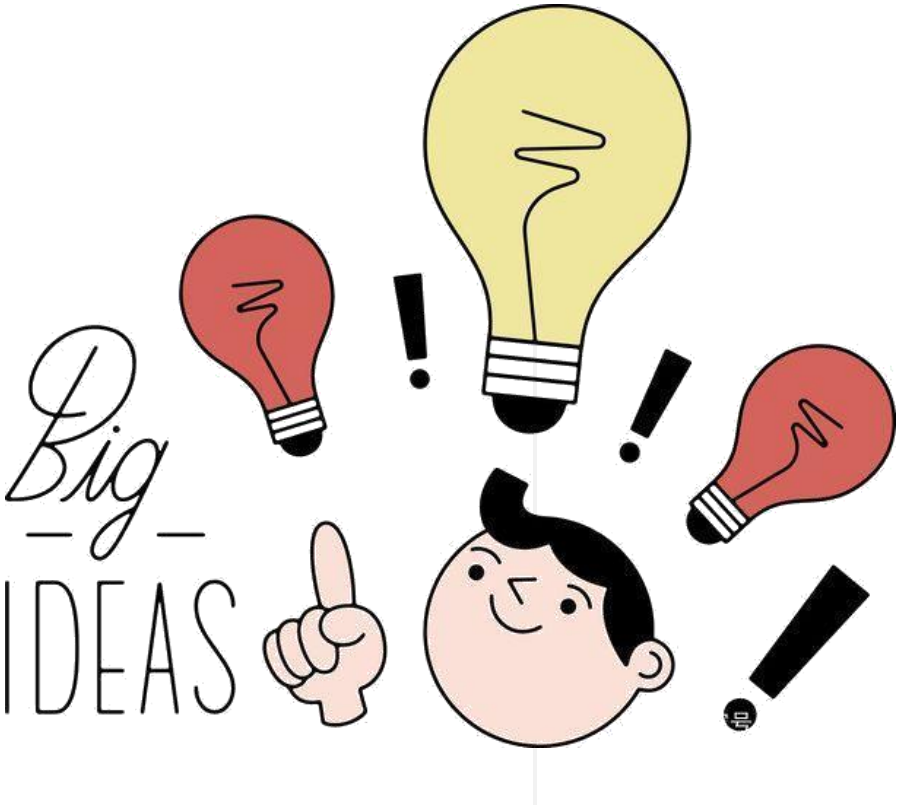
Automatic Generation of Smell-free Unit

汇报人　　李博诺

01

Key Word

关键词



关键词

01

test smell

bad programming practice

用来描述测试代码中可能存在的设计问题、代码质量问题或测试方法不当的情况，这些问题可能导致测试代码的可读性、可维护性和可靠性下降。

02

Evosuite

an automated software test case generation tool

本研究就是基于Evosuite已有的搜索算法设计新的方法来辅助其生成无异味的测试用例。

03

secondary criteria

test smell metrics

将经过筛选的测试异味指标纳入Evosuite的二级标准中对原有二级标准进行优化，来提高自动生成的测试用例的质量，并确保生成的测试代码符合良好的编程实践。

02

test smell metrics

测试异味指标

INDICATORS

Metric
AssertionRoulette
DuplicateAssert
EagerTest
IndirectTesting
LackOfCohesionOfMethods
LazyTest
LikelyIneffectiveObjectComparison
ObscureInlineSetup
Overreferencing
RedundantAssertion
RottenGreenTests
TestRedundancy
UnknownTest
UnrelatedAssertions
UnusedInputs
VerboseTest

