Unit 2: Bayesian Learning

2. Learning by Bayesian inference

10/6/2020

Learning by Bayesian inference

1. Bayesian inference provides a framework for causal learning

2. The size principle embodies an assumption about generating processes that leads to stronger inference

3. Graphical models are a powerful and flexible notation for describing Bayesian Models

The number game (Tenenbaum, 2000)

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

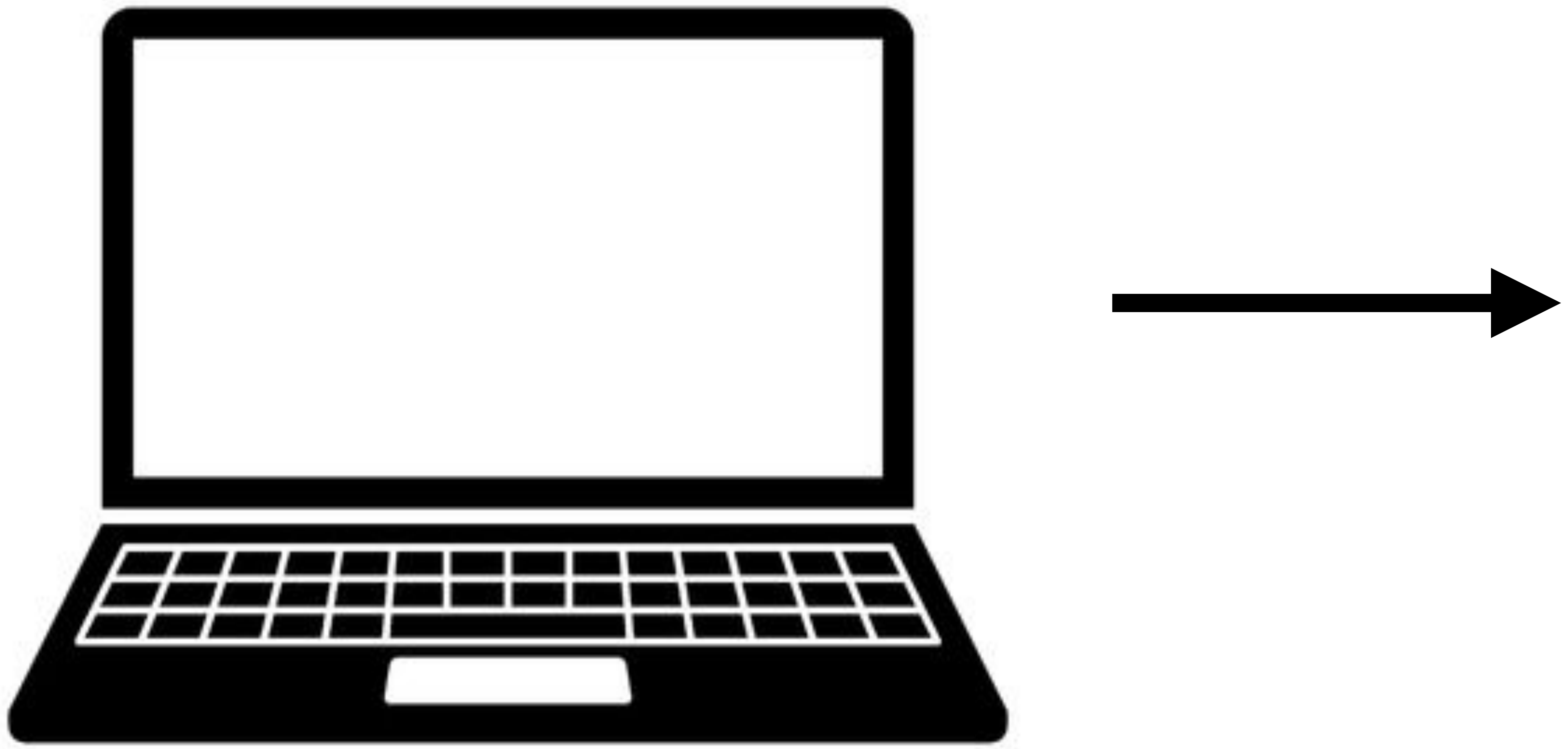
→ **60**

What other numbers will this program generate?

51? **58?** **20?**

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 number game

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

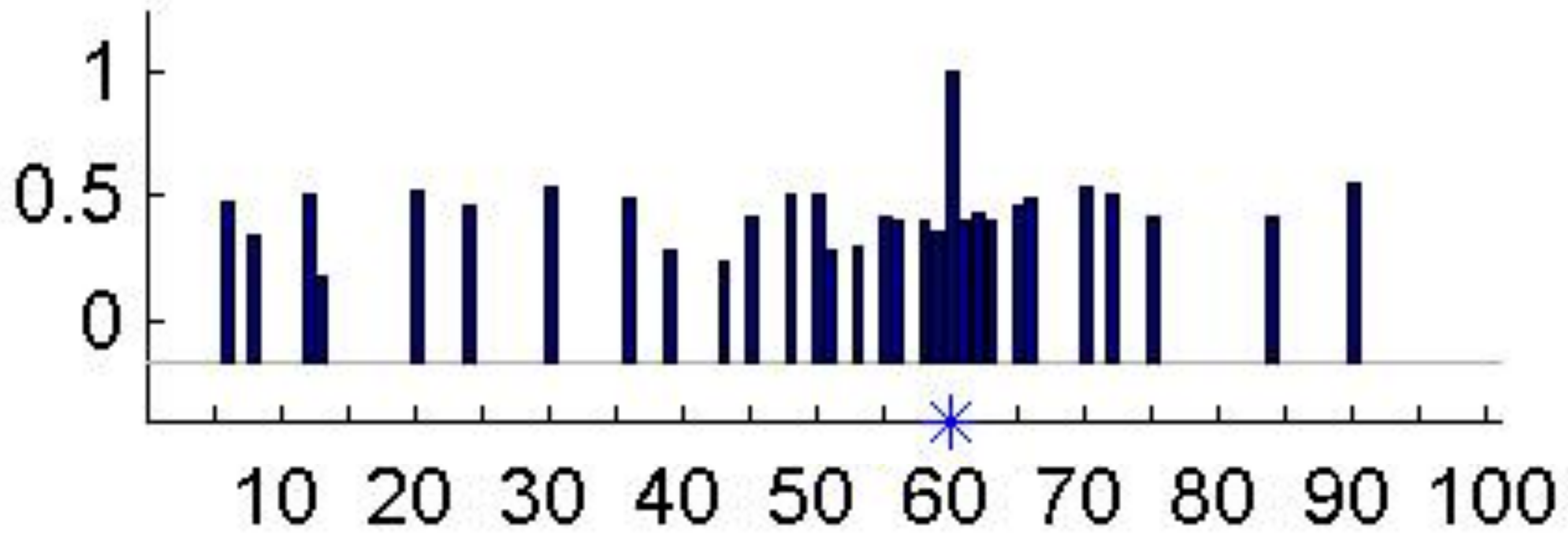
60 52 57 55

What other numbers will this program generate?

51?   58?   20?

