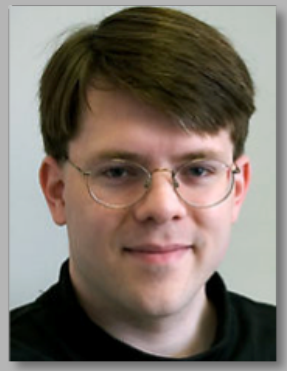
REPRESENTATIONS AND OPERATORS FOR IMPROVING EVOLUTIONARY SOFTWARE REPAIR

Claire Le Goues



Westley Weimer

Stephanie Forrest

THE UNIVERSITY of NEW MEXICO

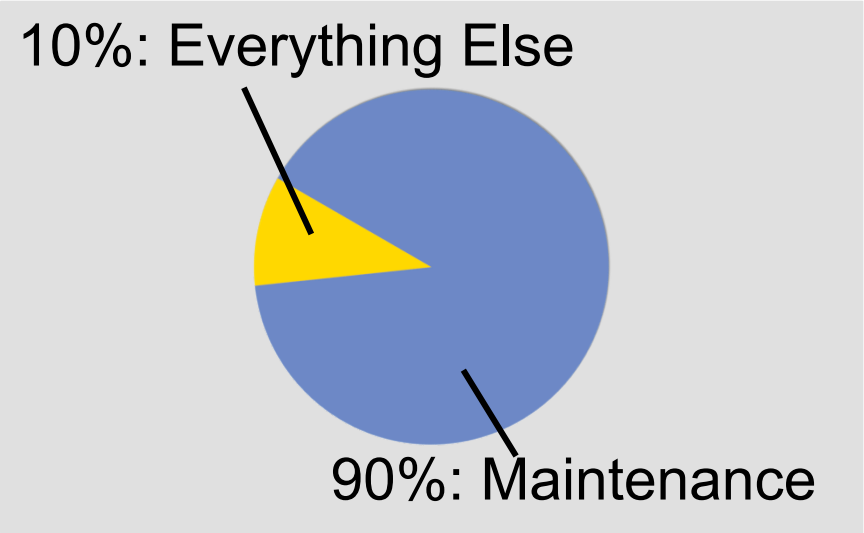
"Everyday, almost 300 bugs appear […] far too many for only the Mozilla programmers to handle."

– *Mozilla Developer, 2005*

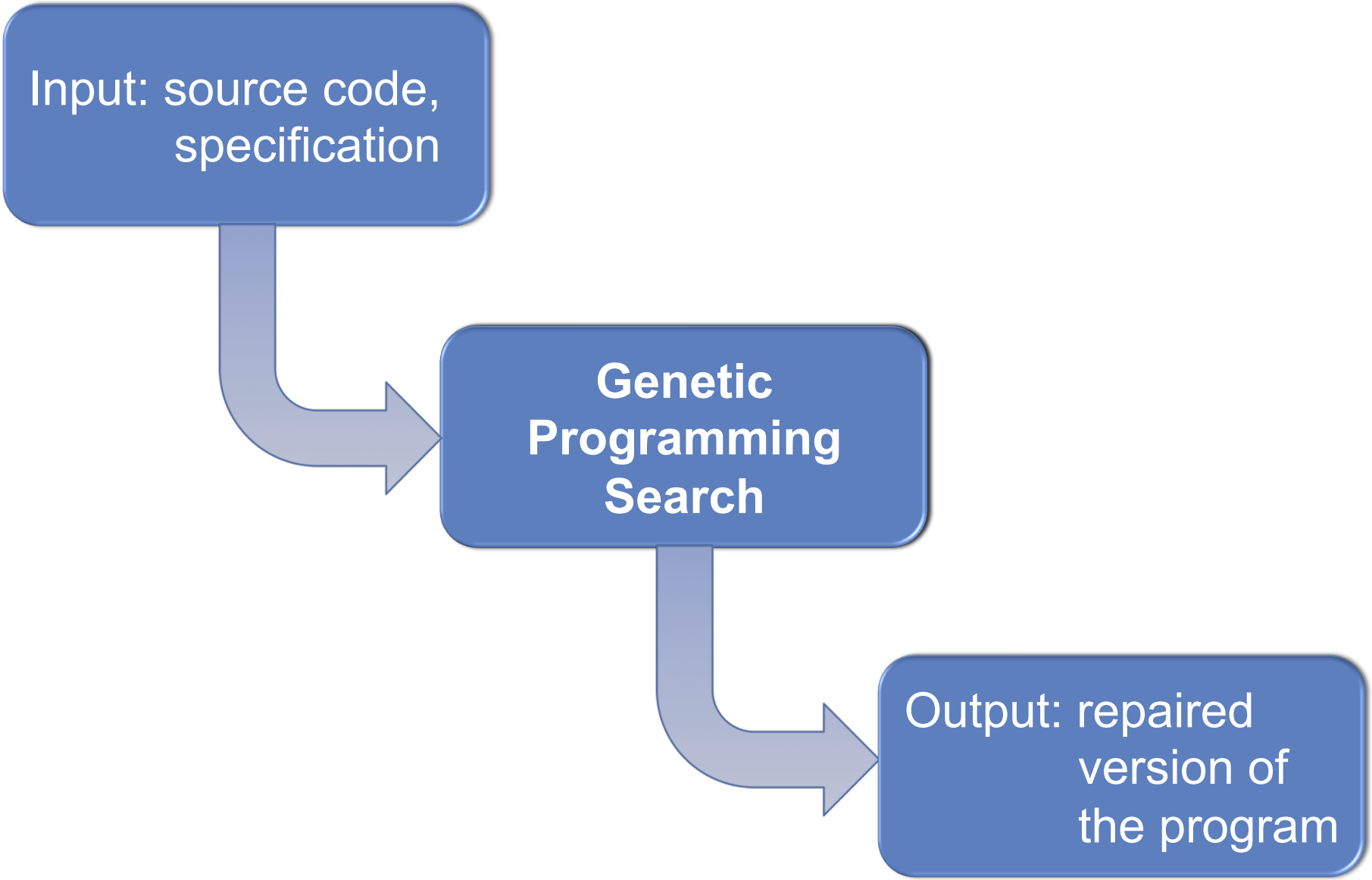
Annual cost of software errors in the US: $59.5 billion (0.6% of GDP).

PROBLEM: BUGGY SOFTWARE

Average time to fix a security-critical error: 28 days.



