(How) does the brain do Bayesian inference?

Sampling, search, and conditional probability in the mind

Kim Scott
Probcomp tutorial 11/1/2012

Marr's levels of analysis for Bayesian inference

Computation

$$p(h \mid X) = \frac{p(X \mid h)p(h)}{p(X)}$$

$$= \frac{p(X \mid h)p(h)}{\sum_{h' \in H} p(X \mid h')p(h')}.$$

Algorithm

- **Markov chain?**
- **Monte Carlo?**

Implementation

- a case for biological plausibility
- inspiration and encouragement for hardware
- Boesing et al.

Today: A review of literature relevant to the algorithmic level, & discussion of potential directions.

Hypotheses: from conscious states to percepts

I appeal to anyone's experience whether upon sight of an OBJECT he computes its distance by the bigness of the ANGLE made by the meeting of the two OPTIC AXES? […] In vain shall all the MATHEMATICIANS in the world tell me, that I perceive certain LINES and ANGLES which introduce into my mind the various IDEAS of DISTANCE, so long as I myself am conscious of no such thing.

(Berkeley, 1709, "An essay towards a new theory of vision")

In the ordinary acts of vision this knowledge of optics is lacking. Still it may be permissible to speak of the psychic acts of ordinary perception as *unconscious conclusions,* thereby making a distinction of some sort between them and the common so-called conscious conclusions. And while it is true that there has been […] a measure of doubt as to the similarity of the psychic activity in the two cases, there can be no doubt as to the similarity between the results […]

(Helmholtz, 1924, Treatise on Physiological Optics)

MC(?) MC(?) in the mind overview

1. Brief motivation
2. Examples of people “doing Bayesian inference”
3. Evidence for computational framing
4. MCMC for Bayes net demo
5. Evidence for sampling
6. Evidence for Markov chains

Why movement through a hypothesis space?

“Yet I say again that learning must be nondemonstrative inference; there is nothing else for it to be. And the only model of a nondemonstrative inference that has ever been proposed anywhere by anyone is hypothesis formation and confirmation.”

(Fodor, “Fixation of Belief and Concept Acquisition”)

1. We really don’t have anything else
2. Subjective familiarity of the analogy for explicit problem-solving
3. “One state at a time”

Why care about algorithms?

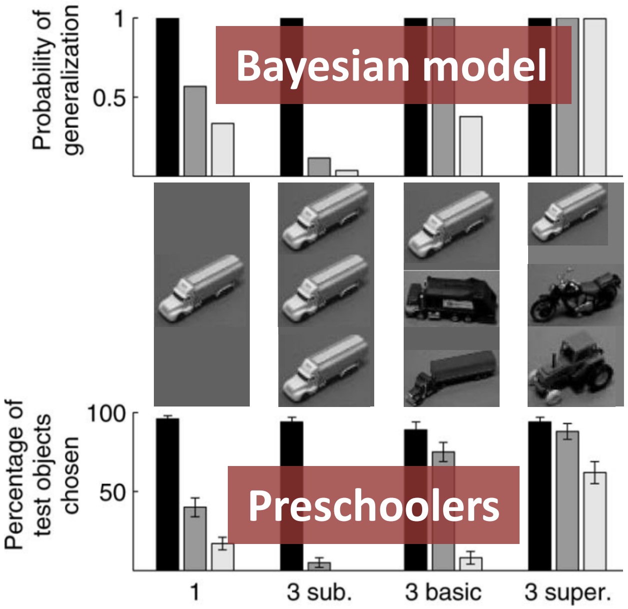
[In] most distributional learning procedures there are vast numbers of properties that a learner could record, and since the child is looking for correlations among these properties, he or she faces a combinatorial explosion of possibilities. […] To be sure, the inappropriate properties will correlate with no others and hence will eventually be ignored […], but only after astronomical amounts of memory space, computation, or both.

(Pinker, Language Learnability and Language Development)

In addition to standard curiosity…

1. Getting from behavioral data to **representation** of hypotheses and what is **actually being learned** requires assumptions about algorithms.
2. As inspiration for engineering systems for inference
3. To find out whether Bayesian inference is actually applied to varied problems in the same way

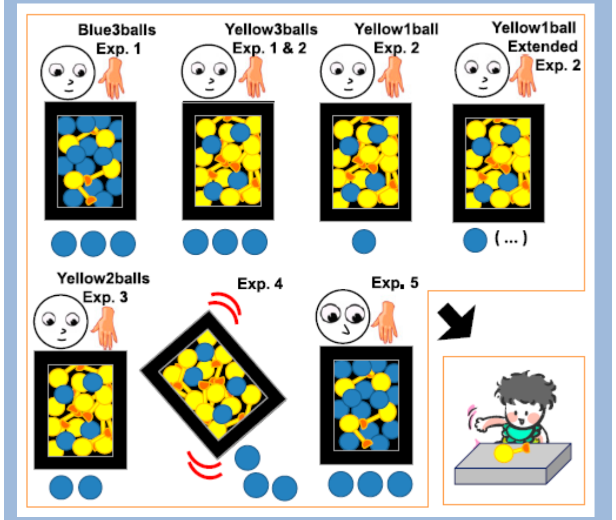
Word learning



Preschoolers constrain generalization of a new label when more examples are given

Xu & Tenenbaum 2007

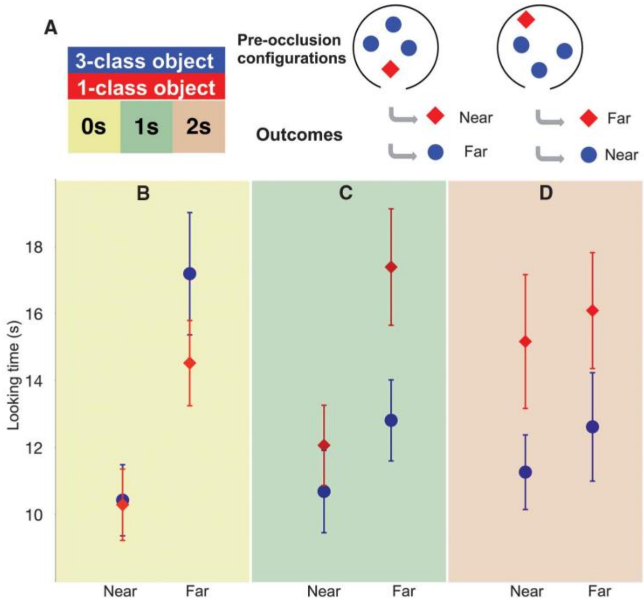
Property generalization



Toddlers use both the sample and sampling process to generalize properties

Gweon, Tenenbaum, & Schulz 2010

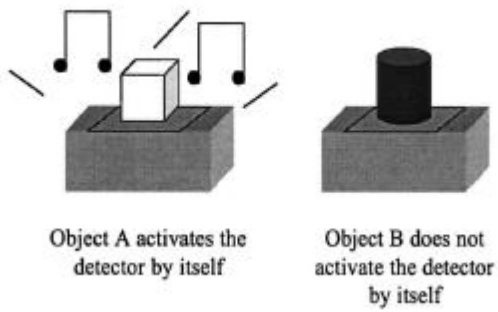
Physical events



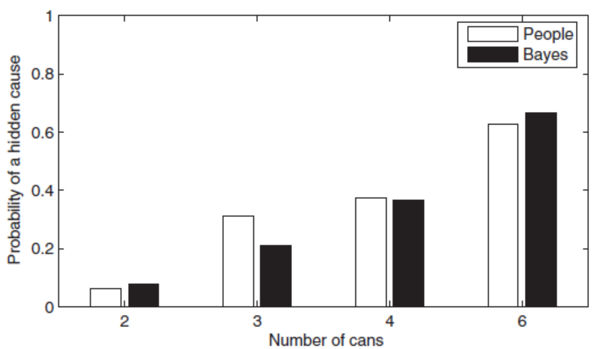
Graded infant looking times show effects of both frequency and arrangement, dependent on time

