Princeton Computational Memory Lab

CSBMB

Detecting distributed patterns in an fMRI study of free recall.

Sean M. Polyn, Jonathan D. Cohen & Kenneth A. Norman
Department of Psychology, Princeton University, Princeton, NJ, USA

79.14

Introduction

Both theoretical and intuitive accounts of episodic recall focus on the importance of reinstating context. These theories posit that – at study – individual stimuli are associated with more stable, "contextual" aspects of the study episode (relating to general characteristics of items, how they were presented, and how they were processed). At test, subjects use reinstated contextual information to cue for specific studied details. According to this view, retrieval success at test should be a function of how well subjects are able reinstate their neural representation of the study context.

We tested this idea using an fMRI study of free recall, in which subjects were scanned during both the study and retrieval phases. Three categories of items were studied (faces, locations & objects). At the end of the experiment, there was a free recall period in which subjects verbally recalled the studied items in any order.

We used a neural network classifier algorithm to isolate characteristic patterns of neural activity relating to studying faces vs. locations vs. objects. Then, we used the trained classifier to track, on a second-to-second basis, how well subjects were able to reinstate these patterns of neural activity at test.

We show that our neural measure of "contextual reinstatement" is highly predictive of when subjects will successfully retrieve studied items. At a more general level, we also discuss how classifiers – neural network and otherwise – can be used to isolate patterns of neural activity and track retrieval dynamics in fMRI experiments.

The paradigm

The basic memory experiment is repeated three times. Subjects study a list of 30 items, and then are asked to recall these items. After the three study / recall blocks there is a final free recall block, during which subjects recall all 90 items learned during the experiment in any order.

The study period. The 30 items are drawn from three categories: famous faces, famous locations, and common objects. Each study list is composed of 10 items from each category. Each study item appears for 4.5 sec, followed by a 2.7 sec period in which a judgment is made on the item. An arithmetic task (lasting about 10 sec) follows each item.

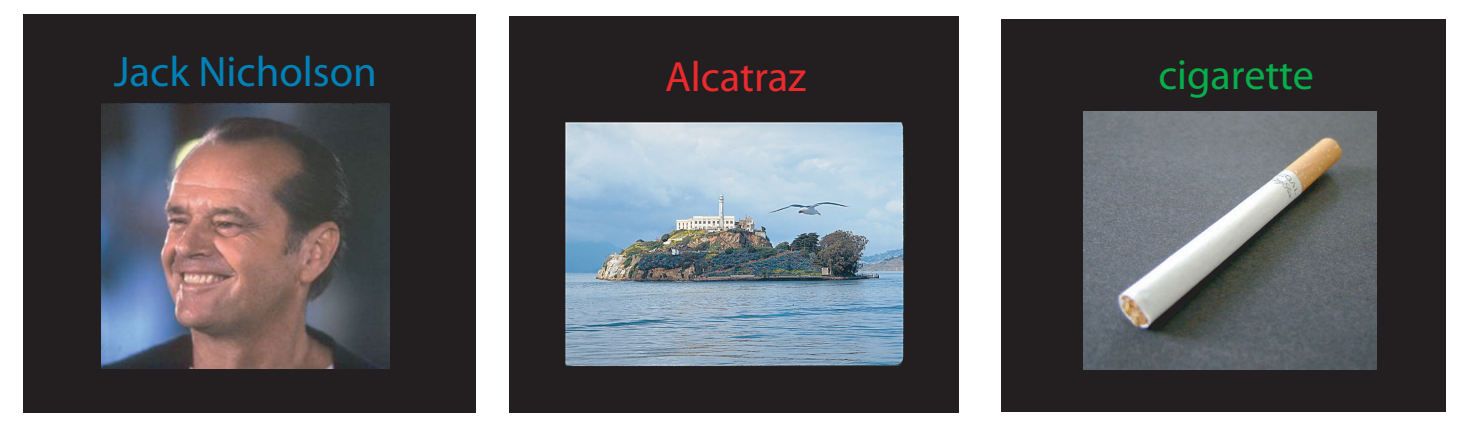


Figure 1. Example items from each of the categories.

