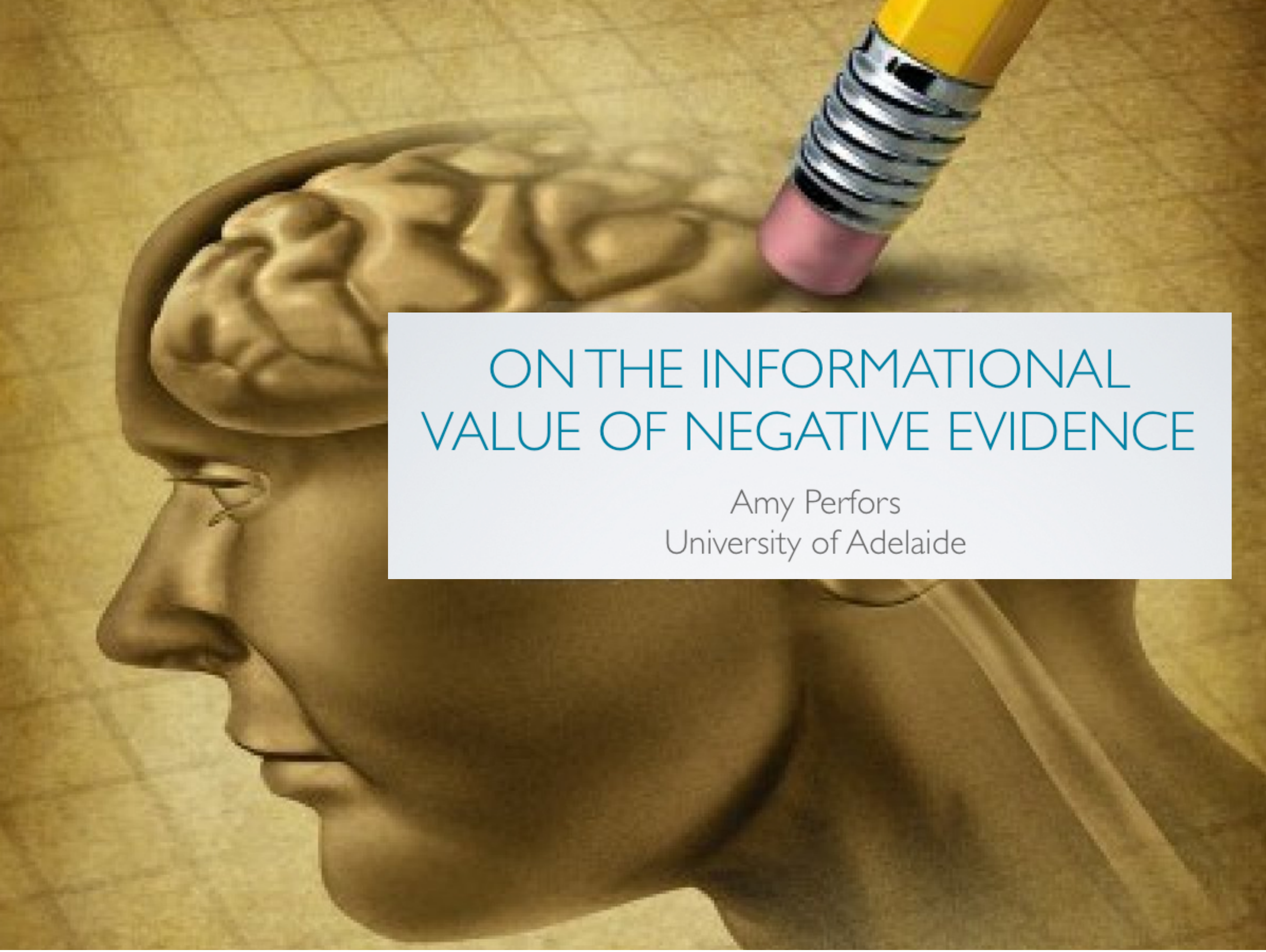


ON THE INFORMATIONAL VALUE OF NEGATIVE EVIDENCE

Amy Perfors
University of Adelaide

YOU FIND YOURSELF IN THIS WORLD...

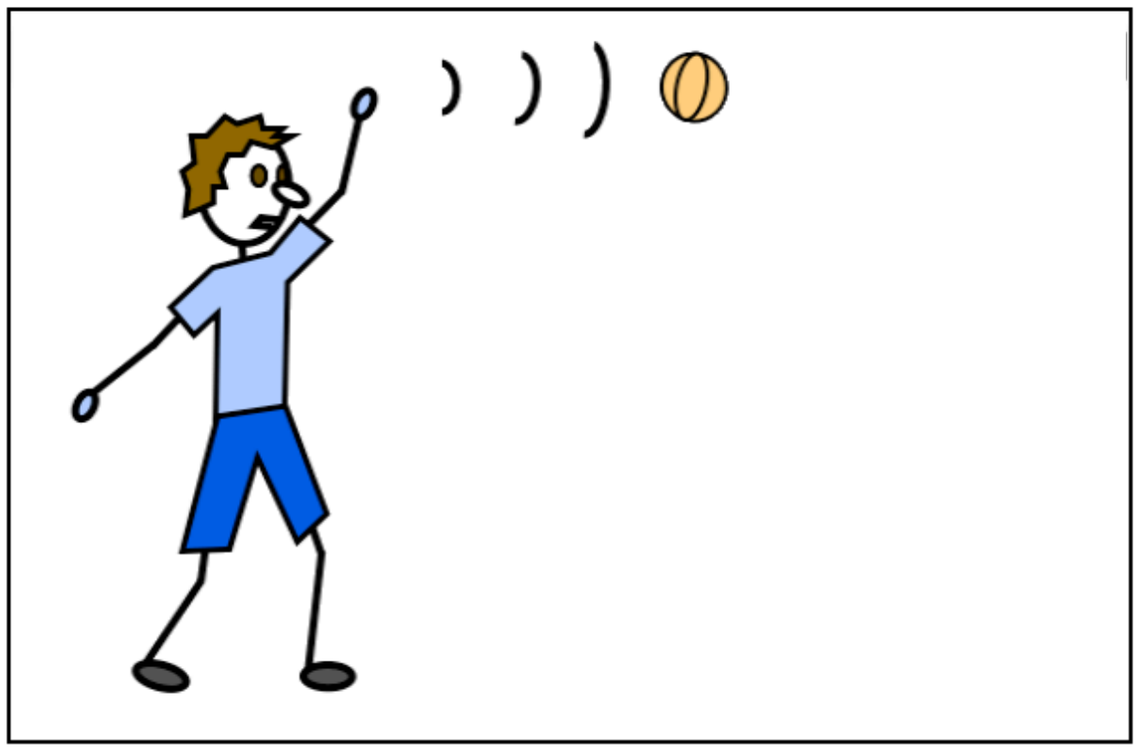
YOU FIND YOURSELF IN THIS WORLD...

YOU FIND YOURSELF IN THIS WORLD...

HOW DO YOU LEARN TO TALK ABOUT IT?

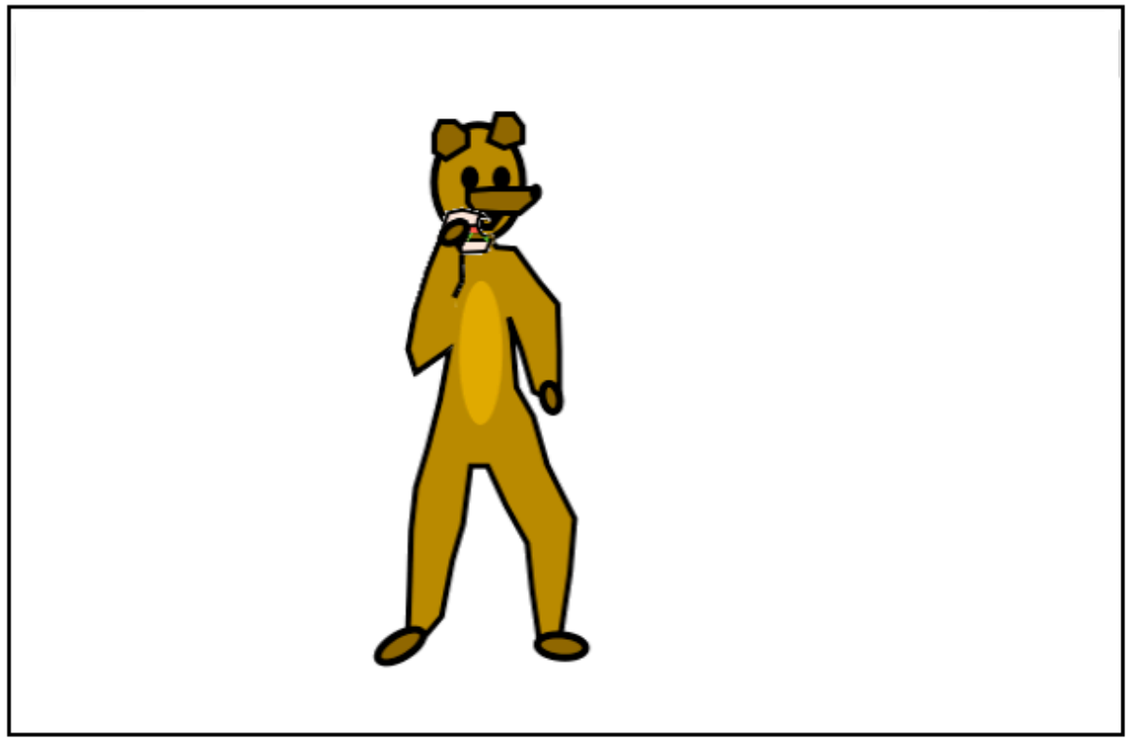
Positive example
(a sentence from the language)

CORRECT: SHIGU JOOM SEMP



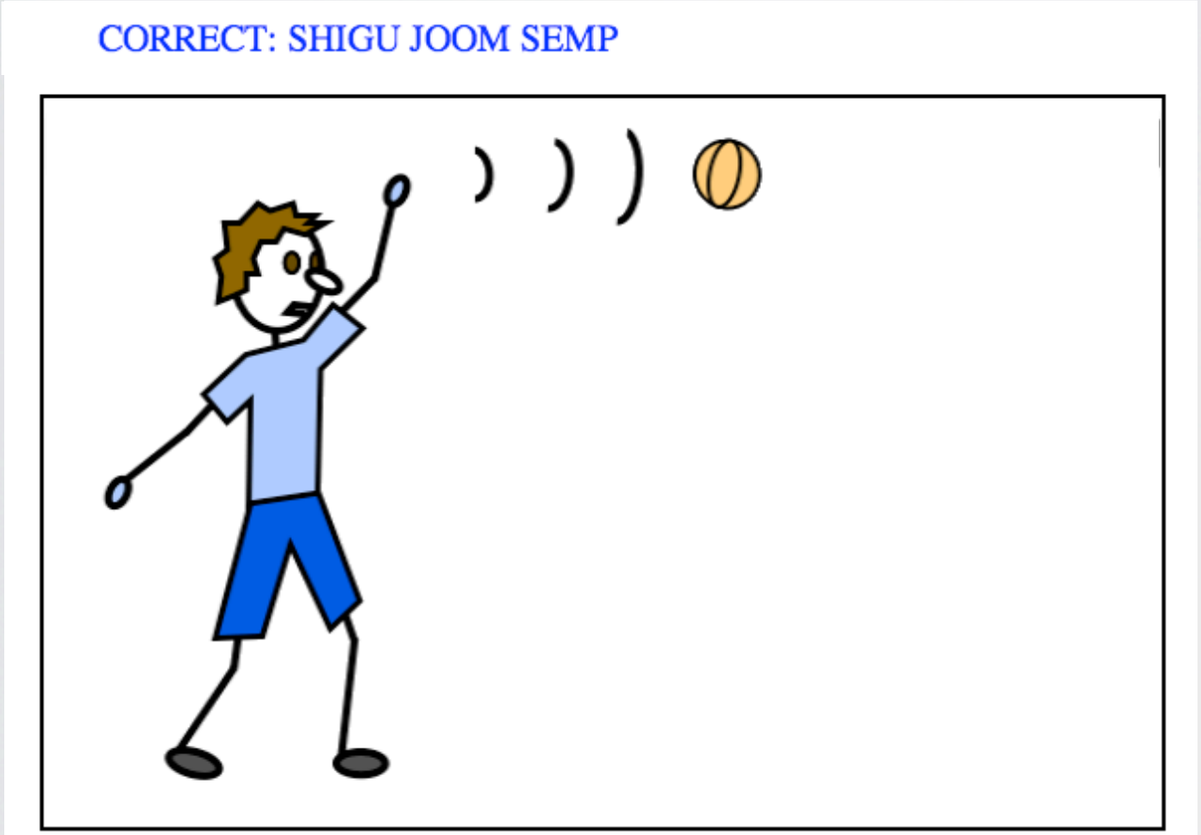
Negative example
(a sentence *not* from the language)

INCORRECT: VANTZO POOK



HOW DO YOU LEARN TO TALK ABOUT IT?

Positive example
(a sentence from the language)

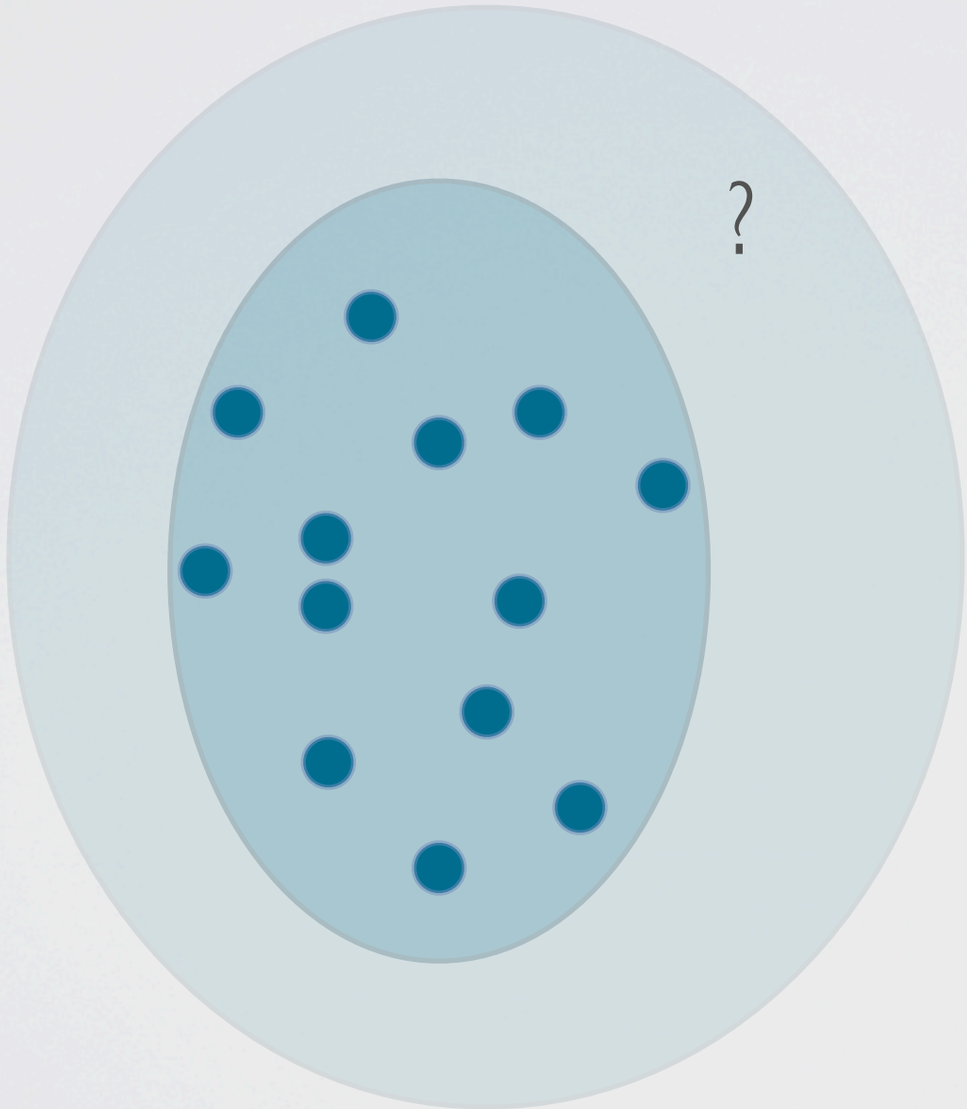


← Intuitively, this is better

- Why is this?
- What does "better" mean?
- And *when* is it better?

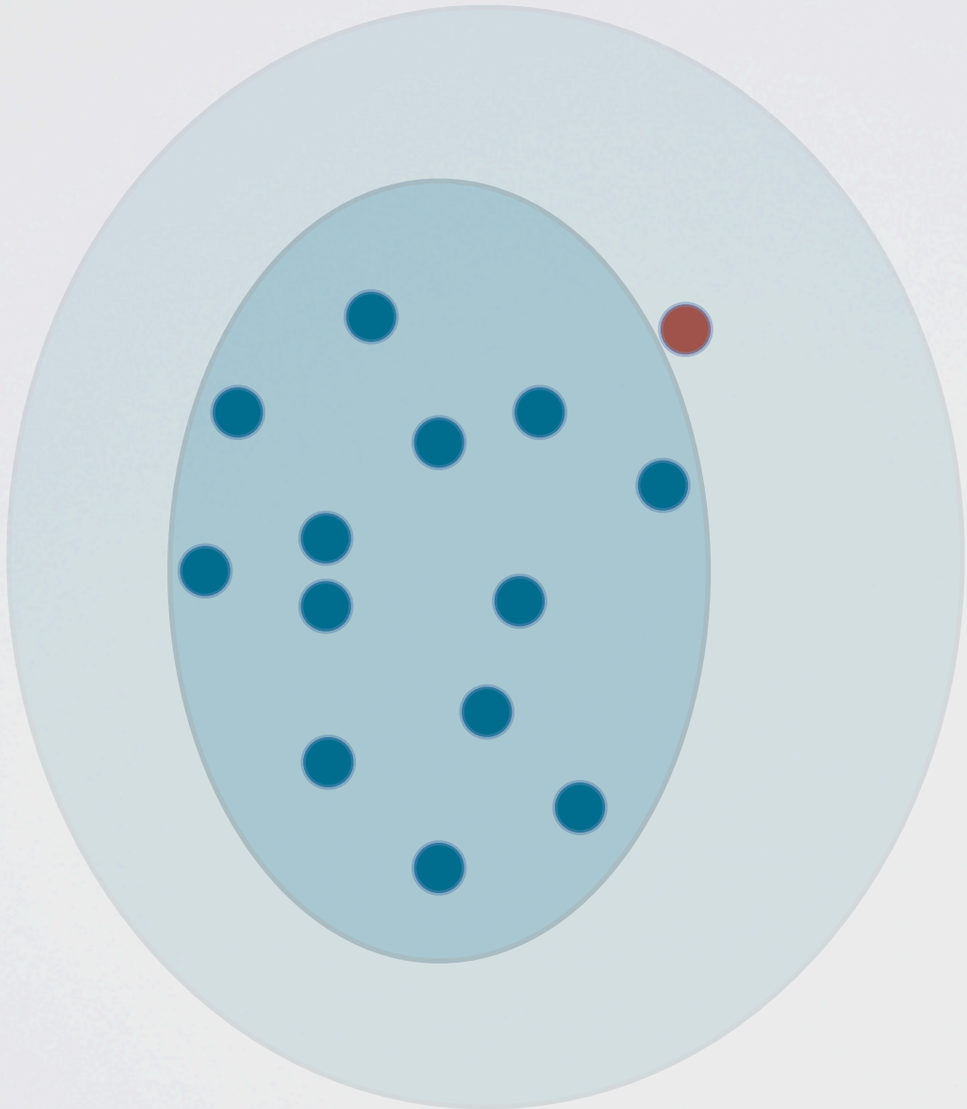
WHY DOES THIS MATTER?

Gold: need negative evidence to rule out superset grammars with 100% certainty



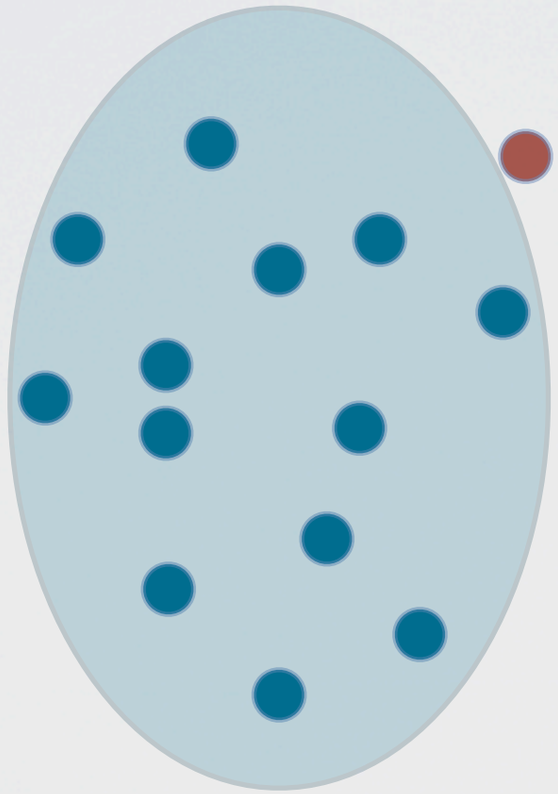
WHY DOES THIS MATTER?

Gold: need negative evidence to rule out superset grammars with 100% certainty



WHY DOES THIS MATTER?

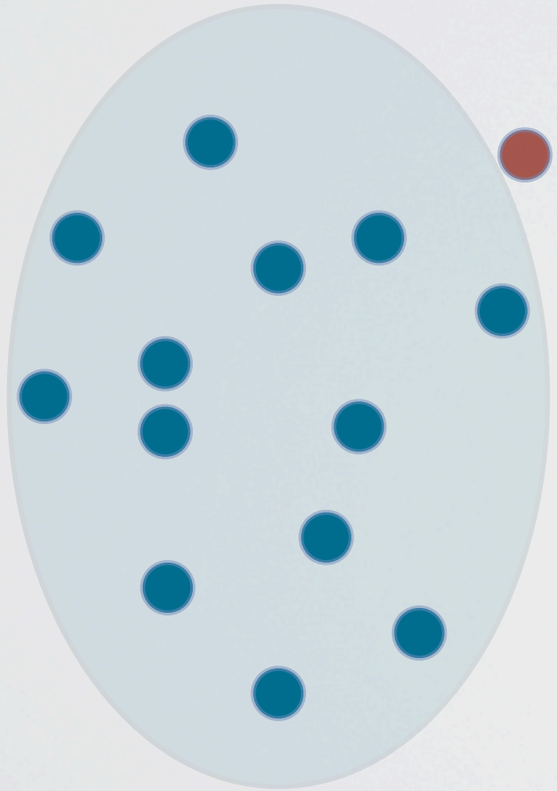
Gold: need negative evidence to rule out superset grammars with 100% certainty



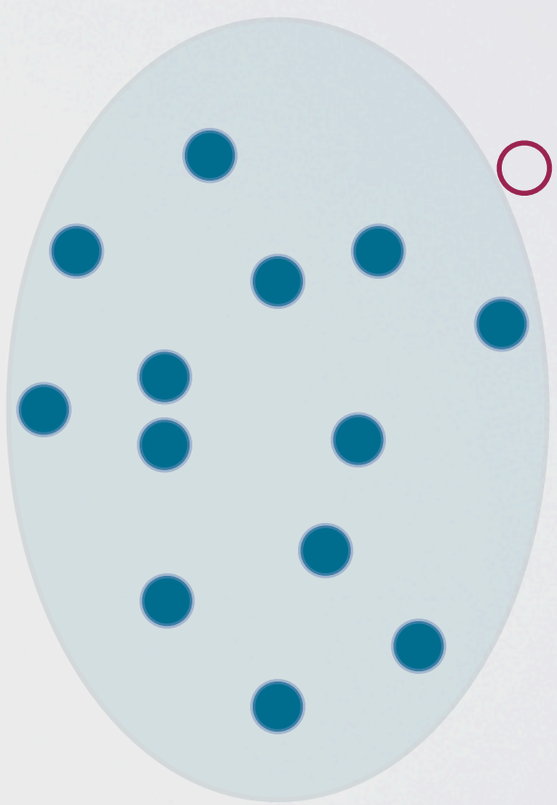
WHY DOES THIS MATTER?

Note that this proof makes no distinction between negative *examples* and negative *feedback*

Example: a sentence not from the language



Feedback: informing the learner when they produced an incorrect sentence



WHY DOES THIS MATTER?

Note that this proof makes no distinction between negative *examples* and negative *feedback*

Example: a sentence not from the language

Almost never occurs in speech to children

Child: I can't sing no songs of yours.
Adult: *Any* songs.

Feedback: informing the learner when they produced an incorrect sentence

Sometimes occurs, although not always (and often in the form of reformulations)

Child: I telled on him.
Adult: You *told* on him.

WHY DOES THIS MATTER?

Note that this proof makes no distinction between negative *examples* and negative *feedback*

Example: a sentence not from the language

Almost never occurs in speech to children

Feedback: informing the learner when they produced an incorrect sentence

Sometimes occurs, although not always (and often in the form of reformulations)

Measures of the frequency of negative feedback (as explicit feedback or reformulations) range between **10-67%** of ungrammatical utterances

(Saxton, 2000; Bohannon & Stanowicz, 1988, Farrar 1992, Chouinard & Clark, 2003)

QUESTIONS

- Why are negative examples not used, but negative feedback does occur to at least some extent?
- Is it more informationally optimal to receive negative feedback all of the time, unlike what children actually receive?
- Do we gain anything by thinking about negative evidence in terms of utility, rather than -- like Gold -- in terms of logical necessity?

GOAL OF THE TALK

I will be exploring the *utility* (in a probabilistic or information-theoretic sense) of different kinds of negative evidence.

By doing so we can learn more about:
(a) why negative evidence is often useless
(b) what kinds of negative evidence can be useful
(c) when that negative evidence is useful

.. and compare it to the prevalence of negative evidence received by children

TAKE-HOME POINTS

Negative *examples* are always useless

Negative *feedback* is sometimes useful, but sometimes it is more informative to instead receive positive examples from the language

This occurs due to the nature of linguistic hypotheses and assumptions about how the data are generated (which there is evidence people are sensitive to)

The prevalence of negative feedback predicted as optimal in our analysis is similar to the observed prevalence of negative feedback to children

KEY ASSUMPTION: GRADIENCE

Our understanding of languages is not all-or-none and can be represented as a probability distribution over possible hypotheses about the correct language.

Data contains *information* whose utility we can measure by looking at how it changes the probability associated with the correct hypothesis.

THE PLAN

Part 1: Negative examples

- Experiment: they are totally useless
- Theory: why are they useless?

