Replications Studies in Software Engineering Research

Tamer Abdou, Ph.D.

Industrial Engineering

Ryerson University

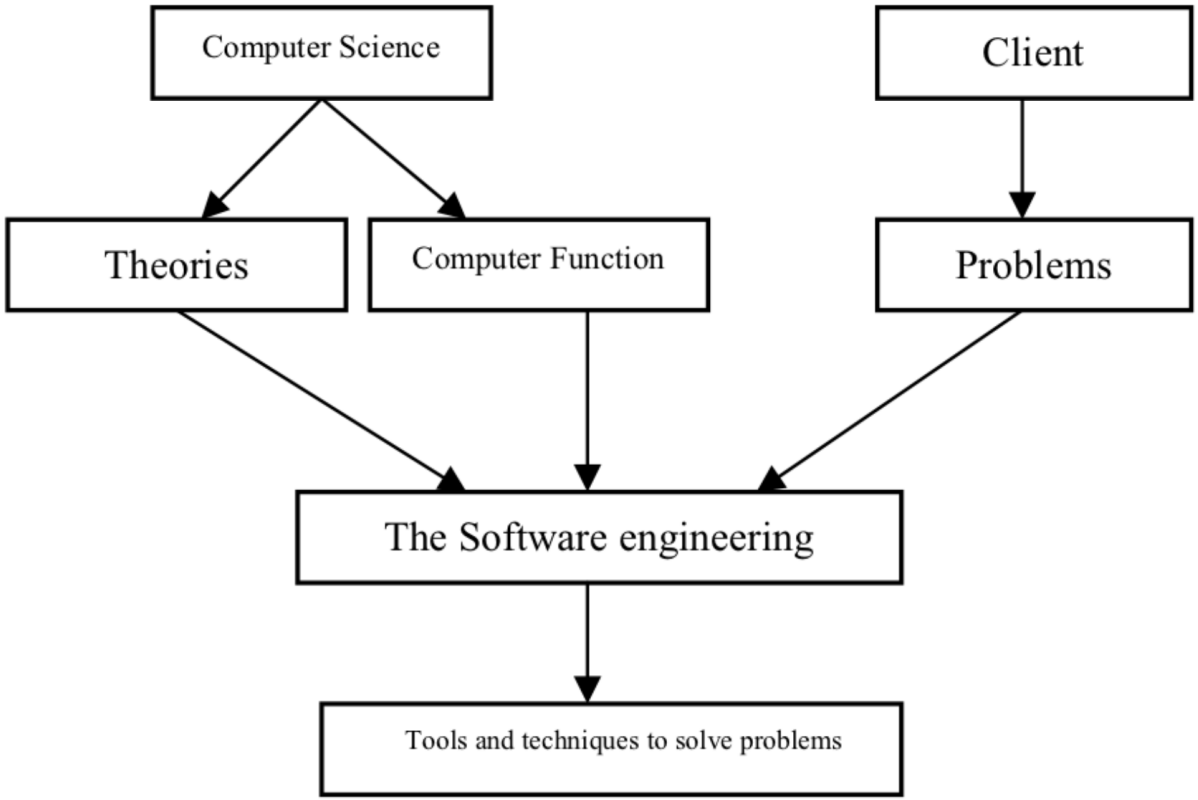
Toronto, Ontario

Canada

Wednesday, July 11, 2018

Outline

Software Engineering Conception



Mohammed, N., Munassar, A. and Govardhan, A. (2010) A Comparison Between Five Models Of Software Engineering, International Journal of Computer Science Issues, 7(5), pp. 94-101.

Replication Study

- A study that involves **sharing information/knowledge** so as to ensure **consistency** between **redundant resources**, such as software or hardware components[1].
- A study based on the **design**, **methodology** and **results** of **previously** published research papers[2].