The recall period. Three recall periods follow each study list. Subjects are asked to verbally recall items by category from the current list. Subjects have a minute to recall words from each category.

Final free recall. During the final free recall period, subjects are asked to recall as many items as they can from the entire experiment, in any order. The final free recall period lasts about three minutes.

Imaging methods

Five subjects were run in the current experiment.

Data was acquired from a Siemens Allegra 3T scanner: TR = 1.8; TE = 30ms; 30 oblique slices (whole brain coverage); 3mm by 3mm by 4mm.

AFNI preprocessing. The data was motion corrected and despiked. Linear and quadratic trends were removed. 4mm spatial smoothing was applied. A whole-brain mask was created to select the voxels that coincided with brain tissue. *Matlab preprocessing.* A z-score normalization was applied to the individual voxel timecourses by run.

Feature selection. An ANOVA was used to select the voxels whose signal showed significant variation ($p>0.05$) across categories. The number of voxels by subject that passed the feature selection process: 6482, 10008, 9867, 7809, 8311.

Classification methods

Details of the classifier. We used a neural network classifier trained with the backpropagation algorithm. The classifier is given patterns from the study phase and is trained to discriminate between the face study, location study, and object study conditions. The input layer of each network has one unit for each voxel that passes the feature selection process described above. The output layer consists of three units, one for each of the study categories. Each output unit receives a weight from every input unit.

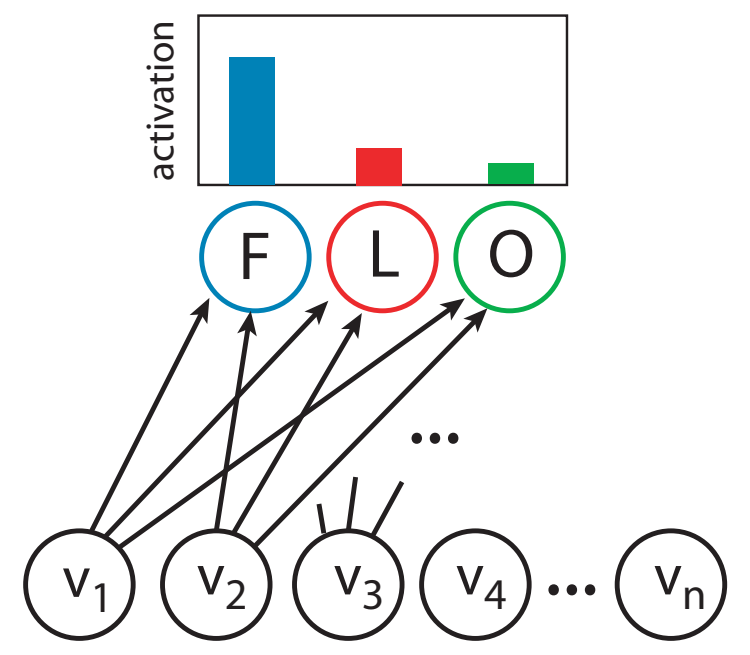


Figure 2. A schematic of the neural network classifier. Each input unit (v1 to vn) corresponds to a voxel. Each output unit corresponds to a study condition.

To increase reliability of classifier output, we re-ran the classifier 20 times with different starting weights. The three output values of the classifier are then normalized to sum to one using the Luce Choice Ratio.

Applying the classifier to data from the final recall period. We use the trained classifier to read out, second-by-second, how well neural activity during the final free recall period resembles the face study, location study, and object study conditions. Each of the output units of the classifier returns a scalar value at each timepoint indicating the degree of study / test match.

Interpreting the weight structure of the network. Given a trained network, it is possible to read out (based on the network's weight structure) which voxels are important in representing each category. We obtain a measure of 'path strength' by multiplying the average weight from a given voxel (over the 20 networks) by the average activity value of that voxel over the training set. The sign of the path strength says whether this voxel turns a given output unit on or off. See Polyn et al (2004) for details of this procedure.

