

Download slides: <https://polgreen.github.io/pdfs/synth.pdf>

Program Synthesis

(A brief introduction)

Elizabeth Polgreen, 10th October



**Automatically generating code
that satisfies the user's
specification**

**Hasn't this all
been solved?**



GitHub
Copilot

GitHub Copilot: A potential security risk?

By amavsharma

MAY 10, 2023



(Submitted on 20 Aug 2021 (v1), last revised 16 Dec 2021 (this version, v3))

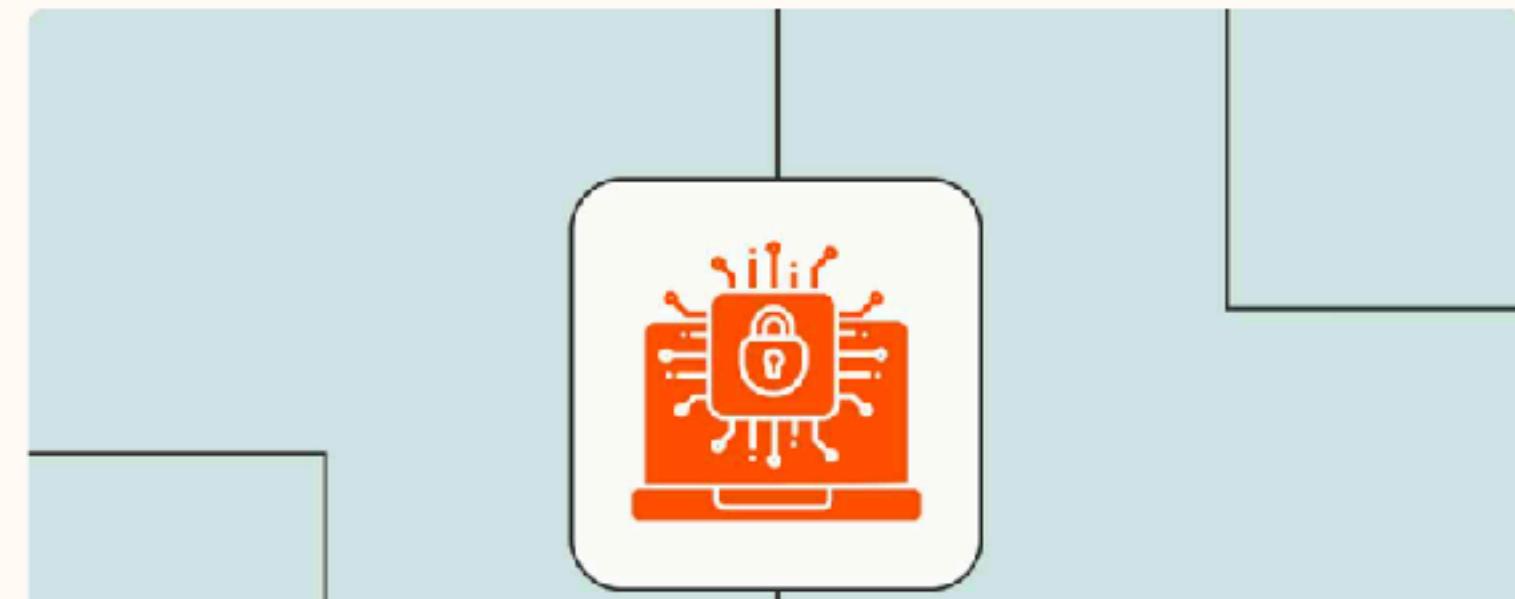
Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions

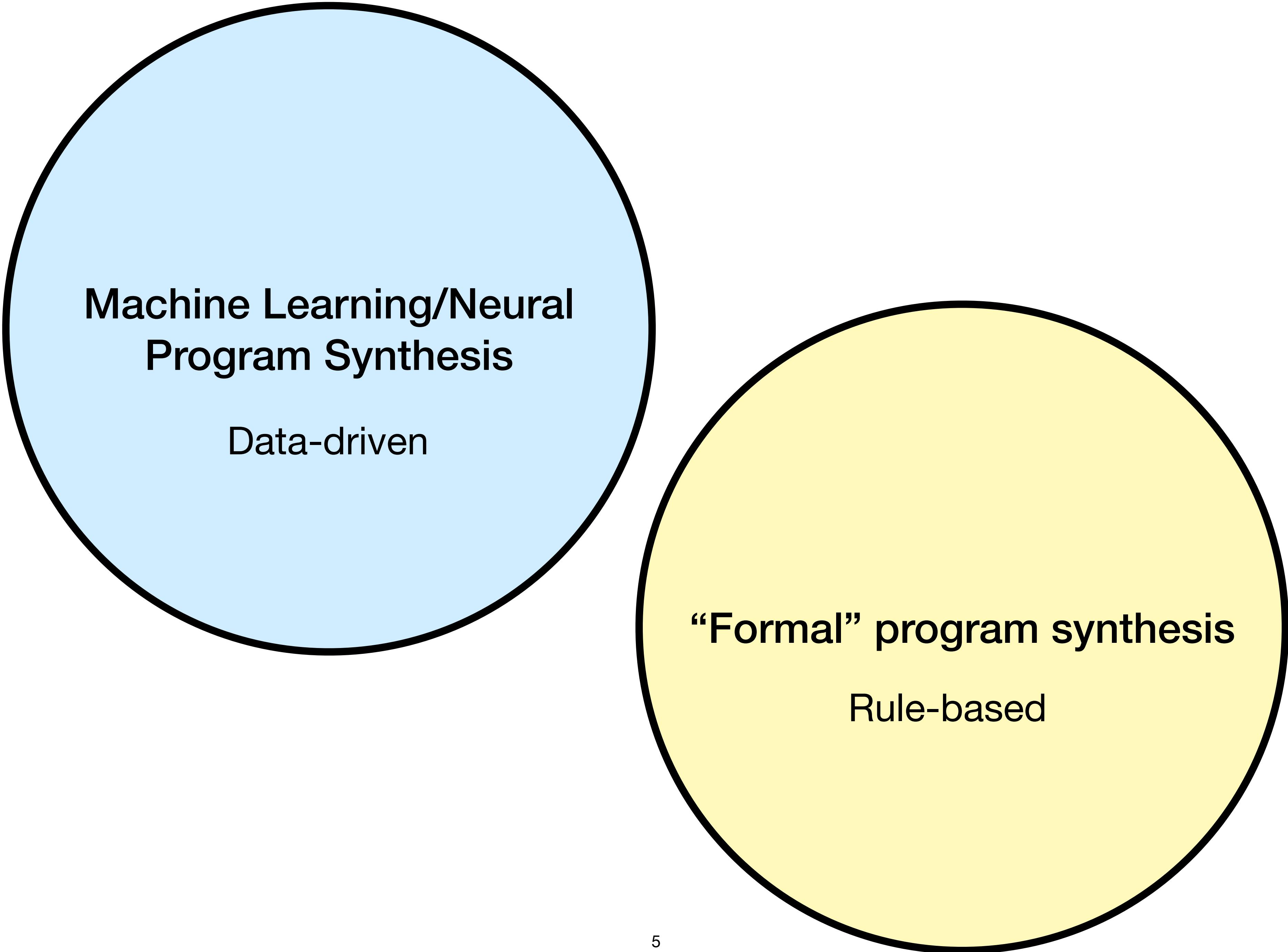
Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, Ramesh Karri

There is burgeoning interest in designing AI-based systems to assist humans in designing computing systems, including tools that automatically generate computer code. The most notable of these comes in the form of the first self-described 'AI pair programmer', GitHub Copilot, a language model trained over open-source GitHub code. However, code often contains bugs – and so, given the vast quantity of unvetted code that Copilot has processed, it is certain that the language model will have learned from exploitable, buggy code. This raises concerns on the security of Copilot's code contributions. In this work, we systematically investigate the prevalence and conditions that can cause GitHub Copilot to recommend insecure code. To perform this analysis we prompt Copilot to generate code in scenarios relevant to high-risk CWEs (e.g. those from MITRE's "Top 25" list). We explore Copilot's performance on three distinct code generation axes -- examining how it performs given diversity of weaknesses, diversity of prompts, and diversity of domains. In total, we produce 89 different scenarios for Copilot to complete, producing 1,689 programs. Of these, we found approximately 40% to be vulnerable.

5 security risks of generative AI and how to prepare for them

By Elisa Silverman · June 14, 2023





Machine Learning/Neural Program Synthesis

Data-driven

“Formal” program synthesis

Rule-based

A Venn diagram consisting of two overlapping circles. The left circle is light blue and labeled "Machine Learning/Neural Program Synthesis". The right circle is light yellow and labeled "'Formal'" program synthesis". The overlapping region between the two circles is shaded green and contains the text "Data-driven" from the blue circle and "Rule-based" from the yellow circle.

**Machine Learning/Neural
Program Synthesis**

Data-driven

“Formal” program synthesis

Rule-based

This lecture

- What is formal synthesis
 - CounterExample Guided Inductive Synthesis:
 - Guessing
 - Checking
- Applications of formal synthesis
 - Translating code from one language to another
 - Synthesising invariants/ranking functions
 - Learning from sparse data
 - Helping machine learning to play Atari games

Formal Program Synthesis

$$\exists P \forall x . \sigma(P, x)$$

Does there exist a function P such that, for all possible inputs x , the specification σ will evaluate to true for P and x .

Formal Program Synthesis

$$\exists P \forall x . \sigma(P, x)$$

Does there exist a function P such that, for all possible inputs x , the specification σ will evaluate to true for P and x .

σ is a quantifier free formula in a background theory,
e.g., Linear Integer Arithmetic

NB: we can write specs with input-output examples as quantifier free formula

Formal Program Synthesis

```
int f(int x, int y)
{
    ???
}
```

@ensures: $\text{@ret} \geq x \wedge \text{@ret} \geq y \wedge (\text{@ret} = x \wedge \text{@ret} = y)$

Formal Program Synthesis

```
int f(int x, int y)
{
    ???
}
@ensures: @ret ≥ x ∧ @ret ≥ y ∧ (@ret = x ∨ @ret = y)
```

$$\exists f. \forall x, y. f(x, y) \geq y \wedge f(x, y) \geq x \wedge (f(x, y) = x \vee f(x, y) = y)$$

Defining the search space

Syntax-Guided Synthesis

```
int f(int x, int y)
{
    ???
}

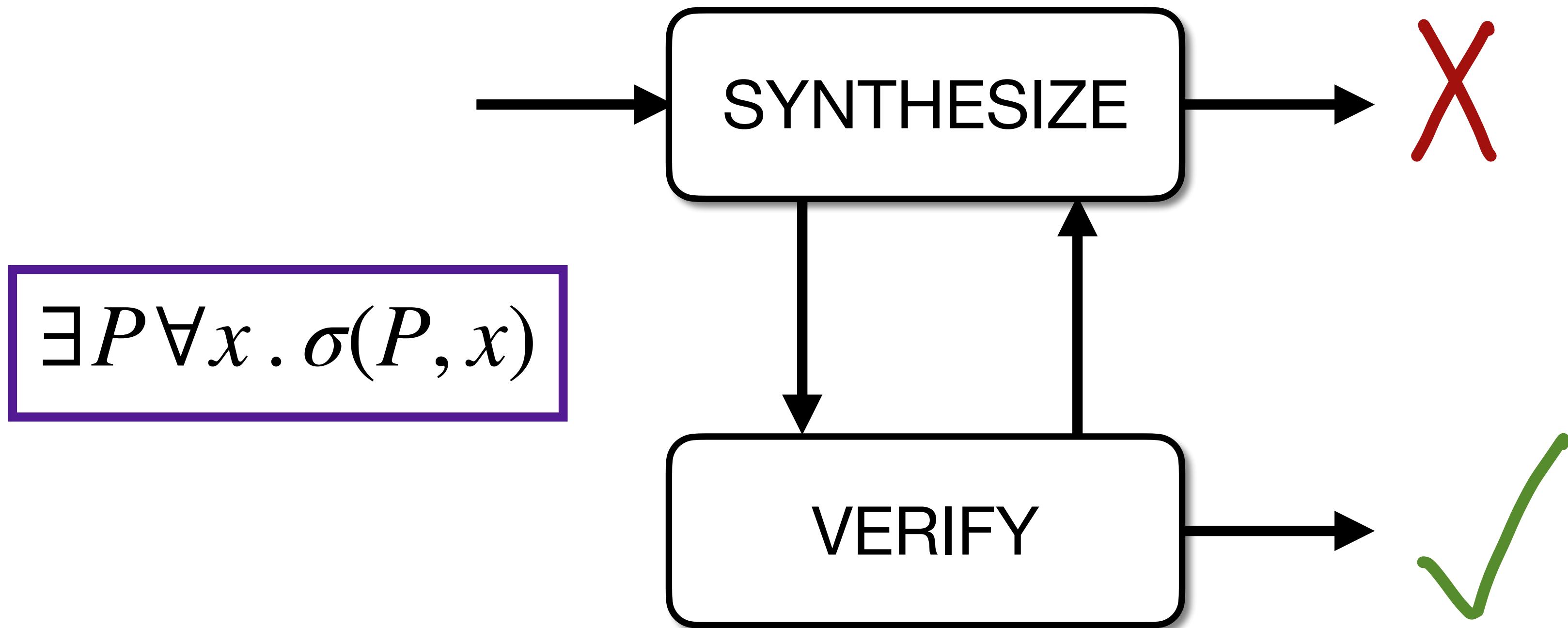
@ensures: @ret ≥ x ∧ @ret ≥ y ∧ (@ret = x ∧ @ret = y)
```

$$\begin{aligned} A \rightarrow & A+A \mid -A \mid x \mid y \mid 0 \mid 1 \mid \text{ite}(B, A, A) \\ B \rightarrow & B \wedge B \mid \neg B \mid A=A \mid A \geq A \mid \perp \end{aligned}$$

