UNSW
AUSTRALIA

Never Stand Still

On the origins of data: How sampling assumptions influence learning, reasoning and decision making

Dan Navarro

Navarro, Dry & Lee (2012). Sampling assumptions in inductive generalization. *Cognitive Science*

Hendrickson, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations

Ransom, Perfors & Navarro (in press). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*

Voorspoels, Navarro, Perfors, Ransom & Storms (2015). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*

Perfors, Navarro & Shafto (in preparation). Stronger evidence isn't always better: The role of social inference in evidence selection and interpretation.

Navarro, Dry & Lee (2012). Sampling assumptions in inductive generalization. *Cognitive Science*

Hendrickson, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations

Ransom, Perfors & Navarro (in press). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*

Voorspoels, Navarro, Perfors, Ransom & Storms (2015). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*

Perfors, Navarro & Shafto (in preparation). Stronger evidence isn't always better: The role of social inference in evidence selection and interpretation.

Navarro, Dry & Lee (2012). Sampling assumptions in inductive generalization. *Cognitive Science*

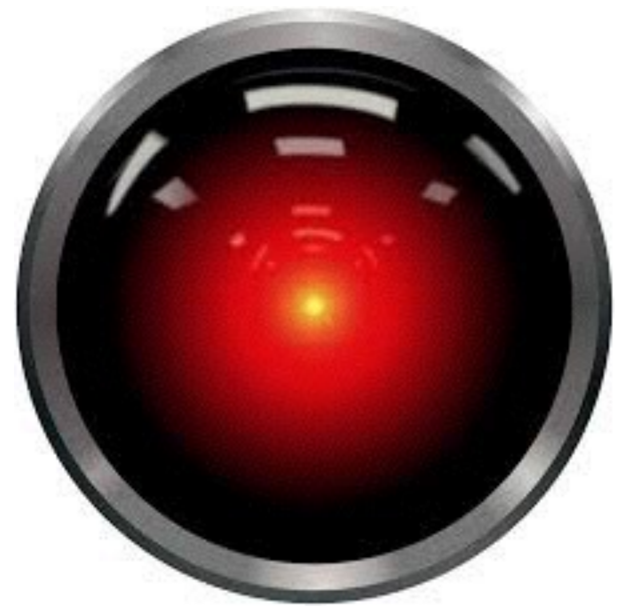
Drew Hendrickson, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations

Keith Ransom, Perfors & Navarro (in press). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*

Wouter Voorspoels, Navarro, Perfors, Ransom & Storms (2015). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*

Perfors, Navarro & Shafto (in preparation). Stronger evidence isn't always better: The role of social inference in evidence selection and interpretation.

So… I want to build a smart machine, and I want it to do human-like inductive reasoning



So… I want to build a smart machine, and I want it to do human-like inductive reasoning

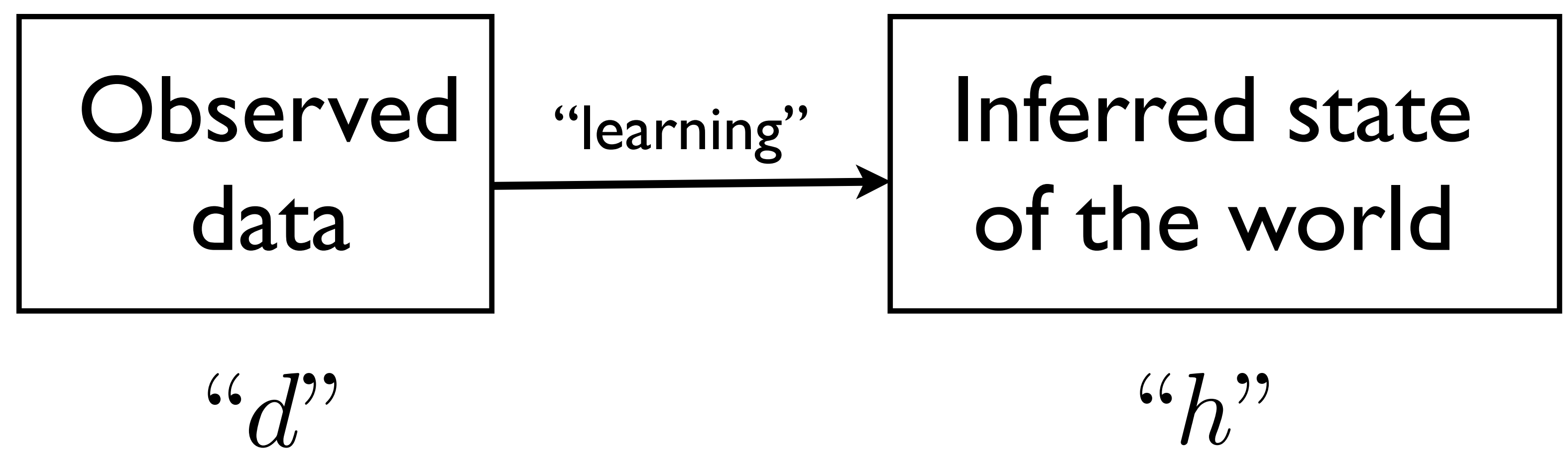
I want it to have **common sense**

Why isn’t inductive inference simple?

Why isn't inductive inference simple?

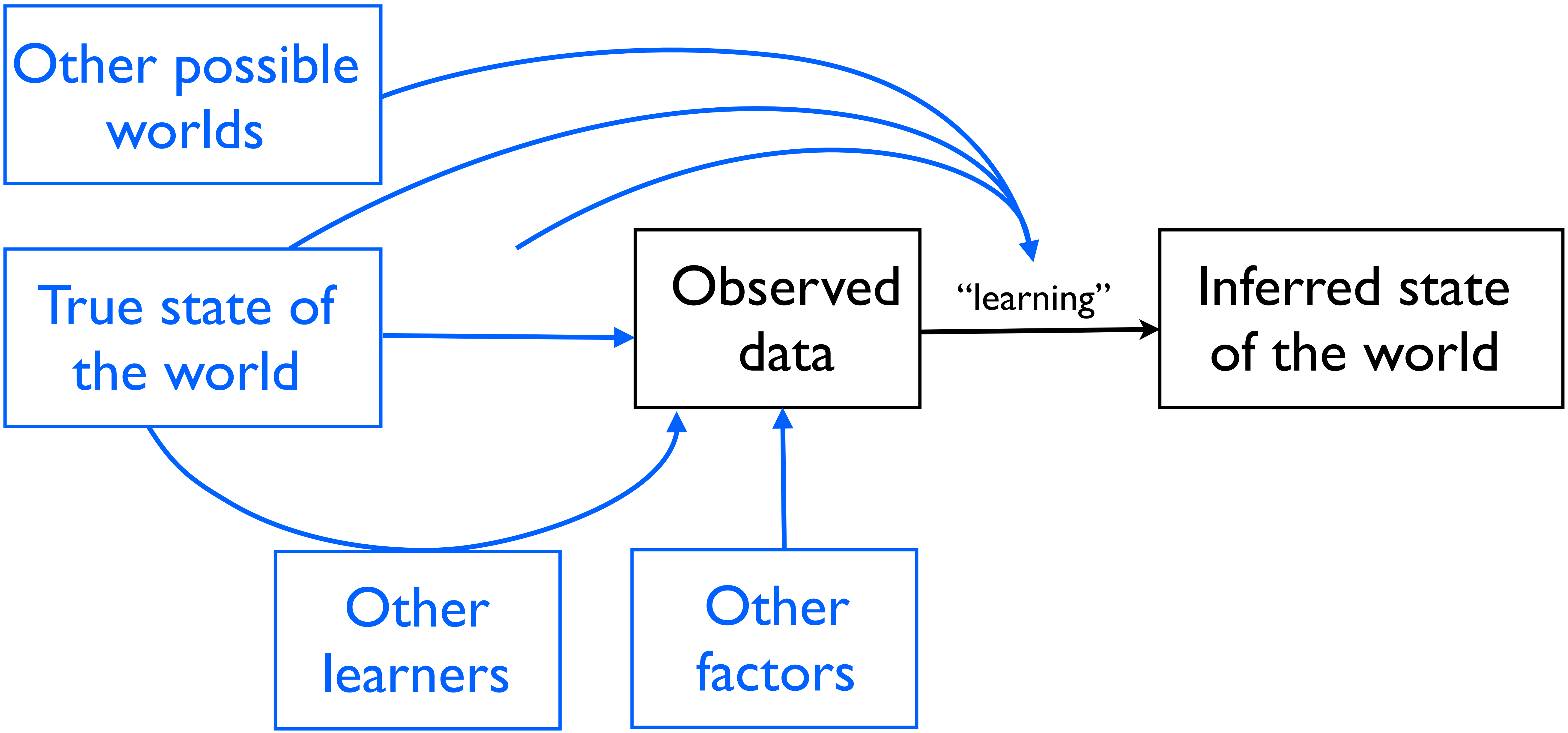
$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

A simple learning rule...

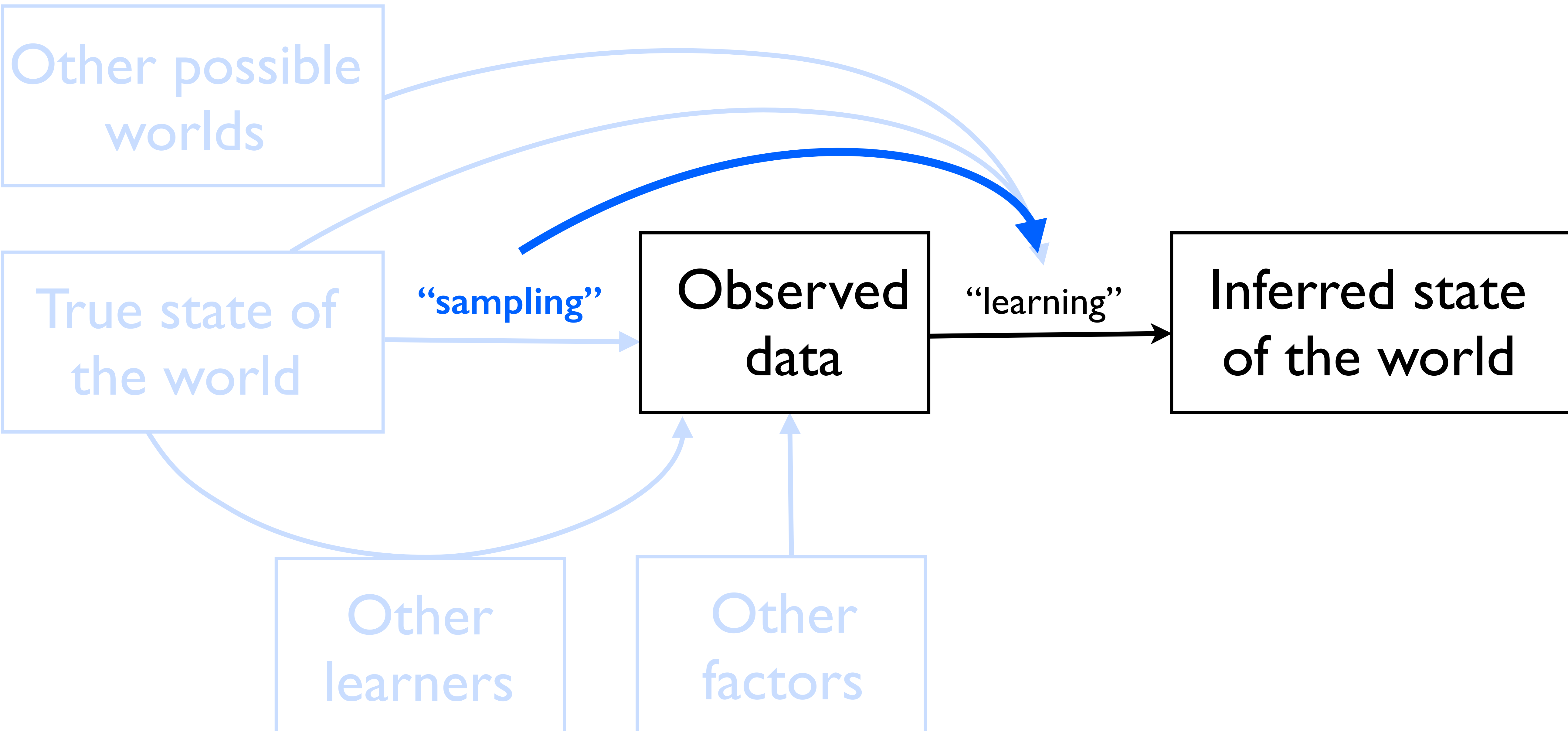


$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

… hides a lot of complexity



The evidentiary value of data depends on where it comes from



What do people assume about the data generating mechanism in simple learning problems?

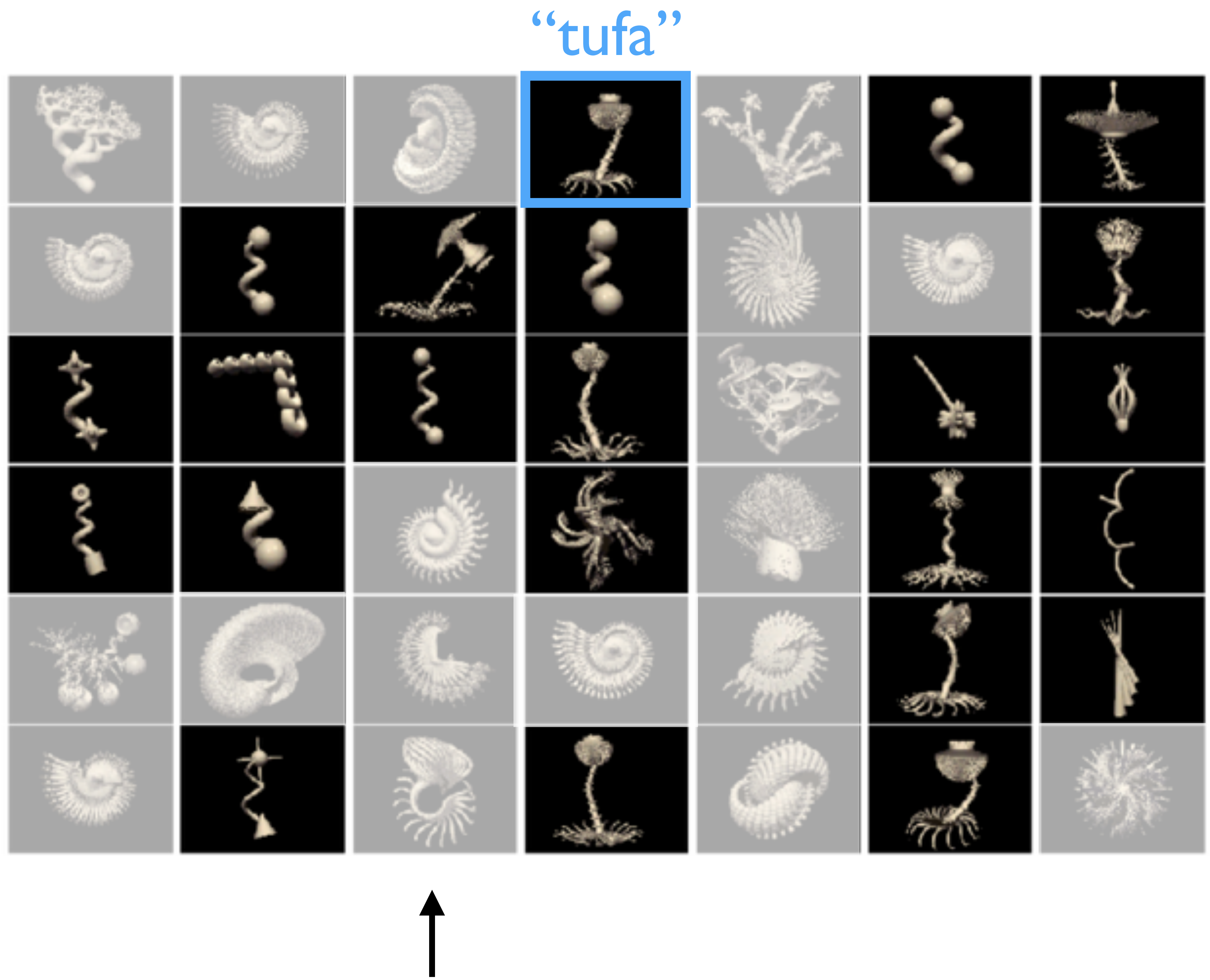
Navarro, Dry & Lee (2012). Sampling assumptions in inductive generalization. *Cognitive Science*

Hendrickson, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations

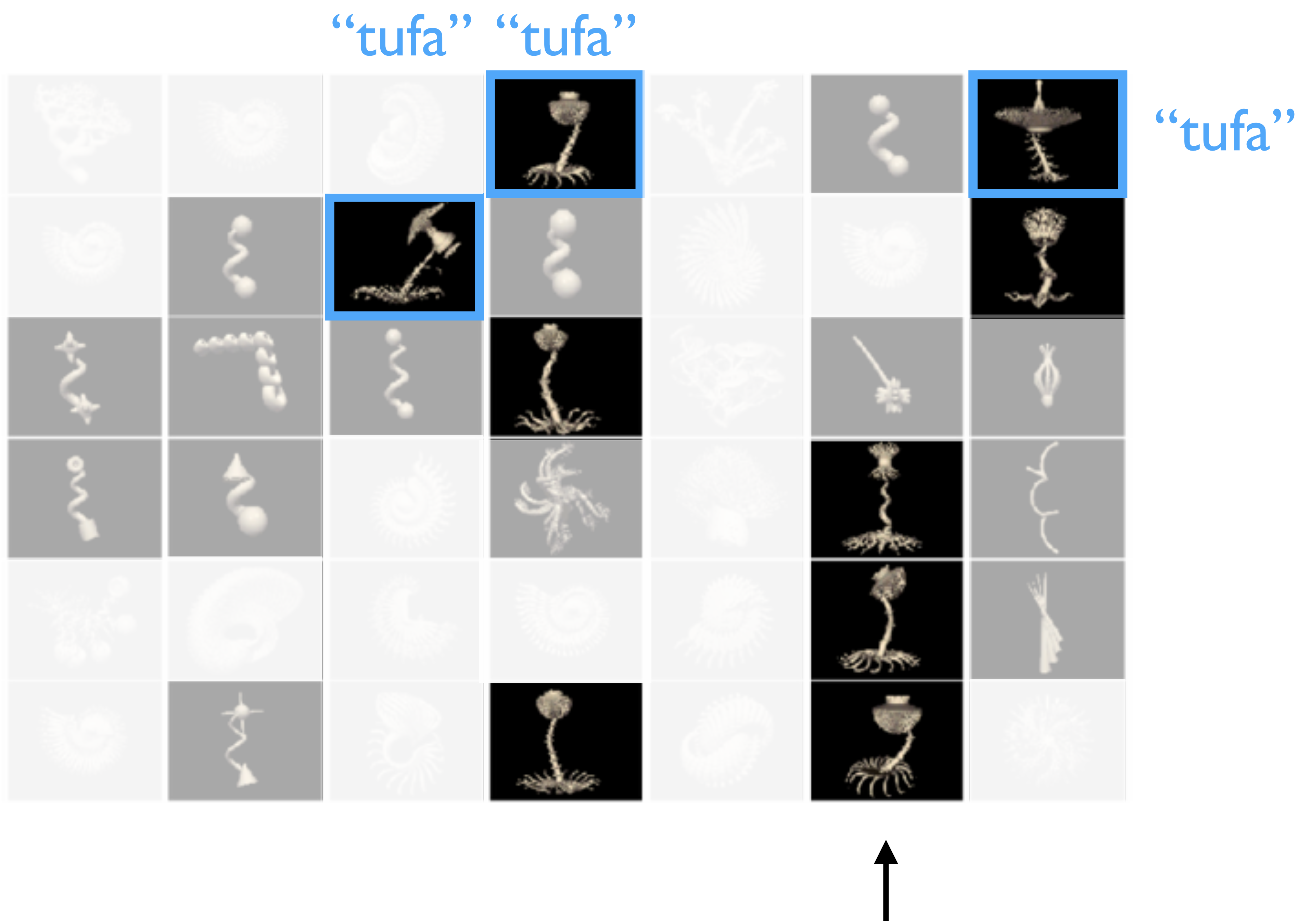
Ransom, Perfors & Navarro (in press). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*

Voorspoels, Navarro, Perfors, Ransom & Storms (under revision). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*

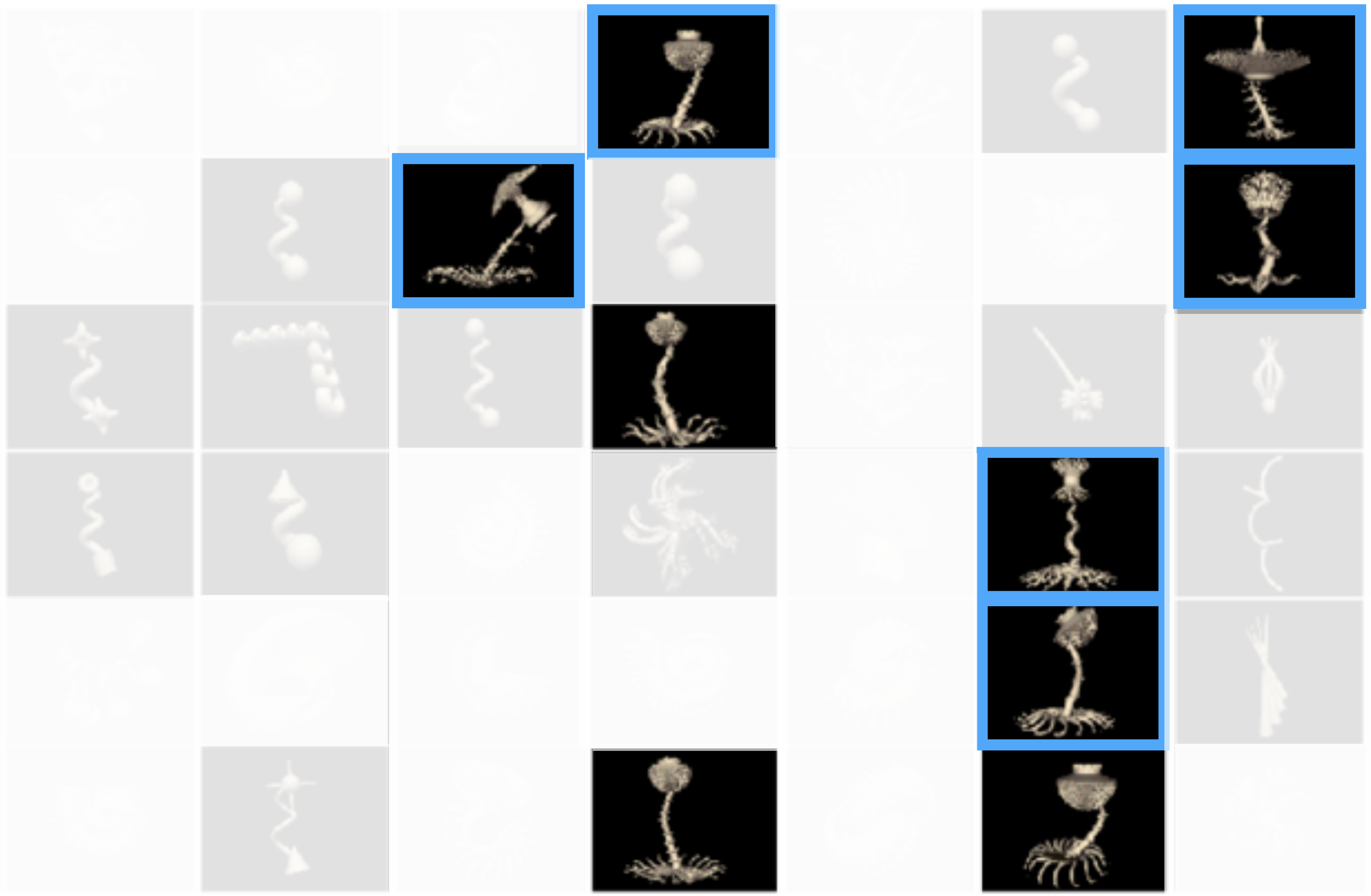
Perfors, Navarro & Shafto (in preparation). Stronger evidence isn't always better: The role of social inference in evidence selection and interpretation. Previously rejected from *JEP:G*



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



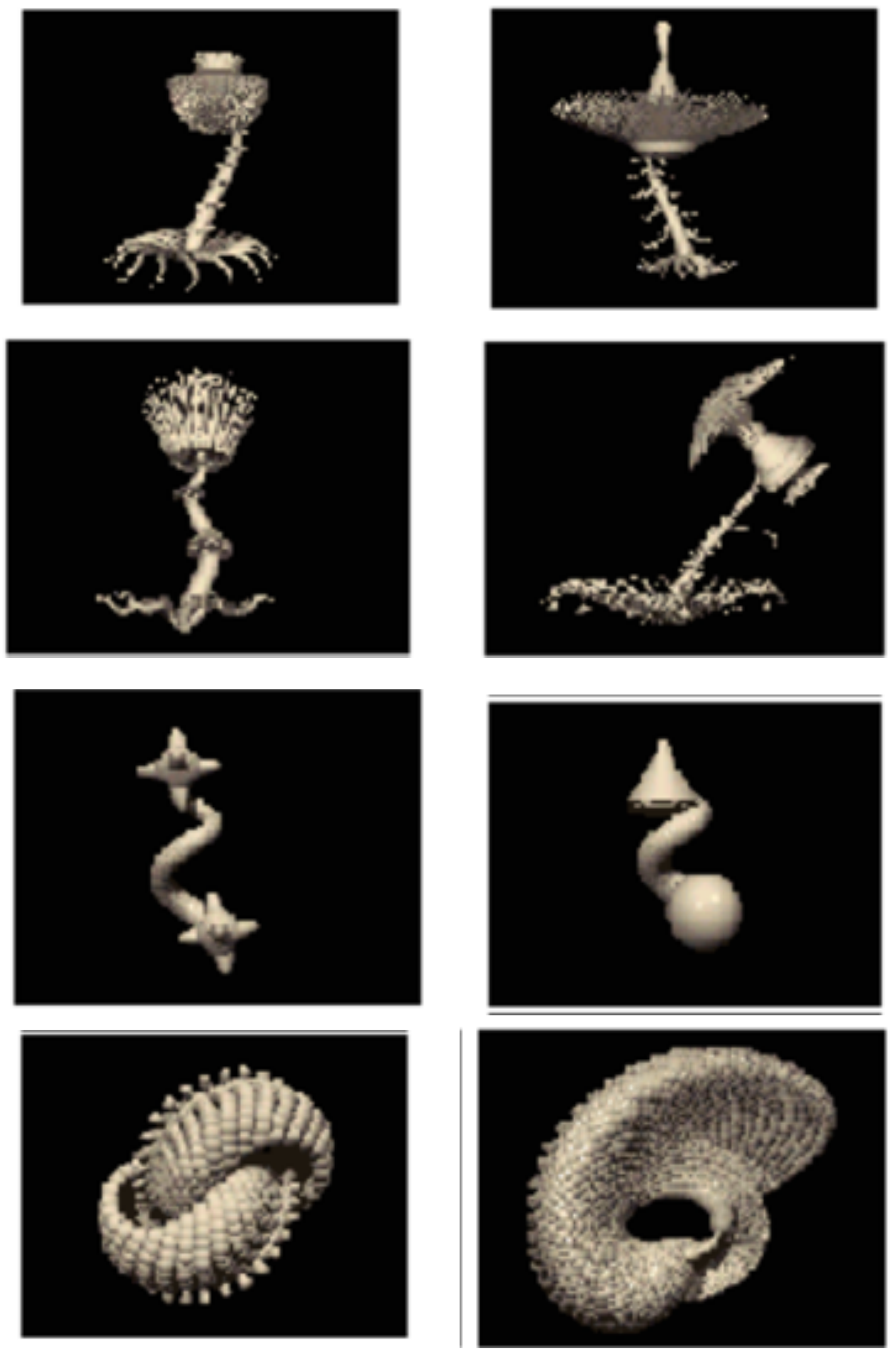
"tufas"

Sir Ronald would like a few words…

"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"



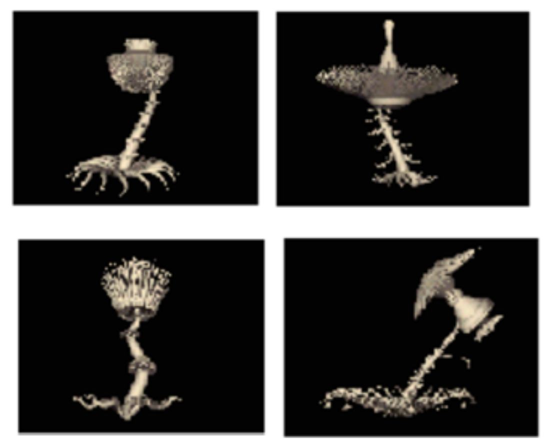
← Things!

Which things are tufas?

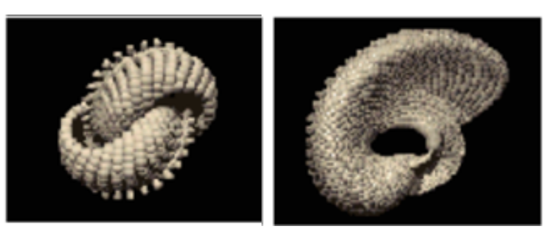
helical things?

mushroom heads?

creepy flowers?

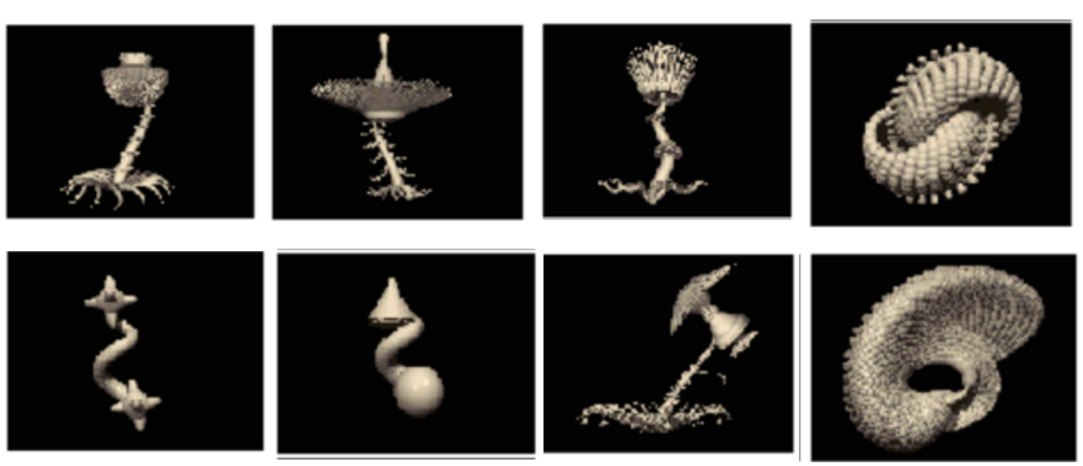


seashell things?

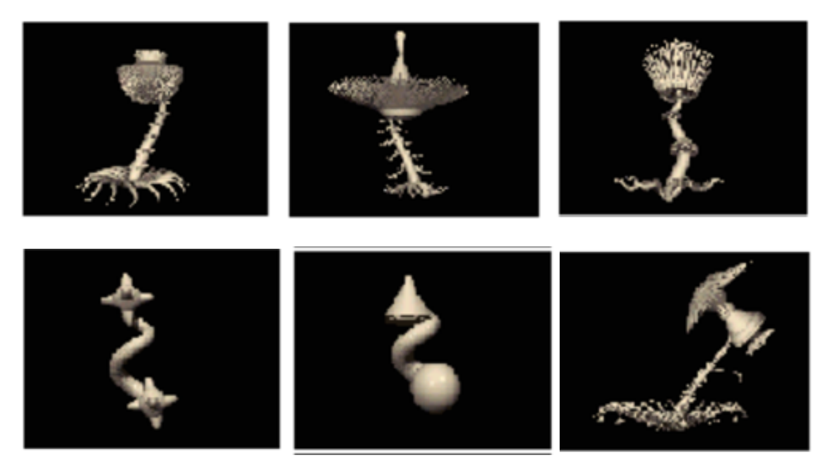


botanical radio telescopes?

they're all tufas!



anything long and narrow

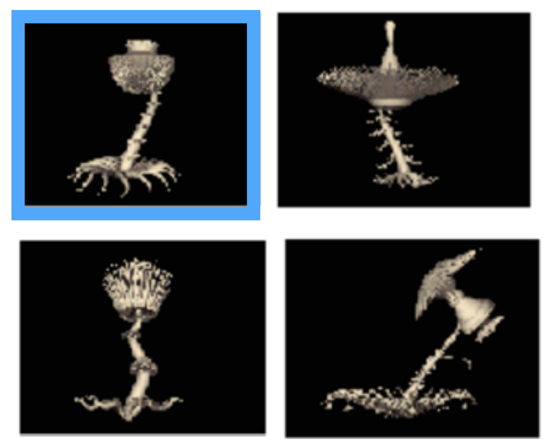


Observe one tufa and falsify…

helical things?

seashell things?

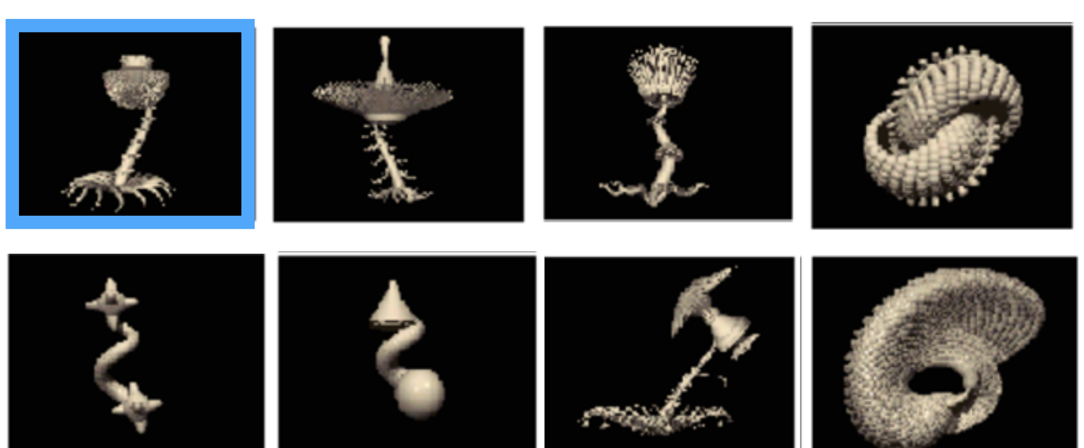
creepy flowers?



