Mechanisms Of The Distributed Practice Effect

Michael Mozer

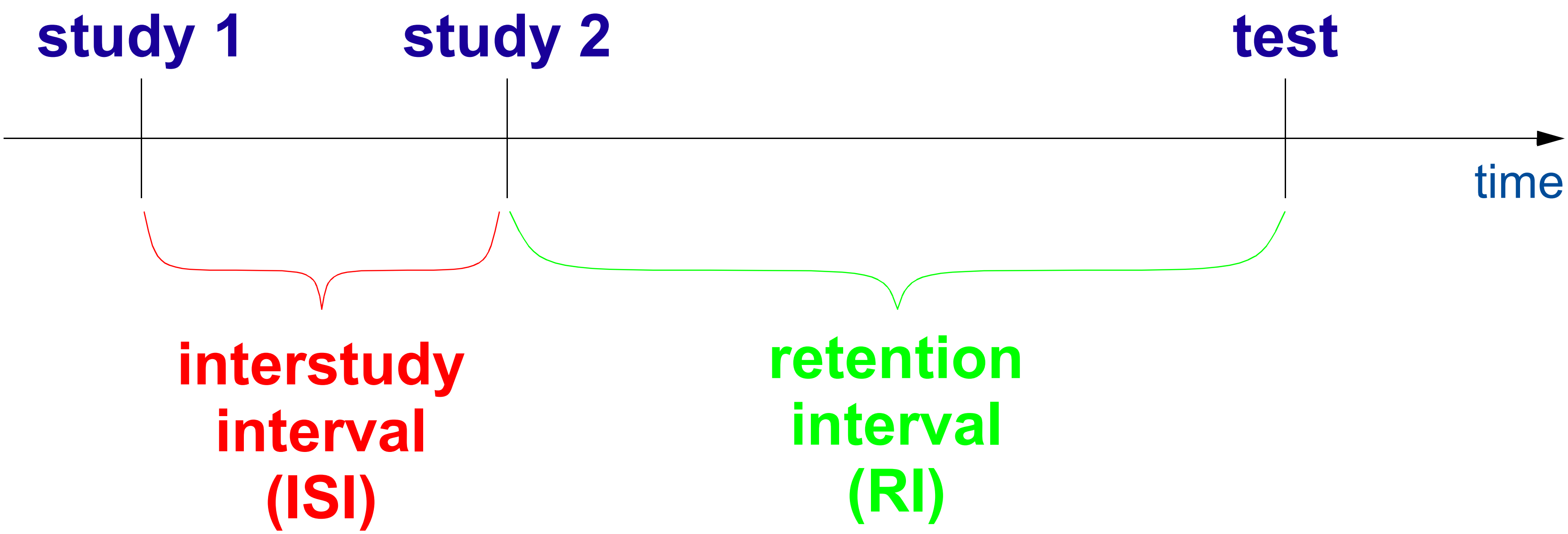
University of Colorado

Harold Pashler

UCSD

Reminder: The Basic Paradigm

Reminder: The Basic Paradigm



Rich Theoretical Literature Attempts to Explain Distributed Practice Effect

Rich Theoretical Literature Attempts to Explain Distributed Practice Effect

- **Encoding variability**
- **Predictive utility**

Rich Theoretical Literature Attempts to Explain Distributed Practice Effect

- **Encoding variability** — Raaijmakkers (2003)
- **Predictive utility** — Staddon, Chelaru, & Higa (2002)

Rich Theoretical Literature Attempts to Explain Distributed Practice Effect

- Encoding variability
- Predictive utility

Raaijmakkers (2003)

+

Staddon, Chelaru, & Higa (2002)

=

cool story about the
temporal dynamics of memory

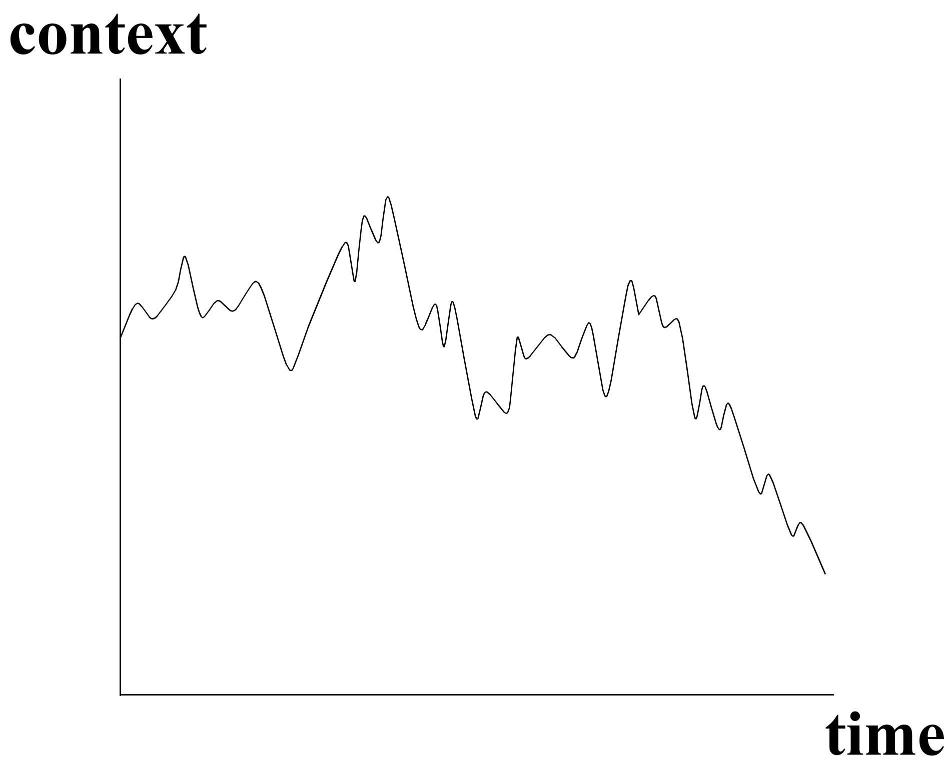
Encoding Variability Theories

Encoding Variability Theories

Each study episode, a separate trace is laid down.

The trace includes a psychological *context.*

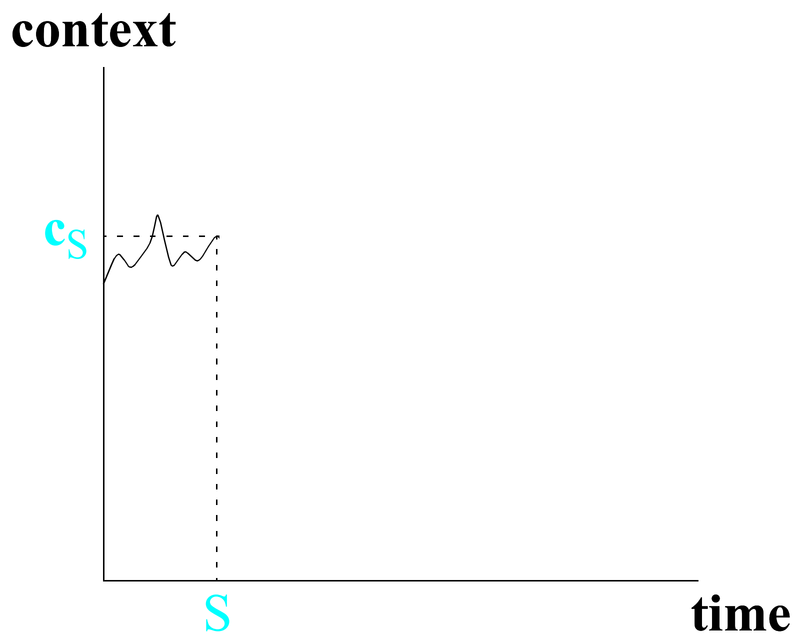
Context wanders over time.



Encoding Variability Explains Forgetting

Encoding Variability Explains Forgetting

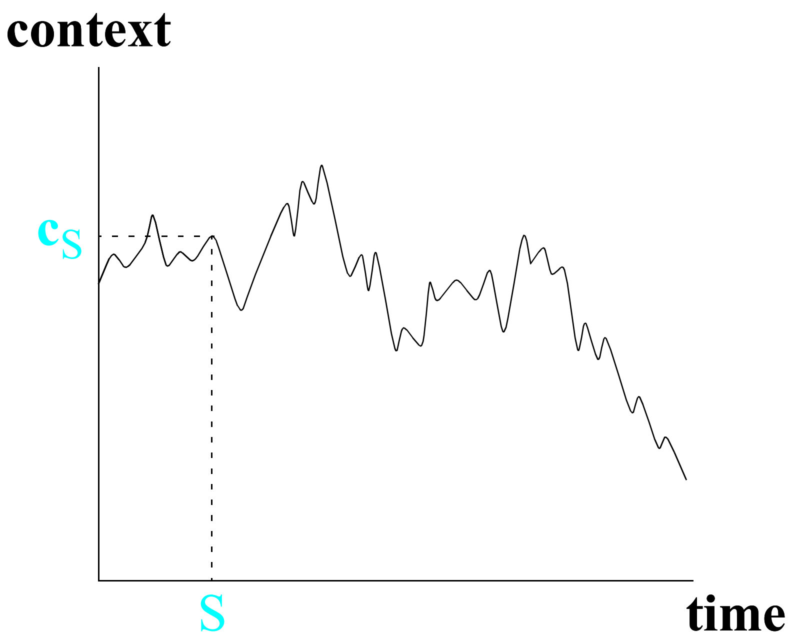
Study item at S



Encoding Variability Explains Forgetting

Study item at S

During retention interval, context wanders



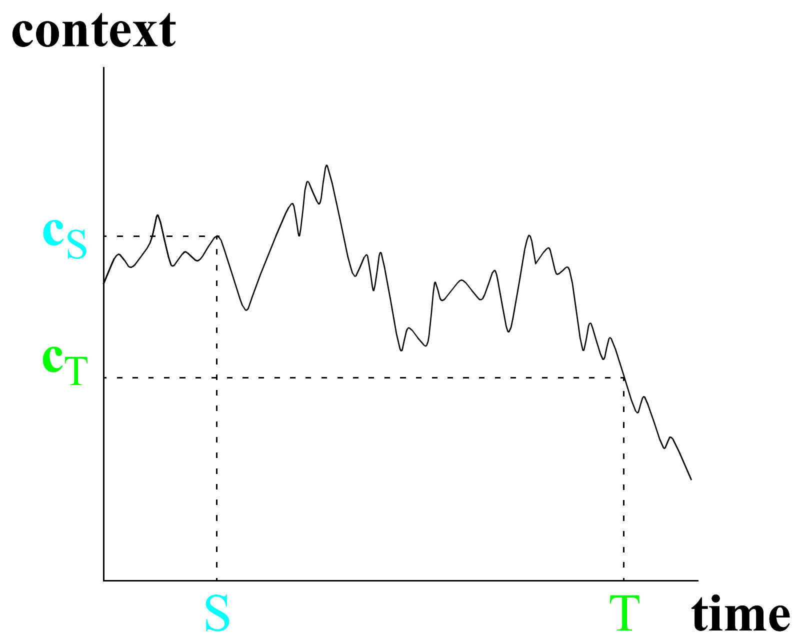
Encoding Variability Explains Forgetting

Study item at S

During retention interval, context wanders

Test at T

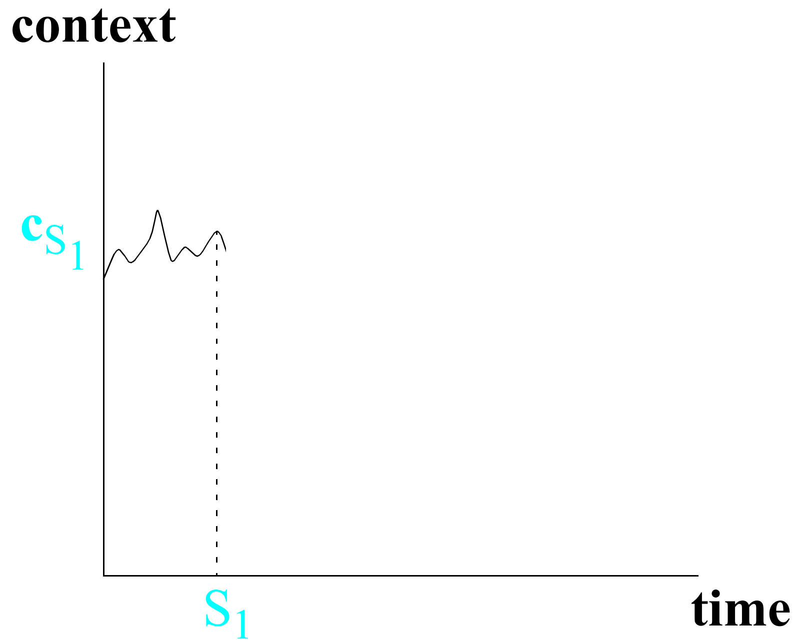
Retrieval success depends on similarity of c_T and c_S