Context Free Grammar (with no cycles)

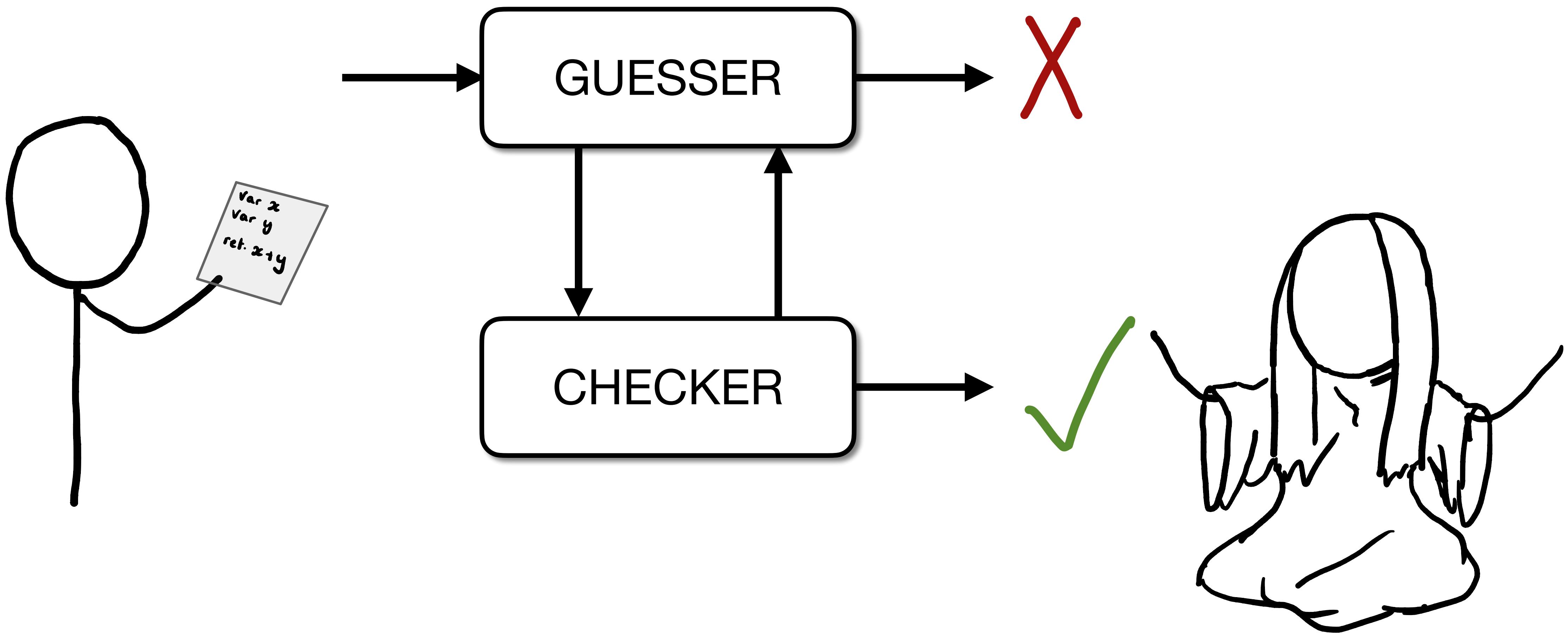
Algorithms for formal synthesis

Counterexample Guided Inductive Synthesis



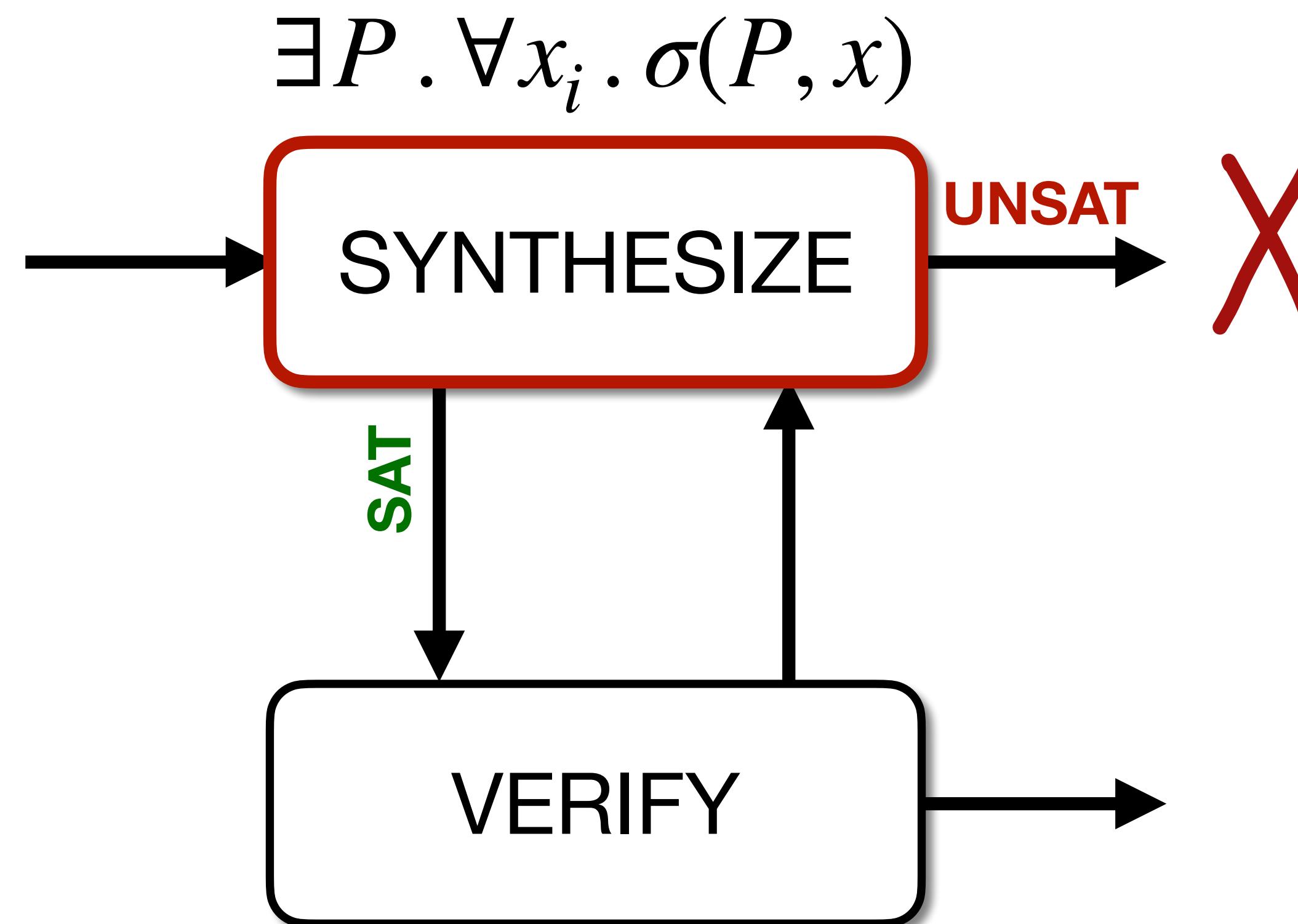
Algorithms for formal synthesis

Counterexample Guided Inductive Synthesis



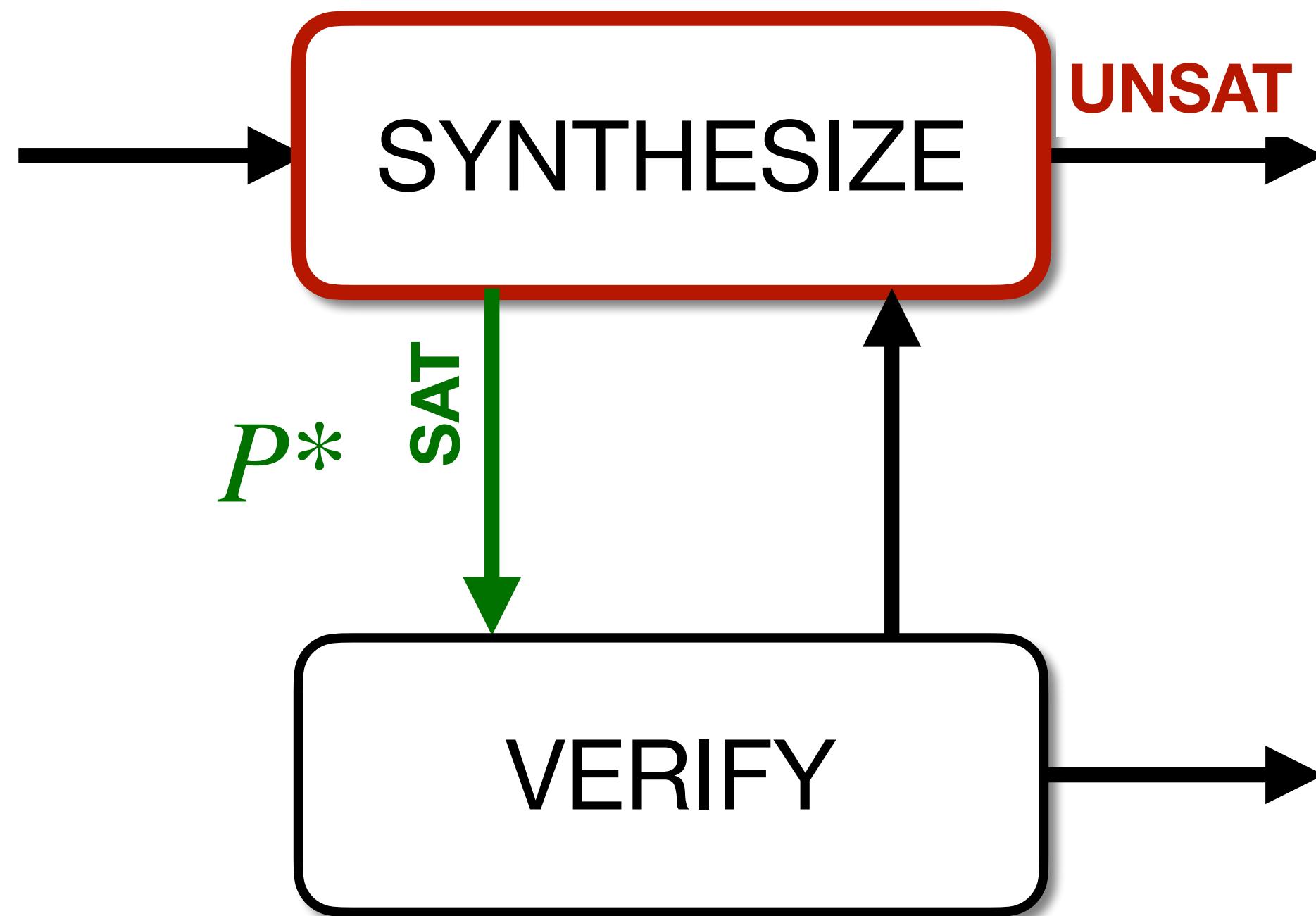
Algorithms for formal synthesis

Counterexample Guided Inductive Synthesis



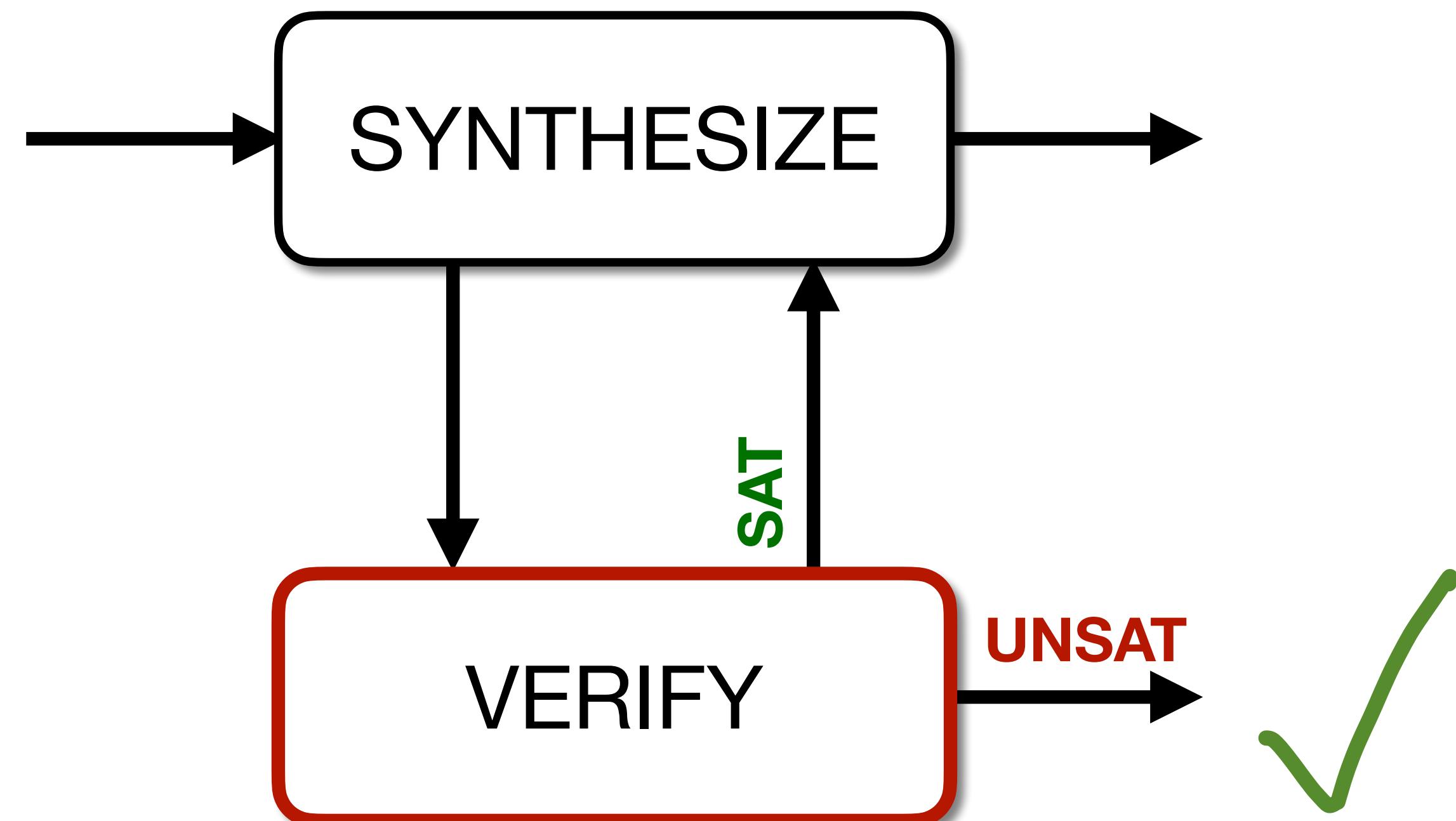
Algorithms for formal synthesis

Counterexample Guided Inductive Synthesis



Algorithms for formal synthesis

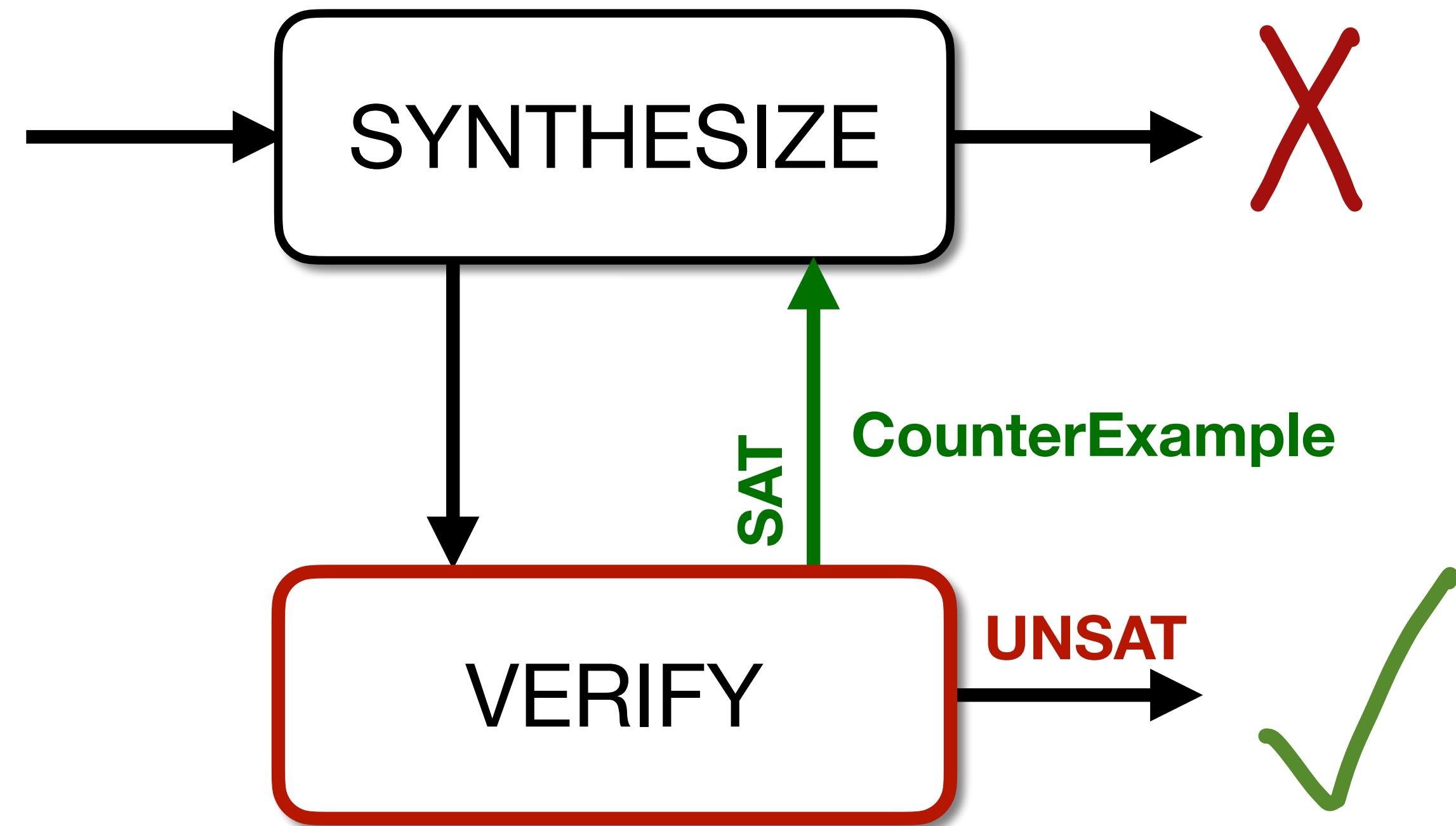
Counterexample Guided Inductive Synthesis



