Automated Detection of Performance Regressions Using Regression Models on Clustered Performance Counters

Weiyi Shang, Ahmed E. Hassan
Queen's University, Kingston, Canada

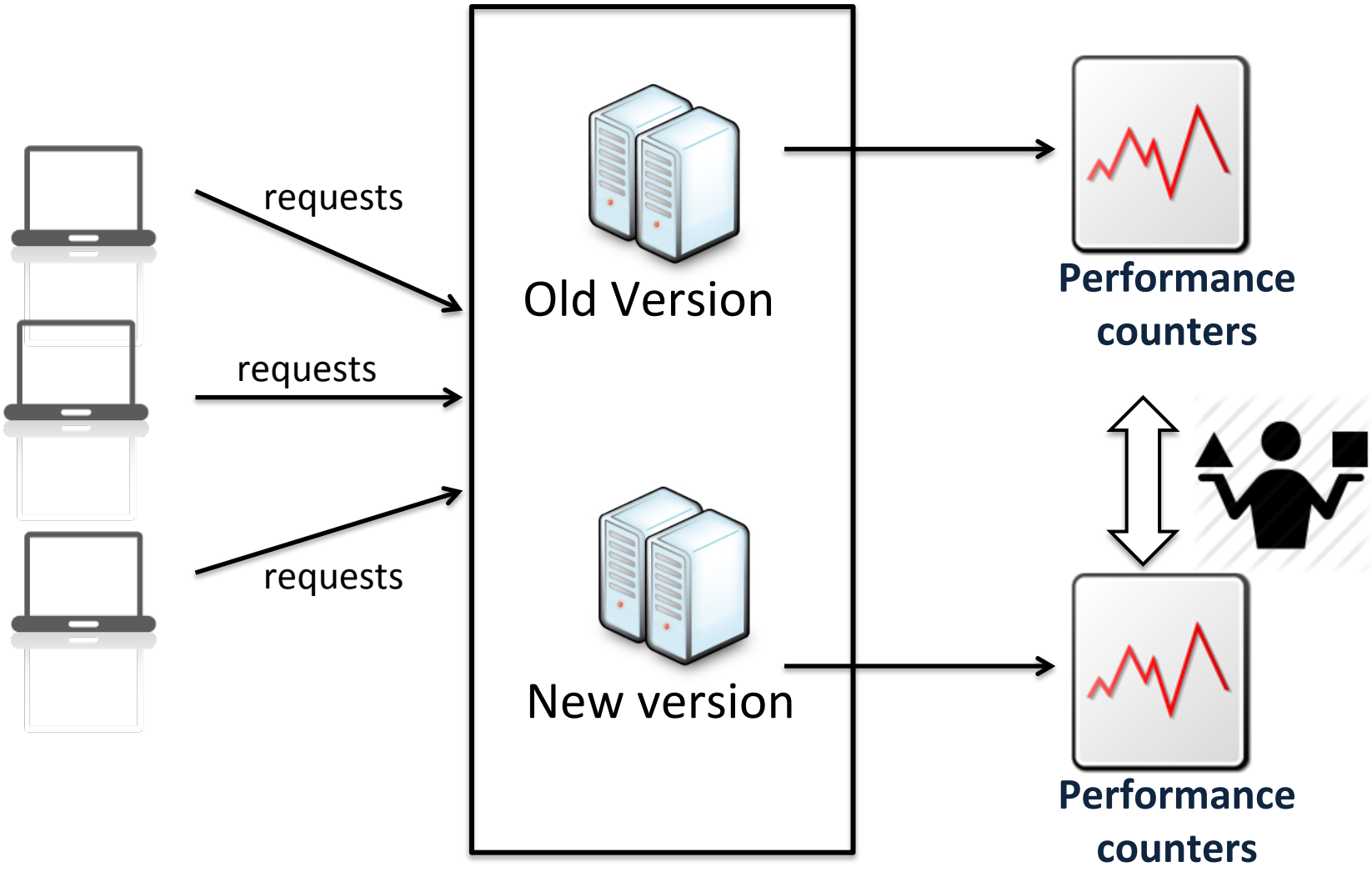
Mohamed Nasser, Parminder Flora
BlackBerry, Waterloo, Canada

What is a performance regression?