Encoding Variability Explains DP Effect

Encoding Variability Explains DP Effect

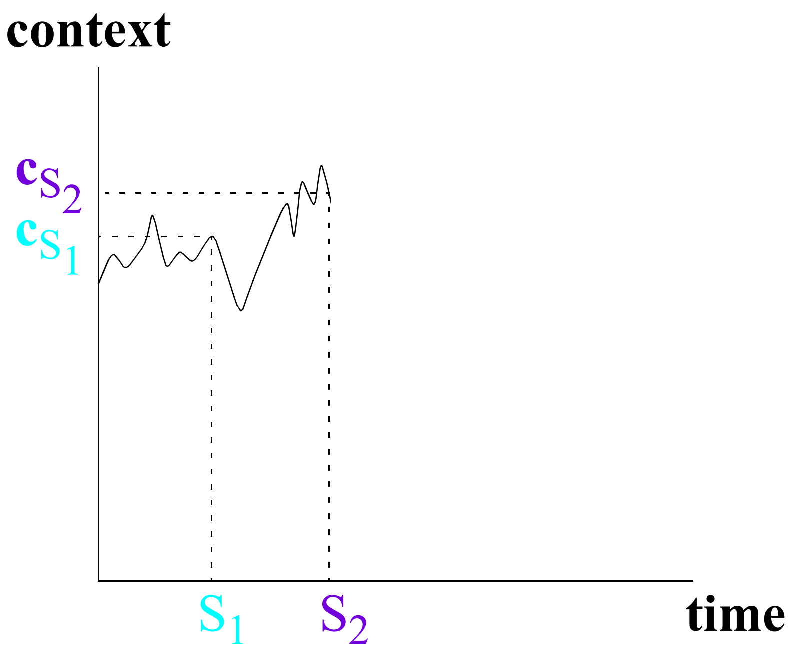
Study item at S1



Encoding Variability Explains DP Effect

Study item at S1

Study item at S2

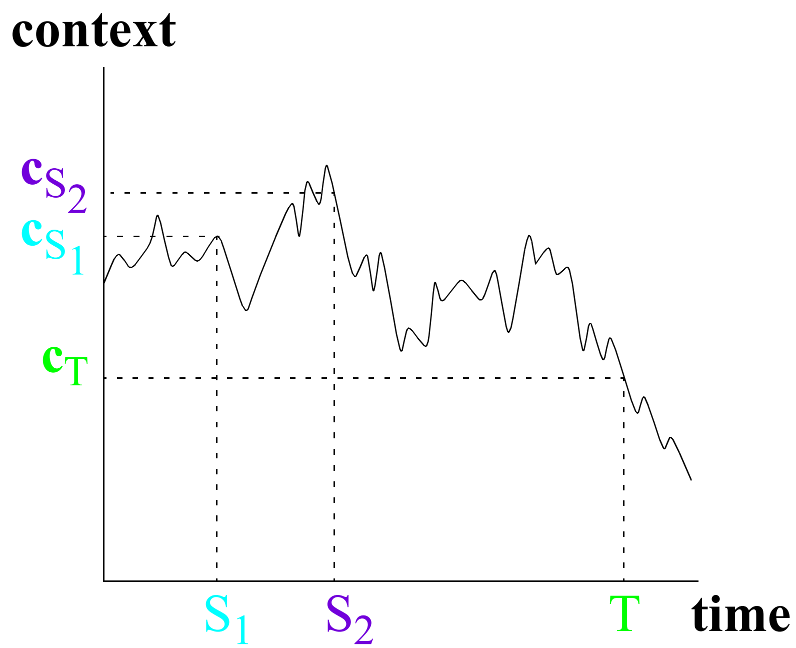


Encoding Variability Explains DP Effect

Study item at S1

Study item at S2

Test at T

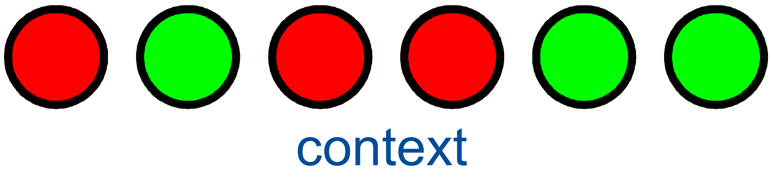


Retrieval success at T depends on similarity of c_T to either c_{S1} or c_{S2}

Disadvantage for small ISIs: redundancy of c_{S1} and c_{S2}.

Raaijmakkers (2003)

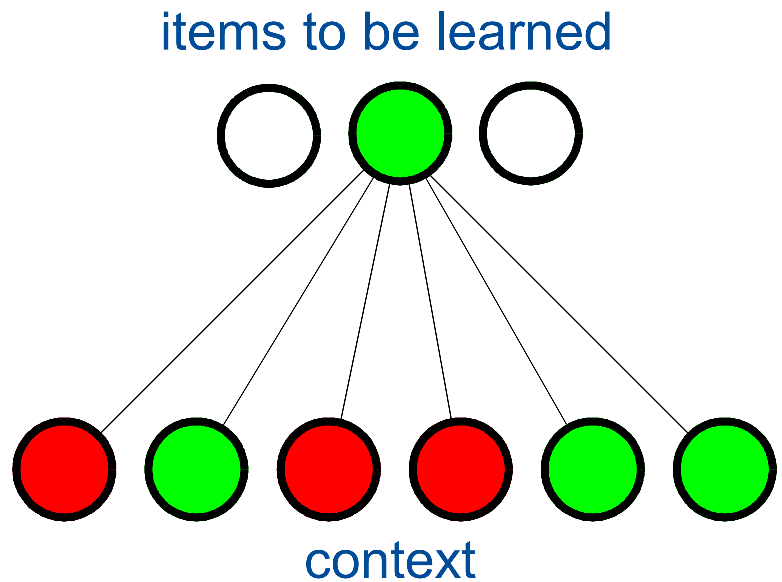
Context is represented by pool of binary valued neurons.



Raaijmakkers (2003)

Context is represented by pool of binary valued neurons.

Each item to be learned represented by an output neuron.

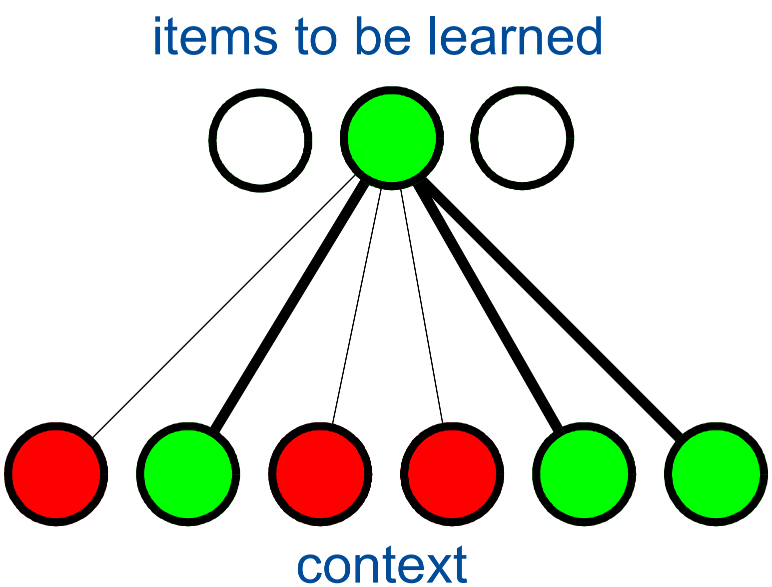


Raaijmakkers (2003)

Context is represented by pool of binary valued neurons.

Each item to be learned represented by an output neuron.

Hebbian learning rule



Raaijmakkers (2003)

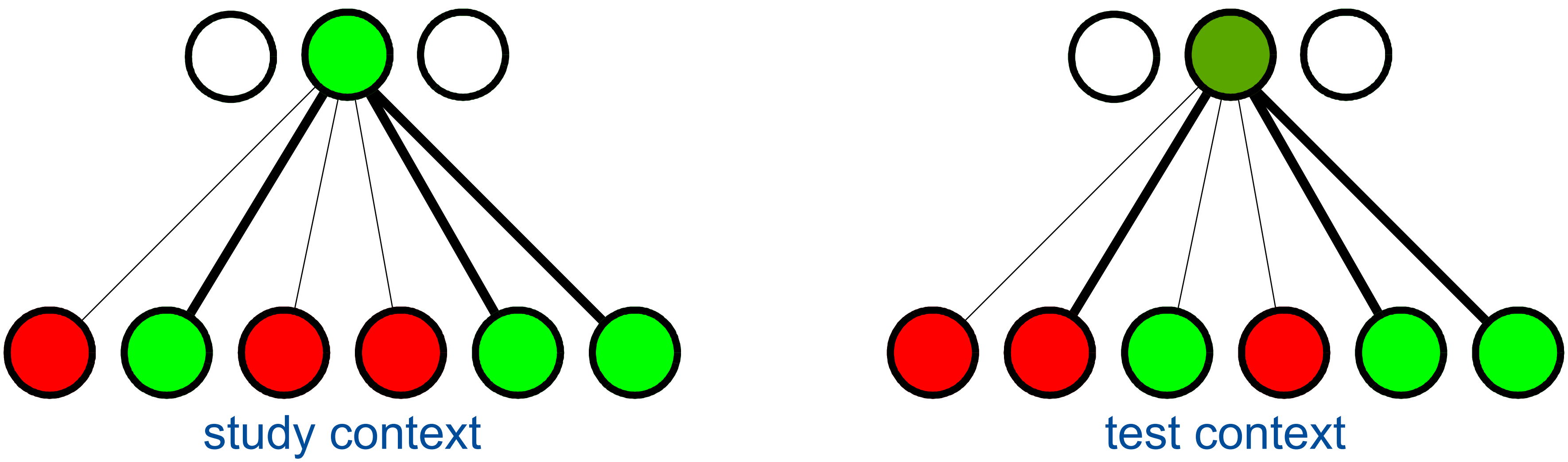
Context is represented by pool of binary valued neurons.

Each item to be learned represented by an output neuron.

Hebbian learning rule

Output activity at test ~ recall probability

depends on similarity of study and test contexts



Raaijmakkers (2003)

Context is represented by pool of binary valued neurons.

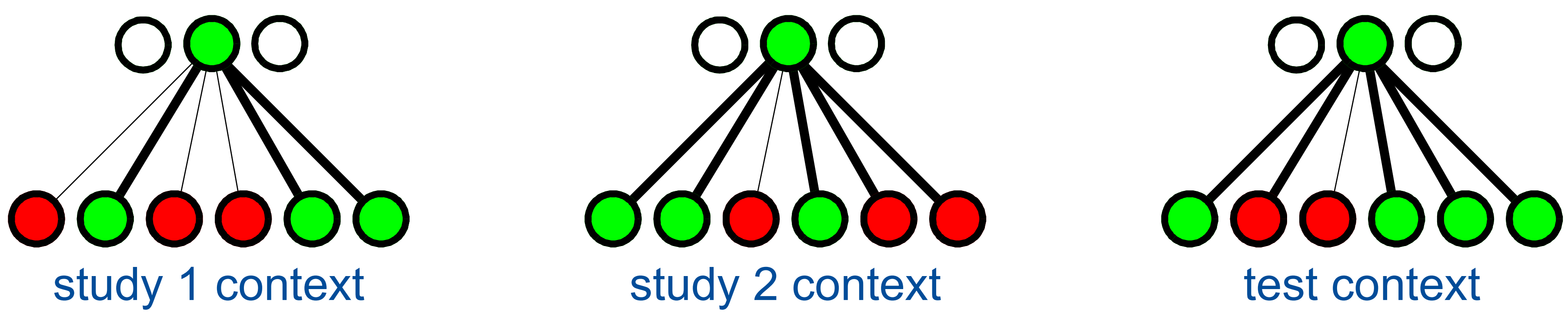
Each item to be learned represented by an output neuron.

Hebbian learning rule

Output activity at test ~ recall probability

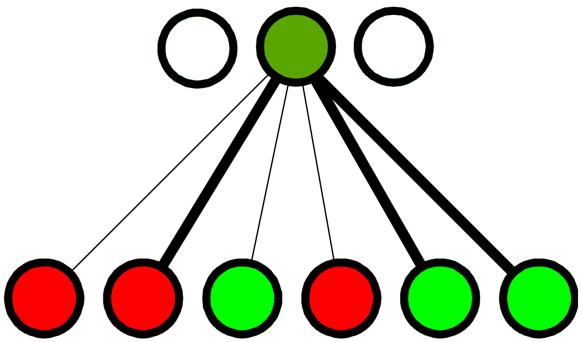
depends on similarity of study and test contexts

Multiple study opportunities ⇒ context variability
⇒ robust recall



Raaijmakkers (2003): Formal Description

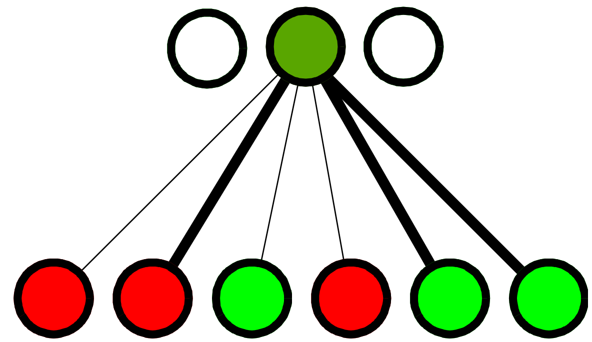
Retrieval at test facilitated when context unit active at both study and test.



Expected output neuron activity ~
P(retrieval) ~ P(C_S = 1 & C_T = 1)

Raaijmakkers (2003): Formal Description

Retrieval at test facilitated when context unit active at both study and test.



Expected output neuron activity ~
P(retrieval) ~ $P(C_S = 1 \,\&\, C_T = 1)$

How does context wander over time?

