LING 1010

Language and Mind

Prof. Jon Sprouse

03.01.21: The logical problem of language acquisition

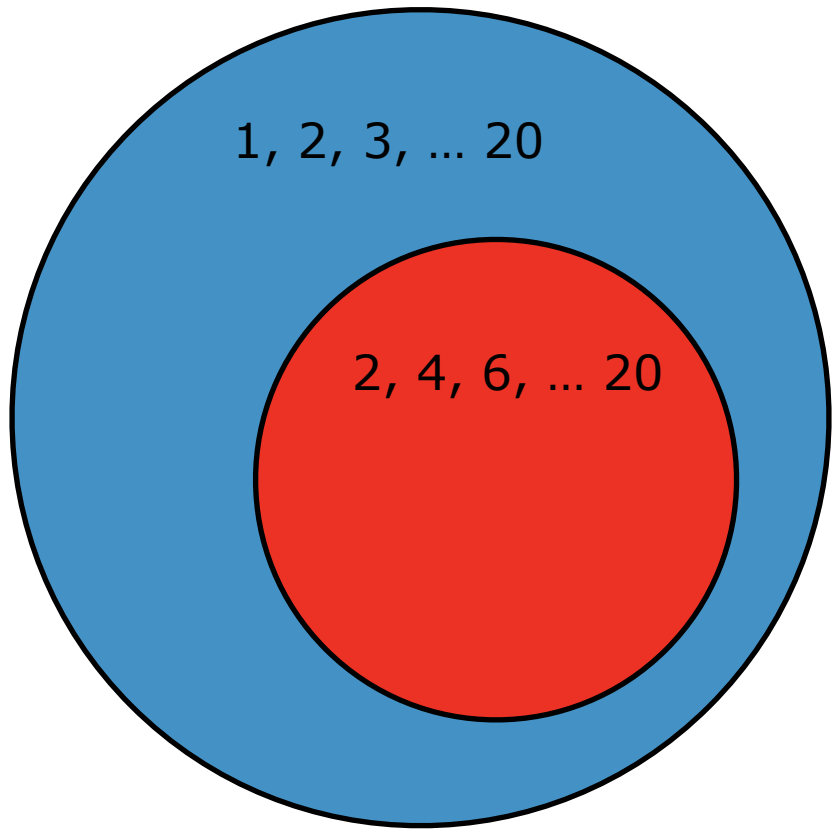
Sets and subsets

A **set** is a collection of objects.

This is the set of positive integers from 1 to 20.

A **subset** is a set (a collection of objects) that is contained within another set.

This is the set of positive, even integers from 2 to 20.



Infinite sets

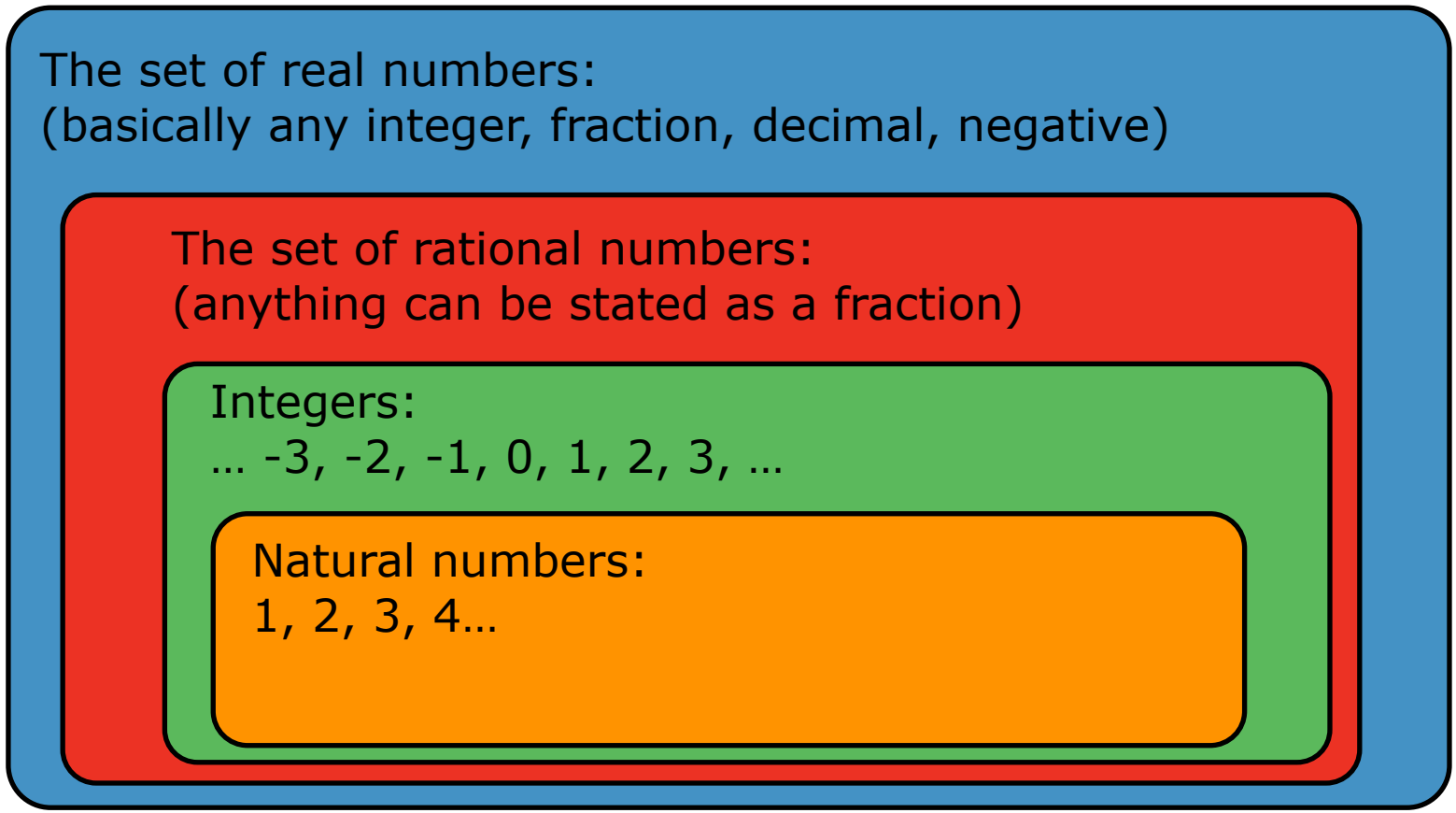
Sets can be **infinite**, which means they can contain an infinite number of items:

