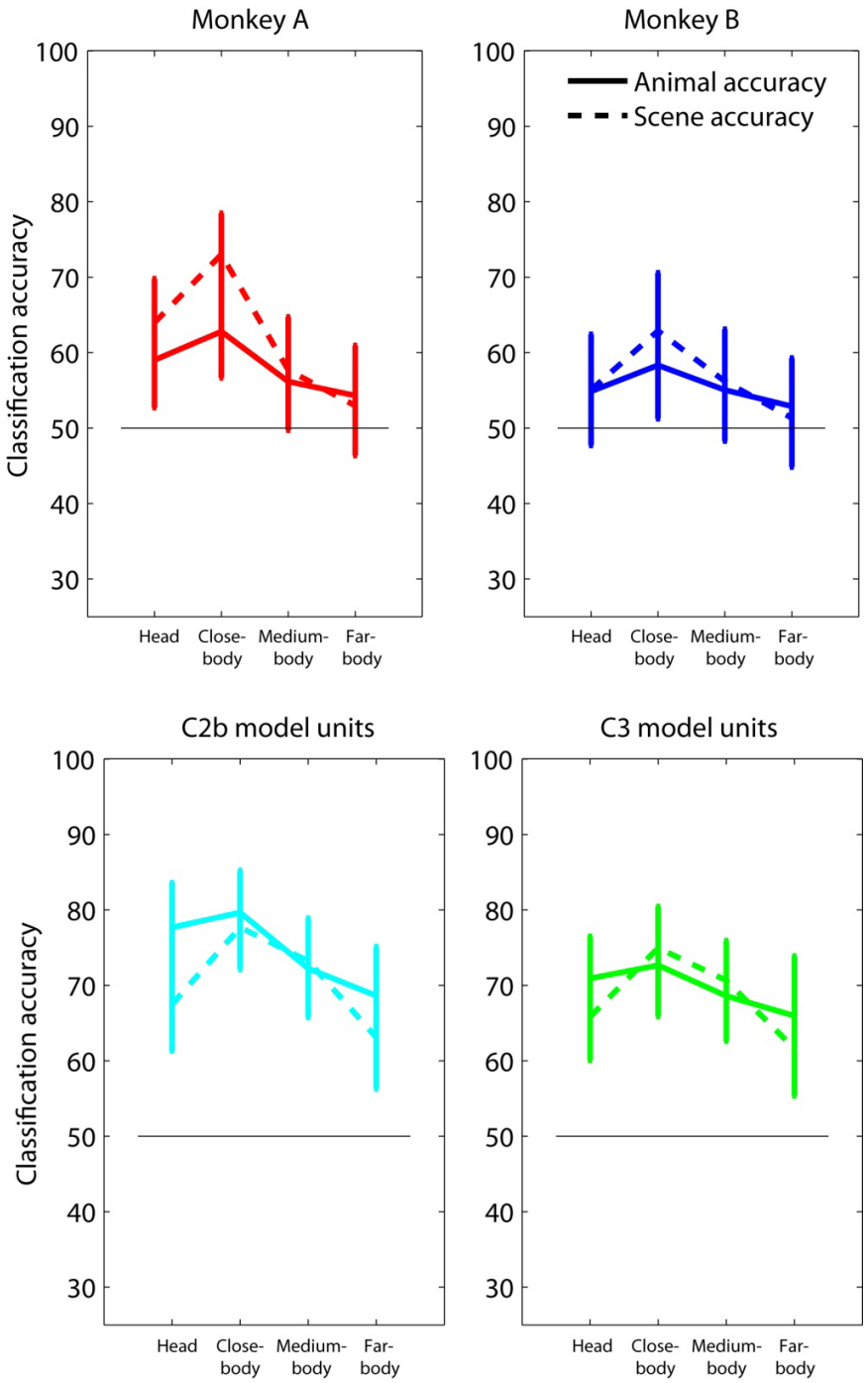
Supplemental table 4.3 Median coefficient of variation (CV = stdev/mean) values from the population of neurons from monkey A, and B using the 77 isolated objects data. For each neuron, the coefficient of variation was calculated either separately for each stimulus shown and then averaged together (sep ID ave), or calculated over trials regardless of the stimulus shown (over all ID). The median values over all the neurons are shown above. P-values using a Mann-Whitney U (which is the same as the Wilcoxon rank-sum) were calculated on the CV values for all pairs monkeys. Results from monkey H (recorded by Hung et al., 2005) had less variability compared to monkey A, and B (data not shown), which could partially account for the higher decoding accuracy seen in that monkey.



Supplemental figure 4.4 Basic animal/non-animal decoding results (same as Figure 4.2) but plotting as in terms of the percent correct classification accuracy rather than as d'.



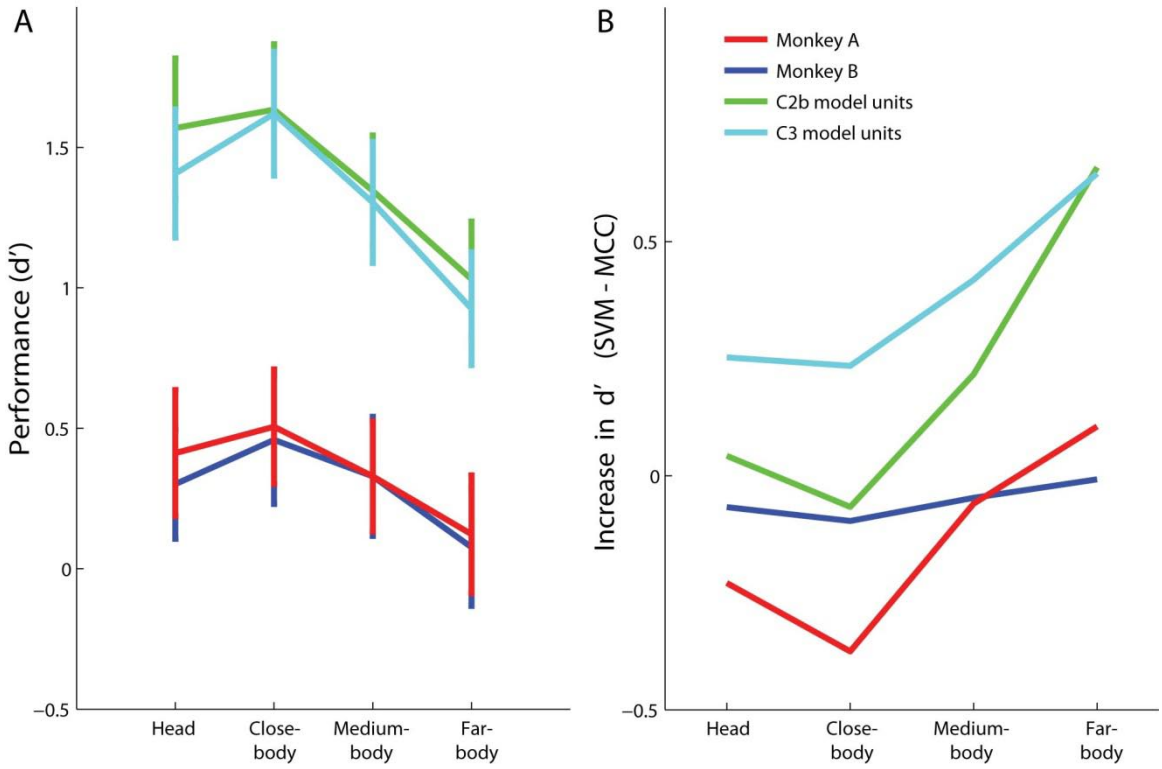
Supplemental figure 4.5 Decoding results from training and testing the classifier separately on the head, close-body, medium body and far-body conditions (rather than training on all 4 conditions jointly as was done throughout most of the paper). The same conventions are used here that were used in Figure 4.2. Notice also that the results look very similar to Figure 4.2A, indicating that training the classifier separately on each distance does not make a large difference in the results that were obtained.



Supplemental figure 4.6 Decoding results from training and testing the classifier separately on the head, close-body, medium body and far-body conditions (rather than training on all 4 conditions jointly as was done throughout most of the paper), and plotting the decoding accuracies separately for the animal and non-animal images. The same conventions are used here that were used in Figure 4.3. Notice again that the model and neural data have the similar trends. However, the far-body distance here is not below chance due to the fact that the classifier was optimized separately for each distance.

	Monkey A	Monkey B	C1	C2	C2b	C3	All units	Rand pix	Human	Serre model
Monkey A		0.48	0.24	0.22	0.26	0.19	0.25	0.09	0.40	0.38
Monkey B	0.45		0.32	0.36	0.34	0.27	0.37	0.07	0.51	0.43
C1	0.25	0.36		0.24	0.25	0.16	0.29	0.00	0.49	0.52
C2	0.24	0.36	0.26		0.68	0.71	0.85	0.02	0.42	0.50
C2b	0.28	0.36	0.27	0.74		0.91	0.88	0.21	0.50	0.56
C3	0.21	0.29	0.18	0.75	0.93		0.88	0.18	0.40	0.45
all units	0.27	0.38	0.32	0.88	0.92	0.92		0.12	0.51	0.58
Rand pix	0.11	0.12	0.04	0.06	0.21	0.19	0.16		0.21	0.15
Human	0.37	0.48	0.51	0.42	0.50	0.42	0.51	0.25		0.72
Serre model	0.36	0.44	0.54	0.52	0.61	0.50	0.62	0.20	0.71	

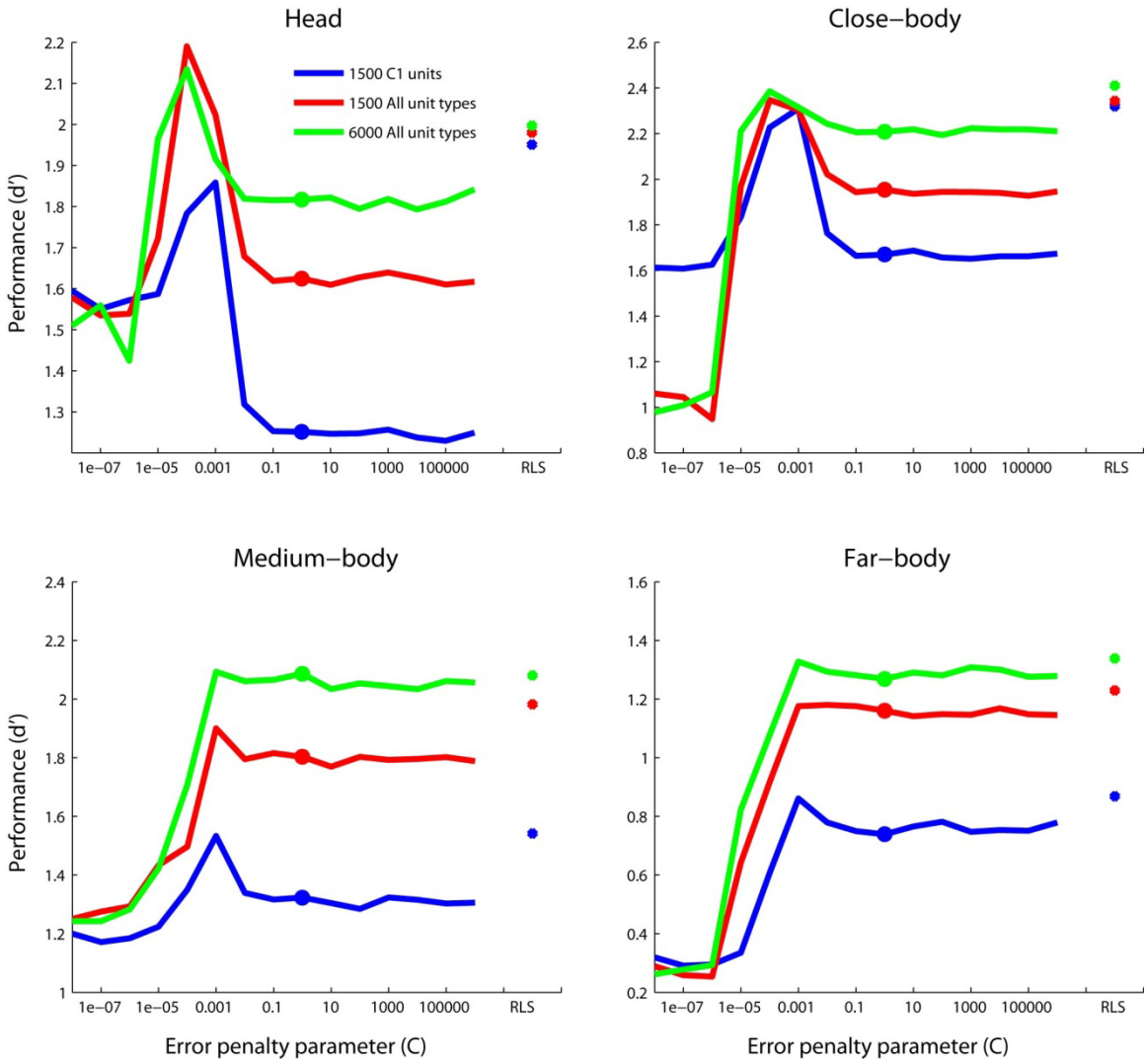
Supplemental table 4.4 Correlation coefficient values between how often each of the 1200 images were predicted as containing an animal based on using human psychophysics results, classification accuracies from neural or computational model units. The results are the same as Table 1 but with additional correlations of C1, C2, units, and random pixel decoding results. Upper triangular results are based on Pearson's correlation coefficient and lower triangular results are based on Spearman's correlation coefficient.



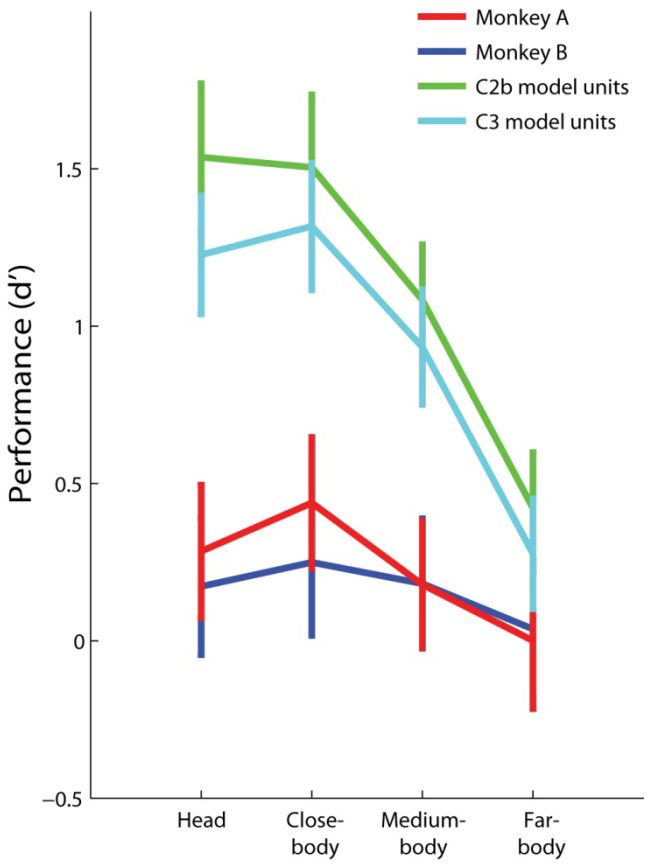
Supplemental figure 4.7 Decoding accuracies for whether an animal is in a cluttered scene image using a support vector machine (SVM) classifier (same as Figure 4.4, but using a SVM instead of an RLS classifier). For these plots, the error penalty constant value was set to C = 1, which is the same value that was used by Serre et al. (2007). A: Results plotted in the same format as Figure 4.2A and Figure 4.4A, except that a SVM classifier is used instead of an MCC or RLS classifier. B: The change in d' classification accuracy when using a SVM classifier compared to when using an MCC classifier (values calculated by subtracting the MCC decoding accuracies in Figure 4.2A from the SVM classification accuracies shown in Figure 4.4A). As can be seen, using a SVM improves the classification performance of the computational model units on the farther image distances, while leading to a decrease neural decoding results for monkey A at the close distances. However overall, the pattern of results look the same (as can be seen in A).

	Monkey A	Monkey B	C1	C2	C2b	C3	All units	Rand pix	Human	Serre model
Monkey A		0.49	0.33	0.35	0.36	0.36	0.36	0.13	0.41	0.39
Monkey B	0.47		0.47	0.47	0.44	0.45	0.47	0.09	0.51	0.44
C1	0.34	0.49		0.94	0.78	0.79	0.91	0.14	0.62	0.70
C2	0.35	0.47	0.94		0.78	0.79	0.92	0.13	0.66	0.74
C2b	0.37	0.46	0.80	0.80		0.95	0.92	0.21	0.66	0.77
C3	0.36	0.47	0.80	0.81	0.93		0.93	0.19	0.66	0.75
all units	0.37	0.49	0.91	0.93	0.92	0.93		0.18	0.69	0.79
Rand pix	0.13	0.09	0.17	0.15	0.21	0.19	0.19		0.24	0.19
Human	0.39	0.49	0.63	0.65	0.66	0.65	0.69	0.25		0.72
Serre model	0.38	0.45	0.71	0.75	0.77	0.76	0.79	0.20	0.71	

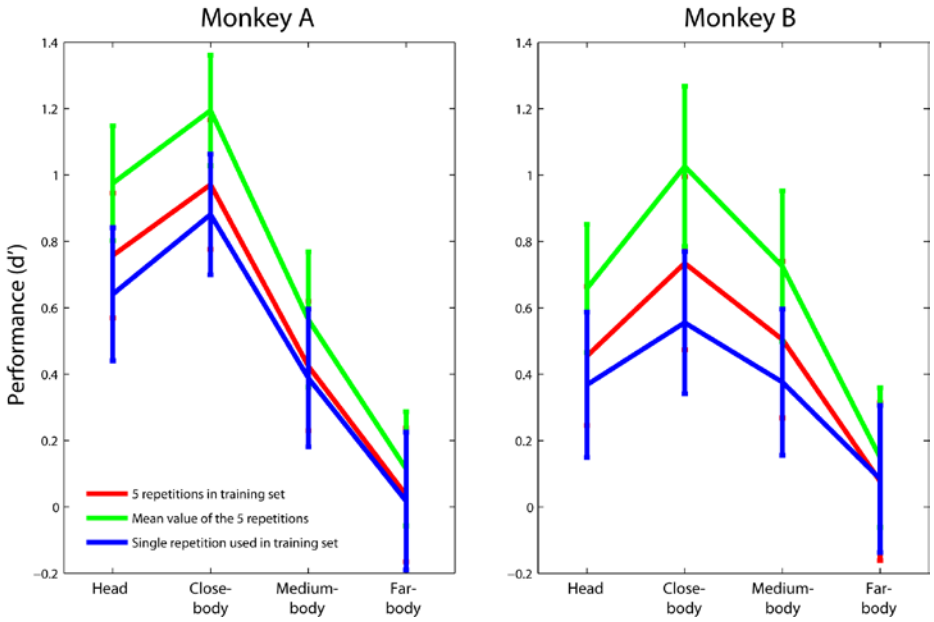
Supplemental table 4.5 Correlation coefficient values between how often each of the 1200 images were predicted as containing animals based on using human psychophysics results, classification accuracies from neural or computational model units when an RLS classifier was used. The results are the same as Table 2 but with additional correlations of C1, C2, units, and random pixels decoding results. Upper triangular results are based on Pearson's correlation coefficient and lower triangular results are based on Spearman's correlation coefficient.