context bits flip from off to on at rate μ_{01}

context bits flip from on to off at rate μ_{10}

$$P(C_S = 1 \,\&\, C_T = 1) = \beta^2 + \beta(1-\beta) \exp(-\alpha \, RI)$$

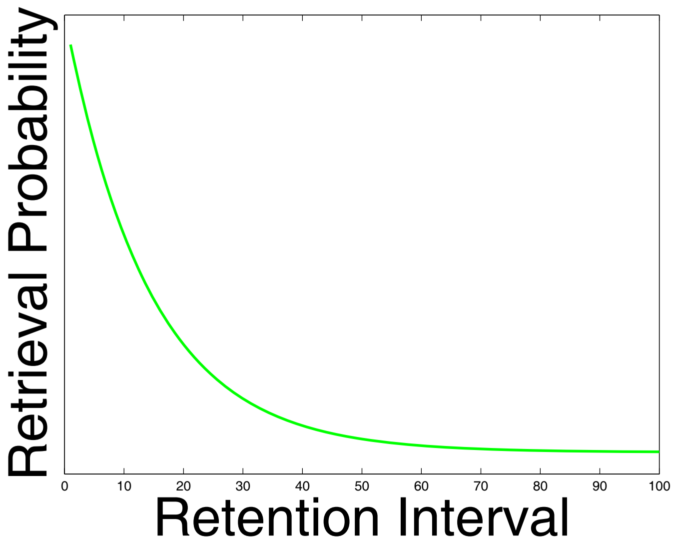
retention interval

flip rate: $\mu_{01} + \mu_{10}$

proportion on : $\mu_{01} / (\mu_{01} + \mu_{10})$

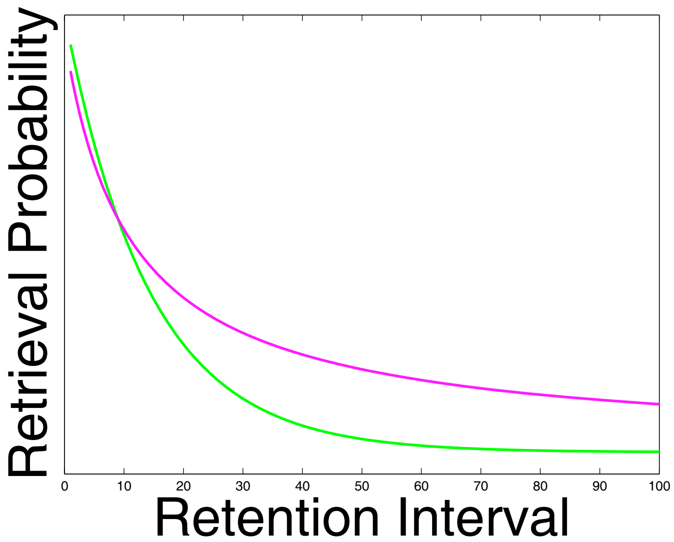
What It Boils Down To

Forgetting function is exponential



What It Boils Down To

Forgetting function is exponential

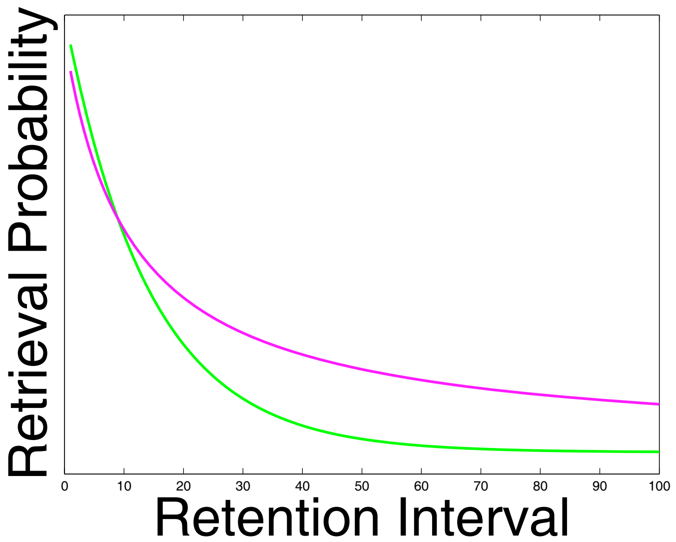


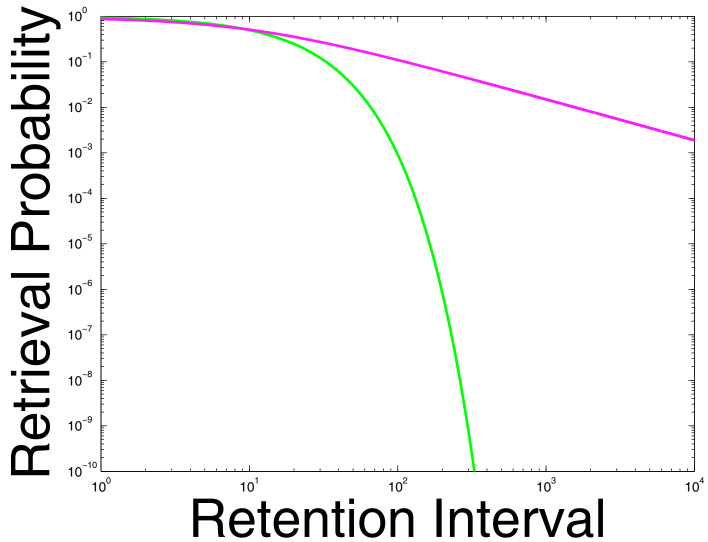
Human forgetting functions follow a power law (Wickelgren, 1974; Wixted & Carpenter, 2007):

$$P(\text{retrieval}) = \lambda(1 + \varphi\,\text{RI})^{-\phi}$$

What It Boils Down To

Forgetting function is exponential





Human forgetting functions follow a power law (Wickelgren, 1974; Wixted & Carpenter, 2007):

$P(\text{retrieval}) = \lambda(1 + \varphi\, RI)^{-\phi}$

Power law shows scale invariance

I.e., memory shows same properties at different time scales

Is it a problem that Raaijmakkers' (2003) model doesn't show scale invariance?

Yes, distributed practice effects are scale invariant.

Model has other problems too.

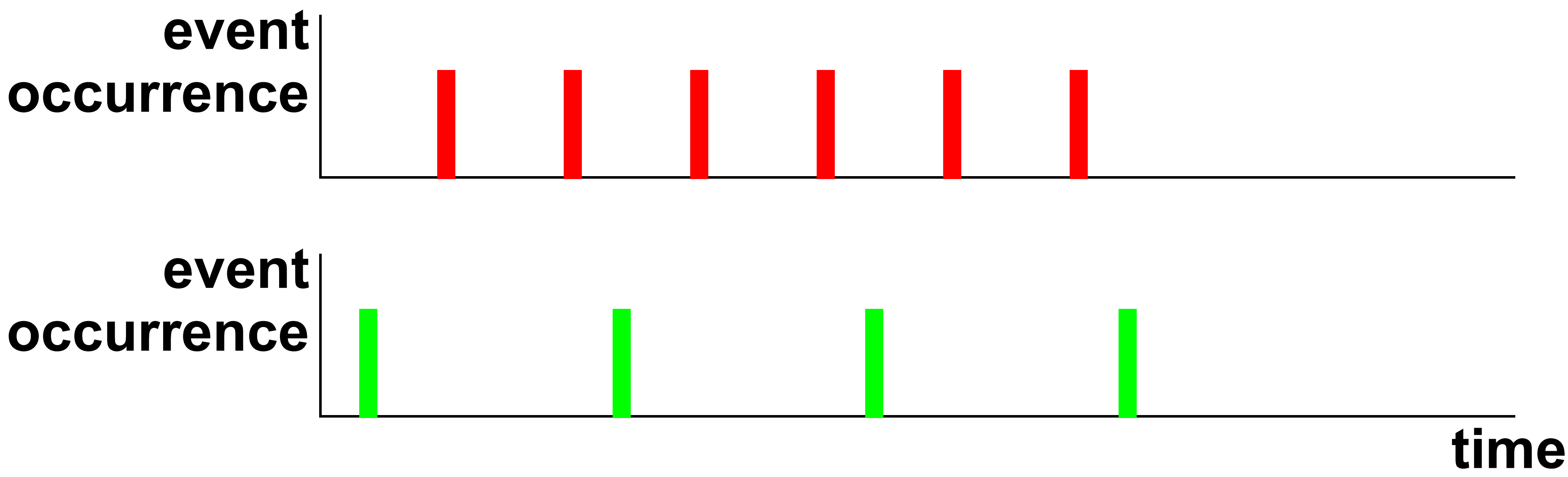
- Many free parameters and ugly hacks
- Doesn't fit data particularly well

Predictive Utility Theories

Predictive Utility Theories

Suppose that memory
- **is limited in capacity, and/or**
- **is imperfect and allows intrusions.**

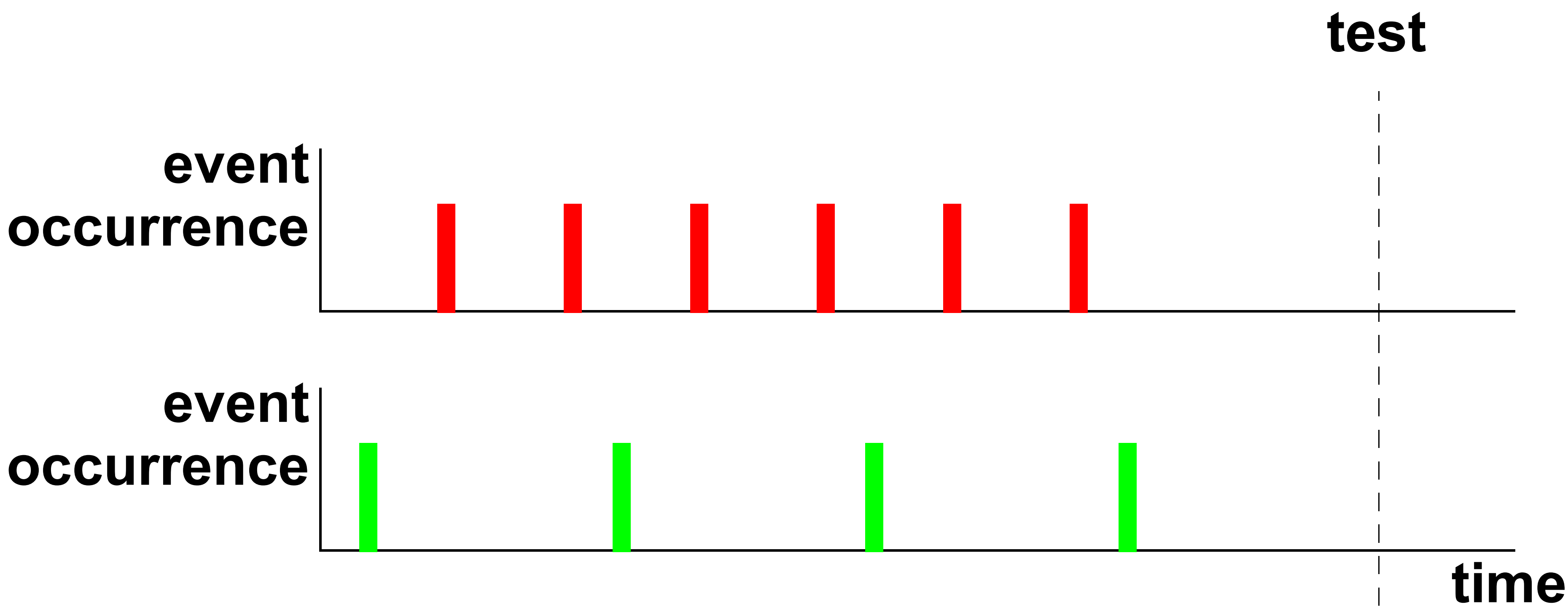
To achieve optimal performance, memories should be erased if they are not likely to be needed in the future.



Predictive Utility Theories

Suppose that memory
- **is limited in capacity, and/or**
- **is imperfect and allows intrusions.**

To achieve optimal performance, memories should be erased if they are not likely to be needed in the future.

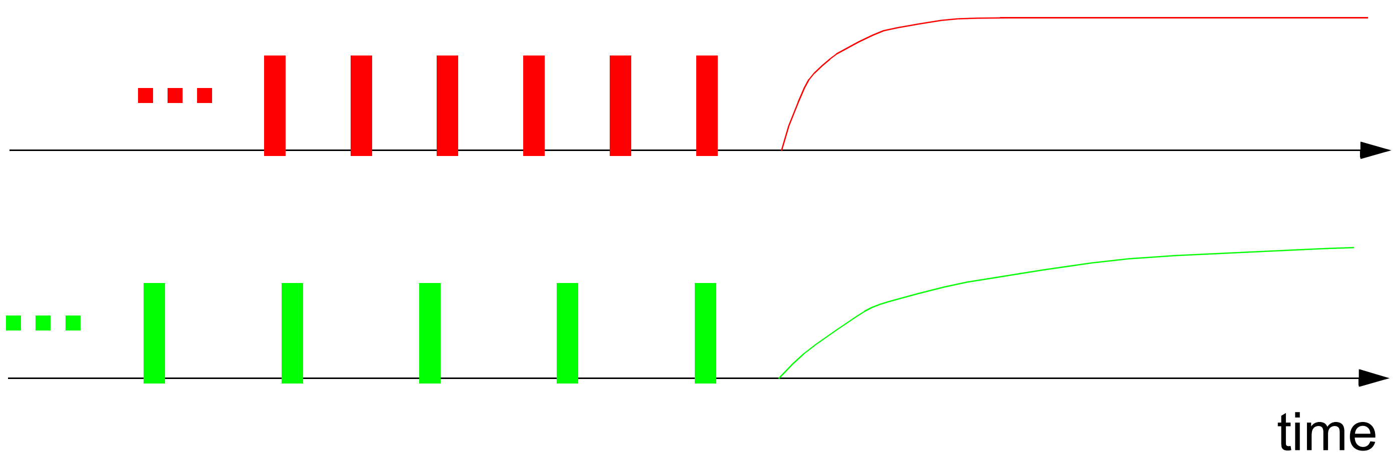


Staddon, Chelaru, & Higa (2002)

Rats habituate to a repeated stream of stimuli.

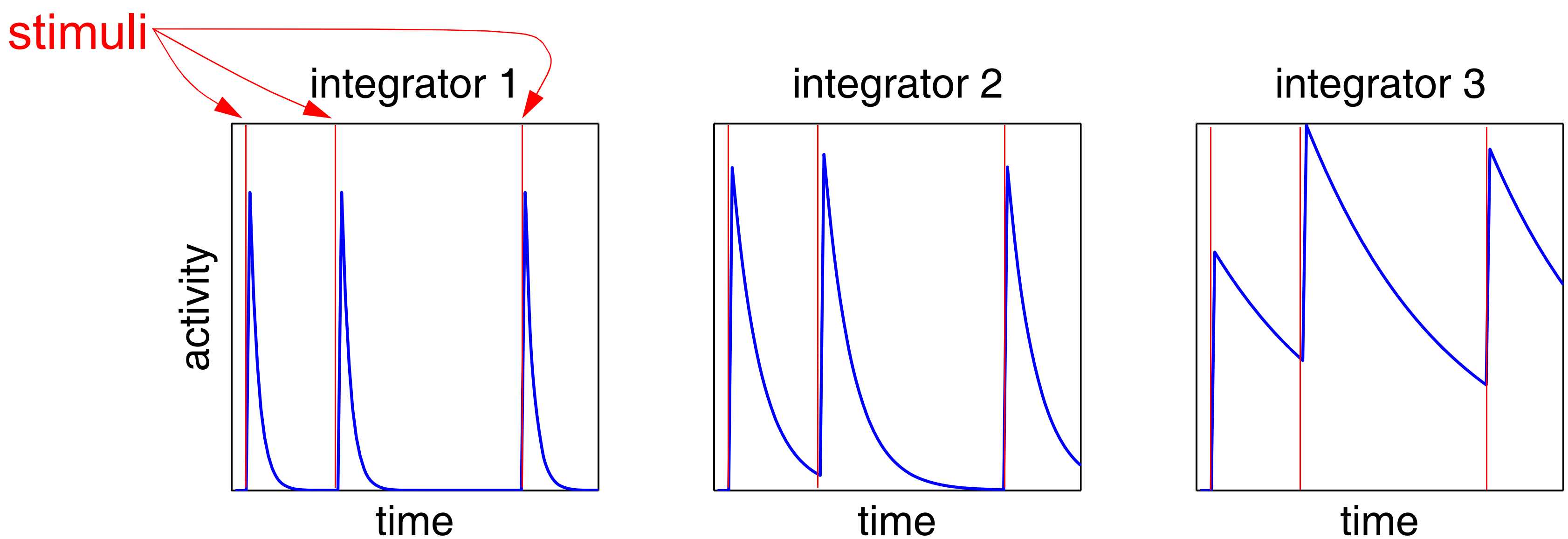
Time for recovery from habituation ~ rate of stimuli

Longer-lasting memory for stimuli delivered at slower rate



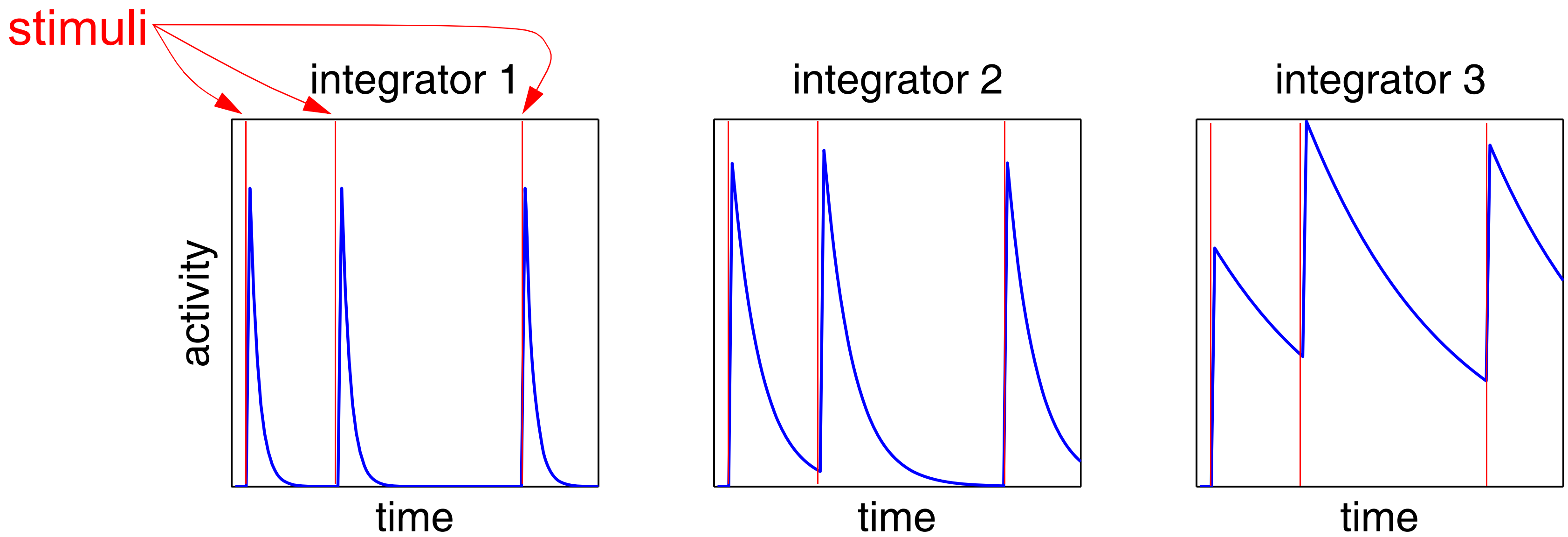
Staddon, Chelaru, & Higa (2002)

Each item to be learned represented by memory consisting of *leaky integrators* at multiple time scales.



