PROGRAM ANALYSIS FUELED SEARCH

Dr. Katie Stolee
Assistant Professor
North Carolina State University

NC STATE UNIVERSITY

85% search for code at least weekly [TOSEM 2014]

DEVELOPERS SEARCH THE WEB FOR CODE ....

Average of 12 queries per day [FSE 2015]

...BUT WEB SEARCH WAS NOT BUILT FOR CODE.

Code searches require more effort. [MSR 2018]

CODE SEARCH IS DIFFERENT.

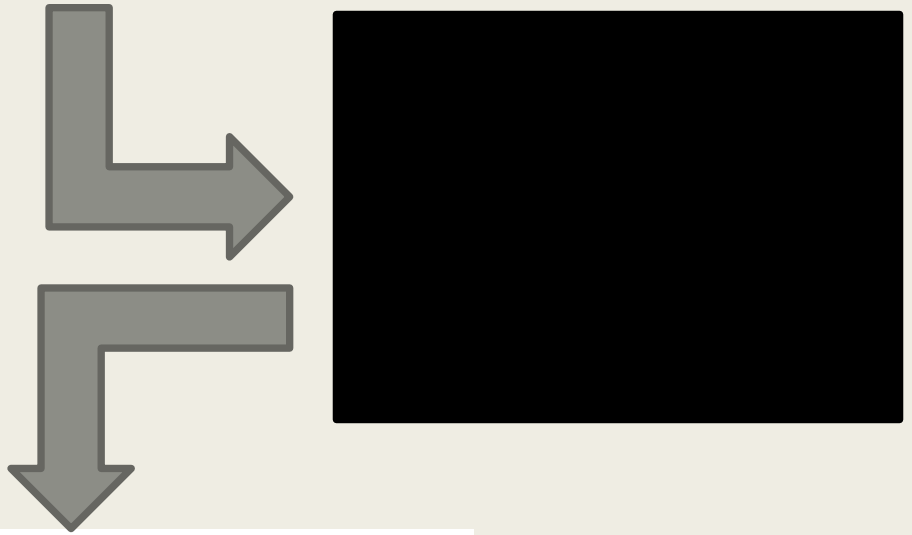
A different kind of search

```
blackbox([6, 2, 3, 4])
```

