Bayesian models of inductive learning

Tom Griffiths
UC Berkeley

Charles Kemp
CMU

Josh Tenenbaum
MIT

Outline

- Morning
  - 9:00-10:30: Introduction: Why Bayes?; Basics of Bayesian inference (Josh)
  - 11:00-12:30: How to build a Bayesian cognitive model (Tom)

- Afternoon
  - 1:30-3:00: Hierarchical Bayesian models and learning structured representations (Charles)
  - 3:30-5:00: Monte Carlo methods and nonparametric Bayesian models (Tom)

What you will get out of this tutorial

- Our view of what Bayesian models have to offer cognitive science

- In-depth examples of basic and advanced models: how the math works & what it buys you

- A sense for how to go about the process of building Bayesian models

- Some (not extensive) comparison to other approaches
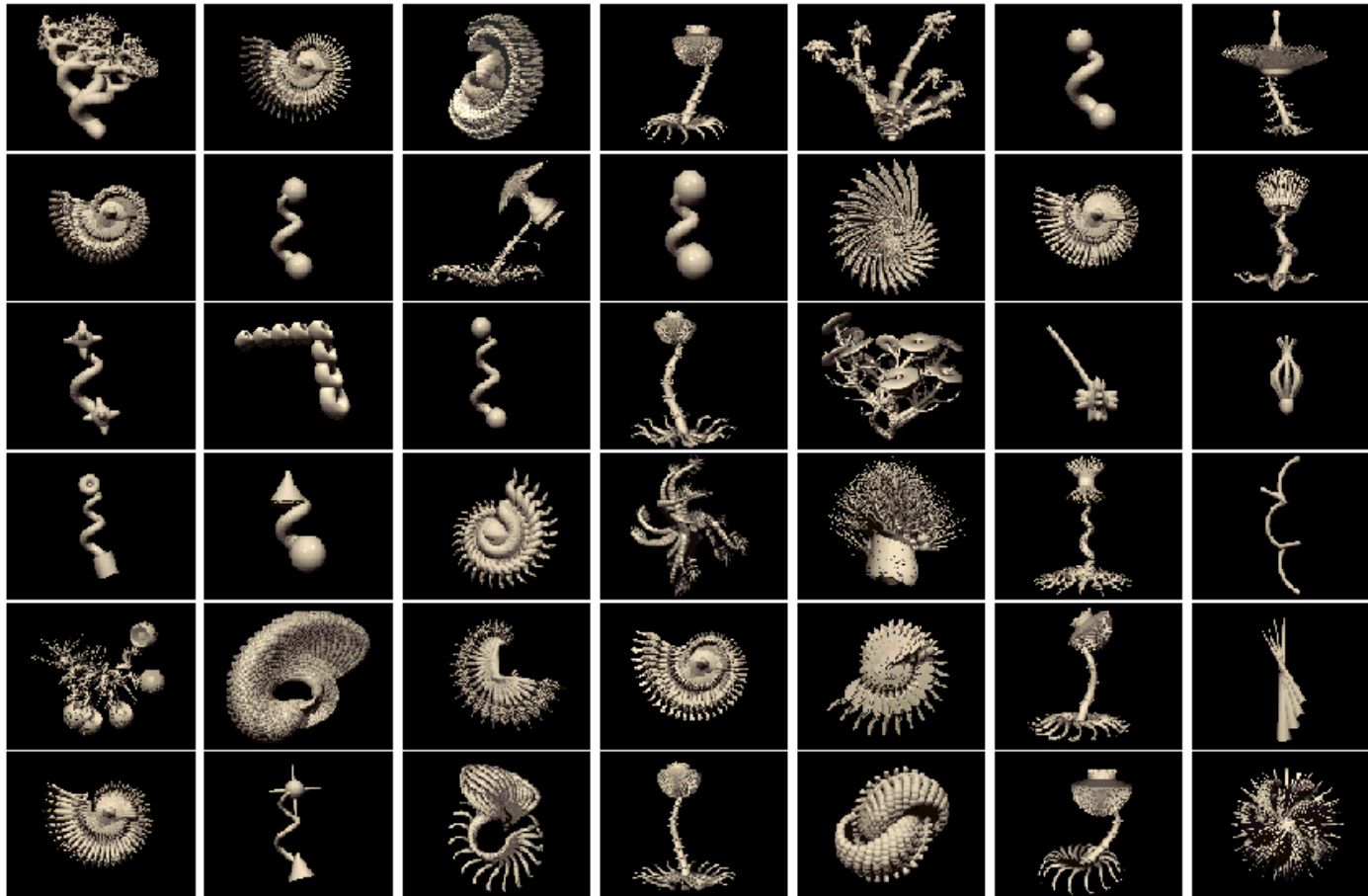
- Opportunities to ask questions

The big question

How does the mind get so much out of so
  little?

Our minds build rich models of the world and make strong
generalizations from input data that is sparse, noisy, and
ambiguous – in many ways far too limited to support the
inferences we make.

How do we do it?

Learning words for objects

Learning words for objects

"tufa"

The big question

How does the mind get so much out of so little?

- – Perceiving the world from sense data
- – Learning about kinds of objects and their properties
- – Inferring causal relations
- – Learning and using words, phrases, and sentences
- – Learning and using intuitive theories of physics, psychology, biology, …
- – Learning social structures, conventions, and rules

The goal: A general-purpose computational framework for understanding how people make these inferences, and how they can be successful.

The problem of induction

Abstract knowledge.

(Constraints / Inductive bias / Priors)

The problems of induction

1. How does abstract knowledge guide inductive learning, inference, and decision-making from sparse, noisy or ambiguous data?

2. What is the form and content of our abstract knowledge of the world?

3. What are the origins of our abstract knowledge? To what extent can it be acquired from experience?

4. How do our mental models grow over a lifetime, balancing simplicity versus data fit (Occam), accommodation versus assimilation (Piaget)?

5. How can learning and inference proceed efficiently and accurately, even in the presence of complex hypothesis spaces?

A toolkit for reverse-engineering induction

1. Bayesian inference in probabilistic generative models

2. Probabilities defined on a range of structured representations: spaces, graphs, grammars, predicate logic, schemas, programs.

3. Hierarchical probabilistic models, with inference at all levels of abstraction

4. Models of unbounded complexity ("nonparametric Bayes" or "infinite models"), which can grow in complexity or change form as observed data dictate.

5. Approximate methods of learning and inference, such as belief propagation, expectation-maximization (EM), Markov chain Monte Carlo (MCMC), and sequential Monte Carlo (particle filtering).
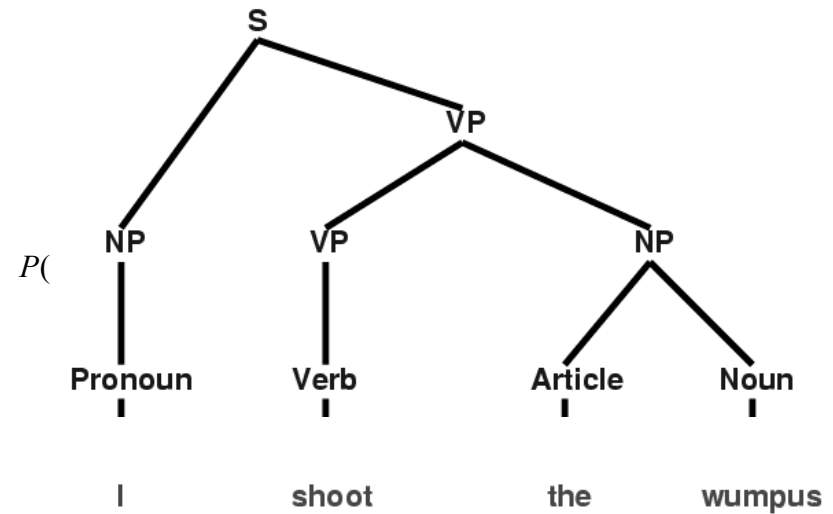
Grammar G

$$S \rightarrow NP\ VP$$
$$NP \rightarrow Det\ [Adj]\ Noun\ [RelClause]$$
$$RelClause \rightarrow [Rel]\ NP\ V$$
$$VP \rightarrow VP\ NP$$
$$VP \rightarrow Verb$$

$\downarrow$ $P(S \mid G)$

Phrase structure S

$\downarrow$ $P(U \mid S)$

Utterance U

$P($

```
            S
           / \
          /   VP
         /   / \
        NP  VP  NP
        |   |   / \
     Pronoun Verb Article Noun
        |    |    |    |
        I  shoot the wumpus
```

$$P(S \mid U, G) \sim P(U \mid S) \times P(S \mid G)$$

Bottom-up   Top-down

"Universal Grammar"

Hierarchical phrase structure grammars (e.g., CFG, HPSG, TAG)

$\downarrow$ $P(\text{grammar} \mid \text{UG})$

Grammar

$$S \rightarrow NP\ VP$$
$$NP \rightarrow Det\ [Adj]\ Noun\ [RelClause]$$
$$RelClause \rightarrow [Rel]\ NP\ V$$
$$VP \rightarrow VP\ NP$$
$$VP \rightarrow Verb$$
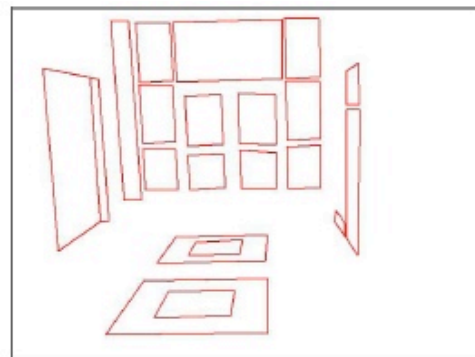
$P(\text{phrase structure} \mid \text{grammar})$

Phrase structure

$P(\text{utterance} \mid \text{phrase structure})$

Utterance

$\downarrow$ $P(\text{speech} \mid \text{utterance})$

Speech signal
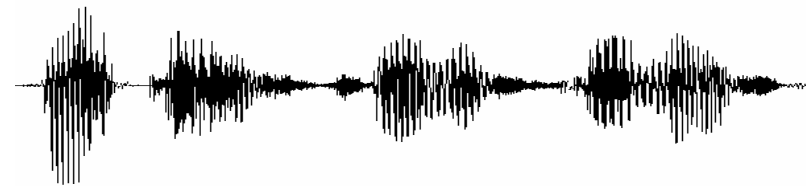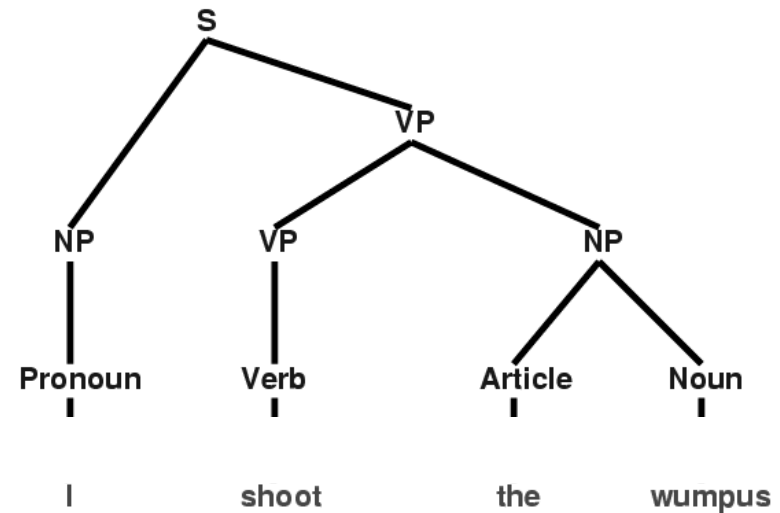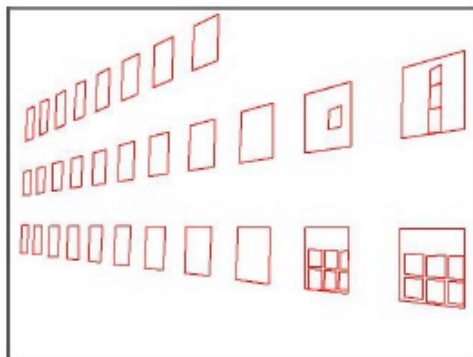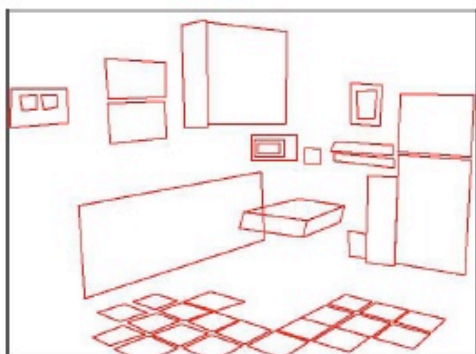
Vision as probabilistic parsing
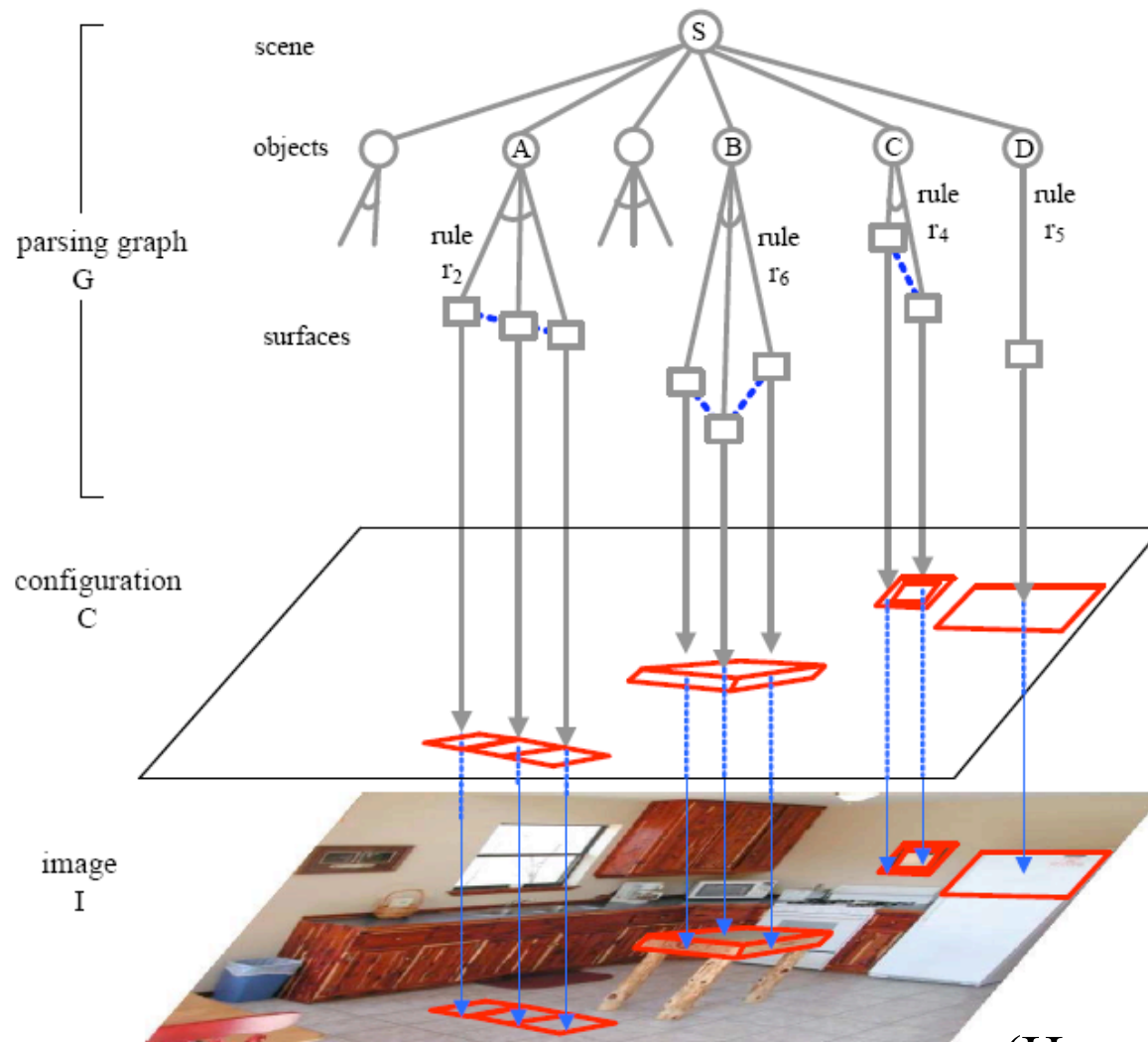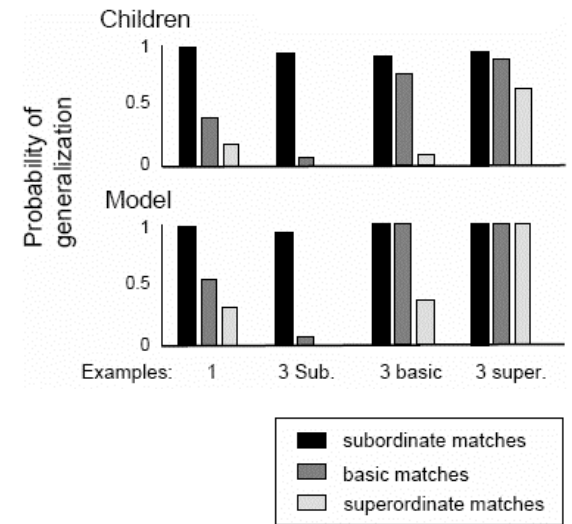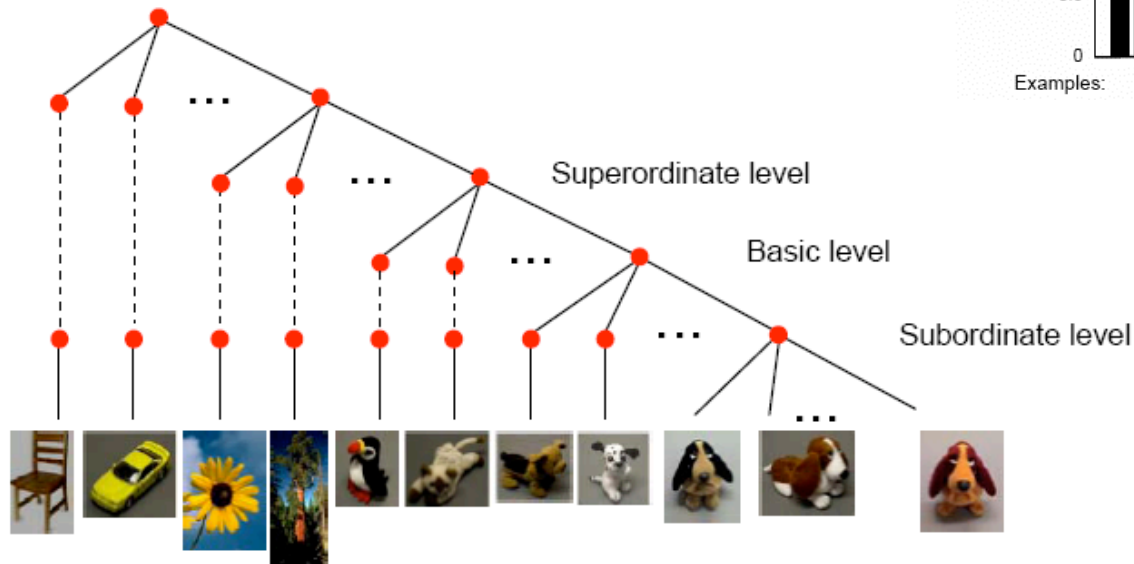

(Han and Zhu, 2006)

Learning word meanings

Principles

Whole-object principle
Shape bias
Taxonomic principle
Contrast principle
Basic-level bias

Structure

Superordinate level

Basic level

Subordinate level

Data

"fep"   "fep"   "fep"   "dax"   "zoog"   "gazzer"   ...

