Multivariate methods for tracking cognitive states

Kenneth A. Norman, Joel R. Quamme, & Ehren L. Newman

Department of Psychology
Princeton University
Green Hall
Washington Road
Princeton, NJ 08540

knorman@princeton.edu
Voice: (609) 258-9694
Fax: (609) 258-1113

To appear in F. Rösler, C. Ranganath, B. Röder, & R. H. Kluwe (Eds.), *Neuroimaging and Psychological Theories of Human Memory*. New York: Oxford University Press.

Draft version, January 28, 2008
This version may differ in some ways from the final published version.

Introduction

Most fMRI studies of memory focus on relating the activity of specific, localized brain regions to task conditions or to behavior. Based on this information, one can make inferences about how these regions contribute to memory, and about cognitive processes more generally.[1] In this paper, we describe a different, complementary approach: Multi-voxel pattern analysis (*MVPA*). Instead of trying to characterize the functional properties of individual brain voxels (volumetric pixels), MVPA involves applying pattern classification algorithms to *multi-voxel* patterns of brain activity, and training these classifiers to detect the spatially distributed neural correlates of specific cognitive states. Once a pattern classifier has been trained to detect the neural manifestation of a particular cognitive state, the classifier can be used to track the comings and goings of that state over time. For recent reviews of MVPA research, see Norman, Polyn, Detre, and Haxby (2006b) and Haynes and Rees (2006). The idea of analyzing multi-voxel patterns has a long history in fMRI data analysis (e.g., Friston & Buchel, 2003; Friston, Harrison, & Penny, 2003; McIntosh, Bookstein, Haxby, & Grady, 1996; McIntosh & Lobaugh, 2004; Calhoun, Adali, Pearlson, & Pekar, 2001). The difference between MVPA and other multivariate analysis methods is, in large part, one of emphasis: Other multivariate techniques have focused on characterizing functional relationships between brain regions, whereas MVPA is more focused on decoding the informational contents of particular brain states.

Importantly, the same pattern classification approach can be applied to other types of neuroimaging data besides fMRI; we discuss applications to EEG in this paper. To accommodate the fact that pattern analysis can be applied to multiple imaging modalities, the term MVPA can be construed more broadly as *multivariate* pattern analysis (to refer to the fact that MVPA factors in multiple aspects of the signal, whatever that signal might be), not just multi-voxel pattern analysis.

This paper is divided into two sections:

- In the first section, we provide a general overview of the MVPA approach, drawing on our recent review of MVPA methods (Norman et al., 2006b).
- In the second section, we show how MVPA can be used to address theoretically

meaningful questions about memory. The key idea here is that having a time-varying readout of the subject's cognitive state makes it possible to more directly test hypotheses about how specific cognitive states are related to behavioral outcomes.

Section 1: Overview of the MVPA approach

Patterns in the brain

The central idea that underlies the MVPA approach is that (to a first approximation) each cognitive state is associated with a characteristic pattern of brain activity. A study by Haxby, Gobbini, Furey, Ishai, Schouten, and Pietrini (2001) provides a useful illustration of how multi-voxel patterns of activity can be used to distinguish between cognitive states. Subjects viewed faces, houses, and a variety of object categories (e.g., chairs, shoes, bottles). The data were split in half for each subject (based on odd vs. even scanner runs), and the multi-voxel pattern of response to each category in ventral temporal (VT) cortex was characterized separately for each half. By correlating the first-half patterns with the second-half patterns (within a particular subject), Haxby et al. (2001) were able to show that each category was associated with a reliable, distinct pattern of activity in VT cortex (e.g., the first-half "shoe" pattern matched the second-half "shoe" pattern more than it matched the patterns associated with other categories; for similar results, see Spiridon & Kanwisher, 2002; Tsao, Freiwald, Knutsen, Mandeville, & Tootell, 2003; Carlson, Schrater, & He, 2003; Cox & Savoy, 2003; Hanson, Matsuka, & Haxby, 2004; O'Toole, Jiang, Abdi, & Haxby, 2005).

Sensitively detecting brain patterns

Given the goal of detecting the presence of a particular mental representation in the brain, the primary advantage of MVPA methods over individual-voxel-based methods is increased sensitivity. Conventional fMRI analysis methods try to find voxels that show a statistically significant response to the experimental conditions. To increase sensitivity to a particular condition, these methods spatially average across voxels that respond significantly to that condition. While this approach reduces noise, it also reduces signal in two important ways: First, voxels with weaker (i.e., nonsignificant) responses to a particular condition might carry some information about the presence/absence of that condition. Second, spatial averaging blurs out fine-grained spatial patterns that might discriminate between experimental conditions (Kriegeskorte, Goebel, & Bandettini, 2006).

Like conventional methods, the MVPA approach also seeks to boost sensitivity by looking at the contributions of multiple voxels. However, to avoid the signal-loss issues mentioned above, MVPA does not routinely involve spatial averaging of voxel responses. Instead, MVPA uses pattern classification algorithms, derived from computer science and statistics, to aggregate the (possibly weak) information that is present in the responses of individual voxels. Because MVPA analyses focus on high-spatial-frequency (and often idiosyncratic) patterns of response, MVPA analyses are typically conducted within individual subjects.

MVPA methods

Insert Figure 1 about here.

The basic MVPA method is a straightforward application of pattern classification techniques, where the patterns to be classified are typically vectors of voxel activity values. To illustrate these standard MVPA procedure, assume (for the purposes of this example) that we want to be able to decode whether the subject is viewing shoes or bottles based on fMRI activity.

The first step of an MVPA analysis is *feature selection*: Deciding which voxels to include in the pattern classification analysis (Figure 1A). As mentioned above, one of the defining features of MVPA is that it can make use of information provided by voxels that (on their own) do not meet conventional criteria for statistical significance. However, there is a cost to being too inclusive: If a voxel is especially noisy, the harmful effects of added noise (from this voxel) might outweigh the beneficial effects of added signal. As such, removing voxels with an especially poor signal-to-noise ratio prior to classification can greatly improve classification performance. There are several approaches to feature selection. Many studies use voxel-wise tests to weed out noisy voxels (e.g., Polyn, Natu, Cohen, & Norman, 2005). Another approach that is gaining popularity is to sweep a spherical “searchlight” around the brain and choose voxels based on whether the pattern of activity within the searchlight discriminates between the conditions of interest (Kriegeskorte et al., 2006). See Norman et al. (2006b) and Mitchell, Hutchinson, Niculescu, Pereira, Wang, Just, and Newman (2004) for additional discussion of feature selection methods.

The second step in an MVPA analysis, *pattern assembly*, involves sorting the data into discrete “brain patterns” corresponding to the pattern of activity across the selected voxels at a particular time in the experiment (Figure 1B). Brain patterns are labeled according to which experimental condition generated the pattern. This labeling procedure needs to account for the fact that the hemodynamic response measured by the scanner is delayed and smeared out in time, relative to the instigating neural event.

The third step, *classifier training*, involves feeding a subset of these labeled patterns into a multivariate pattern classification algorithm. Based on these patterns, the classifier learns a function that maps between voxel activity patterns and the labels (Figure 1C). Most MVPA studies have used *linear classification algorithms* such as linear support vector machines (Kamitani & Tong, 2005; Cox & Savoy, 2003; Mitchell et al., 2004) and neural network classifiers without a hidden layer (Polyn et al., 2005). Linear classifiers compute a weighted sum of voxel activity values. In some classifiers, this weighted sum is then passed through a decision function, which effectively creates a threshold for saying whether or not a category is present. The linear classifiers listed above all adjust weights in order to optimize the network’s ability to predict the labels of the training data; the details of how the weights are adjusted vary from classifier to classifier. For further discussion of how linear classifiers and other (nonlinear) classifiers have been applied to neuroimaging data, see Norman et al. (2006b).

The fourth step is *generalization testing*. In this step, the classifier is given new patterns of brain activity that were not presented at training (and were not used for feature selection). For each pattern, the classifier is asked to generate an estimate of the subject’s cognitive state (Figure 1D). If the new brain patterns have already been labeled (in this example, as shoes vs. bottles), we can evaluate the classifier's performance by seeing whether it predicts the correct label. However, for many MVPA applications the brain patterns in the generalization set have not been labeled (i.e., we do not know the "ground truth" of which cognitive state the subject is in at that moment). In this case, we can evaluate the classifier based on whether its estimate of the subject's cognitive state predicts the subject's behavior. This point is discussed in further detail in *Section 2*.

MVPA examples

Over the past several years, MVPA methods have been applied to a very wide range of problems, ranging from decoding the direction of movement of a viewed field of dots (Kamitani & Tong, in press) to decoding whether a subject intends to perform an addition or subtraction operation on two numbers (Haynes, Sakai, Rees, Gilbert, Frith, & Passingham, 2007). For a more complete listing of MVPA studies, see Norman et al. (2006b) and Haynes and Rees (2006).

Generating a temporal trace

Importantly, the increased sensitivity afforded by MVPA methods makes it possible to measure the presence/absence of cognitive states based on only a few seconds' worth of brain activity. If the cognitive states in question are sufficiently distinct from one another, discrimination can be well above chance based on single brain scans (acquired over a period of approximately 2 to 4 seconds) (Haynes & Rees, 2005a, 2005b; Polyn et al., 2005; Mitchell et al., 2004; O'Toole et al., 2005; Carlson et al., 2003; LaConte, Anderson, Muley, Ashe, Frutiger, Rehm, Hansen, Yacoub, Hu, Rottenberg, & Strother, 2003; LaConte, Strother, Cherkassky, Anderson, & Hu, 2005; Strother, La Conte, Hansen, Anderson, Zhang, Pulapura, & Rottenberg, 2004; Mouro-Miranda, Bokde, Born, Hampel, & Stetter, 2005). This increase in temporal resolution makes it possible to create a temporal trace of the waxing and waning of a particular cognitive state over the course of the experiment, which (in turn) can be related to subjects' ongoing behavior. For example, MVPA has been used to predict ongoing recall behavior in a free recall task (Polyn et al., 2005) (see *Case Study 1*, below), and it has also been used to predict changes in perceived stimulus dominance during a binocular rivalry task (Haynes & Rees, 2005b). The results of the 2006 Pittsburgh brain activity interpretation competition provide another example of how MVPA can be used to predict time-varying aspects of subjects' cognitive state (see http://www.ebc.pitt.edu/2006/competition.html). For this competition, subjects were scanned while they watched 3 episodes of the television show "Home Improvement" and then rated several aspects of their experience (e.g., subjects generated time-varying, real-valued ratings of how amused they were while watching the show). The winning entrants in this competition were able to decode several time-varying aspects of subjects' cognitive state, raging from subjective factors (amusement ratings) to more "objective" factors (whether tools were being used on screen).

Section 2: Testing psychological theories of memory with MVPA

The above discussion of MVPA illustrates how this method can be used to track subjects' cognitive state over time. The rest of the chapter is focused on how we can use this "thought-tracking" ability to test psychological theories of memory.[2] At a high level, psychological theories can be construed as collections of "if-then" statements: *If* the subject is in a particular cognitive state, *then* a particular outcome should take place. The standard, behavioral approach to testing theories is to set up experimental conditions that you expect will bring about the cognitive state of interest, and then look for the predicted outcome. The difficulty with this approach is that the mapping between experimental conditions and cognitive states is not perfect: Within a particular condition, subjects might slip in and out of the cognitive state of interest. As such, when the predicted outcome is not observed, there are always two possible explanations for this failure:

- The first explanation is that the theory is incorrect (i.e., the cognitive state in question does not elicit the predicted outcome).
- The second explanation is that the experiment did not succeed in eliciting the cognitive state of interest. In this case, the experiment's failure to elicit the predicted outcome does not speak to the validity (or lack thereof) of the theory in question.

