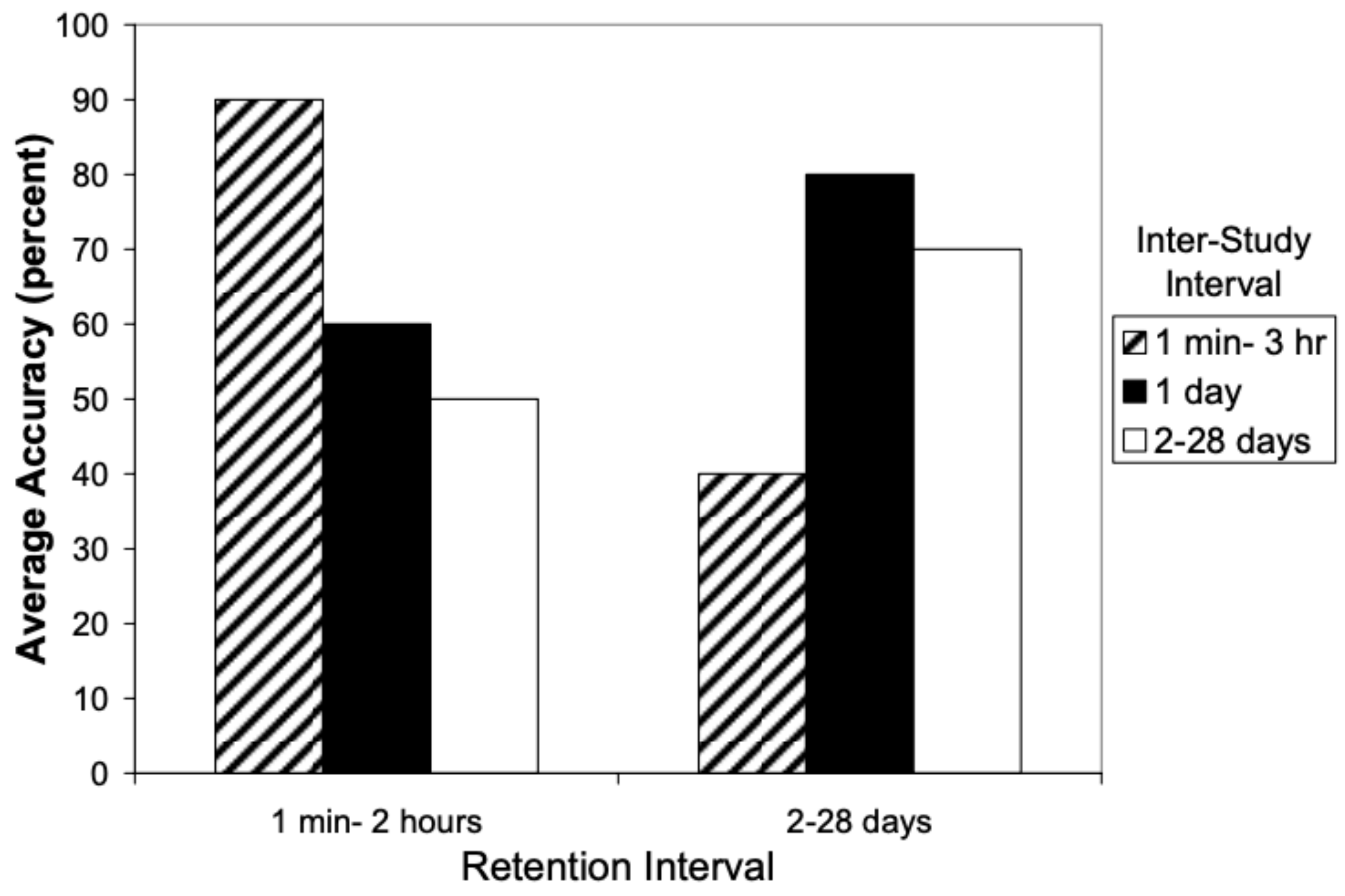
Wrapping up

12/10/2020

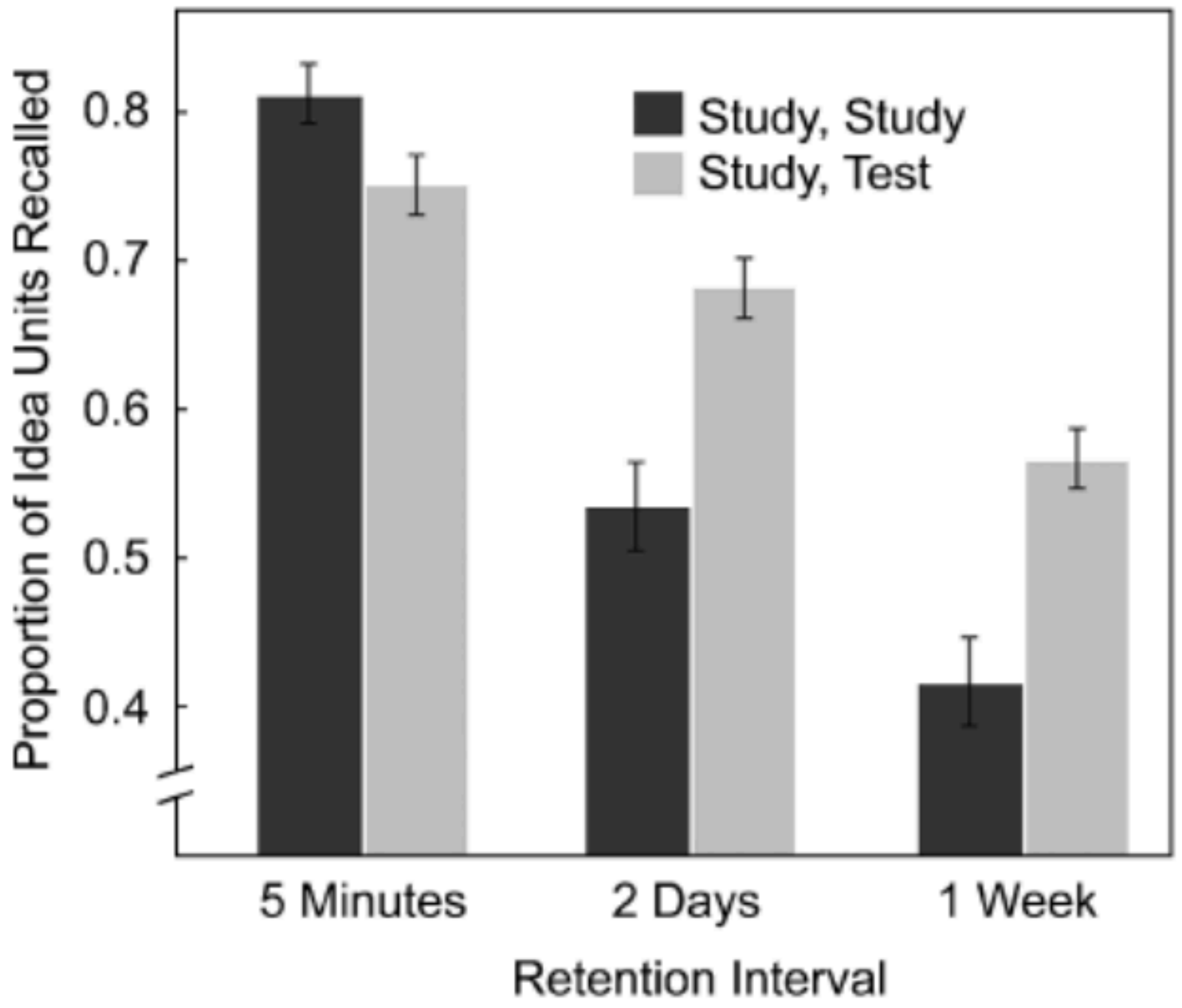
Two important reliable effects in education psychology

Spaced learning is better than massed learning



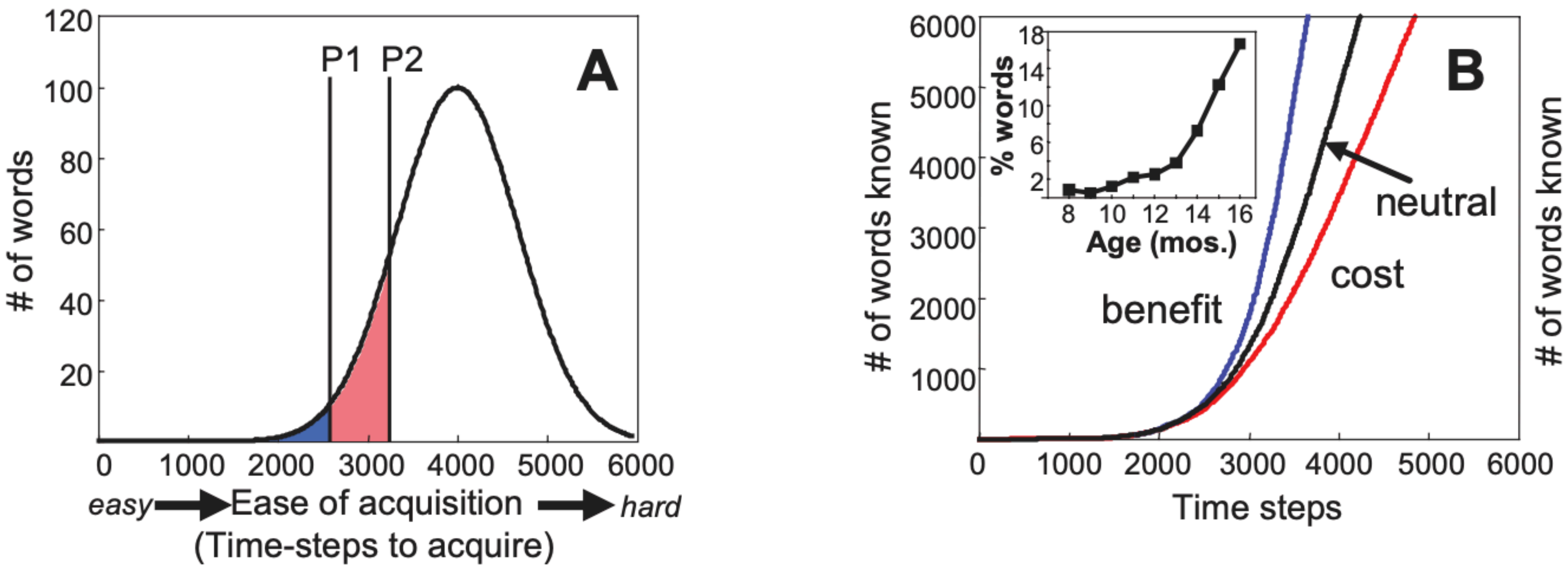
Cepeda et al. (2006)

Testing helps you learn



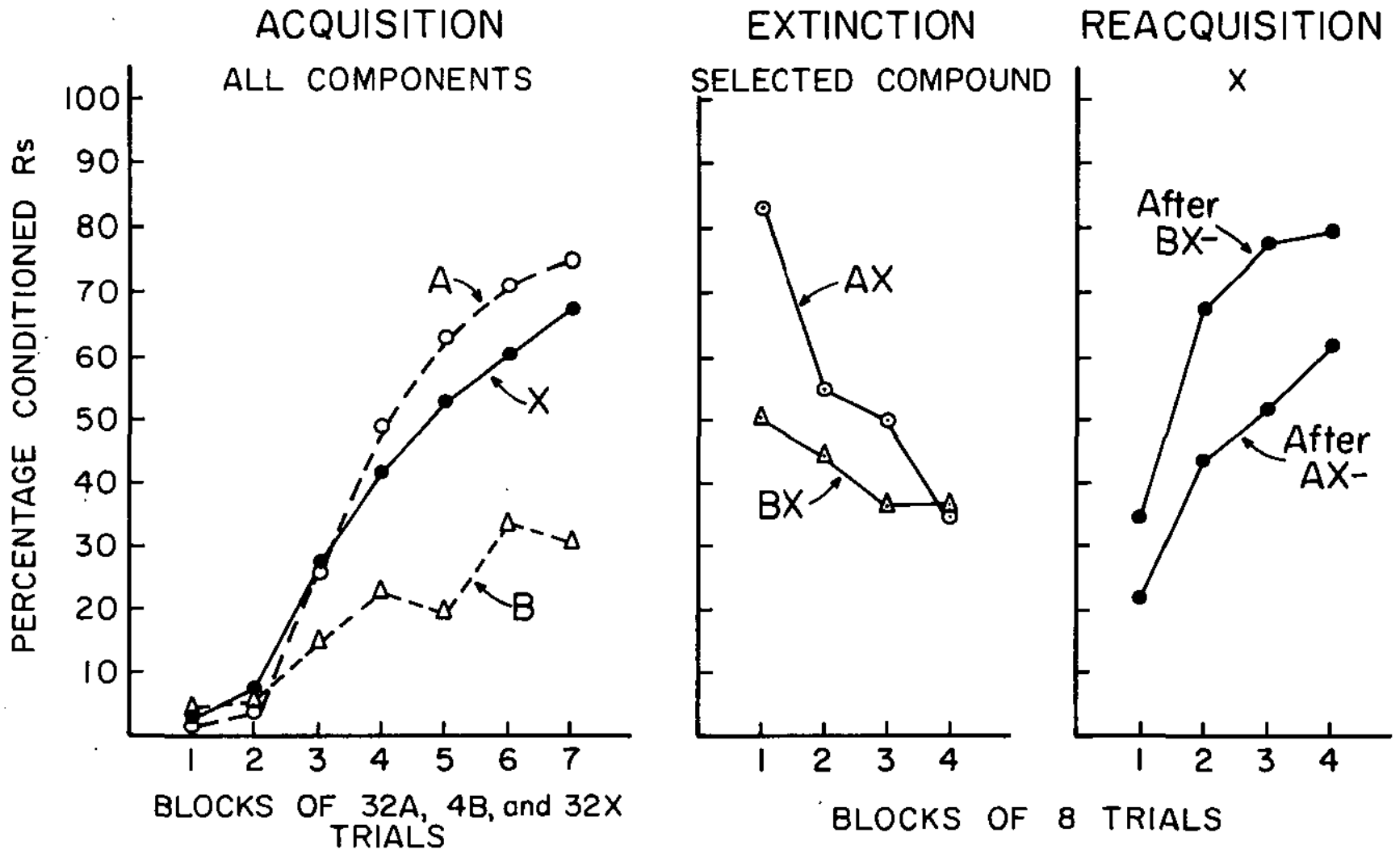
Roediger & Karpicke (2006)

The vocabulary spurt as a null hypothesis

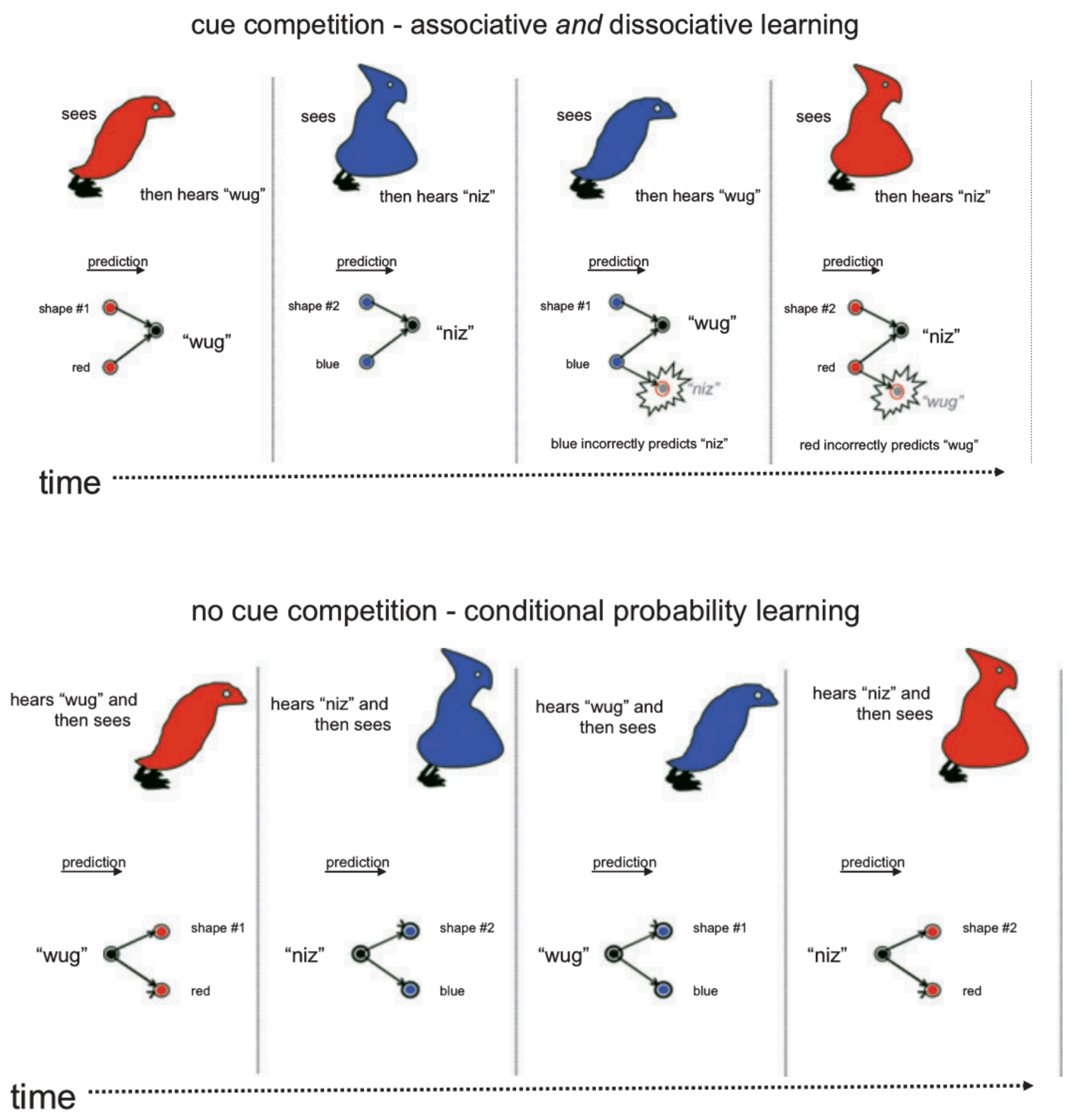


McMurray (2007)