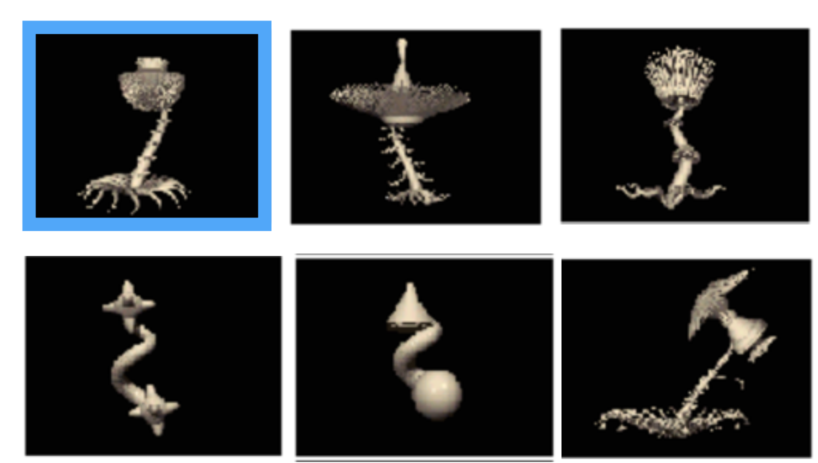
mushroom heads?

botanical radio telescopes?

they're all tufas!



anything long and narrow

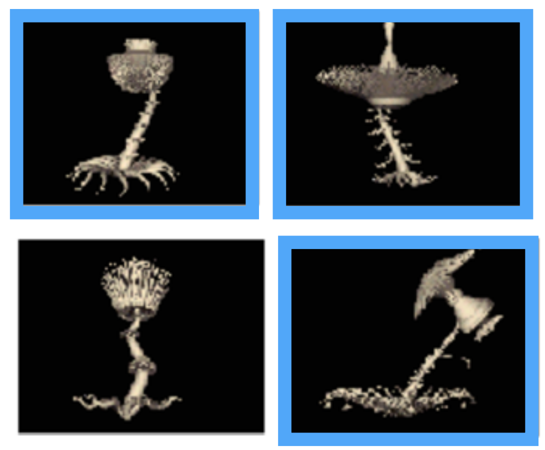


See two more and do nothing????

helical things?

mushroom heads?

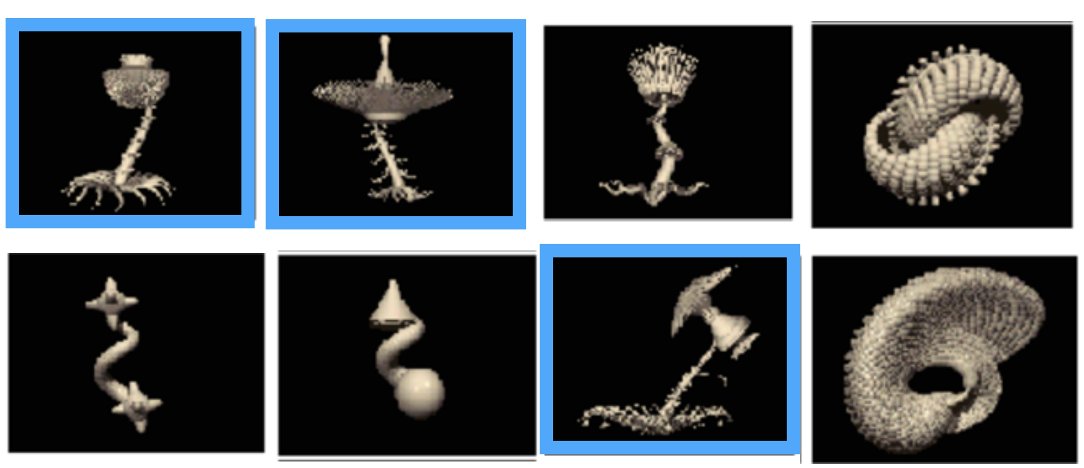
creepy flowers?



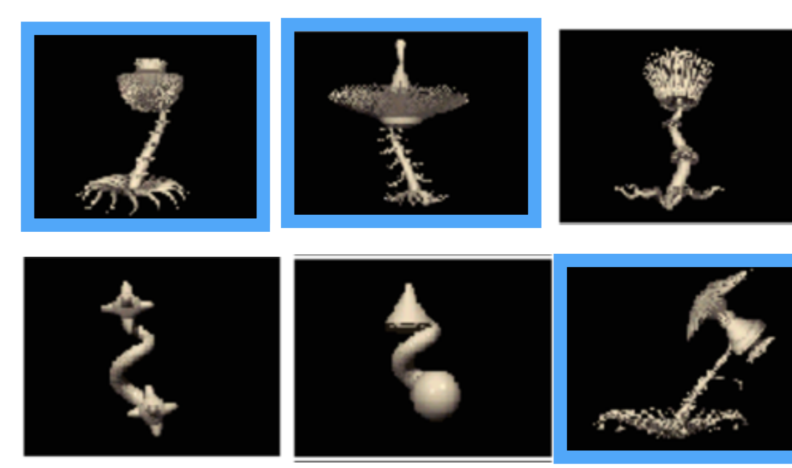
botanical radio
telescopes?

seashell things?

they're all tufas!



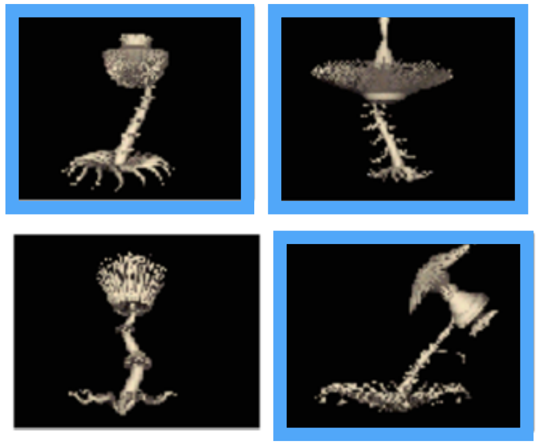
anything long and narrow



Mr. Ockham wishes to discuss tufas with you…

mushroom heads?

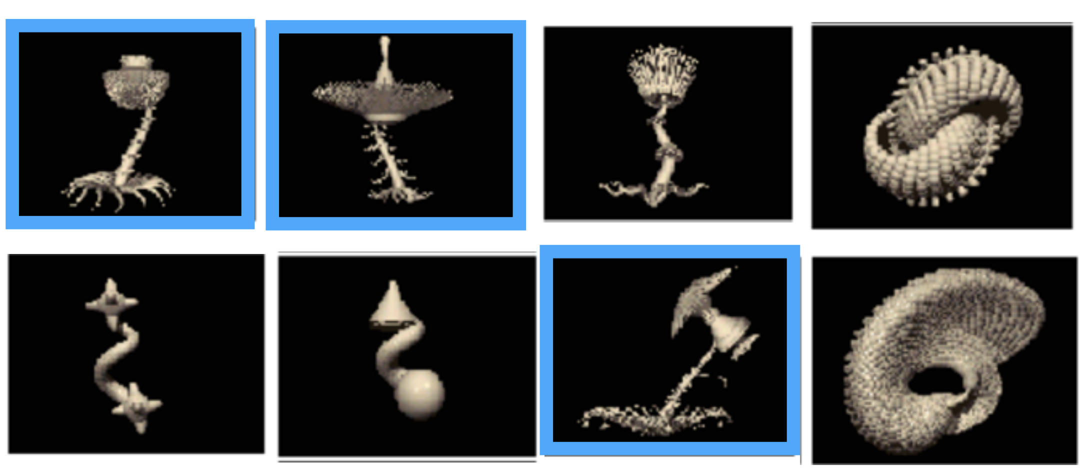
creepy flowers?



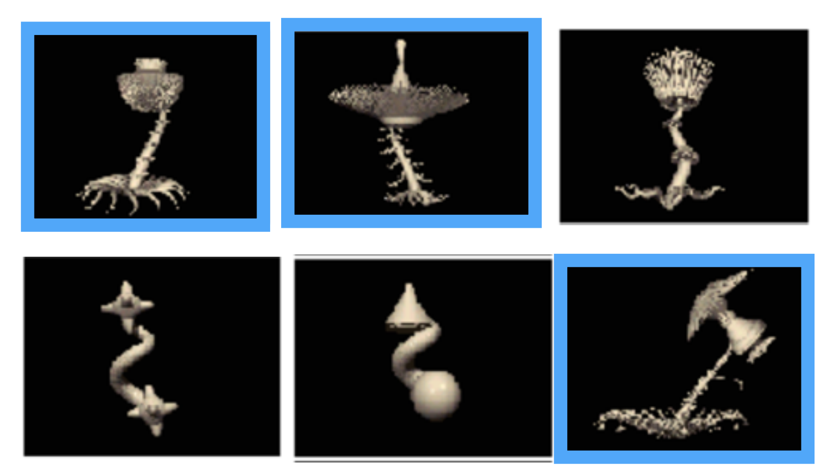
These hypotheses do not require me to believe a bizarre coincidence as to why the only observed tufas are so bloody similar

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

they're all tufas!



anything long and narrow

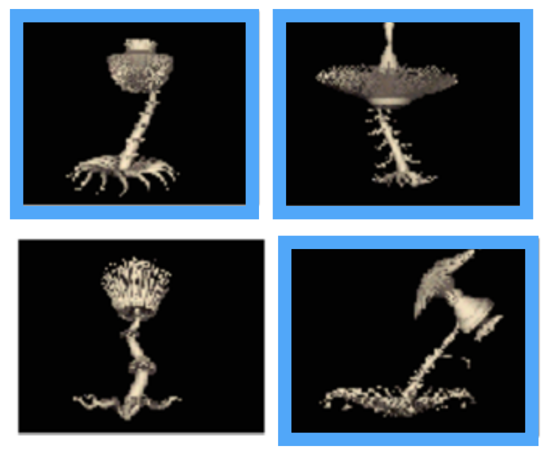


An Ockhamist reasoner has little faith in coincidences

helical things?

mushroom heads?

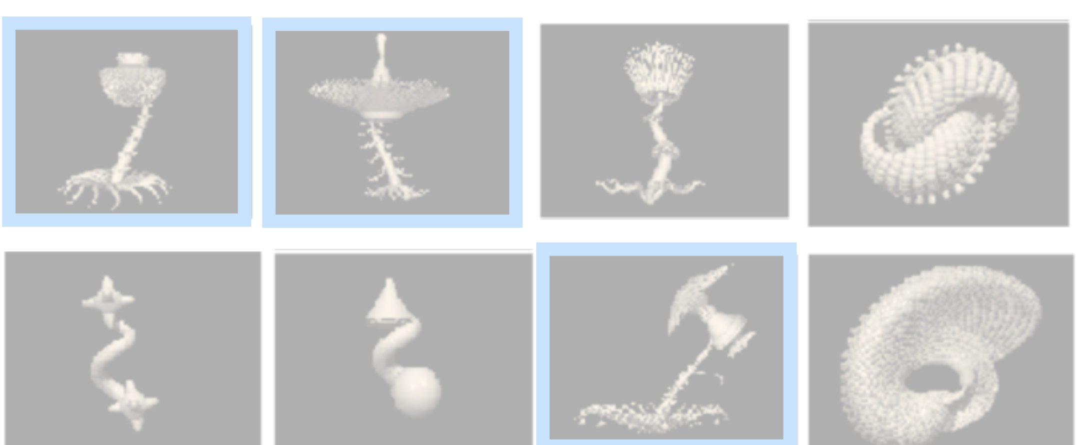
creepy flowers?



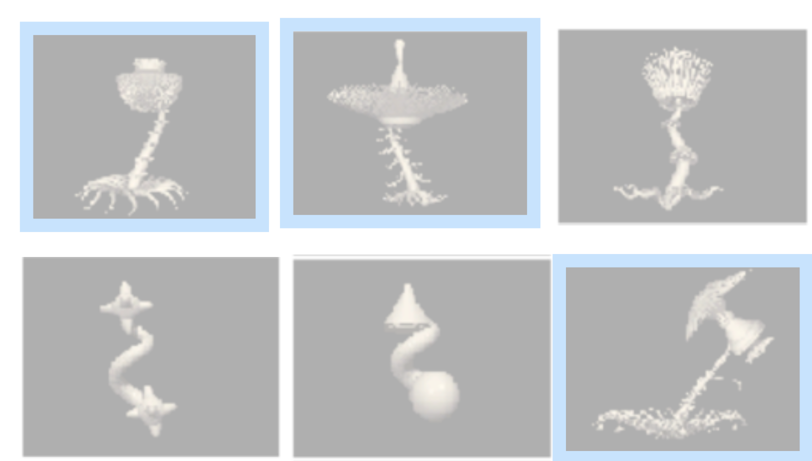
botanical radio telescopes?

seashell things?

they're all tufas!



anything long and narrow



Are these fundamentally distinct?

Or can we express them in a common framework?

Bayes' 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?

$$\downarrow$$

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

The likelihood is the learner's **theory** about the problem they're solving

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

Different theories, different learning

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

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

Two simple theories about the data generating mechanism…

Weak sampling:

“select an item at random and then provide the category label”

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 from 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:

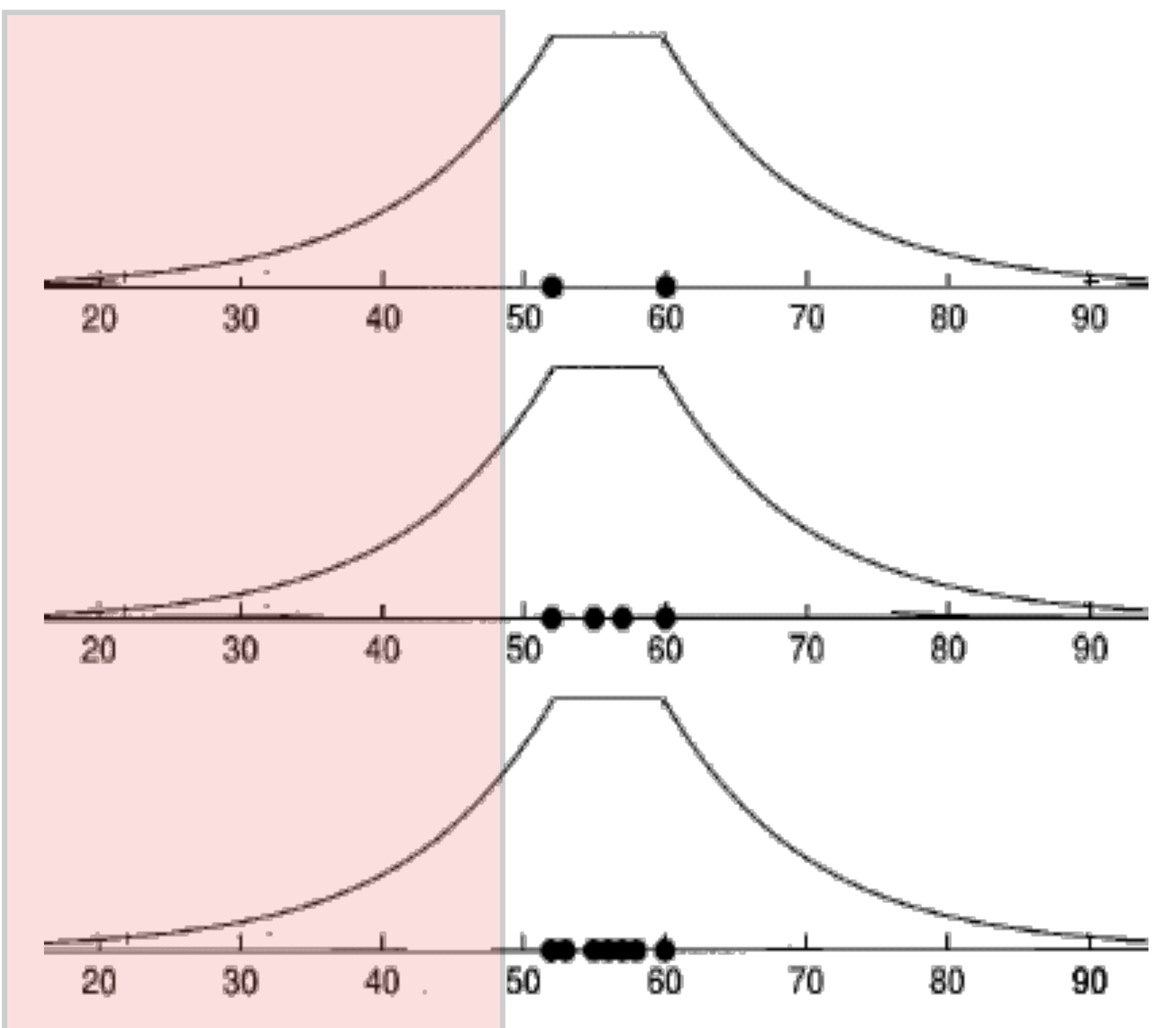
Act like a falsificationist

Strong sampling:

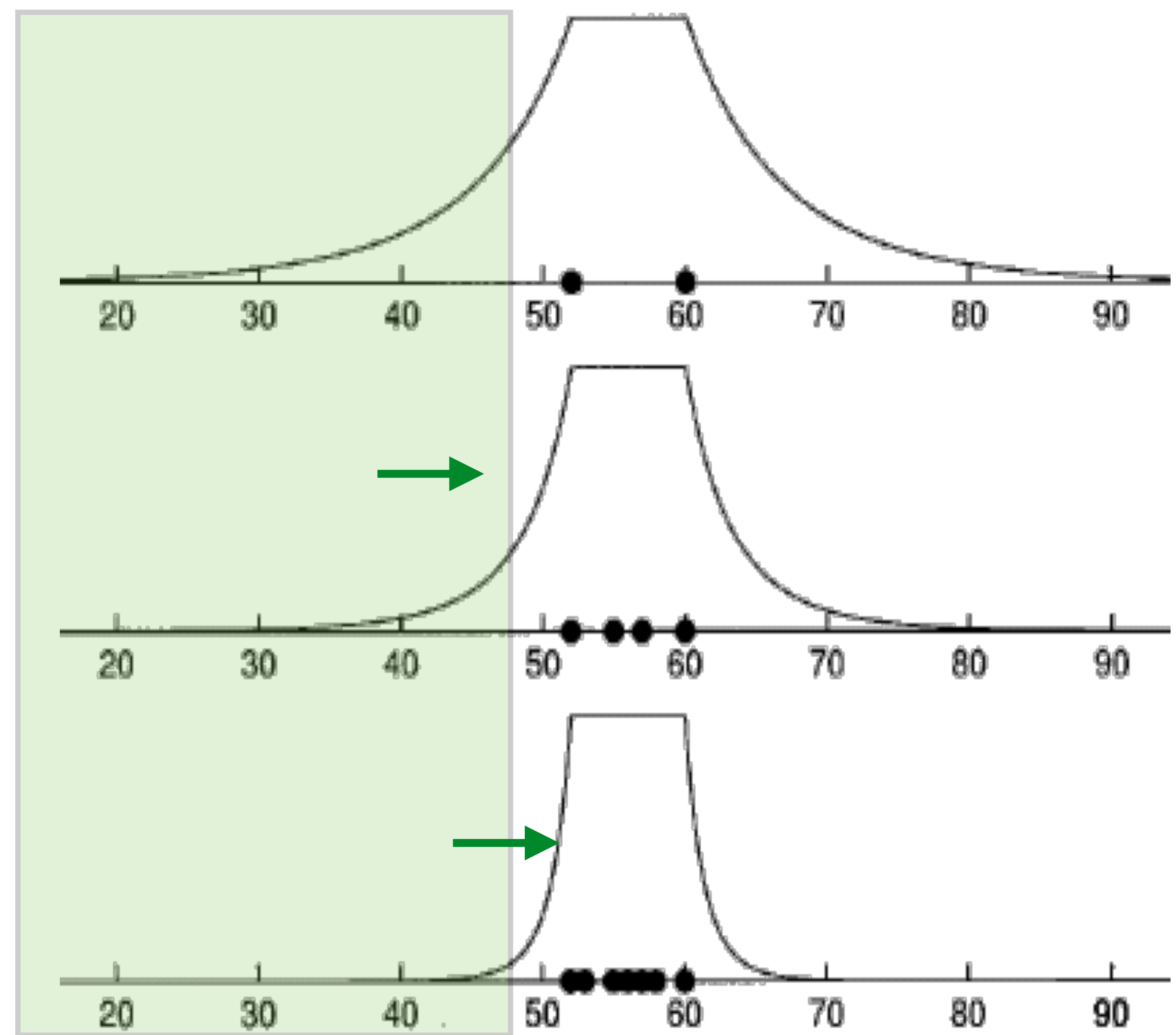
Apply Ockham’s razor: prefer small/simple hypotheses

Here's the testable prediction about generalisation gradients…

weak sampling



strong sampling



And a series of 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

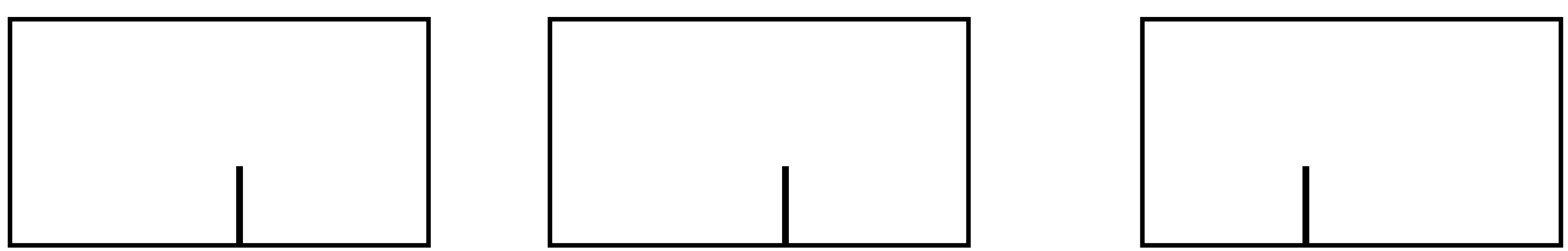
And a series of experimental tests…

- 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

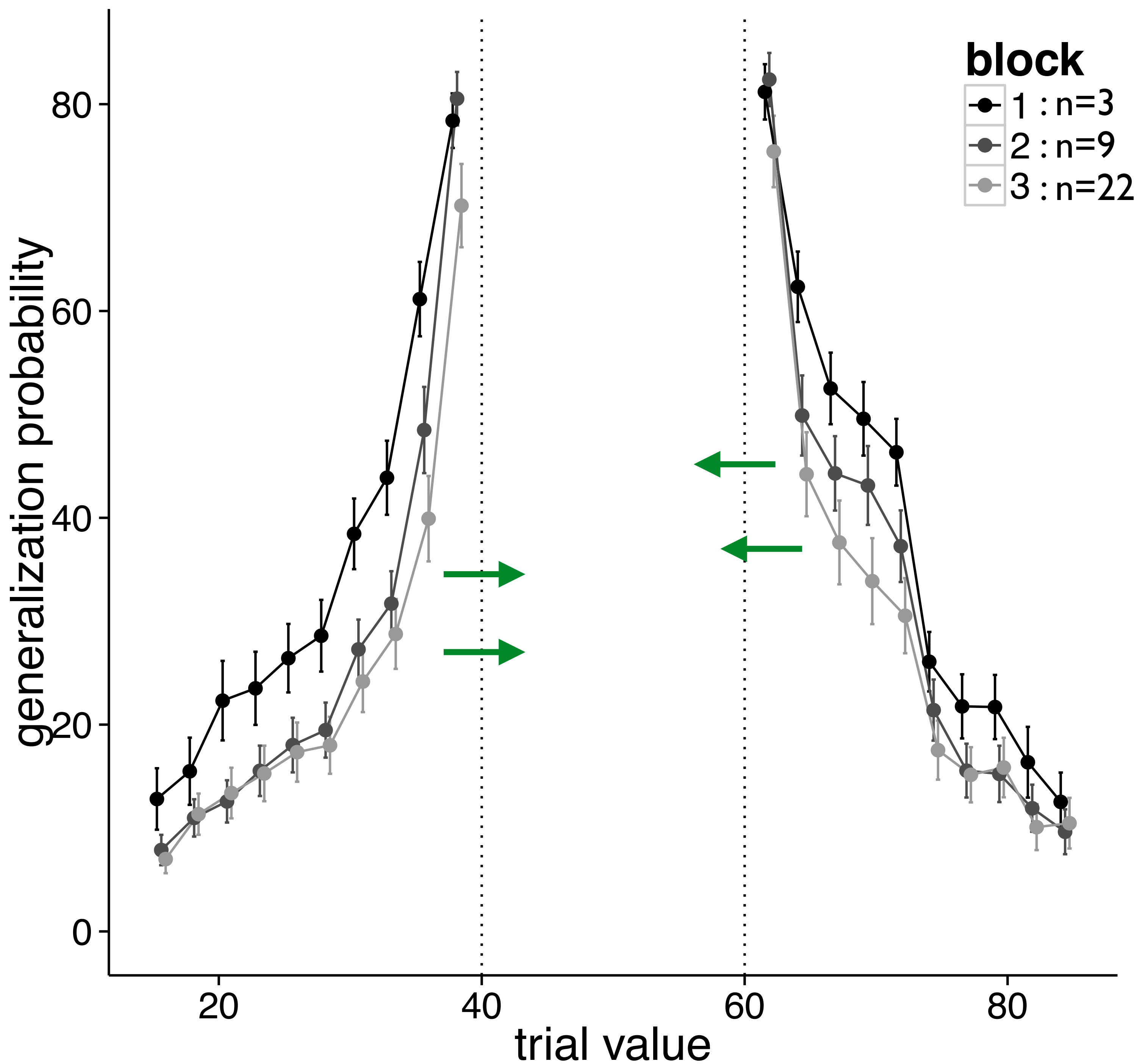
And a series of experimental tests…

- 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:

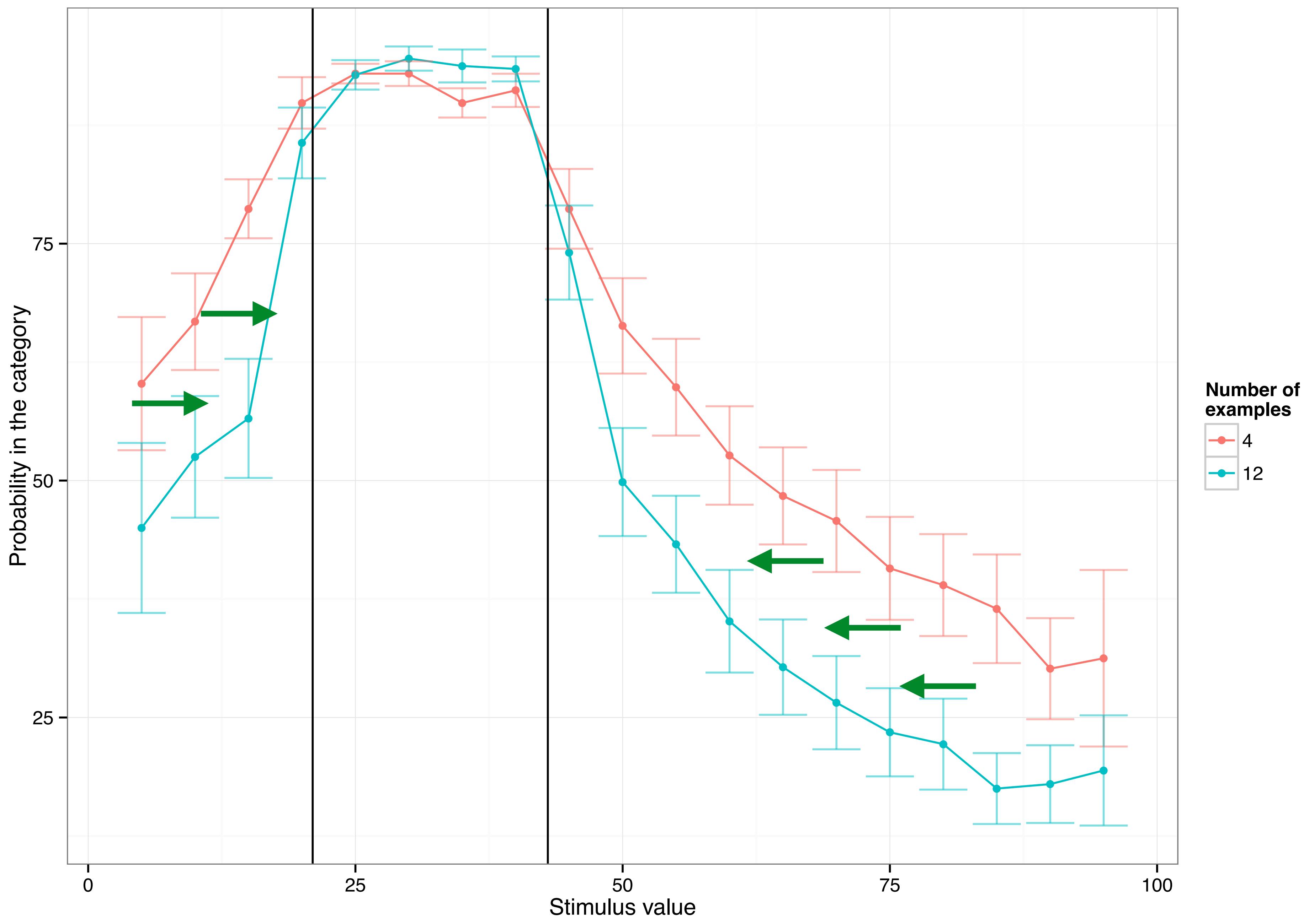


Looks like strong sampling…



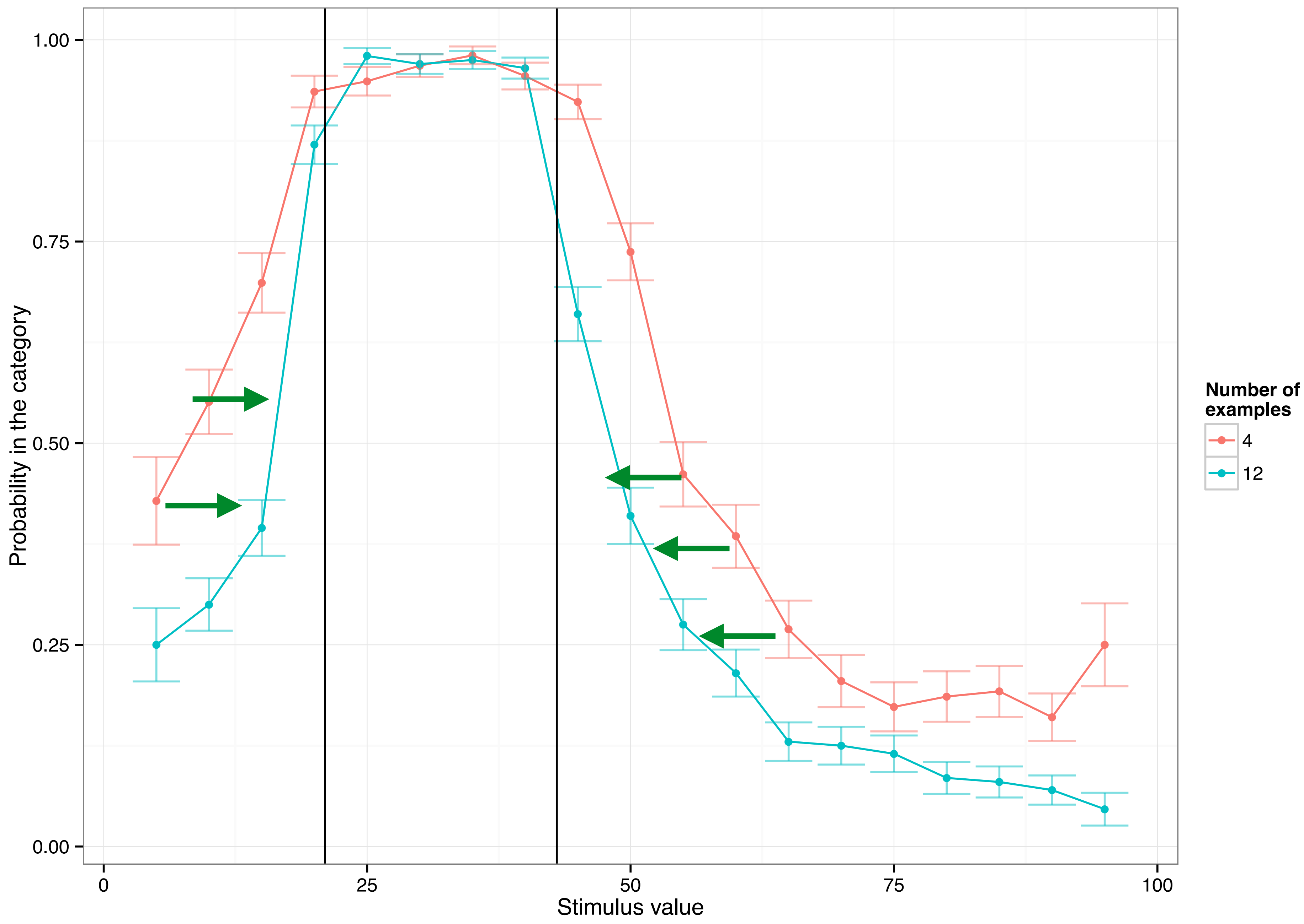
Vong, et al (2013) - probability judgment with “toy” task

