UNSW AUSTRALIA

Never Stand Still

On the origins of data: A Bayesian perspective on inductive generalisation, categorisation and reasoning

Dan Navarro

Amy Perfors

Drew Hendrickson

Keith Ransom

Wouter Voorspoels

Wai Keen Vong

Simon De Deyne

Sean Tauber

Lauren Kennedy

Steve Langsford

also Michael Lee, Matt Dry, Ben Newell, Gert Storms, Christin Schulze, Pat Shafto, Nancy Briggs, Jay Myung, Mark Pitt, Woojae Kim, Ian Fuss, Mark Steyvers, Josh Tenenbaum, Tom Griffiths, Adam Sanborn, Charles Kemp, Rachel Stephens, Dinis Gokaydin, Anastasia Ejova, Matthew Welsh, Luke Maurits, Belinda Bruza, Steve Begg, and more

Amy Perfors

Drew Hendrickson

Keith Ransom

Wouter Voorspoels

Wai Keen Vong

Simon De Deyne

Sean Tauber

Lauren Kennedy

Steve Langsford

also Michael Lee, Matt Dry, Ben Newell, Gert Storms, Christin Schulze, Pat Shafto, Nancy Briggs, Jay Myung, Mark Pitt, Woojae Kim, Ian Fuss, Mark Steyvers, Josh Tenenbaum, Tom Griffiths, Adam Sanborn, Charles Kemp, Rachel Stephens, Dinis Gokaydin, Anastasia Ejova, Matthew Welsh, Luke Maurits, Belinda Bruza, Steve Begg, and more

How do we make choices in an uncertain world?

(judgment & decision making)

How do people acquire new knowledge?

(categorisation & reasoning)

How should psychologists analyse our data?

(math psych & statistics)

How do we make choices in an uncertain world?

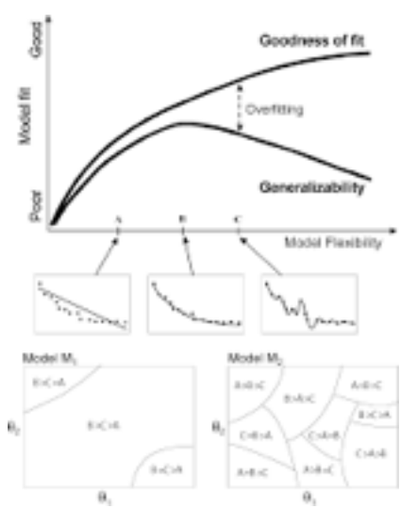
(judgment & decision making)

How do people acquire new knowledge?

(categorisation & reasoning)

How should psychologists analyse our data?

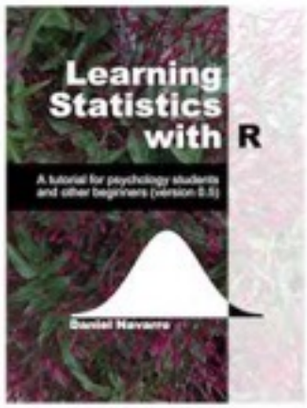
(math psych & statistics)



What criteria should be used to evaluate psychological models?

Kim, Navarro, Pitt & Myung (2006). *Psych Review.*
Navarro, Pitt, Myung (2004) *Cognitive Psychology*

How should psychologists analyse our data?



What tools should we teach to undergrads?

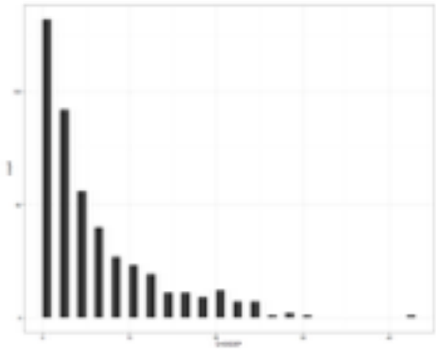
Navarro (2015)??

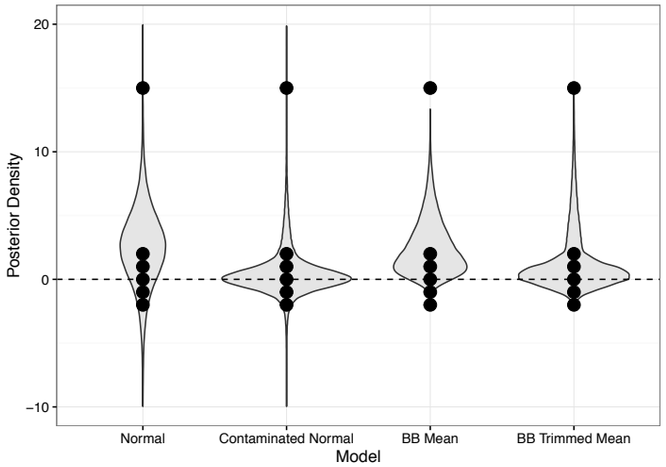
What criteria should be used to evaluate psychological models?

How should psychologists analyse our data?

What tools should we teach to undergrads?

What should we use as our "default" models when our data sets are messy?

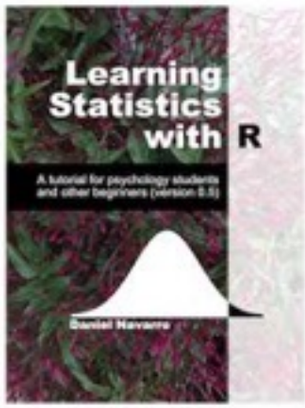




Kennedy, Navarro, Perfors & Briggs (in prep)

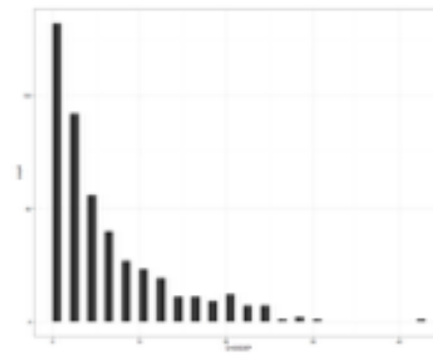
What criteria should be used to evaluate psychological models?

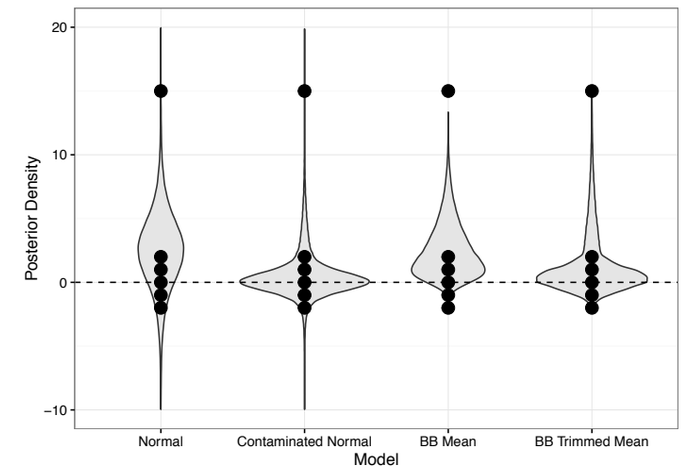
How should psychologists analyse our data?

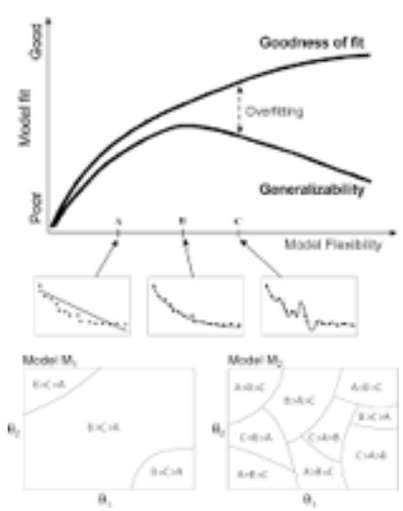


What tools should we teach to undergrads?

What should we use as our "default" models when our data sets are messy?



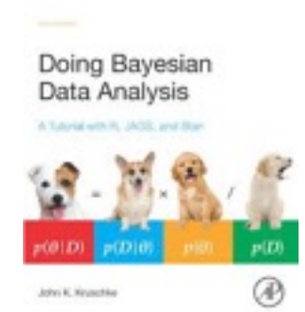




What criteria should be used to evaluate psychological models?

How should psychologists analyse our data?

Should we all be Bayesian?



How do we make choices in an uncertain world?

(judgment & decision making)

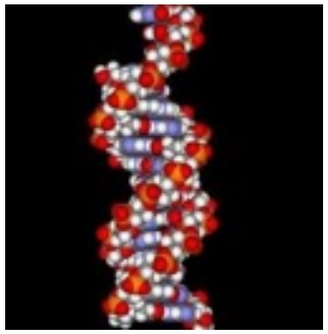
How do people acquire new knowledge?

(categorisation & reasoning)

How should psychologists analyse our data?

(math psych & statistics)

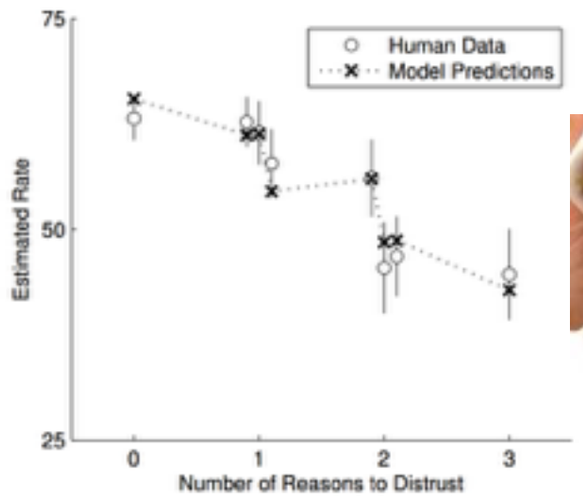
How do we make choices
in an uncertain world?


How do people aggregate prior
knowledge with statistical evidence?

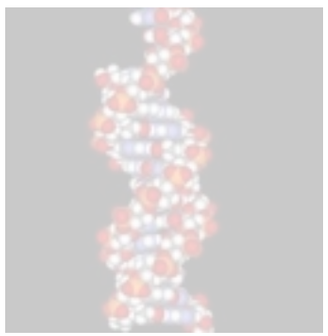
Tauber, Navarro, Perfors & Steyvers
(under revision). *Psych Review*

How do people evaluate the quality of evidence?



Welsh & Navarro (2012). *Org. Behavior & Human Dec. Making*

How do we make choices in an uncertain world?

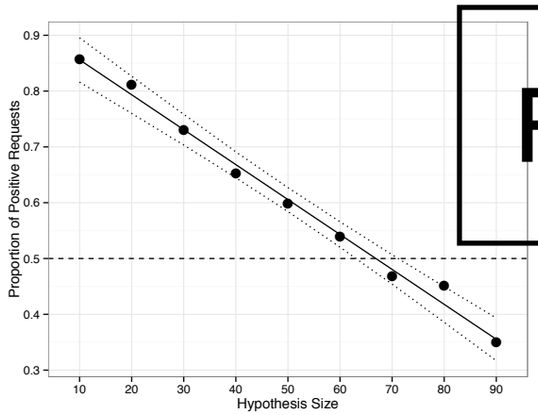


How do people aggregate prior knowledge with statistical evidence?

How do people evaluate the quality of evidence?

Navarro & Perfors (2011). *Psych Review*
Hendrickson, Perfors & Navarro (2016) *Decision*

How do we make choices in an uncertain world?



P D 2 4

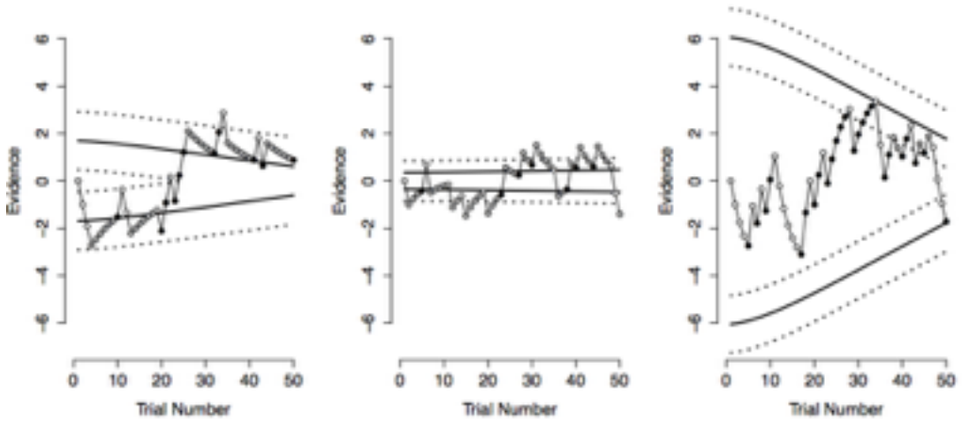
When do the statistics of the task matter?

How do people aggregate prior knowledge with statistical evidence?

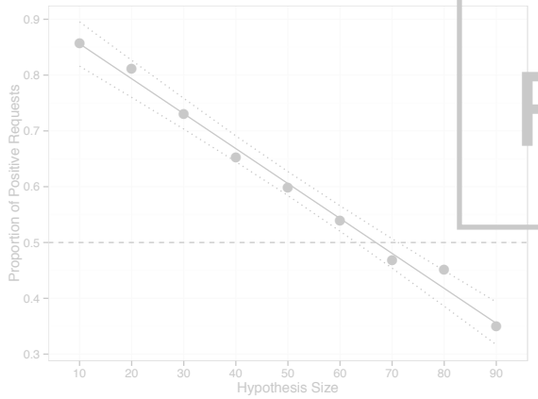
How do we make choices in an uncertain world?

How do people evaluate the quality of evidence?

What psychological processes control how people regulate evidence accrual and choice?



Navarro, Newell & Schulze (2016). *Cogn Psych*
Navarro & Fuss (2009). *J. Math. Psychol.*



P D 2 4

When do the statistics of the task matter?

How do people aggregate prior knowledge with statistical evidence?

How do we make choices in an uncertain world?

(judgment & decision making)

How do people acquire new knowledge?

(categorisation & reasoning)

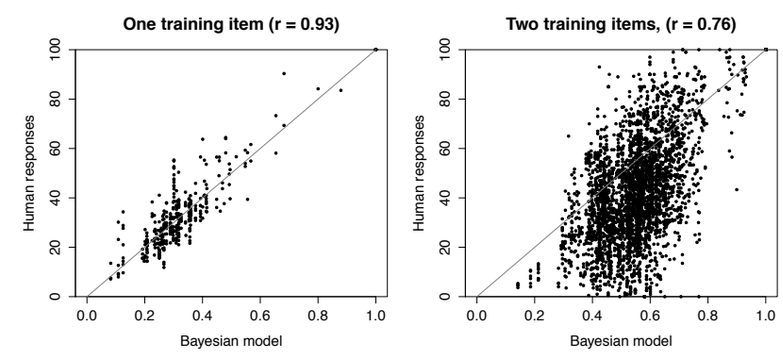
How should psychologists analyse our data?

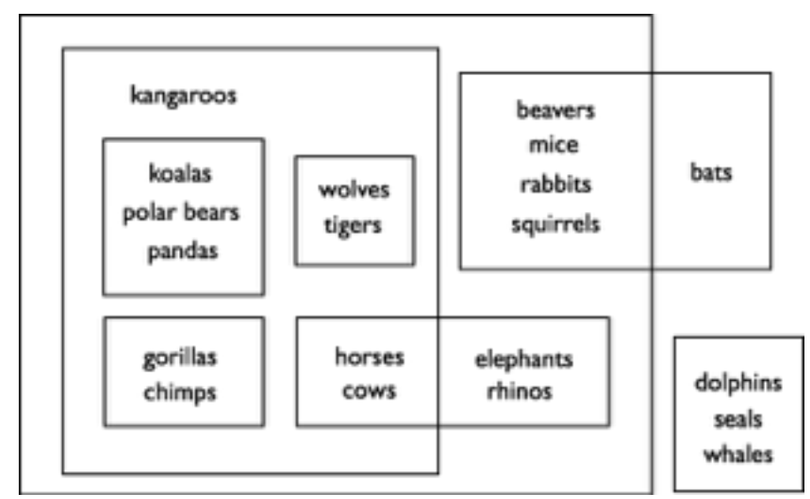
(math psych & statistics)

How do people acquire new knowledge?

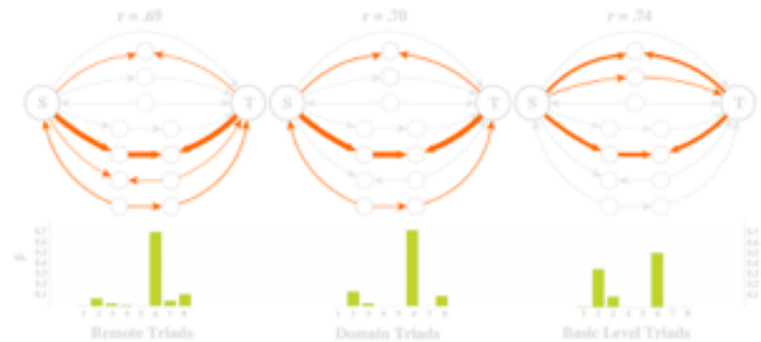
Tauber, Navarro, Perfors & Steyvers (under revision). *Psych. Review*
Navarro & Griffiths (2008). *Neural Computation*
Navarro & Lee (2004) *PB&R*

What mental representations underpin human reasoning?





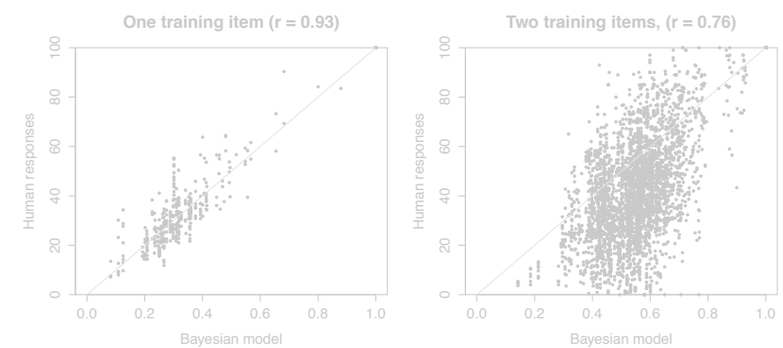
Can we use "big data" to study semantic representation?



De Deyne, Navarro & Storms (2013). *BRM*
De Deyne, Navarro, Perfors & Storms (in press). *JEP:G*

How do people acquire new knowledge?

What mental representations underpin human reasoning?

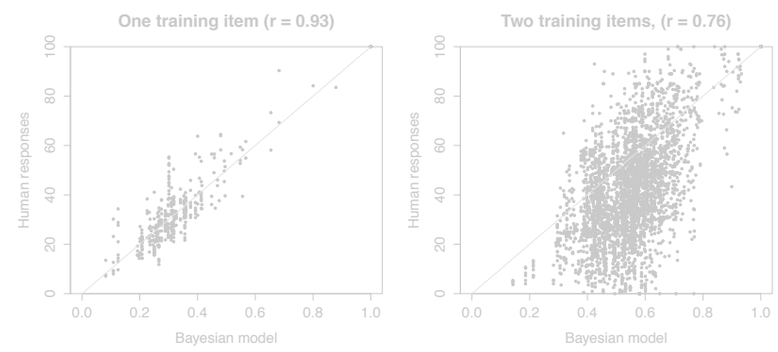


Can we use "big data" to study semantic representation?

How do people acquire new knowledge?

What kinds of inductive biases guide human learning and reasoning?

What mental representations underpin human reasoning?



Inductive bias

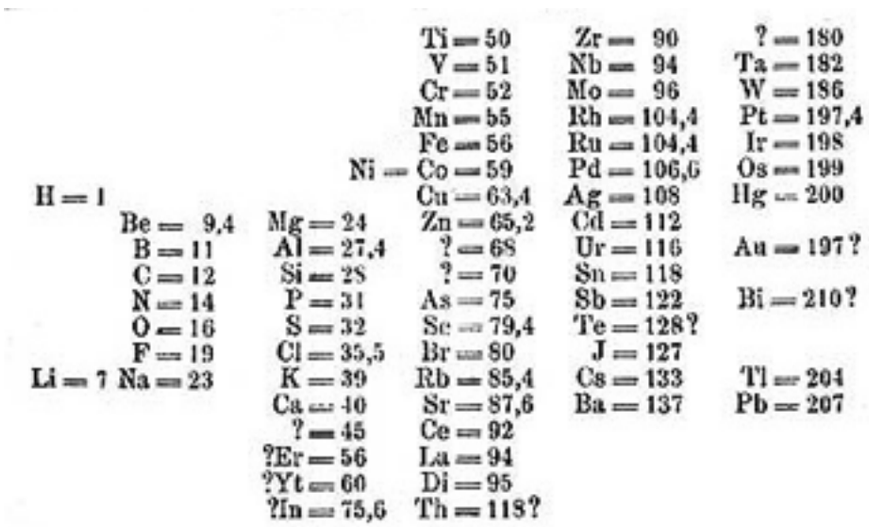
Why do people generalise one way and not another?

Different biases are learned for different ontological kinds

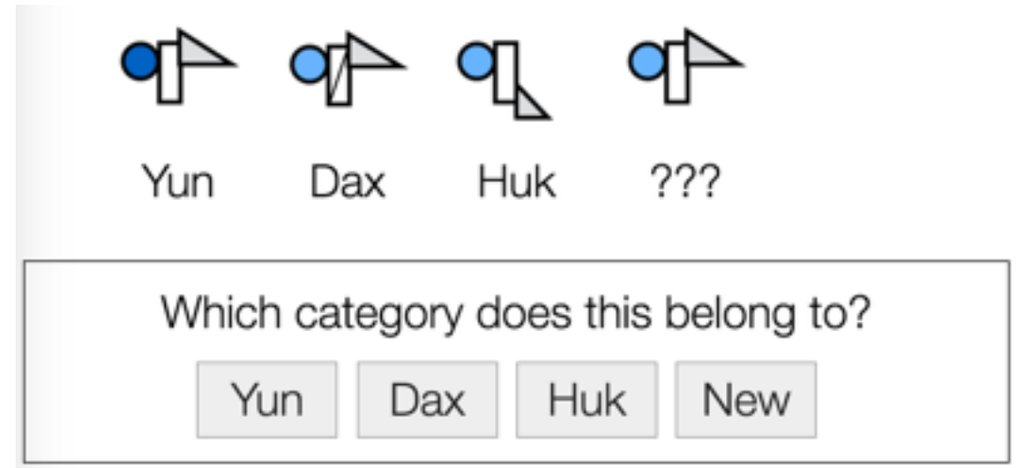
Class	Category	Instance	Shape	Color
Object	Cat	1	*Quadruped*	Tabby
		2	*Quadruped*	White
		3	*Quadruped*	Grey
	Ball	1	*Sphere*	Blue
		2	*Sphere*	Red
		3	*Sphere*	Green
Substance	Mud	1	Pile	*Brown*
		2	Puddle	*Brown*
		3	Spherical	*Brown*
	Flour	1	Pile	*White*
		2	Cloud	*White*
		3	Scatter	*White*

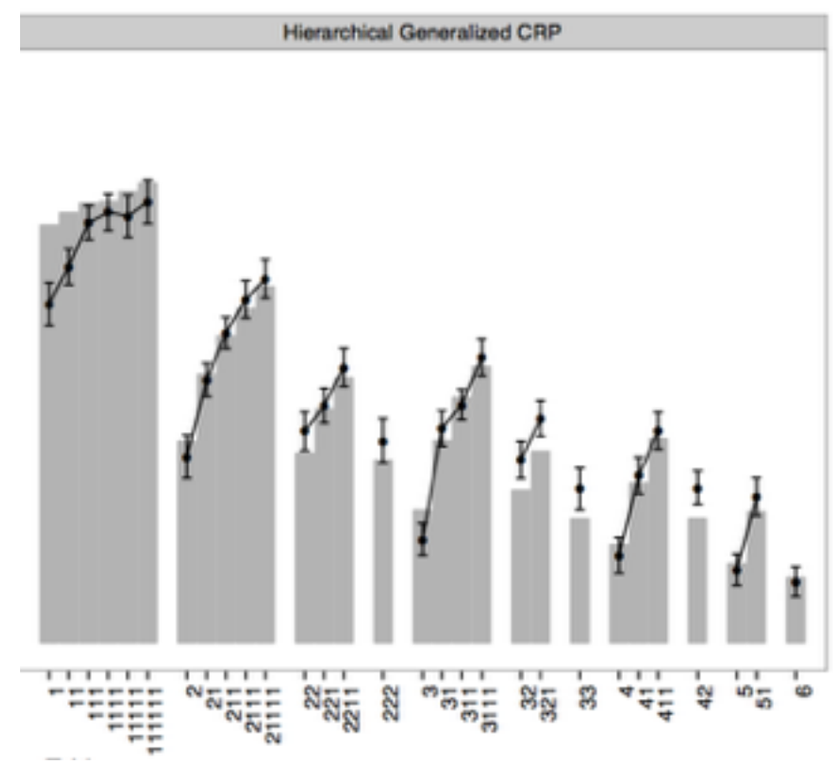
Perfors, Navarro & Tenenbaum (under revision). *Cognitive Science*

Biases influence the discovery of new categories



			Ti = 50	Zr = 90	? = 180
			V = 51	Nb = 94	Ta = 182
			Cr = 52	Mo = 96	W = 186
			Mn = 55	Rh = 104,4	Pt = 197,4
			Fe = 56	Ru = 104,4	Ir = 198
		Ni =	Co = 59	Pd = 106,6	Os = 199
H = 1			Cu = 63,4	Ag = 108	Hg = 200
	Be = 9,4	Mg = 24	Zn = 65,2	Cd = 112	
	B = 11	Al = 27,4	? = 68	Ur = 116	Au = 197?
	C = 12	Si = 28	? = 70	Sn = 118	
	N = 14	P = 31	As = 75	Sb = 122	Bi = 210?
	O = 16	S = 32	Se = 79,4	Te = 128?	
	F = 19	Cl = 35,5	Br = 80	J = 127	
Li = 7	Na = 23	K = 39	Rb = 85,4	Cs = 133	Tl = 204
		Ca = 40	Sr = 87,6	Ba = 137	Pb = 207
		? = 45	Ce = 92		
		?Er = 56	La = 94		
		?Yt = 60	Di = 95		
		?In = 75,6	Th = 118?		

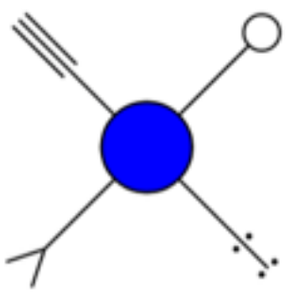




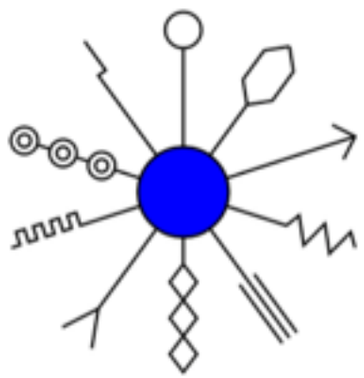
Navarro & Kemp (under revision). *Psych. Review*

Biases constrain & shape learning

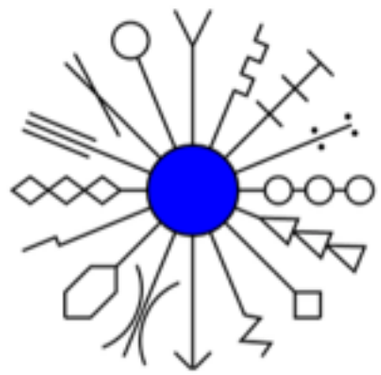
4

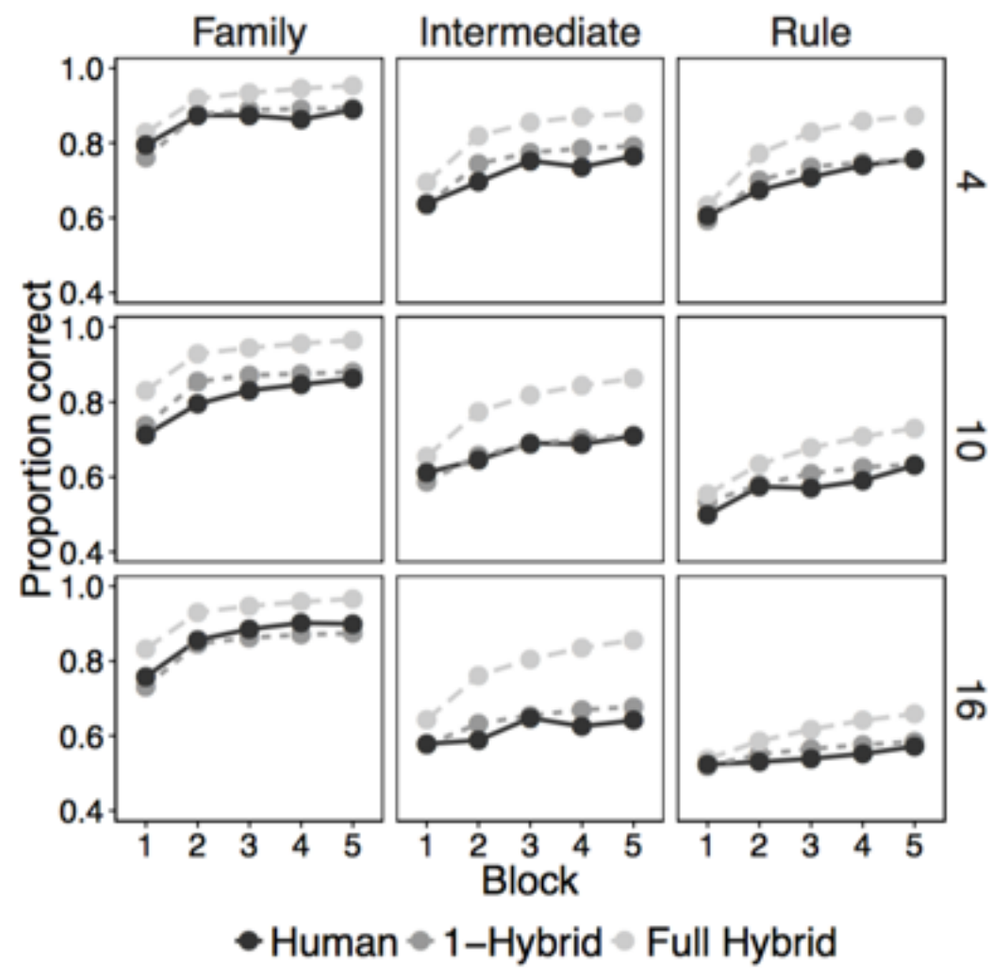


8



16

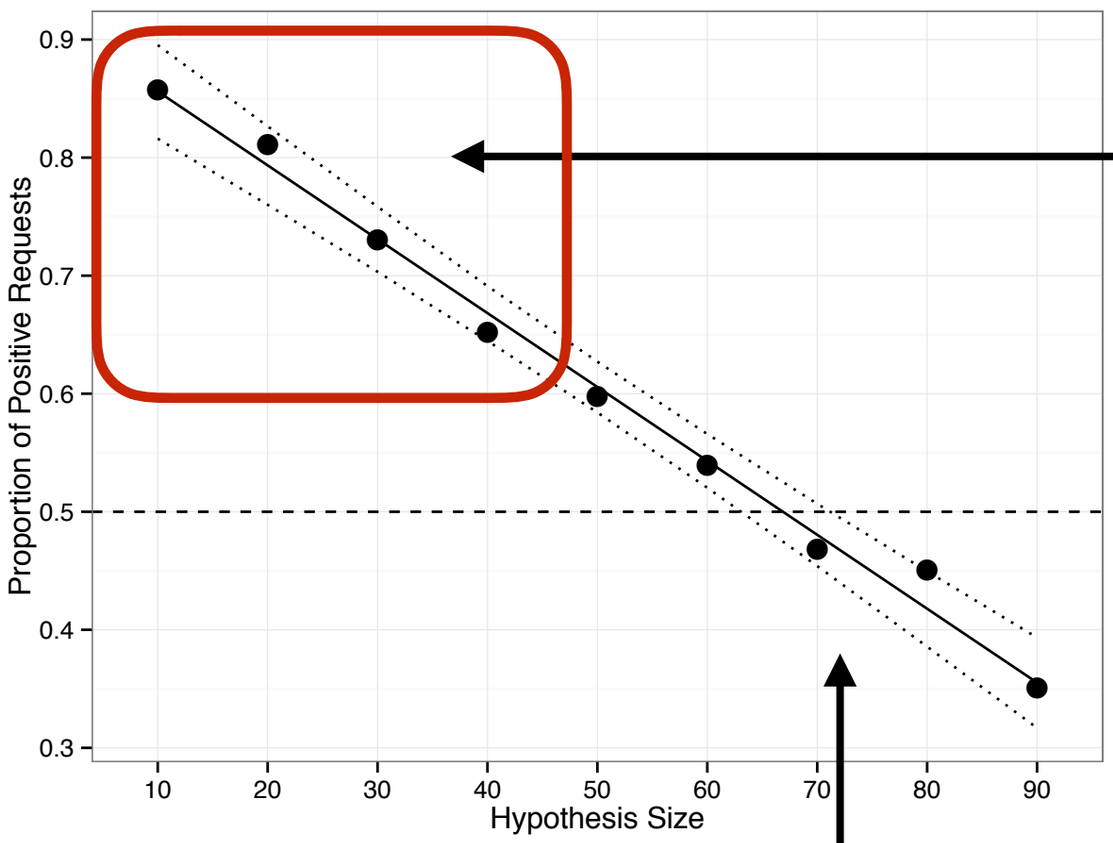




Limited capacity statistical learning?

Vong, Hendrickson, Perfors & Navarro (in press). *Proc. Cog. Sci. Soc.*

Biases shape information search



"Confirmatory evidence" is informative when hypotheses are consistent with only a minority of possible observations

When confirmation is uninformative, people prefer to search for disconfirmatory evidence

Navarro & Perfors (2011). *Psych Review*
Hendrickson, Perfors & Navarro (2016) *Decision*

Why study inductive bias using <u>computational models</u>?

In the case where $n = 1$ we observe that,

$$\begin{aligned}
\int_{\mathcal{R}} P(x_1, x_1 \in r_t \,|\, r_t = r)\; dr &= \int_0^{z_l} \int_{z_u}^1 P(x_1 \,|\, r_t = [l, u])\; du\; dl \\
&= \int_0^{z_l} \int_{z_u}^1 (u - l)^{-1}\; du\; dl \\
&= \int_0^{z_l} [\ln(u - l)]_{z_u}^1\; dl \\
&= \int_0^{z_l} \ln(1 - l) - \ln(z_u - l)\; dl \\
&= [(l - 1)\ln(1 - l) - l]_0^{z_l} - [(l - z_u)\ln(z_u - l) - l]_0^{z_l} \\
&= ((z_l - 1)\ln(1 - z_l) - z_l) - ((z_l - z_u)\ln(z_u - z_l) - z_l + z_u \ln z_u) \\
&= (z_u - z_l)\ln(z_u - z_l) - (1 - z_l)\ln(1 - z_l) - z_u \ln z_u \qquad (24)
\end{aligned}$$

Applying the same procedure as before yields the expression

$$P(y \in r_t \,|\, x_1, x_1 \in r_t) = \begin{cases} \dfrac{(z_u - y)\ln(z_u - y) - (1 - y)\ln(1 - y) - z_u \ln z_u}{(z_u - z_l)\ln(z_u - z_l) - (1 - z_l)\ln(1 - z_l) - z_u \ln z_u} & \text{if} \quad y < z_l \\ 1 & \text{if} \quad z_l \leq y \leq z_u \\ \dfrac{(y - z_l)\ln(y - z_l) - (1 - z_l)\ln(1 - z_l) - y \ln y}{(z_u - z_l)\ln(z_u - z_l) - (1 - z_l)\ln(1 - z_l) - z_u \ln z_u} & \text{if} \quad z_u < y \end{cases} \qquad (25)$$

In this case, however, the expression can be further simplified since $z_l = z_u = x_1$:

$$P(y \in r_t \,|\, x_1, x_1 \in r_t) = \begin{cases} \dfrac{(1 - y)\ln(1 - y) + x_1 \ln x_1 - (x - y)\ln(x_1 - y)}{(1 - x_1)\ln(1 - x_1) + x_1 \ln x_1} & \text{if} \quad y < x_1 \\ 1 & \text{if} \quad y = x_1 \\ \dfrac{(1 - x_1)\ln(1 - x_1) + y \ln y - (y - x_1)\ln(y - x_1)}{(1 - x_1)\ln(1 - x_1) + x_1 \ln x_1} & \text{if} \quad x_1 < y \end{cases} \qquad (26)$$

(Obviously, this expression could be derived directly, rather than found as a special case

In the case where $n = 1$ we observe that,

$$\begin{aligned}\int_{\mathcal{R}} P(x_1, x_1 \in r_t \,|\, r_t = r)\; dr &= \int_0^{z_l} \int_{z_u}^1 P(x_1 \,|\, r_t = [l, u])\; du\; dl \\ &= \int_0^{z_l} \int_{z_u}^1 (u - l)^{-1}\; du\; dl\end{aligned}$$

$$- z_l + z_u \ln z_u) \qquad (24)$$

Applying the same proced

$$P(y \in r_t \,|\, x_1, x_1 \in r_t) = \begin{cases} & y < z_l \\ & z_l \leq y \leq z_u \\ & z_u < y \end{cases} \qquad (25)$$

In this case, however, the

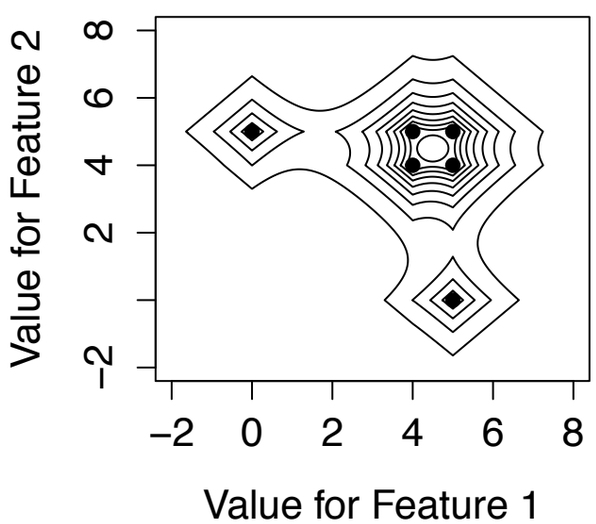
$$P(y \in r_t \,|\, x_1, x_1 \in r_t) = \begin{cases} \dfrac{(1-y)\ln(1-y) + x_1 \ln x_1 - (x-y)\ln(x_1 - y)}{(1-x_1)\ln(1-x_1) + x_1 \ln x_1} & \text{if} \quad y < x_1 \\ 1 & \text{if} \quad y = x_1 \\ \dfrac{(1-x_1)\ln(1-x_1) + y \ln y - (y - x_1)\ln(y - x_1)}{(1-x_1)\ln(1-x_1) + x_1 \ln x_1} & \text{if} \quad x_1 < y \end{cases} \qquad (26)$$

(Obviously, this expression could be derived directly, rather than found as a special case

Computational models make it easier to be precise about one's theories

??