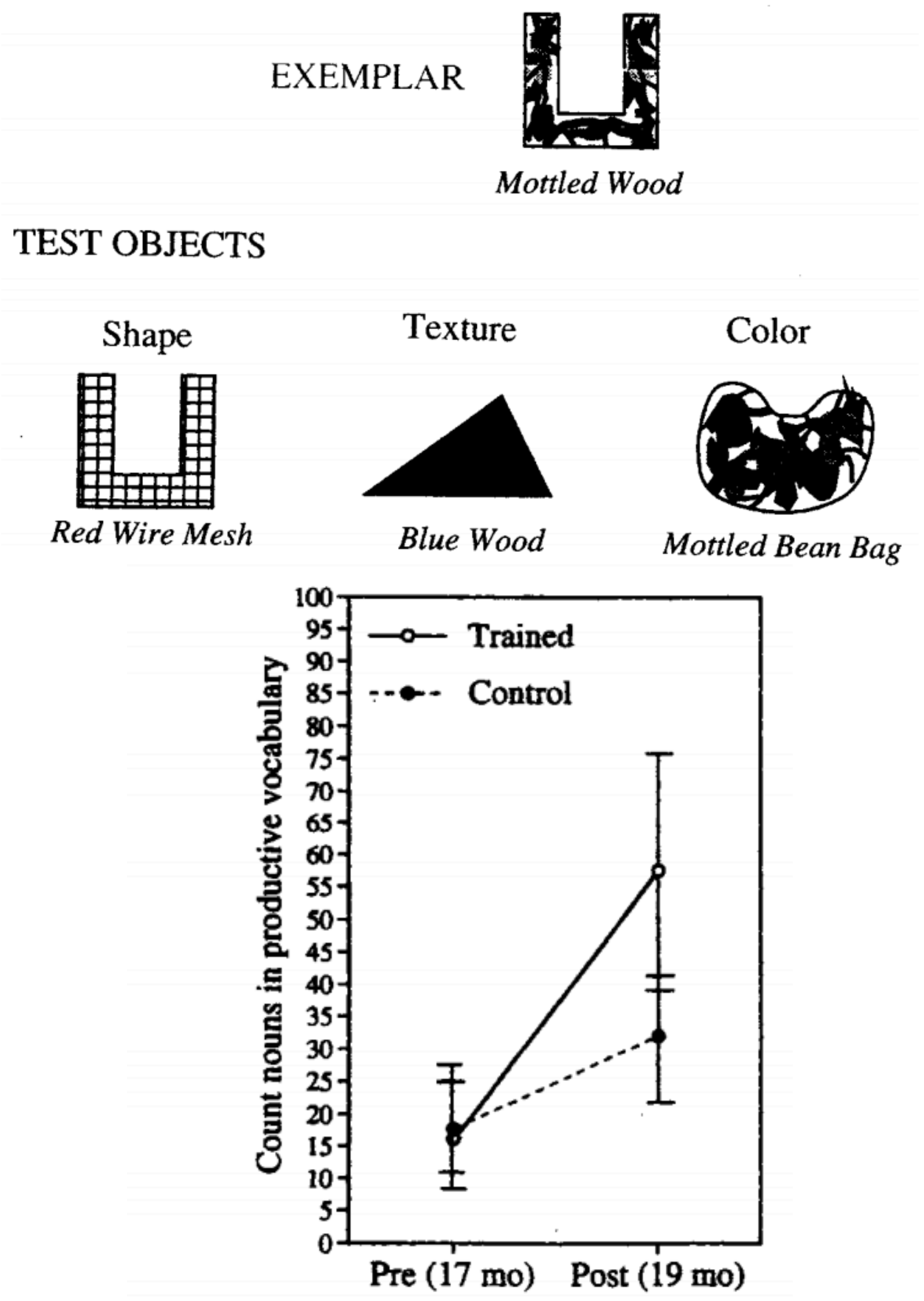
The key experiment in Rescorla & Wagner (1972)



Associative learning can explain some interesting phenomena

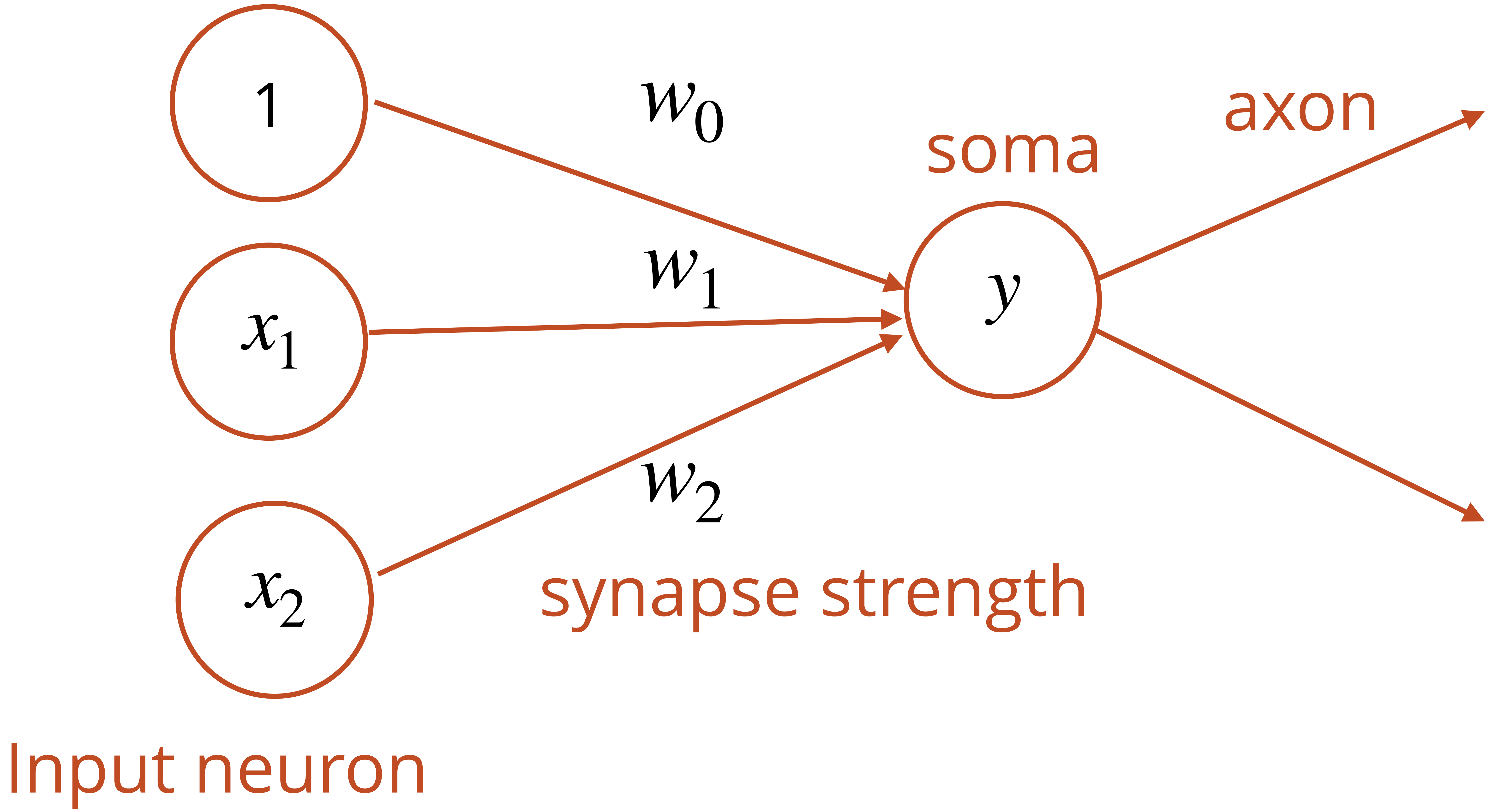


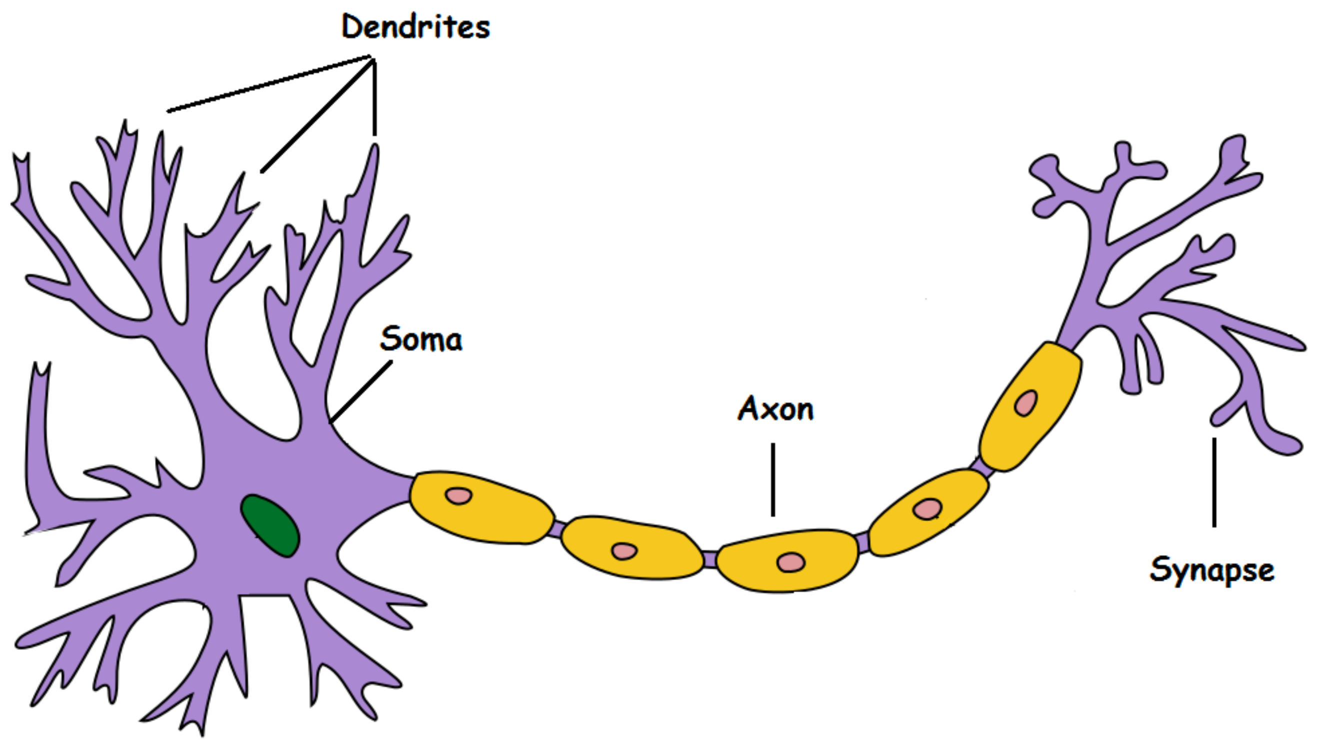
Ramscar et al. (2010)



Smith (2000)

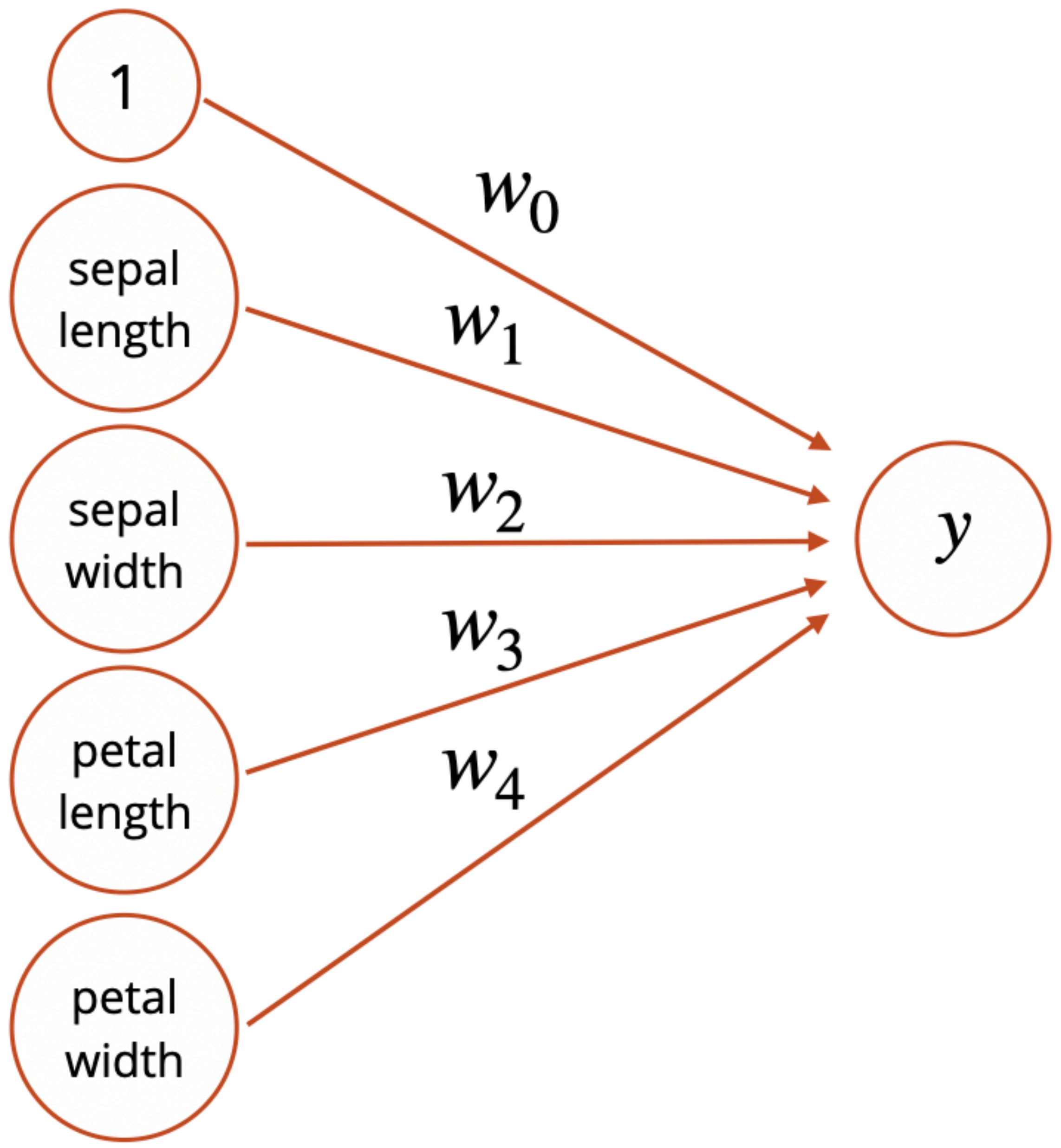
An artificial neuron





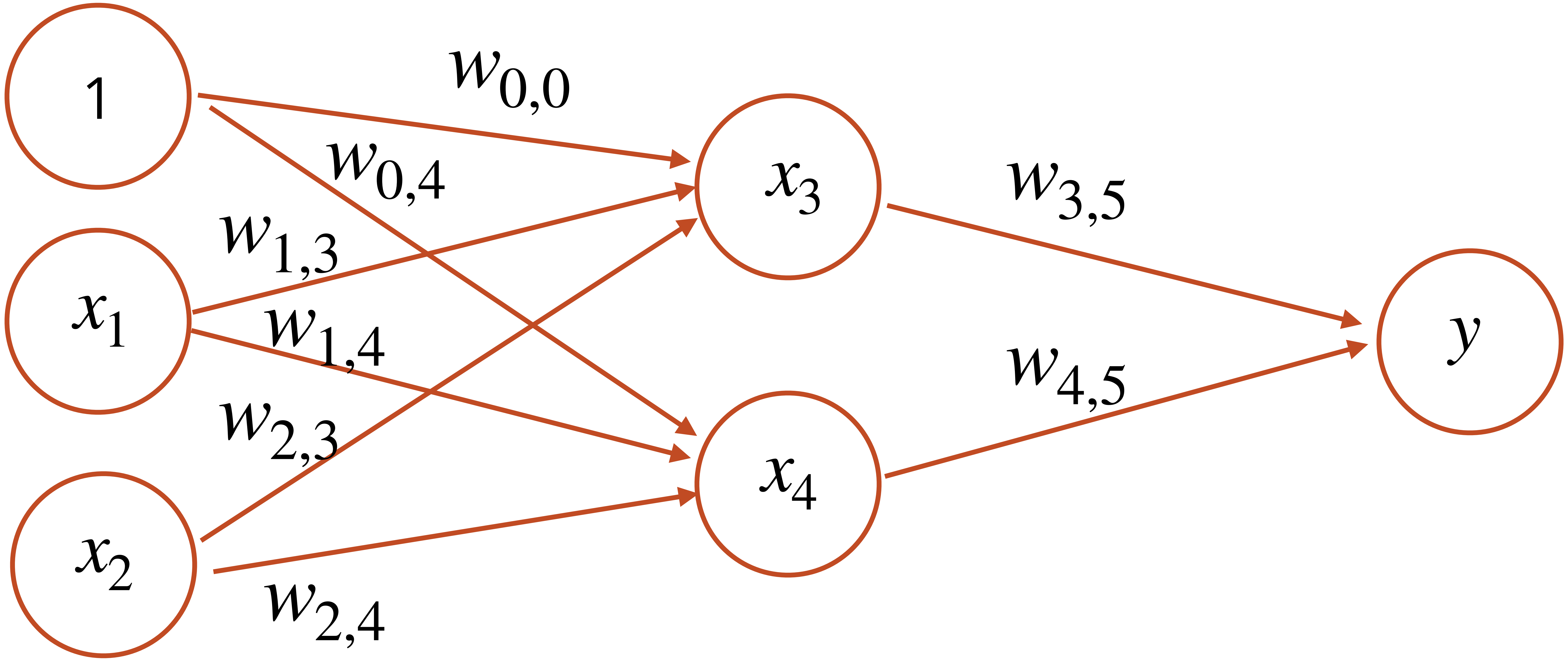
Logistic regression as an iris classifier

```
glm(Species ~ Sepal.Length + Sepal.Width
              + Petal.Length + Petal.Width,
              family = “binomial”)
```



term	estimate	std.error	statistic	p.value
(Intercept)	-42.638	25.707	-1.659	.097
Sepal.Length	-2.465	2.394	-1.030	.303
Sepal.Width	-6.681	4.480	-1.491	.136
Petal.Length	9.429	4.737	1.991	.047
Petal.Width	18.286	9.743	1.877	.061