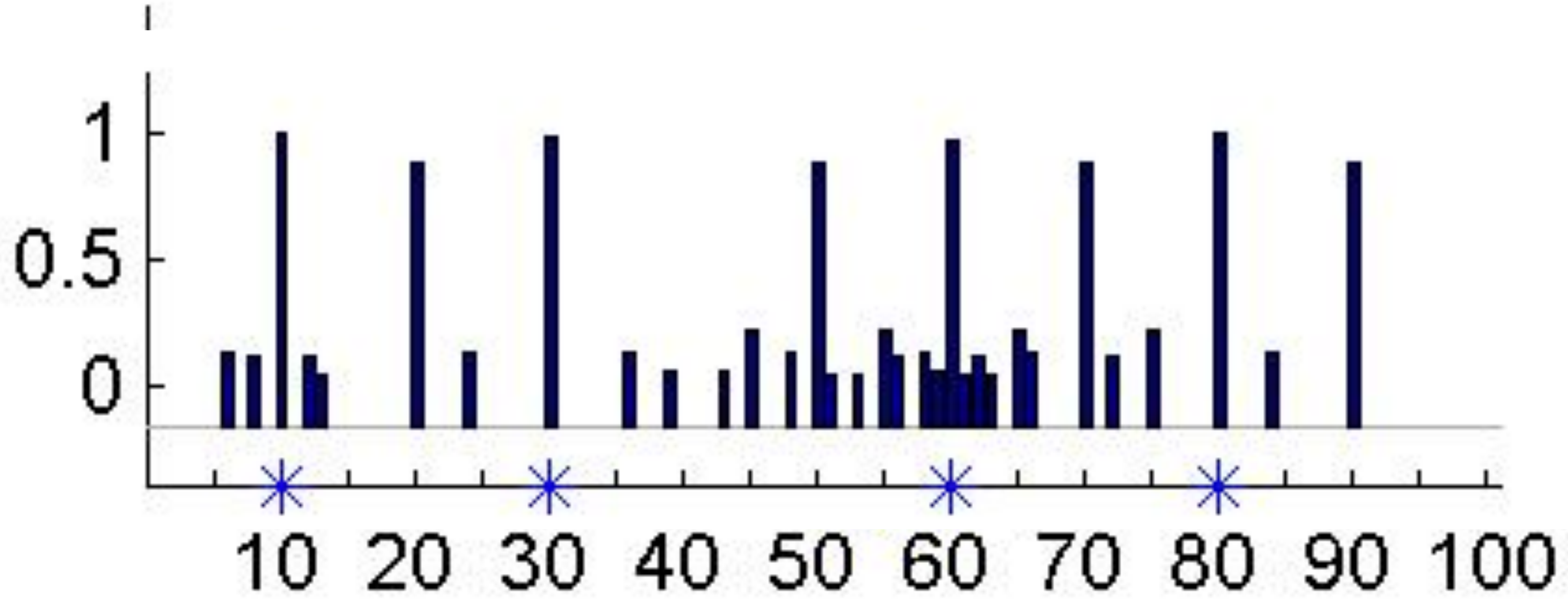
Human judgments in the number game

60



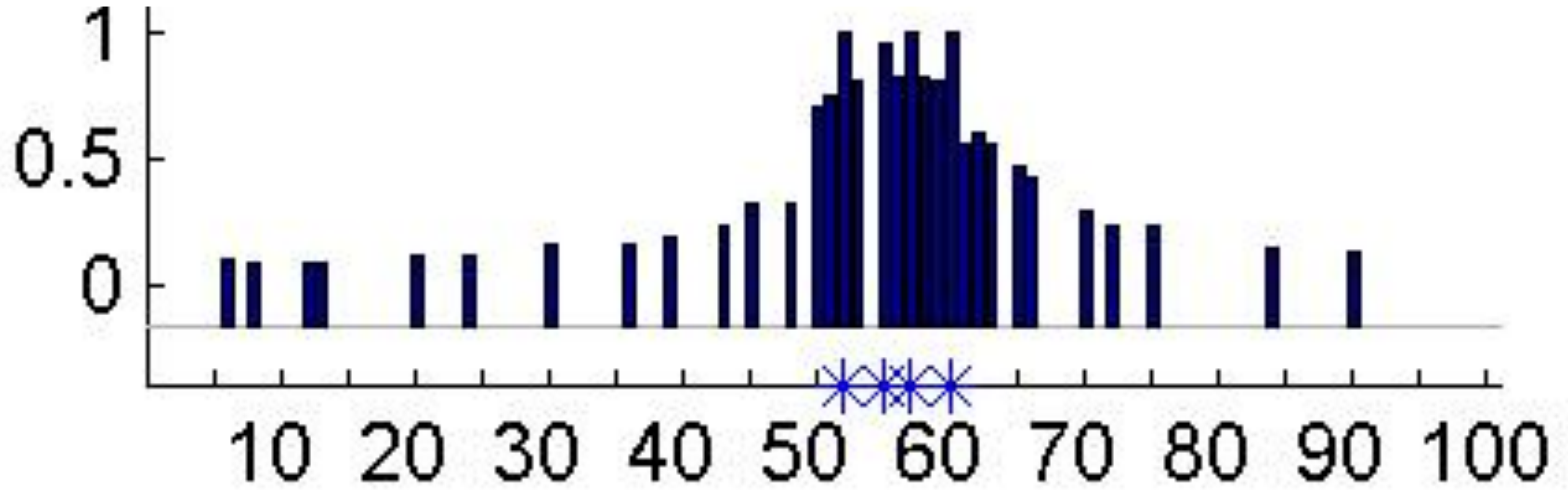
Diffuse similarity

60 80 10 30



Multiples of 10

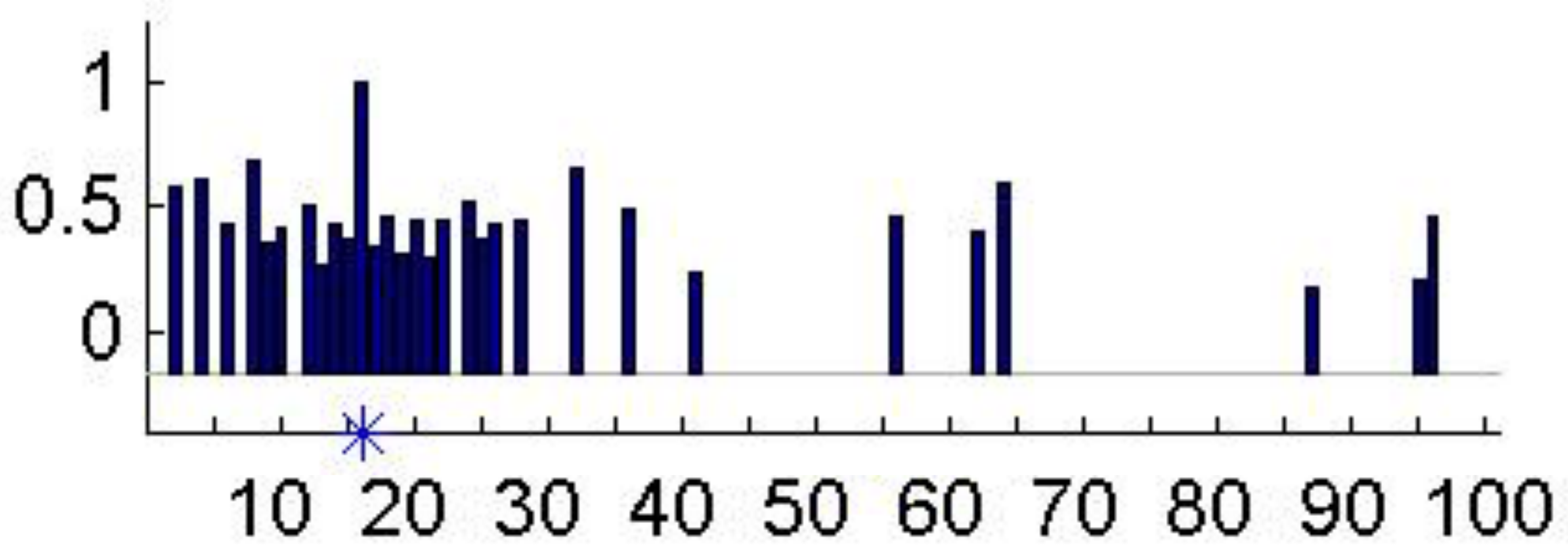
60 52 57 55



Focused similarity
(Near 50-60)

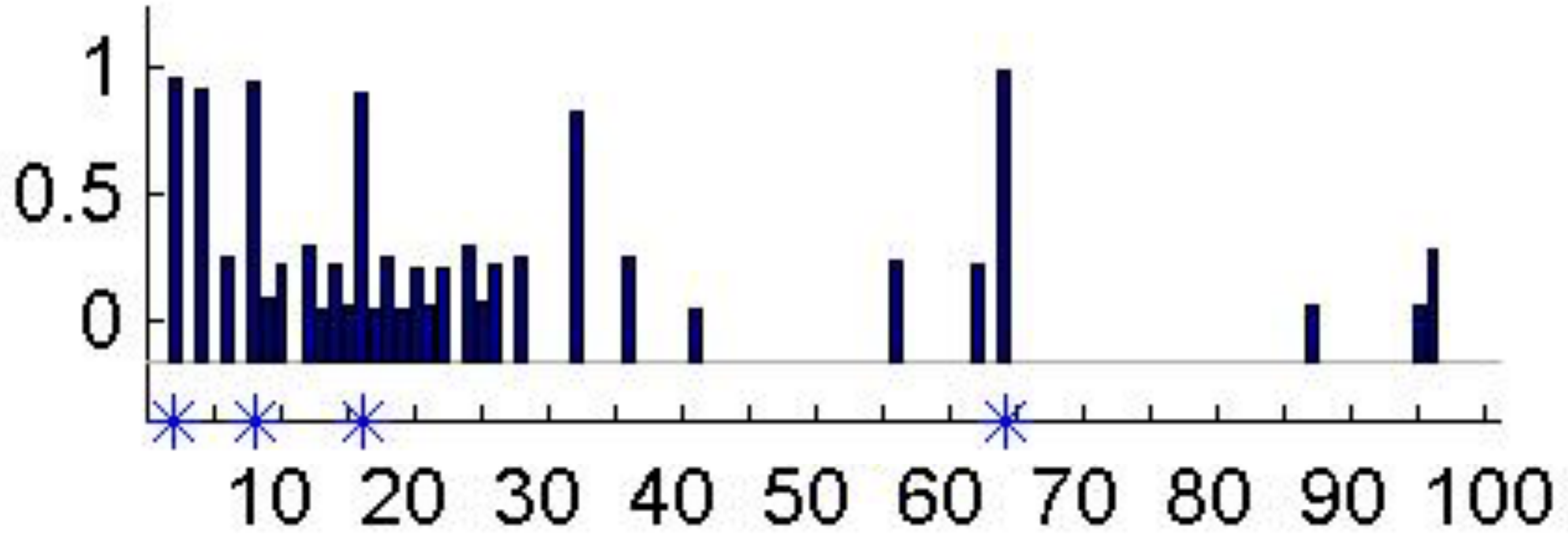
Human judgments in the number game

16



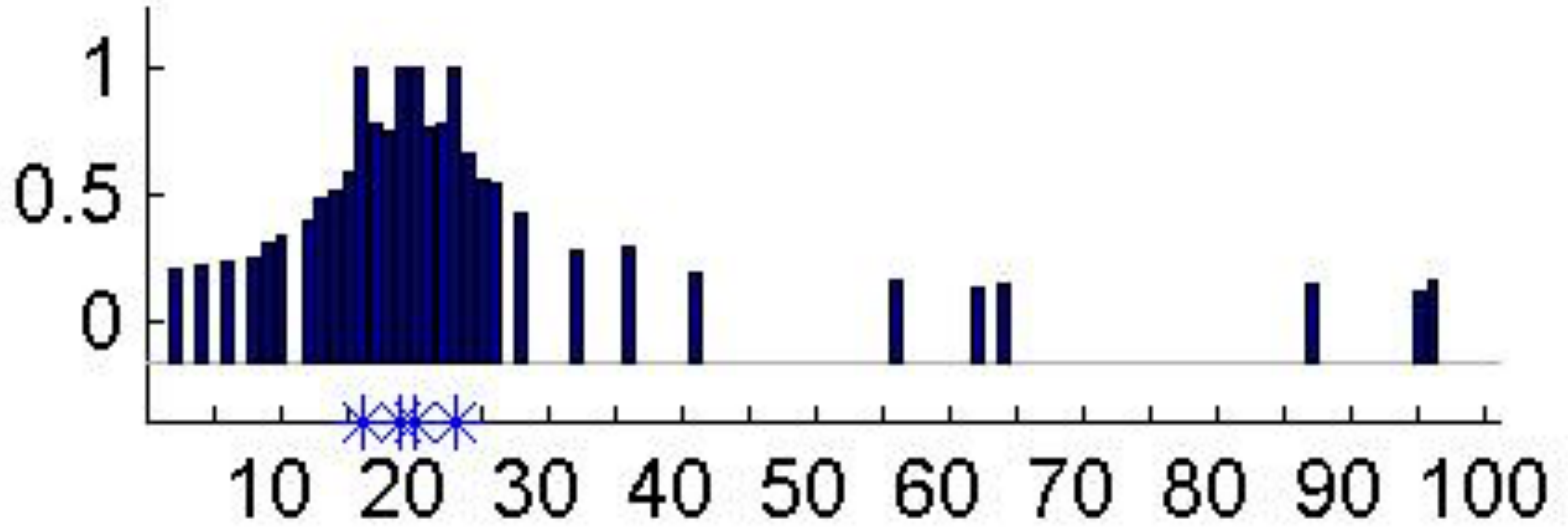
Diffuse similarity

16 8 2 64



Powers of 2

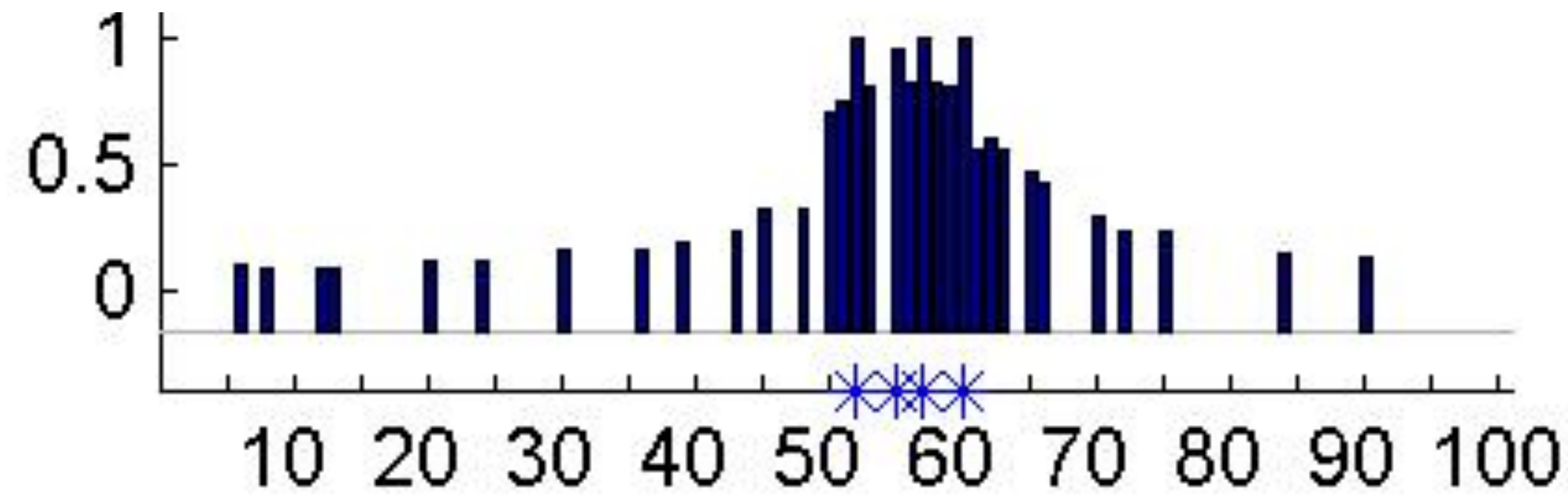
16 23 19 20



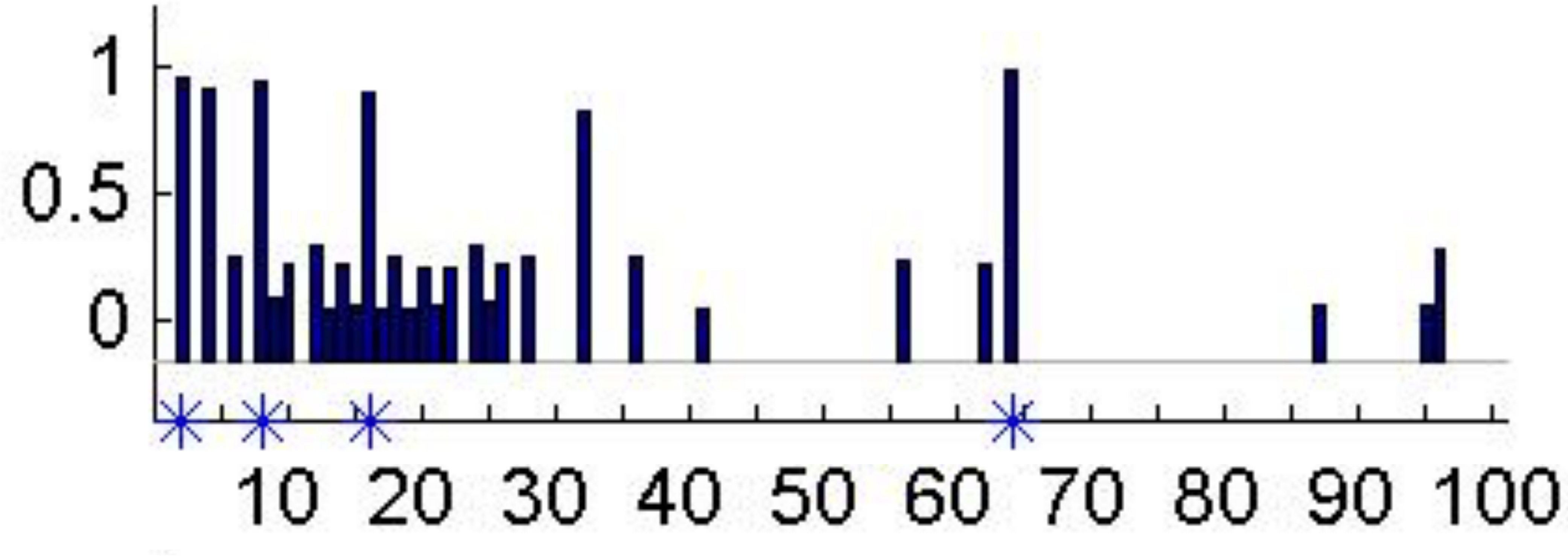
Focused similarity
(Near 20)

