Understanding Long-Term Priming via a Probabilistic Model

Michael C. Mozer
Michael Colagrosso
David E. Huber

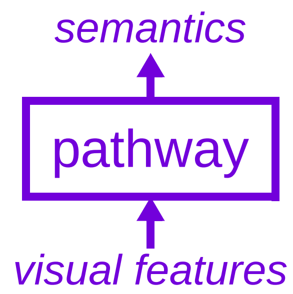
Institute of Cognitive Science and
Department of Computer Science
University of Colorado, Boulder

Goals

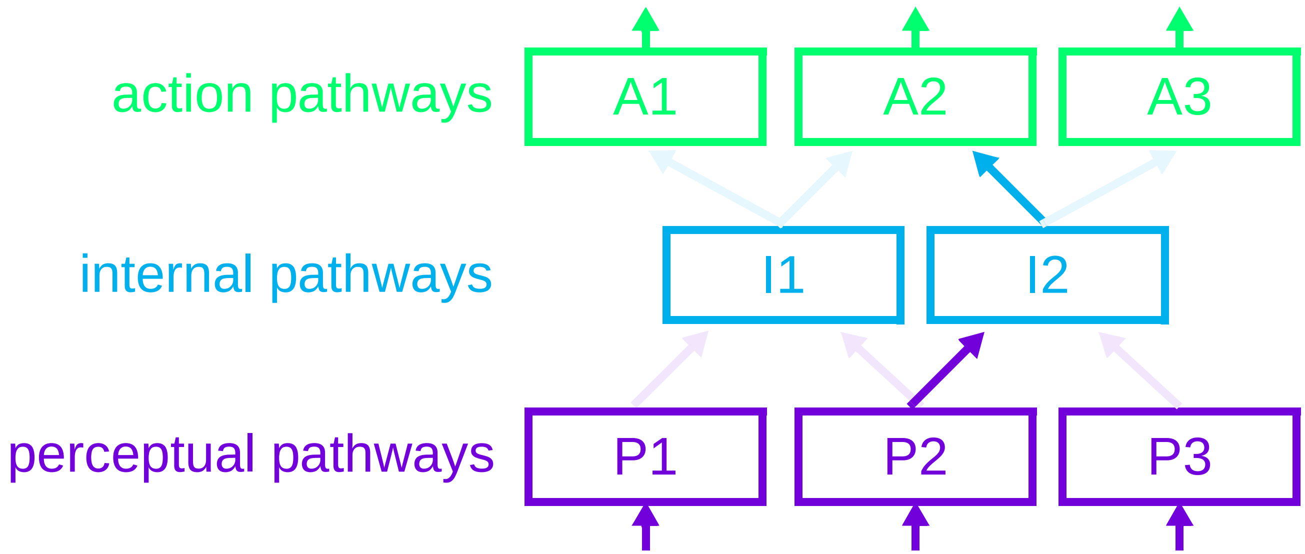
- **Characterize brain computation in terms of information transmission along cortical *pathways.***
- **Characterize skill acquisition/practice/rehearsal (*learning*) in terms of changes in information flow.**
- **Propose and evaluate learning mechanisms that underlie these changes.**
- **Model data on repetition priming and skill acquisition.**

Pathway Architecture

- **Cortical computation is performed by a set of functionally specialized *pathways.***

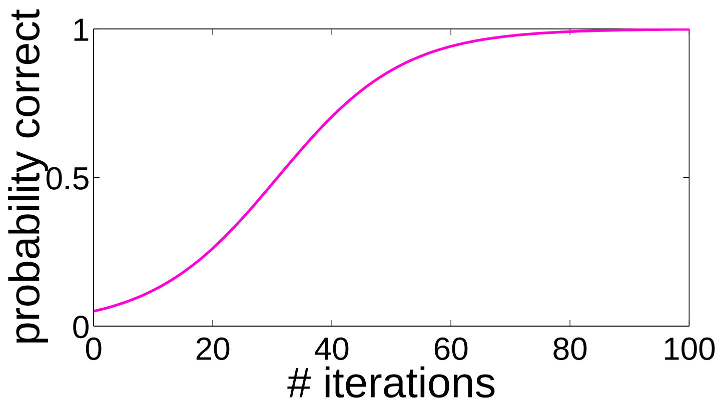


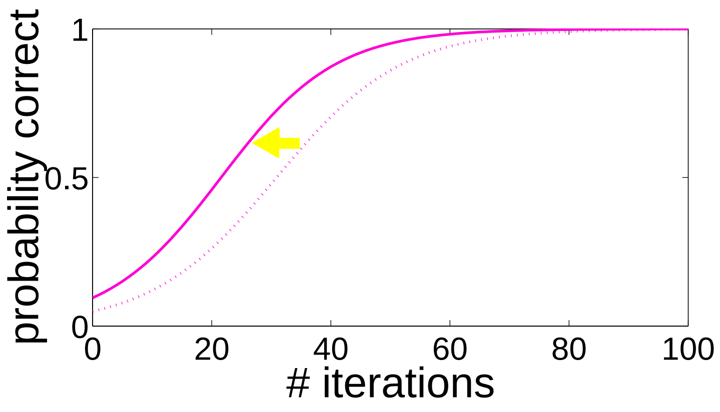
- **Cognition arises from dynamic, task dependent interconnections among pathways.**



Pathway Architecture

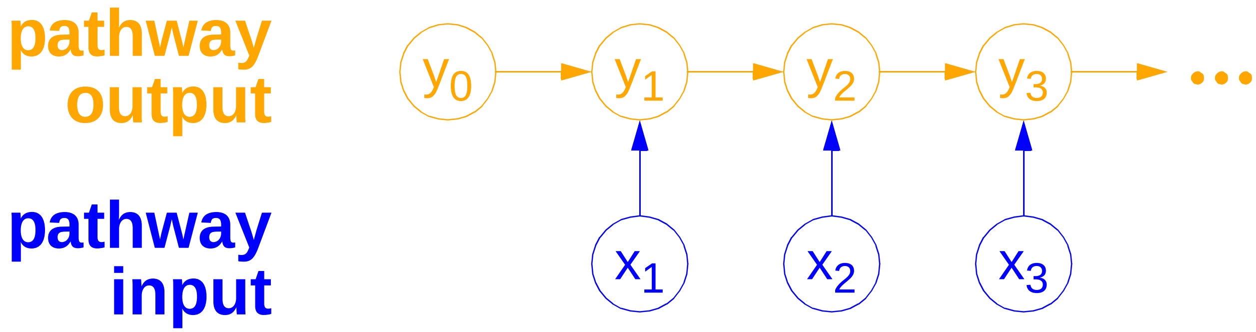
- **Pathway operation shows a speed-accuracy trade off.**





- **With each experience, pathway tends to produce its response more rapidly.**
- **Output of one pathway continuously transmitted to input of another, but no other communication allowed.**
- **Subject to these constraints and knowledge limitations, inference in a pathway is optimal (the *rational perspective*).**

Pathway as a Dynamic Belief Net



Informal description

Input at time t combines with output at time t−1 to determine output at time t.

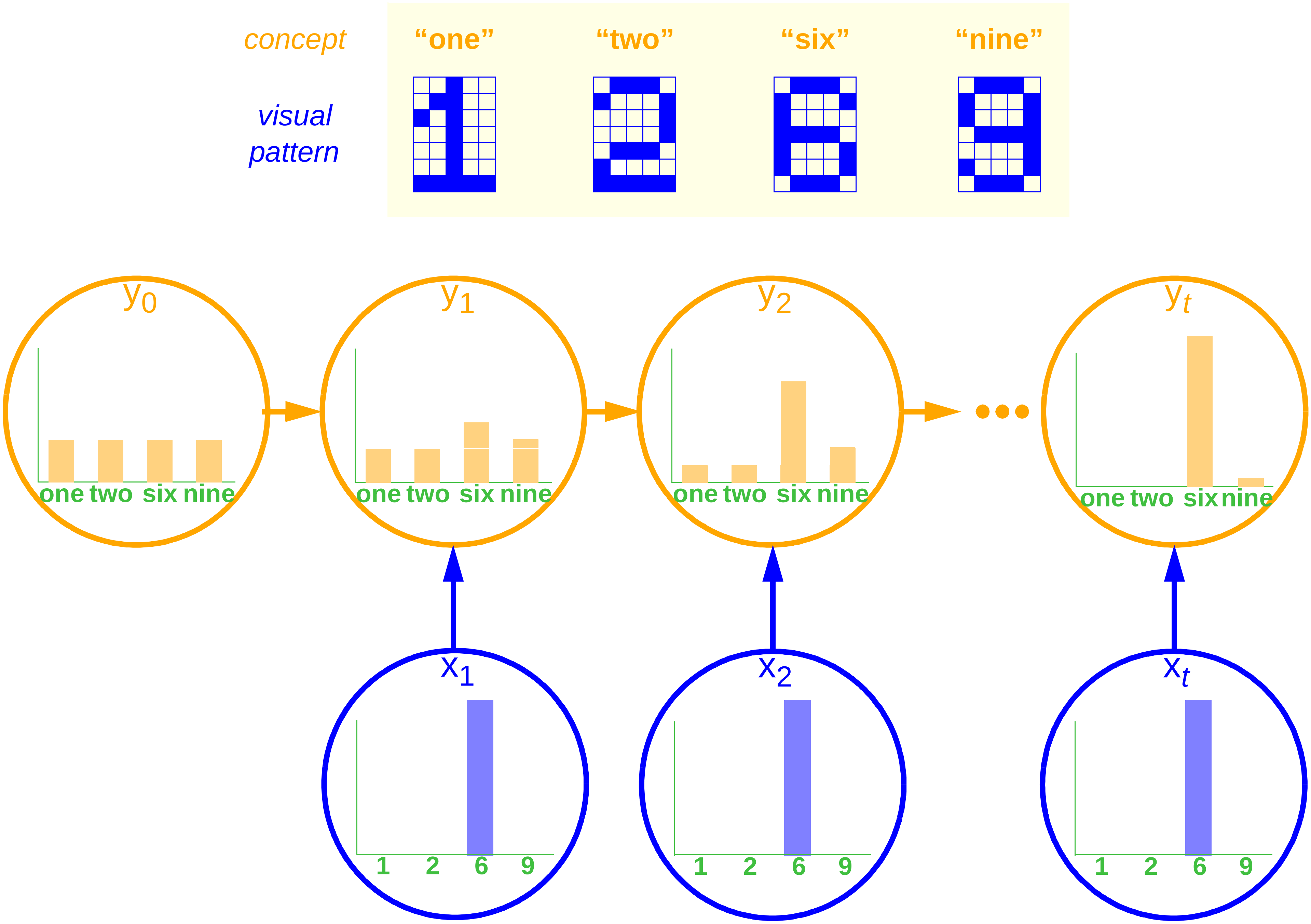
Formal description

Each node is a discrete random variable.

