

PRINCETON COMPUTATIONAL MEMORY LAB

How do retrieval dynamics drive learning? Insights from fMRI and computational models

Ken Norman

Department of Psychology and
Princeton Neuroscience Institute

Princeton University

March 17, 2025



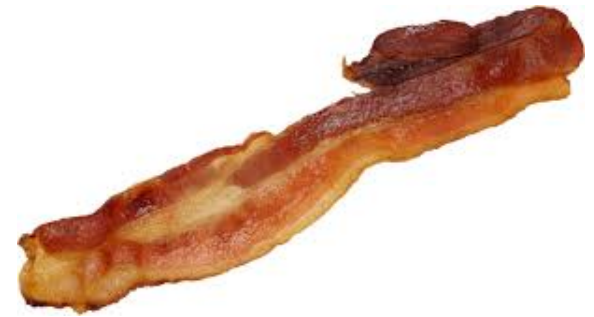
Mental representations compete to
become active



duck? rabbit?



What did you have for breakfast today?

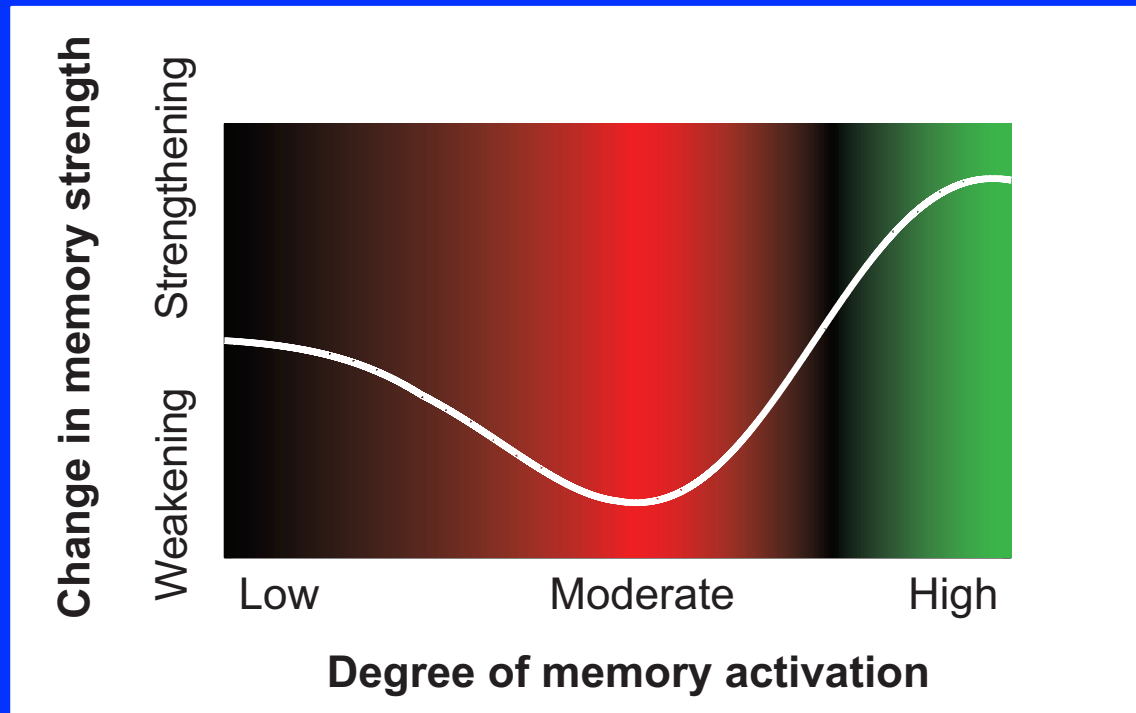


Reducing Competition Through Learning

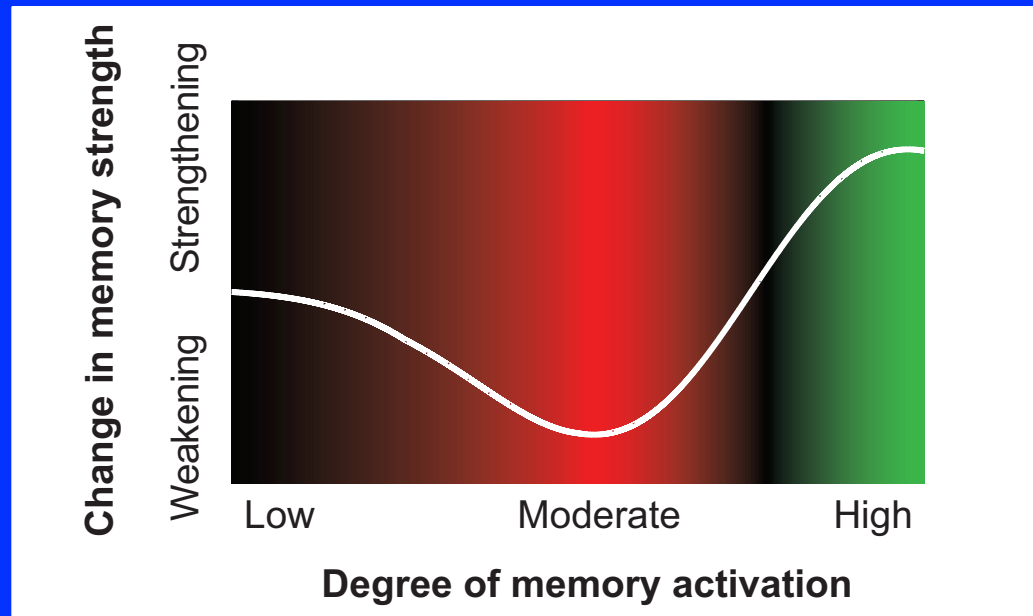
- To function properly, we need to be able to **quickly** and **reliably** access relevant knowledge
- I will describe a **simple, biologically-grounded** learning principle that optimizes memory retrieval by **detecting** and then **reducing** competition

Reducing Competition Through Learning

Nonmonotonic Plasticity Hypothesis (NMPH)

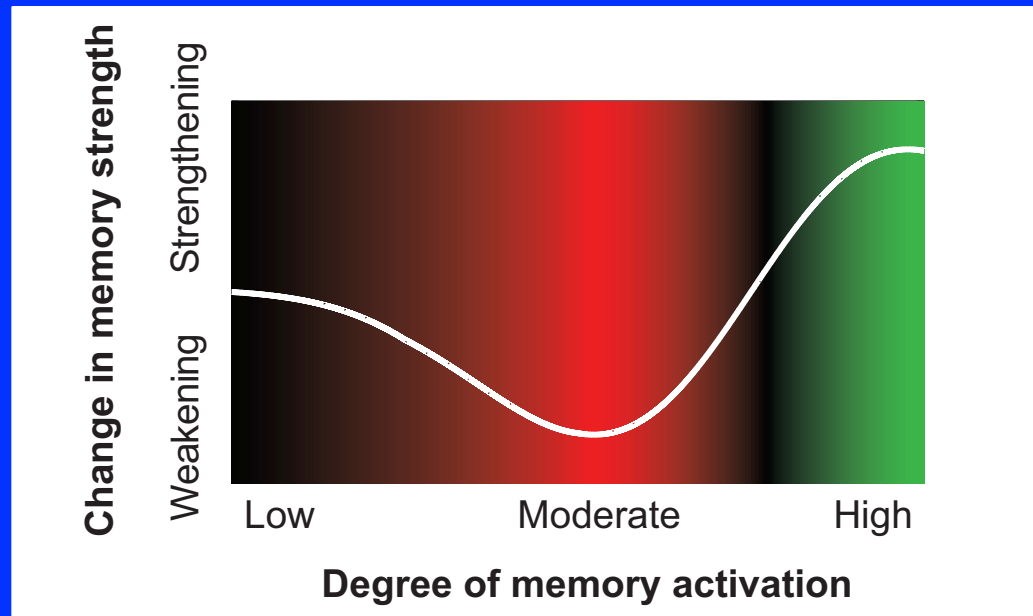


Evidence for Nonmonotonic Plasticity

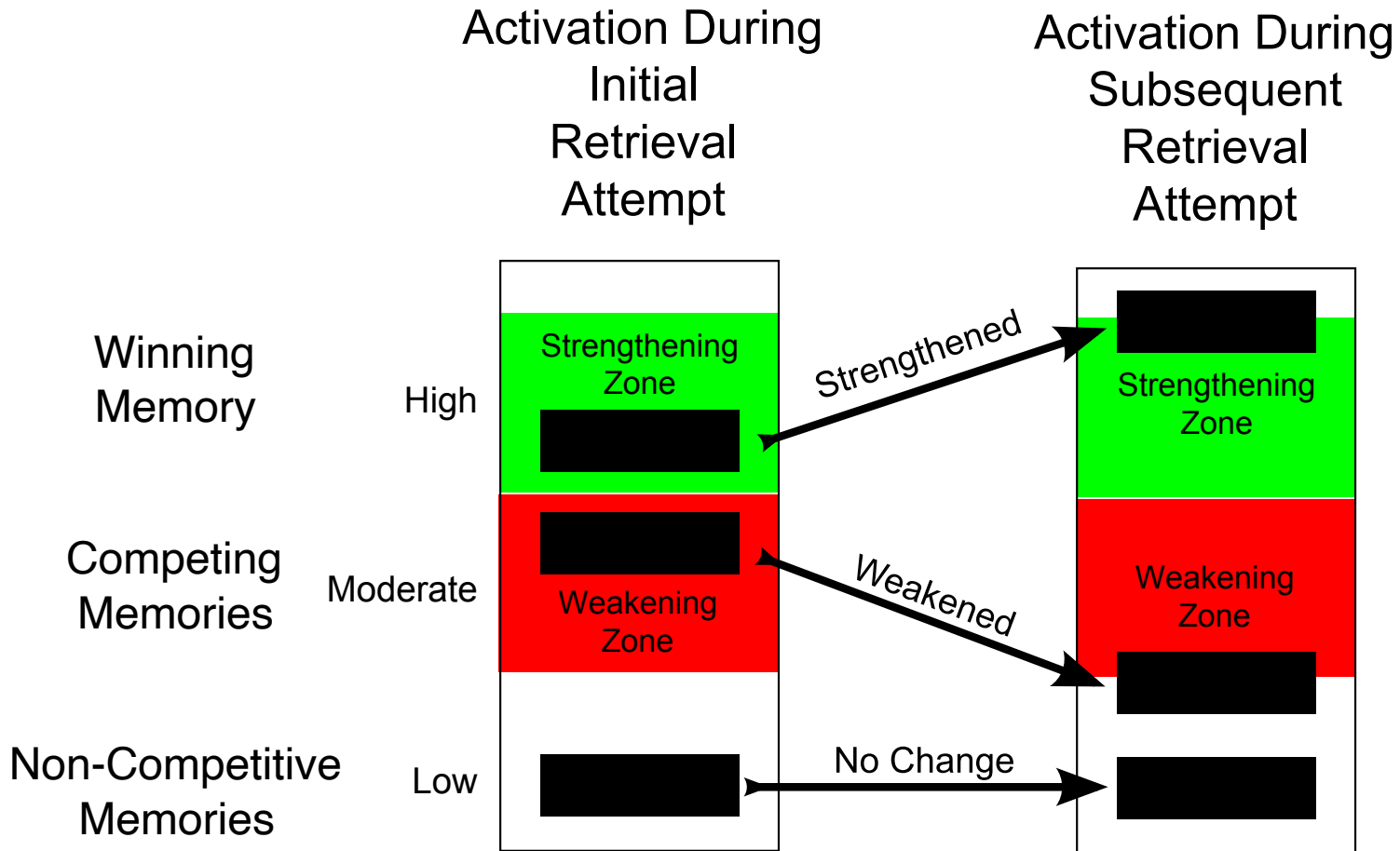


- Neural evidence:
 - At the synaptic level, moderate excitation of the postsynaptic neuron leads to synaptic weakening (LTD)
 - Higher levels of excitation lead to synaptic strengthening (LTP; e.g., Artola et al., 1990)

Evidence for Nonmonotonic Plasticity



- Computational modeling work in my lab suggests that this synaptic-level principle should “scale up” to the level of **cognitive representations** (Norman et al., 2006, *Neural Computation*; Norman et al., 2007, *Psychological Review*)
- Long tradition of neural network modeling work using learning rules with this form (e.g., Bienenstock, Cooper, & Munro, 1982)



Outline

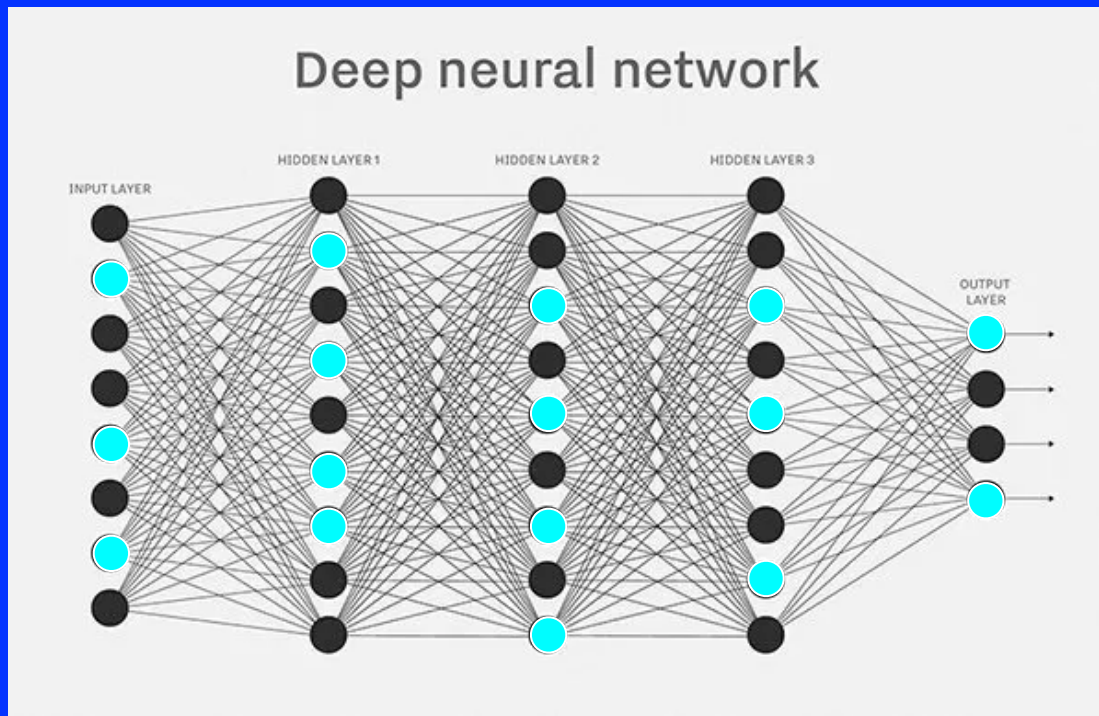
- Situate NMPH relative to other kinds of learning
- Implications of NMPH for **memory weakening**
 - Key prediction: Moderate activation leads to weakening of competing memories
- Implications of NMPH for the **similarity structure** of memories
 - Key prediction: Moderate activation leads to differentiation of competing memories
- Current directions
 - Role of NMPH in learning during sleep
 - Using neurofeedback to promote discrimination learning

Outline

- **Situate NMPH relative to other kinds of learning**
- Implications of NMPH for **memory weakening**
 - Key prediction: Moderate activation leads to weakening of competing memories
- Implications of NMPH for the **similarity structure** of memories
 - Key prediction: Moderate activation leads to differentiation of competing memories
- Current directions
 - Role of NMPH in learning during sleep
 - Using neurofeedback to promote discrimination learning

High-Level Overview of Neural Learning Rules

- Key goal of learning: build an **internal model** of the world that allows you to make accurate inferences / responses given a particular input



squirrel-relevant
predictions (climb
trees, eat nuts, etc)

“this is a squirrel”

High-Level Overview of Neural Learning Rules

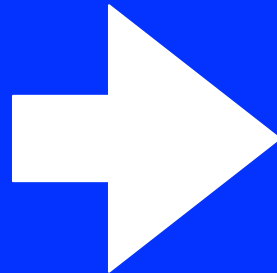
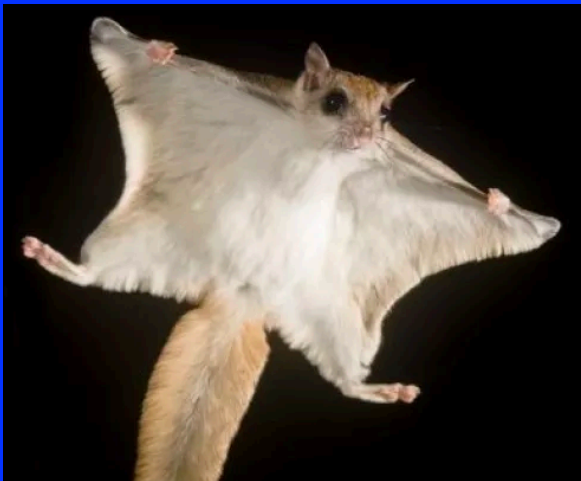
- Researchers have come up with several ideas about how we learn this internal model
 - **Supervised**: compare your guess to the right answer
 - **Unsupervised**: where you don't have access to the right answer

Supervised Learning

- Make a guess; compare to outcome; adjust weights to minimize discrepancy between guess and outcome
- Workhorse of deep learning (e.g., backpropagation)
- Extremely powerful — the magic of LLMs depends on this, as do advances in computer vision
- There's widespread consensus that the brain needs to be doing **some** kind of supervised learning
- Lots of recent progress in thinking about how this can be implemented in the brain (e.g., Richards et al., 2019; Lillicrap et al., 2020; Whittington & Bogacz, 2109)

Unsupervised Learning

- Supervised learning requires us knowing the correct answer, which we can compare to our guess
- Importantly, there are many useful kinds of learning that we can do, even when we don't know the correct answer



IF two things are known to be the **same** (in some sense)
THEN push their internal representations together

Unsupervised Learning

- This kind of idea...
 - pushing together representations of the “same” things
 - pulling apart representations of “different” things
- ... has become very popular in computer vision

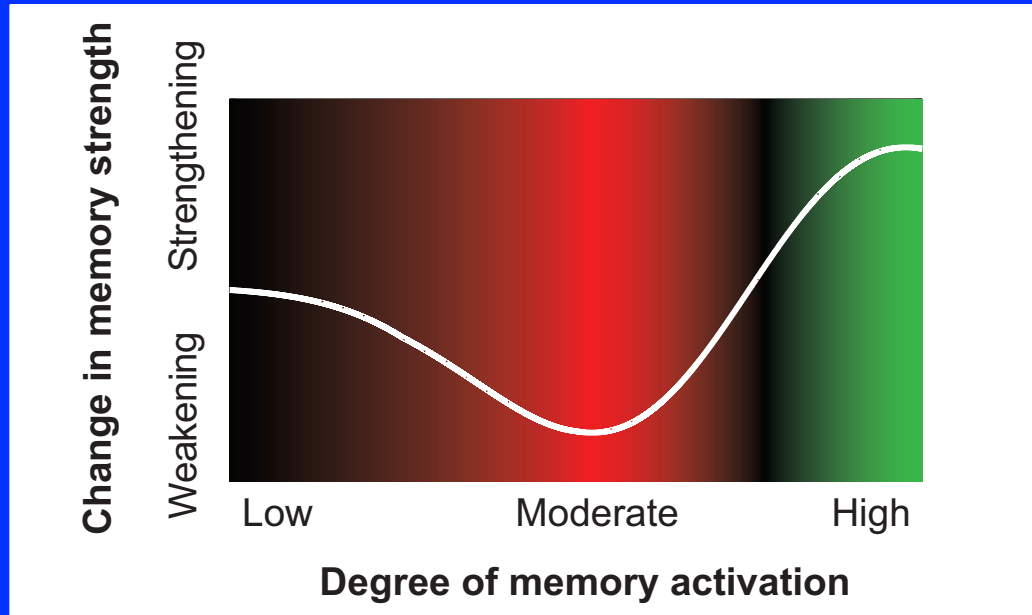
Unsupervised Learning

Unsupervised neural network models of the ventral visual stream

Chengxu Zhuang^{a,1}, Siming Yan^b, Aran Nayebi^c, Martin Schrimpf^d, Michael C. Frank^a, James J. DiCarlo^d, and Daniel L. K. Yamins^{a,e,f}

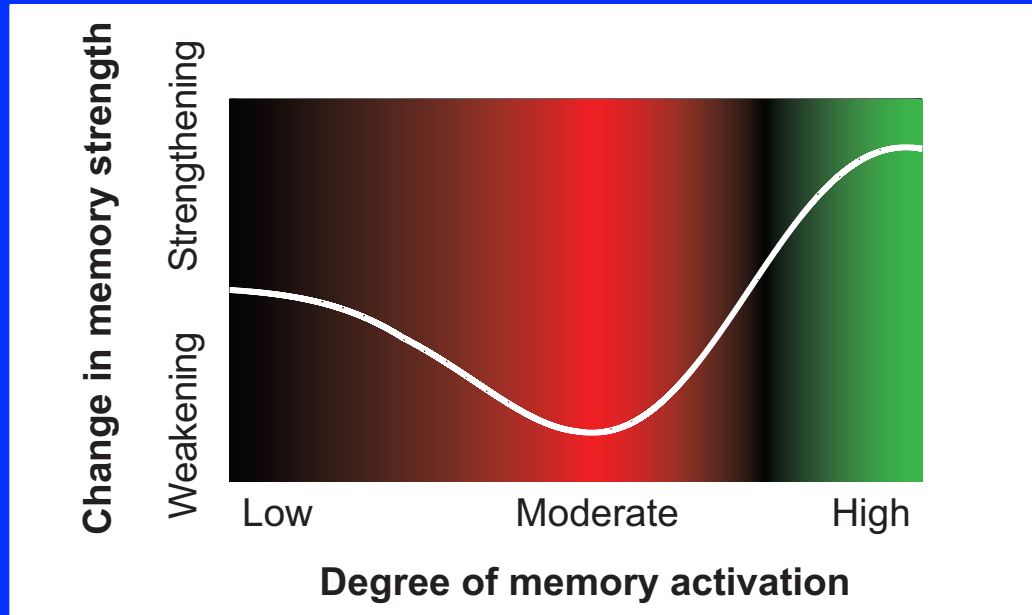
- Networks imbued with this kind of learning (in addition to supervised learning):
 - Perform better on object recognition benchmarks
 - Develop visual representations that map more closely onto neural data
- While this supports the idea that the brain is doing **something** like this, **how exactly** the brain does this is a topic of active investigation

Back to the NMPH



- Definitely a form of unsupervised learning — no need to specify the “right answer”
- Simpler than the form of unsupervised learning I just discussed — no need to mark inputs as “same” or “different”
- Clear biological basis

Back to the NMPH



- Goal of our lab's work: See how much we can explain with this very simple form of plasticity
- Important point: Not mutually exclusive with other ideas about plasticity

Outline

- Situate NMPH relative to other kinds of learning
- **Implications of NMPH for memory weakening**
 - **Key prediction: Moderate activation leads to weakening of competing memories**
- Implications of NMPH for the **similarity structure** of memories
 - Key prediction: Moderate activation leads to differentiation of competing memories
- Current directions
 - Role of NMPH in learning during sleep
 - Using neurofeedback to promote discrimination learning

How do competitive dynamics drive learning?

- When representations compete, the representation that **wins** the competition becomes easier to retrieve later
 - e.g., testing effects (Karpicke & Roediger, 2008)
- What happens to memories that **compete** but **lose** the competition?
- Several lines of research suggest that people can **inhibit** these irrelevant memories in a lasting fashion
- “Inhibit” = something happens to these memories that makes them harder to retrieve in the future

Retrieval-Induced Forgetting (RIF)

(Levy & Anderson, 2002)

STUDY

Fruit – Apple
Fruit – Pear
Animal – Sheep
Animal – Cow



PRACTICE

Fruit – Pe__



TEST

Fruit – A__
Fruit – P__
Animal – S__
Animal – C__

Retrieval-Induced Forgetting (RIF)

(Levy & Anderson, 2002)

STUDY

Fruit – Apple
Fruit – Pear
Animal – Sheep
Animal – Cow



PRACTICE

Fruit – Pe__



TEST

Fruit – A__
Fruit – P__
Animal – S__
Animal – C__



- Retrieval practice helps the practiced item (Pear) & hurts non-practiced items from the practiced category (Apple)

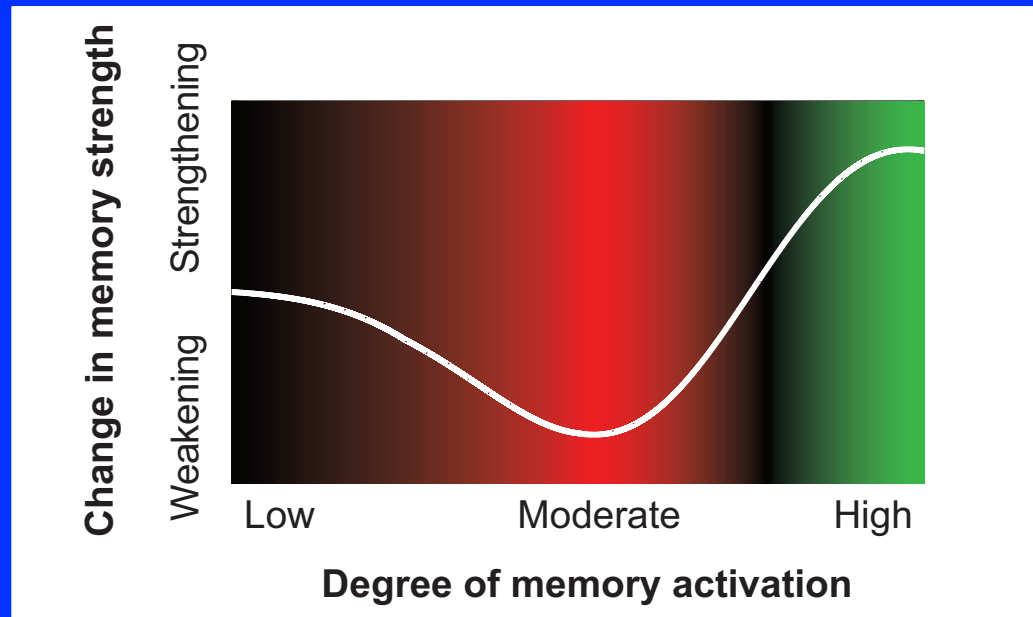
Think/No-Think (Anderson & Green, 2001)

- Study phase: Learn word pairs like “elephant-wrench”
- Later, participants are given cue words and told **not** to think of the studied associate
- At the end of the experiment, participants are given a cued recall memory test for studied items
- The no-think procedure leads to impaired recall

Variability

- These “inhibitory” memory effects are highly variable
- While a large number of studies have replicated Anderson’s think/no-think effect...
- .. there have also been several published failures to replicate these results (e.g., Bulevich et al., 2006; Bergstrom et al., 2007)
- Same thing with other inhibitory effects (e.g., retrieval-induced forgetting)

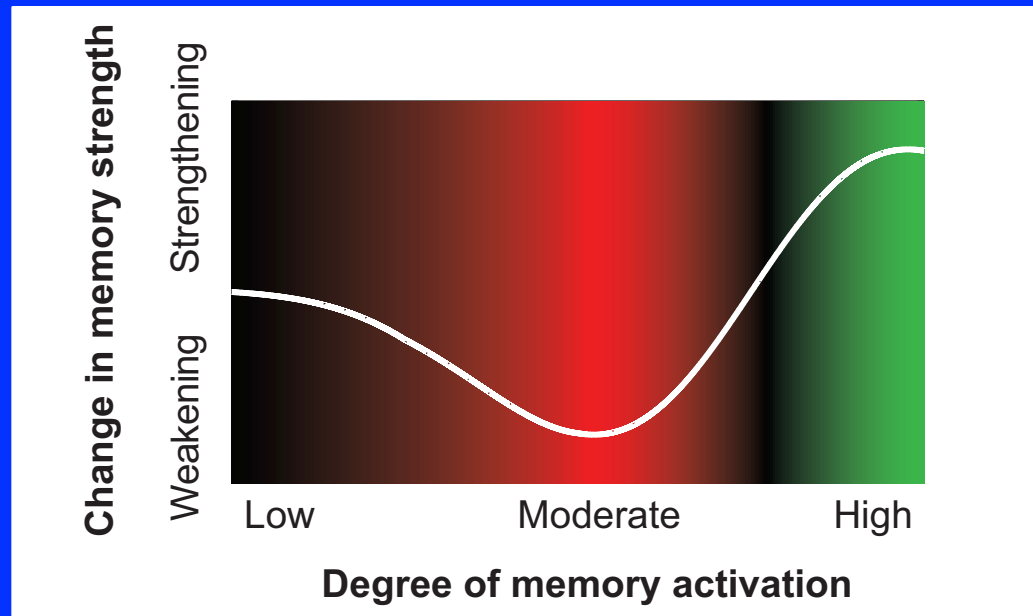
Explaining Inhibitory Memory Effects



Wrench

- No think trial: Don't think of what went with Elephant

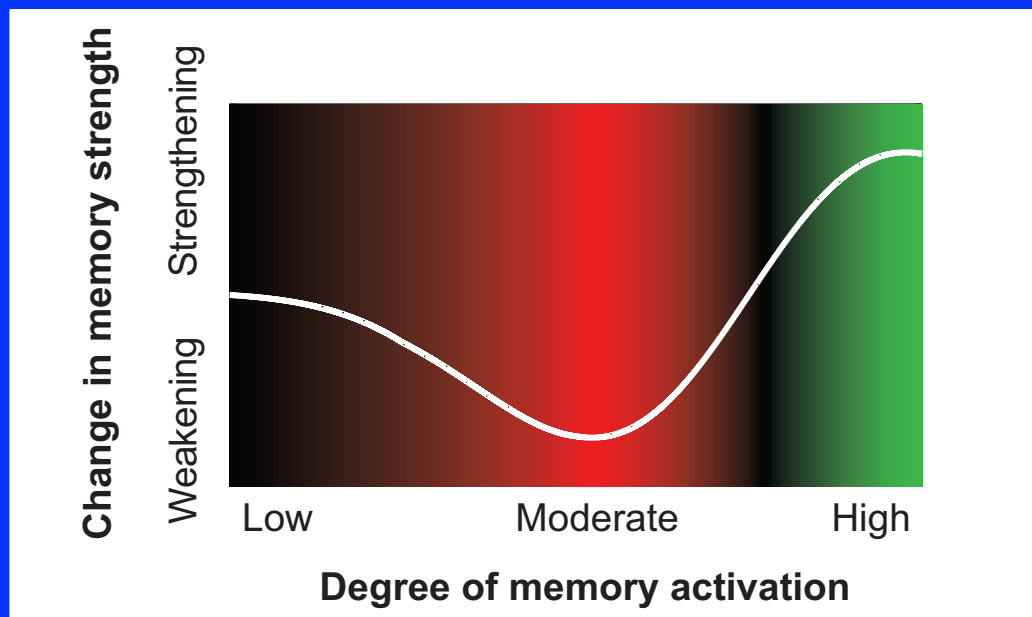
Explaining Inhibitory Memory Effects



Wrench

- No think trial: Don't think of what went with Elephant

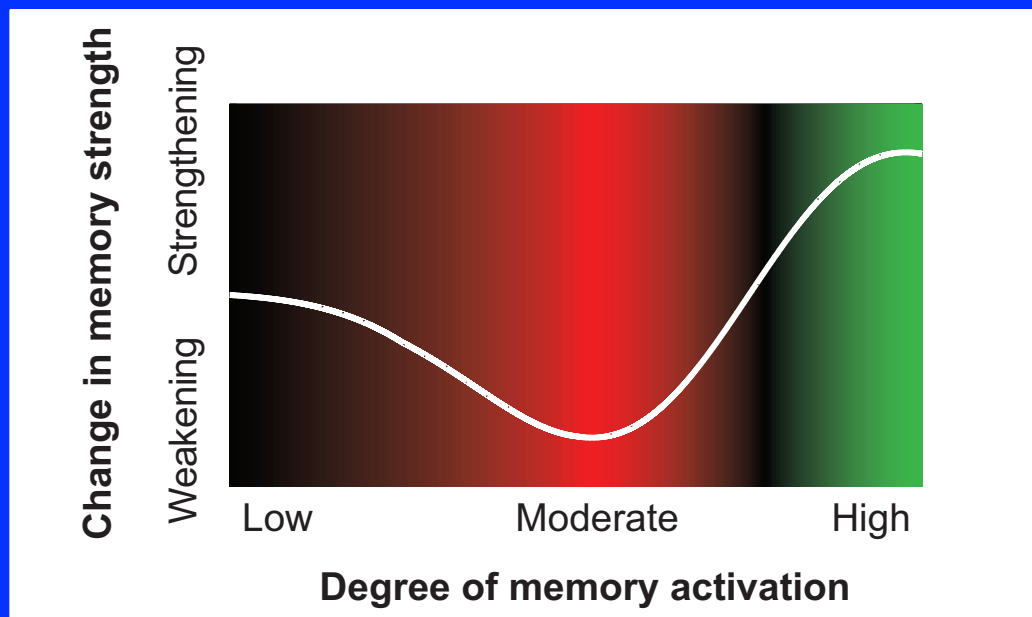
Explaining Inhibitory Memory Effects



Wrench

- No think trial: Don't think of what went with Elephant

Explaining Inhibitory Memory Effects



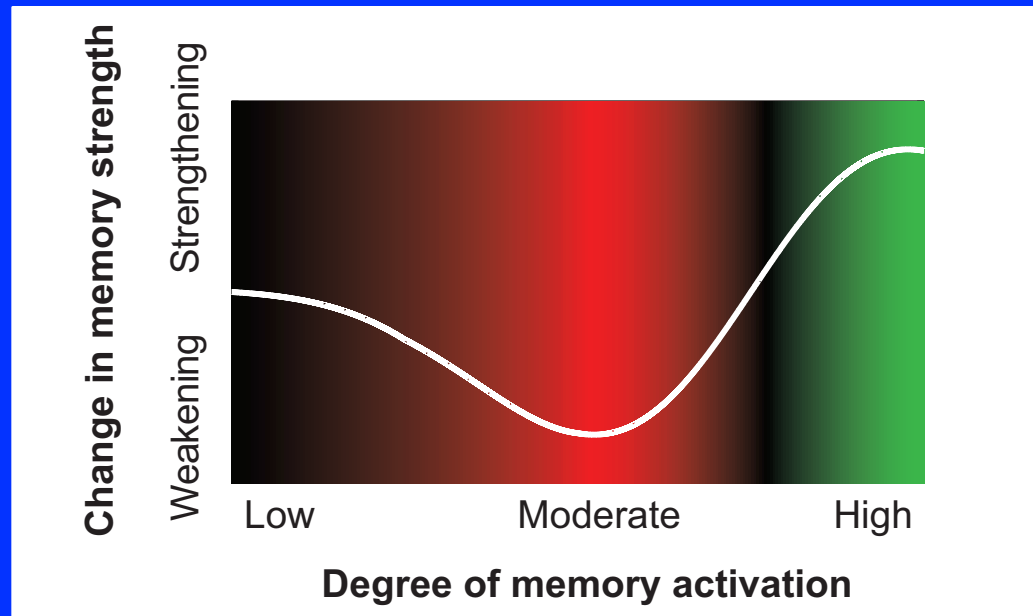
Wrench

Wrench

Wrench

- This theory makes it clear how inhibition can arise in this paradigm, and also why it is **very difficult** to get this effect...

Explaining Inhibitory Memory Effects



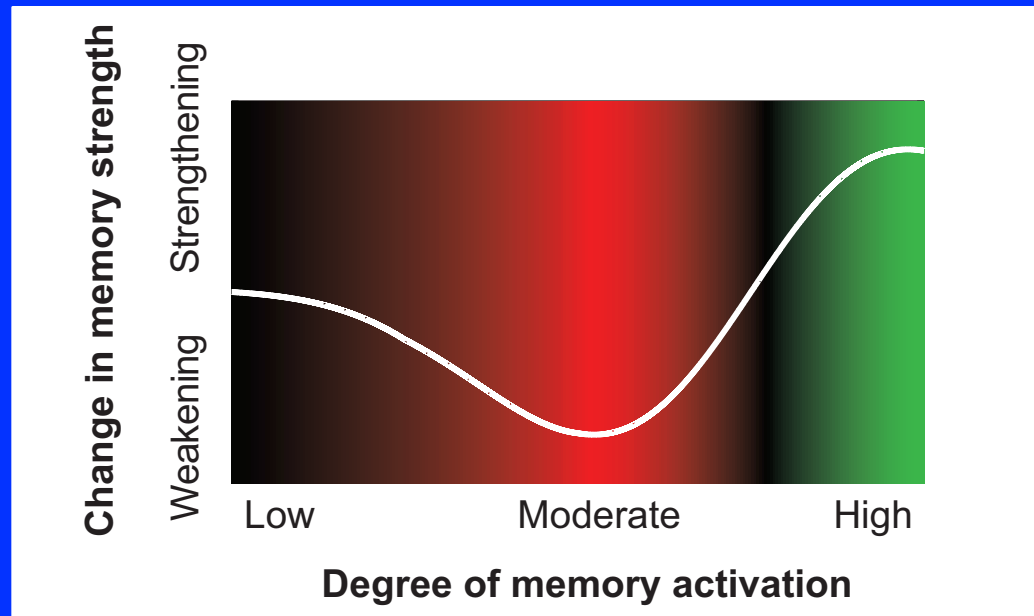
Wrench

Wrench

Wrench

- How do we test this idea?