Looks like strong sampling…



Hendrickson, et al (in prep) - probability judgments

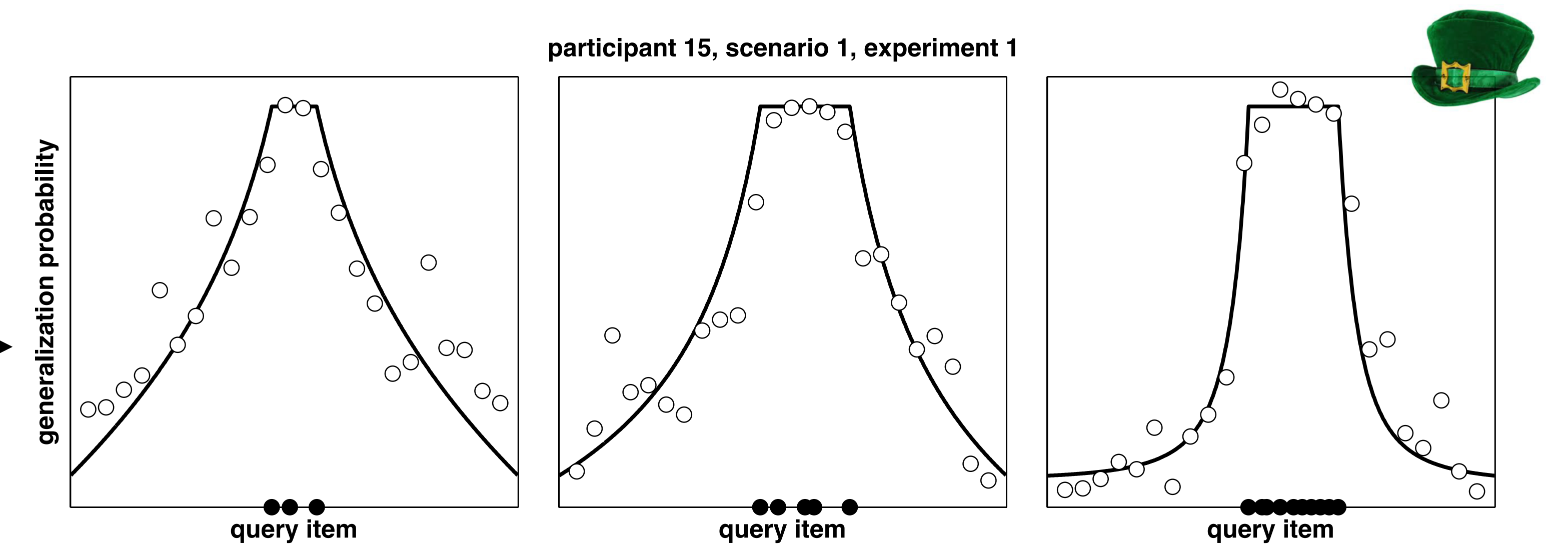
Looks like strong sampling…



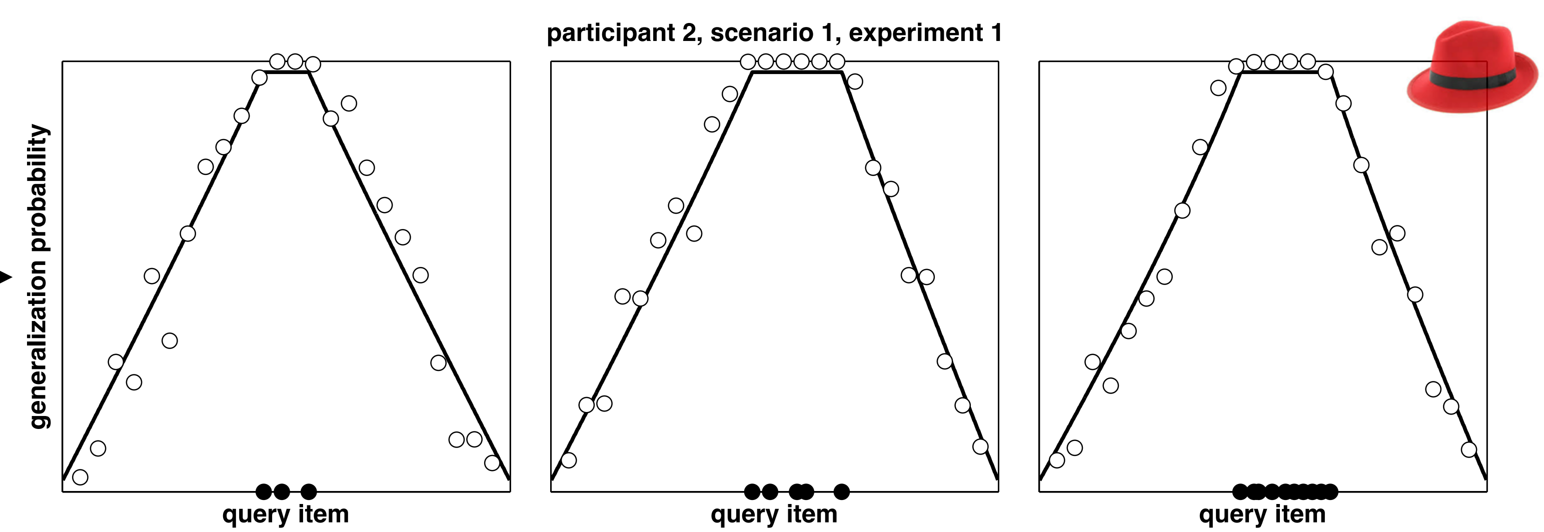
Hendrickson, et al (in prep) - categorisation data

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

Sensitivity to sample size in simple generalisation

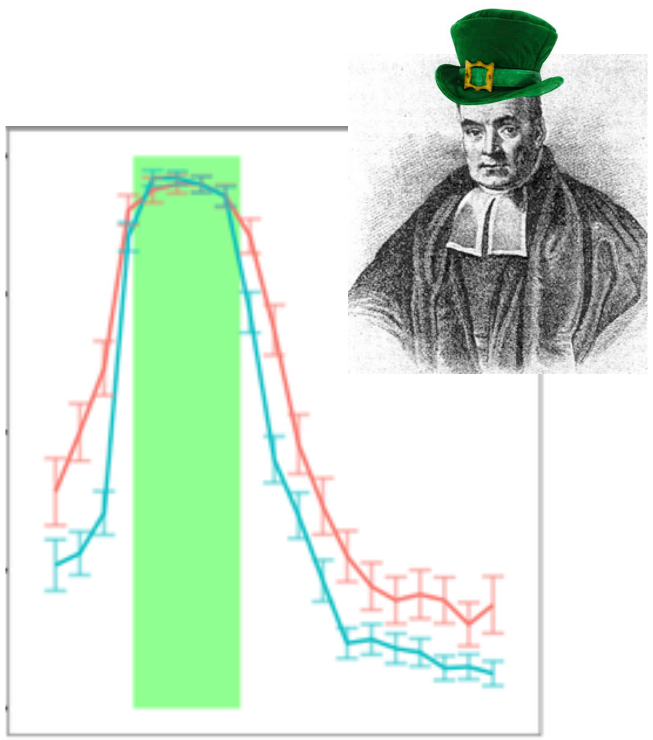


Insensitivity to sample size in simple generalisation



Navarro et al (2012)

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

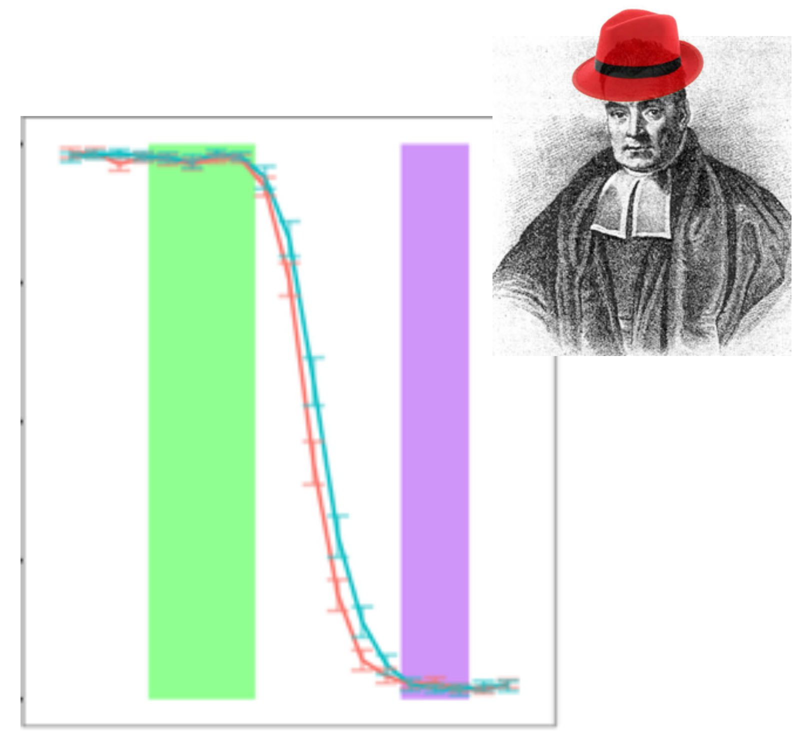


"Concept learning" designs where people see positive examples from one category produce the strong sampling "tightening" effect

Hendrickson, et al (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, et al (in prep)

- The tightening effect predicted by strong sampling does happen
- But there are differences across individuals and across tasks

- The task differences make sense if you assume people are forming theories about how the experiment(er) designed the task
- This starts to feel like <u>social</u> cognition…

Manipulating the sampling assumption in an inductive reasoning task

Navarro, Dry & Lee (2012). Sampling assumptions in inductive generalization. *Cognitive Science*

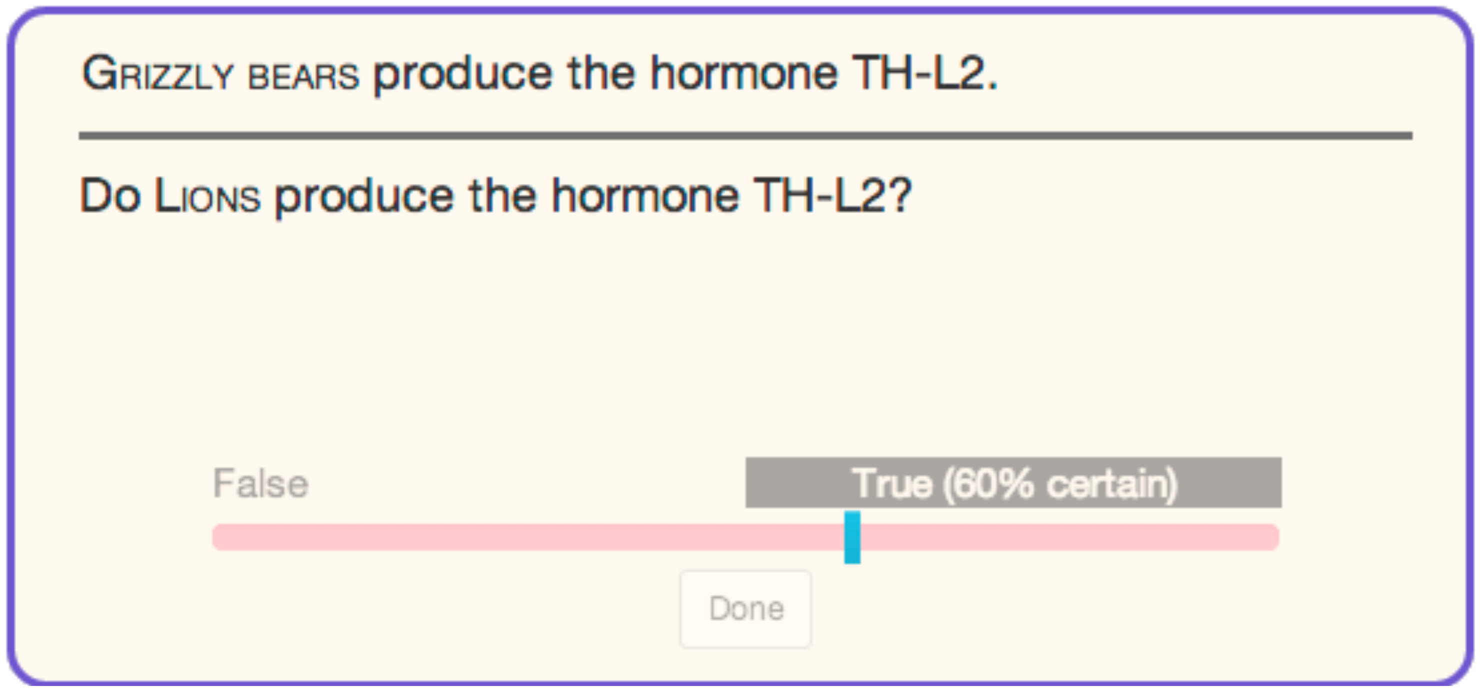
Hendrickson, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations

Ransom, Perfors & Navarro (in press). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*

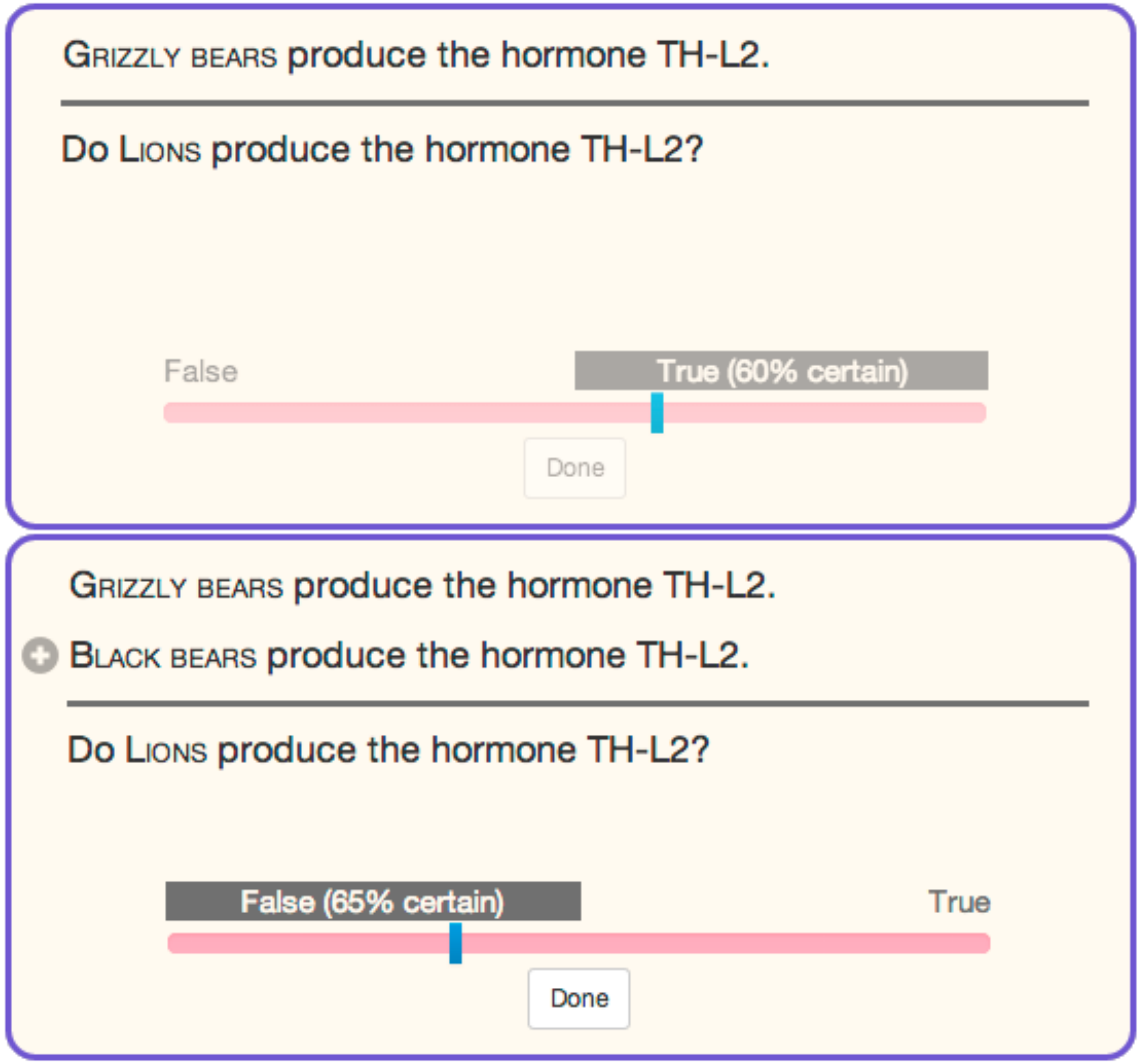
Voorspoels, Navarro, Perfors, Ransom & Storms (under revision). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*

Perfors, Navarro & Shafto (in preparation). Stronger evidence isn't always better: The role of social inference in evidence selection and interpretation. Previously rejected from *JEP:G*

Property induction tasks



Property induction tasks



Grizzly Bears → Lions

Grizzly Bears + Black Bears → Lions

Adding the "Black Bears" premise weakens the argument?

Grizzly Bears → Lions

Grizzly Bears + Black Bears → Lions

Tigers → Ferrets

Tigers + Lions → Ferrets

↑

Same thing with the "Lions" premise

Grizzly Bears → Lions
Grizzly Bears + Black Bears → Lions

Tigers → Ferrets
Tigers + Lions → Ferrets

Orangutans → Gorillas
Orangutans + Chimpanzees → Gorillas

Conversely, the "Chimpanzee" premise strengthens the argument here

Grizzly Bears + Black Bears
Tigers + Lions
Orangutans + Chimpanzees

In all cases the additional premise concentrates beliefs around a target category, e.g. bears, cats, primates

Is this "tightening" effect related to the tightening in the "tufa" generalisation problem?

Is this “tightening” effect related to the tightening in the “tufa” generalisation problem?

Does it depend on the learner’s theory about how the argument was constructed?

Is this “tightening” effect related to the tightening in the “tufa” generalisation problem?

Does it depend on the learner’s theory about how the argument was constructed?

Can we produce qualitative shifts in people’s reasoning by manipulating their theory about how the argument was made?

Cover story?	Previous experience? (filler trials)	
	Relevant cover story, Relevant fillers	
	Neutral cover story, Relevant fillers	Neutral cover story, Random fillers
		Random cover story, Random fillers

Cover story manipulation

- Relevant: people were told that the “additional” premise was chosen by a helpful teacher
- Neutral: people were told nothing about how the second premise was generated

- Random: people were told that the second premise was selected at random from the set of true facts.

Eagles → Doves
Elephants → Deer
Kangaroos → Wombats

Three "filler" arguments

Eagles → Doves
Elephants → Deer
Kangaroos → Wombats

Eagles + Hawks → Doves
Elephants + Cows → Deer
Kangaroos + Koalas → Wombats

… with a relevant second premise

(positive premises from the same category suggest strong sampling)

Eagles → Doves
Elephants → Deer
Kangaroos → Wombats

Eagles + Hawks → Doves
Elephants + Cows → Deer
Kangaroos + Koalas → Wombats

Eagles - Tortoises → Doves
Elephants + Anteaters → Deer
Kangaroos - Flamingos → Wombats

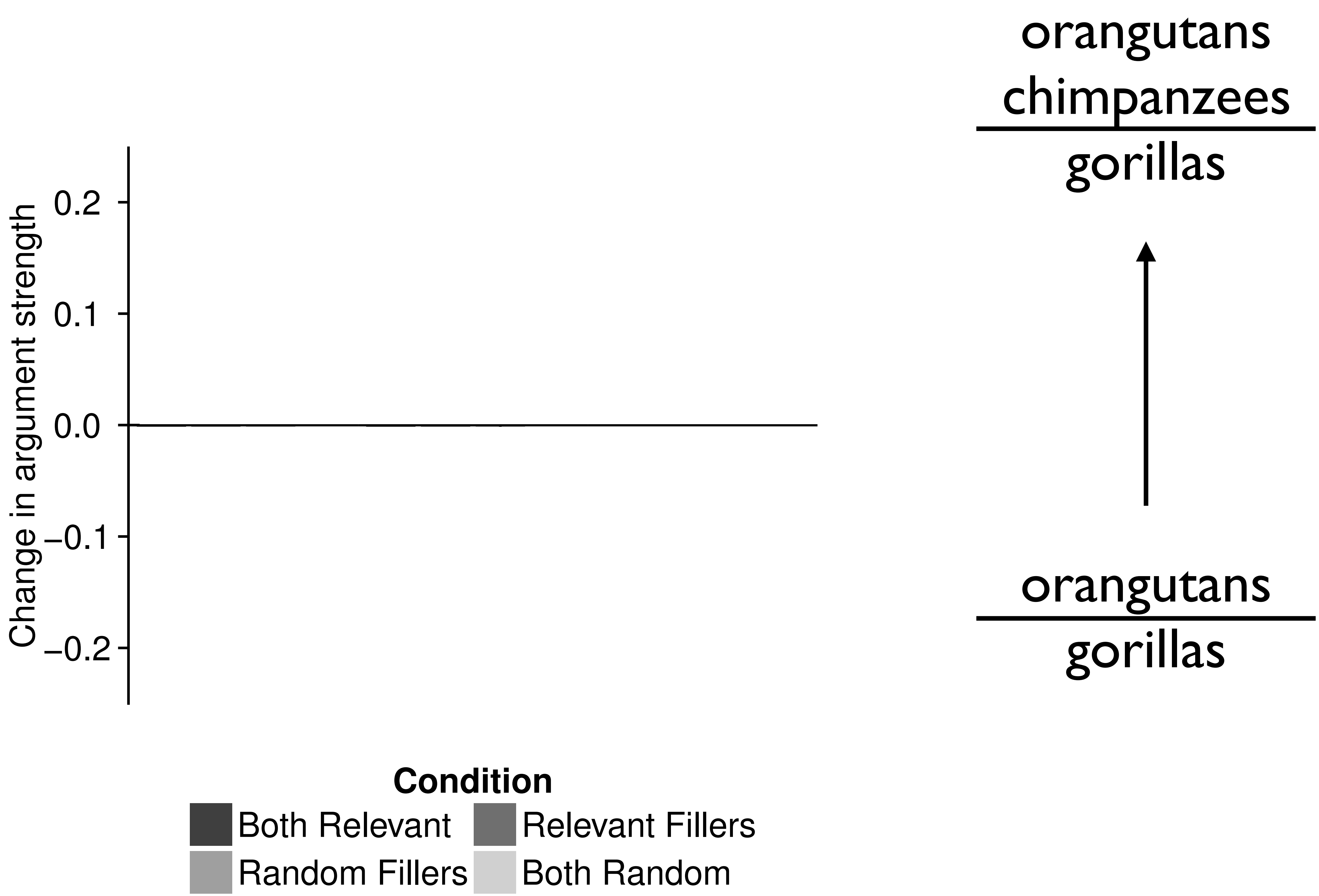
… or a random one

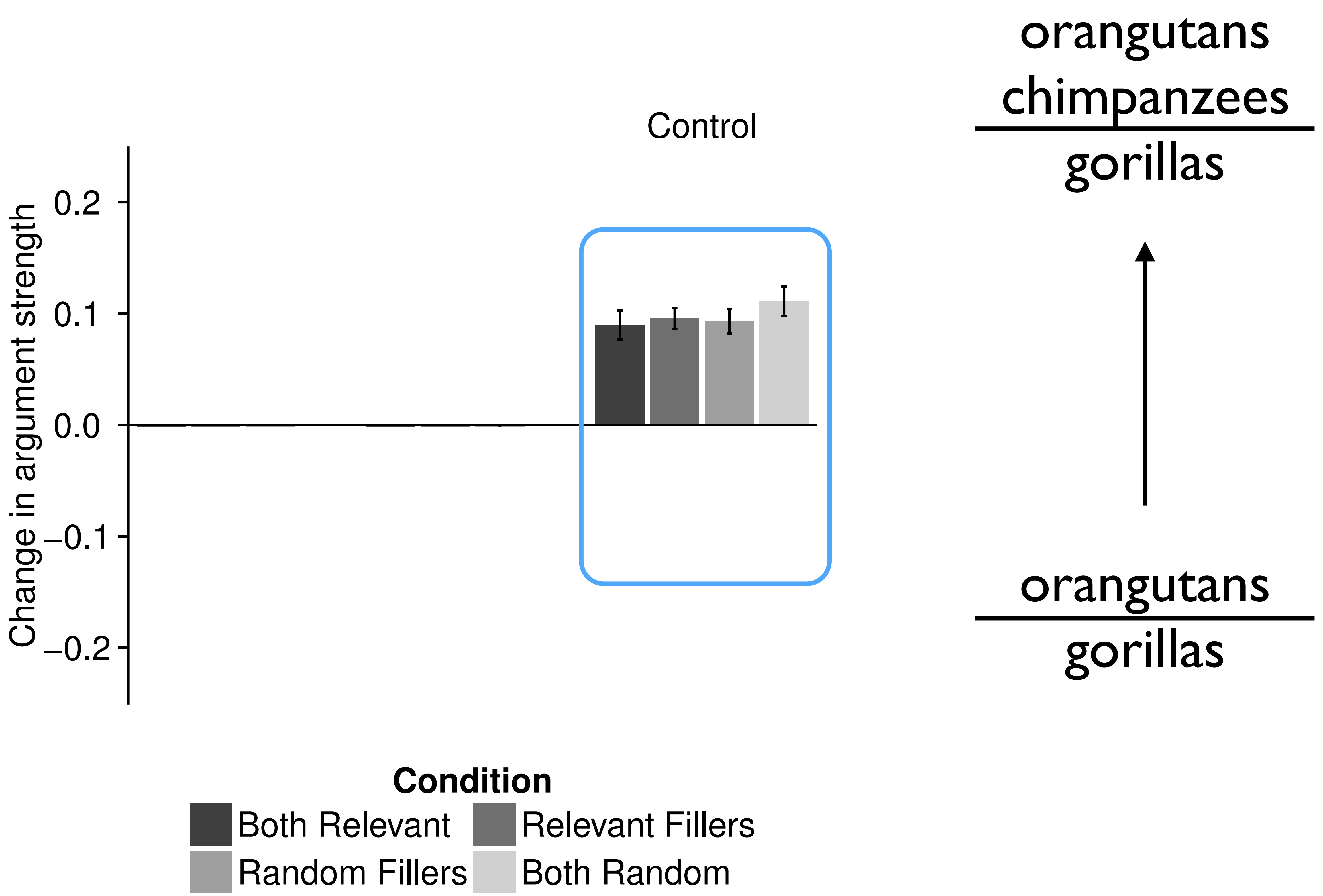
(negated premises
unrelated to the topic
suggest weak sampling)

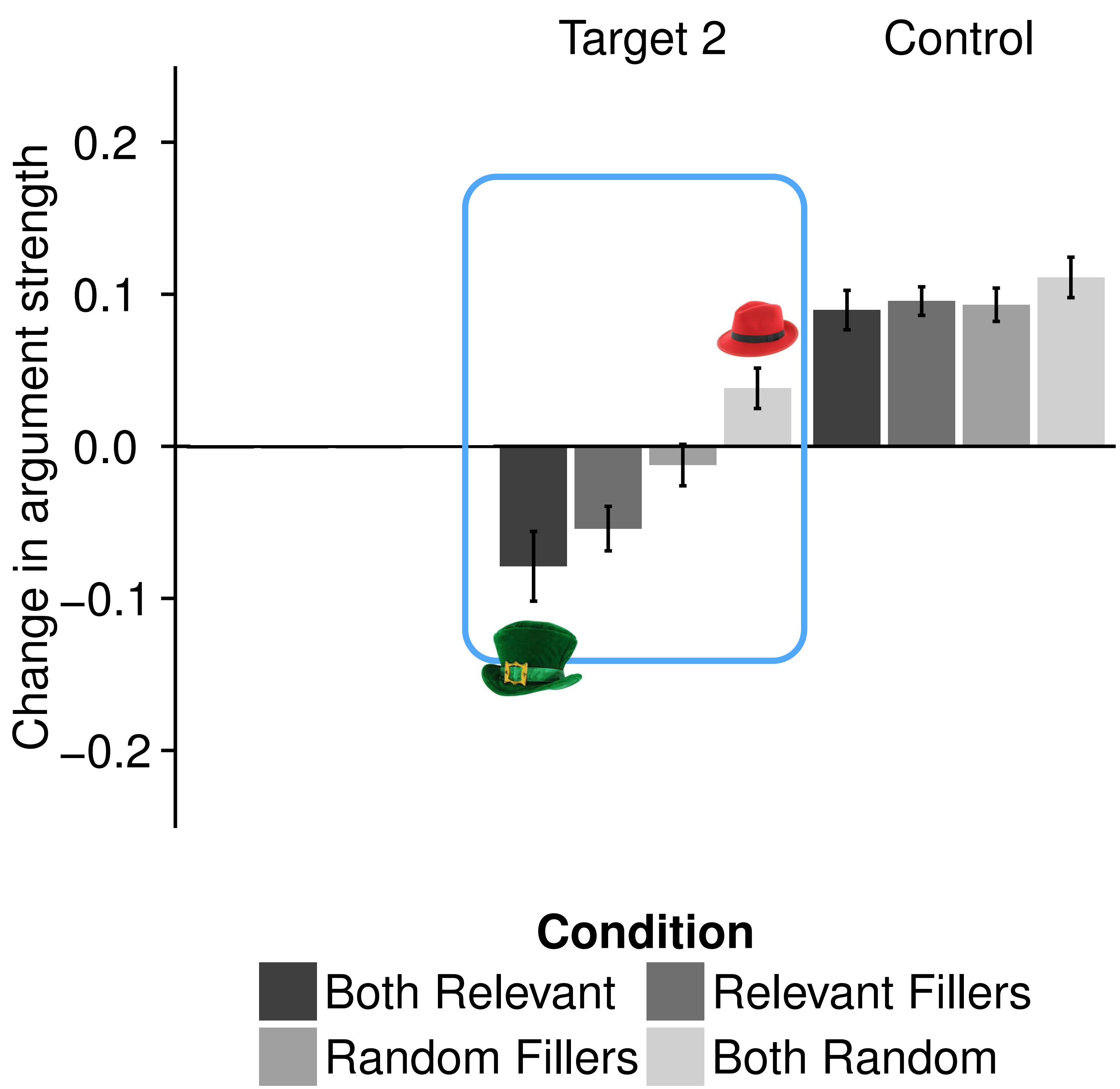
Stimulus ordering was fixed and designed to ensure that fillers (mostly) preceded targets:

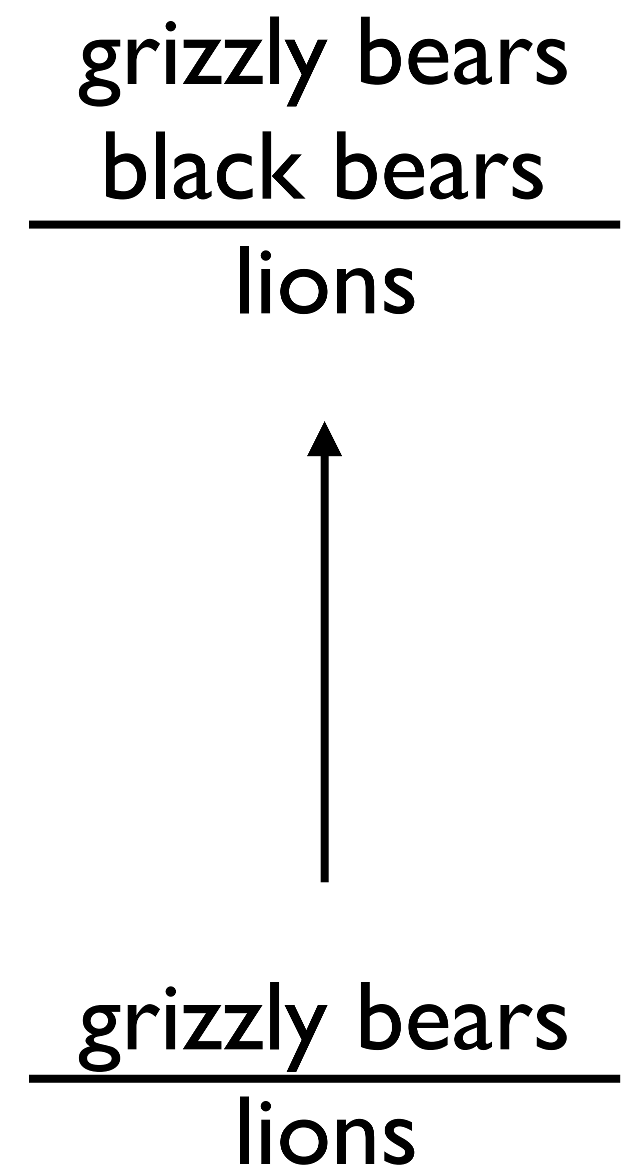
	Additional example	
First generalisation	HELPFUL	RANDOM
EAGLES → DOVES	+HAWKS	−TORTOISES
ELEPHANTS → DEERS	+COWS	+ANTEATERS
TIGERS → FERRETS	+LIONS	+LIONS
KANGAROOS → WOMBATS	+KOALAS	−FLAMINGOS
GRIZZLY BEARS → LIONS	+BLACK BEARS	+BLACK BEARS
ORANGUTANS → GORILLAS	+CHIMPANZEES	+CHIMPANZEES