How would a network solve xor?



x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

Updating one weight

Terms:

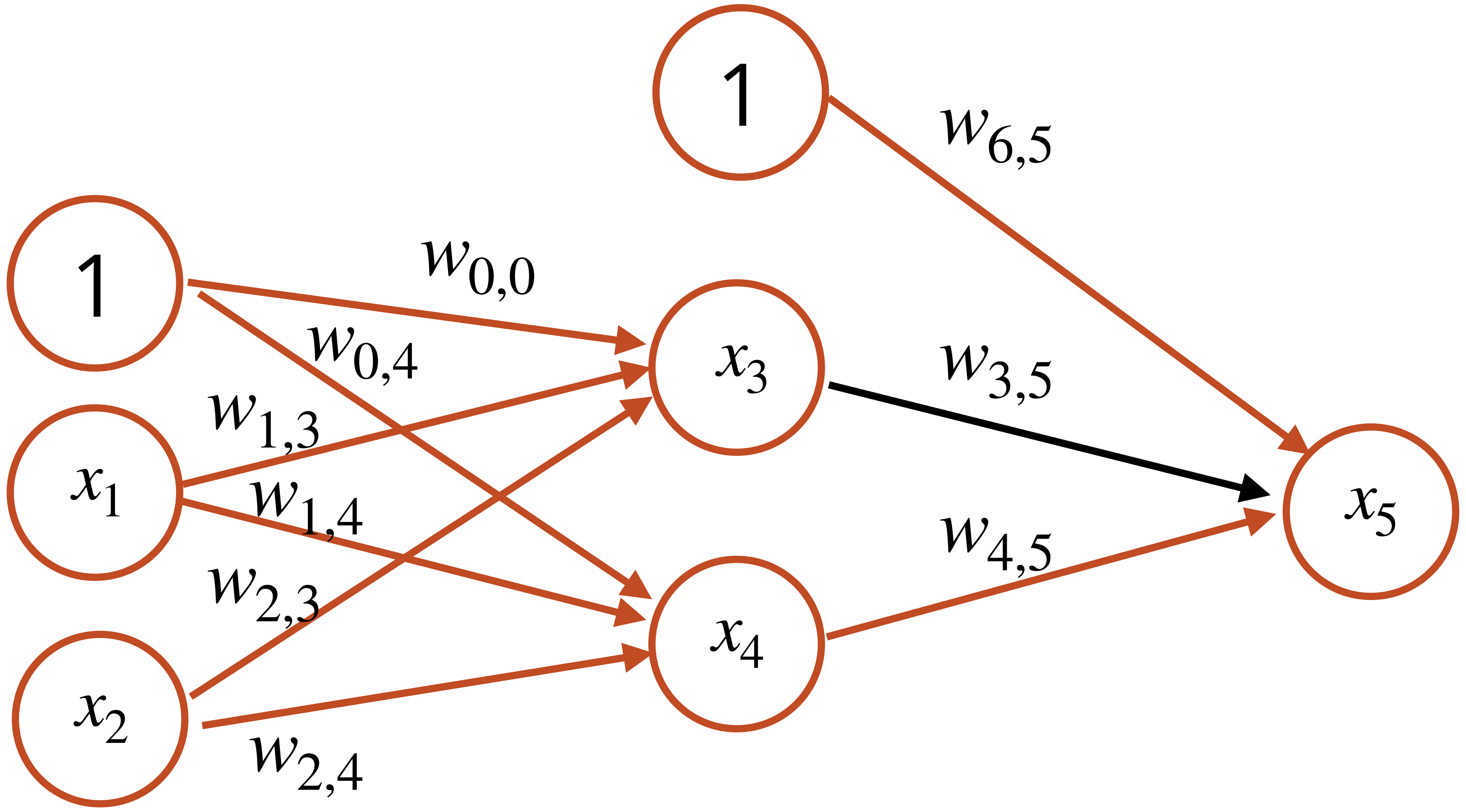
E Squared Prediction error

x_5 The summed input to $x_5 = w_{6,5} + w_{3,5} \cdot x_3 + w_{4,5} \cdot x_4$

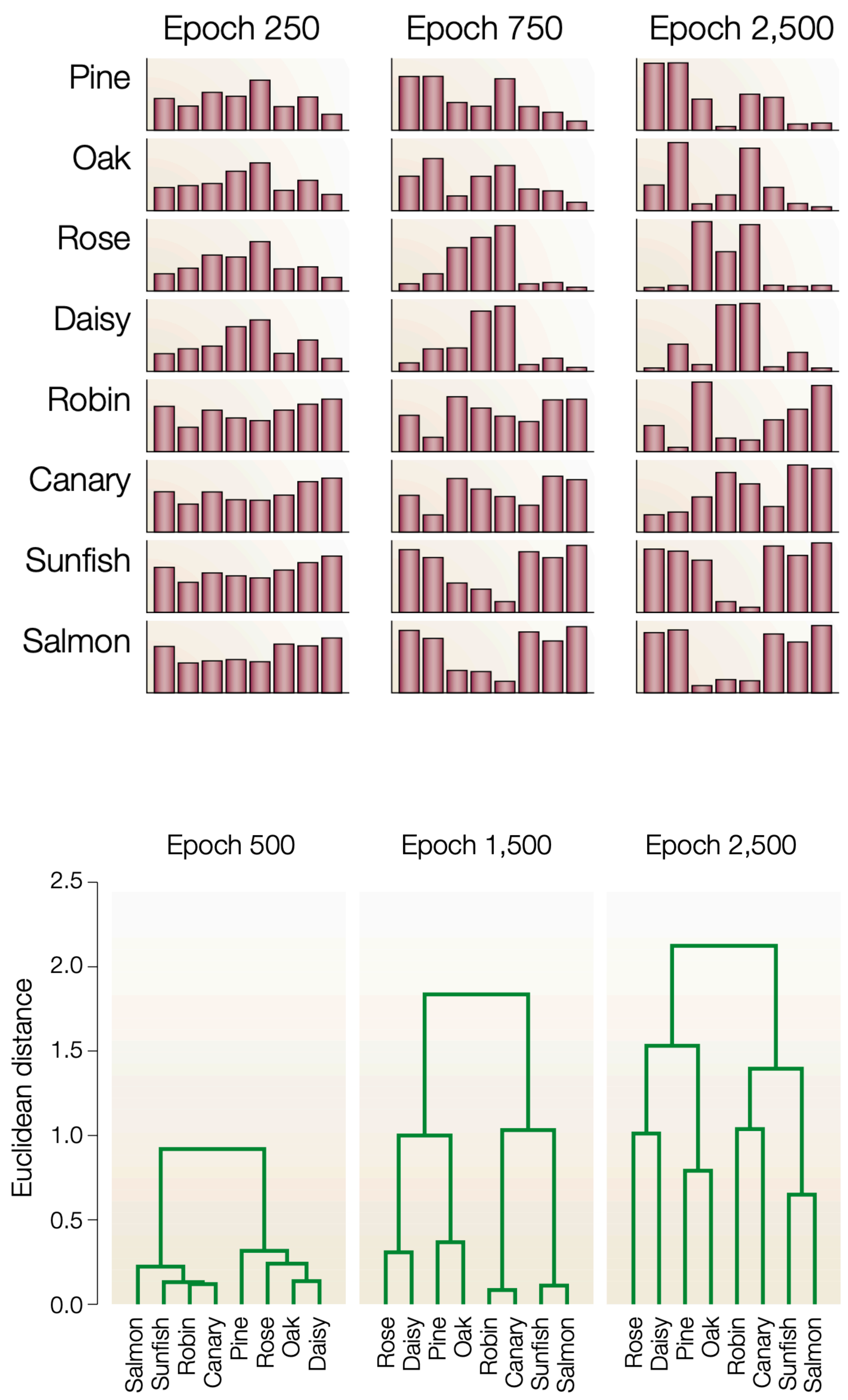
a_{x_5} The activation of $x_5 = \dfrac{1}{1 - e^{x_5}} = \sigma\left(x_5\right)$

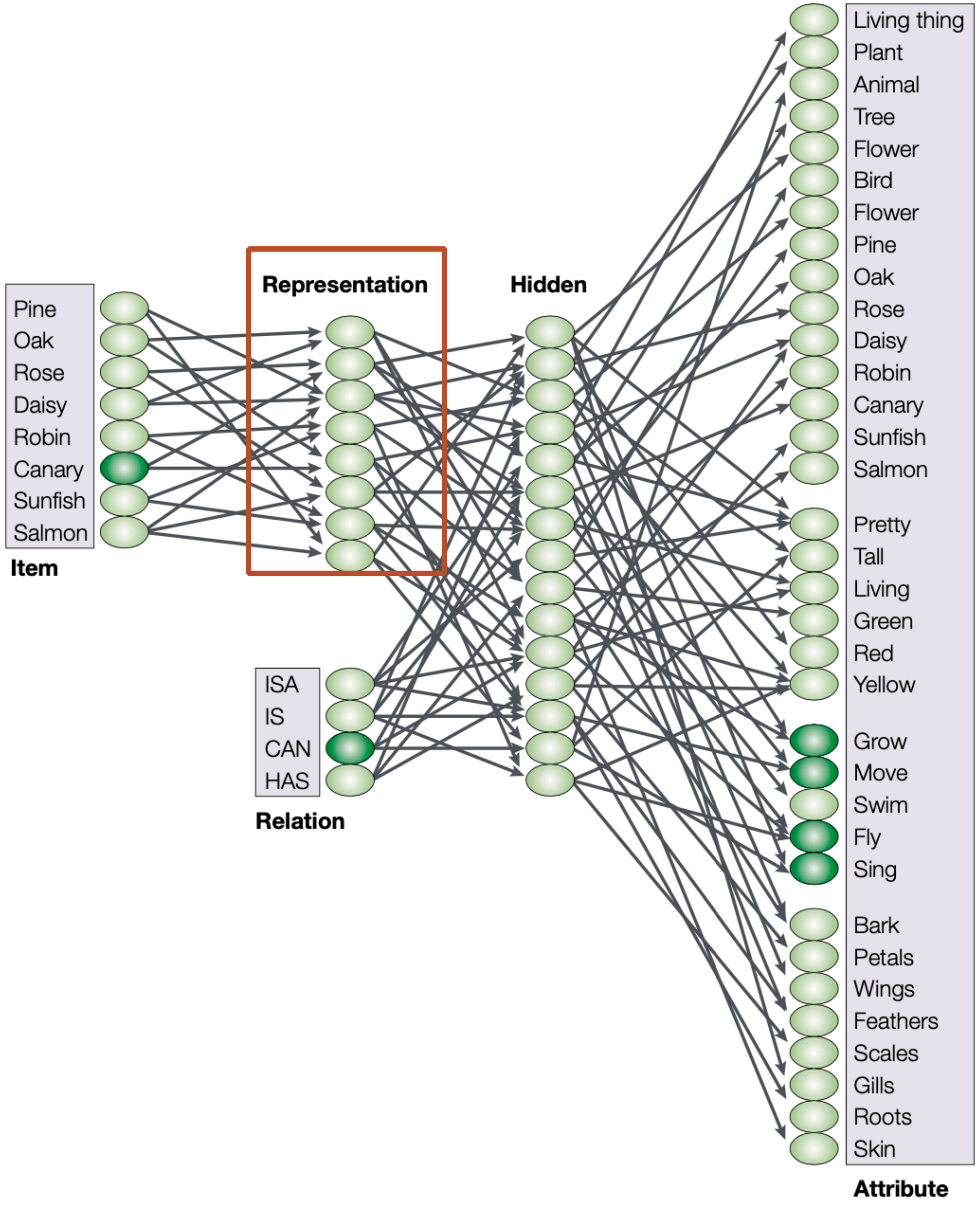
$$\frac{\partial E}{\partial w_{3,5}} = \frac{\partial x_5}{\partial w_{3,5}} \frac{\partial a_{x_5}}{\partial x_5} \frac{\partial E}{\partial a_{x_5}}$$

By the chain rule

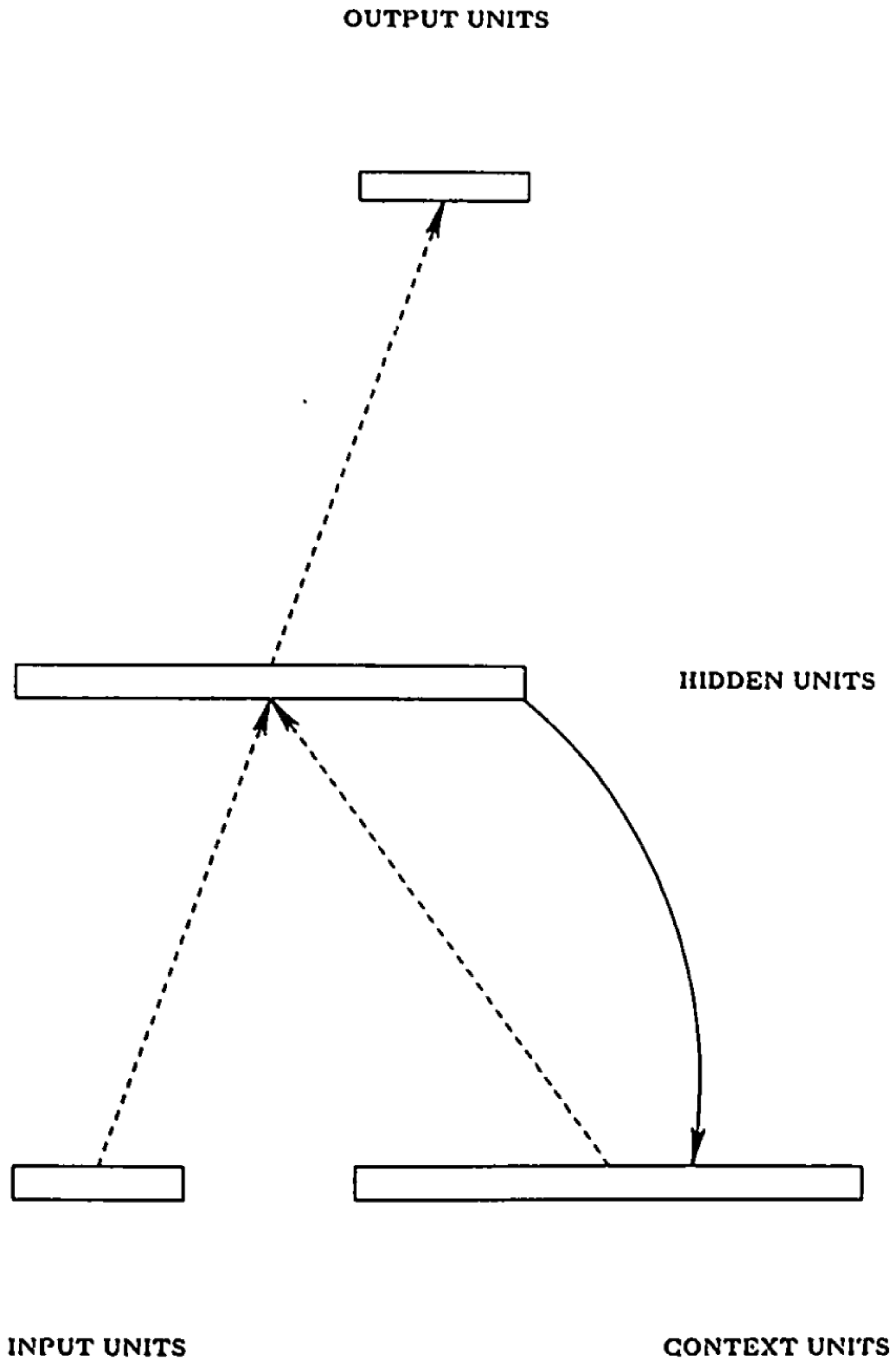


Learning semantic relations through backpropagation





Simple Recurrent Networks (Elman Networks - Elman, 1990)