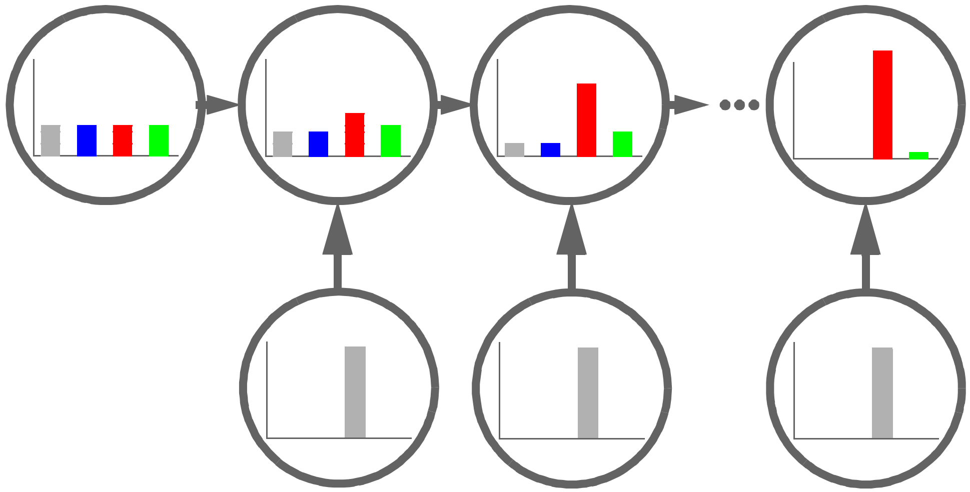
Each arrow indicates a conditional dependency.

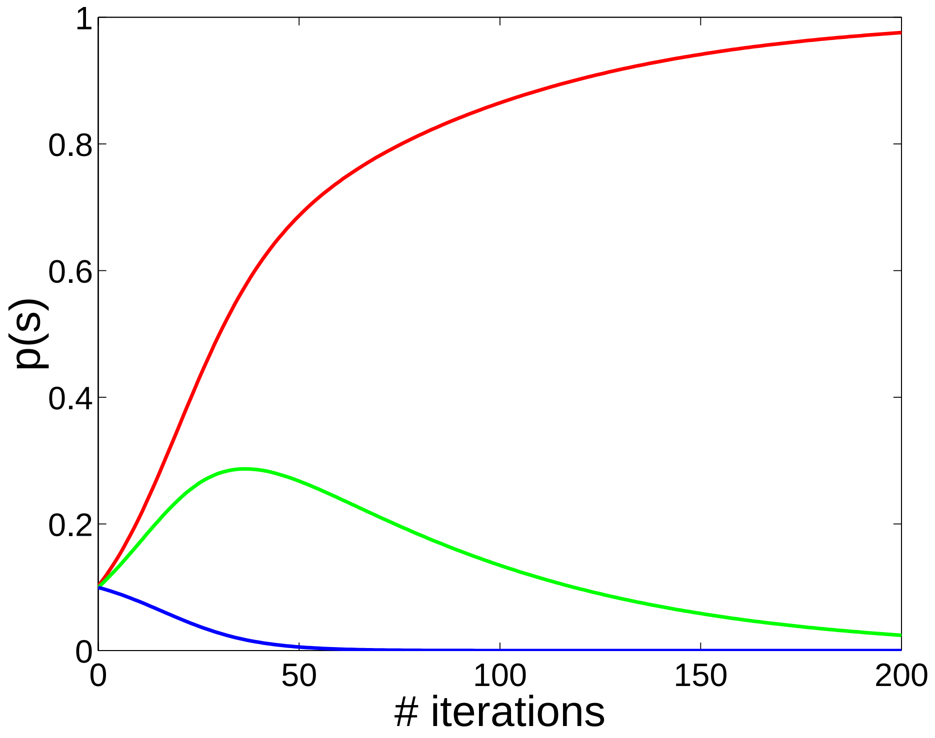
$p(y_3 \mid x_1, x_2, x_3, y_0, y_1, y_2) = p(y_3 \mid x_3, y_2)$

Illustration of Pathway Behavior

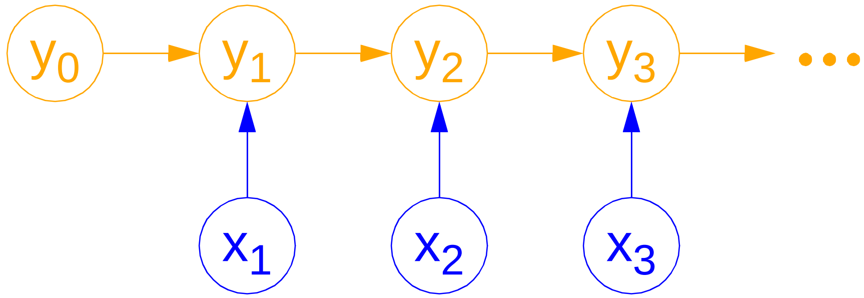


Simple Illustration of Pathway Behavior

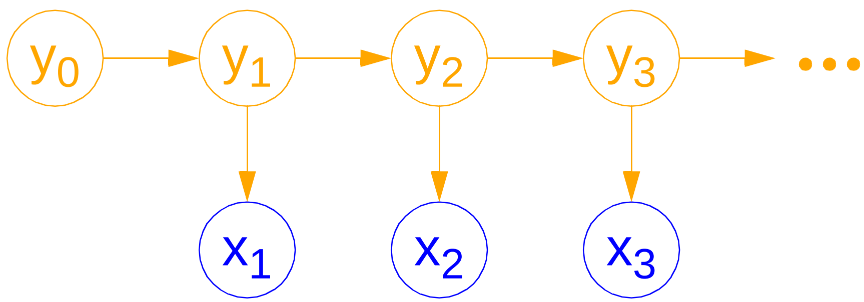




Recognition framework

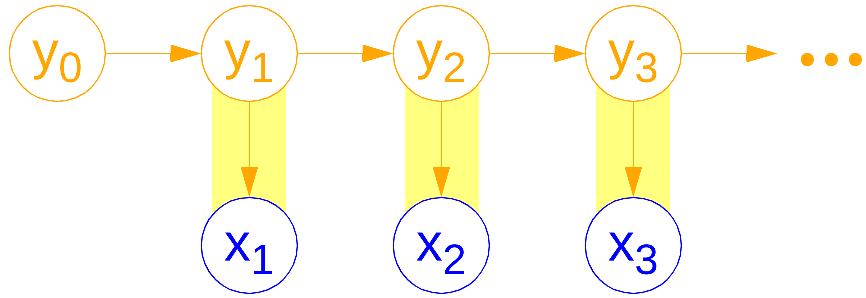


Generative framework



Inference can be performed in any direction regardless of direction of arrows.

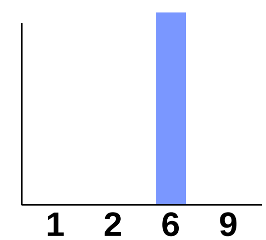
Input-Output Association



Strong one-to-one association

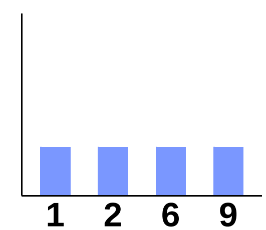
$$p(x_t = i \mid y_t = j) = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{otherwise} \end{cases}$$

“six”



No association

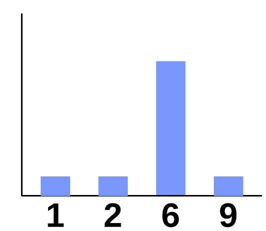
$$p(x_t = i \mid y_t = j) = \frac{1}{N_x}$$



Intermediate association strength

$$p(x_t = i \mid y_t = j) \sim \begin{cases} \varepsilon + \alpha & \text{if } i=j \\ \varepsilon & \text{otherwise} \end{cases}$$

ε	task difficulty
α	task familiarity



Input-Output Association

General formulation

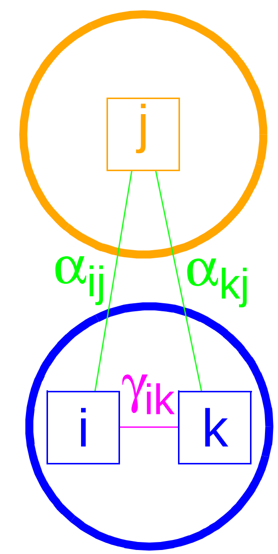
$$p(x_t = i \mid y_t = j) \sim \varepsilon + \alpha_{ij}$$

ε	task difficulty
α_{ij}	familiarity with x_i–y_j association

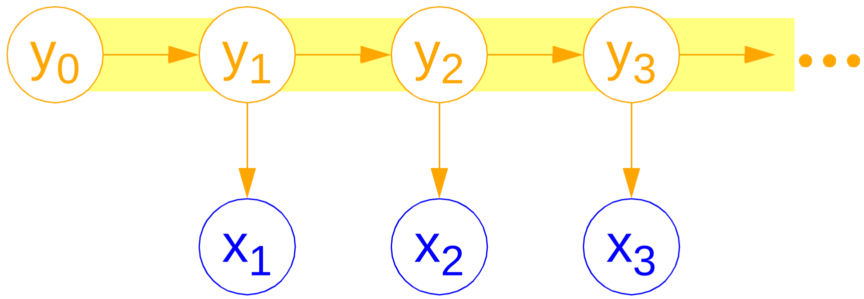
Similarity structure among inputs

$$p(x_t = i \mid y_t = j) \sim \varepsilon + \sum_k \gamma_{ik}\alpha_{kj}$$

γ_{ik}	similarity between i and k



Output Transitions



Perfect memory

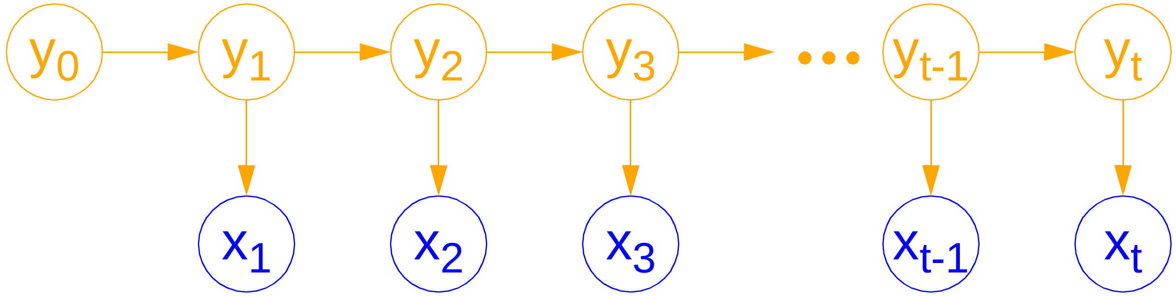
$$p(y_t = i \mid y_{t-1} = j) = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{otherwise} \end{cases}$$