1. https://en.wikipedia.org/wiki/Replication_(computing)
2. Capiluppi, A. and da Silva, F. Q. B. (2018) Guest Editors introduction to the special issue on replication studies in software engineering, Journal of Systems and Software, pp. 137-138.

Replication Study

- A study that involves **sharing information/knowledge** so as to ensure **consistency** between **redundant resources**, such as software or hardware components[1].
- A study based on the **design**, **methodology** and **results** of **previously** published research papers[2].

1. https://en.wikipedia.org/wiki/Replication_(computing)
2. Capiluppi, A. and da Silva, F. Q. B. (2018) Guest Editors introduction to the special issue on replication studies in software engineering, Journal of Systems and Software, pp. 137-138.

Do We Need to Replicate Studies in SE?

- Concerns about the **reliability** of empirical research results are fast becoming endemic and software engineering is no exception.
- **False discoveries** and how likely published experiments report erroneous results.
- Researchers questioned the **prevalence** of reported **p-values**
- Concerns about the **variability of results** depending upon which research **team performs the work**.
- Some studies are **selectively published** based on preferences for particular results.
- There is both a **low probability** of discovering a **true effect** and the parameter of interest has high variance.

Shepperd, M., Ajienka, N. and Counsell, S. (2018) The role and value of replication in empirical software engineering results, Information and Software Technology, 99, pp. 120-132.

Do We Need to Replicate Studies in SE?

- Concerns about the **reliability** of empirical research results are fast becoming endemic and software engineering is no exception.
- **False discoveries** and how likely published experiments report erroneous results.
- Researchers questioned the **prevalence** of reported **p-values**
- Concerns about the **variability of results** depending upon which research **team performs the work**.
- Some studies are **selectively published** based on preferences for particular results.
- There is both a **low probability** of discovering a **true effect** and the parameter of interest has high variance.

Shepperd, M., Ajienka, N. and Counsell, S. (2018) The role and value of replication in empirical software engineering results, Information and Software Technology, 99, pp. 120-132.

Do We Need to Replicate Studies in SE?

- Concerns about the **reliability** of empirical research results are fast becoming endemic and software engineering is no exception.
- **False discoveries** and how likely published experiments report erroneous results.
- Researchers questioned the **prevalence** of reported **p-values**
- Concerns about the **variability of results** depending upon which research **team performs the work**.
- Some studies are **selectively published** based on preferences for particular results.
- There is both a **low probability** of discovering a **true effect** and the parameter of interest has high variance.

Shepperd, M., Ajienka, N. and Counsell, S. (2018) The role and value of replication in empirical software engineering results, Information and Software Technology, 99, pp. 120-132.

Do We Need to Replicate Studies in SE?

- Concerns about the **reliability** of empirical research results are fast becoming endemic and software engineering is no exception.
- **False discoveries** and how likely published experiments report erroneous results.
- Researchers questioned the **prevalence** of reported **p-values**
- Concerns about the **variability of results** depending upon which research **team performs the work**.
- Some studies are **selectively published** based on preferences for particular results.
- There is both a **low probability** of discovering a **true effect** and the parameter of interest has high variance.

Shepperd, M., Ajienka, N. and Counsell, S. (2018) The role and value of replication in empirical software engineering results, Information and Software Technology, 99, pp. 120-132.

Do We Need to Replicate Studies in SE?

- Concerns about the **reliability** of empirical research results are fast becoming endemic and software engineering is no exception.
- **False discoveries** and how likely published experiments report erroneous results.
- Researchers questioned the **prevalence** of reported **p-values**
- Concerns about the **variability of results** depending upon which research **team performs the work**.
- Some studies are **selectively published** based on preferences for particular results.
- There is both a **low probability** of discovering a **true effect** and the parameter of interest has high variance.