Part 2: Negative feedback and change over time

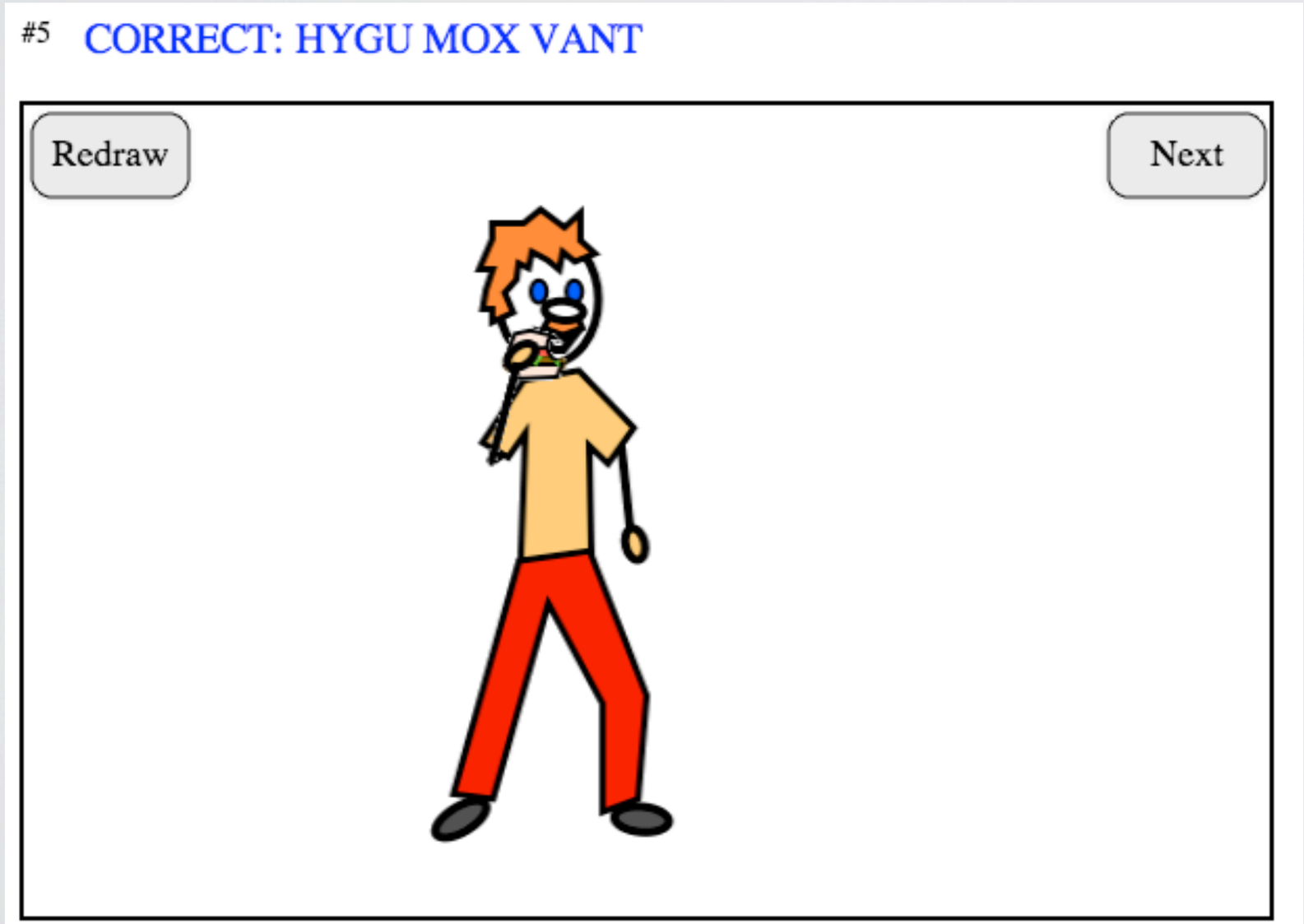
- Why is it more useful than examples
- How does this change as the learner acquires some language?
- Dependence on other assumptions about data

PART I

NEGATIVE EXAMPLES

NEGATIVE EXAMPLES: THE EXPERIMENT

Presented people with an artificial language



6 agents
(boy, girl, man, lady, bear, alien)

2 noun particles (one for humans, one for non-humans)

6 objects (apple, banana, hamburger, ball, flower, book)

6 verbs (give, throw, kick, eat, carry, hit)

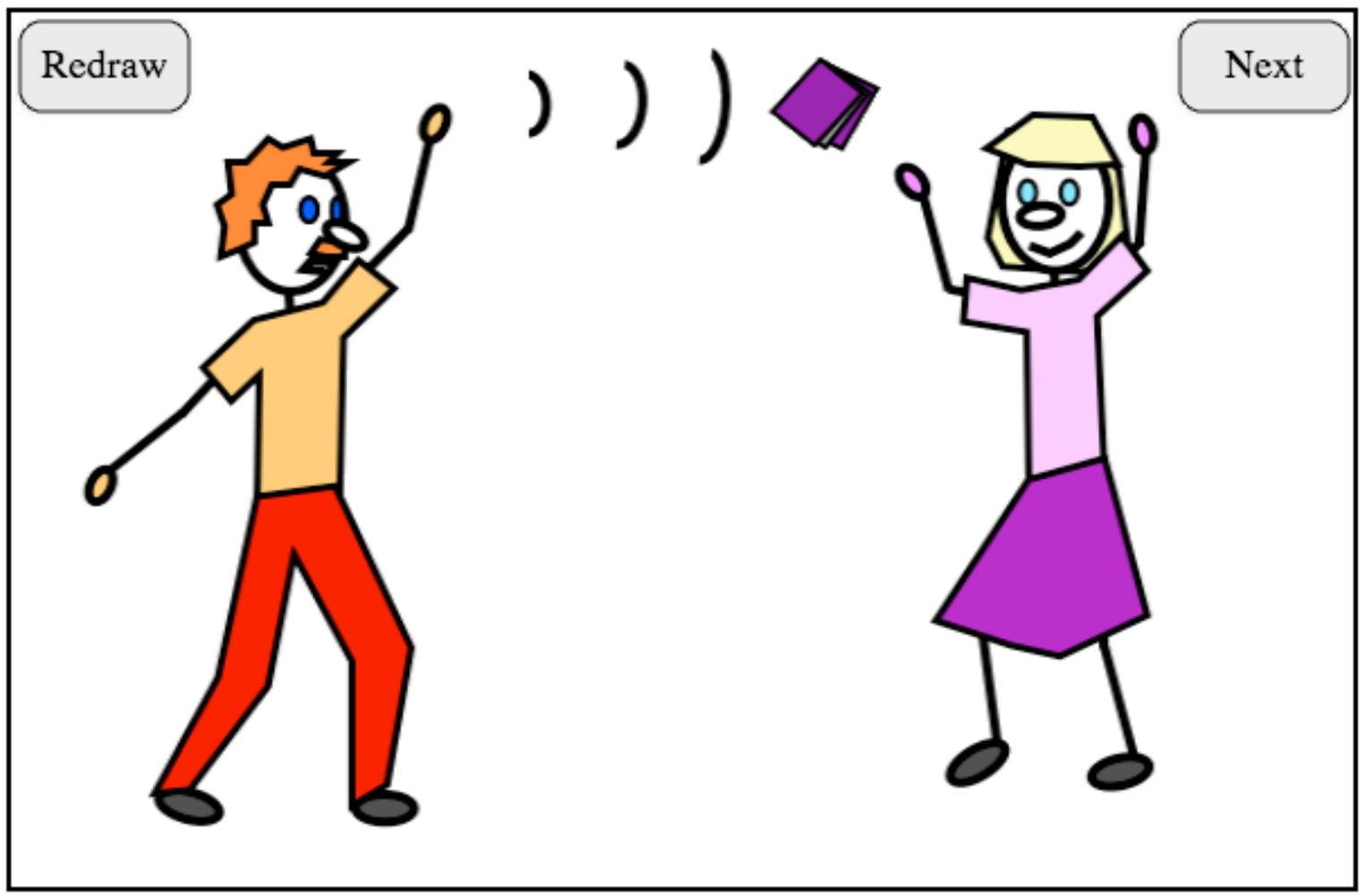
1 particle indicating recipient of an action

SOV word order

NEGATIVE EXAMPLES: THE EXPERIMENT

Presented people with an artificial language

#7 CORRECT: HYGU NEGULI WUZ SEMP



6 agents
(boy, girl, man, lady, bear, alien)

2 noun particles (one for humans, one for non-humans)

6 objects (apple, banana, hamburger, ball, flower, book)

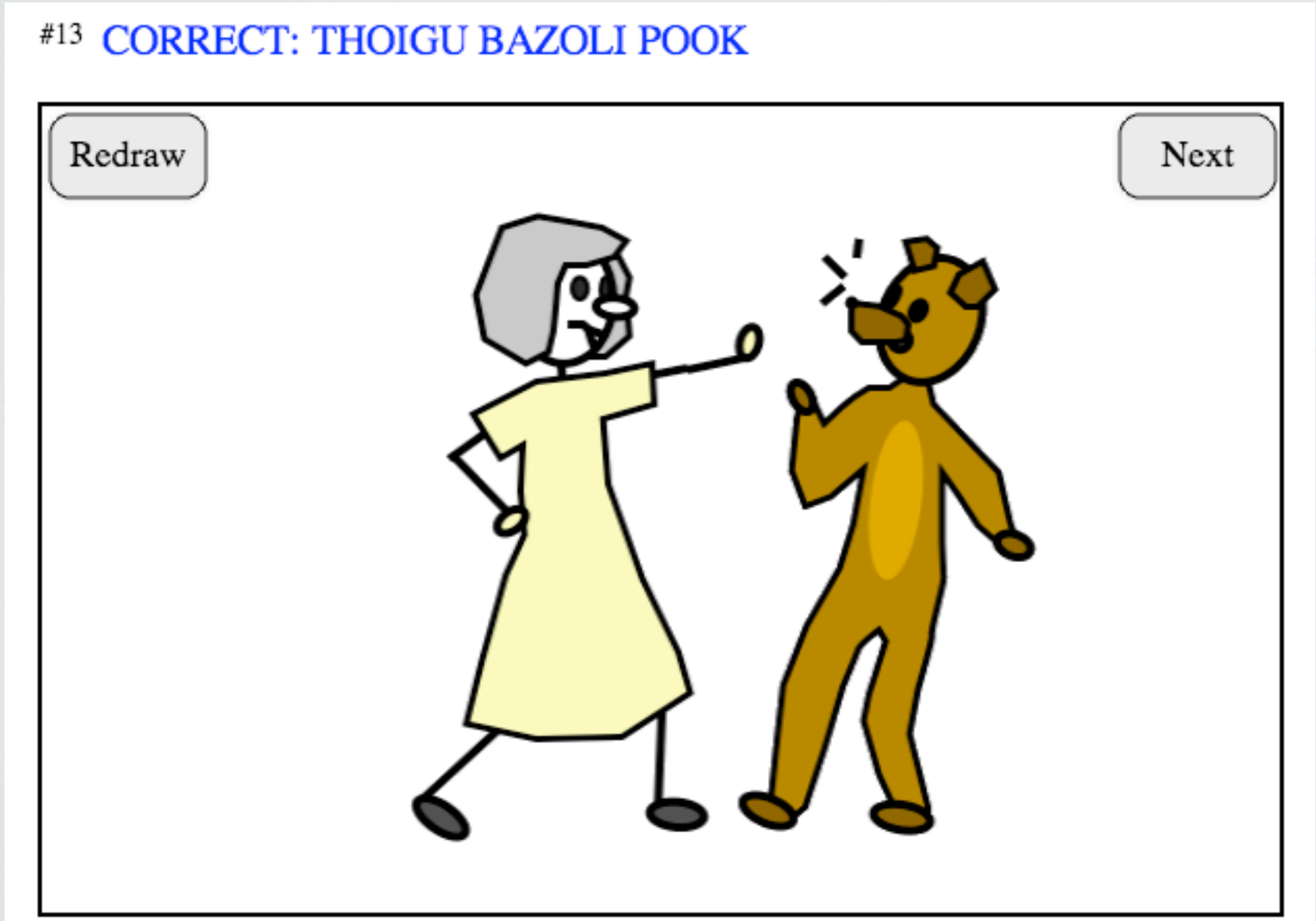
6 verbs (give, throw, kick, eat, carry, hit)

1 particle indicating recipient of an action

SOV word order

NEGATIVE EXAMPLES: THE EXPERIMENT

Presented people with an artificial language



6 agents
(boy, girl, man, lady, bear, alien)

2 noun particles (one for humans, one for non-humans)

6 objects (apple, banana, hamburger, ball, flower, book)

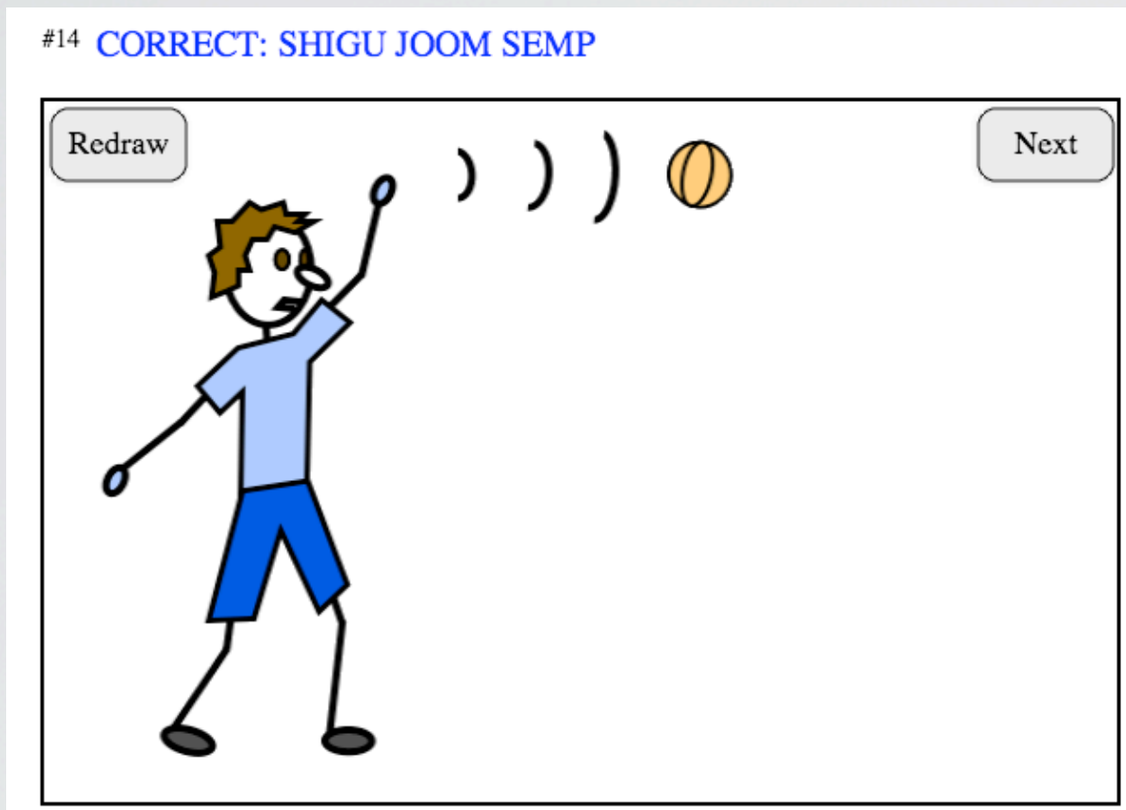
6 verbs (give, throw, kick, eat, carry, hit)

1 particle indicating recipient of an action

SOV word order

EXPERIMENT STRUCTURE

Positive examples

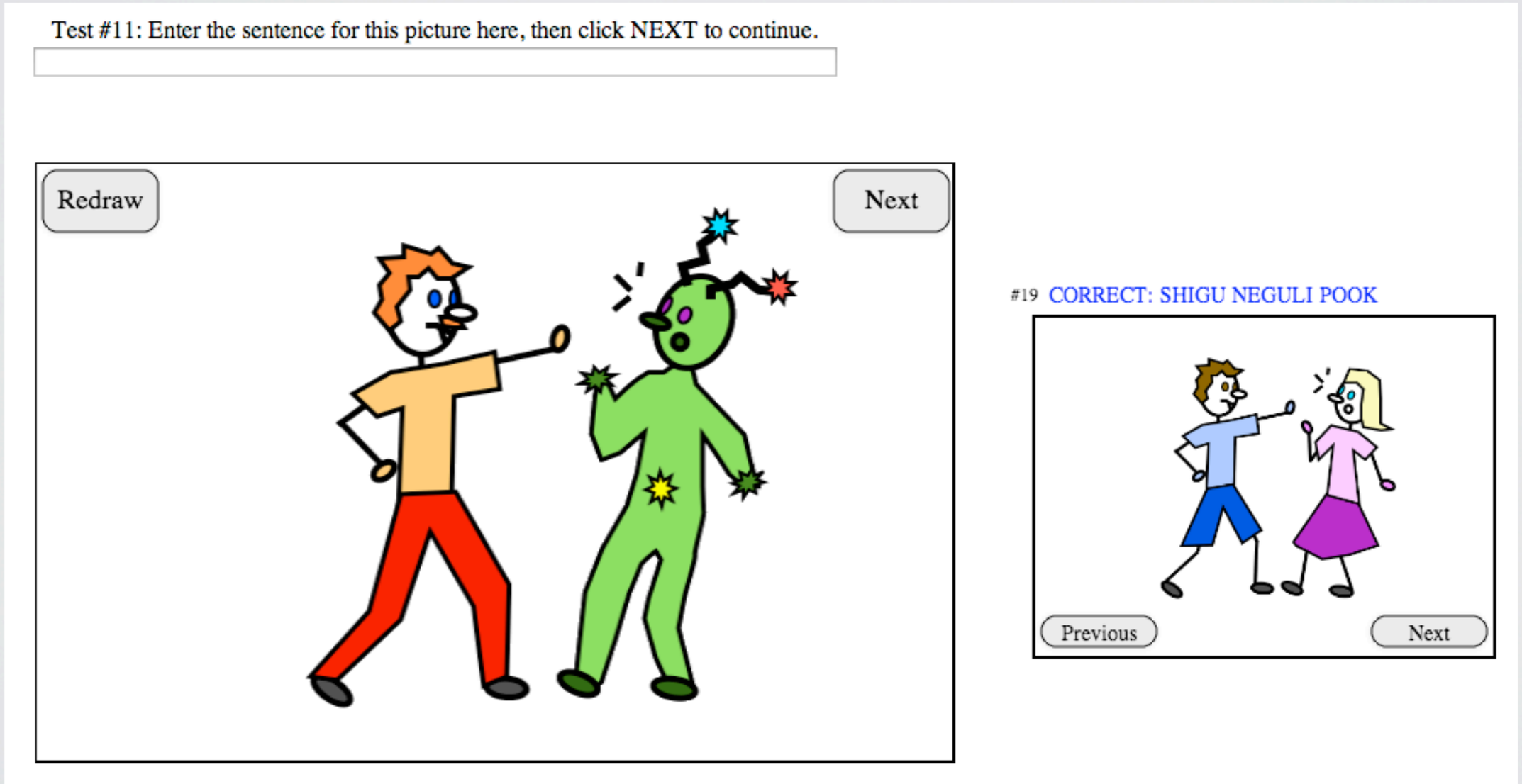


Negative examples

Block 1	Block 2	Block 3	Block 4
10 sentences train 10 sentences test	10 sentences train 10 sentences test	10 sentences train 10 sentences test	10 sentences train 10 sentences test

EXPERIMENT STRUCTURE

Test: To make it not about memory, on test people could go through all the previous sentences they had seen. No feedback was given.

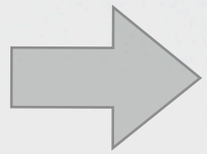


CONDITIONS

Varied how many positive and negative examples each participant saw

100% positive, 0% negative

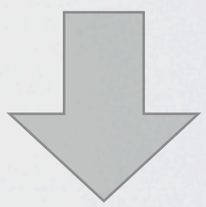
50% positive, 50% negative



100% positive, 0% negative

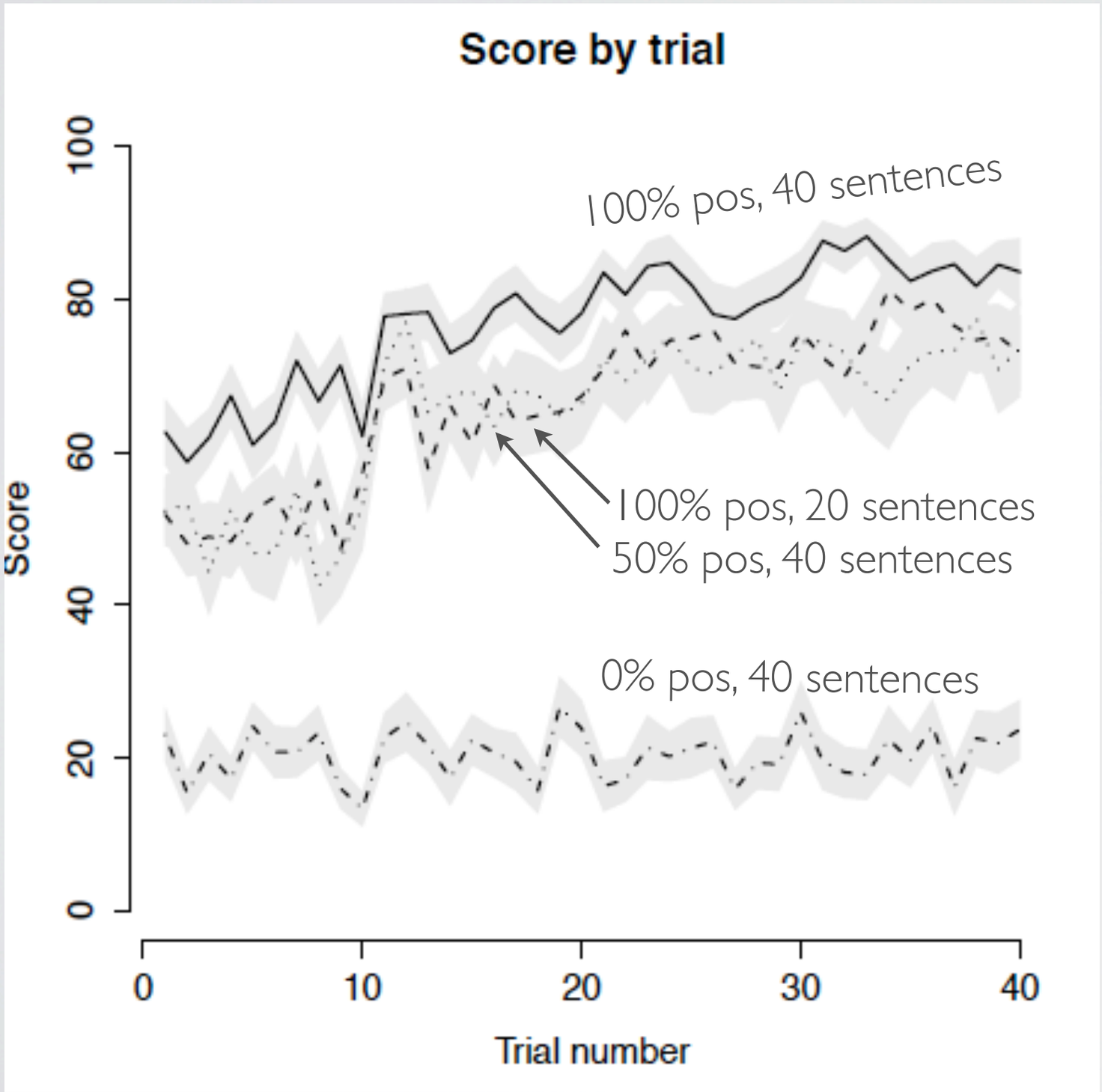
0% positive, 100% negative

40 total sentences



20 total sentences

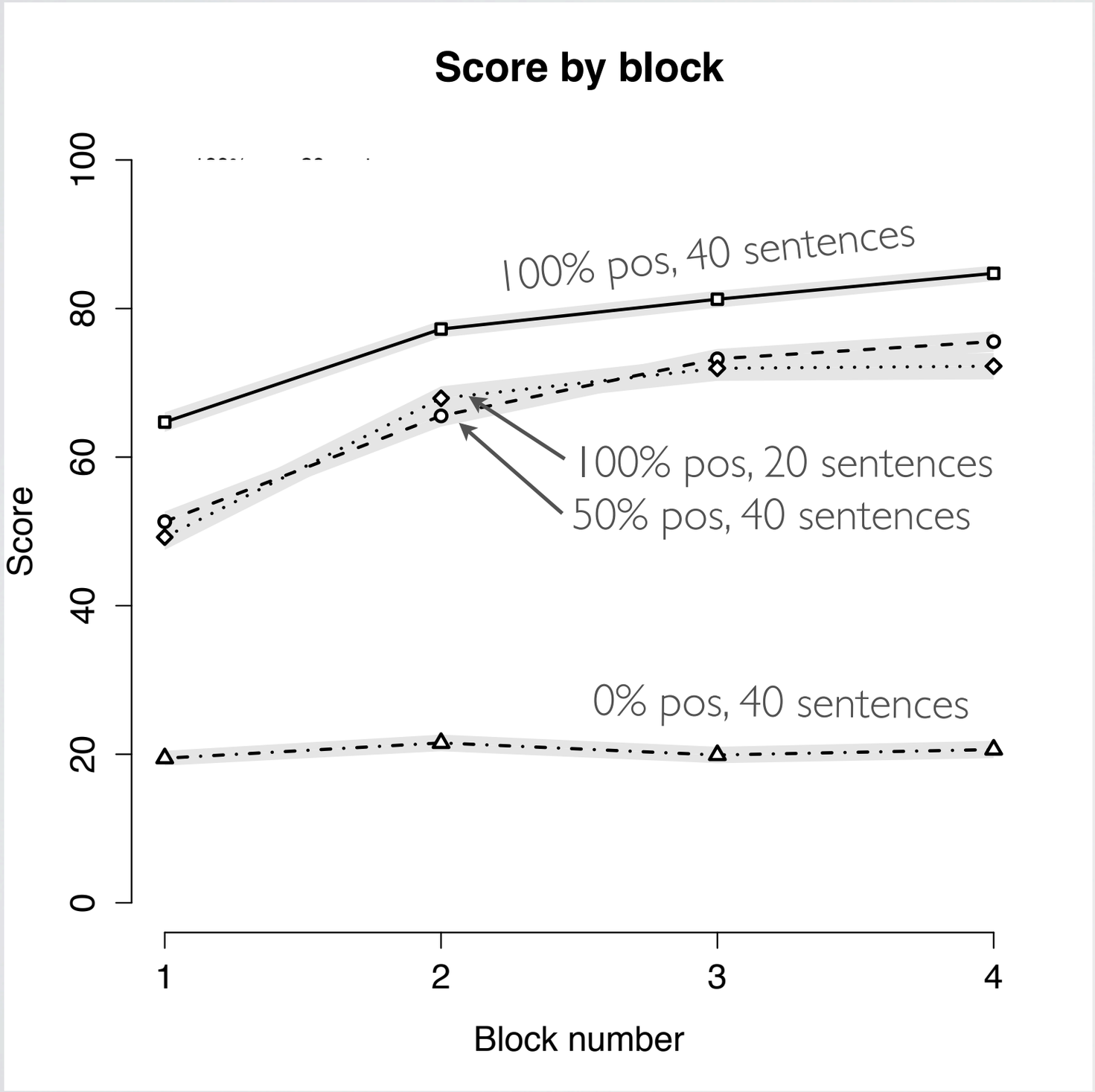
NEGATIVE EVIDENCE DIDN'T HELP AT ALL



$$\text{score} = \frac{100(w_c + w_{co} + m_c + m_{co})}{2w + 2m}$$

w_c = # words correct
w_{co} = # words in correct order
m_c = # morphemes correct
m_{co} = # morphemes in correct order
w = total # words
m = total # morphemes

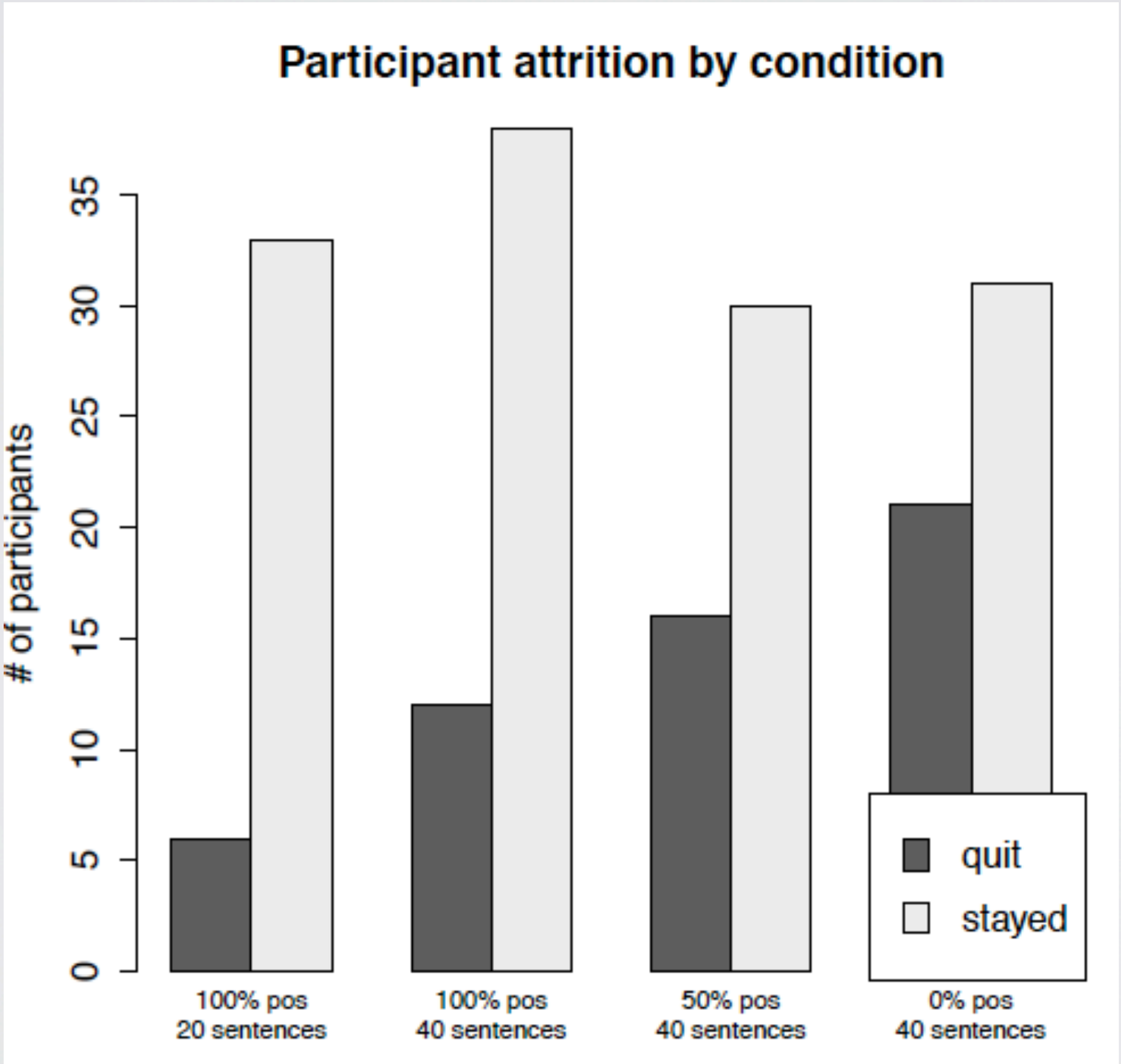
NEGATIVE EVIDENCE DIDN'T HELP AT ALL



$$\text{score} = \frac{100(w_c + w_{co} + m_c + m_{co})}{2w + 2m}$$

w_c = # words correct
w_{co} = # words in correct order
m_c = # morphemes correct
m_{co} = # morphemes in correct order
w = total # words
m = total # morphemes

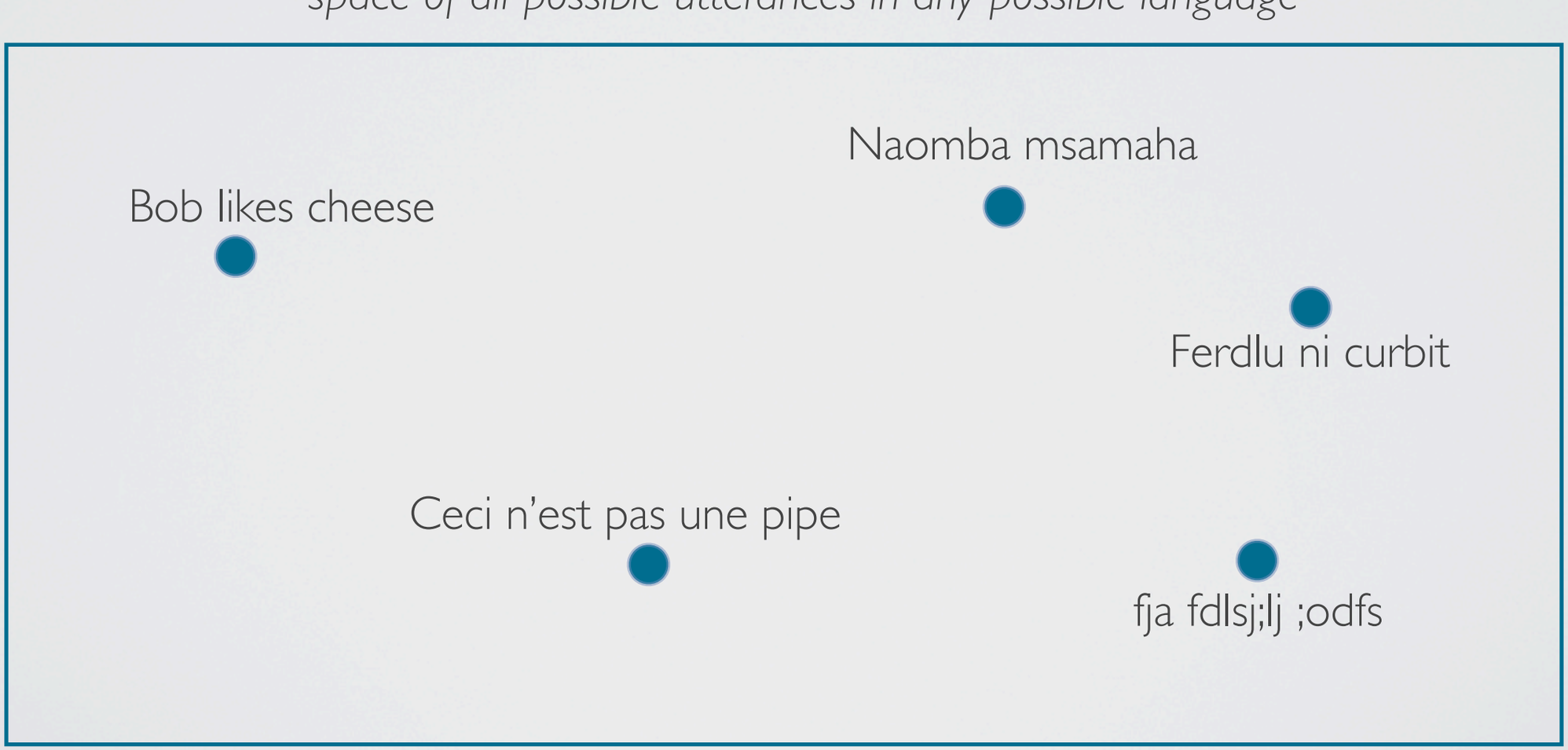
IT ALSO MADE PEOPLE VERY FRUSTRATED



WHY ARE NEGATIVE EXAMPLES USELESS?