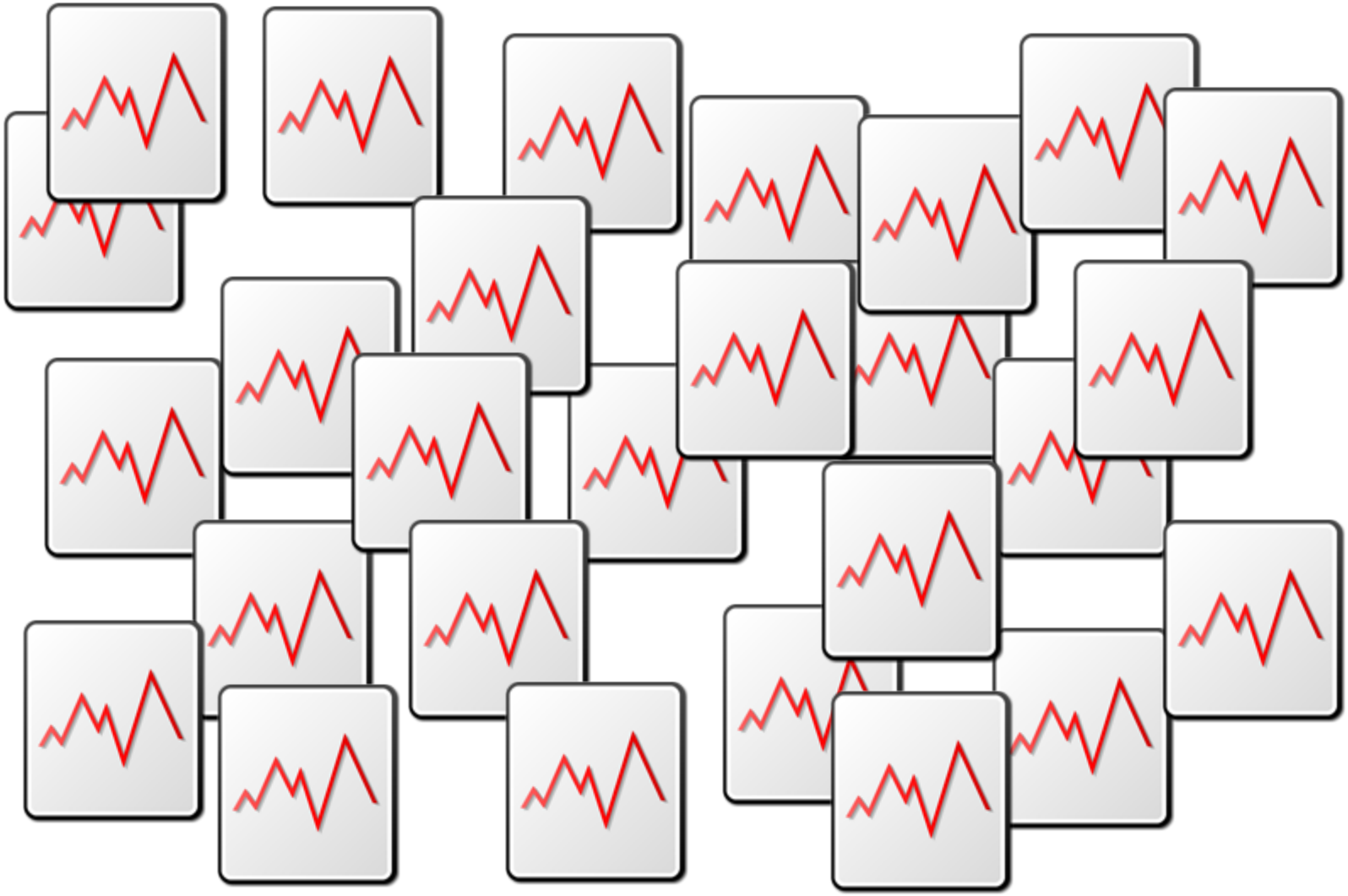
Old rsion

Does the new version have worse performance than the old version?