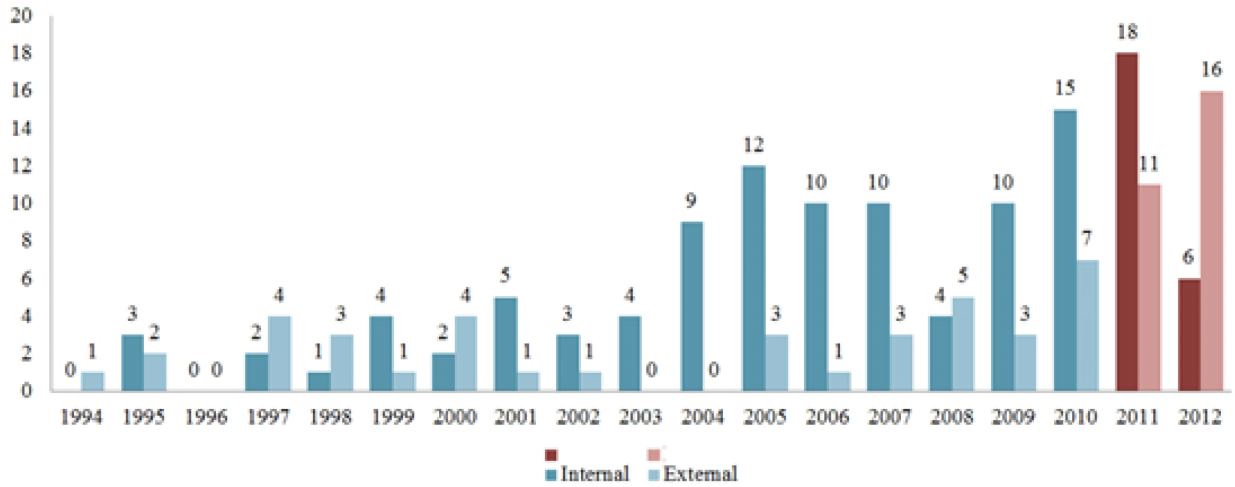
Shepperd, M., Ajienka, N. and Counsell, S. (2018) The role and value of replication in empirical software engineering results, Information and Software Technology, 99, pp. 120-132.

Do We Need to Replicate Studies in SE?

- Concerns about the **reliability** of empirical research results are fast becoming endemic and software engineering is no exception.
- **False discoveries** and how likely published experiments report erroneous results.
- Researchers questioned the **prevalence** of reported **p-values**
- Concerns about the **variability of results** depending upon which research **team performs the work**.
- Some studies are **selectively published** based on preferences for particular results.
- There is both a **low probability** of discovering a **true effect** and the parameter of interest has high variance.

Shepperd, M., Ajienka, N. and Counsell, S. (2018) The role and value of replication in empirical software engineering results, Information and Software Technology, 99, pp. 120-132.

Evolution of Replications over Years



1994-2003 An average of 4.1 studies published per year.

2004-2009 An average of 11.7 studies published per year.

2004-2012 An average of 24.3 studies published per year.

Bezerra, R. M. M., Da Silva, F. Q. B., Santana, A. M., Magalhes, C. V. C. and Santos, R. E. S. (2015) 'Replication of Empirical Studies in Software Engineering : An Update of a Systematic Mapping Study', in International Symposium on Empirical Software Engineering and Measurement, pp.1-4.

Topics of Interest

PROMISE'18 Replication and repeatability of previous work using predictive modelling and data analytics in software engineering
International conference on Predictive Models and Data Analytics in Software Engineering

JSS'18 Replication of empirical studies and families of studies.
Journal of Systems and Software

ESEM'18 Replication of software engineering studies.
Empirical Software Engineering and Measurement

Guidelines to follow

1. Information about the original study
2. Information about the replication
3. Comparison of results to original
4. Conclusion Across Studies

Carver, J. C. (2010) Towards Reporting Guidelines for Experimental Replications: A Proposal, in 1st International Workshop on Replication in Empirical Software Engineering. Cape Town, South Africa.

1- Research Questions

A description of the research question(s) that was the basis for the original design.

