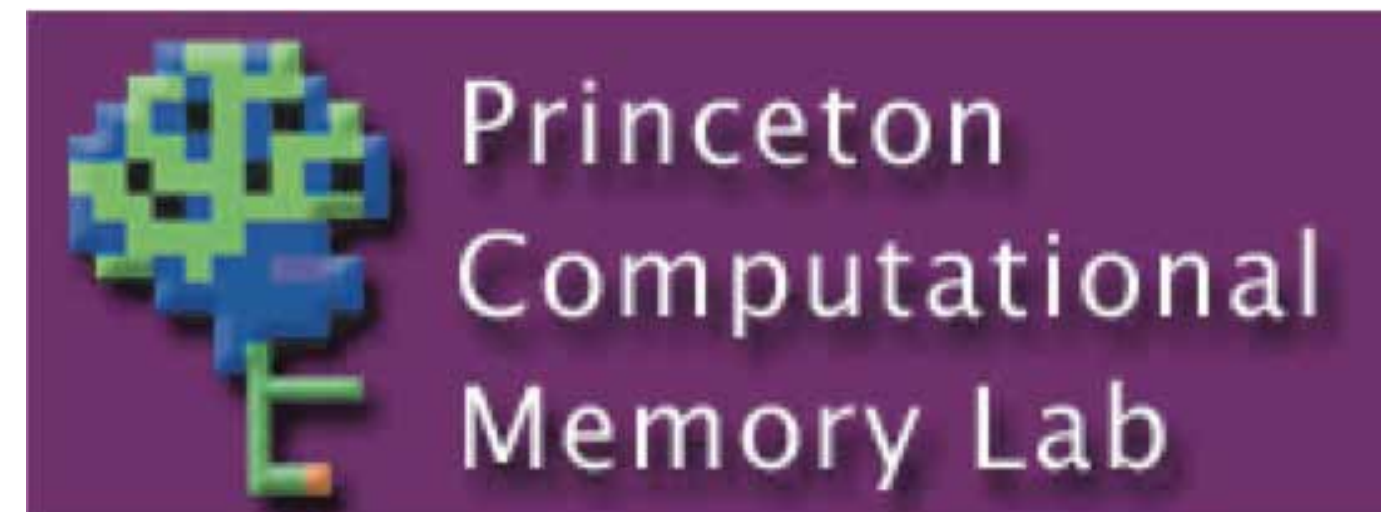# Deactivation of Items in Working Memory Can Weaken Long-Term Memory
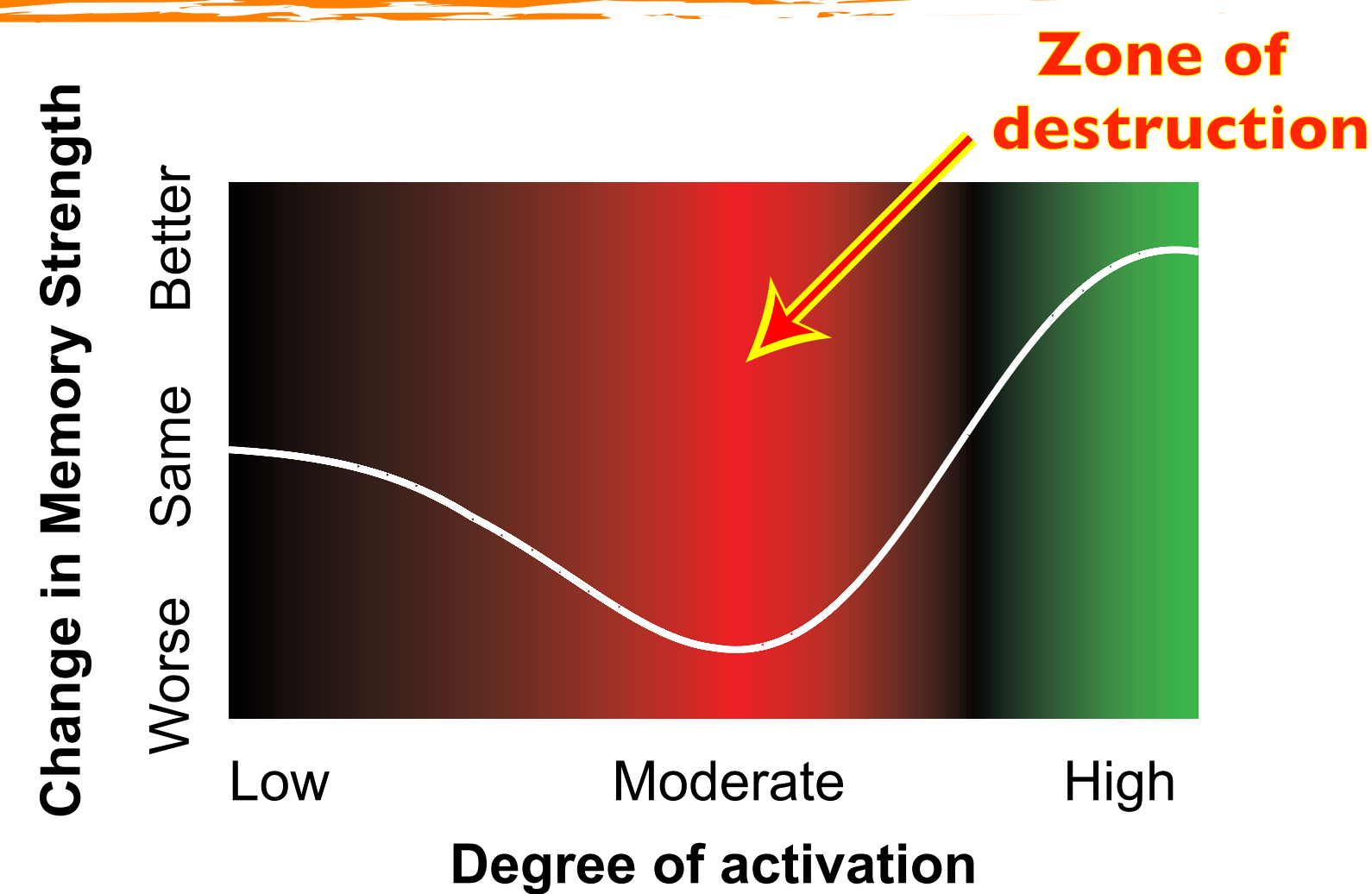
Jarrod A. Lewis-Peacock and Kenneth A. Norman
Princeton Neuroscience Institute and Department of Psychology, Princeton University