In a nutshell: because languages are *sparse*

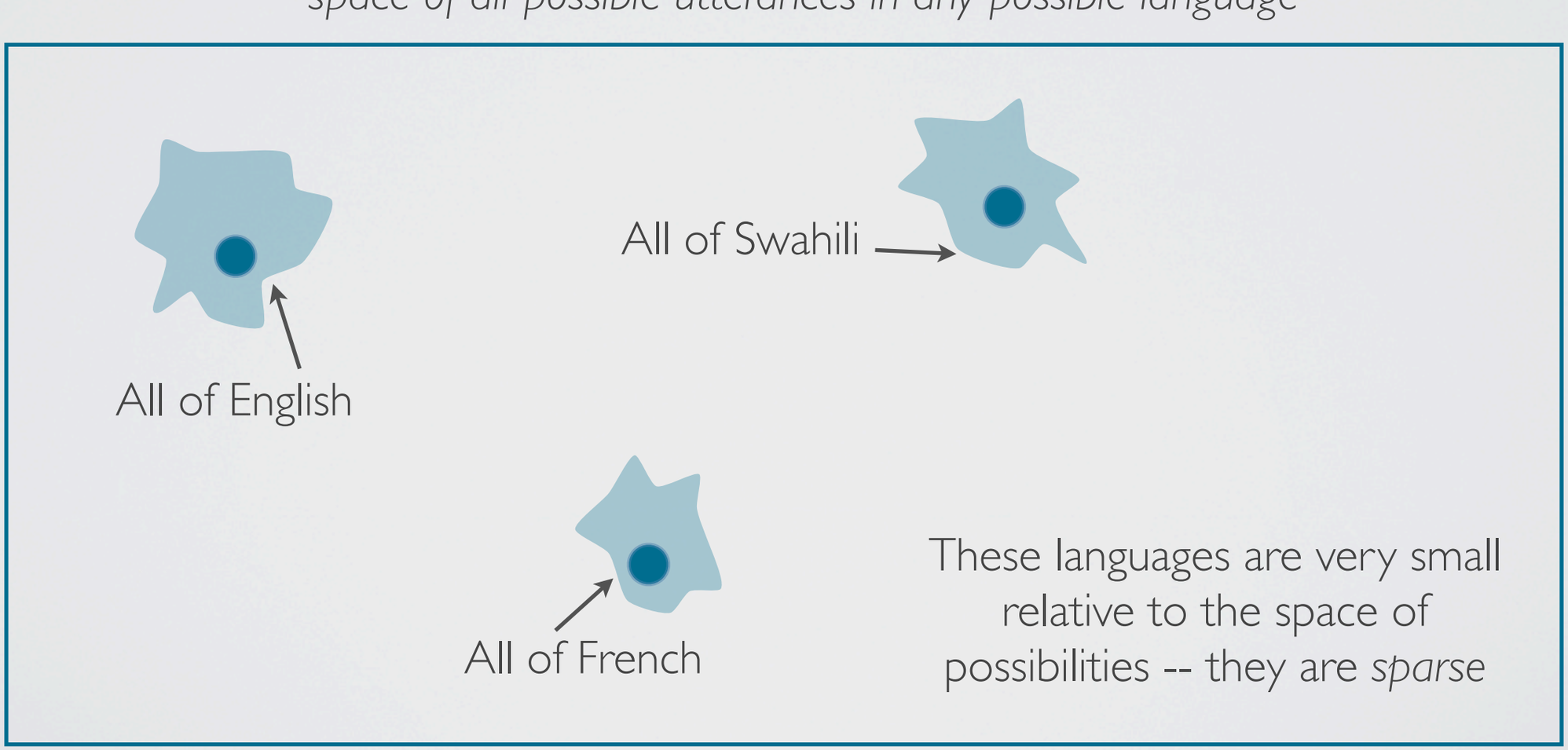
- Sparse hypotheses are true for a minority of entities
  - Few possible sentences are grammatical in a language
  - Few animals are PETS



WHY ARE NEGATIVE EXAMPLES USELESS?

In a nutshell: because languages are *sparse*

- Sparse hypotheses are true for a minority of entities
  - Few possible sentences are grammatical in a language
  - Few animals are PETS

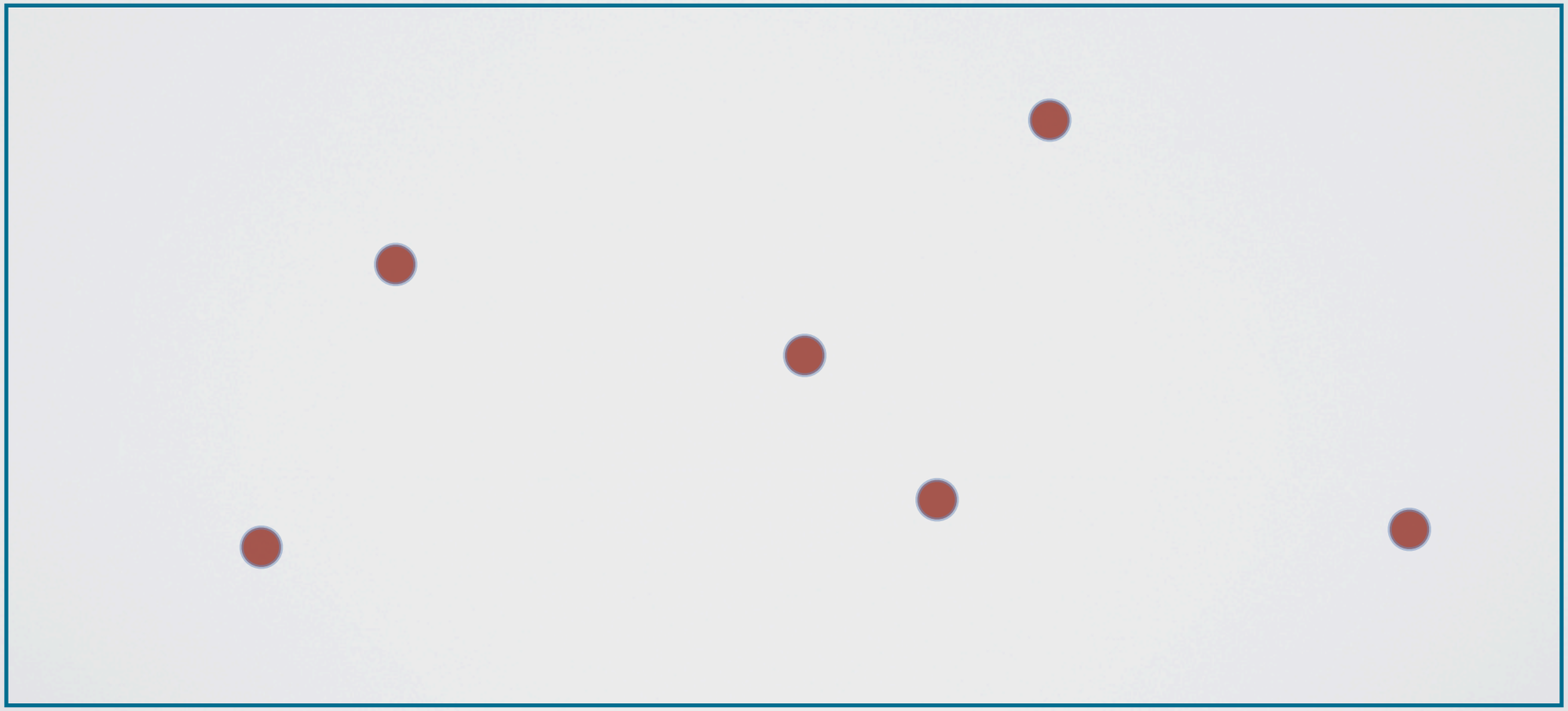


WHY ARE NEGATIVE EXAMPLES USELESS?

In a nutshell: because languages are *sparse*

When a hypothesis is sparse, positive evidence (about what sentences are grammatical) is vastly more informative than negative evidence (about what sentences are not)

space of all possible utterances in any possible language



WHY ARE NEGATIVE EXAMPLES USELESS?

In a nutshell: because languages are *sparse*

When a hypothesis is sparse, positive evidence (about what sentences are grammatical) is vastly more informative than negative evidence (about what sentences are not)

space of all possible utterances in any possible language

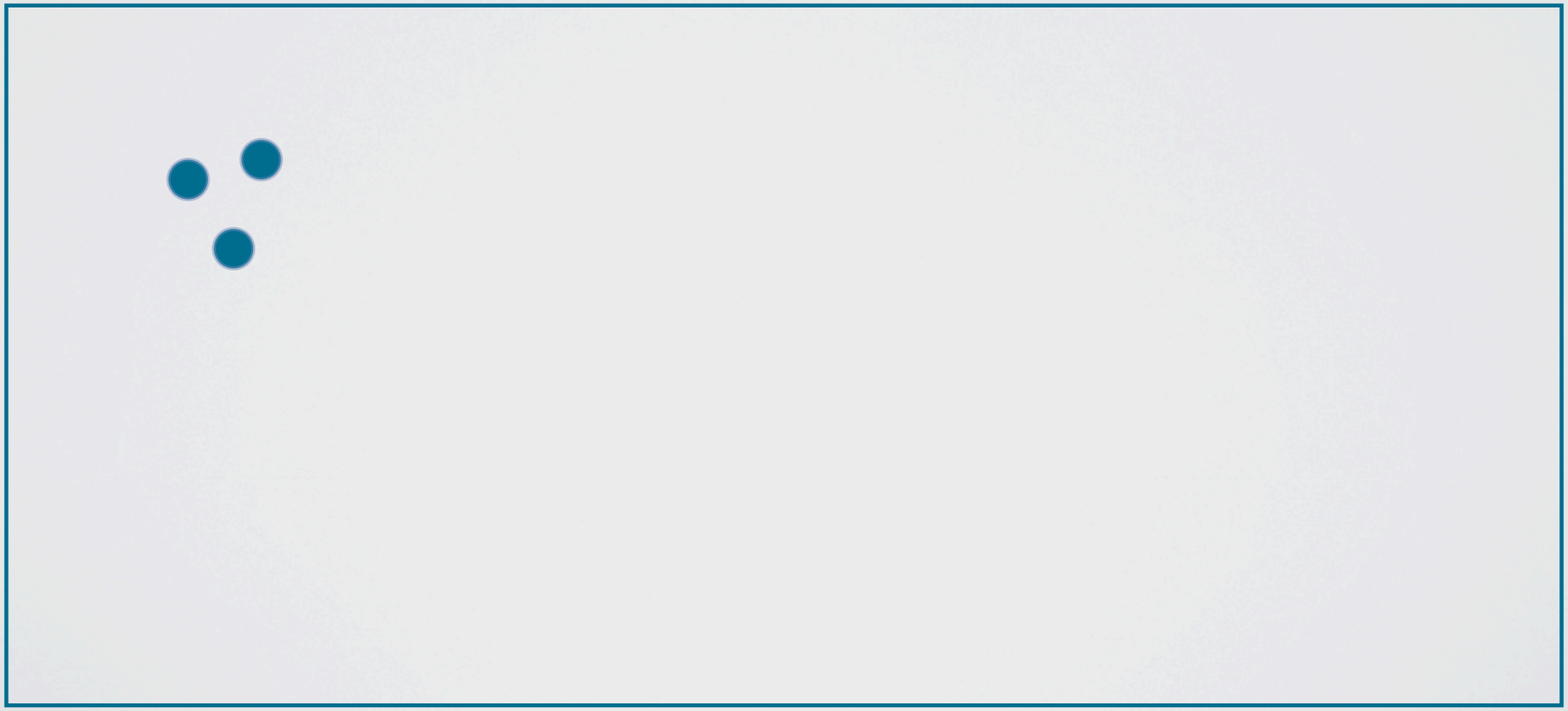
by contrast...

WHY ARE NEGATIVE EXAMPLES USELESS?

In a nutshell: because languages are *sparse*

When a hypothesis is sparse, positive evidence (about what sentences are grammatical) is vastly more informative than negative evidence (about what sentences are not)

space of all possible utterances in any possible language



WHY ARE NEGATIVE EXAMPLES USELESS?

When a hypothesis is sparse, positive evidence (about what sentences are grammatical) is vastly more informative than negative evidence (about what sentences are not)

We can prove this mathematically....

To learn the true hypothesis fastest, one seeks to minimize the expected entropy of the distribution over the *m* possible hypotheses.

Straightforward to show that this depends on the number *m(x)* of not-yet-falsified hypotheses that are consistent with the data x:

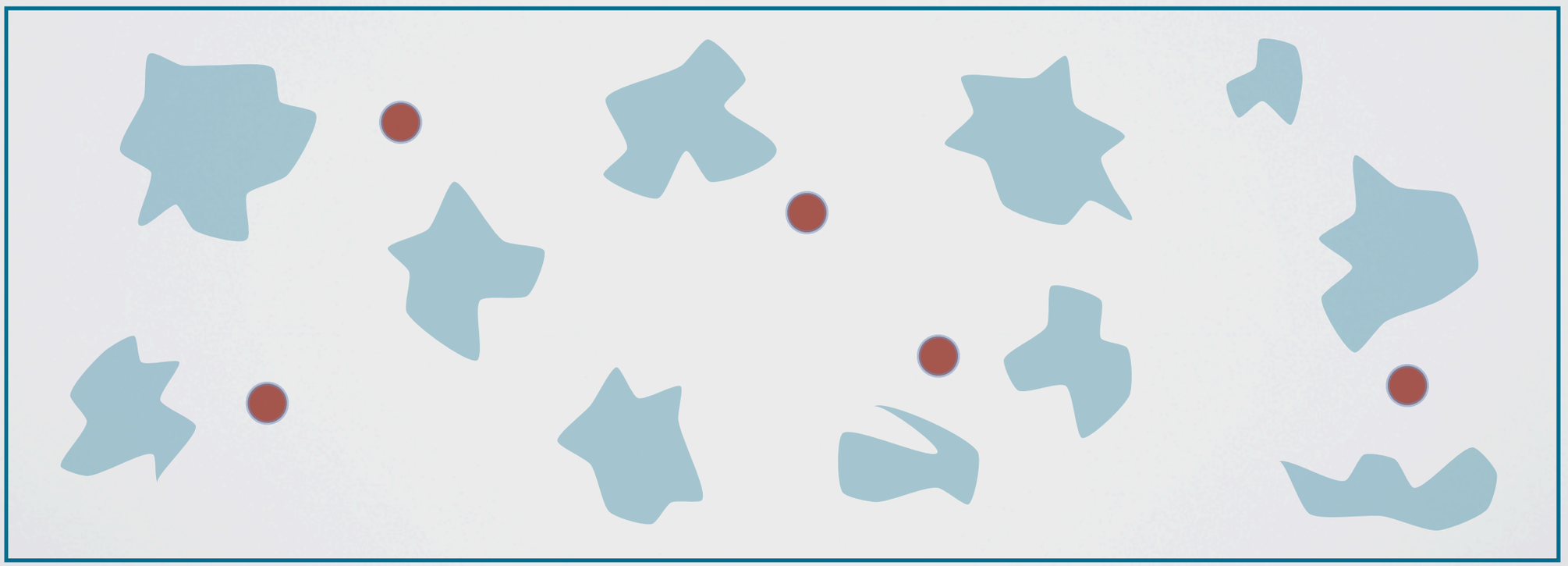
$$
\begin{aligned}
E[H(\mathcal{O}\,|\,x)] &= -E_{\mathcal{O}(x)}\left[\sum_i \Pr(\mathcal{O}\to r_i\,|\,\mathcal{O}(x))\ln\Pr(\mathcal{O}\to r_i\,|\,\mathcal{O}(x))\right] \\
&= -\Pr(\mathcal{O}(x)=0)\left(\sum_i \Pr(\mathcal{O}\to r_i\,|\,\mathcal{O}(x)=0)\ln\Pr(\mathcal{O}\to r_i\,|\,\mathcal{O}(x)=0)\right) \\
&\quad -\Pr(\mathcal{O}(x)=1)\left(\sum_i \Pr(\mathcal{O}\to r_i\,|\,\mathcal{O}(x)=1)\ln\Pr(\mathcal{O}\to r_i\,|\,\mathcal{O}(x)=1)\right) \\
&= -\frac{m(\neg x)}{m}\left(\sum_{i|r_i(x)=0}\frac{1}{m(\neg x)}\ln\frac{1}{m(\neg x)}\right) - \frac{m(x)}{m}\left(\sum_{i|r_i(x)=1}\frac{1}{m(x)}\ln\frac{1}{m(x)}\right) \\
&= \frac{m(\neg x)\ln m(\neg x) + m(x)\ln m(x)}{m}
\end{aligned}
$$

Navarro & Perfors, 2010

WHY ARE NEGATIVE EXAMPLES USELESS?

When a hypothesis is sparse, positive evidence (about what sentences are grammatical) is vastly more informative than negative evidence (about what sentences are not)

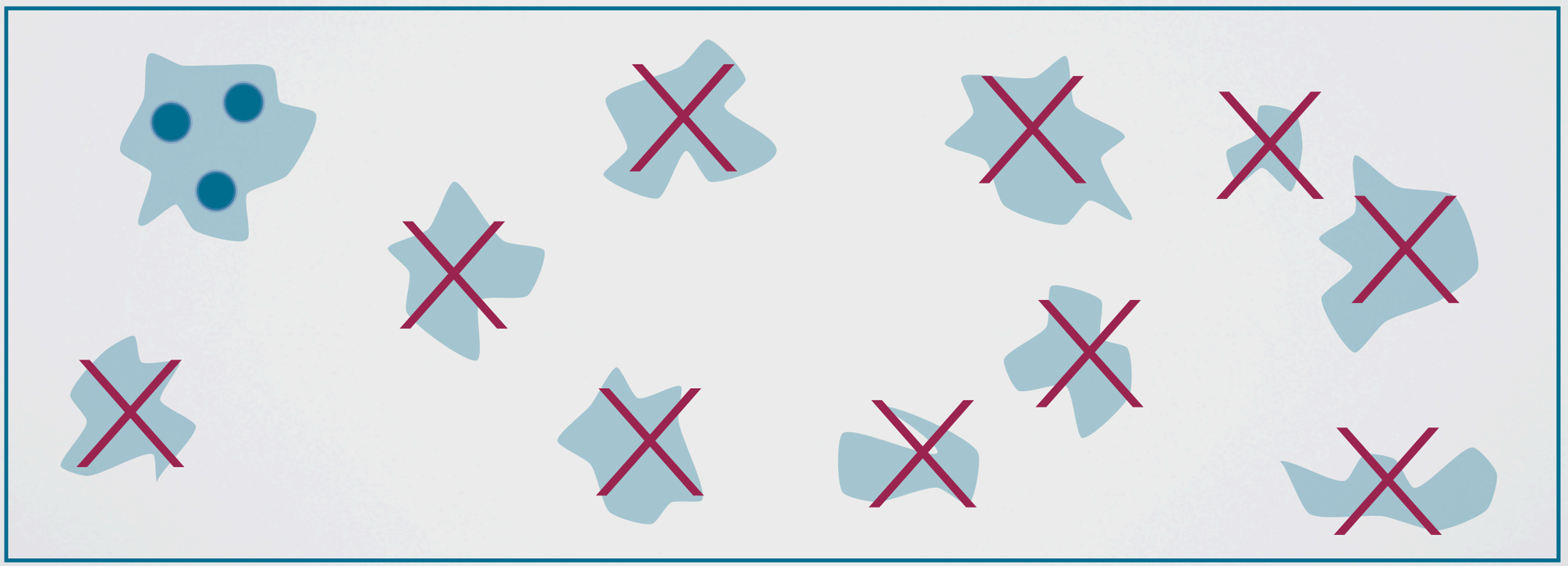
When most of the hypotheses are sparse, positive evidence eliminates most of them, and thus is highly informative. Negative evidence eliminates very few, and thus is very uninformative.



WHY ARE NEGATIVE EXAMPLES USELESS?

When a hypothesis is sparse, positive evidence (about what sentences are grammatical) is vastly more informative than negative evidence (about what sentences are not)

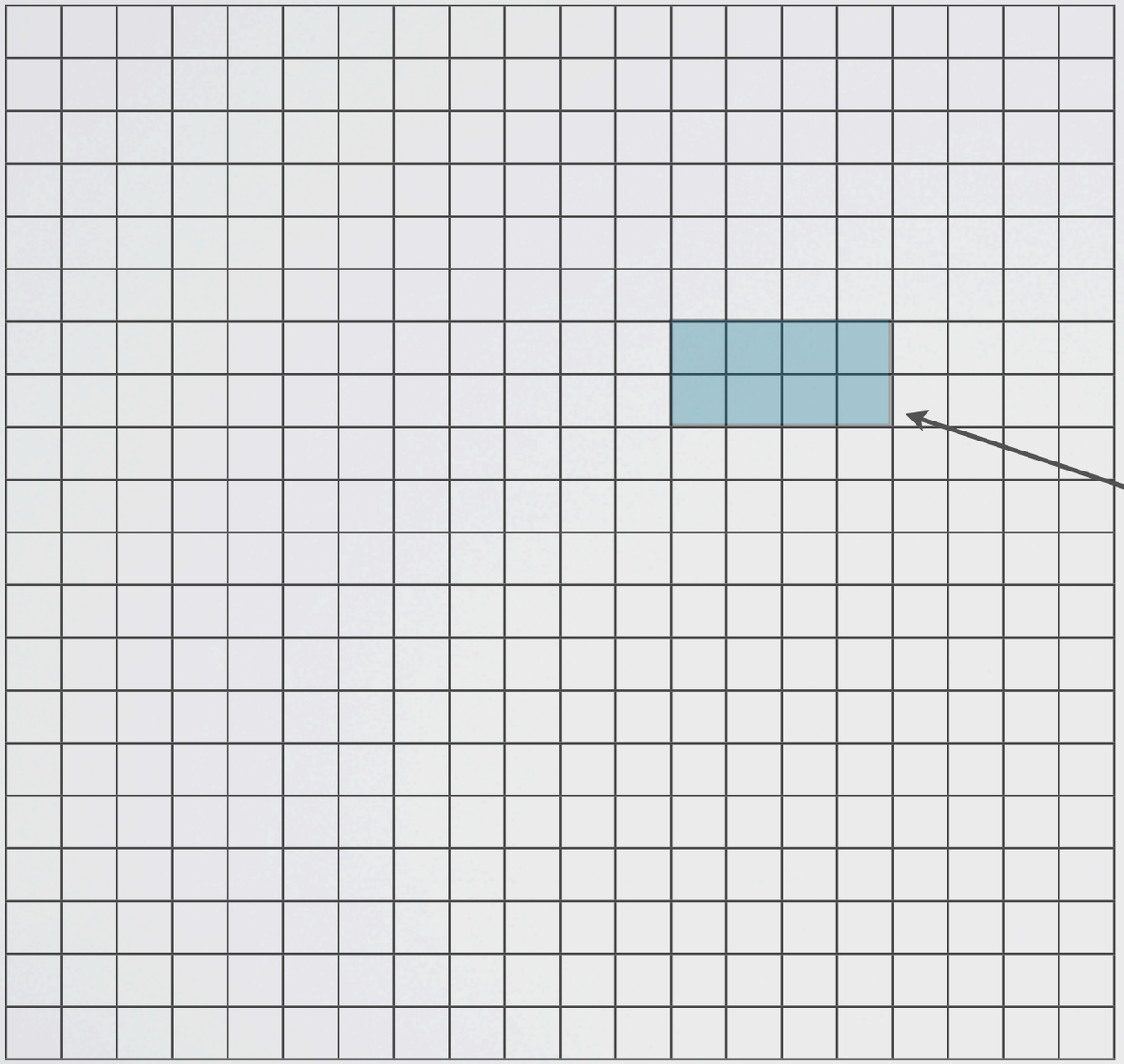
When most of the hypotheses are sparse, positive evidence eliminates most of them, and thus is highly informative. Negative evidence eliminates very few, and thus is very uninformative.



We can confirm this (and explore more thoroughly what it means) via simulation...

SIMULATION RESULTS

Simulations take place in a "rectangle world"



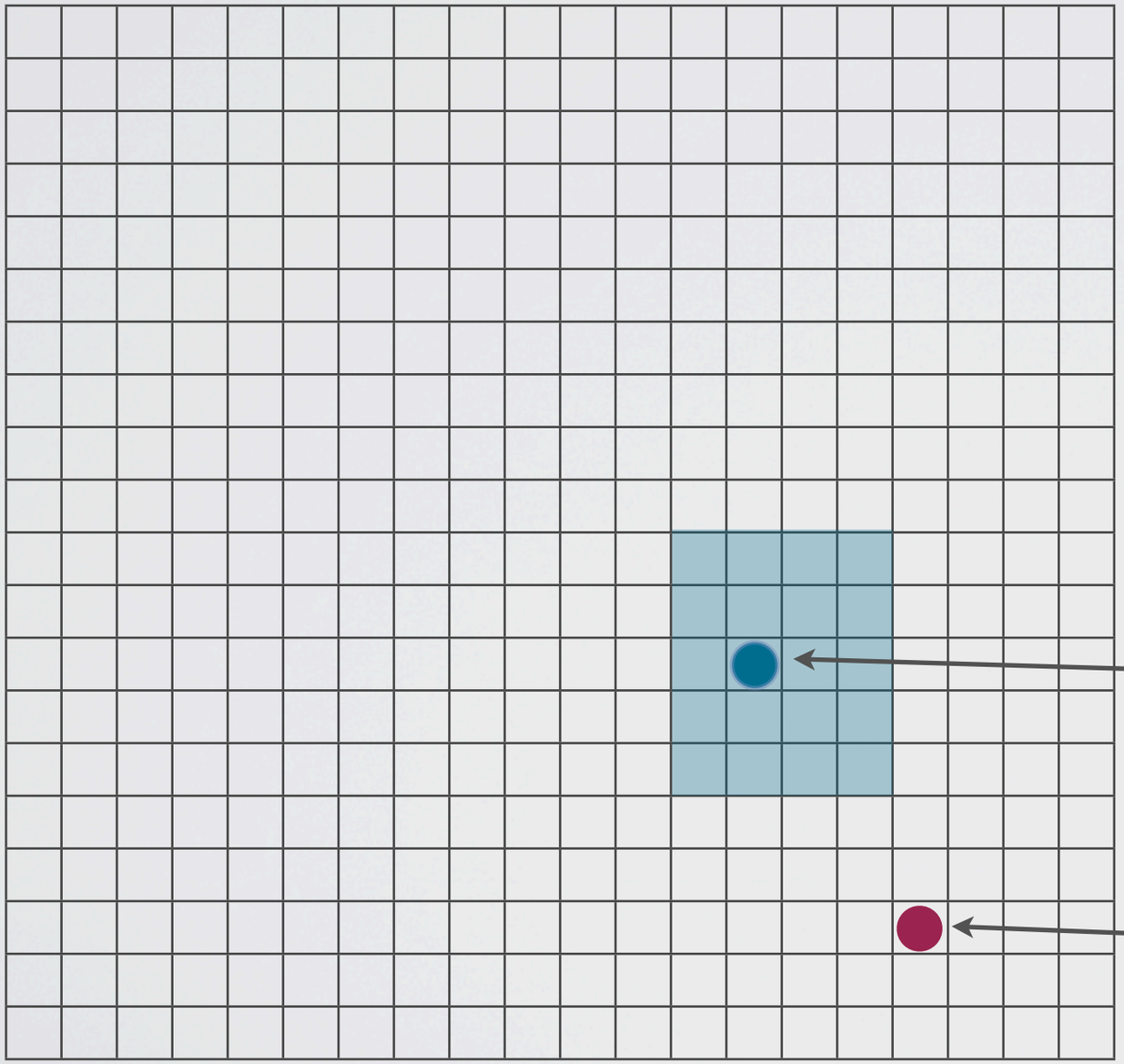
The world is a 2D grid (here of size 20x20)

Each hypothesis is a rectangle in the grid. This one has a sparsity of 2% (i.e., 8/400)

The hypothesis space is the set of all possible rectangles (a huge number!)

SIMULATION RESULTS

Simulations take place in a "rectangle world"



Assume a learner who considers possible hypotheses, and calculates their probability according to Bayes' Rule.

Prior is uniform for all hypotheses

If positive, likelihood is relative to the size of the hypothesis

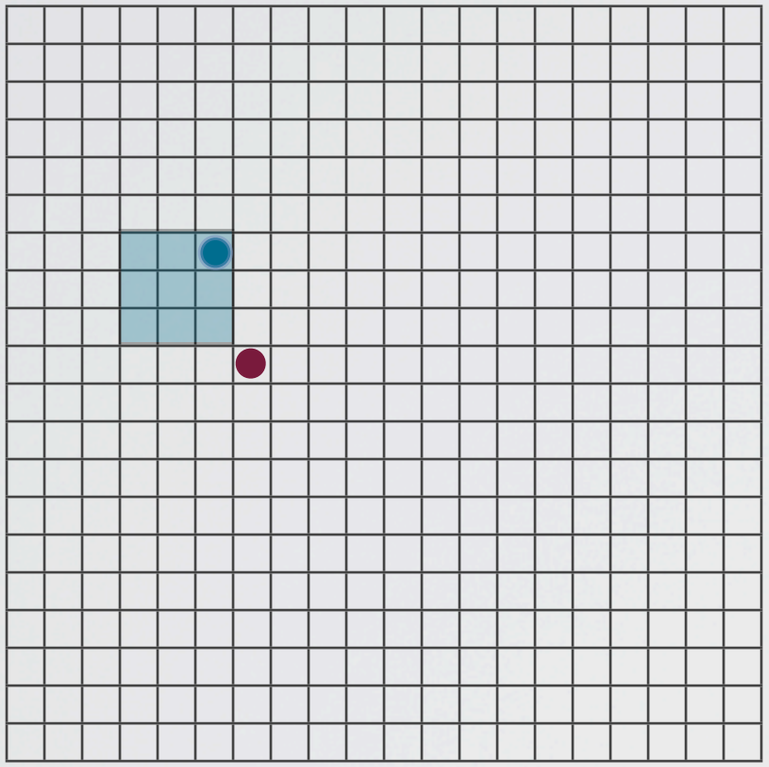
$p(d|h) = 1/20 = 5\%$

If negative, likelihood is relative to size of world minus hypothesis

$p(d|h) = 1/(400\text{-}20) = 0.26\%$

SIMULATION RESULTS

Goal: evaluate which data is most useful to a learner



- $p(h_c|d) = 0.48$
- $p(h_c|d) = 0.34$

1. Randomly select one hypothesis to be the "true" language

2. For each possible positive example, calculate how much receiving it would increase the learner's probability of the *correct* hypothesis.