APPROACH: EVOLUTIONARY COMPUTATION


Input: source code, specification
Genetic Programming Search
Output: repaired version of the program

SEARCH SPACE

OTHER GP PROBLEMS

GECCO GP-TRACK BEST PAPERS

Learning: expression trees or lists

Population: 64 – 2500

Iterations: 50 – 10000

Max variant size: 16 operations, 48 operations, 17 levels, 11 levels

PROGRAM REPAIR

Learning: patches or repaired programs

Population: 40

Iterations: 10

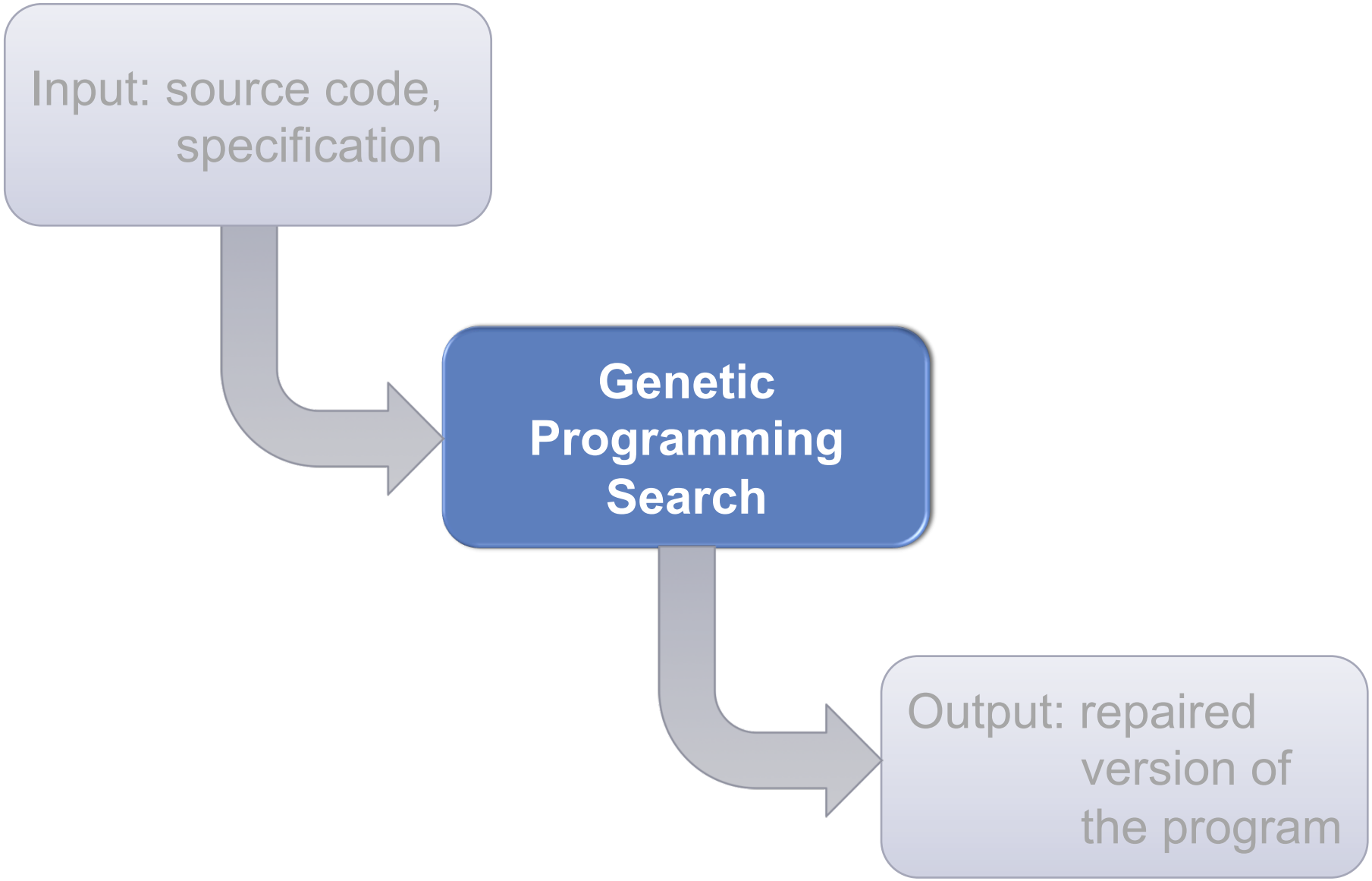
Max variant size: unbounded

Largest benchmark program: 2.8 million lines of C code.

EC-based repair starts with a large genome.