One way of summarizing this point is that there is almost always variability in subjects’ cognitive state, above and beyond the variability that is directly driven by the experimental manipulation. In analyses that focus on comparing experimental conditions, this extra variability is treated as a source of noise and makes it harder to see the predicted effect.

MVPA gives us a way of addressing this problem: Instead of simply assuming that experimental conditions are effective in eliciting the cognitive state of interest, we can use MVPA to track that cognitive state and relate it to outcomes of interest. In paradigms where there is extensive uncontrolled variance in subjects’ cognitive state, this approach gives us a much more sensitive way of testing theories of how cognitive states drive behavior. Another benefit of MVPA is that it allows for more unconstrained designs: Instead of trying to lock in subjects’ cognitive state, MVPA gives us the option of letting subjects’ cognitive state “float” more naturally. So long as the classifier has been trained to detect fluctuations in the cognitive states of interest, we can use the classifier to soak up variance in the subject’s cognitive state and explore the consequences of these fluctuations.

Case studies

In the remaining part of section, we will present three case studies from our laboratory of how MVPA can be used to test psychological theories of memory.

All three experiments consist of two distinct parts:

- *Data for classifier training*: One part of the experiment is devoted to strongly and unambiguously eliciting the cognitive states of interest. Data from this part of the experiment is used to train the classifier to recognize these cognitive states.
- *Data for theory testing*: In the other part of the experiment, the trained classifier is then applied to new data (not presented at training) where the cognitive state(s) of interest are more variable. We then relate classifier’s readout of the subject’s cognitive state during this period to the subject’s behavior, in order to test whether the subject’s cognitive state predicts behavior in the manner predicted by the theory being tested.

The three case studies are as follows:

- In the first case study, we discuss how MVPA can be used to evaluate *contextual reinstatement* theories of recall (Polyn et al., 2005).
- In the second case study, we use MVPA to test some basic predictions of dual-process models of recognition (Quamme & Norman, 2006).
- In the third case study, we discuss how pattern classification methods can be applied to EEG data to track the fine-grained temporal dynamics of competition between mental representations (Newman & Norman, 2006). We also discuss how this approach can be used to test theories of how competition drives learning.

Note that the latter two case studies report preliminary data. Our focus here is primarily on explaining the logic of the studies and demonstrating the feasibility of the MVPA approach to theory-testing.

Case study 1: Testing contextual reinstatement

Insert Figure 2 about here.

A recent study by Polyn et al. (2005) set out to test the *contextual reinstatement* hypothesis of memory search (Tulving & Thompson, 1973; Bartlett, 1932). This hypothesis states that subjects target memories from a particular episode (or type of episode) by trying to reactivate characteristic patterns of mental activity from the to-be-remembered event. To the extent that subjects succeed in aligning the pattern of mental activity at recall with the general pattern of mental activity that was present at study, this will trigger recall of specific details from the event. The contextual reinstatement hypothesis can be framed as an “if-then” statement in the following manner: If the subject’s cognitive state at test matches the general properties of their cognitive state at study, then specific details should come to mind.

To test this hypothesis, MVPA methods were used to calculate the degree to which patterns of brain activity recorded during recall matched those seen during the initial encoding phase, on a time-varying basis. During the initial part of the experiment, subjects studied celebrity faces, famous locations, and common objects. Each stimulus category was studied using a different encoding task (for faces, subjects were asked how much they liked the celebrity; for locations, subjects were asked how much they would like to visit that location; and for objects subjects were asked how often they encounter that object). A neural network classifier was trained (separately for each subject) to recognize the pattern of brain activity corresponding to studying faces, locations, and objects. Then, subjects were asked to recall (in any order they liked, over a three minute period) the names of all of the faces, locations, and objects that they had studied earlier in the experiment, and the classifier was used to track the re-emergence (during this recall period) of brain patterns from the study phase.

This design conforms to the general design principles outlined at the beginning of *Section 2*: The classifier is trained using data from a part of the experiment where cognitive states are relatively well-controlled (the study phase), and the trained classifier is used to track mental activity from a part of the experiment where cognitive states are more variable (the recall phase).

There were two key predictions:

- During the recall period, subjects' brain state should come into alignment with brain states associated with studying faces, locations, and objects
- Reinstatement of study-phase activity associated with a particular category should predict recall of specific items from that category. Also, to the extent to that reinstatement is (at least in part) causing recall of specific items, reinstatement of category-specific study-phase activity should start to occur before recall of items from that category.

One thing to note about this design is that (because of MVPA) we can test the contextual reinstatement hypothesis without specifically asking subjects to reinstate context: Rather, we can let subjects' cognitive state fluctuate and explore the extent to which contextual reinstatement occurs naturally.

In keeping with the idea that subjects think about general event properties in order to remember specific details, Polyn et al. (2005) found that fluctuations in the strength of "neural reinstatement" over time were highly correlated with subjects' recall behavior. Figure 2a illustrates the close correspondence between classifier estimates of category-specific reinstatement and recall behavior in a single representative subject. Also, in keeping with the idea that reinstatement precedes (and triggers) recall, Polyn et al. (2005) found that — on average — category-specific patterns of brain activity (associated with studying faces, locations, and objects) started to emerge approximately 5 seconds before recall of specific items from that category (Figure 2b).

This study is not the first to show reinstatement of study-phase brain activity during recall (Wheeler, Petersen, & Buckner, 2000; Nyberg, Habib, & Tulving, 2000; Wheeler & Buckner, 2003; Kahn, Davachi, & Wagner, 2004; Smith, Henson, Dolan, & Rugg, 2004). The main difference between the Polyn et al. study and these other studies is that, because of the increased sensitivity of the MVPA approach, Polyn et al. were able to track the temporal dynamics of reinstatement over the course of the recall period and relate these dynamics to second-by-second changes in behavior.

Methodologically, the Polyn et al. (2005) study is significant insofar as it provides "proof of concept" that we can track cognitive states during an unconstrained memory retrieval task. Theoretically, the finding that reinstatement precedes recall provides some initial evidence in support of the contextual reinstatement hypothesis. However, more work is needed to evaluate this hypothesis. Insofar as the to-be-recalled items in the Polyn et al. (2005) study came from different semantic categories, it is possible that reinstatement effects in that study simply reflect subjects thinking about the semantic (categorical) properties of the items themselves, as opposed to subjects reinstating their "mindset" from the study phase (which should include information about how items were processed at study and how they were presented, in addition to core semantic features of the items). A much stronger test of the contextual reinstatement hypothesis would be to design an experiment where the same types of items are presented in different "contexts" (e.g., stimuli could be randomly selected words), and the only thing that differs across contexts is how the items are presented perceptually at study (e.g., words could be presented on top of different backgrounds) and/or how they are processed at study (e.g., different encoding tasks could be used for the different contexts). Experiments that fit this description are presently underway in our laboratory. We are also running studies that track contextual reinstatement in paradigms other than free recall (Frankel, Robison, & Norman, 2006).

Case study 2: Testing dual-process models of recognition

In the second case study, we explore how MVPA can be used to help test *dual-process* theories of recognition memory. The basic idea behind these theories is that recognition judgments can be driven by two distinct sources of information:

- Recollection of specific studied details
- Nonspecific feelings of familiarity

For a review of dual-process theories, see Yonelinas (2002). In recent years, researchers have started to develop computational models of recollection and familiarity, such as the Complementary Learning Systems (*CLS*) model (Norman & O’Reilly, 2003) and the Source of Activation Confusion (*SAC*) model (Reder, Nhouyvanisvong, Schunn, Ayers, Angstadt, & Hiraki, 2000). These models can be used to generate specific predictions about how a given manipulation will affect recollection and familiarity.

Dual-process decision-making

The major challenge that arises in testing the predictions of dual-process models of recognition is deciding how to combine the recollection and familiarity signals, in order to generate predictions about overall recognition performance (Wixted & Stretch, 2004). Put another way, how much should subjects “weight” recollection vs. familiarity when making a recognition decision?

Most dual-process models use very simple decision-making rules, where subjects’ use of recollection vs. familiarity does not vary as a function of situational factors. For example, Jacoby, Yonelinas, & Jennings, 1997 and Norman & O’Reilly, 2003 use a decision rule whereby subjects always consult recollection first; if the level of recollection is below a pre-specified threshold, then subjects consult familiarity. However, contrary to this view of dual-process decision-making (whereby recollection always takes precedence over familiarity), extant data suggest that numerous situational factors can influence the extent to which subjects rely on recollection. For example, Malmberg and Xu (2007) explored subjects’ utilization of recollection in an associative recognition paradigm, where subjects have to discriminate studied word pairs from *re-paired lures* generated by re-combining words from studied pairs. In this paradigm, familiarity and recollection have opposing effects on false recognition of re-paired lures: The fact that the individual items in the pair are familiar pushes subjects to say “old” but recollection of the actual pairs that were studied (i.e., “I studied window-banana, not window-shoebox”) pushes subjects to say “new”. To measure how strongly subjects were relying on recollection vs. familiarity, Malmberg and Xu (2007) measured how repeating pairs at study affects false recognition of re-paired lures: To the extent that subjects rely on familiarity, repeating pairs at study should boost false alarms (by making the items in re-paired lures more familiar). However, to the extent that subjects utilize recollection, repeating pairs at study should reduce false alarms (by increasing the odds that subjects will recollect the pairs they actually studied when given a re-paired lure at test).

The key finding from Malmberg and Xu (2007) was that subjects’ use of recollection was modulated by various aspects of the test procedure: Asking subjects to give confidence ratings at test and asking subjects to delay their responses both increased subjects’ use of recollection. Also, adding novel-item lures to the test (in addition to re-paired lures) reduced the extent to which subjects relied on recollection for the re-paired lures. Intuitively, the presence of novel items at test makes familiarity more useful (overall) as a basis for discriminating studied items vs. lures, reducing subjects’ incentive to use recollection.

Variability in subjects’ use of recollection can be explained in terms of two ideas: First, several studies have demonstrated that (over the course of a retrieval attempt) information about stimulus familiarity becomes available more quickly than recollected details (e.g., Hintzman & Curran, 1994; Gronlund & Ratcliff, 1989; Rotello & Heit, 1999). As such, recollection should play less of a role when subjects are responding relatively quickly. Second, using recollection requires more cognitive effort than using familiarity. For example Gruppuso, Lindsay, and Kelley (1997) found that dual-task demands hurt recollection-based responding more than familiarity-based responding. The idea that there is an “effort cost” associated with recollection-based responding implies that subjects will only draw upon recollection to the extent that the benefits (in terms of increased performance) outweigh the costs (in terms of increased effort and time).