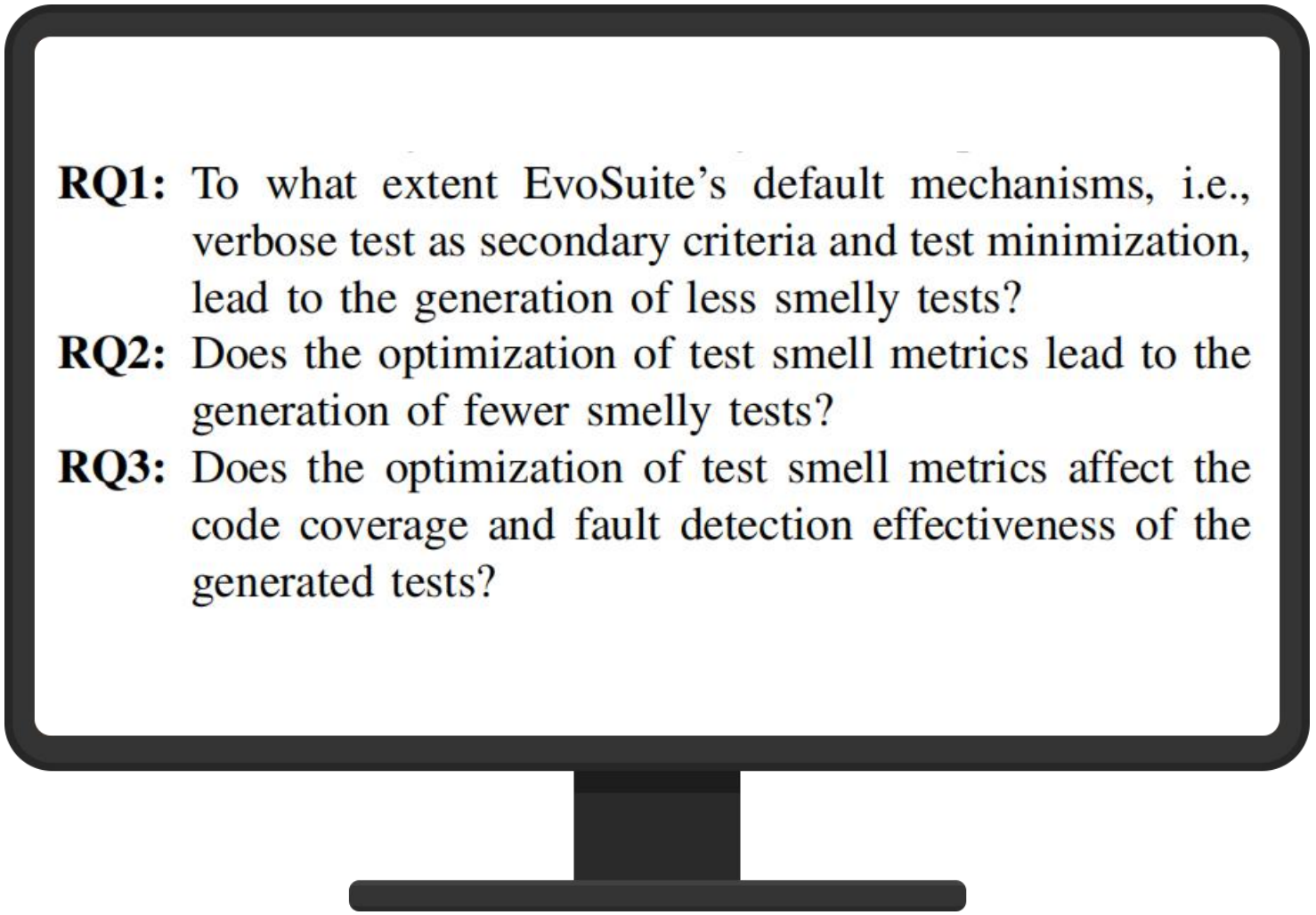
03

实例研究

1.EvoSuite’s default mechanisms

2.the effect of the optimization of test smell metrics

Research Questions



研究过程

01

02

03

实验对象

使用一个从117个开源项目中提取的共346个非平凡Java类的集合

实验指标

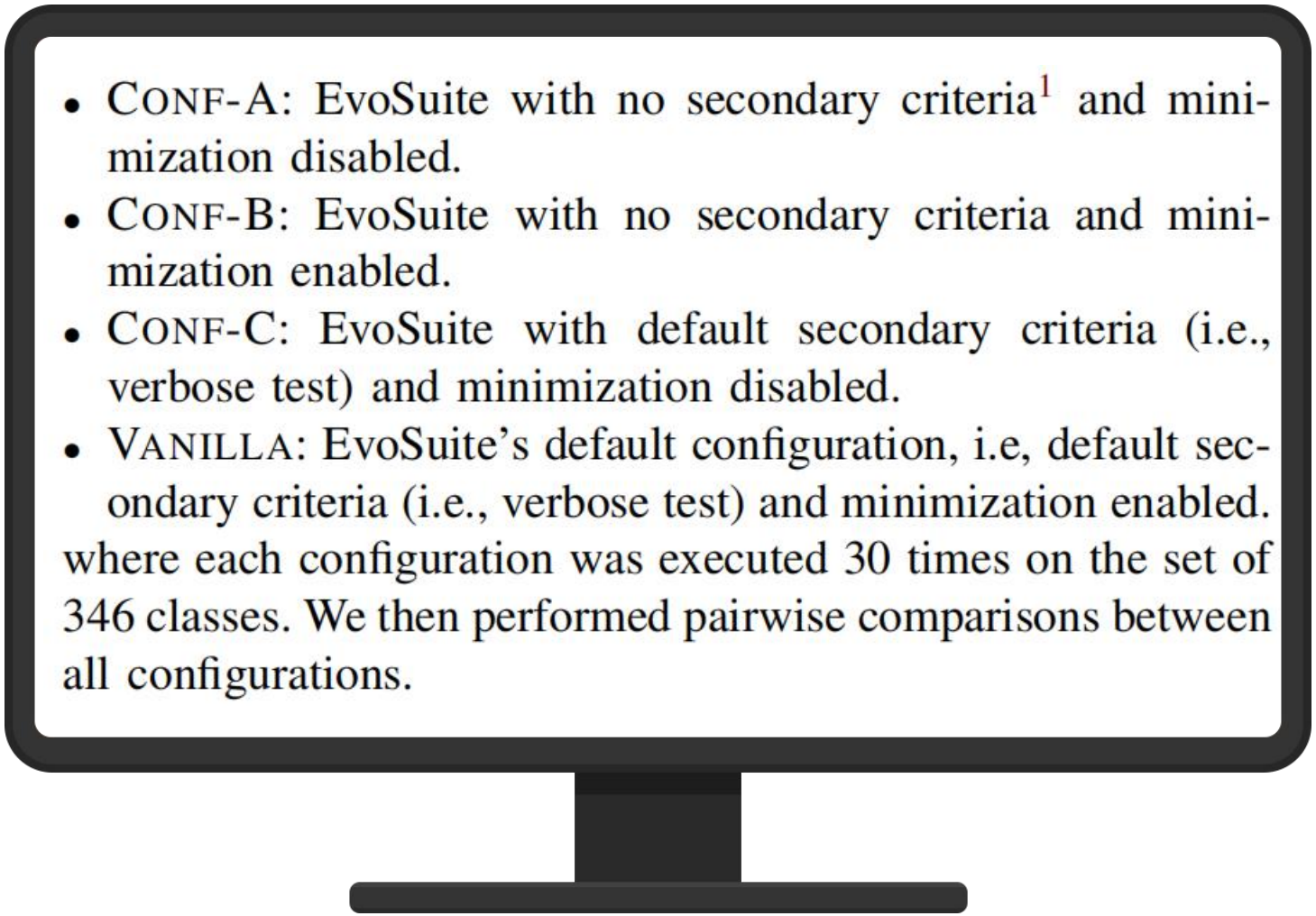
在Evosuite的每次执行中，我们收集生成测试的代码覆盖率、突变分数以及每个异味指标的值

实验内容

在Evosuite的默认设置（搜索算法为DynaMOSA）中考虑不同的配置解决不同的问题

RQ1

To what extent EvoSuite's default mechanisms, i.e., verbose test as secondary criteria and test minimization, lead to the generation of less smelly tests?