The set of real numbers:
(basically any integer, fraction, decimal, negative)

This is easiest to demonstrate with numbers. You know that the set of possible numbers is infinite.

Infinite subsets

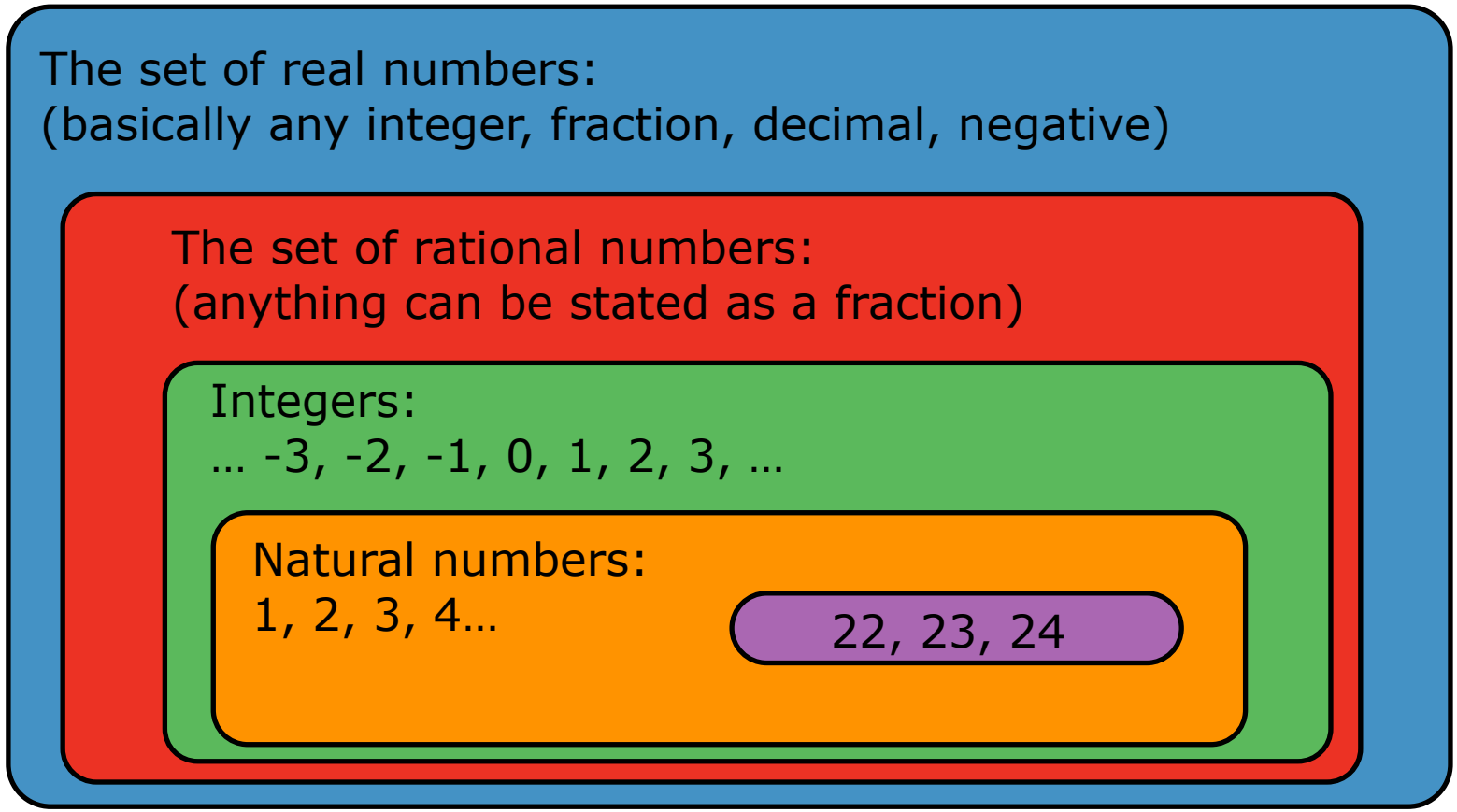
The **subsets of an infinite set** can be infinite too. Again this is easiest to see with numbers:



You know that in math we can define subsets of numbers that are themselves infinite. For example, the natural numbers (or counting numbers) are a subset of the full set of possible numbers, and they are infinite.

Finite subsets of infinite sets

Obviously, you can also have a **finite subset of an infinite set**. The purple set below is finite (it only contains 3 items).

