Parameter Tuning for Search-Based Test-Data Generation Revisited

Support for Previous Results

Anton Kotelyanskii

Gregory M. Kapfhammer

Software Testing

Software Testing

Test Suites

Software Testing

Test Suites

Automatic Generation

Software Testing

Test Suites

Automatic Generation

Confronting Challenges

Software Testing

Test Suites

Automatic Generation

Confronting Challenges

Evaluation Strategies

Empirical Studies

Empirical Studies

Challenges

Empirical Studies

Challenges

Importance

Empirical Studies

Challenges
Importance
Replication

Empirical Studies

Challenges
Importance
Replication
Rarity


01010

EvoSuite

EvoSuite

Amazing test suite generator

creative commons licensed (BY-SA) flickr photo shared by mcclanahoochie

EvoSuite

Amazing test suite generator

Uses a genetic algorithm

EvoSuite

Amazing test suite generator

Uses a genetic algorithm

Input: A Java class

EvoSuite

Amazing test suite generator

Uses a genetic algorithm

Input: A Java class

Output: A JUnit test suite

EvoSuite

Amazing test suite generator

Uses a genetic algorithm

Input: A Java class

Output: A JUnit test suite

http://www.evosuite.org/

Parameter Tuning

Parameter Tuning

RSM: Response surface methodology

Parameter Tuning

RSM: Response surface methodology

SPOT: Sequential parameter optimization toolbox

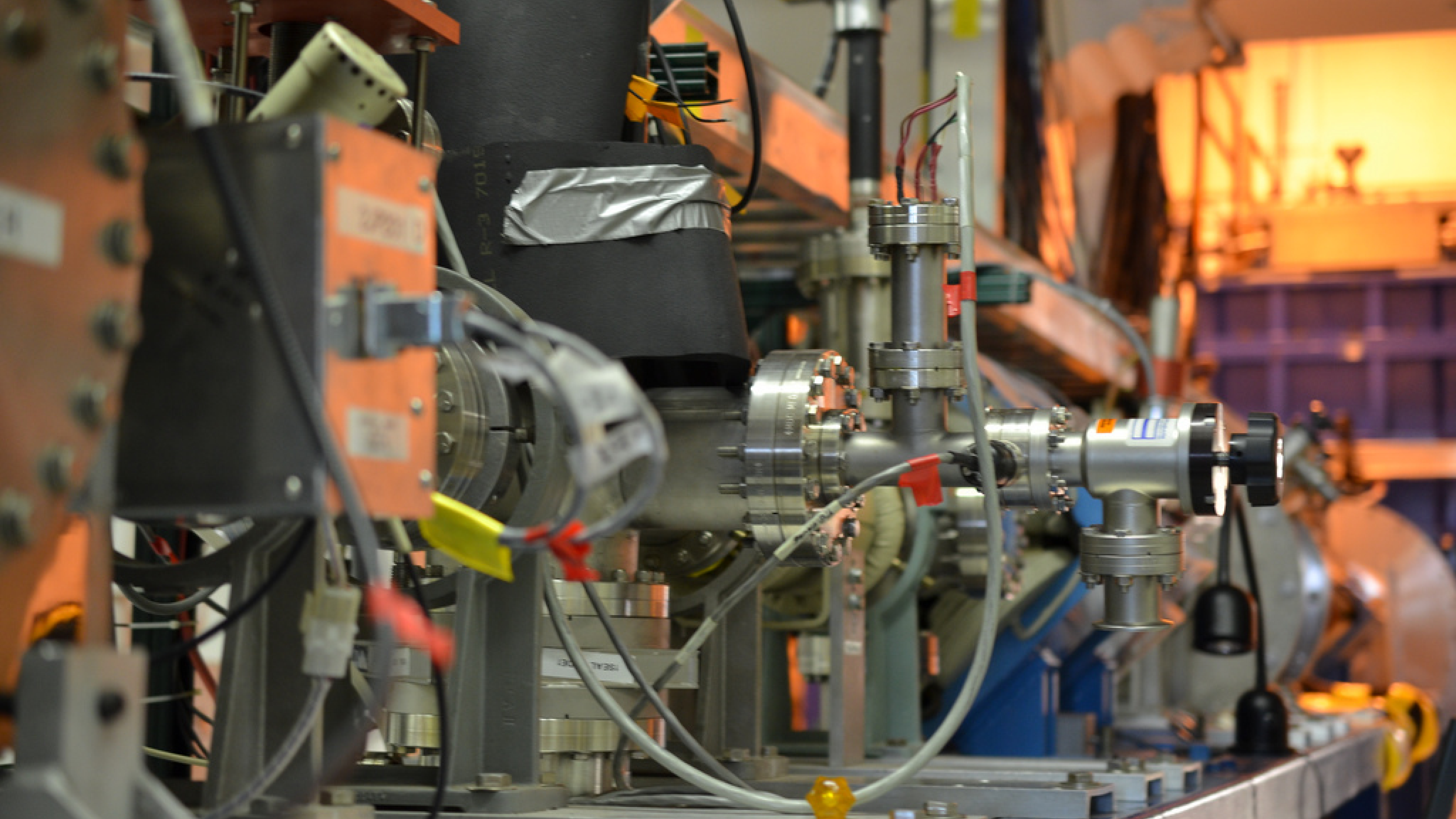
Parameter Tuning

RSM: Response surface methodology

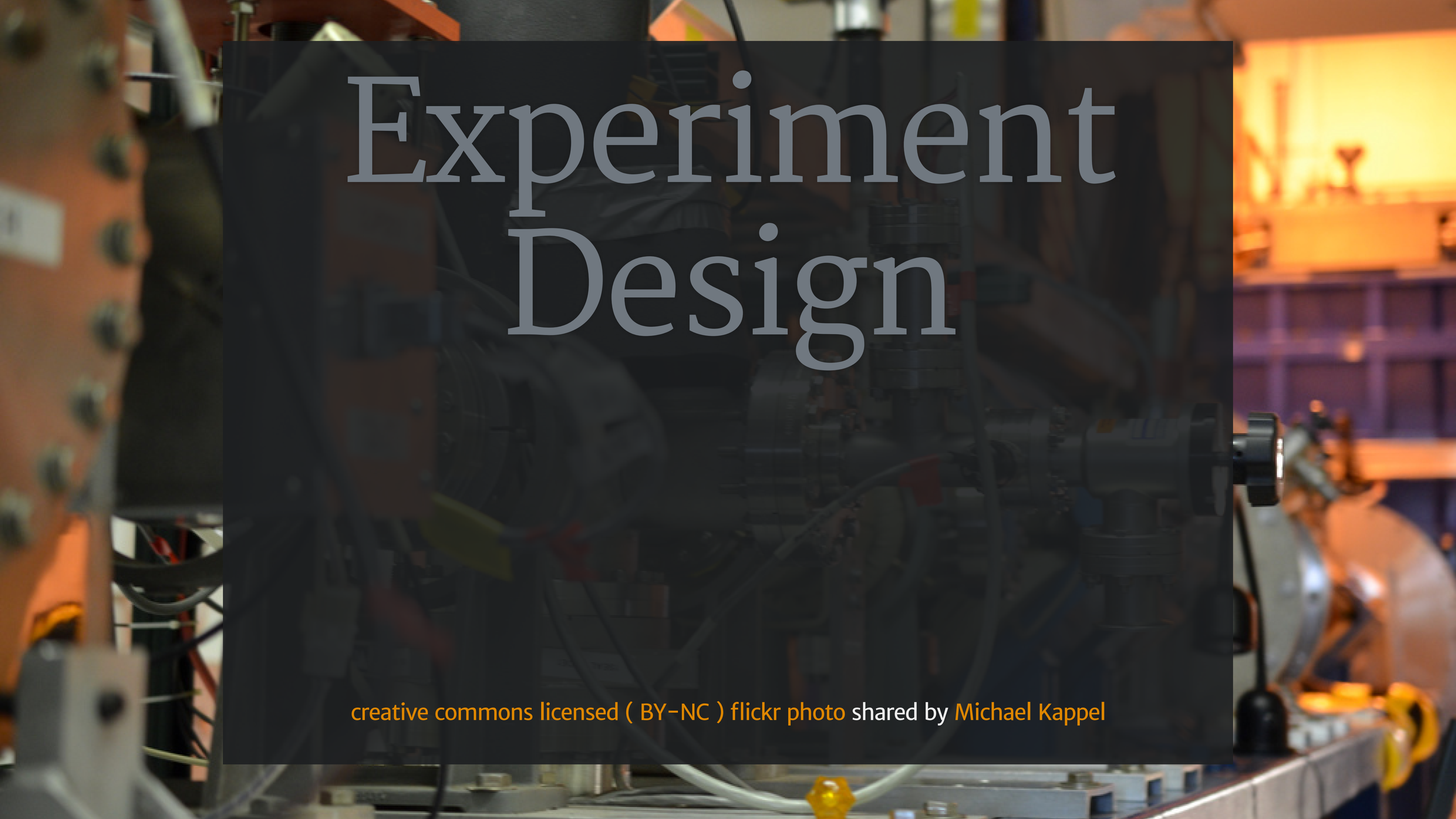
SPOT: Sequential parameter optimization toolbox

Successfully applied to many diverse problems!

