Week 4 — Bayesian Generalization

Friday, May 22, 2026

Prof. Joseph Austerweil

Agenda

Welcome + Clusters walkthrough	0:00
The generalization problem	0:10
The Bayesian generalization framework + size principle	0:18
Rectangle game + number game	0:40
Break	1:08
Student presentation — Shohei	1:15
No Free Lunch	1:40
Hierarchical Bayes + close	1:50

Assignment 1 — Clusters

Clusters — what and when

Assignment 1: Gaussians, Categories, and Clusters

- Due **Fri Jun 5, 2026, 8:00 PM** — worth **7.5%**
- **Problem 1** — Gaussian-Gaussian conjugate model → *you can do this today* (it's Week 3 material)
- **Problem 2** — Gaussian mixture / categorization → Bayes' rule over two category Gaussians
- **Problem 3** — clustering → leans on this week + T3 Ch 5

Clusters — which notebook

- **`clusters.ipynb` is the canonical stencil** — the GenJAX path
- `clusters_python.ipynb` and `clusters_nosoln.Rmd` — non-GenJAX paths, *same math, same credit*
- Matlab available on request
- **“Open in Colab”** links are live on the assignments page — one click, no local setup
- Read the **assignment PDF first** — it has the problem statements and all the math

Assignments page: hml.chibatech.dev/assignments.html

Where we are

Welcome + Clusters walkthrough	0:00
The generalization problem	0:10
The Bayesian generalization framework + size principle	0:18
Rectangle game + number game	0:40
Break	1:08
Student presentation — Shohei	1:15
No Free Lunch	1:40
Hierarchical Bayes + close	1:50

The generalization problem

Chibany’s lunches

Chibany has had three tonkatsu lunches this week. Each weighed about **500 g**.

Today a lunch arrives weighing **480 g**. Is it tonkatsu?

What about **700 g**? What about **350 g**?

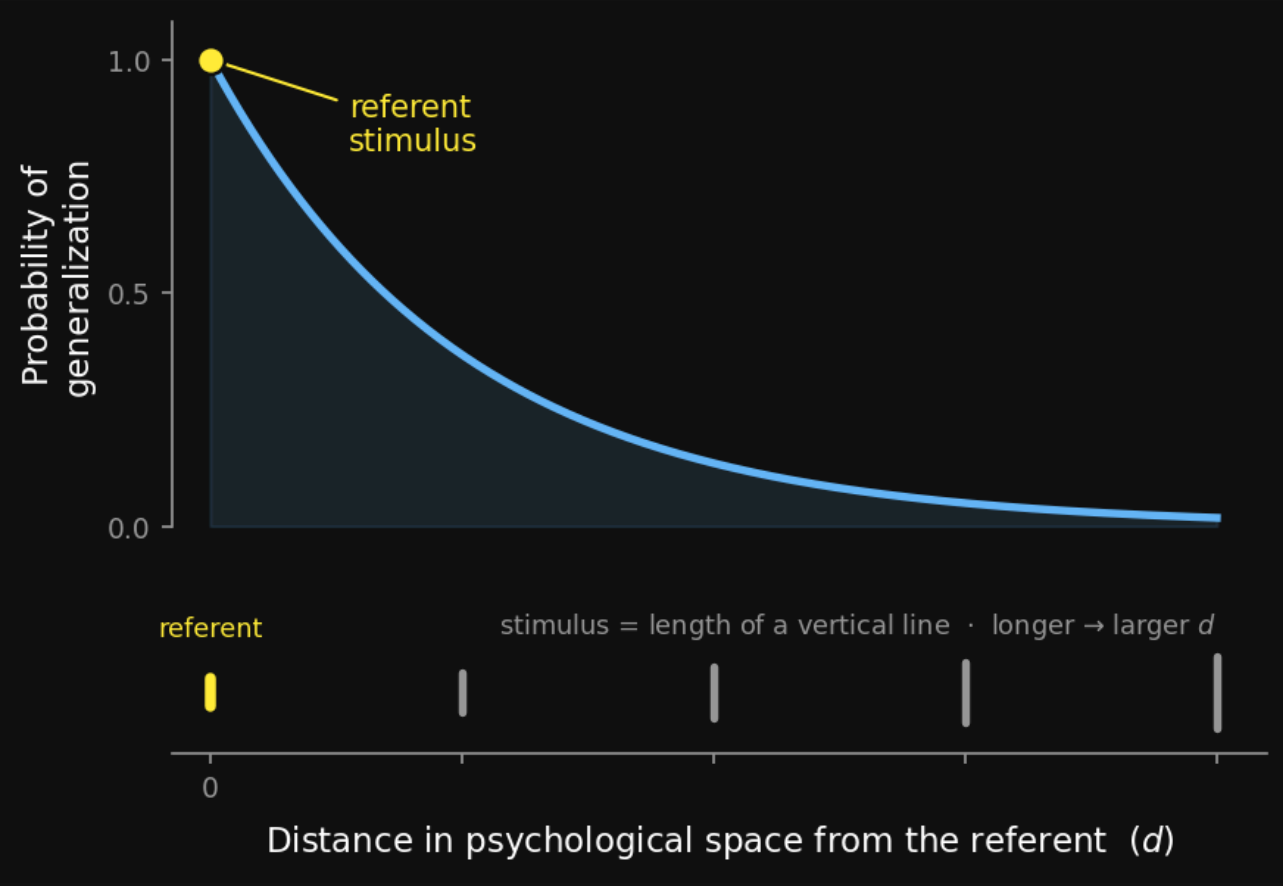
What just happened

Generalization — deciding when to extend a property from observed examples to a *novel* stimulus.

- No two stimuli are ever identical → you *must* generalize to act at all
- It is everywhere: word learning, categorization, object recognition, property induction, stereotypes
- It is the core problem of inductive inference

Shepard’s universal law

Shepard (1987) — across species and domains, the probability of generalization **decays exponentially** with distance in **psychological space** (*the perceived distance, after the mind represents the stimulus — not raw physical space*).



One law, one equation:

$$g(d) = e^{-d}$$

- g — probability of generalization
- d — distance in psychological space from the referent

The same curve holds across **species, senses, and stimulus types** — pitch, colour, line length, faces. Shepard called it *universal* for that reason.

But the law is **descriptive**: it says generalization decays exponentially, not *why*. Today’s Bayesian framework will **derive** that exponential — it falls out of the posterior.

Poll — Shepard’s universal law

Shepard’s universal law of generalization says generalization …

- **A.** decays exponentially in psychological space
- **B.** decays exponentially in stimulus space
- **C.** grows exponentially in psychological space
- **D.** grows exponentially in stimulus space

Poll — answer

A. Decays exponentially in psychological space.

“Stimulus space” is the trap — generalization isn’t governed by physical distance but by *perceived* distance. A model of generalization therefore needs a model of the **psychological space**. That is exactly what the Bayesian framework supplies next.

Where we are

Welcome + Clusters walkthrough	0:00
The generalization problem	0:10
The Bayesian generalization framework + size principle	0:18
Rectangle game + number game	0:40
Break	1:08
Student presentation — Shohei	1:15
No Free Lunch	1:40
Hierarchical Bayes + close	1:50

The Bayesian generalization framework

The idea — concepts as hypotheses

Instead of measuring distance directly, posit a **space of candidate concepts** and let Bayes do the generalizing.

- A concept = a **set of stimuli** that share the property
- Shepard's term: a **consequential subset** — the subset of things the property "applies to"
- A **feature** is a hypothesis too: "has stripes" picks out a set of stimuli — generalization is just asking which feature-sets the examples imply
- Generalization becomes: *which concepts are consistent with the examples, and do they contain the new stimulus?*

Notation lock-in

- h — a **hypothesis**: one candidate concept, i.e. a *set* of stimuli
- $\mathcal{H}$ — the **hypothesis space**: all candidate h
- $X = \{x_1, \ldots, x_n\}$ — the observed **examples** of the concept
- y — a **novel stimulus** we must judge
- C — the (unknown) true concept

The three ingredients

Prior $p(h)$ — domain knowledge: which concepts are *natural* before any data.

Likelihood $p(X \mid h)$ — how probable the examples are if h is the true concept.

Posterior $p(h \mid X)$ — belief in h after seeing the examples:

$$p(h \mid X) \;\propto\; p(X \mid h)\, p(h)$$

The hypothesis space IS a prior

Choosing $\mathcal{H}$ is already a **strong prior**.

Any concept not in $\mathcal{H}$ has $p(h) = 0$ — it can never be learned, no matter the data.

So: a learner's inductive bias lives in *which hypotheses it even considers*.