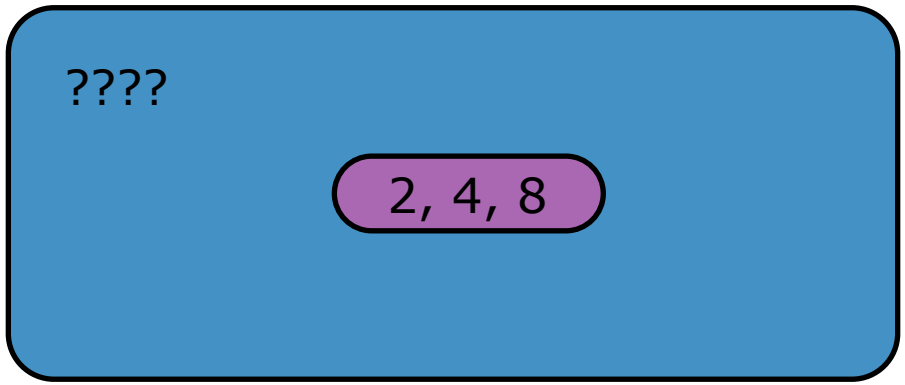


Notice that this finite subset is a member of all of the infinite sets that we've discussed so far. This is crucial, because it gives us our first glimpse of the learning problem that we face: the finite subset is not enough information to pick one (and only one) infinite set.

The learning problem

Imagine that somebody gave you the finites set: 2, 4, 8

They tell you that this **finite set** comes from an **infinite set** of numbers. It is a finite subset of an infinite set. Your task is to figure out which infinite set this finite subset comes from.



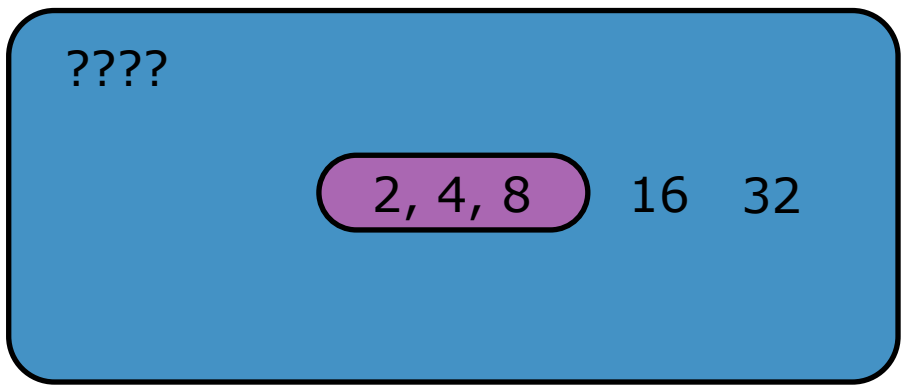
You are allowed to ask them for more information (but not the name of the infinite set). **What would you ask them?**

The moral of this example: Learning infinite sets from finite subsets is impossible without a help. And, only a certain type of help is actually helpful.

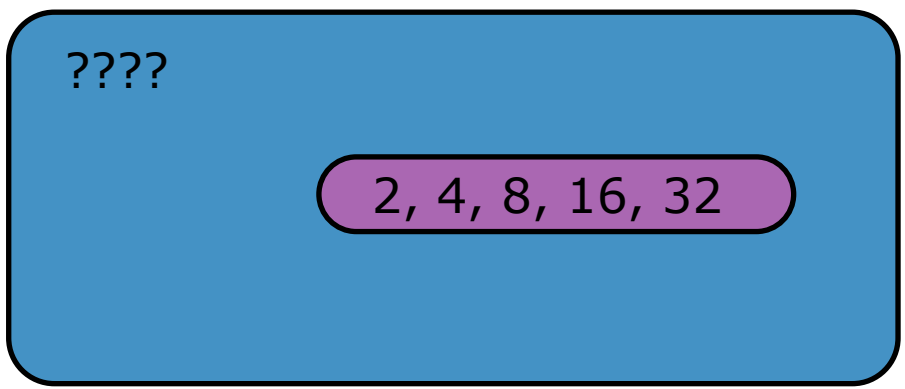
Let's try **positive evidence**

Option 1: You could ask them to produce more numbers that are in the set

We call this **positive evidence.** Positive evidence is evidence about which items are **present** in the infinite set.



Notice that this is mathematically equivalent to making the finite subset larger, so you can already see that it is probably not going to help:

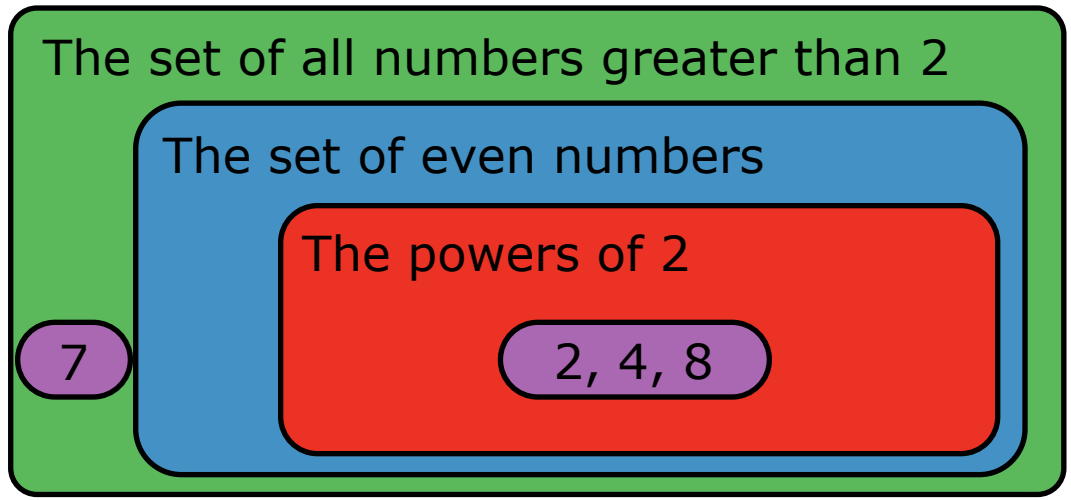


Positive evidence sometimes helps

Positive evidence can help in some cases. But it cannot guarantee successful learning.