Defaults or Tuned Values?



Experiment Design



Experiment Design

Eight EvoSuite parameters

Experiment Design

Eight EvoSuite parameters

Ten projects from SF100

Experiment Design

Eight EvoSuite parameters

Ten projects from SF100

475 Java classes for subjects

Experiment Design

Eight EvoSuite parameters
Ten projects from SF100
475 Java classes for subjects
100 trials after parameter tuning

Experiment Design

Eight EvoSuite parameters

Ten projects from SF100

475 Java classes for subjects

100 trials after parameter tuning

Aiming to improve statement coverage

Parameters

Parameter Name	Minimum	Maximum
Population Size	5	99
Chromosome Length	5	99
Rank Bias	1.01	1.99
Number of Mutations	1	10
Max Initial Test Count	1	10
Crossover Rate	0.01	0.99
Constant Pool Use Probability	0.01	0.99
Test Insertion Probability	0.01	0.99

Experiments

Experiments

184 days of computation time estimated

Experiments

184 days of computation time estimated

Cluster of 70 computers running for weeks

Experiments

184 days of computation time estimated

Cluster of 70 computers running for weeks

Identified 139 "easy" and 21 "hard" classes

Experiments

184 days of computation time estimated

Cluster of 70 computers running for weeks

Identified 139 "easy" and 21 "hard" classes

Mann-Whitney U-test *and*

Experiments

184 days of computation time estimated
Cluster of 70 computers running for weeks
Identified 139 "easy" and 21 "hard" classes
Mann-Whitney U-test *and*
Vargha-Delaney effect size

Results

Category	Effect Size	p-value
Results Across Trials and Classes	0.5029	0.1045
No "Easy" and "Hard" Classes	0.5048	0.0314

Results

Category	Effect Size	p-value
Results Across Trials and Classes	0.5029	0.1045
No "Easy" and "Hard" Classes	0.5048	0.0314

Using *lower-is-better* inverse statement coverage

Results

Category	**Effect Size**	**p-value**
Results Across Trials and Classes	0.5029	0.1045
No "Easy" and "Hard" Classes	0.5048	0.0314

Using *lower-is-better* inverse statement coverage

Effect size *greater* than 0.5 means that tuning is *worse*

Results

Category	Effect Size	p-value
Results Across Trials and Classes	0.5029	0.1045
No "Easy" and "Hard" Classes	0.5048	0.0314

Using *lower-is-better* inverse statement coverage

Effect size *greater* than 0.5 means that tuning is *worse*

Testing shows we do not *always* reject the null hypothesis

Results

Category	Effect Size	p-value
Results Across Trials and Classes	0.5029	0.1045
No "Easy" and "Hard" Classes	0.5048	0.0314

Using *lower-is-better* inverse statement coverage

Effect size *greater* than 0.5 means that tuning is *worse*

Testing shows we do not *always* reject the null hypothesis

Additional empirical results in the QSIC 2014 paper!

Discussion

Discussion

Tuning improved scores for 11 classes

Discussion

Tuning improved scores for 11 classes

Otherwise, same as or worse than defaults

Discussion

Tuning improved scores for 11 classes

Otherwise, same as or worse than defaults

A "soft floor" may exist for parameter tuning

Discussion

Tuning improved scores for 11 classes

Otherwise, same as or worse than defaults

A "soft floor" may exist for parameter tuning

Additional details in the QSIC 2014 paper!

Practical Implications

Practical Implications

Fundamental Challenges

Practical Implications

Fundamental Challenges
Tremendous Confidence

Practical Implications

Fundamental Challenges
Tremendous Confidence
Great Opportunities


33
N
3
6
W
E
24
21
S
15
12
VACUUM
PUSH

Important Contributions

Important Contributions

Comprehensive Experiments

Important Contributions

Comprehensive Experiments

Conclusive Confirmation

Important Contributions

Comprehensive Experiments

Conclusive Confirmation

For EvoSuite, *Defaults* = *Tuned*