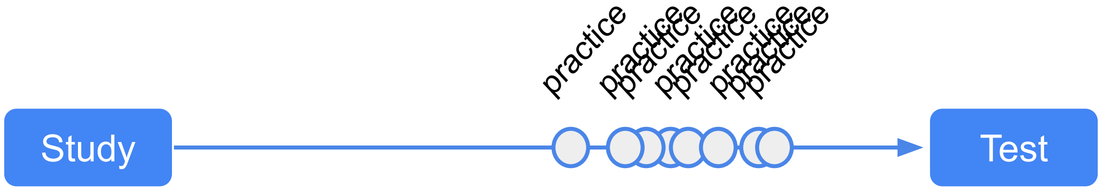
A Free Energy Interpretation of the Spacing Effect

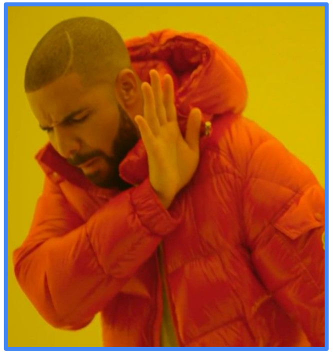
Andrea Stocco & Christian Lebiere

$$A_t(m) = \log \sum_i (t - t_i)^{-d(i)}$$

$$d(i) = e^{A(m)} + \varphi$$

The equation's origin story: The Spacing Effect





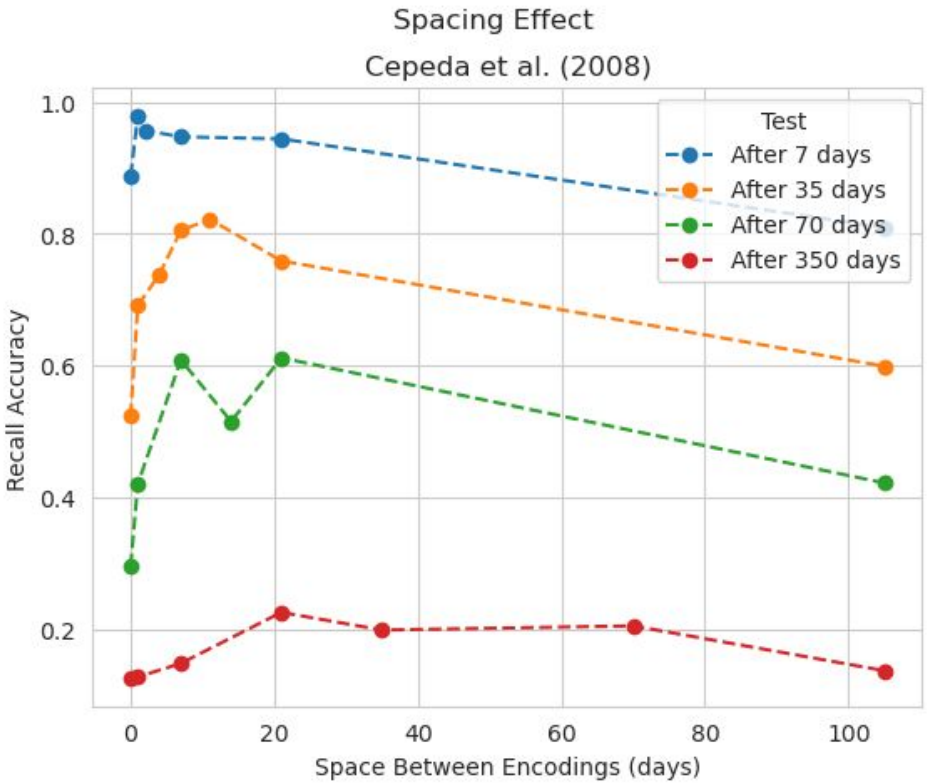
The spacing effect /2

- AKA the **spaced practice** benefit, **spaced repetition** effect, etc.
- Was first observed by Ebbinghaus himself in 1884
  - First modern study of memory ever
- Mostly studied for declarative memories, but…
- … Has been shown for **procedural** memories as well
  - Real-life skills, like CPR
  - Complex skills, like surgical practice
  - Motor skills in athletes
- We have good descriptions, no good explanations

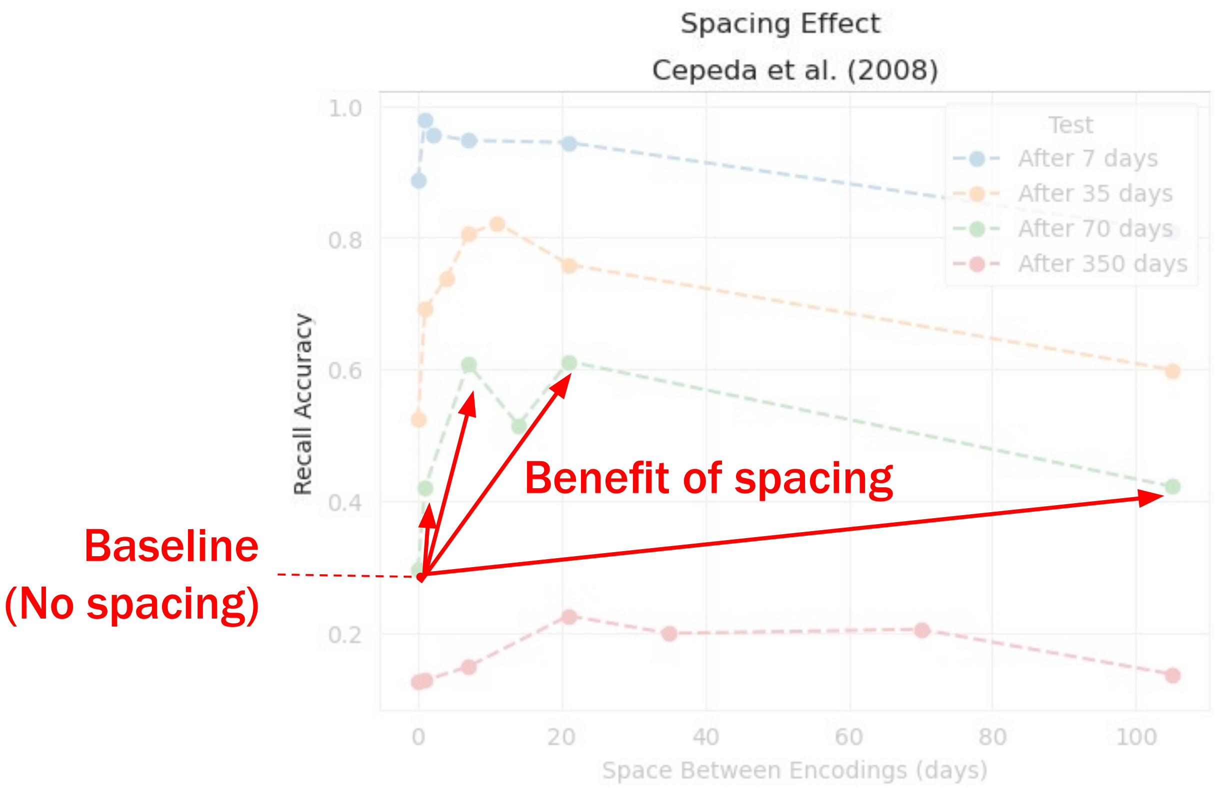
Example: Nick Cepeda's experiment (2008)