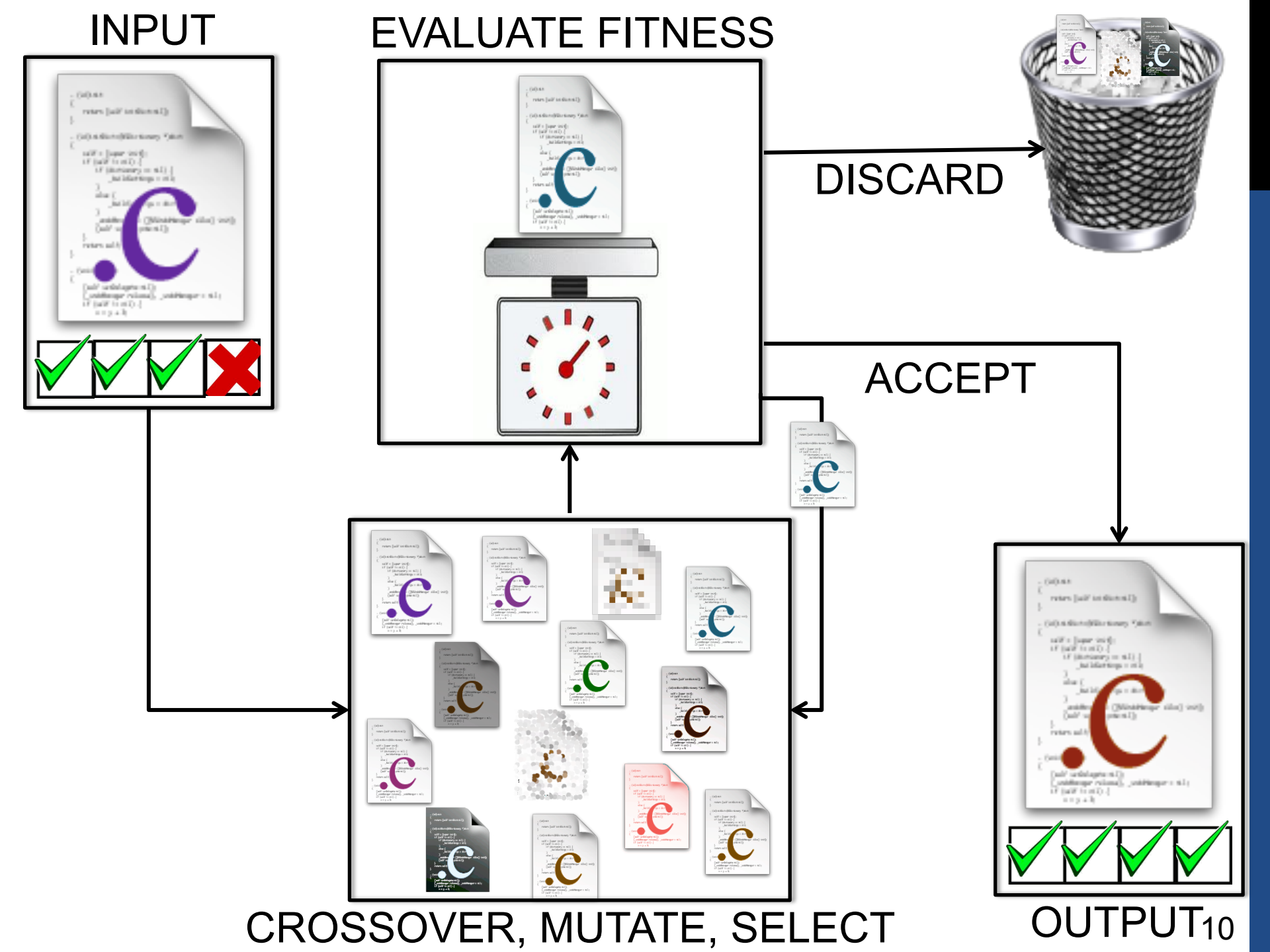
SEARCH SPACE

The starting individual is mostly correct.


Input: source code, specification
Genetic Programming Search
Output: repaired version of the program

