Locally Bayesian Learning

John K. Kruschke
Indiana University

Bayesian Prediction & Estimation

$$p(y \mid x, \theta)$$

Hypothesized models, parameterized by θ, map each x value to a probability distribution over y values.

Bayesian Prediction & Estimation

$$p(y \mid x, \theta)$$

$$p(\theta)$$

There is a distribution of probabilities regarding values of θ.

Bayesian Prediction & Estimation

$p(y \mid x, \theta)$

$p(\theta)$

For a given x, we predict y by marginalizing over parameter values.

$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta)\, d\theta$$

$$\text{For SSE loss, } \hat{y} = \int y \; p(y \mid x)\, dy$$

Bayesian Prediction & Estimation

$$p(y \mid x, \theta)$$

$$p(\theta)$$

For a given x,y pair, we estimate parameters by Bayes' rule:

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)\, p(\theta)}{\int p(y \mid x, \theta)\, p(\theta) d\theta}$$

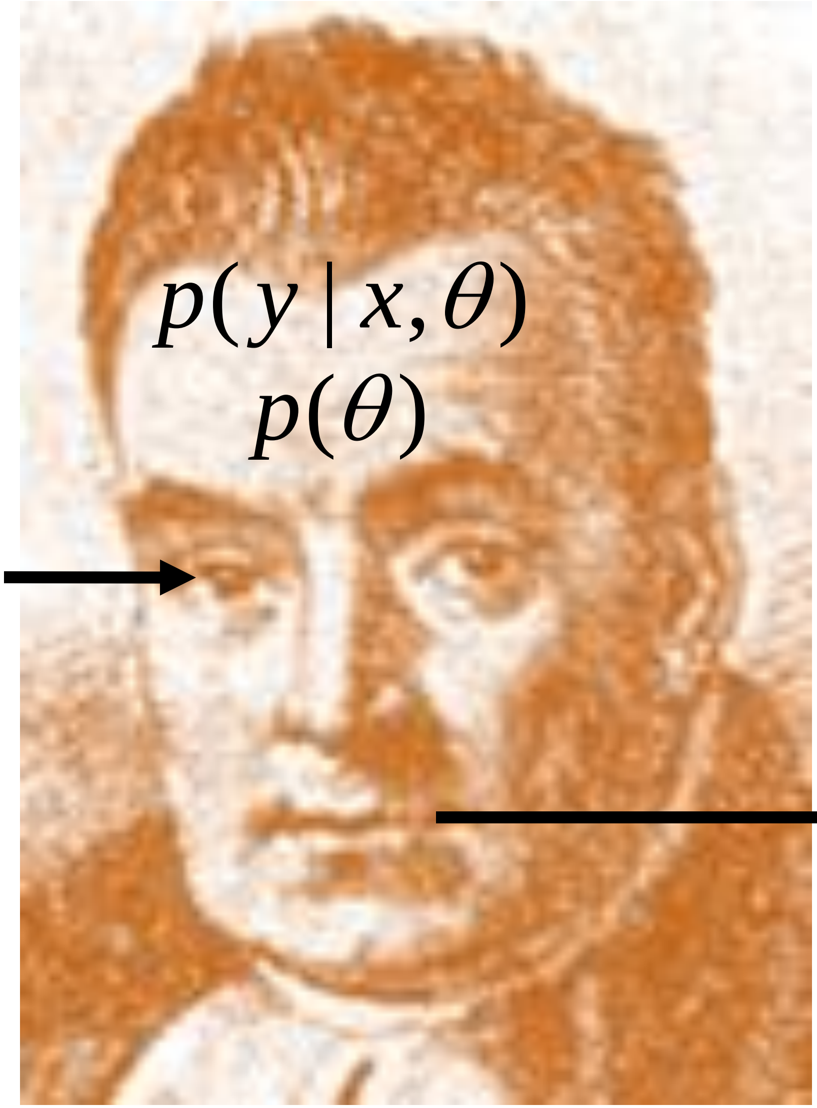
Bayesian Prediction & Estimation

$$p(y \mid x, \theta)$$

$$p(\theta)$$

Formalism doesn't care what it refers to in the world. Suppose that
x is a stimulus,
y is a response, and
θ is a hypothesis.

Bayesian Prediction



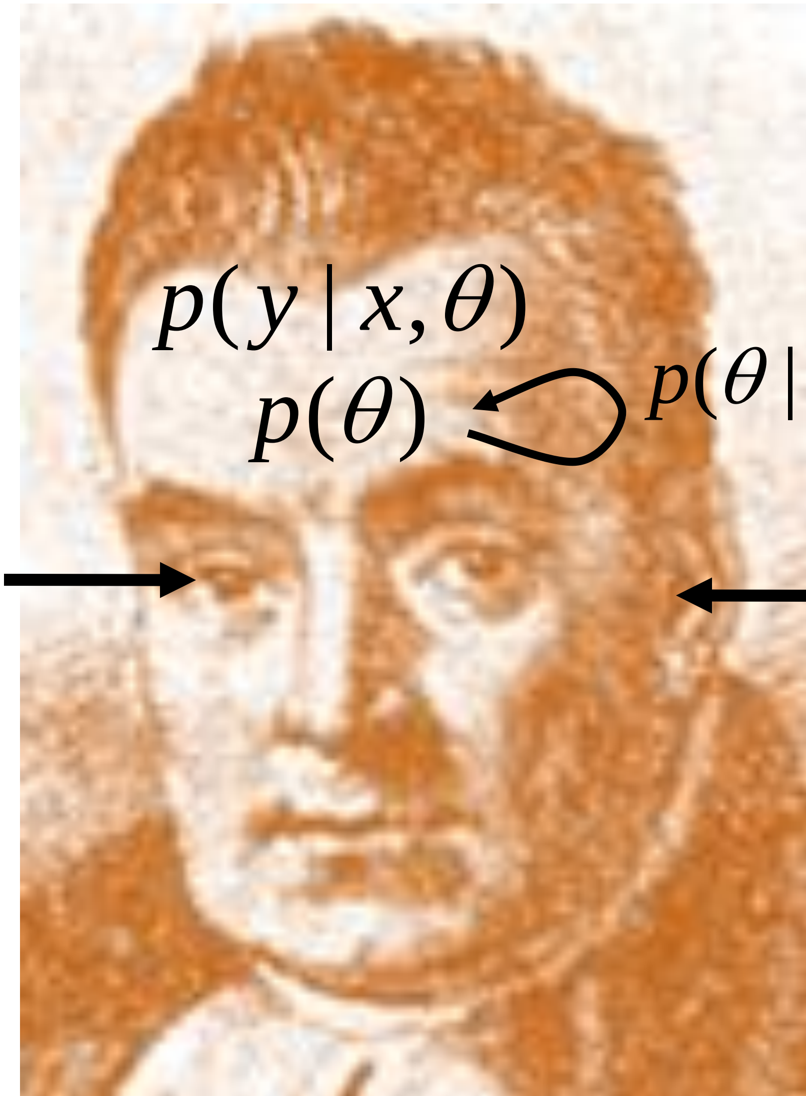
x

Then θ, $p(\theta)$, and $p(y|x,\theta)$ are in (or refer to) the mind.

$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta)\, d\theta$$

$$\hat{y} = \int y\; p(y \mid x)\, dy$$

Bayesian Estimation = Learning



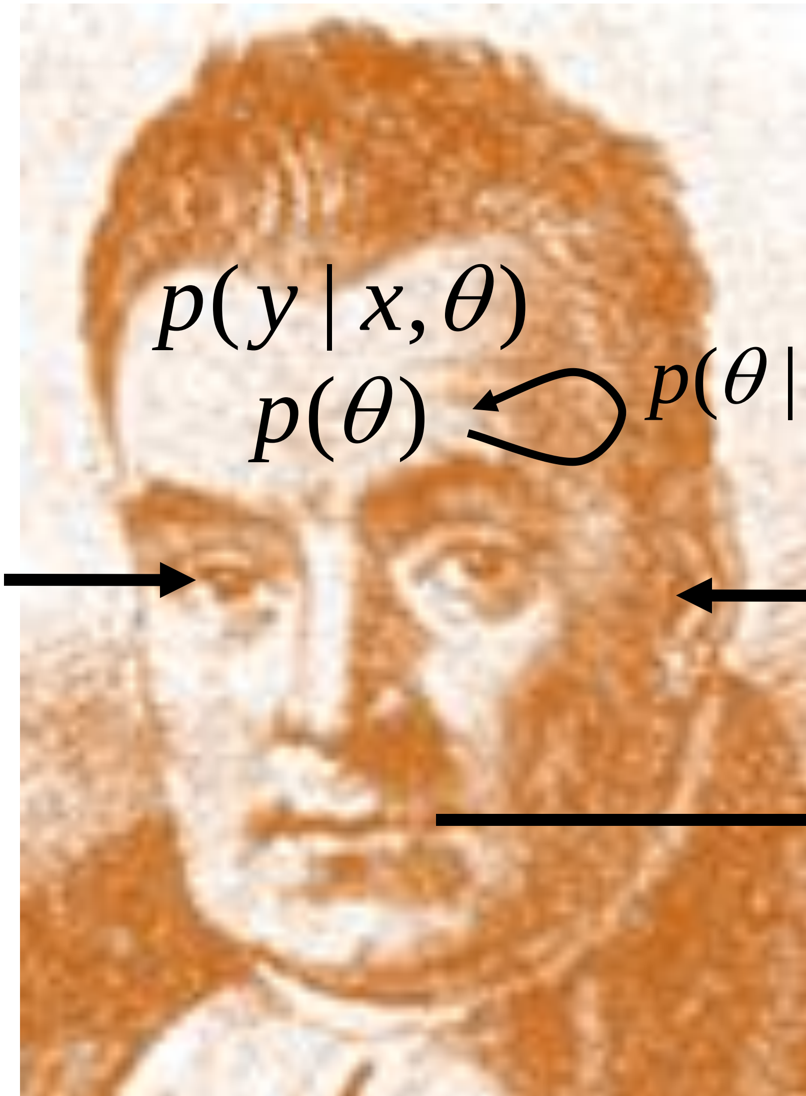
$$p(y \mid x, \theta)$$

$$p(\theta)$$

x → ← y

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)\, p(\theta)}{\int p(y \mid x, \theta)\, p(\theta)\, d\theta}$$

Bayesian Cognition



$$p(y \mid x, \theta)$$

$$p(\theta)$$

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)\, p(\theta)}{\int p(y \mid x, \theta)\, p(\theta)\, d\theta}$$

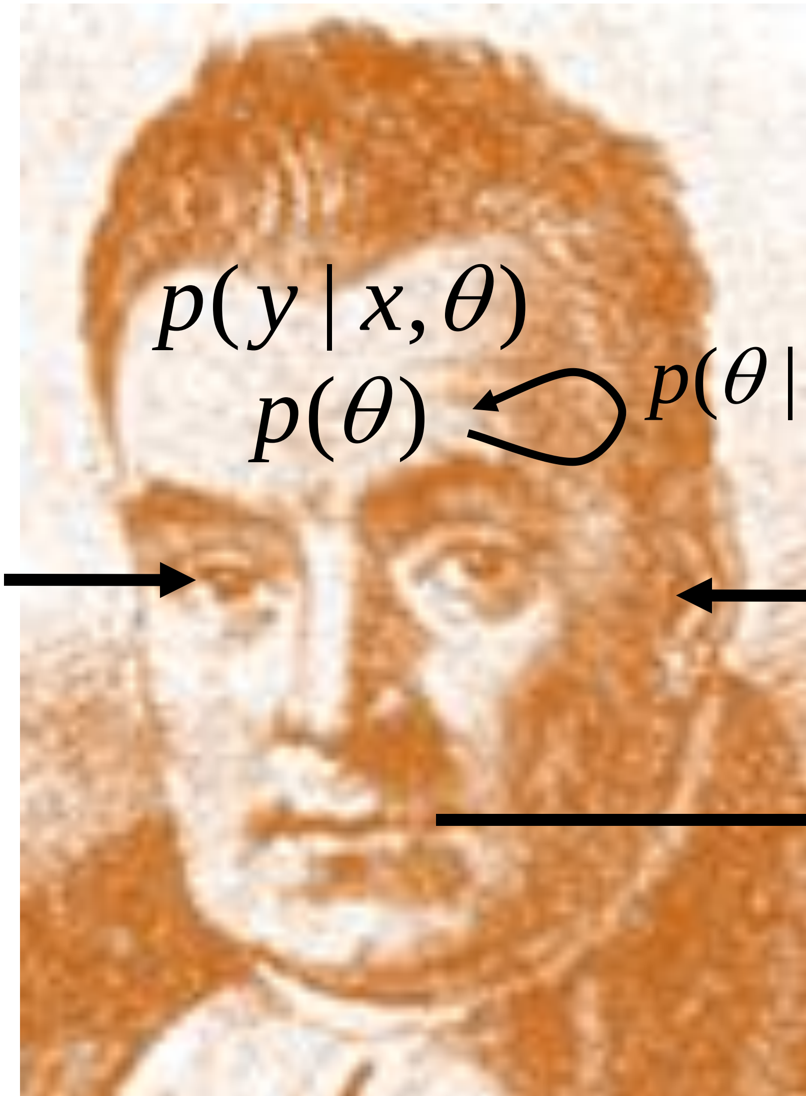
x

y

$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta)\, d\theta$$

$$\hat{y} = \int y \; p(y \mid x)\, dy$$

Not only cognition *by* Bayes...



$$p(y \mid x, \theta)$$

$$p(\theta)$$

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)\, p(\theta)}{\int p(y \mid x, \theta)\, p(\theta)\, d\theta}$$

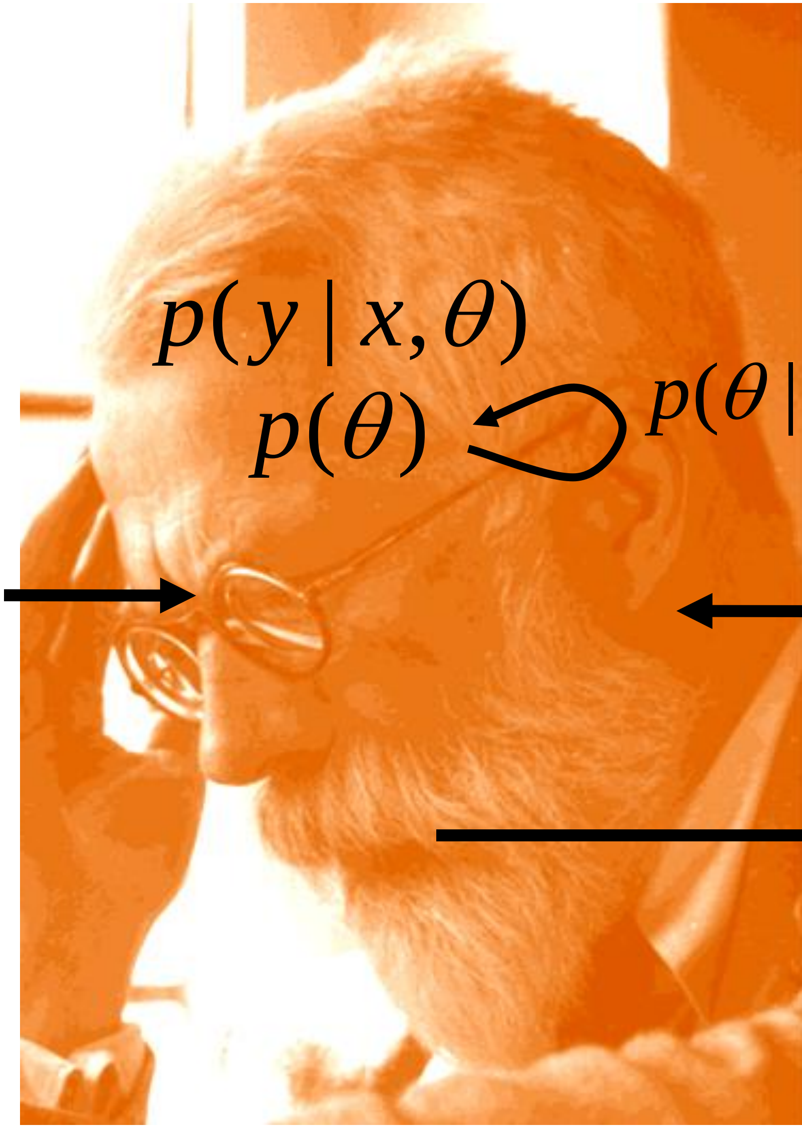
x

y

$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta)\, d\theta$$

$$\hat{y} = \int y \; p(y \mid x)\, dy$$

Bayesian cognition by others, too



$p(y \mid x, \theta)$

$p(\theta)$

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)\, p(\theta)}{\int p(y \mid x, \theta)\, p(\theta)\, d\theta}$$

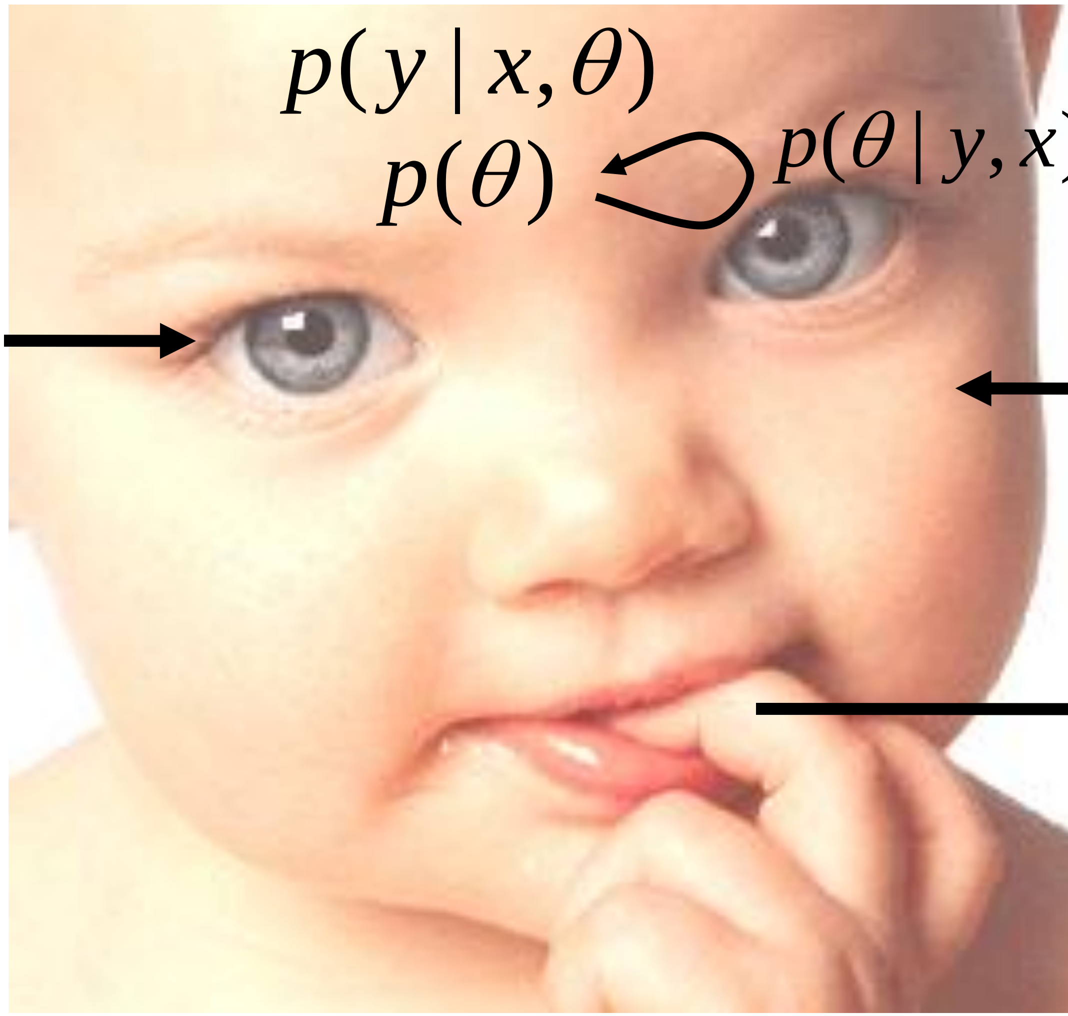
x

y

$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta)\, d\theta$$

$$\hat{y} = \int y \; p(y \mid x)\, dy$$

Bayesian Cognition?



$$p(y \mid x, \theta)$$

$$p(\theta)$$

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)\, p(\theta)}{\int p(y \mid x, \theta)\, p(\theta) d\theta}$$

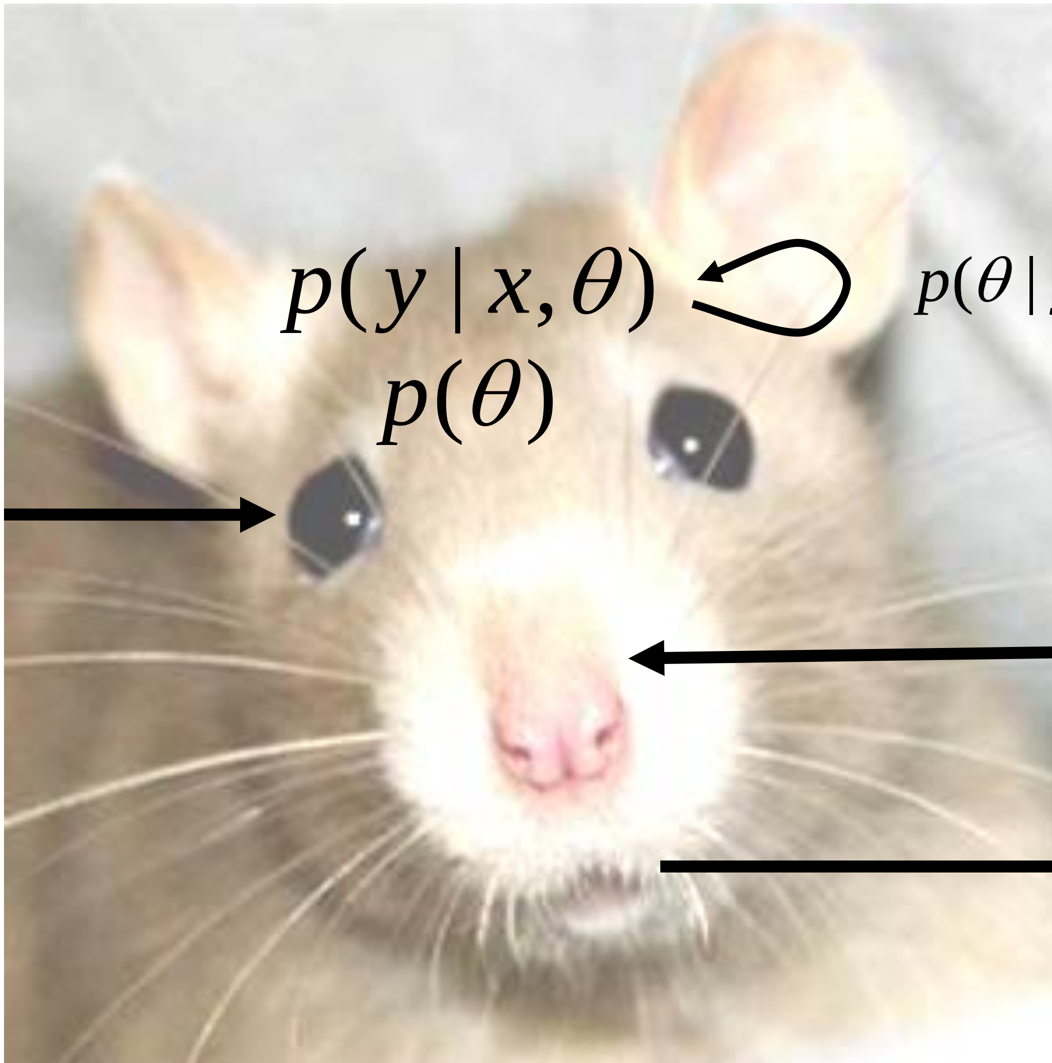
x

y

$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta) d\theta$$

$$\hat{y} = \int y \; p(y \mid x) dy$$

Bayesian Cognition?



$$p(y \mid x, \theta)$$

$$p(\theta)$$

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)\, p(\theta)}{\int p(y \mid x, \theta)\, p(\theta)\, d\theta}$$

x

y

$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta)\, d\theta$$

$$\hat{y} = \int y \; p(y \mid x)\, dy$$

Bayesian Cognition?

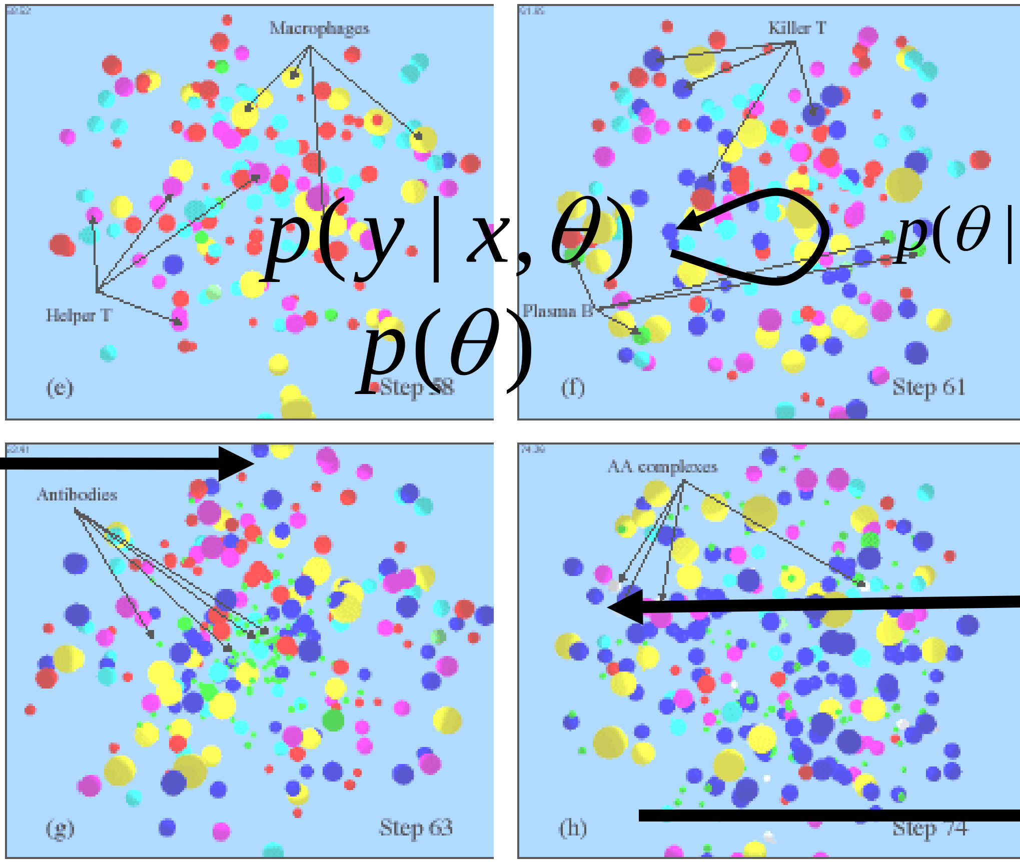


Image from Jacob, Litorco & Lee (2004)

$p(y \mid x, \theta)$

$p(\theta)$

x

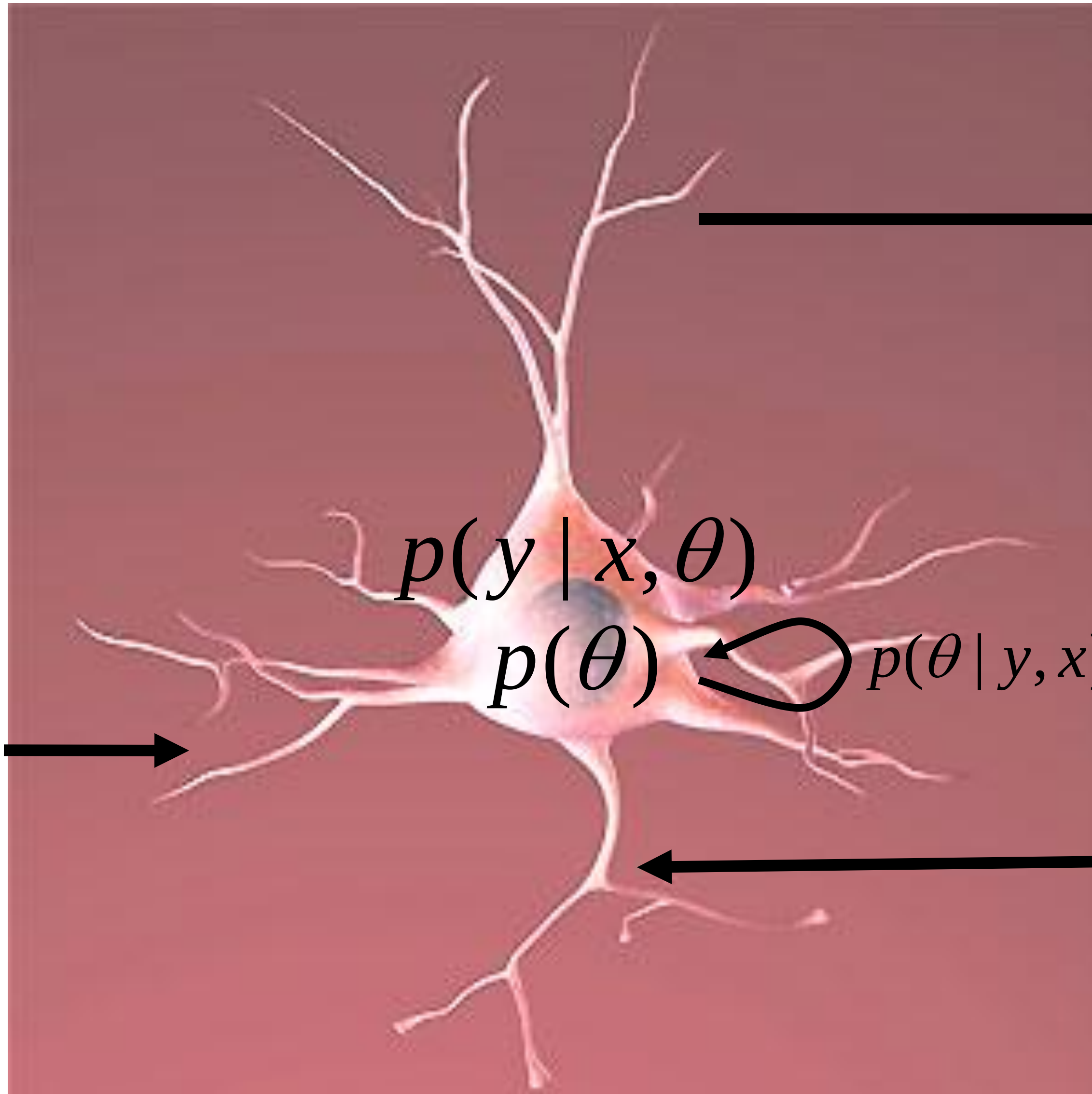
y

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)\, p(\theta)}{\int p(y \mid x, \theta)\, p(\theta)\, d\theta}$$

$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta)\, d\theta$$

$$\hat{y} = \int y \; p(y \mid x)\, dy$$

Bayesian Cognition?



$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta)\, d\theta$$

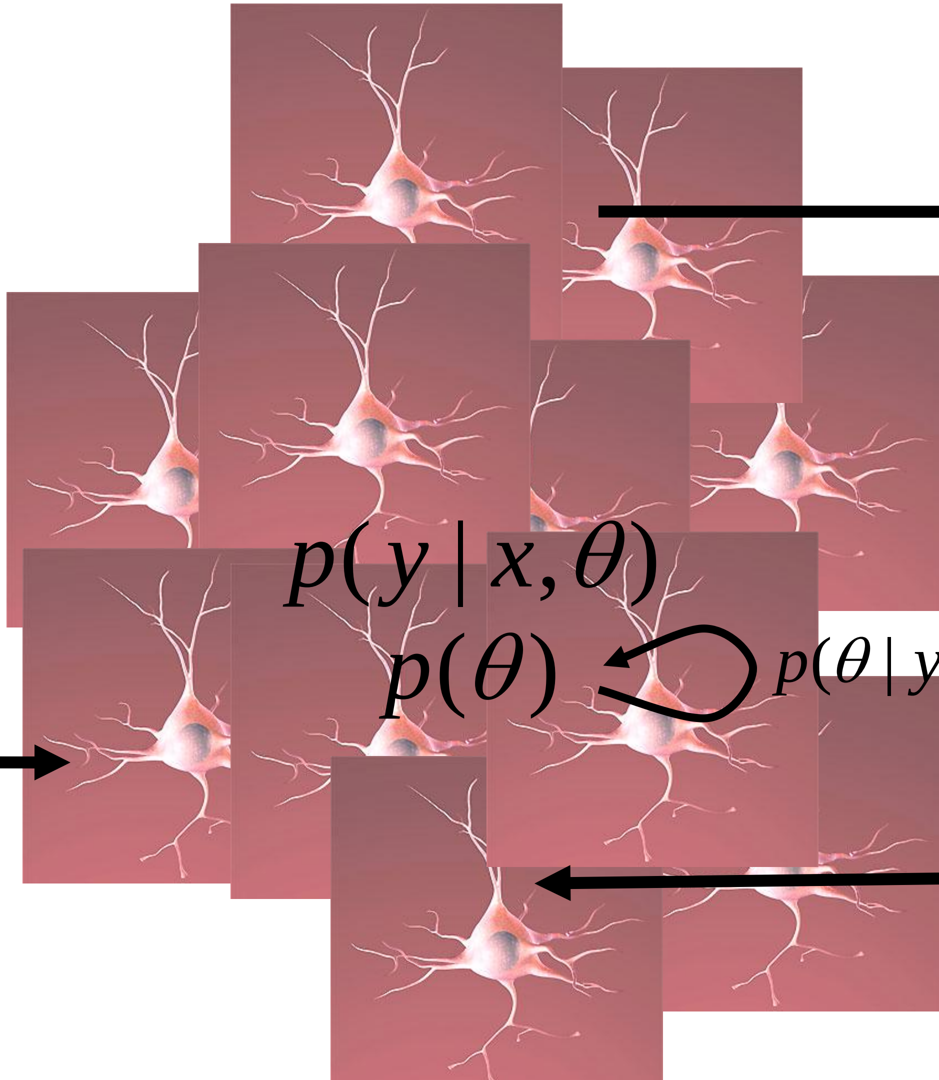
$$\hat{y} = \int y\; p(y \mid x)\, dy$$

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)\, p(\theta)}{\int p(y \mid x, \theta)\, p(\theta)\, d\theta}$$

x

y

Bayesian Cognition?



$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta)\, d\theta$$

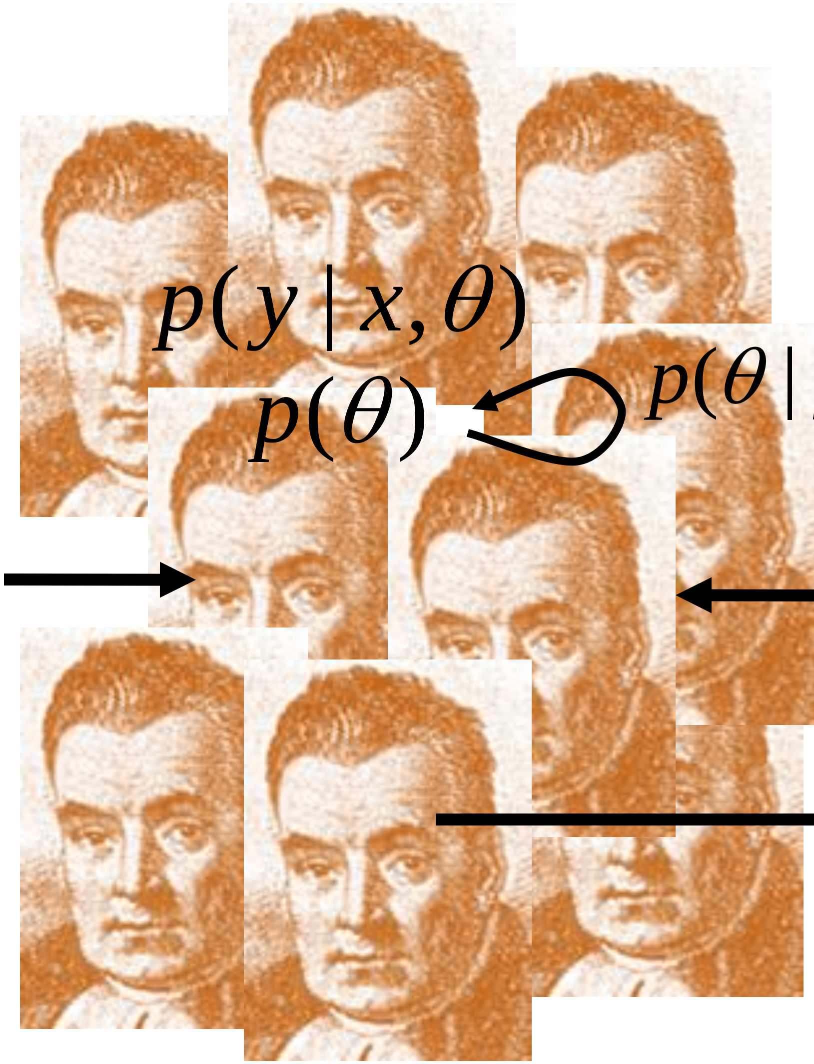
$$\hat{y} = \int y \; p(y \mid x)\, dy$$

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)\, p(\theta)}{\int p(y \mid x, \theta)\, p(\theta)\, d\theta}$$

x

y

Bayesian Cognition?



$p(y \mid x, \theta)$

$p(\theta)$

x

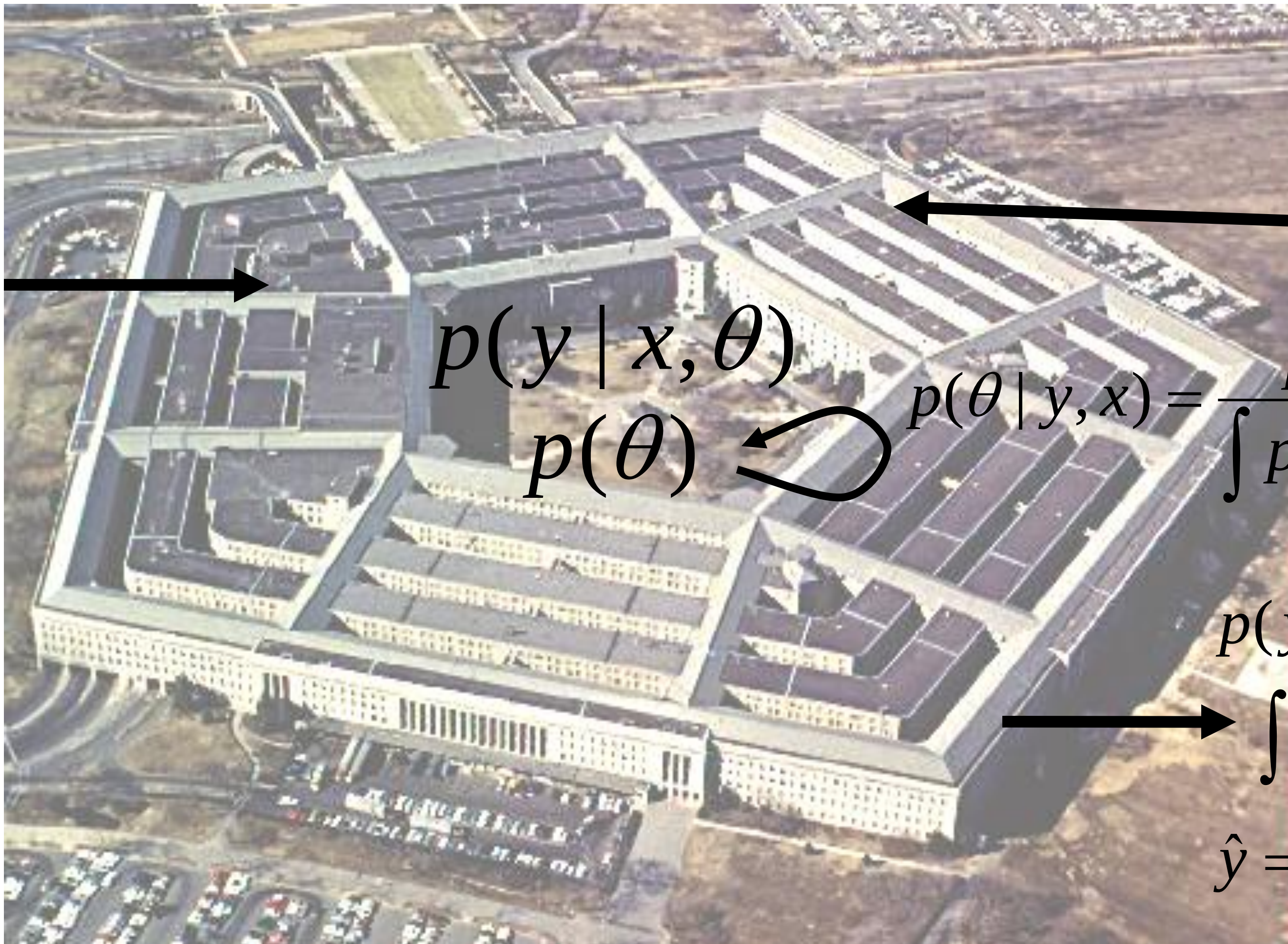
y

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)\, p(\theta)}{\int p(y \mid x, \theta)\, p(\theta) d\theta}$$

$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta) d\theta$$

$$\hat{y} = \int y \; p(y \mid x) dy$$

Bayesian Cognition?



x

y

$$p(y \mid x, \theta)$$

$$p(\theta)$$

$$p(\theta \mid y, x) = \frac{p(y \mid x, \theta)\, p(\theta)}{\int p(y \mid x, \theta)\, p(\theta)\, d\theta}$$

$$p(y \mid x) = \int p(y \mid x, \theta)\, p(\theta)\, d\theta$$

$$\hat{y} = \int y \; p(y \mid x)\, dy$$

To Ponder:

- For a Bayesian model of "cognitive behavior", what level of analysis is appropriate?
- If a system is Bayesian at one level of analysis, is it Bayesian at other levels?