Generalization = a posterior-weighted vote

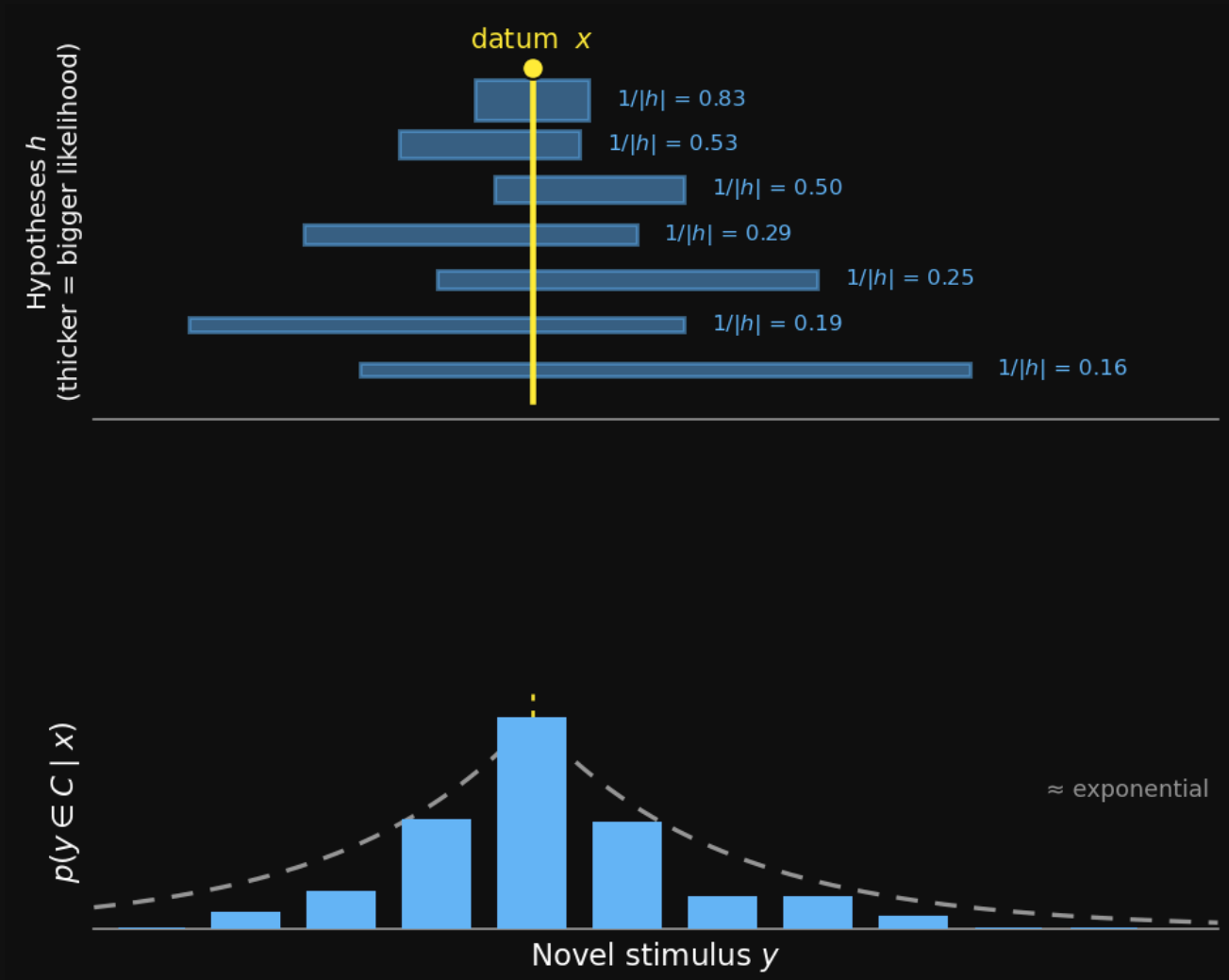
Probability that the novel stimulus y has the property:

$$p(y \in C \mid X) \;=\; \sum_{h \in \mathcal{H}} \mathbf{1}[\, y \in h \,]\; p(h \mid X)$$

Every hypothesis votes. Its vote is **its posterior weight**, and it votes “yes” only if it *contains* y.

$\mathbf{1}[\, y \in h \,]$ — the indicator: 1 if y is in the set h, else 0.

One datum → the posterior-weighted vote



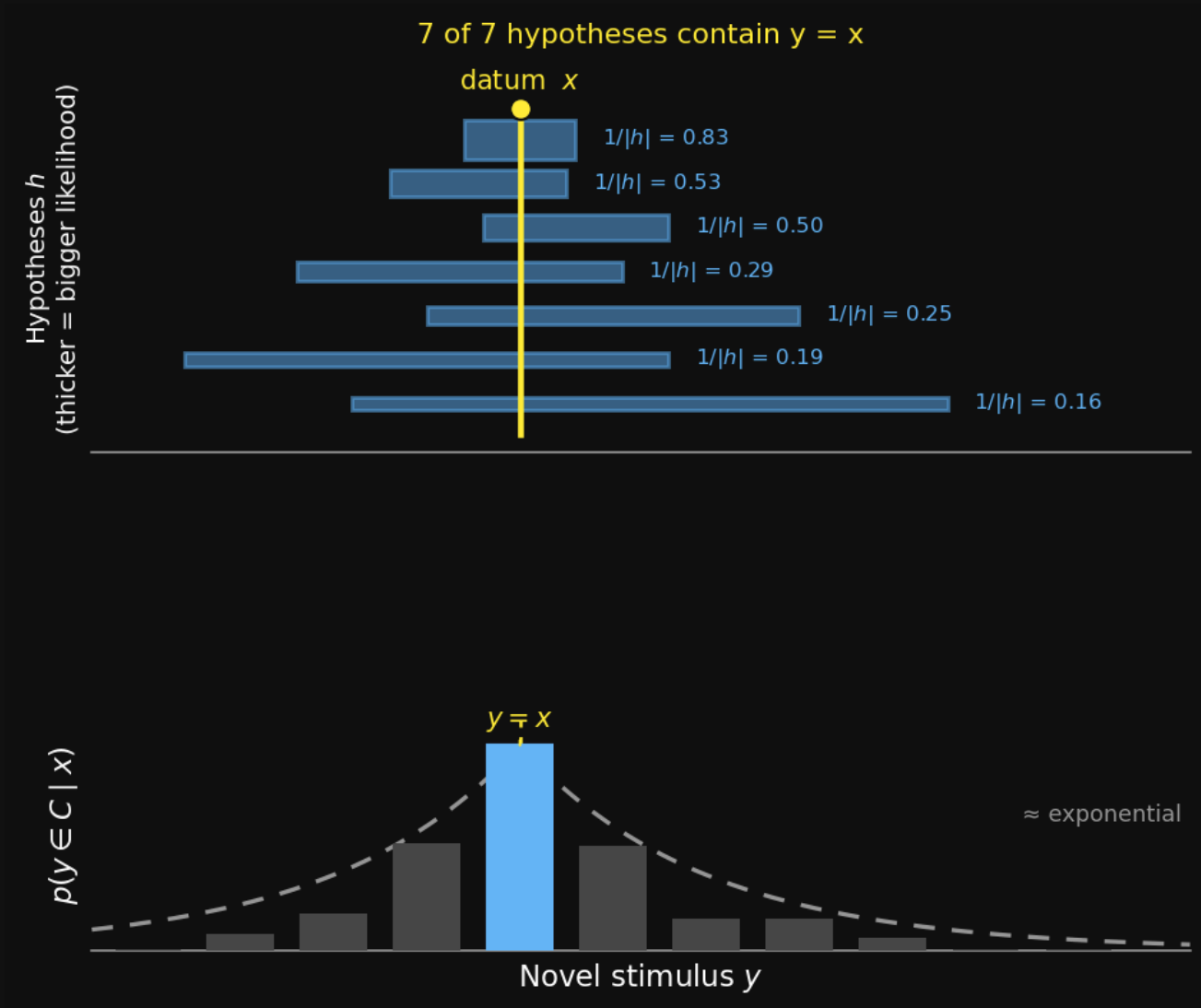
The hypotheses. Observe one datum x. Every interval **containing** x is a live hypothesis h. Bar thickness $= 1/|h|$ (the strong-sampling likelihood); flat prior, so posterior $\propto$ likelihood.

The vote. For each candidate y, sum the posterior of every h that contains it:

$$p(y \in C \mid x) = \sum_h \mathbf{1}[\, y \in h \,]\, p(h \mid x)$$

Next three slides: walk y outward — x, $x{+}1$, $x{+}2$.

Vote for $y = x$



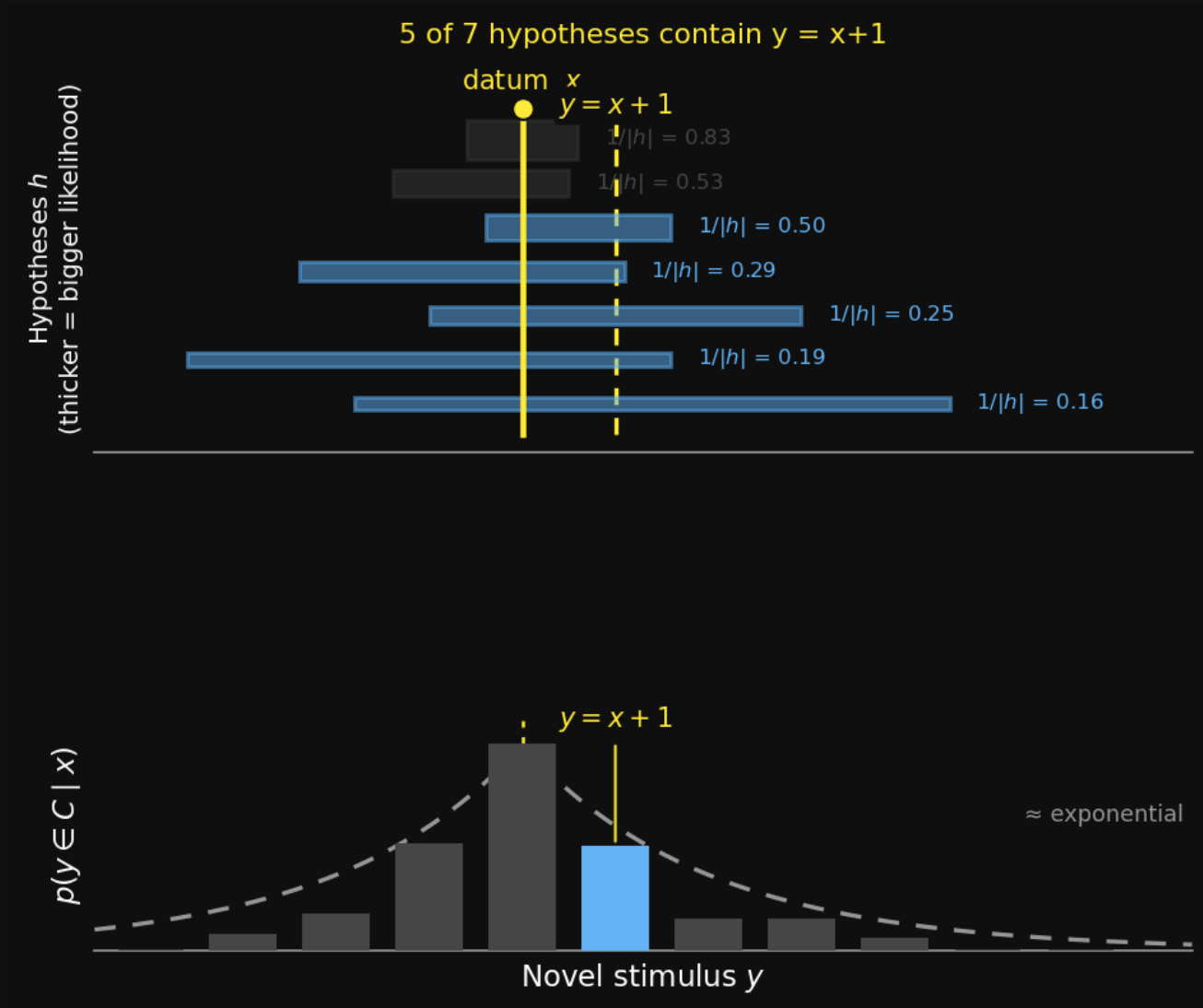
The question. Does the property generalize to $y = x$, the observed datum itself?