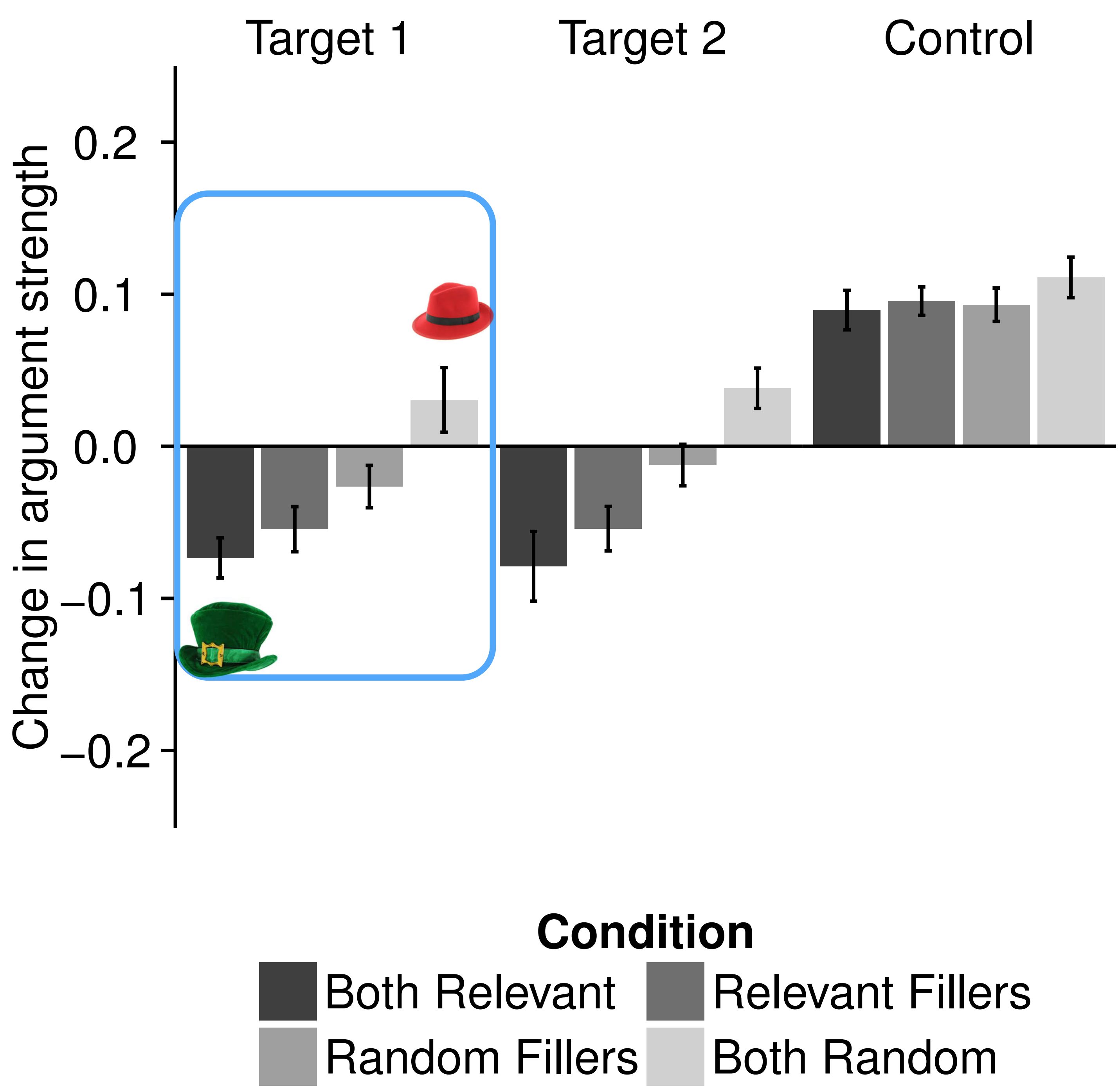
Participants were 296 people recruited through MTurk

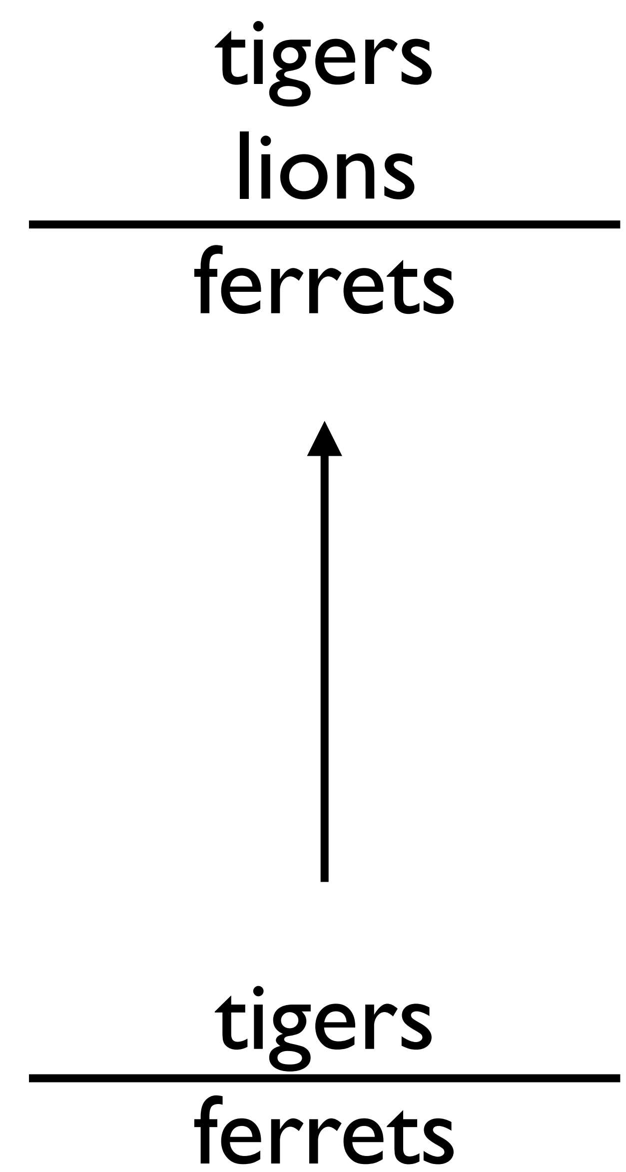

Change in argument strength
0.2
0.1
0.0
–0.1
–0.2
Condition
Both Relevant
Relevant Fillers
Random Fillers
Both Random
orangutans
chimpanzees
gorillas
orangutans
gorillas


Control
Change in argument strength
0.2
0.1
0.0
−0.1
−0.2
Condition
Both Relevant
Relevant Fillers
Random Fillers
Both Random
orangutans
chimpanzees
gorillas
orangutans
gorillas


Target 2
Control
Change in argument strength
0.2
0.1
0.0
−0.1
−0.2
Condition
Both Relevant
Relevant Fillers
Random Fillers
Both Random


grizzly bears
black bears
lions
grizzly bears
lions


Target 1
Target 2
Control
Change in argument strength
0.2
0.1
0.0
−0.1
−0.2
Condition
Both Relevant
Relevant Fillers
Random Fillers
Both Random


tigers
lions
ferrets
tigers
ferrets

(Bayesian) data analysis: hypothesis tests for order restricted models

Model	Order restrictions
NO EFFECT	$\mu_1 = \mu_2 = \mu_3 = \mu_4$
FILLERS ONLY	$\mu_1 = \mu_2 < \mu_3 = \mu_4$
STORY ONLY	$\mu_1 < \mu_2 = \mu_3 < \mu_4$
BOTH	$\mu_1 < \mu_2 < \mu_3 < \mu_4$
RANDOM EFFECT	$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$

Clear effect of cover story on targets, possibly also an effect of filler type

		Bayes Factor (: NO EFFECT)	
Model	Order restrictions	Target 1	Target 2
NO EFFECT	$\mu_1 = \mu_2 = \mu_3 = \mu_4$	-	-
FILLERS ONLY	$\mu_1 = \mu_2 < \mu_3 = \mu_4$	740:1	12,000:1
STORY ONLY	$\mu_1 < \mu_2 = \mu_3 < \mu_4$	4,100:1	17,000:1
BOTH	$\mu_1 < \mu_2 < \mu_3 < \mu_4$	2,900:1	30,000:1
RANDOM EFFECT	$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$	520:1	4,600:1

Null effect for the control item

		Bayes Factor (: NO EFFECT)
Model	Order restrictions	Control
NO EFFECT	$\mu_1 = \mu_2 = \mu_3 = \mu_4$	-
FILLERS ONLY	$\mu_1 = \mu_2 < \mu_3 = \mu_4$	$< 1 : 1$
STORY ONLY	$\mu_1 < \mu_2 = \mu_3 < \mu_4$	$< 1 : 1$
BOTH	$\mu_1 < \mu_2 < \mu_3 < \mu_4$	$< 1 : 1$
RANDOM EFFECT	$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$	$< 1 : 1$

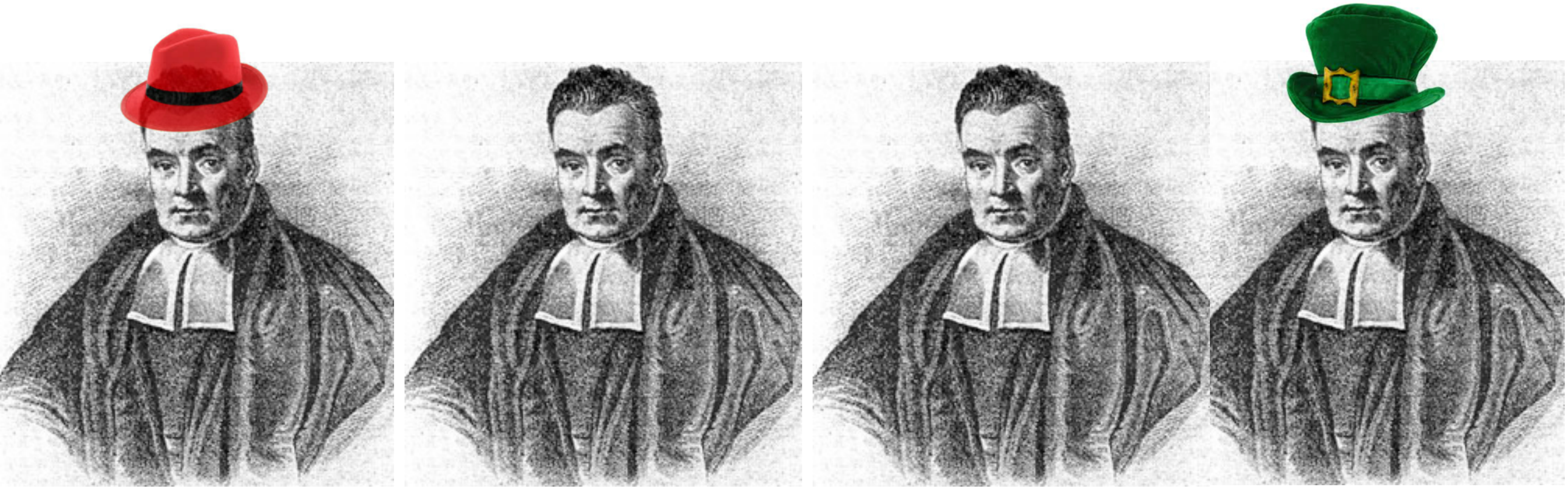
Can we accommodate this pattern using Bayesian models?

Should we model this as a difference between two Bayesian learners?

A weakly sampling falsificationist

A strongly sampling Ockhamist

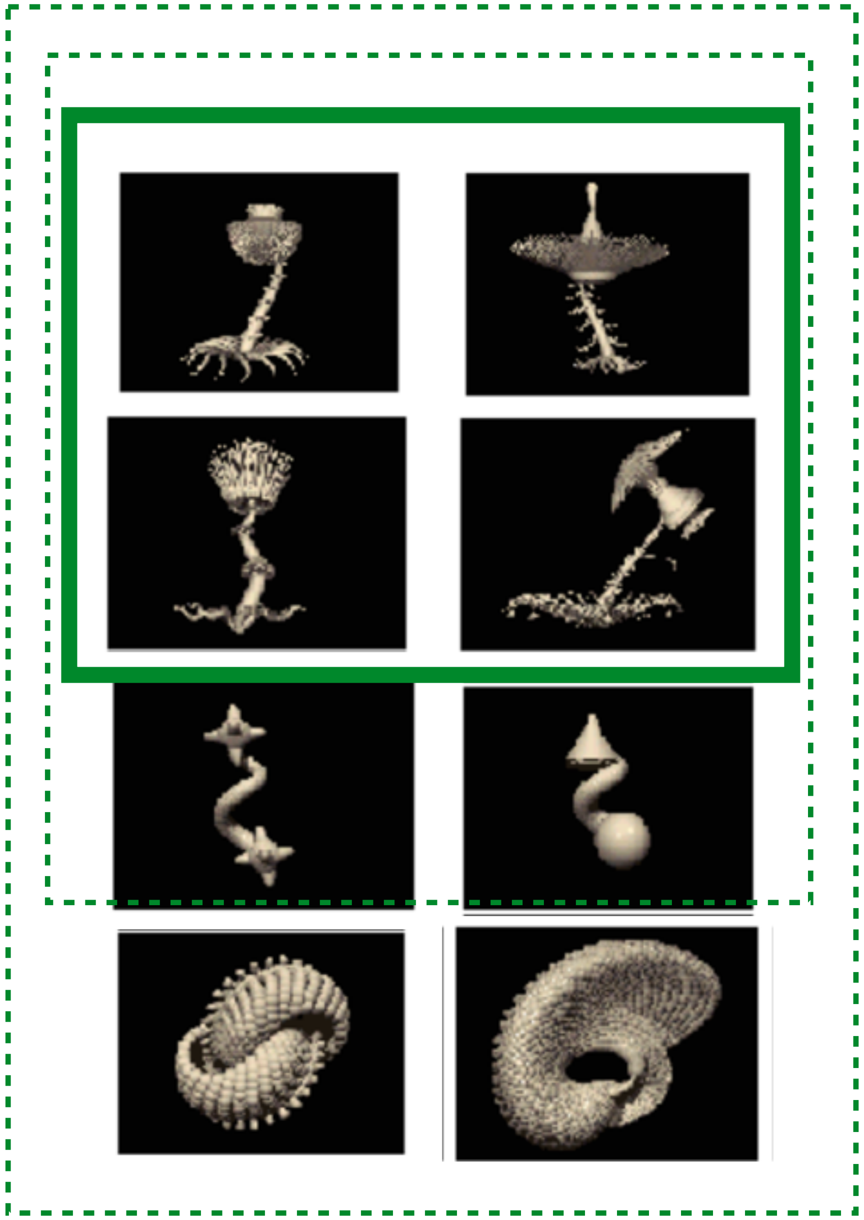
Or posit a continuum of Bayesians?



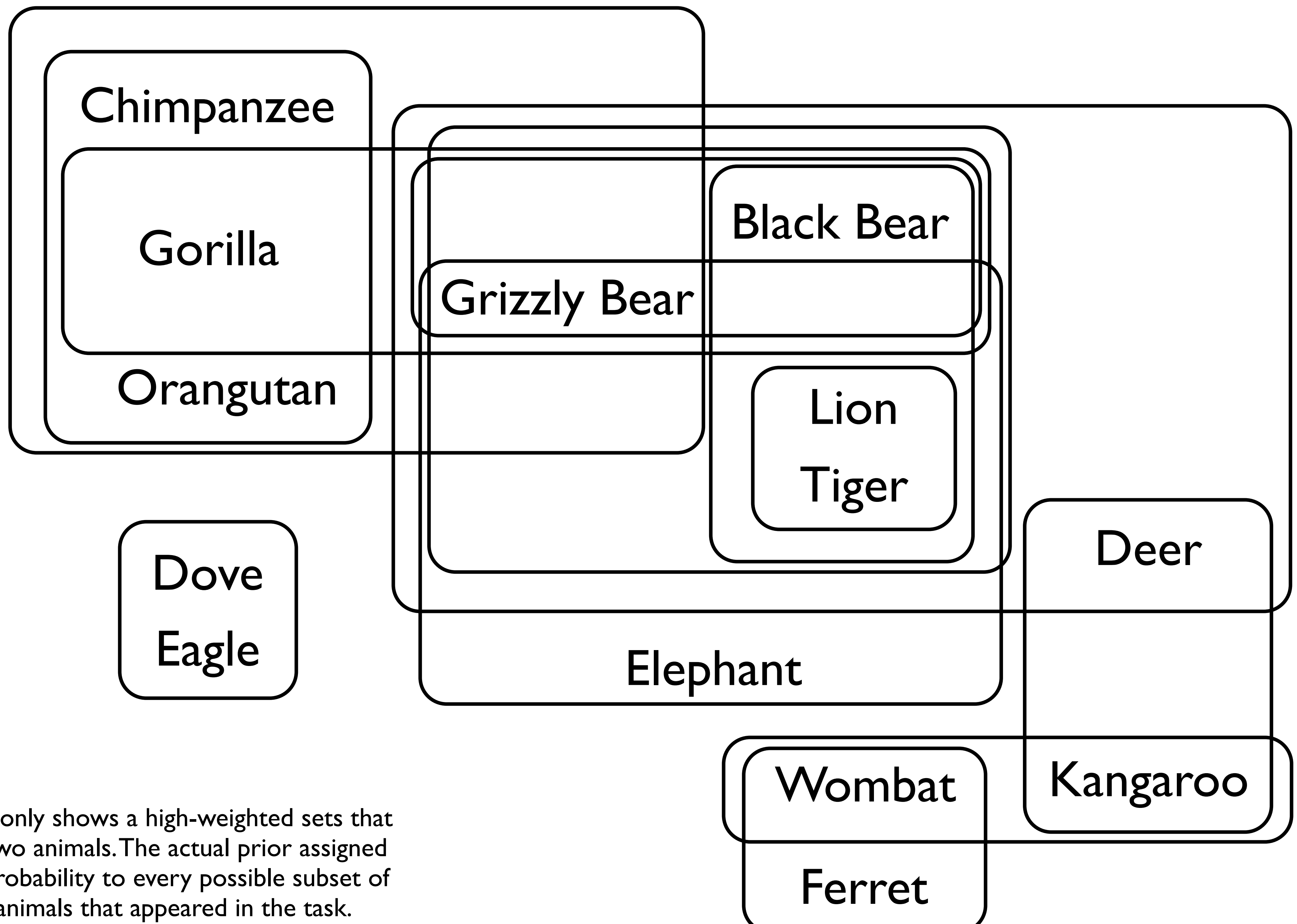
θ=0 θ=.33 θ=.67 θ=1

"weak" "strong"

And what shall our Bayesians use for their hypotheses?

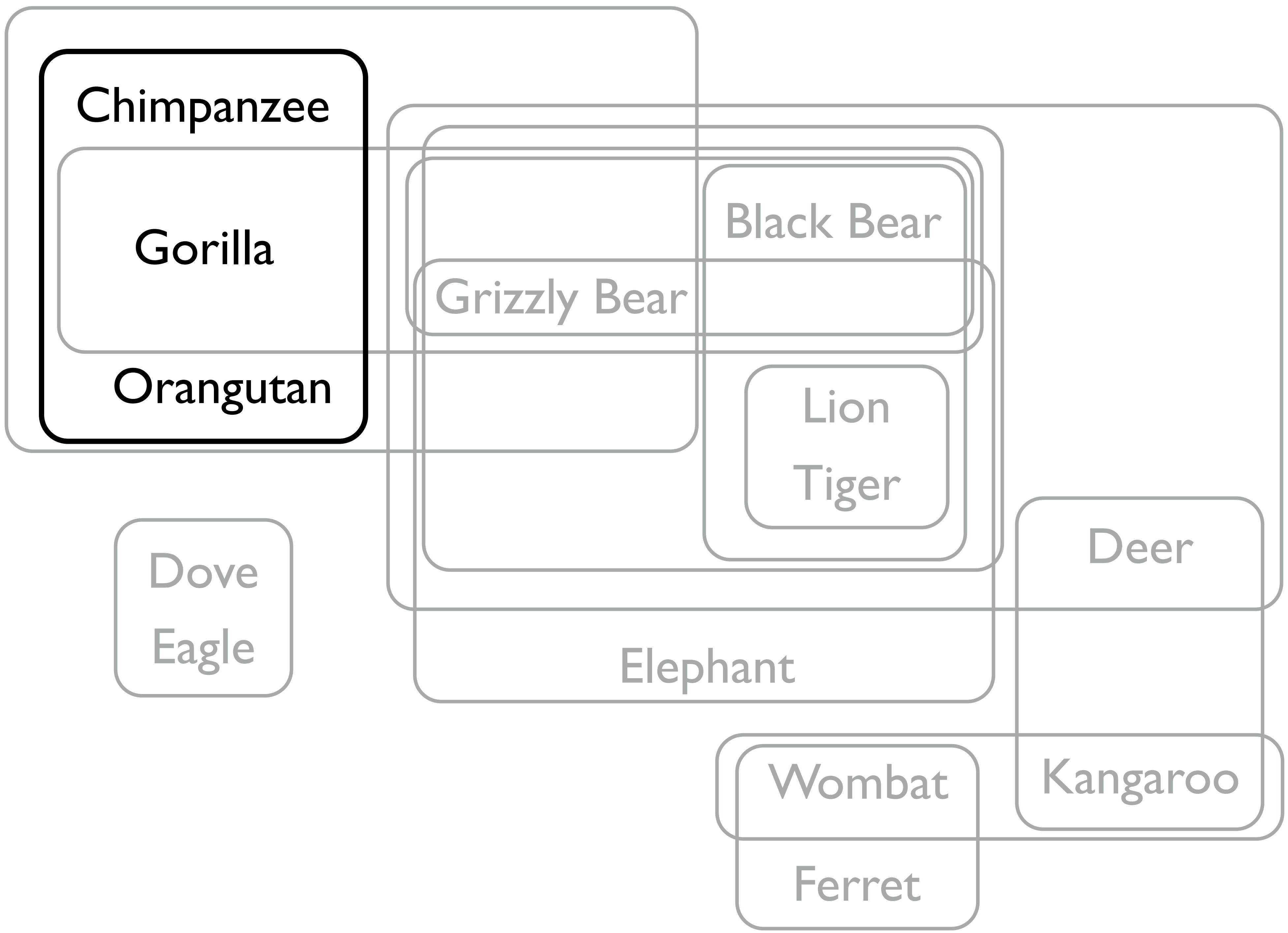


Hypotheses inferred from a separate data set

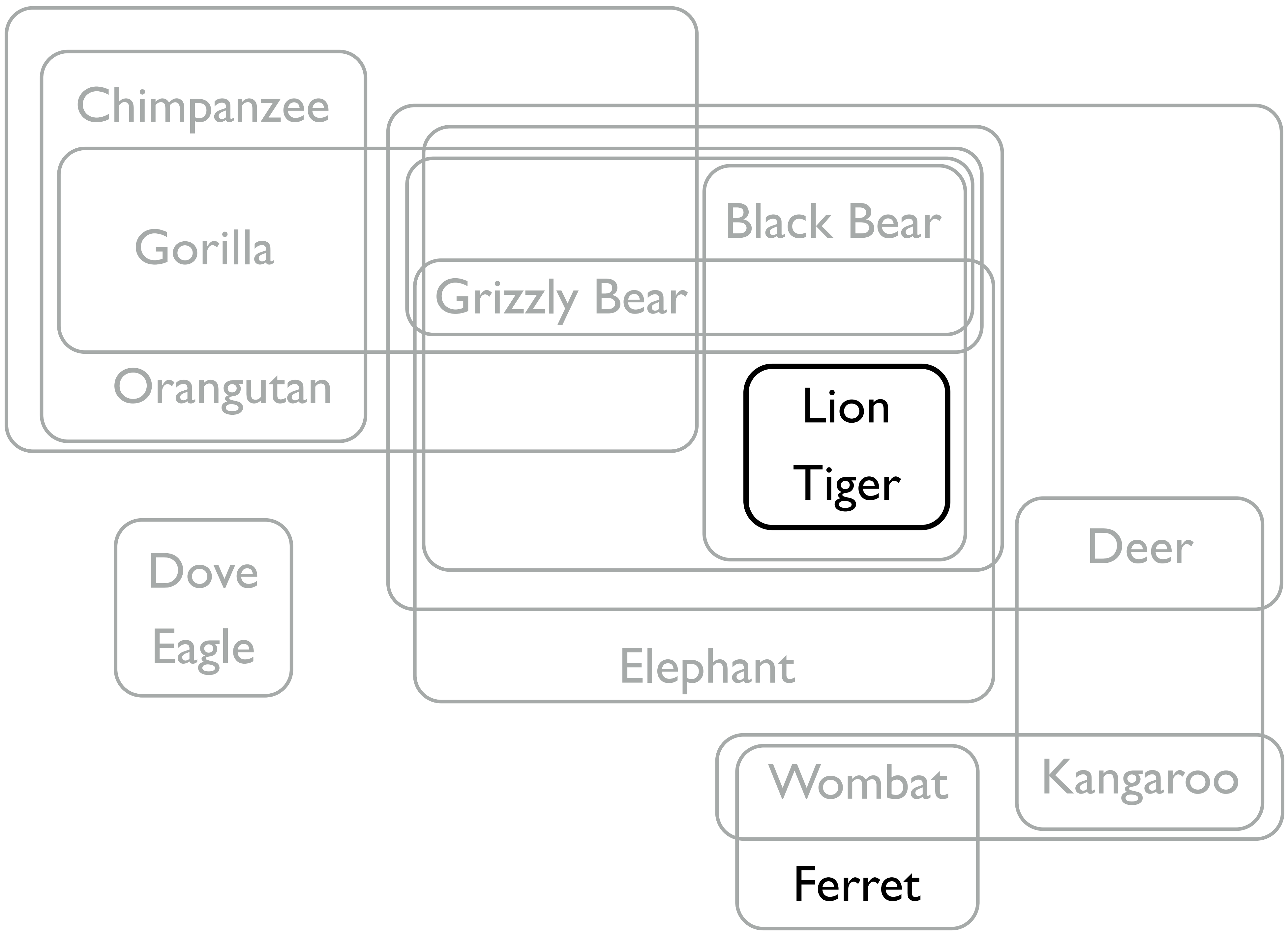


*This illustration only shows a high-weighted sets that contain at least two animals. The actual prior assigned non-zero prior probability to every possible subset of the set of all animals that appeared in the task. Qualitative features of model predictions are robust to the specific choice of prior: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

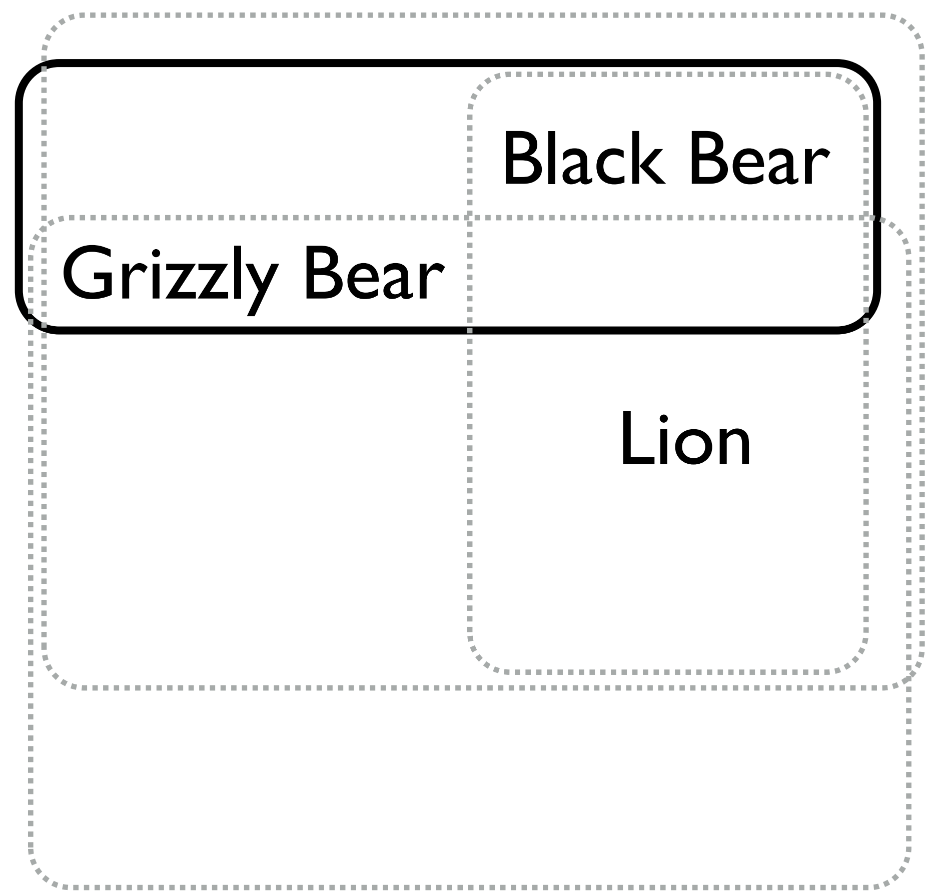
(Chimpanzee, Gorilla, Orangutan)



(Lions, Tigers) but not Ferrets

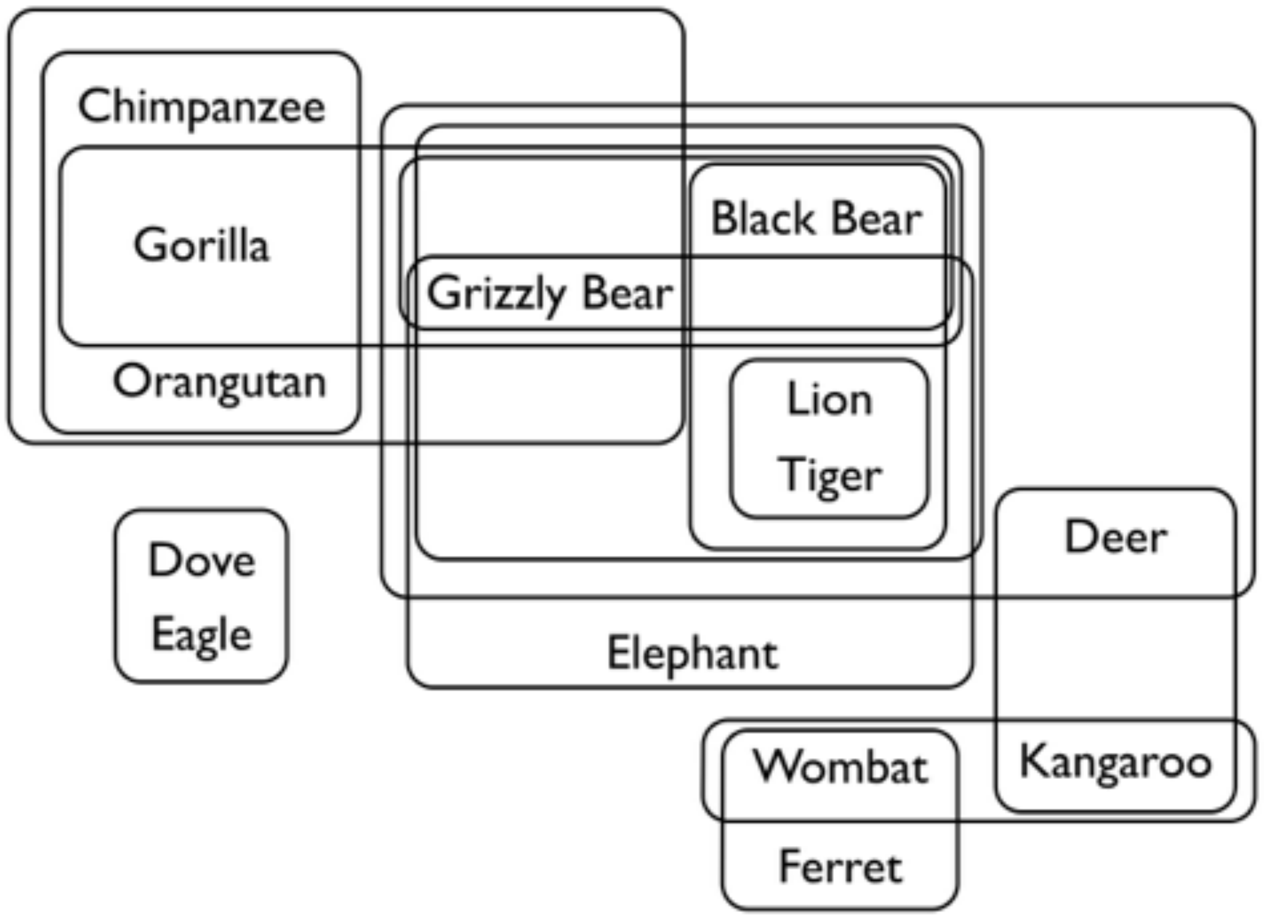


(Grizzly Bears, Black Bears) but not Lions?