Inference is fast, flexible, and can be "rule like" or similarity-based

60 52 57 55



16 8 2 64



A Bayesian model of the number game

Observations: $X = \left\{x_1, \ldots, x_2\right\}$

A set of hypotheses: $h \in H$

- even numbers: $h_1 = \left\{2, 4, 6, \ \ldots \ 96, 98, 100\right\}$
- multiples of 10: $h_2 = \{10, 20, 30, \ \ldots \ 80, 90, 100\}$
- powers of 2: $h_3 = \left\{2, 4, 8, 16, 32, 64\right\}$
- between 50—60: $h_4 = \left\{50, 51, 52, \ \ldots, 58, 59, 60\right\}$
- …

A Bayesian model of the number game

Observations: $X = \left\{x_1, \ldots, x_2\right\}$

A set of hypotheses:

- **Mathematical hypotheses**:
  - odd numbers,
  - even numbers,
  - square numbers,
  - cube numbers,
  - primes,
  - multiples of n $(3 \leq n \leq 12)$
  - powers of n $(2 \leq n \leq 10)$

- **Interval hypotheses**:
  - Decades

    $\{1 - 10,\ 10 - 20,\ \ldots\}$
  - Any range

    $1 \leq n \leq 100$

    $n \leq m \leq 100$

    $\{n - m\}$

A Bayesian model of the number game

Observations: $X = \{x_1, \dots, x_2\}$

A set of hypotheses: $h \in H$

A prior: $P(h) = \begin{cases} \frac{\lambda}{N}, & N \text{ mathematical hypotheses} \\ \frac{(1-\lambda)}{M}, & M \text{ interval hypotheses} \end{cases}$

Likelihood: $P(X|h) = \prod_{x}^{x} P(x|h)$

The size principle

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	14	46	48	50
52	24	56	58	60
62	34	66	68	70
72	74	76	78	80
82	84	86	88	90
92	94	96	98	100

Likelihood:

$$P\left(x \mid h\right) = \begin{cases} \frac{1}{|h|}, & x \in h \\ 0 & \text{otherwise} \end{cases}$$

60: slightly more likely powers of 10

10 30 60 80:
much more likely powers of 10

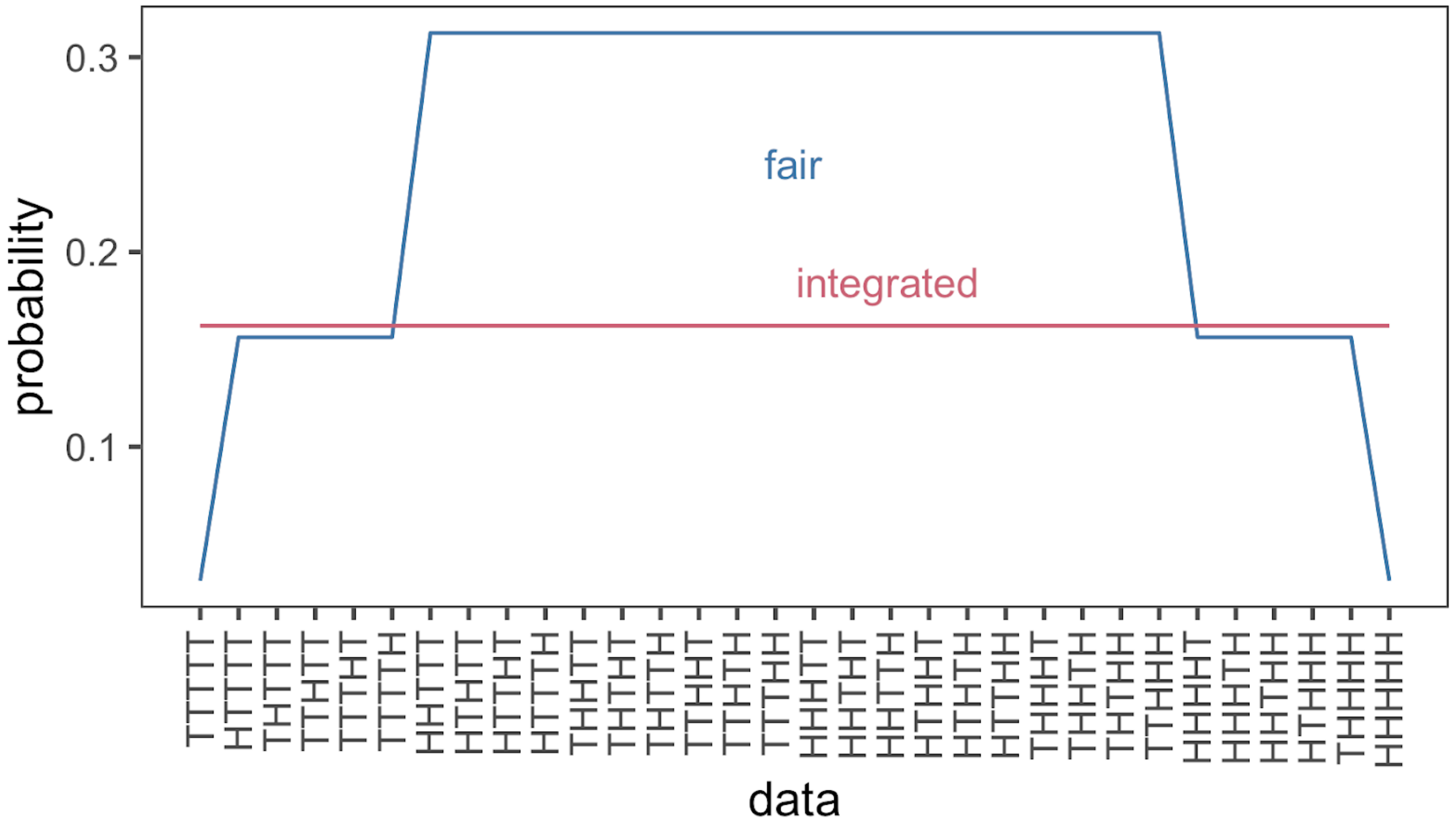
Bayesian Occam's Razor

Just like our biased coin example!

Simple vs. complex hypotheses:

H_1 : Fair Coin — $P(H) = .5$

H_2 : Biased Coin — $P(H) = p$



Law of conservation of belief

A Bayesian model of the number game

Observations: $X = \{x_1, \ldots, x_2\}$

A set of hypotheses: $h \in H$

A prior: $P(h) = \begin{cases} \frac{\lambda}{N}, & N \text{ mathematical hypotheses} \\ \frac{(1-\lambda)}{M}, & M \text{ interval hypotheses} \end{cases}$

Likelihood: $P(x|h) = \begin{cases} \frac{1}{|h|}, & x \in h \\ 0 & \text{otherwise} \end{cases}$

Posterior: $$P(h|X) = \frac{P(X|h)\,P(h)}{\sum_{h' \in H} P(X|h')\,P(h')}$$

Making predictions about new numbers

Observations: $X = \{x_1, \ldots, x_2\}$ **Posterior:** $P(h|X) = \frac{P(X|h)\,P(h)}{\sum_{h' \in H} P(X|h')\,P(h')}$

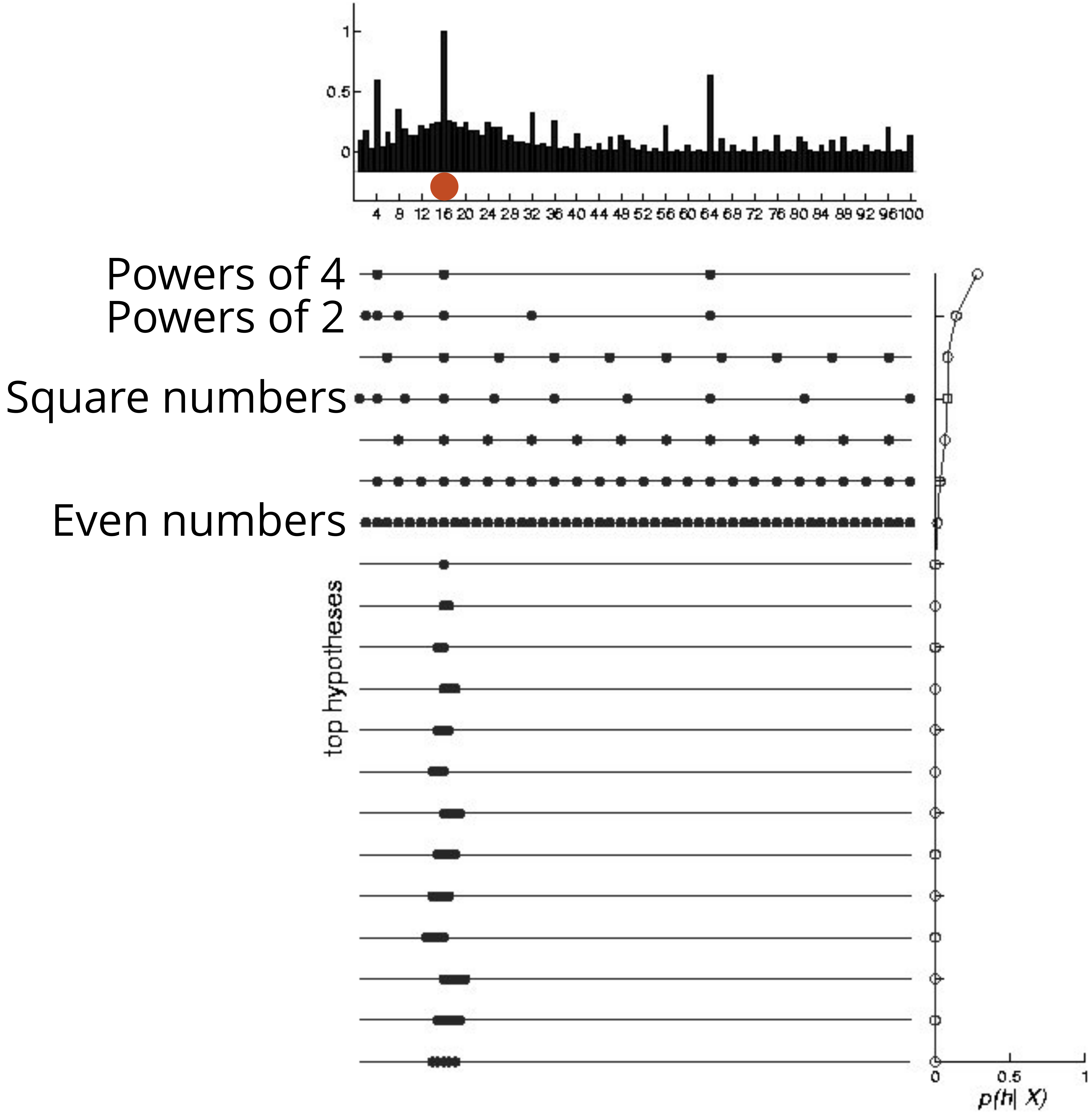
What about a new number? $P(y \in C|X)$

Posterior prediction: $P(y \in C|X) = \sum_{h \in H} P(y \in C|h)\,P(h|X)$

Bayesian hypothesis averaging: To make optimal predictions, average over all possible hypotheses, weighted by their posterior