2- Participants

The number of participants and any relevant characteristics of the participants.

3- Design

A graphical (or textual) description of the experimental design.

4- Artifacts

A description of and/or links to the artifacts used.

5- Context variables

Any important context variables that affected the design of the study or interpretation of the results

6- Summary of results

A brief overview of the major findings

1- Motivation

a description of why the replication was conducted.

2- Level of interaction

The level of interaction the replicators had with the original experimenter should be reported.

3- Changes to the original experiment

Any changes made to the design, participants, artifacts, procedures, data collected and/or analysis techniques should be discussed here.

1- Similarities in results

Replication results that supported results from the original study.

2- Differences in results

Results from the replication that did not coincide with the results from the original study.

Final Conclusion

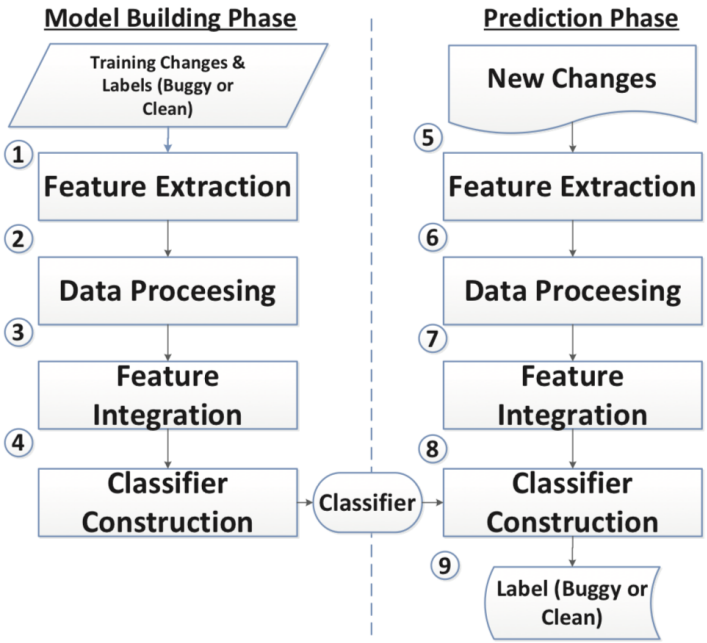
The authors should provide a discussion of the current state of knowledge.

1- Same experiment & Same objects

Goal: Evaluating the certainty of current knowledge (i.e., confirming or disputing previous results).

1- Same experiment & Same objects

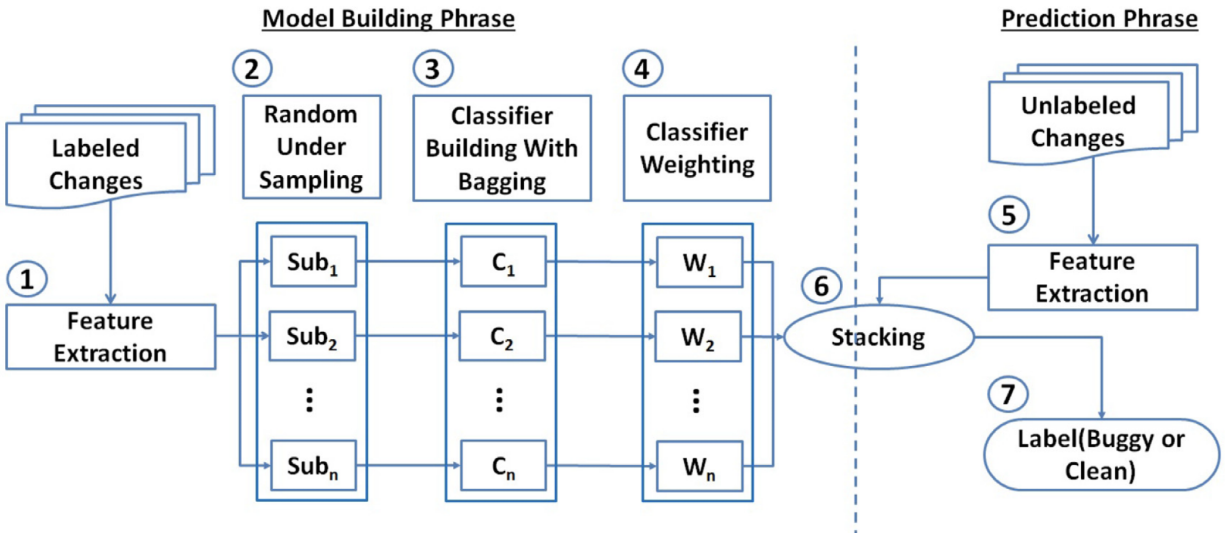
Example (Original Study - Deeper Model)