Children

Probability of generalization

1

0.5

0

Model

1

0.5

0

Examples: 1   3 Sub.   3 basic   3 super.

■ subordinate matches
■ basic matches
□ superordinate matches

Causal learning and reasoning

Principles

Classes: {R, D, S} (Risks, Diseases, Symptoms)
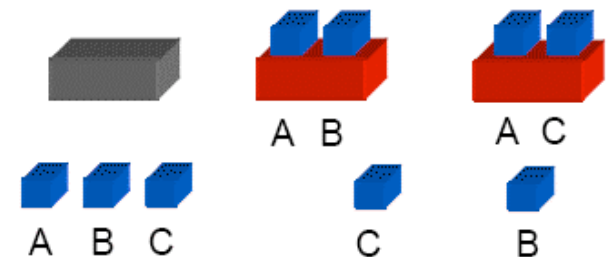Causal laws: R → D, D → S

Objects can activate Machines
Activation requires contact
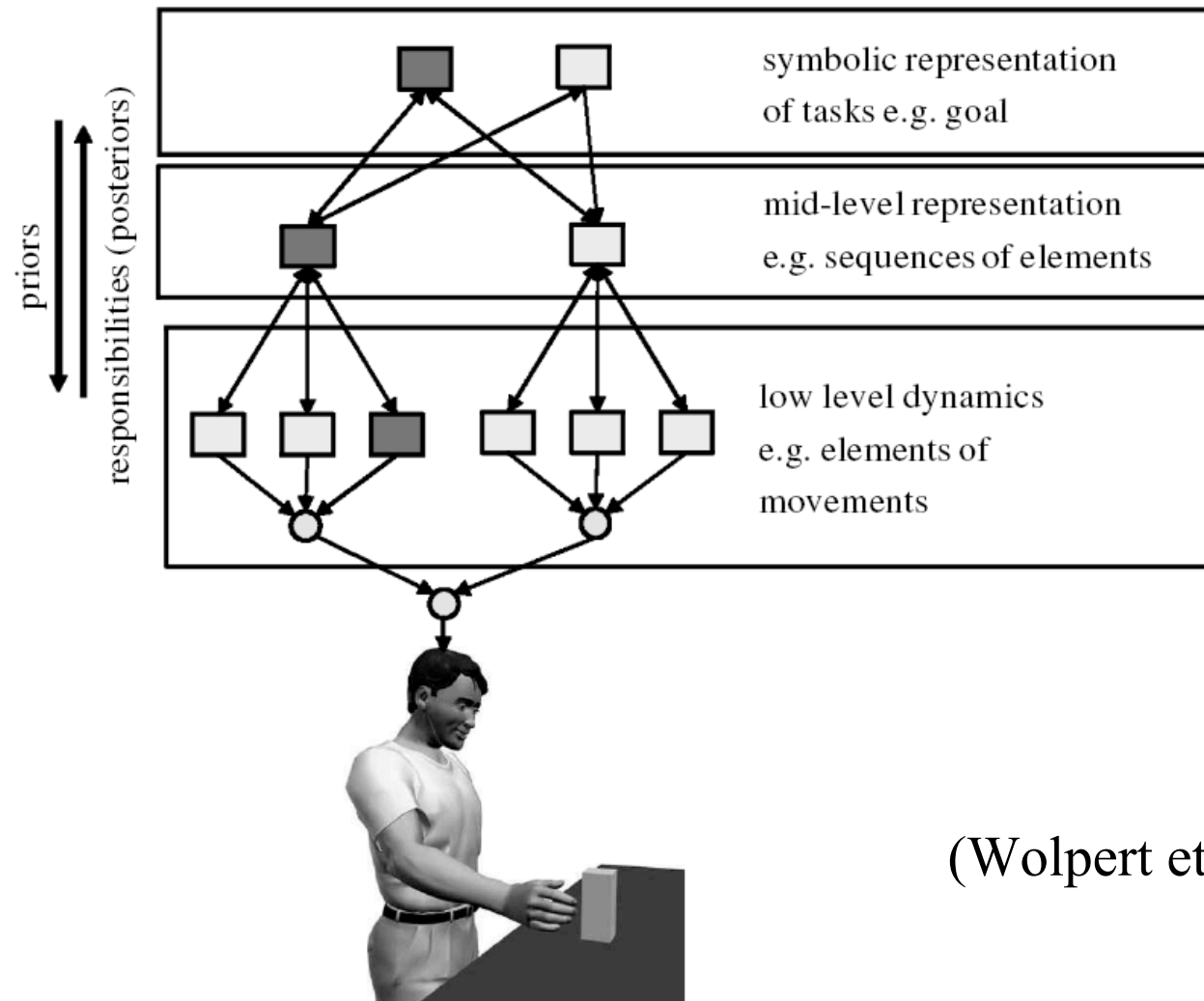Machines are (near) deterministic

Structure
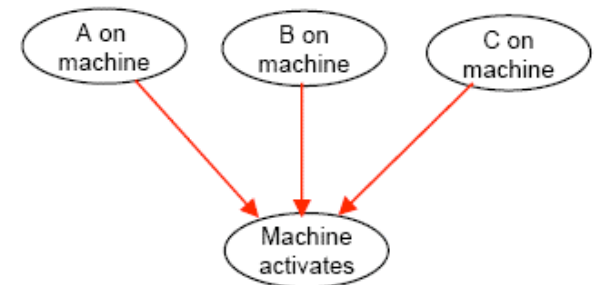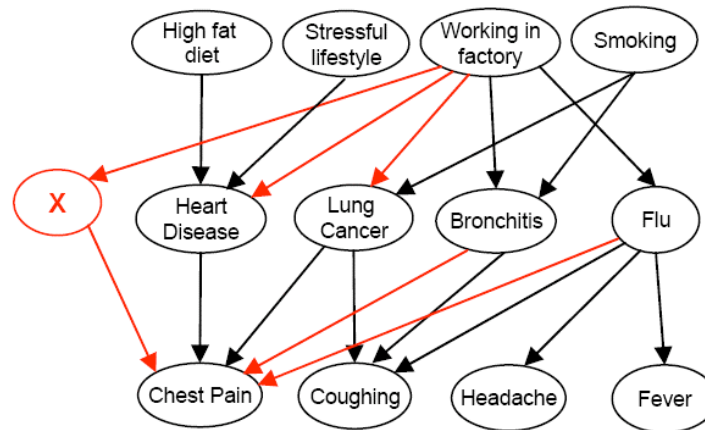
Data

Patient 1: Stressful lifestyle
 Chest Pain
Patient 2: Smoking
 Coughing
Patient 3: Working in factory
 Chest Pain
…

Goal-directed action
(production and comprehension)

(Wolpert et al., 2003)

Why Bayesian models of cognition?

- A framework for understanding how the mind can solve fundamental problems of induction.

- Strong, principled quantitative models of human cognition.

- Tools for studying people's implicit knowledge of the world.

- Beyond classic limiting dichotomies: "rules vs. statistics", "nature vs. nurture", "domain-general vs. domain-specific" .

- A unifying mathematical language for all of the cognitive sciences: AI, machine learning and statistics, psychology, neuroscience, philosophy, linguistics…. A bridge between engineering and "reverse-engineering".
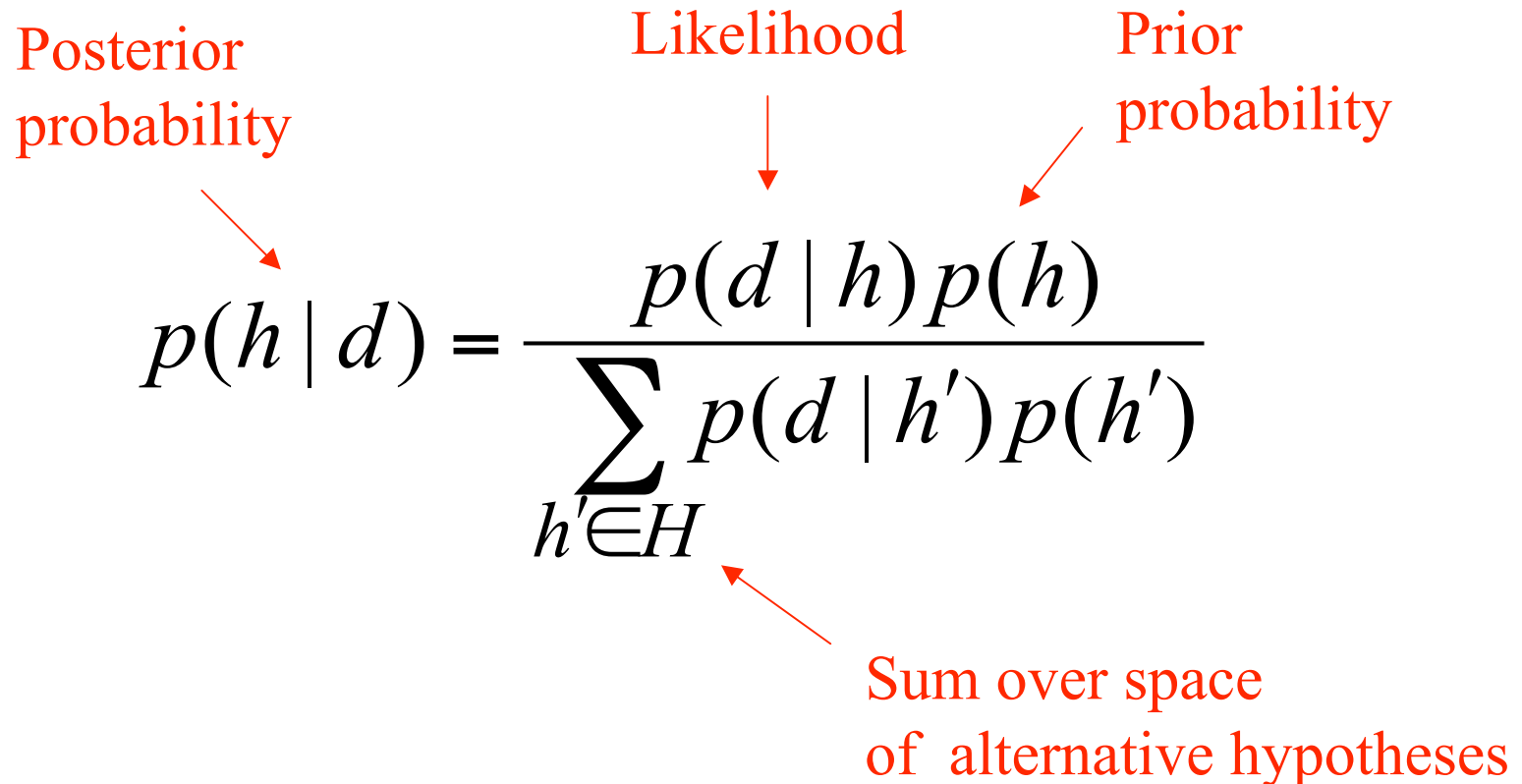
Why now? Much recent progress, in computational resources, theoretical tools, and interdisciplinary connections.

Outline

- ## Morning
  - Introduction: Why Bayes? (Josh)
  - Basics of Bayesian inference (Josh)
  - How to build a Bayesian cognitive model (Tom)

- ## Afternoon
  - Hierarchical Bayesian models and learning structured representations (Charles)
  - Monte Carlo methods and nonparametric Bayesian models (Tom)

Bayes' rule

For any hypothesis *h* and data *d*,

Posterior probability

Likelihood

Prior probability

$$p(h \mid d) = \frac{p(d \mid h) \, p(h)}{\displaystyle\sum_{h' \in H} p(d \mid h') \, p(h')}$$

Sum over space of alternative hypotheses

Bayesian inference

- Bayes' rule: $P(h|d) = \dfrac{P(h)P(d|h)}{\sum\limits_{h_i} P(h_i)P(d|h_i)}$

- An example

  - Data: John is coughing

  - Some hypotheses:

    1. John has a cold

    2. John has lung cancer

    3. John has a stomach flu

  - Prior $P(h)$ favors 1 and 3 over 2

  - Likelihood $P(d|h)$ favors 1 and 2 over 3

  - Posterior $P(h|d)$ favors 1 over 2 and 3

Plan for this lecture

- Some basic aspects of Bayesian statistics
  - Comparing two hypotheses
  - Model fitting
  - Model selection
- Two (very brief) case studies in modeling human inductive learning
  - Causal learning
  - Concept learning

Coin flipping

- Basic Bayes
  - data = `HHTHT` or `HHHHH`
  - compare two hypotheses:
    $P(\texttt{H}) = 0.5$ vs. $P(\texttt{H}) = 1.0$

- Parameter estimation (Model fitting)
  - compare many hypotheses in a parameterized family
    $P(\texttt{H}) = \theta$: Infer θ

- Model selection
  - compare qualitatively different hypotheses, often varying in complexity:
    $P(\texttt{H}) = 0.5$ vs. $P(\texttt{H}) = \theta$

Coin flipping

HHTHT

HHHHH

What process produced these sequences?

Comparing two hypotheses

- Contrast simple hypotheses:
  - h_1: "fair coin", $P(\texttt{H}) = 0.5$
  - h_2: "always heads", $P(\texttt{H}) = 1.0$
- Bayes' rule:

$$P(h \mid d) = \frac{P(h)P(d \mid h)}{\sum_{h_i} P(h_i)P(d \mid h_i)}$$

- With two hypotheses, use odds form

Comparing two hypotheses

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D: HHTHT

H_1, H_2:   "fair coin", "always heads"

$P(D|H_1) =$ $1/2^5$ $P(H_1) =$?

$P(D|H_2) =$ 0 $P(H_2) =$ 1-?

Comparing two hypotheses

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D: HHTHT

H_1, H_2: "fair coin", "always heads"

$P(D|H_1) =$ $1/2^5$ $\qquad$ $P(H_1) =$ 999/1000

$P(D|H_2) =$ 0 $\qquad$ $P(H_2) =$ 1/1000

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{1/32}{0} \times \frac{999}{1} = \text{infinity}$$

Comparing two hypotheses

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D: HHHHH

H_1, H_2: "fair coin", "always heads"

$P(D|H_1) = 1/2^5$ $\qquad$ $P(H_1) = 999/1000$

$P(D|H_2) = 1$ $\qquad\qquad$ $P(H_2) = 1/1000$

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{1/32}{1} \times \frac{999}{1} \approx 30$$

Comparing two hypotheses

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D: HHHHHHHHHH

H_1, H_2:   "fair coin", "always heads"

$P(D|H_1) = $ $1/2^{10}$ $P(H_1) = $ 999/1000

$P(D|H_2) = $ 1 $P(H_2) = $ 1/1000

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{1/1024}{1} \times \frac{999}{1} \approx 1$$

Measuring prior knowledge

1. The fact that `HHHHH` looks like a "mere coincidence", without making us suspicious that the coin is unfair, while `HHHHHHHHHH` does begin to make us suspicious, measures the strength of our prior belief that the coin is fair.

   - If θ is the threshold for suspicion in the posterior odds, and D^* is the shortest suspicious sequence, the prior odds for a fair coin is roughly $\theta/P(D^*|$"fair coin"$)$.
   - If $\theta \sim 1$ and D^* is between 10 and 20 heads, prior odds are roughly between 1/1,000 and 1/1,000,000.