Memory with diffusion

$$p(y_t = i \mid y_{t-1} = j) = \begin{cases} (1-\beta) + \beta / N_y & \text{if } i=j \\ \beta / N_y & \text{otherwise} \end{cases}$$

β	diffusion rate (forgetting)

Inference in a Dynamic Belief Network



Given

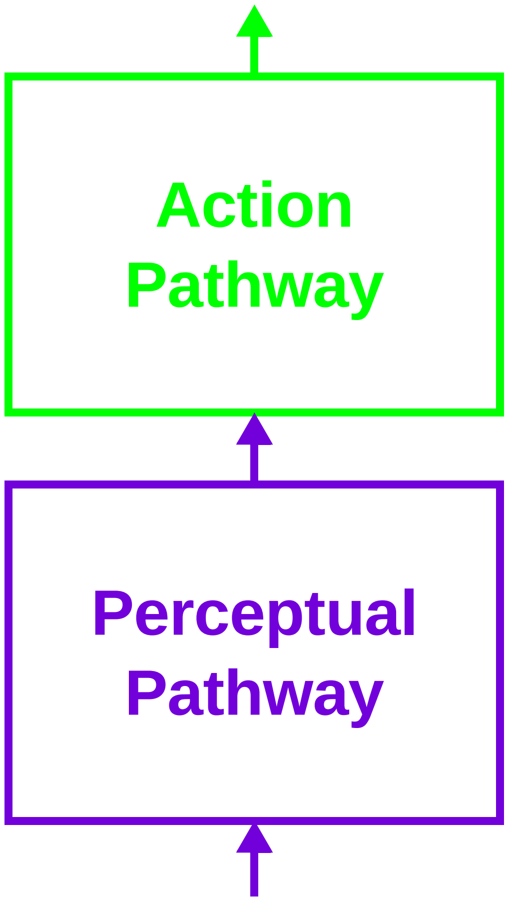
- **input-output associations, $p(x_t|y_t)$**
- **output transitions, $p(y_t|y_{t-1})$**
- **prior distribution, $p(y_0)$**

can compute y_t based on x_1 ... x_t:

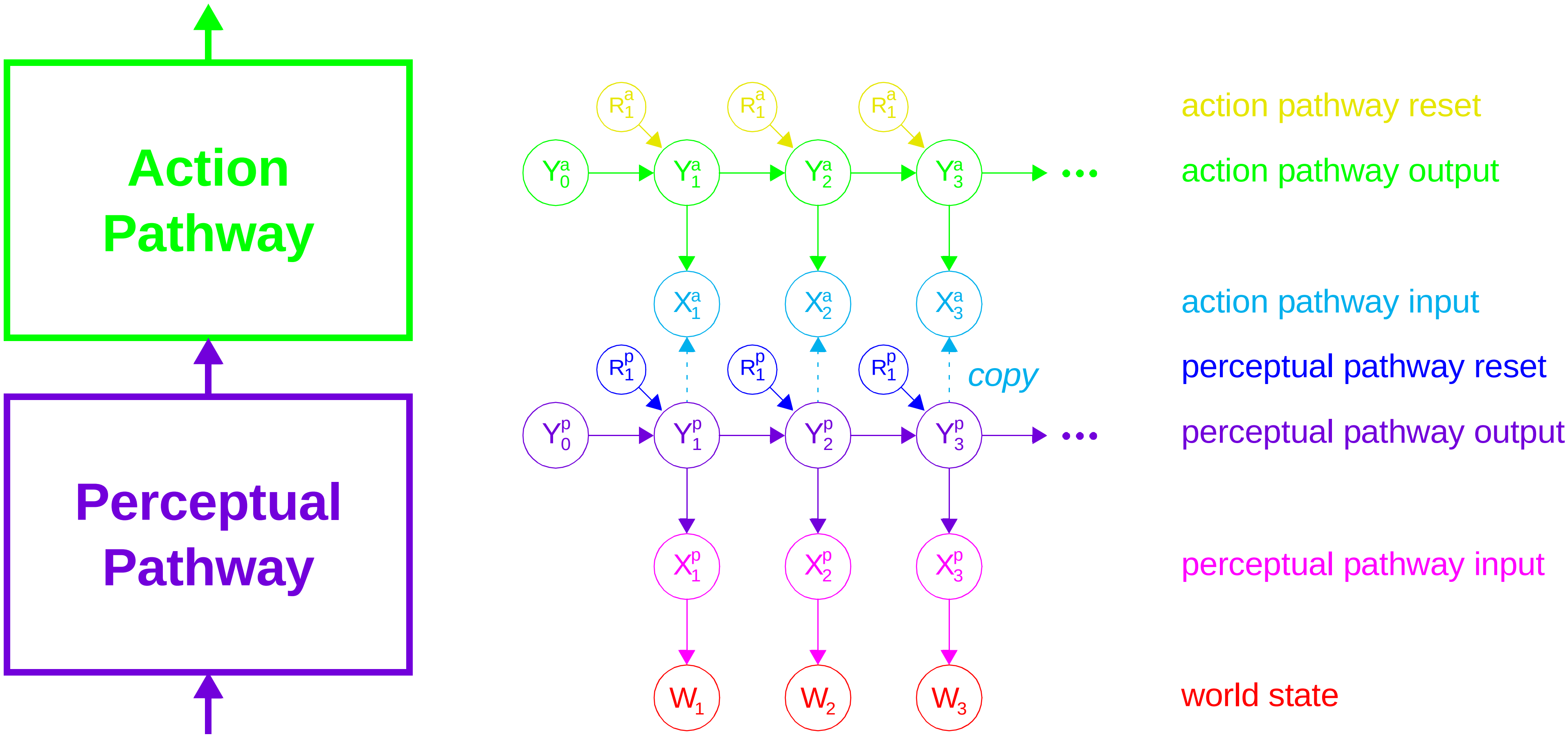
$$p(y_t = j|x_1, x_2, x_3, \ldots, x_{t-1}, x_t) \sim$$

$$\left[\sum_{k=1}^{N_y} p(y_{t-1} = k)p(y_t = j|y_{t-1} = k)\right]\left[\sum_{i=1}^{N_x} p(x_t = i)p(x_t = i|y_t = j)\right]$$

Multiple Pathways



Multiple Pathways

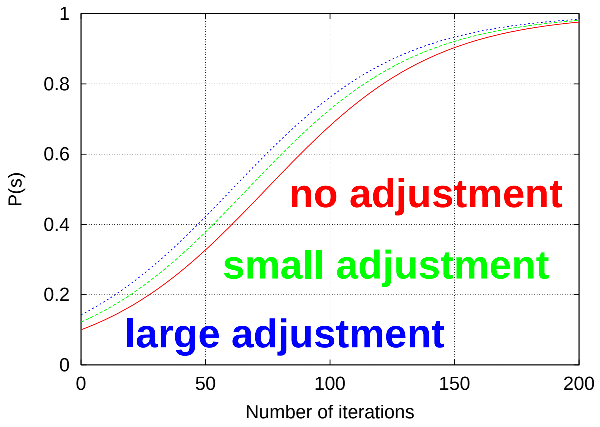


Decoupled inference due to modularity assumption

Learning Mechanisms for Skill Acquisition

Experience with input i leading to output j could cause:

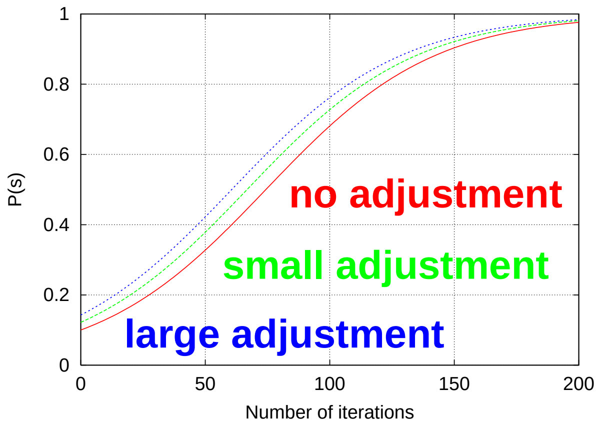
revising priors, $y_j(0)$
(model of environment)



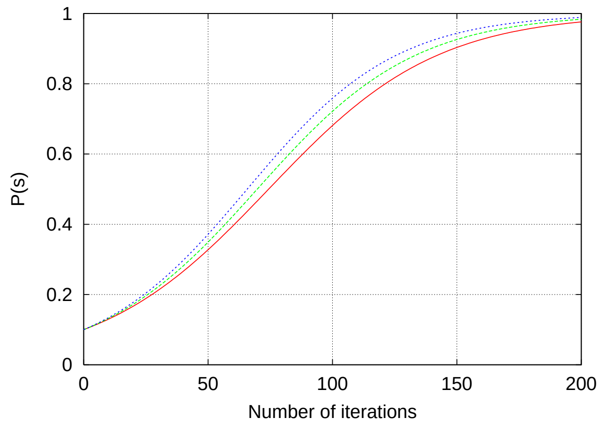
Learning Mechanisms for Skill Acquisition

Experience with input i leading to output j could cause:

revising priors, $y_j(0)$
(model of environment)



