CS340

Bayesian concept learning cont'd

Kevin Murphy

Bayesian inference

- H: Hypothesis space of possible concepts:
- $X = \{x_1, \ldots, x_n\}$: n examples of a concept C.
- Evaluate hypotheses given data using Bayes' rule:

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

  - $p(h)$ ["prior"]: domain knowledge, pre-existing biases
  - $p(X|h)$ ["likelihood"]: statistical information in examples.
  - $p(h|X)$ ["posterior"]: degree of belief that h is the true extension of C.

Hypothesis space

- Mathematical properties (~50):
  - odd, even, square, cube, prime, …
  - multiples of small integers
  - powers of small integers
  - same first (or last) digit
- Magnitude intervals (~5000):
  - all intervals of integers with endpoints between 1 and 100
- Hypothesis can be defined by its **extension**

$$h = \{x : h(x) = 1,\ x = 1, 2, \ldots, 100\}$$

Likelihood p(X|h)

- Assume samples are iid, so $p(X|h) = \prod_{i=1}^{n} p(x_i|h)$
- **Size principle**: Smaller hypotheses receive greater likelihood, and exponentially more so as *n* increases.

$$p(X|h) = \begin{cases} \frac{1}{|size(h)|^n} & \text{if all } x_1, \ldots, x_n \in h \\ 0 & \text{if any } x_i \notin h \end{cases}$$

- This is the likelihood of the *ordered sequence* $x_1, \ldots, x_n$ sampled randomly (with replacement) from h (**strong sampling assumption**).
- Captures the intuition of a representative sample.

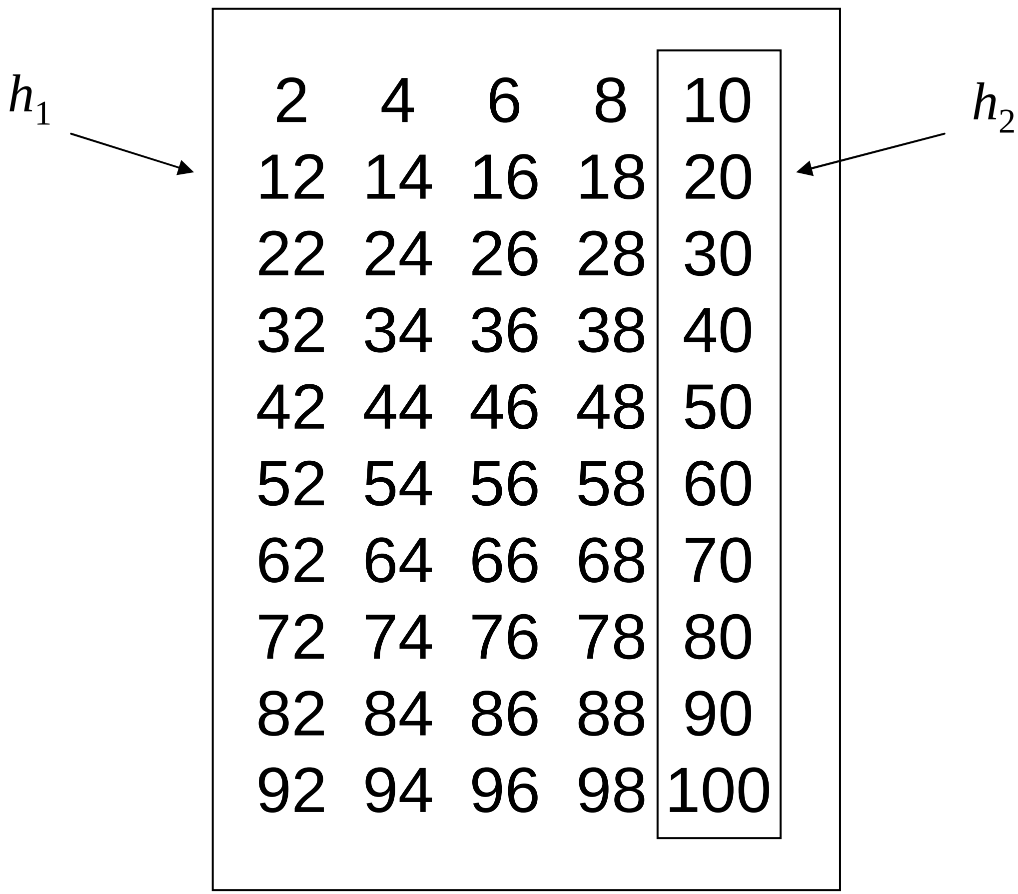
Likelihood function

- Since $p(\vec{x}|h)$ is a distribution over vectors of length n, we require that, for all h, $\sum_{\vec{x}} p(x|h) = 1$
- It is easy to see this is true, e.g., for h=even numbers, n=2