Who votes. Every hypothesis was built to contain x, so **all 7 of 7 vote "yes"**.

The bar. The vote sums *all* the posterior weight — $\sum_h p(h \mid x) = 1$. The **tallest bar the gradient can have**.

The baseline. This peak is what every other y is measured against — step away and the bar can only fall.

Vote for $y = x + 1$



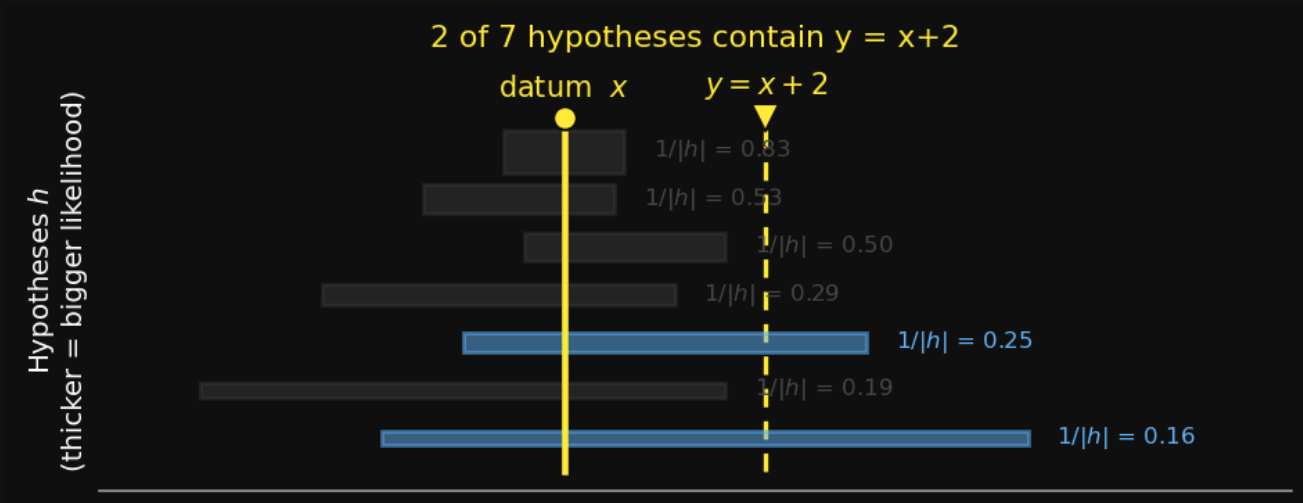
The question. Step one unit out: does the property generalize to $y = x + 1$?

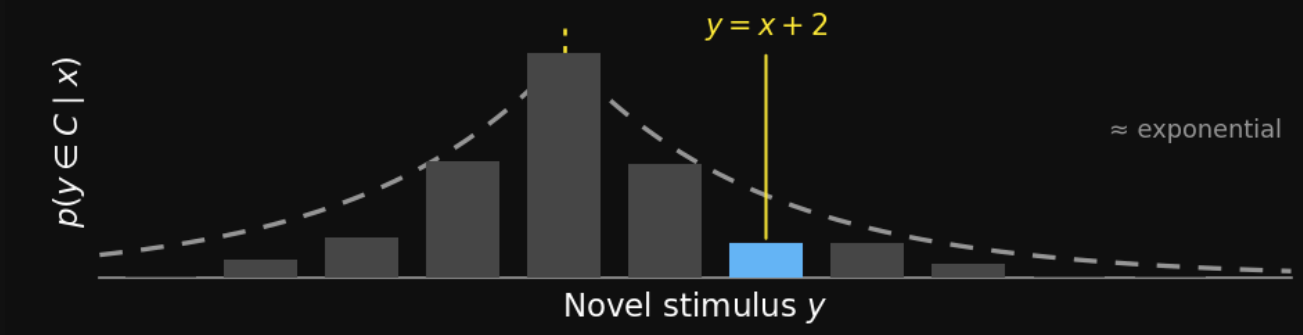
Who drops. The two smallest intervals no longer reach y — they grey out and vote “no”. **5 of 7 still vote “yes”.**

The bar. Fewer hypotheses contribute → a **shorter** bar than at x.

Why it falls fast. The dropouts are the **smallest** intervals — by the size principle the **heaviest**-weight votes. Losing those first is why the gradient decays steeply near x.

Vote for $y = x + 2$





The question. Two steps out: does the property generalize to $y = x + 2$?

Who's left. Only the 2 widest intervals still reach y — and those are the *least* likely (thinnest bars). **2 of 7 vote.**

The bar. Few votes, and only low-weight ones → the bar is short.

The payoff. Sweep y across every point and the bar heights trace an **approximately exponential** decay — Shepard's universal law, **derived** from the model, not assumed.

The size principle

Where did the examples come from?

The likelihood $p(X \mid h)$ depends on *how you assume the examples were generated*. Two assumptions:

Weak sampling — examples generated some other way, then *labeled* by whether they fall in h.

Strong sampling — each example drawn *uniformly at random from within* h.

The two likelihoods

Weak sampling

$$p(X \mid h) = \begin{cases} 1 & \text{all } x_i \in h \\ 0 & \text{else} \end{cases}$$

Size-blind: a hypothesis either contains the data or it doesn't.

Does not depend on $|h|$.

Strong sampling

$$p(x \mid h) = \frac{1}{|h|} \quad \Rightarrow \quad p(X \mid h) = \left(\frac{1}{|h|}\right)^n$$

Each example drawn *uniformly from inside* h.

Smaller $|h|$ → higher likelihood, exponentially so in n.

$|h|$ — the size of hypothesis h (how many stimuli it contains).
n — number of examples.

The size principle

Under strong sampling: **smaller hypotheses get higher likelihood** — and *exponentially* more so as the number of examples n grows.

$$p(X \mid h) = \left(\frac{1}{|h|}\right)^n$$

So a small $|h|$ wins — and **wins faster with more data**, because the exponent n magnifies any size advantage.

The mechanism behind both games today.

Why — the suspicious coincidence

You see the examples $\{60, 80, 10, 30\}$.

- If the concept is **“multiples of 10”** — unremarkable, that’s just what its members look like.
- If the concept is **“even numbers”** — it’s a **suspicious coincidence** that not one of the four was 2, 4, 6, 8, …

Strong sampling *penalizes* the big hypothesis for failing to predict the tight clustering you actually saw.

Poll — strong sampling

What is **strong sampling**?

- **A.** Each stimulus is generated uniformly at random from the true hypothesis
- **B.** A stimulus has probability one given the true hypothesis
- **C.** Larger hypotheses are given smaller prior probability
- **D.** Smaller hypotheses are given smaller prior probability

Poll — answer

A. Uniformly at random from the true hypothesis.

B describes *weak* sampling — membership gives likelihood 1, regardless of size. **C / D** describe a *prior* over hypotheses; the size principle is about the **likelihood**, not the prior. Strong sampling = "the examples were drawn *from inside* the concept" — and that is what makes size matter.

Where we are

Welcome + Clusters walkthrough	0:00
The generalization problem	0:10
The Bayesian generalization framework + size principle	0:18
Rectangle game + number game	0:40
Break	1:08
Student presentation — Shohei	1:15
No Free Lunch	1:40
Hierarchical Bayes + close	1:50