Princeton Neuroscience INSTITUTE

Princeton Computational Memory Lab

1 Introduction: Nonmonotonic Plasticity

Zone of destruction

Hypothesis: the relationship between memory activation and learning is **nonmonotonic**

If a memory is strongly activated, it gets **strengthened**
If a memory has very low activation (or none at all), nothing happens
If a memory activates to a moderate degree, it gets **weakened**

This nonmonotonic relationship is predicted by computational models of learning (e.g., Bienenstock, Cooper, & Munro, 1982; Norman, Newman, Detre, & Polyn, 2006)

The nonmonotonic pattern has been found at the synaptic level (post-synaptic potential: Artola et al., 1990; post-synaptic Ca2+ concentration: Hansel et al., 1996)

We want to see if this pattern occurs at the level of **memory representations**

## 2 Prior Work: Newman & Norman (2010)

Newman & Norman set out to test the nonmonotonic plasticity hypothesis using a negative priming paradigm

Negative priming effect: Ignoring a distracting stimulus makes you slower to respond to it later (Tipper, 1985)

According to the nonmonotonic plasticity hypothesis, moderate activation of the distractor should weaken the distractor, leading to negative priming

Approach: Use pattern classifiers to track activity of the distractor. Relate this neural measure of distractor activity to priming effects