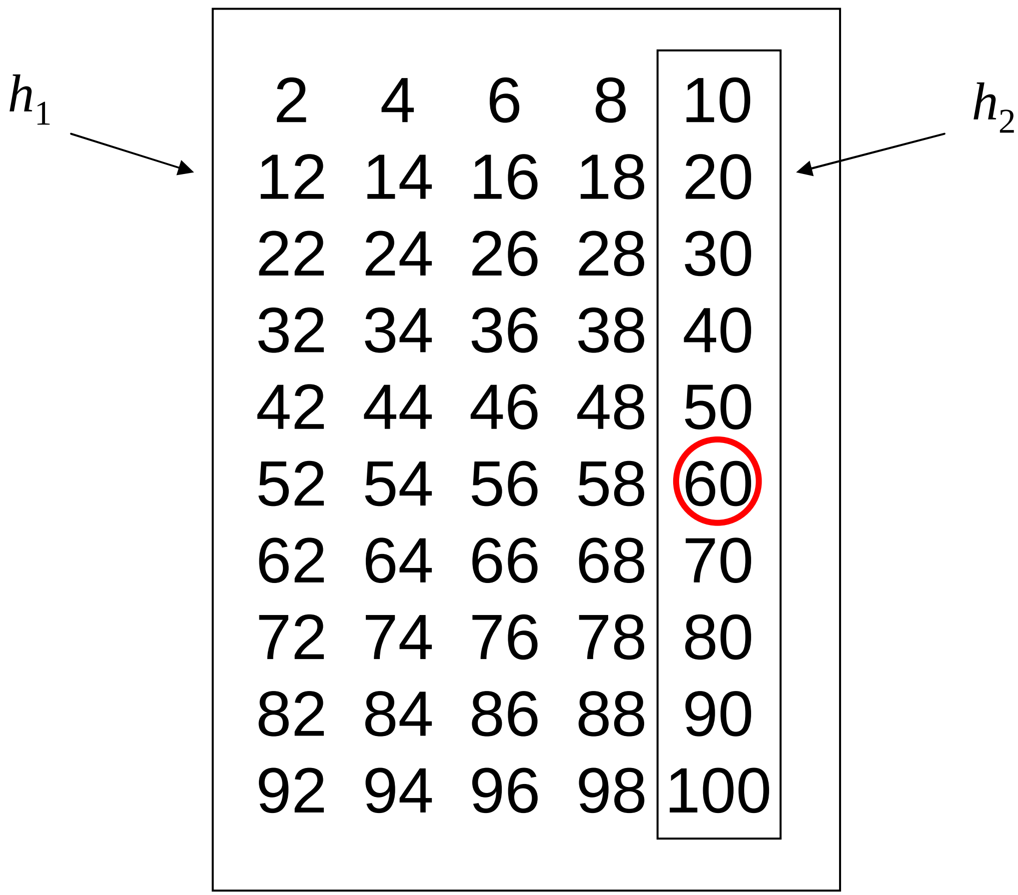
$$\sum_{x_1=1}^{100} \sum_{x_2=1}^{100} p(x_1, x_2|h) = \sum_{x_1=1}^{100} \sum_{x_2=1}^{100} p(x_1|h)p(x_2|h) = \sum_{x_1 \in even} \sum_{x_2 \in even} \frac{1}{50}\frac{1}{50} = 1$$

- If x is fixed, we do not require $\sum_{h} p(X|h) = 1$
- Hence we are free to multiply the likelihood by any constant independent of h