Measuring Memory Activation using fMRI



Wrench

Wrench

Wrench

- Our approach: use pattern classifiers, applied to fMRI and EEG data, to **measure** how strongly memories are coming to mind on individual trials
- We then relate this **covert neural measure of retrieval** to performance on the final memory test

fMRI Study of Think / No-Think (Detre et al., 2013)



Greg Detre



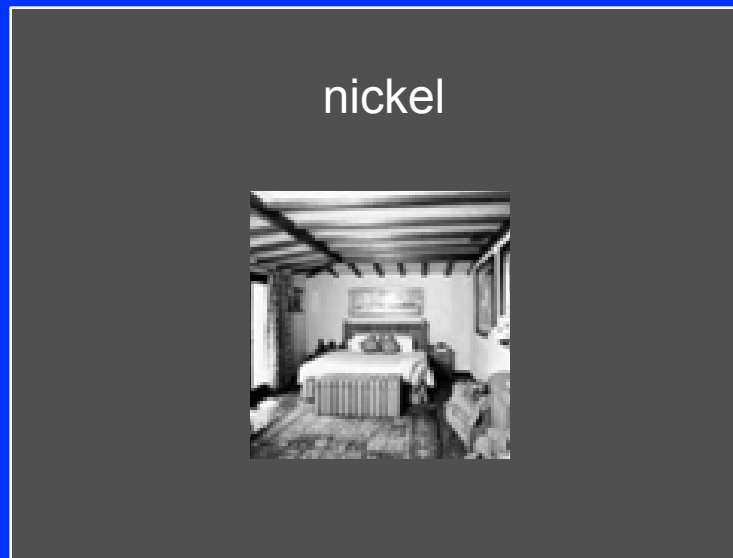
Malai Natarajan



Sam Gershman

Detre et al. (2013)

- As in Anderson & Green (2001), participants studied novel paired associates
- Instead of word-word pairs (like elephant-wrench), we used word-picture pairs
- Pictures were drawn from 4 categories: Face, Scene, Car, Shoe



Detre et al. (2013)

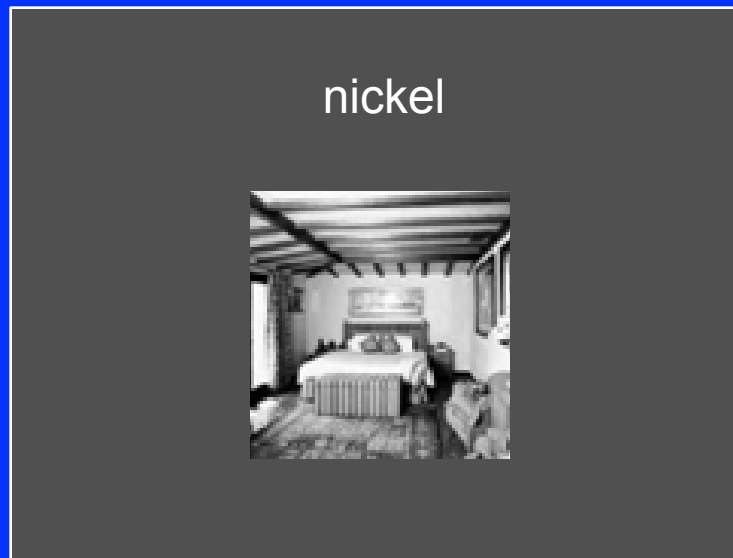
- Standard 3-phase design
 - study word-picture pairs
 - Think / no-think phase
 - memory test for studied pairs

nickel

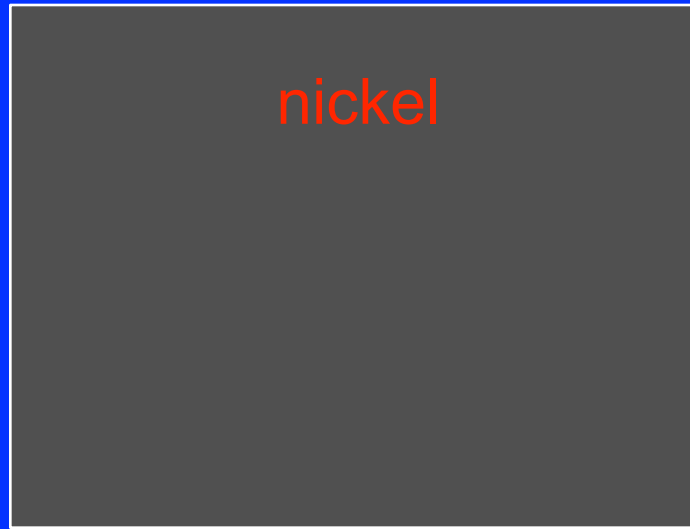


Detre et al. (2013)

- Pattern classification approach:
- We trained fMRI pattern classifiers to track activation relating to the four categories (Face, Scene, Car, Shoe)
- We used these classifiers to covertly track recall on no-think trials



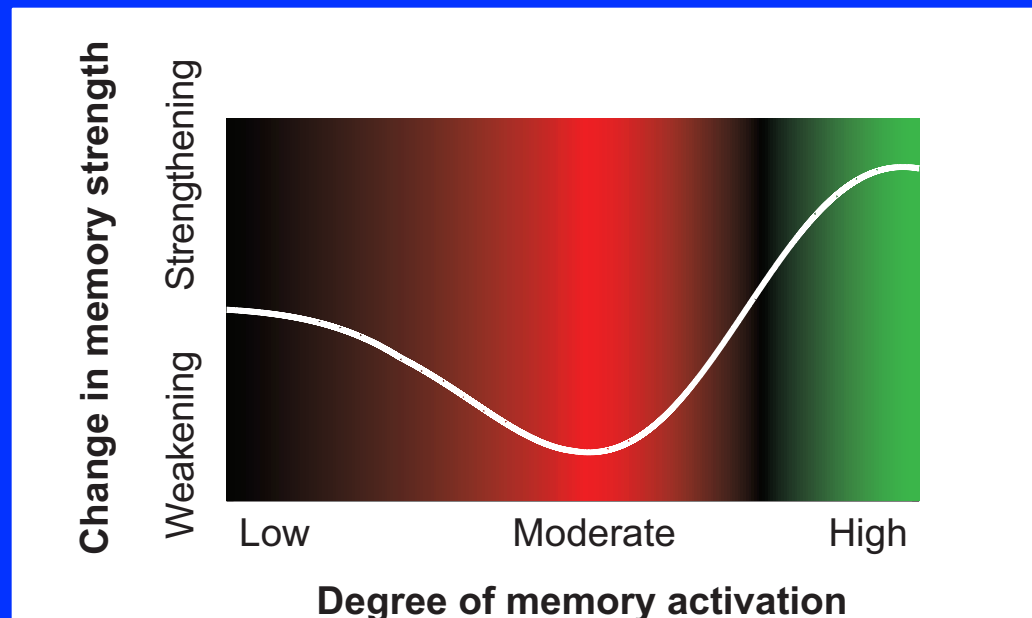
Sample No-Think Trial



- If nickel was paired with a scene at study, we would use the **scene** classifier on this trial to measure the extent to which the scene associate was coming to mind
- Prediction for this trial: Moderate levels of scene activity should be associated with forgetting, higher levels of scene activity should be associated with improved memory

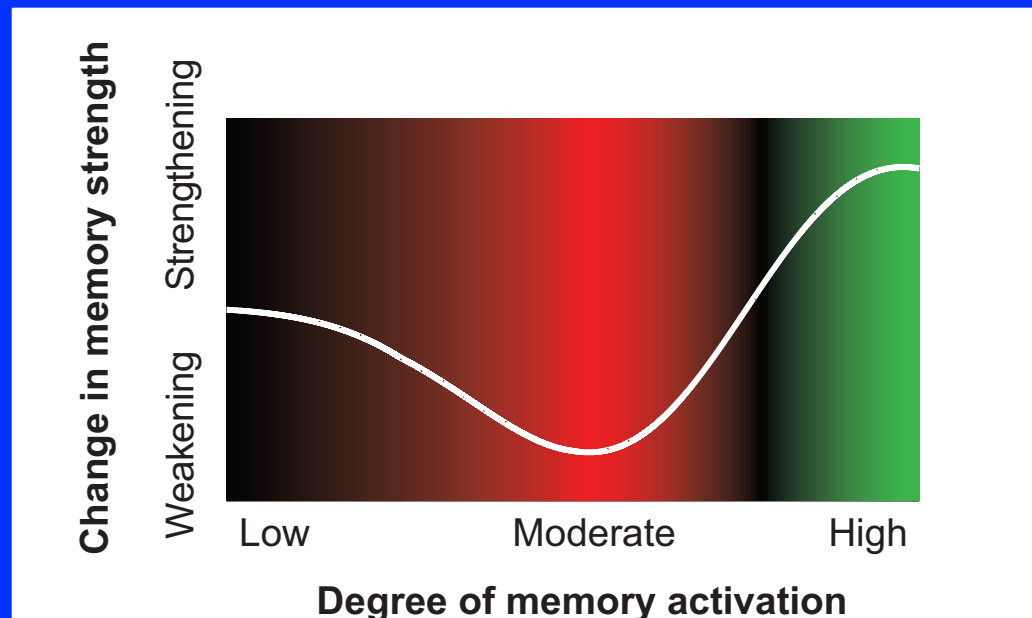
Analysis of fMRI Data

- Analysis goal: Estimate the shape of the function relating:
 - **how strongly a memory activated during the no-think phase** (as measured by the classifier), and
 - **how well that memory was recalled on the final test**
- To map out the shape of this function, we used an algorithm developed in my lab: P-CIT
- Probabilistic Curve Induction and Testing (Detre et al., 2013)



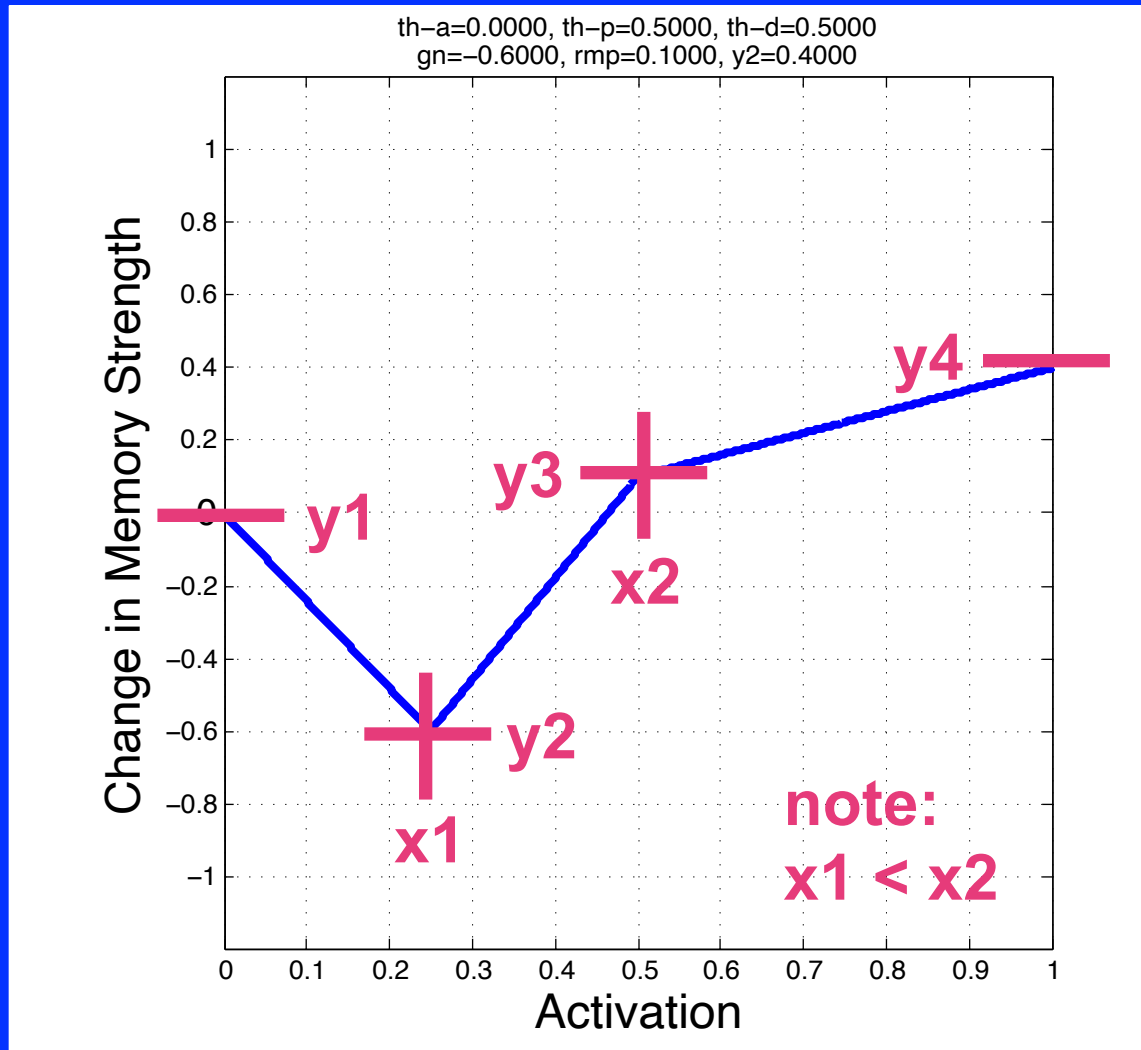
How P-CIT Works

- Step 1: Randomly generate a large number of curves
- Step 2: For each randomly generated curve, **evaluate** how well it explains our data
- Step 3: Compute a **weighted average** of the curves, where each curve is weighted by how well it explains the data

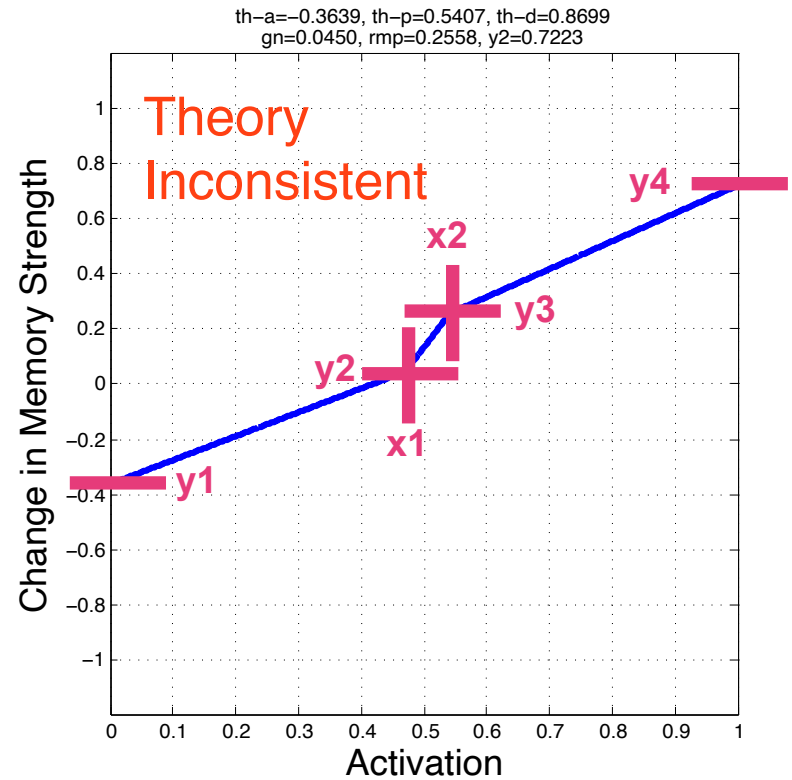
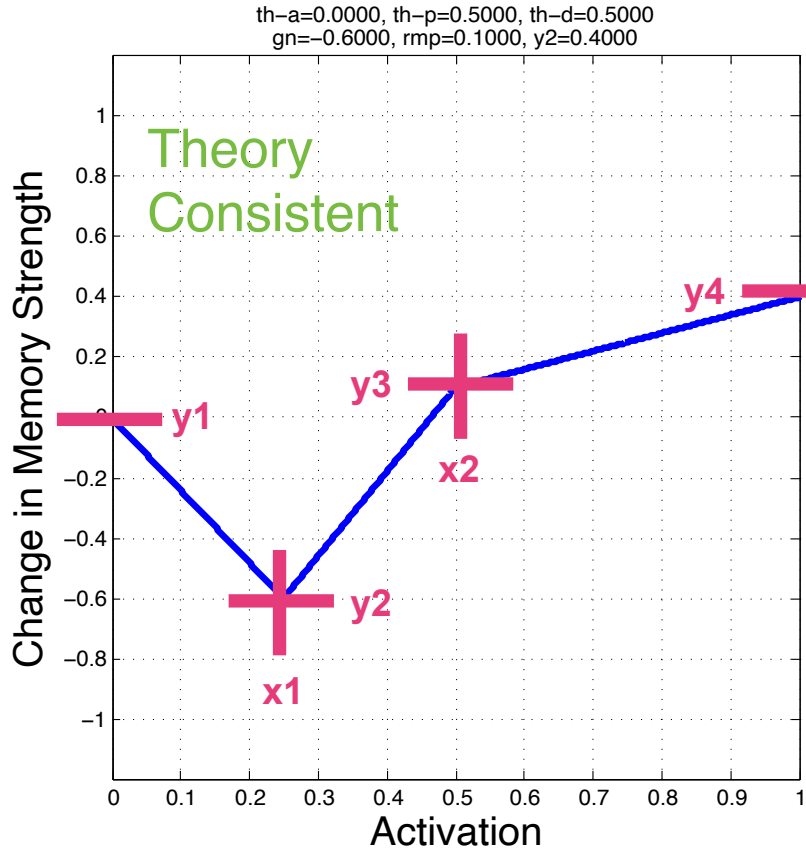


Parameterized Curve

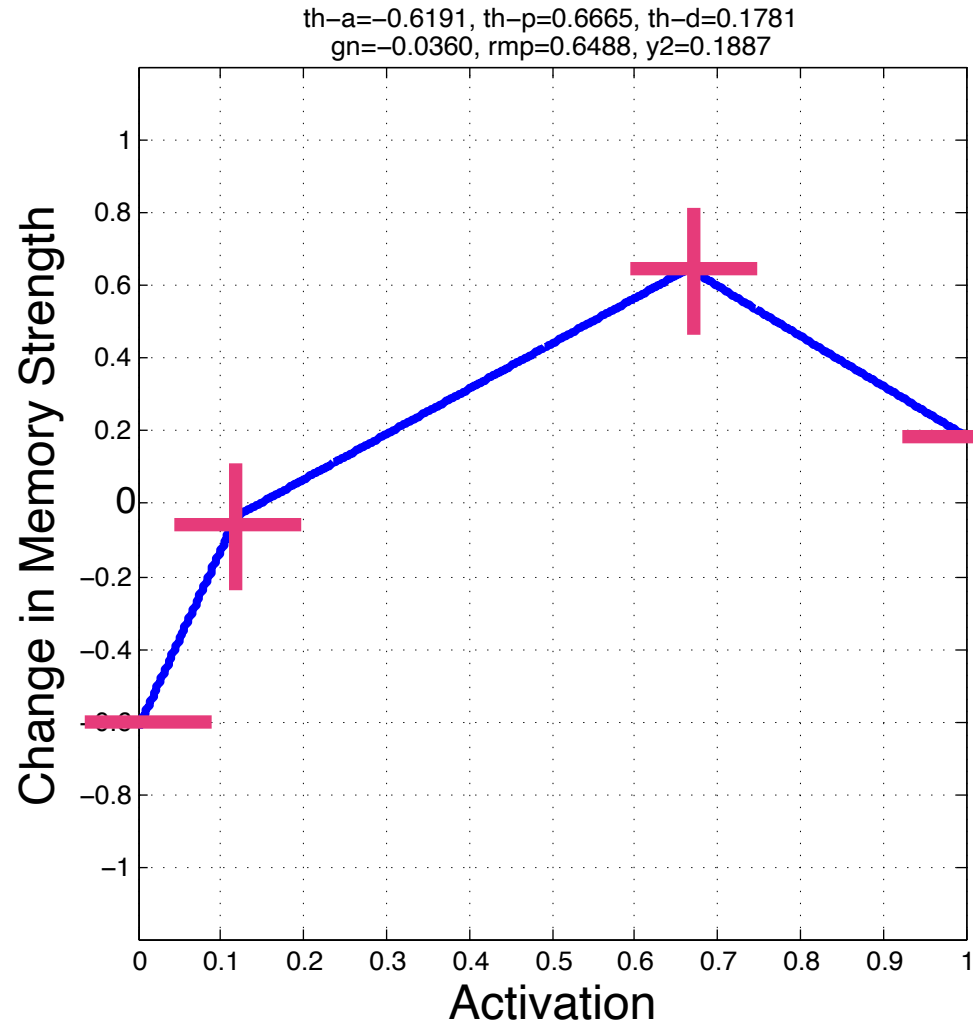
- Problem: The space of all possible curves is too big
- Solution: Use parameterized curves



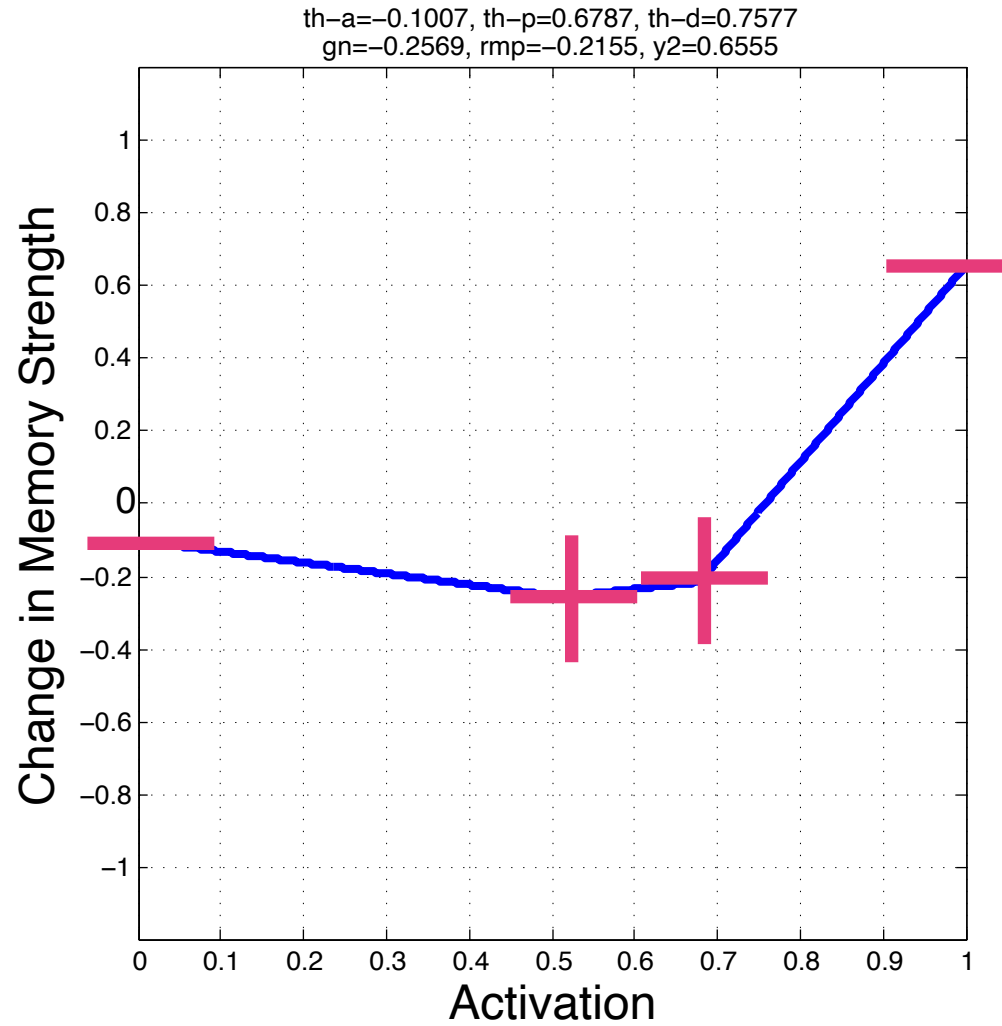
Examples of Curves



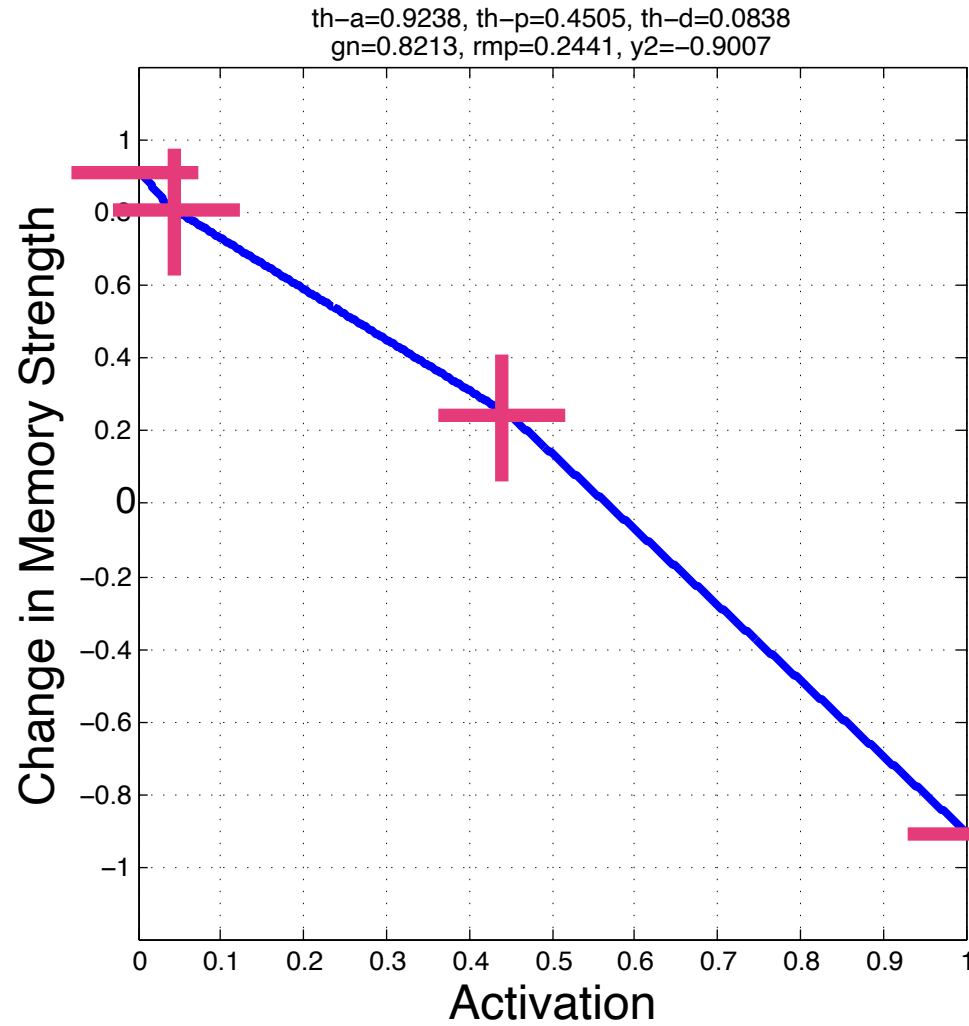
Curve Examples



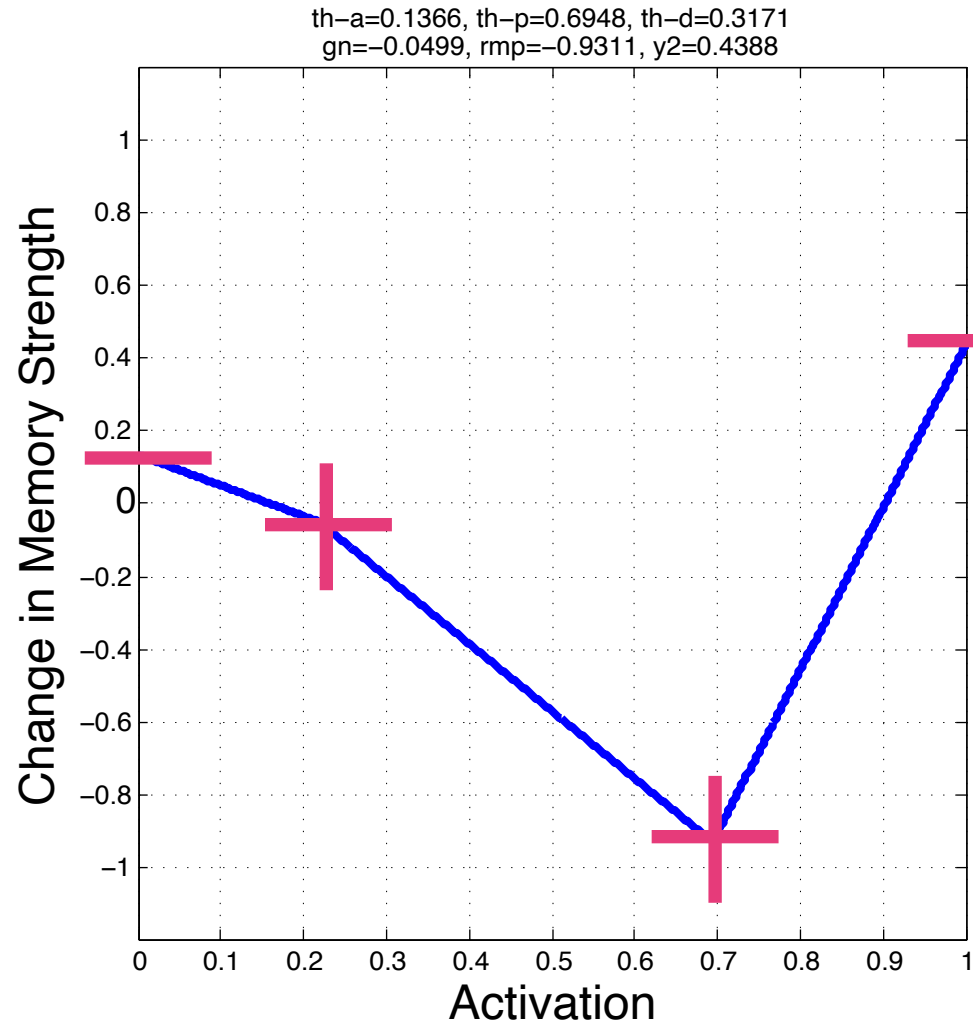
Curve Examples



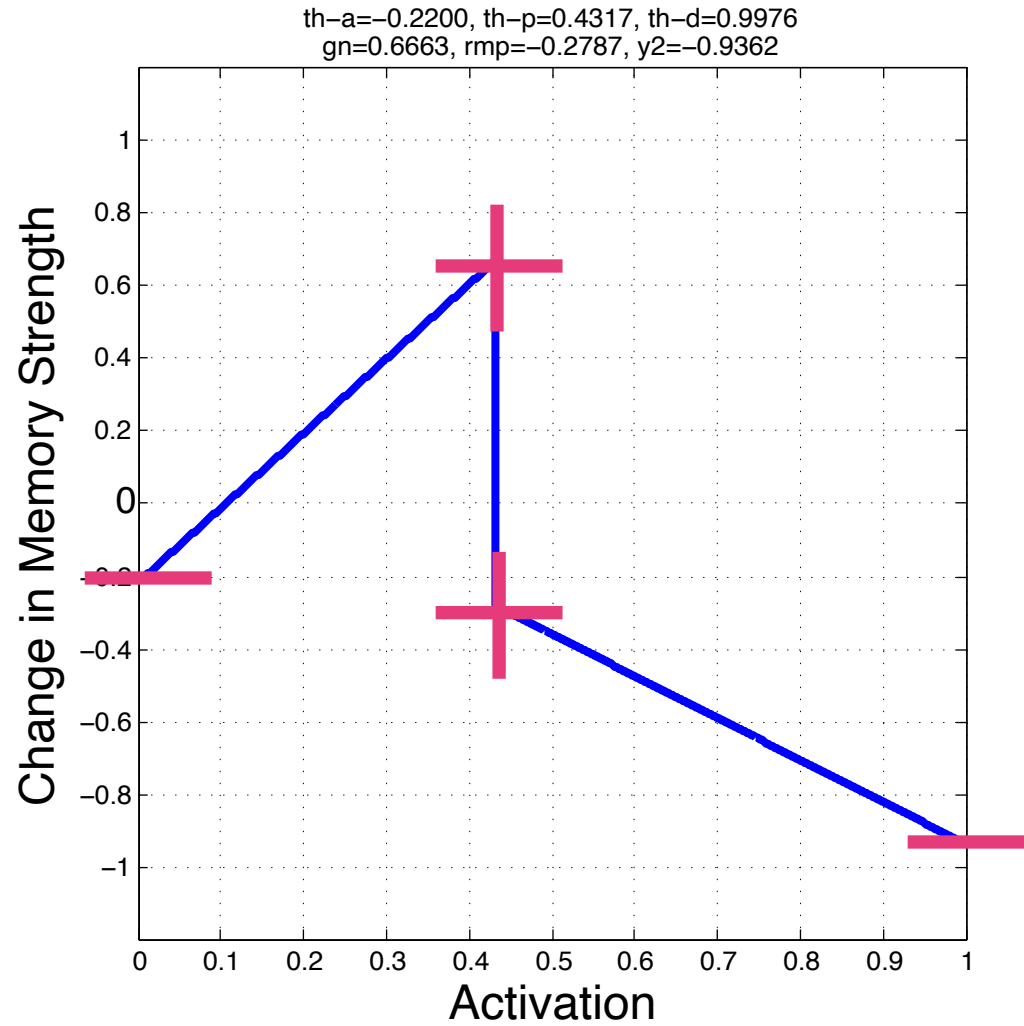
Curve Examples



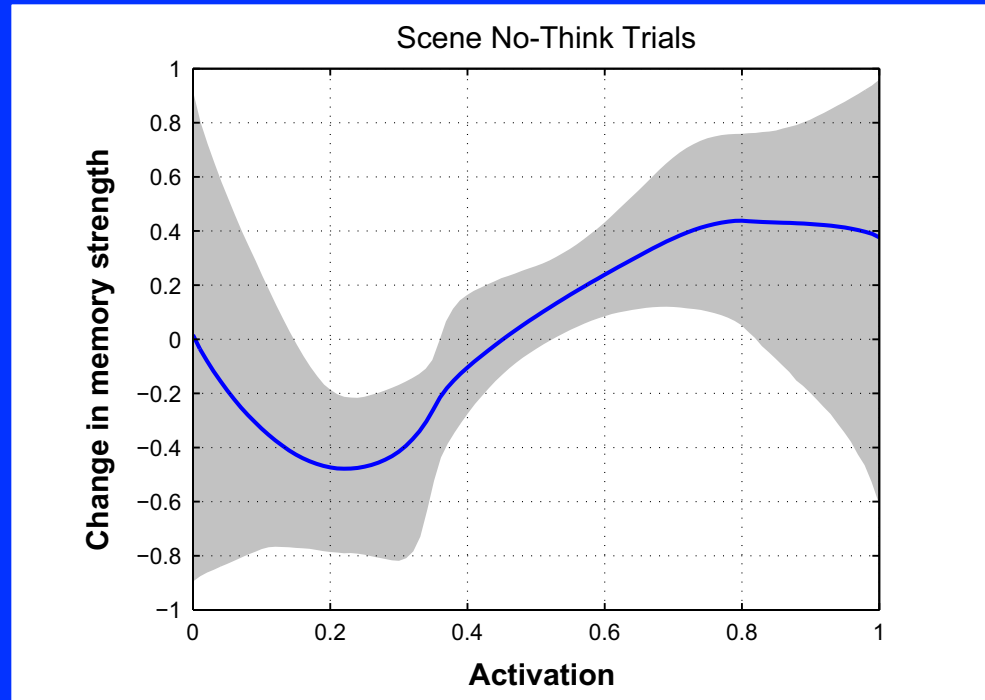
Curve Examples



Curve Examples



Estimated curve

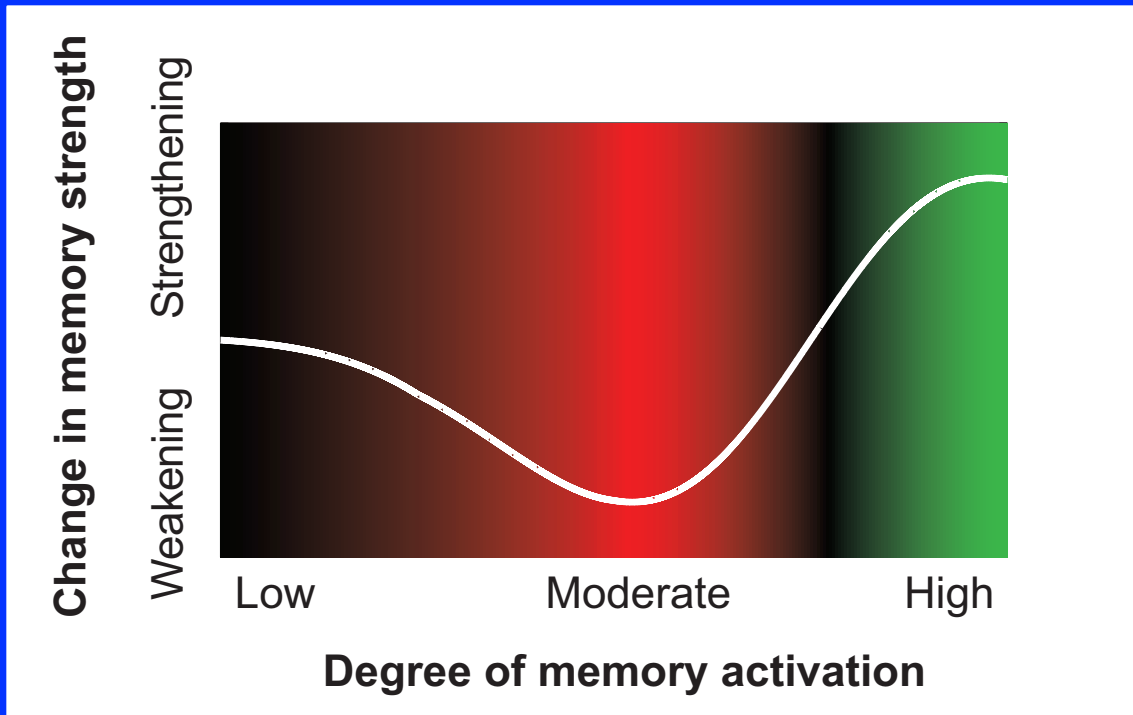


- Results were consistent with the nonmonotonic plasticity hypothesis

Summary Thus Far...