How to detect performance regression?

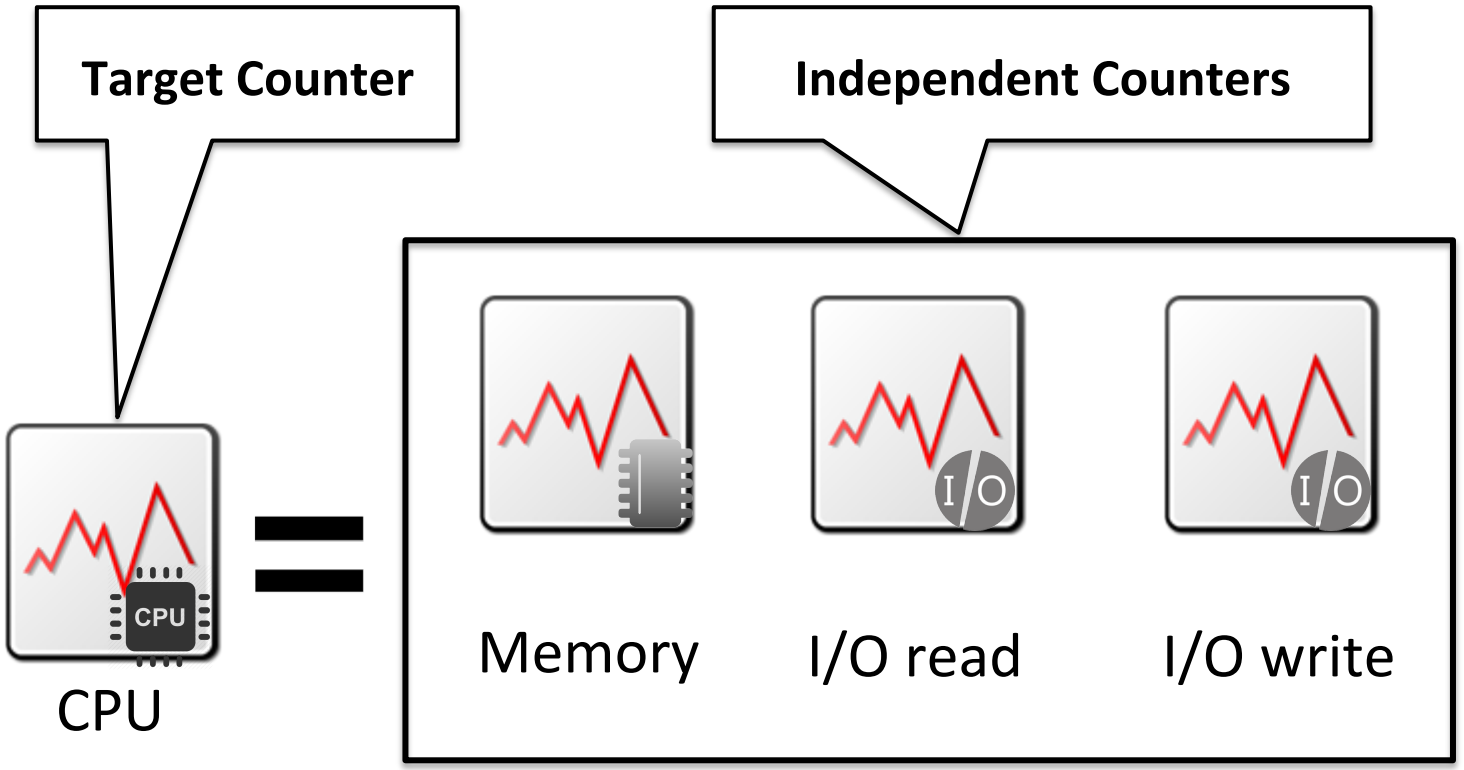


Large software systems generate large amounts of performance counters

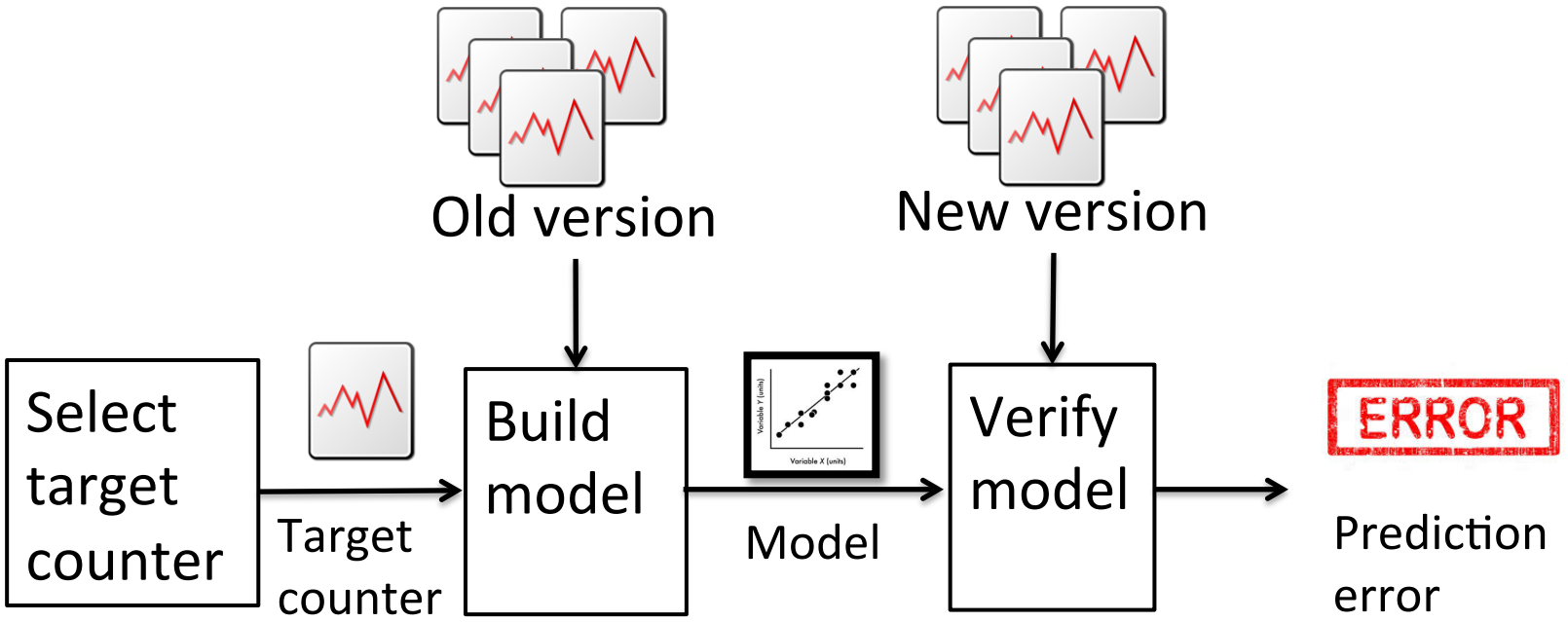