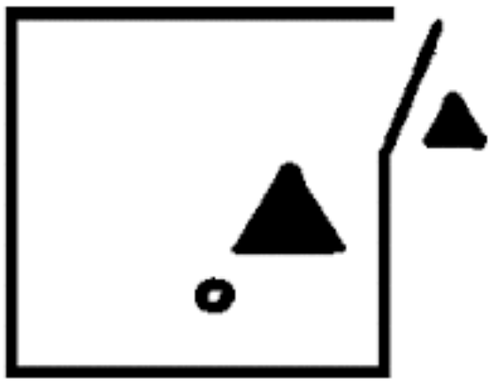
categorisation is sort of related to similarity I guess?



categorisation probability is proportional to the sum of similarities to previous exemplars

Formal descriptions of human inductive biases are useful for machine learning

triangles are jerks



inferring intention from actions

"I'm not driving"

understanding the relevance of utterances to context

teapot death star?

constructing categories from instances

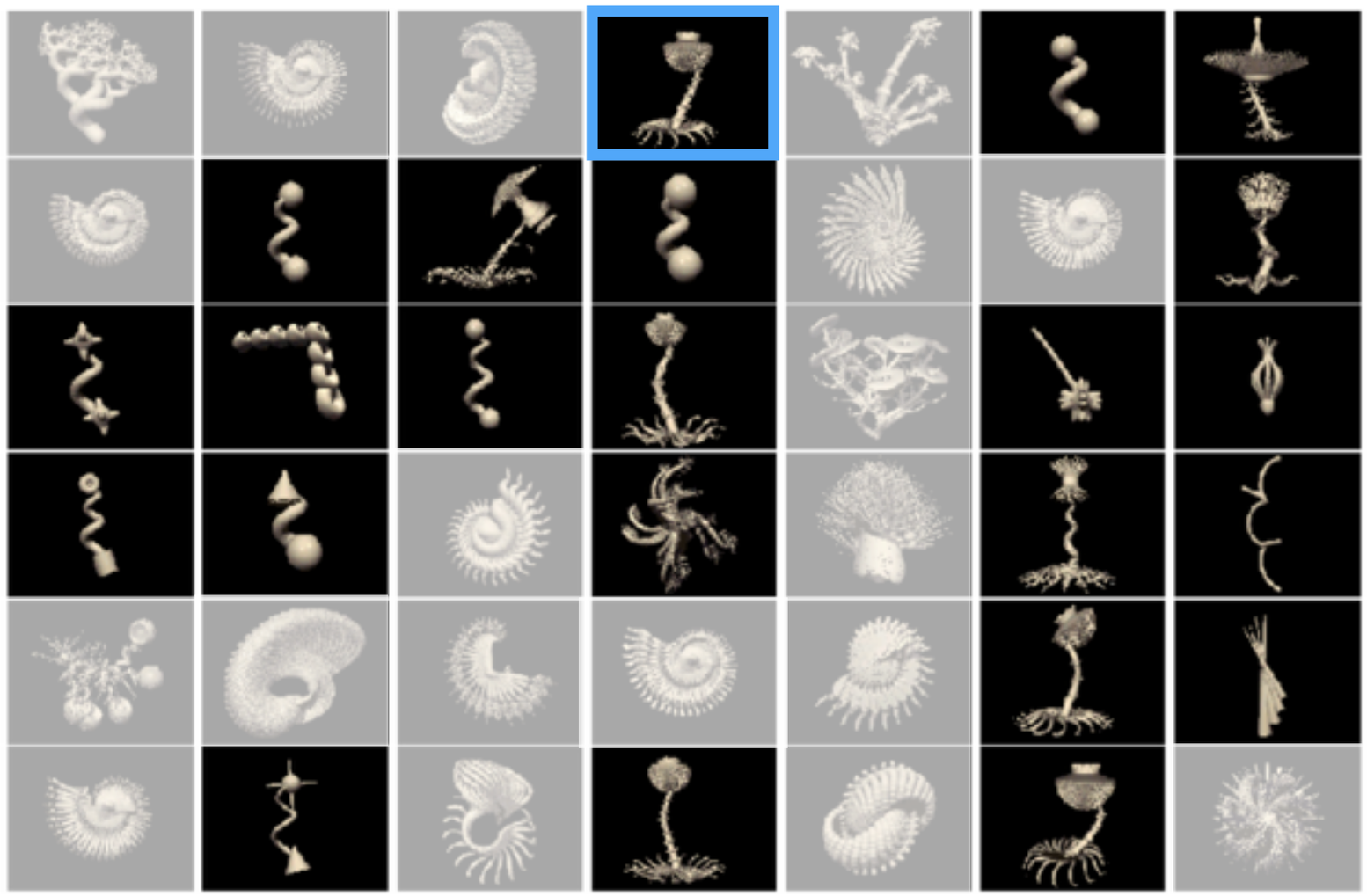
Machine agents need to interact with humans, so they need to understand us

changing speech patterns when the speech recognition system doesn't work

autonomous vehicles need to understand human drivers

Sampling assumptions in inductive generalisation

“tufa”



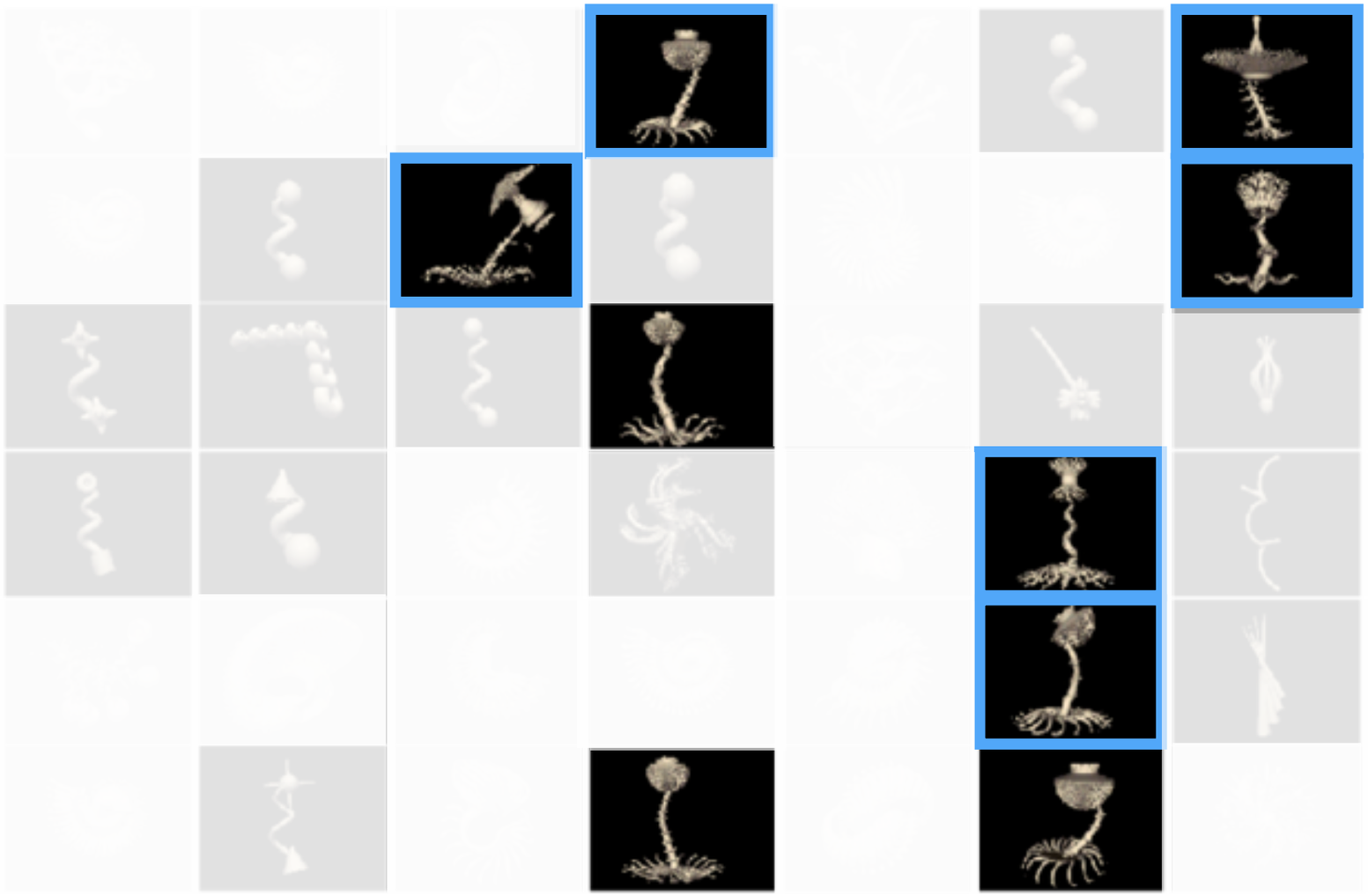
very dissimilar things are
probably not tufas

“tufa” “tufa”

“tufa”

it’s such a weird coincidence that ALL THREE have the SAME shape, right?

Generalisations narrow as this "coincidence" becomes suspicious

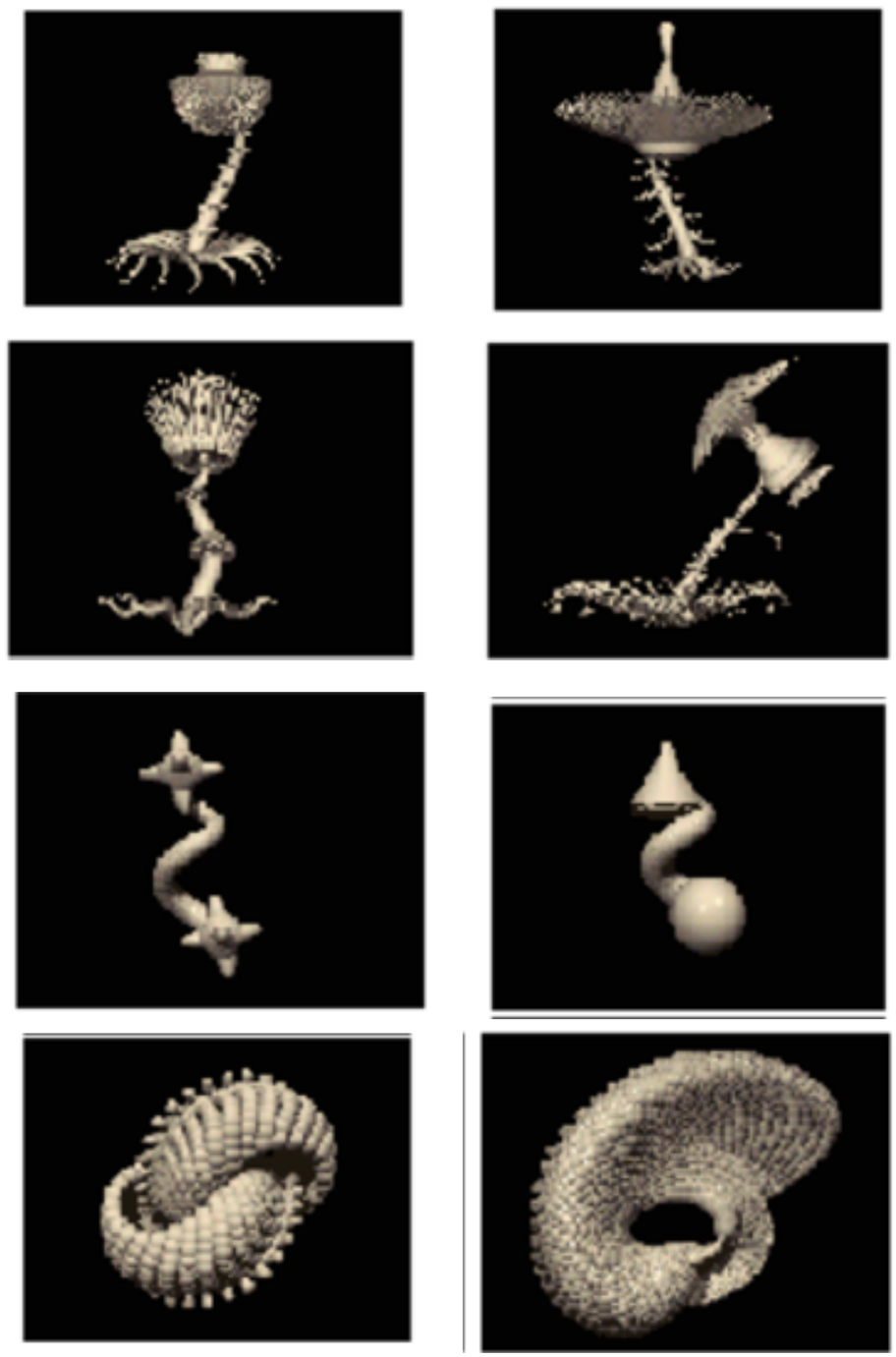


"tufas"

"The null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist **only to give the facts a chance of disproving the null hypothesis**."

- R. A. Fisher

Let's play… "Do what Fisher says"



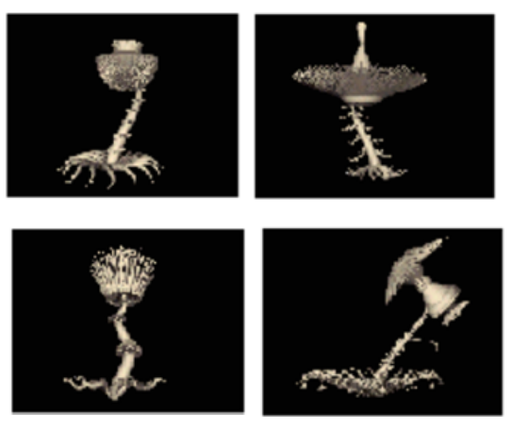
← Things!

Which things are tufas?

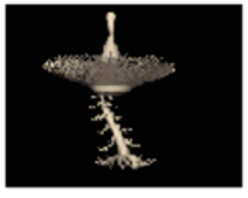
helical things?

mushroom heads?

creepy flowers?

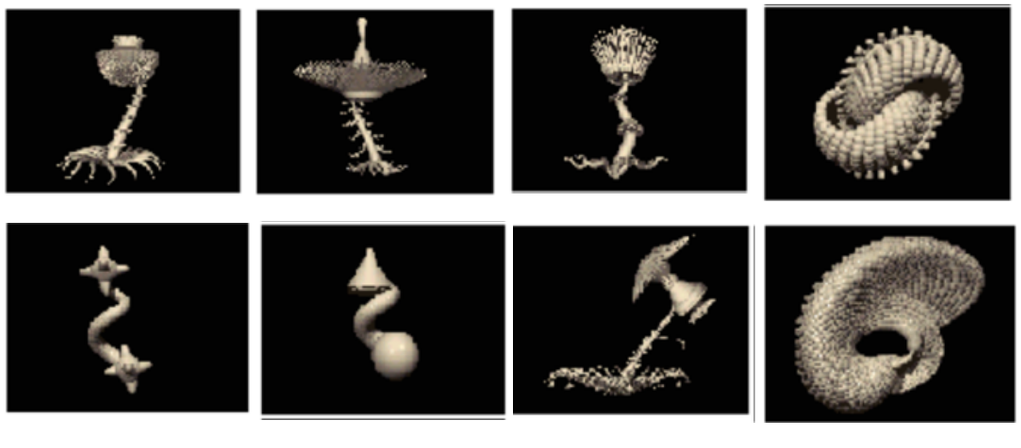


botanical radio telescopes?

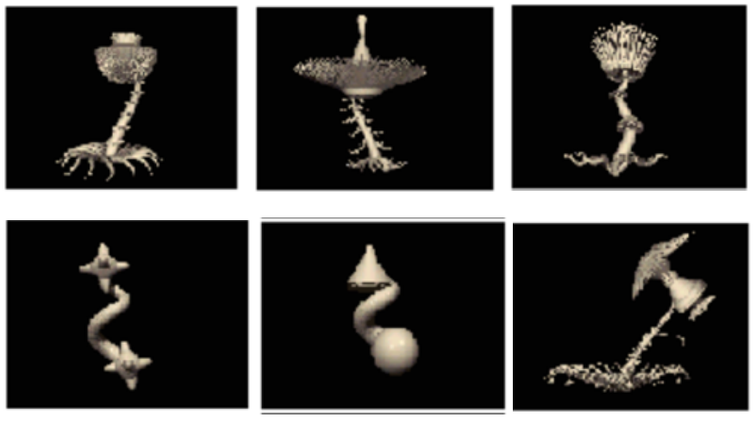


seashell things?

they're all tufas!



anything long and narrow

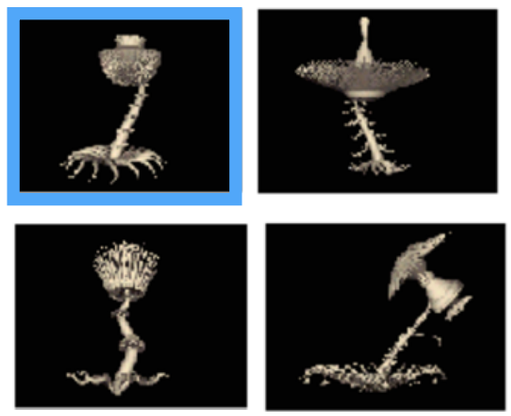


Observe one tufa and falsify…

helical things?

mushroom heads?

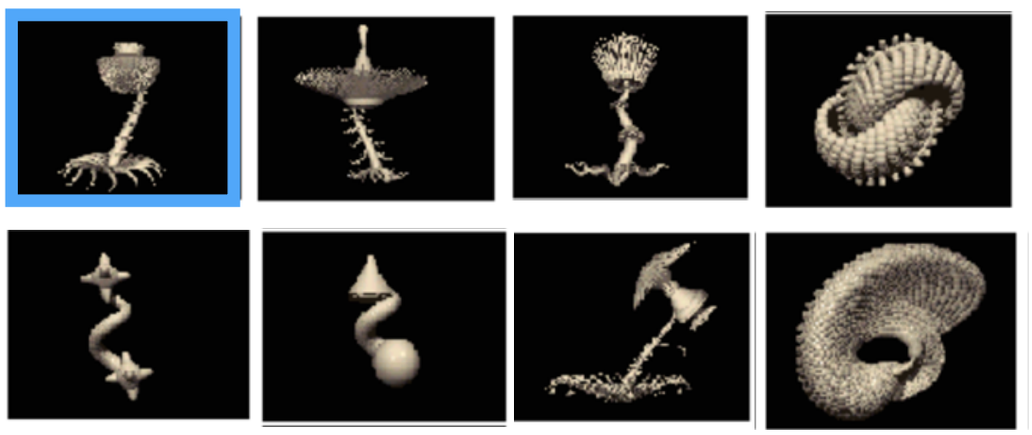
creepy flowers?



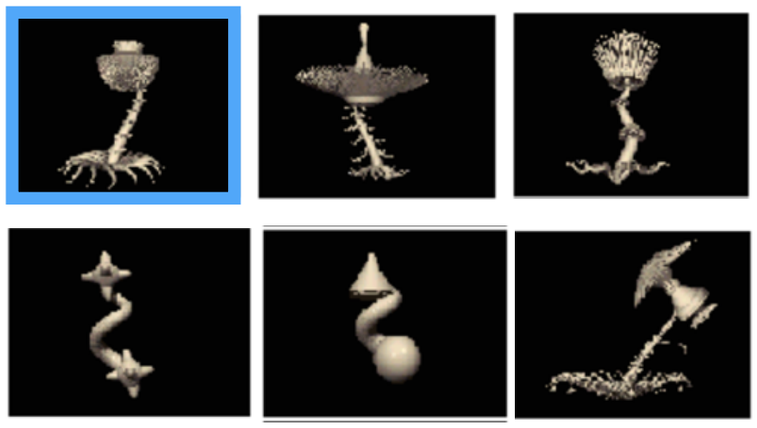
seashell things?

botanical radio telescopes?

they're all tufas!



anything long and narrow

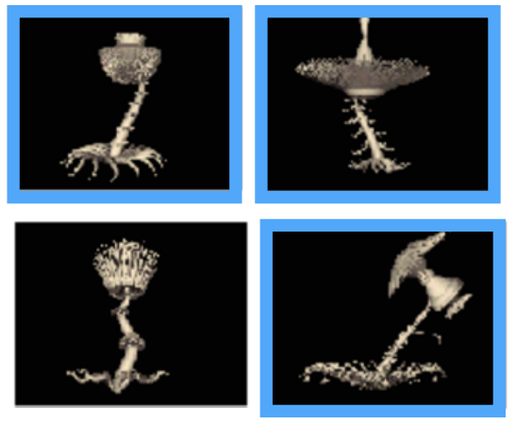


The next two tufas falsify no hypotheses, so we learn nothing?

helical things?

mushroom heads?

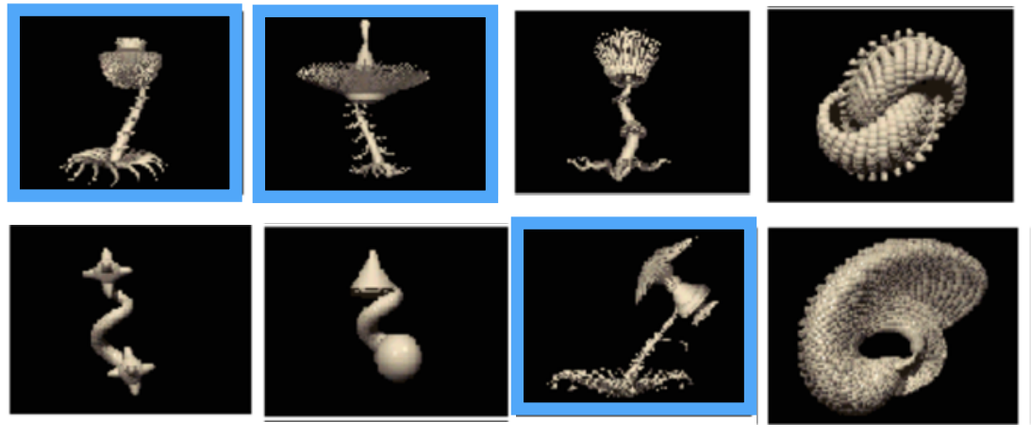
creepy flowers?



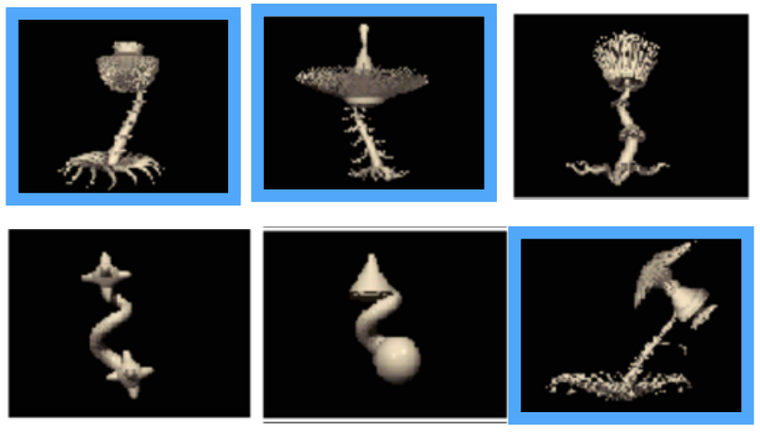
botanical radio telescopes?

seashell things?

they're all tufas!



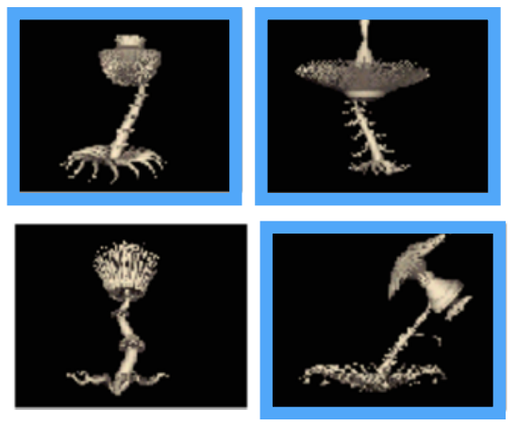
anything long and narrow



Mr. Ockham wishes to discuss tufas with you…

mushroom heads?

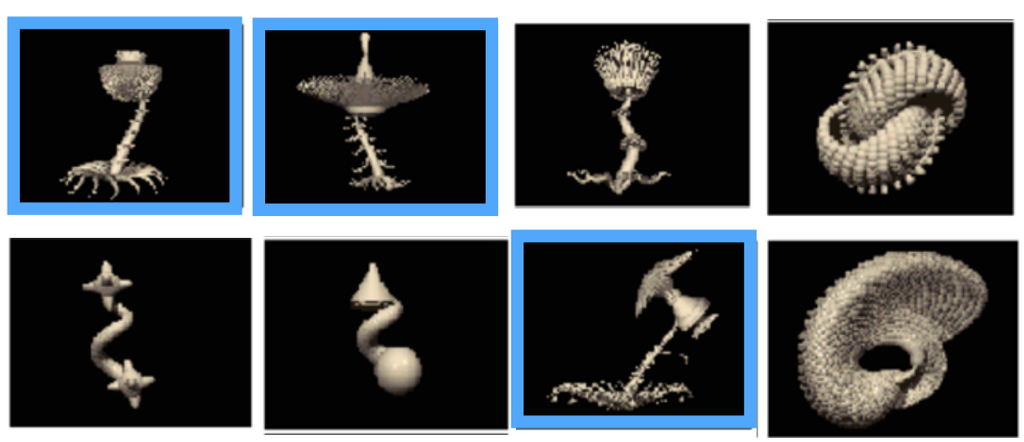
creepy flowers?



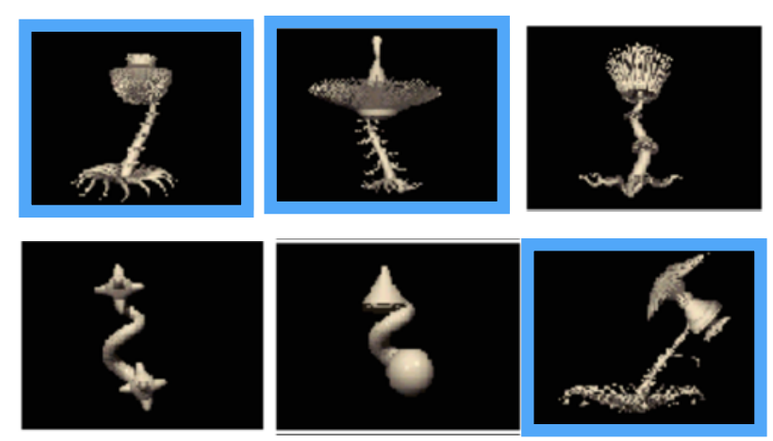
These hypotheses do not require me to believe a bizarre coincidence as to why the only observed tufas are so bloody similar

For these to be plausible, I require an <u>additional</u> explanation as to why the only tufas I have <u>seen</u> are flower-like

they're all tufas!



anything long and narrow

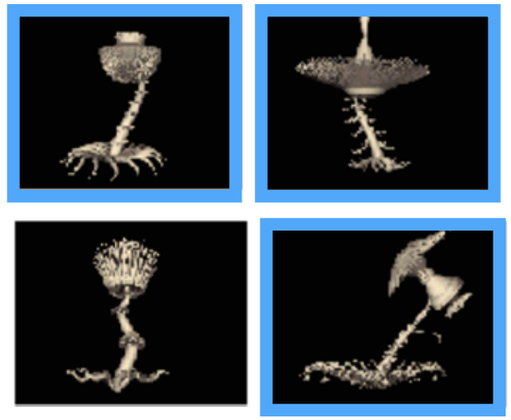


An Ockhamist reasoner prefers the smallest hypotheses consistent with the observations

helical things?

mushroom heads?

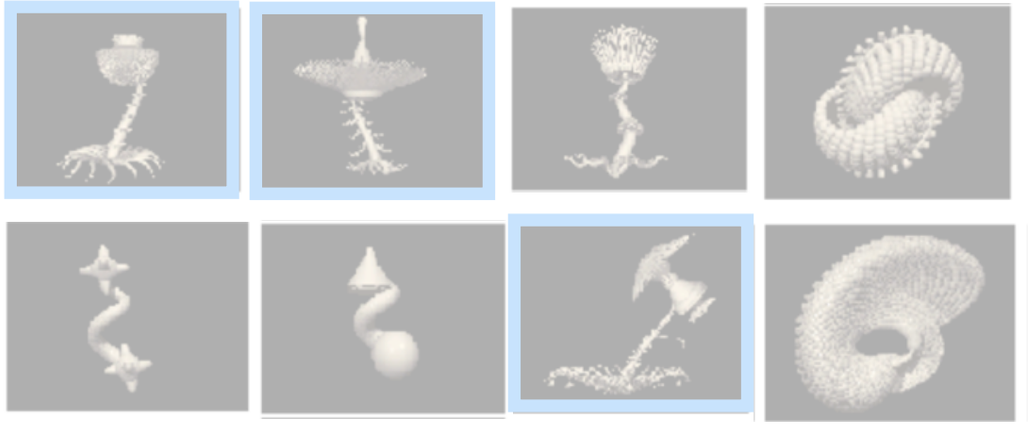
creepy flowers?



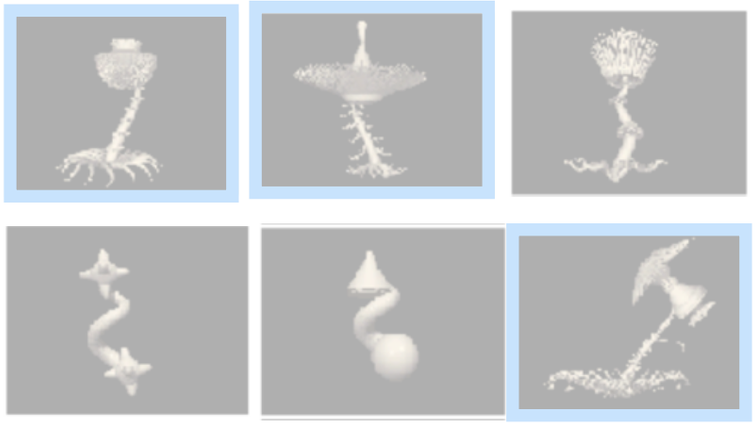
botanical radio telescopes?

seashell things?

they're all tufas!



anything long and narrow



Are these fundamentally distinct?

Or can we express them in a common framework?

Bayes #1?

Bayes #2?

Bayes' rule: $$P(h|x) \propto P(x|h)P(h)$$

Posterior degree of belief

Prior degree of belief

A Bayesian "scores" hypotheses by asking how likely they think it is that we data x would be if hypothesis h were true?

$$P(h|x) \propto P(x|h)P(h)$$

The likelihood supplies the learner with a **theory** about the data generating process

$$P(h|x) \propto P(x|h)P(h)$$

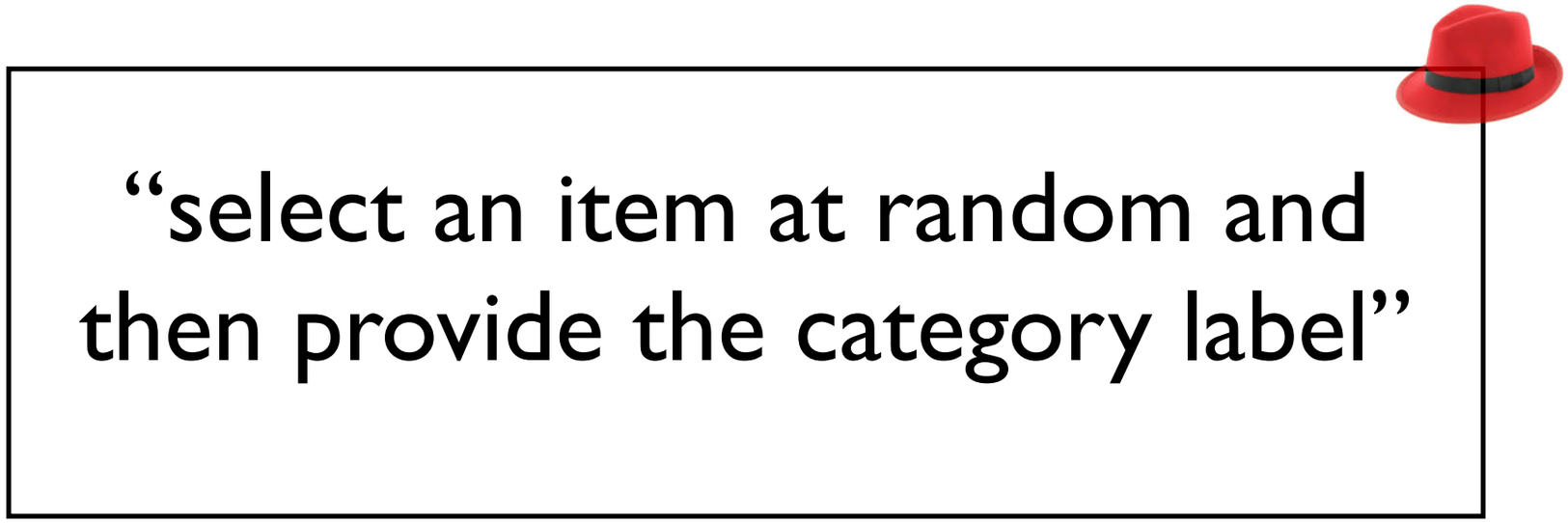
Different theories, different learning

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

Two simple theories about the data generating mechanism…

Weak sampling:



Two simple theories about the data generating mechanism…

Weak sampling:

"select an item at random and then provide the category label"

Strong sampling:

"select items <u>from</u> the target category"

… produce two different learning rules

Weak sampling:

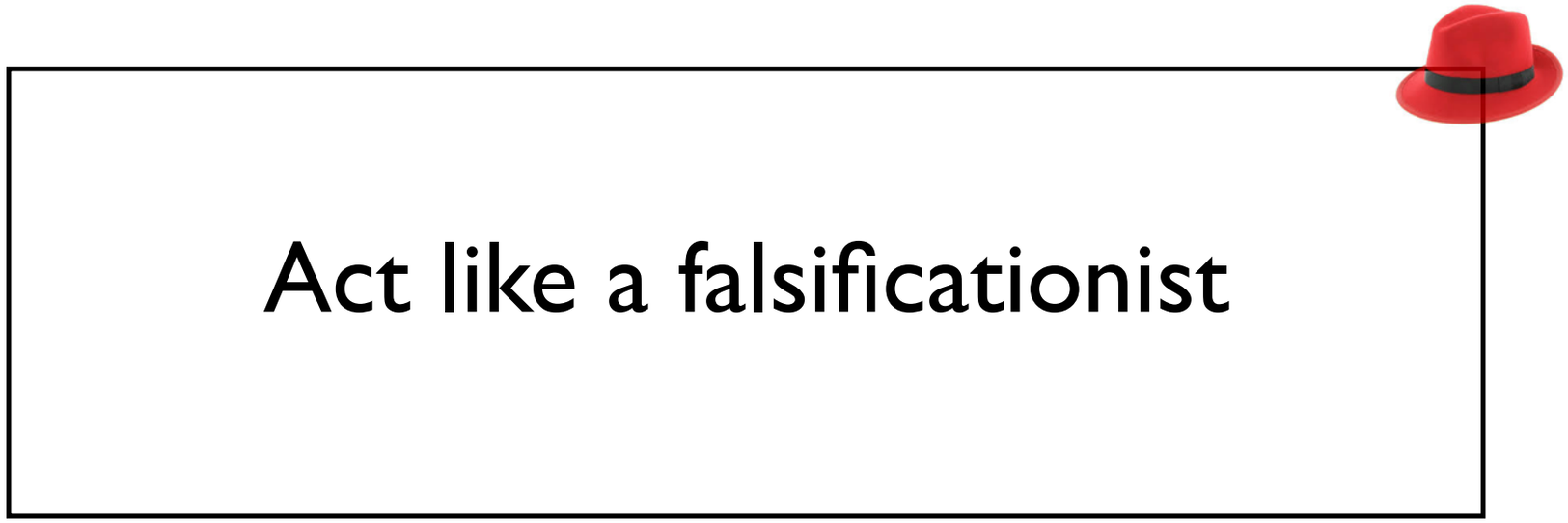
$$P(x|h) \propto \begin{cases} 1 & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases}$$

Strong sampling:

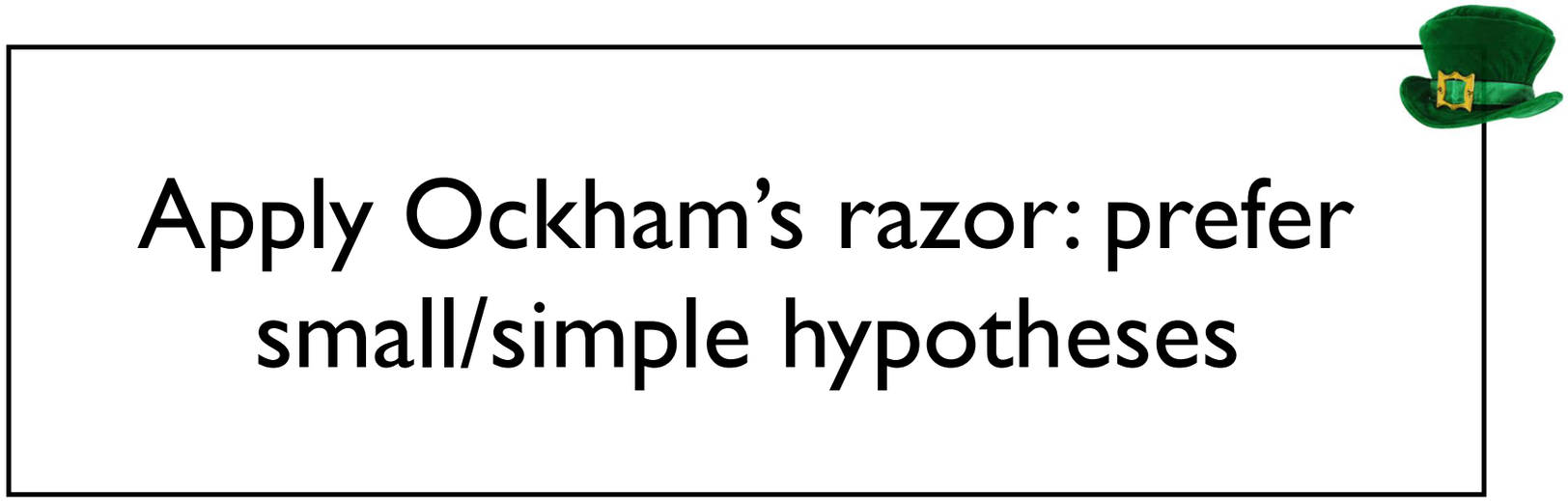
$$P(x|h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases}$$