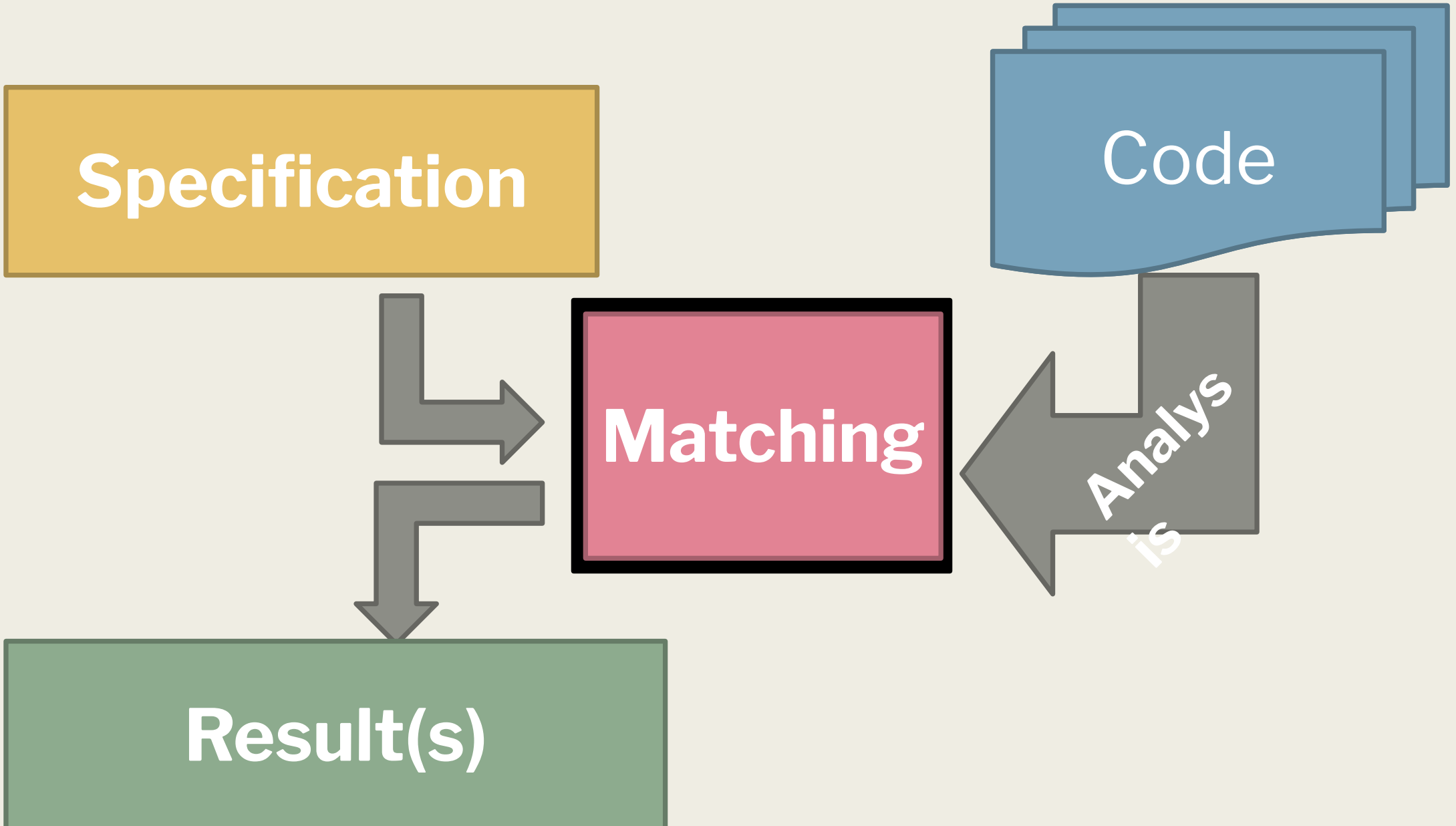
```
2
```

Example input

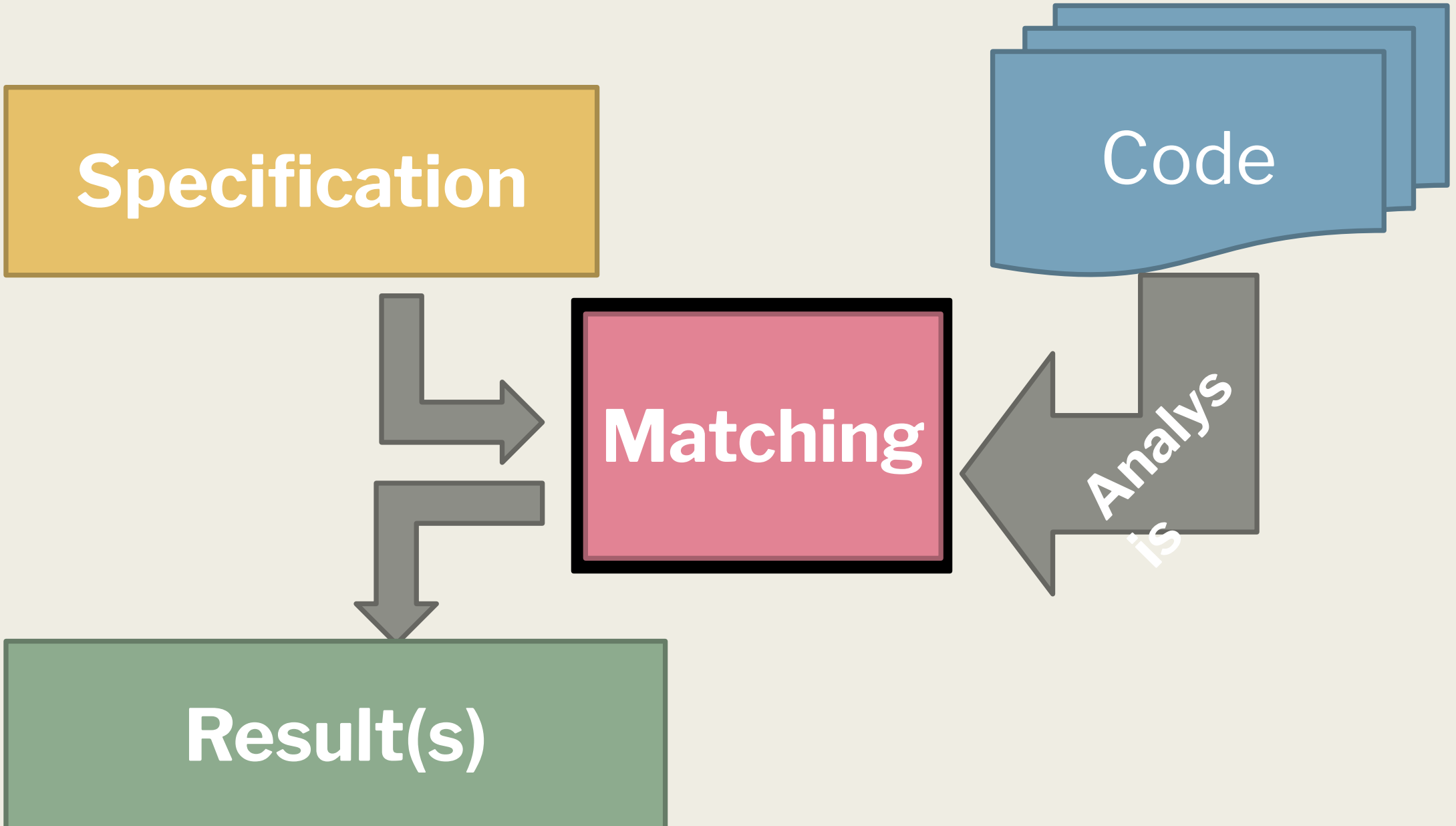
Example output

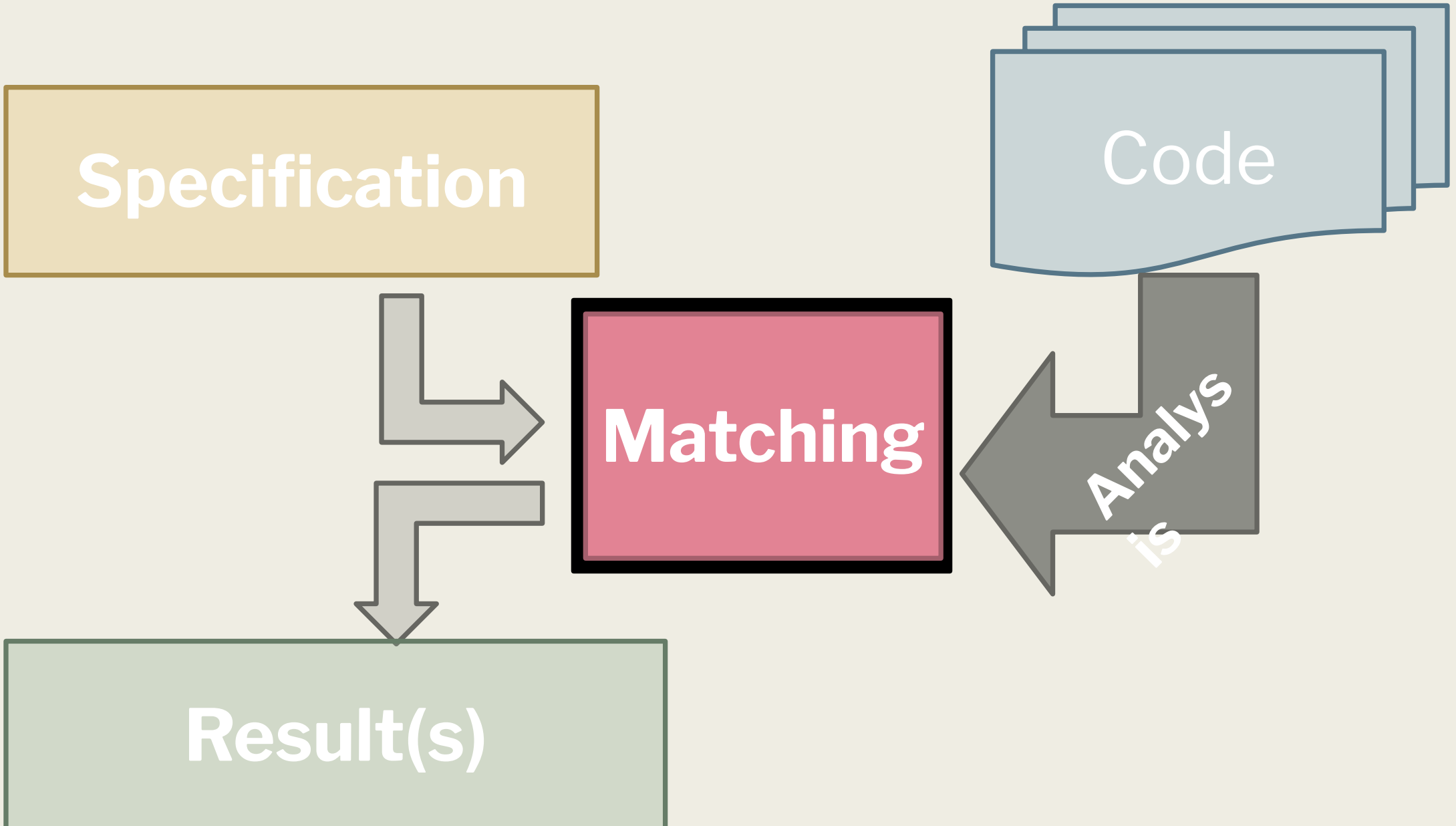


```
def min_py2(a):
    return sorted(a)[0]
```

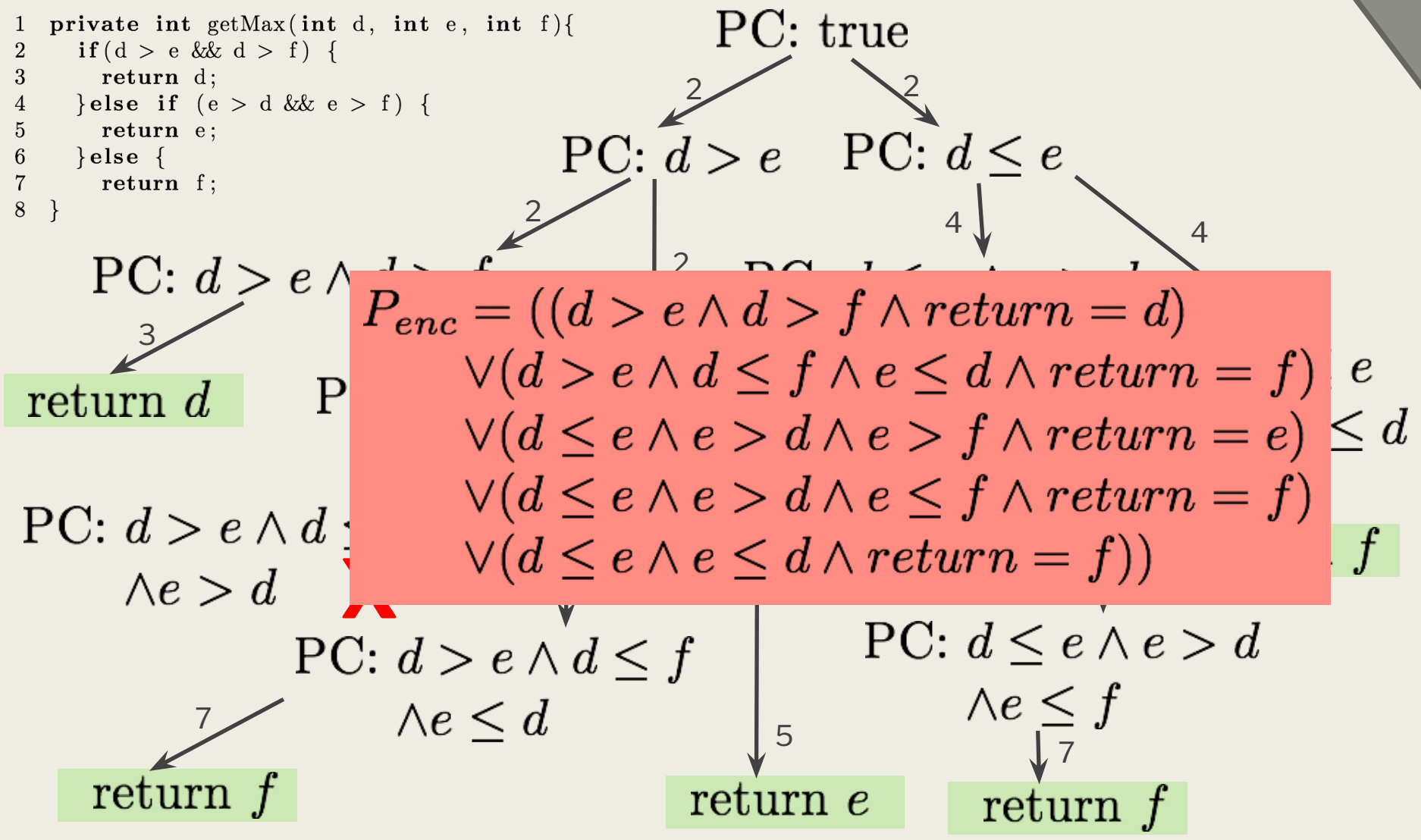

Specification
Code
Matching
Analysis
Result(s)

FUELED BY: STATIC ANALYSIS


Specification
Code
Matching
Analysis
Result(s)


Specification
Code
Matching
Analysis
Result(s)

Symbolic Execution



SMT Solvers

Matching

Satisfiability **M**odulo **T**heory solvers determine if a logical formula is satisfiable

Facts	Assertions
`a >= 0`	`(assert (>= a 0))`
`b = 2`	`(assert (= b 2))`
`c = 2`	`(assert (= c 2))`
`c = a * b`	`(assert (= ( * a b) c))`