Rectangle game — continuous concepts

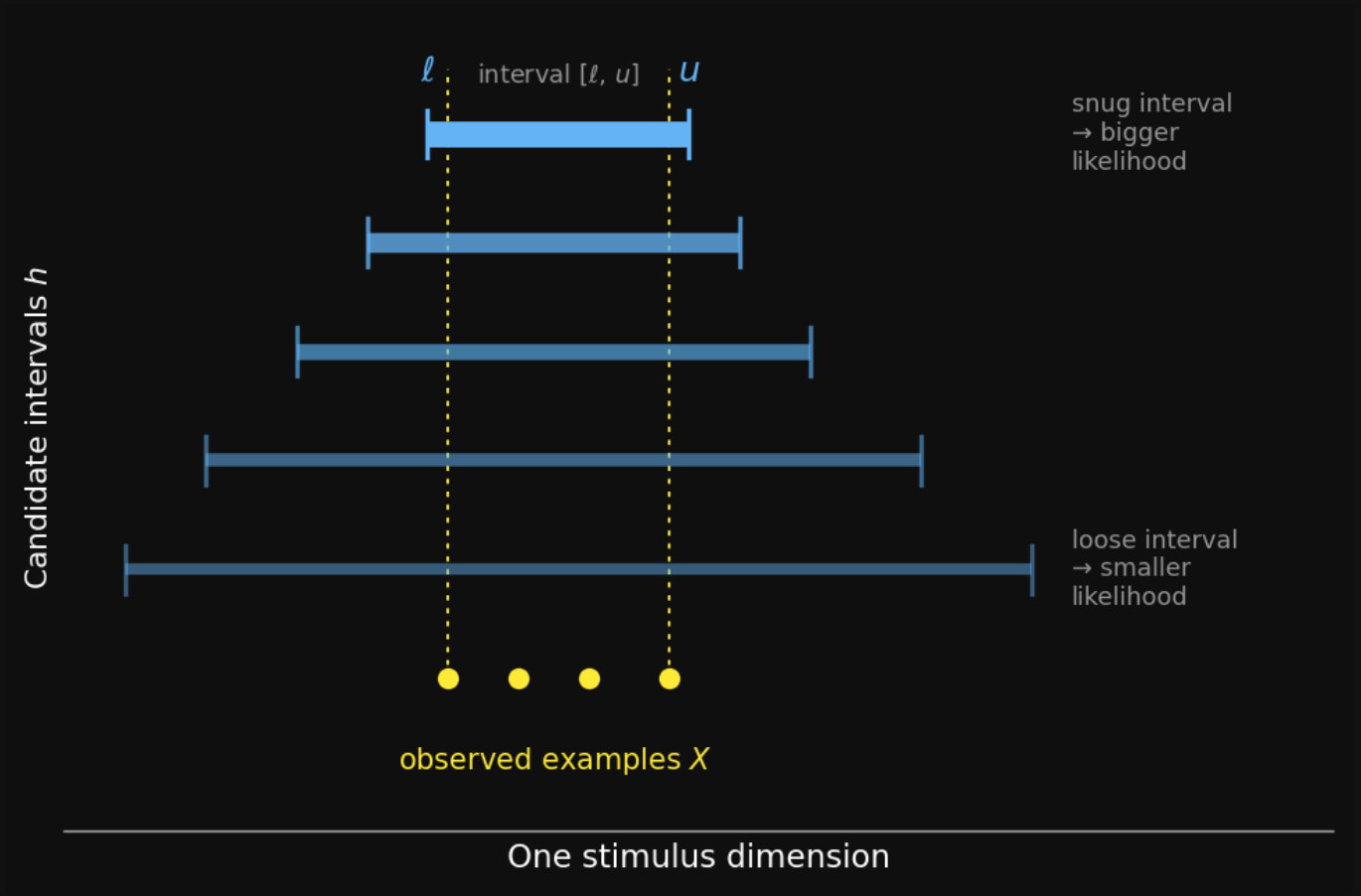
Gnarbles

A **gnarble** is a rectangle whose dimensions fall in some interval.

Tenenbaum's cover story: a **2-D concept** is an axis-aligned rectangle in a feature space — e.g. *healthy levels* of insulin × cholesterol.

You observe a few examples drawn from inside the rectangle. **Which other points are gnarbles?**

Start in 1-D



Concept = an **interval** $[\ell, u]$ on one dimension.

- $\mathcal{H}$ = *all* intervals
- Strong sampling → likelihood $\propto \left(\frac{1}{u-\ell}\right)^n$
- Hypothesis size = the **length** $u - \ell$

Every interval that contains the data is a hypothesis; the **snug** ones get the most posterior weight.

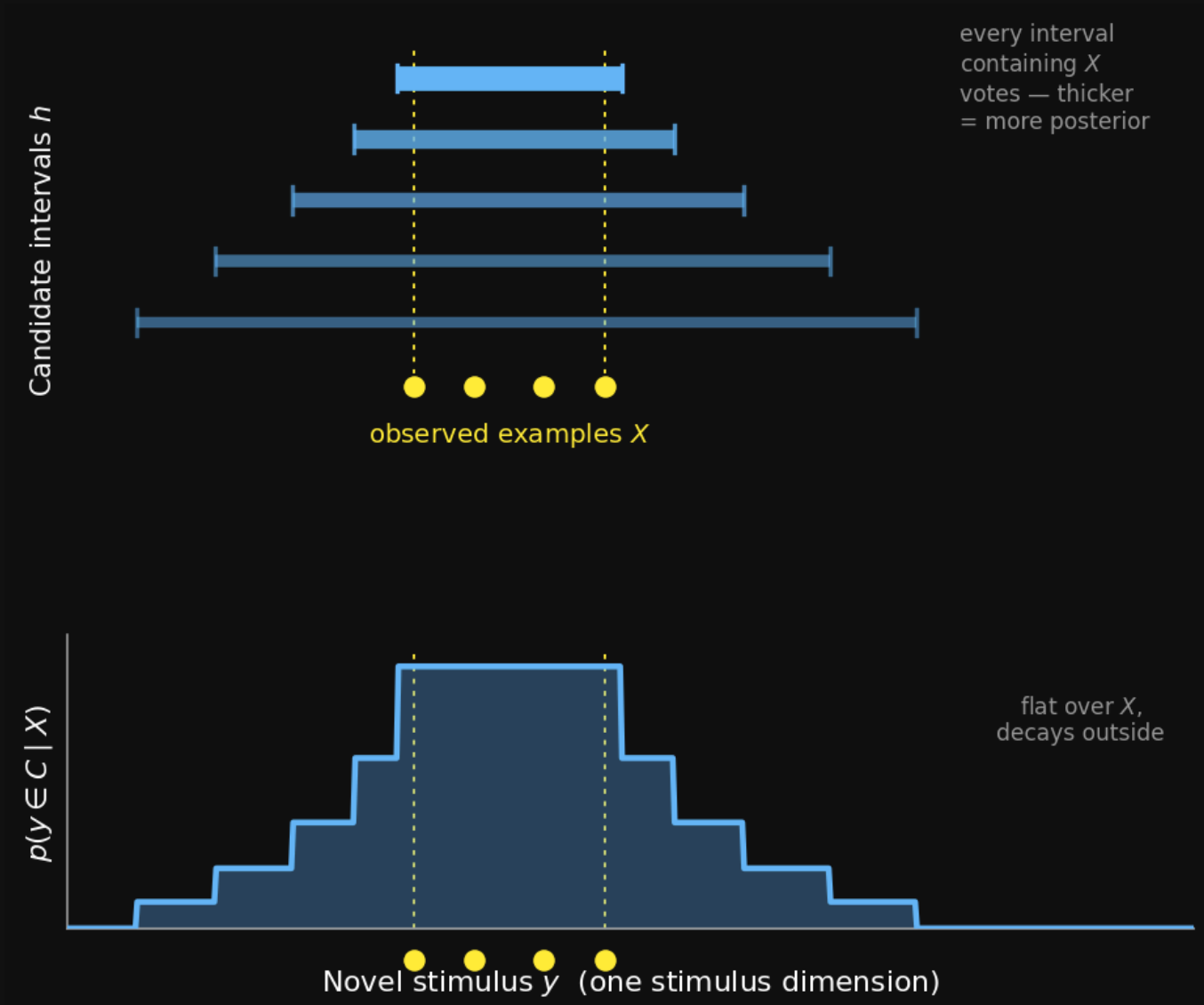
The generalization gradient

Plot $p(y \in C \mid X)$ as y moves along the dimension:

- **High** inside the range of the examples
- **Decays** as y moves outside that range

The decay is *exponential* — Shepard's universal law, now **derived** rather than assumed.

The 1-D gradient — built from votes



The same posterior-weighted vote as the number-line construction — now with **several examples** X.

Who votes. Every interval that **contains all of** X is a live hypothesis; a snugger interval is thicker (more posterior).

The gradient. Sum the posterior of every interval that contains y:

$$p(y \in C \mid X) = \sum_h \mathbf{1}[y \in h]\, p(h \mid X)$$

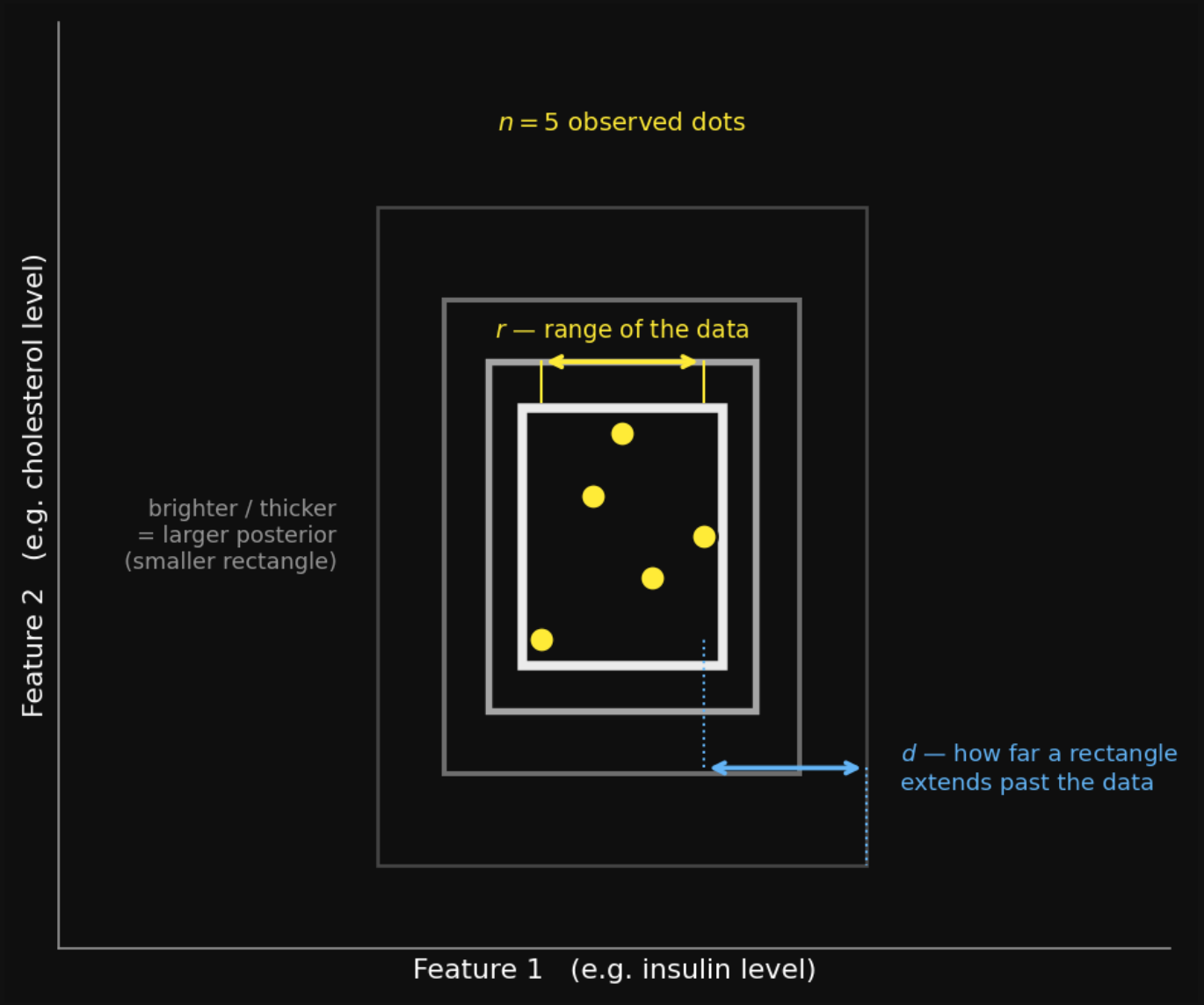
Flat across the data, decaying outside — the gradient *is* the posterior-weighted vote, one y at a time.