And yield qualitatively different behaviour

Weak sampling:

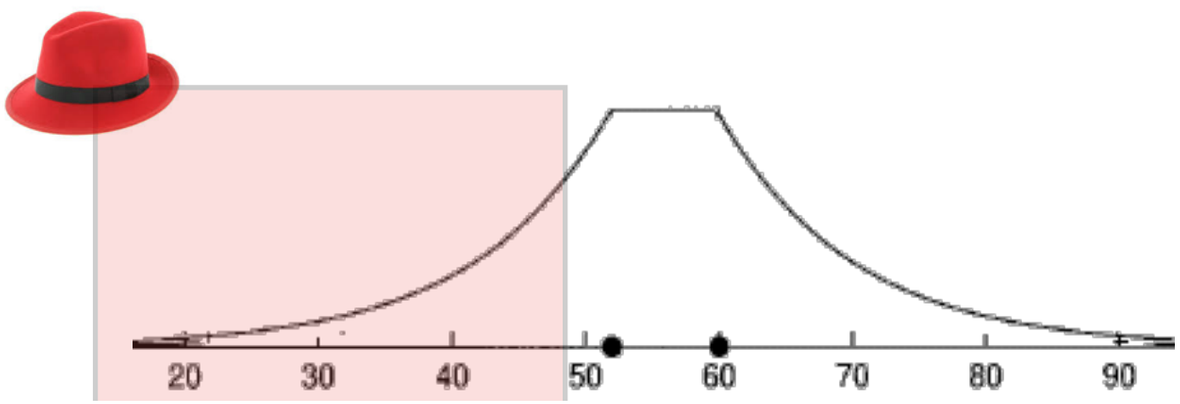


Strong sampling:

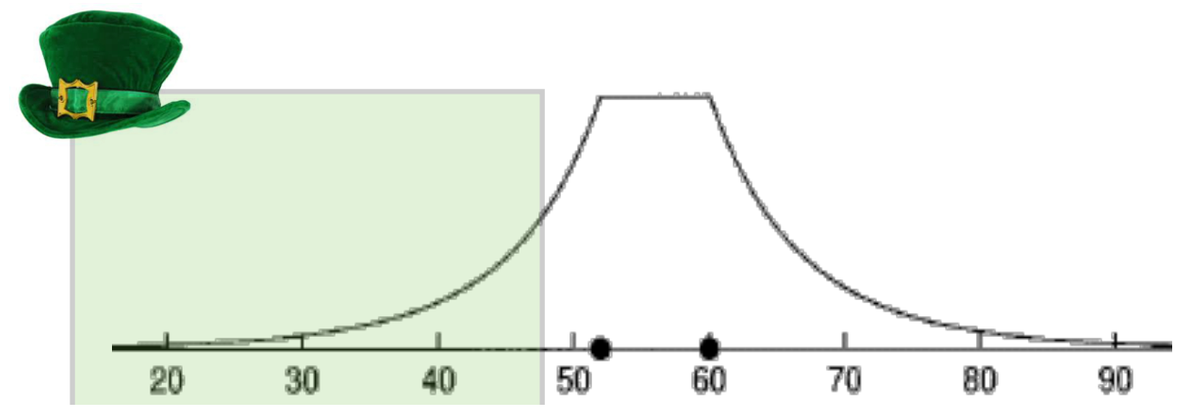


Here's the testable prediction about generalisation gradients…

weak sampling



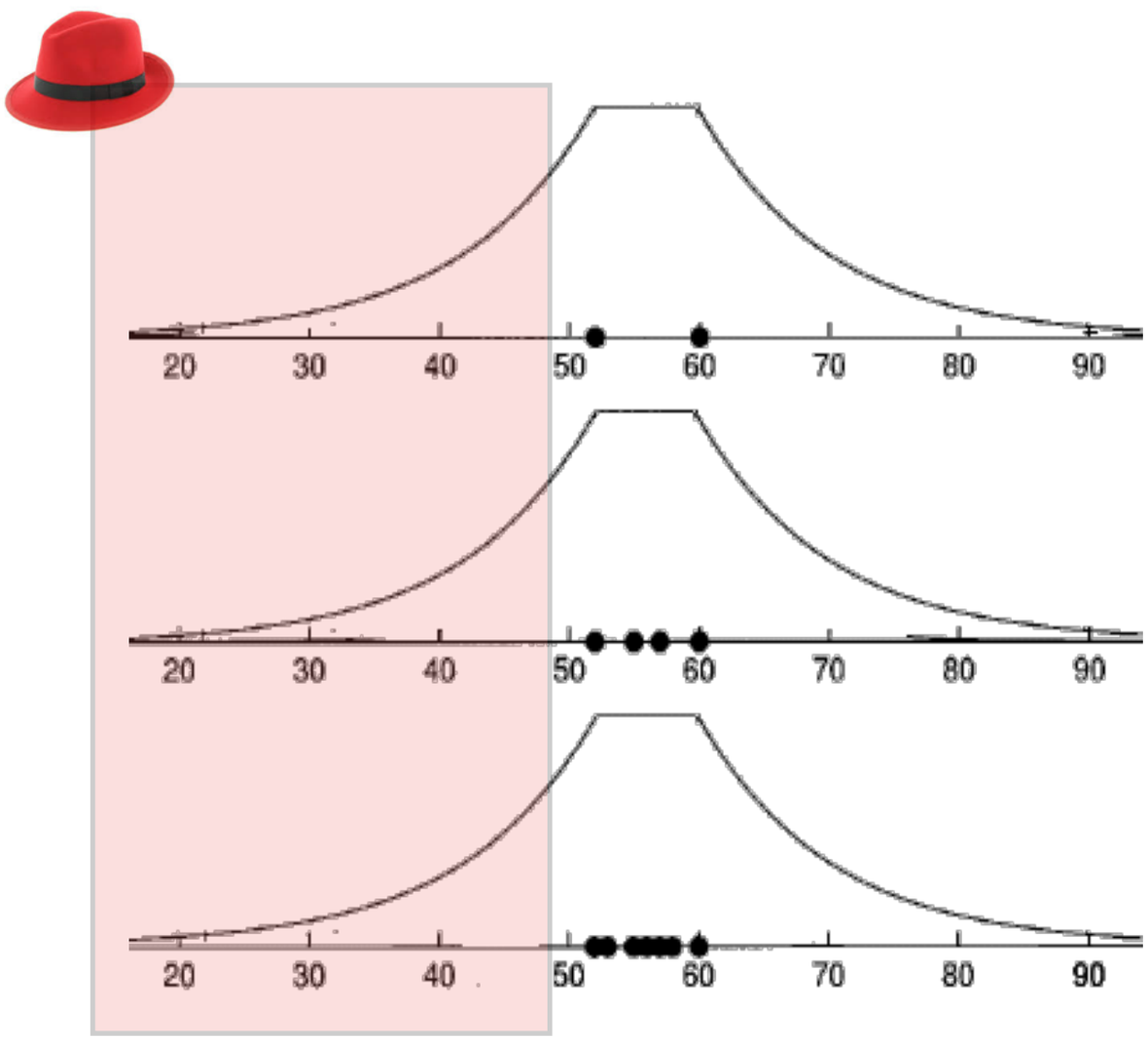
strong sampling



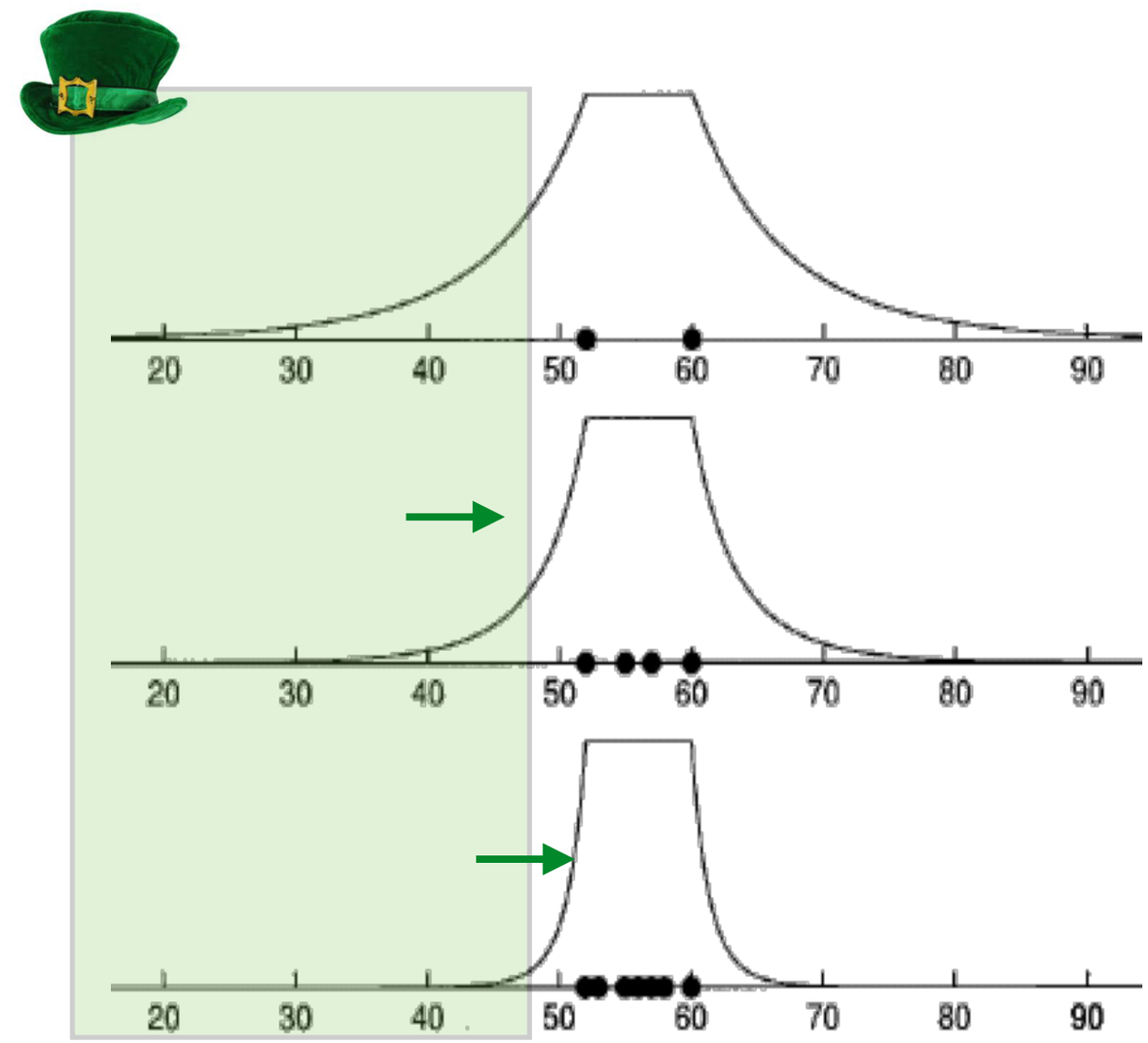
Tenenbaum & Griffiths (2001)

Here's the testable prediction about generalisation gradients…

weak sampling

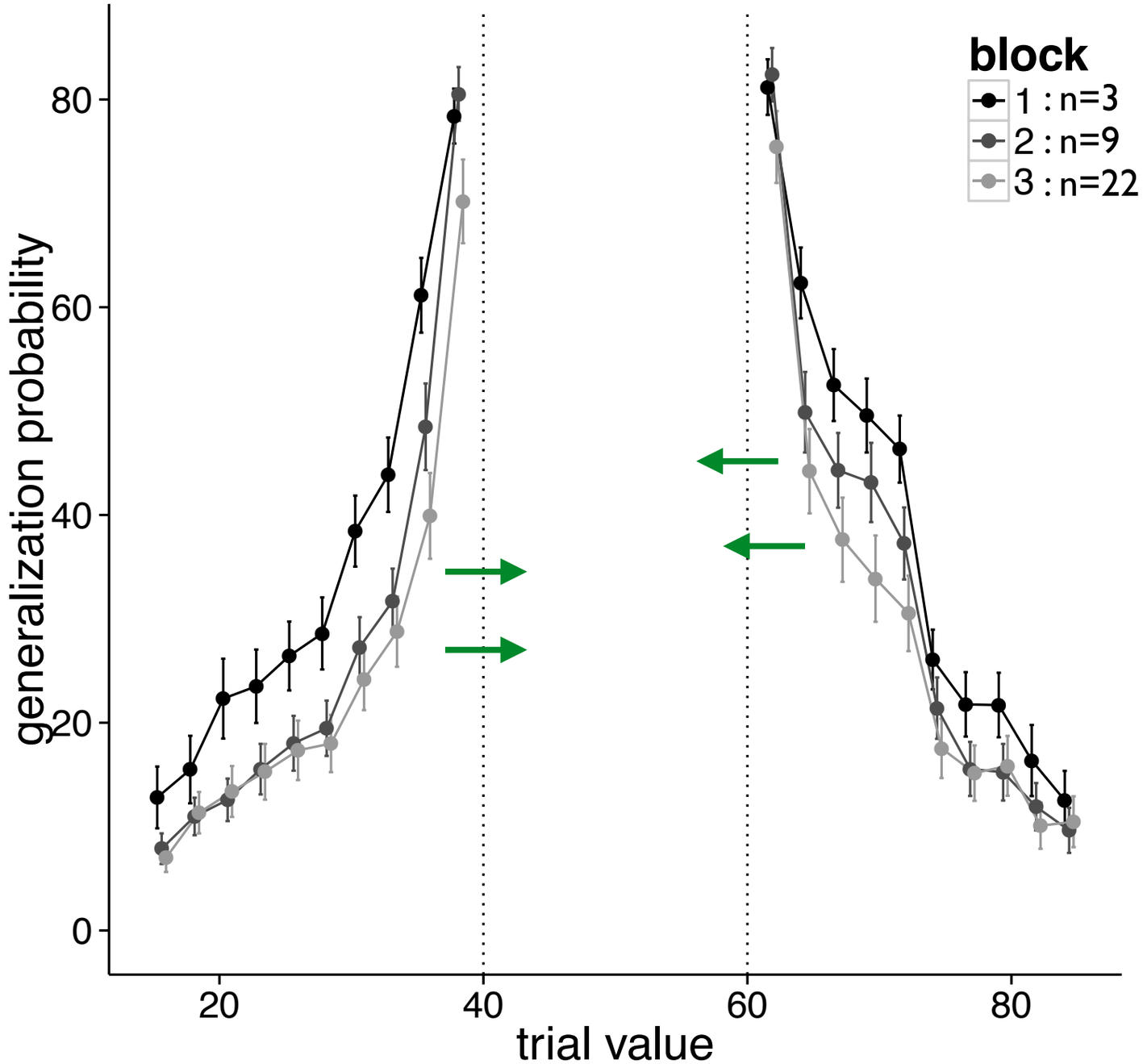


strong sampling



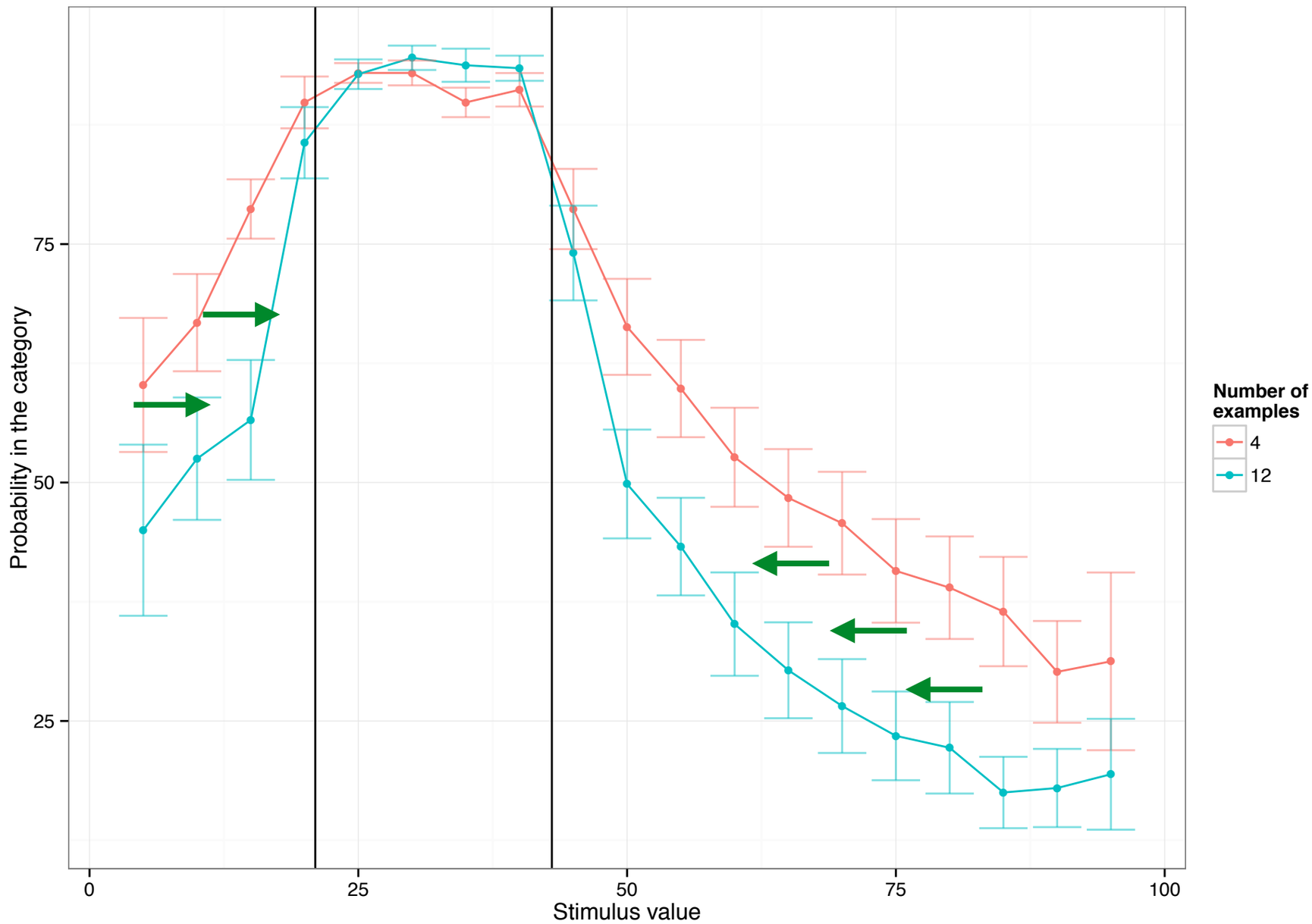
Tenenbaum & Griffiths (2001)

Looks like strong sampling…



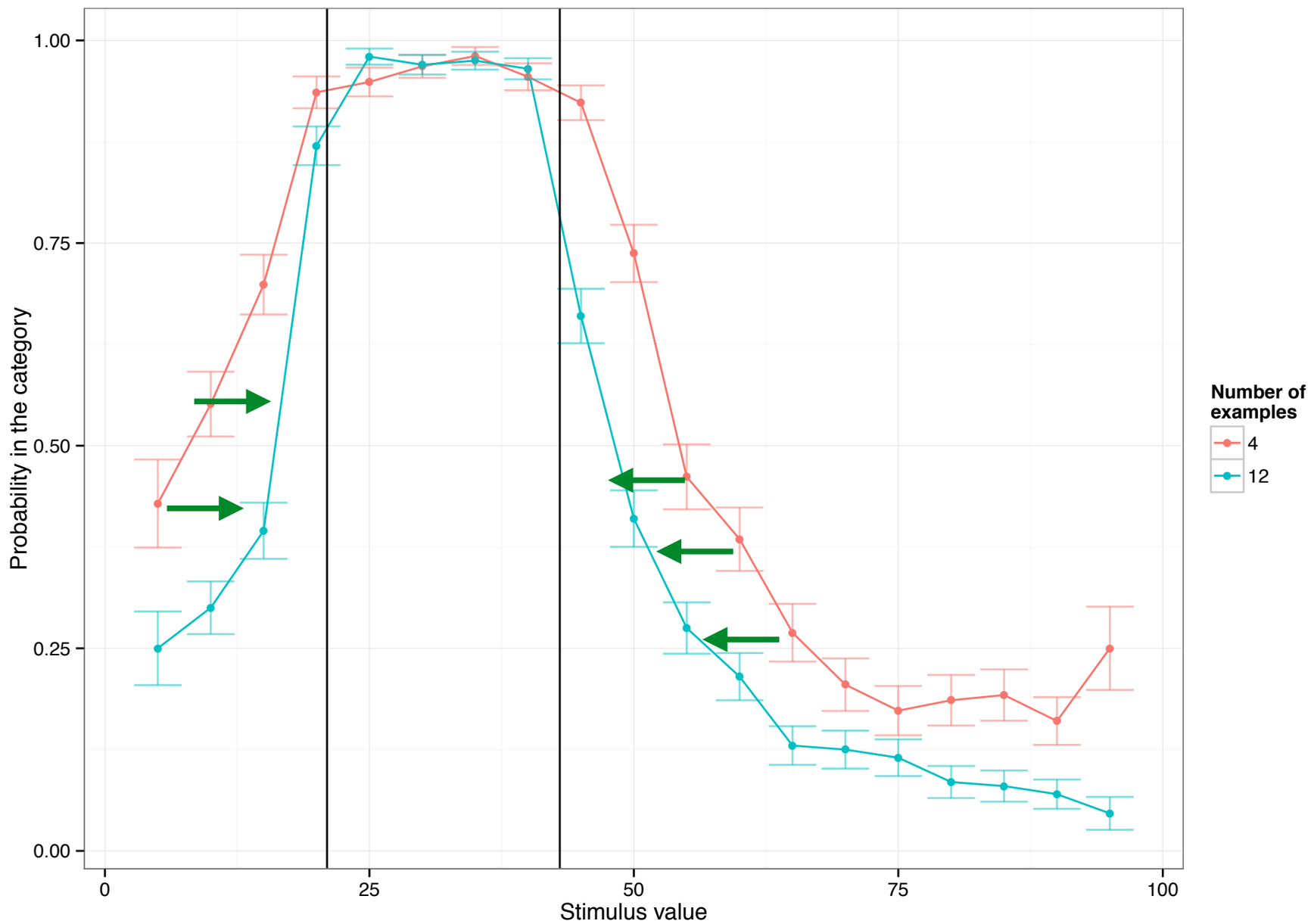
Vong, Hendrickson, Perfors & Navarro (2013). *Proc. Cog. Sci. Soc.*

Looks like strong sampling…



Hendrickson, Perfors & Navarro (in prep)

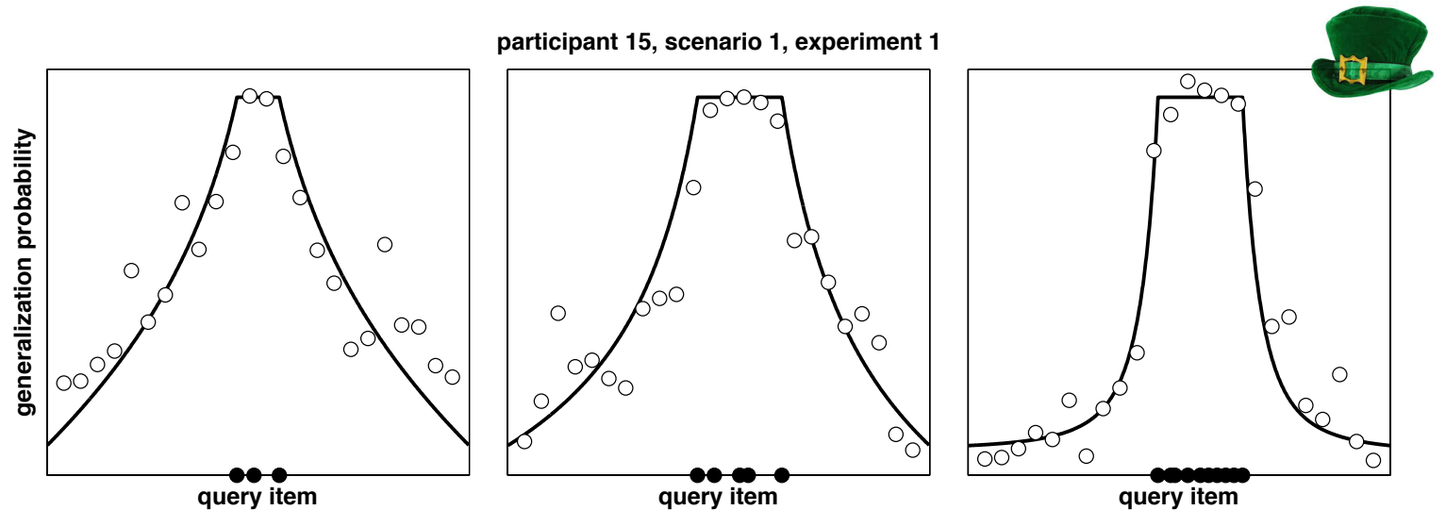
Looks like strong sampling…



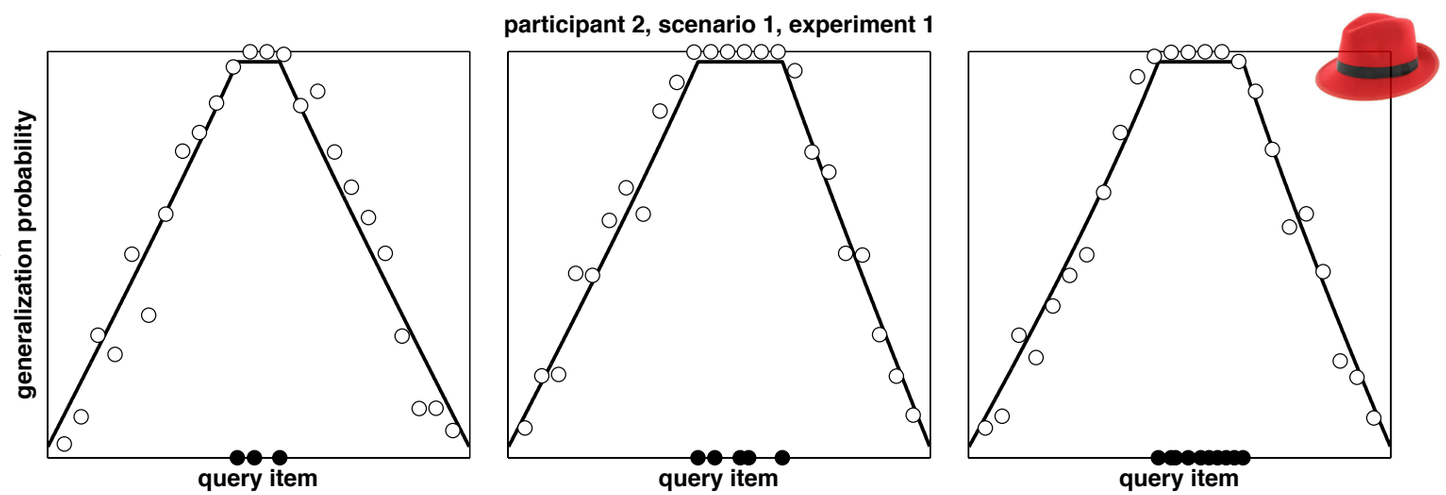
Hendrickson, Perfors & Navarro (in prep)

<u>But</u> there are individual differences:

Sensitivity to sample size in simple generalisation



Insensitivity to sample size in simple generalisation



Navarro, Dry & Lee (2012). *Cognitive Science*

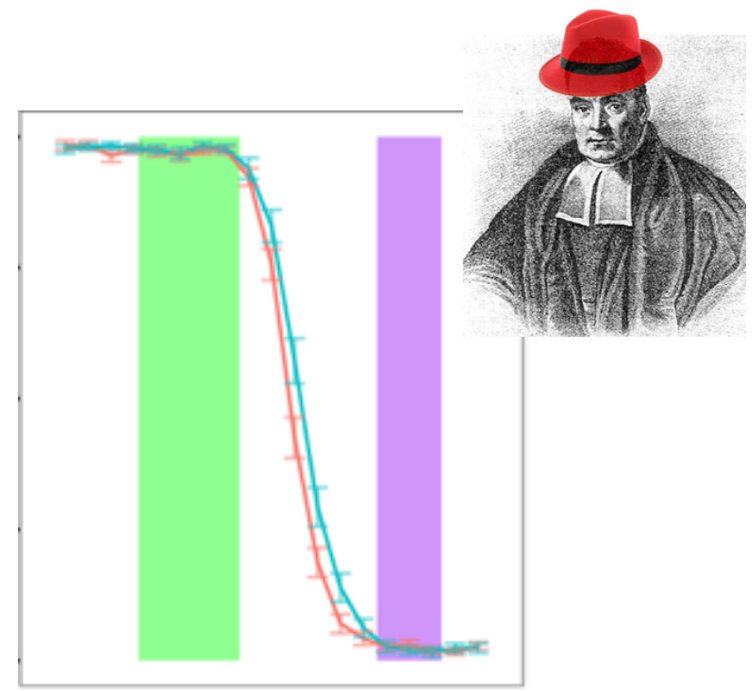
And there are <u>task</u> differences:

"Concept learning" designs where people see positive examples from one category produce the strong sampling "tightening" effect

Hendrickson, Perfors & Navarro (in prep)

And there are <u>task</u> differences:

"Concept learning" designs where people see positive examples from one category produce the strong sampling "tightening" effect



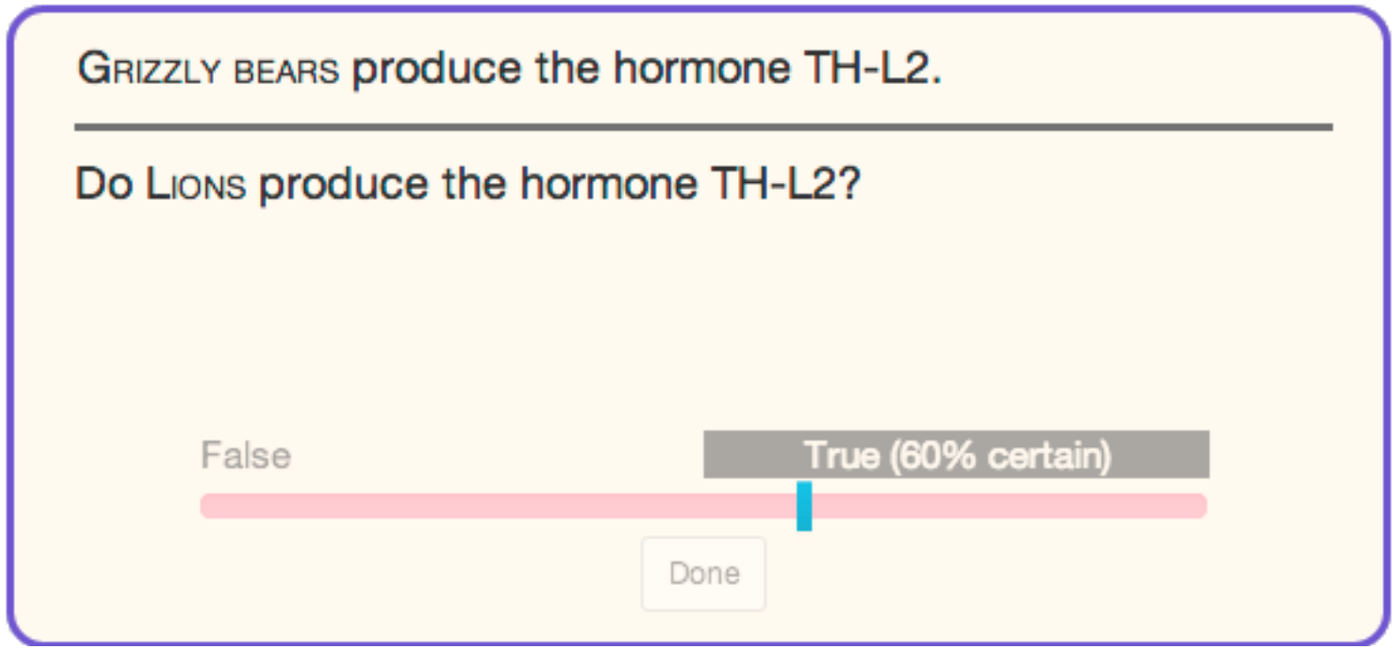
"Classification" designs where people see labelled examples from two categories show no tightening, only a weak base rate effect (in the opposite direction)

Hendrickson, Perfors & Navarro (in prep)

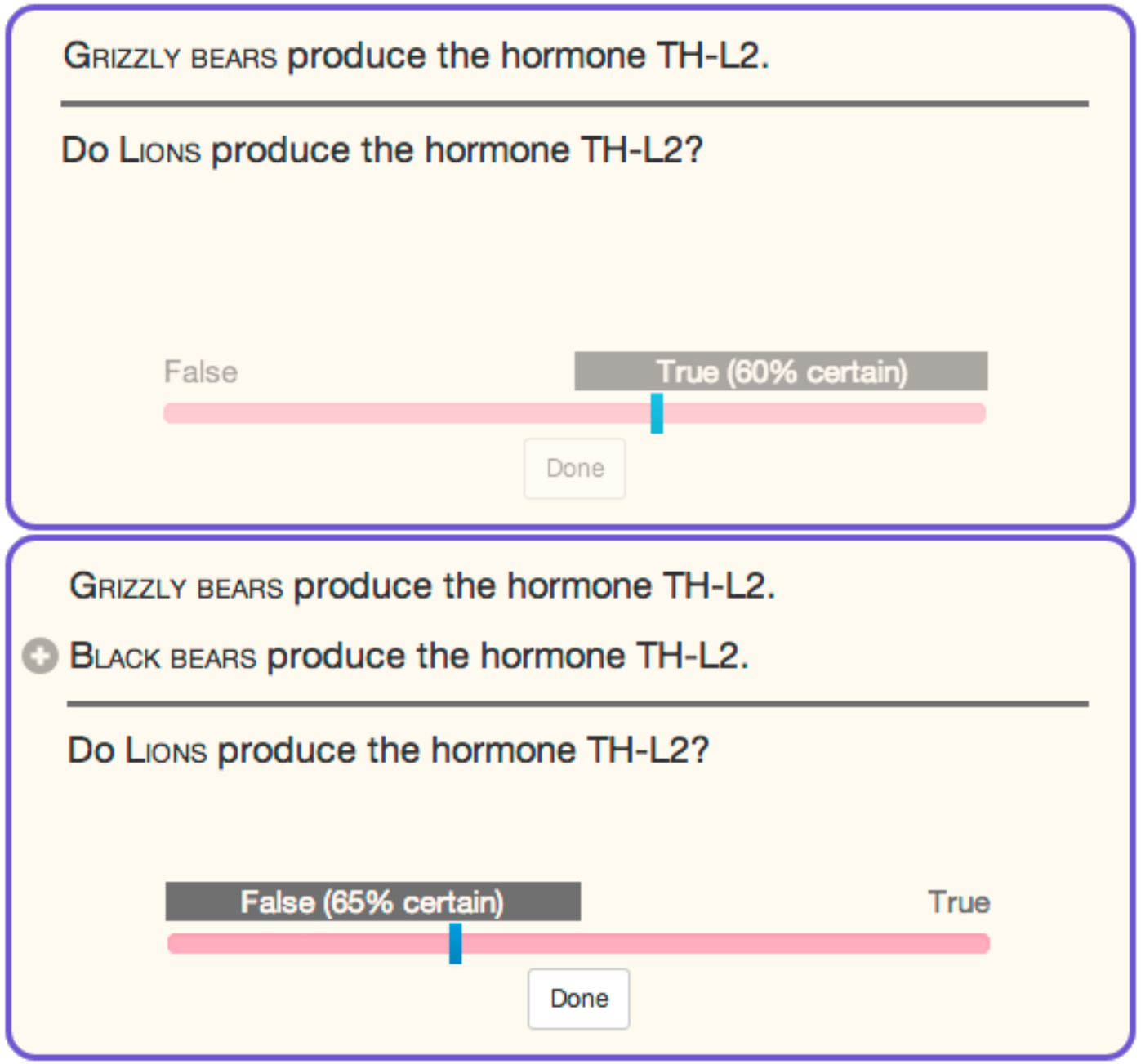
Manipulating the sampling assumption in an inductive reasoning task

Ransom, Perfors & Navarro (in press). *Cognitive Science*

Property induction tasks



Property induction tasks



#1 Grizzly Bears → Lions
#2 Grizzly Bears + Black Bears → Lions

<u>#2</u> … adding the "black bears" premise doesn't falsify any (reasonable) hypothesis

<u>#1</u> … it's a suspicious coincidence that both exemplars are bears

#1 Grizzly Bears → Lions

#2 Grizzly Bears + Black Bears → Lions

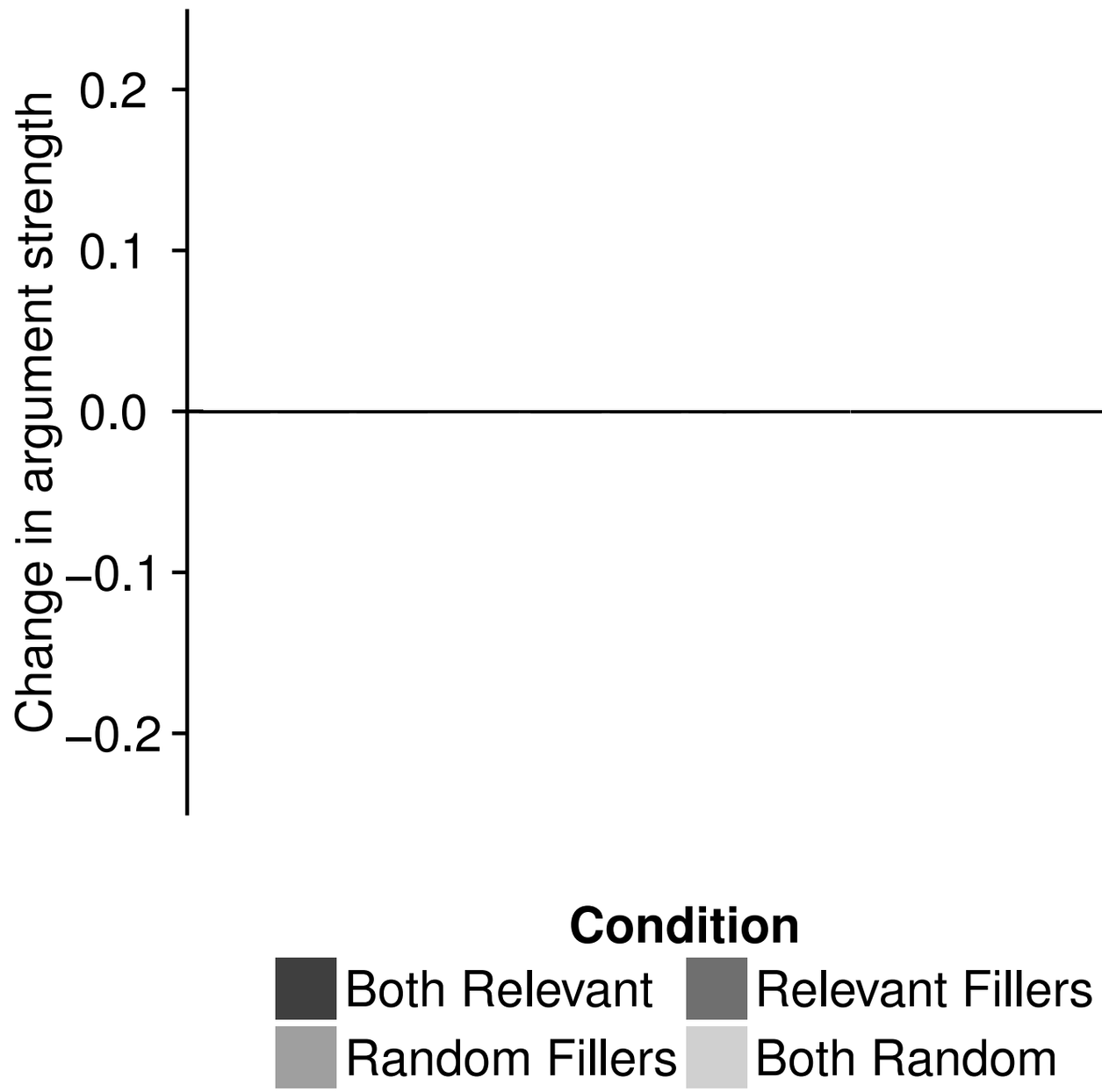
Weak sampling: If the premises are selected at random ***from the set of all true facts***, it's just bad luck that the two exemplars happen to be bears and it is not relevant to the inductive inference

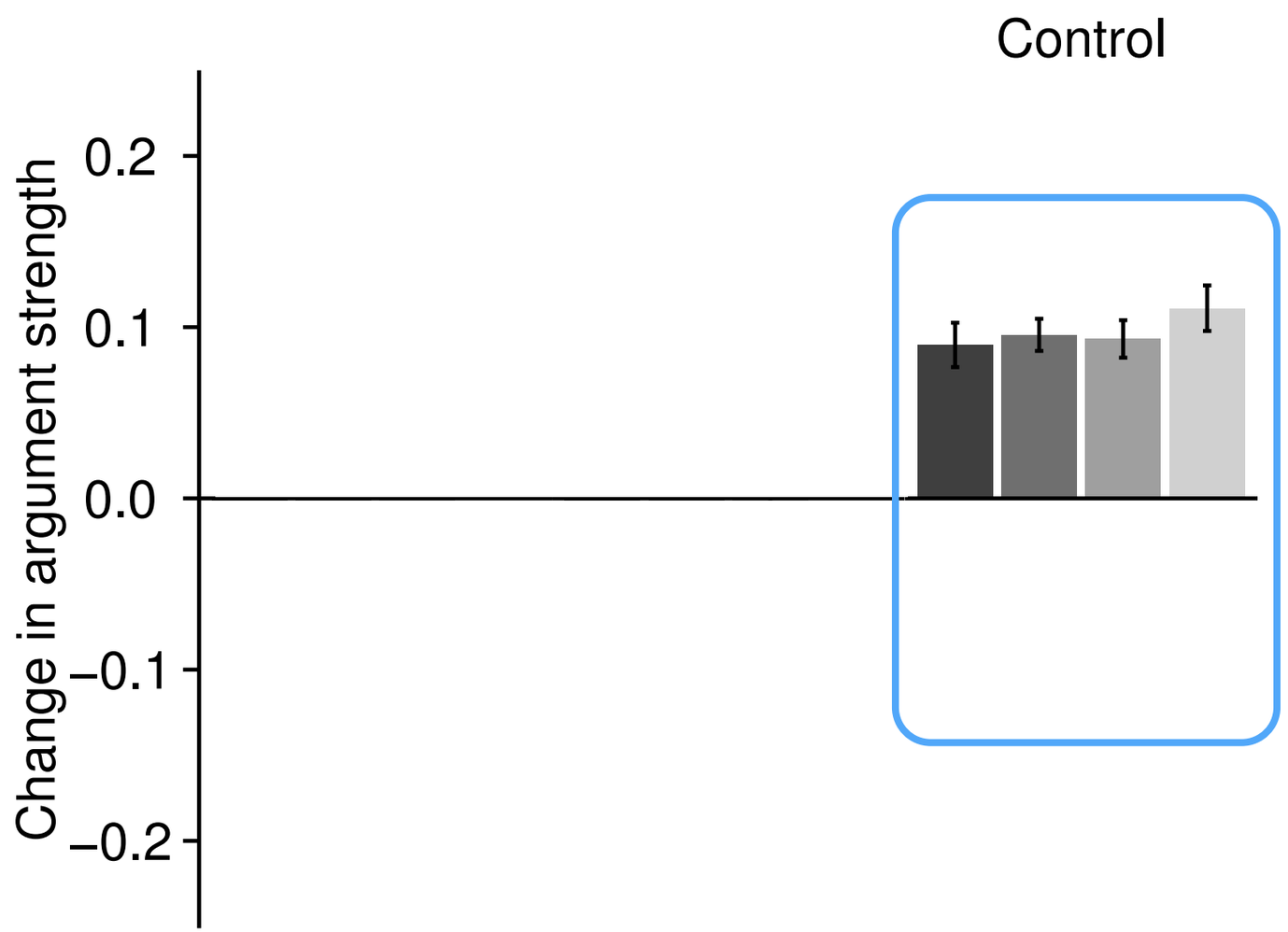
Strong sampling: If the premises are selected by sampling ***from a target category*** then the fact that both exemplars are both bears is entirely relevant and very informative.