Performance engineers pick counters to compare using T-test

Performance engineers build a model using all counters

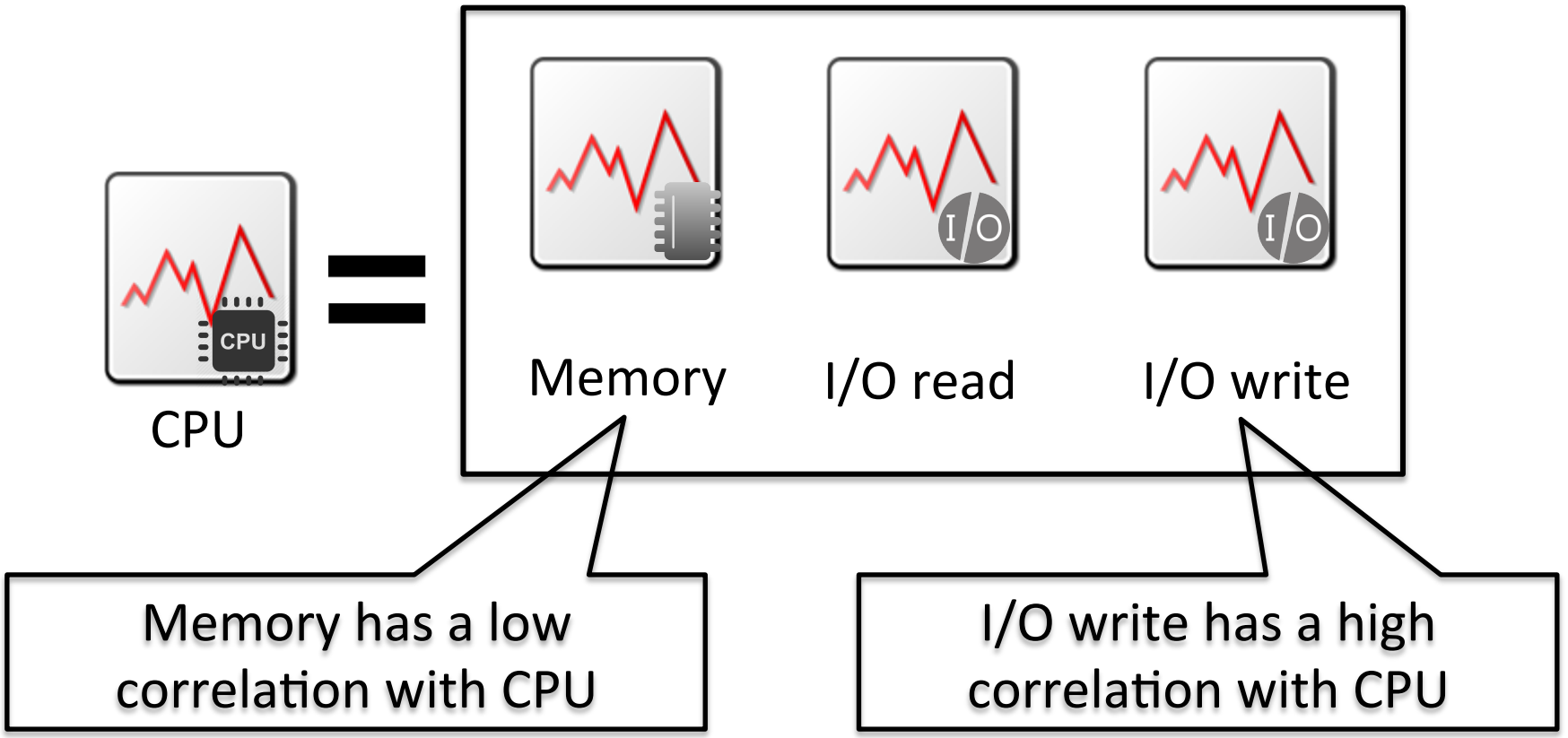