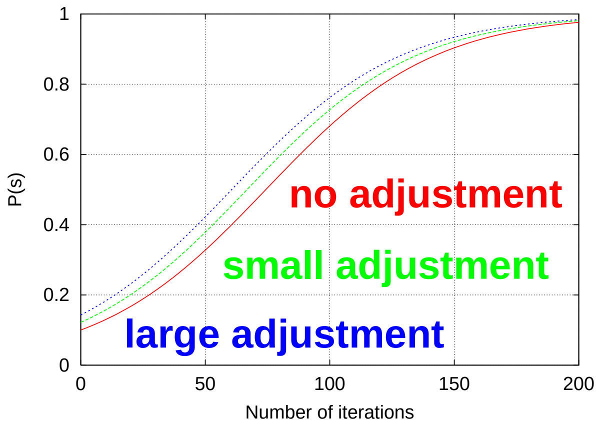
more efficient
signal transmission, α_{ij}

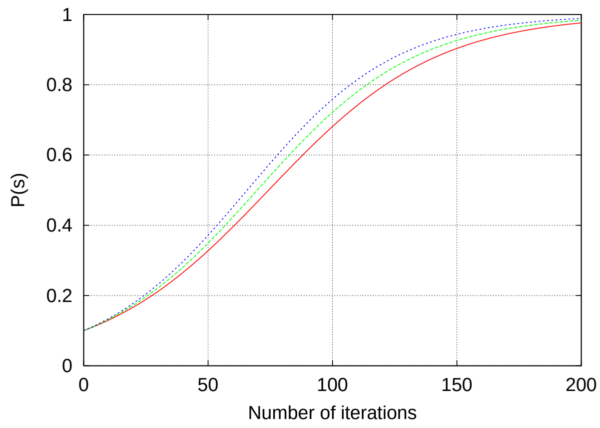


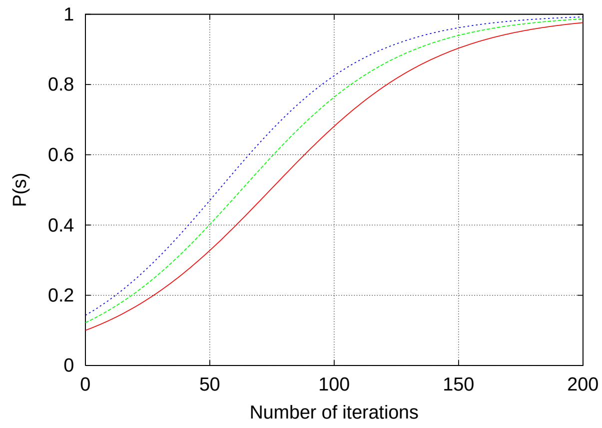
Learning Mechanisms for Skill Acquisition

Experience with input i leading to output j could cause:

revising priors, $y_j(0)$ (model of environment)	more efficient signal transmission, α_{ij}	both







Two distinct mechanisms, possibly differing in persistence.

Both shift the speed-accuracy curve with experience.

Learning Mechanisms for Skill Acquisition

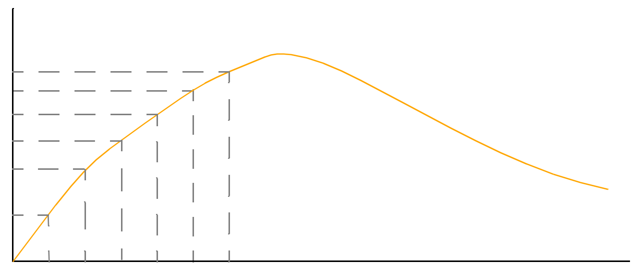
Efficiency of signal transmission

$$p(x_t = i \mid y_t = j) \sim \varepsilon + \alpha_{ij}$$

Hebbian temporal-difference rule:

$$\Delta\alpha_{ij} = (1-\kappa)\alpha_{ij} + \eta \max(0, p(x_t = i)p(y_t = j) - p(x_{t-1} = i)p(y_{t-1} = j))$$

κ	decay
η	learning rate



Priors (model of environment)

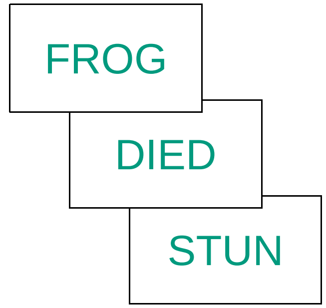
$$p(y_0 = j) \sim \hat{\varepsilon} + \hat{\alpha}_j$$

Temporal-difference rule:

$$\Delta\hat{\alpha}_j = (1-\hat{\kappa})\hat{\alpha}_j + \hat{\eta} \max(0, p(y_t = j) - p(y_{t-1} = j))$$

Ratcliff and McKoon (1997) Experiment 3

Phase 1: Study list

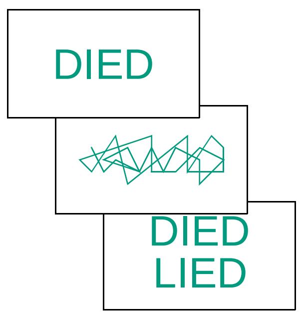


Phase 2: perceptual identification test

target (variable duration)

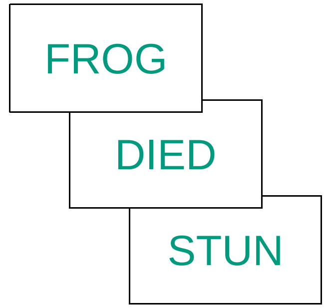
mask

two alternative
forced choice



Ratcliff and McKoon (1997) Experiment 3

Phase 1: Study list

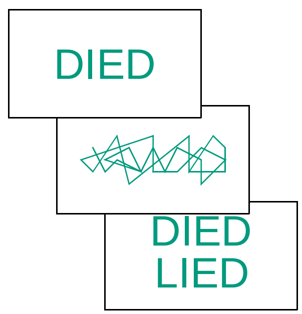


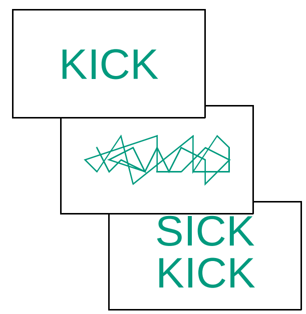
Phase 2: perceptual identification test

target (variable duration)

mask

two alternative
forced choice

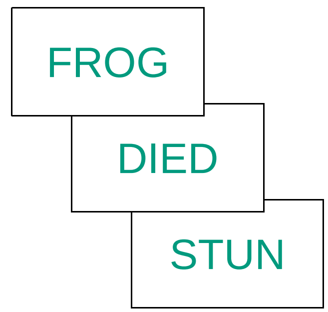




neutral
condition

Ratcliff and McKoon (1997) Experiment 3

Phase 1: Study list

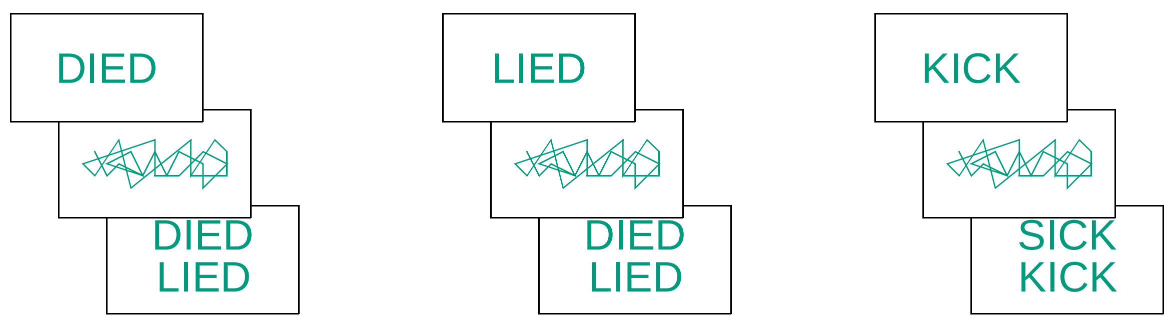


Phase 2: perceptual identification test

target (variable duration)

mask

two alternative
forced choice



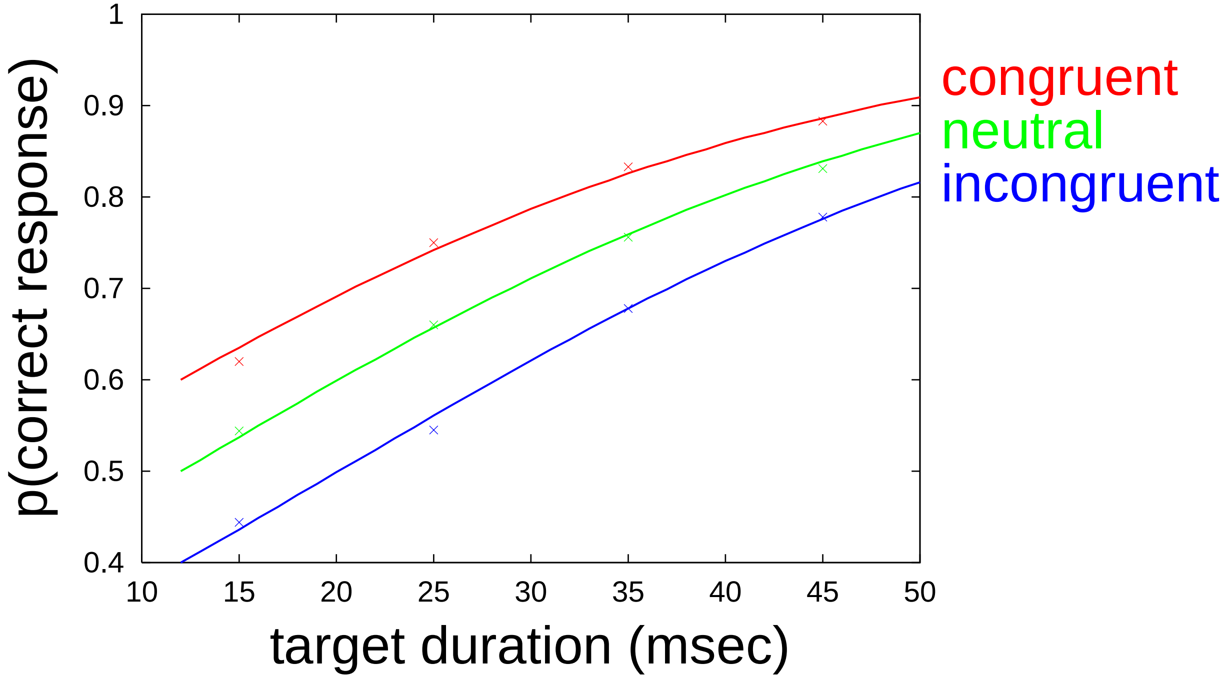
congruent
condition

incongruent
condition

neutral
condition

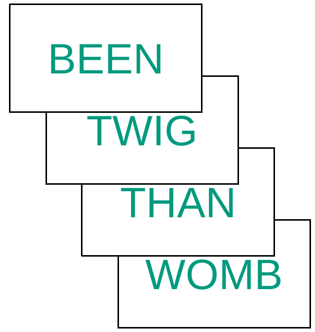
Ratcliff & McKoon (1997) Experiment 3

Results

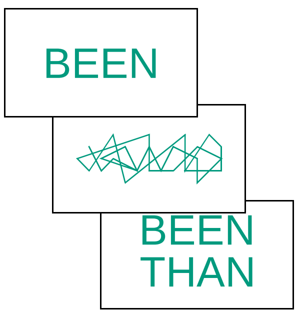


Bowers (1999); McKoon and Ratcliffe (2001)