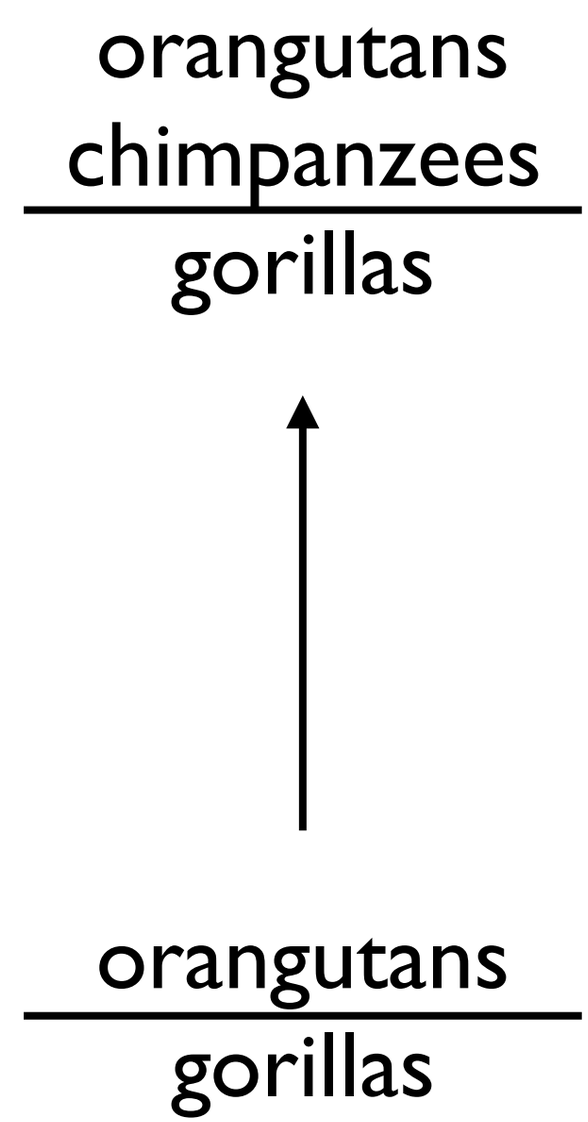
Previous experience?
(filler trials)

Cover story?

Relevant cover story, Relevant fillers	
Neutral cover story, Relevant fillers	Neutral cover story, Random fillers
	Random cover story, Random fillers

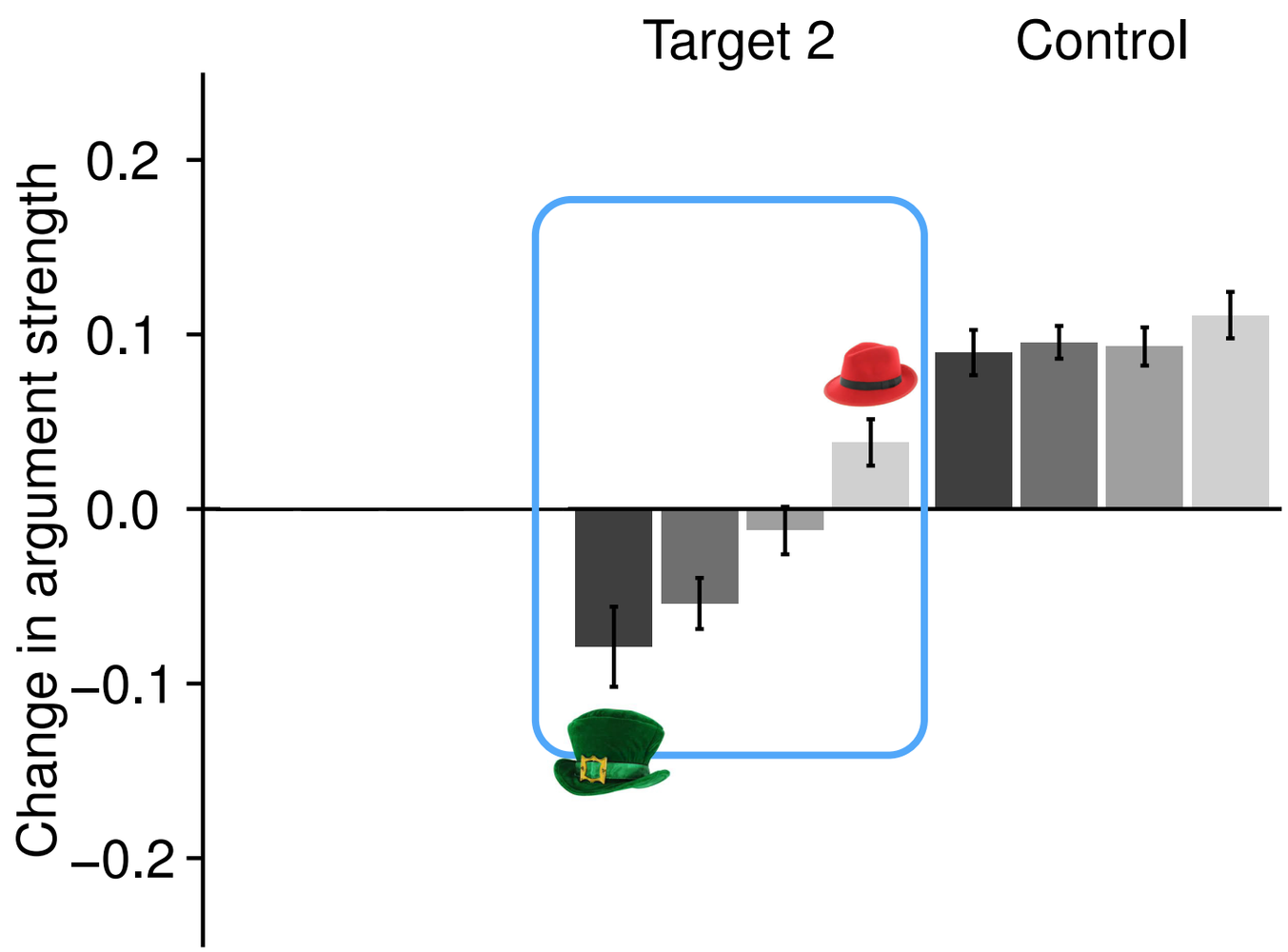

Change in argument strength
0.2
0.1
0.0
−0.1
−0.2
Condition
Both Relevant
Relevant Fillers
Random Fillers
Both Random

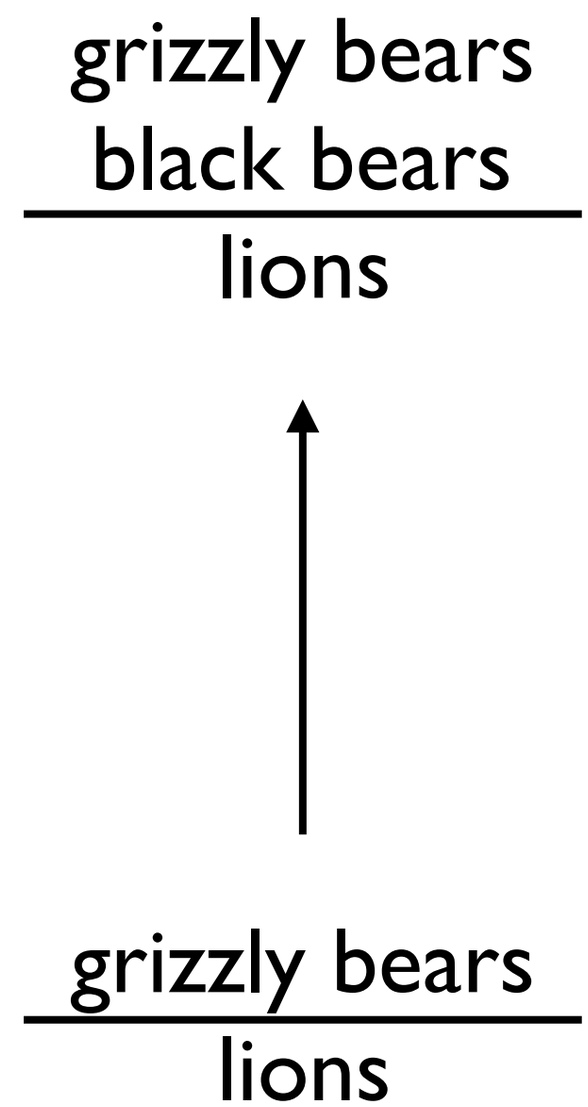




Condition

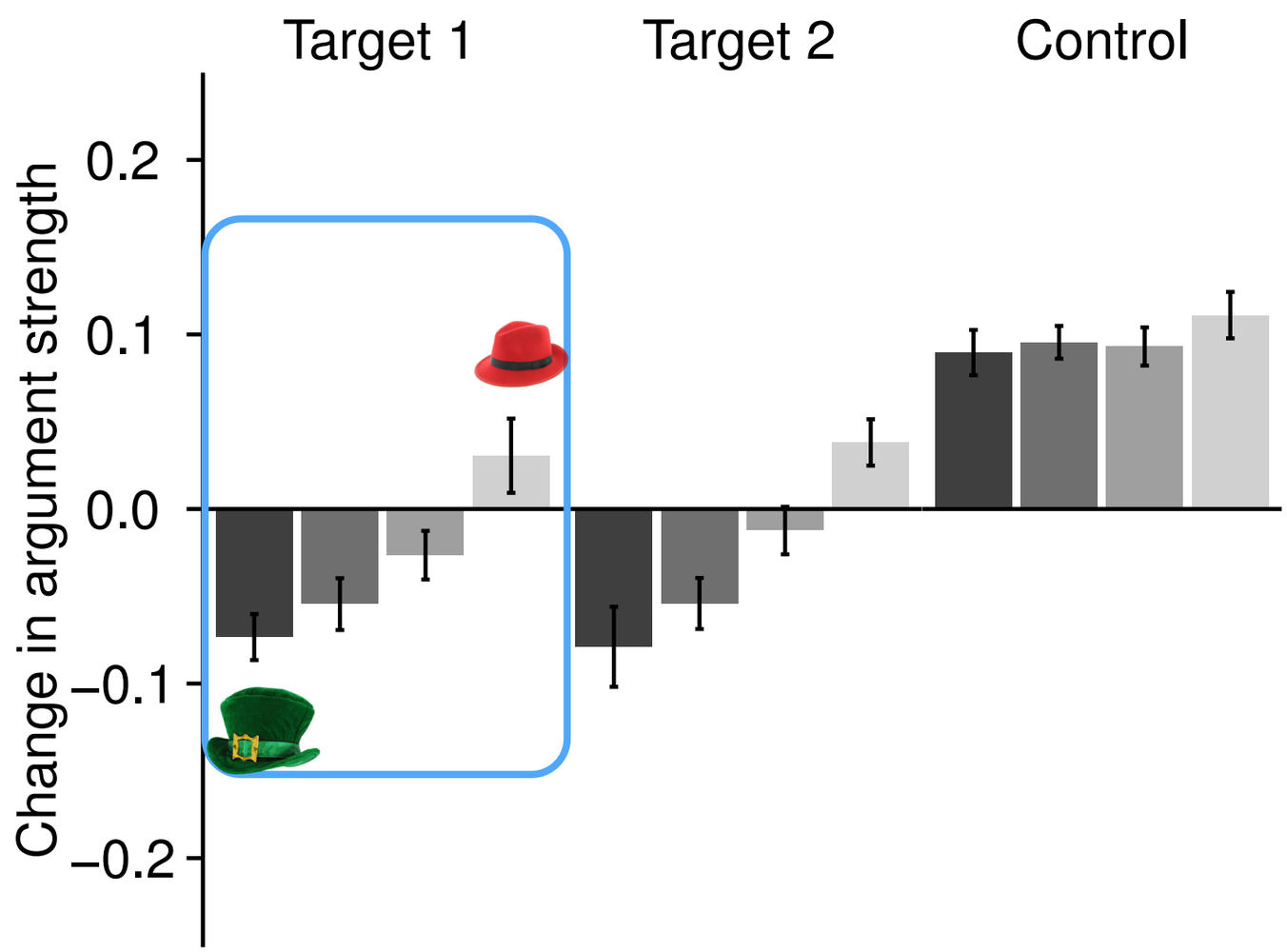
- Both Relevant
- Relevant Fillers
- Random Fillers
- Both Random

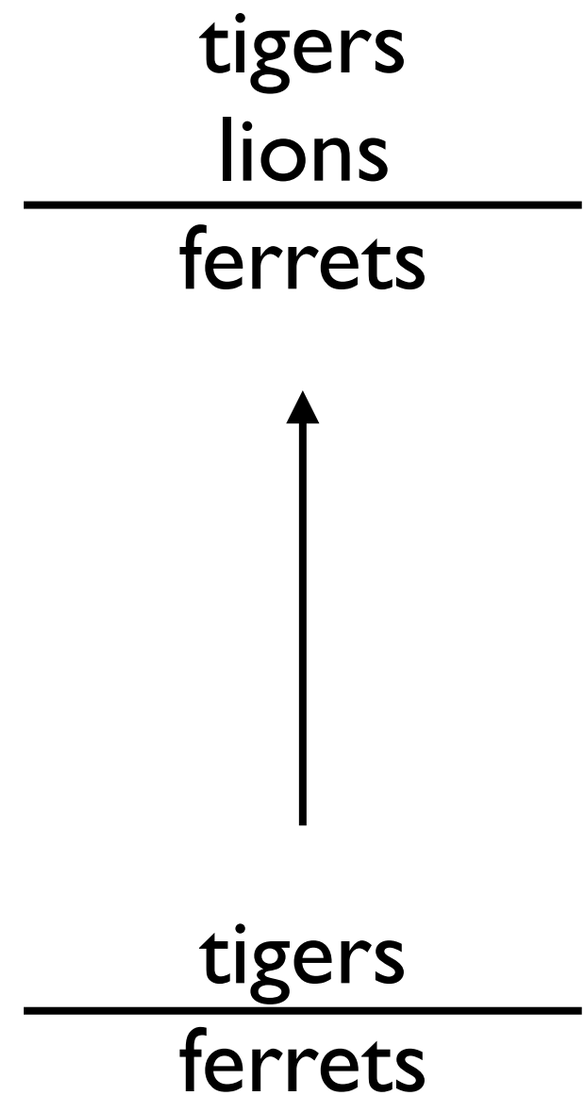




Condition

- Both Relevant
- Relevant Fillers
- Random Fillers
- Both Random



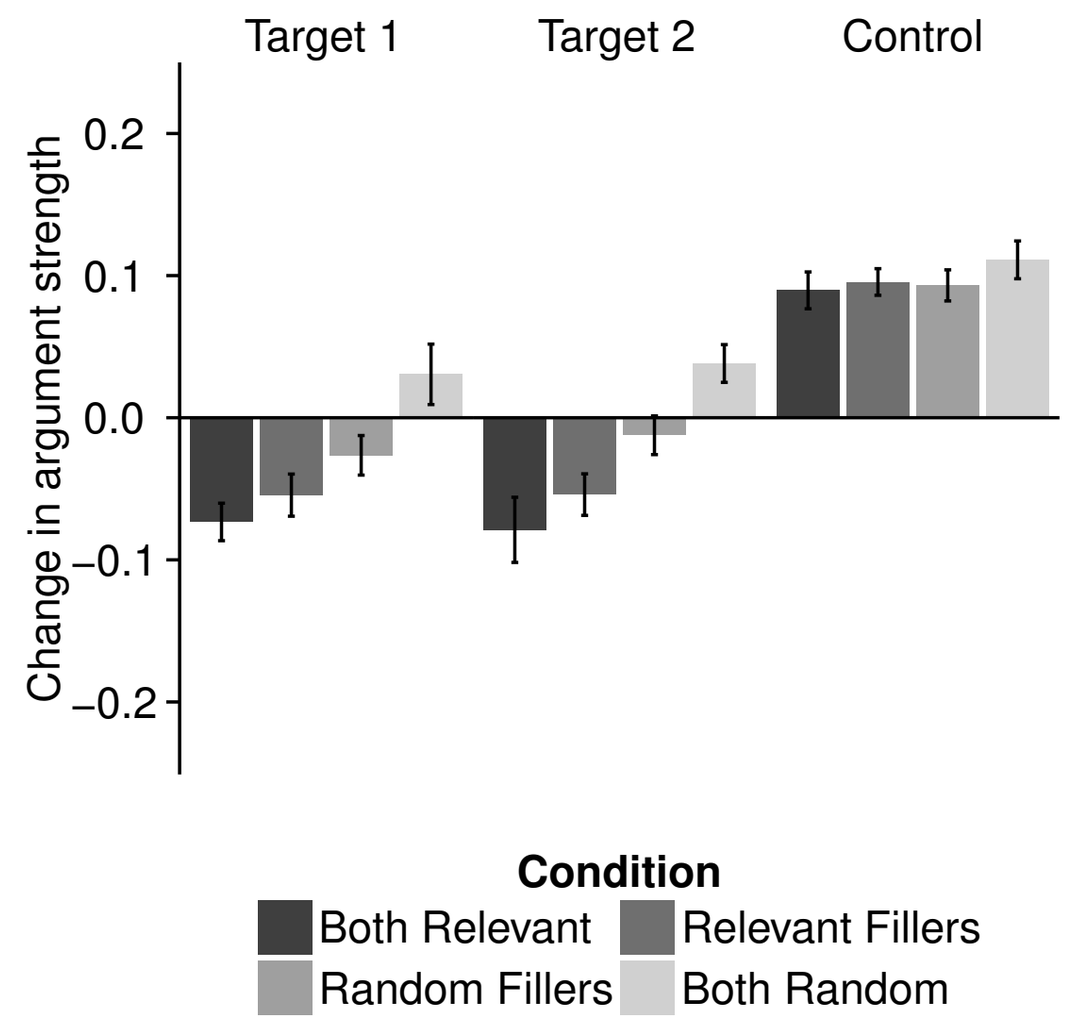


Condition

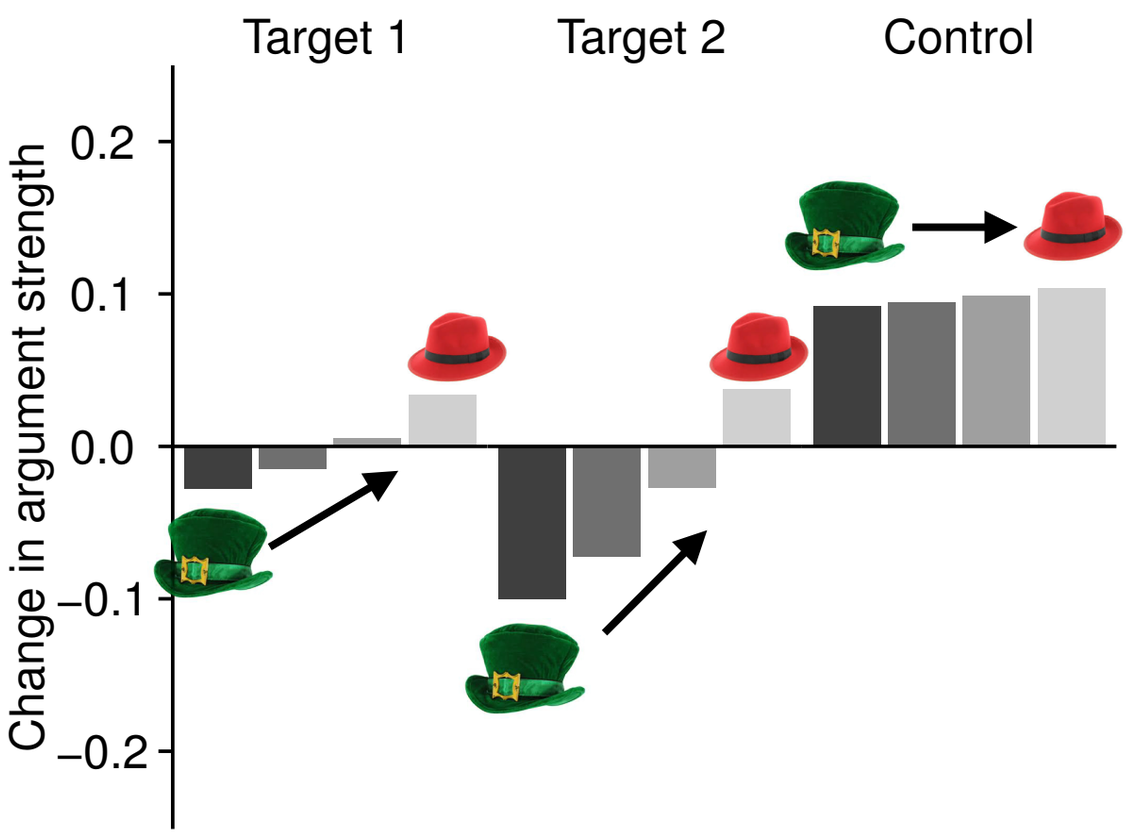
- Both Relevant
- Relevant Fillers
- Random Fillers
- Both Random

Empirical data

Is this consistent with Bayesian generalisation?



Yep.

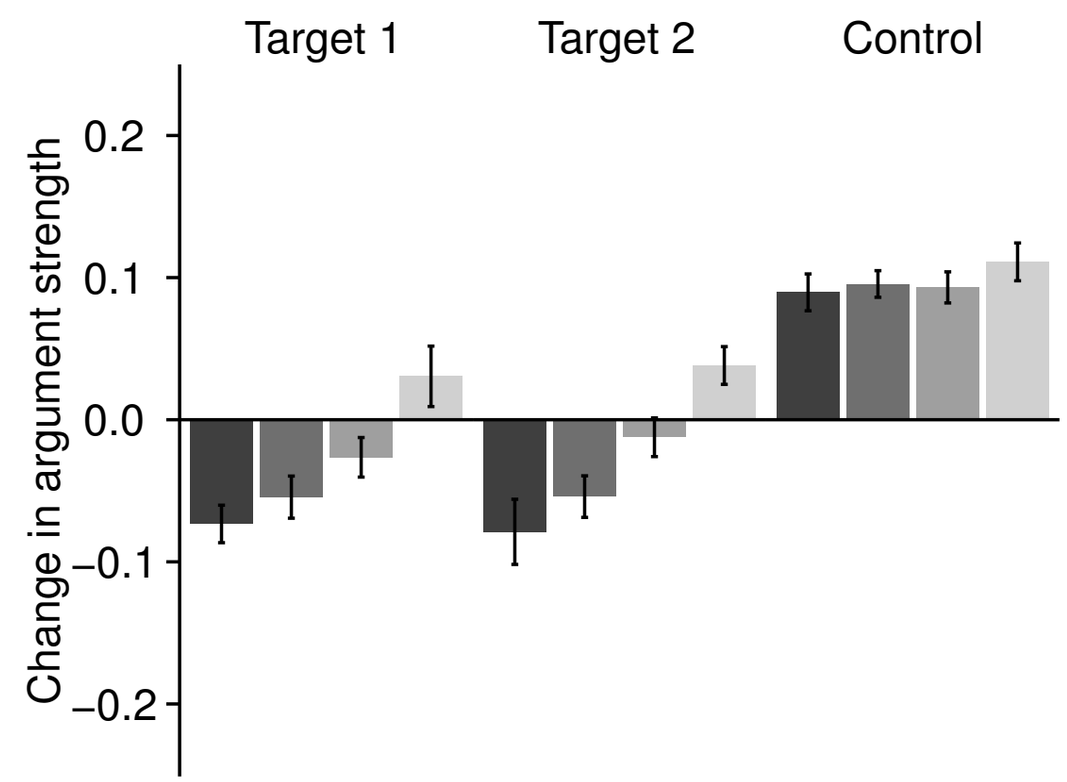


Mixed Sampling

$\theta = 0.31$ $\theta = 0.22$

$\theta = 0.11$ $\theta = 0$

Empirical data



Condition

Both Relevant | Relevant Fillers

Random Fillers | Both Random

Inductive reasoning depends on more than mere facts

- It's not just about the evidence that facts provide for a conclusion, it's also about <u>how</u> you think those facts were put together

- Bayesian models explain this as a shift in the sampling assumption

How to take a helpful hint…
(the curious power of negative evidence)

Voorspoels, Navarro, Perfors, Ransom & Storms (2015). *Cognitive Psychology*

You want to infer whether <u>all ravens are black</u>.
Which of these observations is more helpful?

Law of contraposition makes these two statements logically equivalent

Raven → Black

¬Black → ¬Raven

Okaaaay…. apparently these are the same?

Raven → Black

(raven, black)

¬Black → ¬Raven

(¬black, ¬raven)

Hempel (1945)

	Raven	¬Raven
Black		
¬Black	???	

	Raven	¬Raven
Black		
¬Black	???	

Category size/frequency matters, theoretically & empirically

- Positive (labelled) categories are small
  - Oaksford & Chater (1998), Navarro & Perfors (2011), etc.

- Sampling from a small category is more powerful
  - Good (1960), Klayman & Ha (1987), Oaksford & Chater (1998), Navarro & Perfors (2011), Austerweil & Griffiths (2011), etc

- People treat positive evidence as more informative than negative evidence
  - Wason (1960, 1968), many many others…
  - So it all makes sense! And…

Paradox resolved!

A black raven is very informative

A non-black non-raven has very modest evidentiary value

So we'll just some empirical work, with some *obviously* predictable results…

Mozart produces alpha waves

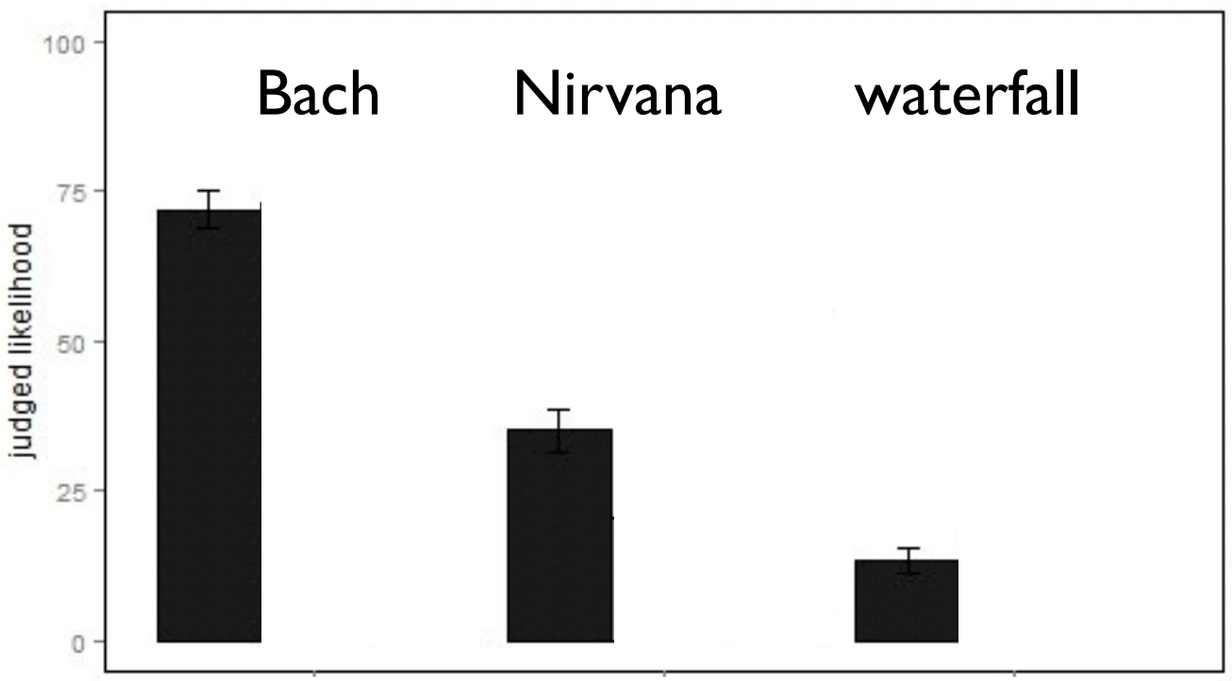
The sound of a falling rock does not

	music	¬music
alpha		
¬alpha	???	 This ought to be about as utterly useless as the green shoes thing

Okay, we start by telling people that Mozart does produce alpha waves…

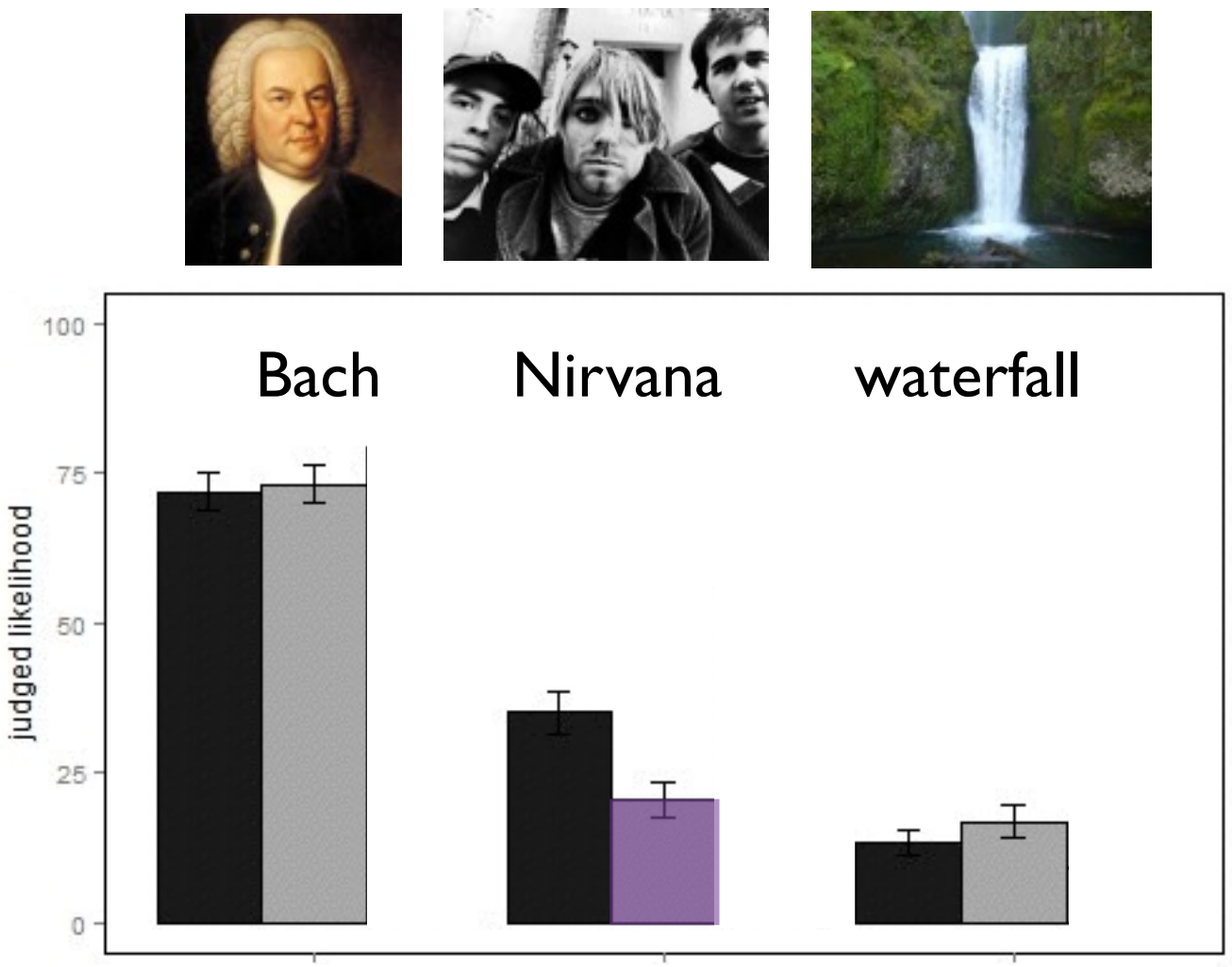
+Mozart





+Mozart

… and they generalise in a way that seems terribly sensible

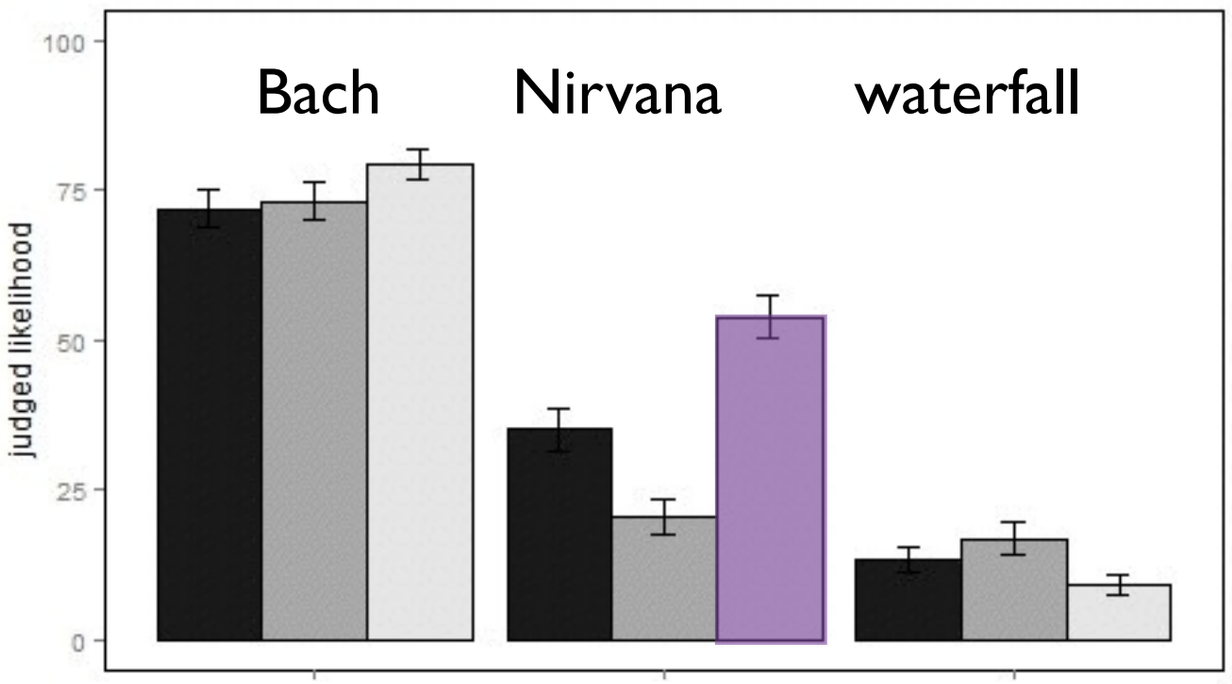


+Mozart

-Metallica

Adding Metallica as a negative example has a modest, sensible effect on inferences about Nirvana