A set of **context units** that are an exact copy of the hidden layer at $t-1$

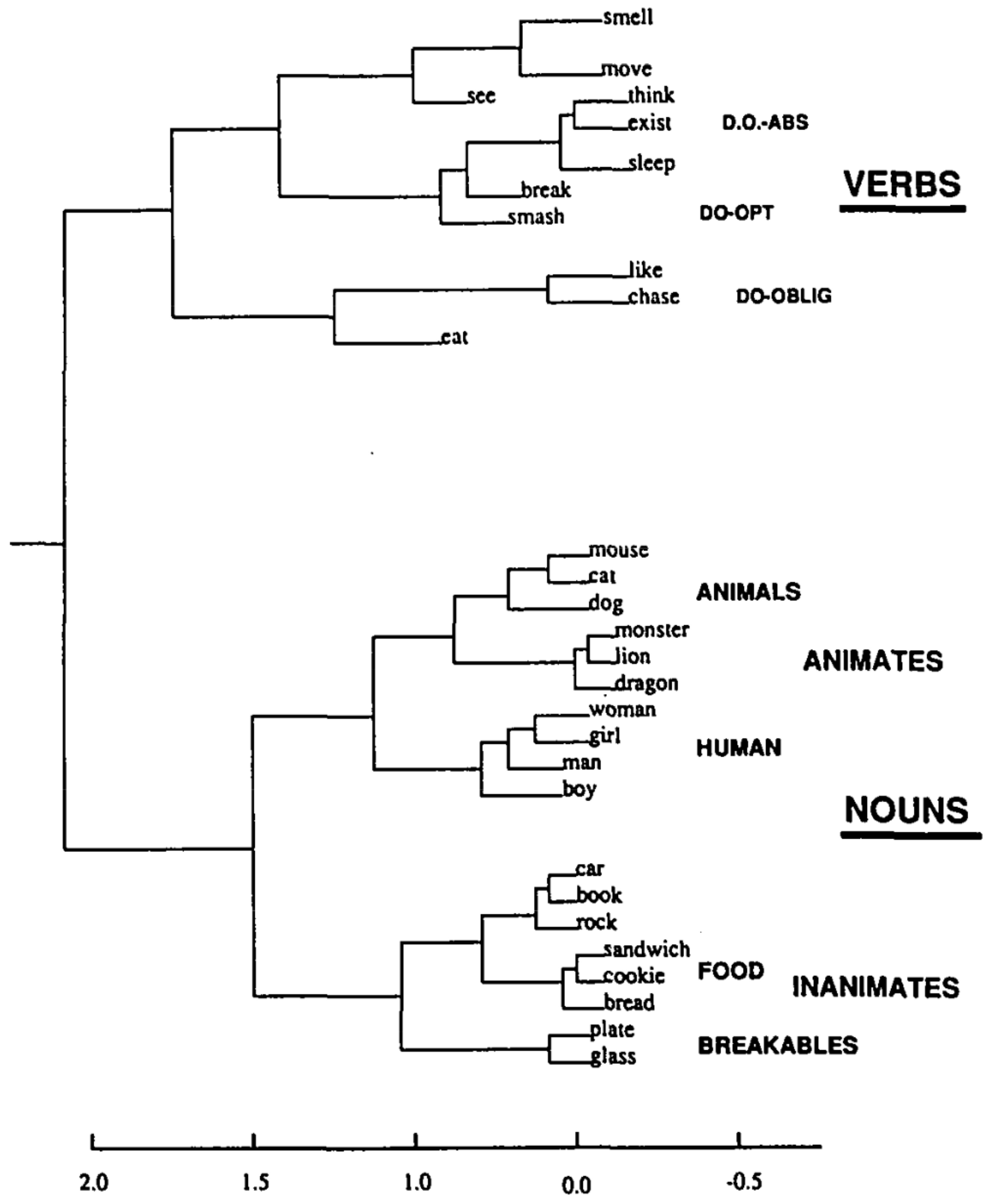
The hidden layer at time t gets input from both the **input units** and the **context units**

Syntactic structure through prediction error

Hierarchically clustering the hidden layer activations for words reveals structure!

The network learns **syntactic** and **semantic** roles

Why?



Strengths of connectionism

1. Each unit of the network is a simple computer, but the network as a whole can give rise to complex phenomena.

2. The framework is general—you don't need a separate model for every domain (sort of).

3. Blurs the hardware/software distinction

But what about rules? And other problems

Exp.	Mean listening time (s) (SE)		Repeated measures analysis of variance
	Consistent sentences	Inconsistent sentences	
1	6.3 (0.65)	9.0 (0.54)	$F(14) = 25.7, P < 0.001$
2	5.6 (0.47)	7.35 (0.68)	$F(14) = 25.6, P < 0.005$
3	6.4 (0.38)	8.5 (0.5)	$F(14) = 40.3, P < 0.001$

Marcus (1999)

For most problems where deep learning has enabled transformationally better solutions (vision, speech), we've entered diminishing returns territory in 2016-2017.

François Chollet, Google, author of Keras
neural network library
December 18, 2017

'Science progresses one funeral at a time.' The future depends on some graduate student who is deeply suspicious of everything I have said.

Geoff Hinton, grandfather of deep learning
September 15, 2017

Marcus (2018)

But how should you form your beliefs?

In practice, we don't want to say you can have any old belief. We want to talk about the belief that a **rational** agent should have after observing some data

Likelihood
(What the data say)

Prior probability
(What you used to believe)

Bayes rule: $$P\left(H|D\right) = \frac{P\left(D|H\right)P\left(H\right)}{P\left(D\right)}$$

Posterior probability
(What you should believe)

The number game

An unknown computer program that generates from 1 to 100.
You get some random examples from this program.

60 80 10 30

What other numbers will this program generate?

51? 58? 20?

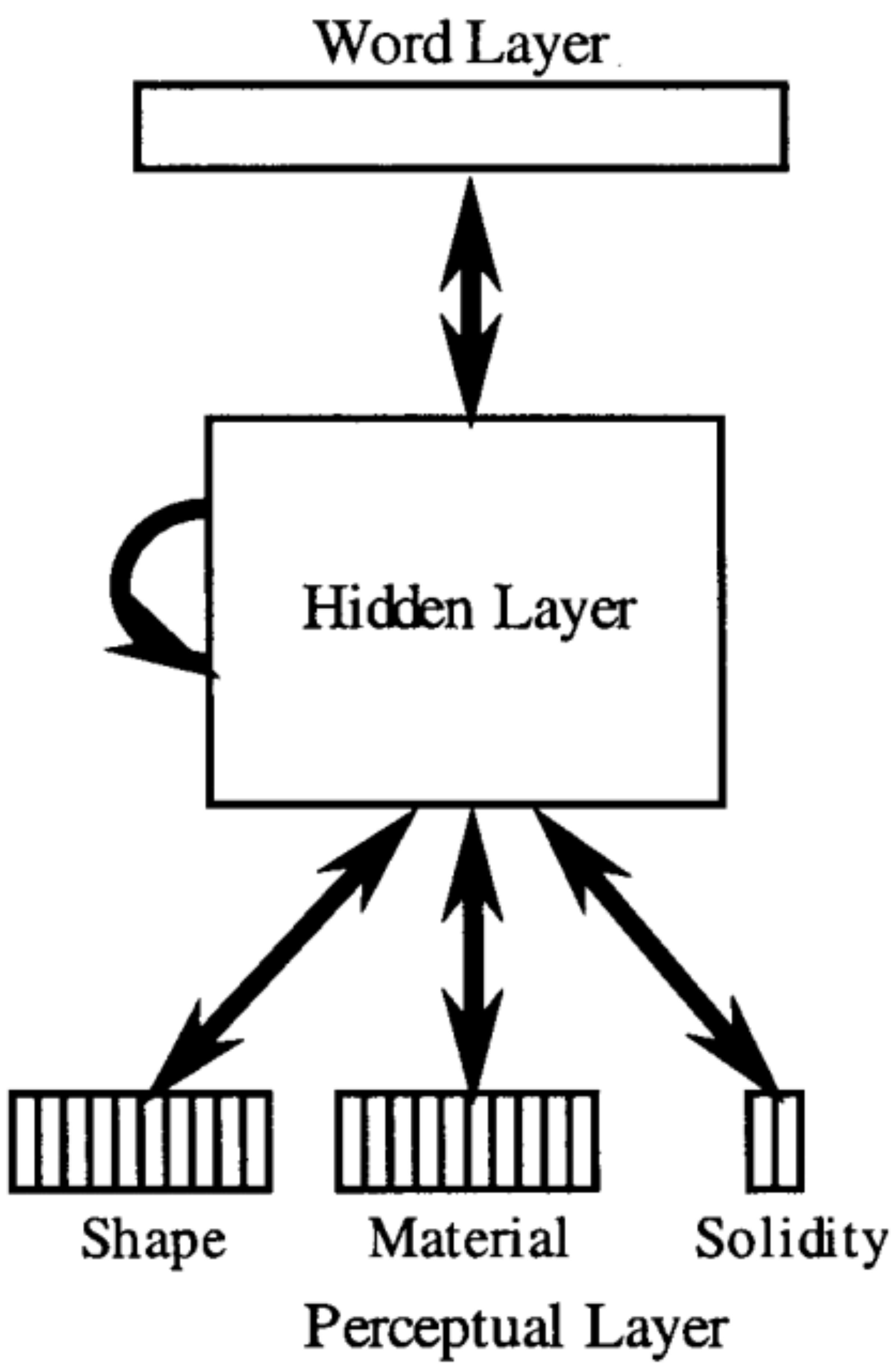
The size principle!

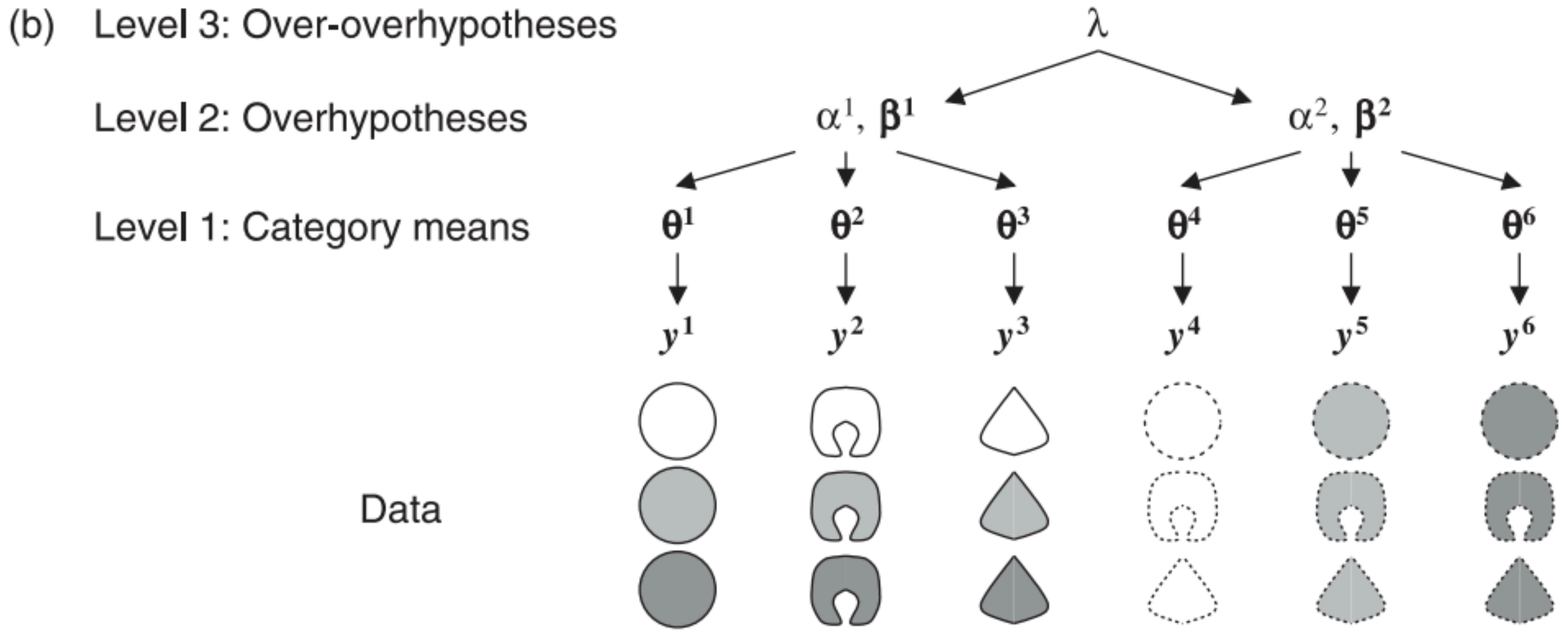
P( | dog)

<

P( | dalmation)

Models at different levels





Colunga & Smith (2005)

Kemp et al. (2007)

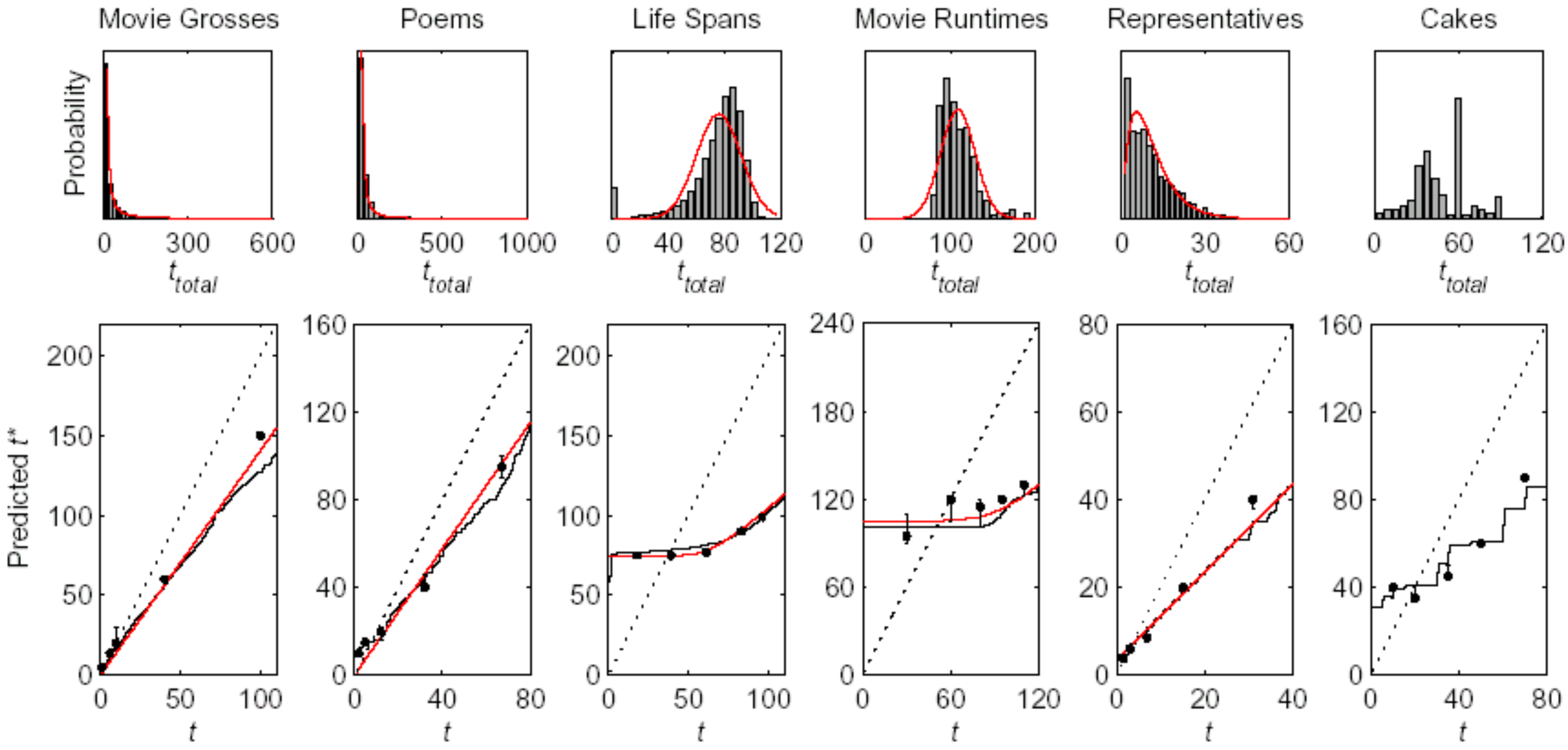
Rational analysis

For a given computational problem, there is an *optimal solution*. Whatever it is, we have evolved to approximate it.

Figure out the optimal solution, and you'll know a lot about what people do.