Model predictions

16



Model predictions

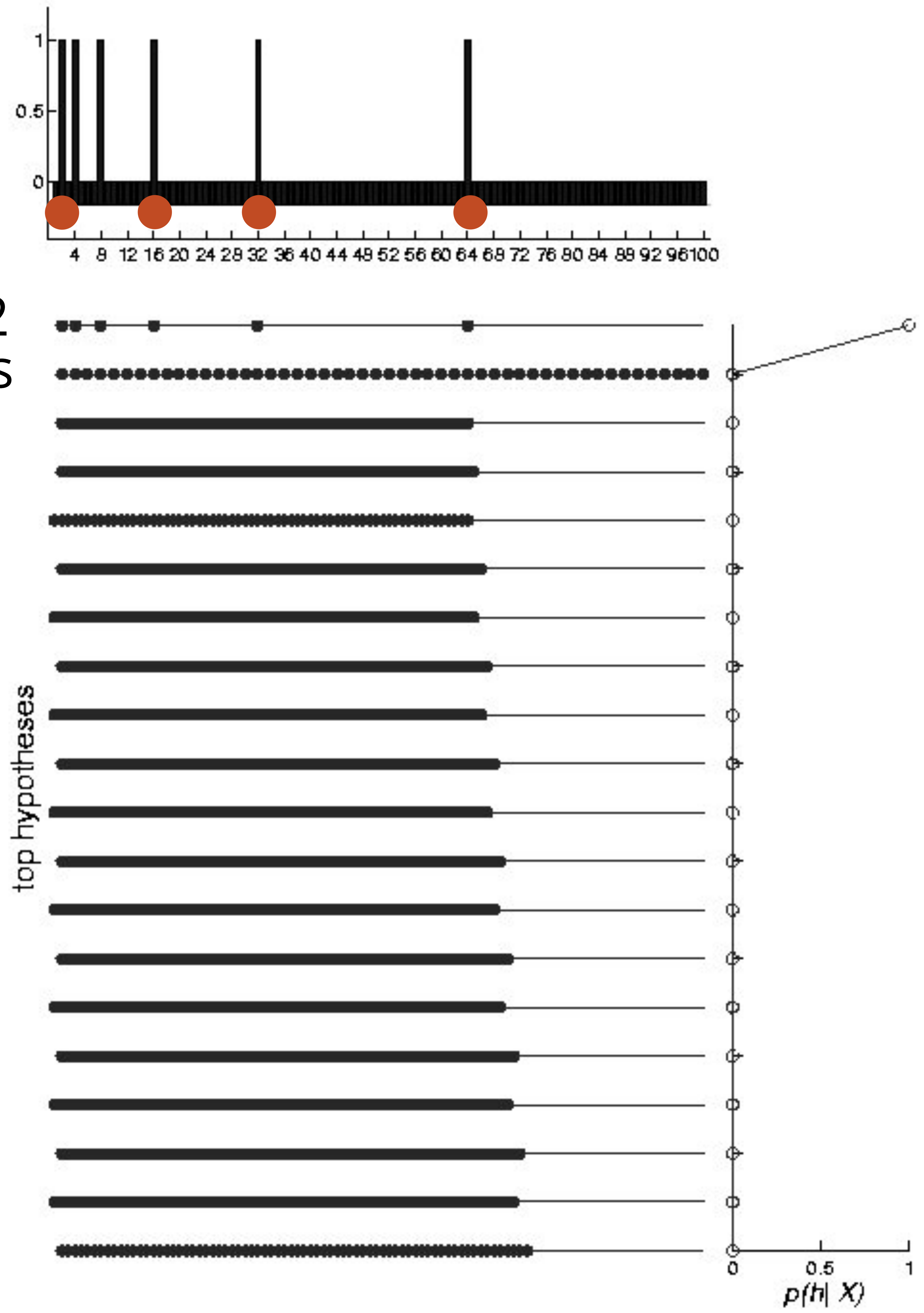
16

8

2

64

Powers of 2
Even numbers



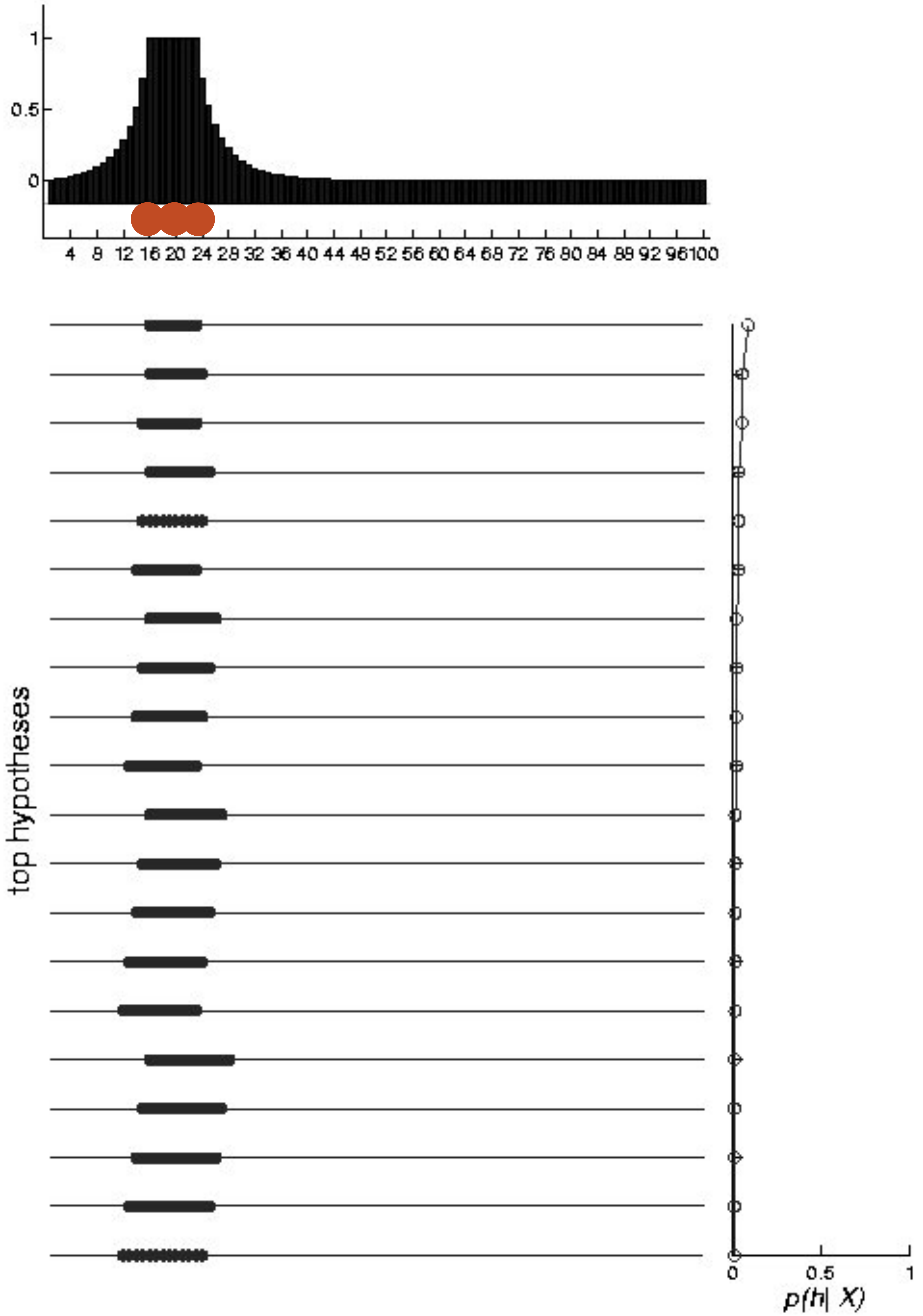
Model predictions

16

23

19

20



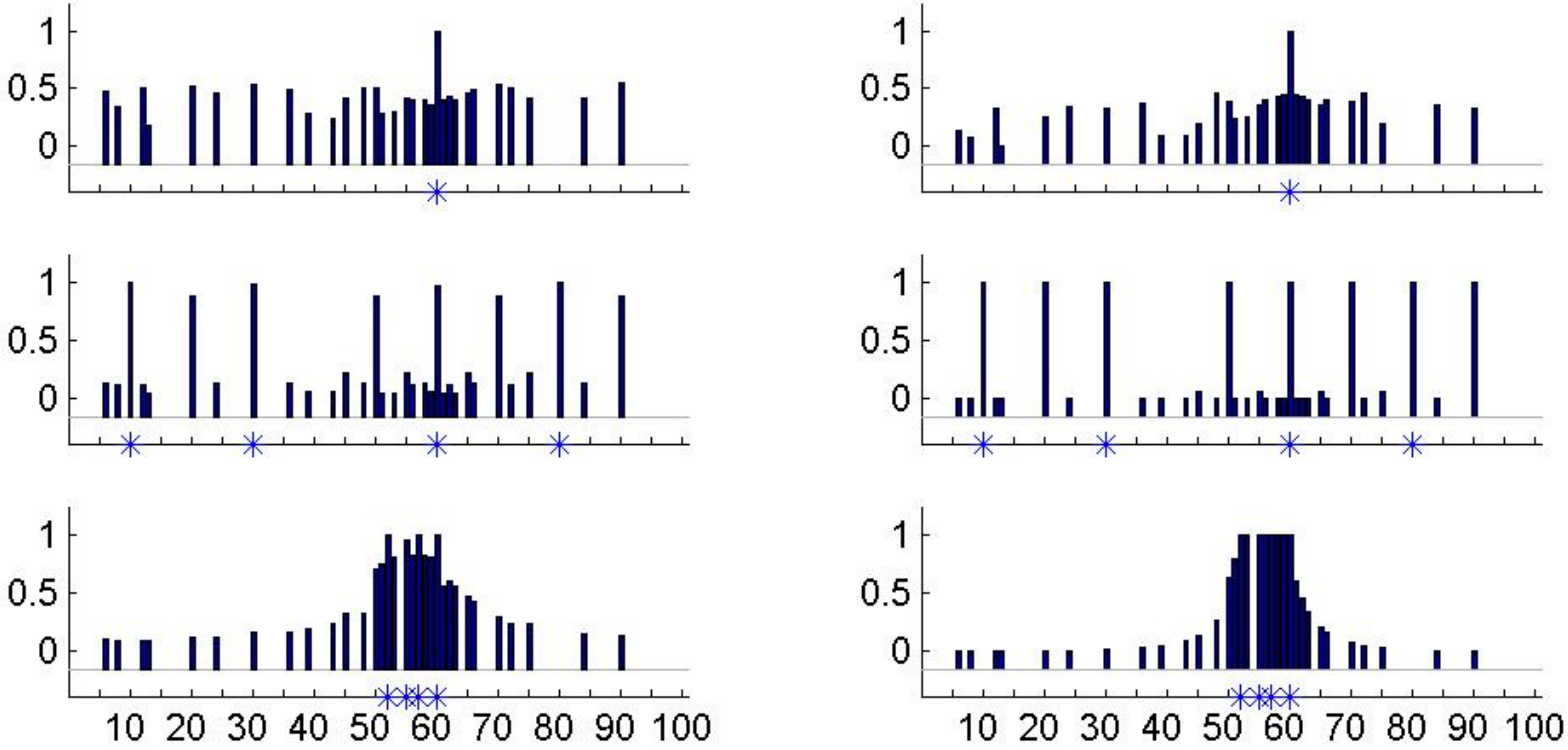
Model fits

Humans | **Model**

60

60 80 10 30

60 52 57 55



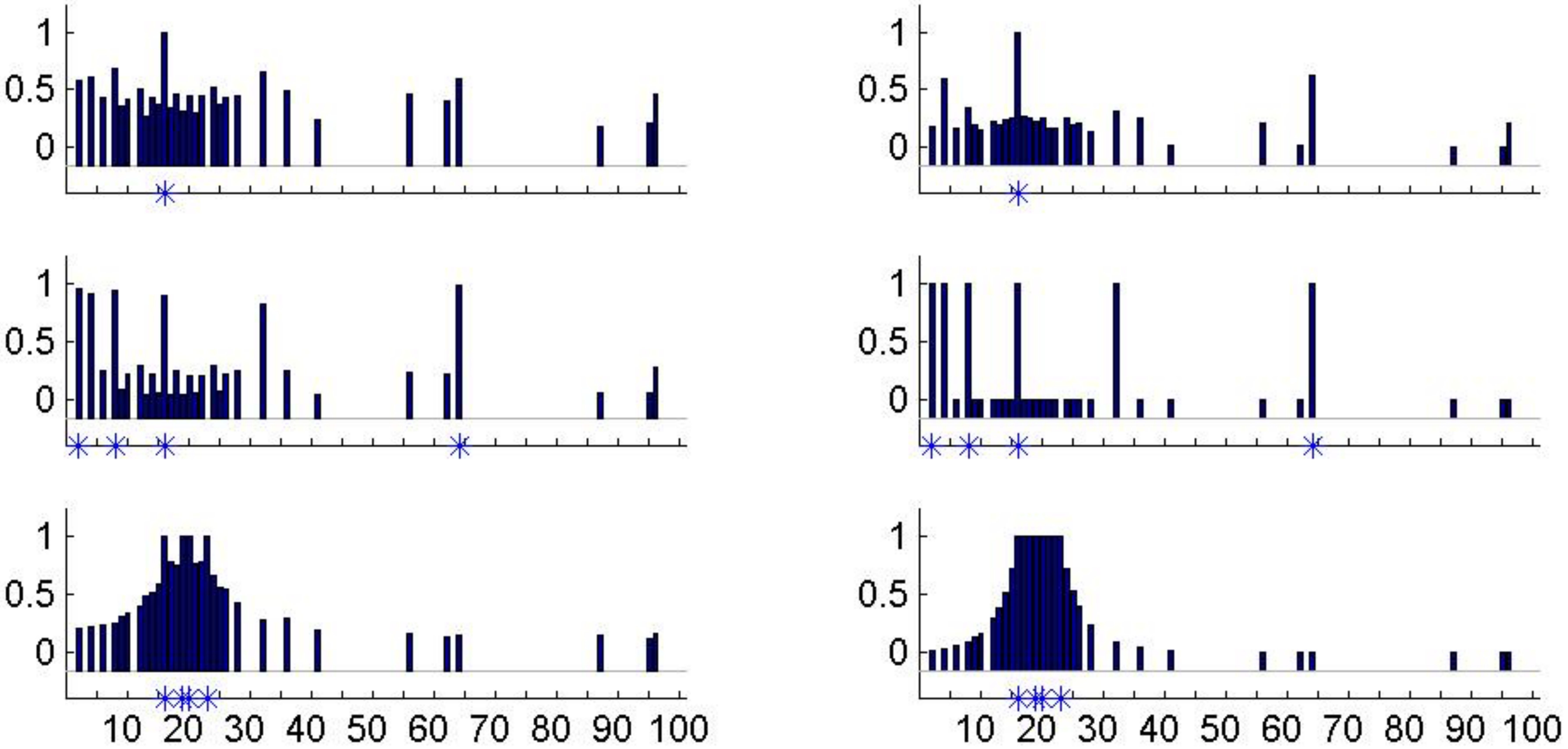
Model fits

Humans **Model**

16

16 8 2 64

16 23 19 20



The gavagai problem

Quine (1960)

Let’s try it out

dalmatian

dog

animal

dax

dalmatian

dog

animal

dax dax dax

Xu & Tenenbaum (2007)

What’s going on here?

$$P(H|D) \propto P(D|H)P(H)$$

P(dog | ) ∝ P( | dog)P(dog)

P(dalmation | ) ∝ P( | dalmation)P(dalmation)

What is P(dog)? What is P(dalmation)?

So maybe P(dog) > P(dalmation)

The size principle!

P( | dog)

<

P( | dalmation)

What’s going on here?