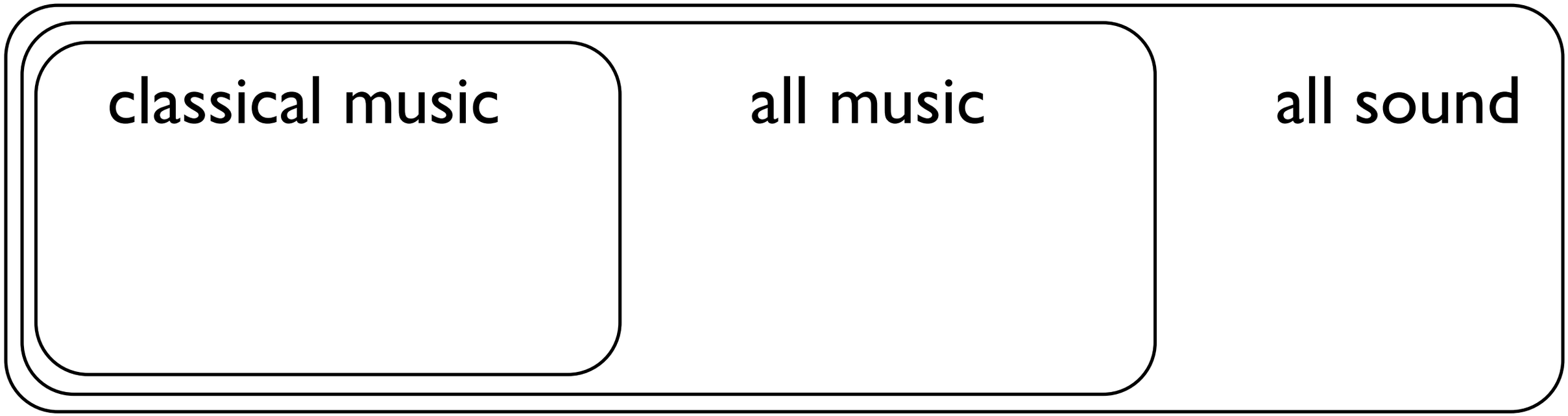




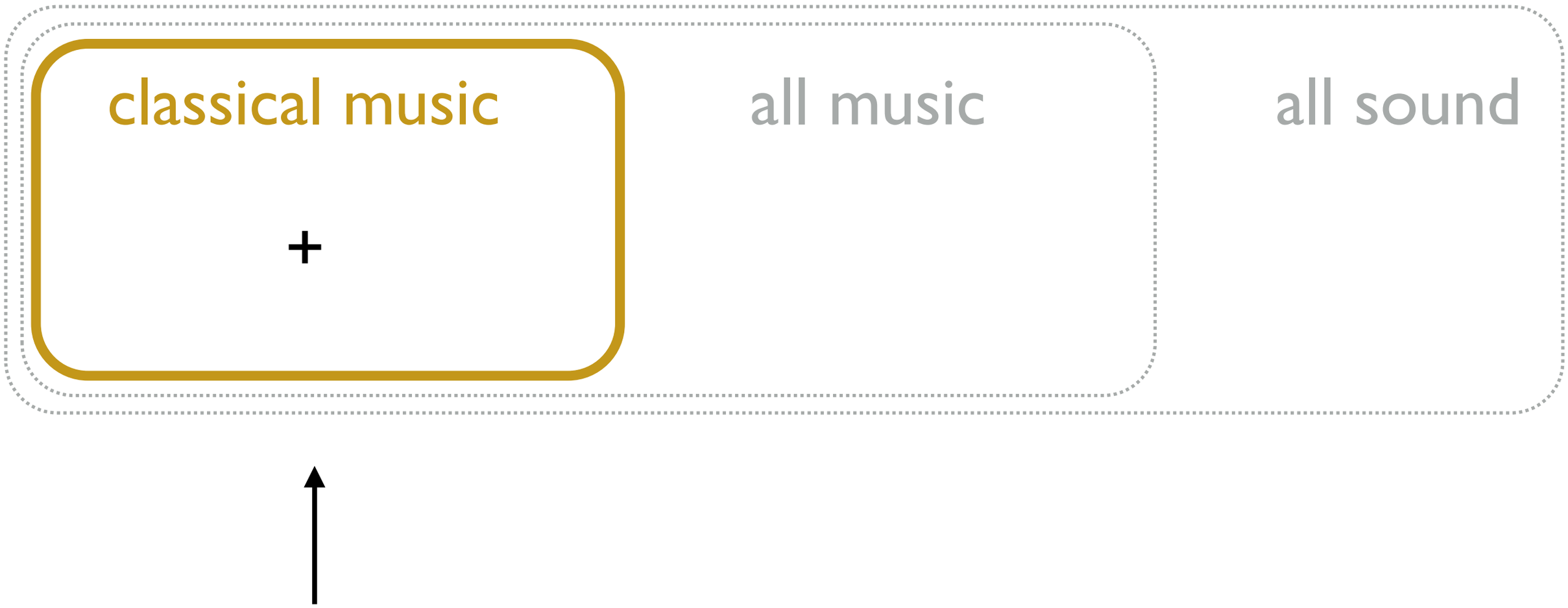
sigh.

+Mozart

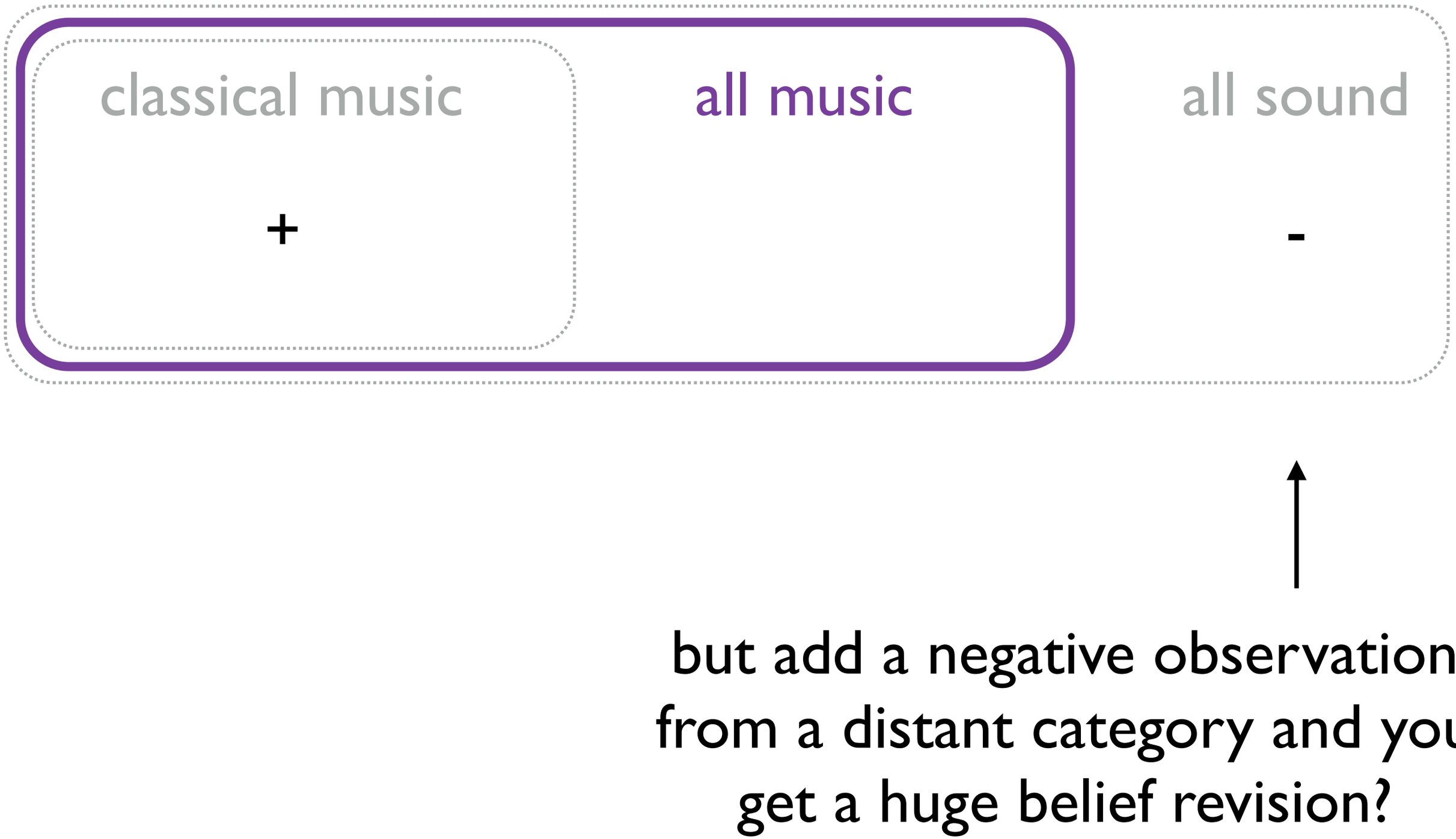
-Falling rock



three relevant hypotheses for the extension of the alpha waves property

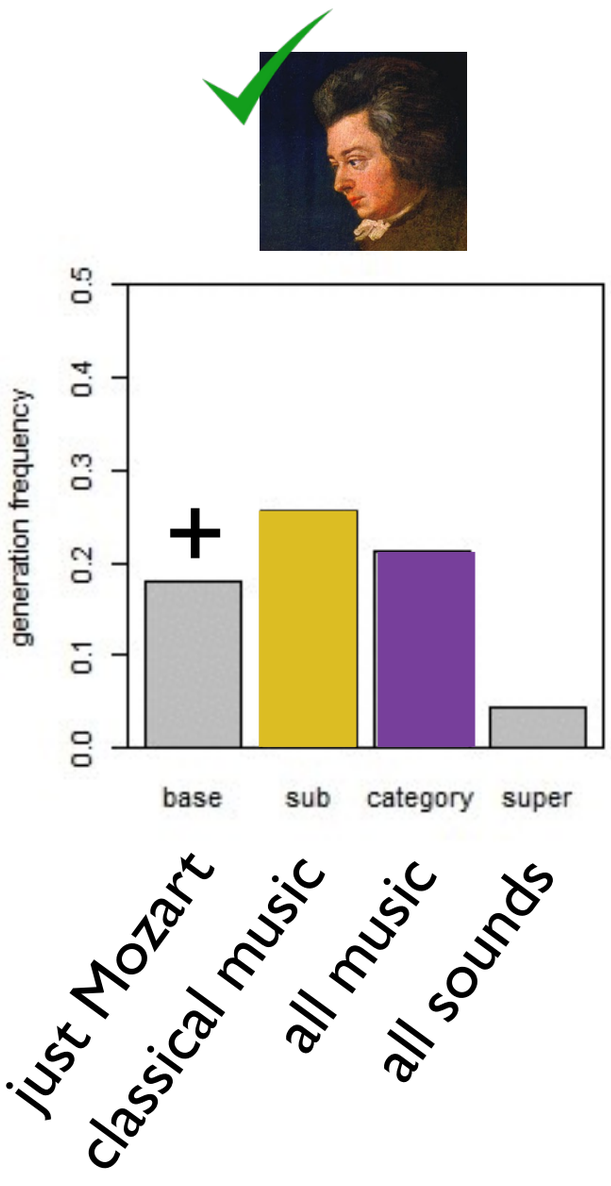


positive example of classical music means people strongly endorse the narrow category

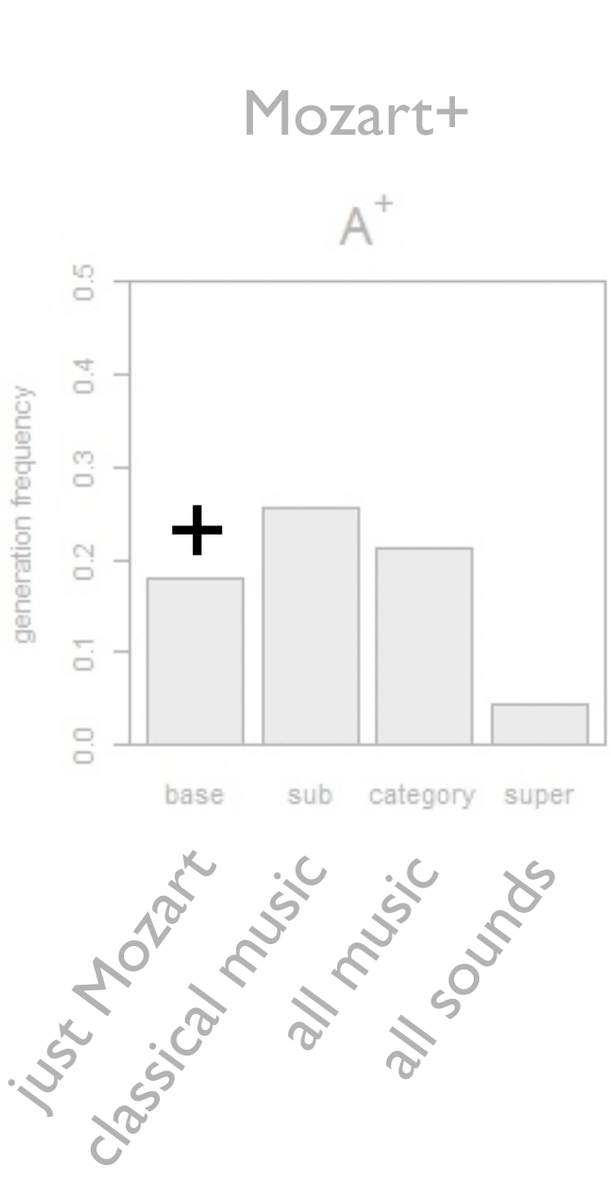

classical music
all music
all sound
+
-
but add a negative observation
from a distant category and you
get a huge belief revision?

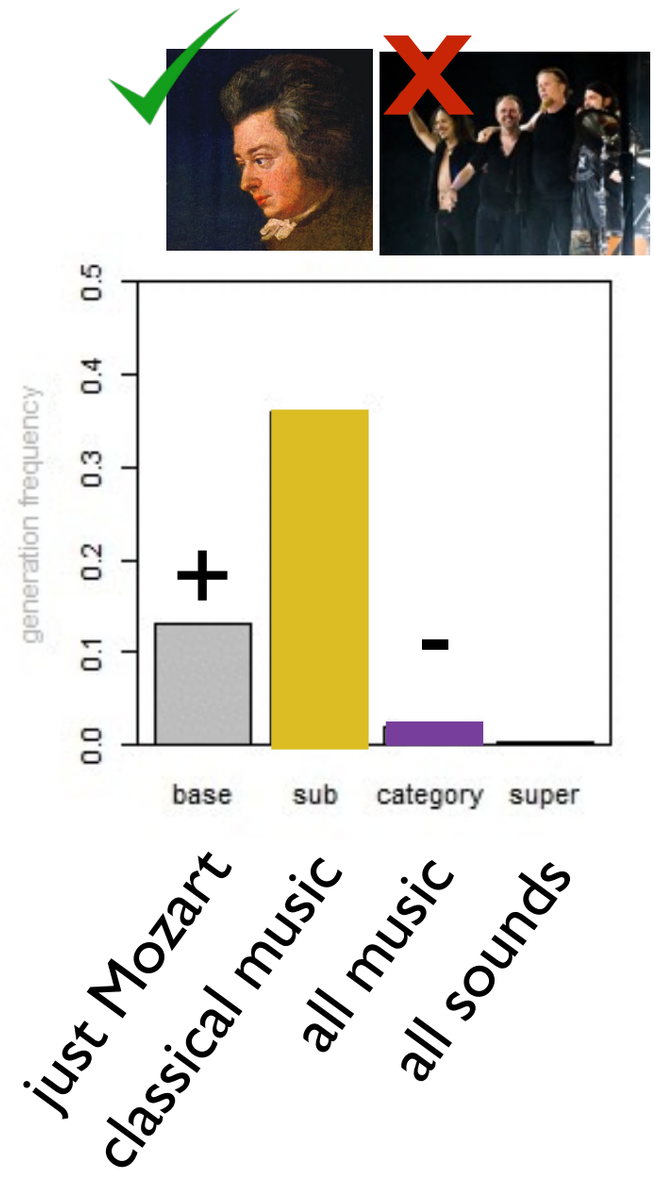
Apparently people make a (pragmatic?) inference that the negative observation is used to demarcate the category boundary

Well, let's ask them what they think the true extension of the property is…

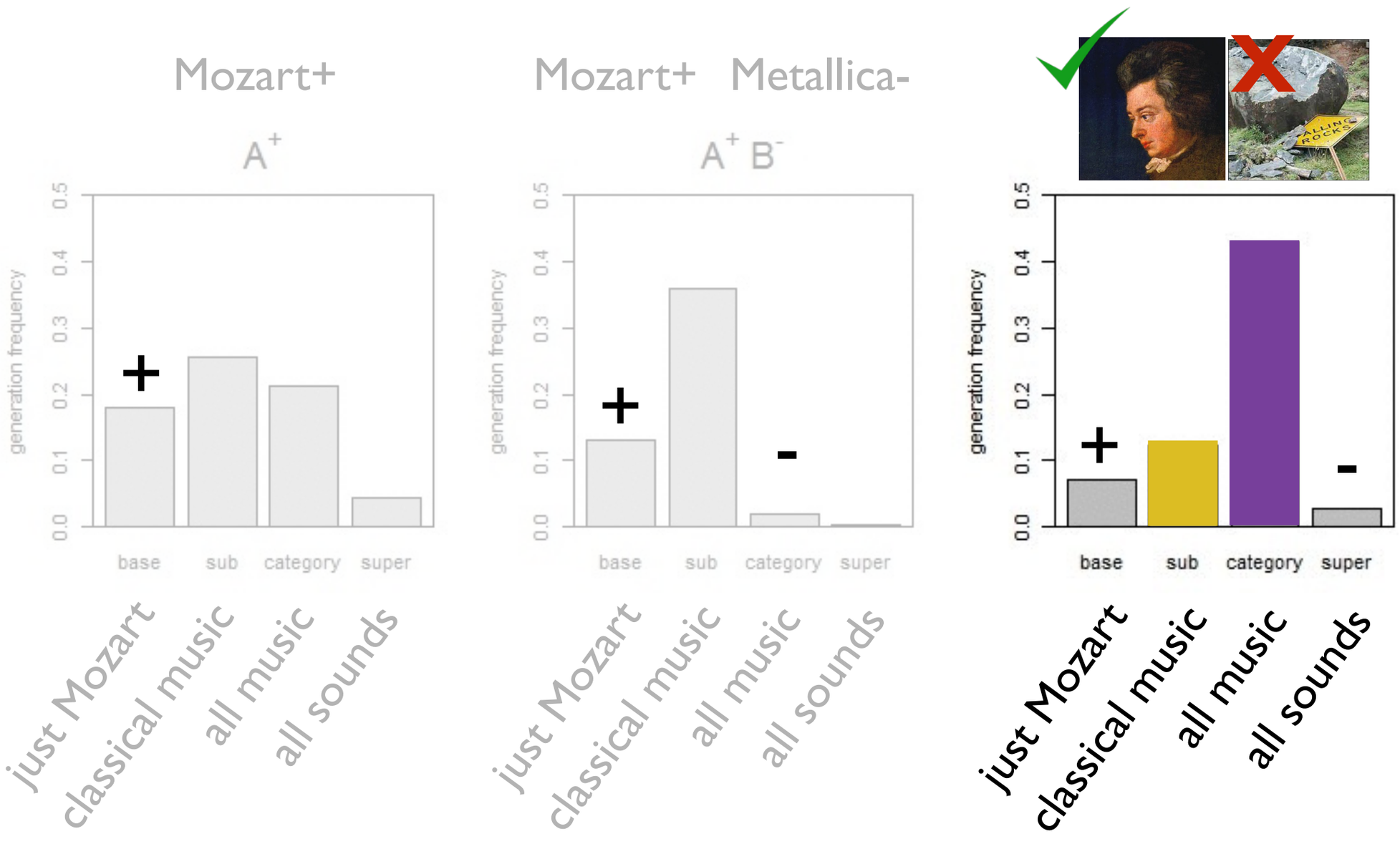


Well, let's ask them what they think the true extension of the property is…

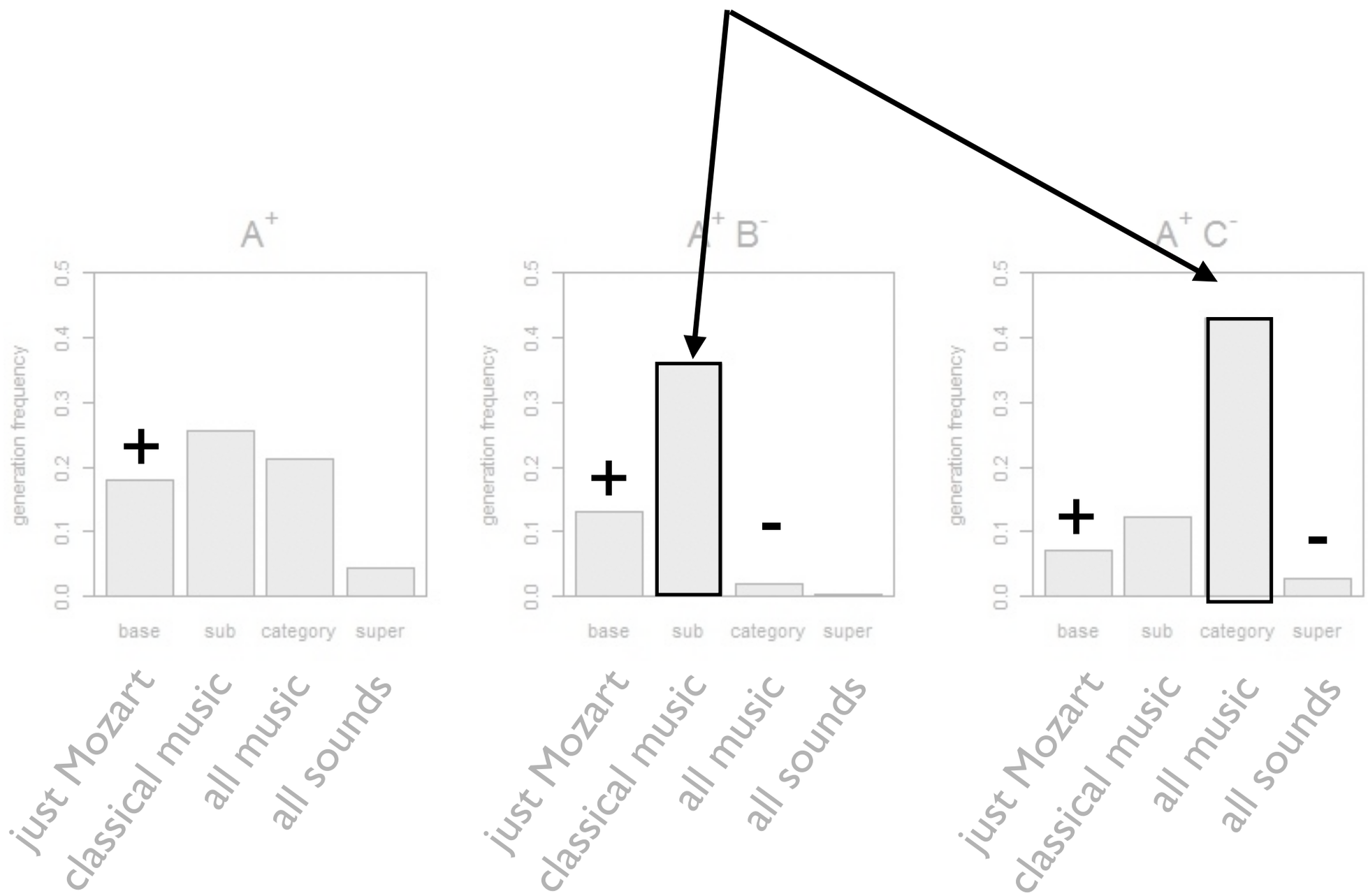




Yep


Mozart+
A+
generation frequency
0.0
0.1
0.2
0.3
0.4
0.5
+
base
sub
category
super
just Mozart
classical music
all music
all sounds
Mozart+ Metallica-
A+ B-
generation frequency
0.0
0.1
0.2
0.3
0.4
0.5
+
-
base
sub
category
super
just Mozart
classical music
all music
all sounds
generation frequency
0.0
0.1
0.2
0.3
0.4
0.5
+
-
base
sub
category
super
just Mozart
classical music
all music
all sounds

The negative observation shifts belief to the largest category that excludes it



(aside: the actual experiment used many different arguments)

topic	premises			conclusions		
	subcat A	subcat B	cat C	A-member	B-member	C-member
MUSIC	Mozart	Metallica	falling rock	Bach	Nirvana	waterfall
PAINTERS	Rubens	Dahli	woodcarver	Van Eyck	Warhol	sculpturer
PUBLIC FIGURES	actors	librarians	moles	politicians	programmers	pheasants
SHIPS	freight ships	hovercrafts	cars	cruise ships	sail boats	rocks
GLASS	window glass	bottle glass	art glass	car glass	drinking glass	jewelry glass
DISPLAYS	LCD	television	paintings	plasma	traffic signs	book page
WATER BODIES	Atlantic	Balaton	mustard gass	Mediterranean	Silverlake	olive oil
WIND	flute	guitar	crying child	clarinet	violin	door
FRUIT	strawberries	banana's	grass blades	cranberries	apples	oak leafs
WATER BIRDS	ducks	sparrows	elephants	seagulls	blackbirds	camels
INSECTS	moths	spiders	lizzards	flies	centipede	goldfish
POLAR ANIMALS	polar bears	deer	sow bug	pinguins	parakeet	ant

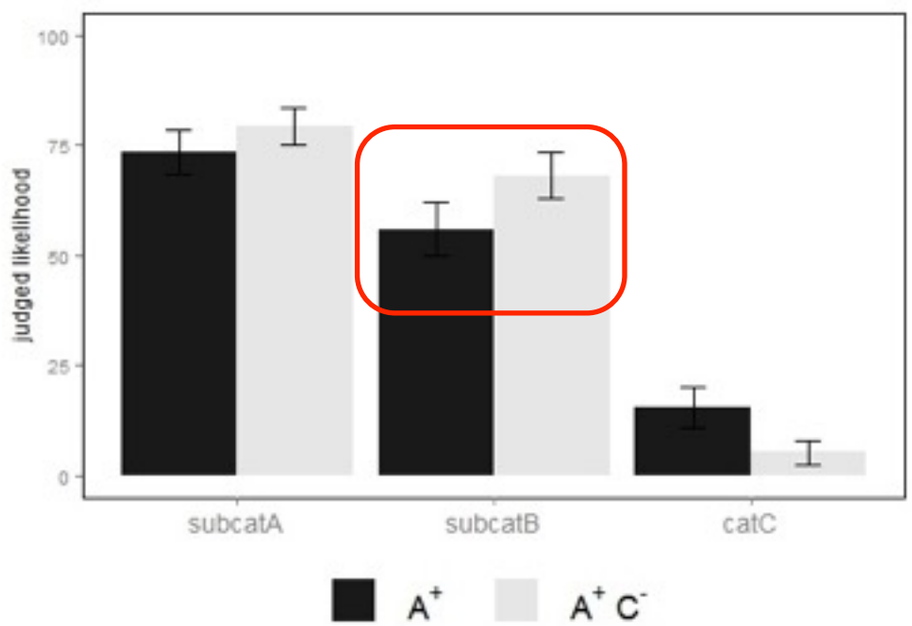
(aside: the actual experiment used many different arguments)

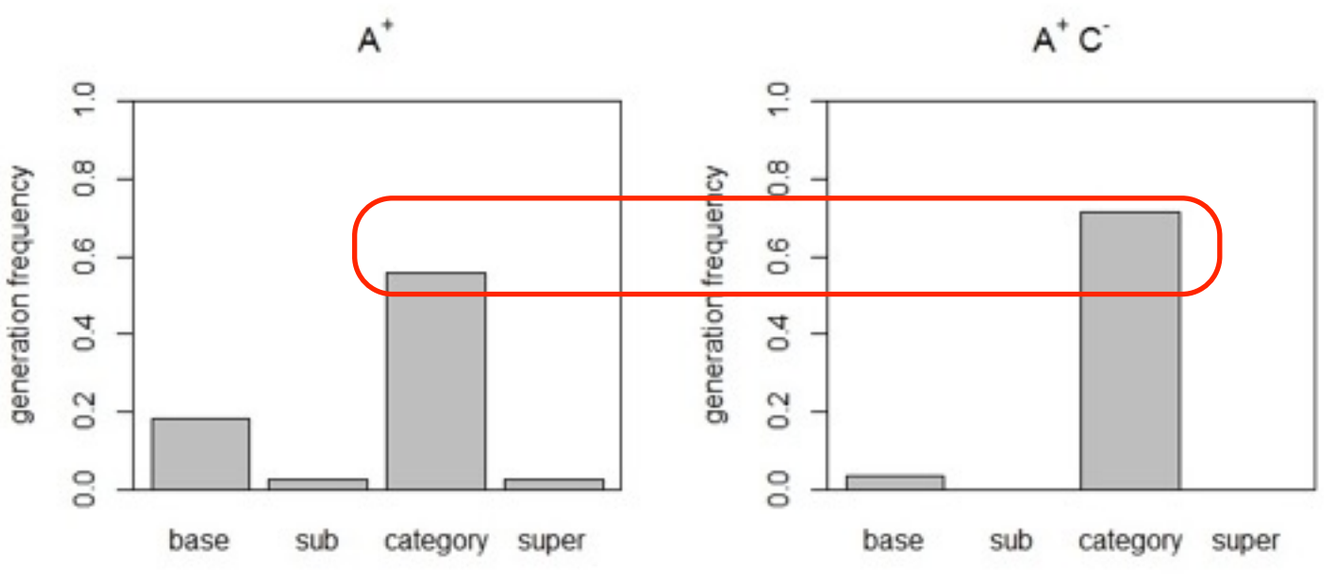
topic	premises			conclusions		
	subcat A	subcat B	cat C	A-member	B-member	C-member
MUSIC	Mozart	Metallica	falling rock	Bach	Nirvana	waterfall
PAINTERS	Rubens	Dahli	woodcarver	Van Eyck	Warhol	sculpturer
PUBLIC FIGURES	actors	librarians	moles	politicians	programmers	pheasants
SHIPS	freight ships	hovercrafts	cars	cruise ships	sail boats	rocks
GLASS	window glass	bottle glass	art glass	car glass	drinking glass	jewelry glass
DISPLAYS	LCD	television	paintings	plasma	traffic signs	book page
WATER BODIES	Atlantic	Balaton	mustard gass	Mediterranean	Silverlake	olive oil
WIND	flute	guitar	crying child	clarinet	violin	door
FRUIT	strawberries	banana's	grass blades	cranberries	apples	oak leafs
WATER BIRDS	ducks	sparrows	elephants	seagulls	blackbirds	camels
INSECTS	moths	spiders	lizzards	flies	centipede	goldfish
POLAR ANIMALS	polar bears	deer	sow bug	pinguins	parakeet	ant

plus we ran an entire pseudo-replication with different items

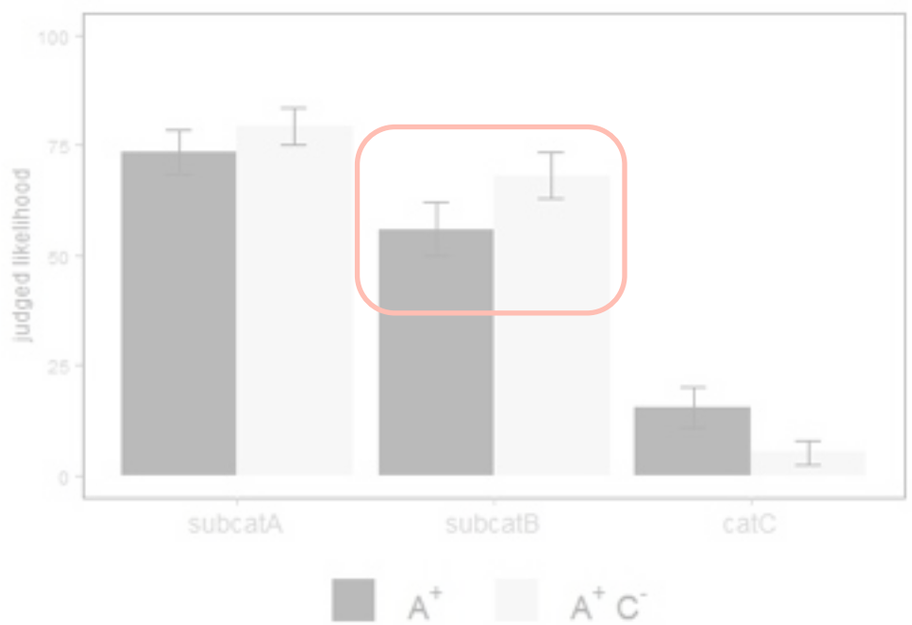
topic	premises			conclusions	
	subcat A	cat C	A-member	B-member	C-member
MAMMALS	dog (+)	magpie (-)	wolf	donkey	blackbird
BIRDS	crow (+)	tuna fish (-)	raven	swan	halibot
FISH	salmon (+)	lizzard (-)	codfish	goldfish	snake
INSECTS	bee (+)	sparrow (-)	ant	cricket	pigeon

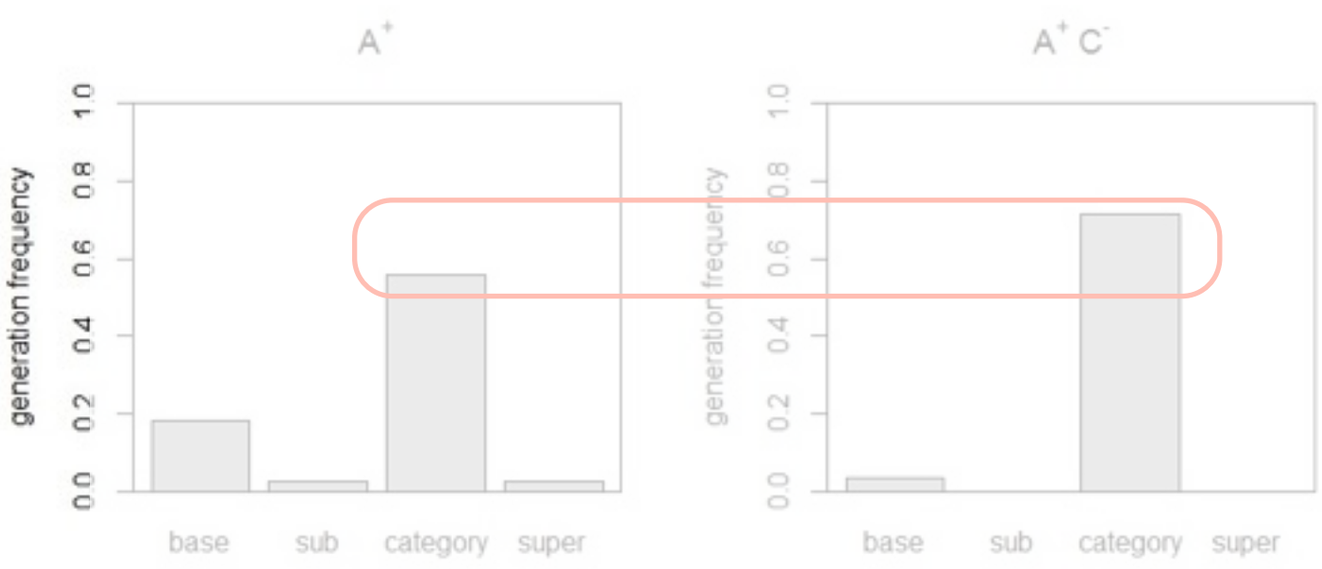
(and yes, the replication worked)





(and yes, the replication worked)

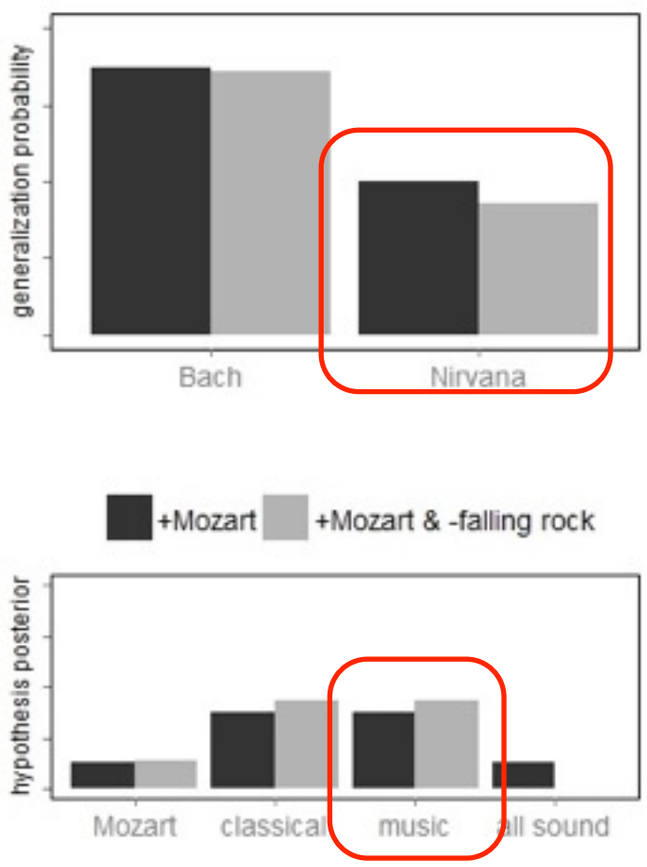




The big question is <u>how</u> to account for the results…

Does the weak sampling model capture the effect?