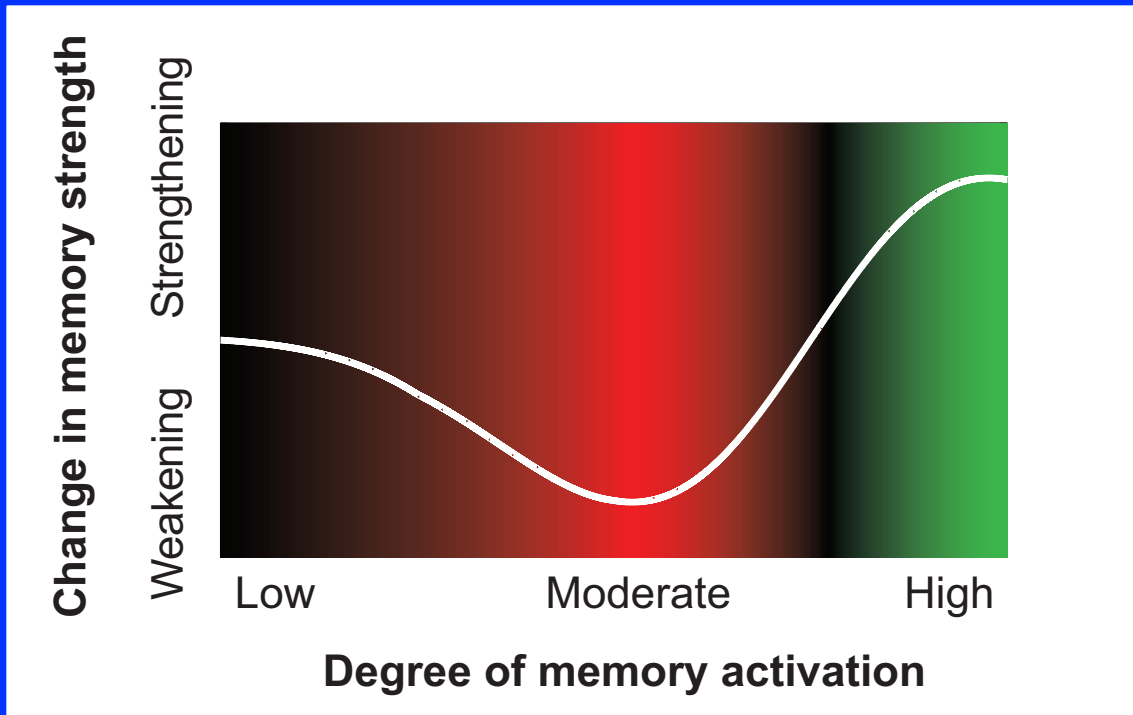
- U-shaped result in think-no think
- ... and in several other studies, not shown here
- Think-no think involves **use of executive control**
- Michael Anderson has argued that use of executive control is **necessary** to get these effects (e.g., Anderson & Huddleston, 2012)
- To the contrary, our model posits that executive control is not necessary for memory inhibition

Role of Executive Function in Causing Forgetting



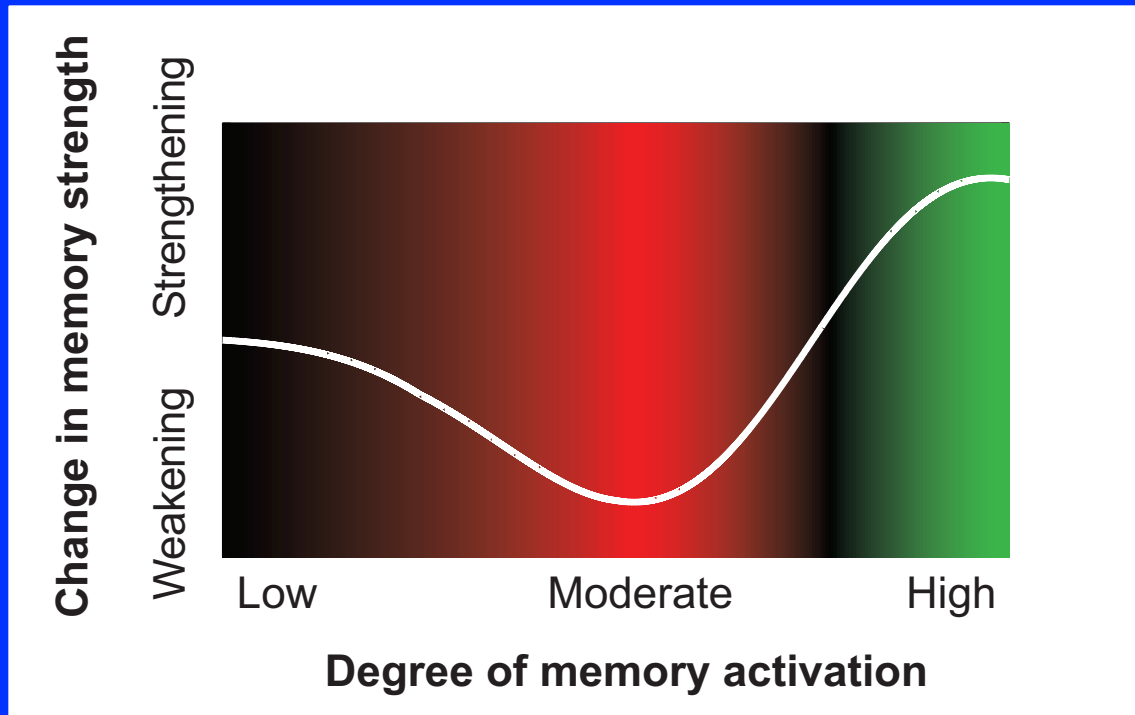
- According to our theory, the **level of activation** of a representation determines the learning that takes place
- Top-down control processes can **indirectly** affect learning by affecting the **level of activation** of the memories

Role of Executive Function in Causing Forgetting



- For example, in the think-no think paradigm, top-down control processes can boost forgetting by taking an item that would normally fall in the green zone and pushing it down into the red zone

Role of Executive Function in Causing Forgetting



- Key implication of our model: While top-down control can **promote** memory inhibition, top-down control is not (strictly speaking) **necessary** in order to get memory inhibition
- If a memory activates moderately in the **absence** of top-down control, forgetting should occur

Statistical Learning



Ghootae
Kim



Jarrod
Lewis-
Peacock



Nick
Turk-
Browne















**Later: surprise recognition test for
faces and scenes**

Design in encoding phase

Initial triplet

Repeated triplet



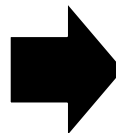
Face



Face



Scene



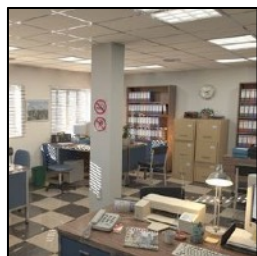
Face



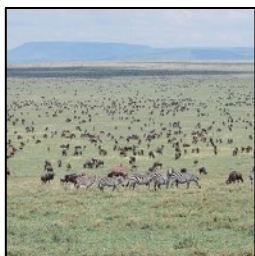
Face



Face



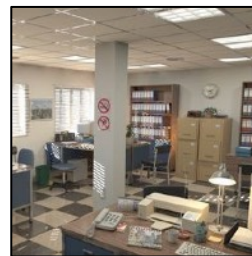
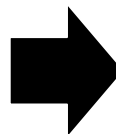
Scene



Scene



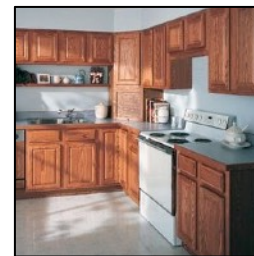
Face



Scene



Scene



Scene



Face



Scene

Control

Design in encoding phase

Initial triplet

R



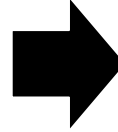
Face



Face



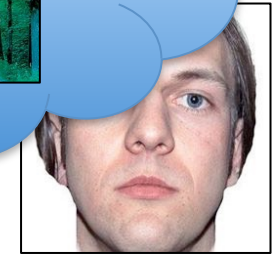
Scene



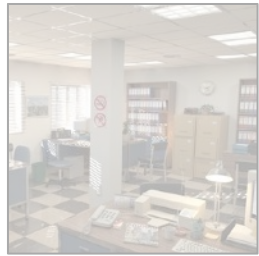
Face



Face



Face



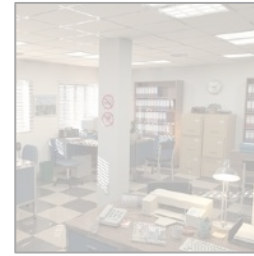
Scene



Scene



Face



Scene



Scene



Scene

Question: How does prediction of the scene affect subsequent memory for the scene?

Design in encoding phase

Initial triplet

R



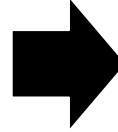
Face



Face



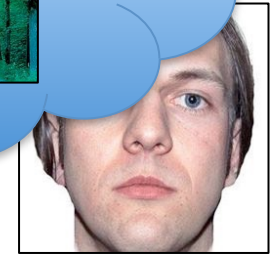
Scene



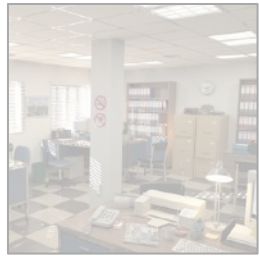
Face



Face



Face



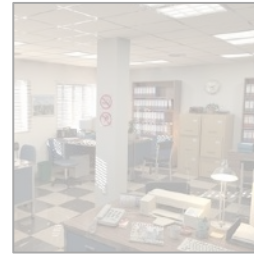
Scene



Scene



Face



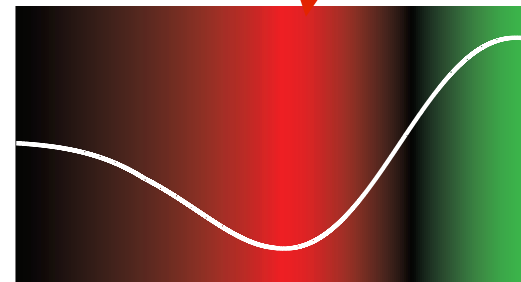
Scene



Scene



Scene



Design in encoding phase

Initial triplet

R



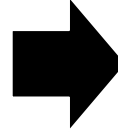
Face



Face



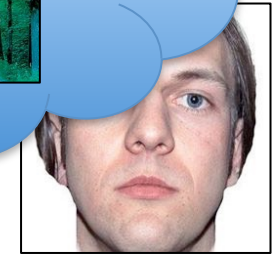
Scene



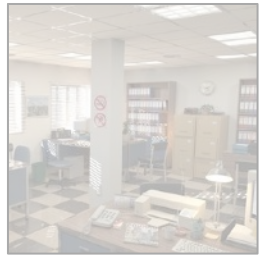
Face



Face



Face



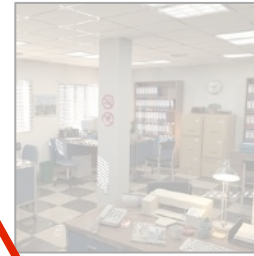
Scene



Scene



Face



Scene

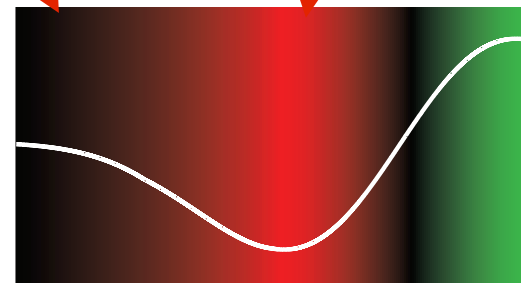


Scene



Scene

Question: How does perception of the scene during the initial triplet affect subsequent memory?



Design in encoding phase

Initial triplet

Repeated triplet



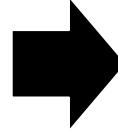
Face



Face



Scene



Face



Face



Face



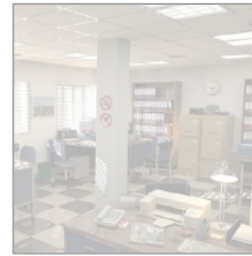
Scene



Scene



Face



Scene



Scene



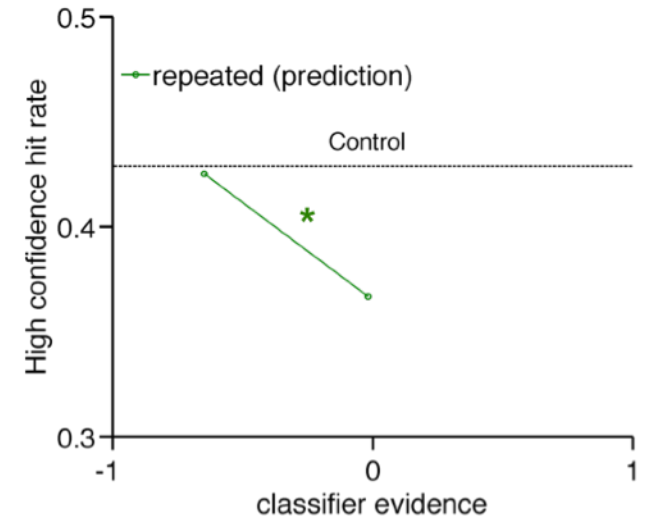
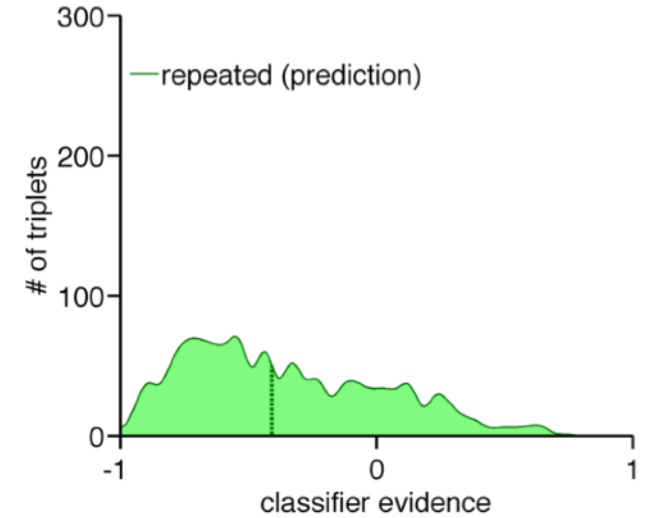
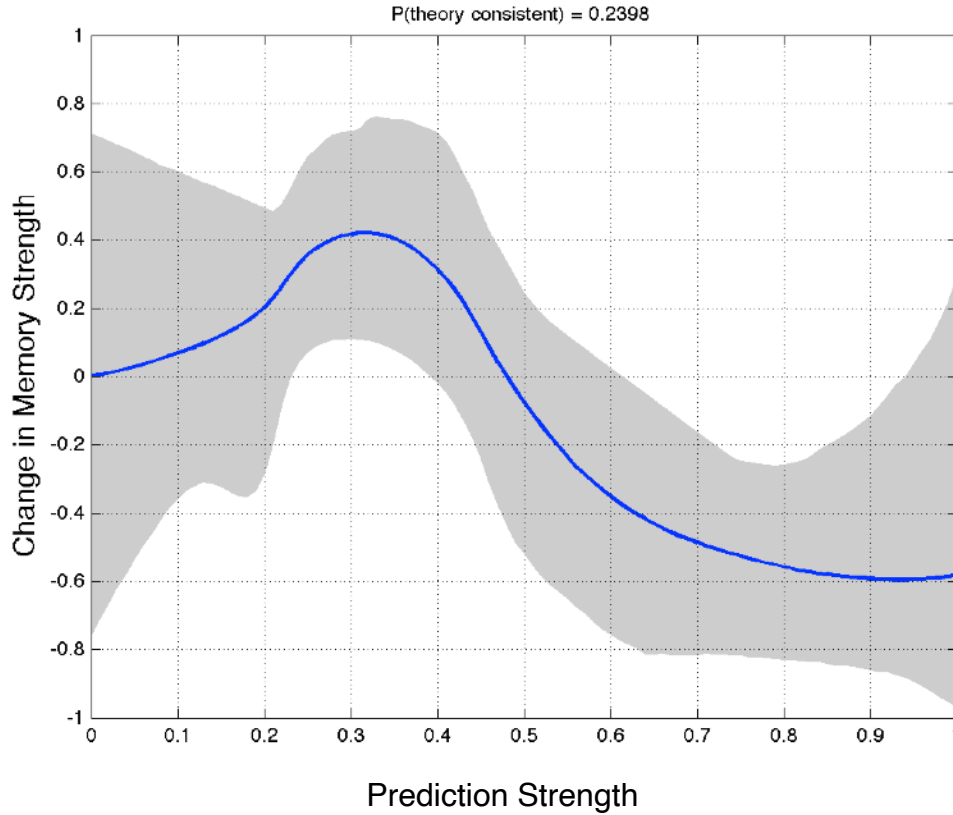
Scene

Analysis strategy:

Use pattern classifiers to measure activity relating to **perception** and **prediction**

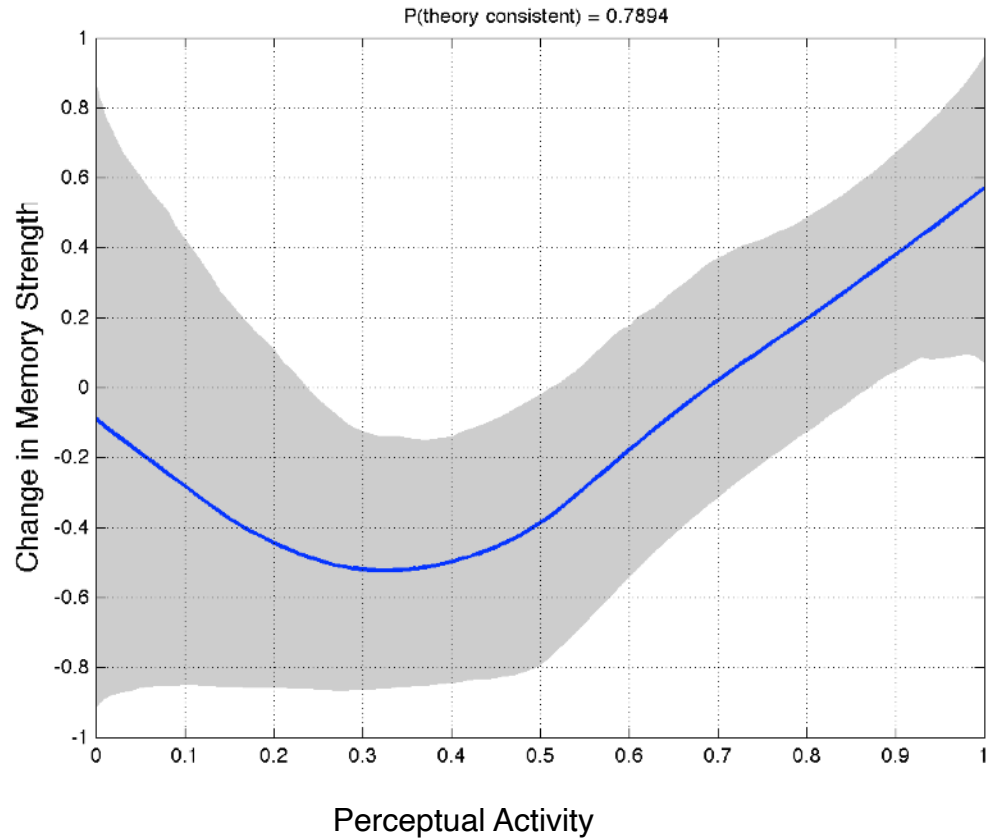
Relate these activity measurements to subsequent memory

Relationship between prediction strength and subsequent memory

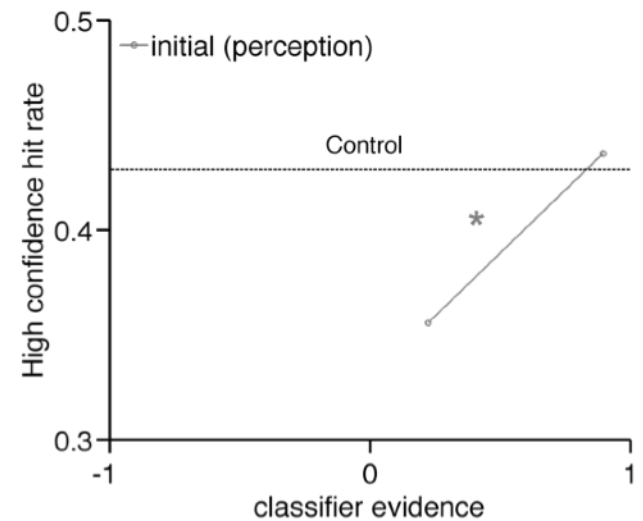
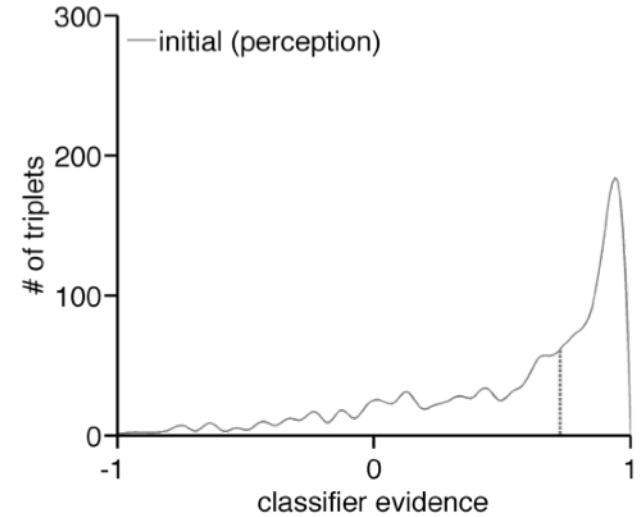


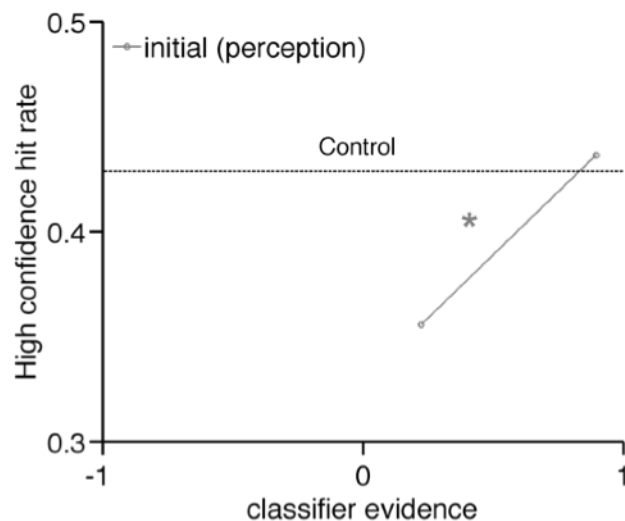
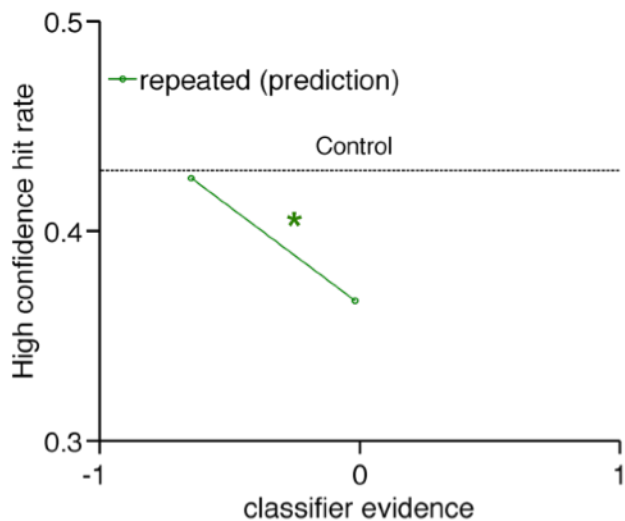
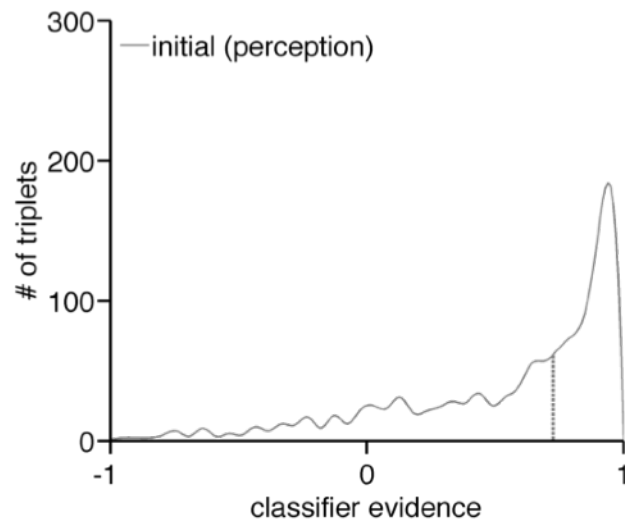
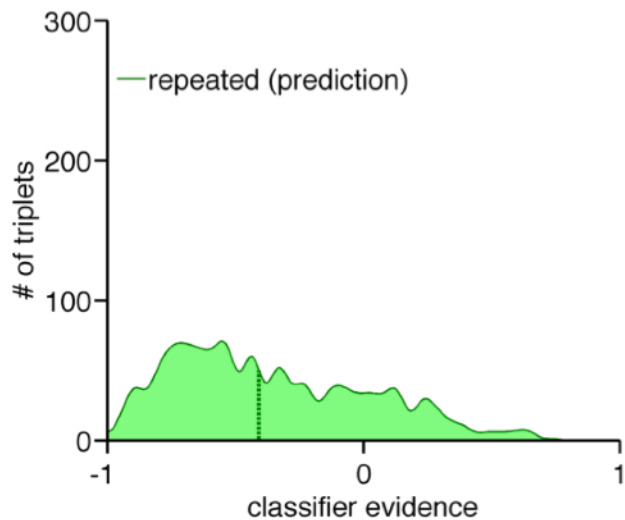
Higher levels of prediction strength were associated with **worse** subsequent memory

Relationship between perceptual activity and subsequent memory



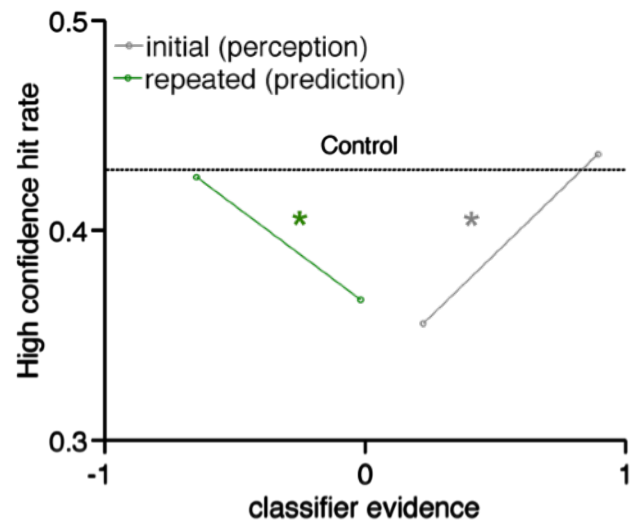
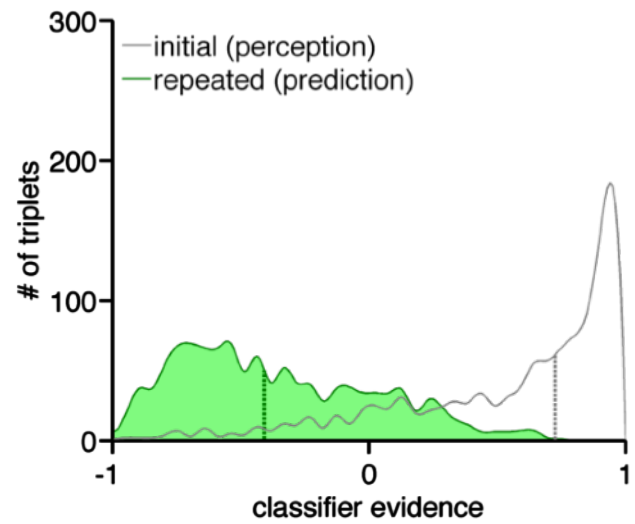
Higher levels of perceptual activity were associated with **better** subsequent memory

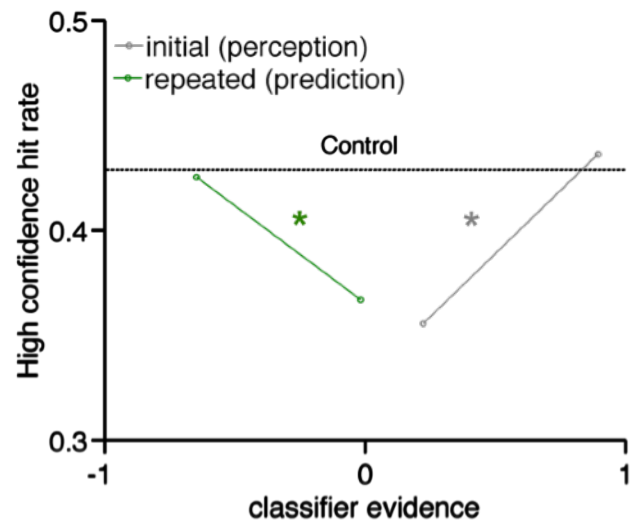
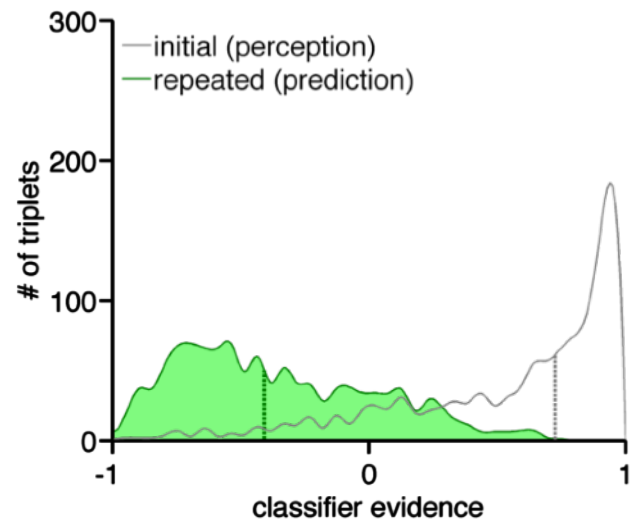
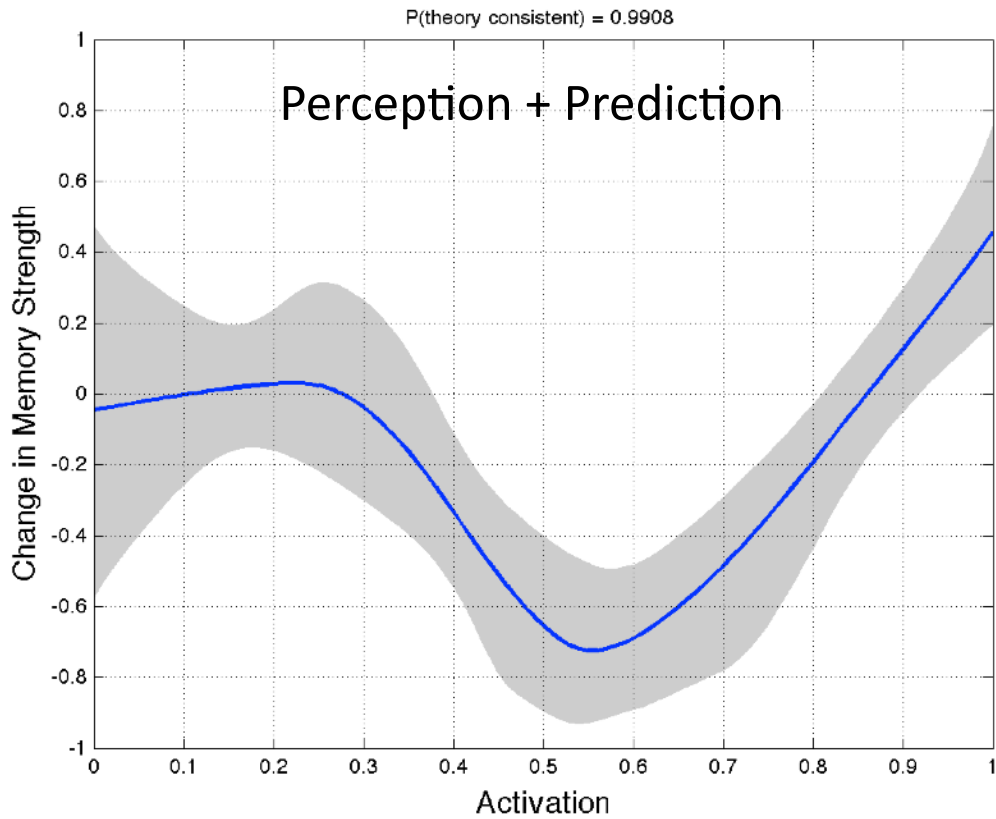




Prediction

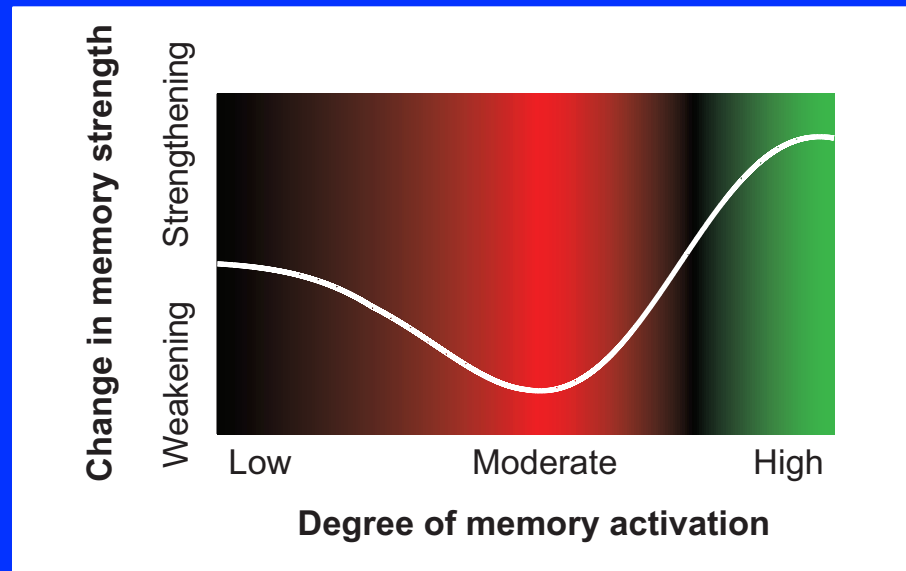
Perception





Summary: Statistical Learning

- We used a pattern classifier to relate activity during **perception** and **prediction** to subsequent memory
- We found a consistent relationship where, for both **perception** and **prediction**, moderate levels of activation were associated with worse memory
- These results fit with the nonmonotonic plasticity hypothesis



