Computational Modeling of Human Cognition

Michael C. Mozer

Department of Computer Science and
Institute of Cognitive Science
University of Colorado, Boulder

Lecture for Issues and Methods
February 23, 2006

Computational Modeling

Computer simulation of neural and/or cognitive processes that underlie performance on a task

Goals

- Understand mechanisms of information processing in the brain
- Explain behavioral, neuropsychological, and neuroscientific data
- Suggest techniques for remediation of cognitive deficits due to brain injury and developmental disorders
- Suggest techniques for facilitating learning in normal cognition
- Construct computer architectures to mimic human-like intelligence

Why Build Models?

- Forces you to be explicit about hypotheses and assumptions
- Provides a framework for integrating knowledge from various fields
- Allows you to observe complex interactions among hypotheses
- Provides ultimate in controlled experiment
- Leads to empirical predictions
- A mechanistic framework will ultimately be required to provide a unified theory of cortex.

Levels of Modeling

Single cell

ion flow, membrane depolarization, neurotransmitter release, action potentials, neuromodulatory interactions

Network

neurophysiology and neuroanatomy of cortical regions, cell firing patterns, inhibitory interactions, mechanisms of learning

Functional

operation and interaction of cortical areas, transformation of representations

Computational

input-output behavior, mathematical characterization of computation

Levels of Modeling

Single cell

computational neuroscience

Network

connectionist (neural network)

Functional

production system (symbolic)

Computational

probabilistic (Bayesian)

Overview

Computational modeling

Connectionist models

Bayesian models

Comparison of connectionist and Bayesian models

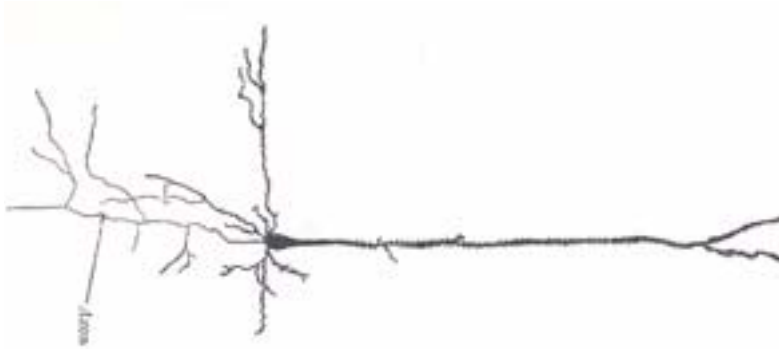
the neighborhood™ Jerry Van Amerongen

how the brain works

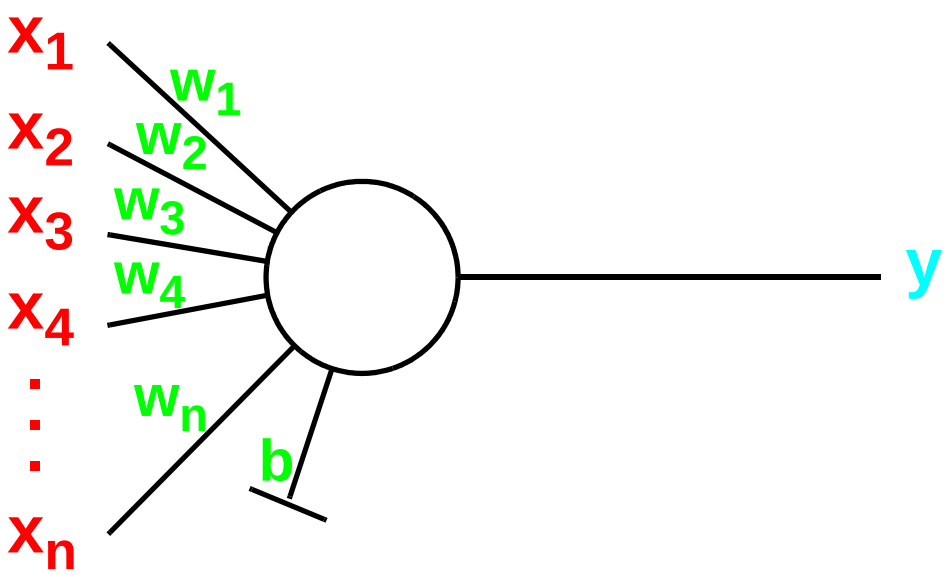
Key Features of Cortical Computation

- Neurons are slow (10^{-3} – 10^{-2} propagation time)
- Large number of neurons (10^{10} – 10^{11})
- No central controller (CPU)
- Neurons receive input from a large number of other neurons (10^4 fan-in and fan-out of cortical pyramidal cells)
- Communication via excitation and inhibition
- Statistical decision making (neurons that single-handedly turn on/off other neurons are rare)
- Learning involves modifying coupling strengths (the tendency of one cell to excite/inhibit another)
- Neural hardware is dedicated to particular tasks (vs. conventional computer memory)
- Information is conveyed by mean firing rate of neuron, a.k.a. *activation*

Modeling Individual Neurons



flow of activation



input activities

weights and bias

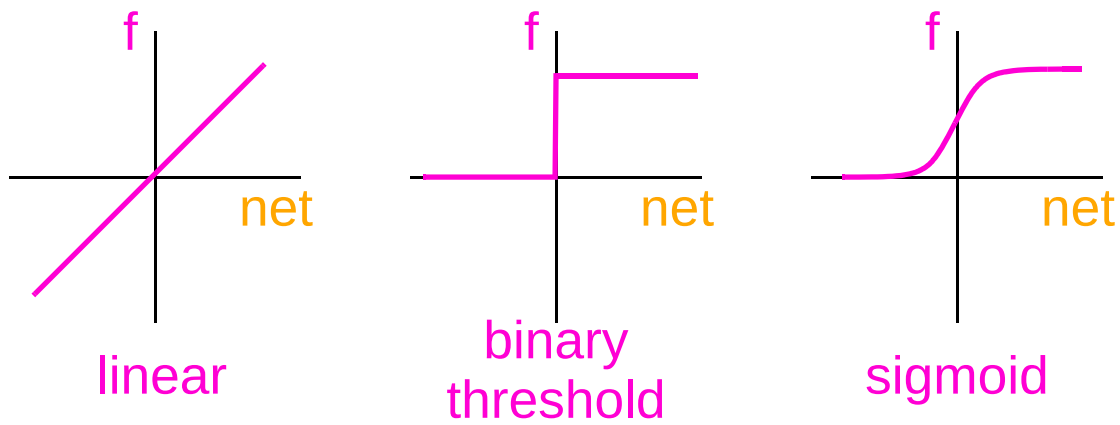
output activity

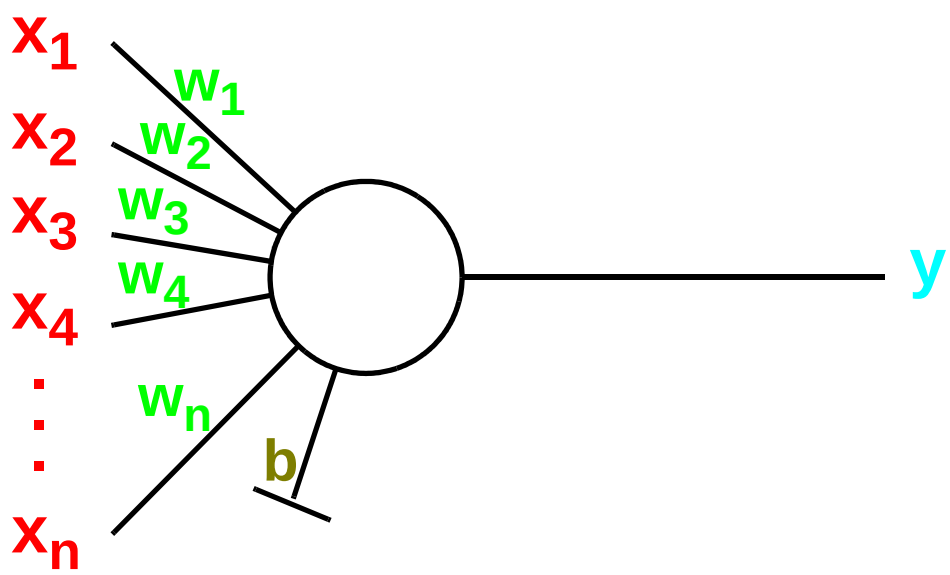
Modeling Individual Neurons

Activation function

$$net = \sum_i w_i x_i + b$$

$$y = f(net)$$



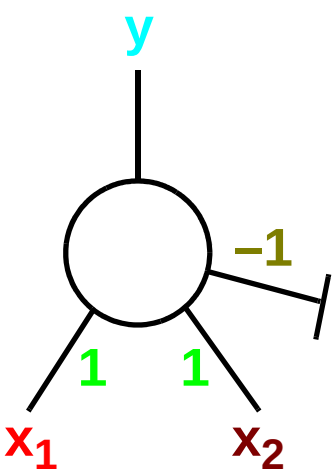


input activities **weights and bias** **output activity**

Computation With a Binary Threshold Unit

"And" gate

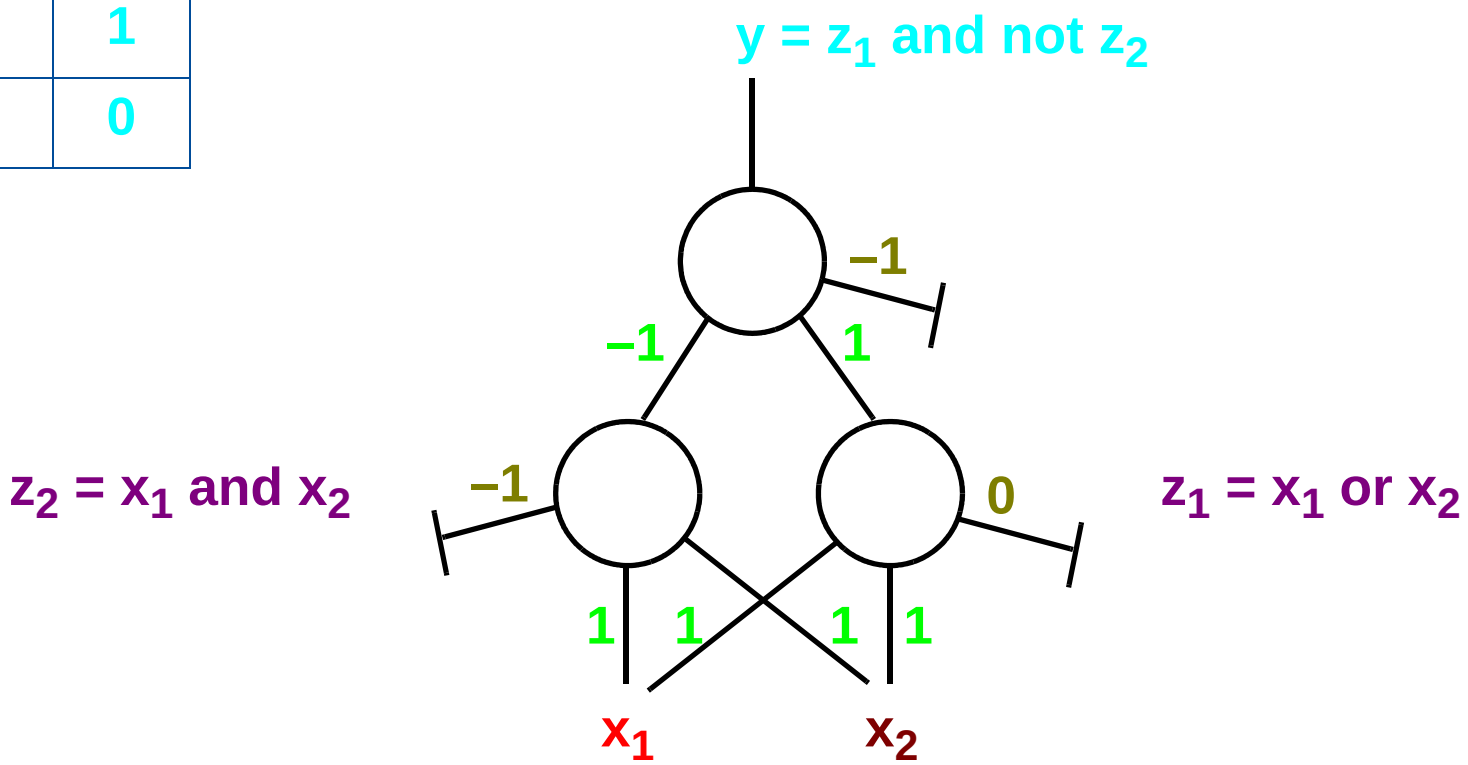
x1	x2	y
0	0	0
0	1	0
1	0	0
1	1	1



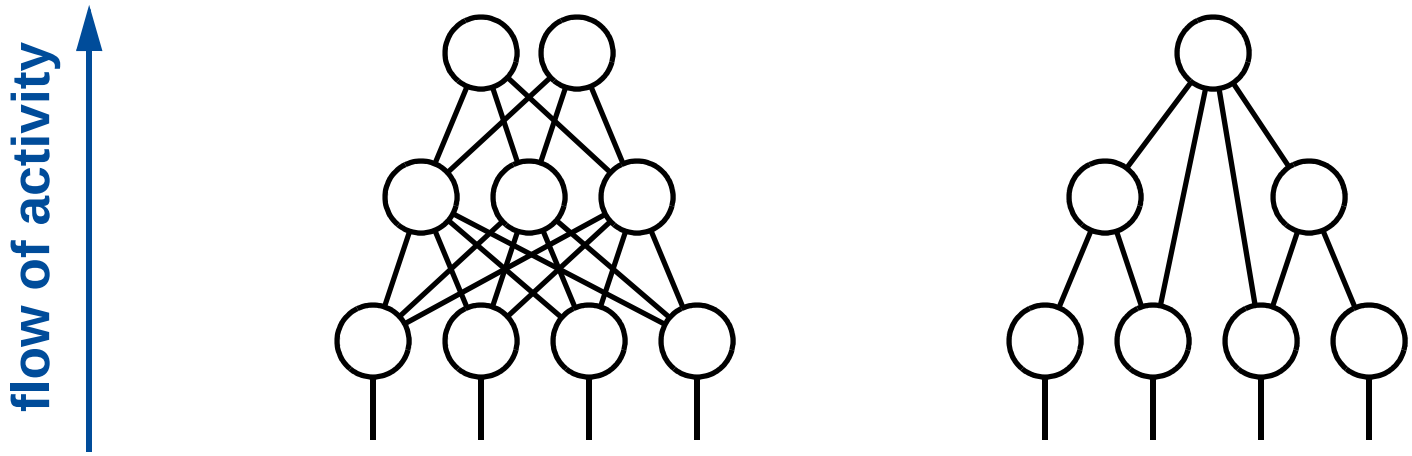
Computation With a Binary Threshold Unit

"Exclusive or" gate

x1	x2	y
0	0	0
0	1	1
1	0	1
1	1	0



Feedforward Architectures



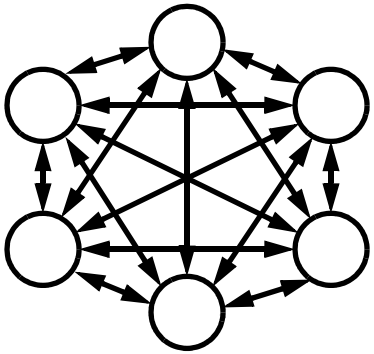
Activation flows in one direction; no closed loops

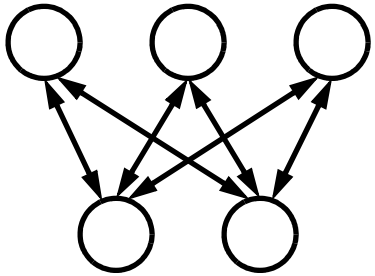
Performs association from input pattern to output pattern

big, hairy, stinky è run away
small, round, orange è eat
big, round, soft è eat
small, orange, hairy è run away
stinky, yellow è eat

Learning: adjust connections to achieve input-output mapping

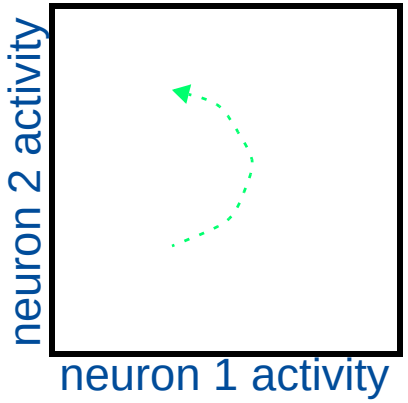
Recurrent Architectures



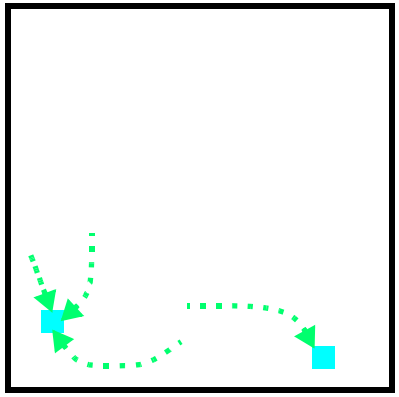


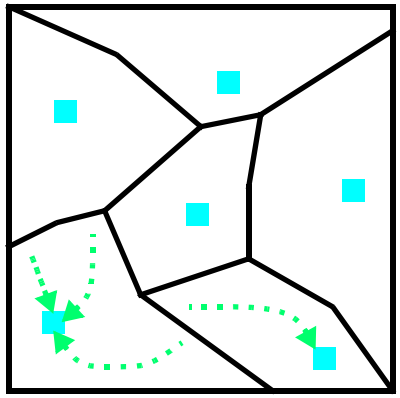
Achieves *best interpretation* of partial or noisy patterns, e.g.,
MAR – – M – LLOW

State space dynamics



Attractor dynamics





Learning: establishes new attractors and shifts attractor basin boundaries

Supervised Learning in Neural Networks

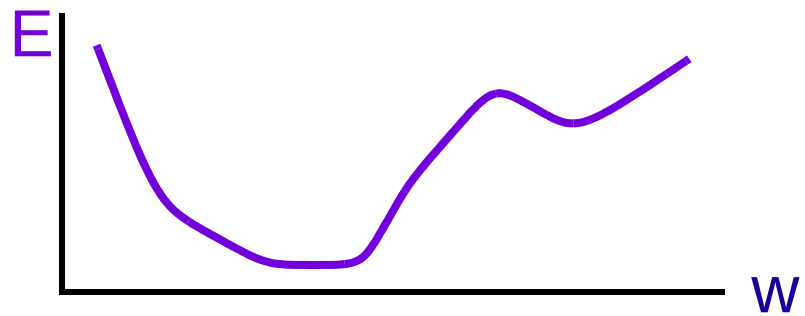
1. Assume a set of training examples, $\{x^i, t^i\}$

e.g., MAR – – M – LLOW è MARSHMALLOW

e.g., big, hairy, stinky è run away

2. Define a measure of network performance, e.g.,

$$E = \sum_i \|d^i - \mathbf{y}^i\|^2$$



3. Make small incremental changes to weights to decrease error (*gradient descent*), i.e.,

$$\Delta w_{ji} \sim -\partial E / \partial w_{ji}$$

For multilayered sigmoidal neural networks, gradient descent update has a simple *local* form (depends on activity of neuron i and error associated with neuron j)

Using Testing to Enhance Learning: A Comparison of Two Hypotheses

Michael C. Mozer
Michael Howe

Department of Computer Science and
Institute of Cognitive Science
University of Colorado, Boulder

Harold Pashler

Department of Psychology
University of California, San Diego

Fact Learning

E.g., foreign language vocabulary

French word for dog is *chien.*

E.g., history trivia

The University of Colorado was founded in 1876.