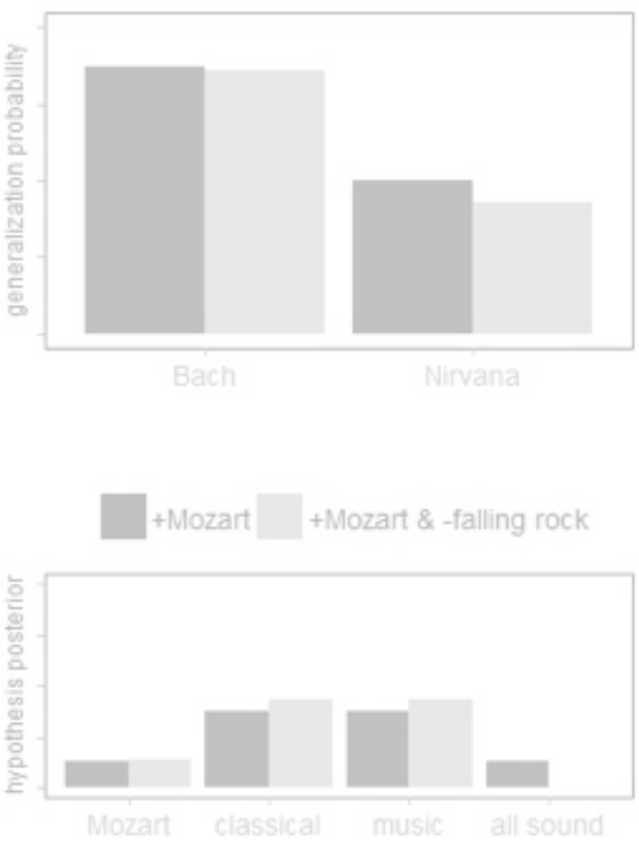
Weak sampling



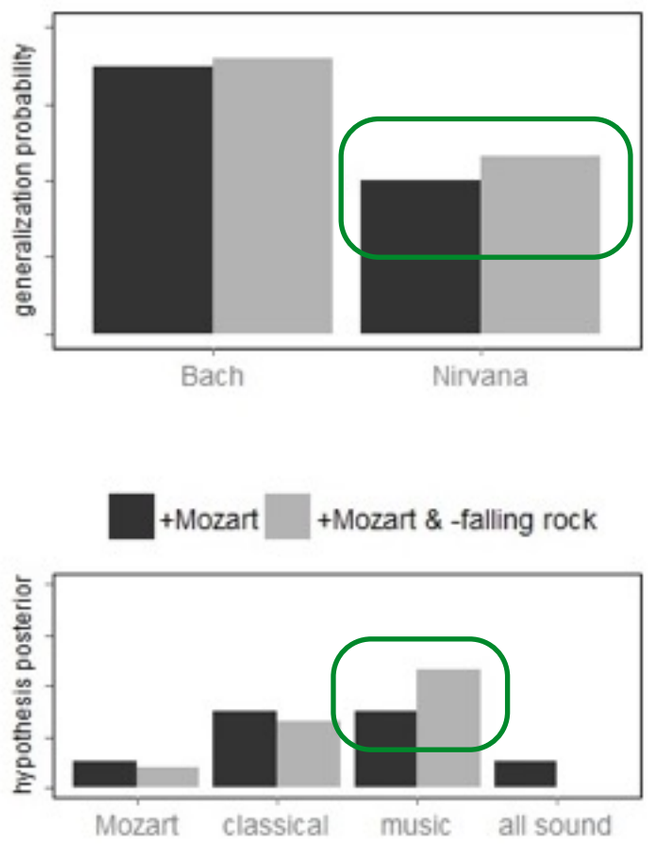
← Nope.

Okay, does the "strong sampling" model capture the effect?

Weak sampling



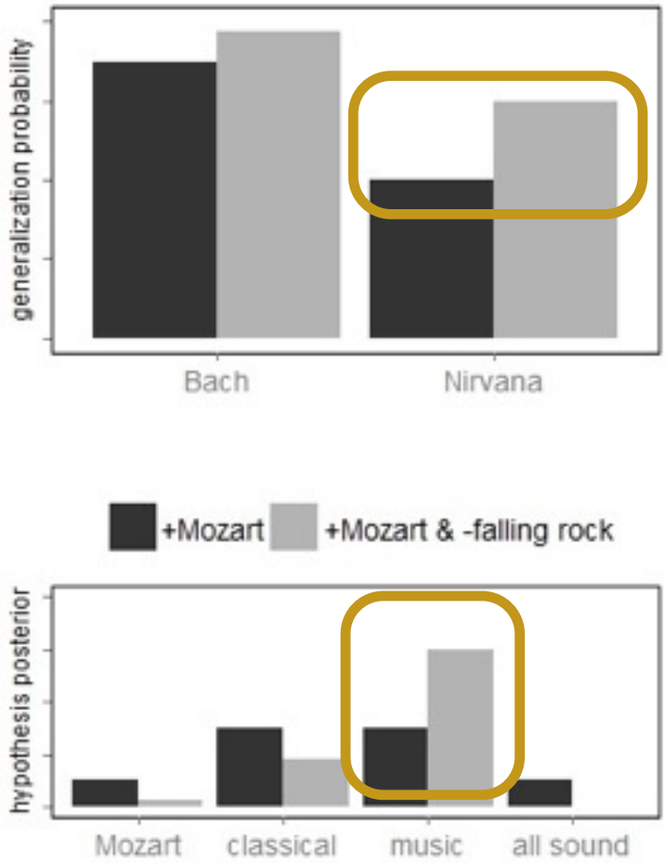
Strong sampling



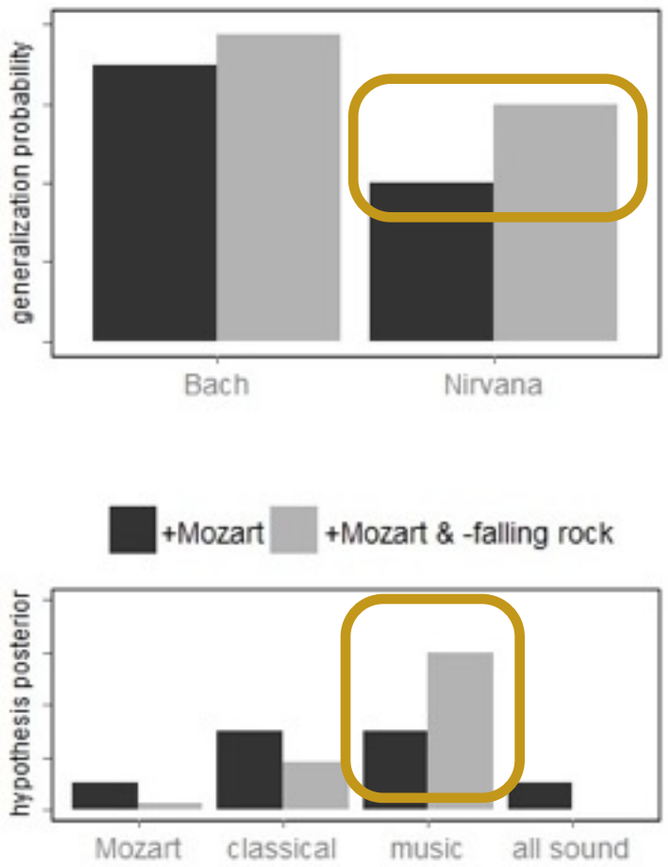
← Meh.

(Out-Bayesing Bayes??!)

Here's a model that gets the effect size right…



But we're going to need a bigger hat.



Weak sampling

An argument consists of random true statements about the world

Weak sampling

An argument consists of random true statements about the world

Strong sampling

An argument consists of randomly selected facts particular to a target category

Weak sampling

An argument consists of random true statements about the world

Strong sampling

An argument consists of randomly selected facts particular to a target category

Pedagogical / persuasive sampling

An argument consists of <u>purposefully</u> chosen facts designed to convince an intelligent reasoner of the truth of some proposition

$$P(x|h) \propto P(h|x)^{\alpha}$$

The data x selected by the <u>communicator</u>…

… is designed to maximise the <u>learner's</u> posterior degree of belief in hypothesis h

If that's right, then the same manipulation we used in the previous study should work…

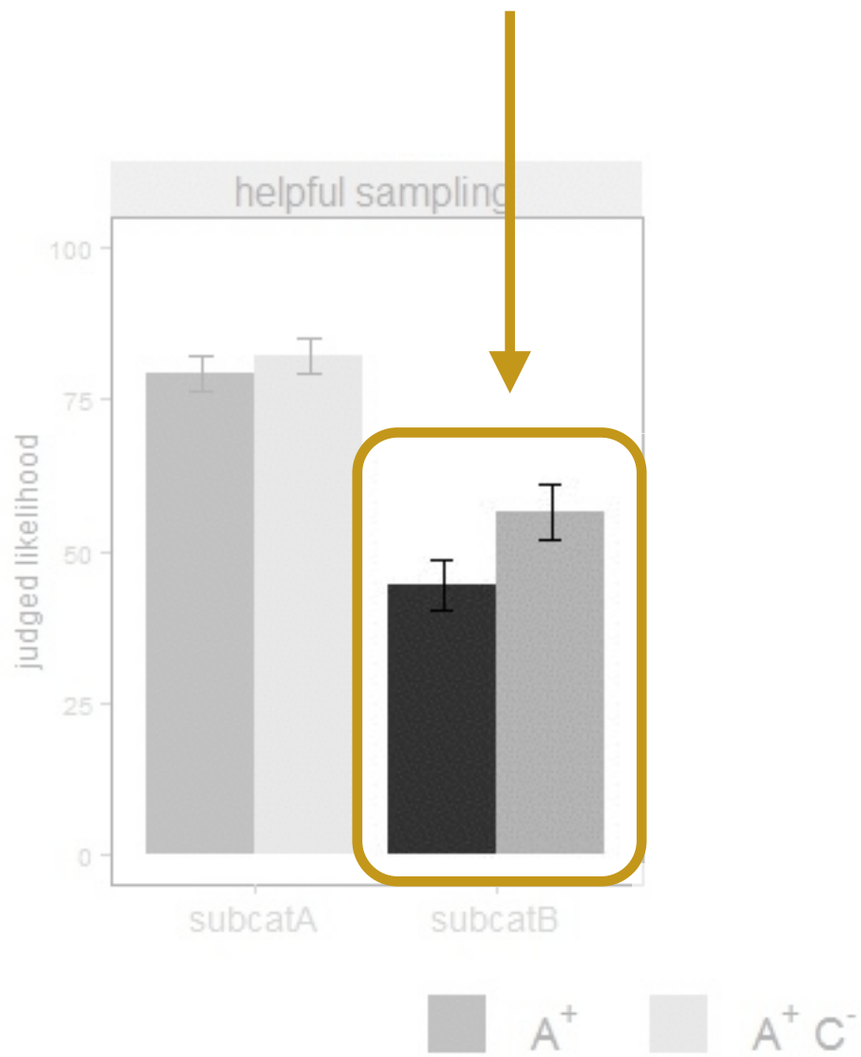
If the negative example is perceived as a "helpful hint" we should continue to get the effect

If it is construed as an arbitrary fact, the effect should vanish

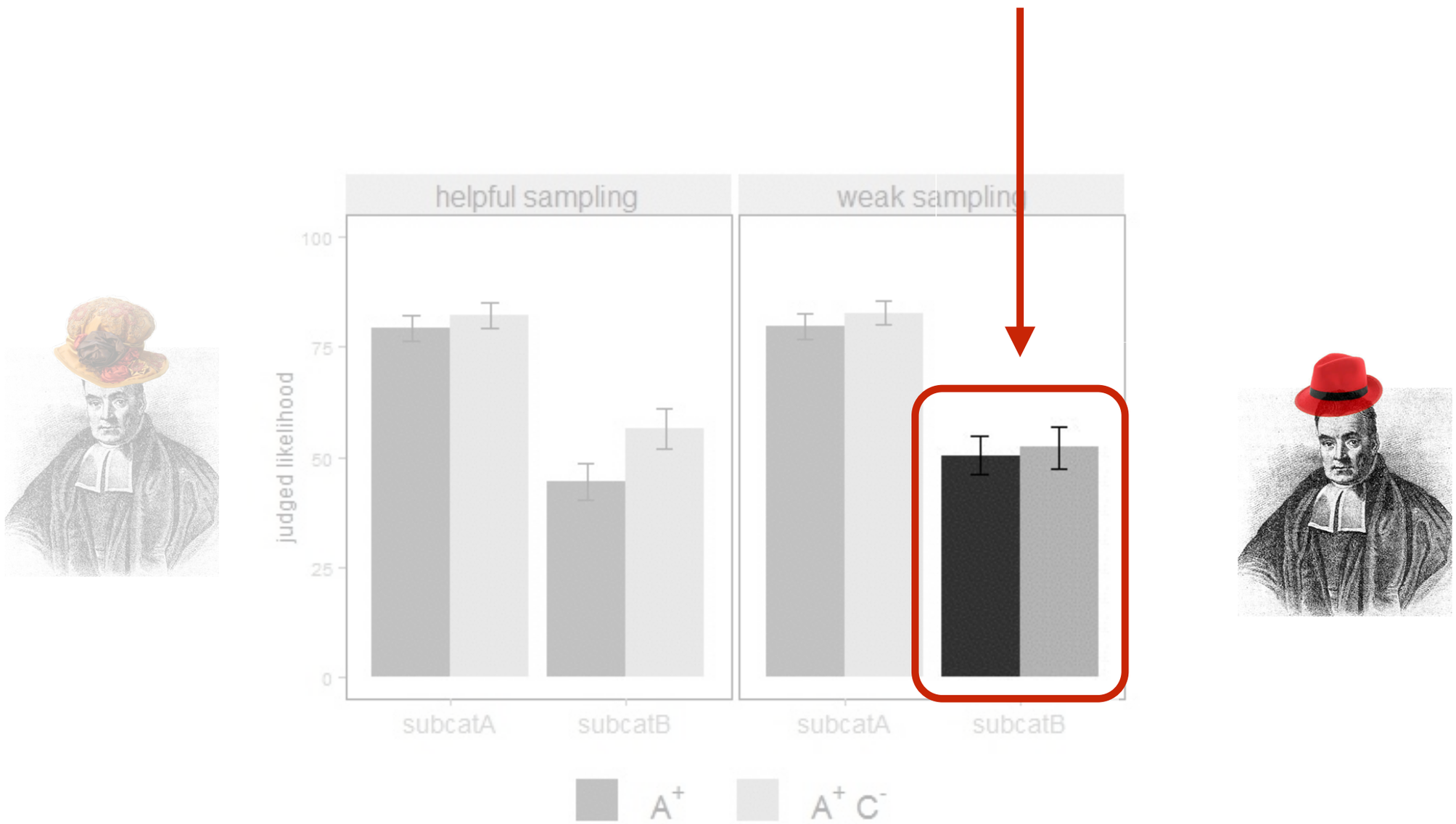
		topics		
	premise 1 (+)	premise 2 (-)	A-member	B-member
MUSIC	Mozart	waterfall	Bach	Nirvana
FRUIT	strawberries	grass blades	blackberry	apple
BIRDS	ducks	elephants	swan	blackbird
TYPES OF WATER	Atlantic ocean	tap water	Mediterranean	Lake Balaton
		fillers weak sampling		
	premise 1	premise 2	conclusion 1	conclusion 2
EXAMPLE	sheep (+)	dogs (-)	horses	chickens
TRIAL 1	aluminium (+)	lead (+)	copper	tin
TRIAL 2	Earth (+)	weather satelite (-)	Uranus	Sun
FILLER	physicists (+)	engineers (+)	mathematicians	carpenters
FILLER	cobras (+)	iguanas (-)	pythons	sea turtles
		fillers helpful sampling		
	premise 1	premise 2	conclusion 1	conclusion 2
EXAMPLE	sheep (+)	cows (+)	horses	pigs
TRIAL 1	aluminium (+)	brass (-)	copper	lead
TRIAL 2	Earth (+)	Mars (+)	Uranus	Sun
FILLER	cobras (+)	pythons (-)	vipers	anacondas
FILLER	physicists (+)	mathematicians (+)	chemists	carpenters

200 participants on MTurk

Negative evidence framed as a “hint” produces the effect



Arbitrary negative evidence does not



- The social aspect to inductive reasoning is central
  - By default, people seem to "read" an inductive argument as if it were put together for a <u>purpose</u>

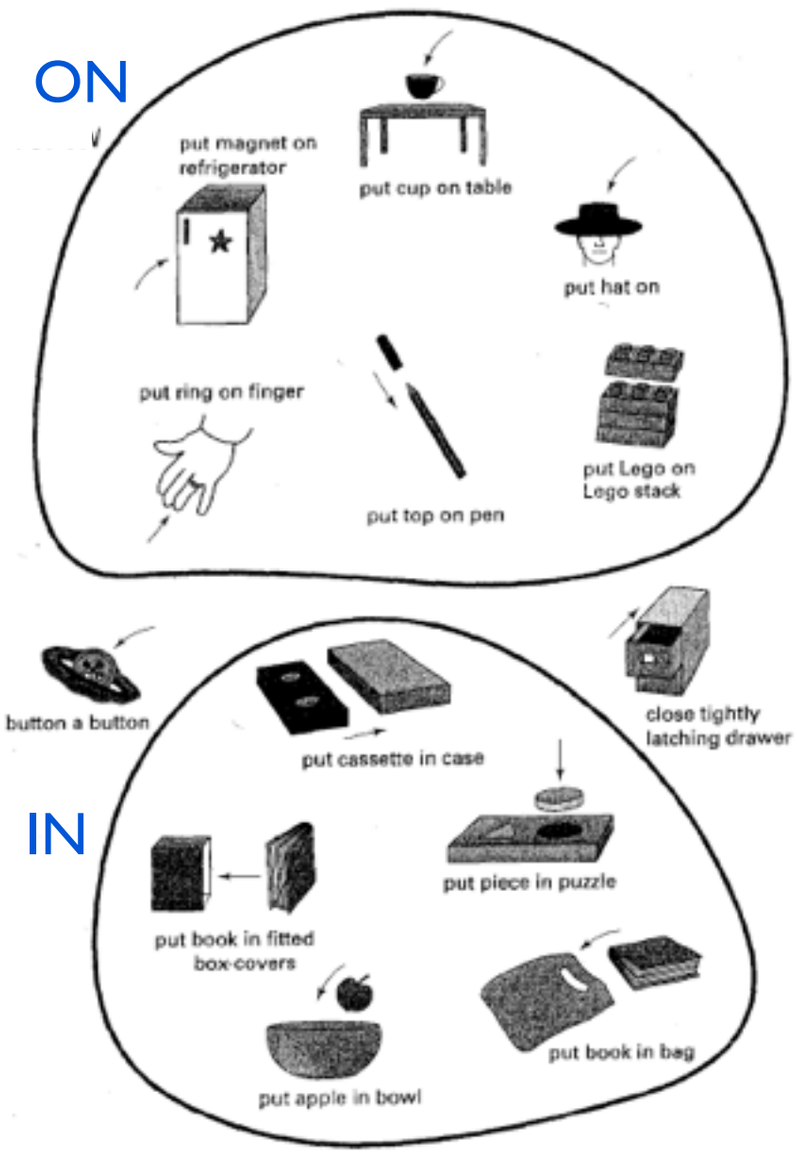
- Pedagogical sampling as normative standard
  - In real life, arguments <u>aren't</u> collections of facts
  - They're acts of <u>persuasion</u>
  - If so, shouldn't "normative" accounts reflect that?

Inductive biases reflect the structure of the mind *and* the environment

Perfors & Navarro (2014). *Cognitive Science*

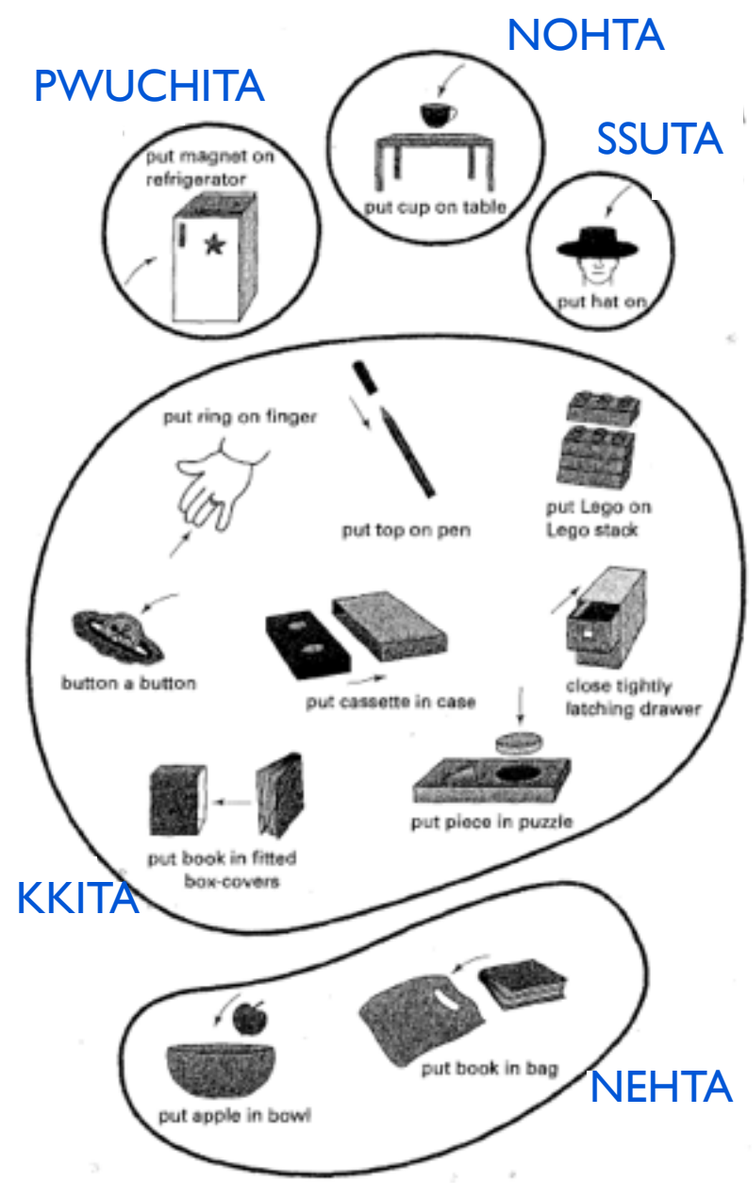
Languages carve the world into categories in different ways

English



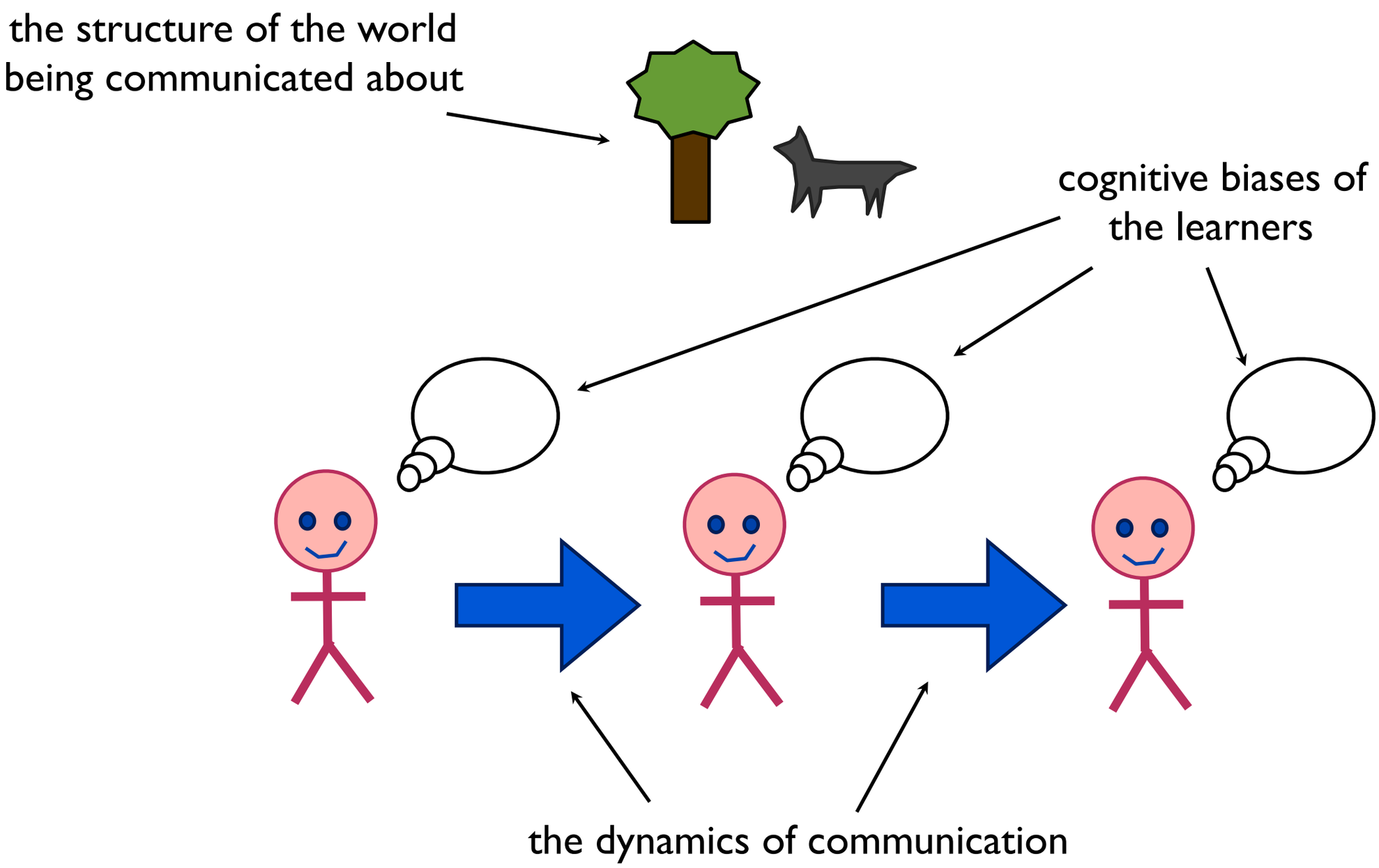
a. English

Korean



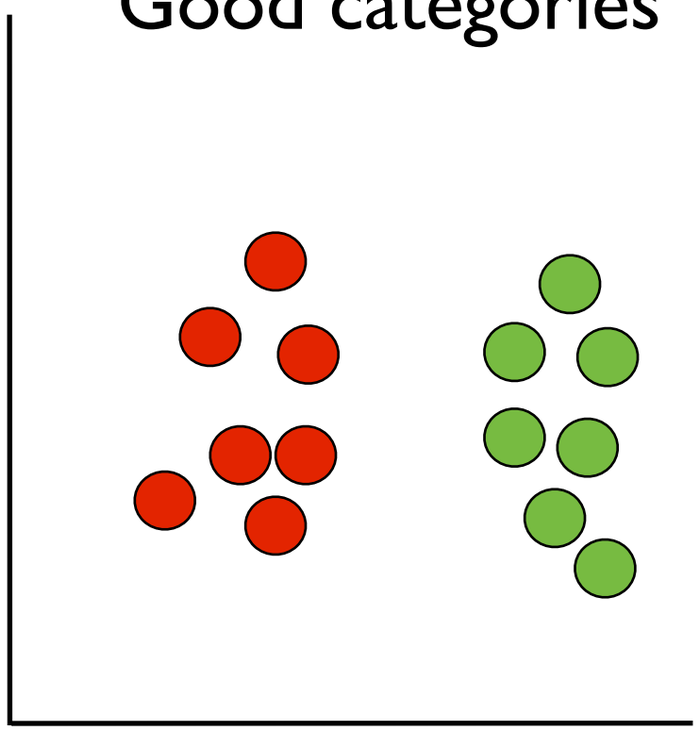
b. Korean

How do naming systems evolve?

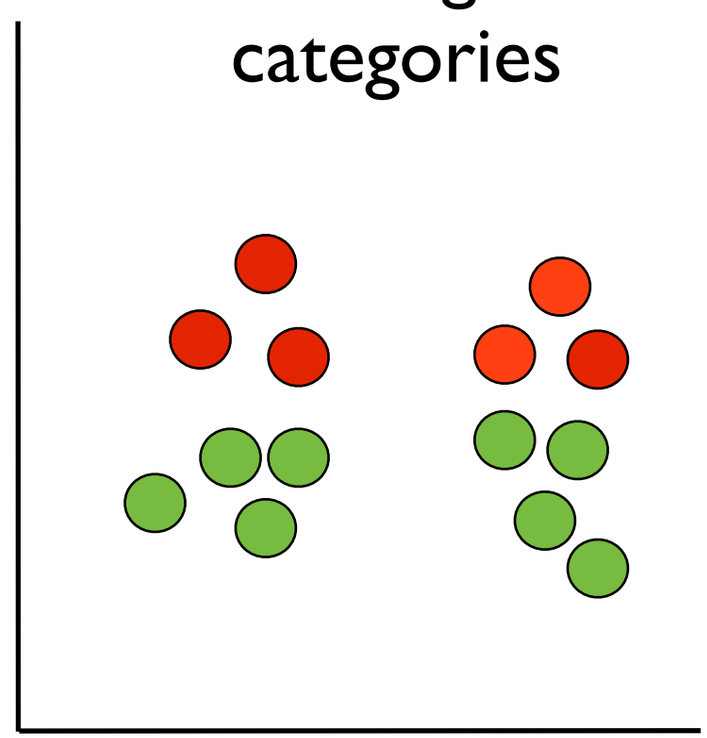


The usual idea in categorisation: stimulus distribution (i.e., <u>the world</u>) shapes inferences

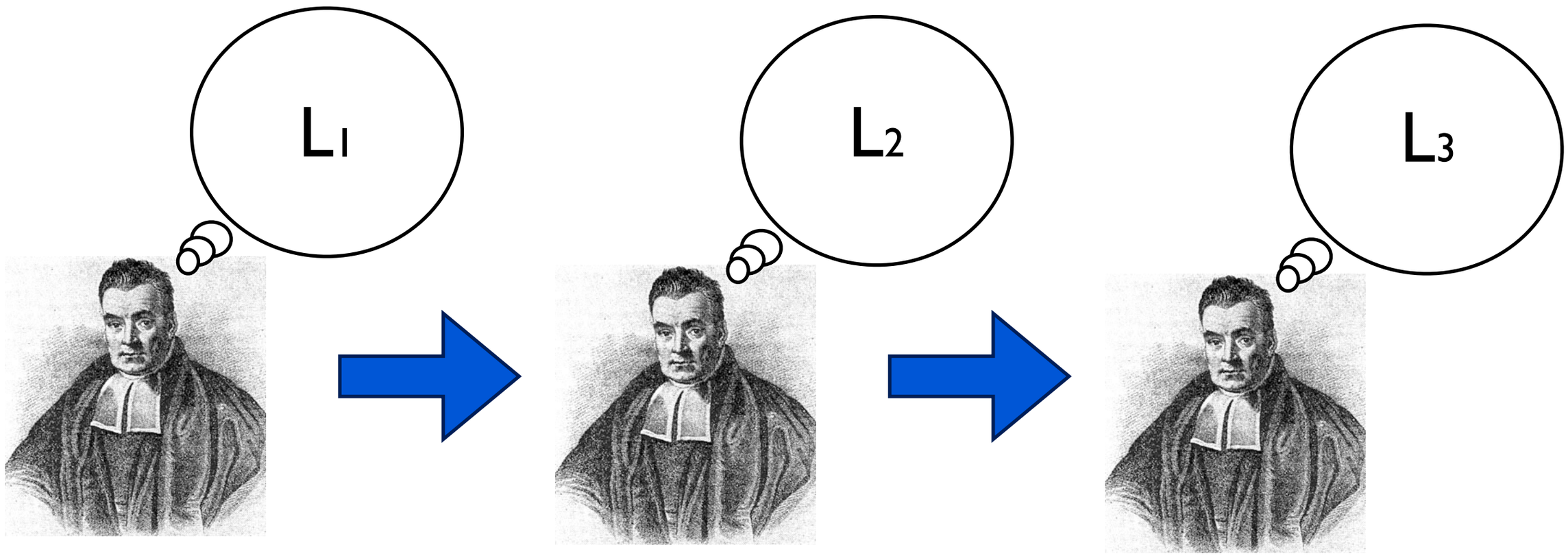
Good categories



Not so good categories



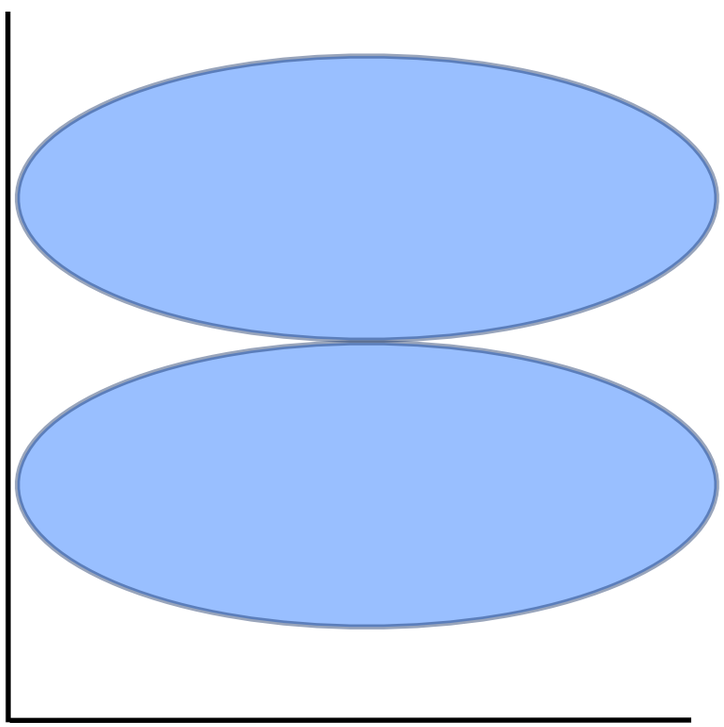
Language evolution as "iterated learning" places the inductive bias within the learner



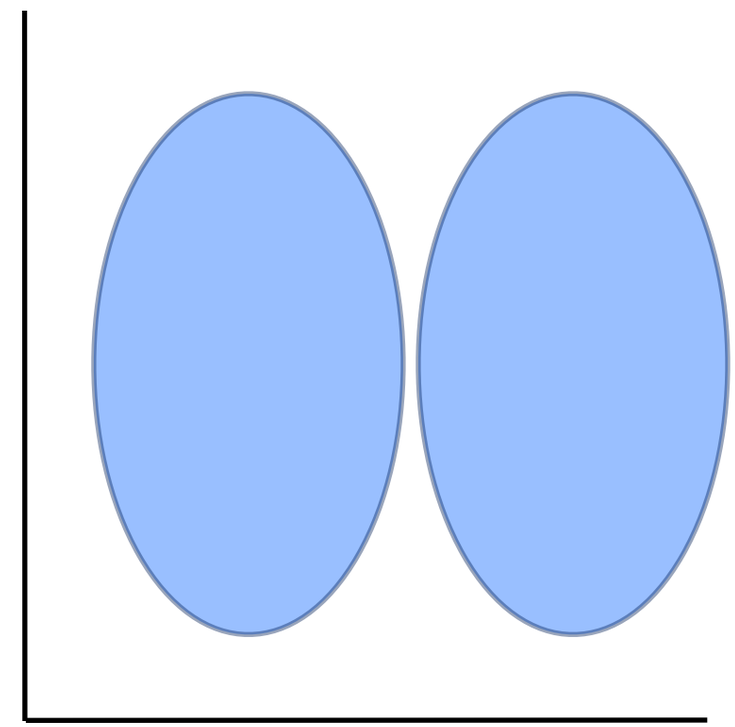
A sequence of Bayesian learners, each learning from the language output of the last one, and then generating the input for the next one... **converges to the learner's <u>prior</u>.**

Griffiths & Kalish (2005, 2007)

This is *weird*. Suppose this is my prior:



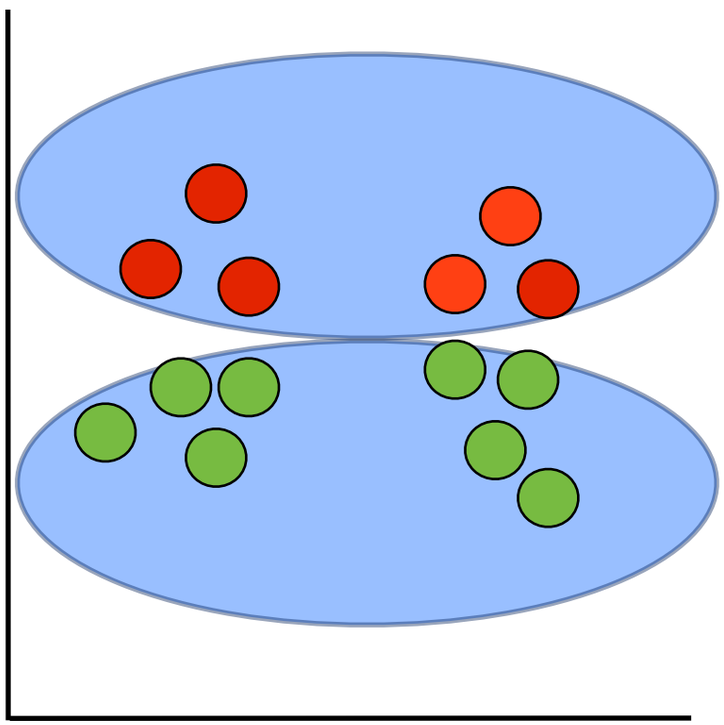
High probability



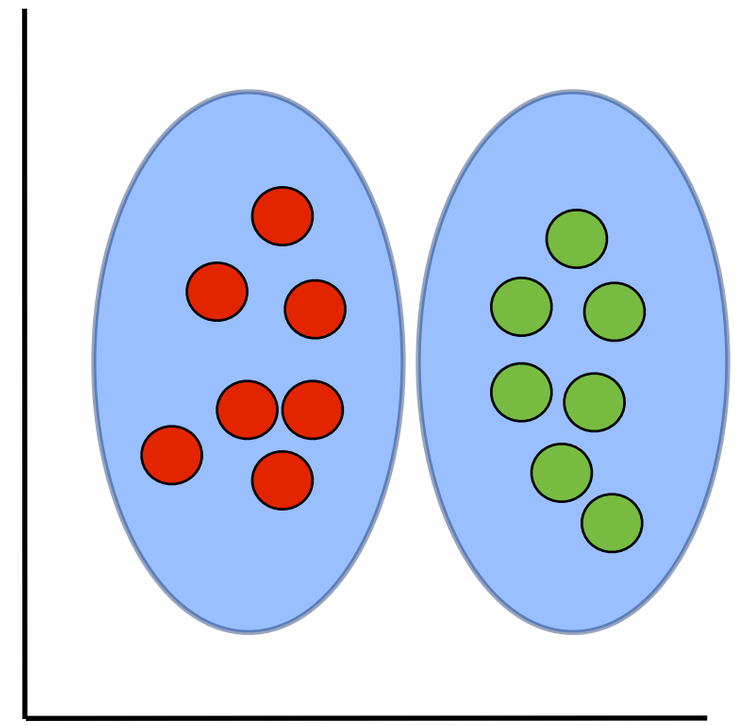
Low probability

And these are entities that need names

Convergence to the prior is *absurd*.