Visualizing the dynamics of recall

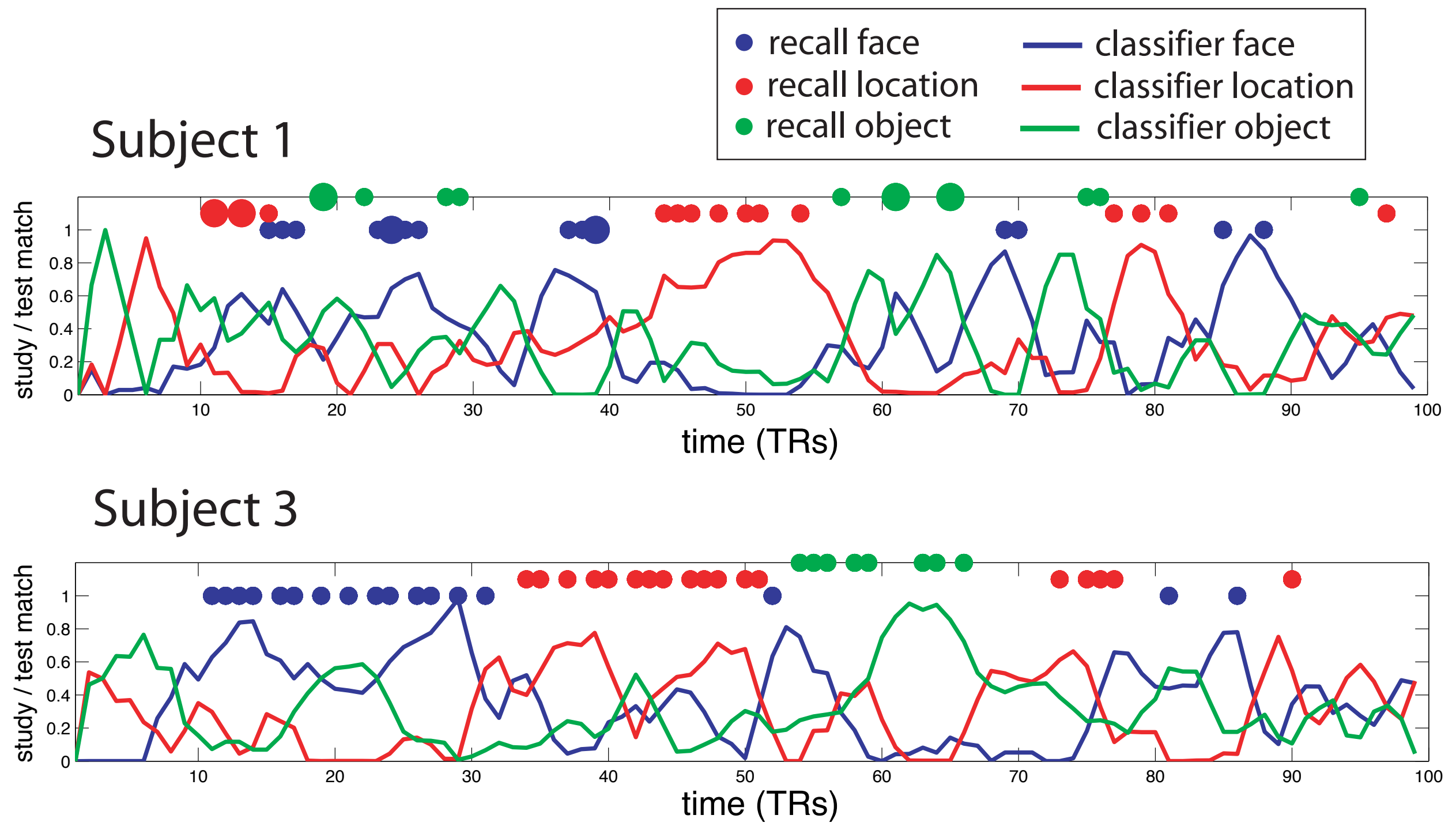


Figure 3. The classifier output during final free recall, for subjects 1 and 3 (See classification methods for details). The colored dots label the timepoints during which recalls were made from each of the categories. A small dot means that one response was made during this time period. A larger dot means that two responses were made. The dots are shifted ahead 3 timepoints to account for hemodynamic lag. The classifier traces have been slightly blurred for visualization purposes. The statistics below are performed on the unblurred traces.