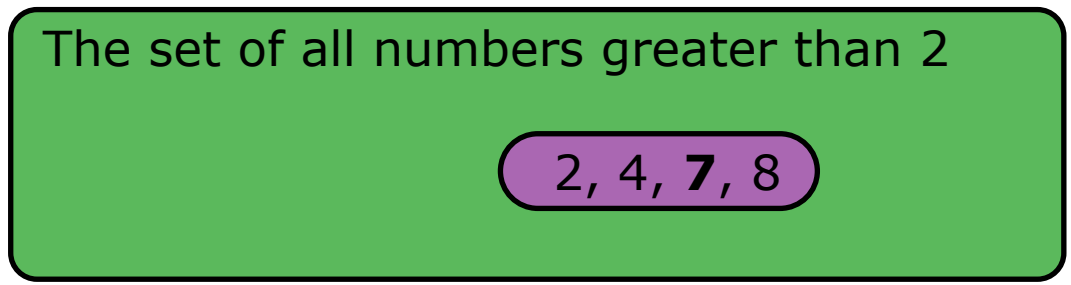
Here is a case where it helps:

Let's say you have 2,4,8, and then I add 7 to the finite set.



Here I had to draw 7 separately because it is not part of the set of powers of 2, and it is not part of the set of even numbers. It is only part of the set of all numbers greater than 2.

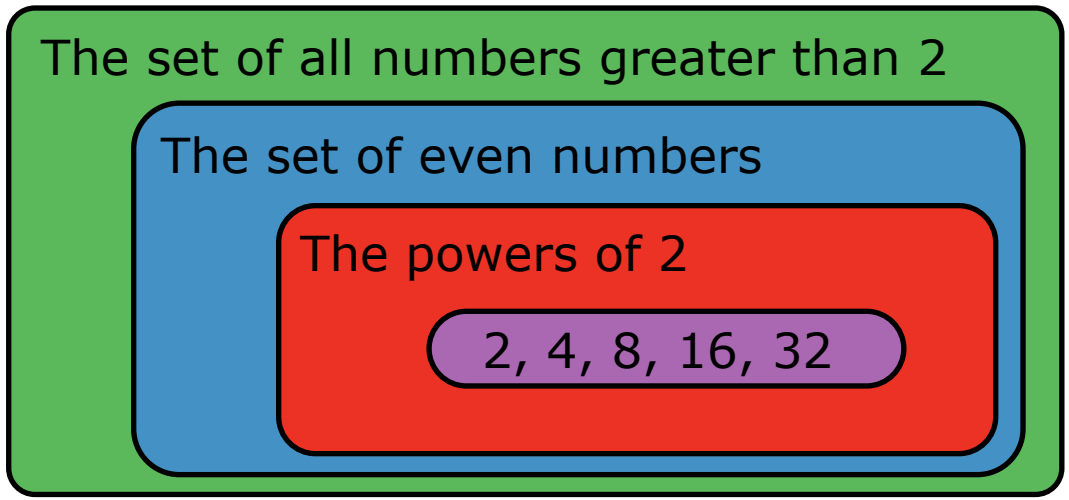
This allows us to eliminate those two hypotheses, leaving behind just the one.



But there is no guarantee

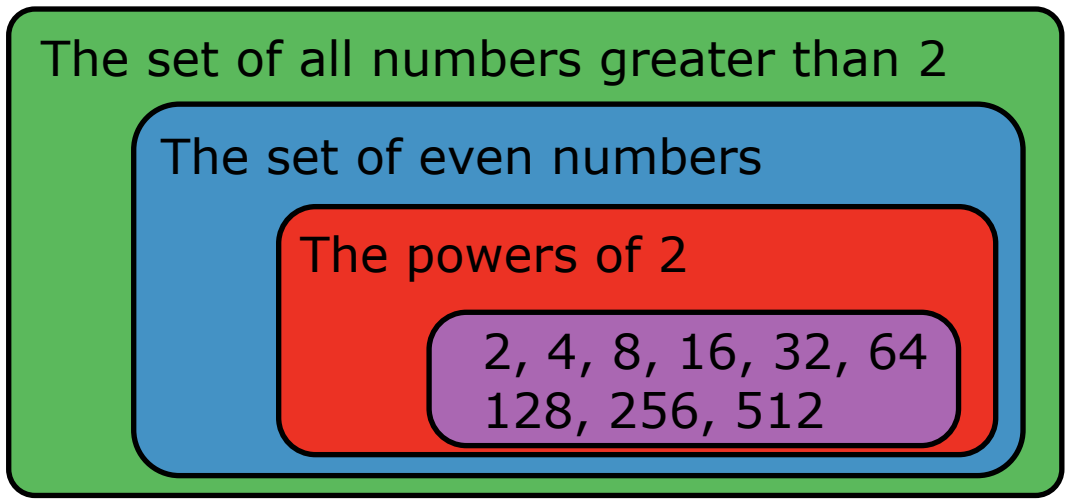
The problem with positive evidence is that there is no guarantee that we will get the relevant evidence (i.e., evidence that allows us to eliminate the incorrect sets and settle on the correct one). Here is a concrete example:

The five numbers we have so far can fit in any of these three sets.