$$\exists x . \neg \sigma(P^*, x)$$

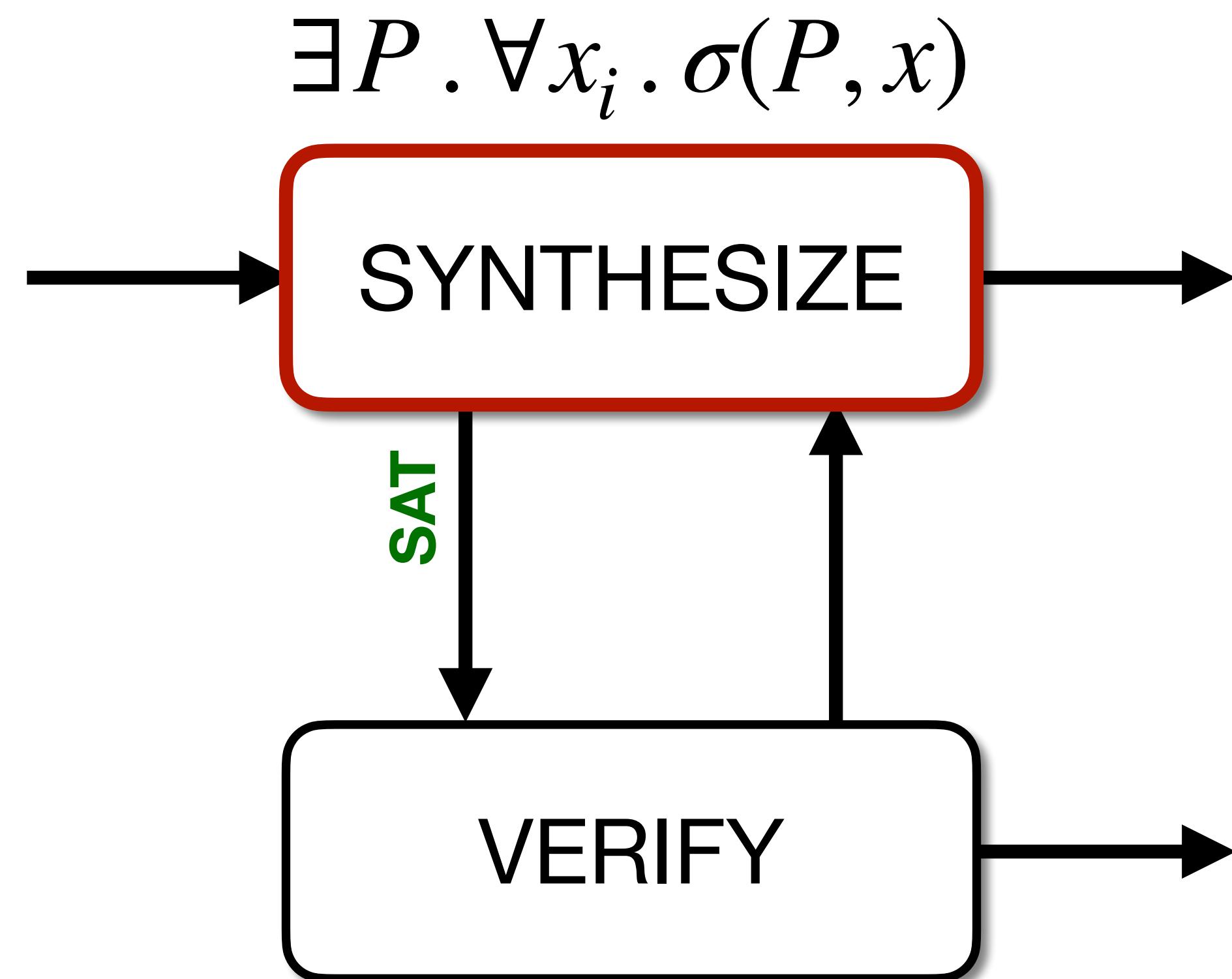
Algorithms for formal synthesis

Counterexample Guided Inductive Synthesis



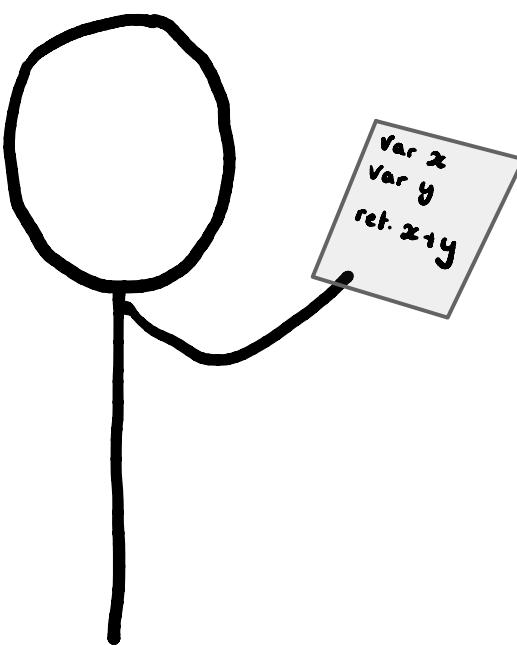
Algorithms for formal synthesis

Counterexample Guided Inductive Synthesis



Max

$$\exists f. \forall x, y. f(x, y) \geq y \wedge f(x, y) \geq x \wedge (f(x, y) = x \vee f(x, y) = y)$$


 $\exists P. \forall x_i. \sigma(P, x)$

Guess	Counterexamples
x	

$$\exists x. \neg \sigma(P^*, x)$$

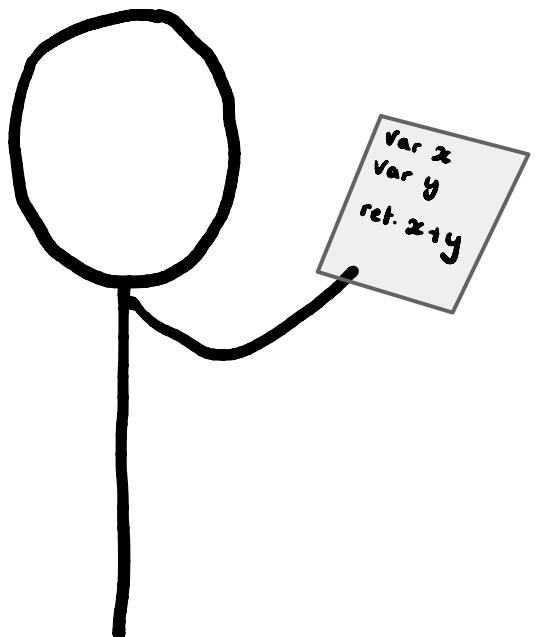


Max

$$\exists f. \forall x, y. f(x, y) \geq y \wedge f(x, y) \geq x \wedge (f(x, y) = x \vee f(x, y) = y)$$