Results (shown at right) fit with the nonmonotonic plasticity hypothesis

* $p < 0.05$
** $p < 0.01$
Distractor processing during prime display
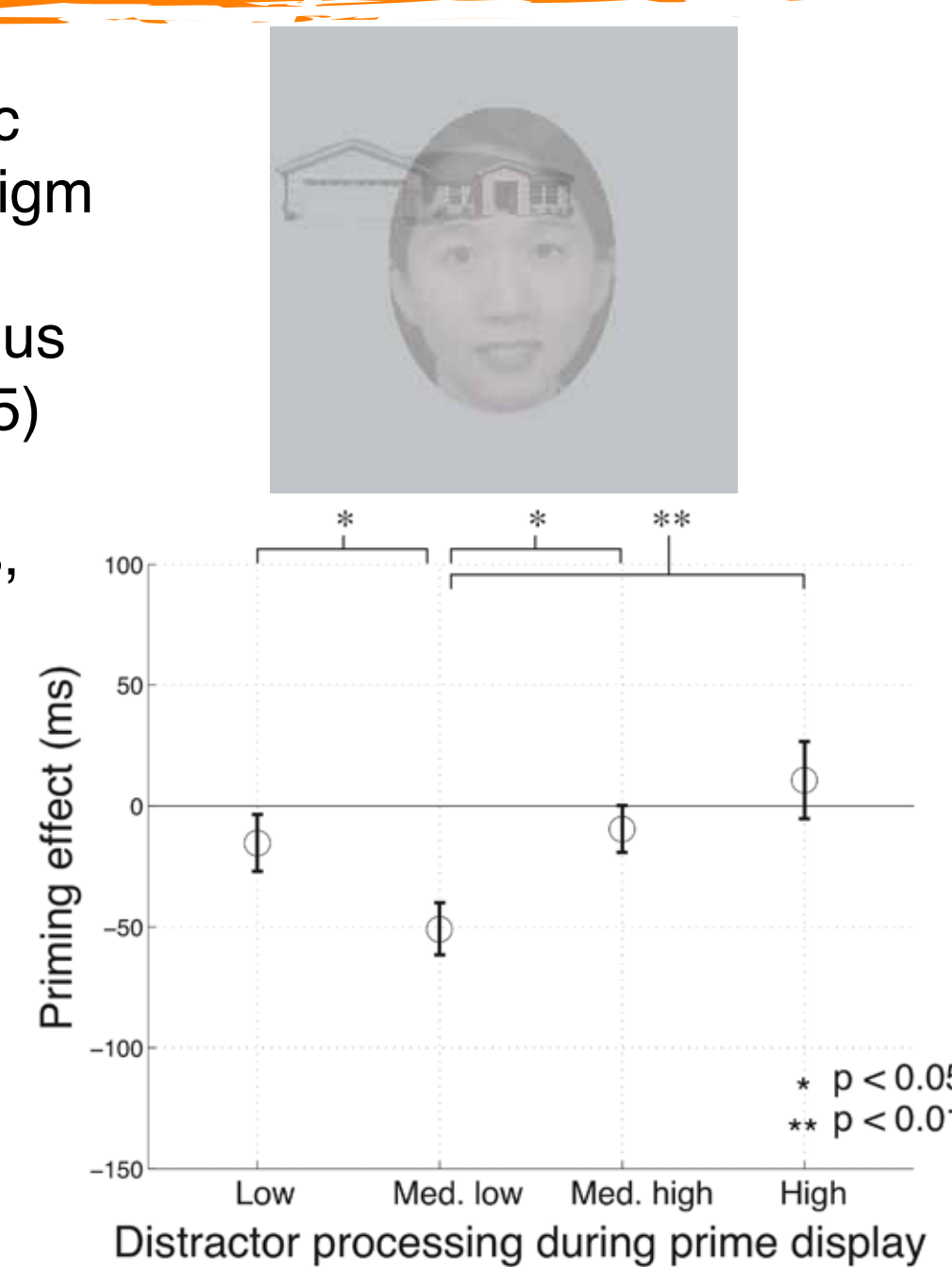
3 Background: Memory Deactivation

You are faster to respond to a memory probe if given enough time (~1 sec per item) to remove your attention from a subset of irrelevant items in working memory (Oberauer, 2001)

Lewis-Peacock, Drysdale, Oberauer & Postle (2011) showed that removing attention from an item in working memory results in the neural deactivation of that item