There are many high weighted features involving these three, but overall the prior puts the bears together more often

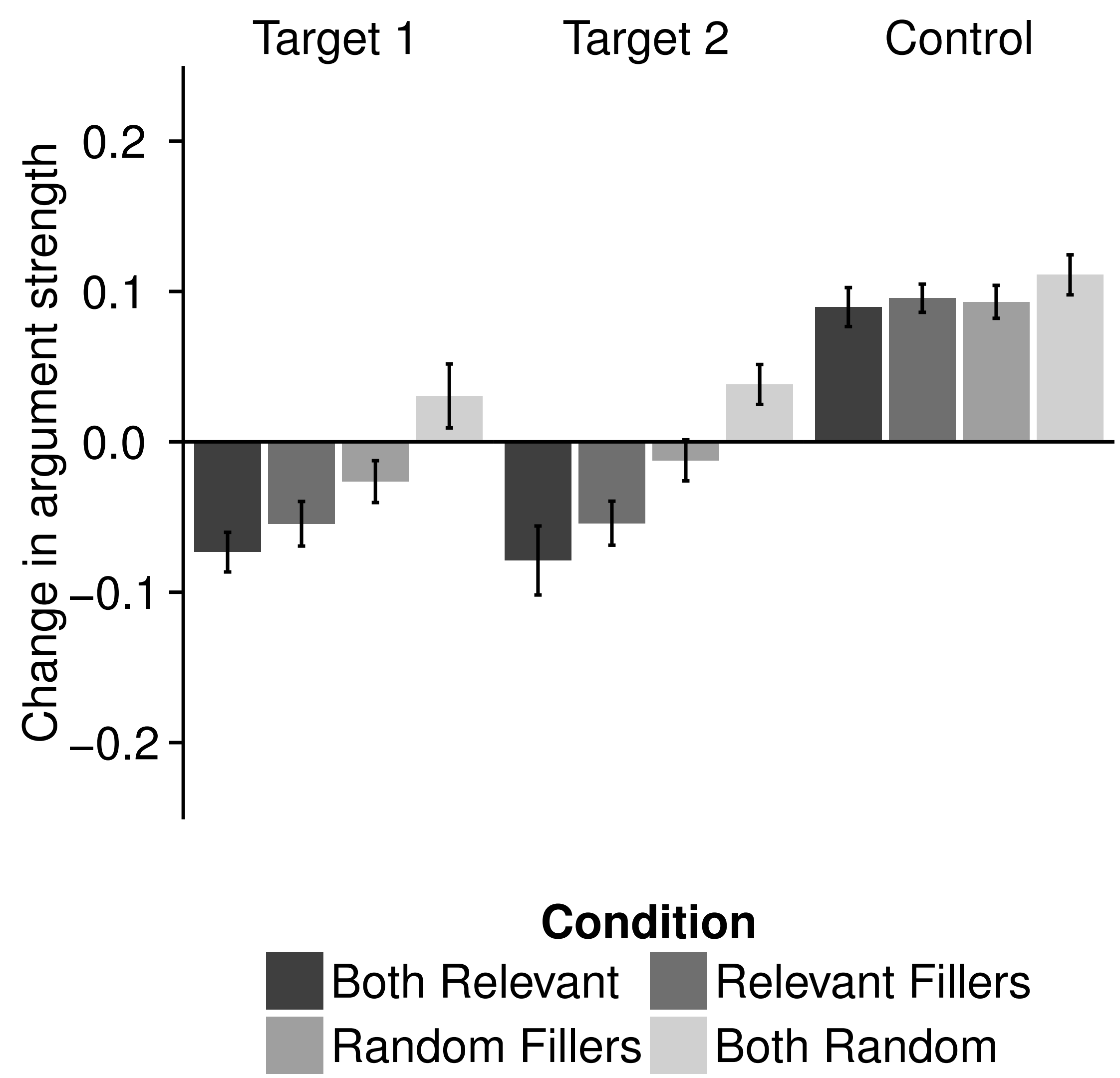
The prior explains why there are structural differences between the targets and the control



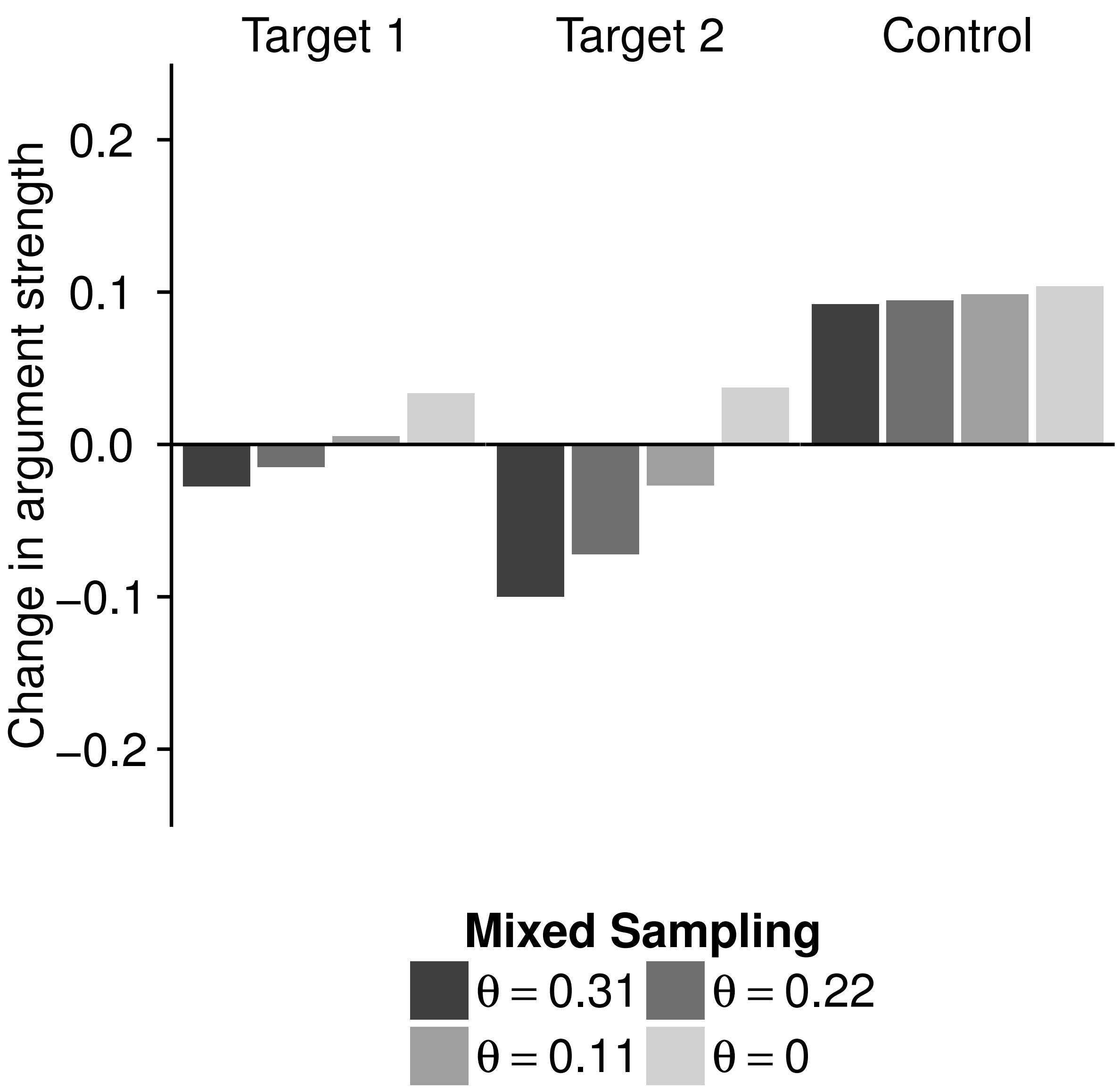
The likelihood describes how “adding more premises” can have different effects across conditions

Does the model work???

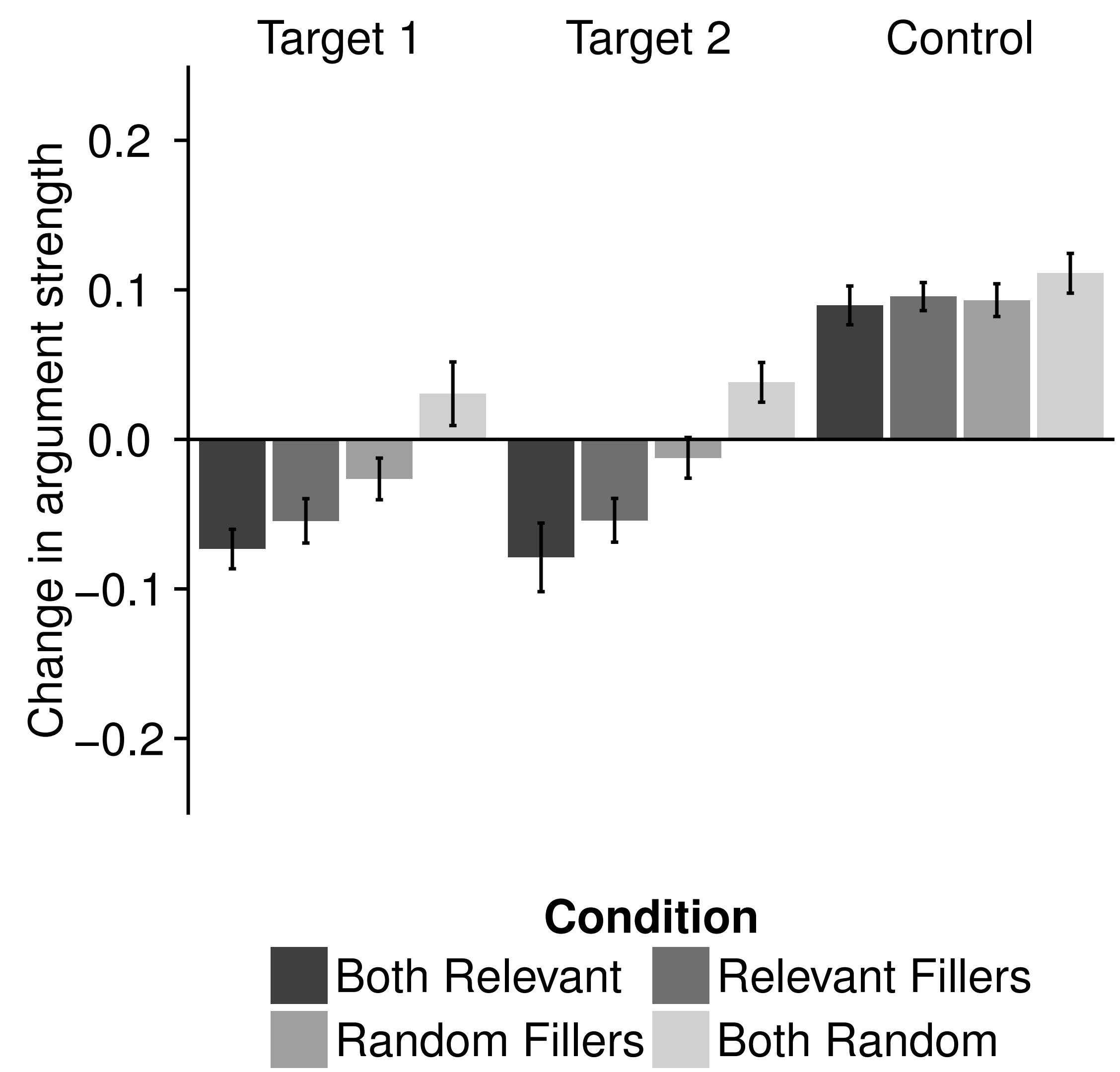
Empirical data



Yep.



Empirical data



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

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

- This is encouraging, so…

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

Navarro, Dry & Lee (2012). Sampling assumptions in inductive generalization. *Cognitive Science*

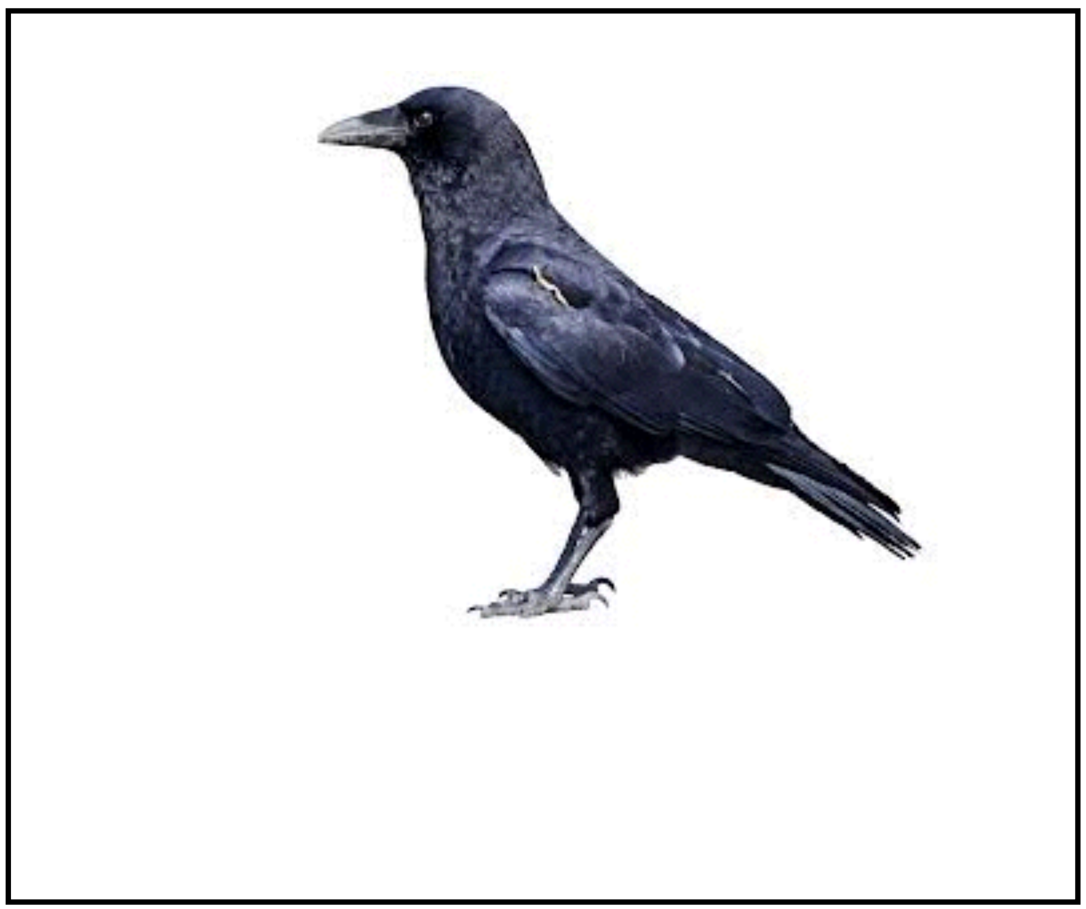
Hendrickson, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations

Ransom, Perfors & Navarro (in press). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*

Voorspoels, Navarro, Perfors, Ransom & Storms (under revision). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*

Perfors, Navarro & Shafto (in preparation). Stronger evidence isn't always better: The role of social inference in evidence selection and interpretation. Previously rejected from *JEP:G*

You want to infer whether all ravens are black.
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?

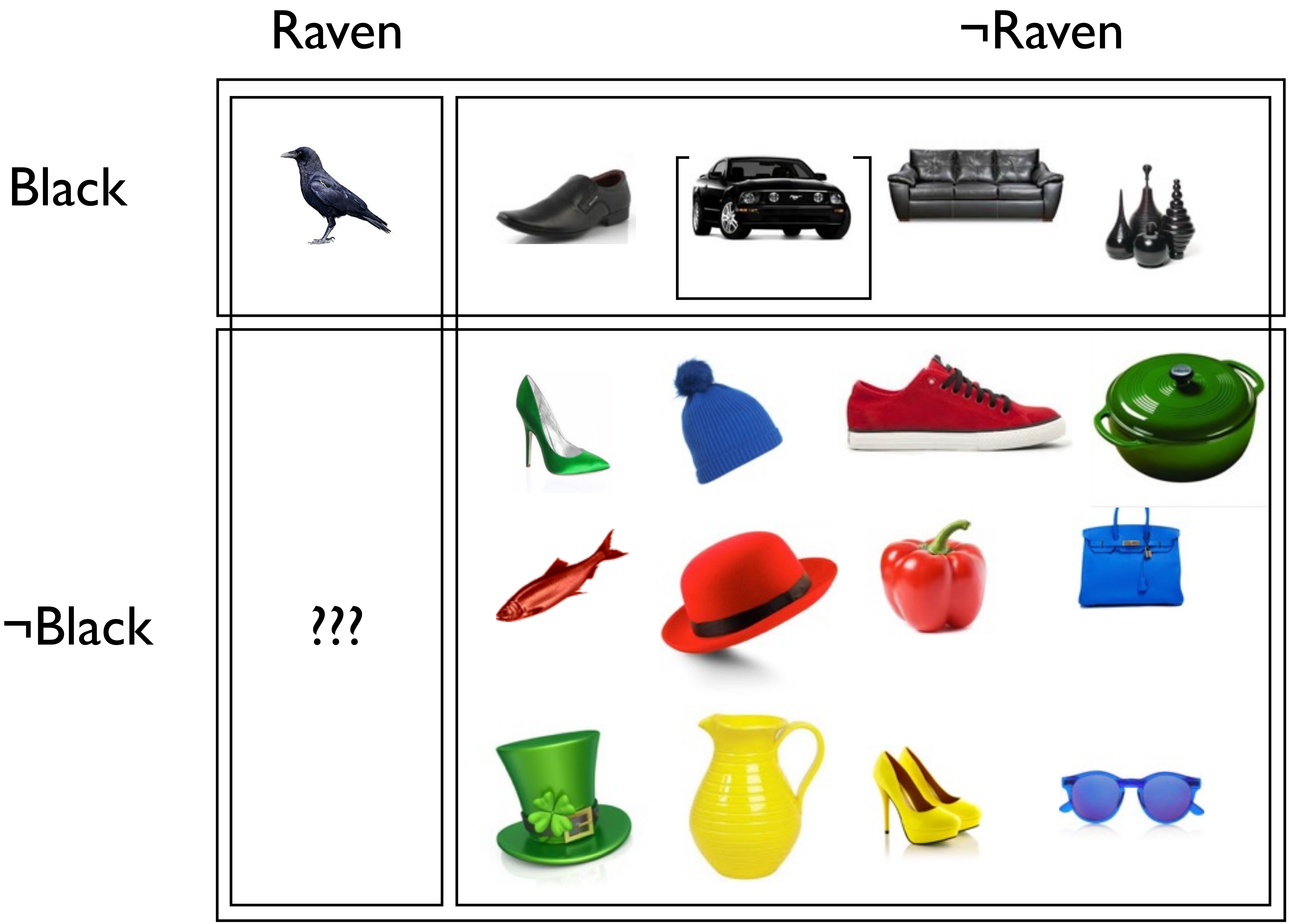
$\text{Raven} \rightarrow \text{Black}$ $\neg\text{Black} \rightarrow \neg\text{Raven}$

(raven, black)

(¬black, ¬raven)

(Hempel's paradox)

	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!

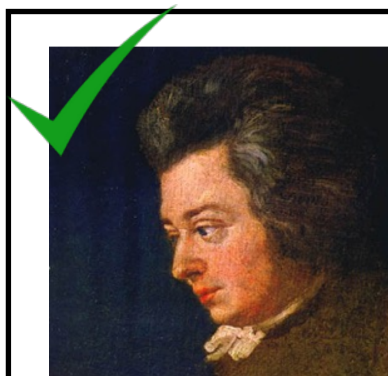
A black raven is very informative

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

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

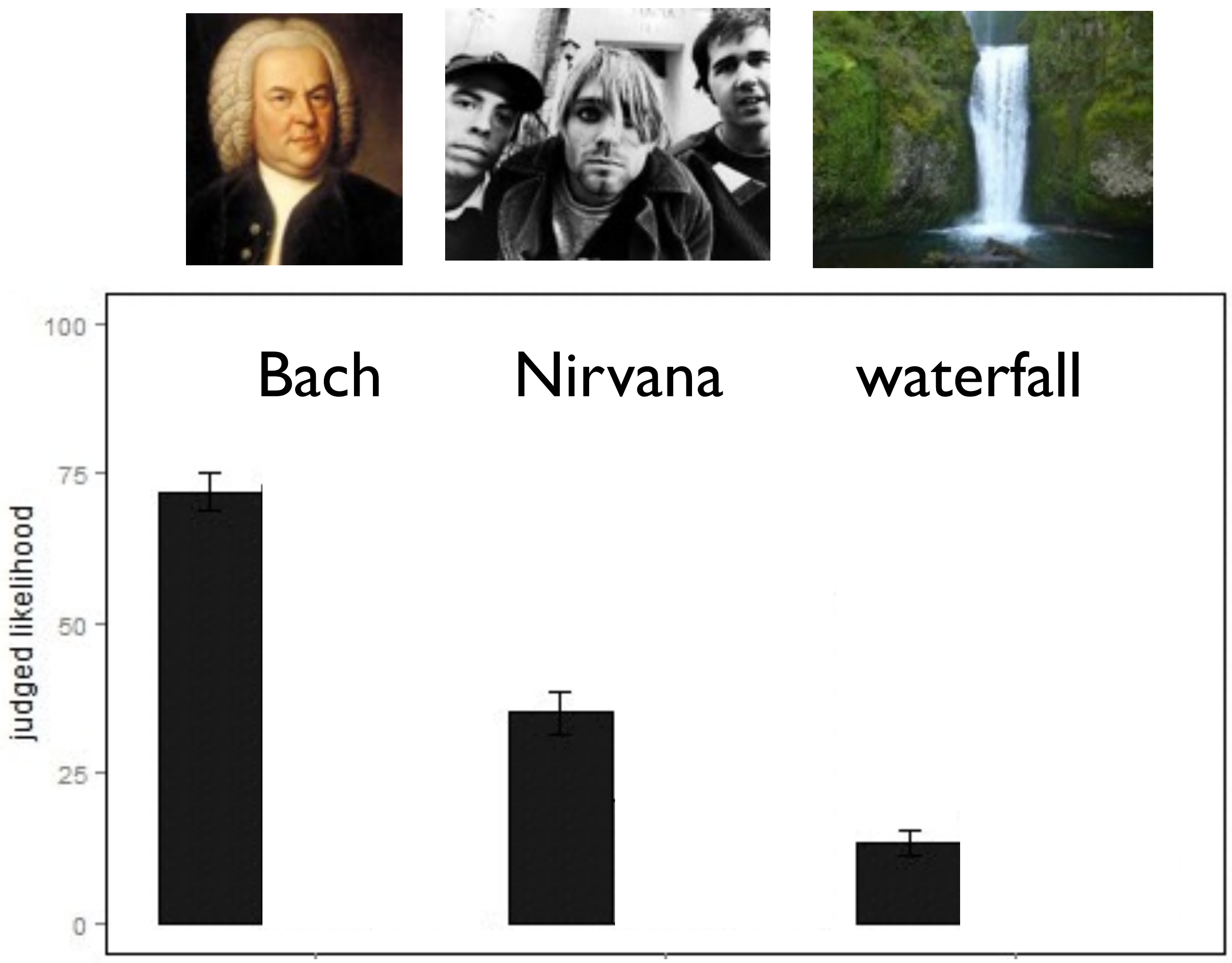
Mozart produces alpha waves

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

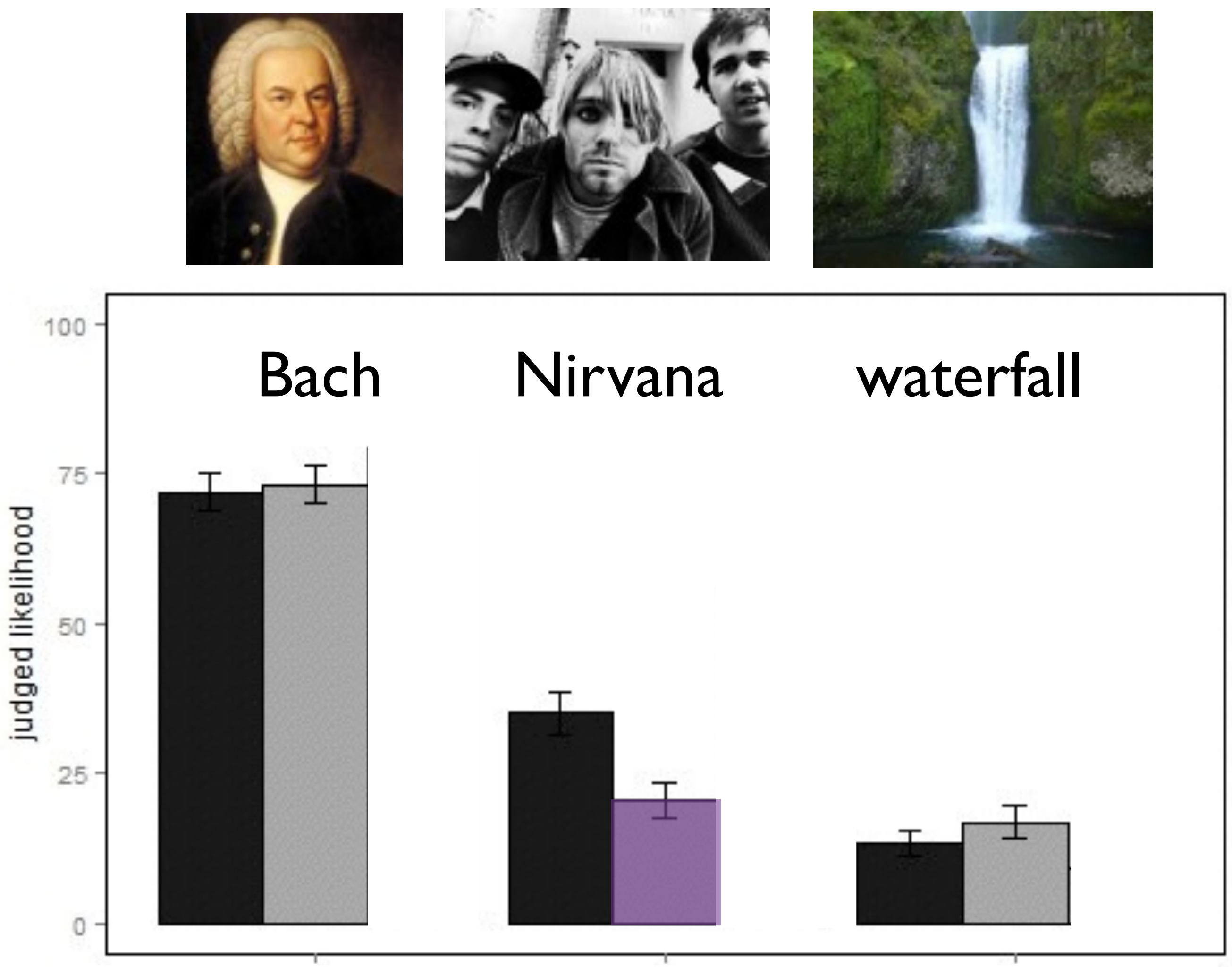
Okay, we start by telling people that Mozart does produce alpha waves…

+Mozart



+Mozart

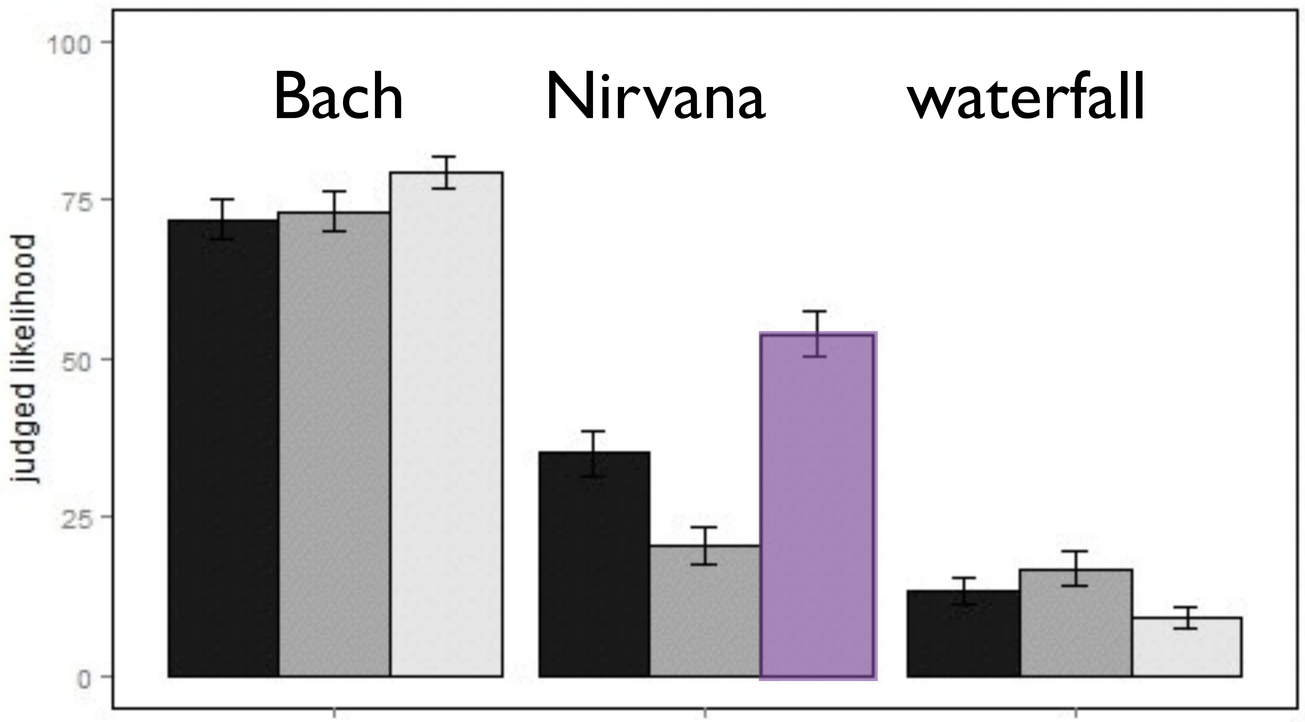
… and they generalise in a way that seems terribly sensible



+Mozart

-Metallica

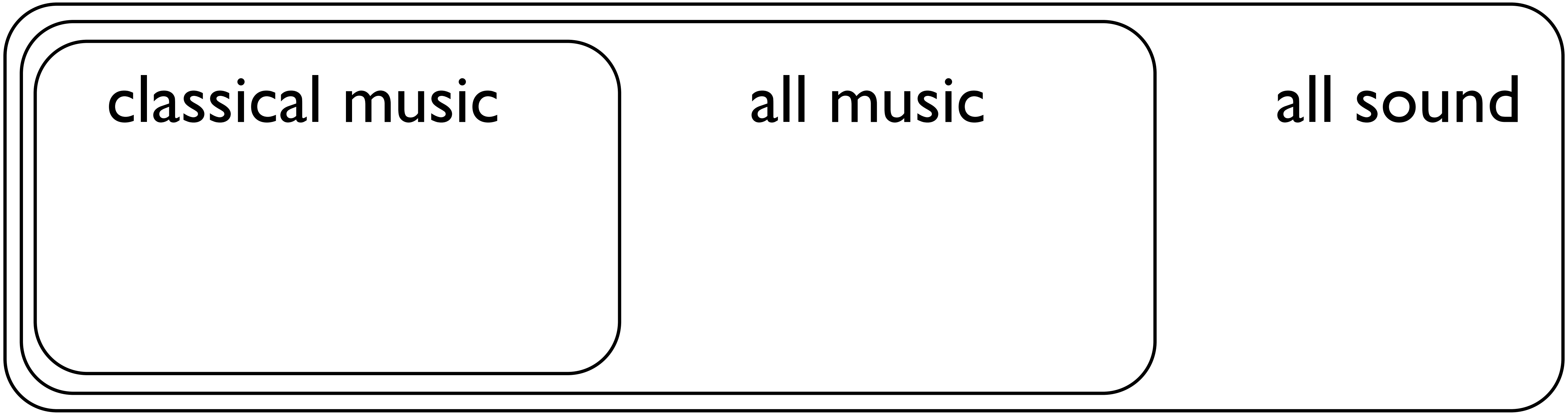
Adding Metallica as a negative example has a modest, sensible effect on inferences about Nirvana



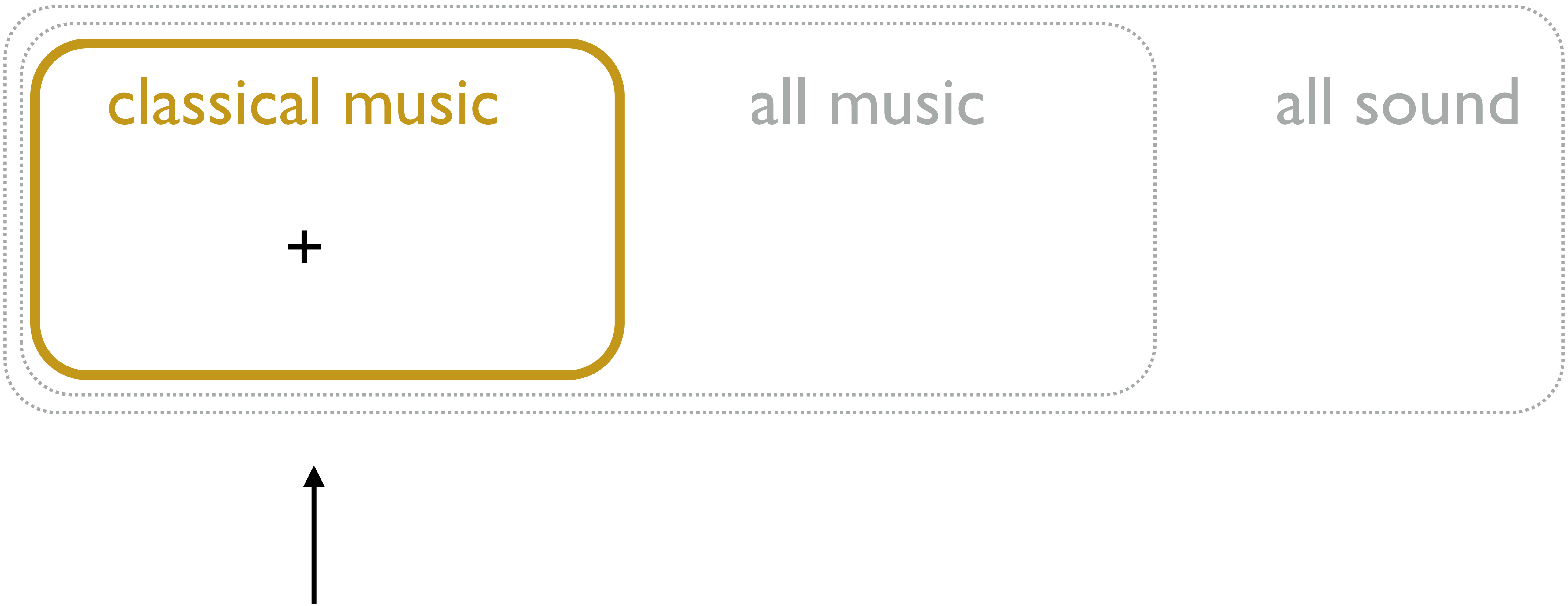
+Mozart

-Falling rock

sigh.



three relevant hypotheses for the extension of the alpha waves property

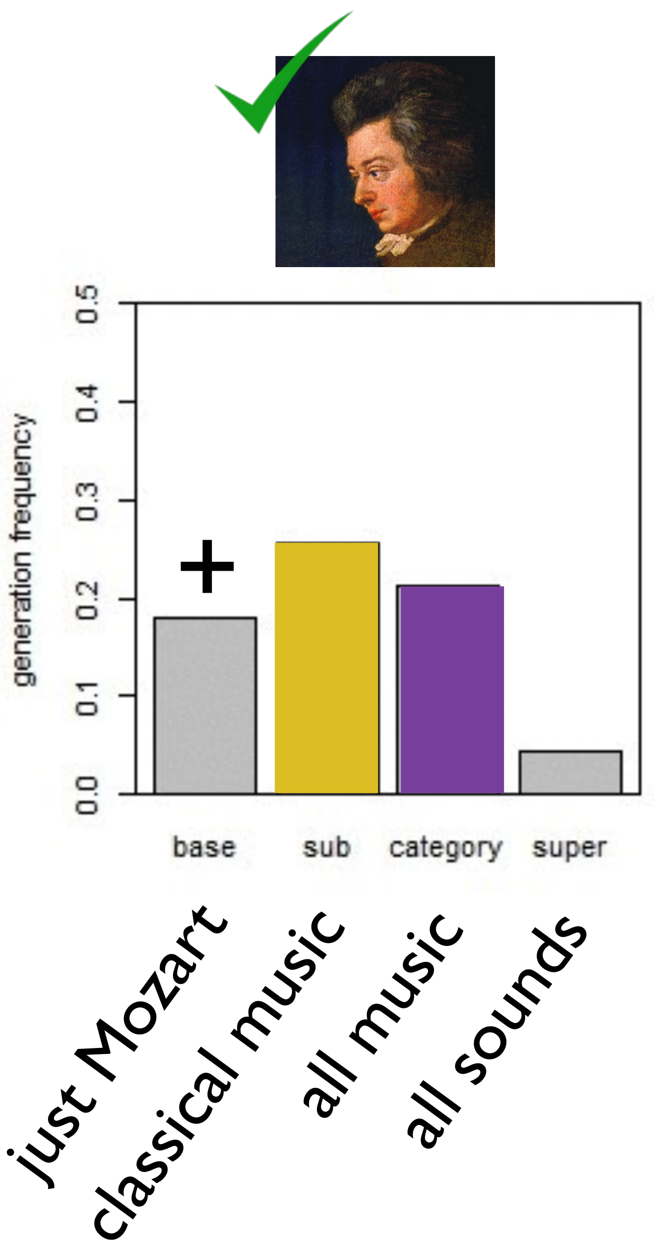


positive example of classical music means people strongly endorse the narrow category

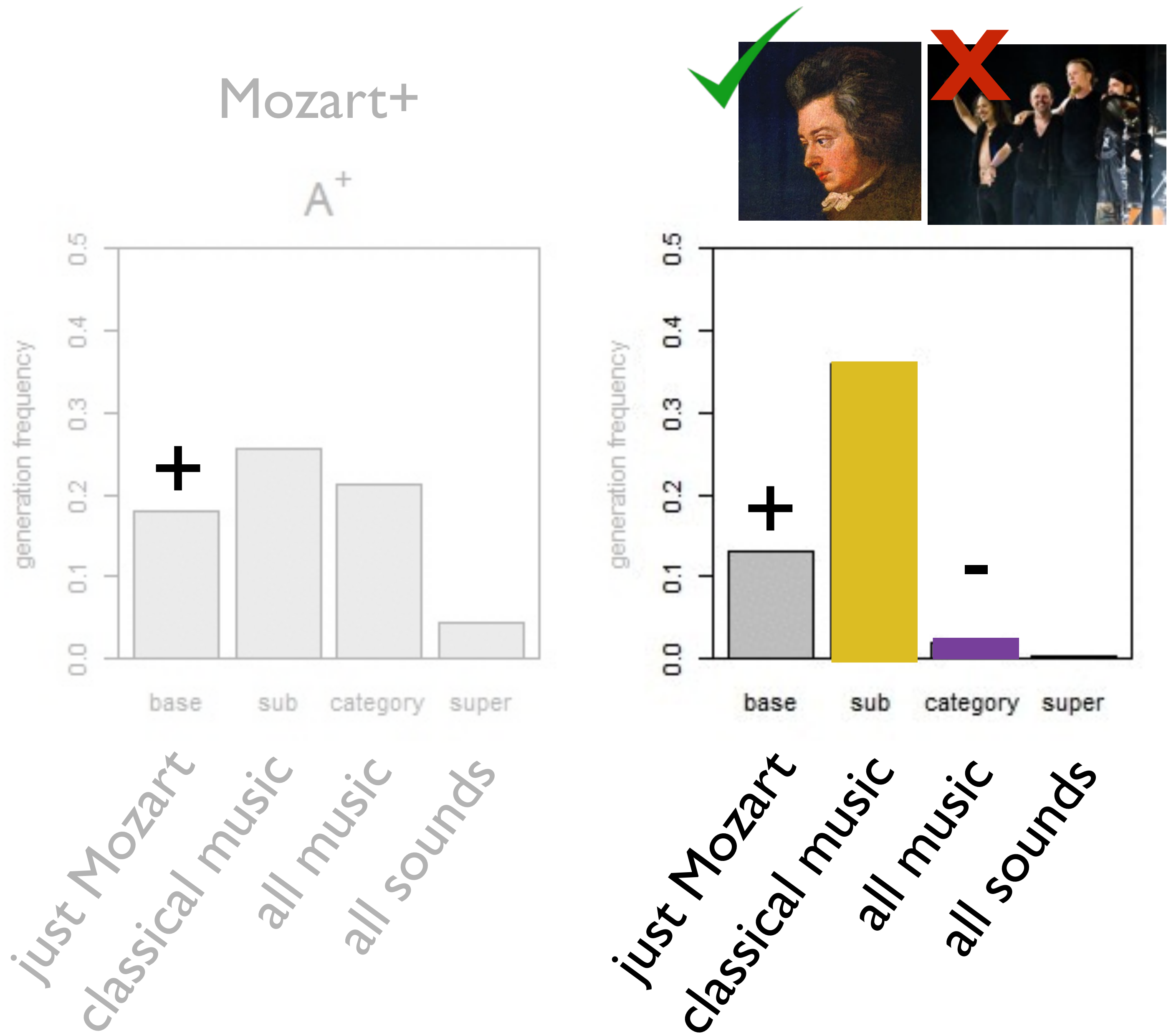
but add a negative observation from a distant category and you get a huge belief revision?