Bayesian Cognition?

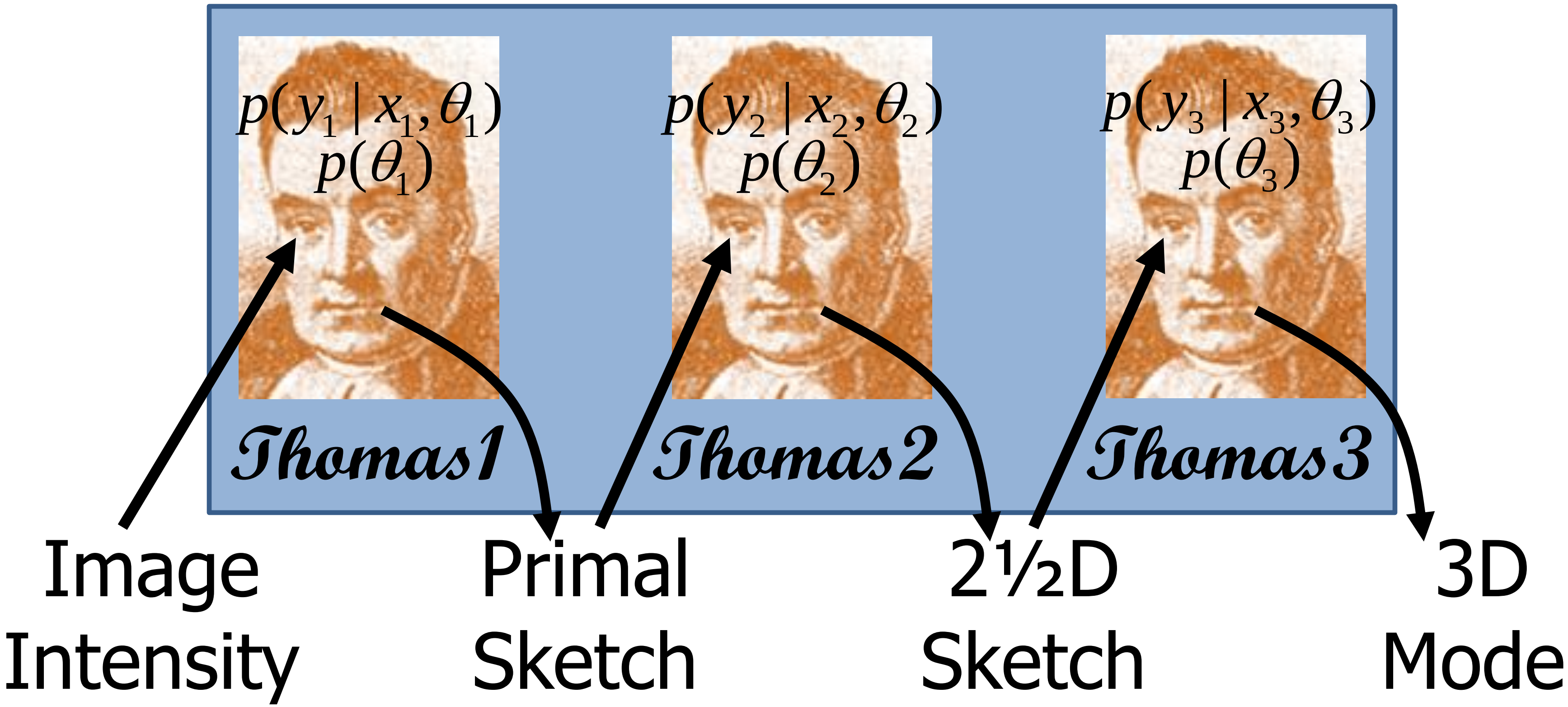
Marr (1982):

Image Intensity → Primal Sketch → 2½D Sketch → 3D Model

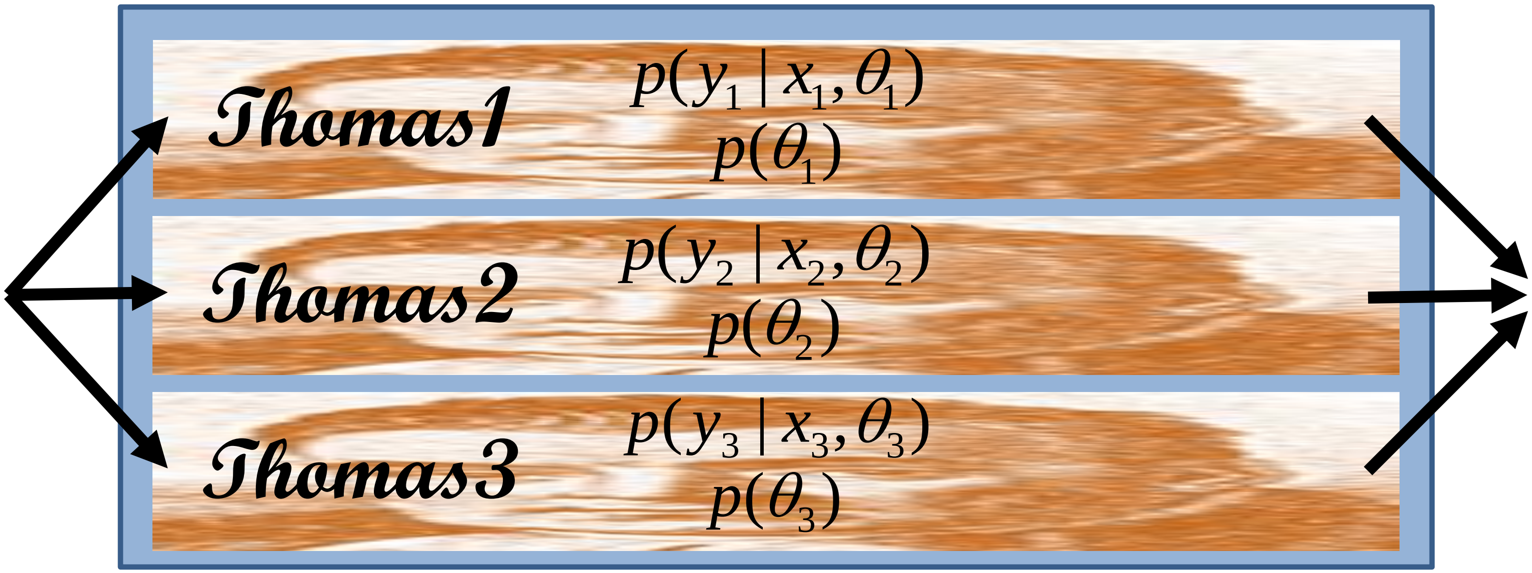
Is the overall mapping, from image to 3D model, Bayesian?

Is each component Bayesian?

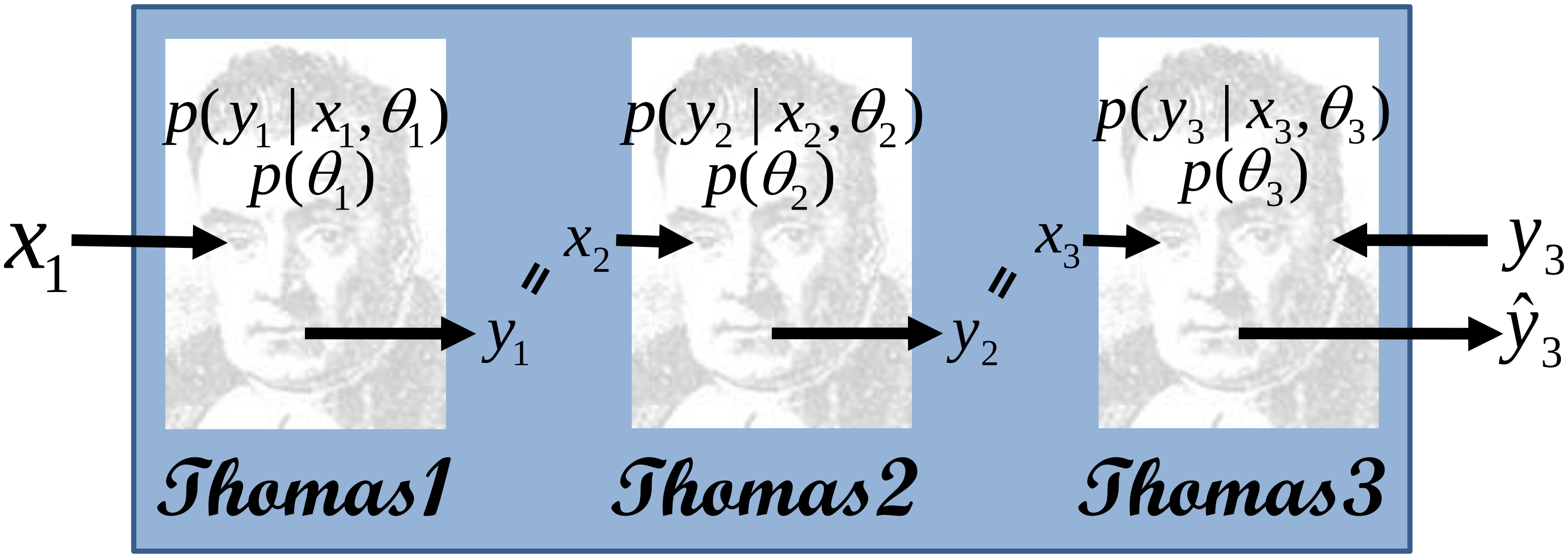
Consider a Chain of Bayesians



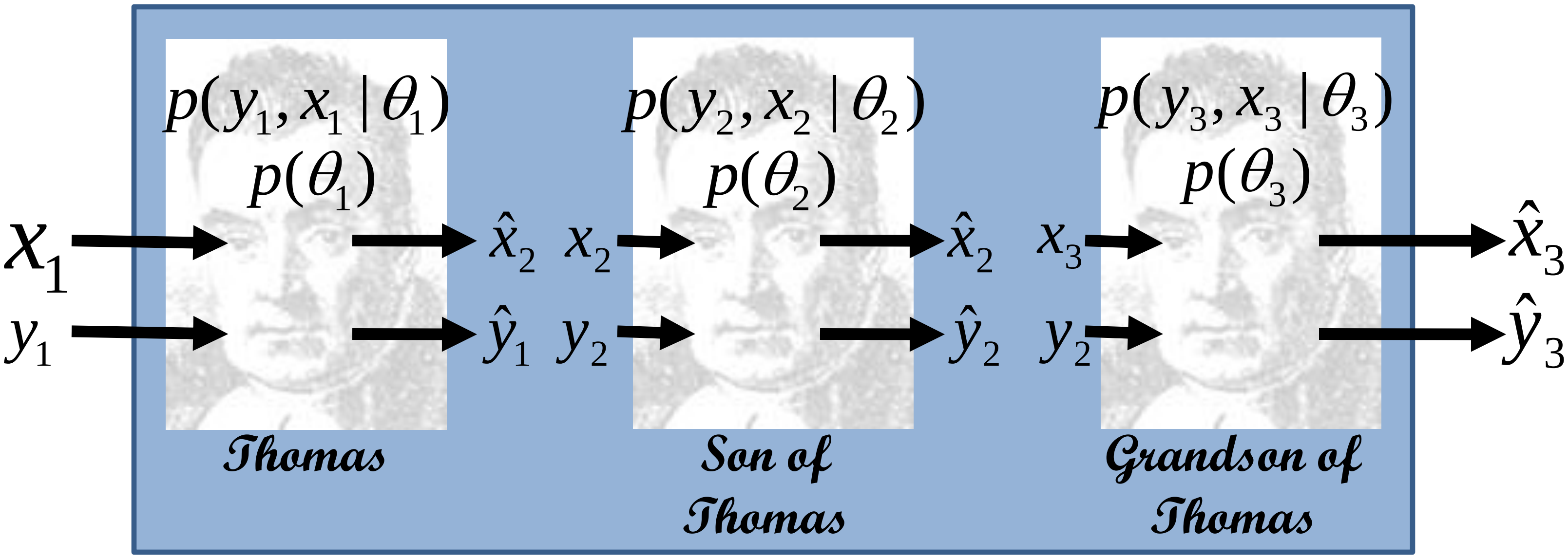
Not Parallel Bayesians



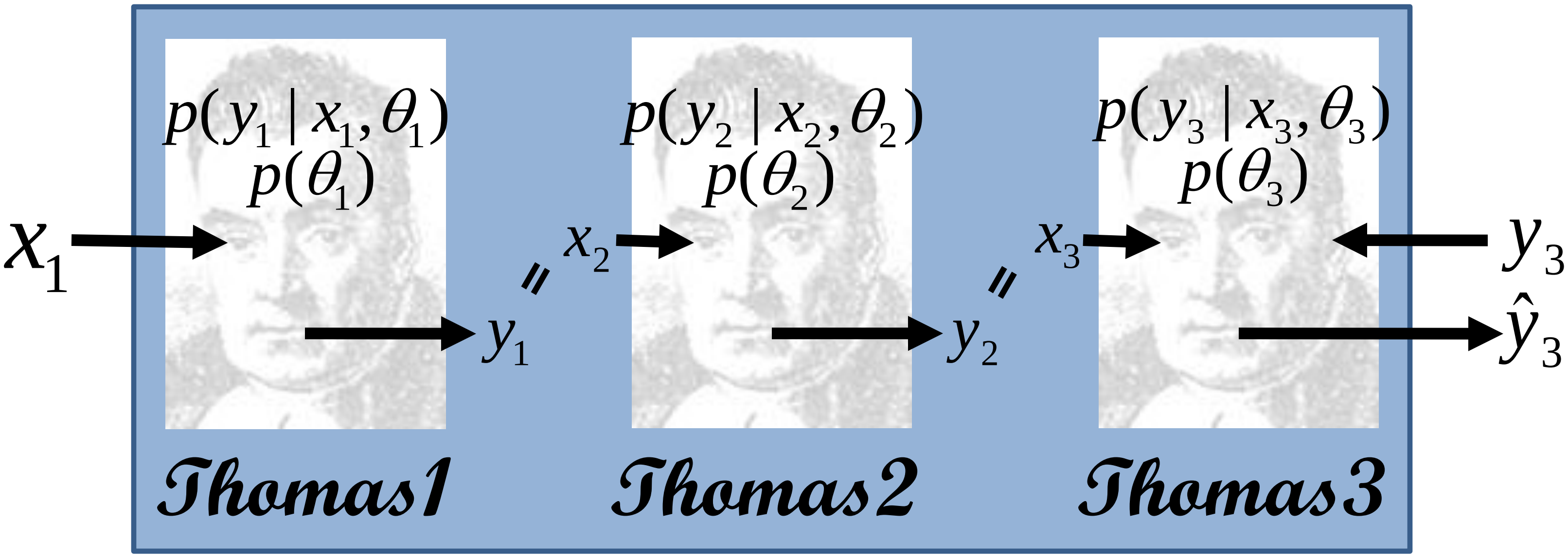
A Chain of Bayesians



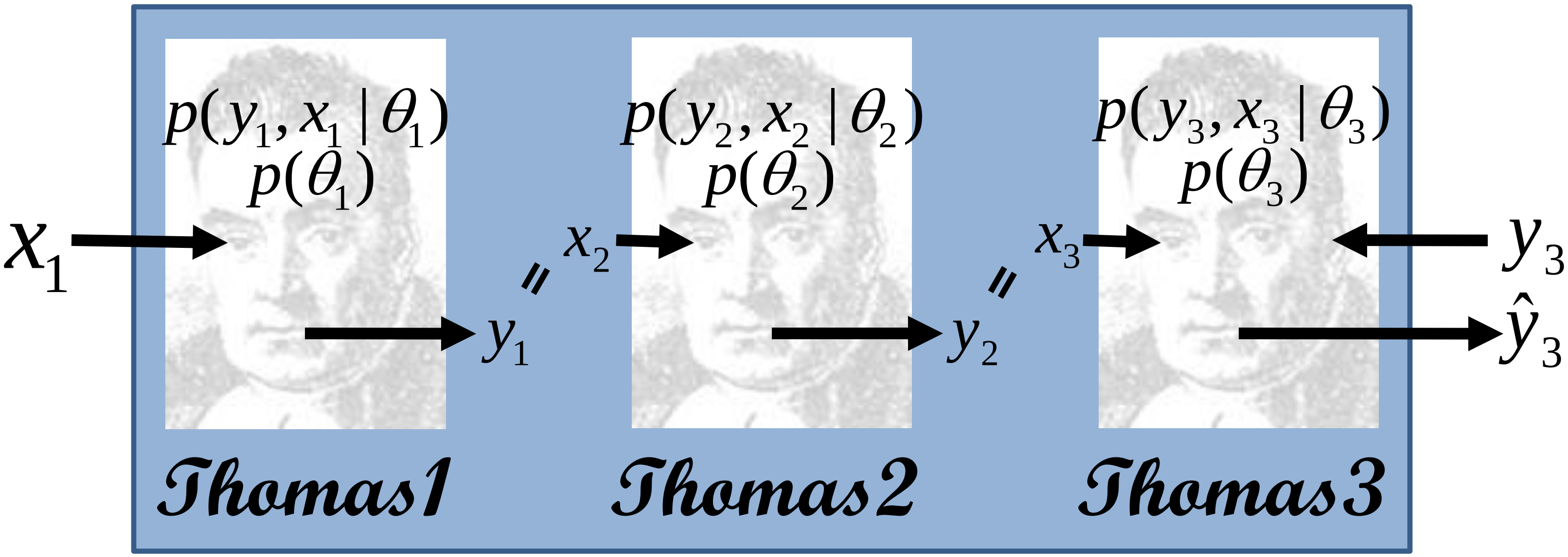
Not Iterated Bayesians



A Chain of Bayesians

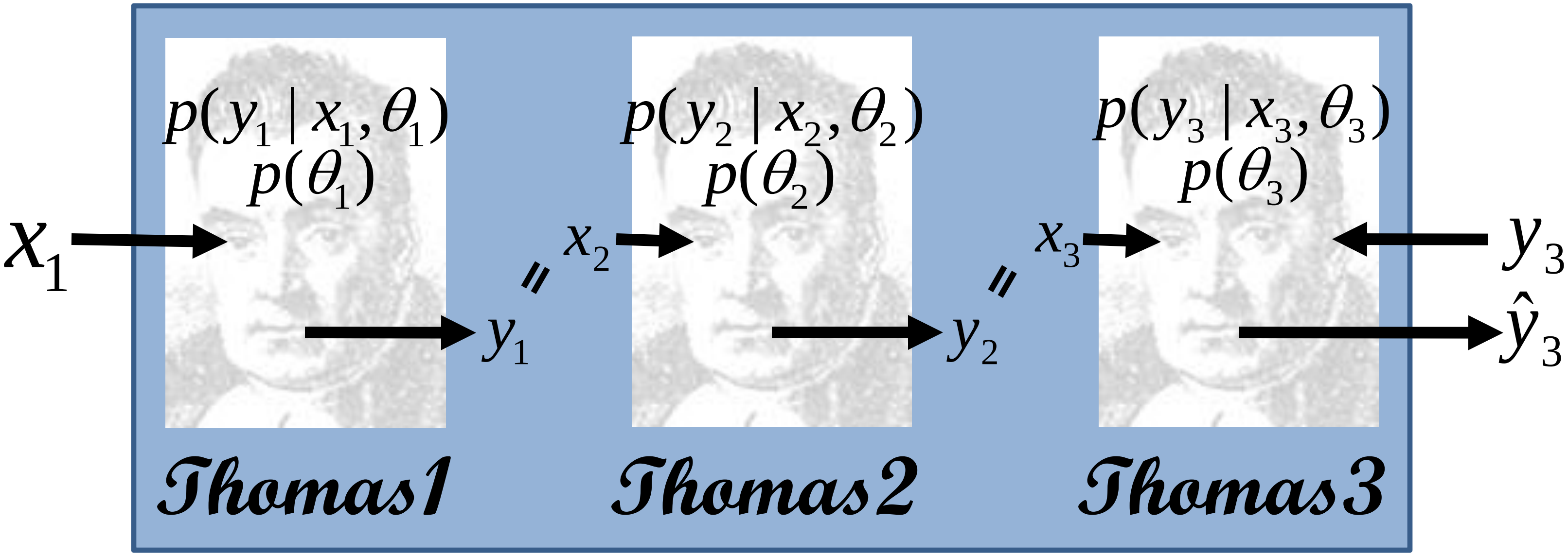


Could Be Generative Bayesians

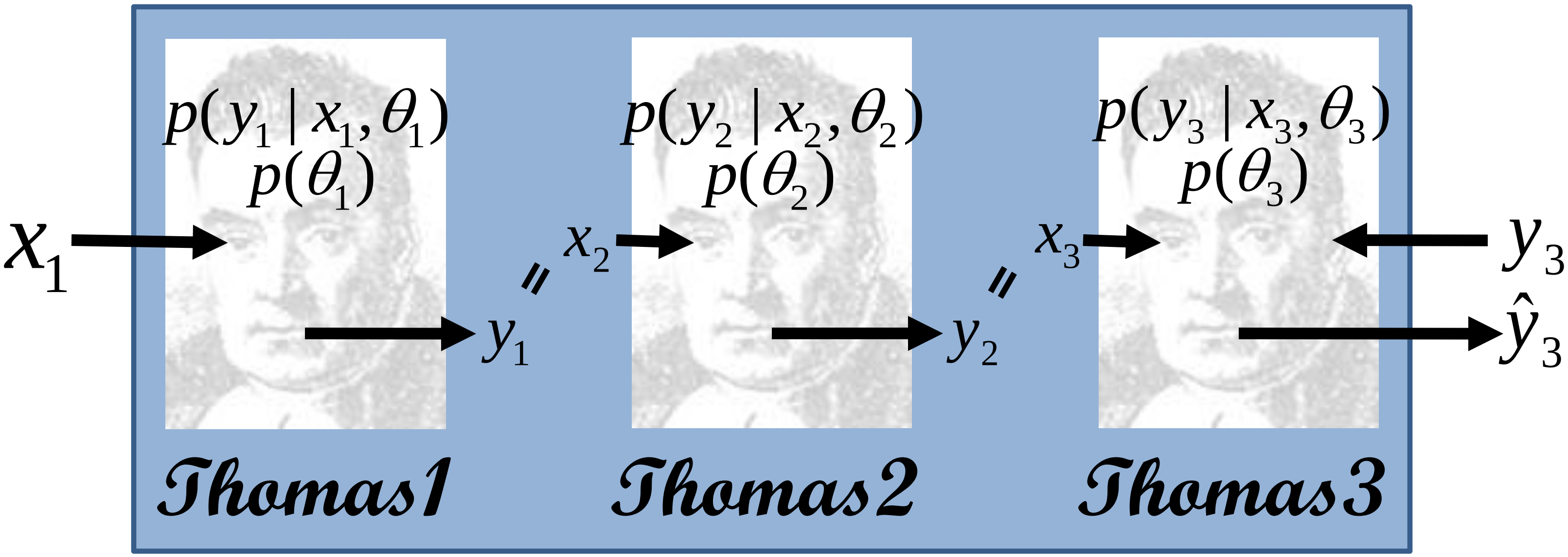


But not pursued here.

A Chain of Bayesians

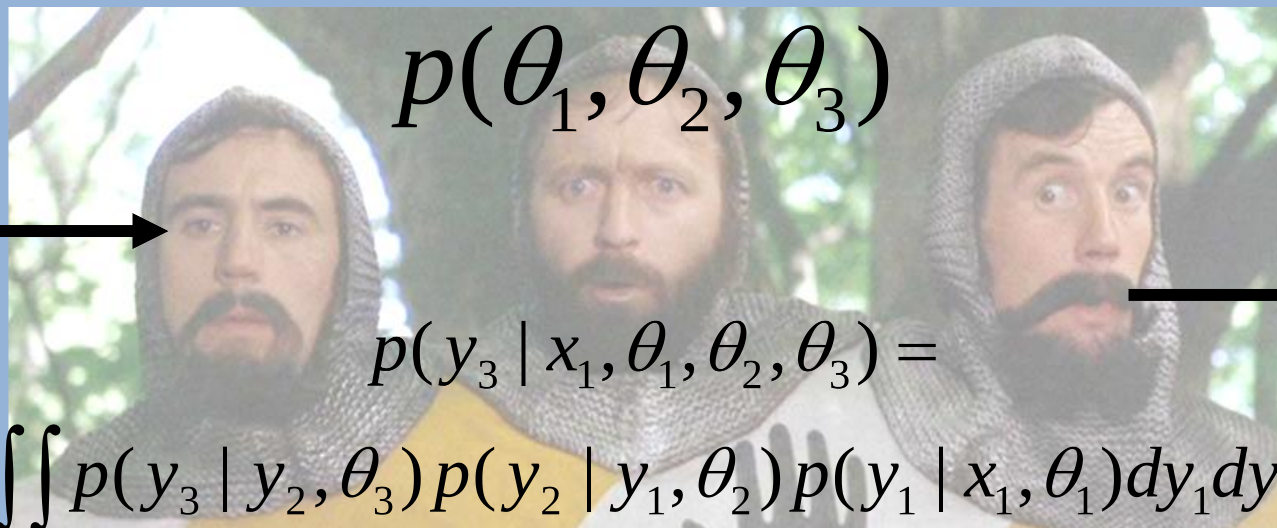


A Chain of Bayesians



The standard approach: The three heads are conjoined over a joint parameter space.

The Globally Bayesian Approach



$$p(\theta_1, \theta_2, \theta_3)$$

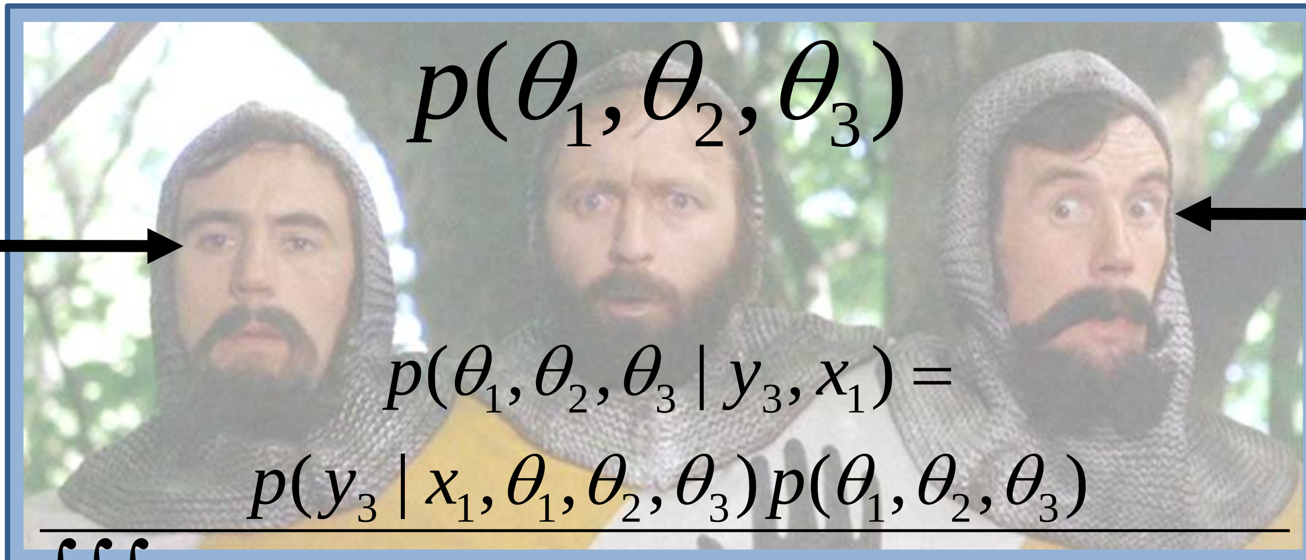
x_1

$\hat{y}_3$

$$p(y_3 \mid x_1, \theta_1, \theta_2, \theta_3) =$$
$$\iint p(y_3 \mid y_2, \theta_3)\, p(y_2 \mid y_1, \theta_2)\, p(y_1 \mid x_1, \theta_1)\, dy_1\, dy_2$$

$$p(y_3 \mid x_1) = \iiint p(y_3 \mid x_1, \theta_1, \theta_2, \theta_3)\, p(\theta_1, \theta_2, \theta_3)\, d\theta_1\, d\theta_2\, d\theta_3$$

The Globally Bayesian Approach



$$p(\theta_1, \theta_2, \theta_3)$$

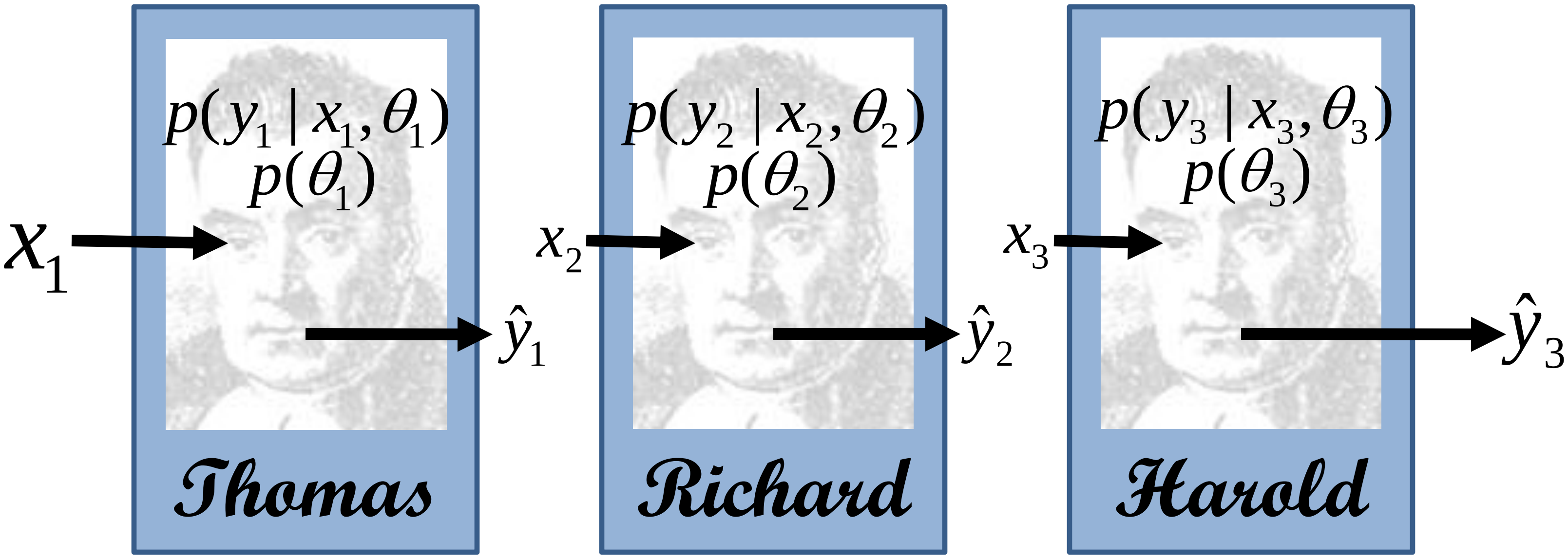
$$x_1 \qquad\qquad y_3$$

$$p(\theta_1, \theta_2, \theta_3 \mid y_3, x_1) = \frac{p(y_3 \mid x_1, \theta_1, \theta_2, \theta_3)\, p(\theta_1, \theta_2, \theta_3)}{\iiint p(y_3 \mid x_1, \theta_1, \theta_2, \theta_3)\, p(\theta_1, \theta_2, \theta_3)\, d\theta_1 d\theta_2 d\theta_3}$$

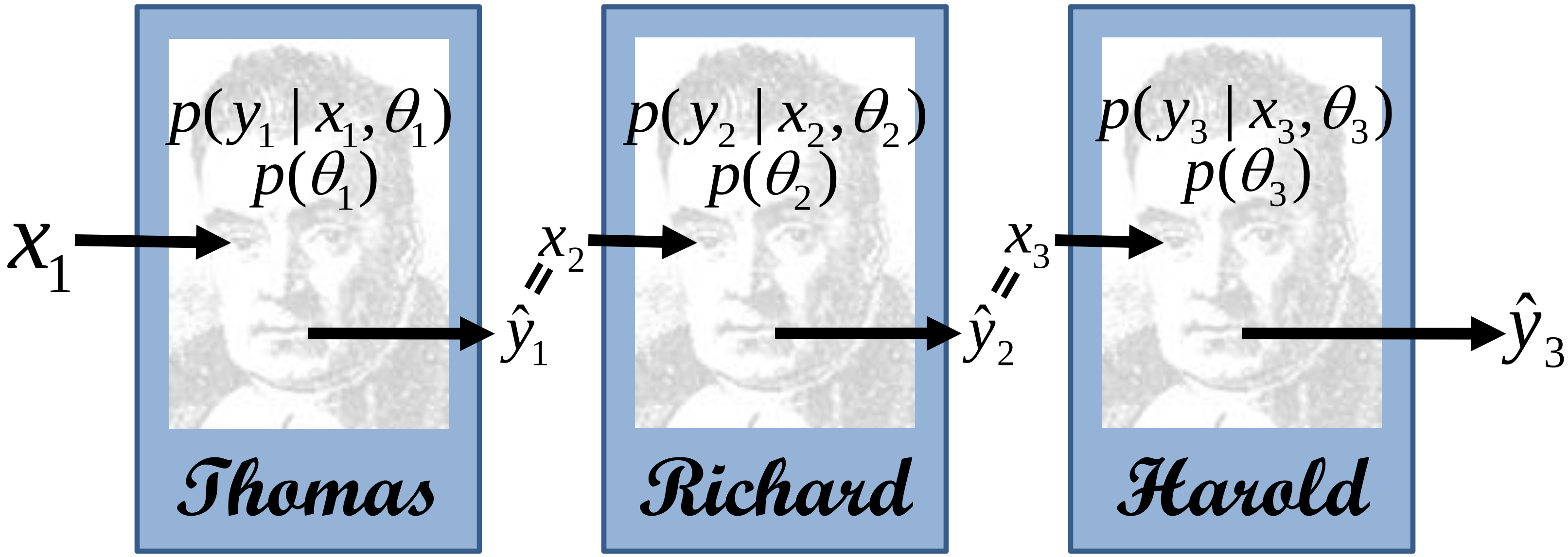
The Locally Bayesian Approach

You are all individuals!

Yes, we *are* all individuals!



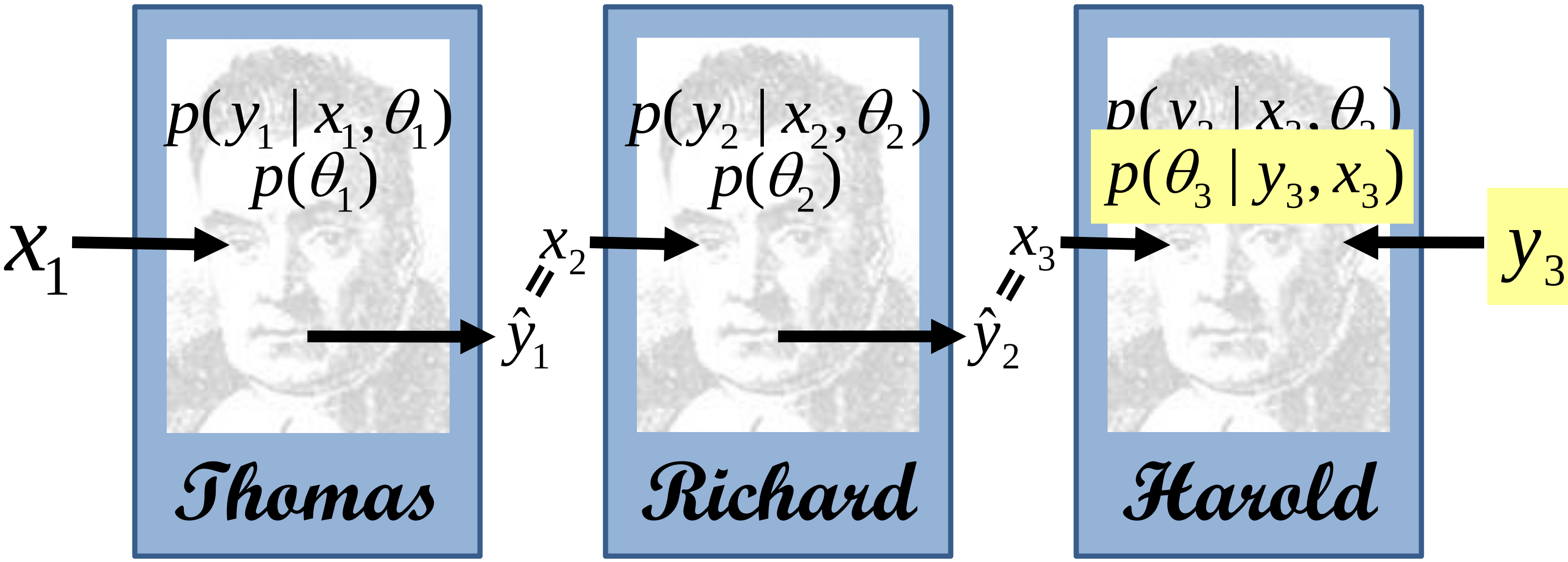
Locally Bayesian *Prediction*



Each Bayesian agent computes its best prediction, and propagates it forward.

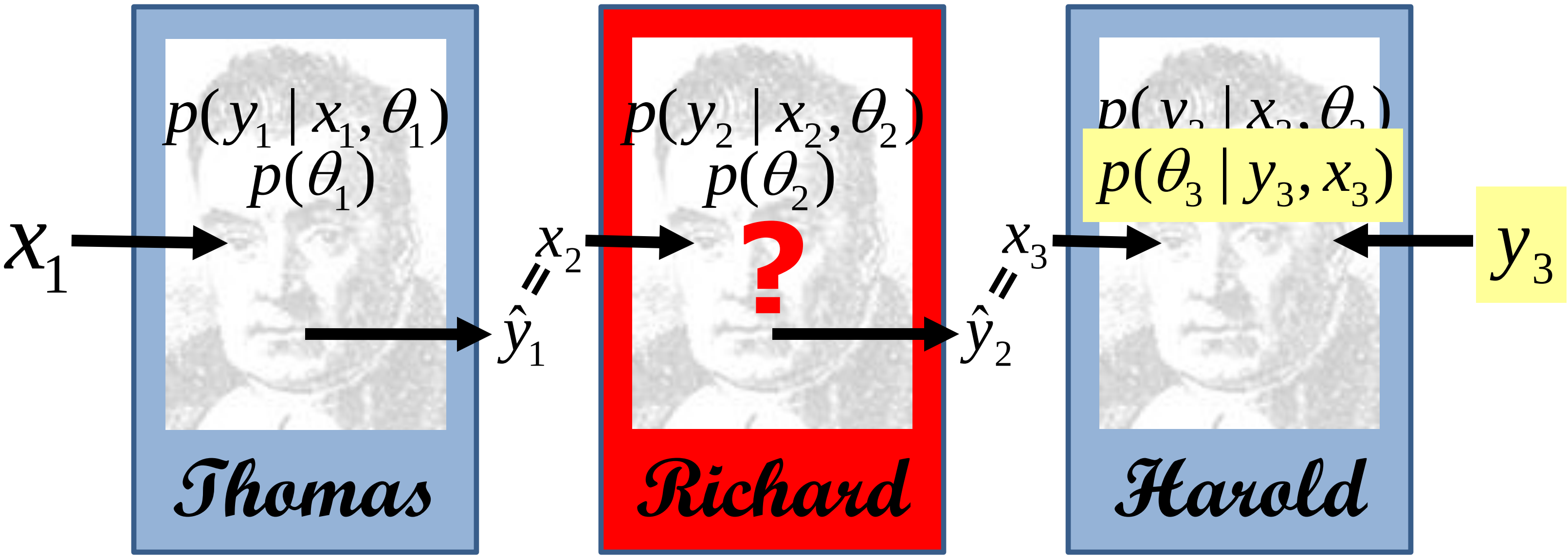
This process needs integrals over only the individual parameter spaces.