- Participants studied trivia
  - E.g., "which country consumes the most hot sauce per capita?"/"Norway"
- Each trivia was studied **twice**
- Systematically varied the **spacing (7 levels)** between the two study sessions and the **retention interval (4 levels)** before test
- 7 x 4 Between-group design = 1,350 participants

Cepeda et al., 2008 Results

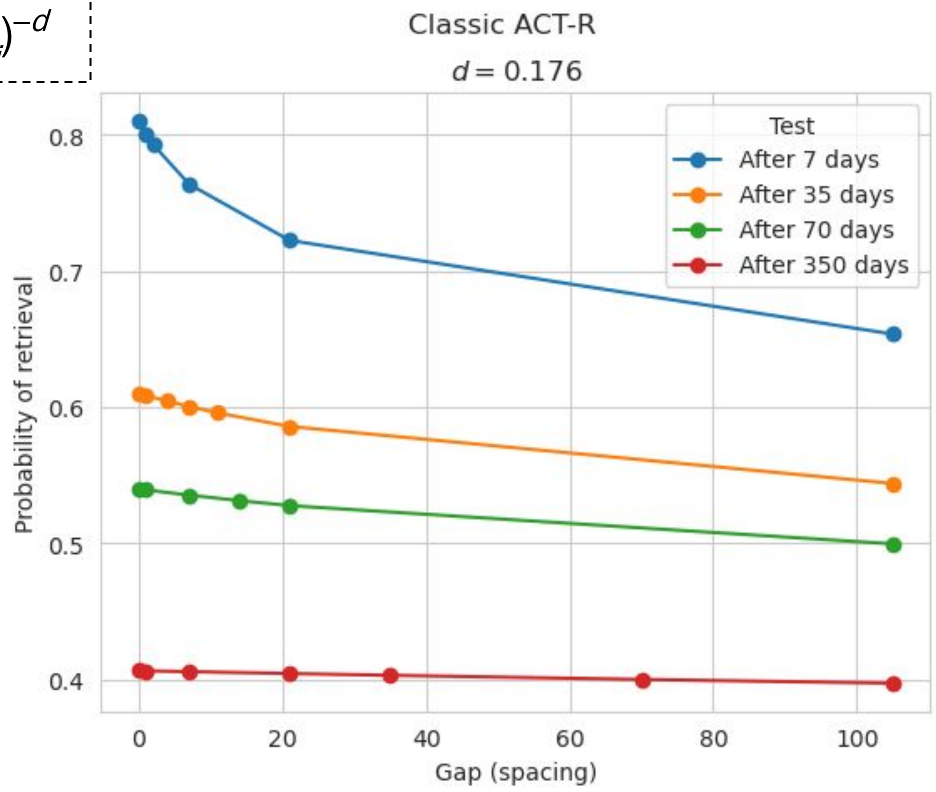


Cepeda et al., 2008 Results



Before Pavlik, ACT-R could not produce a spacing effect

$$A_t(m) = \log \sum_i (t - t_i)^{-d}$$



Classic ACT-R (1991) → **Pavlik & Anderson (2005)**

Classic ACT-R (1991)

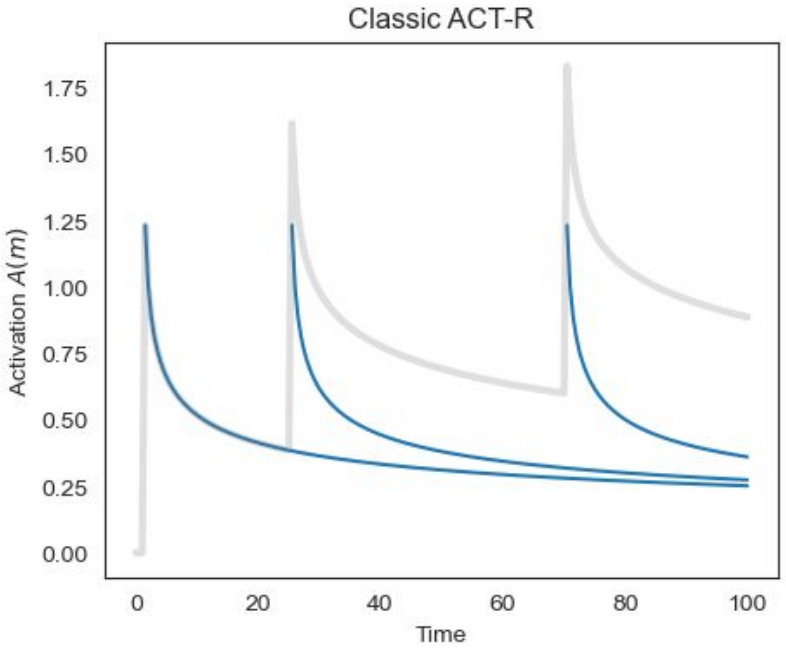
$$A(m) = \log(\, t_1^{-d} + t_2^{-d} + \ldots + t_N^{-d} \,)$$

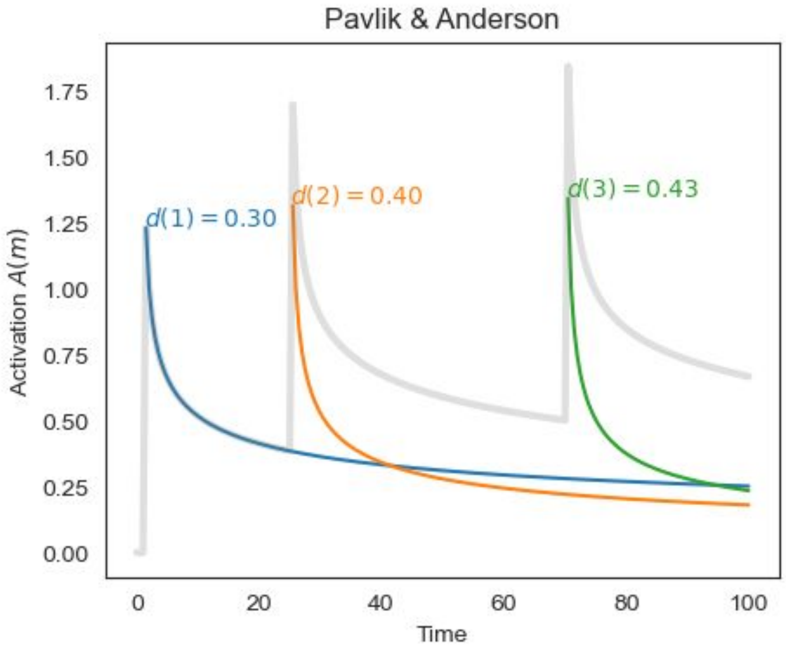
$$d = \text{constant}$$

Pavlik & Anderson (2005)