Teglas et al 2011

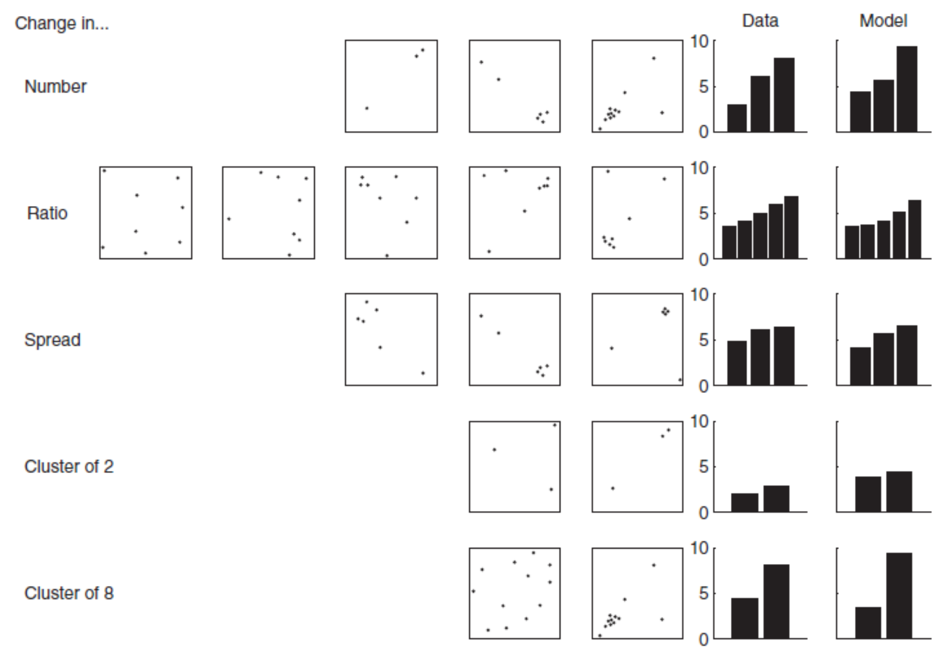
Causal inference



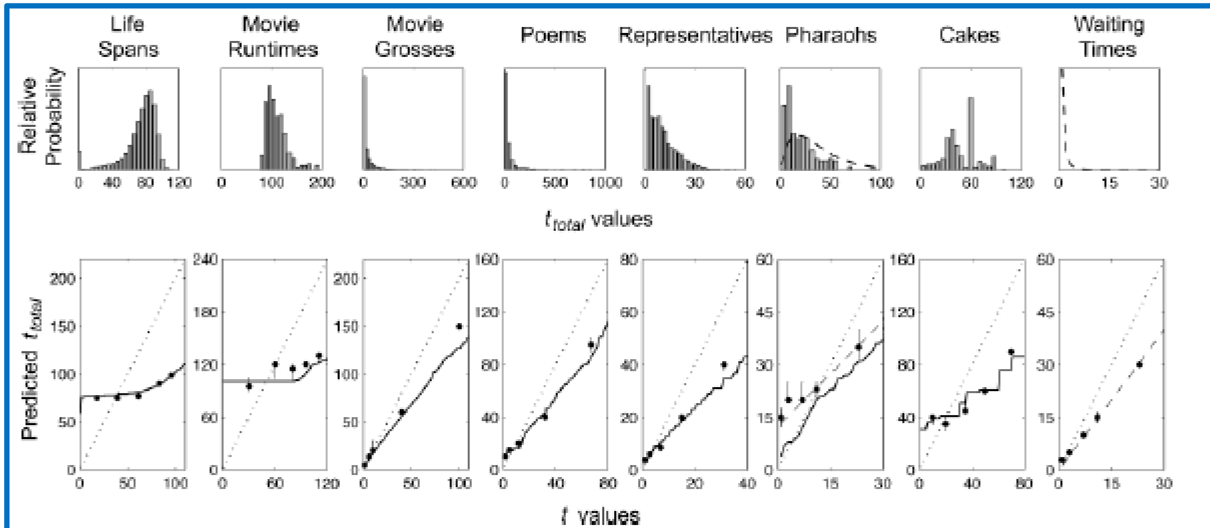
Gopnik et al 2004



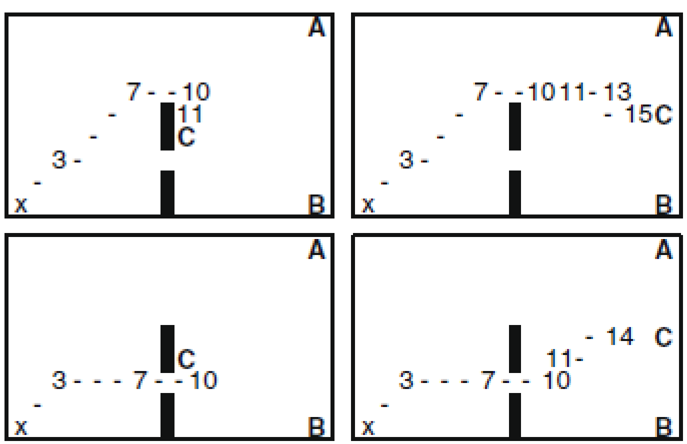
Griffiths et al 2004



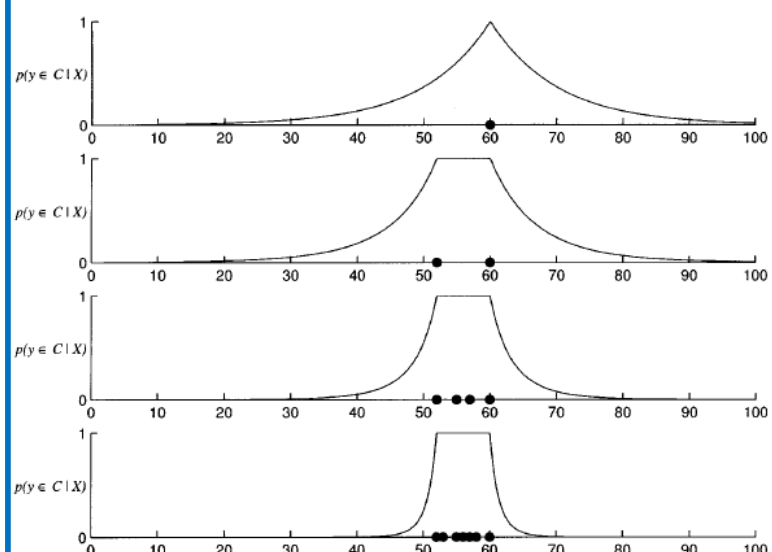
Griffiths & Tenenbaum 2007



Griffiths & Tenenbaum 2006

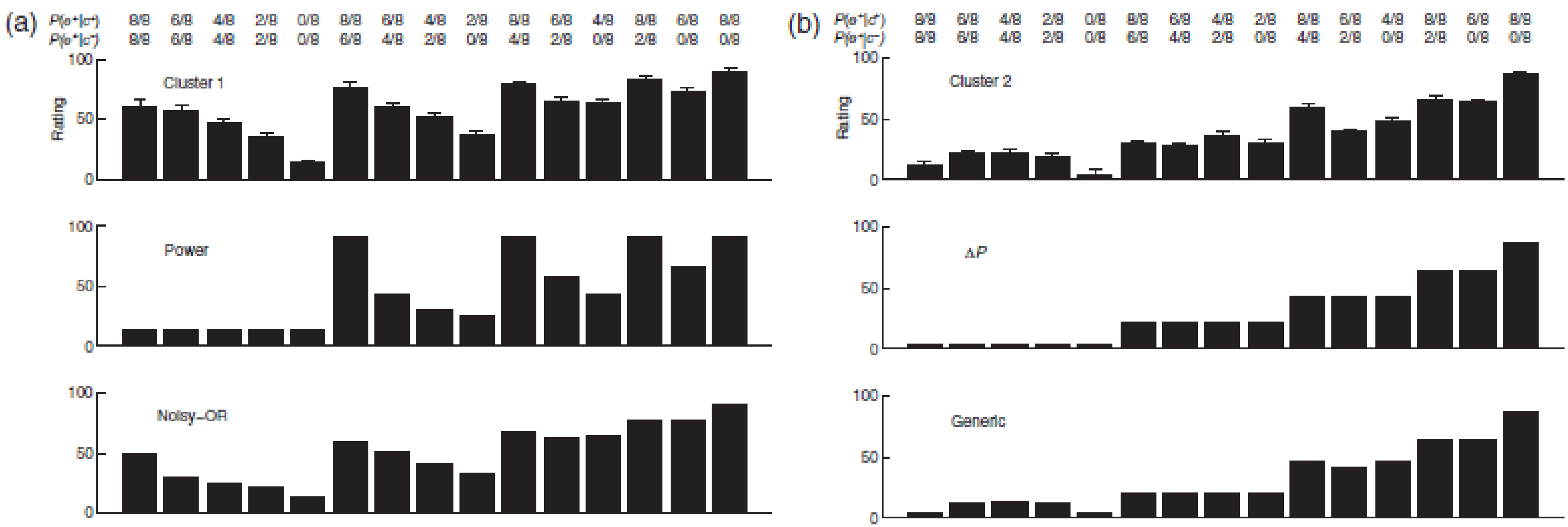


Baker, Saxe, & Tenenbaum 2009



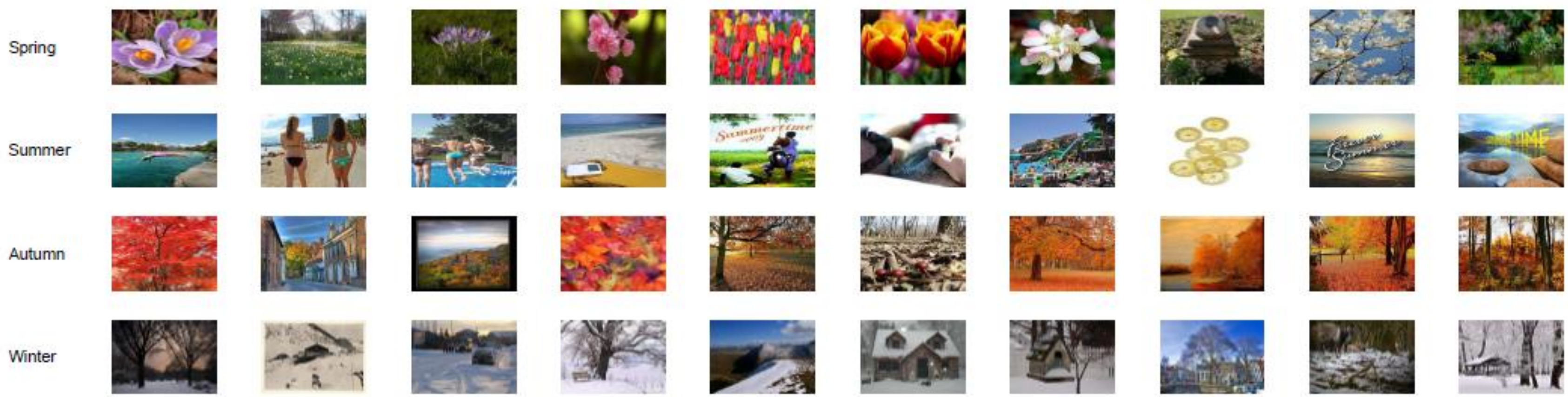
Tenenbaum & Griffiths 2001

Computational-level evidence: psychological reality of priors



Griffiths & Tenenbaum 2009

Computational-level evidence: MCMC with people



Idea: use people's 2AFC category-membership choices as acceptance function for Markov chain so it converges to P(x|c)

$$P_{\text{choice}}(x';x|c) = \frac{p(x'|c)}{p(x'|c) + p(x|c)}$$

Computational-level evidence

- Priming affects spontaneously generated explanations, but not evaluation of given hypotheses
  – Bonawitz & Griffiths 2008: “Deconfounding hypothesis generation and evaluation in Bayesian models”

- Reading time ~ log probability of word (Smith & Levy 2008)