2. The fact that `HHTHT` looks representative of a fair coin, and `HHHHH` does not, reflects our prior knowledge about possible causal mechanisms in the world.

   - Easy to imagine how a trick all-heads coin could work: low (but not negligible) prior probability.
   - Hard to imagine how a trick "`HHTHT`" coin could work: extremely low (negligible) prior probability.

Plan for this lecture

- Some basic aspects of Bayesian statistics
  - Comparing two hypotheses
  - Model fitting
  - Model selection
- Two (very brief) case studies in modeling human inductive learning
  - Causal learning
  - Concept learning

Coin flipping

- Basic Bayes
  - data = `HHTHT` or `HHHHH`
  - compare two hypotheses:
    $P(\mathtt{H}) = 0.5$ vs. $P(\mathtt{H}) = 1.0$

- Parameter estimation (Model fitting)
  - compare many hypotheses in a parameterized family
    $P(\mathtt{H}) = \theta$: Infer θ

- Model selection
  - compare qualitatively different hypotheses, often varying in complexity:
    $P(\mathtt{H}) = 0.5$ vs. $P(\mathtt{H}) = \theta$

Model fitting (Parameter estimation)

- Assume data are generated from a parameterized model:
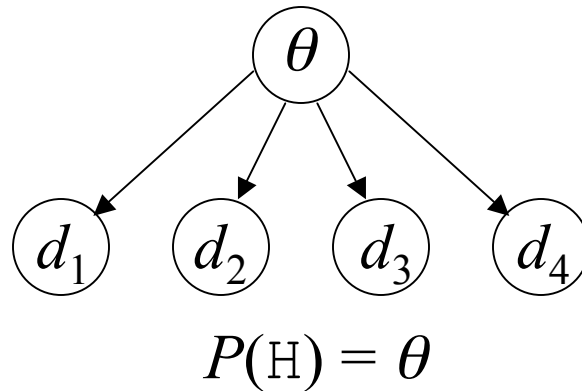
$$P(\mathtt{H}) = \theta$$

- What is the value of θ ?
  - each value of θ is a hypothesis H
  - requires inference over infinitely many hypotheses
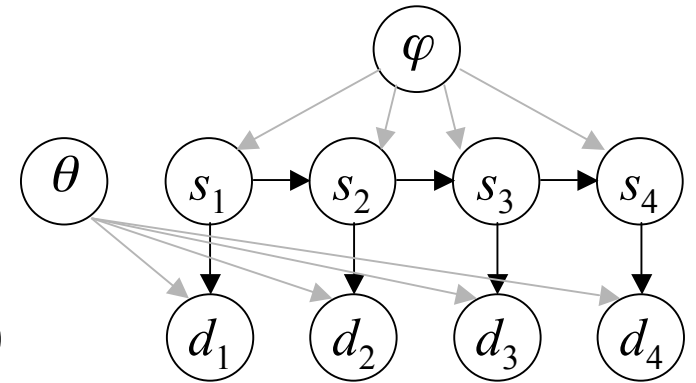
Model selection

- Assume hypothesis space of possible models:
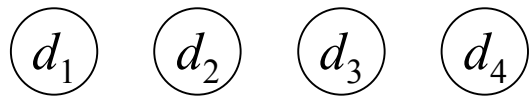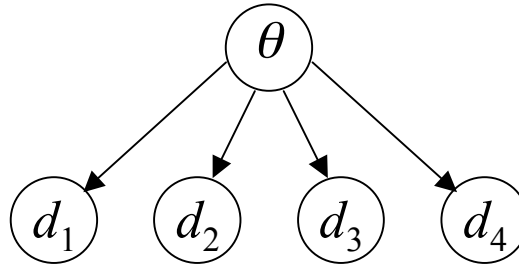
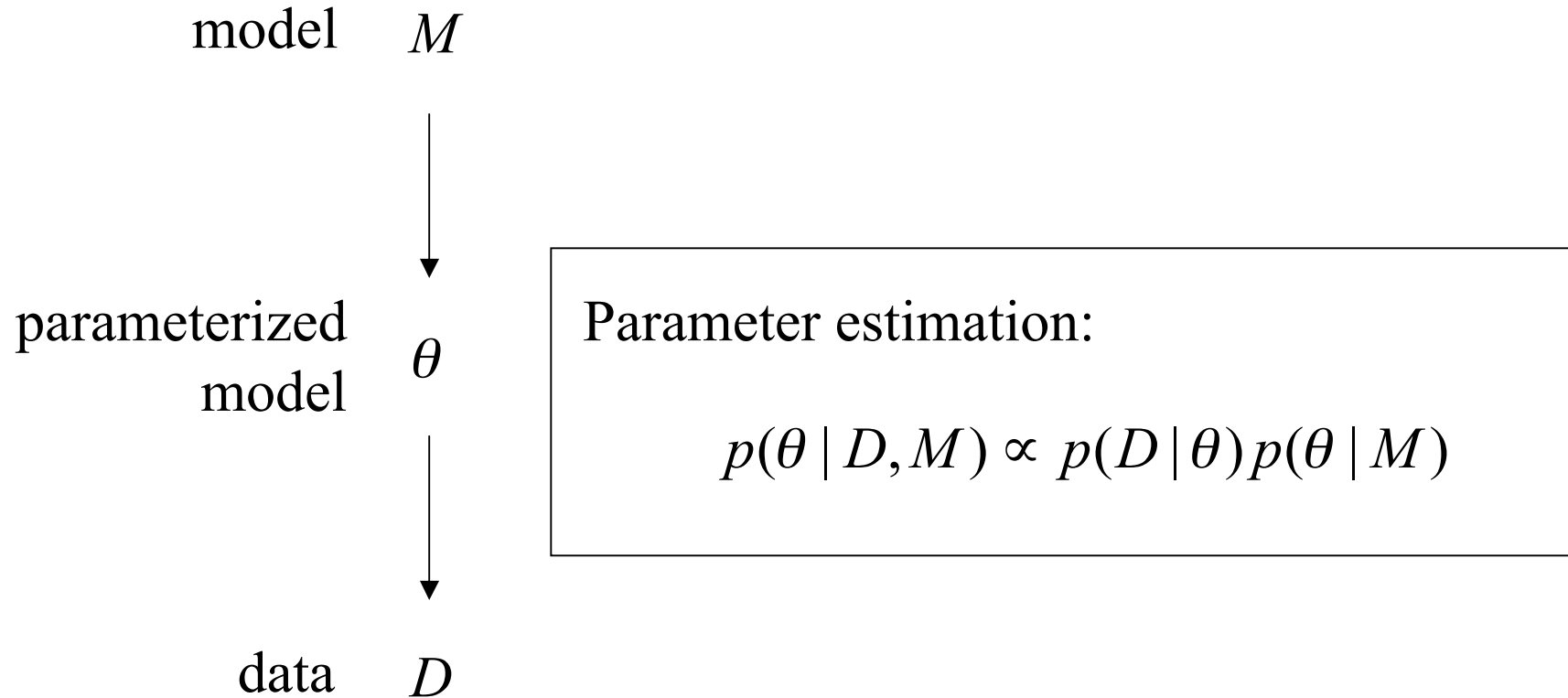Fair coin: $P(\text{H}) = 0.5$ $P(\text{H}) = \theta$ Hidden Markov model:
$s_i \in \{\text{Fair coin, Trick coin}\}$

- Which model generated the data?
  - requires summing out hidden variables
  - requires some form of Occam's razor to trade off complexity with fit to the data.
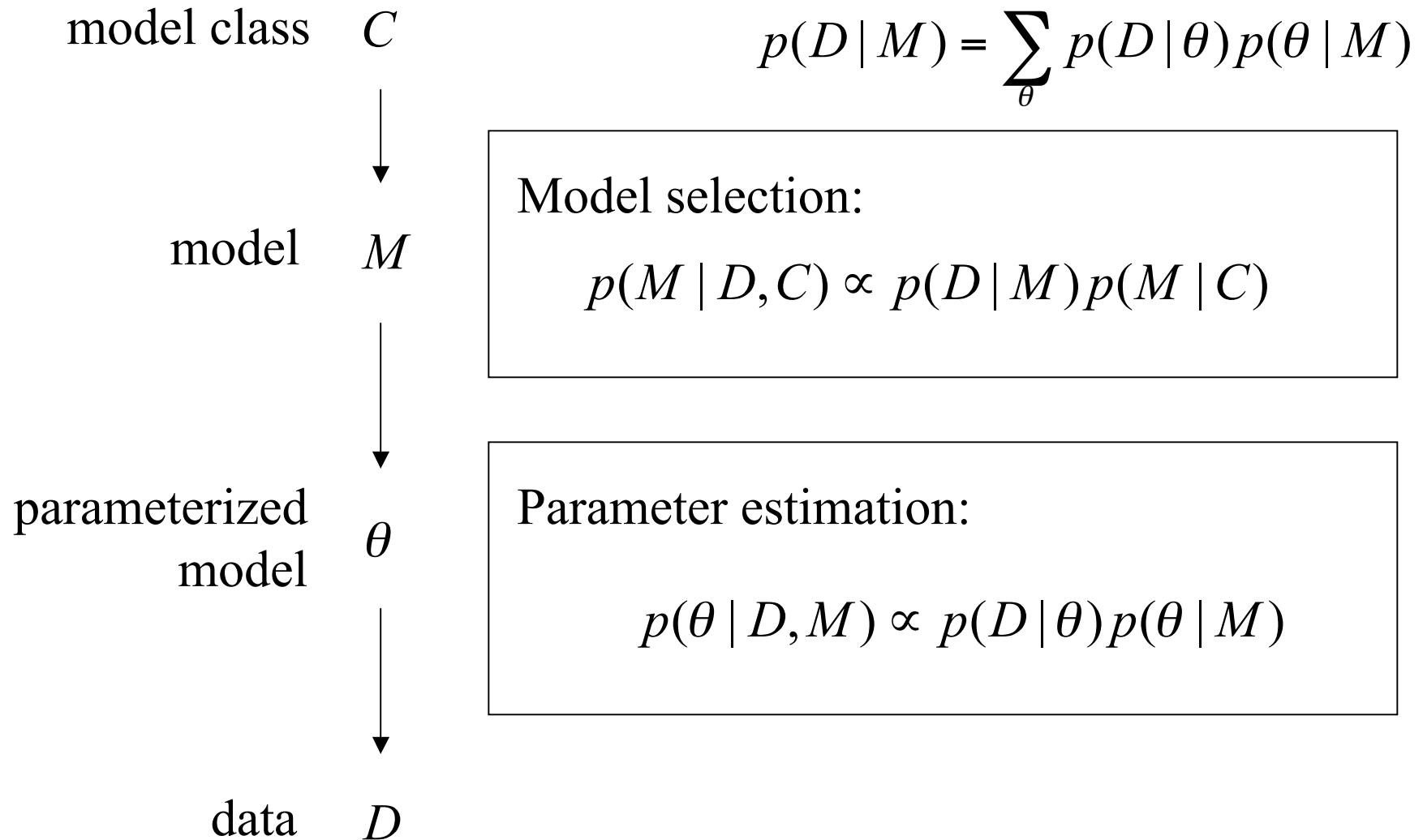
Parameter estimation *vs.* Model selection across learning and development

- *Causality:* learning the strength of a relation *vs.* learning the existence and form of a relation

- *Perception:* learning the strength of a cue *vs.* learning the existence of a cue, in sensory cue combination

- *Language acquisition:* learning a speaker's accent, or frequencies of different words *vs.* learning a new tense or syntactic rule (or learning a new language, or the existence of different languages)

- *Concepts:* learning what horses look like *vs.* learning that there is a new species (or learning that there *are* species)

- *Intuitive physics:* learning the mass of an object *vs.* learning about the existence of a force (e.g., gravity, magnetism)
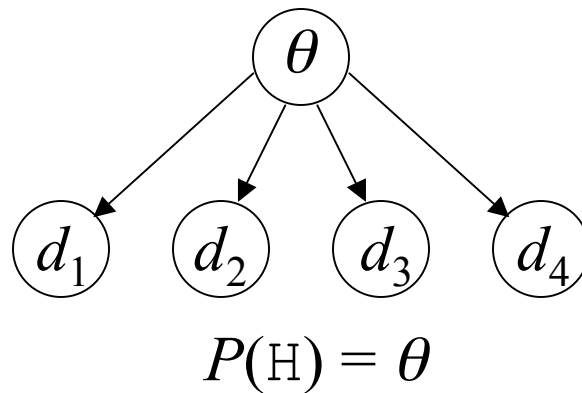
A hierarchical learning framework

model M

parameterized
model θ

data D

Parameter estimation:

$$p(\theta \,|\, D, M) \propto p(D \,|\, \theta)\, p(\theta \,|\, M)$$

A hierarchical learning framework

model class C

$$p(D\,|\,M) = \sum_{\theta} p(D\,|\,\theta)\,p(\theta\,|\,M)$$

model M

Model selection:
$$p(M\,|\,D,C) \propto p(D\,|\,M)\,p(M\,|\,C)$$

parameterized model θ

Parameter estimation:
$$p(\theta\,|\,D,M) \propto p(D\,|\,\theta)\,p(\theta\,|\,M)$$

data D

Bayesian parameter estimation

- Assume data are generated from a model:

$$P(\mathrm{H}) = \theta$$

- What is the value of θ ?
  - each value of θ is a hypothesis H
  - requires inference over infinitely many hypotheses

Some intuitions

- $D = 10$ flips, with 5 heads and 5 tails.
- $\theta = P(\texttt{H})$ on next flip? 50%
- Why? 50% = 5 / (5+5) = 5/10.
- Why? "The future will be like the past"

- Suppose we had seen 4 heads and 6 tails.
- $P(\texttt{H})$ on next flip? Closer to 50% than to 40%.
- Why? Prior knowledge.

Integrating prior knowledge and data

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{\int p(D \mid \theta')\, p(\theta')\, d\theta'}$$

- Posterior distribution $P(\theta \mid D)$ is a probability density over $\theta = P(\texttt{H})$
- Need to specify likelihood $P(D \mid \theta)$ and prior distribution $P(\theta)$.

Likelihood and prior

- Likelihood: Bernoulli distribution

$$P(D \mid \theta) = \theta^{N_H} (1-\theta)^{N_T}$$

  - N_H: number of heads
 - N_T: number of tails

- Prior:

$$P(\theta) \propto \quad ?$$

Some intuitions

- $D = 10$ flips, with 5 heads and 5 tails.
- $\theta = P(\texttt{H})$ on next flip? 50%
- Why? 50% = 5 / (5+5) = 5/10.
- Why? *Maximum likelihood:* $\hat{\theta} = \arg\max_{\theta} P(D \mid \theta)$

- Suppose we had seen 4 heads and 6 tails.
- $P(\texttt{H})$ on next flip? Closer to 50% than to 40%.
- Why? Prior knowledge.

A simple method of specifying priors

- Imagine some fictitious trials, reflecting a set of previous experiences
  - strategy often used with neural networks or building invariance into machine vision.

- e.g., $F = \{1000$ heads, 1000 tails$\}$ ~ strong expectation that any new coin will be fair

- In fact, this is a sensible statistical idea...