Guess	Counterexamples
x	$x=1, y=2$

$$\exists P. \forall x_i. \sigma(P, x)$$



$$\exists x. \neg \sigma(P^*, x)$$

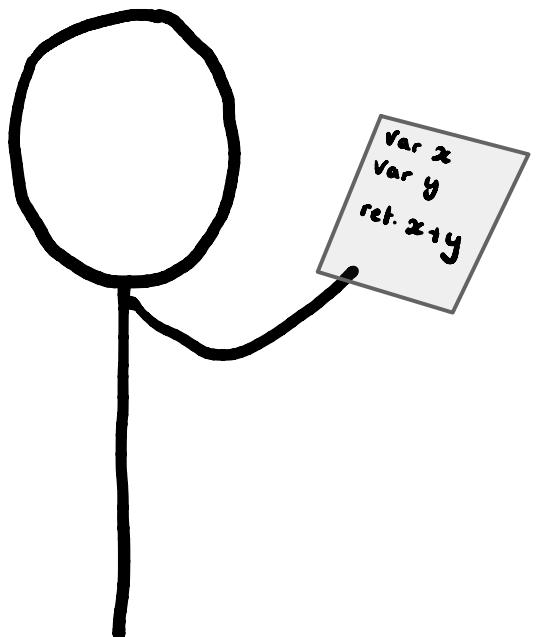


Max

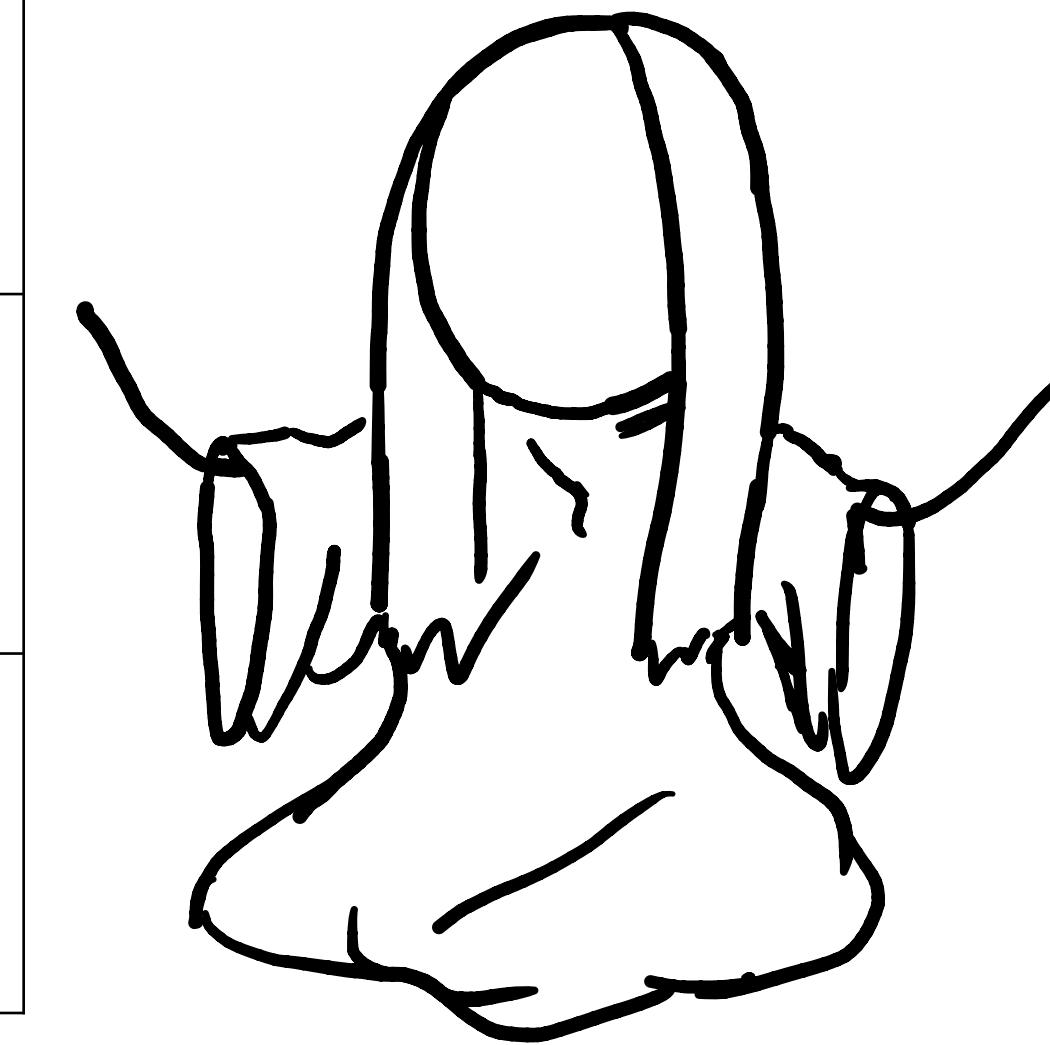
$$\exists f. \forall x, y. f(x, y) \geq y \wedge f(x, y) \geq x \wedge (f(x, y) = x \vee f(x, y) = y)$$

Guess	Counterexamples
x	$x=1, y=2$
y	

$$\exists P. \forall x_i. \sigma(P, x)$$



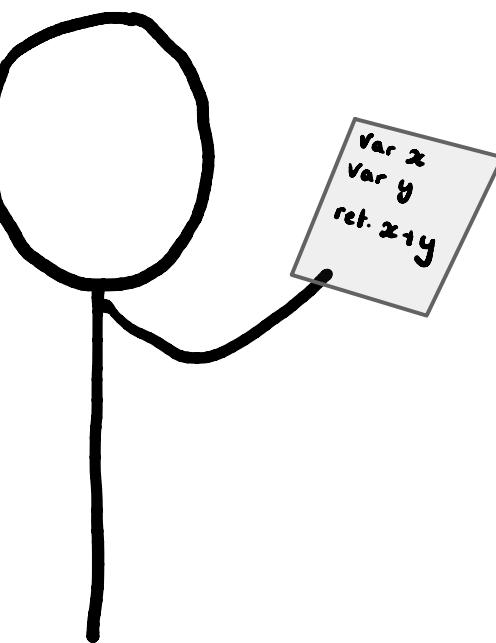
$$\exists x. \neg \sigma(P^*, x)$$



Max

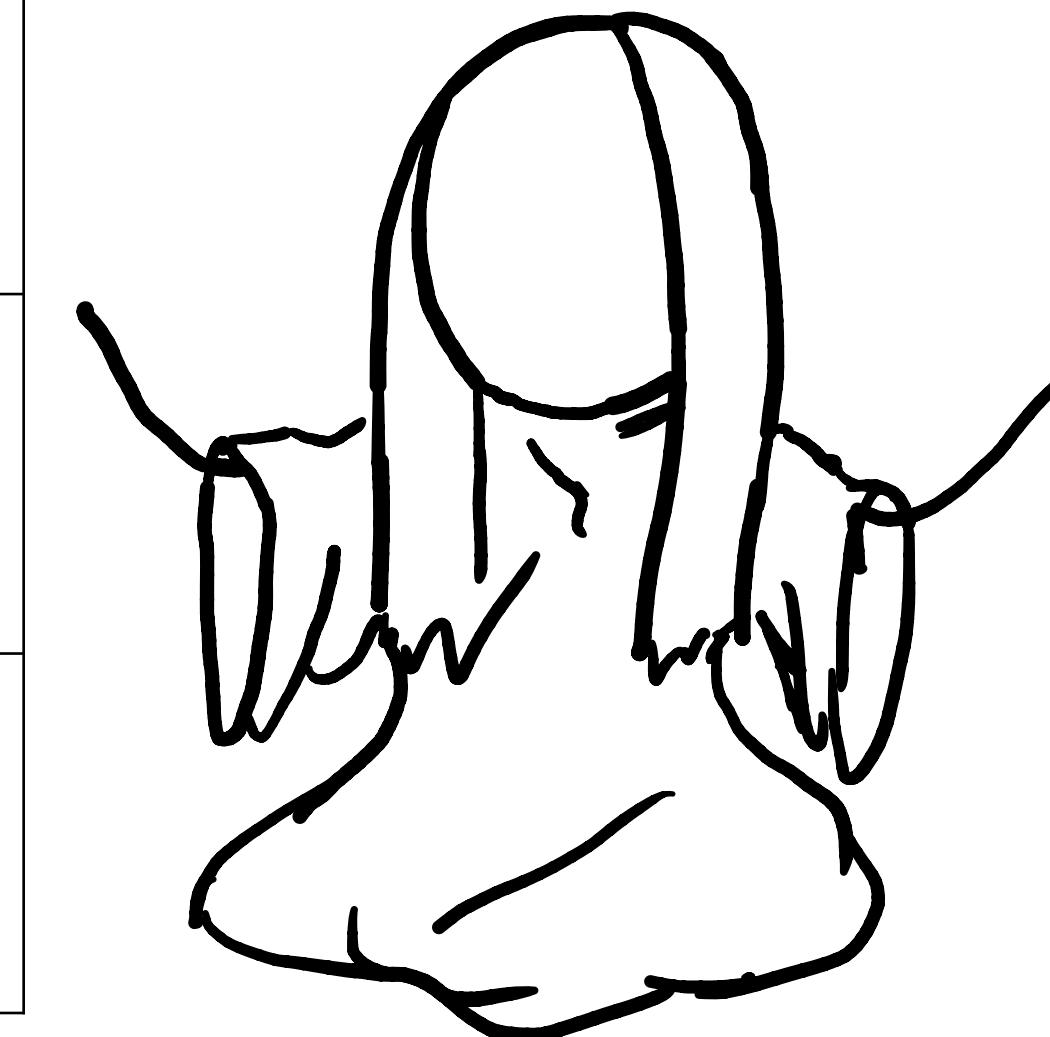
$$\exists f. \forall x, y. f(x, y) \geq y \wedge f(x, y) \geq x \wedge (f(x, y) = x \vee f(x, y) = y)$$

$\exists P. \forall x_i. \sigma(P, x)$



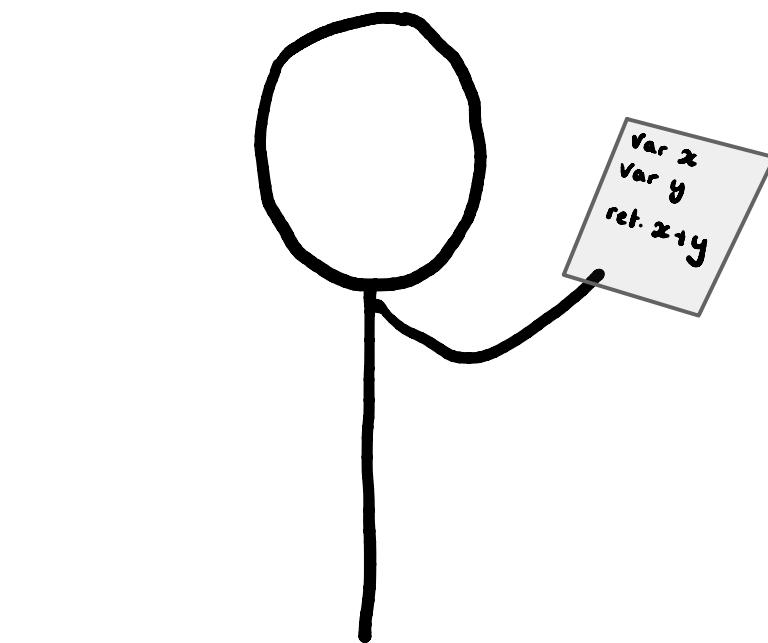
Guess	Counterexamples
x	$x=1, y=2$
y	$x=2, y=1$

$$\exists x. \neg \sigma(P^*, x)$$



Max

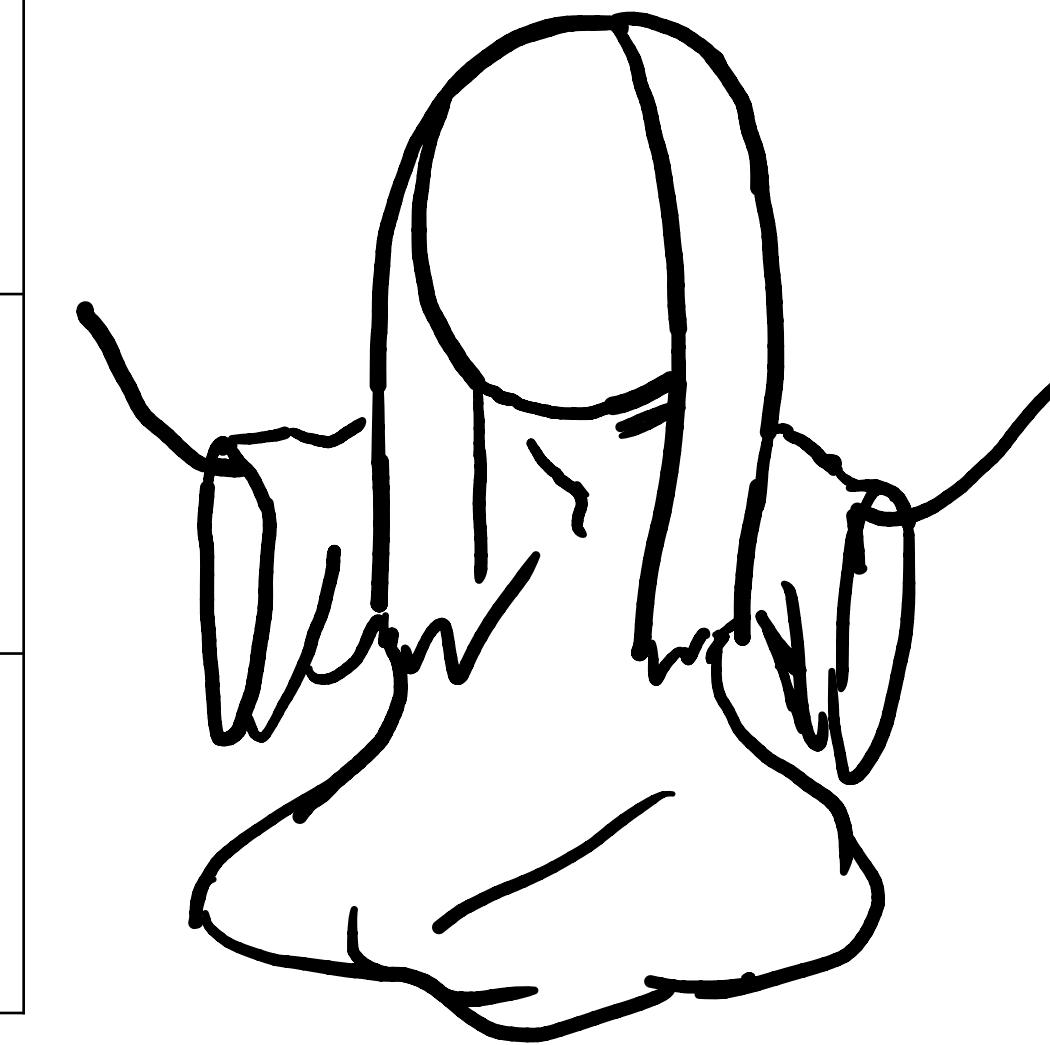
$$\exists f. \forall x, y. f(x, y) \geq y \wedge f(x, y) \geq x \wedge (f(x, y) = x \vee f(x, y) = y)$$



$$\exists P. \forall x_i. \sigma(P, x)$$

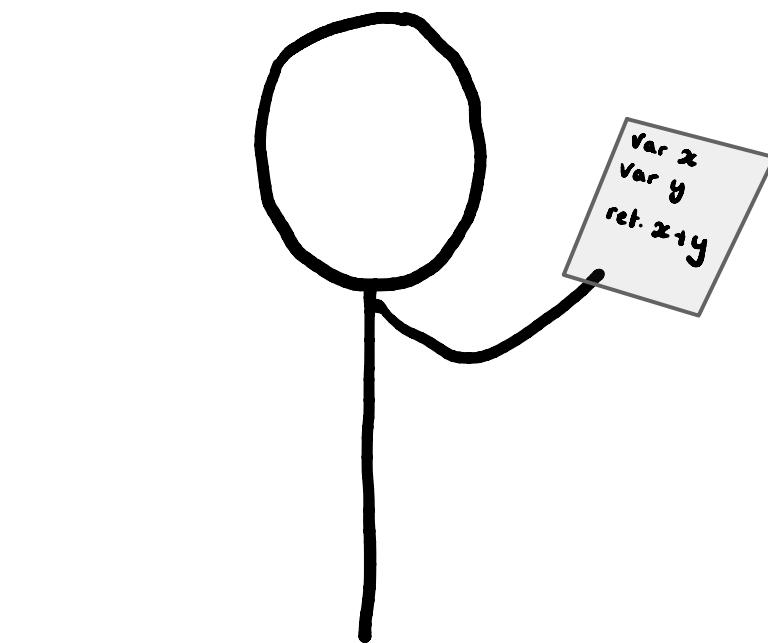
Guess	Counterexamples
x	$x=1, y=2$
y	$x=2, y=1$
$(x \geq y)?x : y$	

$$\exists x. \neg \sigma(P^*, x)$$



Max

$$\exists f. \forall x, y. f(x, y) \geq y \wedge f(x, y) \geq x \wedge (f(x, y) = x \vee f(x, y) = y)$$



$$\exists P. \forall x_i. \sigma(P, x)$$

Guess	Counterexamples
x	$x=1, y=2$
y	$x=2, y=1$
$(x \geq y)?x : y$	Correct!