Implications for theory-testing

The fact that subjects can strategically vary their use of recollection makes it difficult to test behavioral predictions of dual-process models. For example, the Complementary Learning Systems (CLS) model predicts that increasing *list strength* (i.e., strengthening some items on the list but not others) should impair recollection of nonstrengthened studied items, but it should not impair subjects' ability to discriminate nonstrengthened studied items from lures based on familiarity. To the extent that both recollection and familiarity contribute to recognition, and increasing list strength impairs recollection, this implies that list strength should also impair overall recognition sensitivity. However, several studies have failed to find a list strength effect for overall recognition sensitivity (e.g., Ratcliff, Clark, & Shiffrin, 1990). As discussed by Norman (2002), there are two possible interpretations of this finding:

- The first possibility is that the model is wrong, and that list strength does not affect recollection or familiarity.
- The second possibility is that the model is correct (i.e., list strength does affect recollection) but, for whatever reason, subjects were not making use of recollection in the studies that failed to find a list strength effect.

Put another way: "Use of recollection" is an uncontrolled variable in these studies and this makes it difficult to evaluate predictions about the properties of recollection (when it is being used). To address this problem, it is necessary to take steps to eliminate this uncontrolled variance.

There are two ways to address this uncontrolled variance: The standard approach is to adjust the paradigm in order to boost subjects' use of recollection. For example, to specifically address how list strength affects recollection, Norman (2002) explored list strength effects using a *plurality recognition* paradigm. In this paradigm, subjects have to discriminate between studied items, unrelated lures, and also switched-plurality lures (e.g., study "rats", test with "rat"). Prior work with this paradigm has established that discrimination of studied items and switched-plurality lures relies heavily on recollection of plurality information (familiarity is not useful insofar as switched-plurality lures are also familiar; Hintzman, Curran, & Oppy, 1992; Hintzman & Curran, 1994; Curran, 2000). To the extent that plurality discrimination depends on recollection, and list strength impairs recollection, increasing list strength should impair plurality discrimination. This prediction was confirmed by Norman (2002).

A different approach to the problem of uncontrolled variance in “use of recollection” is to use MVPA to extract a time-varying measure (based on brain activity) of whether subjects are using recollection. This approach potentially has several advantages over the first approach (i.e., adjusting the paradigm to boost subjects’ reliance on recollection):

- The first advantage is that MVPA can be used to study the properties of recollection in a wider range of situations. So long as subjects are using recollection on some fraction of the test trials, we can use the classifier’s readout of “use of recollection” to restrict the analysis to those trials.
- A second, related advantage is that this approach lets us collect data on how subjects vary their use of recollection on their own (i.e., when their strategies are not being strongly constrained), which will help us refine our theories of dual-process decision-making.
- A final point is that there are limits on our ability to control recollection: Even in the plurality paradigm, it seems likely that the level of effort that subjects expend on trying to retrieve specific details will wax and wane over time, which has implications for their behavior.

Paradigm details

Here, we present results from our initial attempt to use MVPA to track subjects’ use of recollection (Quamme & Norman, 2006). Our long-term goal is to use this technology to test sophisticated predictions of dual-process models, such as the list strength prediction described above. However, for our initial foray into this area, we decided to focus on a basic and relatively uncontroversial prediction of dual-process models: the idea (discussed above) that recollection of studied details can be used to oppose the familiarity of lures that are similar to studied items, thereby helping subjects avoid false recognition of these items.

To explore this idea, we used the plurality recognition paradigm described above (Hintzman et al., 1992). In this paradigm, familiarity pushes subjects to respond “old” to switched plurality lures, but recollection of studied plurality information pushes subjects to say “new” to these items. Thus, the straightforward prediction is that *if* subjects are using recollection, *then* they will be less likely to false alarm to switched-plurality lures.

As with our previous case study, this study used a two-phase design:

Phase 1: Classifier training

The goal of phase 1 was to train the classifier to recognize brain states associated with intentionally using recollection to make recognition judgments, vs. making recognition judgments based on familiarity. Subjects studied singular and plural words. For each stimulus, subjects were asked to mentally picture multiple objects if the word was plural and single objects if the word was singular (e.g., picture multiple shoes for the word “shoes” and single shoe for the word “shoe”). After the study phase, subjects were scanned while they were given recognition tests comprised of studied items (rats) and unrelated lures (bicycle); subjects were not given switched-plurality lures during this phase of the experiment. The key manipulation was to divide up the test into *recollection blocks* and *familiarity blocks*. For recollection blocks, subjects were told that they should try to recall specific details of the metal image they formed at study, and respond “yes” only if they were successful. For familiarity blocks, subjects were instructed to say “yes” if the word seemed familiar, and to ignore any details that they might recollect from the study phase. The classifier was trained to discriminate between brain patterns from recollection blocks and brain patterns from familiarity blocks.

Note that, although subjects were asked to focus on either recollection or familiarity (but not both) during phase 1, the classifier training procedure does not assume that recollection and familiarity are mutually exclusive. The only assumption that we make is that subjects rely relatively more on recollection during recollection blocks vs. familiarity blocks. The output of a classifier trained using this procedure indicates the *relative extent* to which subjects are relying on recollection vs. familiarity (i.e., does the pattern of brain activity more closely resemble the pattern associated with *relatively high* use of recollection, or does it more closely resemble the pattern associated with *relatively low* use of recollection).

Phase 2: Generalization testing

The goal of phase 2 was to use the trained classifier to explore subjects' use of recollection and familiarity, and to relate the classifier activity to behavior. Here, subjects were scanned while they were given a recognition test containing studied items, unrelated lures, and switched-plurality lures. Crucially, during this phase, subjects were not given any specific advice about whether to use recollection vs. familiarity to make their judgments. The trained classifier was used to estimate (on a scan-by-scan basis) how closely the subject's brain state resembled their brain state during recollection blocks vs. familiarity blocks from phase 1.

Predictions

As discussed above, we predicted that subjects would be more likely to falsely recognize switched-plurality lures when their brain is in a familiarity state vs. when their brain is in a recollection state. Importantly, while we expected an effect of familiarity vs. recollection state on responding to switched-plurality lures, we did not expect to see an effect of familiarity vs. recollection state on responding to studied items. For studied items, familiarity and recollection push responding in the same direction (i.e., they both push subjects to make an "old" response), so responding to studied items should be similar regardless of how much subjects are utilizing familiarity vs. recollection. The same logic applies to unrelated lures: These items are associated with low familiarity values and low levels of recollection. Both of these factors should push subjects to say "new", so responding to unrelated lures should be generally similar when subjects are utilizing familiarity vs. recollection.

Results

The first step in the classification analysis was to assess whether the classifier was able to reliably discriminate between brain states associated with recollection blocks vs. familiarity blocks in phase 1. If the classifier is unable to discriminate between recollection and familiarity blocks in phase 1, there is no reason to expect that the classifier will be able to accurately track subjects' use of recollection vs. familiarity in phase 2. To assess phase 1 accuracy, we trained the classifier on 3/4 of the phase 1 data and tested its ability to classify individual brain scans from the remaining 1/4 of the data. Across all 10 subjects, average classification accuracy was .59, which was significantly above .50 (chance), p < .01. Inspection of individual accuracy scores revealed that classification was well above chance for 6/10 subjects (accuracy > .60), and classification was basically at chance (accuracy between .48 and .53) for the remaining 4/10 subjects.[3] All subsequent analyses (exploring how a classifier trained on phase 1 generalizes to phase 2) were only run on the six subjects who showed above-chance classification performance on the phase 1 data.

Insert Figure 3 about here.

Figure 3 shows representative results from phase 2 (the plurals test) from one of these six subjects. Part A plots (over time) the classifier's readout of whether the subject's current brain state more closely resembles the "familiarity" brain state from phase 1 or the "recollection" brain state from phase 1. Part B plots (for this subject) the rate of saying "old" to studied items, switched-plurality lures, and unrelated lures, as a function of whether (according to the classifier) the subject was relying on recollection vs. familiarity. As predicted, false recognition of switched-plurality lures was higher when the subject was in a familiarity brain state vs. a recollection brain state, but responding to studied items and unrelated lures was relatively unaffected by whether the subject was in a familiarity brain state vs. a recollection brain state. We used non-parametric Monte Carlo statistical procedures to test the significance of individual-subject results. These procedures involve randomly scrambling the data and assessing the likelihood of obtaining the observed differences between "recollection state" vs. 'familiarity state" behavior, assuming no actual difference between conditions (for additional details regarding our non-parametric statistical procedures, see Polyn et al., 2005). For the subject shown in Figure 3, the difference in switched-plurality false alarms was significant (p = .014), as was the interaction between trial type (studied item, switched-plurality lure, and unrelated lure) and recollection/familiarity state (p = .025), indicating that recollection/familiarity state differentially affects responding to switched-plurality lures. We also ran a group analysis (across the 6 subjects who showed above-chance phase 1 classification) using standard parametric statistics (an ANOVA on the per-subject means) and obtained the same pattern of results: There was a significant effect of recollection/familiarity state on switched-plurality false alarms and a significant trial type X recollection/familiarity interaction; the effect of recollection/familiarity state on responding to studied items and unrelated lures was not significant.

Discussion

The pilot results presented above provide preliminary evidence that we can track subjects' use of recollection. Here, we discuss two current & future directions for this work. One major direction is *functional localization*: using MVPA to map out which brain regions contribute subjects' use of recollection vs. familiarity, and how these regions contribute. We also briefly describe how we can extend our analysis procedure to address more complex types of strategic variability.

With regard to functional localization: Numerous studies have used conventional, individual-voxel-based fMRI analysis methods to identify brain regions that are differentially activated when subjects are orienting to recollected details (e.g., judging the source of an item) vs. responding to item familiarity (see Wagner et al., 2005 for a review). These studies have identified a network of parietal regions (including the precuneus, retrosplenial cortex, posterior cingulate, and lateral parietal areas in and around the intraparietal sulcus) and frontal regions that are differentially recruited by tasks that place demands on recollection vs. familiarity. At this point in time, however, it is unclear *how* these regions are contributing. As discussed by Wagner et al. (2005), there are at least two reasons why a brain region might activate more strongly when subjects are trying to recollect details: One possibility is that the region helps to establish an *internally directed attentional state* ("listening for recollection") that amplifies hippocampal output. Another possibility is that the region implements processes that *operate on retrieved information*. For example, several researchers have argued that parietal regions may serve to accumulate evidence during decision-making (e.g., Huk & Shadlen, 2005; Ploran et al., 2007; Shadlen & Newsome, 2001).

To specify which regions contribute and how they contribute, we are currently running a variant of the pilot study described above, where – instead of applying the classifier to whole-brain patterns of activity – we are applying the classifier to patterns of activity from localized brain regions. Specifically, we are using the "searchlight" procedure mentioned in the *Feature Selection* section above (Kriegeskorte et al., 2006). This procedure involves sweeping a spherical searchlight (radius = 3 voxels) around the brain. For each location of the searchlight, we apply our two-phase classifier analysis (train on phase 1, generalize to phase 2) to the pattern of activity within the searchlight. The goal of this analysis is to find searchlight locations where the pattern of activity within the searchlight reliably discriminates between recollection vs. familiarity blocks during phase 1, and where the output of the classifier during phase 2 predicts behavior in the manner specified previously (i.e., classifier output indicating "recollection state" is associated with a decrease in false alarms to switched-plurality lures, but hits and unrelated-lure false alarms are relatively unaffected).