Staddon, Chelaru, & Higa (2002)

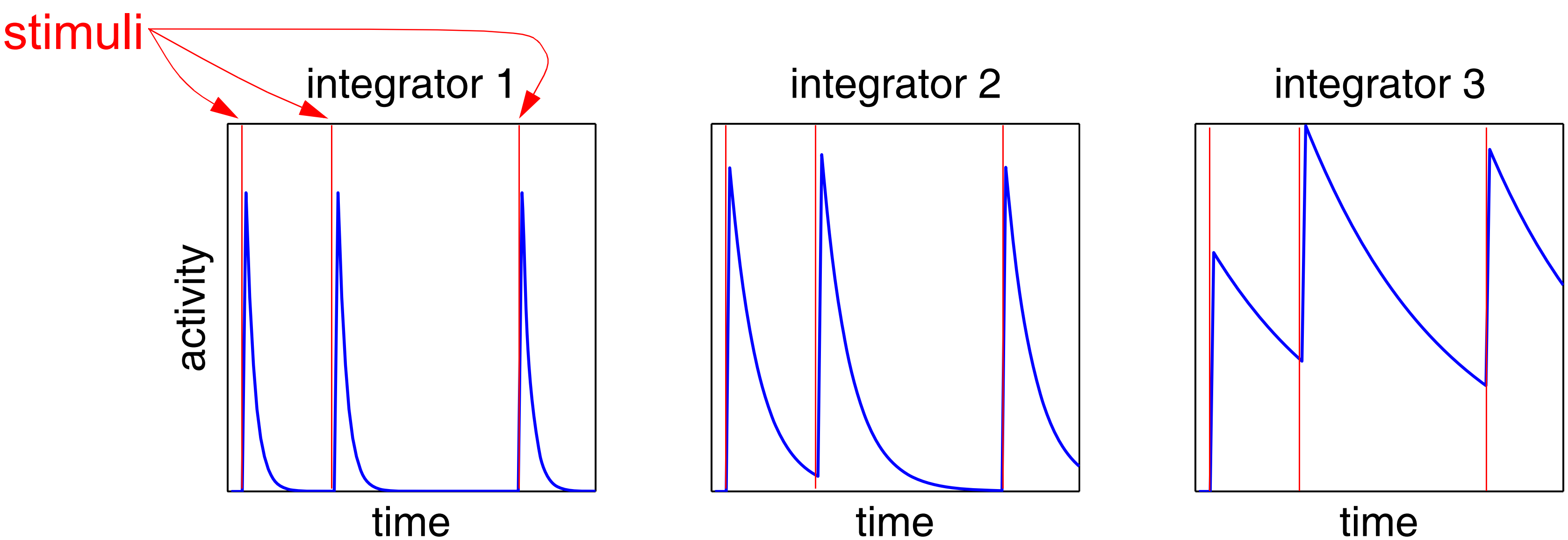
Each item to be learned represented by memory consisting of *leaky integrators* at multiple time scales.



Memory trace is the sum of the integrator activities.

Staddon, Chelaru, & Higa (2002)

Each item to be learned represented by memory consisting of *leaky integrators* at multiple time scales.



Memory trace is the sum of the integrator activities.

Memory storage rule

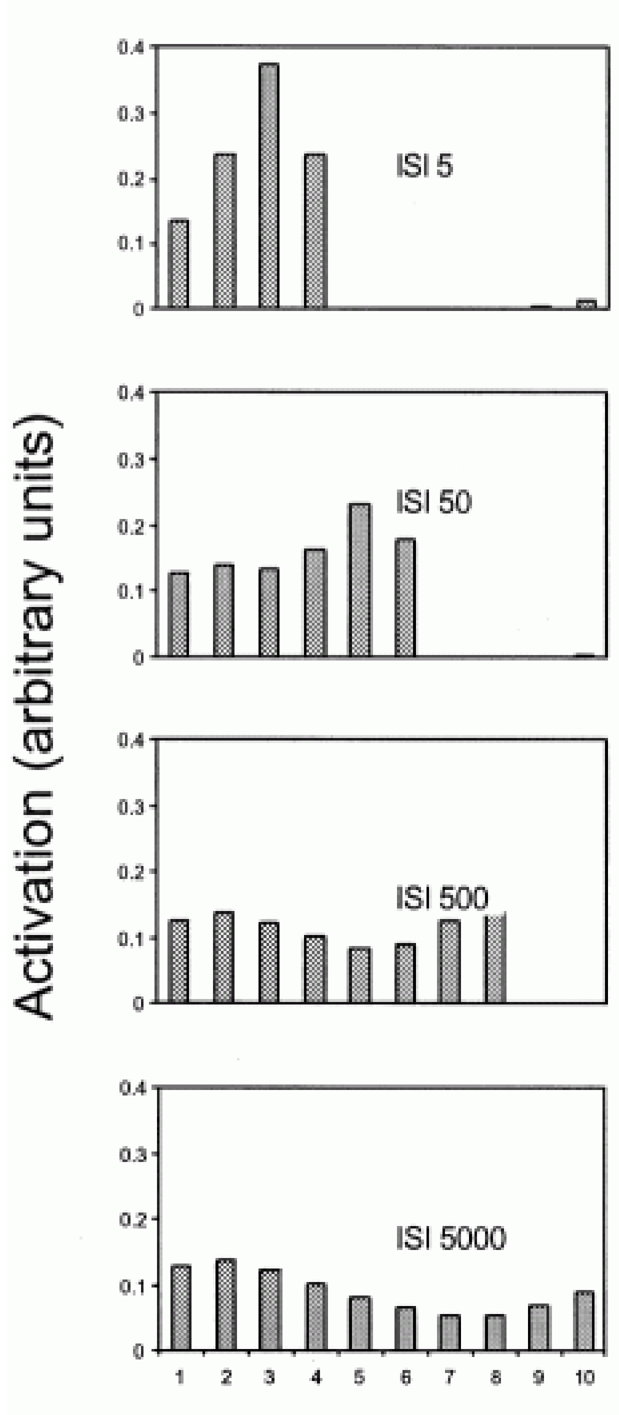
Integrators with long time constants get activated only when integrators with short time constants have decayed.

Example

10 integrators

Stimulus repeatedly presented at various ISIs

Greater spacing ⇒ memory shifts to longer time-scale integrators ⇒ more durable memory



Example

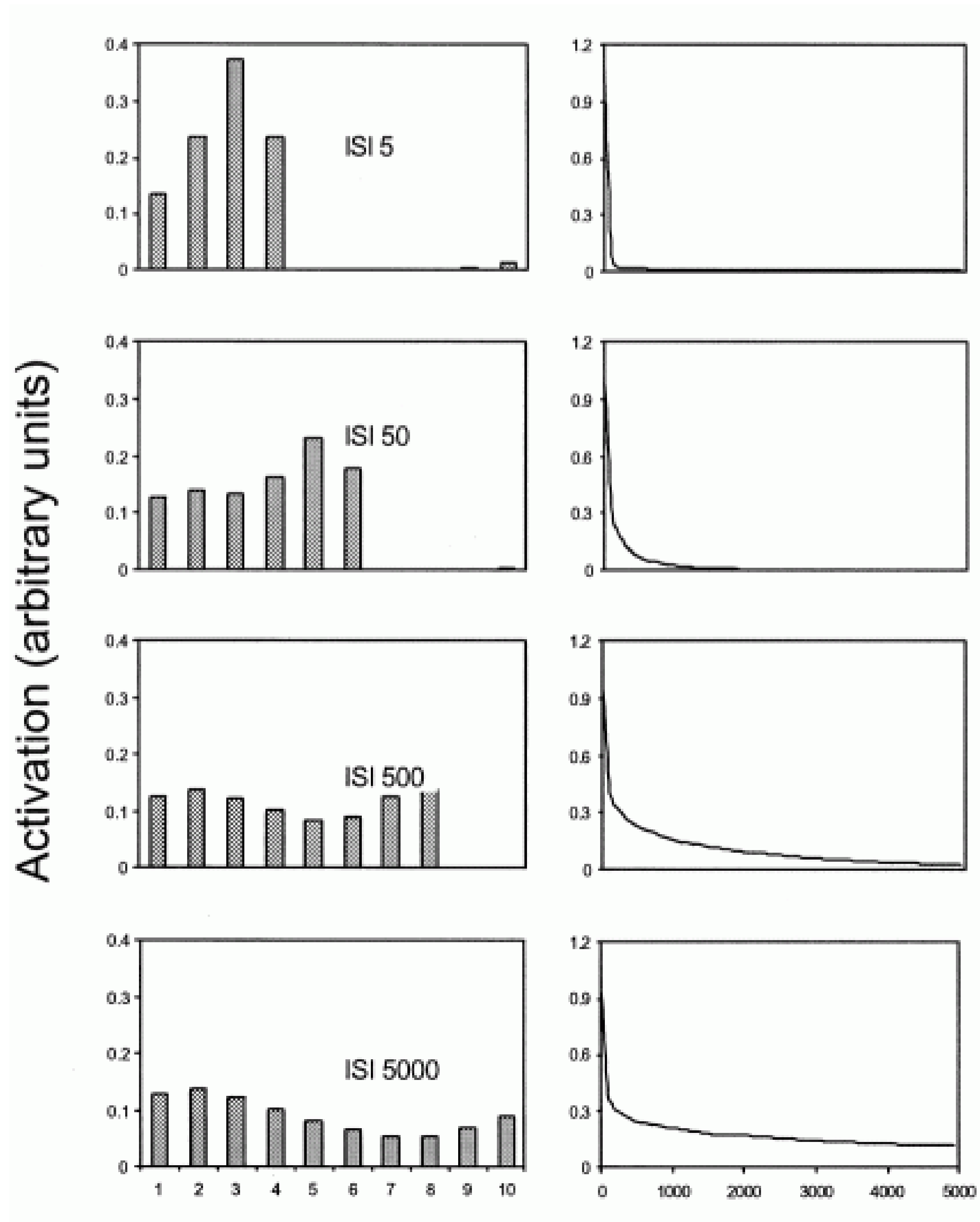
10 integrators

Stimulus repeatedly presented at various ISIs

Greater spacing ⇒ memory shifts to longer time-scale integrators ⇒ more durable memory

Model is sensitive to predictive utility

Slower forgetting following longer ISI stimulus sequences.

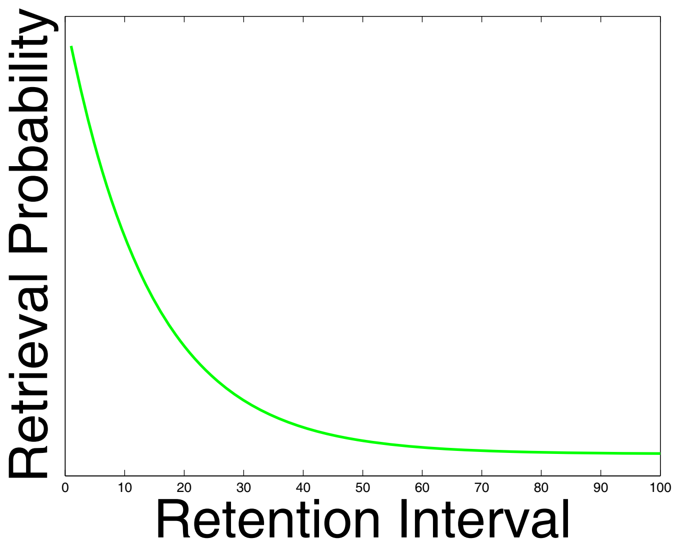


Model was fit to rat habituation and interval timing data,

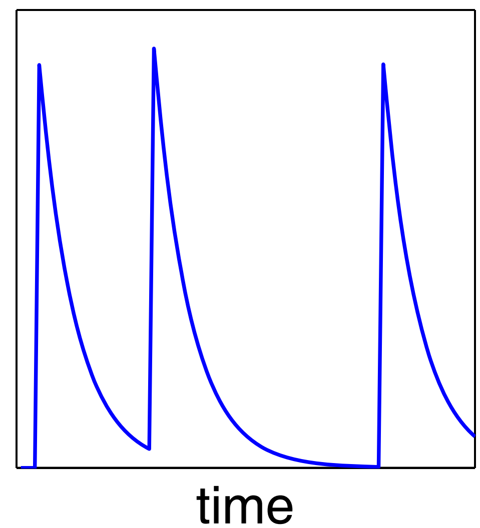
... but isn't sufficiently well specified to explain human studies of distributed practice.

Two Models Share Key Property: Exponential Decay of Internal Representations

encoding variability model
via context drift



predictive utility model
via leaky integrator



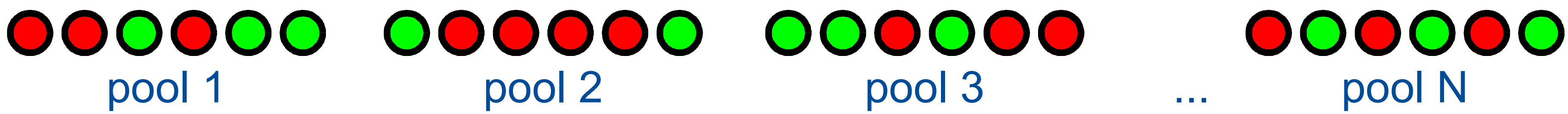
This commonality allows us to integrate the two models.