$$\exists x. \neg \sigma(P^*, x)$$



The Checker

$$\exists x . \neg \sigma(P^*, x)$$

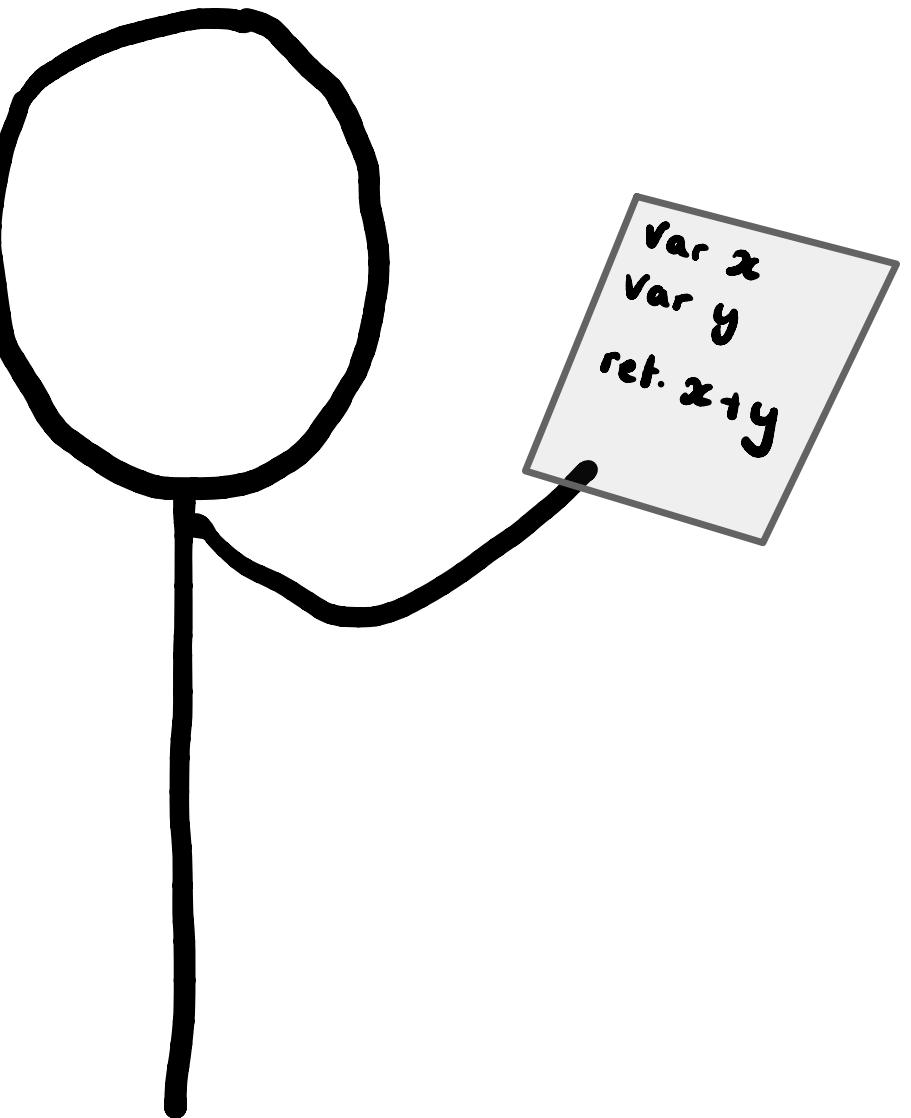

The Checker

- Checking against input-output specifications is easy: just execute the code
- Checking against arbitrary formula is harder:
 - Recall Inf2D: DPLL
 - We can extend DPLL to handle theories (SMT solvers)

$$\exists x . \neg \sigma(P^*, x)$$

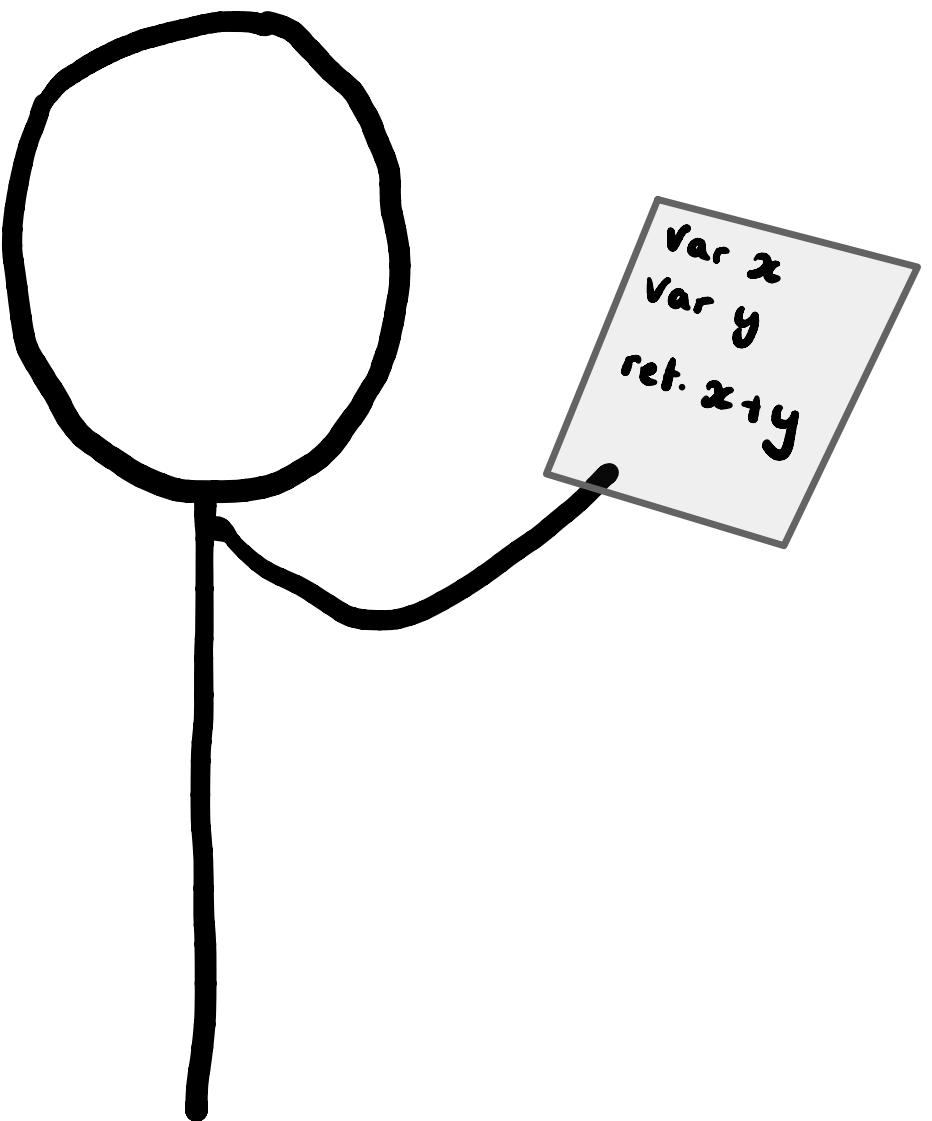


The Guesser

$$\exists P . \forall x_i . \sigma(P, x)$$


- Enumerate through the grammar (with heuristics)
- Symbolic encoding

The Guesser

$$\exists P . \forall x_i . \sigma(P, x)$$


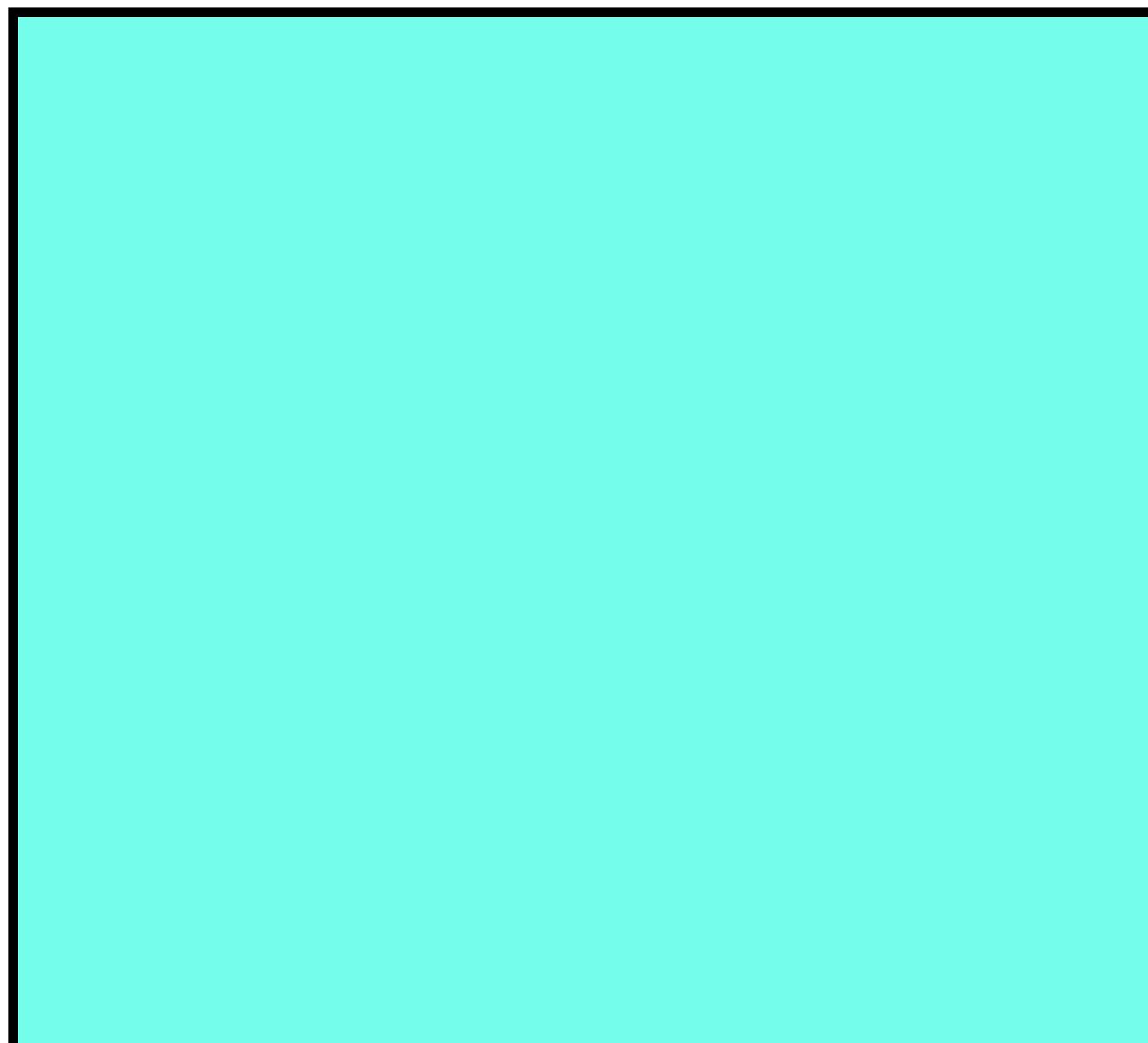
- Enumerate through the grammar (with heuristics)

```
A -> A+A | -A | x | y | 0 | 1 | ite (B, A, A)  
B -> B&B | ¬B | A=A | A≥A | ⊥
```

The Guesser

```
A->A+A|-A|x|y|0|1|ite(B,A,A)  
B->B&B|¬B|A=A|A≥A|⊥
```

Programs so far



Programs of length 1:

X Y 0 1 ⊥

The Guesser

```
A->A+A|-A|x|y|0|1|ite(B,A,A)  
B->B&B|¬B|A=A|A≥A|⊥
```

Programs so far

```
X Y 0 1 ⊥
```

Programs of length 2:

X+X Y+Y X+0 X+1 Y+0 Y+1 X+Y

-X -Y -0 -1

ite(⊥, X, X) ite(⊥, X, Y) ite(⊥, X, 0). ...

X=X X=Y Y=Y Y=0 Y=1 X=1 X=0

...