More examples → tighter generalization

- **One example** → broad, diffuse generalization (many interval sizes survive)
- **Many examples** → tight generalization, hugging the data range

Why: the size principle. With large n, big intervals lose likelihood *exponentially* fast — only the small, snug intervals keep posterior mass.

Into 2-D — the rectangle game



Same machinery, one dimension up: a concept is an axis-aligned **rectangle**.

- Observe n dots inside the true rectangle
- Every rectangle that **encloses all** n is a hypothesis
- Smaller rectangle → bigger likelihood (brighter / thicker)

r — the range the data spans.

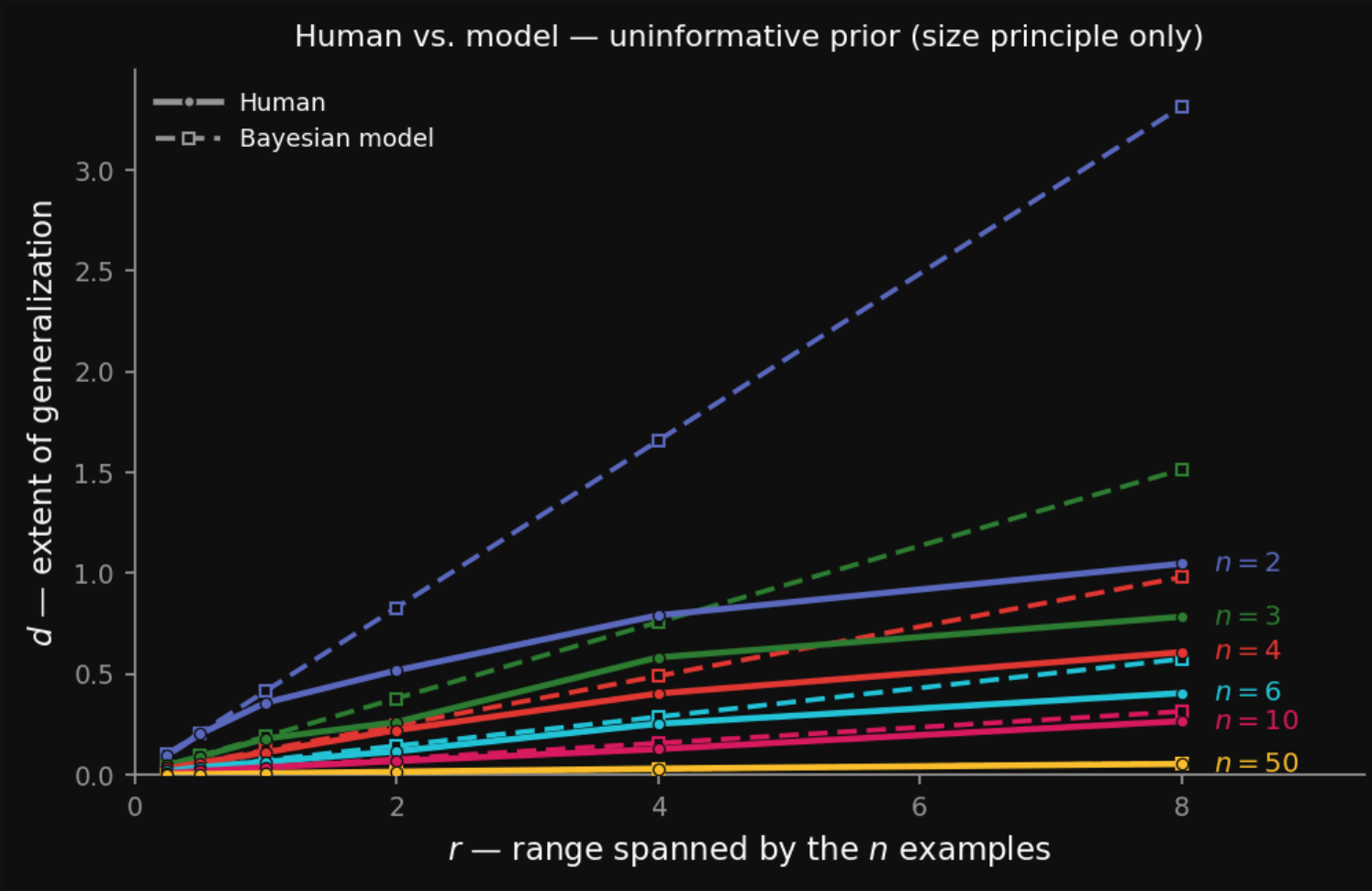
d — how far a rectangle extends past that range.

The rectangle experiment

Tenenbaum (1999) ran this as a behavioural experiment.

- On each trial, subjects saw n **dots** drawn from “an arbitrary rectangle of healthy insulin / cholesterol levels”
- They drew the rectangle they thought the dots came from
- n varied from **2 to 50**; the data range r varied across trials
- The measure: d — how far past the data range r the drawn rectangle extends

The result — d vs. r, by n



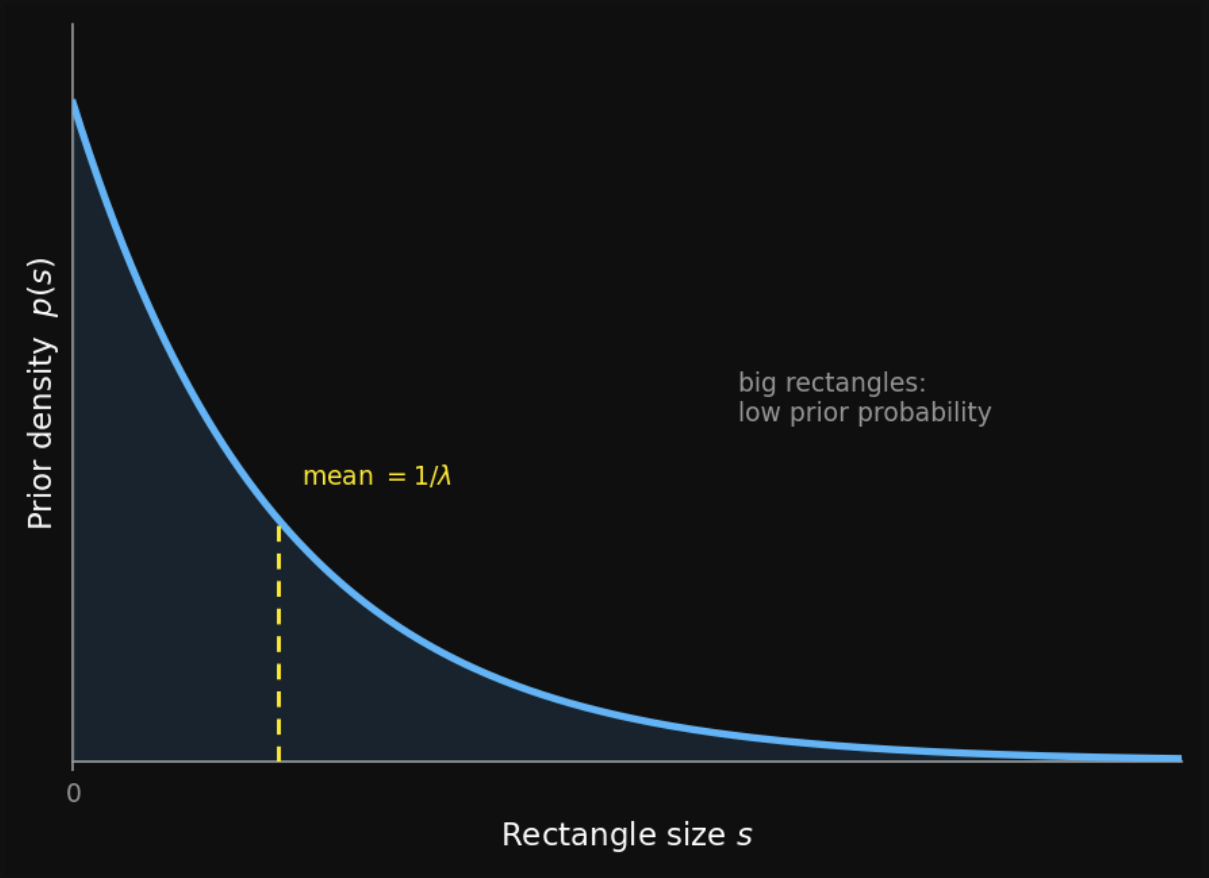
Solid = human, dashed = model. One colour per n.

The human pattern. Fewer examples → generalize further ($n = 2$ on top); d rises with r but **saturates**.

The model — likelihood only. The size principle with a *flat* (uninformative) prior. It captures the n-ordering…

…but the curves run **straight** — the model **over-extends**, badly for small n and large r. It misses the human saturation.