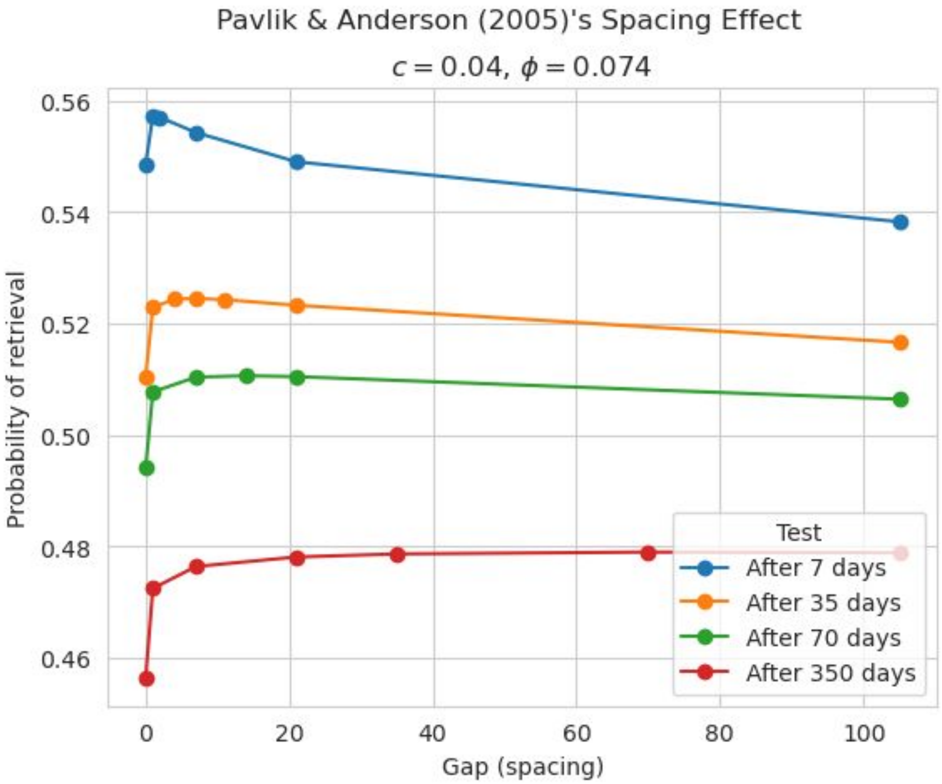
$$A(m) = \log(\, t_1^{-d(1)} + t_2^{-d(2)} + \ldots + t_N^{-d(N)} \,)$$

$$d(i\,) = c\, e^{A(m)} + \varphi$$





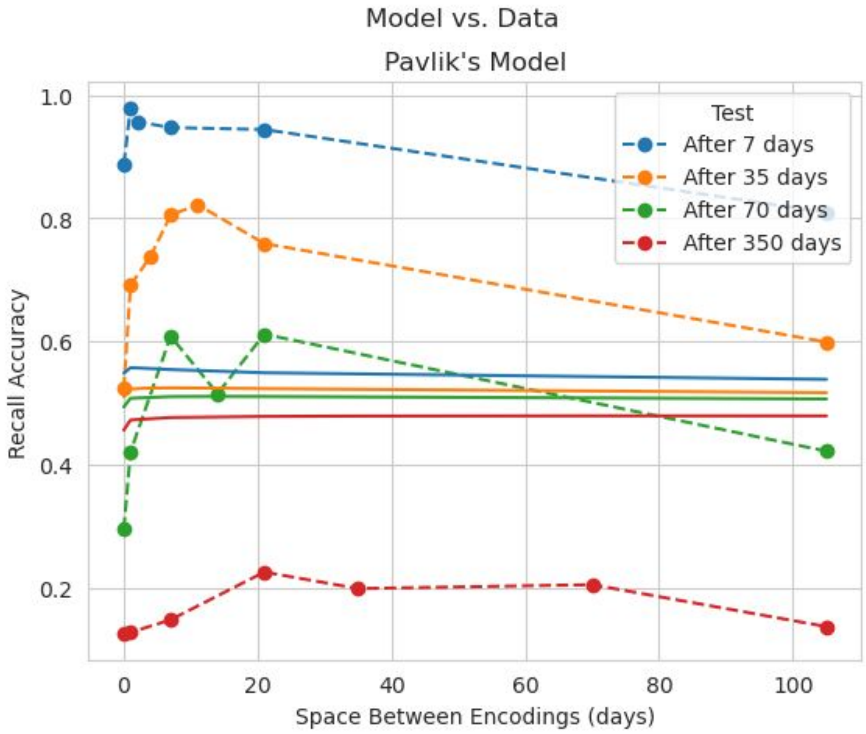
Pavlik & Anderson does produce a spacing effect




IF IT AIN'T BROKE,
DON'T FIX IT.

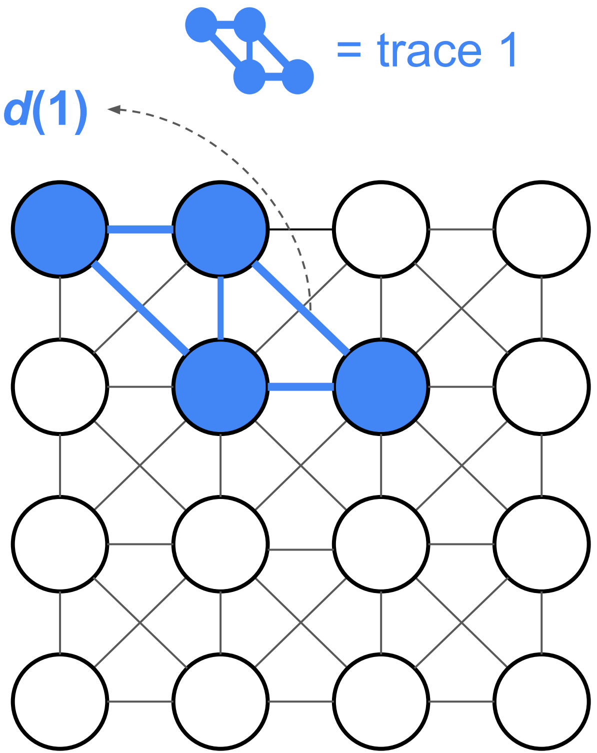

IF IT AIN'T BROKE
DON'T FIX IT

Problem 1: P&A does not work well for long intervals



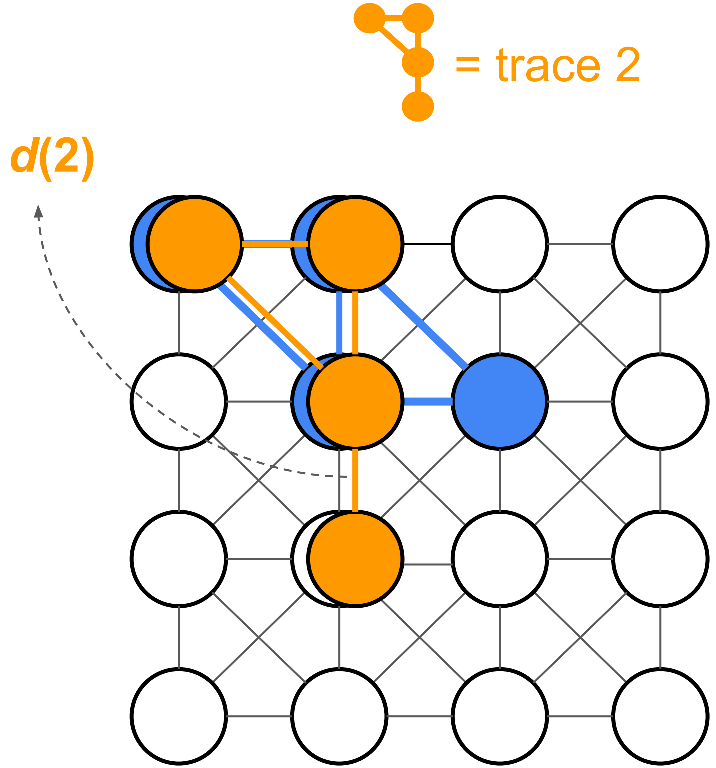
Problem 2: Interpretation

- **Memories** are accumulation of traces
- **Traces** are patterns of active neurons
- Memories are strengthened by increasing **synaptic weights**
- Decay *d* approximates the speed at which synapses are lost