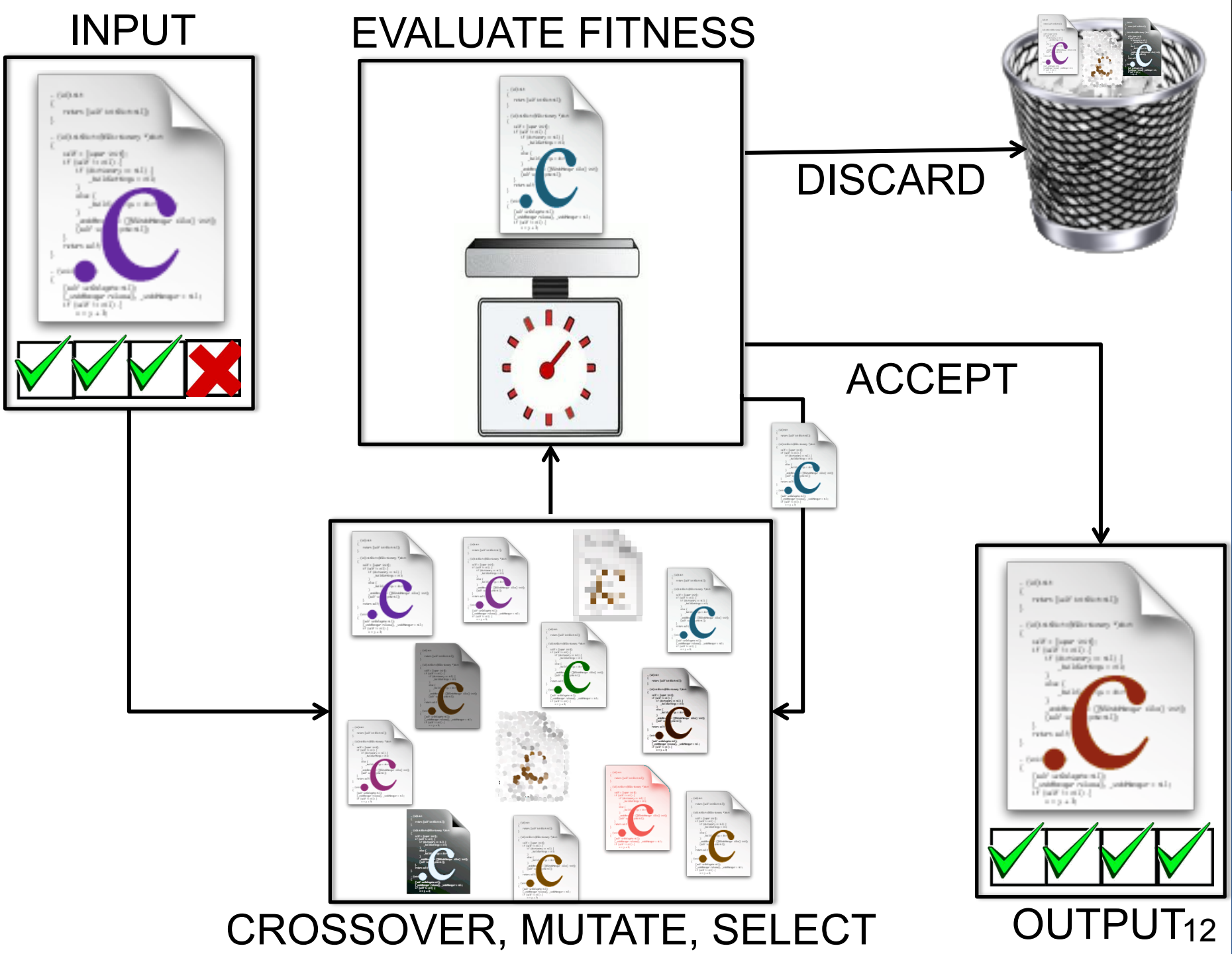
OUR GOAL

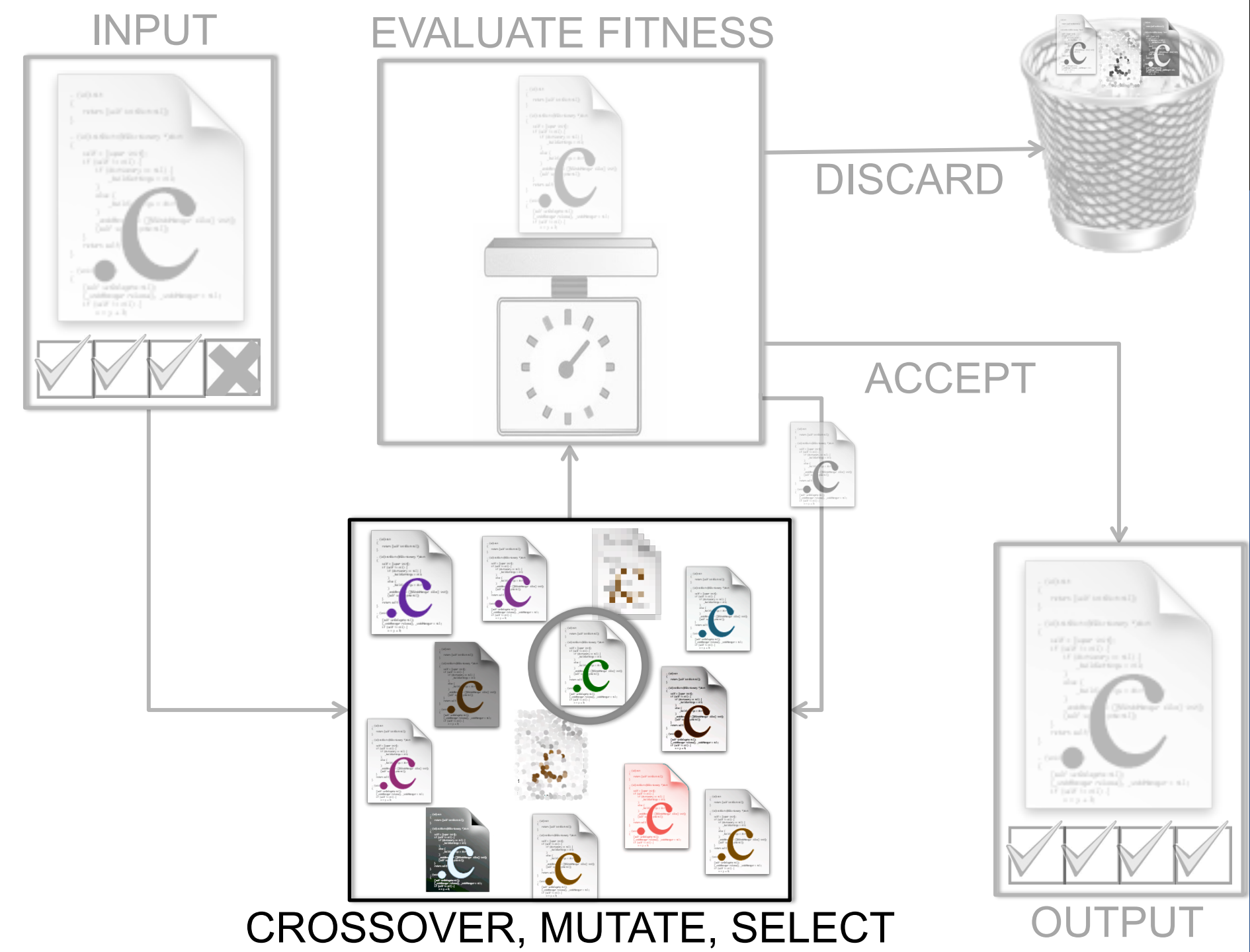
IN-DEPTH STUDY OF REPRESENTATION AND OPERATORS FOR EVOLUTIONARY PROGRAM REPAIR.

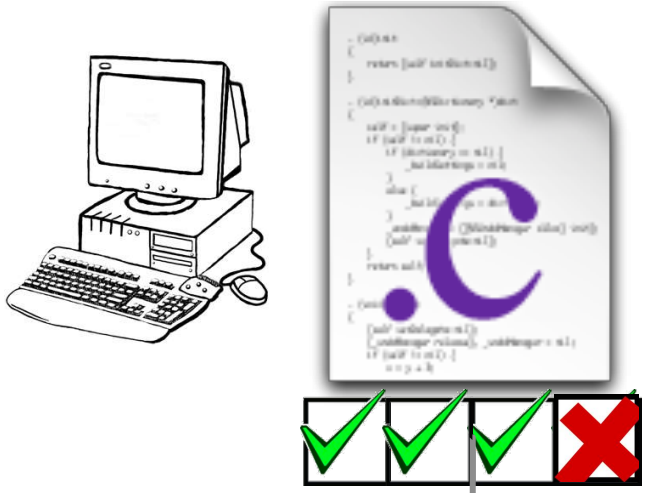

INPUT
EVALUATE FITNESS
DISCARD
ACCEPT
CROSSOVER, MUTATE, SELECT
OUTPUT10

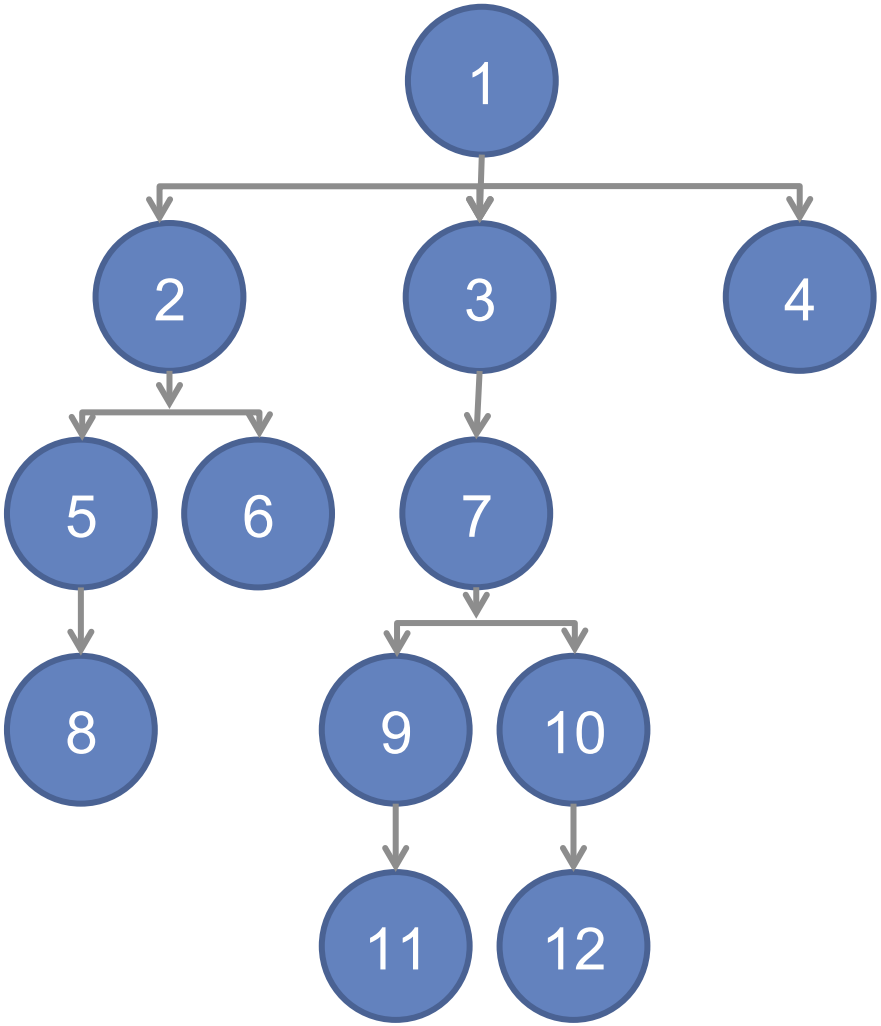
OUR GOAL

IN-DEPTH STUDY OF REPRESENTATION AND OPERATORS FOR EVOLUTIONARY PROGRAM REPAIR.