Summary: Statistical Learning

- Importantly, participants reported being **completely unaware** that the contexts were repeating
- It appears that forgetting was occurring as a result of **implicit predictions** triggered by context; there was no explicit intent to retrieve and no awareness of retrieval
- This suggests that memory inhibition effects do not require an intent to retrieve or an intent to suppress

Outline

- Situate NMPH relative to other kinds of learning
- Implications of NMPH for memory weakening
 - Key prediction: Moderate activation leads to weakening of competing memories
- **Implications of NMPH for the similarity structure of memories**
 - Key prediction: Moderate activation leads to differentiation of competing memories
- Current directions
 - Role of NMPH in learning during sleep
 - Using neurofeedback to promote discrimination learning

Representational Change

- Up to this point, I have been discussing coarse-grained predictions about **strengthening and weakening** of entire memories
- Now I will discuss finer-grained predictions, dealing with how neural representations grow **more or less similar** as a function of training
- Our initial investigations looked at this in the context of **interleaved learning**
 - What happens if you have two competing stimuli, and you alternate back and forth between studying them



Princeton Computational Memory Lab

Department of Psychology | Princeton Neuroscience Institute | Princeton University

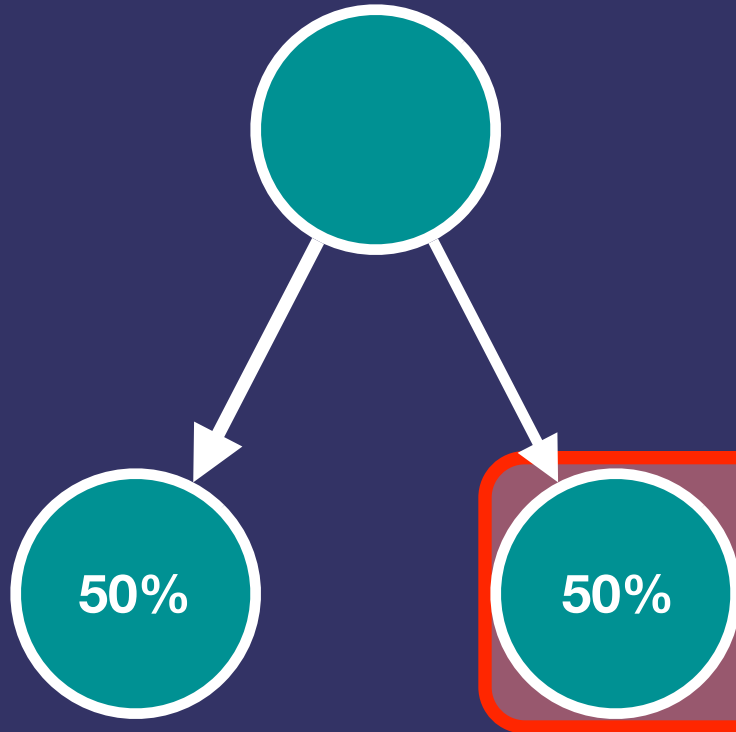


Jared

Page last modified on May 09, 2011, at 01:18 PM

NRP'ed CATEGORY

FRUITS

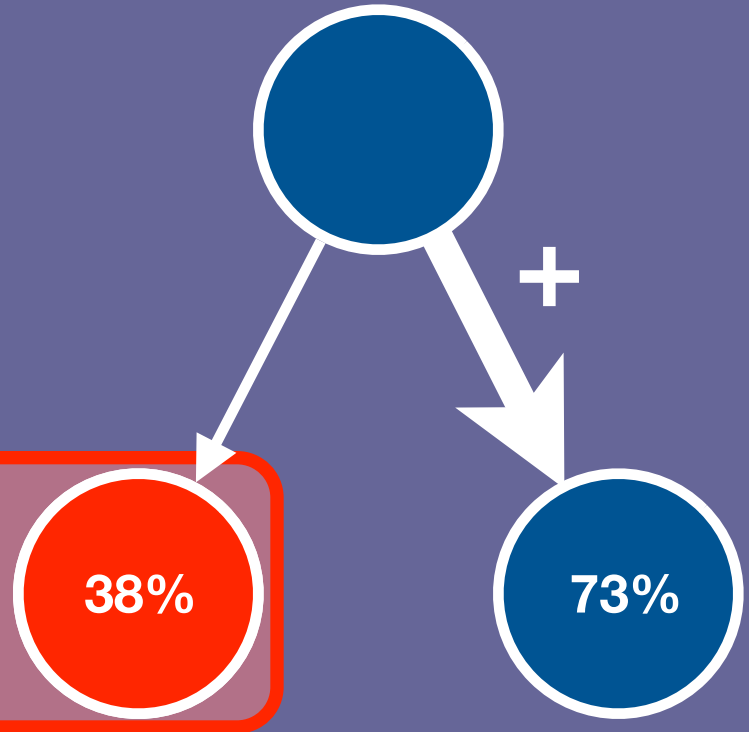


LEMON
Nrp

LIME
Nrp

RP'ed CATEGORY

POSTDOCS



JUSTIN
Rp-

JEREMY
Rp+

The Retrieval Practice Paradigm

Hypothetical Data



Initial Study

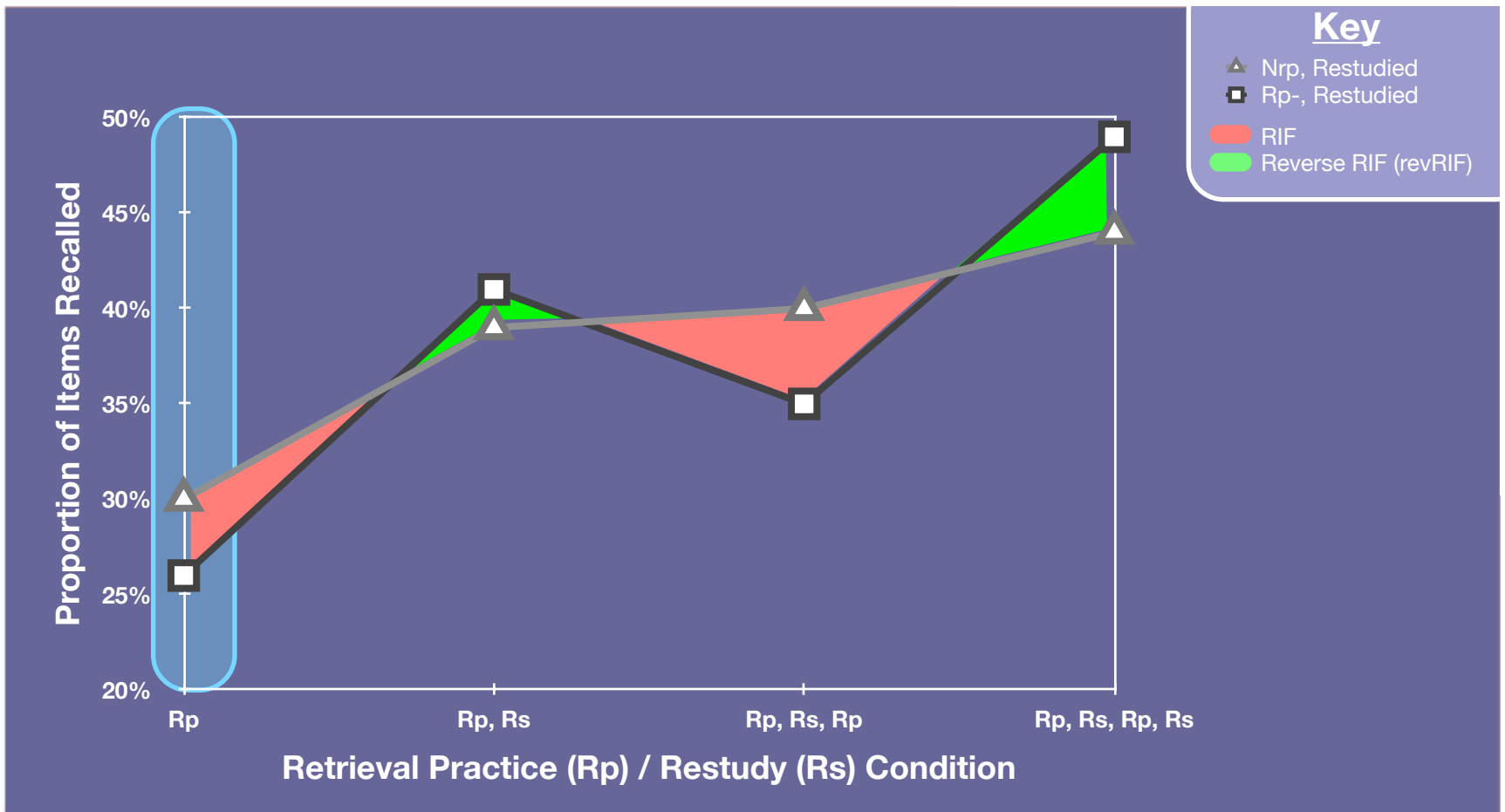
FRUITS-**LEMON**
POSTDOCS-**JEREMY**
FRUITS-**LIME**
POSTDOCS-**JUSTIN**

Retrieval Practice (RP)

POSTDOCS-**JE**___

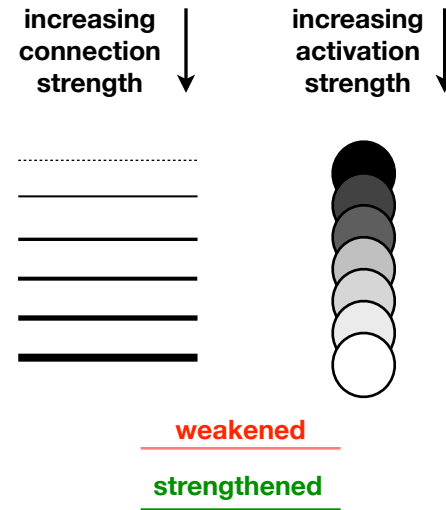
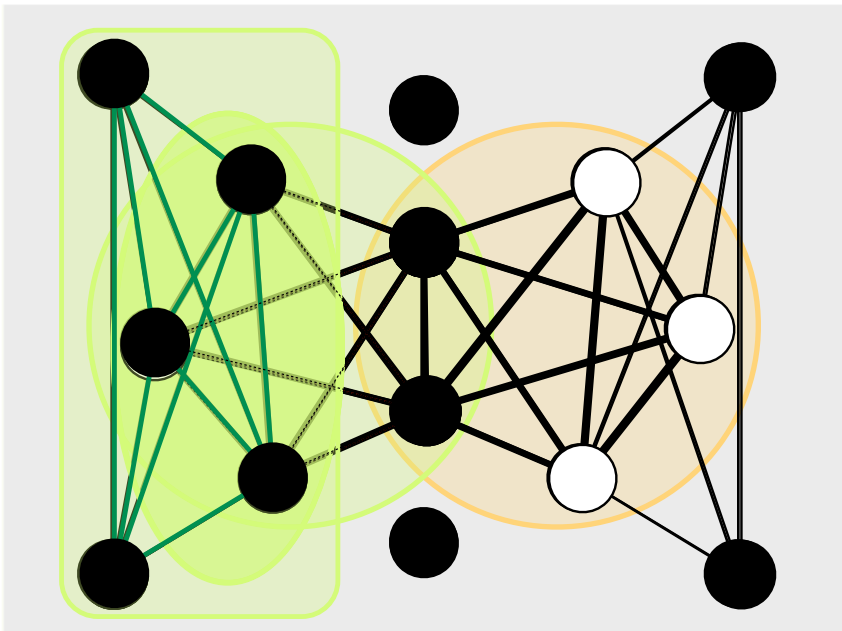
Final Test

FRUITS-**LE**___
POSTDOCS-**JU**___
FRUITS-**LI**___
POSTDOCS-**JE**___



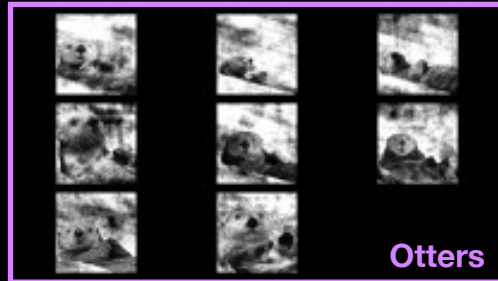
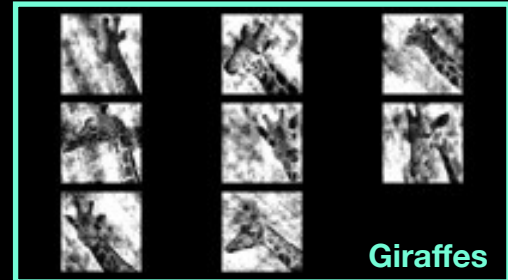
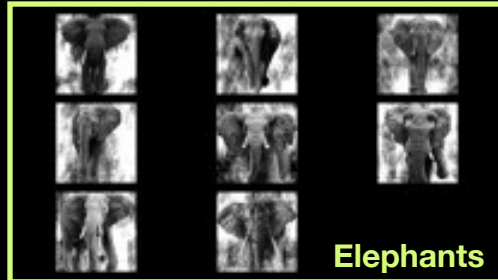
Effects of Interleaved Rp and Restudy

JEREMY (R_{p+})
JUSTIN (R_{p-})



Weights After Restudy

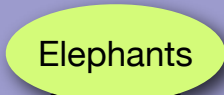
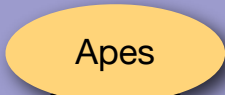
Representational Differentiation



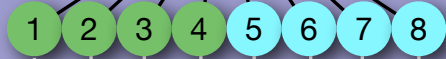
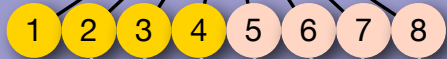
The Cast of Characters

Rp'ed

categories:



exemplars:



names:

Abner Alden Amin Amos Angus Anton Arlen Asher

Edbert Egan Emmet Eton Efren Elden Ennis Ewan

Garret Gino Godwin Griswold Gifford Glenworth Gordie Gunther

Rp+

Rp-

Rp+

Rp-

Rp+

Rp-

Nrp'ed

categories:



exemplars:



names:

Landin Lester Linton Luka Leroy Levar Lonnie Luther

Oakley Olaf Orrin Osgood Odin Omar Orson Oxley

Parrish Paxton Percy Prescott Patrice Pearson Perry Pryor

Nrp_a

Nrp_b

Nrp_a


Nrp_b


Nrp_a

Nrp_b

The Cast of Characters

Rp+ (competitive)

RP  _____ the _____

RS  Abner the Ape

Rp- & Nrp (non-competitive)

RS  Amin the _____

RP  _____ the Ape

RS  Odin the _____

RP  _____ the Otter

2) Interleaved Retrieval Practice & Restudy

4 rounds
2x per item

Feedback Trial for Rp+ items (Competitive Retrieval + Restudy) *Feedforward Trial for all other items (Restudy + Working Memory Dump)*

FEEDBACK or FEEDFORWARD

 Abner  Abner the Ape +  Odin the Otter  _____ the Otter ...

2s 2s jittered (1-7.5s) 2s 2s

3) Post-Manipulation Acquisition

1 round
1x per item

RESTUDY

 Olaf the Otter 8+9  Abner the Ape 5+3  Odin the Otter ...

2s 6s 2s 6s 2s

4) Final Behavioral Test

1 round
1x per item

FINAL RECALL

 _____ the Otter +  Olaf +  Abner ...

4s 2s 4s 2s 4s

TIME →

The Design

Scanning

Naming Practice

RS1

RP1

RP2

RP3

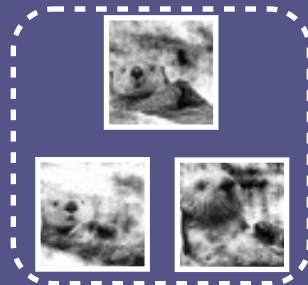
RP4

RS2

Final Test



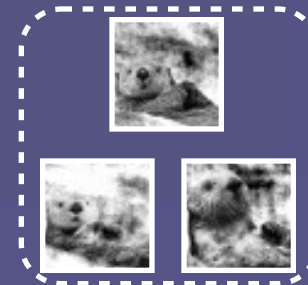
Nrp'ed



Rp'ed



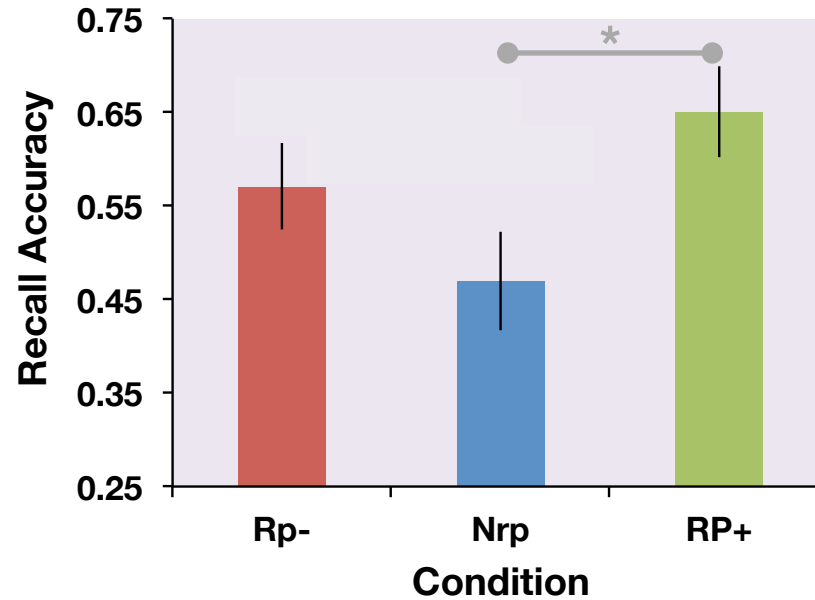
Nrp'ed



Rp'ed

Neural Learning Score: $(Nrp'ed_2 - Nrp'ed_1) - (Rp'ed_2 - Rp'ed_1)$

Final Test Performance



Behavioral Results

Key prediction:

IF revRIF occurs because of neural differentiation,
THEN

Across participants, the amount of neural
differentiation should predict the amount of
behavioral revRIF

Neural Learning & revRIF

Region	Neural Differentiation & revRIF
Both Hippocampi	$r=.34$ ($p=0.051$)~
Left Hippocampus	$r=.43$ ($p=0.017$)*
Right Hippocampus	$r=.26$ ($p=0.108$)

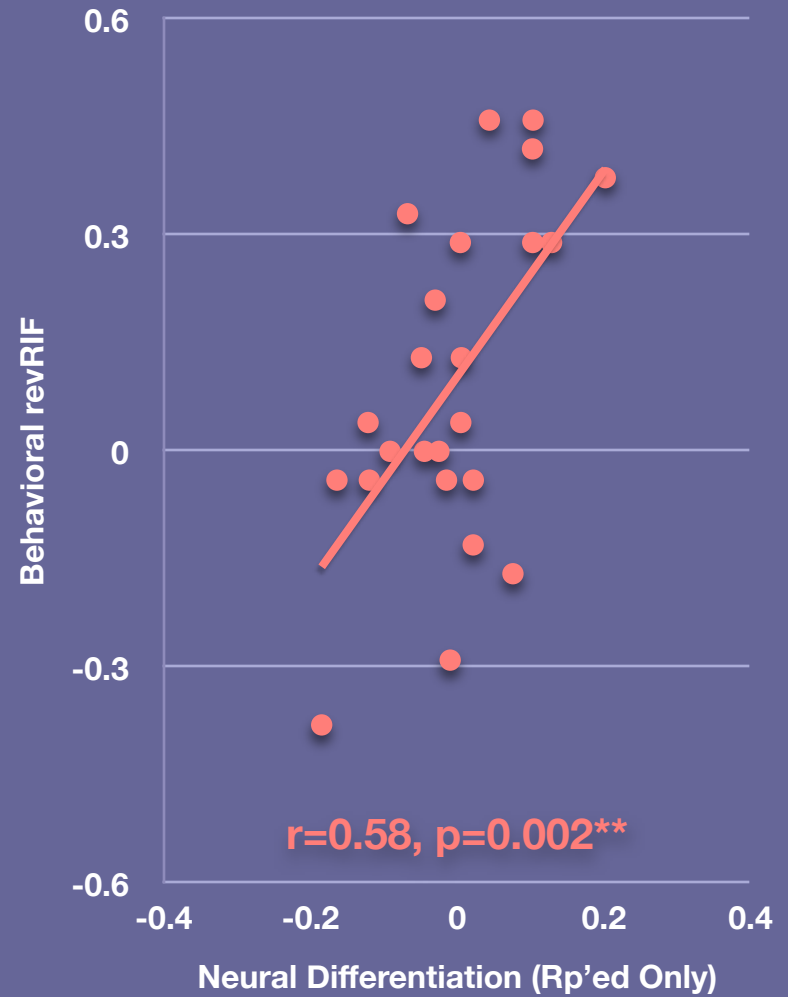
Neural Learning & revRIF

- Both Hippocampi
- Left Hippocampus
- Right Hippocampus

Neural Learning Score: $(Nrp'ed_2 - Nrp'ed_1) - (Rp'ed_2 - Rp'ed_1)$

Nrp'ed Categories

Rp'ed Categories



● Left Hippocampus



- ▶ As predicted by a neural network model of memory:
 - We observed a trend towards **revRIF**
 - **Differentiation** in the left hippocampus predicted revRIF
- ▶ Without differentiation, memory is a zero-sum game
- ▶ Differentiation provides a way out of this zero-sum trap
 - Pulling representations apart reduces competition while still preserving access to all of the memories

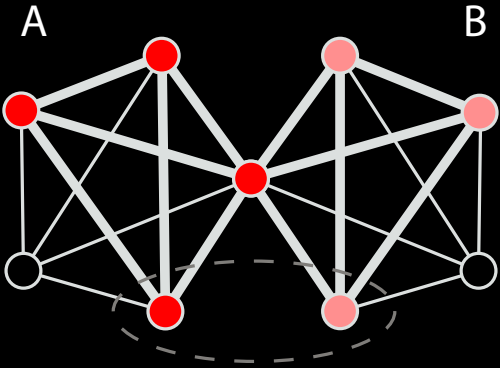
Conclusions

Further Tests of Differentiation Prediction

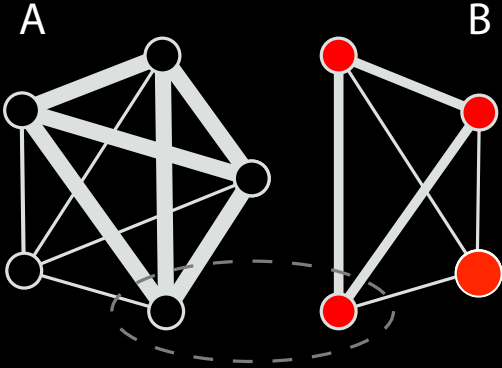
- Hulbert & Norman (2015) showed that differentiation can occur after competition & restudy
- Crucially, the NMPH predicts that differentiation will depend on the **level of activation of the competing memory**

Differentiation after moderate activation

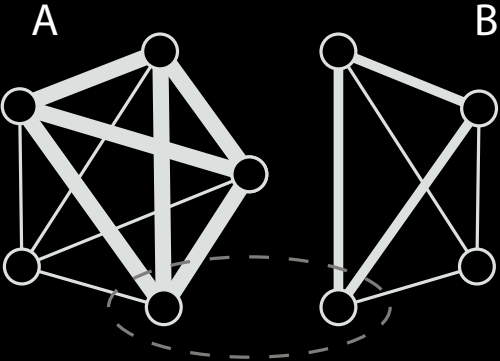
Competition trial



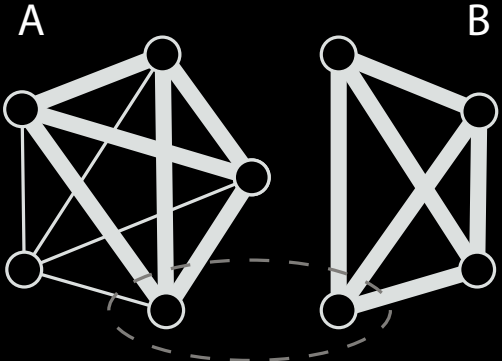
Restudy trial



resulting weight changes

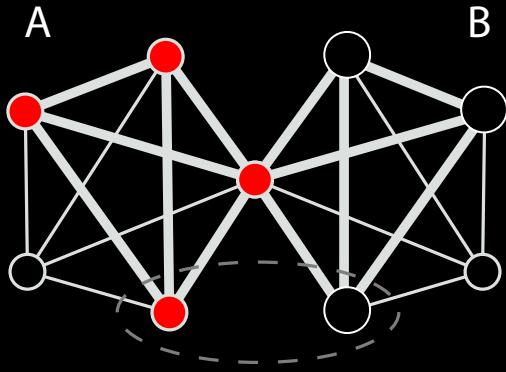


resulting weight changes

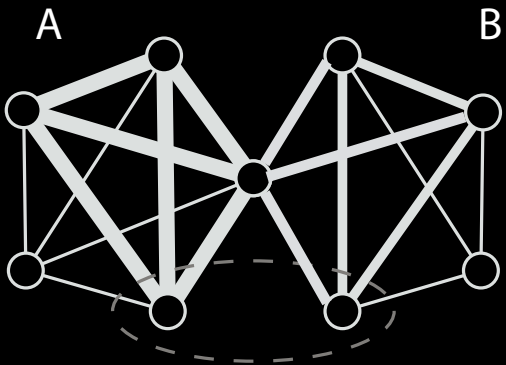


No competitor activation = no learning

Competition trial

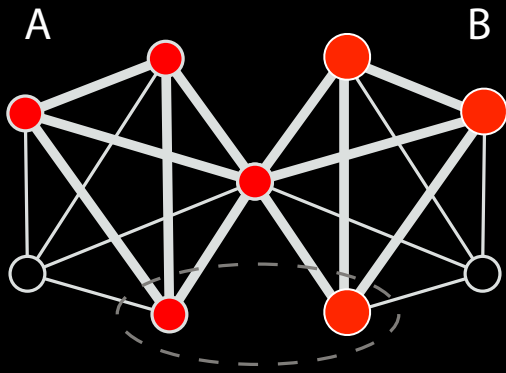


resulting weight changes

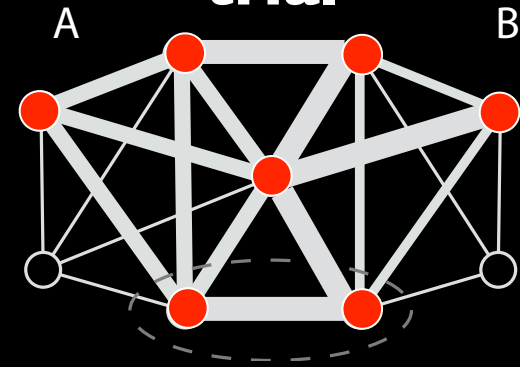


Strong competitor activation = integration

Competition trial



Restudy trial



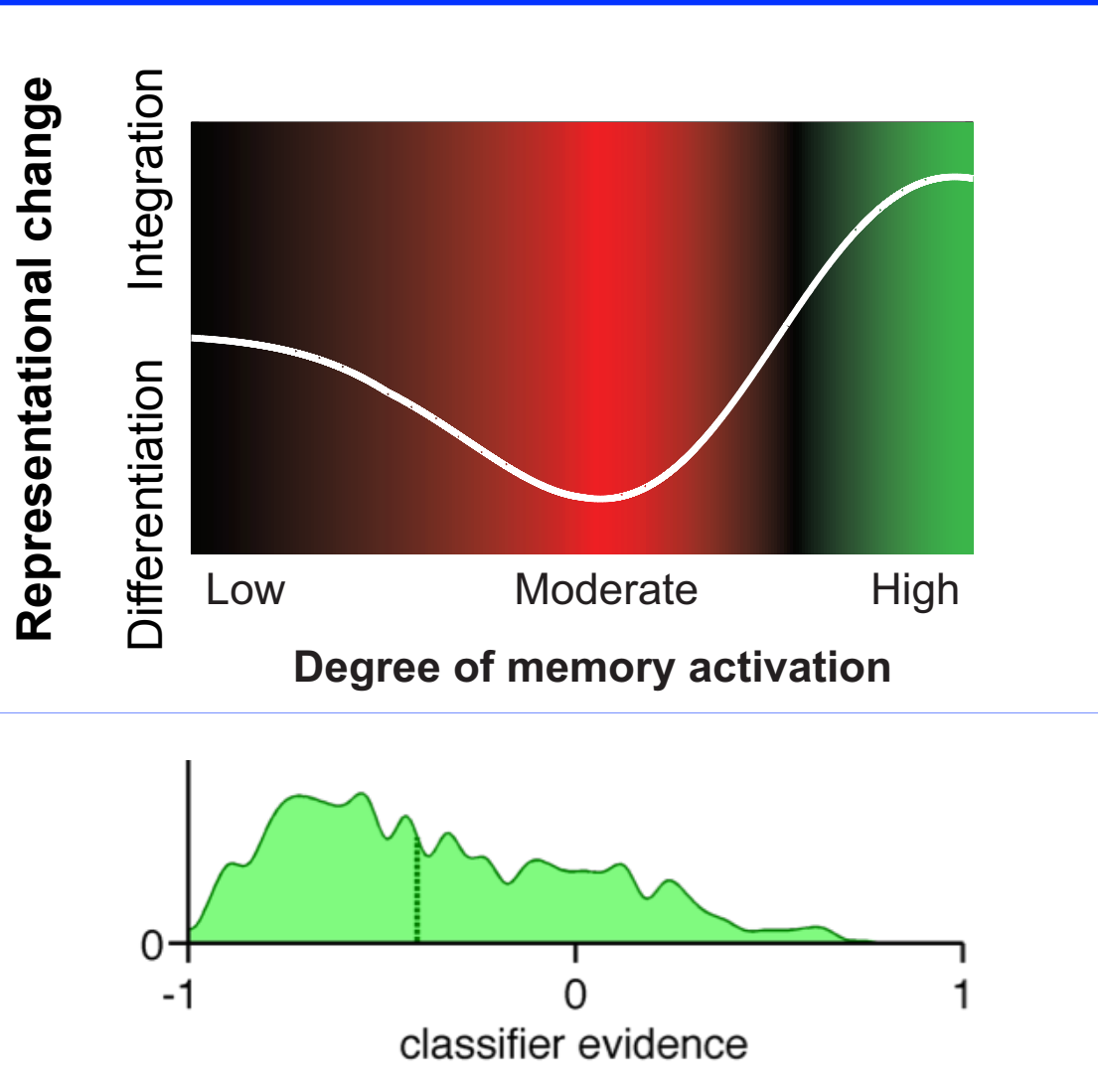
resulting weight changes



Further Tests of Differentiation Prediction

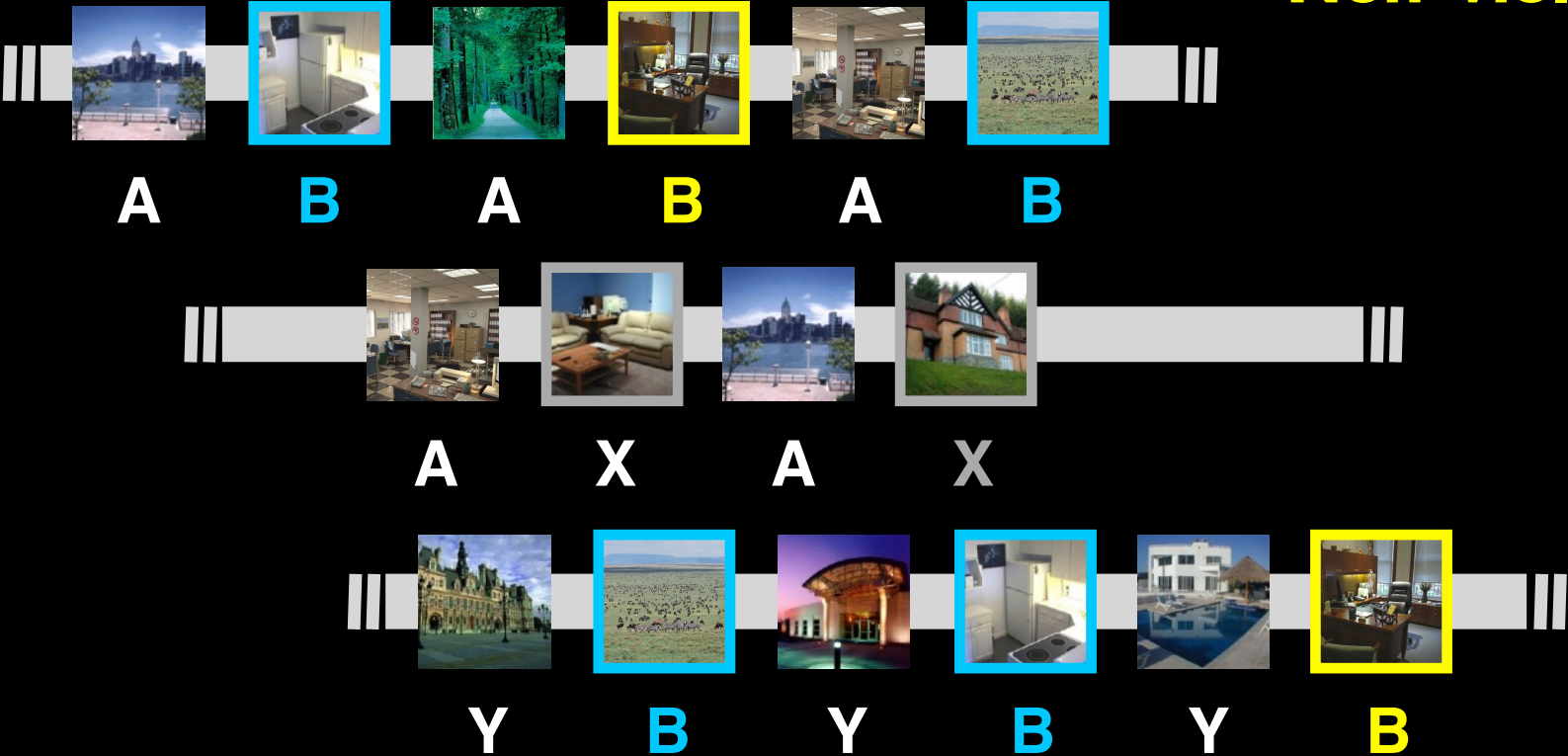
- To test the prediction that differentiation is activation-dependent, Kim, Norman, & Turk-Browne (2017, *J. Neurosci.*) used a statistical learning paradigm