We can add more numbers, but they could still all fit in all three sets. The extra numbers don't help us.

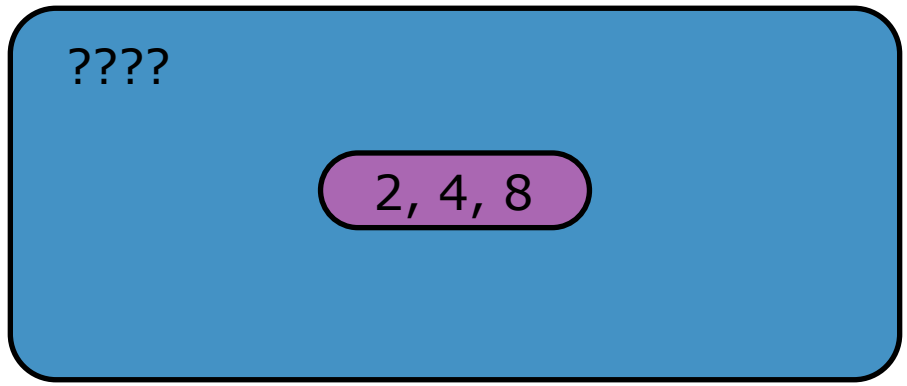
(When humans do this, it feels mean. But in nature, it could just be an accident.)



Let's try **negative evidence**

Option 2: You could ask for number that are not in the set.

We call this **negative evidence.** Negative evidence is evidence about which items are **absent** from the infinite set.



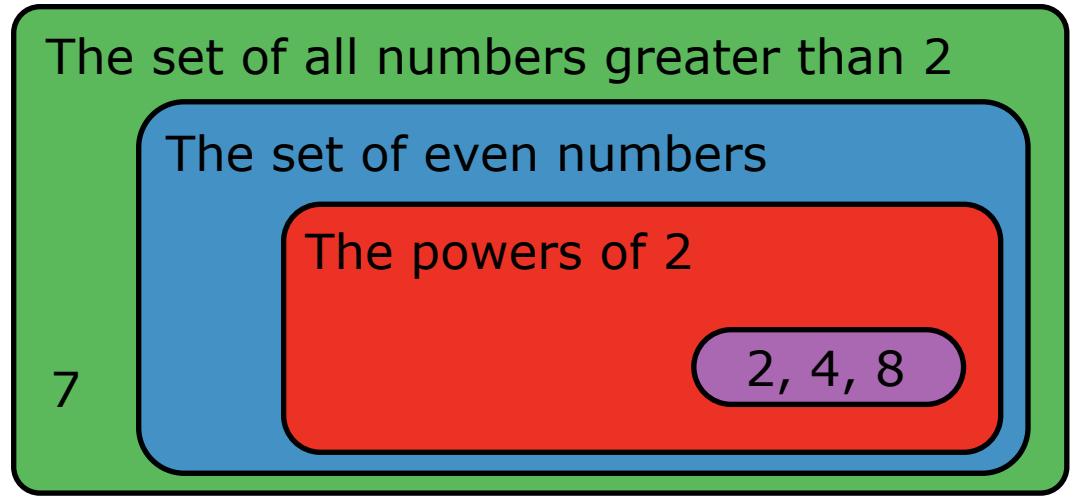
What I want to draw your attention to here is that by adding the option to have negative evidence, we can now engage in **hypothesis testing** intentionally - we can postulate some hypotheses, and then choose numbers that can tease apart the hypotheses.

If you are strategic about this, you can eliminate potential infinite sets from consideration. But it is only possible if negative evidence is available.

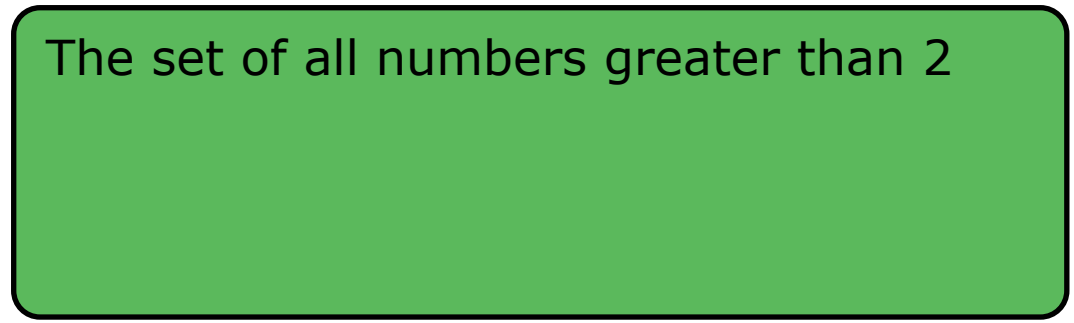
Hypothesis testing with negative evidence

Let's say we have the positive evidence 2, 4, 8. This leads us to create three hypotheses:

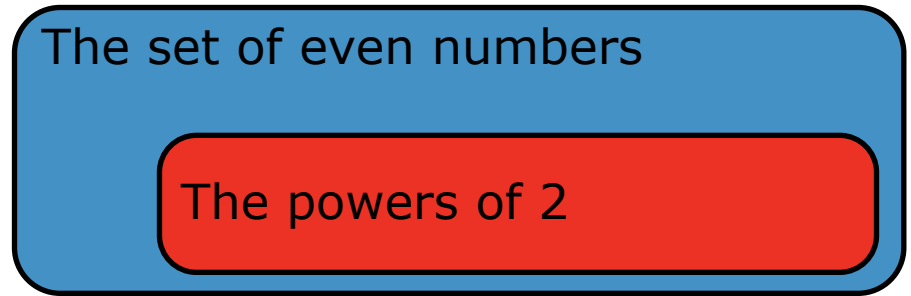
Here is what I do: I choose a number that would be in the outer set but not in the subsets, like the number 7, and ask for feedback.