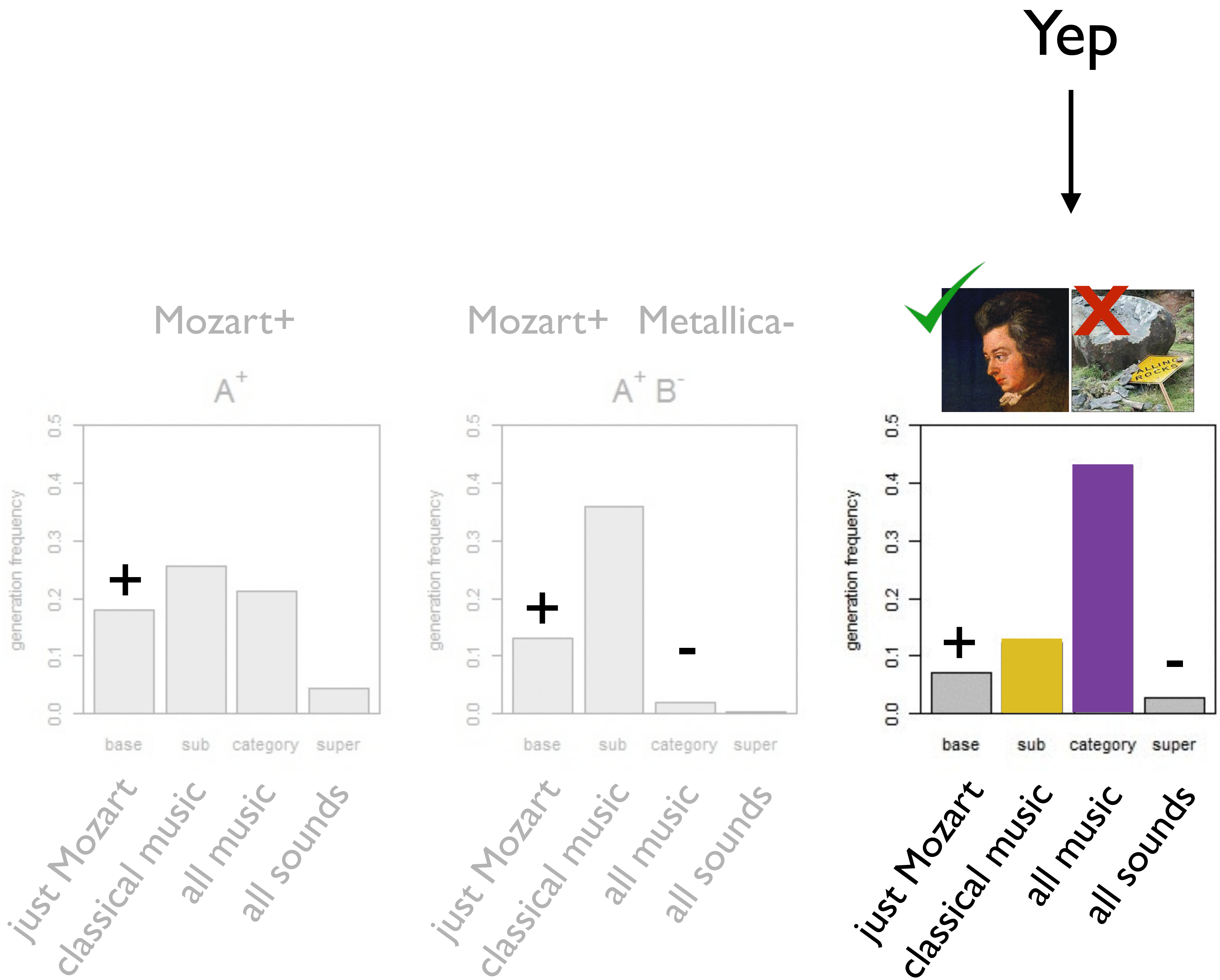
Apparently people make a (pragmatic?) inference that the negative observation is used to demarcate the category boundary

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

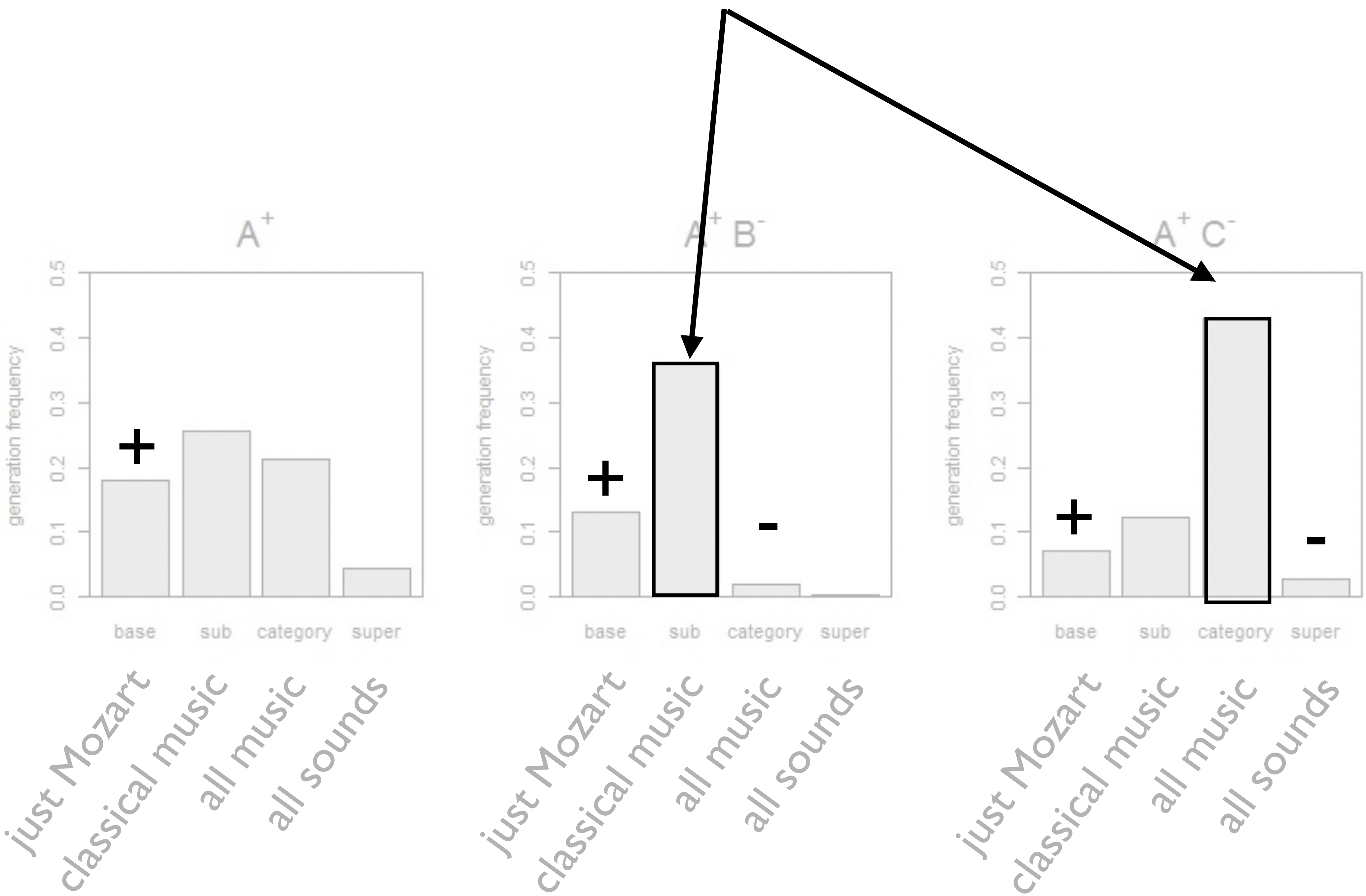


Well, let's ask them what they think the true extension of the property is…




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

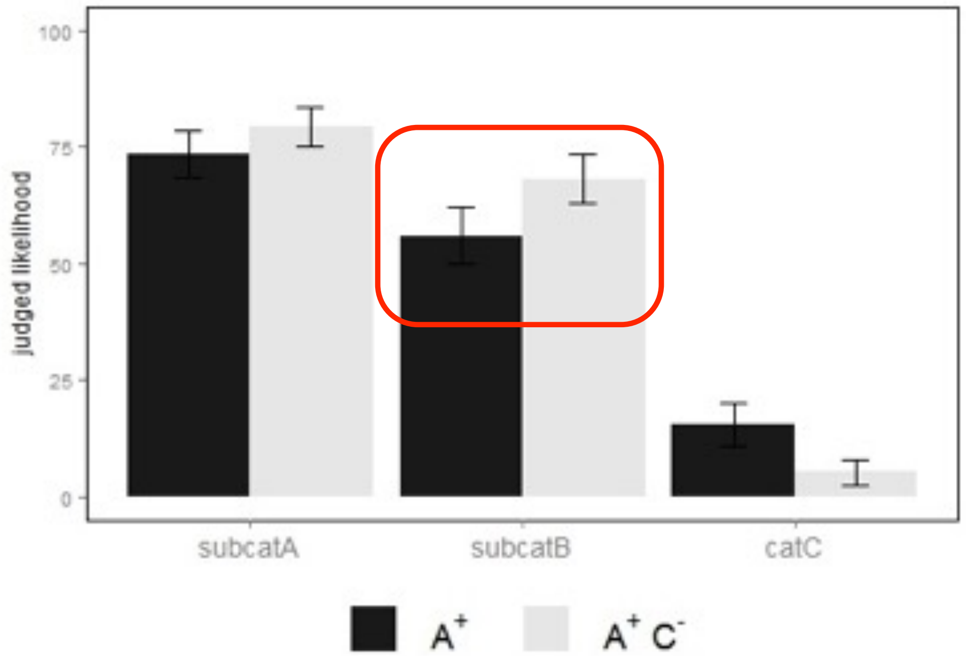
(aside: the actual experiment used many different arguments)

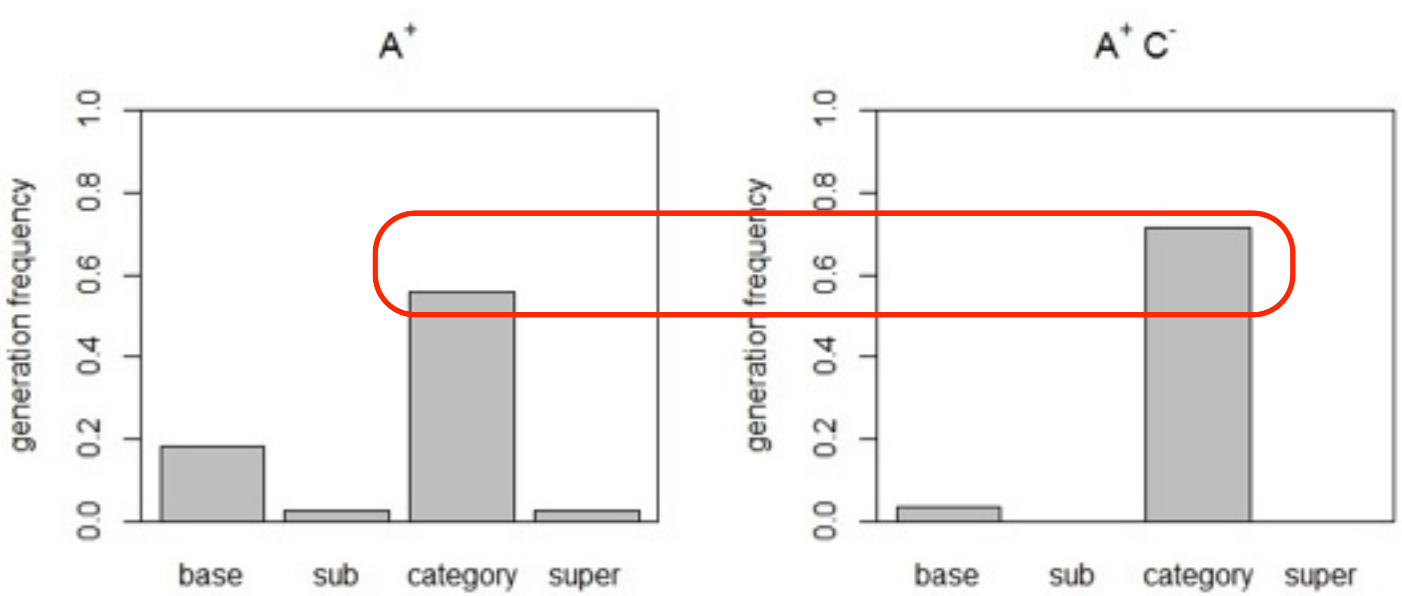
topic	premises			conclusions		
	subcat A	subcat B	cat C	A-member	B-member	C-member
MUSIC	Mozart	Metallica	falling rock	Bach	Nirvana	waterfall
PAINTERS	Rubens	Dahli	woodcarver	Van Eyck	Warhol	sculpturer
PUBLIC FIGURES	actors	librarians	moles	politicians	programmers	pheasants
SHIPS	freight ships	hovercrafts	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

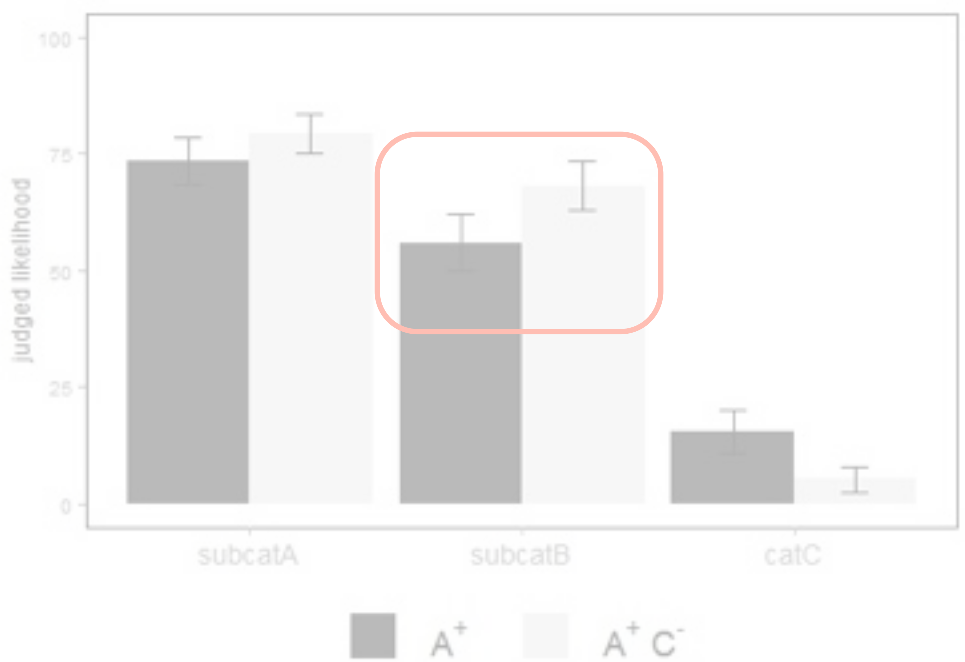
topic	premises		conclusions		
	subcat A	cat C	A-member	B-member	C-member
MAMMALS	dog (+)	magpie (-)	wolf	donkey	blackbird
BIRDS	crow (+)	tuna fish (-)	raven	swan	halibot
FISH	salmon (+)	lizzard (-)	codfish	goldfish	snake
INSECTS	bee (+)	sparrow (-)	ant	cricket	pigeon

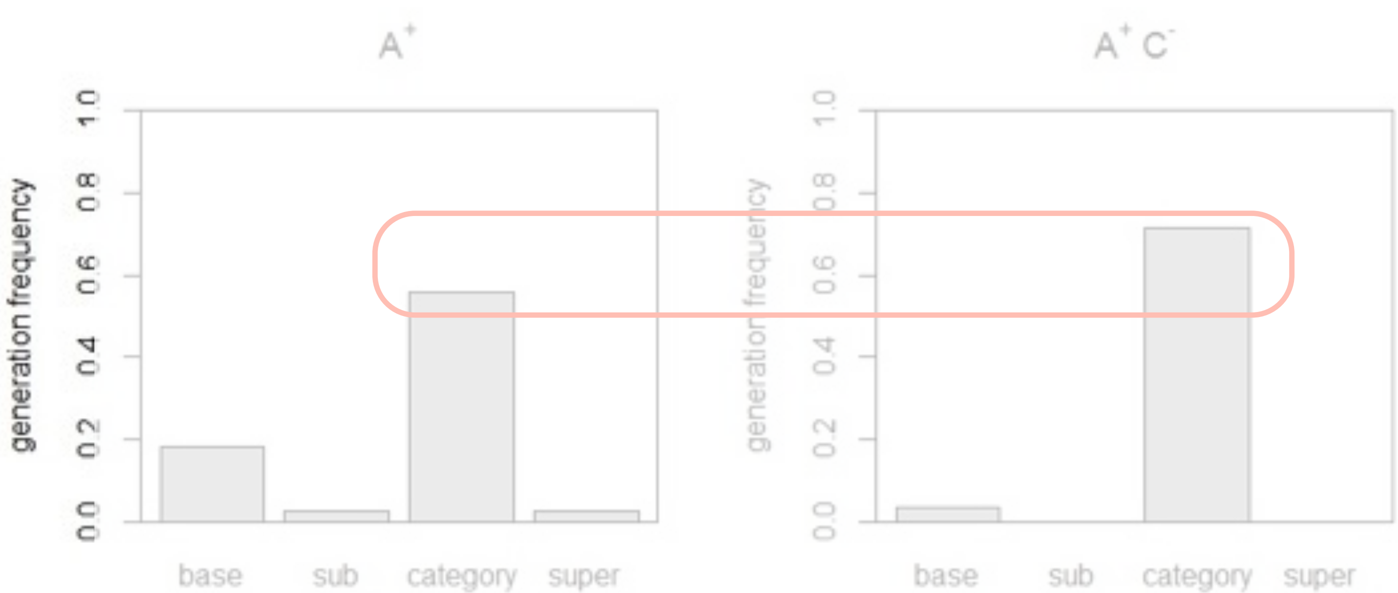
(and yes, the replication worked)





(and yes, the replication worked)

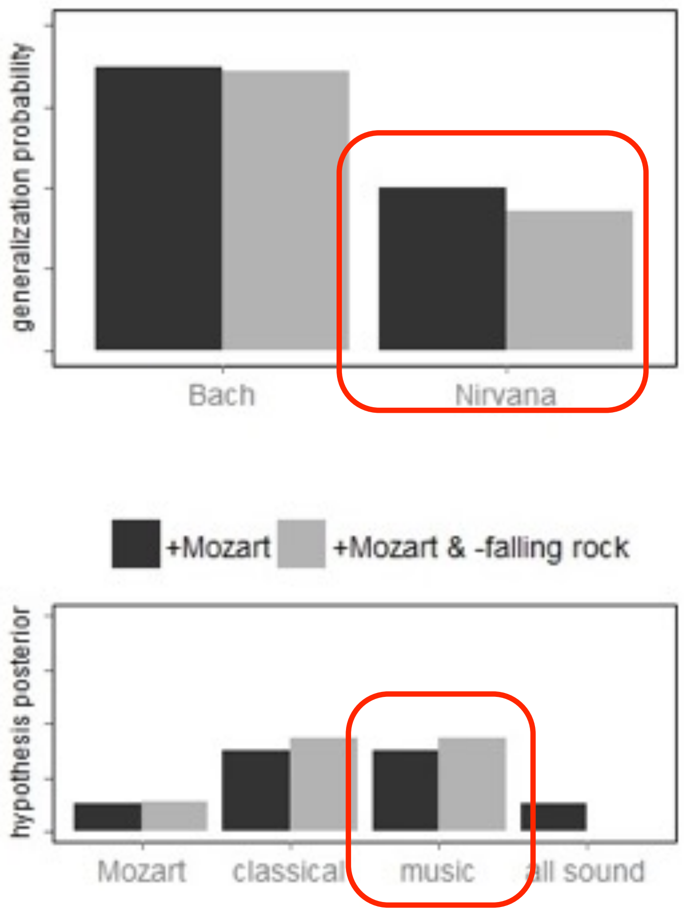




The big question is <u>how</u> to account for the results…

Does the weak sampling model capture the effect?

Weak sampling

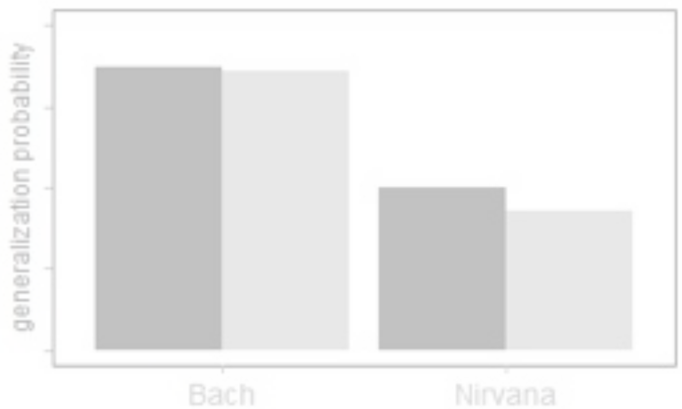


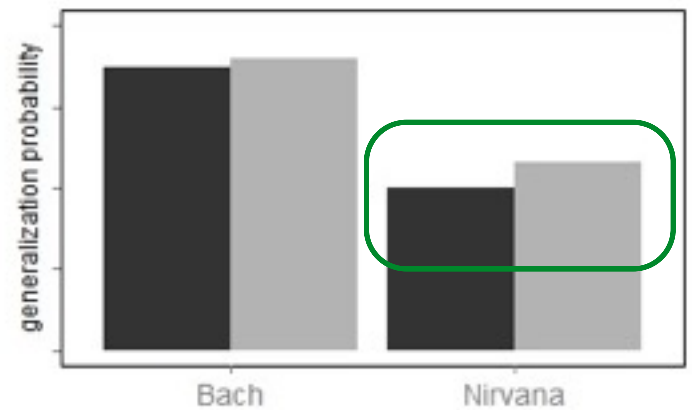
Nope.

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

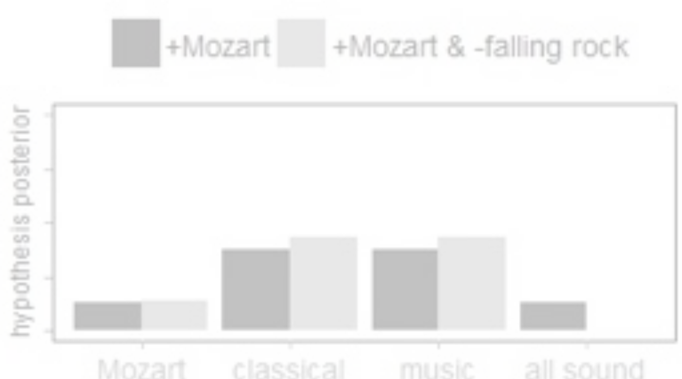
Weak sampling

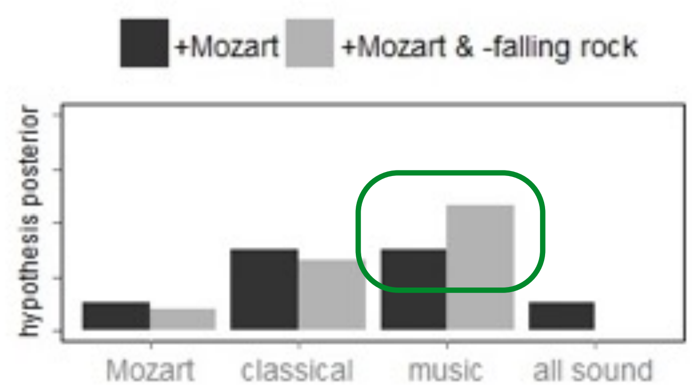
Strong sampling





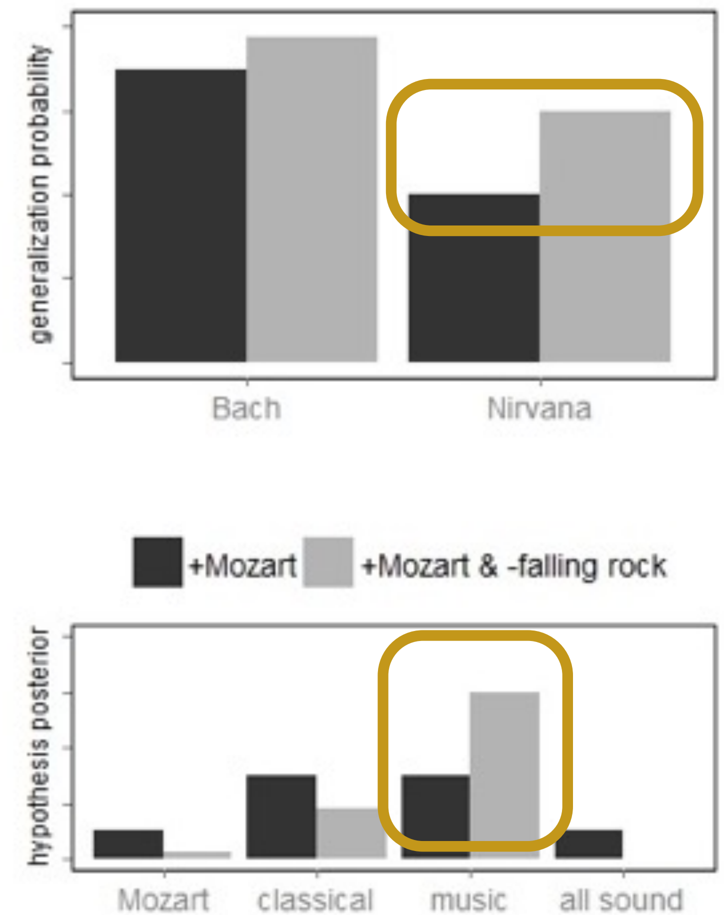
Meh.



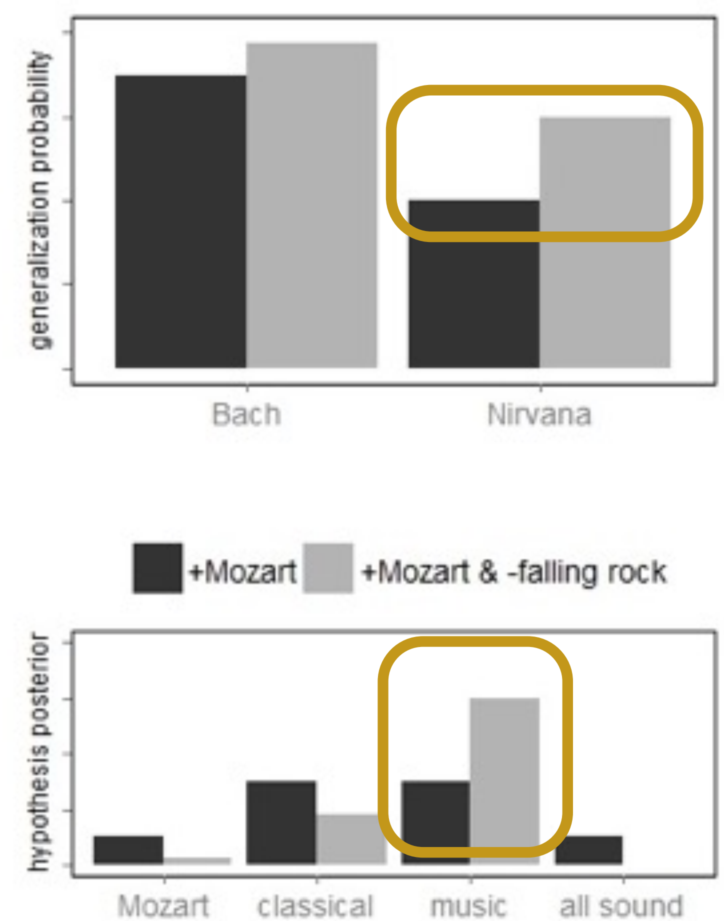


(Out-Bayesing Bayes??!)