- CONF-A: EvoSuite with no secondary criteria[1] and minimization disabled.
- CONF-B: EvoSuite with no secondary criteria and minimization enabled.
- CONF-C: EvoSuite with default secondary criteria (i.e., verbose test) and minimization disabled.
- VANILLA: EvoSuite's default configuration, i.e, default secondary criteria (i.e., verbose test) and minimization enabled.

where each configuration was executed 30 times on the set of 346 classes. We then performed pairwise comparisons between all configurations.

Remaining Research Questions

01

RQ2

one additional configuration of EvoSuite:

• $S_{MELLESS}$: EvoSuite's secondary criteria configured with the combination of all smell metrics that could be optimized as a secondary criteria (i.e., Eager Test, Indirect Testing,Likely Ineffective Object Comparison,Obscure InlineSetup,Overreferencing, Rotten Green Tests,and Verbose Test)

02

RQ3

perform a pairwise comparison between V_{ANILLA} and $S_{MELLESS}$
and assess whether the $S_{MELLESS}$ configuration generates tests that are as effective (in terms of coverage and mutation score) as those generated by the V_{ANILLA} configuration.

04

实验结果评估

Table III: Diffusion of test smells on the tests generated by the CONF-A configuration vs. CONF-B, CONF-C, and VANILLA configurations, on the 63 classes under test for which all configurations achieved similar coverage. Column $\bar{x}$ reports the ratio of smelly tests generated by each configuration. $\hat{A}_{12}$ reports the effect size of X vs. Y. Note that statistically significantly effect size values, i.e., p-value ≤ 0.05, are annotated in **bold face**. Column 'Rel. impr.' reports the relative improvement of X over Y regarding the percentage of smelly tests generated by both configurations.

	CONF-A	CONF-B			CONF-C			VANILLA		
Metric	$\bar{x}$	$\bar{x}$	$\hat{A}_{12}$	Rel. impr.	$\bar{x}$	$\hat{A}_{12}$	Rel. impr.	$\bar{x}$	$\hat{A}_{12}$	Rel. impr.
AssertionRoulette	0.90%	2.37%	0.44	163.47%	1.82%	0.43	102.10%	3.24%	0.44	260.29%
DuplicateAssert	1.71%	0.49%	0.60	-71.14%	0.88%	0.55	-48.90%	0.36%	0.62	-79.23%
EagerTest	51.47%	6.57%	**0.88**	-87.23%	20.51%	**0.87**	-60.15%	3.36%	**0.89**	-93.47%
IndirectTesting	84.98%	51.20%	**0.95**	-39.76%	56.08%	**0.95**	-34.00%	37.90%	**0.98**	-55.41%
LackOfCohesionOfMethods	0.00%	0.00%	**0.50**	0.00%	0.00%	**0.50**	0.00%	0.00%	**0.50**	0.00%
LazyTest	0.00%	0.00%	**0.50**	0.00%	0.00%	**0.50**	0.00%	0.00%	**0.50**	0.00%
LikelyIneffectiveObjectComparison	0.24%	0.05%	**0.51**	-77.18%	0.06%	**0.51**	-73.99%	0.03%	**0.51**	-85.97%
ObscureInlineSetup	84.44%	6.21%	**1.00**	-92.64%	32.06%	**0.99**	-62.03%	1.15%	**1.00**	-98.64%
Overreferencing	51.60%	18.99%	**0.94**	-63.20%	20.62%	**0.95**	-60.05%	5.16%	**0.97**	-89.99%
RedundantAssertion	1.74%	0.00%	**0.69**	-99.84%	1.04%	0.54	-40.07%	0.00%	**0.69**	-99.85%
RottenGreenTests	57.98%	7.08%	**0.97**	-87.79%	27.63%	**0.95**	-52.35%	1.51%	**0.98**	-97.40%
TestRedundancy	0.00%	0.00%	**0.50**	0.00%	0.00%	**0.50**	0.00%	0.00%	**0.50**	0.00%
UnknownTest	73.62%	51.17%	0.90	-30.49%	56.58%	**0.90**	-23.14%	46.42%	**0.92**	-36.95%
UnrelatedAssertions	9.55%	16.14%	0.30	68.93%	14.32%	0.31	49.95%	16.77%	0.31	75.57%
UnusedInputs	84.89%	42.19%	**0.95**	-50.31%	56.00%	**0.96**	-34.03%	29.22%	**0.96**	-65.59%
VerboseTest	91.00%	6.29%	**1.00**	-93.09%	37.31%	**1.00**	-59.00%	0.94%	**1.00**	-98.97%
Average	37.13%	13.05%	**0.73**	-35.02%	20.31%	**0.71**	-24.73%	9.13%	**0.74**	-35.35%