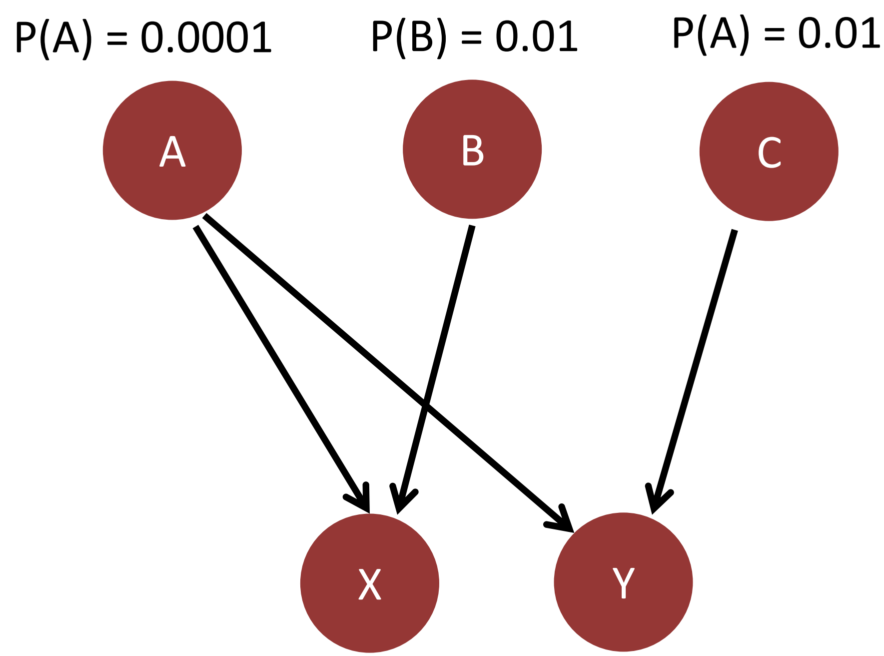
Algorithmic level: plausibility of MCMC

- Alternatives?
  - Importance sampling
  - Magic to represent hypothesis space exponential in parameters in parallel… phase relative to a vector of frequencies?
- To model exact Bayesian inference (computing the posterior distribution), we have to make approximations, e.g. MCMC methods.
  - …maybe the system we're modeling does exactly the same thing.
  - Unfounded, but maybe still true.
  - And that would be great news about samplers!
- If we buy into this framework enough to consider specific algorithms, we want to be able to identify…
  - What is the hypothesis space?
  - How do we move from one state to another?
  - What does a percept or judgment correspond to; how many samples does it use?

1. Demo
2. Monte Carlo: Evidence for sampling
3. Markov chain: Evidence for movement through a hypothesis space

Demo: Diagnosis net

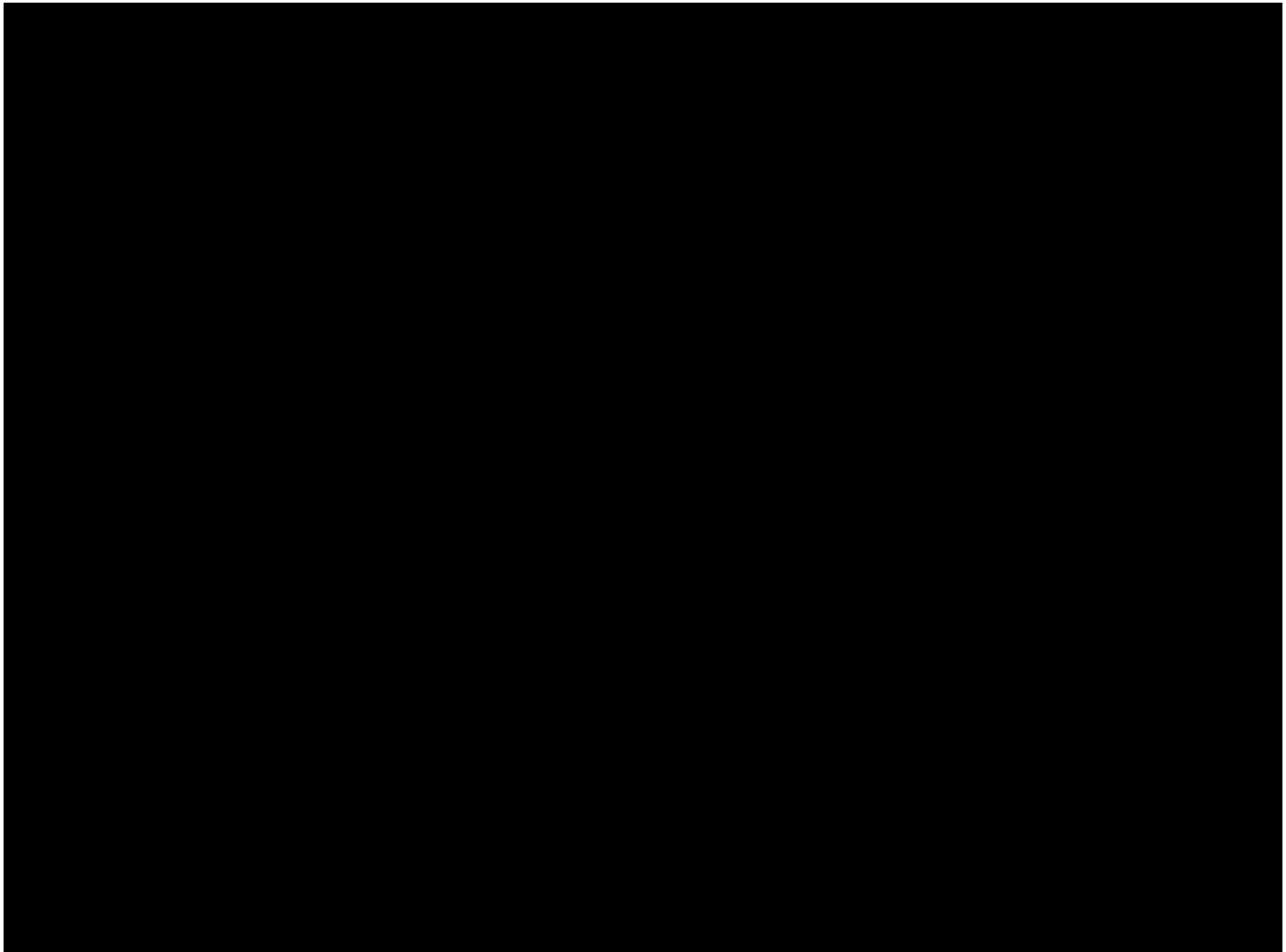
- Gibbs sampler for "medical diagnosis" Bayes net
- Binary nodes, single layer
- Observes effects, uses (correct!) structure of net to wander towards posterior distribution