One fix — an exponential prior



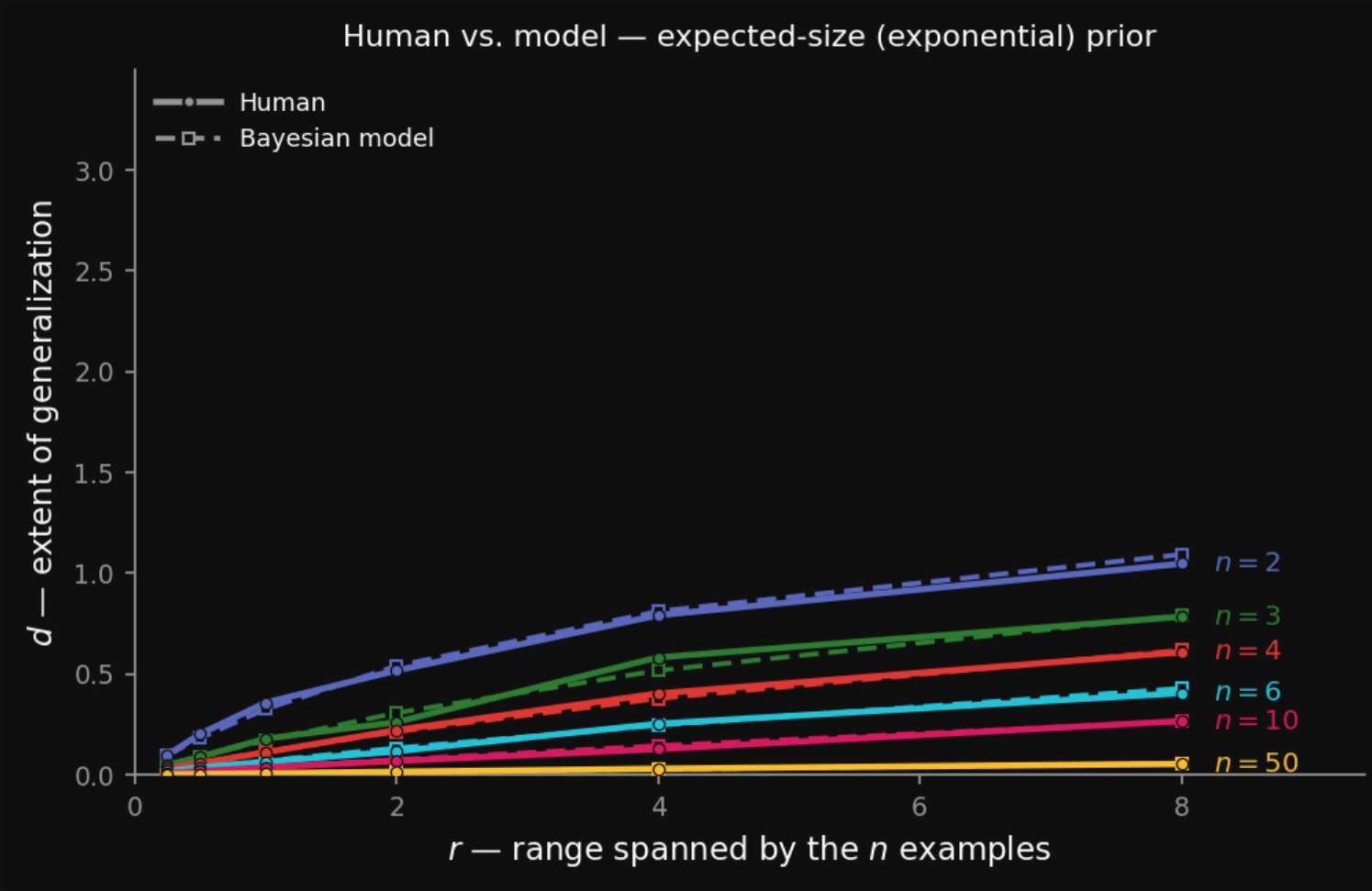
A flat prior lets the rectangle run **straight** — it **over-extends**. The fix: a prior that makes large rectangles less likely.

The exponential distribution — our first new distribution today. For a size $s \geq 0$:

$$p(s) = \lambda\, e^{-\lambda s}$$

- Always decreasing — **small s favoured**
- One parameter $\lambda > 0$; the mean is $1/\lambda$
- Larger $\lambda \rightarrow$ faster decay $\rightarrow$ stronger pull toward small rectangles

With the prior — the fit



Same axes, same data — now the model carries the exponential prior over rectangle size.

The dashed curves bend. The straight, over-extending lines collapse onto the human curves — d now **saturates** with r, just as people do.

Likelihood (size principle) **and** prior together — neither alone — give the fit.

Same lesson for every paper: lay the **model on top of the human data** and read off where it bends.

The rectangle game in one line

The continuous concept learner is **just the framework equation** with $\mathcal{H}$ = intervals / rectangles.

$$p(y \in C \mid X) = \sum_{h} \mathbf{1}[\, y \in h \,]\; p(h \mid X)$$

Nothing new — only the choice of $\mathcal{H}$.

Number game — discrete concepts

The number game

A simple task (Tenenbaum, 1999):

- I have a concept — a set of numbers between 1 and 100
- You see one or more **“yes” examples**
- You judge: is some other number a “yes”?

Watch what your own judgments do as examples come in.

What people actually do

Human generalization judgments (Tenenbaum's $N = 20$ subjects):

- Examples $\{60\} \rightarrow$ **diffuse** similarity: many numbers get moderate "yes"
- Examples $\{60, 80, 10, 30\} \rightarrow$ sharp **"multiples of 10"**
- Examples $\{60, 52, 57, 55\} \rightarrow$ sharp **"numbers near 60"**

One example $\rightarrow$ graded. Four examples $\rightarrow$ a crisp **rule**.

Two things to explain

1. Generalization can look **similarity-based (graded)** *or* **rule-based (all-or-none)** — and people switch between them.
2. People learn a concept from **just a few examples**.

One model — the Bayesian framework — produces *both*, with no extra machinery.

The discrete hypothesis space

$\mathcal{H}$ for the number game has two kinds of hypothesis:

Mathematical properties (~24) — even, odd, primes, squares, cubes, multiples of k, powers of k.

Magnitude intervals — "numbers in $[a, b]$": e.g. 10–20, 30–45.

The **prior** $p(h)$ weights these families against each other.

Size principle, by the numbers

Two candidate concepts for numbers in 1–100:

- **Multiples of 2** — 50 numbers → $p(x \mid h) = \frac{1}{50} = 2\%$ each
- **Multiples of 10** — 10 numbers → $p(x \mid h) = \frac{1}{10} = 10\%$ each

One example: $x = 60$

$$p(60 \mid \text{mult-2}) = \frac{1}{50} \qquad p(60 \mid \text{mult-10}) = \frac{1}{10}$$

Multiples of 10 is already **5×** more likely — but only 5×. With one example, many hypotheses stay in contention → **graded** generalization.

Four examples: $\{10, 30, 60, 80\}$

$$p(X \mid \text{mult-2}) = \left(\tfrac{1}{50}\right)^4 \approx 1.6 \times 10^{-7}$$

$$p(X \mid \text{mult-10}) = \left(\tfrac{1}{10}\right)^4 = 10^{-4}$$

Now multiples of 10 is **~625×** more likely. The 5× edge got raised to the 4th power → a crisp **rule**.

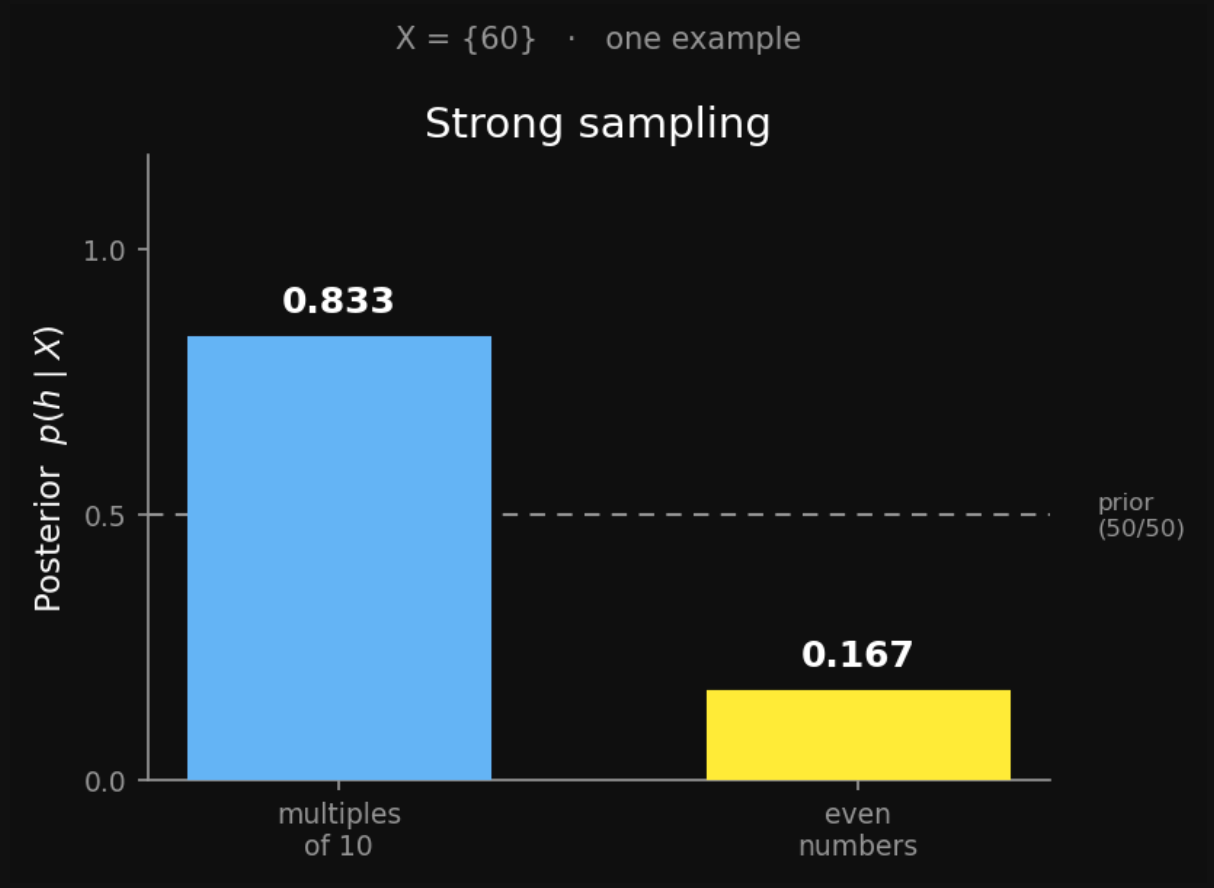
From likelihood to posterior

Likelihoods aren’t beliefs yet. Take a **two-hypothesis** model — $\mathcal{H} = \{$ **multiples of 10**, **even numbers** $\}$, both containing every example — with a flat prior, and turn the crank:

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

Next: the posterior under **strong** vs. **weak** sampling, for $X = \{60\}$ then $X = \{60, 80, 10, 30\}$.