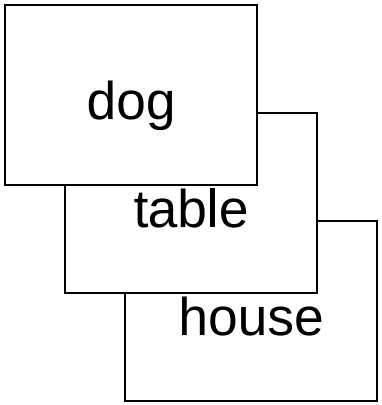
Facts can be framed as *cue – response* pairs.

e.g., dog – chien

e.g., Founding of University of Colorado – 1876

a.k.a. paired associate learning

Self Testing



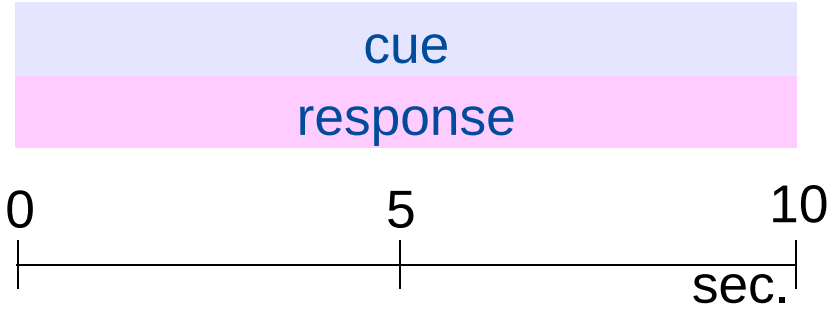
Does Self Testing Foster Learning?

Long history of empirical demonstrations, but many methodological difficulties.

Carrier and Pashler (1992)

Pure study (PS)

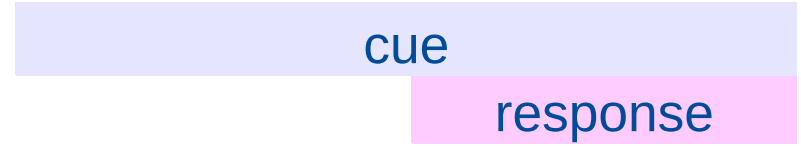
cue-response pair presented together for 10 sec.



Self testing (ST)

cue presented alone for 5 sec., during which response is to be retrieved

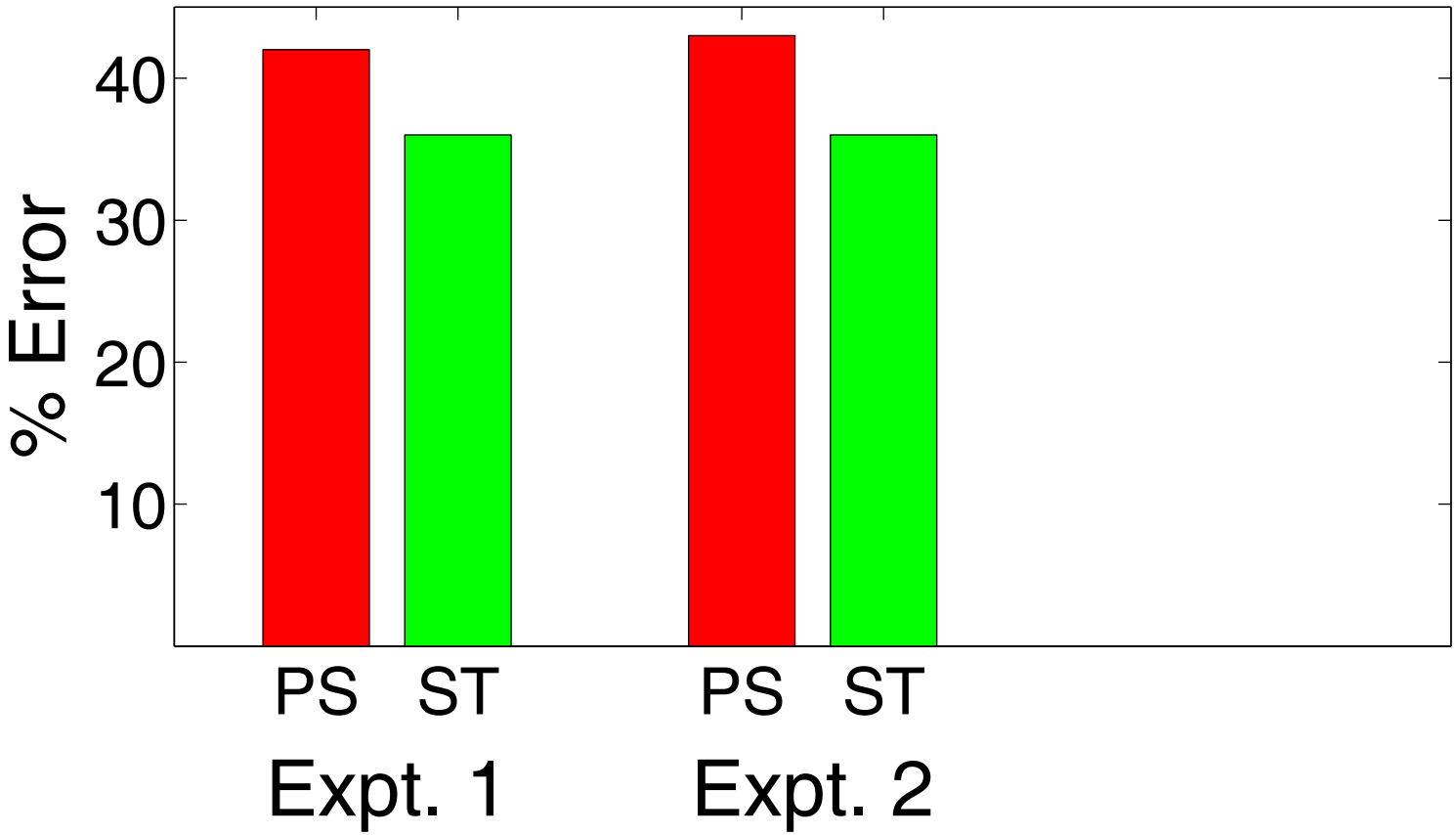
cue and response together for 5 sec., during which response is to be studied



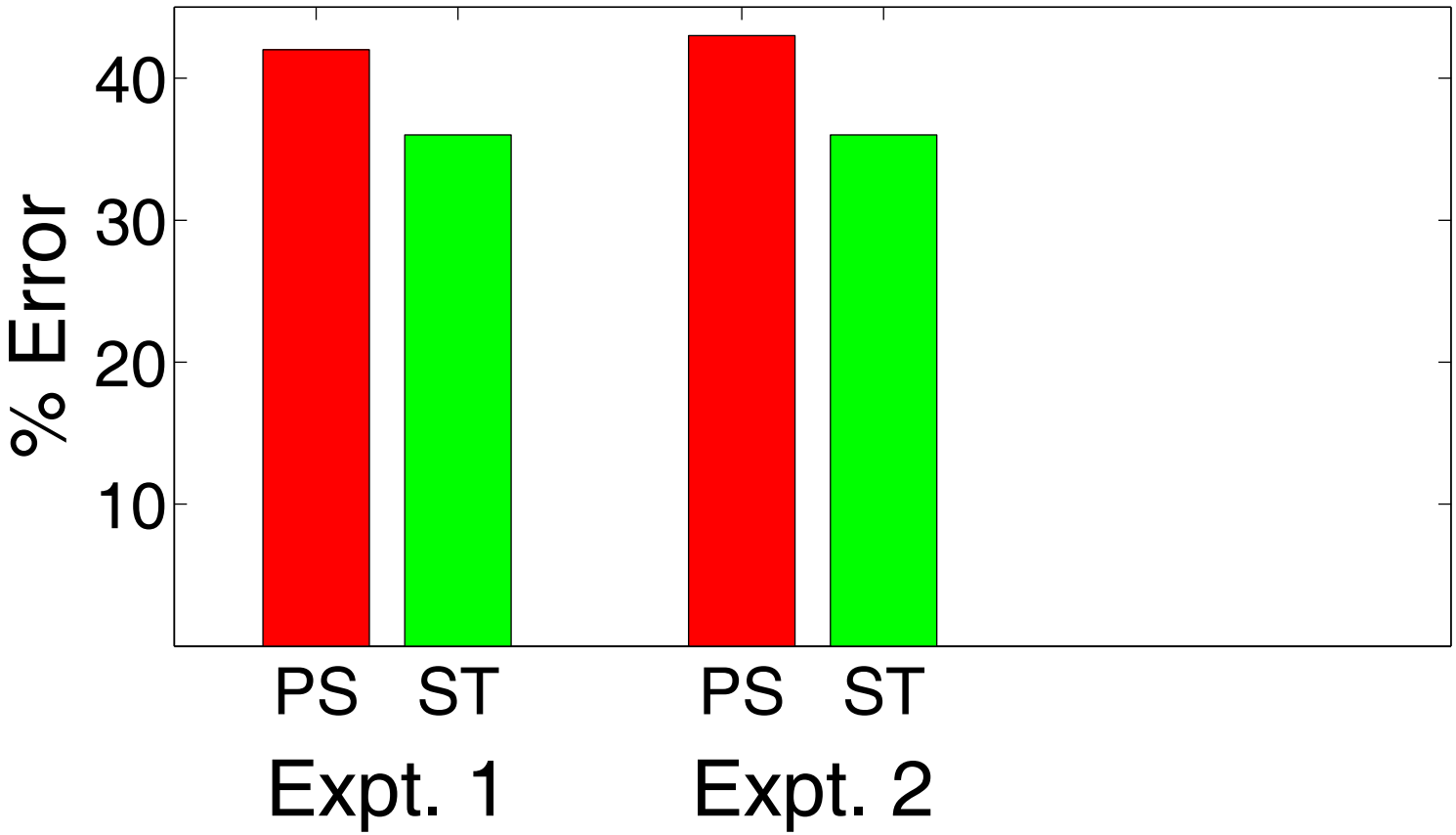
Design

- 20 items designated for PS, and 20 for ST
- 3 training blocks; all items studied in block 1
- final test phase; evaluation via cued recall
- Experiment 1: consonant trigrams – two-digit numbers
- Experiment 2: English – Siberian Eskimo Yupik word translation

Carrier and Pashler (1992)



Carrier and Pashler (1992)

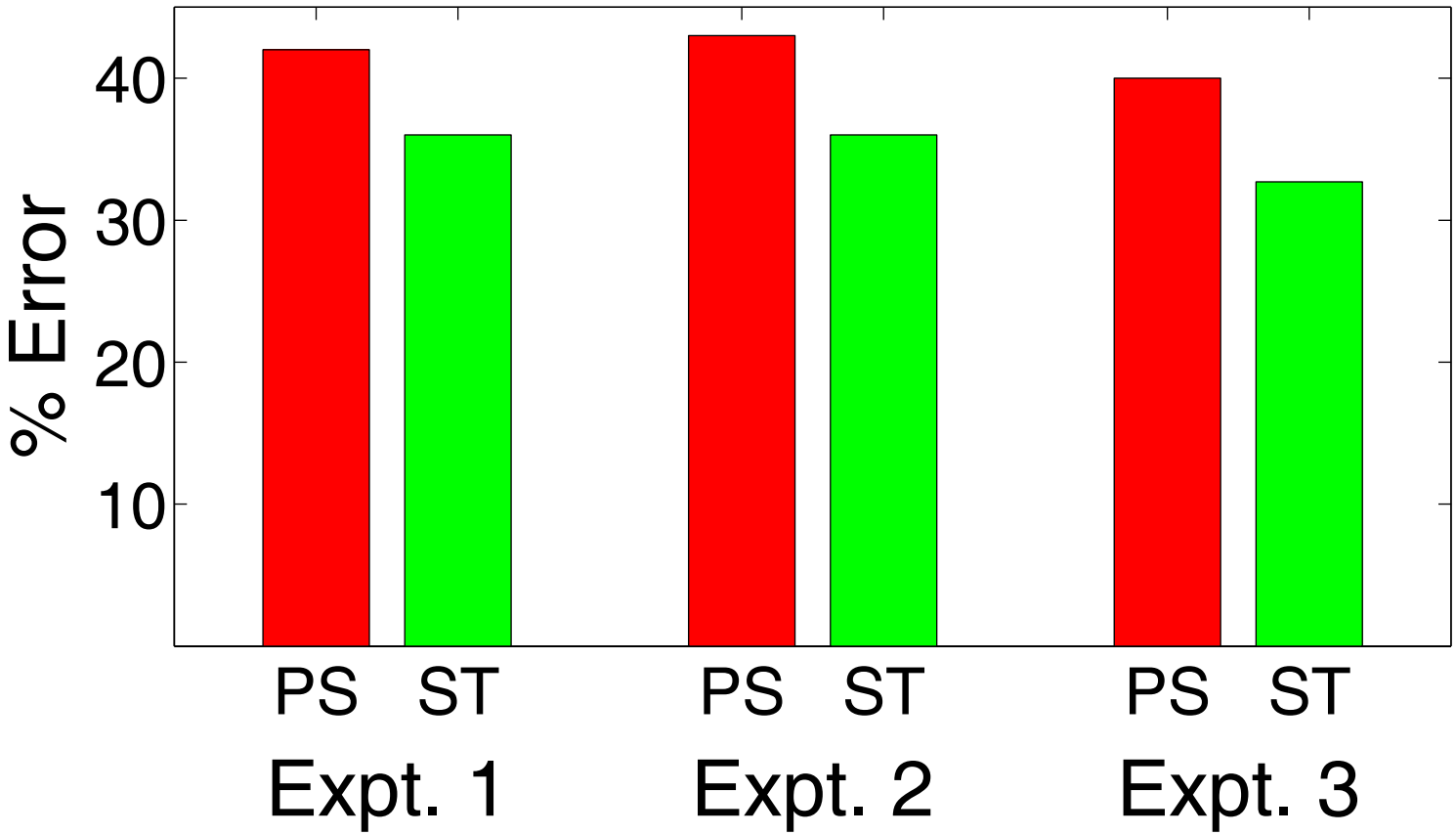


Possible explanation

Subjects used first self-test trial to assess item difficulty, and increased encoding effort on second self-test trial.

Expt. 3 same as Expt 2. except all items studied in first *two* training blocks

Carrier and Pashler (1992)



Some Explanations of Self-Testing Benefit

Landauer and Bjork (1978)

Retrieval attempts provide general boost to performance at a future time.

Incorrectly predicts that ST and PS items should benefit equally

Mandler (1979)

Cued recall strengthens structural, integrative information about an item.

Because cue and response are simultaneously activated for both ST and PS items, not clear why they wouldn't both benefit.

Bjork (1975)

Act of retrieval strengthens existing retrieval routes to the response.

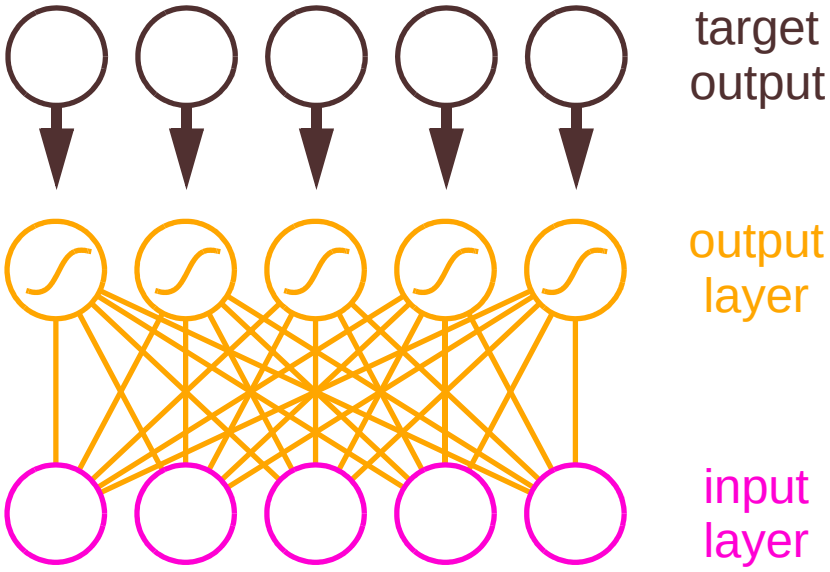
Consistent with data, but seems to require novel learning mechanisms

Basic Approach

Use a common, relatively noncontroversial architecture

Feedforward neural network

- Input layer connected to output layer
- Standard sigmoidal activation function
- Error correction learning
  - Widrow-Hoff (a.k.a. LMS)
    network generates *actual* output
    teacher provides *target* output
    training depends on actual – target



Basic Approach

Use a common, relatively noncontroversial architecture

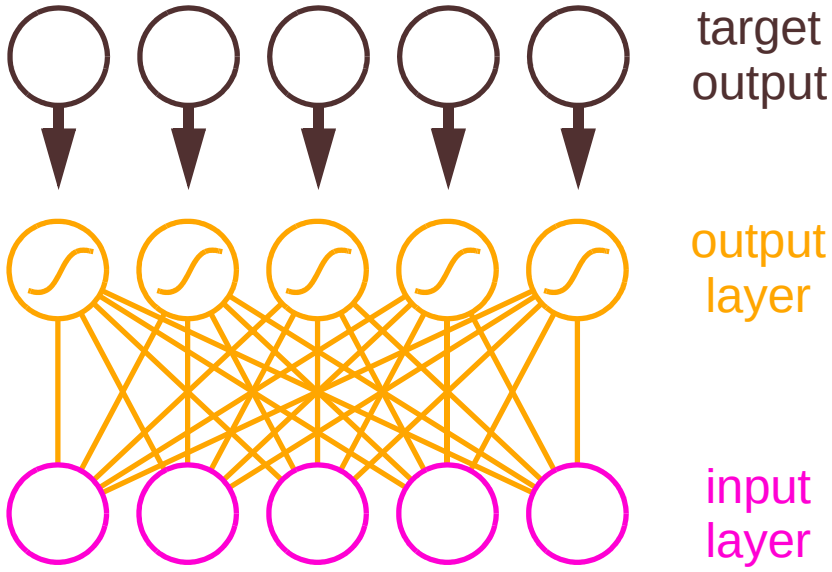
Feedforward neural network

Input layer connected to output layer

Standard sigmoidal activation function

Error correction learning

Widrow-Hoff (a.k.a. LMS)
network generates *actual* output
teacher provides *target* output
training depends on actual – target



Training of neural net often viewed as abstract procedure for loading knowledge into net.

Here, we make a stronger claim.

One training trial in neural net ~ one experimental trial

Hypothesis 1: Self-Generated Training Targets

Guthrie (1952)

One learns what one does.

ST item

Self test ⇨candidate response ⇨ target for error-correction learning

Study ⇨correct response ⇨ target for error-correction learning

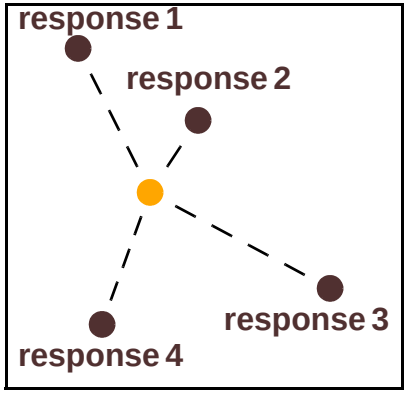
Both candidate and correct response are trained.

PS item

Only correct response is trained.