$$P(H|D) \propto P(D|H)P(H)$$

P(dog | ) $\propto$ P( | dog)P(dog)

P(dalmation | ) $\propto$ P( | dalmation)P(dalmation)

What is P( | dog)?

3 dalmatians from the dog category? A suspicious coincidence!

$$P(H|D) \propto P(D|H)P(H)$$

P(dog |  ,  , ) ∝

P( ,  ,  | dog)P(dog)

P(dalmation |  ,  , ) ∝

P( ,  ,  | dalmation)P(dalmation)

The size principle!

P(, ,  |dog)

If I'm picking examples from the dog category,
it's **really** unlikely to pick three dalmations

Let's try it out

dalmatian

dog

animal

dax

dalmatian

dog

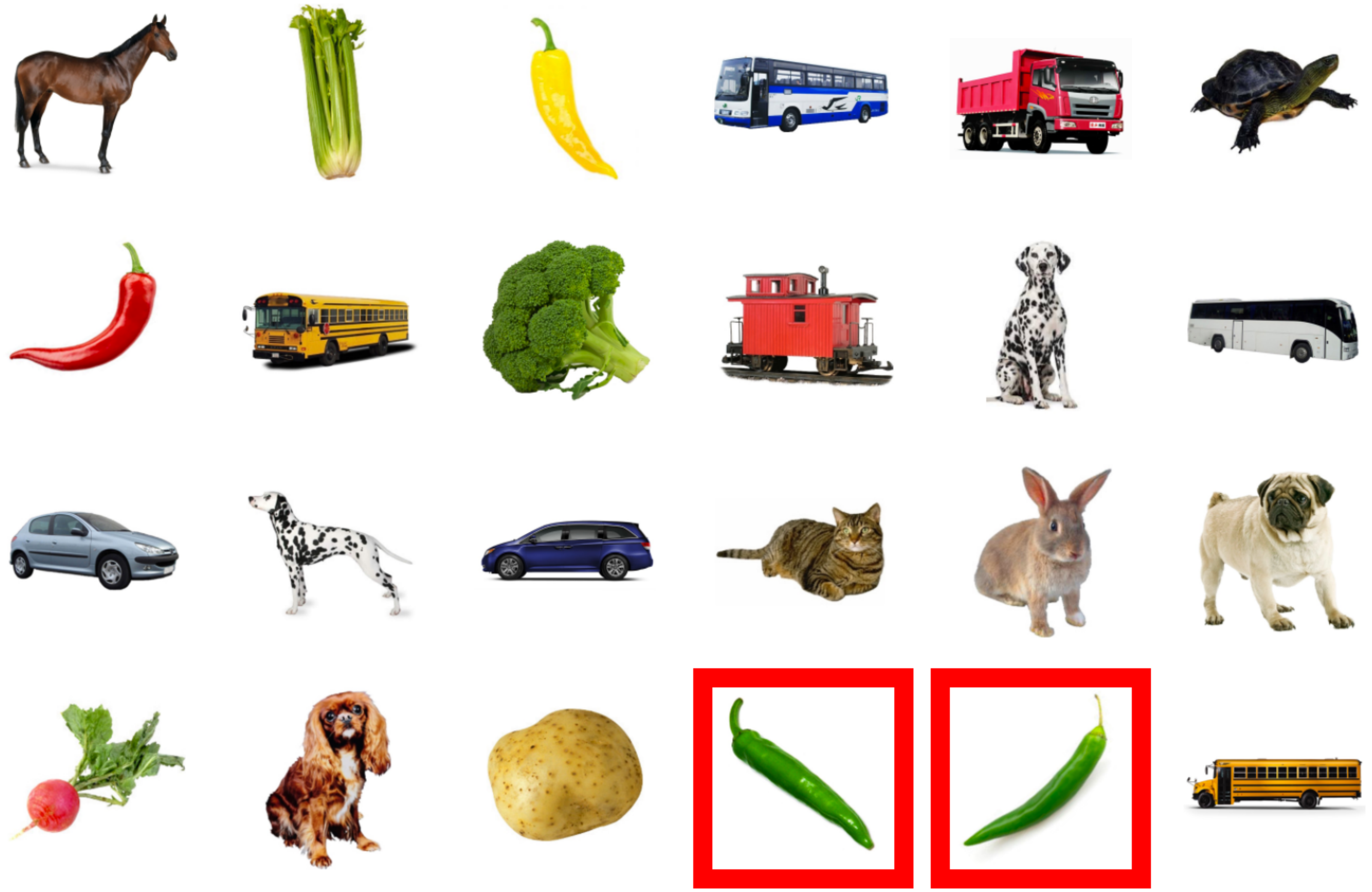
animal

dax dax dax

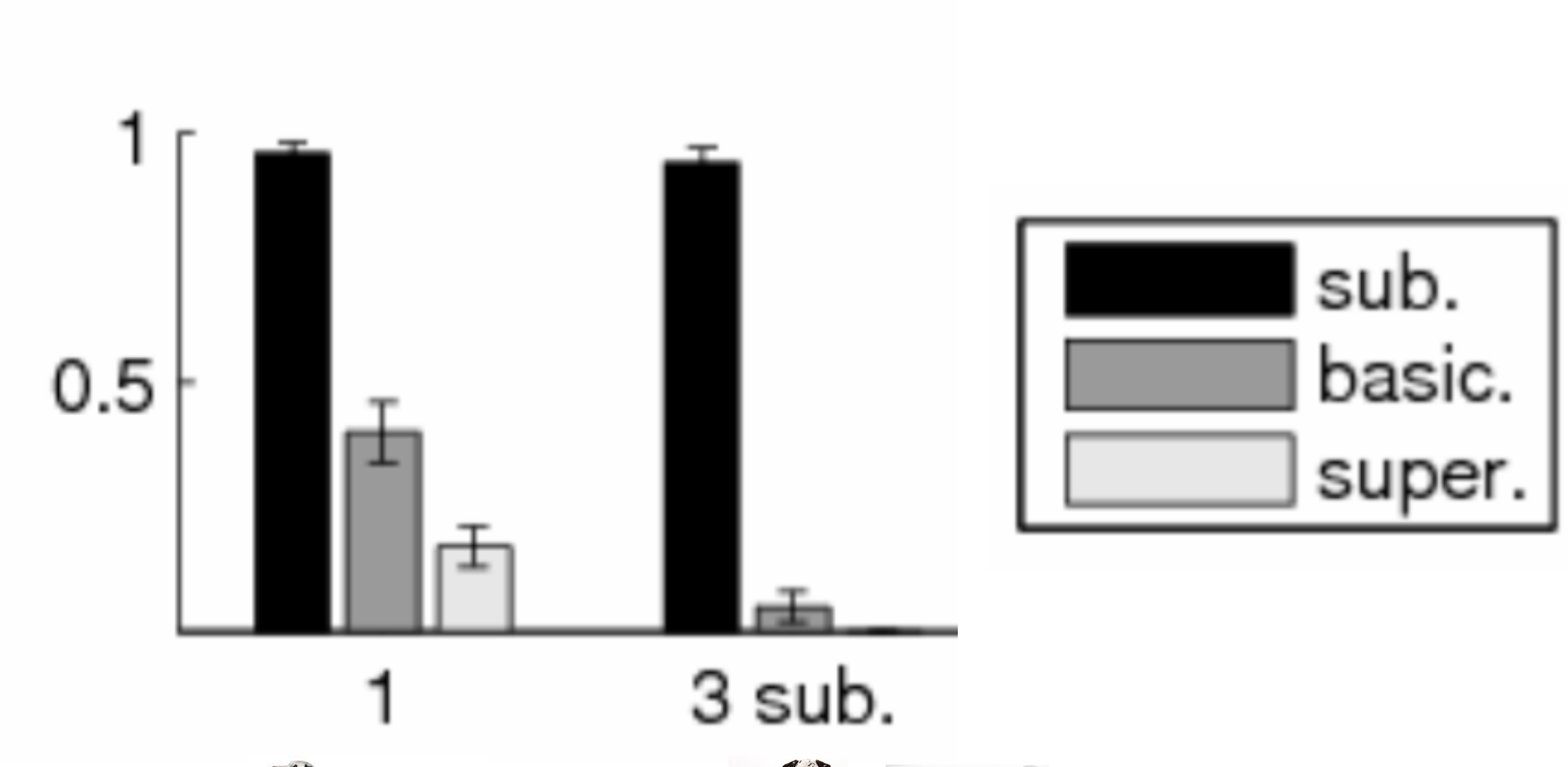
Testing the suspicious coincidence

Here are three sibs. Can you give Mr. Frog all the other sibs?

To give a sib, click on it below. When you have given all the sibs, click the Next button.