This searchlight procedure tells us, in an unbiased fashion, which regions carry information about subjects' use of recollection. Importantly, we should also be able to gain insight into how these regions contribute by examining *when* (relative to stimulus onset) classifier activity predicts behavior. If a brain region contributes to internally-directed attention, it should be possible to use the pattern of activity *prior to stimulus onset* to determine whether the subject is "listening to recollection" at that point in time. This information should (in turn) give us some ability to predict how the subject will respond to a test item, *before the test item actually appears*. In contrast, if a brain region is involved in processing recollected information, then the pattern of activity in that region should predict behavior *after* stimulus onset but not *before* stimulus onset. Importantly, it may be that some of the brain regions identified in the Wagner et al. (2005) review show timing profiles consistent with internally directed attention, and other regions show timing profiles that are more consistent with some kind of post-retrieval processing. We are examining these possibilities in our current work (Quamme, Weiss, & Norman, 2007).

In addition to exploring functional localization, we also plan to explore more complex models of strategic variability. As described earlier, our current paradigm allows us to measure the relative extent to which subjects are relying on recollection vs. familiarity. This measurement procedure assumes that subjects' recognition strategies vary along a single dimension (indicating the relative mix of recollection vs. familiarity). We consider this simple model to be a good starting point for our investigations of strategic processes in recognition memory, but we also acknowledge the possibility that subjects' strategies may vary along multiple dimensions. In particular, subjects may be able to independently vary their use of recollection and their use of familiarity (see, e.g., Wixted & Stretch, 2004). To accommodate this more complex model, we could add a third, "baseline" condition to phase 1. During baseline blocks, subjects would be asked to make simple perceptual judgments about studied and nonstudied words instead of recognition memory judgments. The presence of this third condition would force the classifier to discriminate recollection and familiarity states (individually) from the brain pattern that is present when subjects are not trying to use recollection *or* familiarity. During phase 2, a classifier trained in this fashion should be able to separately compute the strength of the recollection pattern vs. baseline and the strength of the familiarity pattern vs. baseline.

One final point regarding this case study is that, while we have focused on recollection and familiarity, the approach described here is quite general: In principle, it can be applied to any situation where there are multiple sources of information that could be used in making a decision, and subjects can choose to rely on some sources of information more than others.

Case study 3: Classifying EEG and tracking competitive dynamics

The fMRI classification methods described above are appropriate for tracking cognitive processes that vary on the order of seconds. This level of temporal resolution makes it possible to test theories about how cognitive processes vary across trials (or how cognitive processes vary across an extended recall attempt, as in the Polyn et al., 2005 study). However, temporal resolution of fMRI pattern classification is insufficient to address hypotheses about within-trial dynamics. For example, building on work by Anderson (2003) and others, Norman, Newman, and Detre (2007) developed a computational model of how competition between stored memory representations during a retrieval attempt can drive strengthening or weakening of the competing memories. To directly test this theory, we need a way of tracking the activation of memory representations as they compete, over the course of a single trial. Insofar as the retrieval competition plays out on the order of tens of milliseconds (as opposed to seconds), there is no easy way to accomplish this goal using fMRI.

To address this problem, we have started to explore ways of extending our pattern classification methods to other imaging modalities with better temporal resolution than fMRI (in particular, EEG). Pattern classification of EEG has a long history; most applications of EEG pattern-classification have focused on decoding movement-related activity (e.g., Peters, Pfurtscheller, & Flyvbjerb, 1998; Parra, Alvina, Tang, Pearlmutter, Yeung, Osman, & Sajda, 2002; Muller-Putz, Scherer, Pfurtscheller, & Rupp, 2005; Vallabhaineni & He, 2004; Wang, Deng, & He, 2004), although a few recent studies have used pattern classifiers to decode perceptually-related cognitive states (Philiastides & Sajda, 2006; Philiastides, Ratcliff, & Sajda, 2006).

This case study is divided up into two parts:

- First, we describe our preliminary attempts to classify subjects' cognitive state based on EEG data.
- Second, we describe how we plan to use these methods to test theoretical accounts of how competitive dynamics drive learning.

All of the results presented below were initially reported by Newman and Norman (2006).

Classifying EEG

Insert Figure 4 about here.

In our initial explorations of EEG classification, we used a delayed match to sample task, where subjects saw a *sample* stimulus (a photo of a face, a house, a chair, or a shoe), followed by a 500 ms mask, followed by a *probe* stimulus (see Figure 4 parts a and b). When the probe stimulus appeared, subjects had to judge whether the probe stimulus matched the sample stimulus. The probe was either the same photo (in which case subjects were instructed to respond "yes") or a photo of a different item from the same category (in which case subjects were instructed to respond "no").

The goal of our preliminary analyses was to see whether we could train a classifier (based on EEG data collected during the sample stimulus presentation) to discriminate between trials where subjects were viewing a face, a house, a chair, and a shoe (for related work, see Philiastides & Sajda, 2006 and Philiastides et al., 2006, who showed that it is possible to decode whether a subject is viewing a face or a car based on single-trial EEG). EEG data were collected using a 79 electrode cap, using a 1000 Hz sampling rate. After removing trials with excessive noise or blinks, we ran a wavelet decomposition on the data for each electrode to extract (for each EEG sample) oscillatory power at 49 frequency bands between 2 and 128 Hz. Then, for each trial, we computed the average oscillatory power value (for each frequency/electrode combination) for each of the 20ms "time bins" relative to the onset of the stimulus.

In the fMRI classification analyses described in *Case Studies 1 and 2*, the "brain patterns" that we fed into the classifier were vectors of voxel activity values (see Figure 1). For our EEG classification analyses, we applied a classifier to vectors of oscillatory power values, where each "feature" in the vector corresponds oscillatory power at a particular frequency, electrode, and time bin (relative to stimulus onset). As in our fMRI analyses, we did not run our classification analyses on the entire feature set: Only features that individually showed significant discrimination between categories (as indexed by a nonparametric statistical procedure) were used for classification. Finally, in keeping with the idea that different features could discriminate at different time points in the trial, we trained a separate classifier for each time bin (so, one classifier was trained to discriminate between stimulus categories based on data collected 0-20ms post-stimulus-onset; another classifier was trained to discriminate based on data collected 20-40ms post-stimulus-onset; and so on).

To test the classifier's generalization performance, we used a cross-validation procedure where the classifier was trained on 9/10 of the data and then tested on the remaining 1/10 of the data. Figure 4c plots generalization accuracy (averaged across 9 subjects) as a function of time bin. Average classification accuracy peaked at approximately .50 (chance = .25) at around 200ms post-stimulus-onset. In light of data showing that face stimuli elicit distinctive EEG patterns (Jeffreys, 1989; Itier & Taylor, 2004; Philiastides & Sajda, 2006), one might speculate that classification performance was being driven entirely by the face/non-face distinction (e.g., perfect face/non-face discrimination, but no ability to discriminate between non-face categories, would yield perfect accuracy for faces and .33 accuracy for the 3 non-face categories, leading to .50 overall accuracy). To address this hypothesis, we ran a follow-up analysis where we computed accuracy by category; average accuracy for the non-face categories was .42. This result shows that the classifier was picking up some information about the non-face categories (albeit less than it was picking up for faces).

Testing the competitor weakening hypothesis: Applications to negative priming

The above results show that we can decode category information with well-above-chance accuracy based on 20ms time bins of EEG data. Here, we discuss how this ability to track activation dynamics at a fine time scale can be used to test theories of how competitive dynamics affect learning.

Over the past decade, several researchers (see Anderson, 2003) have argued that retrieving a memory can have lasting consequences on memory strength, whereby the retrieved memory (the "winner" of the competition) is strengthened and other, "losing" memories are weakened. Crucially, Anderson has argued that this weakening effect is *competition-dependent*, such that the degree of weakening for a particular "losing" memory is proportional to how strongly it competes at retrieval. Recently, Norman et al. (2007) presented a neural network model that provides a concrete neural mechanistic account of competition-dependent forgetting, and relates this phenomenon to neural oscillations (see also Norman, Newman, Detre, & Polyn, 2006a for a discussion of how competition-dependent forgetting can boost the capacity of neural networks).

A large number of semantic memory and episodic memory findings can be explained in terms of competition-dependent weakening (for reviews, see Anderson, 2003, and Norman et al., 2007). For example, Anderson, Bjork, and Bjork (1994) had subjects study word pairs like Fruit-Apple, Fruit-Kiwi, and Fruit-Pear; they found that practicing retrieval of Pear (using the cue Fruit-Pe) impaired subsequent retrieval of Apple (a taxonomically strong Fruit) but not Kiwi (a taxonomically weak Fruit). Anderson et al. (1994) explained this finding in terms of the idea that taxonomically strong exemplars like Apple compete more strongly when subjects are trying to retrieve Pear, thus they suffer more weakening.

A key prediction of the competitive learning hypothesis is that, if a non-target (competing) memory *wins* the competition, it should be strengthened, not weakened. A retrieval-induced forgetting study conducted by Johnson and Anderson (2004) provides some support for this view. In the Johnson and Anderson study, subjects were asked to practice retrieving the subordinate meaning of a homograph (e.g., given the cue “prune”, subjects were asked to retrieve the non-dominant verb meaning, “trim”, over the dominant noun meaning, “fruit”). Johnson and Anderson found that, in some conditions, practicing retrieval of the subordinate meaning led to strengthening of the dominant meaning (for a similar result, see Shivde & Anderson, 2001). To explain this finding, Johnson and Anderson argued that (initially) the dominant meaning is so strong that subjects inadvertently recall it when trying to recall the subordinate meaning; since the dominant meaning wins the competition on these trials, it undergoes strengthening instead of weakening.

Another phenomenon that can potentially be explained in terms of competitor weakening is *negative priming* (e.g., Tipper, 1985; Fox, 1995). In a typical negative priming experiment, subjects are given stimulus displays consisting of two stimuli, a *target* stimulus and a *competitor* stimulus. Subjects are instructed to attend to the target stimulus and to ignore the competitor stimulus (e.g., the experiment might be set up such that the target stimulus is always tinted red, and subjects are asked to attend to the red stimulus). The key manipulation is that stimuli that serve as competitors on one trial sometimes appear as the target stimulus on later trials. Negative priming studies have found that, relative to stimuli that are being presented for the first time, subjects are faster to respond to stimuli that were previously attended and slower to respond to stimuli that were previously ignored (Tipper, 1985).

The basic pattern of negative priming results fits nicely with the competitive learning theory outlined above: Target items win the competition, so they are strengthened. Competitor items receive some support from the stimulus display (but not enough to win the competition) so they are weakened.[4] The competitive learning account may also provide a way of explaining variance in the size of the negative priming effect. For example, Fox (1994) found that reducing the spacing between the target and competitor stimuli significantly increased the magnitude of the resulting negative priming effect. Moving the competitor closer to the target should make it compete more strongly; according to the competitive learning account, this increase in competition should lead to greater suppression. Grison and Strayer (2001) found that degrading the perceptual quality of the competitor reduced the negative priming effect, supporting the idea that weaker competitors are suppressed less (see also Fox, 1998 and Strayer and Grison, 1999 for additional relevant findings). There is also evidence that, if the competitor becomes strong enough to win the competition, facilitation occurs instead of suppression. For example, Fuentes, Humphreys, Agis, Encarna, and Catena (1998) found that unifying the competitor and the target into a single visual object (a manipulation thought to increase processing of the competitor) caused the otherwise significant negative priming effect to reverse and become a significant positive priming effect.