INPUT
EVALUATE FITNESS
DISCARD
ACCEPT
CROSSOVER, MUTATE, SELECT
OUTPUT12

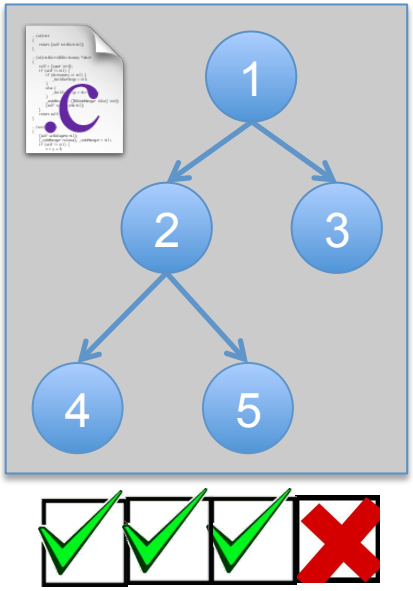

INPUT
EVALUATE FITNESS
DISCARD
ACCEPT
CROSSOVER, MUTATE, SELECT
OUTPUT



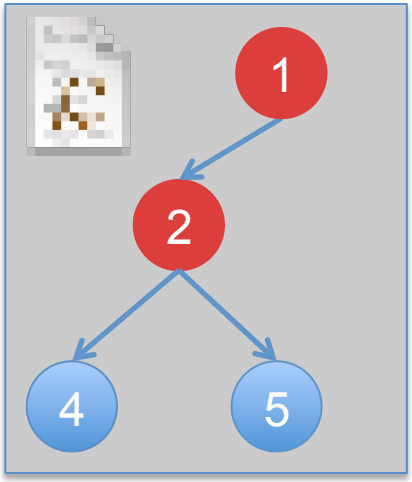

1
2
3
4
5
6
7
8
9
10
11
12

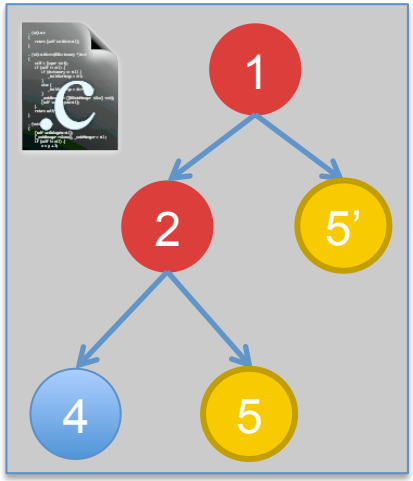
REPRESENTATION

Input:



AST/WP:





Patch:

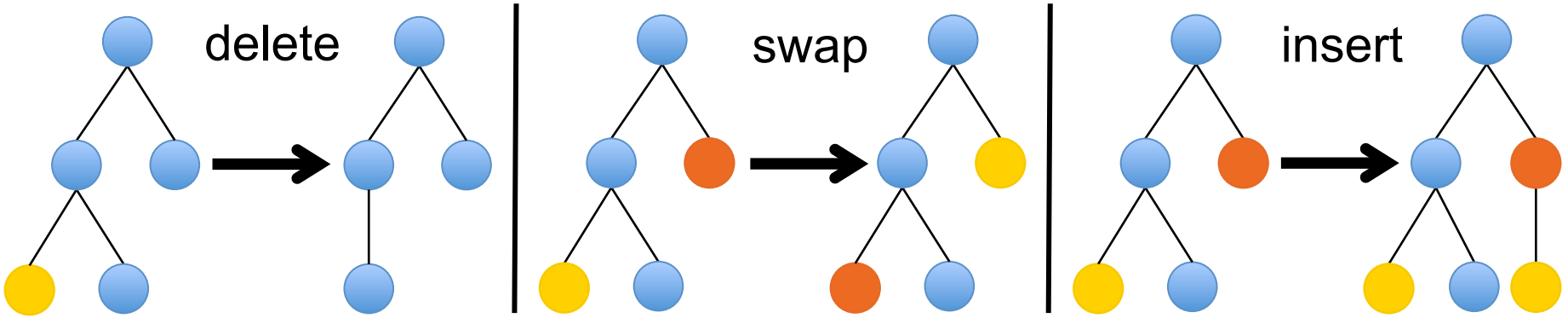

Delete(3)


Delete(3) Insert(2,4) Replace(5,1)


Insert(5,4) Insert(3,3) Delete(4) …


Replace(3,5)

GENETIC OPERATORS



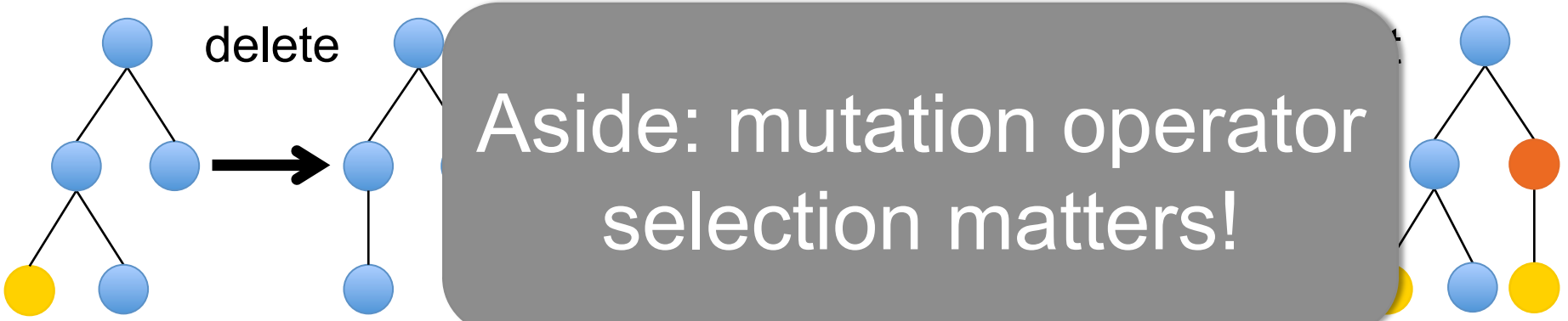
Mutation operators

- Manipulate only existing genetic material.
- Semantic checking improves probability that mutation is viable.

Crossover:

- One-point: on the weighted path or edit list.
- Patch-subset: uniform, on the edit list.