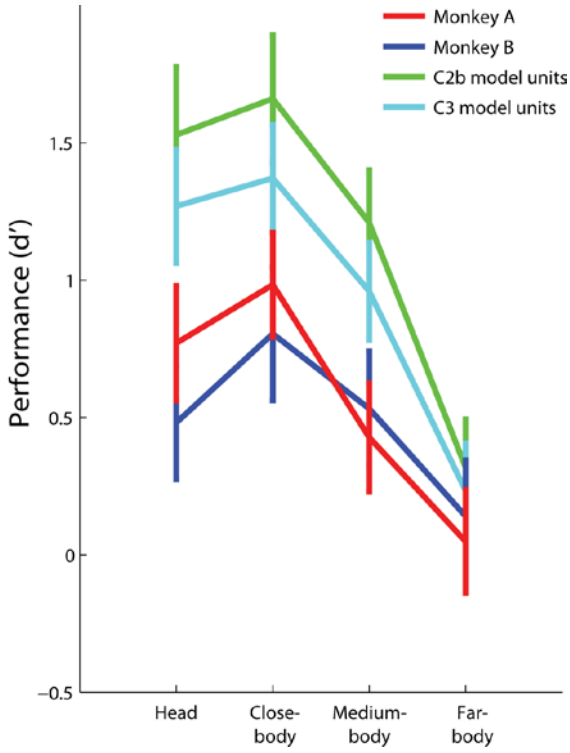
Supplemental figure 4.8 SVM animal/non-animal decoding results as a function of the error penalty parameter (C), plotting separately for the head, close-body, medium-body and far-body conditions. As can be seen, the best regularization constant value is around 0.001 for most image distances and model unit conditions. For the close-body condition at the peak error penalty constant value there is not much difference in the decoding results between using C1 and model units of all types. For the other conditions, generally there is a slight advantage to using all unit types and using more units (although this advantage tends to be smaller around the optimal error penalty constant value, than when compared to the default error penalty constant value C=1 that was used by Serre et al. (2007)). Also shown on the right of each plot are the decoding RLS results (same as Figure 4.6B) that were found by optimizing the error penalty constant using only the training data.



Supplemental figure 4.9 Results from decoding each block of 120 images separately and then combining the results together (errorbars are stdevs over all decoding results from all blocks combined). Since the neural data had more similar firing rates within each block of 120 images compared to firing rates across blocks of images (which could be due to either electrode drift or to associations formed by repeatedly showing the same group of images together), we thought it might be possible to achieve higher accuracy on the neural data by separately on each block since it would eliminate the within block similarity confound. However, the results show that if anything, training separately on each block led to lower decoding accuracy of the neural data, with the results from the computational units remaining largely the same. We speculate that perhaps the neural results were lower in the blocked readout paradigm because there were fewer training points used on each decoding block, although this does not explain why the computational model unit results remain largely unchanged (although perhaps because the computational model units are less variable in their response so a smaller training set is sufficient).



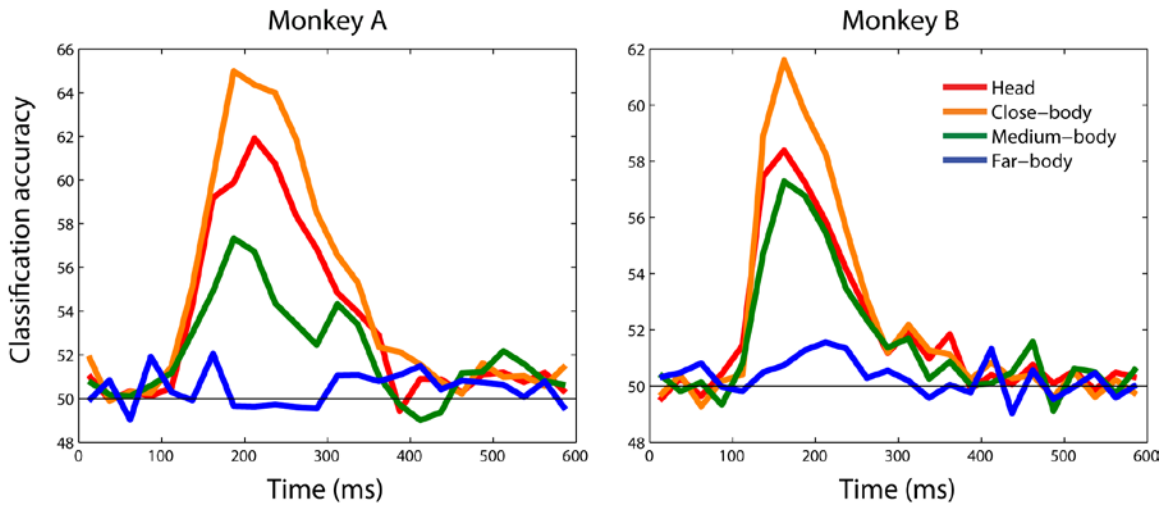
Supplemental figure 4.10 Comparison of animal/non-animal decoding results using different number of training and testing points from 100 neurons in AITd using a MCC classifier for Monkey A (left plot) and Monkey B (right plot). Red traces shows the results when data from 5 presentations of each image were in the training set, giving rise to 3000 training points and 600 test points on each bootstrap-like iteration (this is the same paradigm that was used for all neural decoding results in the paper). Blue trace shows the decoding results when data from only a single trial was in either the training or test set, giving rise to 600 training points and 600 test points. The green trace shows the results from first averaging together all 5 repetitions for each image, and then applying the full decoding paradigm using the averaged data with 600 training and 600 test images on each bootstrap-like iteration. As can be seen, using data from 5 repetitions for each image in the training (red trace) set has a slightly better performance than using data from only 1 repetition of each image type (blue trace). Averaging together the results for all 5 trials for each stimulus and then applying the decoding procedure (green trace) led to slightly higher results. This is not surprising since averaging the results reduces the large amount of noise that can be present on a specific individual trial – however given the fact that such averaging is not representative of the amount of information that is available on actual individual trials, we used the more realistic analysis of decoding data from single trials in the body of the paper.



Supplemental figure 4.11 Animal/non-animal decoding with data within each block of 120 images z-score normalized. Examining the neural data carefully revealed that it contained slow temporal trends that which resulted in the slow increases and decreases in the mean firing rates of neurons that seemed to be unrelated to the stimuli being presented. These slow trends, combined with the block design used, resulted in images within a block being biased to have slightly more similar firing rates than images in different blocks. To see if these slow trends had a large affect on decoding accuracy we normalized the firing rates for all trials that occurred within a block to have a mean of zero and a standard deviation of one (we also applied this normalization to the Model units above to be consistent). We then applied the same decoding procedure used in Figure 4.2A. The results plotted above shown that overall the decoding accuracy for the neural data was slight higher when this normalization was applied, but overall the results are very similar.

	Monkey A	Monkey B	C2b	C3	Human	Serre full model
Monkey A		0.46	0.20	0.13	0.36	0.33
Monkey B	0.55		0.30	0.23	0.48	0.42
C1	0.31	0.40		0.91	0.45	0.52
C3	0.24	0.33	0.93		0.36	0.42
Human	0.45	0.56	0.54	0.46		0.70
Serre full model	0.43	0.51	0.60	0.51	0.75	

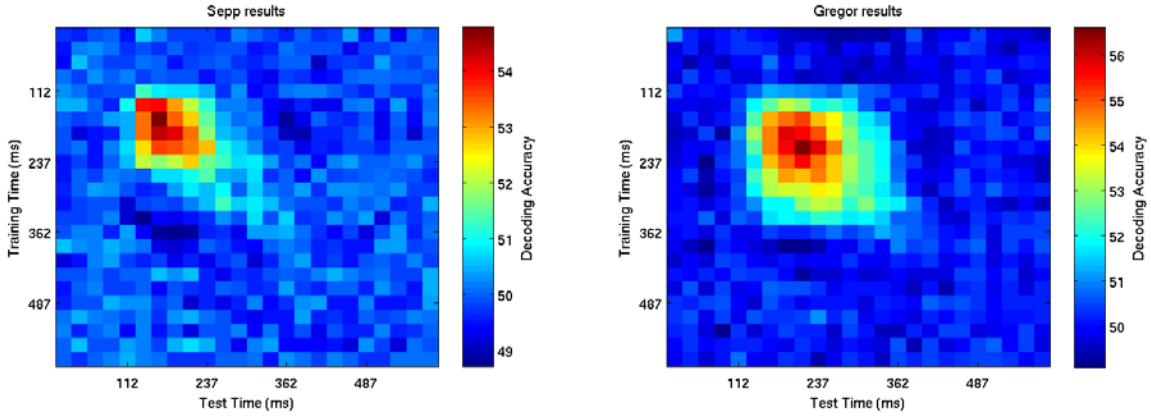
Supplemental table 4.6 The same conventions as table 1 but using the data that was z-score normalized for each block prior to the decoding algorithm was run. Z-score normalizing the data removed some of the 'noise' from the neural signal that was due to slow changes in firing rate that were unrelated to the stimuli. This lead to higher correlations between the monkeys in terms of the pattern of classification mistakes made, although the correlations between the monkeys and the Model remained about the same.



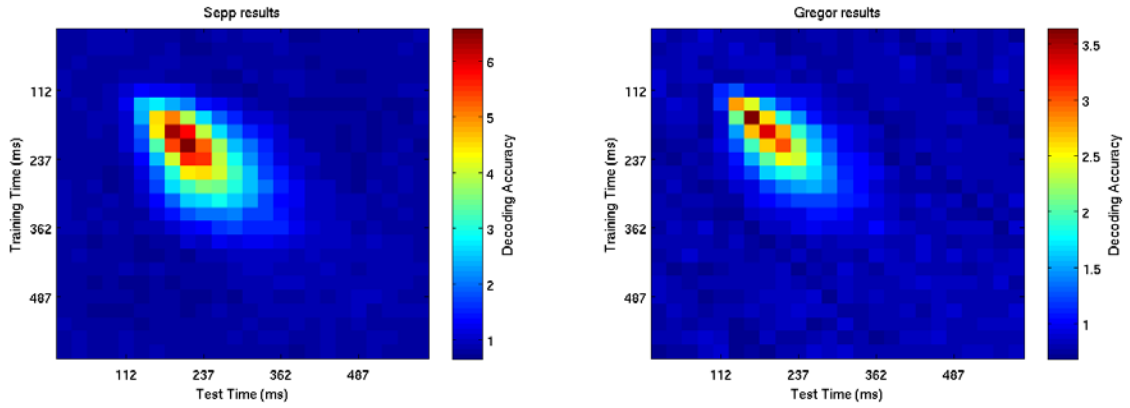
Supplemental figure 4.12 Decoding accuracies as a function of time using data that has been block z-score normalized (the same as Figure 4.2C, but the data has been block z-score normalized). As can be seen, the results look similar to Figure 2C except that the decoding accuracy is slightly higher due to the block z-score normalization which removed some slow temporal noise from the data.

	Monkey A	Monkey B	C2b	C3	Human	Serre full model
Monkey A		0.43	0.33	0.31	0.39	0.35
Monkey B	0.52		0.43	0.42	0.49	0.43
C1	0.42	0.52		0.94	0.63	0.74
C3	0.41	0.51	0.95		0.62	0.72
Human	0.49	0.57	0.70	0.69		0.70
Serre full model	0.45	0.52	0.79	0.77	0.75	

Supplemental table 4.7 The same conventions as table 1 (and table 2) but using the data that was z-score normalized for each block prior to the decoding algorithm was run and using an RLS classifier. Z-score normalizing the data again led to higher correlations between the monkeys in terms of the pattern of classification mistakes made. Here we see that the results from Monkey B are about equally well explained by either Monkey B or by the Model units (although the results from Monkey B seem to best match the human psychophysics results).



Additional supplemental material 4.1 Examining the dynamic representation of animal/non-animal information. Here we show the results from figure 4.2C (averaged over all four distances), when training the classifier at time 1 (indicated by the y-axis) and testing the classifier at time 2 (indicated by the x-axis), as was done in figure 3.6, for monkey A (left) and monkey B (right). Overall it appears that the representation is rather static, although given that the stimuli were only presented for a short amount of time, strong conclusions should not be drawn.



Additional supplemental material 4.2 Examining the dynamic representation of exact stimulus identity decoding. Here we show the results from Supplemental figure 9 (i.e., decoding the identity of the 120 images shown in each block, and then averaged the results over the 10 blocks), when training the classifier at time 1 (indicated by the y-axis) and testing the classifier at time 2 (indicated by the x-axis), as was done in figure 3.6, for monkey A (left) and monkey B (right). Overall it appears that the representation is has some slight dynamics, although given that the stimuli were only presented for a short amount of time, strong conclusions should not be drawn.

Chapter 5: Object decoding with attention in inferior temporal cortex

This work in this chapter was done as a collaboration between the Desimone lab and the Poggio lab. Ying Zhang recorded the data with the help of Narcisse Bichot. I analyzed the data and helped write the manuscript along with Robert Desimone, Tomaso Poggio. A modified version of this work has recently been submitted for publication.

Abstract

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[1-6], and have suggested that these physiological changes should, in principle, be beneficial for visual information processing. However a more computational understanding of how the visual system recognizes objects in clutter and how attention contributes has not been developed. Here, we develop a deeper computational understanding of how attention improves object recognition by assuming that visual objects are represented by patterns of neural activity in the inferior temporal (IT) cortex, and examining how attention influences these representations. 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 from populations of IT neurons[7,8]. The results show that before attention was deployed, information about 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 across the population was restored toward the pattern representing the isolated object, increasing the amount of information about this object. Increasing the saliency of nonattended objects overrode these attentional enhancements. Thus we find that specific firing rate changes can have a significant impact on the information present in IT cortex. By taking a computational perspective, this work brings us closer to an algorithmic level understanding of how attention affects object recognition, and also provides insight into which attention related physiological changes are directly related to information processing, and which are byproducts of the particular mechanisms/(implementation) that is used by the brain.

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 (see Figure 1 and Supplemental figure 5.1). The three stimuli were positioned so that each was likely to be contained within a different receptive field (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 distracters 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 1260 ms after cue onset. On half of the trials, one of the distracter 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 a distracter color change were made on only ~1% of trials (with additional fixation errors accounting for remainder of the performance).

To understand how information about objects is represented by populations of IT neurons, we applied population decoding methods[7,8] to the firing rates of pseudo-populations of 187 neurons from two monkeys on a first stimulus set (similar results were obtained from each monkey so the data were combined, see Supplemental figure 5.2) and on a second stimulus set shown to monkey 2 (Supplemental figure 5.5). 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 (see methods). Figure 2a 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 ROC curve (AUROC) of .83 +/-.022 at 225 ms after stimulus onset, while 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 .62 +/-

.014 prior to the onset of the attentional cue. An AUROC of 0.5 represents chance performance. Thus, before the attentional cue, the amount of information about each object in the three-object displays was greatly reduced compared to when these objects were shown in isolation, showing that clutter has a significant impact on the amount of information about specific objects in IT (also see Supplemental figure 5.2). After the attentional cue was displayed, information about the attended object (red trace) rose, reaching an AUROC value of .64 +/- .017 400 ms after the onset of the cue which was similar to the value of .68 +/- .024 for decoding isolated object trials during the same trial period, while information about the nonattended stimuli (green trace) decreased to a value of .56 +/- .010. Thus location-directed attention can have a significant impact on the amount of information about specific objects in IT. These attention related changes can also be observed in the firing rate of the population of neurons to preferred and non-preferred stimuli (Supplemental figure 5.3).

In addition to identity information, position information was also enhanced (Figure 5.1c). When this position enhancement was examined using more conventional analyses of firing rate tuning curves for the best to worst stimuli[1], this position enhancement appeared as a constant offset in the tuning curves of individual neurons (Figure 5.1d). However, this upward shift in tuning curves is a consequence of aligning all neuronal responses to their preferred and non-preferred stimuli – for a given stimulus and location, these position related attention effects once again created increases and decreases in activity across the population of cells, leading to a distributed pattern of activity for position information.

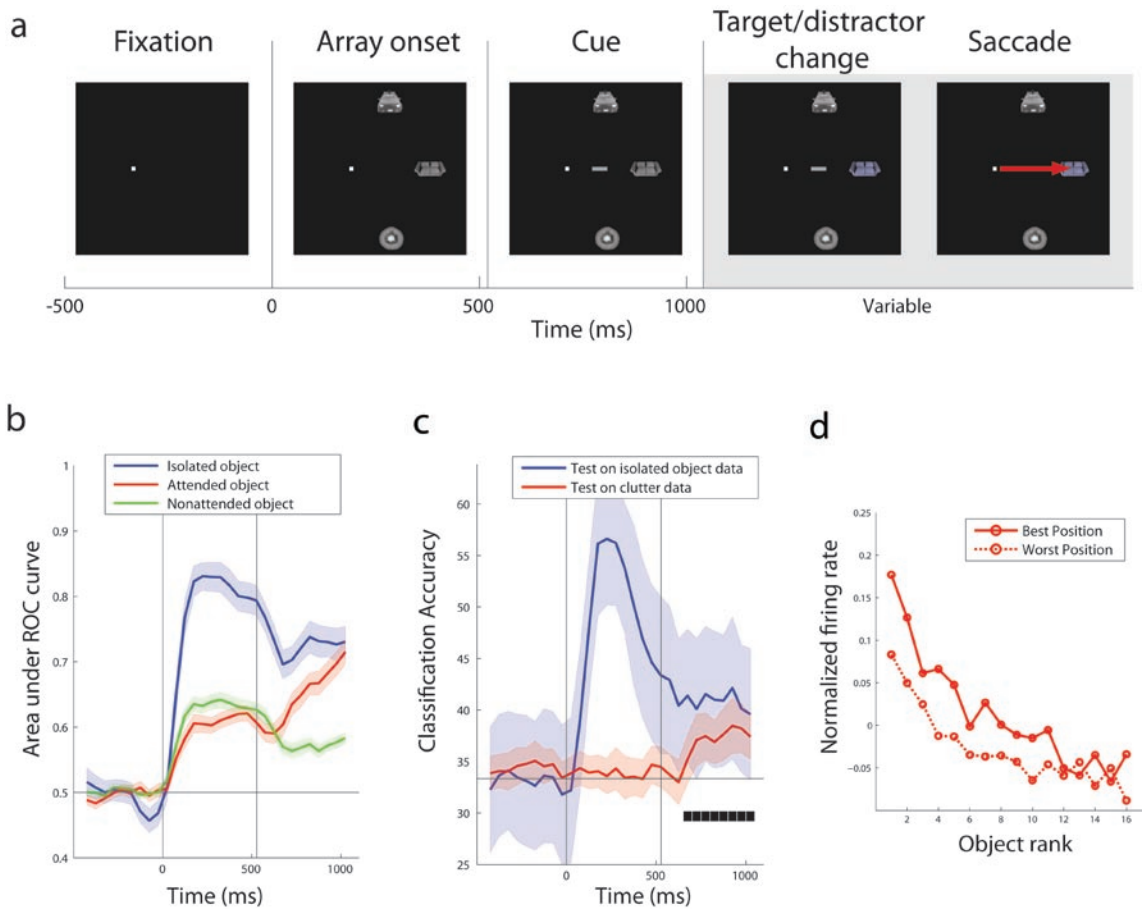


Figure 5.1. Effects of attention on decoding accuracy. a, Timeline for 3-object trials. Single object trials had the same time line except only one object was displayed. **b,** Decoding accuracies for which object was shown on isolated object trials (blue traces), and the attended object (red trace) and nonattended objects (green trace) in the 3-object displays. Vertical lines indicate the times of stimulus onset and cue onset. Colored shaded regions indicates +/- 1 stderr of the decoding results (see methods). **c**, Decoding accuracies for the position of isolated stimulus (blue trace) and the attended stimulus (red trace). Black square boxes indicate times where the decoding accuracy for the position of the attended object was above what would be expected by chance (chance performance is 33%). **d**, 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.