Quantifying the dynamics of recall.

Correlation. We run all the pairwise correlations between the classifier output traces and the recall records to gain a sense of how well the classifier is tracking the neural process underlying recall. These correlations are performed on the +3 shifted recall records.

Subject 1

recall	classifier		
	—	—	—
●	.4029	-.1224	-.2735
●	-.1728	.2543	-.0827
●	.0639	-.2022	.1378

Subject 3

recall	classifier		
	—	—	—
●	.3721	-.2444	-.1643
●	-.1712	.3442	-.1856
●	-.0864	-.0952	.2094

Over all subjects

recall	classifier		
	—	—	—
●	.3553	-.1204	-.2458
●	-.0904	.1021	-.0103
●	-.0467	-.1007	.1560

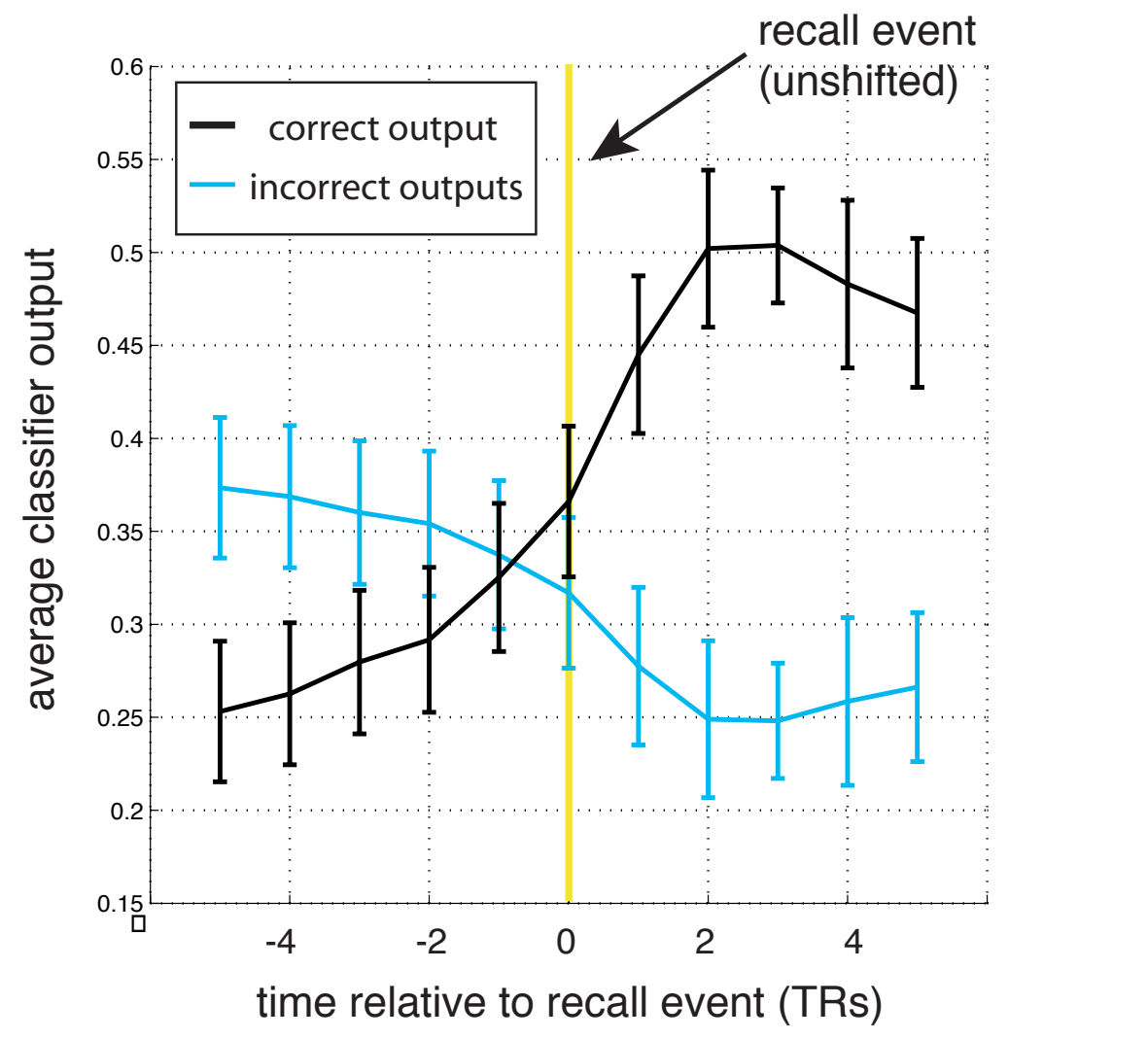
Prediction. Given a recall event, the classifier is used to predict which category the recall came from. This result is performed on the +3 shifted recall records. Chance = 33%.