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 purposefully 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
communicator…

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

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

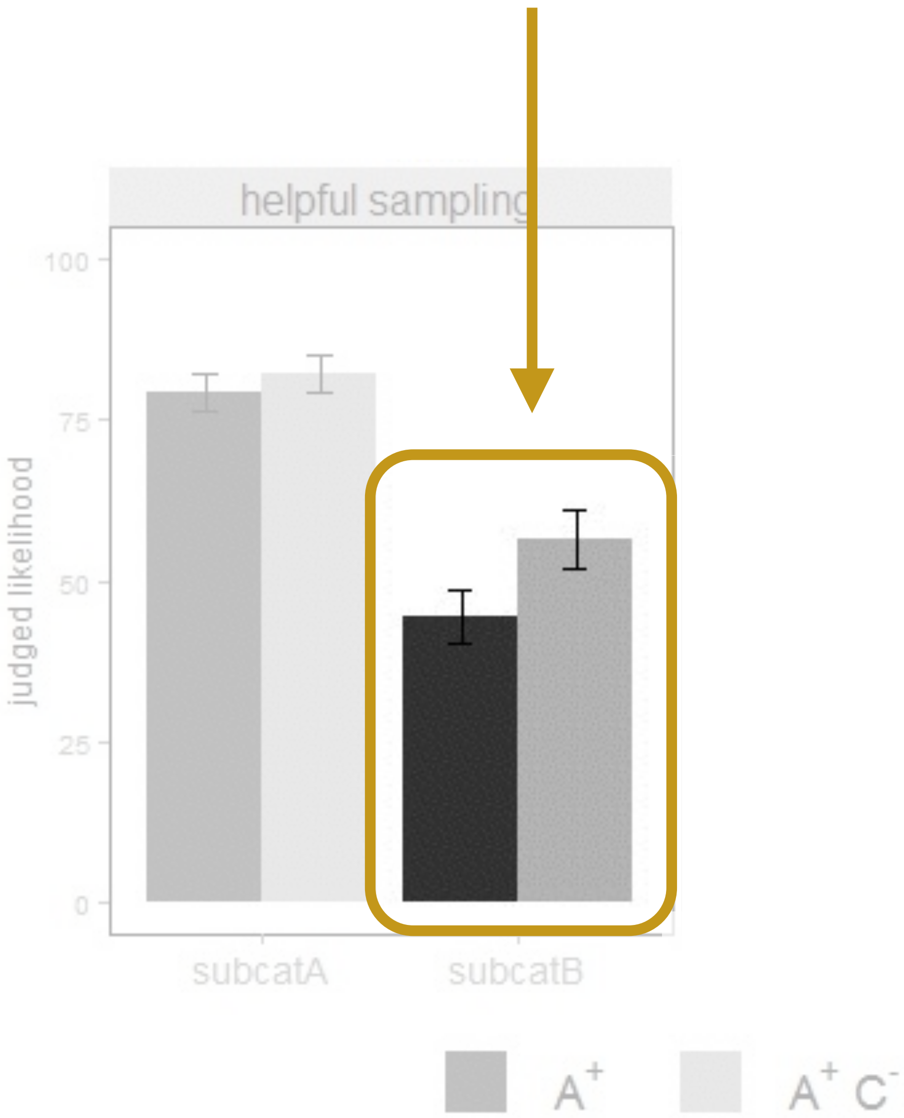
If the negative example is perceived as a "helpful hint" we should continue to get the effect

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

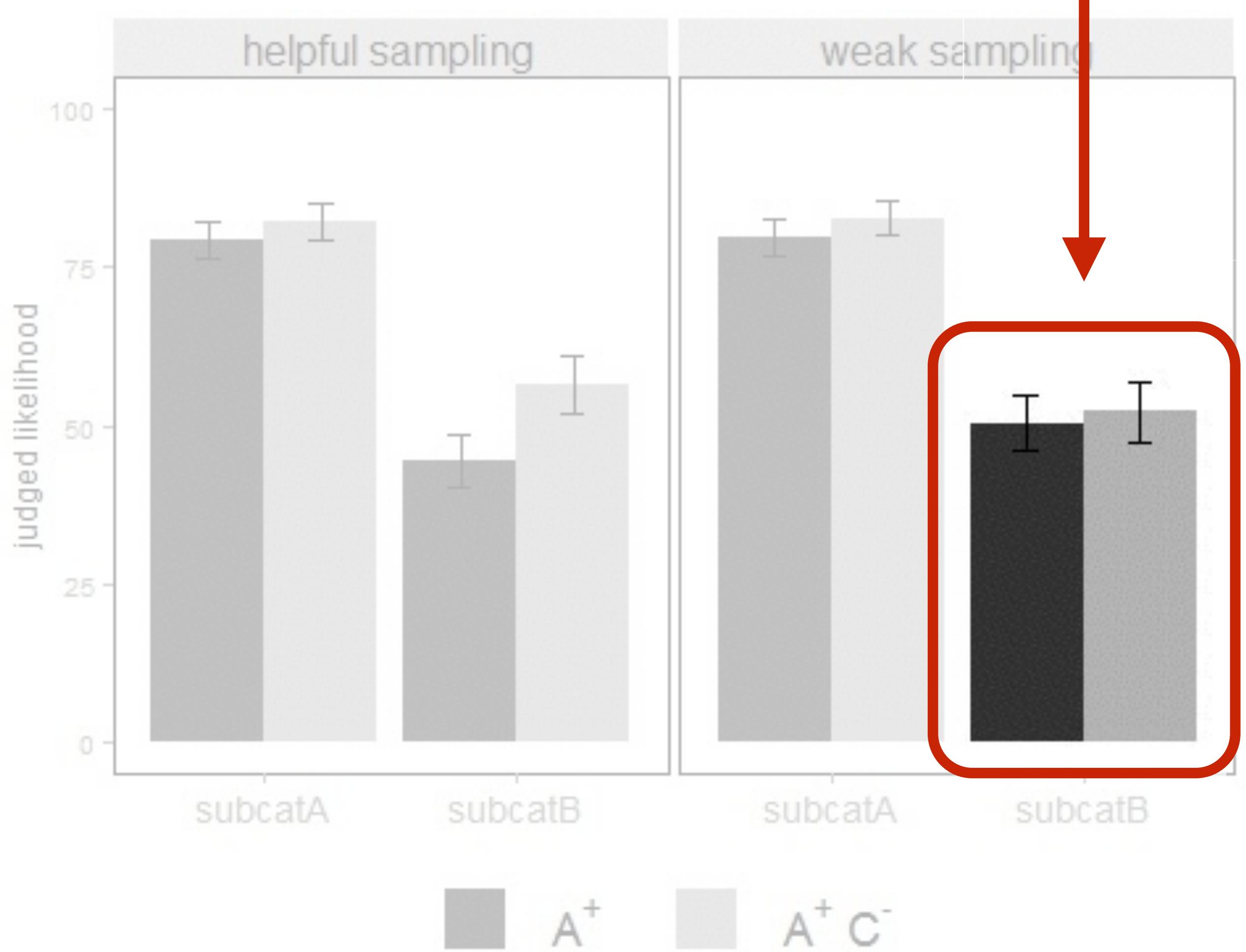
		topics		
	premise 1 (+)	premise 2 (-)	A-member	B-member
MUSIC	Mozart	waterfall	Bach	Nirvana
FRUIT	strawberries	grass blades	blackberry	apple
BIRDS	ducks	elephants	swan	blackbird
TYPES OF WATER	Atlantic ocean	tap water	Mediterranean	Lake Balaton
		fillers weak sampling		
	premise 1	premise 2	conclusion 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?



Let’s make the social aspect explicit:

The role of goals and social reasoning when aggregating expert opinions

Navarro, Dry & Lee (2012). Sampling assumptions in inductive generalization. *Cognitive Science*

Hendrickson, Perfors & Navarro (in preparation). One cat, two cats: Sampling models in different category learning tasks produce qualitatively different inductive generalisations

Ransom, Perfors & Navarro (in press). Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*

Voorspoels, Navarro, Perfors, Ransom & Storms (under revision). How do people learn from negative evidence? Non-monotonic generalizations and sampling assumptions in inductive reasoning. *Cognitive Psychology*

Perfors, Navarro & Shafto (in preparation). Stronger evidence isn't always better: The role of social inference in evidence selection and interpretation. Previously rejected from *JEP:G*

You're a journalist writing an article about expert opinions about climate change…

You're a journalist writing an article about expert opinions about climate change…

93% likely

95% likely

97% likely

92% likely

99% likely

91% likely

89% likely

You're a journalist writing an article about expert opinions about climate change…

93% likely

95% likely

97% likely

5% likely

92% likely

99% likely

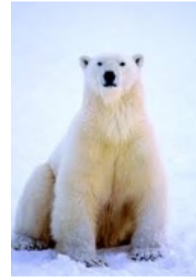
91% likely

89% likely


93


95


97


92


99

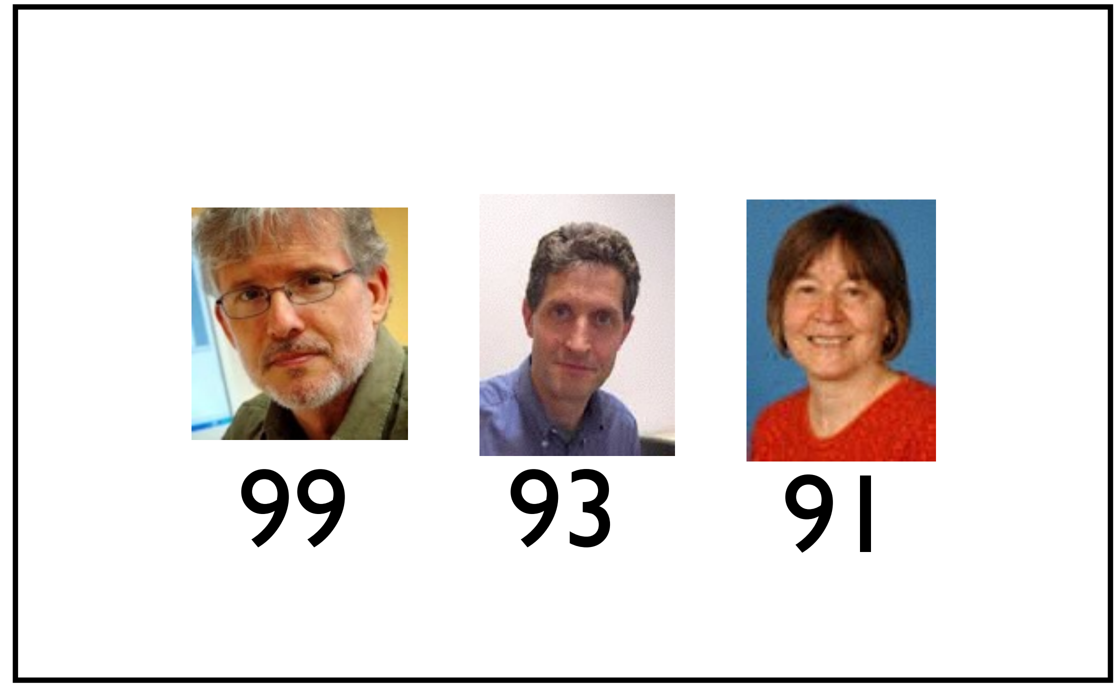

91


89

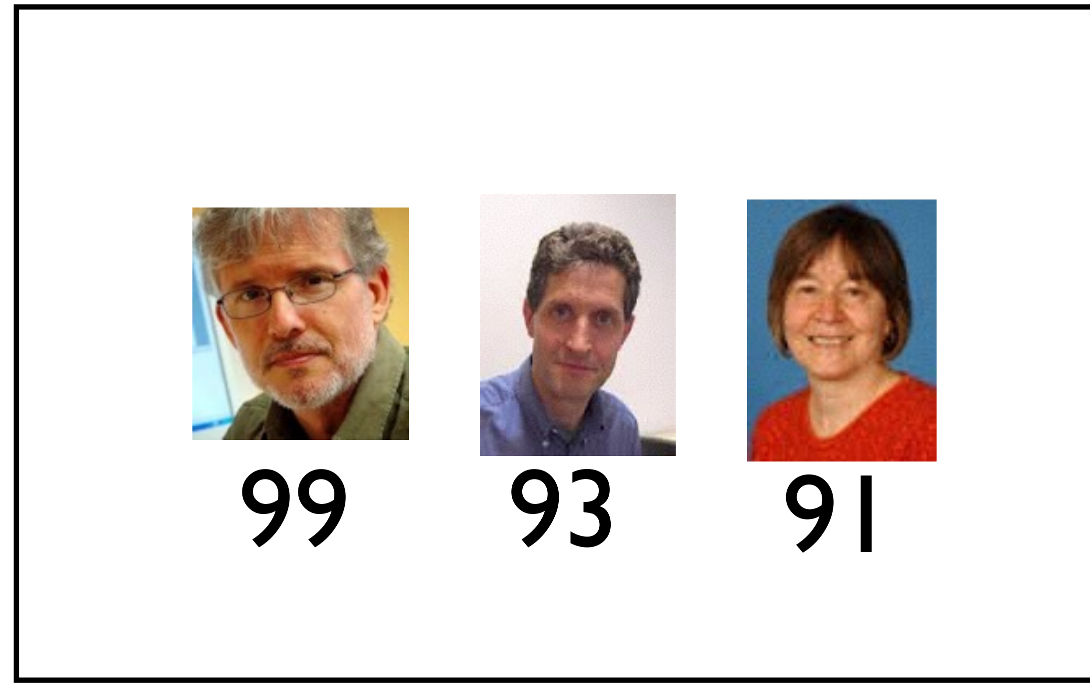

5

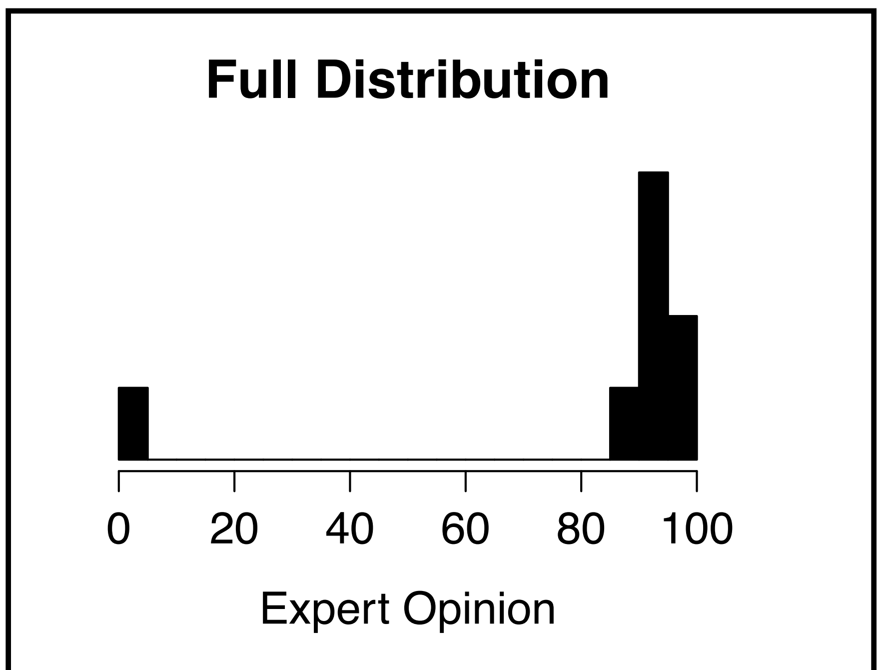
Here's your full distribution of expert opinion

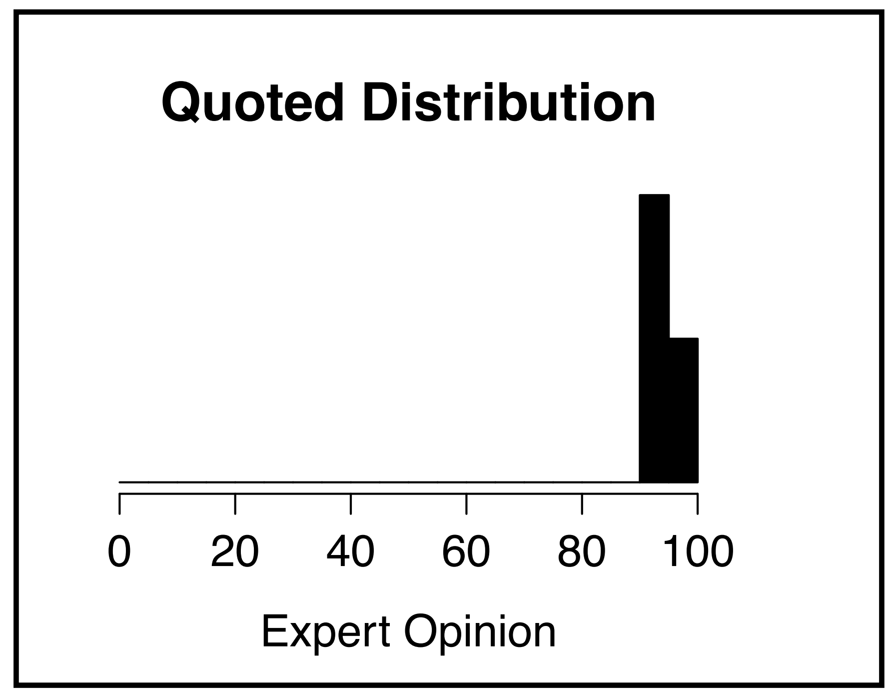
Your editor says the article only has room for (at most) three quotes. Who to choose??



Do you quote only from the consensus?

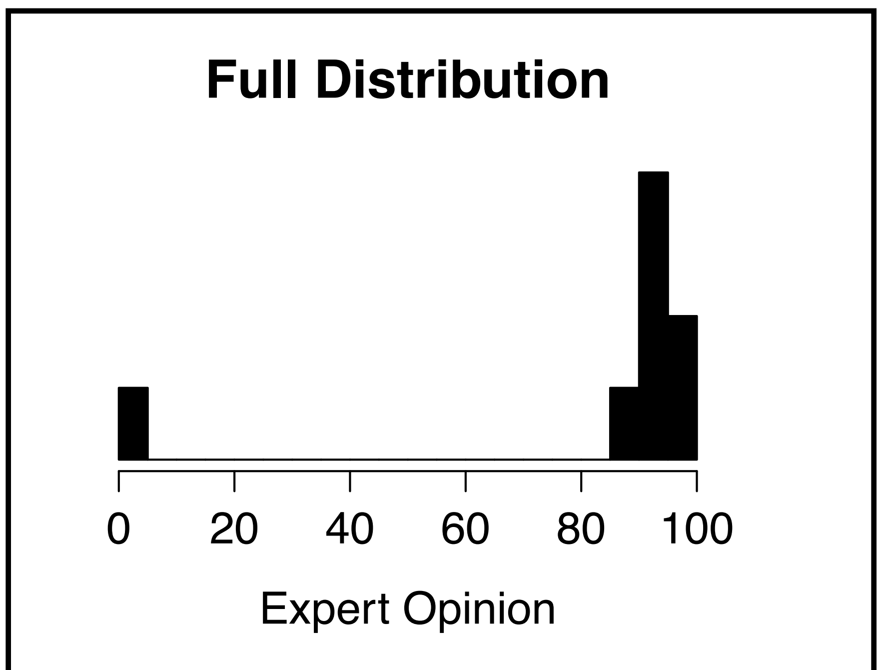


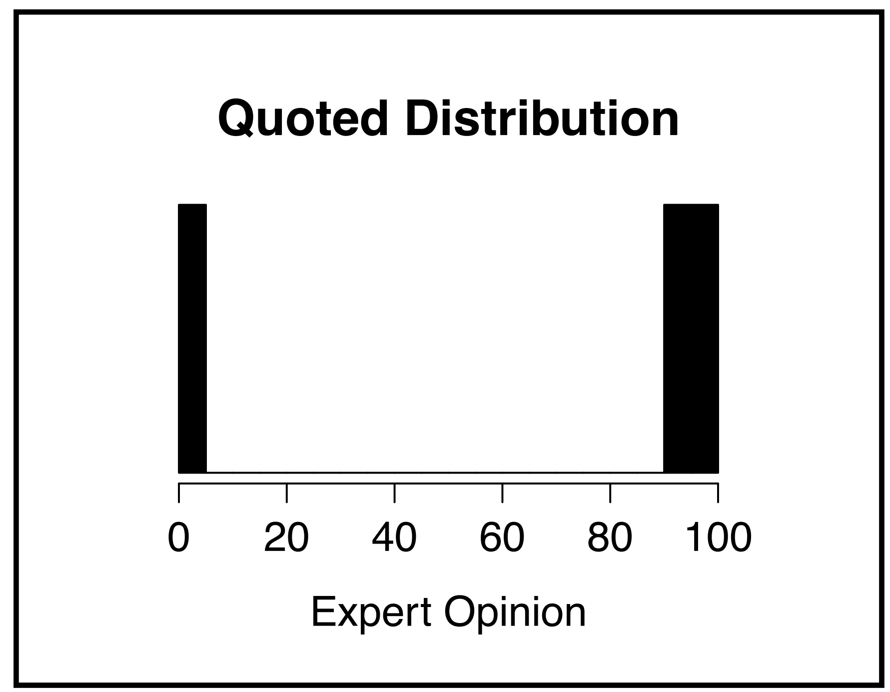




(maximises distributional similarity)

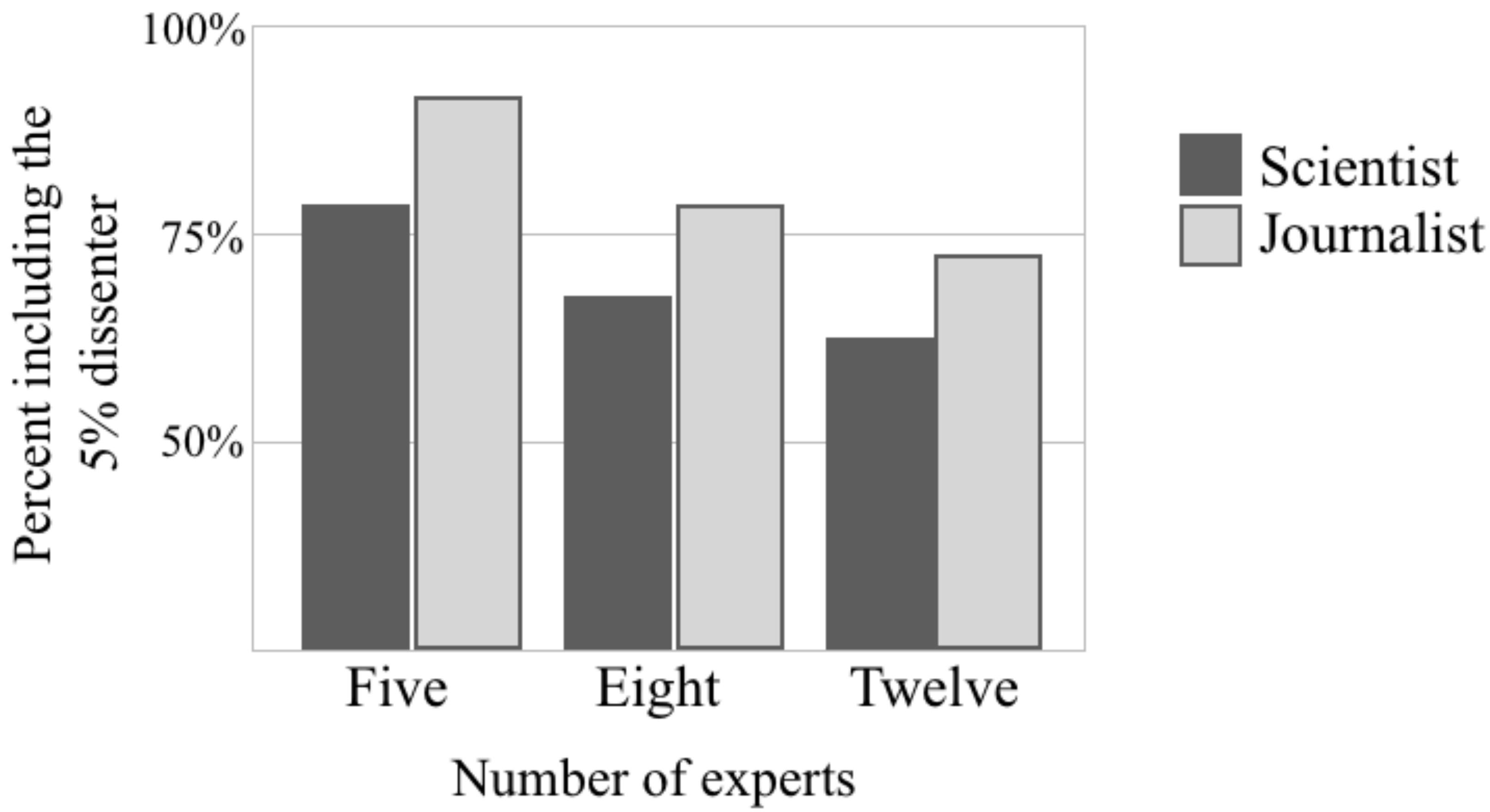
Or do you include the dissenter?

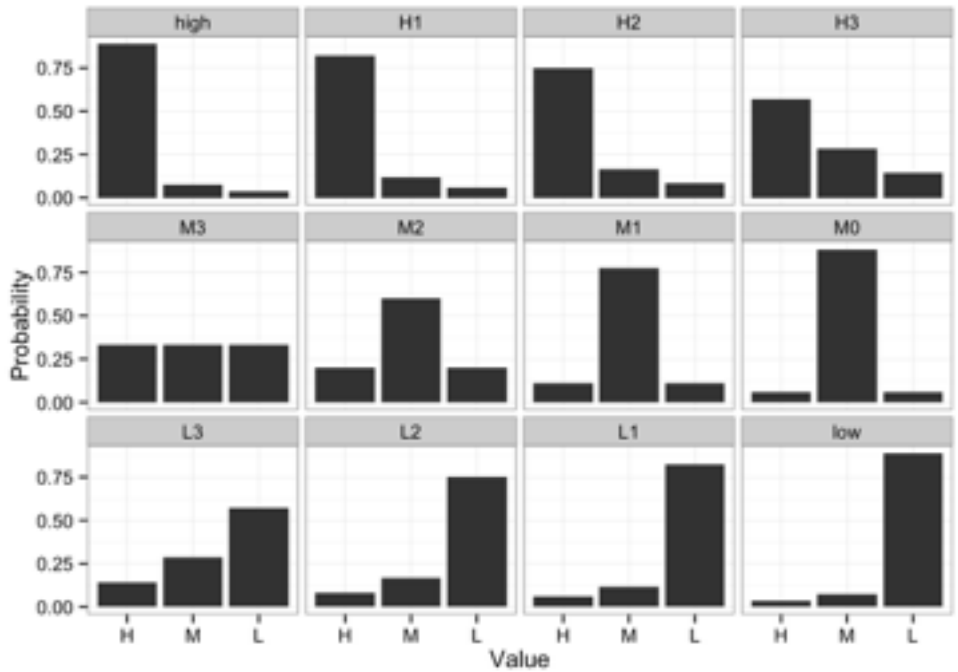




("full spectrum" but terrible approximation)

Some empirical data. Even when outnumbered 11 to 1, most people choose to quote the contrarian.





A hypothesis space of possible expert distributions

A hypothesis space of possible expert distributions

A hypothesis space of possible journalistic agendas

"Helpful" → Communicate the true distribution

"Bias high"

"Bias low"

→ Communicate a distribution with highest/lowest mean

Select evidence to manipulate
the reader's beliefs

Bayesian writer

Bayesian reader

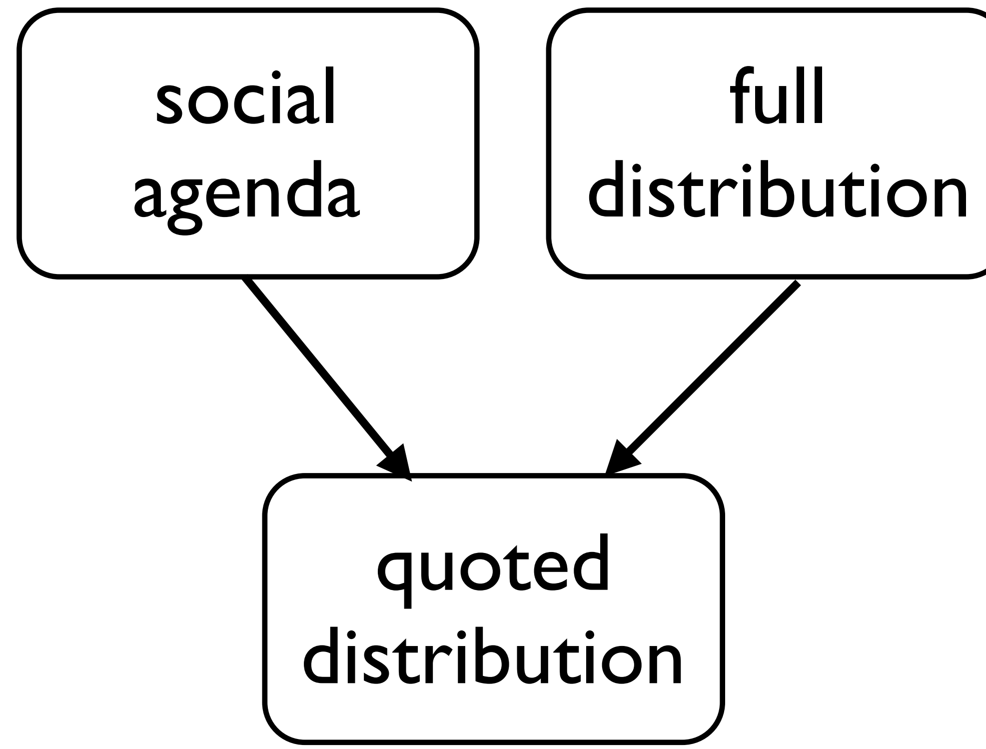
Select evidence to manipulate
the reader's beliefs

Bayesian writer

Bayesian reader

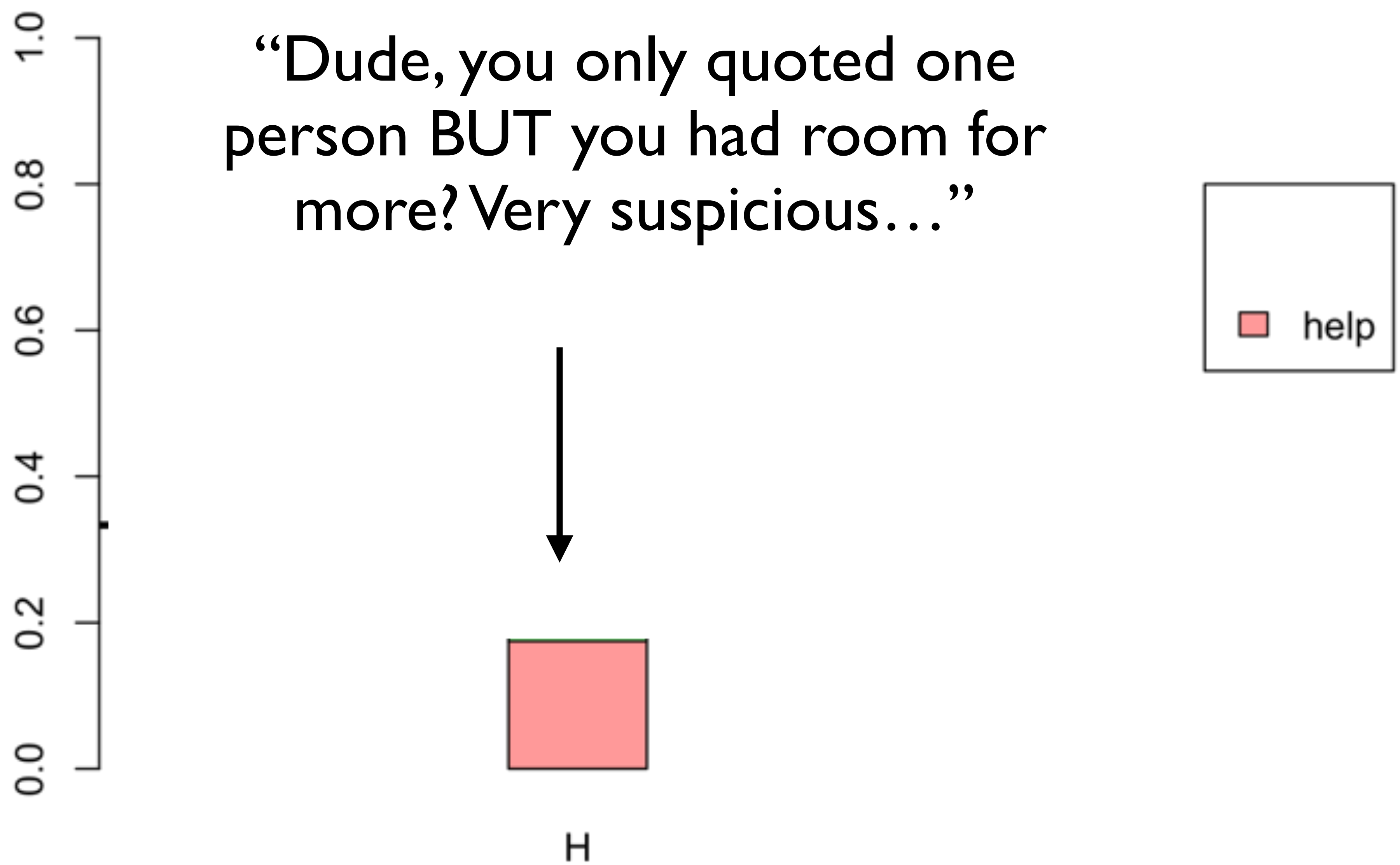
Guess the true distribution <u>AND</u>
infer the journalistic agenda

So what does a Bayesian reader infer about the Bayesian writer?