Yang, X., Lo, D., Xia, X., Zhang, Y. and Sun, J. (2015) Deep Learning for Just-in-Time Defect Prediction, in Proceedings - 2015 IEEE International Conference on Software Quality, Reliability and Security, QRS 2015, pp. 1726.

1- Same experiment & Same objects

Example (Original Study - TLEL Model)



Yang, X., Lo, D., Xia, X. and Sun, J. (2017) TLEL: A Two-layer Ensemble Learning Approach for Just-in-time Defect Prediction, in Information and Software Technology, pp. 206-220.

1- Same experiment & Same objects

Example (Same Datasets)

	Period	The total number of changes		Average LOC File	Average LOC Change	# of modified files per change	# of changes per day	# dev. per file Max	# dev. per file Avg
Bugzilla	08/1998 - 12/2006	4,620	(36%)	389.8	37.5	2.3	1.5	37	8.4
Columba	11/2002 - 07/2006	4,455	(31%)	125.0	149.4	6.2	3.3	10	1.6
Eclipse JDT	05/2001 - 12/2007	35,386	(14%)	260.1	71.4	4.3	14.7	19	4.0
Eclipse Platform	05/2001 - 12/2007	64,250	(14%)	231.6	72.2	4.3	26.7	28	2.8
Mozilla	01/2000 - 12/2006	98,275	(5%)	360.2	106.5	5.3	38.9	155	6.4
PostgreSQL	07/1996 - 05/2010	20,431	(25%)	563.0	101.3	4.5	4.0	20	4.0
OSS-Median	-	27,909	(20%)	310.1	86.7	4.4	9.4	24	4.0
C-1	10/2000 - 12/2009	4,096		-	16.4	2.0	1.2	-	-
C-2	10/2000 - 12/2009	9,277		-	19.2	2.4	2.8	-	-
C-3	07/2002 - 12/2009	3,586		-	16.6	2.0	1.3	-	-
C-4	12/2003 - 12/2009	5,182		-	12.9	1.8	2.4	-	-
C-5	10/1982 - 12/1995	10,961		303.0	39.0	4.8	2.3		-
COM-Median	-	5,182		-	16.6	2.0	2.3	-	-

Kamei, Y., Shihab, E., Adams, B., Hassan, A. E., Mockus, A., Sinha, A. and Ubayashi, N. (2013) A Large-scale Empirical Study of Just-in-time Quality Assurance, IEEE Transactions on Software Engineering, 39(6), pp. 757773.

1- Same experiment & Same objects

Example 1 (Replication Study)

Project	Deeper Original	Deeper Replicated	TLEL Original	TLEL Replicated	DSL
Bugzilla	0.6292	0.6348	0.6850	0.6722	**0.6730**
Columba	0.5606	0.5641	0.6065	0.6050	**0.6090**
JDT	0.3779	0.3762	0.4194	0.4125	**0.4233**
Mozilla	0.2215	0.2127	0.2625	0.2561	**0.2582**
Platform	0.3822	0.3910	0.4471	0.4381	**0.4425**
PostgreSQL	0.5509	0.5485	0.6052	0.5958	**0.5994**
Average	0.4537	0.4546	0.5043	0.4966	**0.5009**