The goal of our negative priming research is to provide more direct evidence in support of the competitive learning theory. Instead of making assumptions about competitor activation in a particular condition, we can use a pattern classifier (applied to EEG data) to directly measure target and competitor activation on a trial-by-trial basis, and then use these activation values to predict subsequent reaction times to the competitor. If the competitor activates strongly (i.e., the competitor happens to “win” on that trial), we expect to see positive priming. If the competitor activates weakly (i.e., the competitor is active, but “loses” to the target), we expect to see negative priming. The above predictions make it clear that, according to the competitive learning theory, negative priming should only occur when competitor activation falls within a narrowly defined range; too much or too little competitor activation will reduce the weakening effect or even lead to strengthening. This situation may help to explain why negative priming effects tend to be small (on the order of 20 ms). Using the classifier to focus on the set of trials where competitor activation falls in the correct range (not too high or too low) may allow us to observe a much more robust negative priming effect.

At present, this research is still ongoing and we are not yet in a position to make firm conclusions about the relationship between classifier activity and reaction time. Nonetheless, we think that our design for the study is instructive regarding the kinds of questions that one can address using MVPA (especially as applied to EEG), so we describe it below.

Negative priming experiment design

Our negative priming paradigm resembles the paradigm that we used in the preliminary classification study (described above). Subjects perform a delayed-match-to-sample task using face, house, shoe, and chair stimuli. However, in this version, each display contains two items: a *target* item (tinted red) overlaid on top of a *competitor* item (presented in grayscale) from a different category (e.g., a shoe overlaid on top of a face; see Figure 4d). Subjects are asked to make their match judgments based on the target item and to ignore the competitor item. Since the target and the competitor are from different categories, we can use the classifier to derive separate readouts of target and competitor activation (e.g., if the target is a shoe and the competitor is a face, we can use the classifier's readout of "shoe" as a proxy for target activation, and we can use the classifier's readout of "face" as a proxy for competitor activation). We then use these readouts of competitor and target activation to predict reaction time when the competitor stimulus used on this trial (e.g., the face) subsequently re-appears as a target.

Here, as in the other two case studies, we use the general approach of training on stable cognitive states and generalizing to situations where cognitive states are more variable. We reasoned that target stimuli would elicit stronger and more stable representations than competitor stimuli. As such, our analysis procedure involves training the classifier to recognize the category of the target stimulus, and then using the classifier (trained on a subset of trials) to measure the activity of the target category and the competitor category on other trials.

This analysis procedure assumes that classifiers trained to detect the target category can also detect the competitor category. Preliminary results support this view. Figure 4e plots (for a single subject) the average classifier output associated with the target category, the competitor category, and the two other categories not being presented on a given trial. Classifier output for the target and competitor categories was well above the level of classifier output associated with the two other categories. This finding (which was consistently present across subjects) suggests that we will be able to derive separate readouts of target and competitor activation for our negative priming study.[5]

Section 3: Discussion

In this paper, we presented three case studies of how MVPA can be used to test theories. In all three cases, the theory being tested could be described as an if-then rule:

- Case study 1: If *study-phase brain activity is reinstated at test*, then subjects will be more likely to recall additional studied details.
- Case study 2: If *subjects use recollection*, then subjects will be less likely to falsely recognize similar lures.
- Case study 3: If *the competing representation activates but then loses the competition*, then competitor weakening will occur (leading to a negative priming effect).

Testing these theories using MVPA involves a three-step process:

- First, we train the classifier to predict the cognitive state(s) of interest, using data from a part of the experiment where subjects' cognitive state is relatively well-controlled.
- Next, using data from a different part of the experiment (where subjects' cognitive state is less well controlled), we use the classifier to track the comings and goings of the cognitive states of interest.
- Finally, we plug these classifier estimates into theories (the if-then statements above) to generate behavioral predictions.

Effectively, this procedure is a form of bootstrapping: There are always going to be some situations where we have a relatively good understanding of what subjects are thinking, and other situations where we have a relatively poor understanding of what subjects are thinking. The goal of the multi-step procedure outlined above is to leverage our good understanding of subjects' cognitive state in one situation in order to gain insight into another (more murky) situation.

One of the key ideas motivating this approach is that the mapping between experimental conditions and cognitive states can be noisy: For example, we discussed in *Case Study 3* how it is very difficult to ensure that competitor activation falls within the range that is predicted to yield negative priming (i.e., not too high and not too low). By directly measuring the cognitive state of interest, we can soak up some of this within-condition variance that would otherwise be attributed to error. Another key benefit of MVPA is that it allows us to *covertly* measure cognitive variables of interest in situations where overtly asking about these variables might affect their strategies (e.g., asking subjects to directly report their use of recollection in *Case Study 2* might make them more likely to use recollection).

It is also worth discussing how our approach relates to concept of *reverse inference* (Poldrack, 2006). In the 2006 paper, Poldrack cautions against inferring that a cognitive process is present based on activation of a particular brain region. The problem with reverse inference is that a brain region might be activated by multiple different cognitive processes (other than the one of interest), making it impossible to ascertain which of these cognitive processes is causing the activation. To be clear, our MVPA analyses are a form of reverse inference. However, there are two properties of MVPA that mitigate the usual concerns about reverse inference. First, the patterns of activity that are detected by the classifier in MVPA are much more specific: While it is possible for a particular voxel to be involved in multiple disparate cognitive processes, the odds that a particular multi-voxel pattern will also be involved in very different kinds of cognitive processes are correspondingly lower (although perhaps not zero, depending on the pattern). The second, more important point is that we can obtain independent validation for our claims about the cognitive "significance" of a particular brain pattern by measuring whether the presence/absence of that pattern predicts subjects' behavior (in a manner consistent with our theory).

Brain-mapping with MVPA

Most of the MVPA examples in this paper have focused on using MVPA to track cognitive states, without much discussion of the specific neural instantiation of these states. However, as discussed at the end of *Case Study 2*, MVPA can also serve as a powerful tool for mapping cognitive functions onto brain structures, with strengths and weaknesses that are complementary to standard brain-mapping approaches. As discussed by Norman et al. (2006b), MVPA is not ideally suited to characterizing the roles of individual voxels; if your goal is to determine whether a particular voxel, on its own, contributes to a particular cognitive state, the best approach is to run a univariate analysis that focuses on that voxel. The chief advantage of MVPA, with regard to brain mapping, is that it can be used to sensitively assess whether a particular set of voxels (in aggregate) contains information about the cognitive states of interest (Kriegeskorte et al., 2006).

The future of MVPA

At this point in time, we are still at a very early stage in the development of MVPA methods. While individual-trial classifier performance is above chance in the examples described here, our ability to decode a subject's "instantaneous cognitive state" is still very far from 100% accurate. We typically need to look at data from a large number of trials in order to establish a link between classifier output and behavior. Also, there are limits on the kinds of cognitive states that can be resolved with extant MVPA methods. The Polyn et al. (2005) study described in *Case Study 1* constitutes a "best case scenario" for tracking cognitive states over time, insofar as Polyn et al. (2005) intentionally chose cognitive states (thinking about faces, locations, and objects) with highly discriminable neural substrates. As we increase the similarity of the cognitive states that are being studied, it stands to reason that our ability to track those states (at a fine time scale) should decrease accordingly.

Over time, we expect that improvements in imaging hardware and data analysis methods will boost the signal-to-noise ratio in MVPA analyses. Initial results from studies that have applied MVPA to high-resolution fMRI data are very promising (e.g., Sayres, Ress, & Grill-Spector, 2006). Also, better signal processing should help: There are several important properties of the fMRI signal relating to spatial structure (i.e., nearby voxels tend to represent similar things) and temporal structure (i.e., nearby time points tend to show similar patterns of activation) that are not routinely factored into MVPA analyses. Also, we expect that different cognitive states will fluctuate over time at different rates, and will vary in how they are instantiated in the brain (e.g., some cognitive states will have more localized neural substrates and other cognitive states will have more distributed neural substrates). As a general principle, giving the classifier better "priors" about what a particular cognitive state should look like in the brain (and how it should vary over time) should improve the classifier’s ability to track that cognitive state. Finally, one of the main factors limiting classifier performance in extant MVPA studies is lack of training data; there is only so much data that one can collect from a single subject in a single experiment. As such, classifier performance stands to benefit tremendously if we can develop ways of leveraging data from multiple subjects to constrain classifier estimates for a particular subject. For example, even though the “shoe” brain pattern might vary from subject to subject, it might be possible to use multi-subject data to develop better priors on how the “shoe” pattern might manifest itself in any one subject’s brain.

Tracking parameters

While the technical developments listed above are important, it is easy to get caught up in trying to optimize classifier performance and then lose track of the main question addressed in this paper, namely: “How can MVPA be used to test theories of memory? ”. The framework presented in this paper (where MVPA is used to test the validity of “if-then” statements relating cognitive states and outcomes) is a useful start. However, this “if-then” way of framing theories does not come close to capturing the nuance and complexity of extant computational models of memory. These models contain numerous parameters and state variables, and make precise quantitative predictions about how behavior should vary as a function of these factors (see Norman, Detre, & Polyn, 2008 and Raaijmakers, 2005 for reviews of extant computational models of memory).

In principle, it should be possible to extend the approach described in this paper in order to directly “read out” the parameters of quantitative models based on brain activity. We can scan subjects in situations where (according to the theory) the parameter is likely to be high, and situations where (according to the theory) the parameter is likely to be low. Then, we can train the classifier to discriminate between brain states associated with high and low values of this parameter. Once we have trained classifiers to read out values associated with key parameters of the model, we can track the values of multiple parameters and plug these values back into the model to generate quantitative predictions about subjects’ behavior.

Summary

Cognitive neuroscience theories, at their core, are about how the brain represents and processes information. One can translate these predictions about information processing into predictions about the overall level of activation in a particular brain region, and this approach has been highly productive (as described in other contributions to this volume). However, fine-grained information about the subject’s cognitive state gets lost in this translation process (insofar as two meaningfully different representational states can result in the same overall amount of activity). The main benefit of MVPA is that it allows us to skip this translation step and directly test predictions about the information that should be present in the subject’s brain at a particular point in time, and how this relates to behavior. MVPA has a long way to go before it fully delivers on this promise, but the potential payoff is extremely high: By eliminating the need to translate model predictions into predictions about overall activity, MVPA promises to provide a much more transparent and “higher-bandwidth” interface between theories and brain data.[6]

References

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*, 415–445.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *5*, 1063–1087.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Calhoun, V. D., Adali, T., Pearlson, G. D., & Pekar, J. J. (2001). Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. *Human Brain Mapping*, *13*, 43–53.

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, *15*, 704–717.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*, *19(2 Pt1)*, 261–70.

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory and Cognition*, *28*, 923.

Fox, E. (1994). Interference and negative priming from ignored distractors: The role of selection difficulty. *Perception & Psychophysics*, 56(5), 565-574.

Fox, E. (1995). Negative priming from ignored distractors in visual selection. *Psychonomic Bulletin and Review*, *2*, 145–173.

Fox, E. (1998). Perceptual grouping and visual selective attention. *Perception & Psychophysics*, 60(6), 1004-1021.

Frankel, H. C., Robison, S. G., & Norman, K. A. (2006). fMRI correlates of retrieval orientation: Tracking contextual reinstatement using pattern classification. Program No. 365.14. *2006 Neuroscience Meeting Planner*. Atlanta, GA: Society for Neuroscience. Online.

Friston, K., & Buchel, C. (2003). Functional connectivity. In R. Frackowiak, K.