Illustrating the size principle

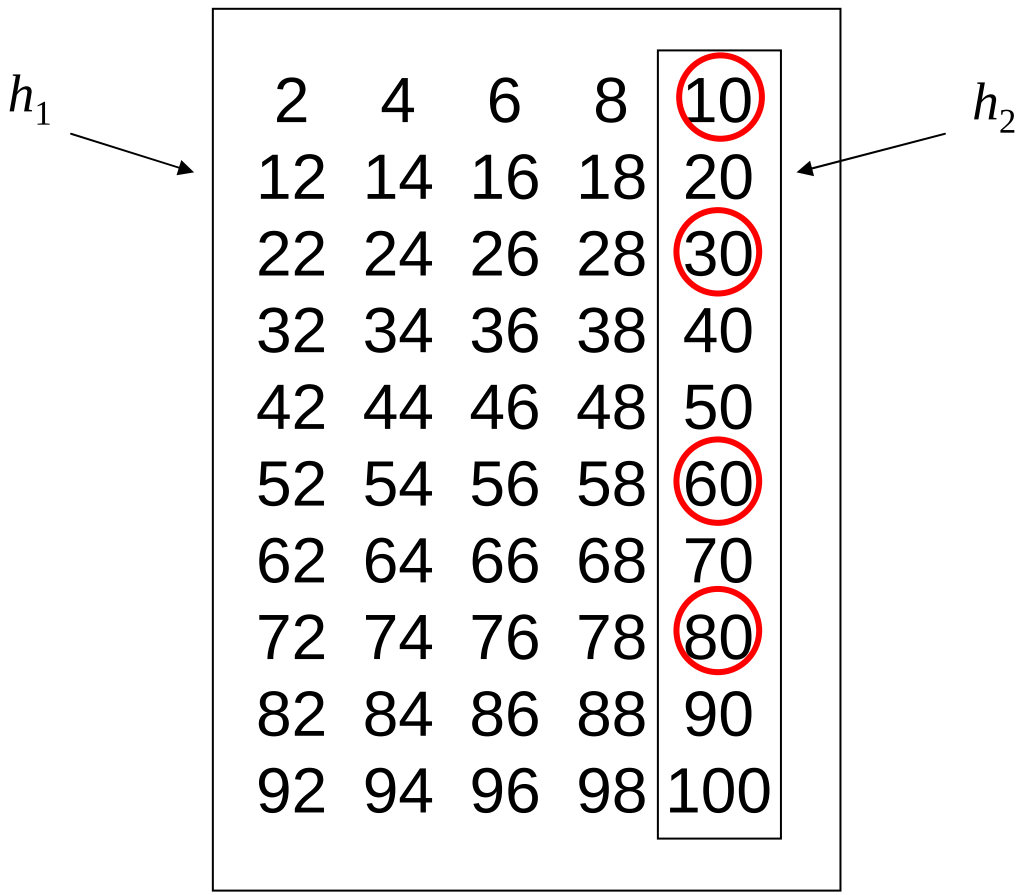


Illustrating the size principle



Data slightly more of a coincidence under h_1

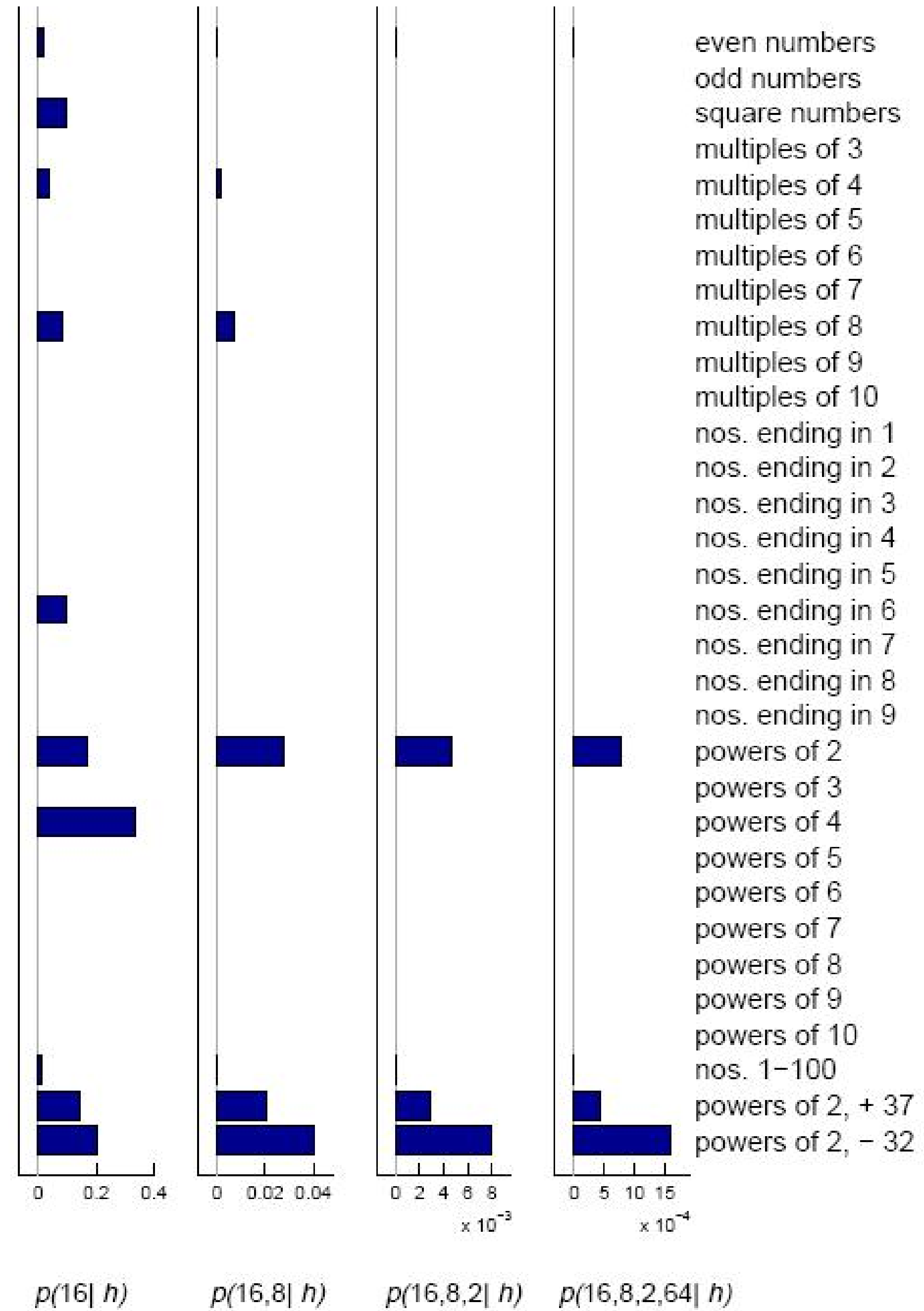
Illustrating the size principle



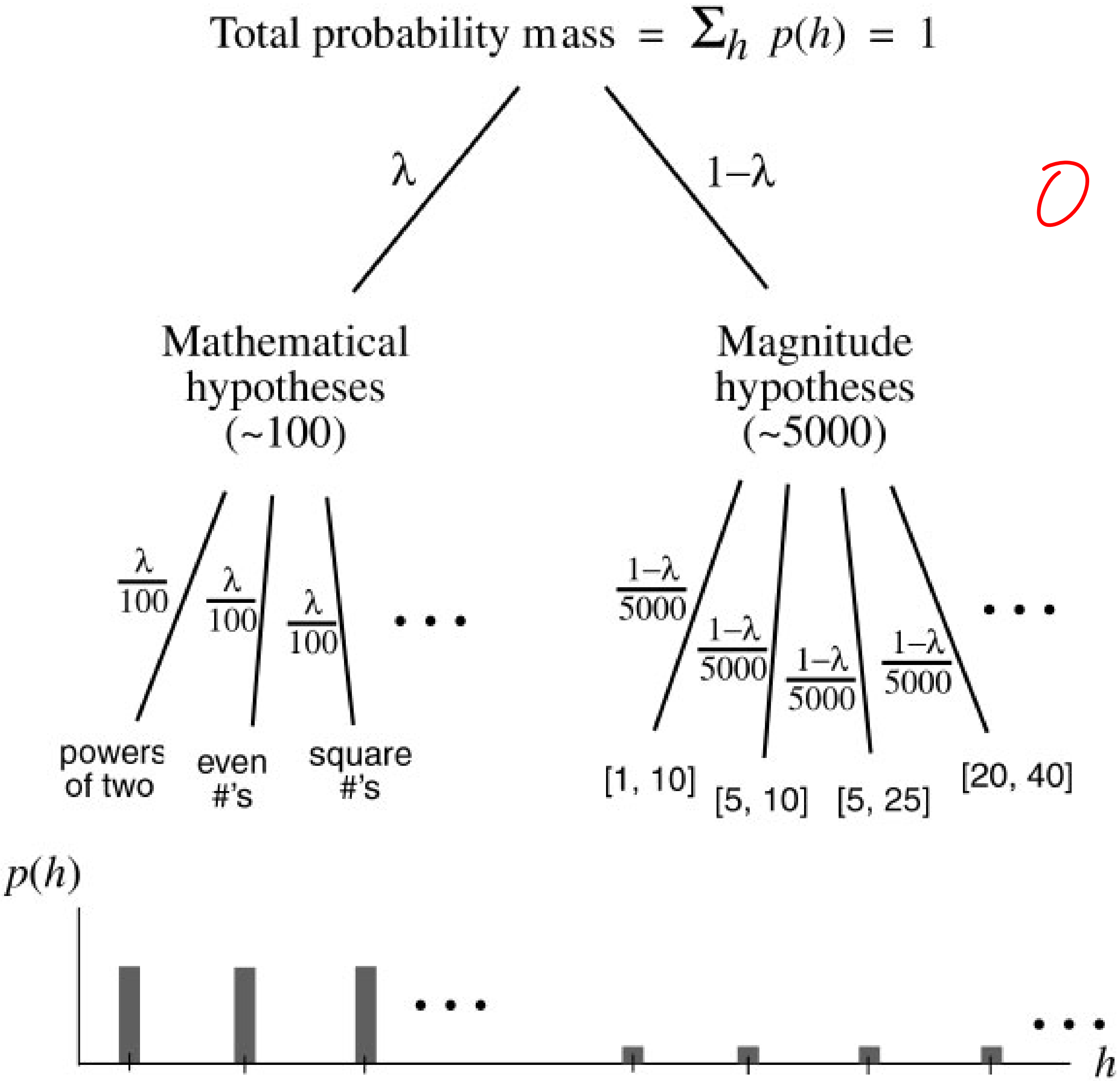
Data *much* more of a coincidence under h_1

Example of likelihood

- X={20,40,60}
- H1 = multiples of 10 = {10,20,…,100}
- H2 = even numbers = {2,4,…,100}
- H3 = odd numbers = {1,3,…,99}
- P(X|H1) = 1/10 * 1/10 * 1/10
- p(X|H2) = 1/50 * 1/50 * 1/50
- P(X|H3)   = 0


even numbers
odd numbers
square numbers
multiples of 3
multiples of 4
multiples of 5
multiples of 6
multiples of 7
multiples of 8
multiples of 9
multiples of 10
nos. ending in 1
nos. ending in 2
nos. ending in 3
nos. ending in 4
nos. ending in 5
nos. ending in 6
nos. ending in 7
nos. ending in 8
nos. ending in 9
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
powers of 10
nos. 1−100
powers of 2, + 37
powers of 2, − 32
0 0.2 0.4
0 0.02 0.04
0 2 4 6 8
x 10^-3
0 5 10 15
x 10^-4
p(16| h)
p(16,8| h)
p(16,8,2| h)
p(16,8,2,64| h)