Result: **sat** $a \mapsto 1$

SMT Solvers

Matching

Satisfiability **M**odulo **T**heory solvers determine if a logical formula is satisfiable

Facts	Assertions
`a >= 0`	`(assert (>= a 0))`
`b = ?`	**`(assert (= b ?))`**
`c = 2`	`(assert (= c 2))`
`c = a * b`	`(assert (= ( * a b) c))`

Result: **sat**

$$a \mapsto 2 \wedge b \mapsto 2$$

SMT Solvers

Matching

Satisfiability **M**odulo **T**heory solvers determine if a logical formula is satisfiable

Facts	Assertions
a = 0	**(assert (= a 0))**
b = ?	(assert (= b ?))
c = 2	(assert (= c 2))
c = a * b	(assert (= (* a b) c))

Result: **unsat**

SMT Matching

Encoding

Matching

```
(declare-fun a () Int)
(declare-fun b () Int)
(declare-fun c () Int)
(declare-fun return () Int)
(assert (= return (+ (+ a b) c)))
(assert (and (= a 3) (= b 0)
    (= c 0)))
(assert (= return 3))
```

```
private int getsum(int a, int b, int c){
  return a + b + c;
}
```

Potential Search Result

Query

Input	Output	Result
3, 4, 3	4	unsat
3, 0, 0	3	

Not a Result!

Matching

Matching

```
1  private int getMax(int d, int e, int f){
2    if(d > e && d > f) {
3      return d;
4    }else if (e > d && e > f) {
5      return e;
6    }else {
7      return f;
8  }
```

PC1: $d > e \wedge d > f \wedge return = d$
PC2: $d > e \wedge d \leq f \wedge e \leq d \wedge return = f$
PC3: $d \leq e \wedge e > d \wedge e > f \wedge return = e$
PC4: $d \leq e \wedge e > d \wedge e \leq f \wedge return = f$
PC5: $d \leq e \wedge e \leq d \wedge return = f$

Query

Input	Output	Result
3, 4, 3	4	sa
3, 0, 0	3	sa
		t

This is a result!

SMT-based search

Search

Get email alias in Java

Satsy

Input: sporty@spice.com

Output: sporty

Satsy often returns more relevant search results than Google!

[Stolee, Elbaum, and Dobos. “Solving the Search for Source Code.” TOSEM 2014.]
[Stolee, Elbaum and Dwyer. “Code Search with input/output queries: Generalizing, ranking, and assessment.”

```
blackbox
2
```

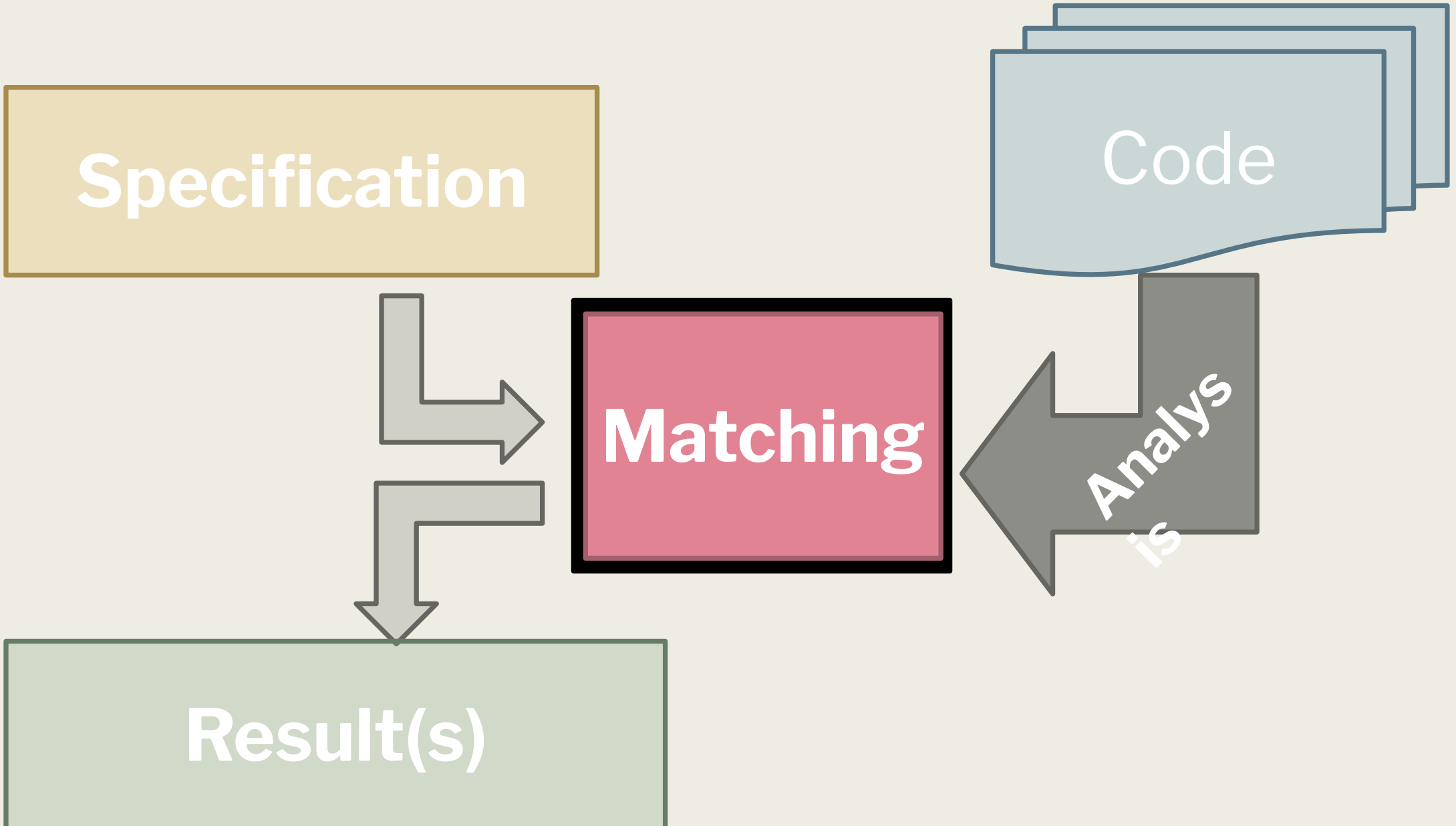
```
] a) {

ength; i++) {
i] < m)
```

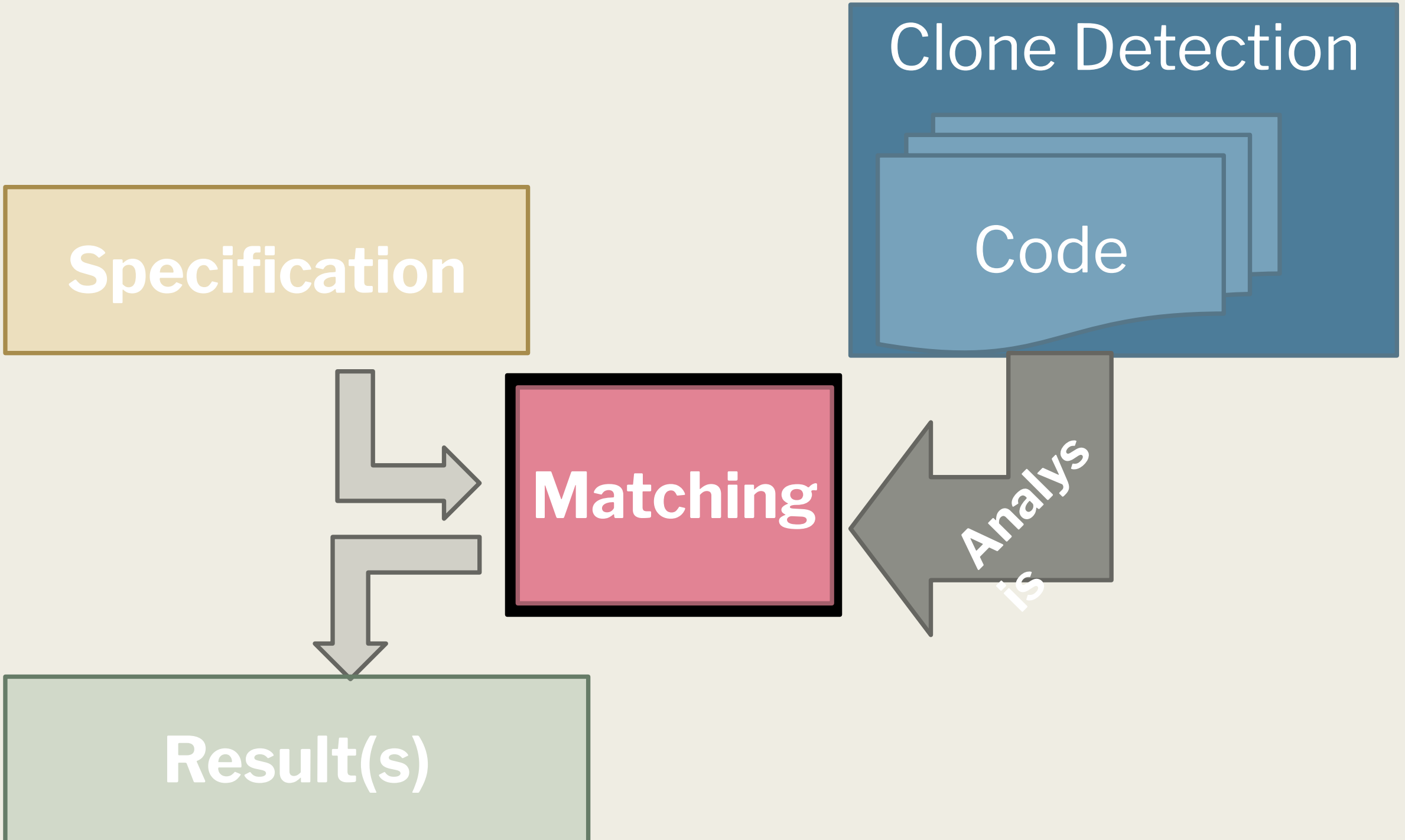
This gets expensive!

```
def min_py1(ar
    least = arr[
    for a in arr[1:]:
      if a < least:
        least = a
    return least
```