Strong sampling — one example



The data: $X = \{60\}$ — a single example. Strong sampling:

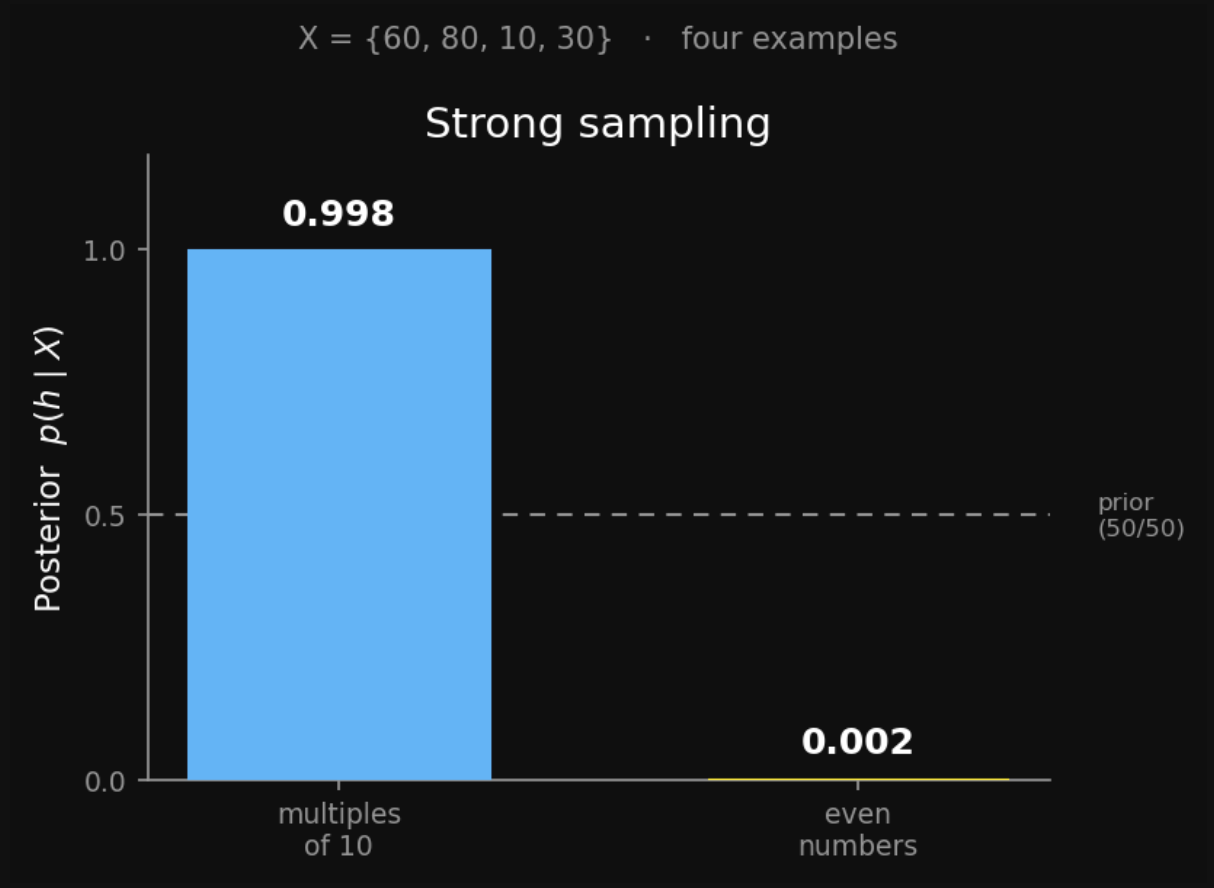
$$p(h \mid X) \propto \left(\frac{1}{|h|}\right)^{1} \cdot 0.5$$

- mult-10: likelihood $\frac{1}{10}$ · even: $\frac{1}{50}$
- Normalize the two → **0.83 vs. 0.17**

Read it: one example already **tilts** belief toward the smaller hypothesis — but only gently. A 5:1 likelihood ratio is not decisive, so plenty of posterior mass still sits on "even numbers".

This is the **graded** regime — belief shifts, but no rule yet.

Strong sampling — four examples



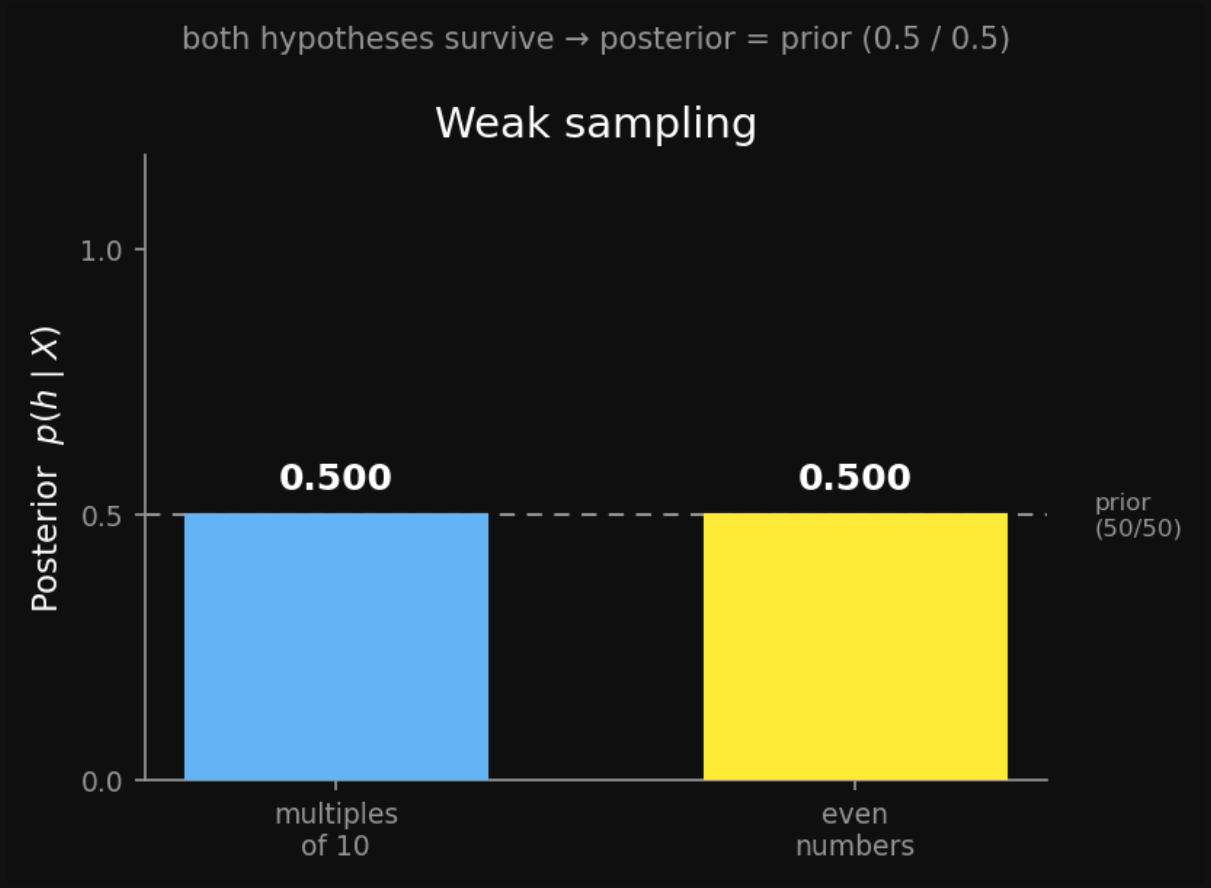
$X = \{60, 80, 10, 30\}$, strong sampling:

$$p(h \mid X) \propto \left(\frac{1}{|h|}\right)^4 \cdot 0.5$$

- The $5{:}1$ ratio is now $5^4 = 625{:}1$
- Normalize → **0.998 vs. 0.002**

Four “even” numbers, none of them odd-looking — that’s the **suspicious coincidence**. Strong sampling all but rules “even” out.

Weak sampling — eliminate, but don’t rank



Weak-sampling likelihood is 1 **for any h that contains the data**, 0 otherwise — *no $|h|$ dependence*.

What it can do: a datum *outside h* kills it — likelihood 0. Weak sampling **does** move the posterior, by **ruling hypotheses out**.

What it can’t do: among the hypotheses that *all still contain the data*, it has no preference — every survivor keeps likelihood 1, so the posterior over them stays at the **prior**.

Here both hypotheses contain every example — nothing is ruled out:

$$p(h \mid X) = \frac{1 \cdot 0.5}{1 \cdot 0.5 + 1 \cdot 0.5} = 0.5$$

So weak sampling can’t see the suspicious coincidence — it can’t tell “multiples of 10” from “even numbers”.