Subject 1 percent correct: **73%**

Subject 3 percent correct: **65%**

All subjects percent correct: **60%**
min: **48%** max: **73%**

Reinstated activity precedes recall



Does classifier activity reflect cue construction, retrieved information, or both? As a first pass at addressing this question, we created an event-related average of classifier output, centered on the time when recall occurred. We chose recall events without same-category recalls in the prior 5 TRs (9 seconds) to ensure that the classifier output was uncontaminated by recall-related activity.

The correct classifier unit ramps up activity before the recall occurs. On average, recall occurs once the correct unit's activity exceeds the average activity of the incorrect units.

Figure 4. The traces are an average of classifier activity surrounding 54 recall events across all 5 subjects. The black line is the average value of the correct output unit, while the cyan line is the average value of the two incorrect output units. Error bars are standard error of the mean.

Extracting brain maps from the classifier

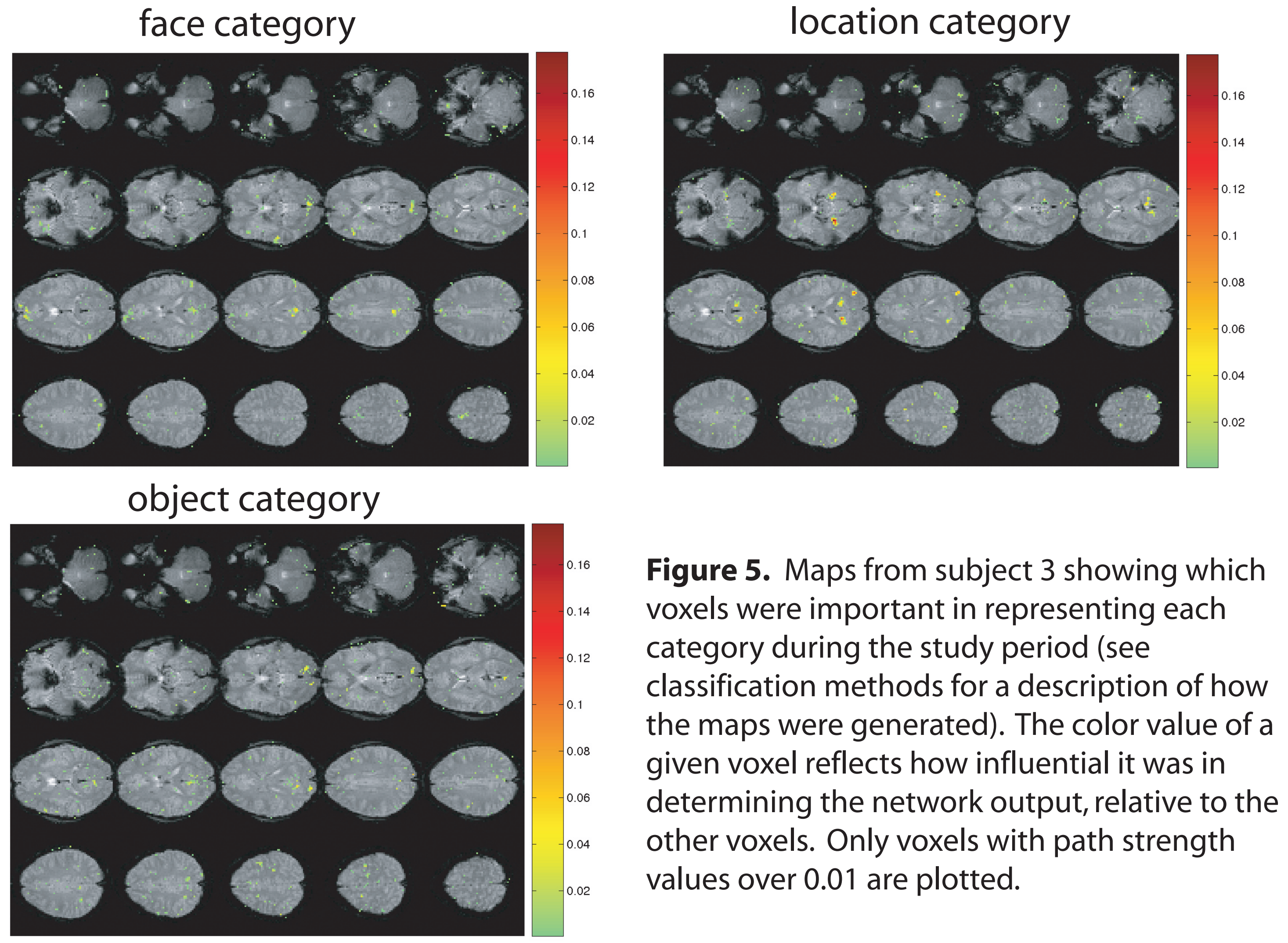


Figure 5. Maps from subject 3 showing which voxels were important in representing each category during the study period (see classification methods for a description of how the maps were generated). The color value of a given voxel reflects how influential it was in determining the network output, relative to the other voxels. Only voxels with path strength values over 0.01 are plotted.

Investigating frontal contributions

What can we learn about the role of prefrontal cortex (PFC) in episodic memory from this experiment?