Likelihood and prior

- Likelihood: Bernoulli(θ) distribution

$$P(D \mid \theta) = \theta^{N_{\text{H}}} (1-\theta)^{N_{\text{T}}}$$

  - N_{H}: number of heads observed
  - N_{T}: number of tails observed

- Prior: Beta($F_{\text{H}}, F_{\text{T}}$) distribution

$$P(\theta) \propto \theta^{F_{\text{H}}-1} (1-\theta)^{F_{\text{T}}-1}$$

  - F_{H}: fictional observations of heads
  - F_{T}: fictional observations of tails

Shape of the Beta prior
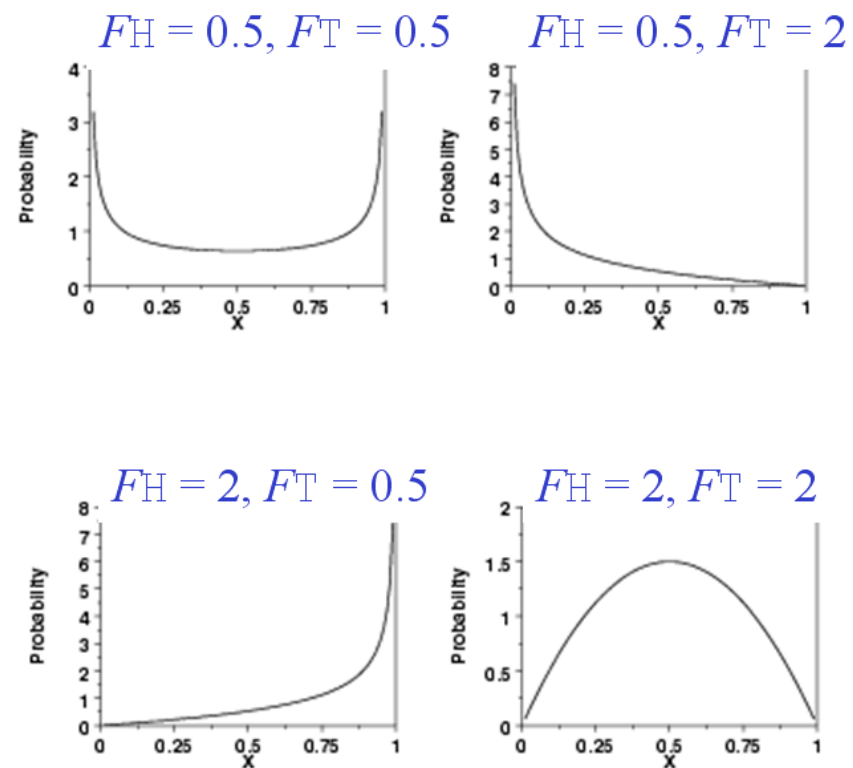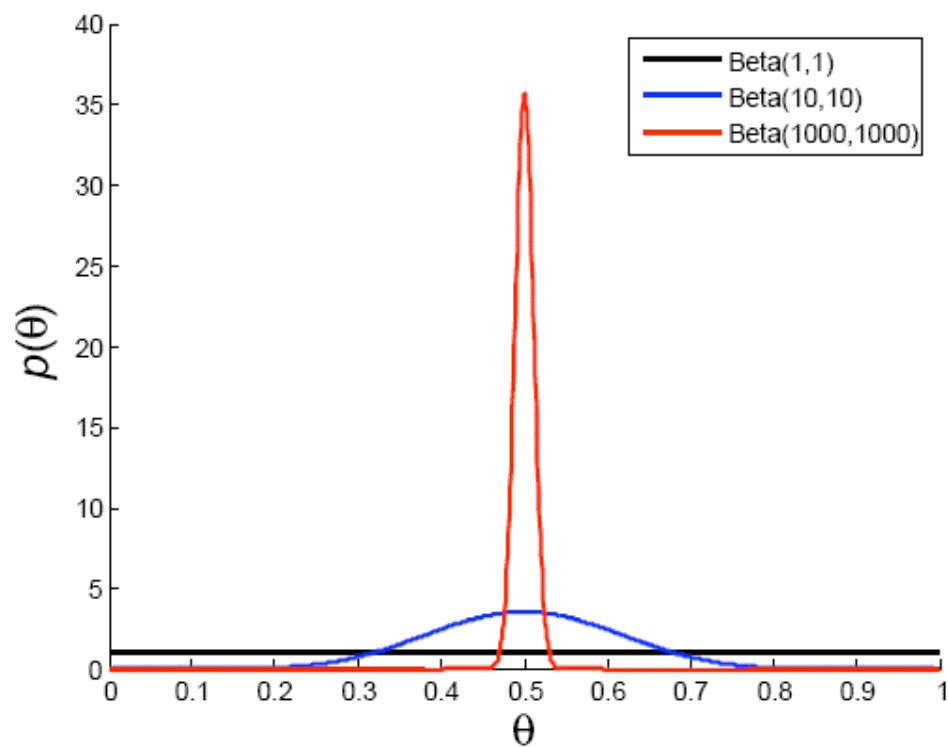
Bayesian parameter estimation

$$P(\theta \mid D) \propto P(D \mid \theta)\, P(\theta) = \theta^{\,N_H + F_H - 1} (1-\theta)^{\,N_T + F_T - 1}$$

- Posterior is Beta($N_H + F_H, N_T + F_T$)
  - same form as prior!

Conjugate priors

- A prior $p(\theta)$ is *conjugate* to a likelihood function $p(D \mid \theta)$ if the posterior has the same functional form of the prior.

  - Parameter values in the prior can be thought of as a summary of "fictitious observations".

  - Different parameter values in the prior and posterior reflect the impact of observed data.

  - Conjugate priors exist for many standard models (e.g., all exponential family models)

Bayesian parameter estimation
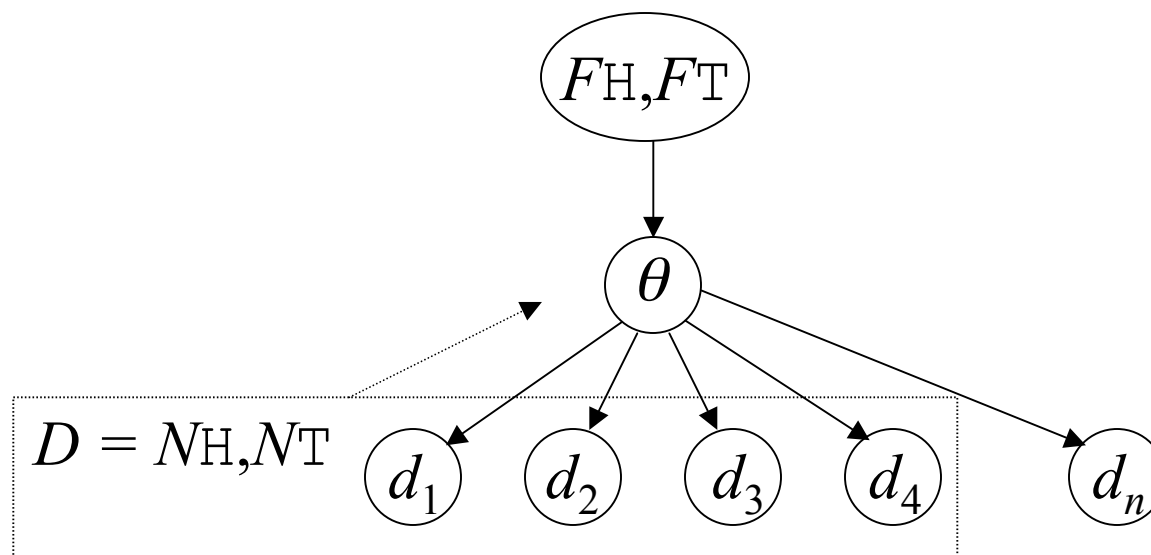
$$P(\theta \mid D) \propto P(D \mid \theta) \, P(\theta) = \theta^{\, N_{\mathrm{H}}+F_{\mathrm{H}}-1} \, (1-\theta)^{\, N_{\mathrm{T}}+F_{\mathrm{T}}-1}$$

- Posterior predictive distribution:

$$P(d_n = \mathrm{H} \mid D, F_{\mathrm{H}}, F_{\mathrm{T}}) = \int_0^1 P(\mathrm{H} \mid \theta) \, P(\theta \mid D, F_{\mathrm{H}}, F_{\mathrm{T}}) \, d\theta$$

"hypothesis averaging"

Bayesian parameter estimation

$$P(\theta \mid D) \propto P(D \mid \theta) \, P(\theta) = \theta^{\,N\mathrm{H}+F\mathrm{H}-1} (1-\theta)^{\,N\mathrm{T}+F\mathrm{T}-1}$$
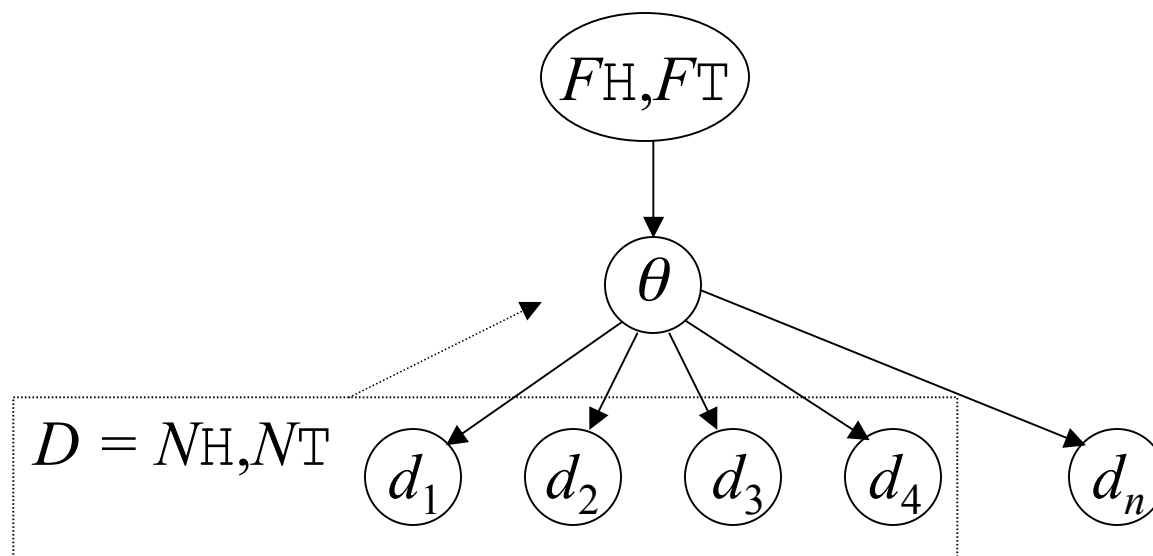
- Posterior predictive distribution:

$$P(d_n = \mathrm{H} \mid D, F\mathrm{H}, F\mathrm{T}) = \frac{(N\mathrm{H}+F\mathrm{H})}{(N\mathrm{H}+F\mathrm{H}+N\mathrm{T}+F\mathrm{T})}$$

Example: coin fresh from bank

- e.g., $F = \{1000$ heads, 1000 tails$\} \sim$ strong expectation that any new coin will be fair

- After seeing 4 heads, 6 tails, $P(\texttt{H})$ on next flip $= 1004 / (1004+1006) = 49.95\%$

- Compare: $F = \{3$ heads, 3 tails$\} \sim$ weak expectation that any new coin will be fair

- After seeing 4 heads, 6 tails, $P(\texttt{H})$ on next flip $= 7 / (7+9) = 43.75\%$

Example: thumbtack
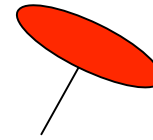
- e.g., $F = \{5 \text{ heads, 3 tails}\}$ ~ weak expectation that tacks are slightly biased towards heads
- After seeing 2 heads, 0 tails, $P(\texttt{H})$ on next flip $= 7 / (7+3) = 70\%$

- Some prior knowledge is always necessary to avoid jumping to hasty conclusions...
- Suppose F = { }: After seeing 1 heads, 0 tails, $P(\texttt{H})$ on next flip $= 1 / (1+0) = 100\%$

Origin of prior knowledge

- Tempting answer: prior experience
- Suppose you have previously seen 2000 coin flips: 1000 heads, 1000 tails

Problems with simple empiricism

- Haven't really seen 2000 coin flips, or *any* flips of a thumbtack
  - Prior knowledge is stronger than raw experience justifies
- Haven't seen exactly equal number of heads and tails
  - Prior knowledge is smoother than raw experience justifies
- Should be a difference between observing 2000 flips of a single coin versus observing 10 flips each for 200 coins, or 1 flip each for 2000 coins
  - Prior knowledge is more structured than raw experience

A simple theory

- "Coins are manufactured by a standardized procedure that is effective but not perfect, and not in principle biased toward heads or tails."
  - Justifies generalizing from previous coins to the present coin.
  - Justifies smoother and stronger prior than raw experience alone.
  - Explains why seeing 10 flips each for 200 coins is more valuable than seeing 2000 flips of one coin.

A hierarchical Bayesian model
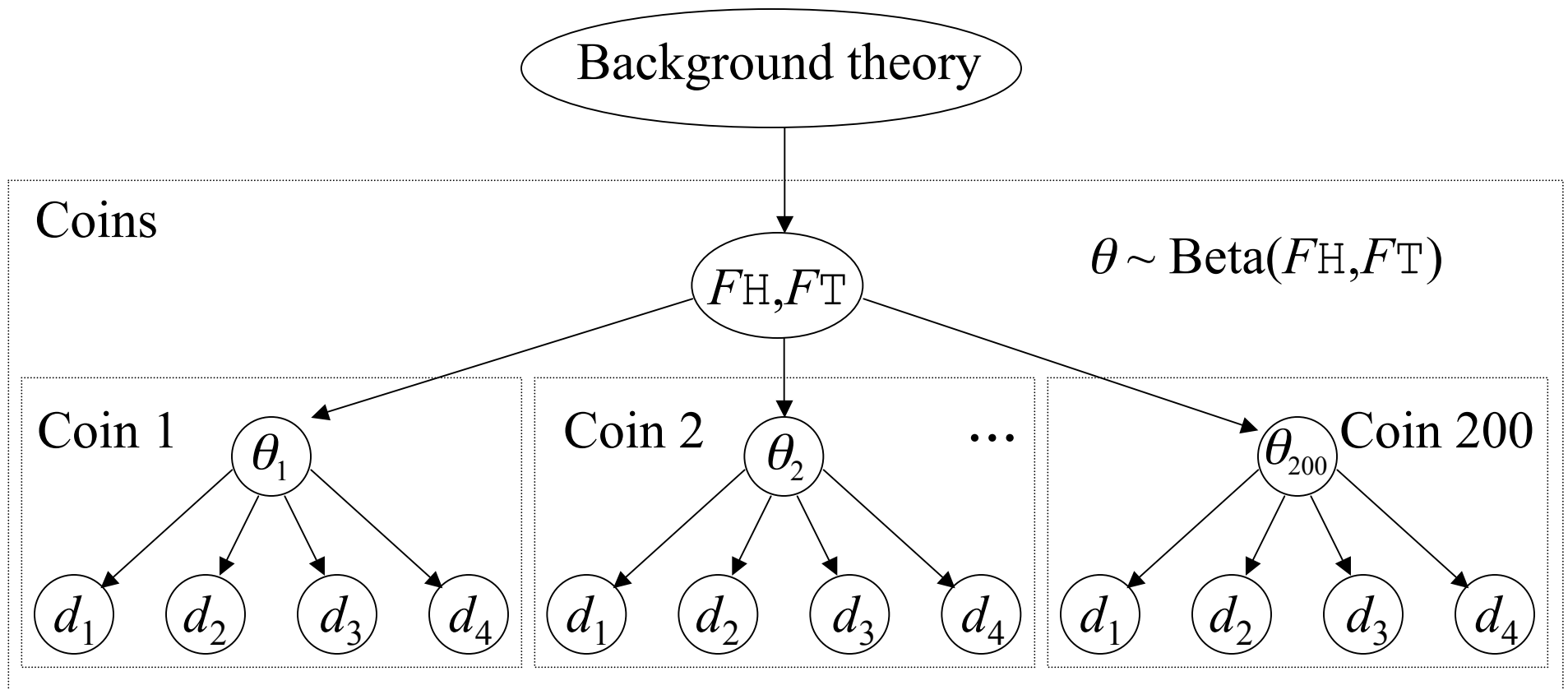
- Qualitative prior knowledge (e.g., symmetry) can influence estimates of continuous parameters (F_H, F_T).

- Explains why 10 flips of 200 coins are better than 2000 flips of a single coin: more informative about F_H, F_T.

Summary: Bayesian parameter estimation

- Learning the parameters of a generative model as Bayesian inference.

- Prediction by Bayesian hypothesis averaging.

- Conjugate priors
  - an elegant way to represent simple kinds of prior knowledge.

- Hierarchical Bayesian models
  - integrate knowledge across instances of a system, or different systems within a domain, to explain the origins of priors.

A hierarchical learning framework

model class C

$$p(D\,|\,M) = \sum_{\theta} p(D\,|\,\theta)\,p(\theta\,|\,M)$$

model M

Model selection:

$$p(M\,|\,D,C) \propto p(D\,|\,M)\,p(M\,|\,C)$$

parameterized model θ

Parameter estimation:

$$p(\theta\,|\,D,M) \propto p(D\,|\,\theta)\,p(\theta\,|\,M)$$

data D
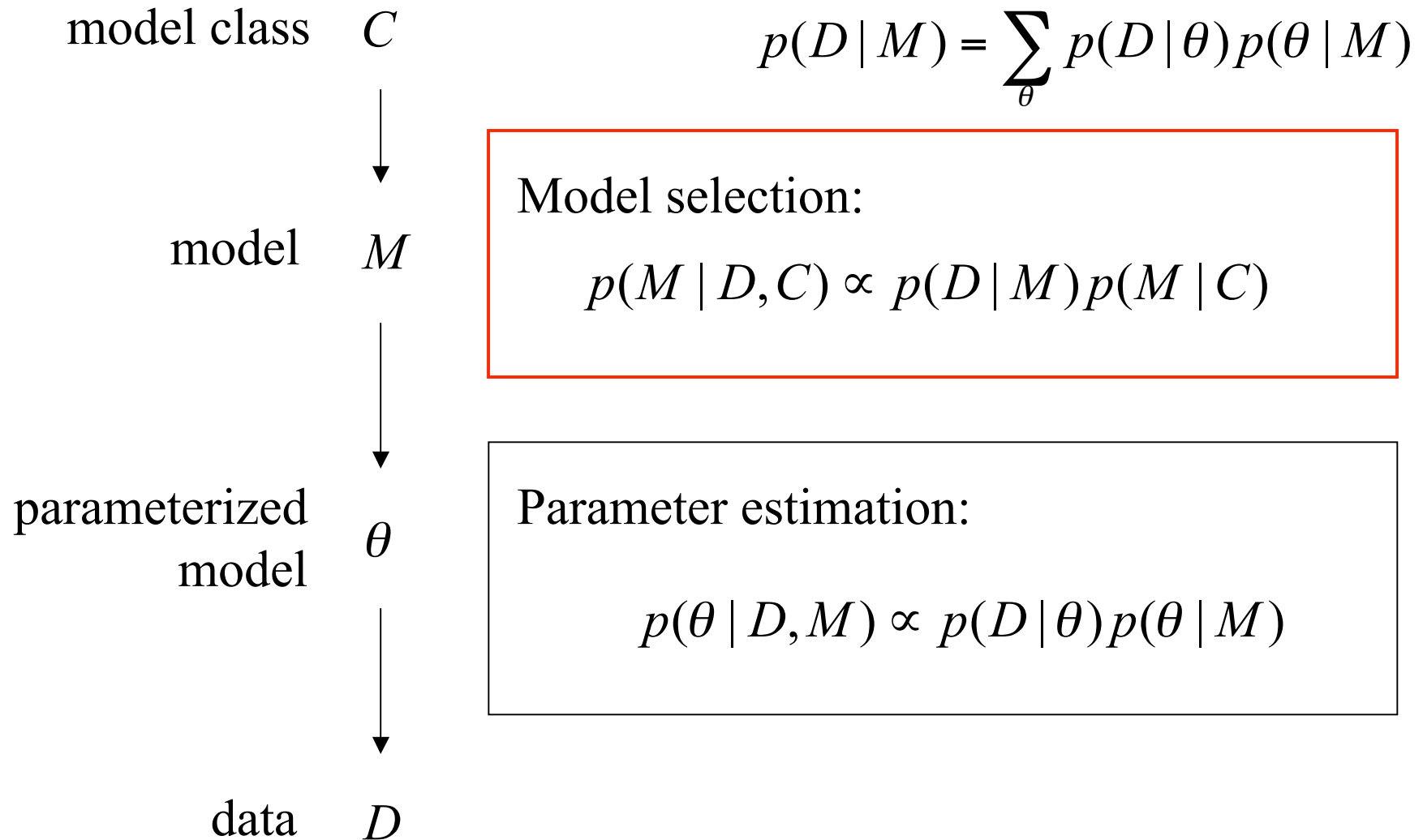
Stability versus Flexibility

- Can all domain knowledge be represented with conjugate priors?