"The predictions flow from the statistical structure of the environment and **not** the assumed structure of the mind." (Anderson, 1991)

Evaluating peoples' predictions

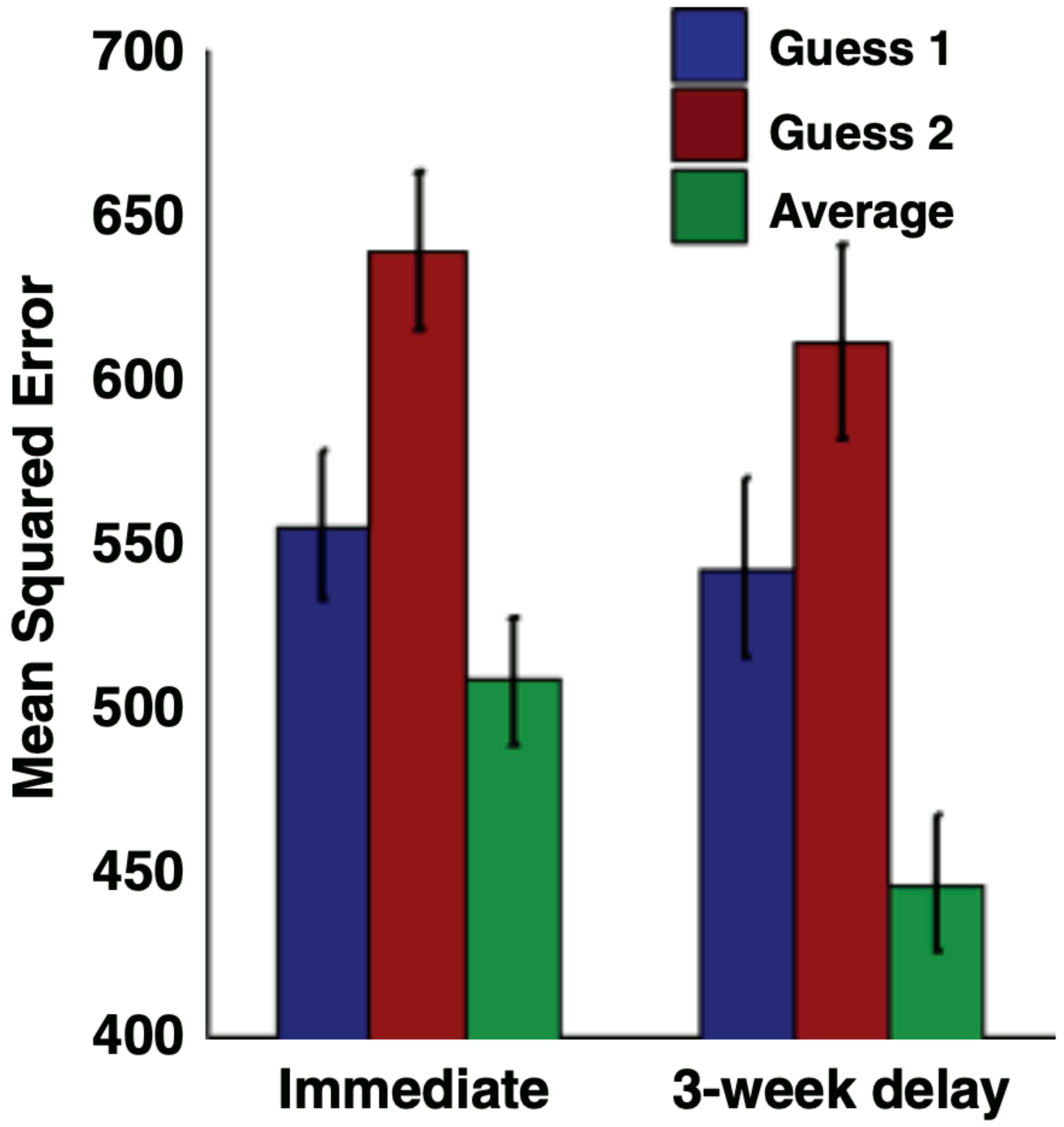


The contract

Contract: King Markov must visit each island in proportion to its population size

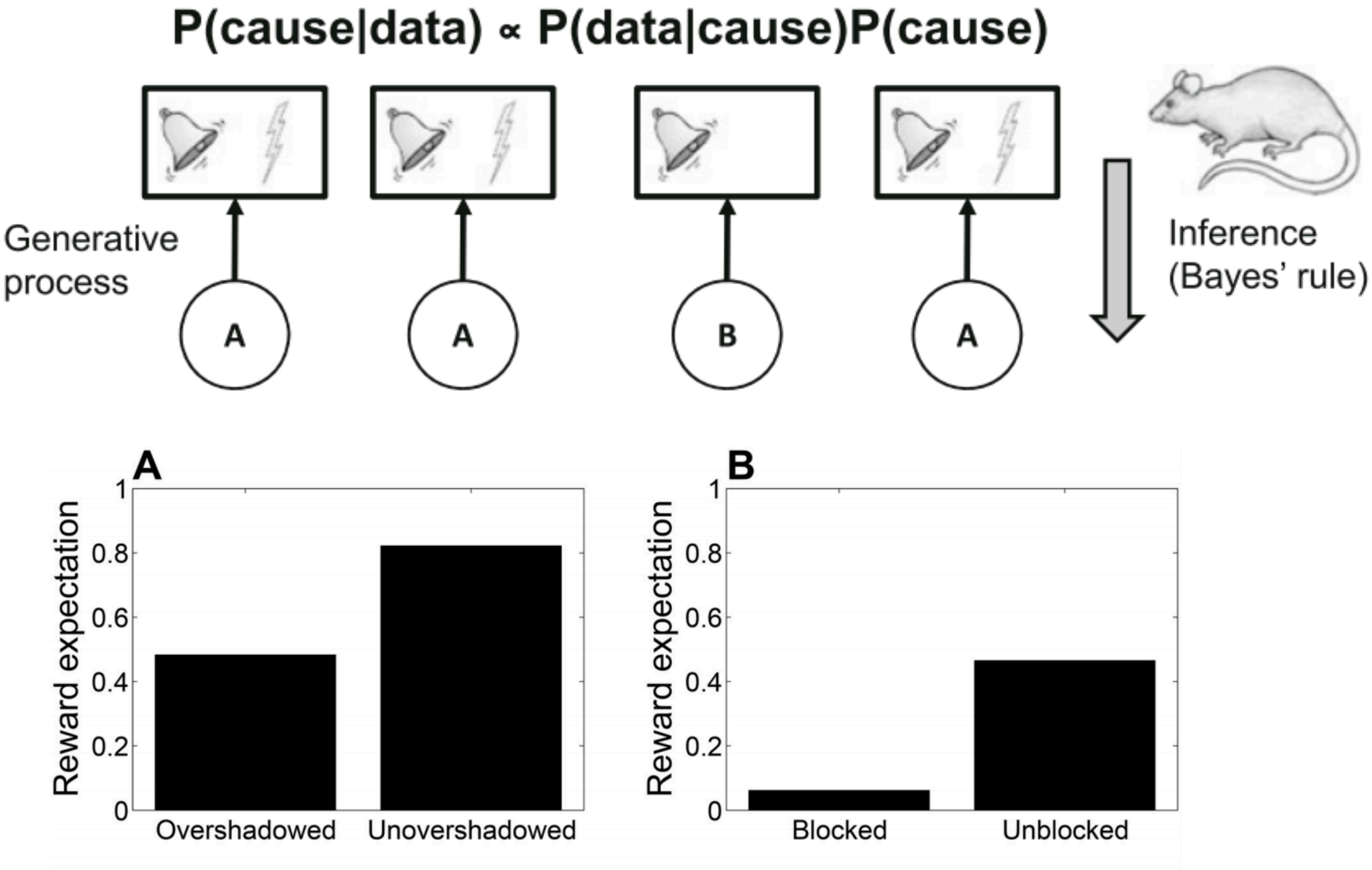
The Metropolis Archipelago

Averaging two guess from the same person is better than their best guess



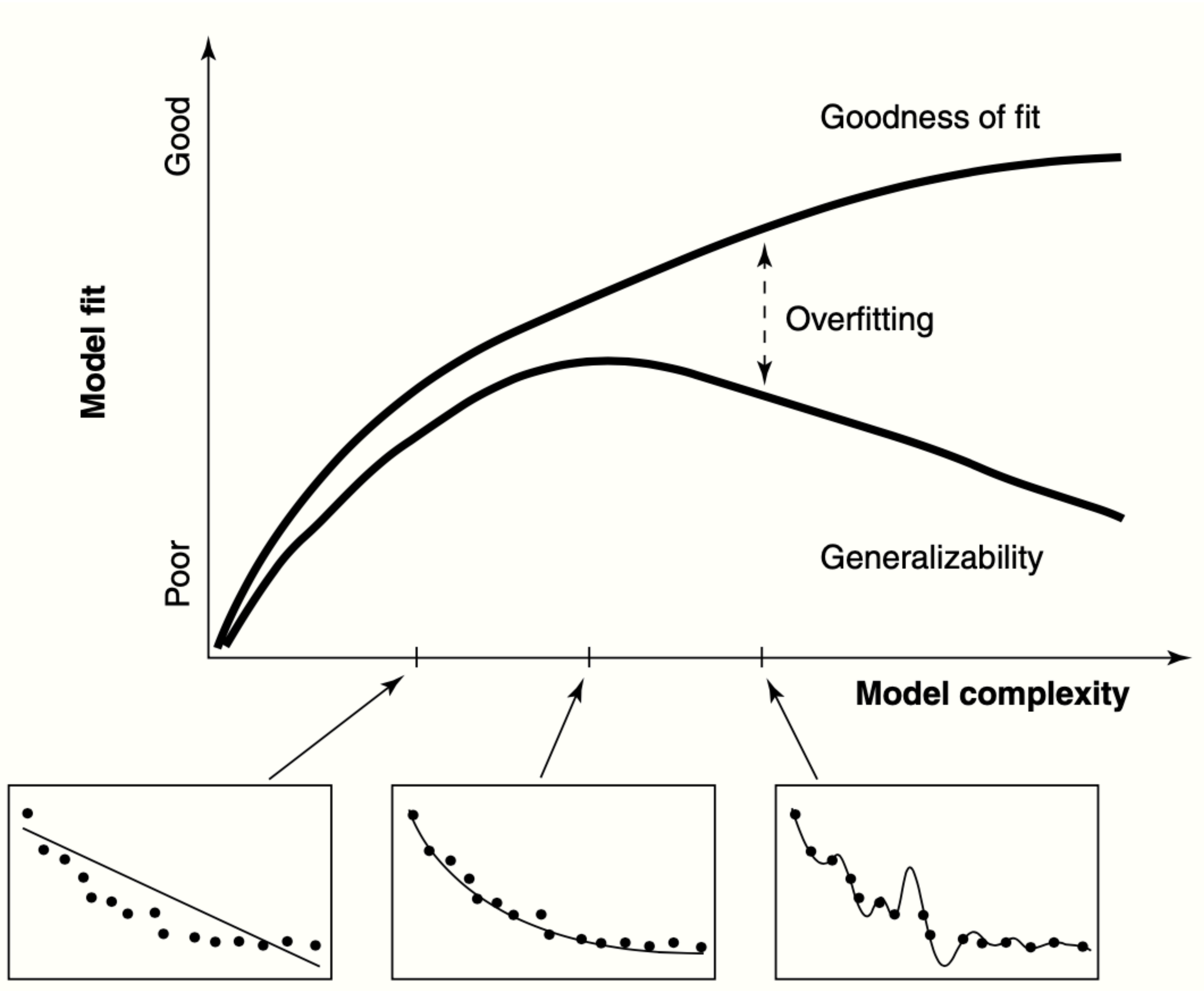
Why is asking after a longer delay better?

Bayesian associative learning

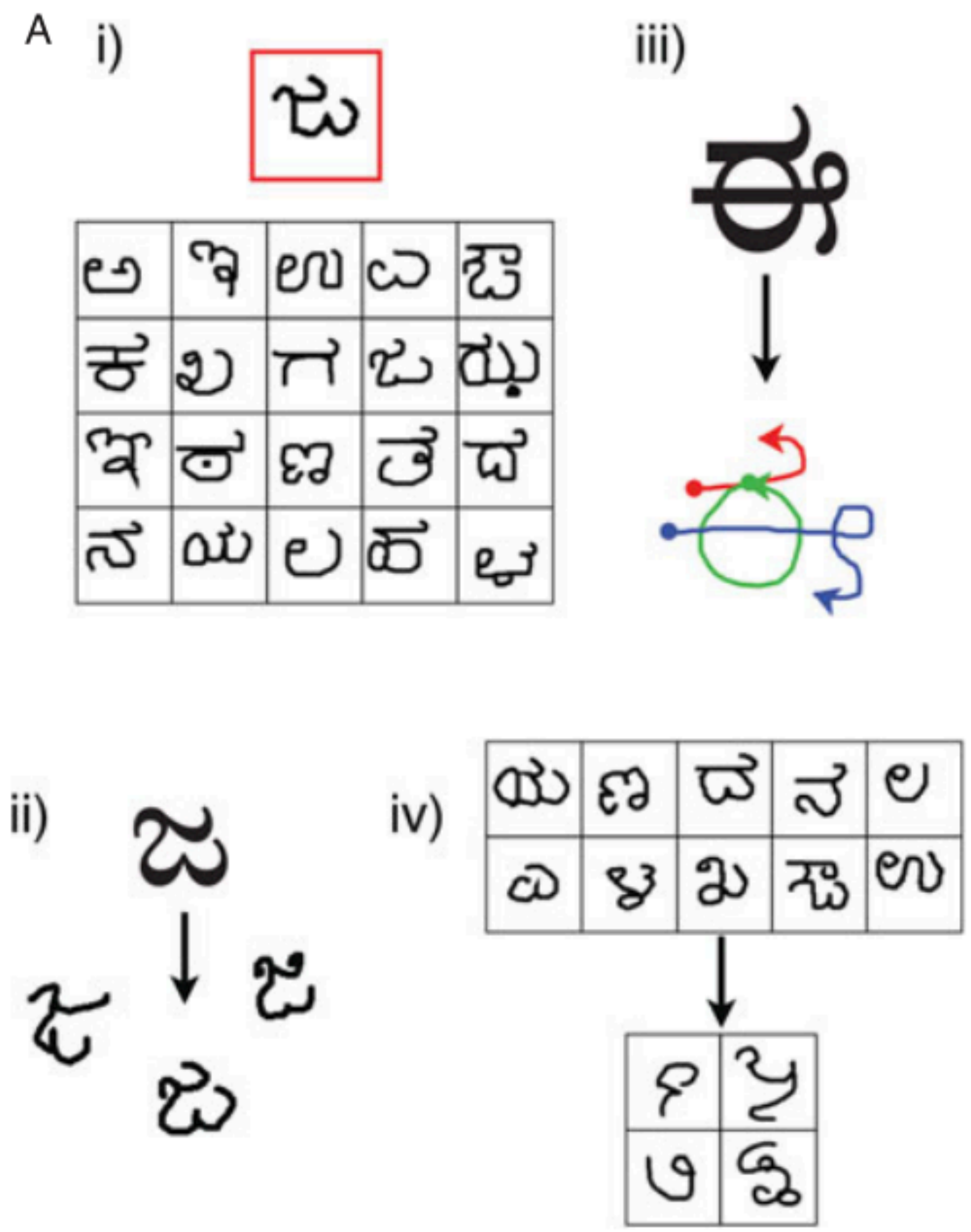


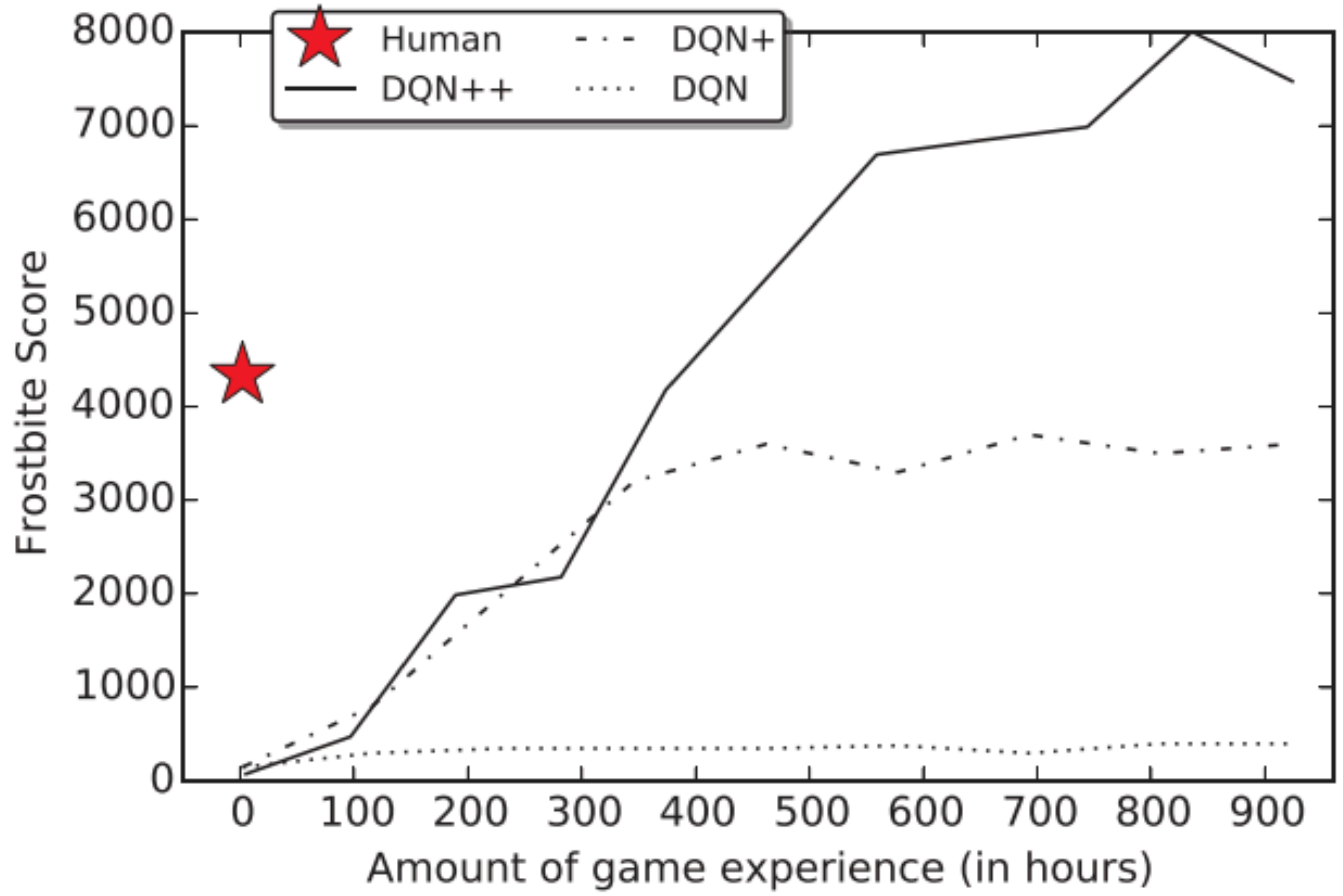
Gershman & Niv(2012); Gershman (2015)