Locally Bayesian *Learning*



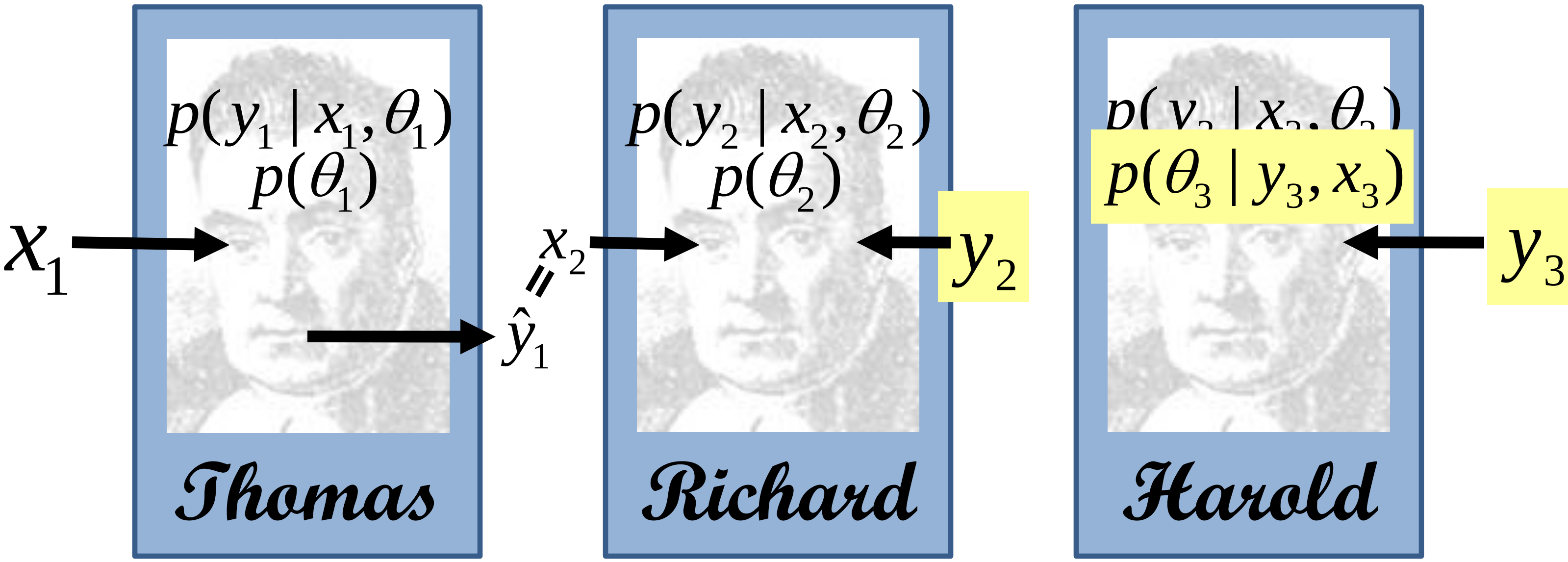
Update $p(\theta_3|y_3,x_3)$ by Bayes' rule.
Involves integrating only over the θ_3 parameter space.

Locally Bayesian Learning



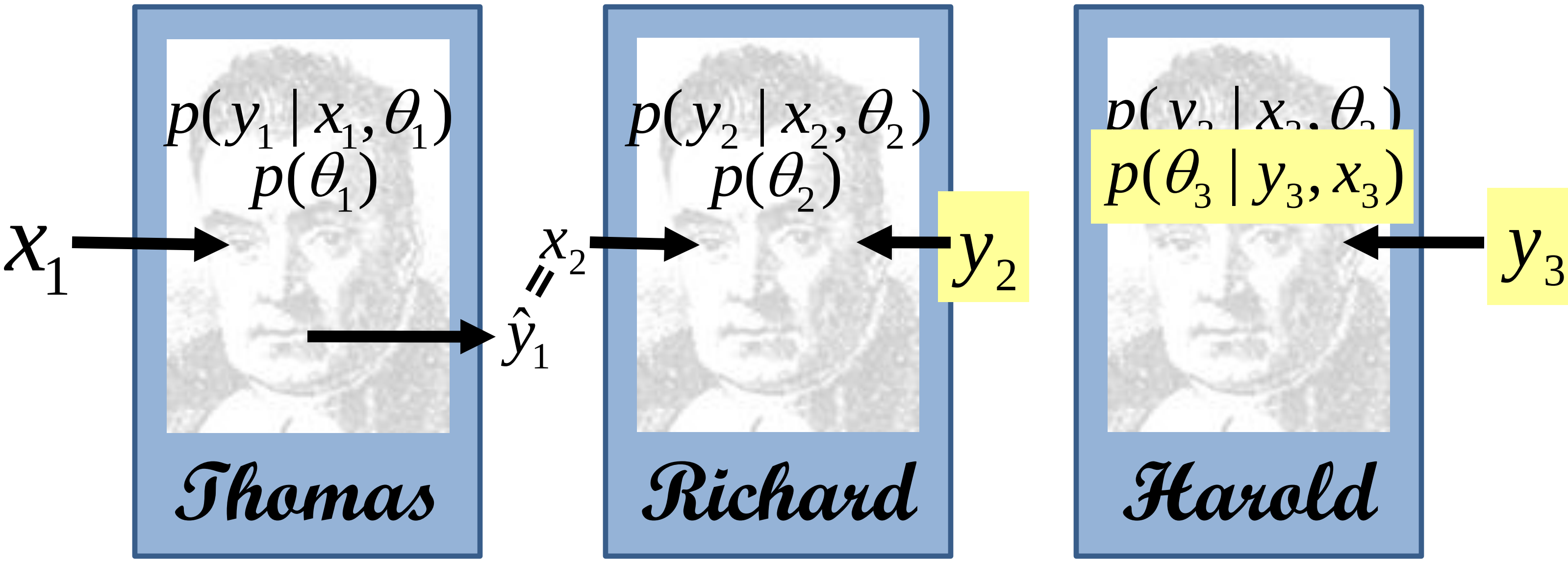
But how should poor Richard update his beliefs about θ_2? He needs a y_2 value to learn about!

Locally Bayesian Learning



$$\text{Let } y_2 = \underset{x_3^*}{\operatorname{argmax}} \; p(y_3 \mid x_3^*)$$

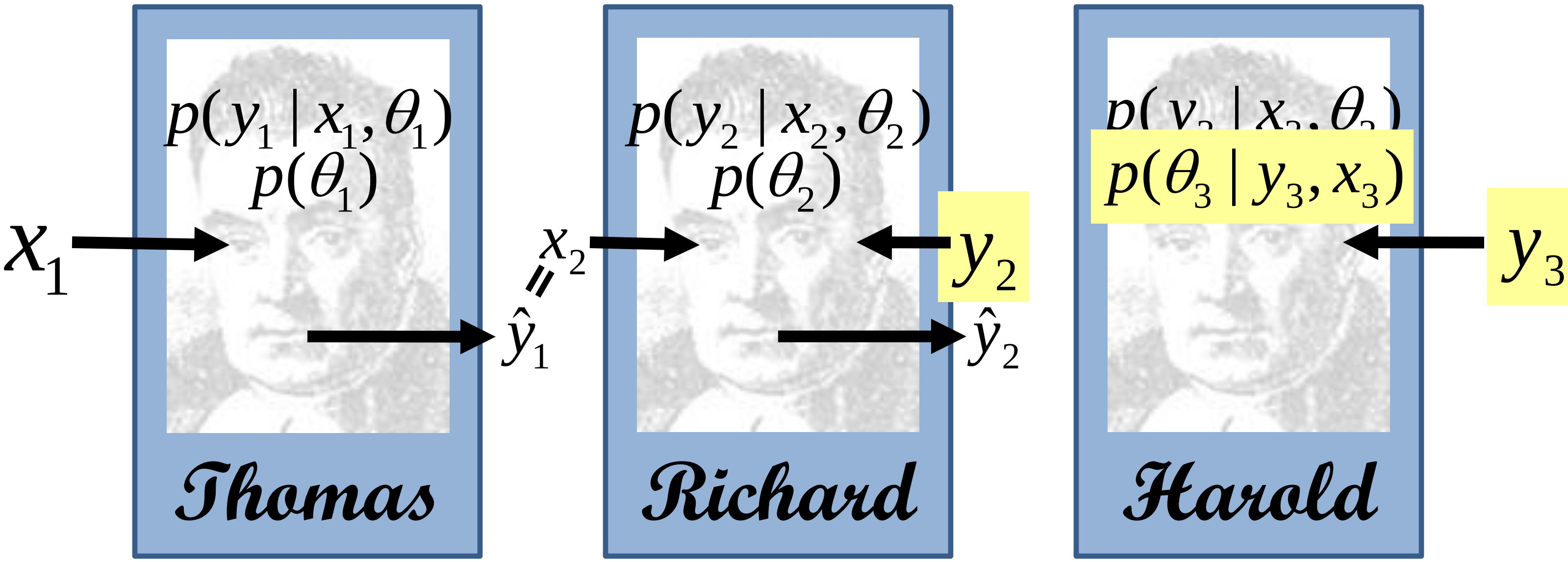
Locally Bayesian Learning



$$\text{Let } y_2 = \underset{x_3^*}{\operatorname{argmax}}\ p(y_3 \mid x_3^*)$$

Harold tells Richard to produce a value that is consistent with Harold's beliefs!

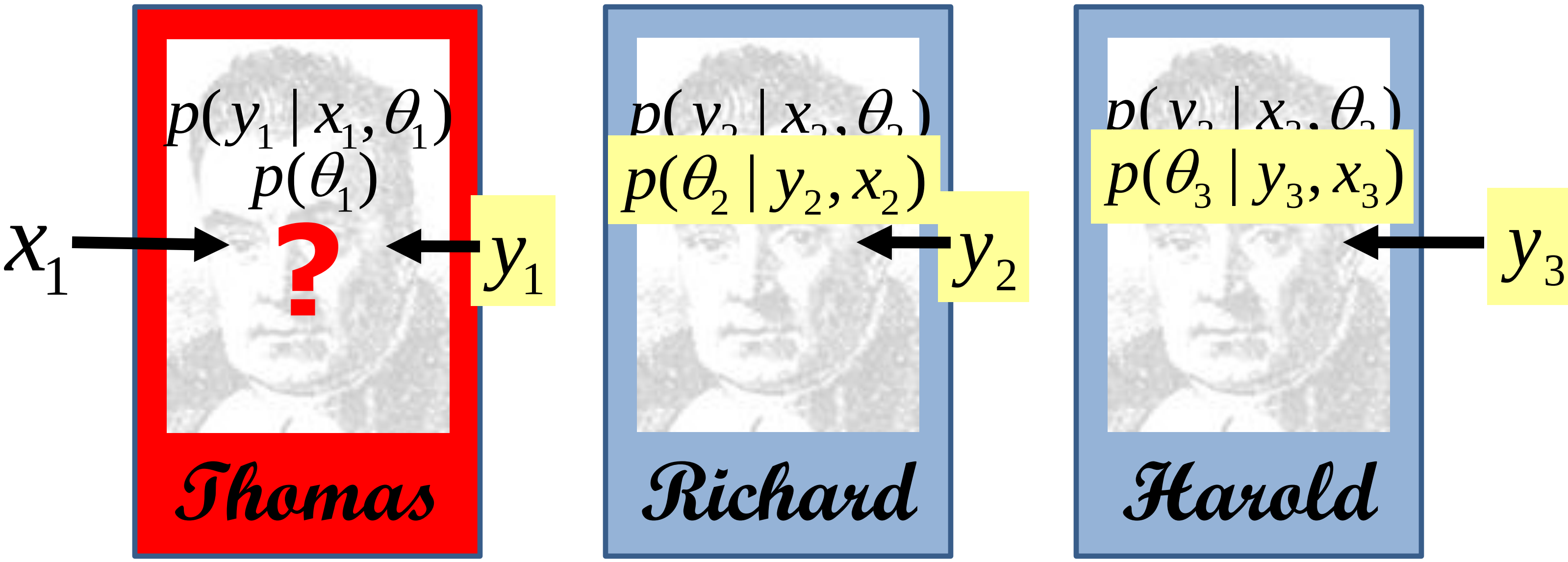
Locally Bayesian Learning



$$\text{Let } y_2 = \underset{x_3^*}{\operatorname{argmax}}\; p(y_3 \mid x_3^*)$$

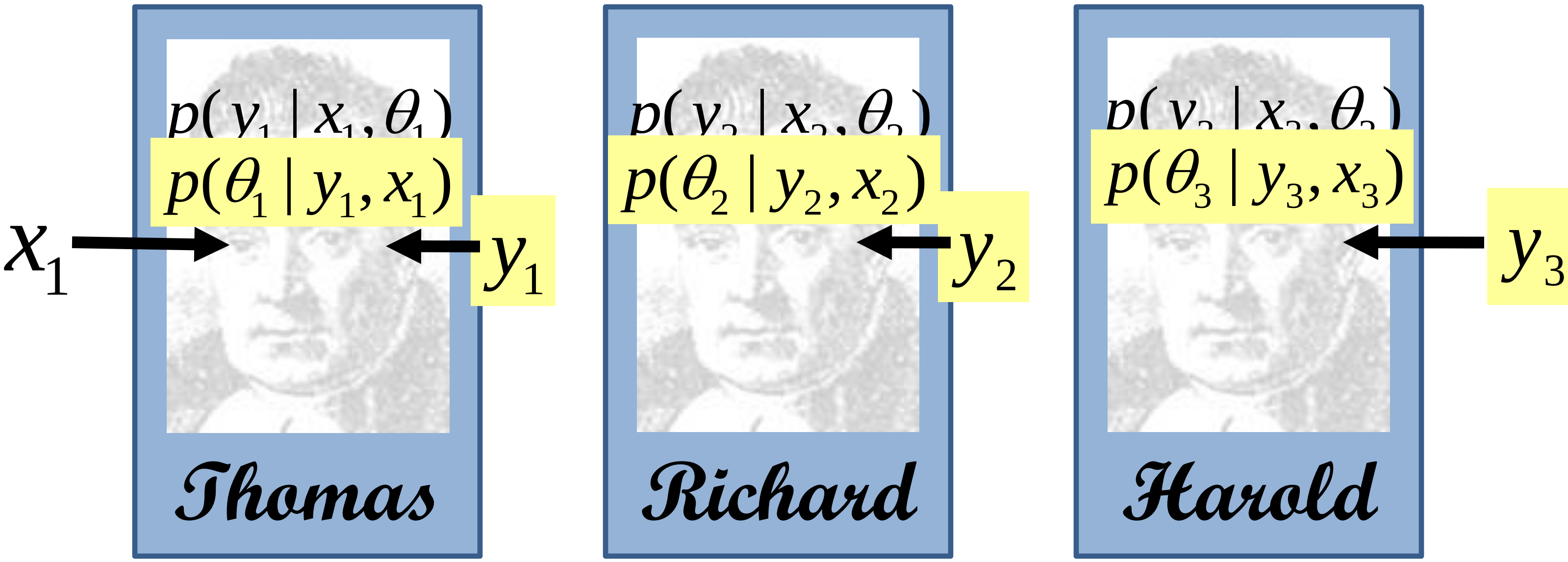
In practice, don't need to maximize; just get a value of y_2 with $p(y_3 \mid y_2) > p(y_3 \mid \hat{y}_2)$

Locally Bayesian Learning



$$\text{Let } y_1 = \underset{x_2^*}{\operatorname{argmax}}\ p(y_2 \mid x_2^*)$$

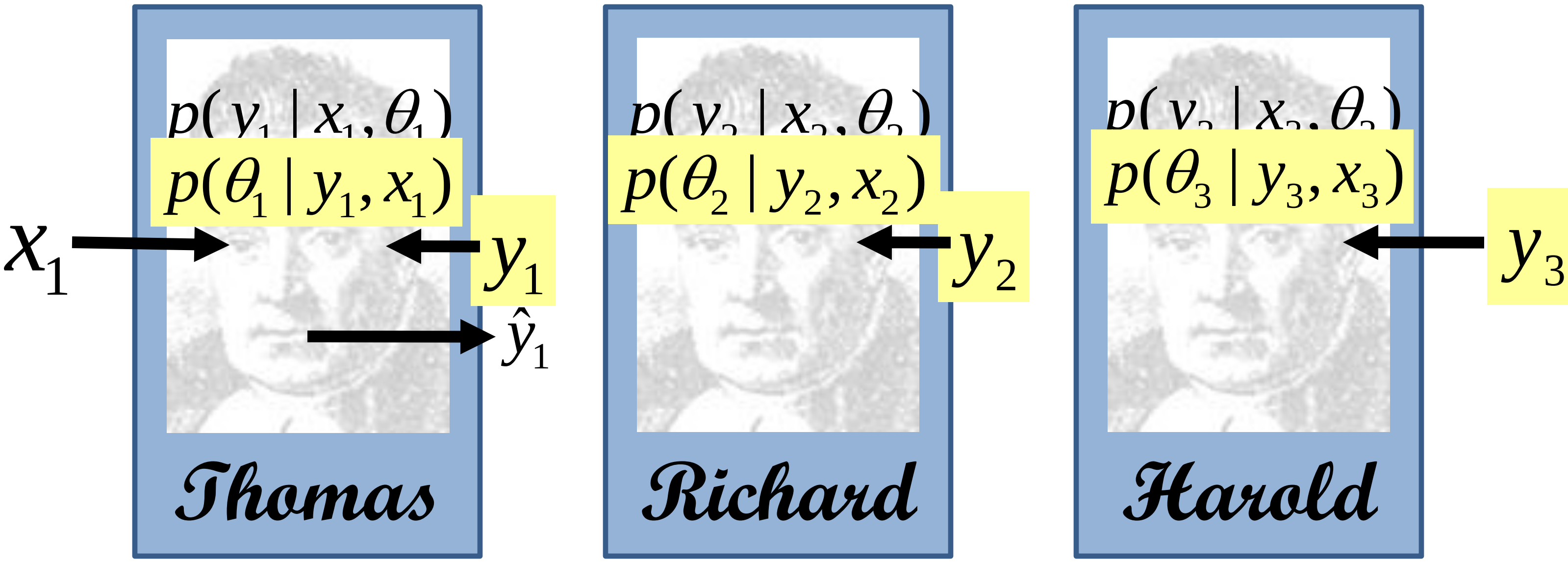
Locally Bayesian Learning



$$\text{Let } y_1 = \underset{x_2^*}{\operatorname{argmax}} \; p(y_2 \mid x_2^*)$$

Richard tells Thomas to produce a value that is consistent with Richard's beliefs!

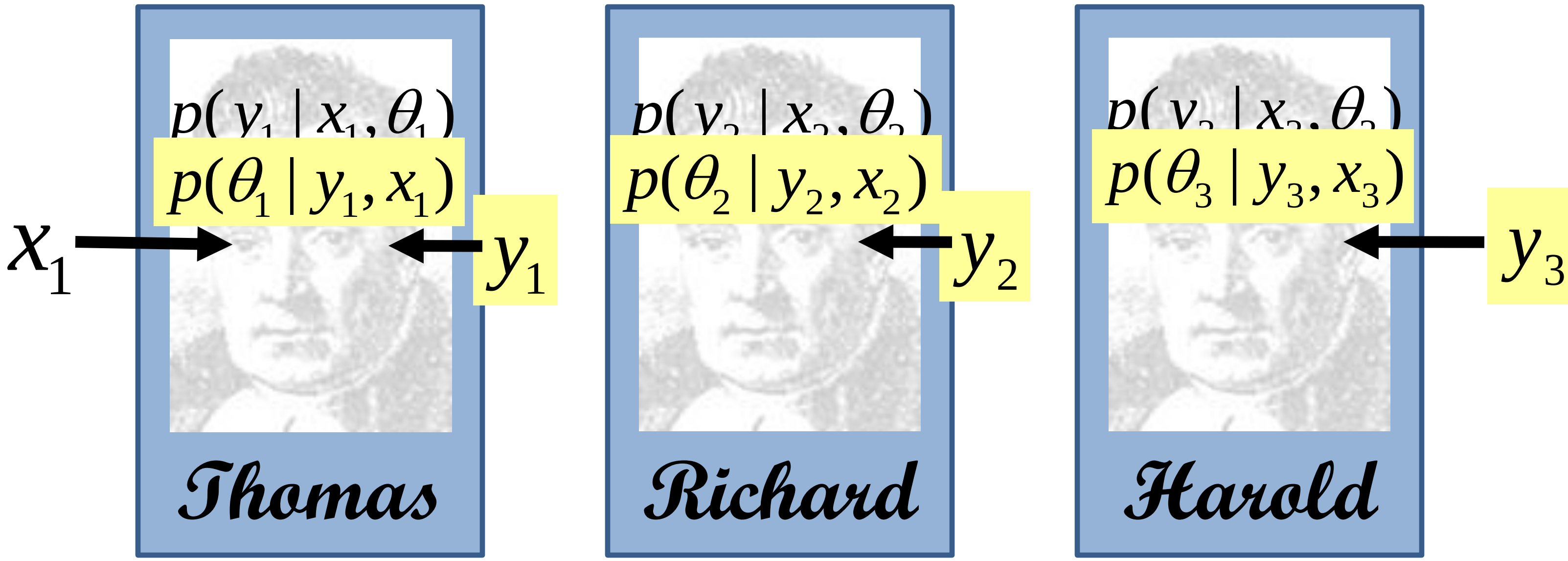
Locally Bayesian Learning



$$\text{Let } y_1 = \underset{x_2^*}{\text{argmax}}\ p(y_2 \mid x_2^*)$$

In practice, don't need to maximize; just get a value of y_1 with $p(y_2 \mid y_1) > p(y_2 \mid \hat{y}_1)$

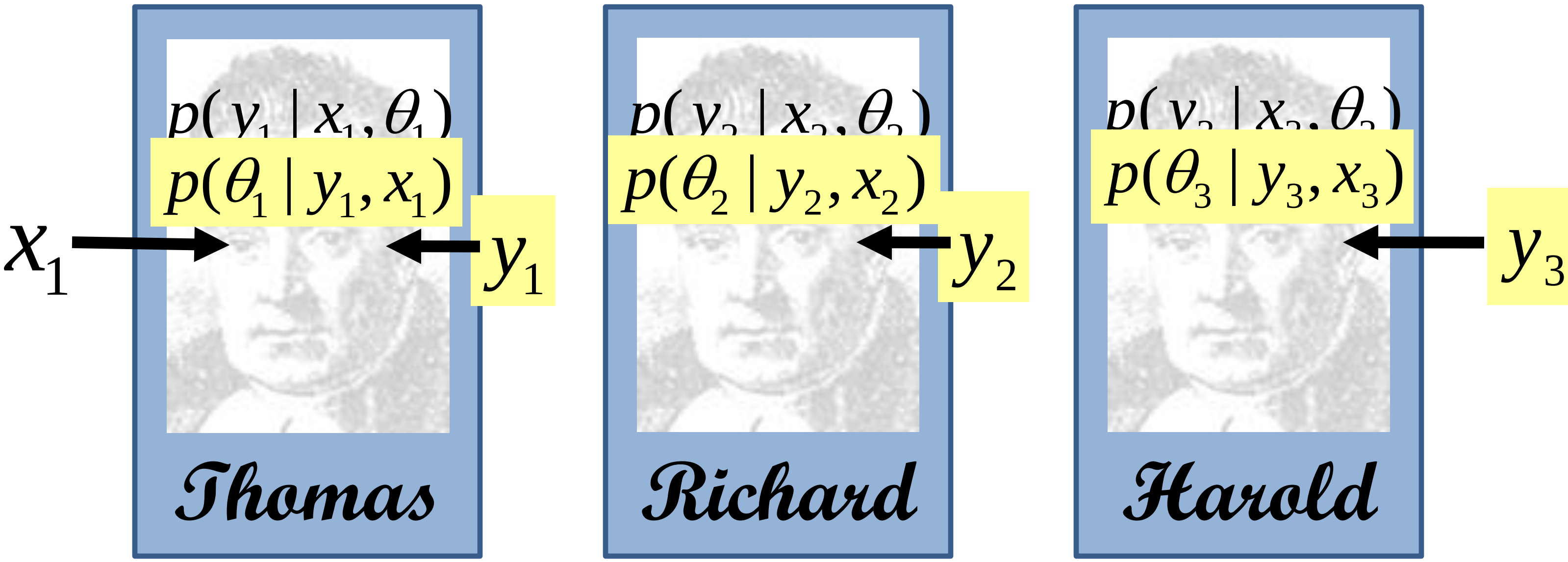
Locally Bayesian Learning



Other updating dynamics are possible.

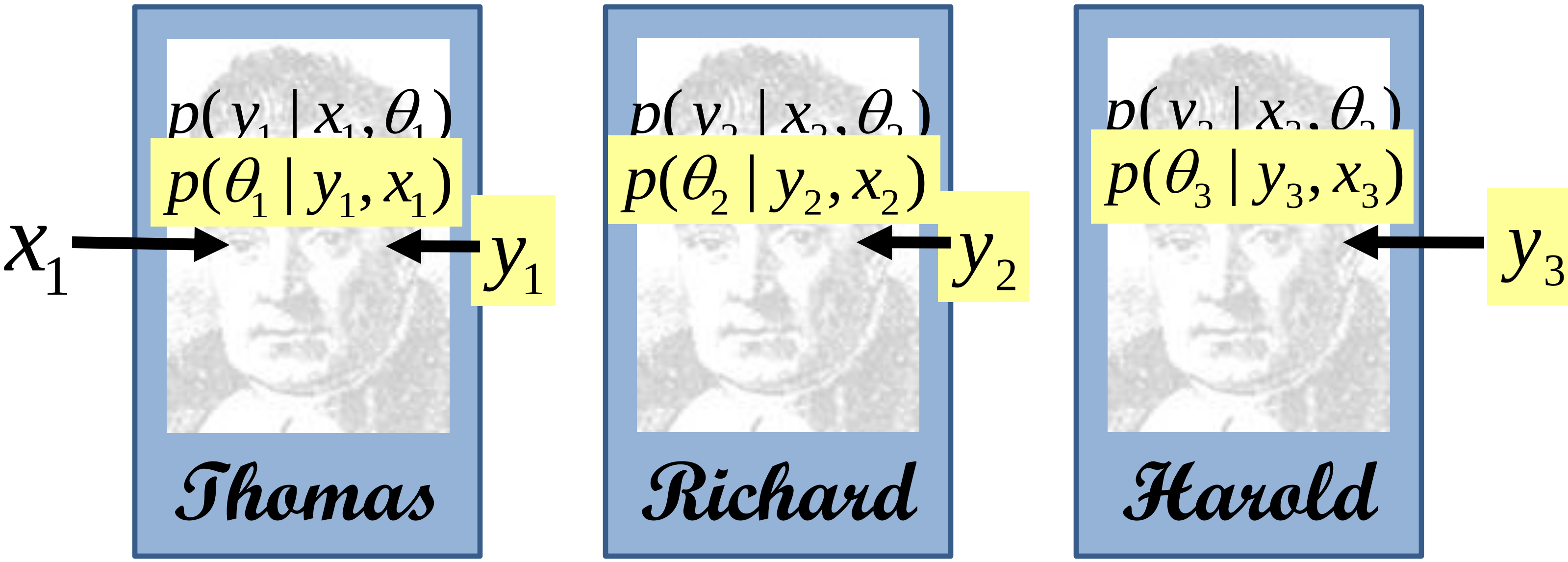
E.g., first propagate y_3 all the way back to the first agent, and update $p(\theta_1|y_1,x_1)$. Then compute predicted $\hat{y}_1$. Then update $p(\theta_2|y_2, \hat{y}_1)$. And so on.

Locally Bayesian Learning



Each agent is told by its superior to learn a datum that is maximally consistent (or minimally inconsistent) with the superior's current beliefs.

Locally Bayesian Learning



This process protects the superior's beliefs from disconfirmation! The inferior will learn to "distort the data" to avoid disconfirming the superior.

Locally Bayesian Learning (LBL)

LBL preserves current beliefs and creates "epicycles" for new data. Perhaps not perfectly optimal, but then, are real systems?

Put your models where your data are...

- Some real behavior, in the domain of associative learning, to which Locally Bayesian Learning can be applied.

Typical Learning Task

Stimulus presentation and response collection:

Typical Learning Task

Corrective feedback:

Phenomena Suggestive of Attention in Learning

- Fewer relevant cues ➔ faster learning.
- Intradimensional shifts are faster than extradimensional.
- Attenuated learning after blocking.
- Overshadowing.
- Context-specific attention.
- Highlighting.
- Et cetera!

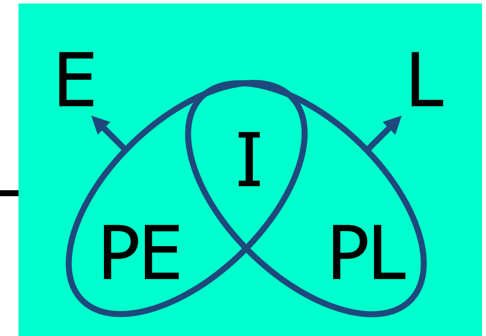
Highlighting:

Early Training:	I.PE→E	
Late Training:	I.PE→E	I.PL→L
Testing Results:	I→? (E!) PE.PL→? (L!)	

Highlighting:

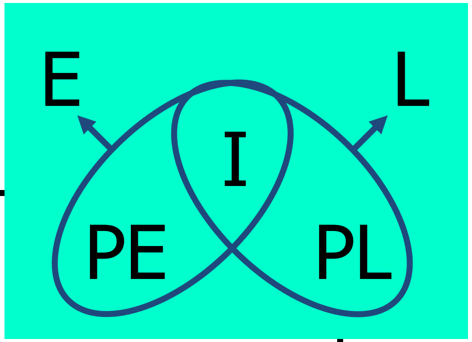
Early Training:	I.PE→E	
Late Training:	I.PE→E	I.PL→L
Testing Results:	I→? (E!)	
	PE.PL→? (L!)	

Highlighting:



Early Training:	I.PE→E	
Late Training:	I.PE→E	I.PL→L
Testing Results:	I→? (E!) PE.PL→? (L!)	

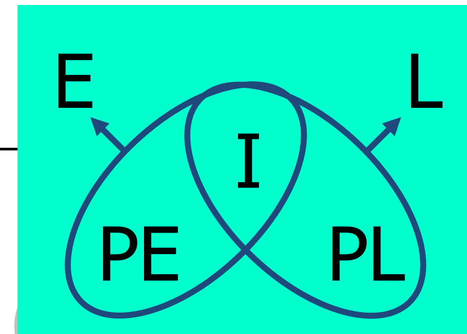
Highlighting:



Early Training:	I.PE→E
Late Training:	I.PE→E I.PL→L
Testing Results:	I→? (E!) PE.PL→? (L!)

Design: Highlighting

Phase	Cues→Outcome	
Initial Training:	(2x) I1.PE1→E1	E2
3:1 base-rate Training:	(3x) I1.PE1→E1 (1x) I1.PL1→L1	E2 2
1:3 base-rate Training:	(1x) I1.PE1→E1 (3x) I1.PL1→L1	(1x) I2.PE2→E2 (3x) I2.PL2→L2
Testing:	PE.PL→?, etc.	



Design: Highlighting

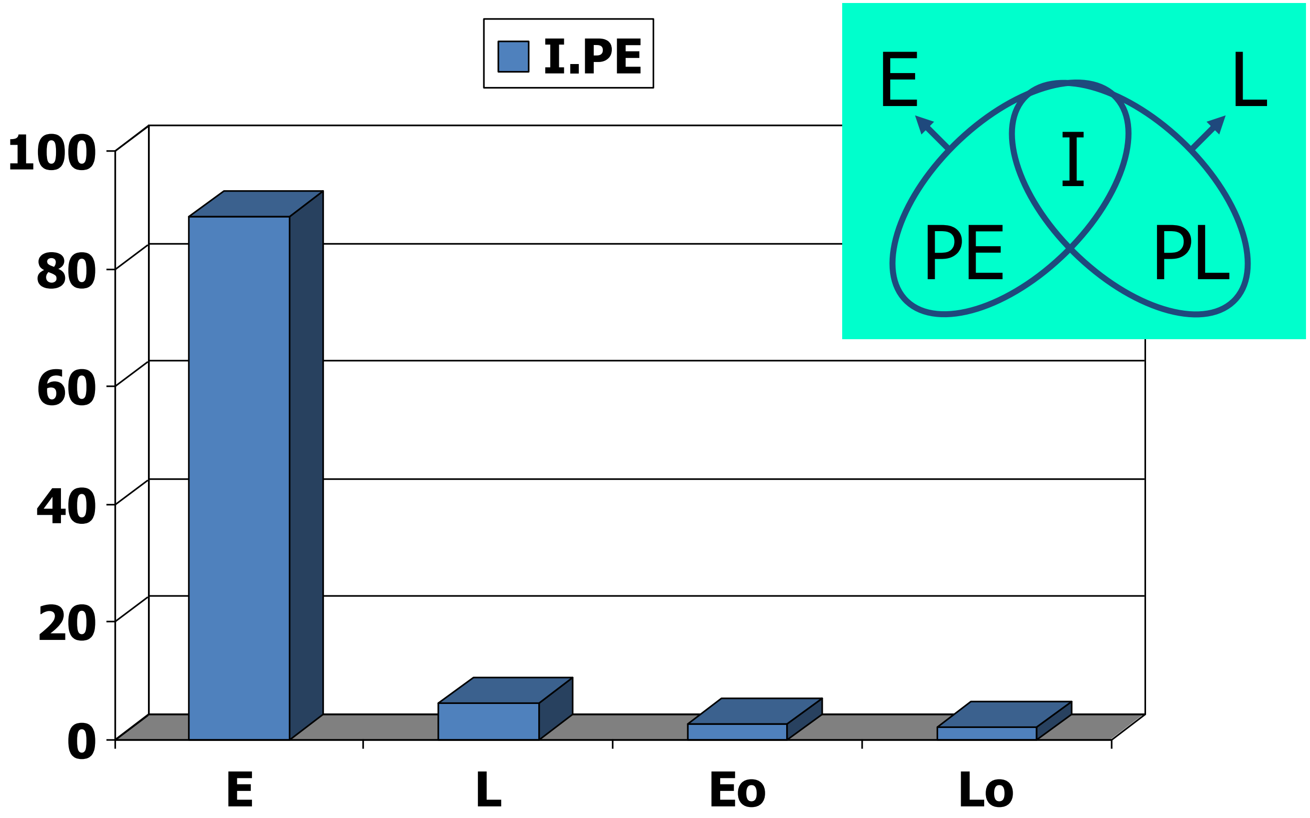
Phase	Cues→Outcome	
Initial Training:	(2x) I1.PE1→E1	(2x) I2.PE2→E2
3:1 base-rate Training:	(3x) I1.PE1→E1 (1x) I1.PL1→L1	(3x) I2.PE2→E2 (1x) I2.PL2→L2
1:3 base-rate Training:	(1x) I1.PE1→E1 (3x) I1.PL1→L1	(1x) I2.PE2→E2 (3x) I2.PL2→L2
Testing:	PE.PL→?, etc.	

“Canonical” Design: Highlighting

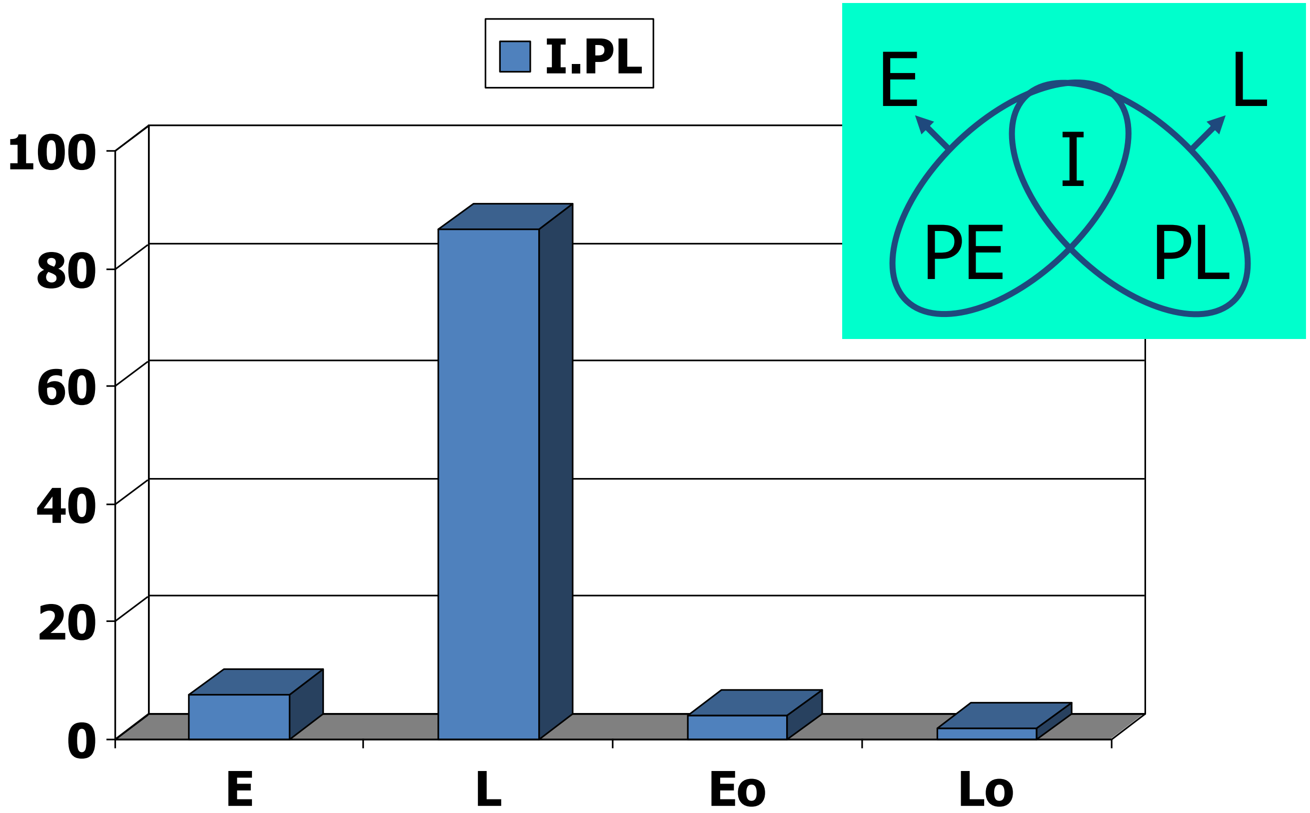
# Blocks	Cues→Outcome
N1:	(2x) I1.PE1→E1 (2x) I2.PE2→E2
N2:	(3x) I1.PE1→E1 (3x) I2.PE2→E2 (1x) I1.PL1→L1 (1x) I2.PL2→L2
N1+N2:	(1x) I1.PE1→E1 (1x) I2.PE2→E2 (3x) I1.PL1→L1 (3x) I2.PL2→L2

Frequency of I.PE→E trials *equals* frequency of I.PL→L trials.