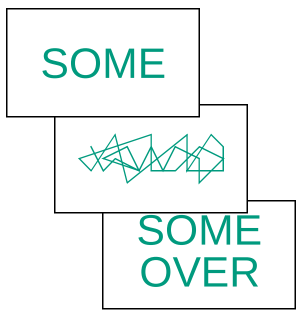
Study phase



Perceptual identification task



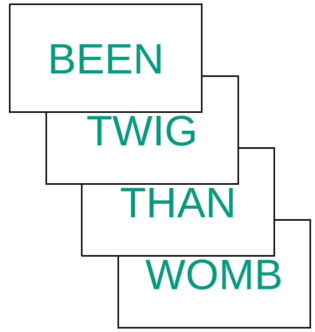
both studied



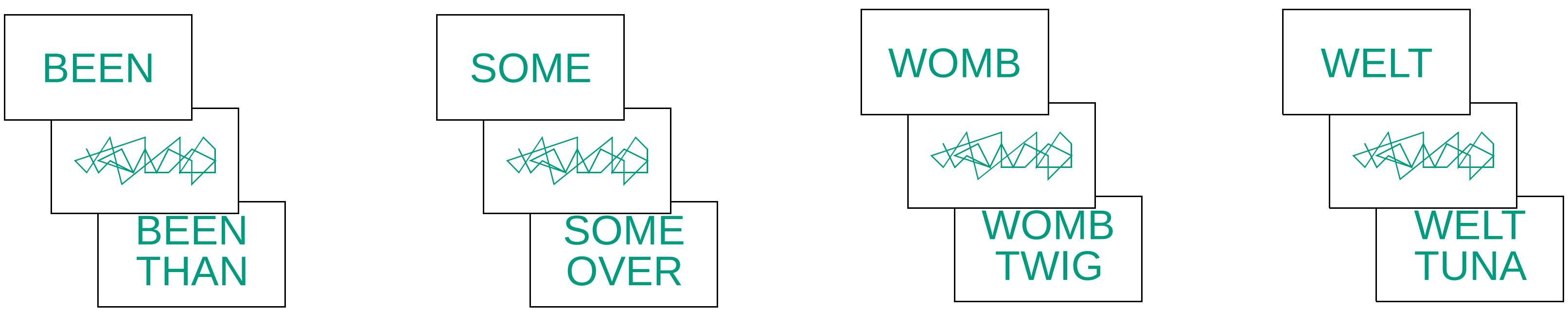
neither studied

Bowers (1999); McKoon and Ratcliffe (2001)

Study phase



Perceptual identification task

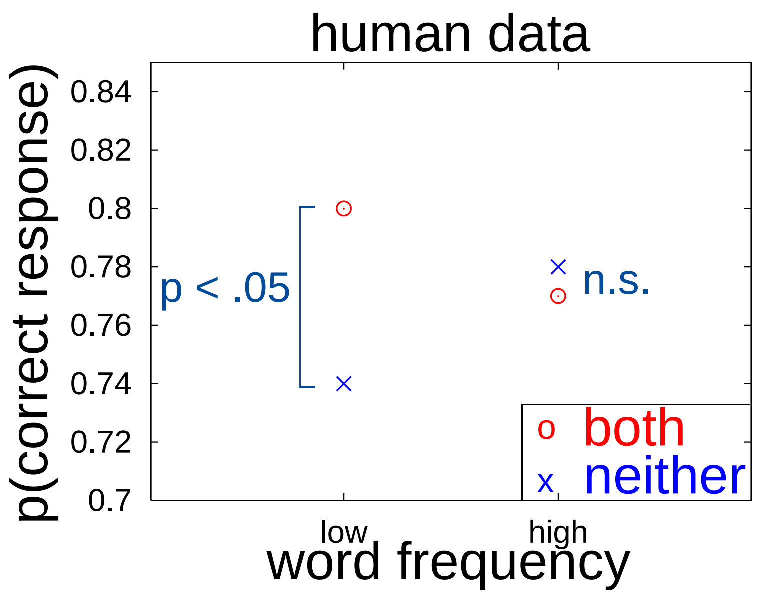


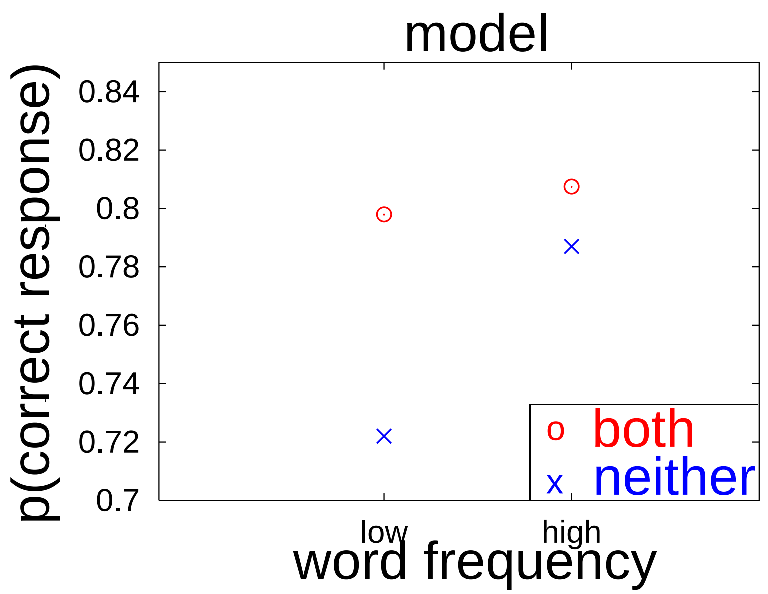
both studied   neither studied   both studied   neither studied

high frequency   low frequency

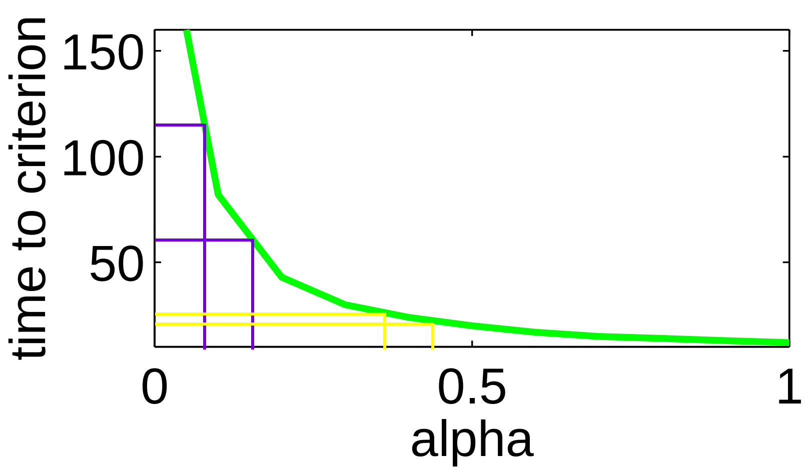
Simulation

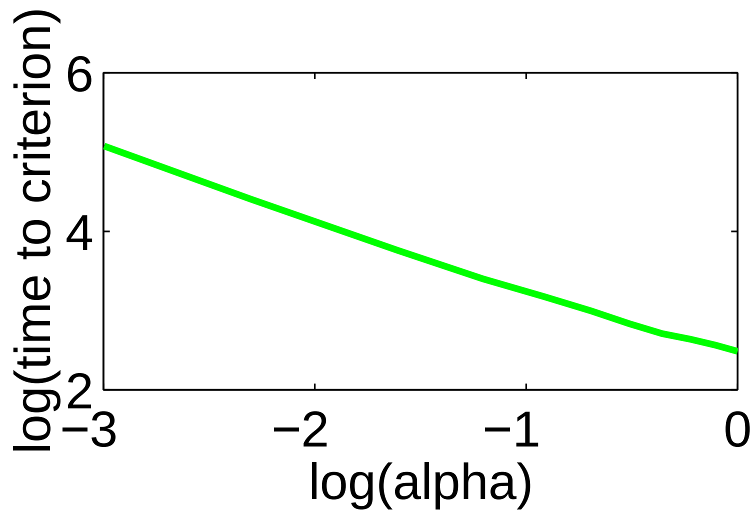
Two free parameters: $\alpha_{\text{low-freq}}$, $\Delta\alpha$.