1st cued / 2nd cued / baseline
$p < 0.002$
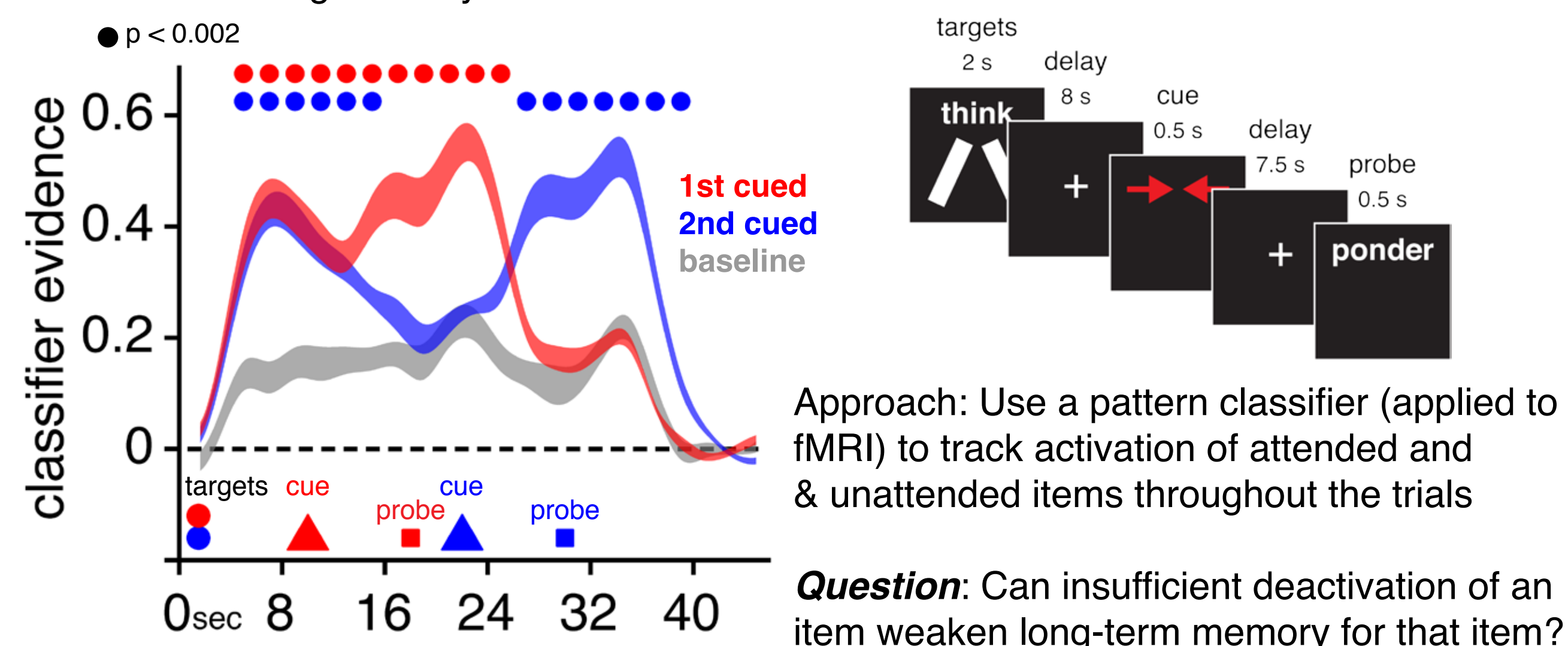
Approach: Use a pattern classifier (applied to fMRI) to track activation of attended and & unattended items throughout the trials
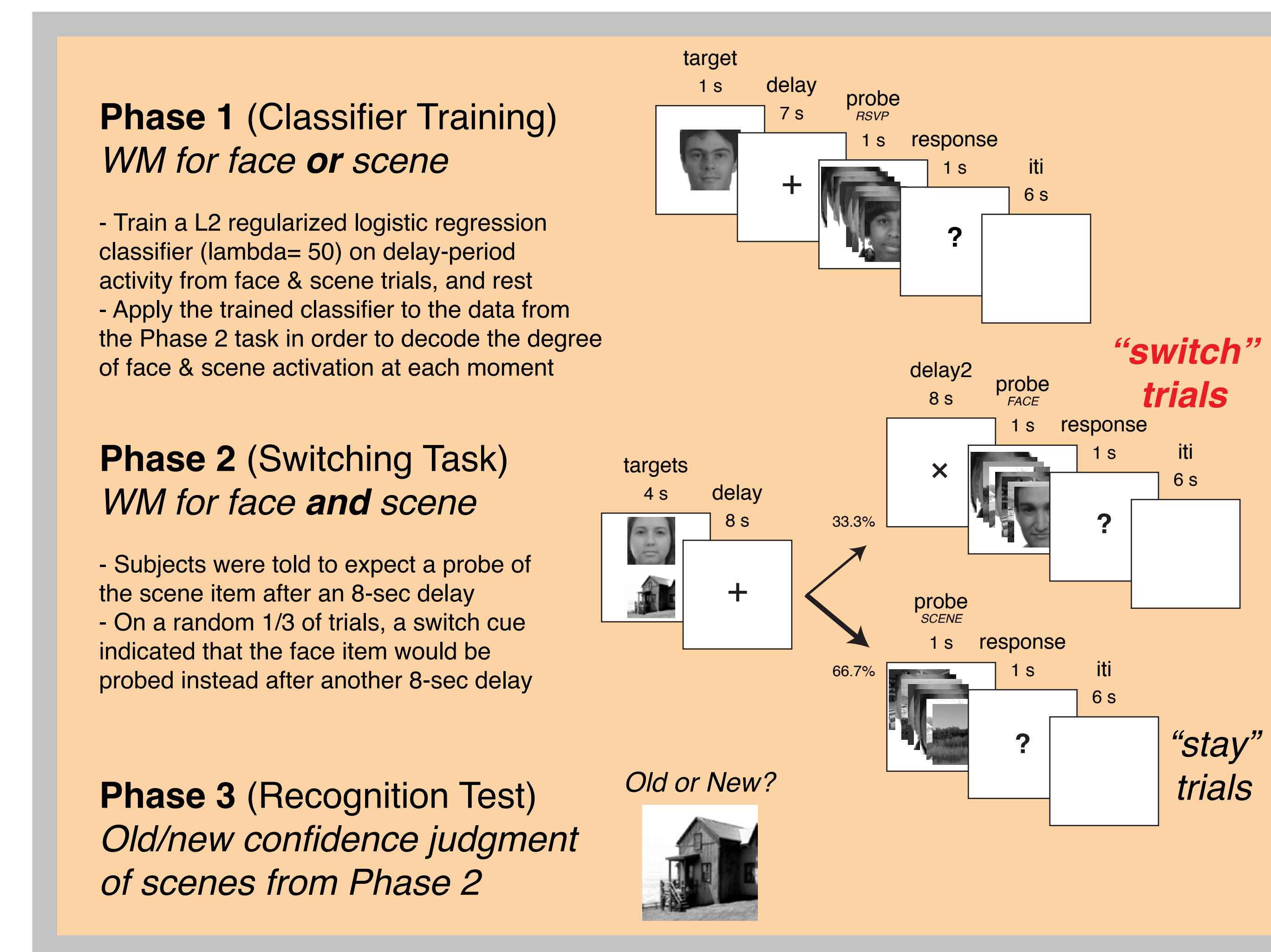
Question: Can insufficient deactivation of an item weaken long-term memory for that item?