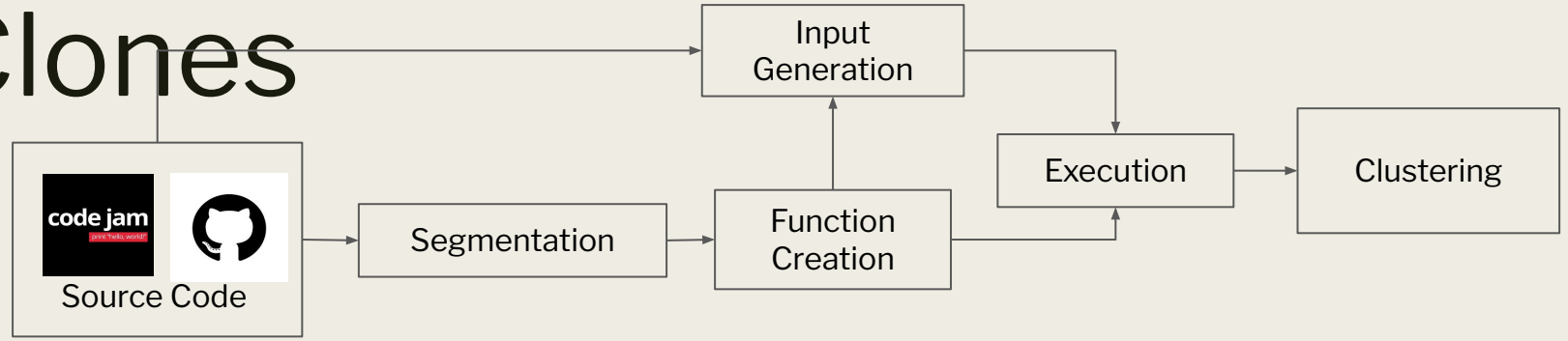
```
def min_py2(a):
    return sorted(a)[0]
```

FUELED BY: DYNAMIC ANALYSIS


Specification
Code
Matching
Analysis
Result(s)


Clone Detection
Code
Specification
Matching
Analysis
Result(s)

Cross-Language Behavioral Clones

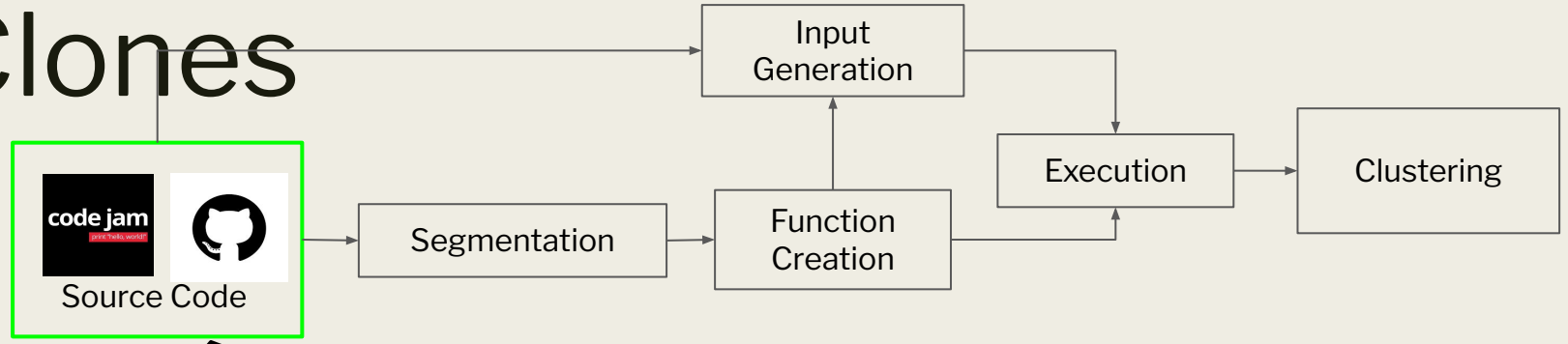


This Is Georg e.

Find him at ICSE 2020!

[Mathew, Parnin and Stolee. “SLACC: Simion-based Language Agnostic Code Clones.” ICSE 2020.]

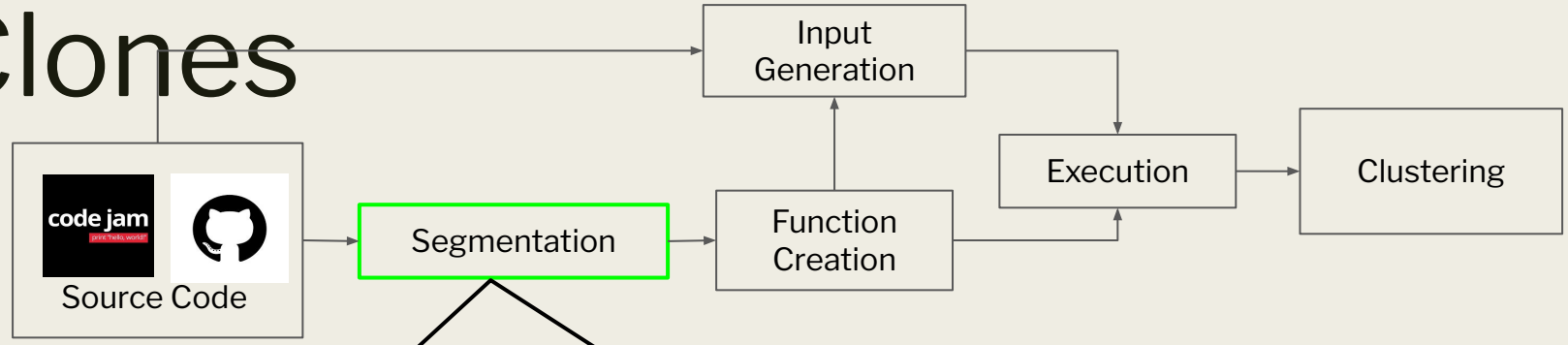
Cross-Language Behavioral Clones




```
public String interleave(int[] a, int[] b) {
  String result = "";
  int i = 0;
  for( i = 0; i < a.length && i < b.length; i++ ) {
    result += a[i];
    result += b[i];
  }
  int[] remaining = a.length < b.length ? b : a;
  for( int j = i; j < remaining.length; j++ ) {
    result += remaining[j];
  }
  return result;
}
```

[Mathew, Parnin and Stolee. “SLACC: Simion-based Language Agnostic Code Clones.” ICSE 2020.]

Cross-Language Behavioral Clones

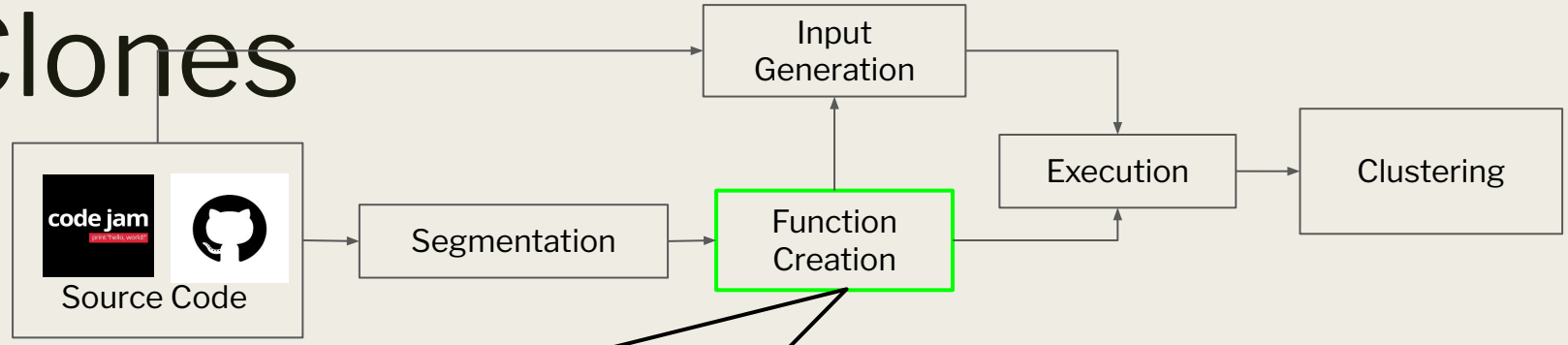



```
String result = "";
int i = 0;
for( i = 0; i < a.length && i < b.length; i++ ) {
  result += a[i];
  result += b[i];
}
```

```
int[] remaining = a.length < b.length ? b : a;
for( int j = i; j < remaining.length; j++ ) {
  result += remaining[j];
}
return result;
```

[Mathew, Parnin and Stolee. “SLACC: Simion-based Language Agnostic Code Clones.” ICSE 2020.]