If the answer is "yes", it is positive evidence for the green set. This is like our earlier positive example, but here we guaranteed success by using hypothesis testing.



If the answer is "no", it is negative evidence for the green set. This lets us eliminate the green hypothesis from consideration!



Positive Evidence and Negative Evidence

Positive Evidence: Evidence about which items are **present** in the infinite set.

The problem with positive evidence is that there is no guarantee that the relevant evidence will be given. It might happen; it might not.

Negative Evidence: Evidence about which items are **absent** from the infinite set.

The benefit of adding negative evidence is that you can engage in intentional hypothesis testing. This means you can strategically test hypotheses to eliminate the incorrect ones from consideration!

How does this matter for language?

Language learning is the generalization from a finite subset to an infinite set

Fact 1: All human languages can be characterized as an infinite set of sentences

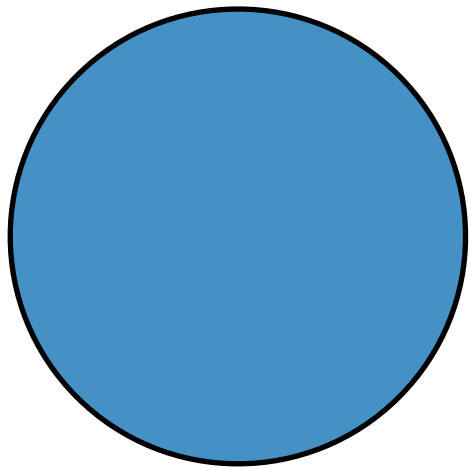
Sarah wrote a novel.

Lisa claims that Sarah wrote a novel.

Mary thinks that Lisa claims that Sarah wrote a novel.

John said that Mary thinks that Lisa claims that Sarah wrote a novel.

Language is an infinite set of sentences



Language learning is the generalization from a finite subset to an infinite set

Fact 1: All human languages can be characterized as an infinite set of sentences

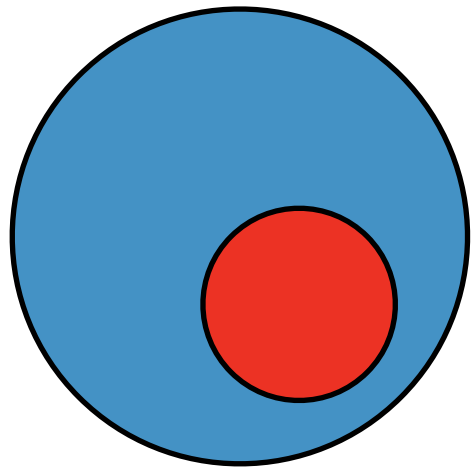
John said that Mary thinks that Lisa claims that …

Fact 2: The input that children receive when learning their language is finite.

Fact 2 has to be true because human learn language in a finite amount of time (maximally 14 years, more likely ~6 years).

Language is an infinite set of sentences

language input is a finite set of sentences that is a subset of the infinite set of the language



Language learning is the generalization from a finite subset to an infinite set

Fact 1: All human languages can be characterized as an infinite set of sentences

John said that Mary thinks that Lisa claims that …

Fact 2: The input that children receive when learning their language is finite.

Fact 3: All children succeed at language acquisition, except for atypical circumstances (diseases, disorders, imprisonment, etc).

As one professor once remarked to me: "We talk about the literacy rates of different countries. But have you ever heard anyone talk about the speaking rates of different countries?"

Of course not. That is because **everybody succeeds** at learning language.

Language learning is the generalization from a finite subset to an infinite set

Fact 1: All human languages can be characterized as an infinite set of sentences

John said that Mary thinks that Lisa claims that …

Fact 2: The input that children receive when learning their language is finite.

Fact 3: All children succeed at language acquisition, except for atypical circumstances (diseases, disorders, imprisonment, etc).

Fact 4: Combining positive and negative evidence would allow for hypothesis testing, and guarantee that the infinite set can be learned from a finite subset.

Obviously, children have positive evidence

It is fairly uncontroversial to claim that children have access to **positive evidence.** After all, positive evidence is just another name for the actual items in the set, and people do speak to (and around) children. So children definitely see examples of the items in the infinite set!

There is a freely available corpus (corpus means body or collection) of transcripts of children engaging in conversation across a wide range of ages (and several languages). It is called the Child Language Data Exchange System, or CHILDES:

http://childes.psy.cmu.edu/

You can use this corpus to see the positive evidence that children receive while learning their native languages

But do they have negative evidence?

The trickier question is wether there is negative evidence available to children.

Negative evidence in language would be some sort of response by the parent after a child produces an ungrammatical sentence. But crucially not after a grammatical sentence.

This response need not be an explicit correction. It could take any number of forms (these are adapted from Marcus 1993):

Explicit disapproval:	Parent says no or shakes head.
Non sequiturs:	Parent fails to understand child.
Repetitions:	Parent repeats the child utterance.
Recasts:	Parent corrects the child utterance.
Questions:	Parent asks for more information.