Memory as a Hopfield network

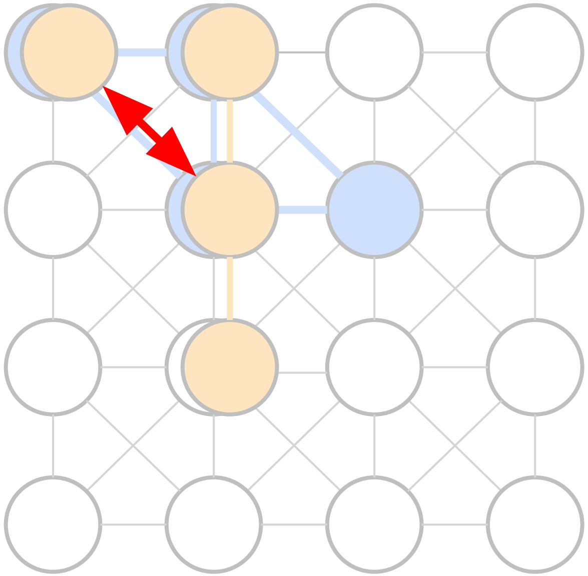
- **Memories** are acculation of traces
- **Traces** are patterns of active neurons
- Memories are strengthened by increasing **synaptic weights**
- Decay *d* approximates the speed at which synapses are lost
- … But this causes a problem



Memory as a Hopfield network

- **Memories** are acculation of traces
- **Traces** are patterns of active neurons
- Memories are strengthened by increasing **synaptic weights**
- Decay *d* approximates the speed at which synapses are lost
- … But this causes a problem

d(1) or *d*(2)?



A New Approach

Alternative model

Pavlik's idea: traces decay at different rates

$$A(m) = \log(\ t_1^{-d(1)} + t_2^{-d(2)} + \ldots + t_N^{-d(N)}\ )$$

Alternative: Different traces are **weighted different**

$$A(m) = \log(\ w_1 t_1^{-d} + w_2 t_2^{-d} + \ldots + w_N t_N^{-d}\ )$$

But how is *w* computed?

This is where Free Energy comes in place

- **Free Energy Principle (FEP: Friston, 2010):** The brain maintains homeostasis by minimizing the **surprisal (-log *P*) of new stimuli:**
  - Allocates neural resources efficiently
  - Can be shown
- FEP is related to **Predictive Coding (PC: Rao, 1999):** The brain is trying to maximize successful predictions of the next events
  - Encode neural information so that it predicts future states
- Both FEP solves the problem of **optimal encoding:** How many neural resources should you invest in processing a new new trace?
- You can think of it as a neural extension of **rational analysis**


ANDREA
& CHRISTIAN
PAVLIK
& ANDERSON
FREE ENERGY
imgflip.com

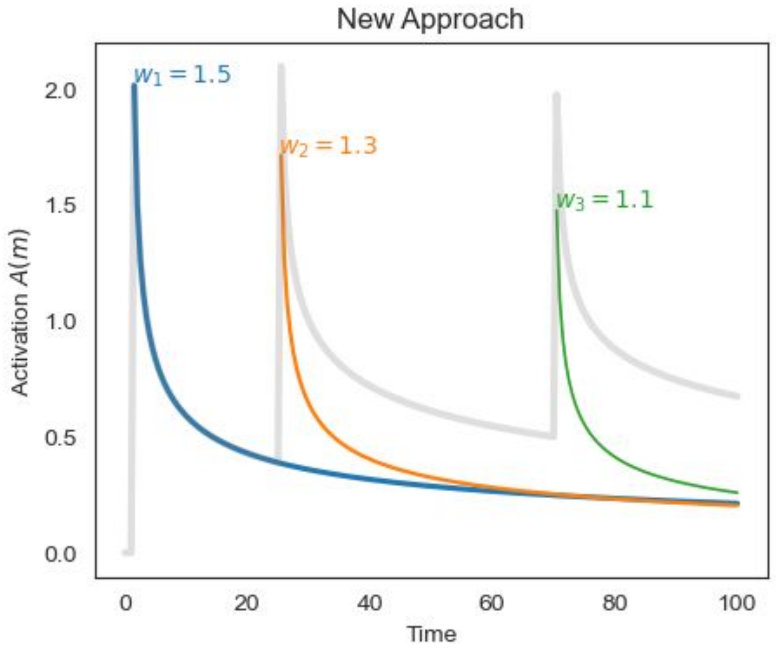
New Approach

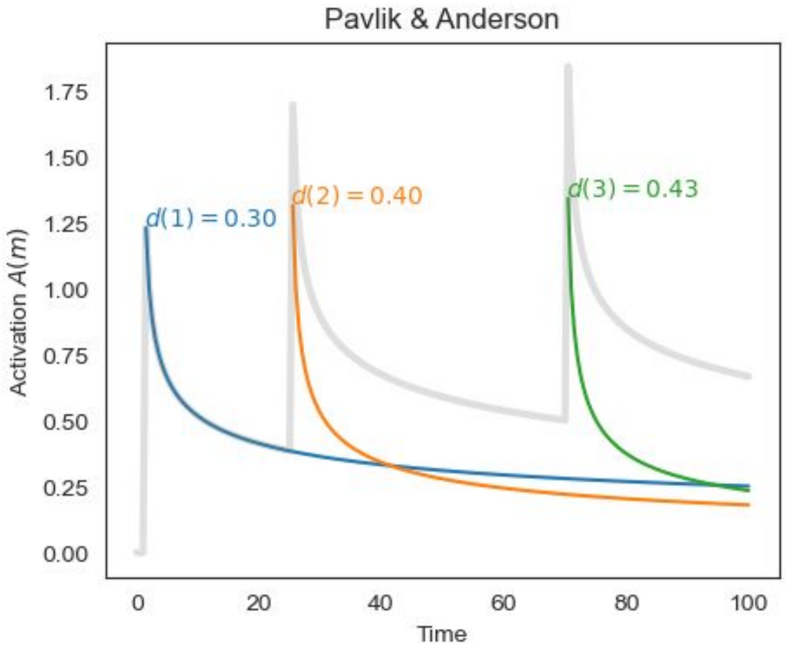
$$A(m) = \log(\, w_1 t_1^{-d} + w_2 t_2^{-d} + \ldots + w_3 t_N^{-d} \,)$$