Too much flexibility leads to overfitting (Pitt & Myung, 2002)



Machines that learn like people

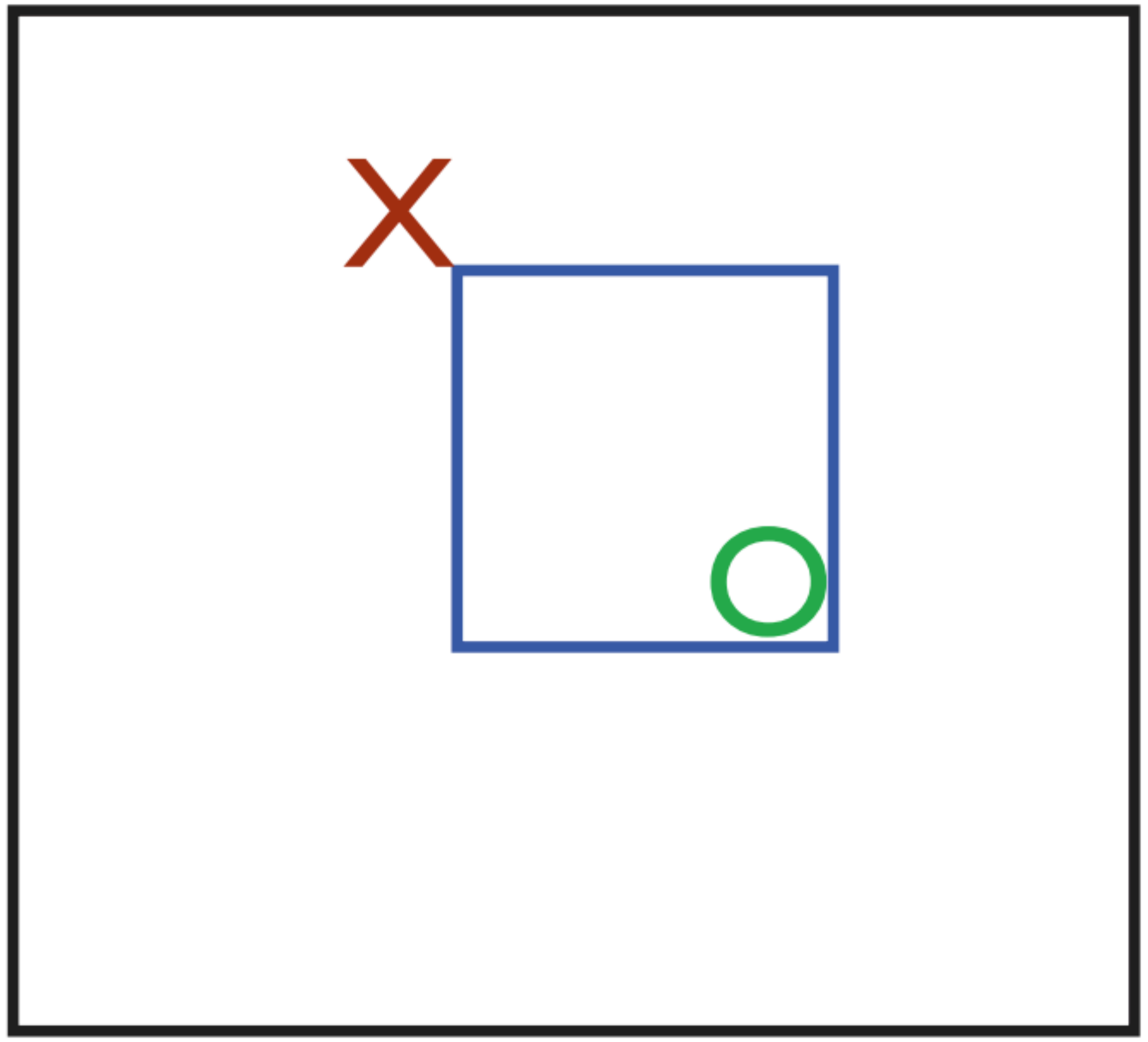




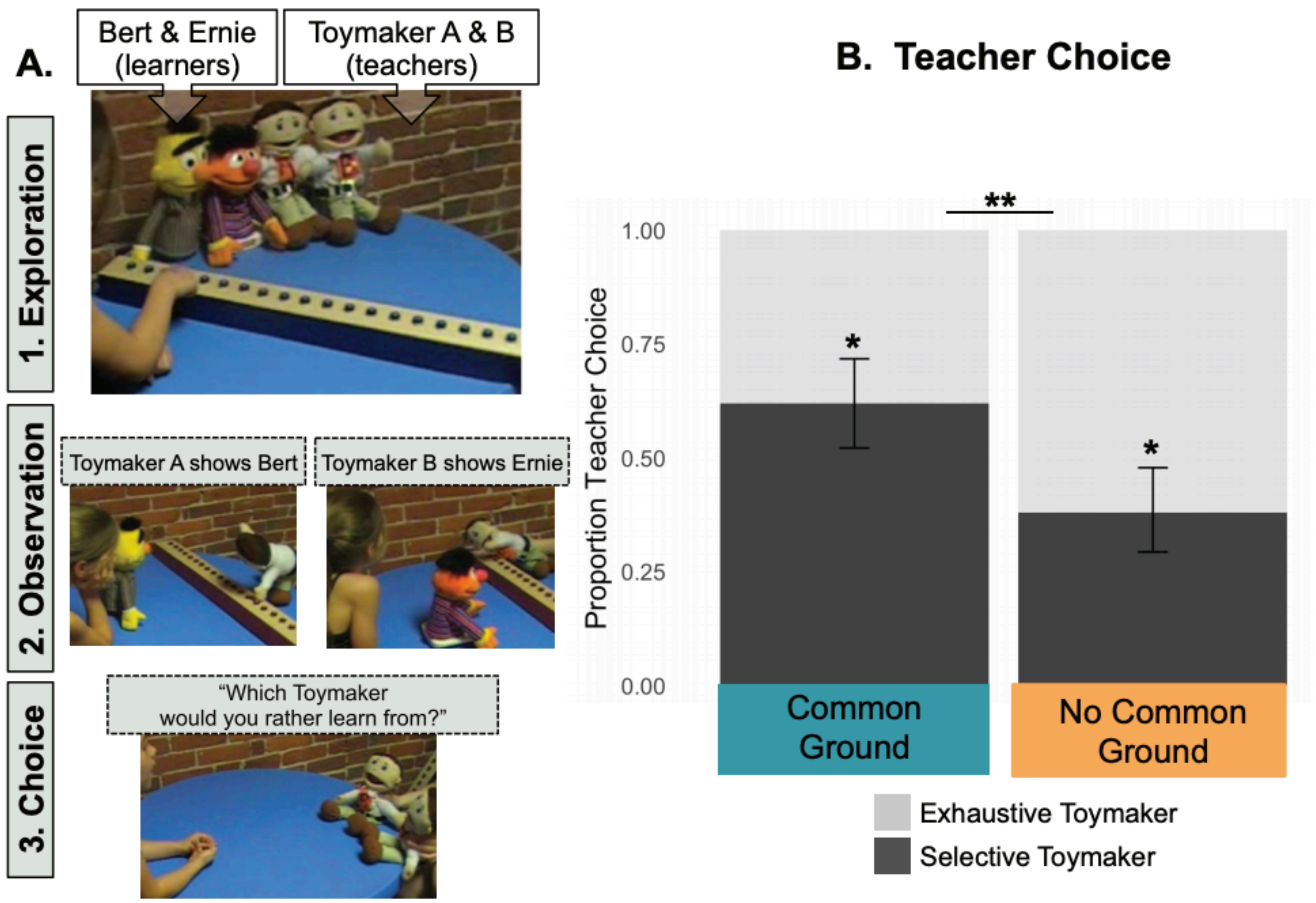
Lake et al. (2017)

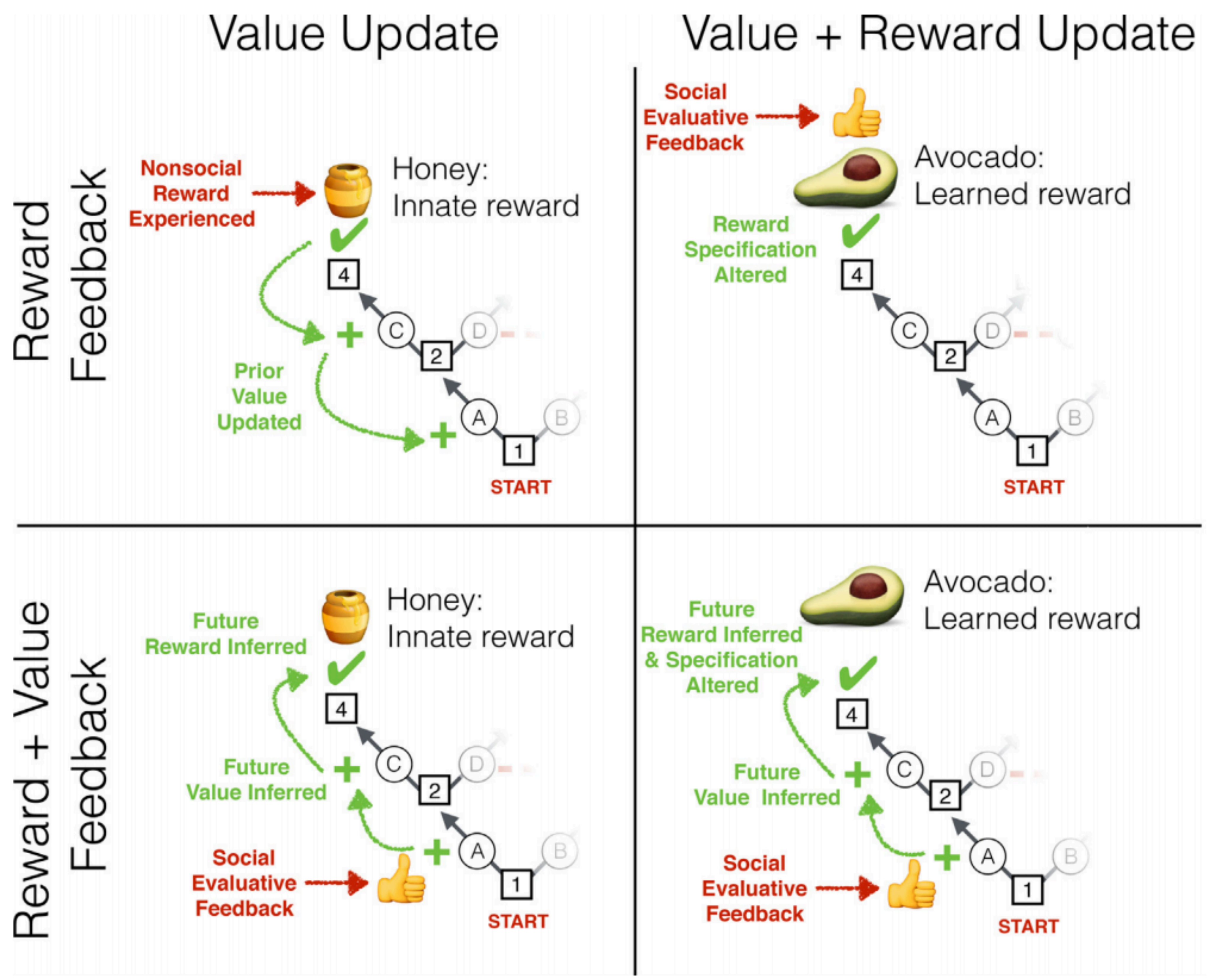
People should choose examples near the edges

Given hypothesis



What makes a good teacher?





Gweon et al. (2017)

Ho et al. (2017)

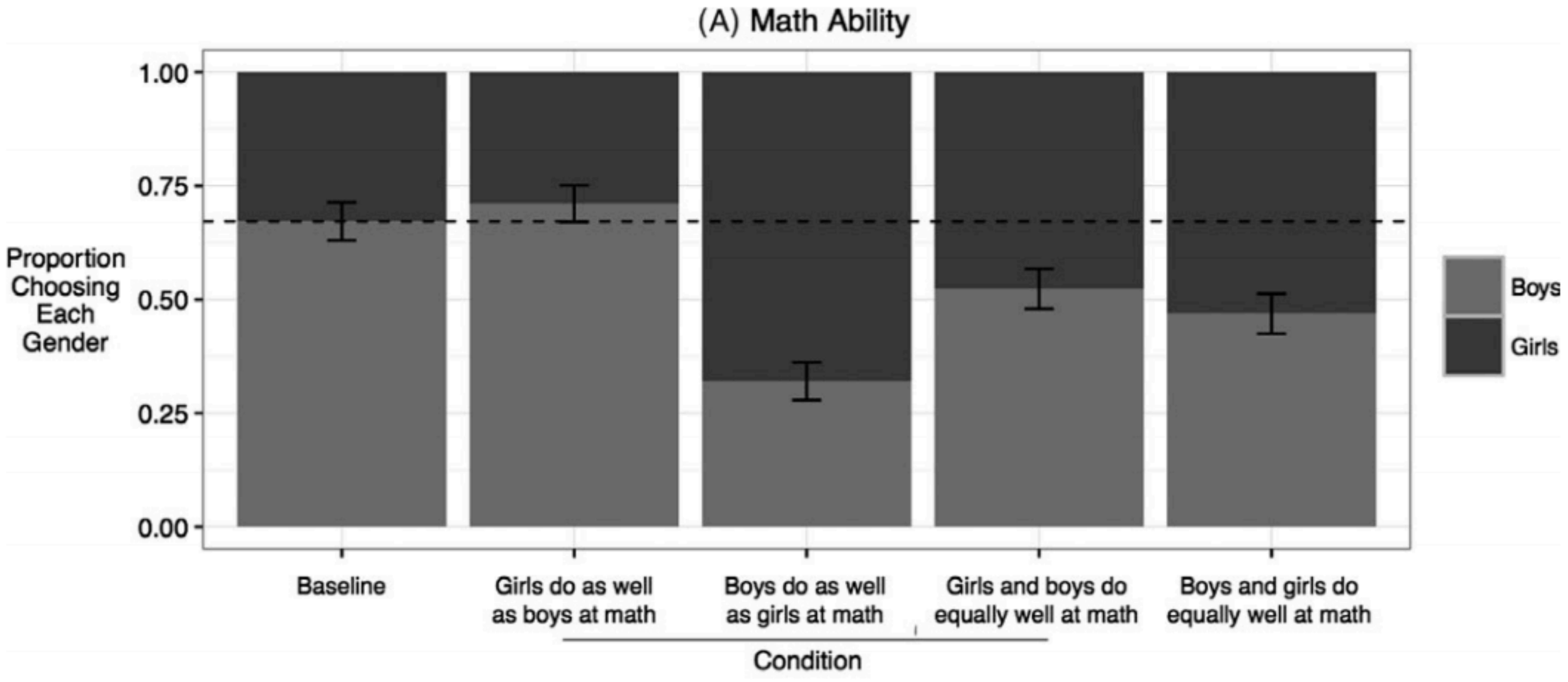
Pragmatic inference (Goodman & Frank, 2016)

Suppose you heard me say: **“My friend has glasses”**

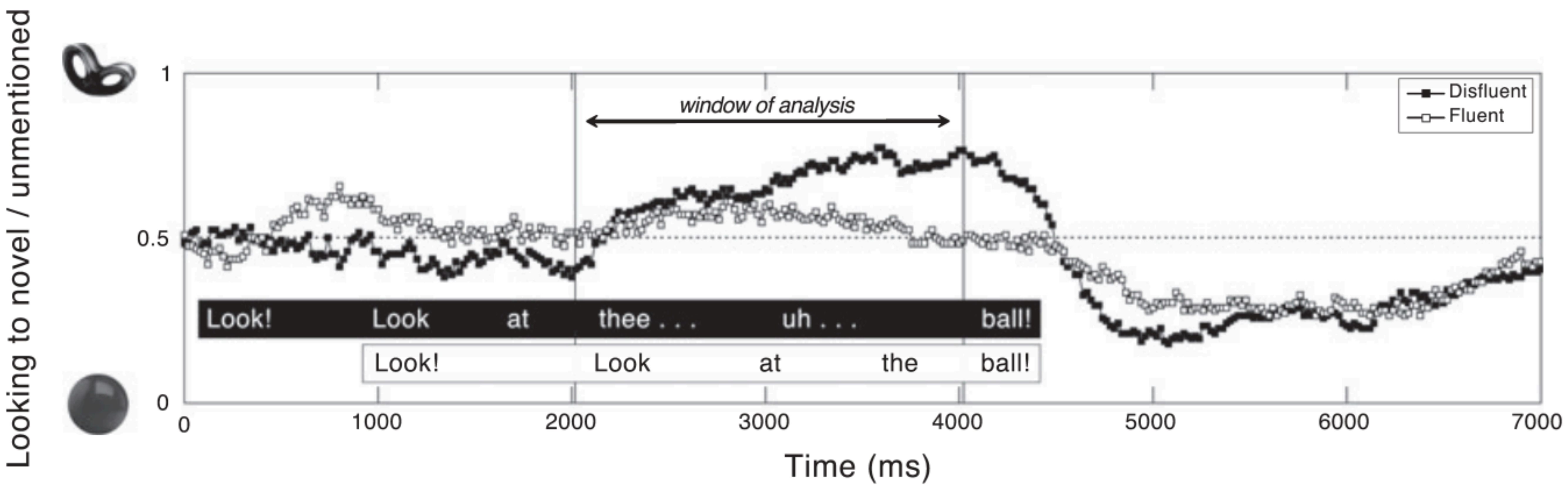
Which one of these people is my friend?