4 Hypothesis and Experimental Approach

Key prediction: Forgetting will happen when an item gets "stuck" in the **moderate activity** range while it is being deactivated from working memory

Strategy for testing this prediction: Use fMRI pattern classifiers to read out the deactivation of an item following the attention cue in a working memory switching task

Use this neural measure to predict recall of the item on a final memory test

Phase 1 (Classifier Training)
WM for face or scene
- Train a L2 regularized logistic regression classifier (lambda= 50) on delay-period activity from face & scene trials, and rest
- Apply the trained classifier to the data from the Phase 2 task in order to decode the degree of face & scene activation at each moment

Phase 2 (Switching Task)
WM for face and scene
- Subjects were told to expect a probe of the scene item after an 8-sec delay
- On a random 1/3 of trials, a switch cue indicated that the face item would be probed instead after another 8-sec delay
"switch" trials (33.3%)
"stay" trials (66.7%)

Phase 3 (Recognition Test)
Old/new confidence judgment of scenes from Phase 2
Old or New?

Note: Stimuli were selected based on moderate memorability ratings, as assessed by a stimulus evaluation experiment conducted through Amazon.com's Mechanical Turk

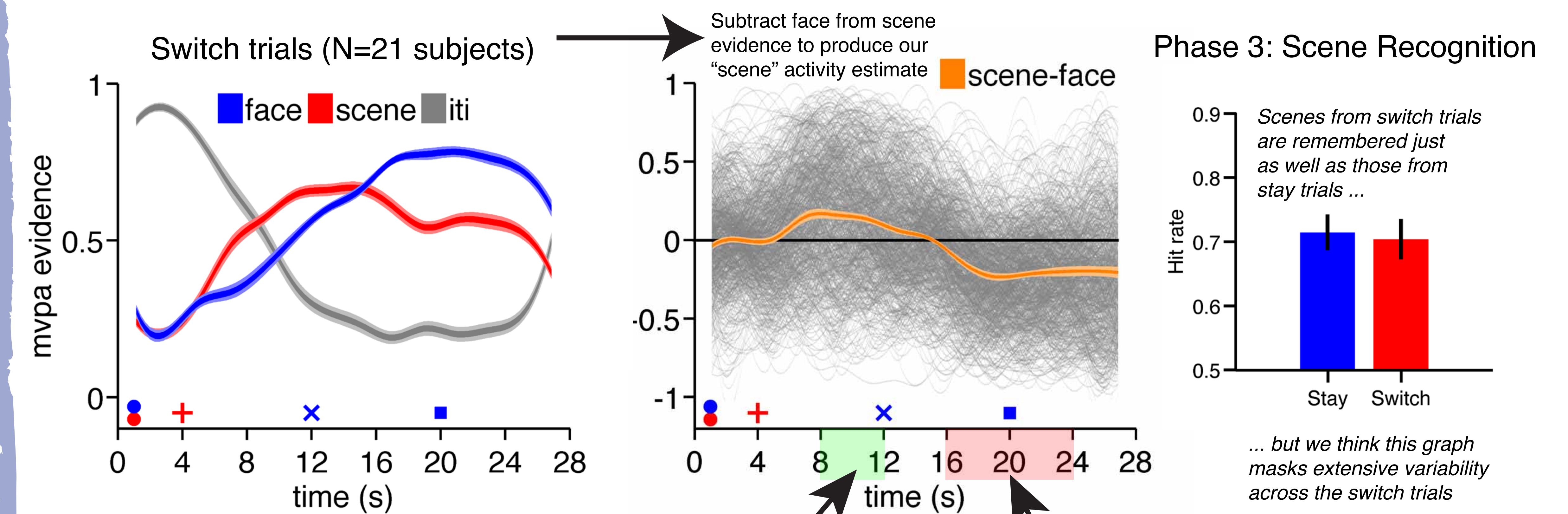Predictions from nonmonotonic plasticity hypothesis:
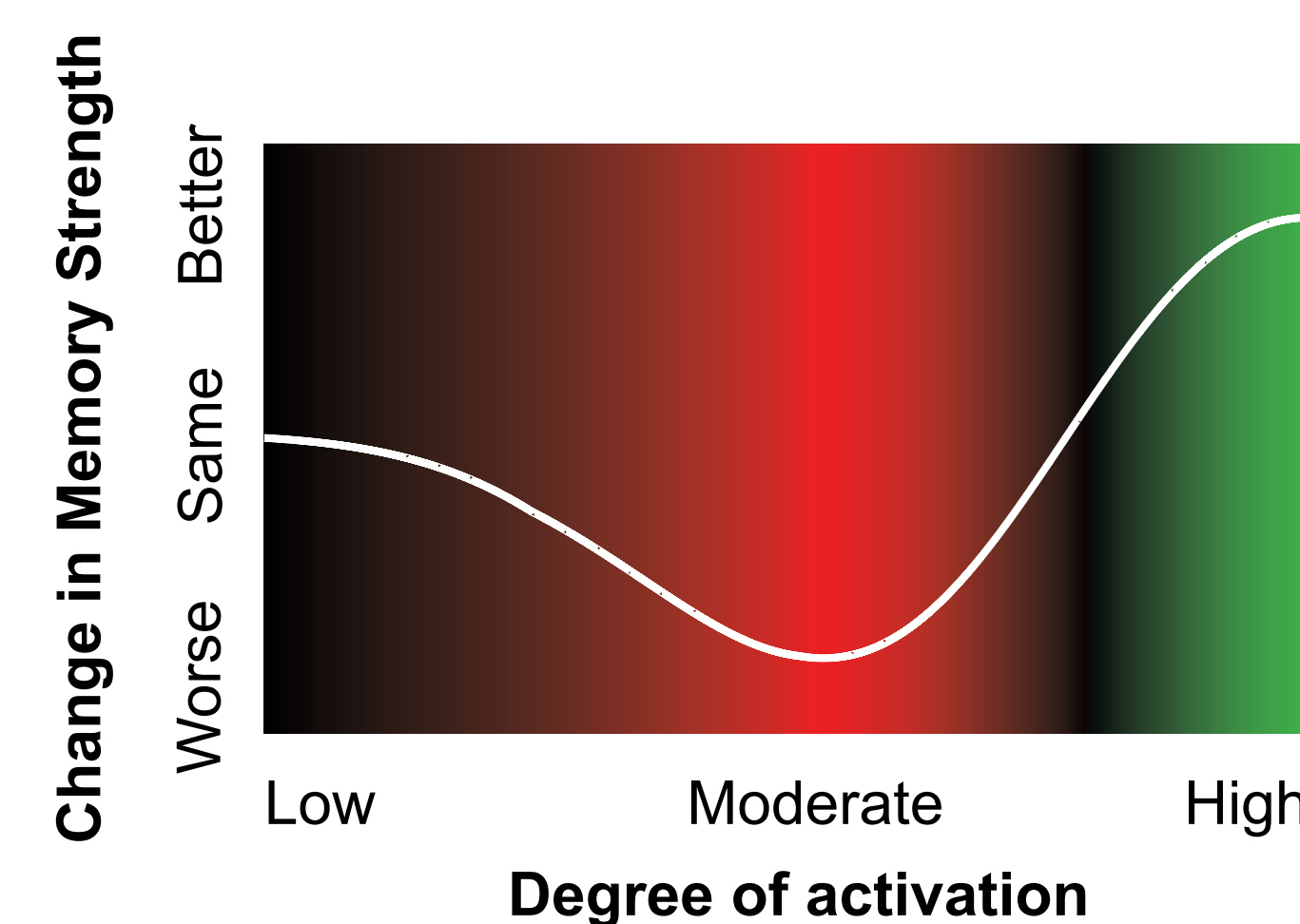
If the scene memory is deactivated efficiently, resulting in **low scene activity**, the memory will be mostly unharmed