Pavlik & Anderson (2005)

$$A(m) = \log(\, t_1^{-d(1)} + t_2^{-d(2)} + \ldots + t_N^{-d(N)} \,)$$

$$d(i) = c\, e^{A(m)} + \varphi$$





Christian's solution

Current activation at rehearsal time t:

$$A(m,t) = \log(\, w_1 t_1^{-d} + \ldots + w_{N\text{-}1} t_{N\text{-}1}^{-d}\,)$$

Compute w_N such that activation at time $t + \tau$ is constant k:

$$\log(\, w_1\, (t_1 + \tau\,)^{-d} + \ldots + w_N\, t^{-d}\,) = k$$

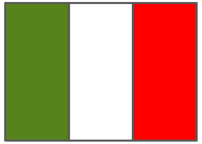
When $\tau = 1$, w_1 independent of d

When $k = 0$, $w_1 = 1$: full rehearsal

Prevents **out-of-control activation** buildup and limit **winner-take-all dynamics**

Like Belgian drinking, you try to keep the buzz constant as more beers arrive

Andrea's solution



Weights are proportional to **surprisal**

w = surprisal of $m = -\log P(m)$

$= -\log\,[e^{A(m)} / (1 + e^{A(m)})]$

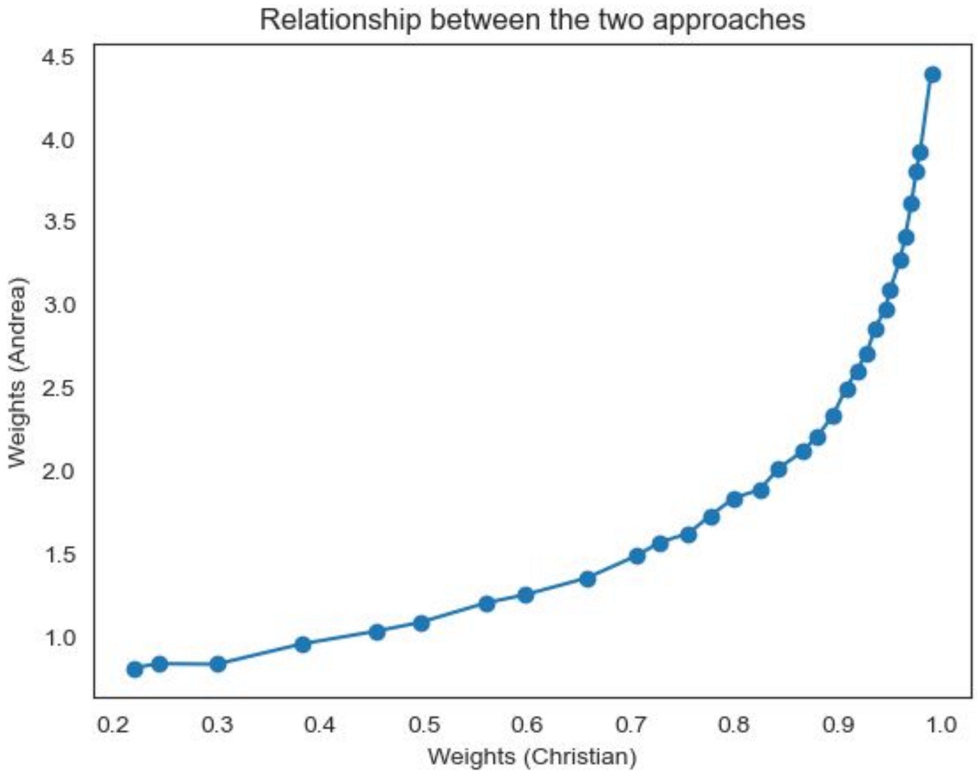
$= -\log\,[1 / (1 + e^{-A(m)})]$

$= \log(1 + e^{-A(m)})$

You can think of it as **minimizing free energy**

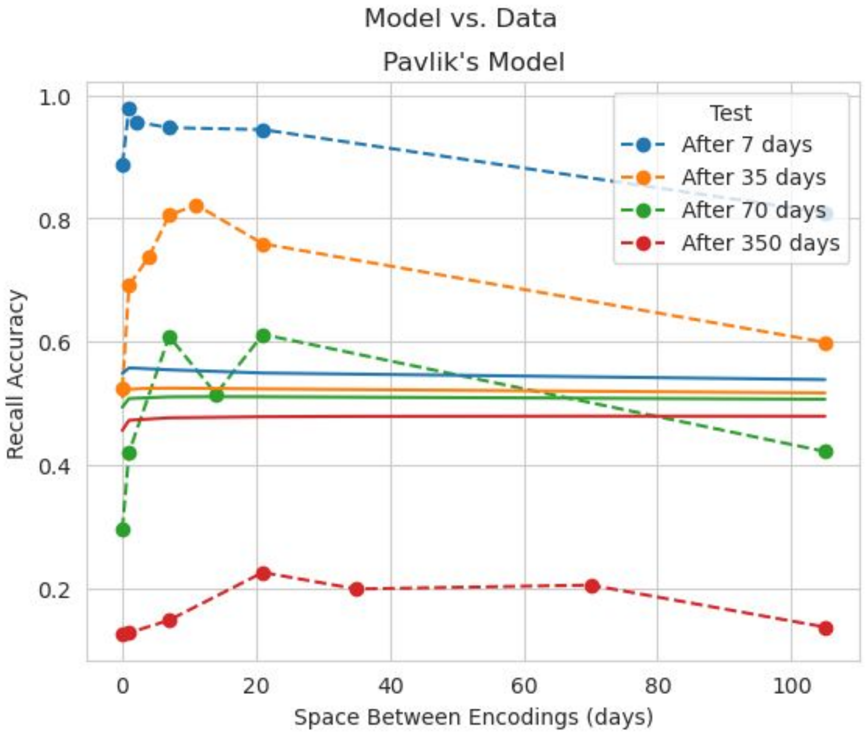
Like Italian wine, you try to match it the food that it expected in a full meal

The two approaches are, in fact, related

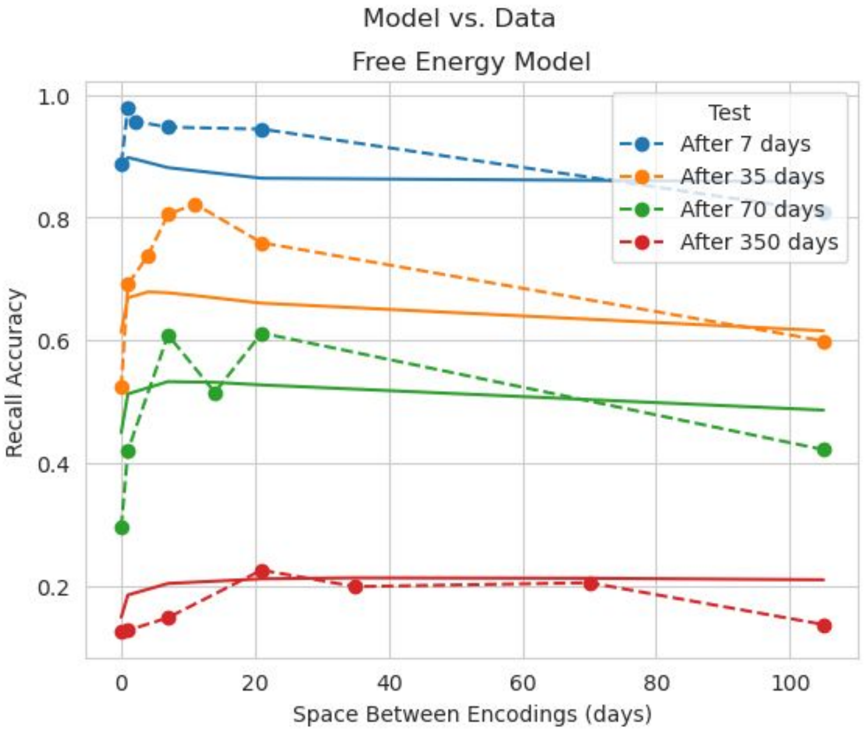


Is the new approach better?