- Suppose you flip a coin 25 times and get all heads. *Something funny is going on …*

- But with $F =$ {1000 heads, 1000 tails}, P(heads) on next flip $= 1025 / (1025+1000)$ $= 50.6\%$. *Looks like nothing unusual.*

- How do we balance stability and flexibility?
  - Stability: 6 heads, 4 tails $\longrightarrow$ $\theta \sim 0.5$
  - Flexibility: 25 heads, 0 tails $\longrightarrow$ $\theta \sim 1$

Bayesian model selection

$$d_1 \quad d_2 \quad d_3 \quad d_4 \qquad \text{vs.} \qquad \theta \rightarrow d_1 \quad d_2 \quad d_3 \quad d_4$$
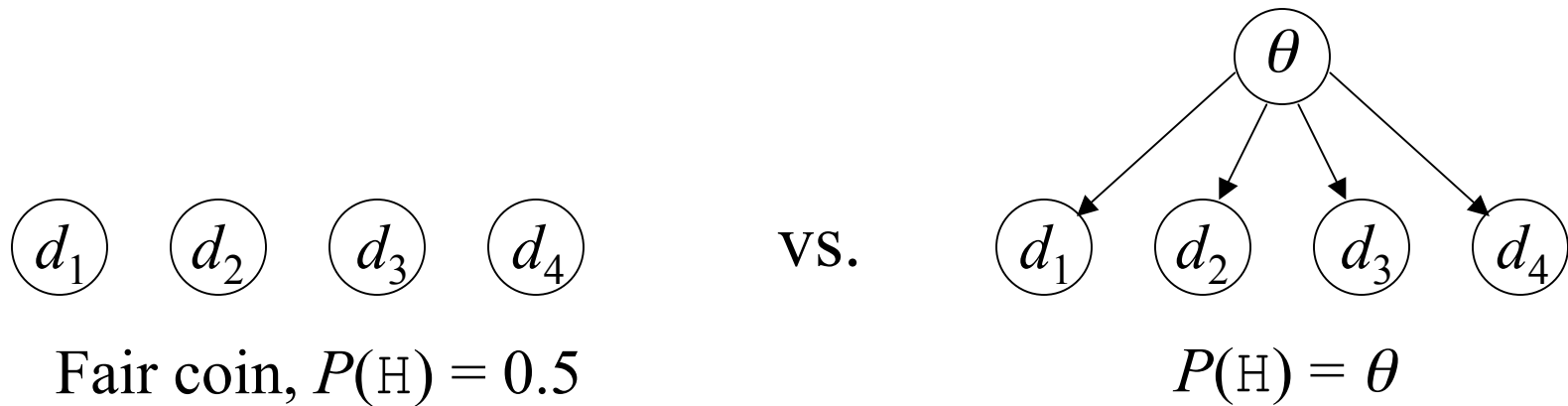
Fair coin, $P(\text{H}) = 0.5$ $\qquad\qquad\qquad\qquad$ $P(\text{H}) = \theta$

- Which provides a better account of the data: the simple hypothesis of a fair coin, or the complex hypothesis that $P(\text{H}) = \theta$?

Comparing simple and complex hypotheses: the need for Occam's razor

- $P(\text{H}) = \theta$ is more complex than $P(\text{H}) = 0.5$ in two ways:
    - $P(\text{H}) = 0.5$ is a special case of $P(\text{H}) = \theta$
    - for any observed sequence D, we can choose θ such that D is more probable than if $P(\text{H}) = 0.5$

Comparing simple and complex hypotheses: the need for Occam's razor
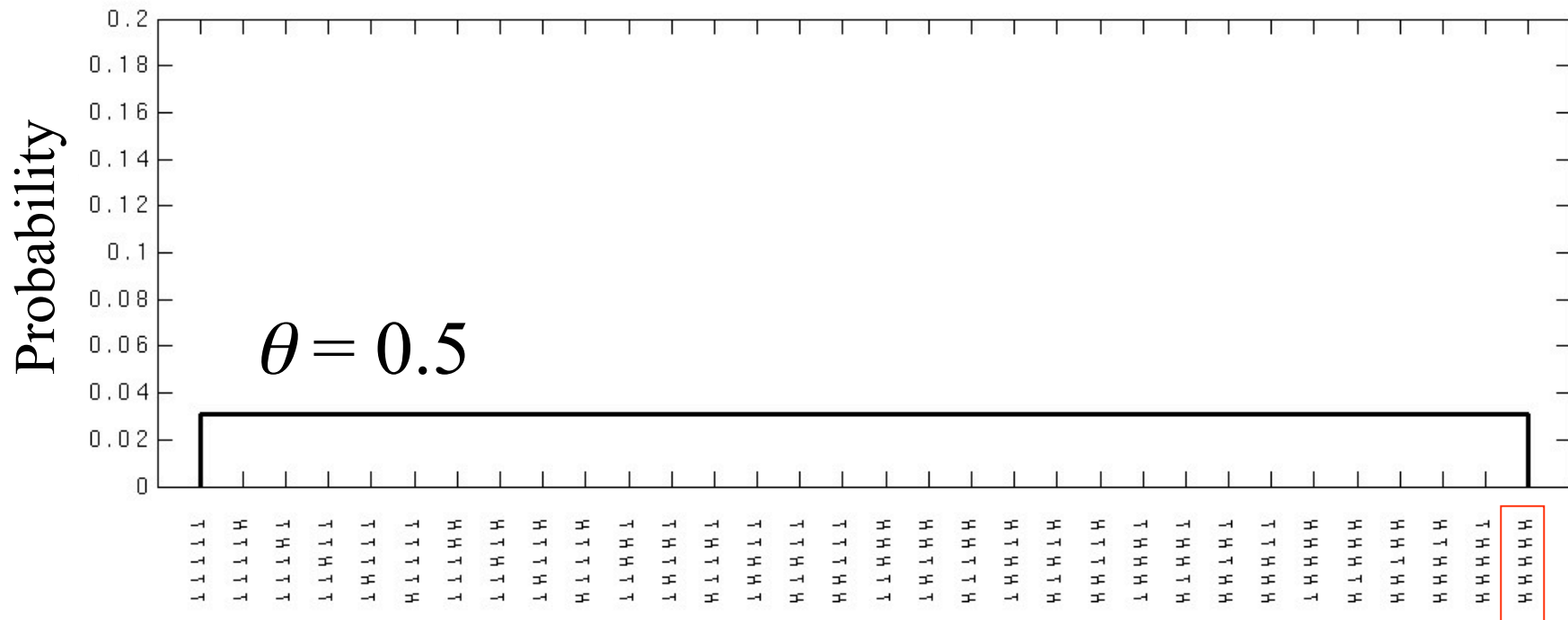
$$P(D \mid \theta) = \theta^n (1-\theta)^{N-n}$$

$\theta = 0.5$

$D = \text{HHHHH}$

Comparing simple and complex hypotheses: the need for Occam's razor

$$P(D \mid \theta) = \theta^n (1-\theta)^{N-n}$$

$\theta = 1.0$

$\theta = 0.5$
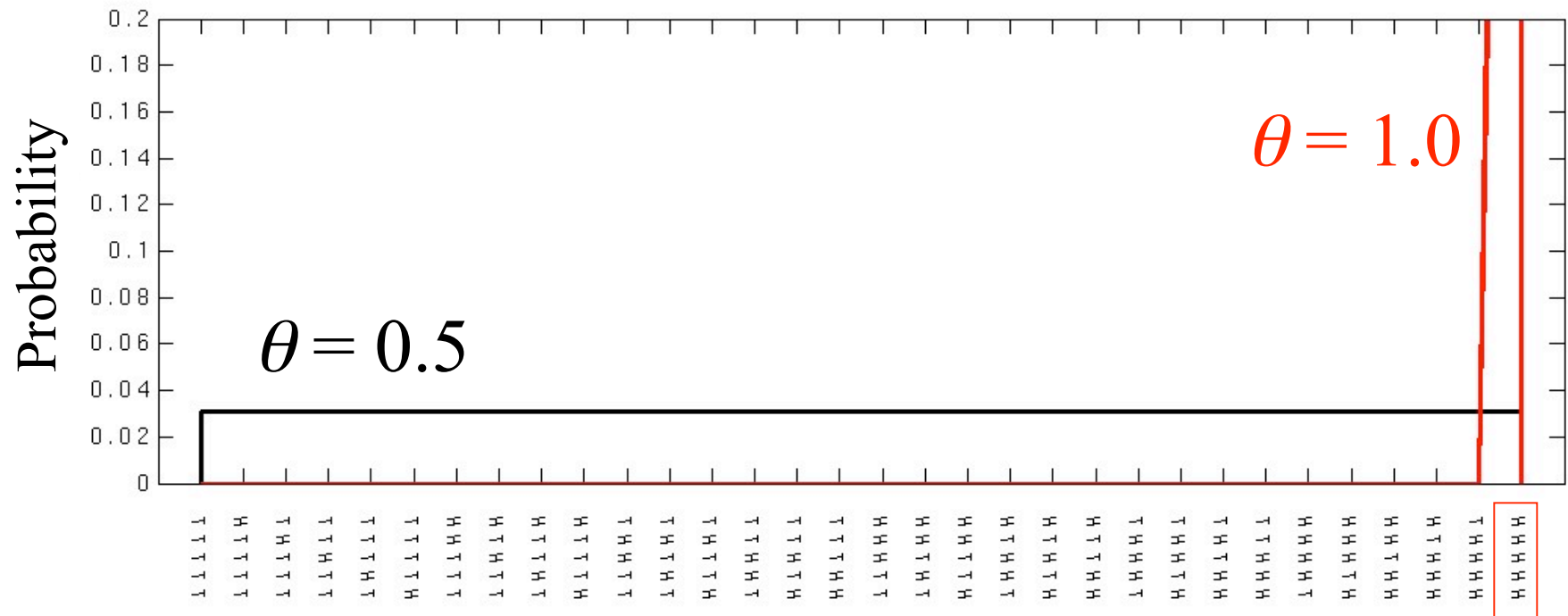
$D = \texttt{HHHHH}$

Comparing simple and complex hypotheses: the need for Occam's razor

$$P(D \mid \theta) = \theta^n (1-\theta)^{N-n}$$
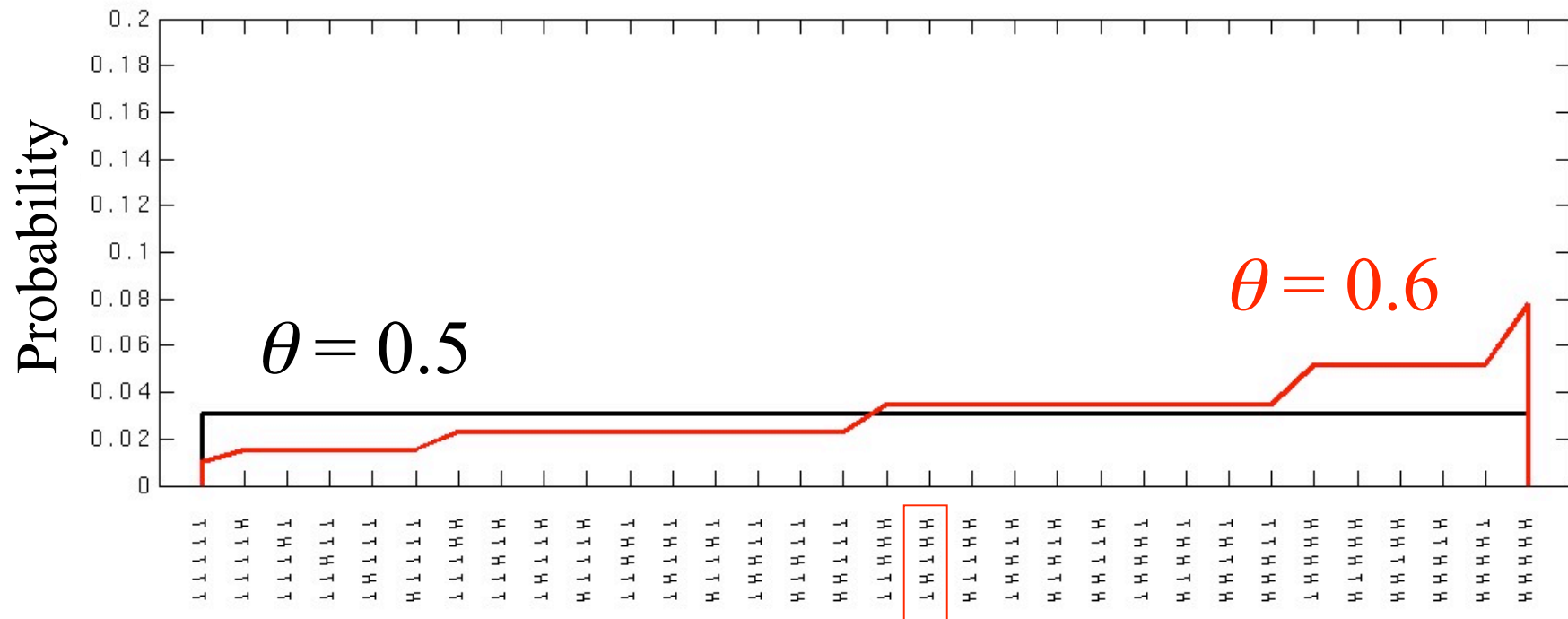
$D = \texttt{HHTHT}$

Comparing simple and complex hypotheses: the need for Occam's razor

- $P(\text{H}) = \theta$ is more complex than $P(\text{H}) = 0.5$ in two ways:
  - $P(\text{H}) = 0.5$ is a special case of $P(\text{H}) = \theta$
  - for any observed sequence X, we can choose θ such that X is more probable than if $P(\text{H}) = 0.5$
- How can we deal with this?
  - Some version of Occam's razor?
  - Bayes: automatic version of Occam's razor follows from the "law of conservation of belief".