Choosing candidate response

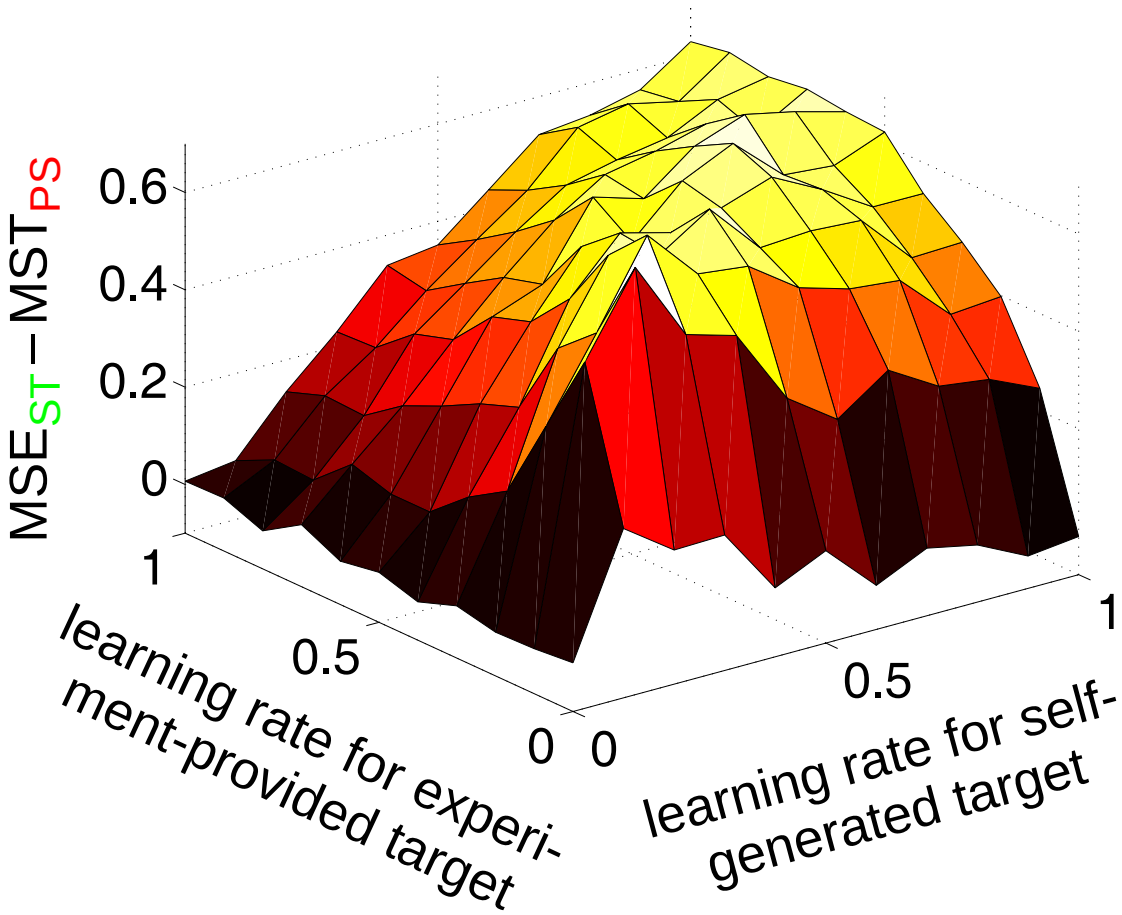
Probabilistic selection with Luce Choice Rule (a.k.a. Boltzmann distribution)



Hypothesis 1: Simulation Result

No parameter settings found that yield an enhancement of learning by testing.

In final test, mean-squared error (MSE) significantly higher for ST than PS items.



Hypothesis 2: Interruption of Cue Processing

Carrier and Pashler (1992)

Presentation of the response simultaneously with cue interrupts processing of the cue, making learning less efficient.

Hypothesis 2: Interruption of Cue Processing

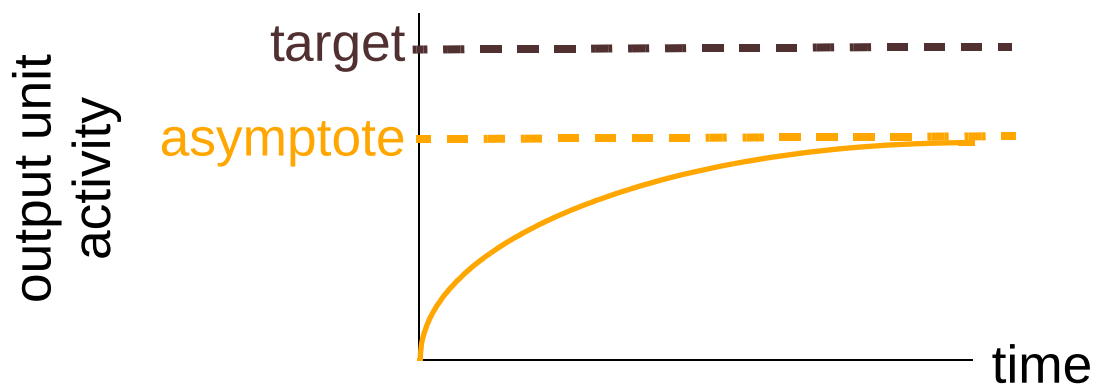
Carrier and Pashler (1992)

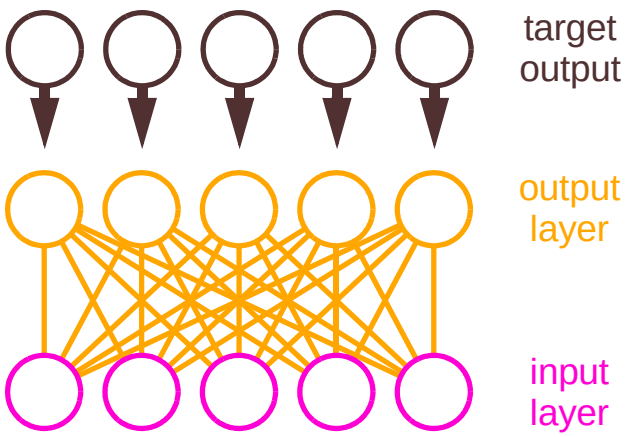
Presentation of the response simultaneously with cue interrupts processing of the cue, making learning less efficient.

Our interpretation

Units in neural net have temporal dynamics.

Leaky integrator model: $y_i(t) = \tau y_i(t-1) + (1-\tau) f(net_i(t))$





Hypothesis 2: Interruption of Cue Processing

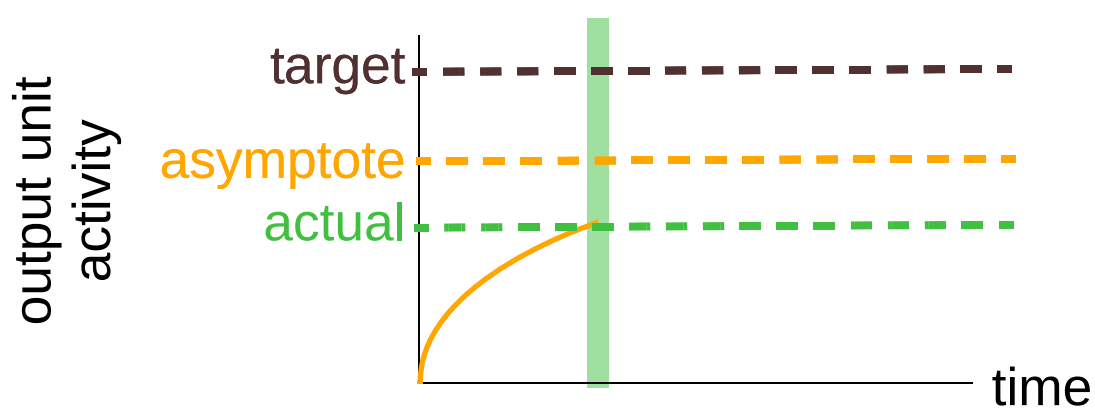
Carrier and Pashler (1992)

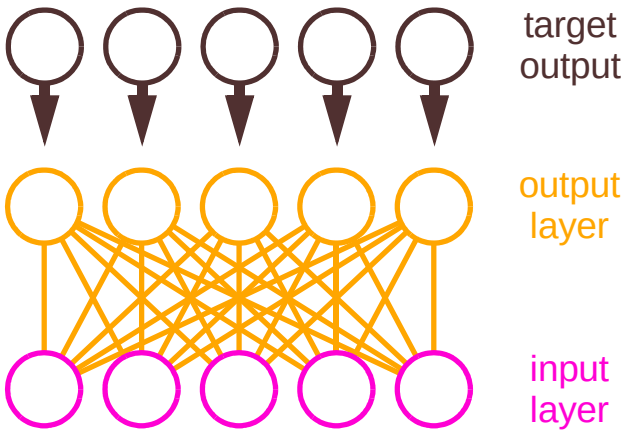
Presentation of the response simultaneously with cue interrupts processing of the cue, making learning less efficient.

Our interpretation

Units in neural net have temporal dynamics.

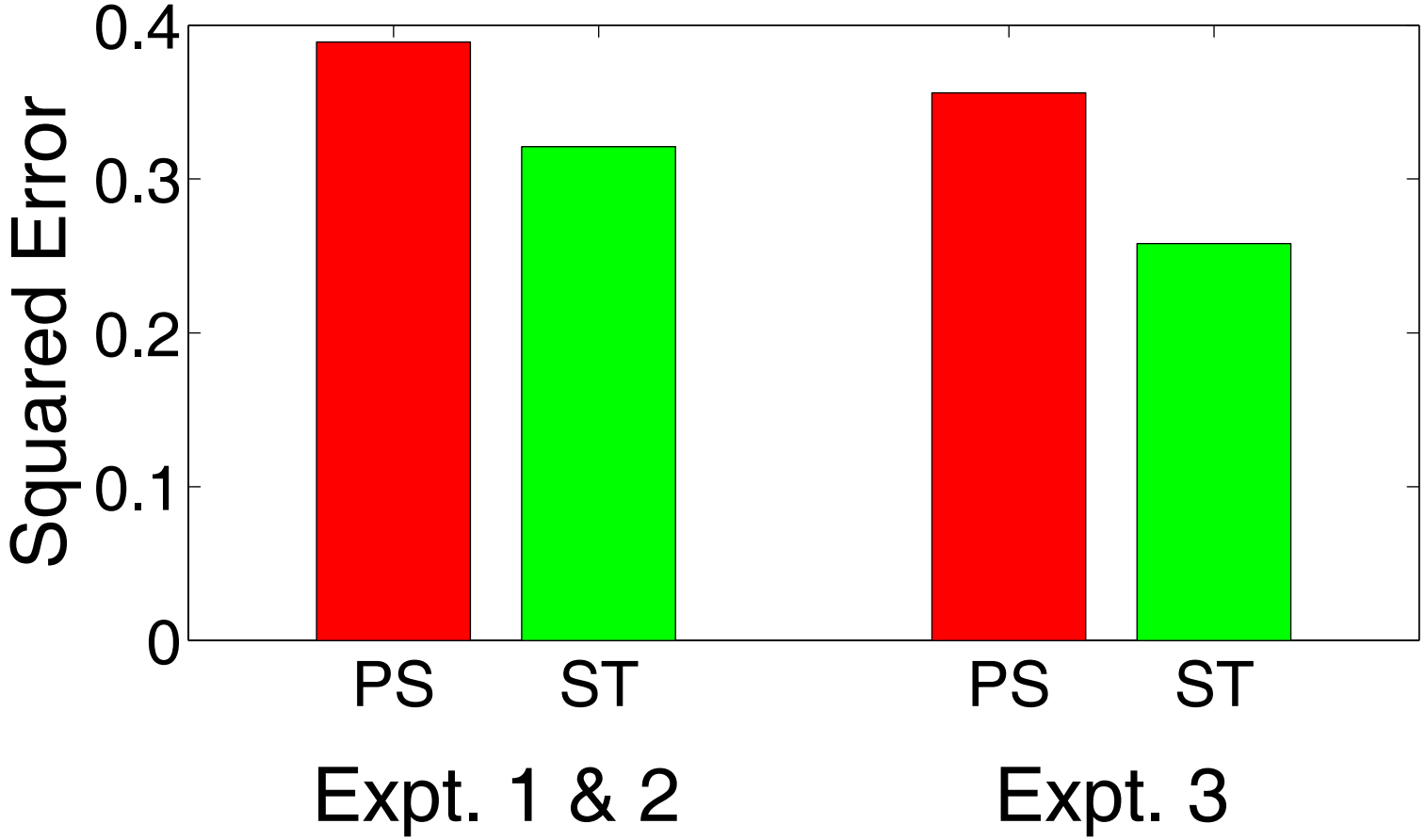
Leaky integrator model: $y_i(t) = \tau y_i(t-1) + (1-\tau) f(net_i(t))$





Presentation of correct response ⇨ premature termination of processing ⇨ incorrect output ⇨ incorrect error signal

Hypothesis 2: Simulation Result



Summary

Goal

Explain the enhancement of learning through testing

Approach

In the context of a simple neural network model, we explored two alternative hypotheses.

(1) Testing obtains a candidate response whose association to the cue is strengthened, making the association less vulnerable to decay or interference.

(2) Error-correction learning requires comparison of the correct response to a candidate response. Testing forces a candidate response to be generated, whereas pure study does not.

Result

Simulations supported hypothesis 2, not hypothesis 1.

Overview

Computational modeling

Connectionist models

Bayesian models

Comparison of connectionist and Bayesian models

Learning Formalism

Supervised learning problem with a set of training examples

x_i: training example i input

y_i: training example i output

Data

$$D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots\}$$

Model space

$$M = \{m_1, m_2, m_3, \ldots\}$$

models can vary in complexity (e.g., neural nets with different numbers of hidden units)

models can vary in specific parameters (e.g., neural net weights)

model space is generally infinite

models are probabilistic, i.e., using model one can compute $p(y|x, m)$

generalization of deterministic models (e.g., neural nets)

Bayesian Approach to Learning

Standard approach to learning

Find the single model that best fits the data

maximum likelihood approach: $m^* = \max_m p(D|m) = \max_m \left(\prod_i p(y_i|m, x_i) \right)$

Given input x_q, prediction y_q is obtained as $p(y_q|x_q, m^*)$

Bayesian approach

Consider *all* possible models in parallel and determine their plausibility based on how well they fit the data.

Given input x_q, prediction y_q is obtained by taking a weighted average of predictions from all models, weighted by the model plausibility.

model averaging:

$$p(y_q|D, x_q) = \sum_j p(y_q|D, x_q, m_j) p(m_j|D)$$

Intuitive Example

Coin with unknown bias, ρ = probability of heads

Sequence of observations: H T T H T T T H

Maximum likelihood approach

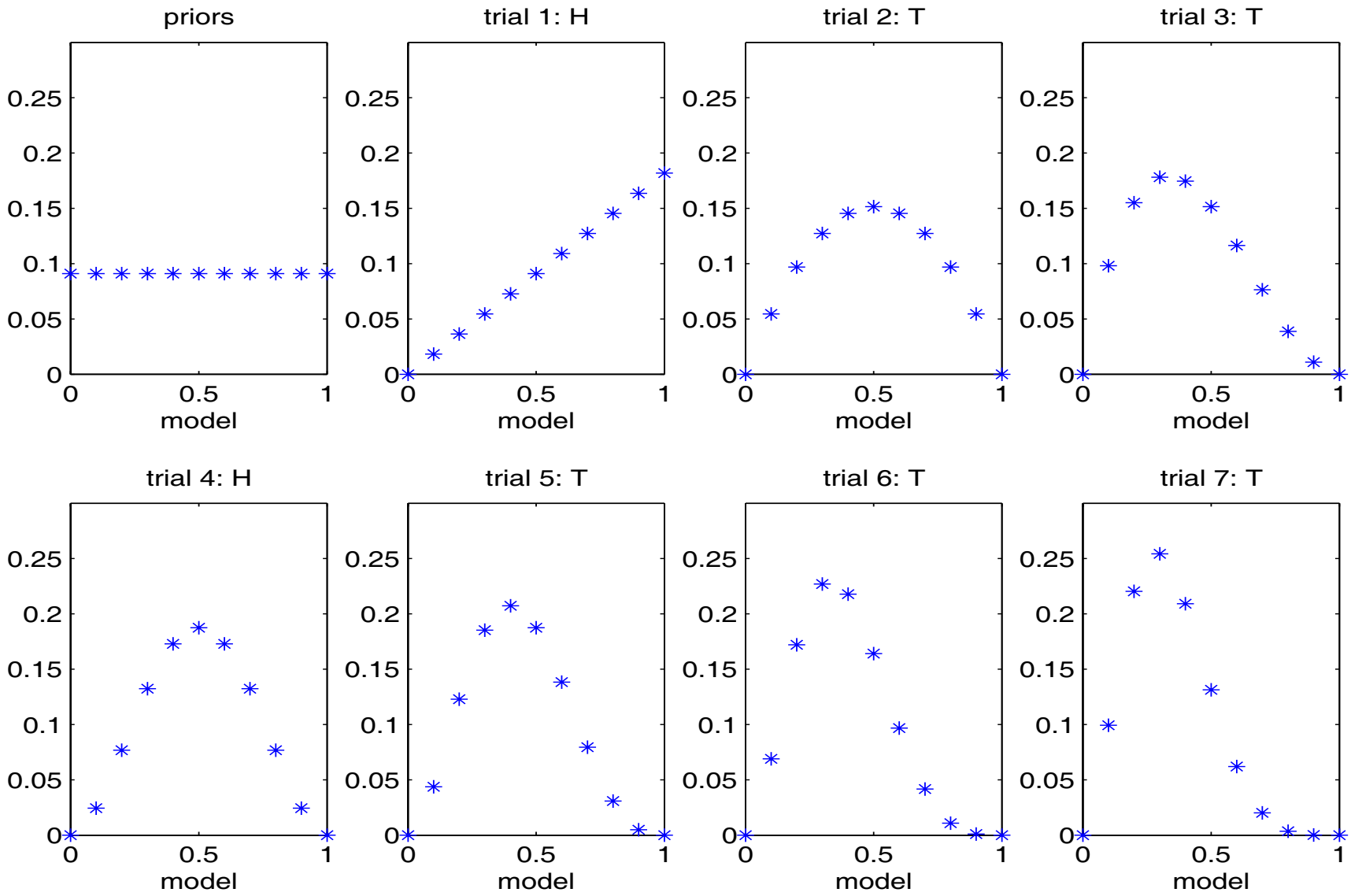
$\rho = 3 / 8$

Bayesian approach

set of models $M = \{m_\rho\}$, where probability associated with m_ρ is ρ

e.g., $M = \{m_{0.0}, m_{0.1}, m_{0.2}, \ldots, m_{1.0}\}$

Coin Flip Sequence: H T T H T T T



Bayesian Model Updates

Bayes rule

posterior likelihood prior

$$p(m|D) = \frac{p(D|m)p(m)}{p(D)}$$

Likelihood model

$$p(\text{head}|m_\rho) = \rho$$

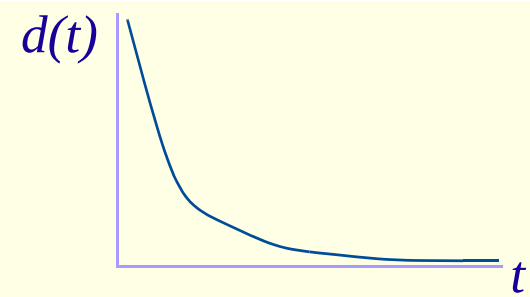
$$p(\text{tail}|m_\rho) = 1 - \rho$$

Priors

$$p(m_\rho) = \frac{1}{11}$$

μ_j : center of attractor *j*

β_j : width of attractor *j*



$$d_i(t) = 1 - \bar{e}_i(t-1)/e_i(t)$$

$$\bar{e}_i(t) = \alpha\, e_i(t) + (1-\alpha)\bar{e}_i(t-1)$$

Infinite Model Spaces

This all sounds great if you have just a few models, but what if you have infinite models?