Data fit

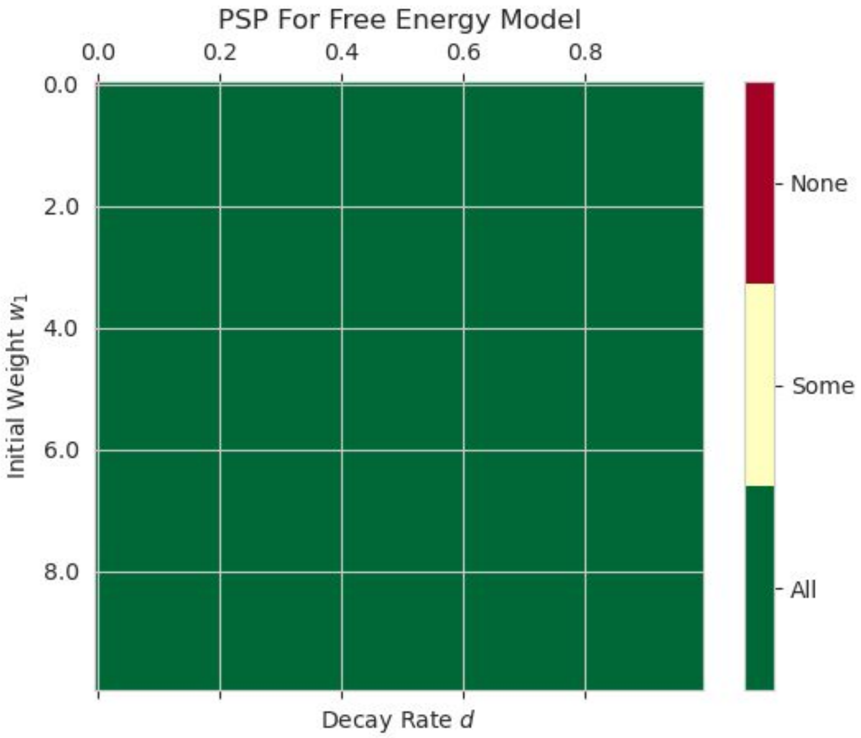


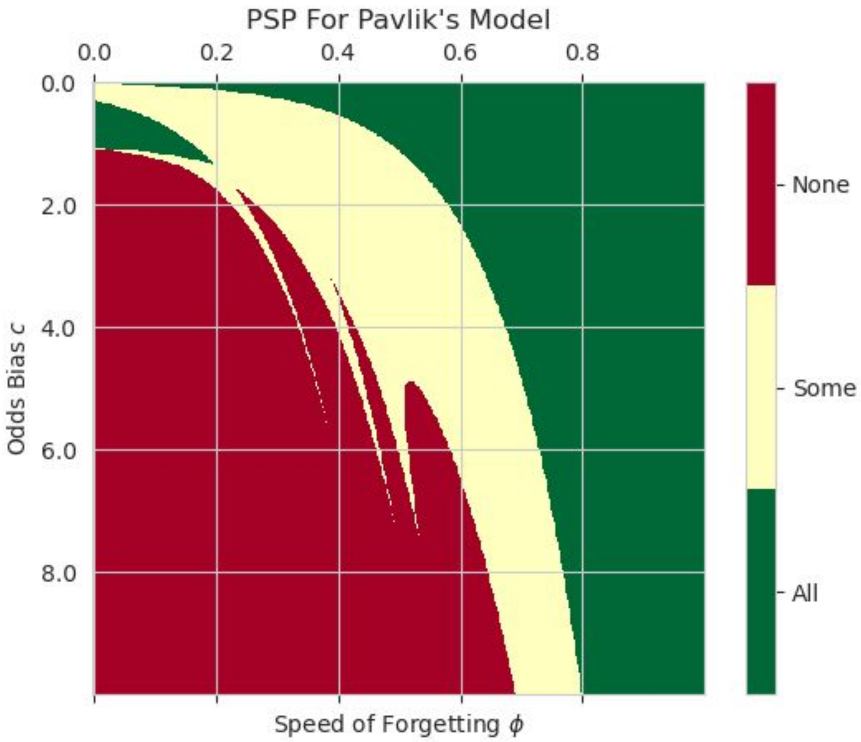
Model vs. Data
Pavlik's Model
Test
After 7 days
After 35 days
After 70 days
After 350 days
Recall Accuracy
Space Between Encodings (days)



Model vs. Data
Free Energy Model
Test
After 7 days
After 35 days
After 70 days
After 350 days
Recall Accuracy
Space Between Encodings (days)

Parameter Space Partitioning

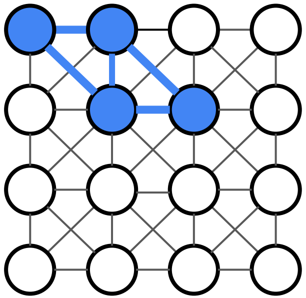


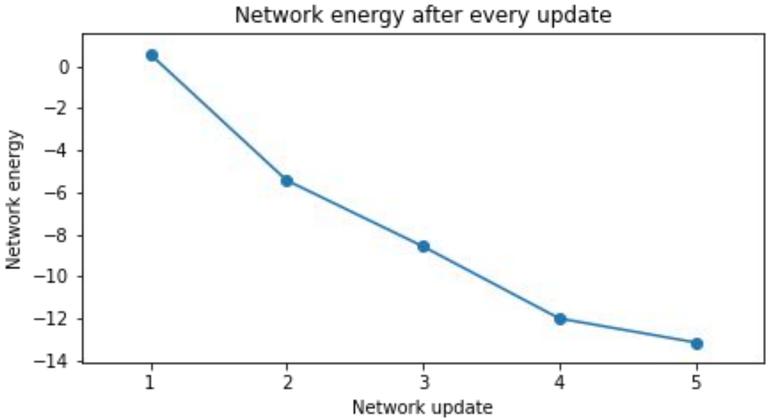


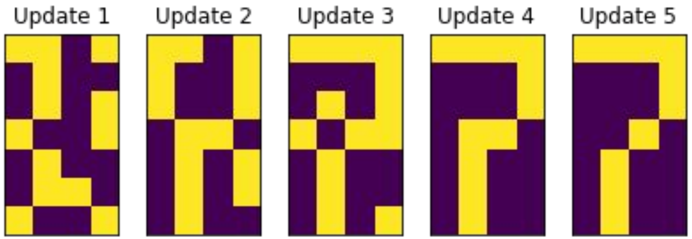
How about the brain?