By training the classifier with data from isolated object trials and then evaluating the classifier with data from cluttered trials, we tested 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 trained the classifier on either cluttered display data or on isolated object data and compared the classification accuracies for decoding objects in the cluttered display data. If attention added additional information, then there should be higher accuracy when training with cluttered data than when training with isolated object data. Figure 5.2 shows that when the same amount of training data was used, training on isolated object trial data (blue trace) was always better than training on cluttered trial data (red trace). Thus, attention seemed to restore the population of neural activity to a state that was similar to when the attended object was shown in isolation.

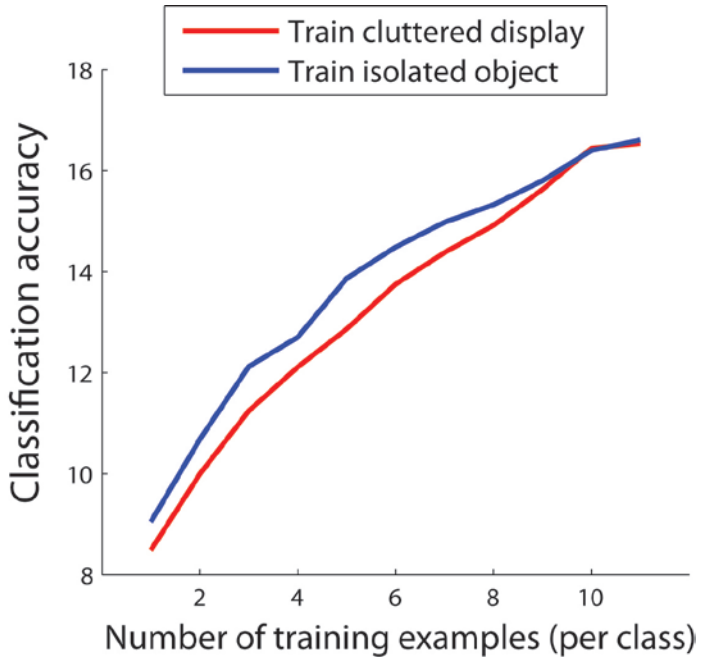


Figure 5.2. Attention restores neural activity to a state that is similar to when the attended object is presented alone. We trained a classifier using either data from isolated object trials (blue trace) as was done in figure 1, or using data from cluttered trials using the identity of the attended object (red trace). The results are plotted as a function of the number of training examples used, and the data are from the cue period (200-500 ms after cue onset). No matter how many training examples were used, the results for training with attended object in cluttered displays were never better than training with data from isolated object trials. Thus it appears that the effect of attention was to restore the neural representation to a state that was similar to when the attended object was shown alone rather than creating a new 'attention-based' representation. Chance decoding accuracy is 1/16 or 6.25%.

The above results show that top-down attention has a large impact on what information is represented in IT. However, it was not clear how these representations would be affected by "bottom-up" salient changes in distracter objects. We therefore 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 Figure 5.3, 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 briefly switched to the distractor object (light green trace), before returning to the attended object representation (red trace). Thus bottom-up changes in the saliency of the

distractor objects overrode the top-down attention induced enhancements of particular objects. An examination of behavioral data (Supplemental figure 5.4), revealed that reaction times were longer when the target changed soon after the distractor change, suggesting that the monkeys' behavior is likely to be influenced by which objects are being represented in IT.

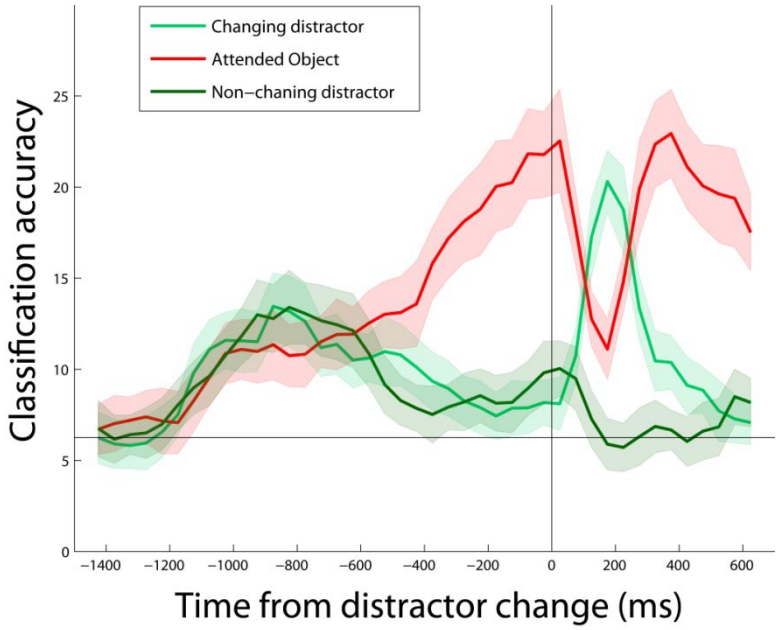


Figure 5.3 Changes in the salience of distractor stimuli dominate over attention related enhancements. A comparison of the decoding accuracies for the attended stimulus (red trace), to the distractor that underwent a color change (green trace), and the distractor that did not undergo a color change (cyan 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%.

Previous work at earlier levels of the ventral stream has shown that attention to a stimulus in the RF is correlated with increases in firing rates or effective contrast, increase in gamma synchronization, and decreases in the Fano factor and noise correlation, compared to when attention is directed outside the RF[1-4,9-12]. However, because these effects are relatively modest, and because the results have not been interpreted in terms of the computational functions of downstream brain regions, it has been unclear how object recognition at higher stages might be impacted by these physiological changes. Our

results show that in cluttered scenes, the amount of information about behaviorally relevant objects in IT is significantly increased by attention related firing rate changes. Also, because our decoding method was invariant to the overall level of firing rate in the population as a whole, we believe that on the single neuron level, the differential effects on attended and unattended objects in the same IT RF[4,13-15] captured by "biased competition" models[9,16,17] are a better description of the important information processing operations underlying object recognition, as opposed to global increases in firing rates[1] or synchrony[3] with attention. With clutter in the RF, firing rates for a given stimulus may increase or decrease with attention, but the important result is that the pattern of activity across the population is restored to the pattern that would have been obtained by the object in isolation.

One limitation of this study is that most of the neurons were not recoded simultaneously, and thus it was not possible for us to directly test how noise correlations or synchrony would impact the population decoding results. However when we artificially added noise correlations to our pseudo-populations, in a way that replicated the noise correlations seen in several experiments, we found the results were largely unchanged. Thus, overall our results support the view that the main goal of attention is to suppress clutter in order to allow higher modules to recognize an object in clutter after learning its appearance from isolated presentations (or in a different clutter).

Methods

Experimental procedures. Procedures were done according to NIH guidelines and were approved by the MIT Animal Care and Use Committee. Single 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 supplemental figure 1. The stimuli were 2.3˚

x 2.3˚ in size and were shown at an eccentricity of 5.5˚ from fixation, at angles of +60˚, 0˚ and -60˚ relative to fixation. The stimulus size/locations were chosen such that there would be little overlap between the three simultaneously presented stimuli in terms of most V4 neurons' RFs[27]. For the 3 object displays, 864 configurations were chosen (out of the possible 3,360 permutations), two-thirds of which consisted of three objects from the same category. A second set of 7 stimuli were also shown to the second monkey for additional results presented in Supplemental figure 5.5; all configurations of the 630 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 (Supplemental figure 5.3, Supplemental figure 5.5d). For the population decoding analyses, all neurons that had at least 12 representations of the isolated objects and 800 presentations of cluttered trials were included; this gave 75 neurons from monkey 1 and 112 neurons from monkey 2. Since different three-object images were shown to different neurons, we only used the three-object images that had been shown to all neurons, which gave data from 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.

Decoding analyses. The decoding results were based on a cross-validation procedure that has previously been described[8]. Briefly, the decoding algorithm works by: 1) randomly selecting a number of trials for each stimulus from each neuron and creating pseudo-populations responses (i.e., fake 'populations' responses from neurons that were recorded independently but treated as if they had been recorded simultaneously); 2) dividing these pseudo-populations responses into a *k* different splits, with each split have at least one pseudo-population response to each stimulus; 3) creating a training set using *k-1* of these splits and a test set using the remaining split; 4) z-score normalizing the responses of each neuron of each trial using the mean and standard deviation of each neuron over all trials in the training set (this ensures that neurons with high firing rates do

not completely dominate the decoding); 5) training a pattern classifier to discriminate between the different experimental stimuli/conditions using the training set and testing the classifier's performance on the test set. 6) repeating this procedure from step 3 *k* times, using a different split for the test set each time, and 7) repeating this whole procedure from step 1 fifty times with different pseudo-populations each time. Standard errors for the decoding accuracy were estimated by repeating the decoding procedure but creating pseudo-populations by sampling neurons with replacement (being careful not to include any of the same data in the training and test sets) and then taking the standard deviation of the mean decoding accuracy over the 50 bootstrap runs, in order to estimate the variability that is present if a different subset of neurons had been selected from a similar population (this procedure creates a slightly negatively biased estimate of the actual decoding accuracy of the larger population, so we use a procedure that samples without replacement for estimating true decoding accuracy, and only sample with replacement when estimating the standard error).

In order to make a fair comparison between the decoding accuracies on the isolated object trials and the three object trials, a decoding accuracy measure based on the area under the ROC curve (AUROC) was used. This measure was calculated separately for each class *i* by: 1) computing a vector v_i that was the average of all the training points from class *i*, 2) calculating the correlation coefficient between v_i and all test points, 3) repeating this for all cross-validation splits to get a large collection of correlation coefficients, 4) calculating the ROC curve from the correlation coefficient values from test points that were in class *i* and the values of test points that were not from class *i*, and 5) averaging the results over all classes.

For decoding the isolated object decoding (blue trace in figure 5.1b), a twelve-fold cross validation procedure was used in which the classifier was trained on 11 pseudo-population examples of each stimulus and tested on 1 example of each stimulus on each cross-validation run. For the cluttered trial decoding (figure 1b red and green traces), the classifier was trained on 11 pseudo-population examples of each object on the isolated object trials, and testing on the 635 three object trials pseudo-population responses. For

figure 1c (Supplemental figure 5.5c), the classifier was trained using 2 (18) examples of each stimulus at all three locations. Statistical significance was assessed by running the full decoding procedure 200 times with randomly shuffled labels and finding all time bins in which less than 1% of the randomly shuffled runs were higher than the real decoding accuracy. For figure 5.2, the classifier was either trained on *n* training examples of the attended object (displayed with two other random objects) and was tested on all the remaining cluttered trial data (red trace), or the classifier was trained on *n* examples of isolated objects, and then was tested on all the cluttered trials (blue trace), where *n* is the value given on the x-axis of that figure. For figure 5.3, the classifier was again trained on 11 examples from isolated object trials and tested on all clutter trials in which one of the distractors changed color. The decoding accuracy for figures 5.1b and 5.3 are based on training the classifier on isolated object data using 500ms period that started 85ms after stimulus onset. The training and test data for figure 5.2, and 5.1c, are based on firing rates in a 300ms bin that started 200ms after the onset of the attentional cue. In general the results were very robust to the parameters used in the decoding procedure. Figure 5.1d was created by using isolated trials to find the position that elicited the highest and lowest firing rate for each neuron, and then assessing the best to worst stimulus again using 500ms of data from the array period. These tuning curves were then plotted for the attended object using data from cluttered trials when attention was directed to best or worst position.

References

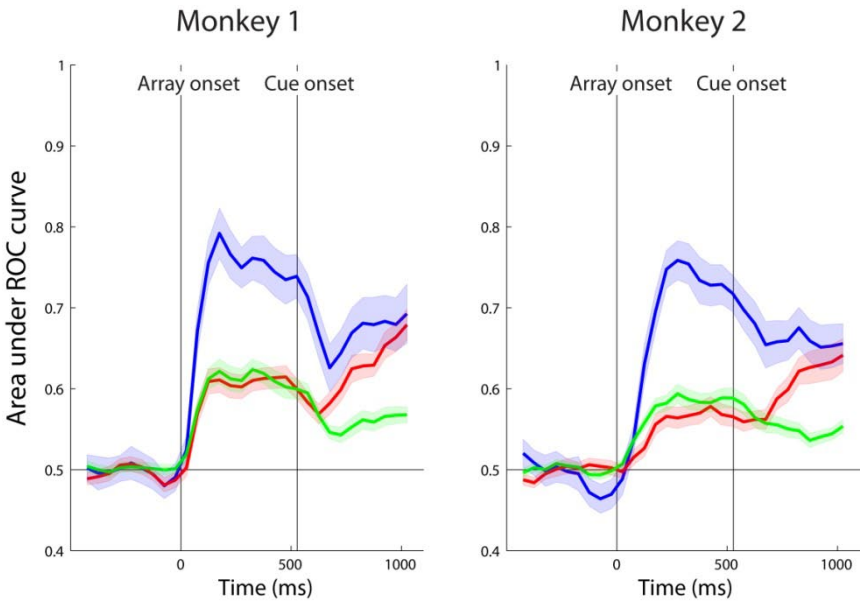
1. McAdams, C.J. & Maunsell, J.H. Effects of attention on orientation-tuning functions of single neurons in macaque cortical area V4. *J. Neurosci* **19**, 431-441 (1999).
2. Cohen, M.R. & Maunsell, J.H.R. Attention improves performance primarily by reducing interneuronal correlations. *Nat. Neurosci* **12**, 1594-1600 (2009).
3. Fries, P., Reynolds, J.H., Rorie, A.E. & Desimone, R. Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* **291**, 1560-1563 (2001).
4. Moran, J. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782-784 (1985).
5. Motter, B.C. 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 (1993).
6. Hayden, B.Y. & Gallant, J.L. Combined effects of spatial and feature-based attention on responses of V4 neurons. *Vision Res* **49**, 1182-1187 (2009).
7. Hung, C., Kreiman, G., Poggio, T. & DiCarlo, J. Fast readout of object identity from macaque inferior temporal cortex. *Science* **310**, 863-866 (2005).
8. Meyers, E.M., Freedman, D.J., Kreiman, G., Miller, E.K. & Poggio, T. Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J Neurophysiol* **100**, 1407-19 (2008).
9. Reynolds, J.H., Chelazzi, L. & Desimone, R. Competitive Mechanisms Subserve Attention in Macaque Areas V2 and V4. *J. Neurosci.* **19**, 1736-1753 (1999).
10. Mitchell, J.F., Sundberg, K.A. & Reynolds, J.H. Spatial Attention Decorrelates Intrinsic Activity Fluctuations in Macaque Area V4. *Neuron* **63**, 879-888 (2009).
11. Mitchell, J., Sundberg, K. & Reynolds, J. Differential Attention-Dependent Response Modulation across Cell Classes in Macaque Visual Area V4. *Neuron* **55**, 131-141 (2007).
12. Cohen, M.R. & Maunsell, J.H.R. Attention improves performance primarily by reducing interneuronal correlations. *Nat Neurosci* **12**, 1594-1600 (2009).
13. Chelazzi, L., Duncan, J., Miller, E.K. & Desimone, R. Responses of Neurons in Inferior Temporal Cortex During Memory-Guided Visual Search. *J Neurophysiol* **80**, 2918-2940 (1998).
14. Reddy, L., Kanwisher, N.G. & VanRullen, R. Attention and biased competition in multi-voxel object representations. *Proceedings of the National Academy of Sciences* **106**, 21447-21452 (2009).
15. Sheinberg, D.L. & Logothetis, N.K. Noticing Familiar Objects in Real World Scenes: The Role of Temporal Cortical Neurons in Natural Vision. *J. Neurosci.* **21**, 1340-1350 (2001).
16. Lee, J. & Maunsell, J.H.R. A Normalization Model of Attentional Modulation of Single Unit Responses. *PLoS ONE* **4**, e4651 (2009).
17. Desimone, R. & Duncan, J. Neural Mechanisms of Selective Visual Attention. *Annu. Rev. Neurosci.* **18**, 193-222 (1995).
18. Chikkerur, S.S., Serre, T., Tan, C. & Poggio, T. What and where: A Bayesian inference theory of attention. *Vision Res* (2010).doi:10.1016/j.visres.2010.05.013

19. Reynolds, J.H. & Heeger, D.J. The Normalization Model of Attention. *Neuron* **61**, 168-185 (2009).
20. Reynolds, J.H. & Desimone, R. The role of neural mechanisms of attention in solving the binding problem. *Neuron* **24**, 19-29, 111-125 (1999).
21. Lee, D.K., Itti, L., Koch, C. & Braun, J. Attention activates winner-take-all competition among visual filters. *Nat. Neurosci* **2**, 375-381 (1999).
22. Hamker, F.H. The Reentry Hypothesis: The Putative Interaction of the Frontal Eye Field, Ventrolateral Prefrontal Cortex, and Areas V4, IT for Attention and Eye Movement. *Cerebral Cortex* **15**, 431 -447 (2005).
23. Ardid, S., Wang, X. & Compte, A. An Integrated Microcircuit Model of Attentional Processing in the Neocortex. *J. Neurosci.* **27**, 8486-8495 (2007).
24. Börgers, C., Epstein, S. & Kopell, N.J. Gamma oscillations mediate stimulus competition and attentional selection in a cortical network model. *Proceedings of the National Academy of Sciences* **105**, 18023 -18028 (2008).
25. Tiesinga, P.H., Fellous, J., Salinas, E., José, J.V. & Sejnowski, T.J. Inhibitory synchrony as a mechanism for attentional gain modulation. *J. Physiol. Paris* **98**, 296-314 (2004).
26. Tsotsos, J.K. How Does Human Vision beat the Computational Complexity of Visual Perception? *Computational Processes in Human Vision: An Interdisciplinary Perspective* 286 - 338 (1988).
27. Gattass, R., Sousa, A. & Gross, C. Visuotopic organization and extent of V3 and V4 of the macaque. *J. Neurosci.* **8**, 1831-1845 (1988).