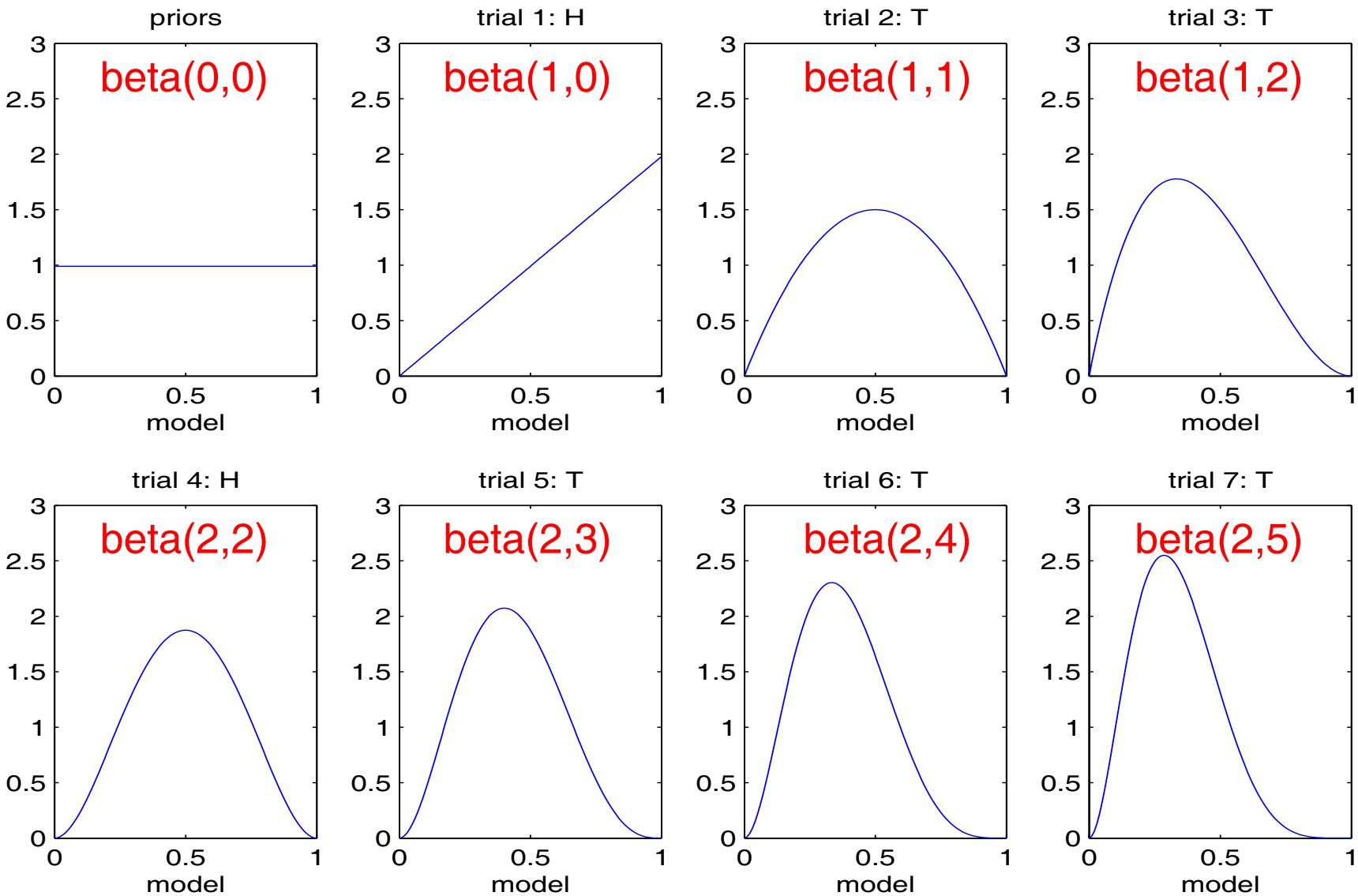
e.g., ρ continuous in [0, 1]

If you limit the form of the probability distributions, you can often do so efficiently.

e.g., beta distribution to represent priors and posterior in coin flip example

Requires only *two* parameters to update, one representing count of heads, one representing count of tails.

Coin Flip Sequence: H T T H T T T



Control of Visual Attention: A Rational Account

Michael C. Mozer
Institute of Cognitive Science and
Department of Computer Science
University of Colorado, Boulder

Michael Shettel
Department of Computer Science
University of Colorado, Boulder

Shaun Vecera
Department of Psychology
University of Iowa

Visual Search

Find wallet in office

Find Jim in UMC

Find car in parking lot

How is the visual system reconfigured to perform such a remarkable variety of tasks?

Top-Down Control of Visual Attention

Top-Down Control of Visual Attention

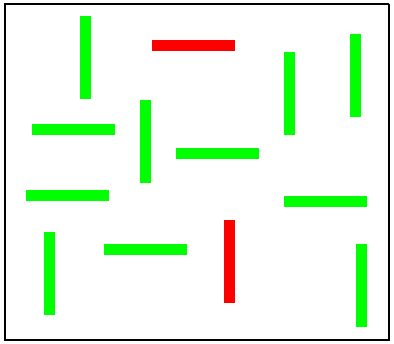
1. Deploying attention in a task- or goal-dependent manner

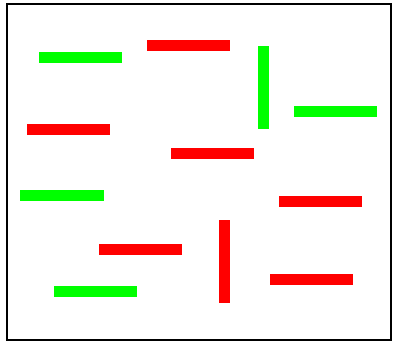
Top-Down Control of Visual Attention

1. Deploying attention in a task- or goal-dependent manner

2. Fine tuning attention to the environment

e.g., find the red vertical line



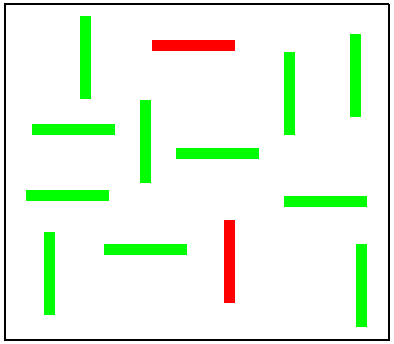


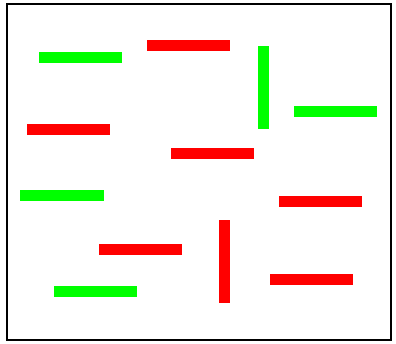
Top-Down Control of Visual Attention

1. Deploying attention in a task- or goal-dependent manner

2. Fine tuning attention to the environment

e.g., find the red vertical line



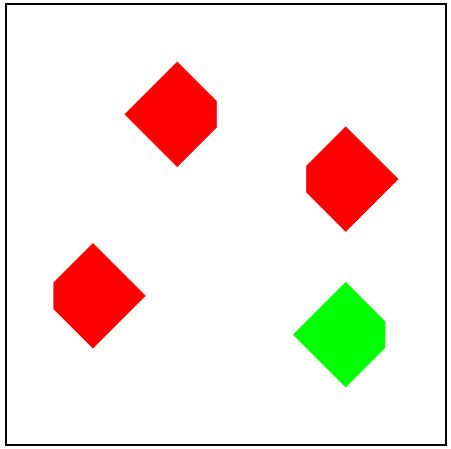


Two distinct problems?

Strategy: Study the latter to get a handle on the former

Attentional Adaptation
(Maljkovic & Nakayama, 1994)

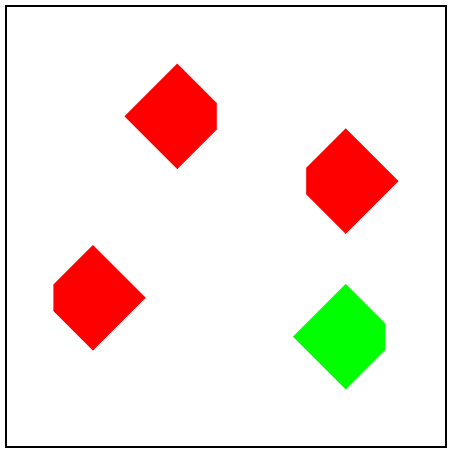
Is odd colored diamond notched on the left or right?



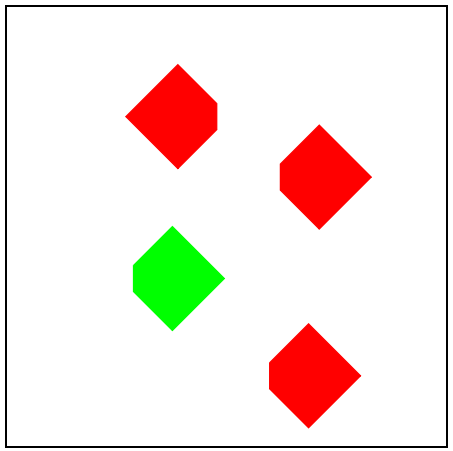
Attentional Adaptation (Maljkovic & Nakayama, 1994)

Is odd colored diamond notched on the left or right?

trial *n*



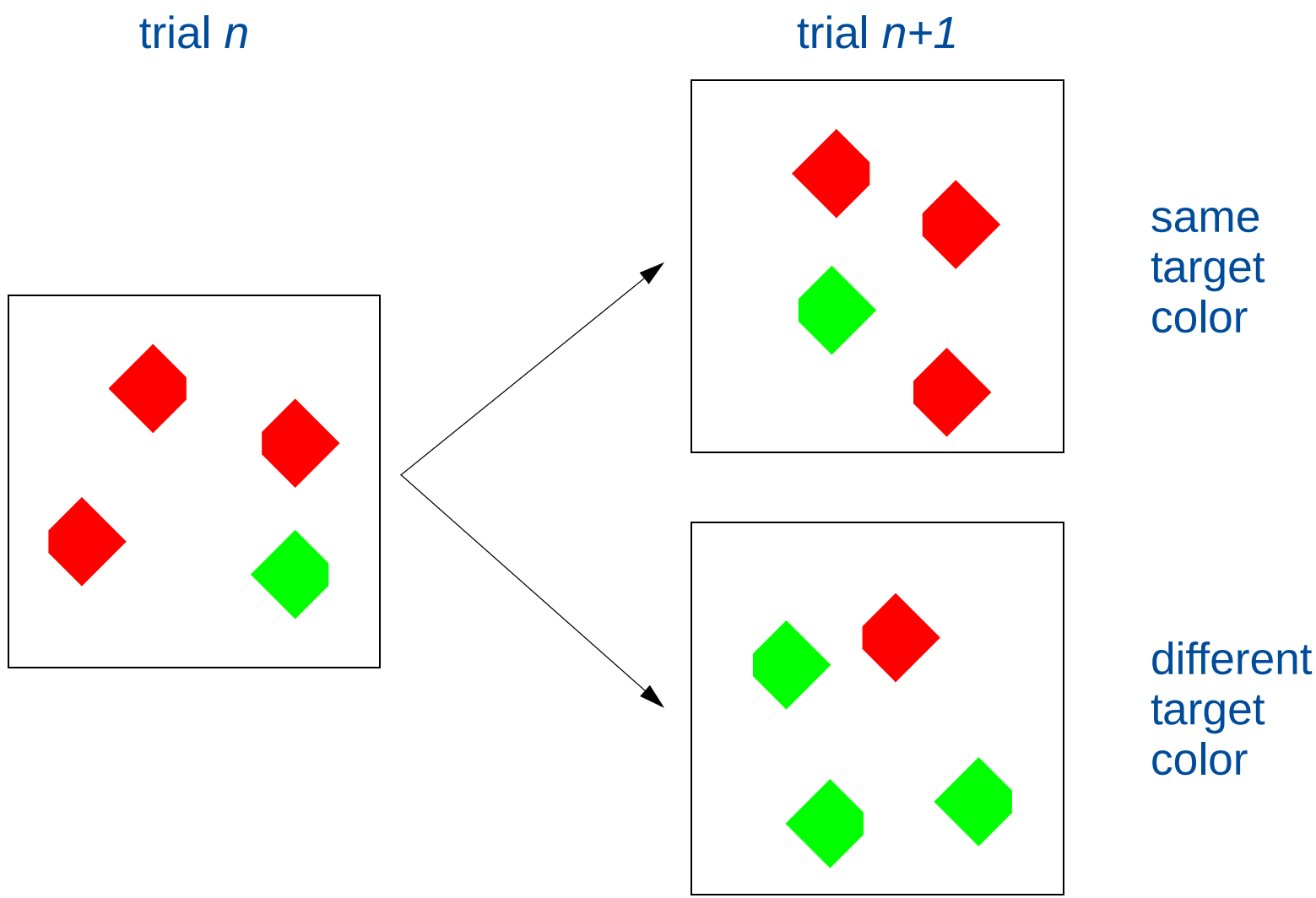
trial *n+1*



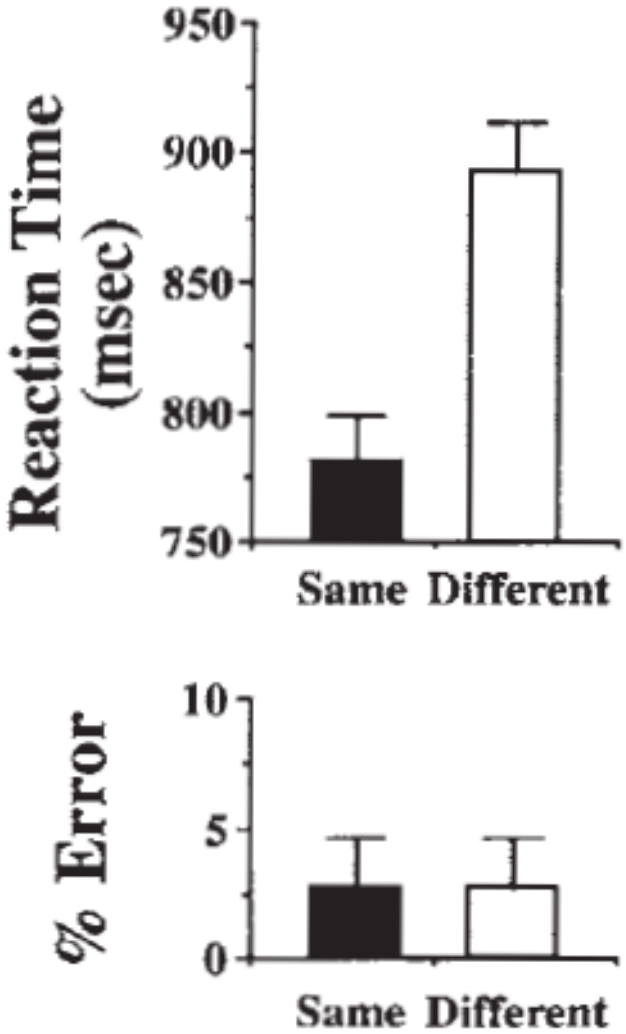
same target color

Attentional Adaptation (Maljkovic & Nakayama, 1994)

Is odd colored diamond notched on the left or right?



Attentional Adaptation
(Maljkovic & Nakayama, 1994)



Why Does Repetition Facilitate Performance?

We view attentional control as optimizing performance to the environment in which an individual is operating.

Why Does Repetition Facilitate Performance?

We view attentional control as optimizing performance to the environment in which an individual is operating.

Two-Stage Process

1. Construct predictive (probabilistic) model of the environment based on past experience.
2. Tune control parameters of attention to optimize performance under the current environmental model.

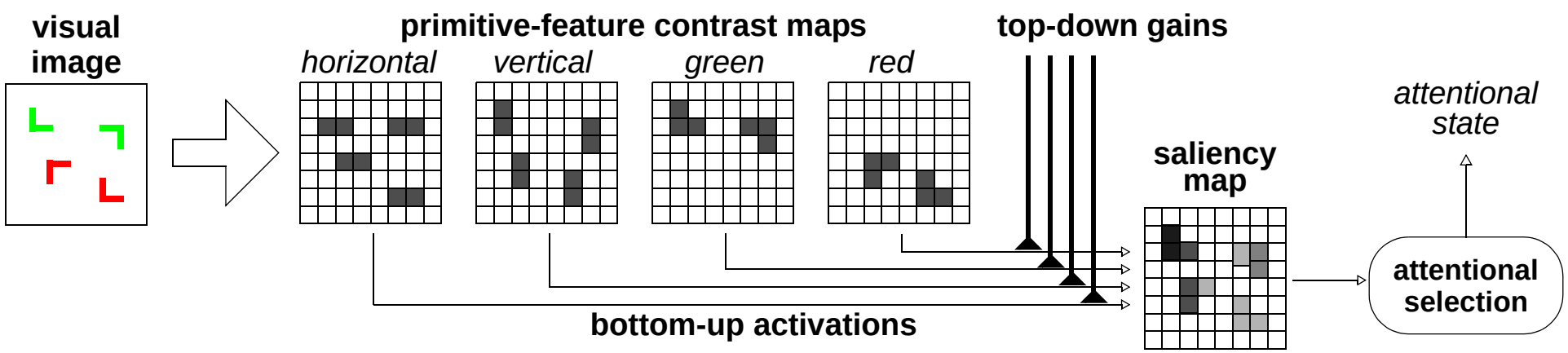
Why Does Repetition Facilitate Performance?

We view attentional control as optimizing performance to the environment in which an individual is operating.

Two-Stage Process

1. Construct predictive (probabilistic) model of the environment based on past experience.
2. Tune control parameters of attention to optimize performance under the current environmental model.

e.g., Itti et al. (1998); Koch & Ullman (1985); Mozer (1991); Wolfe (1994)



Why Does Repetition Facilitate Performance?

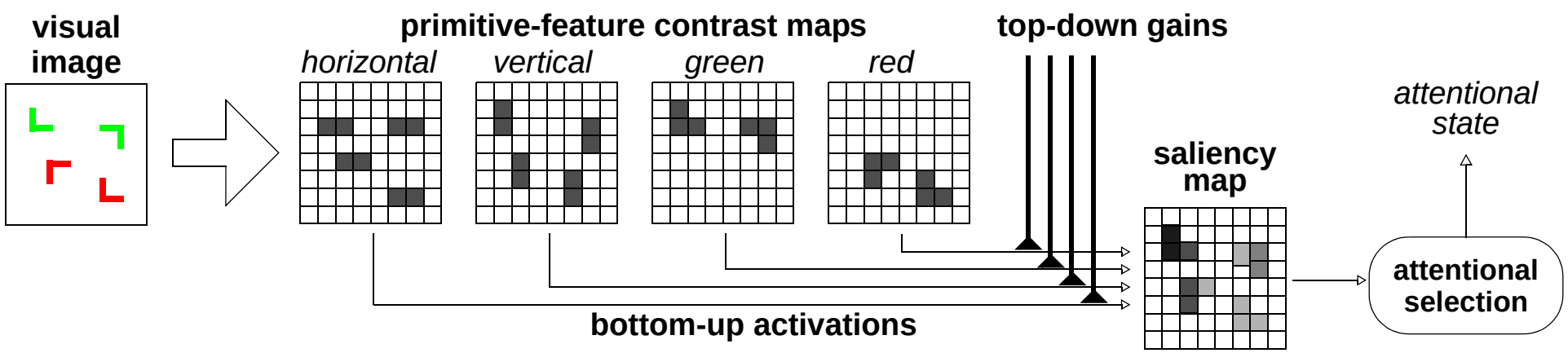
We view attentional control as optimizing performance to the environment in which an individual is operating.