3- and 4-year-olds make this inference

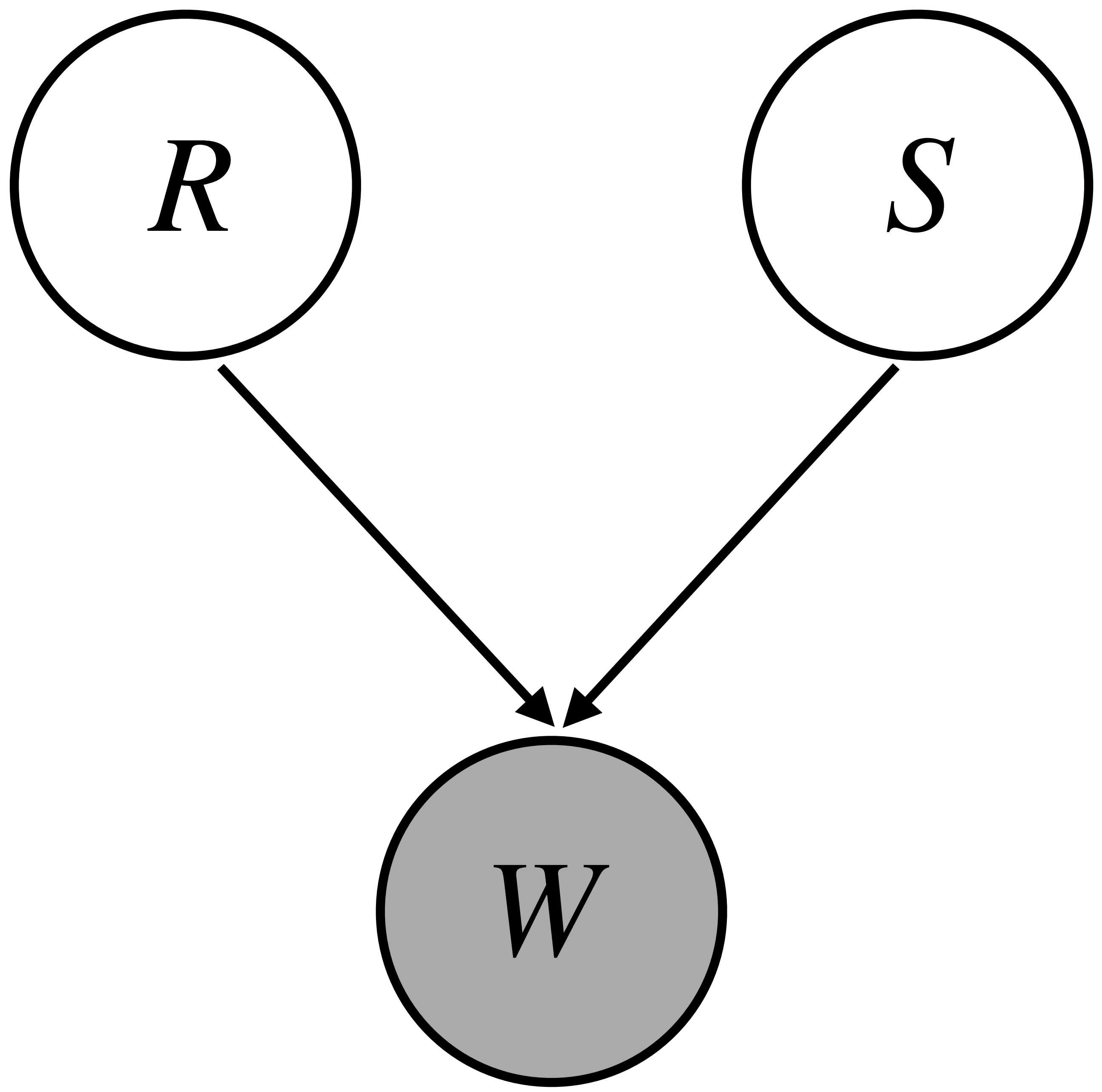


Graphical models

Graphical models are a visual notation for expressing the probabilistic relationships among a set of variables.

Components:

1. **Vertices** that represent the variables
2. **Edges** that represent statistical dependencies between the vertices
3. A set of **probability distributions** that describe these dependencies

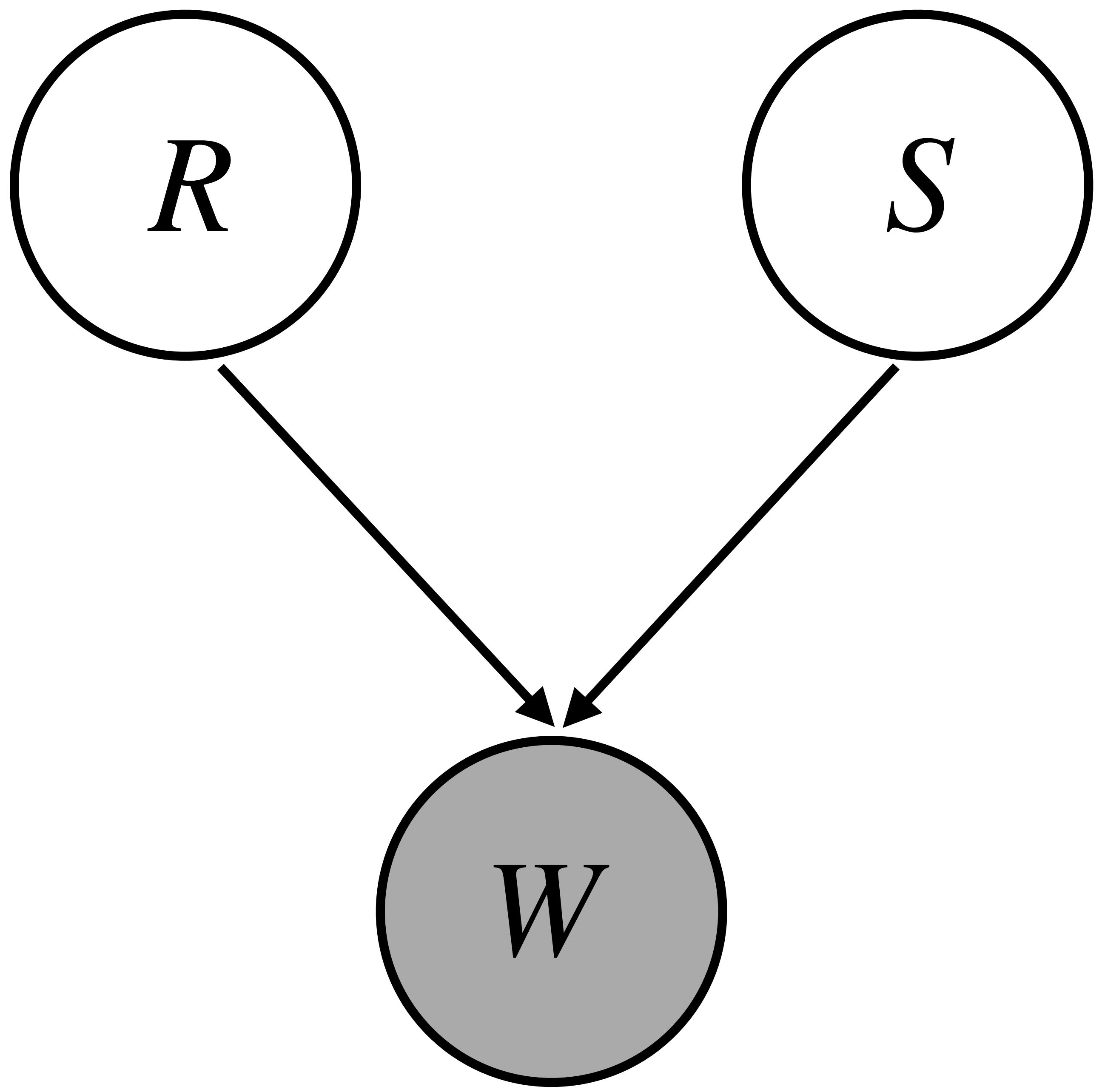


Latent and Observed Variables

Vertices represent two kinds of variables:

Components:

1. **Observed variables** (filled circles) are variables whose values we see directly.
2. **Latent variables** (empty circles) are variables that we do not see, but that explain the process that generated the observed variables.



Typically, we want to infer the values of the latent variables from the observed variables in our data

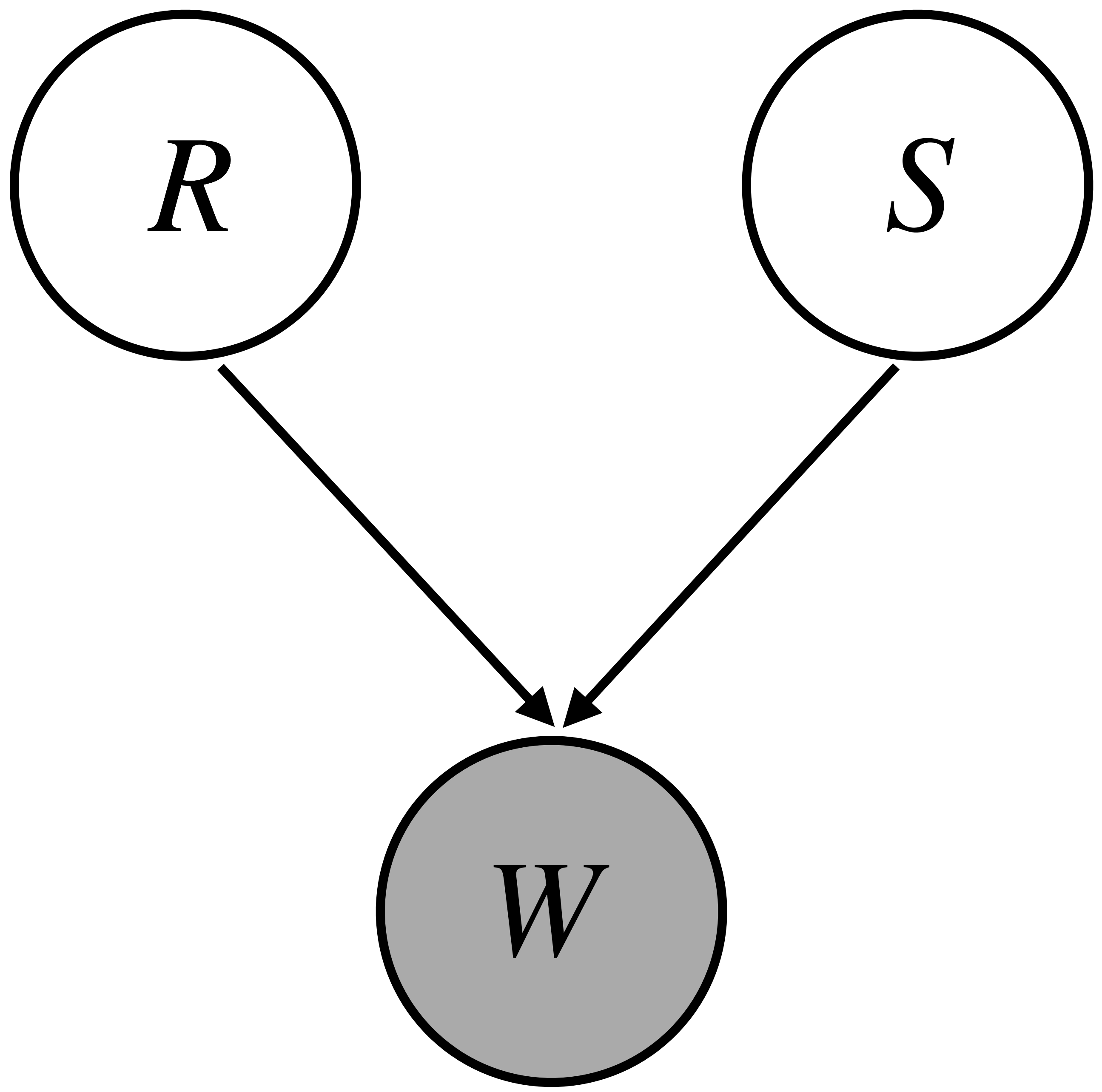
A graphical model for wet grass

This simple model describes how grass might get wet

W denotes whether grass is wet or dry. We is an observed variables because we get to see it

R (rain) and S (sprinklers) are potential causes of wet grass. They are latent because we don't get to observe them