Cross-Language Behavioral Clones

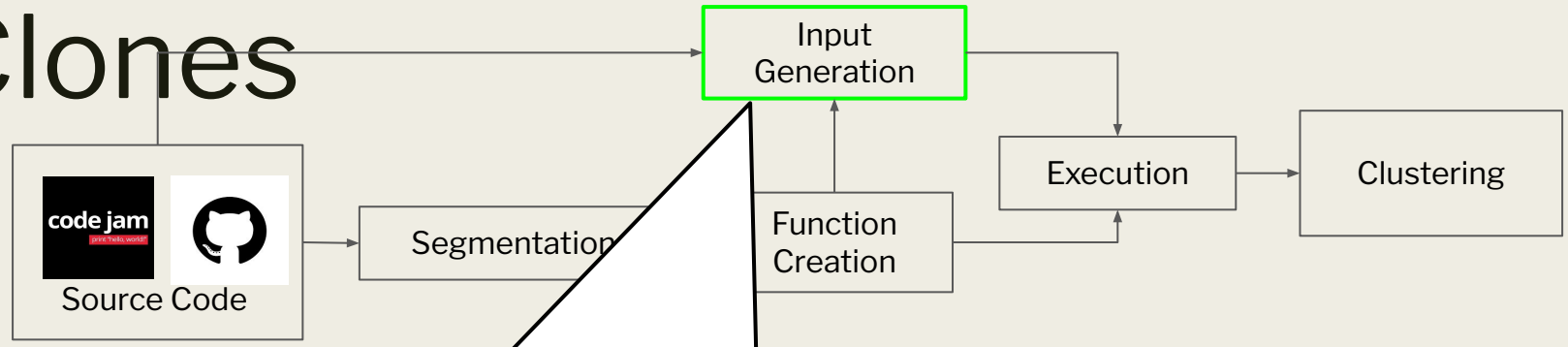



```
public String func_b15f(int[] a, int[] b) {
  String result = "";
  int i = 0;
  for( i = 0; i < a.length && i < b.length; i++ ) {
    result += a[i];
    result += b[i];
  }
  return result;
}
```

```
public String func_ea72(int[] a, int[] b, int i, String result) {
  int[] remaining = a.length < b.length ? b : a;
  for( int j = i; j < remaining.length; j++ ) {
    result += remaining[j];
  }
  return result;
}
```

[Mathew, Parnin and Stolee. “SLACC: Simion-based Language Agnostic Code Clones.” ICSE 2020.]

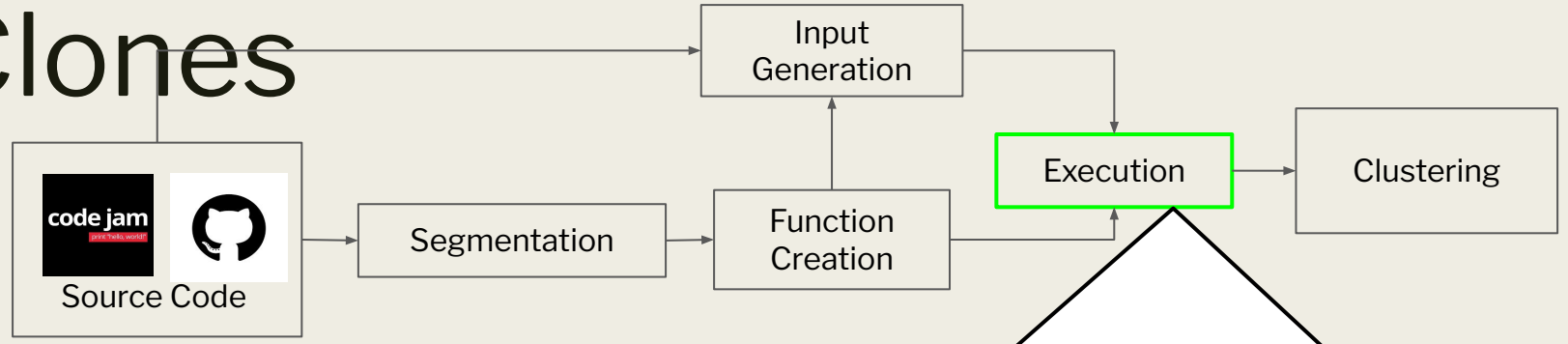
Cross-Language Behavioral Clones




```
int[], int[]
• [1,2,3], [45, 16]
• [ ], [3,2,1]
  ⋮
  256 times
  ⋮
• [4, 5, 6, 7, 8, 99], [ ]
int[], int[], int, String
• [1,2,3], [45, 16], 0, “Hello
  World”
  ⋮
  256 times
  ⋮
```

ed Language Agnostic Code Clones.”
ICSE 2020.]

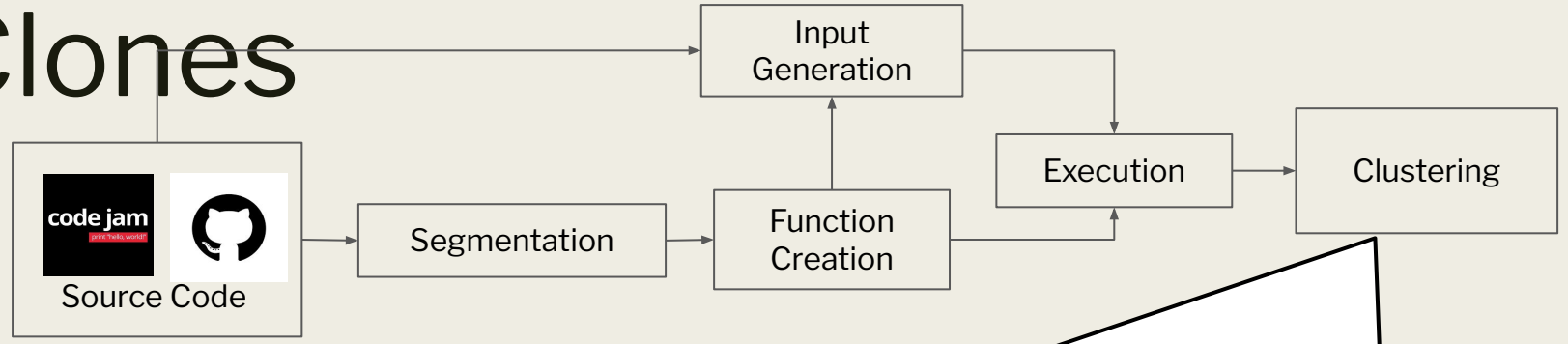
Cross-Language Behavioral Clones




```
func_b15f(int[] a, int[] b)
• func_b15f([1,2,3], [45, 16]) ⇒ 1452163
• func_b15f([ ], [3,2,1]) ⇒ 321
  ⋮
  256 times
  ⋮
• func_b15f([4, 5, 6, 7, 8, 99], [ ]) ⇒ 4567899
• func_b15f([3], [2]) ⇒ 32
func_ea72(int[] a, int[] b, int i, String result)
• func_ea72([1,2,3], [45, 16], 0, "Hello World") ⇒
  Hello World123
  ⋮
  256 times
```

Agnostic Code Clones."
ICSE 2020.]

Cross-Language Behavioral Clones




```java
public String func_b15f(int[] a, int[] b) {
  String result = "";
  int i = 0;
  for( i = 0; i < a.length && i < b.length; i++ ) {
    result += a[i];
    result += b[i];
  }
  return result;
}
```

```python
def func_9f34(l1, l2):
    result = ""
    for (e1, e2) in zip(l1, l2):
        result += str(e1)
        result += str(e2)
    return result
```

```python
def func_15e8(l1, l2):
    from itertools import chain
    return "".join([str(x)
            for x in chain.from_iterable(zip(l1, l2))])
```

[... Agnostic Code Clones." ICSE 2020.]