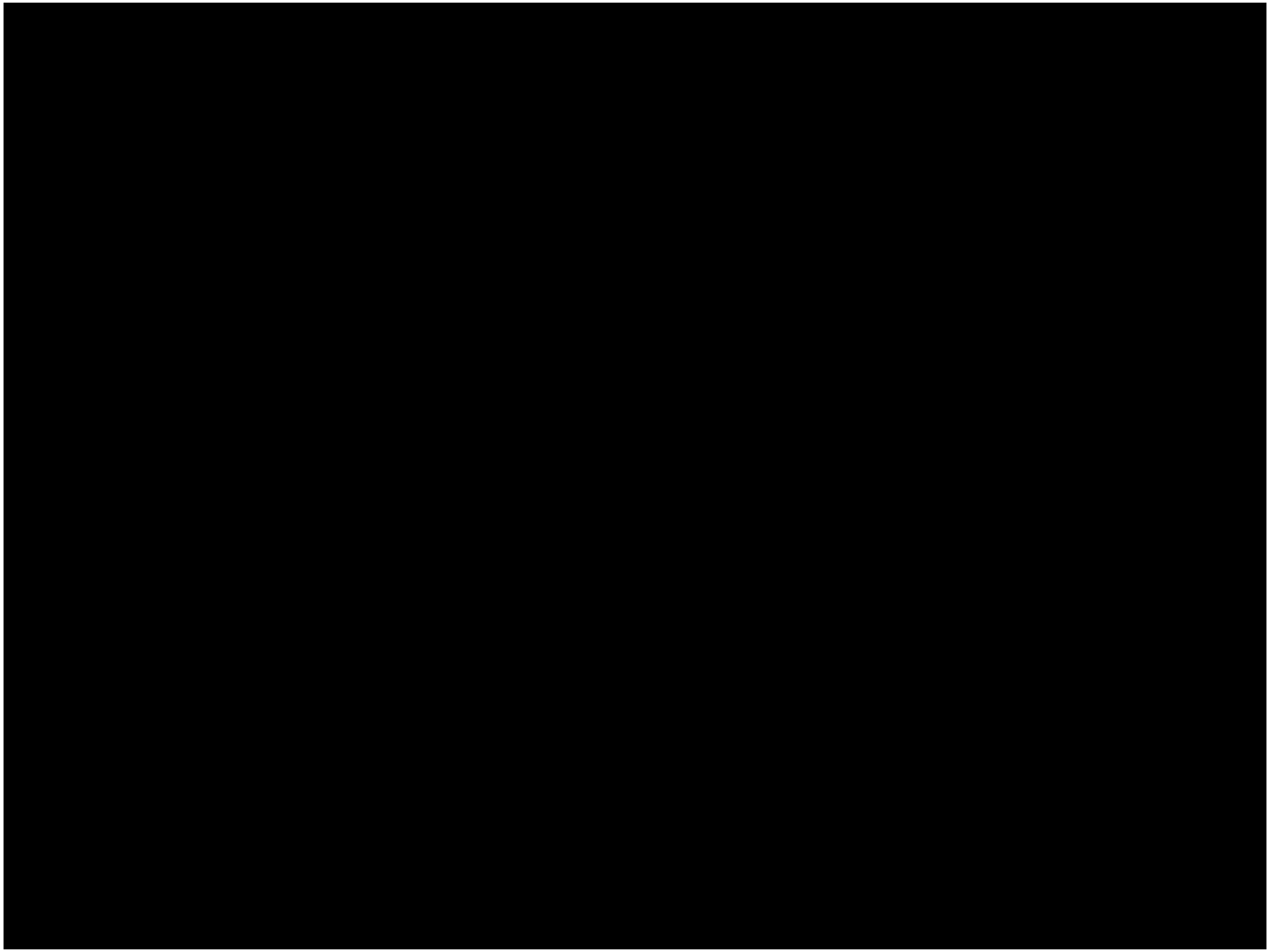
	~A	A
~B	.001	0.99
B	0.99	0.995

	~A	A
~C	.001	0.99
C	0.99	0.995

Diagnosis net example



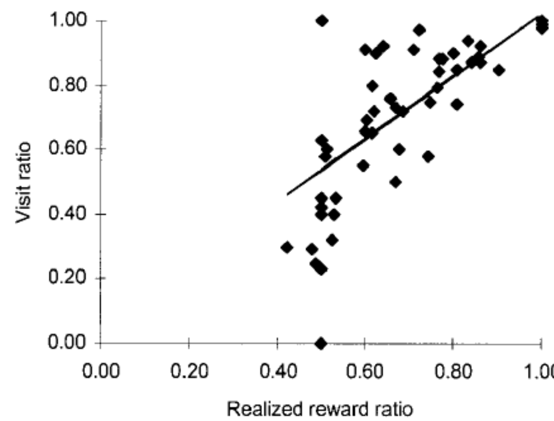
Diagnosis net: simple “causal” net

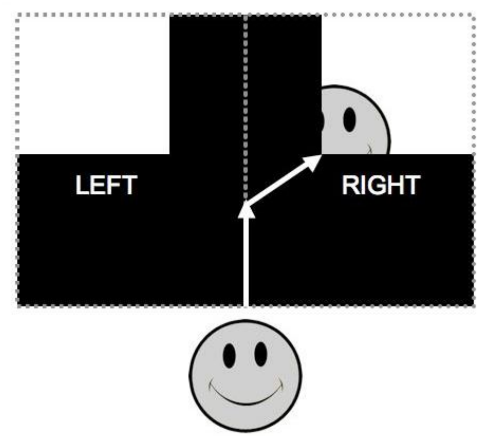


15 causes, 50 effects, ~4 causes/effect.   P(effect|no cause) = 0.1, P(cause) = 0.01

Sampling in human cognition

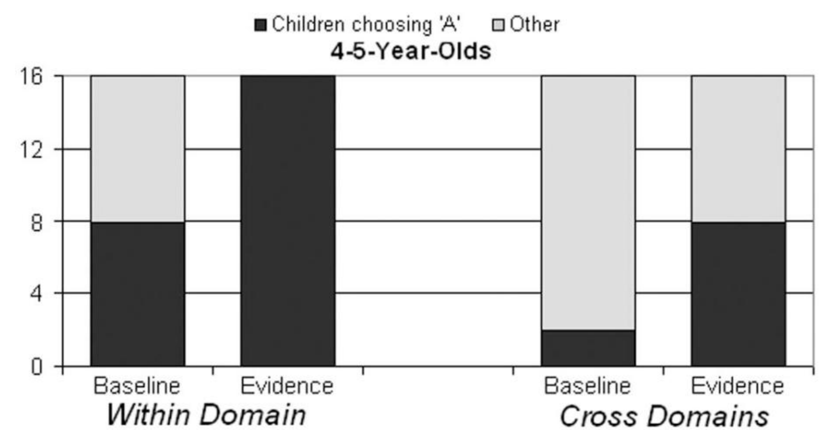
- Interpretations:
  - Explicit responses *are* individual samples
  - Monte Carlo: approximate a distribution by a finite number of samples
- Probability matching
  - Phylogenetically old foraging behavior: Bees in two-armed bandits (Keaser et al 2002)
  - Adults often probability-match rather than maximizing (Gardner 1957); children tend to maximize more (e.g. Hudson Kam & Newport 2009, in language learning)
  - But even ten-month-olds are capable of probability matching (Davis, Newport, & Aslin 2009)
  - Evidence of sampling or separate faculty?