Further Tests of Differentiation Prediction



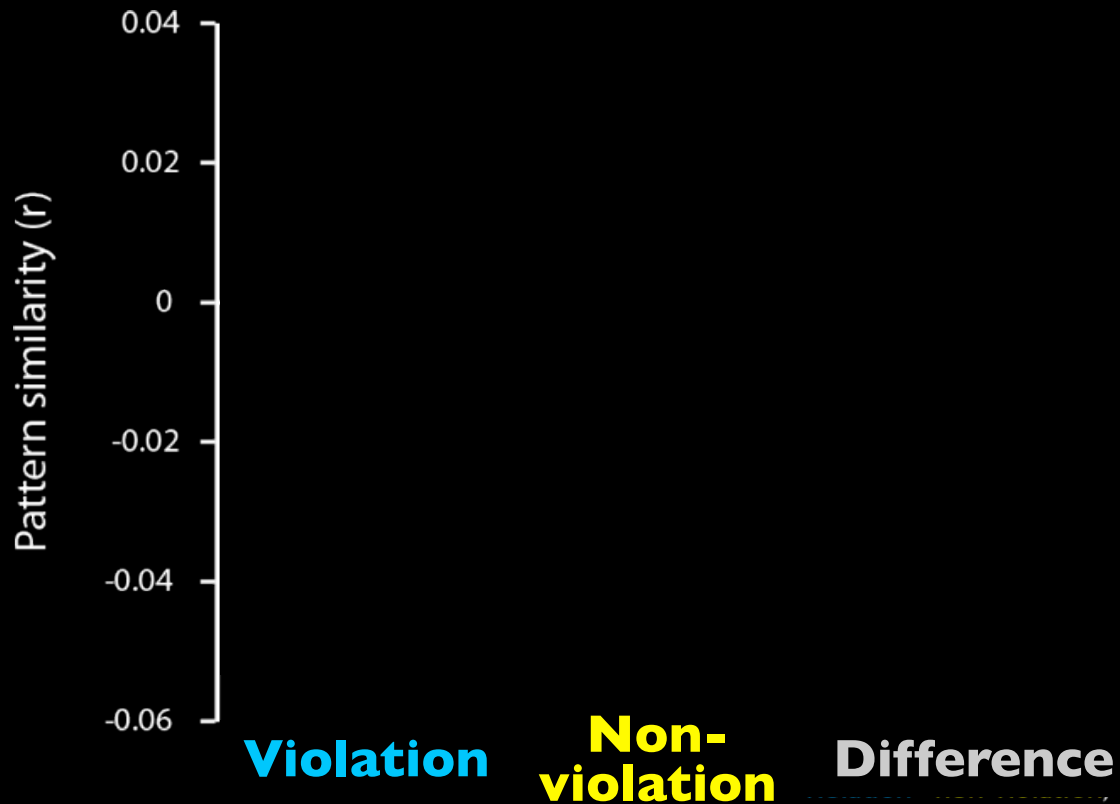
Design

Violation
Non-violation



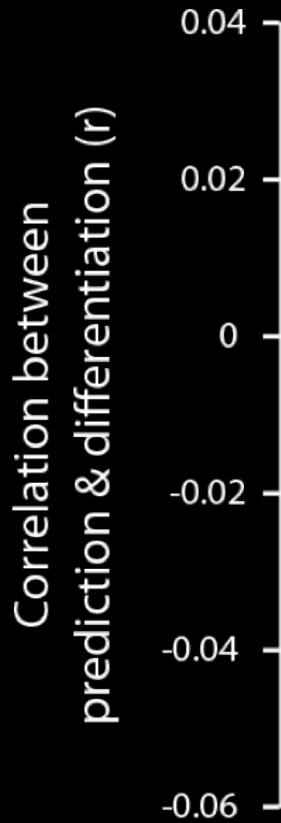
Results

Pre-post pattern similarity between A and B in hippocampus



Results

Relationship of B item prediction during violation to pre-post AB pattern similarity

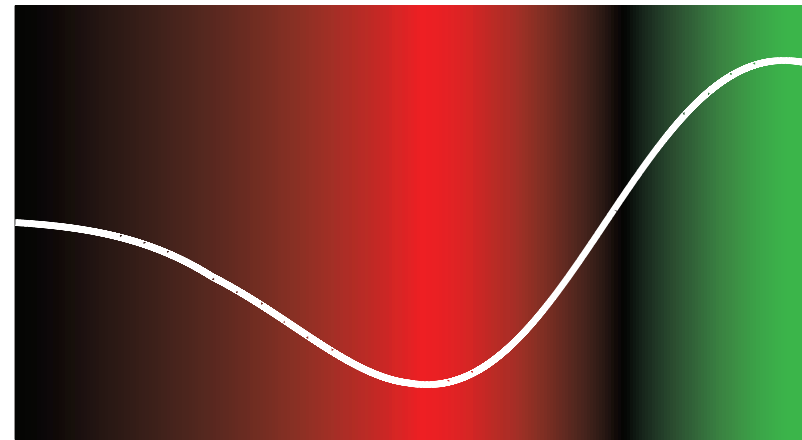


Further Tests of Differentiation Prediction

- Kim, Norman, & Turk-Browne (2017, *J. Neurosci.*) showed activity dependence of differentiation, but they did not show the “full U”
- In our next study, we set out to do this

Representational change

Integration
Differentiation

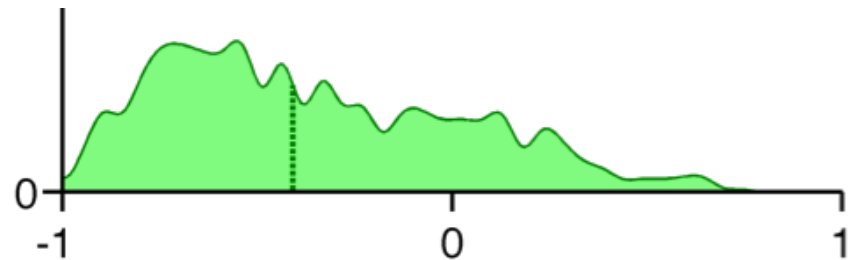


Low

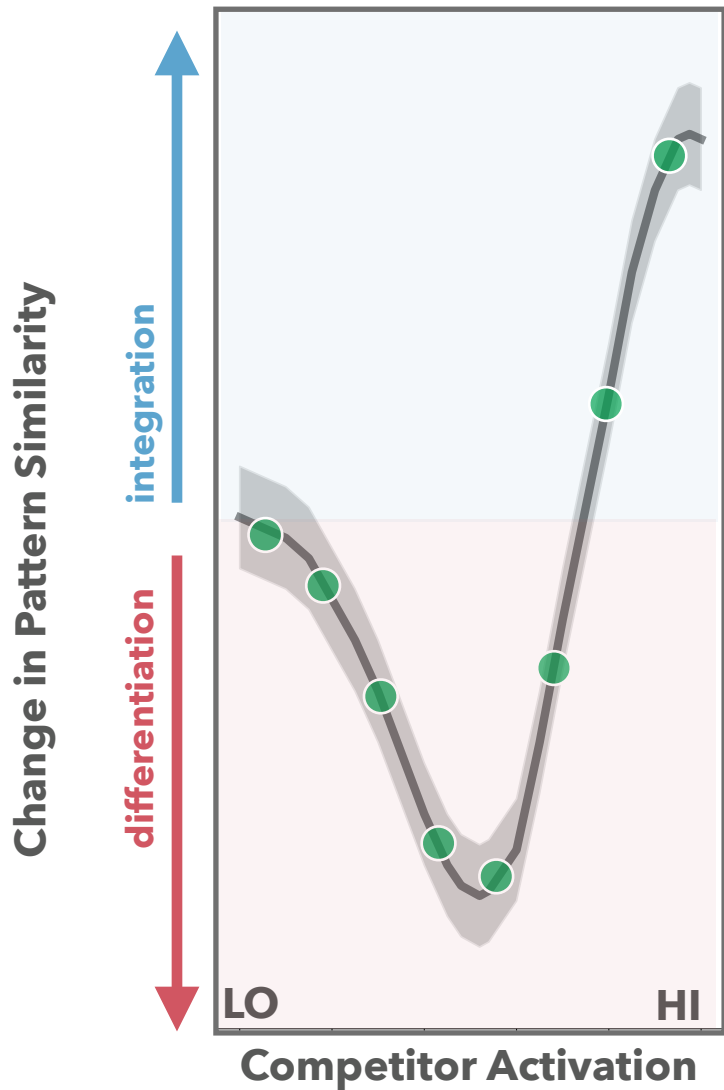
Moderate

High

Degree of memory activation



classifier evidence



How do we sample the full activity continuum from low to high?

Key idea: Parametrically manipulate the **visual similarity** of the competing items

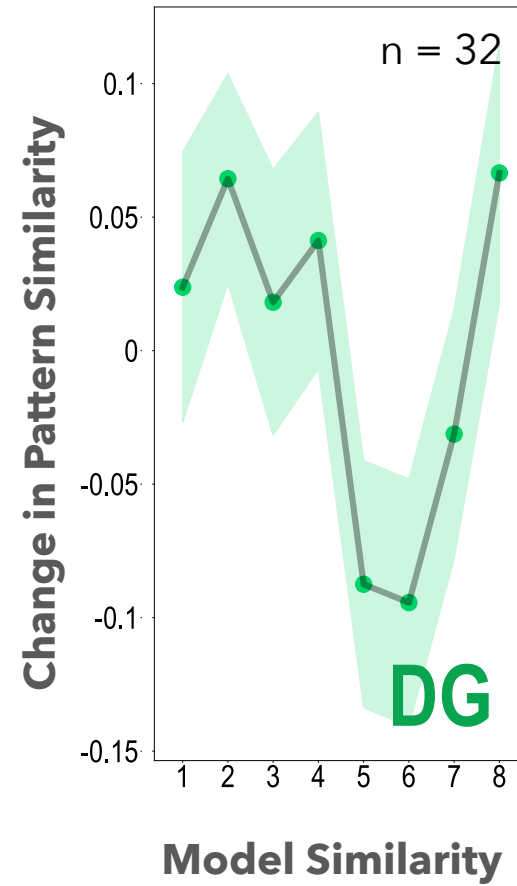
Higher levels of visual similarity
=> more representational overlap
=> more competitor activation

Kim et al. (2017) used unrelated scene images

By increasing visual similarity, we can access higher levels of competitor activation and trace out the full U

Slightly different logic from previous studies: Instead of **measuring** competitor activation, **manipulate** it in a graded fashion (8 levels of visual similarity)





Solving Puzzles Relating to Representational Change



Victoria
Ritvo



Alex
Nguyen



Nick
Turk-
Browne

Solving Puzzles Relating to Representational Change

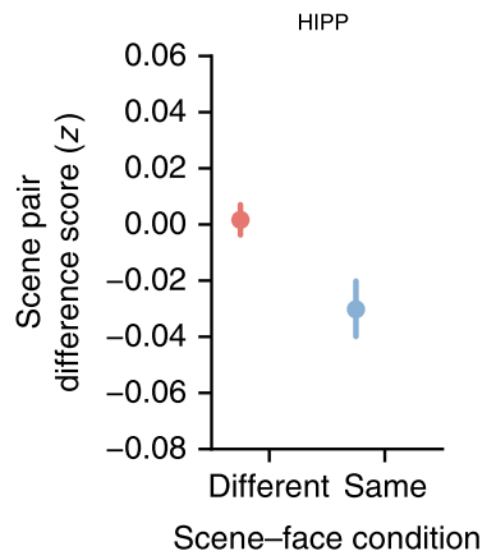
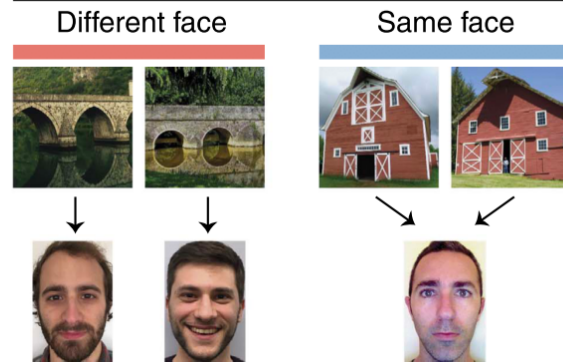
- Supervised learning models that I mentioned earlier adjust internal representations to minimize prediction error
- According to these models:
 - Stimuli with **similar predictive consequences** end up with more similar internal representations
 - Stimuli with **different predictive consequences** end up with more distinct (differentiated) internal representations
- Lots of fMRI data consistent with this (e.g., Schapiro et al., 2013, 2016; Tompary & Davachi, 2017)

Solving Puzzles Relating to Representational Change

- At the same time, some recent fMRI findings have challenged the supervised-learning account (Favila et al., 2016; Schlichting et al., 2015; Molitor et al., 2021)
- All of these findings have the property that linking two stimuli to the same associate makes their representations **less** similar, not **more** similar
 - for related findings, see Chanales et al. (2017); Dimsdale-Zucker et al. (2018); Ballard et al. (2019); Wanjia et al. (2021); Zeithamova et al. (2018); Jiang et al. (2020); Fernandez et al. (2023)

(a)

Scene-face learning



Solving Puzzles Relating to Representational Change

Trends in Cognitive Sciences

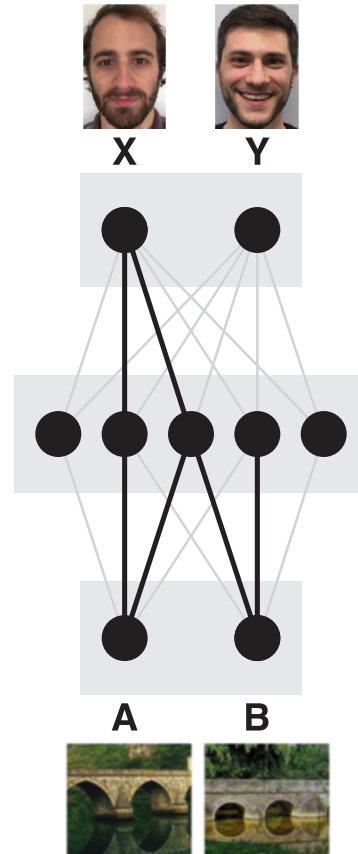
CellPress
REVIEWS

Opinion

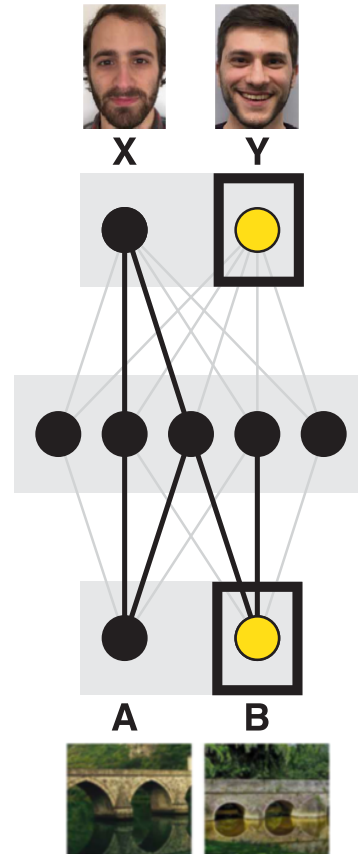
Nonmonotonic Plasticity: How Memory Retrieval Drives Learning

Victoria J.H. Ritvo,¹ Nicholas B. Turk-Browne,² and Kenneth A. Norman^{1,3,*}

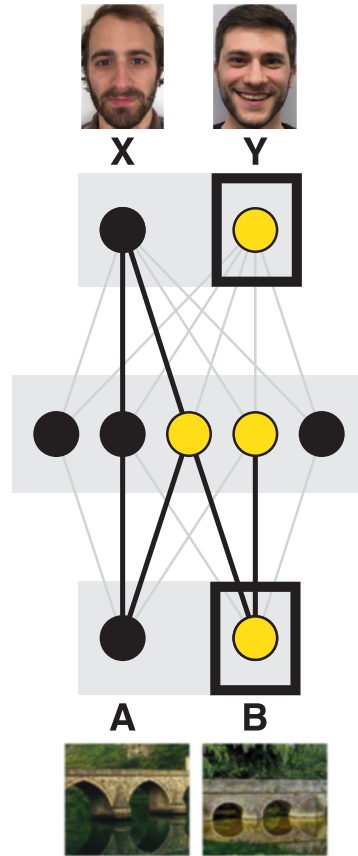
Different Face Condition



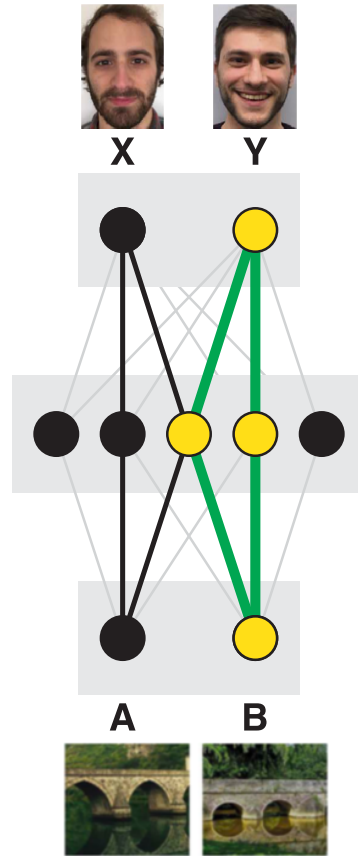
Different Face Condition



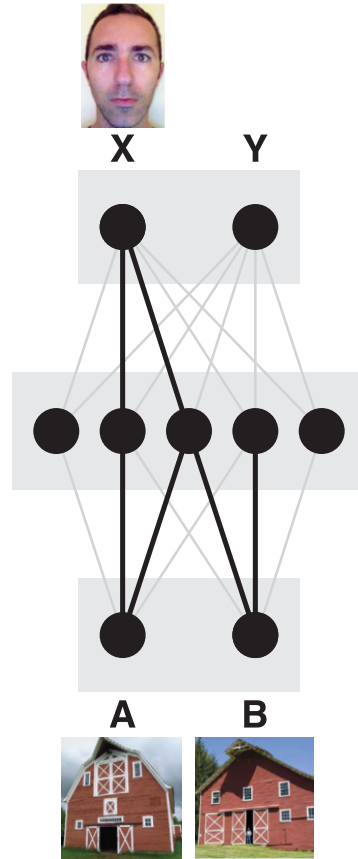
Different Face Condition



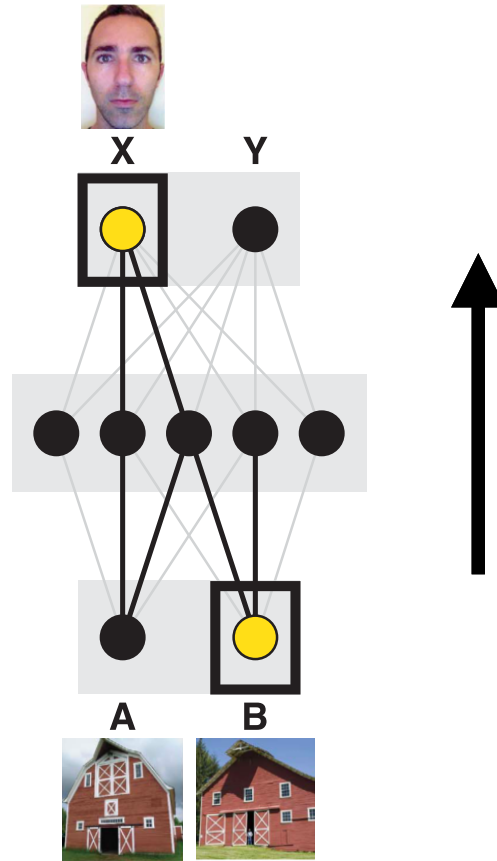
Different Face Condition



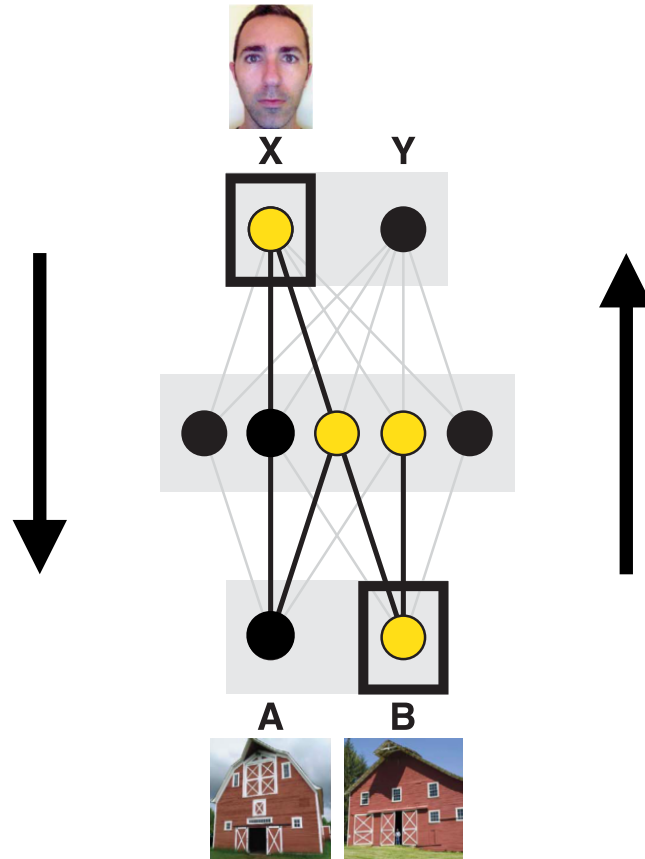
Same Face Condition



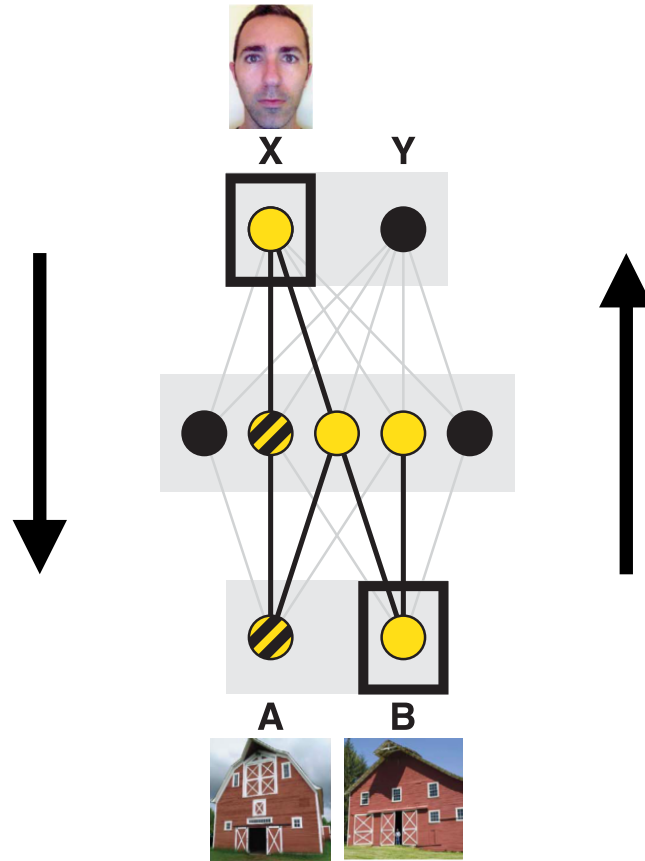
Same Face Condition



Same Face Condition

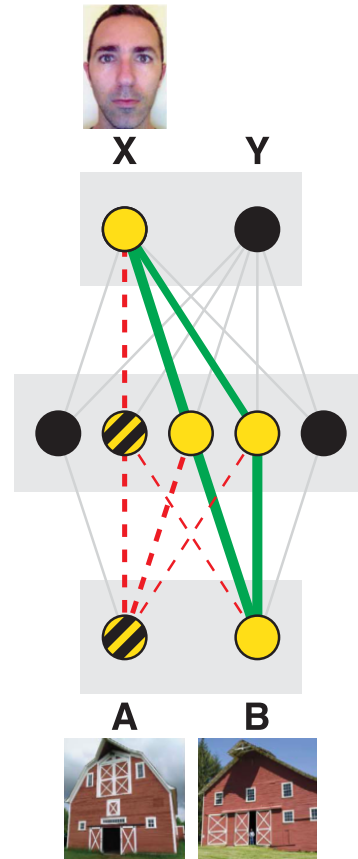


Same Face Condition



Moderate Competitor Activity

Same Face Condition



Research Article

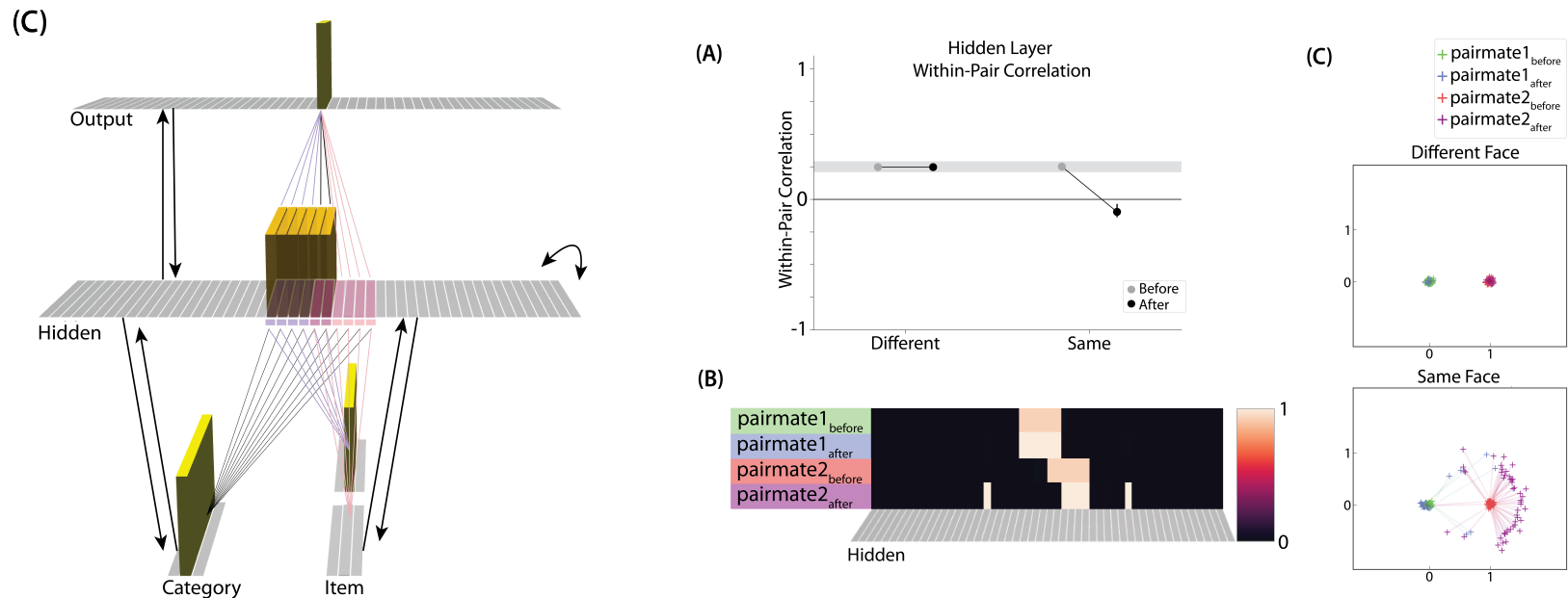
Neuroscience

A neural network model of differentiation and integration of competing memories

Victoria JH Ritvo, Alex Nguyen, Nicholas B Turk-Browne, Kenneth A Norman

Department of Psychology, Princeton University, United States; Princeton Neuroscience Institute, Princeton University, United States; Department of Psychology, Yale University, United States; Wu Tsai Institute, Yale University, United States

Sep 25, 2024 • <https://doi.org/10.7554/eLife.88608.3>



Relating the NMPH and Supervised Learning

- Supervised learning can't explain some fMRI data on representational change (e.g., Favila et al., 2016)
- NMPH can fill in this gap
- Importantly, the NMPH is not meant to *replace* supervised learning
 - The ability to learn to predict specific outcomes from inputs is too valuable to replace
- Rather, we think of NMPH learning as *supplementing* supervised learning

Relating the NMPH and Supervised Learning

- **From a functional perspective**, why do we need both NMPH and supervised learning?
- Supervised learning can train complex neural nets to make useful predictions...
- ... but these networks can suffer from major problems relating to **competition**
- If the network vacillates between nearby states without fully settling into one, this can prevent the network from acting decisively in response to the current stimulus

Relating the NMPH and Supervised Learning

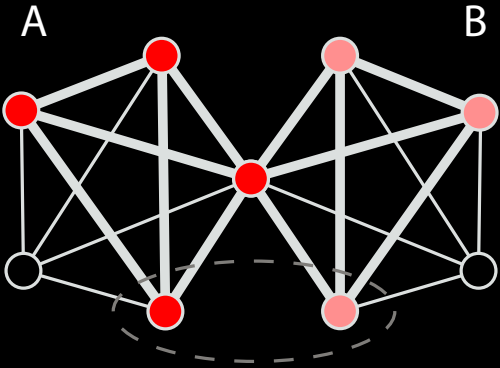
- Key claim: NMPH does **housekeeping** in the brain
- When memories compete, NMPH restructures memories:
 - if the competing memory activates **moderately**, NMPH pushes it away, so it competes less
 - if the competing memory activates **strongly**, NMPH integrates the memories together, so they compete less

Differentiation and the Hippocampus

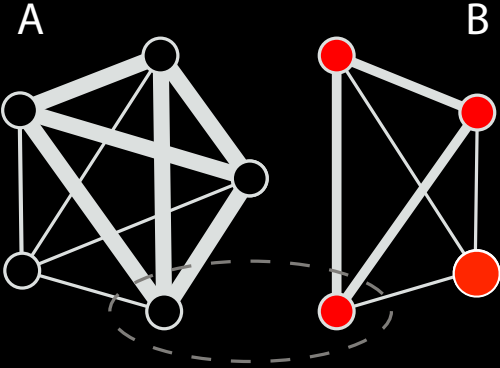
- Common thread: Differentiation effects are most reliably observed in the hippocampus (although they are sometimes observed elsewhere)
- What explains this?
- One factor: hippocampal representations are sparse (not many neurons can be active at once)...
- ... so it's hard to activate competitors strongly
- Put another way: Activity might be “confined” to the moderate activation range that gives rise to differentiation
- Another factor: Hippocampus has a **large learning rate**

Differentiation after moderate activation

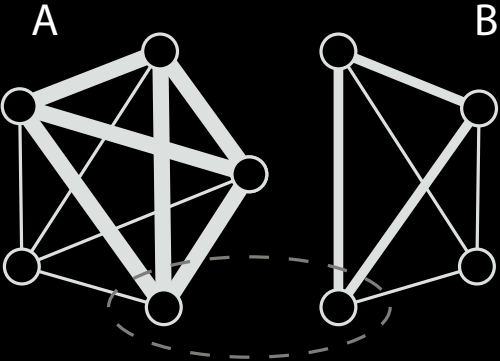
Competition trial



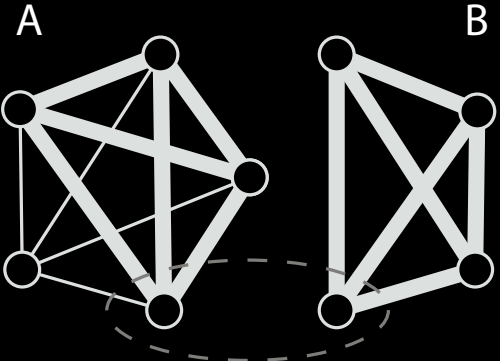
Restudy trial



resulting weight changes

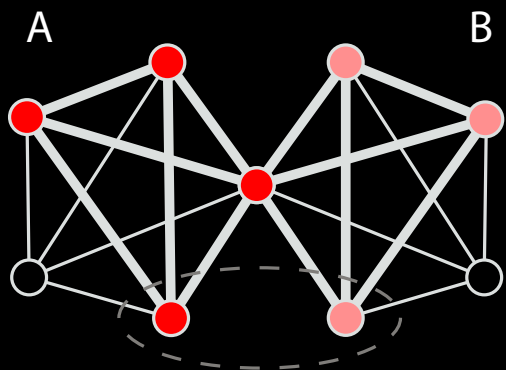


resulting weight changes

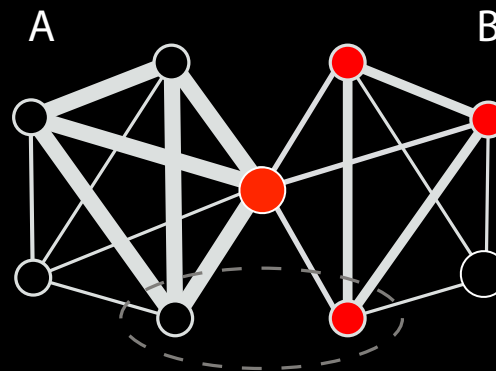


What happens with a smaller learning rate

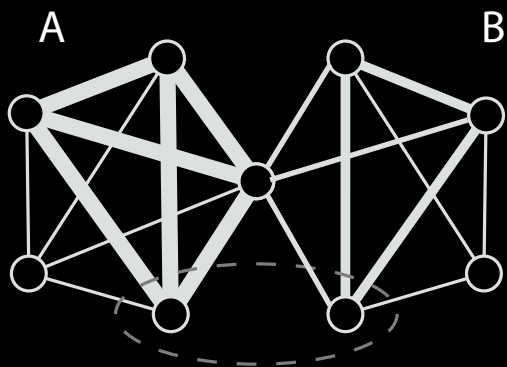
Competition trial



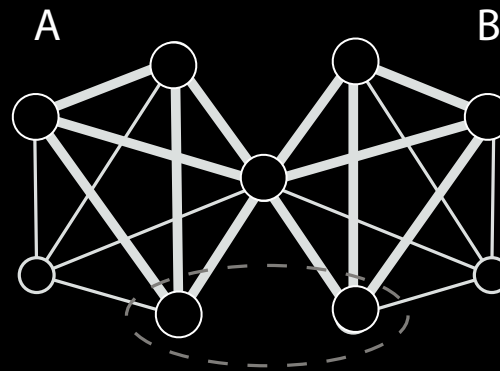
Restudy trial



resulting weight changes



resulting weight changes



What is the **purpose** of hippocampal differentiation?

- Standard story about hippocampal functioning: It is wired to try to assign **orthogonal** representations to similar stimuli (“pattern separation”; Yassa & Stark, 2011; Marr, 1971)
- Key point: The hippocampus’ pattern separation abilities are limited — when inputs to the hippo are very similar, there might end up being enough overlap to cause competition and degrade performance
- In this case, it is useful to have an extra “emergency” mechanism that **completely** removes overlap between the hippocampal representations
- ... and that is what we think the NMPH does

What is the **purpose** of hippocampal differentiation?

- You wouldn't want to use this “emergency differentiation” mechanism routinely
- If the hippocampus used completely non-overlapping codes routinely it would rapidly run out of space
- But we think this is a useful mechanism to deploy on an “as needed” basis to mitigate competition

What is the **purpose** of hippocampal differentiation?