WHAT DOES THIS MEAN FOR SEARCH?

```
blackbox([6, 2, 3, 4])
```

```
2
```

1

```
public Integer minJ(int[] a) {
    Integer m = null;
    for (int i=0; i < a.length; i++) {
      if (m == null || a[i] < m)
        m = a[i];
    }
    return m;
}
```

2

```
def min_py2(a):
    return sorted(a)[0]
```

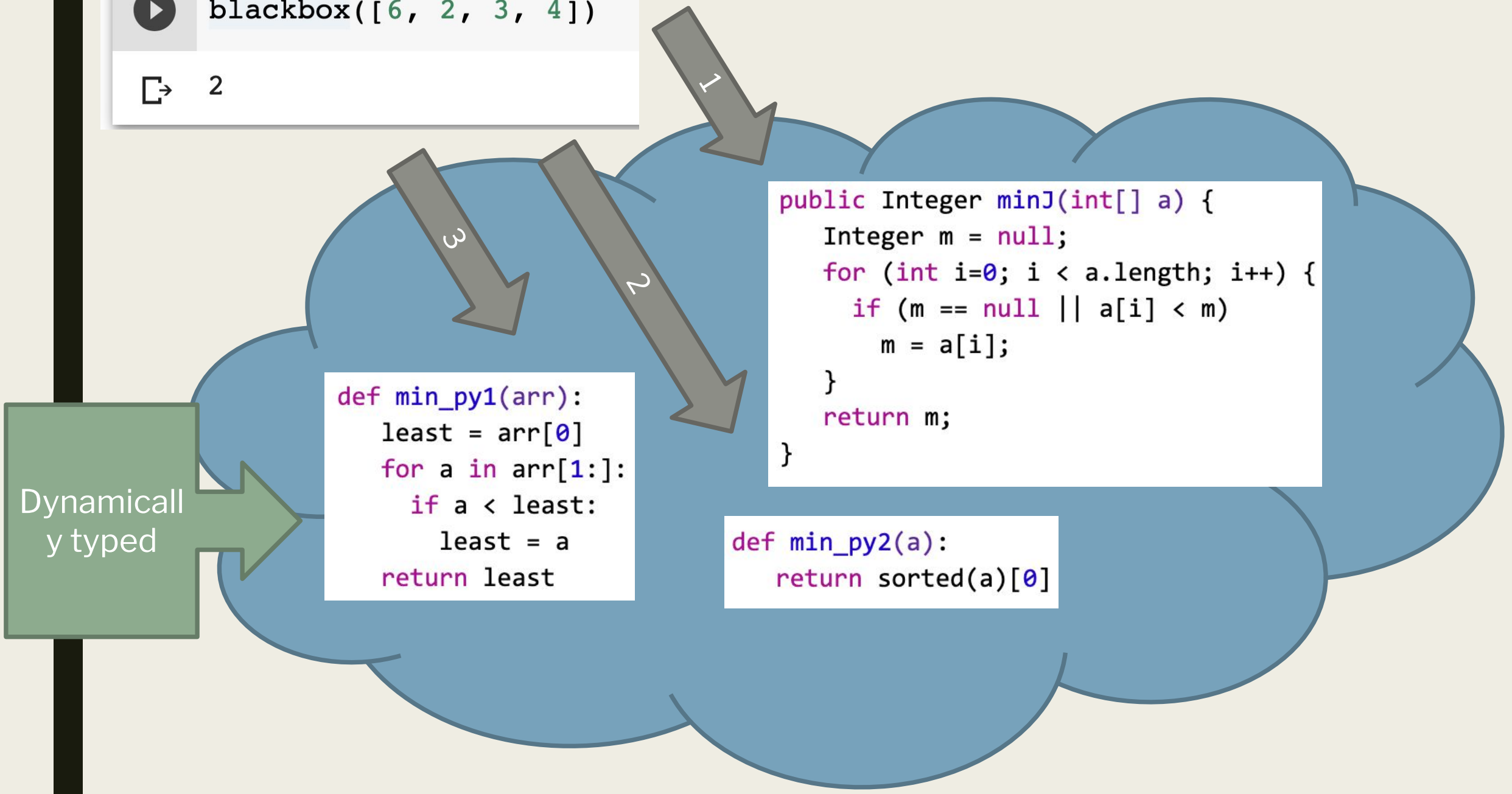
3

```
def min_py1(arr):
    least = arr[0]
    for a in arr[1:]:
      if a < least:
        least = a
    return least
```

[Mathew, Parnin and Stolee. “SLACC: Simion-based Language Agnostic Code Clones.” ICSE 2020.]

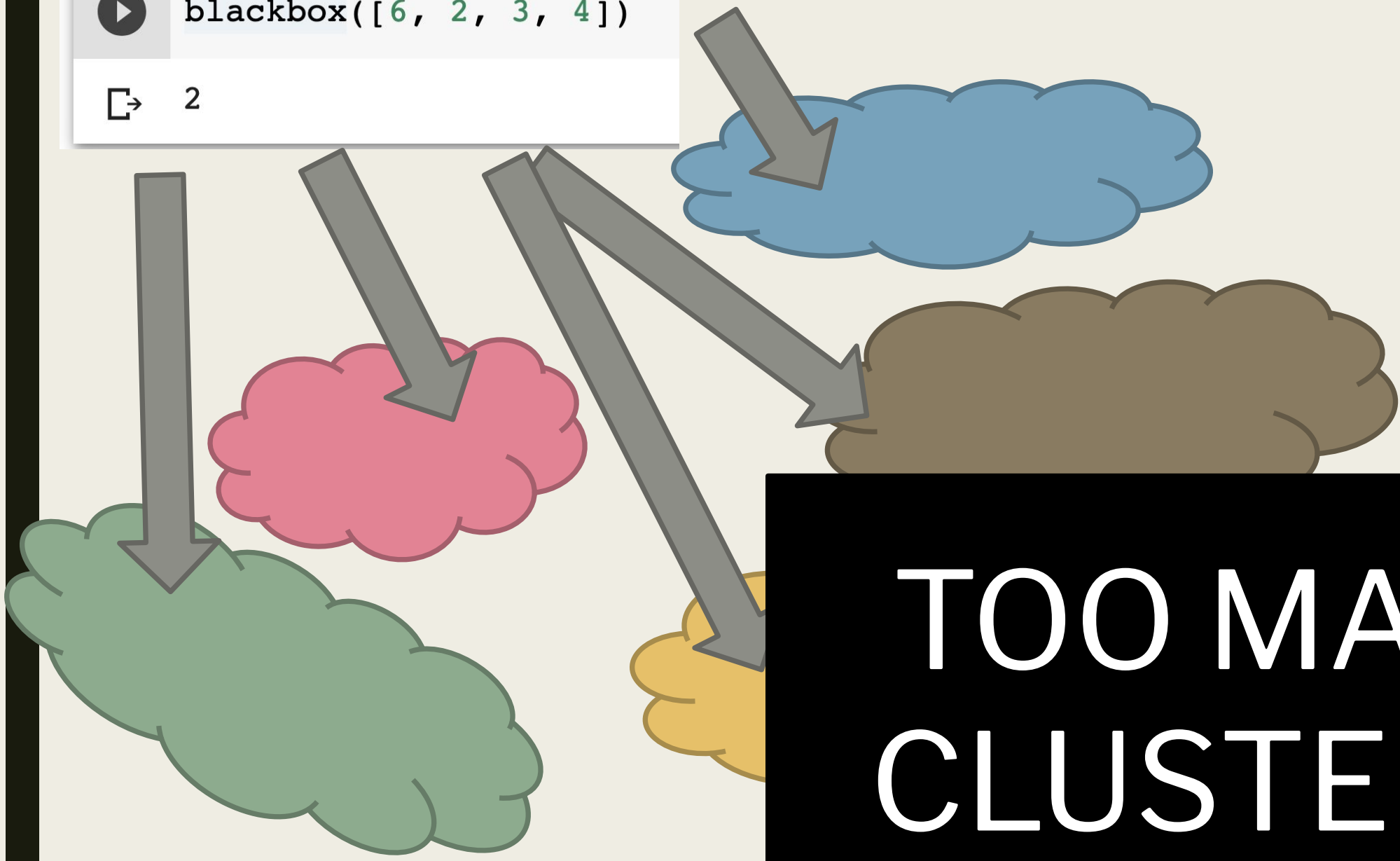
```
blackbox([6, 2, 3, 4])
```

```
2
```



```
blackbox([6, 2, 3, 4])
```

```
2
```



TOO MANY CLUSTERS?