3. Choose the best positive example.

4. Repeat the process with all possible negative examples.

5. Select which of the best positive and negative examples will help the learner the most.

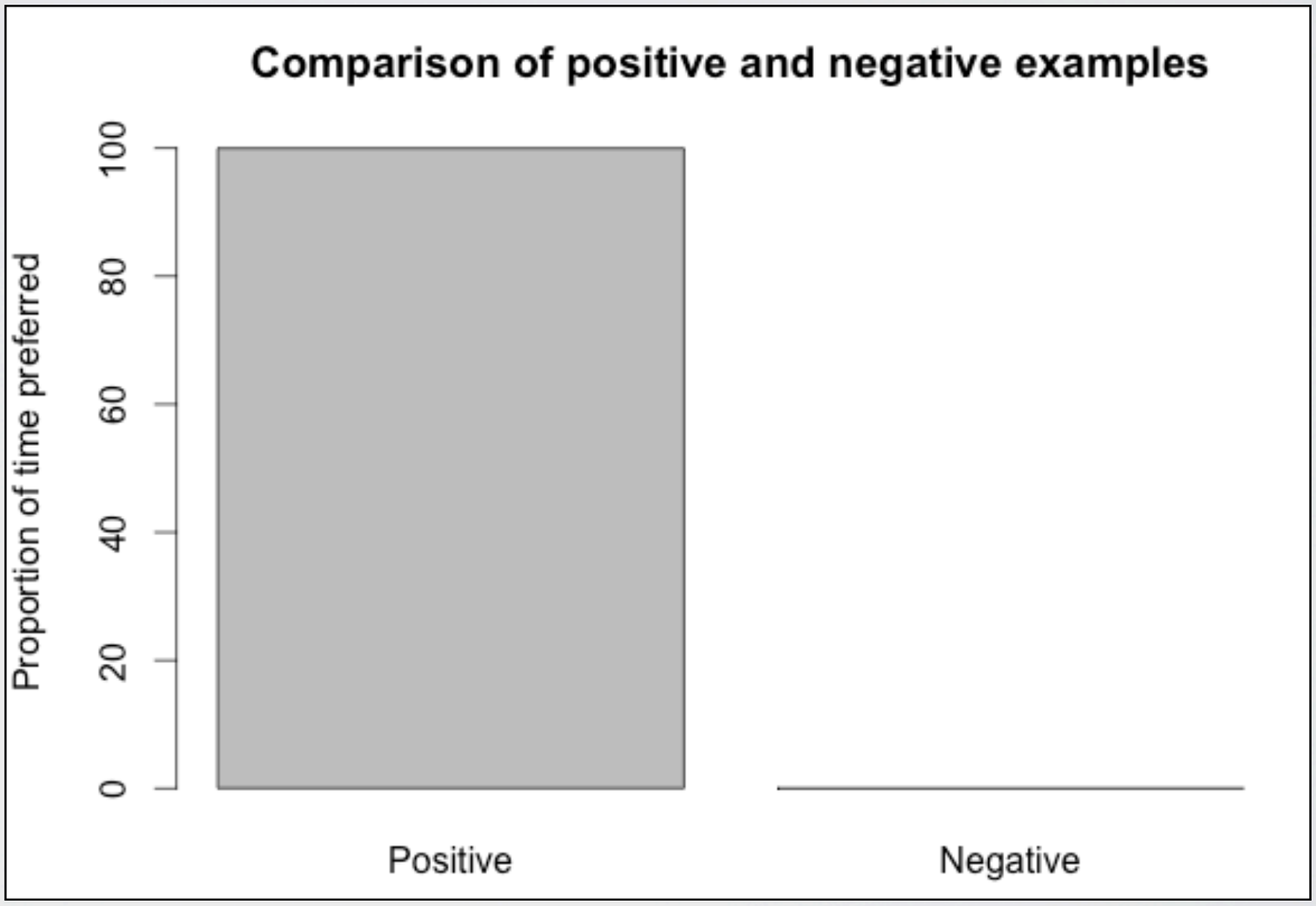
SIMULATION RESULTS

Positive examples were *always* preferred!

Varied, with thousands of runs each:

Size of world:
15x15
20x20
30x30

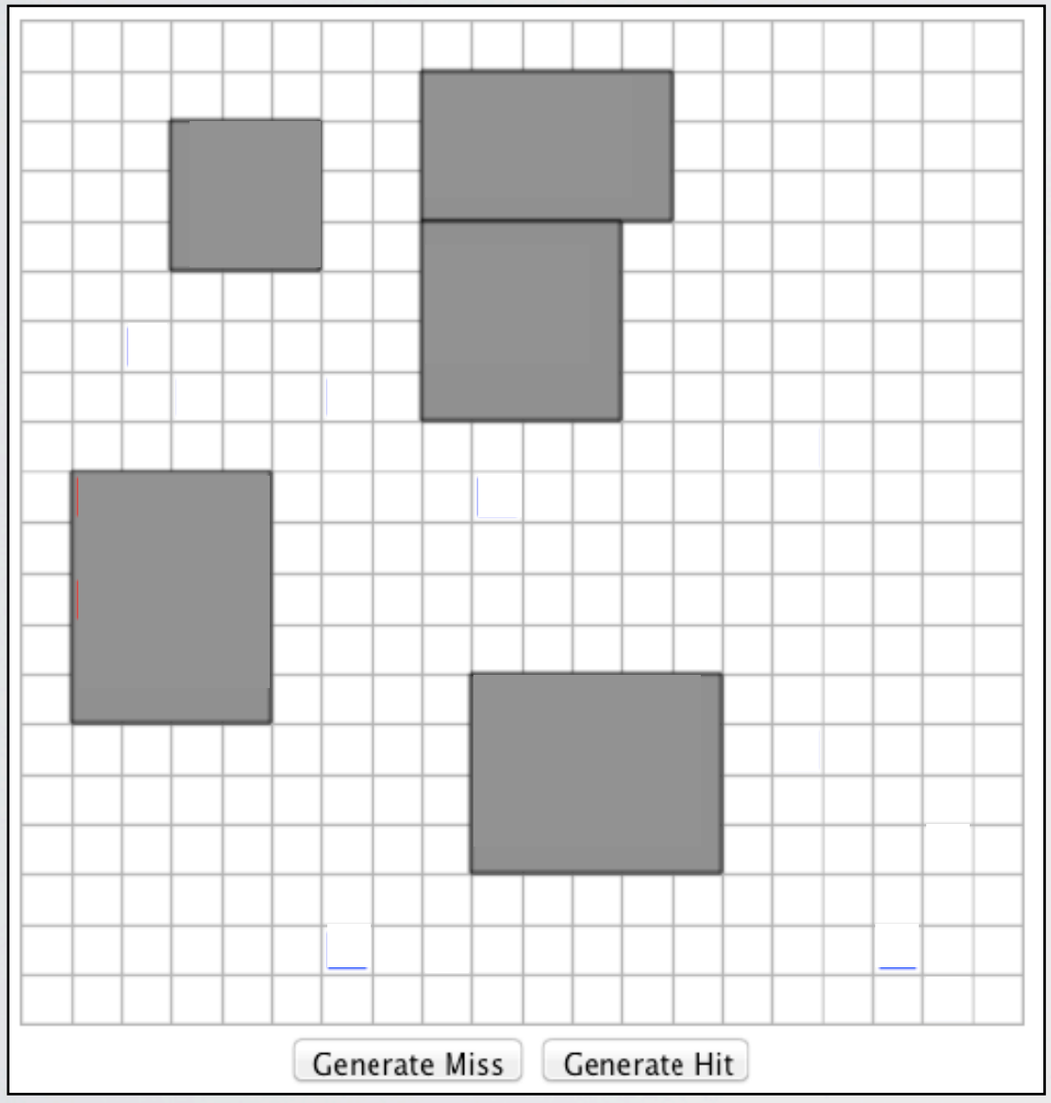
Sparsity:
0.15%
0.25%
0.5%
1%
2%
4%
8%
16%

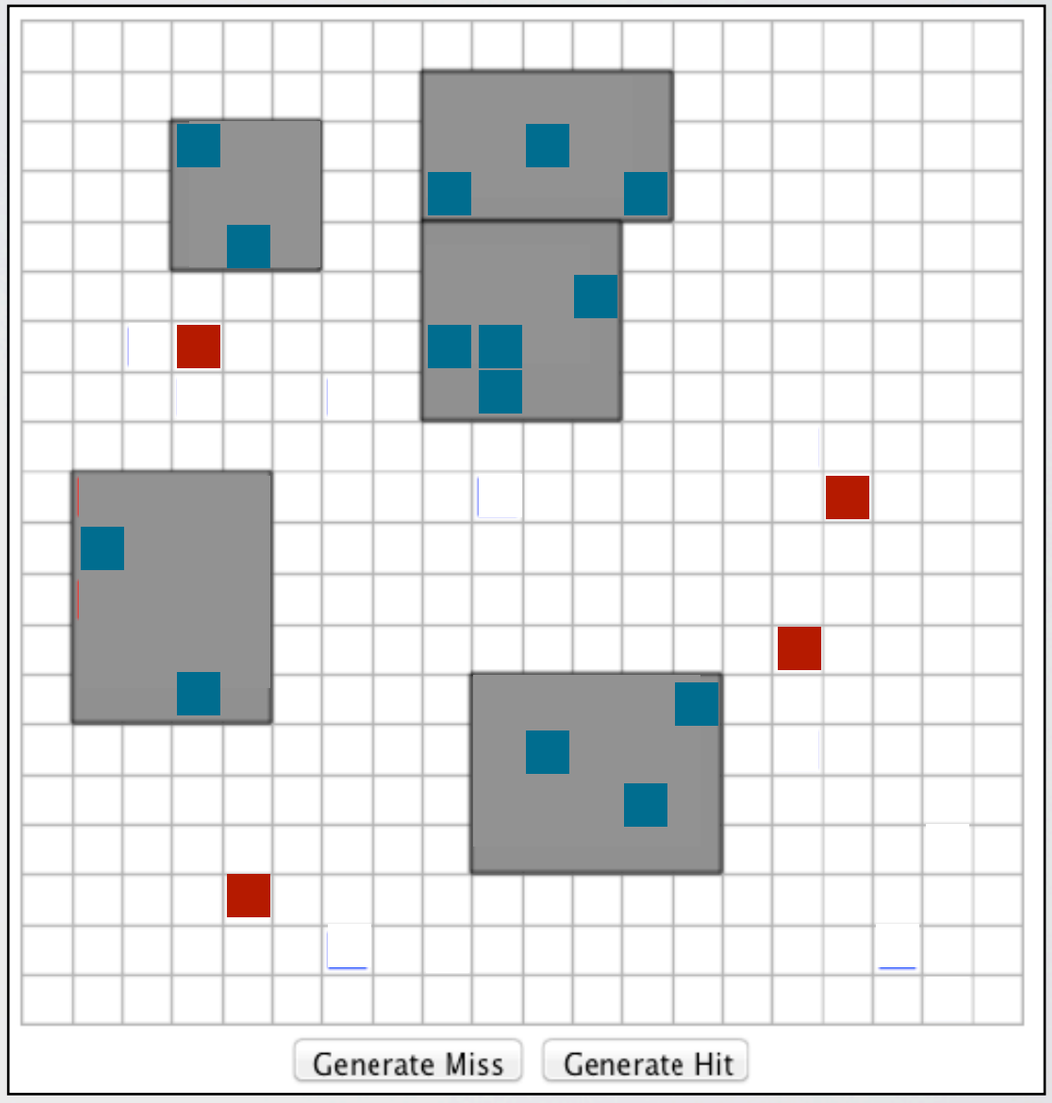


So, if sparsity matters so much...
do people realize this?

PEOPLE'S SENSITIVITY TO SPARSITY

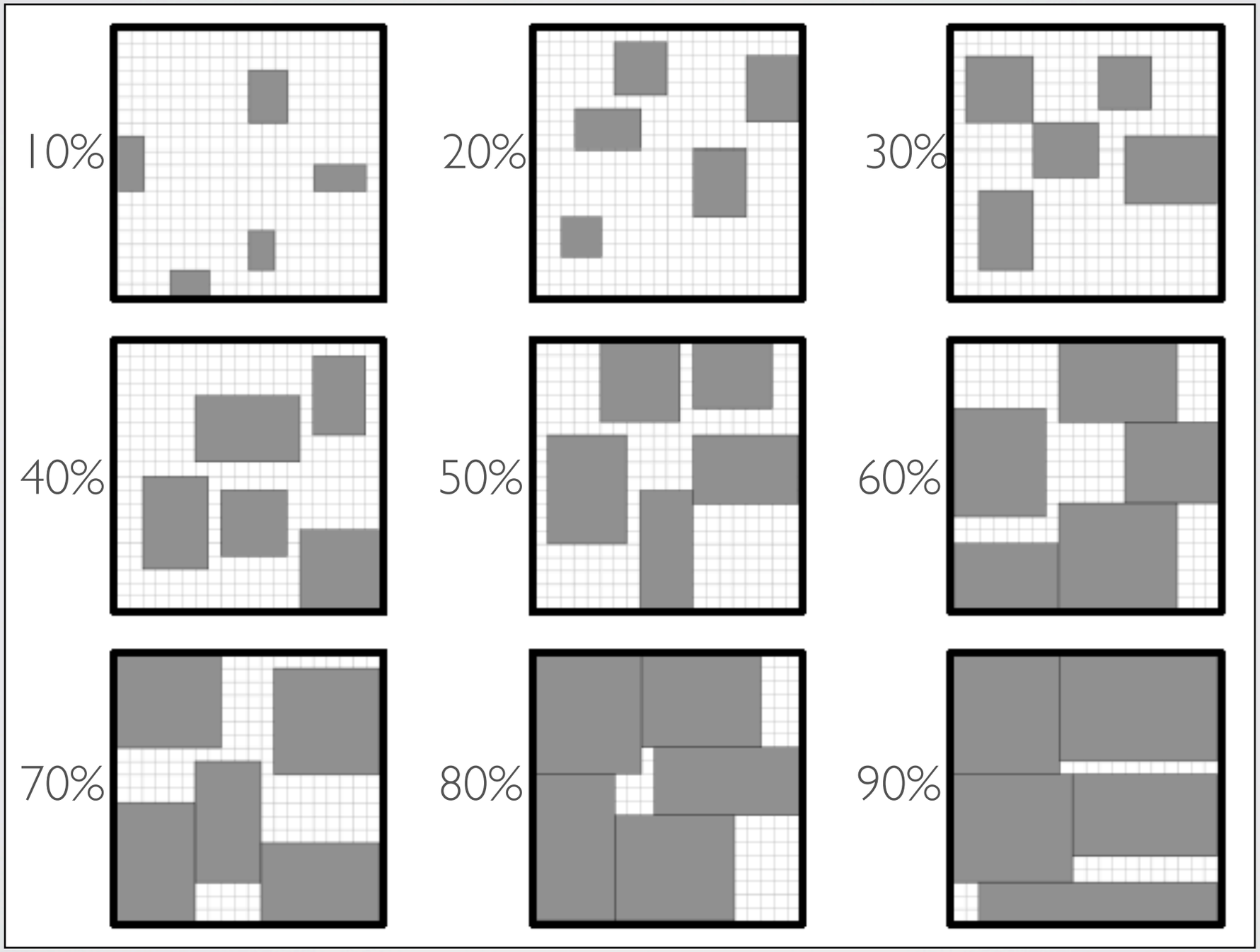
Battleships task: people had to find the location of hidden battleships by asking for hits (positive) or misses (negative)



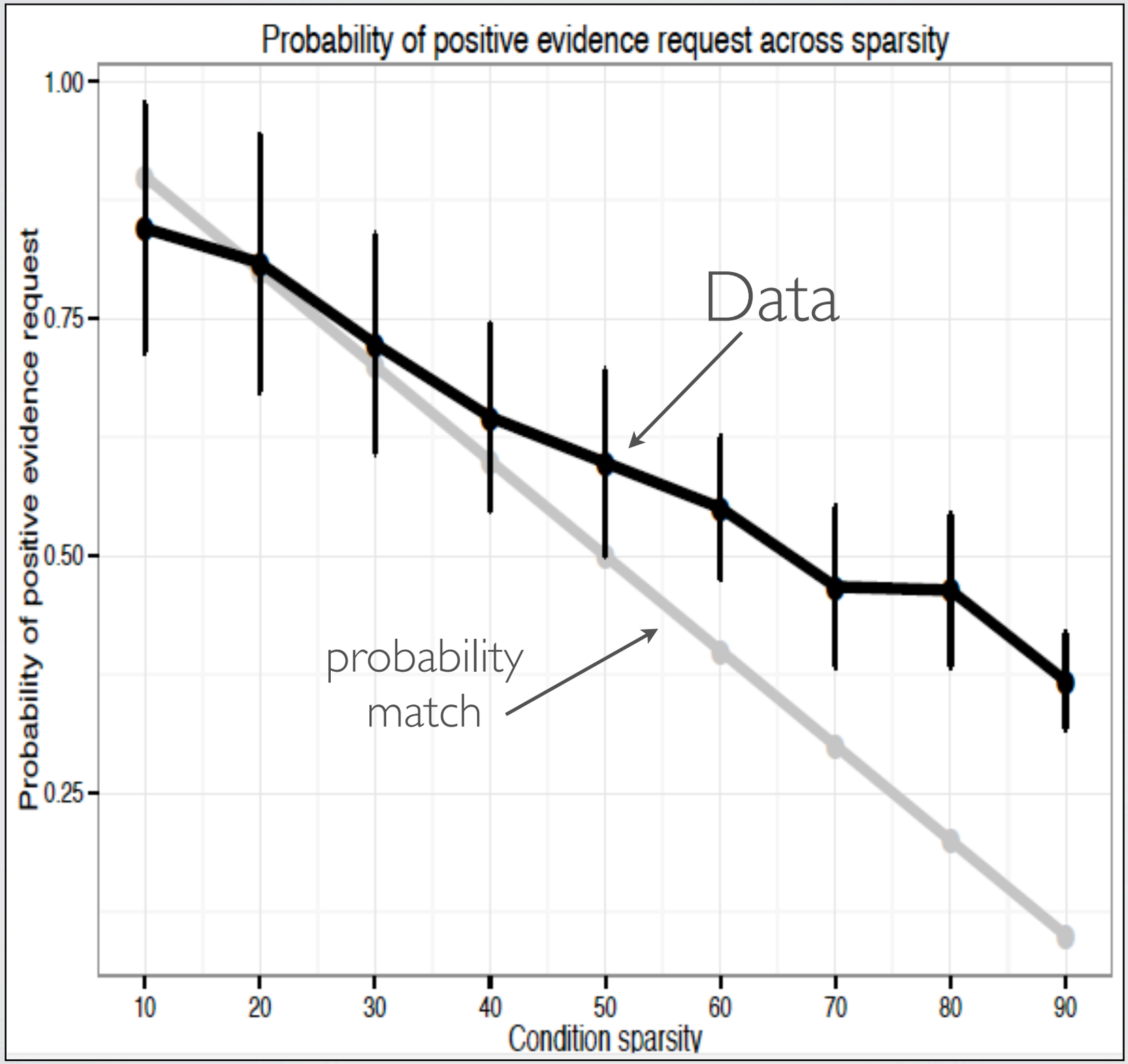


PEOPLE'S SENSITIVITY TO SPARSITY

Varied sparsity by varying the size of the ships



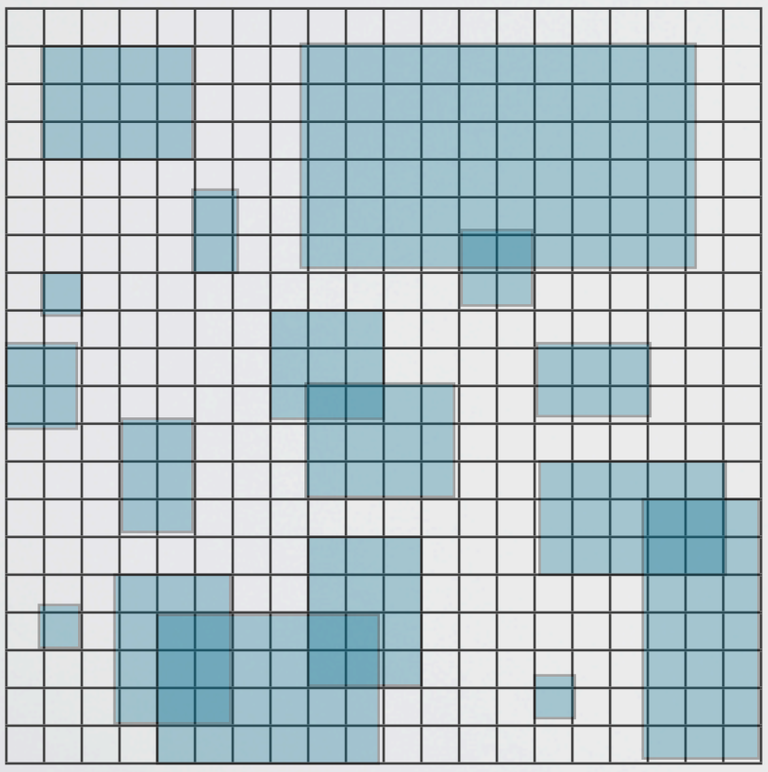
PEOPLE ARE SENSITIVE TO SPARSITY



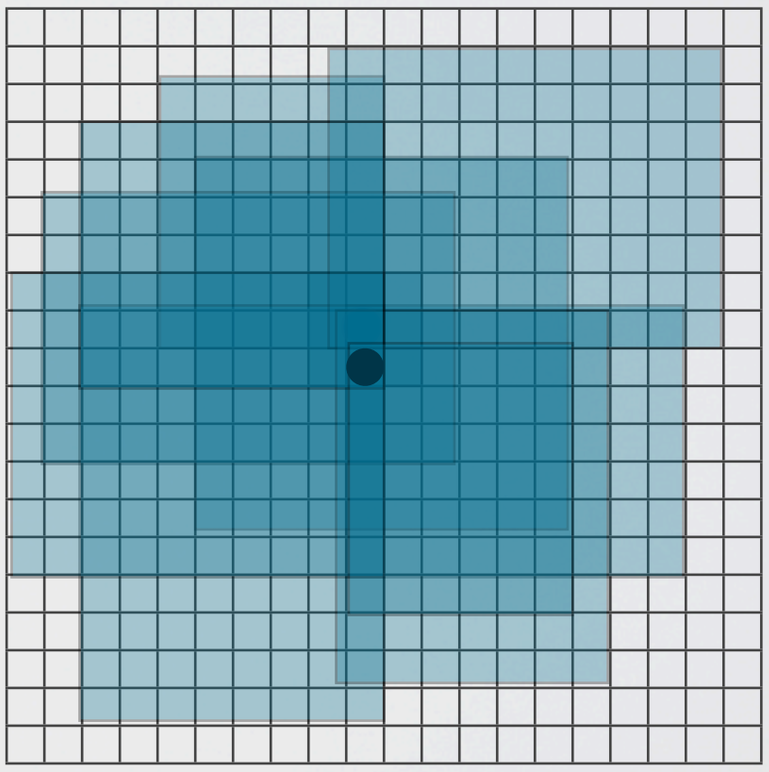
OKAY, BUT...

1. This is just the first data point. After some hypotheses have been eliminated, the nature of the hypothesis space changes. Shouldn't this change what kind of evidence is best?

Before:
many hypotheses,
many sparse

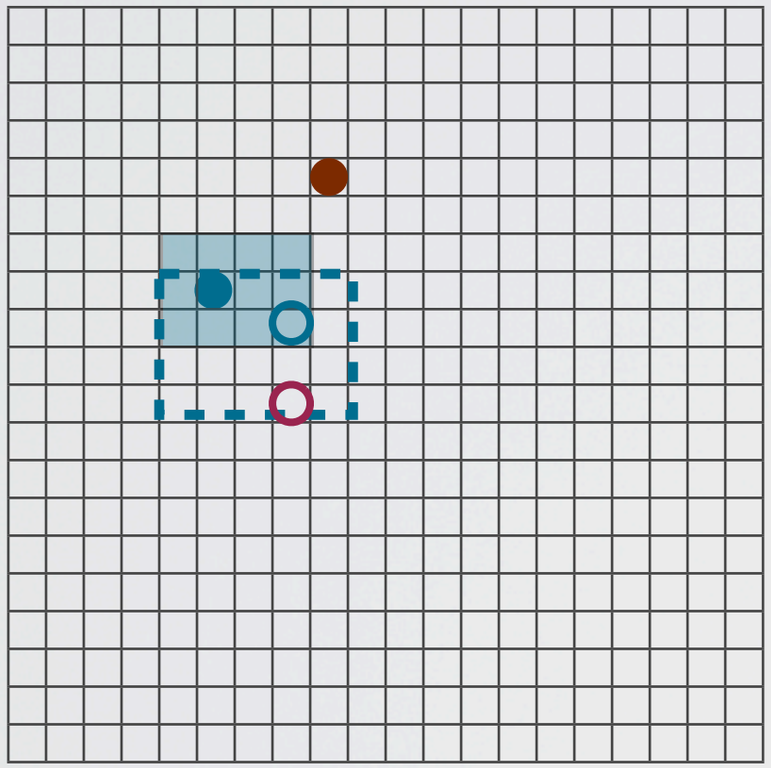


After one data point:
fewer hypotheses, all
overlapping, many not sparse



OKAY, BUT...

2. So far we've just looked at positive and negative examples. What about positive and negative *feedback*? Is this different?



Example generated by parent

- Positive: Sentence from the language

  Bob eats cheese

- Negative: Sentence not in the language

  Nixim mot ferdlu

Feedback to child

- Child's sentence is correct

  Yes, that's right!

- Child's sentence is wrong

  No, that's wrong!

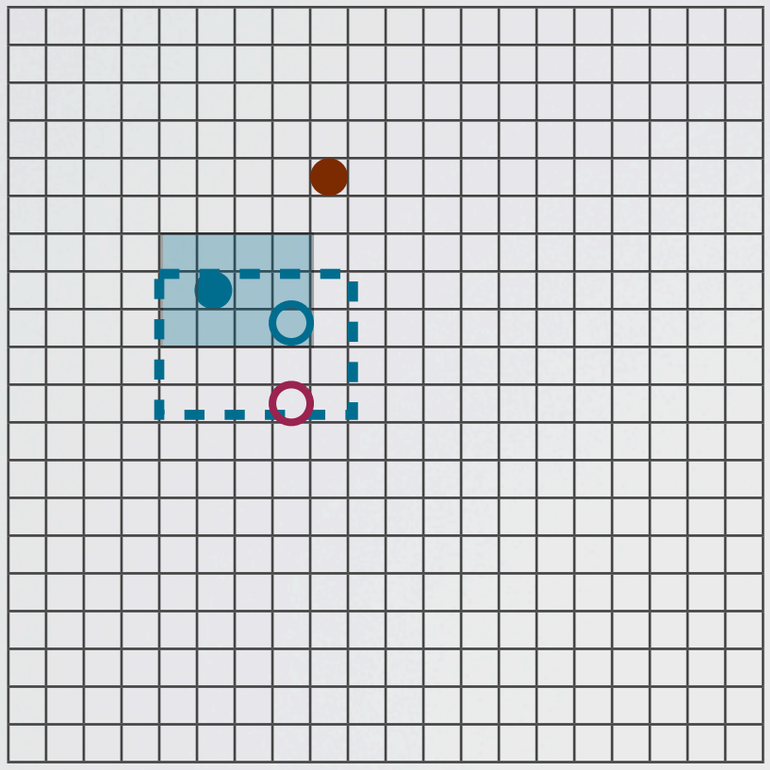
PART 2

OTHER KINDS OF NEGATIVE EVIDENCE

CHANGE OVER TIME

ABOUT NEGATIVE FEEDBACK

It actually is different than negative examples, because the data is generated in a different way -- by the child.



Example generated by parent

- Positive: Sentence from the language

  p(d|h) = 1/n *Bob eats cheese*

- Negative: Sentence not in the language

p(d|h) = 1/(H-n) *Nixim mot ferdlu*

Feedback to child

- Child's sentence is correct

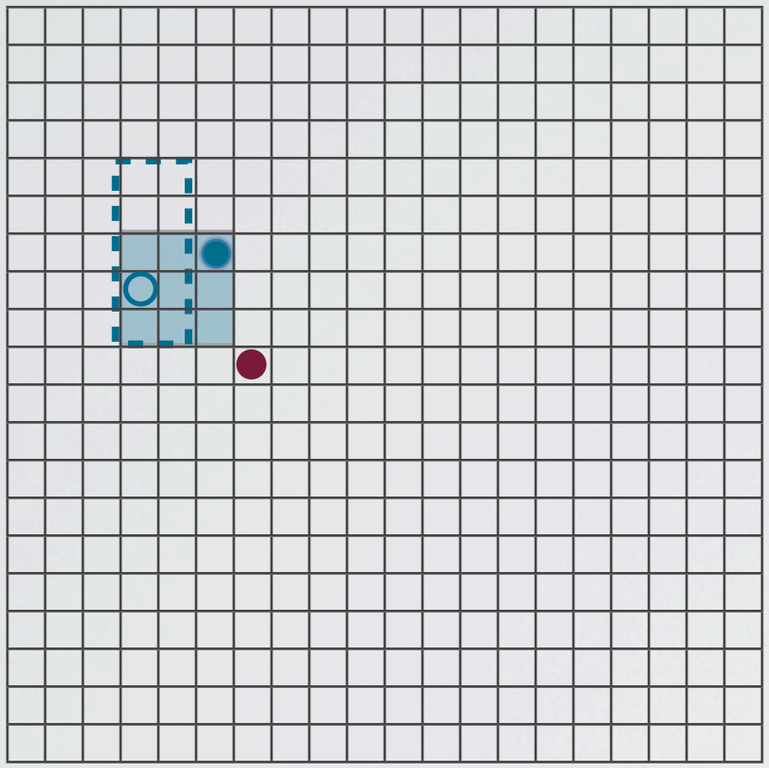
  p(d|h) = 1 *Yes, that's right!*

- Child's sentence is wrong

  p(d|h) = 0 *No, that's wrong!*

INVESTIGATING NEGATIVE FEEDBACK

Same simulations with negative and positive feedback added



- $p(h_c|d) = 0.48$
- $p(h_c|d) = 0.34$
- $p(h_c|d) = 0.37$

1. Randomly select one hypothesis to be the "true" language

2. For each possible positive example, calculate how much receiving it would increase the learner's probability of the *correct* hypothesis.