- Rapidly-formed, differentiated hippocampal codes can act as a kind of “pry bar” to help pull apart representations elsewhere in the brain (e.g., cortex)
- When studies report differentiation effects in cortex, we think this may reflect the influence of projections from differentiated representations in the hippocampus
- By replaying memories over time, hippocampus can teach cortex to distinguish these representations on its own

Outline

- Situate NMPH relative to other kinds of learning
- Implications of NMPH for memory weakening
 - Key prediction: Moderate activation leads to weakening of competing memories
- Implications of NMPH for the similarity structure of memories
 - Key prediction: Moderate activation leads to differentiation of competing memories
- **Current directions**
 - Role of NMPH in learning during sleep
 - Using neurofeedback to promote discrimination learning

NMPH Learning During Sleep

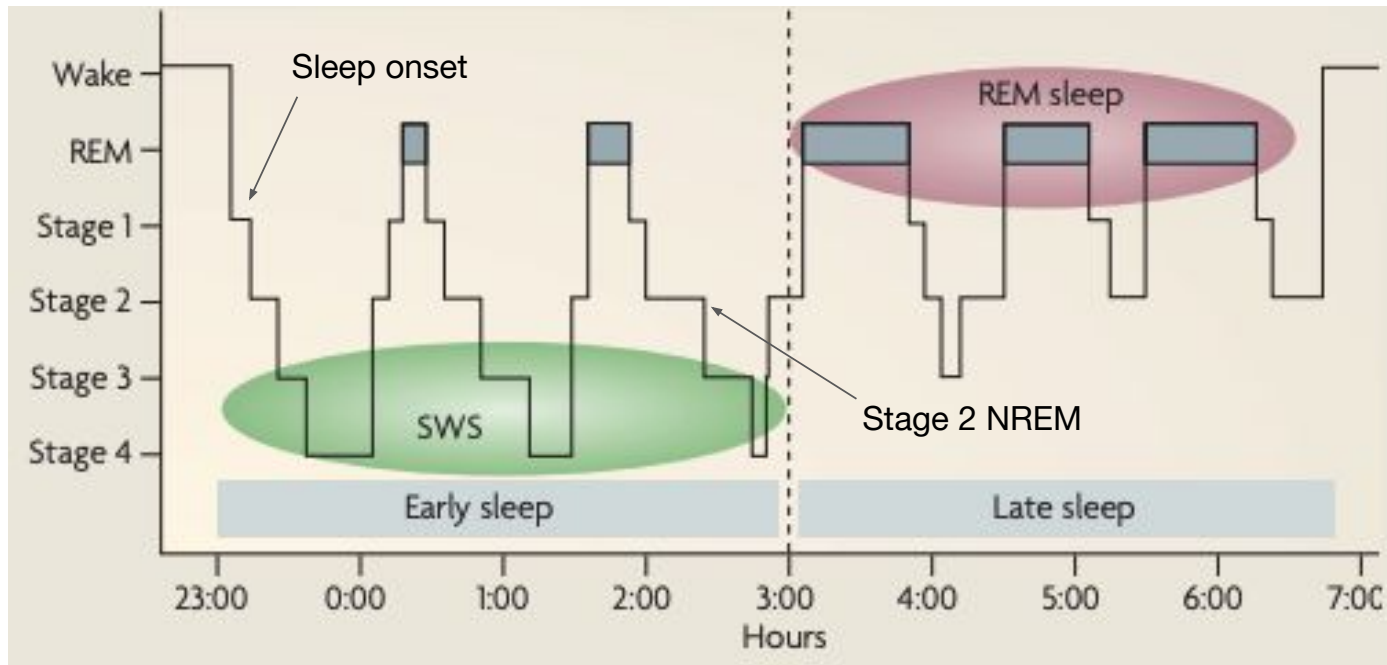
- During sleep, the brain is cut off from the world and sensory input can not provide a “target” for supervised learning
- ... so the brain needs to rely on unsupervised learning rules like the NMPH
- Sleep consists of multiple, neurophysiologically distinct stages

Stages of Sleep

- 3 stages of non-REM (NREM) sleep
 - Stage 1
 - Stage 2
 - Stages 3 and 4: together known as slow wave sleep (SWS)
- Rapid eye movement (REM) sleep



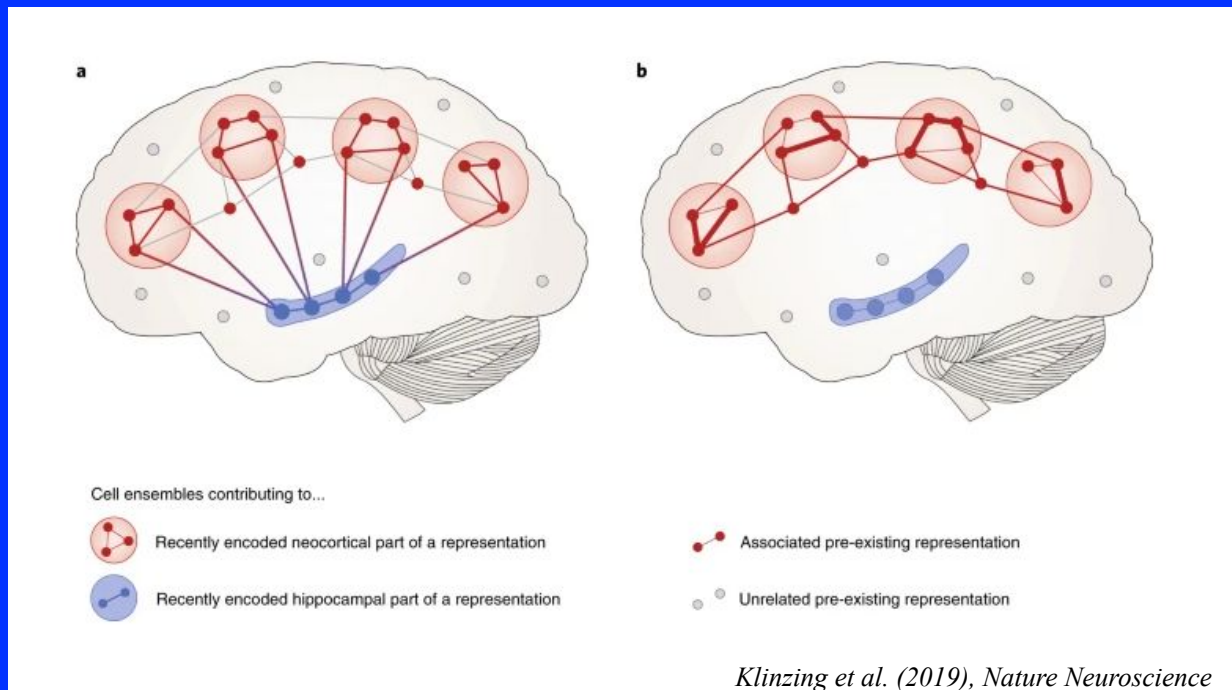
A Good Night's Sleep



- SWS is more prevalent toward the beginning of the night
- REM is more prevalent toward the end of the night

Slow wave sleep (SWS)

- Strong coupling between hippocampus and cortex
- Strong evidence for **neural replay** of recently experienced events
- Thought to be involved in **systems consolidation**

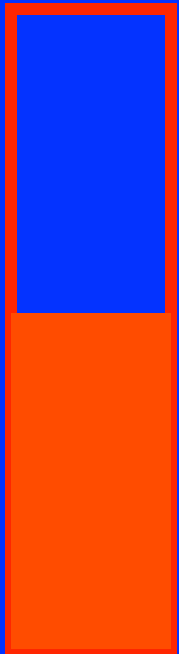


REM sleep

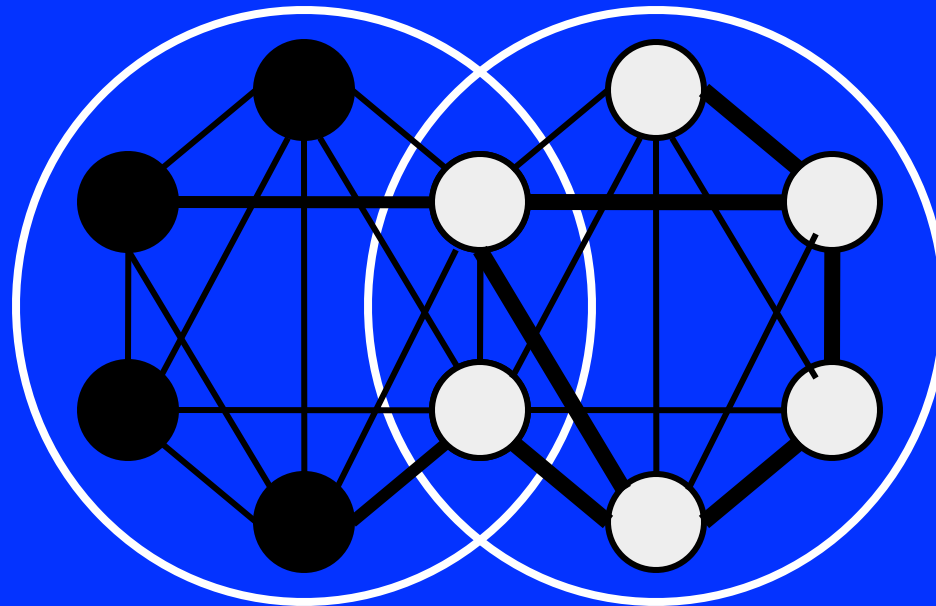
- Less coupling between hippocampus and cortex
- Less evidence for replay (but still some)
- Function is less well understood
- Hypothesis: **REM sleep** may play a key role in representational change
 - During REM, the brain settles on newly acquired memories and activation is allowed to spread outward
 - Representations are adjusted using NMPH
 - Theta oscillations during REM may play a role in promoting plasticity (Boyce et al., 2016; Poe et al., 2000)

Oscillations and Learning

Inhibition
meter

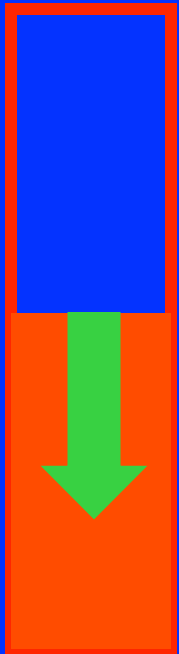


Apple Pear

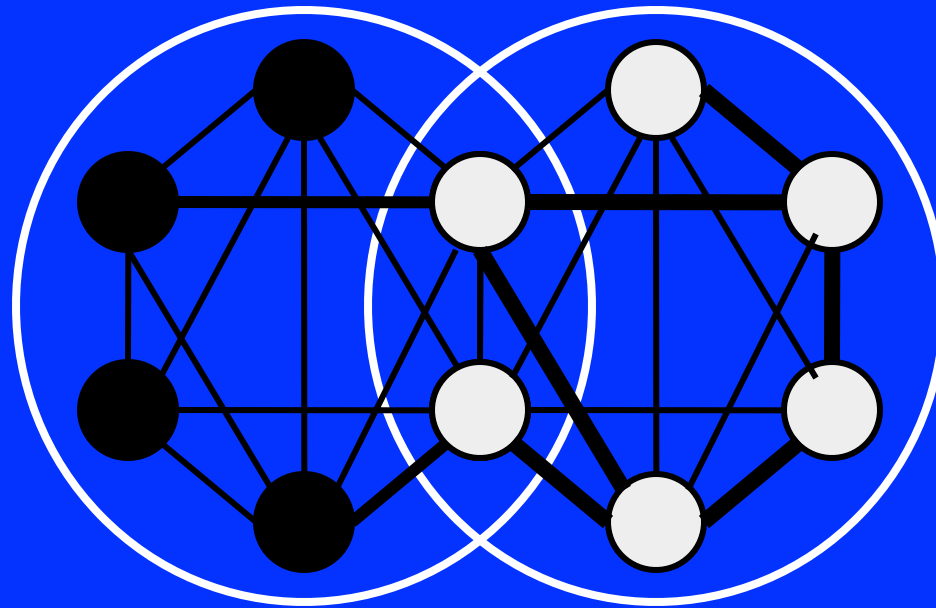


Oscillations and Learning

Inhibition
meter

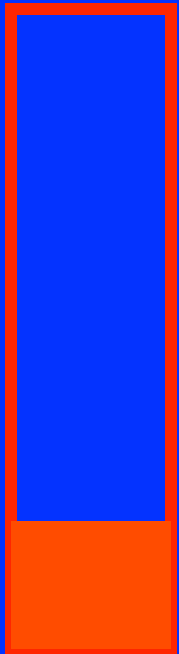


Apple Pear

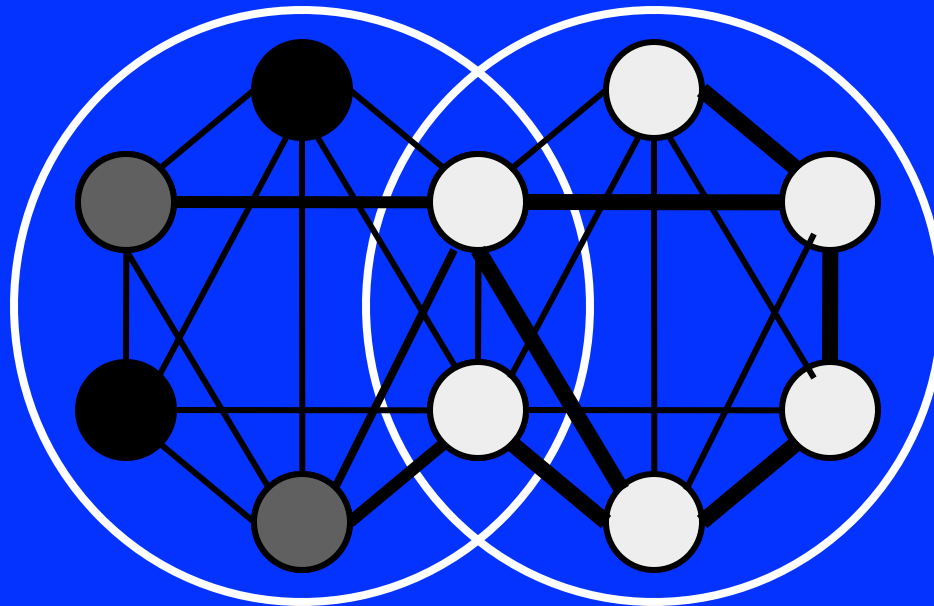


Oscillations and Learning

Inhibition
meter

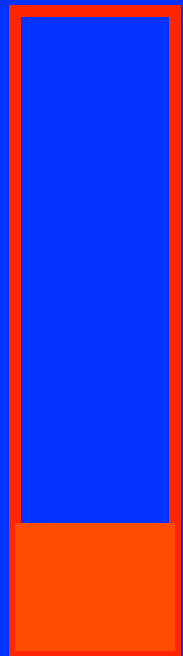


Apple Pear

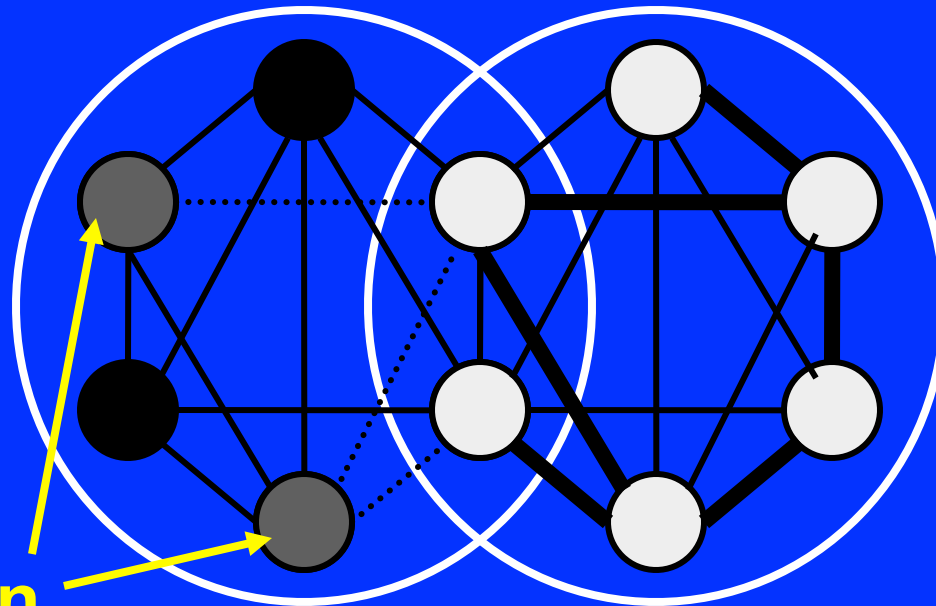


Oscillations and Learning

Inhibition
meter



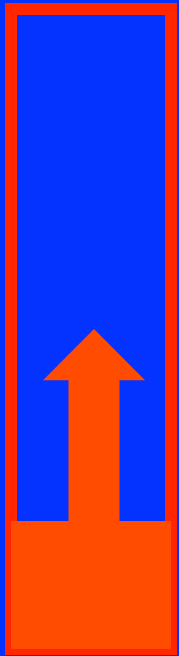
Apple Pear



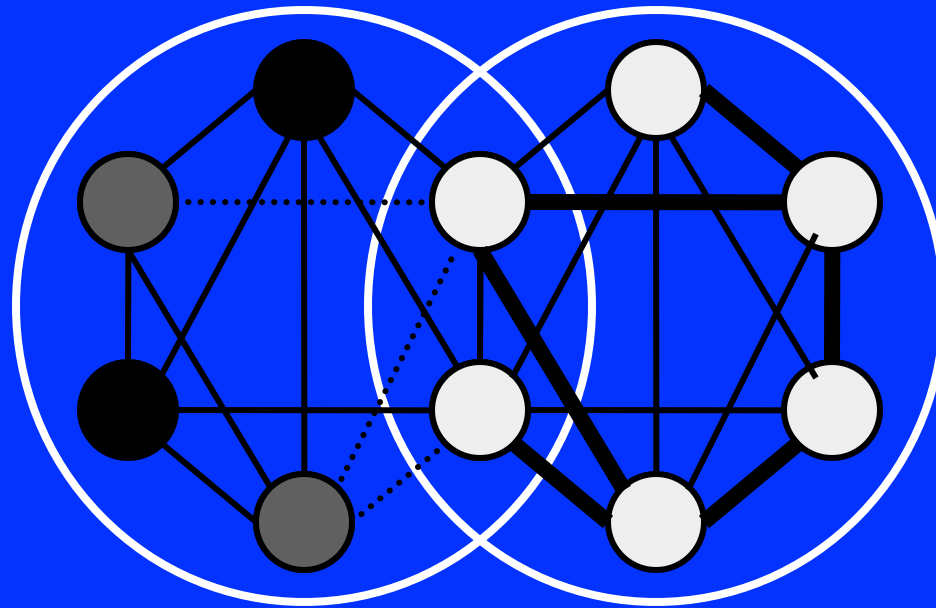
Weaken

Oscillations and Learning

Inhibition
meter

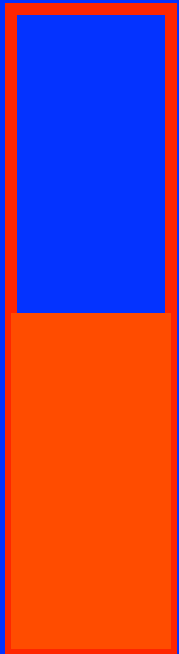


Apple Pear

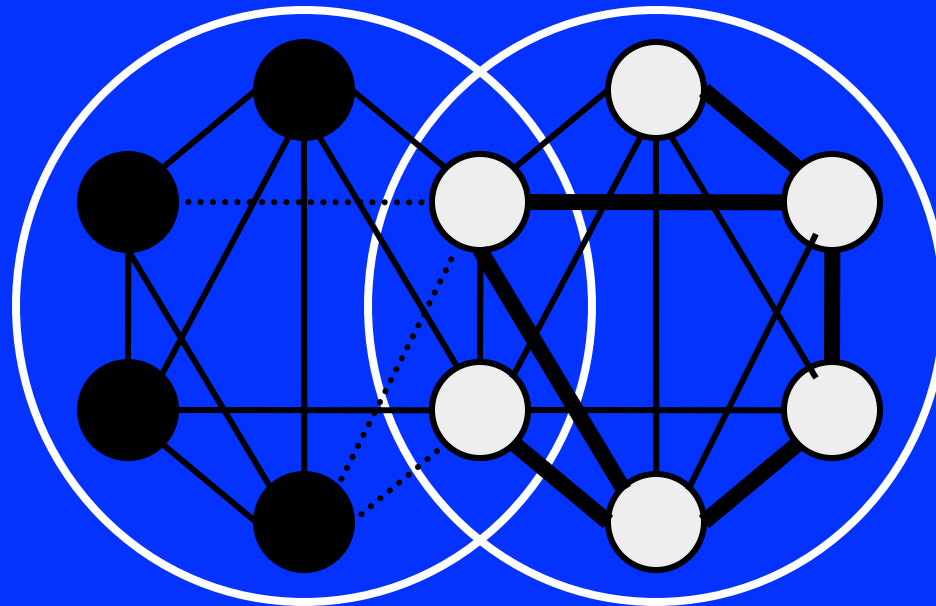


Oscillations and Learning

Inhibition
meter



Apple Pear



NMPH Learning During Sleep

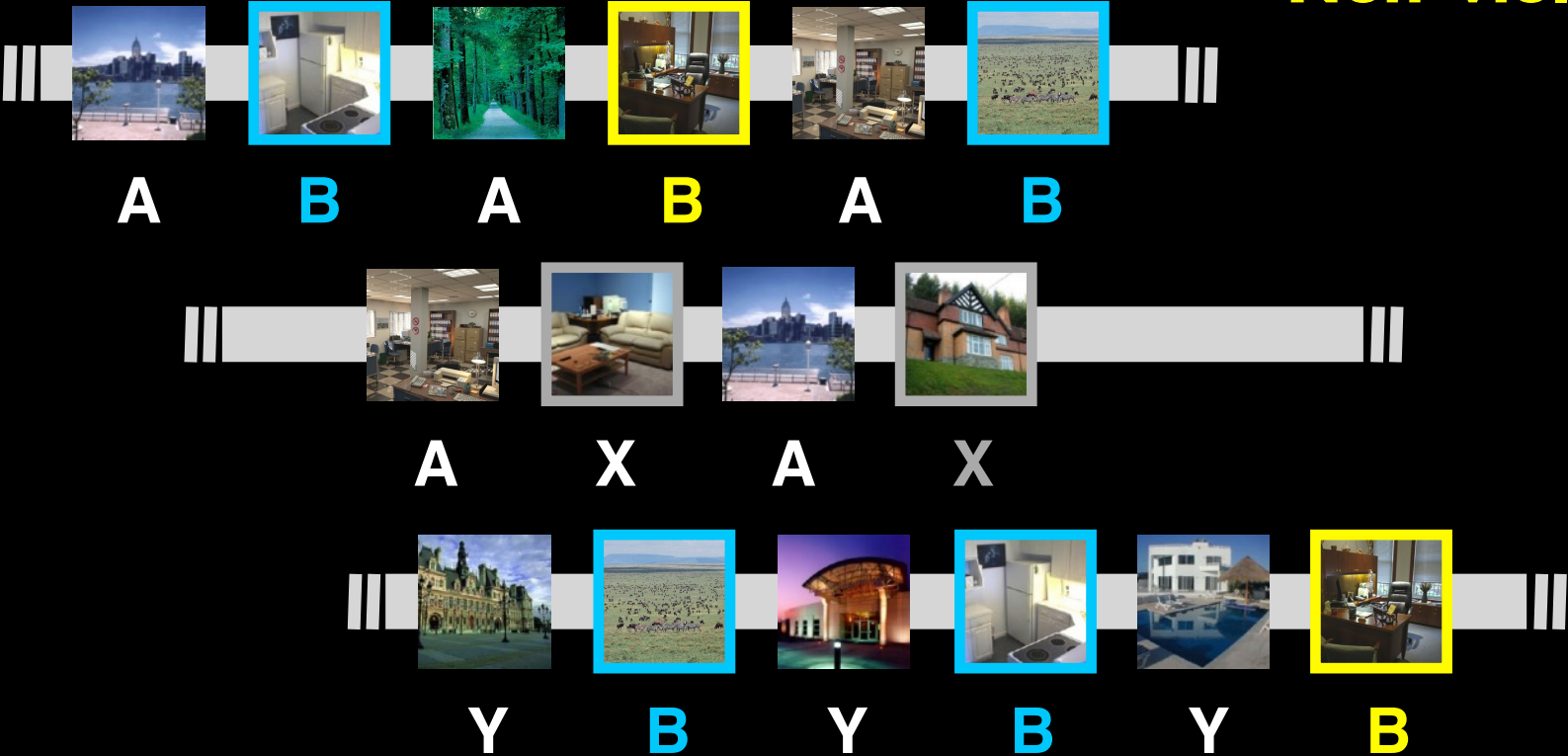
- This ability to re-evoked competitors **repeatedly** through oscillations suggests REM may be particularly useful for driving representational change
- In some situations, waking plasticity alone may not be sufficient to cause differentiation...
- ... in which case you may also need REM sleep to show these effects

NMPH Learning During Sleep

- Sleep-dependent plasticity may have (inadvertently) played a role in the Kim et al. (2017) statistical learning study that I mentioned earlier

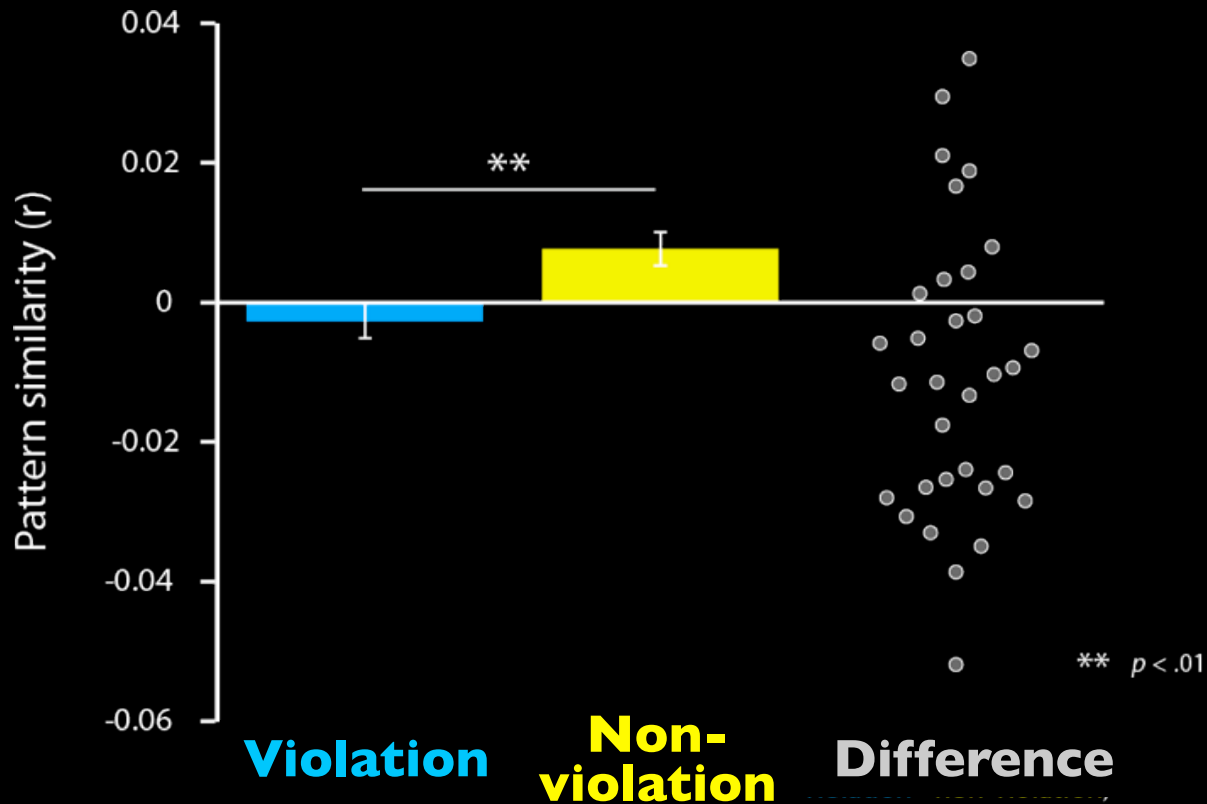
Design

Violation
Non-violation



Results

Pre-post pattern similarity between A and B in hippocampus



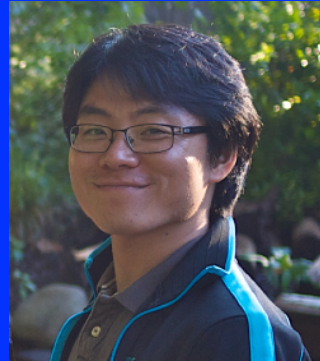
NMPH Learning During Sleep

- Crucially, there was a **1-day gap** between the statistical learning task and the imaging session where we took post-learning “snapshots” of neural representations of the scenes
- Presumably, participants slept during this 1-day gap
- That period of sleep might (or might not) have contributed to the hippocampal differentiation effect that we observed

NMPH Learning During Sleep



Lizzie
McDevitt

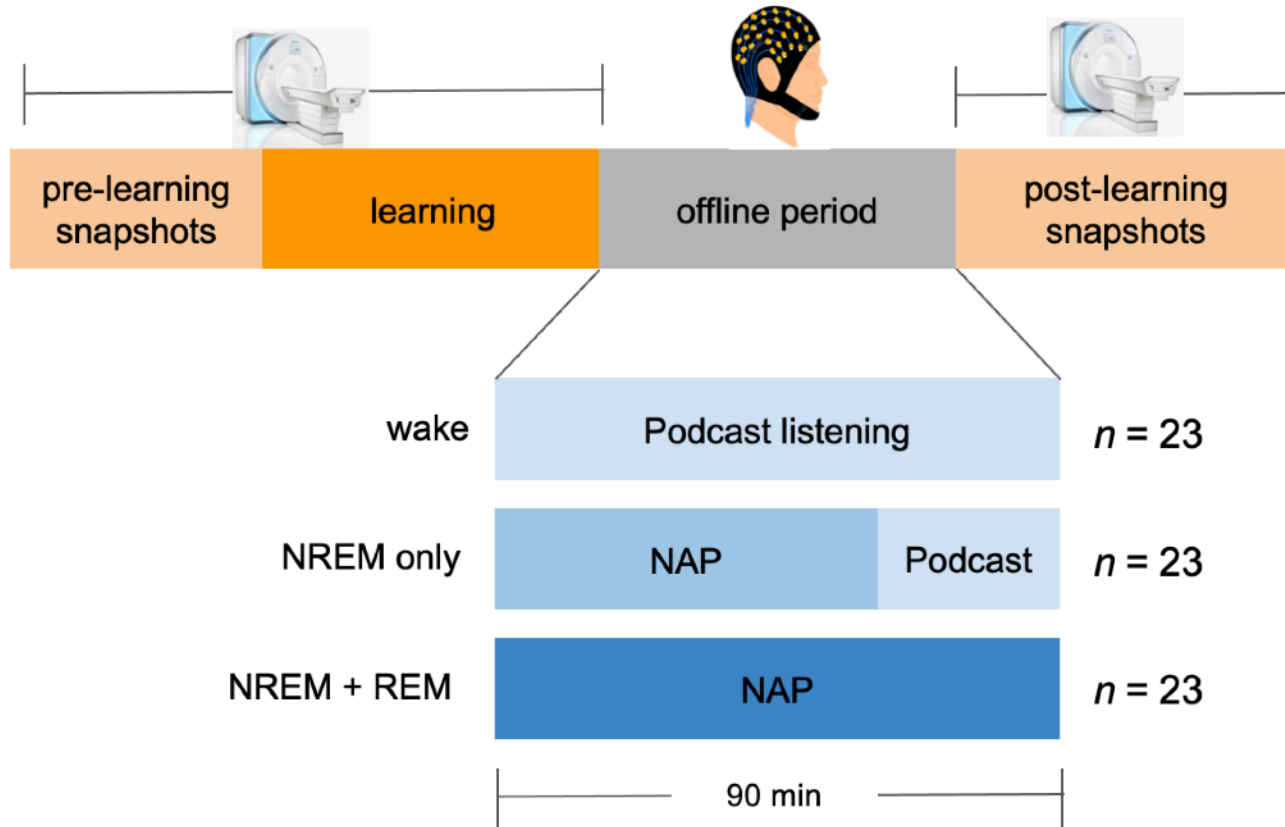


Ghootae
Kim



Nick
Turk-
Browne

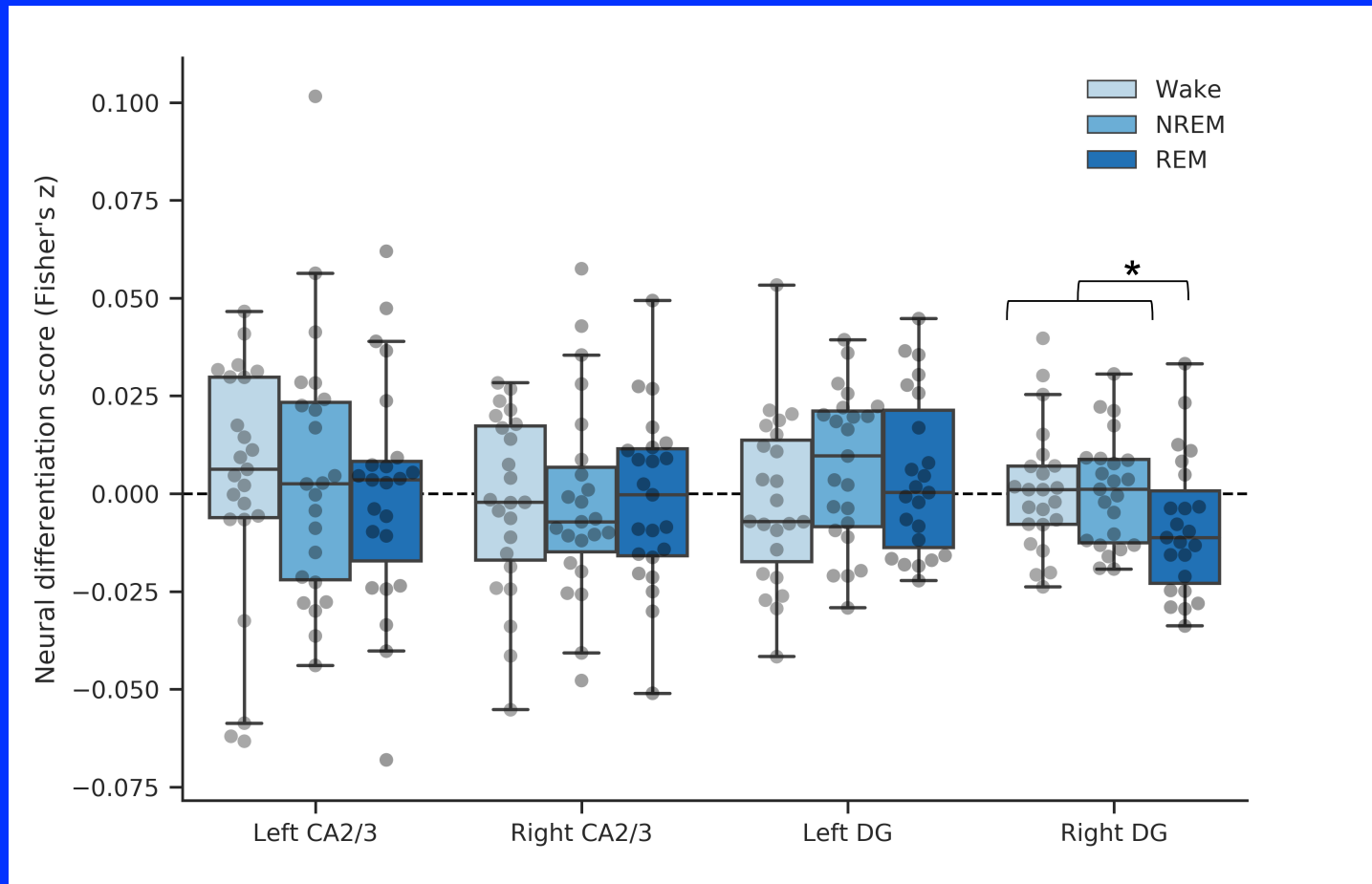
NMPH Learning During Sleep



Total $N = 69$

- Preregistered prediction: More hippo differentiation after REM

NMPH Learning During Sleep



Modeling Learning During Sleep



Anna Schapiro




PNAS

SPECIAL FEATURE

PSYCHOLOGICAL AND COGNITIVE SCIENCES
NEUROSCIENCE



A model of autonomous interactions between hippocampus and neocortex driving sleep-dependent memory consolidation

Dhairyya Singh^{a,1} , Kenneth A. Norman^{b,c} , and Anna C. Schapiro^{a,1} 

Neurofeedback and the NMPH

- Can we use neurofeedback, in concert with the NMPH, to facilitate learning of difficult tasks?
- Focus on **discrimination learning**:
 - Learning to give different responses to similar stimuli
- Discrimination learning is challenging because similar stimuli tend to **strongly coactivate** (as in Wammes et al., 2022, above)
- This strong coactivation will lead to integration (which hurts discrim. learning) instead of differentiation (which would help)
- How do we “turn the tide” from integration to differentiation?