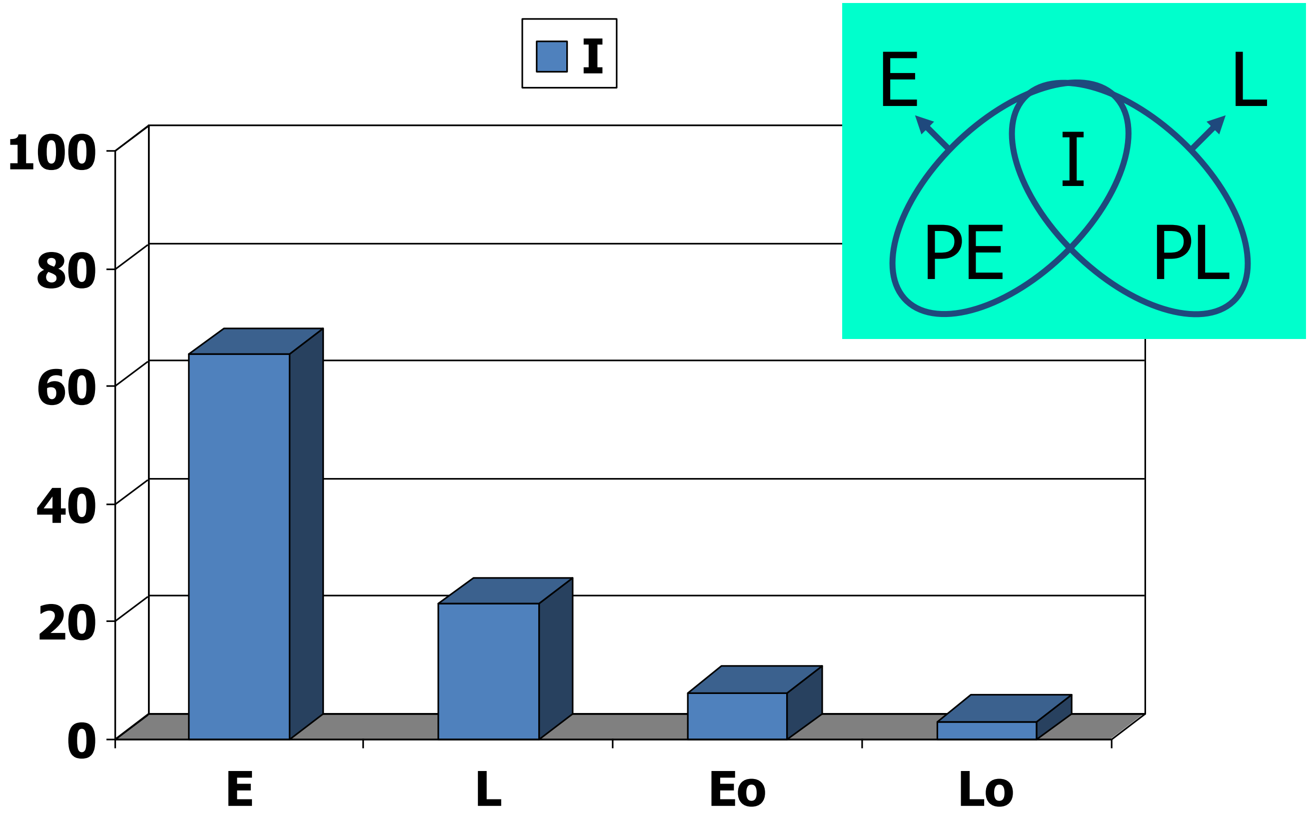
Highlighting: Results I.PE



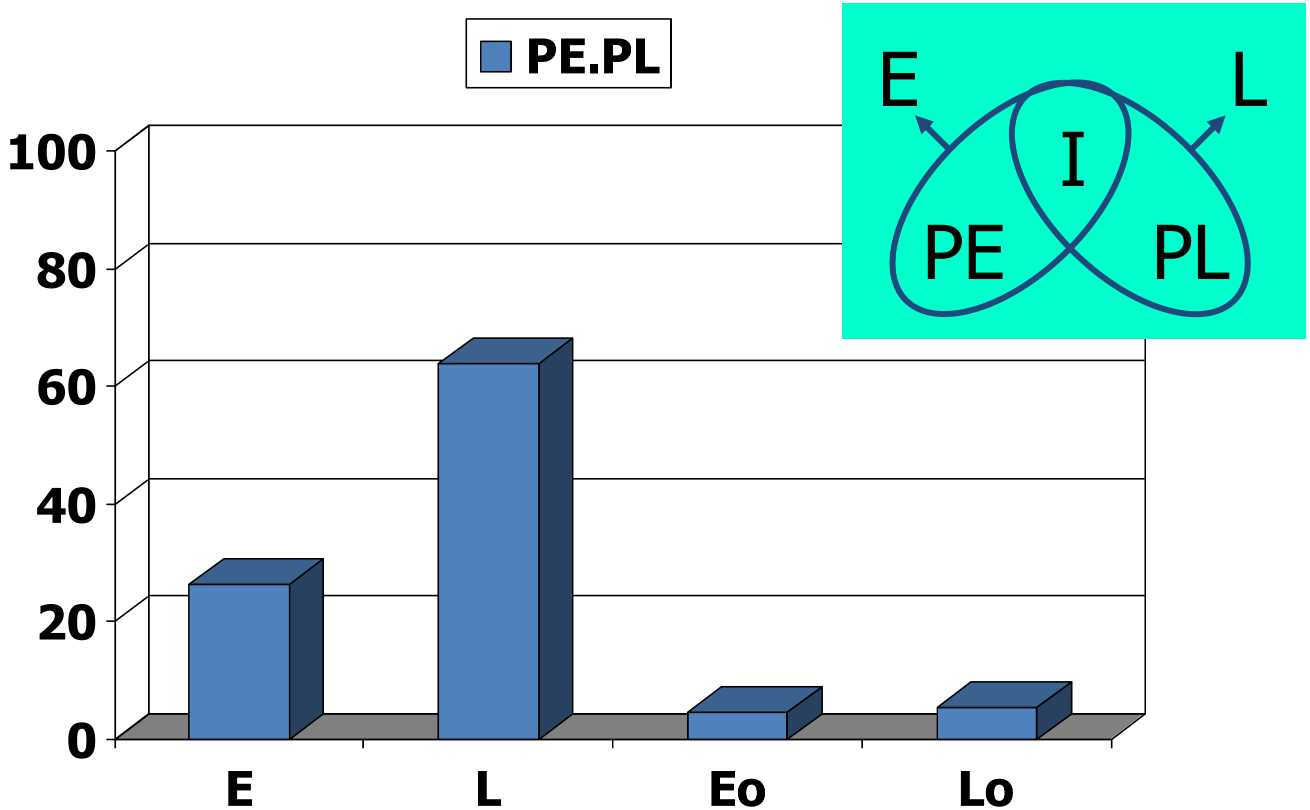
Highlighting: Results I.PL



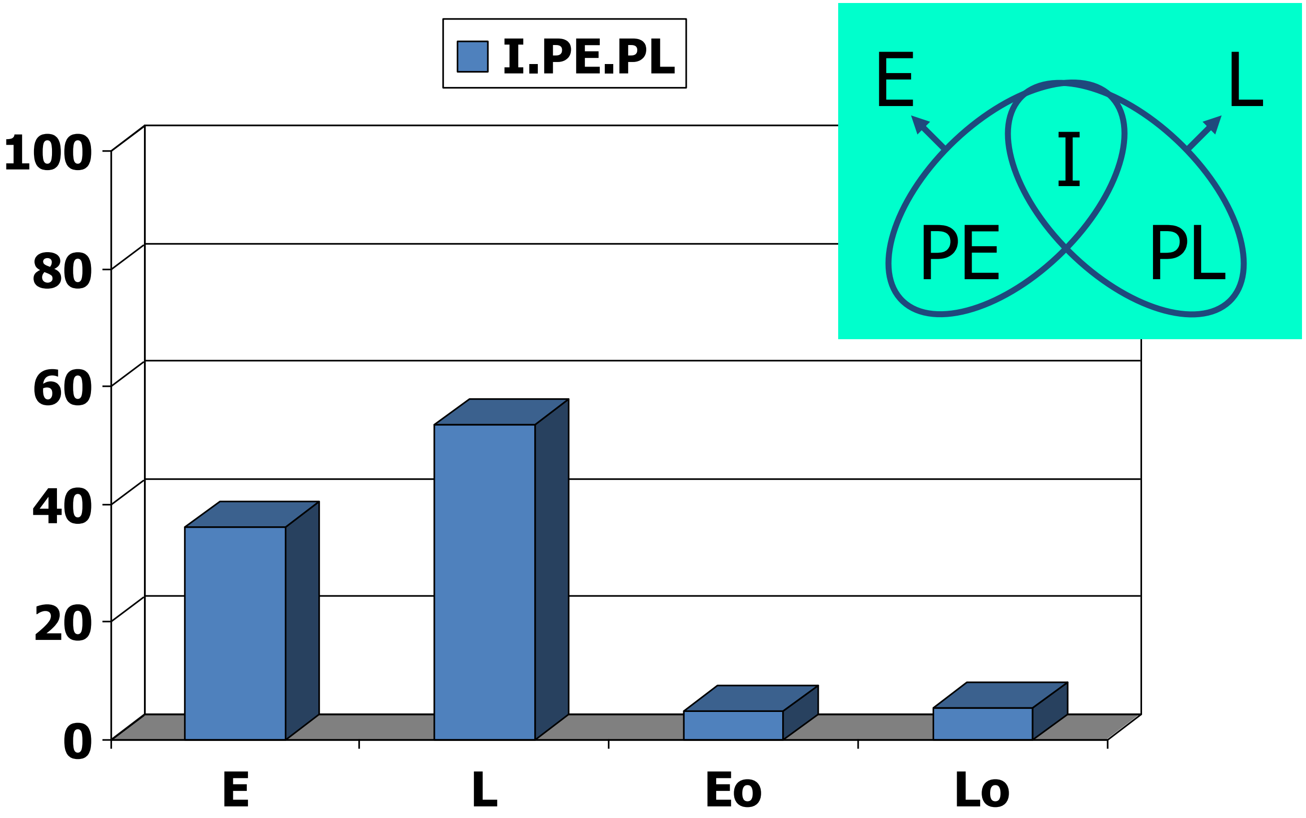
Highlighting: Results I



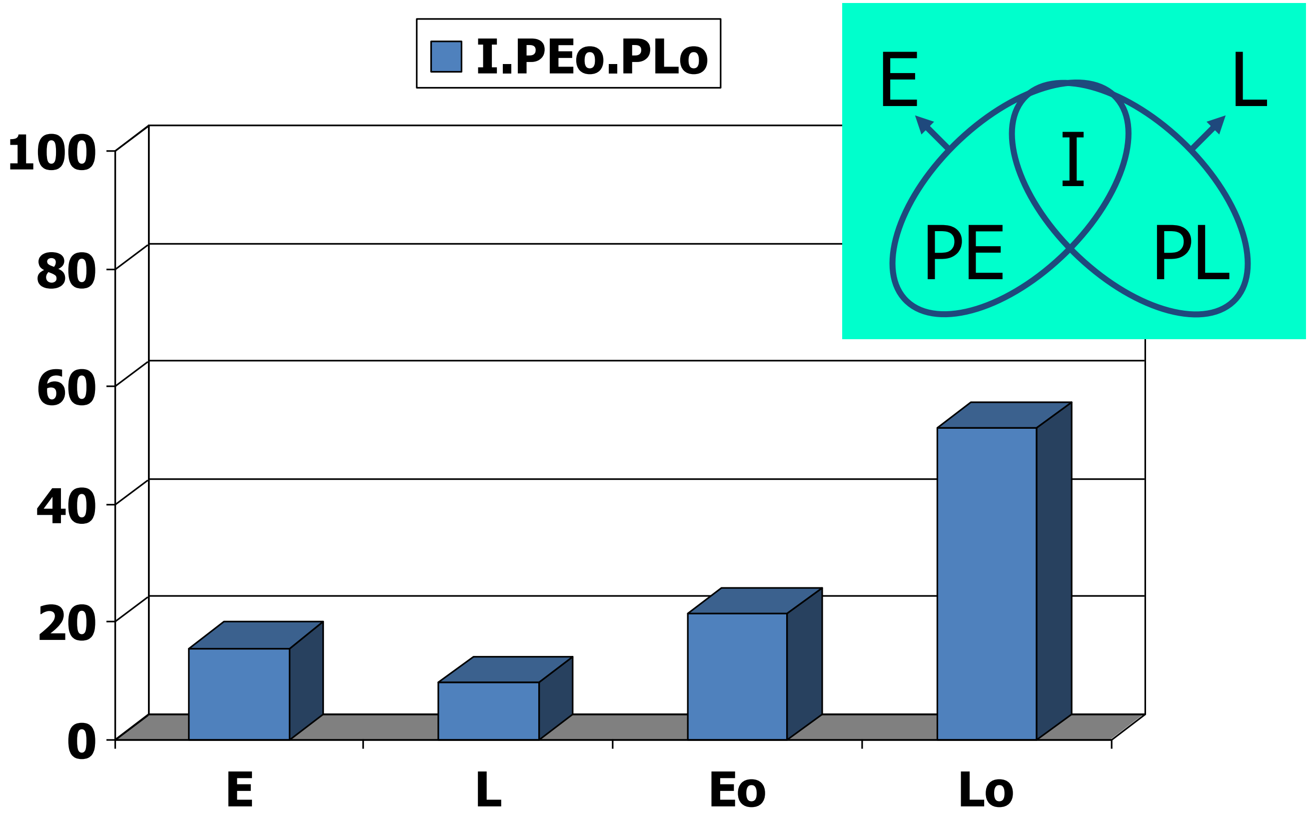
Highlighting: Results PE.PL



Highlighting: Results I.PE.PL



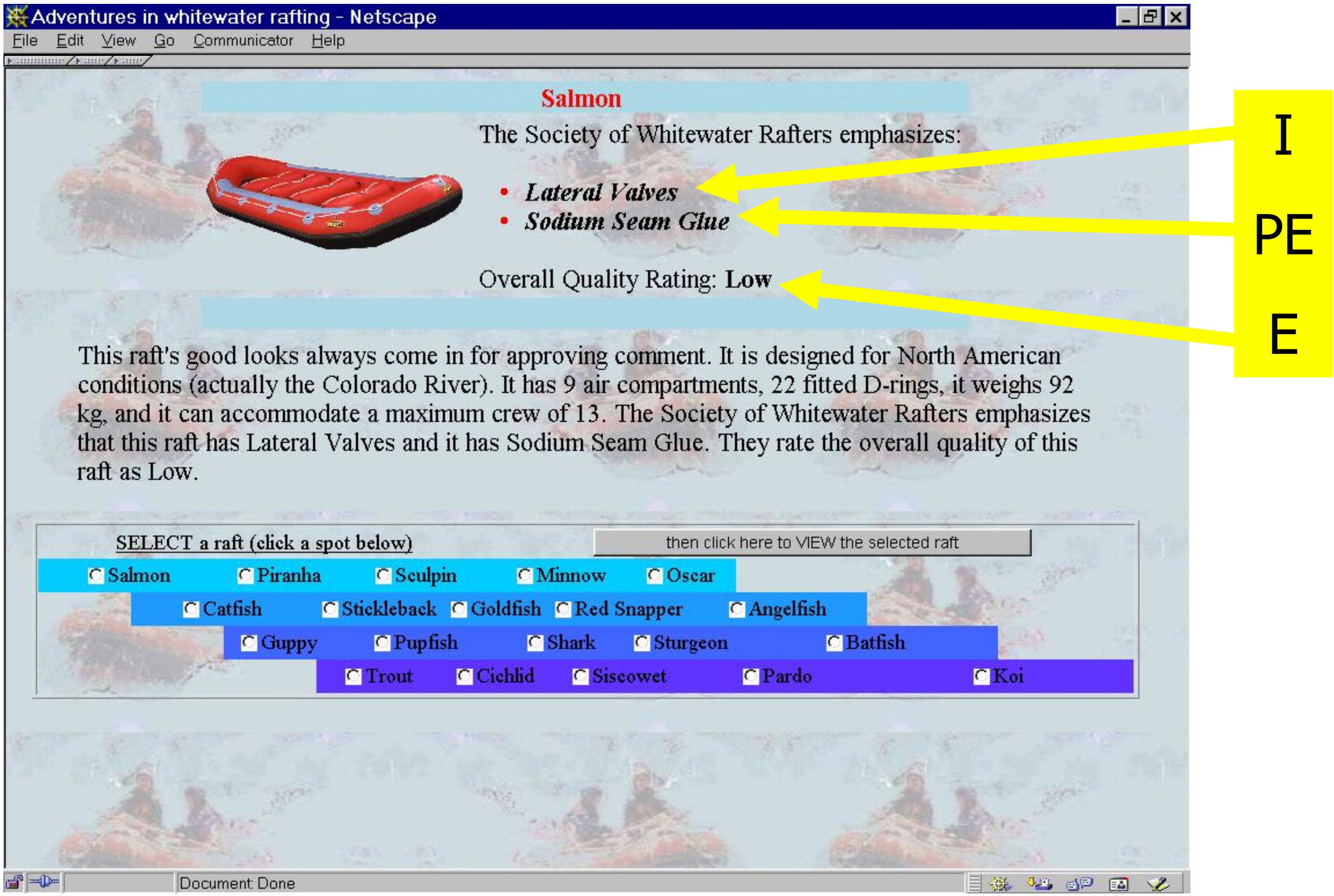
Highlighting: Results I.PEo.PLo



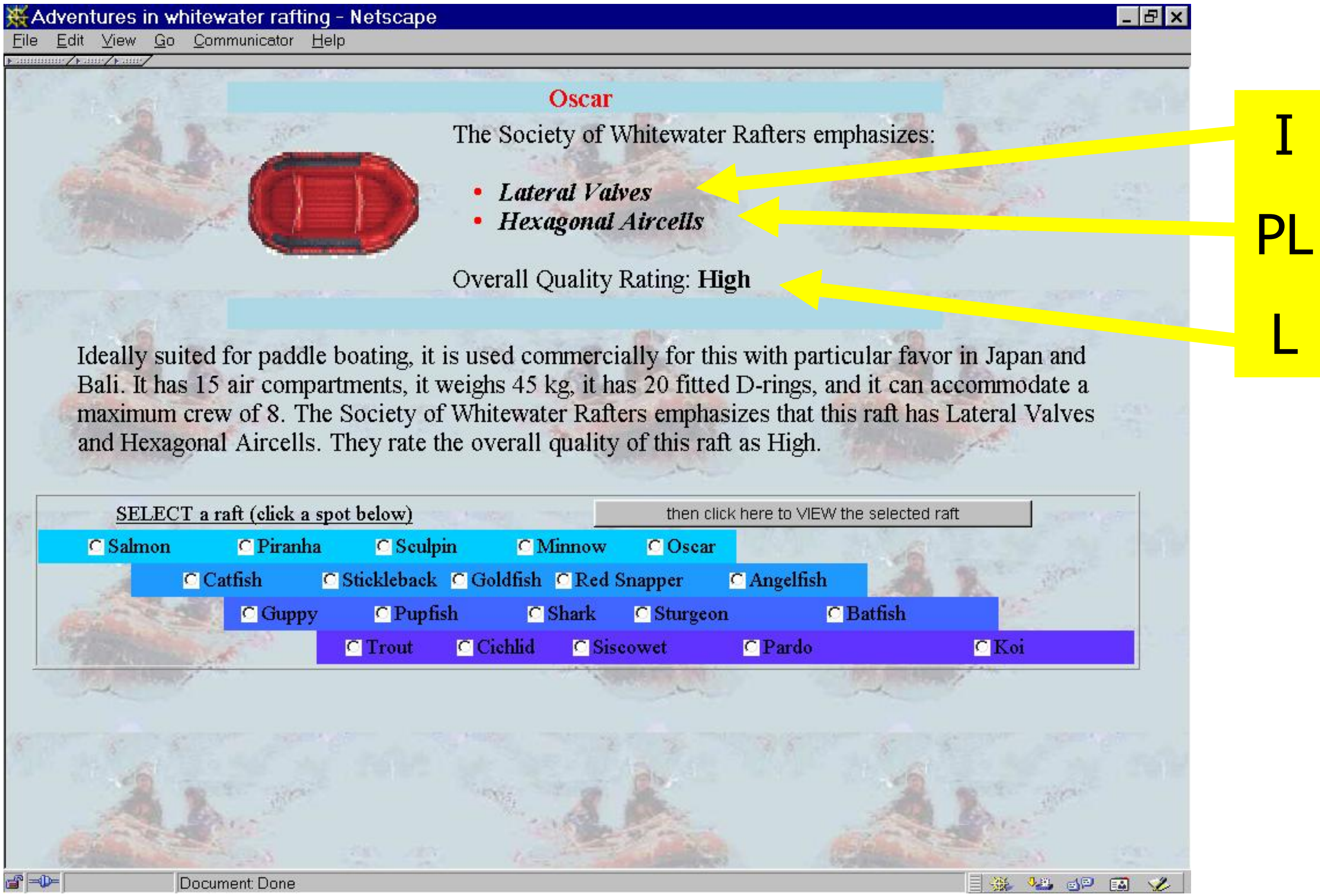
Not just for meaningless associations...

- Highlighting also happens in meaningful domains...

An Application:
Highlighting while web browsing.

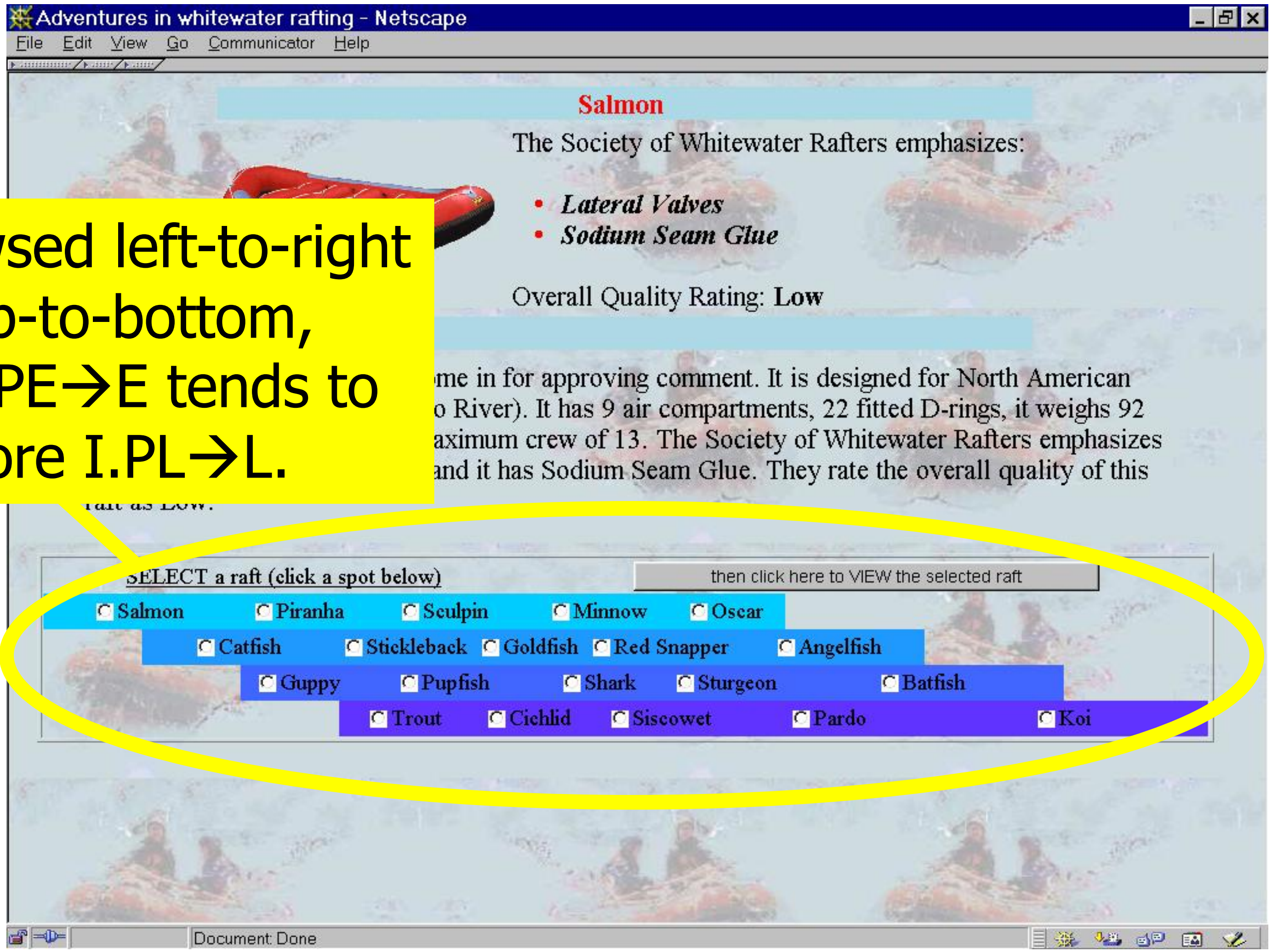


An Application:
Highlighting while web browsing.

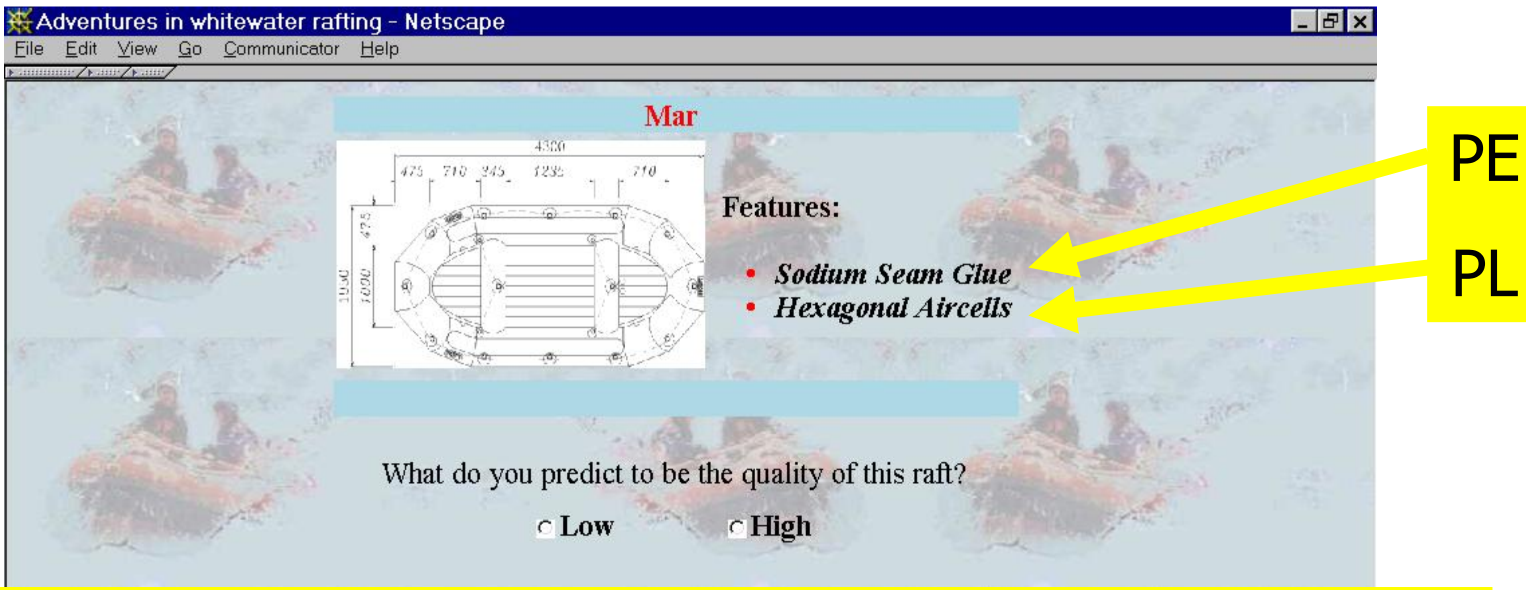


An Application:
Highlighting while web browsing.

If browsed left-to-right and top-to-bottom, then I.PE→E tends to be before I.PL→L.



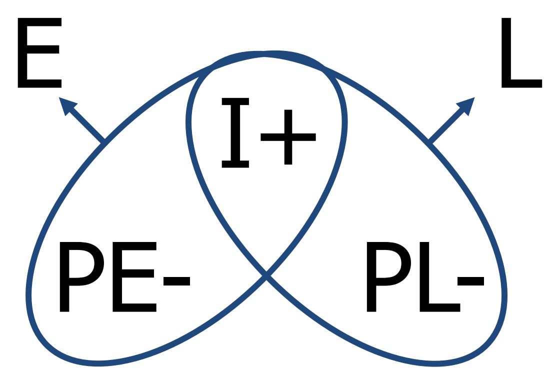
Test items



Results:
I yields strong preference for Early quality;
PE.PL yields strong preference for Later quality.

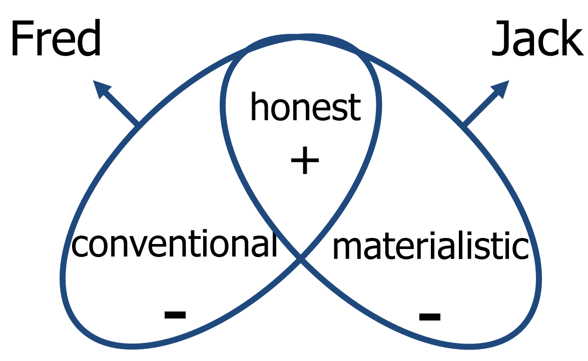
An Application:
Highlighting of personal attributes.

Early Training:	honest(+) & conventional(-) → Fred
Late Training:	honest(+) & conventional(-) → Fred honest(+) & materialistic(-) → Jack



An Application:
Highlighting of personal attributes.

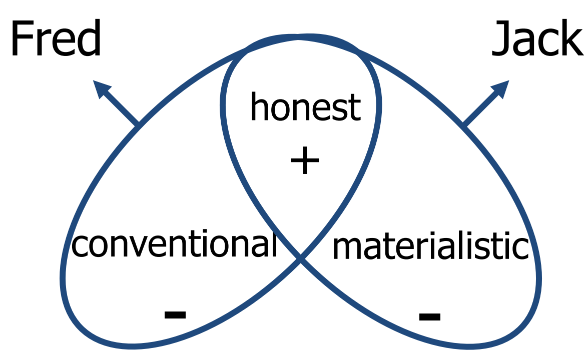
Early Training:	honest(+) & conventional(-) → Fred
Late Training:	honest(+) & conventional(-) → Fred honest(+) & materialistic(-) → Jack



An Application:
Highlighting of personal attributes.

Early Training:	honest(+) & conventional(-) → Fred
Late Training:	honest(+) & conventional(-) → Fred honest(+) & materialistic(-) → Jack

Likability: 6.47



Likability: 5.60

What causes highlighting?

- Can *your* favorite model of learning account for highlighting?
- How about various Bayesian approaches?
  - Only candidates are Bayesian approaches with sensitivity to time or trial order

Rational Model
(J. R. Anderson 1990)

- Representation:
  - There are internal clusters that represent subsets of training items.
  - Each cluster has its own set of Dirichlet distributions over beliefs about feature probabilities.
- Learning:
  - For each item presented, the item is assigned to the cluster that is most probable.
  - The Dirichlet parameters of that cluster are Bayesian updated.

Rational Model Does Not Show Highlighting:

Rational Model, c=0.3

Data entered:

PE	I	PL	E
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
0	1	1	0
1	1	0	1
1	1	0	1
1	1	0	1
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
1	1	0	1
0	1	1	0
0	1	1	0
0	1	1	0
1	1	0	1
0	1	1	0
0	1	1	0
0	1	1	0
1	1	0	1

Internal Clusters
Dirichlet Parameters

Cluster 1:

1	1	12	1
12	12	1	12

Cluster 2:

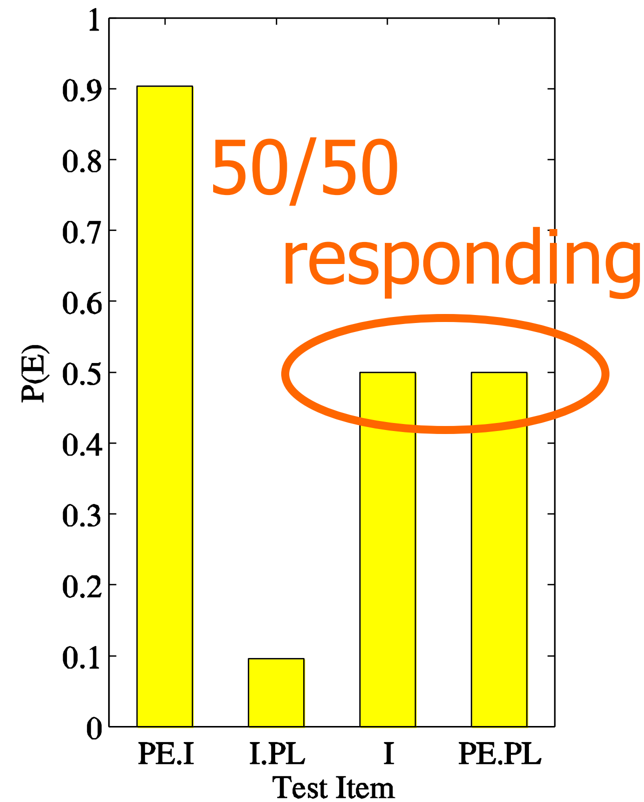
12	1	1	12
1	12	12	1

Cluster 3:

1	1	1	1
1	1	1	1

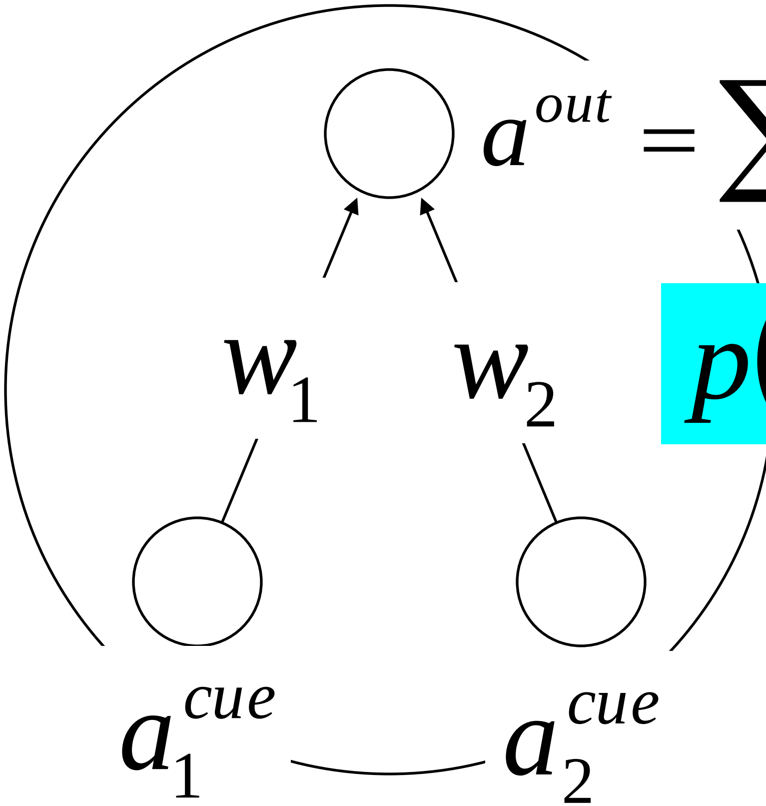
- Cluster parameters are symmetric.

Overt Behavior



Kalman Filter
(Sutton 1992; Dayan, Kakade et al. 2000+)

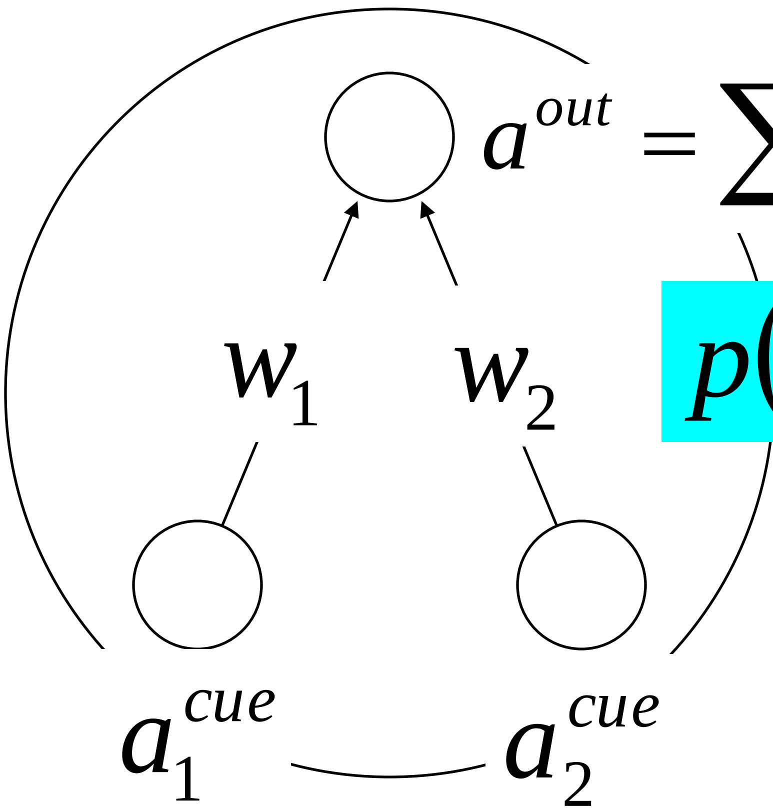
$$p(t) \sim N\left(t \mid a^{out}, v\right)$$



$$a^{out} = \sum_i w_i a_i^{cue}$$

$$p(\vec{w}) \sim N\left(\vec{w} \mid \vec{\mu}, C\right)$$

Kalman Filter Updating:
Step 1. Linear Dynamics



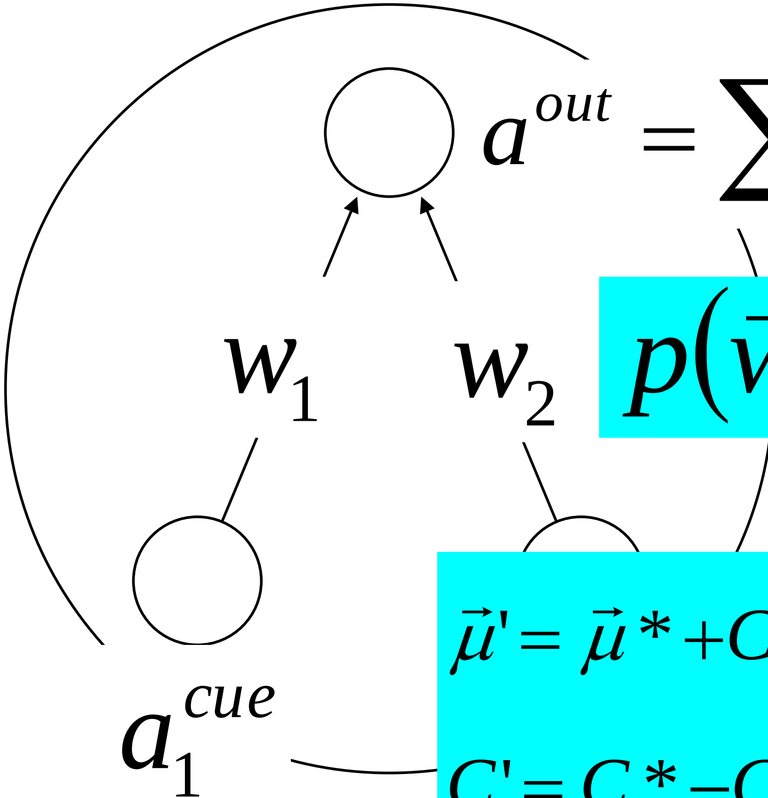
$$a^{out} = \sum_i w_i a_i^{cue}$$

$$p(\vec{w}) \sim N(\vec{w} \mid \vec{\mu}, C)$$

$$\vec{\mu}^* = D\vec{\mu}$$

$$C^* = DCD^T + U$$

Kalman Filter Updating:
Step 2. Bayesian Learning



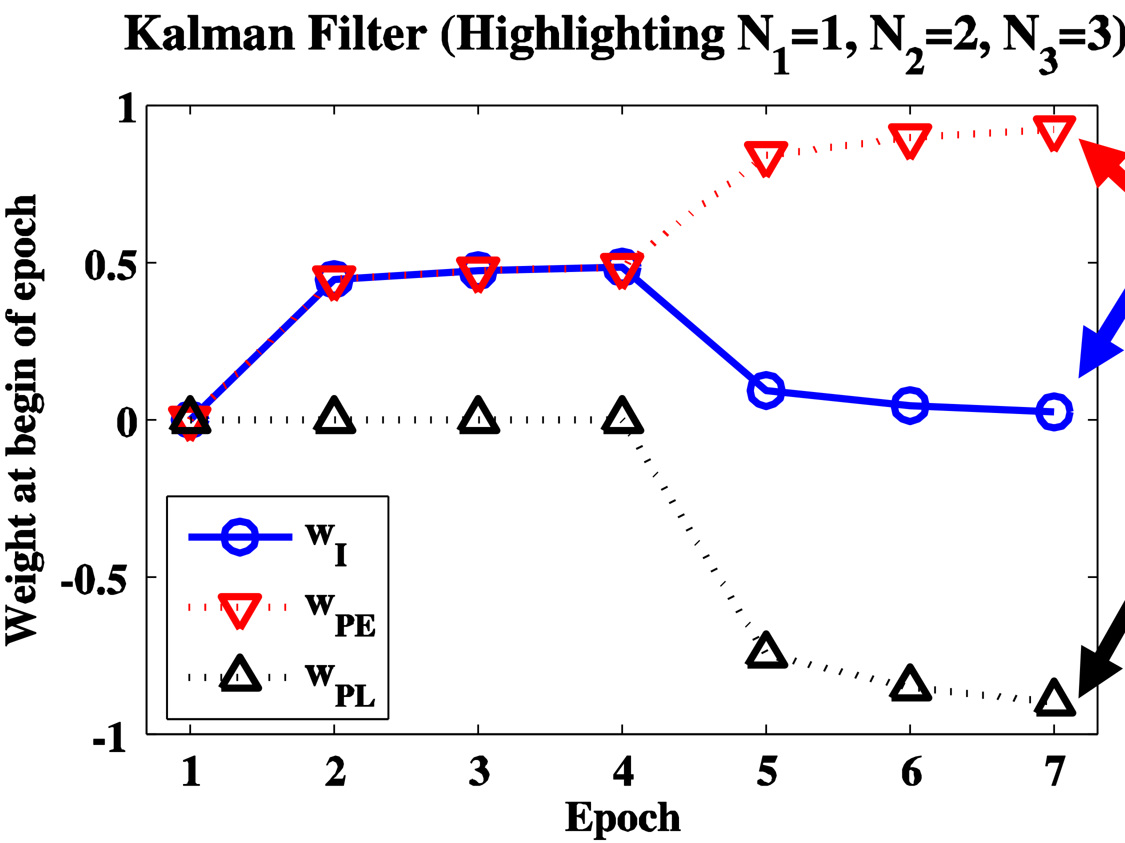
$$a^{out} = \sum_i w_i a_i^{cue}$$

$$p(\vec{w}) \sim N(\vec{w} \mid \vec{\mu}^*, C^*)$$

$$\vec{\mu}' = \vec{\mu}^* + C^* \vec{a}^{cue} \left[v + \vec{a}^{cue^T} C^* \vec{a}^{cue} \right]^{-1} \left(t - \vec{a}^{cue^T} \vec{\mu} \right)$$

$$C' = C^* - C^* \vec{a}^{cue} \left[v + \vec{a}^{cue^T} C^* \vec{a}^{cue} \right]^{-1} \vec{a}^{cue^T} C^*$$

Kalman Filter Does Not Show Highlighting:

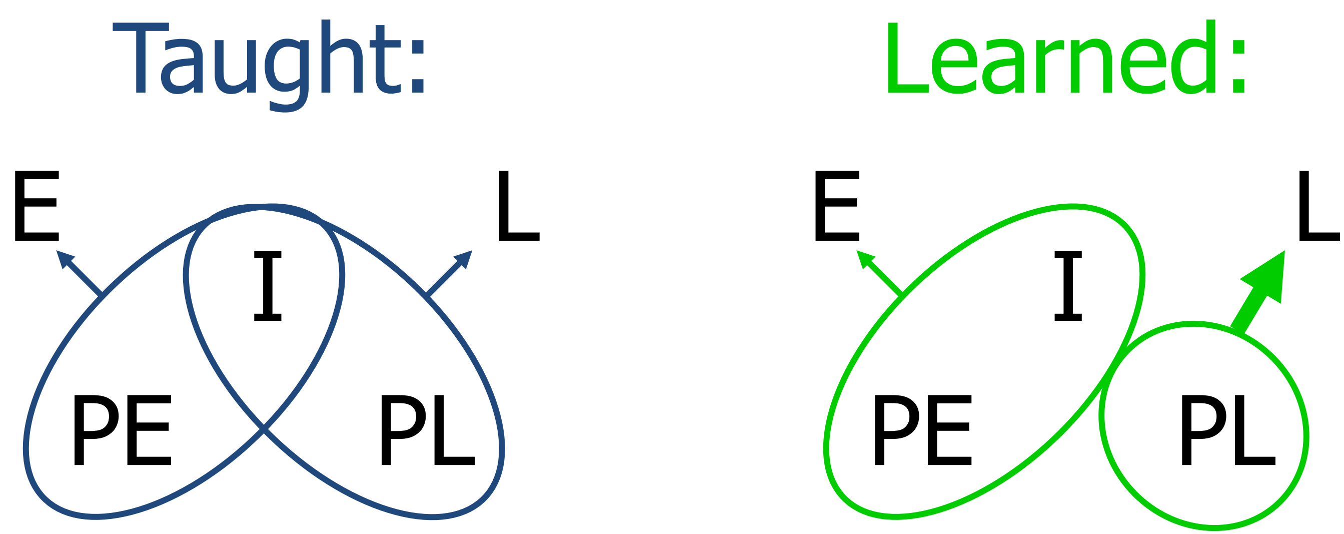


Symmetric weights:

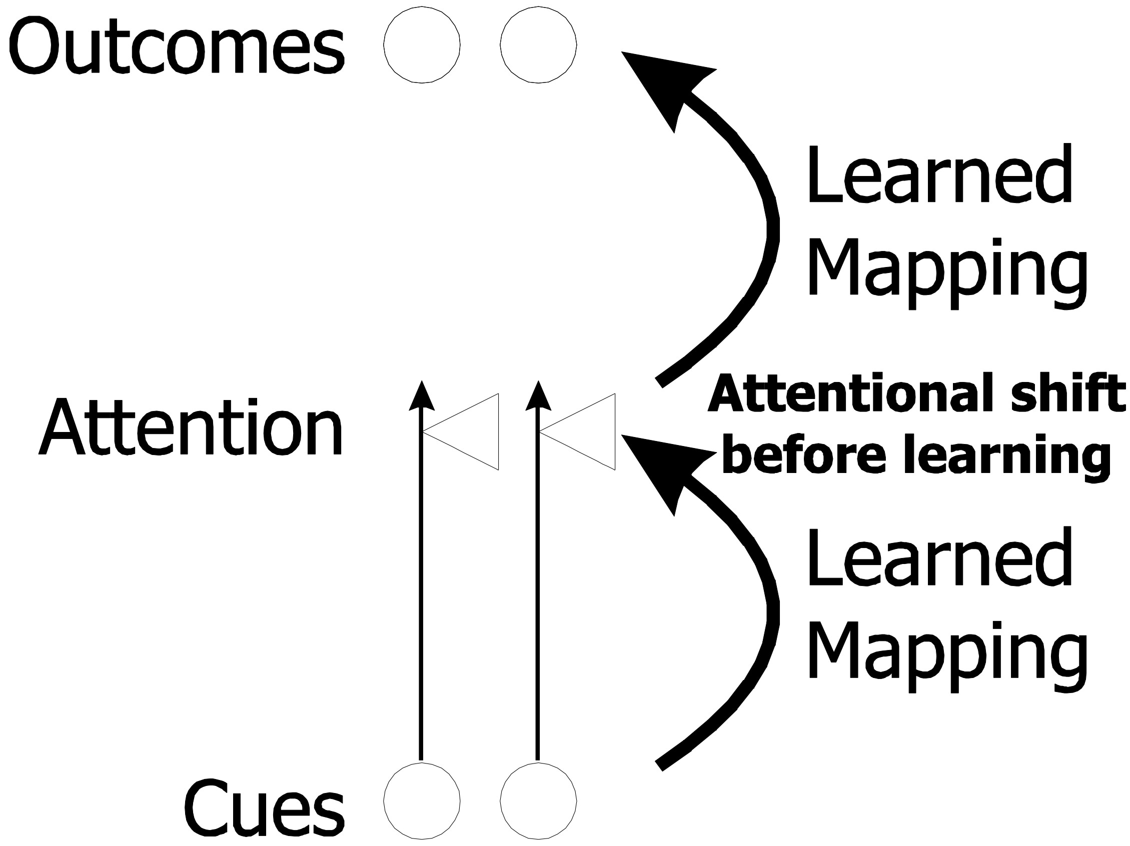
- Weight from cue I is near zero.
- Weights from PE and PL are equal and opposite.