Combine

- **multiple time-scale representation of Staddon’s model**
- **contextual drift of Raaijmakkers’ model**

→ Multiscale Context Model

Multiscale Context Model (MCM)

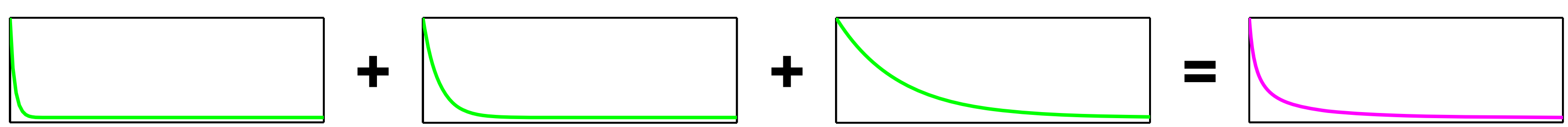


In pool *p*, all units flip state at rate α_p.

The pools can be different sizes:
the relative proportion of units in pool p is γ_p.

Retrieval function is a mixture of exponentials.

$$P(\text{retrieval}) \sim \sum_p \gamma_p \exp(-\alpha_p \mathrm{RI})$$

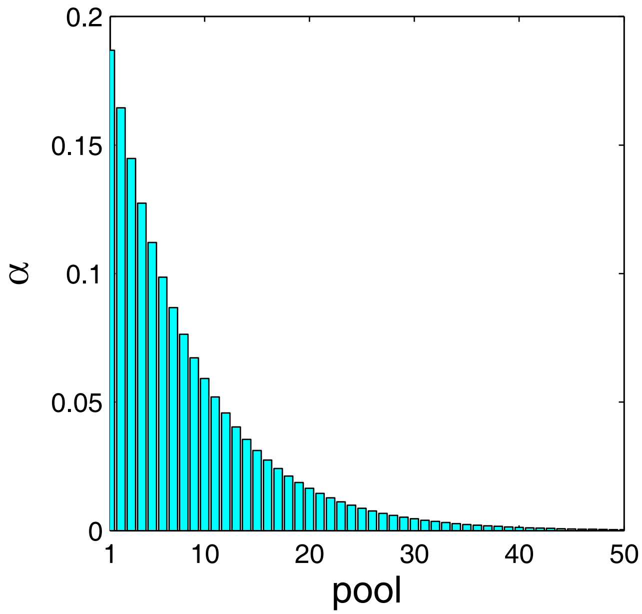


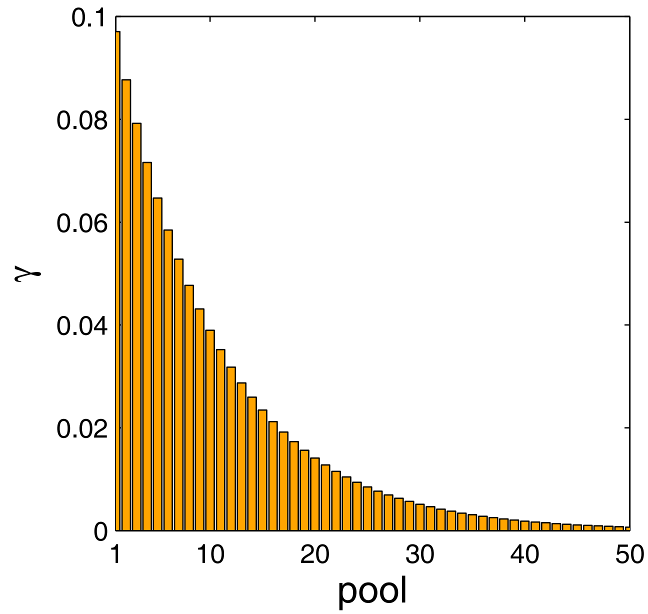
Mixture of exponentials can approximate human forgetting functions (Wixted).

Use Simple Formula to Pick Pool Size (γ) and Rate (α)

$\alpha_p = \mu\, \nu^p$ for $p \in [1, N]$

$\gamma_p = \omega^p$



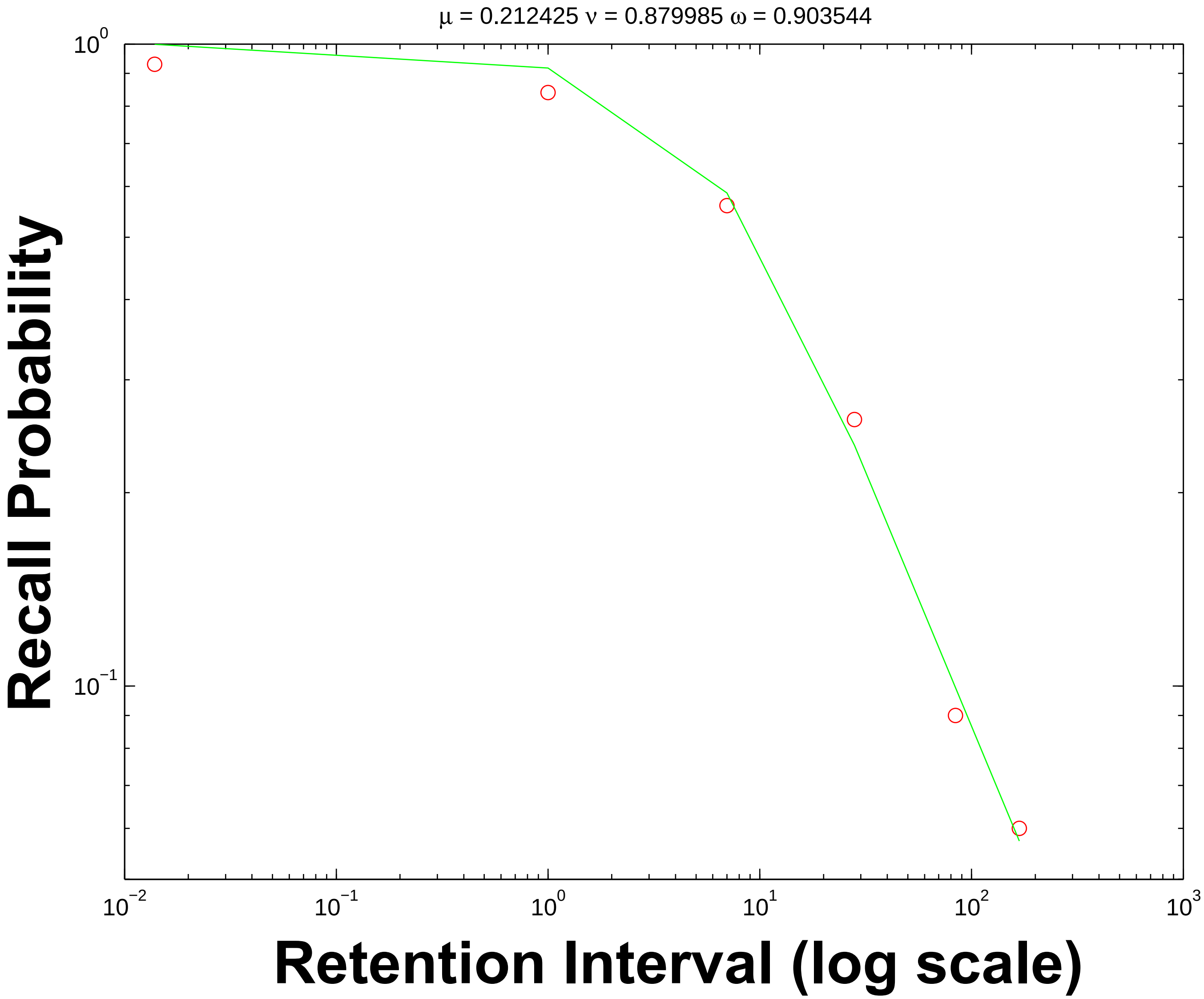


MCM has four free parameters (μ, ν, ω, + one more)

Can we select these parameters such that resulting model yields power law forgetting function and good fits to human data?