Friston, C. Frith, R. Dolan, K. Friston, C. Price, S. Zeki, J. Ashburner, & W. Penny (Eds.), *Human Brain Function.* Academic Press, 2nd edition.

Fuentes, L. J., Humphreys, G. W., Agis, I. F., Encarna, C., & Catena, A. (1998). Object-based perceptual grouping affects negative priming. *Journal of Experimental Psychology: Human Perception and Performance*, 24(2), 664-672.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, *19*, 1273–302.

Gotts, S. J., & Plaut, D. C. (2005, April). Neural mechanisms underlying positive and negative repetition priming. *Poster presented at the 12th Annual Meeting of the Cognitive Neuroscience Society*.

Grison, S., & Strayer, D. L. (2001). Negative priming and perceptual fluency: more than what meets the eye. *Perception & Psychophysics*, 63(6), 1063-71.

Gronlund, S. D., & Ratcliff, R. (1989). Time course of item and associative information: Implications for global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 846–858.

Gruppuso, V., Lindsay, D. S., & Kelley, C. M. (1997). The process-dissociation procedure and similarity: Defining and estimating recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 259.

Hanson, S. J., Matsuka, T., & Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *Neuroimage*, *23*, 156–166.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425–2429.

Haynes, J.-D., & Rees, G. (2005a). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, *8*, 686–691.

Haynes, J.-D., & Rees, G. (2005b). Predicting the stream of consciousness from activity in human visual cortex. *Current Biology*, *15*(14), 1301–7.

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat Rev Neurosci*, *7*(7), 523–534.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Curr Biol*, *17*(4), 323–328.

Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? *Q J Exp Psychol A*, *58*(2), 193–233.

Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, *33*, 1–18.

Hintzman, D. L., Curran, T., & Oppy, B. (1992). Effects of similarity and repetition on memory: Registration without learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 667–680.

Huk. A. C. & Shadlen, M. N. (2005). Neural activity in macaque parietal cortex reflects temporal integration of visual motion signals during perceptual decision making. *Journal of Neuroscience, 25*(45), 10420-10436.

Houghton, G., & Tipper, S. P. (1994). A model of inhibitory mechanisms in selective attention. In D. Dagenbach, & T. H. Carr (Eds.), *Inhibitory processes in attention, memory, and language* (pp. 53–112). San Diego, CA: Academic Press.

Itier, R. J., & Taylor, M. J. (2004). N170 or N1? spatiotemporal differences between object and face processing using ERPs. *Cereb Cortex*, *14*(2), 132–142.

Jacoby, L. L., Yonelinas, A. P., & Jennings, J. M. (1997). The relation between conscious and unconscious (automatic) influences: A declaration of independence. In J. D. Cohen, & J. W. Schooler (Eds.), *Scientific approaches to consciousness* (pp. 13–47). Mahwah, NJ: Lawrence Erlbaum Associates.

Jeffreys, D. A. (1989). A face-responsive potential recorded from the human scalp. *Exp Brain Res*, *78*(1), 193–202.

Johnson, S. K. & Anderson, M. C. (2004). The role of inhibitory control in forgetting semantic knowledge. *Psychological Science, 15*(7), 448-453.

Kahn, I., Davachi, L., & Wagner, A. D. (2004). Functional-neuroanatomic correlates of recollection: Implications for models of recognition memory. *Journal of*

Neuroscience, 24, 4172–4180.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience, 8,* 679–85.

Kamitani, Y., & Tong, F. (in press). Decoding seen and attended motion directions from activity in the human visual cortex. *Current Biology*.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868.

LaConte, S., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L. K., Yacoub, E., Hu, X., Rottenberg, D., & Strother, S. (2003). The evaluation of preprocessing choices in single-subject BOLD fMRI using NPAIRS performance metrics. *Neuroimage*, *18*(1), 10–27.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., & Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. *Neuroimage*, *26*(2), 317–29.

Malmberg, K. J., & Xu, J. (2007). On the flexibility and fallibility of associative memory. *Memory and Cognition, 35*(3), 545-556.

McIntosh, A. R., Bookstein, F. L., Haxby, J. V., & Grady, C. L. (1996). Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage*, *3*(3 Pt 1), 143–57.

McIntosh, A. R., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: applications and advances. *Neuroimage*, *23,* S250–63.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine Learning, 5,* 145–175.

Mourao-Miranda, J., Bokde, A. L., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. *Neuroimage*, *28*(4), 980–95.

Muller-Putz, G. R., Scherer, R., Pfurtscheller, G., & Rupp, R. (2005). EEG-based neuroprosthesis control: a step towards clinical practice. *Neuroscience Letters*, *382*, 169–174.

Newman, E. L., & Norman, K. A. (2006). Tracking the sub-trial dynamics of cognitive competition. Program No. 365.2. *2006 Neuroscience Meeting Planner*. Atlanta, GA: Society for Neuroscience. Online.

Norman, K. A. (2002). Differential effects of list strength on recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28(6)*, 1083–1094.

Norman, K. A., Detre, G. J., & Polyn, S. M. (2008). Computational models of episodic memory. In R. Sun (Ed.), *The Cambridge handbook of computational psychology*. New York: Cambridge University Press.

Norman, K. A., Newman, E. L., & Detre, G. J. (2007). A neural network model of retrieval-induced forgetting. *Psychological Review, 114*(4), 887-953.

Norman, K. A., Newman, E. L., Detre, G. J., & Polyn, S. M. (2006a). How inhibitory oscillations can train neural networks and punish competitors. *Neural Computation*, *18*(7), 1577–1610.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review, 104*, 611–646.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006b). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci*, *10*(9), 424–430.

Nyberg, L., Habib, R., & Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *Proceedings of the National Academy of Sciences*, *97*, 11120.

O'Toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. (2005). Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of Cognitive Neuroscience*, *17*, 580–590.

Parra, L., Alvina, C., Tang, A., Pearlmutter, B., Yeung, N., Osman, A., & Sajda, P. (2002). Linear spatial integration for single-trial detection in encephalography. *Neuroimage*, *17*, 223–230.

Peters, B. O., Pfurtscheller, G., & Flyvbjerb, H. (1998). Mining multi-channel EEG for its information content: an ANN-based method for a brain-computer interface. *Neural Networks, 11*, 1429–1433.

Philiastides, M. G., Ratcliff, R., & Sajda, P. (2006). Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. *J Neurosci*, *26*(35), 8965–8975.

Philiastides, M. G., & Sajda, P. (2006). Temporal characterization of the neural correlates of perceptual decision making in the human brain. *Cerebral Cortex*, *16*, 509–518.

Ploran, E. J., Nelson, S. M., Velanova, K., Donaldson, D. I., Petersen, S. E., & Wheeler, M. E. (2007). Evidence accumulation and the moment of recognition: Dissociating perceptual recognition processes using fMRI. *Journal of Neuroscience*, *27*(44), 11912-11924.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn Sci*, *10*(2), 59–63.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes recall during memory search. *Science*.

Quamme, J. R., & Norman, K. A. (2006). Using fMRI pattern classification of recollection and familiarity to predict false alarms in recognition memory. Program No. 365.6. *2006 Neuroscience Meeting Planner*. Atlanta, GA: Society for Neuroscience. Online.

Quamme, J. R., Weiss, D. J., & Norman, K. A. (2007, November). Pattern classification of fMRI retrieval states in recognition memory. *Poster presented at the 48th Annual Meeting of the Psychonomic Society, Long Beach, CA.*

Raaijmakers, J. G. W. (2005). Modeling implicit and explicit memory. In C. Izawa, & N. Ohta (Eds.), *Human learning and memory: Advances in theory and application* (pp. 85–105). Mahwah, NJ: Erlbaum.

Ratcliff, R., Clark, S., & Shiffrin, R. M. (1990). The list strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 163–178.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. A. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 294–320.

Rotello, C. M., & Heit, E. (1999). Two-process models of recognition memory: Evidence for recall-to-reject. *Journal of Memory and Language*, *40*, 432.

Sayres, R., Ress, D., & Grill-Spector, K. (2006). Identifying distributed object representations in human extrastriate cortex. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (pp. 1169-1176). Cambridge, MA: MIT Press.

Shadlen, M. N. & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiol.*, *86*, 1916-1936.

Shivde, G. & Anderson, M. C. (2001). The role of inhibition in meaning selection: Insights from retrieval-induced forgetting. In D. S. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 175–190). Washington, DC: American Psychological Association.

Smith, A. P. R., Henson, R. N. A., Dolan, R. J., & Rugg, M. D. (2004). fMRI correlates of the episodic retrieval of emotional contexts. *NeuroImage*, *22*, 868–878.

Spiridon, M., & Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? an fMRI study. *Neuron*, *35*(6), 1157–65.

Strother, S., La Conte, S., Hansen, L., Anderson, J., Zhang, J., Pulapura, S., & Rottenberg, D. (2004). Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. a preliminary group analysis. *Neuroimage*, *23 Suppl 1*, S196–207.

Strayer, D. L., & Grison, S. (1999). Negative identity priming is contingent on stimulus repetition. *Journal of Experimental Psychology: Human Perception* *and Performance, 25*(1), 24-38.

Tipper, S. P. (1985). The negative priming effect: inhibitory priming by ignored objects. *Q J Exp Psychol A, 37*(4), 571–590.

Tipper, S. P. (2001). Does negative priming reflect inhibitory mechanisms? a review and integration of conflicting views. *Q J Exp Psychol A, 54*(2), 321–343.

Tipper, S. P., Bourque, T. A., Anderson, S. H., & Brehaut, J. C. (1989). Mechanisms of attention: a developmental study. *J Exp Child Psychol, 48*(3), 353–378.

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., & Tootell, R. B. (2003). Faces and objects in macaque cerebral cortex. *Nature Neuroscience, 6*(9), 989–95.

Tulving, E., & Thompson, D. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352–373.

Vallabhaineni, A., & He, B. (2004). Motor imagery task classification for brain computer interface applications using spatiotemporal principle component analysis. *Neurological Research, 26*, 282–287.

Wagner, A. D., Shannon, B. J., Kahn, I., & Buckner, R. L. (2005). Parietal lobe contributions to episodic memory retrieval. *Trends in Cognitive Sciences, 9*, 445-453.

Wang, T., Deng, J., & He, B. (2004). Classifying EEG-based motor imagery tasks by means of time-frequency synthesized spatial patterns. *Clinical Neurophysiology, 115,* 2744–2753.

Wheeler, M. E., & Buckner, R. L. (2003). Functional dissociation among components of remembering: control, perceived oldness, and content. *Journal of Neuroscience, 23,* 3869–3880.

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory’s echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences, 97,* 11125.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin and Review, 11,* 616–41.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.

Author Note

This research was supported by NIH grants R01MH069456 and P50MH062196 awarded to KAN, and NIH NRSA grant F31MH077649 awarded to ELN. The authors would also like to thank Jim Haxby and members of the Princeton Neuroimaging Analysis research group for helping us to develop these ideas.

Footnotes

[1] For discussion of the kinds of inferences that one can (and can not) make about cognitive processes based on localized fMRI activations, see Henson (2005) and Poldrack (2006).

[2] There are other ways to apply MVPA to theory-testing, other than tracking thoughts over time. For example, MVPA can be used to test theories regarding the *similarity-structure* of cognitive states. To a first approximation, similar cognitive states should be associated with similar brain states. As such, it should be possible to make inferences about cognitive similarity (e.g., whether bottles are more similar to scissors than to faces) based on the similarity of the multi-voxel patterns associated with these states. For an example of this approach, see O'Toole et al. (2005), and for a review of relevant studies see Norman et al. (2006b).