Explanation of Highlighting:

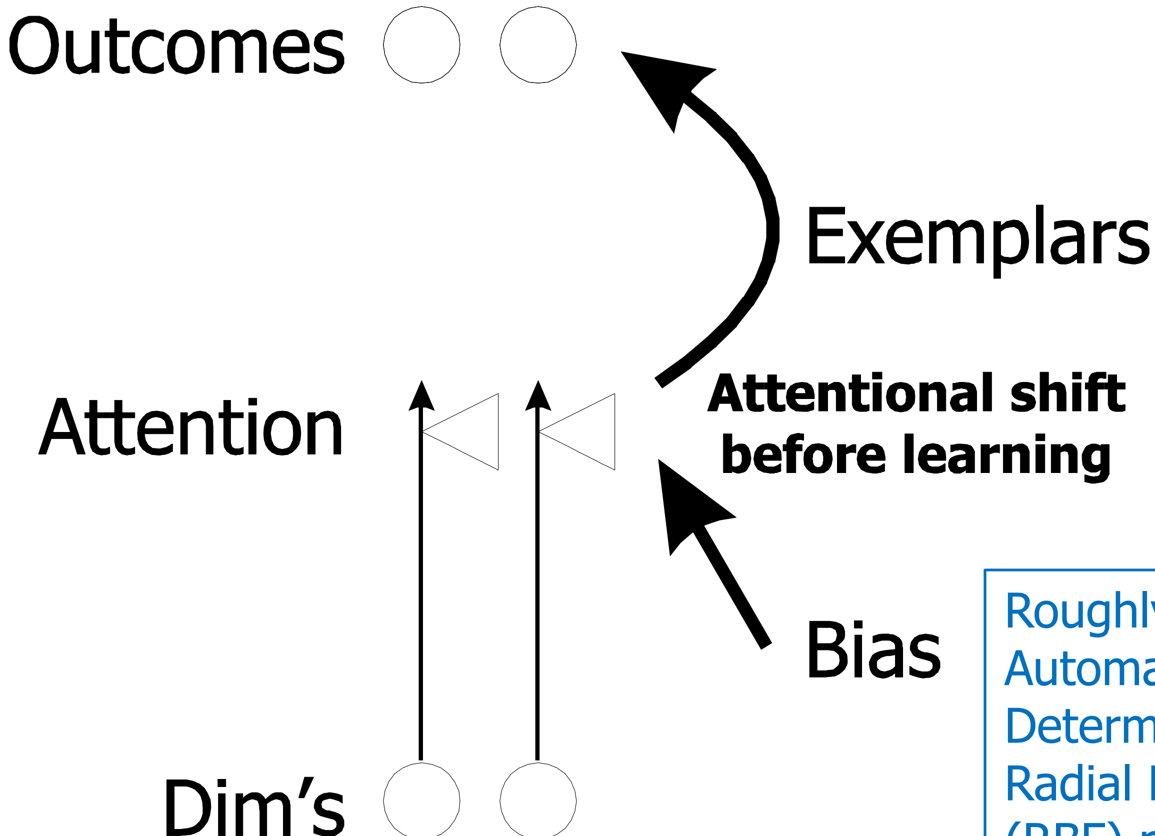
- Attention rapidly shifts to the distinctive feature of the later learned outcome.



Models of Attention Shifting: General Framework



Models of Attention Shifting: RASHNL (/ALCOVE)

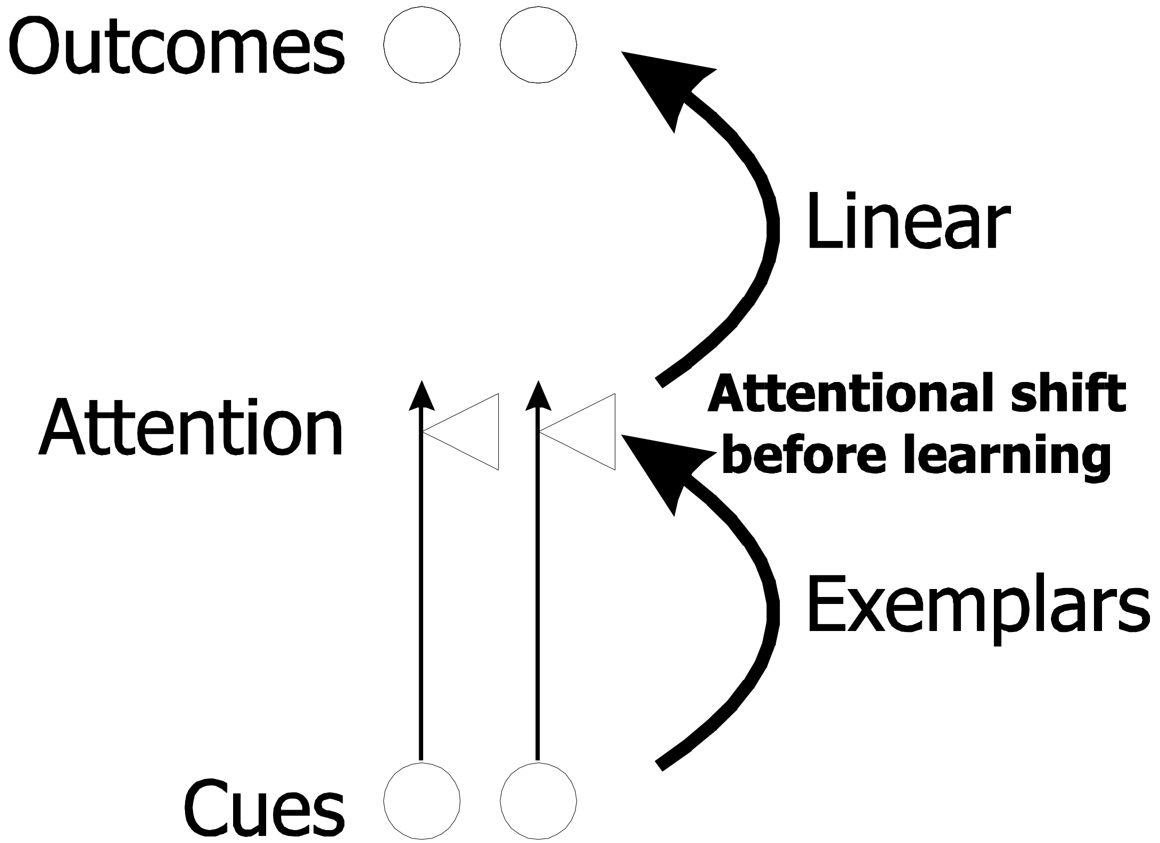


Roughly analogous to Automatic Relevance Determination (ARD) in Radial Basis Function (RBF) networks.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. ***Psychological Review***, 99, 22-44.

Kruschke, J. K. & Johansen, M. K. (1999). A model of probabilistic category learning. ***Journal of Experimental Psychology: Learning, Memory and Cognition***, 25, 1083-1119.

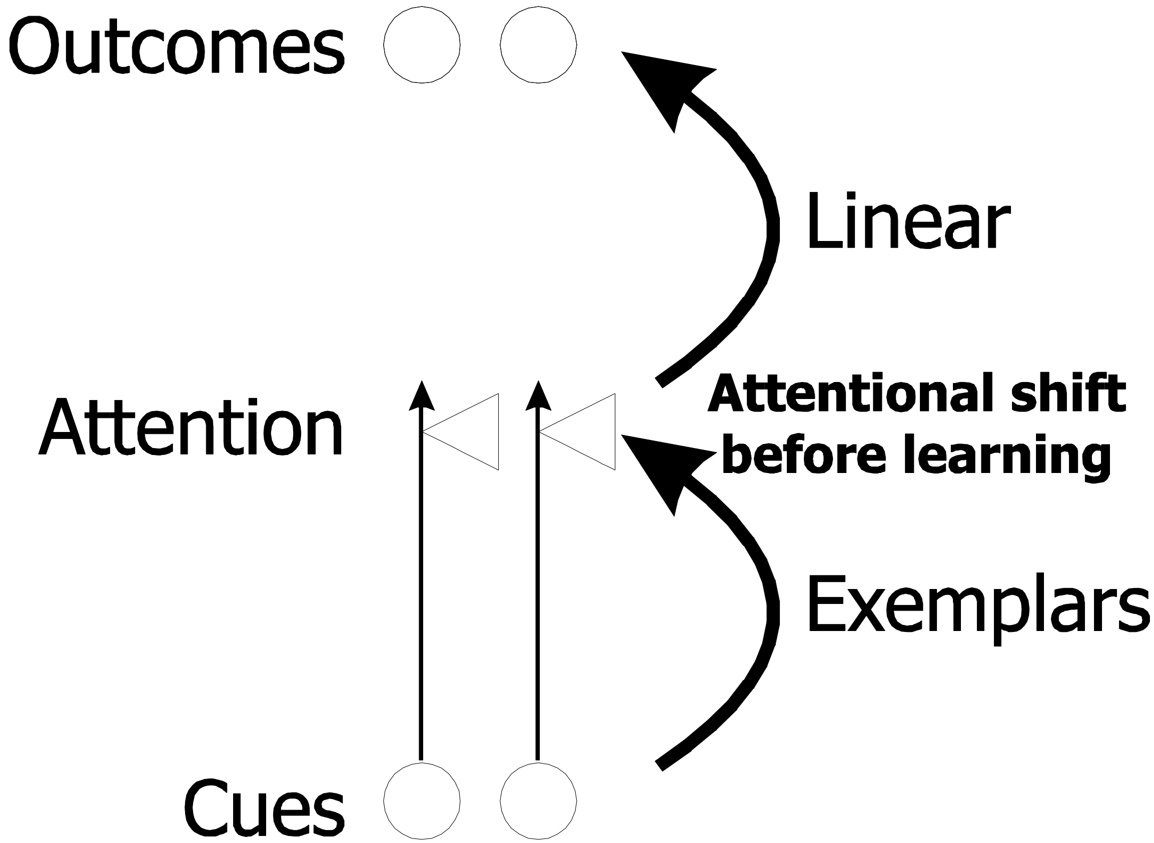
Models of Attention Shifting: EXIT (/ADIT)



Kruschke, J. K. (1996). Base rates in category learning. ***Journal of Experimental Psychology: Learning, Memory and Cognition***, 22, 3-26.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. ***Journal of Mathematical Psychology***, 45, 812-863.

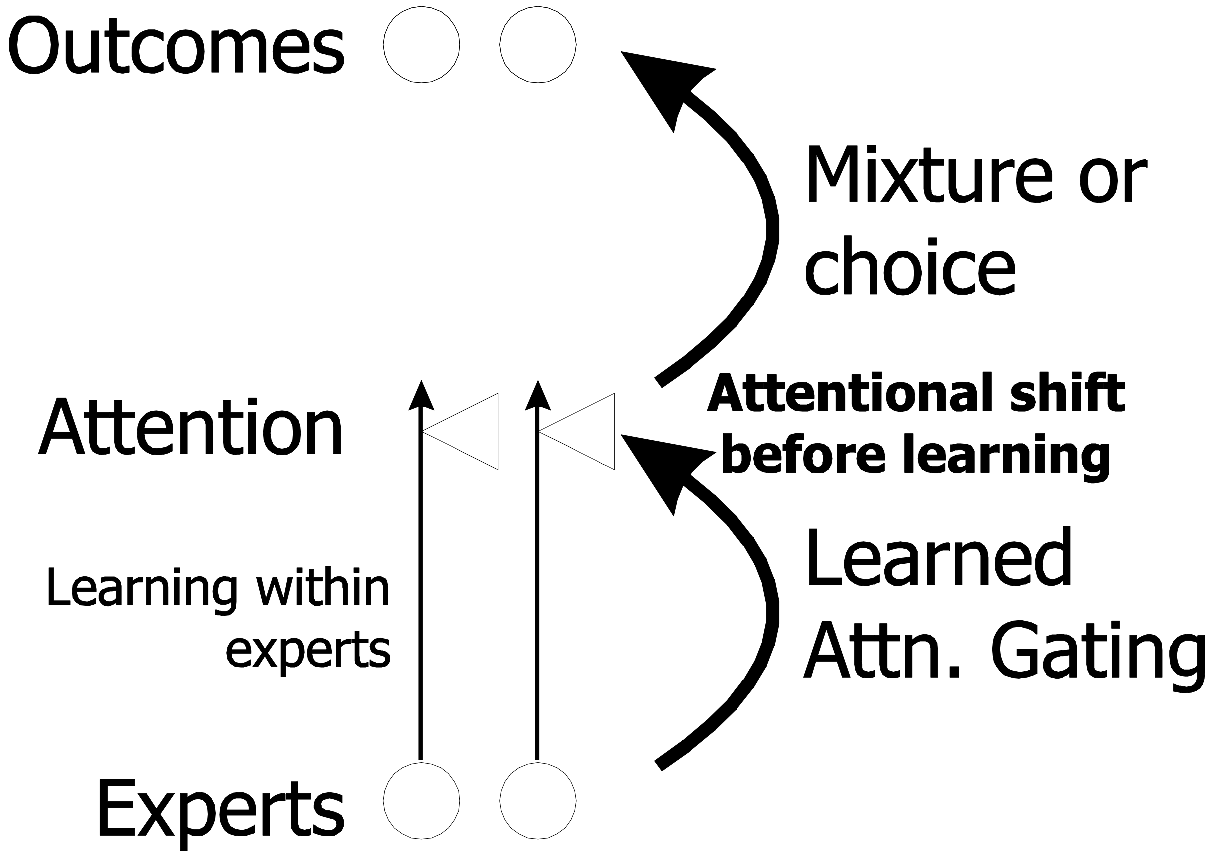
Models of Attention Shifting: EXIT (/ADIT)



Kruschke, J. K. (1996). Base rates in category learning. ***Journal of Experimental Psychology: Learning, Memory and Cognition***, 22, 3-26.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. ***Journal of Mathematical Psychology***, 45, 812-863.

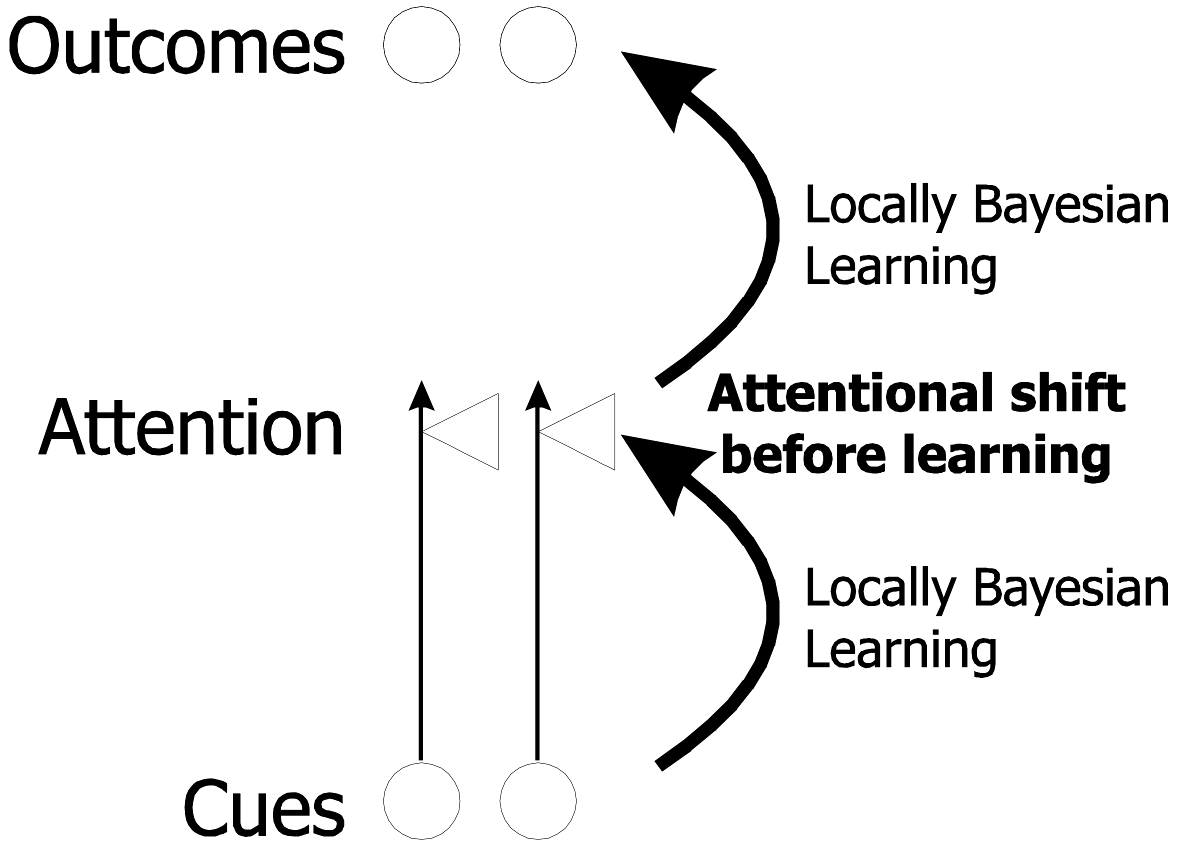
Models of Attention Shifting: ATRIUM & POLE



Kalish, M. L., Lewandowsky, S., and Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. **Psychological Review**, 111(4), 1072-1099.

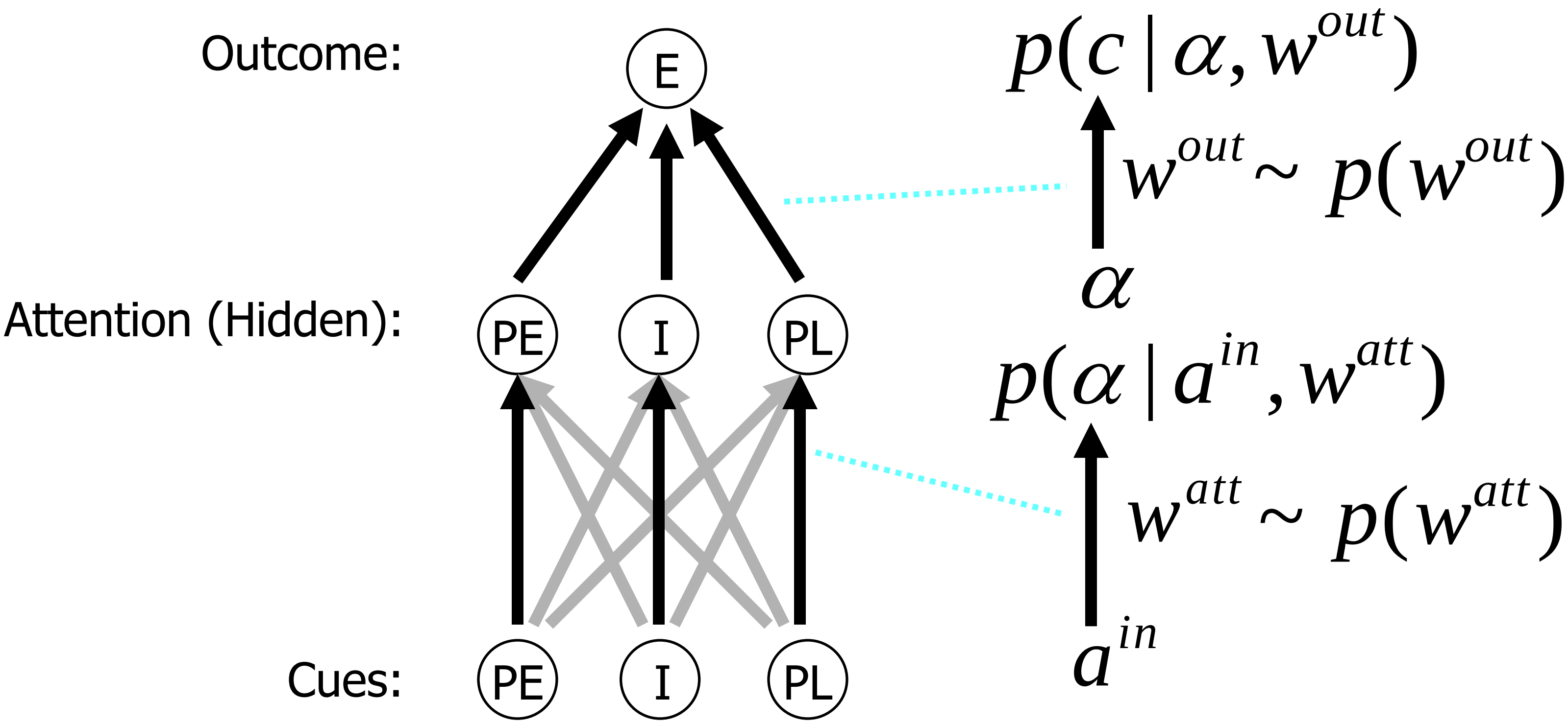
Erickson, M. A. & Kruschke, J. K. (1998). Rules and Exemplars in Category Learning. **Journal of Experimental Psychology: General**, 127, 107-140.

Models of Attention Shifting: Locally Bayesian

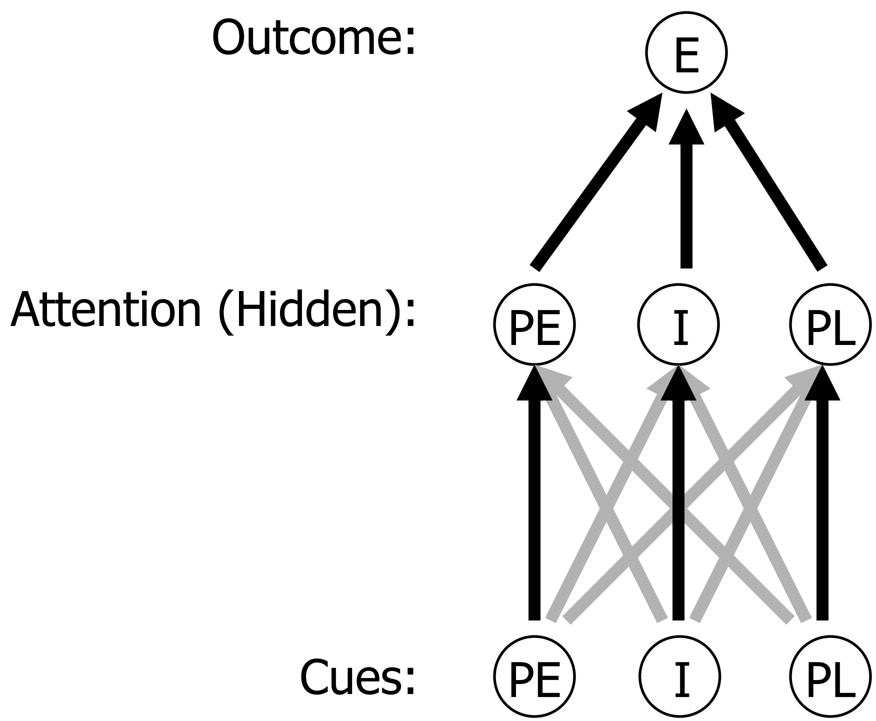


Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. **Psychological Review, 113,** 677-699.

Locally Bayesian Learning Implemented in an Attentional Learning Model



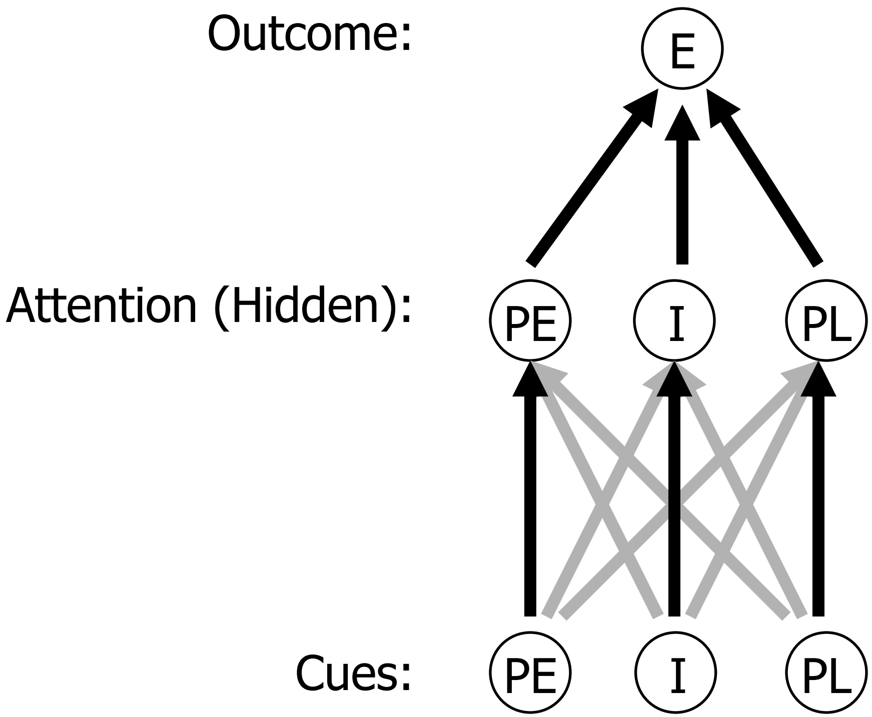
Locally Bayesian Learning Implemented in an Attentional Learning Model



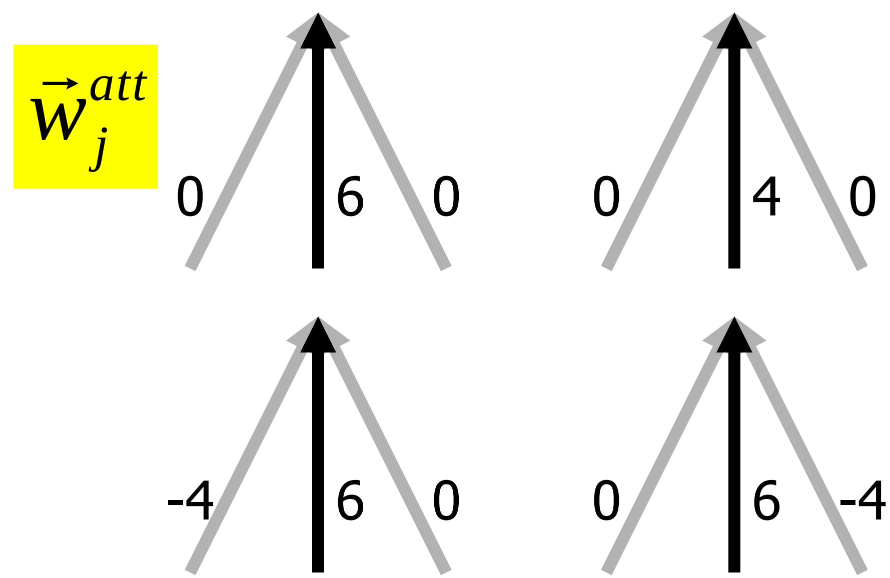
$$p(\alpha_j = 1) = sig(\vec{w}_j^{att} \vec{a}^{in})^6$$

$$a_i^{in} = \begin{cases} 1 \text{ if cue is present} \\ 0 \text{ otherwise} \end{cases}$$

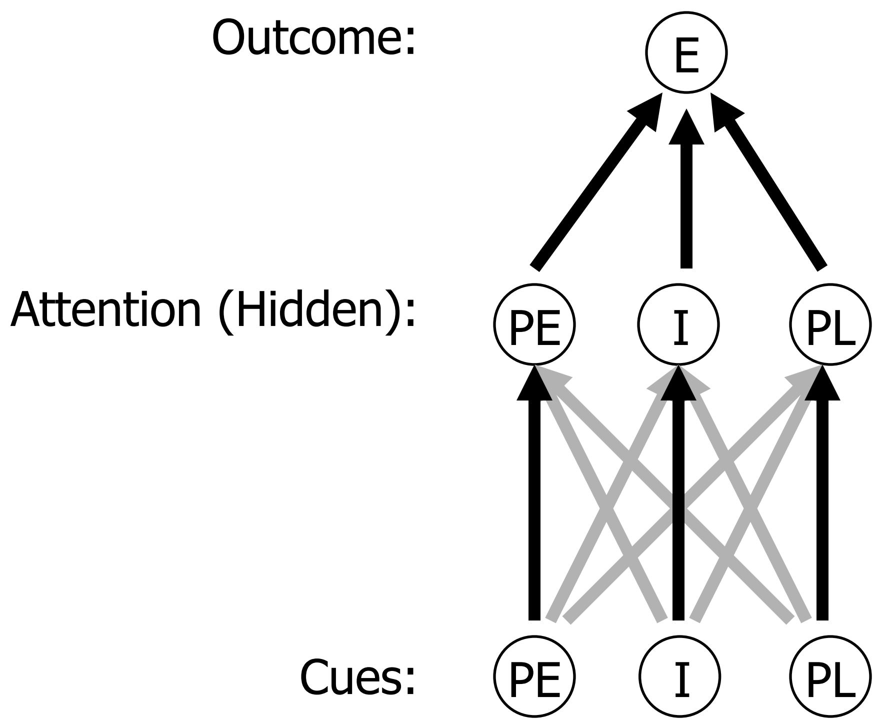
Locally Bayesian Learning Implemented in an Attentional Learning Model



Hidden activations are attentionally filtered copies of input activations.

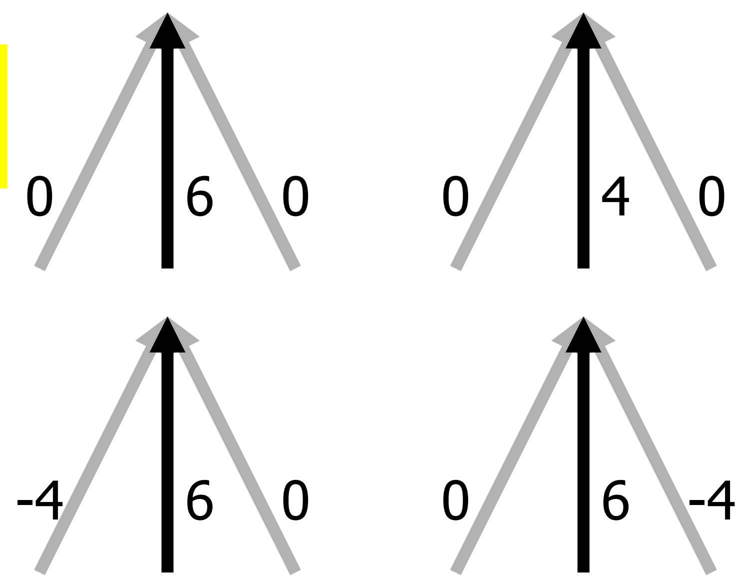


Locally Bayesian Learning Implemented in an Attentional Learning Model

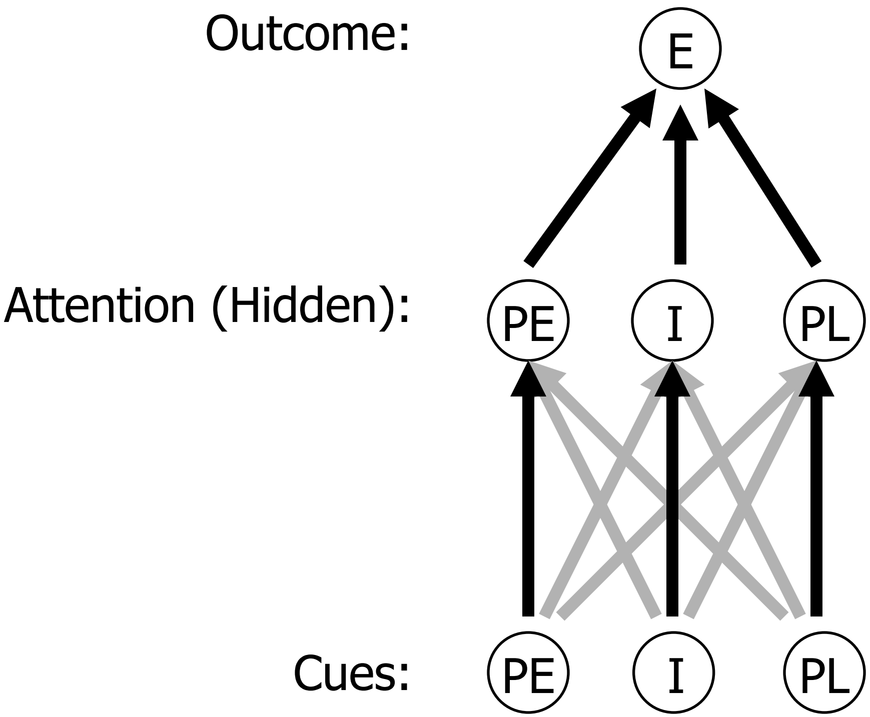


Each combination of weights constitutes a hypothesis. They are symmetrically distributed with uniform prior.