Population responses as samples

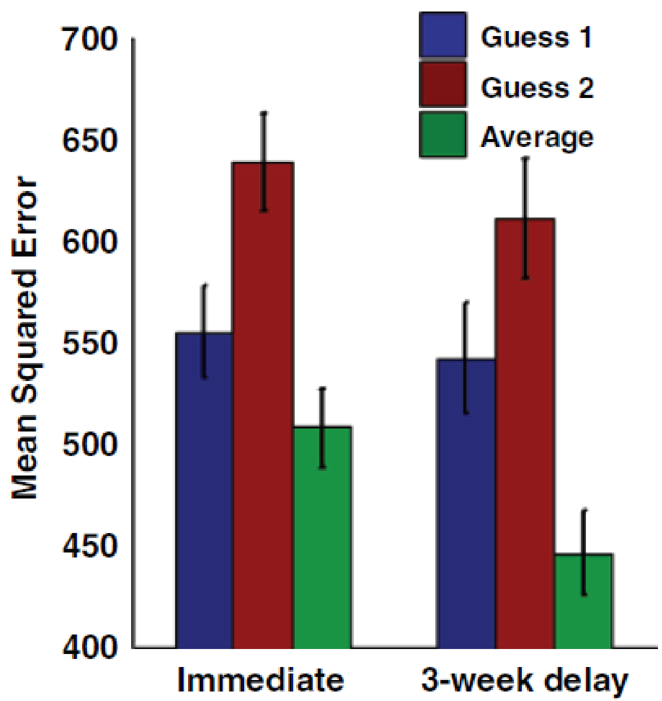
- Sampling hypothesis: variation in judgments reflects the true distribution
- Population level: graded fractions of correct responses as indirect evidence



Schulz, Bonawitz, & Griffiths 2007

Within-subject responses as samples

"What percentage of the world's airports are in the United States?"



Vul & Pashler 2008: "the crowd within"
Analogous results for visual attention (Vul, Hanus, & Kanwisher 2010)

Denison et al 2009: "Preschoolers sample from probability distributions"

Responses	Expectation	*Long Wait*	*Short Wait*
red,red,red	.512	10	1
red,red,blue	.128	1	1
red,blue,red	.128	2	10
red,blue,blue	.032	3	0
blue,red,red	.128	0	1
blue,red,blue	.032	1	6
blue,blue,red	.032	1	1
blue,blue,blue	.008	2	0

Bonawitz et al. "Rational randomness"

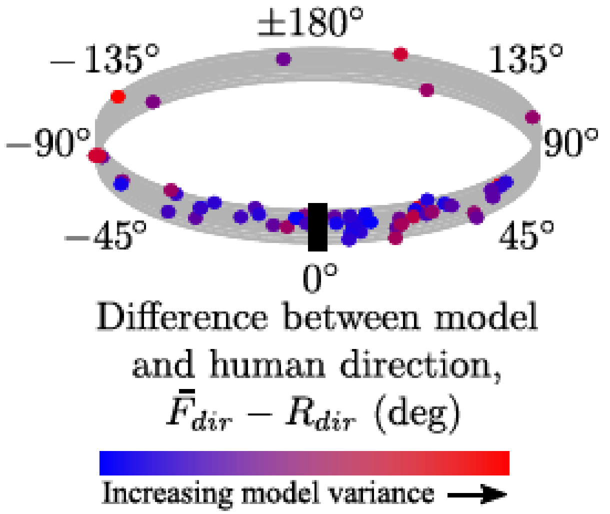
- Follow-up experiments showed children were not just doing probability matching to chip frequencies
- Correlation between hypotheses consistent with win-stay lose-shift mechanism but not independent sampling

Sampling in intuitive physics?

What would sampling (more uniquely) predict?

- Dropoff in accuracy with limited resources, consistent with discrete jumps from n to n-1 samples
- Rare outcomes should (rarely) skew or (usually) not affect estimates
- Precision of posthoc judgment of a conditional probability should depend on conditional probability
- Potential improved precision over time if objects pulled toward some location, in contrast with simple propagation of uncertainty

B. Trajectory, Direction: model v. human

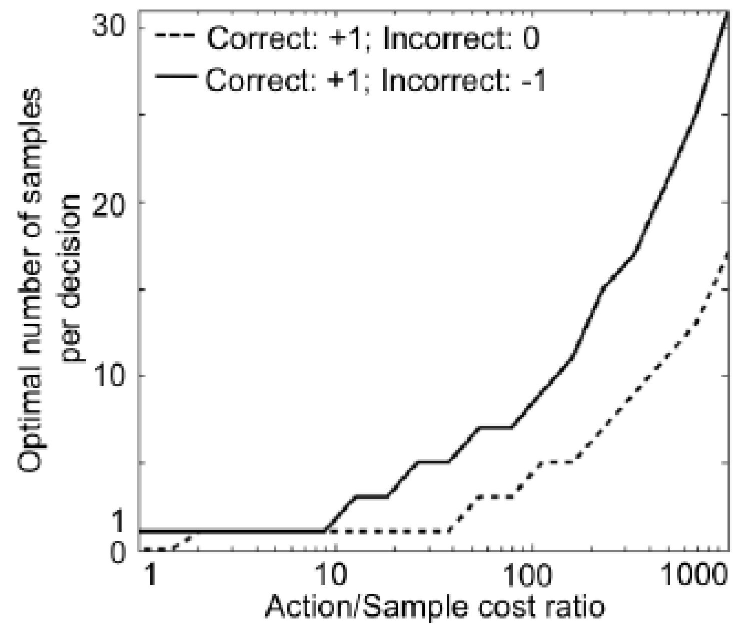


Hamrick, Battaglia, & Tenenbaum 2011

Monte Carlo estimates: a caveat

- Often just a few samples is plenty for practical purposes
- Adding any cost to sampling can even make getting just one rational
- So how can we situate ourselves to grab a *good* “just one”?