Comparing simple and complex hypotheses: the need for Occam's razor

$$\frac{P(h_1|D)}{P(h_0|D)} = \frac{P(D|h_1)}{P(D|h_0)} \times \frac{P(h_1)}{P(h_0)}$$

$$P(D \mid h_0) = (1/2)^n (1-1/2)^{N-n} = 1/2^N$$

$$P(D \mid h_1) = \int_0^1 P(D \mid \theta, h_1) p(\theta \mid h_1) d\theta$$

The "evidence" or "marginal likelihood": The probability that *randomly selected* parameters from the prior would generate the data.

$$\log \frac{P(D \mid h_1)}{P(D \mid h_0)}$$

Number of heads (N_H)
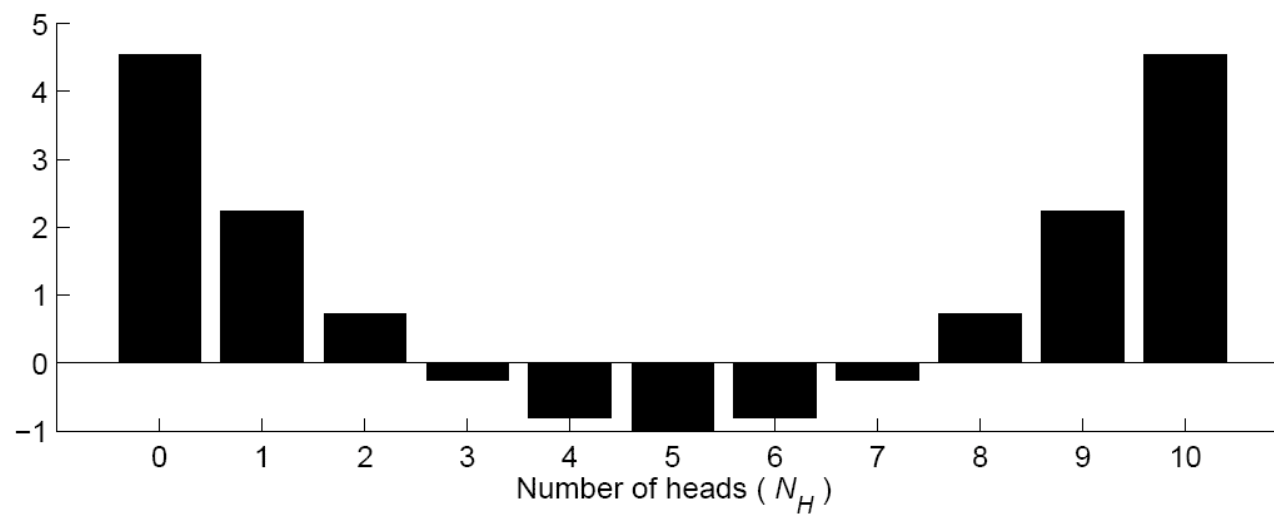
$$P(D \mid h_1) = \int_0^1 P(D \mid \theta, h_1) p(\theta \mid h_1) d\theta$$

$$P(D \mid h_0) = 1/2^N$$

P(HHTHTTHHHT | θ, h_1)

P(HHTHHHTHHH | θ, h_1)

P(d | h_0)

Probability

θ
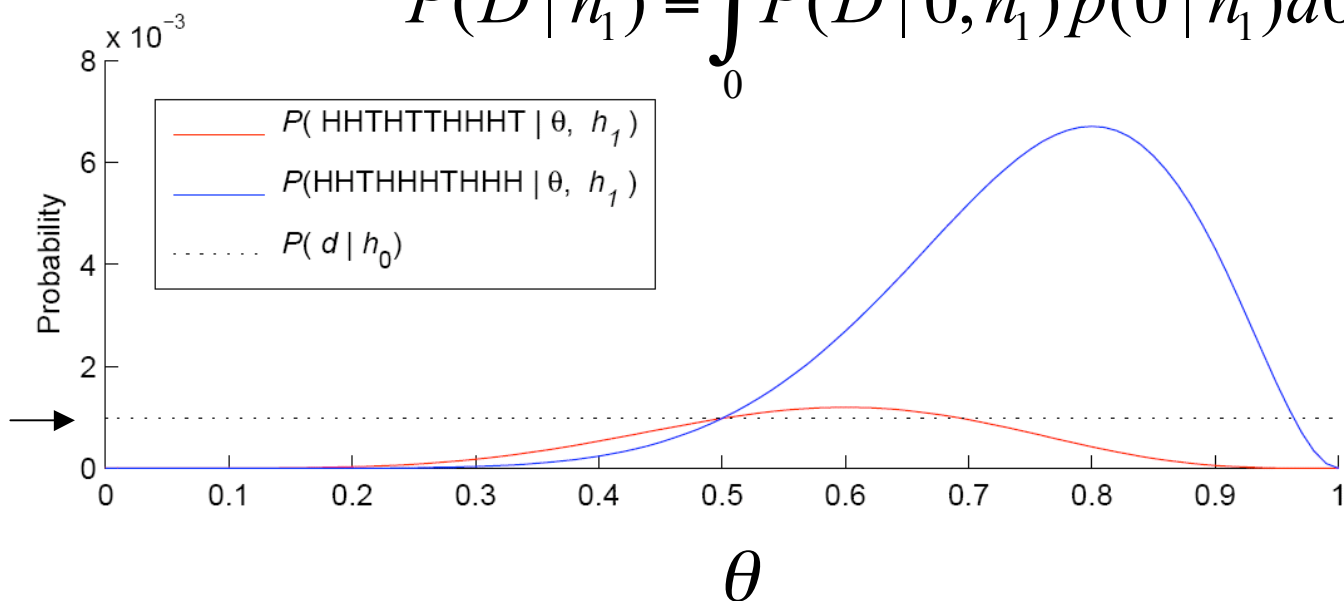
Stability versus Flexibility revisited

- Model class hypothesis: is this coin fair or unfair?

- Example probabilities:
  - $P(\text{fair}) = 0.999$
  - $P(\theta\,|\text{fair})$ is Beta(1000,1000)
  - $P(\theta\,|\text{unfair})$ is Beta(1,1)

- 25 heads in a row propagates up, affecting θ and then $P(\text{fair}|D)$

fair/unfair?

F_H, F_T

θ

d_1 d_2 d_3 d_4
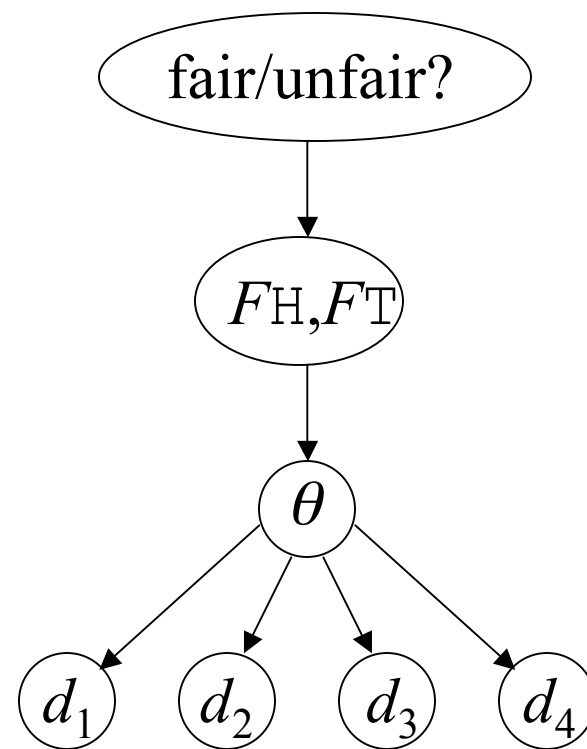
$$\frac{P(\text{fair}|25\text{ heads})}{P(\text{unfair}|25\text{ heads})} = \frac{P(25\text{ heads}|\text{fair})}{P(25\text{ heads}|\text{unfair})} \; \frac{P(\text{fair})}{P(\text{unfair})} \sim 0.001$$

Bayesian Occam's Razor
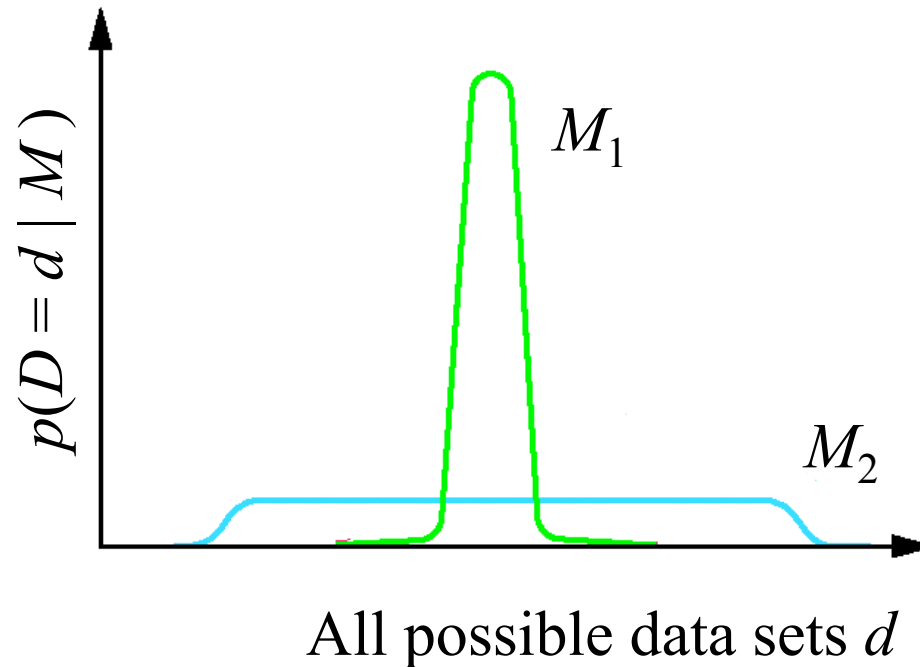
$$\text{For any model } M, \quad \sum_{\text{all } d \in D} p(D = d \mid M) = 1$$

Law of "conservation of belief": A model that can predict many possible data sets must assign each of them low probability.

Occam's Razor in curve fitting
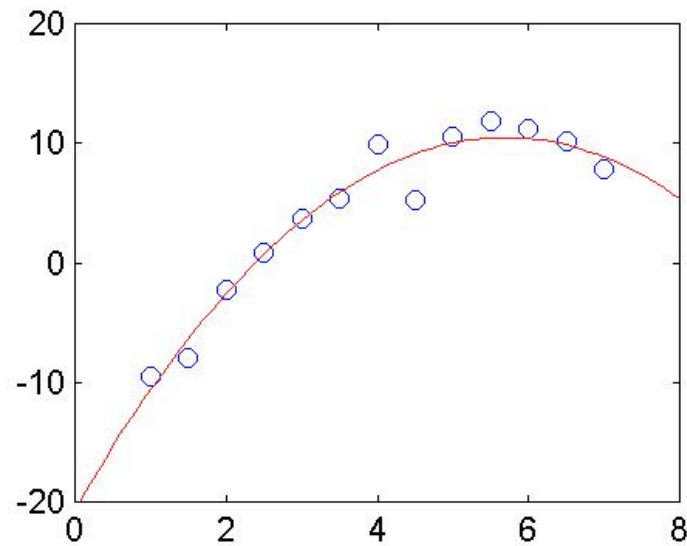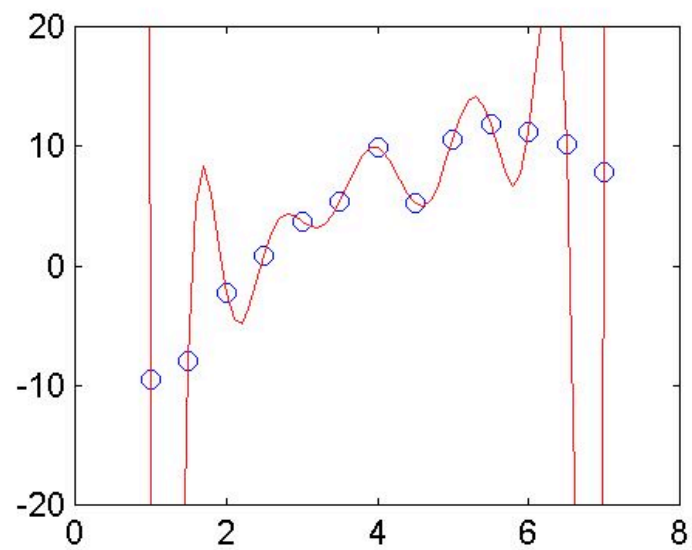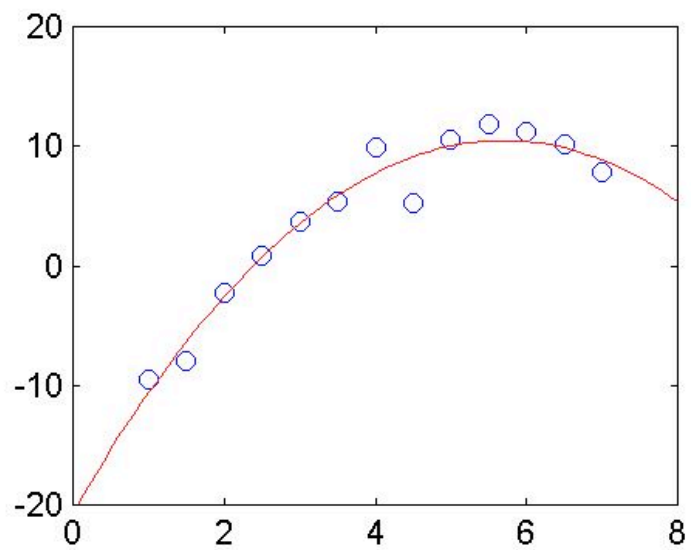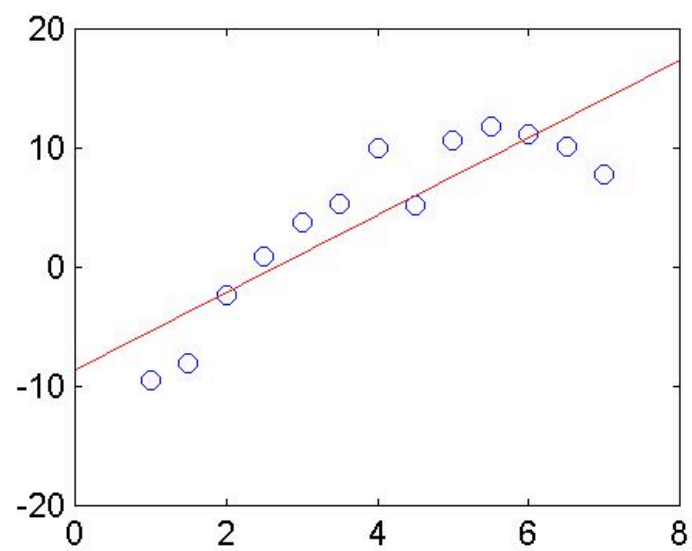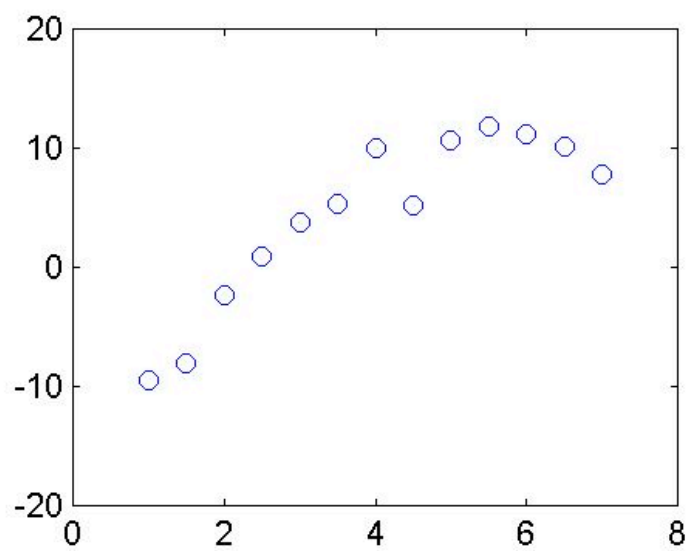
$$\sum_{\text{all } d \in D} p(D = d \mid M) = 1$$
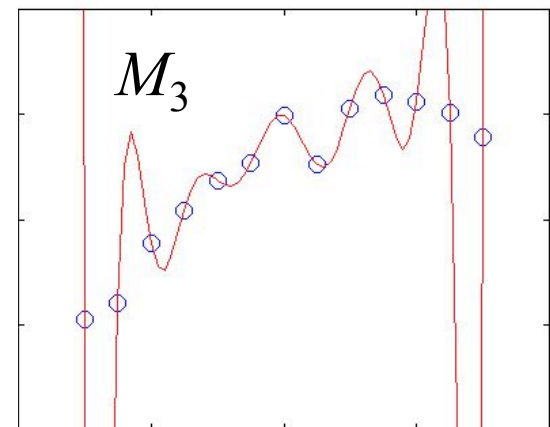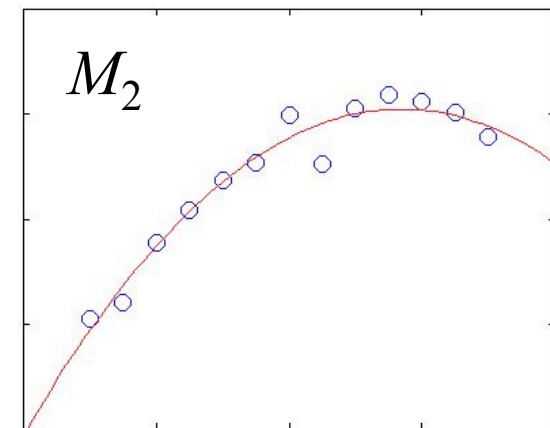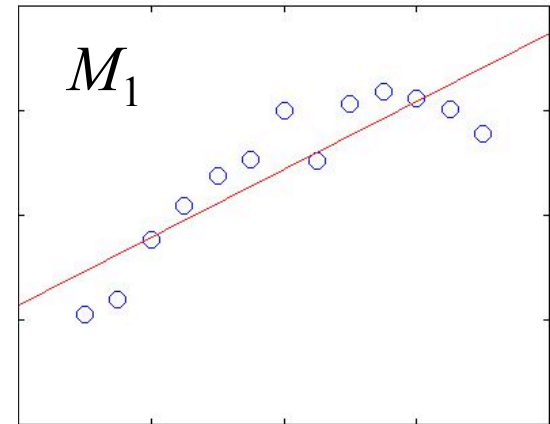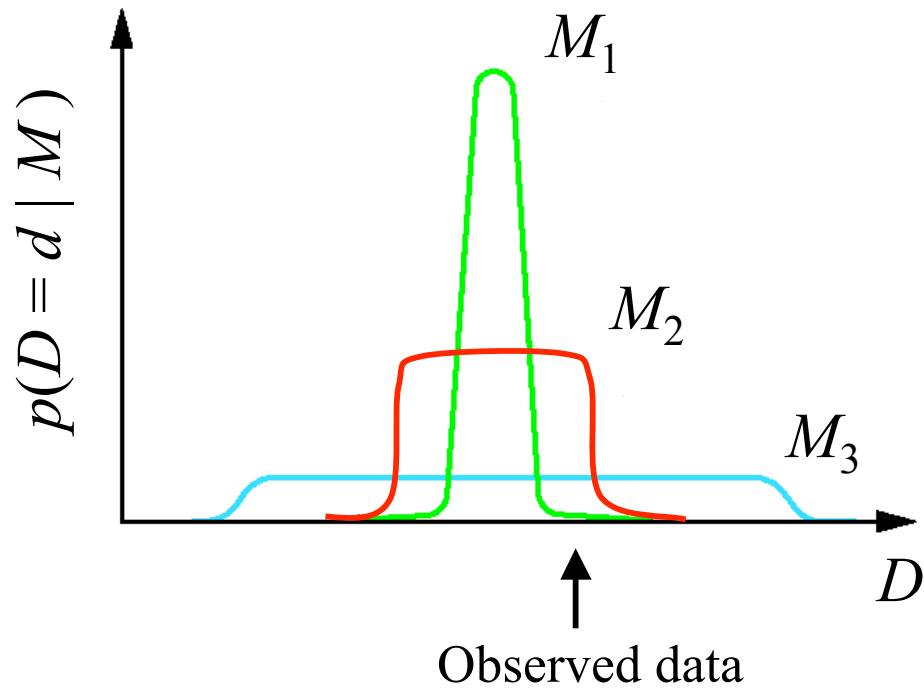
M_1: A model that is *too simple* is unlikely to generate the data.