$\vec{w}_j^{att}$



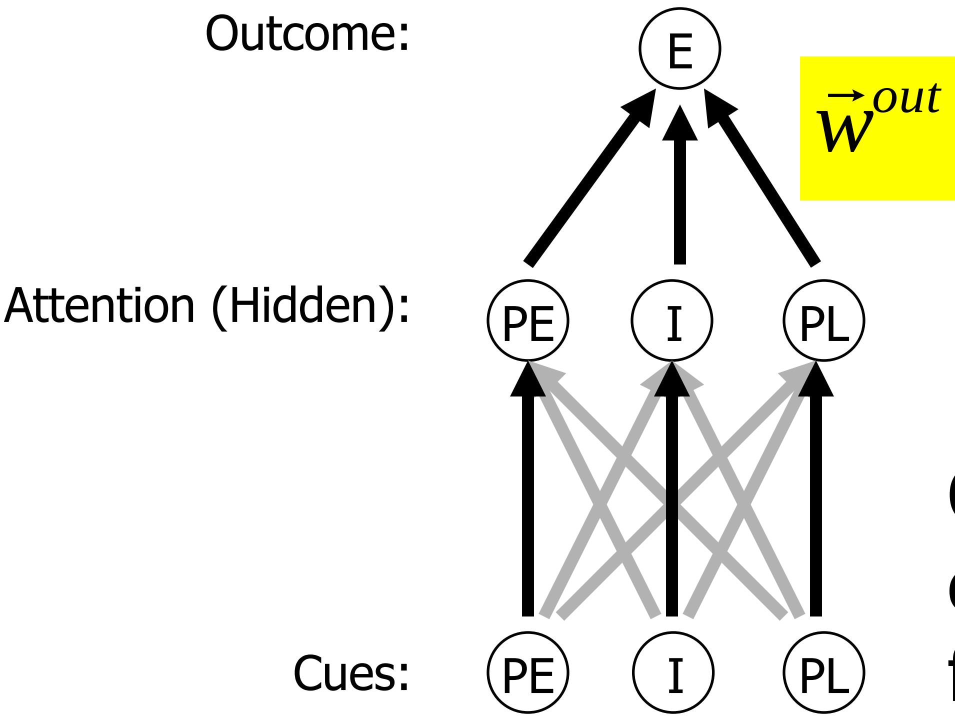
Locally Bayesian Learning Implemented in an Attentional Learning Model

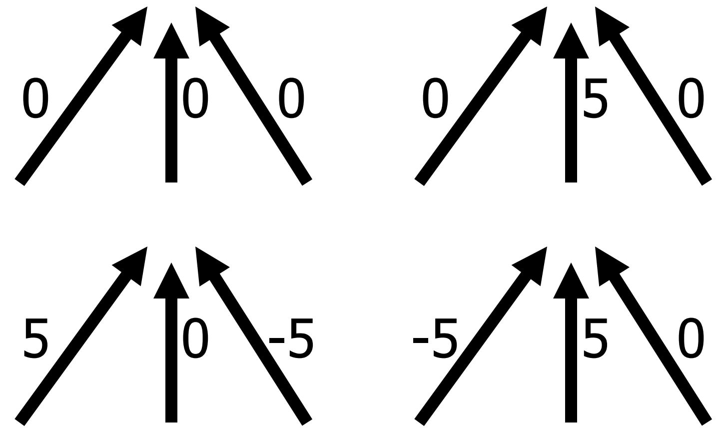


$$p(E=1) = sig(\vec{w}^{out}\vec{\alpha})$$

$$\hat{\alpha}_j = \sum_{\alpha \in \{0,1\}} \alpha \, p(\alpha_j = \alpha)$$

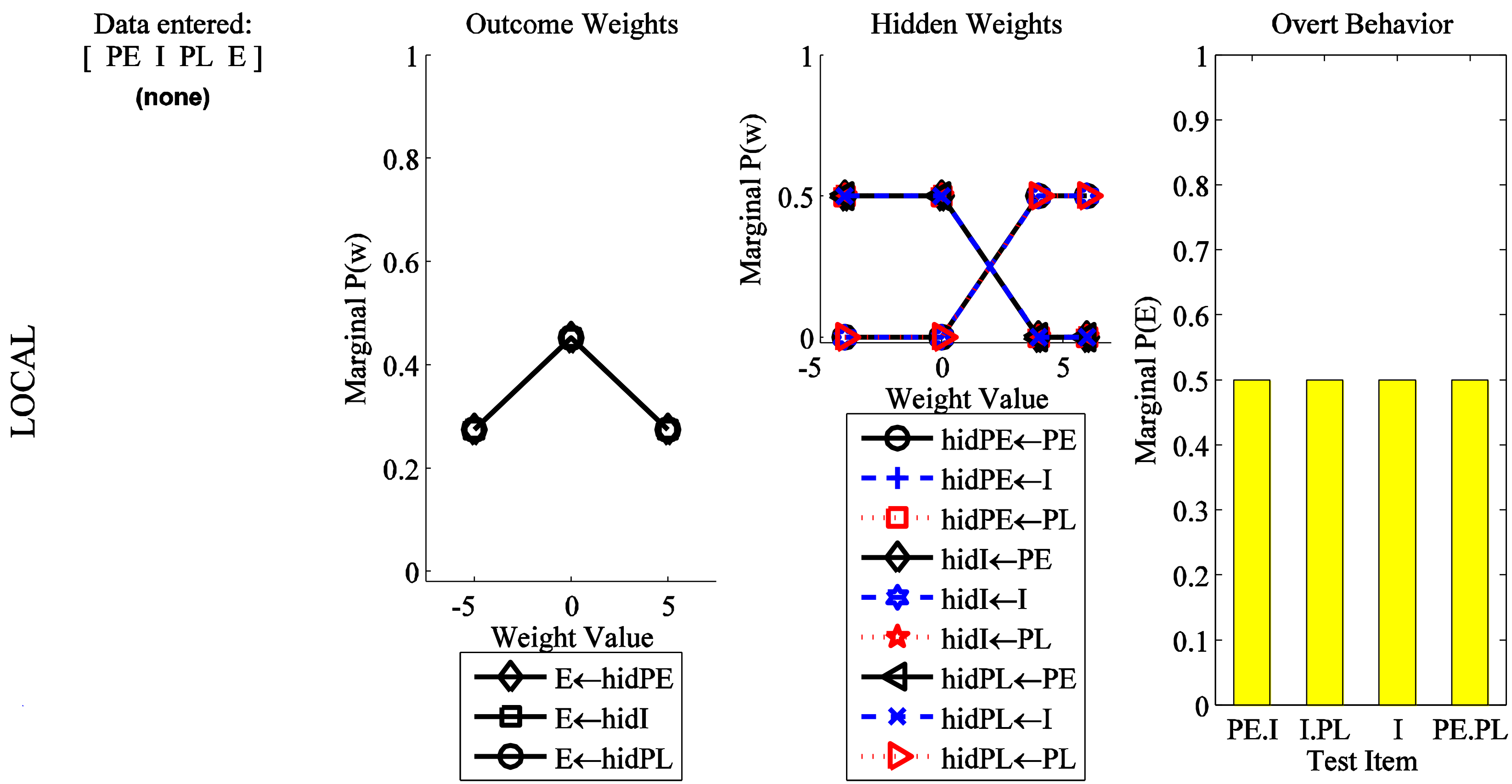
Locally Bayesian Learning Implemented in an Attentional Learning Model



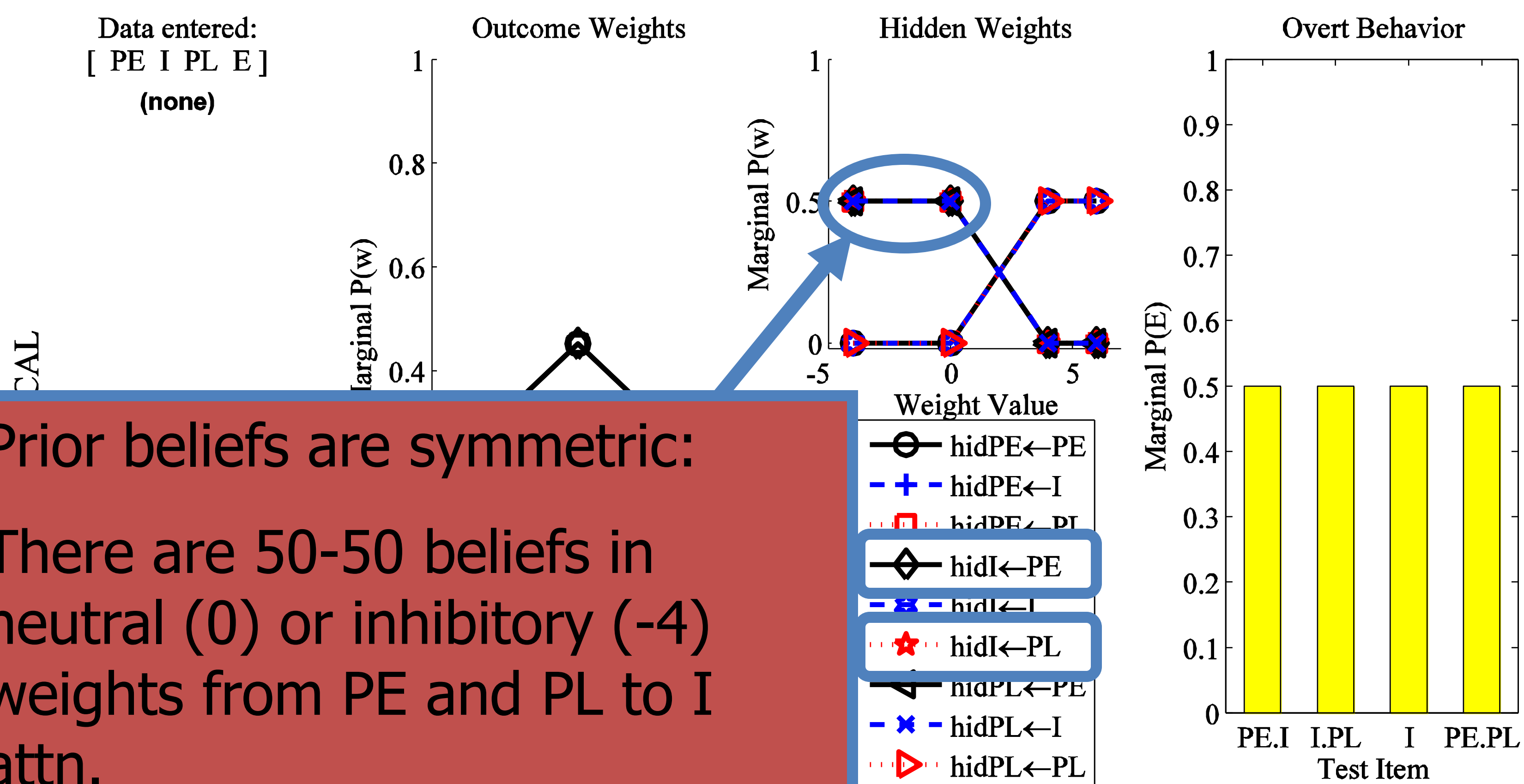


Outcome is arbitrary combination of cues. Prior favors all zeros; symmetrically distributed.

Highlighting: Prior Distribution



Highlighting: Prior Distribution



Prior beliefs are symmetric:

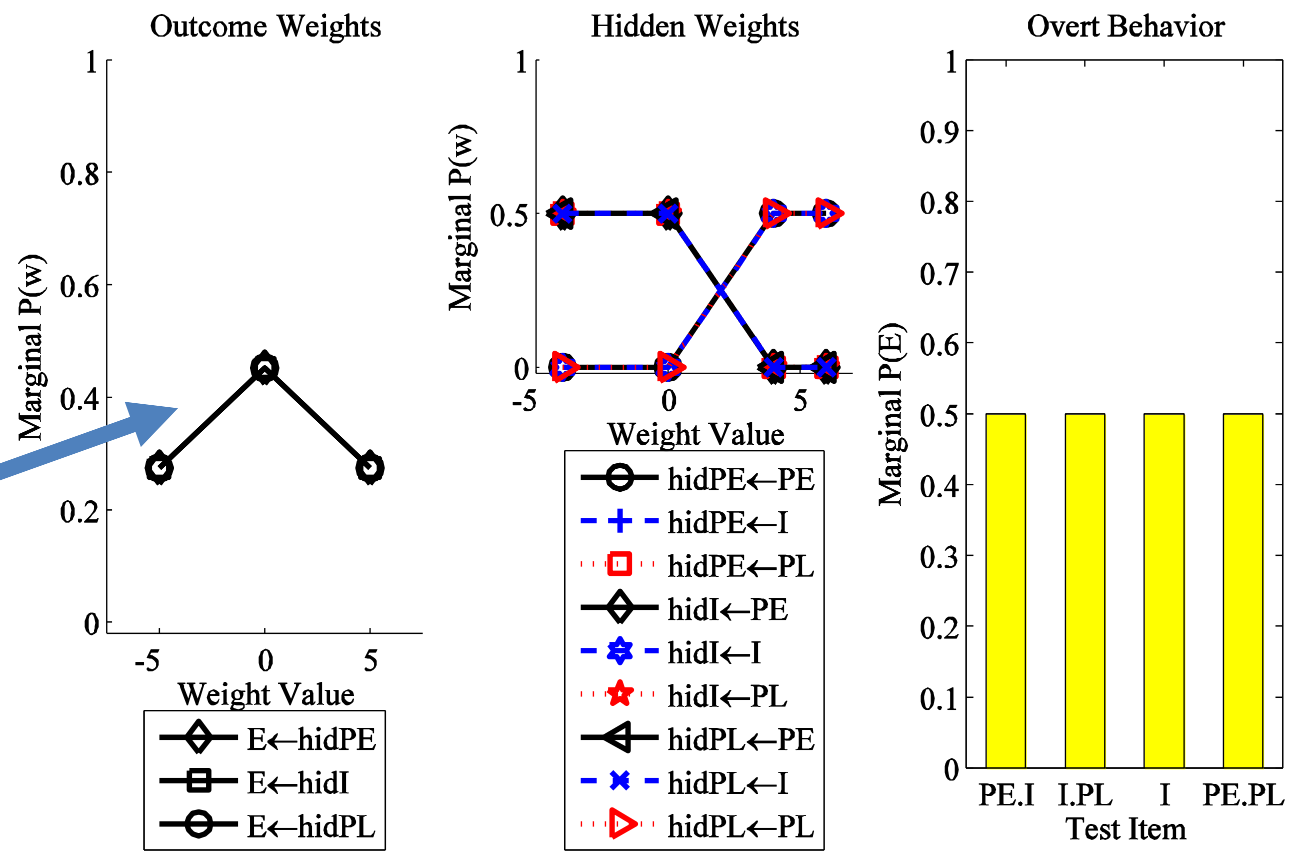
There are 50-50 beliefs in neutral (0) or inhibitory (-4) weights from PE and PL to I attn.

Highlighting: Prior Distribution

Data entered:
[PE I PL E]
(none)

Prior beliefs are symmetric:

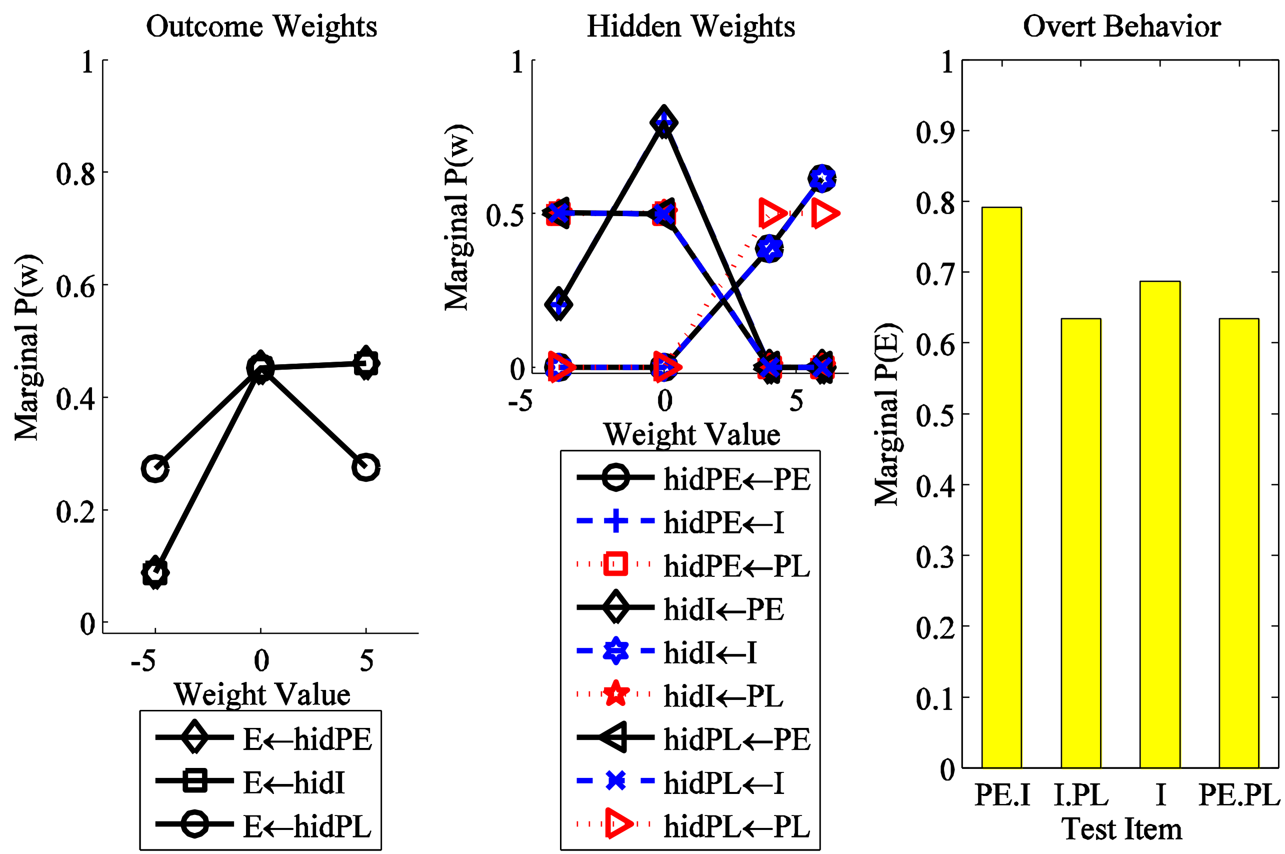
Beliefs about all cues are neutral.



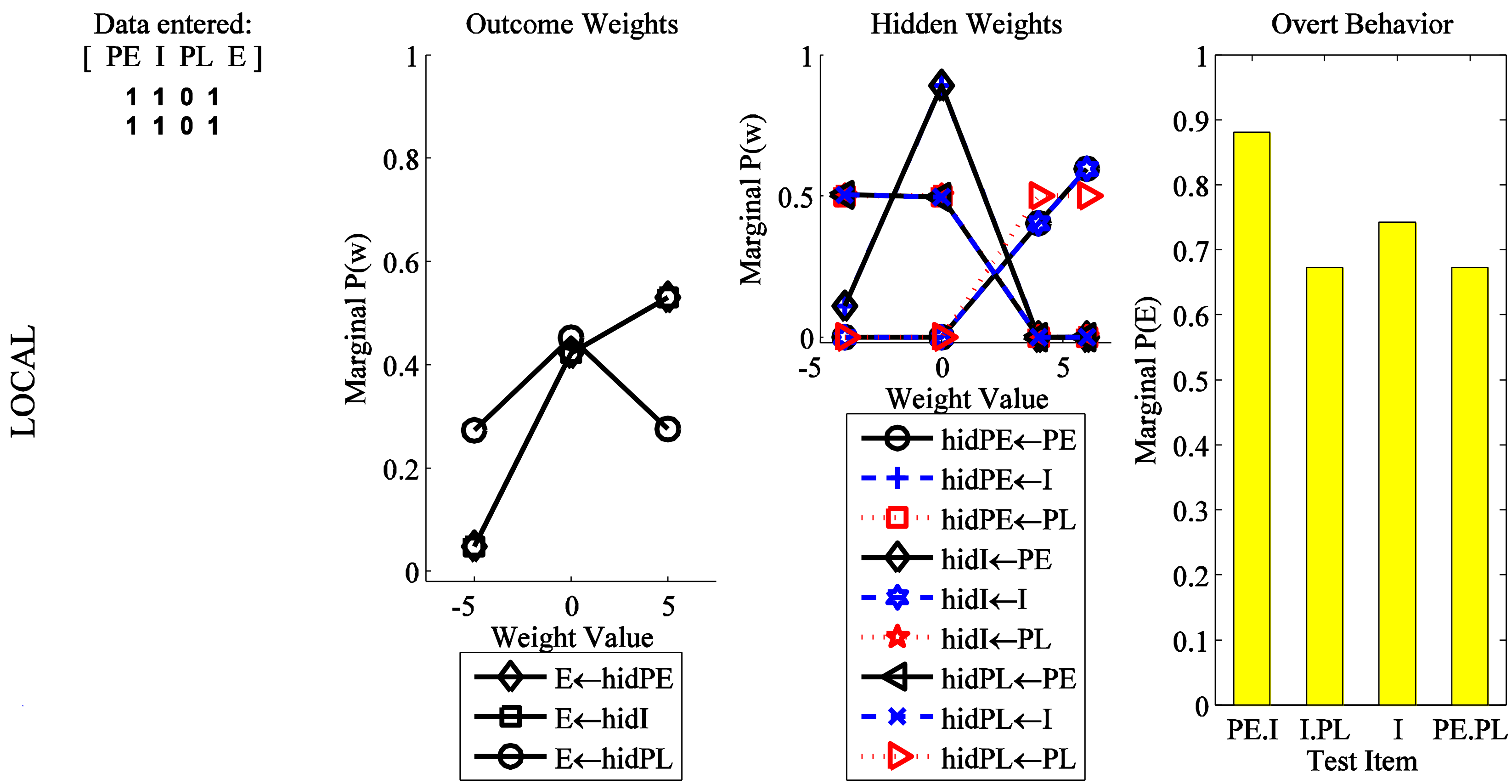
Highlighting: During training...

Data entered:
[PE I PL E]
1 1 0 1

LOCAL



Highlighting: During training...

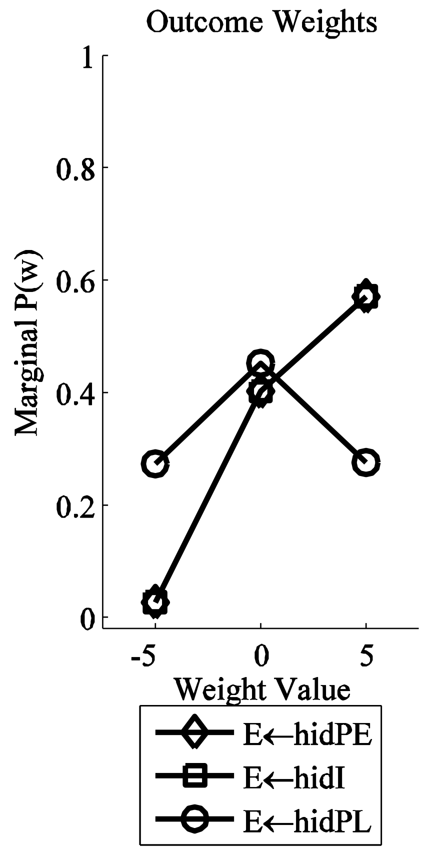


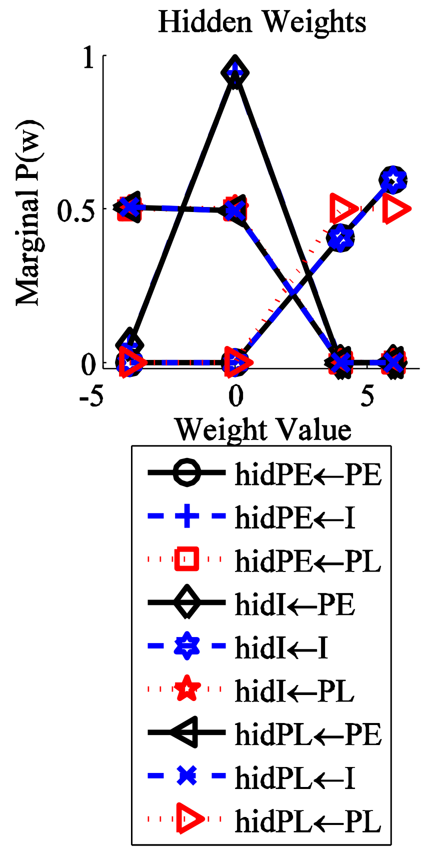
Highlighting: During training...

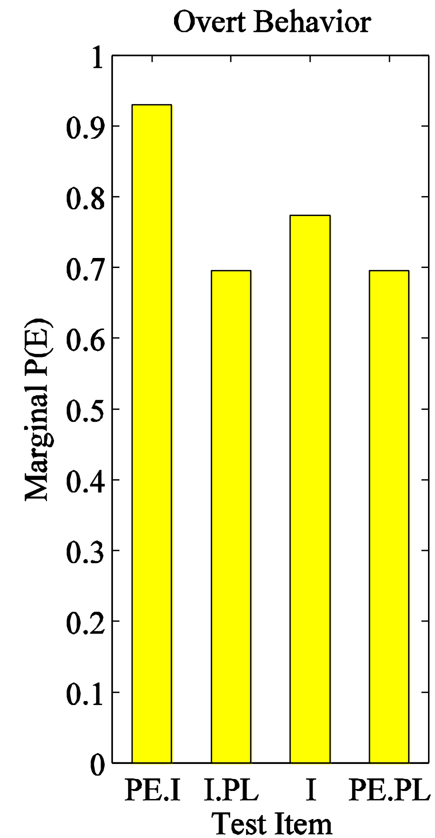
Data entered:
[PE I PL E]

1 1 0 1
1 1 0 1
1 1 0 1

LOCAL







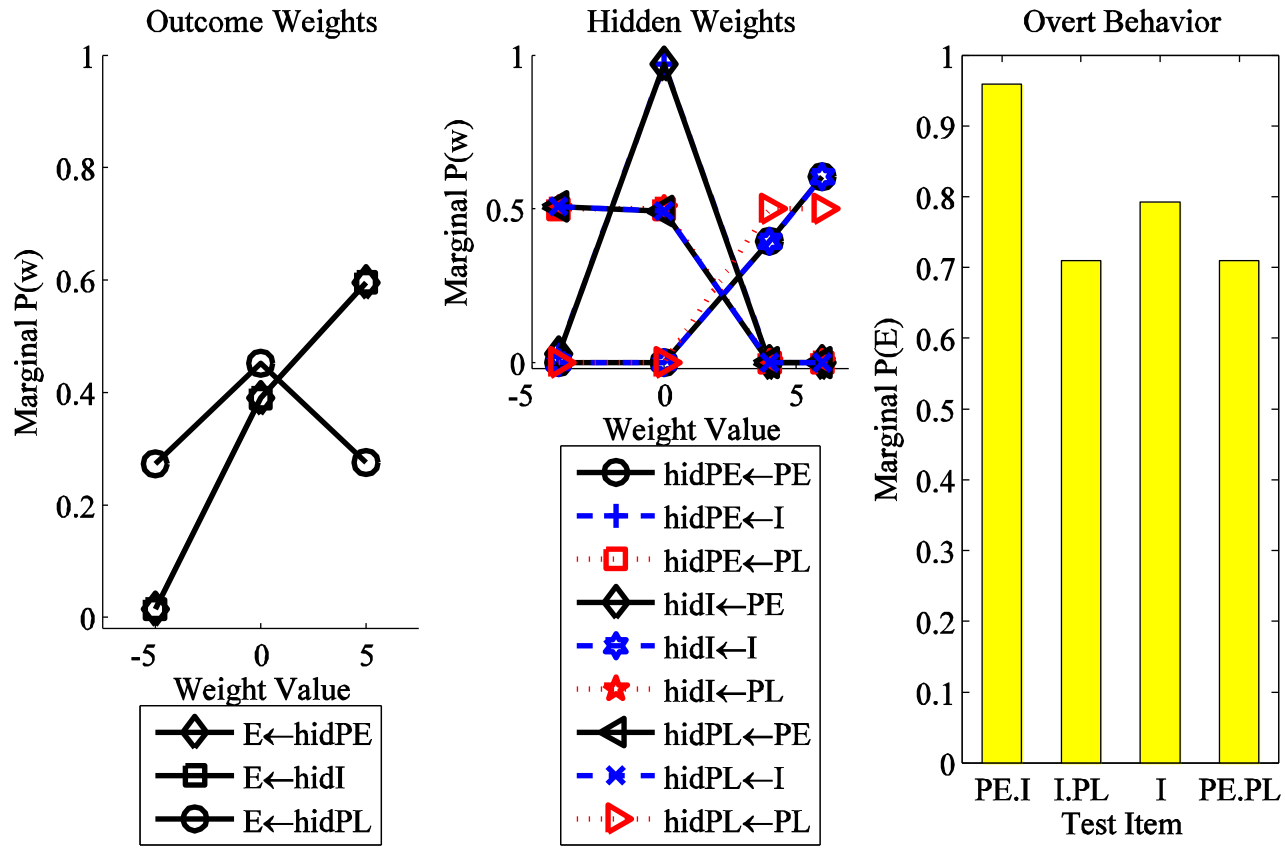
Highlighting: During training...

Data entered:

[PE I PL E]

1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1

LOCAL

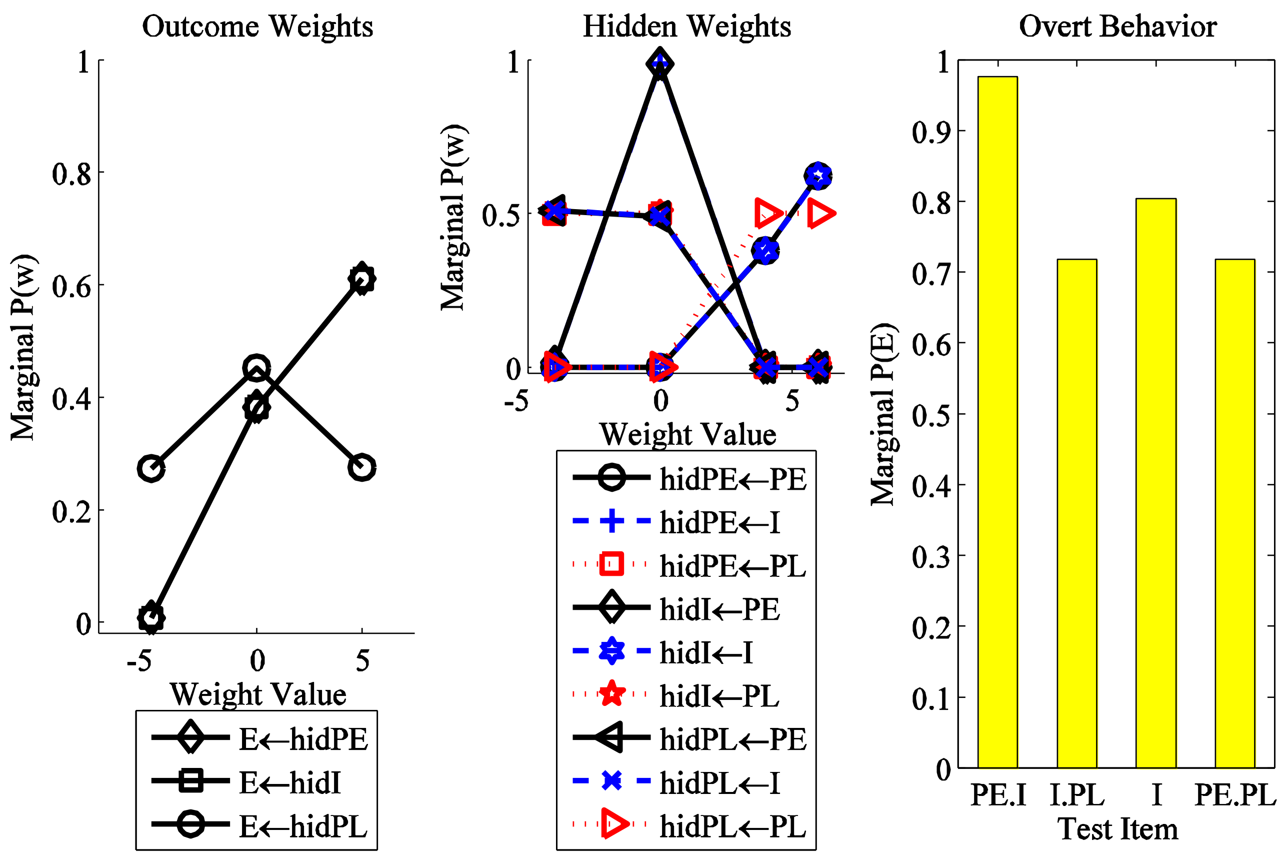


Highlighting: During training...

LOCAL

Data entered:

PE	I	PL	E
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1

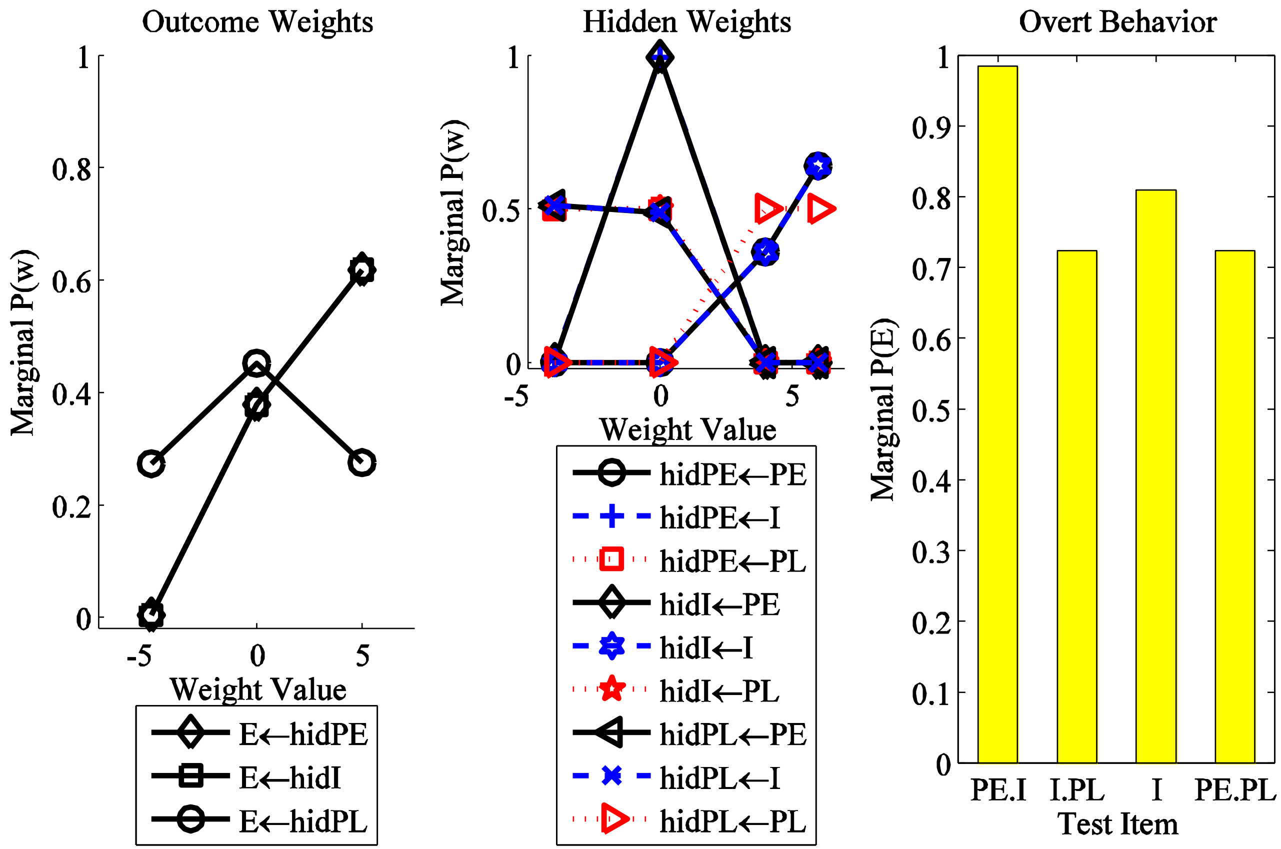


Highlighting: During training...

LOCAL

Data entered:
[PE I PL E]

1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1

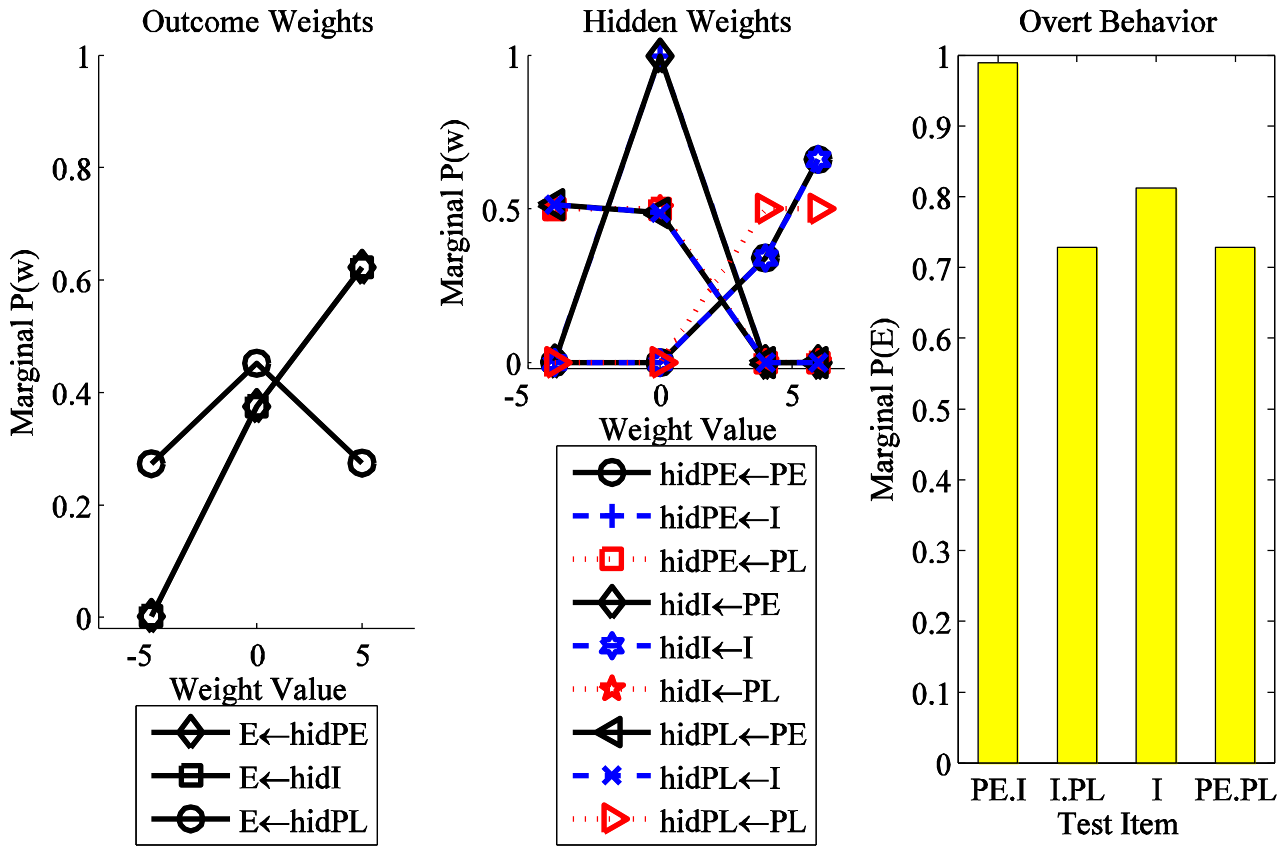


Highlighting: During training…

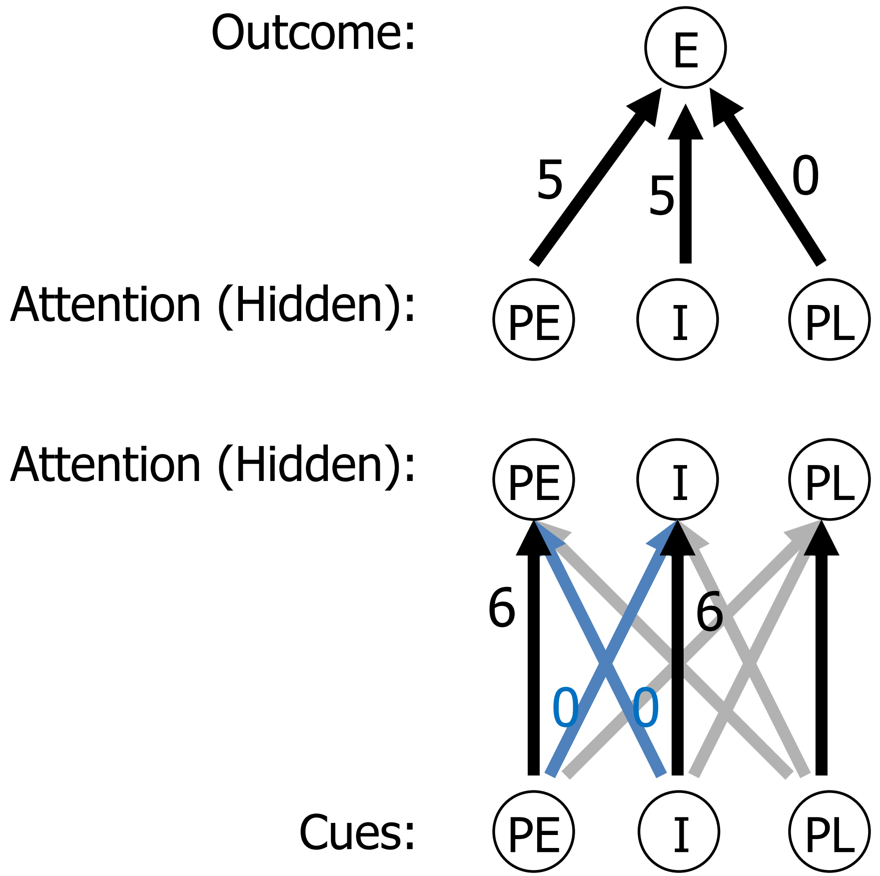
Data entered:
[PE I PL E]

1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1

LOCAL



Hypotheses After Initial Learning of PE.I → E

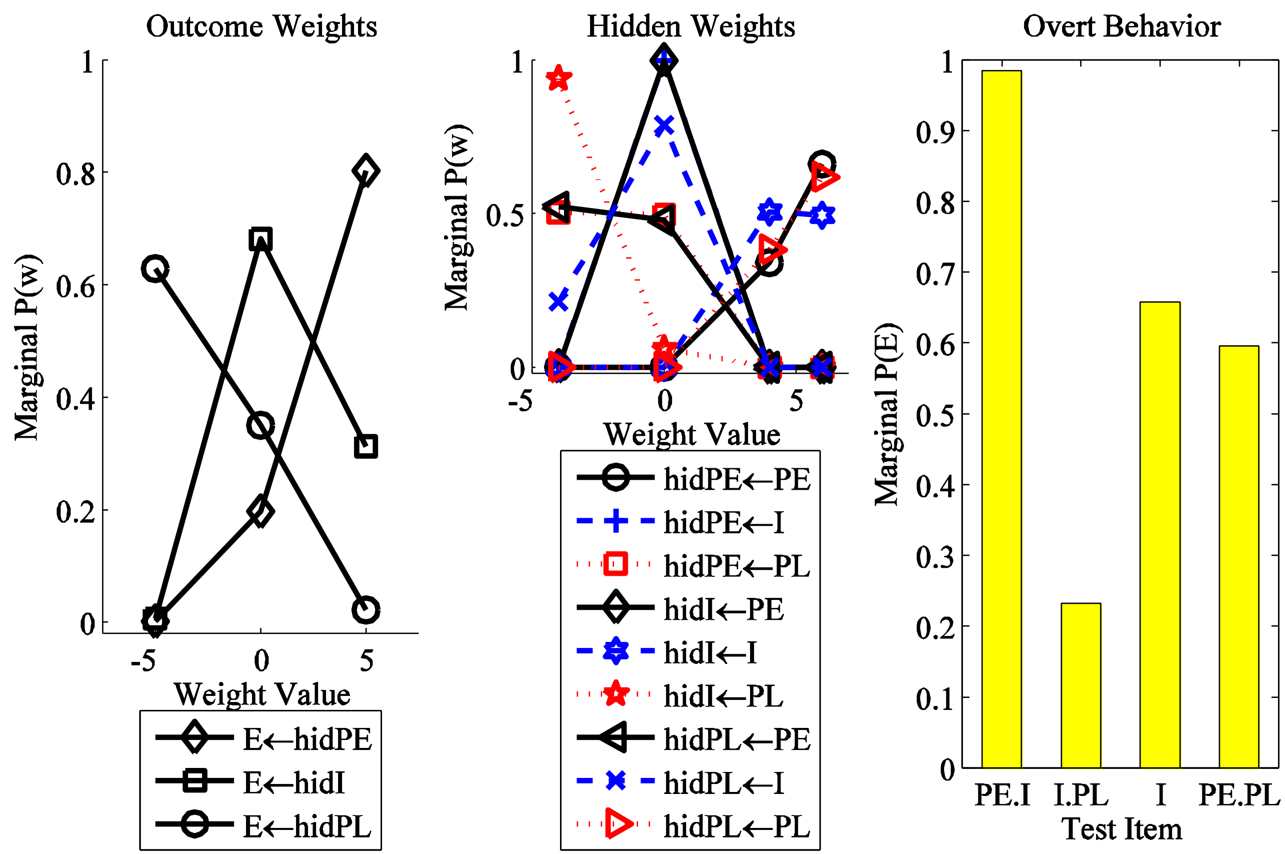


Highlighting: During training...