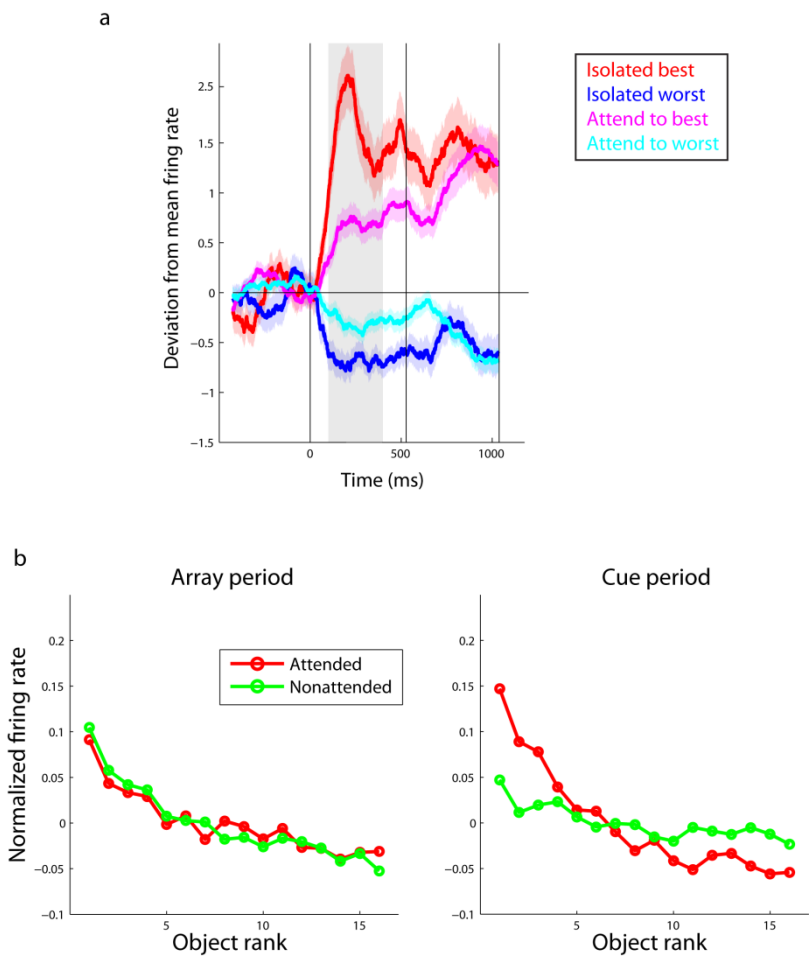
Acknowledgments This research was sponsored by grants from DARPA (IPTO and DSO), National Science Foundation (NSF-0640097, NSF-0827427), and NEI (R01EY017292). Additional support was provided by: Adobe, Honda Research Institute USA, King Abdullah University Science and Technology grant to B. DeVore, NEC, Sony and especially by the Eugene McDermott Foundation. EM was supported by an NDSEG graduate research fellowship and by the Herbert Schoemaker fellowship. Contributions to this work were made by YZ and NB who conducted the neurophysiological recordings, EM who performed the decoding analyses, and EM, RD, TP who were primarily responsible for writing the manuscript. Thomas Serre was involved in some of the initial discussion about the experimental design.

Supplemental figures

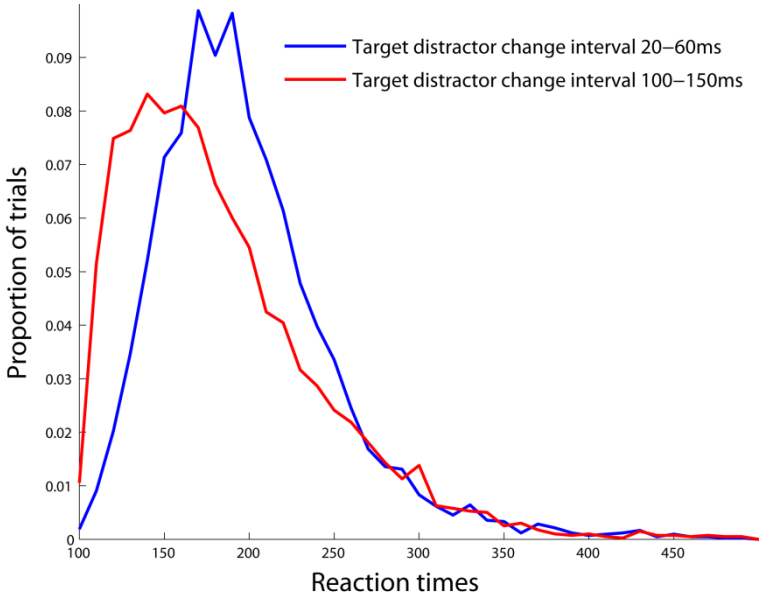
Supplemental figure 5.1 Stimulus sets. The stimuli came from four categories (face, couch, car or fruit). Two-thirds of the multiple object trials consisted of all three images from the same category and one-third of the trials consisted of images from different categories. For all analyses reported in this paper, all images were treated the same (i.e., the category of the stimuli was ignored).



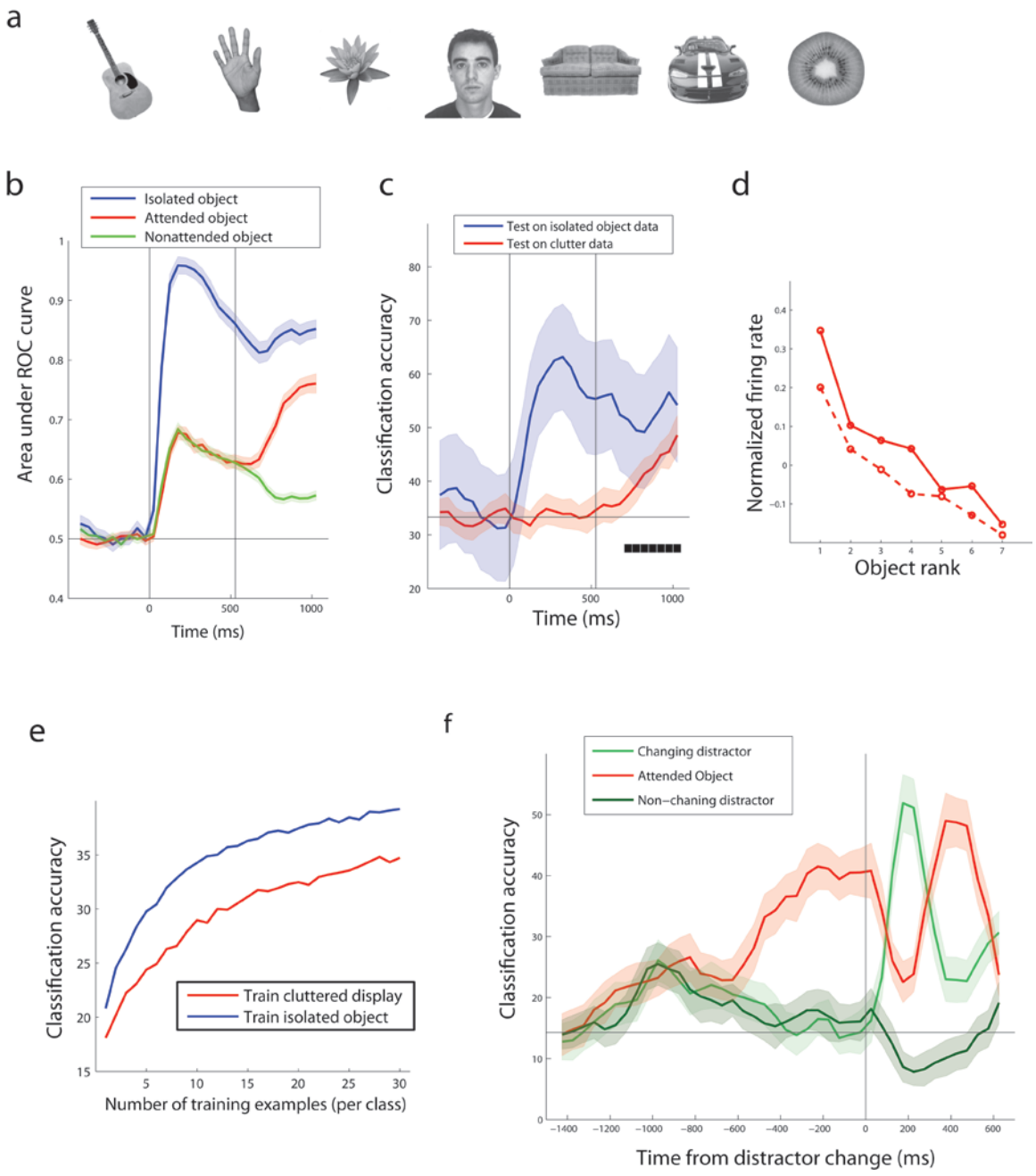
Supplemental figure 5.2. Figures showing decoding results for both monkeys separately. Both monkeys show a similar pattern of reduced decoding accuracy when multiple objects are presented (red and green traces) compared to when only a single object is presented (blue trace) prior to the onset of the attentional cue. After the onset of the attentional cue (vertical line at ~500ms), information about the attended object increased (red trace), while information about the attended object decreased (green trace) for both monkeys. Because the results were similar for both monkeys, we combined data from both monkeys for all other analyses. It should be noted that the impact of clutter in our study is larger than has been reported in a recent paper by Li et al. (2009), which is most likely due to the fact that Li et al., used the same images when training and testing the classifier, which allows the classifier to also exploit visual features related to the configurations of the objects (rather than learning a representation that was completely invariant to the surrounding clutter). Our findings are more in line with Agram et al. (2010), which also showed decreased performance when multiple objects were present.



Supplemental figure 5.3. Effects of attention on firing rates averaged across the population of cells. **a**, Firing rates (with mean subtracted) to the 'best stimulus' (the stimulus that elicited the highest firing rate) and the 'worst stimulus' (the stimulus that elicited the lowest firing rate) on isolated object trials (red and blue traces), and on three object trials (magenta and cyan traces). Before the attentional cue was presented on three object trials, the best stimulus had a higher firing rate than the worst stimulus, and this difference increased after attention had been deployed, with the best stimulus and worst stimulus trials matching the firing rates seen on isolated object trials. The best and worst stimuli were found using data on isolated object trials using a time period from 100-400 ms after stimulus onset (gray shaded region). To correct for selection biases on the isolated object results, we randomly shuffled the labels of the stimuli, found the 'best' and 'worst' stimuli on these shuffled data and subtracted these randomly shuffled best to worst firing rates from the best and worst firing rates obtained from the real stimuli. The results were averaged over all neurons and the colored shaded regions are one standard error of the mean. **b**, Z-score normalized firing rates for the stimulus that was cued on the three object trials, sorted from the best to worst stimulus as determined on isolated object trials. Before attention was deployed, the ranking seen on isolated object trials were preserved (left panel) and that after attention was deployed this ranking was accentuated for the attended stimulus and largely abolished for the non-attended stimulus.



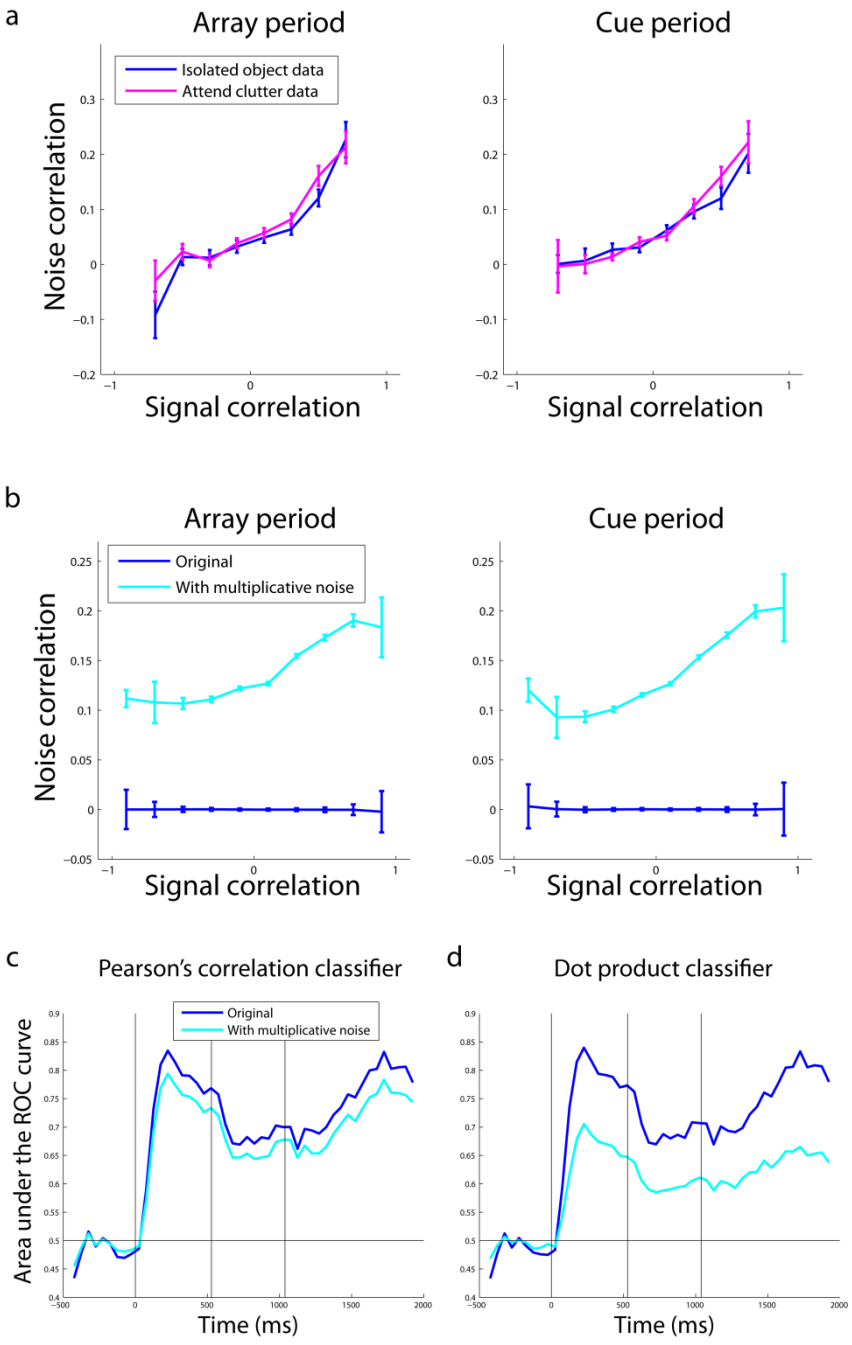
Supplemental figure 5.4. Reaction times were slower when the target changed color soon after the distractor changed color. The distribution of reaction times on trials when the time difference between the target and distractor color change was 20-60ms (blue trace) was compared to when the time difference was 100-150ms (red trace). As can be seen, the distributions were shifted to longer reaction times when the time between the target and distractor change was short. This increase in reaction time could be related to the fact that changes in distractor saliency caused information in IT to be dominated by properties related to the distractor immediately following the distractor color change.



Supplemental figure 5.5. Replication of the results on a second stimulus set. a, We replicated the experiment on the second monkey using a new stimulus set that consisted of 7 unique objects **b-d**, The decoding results were similar to those with 16 stimuli, except the decoding accuracy for the attended object did not reach the accuracy seen for the isolated object. **e**, Replication of the results figure 2 using the 7 stimulus set. **f**, Replication of the results in figure 3 using the 7 stimulus set. Chance performance is 1/7, or 14.29%.