	A	B
1	**Input**	**Output**
2	3	21
3	5	
4	6	
5	7	

=sum(A2:A5)
=sum(A2:A12)
=A2*A5
=3*A3+A4
=2*(A4+A5)-A3
=3*A4+2*A3-A5
=(A2*A3*A4)-(A3*13)-4
=if(A2>2,sum(A2:A5),0)
=A4*A5/2
=(A3*A4)-(A2+A4)
=(A2*A2*A2)-A4

This is the *solution space*

[Shriver, Elbaum and Stolee. "At the end of synthesis narrowing program candidates." ICSE-NIER 2017]

	A	B
1	**Input**	**Output**
2	3	21
3	5	
4	6	
5	7	

Example 1 : 11 programs

```
=sum(A2:A5)
=sum(A2:A12)
=A2*A5
=3*A3+A4
=2*(A4+A5)-A3
=3*A4+2*A3-A5
=(A2*A3*A4)-(A3*13)-4
=if(A2>2,sum(A2:A5),0)
=A4*A5/2
=(A3*A4)-(A2+A4)
=(A2*A2*A2)-A4
```

[Shriver, Elbaum and Stolee. “At the end of synthesis narrowing program candidates.” ICSE-NIER 2017]

	A	B
1	**Input**	**Output**
2	3	24
3	6	
4	7	
5	8	

Example 1 : 11 programs

Example 2

User Gives More Examples

=sum(A2:A5)

=sum(A2:A12)

=A2*A5

=3*A3+A4

=2*(A4+A5)-A3

=3*A4+2*A3-A5

=(A2*A3*A4)-(A3*13)-4

=if(A2>2,sum(A2:A5),0)

=A4*A5/2

=(A3*A4)-(A2+A4)

=(A2*A2*A2)-A4

[Shriver, Elbaum and Stolee. “At the end of synthesis narrowing program candidates.” ICSE-NIER 2017]

	A	B
1	**Input**	**Output**
2	3	24
3	6	
4	7	
5	8	

Example 1 : 11 programs

Example 2 : 5 programs

User Gives More Examples

```
=sum(A2:A5)
=sum(A2:A12)
=A2*A5
=3*A3+A4
=2*(A4+A5)-A3
=3*A4+2*A3-A5
=(A2*A3*A4)-(A3*13)-4
=if(A2>2,sum(A2:A5),0)
=A4*A5/2
=(A3*A4)-(A2+A4)
=(A2*A2*A2)-A4
```

[Shriver, Elbaum and Stolee. “At the end of synthesis narrowing program candidates.” ICSE-NIER 2017]

	A	B
1	**Input**	**Output**
2	3	27
3	7	
4	8	
5	9	

User Gives More Examples

Example 1 : 11 programs
Example 2 : 5 programs
Example 3

=sum(A2:A5)
=sum(A2:A12)
=A2*A5
=3*A3+A4
=2*(A4+A5)-A3
=3*A4+2*A3-A5
=(A2*A3*A4)-(A3*13)-4
=if(A2>2,sum(A2:A5),0)
=A4*A5/2
=(A3*A4)-(A2+A4)
=(A2*A2*A2)-A4

[Shriver, Elbaum and Stolee. "At the end of synthesis narrowing program candidates." ICSE-NIER 2017]

	A	B
1	**Input**	**Output**
2	3	27
3	7	
4	8	
5	9	

=sum(A2:A5)

=sum(A2:A12)

=A2*A5

=3*A3+A4

=2*(A4+A5)-A3

=3*A4+2*A3-A5

=(A3*A4)-(A3*13)-4

=IF(A2,sum(A2:A5),0)

=(A2+A4)

=(A2*A2*A2)-A4

User Gives More Examples

Example 1 : 11 programs

Example 2 : 5 programs

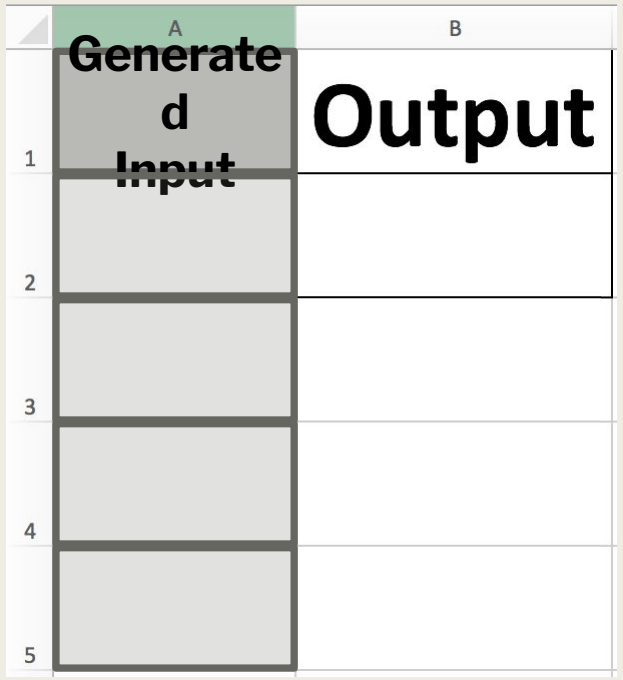
Example 3 : 5 program

!

No Programs Eliminated

[Shriver, Elbaum and Stolee. “At the end of synthesis narrowing program candidates.” ICSE-NIER 2017]

HOW CAN WE MAKE IT EASIER *FOR THE USER*?



Example 1 : 11 programs

```
=sum(A2:A5)
=sum(A2:A12)
=A2*A5
=3*A3+A4
=2*(A4+A5)-A3
=3*A4+2*A3-A5
=(A2*A3*A4)-(A3*13)-4
=if(A2>2,sum(A2:A5),0)
=A4*A5/2
=(A3*A4)-(A2+A4)
=(A2*A2*A2)-A4
```

[Shriver, Elbaum and Stolee. “At the end of synthesis narrowing program candidates.” ICSE-NIER 2017]

User gives output only

	A	
1	Generated Input	Output
2	3	15
3	6	
4	7	
5	5	

Example 1 : 11 programs

Example 2

=sum(A2:A5)

=sum(A2:A12)

=A2*A5

=3*A3+A4

=2*(A4+A5)-A3

=3*A4+2*A3-A5

=(A2*A3*A4)-(A3*13)-4

=if(A2>2,sum(A2:A5),0)

=A4*A5/2

=(A3*A4)-(A2+A4)

=(A2*A2*A2)-A4

[Shriver, Elbaum and Stolee. “At the end of synthesis narrowing program candidates.” ICSE-NIER 2017]

User gives output only

	A	
1	Generated Input	Output
2	3	15
3	6	
4	7	
5	5	

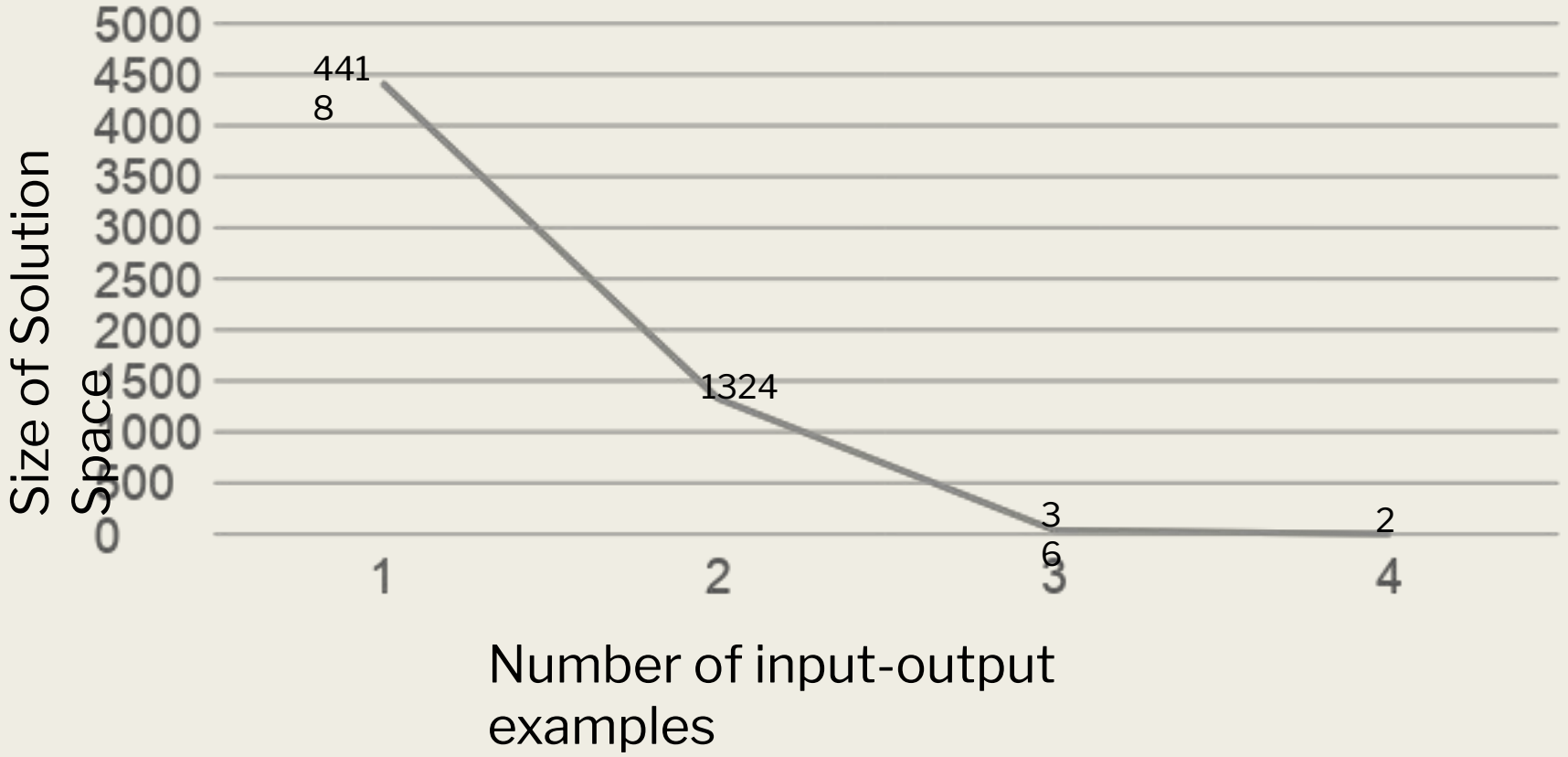
=sum(A2:A5)
=sum(A2:A12)
=A2*A5
=3*A3+A4
=2*(A4+A5)-A3
=3*A4+2*A3-A5
=(A2*A3*A4)-(A3*13)-4
=if(A2>2,sum(A2:A5),0)
=A4*A5/2
=(A3*A4)-(A2+A4)
=(A2*A2*A2)-A4