LOCAL

Data entered:

PE	I	PL	E
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
0	1	1	0

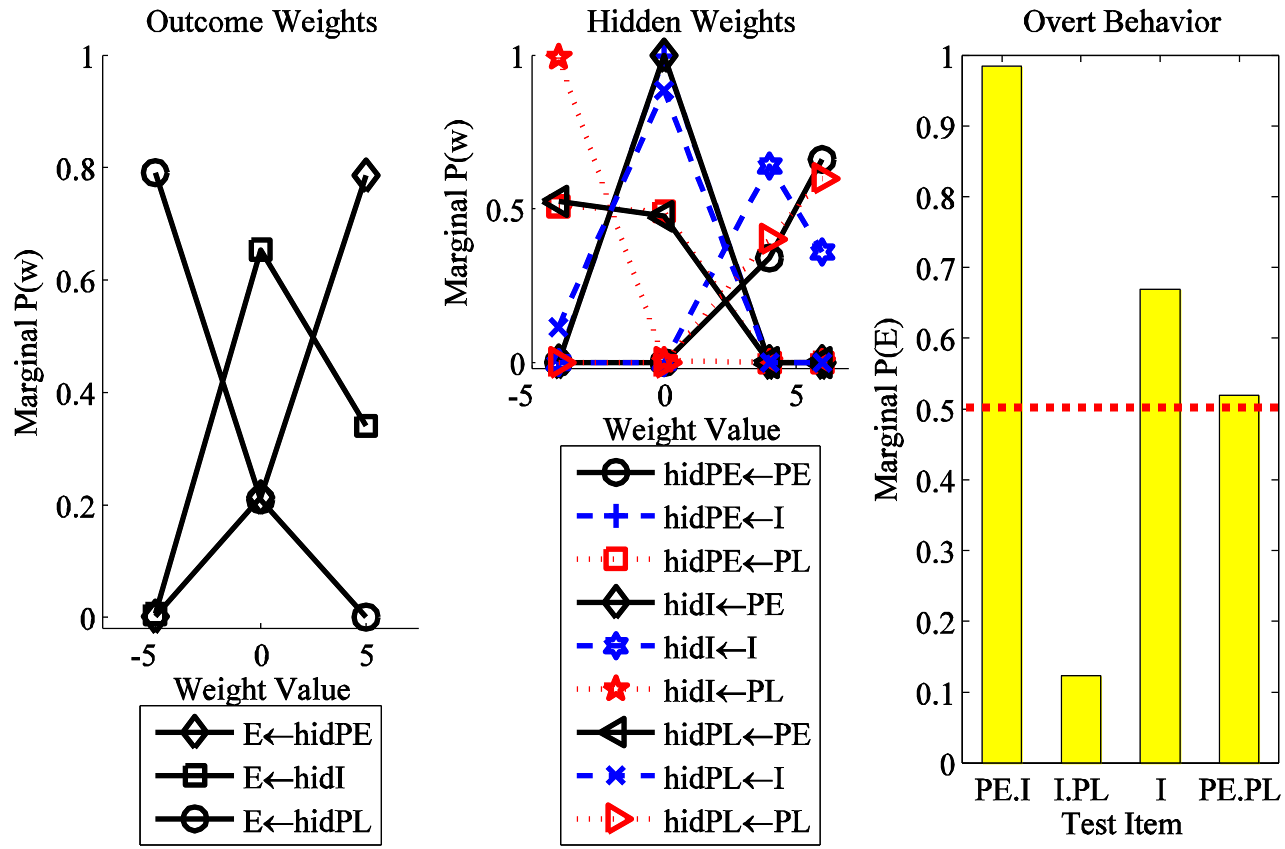


Highlighting: During training...

LOCAL

Data entered:

PE	I	PL	E
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
0	1	1	0
0	1	1	0

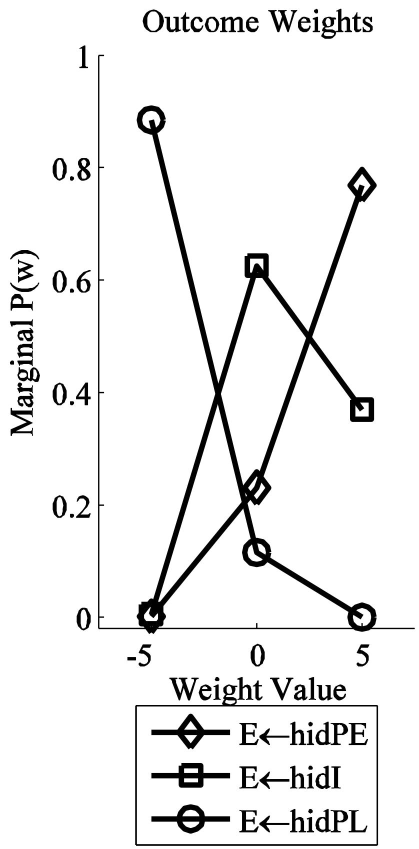


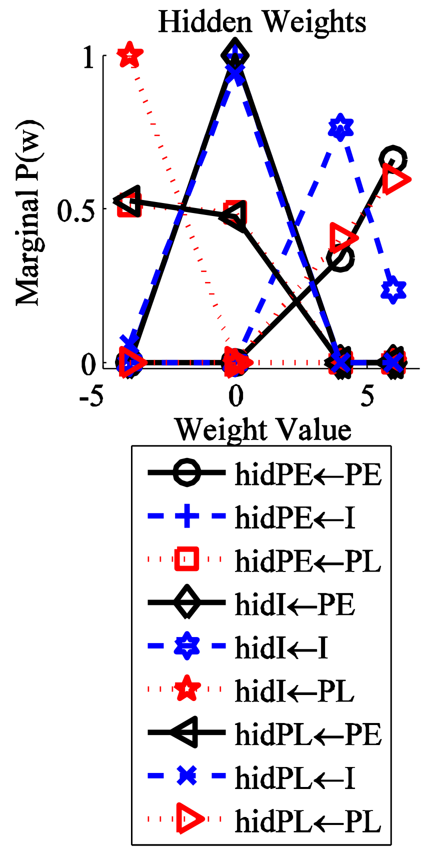
Highlighting: During training...

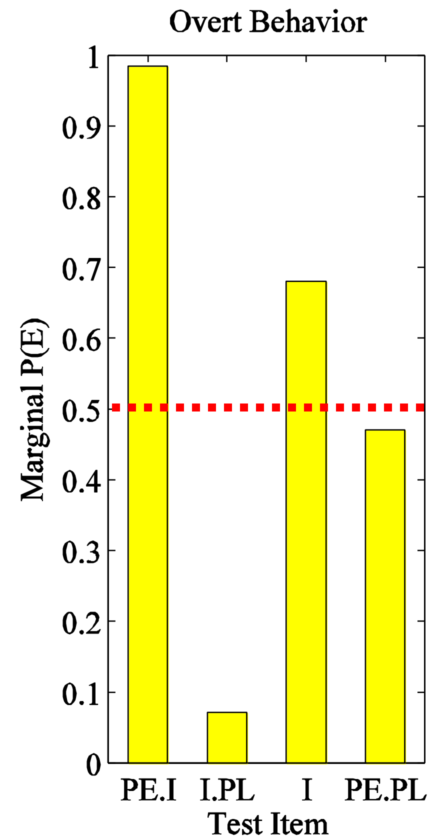
LOCAL

Data entered:

[PE	I	PL	E]
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
0	1	1	0
0	1	1	0
0	1	1	0





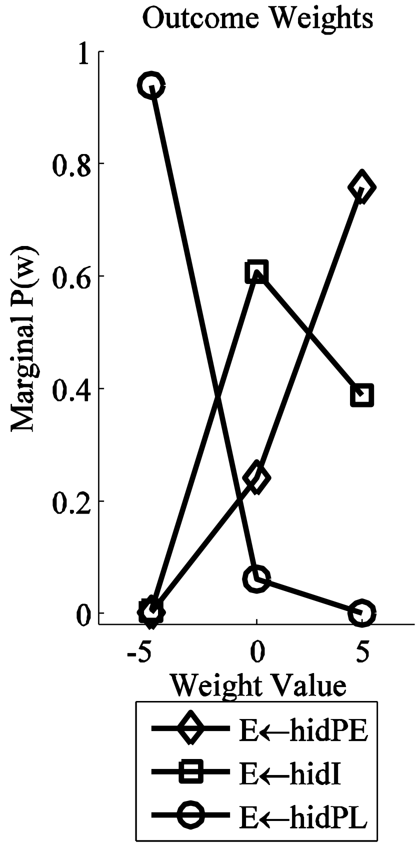


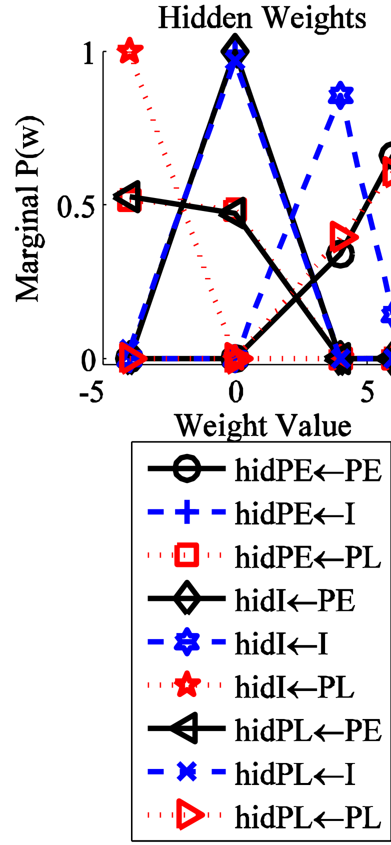
Highlighting: During training...

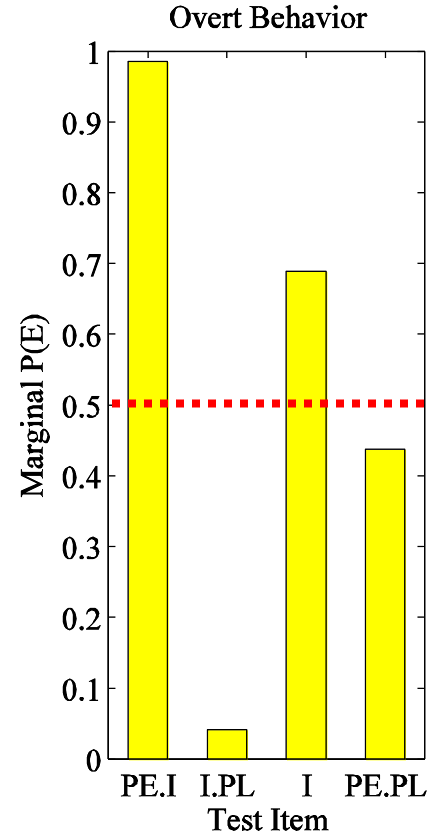
LOCAL

Data entered:

PE	I	PL	E
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0





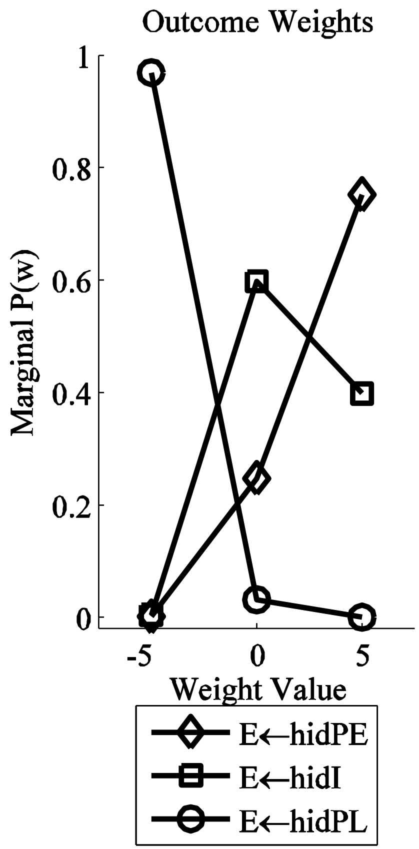


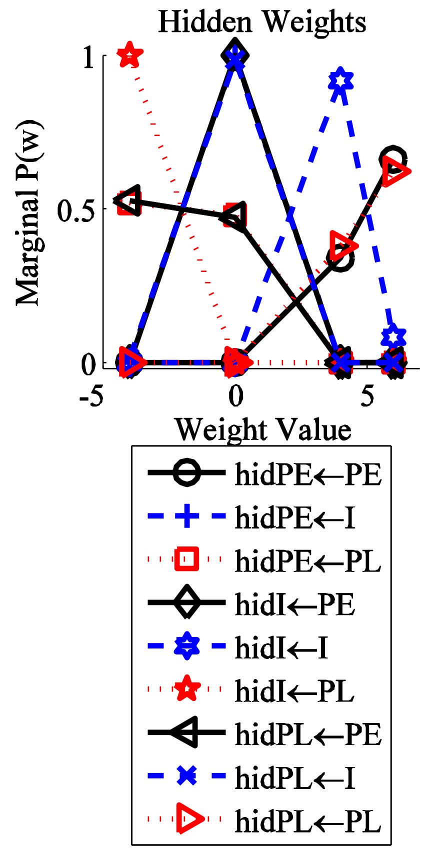
Highlighting: During training...

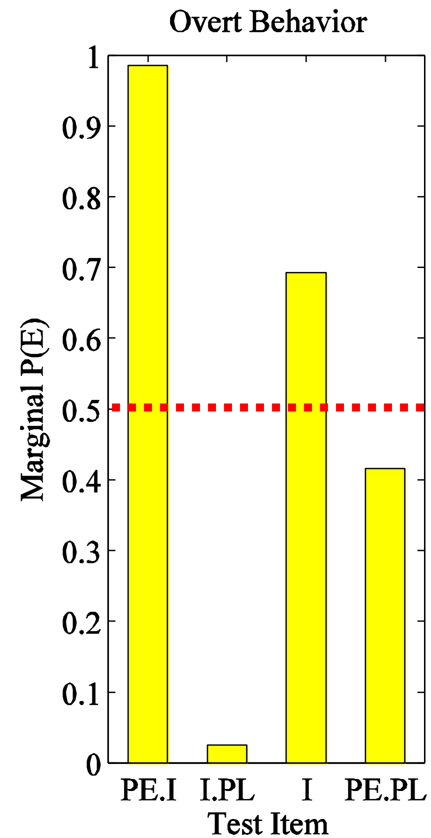
LOCAL

Data entered:

PE	I	PL	E
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0





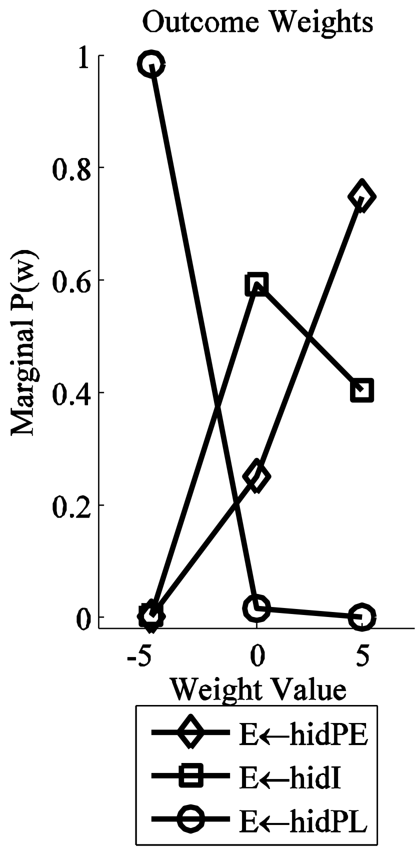


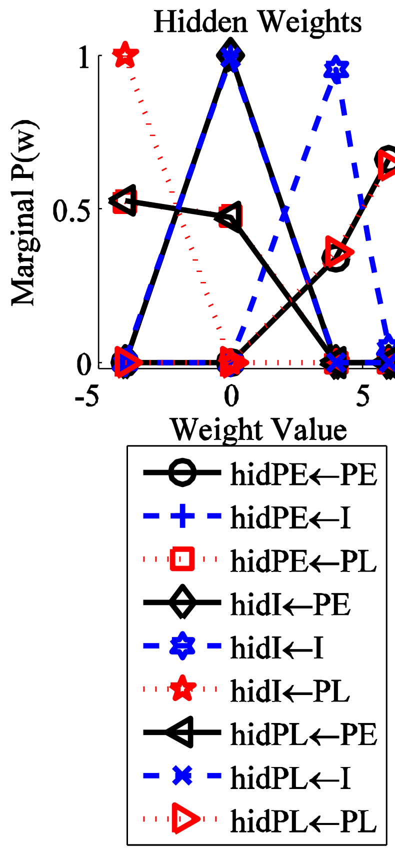
Highlighting: During training…

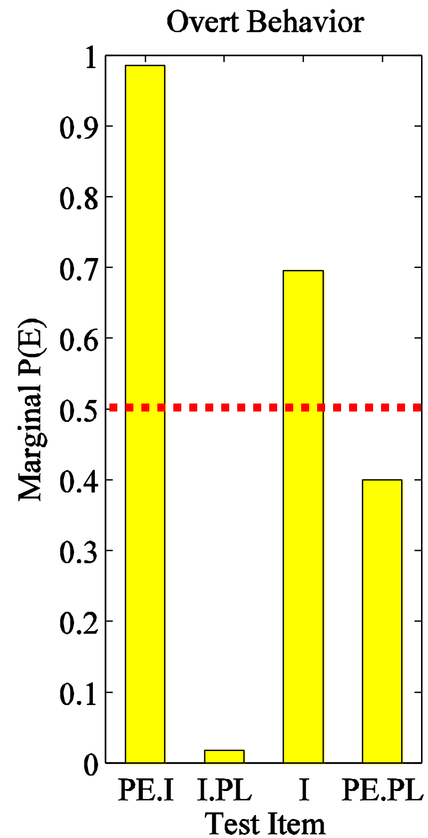
LOCAL

Data entered:

PE	I	PL	E
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0





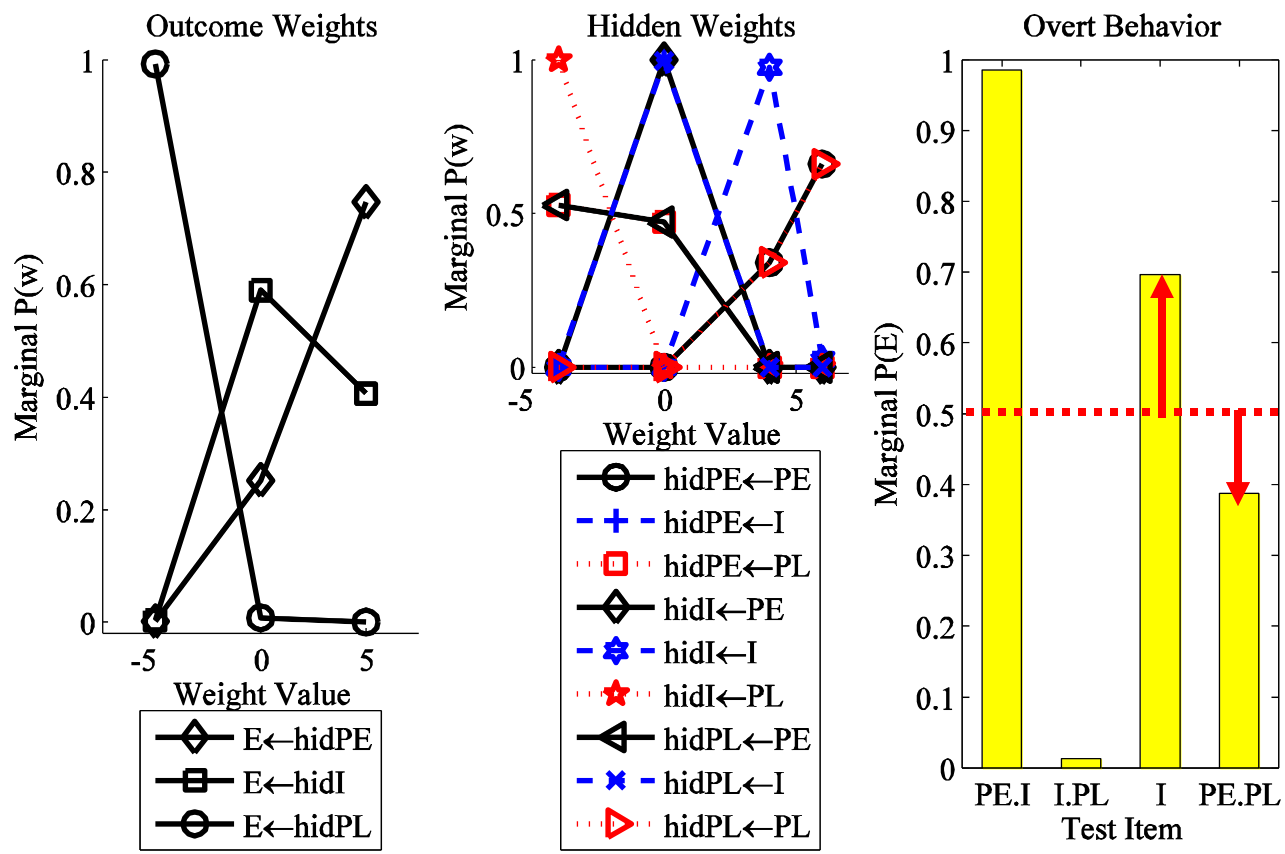


Highlighting: End of training

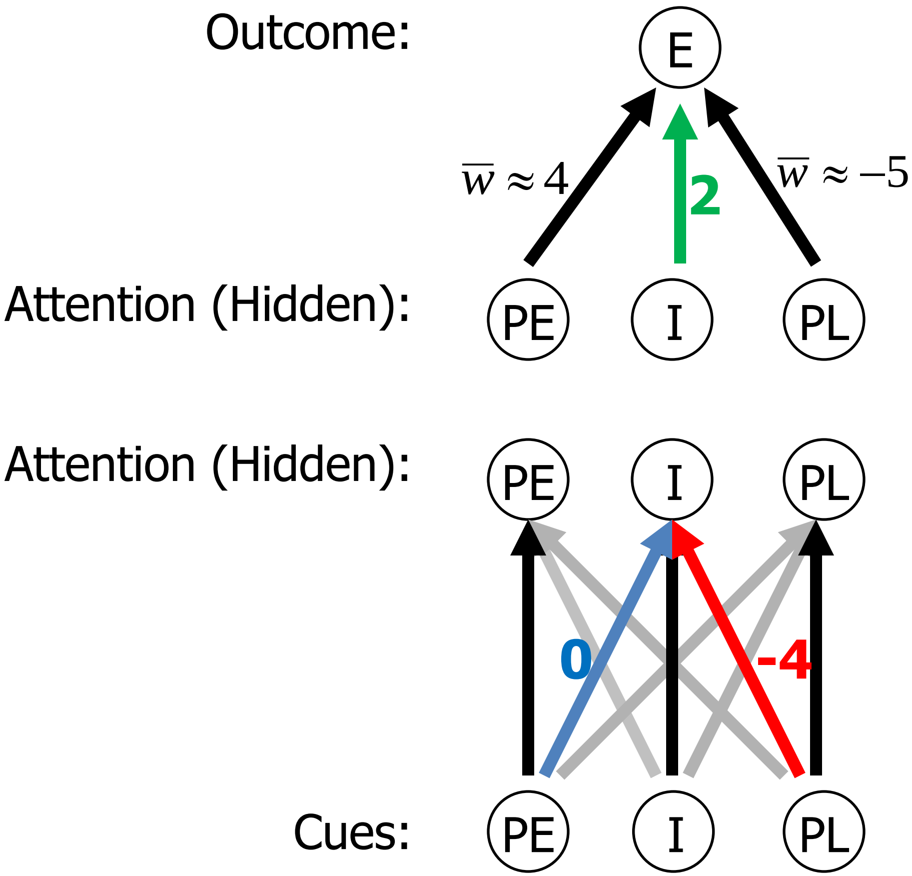
LOCAL

Data entered:

PE	I	PL	E
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0



Hypotheses After All Learning, PE.I → E and I.PL → L



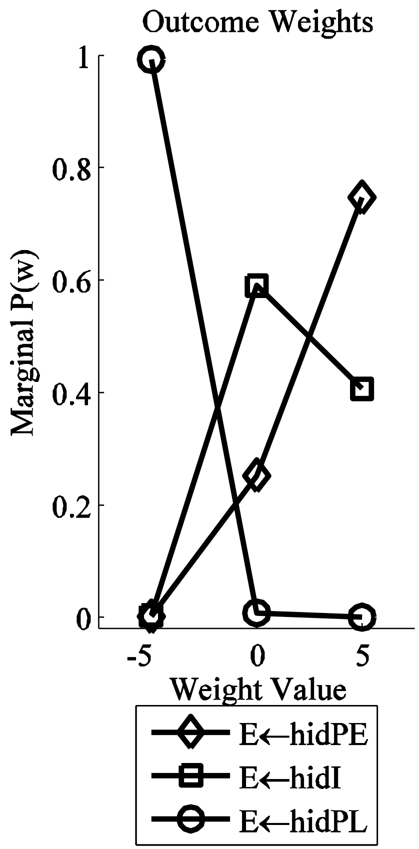
Inhibition of I by PL **prevents disconfirmation** of previous learning that I→E.

Highlighting: End of training

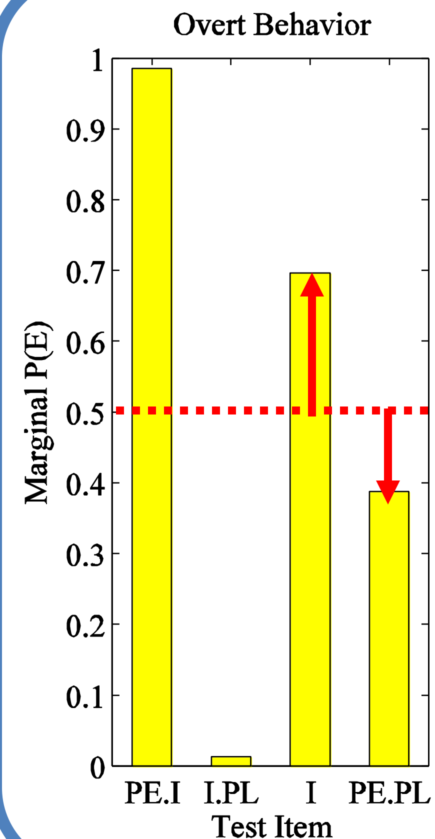
LOCAL

Data entered:

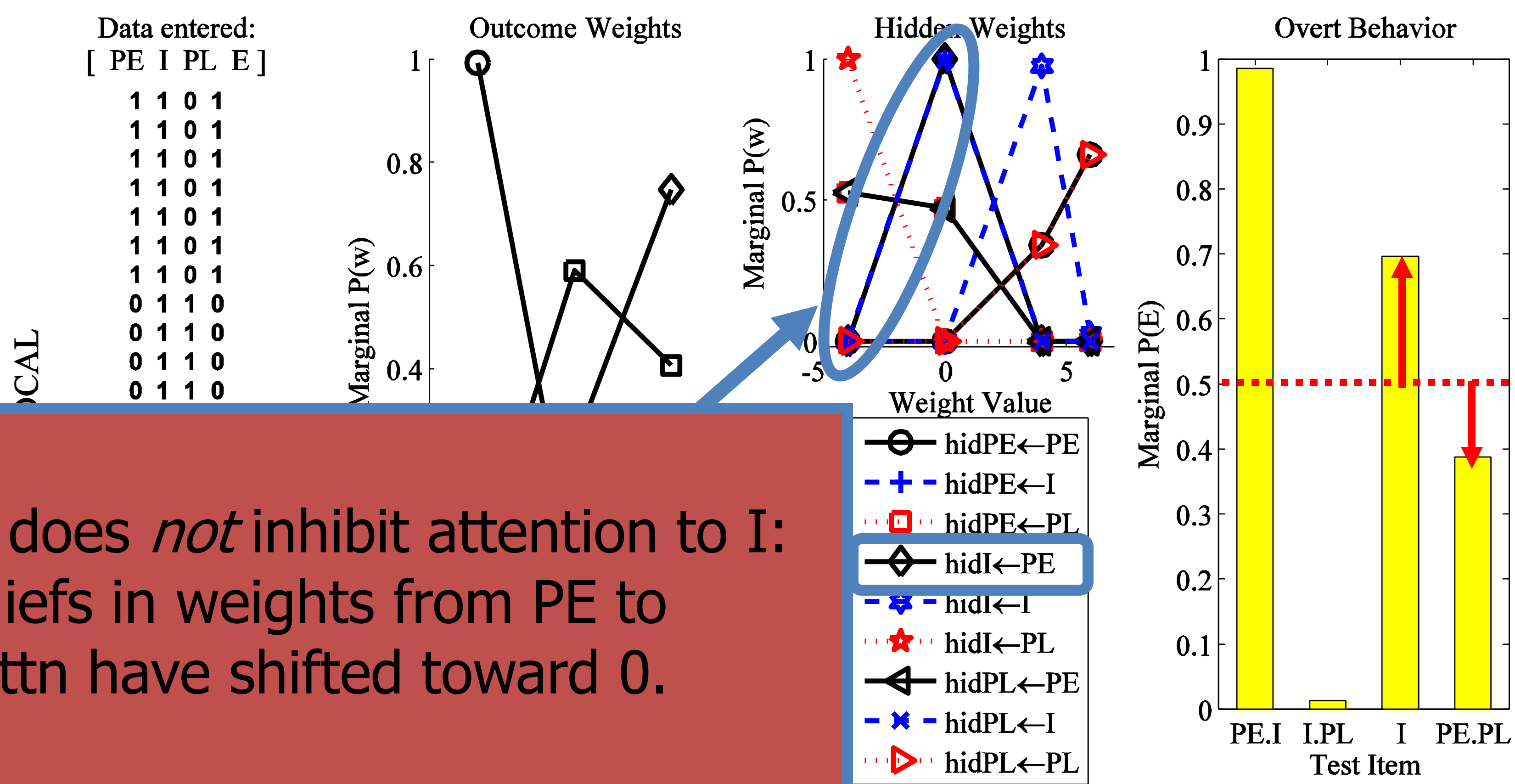
PE	I	PL	E
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
1	1	0	1
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0



Model mimics human preferences

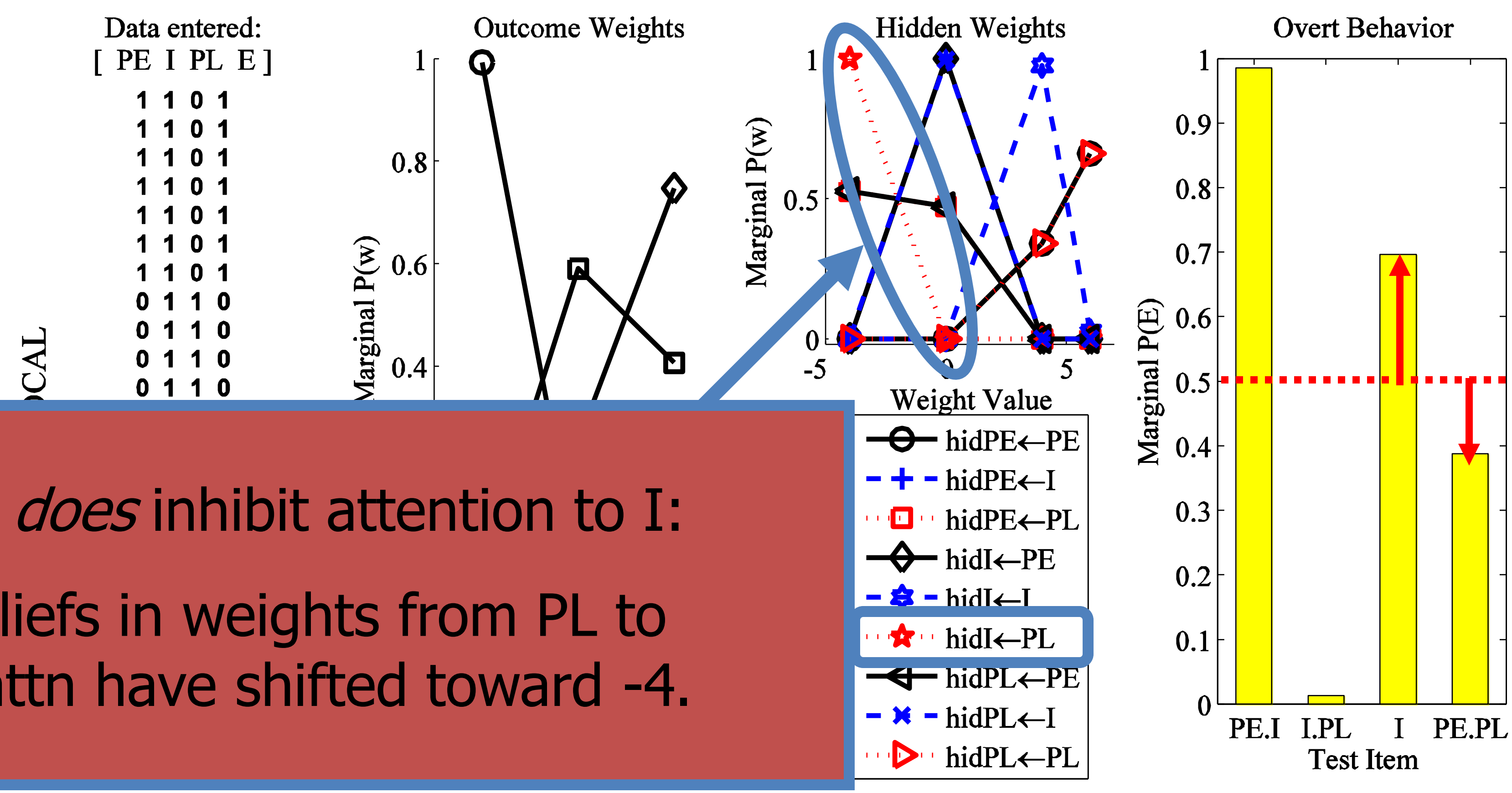


Highlighting: End of training



PE does *not* inhibit attention to I:
Beliefs in weights from PE to I-attn have shifted toward 0.

Highlighting: End of training



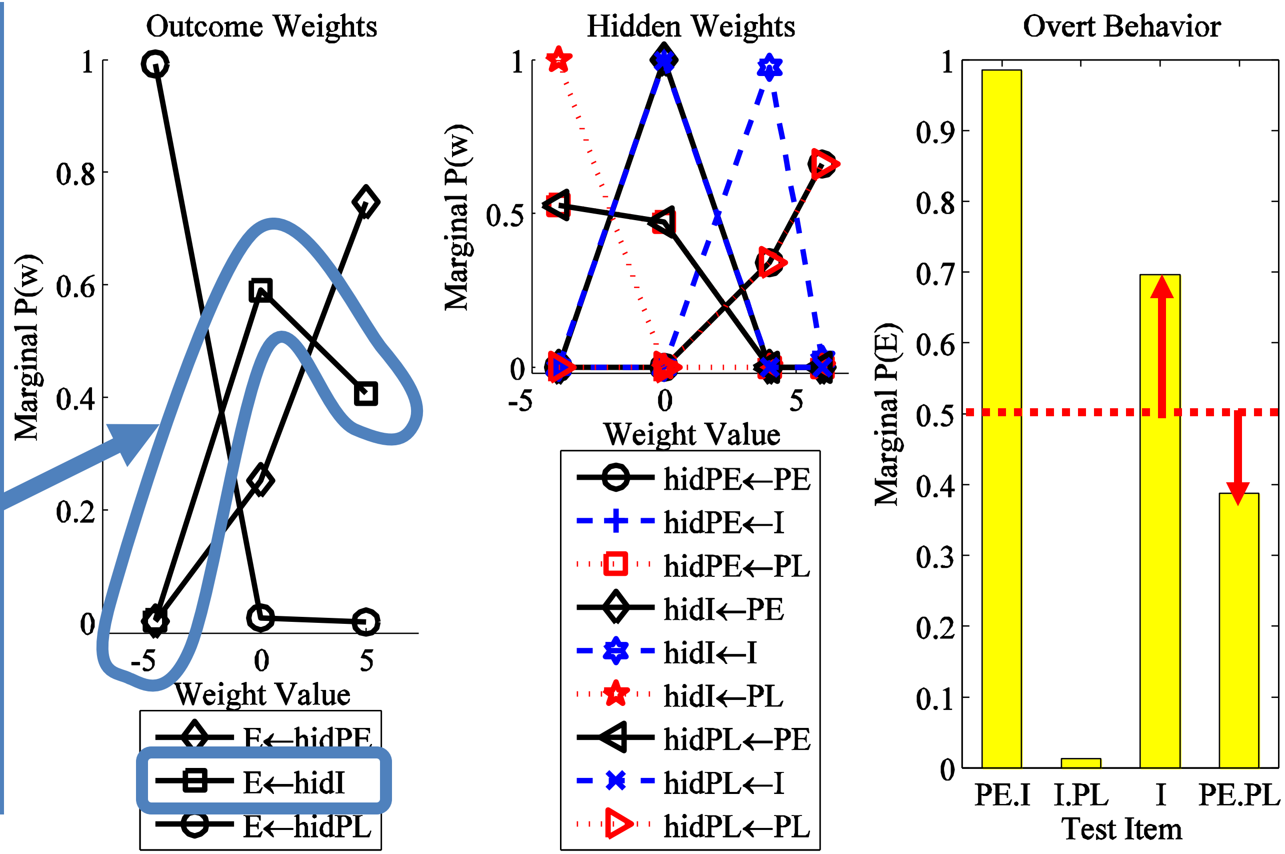
PL *does* inhibit attention to I:

Beliefs in weights from PL to I-attn have shifted toward -4.

Highlighting: End of training

Beliefs about I are asymmetric:

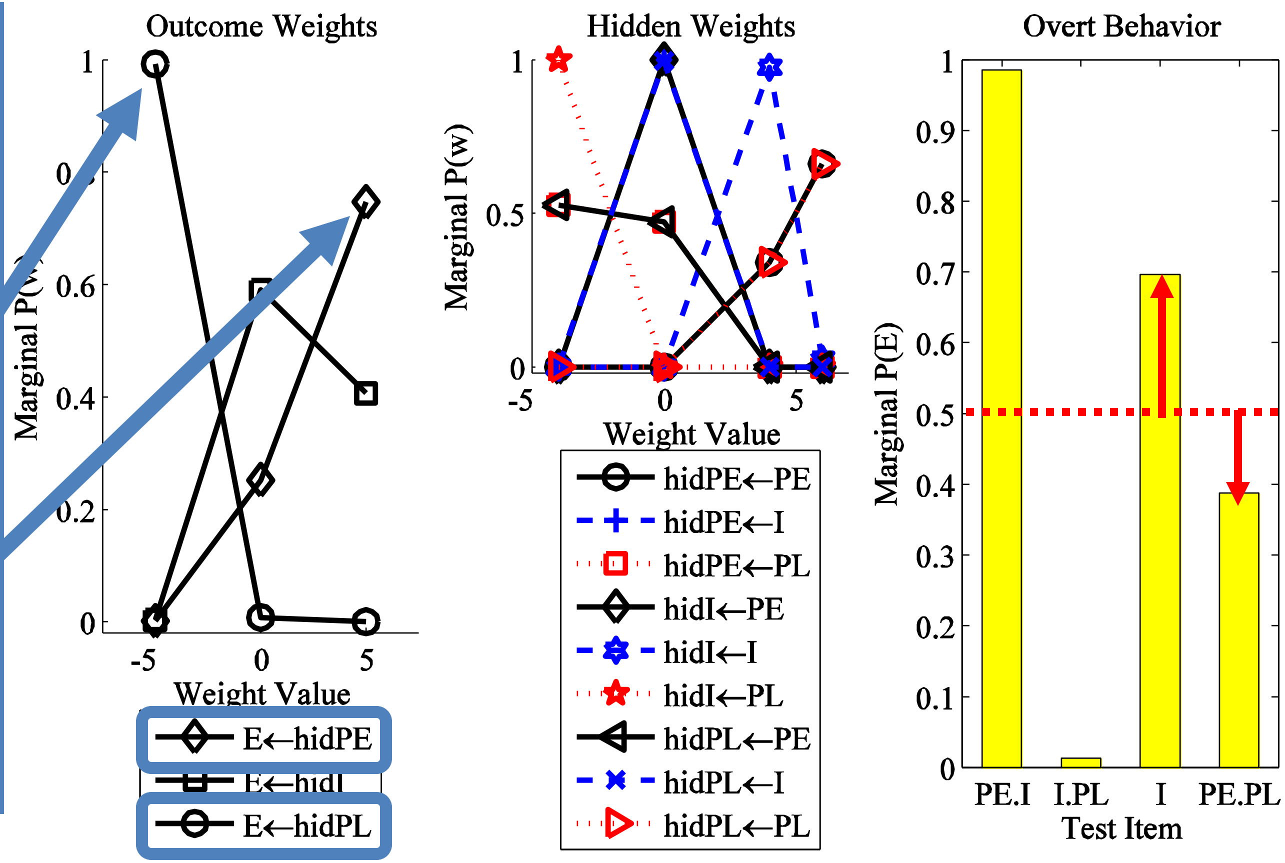
Stronger beliefs in +5 weights than -5 weights.