[3] We hypothesize that chance performance in these subjects was due to high rates of involuntary recollection during "familiarity blocks". In more recent versions of the experiment, we have tried to reduce involuntary recollection by speeding up the stimulus presentation rate at test during phase 1. This change has led to a substantial improvement in classification accuracy (in the updated version of the experiment, mean accuracy = .71 across 12 subjects, *SEM* = .03).

[4] For additional discussion of the idea that negative priming effects are competition-dependent, see Tipper (2001), Houghton and Tipper (1994), and Gotts and Plaut (2005).

[5] The results shown in Figure 4e are from a version of the experiment where we trained the classifier on single-image stimuli (i.e., where no competitor was present) and then applied the classifier to superimposed target-competitor images. The overall pattern of results is the same when we train the classifier to recognize the target category from superimposed target-competitor images, and then we apply the classifier to other

superimposed target-competitor images.

[6] The Matlab scripts that we use to run MVPA analyses are available for public download at http://www.csbmb.princeton.edu/mvpa

Figure Captions

Figure 1. Illustration of a hypothetical experiment and how it could be analyzed using MVPA. a) Subjects view stimuli from two object categories (bottles and shoes). A *feature selection* procedure is used to determine which voxels will be included in the classification analysis. b) The fMRI time series is decomposed into discrete *brain patterns* that correspond to the pattern of activity across the selected voxels at a particular point in time. Each brain pattern is labeled according to the corresponding experimental condition (bottle vs. shoe). The patterns are divided into a training set and a testing set. c) Patterns from the training set are used to train a classifier function that maps between brain patterns and experimental conditions. d) The trained classifier function $f(\vec{v})$ defines a decision boundary (red dashed line) in the high-dimensional space of voxel patterns (collapsed here to 2-D for illustrative purposes). Each dot corresponds to a pattern, and the color of the dot indicates its category. The background color of the figure corresponds to the guess the classifier makes for patterns in that region. The trained classifier is used to predict category membership for patterns from the test set. The figure shows one example of the classifier correctly identifying a bottle pattern (green dot) as a bottle, and one example of the classifier misidentifying a shoe pattern (blue dot) as a bottle. Figure reprinted with permission from Norman et al. (2006b).

Figure 2. Results from the Polyn et al. (2005) free recall study. a) Illustration of how brain activity during recall relates to recall behavior, in a single subject. Each point on the x-axis corresponds to a single brain scan (acquired over a period of 1.8 seconds, during the 3 minute recall period). The blue, red, and green lines correspond to the classifier’s estimate as to how strongly the subject is reinstating brain patterns characteristic of face-study, location-study, and object-study at that point in time. The blue, red, and green dots indicate time points where subjects recalled faces, locations, and objects; the dots were shifted forward by three time-points, to account for the lag in the peak hemodynamic response. The graph illustrates the strong correspondence between the classifier's estimate of category-specific brain activity, and the subject's actual recall behavior. b) Event-related average (incorporating data from 9 subjects) of the classifier's estimates of category-specific brain activity, for the time intervals surrounding recall events. This graph shows that the classifier starts to detect the to-be-recalled category several seconds before recall occurs. The dotted line at $t = 0$ represents the time point at which the verbal recall was made. The Currently Recalled plot (black line) shows average classifier activity for the category that was recalled at $t = 0$. The Baseline plot (purple line) shows average classifier activity for the two categories that were *not* recalled at $t = 0$. Points marked with stars and circles differ from baseline at $p < 0.01$ and $p < 0.05$, respectively. For additional details of how the plot was computed see Polyn et al. (2005). Parts a and b both adapted with permission from Polyn et al. (2005).

Figure 3. Classification results for phase two (plurals test) data from a single subject. a) Classifier output as a function of time for three runs of the plurality recognition task (each time point corresponds to a single brain scan, acquired over a 2 second period) The output measure plotted on the y-axis is the classifier's estimate of how well brain activity matches the "familiarity state" from phase 1, minus the classifier's estimate of how well brain activity matches the "recollection state" from phase 1. Blue regions indicate time points where, according to the classifier, the subject is in a familiarity state (familiarity > recollection); red regions indicate time points where, according to the classifier, the subject is in a recollection state (recollection > familiarity). Symbols ("o" and "x") at the top of each panel indicate time points when switched-plurality lures were presented. An "o" indicates a correct rejection by the subject and an "x" indicates a false alarm. These labels have been shifted forward by three time-points, to account for the lag in the peak hemodynamic response. b) Bar graph showing the proportion of hits (correct "old" responses to studied items) and false alarms (*FA*; incorrect "old" responses to switched plurality lures and unrelated lures), as a function of whether (according to the classifier)

the subject was in a familiarity state or a recollection state for that item. The figure shows that false alarms for switched-plurality lures were greater when the subject was in a familiarity state vs. a recollection state, but that no such difference was present for hits or for unrelated lure false alarms.

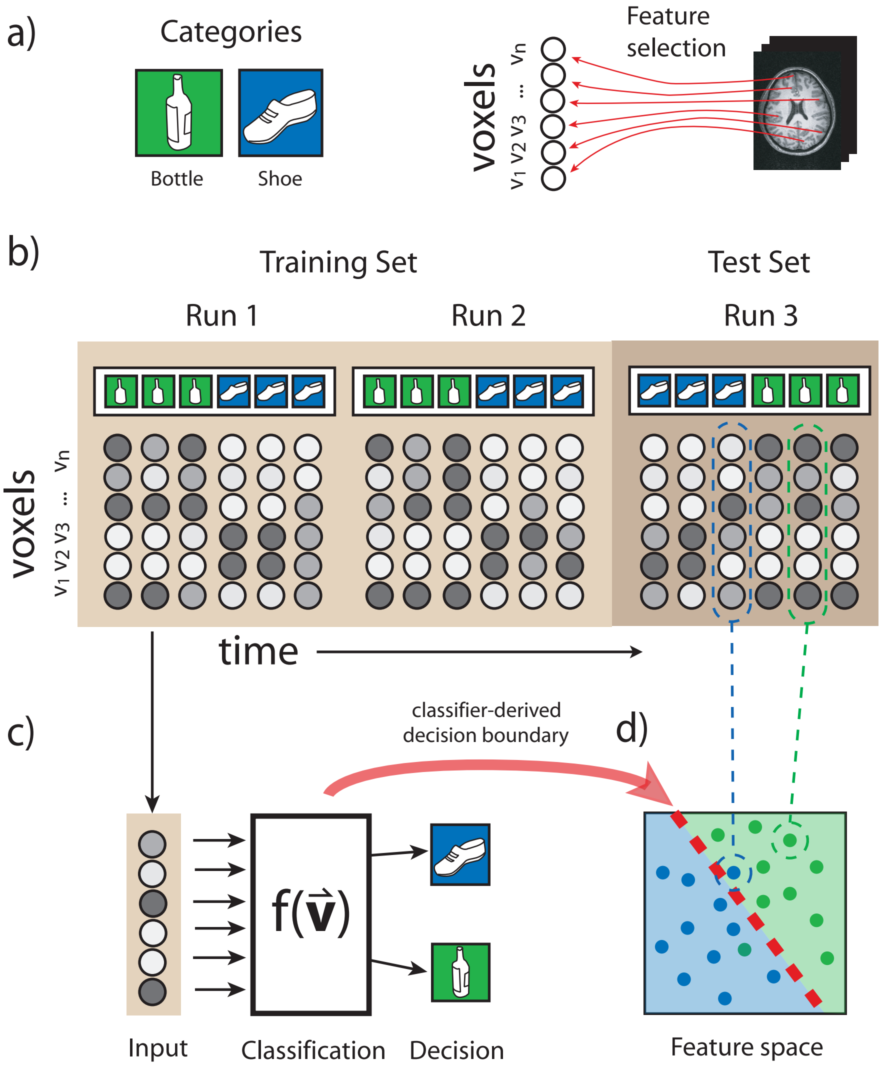
Figure 4. Illustration of stimuli and tasks used for EEG classification along with representative results. a) Our studies used four categories of images: Faces, Houses, Shoes, and Chairs. b) Illustration of the delayed-match-to-sample task used to present images to subjects. Only the EEG collected during the presentation of the sample stimulus was used to train and test the classifiers. c) Average classifier generalization accuracy over 9 subjects. The error bars indicate the standard error of the mean (across subjects). Classification rapidly increased to its peak accuracy value by approximately 200ms and then dropped back down to chance by 600ms (the fact that accuracy was above chance at $t = 0$ is an artefact of the wavelet decomposition procedure). d) Illustration of the negative priming task. Subjects perform a delayed-match-to-sample task; they are told to focus exclusively on red images (targets) and to ignore superimposed black and white objects (competitors). On approximately 15% of trials, subjects are required to respond to the object they just ignored (e.g., the bottom example shows a trial where subjects ignore a face and then have to respond to the face). e) Classifier output from a single representative subject for the negative priming task, showing the average activation of the target (to-be-attended) category and the competitor (to-be-ignored) category. The graph shows that both the target and competitor categories were more active, on average, than categories that were not present on screen.