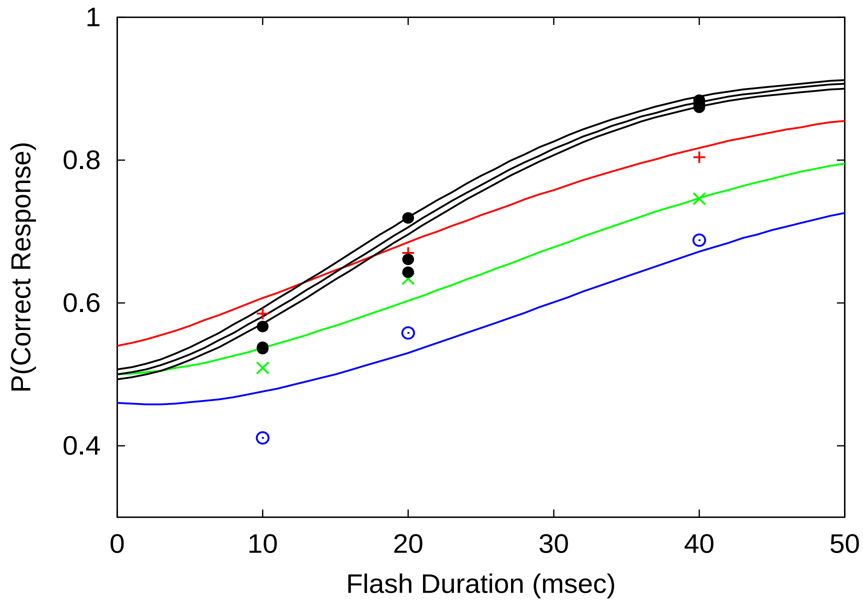
Explanation



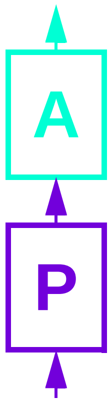


Ratcliff and McKoon (1997) Experiment 4

Human and simulation results



Simulation details



two pathways—*perception* and *action*—with 1-1 mappings

same similarity structure, but lower similarity coefficient, γ, in action pathway

Diffusion, β, necessary in perceptual pathway

inclusion of action pathway allowed for elimination of arbitrary lag term, k

Ratcliff and McKoon (1997) Experiments 6 and 7

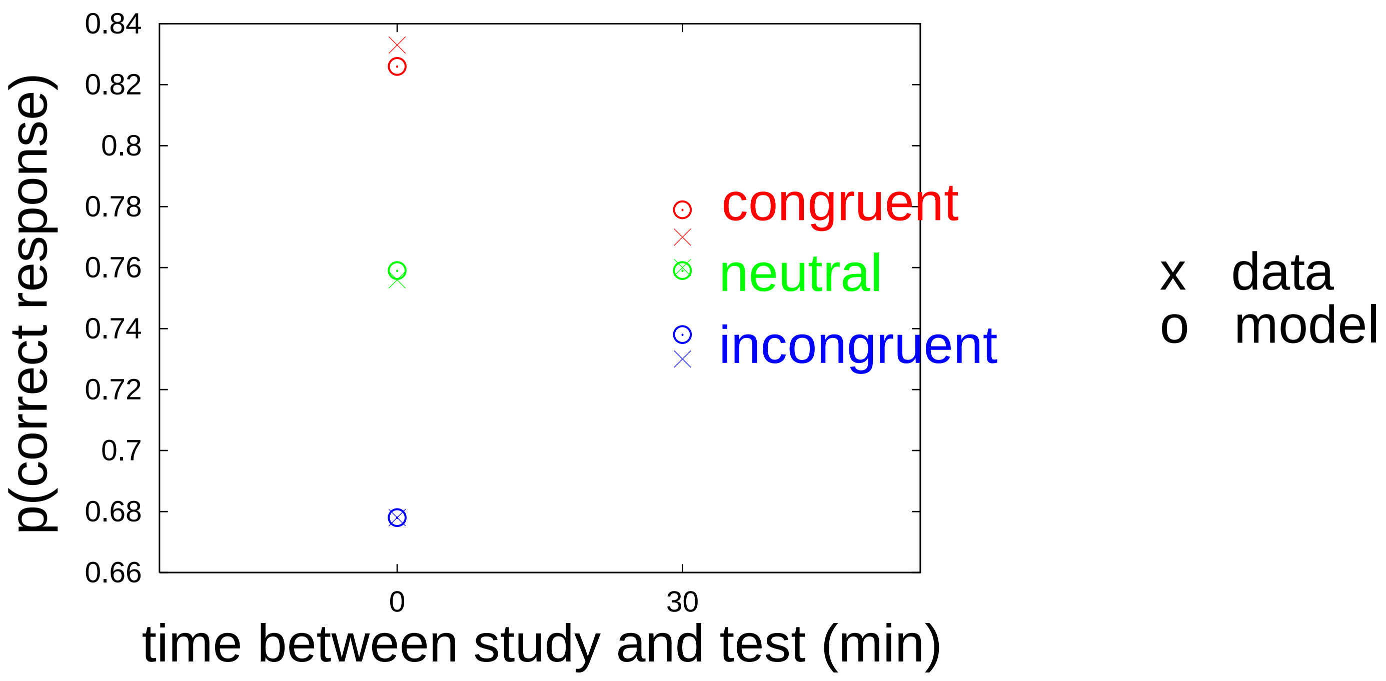
Task

Same as previous study, but vary time between study and test

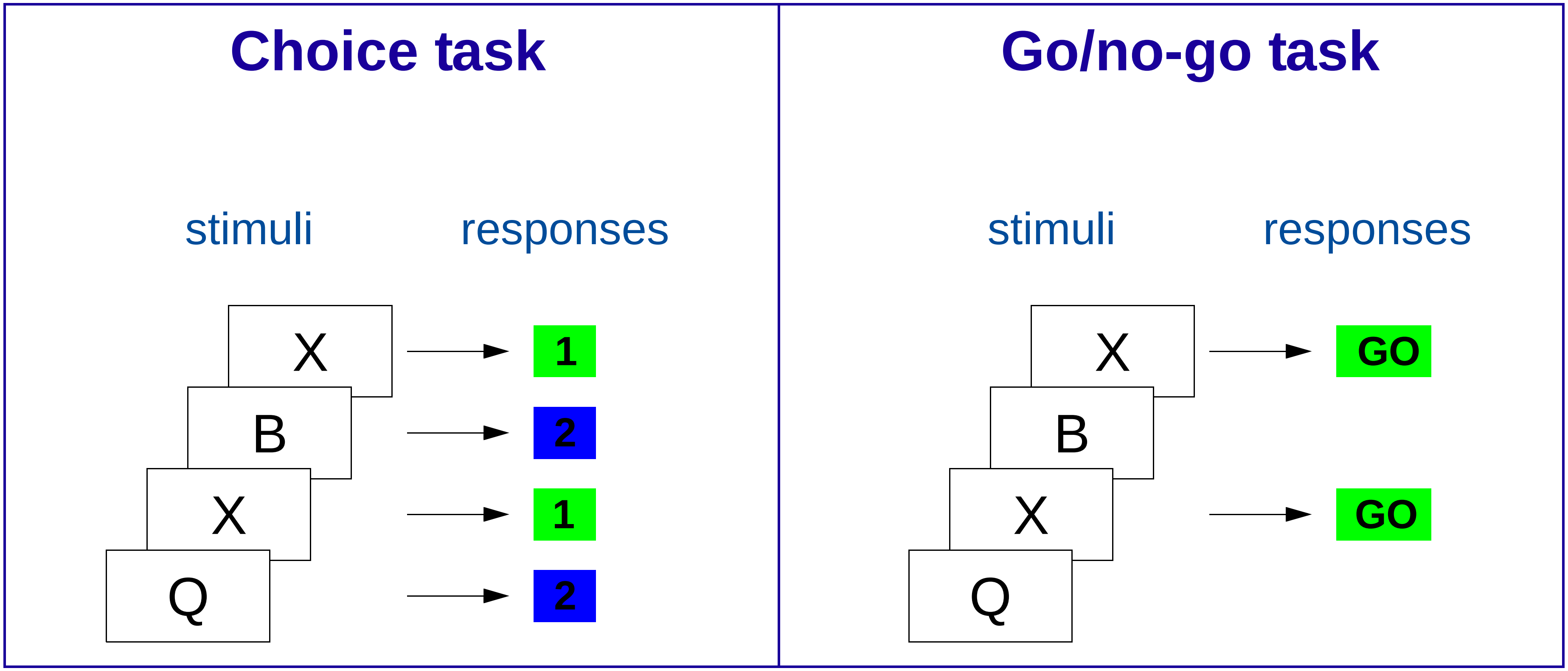
Simulation

One free parameter: ρ at 30 min

Results



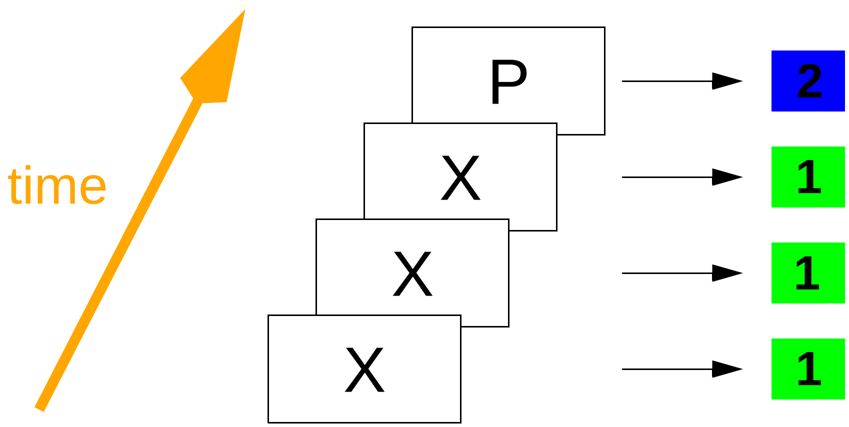
Speeded Discrimination
(Jones, Braver, Cho, Cohen, & Nystrom, 2001)



Response priming introduced by manipulating the relative frequency of X trials: 17%, 50%, or 83% of trials

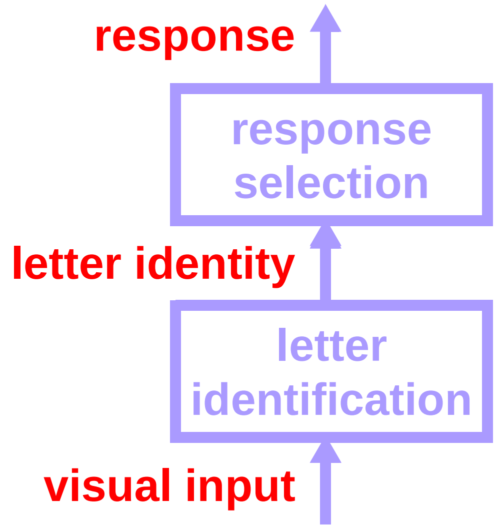
Response conflict

Response conflict arises when two response classes are not balanced in frequency and stimulus invoking low-frequency response is presented.