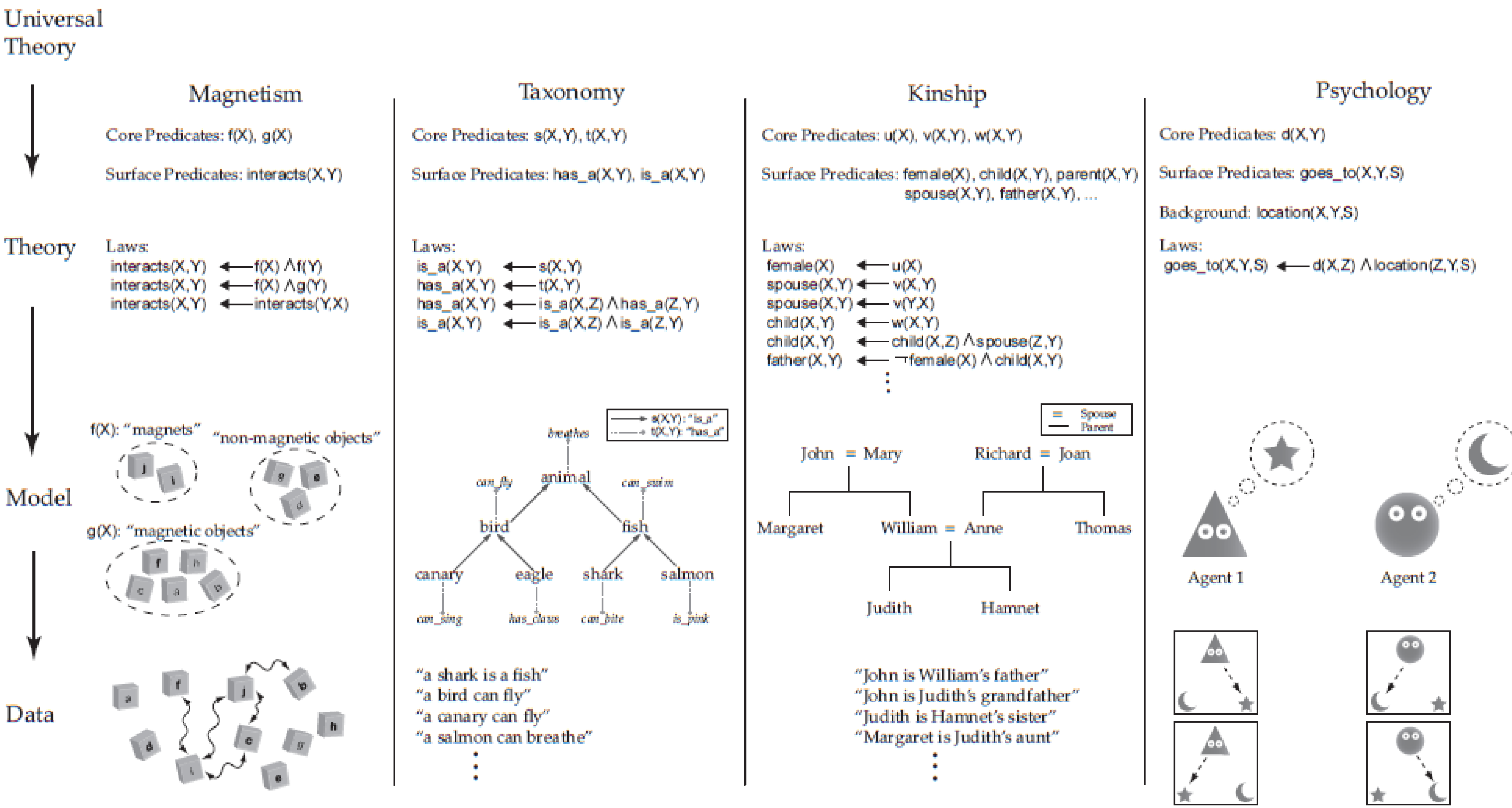
“One and Done”



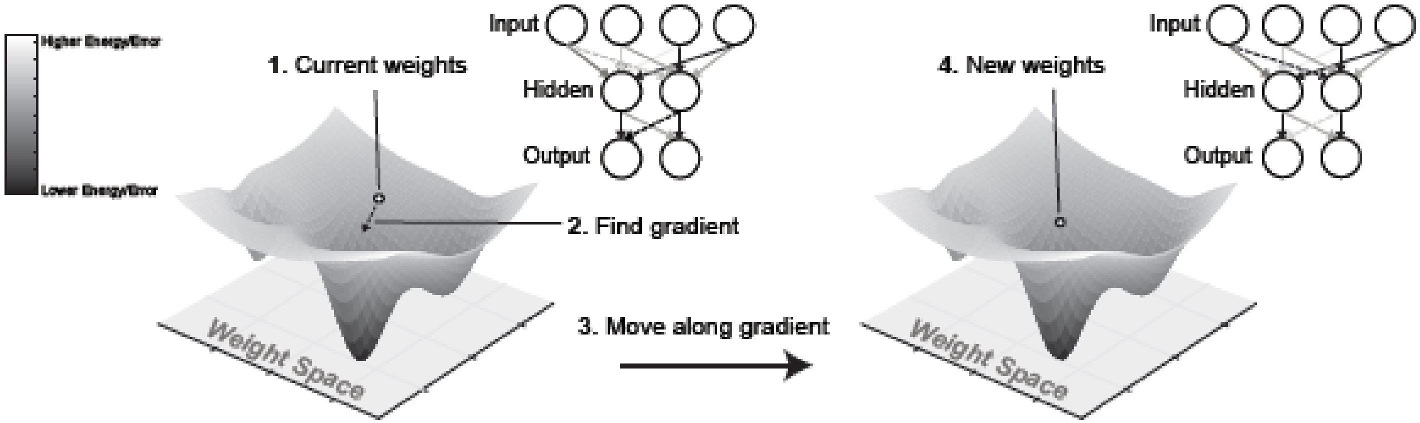
•Samples are from a Bernoulli distribution, p ~ uniform
•Action is prediction of next outcome