Hopfield networks: How do they remember?

- Common model for hippocampus
- Memories are states with associated **energy *H***:
    - $H(m) = -\sum_i \sum_j w_{i,j}\, x_i\, x_j$
- During **retrieval**, networks move to the closest **minimum energy state.**







Similarities between Hopfield networks and ACT-R

Energy $H \approx$ Activation

- Memories have an intrinsic "energy" H

$$H(m) = -\sum_i \sum_j w_{i,j}\, x_i\, x_j$$

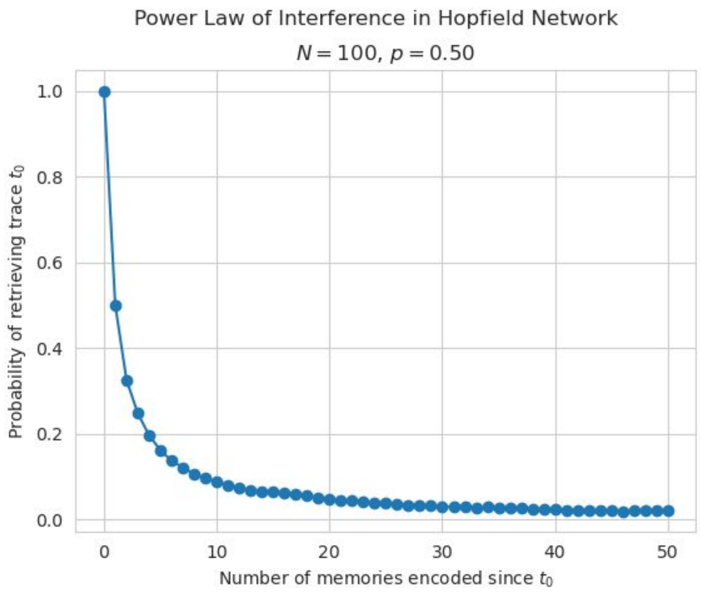
- Probability of retrieving m is inversely proportional to its **energy *H***