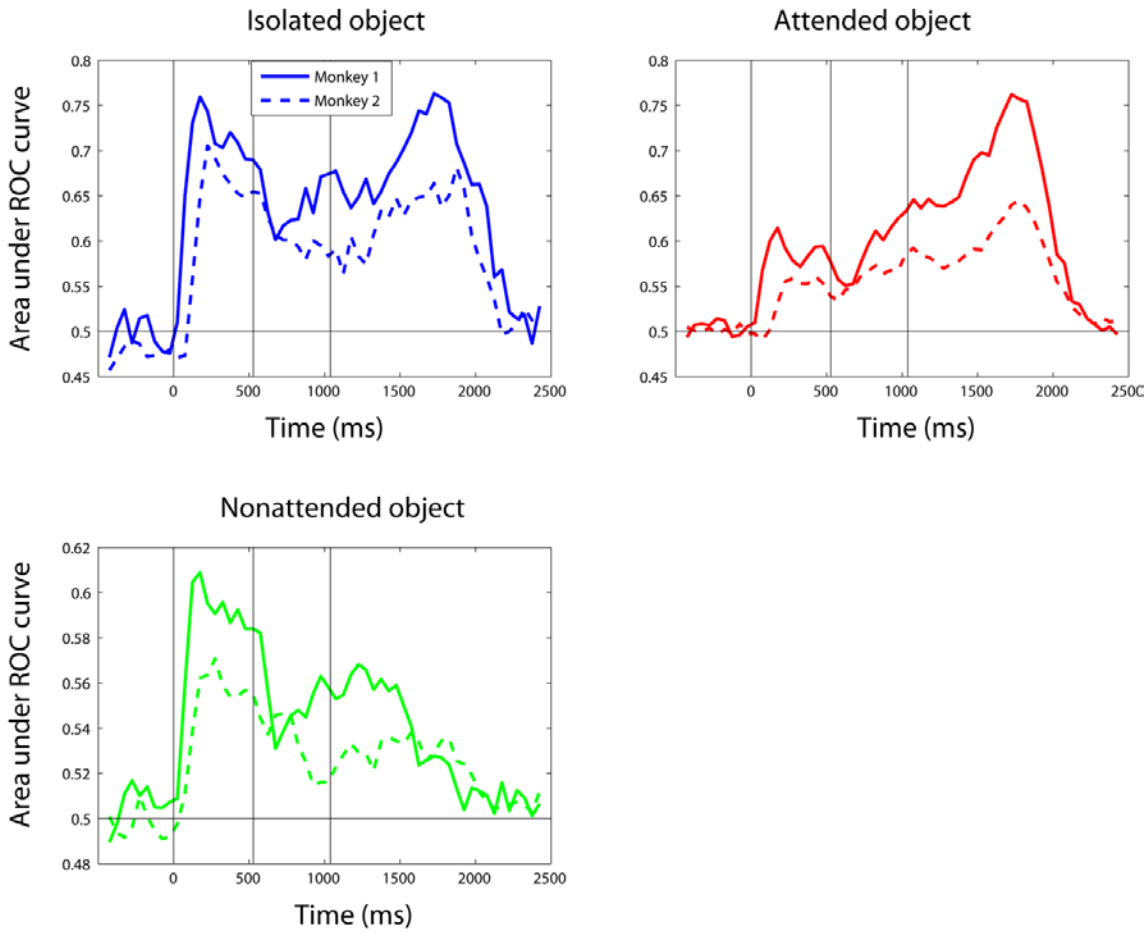
Additional Supplemental Material

Additional results related to the attention and IT

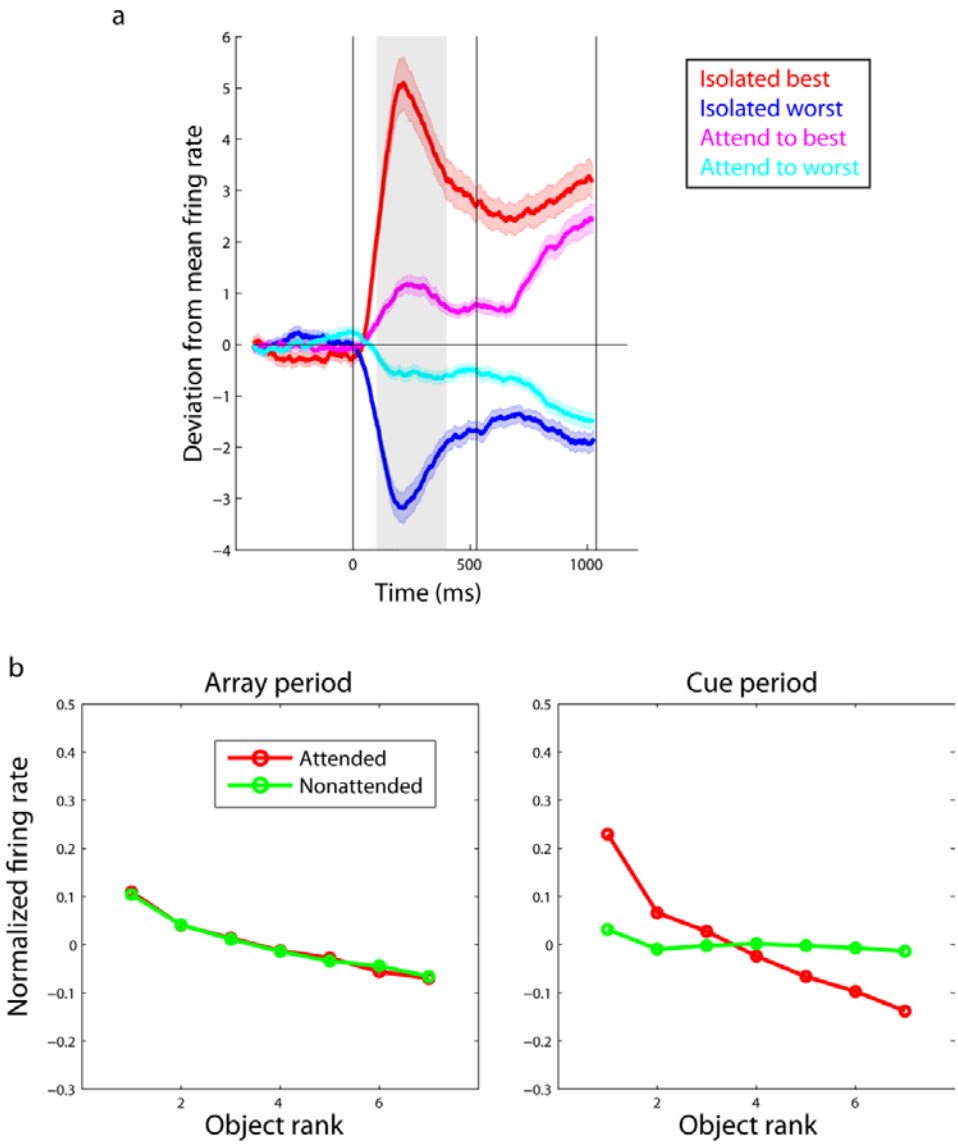


Additional supplemental material 5.1 Adding correlation noise only slightly impact on population decoding. a, Noise correlations plotted as a function of signal correlation for the isolated object trials (blue trace) and for the attended object in the three object trials (magenta

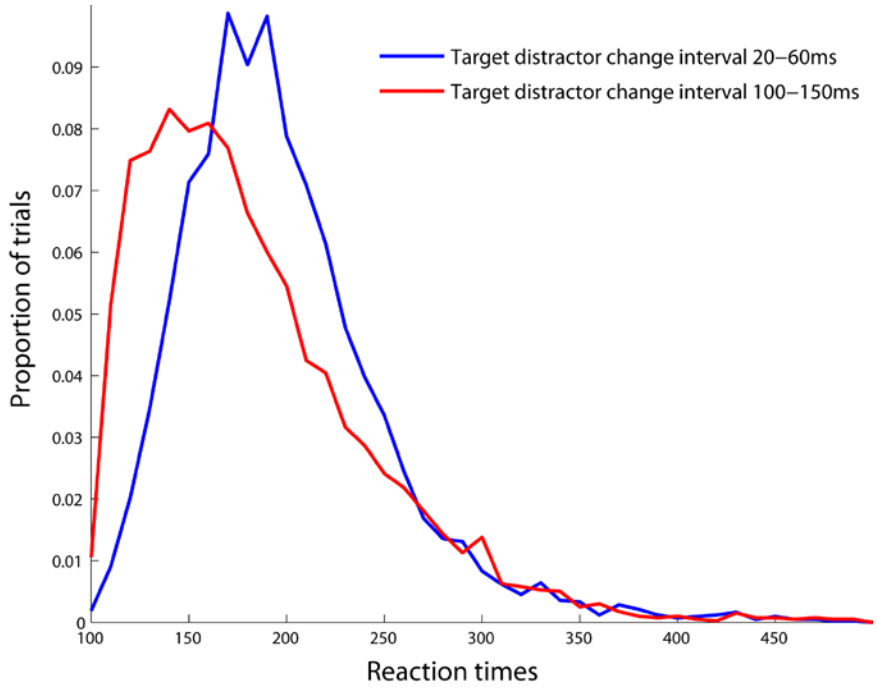
trace), for the array period (left) and cue period (right). As can be seen, noise correlations are similar in the array and the cue periods for both the isolated object and the attended object, thus we do not see any significant changes of noise correlation with attention. We use common methods for calculating these functions to be consistent with the rest of the literature[28], however, it should be noted the functions of increasing noise correlations with signal correlations could be a function of misestimating the signal correlations for neurons that have strong noise correlations, and also that the errorbars are likely underestimated due to the fact that we do not correct for the fact that each neuron contributes to many noise correlation values. b, Results showing that our pseduo-populations (blue traces) do not have any noise correlations, however we can add noise correlations to our data (cyan traces) that resemble those seen in the real data by multiplying the pseudo-population vectors by a constant from 1-10 chosen from a uniform distribution. c, decoding results for a Pearson's correlation coefficient classifier (which is the classifier used in the paper), before noise correlations have been added (blue trace), and after noise correlations have been added (cyan trace). As can be seen, adding correlated to the noise had little effect on the classification accuracy from this classifier. d, the same results as in c but with a dot product classifier. Here the results are more effected by noise, although still a high level of accuracy can be achieved. It should also be noted that similar noise correlations can also be induced by adding multivariate Gaussian noise using the signal correlations as the covariance matrix, however for this type of noise both classifiers maintained their high classification performance. Based on the difference in the results between c and d, it is interesting to speculate that perhaps neural circuits operate in a way that is analogous to a Pearson's correlation classifier, with the normalization operation that is commonly seen with the onset of stimuli1 and used to describe the effects of attention[29,30] being analogous to the normalization operation in the denominator when calculating Pearson's correlation coefficient. In such a scenario, we could also speculate that when no stimulus is being represented by a given brain region (i.e., when the brain region is in an ideal state), that neurons in that brain region can fluctuate in a correlated way, and that when something is being represented, the activity in a brain region is normalized to be in a consistent range (perhaps via divisive inhibition), to that a downstream area can readout the information. This interpretation thus gives one possible explanation for the results seen in Cohen and Maunsell (2010).



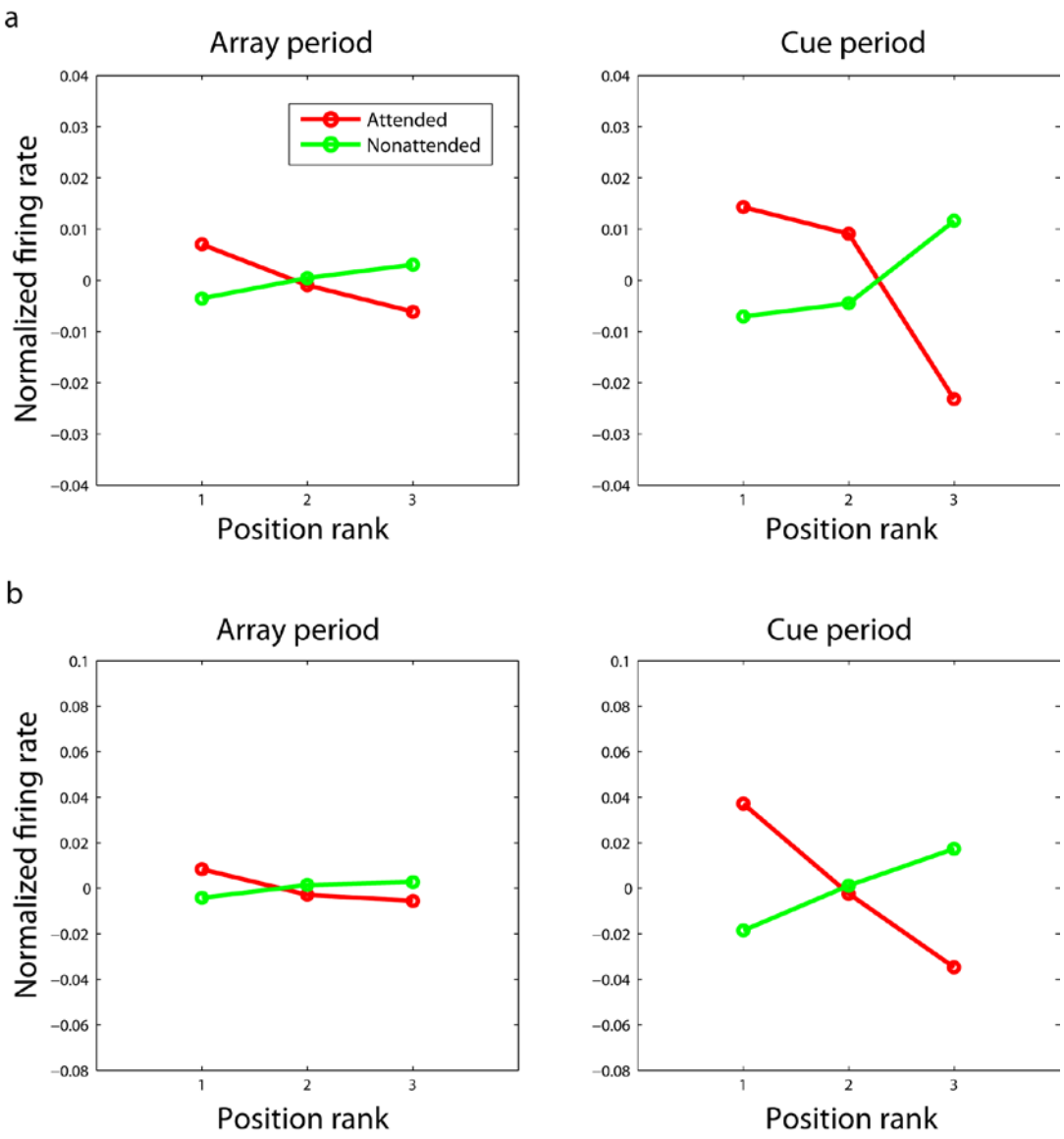
Additional supplemental material 5.2 Results comparing the decoding accuracy for the two monkeys on the same plot. These results are similar to the results plotted in supplemental figure 2, except here we put the results from the two monkeys on the same figure, using the same number of bootstrap neurons (65 neurons on each iteration). As can be seen, the results from monkey 1 are slightly higher than the results from monkey two, but in general they show the same trends.



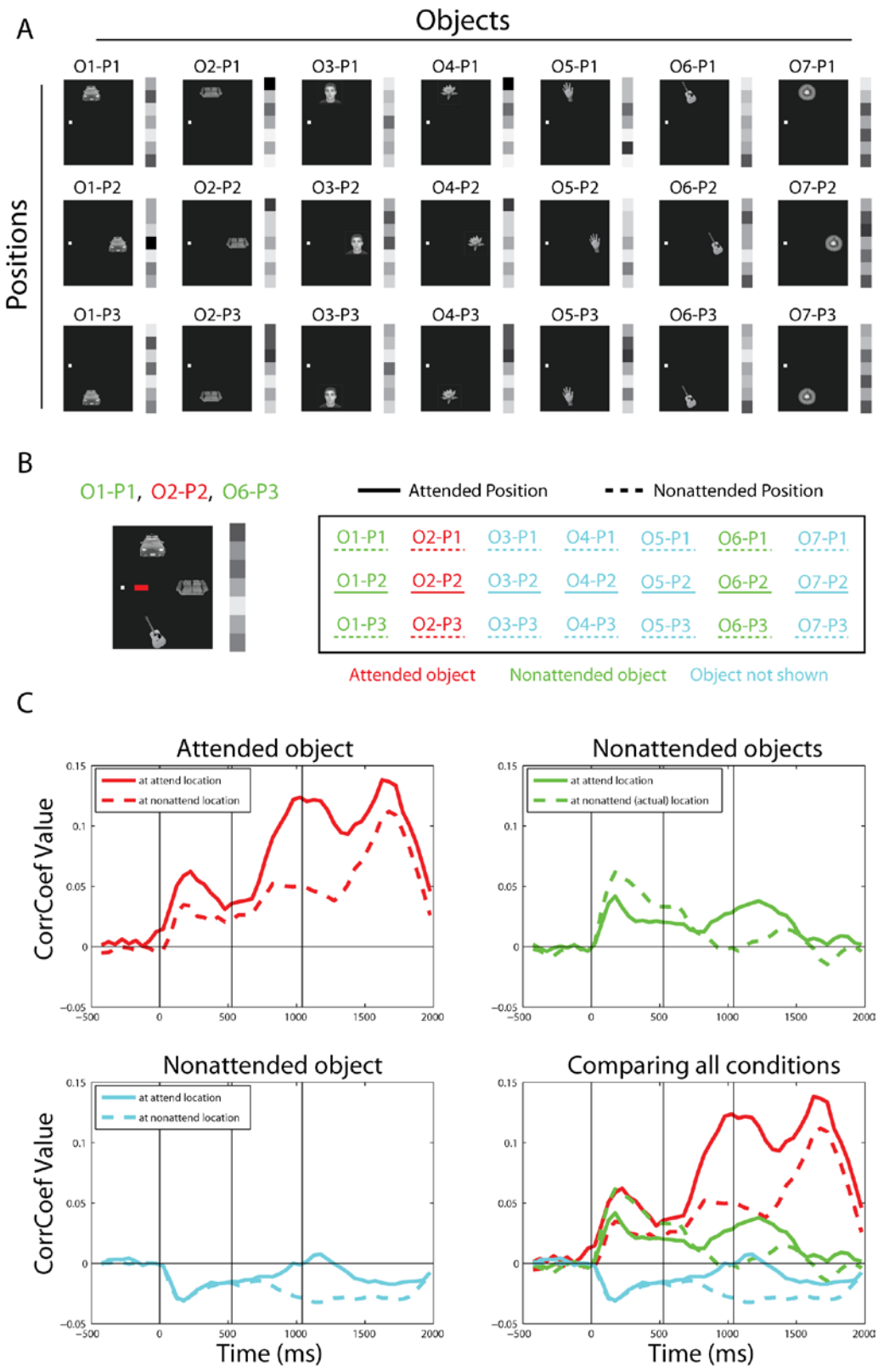
Additional supplemental material 5.3 Effects of attention on firing rates averaged across the population of cells using the 7 unique object stimulus set. This figure is the same as supplemental figure 3, except the data comes from the second monkey using the 7 unique object stimulus set (see supplemental figure 3 for more information about how these figures were made).



Additional supplemental material 5.4 Results from the 7 unique object stimulus set, showing that reaction times were again slower when the target changed soon after the distractor changed color. This figure is the same as supplemental figure 4, except the data comes from the second monkey using the 7 unique object stimulus set (see supplemental figure 4 for more information about how these figures were made).

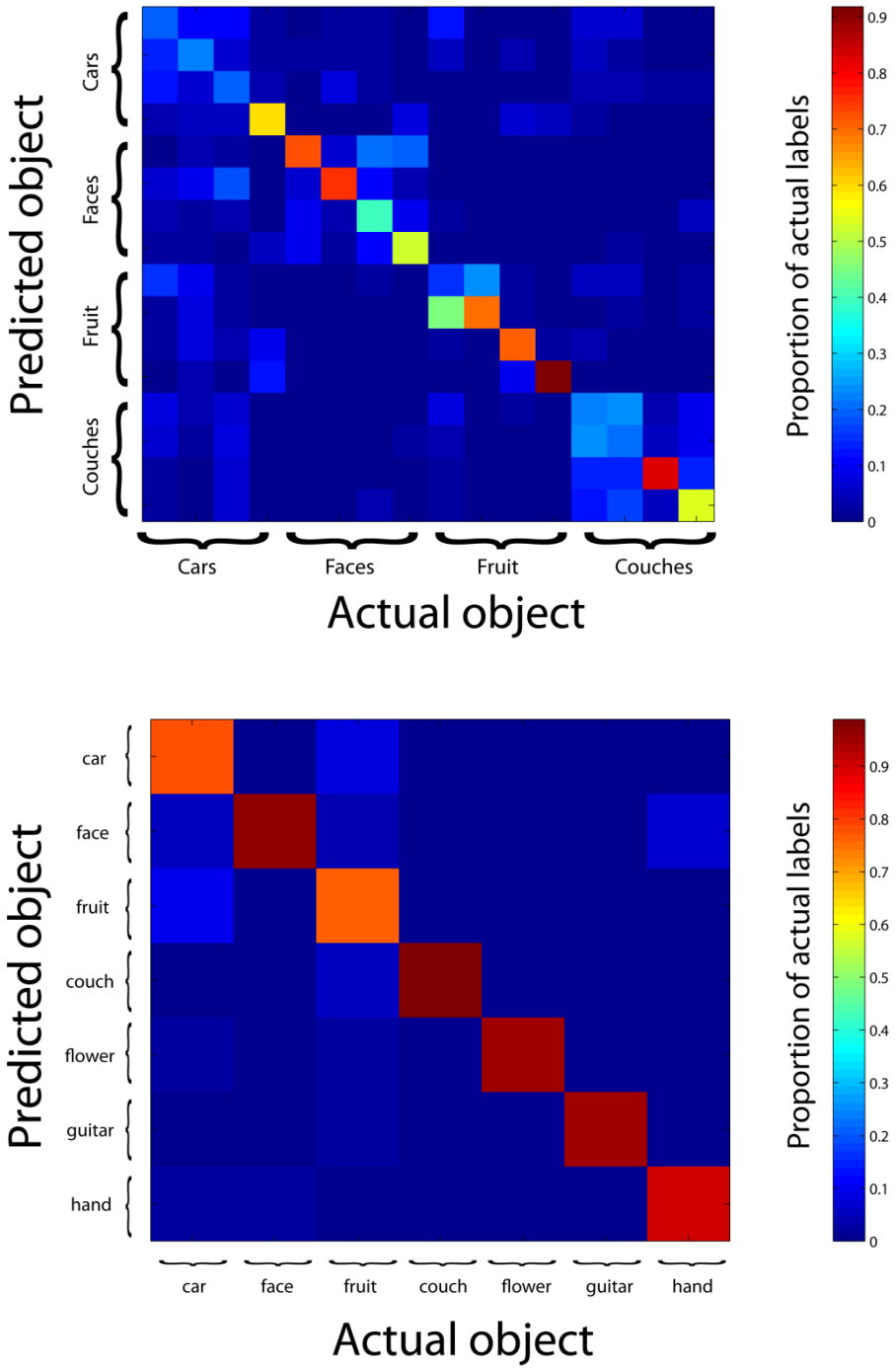


Additional supplemental material 5.5 Population averaged z-score normalized firing rate to each position, ranked according to each neuron's the isolated object position preference for both monkey on 16 object stimulus set (a) and for monkey 2 on the 7 object stimulus set (b). Before attention is deployed this is no difference between the attended and nonattended isolated object position preference, however after attention has been deployed, firing rates increase when the monkey is attending to neurons' 'best' positions and decrease when the monkey is attending to neurons's worst positions.