The Guesser

```
A->A+A|-A|x|y|0|1|ite(B,A,A)  
B->B&B|¬B|A=A|A≥A|⊥
```

Programs so far

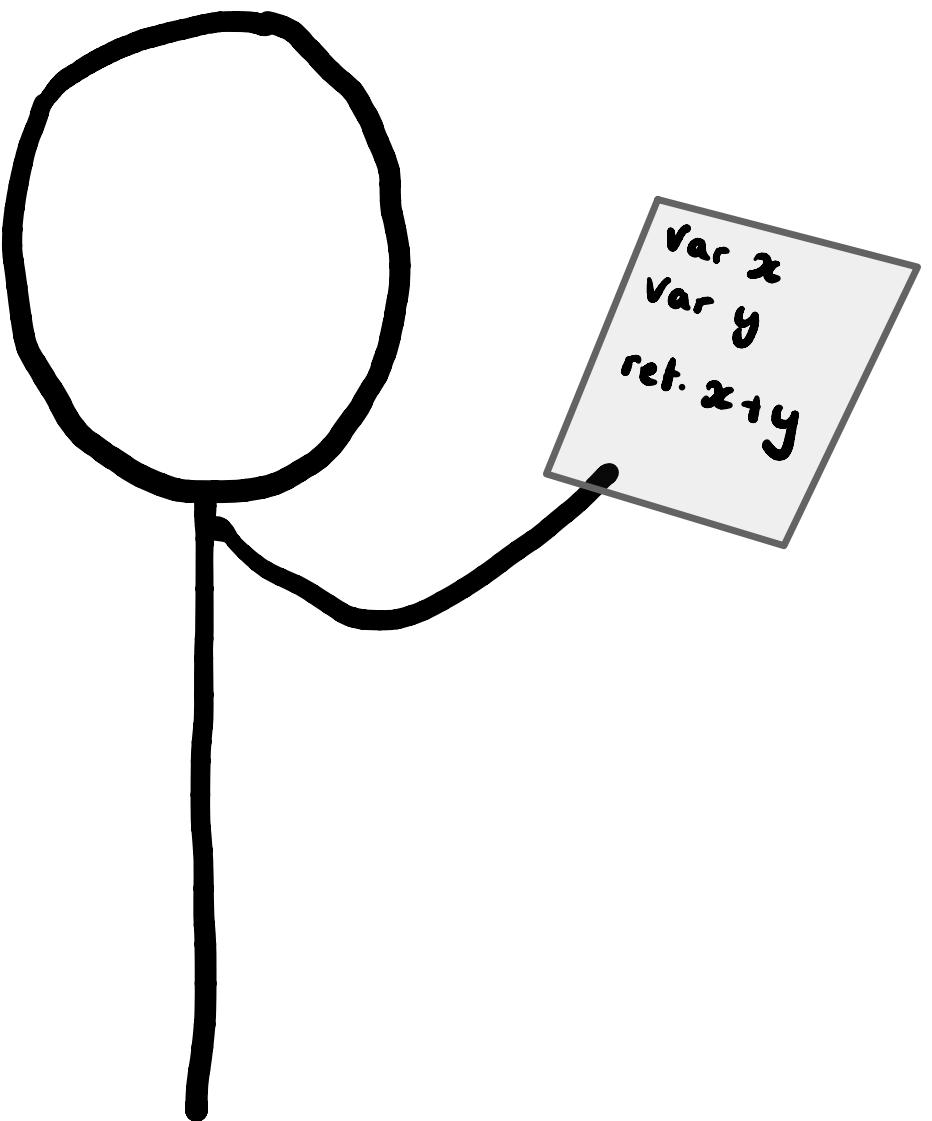
X	Y	0	1	⊥
X+X	Y+Y	X+0	X+1	X=X
Y+0	Y+1	X+Y		X=Y
-X	-Y	-0	-1	Y=Y
ite(⊥, X, X)	ite(⊥, X, Y)	ite(⊥, X, 0).		Y=0 Y=1 X=1 X=0
...				

Programs of length 3:

...

The Guesser

$$\exists P . \forall x_i . \sigma(P, x)$$

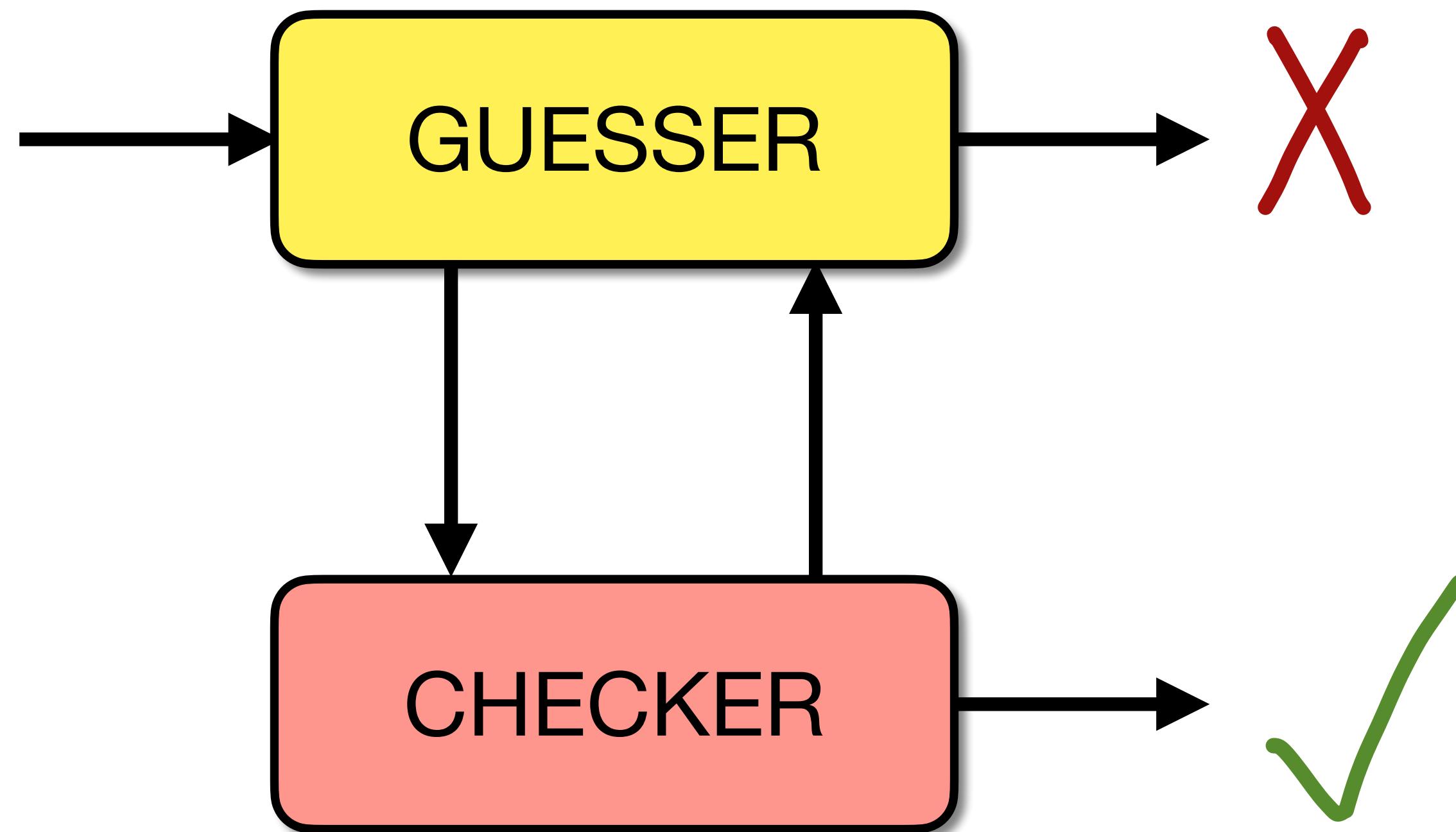


- Enumerate through the grammar (with heuristics)
- Symbolic encoding:
 - use selector variables to choose which parts of the grammar to use.
 - Ask an SMT solver to assign values to those selector variables.

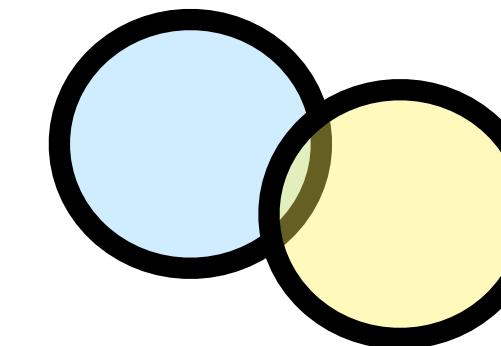
A->A+A | -A | x | y | 0 | 1 | ite (B, A, A)

B->B&B | \neg B | A=A | A>A | ⊥

New synthesis algorithms:



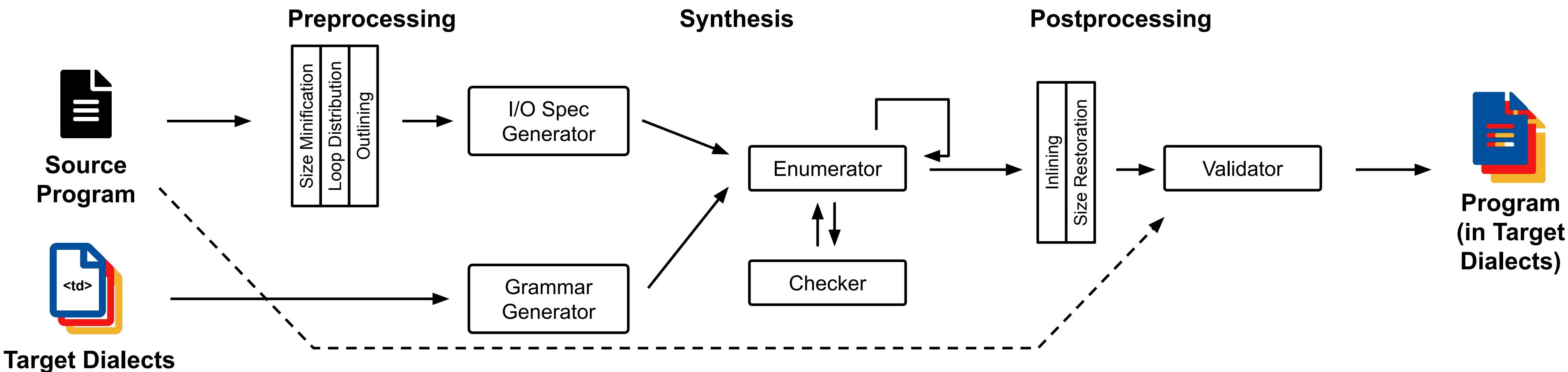
- New guessers: clever heuristics, use machine learning
 - Type directed
 - Reinforcement learning
 - Probabilistic grammars
- New checkers: pass more information back to the guesser
 - “Can this partial program ever satisfy the specification”
- New ways to break the problem into smaller problems



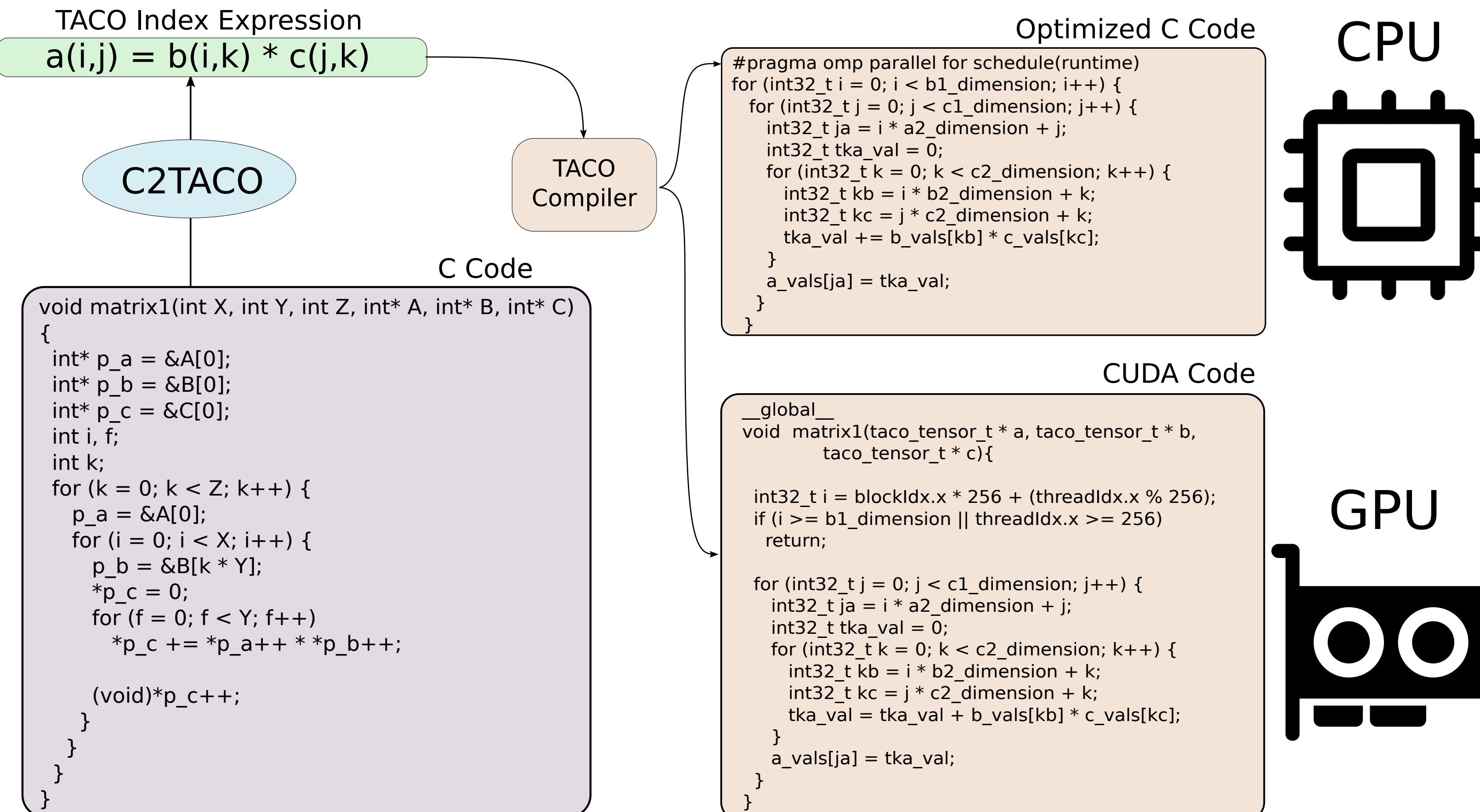
Applications

Lifting Code

- Translating code from one language to another
- Without manually writing a syntactic transpiler
- Improved performance



Lifting Code



C2TACO: Lifting Tensor Code to TACO - José Wesley de Souza Magalhães,
Jackson Woodruff, Elizabeth Polgreen, Michael O'Boyle

Synthesizing Invariants

```
int a=0;  
  
int b=1;  
  
while (!button_pressed)  
{  
    a' = b;  
    b' = a+b;  
    a = a'; b=b'  
}  
assert(a <= b);
```

How can we prove this property?