Neurofeedback and the NMPH

- Intuition: if we can somehow engineer a situation where the to-be-discriminated stimuli **moderately** coactivate (instead of strongly coactivating)...
- ... this will trigger neural differentiation, which will facilitate differential responding to the two stimuli

Neurofeedback and the NMPH

- We set out to test this idea in a study led by Coraline Iordan (Iordan et al., 2024, PNAS)



Coraline Iordan



Jon Cohen

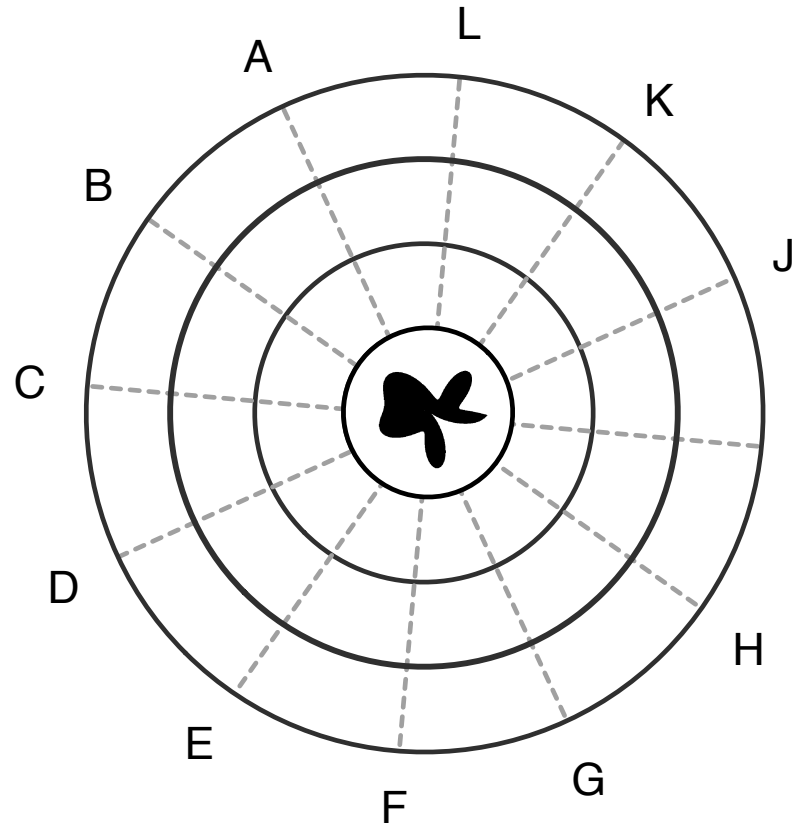


Nick Turk-Browne

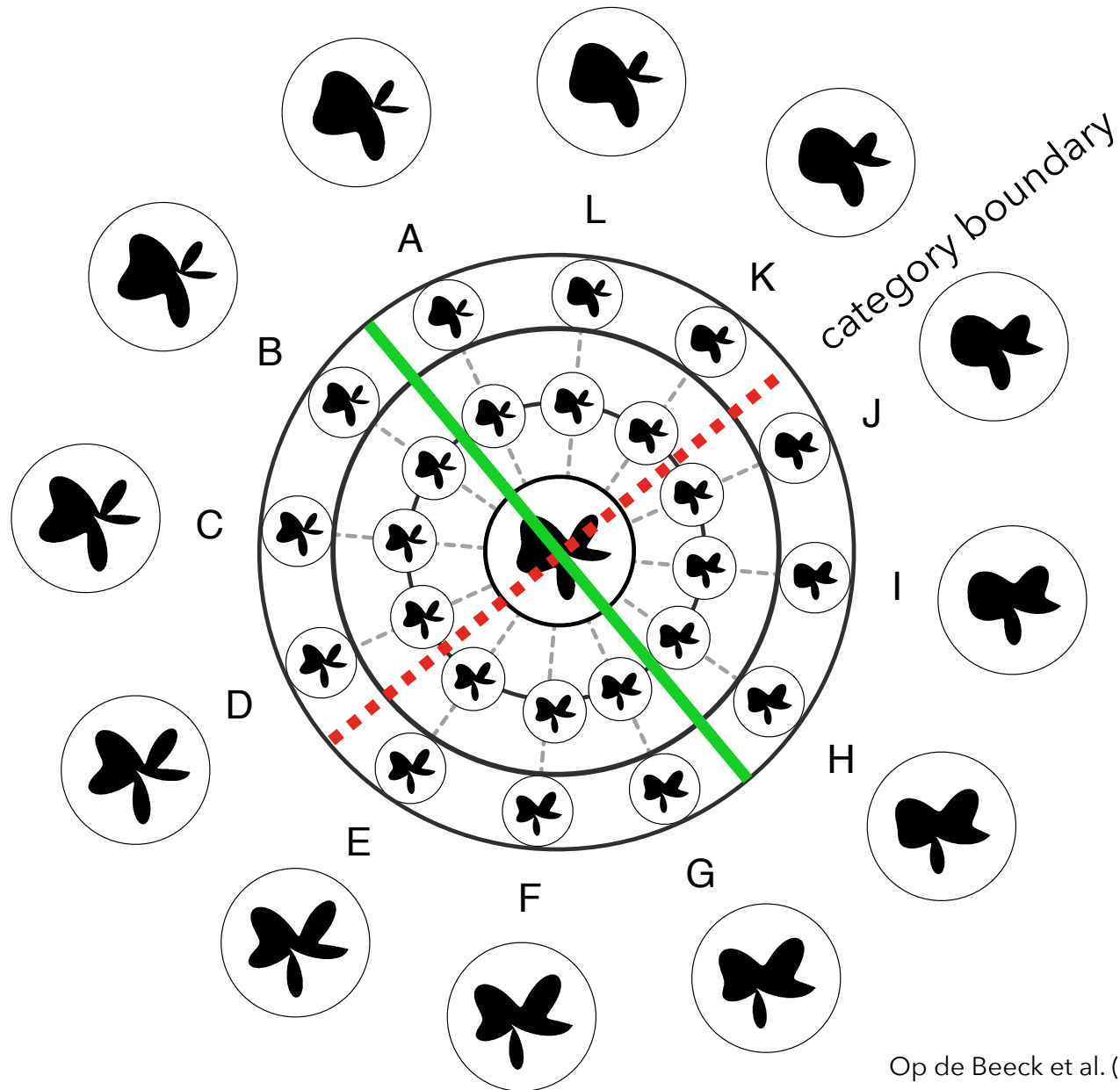


Victoria Ritvo

Stimulus Set: Fourier-Descriptor Shape Space

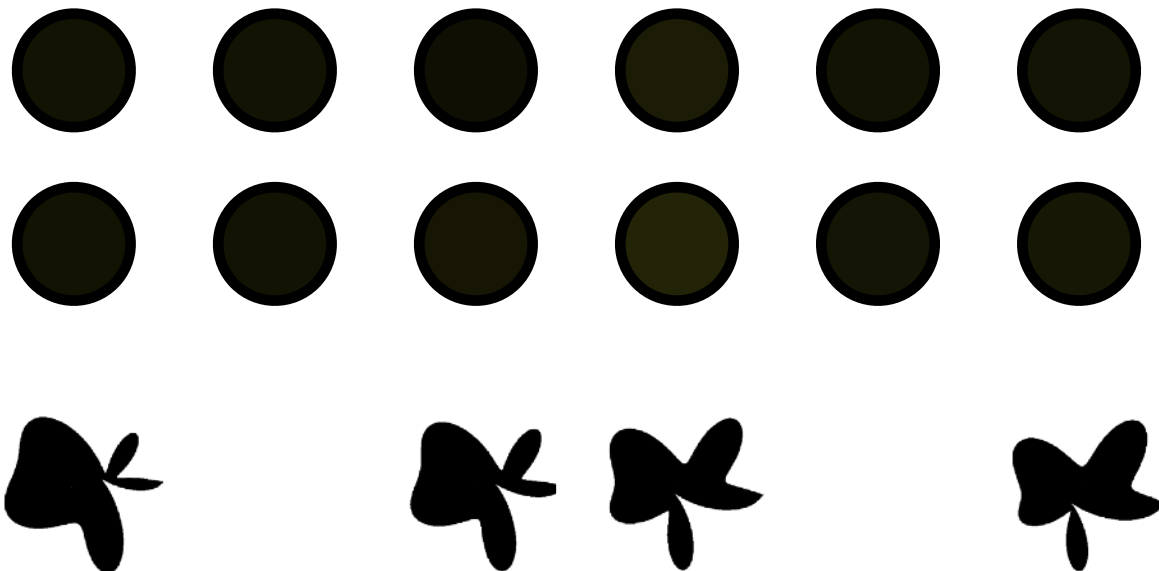


Stimulus Set: Fourier-Descriptor Shape Space

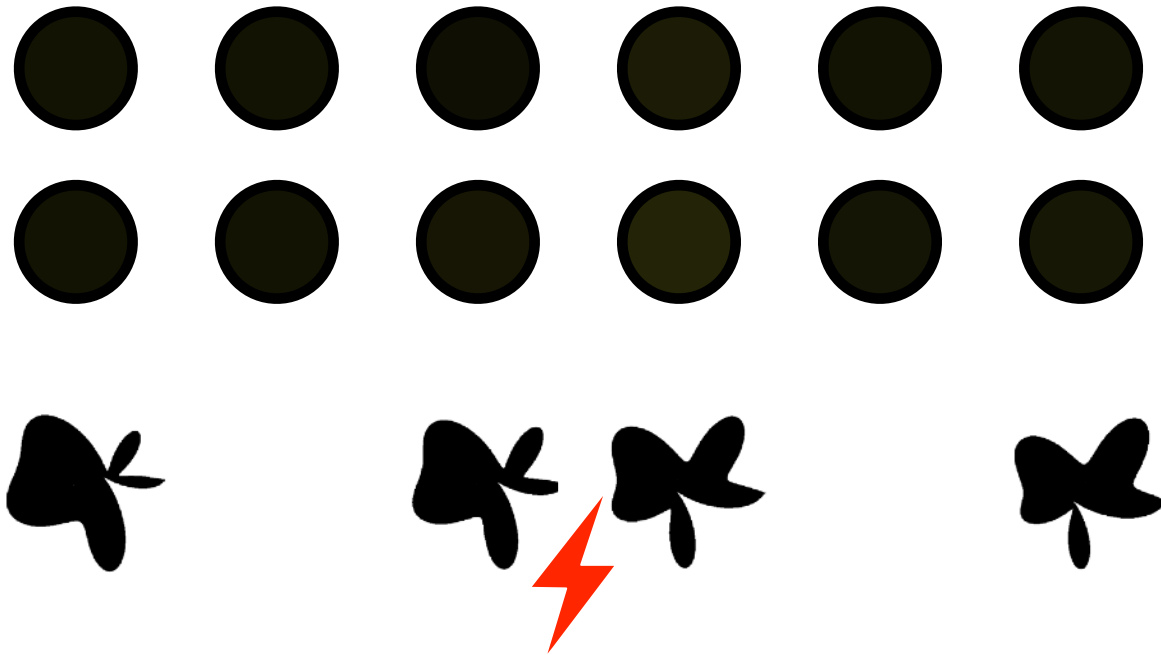


Op de Beeck et al. (2001), Kok et al. (2018)

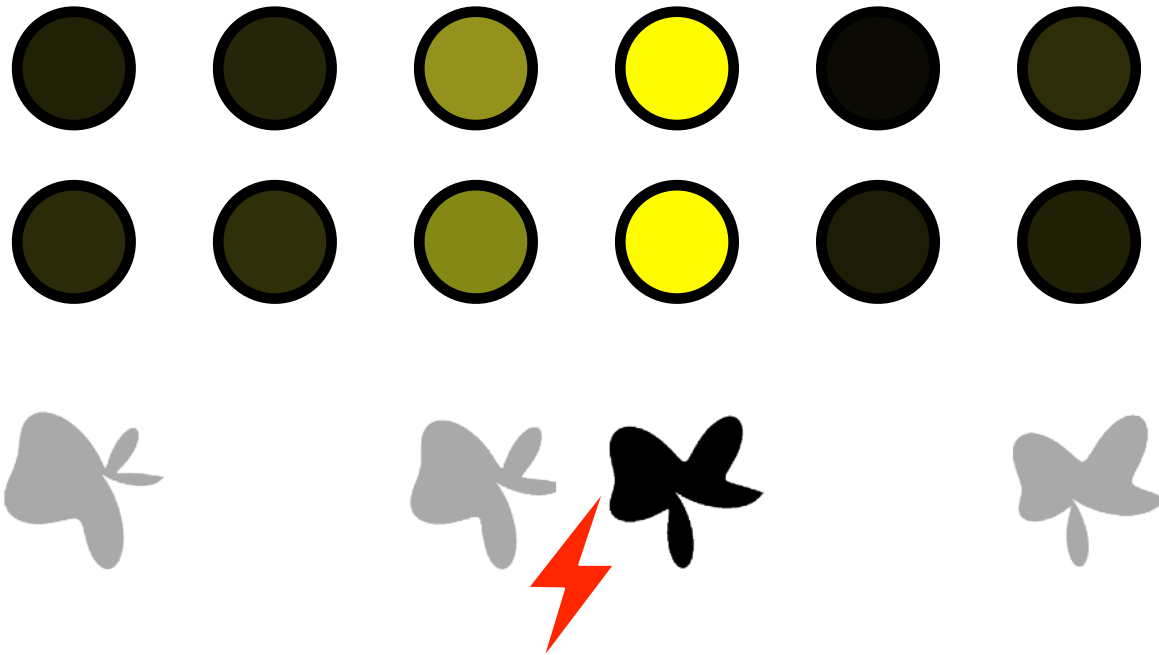
Manipulating stimulus representations with the NMPH



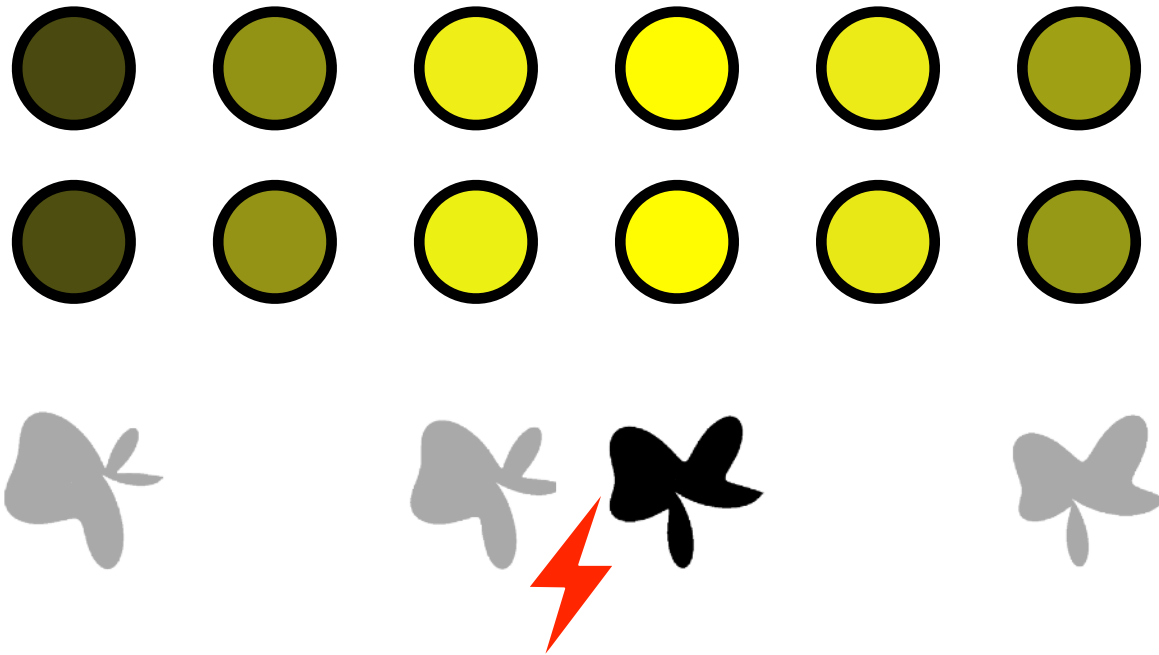
Manipulating stimulus representations with the NMPH



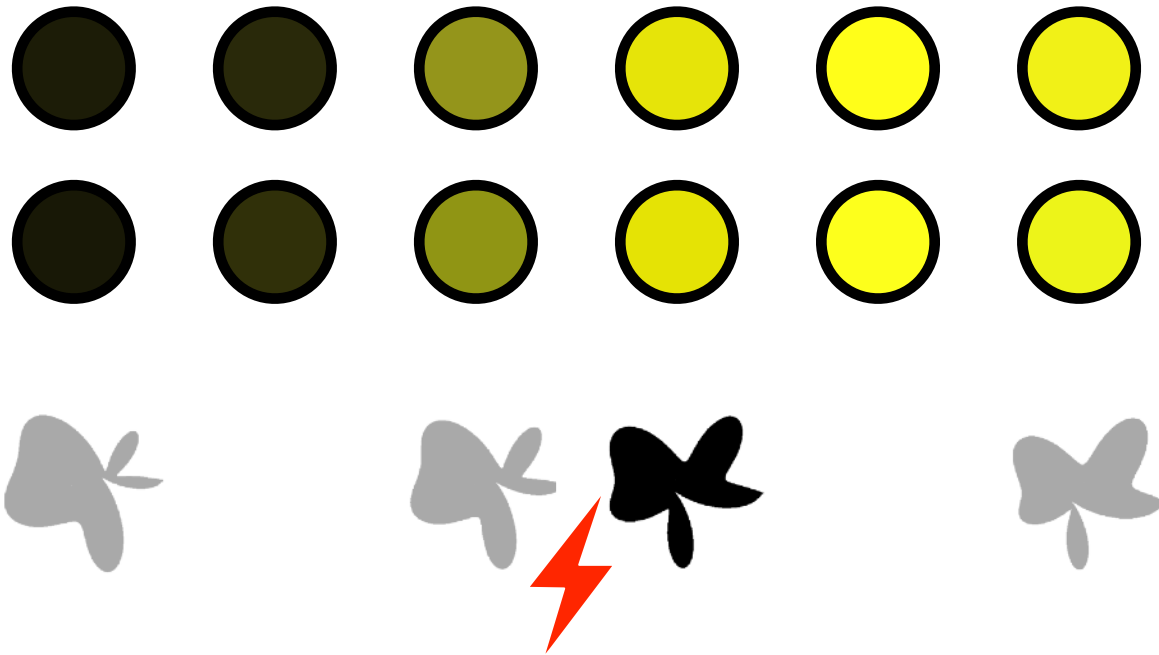
Manipulating stimulus representations with the NMPH



Manipulating stimulus representations with the NMPH

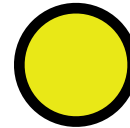
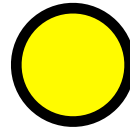
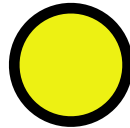
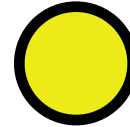
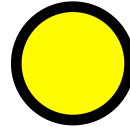
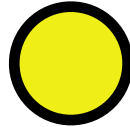
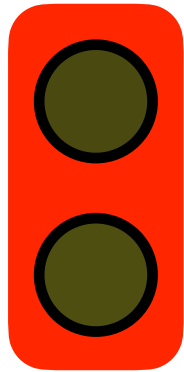


Manipulating stimulus representations with the NMPH

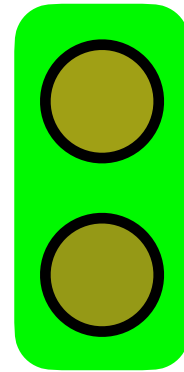


Manipulating stimulus representations with the NMPH

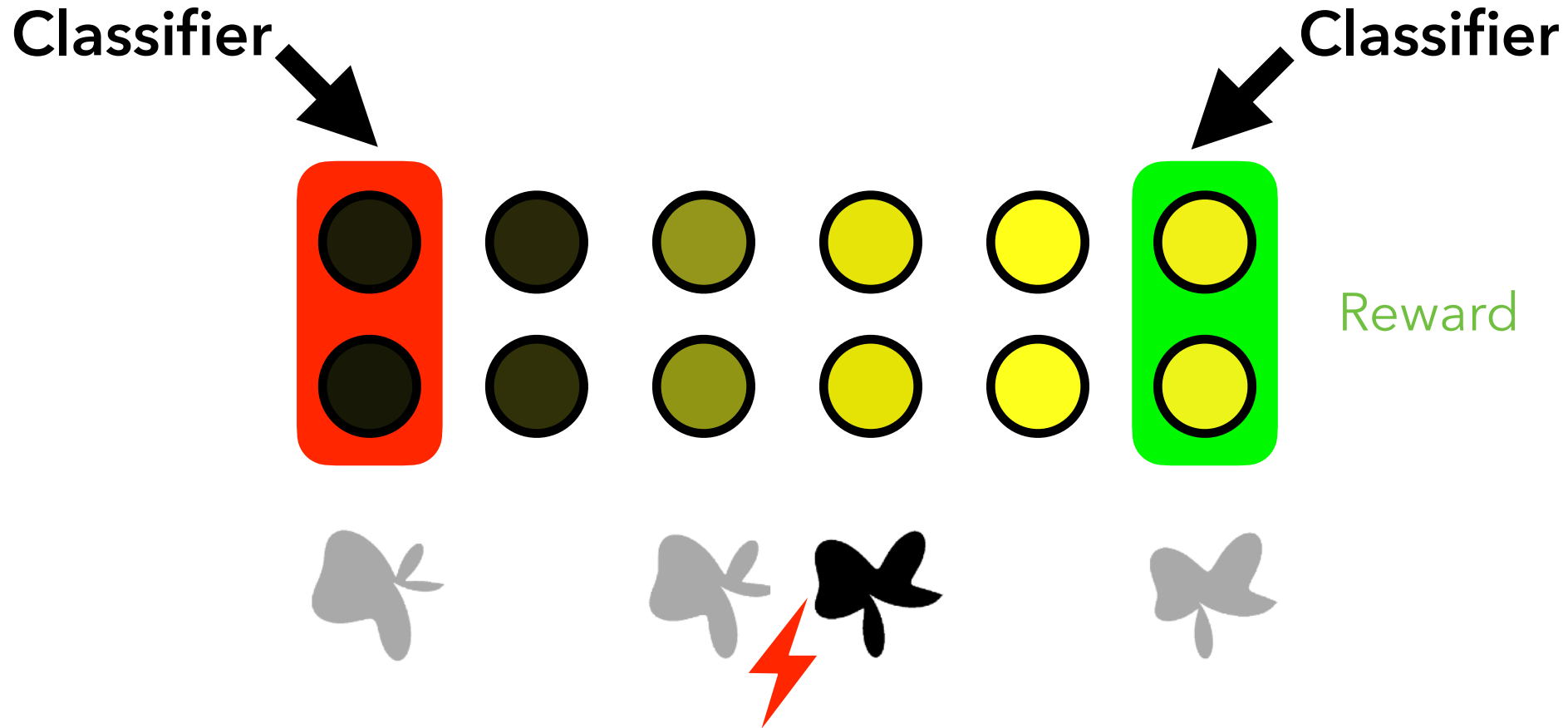
Classifier



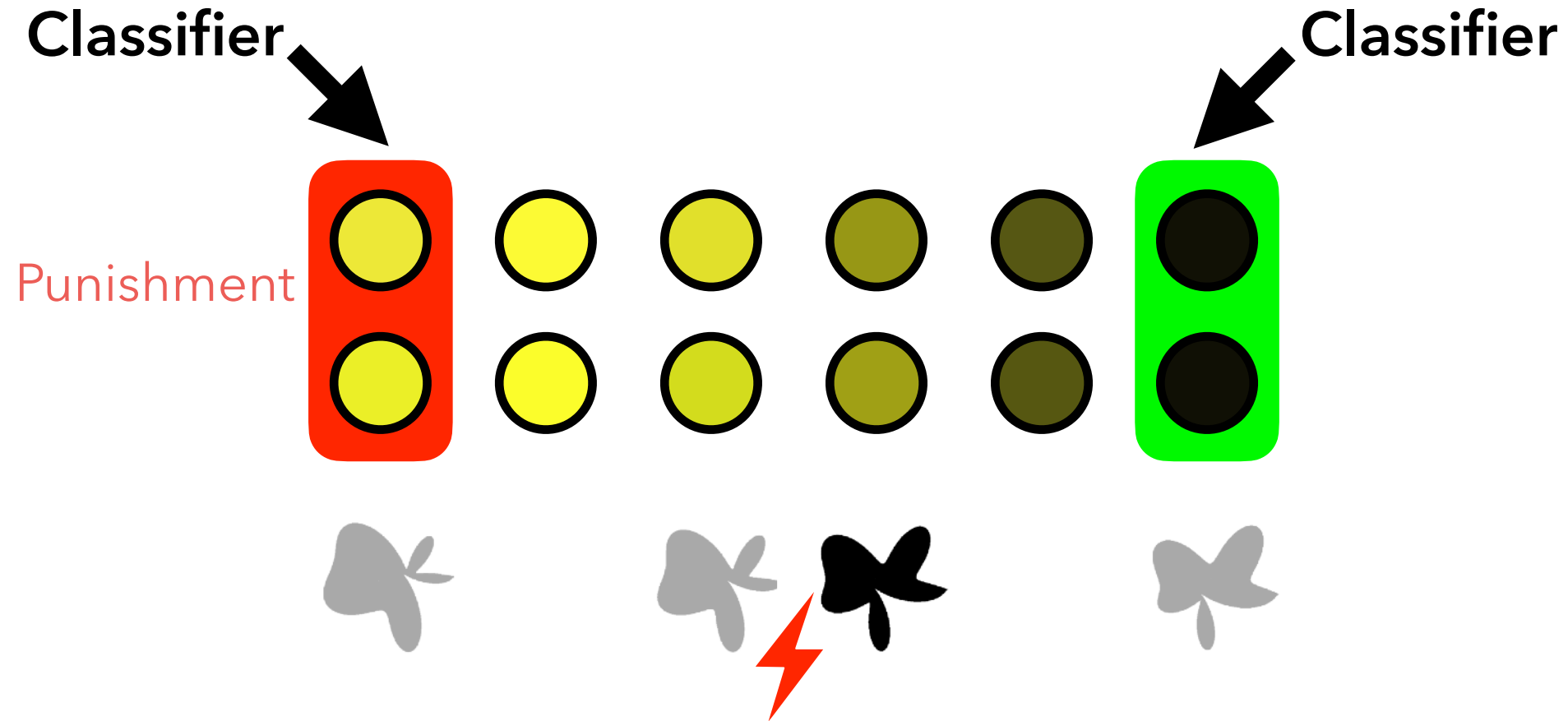
Classifier



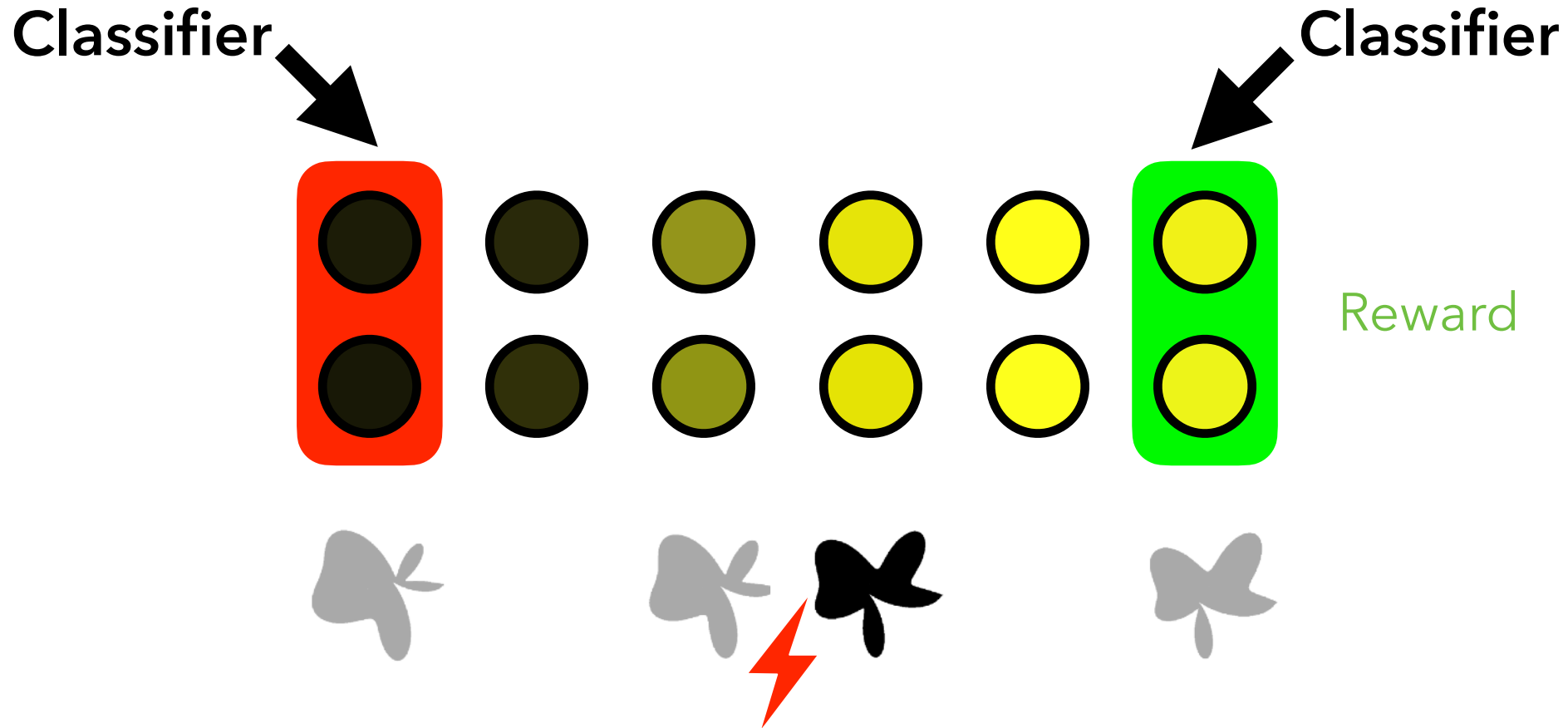
Manipulating stimulus representations with the NMPH



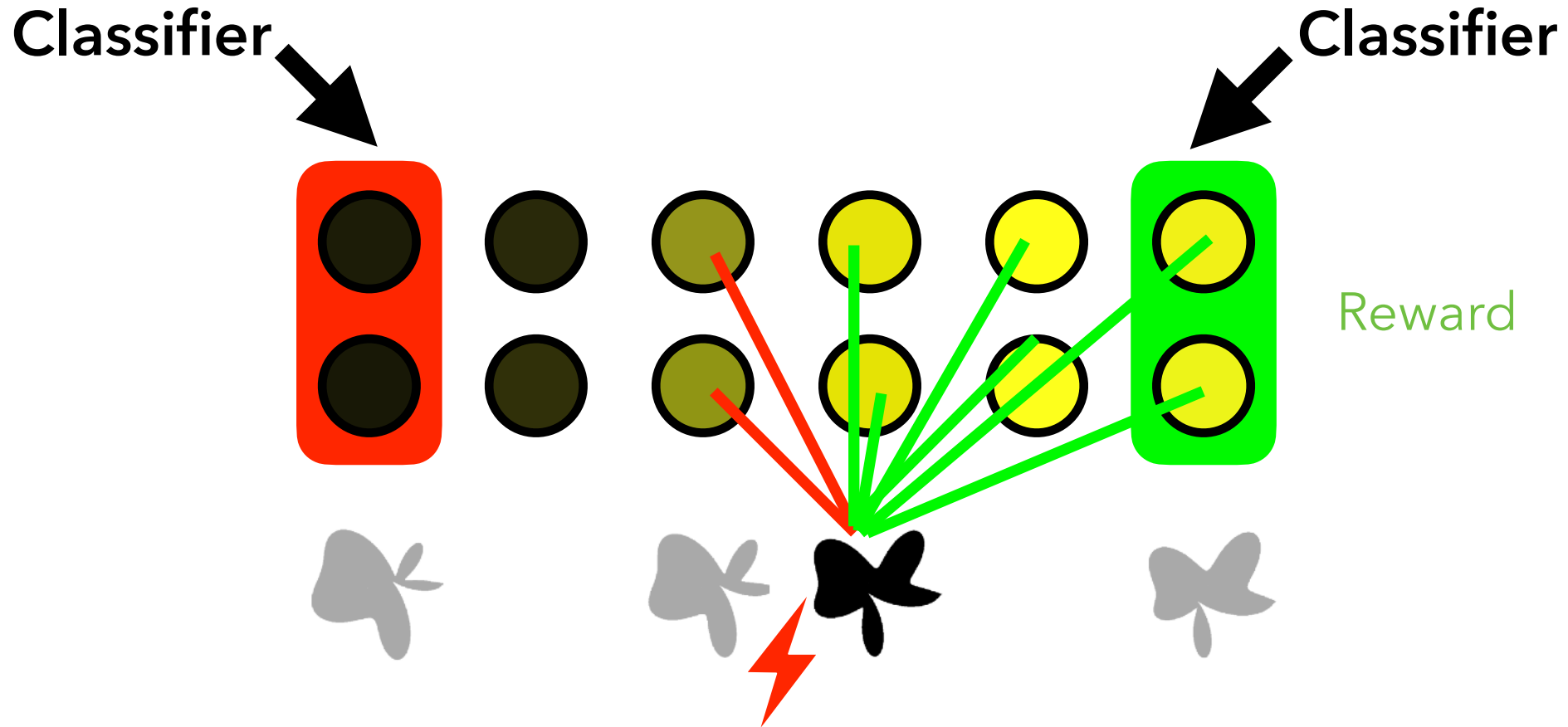
Manipulating stimulus representations with the NMPH



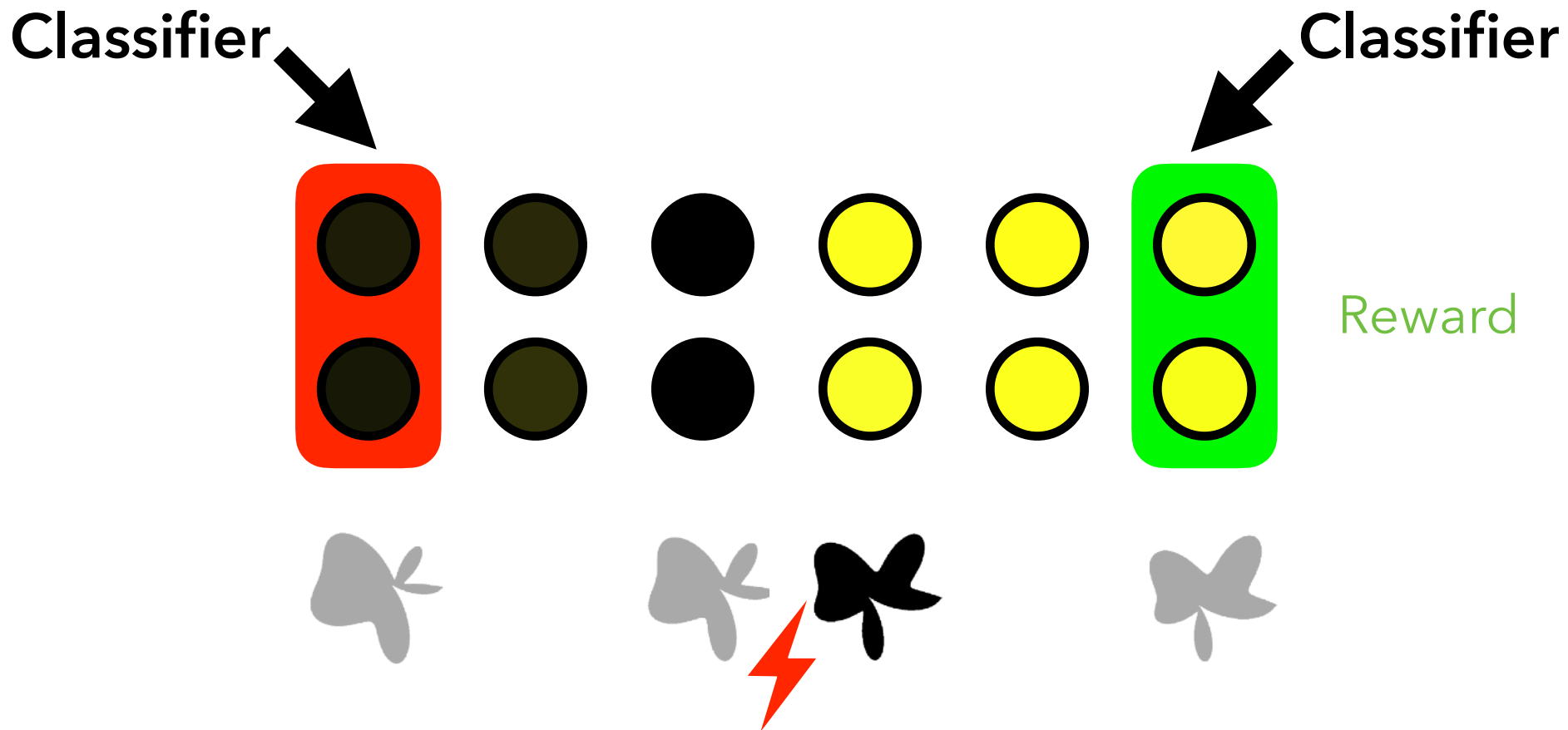
Manipulating stimulus representations with the NMPH



Manipulating stimulus representations with the NMPH

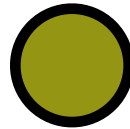
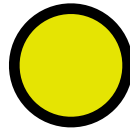
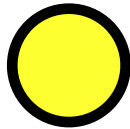
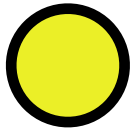
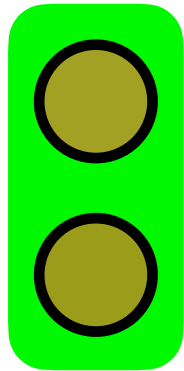


Manipulating stimulus representations with the NMPH

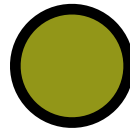
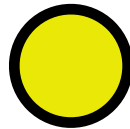
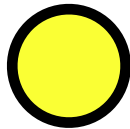
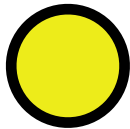
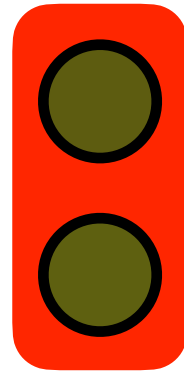


Manipulating stimulus representations with the NMPH

Classifier

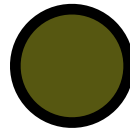
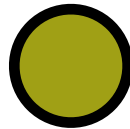
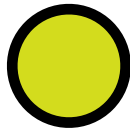
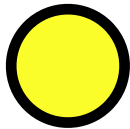
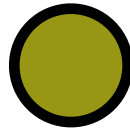
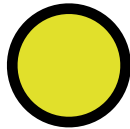
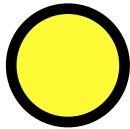
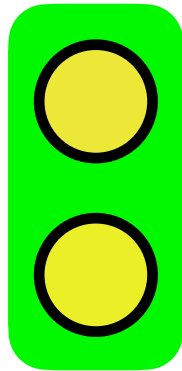