3. Choose the best positive example.

4. Repeat the process with all possible negative examples.

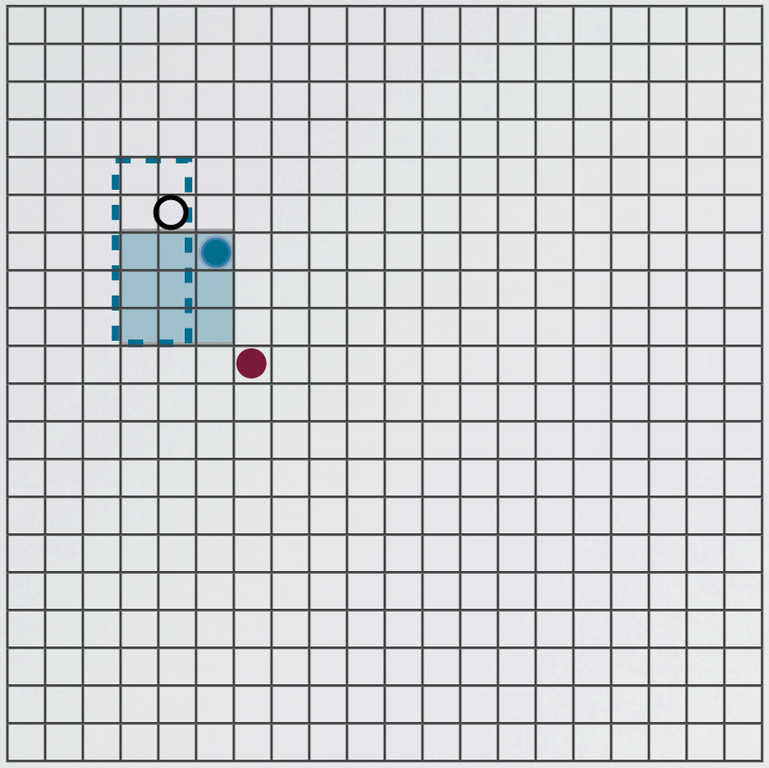
5. Learner selects a language proportional to its probability, and generates an utterance from that language.

6. If that utterance is correct, we calculate how much receiving positive feedback about it would increase the learner's probability of the correct hypothesis.

7. Select which of the positive example, negative example, and positive feedback was best

INVESTIGATING NEGATIVE FEEDBACK

Same simulations with negative and positive feedback added



1. Randomly select one hypothesis to be the "true" language

2. For each possible positive example, calculate how much receiving it would increase the learner's probability of the *correct* hypothesis.

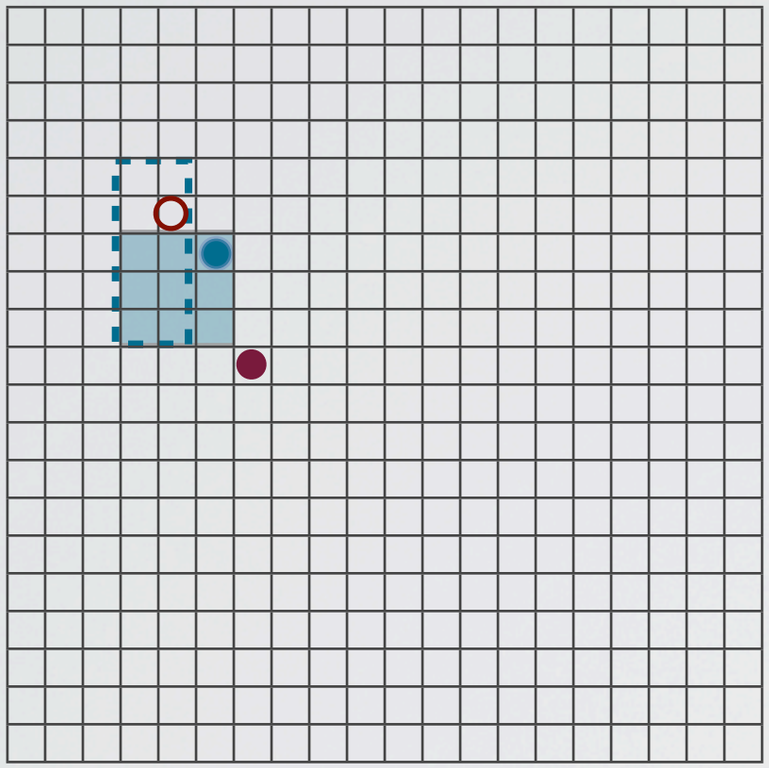
3. Choose the best positive example.

4. Repeat the process with all possible negative examples.

... or ...

INVESTIGATING NEGATIVE FEEDBACK

Same simulations with negative and positive feedback added



- $p(h_c|d) = 0.48$
- $p(h_c|d) = 0.34$
- $p(h_c|d) = 0.45$

1. Randomly select one hypothesis to be the "true" language

2. For each possible positive example, calculate how much receiving it would increase the learner's probability of the *correct* hypothesis.

3. Choose the best positive example.

4. Repeat the process with all possible negative examples.

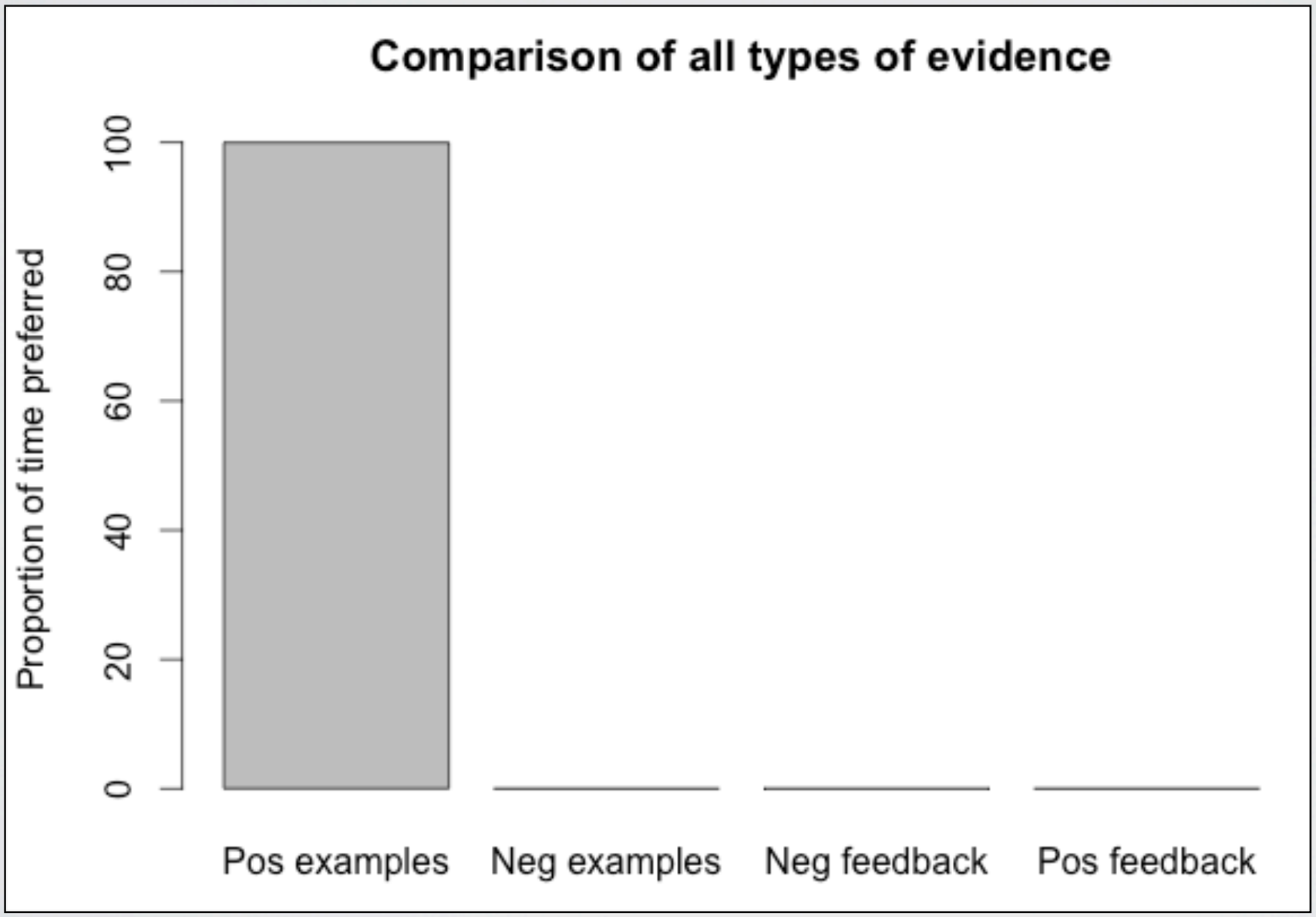
5. Learner selects a language proportional to its probability, and generates an utterance from that language.

6. If it is incorrect, we calculate how much receiving negative feedback about it would increase the learner's probability of the correct hypothesis.

7. Select which of the positive example, negative example, and negative feedback was best

INVESTIGATING NEGATIVE FEEDBACK

Positive examples still always preferred!

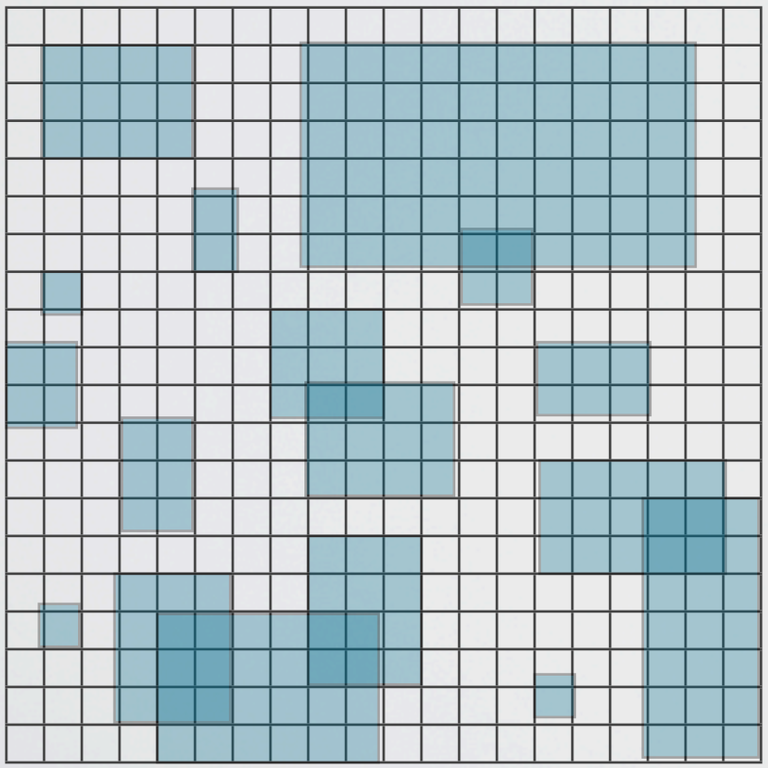


As before, this is because of the sparsity of the hypotheses (at the very beginning)

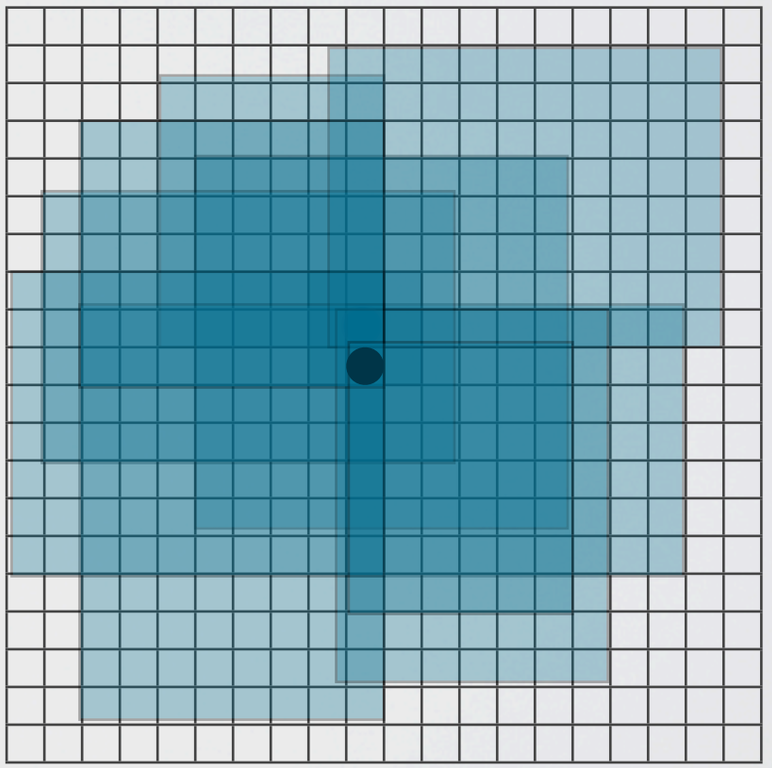
HOW ABOUT OVER TIME, THEN?

Sparsity changes when hypotheses are ruled out; which kind of evidence is best as this occurs?

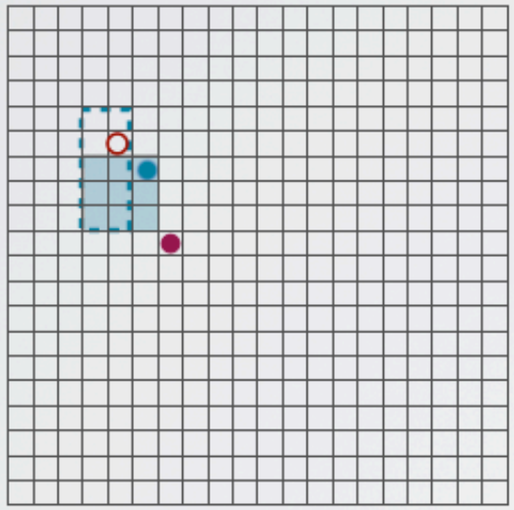
Before:
many hypotheses,
many sparse



After one data point:
fewer overlapping hypotheses,
many not sparse



BEST EVIDENCE OVER TIME



- $p(h_c|d) = 0.48$
- $p(h_c|d) = 0.34$
- $p(h_c|d) = 0.45$

1. Randomly select one hypothesis to be the "true" language

2. For each possible positive example, calculate how much receiving it would increase the learner's probability of the *correct* hypothesis.

3. Choose the best positive example.

4. Repeat the process with all possible negative examples.

5. Learner selects a language proportional to its probability, and generates an utterance from that language.

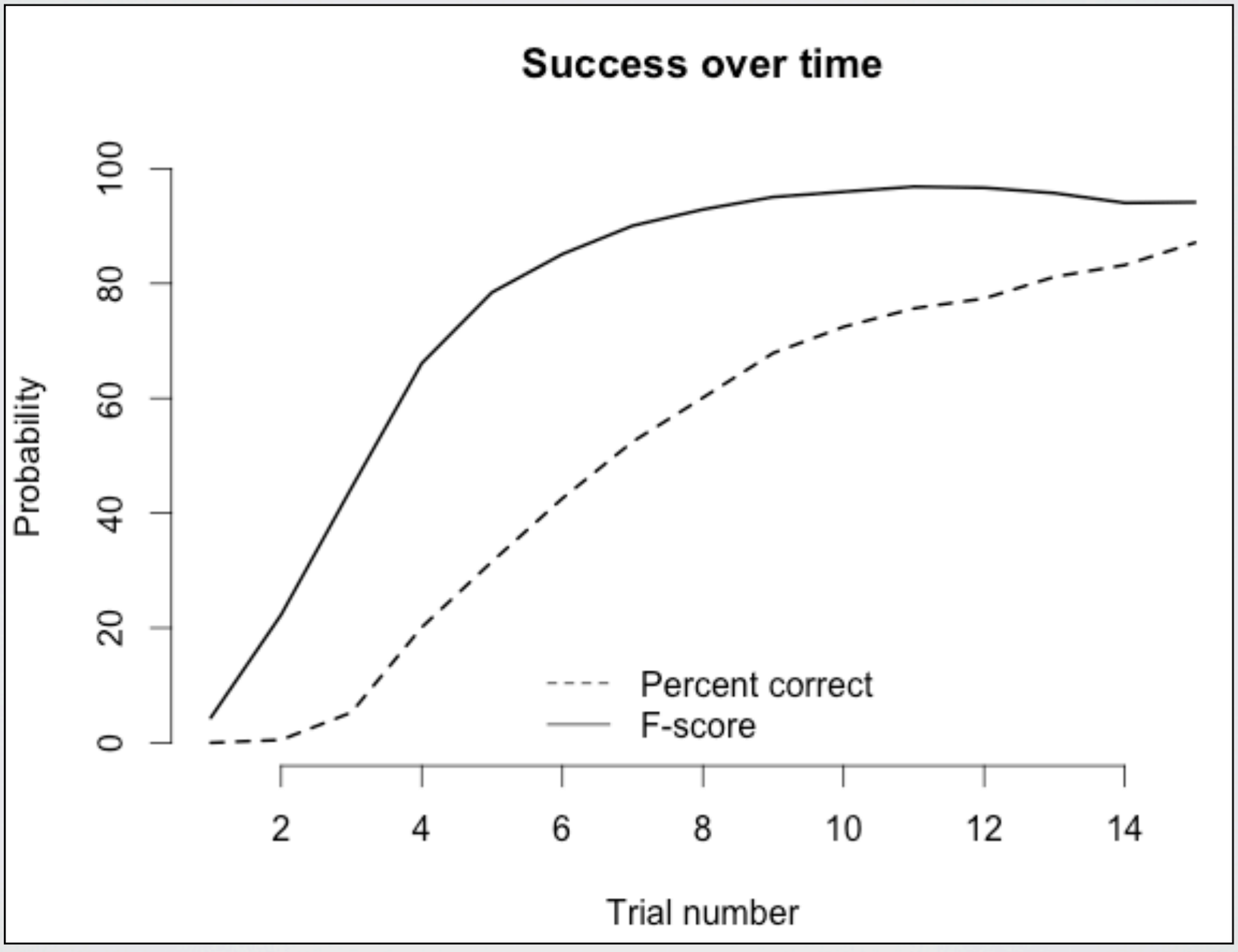
6. If it is incorrect, we calculate how much receiving negative feedback about it would increase the learner's probability of the correct hypothesis.

7. Select which of the positive example, negative example, and negative feedback was best

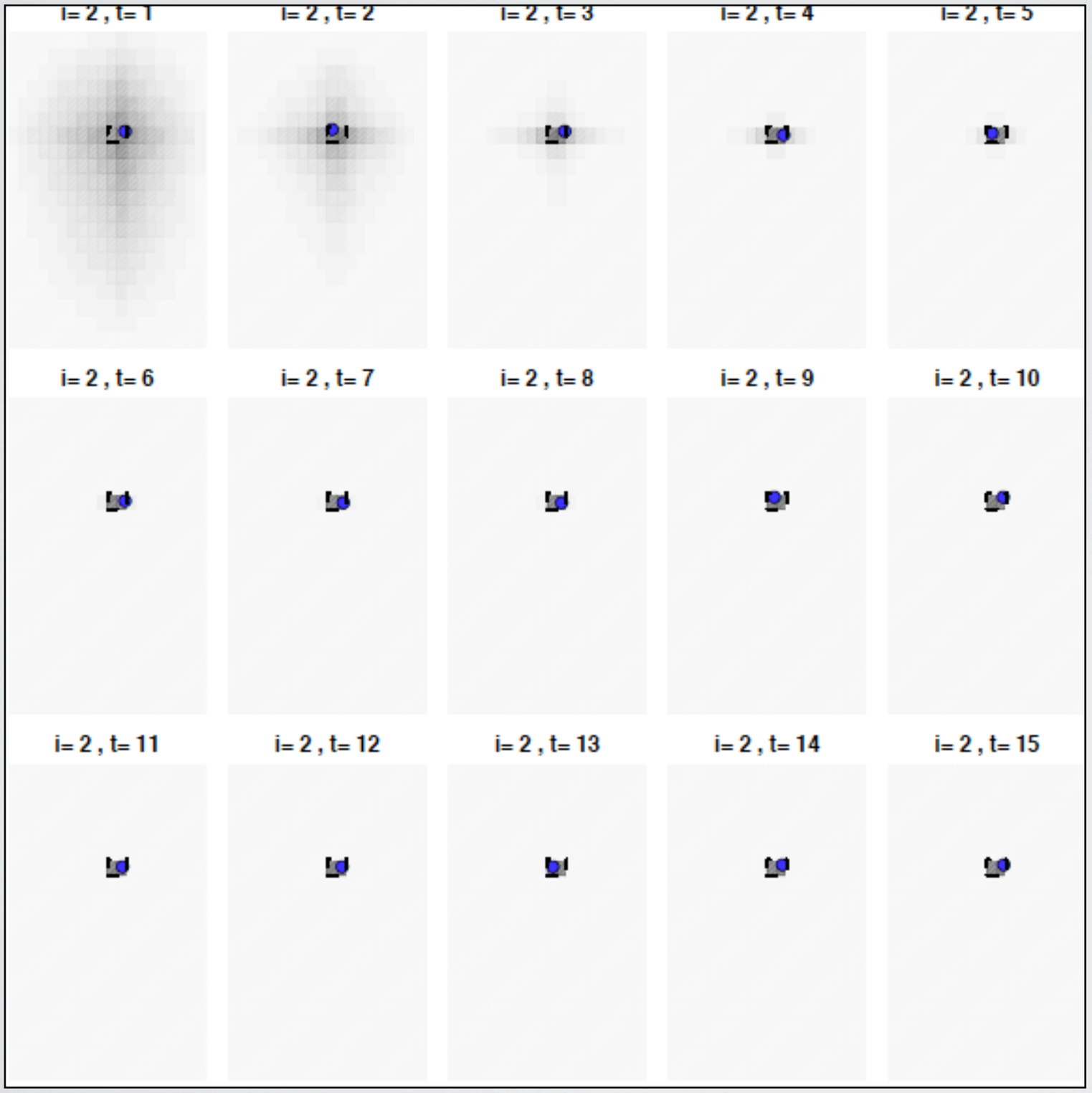
Repeat 15 times, or until the learner believes the correct language with 99.9% probability.

BEST EVIDENCE OVER TIME

The learner in this situation does converge on the correct hypotheses reasonably well...

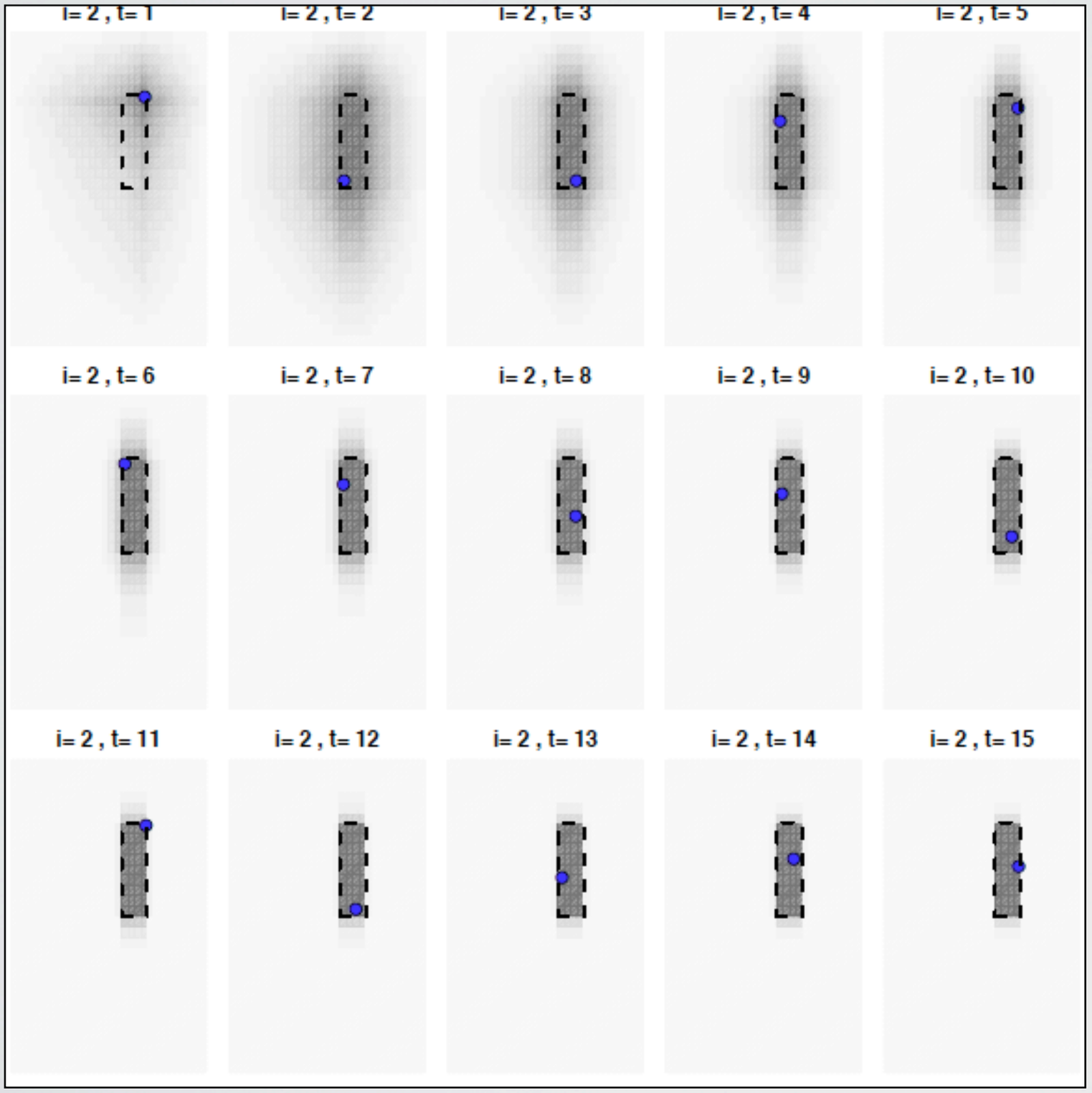


EXAMPLE RUNS...



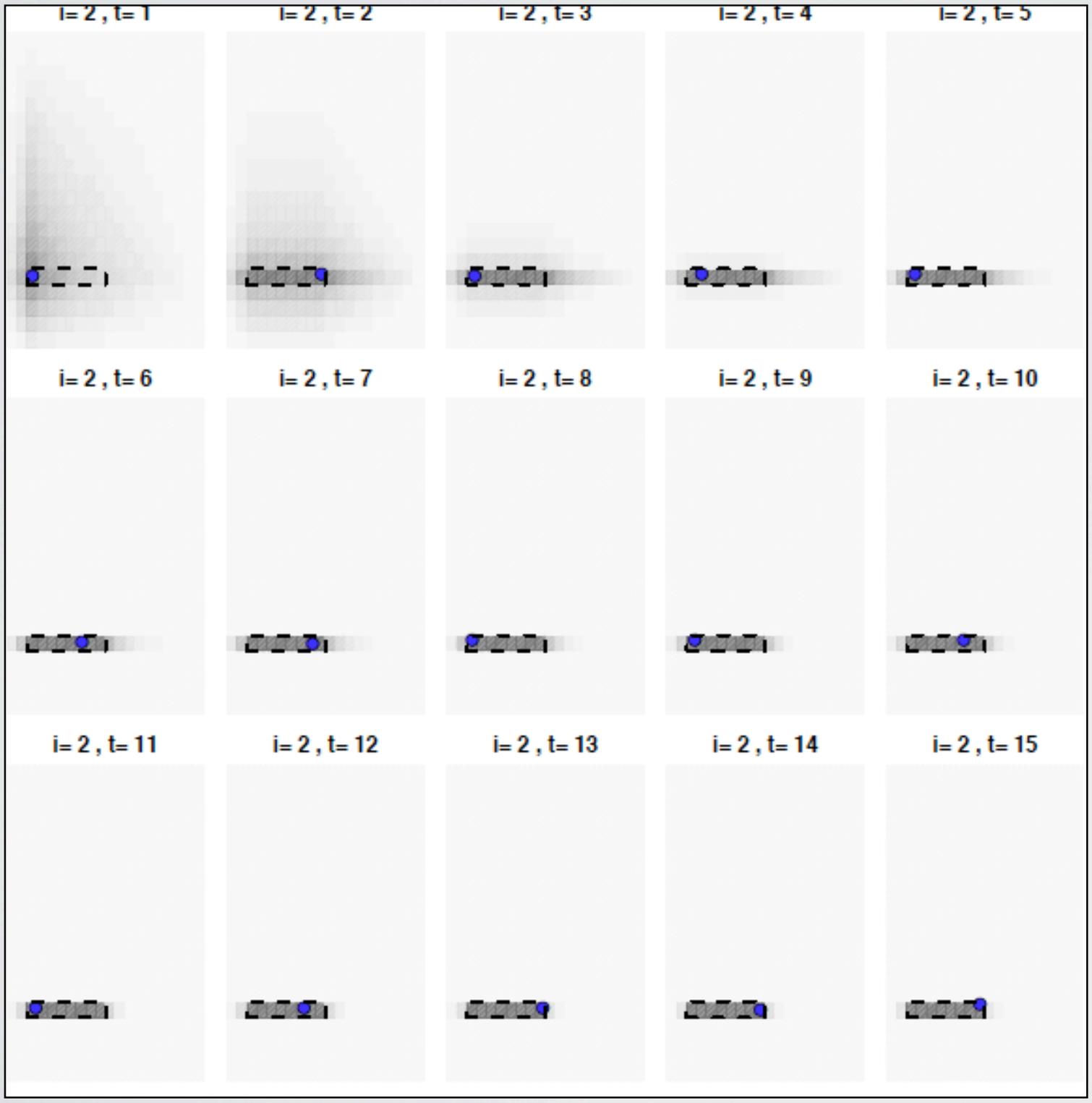
world size 20x20
sparsity 0.5%

EXAMPLE RUNS...