Your language should do this? (high prior)



Your language should <u>not</u> do this? (low prior)

Where did things go astray?
The <u>sampling</u> assumption…

$$\ell = P(y|x)$$

A language provides labels y for entities x

It says <u>nothing</u> about which entities x will be observed

Griffiths & Kalish (2005, 2007)

"laser"

My language has a word for laser… does that really tell me <u>nothing at all</u> about my chances of encountering one?

Where did things go astray?
The <u>sampling</u> assumption…

$$\ell = P(y|x)$$

A language provides labels y for entities x

It says nothing about which entities x will be observed

Griffiths & Kalish (2005, 2007)

$$\ell = P(y, x)$$

A language provides labels y for entities x, but it <u>also</u> makes assumptions about which entities x are likely to appear

Perfors & Navarro (2011, 2014)

Leads to different ideas about learning

$$\ell = P(y|x)$$

A language provides labels y for entities x

It says nothing about which entities x will be observed

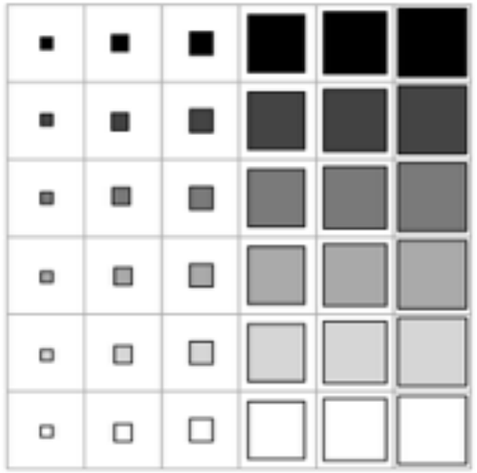
$$\ell = P(y, x)$$

A language provides labels y for entities x, but it <u>also</u> makes assumptions about which entities x are likely to appear

Labels converge to the <u>prior</u> distribution

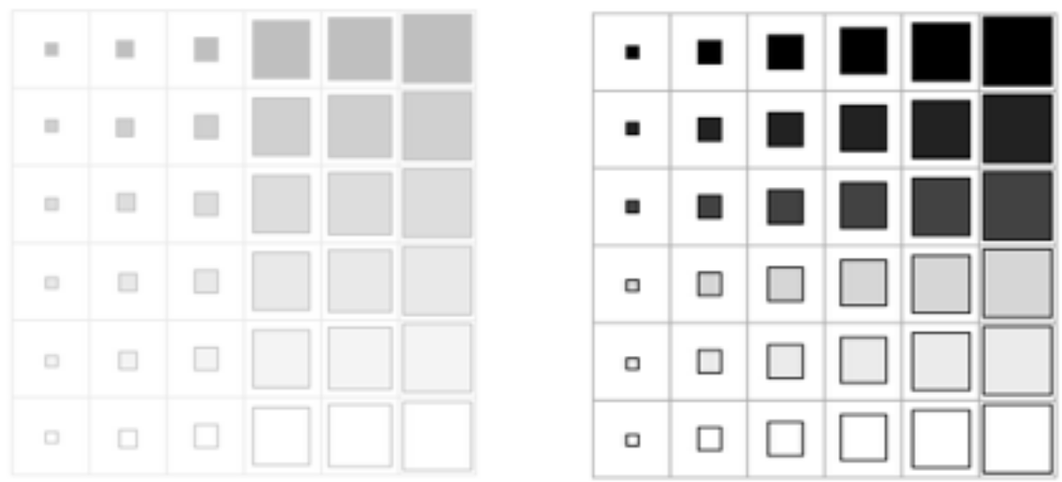
Labels converge to the expected <u>posterior</u> given the entities (sort of)

Three stimulus domains



Attend to size

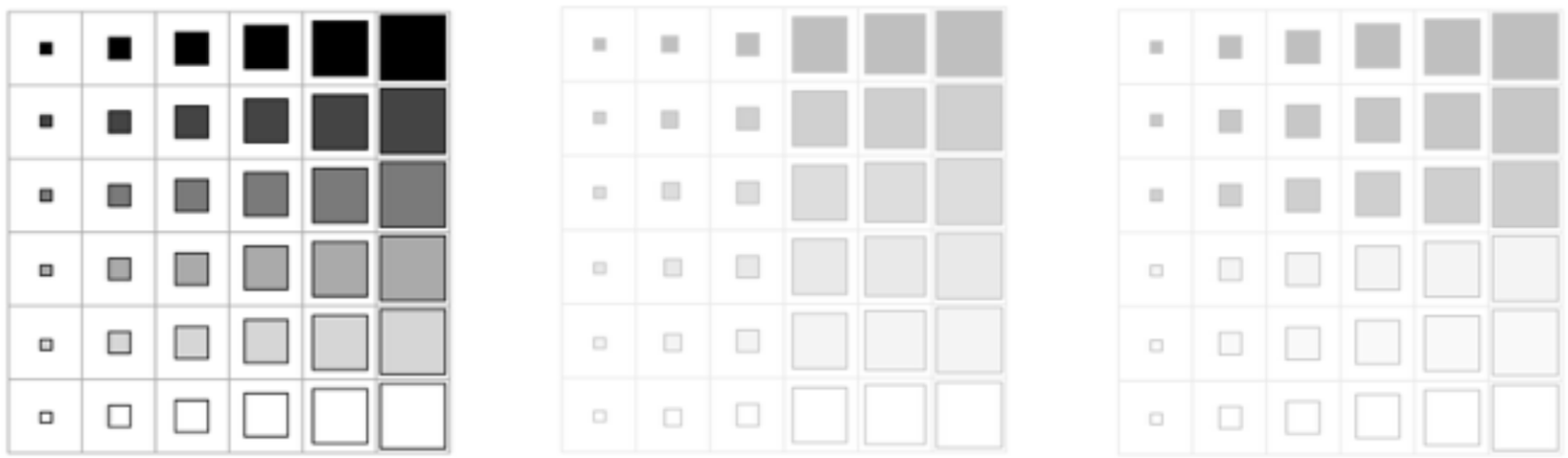
Three stimulus domains



Attend to size

Attend to colour

Three stimulus domains



Control | Attend to size | Attend to colour

An iterated learning design

LUM	YOP	BIR	SUF	GUG	POF
GIC	RIN	NER	LEP	VOD	CIP
ROK	DUT	KUN	TAF	MIG	DIL
BAP	VAM	MAK	PAG	FOD	YEM
TUS	HEC	DOX	CEN	RUP	KOT
FIM	SOR	PIK	BOM	YIB	MEF

Responses from the participant become the target labels for the next participant

Training

see a random subset of those labels (SEEN) twice each

NER

Testing

half SEEN, half UNSEEN, asked to generate label (no feedback)

Training

see a random subset of those labels (SEEN) twice each

NER

Testing

half SEEN, half UNSEEN, asked to generate label (no feedback)

Training

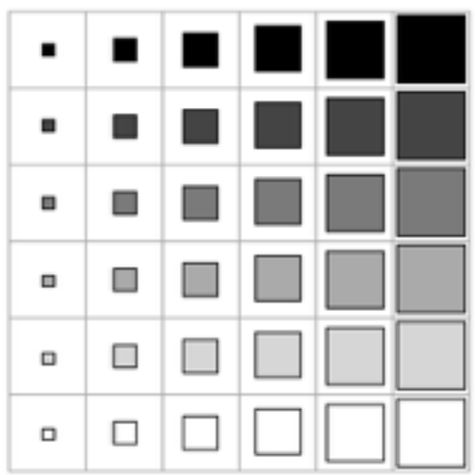
see a random subset of those labels (SEEN) twice each

NER

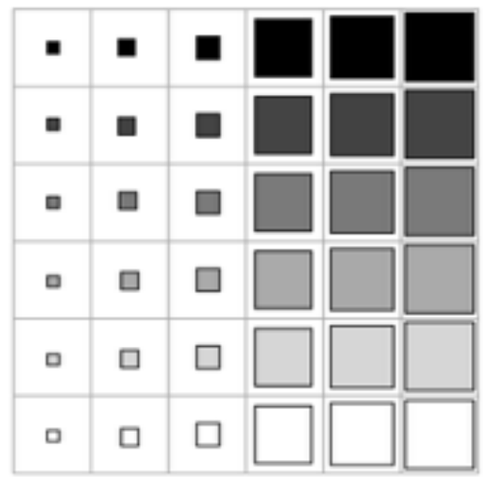
Final Test

all stimuli, asked to generate label (no feedback)

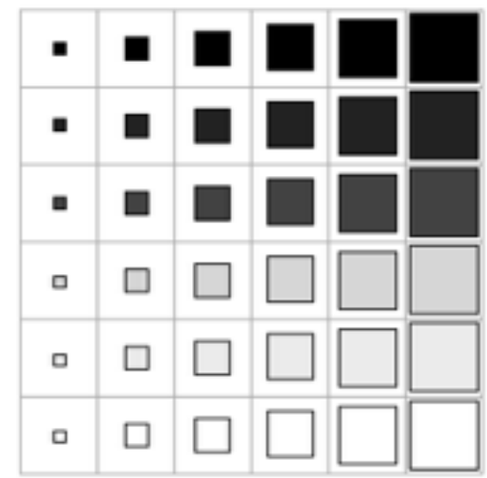
What patterns do we expect?



Control



Size

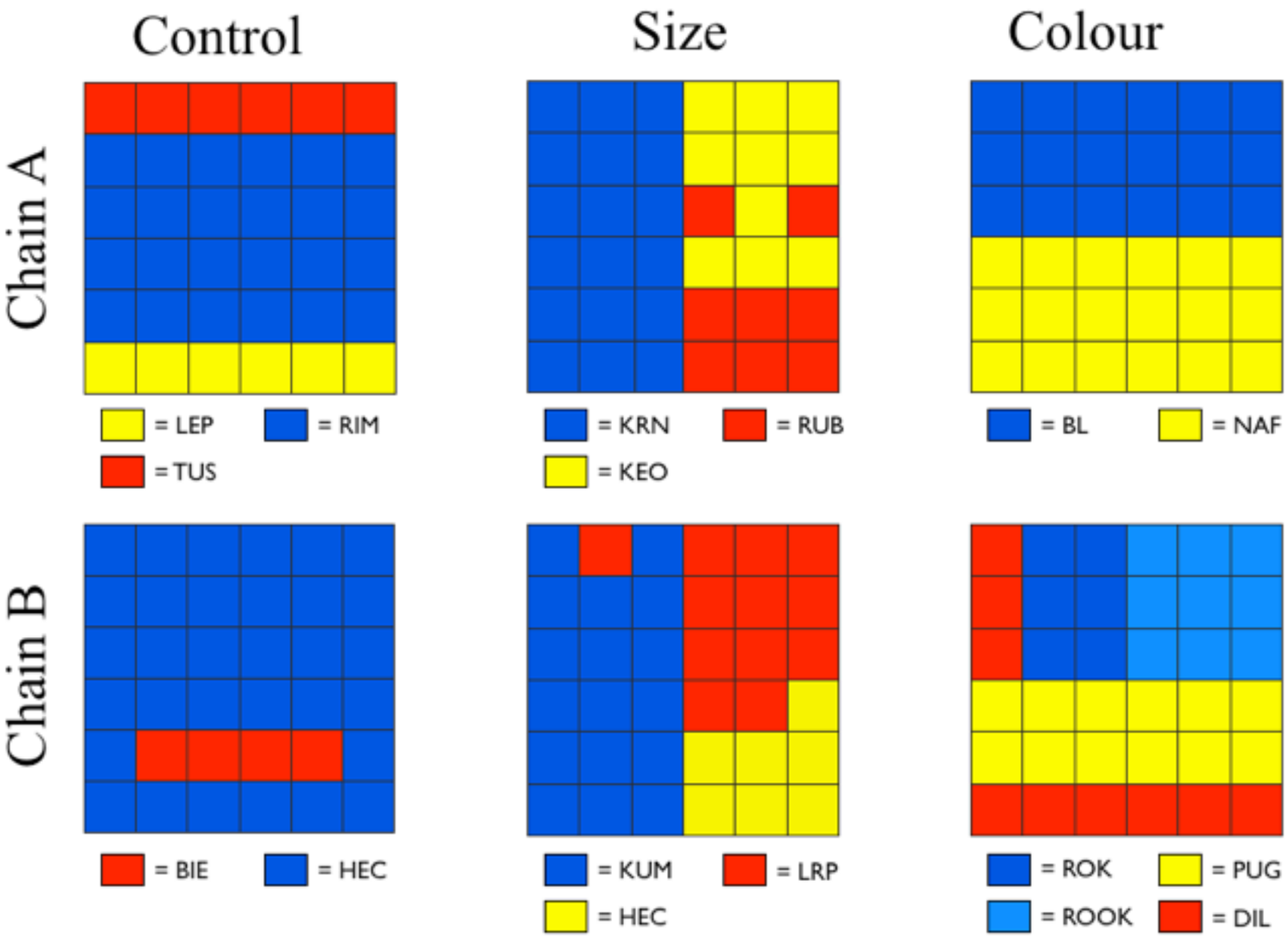


Colour

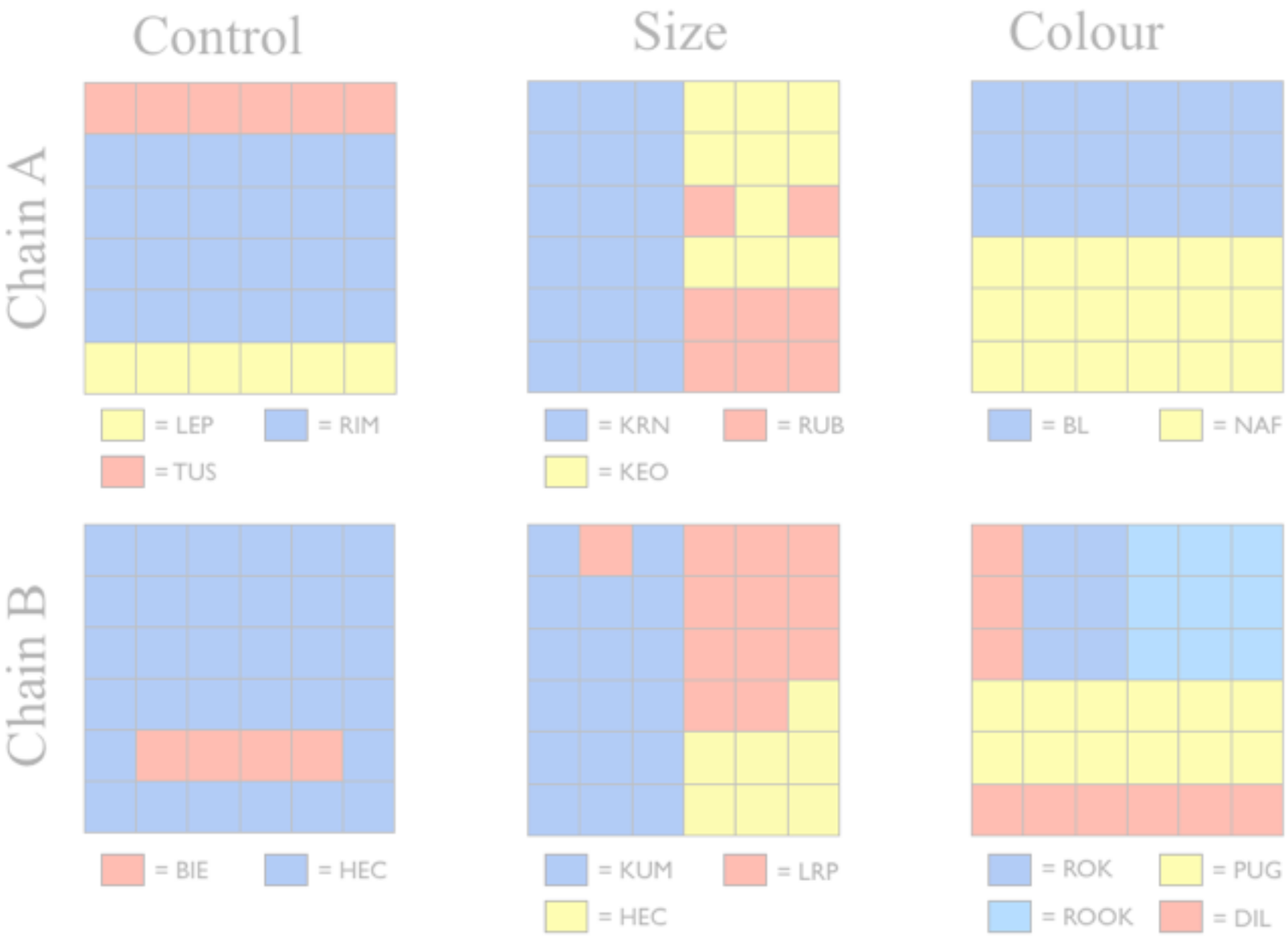
W_1	W_1	W_1	W_2	W_2	W_2
W_1	W_1	W_1	W_2	W_2	W_2
W_1	W_1	W_1	W_2	W_2	W_2
W_1	W_1	W_1	W_2	W_2	W_2
W_1	W_1	W_1	W_2	W_2	W_2
W_1	W_1	W_1	W_2	W_2	W_2

W_1	W_1	W_1	W_1	W_1	W_1
W_1	W_1	W_1	W_1	W_1	W_1
W_1	W_1	W_1	W_1	W_1	W_1
W_2	W_2	W_2	W_2	W_2	W_2
W_2	W_2	W_2	W_2	W_2	W_2
W_2	W_2	W_2	W_2	W_2	W_2

That's exactly what we find



That's exactly what we find



Quantifying the fit:

	Expected Size	Expected Colour
Control	-0.0204	0.0618
Size	0.704	0.079
Colour	0.065	0.696

(*Adjusted Rand index compared to expected patterns for size and colour; 1=identical; 0=random)

Inductive biases are complicated

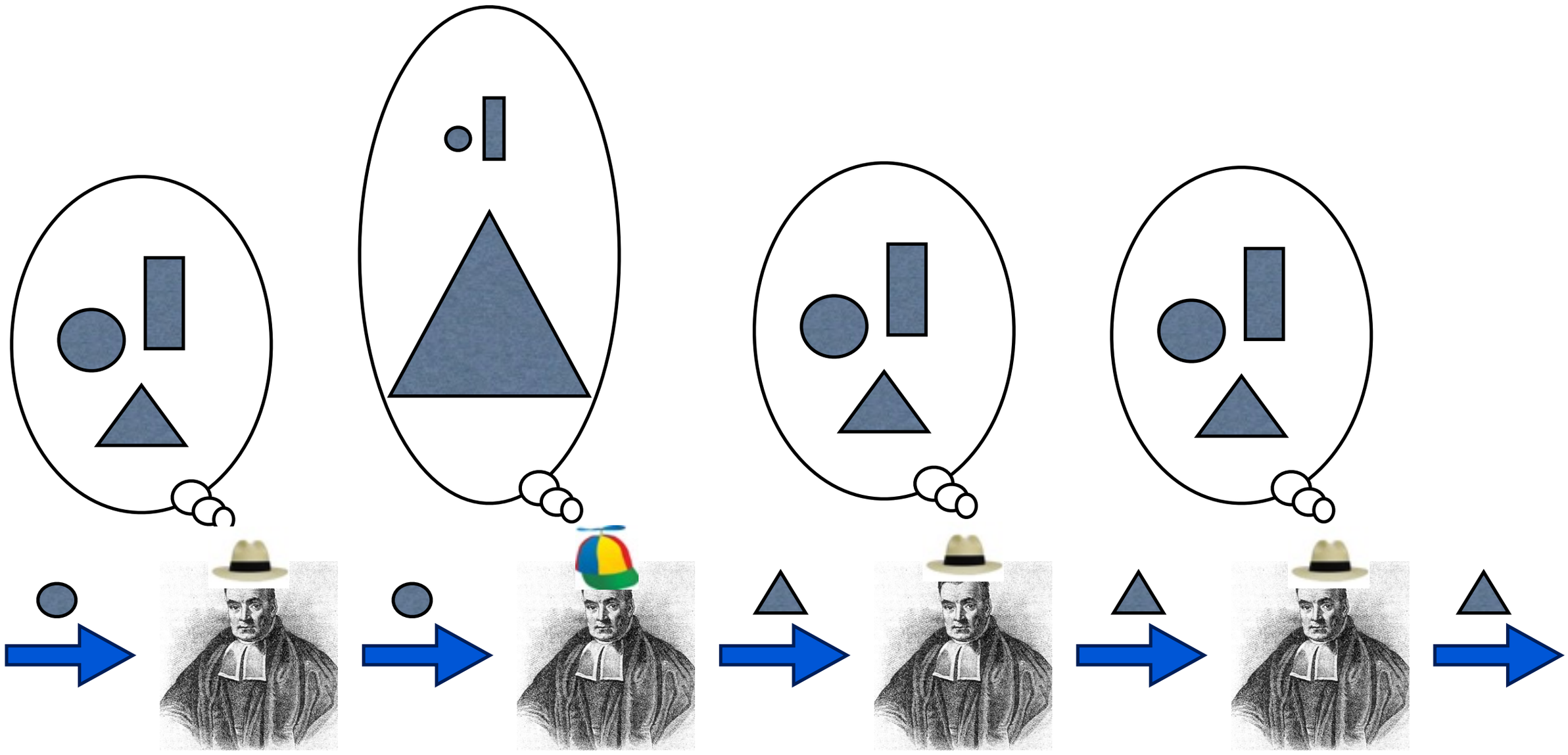
- Convergence to the "*prior*" is deeply suspicious
  - It has been interpreted badly in the literature … taken to mean that the structure of actually existing languages transparently reveal the cognitive biases / universal grammar
  - But this is very wrong

- Convergence to "*something a bit like a posterior conditioned on expected evidence from the environment*" is less sexy
  - But it does have the advantage of being true
  - Well, maybe…

Measuring inductive bias is tricky…

Navarro, Kary, Donkin, Perfors, Brown?? ("in prep")

Iterated learning only works when people share the same biases:



… *and* I can prove it.

The mixed chain misrepresents the mixture of learners

Suppose it were true that "iterated learning reveals the prior" holds even in the presence of mixed learners. Suppose that our two learner types agree on the hypothesis space (i.e., shared values of p and q) above, and disagree only in the prior probability of hypothesis 1, denoted b_1 for learner type 1 and b_2 for learner type 2. If each learner type converges to their own prior (as they do in a homogenous chain), then the observable probability that learner type 1 produces $x = 1$ is given by the prior predictive probability:

$$\begin{aligned} P(x=1|c=1) &= \sum_{h \in \{h_1, h_2\}} P(x=1|h_1)P(h_1|c=1) \\ &= pb_1 + q(1-b_1) \end{aligned}$$

and equivalently

$$P(x=1|c=2) = pb_2 + q(1-b_2)$$

Accordingly, since the proportion of learners of type 1 in the mixed chain is w the marginal probability of a success should be the mixture of prior predictives:

$$\begin{aligned} P(x=1) &= wP(x=1|c=1) + (1-w)P(x=1|c=2) \\ &= w(pb_1 + q(1-b_1)) + (1-w)(pb_2 + q(1-b_2)) \end{aligned}$$

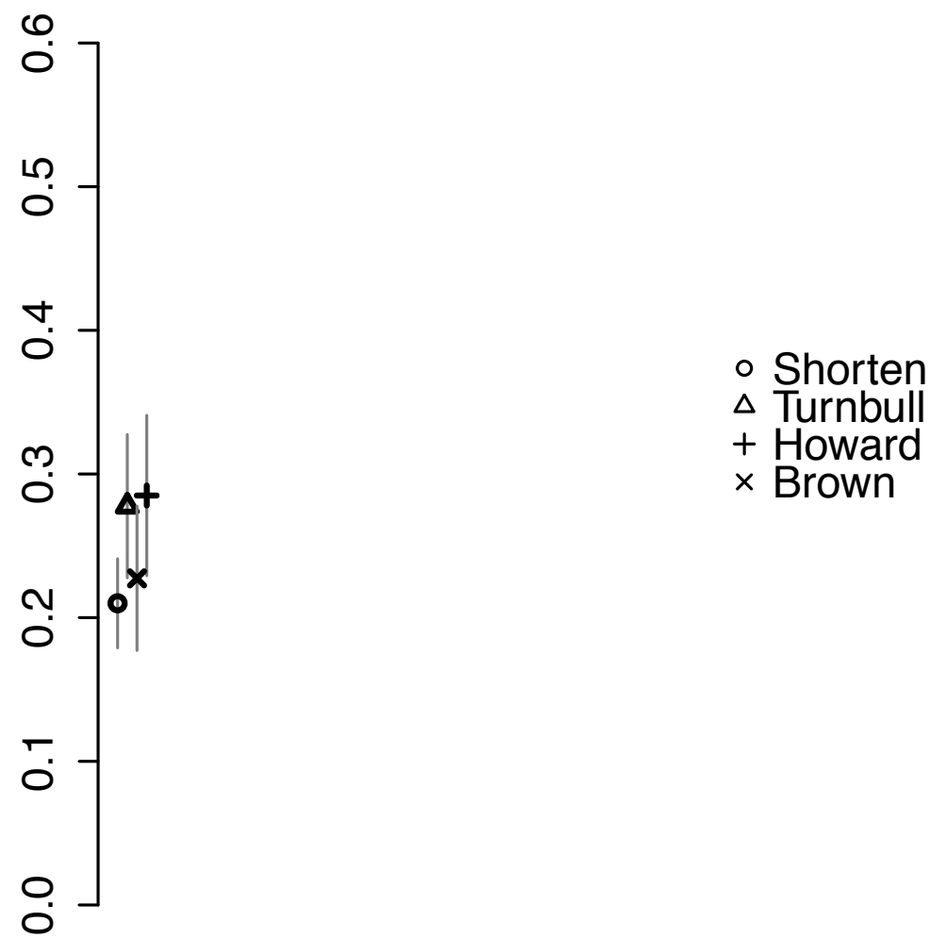
However, substitution into the expressions above reveals that this is far from the case:

$$\begin{aligned} P(x=1) &= u/(u+d) = (1+d/u)^{-1} \\ &= \left(1 + \frac{P(x_i=0|x_{i-1}=1)}{P(x_i=1|x_{i-1}=0)}\right)^{-1} \\ &= \left(1 + \frac{wP(x_i=0|x_{i-1}=1, c_i=1) + (1-w)P(x_i=0|x_{i-1}=1, c_i=2)}{wP(x_i=1|x_{i-1}=0, c_i=1) + (1-w)P(x_i=1|x_{i-1}=0, c_i=2)}\right)^{-1} \\ &= \left(1 + \frac{w\frac{p(1-p)b_1+q(1-q)(1-b_1)}{pb_1+q(1-b_1)} + (1-w)\frac{p(1-p)b_2+q(1-q)(1-b_2)}{pb_1+q(1-b_2)}}{w\frac{p(1-p)b_1+q(1-q)(1-b_1)}{(1-p)b_1+(1-q)(1-b_1)} + (1-w)\frac{p(1-p)b_1+q(1-q)(1-b_2)}{(1-p)b_2+(1-q)(1-b_2)}}\right)^{-1} \end{aligned}$$

Suffice it to say, these are not at all the same except in special cases.

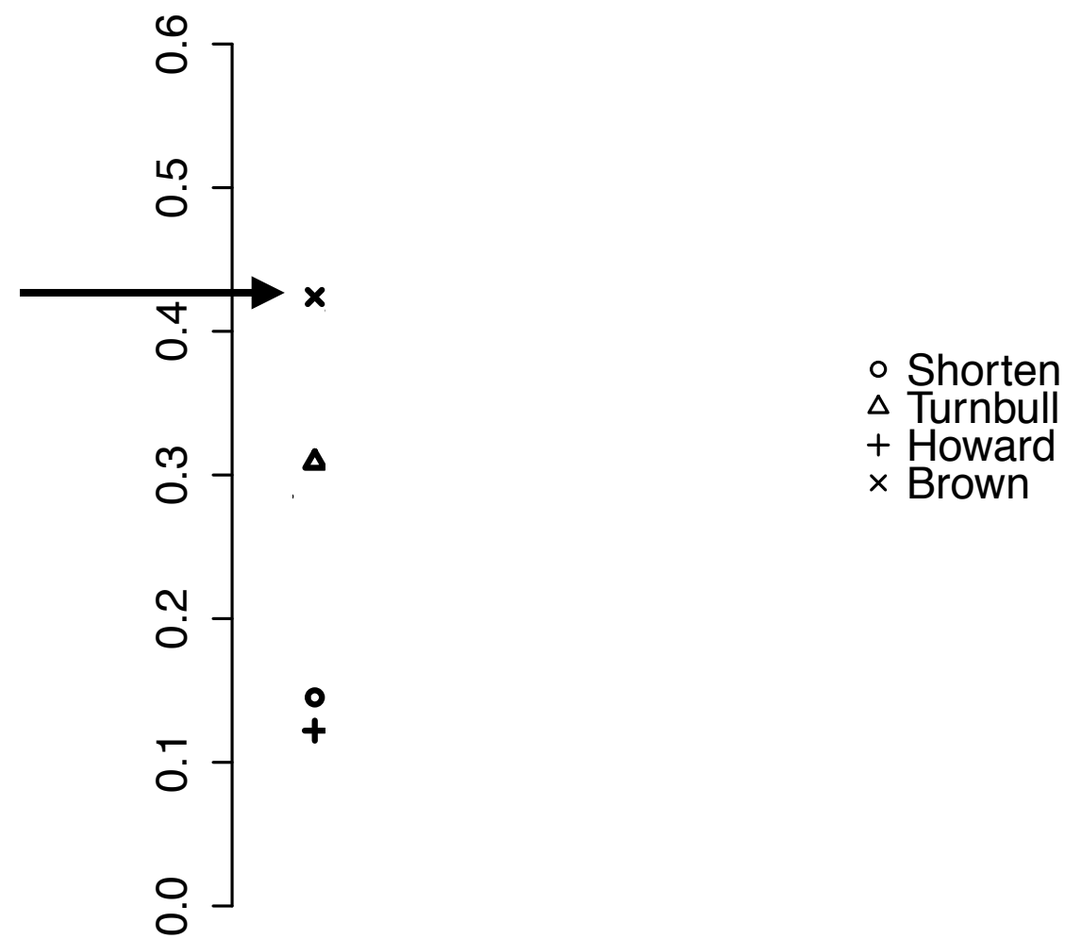
But it's more fun to run dumb experiments that give stupid answers…

Americans *claim* to be totally ignorant about the Oz election…



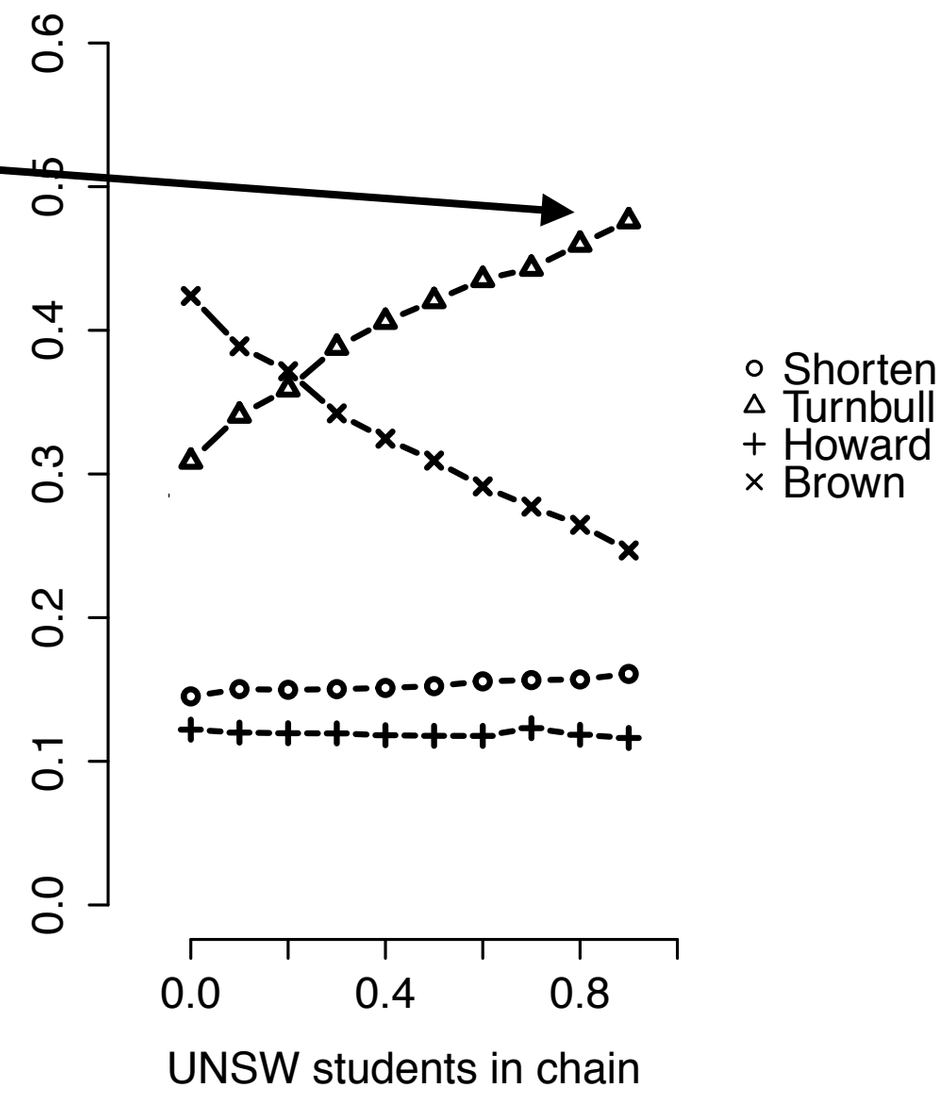
But it's more fun to run dumb experiments that give stupid answers…

… and a "pure" iterated learning chain "reveals" a preference for *Gordon Brown* …



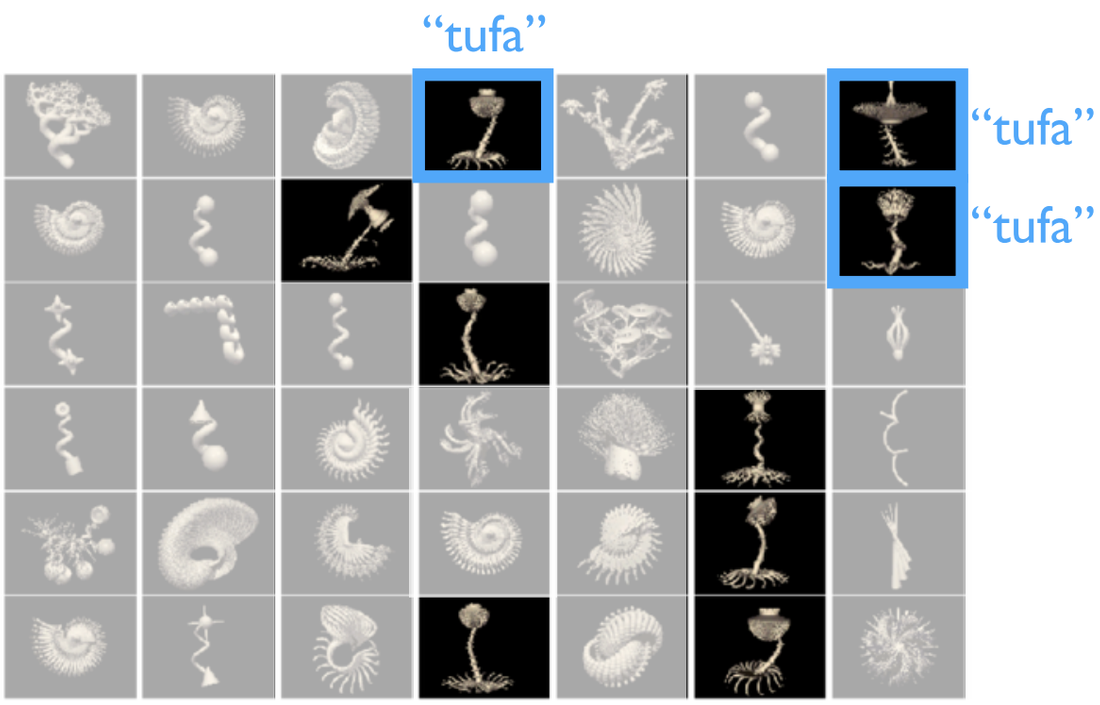
But it's more fun to run dumb experiments that give stupid answers…

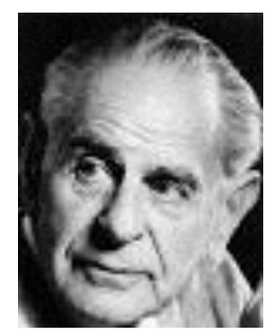
But if we mix some UNSW students into the chain the Americans now appear to back *Malcolm Turnbull*



A few final thoughts about human induction and Bayesian reasoning

Traditional accounts of learning and inference specify norms that implicitly rely on something like falsificationist reasoning

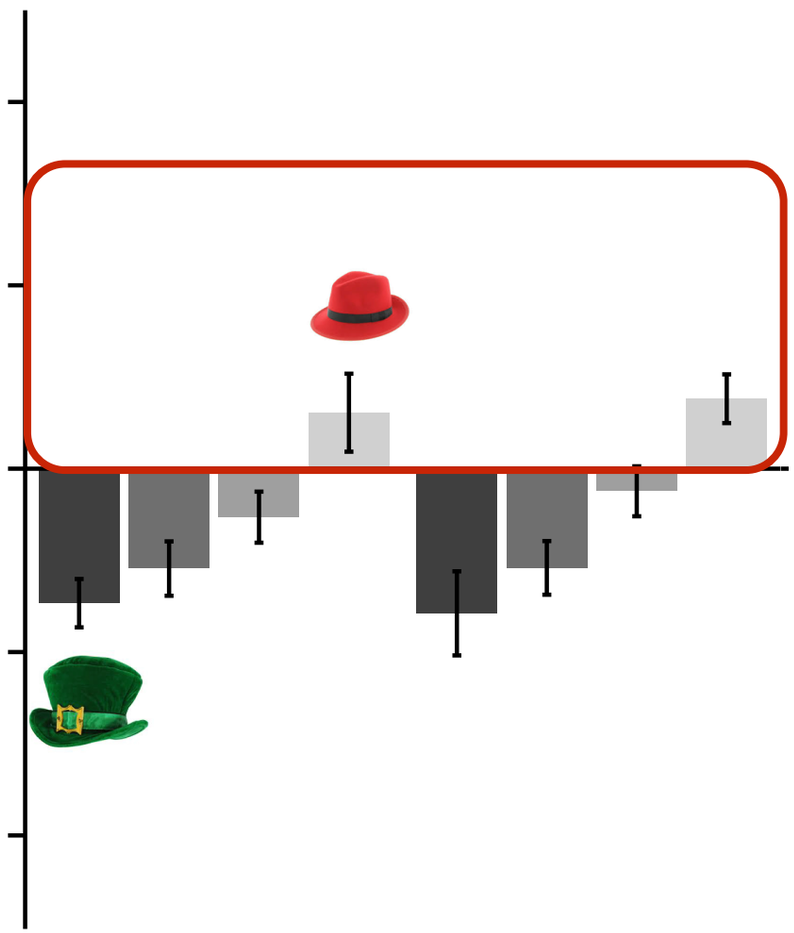




falsificationist learning

weak sampling

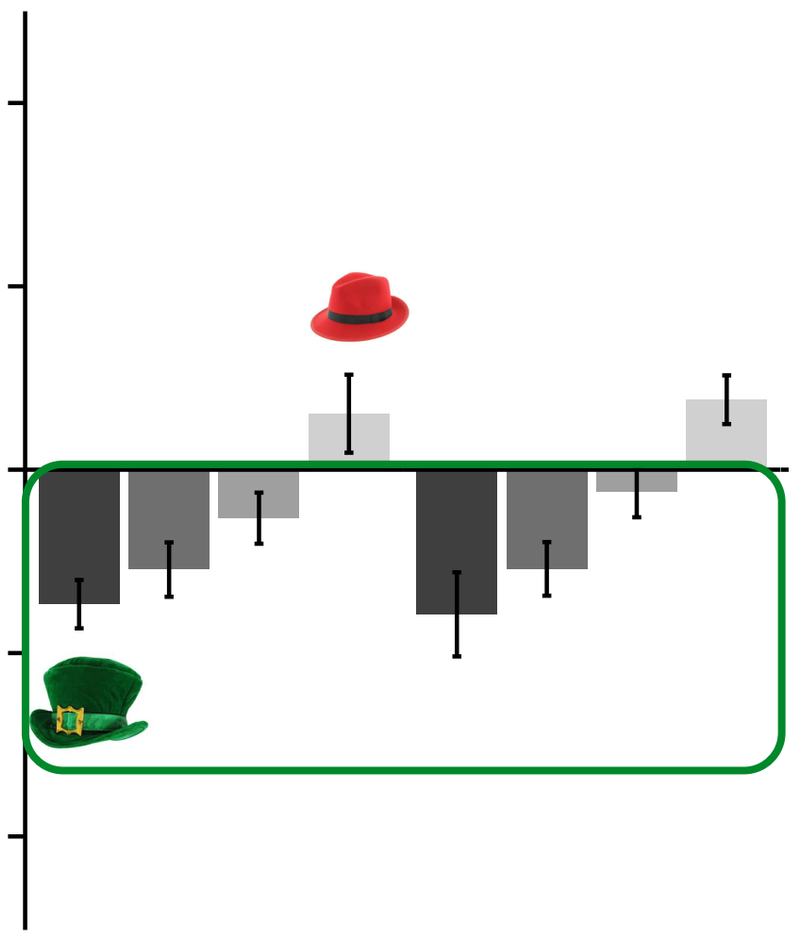
But why?