Two-Stage Process

1. Construct predictive (probabilistic) model of the environment based on past experience.

2. Tune control parameters of attention to optimize performance under the current environmental model.

e.g., Itti et al. (1998); Koch & Ullman (1985); Mozer (1991); Wolfe (1994)



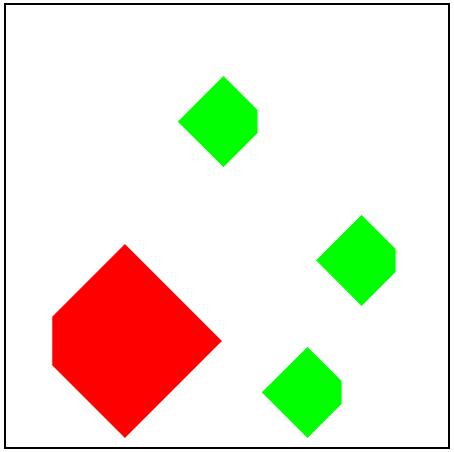
Modeling the Environment

Characterize environment via a probability distribution over configurations of target and distractor features

To simplify presentation, assume distractors are homogeneous.

Example

T_{color}=*red*
T_{size}=*large*
T_{notch}=*left*
D_{color}=*green*
D_{size}=*small*
D_{notch}=*right*



Model

$P(T_{color}, T_{size}, T_{notch}, D_{color}, D_{size}, D_{notch})$

Model 1: Independent Features

$$P(T_{color}, T_{size}, T_{notch}, D_{color}, D_{size}, D_{notch}) =$$
$$P(T_{color})\, P(D_{color})\, P(T_{size})\, P(T_{notch})\, P(D_{size})\, P(D_{notch})$$

Independence assumption is too strong to characterize natural environments.

Model 2: Full Joint Distribution

With 6 features, $2^6 - 1 = 63$ free parameters

T_{color}	D_{color}	T_{size}	T_{notch}	D_{size}	D_{notch}	P(.)
red	*red*	*small*	*left*	*small*	*left*	
green	*red*	*small*	*left*	*small*	*left*	
red	*green*	*small*	*left*	*small*	*left*	
...	*...*	*...*	*...*	*...*	*...*	
green	*green*	*large*	*right*	*large*	*right*	

Requires large amount of experience to obtain accurate probability estimates.

Model 3: Task-Based Architecture

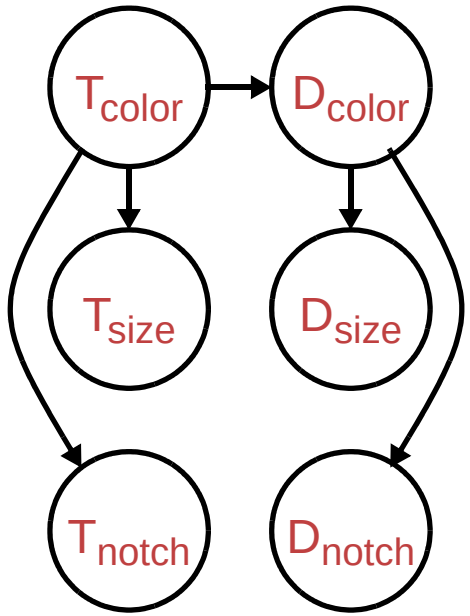
Model 3: Task-Based Architecture

Bayes net

Efficient way of representing high-order probability distributions in terms of low-order distributions

T_{color}	P
red	0.3
green	0.7

T_{size}	T_{color}	P
large	*red*	0.2
small	*red*	0.8
large	*green*	0.6
small	*green*	0.4

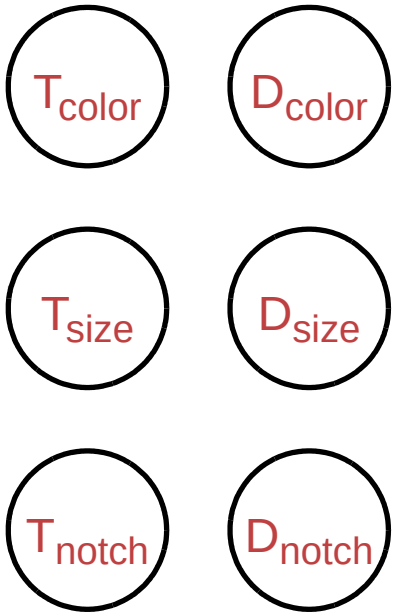


$$P(T_{color}, T_{size}, T_{notch}, D_{color}, D_{size}, D_{notch}) =$$

$$P(T_{color})\, P(D_{color} \mid T_{color})\, P(T_{size} \mid T_{color}) P(T_{notch} \mid T_{color})\, P(D_{size} \mid D_{color})$$

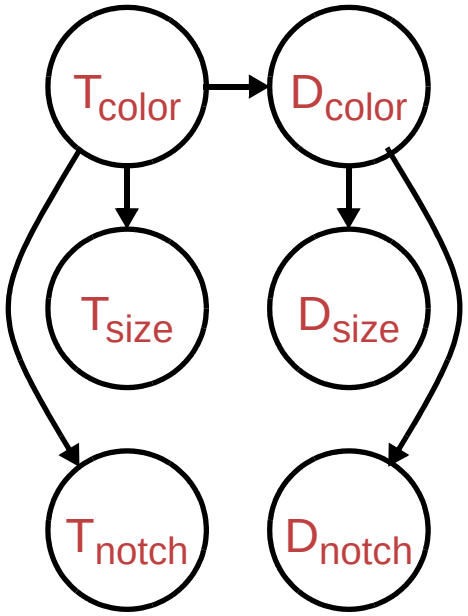
$$P(D_{notch} \mid D_{color})$$

Comparing the Architectures



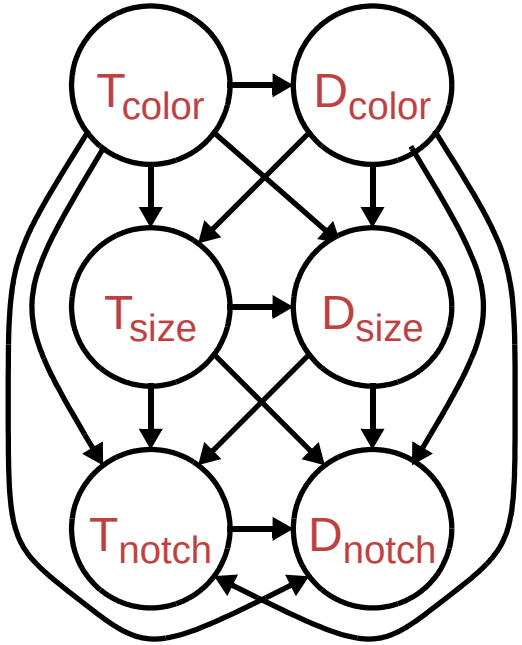
Independence Architecture

6 free parameters



Task-Based Architecture

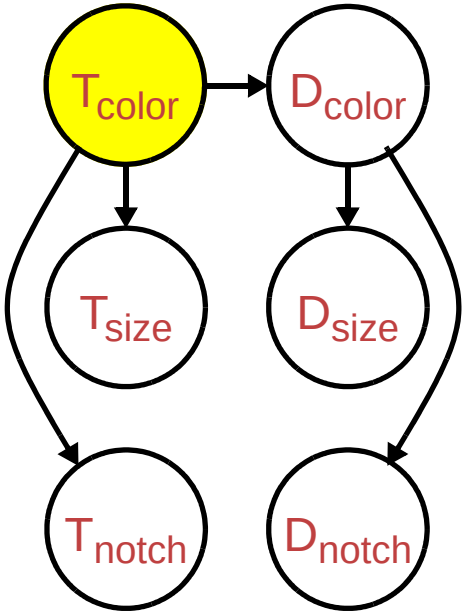
11 free parameters



Full Joint Architecture

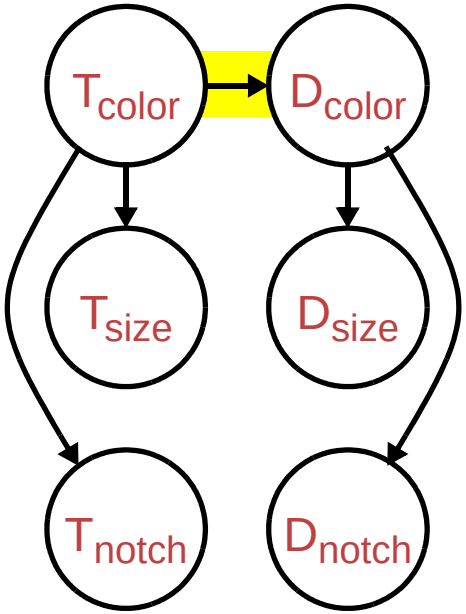
63 free parameters

Key Assumptions of Task-Based Architecture



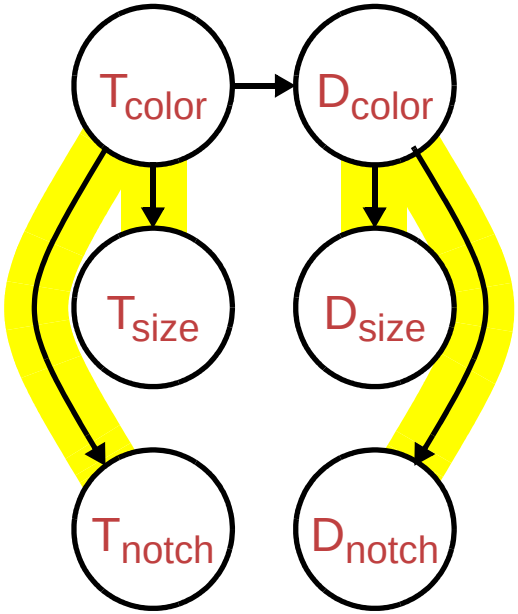
- **Defining feature of target is root of tree.**

Key Assumptions of Task-Based Architecture



- **Defining feature of target is root of tree.**
- **Defining feature of target dominates defining feature of distractor.**

Key Assumptions of Task-Based Architecture



- Defining feature of target is root of tree.
- Defining feature of target dominates defining feature of distractor.
- **Defining feature of target dominates nondefining features of target, and likewise for distractors.**

Simulation of Attentional Adaptation Paradigms

1. **Set up Bayes net for each experiment based on task description.**

Simulation of Attentional Adaptation Paradigms

1. Set up Bayes net for each experiment based on task description.

2. Generate trial sequence that replicates those used in experimental studies.

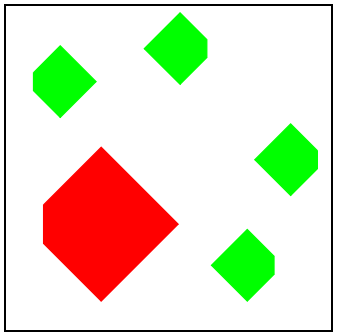
Simulation of Attentional Adaptation Paradigms

1. Set up Bayes net for each experiment based on task description.

2. Generate trial sequence that replicates those used in experimental studies.

3. **Following each trial, update environment model.**

e.g.,



should increase

$P(T_{color} = red)$

$P(T_{size} = large \mid T_{color} = red)$

$P(D_{color} = green \mid T_{color} = red)$

$P(D_{notch} = right \mid D_{color} = green)$ **...**

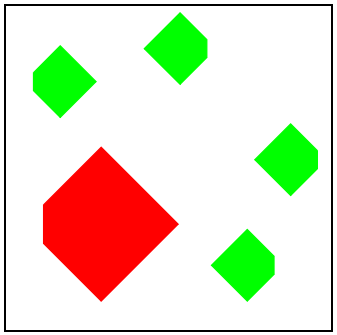
Simulation of Attentional Adaptation Paradigms

1. Set up Bayes net for each experiment based on task description.

2. Generate trial sequence that replicates those used in experimental studies.

3. **Following each trial, update environment model.**

e.g.,



should increase

$P(T_{color} = red)$

$P(T_{size} = large \mid T_{color} = red)$

$P(D_{color} = green \mid T_{color} = red)$

$P(D_{notch} = right \mid D_{color} = green)$ **...**

Simplest scheme: parameter interpolation

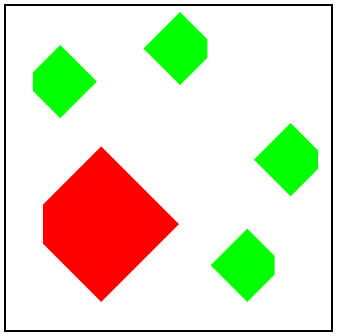
Simulation of Attentional Adaptation Paradigms

1. Set up Bayes net for each experiment based on task description.

2. Generate trial sequence that replicates those used in experimental studies.

3. **Following each trial, update environment model.**

e.g.,



should increase

$P(T_{color} = red)$

$P(T_{size} = large \mid T_{color} = red)$

$P(D_{color} = green \mid T_{color} = red)$

$P(D_{notch} = right \mid D_{color} = green)$ **...**

Simplest scheme: parameter interpolation

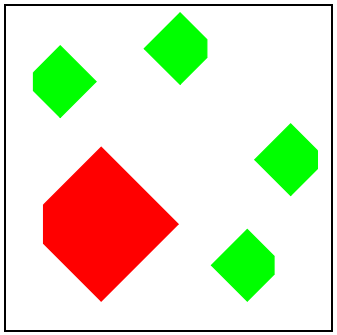
Simulation of Attentional Adaptation Paradigms

1. Set up Bayes net for each experiment based on task description.

2. Generate trial sequence that replicates those used in experimental studies.

3. **Following each trial, update environment model.**

e.g.,



should increase

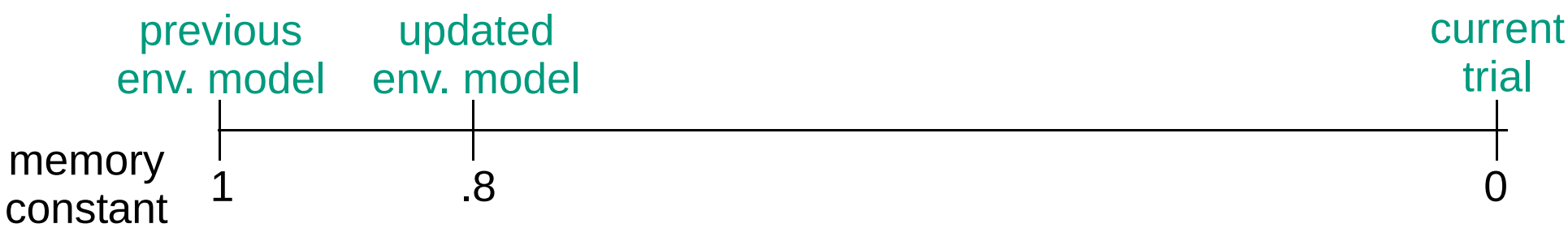
$P(T_{color} = red)$

$P(T_{size} = large \mid T_{color} = red)$

$P(D_{color} = green \mid T_{color} = red)$

$P(D_{notch} = right \mid D_{color} = green)$ **...**

Simplest scheme: parameter interpolation



Simulation of Attentional Adaptation Paradigms

1. **Set up Bayes net for each experiment based on task description.**

2. **Generate trial sequence that replicates those used in experimental studies.**

3. **Following each trial, update environment model.**

4. **Following each update, optimize attentional control to the current environment model.**

Rather than explicitly modeling this optimization process, we assume that it yields RTs that are faster to configurations that have higher probability.

$RT \sim -\log[\, P(T_{color}, T_{size}, T_{notch}, D_{color}, D_{size}, D_{notch}) \,]$

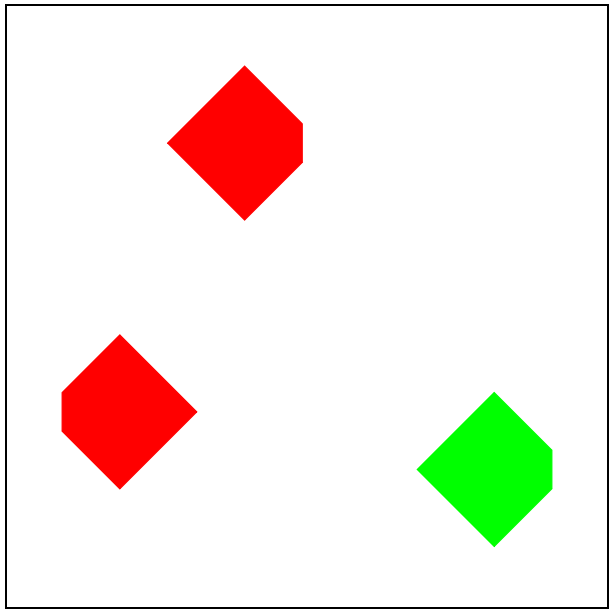
Use this assumption to predict RT on a given trial.

Maljkovic and Nakayama (1994), Experiment 5

Task

Search for color singleton in display of red and green diamonds.

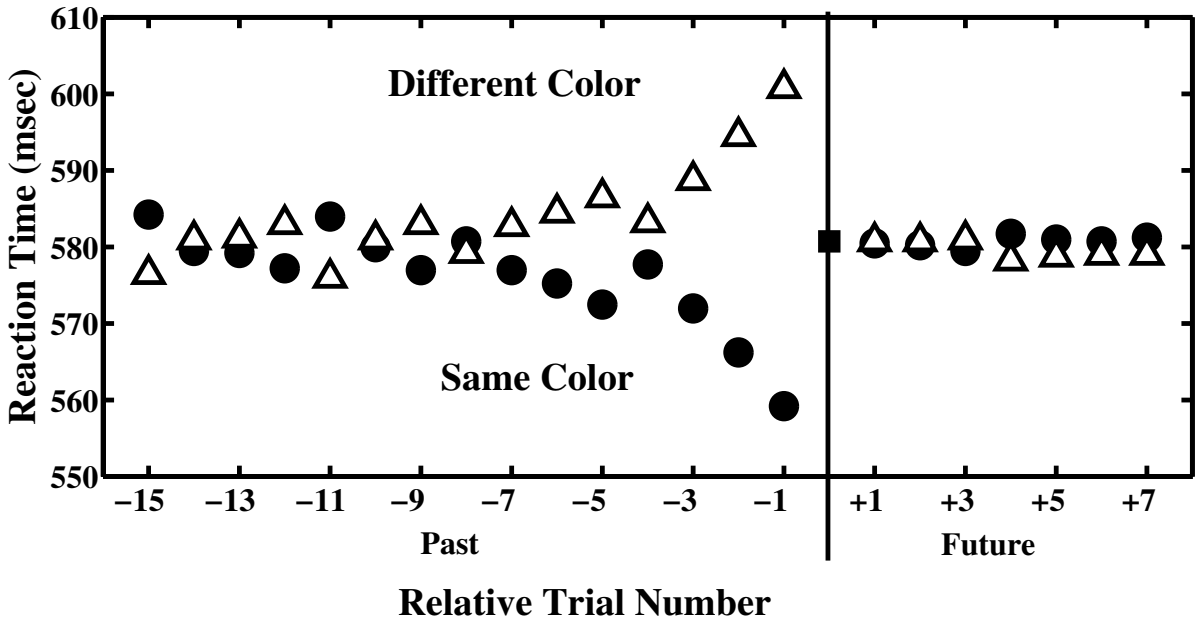
Report whether notch is on left or right.



How does color *k* trials back affect RT on current trial?