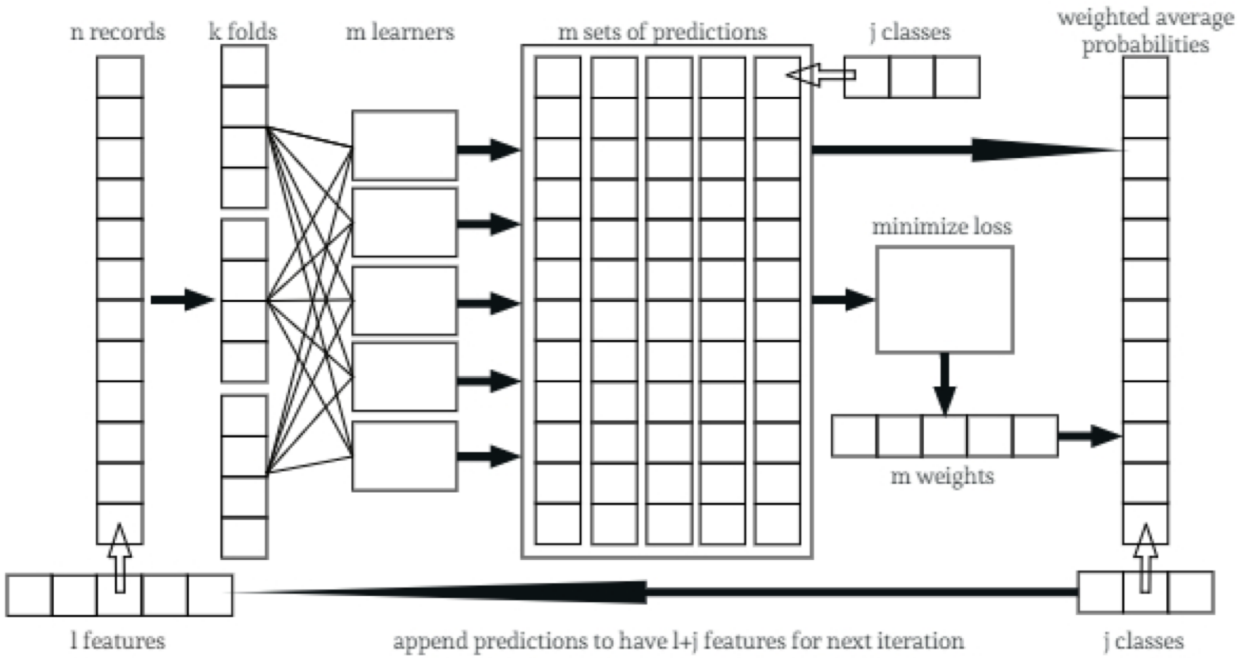
Young, S., Abdou, T. and Bener, A. (2018) A Replication Study: Just-In-Time Defect Prediction with Ensemble Learning, 40th International Conference on Software Engineering (ICSE2018) 6th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE2018), pp. 42-47.

2- Different experiment & Same objects

Goal: Improving the original model and reduce the internal threats to validity (minimize systematic error)

2- Different experiment & Same objects

Example (Different Model - Same Datasets)



Young, S., Abdou, T. and Bener, A. (2018) Deep Super Learner: A Deep Ensemble for Classification Problems, in Proceedings of the 31st Canadian Conference on Artificial Intelligence (CanadianAI-31), pp. 84-95.