Table IV: Diffusion of test smells on the tests generated by the SMELLESS configuration. The rows highlighted in gray correspond to the smells metrics optimized by SMELLESS. Columns $\bar{x}$, standard deviation (σ), and confidence intervals (CI) using bootstrapping at 95% significance level, report the distribution of test smell metrics.

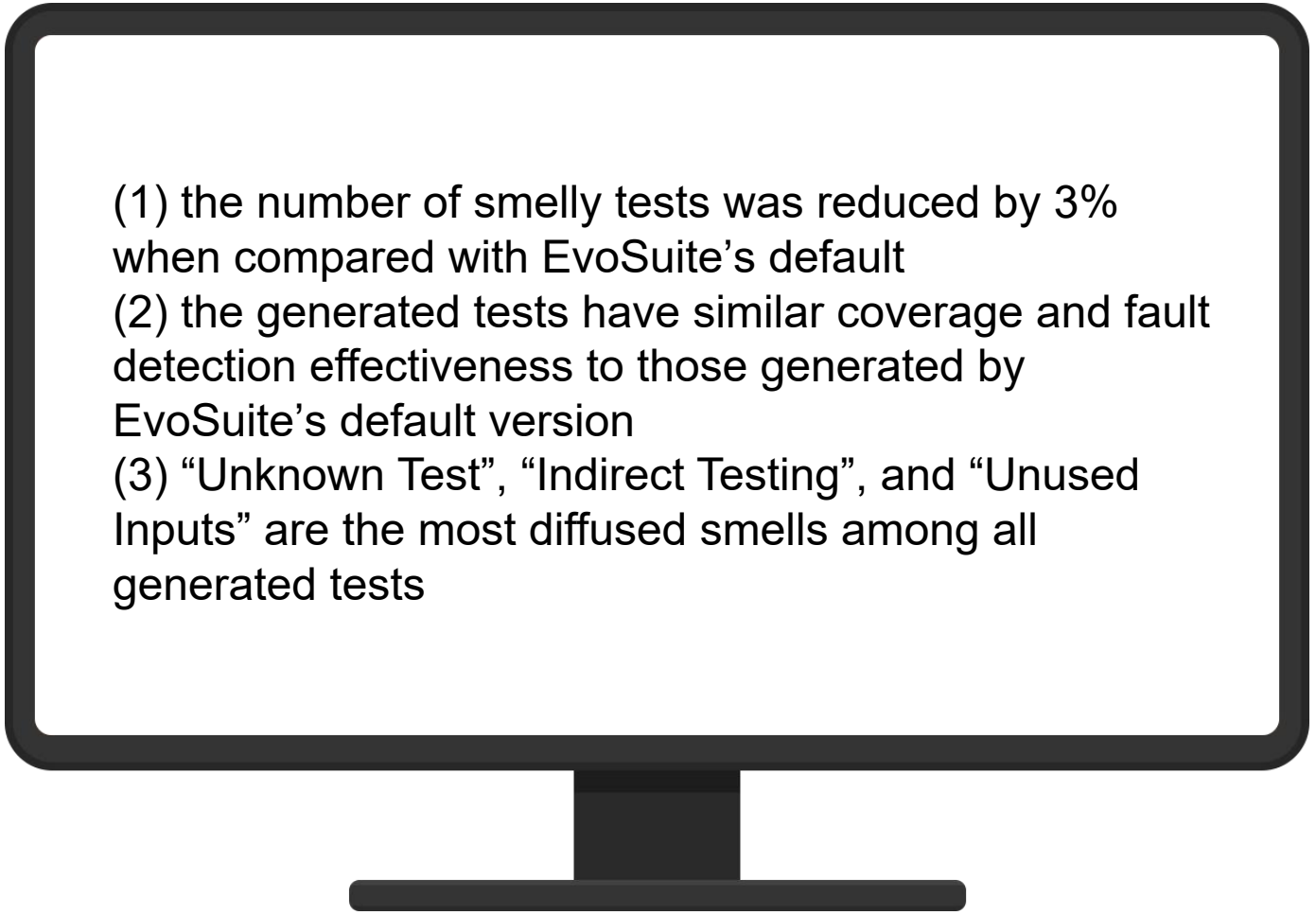
Metric	$\bar{x}$	σ	CI
AssertionRoulette	2.68%	0.09	[0.02, 0.04]
DuplicateAssert	0.62%	0.04	[0.00, 0.01]
EagerTest	3.98%	0.12	[0.03, 0.05]
IndirectTesting	33.05%	0.27	[0.30, 0.36]
LackOfCohesionOfMethods	0.33%	0.06	[-0.01, 0.01]
LazyTest	0.00%	0.00	[0.00, 0.00]
LikelyIneffectiveObjectComparison	0.01%	0.00	[0.00, 0.00]
ObscureInlineSetup	1.58%	0.05	[0.01, 0.02]
Overreferencing	4.69%	0.15	[0.03, 0.06]
RedundantAssertion	0.02%	0.00	[0.00, 0.00]
RottenGreenTests	0.78%	0.03	[0.00, 0.01]
TestRedundancy	0.00%	0.00	[0.00, 0.00]
UnknownTest	45.97%	0.27	[0.43, 0.49]
UnrelatedAssertions	16.31%	0.20	[0.14, 0.18]
UnusedInputs	25.76%	0.24	[0.23, 0.28]
VerboseTest	1.54%	0.06	[0.01, 0.02]
Average (optimized smells)	6.52%	0.10	[0.05, 0.07]
Average (all smells)	8.58%	0.10	[0.07, 0.10]