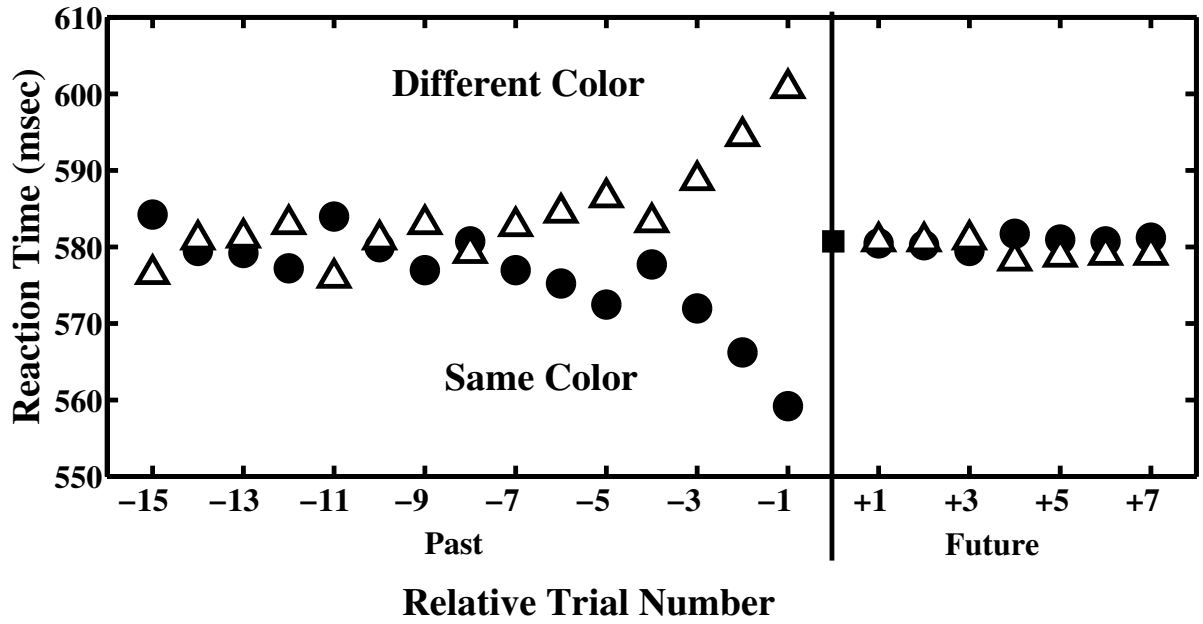
Maljkovic and Nakayama (1994), Experiment 5

Human Data

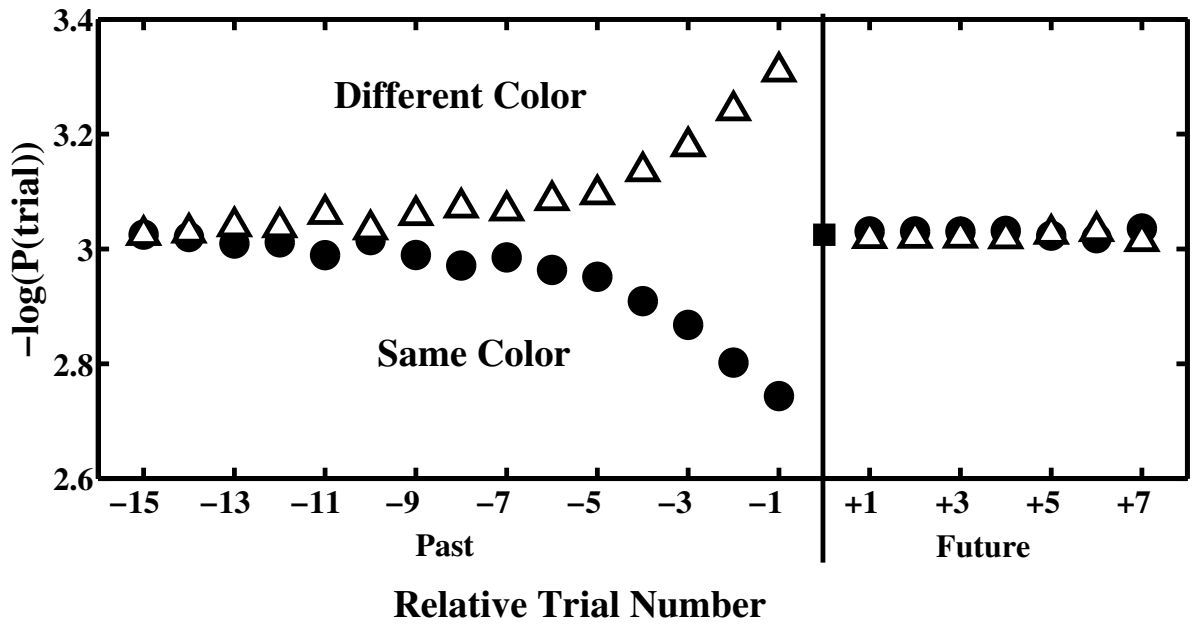


Maljkovic and Nakayama (1994), Experiment 5

Human Data



Simulation

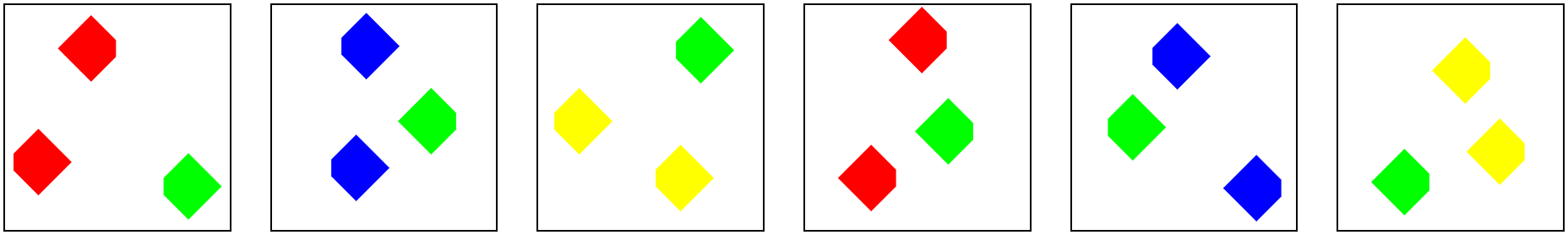


Maljkovic and Nakayama (1994), Experiment 8

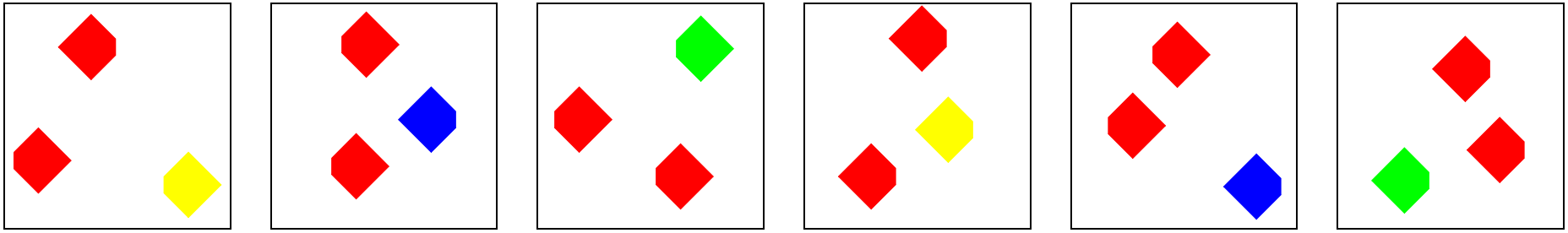
In last experiment, facilitation could be due to repetition of either target *or* distractor color.

In this experiment, four distinct colors.

Repeat target color up to 6 trials, changing distractor color.

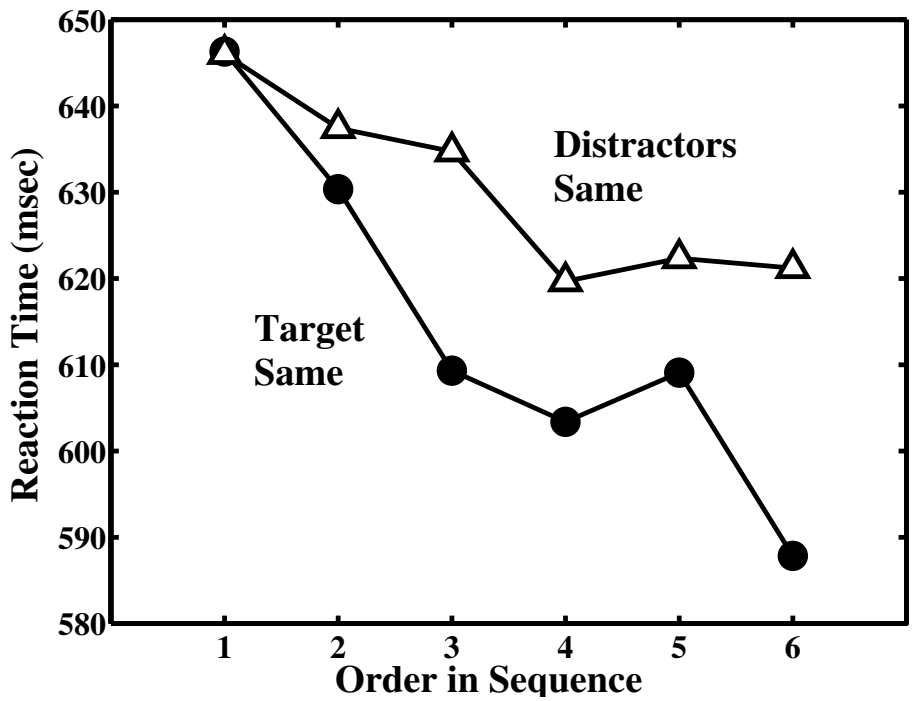


Repeat distractor color up to 6 trials, changing target color.



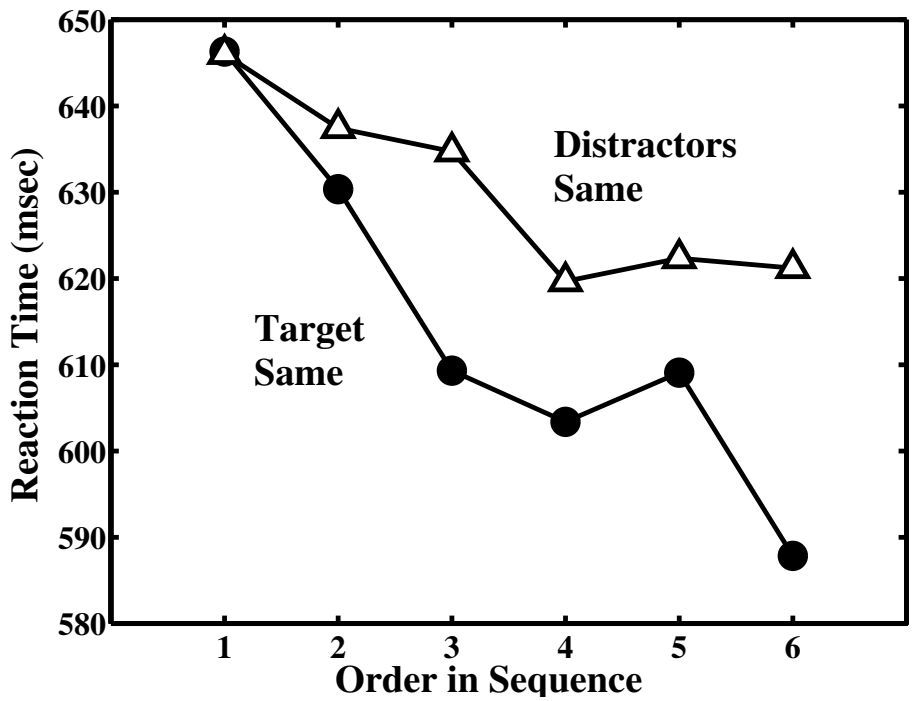
Maljkovic and Nakayama (1994), Experiment 8

Human Data

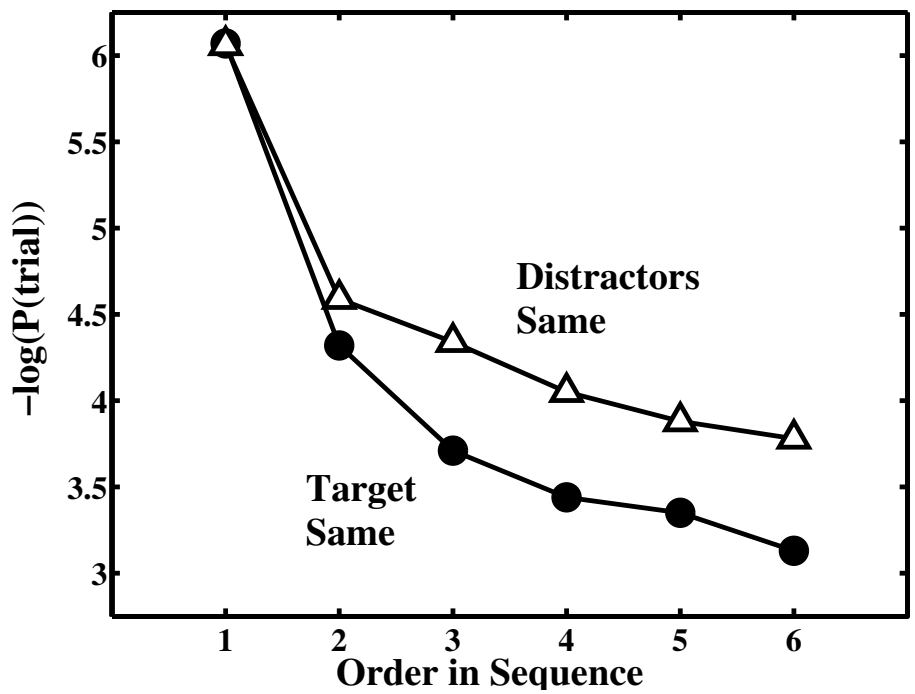


Maljkovic and Nakayama (1994), Experiment 8

Human Data



Simulation



In model, greater effect for target repetition due to dominance of target over distractor.

Huang, Holcombe, and Pashler (2004)

Previous experiments studied only one feature dimension.

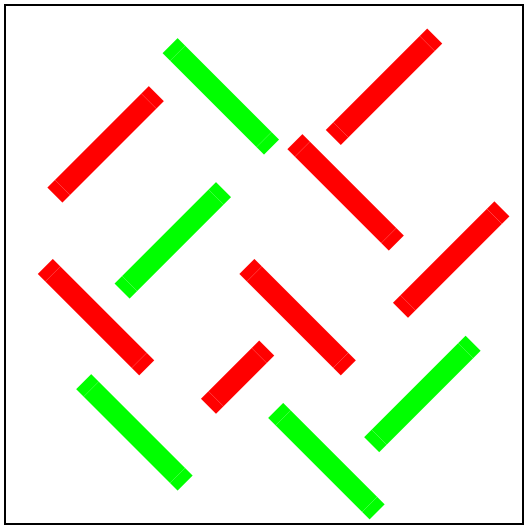
If stimuli vary on multiple dimensions, how do repetitions on one dimension interact with repetitions on another?

Task

Search for singleton in size.

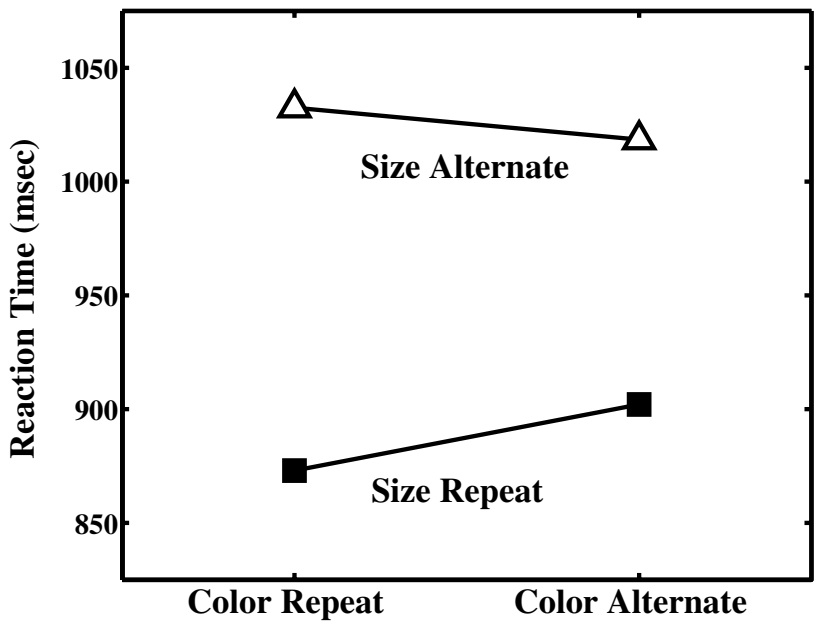
Report slant (left or right).

Color and orientation uncorrelated with size.



Huang, Holcombe, and Pashler (2004)

Human Data

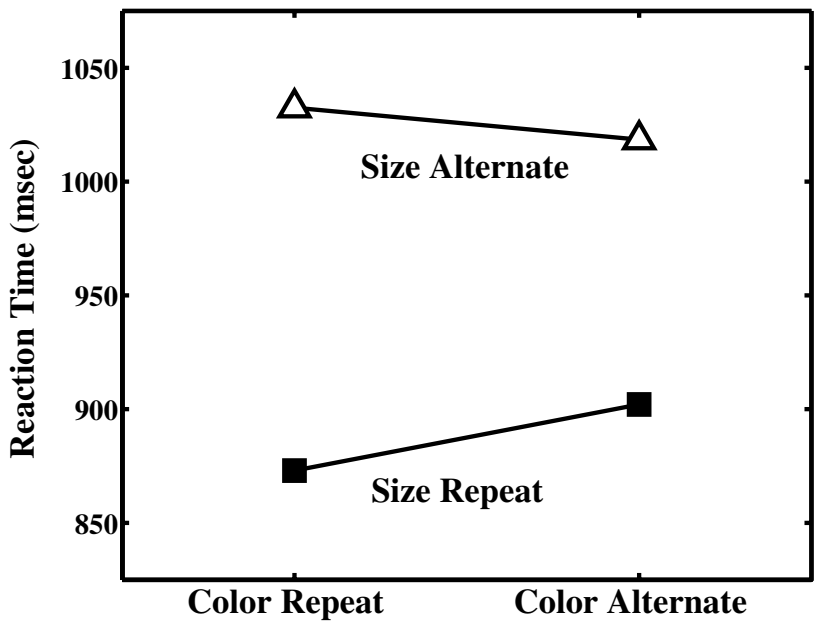


Repetition of defining feature (size) speeds response.

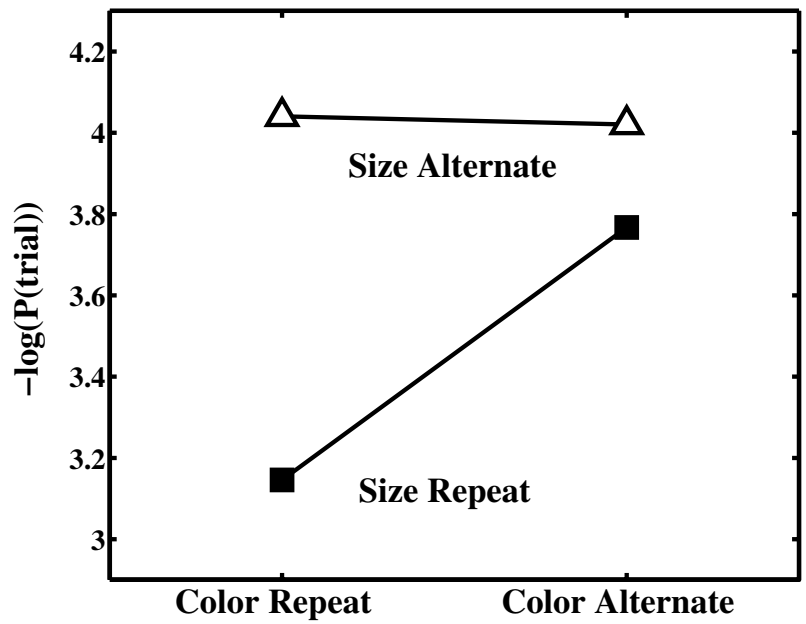
Repetition of nondefining feature (color) speeds response, but only if defining feature is repeated.