Indirectly learning from language

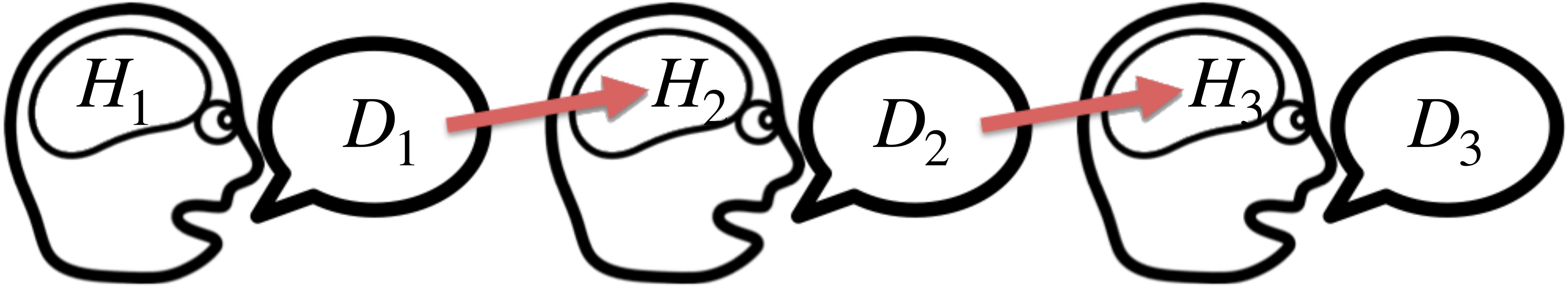
Chestnut & Markman (2018)



Kidd et al. (2011)

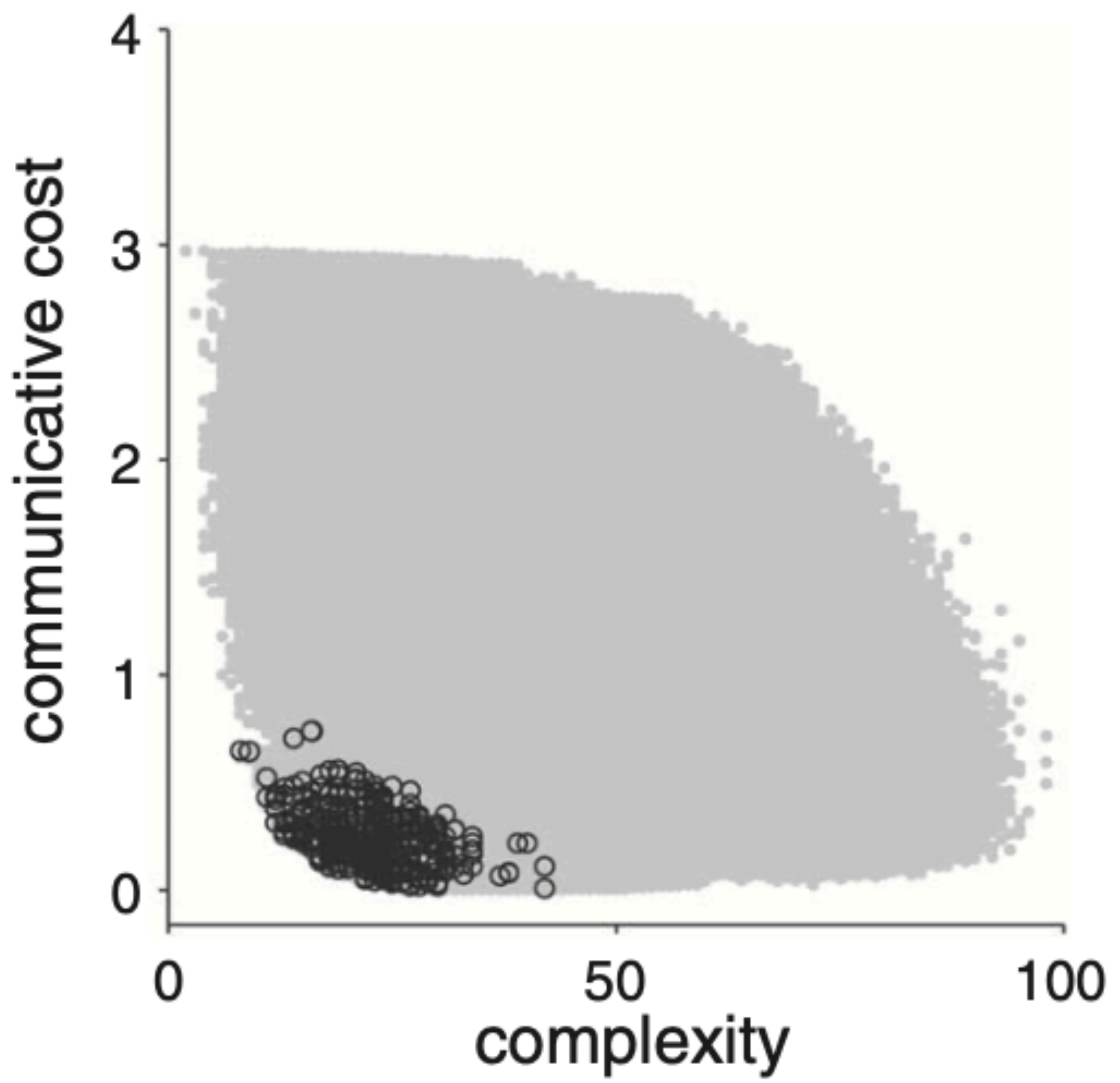


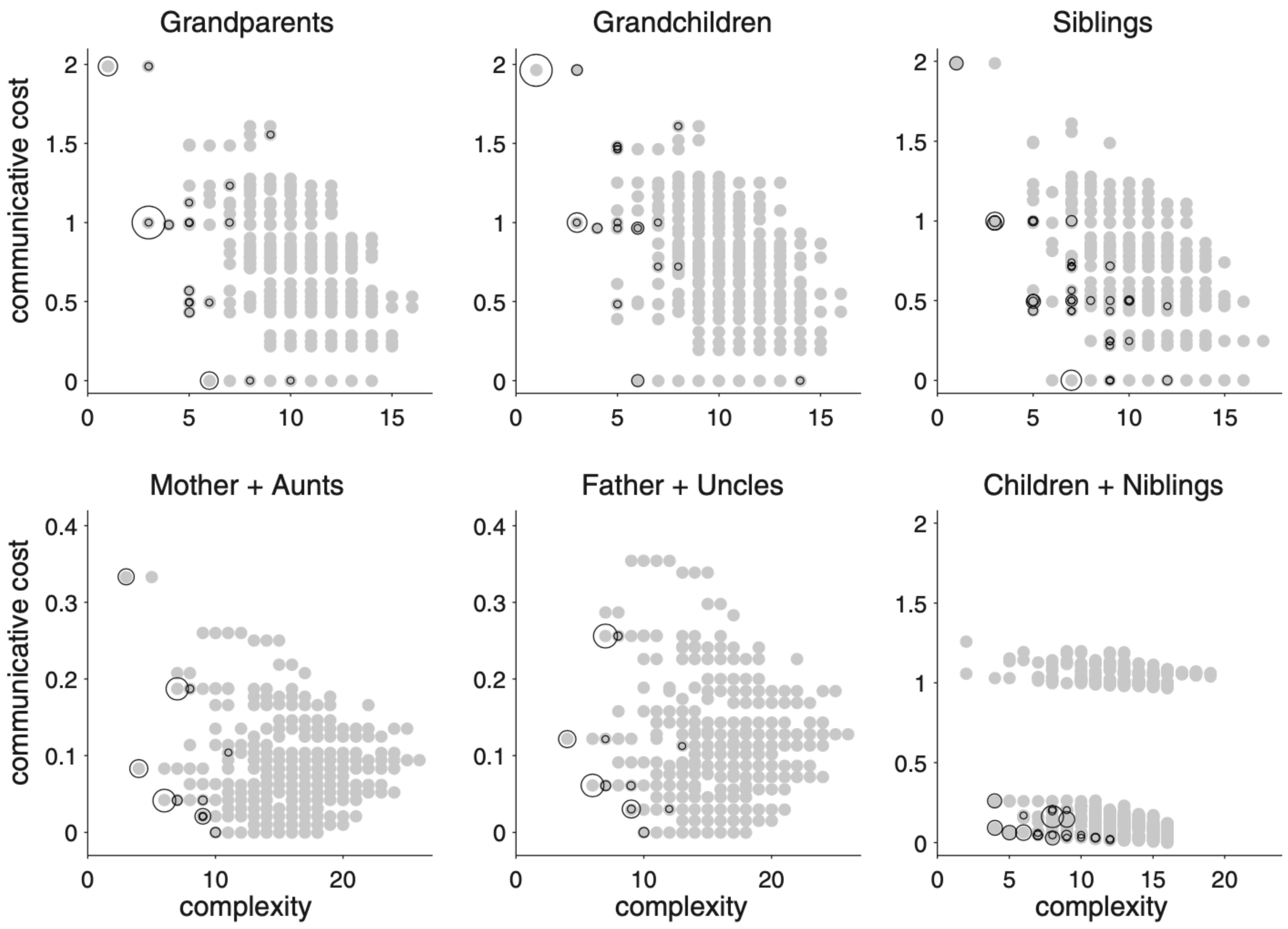
People learn the meaning of gavagai from other people



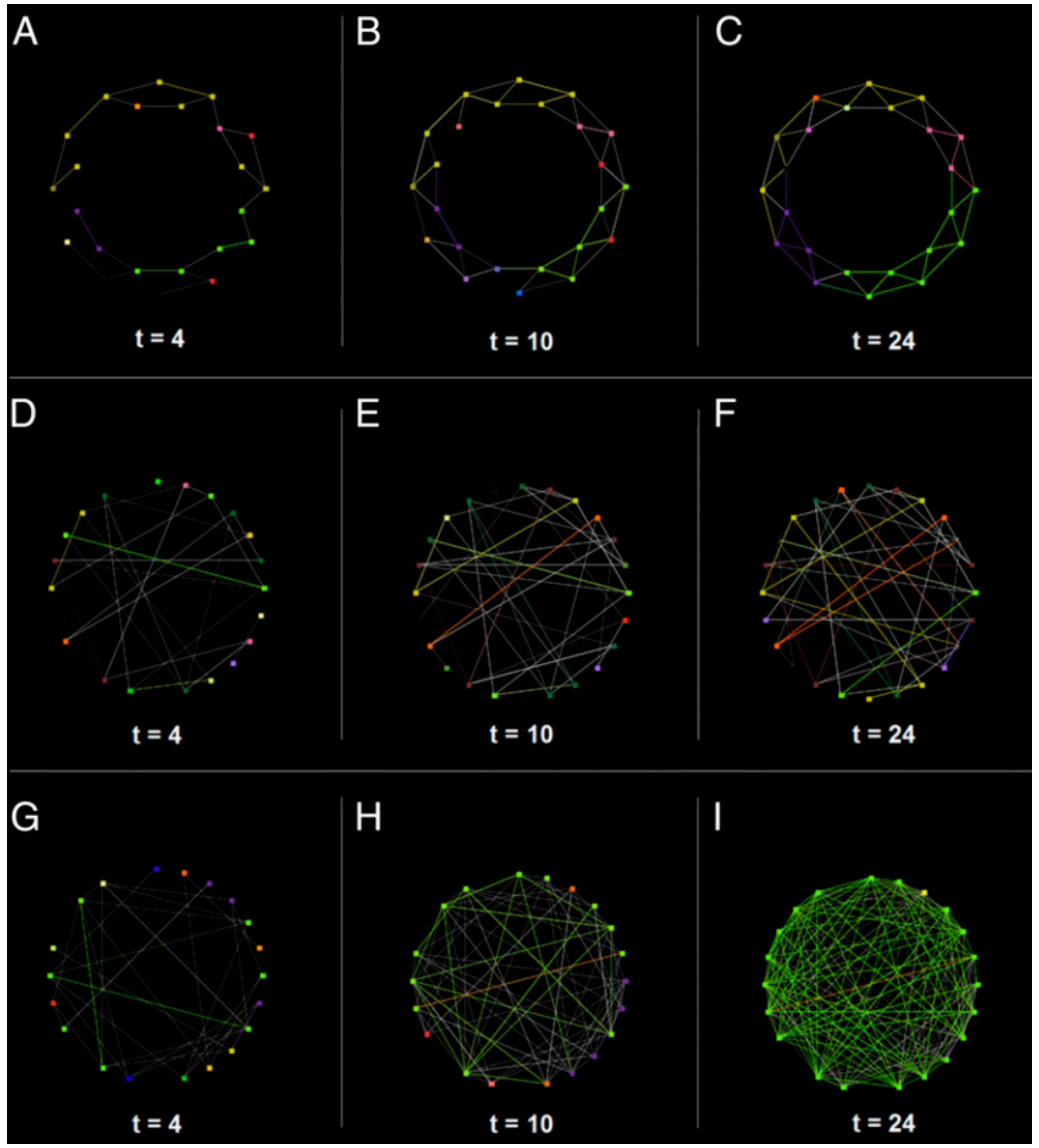
$$P\left(H_{i+1} \mid D_i\right) \propto P\left(D_i \mid H_i\right) P\left(H_i\right)$$

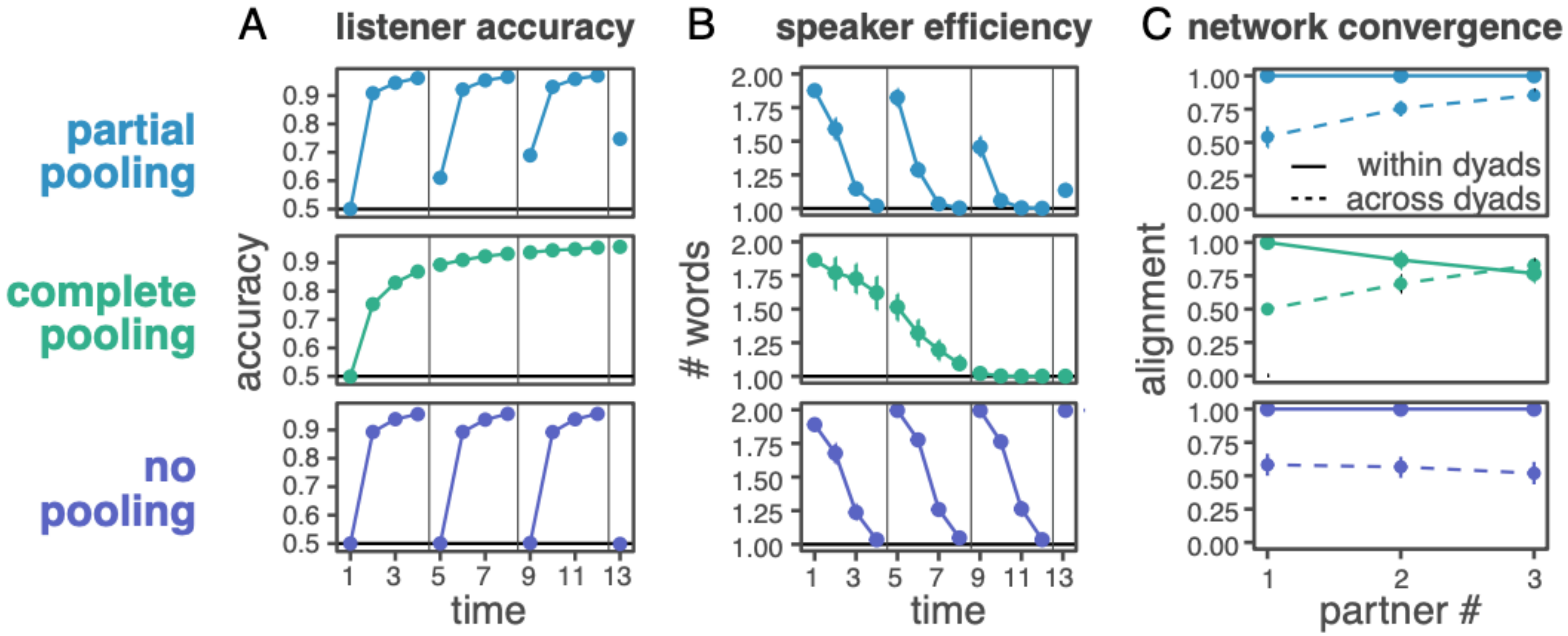
The world's kinship systems are near optimal!