Synthesizing Invariants

```
int a=0;  
int b=1;  
while (!button_pressed)  
{  
    if (a+b<MAX_INT)  
    {  
        a' = b;  
        b' = a+b;  
        a = a'; b=b'  
    }  
}  
assert(a <= b);
```

How can we prove this property?

Synthesizing Invariants

```
int a=0;  
int b=1;  
while (!button_pressed)  
{  
    if (a+b<MAX_INT)  
    {  
        a' = b;  
        b' = a+b;  
        a = a'; b=b'  
    }  
}  
assert(a <= b);
```

How can we prove this property?

$$\exists \text{inv} \forall a, b . \text{init}(a, b) \implies \text{inv}(a, b) \wedge \\ \text{inv}(a, b) \wedge \text{trans}(a, b, a', b') \implies \text{inv}(a', b') \wedge \\ \text{inv}(a, b) \implies a \leq b$$

Synthesizing Invariants

```
int a=0;  
int b=1;  
  
while (!button_pressed)  
{  
    if (a+b<MAX_INT)  
    {  
        a' = b;  
        b' = a+b;  
        a = a'; b=b'  
    }  
}  
assert(a <= b);
```

How can we prove this property?

$$\exists \text{inv} \forall a, b . \text{init}(a, b) \Rightarrow \text{inv}(a, b) \wedge \text{inv}(a, b) \wedge \text{trans}(a, b, a', b') \Rightarrow \text{inv}(a', b') \wedge \text{inv}(a, b) \Rightarrow a \leq b$$

Synthesizing Invariants

```
int a=0;  
int b=1;  
  
while (!button_pressed)  
{  
    if (a+b<MAX_INT)  
    {  
        a' = b;  
        b' = a+b;  
        a = a'; b=b'  
    }  
}  
assert(a <= b);
```

How can we prove this property?

$$\exists \text{inv} \forall a, b. \boxed{a = 0 \wedge b = 1} \Rightarrow \text{inv}(a, b) \wedge$$
$$\text{inv}(a, b) \wedge \boxed{a' = b \wedge b' = a + b} \Rightarrow \text{inv}(a', b') \wedge$$
$$\text{inv}(a, b) \Rightarrow \boxed{a \leq b}$$

This looks a lot like a program synthesis problem

Data wrangling (FlashFill)

The screenshot shows an Excel spreadsheet titled "Raster - Excel". The data is organized into two columns: "Name" and "First" (with "Last" as a header). The first two rows are labeled "Name" and "First", while the third row contains "Ned Lanning" and "Ned". The fourth row contains "Margo Hendrix" and "Margo". A green box highlights the "Ned" and "Margo" entries. A preview window on the right shows the results of the FlashFill operation: "First" and "Last" for the first two rows, followed by "Dianne" under "First" and "Pugh" under "Last" for the third row, and so on. The "Data" tab is selected in the ribbon.

	Name	First	Last
1	Ned Lanning	Ned	
2	Margo Hendrix	Margo	
3	Dianne Pugh	Dianne	Pugh
4	Earlene McCarty	Earlene	
5	Jon Voigt	Jon	
6	Mia Arnold	Mia	
7	Jorge Fellows	Jorge	
8	Rose Winters	Rose	
9	Carmela Hahn	Carmela	
10	Denis Horning	Denis	
11	Johnathan Swope	Johnathan	
12	Delia Cochran	Delia	
13	Marguerite Cervantes	Marguerit	
14	Liliana English	Liliana	
15	Wendy Stephenson	Wendy	

Excel sees
patterns
and
shows a
preview

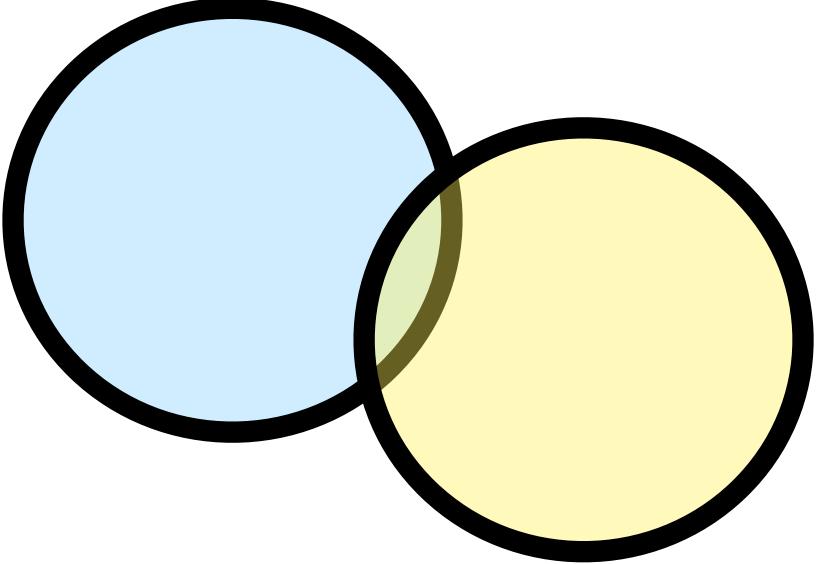
Programming by example:

$$\exists P . P(\text{Ned Lanning}) = \text{Ned} \wedge$$

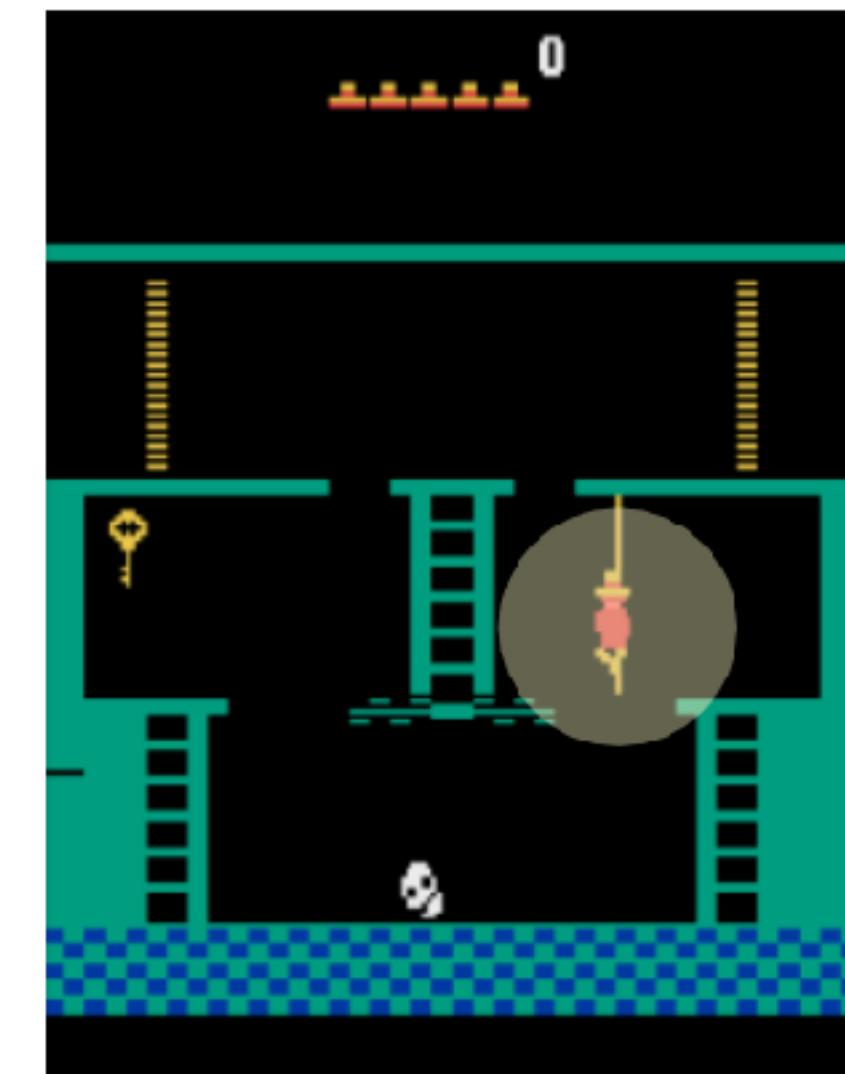
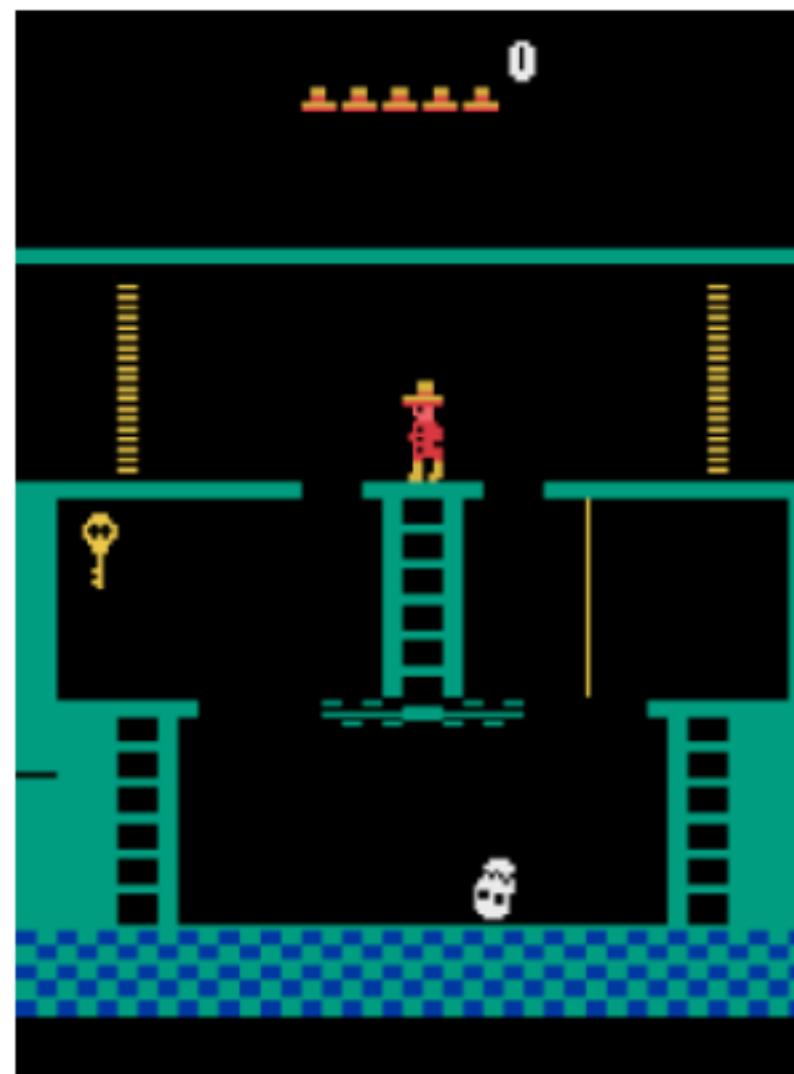
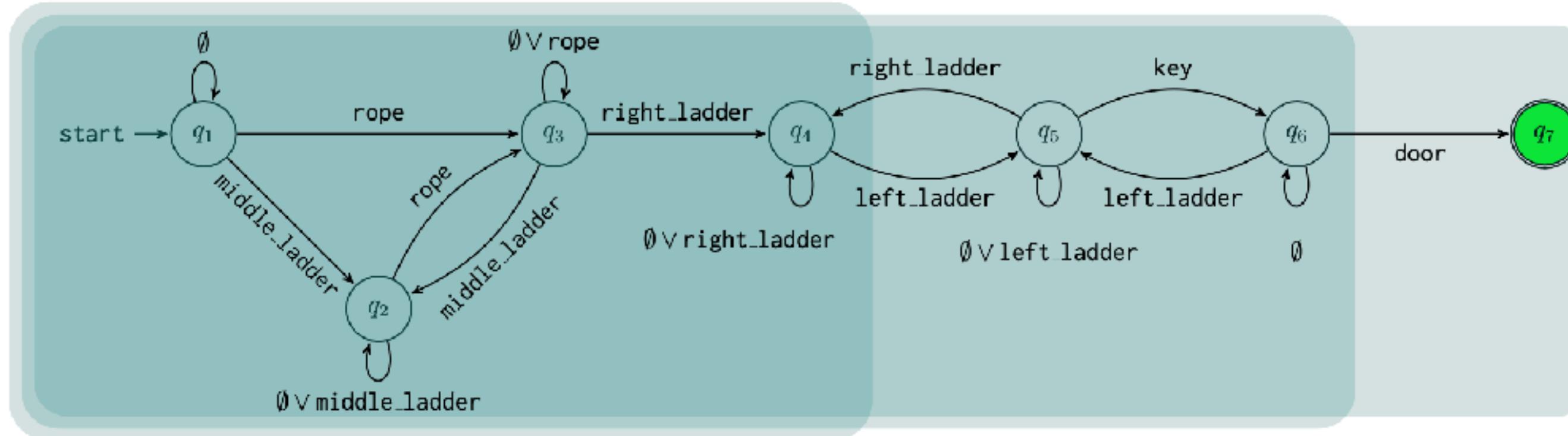
$$P(\text{Margo Hendrix}) = \text{Margo} \wedge$$

...