Huang, Holcombe, and Pashler (2004)

Human Data



Simulation



Repetition of defining feature (size) speeds response.

Repetition of nondefining feature (color) speeds response, but only if defining feature is repeated.

In model, interaction due to dominance of defining feature over nondefining feature

Wolfe, Butcher, Lee, & Hyle (2003)

Task

Detect presence/absence of singleton in display of colored, oriented lines.

Blocks of trials, corresponding to environments of varying complexity.

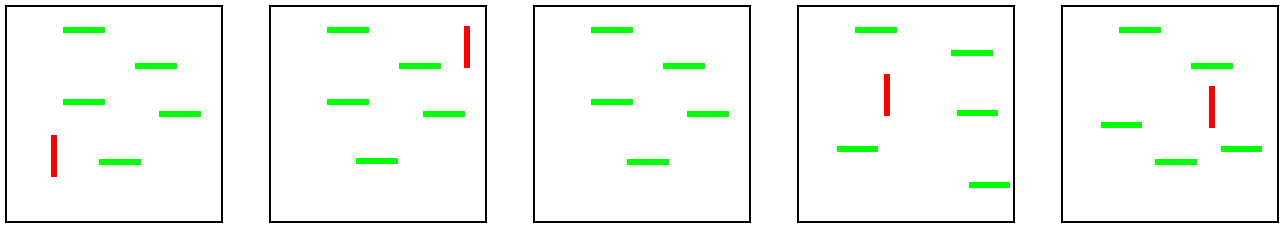
Wolfe, Butcher, Lee, & Hyle (2003)

Task

Detect presence/absence of singleton in display of colored, oriented lines.

Blocks of trials, corresponding to environments of varying complexity.

homogeneous
environment
(red, vertical target)



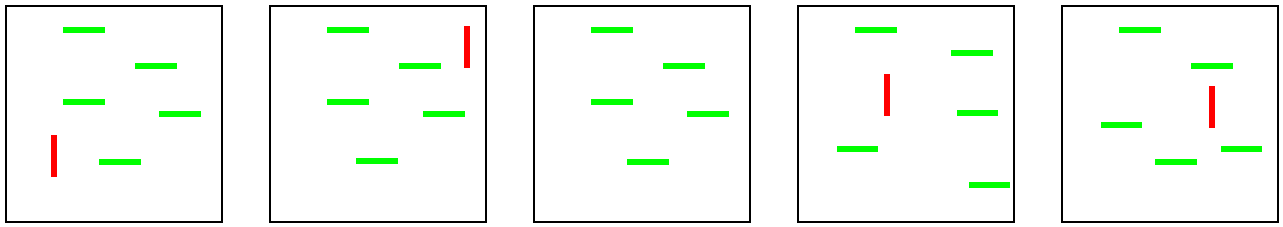
Wolfe, Butcher, Lee, & Hyle (2003)

Task

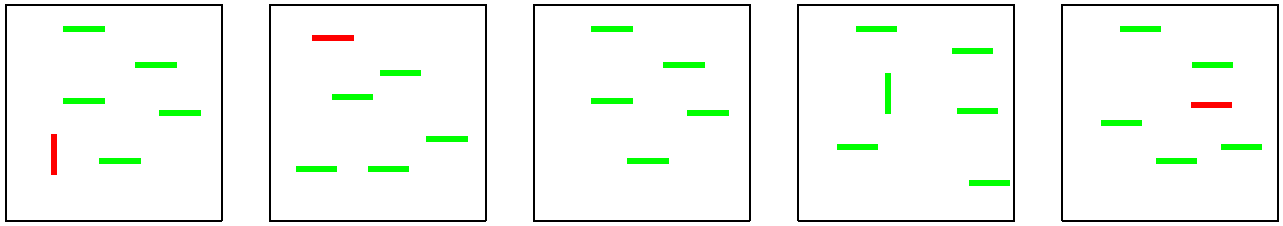
Detect presence/absence of singleton in display of colored, oriented lines.

Blocks of trials, corresponding to environments of varying complexity.

homogeneous environment (red, vertical target)



simple environment (red or vertical target)



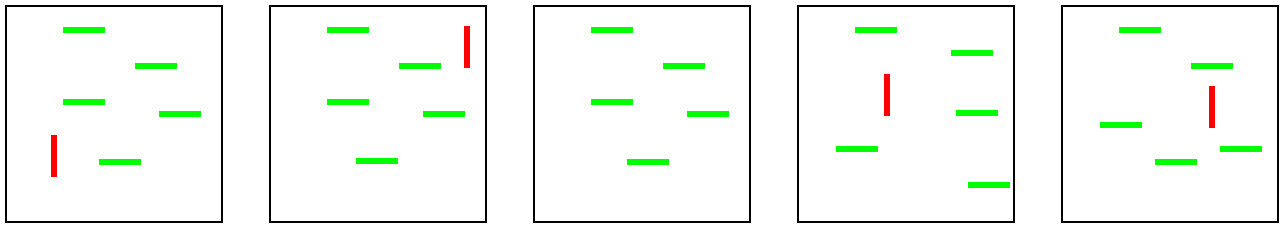
Wolfe, Butcher, Lee, & Hyle (2003)

Task

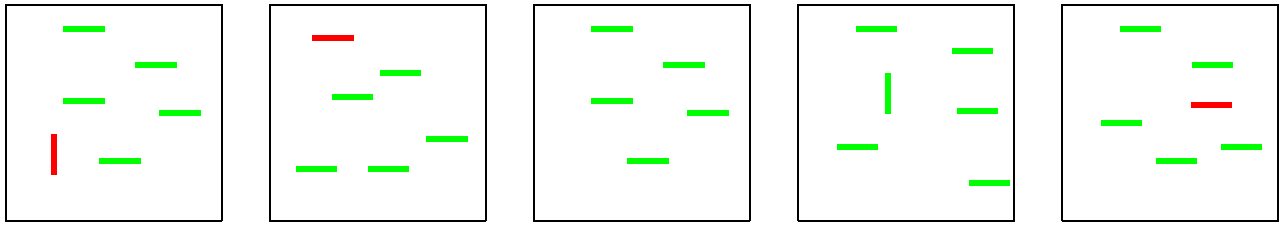
Detect presence/absence of singleton in display of colored, oriented lines.

Blocks of trials, corresponding to environments of varying complexity.

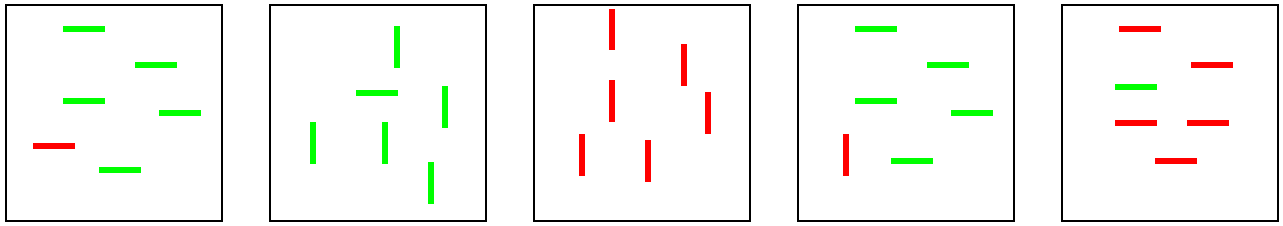
homogeneous environment (red, vertical target)



simple environment (red or vertical target)



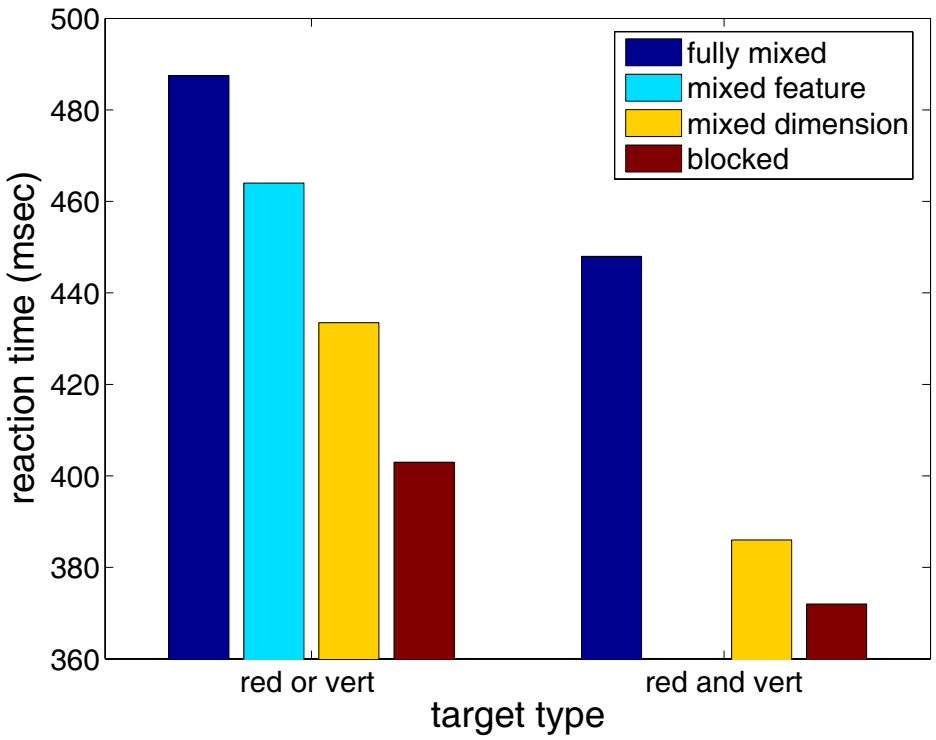
complex environment (target is odd item)



Measure RT on target-present trials.

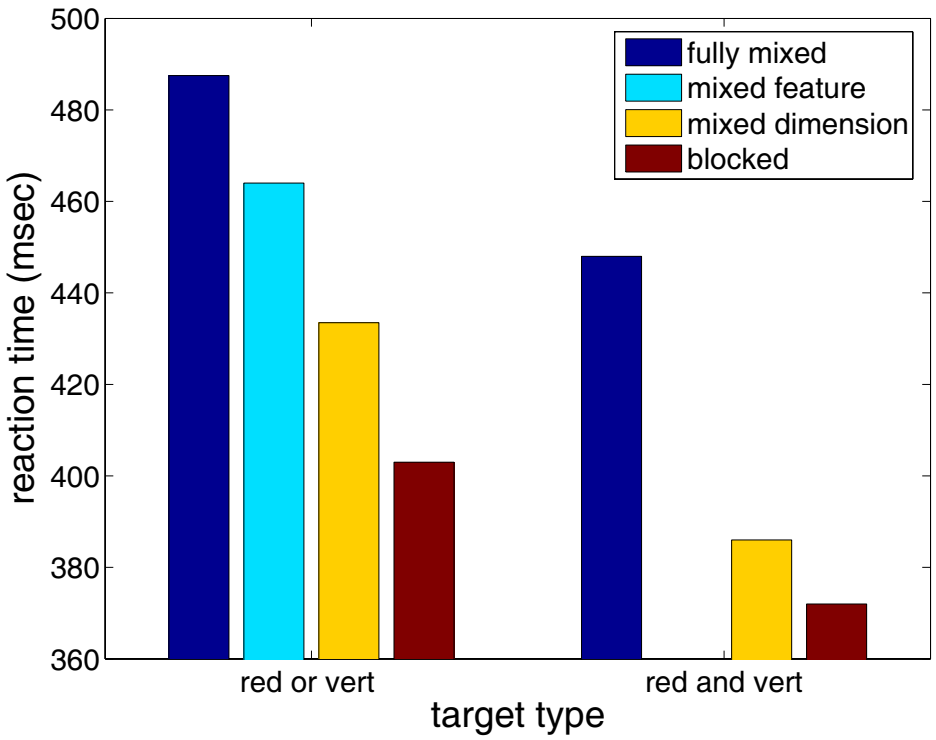
Wolfe, Butcher, Lee, & Hyle (2003), Experiment 1

Human Data

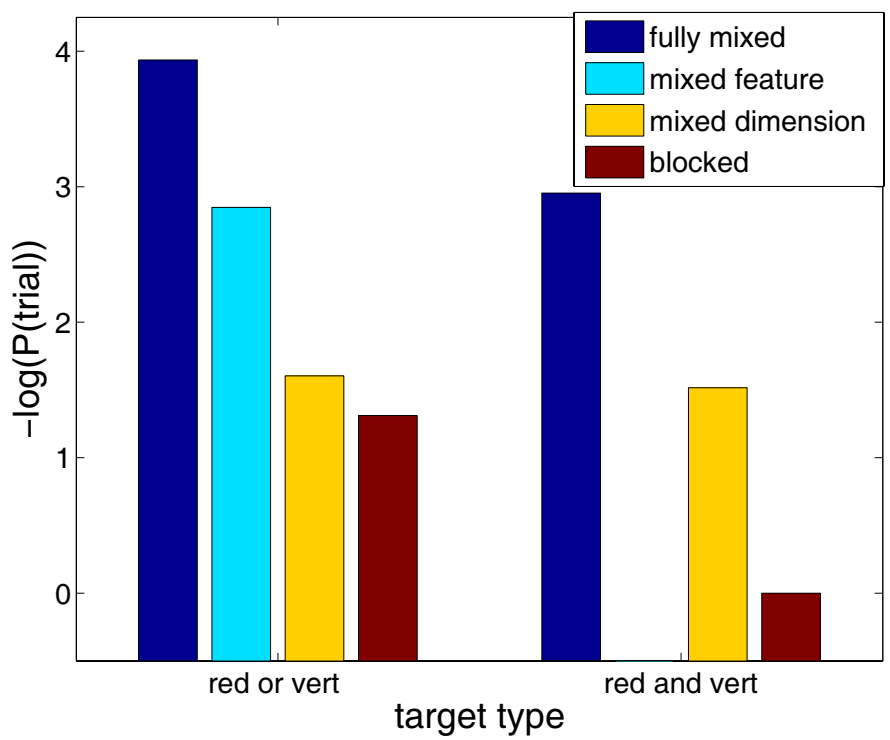


Wolfe, Butcher, Lee, & Hyle (2003), Experiment 1

Human Data



Simulation



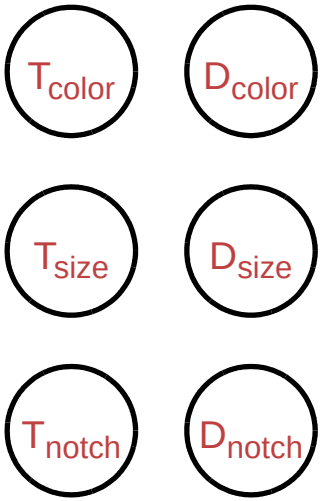
Other Accounts of Attentional Adaptation

Feature-strengthening account (Maljkovic & Nakayama, 1994; Wolfe et al., 2003)

Episodic account (Hillstrom, 2000; Huang et al., 2004)

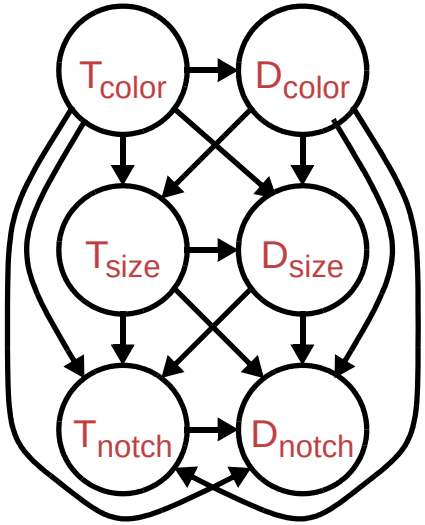
Other Accounts of Attentional Adaptation

Feature-strengthening account (Maljkovic & Nakayama, 1994; Wolfe et al., 2003)



Independence Architecture

Episodic account (Hillstrom, 2000; Huang et al., 2004)

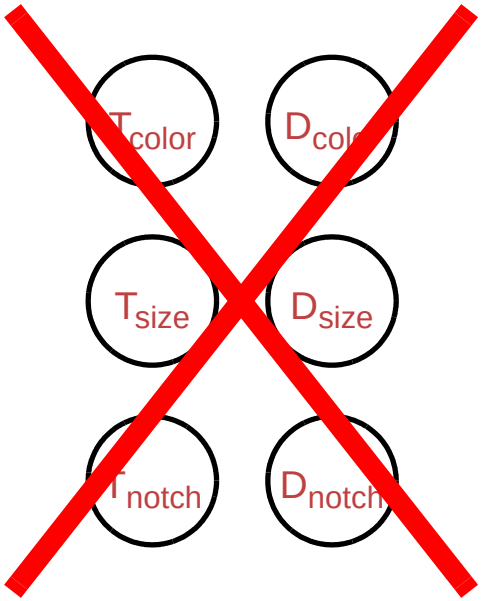


Full Joint Architecture

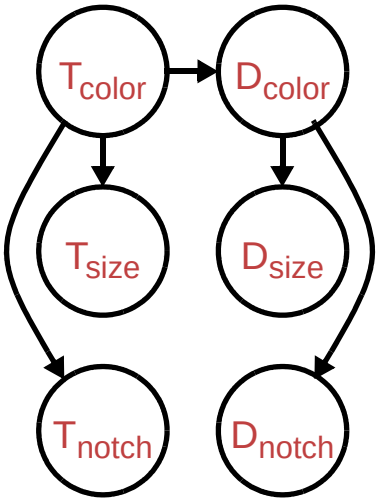
Other Accounts of Attentional Adaptation

Feature-strengthening account (Maljkovic & Nakayama, 1994; Wolfe et al., 2003)

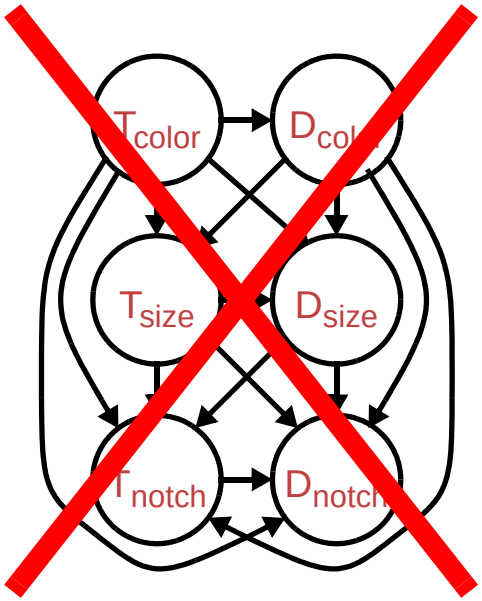
Episodic account (Hillstrom, 2000; Huang et al., 2004)



Independence Architecture



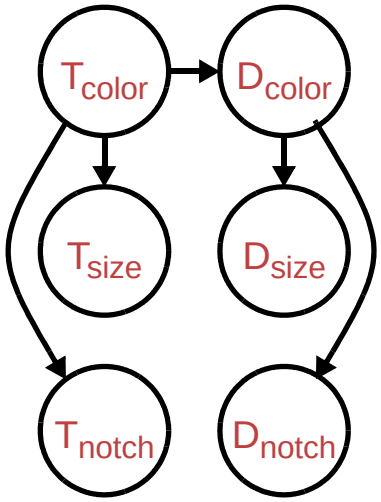
Task-Based Architecture



Full Joint Architecture

Neither is adequate to explain the range of data

Two Ways to View Architecture



- **model of the structure of the environment**
- **model of attentional control**

Rational account

What appears to be cognitive control is a consequence of optimizing performance to the structure of the environment, subject to structural restrictions in the architecture.

Different than traditional view of control as a resource to be allocated.

Pushing the Rational Account Further

Influence of past experience decays rapidly. Why?