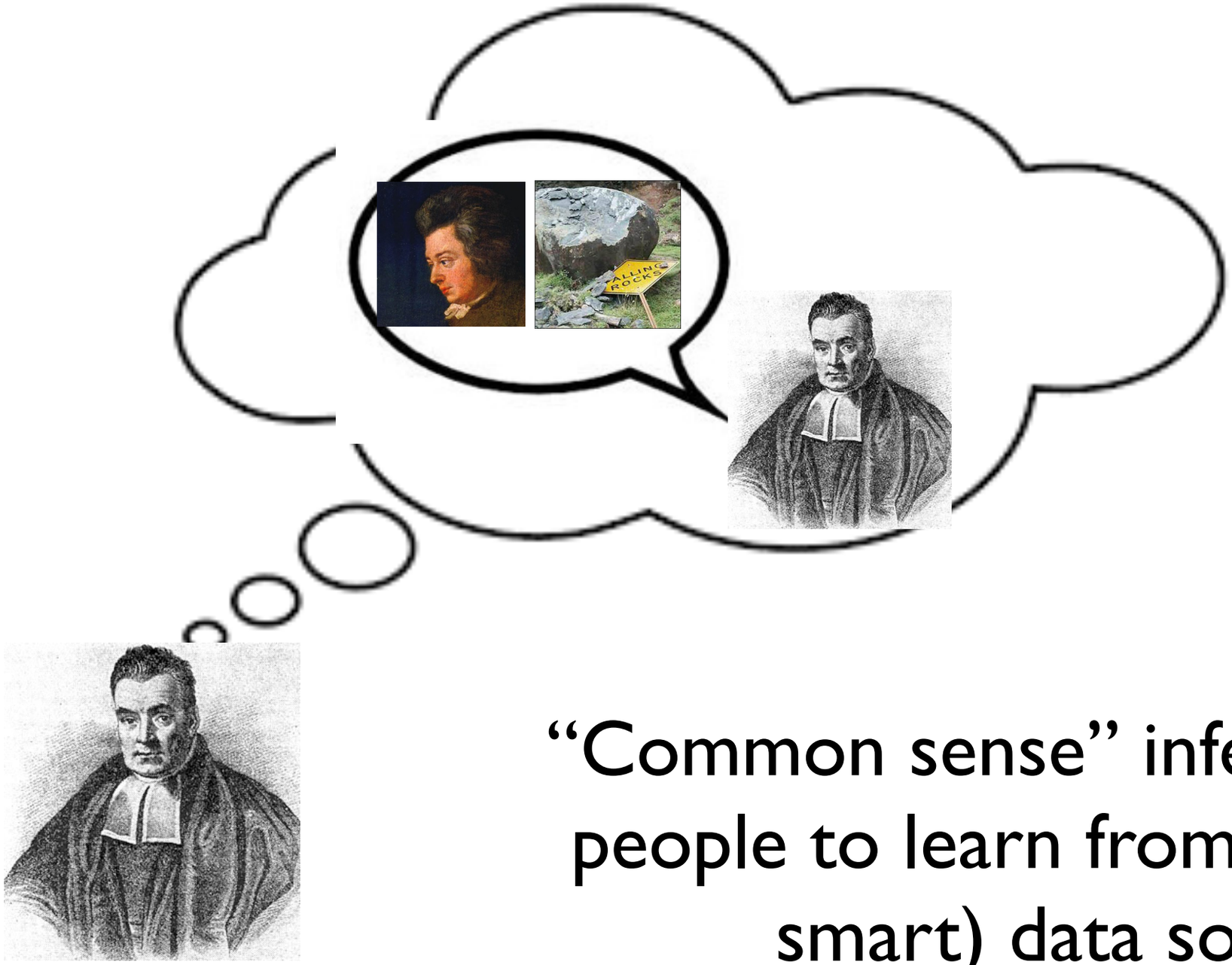
… it **only** makes sense when evidence is selected in an arbitrary and random fashion

Both Random

In real life, isn't
ANYTHING ELSE a more
reasonable theory for the
origin of the data????



Both Relevant
Relevant Fillers
Random Fillers



"Common sense" inference requires people to learn from complex (and smart) data sources…

social agenda

full distribution

quoted distribution

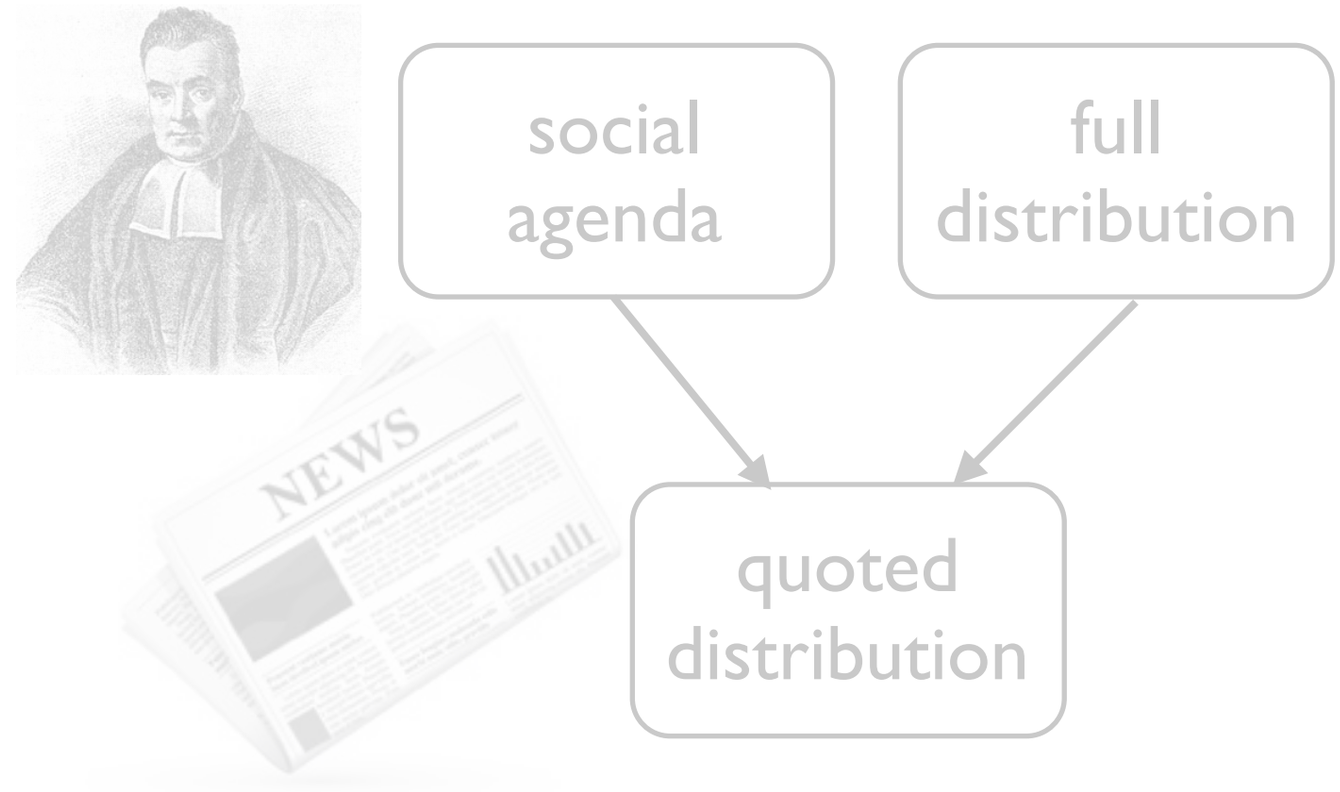
We need to disentangle facts from agendas

We need to detect trickery

social agenda

full distribution

quoted distribution



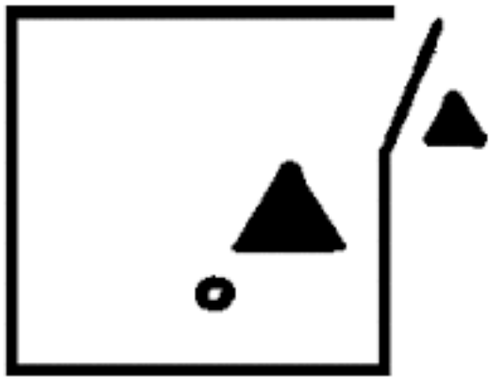
We need to know when to reject the rules/concepts we're given

social agenda

full distribution

quoted distribution

We need to read the intention of other agents



Understanding human common sense reasoning requires something a lot richer

Thanks!

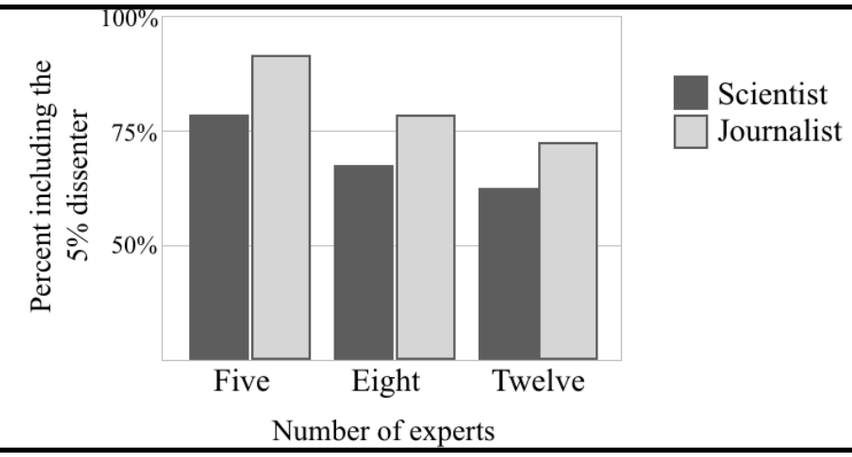




93 95 97 92 5

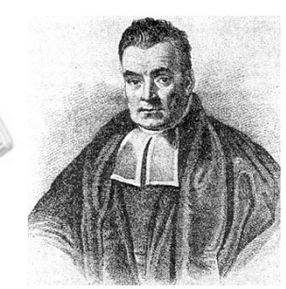
"Expert opinions" about the likelihood that climate change is happening

If an article can only fit three quotes, most people choose to include the dissenter →




A Bayesian journalist who cares about their reputation has a strong motivation to pursue "he says she says" journalism

Because a Bayesian reader can't tell the difference between journalistic bias and expert consensus



Constrained host

Prize location, l	Host choice, o A	B	C
A	0	0.5	0.5
B	0	0	1.0
C	0	1.0	0

Random host

Prize location, l	Host choice, o A	B	C
A	0.33	0.33	0.33
B	0.33	0.33	0.33
C	0.33	0.33	0.33

(a) Uninformative host

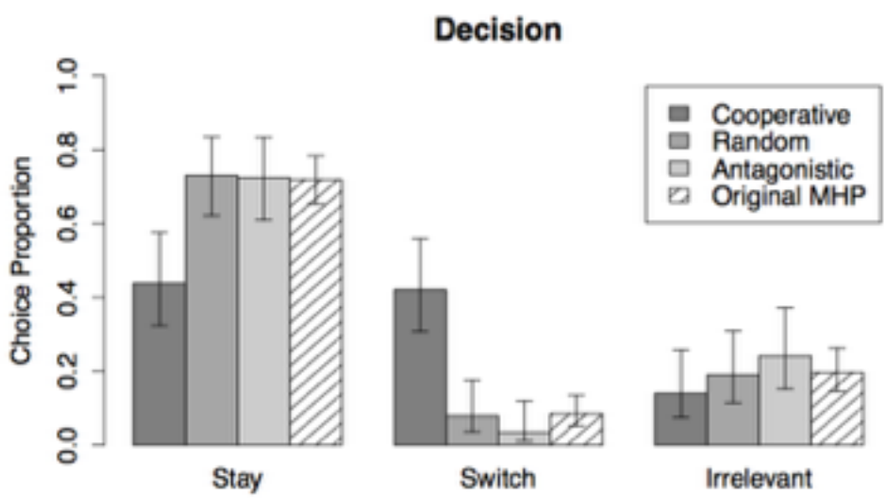
Prize location, l	Host choice, o A	B	C	$\emptyset$
A	0	0.5	0.5	0
B	0	0	0.5	0.5
C	0	0.5	0	0.5

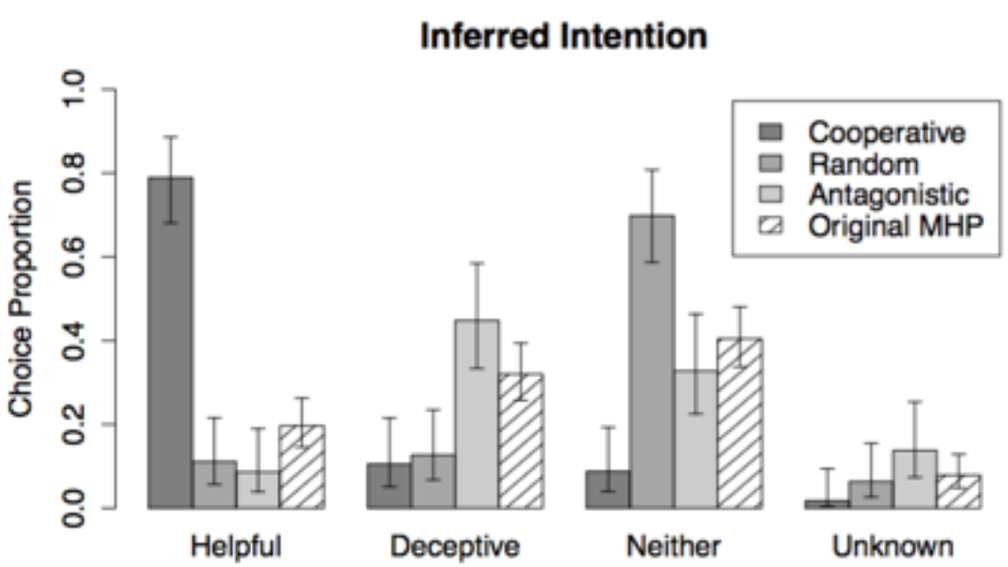
(b) Malicious host

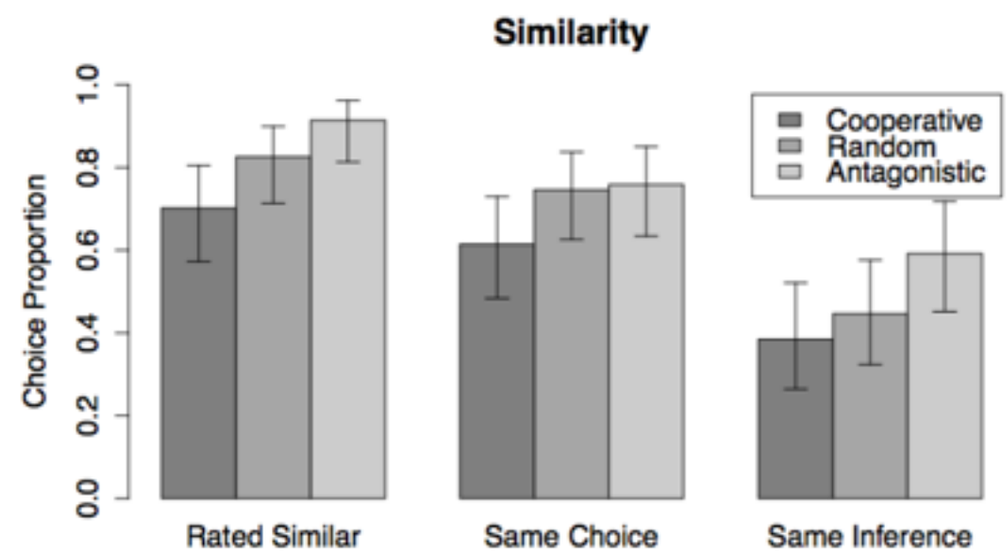
Prize location, l	Host choice, o A	B	C	$\emptyset$
A	0	0.5	0.5	0
B	0	0	0	1
C	0	0	0	1

(c) Helpful host

Prize location, l	Host choice, o A	B	C	$\emptyset$
A	0	0	0	1
B	0	0	1	0
C	0	1	0	0



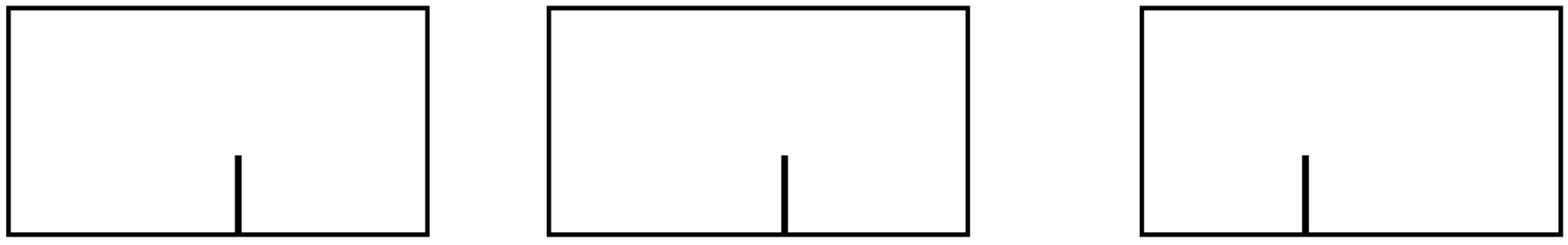




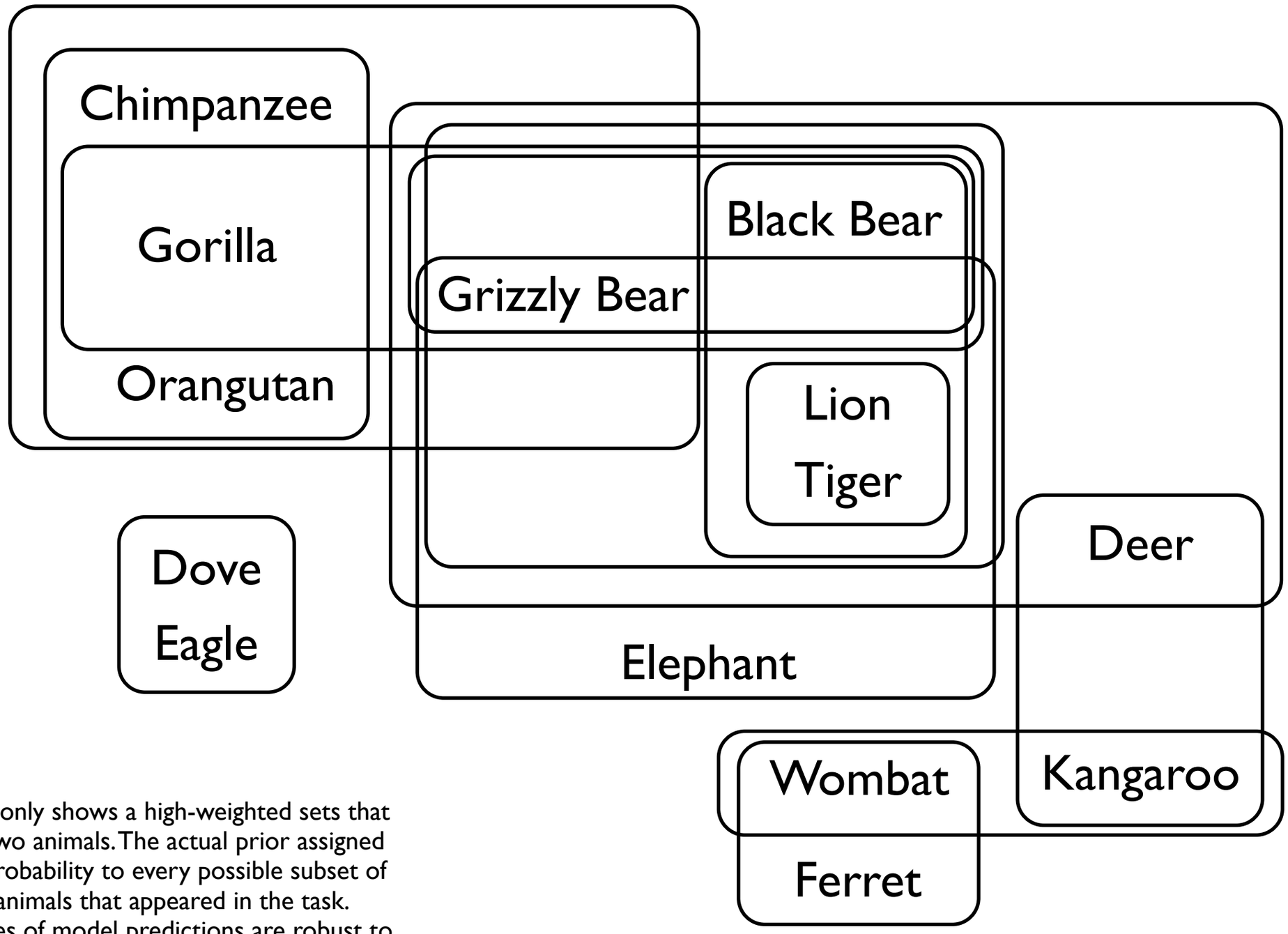
Experimental tests

- Navarro, Dry & Lee (2012):
  - Two experiments, stimuli varied on one dimension
  - N=22 & N=20 undergraduates
  - Non traditional stimulus presentation
  - Response measure: Probability judgments
- Vong, Hendrickson, Perfors & Navarro (2013)
  - As above, but with N=318 workers on AMT
- Hendrickson, Perfors & Navarro (in preparation)
  - One experiment (N=470) on AMT
  - Participants shown traditional categorisation stimuli (below)
  - Response measures: probability judgment & categorisation decisions

stimuli:

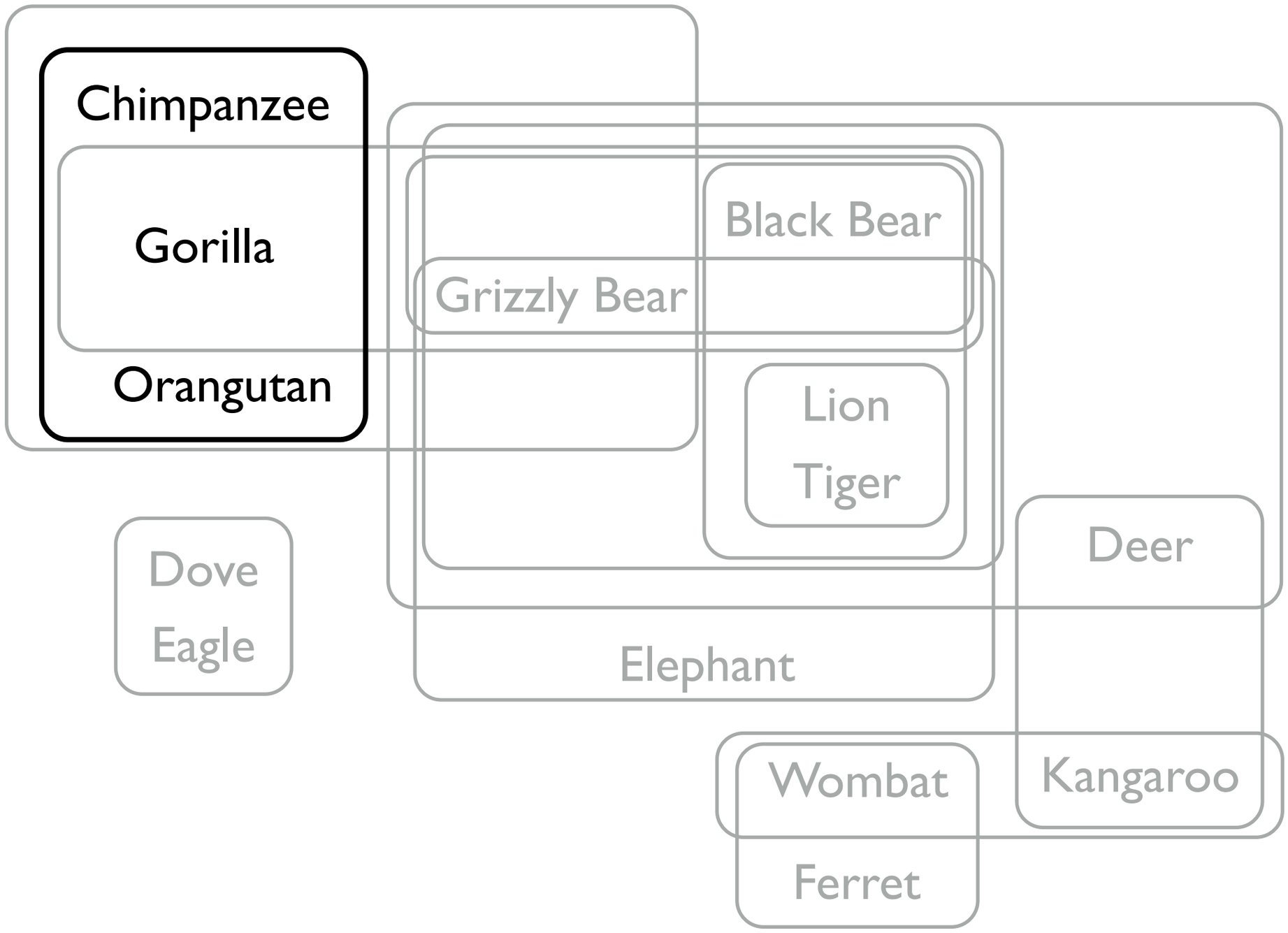


Hypotheses inferred from a separate data set

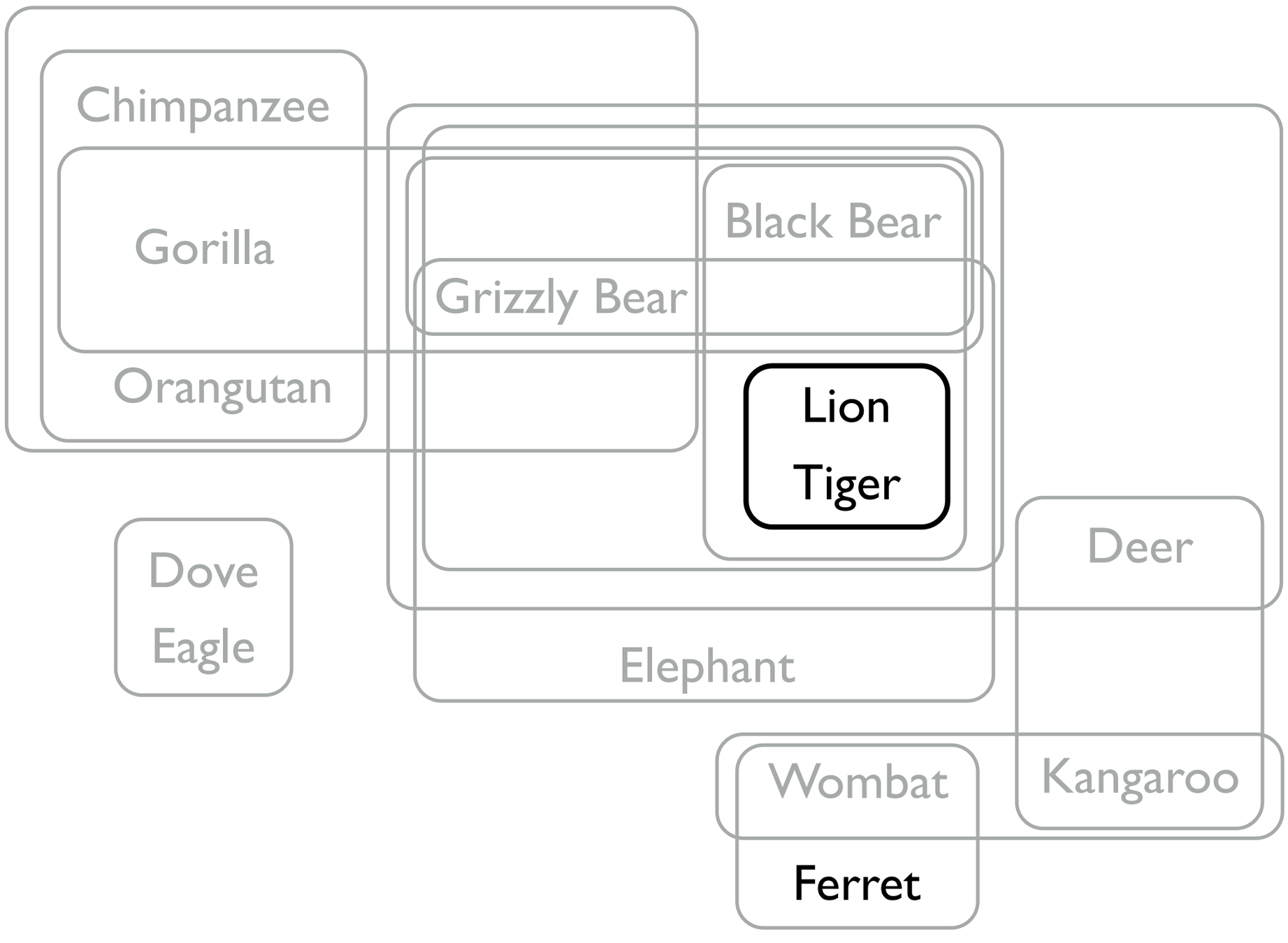


*This illustration only shows a high-weighted sets that contain at least two animals. The actual prior assigned non-zero prior probability to every possible subset of the set of all animals that appeared in the task. Qualitative features of model predictions are robust to the specific choice of prior: anything even semi-reasonable seems to work

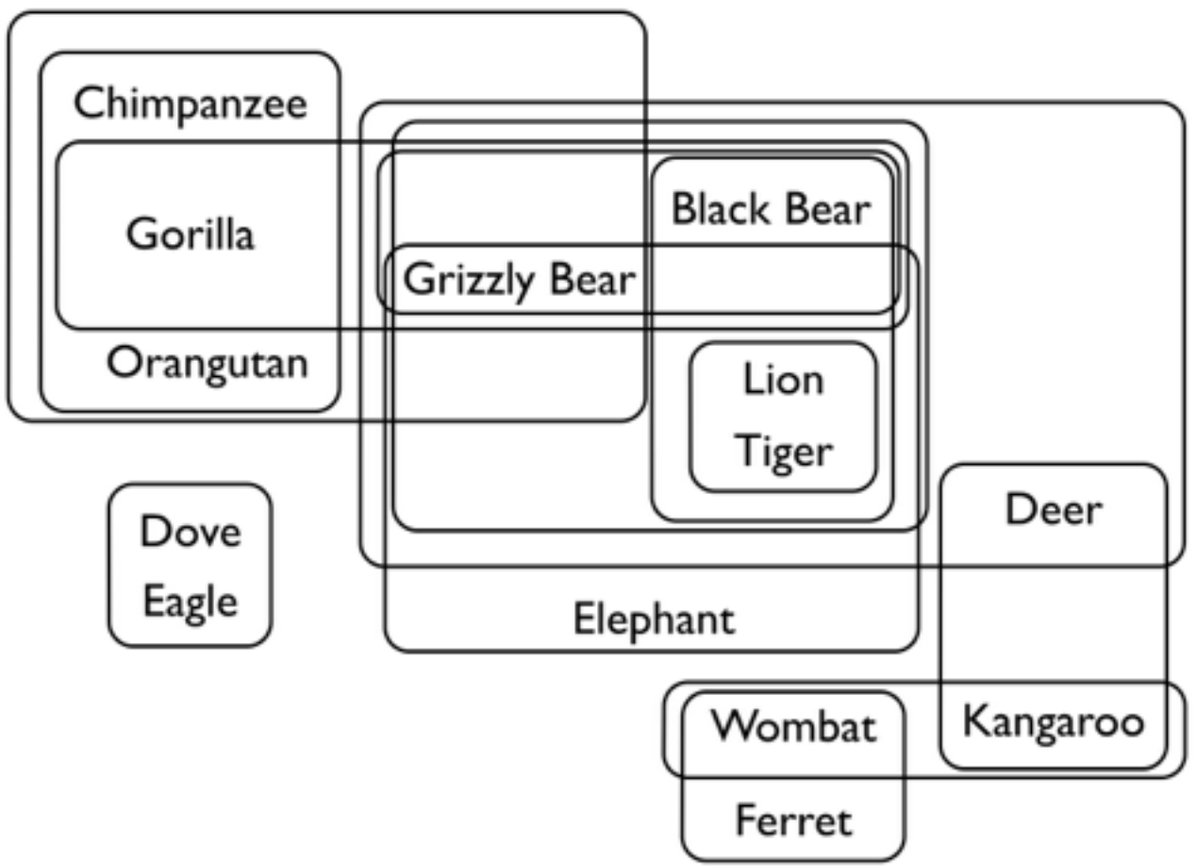
(Chimpanzee, Gorilla, Orangutan)



(Lions, Tigers) but not Ferrets



The prior explains why there are structural differences between the targets and the control



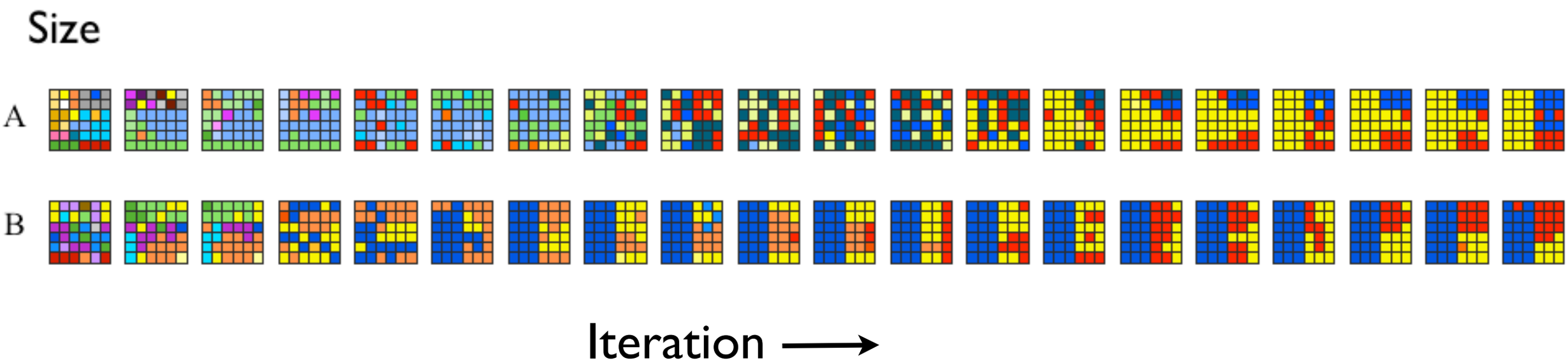
The likelihood describes how "adding more premises" can have different effects across conditions

Stimulus ordering was fixed and designed to ensure that fillers (mostly) preceded targets:

	Additional example	
First generalisation	HELPFUL	RANDOM
EAGLES → DOVES	+HAWKS	−TORTOISES
ELEPHANTS → DEERS	+COWS	+ANTEATERS
TIGERS → FERRETS	+LIONS	+LIONS
KANGAROOS → WOMBATS	+KOALAS	−FLAMINGOS
GRIZZLY BEARS → LIONS	+BLACK BEARS	+BLACK BEARS
ORANGUTANS → GORILLAS	+CHIMPANZEES	+CHIMPANZEES

Participants were 296 people recruited through MTurk

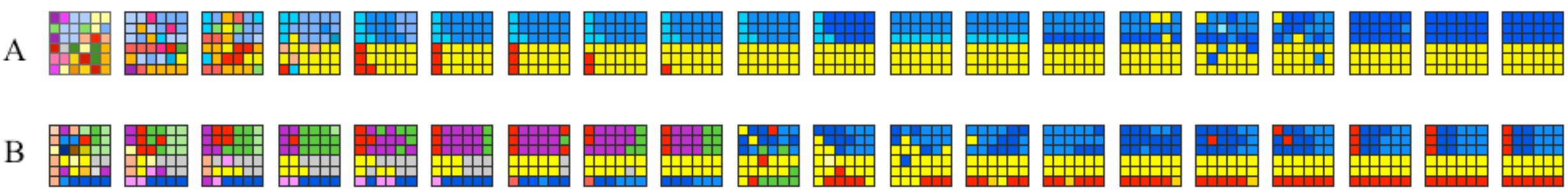
Here's the complete data set...



Here's the complete data set...

Size

Colour



Here's the complete data set...