If the scene memory gets "stuck" in the **moderate activity** range while it is being deactivated, this will lead to **weakening** of the memory, and (consequently) poor recognition memory for that item on the final test

Analysis strategy:
1. Use a pattern classifier (applied to fMRI) to track scene and face activity throughout the switch trials
2. Measure how efficiently participants deactivate the scene on switch trials -- how much scene activity is there after the switch cue?
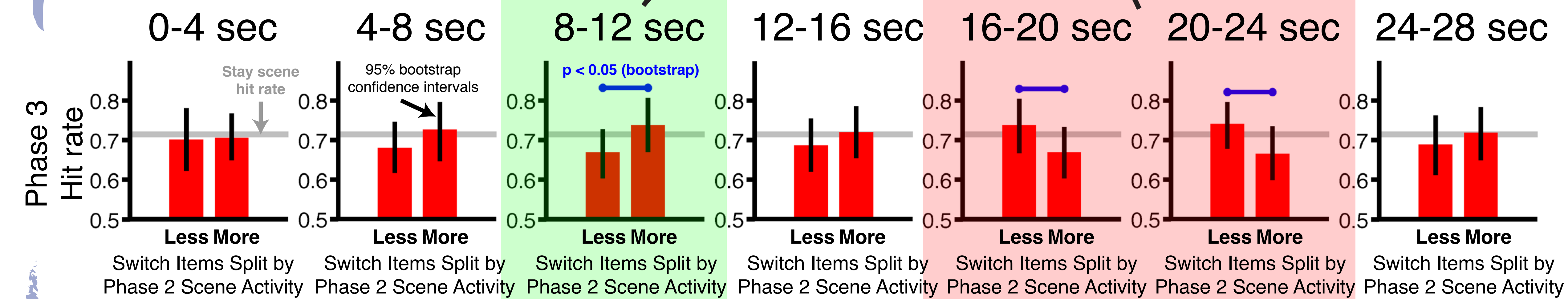3. Relate this **residual scene activity** to subsequent memory for that scene (in the Phase 3 recognition test)

5 Measures of *Scene* Deactivation (in Phase2) and *Scene* Memory (in Phase 3)

Switch trials (N=21 subjects)
face / scene / iti

Subtract face from scene evidence to produce our "scene" activity estimate
scene-face

Phase 3: Scene Recognition
Scenes from switch trials are remembered just as well as those from stay trials ...
Stay / Switch
... but we think this graph masks extensive variability across the switch trials

6 Deactivation Predicts Memory: Post-Switch *Scene* Activity Leads to Forgetting

Analysis: Pool together switch trial activation and memory data from all subjects and do median split of Phase 3 hit rates based on Phase 2 scene activity (at 4-sec time windows)

Bootstrapped the analysis to convert it from a fixed effect into a random effect. Sampled N=21 (with replacement) on 1,000 iterations

0-4 sec | 4-8 sec | 8-12 sec | 12-16 sec | 16-20 sec | 20-24 sec | 24-28 sec
Stay scene hit rate
95% bootstrap confidence intervals
$p < 0.05$ (bootstrap)
Less More — Switch Items Split by Phase 2 Scene Activity

More scene activity before the switch predicts **better** memory

More scene activity after the switch predicts **worse** memory!