Pressures on duration of influence

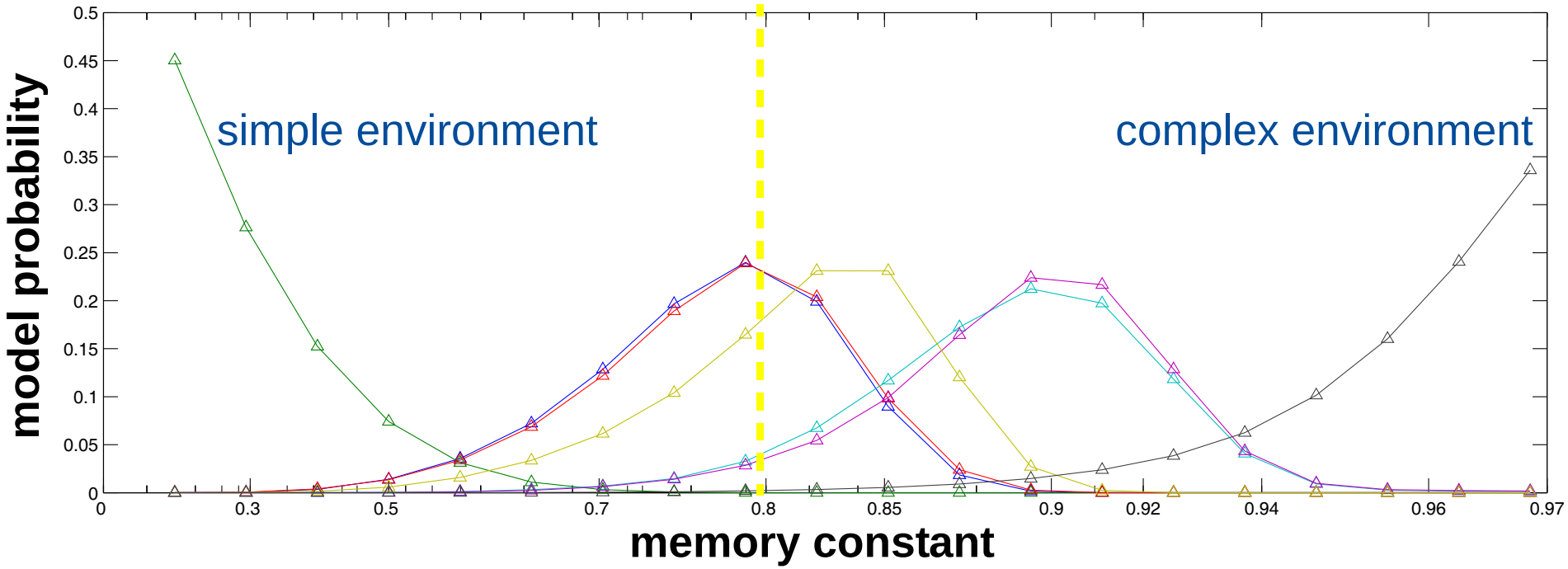
Adapting quickly to new environment $\rightarrow$ short lived influence

Capturing statistics of complex environments $\rightarrow$ long-lived influence

Is observed memory duration optimal?

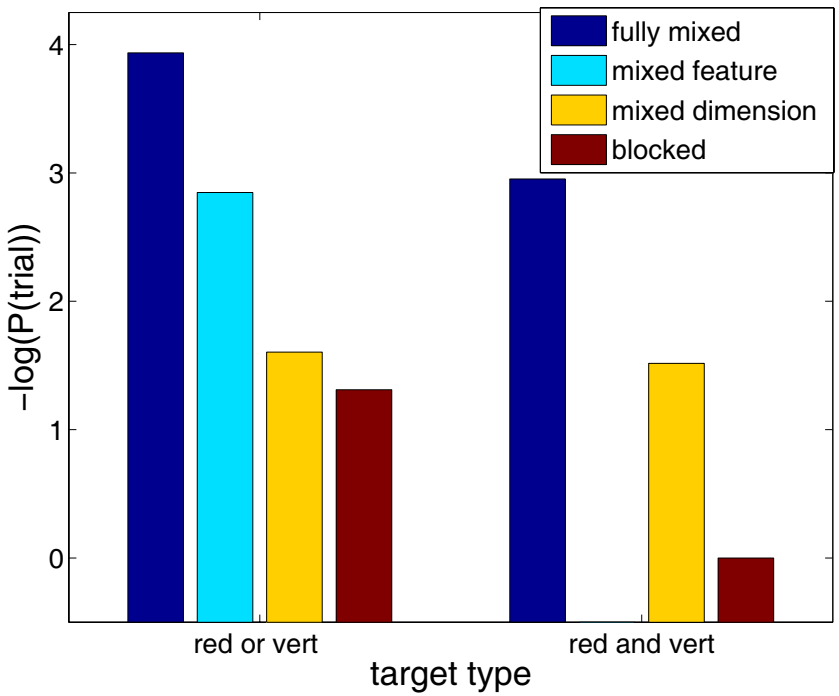
Use Bayesian model selection to determine appropriate memory constant in a given environment.

Posteriors on Memory Constant for Environments of Wolfe et al.

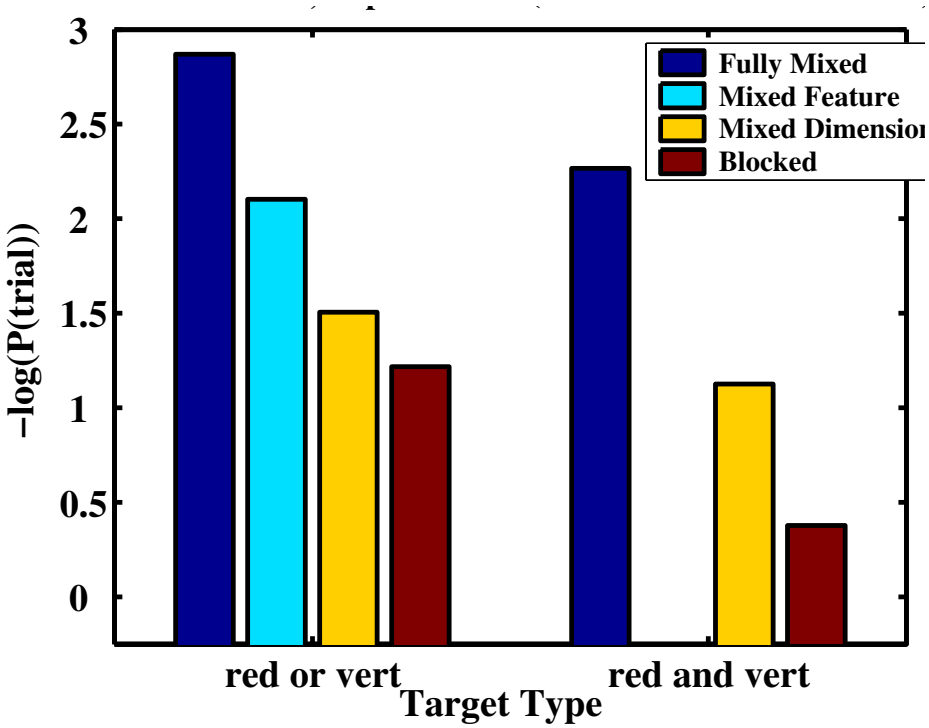


Prediction Via Model Averaging

original simulation

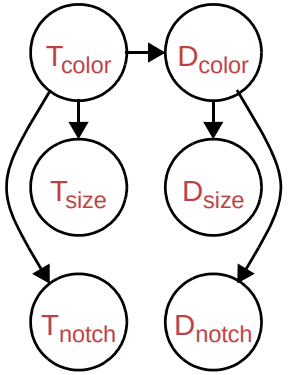


Bayesian model averaging

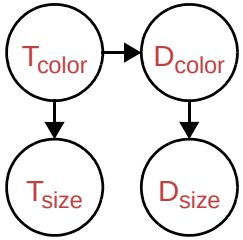


Eliminates the one free parameter of model

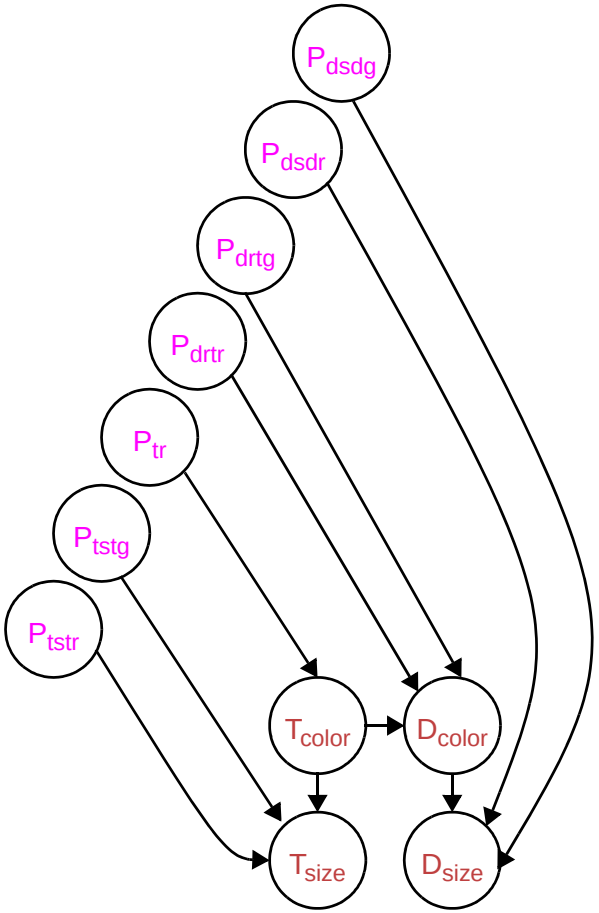
Fancy Version of Model



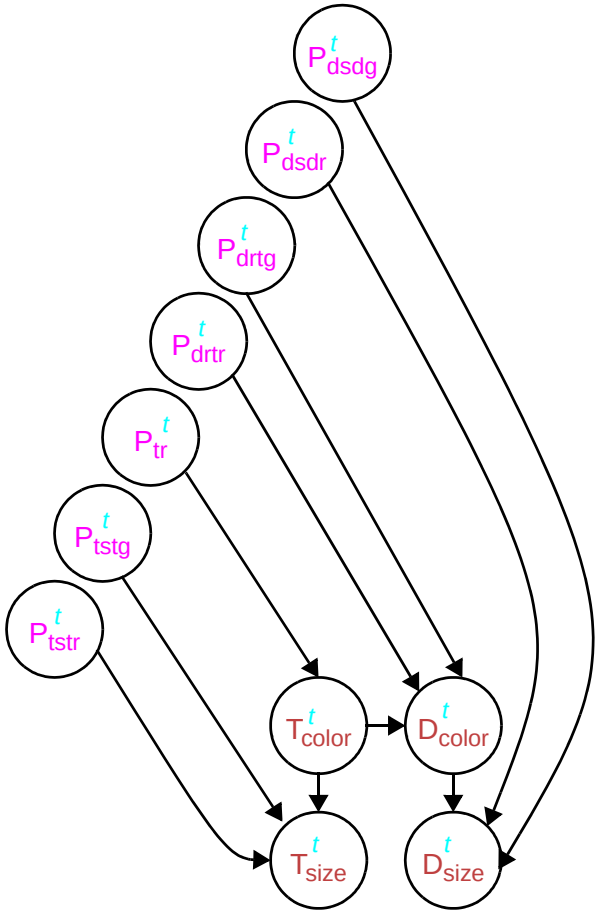
Fancy Version of Model



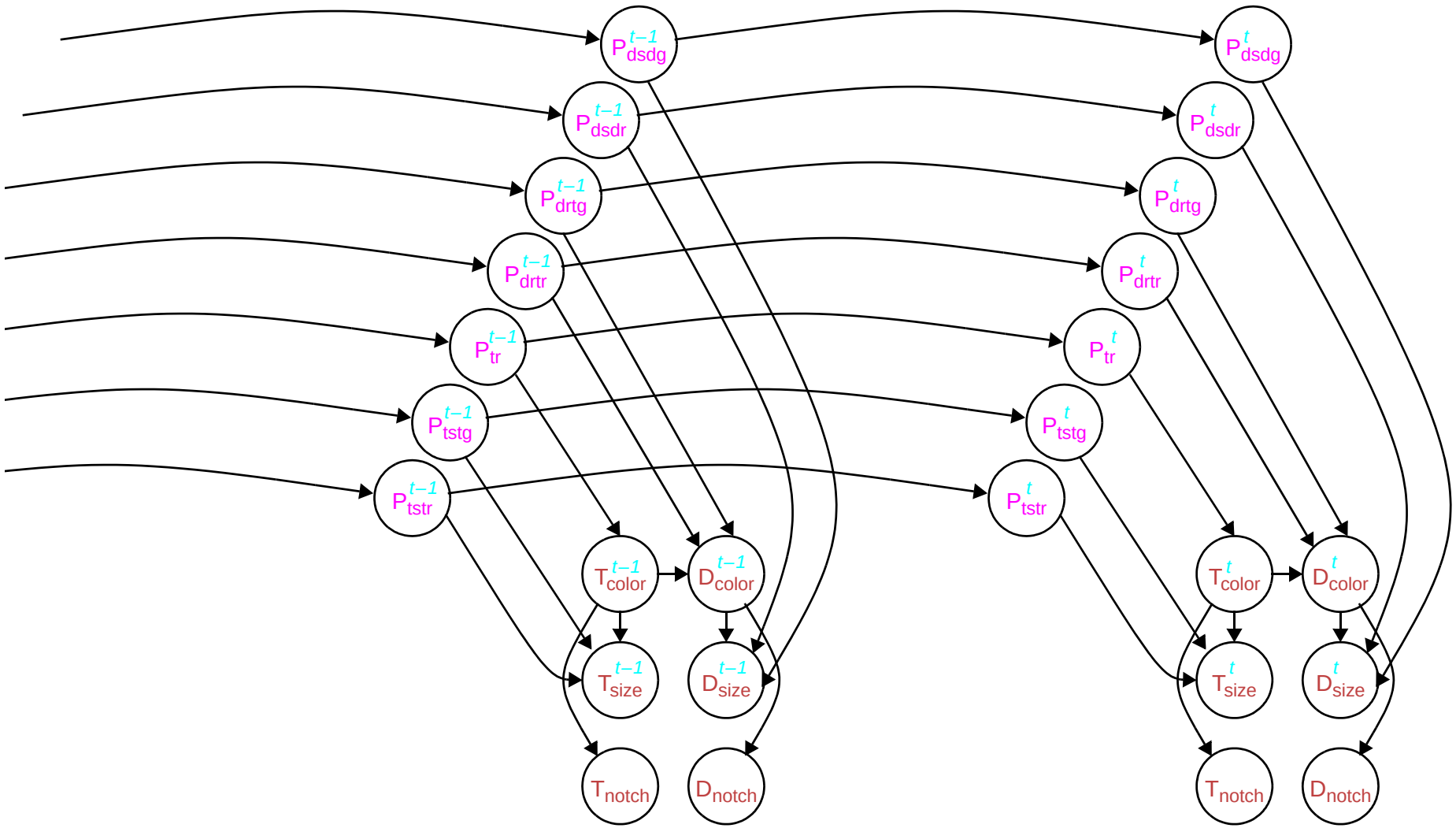
Fancy Version of Model



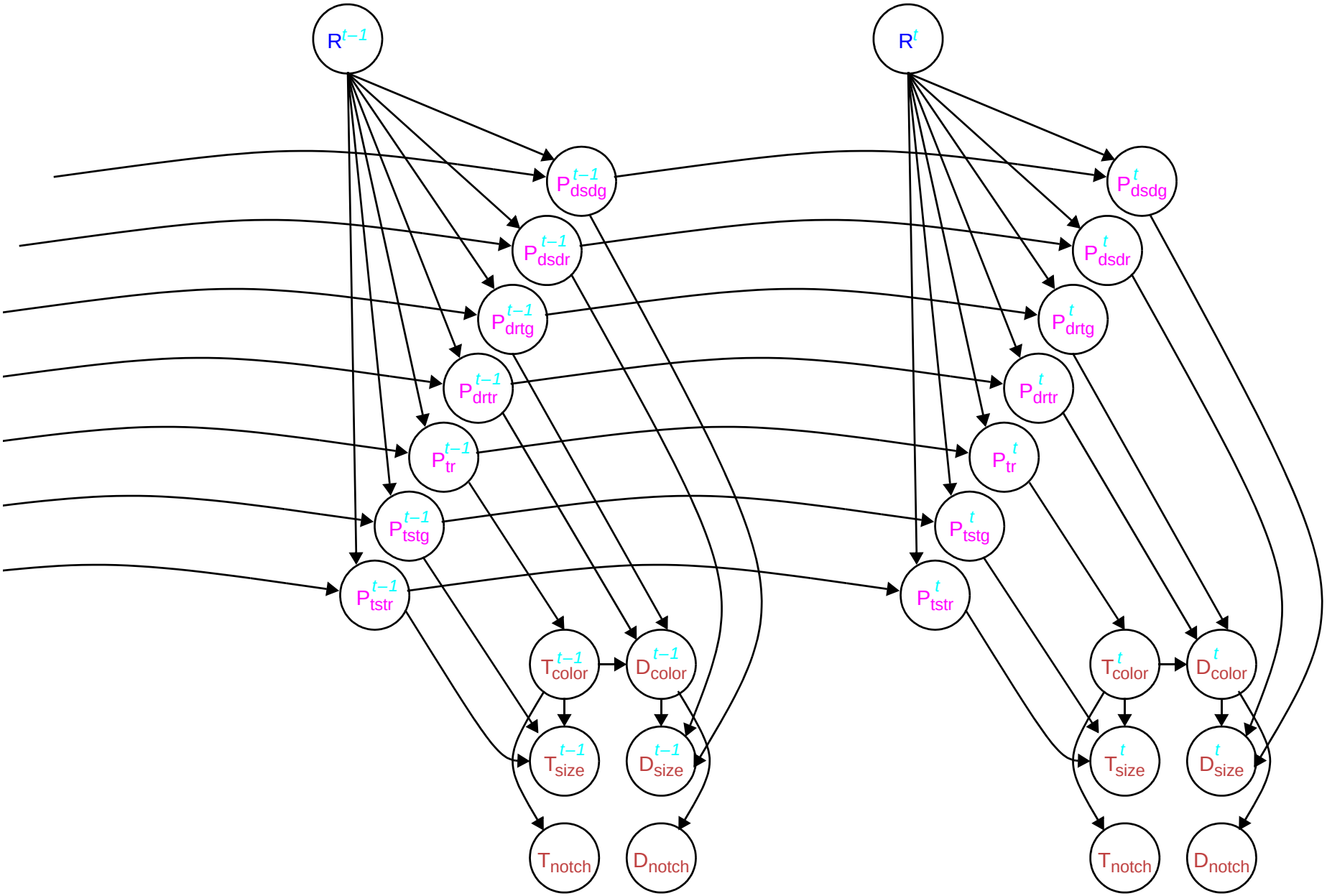
Fancy Version of Model



Fancy Version of Model



Fancy Version of Model



What Does This Work Tell Us About Attentional Control?

What Does This Work Tell Us About Attentional Control?

Tendency in literature to treat control as a resource to be allocated.

Some tasks invoke more top-down control than others (Wolfe et al., 2003)

Accounts often imply homunculus that distributes activation or attention.

Our perspective

Much of what appears to be cognitive control is a consequence of optimizing performance to the structure of the environment.

Control settings can clearly be refined once environment has been experienced.

Comparison of Approaches

	Bayesian	Connectionist
representation of prior knowledge	principled, explicit	ad hoc, implicit
learning/processing mechanisms	principled, elegant; *optimal* given inclusive model space	ad hoc, great freedom for modeler to tweak mechanisms to fit data
incorporation of neurobiological data	difficult	easy
learning from small number of examples	very powerful tool, given appropriate biases	underestimates human learning abilities
statistical learning	easy	easy
rule learning	easy	difficult