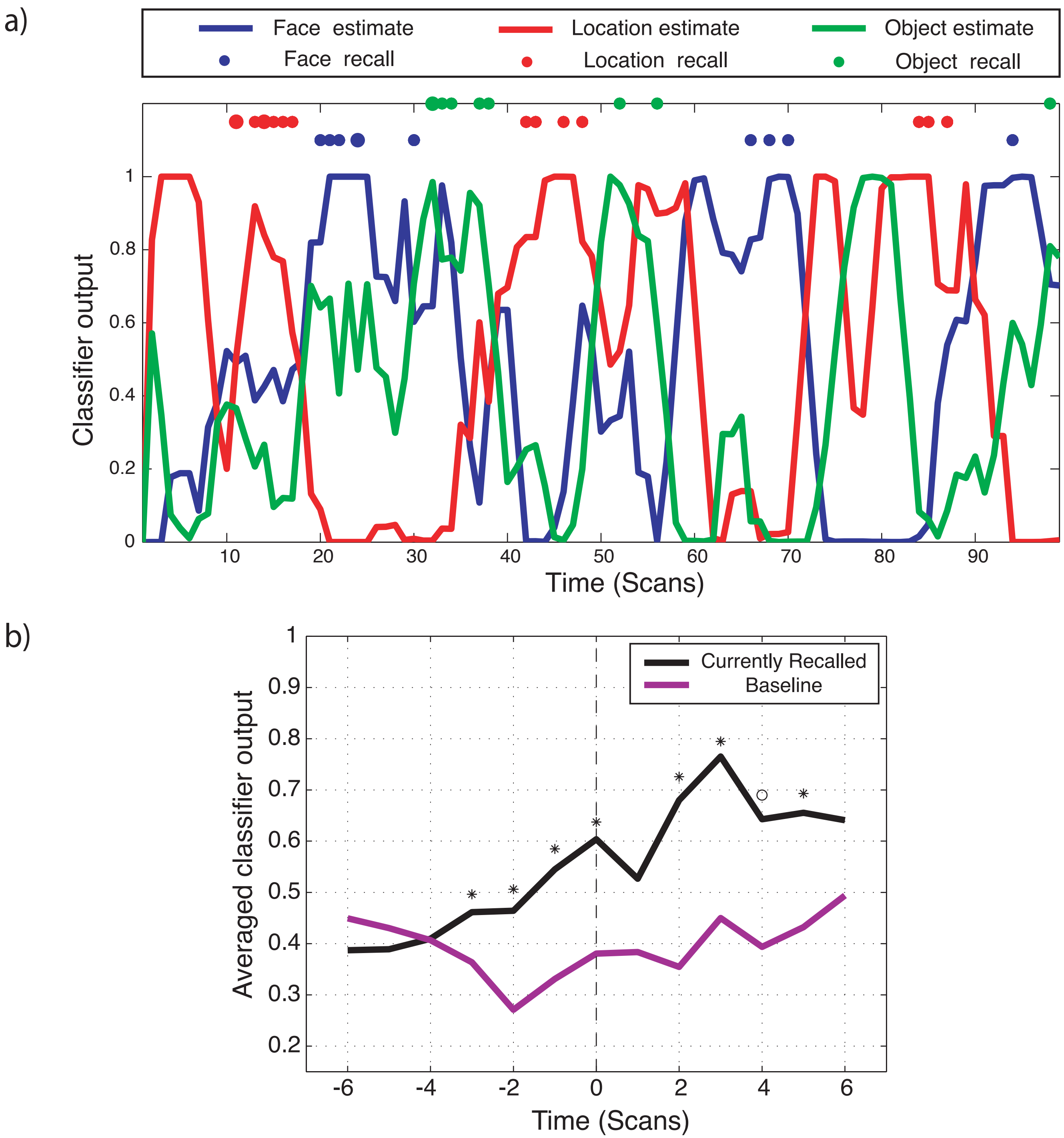
Abbreviations and Acronyms

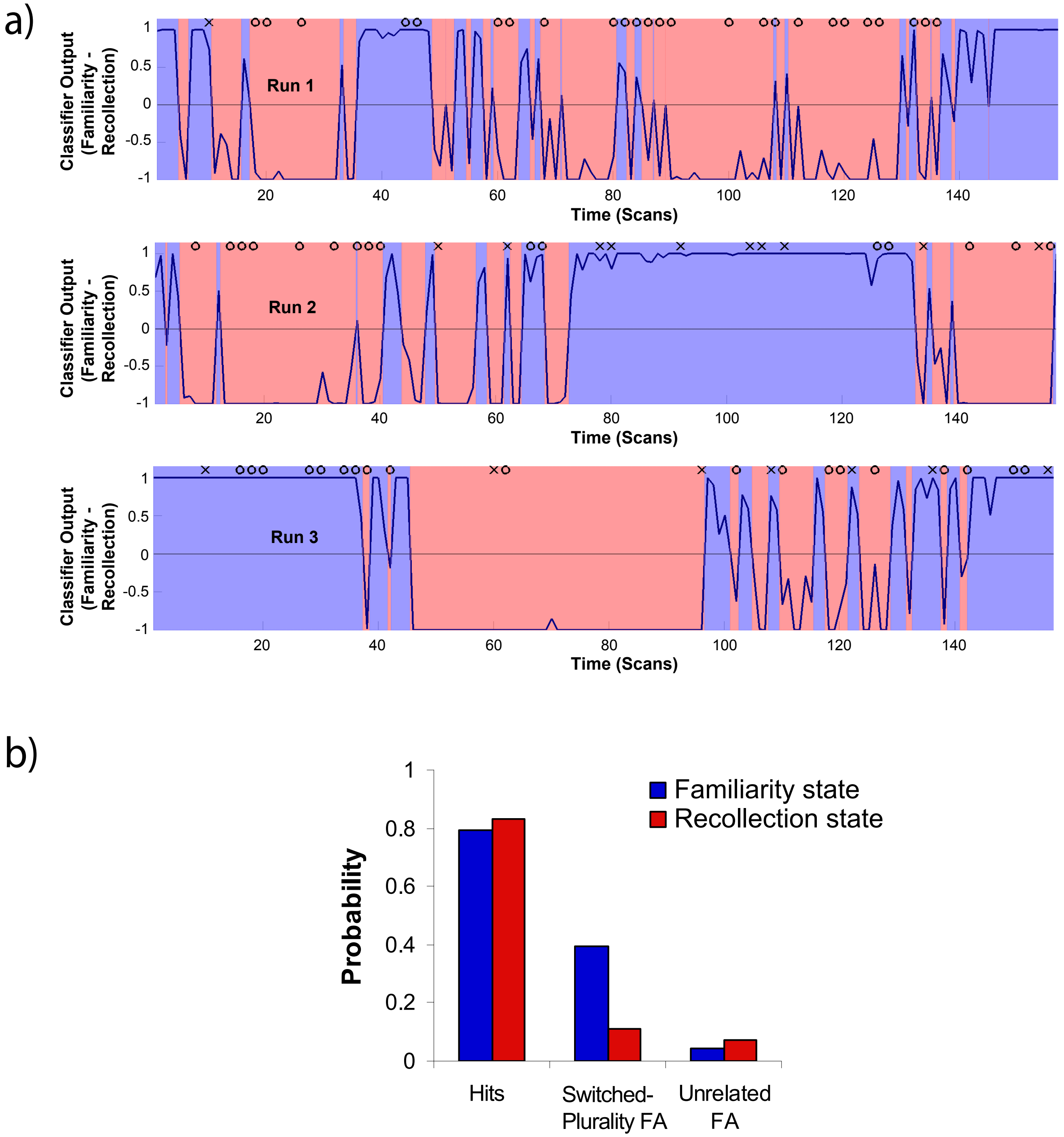
MVPA = multi-voxel pattern analysis
fMRI = functional magnetic resonance imaging
EEG = electroencephalogram
VT = ventral temporal
CLS = Complementary Learning Systems

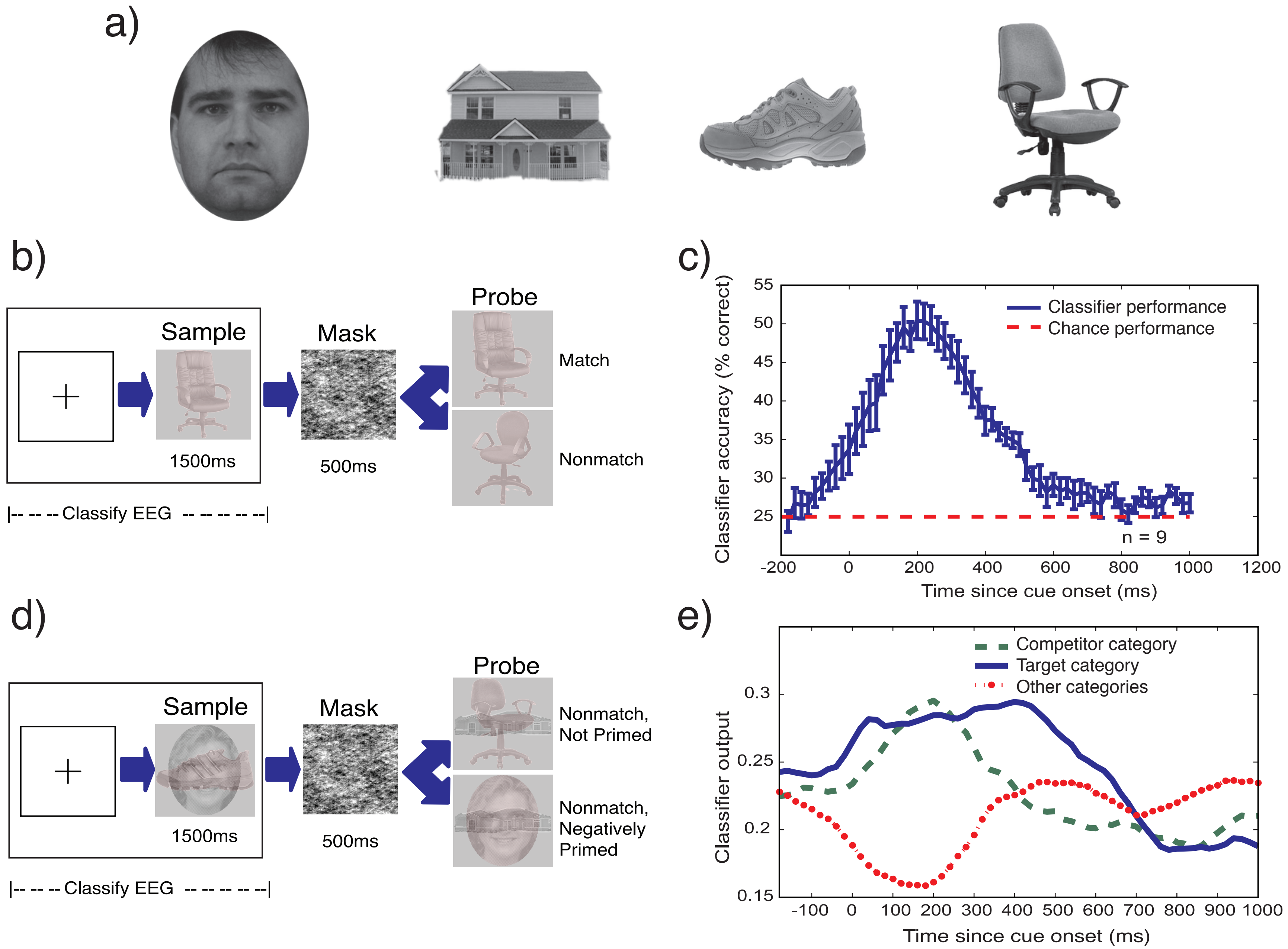
SAC = Source of Activation Confusion

FA = false alarm


a)
Categories
Bottle
Shoe
voxels
V1 V2 V3 ... Vn
Feature selection
b)
Training Set
Test Set
Run 1
Run 2
Run 3
voxels
V1 V2 V3 ... Vn
time
c)
classifier-derived decision boundary
d)
f($\vec{\mathbf{v}}$)
Input
Classification
Decision
Feature space


a)
Face estimate
Location estimate
Object estimate
Face recall
Location recall
Object recall
Classifier output
0
0.2
0.4
0.6
0.8
1
10
20
30
40
50
60
70
80
90
Time (Scans)
b)
Currently Recalled
Baseline
Averaged classifier output
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
-6
-4
-2
0
2
4
6
Time (Scans)


a)
Run 1
Run 2
Run 3
Classifier Output (Familiarity - Recollection)
1
0.5
0
-0.5
-1
20
40
60
80
100
120
140
Time (Scans)
b)
Probability
1
0.8
0.6
0.4
0.2
0
Familiarity state
Recollection state
Hits
Switched-Plurality FA
Unrelated FA


a)
b)
Sample
Mask
Probe
Match
Nonmatch
1500ms
500ms
|-- -- -- Classify EEG -- -- -- -- --|
c)
Classifier accuracy (% correct)
55
50
45
40
35
30
25
Classifier performance
Chance performance
n = 9
-200
0
200
400
600
800
1000
1200
Time since cue onset (ms)
d)
Sample
Mask
Probe
Nonmatch, Not Primed
Nonmatch, Negatively Primed
1500ms
500ms
|-- -- -- Classify EEG -- -- -- -- --|
e)
Classifier output
0.3
0.25
0.2
0.15
Competitor category
Target category
Other categories
-100
0
100
200
300
400
500
600
700
800
900
1000
Time since cue onset (ms)