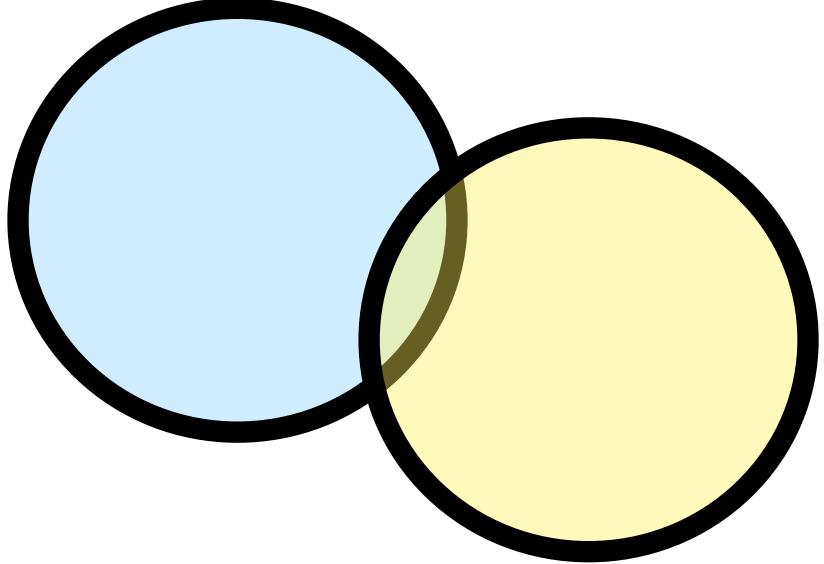
Helping machine learning plan



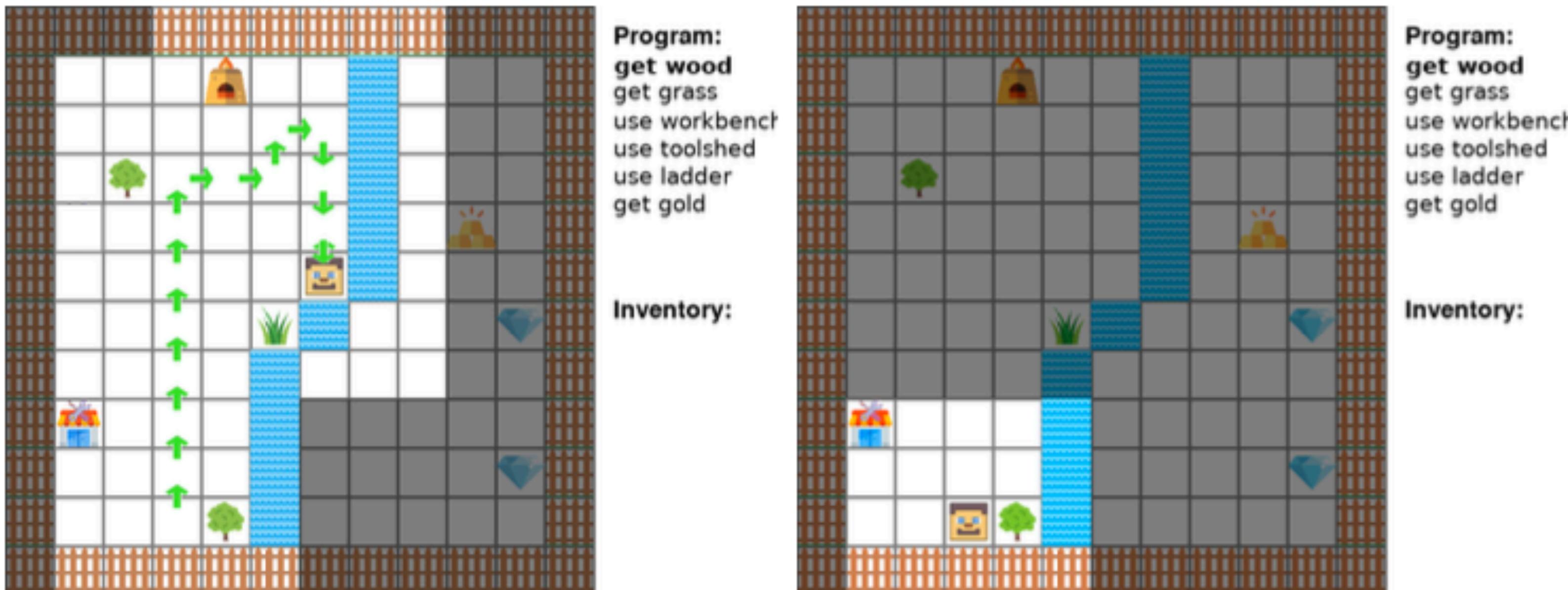
- Use synthesis to learn the order tasks should be done in. Then give SafeRL the ordering



Helping machine learning plan



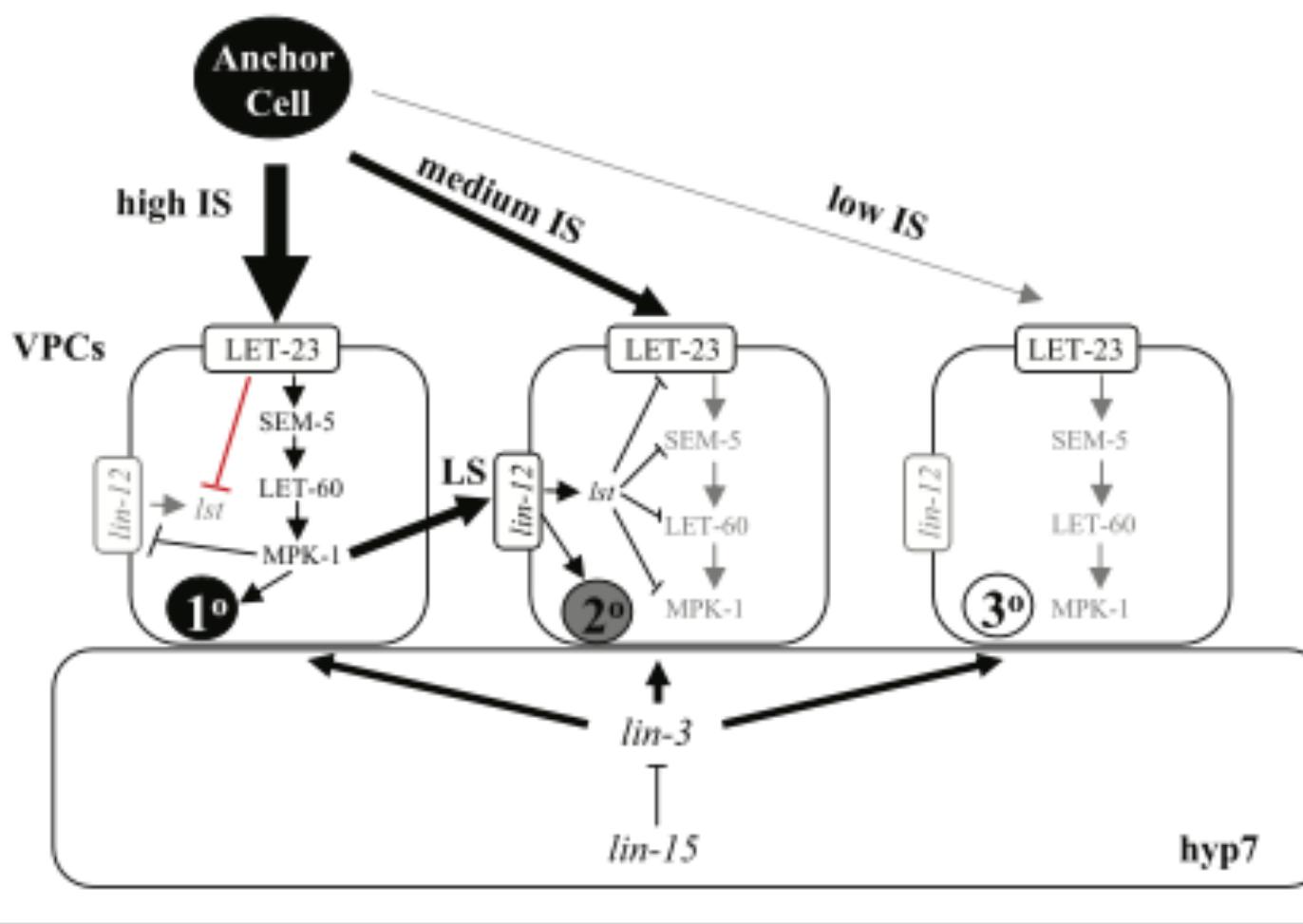
- Model Predictive Program Synthesis: Predict possible worlds based on observations, synthesise a program maximising probability of success, execute program, repeat.



Learning from Sparse Data

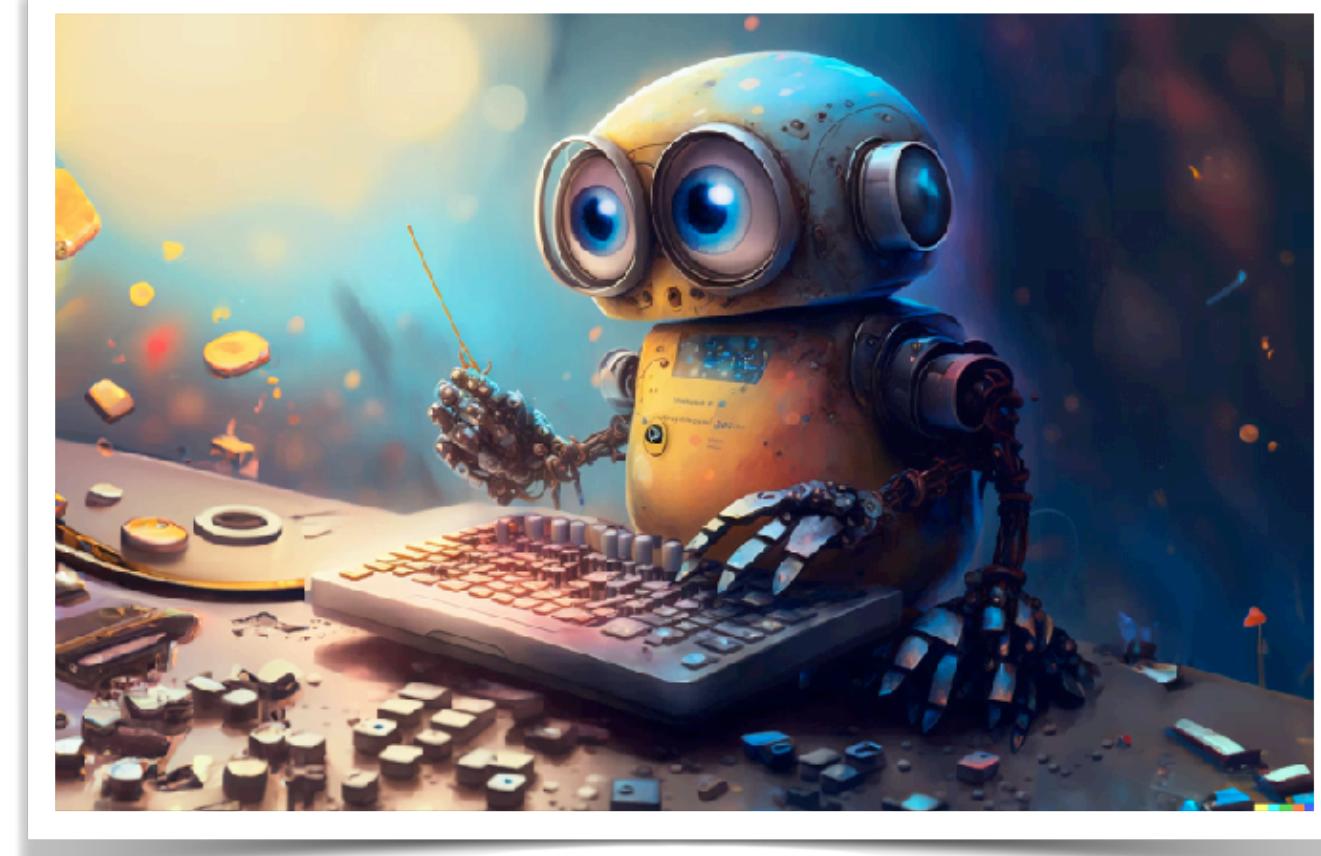
- Program synthesis can learn from less data than classic ML methods
- Resulting program is “human readable”

Cell fate decision in C. Elegans: synthesise the model that matches all observations



Resources

elizabeth.polgreen@ed.ac.uk



- TedX talk by Armando Solar Lezama: <https://youtu.be/RzMk7XWVMWw?si=VnHT-2GR7gs2Zd>
- CAV keynote by Sumit Gulwani: <https://youtu.be/PeeFc9Js9ZU?si=GRXpKbQKXUgAFdwx>
- Course notes from MIT: <https://people.csail.mit.edu/asolar/SynthesisCourse/TOC.htm>
- Fourth year projects? Talk to me
- These slides on <https://polgreen.github.io/pdfs/synth.pdf>