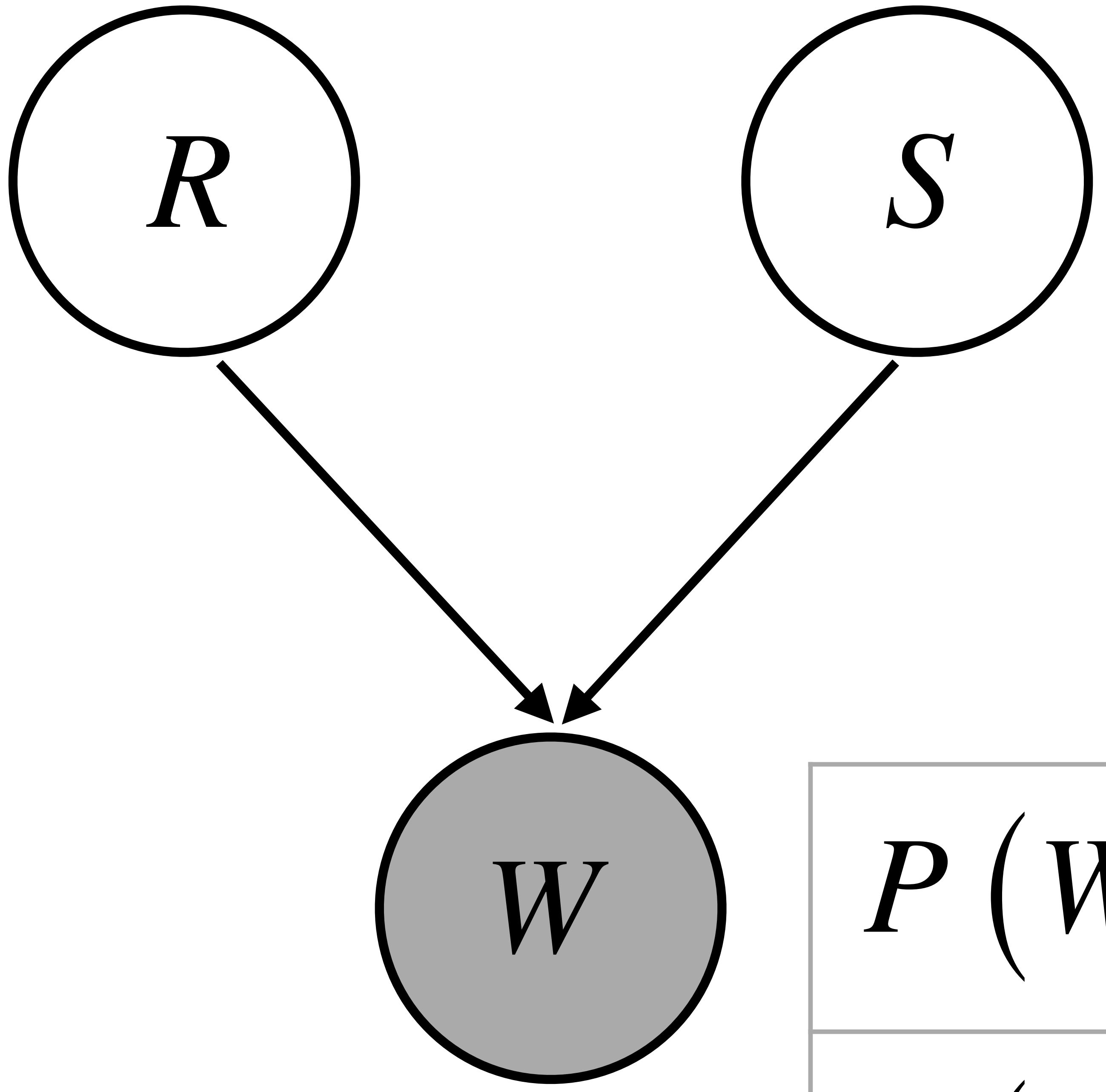
Because there is no arrow between R and S, we know that they are independent



Using the model to reason forward

$$P(R) = .4$$

$$P(S) = .2$$



$$P\left(W|S\,\&\,R\right) = .95$$

$$P\left(W|S\,\&\sim R\right) = .9$$

$$P\left(W|\sim S\,\&\,R\right) = .9$$

$$P\left(W|\sim S\,\&\sim R\right) = .1$$

Suppose we *know* the sprinklers turned on.

What is the probability that the grass is wet?

$$P\left(W|S\right) = P\left(W|S\,\&\,R\right)P(R) + P\left(W|S\,\&\sim R\right)P(\sim R)$$

$$P\left(W|S\right) = .96 \cdot .4 + .9 \cdot .6 = .92$$

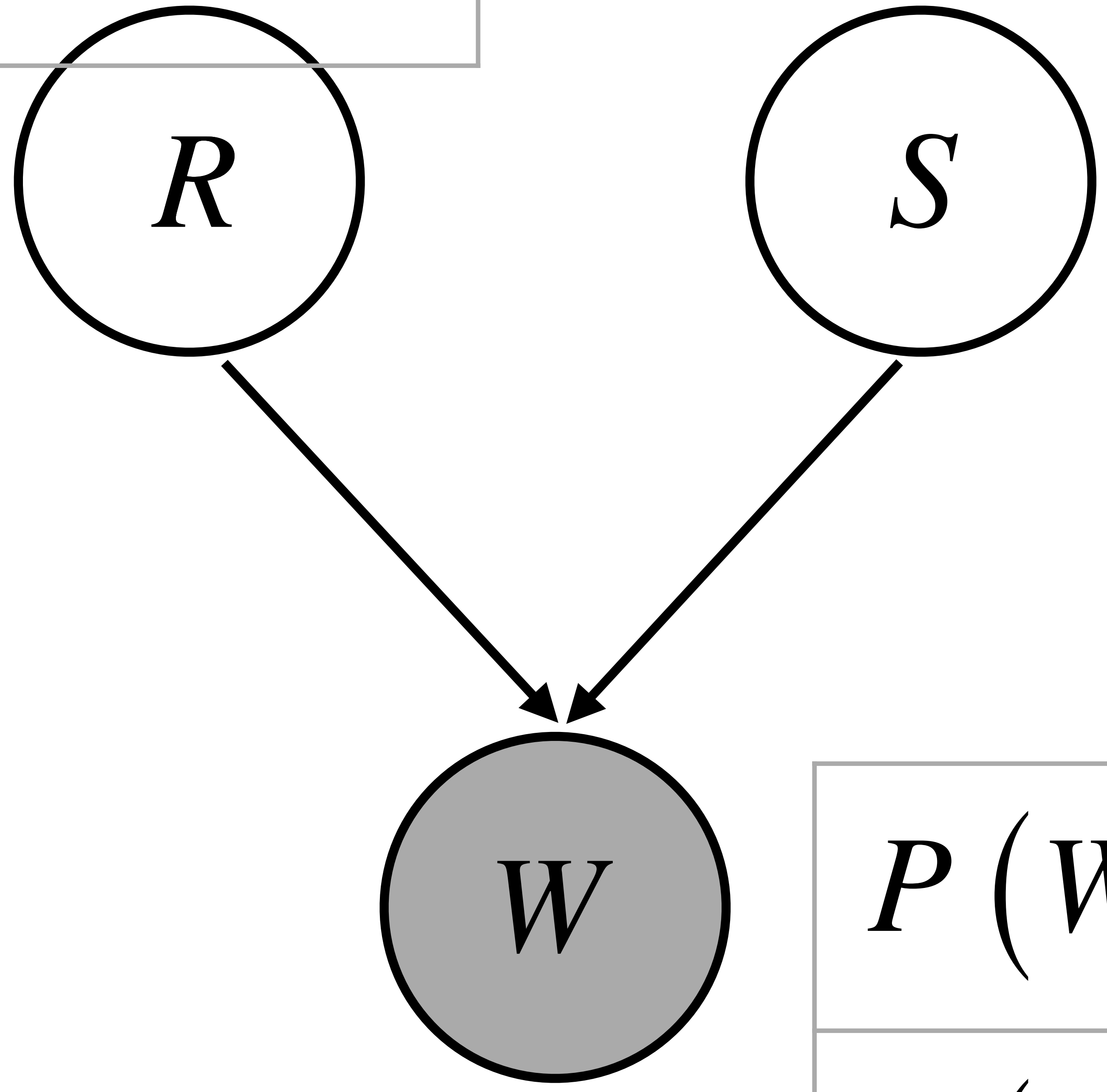
Using the model to reason backward

Suppose we *know* the grass is wet.

What is the probability that the sprinklers are turned on?

$P(R) = .4$

$P(S) = .2$



$$P(W|S \,\&\, R) = .95$$
$$P(W|S \,\&\, \sim R) = .9$$
$$P(W| \sim S \,\&\, R) = .9$$
$$P(W| \sim S \,\&\, \sim R) = .1$$

$$P(S|W) = \frac{P(W|S)\,P(S)}{P(W)} = \frac{.92 \cdot .2}{.52} \approx .35$$

$$\begin{aligned} P(W) = {} & P(W|S \,\&\, R)\,P(S)\,P(R) \\ & + P(W|S \,\&\, \sim R)\,P(S)\,P(\sim R) \\ & + P(W| \sim S \,\&\, R)\,P(\sim S)\,P(R) \\ & + P(W| \sim S \,\&\, \sim R)\,P(\sim S)\,P(\sim R) \end{aligned}$$

$$= .92 \cdot .2 \cdot .4 + .9 \cdot .2 \cdot .6 + .9 \cdot .8 \cdot .4 + .1 \cdot .8 \cdot .6 = .52$$

Using the model to diagnose hidden causes

Suppose we *know* the grass is wet and that it rained.

What is the probability that the sprinklers are turned on?

$$P(S|W \& R) = \frac{P(W \& S \& R)}{P(W \& R)}$$

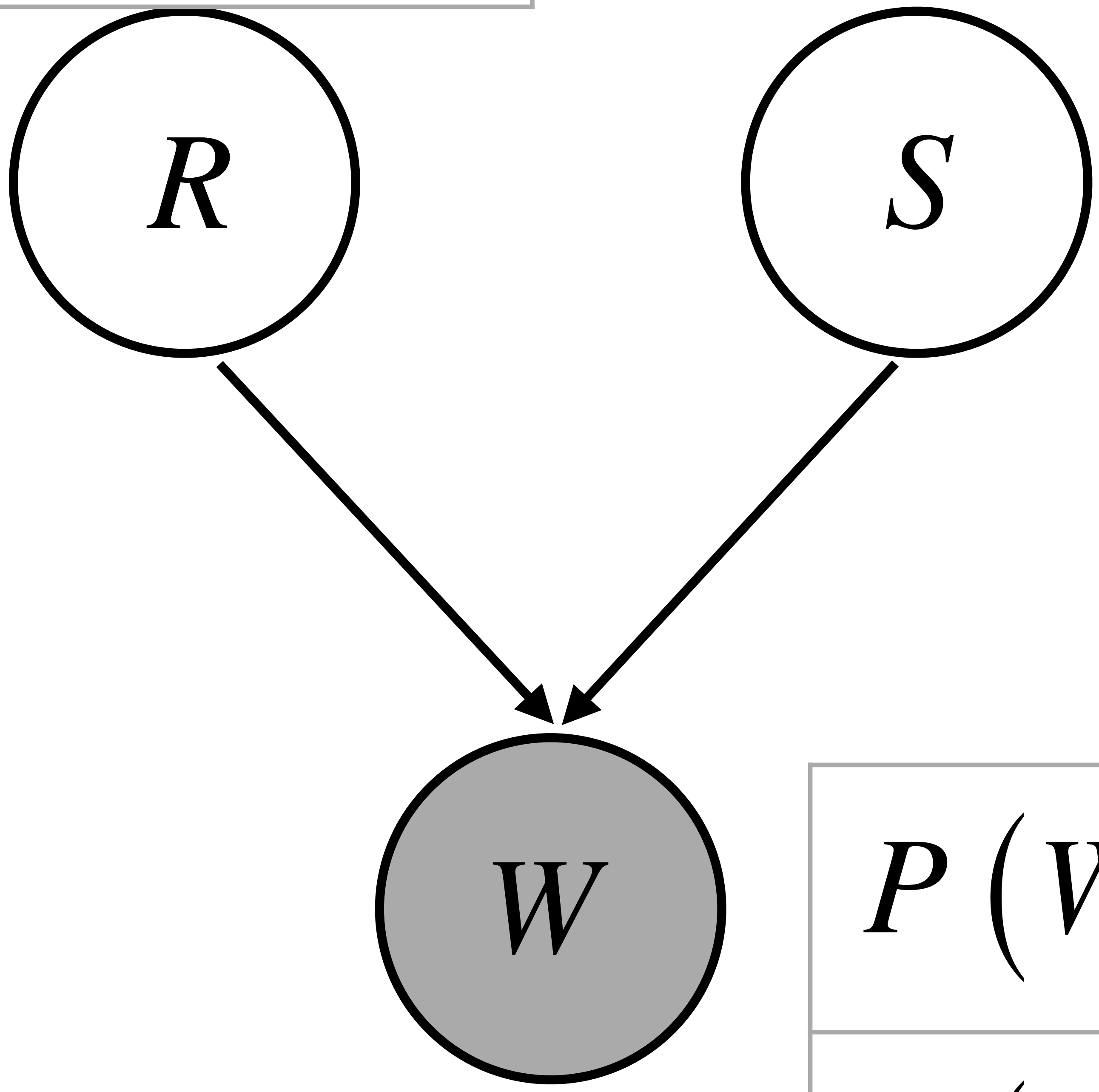
$$= \frac{P(W| \& S \& R)\, P(S \& R)}{P(W \& R)}$$

$$= \frac{P(W| \& S \& R)\, P(S)}{P(W|R)}$$

$$= \frac{P(W| \& S \& R)\, P(S)}{P(W|S \& R)\, P(S) + P(W| \sim S \& R)\, P(\sim S)} = .21$$