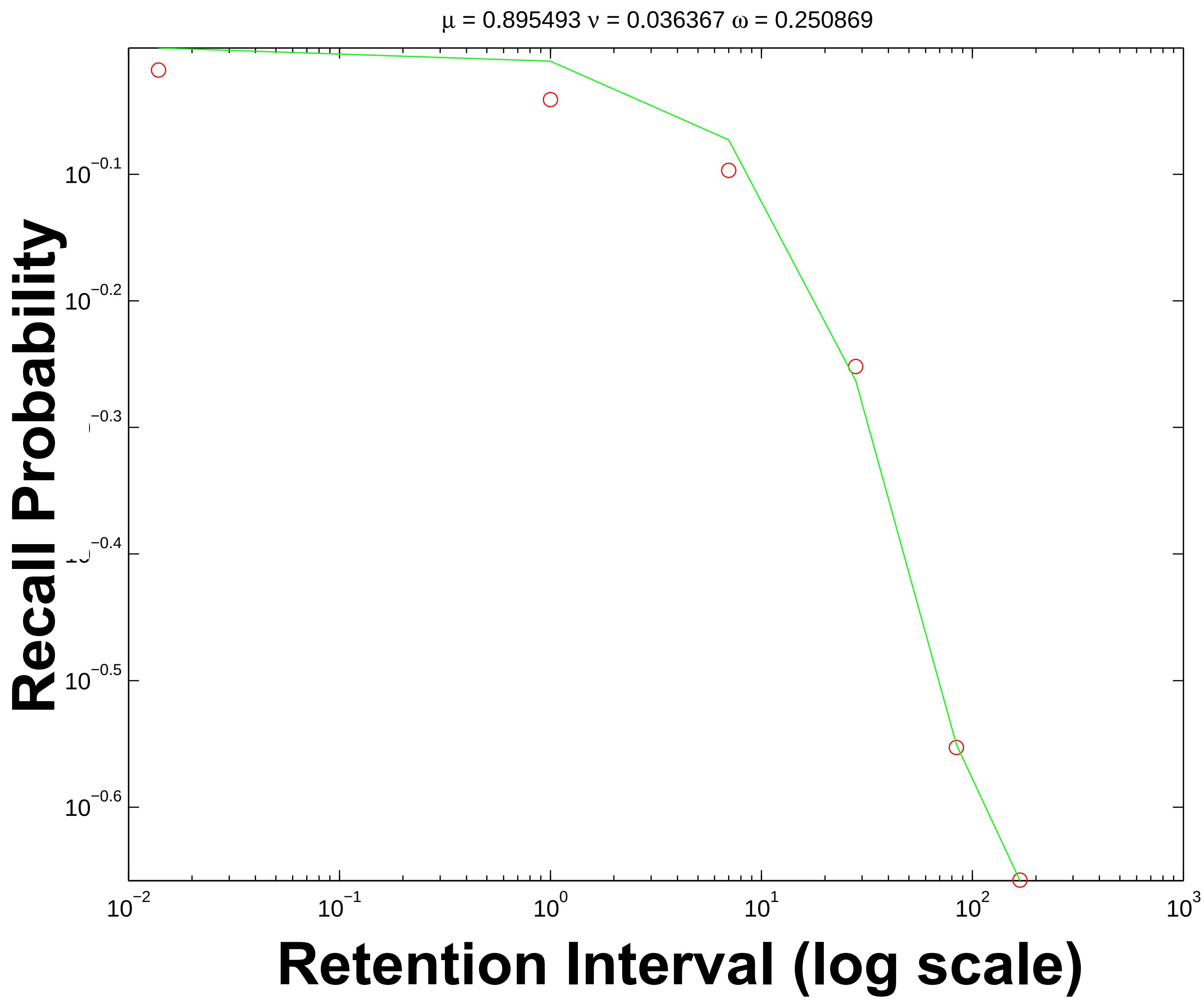
Fitting Forgetting Functions I

Cepeda et al., Expt 2B



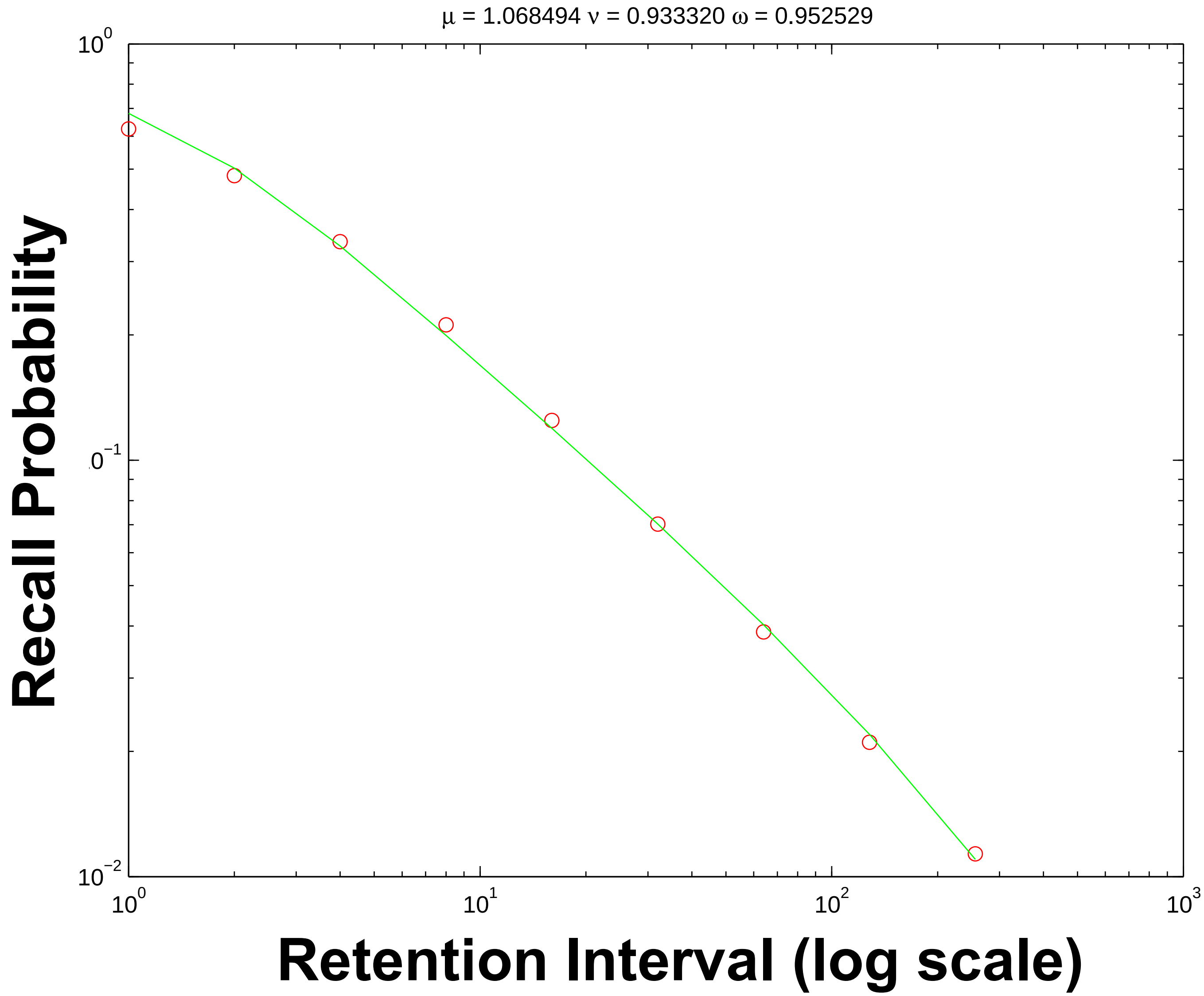
Fitting Forgetting Functions II

Cepeda et al., Expt 2A



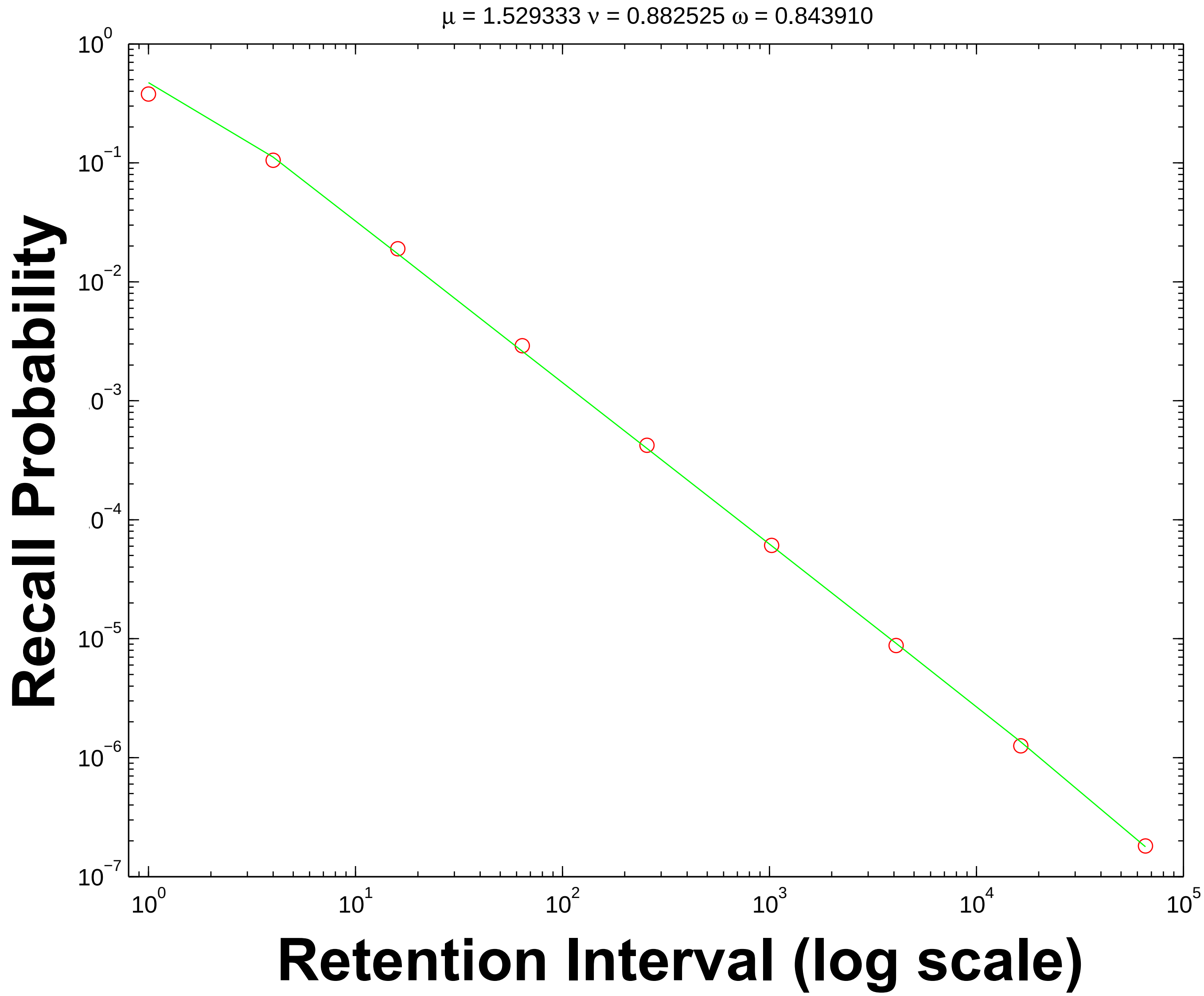
Fitting Forgetting Functions III

P(recall) = .9(1 + .5 t)$^{-0.9}$



Fitting Forgetting Functions IV

$P(\text{recall}) = (1 + t)^{-1.4}$



Multiscale Context Model: A Convergence of Theories

Multiscale Context Model: A Convergence of Theories

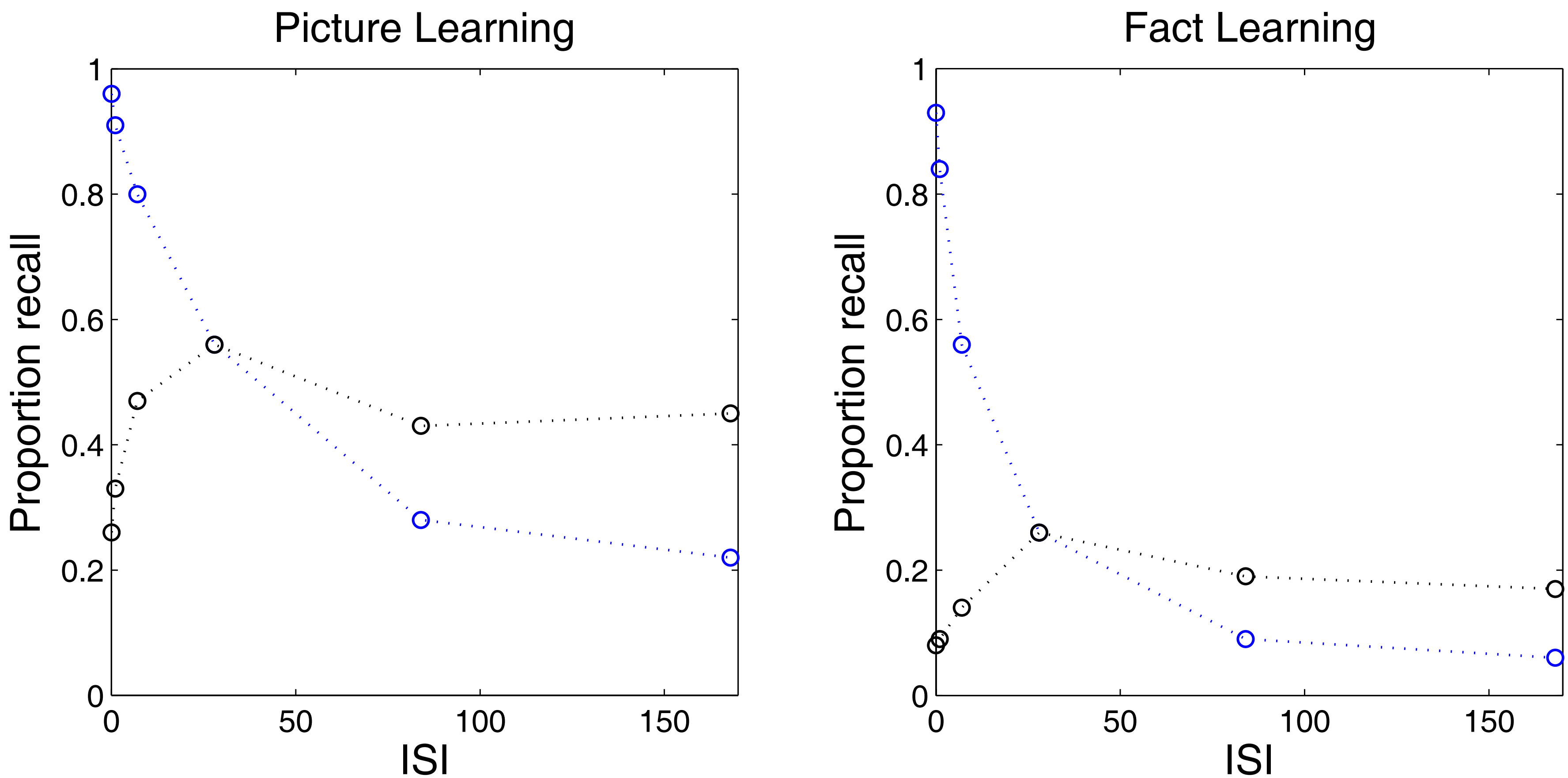
	Raaijmakkers (2003)	Staddon et al. (2002)
context drift	X	
multiple time-scale representation		X
learning rule	X (dependence of learning on retrieval success)	X (cascaded error correction)

Multiscale Context Model: A Convergence of Theories

	Raaijmakkers (2003)	Staddon et al. (2002)	Our Contribution
context drift	X		
multiple time-scale representation		X	
learning rule	X (dependence of learning on retrieval success)	X (cascaded error correction)	
variable pool size			X
parameterization of multiscale constants			X
neural characterization			X

Cepeda, Coburn, Rohrer, Wixted, Mozer, & Pashler (in press)

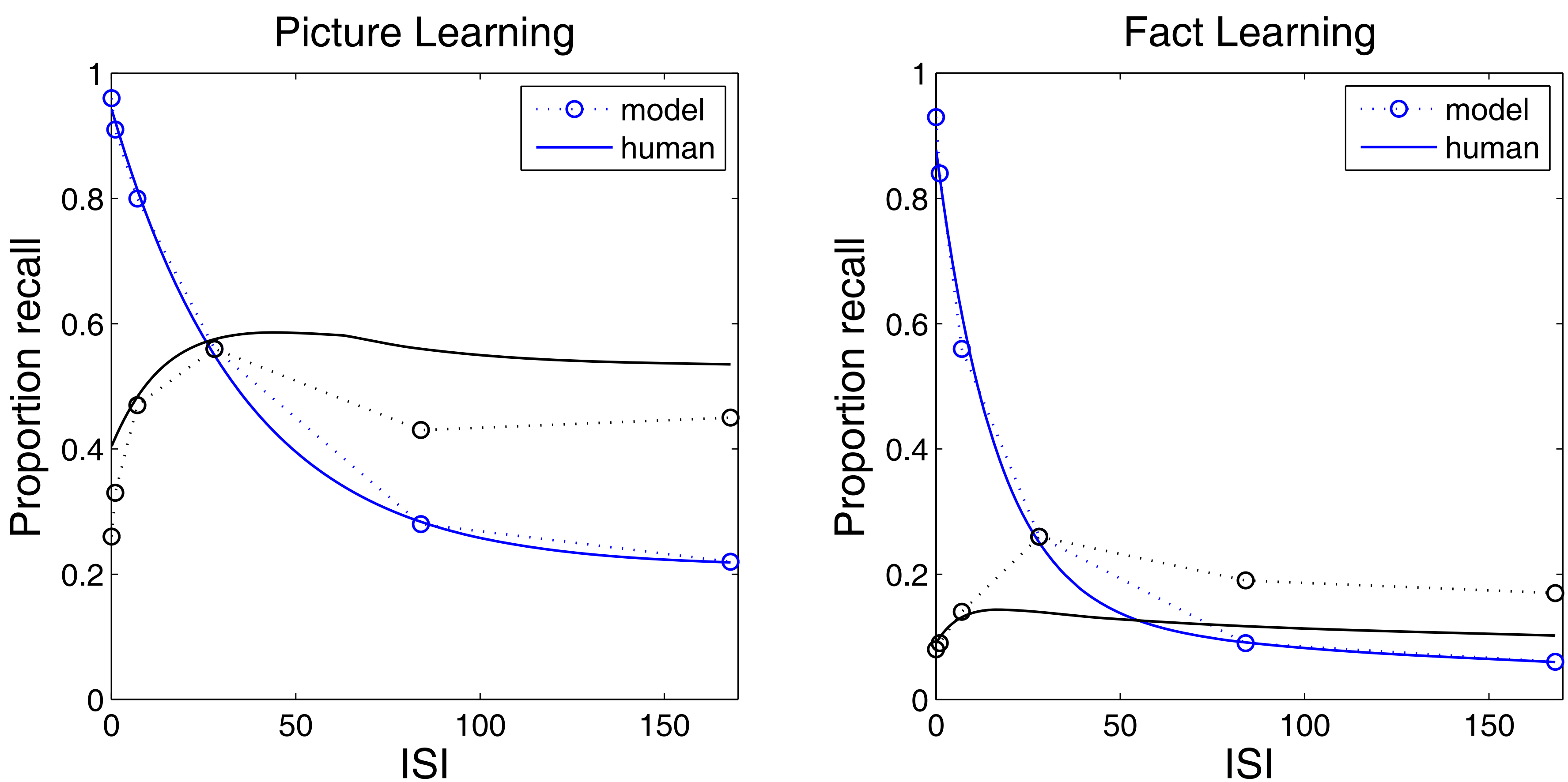
P(recall at study 2)
P(recall at test)



Cepeda, Coburn, Rohrer, Wixted, Mozer, & Pashler (in press)