Table V: Diffusion of test smells on the tests generated by the VANILLA configuration vs. the SMELLESS configuration, on the 165 classes under test for which both configurations achieved similar coverage.

Metric	VANILLA	SMELLESS	$\hat{A}_{12}$	Rel. impr.
AssertionRoulette	2.00%	2.10%	0.49	5.26%
DuplicateAssert	0.26%	0.23%	0.50	-10.31%
EagerTest	3.58%	3.70%	0.49	3.39%
IndirectTesting	35.97%	34.37%	0.57	-4.44%
LackOfCohesionOfMethods	0.00%	0.00%	**0.50**	0.00%
LazyTest	0.00%	0.00%	**0.50**	0.00%
LikelyIneffectiveObjectComparison	0.02%	0.01%	**0.50**	-39.45%
ObscureInlineSetup	1.38%	1.64%	0.48	18.68%
Overreferencing	4.16%	3.63%	0.52	-12.66%
RedundantAssertion	0.03%	0.03%	**0.50**	-6.90%
RottenGreenTests	0.71%	0.68%	0.50	-3.97%
TestRedundancy	0.00%	0.00%	**0.50**	0.00%
UnknownTest	45.97%	45.95%	0.51	-0.03%
UnrelatedAssertions	17.39%	17.51%	0.49	0.68%
UnusedInputs	28.17%	27.76%	0.50	-1.43%
VerboseTest	1.55%	1.69%	0.49	9.46%
Average (optimized smells)	6.77%	6.53%	0.51	-4.14%
Average (all smells)	8.82%	8.71%	0.50	-2.61%

结论

conclusion



感谢聆听

汇报人　　李博诺