$P(R) = .4$

$P(S) = .2$



$P(W\|S \& R) = .95$
$P(W\|S \& \sim R) = .9$
$P(W\| \sim S \& R) = .9$
$P(W\| \sim S \& \sim R) = .1$

Explaining away

We just discovered something interesting!

$$P\left(S\,|\,W\right) = .35$$

$$P\left(S\,|\,W\,\&\,R\right) = .21$$

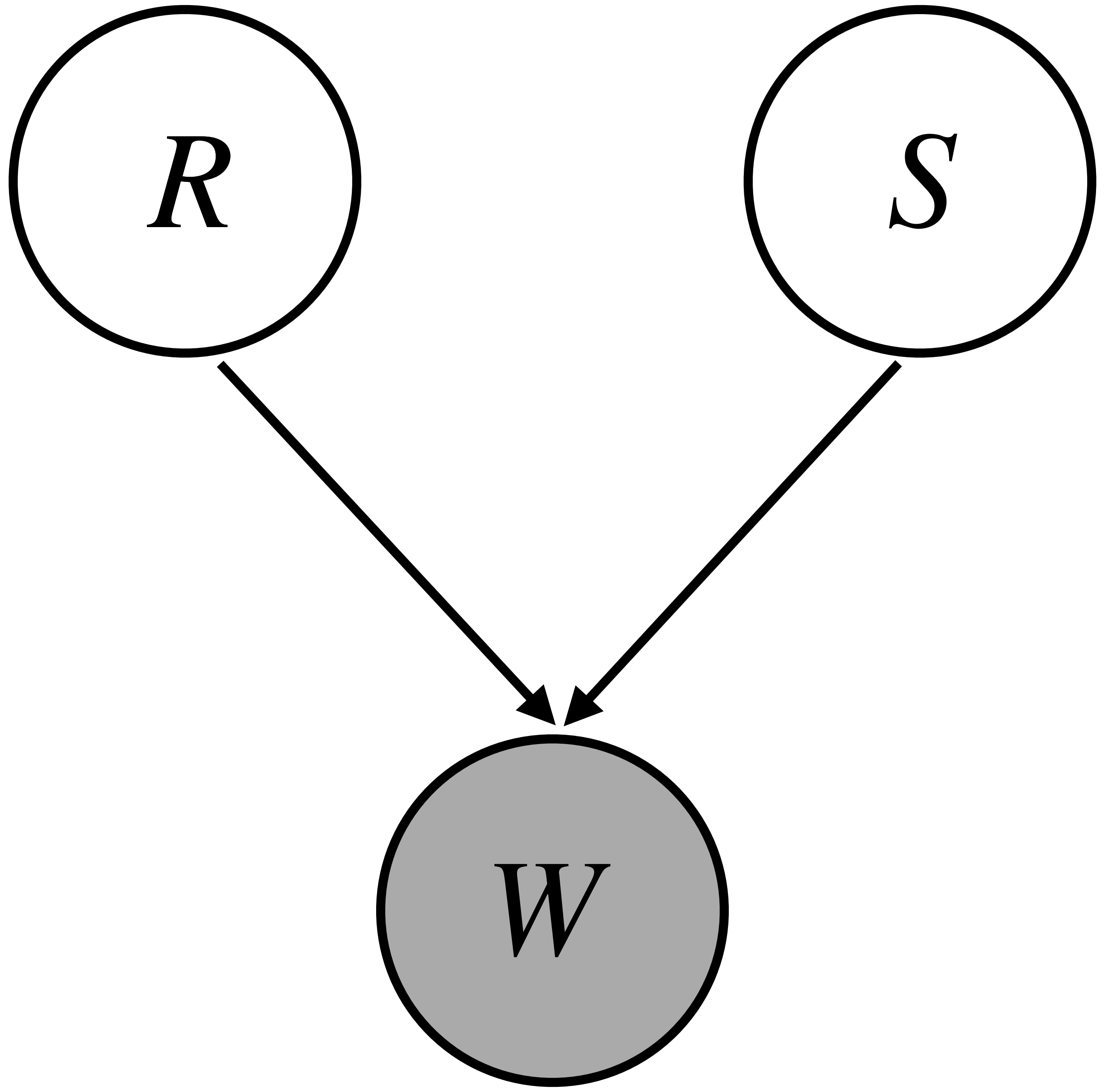
The sprinklers and the rain are independent of each-other.

But they are conditionally-dependent on each other through the wetness of grass

Rain **explains away** sprinklers as a cause of wet grass

$P(R) = .4$

$P(S) = .2$



Conditional independence

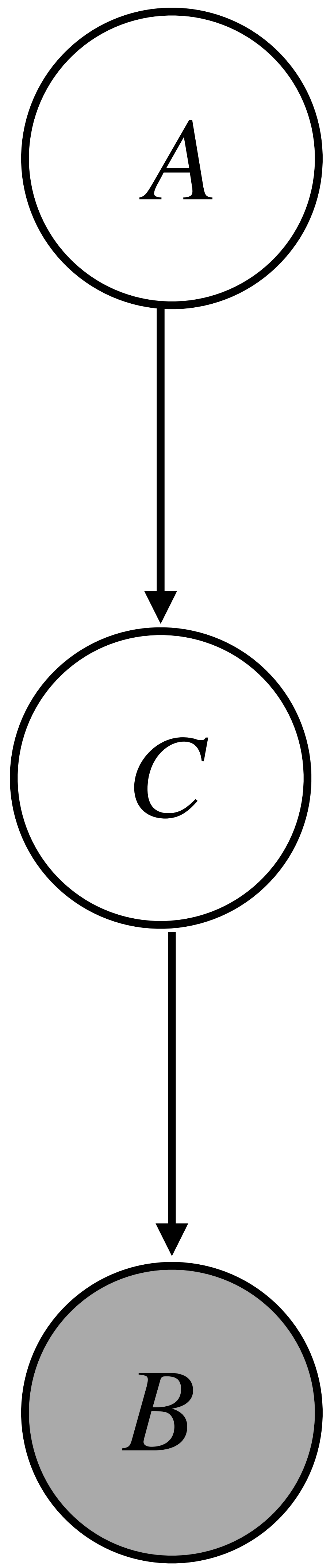
Events A and B are **independent** iff

$$P\,(A\ \&\ B) = P\,(A)\,P\,(B)$$

Events A and B are **conditionally independent** given event C iff

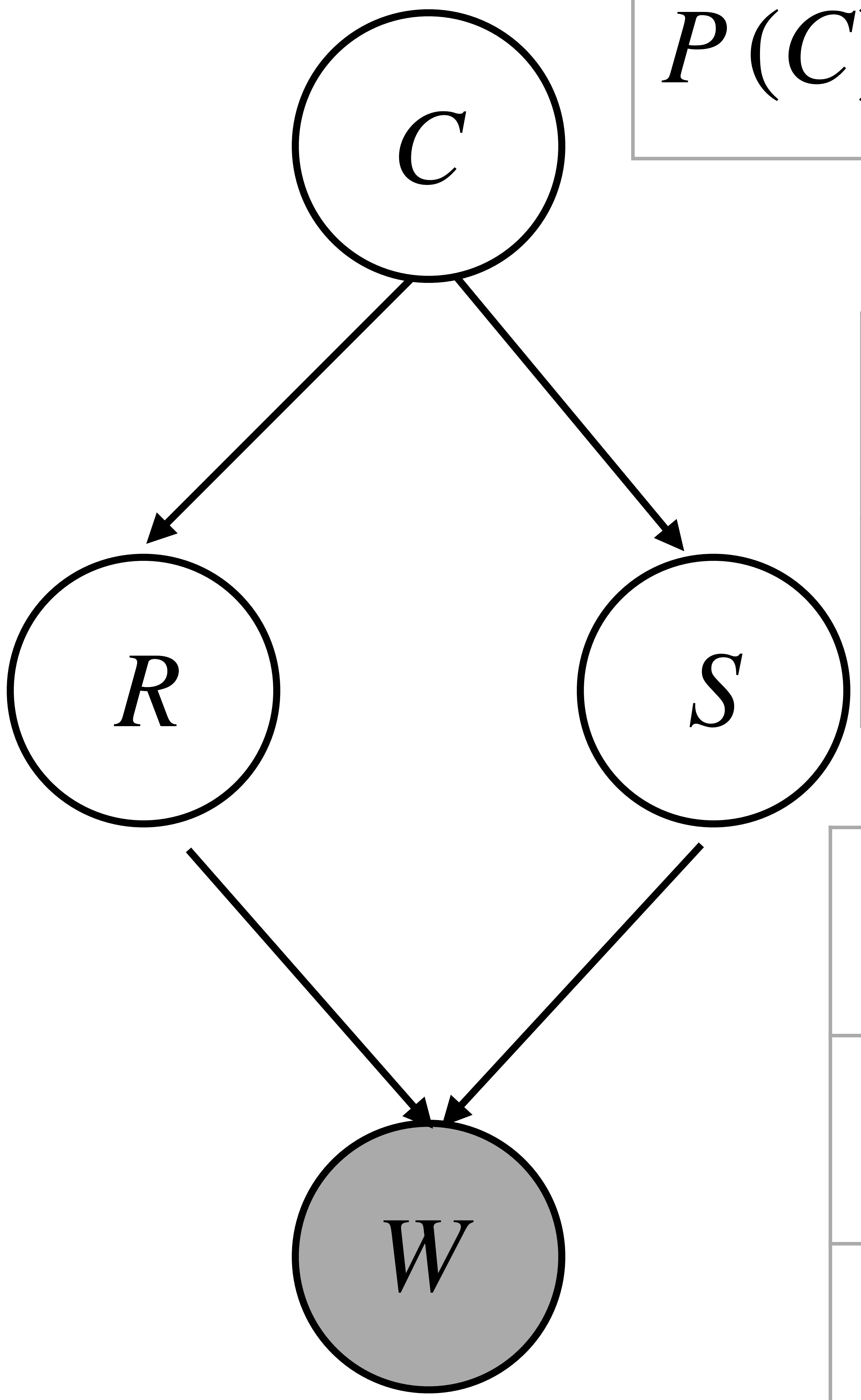
$$P\left(A \mid B\ \&\ C\right) = P\left(A \mid B\right)$$

In a graphical model, grand-children of a vertex are independent of their grandparents given their children



Conditional independence

$$P(C) = .5$$



$$P(R \mid C) = .4$$

$$P(R \mid \sim C) = .1$$

$$P(S \mid C) = .1$$

$$P(S \mid \sim C) = .5$$

$$P(W \mid S \,\&\, R) = .95$$

$$P(W \mid S \,\&\sim R) = .9$$

$$P(W \mid \sim S \,\&\, R) = .9$$

$$P(W \mid \sim S \,\&\sim R) = .1$$

Models at different levels

Read before class on Thursday, September 24, 2020

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychological Review*, *112*, 347—382.

- Read the introduction, Experiments 1-3, and the discussion and conclusion. Your goal should be to understand what the phenemon being modeled is, how the model works, and what the basic results are.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models.. *Developmental Science*, *10*, 307—321.

- You can skip the section on ontological kinds. Your goal should again be to understand what the model is doing and why it produces the results it does.

The primary goal this week is to think about the relationship between these two models. How are they the same? How are they different? Are there reasons to prefer one to the other? Are there some things that one does better than the other?

Learning by Bayesian inference

1. Bayesian inference provides a framework for causal learning
2. The size principle embodies an assumption about generating processes that leads to stronger inference
3. Graphical models are a powerful and flexible notation for describing Bayesian Models