2- Different experiment & Same objects

Example (Replication Study - Same Datasets)

Project	Deeper Original	Deeper Replicated	TLEL Original	TLEL Replicated	DSL
Bugzilla	0.6292	0.6348	0.6850	0.6722	**0.6730**
Columba	0.5606	0.5641	0.6065	0.6050	**0.6090**
JDT	0.3779	0.3762	0.4194	0.4125	**0.4233**
Mozilla	0.2215	0.2127	0.2625	0.2561	**0.2582**
Platform	0.3822	0.3910	0.4471	0.4381	**0.4425**
PostgreSQL	0.5509	0.5485	0.6052	0.5958	**0.5994**
Average	0.4537	0.4546	0.5043	0.4966	**0.5009**

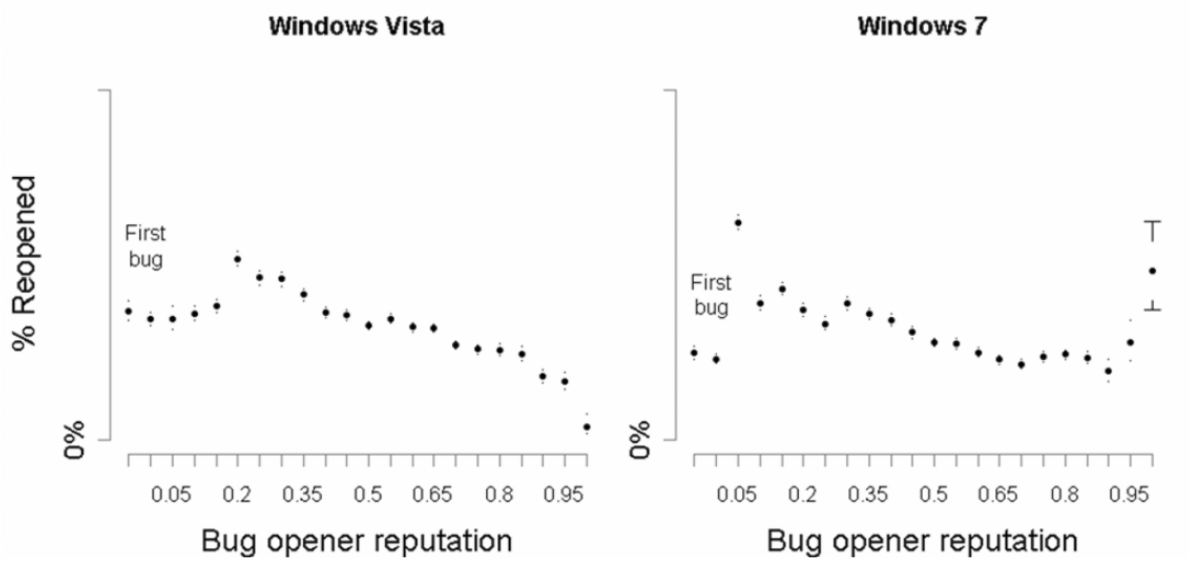
Young, S., Abdou, T. and Bener, A. (2018) A Replication Study: Just-In-Time Defect Prediction with Ensemble Learning, 40th International Conference on Software Engineering (ICSE2018) 6th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE2018), pp. 42-47.

3- Same experiment & Different objects

Goal: Identifying limitations to the generality of the conclusions (or to problems with the objects).