The number game in one line

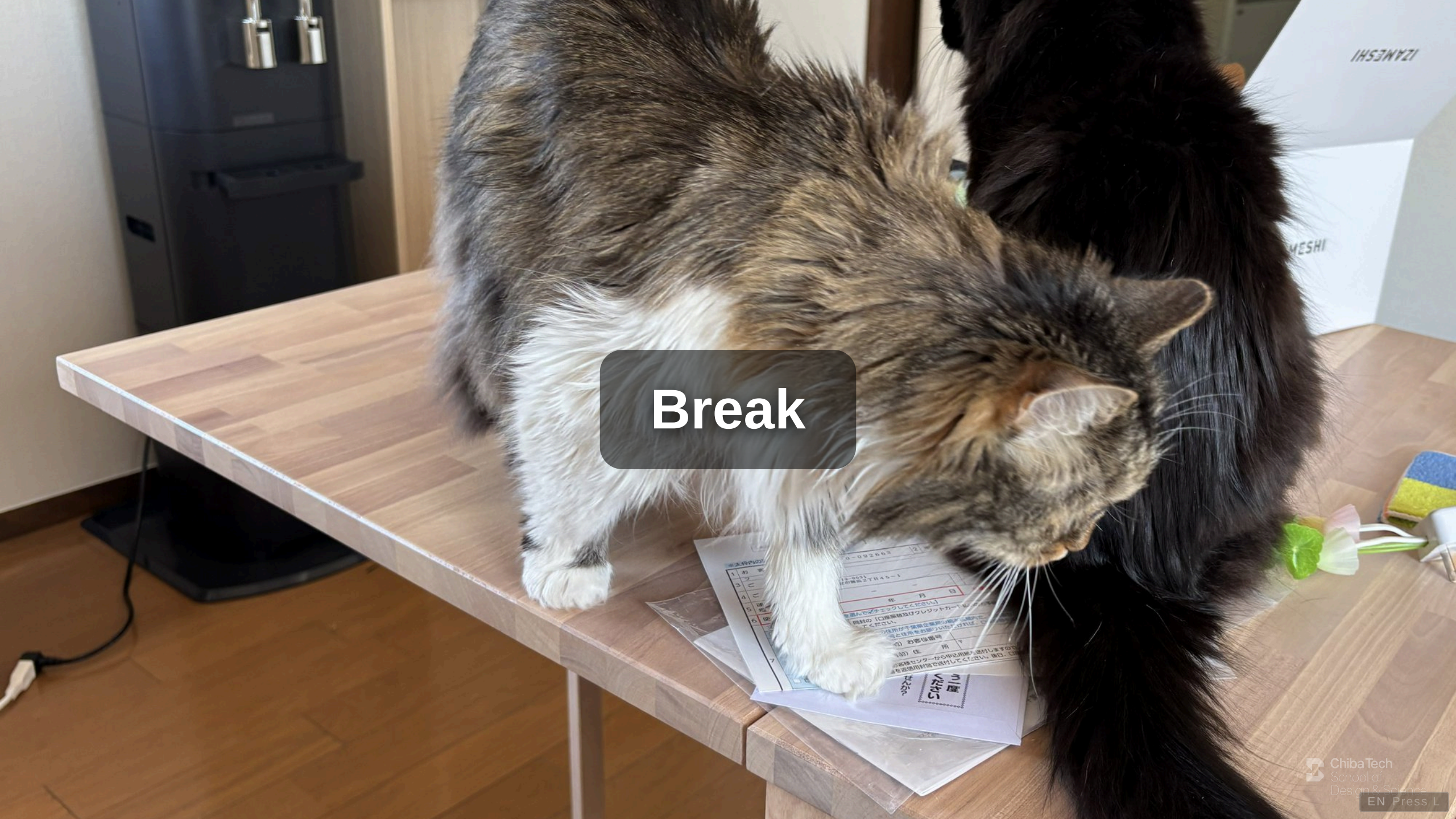
Same equation as the rectangle game — $\mathcal{H}$ is now **discrete**.

$$p(y \in C \mid X) = \sum_{h} \mathbf{1}[\, y \in h \,]\; p(h \mid X)$$

Graded vs. rule-like generalization both **fall out of the posterior** — no extra mechanism. The size principle's exponent does the switching.

Where we are

Welcome + Clusters walkthrough	0:00
The generalization problem	0:10
The Bayesian generalization framework + size principle	0:18
Rectangle game + number game	0:40
Break	1:08
Student presentation — Shohei	1:15
No Free Lunch	1:40
Hierarchical Bayes + close	1:50


Break

Bridge — Shohei's paper

Up next: **Shohei** presents **Tenenbaum & Xu (2000), *Word learning as Bayesian inference***.

A child hears *"this is a dax"* pointing at three Dalmatians. Is a poodle a dax? A cat?

This is the **number game**, with $\mathcal{H}$ = candidate word meanings (subordinate / basic-level / superordinate). The **size principle** explains why three subordinate examples → a subordinate meaning.

Watch for the size principle doing the work.

Student presentation — Shohei

Where we are

Welcome + Clusters walkthrough	0:00
The generalization problem	0:10
The Bayesian generalization framework + size principle	0:18
Rectangle game + number game	0:40
Break	1:08
Student presentation — Shohei	1:15
No Free Lunch	1:40
Hierarchical Bayes + close	1:50

No Free Lunch

The No Free Lunch theorem

Wolpert (1996) — averaged over **all possible worlds**, no learning algorithm beats any other. Here is *why*, concretely.

The task. You see the bits $0, 1$ and must predict x_3.

Your rule. Say it predicts $x_3 = 0$. In the world $0, 1, \mathbf{0}$ it is **right**.

The mirror world. But the world $0, 1, \mathbf{1}$ is *just as possible* — same data $0, 1$, opposite answer. There your rule is **wrong**.

They pair off. Every world where your rule scores a point has a mirror world — identical data, flipped continuation — where it loses one.

Sum the pair. Right + wrong $= 1$ hit out of 2. Average over *all* worlds: every algorithm, every rule, scores **exactly** $1/2$.

No rule can win the average, because the data $0, 1$ says **nothing** about x_3 until you assume some worlds are more likely than others.

What NFL means for us

A learner only works because the distribution over worlds is **constrained** — i.e. because it has a **non-flat prior**.

- Generalization is *impossible* without inductive bias
- Recall: the hypothesis space $\mathcal{H}$ and the prior $p(h)$ from Block 3
- They are not bookkeeping — they are *the entire reason* learning is possible

Poll — No Free Lunch

What is the No Free Lunch theorem (for prediction)?

- **A.** When all hypotheses are possible, there’s nothing you can learn to predict
- **B.** Learning one hypothesis hurts learning other hypotheses
- **C.** If someone gives you lunch for free, they’ll expect something back
- **D.** Generalizing to new stimuli can hurt a learner

Poll — answer

A. When all hypotheses are possible, there’s nothing you can learn to predict.

A flat prior over every possible world means data carries no leverage — every continuation stays 50/50. To learn, you must *commit* to some worlds being more likely than others. That commitment is your prior — and it is why your prior matters.

Where we are

Welcome + Clusters walkthrough	0:00
The generalization problem	0:10
The Bayesian generalization framework + size principle	0:18
Rectangle game + number game	0:40
Break	1:08
Student presentation — Shohei	1:15
No Free Lunch	1:40
Hierarchical Bayes + close	1:50

Hierarchical Bayes

Back to Chibany's class

Every student in Chibany's class has their **own** tonkatsu-vs-hamburger rate θ_i.

The rates aren't identical — but they aren't unrelated either. They're all Chibany's customers.

Should learning Aoi's rate tell us anything about Ben's?

Priors over priors

Put a prior on the **parameters of the prior** — and treat (a, b) themselves as unknown:

$$\theta_i \sim \mathrm{Beta}(a, b)$$

The two-level model: $(a, b) \longrightarrow \theta_i \longrightarrow$ observed bentos.

(a, b) is the *shared* structure; each θ_i is a student.

Why it matters

A hierarchical model lets a learner **learn the prior from data** — exactly the inductive bias No Free Lunch said you cannot do without.

- (a, b) is learned from *all* the students together
- It is how “overhypotheses” get acquired — the shape bias, object vs. substance kinds (Kemp, Perfors & Tenenbaum, 2007)

Three ways to pool the data

Six students, six bento records. How do you estimate their rates?

Approach	Model	Problem
Complete pooling	one shared θ for everyone	ignores that students differ
No pooling	a separate θ_i, all unrelated	ignores that they're all Chibany's customers
Hierarchical	$\theta_i \sim \mathrm{Beta}(a, b)$	*the middle path* — borrows strength

Hierarchical Bayes = **partial pooling**.

The two-level model

Built up, top to bottom:

$$(a, b) \sim \text{prior}$$

$$\theta_i \mid a, b \sim \text{Beta}(a, b)$$

$$k_i \mid \theta_i \sim \text{Binomial}(n_i, \theta_i)$$

k_i — tonkatsu count for student i. n_i — that student's total bentos.

Inference — no closed form

We want the posterior over *everything*:

$$p\big(a, b, \{\theta_i\} \mid \text{data}\big)$$

Unlike the Beta-Binomial of Week 3, this has **no clean closed form.**

This is where sampling comes in — and where **GenJAX** earns its place (the hierarchical `bento_day()` exercise builds exactly this model).

Shrinkage — borrowing strength

For each student, compare:

- their **raw** tonkatsu fraction k_i/n_i
- their **posterior mean** $\hat{\theta}_i$

The posterior means are pulled **toward the group mean** — *a lot* when a student has little data, barely at all when they have a lot.

This is hierarchical Bayes **automatically borrowing strength** across students.

Overhypotheses — learning the bias

With the hierarchy in hand, **overhypotheses** become precise.

The shape bias; the object-vs-substance distinction — these are *second-level* hypotheses, learned as a distribution over *kinds of concept* (Kemp, Perfors & Tenenbaum, 2007).

No Free Lunch said a learner needs inductive bias. The hierarchy is where a learner **acquires** it — instead of being born with it.

Close

Before next week

Read **T3 Ch 5 — Mixture models** before Week 5.

It formalizes the same partial-pooling idea you just saw, and feeds directly into **Problem 3** of the Clusters assignment.

Clusters is due **Fri Jun 5, 8:00 PM**.

Next week — Week 5

Bayes nets + causal Bayes nets.

From “concepts as sets” to **structured probabilistic models** — graphs that encode how variables depend on each other, and what happens when you *intervene*.

Check the readings page for the Week 5 paper + presenter.