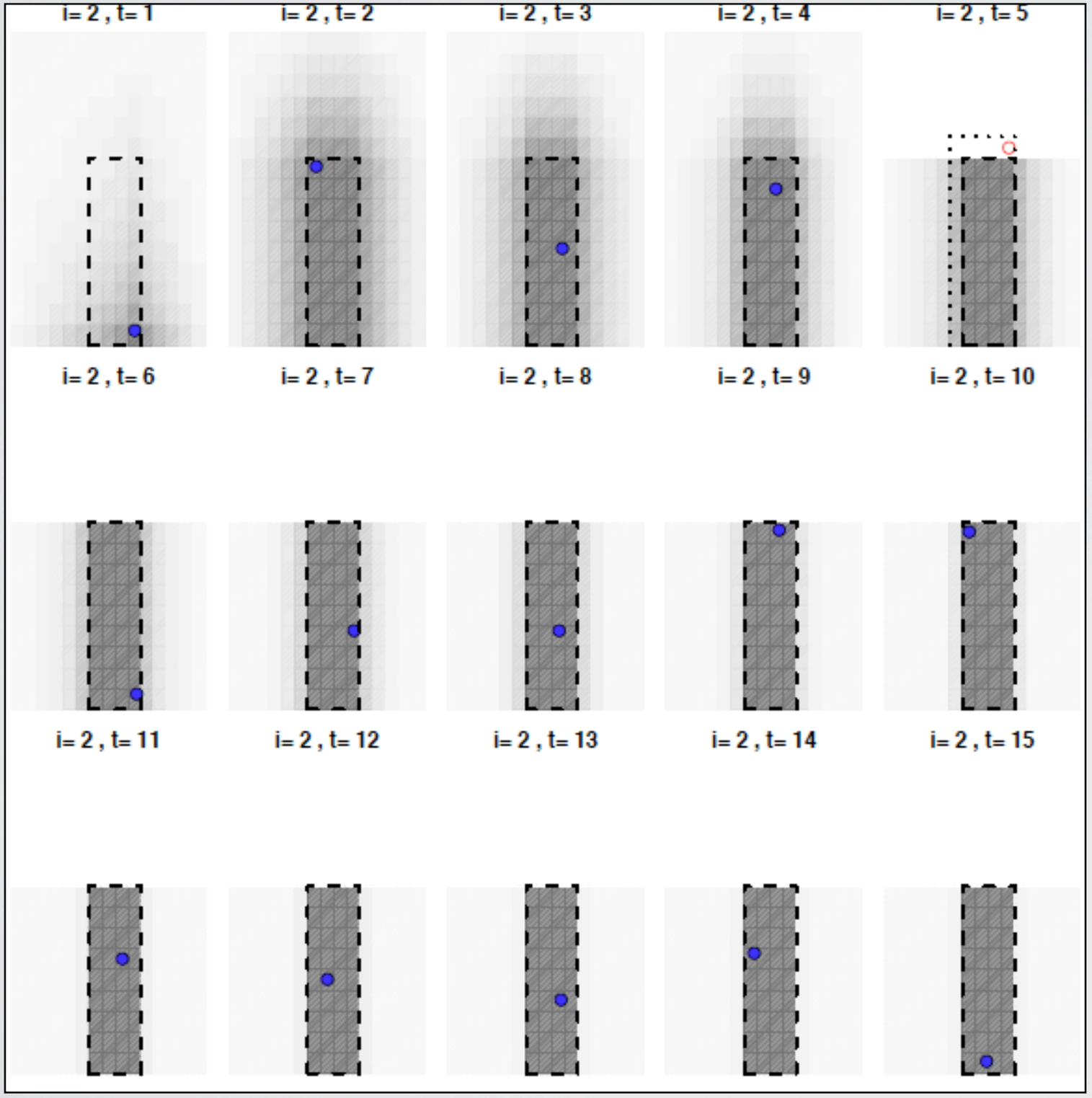
world size 30x30
sparsity 4%

EXAMPLE RUNS...



world size 20x20
sparsity 2%

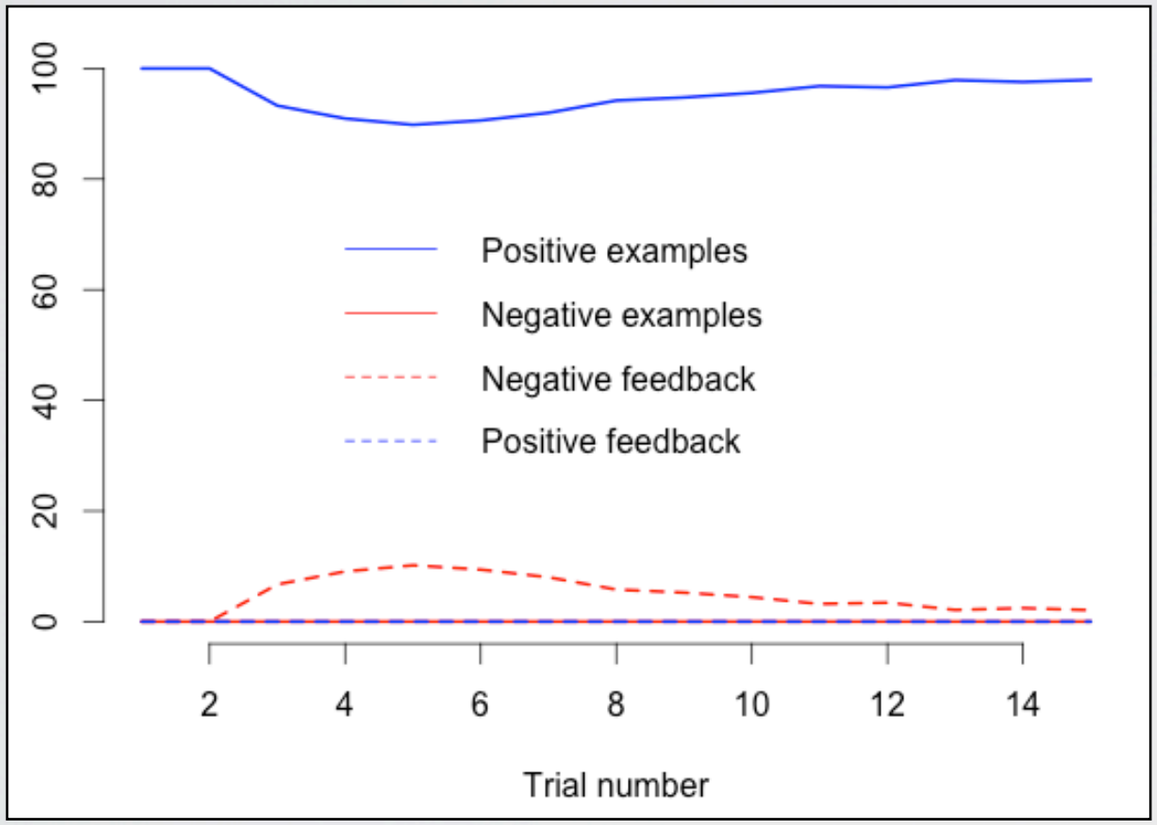
EXAMPLE RUNS...



world size 15x15
sparsity 16%

BEST EVIDENCE OVER TIME

Positive examples are still by far the best, but negative feedback is sometimes better



1. Why is negative feedback better than negative examples?

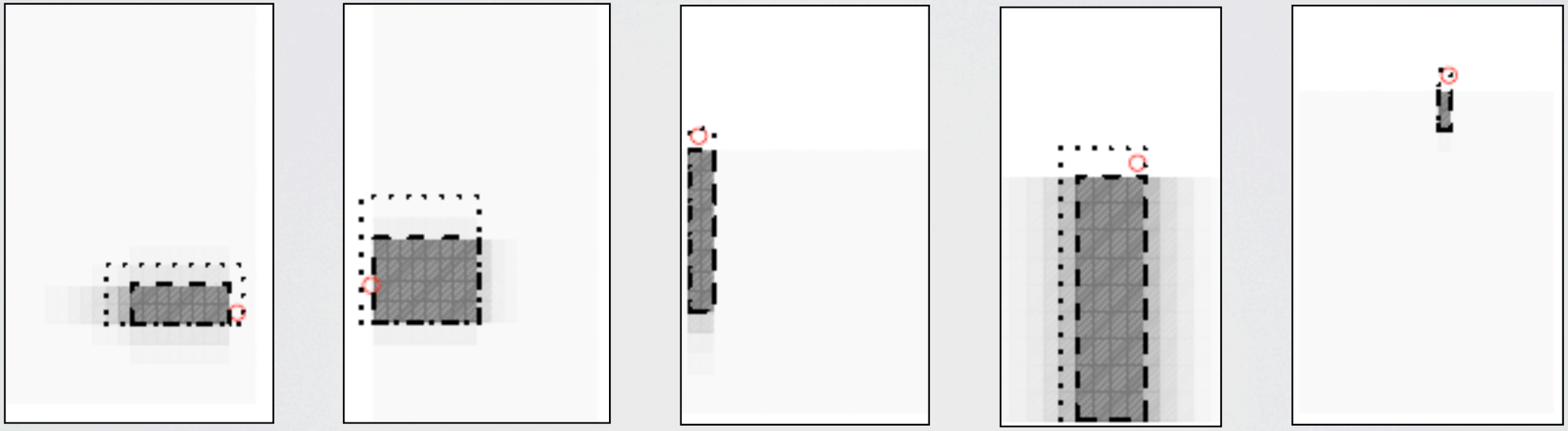
2. Why are positive examples still good so often, even though the hypothesis space is much less sparse?

3. How does this compare to the observed frequency of negative feedback to children?

BEST EVIDENCE OVER TIME

1. Why is negative feedback better than negative examples?

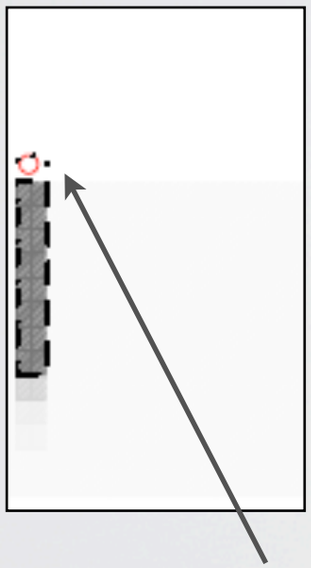
One possibility is that the negative feedback is better targeted (to fix what the child erroneously believes at the moment)



BEST EVIDENCE OVER TIME

1. Why is negative feedback better than negative examples?

Though this might be a factor in the real world, it doesn't explain this: remember we are looking at the informational value of the *best* negative feedback vs the *best* negative example.



This precise data point is better if offered as negative feedback than as a negative example

As feedback, it rules out all hypotheses that contain it but does not change the probability of the correct hypothesis relative to the hypotheses that remain.

$p(d|h,u) = 1$

As an example, it rules the same hypotheses out... but, if the correct hypothesis is sparse, makes it *less* probable relative to superset (non-sparse) hypotheses

$p(d|h) = 1/(H-n) = 1/(400-10) = 0.25\%$

vs larger superset hL of size 300:

$p(d|hL) = 1/(400-300) = 1\%$

BEST EVIDENCE OVER TIME

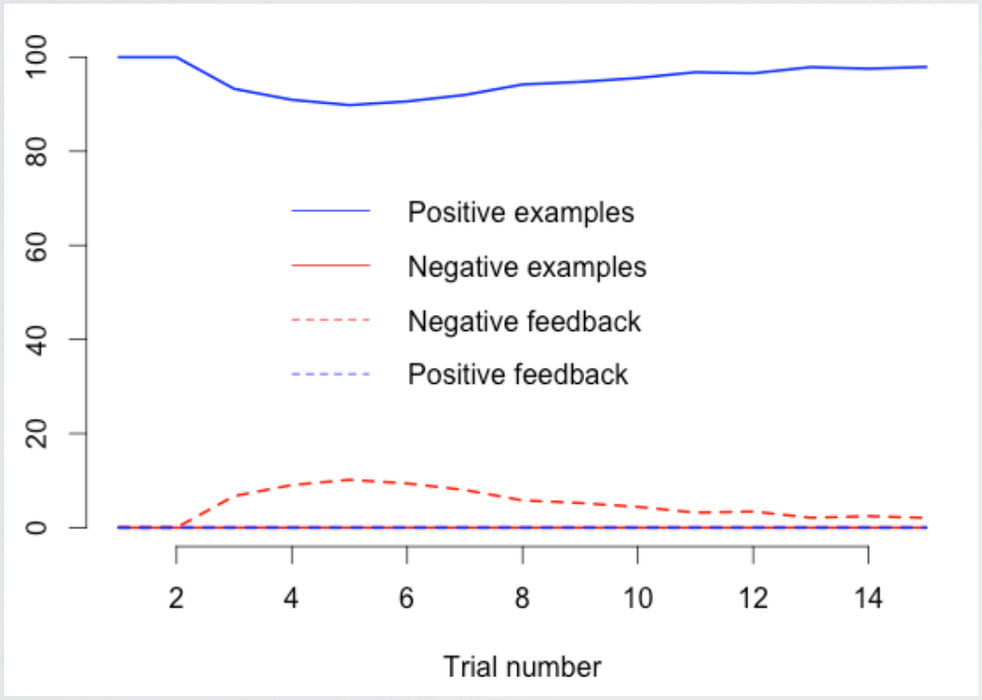
1. Why is negative feedback better than negative examples?

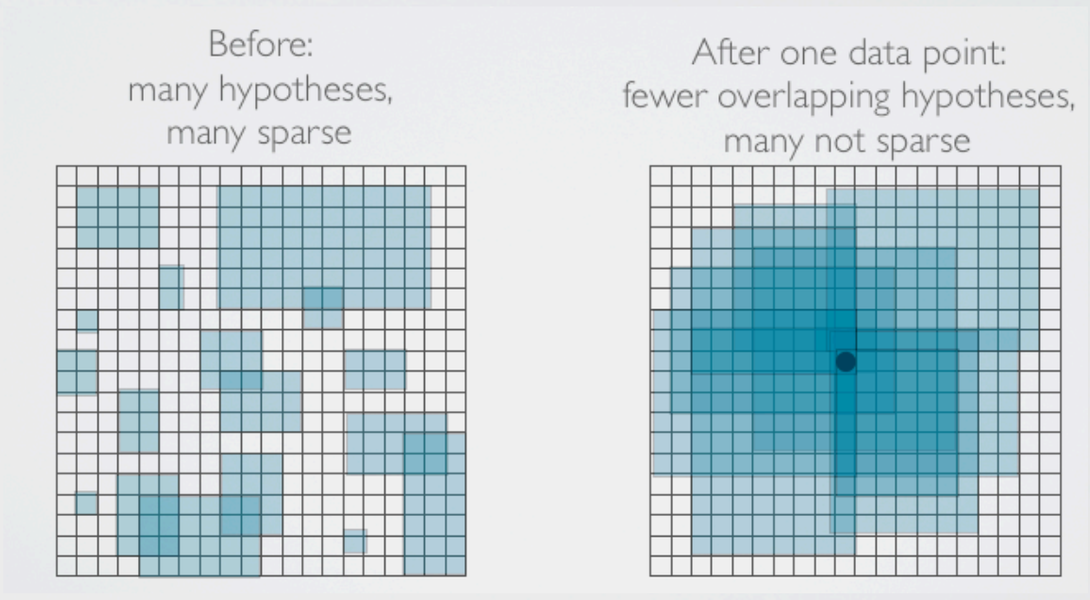
Though this might be a factor in the real world, it doesn't explain this: remember we are looking at the informational value of the *best* negative feedback vs the *best* negative example.

So as long as the correct hypothesis is sparse, negative examples actually incorrectly give more relative probability to incorrect superset hypotheses (even though they also rule some out). Negative feedback rules them out without this problem.

BEST EVIDENCE OVER TIME

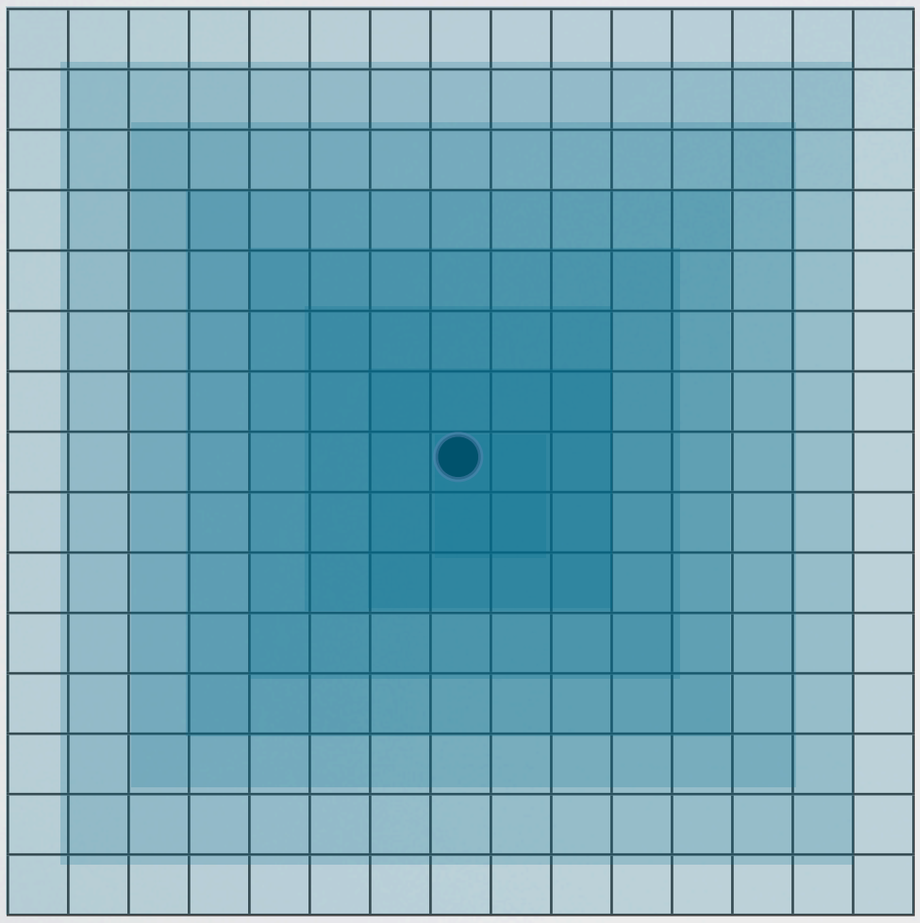
2. Why are positive examples still good so often, even though the hypothesis space is no longer sparse after the first few trials?





BEST EVIDENCE OVER TIME

2. Why are positive examples still good so often, even though the hypothesis space is no longer sparse after the first few trials?



Answer: the size principle

$$p(d|h) = 1/n$$

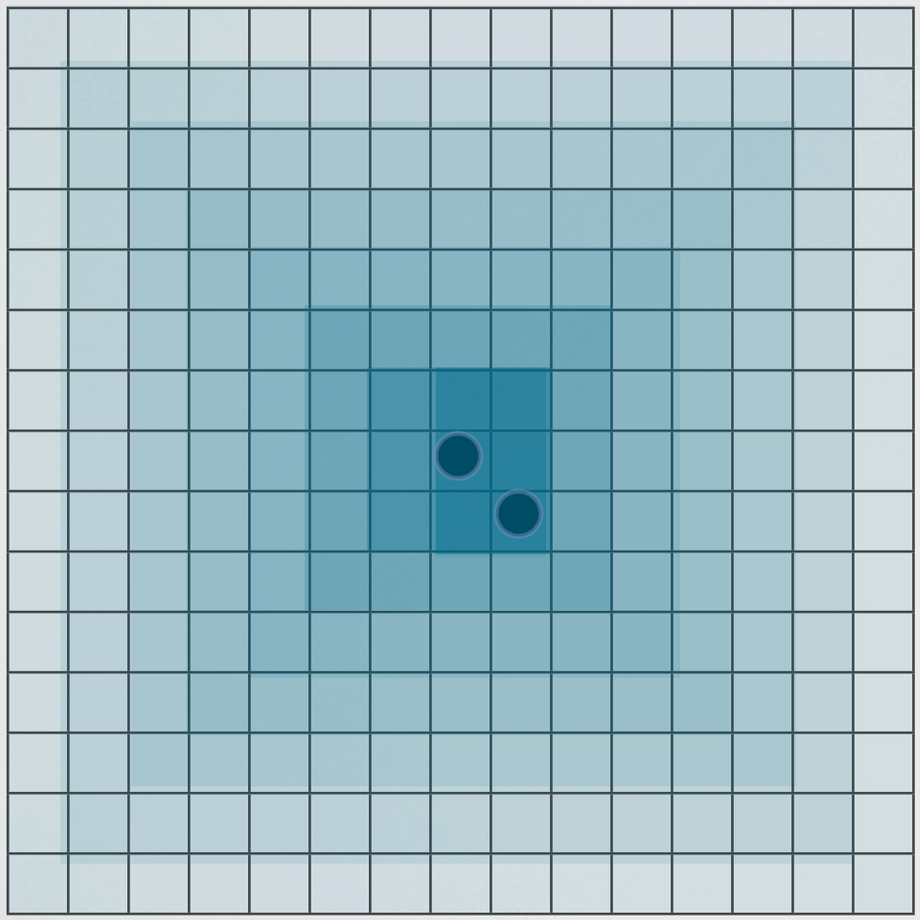
(where n is the size of the hypothesis, i.e., the number of elements in it)

This means that smaller hypotheses are increasingly favoured with additional data

This is the flip side of why negative examples are not useful

BEST EVIDENCE OVER TIME

2. Why are positive examples still good so often, even though the hypothesis space is no longer sparse after the first few trials?



Answer: the size principle

$p(d|h) = 1/n$

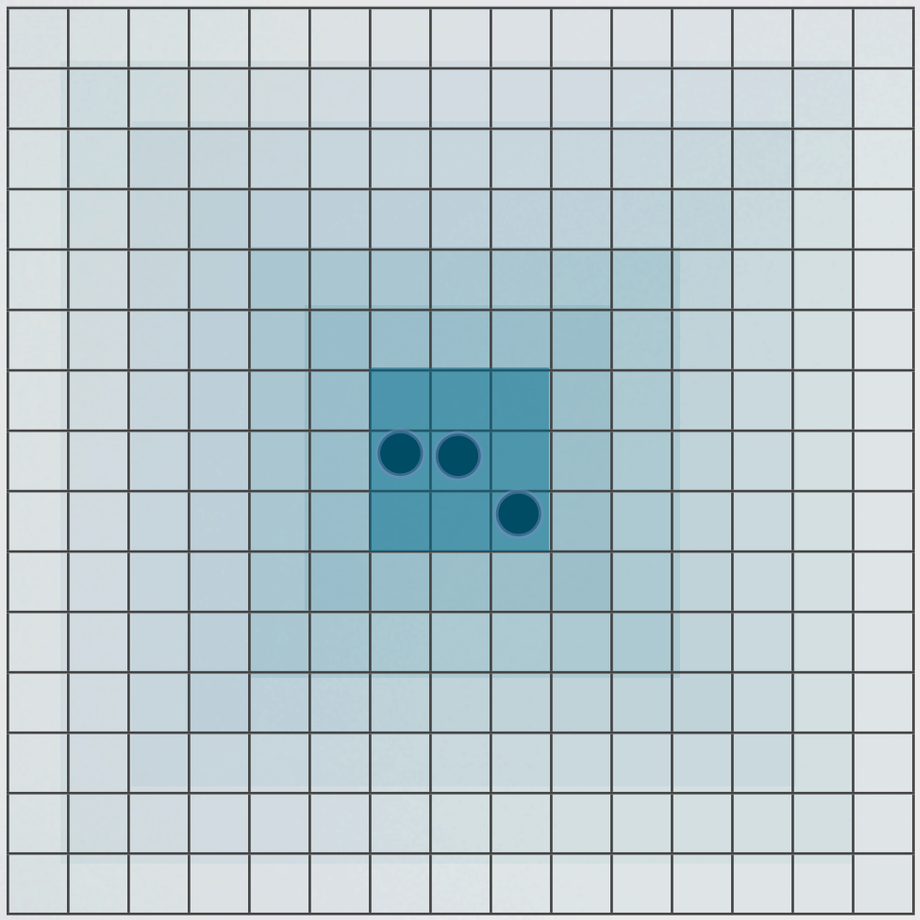
(where n is the size of the hypothesis, i.e., the number of elements in it)

This means that smaller hypotheses are increasingly favoured with additional data

This is the flip side of why negative examples are not useful

BEST EVIDENCE OVER TIME

2. Why are positive examples still good so often, even though the hypothesis space is no longer sparse after the first few trials?



Answer: the size principle

$p(d|h) = 1/n$

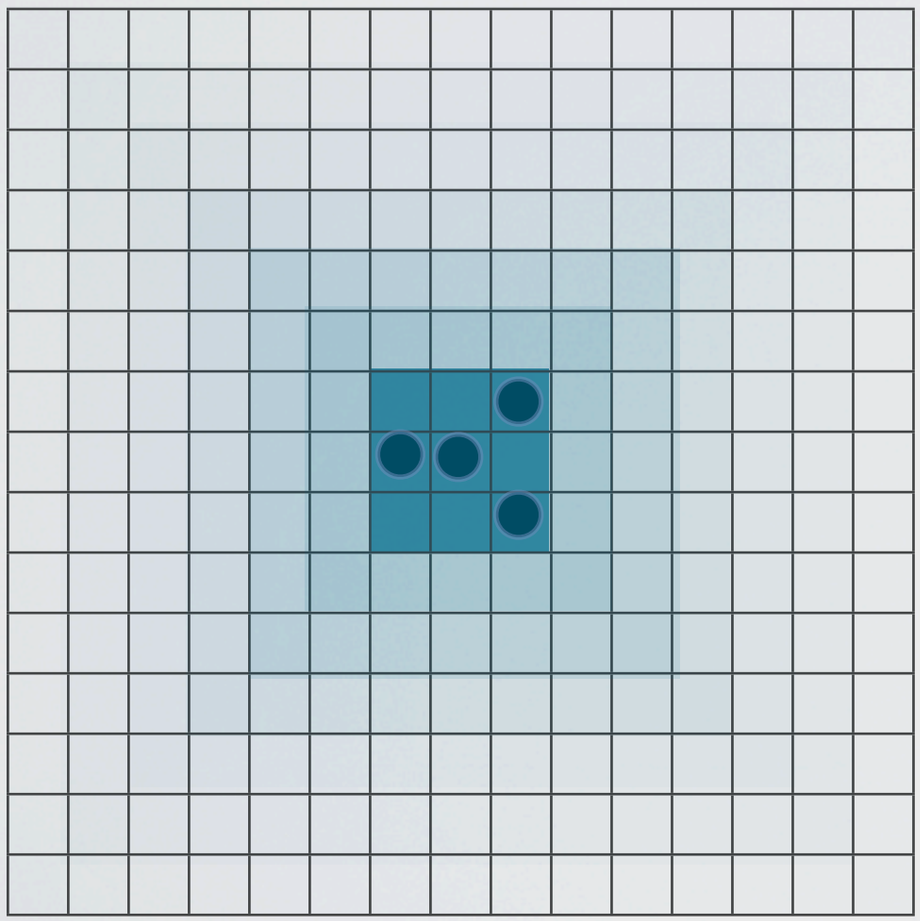
(where n is the size of the hypothesis, i.e., the number of elements in it)

This means that smaller hypotheses are increasingly favoured with additional data

This is the flip side of why negative examples are not useful

BEST EVIDENCE OVER TIME

2. Why are positive examples still good so often, even though the hypothesis space is no longer sparse after the first few trials?



Answer: the size principle

$$p(d|h) = 1/n$$

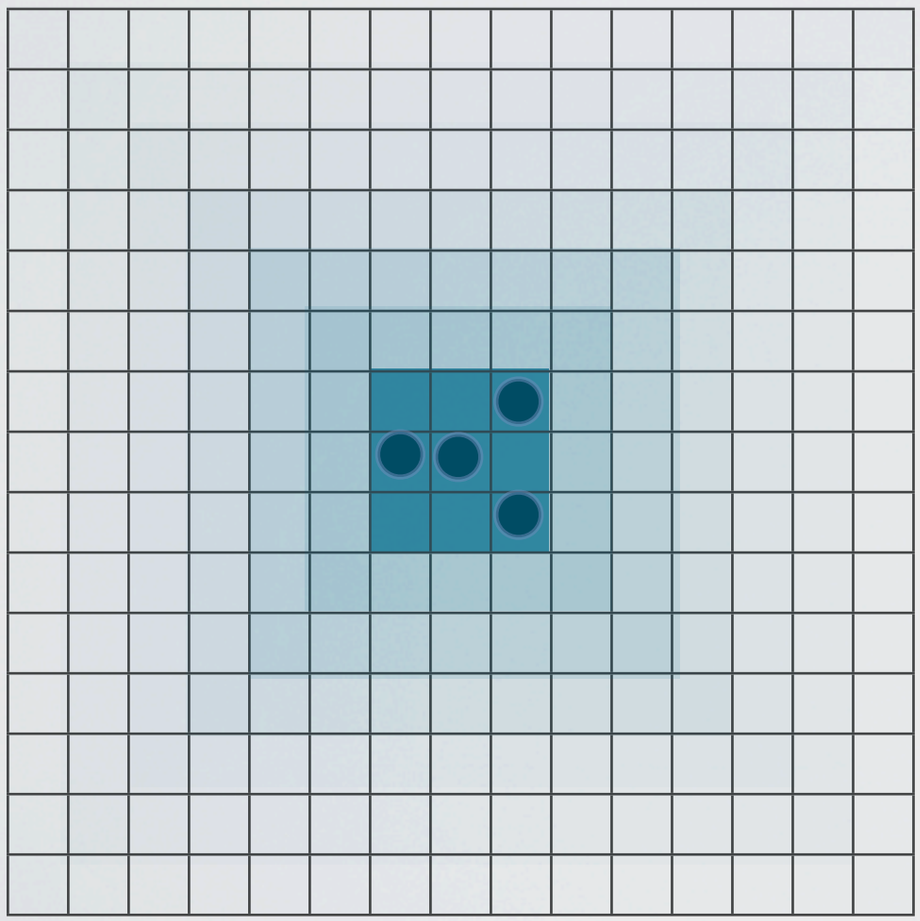
(where n is the size of the hypothesis, i.e., the number of elements in it)

This means that smaller hypotheses are increasingly favoured with additional data

This is the flip side of why negative examples are not useful

BEST EVIDENCE OVER TIME

2. Why are positive examples still good so often, even though the hypothesis space is no longer sparse after the first few trials?



Answer: the size principle

$$p(d|h) = 1/n$$

(where n is the size of the hypothesis, i.e., the number of elements in it)

The smaller the hypothesis is, the more informative any single example from it is, relative to negative evidence.