$$P_{Hopfield}(m) = 1 / (1 + e^{H(m)})$$

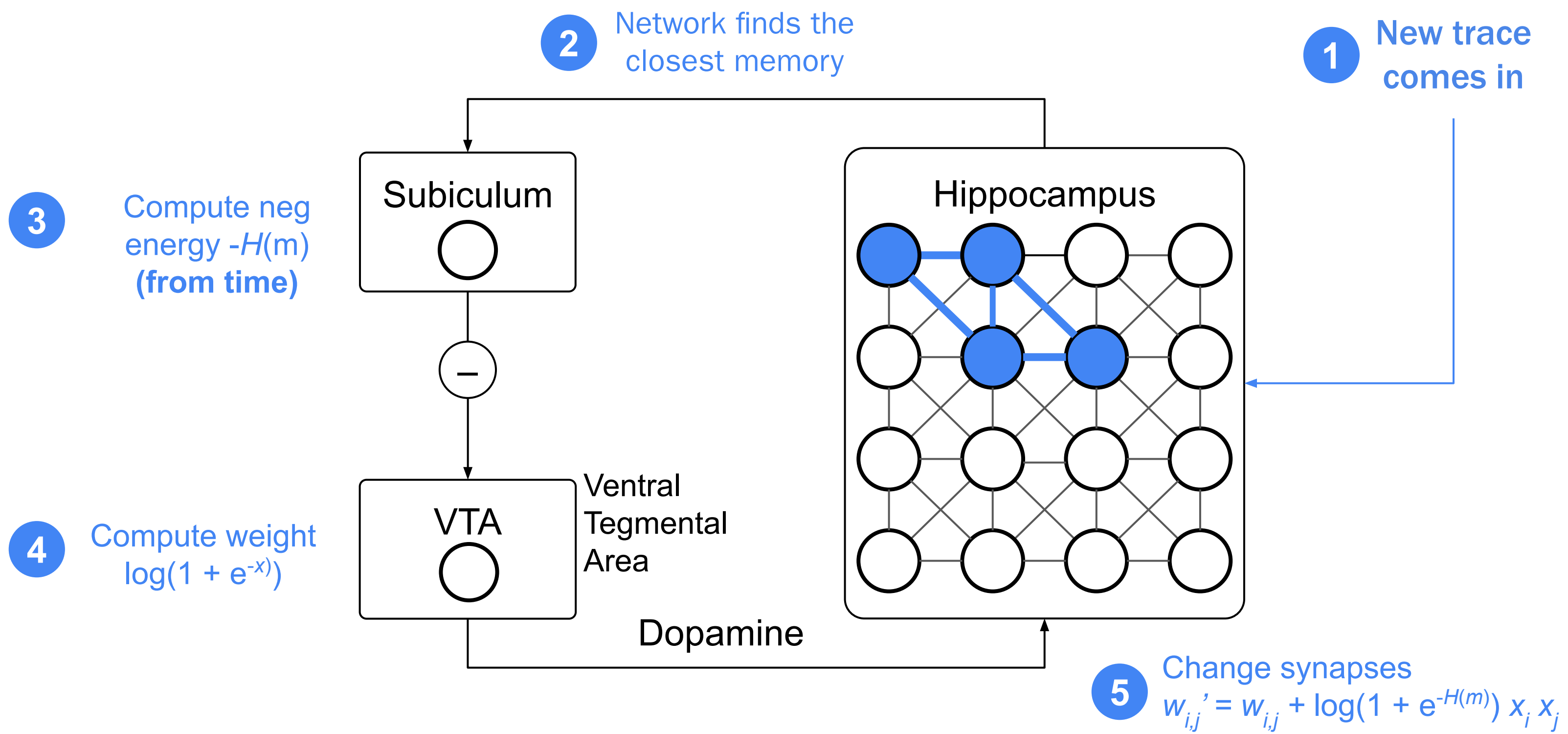
- Analogous to ACT-R, where

$$P_{ACT\text{-}R}(m) = 1 / (1 + e^{-A(m)})$$

Interference ≈ Power Decay

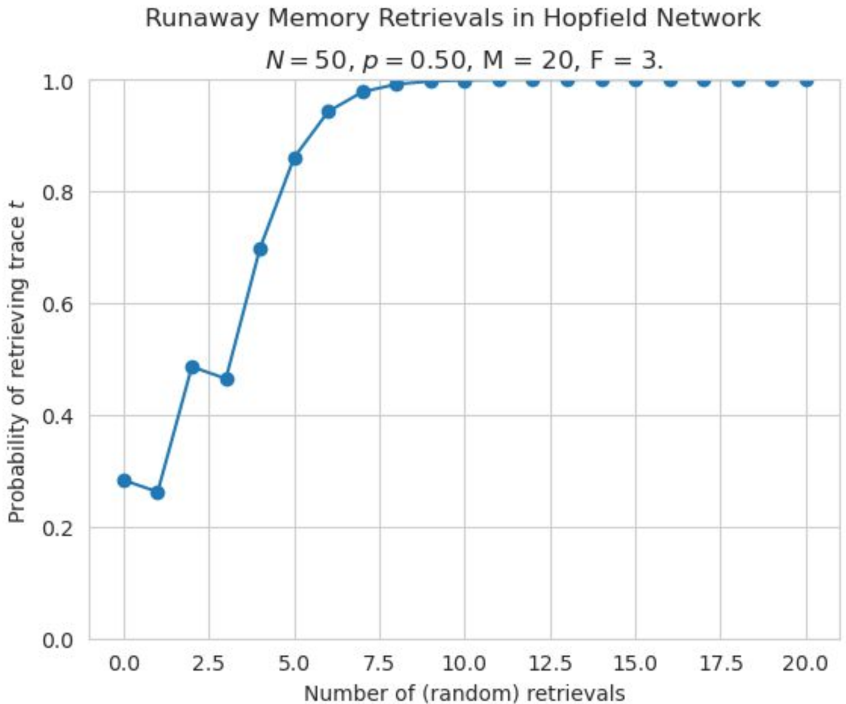


Neural implementation of Free Energy model

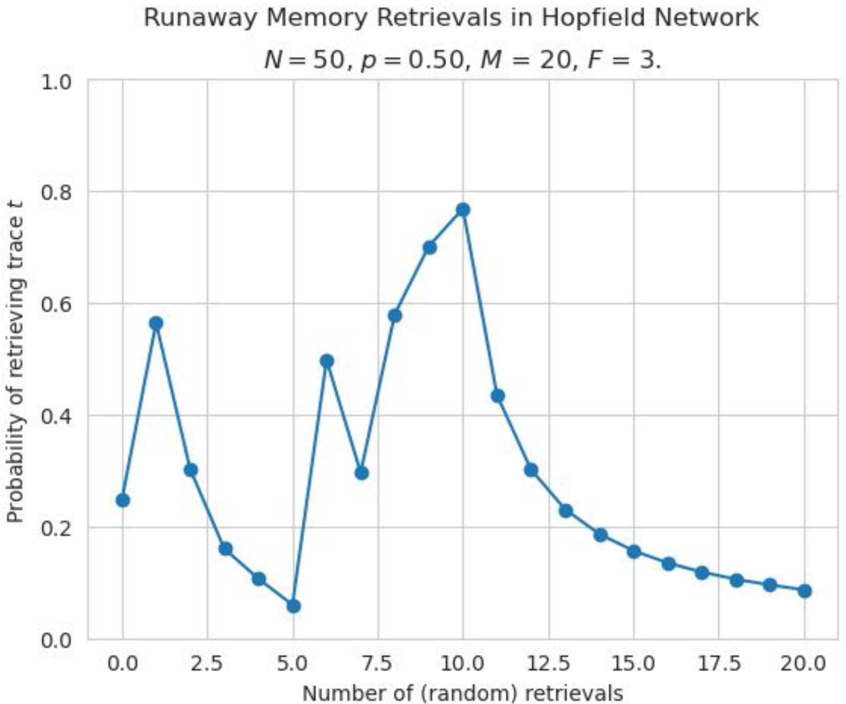


Free energy learning prevents runaway dynamics

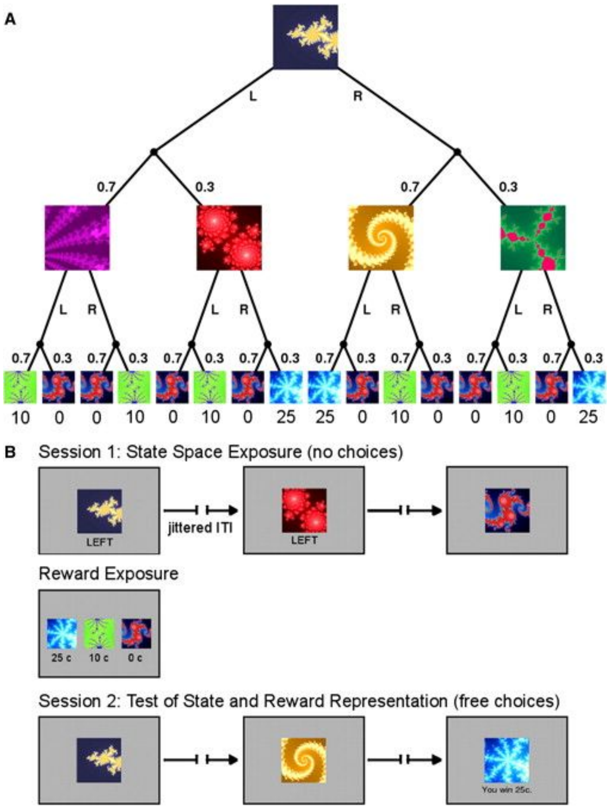
Standard Hopfield



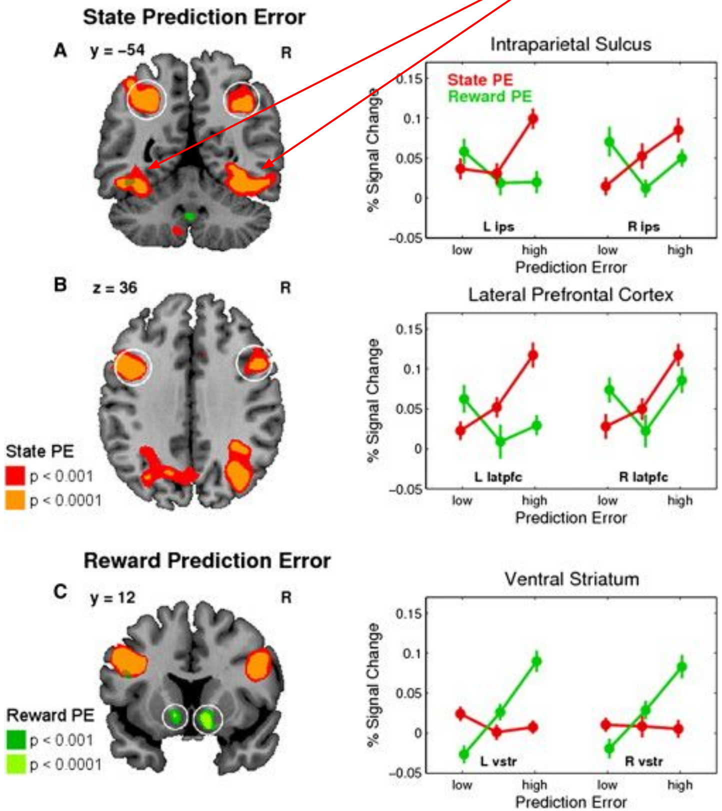
Free Energy Hopfield



Evidence from fMRI



Hippocampus



Next steps

- Find formal relationship between approaches
- Derive optimal training schedule
  - Compare to Pavlik & Anderson
  - Impact on practical applications? MemoryLab?
- Biological: apply to neural data
- Cognitive: model Pavlik and Anderson tasks
- Rational: relate to Anderson and Milson model
- Social: apply to Anderson et al. (2022) Twitter data analysis

Questions?


LEFT
EXIT 12
Pavlik
&
Anderson
Free
Energy
Andrea & Christian
imgflip.com