3- Same experiment & Different objects

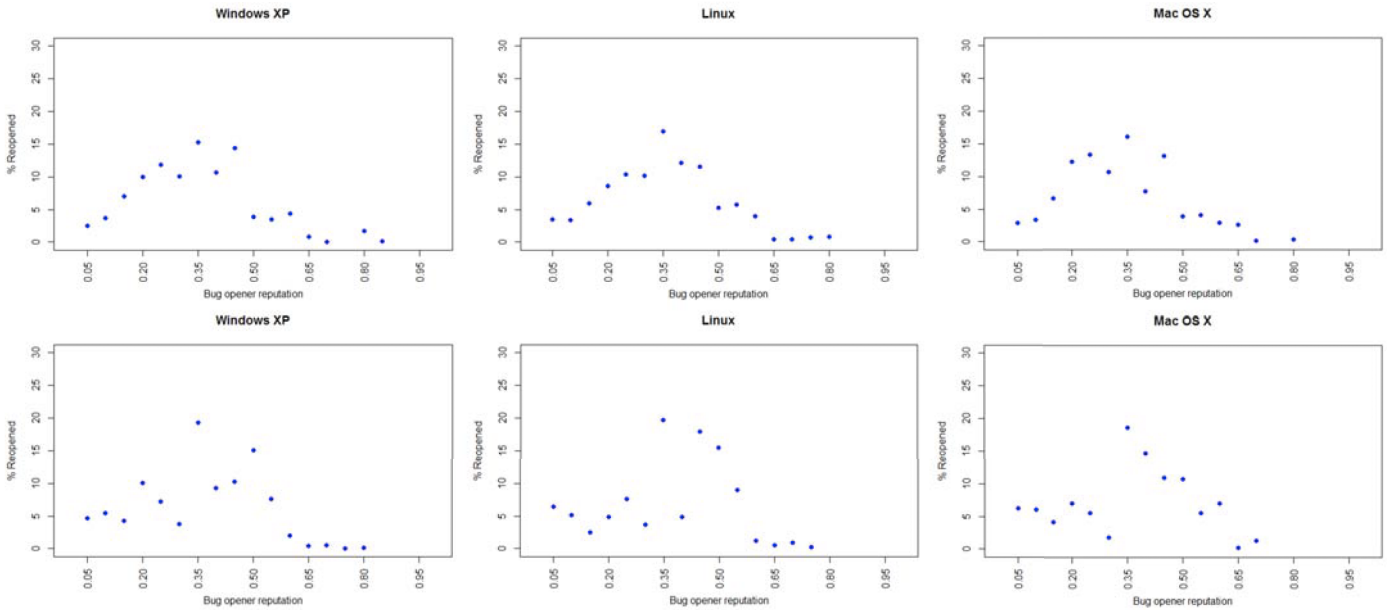
Example (Original Study)



Zimmermann, T., Nagappan, N., Guo, P. J. and Murphy, B. (2012) Characterizing and Predicting Which Bugs Get Reopened, in Proceedings of the 34th International Conference on Software Engineering. Piscataway, NJ, USA: IEEE Press (ICSE 12), pp. 1074-1083.

3- Same experiment & Different objects

Example (Replication Study - Different Datasets)



Abdou, T. and Bener, A. (2016) A Replication Study: Where and When Should Defects be Re-Assigned, in Software Engineering and Advanced Applications (SEAA), 2016 42th Euromicro Conference on, pp. 368-371.

Thank you!


Software
results
Information
International
Concerns
Counsell
high
120-132
Intelligence
Empirical
variance
Ajienka
Researchers
study
work
exception
Different
IEEE
depending
experiments
particular
probability
fast
Prediction
Proceedings
Artificial
per
Guest
Original
introduction
Synergies
Bener
Guidelines
Conclusion
description
team
Ensemble
objects
parameter
p-values
Comparison
based
Conference
prevalence
published
research
software
Engineering
replication
Study
original
reliability
preferences
Technology
Abdou
design
Defect
average
Deep
interest
Journal
participants
Datasets
Editors
Silva
Model
empirical
studies
pp
Replication
variability
effect
performs
experiment
Young
Workshop
Quality
Issue
Learning
Systems
RAISE2018
Capiluppi
Studies
questioned
likely
low
reported
discovering
erroneous
selectively
discoveries
endemic
Across
Just-In-Time
current
Example
False
true
role
value
report
Shepperd
ICSE2018
year
becoming
hardware
Xia
engineering
Just-in-time
papers2
https://en.wikipedia.org/wiki/Replication

Wait ... Did you see this!

Shepperd, M. (2018) Replication studies considered harmful, in 40th International Conference on Software Engineering (ICSE2018) New Ideas and Emerging Results (NIER 2018), pp. 73-76.