Example 1 : 11 programs
Example 2 : 1 program

[Shriver, Elbaum and Stolee. “At the end of synthesis narrowing program candidates.” ICSE-NIER 2017]

FIND INPUT THAT MAXIMIZES THE DIVERSITY OF THE OUTPUT VALUES

Solution Space Reduction on Example

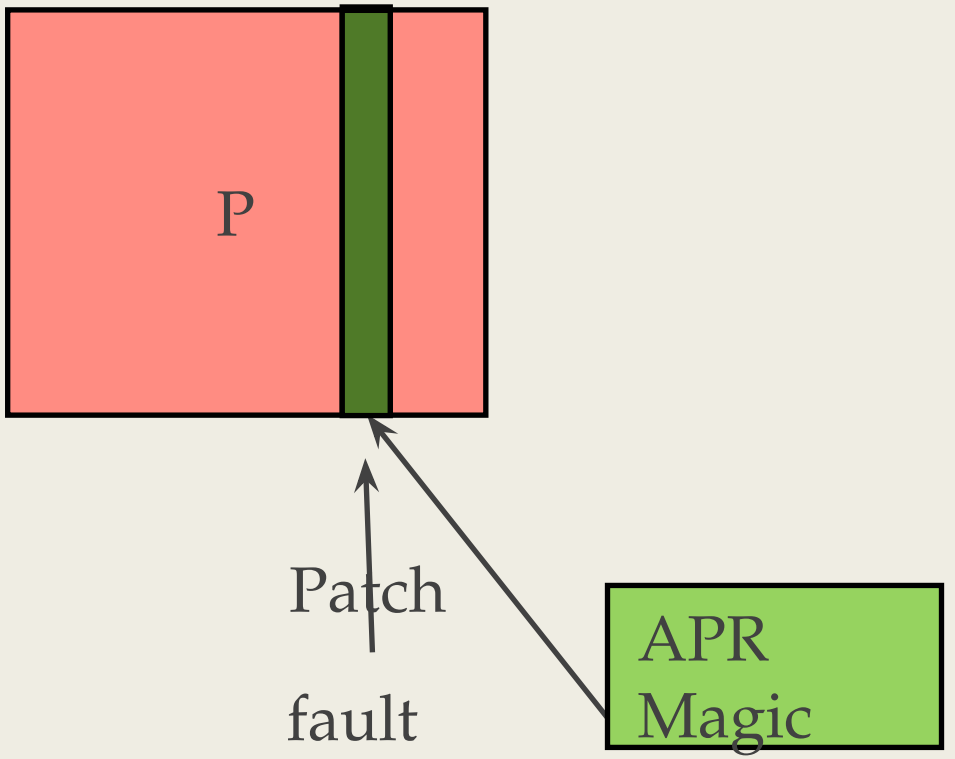


[Shriver, Elbaum and Stolee. “At the end of synthesis narrowing program candidates.” ICSE-NIER 2017]

WHAT WE CAN DO WITH BEHAVIORAL CODE SEARCH

Automated Program Repair

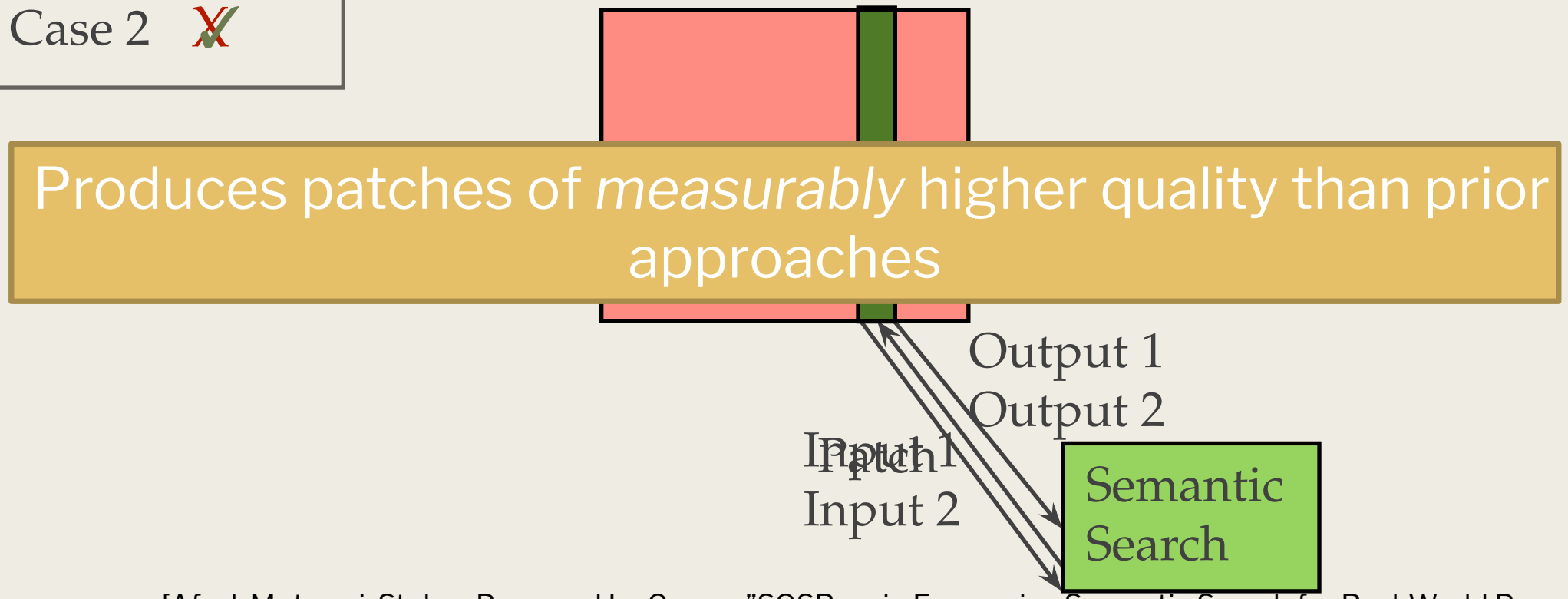
Test Suite	
Test Case 1	✓
Test Case 2	✗✓



[Afzal, Motwani, Stolee, Brun, and Le Goues. "SOSRepair: Expressive Semantic Search for Real-World Program Repair." TSE 2020]

My Automated Program Repa

Test Suite	
Test Case 1	✓
Test Case 2	✗✓



[Afzal, Motwani, Stolee, Brun, and Le Goues. "SOSRepair: Expressive Semantic Search for Real-World Program Repair." TSE 2020]

CAN IT PATCH REAL BUGS?

Python bug #69223

Developer Patch

```
   }
+  if (timeout < 0) {
+    PyErr_SetString(PyExc_ValueError ,
+        "timeout must be non-negative");
+    return NULL;
+  }
   seconds = (long)timeout;
```

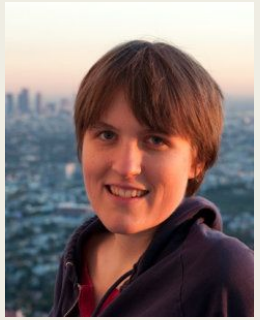
SearchRepair Patch

```
if ( timeout < 0
    PyErr_SetString(PyExc_ValueError ,
       ”read length must be positive");
    return NULL;
}
```

THE FUTURE IS SEARCH

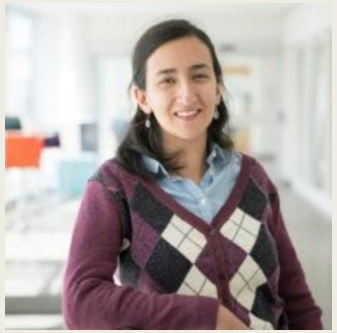
THANKS



- **Funding**:
  - *[2018-2023] NSF CAREER #1749936: On the Foundations of Semantic Code Search*
  - *[2016-2020] NSF SHF Medium #1645136: Collaborative Research: Semi and Fully Automated Program Repair and Synthesis via Semantic Code Search*
  - *[2014-2016] NSF SHF EAGER #1446932: Collaborative Research: Demonstrating the Feasibility of Automatic Program Repair Guided by Semantic Code Search.*




…. All the undergrads: Joshua Kayani, Sydney Paul, Jed Barson, Daniel