Hierarchical prior

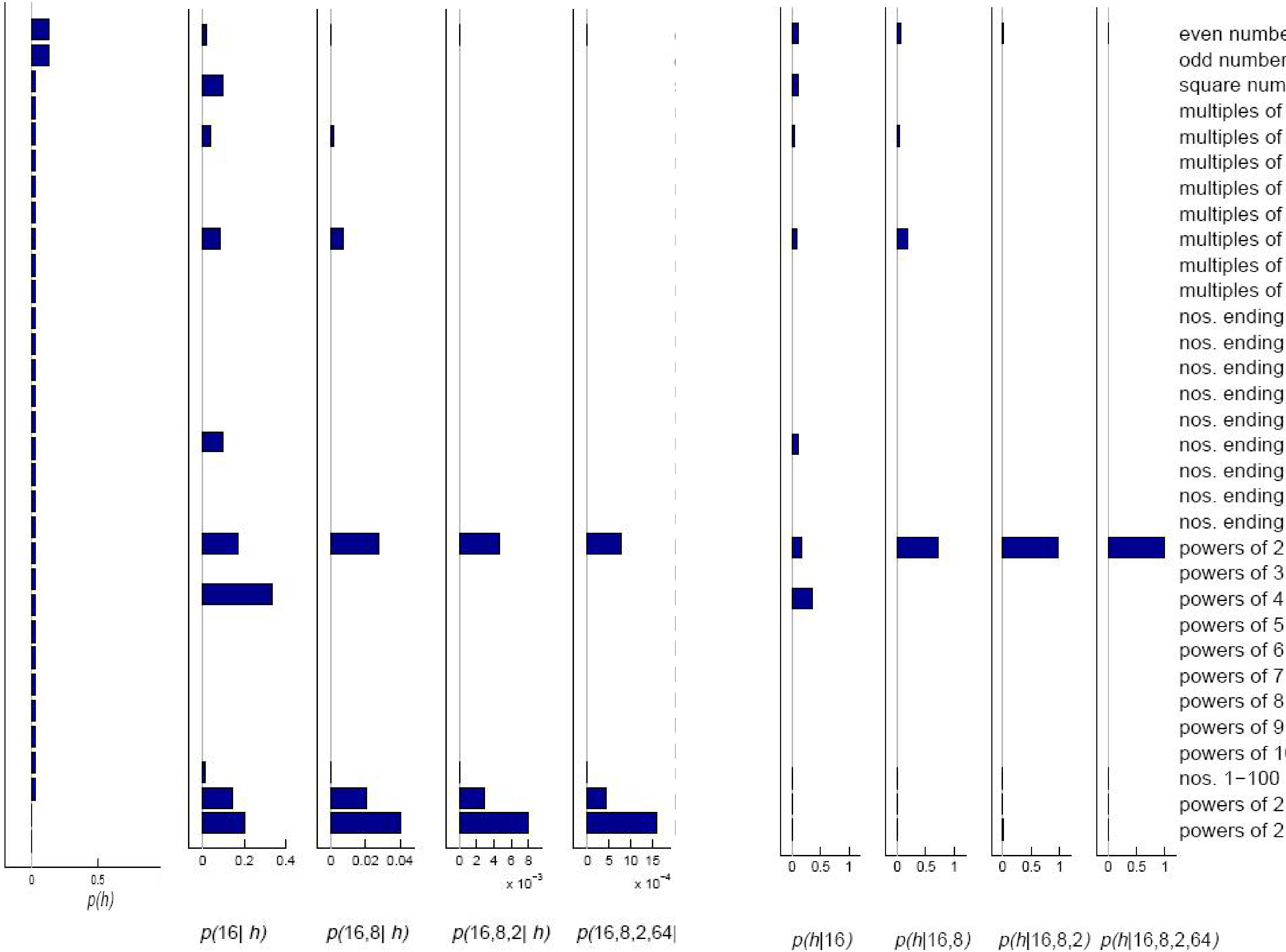


$0 < \lambda < 1$

Computing the posterior

- In this talk, we will not worry about computational issues (we will perform brute force enumeration or derive analytical expressions).

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



p(h)
p(16| h)
p(16,8| h)
p(16,8,2| h)
p(16,8,2,64|
p(h|16)
p(h|16,8)
p(h|16,8,2)
p(h|16,8,2,64)
x 10^-3
x 10^-4
even numbe
odd number
square num
multiples of
nos. ending
powers of 2
powers of 3
powers of 4
powers of 5
powers of 6
powers of 7
powers of 8
powers of 9
nos. 1−100

Generalizing to new objects

Given $p(h|X)$, how do we compute $p(y \in C \mid X)$, the probability that C applies to some new stimulus y?