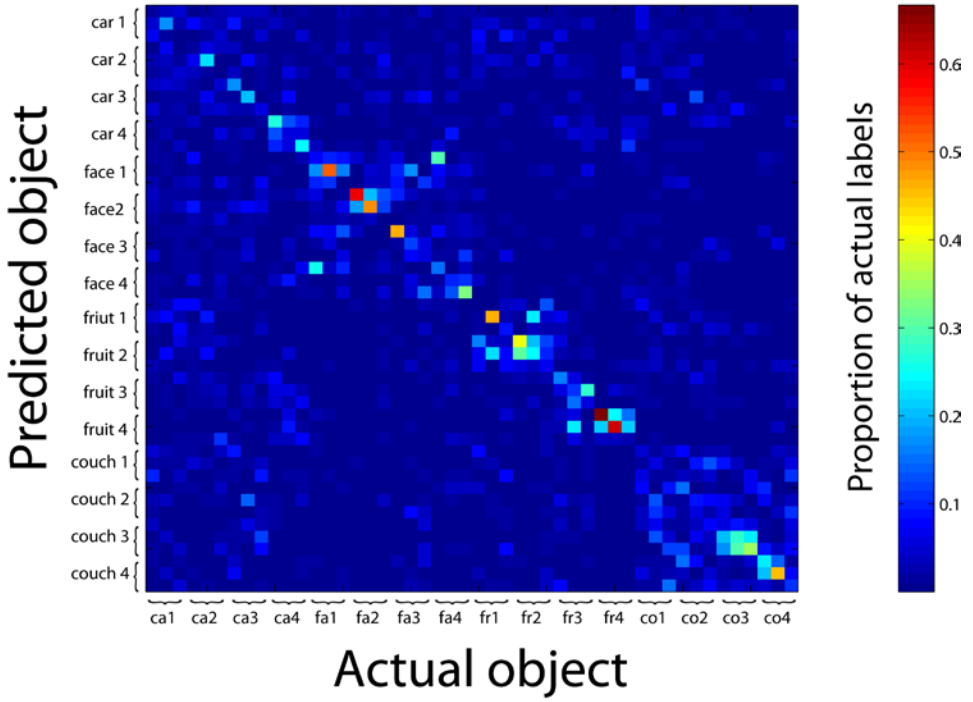


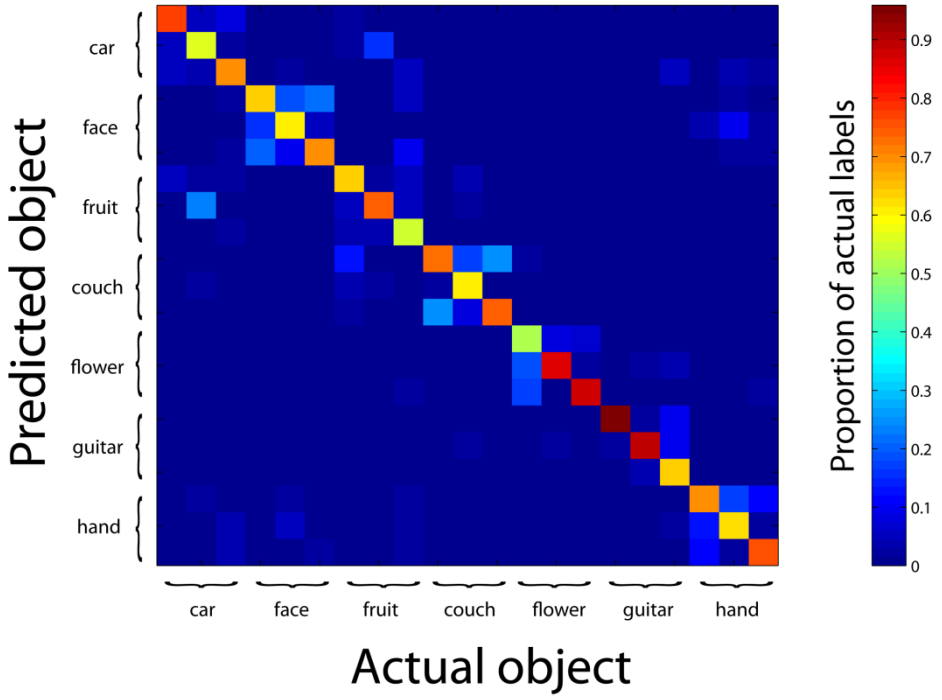
Additional supplemental material 5.6 Results showing both position and identity enhancement with attention using the 7 object stimulus set. A, illustration showing that each

object at each position generates a particular pattern of neural activity, which we call the 'population representation' of a particular object at a particular location. For this analysis we create the population representations of each object at each position using data from the isolated object trials, and then we correlate these population representations with data from cluttered object trials. B, Illustration showing that the cue creates both an 'attended object' and an 'attended location'. In this example, attention is directed to object 2 at position 2. If attention affects object and position information independently, then we might expect the population activity on this cluttered trial to be similar to the population representations that share the same identity as the attended object, and to the population representations that share the same position as the attended object. Thus, for this example, we would expect the population activity to be similar to the population representations that contain object 2 (red representations in the second column on the right of the figure), and to be similar to the population representations that contain an object at position 2 (solid underlined representations in row two on the right of the figure). C, The plot on the upper left shows the correlation coefficient between the cluttered data and population representation of the attended object at the attended location, and also the correlation coefficient between the cluttered data and the population representation of the attended object at the nonattended locations. As can be seen, there is more similarity between cluttered data and the population representation of the attended object at the exact attended location, then there is with the population representations of the attended object at nonattended locations. The plot on the upper right shows the correlation coefficient between the cluttered data and population representations of the nonattended object at the attended location, and also the correlation coefficient between the cluttered data and the population representations of the nonattended object at the actual location these nonattended objects were shown. As can be seen, prior to the onset of the attentional cue, there is a higher correlation between the cluttered data and the population representation of nonattended objects at the actual locations they were shown, than the population representations of the nonattended objects at the location that will be attended. However, after the onset of the attentional cue, the cluttered data more closely resembles the population representations of the nonattended objects at the attended locations, than it does the nonattended objects at the actual locations they are on the screen. The plot on the lower left shows the correlation coefficient between the cluttered data and population representation of the objects that were not shown at the attended location, and also the correlation coefficient between the cluttered data and the population representations of the objects that were not shown at the nonattended locations. As can be seen, after the onset of the attentional cue, the cluttered data more closely resembles the population representations of objects that were not shown at the attended location compared to the population representations of objects that were not shown at the nonattended locations. The plot on the lower right compares all the correlation coefficients shown in the other plots together. It can be clearly seen that after the onset of the attention cue, the cluttered data is more similar to the population representations for the attended object than for the nonattended objects, and that the cluttered data is more similar to the population representations for the nonattended objects than to the population representations for the objects that were not shown (red > green > cyan). Likewise, after the onset of the attention cue, the cluttered data is more similar to the population representations for objects at the attended location (solid lines) compared to the population representations of objects that are at the nonattended location (dashed lines). Thus attention enhances the population representations for both object identity and object position. These figures were created by running trying to decode each object at its exact location using a MCC classifier (training on isolated object data and testing on cluttered data), and then plotting the correlation values separately for the different attention conditions.

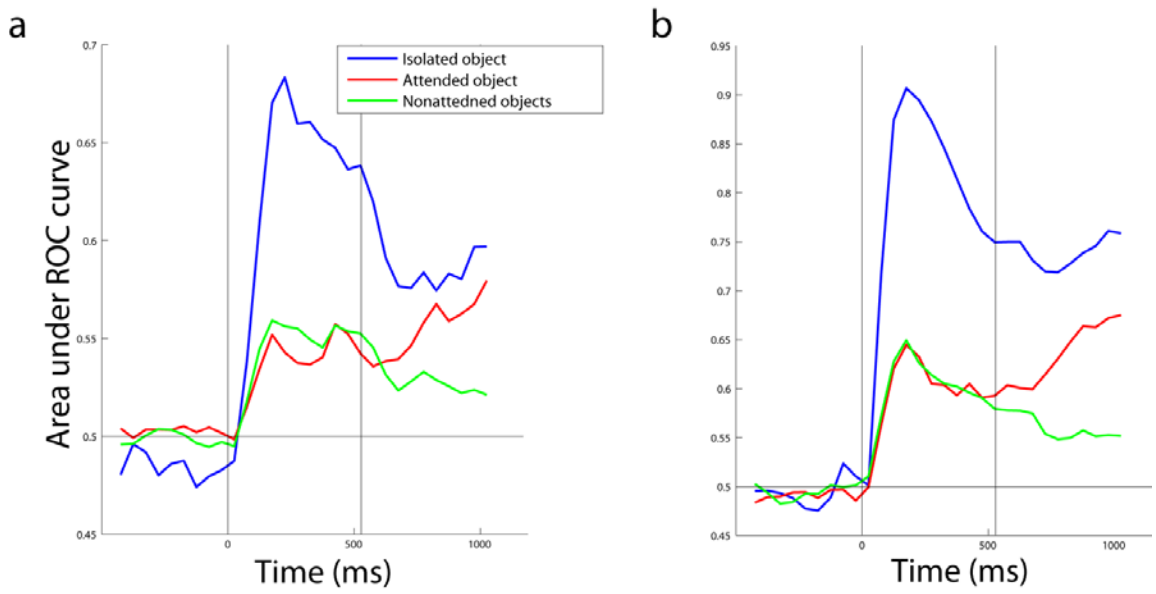


Additional supplemental material 5.7 Classification accuracy confusion matrices. Top plot shows the confusion matrix the 16 object stimulus set (decoding based on combining data from both monkeys), and the bottom plot shows the confusion matrix the 7 object stimulus set (decoding based on data from monkey 2). In general, there is not a clear pattern of mistakes made by the classifier, even on the 16 object stimulus set where the images come for 4 categories - although there are a few trends in this dataset, such as the fact that the image of an apple was often mistaken as an image of an orange, there is more confusion among the couches, and the cars generally had lower decoding accuracies than the other categories of objects.



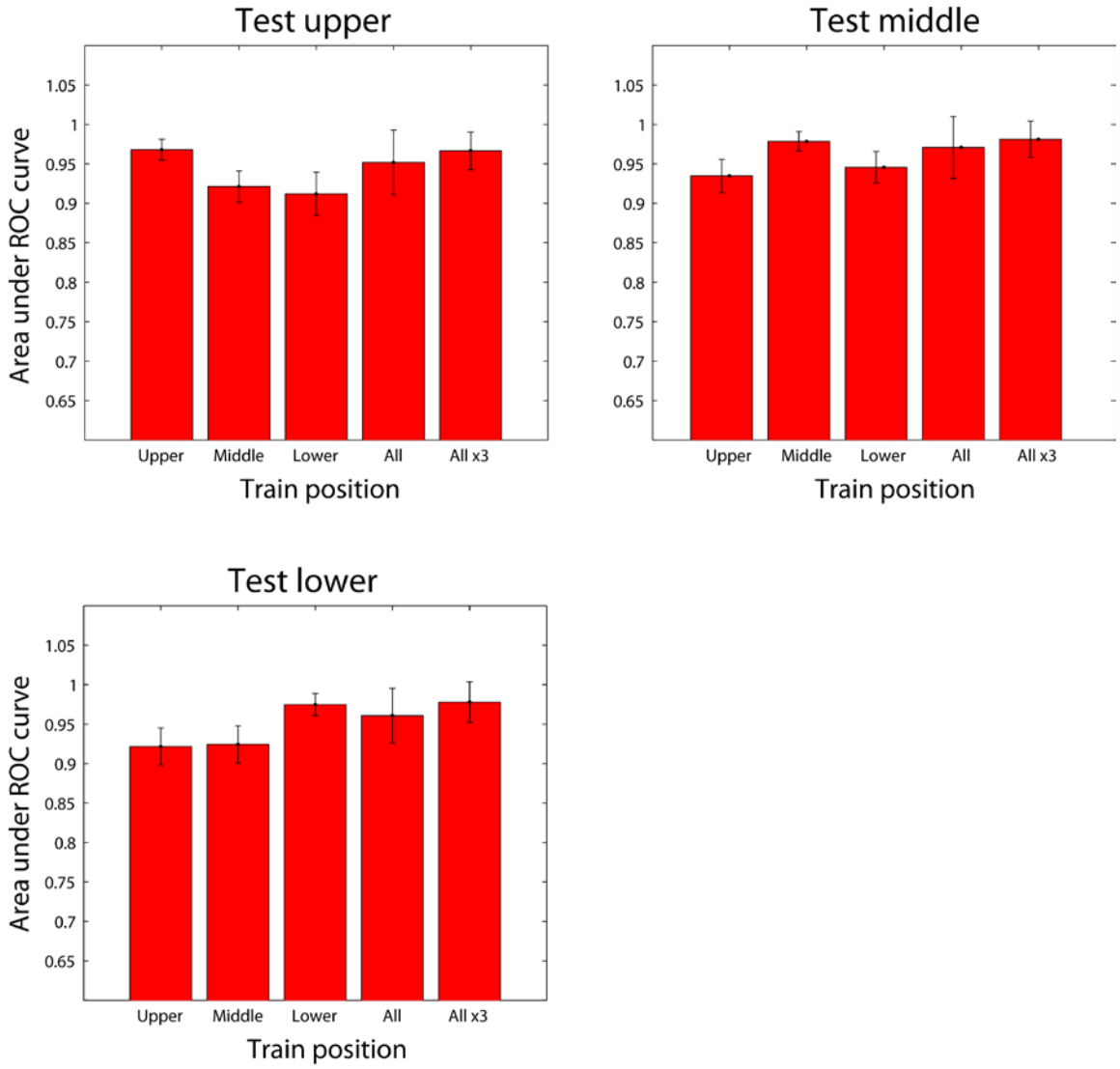


Additional supplemental material 5.8 Classification accuracy confusion matrices for decoding which object was shown at which exact location. Upper plot shows the confusion matrix for the 16 object stimulus set, and the bottom plot shows the confusion matrix the 7 object stimulus set. While some mistakes are made between the same object at different locations, in general there are not too many obvious trends in the pattern of mistakes, and objects at particular locations can be distinguished from each other surprisingly well.

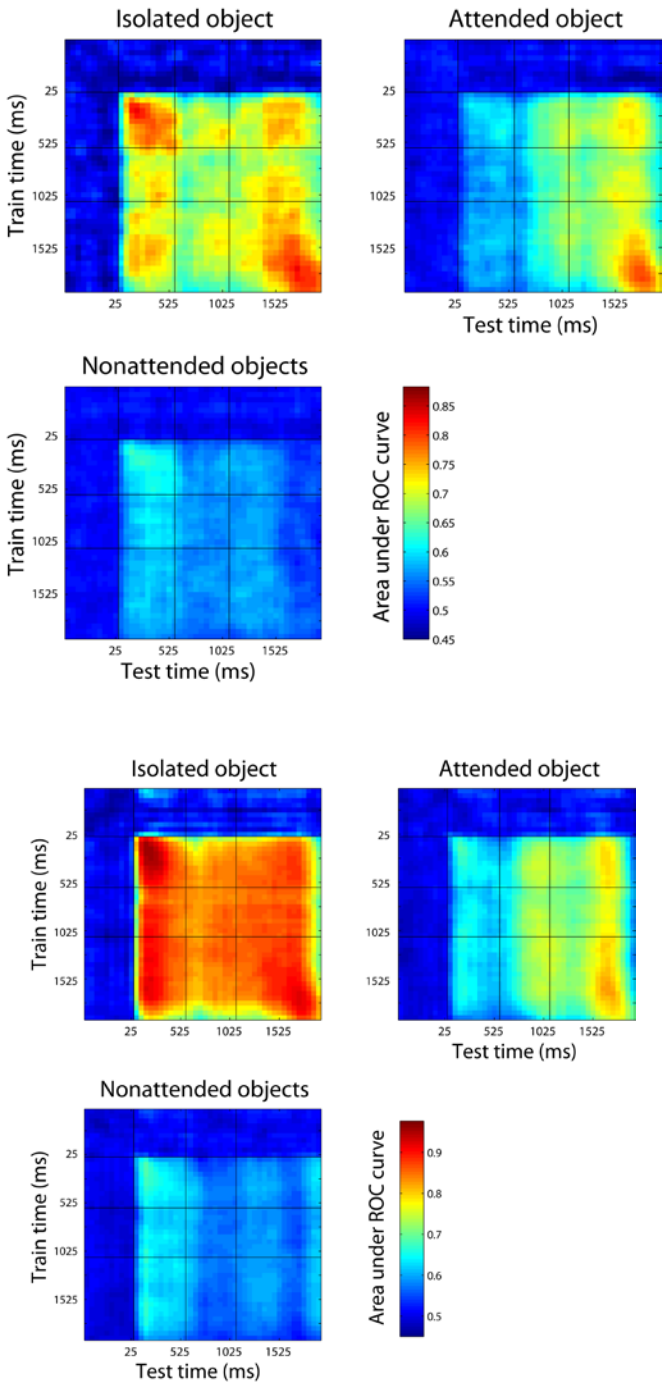


Additional supplemental material 5.9 Results showing that the main attention and isolated object decoding accuracies generalize across position. For this analysis we trained the classifier using data from one location and then tested the classifier with data from when either the isolated object was shown at a different location (blue trace), or when the attended or nonattended object was shown at a different location (red and green traces). The results are averaged over training and testing at all locations. a, results from the 16 object stimulus set (data combined from both monkeys), and b, are the results from the 7 object stimulus set (data from monkey 2).