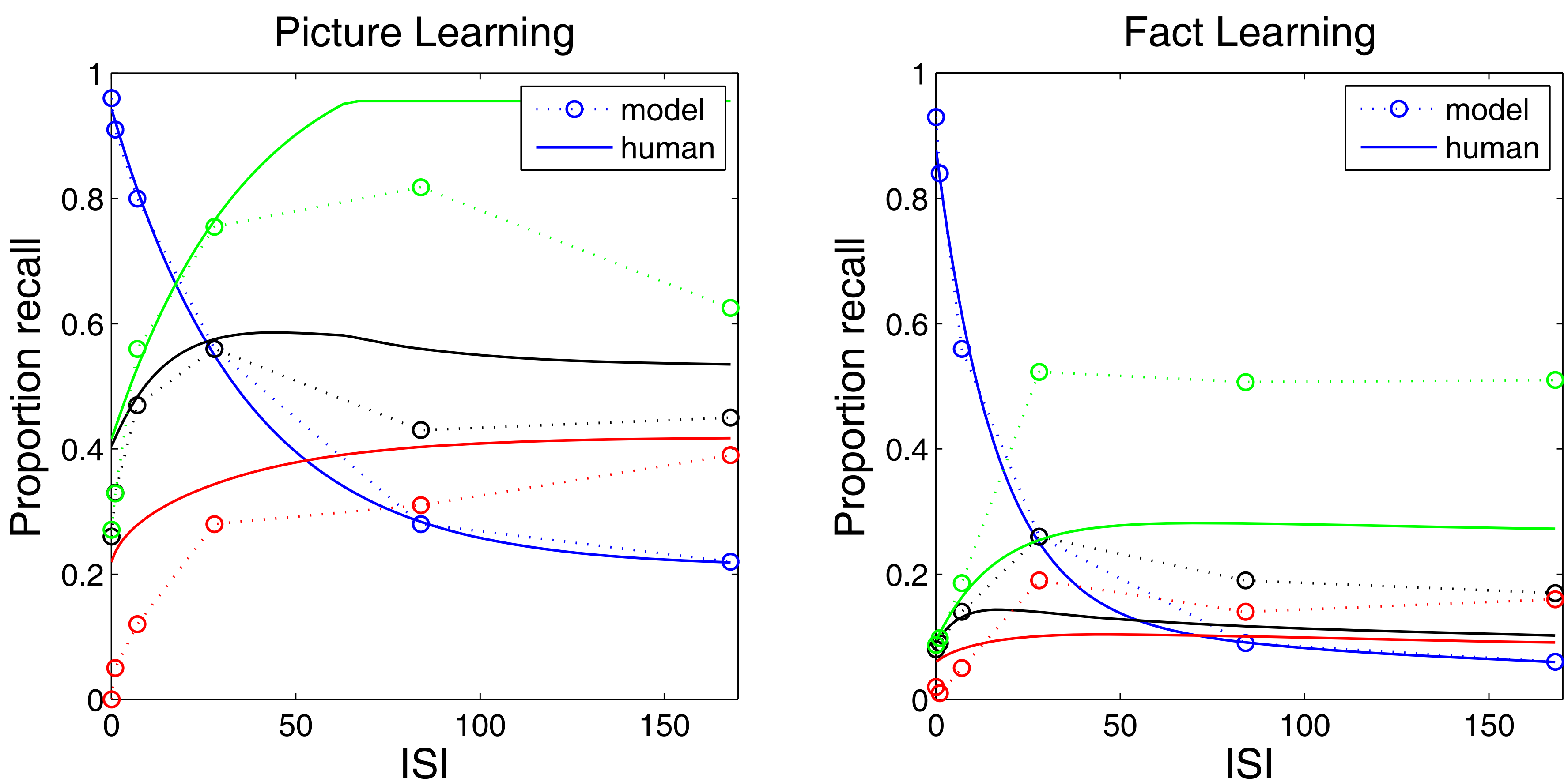
P(recall at study 2)

P(recall at test)



Cepeda, Coburn, Rohrer, Wixted, Mozer, & Pashler (in press)

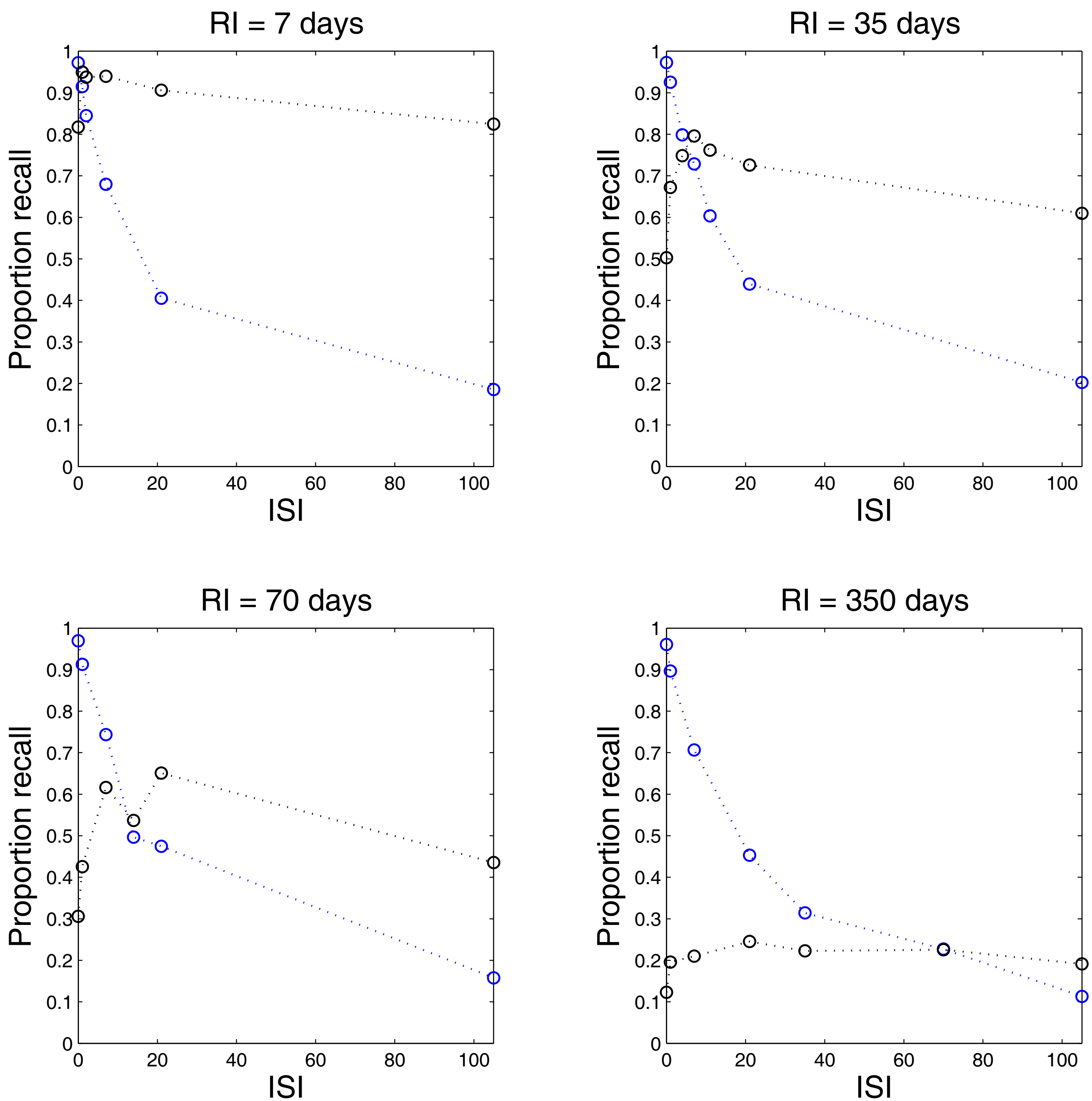
P(recall at study 2)
P(recall at test)
P(recall at test | recall at study 2)
P(recall at test | no recall at study 2)



Simulation of Cepeda, Vul, Rohrer, Wixted, & Pashler (in press)

P(recall at study 2)

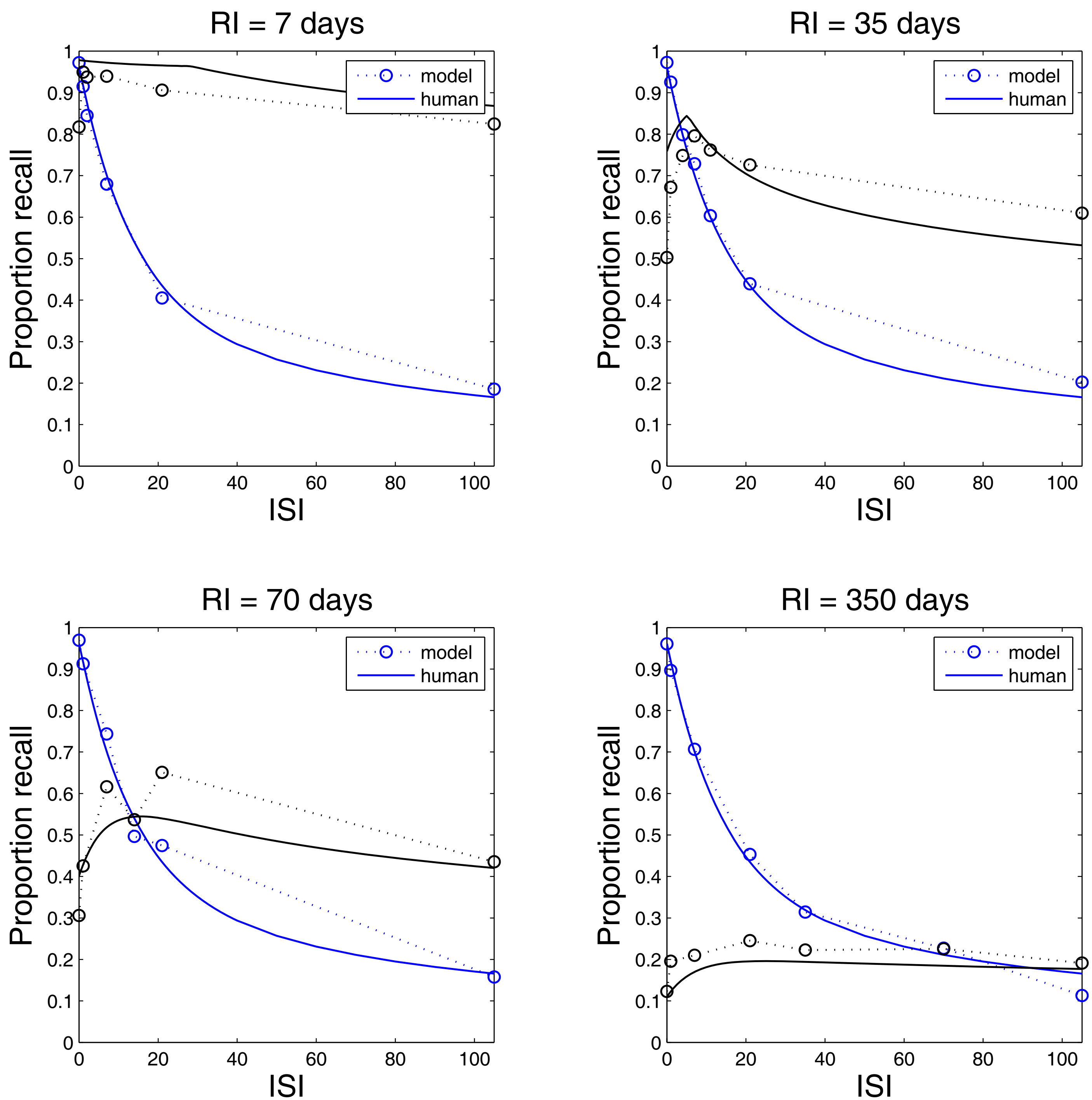
P(recall at test)



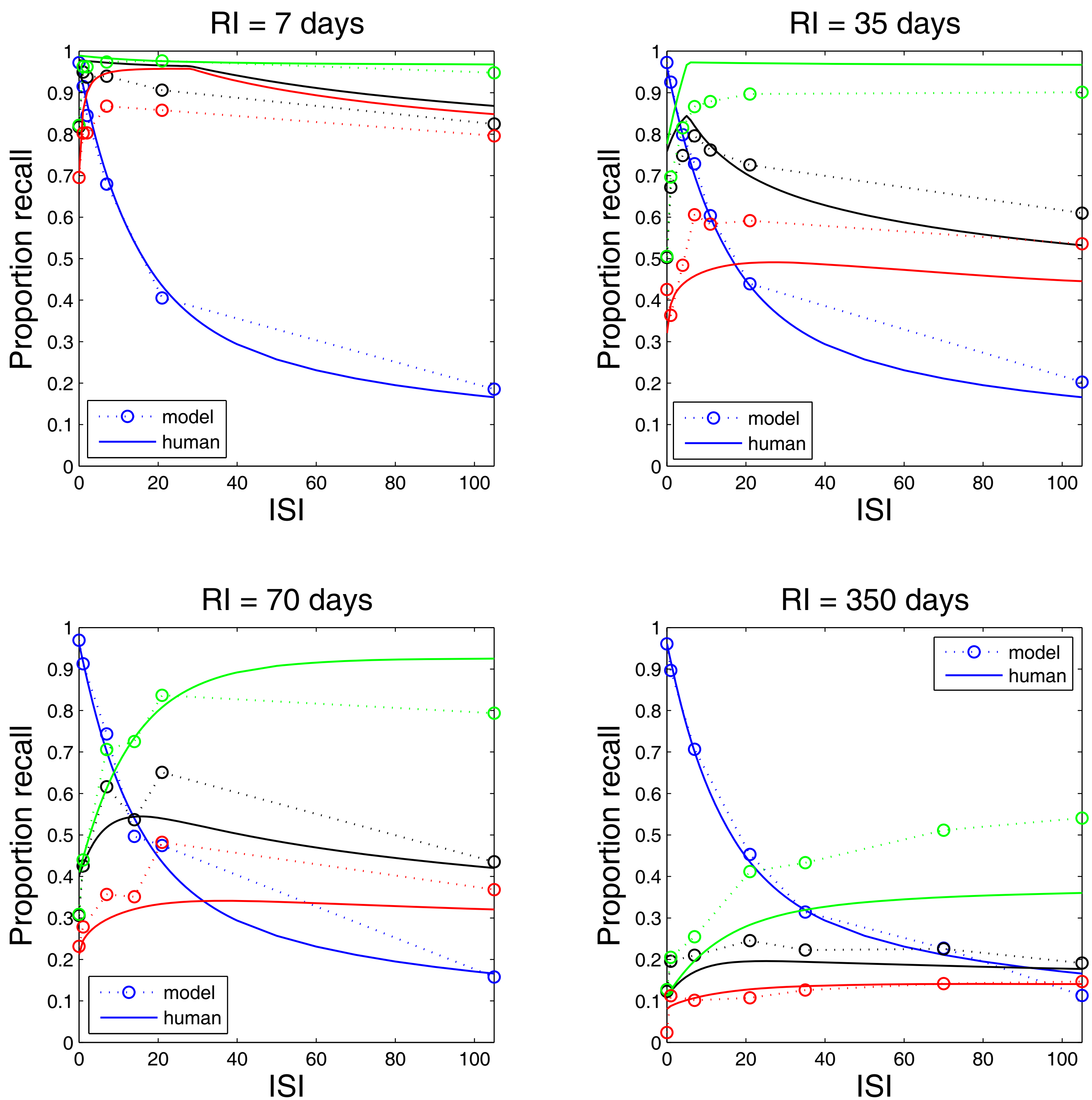
Simulation of Cepeda, Vul, Rohrer, Wixted, & Pashler (in press)

P(recall at study 2)

P(recall at test)



Simulation of Cepeda, Vul, Rohrer, Wixted, & Pashler (in press)



P(recall at study 2)

P(recall at test)