Problem 1: Children **ignore** negative evidence

child:	Want other one spoon, Daddy.
parent:	You mean, you want the other spoon.
child:	Yes, I want other one spoon, please Daddy.
parent:	Can you say “the other spoon”?
child:	Other… one… spoon.
parent:	Say “other”.
child:	Other.
parent:	“Spoon”.
child:	Spoon
parent:	“Other spoon”.
child:	Other… spoon. Now give me other one spoon?

… or **misinterpret** it

child: Nobody don’t like me.

parent: No, say “nobody likes me”.

child: Nobody don’t like me.

parent: No, say “nobody likes me”.

child: Nobody don’t like me.

parent: No, say “nobody likes me”.

child: Nobody don’t like me.

parent: No, say “nobody likes me”.

child: Nobody don’t like me.

parent: No, say “nobody likes me”.

child: Oh! Nobody don’t likes me.

Problem 2: Feedback from parents is **noisy**

Here is another problem: parents provide feedback of various kinds after **both** ungrammatical sentences and grammatical sentences.

A study by Bohannon and Stanowicz 1988 found that parents gave feedback to children after ungrammatical sentences 35% of the time; and they gave feedback to children after grammatical sentences 14% of the time.

This means that children couldn't be sure that the correction was because of the sentence being ungrammatical, or just because parents like to give feedback. In other words, feedback is **noisy**. It is not a clear indicator of ungrammaticality.

I won't go into the math, but Marcus 1993 calculated that the rates of feedback for ungrammatical and grammatical sentences mean that **children would have to repeat a sentence 85 times** in order to determine whether the feedback that they were receiving was because it was ungrammatical, or whether it was because it was grammatical (i.e., to figure out if it is the 35% rate or 14% rate). Obviously, children don't repeat sentences 85 times to figure out if they are part of the language or not.

The logical problem of language acquisition

And now we are ready to lay out the logical problem of language acquisition:

Fact 1: All human languages can be characterized as an infinite set of sentences

Fact 2: The input that children receive when learning their language is finite.

Fact 3: All children succeed in learning language.

Fact 4: Combining positive and negative evidence would allow for hypothesis testing, and guarantee that the infinite set can be learned from a finite subset.

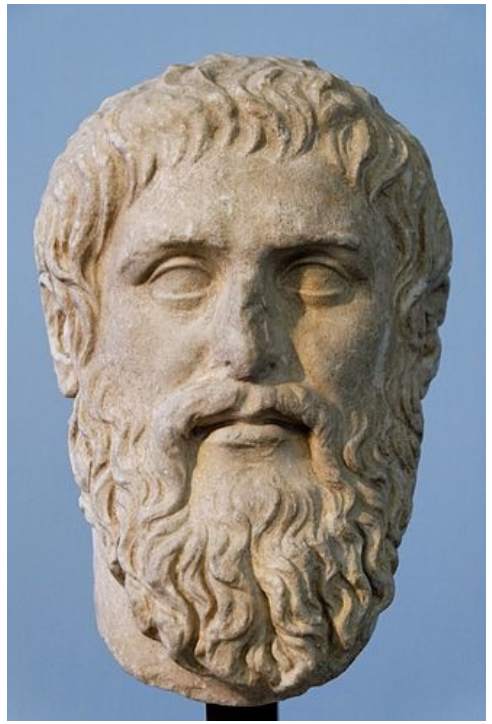
Fact 5: But children do not make use of negative evidence.

Conclusion: Children learn an infinite set from a finite set, but don't use the one method (combining positive and negative evidence in hypothesis testing) that would guarantee the solution.

This is the logical problem of language acquisition: Children are able to learn language despite not having enough evidence to learn it!

This is sometimes called Plato's problem

You've probably heard of Plato before. He was a Greek philosopher/thinker who lived from ~427BC to ~347BC (80 years!). He was a student of Socrates, and the founder of the Academy, most likely the first "university" in the western world.



Plato investigated hundreds of complex questions in his lifetime. One of them was the question of how we humans can have so much knowledge, when the environment provides so little evidence to help us build that knowledge.

If we apply this question to linguistic knowledge, it becomes the **logical problem of language acquisition**. It is also sometimes called the **poverty of the stimulus**, because it highlights the fact that the input (the stimulus) is too poor (impoverished/poverty) to fully specify the knowledge that we learn.

Whatever name you choose, this is a deep mystery about human knowledge. How is it that we learn all that we do from the input that we receive? It has puzzled thinkers for thousands of years, and still puzzles us to this day.

Pure Nativism vs Pure Empiricism

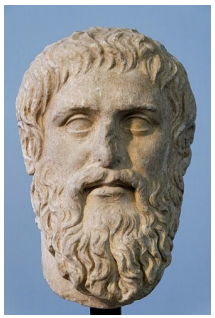
The logical problem of language acquisition is one specific example of a debate between two (extreme) positions about where knowledge comes from:

Pure Nativism

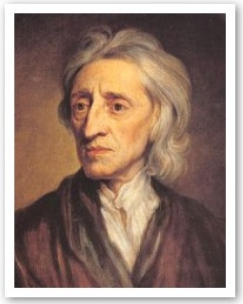
Pure Empiricism

The Nativism solution says that human biology solves the problem. All knowledge is innate (present from birth), and we simply bring that knowledge out as we grow.

Empiricism denies that there is a problem. It says that all knowledge comes from experience (or the input). We just have to figure out how that happens.



Nativism was the solution first proposed by Plato to Plato's problem.