Classifier

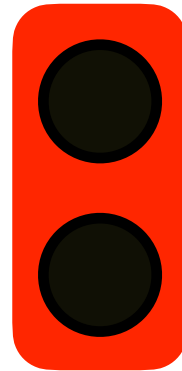


Manipulating stimulus representations with the NMPH

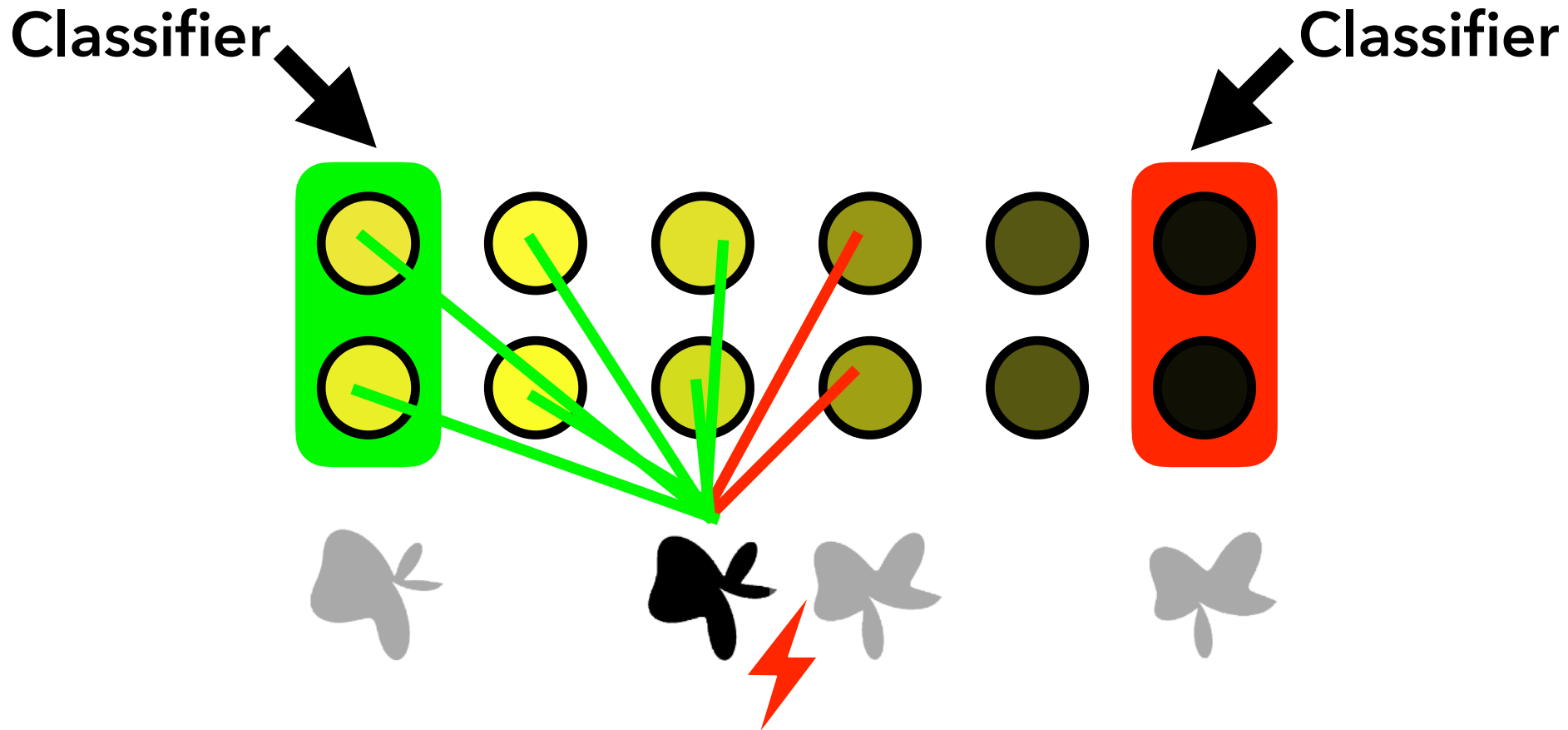
Classifier



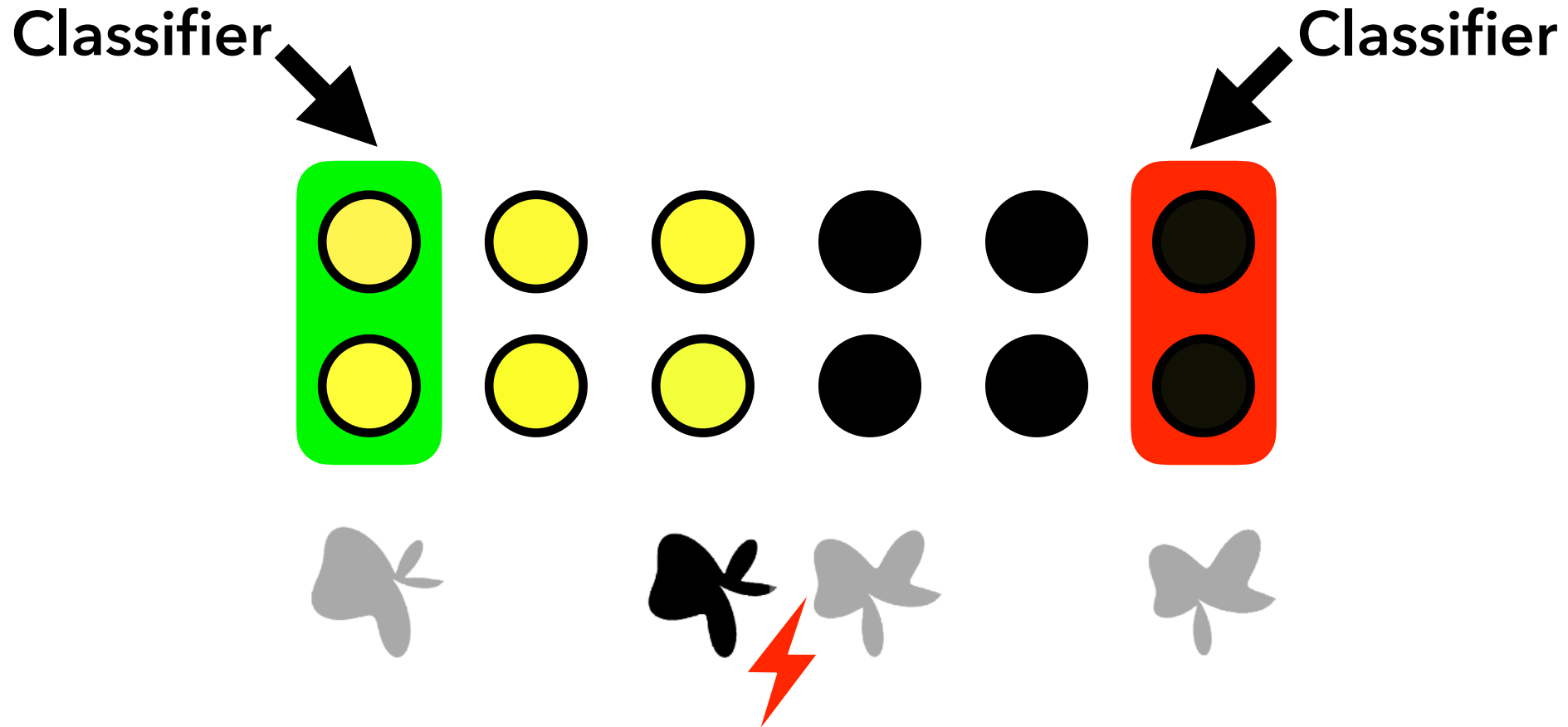
Classifier



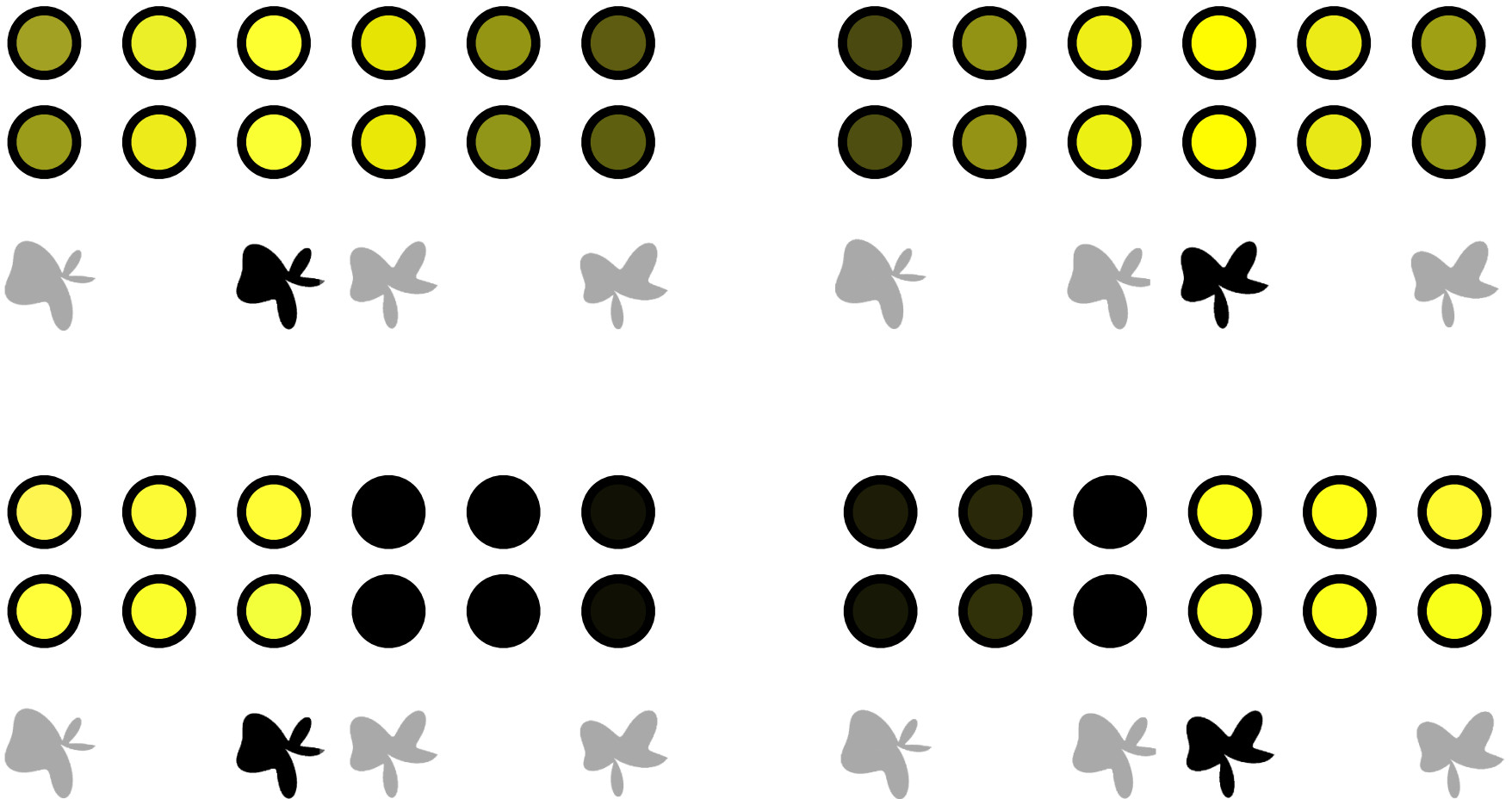
Manipulating stimulus representations with the NMPH



Manipulating stimulus representations with the NMPH



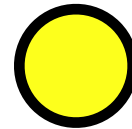
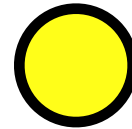
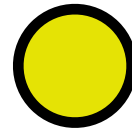
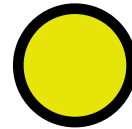
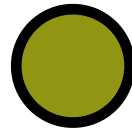
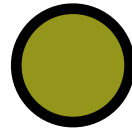
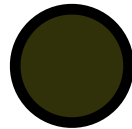
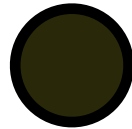
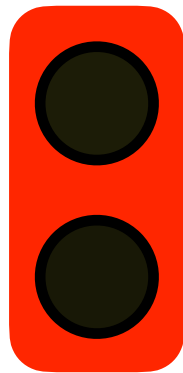
Manipulating stimulus representations with the NMPH



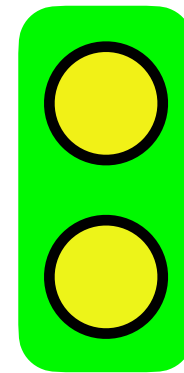
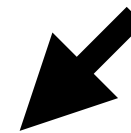
Manipulating stimulus representations with the NMPH

with the NMPH

Wiggle
more



Wiggle
less



Reward

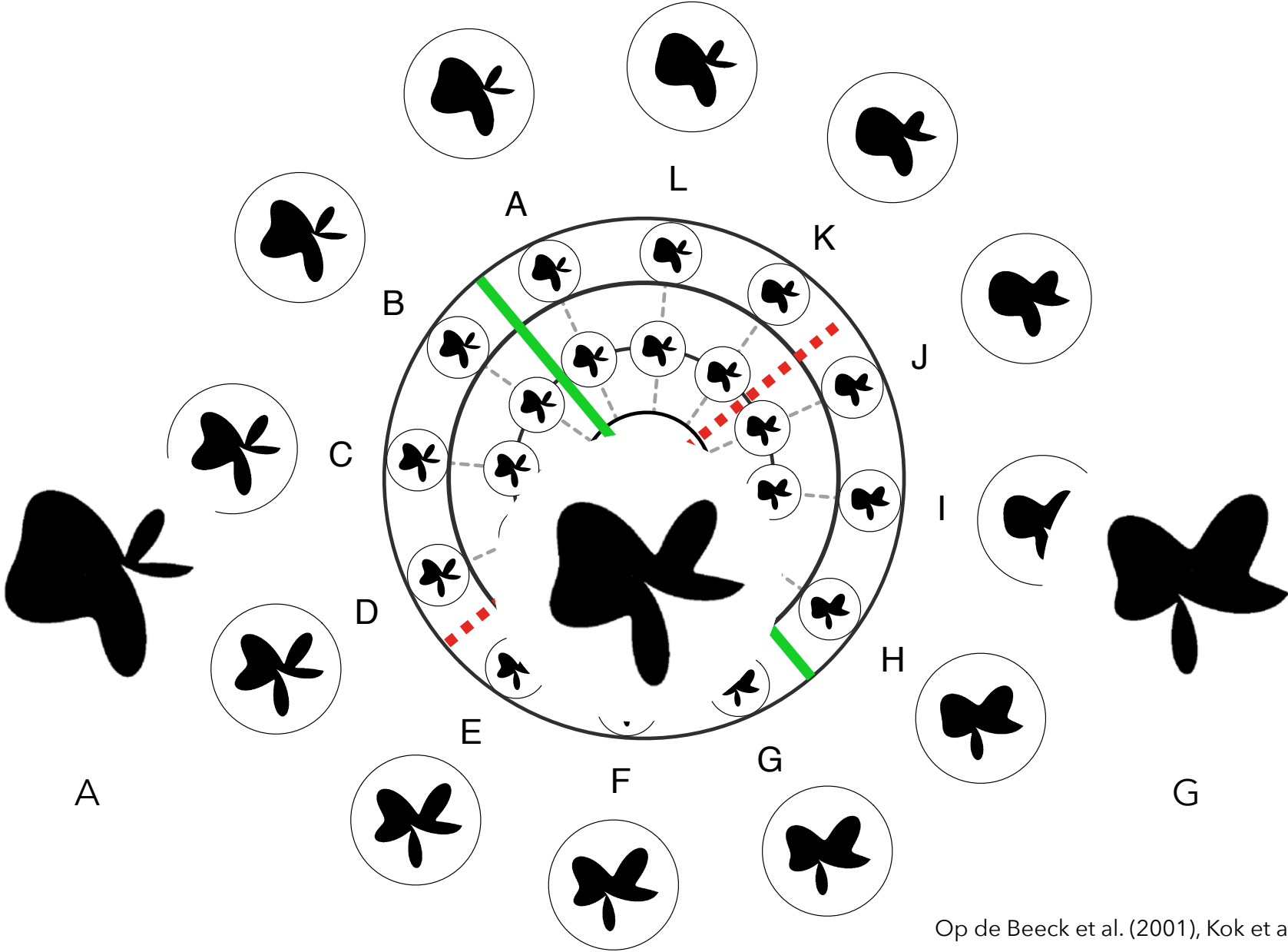
Punishment



<Wiggles>

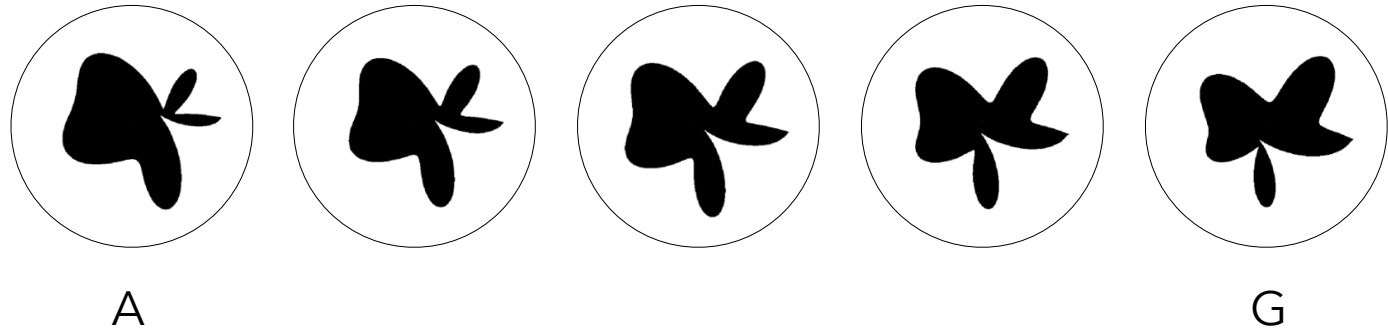
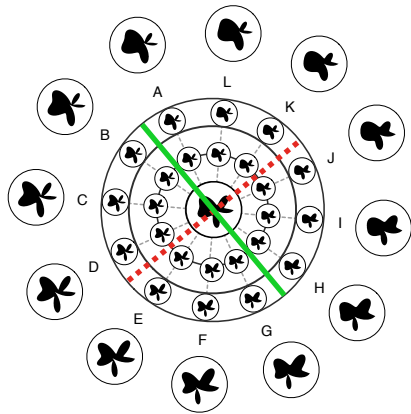


Behavioral Norming: Match-to-Category 2AFC

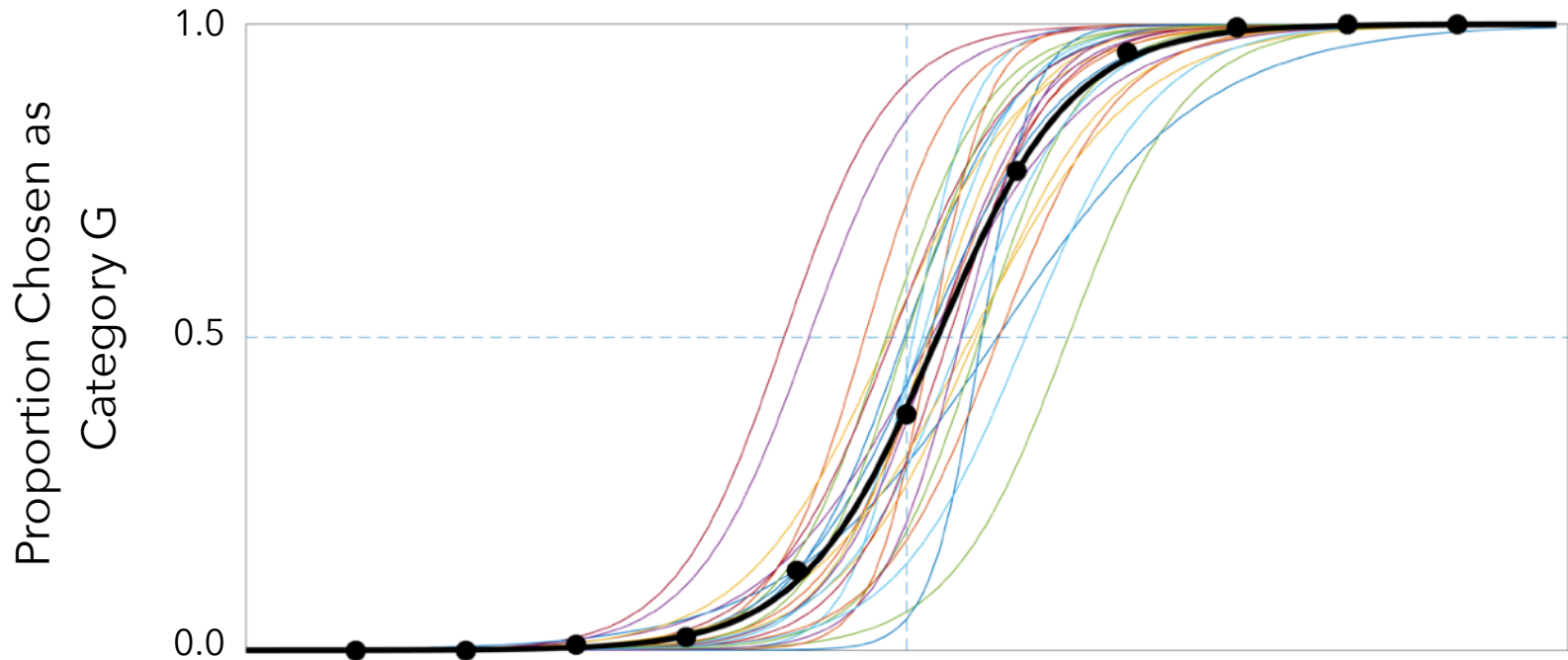


Op de Beeck et al. (2001), Kok et al. (2018)

Behavioral Norming: Match-to-Category 2AFC

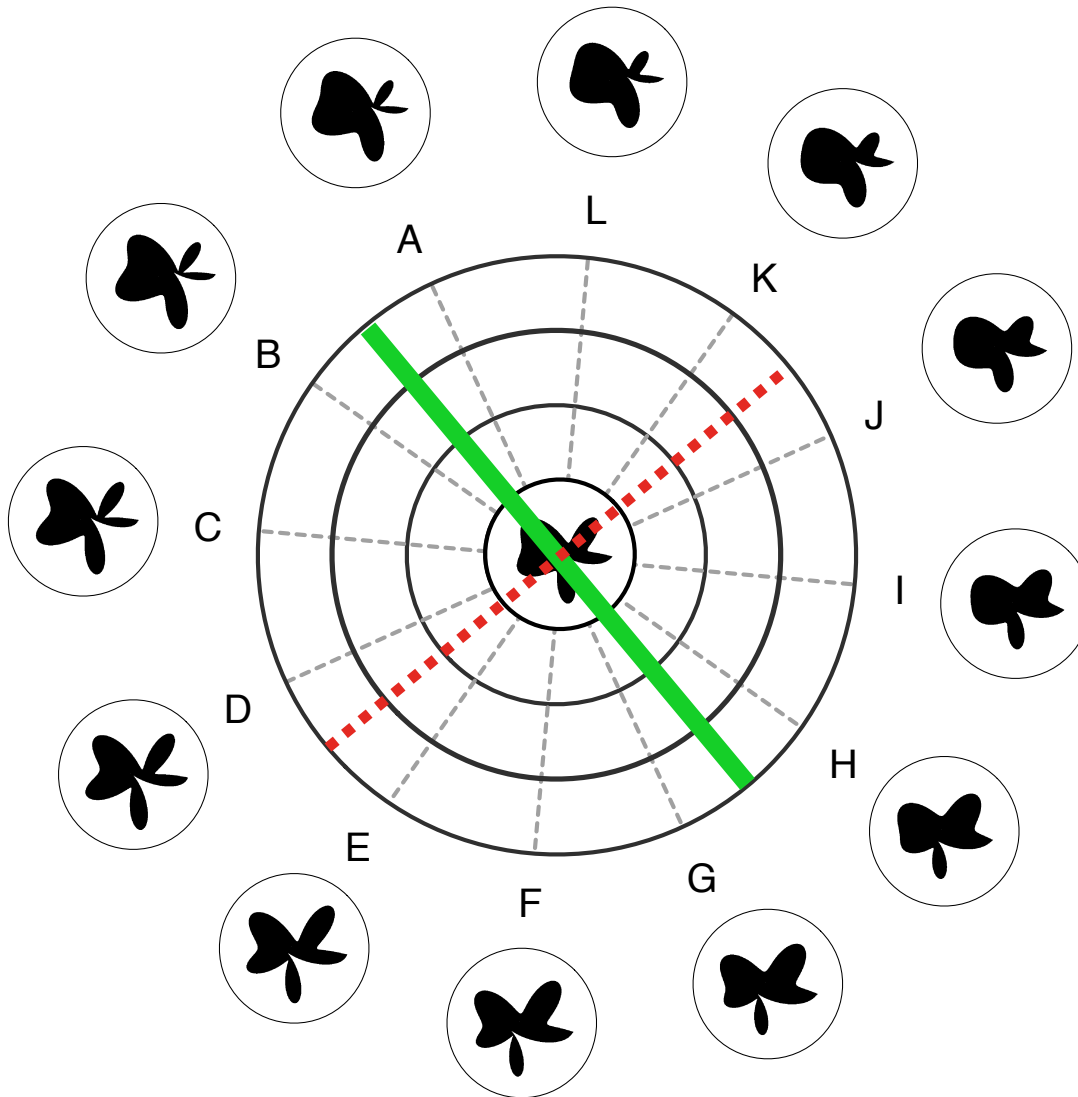


Amazon Mechanical Turk, n=28, 11 shapes x 20 repetitions per shape



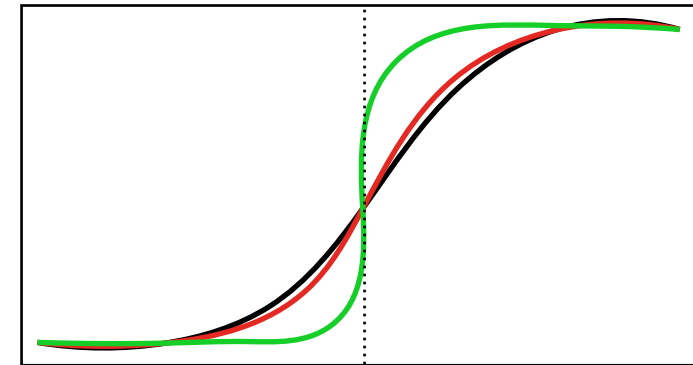
Perceptual Distinctions Enhanced Across Category Boundary

Sharper slope for **trained dimension** compared to **untrained dimensions**

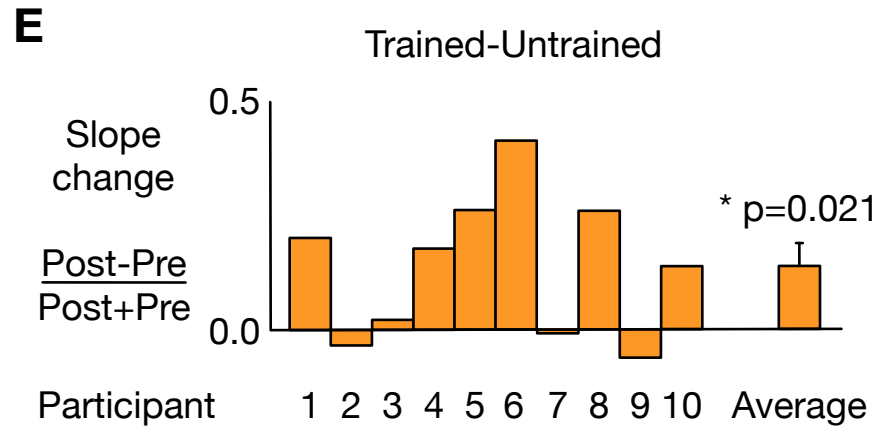
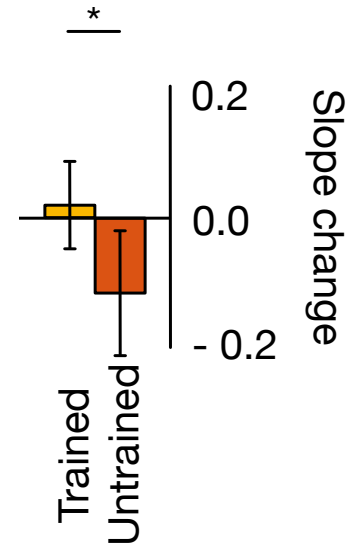
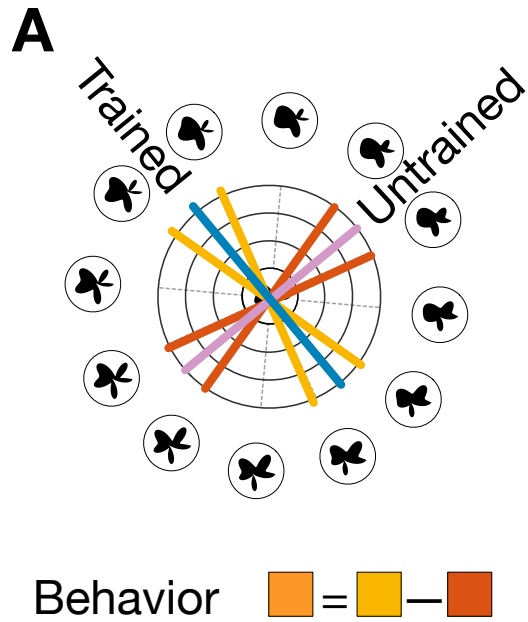


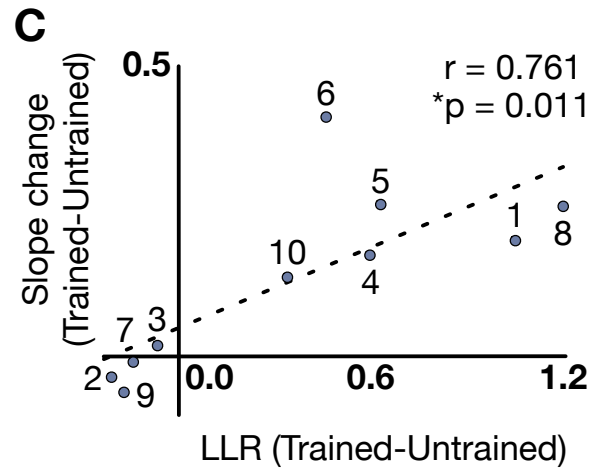
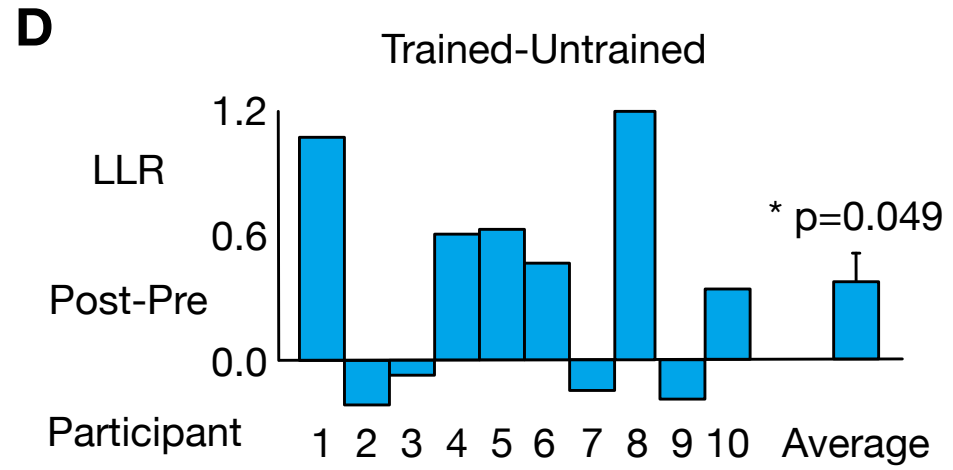
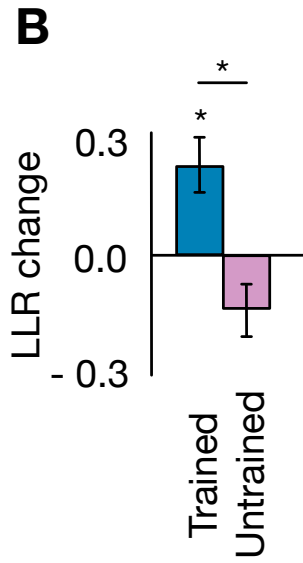
Behavioral Prediction

2AFC task

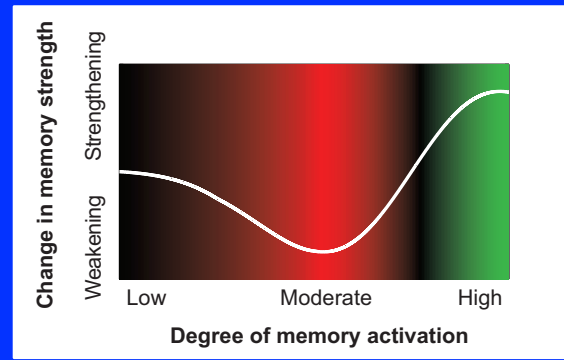


pre-training



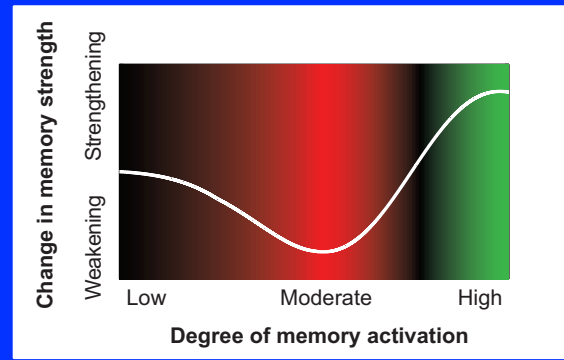


Summary



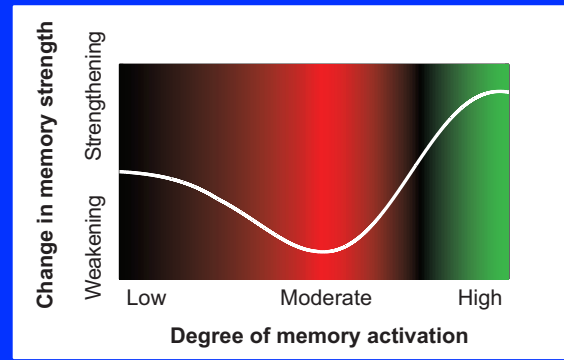
- The flow of activation through neural networks can powerfully sculpt memories that are activated
 - strengthening them or weakening them
 - integrating or differentiating them
- .. in ways that can't always be explained by supervised learning
- The NMPH provides a promising framework for understanding these unsupervised learning effects

Summary



- We are particularly excited to find ways to leverage the NMPH to drive new learning in difficult situations...

Summary



- There is much more work to be done to test & refine these ideas
- Good news: we've gotten much better at measuring **memory activation** and **representational change** with fMRI
- We are making progress in building **computational models** of these NMPH effects



Nick Turk-Browne
Yale University



Jon Cohen
Princeton University



Jarrod Lewis-Peacock
Faculty, UT-Austin



Anna Schapiro
Faculty, UPenn



Sam Gershman
Faculty, Harvard



Justin Hulbert
Faculty, Bates College



Lizzie McDevitt
Postdoc, Princeton



Ghootae Kim
Korea Brain Res. Inst.



Victoria Ritvo
Flatiron Health



Alex Nguyen
Grad, Princeton



Coraline Jordan
Faculty, URochester



Greg Detre
AI Consultant



Malai Natarajan
Philips Inc.

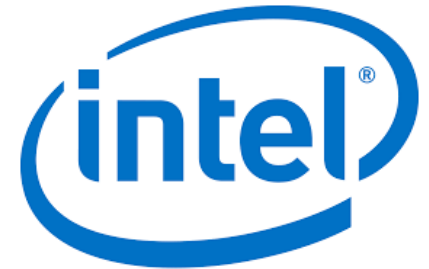


Jeff Wammes
Faculty, Queen's U.



Ehren Newman
Faculty, Indiana

More thanks



25 JOHN TEMPLETON FOUNDATION
YEARS SUPPORTING SCIENCE - INVESTING IN THE BIG QUESTIONS



Questions?