GENETIC OPERATORS



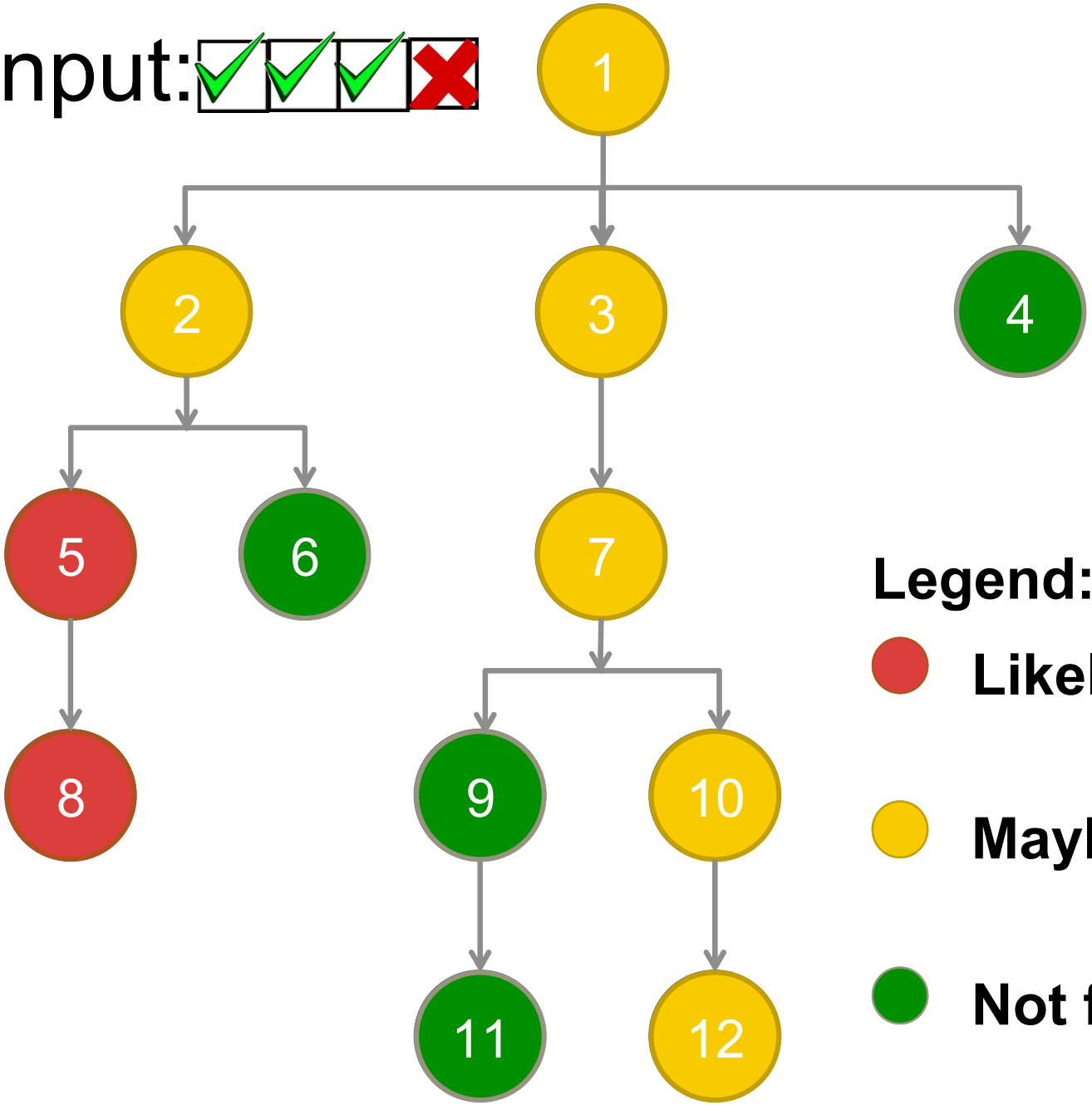
Mutation operators

- Manipulate only existing genetic material.
- Semantic checking improves probability that mutation is viable.

Crossover:

- One-point: on the weighted path or edit list.
- Patch-subset: uniform, on the edit list.

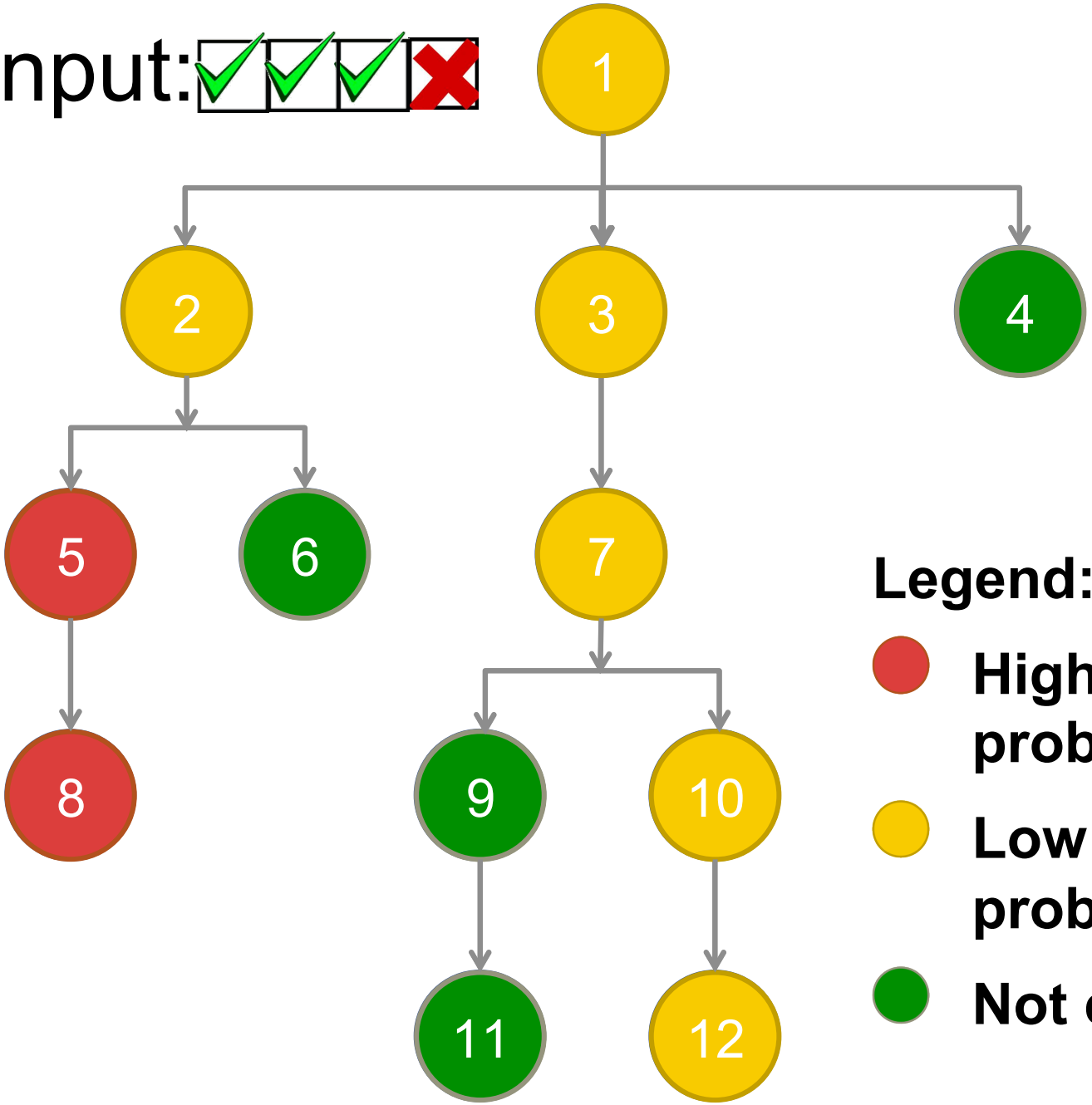
Input: ✓ ✓ ✓ ✗



Legend:

- **Likely faulty.**
- **Maybe faulty.**
- **Not faulty.**

Input: ✓ ✓ ✓ ✗



Legend:

- **High change probability.**
- **Low change probability.**
- **Not changed.**

Input: 

1
4
Default: 10 : 1 ratio
8
9
10
11
12
gend:
High change probability.
Low change probability.
Not changed.