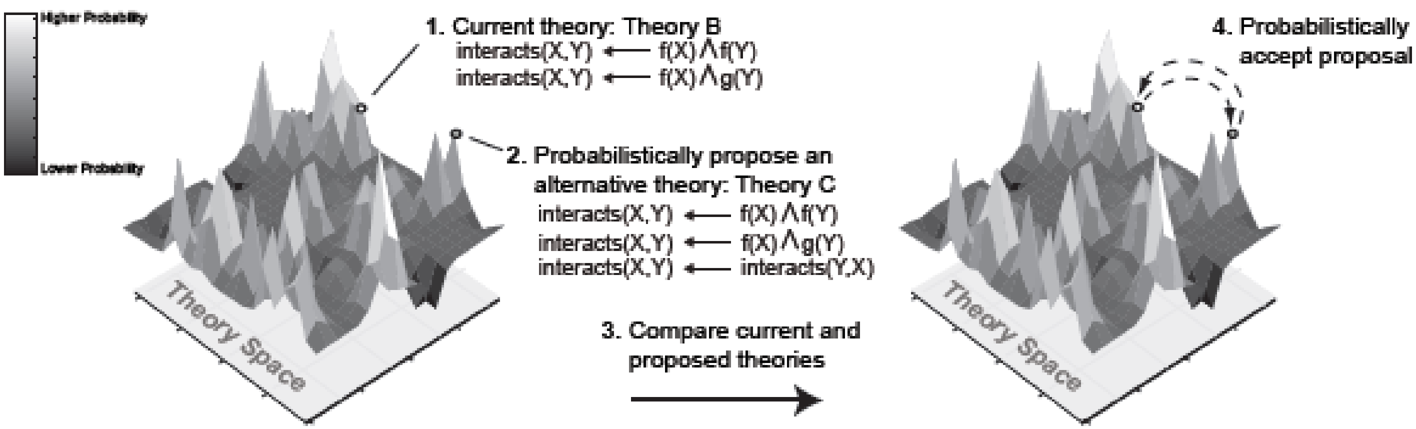
Vul, Goodman, Griffiths, Tenenbaum 2008

Hypothesis space search example



Ullman, Goodman, Tenenbaum 2012





Ullman, Goodman, Tenenbaum 2012

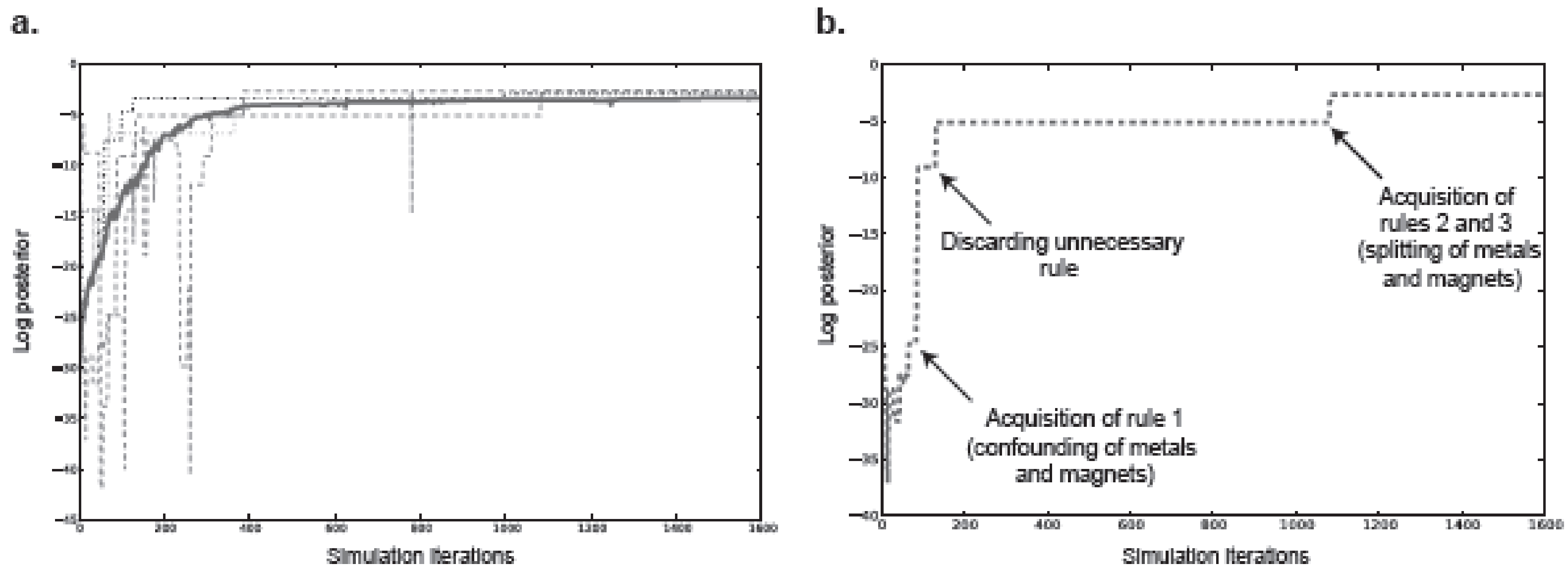


Figure 7: Representative runs of theory learning in Magnetism. (a) Dashed lines show different runs. Solid line is the average across all runs. (b) Highlighting a particular run, showing the acquisition of law 1 and the confounding of magnets and magnetic (but non-magnet) objects, the discarding of an unnecessary law which improves the theory prior, and the acquisition of the final correct theory.

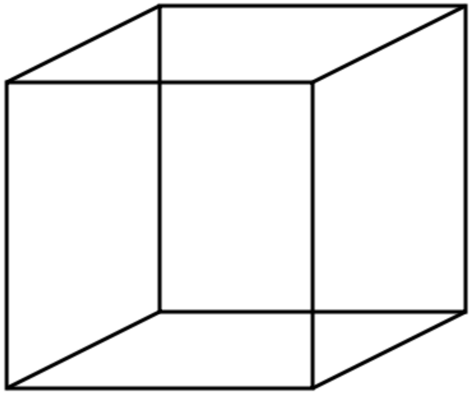
Ullman, Goodman, Tenenbaum 2012

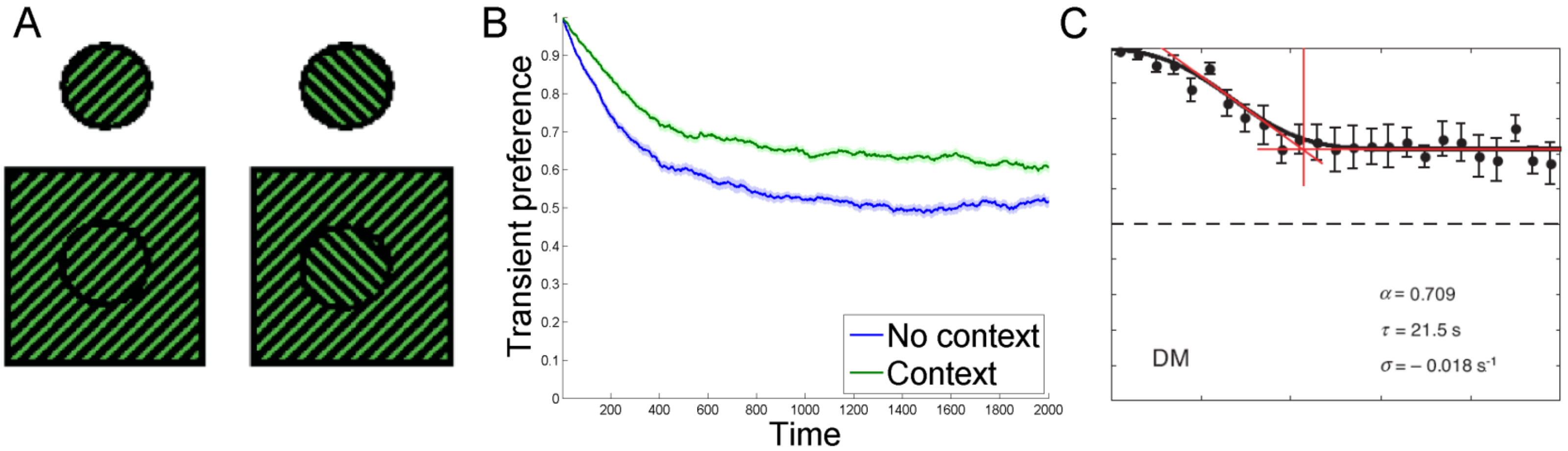
Hypothesis space search: explicit hypotheses

- MCMC with an appropriate grammar can capture some qualitative features of children's learning. What sort of evidence would admit differential predictions?
  - Basic: temporal correlation of hypotheses (often demonstrated)
  - Dependence of likely paths (and perhaps thereby posterior) on grammar used to generate hypotheses
  - Lack of effect of having considered and rejected a hypothesis already (special case of Markov property—no history used)
  - Effects of steepness around an attractive solution, rather than just its likelihood?

Markov chain example in perception: multistable percepts

- Used Markov random field (MRF) lattice model; MCMC to infer hidden cause of image
- Recovered…
  - gamma-distributed dominance times,
  - bias due to context,
  - situations that lead to fusion,
  - switches occurring in travelling waves





Gerschman, Vul, Tenenbaum 2012

Possible directions

- “More cognitive” samplers
  - Allow uncertainty about the data
  - “Focus of attention,” sense of *how* the current hypothesis is lacking
  - Dealing with uncertainty about the model
- Experimental design to test predicted differences in dynamics, performance
- (How) do we constrain the hypothesis space to generate appropriate explanations?