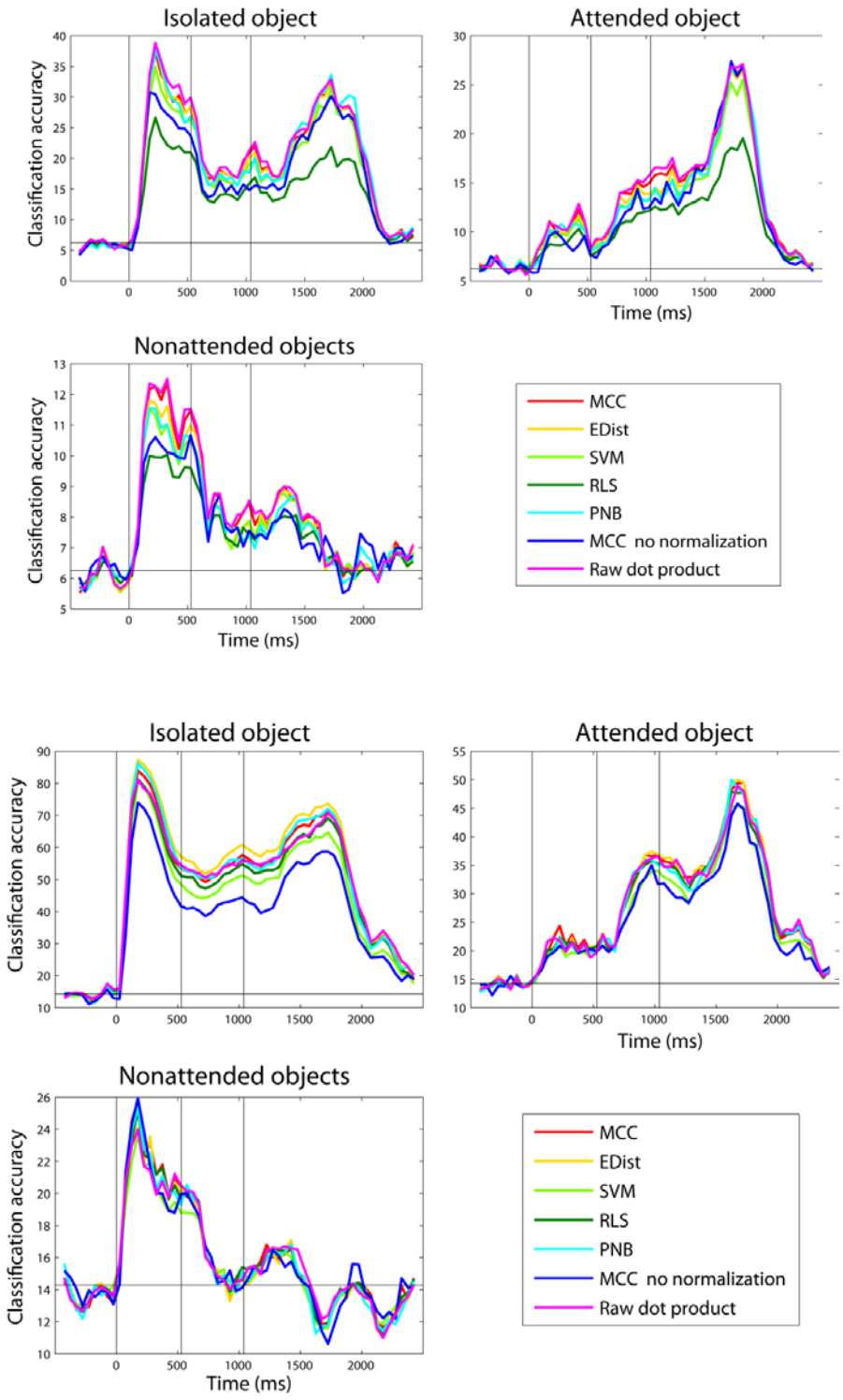
Additional results on IT and showing the robustness of decoding



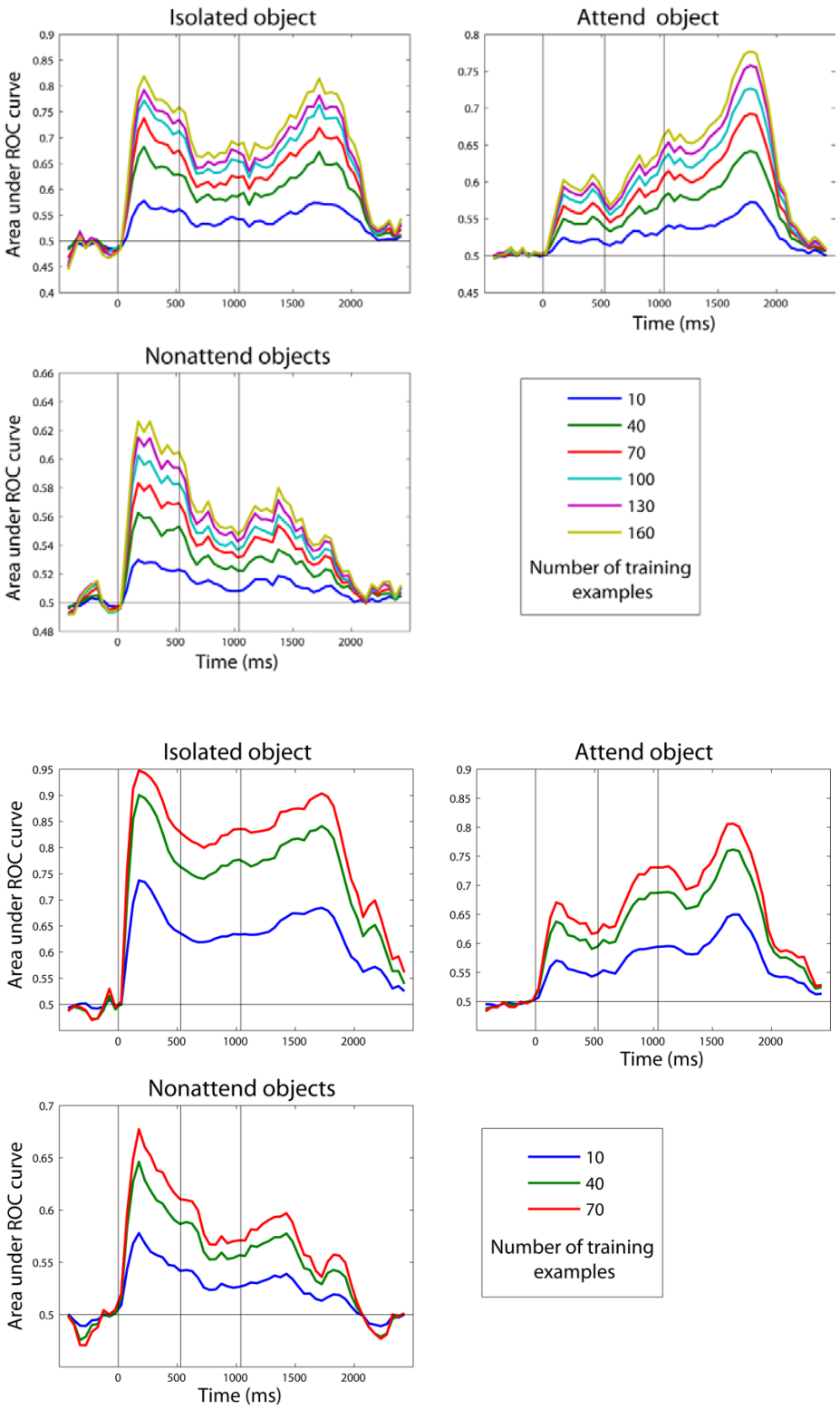
Additional supplemental material 5.10 Results showing that the population response to the isolated objects is similar at the different locations. For this analysis we trained the classifier with data from one location (indicated by the label on the x-axis), and we tested the classifier with data from either the same or a different locations. While in general training and testing at the same location led to slightly better results, the difference was small (as can be seen by comparing the left most three bars on each plot), and even this small difference disappeared when training with data from all locations (right two bars), showing that it is easy to learn a position invariant representation from data in IT. The results in this plot all comes from 7 object stimulus set (from monkey 2).



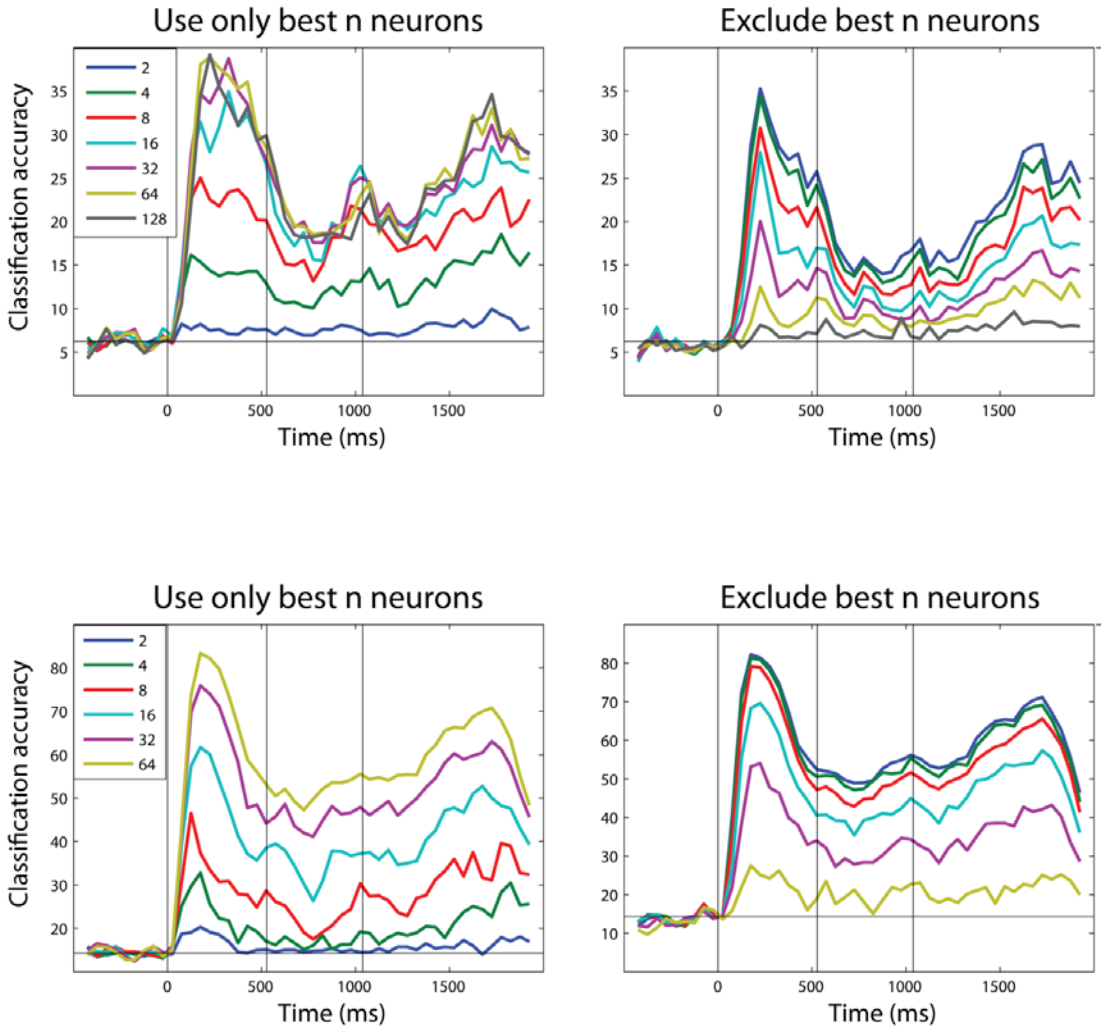
Additional supplemental material 5.11 Results from training the classifier at one time (indicated in the y-axis), and testing the classifier at the same or a different time (indicated by the x-axis). This plot is similar to Figure 3.6. As can be seen, the representation for the identity of the different objects is rather stationary, with one object being represented by the same pattern of activity over all time periods. This is quite different from the abstract category information which changed dynamically in time (see Figure 3.6), although is slightly more similar to the decoding of identity information from the Freedman et al., (2003) data (see Supplemental figure 3.7). These findings are similar to recent results reported by Crowe et al., (2010)[31]. Chapter 6 discusses these findings further.



Additional supplemental material 5.12 A comparison of classification accuracies using different classifiers for the 16 object stimulus set (upper plots), and the 7 object stimulus set (lower plots). As can be seen, all the results are qualitatively similar regardless of which classifier is used, and there are only slight differences in their absolute levels of performance.



Additional supplemental material 5.13 Decoding results plotted as a function of the number of bootstrap neurons used for the 16 object stimulus set (upper plots), and the 7 object stimulus set (lower plots).



Additional supplemental material 5.14 Isolated object decoding results using (left) or excluding) the most selective *k* neurons for 16 object stimulus set (upper plots), and the 7 object stimulus set (lower plots). The best k neurons were selected by applying an ANOVA to each neuron in the training set, and then using only these neurons to train and test the classifier. For the 16 object stimulus set, the results were about as good using only 16 of the most selective neurons compared to using a population of 128 neurons, while for the 7 object stimulus set, the results continued to improve slightly as more neurons were added (the difference between the results is probably due to the fact that the there were many more repetitions of each trial type in the 7 object stimulus set, so the classifier could learn the parameters better, and was thus better able to utilize information even in highly noisy neurons). Performance degraded slowly when the top k neurons were excluded.

Additional references

28. Smith, M.A. & Kohn, A. Spatial and Temporal Scales of Neuronal Correlation in Primary Visual Cortex. *J. Neurosci.* **28**, 12591-12603 (2008).
29. Lee, J. & Maunsell, J.H.R. A Normalization Model of Attentional Modulation of Single Unit Responses. *PLoS ONE* **4**, e4651 (2009).
30. Reynolds, J.H. & Heeger, D.J. The Normalization Model of Attention. *Neuron* **61**, 168-185 (2009).
31. Crowe, D.A., Averbeck, B.B. & Chafee, M.V. Rapid sequences of population activity patterns dynamically encode task-critical spatial information in parietal cortex. *J. Neurosci* **30**, 11640-11653 (2010).

Chapter 6: Conclusions

In this thesis we have developed and applied a set of decoding procedures in order to gain deeper insight into the content and coding of information that is present in high level visual areas. Before this research was started, it was unclear how useful these decoding methods would be because a careful examination of how sensitive these methods were to particular parameter choices had not been conducted. Consequently, I spent a fair amount of time making sure these methods were robust. The findings presented in this thesis demonstrate that across several different datasets, the decoding results are indeed robust to the choice of classifier (see Figure 2.3, Supplemental figure 3.2, Additional supplemental material 3.1, and Additional supplemental material 5.12), the data normalization method used (Additional supplemental material 3.2), and to the various different representations of the neural data (Additional supplemental material 3.10 and Additional supplemental material 3.11). Thus the empirical evaluation of these decoding methods give us a fair amount of confidence that discoveries made using these techniques are capturing underlying consistencies in the data, and are not particularly sensitive to choices of data analysis parameters used.

In addition to showing that these decoding methods are a reliable way to analyze data, we also discovered many new findings about the function of IT cortex and other related brain regions, which demonstrate the usefulness of these methods. In chapter 3, we show there is information about abstract categories in IT, and that this information is contained in a small subset of neurons that change in time. In chapter 4, we see that simple features (i.e., combinations of features that resemble the response properties of V1 neurons), might be sufficient to coarsely identify the presence of an animal in a cluttered scene, and that the latency of information in IT appears to be quite long relative to previously reported behavioral results (Kirchner and Thorpe, 2006; Girard et al., 2008), which gives tentative support to the idea that rapid coarse identification of animals in cluttered scenes might be primarily driven by other areas apart from IT. Finally, in chapter 5, we use these decoding methods to show that one of the main roles of attention is to restore

patterns of neural activity to a state that is similar to when the attended object is presented in isolation, which gives a coherent computational explanation for many of the attention related effects that have been described at the single neuron level. Thus, the work in this thesis makes several significant new findings about the processing that occurs in IT, and while we may still be a long way from understanding the exact computational role of this brain region, the work in this thesis highlights that complex visual processing occurs in IT well beyond the initial feed-forward sweep of visual processing that is often exclusively analyzed in many studies of this brain area.

The findings in chapter 3, showing that abstract category information is coded by changing patterns of neural activity (Figure 3.6, and Figure 3.7), are particularly interesting when compared to the findings other information in IT is coded by more stationary patterns of neural activity (see Additional supplemental material 5.11, and to a lesser degree Supplemental figure 3.7). A recent study by Crowe et al., (2010) used similar population decoding methods to show that there is dynamic coding of relative spatial information in parietal area 7a, while information about task-irrelevant visual stimulus properties were coded by static patterns of neural activity. Based on these findings, Crowe et al., (2010) suggested that such dynamic population coding might mediate task-critical cognitive processing, while static coding might be more related to task irrelevant information. This interpretation seems fairly reasonable to us[26], and additionally, dynamic coding might be related to memory processing, since such dynamic coding is often seen in tasks that have a memory component to them (Baeg et al., 2003; Zaksas and Pasternak, 2006; Pastalkova et al., 2008). Having a dynamic code in memory tasks makes sense because such tasks require that new information is compared to what was previously seen, and if a static code were used, this new information would overwrite the information about previously seen stimuli, making it impossible to complete the task.

[26] Where this interpretation agrees with all our data is questionable however (see Additional supplemental material 3.17 and Additional supplemental material 4.2)

Finally, one issue that arisen in all data analyses in this thesis, is the fact that we use always use pseudo-populations[27] of neural activity rather than examining neural activity that was actually recorded simultaneously. While the previous literature has suggested that correlated neural activity seems to contain only a relatively small amount of information (Panzeri et al., 2003; Averbeck and Lee, 2004; Aggelopoulos et al., 2005; Anderson et al., 2007), we thought it would be useful to do a few initial analyses to address this question. In collaboration with Jim DiCarlo's lab at MIT[28], we recorded several sessions of data using Utah probes that could simultaneously record the activity from 32 channels in area V4. Results comparing pseudo-population decoding accuracies to the decoding accuracies based on simultaneous recordings for natural images patches, color patches, polar, hyperbolic and regular gratings (Gallant et al., 1993), and for isolated objects (Hung et al., 2005) are shown in Figure 6.1. While further analyses need to be done (particular using more complex classifiers), the preliminary results show similar decoding accuracies are obtained using both types of data[29], which gives support to the notion that correlated activity might not contain additional information.