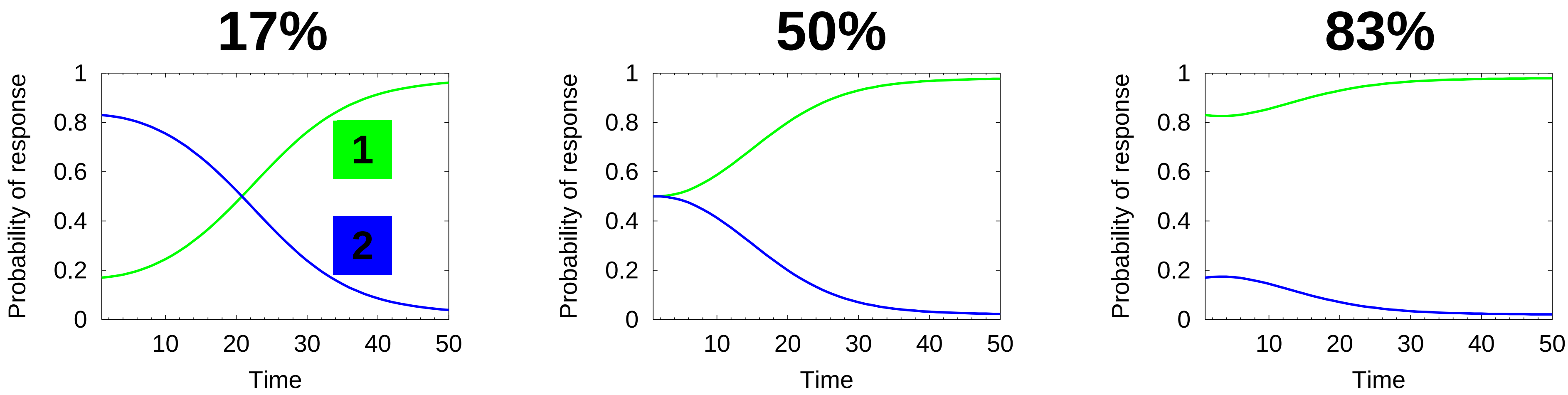


Subjects lower and less accurate on trials invoking low-frequency response.

Modeling the Speeded Discrimination Task



Output of response pathway when X (correct response 1) is presented, for different relative frequencies of X



A Rational View of Cognitive Control

The cognitive system estimates a *response cost*, and initiates a response at the point in time when the cost is minimized.

The cost penalizes inaccurate and slow responses.

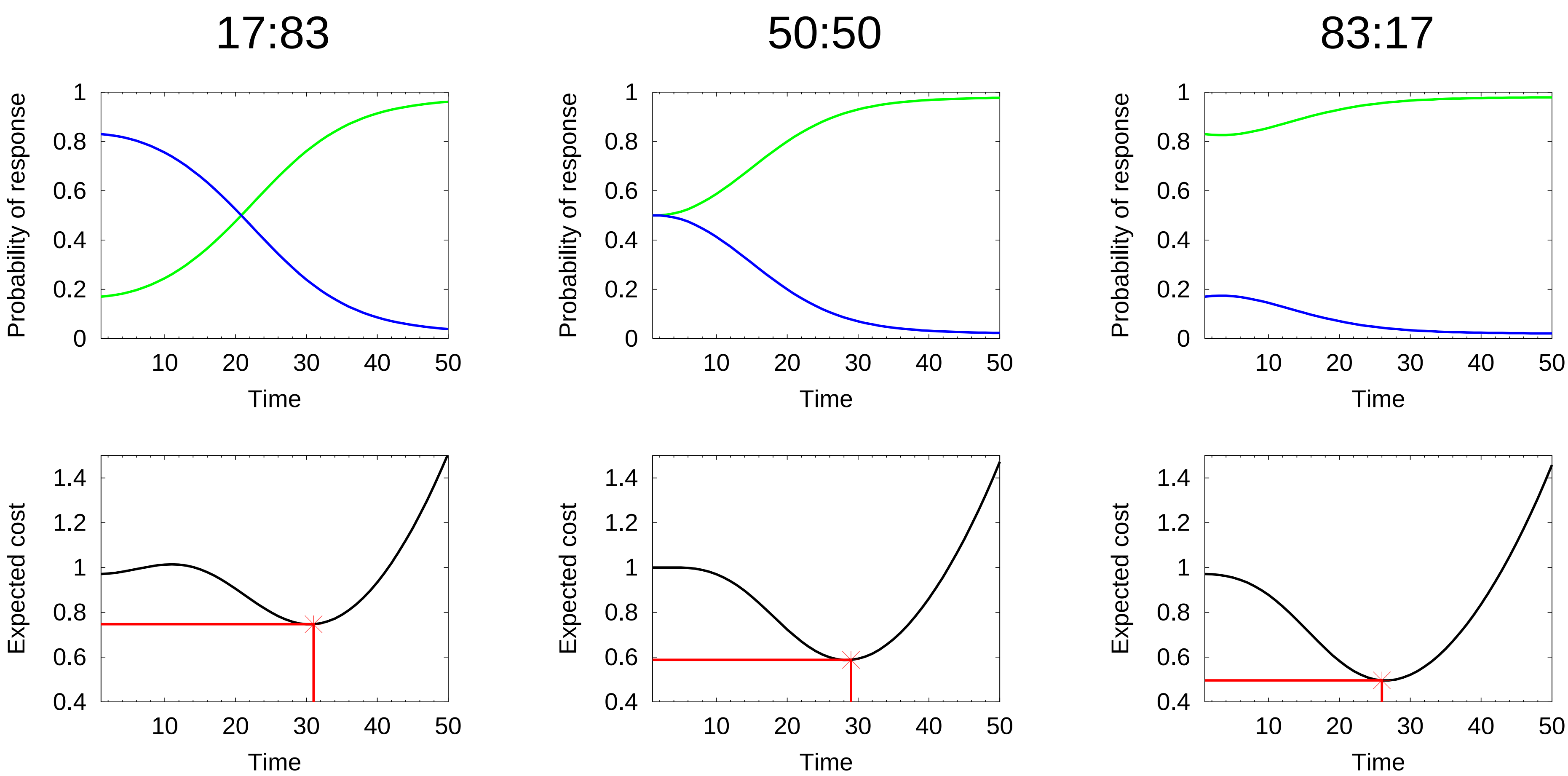
$$\text{cost of response at time t} \;=\; \text{error rate} + \text{constant} \times \text{reaction time}$$

$$c(t|R,S) \;=\; P(Y(t) \neq R|S) \;+\; \kappa_R \; t$$

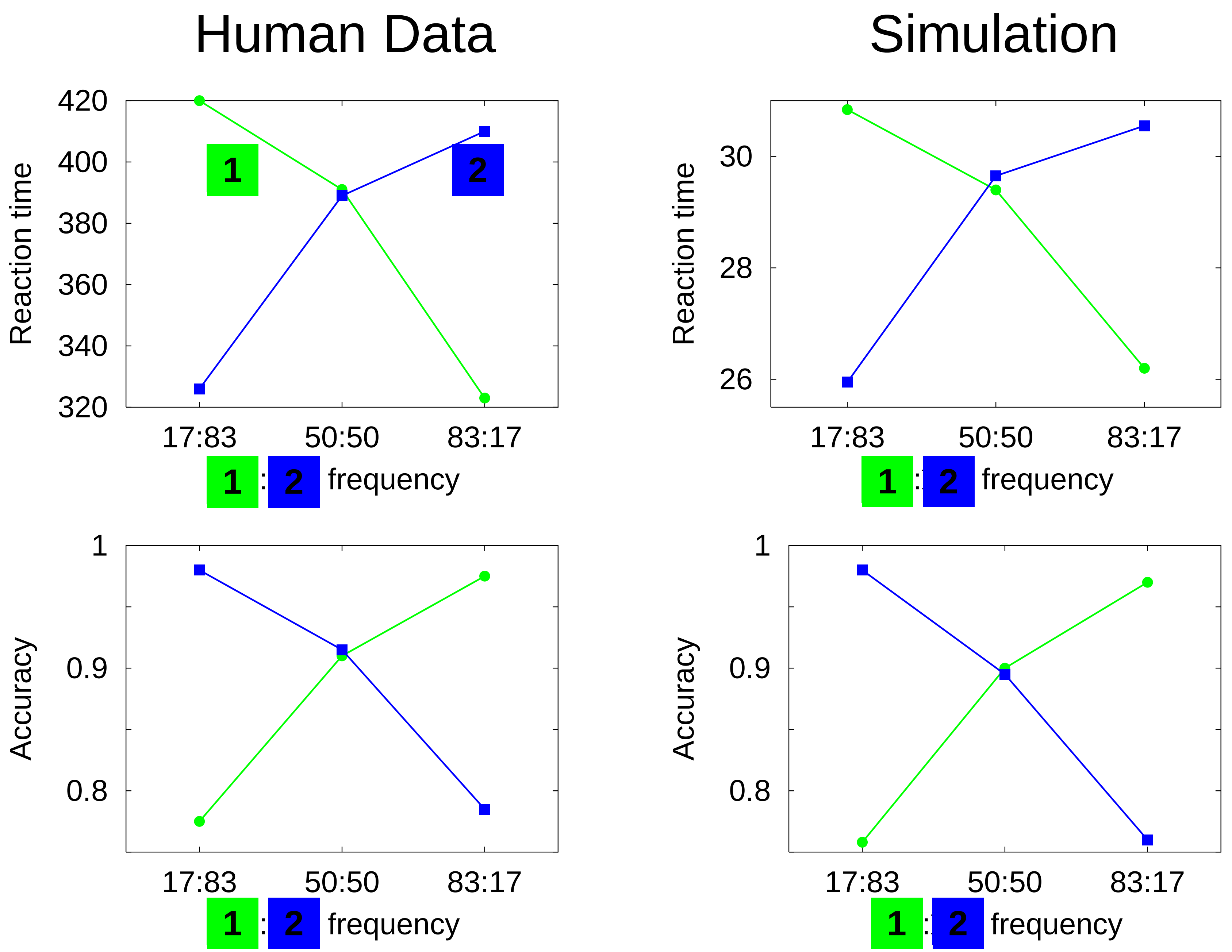
Because the system cannot know the correct response, R, the *expected* cost over all responses is computed instead.

$$\langle c(t|S) \rangle_{\{R\}} \;=\; \sum_{r \in \{R\}} P(Y(t) = r|S) \; c(t|R,S)$$