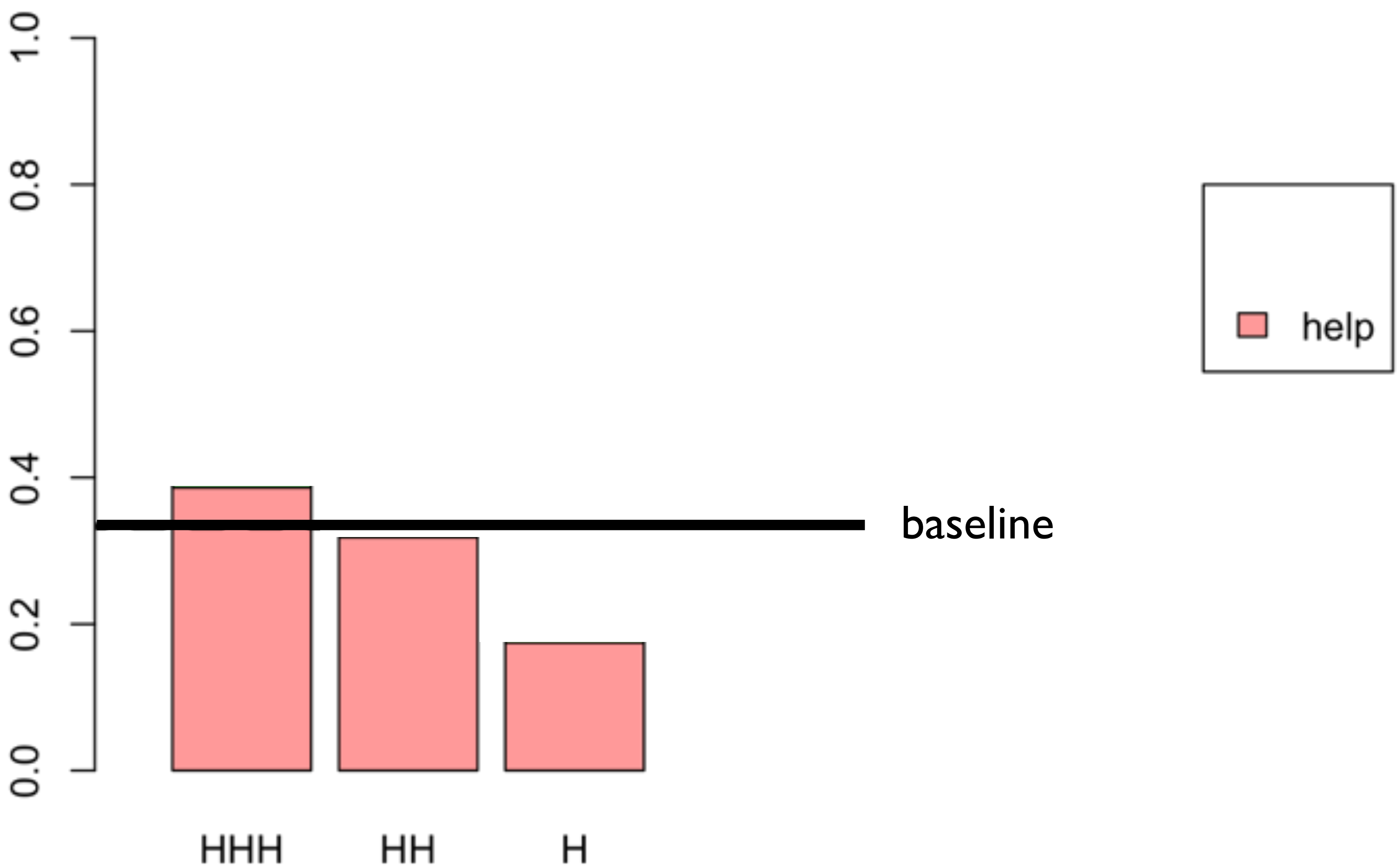


(I'll assume uniform priors over possible agendas and over possible distributional hypotheses)

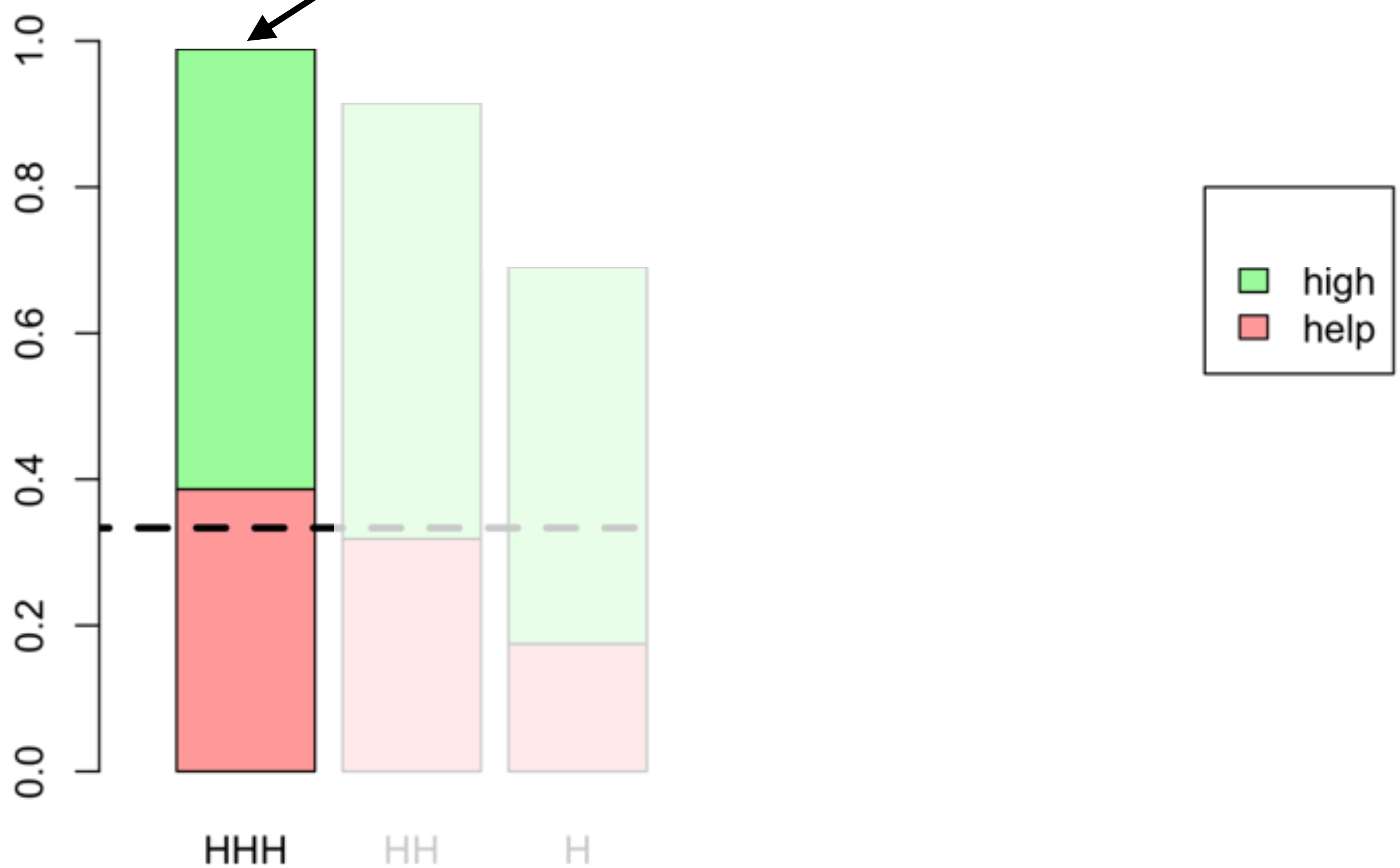
Quoting one expert only looks suspicious



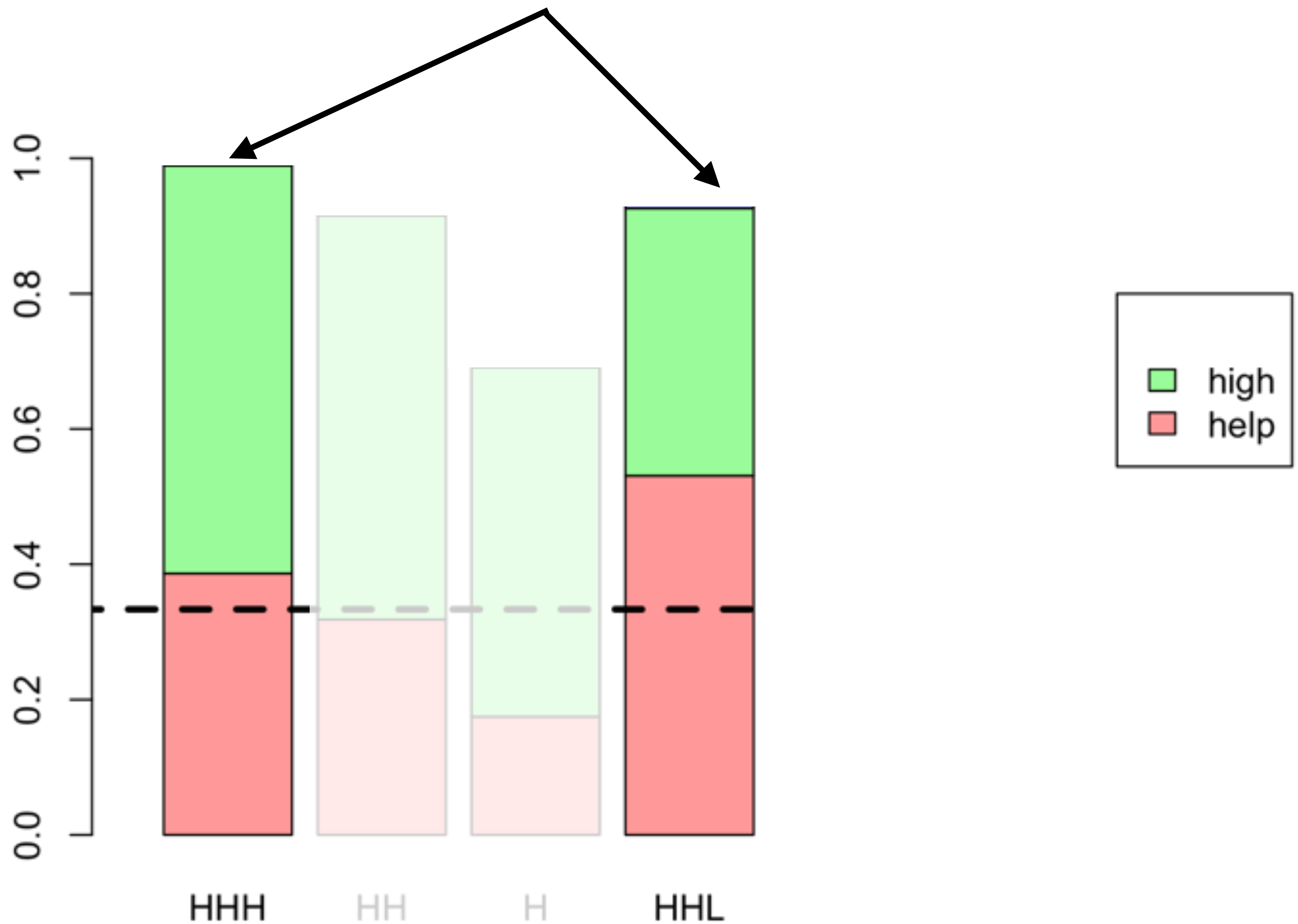
Anything less than maximum number of experts causes a deterioration of trust



But when all the quoted experts agree, the reader thinks you're probably biased



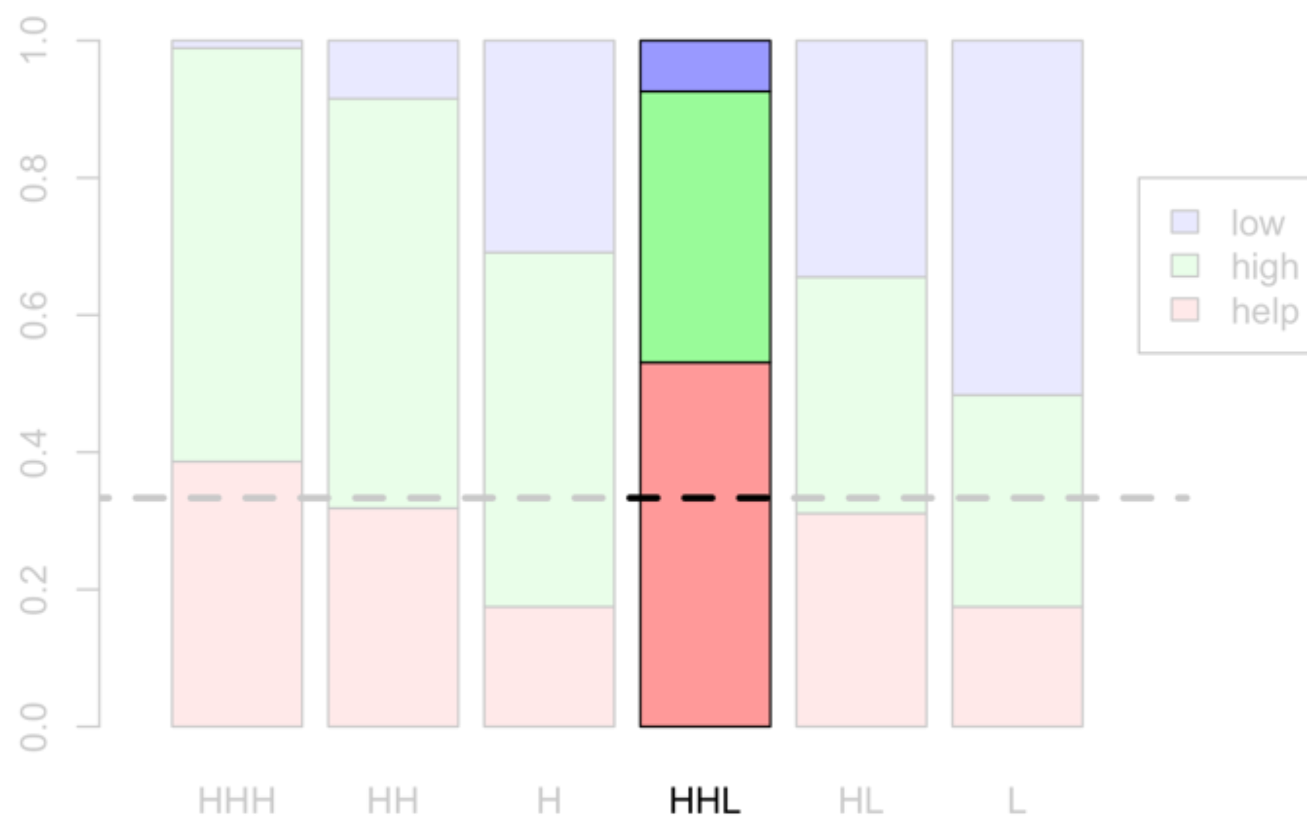
You can increase the reader's trust by including the contrarian

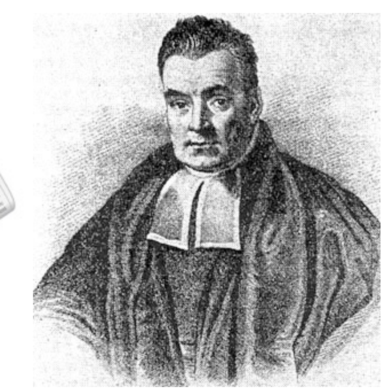


A Bayesian journalist who cares about their reputation has a strong motivation to pursue "he says she says" journalism

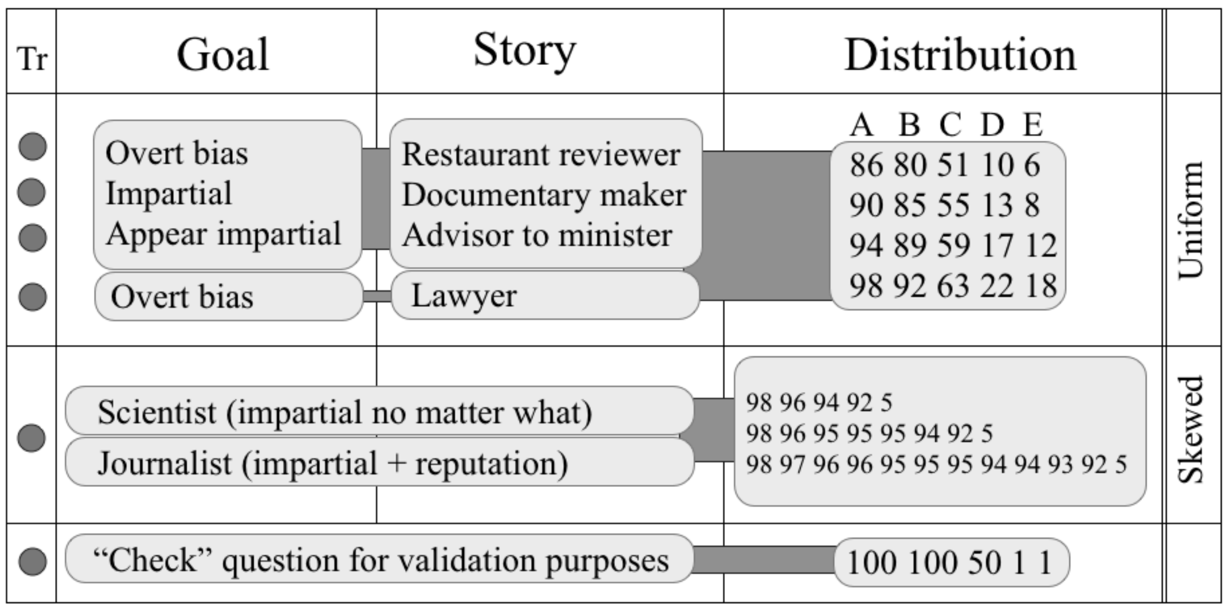

A Bayesian journalist who cares about their reputation has a strong motivation to pursue "he says she says" journalism

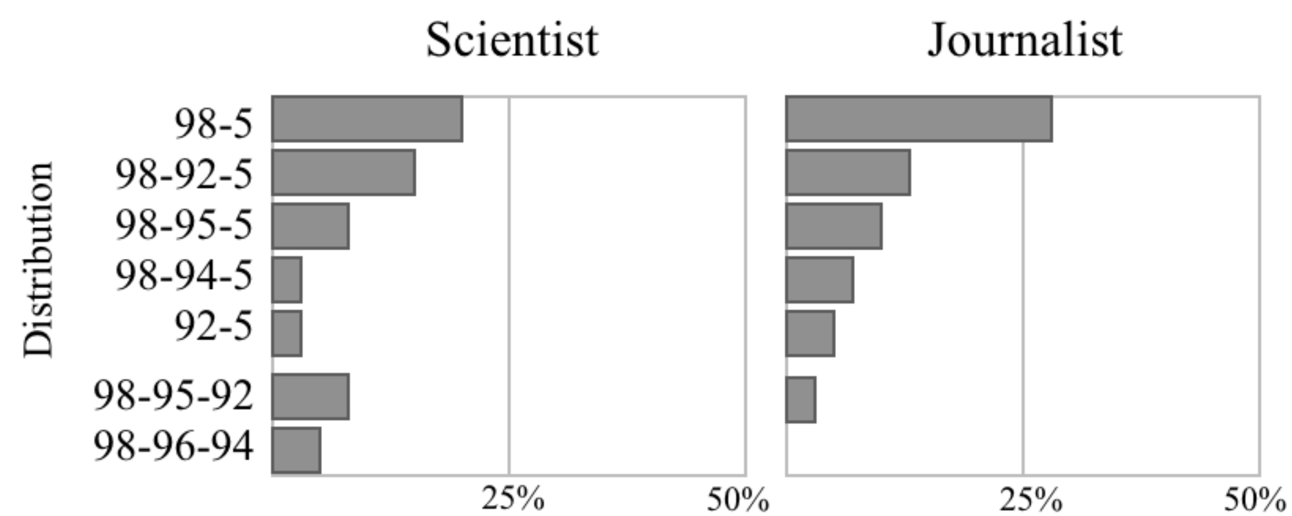


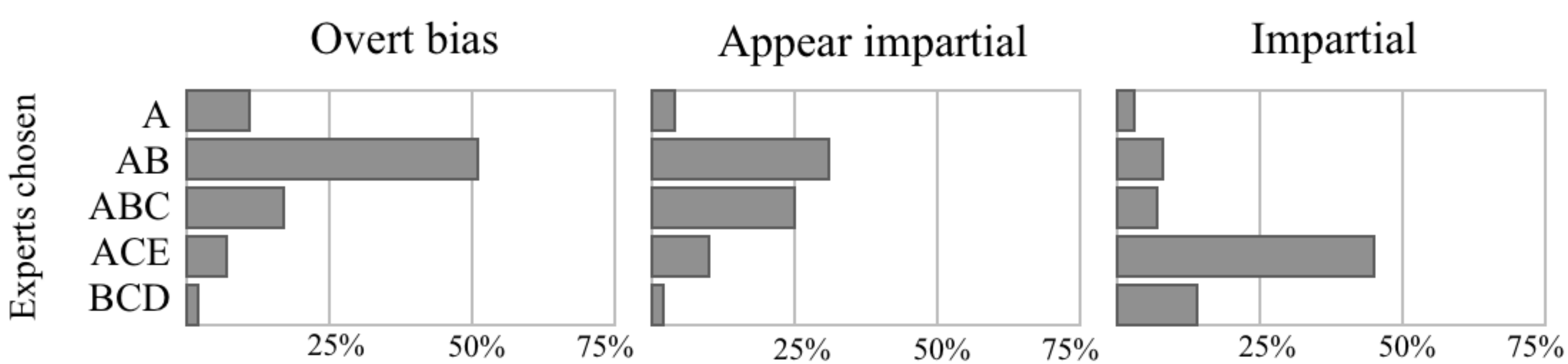


Because a Bayesian reader can't tell the difference between journalistic bias and expert consensus

Oh, and we have a heap of other data and modelling on this too, but I have no time…

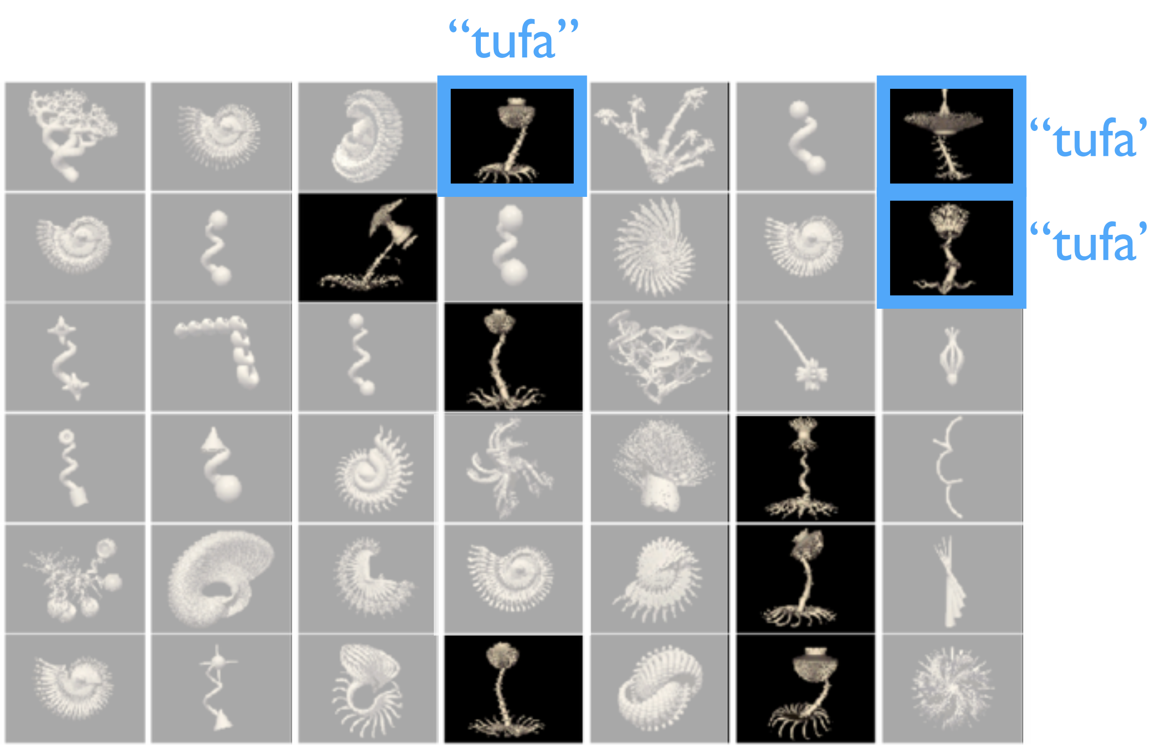


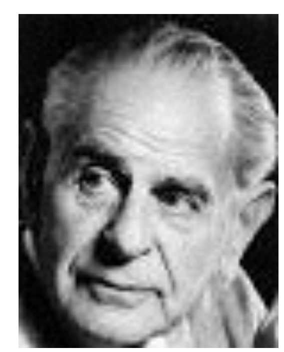




A few final thoughts about human reasoning and Bayesian reasoning

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

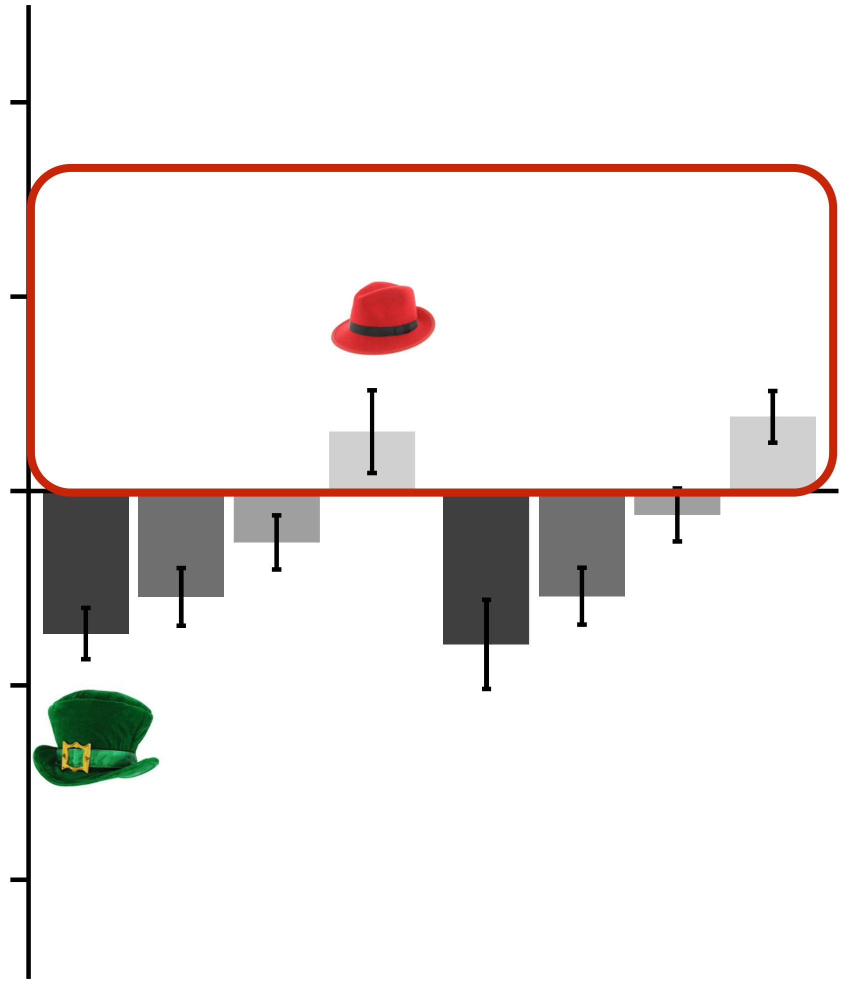




falsificationist learning

weak sampling

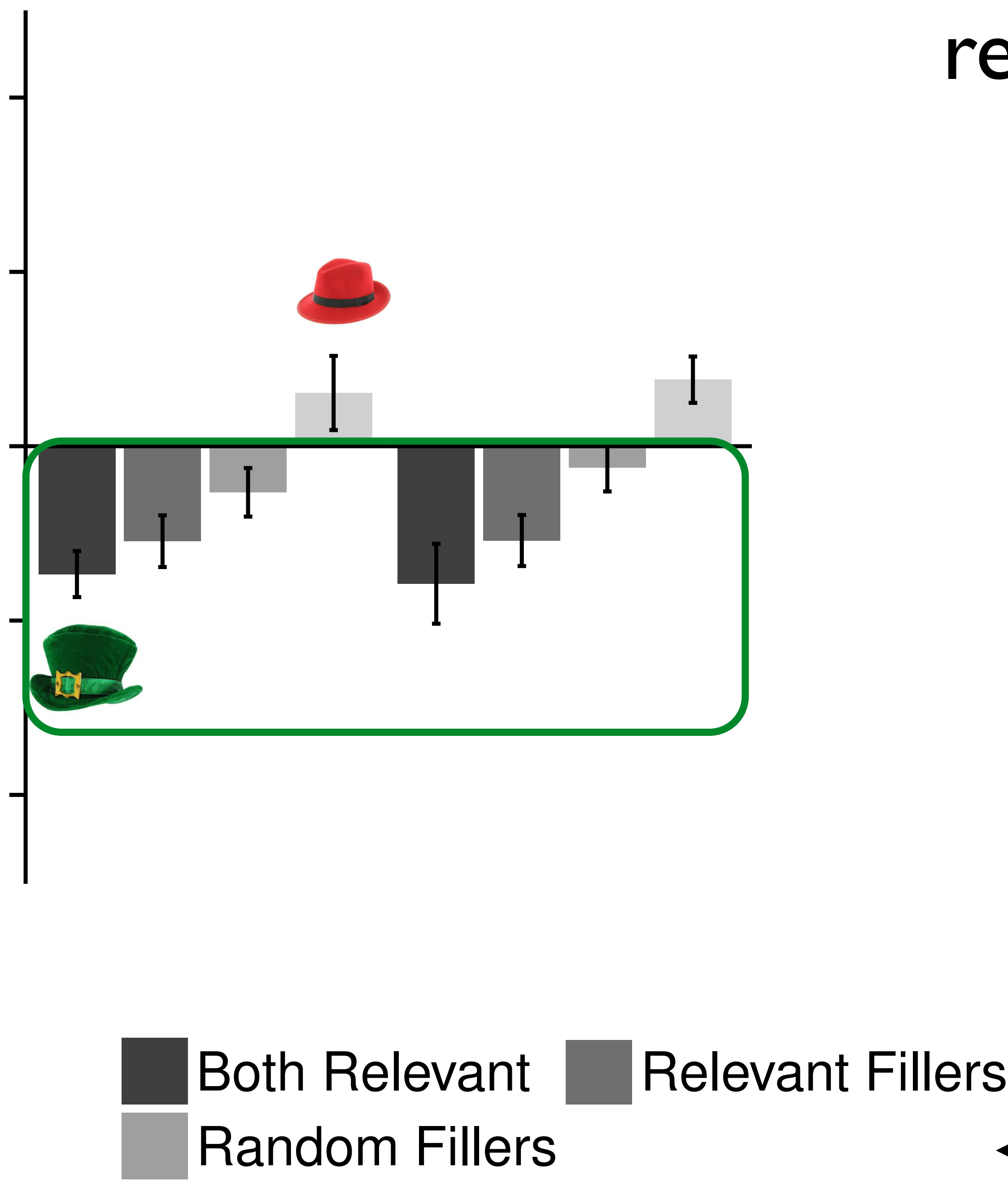
But why?



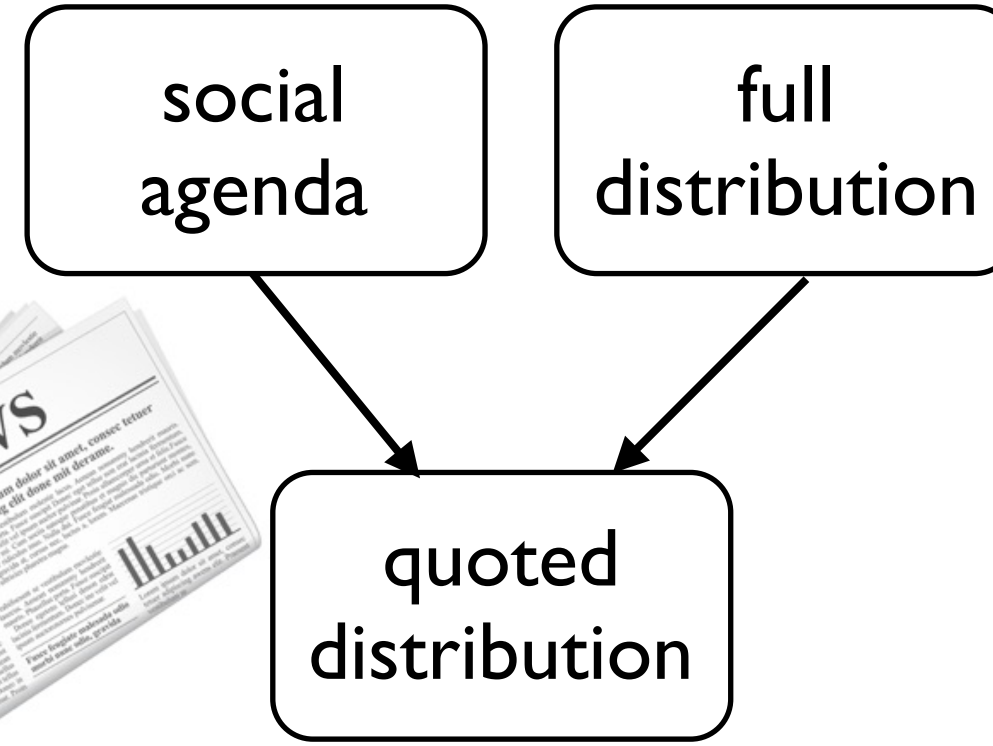
… it **only** makes sense when evidence is selected in an arbitrary and random fashion

Both Random

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



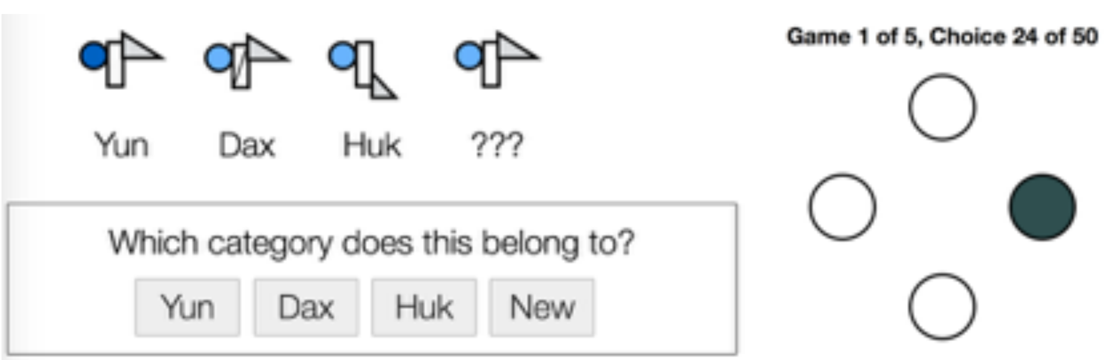
“Common sense” inference requires people to learn from complex (and smart) data sources…



We need to disentangle facts from agendas

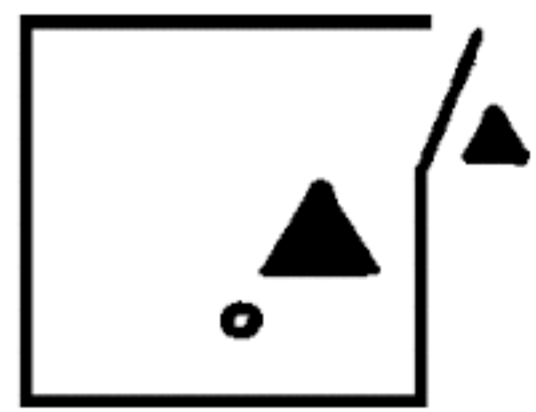
We need to detect trickery

(We actually ran this one. It was fun.)



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

We need to read the intention of other agents



Understanding human common sense reasoning requires something a lot richer

Thanks!