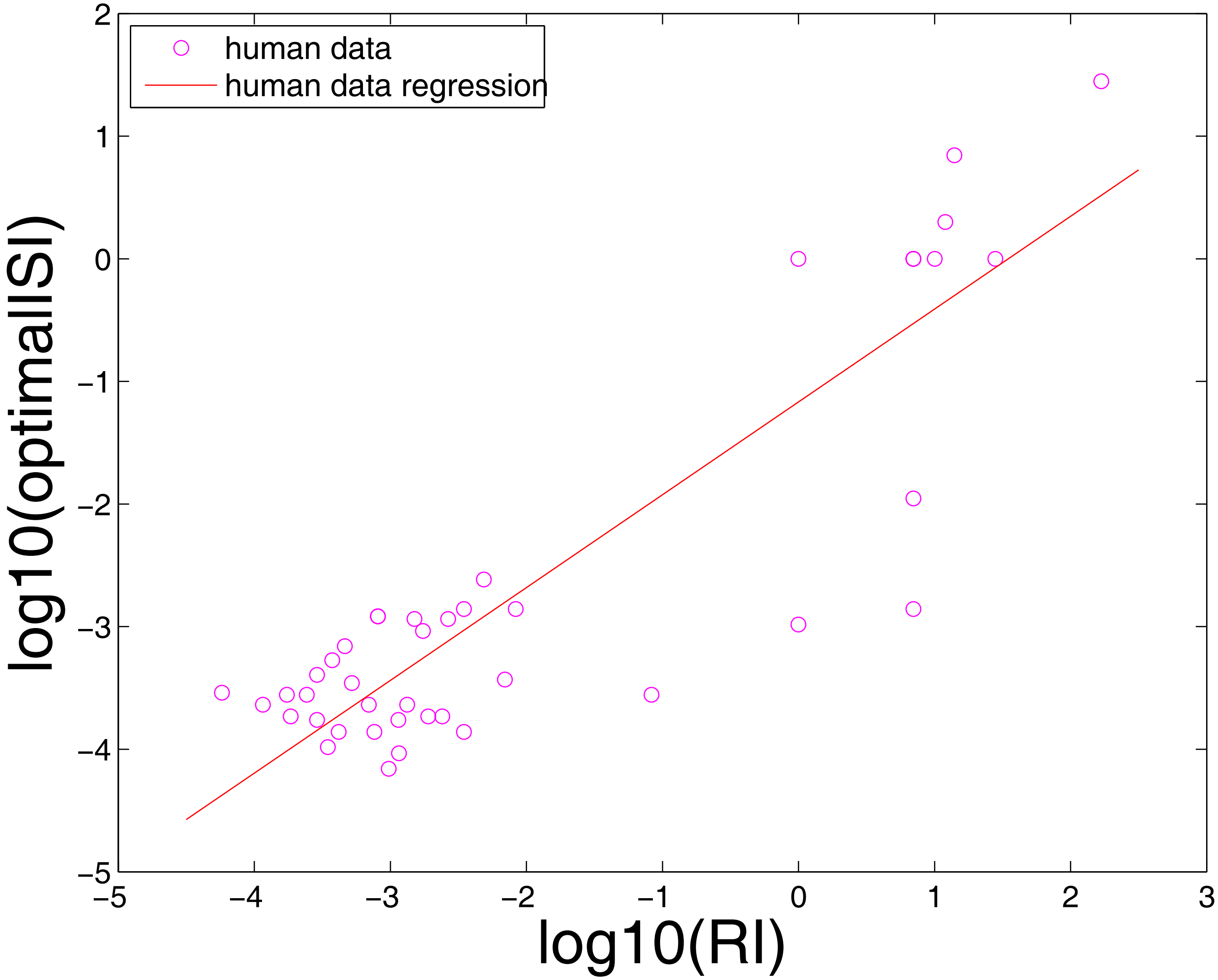
P(recall at test | recall at study 2)

P(recall at test | no recall at study 2)

The Relationship Between RI and Optimal ISI

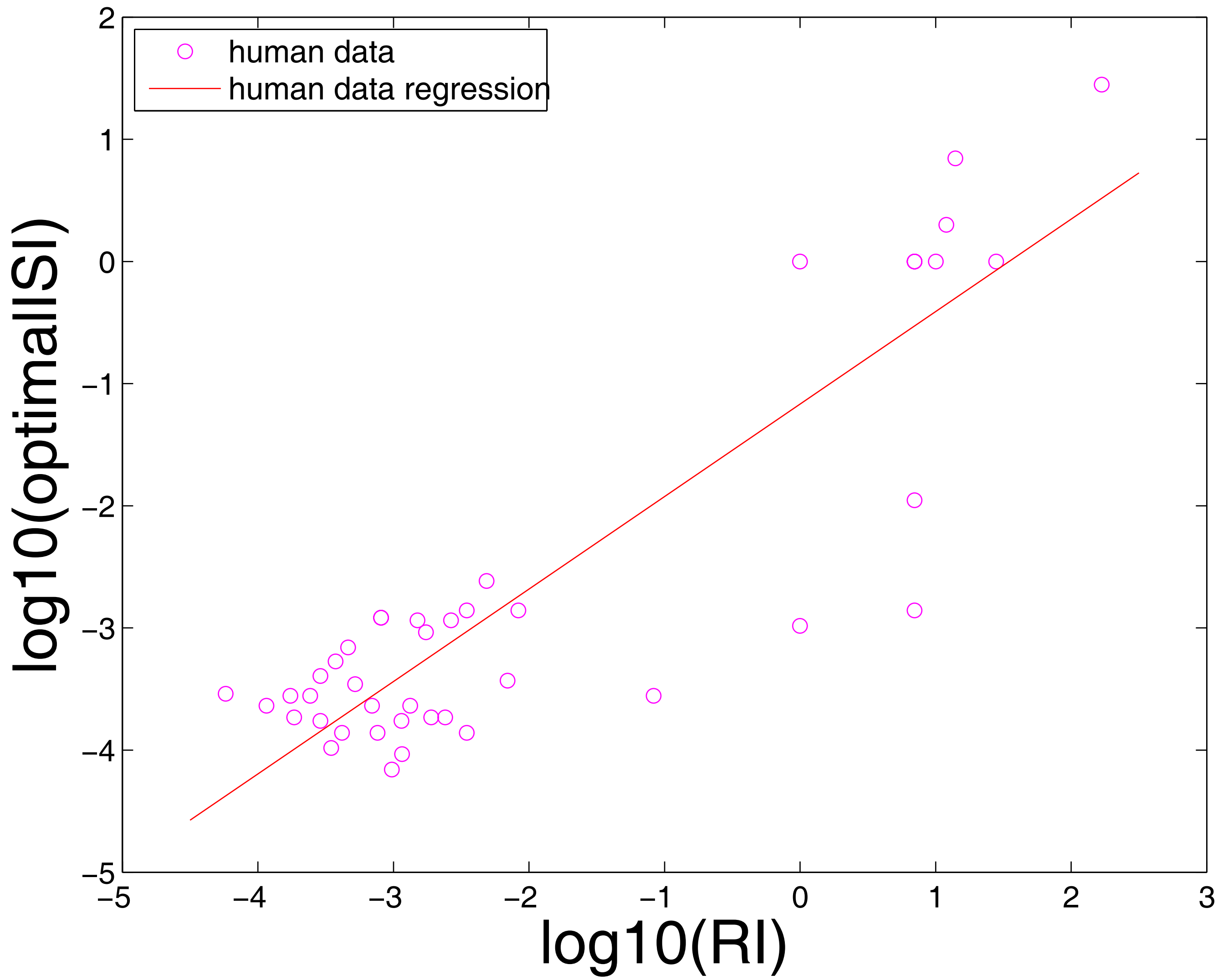
The Relationship Between RI and Optimal ISI

Cepeda et al. metaanalysis



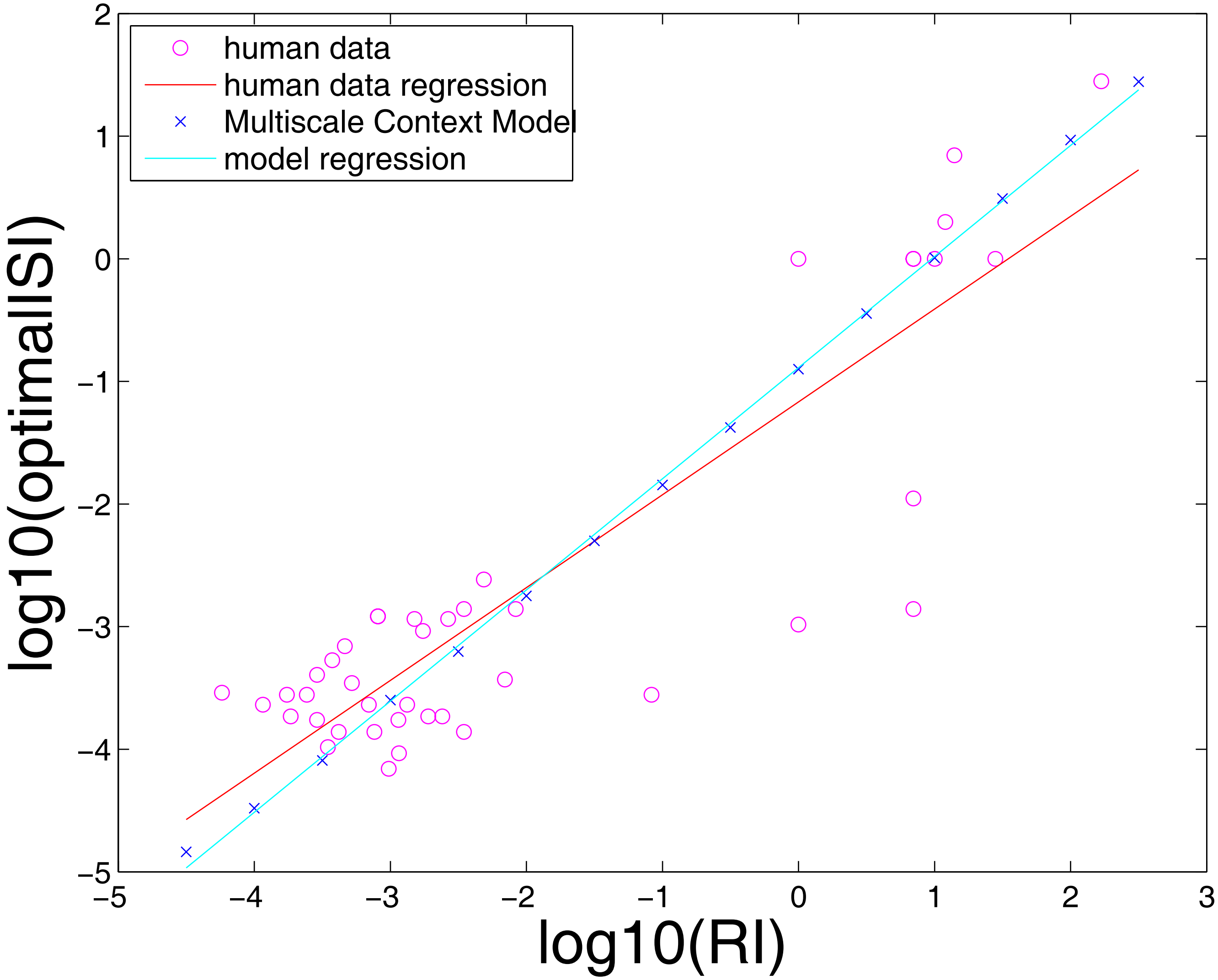
Simulation of Multiscale Context Model

Random parameter settings of model over a large range



Simulation of Multiscale Context Model

Random parameter settings of model over a large range



Why Are We Proposing Yet Another Model?

Previous models

- **have many free parameters, and**
- **obtain only post hoc fits to data.**

Our goal is to develop a truly predictive model.

Few free parameters

Parameters are fully constrained by the forgetting function

Given forgetting function, optimal distribution of practice can be predicted.

Current Research

Current Research

- **Exploring DP effects with three study sessions**

    Human study (Kang, Pashler, and Lindsey)

    Comparing predictions of two different models (Lindsey and Mozer)

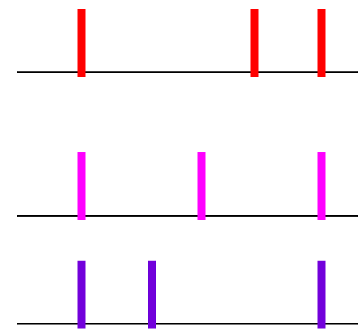
        * MCM: equal spacing is generally best, but dependent on specific materials

        * Pavlik & Anderson: decreasing spacing best

decreasing

equal

increasing



Current Research

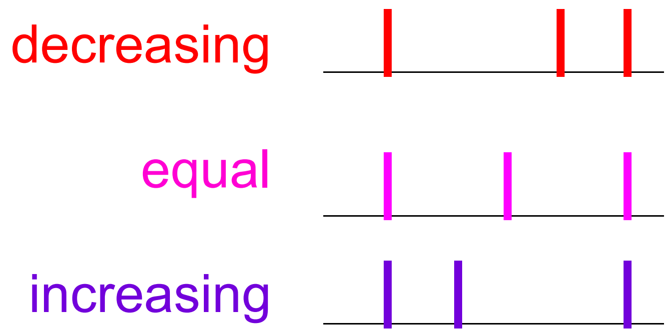
- **Exploring DP effects with three study sessions**

    Human study (Kang, Pashler, and Lindsey)

    Comparing predictions of two different models (Lindsey and Mozer)

    * MCM: equal spacing is generally best, but dependent on specific materials

    * Pavlik & Anderson: decreasing spacing best



- **Exploring DP effects with more complex materials**

    legal, scientific reasoning (Pashler, Coburn, and Carpenter)

Current Research

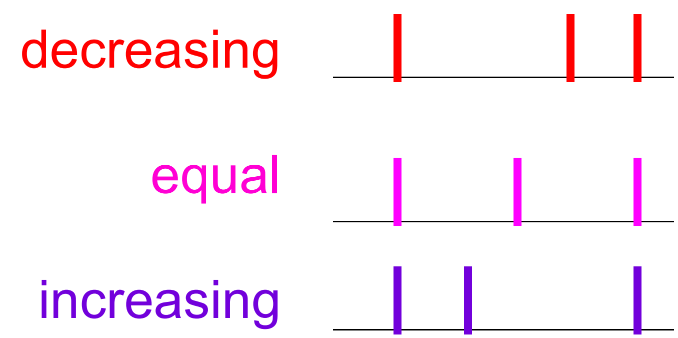
- **Exploring DP effects with three study sessions**

  Human study (Kang, Pashler, and Lindsey)

  Comparing predictions of two different models (Lindsey and Mozer)

  * MCM: equal spacing is generally best, but dependent on specific materials

  * Pavlik & Anderson: decreasing spacing best



- **Exploring DP effects with more complex materials**

  legal, scientific reasoning (Pashler, Coburn, and Carpenter)

- **Developing Facebook app for learning important facts: survival tactics**

  Natural language interface to allow unrestricted answers (Homaei)

  Eventually will use MCM to dynamically optimize study session spacing to promote long-term retention (Lindsey and Mozer)

The End