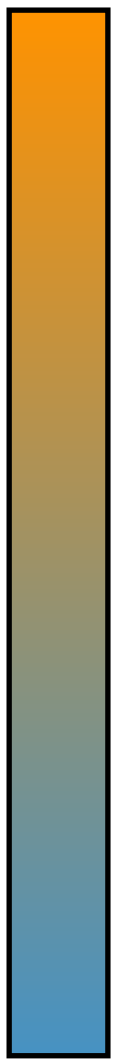


Empiricism was most strongly advocated by John Locke, an English philosopher who lived from 1632-1704.

These are **extreme positions**. Nobody in the modern age of cognitive science believes these are correct. Instead, we explore theories in the middle.

The spectrum of Nativism and Empiricism

One tricky aspect of the debate between Nativism and Empiricism is that the two ideas actually form a spectrum:



Pure Nativism: (Plato)
all knowledge comes from biology,
no knowledge comes from experience

Modern Nativism:
substantial knowledge comes from biology, but experience/input still plays an important role

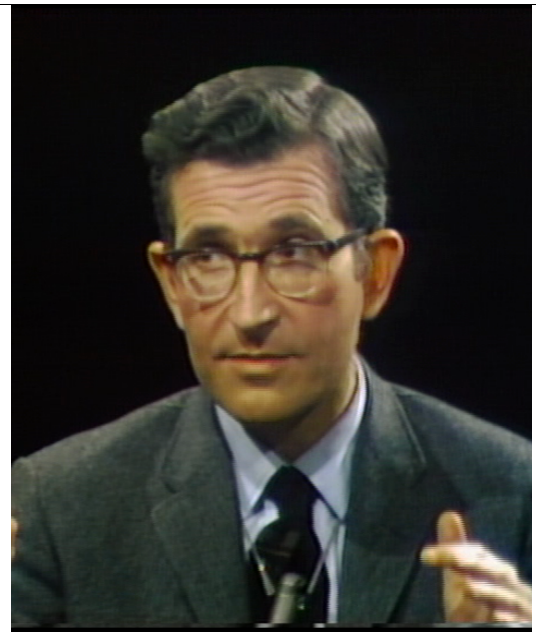
Modern Empiricism:
some knowledge comes from biology, but experience/input plays the most important role

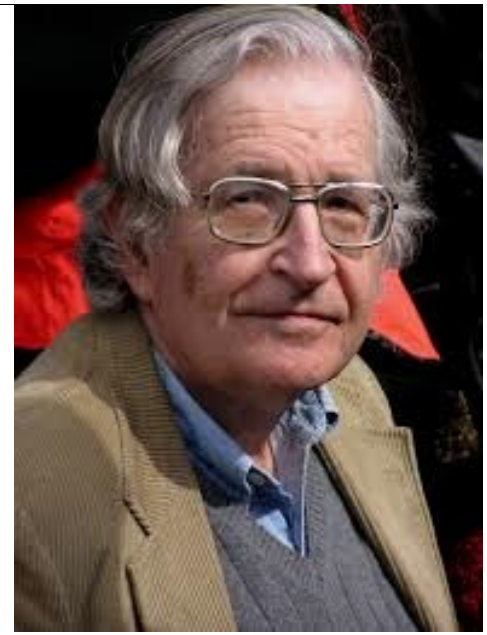
Pure Empiricism: (Locke)
all knowledge comes from experience,
no knowledge comes biology

Noam Chomsky and Modern Nativism

Noam Chomsky is an American linguist, often considered the father of modern linguistics and a major figure in cognitive science.

I am not exaggerating when I say that his ideas have been instrumental in shaping all of the studies we've discussed in this class. He is also a modern proponent of a Nativist approach to language learning.





born in 1928

Noam Chomsky is also known for his writings about political science and government, but in this class we will focus on his contributions to linguistics.

Modern Nativism vs Modern Empiricism

So here is the more moderate views that are debated today… which we can apply to the logical problem of language acquisition in this class.

Modern Nativism

Substantial innate knowledge

Input/experience still plays a role, but less than the role it plays in empiricism.

In short, children come to the problem with a lot of genetic help, and then use experience to hone in on the correct answer.

Modern Empiricism

Minimal innate knowledge

Input/experience plays the largest role in learning

In short, children come to the table with the ability to learn from experience, and use experience to build up all of the complexity of language.

What could that innate knowledge be?

Well, the theory of language that we have built so far already provides a starting point for investigating this. Over the next few lectures, we will explore these ideas by looking at actual facts about child language acquisition!

Principles:	The fact that all languages share certain properties might indicate that those properties are hardwired in some way.
Parameters:	If parameters were built-in, then the learning problem would be simpler: children just need to figure out the right values.
The ability to learn complex rules:	Phonology, morphology, and syntax all seem to be predicated upon complex rules, suggesting that humans have the ability to learn complex rules.
The ability to learn phonemes and morphemes:	Our memorization abilities must be powerful enough to learn the phonemes and morphemes of our languages.

Some Conclusions

Positive Evidence is evidence about which items are **present** in the infinite set. The problem with positive evidence is that there is no guarantee that the relevant evidence will be given. It might happen; it might not.

Negative Evidence is evidence about which items are **absent** in the infinite set. The benefit of negative evidence is that you can engage in intentional hypothesis testing, and eliminate potential infinite sets from consideration.

Children appear to be able to learn language (an infinite set) from finite input (a finite set) without using negative evidence.

This is the logical problem of language acquisition: Children are able to learn language despite not having enough evidence to learn it!

Nativism is the idea that knowledge can be specified by biology (innate knowledge). Modern nativism still allows for a role of input/experience.

Empiricism is the idea that knowledge comes from input/experience (not biology). Modern empiricism still allows for a small role for biology.