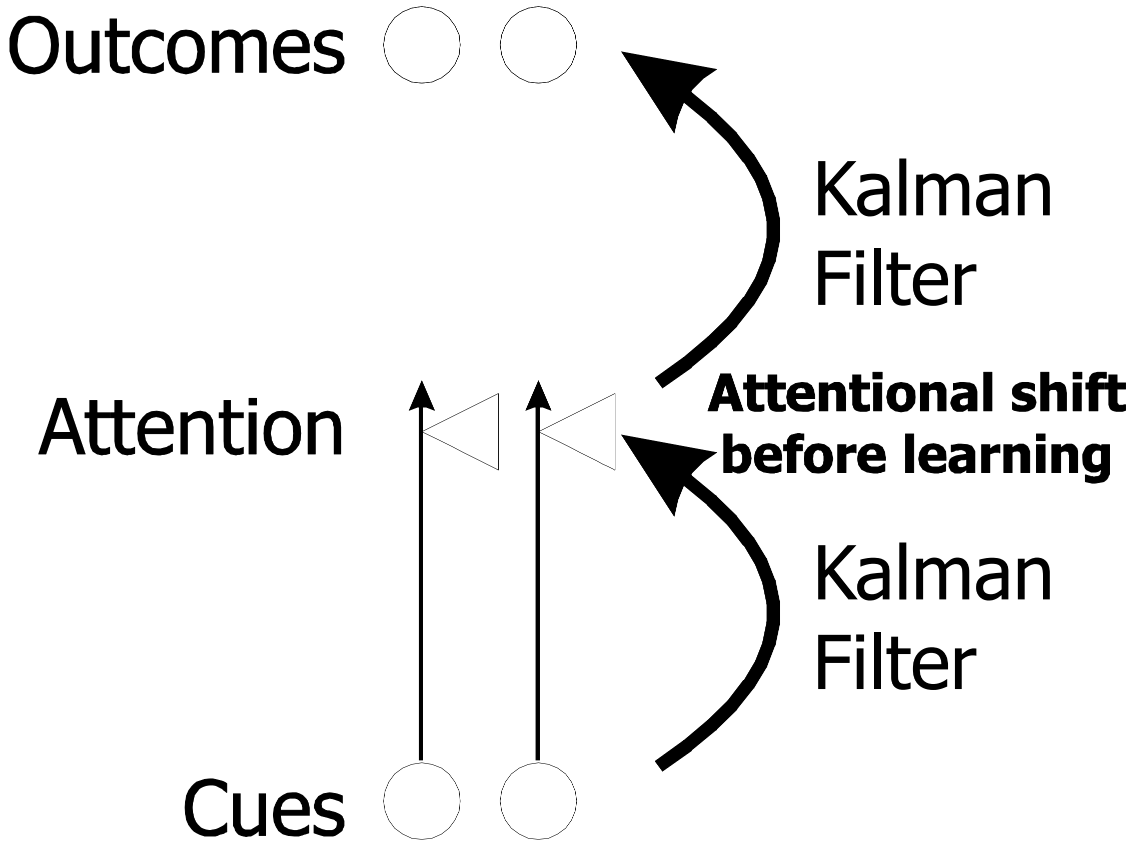
Highlighting: End of training

Beliefs about PE and PL are asymmetric:

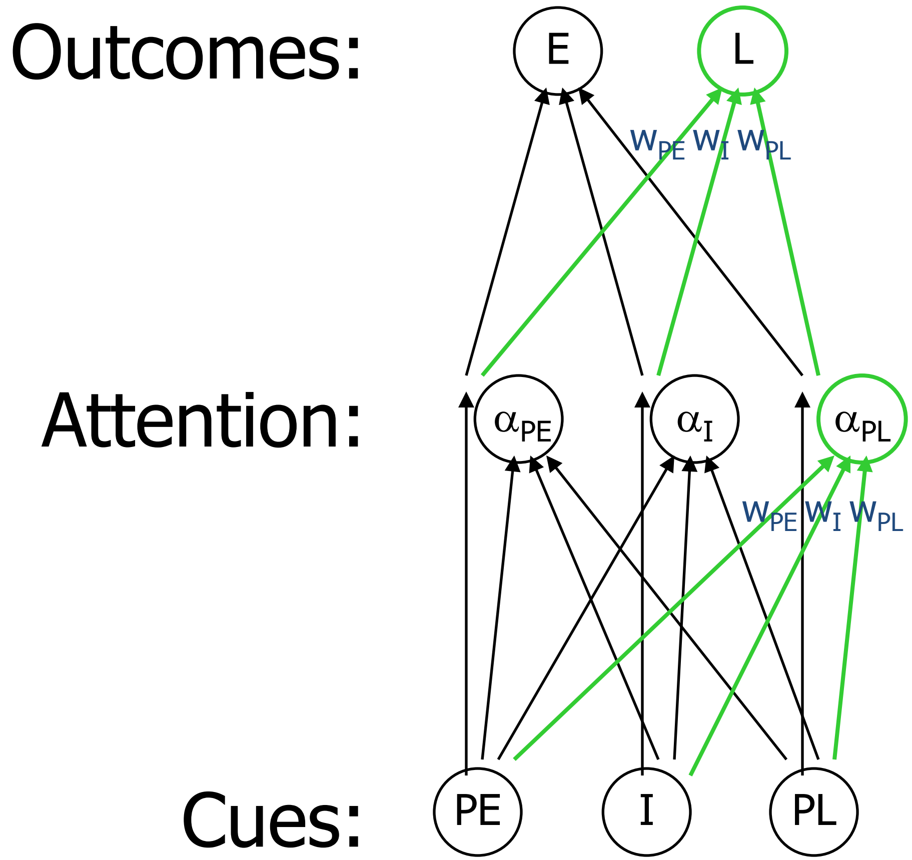
PL beliefs are more extreme than PE beliefs.



Models of Attention Shifting: Locally Bayesian



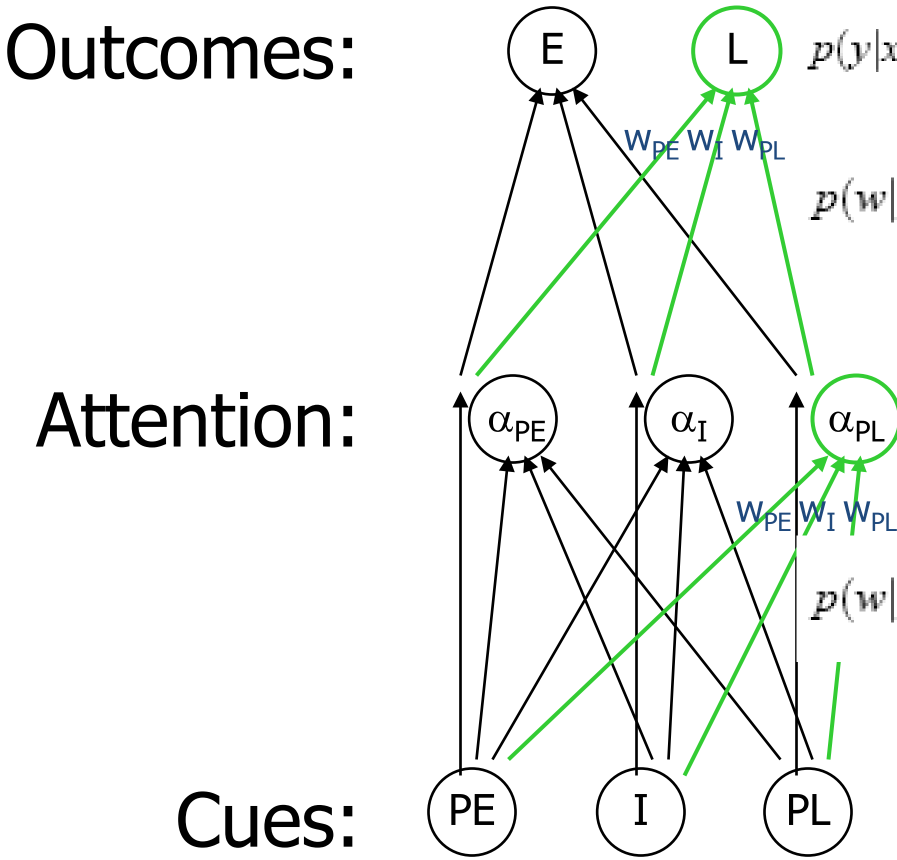
Layers of Kalman Filters
Applied to Highlighting



Kalman Filters

Kalman Filters

Layers of Kalman Filters:
Likelihood and Prior Distributions



$$p(y|x,w,v) = \frac{1}{\sqrt{v}(2\pi)^{1/2}} \exp\left(-.5\frac{(y-w^T x)^2}{v}\right)$$

$$p(w|\mu,C) = \frac{1}{\sqrt{|C|}(2\pi)^{d/2}} \exp\left(-.5(w-\mu)^T C^{-1}(w-\mu)\right)$$

$$p(y|x,w,v) = \frac{1}{\sqrt{v}(2\pi)^{1/2}} \exp\left(-.5\frac{(y-w^T x)^2}{v}\right)$$

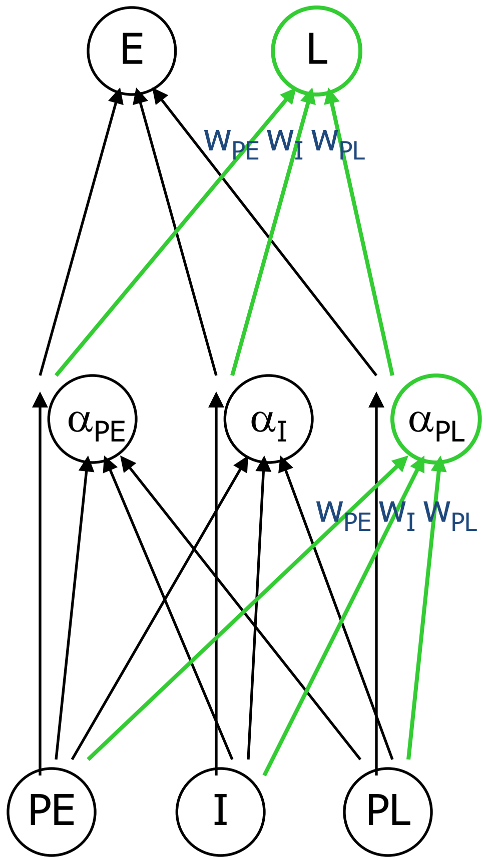
$$p(w|\mu,C) = \frac{1}{\sqrt{|C|}(2\pi)^{d/2}} \exp\left(-.5(w-\mu)^T C^{-1}(w-\mu)\right)$$

Layers of Kalman Filters:
Outcome generation

Outcomes:

Attention:

Cues:



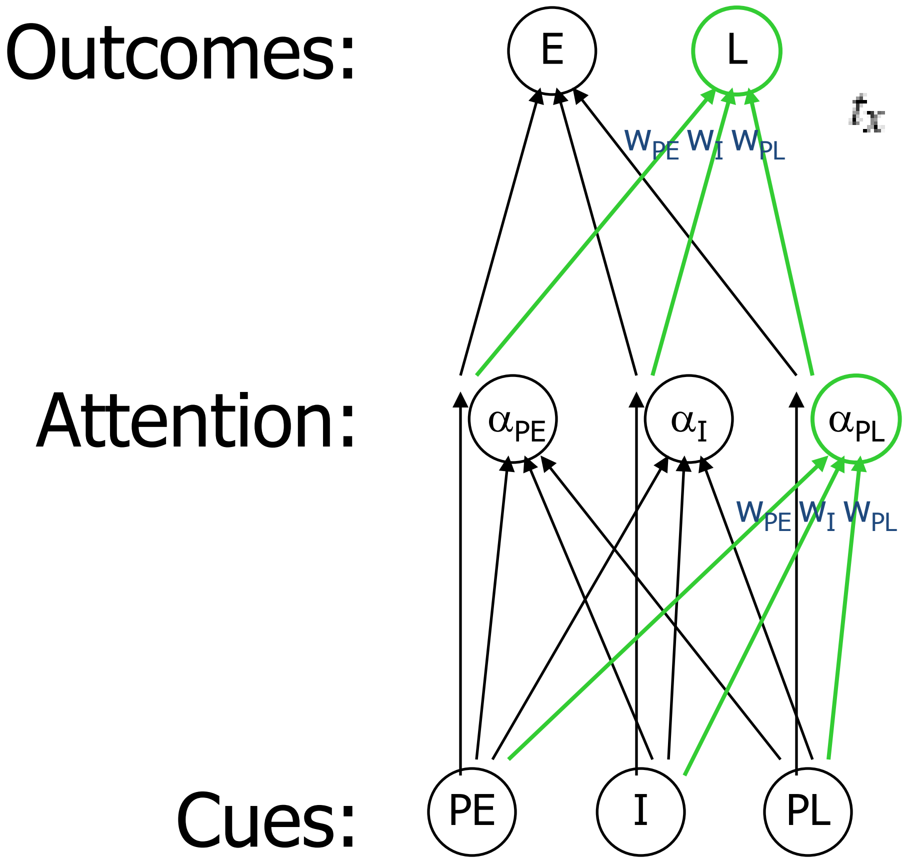
$$\begin{aligned}\bar{y} &= \int dw\, p(w|\mu, C) \int dy\, y\, p(y|x, w, v) \\ &= \mu^T x\end{aligned}$$

$x = \text{input} \cdot \bar{y}$

$$\begin{aligned}\bar{y} &= \int dw\, p(w|\mu, C) \int dy\, y\, p(y|x, w, v) \\ &= \mu^T x\end{aligned}$$

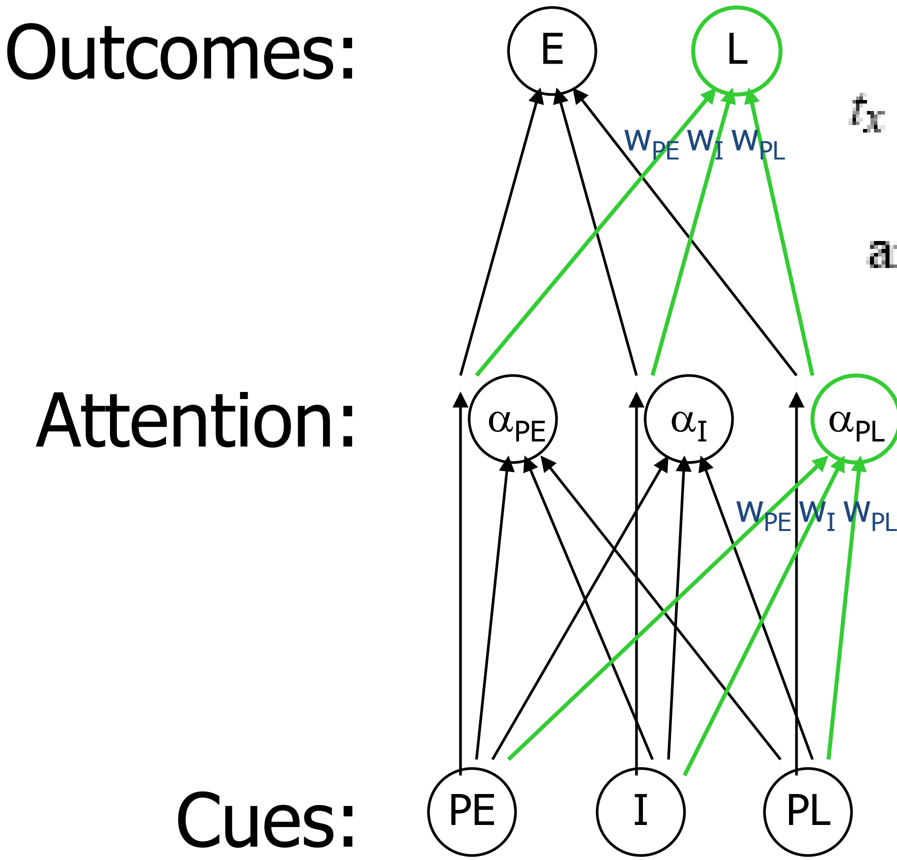
$x = \text{input activation vector}$

Layers of Kalman Filters:
Target for Attention



$$\begin{aligned} t_x &= \underset{x}{\operatorname{argmax}}\; p(t_y|x) \\ &= \underset{x}{\operatorname{argmax}} \int dw\; p(t_y|x,w,v)\; p(w|\mu,C) \end{aligned}$$

Layers of Kalman Filters: **Target for Attention**



$$t_x = \underset{x}{\operatorname{argmax}}\; p(t_y|x)$$

$$\underset{x}{\operatorname{argmax}} \frac{\exp\left(-.5\,(t-x^T\mu)\,[S+x^TCx]^{-1}\,(t-x^T\mu)\right)}{(2\pi)^{d/2}\sqrt{|S+x^TCx|}}$$

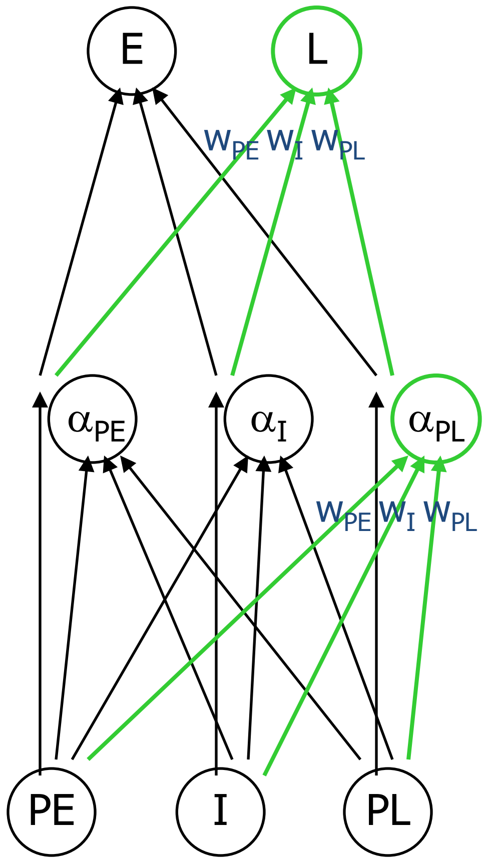
(To determine unique maximum, included tiny cost for unequal attention values, and tiny cost for non-zero attention on absent cue.)

Layers of Kalman Filters:
Dynamics and Bayesian Learning

Outcomes:

Attention:

Cues:



$$\begin{aligned} \mu^* &= D\mu \\ C^* &= DCD^T + U \end{aligned}$$

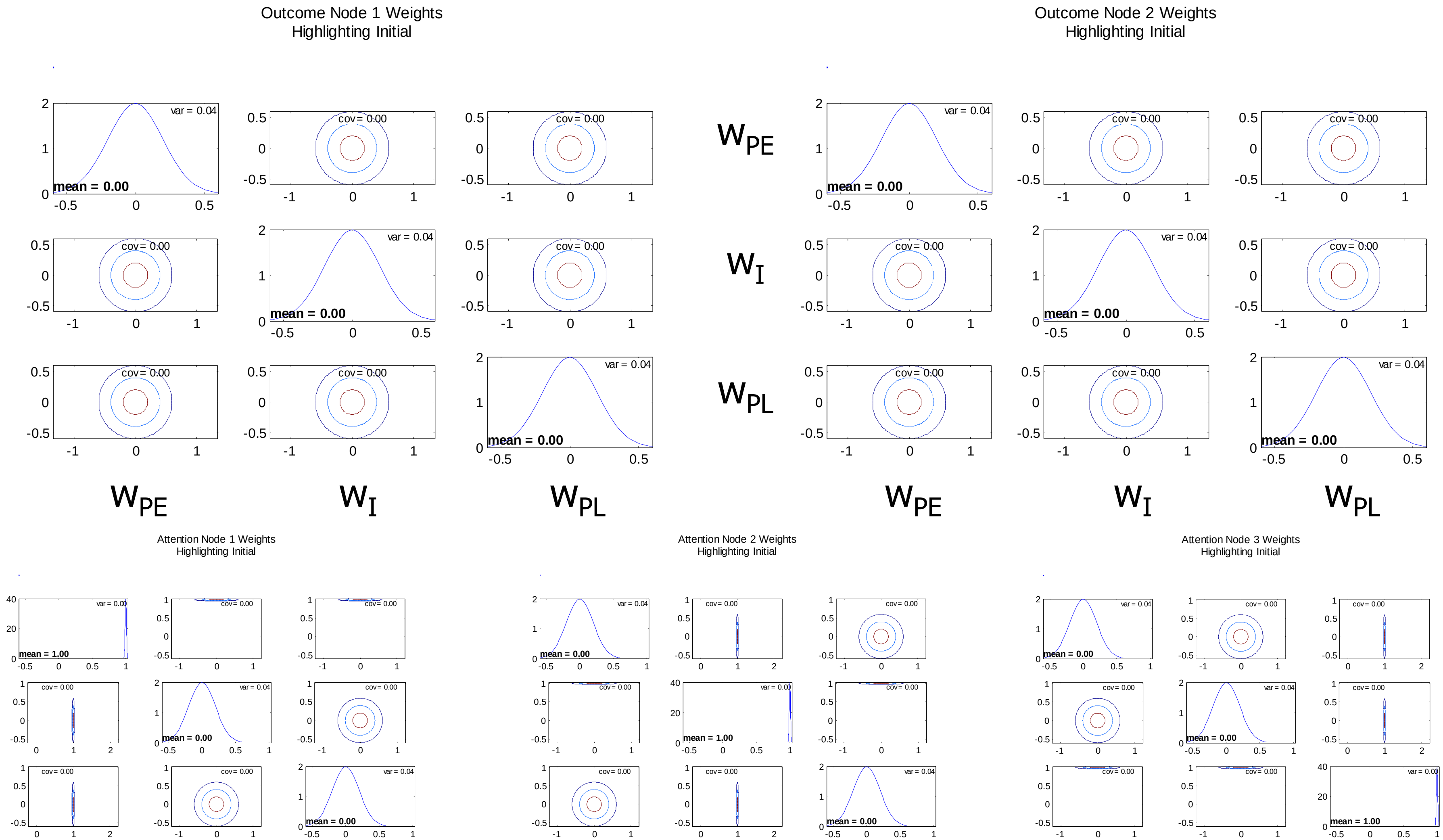
$$\begin{aligned} \mu' &= \mu^* + C^* x \left[v + x^T C^* x\right]^{-1} \left(t - x^T \mu^*\right) \\ C' &= C^* - C^* x \left[v + x^T C^* x\right]^{-1} x^T C^* \end{aligned}$$

$$\begin{aligned} \mu^* &= D\mu \\ C^* &= DCD^T + U \end{aligned}$$

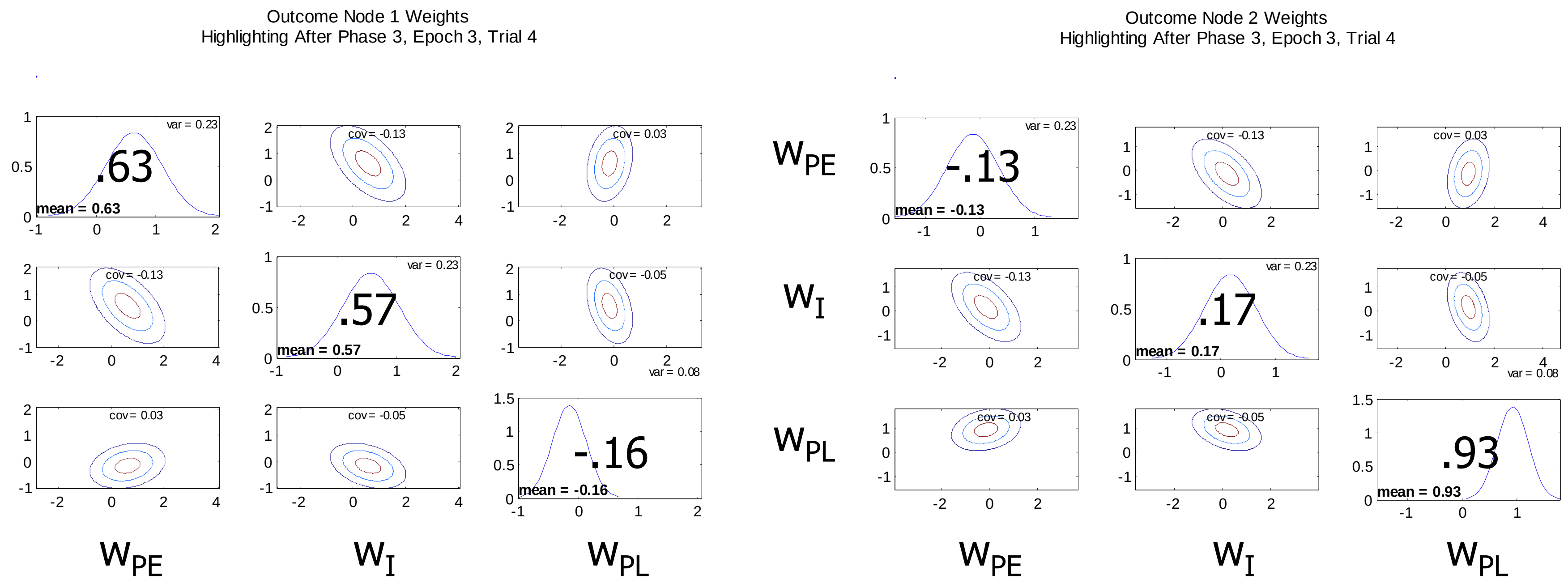
$$\begin{aligned} \mu' &= \mu^* + C^* x \left[v + x^T C^* x\right]^{-1} \left(t - x^T \mu^*\right) \\ C' &= C^* - C^* x \left[v + x^T C^* x\right]^{-1} x^T C^* \end{aligned}$$

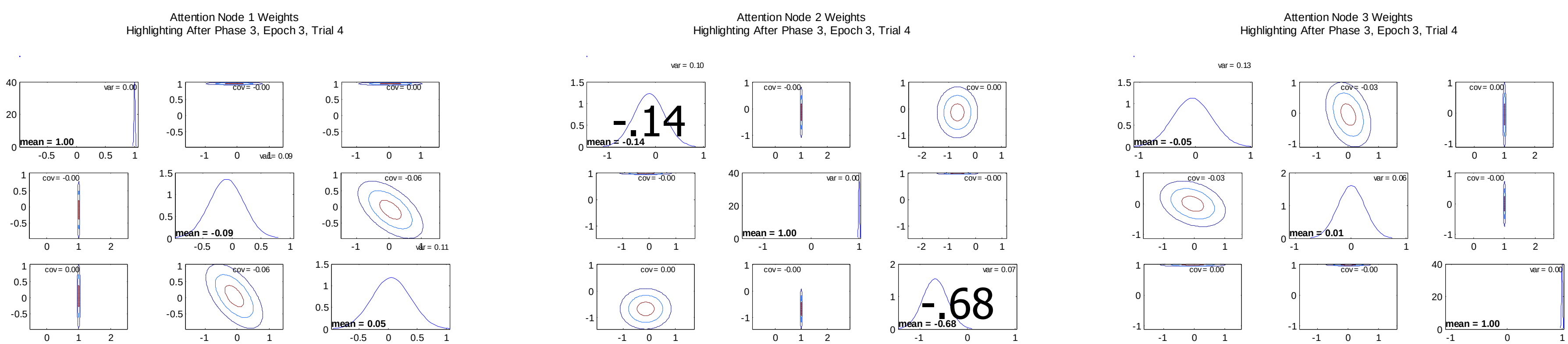
Layers of Kalman Filters Applied to Highlighting:

Initial p(w)



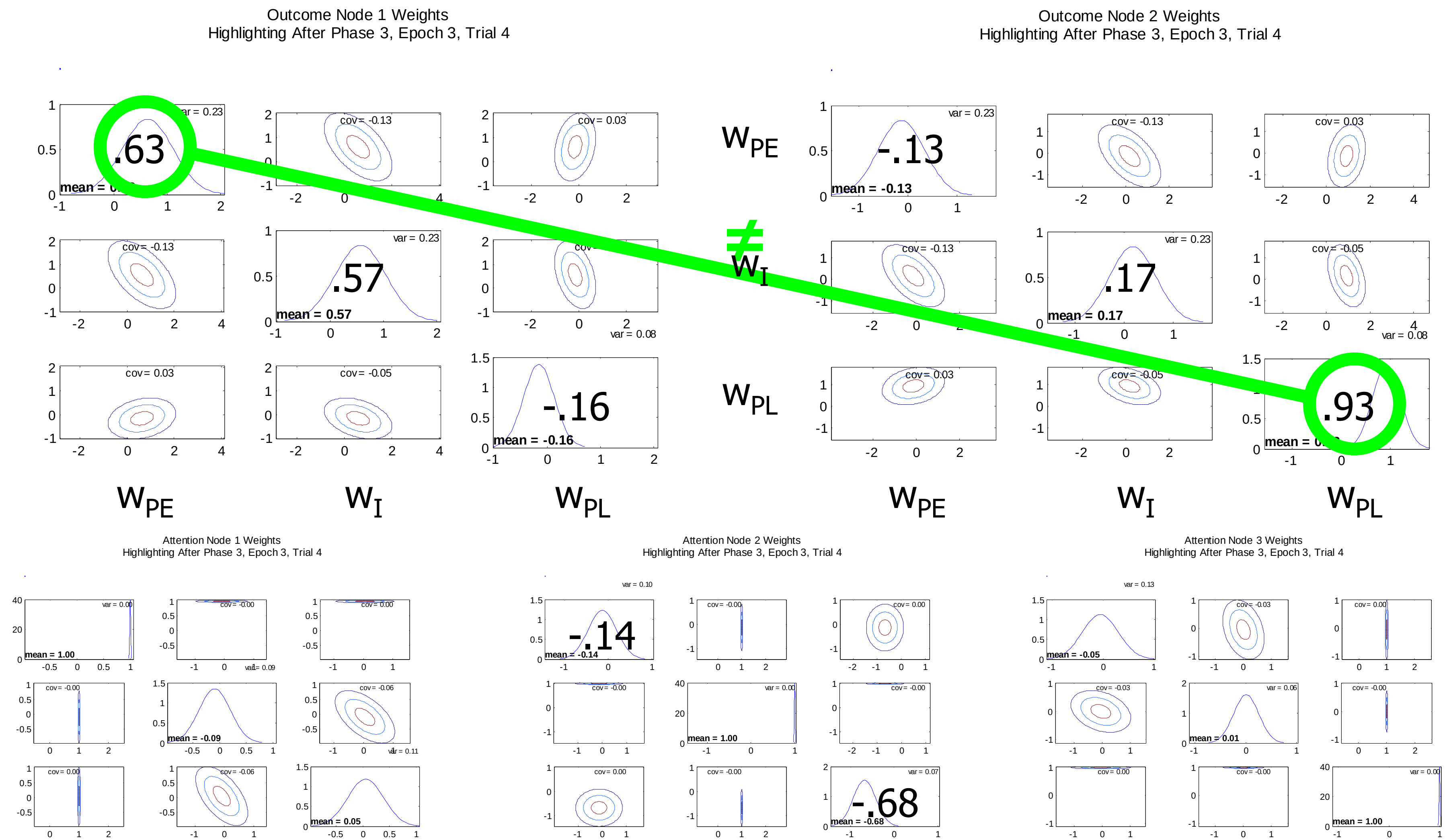
Layers of Kalman Filters Applied to Highlighting:
Final p(w)



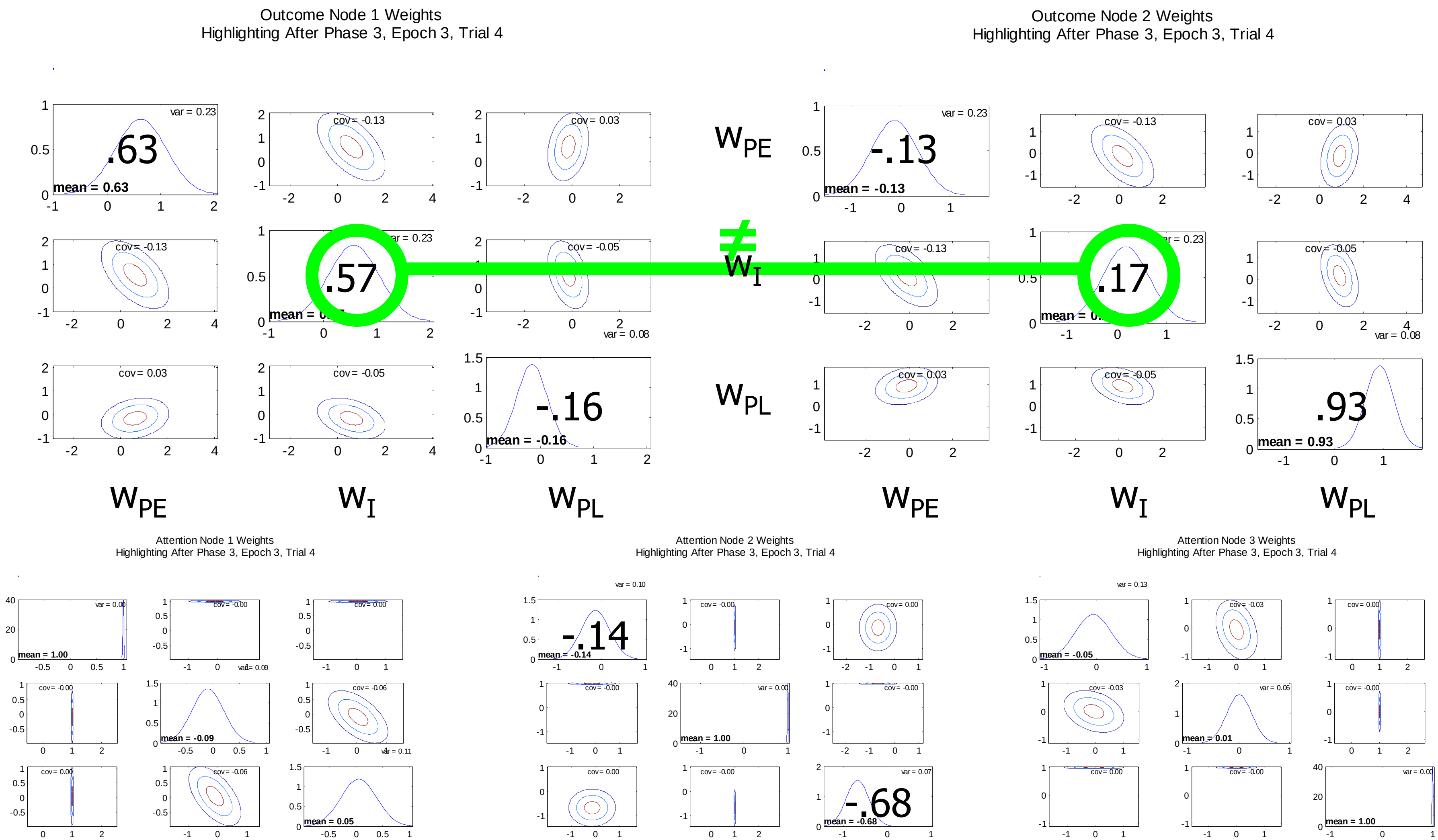


Layers of Kalman Filters Applied to Highlighting:

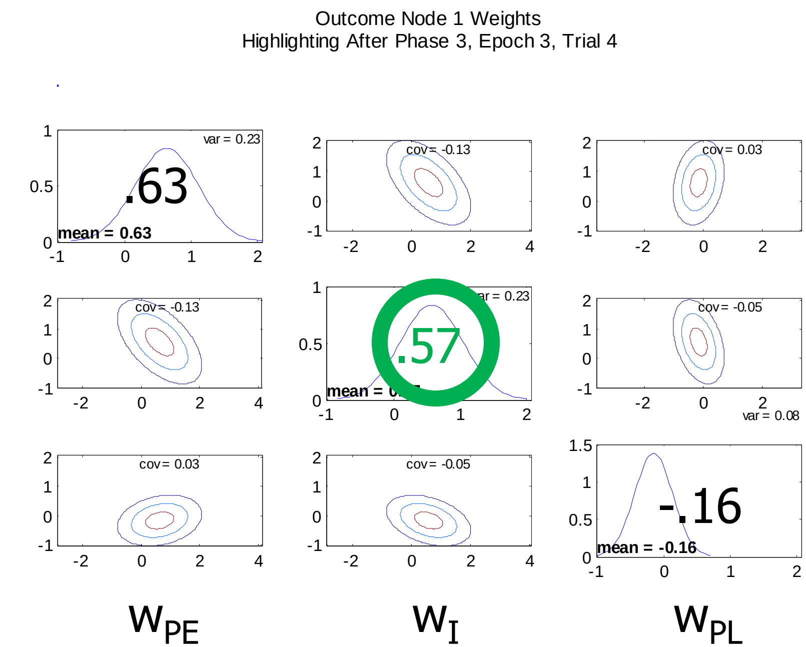
Final p(w)



Layers of Kalman Filters Applied to Highlighting:
Final p(w)



Layers of Kalman Filters Applied to Highlighting:
Final p(w)



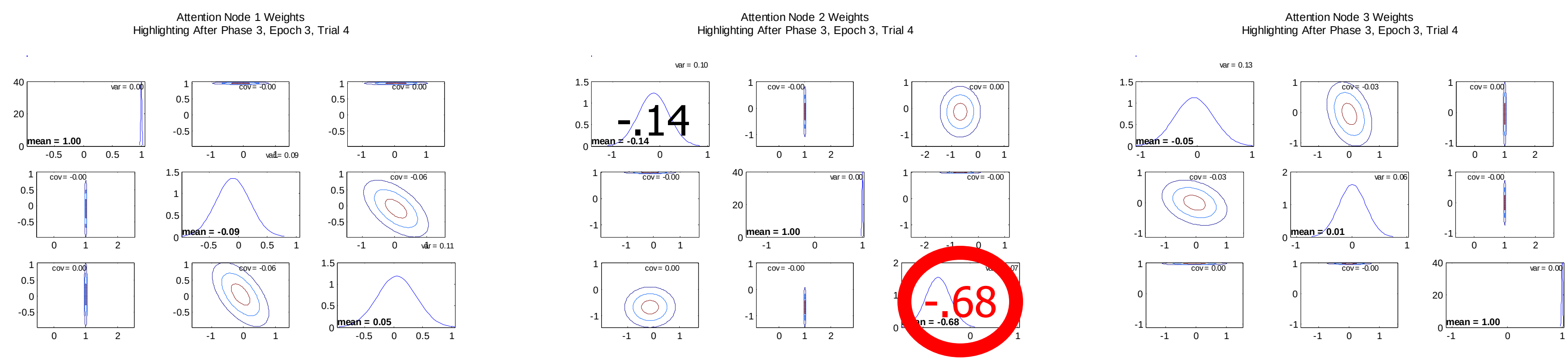
Outcome Node 2 Weights
Highlighting After Phase 3, Epoch 3, Trial 4

W_{PE}

W_I

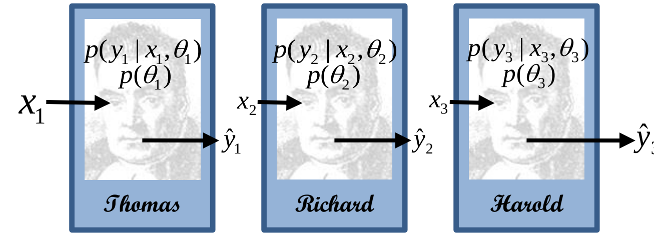
W_{PL}

Inhibition of I by PL **prevents disconfirmation** of previous learning that I→E.

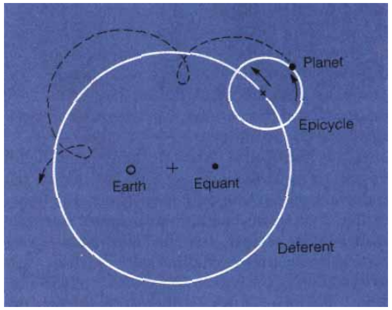


Summary

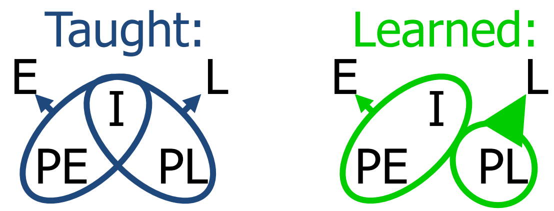
- Different levels of analysis invite possibility of a chain of Bayesian learners.



- Locally Bayesian learning prevents disconfirmation of superior's beliefs and creates distortions in inferior's beliefs.



- Locally Bayesian learning was applied to attentional shifts in associative learning, specifically to account for "highlighting".



Future Directions

- Better models and priors for application to associative learning, to expand scope and quantitatively fit human learning.
- Applications to other domains and phenomena. (Please suggest!)
- Formal analysis of global behavior of system of Bayesian agents.