Counterintuitive result, but makes sense under the nonmonotonic plasticity hypothesis

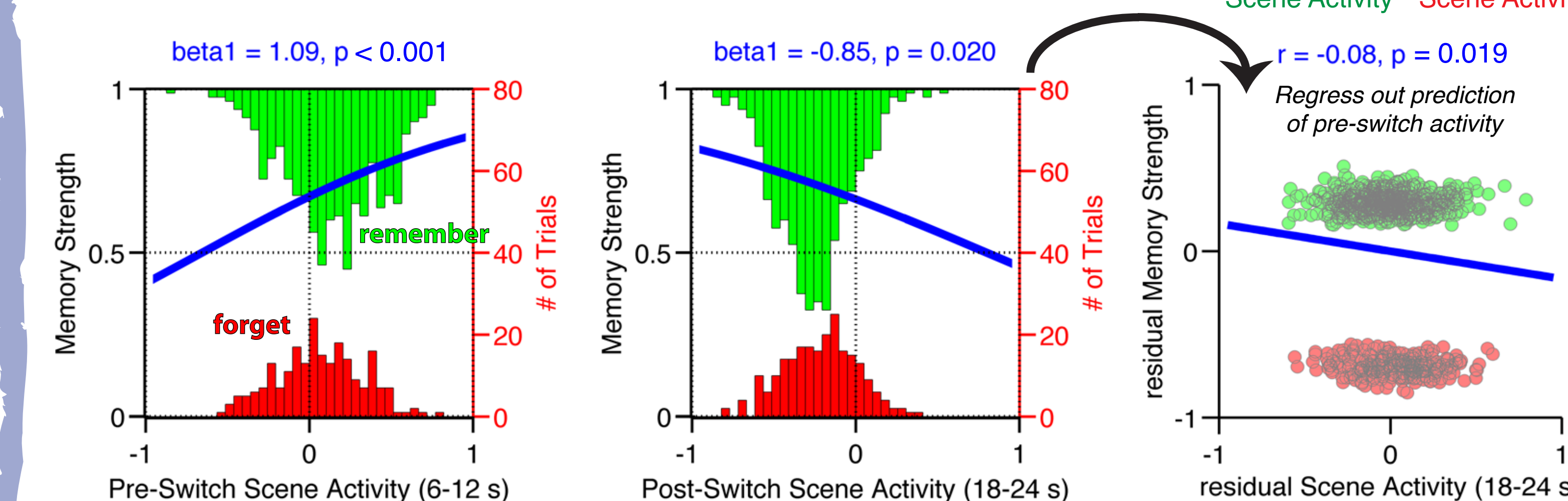## 7 Independent Predictors: Pre- & Post-Switch Activity

Independent?: Post-switch scene activity predicts subsequent memory of that scene independently of its pre-switch activity

Pre-Switch Scene Activity → Scene Memory
Post-Switch Scene Activity → Scene Memory

Mediated?: Cognitive control is a latent variable that mediates the effects of scene activity on subsequent memory

Cognitive Control → Scene Memory
Pre-Switch Scene Activity / Post-Switch Scene Activity

beta1 = 1.09, $p < 0.001$
remember / forget
Pre-Switch Scene Activity (6-12 s)

beta1 = -0.85, $p = 0.020$
Post-Switch Scene Activity (18-24 s)

$r = -0.08$, $p = 0.019$
Regress out prediction of pre-switch activity
residual Scene Activity (18-24 s)

8 Conclusions

1. More activation of the scene **after** the switch cue was associated with subsequent forgetting of that scene

2. Pre- and post-switch scene activity were **independent** predictors of subsequent memory for scenes

3. This finding converges with others from our lab (Think-no think, Negative priming) to suggest that nonmonotonic plasticity is a **general principle** that applies across multiple domains