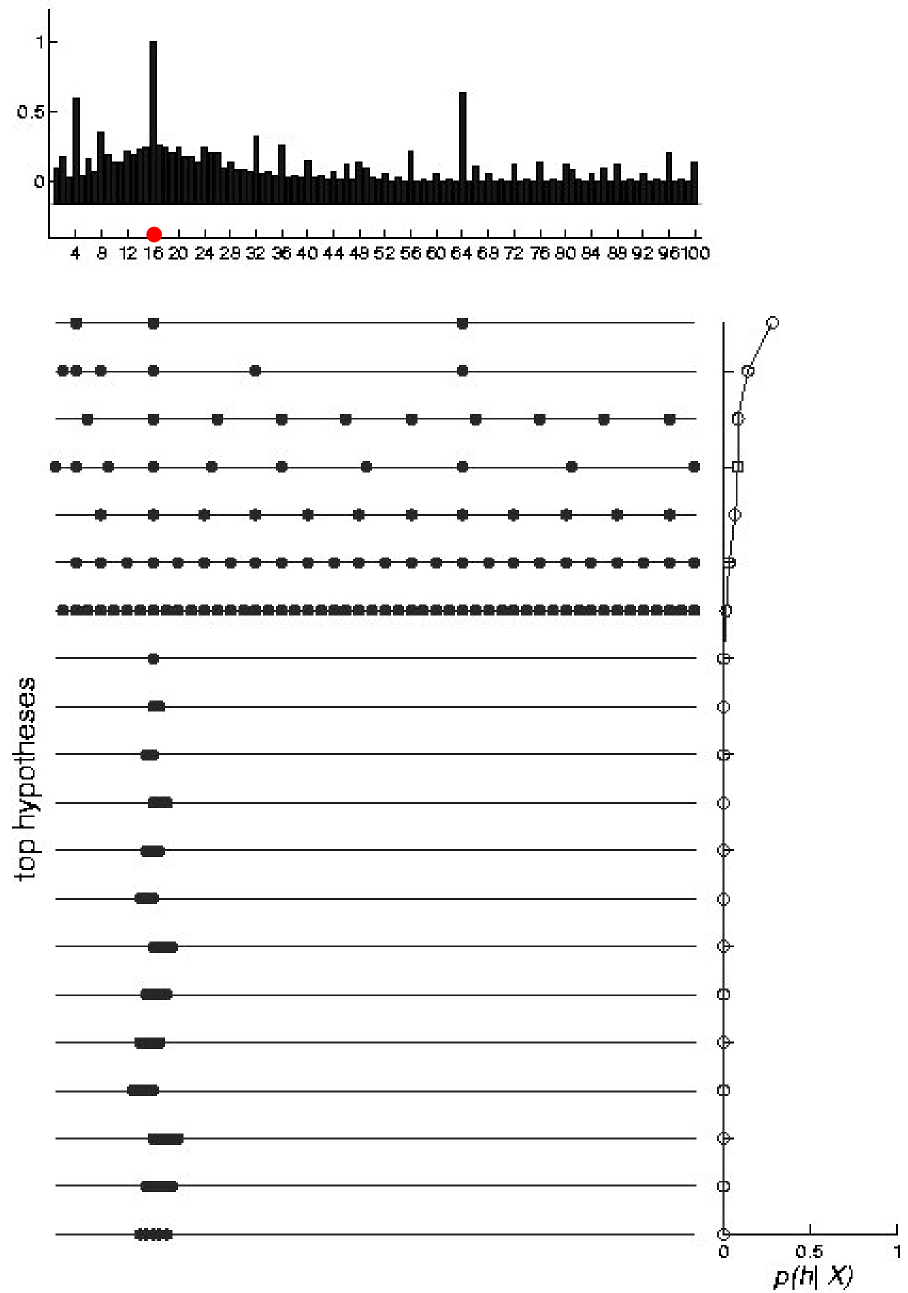
Posterior predictive distribution

Compute the probability that C applies to some new object y by averaging the predictions of all hypotheses h, weighted by $p(h|X)$ (**Bayesian model averaging**):

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

$$= \sum_{h \supset \{y, X\}} p(h \mid X)$$

Examples:
16



1
0.5
0
4 8 12 16 20 24 28 32 36 40 44 48 52 56 60 64 68 72 76 80 84 88 92 96 100
top hypotheses
0 0.5 1
p(h| X)

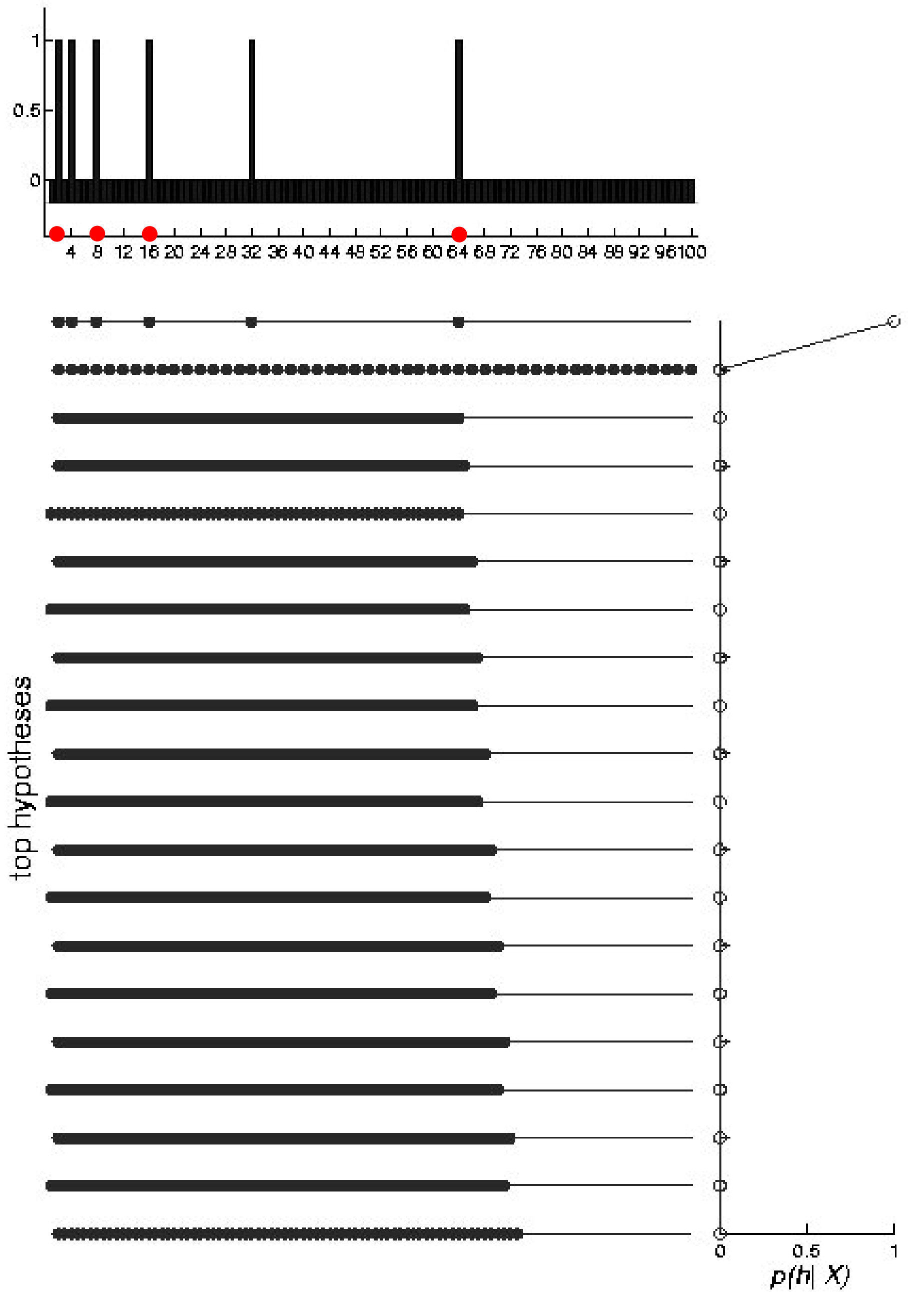
Examples:

16

8

2

64



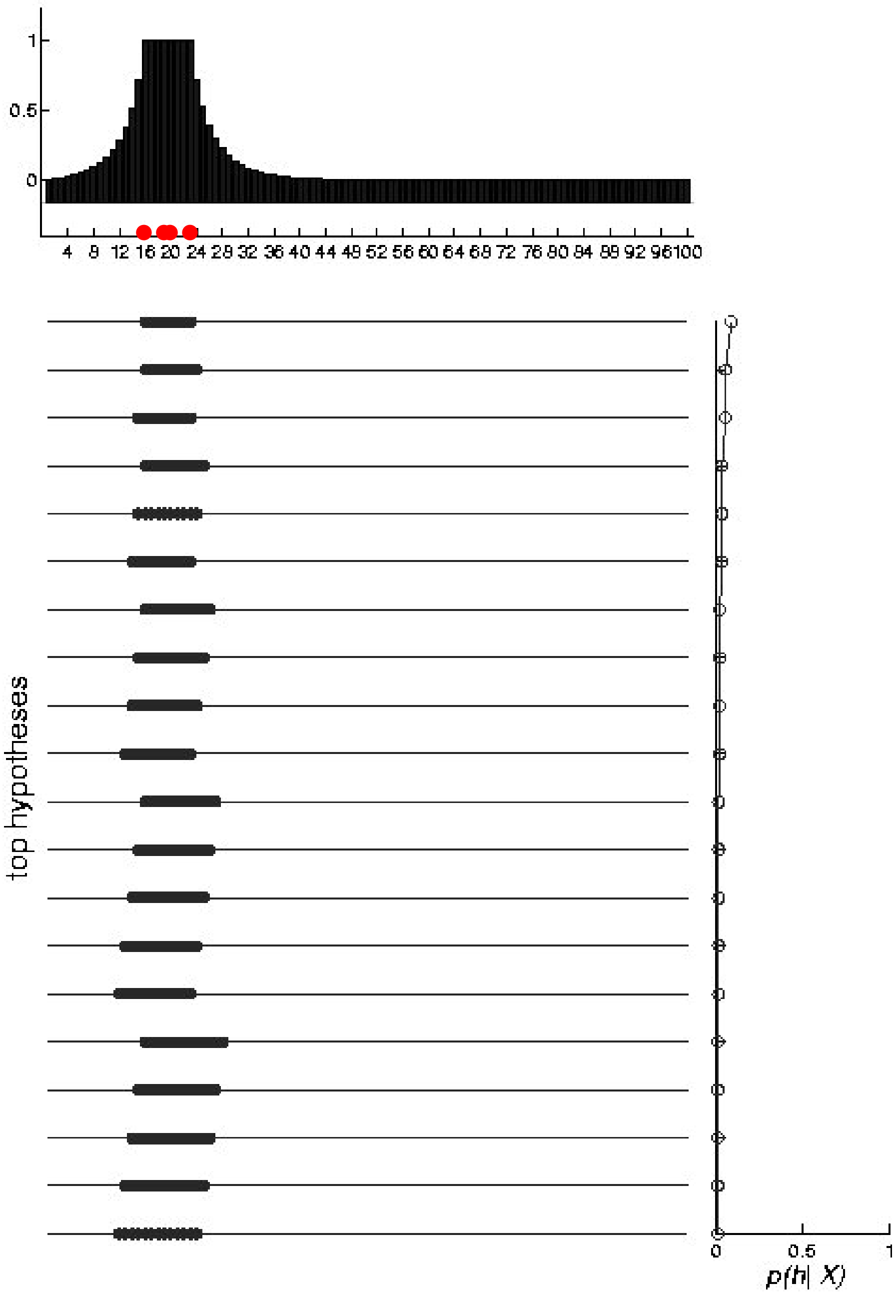
Examples:

16

23

19

20



+ Examples | Human generalization | Bayesian Model

60

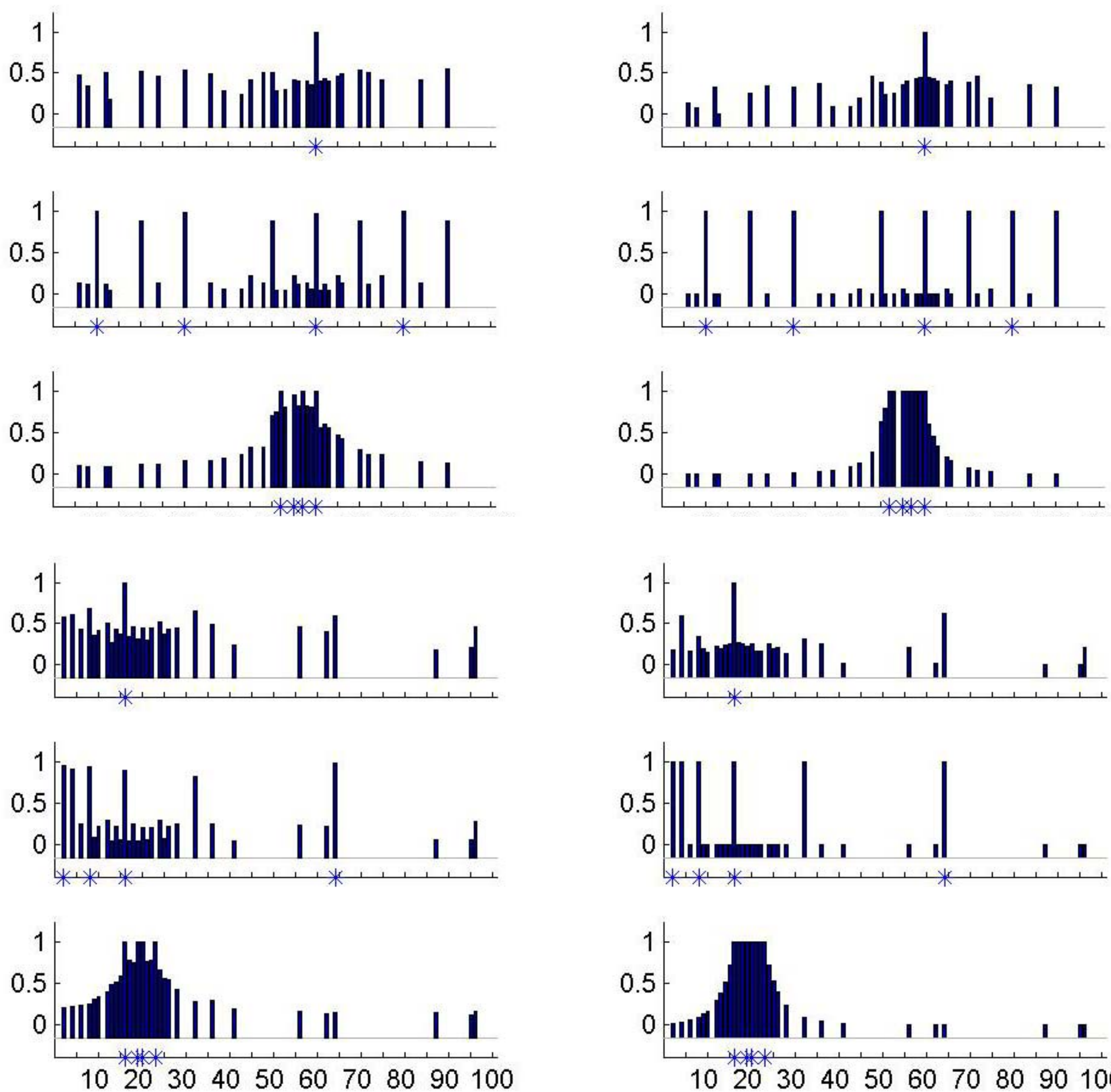
60 80 10 30

60 52 57 55

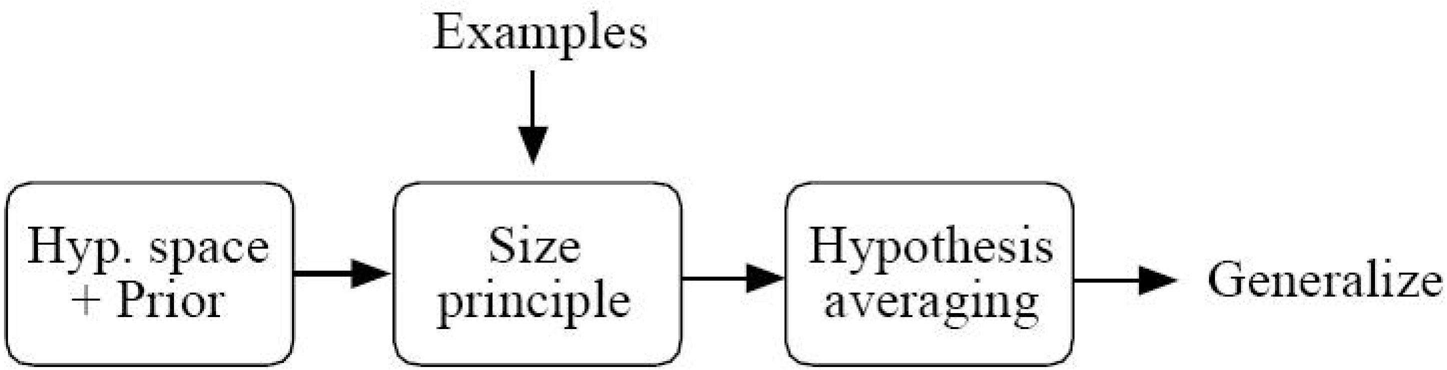
16

16 8 2 64

16 23 19 20



Summary of the Bayesian approach



1. Constrained hypothesis space H
2. Prior p(h)
3. Likelihood p(X|h)
4. Hypothesis (model) averaging:

$$p(y \in C \;\; X) = \sum_{h} p(y \in C|h)p(h|X)$$