[27] By pseudo-populations we mean neurons that were recorded on separate sessions but were treated as if they were recorded simultaneously (see chapter 2, and the methods sections of chapters 3-5 for more details).

[28] Jennie Deutsch, Joel Leibo, and Cheston Tan, helped contribute to collecting this data.

[29] For a few cases, the results from the simultaneous recordings were slightly higher than the results for the pseudo-populations (e.g., on the natural image patches, and the isolated objects in Figure 6.1). These higher decoding accuracies seem to be due to the fact that there is correlated noise in the data (i.e., the whole population fluctuates up and down together), and the normalization process in the MCC classifier leads to slightly higher performance. When a raw cross-correlation classifier is used, these differences go away (and also these differences are not present when other decoding accuracy measures are used, such as the normalized rank, or the area under and ROC curve).

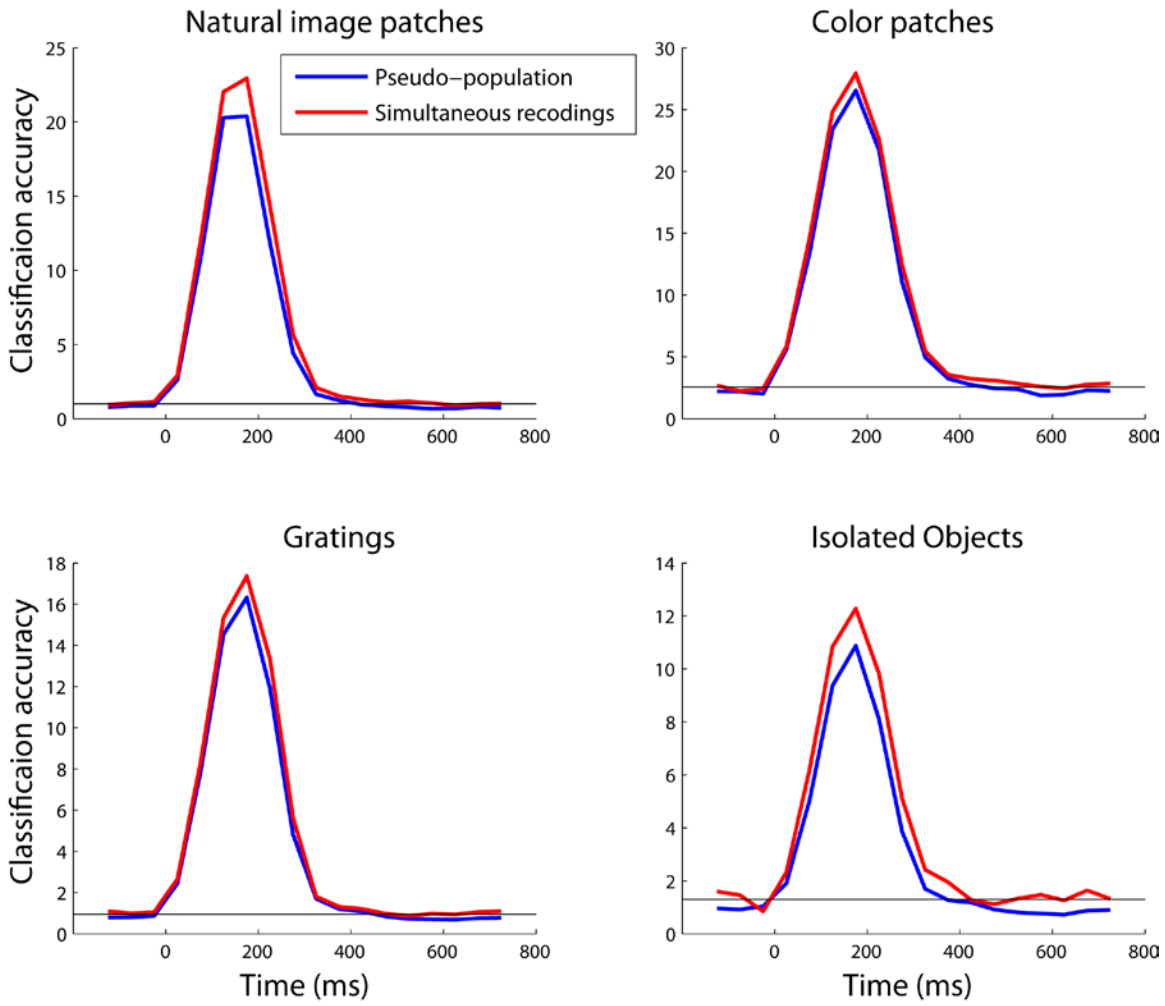


Figure 6.1 Classification accuracies are similar for simultaneously recorded populations and for pseudo-populations. Results are from the four datasets that consist of natural image patches, colored circles, Cartesian and non-Cartesian gratings, and isolated objects. An MCC classifier was using on firing rates calculated in 150ms bin sampled every 50ms. As can be seen, the results for pseudo-populations and simultaneous recordings are very similar.

Advantages of using decoding to analyze neural data

In the beginning of this thesis, we pointed out that neural decoding has many advantages when compared to more commonly used data analysis methods. However, before showing how population decoding can be applied to neural data, it was difficult to give examples of these advantages in a way that could be easily understood. At this point,

however, we feel it is instructive to once again revisit the advantages of neural decoding and illustrate these advantages by highlighting examples from this thesis. Below I list some of the strengths of population decoding, and give examples of how I have used these advantages in the analyses in my thesis.

1) **The ability to examine 'abstract' information.** One of the great strengths of population decoding is its ability to evaluate whether a population of neurons contains information that is abstract from physical properties of stimuli. To evaluate whether a population of neurons contains abstract information, one simply trains a classifier on one set of conditions and then test the classifier on a related set of conditions. For example, to evaluate whether ITC contains information about a visual object's identity that is abstract from the position of the object, one can train the classifier using data from trials when stimuli were shown at one position, and then test whether the classifier can classify these stimuli at a different position (see Figure 2.5 and Hung et al., 2005). Another example where I use this in thesis is to examine whether ITC and PFC contain abstract category information (e.g., information about whether a stimulus belongs learned category, that is separate from the visual properties of the stimulus). To do this analysis I trained the classifier to discriminate between images that belonged to two categories using one set of stimuli, and then tested the classifier on a different set of visual stimuli that were also members of these same categories. Because the visual image used in the training and test set were distinct, the ability of the classifier to perform well on this task had to come from the fact that through the monkey's experience learning this categorization task, neurons in ITC and PFC started responding similarly to the category of the stimuli, regardless of the visual features present in particular images.

The reason that the ability to evaluate whether abstract information is present is important, is because if such information is present, it must have been activity constructed by neural processes, which strongly suggests this abstract information is actually being used by the brain to influence behavior. In contrast, evaluating non-abstract information, (for example, if one is trying to decode the exact same stimuli using data from data from different trials), is less interesting because even if there is a high

amount of information present, it is hard to interpret whether this information is used to influence behavior since this information is inherent in the stimulus, and thus could be passively propagated through the brain. Most conventional data analyses methods do not have the ability to readily evaluate whether a brain region has abstract category information, which highlights a major advantage of neural population decoding.

2) **The ability to examine questions related to neural coding.** Another strong point of population decoding, is that the method allows one to examine questions about how information is coded in neural activity. At the moment it is unclear whether all information is contained in the firing rates of neurons, or whether additional information is contained in inter-spike intervals, synchronized activity of many neurons, or in some other form of neural activity. Population decoding analyses can examine these questions by using different neural features (such as inter-spike interval times or firing rates in different bin sizes) and comparing how high the decoding accuracy is with these different features. If a much higher decoding accuracy can be obtained with one type of neural feature representation than another, this suggests that the brain might be using this type of neural code to transmit information. In Additional supplemental material 3.10 I compare how decoding accuracies when using firing rates calculated in different bin sizes, which is an example of how this method can be applied[30].

3) **The ability to compare different types of data.** Because many different types of data can be fed into pattern classification algorithms, it is possible to compare how much information is in different types of signals. Signals I have looked at in the past include: single unit spiking activity (most data in this thesis), multi-unit spiking activity (e.g., Figure 6.1), computer vision features (see chapter 4), local field potentials, and functional

[30] Results from analyses comparing different bin sizes generally show that larger bin sizes contain more information. However whether the brain averages activity over longer periods still remains an open question due to the fact that the data I analyzed was not collected simultaneously. Thus, this higher decoding accuracy could be a result of the fact that the information was present at different latencies in different trials, and using larger bins is really just averaging out differences between trials rather than saying something about the temporal precision of the neural code (i.e., the neural code could be very precise in terms of a joint activity across many neurons, but this joint activity could occur at different latencies on different trials).

magnetic resonance imaging. Also because population decoding analyses report results as the percentage of conditions correctly classified, it is possible to compare decoding results to human performance on psychophysics tasks when similar measures are used (see chapter 4). Thus it is possible to evaluate how many different types of neural signals compare to the behavioral performance of humans and/or other primates.

4) **Robustness to selection biases (and the ability to examine sparseness/compactness of information).** Selection bias appears to be widespread in the analysis of neural and fMRI (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). One example of where selection bias is likely to be present is when researchers plot examples of individual neuron's activity to show that single neurons have particular effects, as is commonly done in many papers studying the visual properties of neurons. The reason that selection bias can be a problem is that if example neuron was chosen just because it showed desired effect (rather than being selected randomly), then the observed effects could just be due to noise (i.e., it is likely that just by chance one could find a single neuron that shows the desired effect). Another common example where selection can arise, is when researcher first screen a population of neurons for a particular property (such as only using neurons that are visually selective), and then applying additional analyses using only using these neurons. While analyzing subpopulations might make sense in particular circumstances, often the data used to select the subpopulation contains the same information that the subsequent analyses rely on, creating dependencies between these statistical tests, which can lead to incorrect conclusions (see Additional supplemental material 3.7)

The cross-validation procedure used in neural population decoding analyses can help reduce selection bias. By having separate training and test data splits, it is possible to use the training set to select a specific subpopulation of data, and then the test set can be used to validate that the selected data actually contains additional desired properties. For example, one could test if decoding results are affected by non-visually selective neurons by doing the decoding analyses using all the data, and comparing the results to a decoding analysis that first applies data selective to the training set to find the visually selective neurons, and then evaluating the classifier's performance on the test set using

only these visually selective neurons[31]. Additionally, it is possible to use these cross-validation methods to evaluate whether information is coding in a small subset of neurons by selecting only highly selective neurons using the training set data, and then evaluating whether this small subset of neurons contains all the available information on the test set (see Figure 3.4). Determining whether there is a compact set of neurons that contain all available information using convention statistics is difficult because such statistics give a fixed level of false positives (as determined by the alpha level used), and also because conventional statistics applied separately to a set of neurons cannot evaluate whether neurons contain redundant information with one another.

5) **The ability to analyze joint activity of neurons.** Examining the properties of many neurons jointly has several advantages over commonly used analyses that examine each neuron individually. From a purely theoretical point of view, it is widely believed that information is processed in the brain by the joint activity of many neurons, thus this type of data analysis can explore neural activity that potentially more relevant to the neural information processing. From a data analysis perspective, analyzing the joint activity of neurons allows for more robust estimates by pooling the seemingly weak and noisy signals that individual neurons have in order to obtain a more robust signal, thus allowing one to clearly see the flow of information with a relatively high temporal precision (a good example of this is figure 5.3 where we can see that ITC is switching between representing different objects as the monkey changes the focus of attention). Additionally, examining the joint activity of neurons allows one to explore questions that cannot be addressed with single neuron analyses, including, is additional information coded in the joint activity of neurons (see Figure 6.1) and how redundant is the information coded by different neurons in a population (see Figure 3.5).

[31] The analyses I have done examining this question have shown that decoding results are very robust to including non-selective neurons., i.e., including non-visually selective neurons does not appreciably change the results in any of the analyses I have done (See Additional supplemental material 5.14)

6) **A change in perspective from finding significant effects, toward a better understanding of the importance of observed effects.** Many analyses of neural data focus on finding statistically significant differences in neural activity between different conditions, (often by computing different 'selectivity indices' and applying standard statistical tests such as t-tests and ANOVAs to these values). While this hypothesis testing approach is effective in showing that an effect exists, it leaves the task of determining the importance of the observed effect unanswered. Thus if often feels that the field of neuroscience is overrun by facts, while lacking a clear way to piece these facts into a coherent picture that can explain how the observed effects can influence behavior. Neural population decoding can help to give more insight into developing a real computational understanding of neural processing in several ways. First, since population decoding can evaluate the magnitude that different effects contribute to particular tasks, this method can give insight into which effects are most important. Second, population decoding forces one to evaluate data in a way that is potentially more biologically relevant (see chapter 5). Many common data analyses only analyze responses to stimuli that the maximally excite a neuron thus generating a dataset that is highly unrepresentative of the most common neuronal responses (or create indices that have no direct meaning in terms of the activity of a neuron on a single trial). Because population decoding has to make predictions about each stimuli/conditions regardless of whether it is a 'preferred stimulus' or not, population decoding gives a more realistic picture of neural processing that is occurring on individual trials. Third, as mentioned before, population decoding results can be compared to the performance of an animal and to computer vision systems, making the results directly interpretable in terms of computational goals of a biological/computational system (see chapter 4). Finally, population decoding can easily track the state of a population of neurons, which enables one to get a better view of the dynamic computations (see Figure 3.6), which is probably a necessary first step toward algorithmic-like description of neural information processing.

Future directions

To conclude, I thought I would mention a few future directions that I would like to take this work. Many of my future research interests in neural population decoding involve analyzing data that has been recorded simultaneously using chronic recording methods (such as Utah probes), and also they involve creating better ways to share data, analysis code, and results in order to speed up the rate of discovery.

As discussed above, preliminary analyses of simultaneous recordings of V4 data from the DiCarlo lab indicate that there is pseudo-populations contain the same information as simultaneously recorded data, however I would like to examine this question more thoroughly by exploring the results using more complex classifiers and neural coding schemes. I would also like to examine the temporal precision of the neural code using by applying decoding methods to simultaneously recorded data using firing rates calculated over different timescales. While my previous results analyzing pseudo-population responses have shown that there is more information in longer time periods (see Additional supplemental material 3.10), the analysis of simultaneously recorded data could reveal information coded shorter timescales. For example, it is possible that a large amount of the trial-to-trial variability in spiking activity seen in almost all electrophysiological experiments is due to the same information being represented at slightly different times on different trials (i.e., neural activity is not perfectly time-locked to an experimentally chosen event). However, if one looks at activity of a whole population simultaneously using neural decoding, it might be possible to see temporally precise patterns of activity across the population that occur at slightly different times on each trial (e.g., the whole population could undergo state changes on a time scale of say 25ms, but these state changes could occur at different latencies on different trials). If this is the case, it could explain why neurons appear to be able to fire with a high degree of temporal precision and yet it seems that firing rates in longer bins contain the most information. Such a result would have a profound impact on the way electrophysiology data is collected and analyzed since it would indicate that averaging activity over repeated trials (e.g., PSTHs) could be missing essential features of the neural code.

Another direction I am interested in exploring concerns trying to better understand the visual features/dimensions that neurons in IT respond to. A major limitation of past studies exploring this question has been the fact that standard electrophysiological recording methods only allow a limited recording time (on the order of a couple of hours), and thus only a limited stimulus set can be shown to each neuron. By analyzing responses to a much large stimulus set collected over many days, it should be possible to get a much better sense of the visual features neurons are responding to. I am particularly interested in assessing how neurons respond to commonly encountered image transformation (e.g., affine transforms, contrast changes, etc.), and examining how these different images are related to each other in terms of distances in 'neural activity space'. This geometric perspective might reveal whether neurons are responding to lower level visual properties or to more behaviorally relevant factors and might give new insights over standard analyses that typically focus on the 'optimal' stimuli for each neuron.

Finally, to verify that the information that I am decoding is relevant for behavior, I would like to develop decoding algorithms that will allow for real-time decoding experiments in vision. One such experiment that am particularly interested in, involves training a monkey detect a change in either of two stimuli that are shown simultaneously. If the change in the stimulus is made subtle enough, the monkey will have to attend to only one of the stimuli at a time, and by applying population decoding, it should be possible to detect which stimulus the monkey is attending and thus predict the monkey's behavior on a trial-by-trial basis. If such predictions were successful, this would give us increased confidence that we are extracting information that is used by the animal.

Apart from analyzing data, I would also like to develop infrastructure that will allow researchers to better share data and software tools. Currently, there is no agreed upon standard method for analyzing electrophysiological data, so most researchers develop ad hoc methods for each paper that is published. This lack of standard methodology makes it very difficult to interpret the results reported in many papers. If the neuroscience community developed a culture that was more similar to that seen in biology, where data

and analysis tools are almost always shared, it should be much easier to tell which results are real and which are simply artifacts of the data analysis method, which in turn would lead to much more rapid progress. To try to move the field in this direction, I would like to develop common data formats to enable the sharing of data, and I would like to create a database where researchers can submit data, analysis software, and results. Hopefully, this will allow the field to converge on the best methods and allow for more confidence in the accuracy and interpretation of results reported in the literature, which in turn should lead to much more rapid progress in understanding the algorithms that underlie high level visual processing and other functions of the brain.

References

Aggelopoulos N, Franco L, Rolls E. Object perception in natural scenes: Encoding by inferior temporal cortex simultaneously recorded neurons. Journal of Neurophysiology 93: 1342-1357, 2005.

Anderson B, Sanderson M, Sheinberg D. Joint decoding of visual stimuli by IT neurons' spike counts is not improved by simultaneous recording. Experimental Brain Research 176: 1-11, 2007.

Averbeck BB, Lee D. Coding and transmission of information by neural ensembles. Trends in Neurosciences 27: 225-230, 2004.

Baeg E, Kim Y, Huh K, Mook-Jung I, Kim H, Jung M. Dynamics of population code for working memory in the prefrontal cortex. Neuron 40: 177-188, 2003.

Crowe DA, Averbeck BB, Chafee MV. Rapid sequences of population activity patterns dynamically encode task-critical spatial information in parietal cortex. J. Neurosci 30: 11640-11653, 2010.

Gallant J, Braun J, Van Essen D. Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. Science 259: 100-103, 1993.

Girard P, Jouffrais C, Kirchner C. Ultra-rapid categorisation in non-human primates. Animal Cognition 11: 727, 2008.

Hung C, Kreiman G, Poggio T, DiCarlo J. Fast readout of object identity from macaque

inferior temporal cortex. Science 310: 863-866, 2005.

Kirchner H, Thorpe SJ. Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. Vision Res 46: 1762-1776, 2006.

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12: 535-540, 2009.

Panzeri S, Pola G, Petersen R. Coding of sensory signals by neuronal populations: The role of correlated activity. Neuroscientist 9: 175-180, 2003.

Pastalkova E, Itskov V, Amarasingham A, Buzsaki G. Internally Generated Cell Assembly Sequences in the Rat Hippocampus. Science 321: 1322-1327, 2008.

Zaksas D, Pasternak T. Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. Journal of Neuroscience 26: 11726-11742, 2006.