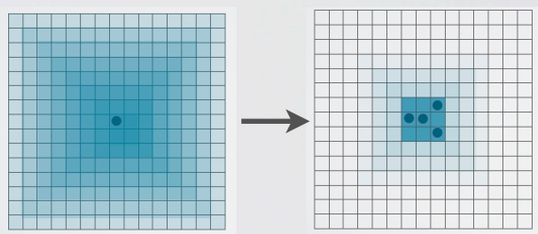
BEST EVIDENCE OVER TIME

Of course, the size principle assumes *strong sampling* -- the presumption that examples are sampled directly from the hypothesis.

If people presume *weak sampling*, in which examples are sampled from the world and then labeled as positive or negative, the probabilities look different.

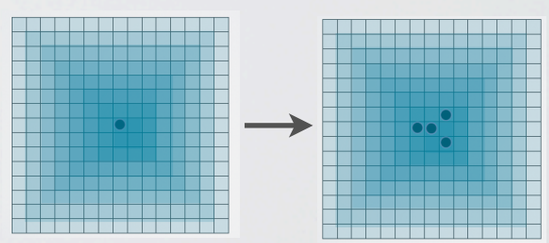
strong sampling:
$p(d|h) = 1/n$

implies rapid "tightening" of hypotheses with additional data



weak sampling:
$p(d|h) = 1$ if allowed
0 if not

implies no "tightening" of hypotheses with additional data



BEST EVIDENCE OVER TIME

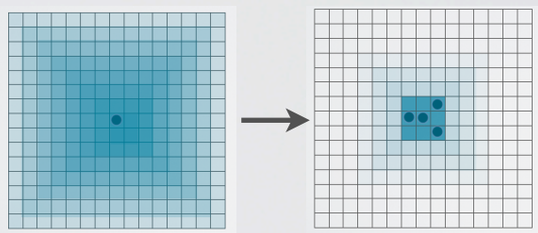
It is also possible to assume some combination of the two:

$p(d|h) = \theta/n + (1-\theta)$ *if allowed*
0 if not

θ is a measure of the degree to which strong sampling is assumed.
$\theta = 1$: strong sampling
$\theta = 0$: weak sampling

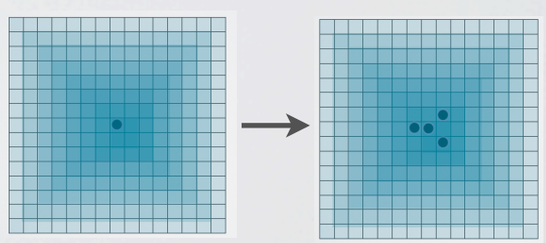
strong sampling:
$p(d|h) = 1/n$

implies rapid "tightening" of hypotheses with additional data



weak sampling:
$p(d|h) = 1$ if allowed
0 if not

implies no "tightening" of hypotheses with additional data



BEST EVIDENCE OVER TIME

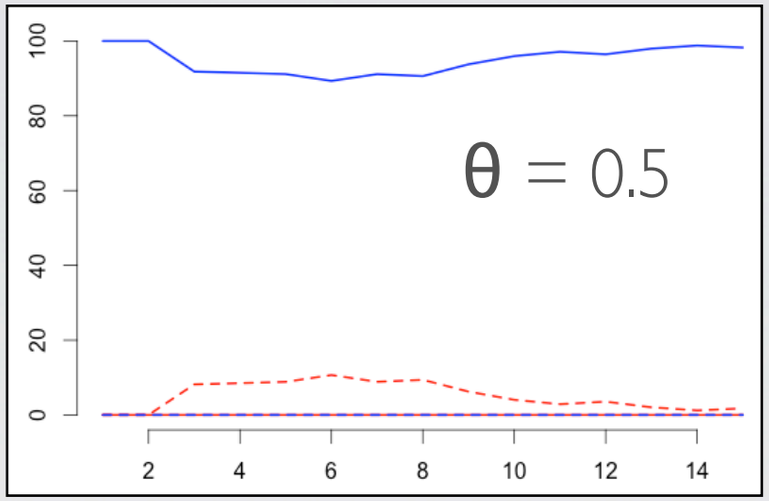
It is also possible to assume some combination of the two:

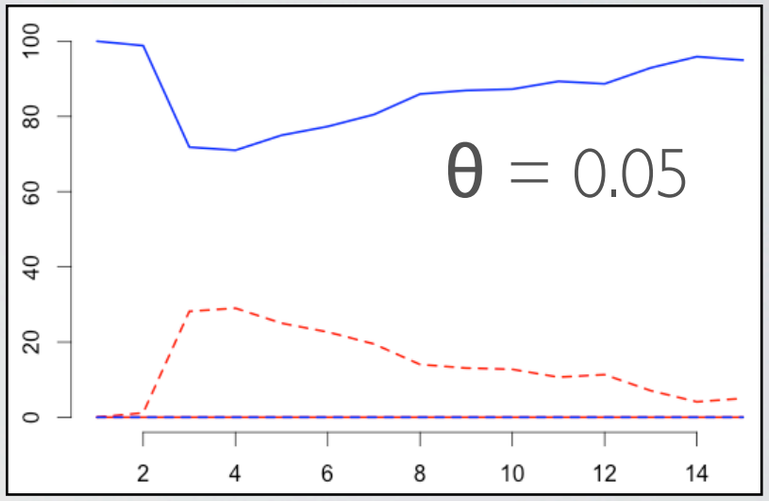
$p(d|h) = \theta/n + (1-\theta)$ *if allowed*
0 if not

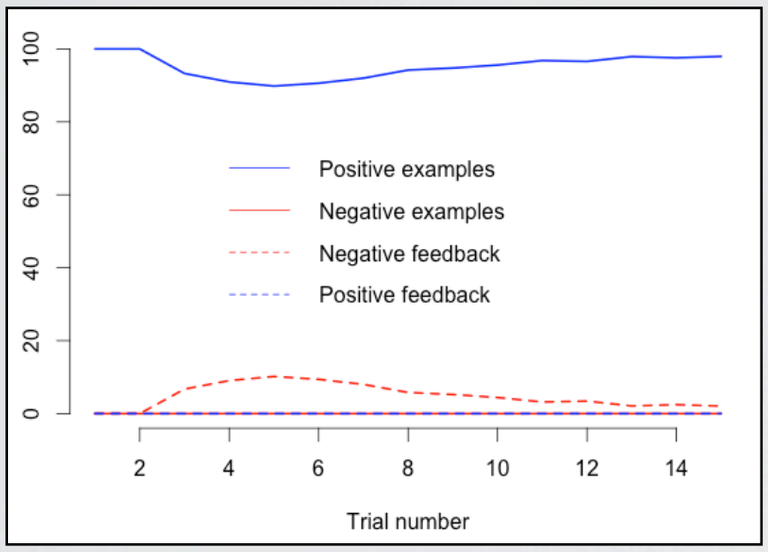
θ is a measure of the degree to which strong sampling is assumed.
$\theta = 1$: strong sampling
$\theta = 0$: weak sampling

How does which evidence is best change as θ changes?

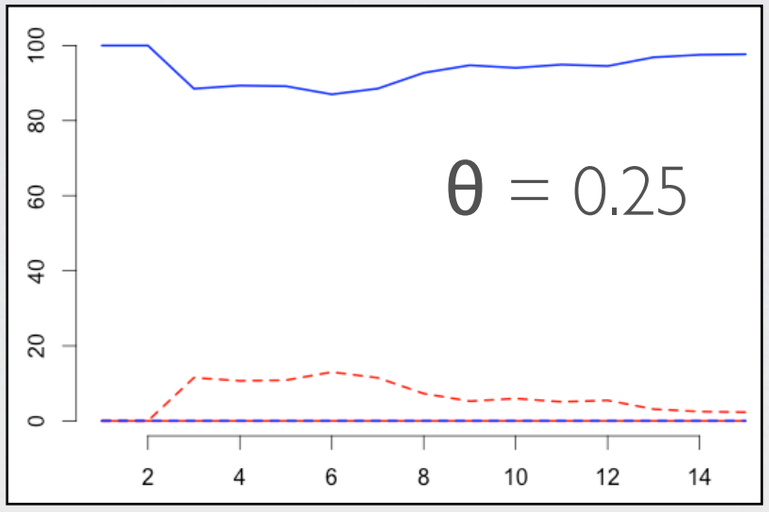
VARYING STRONG SAMPLING ASSUMPTION

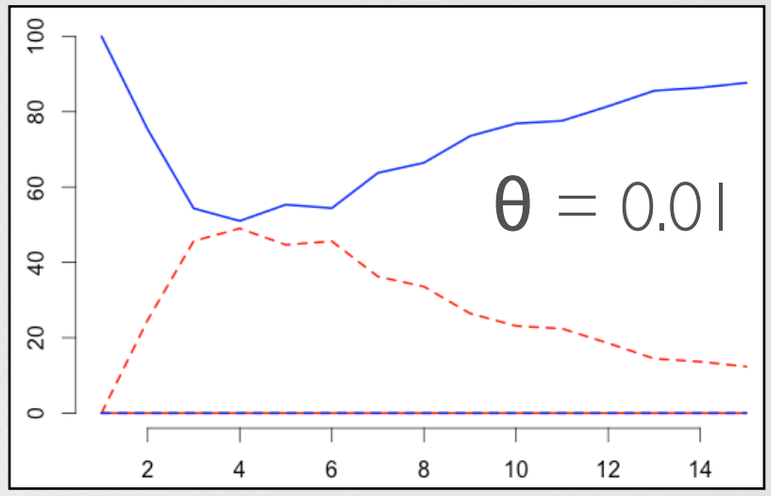


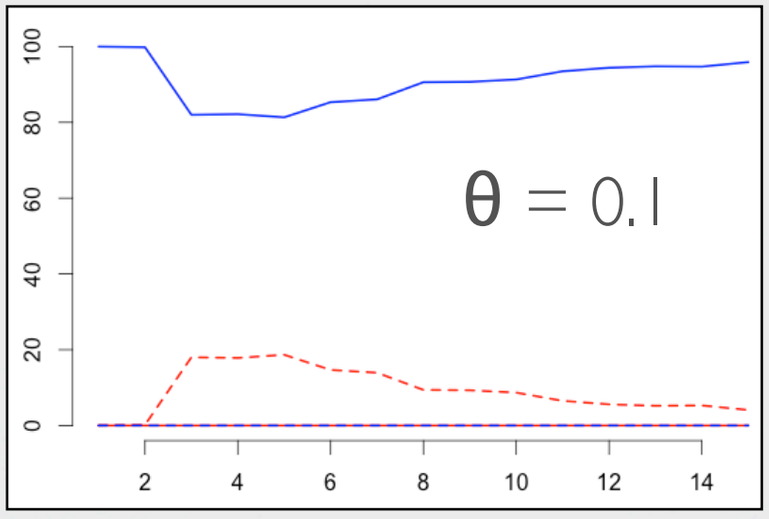


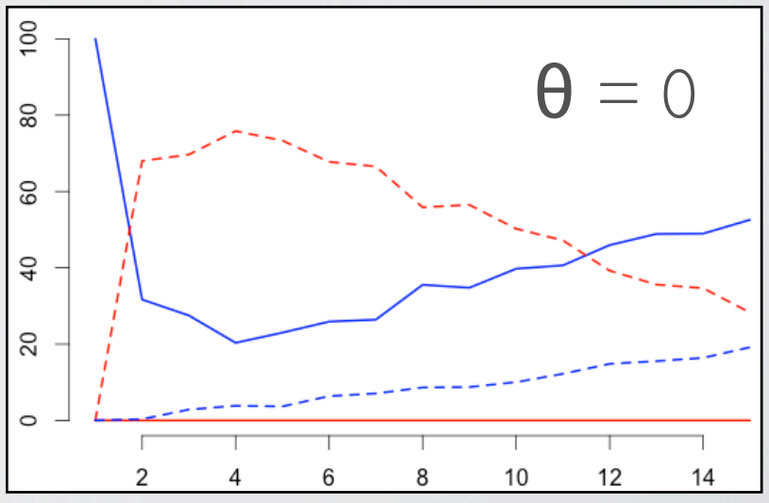


θ = 1 (strong sampling)

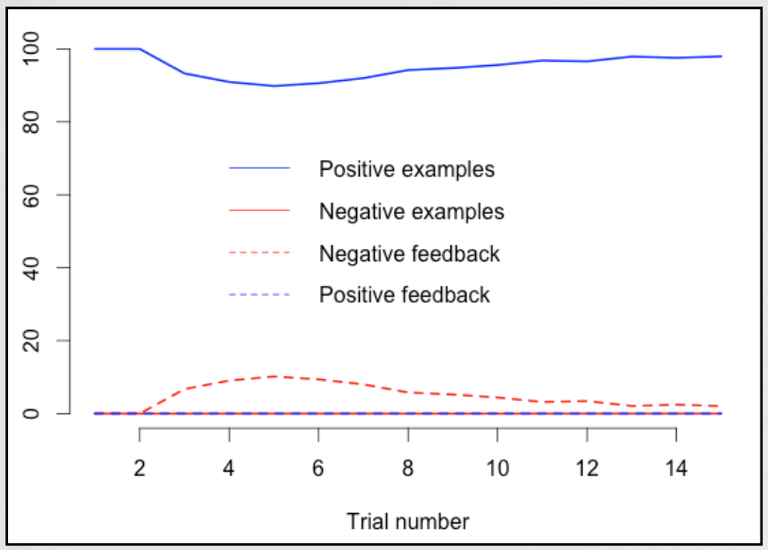








VARYING STRONG SAMPLING ASSUMPTION



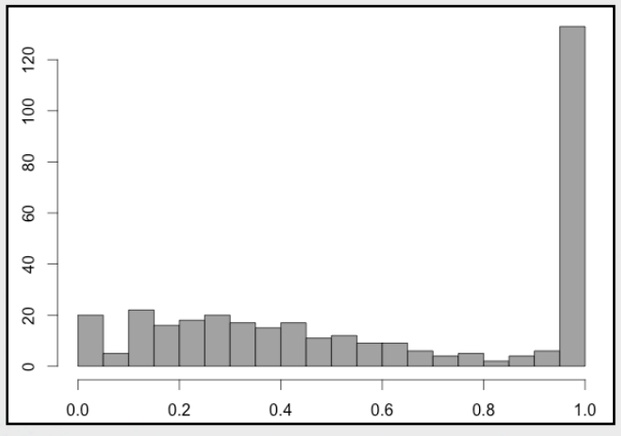
$\theta = 1$ (strong sampling)

Overall, positive evidence is better as long as there is *any* level of tightening with additional data (i.e., as long as people do not completely assume weak sampling)

WHAT DO PEOPLE DO?

There is a great deal of research on this in non-grammar contexts, where it is easier to characterise the "size" (*n*) of a hypothesis and thus precisely measure the extent to which people follow the size principle.

- Lots of evidence that adults and children tighten their generalizations with additional data, following something like the size principle (Sanjana & Tenenbaum, 2003; Fernbach, 2006; Xu & Tenenbaum, 2007; Goodman et al., 2007; Navarro & Perfors, 2010; Perfors et al., 2011; Frank & Tenenbaum, 2011; Frank & Goodman, 2012; Vong et al., 2013; and many more)
- Some work fit individual estimates of θ to participants in a category-learning scenario. mean θ was 0.6, 94.3% had θ greater than 0.05. (Vong et al., 2013)



WHAT DO PEOPLE DO?

There is a great deal of research on this in non-grammar contexts, where it is easier to characterise the "size" (*n*) of a hypothesis and thus precisely measure the extent to which people follow the size principle.

Overall, then, the evidence does seem to suggest that people do tighten their hypotheses with additional data to at least some extent.

It is in this case that positive evidence is usually better than negative evidence, even after most hypotheses have been ruled out.

BEST EVIDENCE OVER TIME

3. How does this compare to the observed frequency of negative feedback to children?

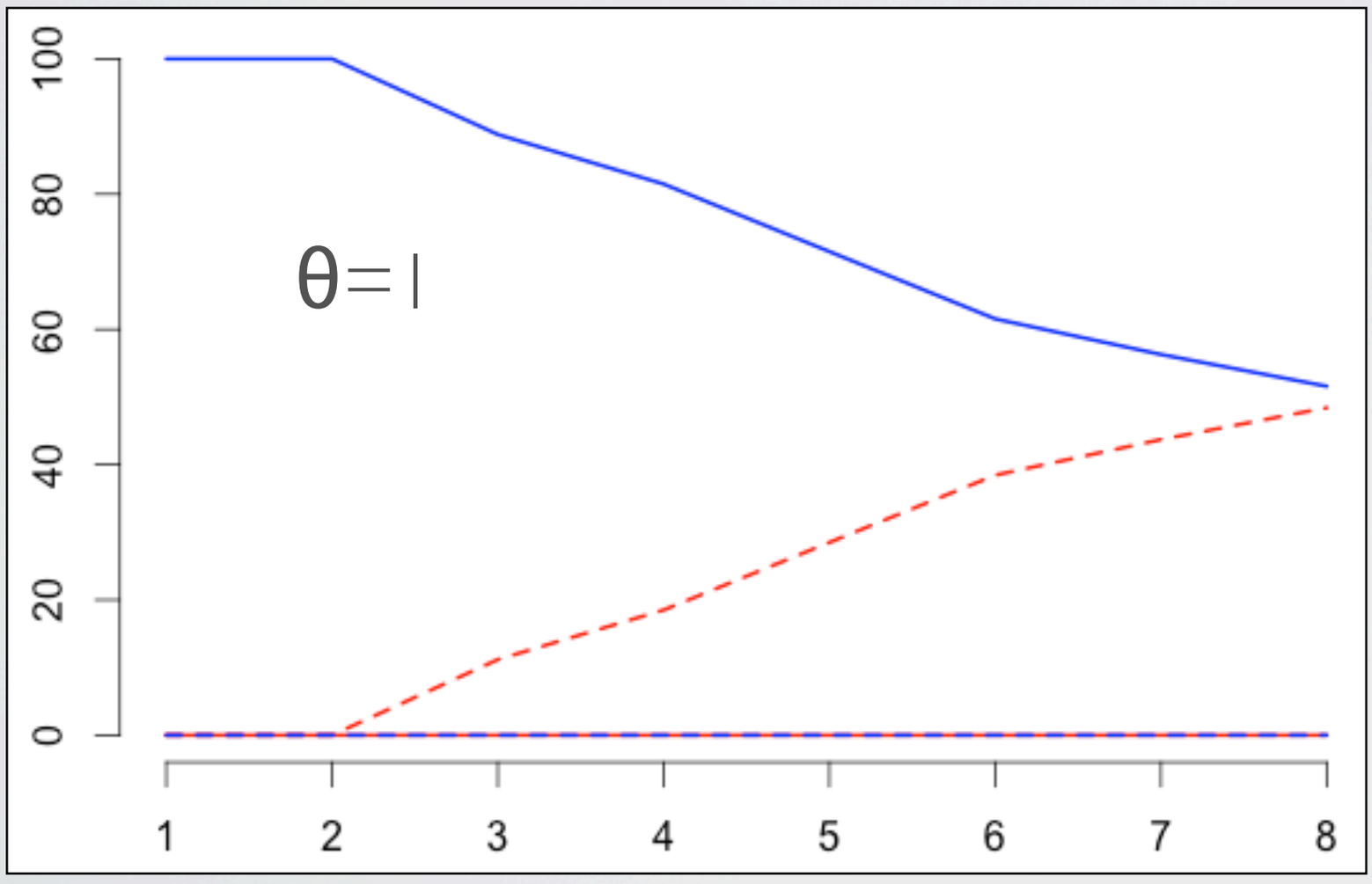
So far I have been reporting the results out of all trials. Perhaps there is relatively little negative feedback preferred simply because our learner makes few mistakes that need correcting?

Indeed, in the literature the quantity of negative evidence (feedback) is generally reported out of the total number of errors children make.

What happens if I restrict the analysis to those trials?

BEST EVIDENCE OVER TIME

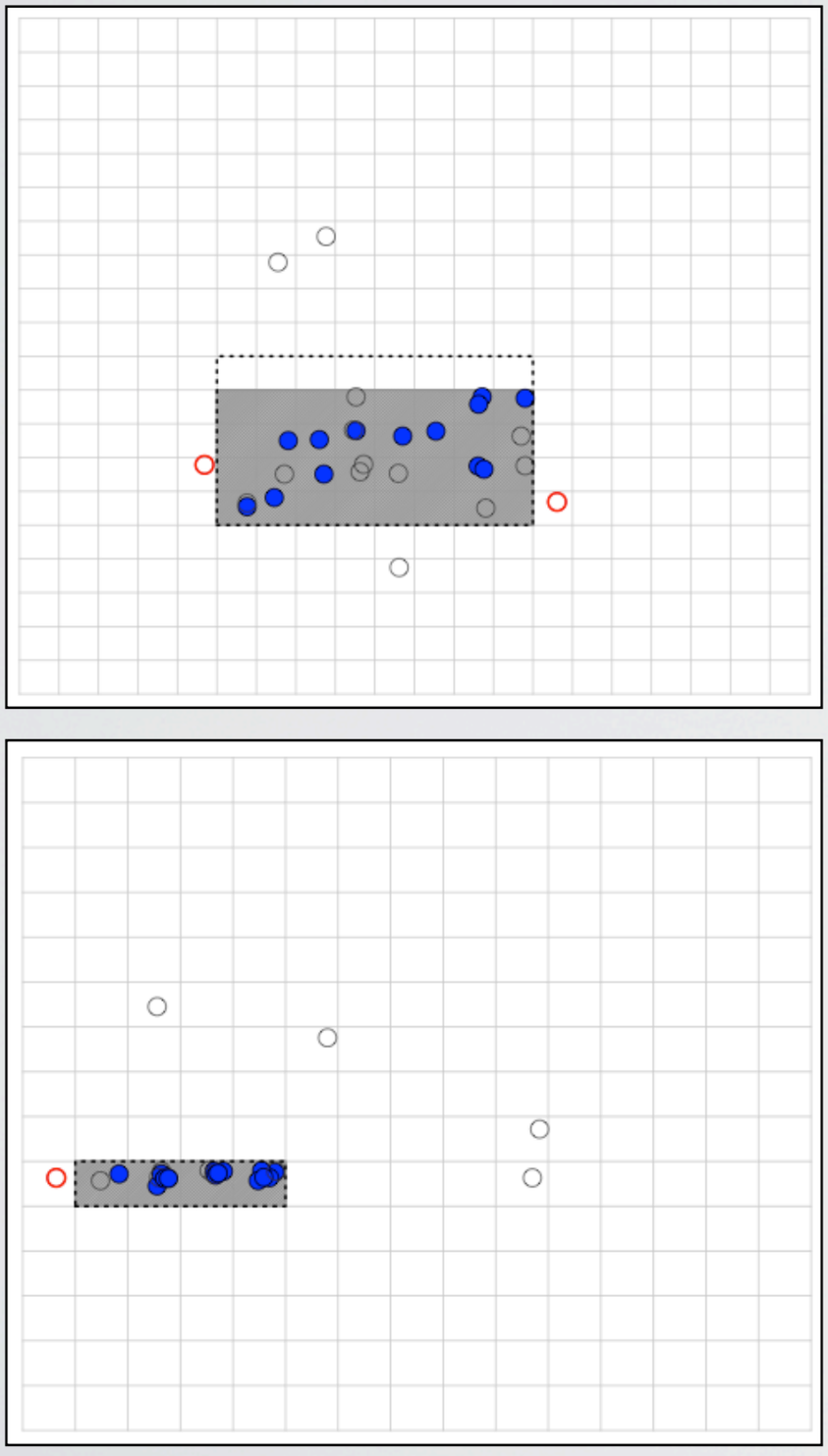
3. How does this compare to the observed frequency of negative feedback to children?

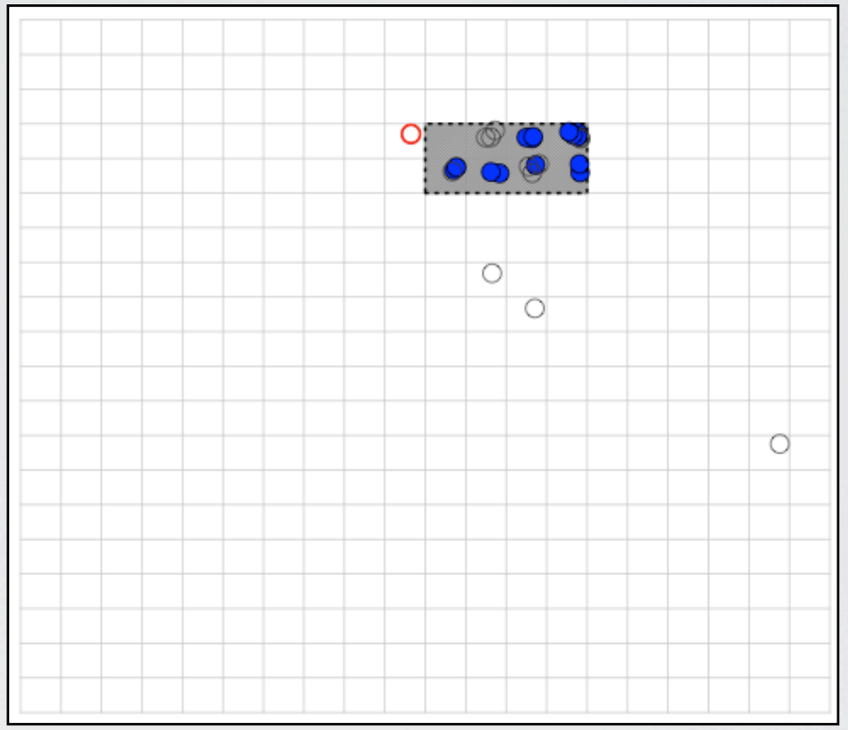


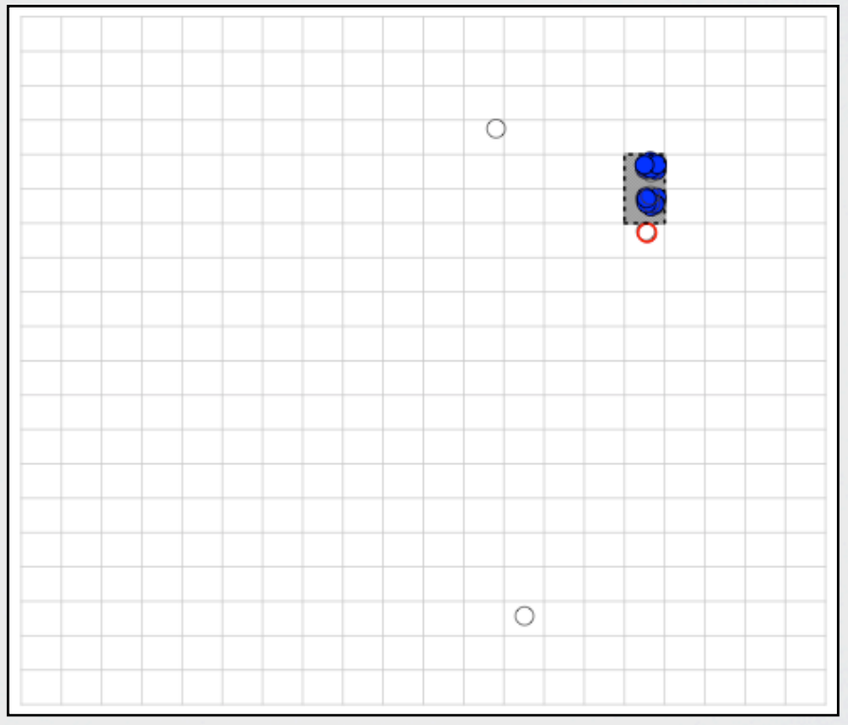
Compares interestingly with the observed empirical frequency of 10-67% of ungrammatical utterances

(Saxton, 2000; Bohannon & Stanowicz, 1988, Farrar 1992, Chouinard & Clark, 2003)

WHEN IS NEGATIVE FEEDBACK BETTER?



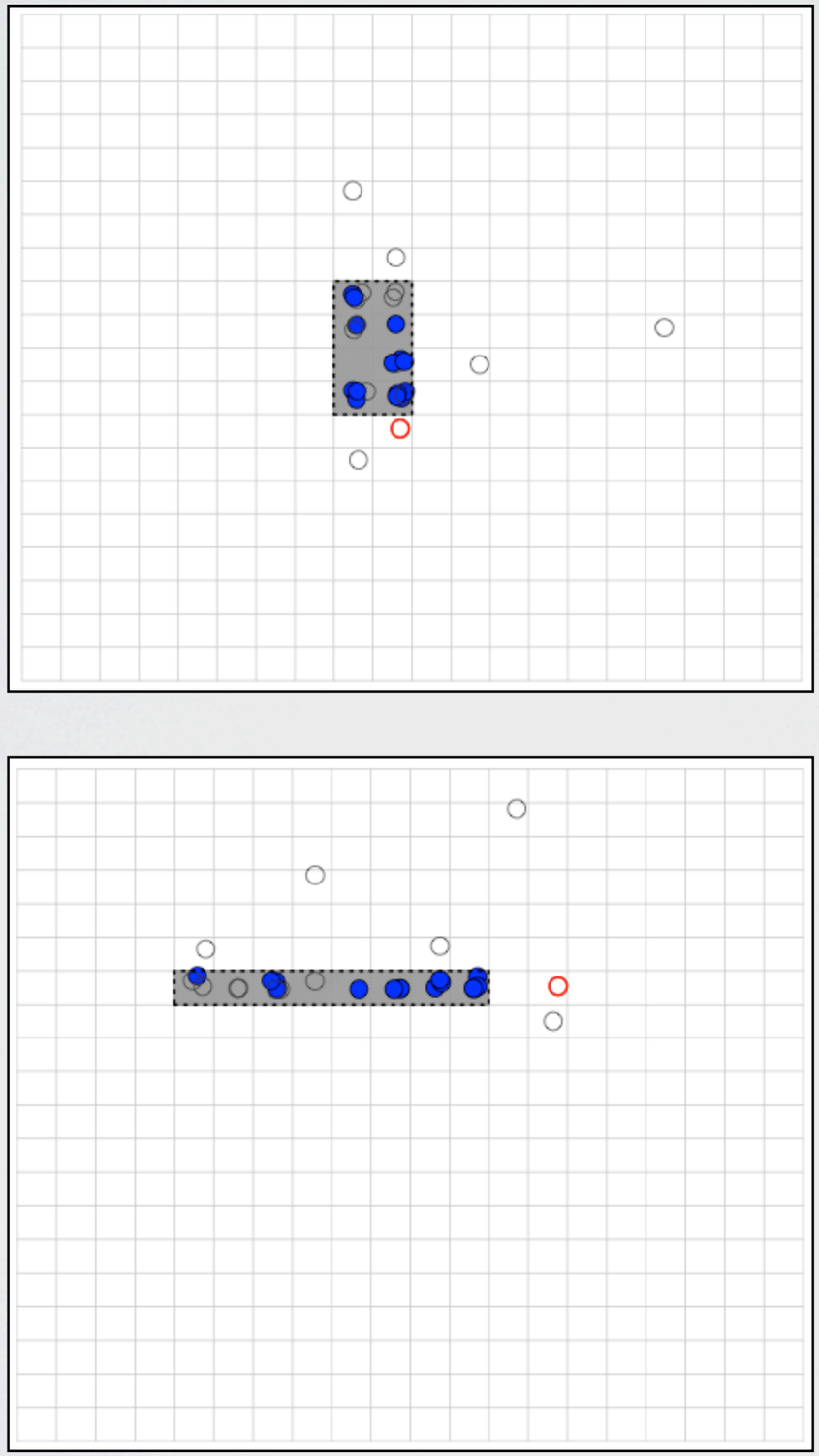




Negative feedback is more useful if the incorrect utterance is close to being correct, but isn't exactly (this way it rules out more hypotheses).