OUR GOAL

IN-DEPTH STUDY OF REPRESENTATION AND OPERATORS FOR EVOLUTIONARY PROGRAM REPAIR.

EXPERIMENTAL SETUP

Benchmarks: 105 bugs in 8 real-world programs.[1]

- 5 million lines of C code, 10,000 test cases.
- Bugs correspond to human-written repairs for regression test failures.

Default parameters, for comparison:

- Patch representation.
- Mutation operators selected with equal random probability. 1 mutation, 1 crossover/individual/iteration.
- Population size: 40. 10 iterations or 12 wall-clock hours, whichever comes first. Tournament size: 2.

[1]Claire Le Goues, Michael Dewey-Vogt, Stephanie Forrest, and Westley Weimer, "A Systematic Study of Automated Program Repair: Fixing 55 out of 105 bugs for $8 Each." International Conference on Software Engineering (ICSE), 2012, pp. 3 – 13.

EXPERIMENTAL SETUP

55/105 bugs repaired using default parameters.

Some bugs are more difficult to repair than others!

- Easy: 100% success rate on default parameters.
- Medium: 50 – 100% success rate on default parameters
- Hard: 1 – 50% success rate on default parameters
- Unfixed: 0% success

Metrics:

- # fitness evaluations to a repair
- GP success rate.

RESEARCH QUESTIONS

Representation:

- Which representation choice gives better results?
- Which representation features contribute most to success?

Crossover: Which crossover operator is best?

Operators:

- Which operators contribute the most to success?
- How should they be selected?

Search space: How should the representation weight program statements to best define the search space?

RESEARCH QUESTIONS

Representation:

- Which representation choice gives better results?
- Which representation features contribute most to success?

Crossover: Which crossover operator is best?

Operators:

- Which operators contribute the most to success?
- How should they be selected?

Search space: How should the representation weight program statements to best define the search space?

RESEARCH QUESTIONS

Representation:

- **Which representation choice gives better results?**
- **Which representation features contribute most to success?**

Crossover: Which crossover operator is best?

Operators:

- Which operators contribute the most to success?
- How should they be selected?

Search space: How should the representation weight program statements to best define the search space?

REPRESENTATION: RESULTS

Procedure:

Compare AST/WP to PATCH on original benchmarks with default parameters.

For both representations, test effectiveness of:

1. Crossover.
2. Semantic check.

Results:

1. Patch outperforms AST/WP (14 – 30%).
2. Semantic check strongly influences success rate of both representations.
3. Crossover also improves results.

RESEARCH QUESTIONS

Representation:

- Which representation choice gives better results?
- Which representation features contribute most to success?

Crossover: Which crossover operator is best?

Operators:

- Which operators contribute the most to success?
- How should they be selected?

Search space: How should the representation weight program statements to best define the search space?

CROSSOVER: RESULTS

Crossover Operator	Success Rate	Fitness evaluations to repair
None	54.4%	82.43
Default/“Uniform”	61.1%	163.05
One-Point/AST-WP	63.7%	114.12
One-Point/Patch	65.2%	118.20

RESEARCH QUESTIONS

Representation:

- Which representation choice gives better results?
- Which representation features contribute most to success?

Crossover: Which crossover operator is best?

Operators:

- Which operators contribute the most to success?
- How should they be selected?

Search space: How should the representation weight program statements to best define the search space?

SEARCH SPACE: SETUP

Hypothesis: statements executed only by the failing test case(s) should be mutated more often than those also executed by the passing test cases.

Procedure: examine that ratio in actual repairs.

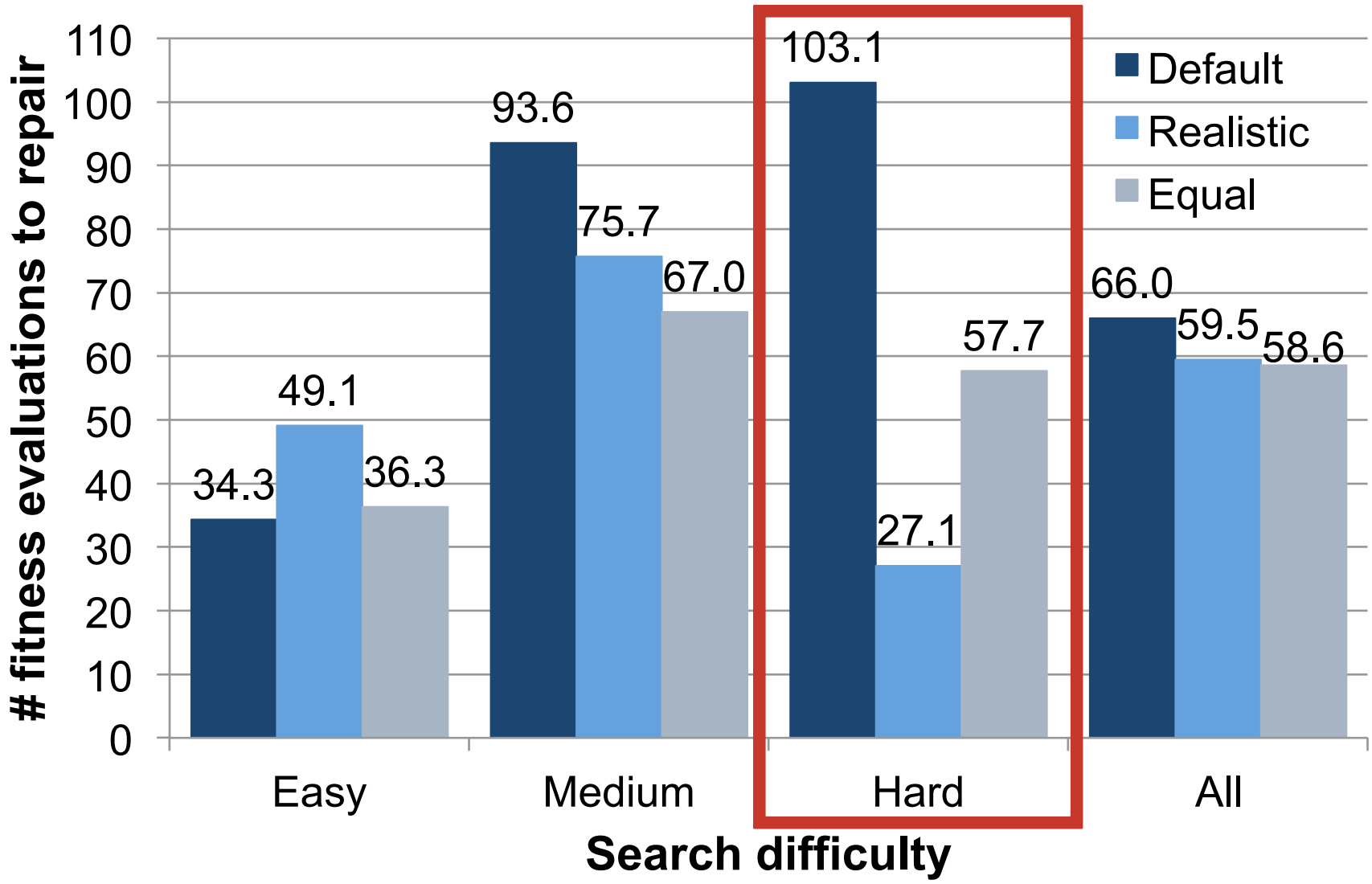
Result:

Expected: 10 : 1

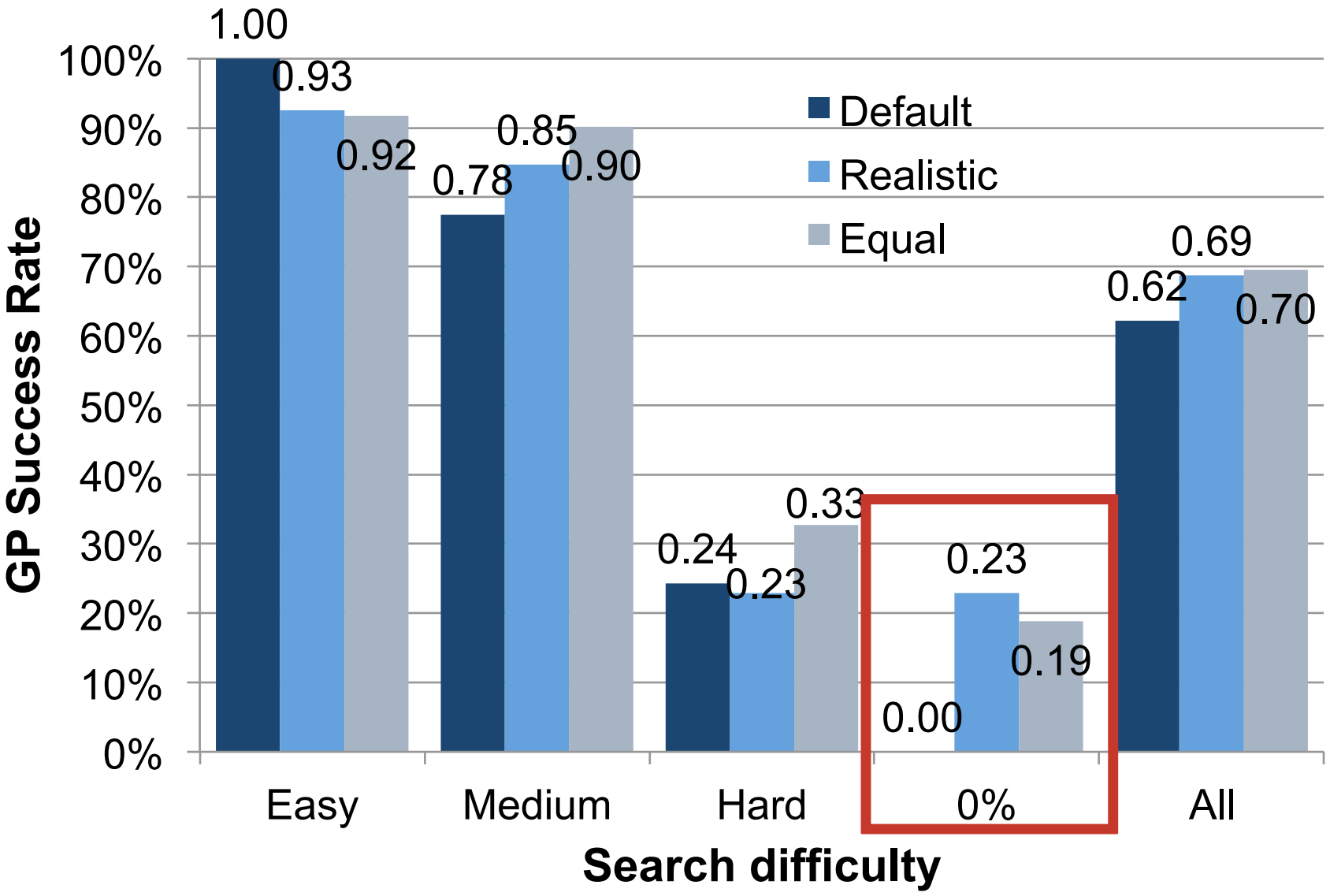
vs.

Actual: 1 : 1.85

SEARCH SPACE: REPAIR TIME



SEARCH SPACE: SUCCESS RATE



CONCLUSIONS/DISCUSSION

This EC problem is atypical; atypical problems warrant study.

We studied representation and operators for EC-based bug repair. These choices matter, especially for difficult bugs.

- Incorporating all recommendations, GenProg repairs 5 new bugs; repair time decreases by 17–43% on difficult bugs.

We don't know why some of these things are true, but:

- We now have lots of interesting data to dig into!
- We are currently (as we sit here) doing more and bigger runs with the new parameters on the as-yet unfixed bugs.

PLEASE ASK QUESTIONS

REPRESENTATION BENCHMARKS

Program	LOC	Tests	Bug	Description
gcd	22	6	infinite loop	Example
uniq-utx	1146	6	segfault	Text processing
look-utx	1169	6	segfault	Dictionary lookup
look-svr	1363	6	infinite loop	Dictionary lookup
units-svr	1504	6	segfault	Metric conversion
deroff-utx	2236	6	segfault	Document processing
nullhttpd	5575	7	buffer exploit	Webserver
indent	9906	6	infinite loop	Source code processing
flex	18775	6	segfault	Lexical analyzer generator
atris	21553	3	buffer exploit	Graphical tetris game
Total	**63249**			

SEARCH SPACE BENCHMARKS

Program	LOC	Tests	Bugs	Description
fbc	97,000	773	3	Language (legacy)
gmp	145,000	146	2	Multiple precision math
gzip	491,000	12	5	Data compression
libtiff	77,000	78	24	Image manipulation
lighttpd	62,000	295	9	Web server
php	1,046,000	8,471	44	Language (web)
python	407,000	355	11	Language (general)
wireshark	2,814,000	63	7	Network packet analyzer
Total	**5,139,000**	**10,193**	**105**	

55/105 bugs repaired using default parameters.