Community effects on learning from others



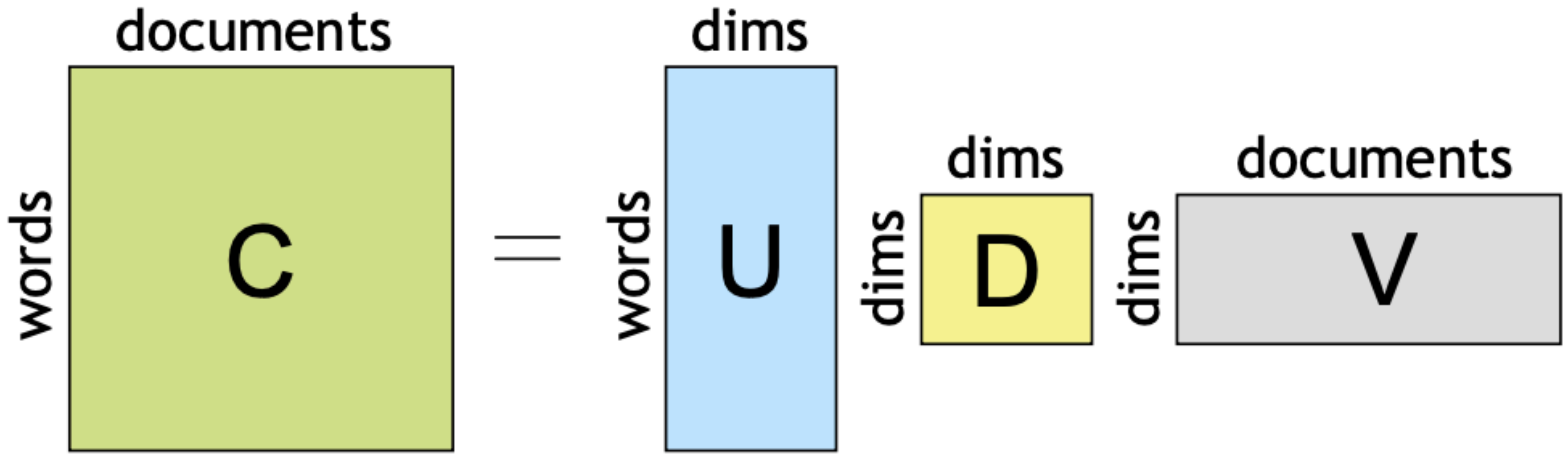


Centolla & Baroncelli (2015)

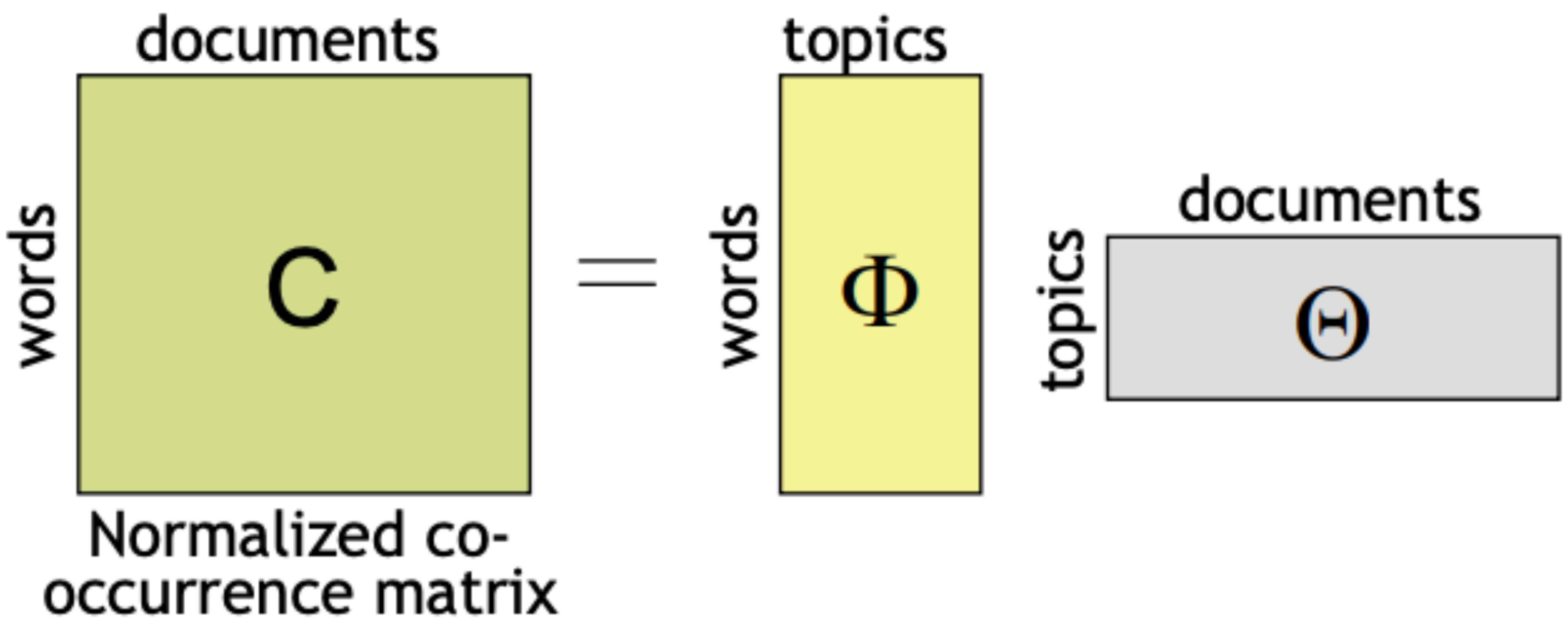
Hawkins et al. (2020)

Comparing LSA and Topic Models

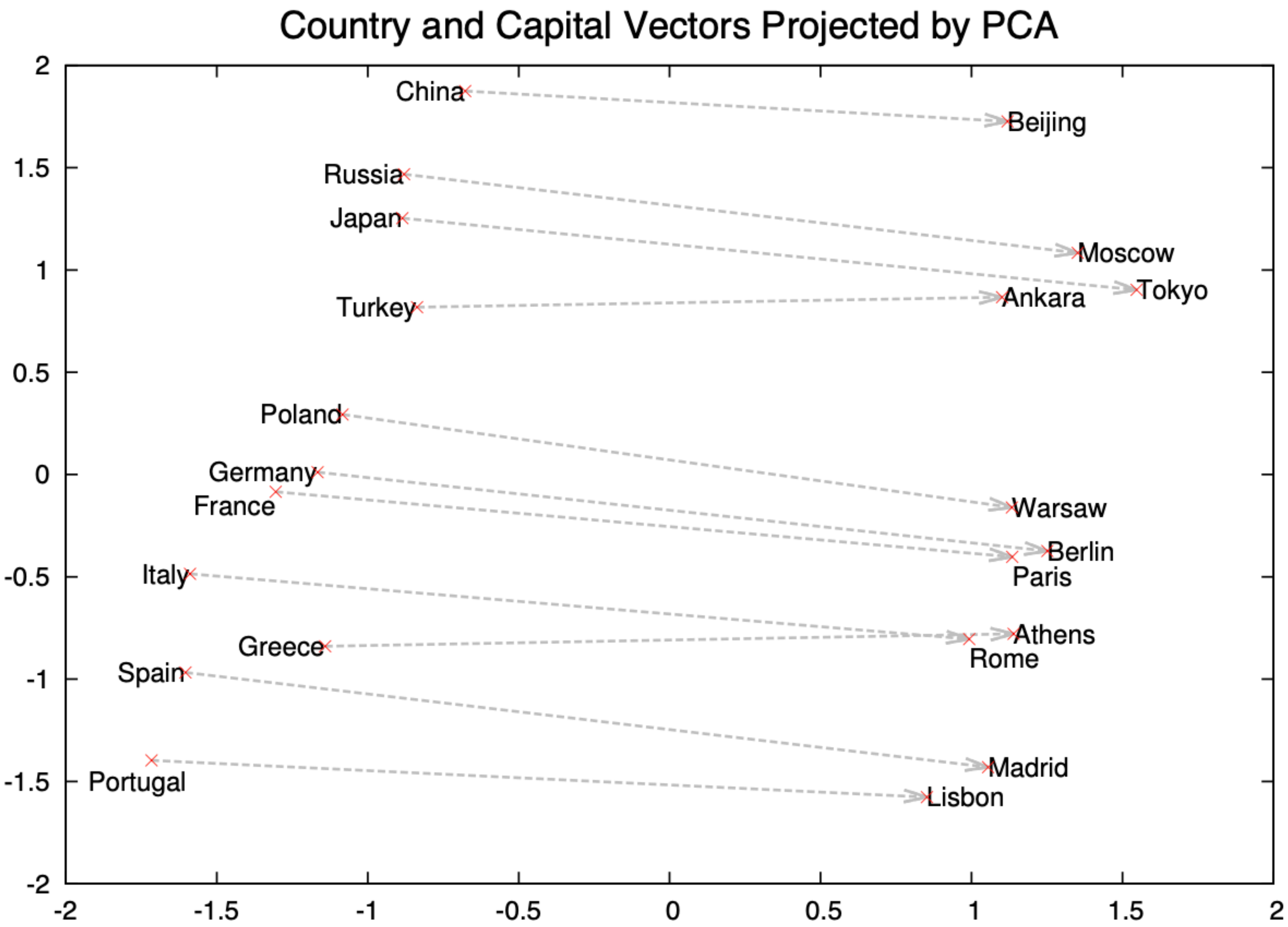
LSA



Topic Model

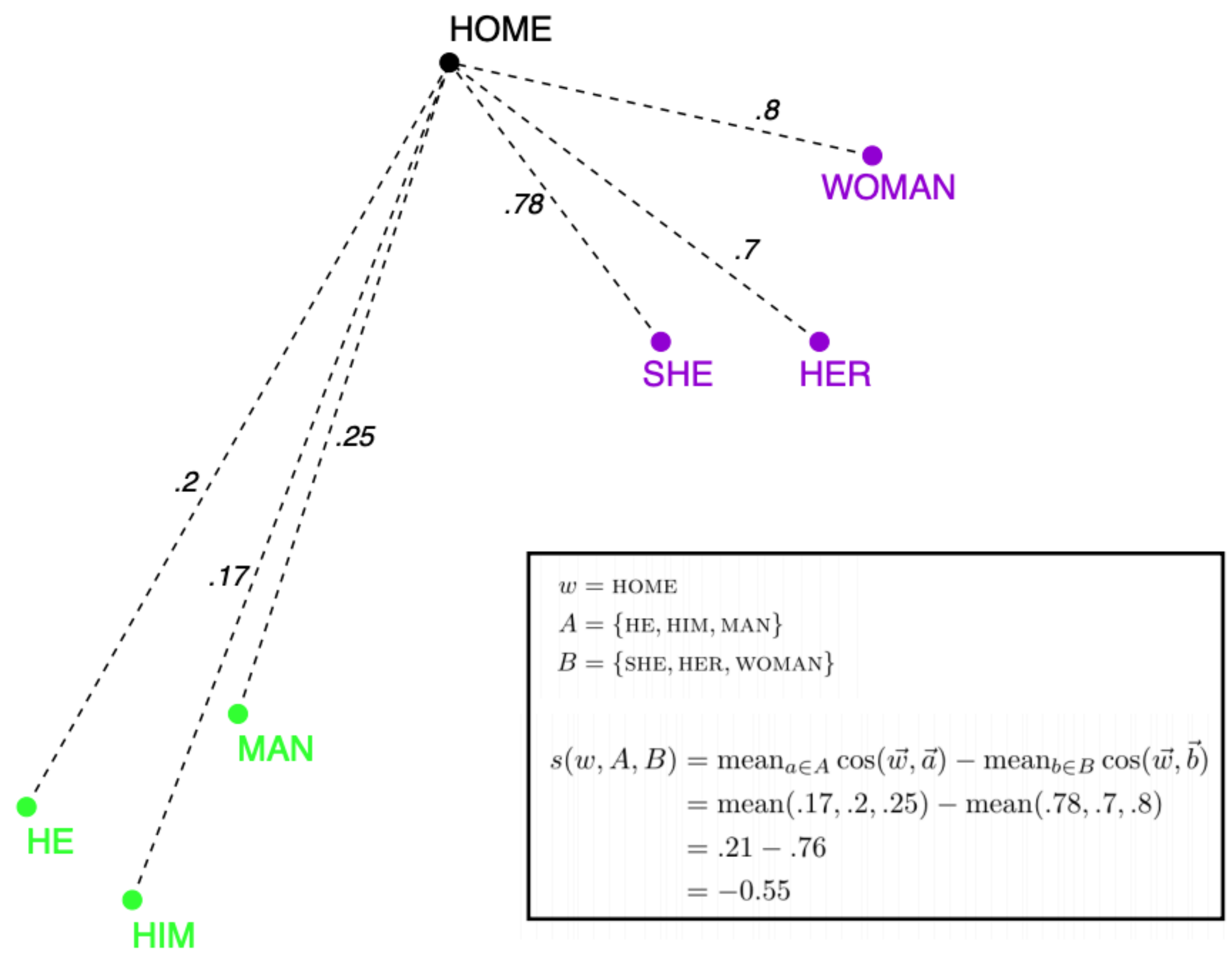


Word2Vec geometry is surprisingly meaningful!

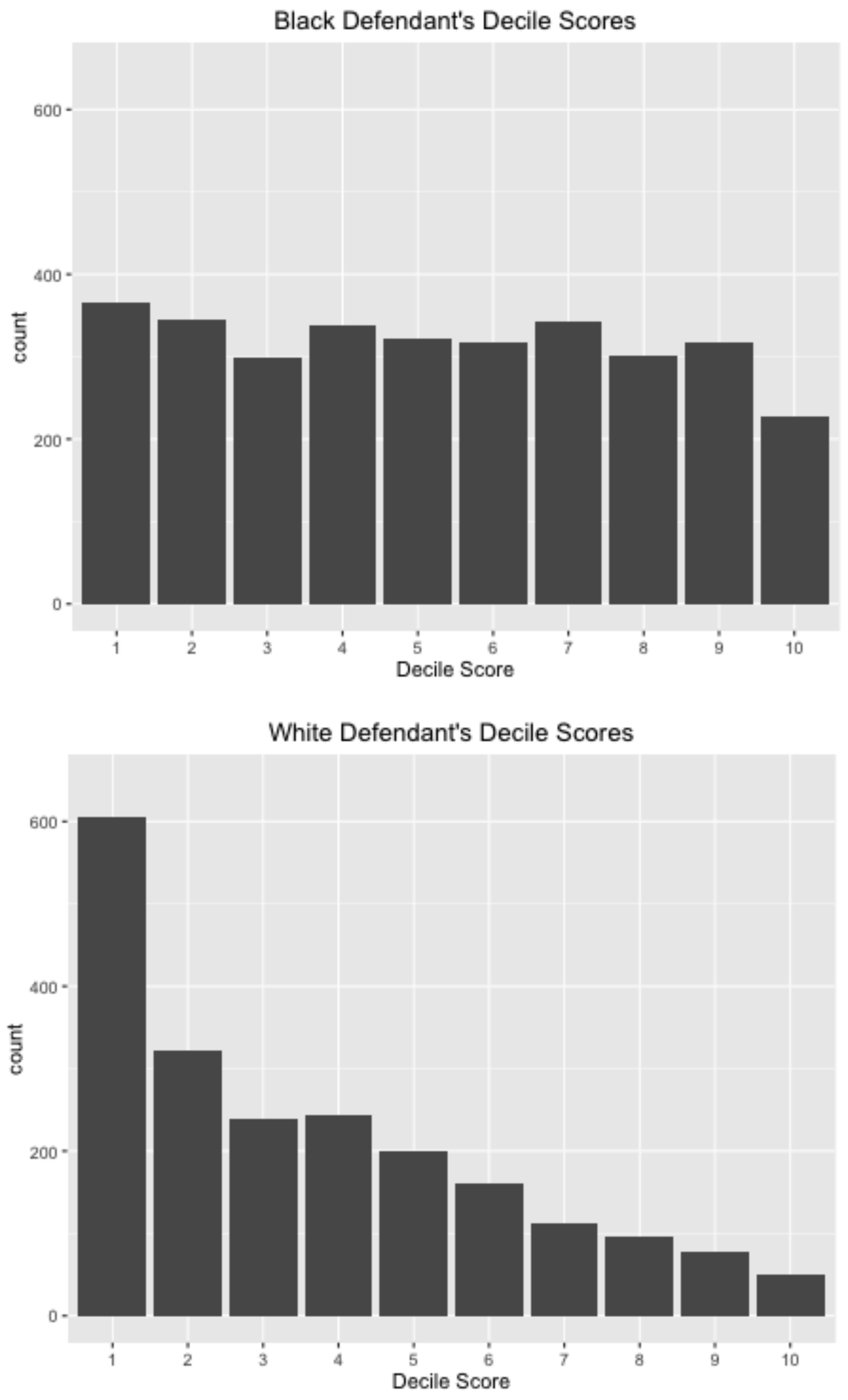


Mikolov et al. (2013)

Why training data matter



Caliskan & Lewis (under review)



Angwin et al. (2016)