M_3: A model that is *too complex* can generate many possible data sets, so it is unlikely to generate this particular data set at random.

Summary so far

- Three kinds of Bayesian inference
  - Comparing two simple hypotheses
  - Parameter estimation
    - The importance and subtlety of prior knowledge
  - Model selection
    - Bayesian Occam's razor, the blessing of abstraction
- Key concepts
  - Probabilistic generative models
  - Hierarchies of abstraction, with statistical inference at all levels
  - Flexibly structured representations

Plan for this lecture

- Some basic aspects of Bayesian statistics
  - Comparing two hypotheses
  - Model fitting
  - Model selection

- Two (very brief) case studies in modeling human inductive learning
  - Causal learning
  - Concept learning

Learning causation from correlation

	C present (c^+)	C absent (c^-)
E present (e^+)	a	c
E absent (e^-)	b	d

"Does C cause E?"
(rate on a scale from 0 to 100)

Learning with graphical models

- **Strength:** how strong is the relationship?

Delta-P, Power PC, ...

w_0 w_1
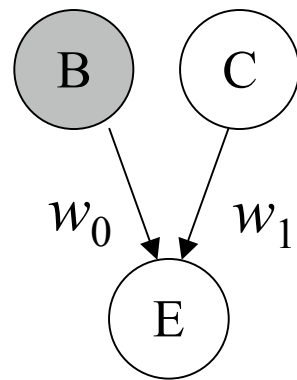
- **Structure:** does a relationship exist?

vs.

h_1 h_0
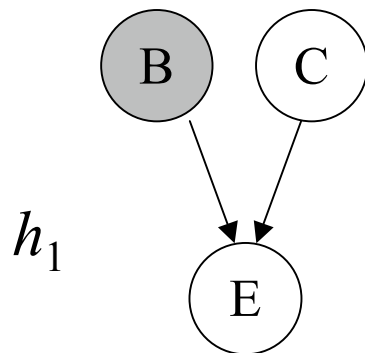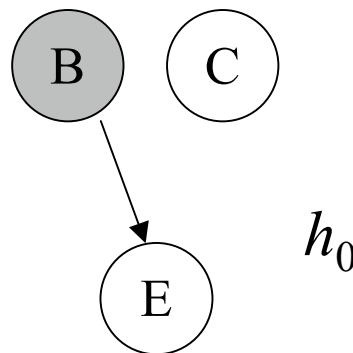
(Griffiths and Tenenbaum, *Cognitive Psychology* 2005)

Bayesian learning of causal structure

- Hypotheses:

h_1 vs. h_0
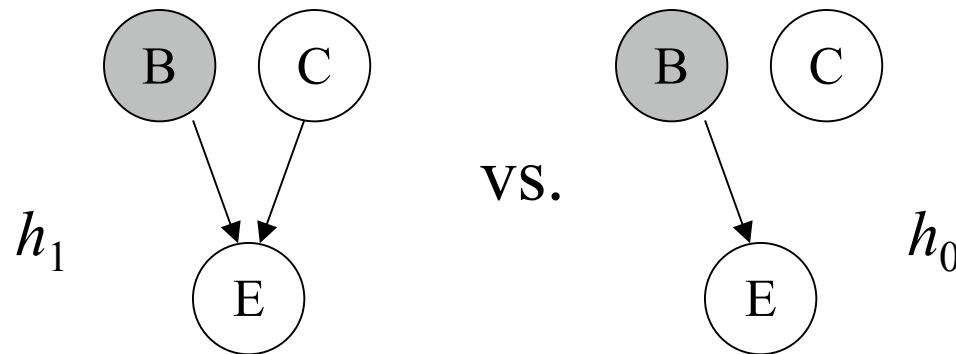
- Bayesian causal inference:

$$\text{support} = \log \frac{P(d|h_1)}{P(d|h_0)}$$

likelihood ratio (Bayes factor) gives evidence in favor of h_1

$$P(d \mid h_1) = \int_0^1 \int_0^1 P(d \mid w_0, w_1)\, p(w_0, w_1 \mid h_1)\, dw_0\, dw_1$$

$$P(d \mid h_0) = \int_0^1 P(d \mid w_0)\, p(w_0 \mid h_0)\, dw_0$$

Bayesian Occam's Razor

$P(d \mid h)$

h_0 (no relationship)

h_1 (positive relationship)

All data sets d
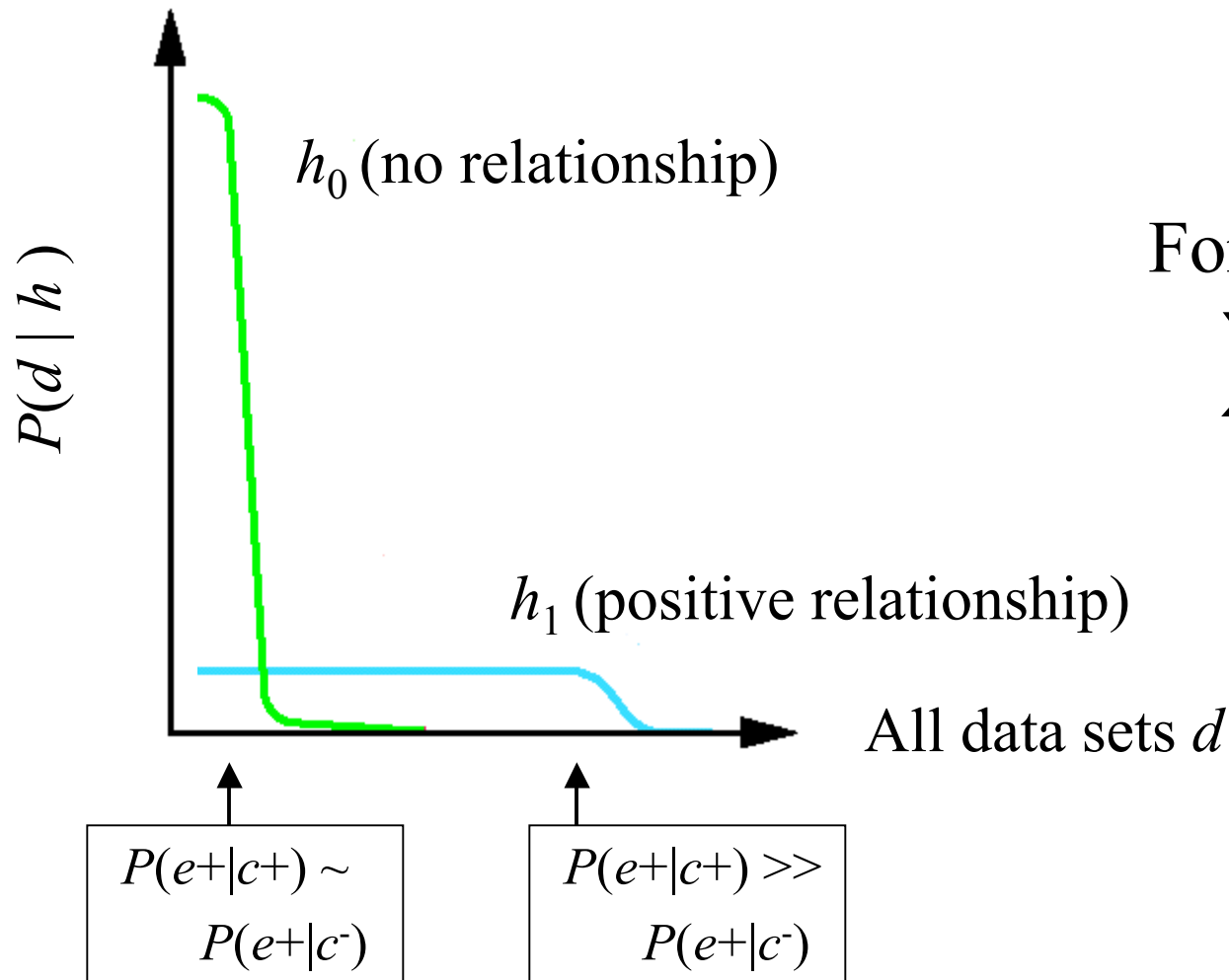
For any model h,

$$\sum_{d} P(d \mid h) = 1$$

$P(e+|c+) \sim P(e+|c^-)$

$P(e+|c+) \gg P(e+|c^-)$

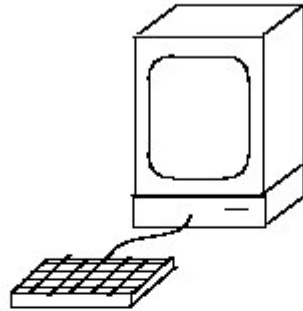# Comparison with human judgments

(Buehner & Cheng, 1997; 2003)
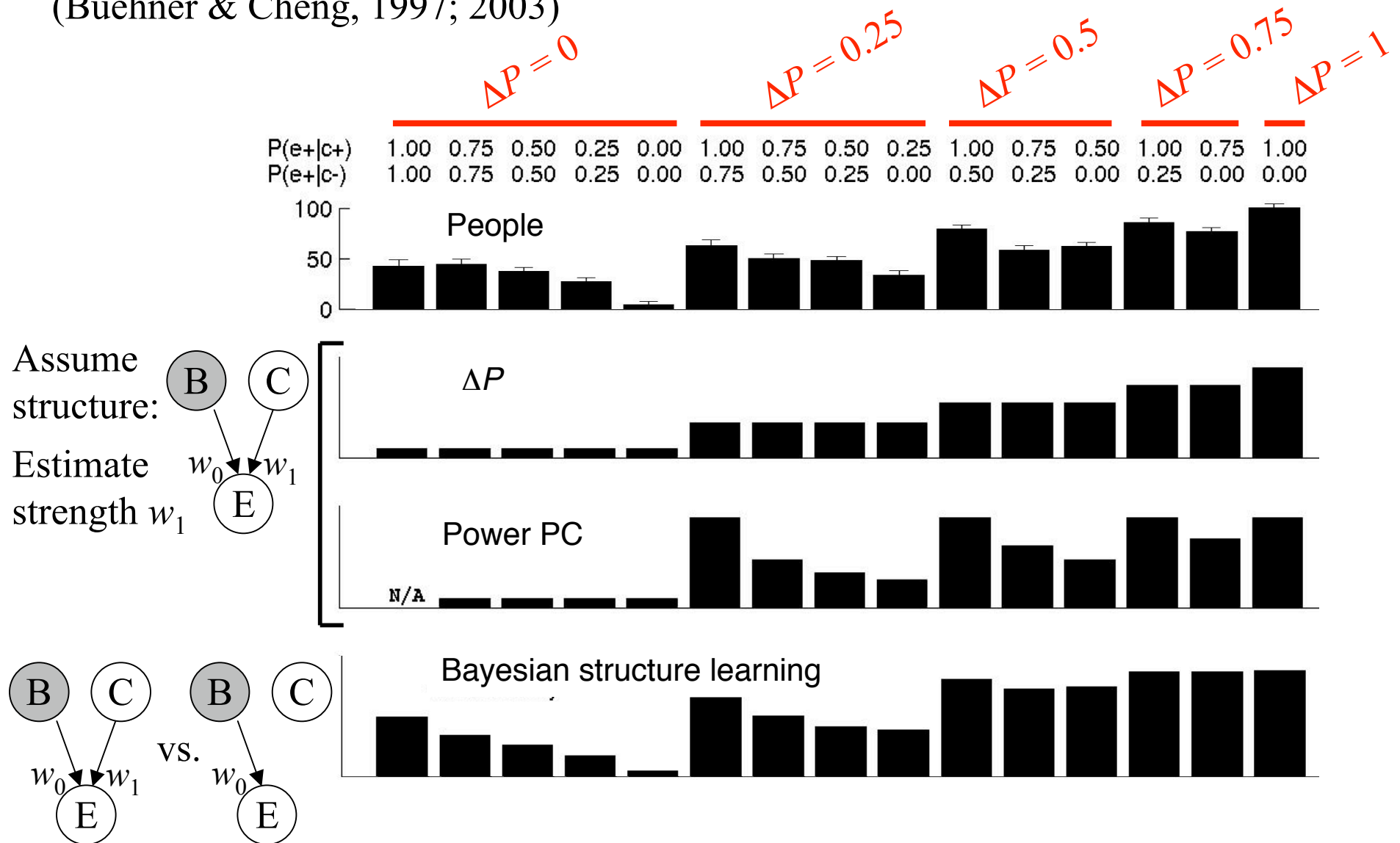
Inferences about causal structure depend on the functional form of causal relations
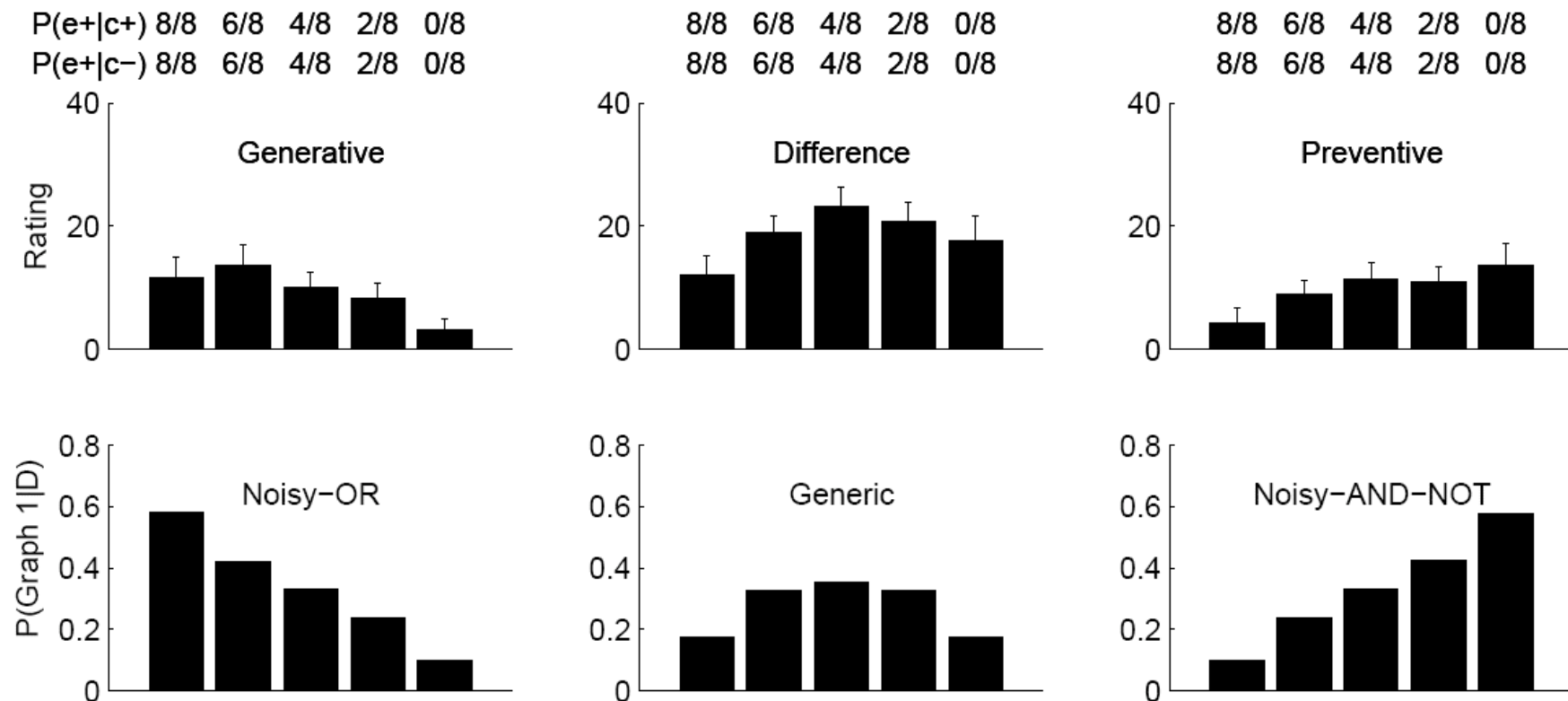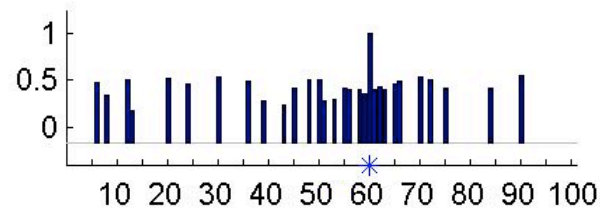
Concept learning: the number game

- Program input: number between 1 and 100

- Program output: "yes" or "no"

- Learning task:

  - Observe one or more positive ("yes") examples.
  - Judge whether other numbers are "yes" or "no".