As a first pass, we retrained the classifier on a restricted set of voxels that excluded most of the posterior regions of the brain (literally the front half of the brain; future analyses will use an ROI derived from anatomy). Subjects 1 & 3 still showed good percent correct prediction of recalled category (51% and 67% respectively). The other subjects did not fare as well (percent correct over all subjects was 46%).

How can we reconcile these results with other fMRI findings showing a strong role for PFC in episodic memory retrieval? One possibility is that - in this paradigm - PFC is more involved in specifying general recall strategies (e.g., "recall by category") than in specifying which category should be recalled. Also, looking at the brain maps above, it appears that posterior areas do a better job than PFC at distinguishing between the three conditions at study. If we used encoding conditions that were associated with distinct patterns of PFC activity at study (e.g. different tasks), then PFC might play a stronger role in targeting specific conditions at retrieval.

Conclusions

- By applying pattern classification techniques to fMRI data, we were able to visualize the process of contextual reinstatement during free recall, and show how this relates to behavioral data.

- More generally, this approach affords us a view into the "black box" of how subjects construct memory cues: What information is contained in the cues, and where is this information represented?

- By running memory retrieval experiments with more subtle categories (e.g. words studied with different encoding tasks), we hope to further our understanding of the role of prefrontal cortex in memory targeting and its interaction with posterior areas.

This research partially supported by an NIMH NRSA fellowship to SMP: MH070177-02
Special thanks to Vaidehi S Natu for assistance with the data analysis.
For an electronic copy of this poster (and other related reprints) visit:
http://compmem.princeton.edu/publications.html