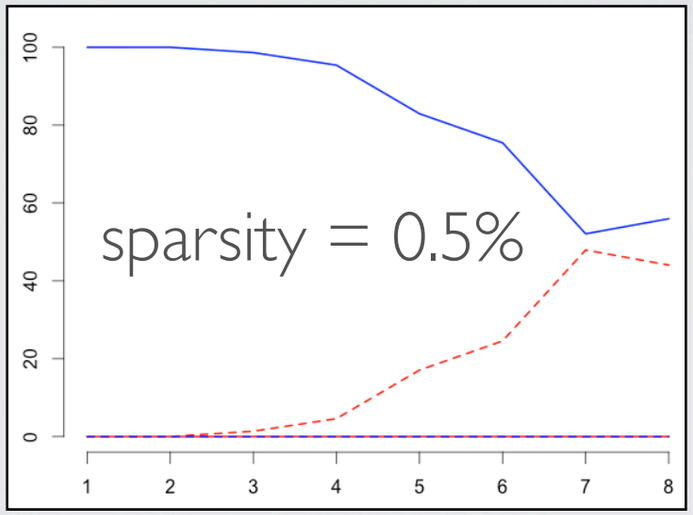
WHEN IS NEGATIVE FEEDBACK BETTER?

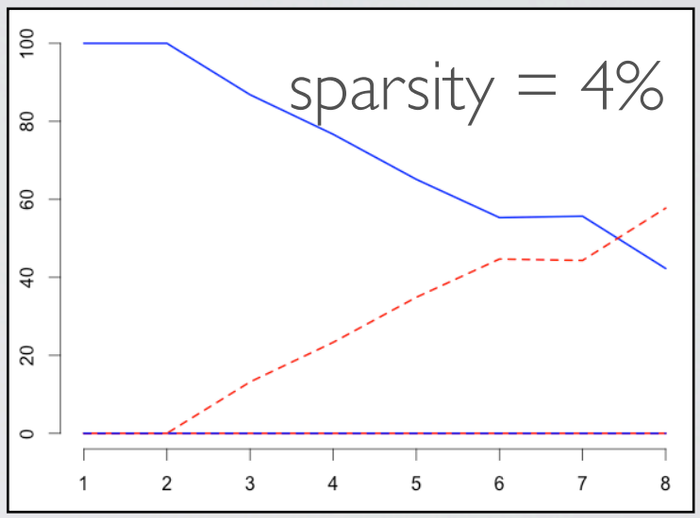
Even so, it is not always better for close-but-incorrect utterances to receive negative feedback

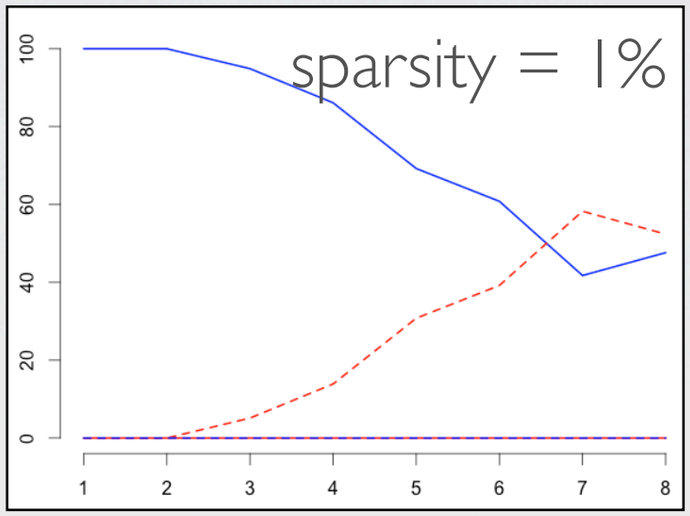


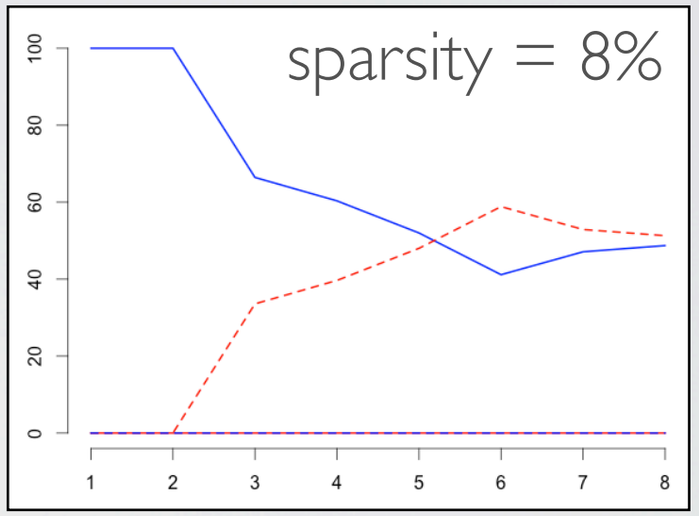
WHEN IS NEGATIVE FEEDBACK BETTER?

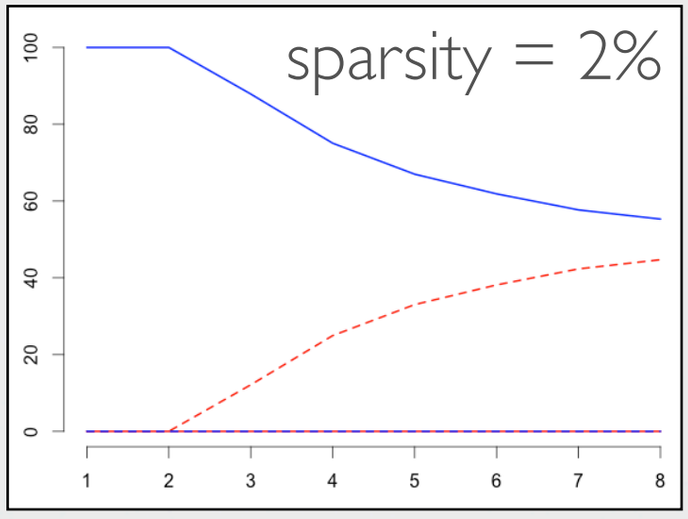
Sparsity plays a key role, as does time

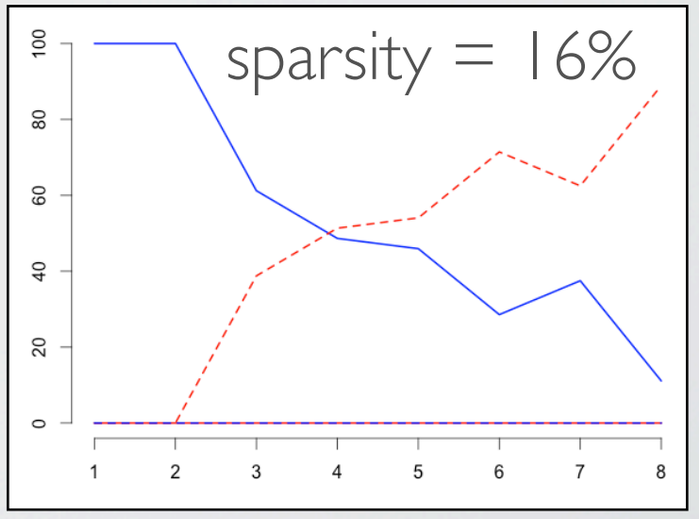






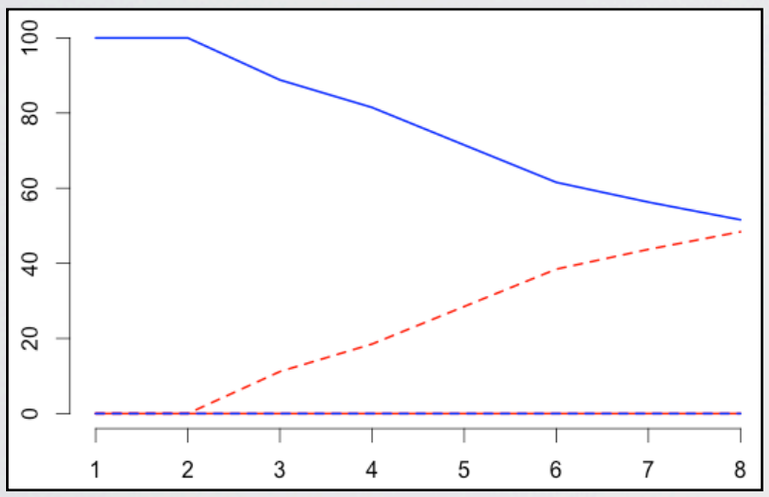




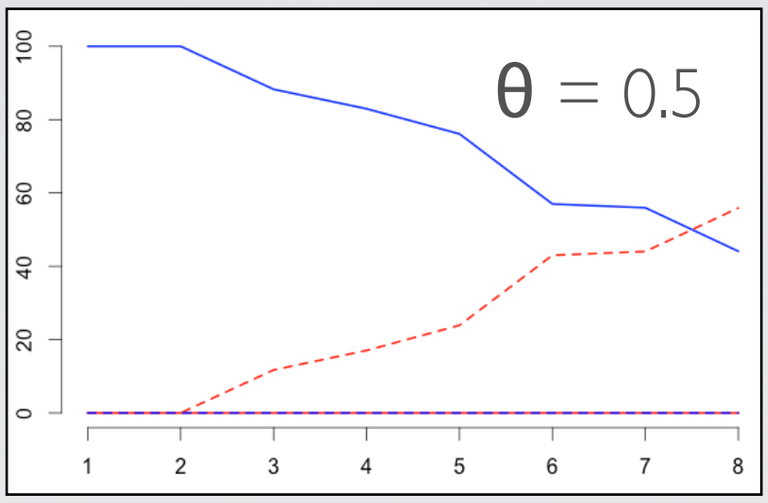


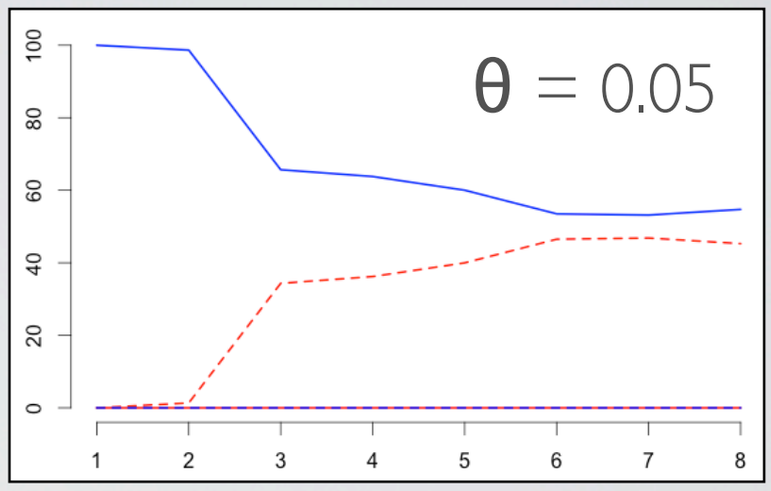
VARYING STRONG SAMPLING ASSUMPTION

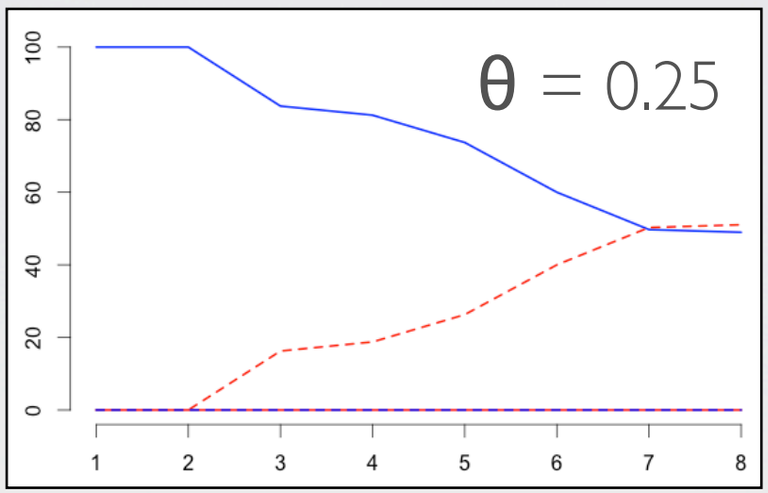
(looking at evidence as a proportion of ungrammatical utterances)

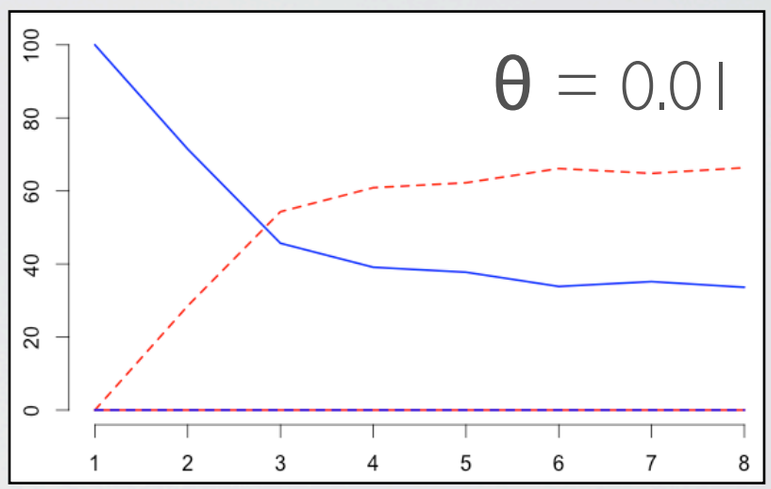


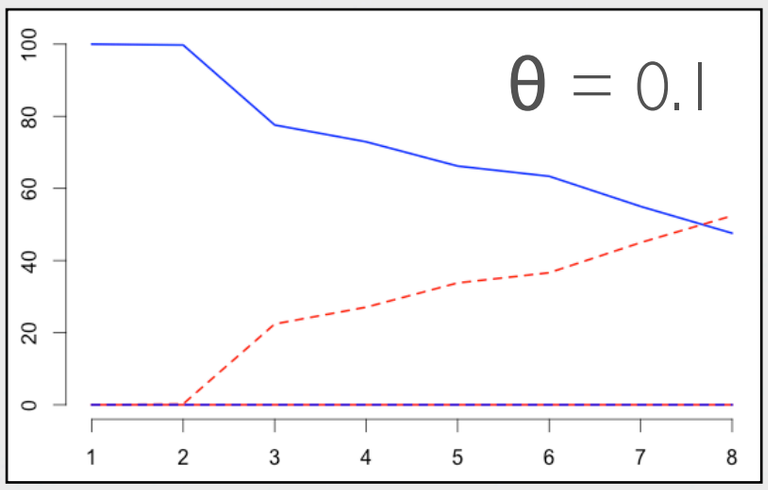
θ = 1 (strong sampling)

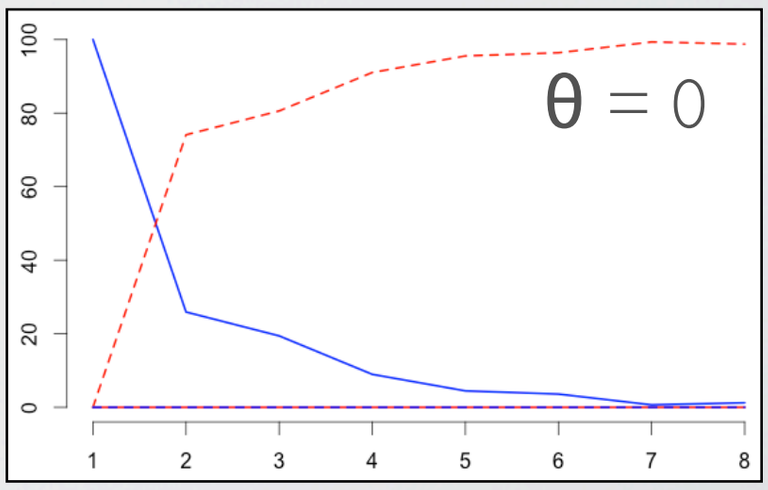






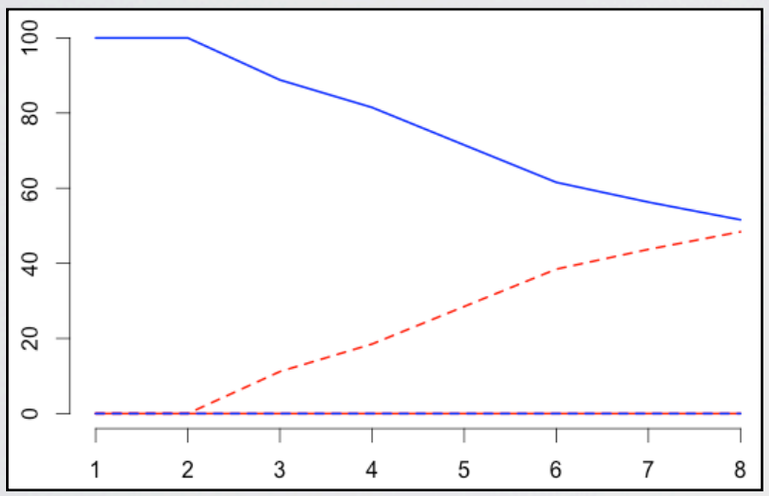






VARYING STRONG SAMPLING ASSUMPTION

(looking at evidence as a proportion of ungrammatical utterances)



θ = 1 (strong sampling)

Overall, as long as there is *any* level of tightening with additional data (i.e., as long as people do not completely assume weak sampling) the optimal level of negative feedback is within the empirically observed range

BACK TO THE BIG PICTURE

Goal was to explore the *utility* (in an information-theoretic sense) of different kinds of negative evidence.

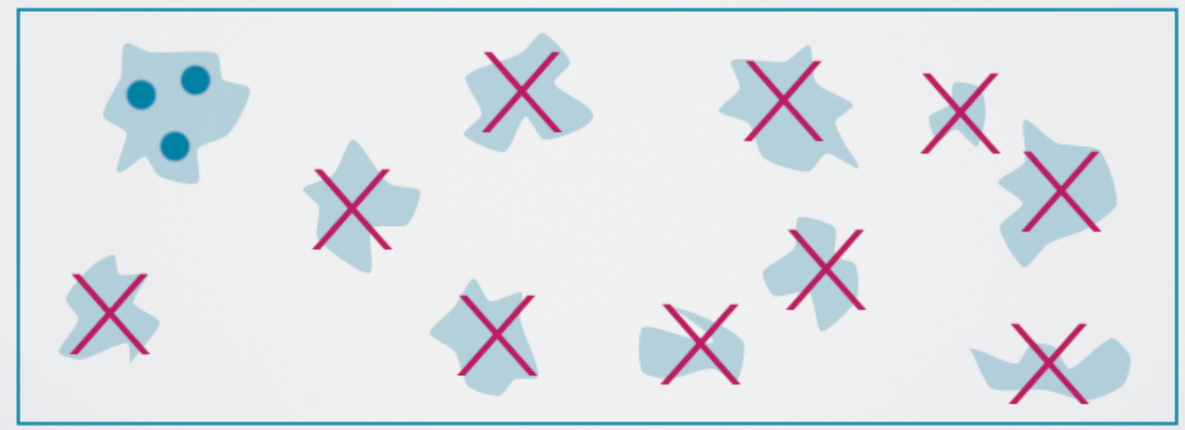
Aim was to understand what evidence is most *useful*, not just logically necessary, and *why*.

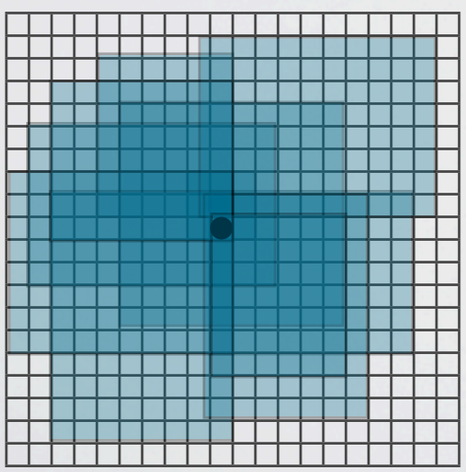
FINDINGS

Negative *examples* are always useless

This is due to language and linguistic hypotheses in general being *sparse*

Negative evidence is vastly less informative than positive evidence, when most of the hypotheses in the space are sparse



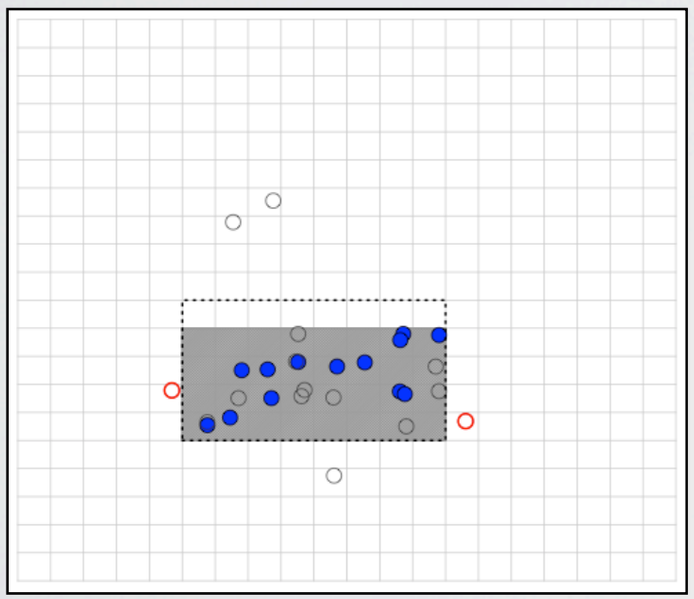


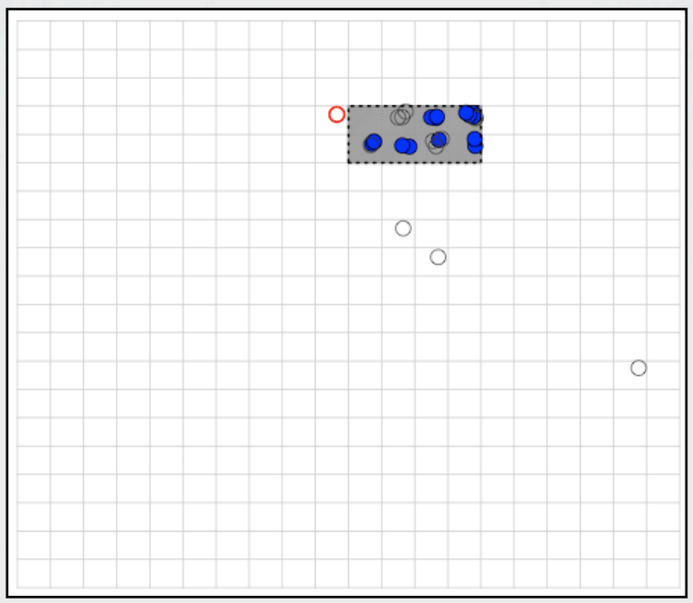
Even when they no longer are, as long as the *correct* hypothesis is sparse and people do not assume complete weak sampling, negative examples lower the probability of the correct hypothesis (relative to negative feedback)

FINDINGS

Negative *examples* are always useless

Negative *feedback* is sometimes useful, but sometimes it is more informative to instead receive positive examples from the language





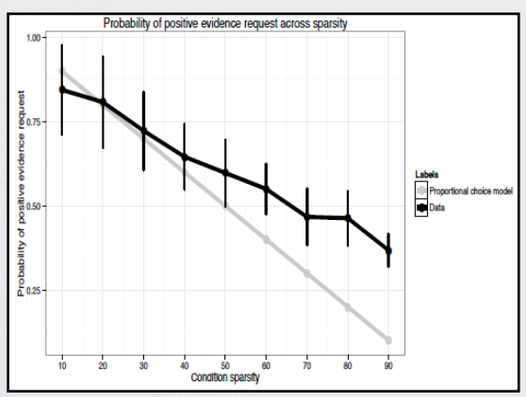
This is again because of the fact that linguistic hypotheses are *sparse*, in combination with non-weak-sampling

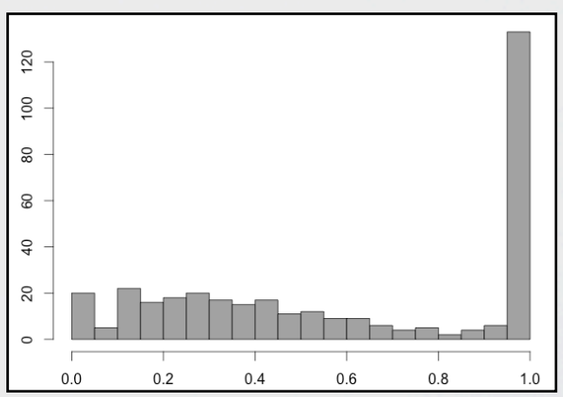
FINDINGS

Negative *examples* are always useless

Negative *feedback* is sometimes useful, but sometimes it is more informative to instead receive positive examples from the language

People are sensitive to sparsity when searching for information, and assume non-weak-sampling (at least insofar as we have been able to measure in non-linguistic contexts)





FINDINGS

Negative *examples* are always useless

Negative *feedback* is sometimes useful, but sometimes it is more informative to instead receive positive examples from the language

People are sensitive to sparsity when searching for information, and assume non-weak-sampling (at least insofar as we have been able to measure in non-linguistic contexts)

The prevalence of negative feedback predicted as optimal by our analysis is similar to the observed prevalence of negative feedback to children

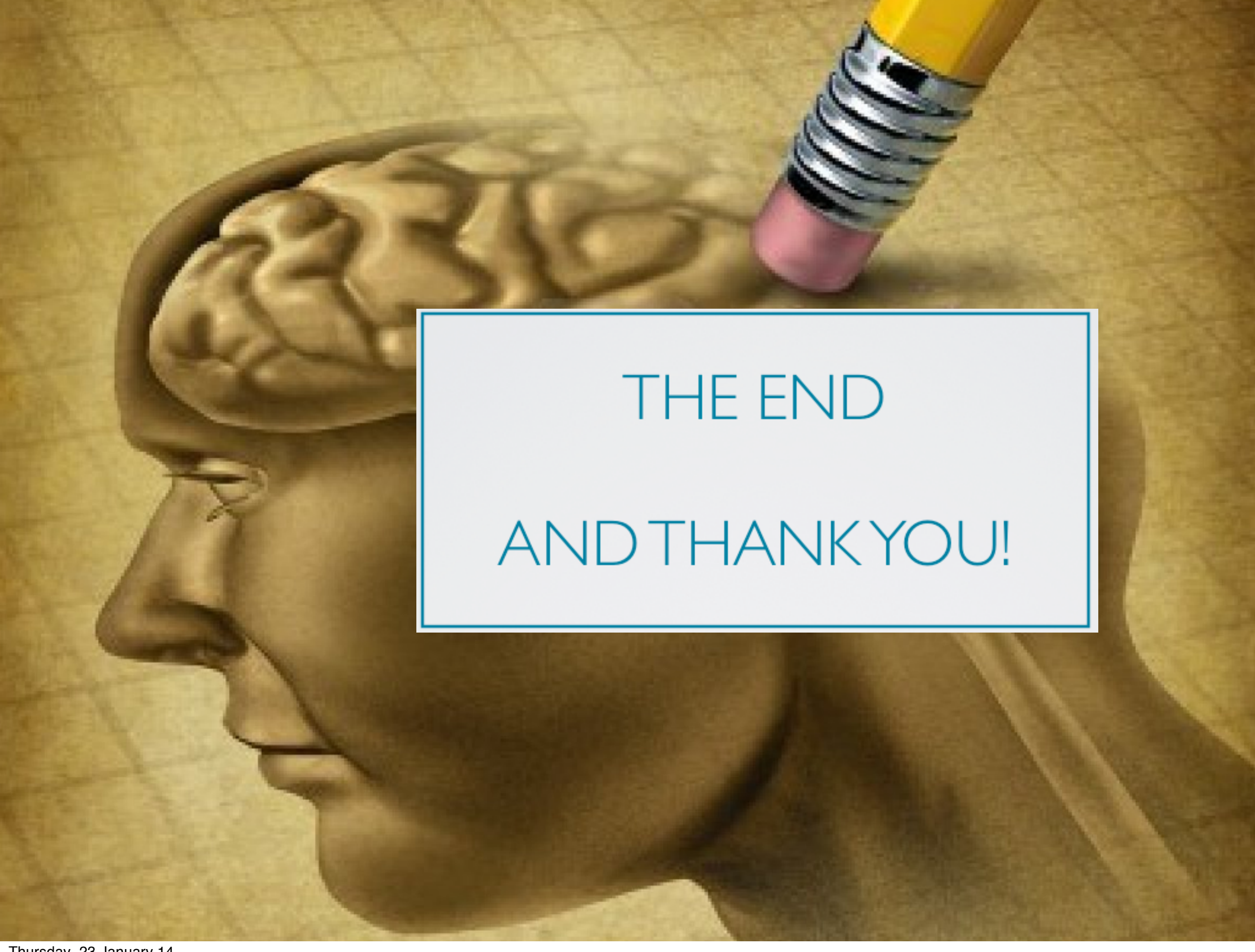
LIMITATIONS AND FUTURE STEPS

- Limitations
  - Analogy from rectangles to languages?
  - Difficulty estimating people's assumptions about sampling (θ) and sensitivity to sparsity in a linguistic context
- Future directions
  - Experimentally investigate whether people are sensitive to informational quality when asking for or giving linguistic evidence
  - Simulations with disjoint rectangles / other shapes
  - Expand to capture reformulations (combo of positive + negative)
  - Model how these results change if the child is a worse learner
  - Look at *average* quality of each kind of evidence, not best

TAKE-HOME MESSAGE

Negative evidence is not only logically unnecessary, it is often more useless (informationally) than you might think

People provide around the same kind and quantity of negative evidence as revealed as optimal by these simulations


THE END
AND THANK YOU!