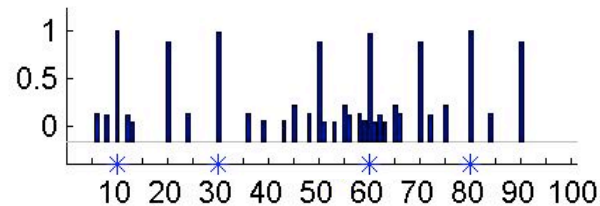# Concept learning: the number game

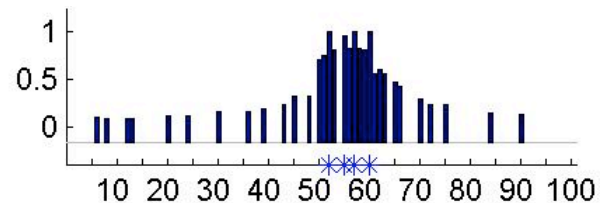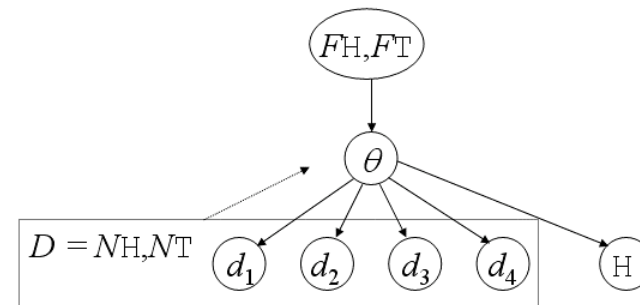| Examples of "yes" numbers	Generalization judgments ($N = 20$)	
60		Diffuse similarity
60 80 10 30		Rule: "multiples of 10"
60 52 57 55		Focused similarity: numbers near 50-60

Baye

- *H*: Hypothesis space of
  - *H*¹: Mathematical properties
  - *H*²: Magnitude: intervals wit

- $X = \{x_1, \ldots, x_n\}$: *n* ex

- *y*: some new object – is

Bayesian parameter estimation

$$P(\theta \mid D) \propto P(D \mid \theta)\, P(\theta) = \theta^{N\text{H}+F\text{H}-1}\,(1-\theta)^{N\text{T}+F\text{T}-1}$$

$$D = N\text{H}, N\text{T}$$

- Posterior predictive distribution:

$$P(\text{H} \mid D, F\text{H}, F\text{T}) = \int_0^1 P(\text{H} \mid \theta)\, P(\theta \mid D, F\text{H}, F\text{T})\, d\theta$$

$$= \frac{(N\text{H}+F\text{H})}{(N\text{H}+F\text{H}+N\text{T}+F\text{T})}$$

Background knowledge *K*

$$p(y \in C \mid X, K) = \sum_{h \in H} p(y \in C \mid h)\ p(h \mid X, K)$$

$X =$
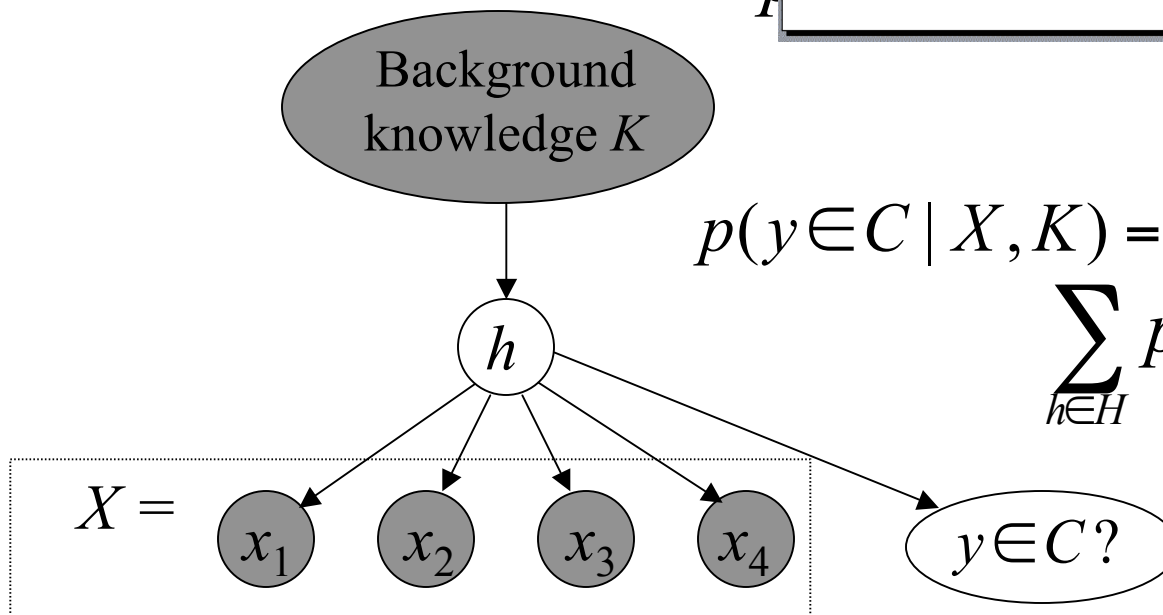
(Tenenbaum 2000; Tenenbaum and Griffiths, *BBS 2001*)

Likelihood: $p(X|h)$

- **Size principle**: Smaller hypotheses receive greater likelihood, and exponentially more so as n increases.

$$p(X \mid h) = \left[\frac{1}{\text{size}(h)} \right]^n \text{ if } x_1, \mathrm{K}, x_n \in h$$

$$= 0 \text{ if any } x_i \notin h$$

h_1 h_2

2	4	6	8	10
12	14	16	18	20
22	24	26	28	30
32	34	36	38	40
42	44	46	48	50
52	54	56	58	60
62	64	66	68	70
72	74	76	78	80
82	84	86	88	90
92	94	96	98	100

- Follows from assumption of randomly sampled examples + law of "conservation of belief":

$$\sum_{\text{all } d \in D} p(D = d \mid M) = 1$$

Likelihood: $p(X|h)$

- **Size principle**: Smaller hypotheses receive greater likelihood, and exponentially more so as n increases.

$$p(X \mid h) = \left[\frac{1}{\text{size}(h)} \right]^n \quad \text{if } x_1, \text{K}, x_n \in h$$

$$= 0 \quad \text{if any } x_i \notin h$$

- Follows from assumption of randomly sampled examples + law of "conservation of belief": $\sum_{\text{all } d \in D} p(D = d \mid M) = 1$

h_1

h_2

2	4	6	8	10
12	14	16	18	20
22	24	26	28	30
32	34	36	38	40
42	44	46	48	50
52	54	56	58	60
62	64	66	68	70
72	74	76	78	80
82	84	86	88	90
92	94	96	98	100

Likelihood: $p(X|h)$

- **Size principle**: Smaller hypotheses receive greater likelihood, and exponentially more so as n increases.

$$p(X \mid h) = \left[\frac{1}{\text{size}(h)} \right]^n \text{ if } x_1, \mathrm{K}, x_n \in h$$

$$= 0 \text{ if any } x_i \notin h$$

- Follows from assumption of randomly sampled examples + law of "conservation of belief":

$$\sum_{\text{all } d \in D} p(D = d \mid M) = 1$$

h_1　h_2

2	4	6	8	(10)
12	14	16	18	20
22	24	26	28	(30)
32	34	36	38	40
42	44	46	48	50
52	54	56	58	(60)
62	64	66	68	70
72	74	76	78	(80)
82	84	86	88	90
92	94	96	98	100

Prior: $p(h)$

- Choice of hypothesis space embodies a strong prior: effectively, $p(h) \sim 0$ for many logically possible but conceptually unnatural hypotheses.

- Prevents overfitting by highly specific but unnatural hypotheses, e.g. "multiples of 10 except 50 and 70".

e.g., $X = \{60\ 80\ 10\ 30\}$:

$$p(X \mid \text{multiples of } 10) = \left[\frac{1}{10}\right]^4 = 0.0001$$

$$p(X \mid \text{multiples of } 10 \text{ except } 50, 70) = \left[\frac{1}{8}\right]^4 = 0.00024$$

Prior: $p(h)$

- Choice of hypothesis space embodies a strong prior: effectively, $p(h) \sim 0$ for many logically possible but conceptually unnatural hypotheses.

- Prevents overfitting by highly specific but unnatural hypotheses, e.g. "multiples of 10 except 50 and 70".

- $p(h)$ encodes relative weights of alternative theories:

H: Total hypothesis space

$p(H^1) = \lambda$ $p(H^2) = 1-\lambda$
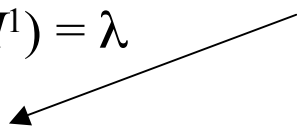
H^1: Mathematical properties (24) H^2: Magnitude intervals (5050)
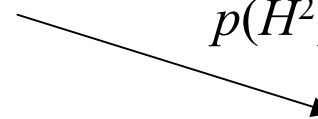
- even numbers   - 10-15
- powers of two   - 20-32
- multiples of three   - 37-54

... $p(h) = \lambda / 24$ ... $p(h) = 1-\lambda / 5050 * Gamma(s; \sigma)$
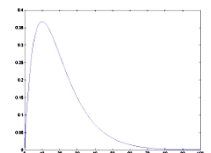
Posterior: $p(h \mid X) = \dfrac{p(X \mid h)p(h)}{\displaystyle\sum_{h' \in H} p(X \mid h')p(h')}$

- $X = \{60, 80, 10, 30\}$

- Why prefer "multiples of 10" over "even numbers"? $p(X|h)$.

- Why prefer "multiples of 10" over "multiples of 10 except 50 and 20"? $p(h)$.

- Why does a good generalization need both high prior and high likelihood? $p(h|X) \sim p(X|h)\,p(h)$
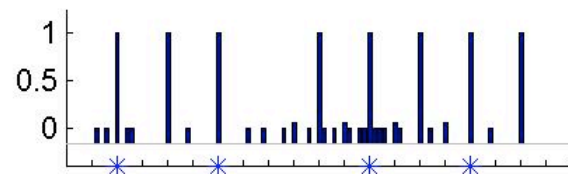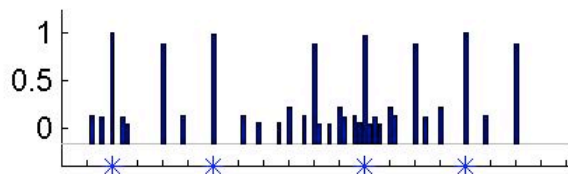
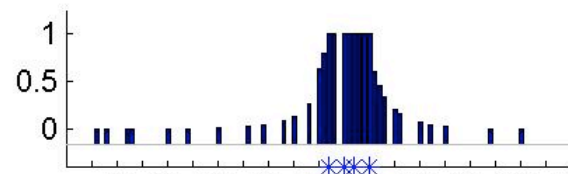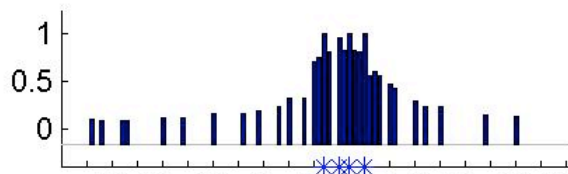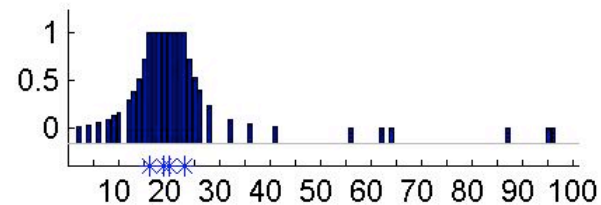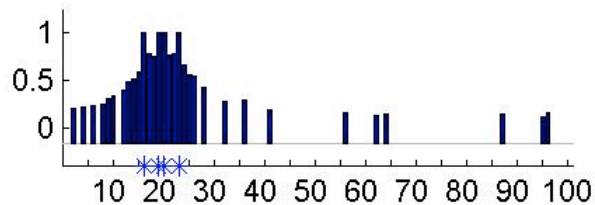| + Examples	Human generalization	Bayesian Model
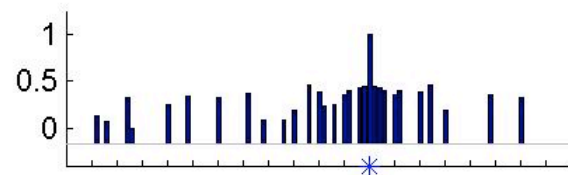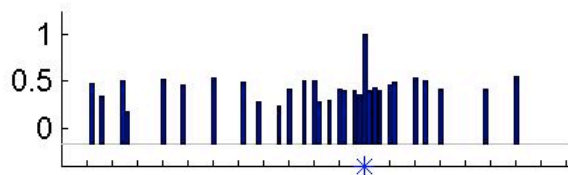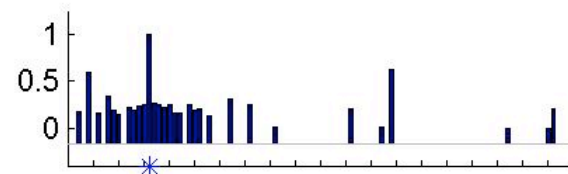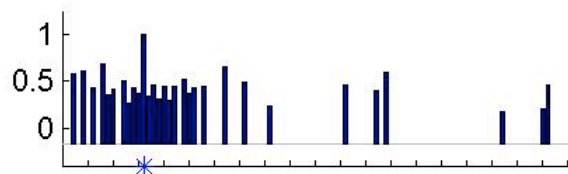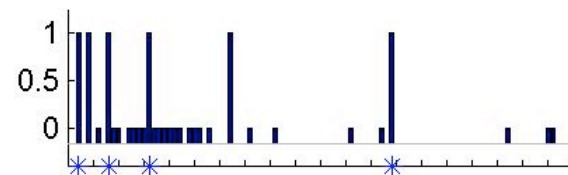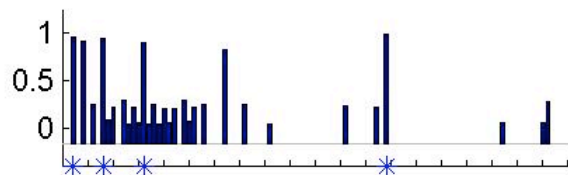
Stability versus Flexibility

- Higher-level hypothesis: is this concept mathematical or magnitude-based?

- Example probabilities:
  - $P(\text{math}) = \lambda$
  - $P(h \mid \text{math})$ …
  - $P(h \mid \text{magnitude})$ …

math/magnitude?

h

$X =$ x_1 x_2 x_3 x_4

- Just a few examples may be sufficient to infer the kind of concept, under the size-principle likelihood
  - if an *a priori* reasonable hypothesis of one kind fits much more tightly than all reasonable hypothesis of the other kind.

- Just a few examples can give all-or-none, "rule-like" generalization or more graded, "similarity-like" generalization.
  - More all-or-none when the smallest consistent hypothesis is much smaller than all other reasonable hypotheses; otherwise more graded.
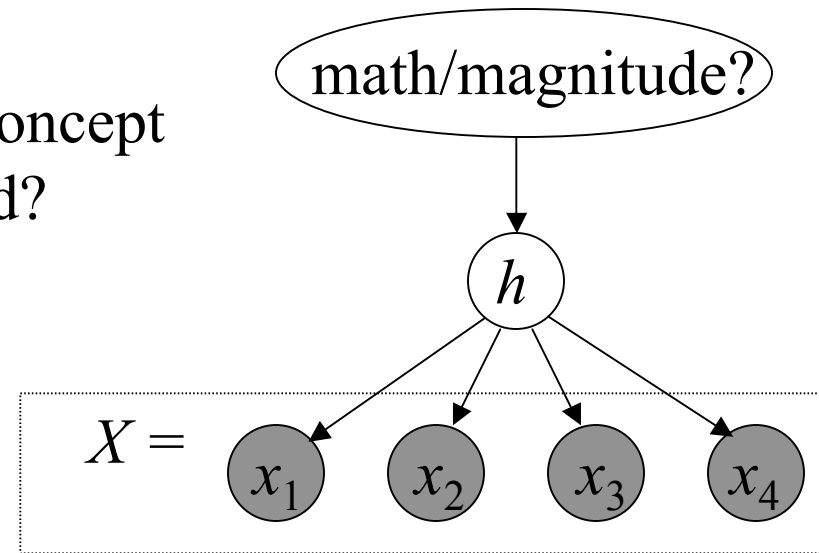
Conclusion:
Contributions of Bayesian models

- A framework for understanding how the mind can solve fundamental problems of induction.

- Strong, principled quantitative models of human cognition.

- Tools for studying people's implicit knowledge of the world.

- Beyond classic limiting dichotomies: "rules vs. statistics", "nature vs. nurture", "domain-general vs. domain-specific" .

- A unifying mathematical language for all of the cognitive sciences: AI, machine learning and statistics, psychology, neuroscience, philosophy, linguistics…. A bridge between engineering and "reverse-engineering".

Resources

- Tutorials chapters and articles:
  - "Bayesian models of cognition" chapter in Ron Sun (ed.), *Cambridge Handbook of Computational Cognitive Modeling.* Cambridge Univ Press.
  - "A tutorial introduction to Bayesian models of cognitive development", by Amy Perfors, JBT, TLG, and Fei Xu.
- Tom's Bayesian reading list:
  - http://cocosci.berkeley.edu/tom/bayes.html
  - tutorial slides will be posted there!
- *Trends in Cognitive Sciences* special issue on probabilistic models of cognition (July 2006, volume 10, issue 7)
- IPAM graduate summer school on probabilistic models of cognition (2007 videos, 2011 forthcoming)
- MLSS 2010 (Machine Learning and Cognitive Science) Sardinia, VideoLectures