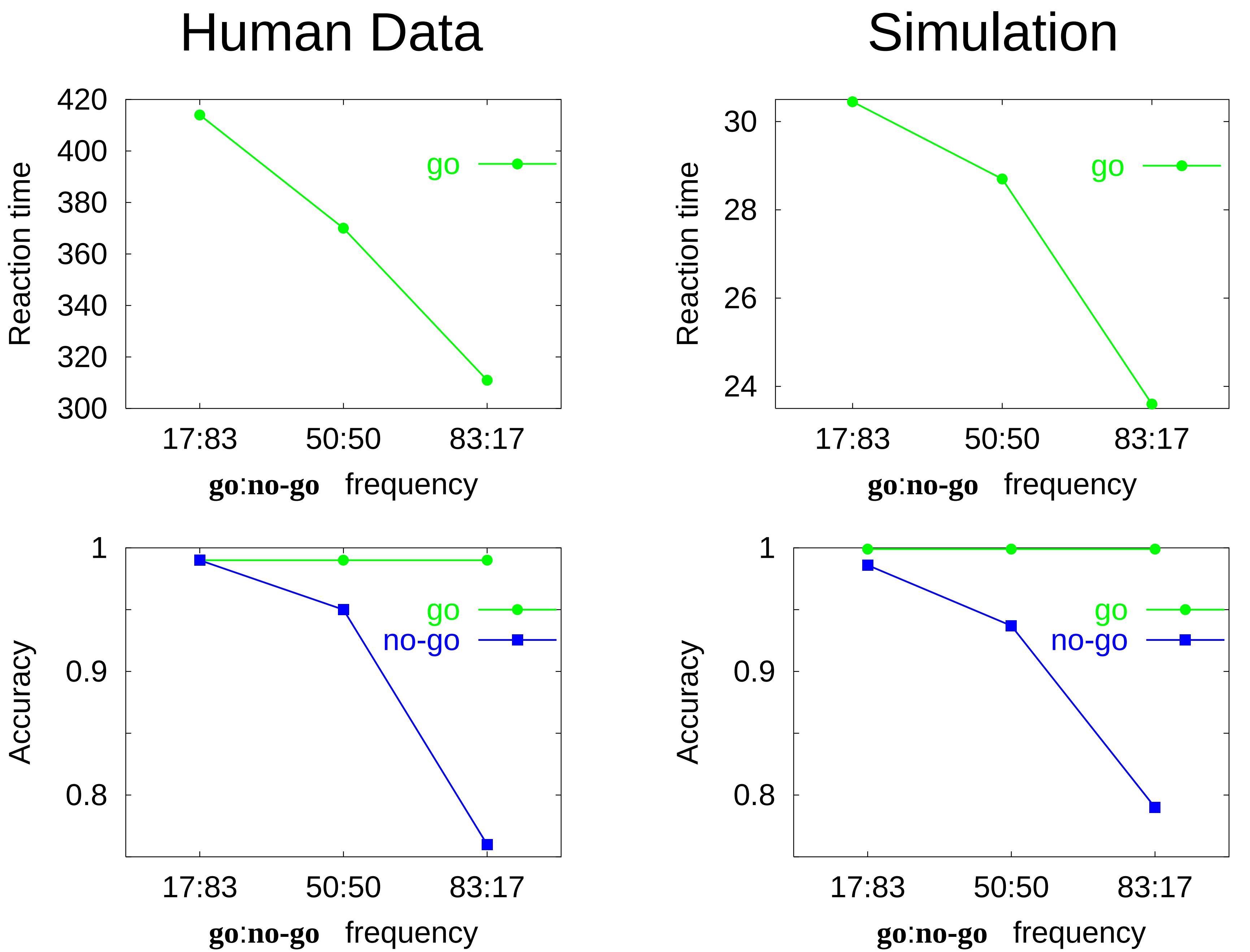
Minimizing the Expected Cost



Results: Choice Task

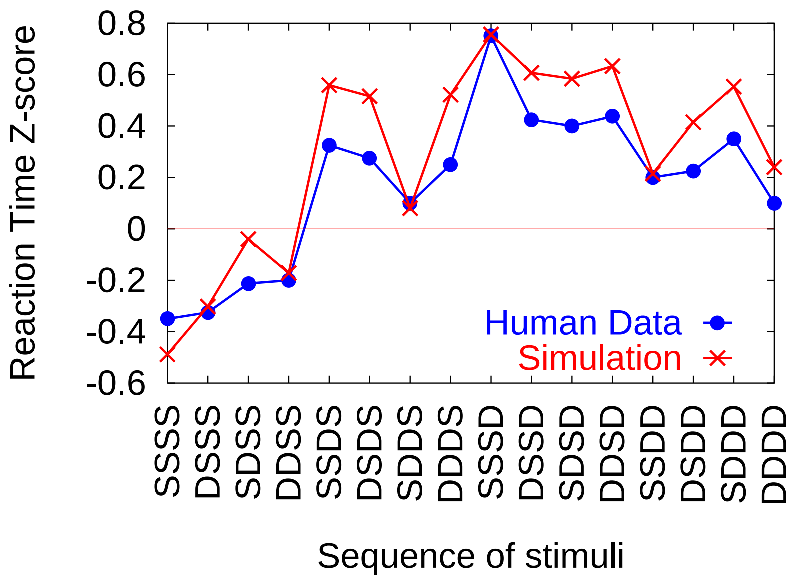


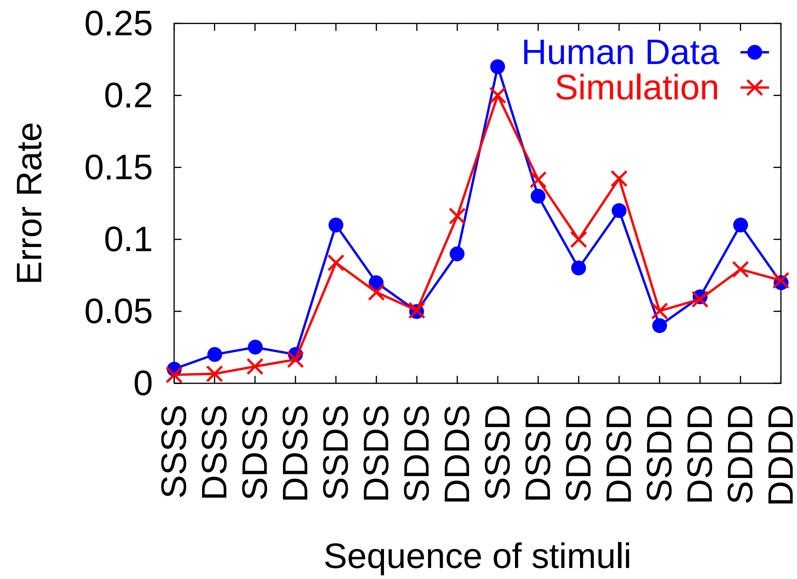
Results: Go/No-Go Task



Trial by Trial Results for Choice Task

Performance on each trial, conditional on the stimulus sequence leading up to the trial





Observations

Fit between human and simulation data is very good (ρ_{RT} = .96, ρ_{err} = .95).

Most variability in performance can be explained by the correct responses for the three previous trials.

Data Explained by Architecture

Priming results in a bias to report primed words. (Ratcliff & McKoon, 1997, Experiment 3)

With low frequency words, priming also results in increased sensitivity to primed words. (Bowers, 1999; McKoon & Ratcliff, 2001)

Bias effects are negligible if response alternatives are dissimilar. (Ratcliff & McKoon, 1997, Experiment 4)

Greater effect of target presentation duration on accuracy if response alternatives are dissimilar. (Ratcliff & McKoon, 1997, Experiment 4)

Decay of bias effect when time between study and test is increased.
(Ratcliff & McKoon, 1997, Experiments 6 & 7)

Interaction between bias effect and response paradigm—naming, forced choice, and yes/no matching (Ratcliff & McKoon, 1997, Experiments 6 & 8; Wagenmakers et al., 2000)

Interaction between bias effect and lexicality (Steyvers et al., 2001)

Sensitivity effects in stem completion (Zeelenberg, et al., 2002)

Effects of trial sequence and response conflict in speeded discrimination (Jones et al., 2002)

Power law of practice

Hick-Hyman law of choice reaction time

Summary

Described pathway architecture based on three basic assumptions

- speed-accuracy trade off
- limited communication between pathways
- speed up with experience

and three rational assumptions

- Inference within a processing pathway is Bayes optimal.
- Priming mechanism accurately estimates prior probabilities of pathway output.
- The goal of the cognitive system is to minimize a cost that depends both on probability of error and reaction time.

Summary (continued)

Virtues of architecture

Architecture has small number of free parameters, each of which has a cognitive interpretation.

Processing dynamics fall out of Bayesian assumption.

Predictive power of the architecture is high.

Architecture formalizes notions such as *quality of representation*, *similarity*, *efficiency of signal transmission*, *systematicity* of a task.

Crisp theoretical account and broad coverage relative to alternative models

Elegant characterization of mechanisms of skill acquisition

Can characterize other phenomena such as working memory and awareness.

Key Design Principles from Mozer Lectures

Build multicomponent, integrated models

- Difficult to justify and understand because they are complex
- Allows a broad understanding of a variety of data from diverse paradigms

Key Design Principles from Mozer Lectures

Build multicomponent, integrated models

“Clean up” of representations at many levels of the cognitive system

- E.g., pull-out net, connectionist pathway, probabilistic pathway
- Restrict representation to the familiar or well-formed states based on domain knowledge
- A type of attentional focusing (competition)
- Necessary to avoid propagating noise to the next stage
- Transformation from a continuous to a discrete set of alternatives is key step of mapping subsymbolic to symbolic representations
  - “...Modeling human cognition requires at least two complementary formal descriptions: a discrete and a continuous one, reflecting two different basic properties—the ability to categorize and the dynamics of the cognitive processes...The symbolic approach is a discrete one while the connectionist approach is a continuous one...” (Kokinov, 1997)

Key Design Principles from Mozer Lectures

Build multicomponent, integrated models

“Clean up” of representations at many levels of the cognitive system

Cognition is optimized with respect to task-related goals and the statistical structure of the environment.

- rational perspective (Anderson, 1990)
- subject to hardware and knowledge limitations
- e.g., visual recognition system is an elegant engineering solution to optimize the trade off between hardware requirements and parallelism
- e.g., perceptual inference is optimal (computes conditional probabilities) given current state of knowledge

Key Design Principles from Mozer Lectures

Build multicomponent, integrated models

“Clean up” of representations at many levels of the cognitive system

Cognition is optimized with respect to task-related goals and the statistical structure of the environment.

Intelligent information processing can emerge from simple, heuristic, approximate processing mechanisms.

- e.g., attentional selection

Key Design Principles from Mozer Lectures

Build multicomponent, integrated models

“Clean up” of representations at many levels of the cognitive system

Cognition is optimized with respect to task-related goals and the statistical structure of the environment.

Intelligent information processing can emerge from simple, heuristic, approximate processing mechanisms.

Continual and rapid adaptation to structure of environment