Model-based performance regression detection



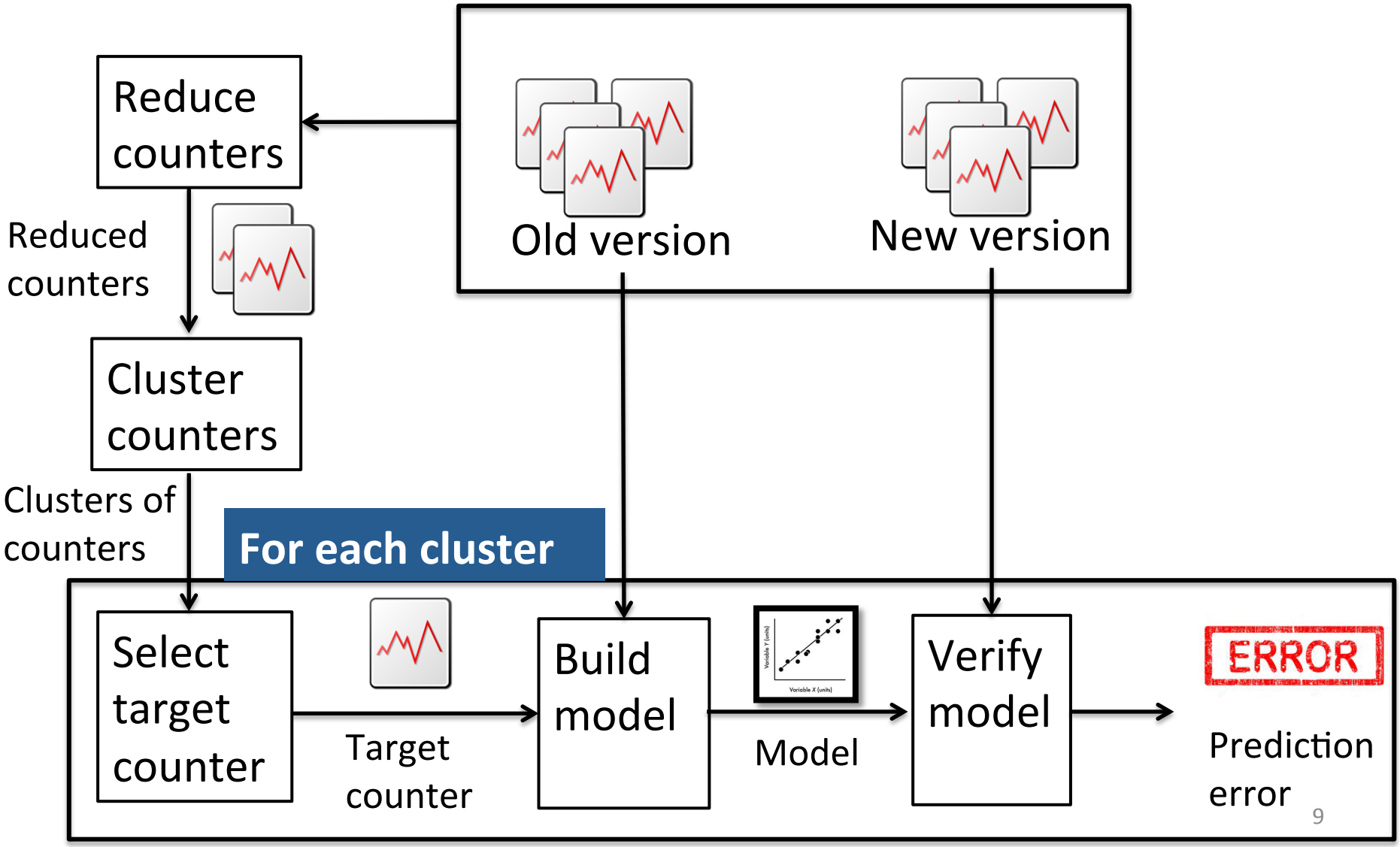
Target counter is selected by experience!

Selecting a wrong target counter will fail to detect performance regression

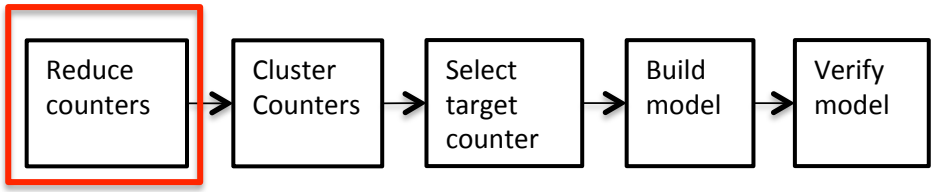


Memory-related performance regressions will not be detected by this model!

Our approach



Step 1: Reduce counters



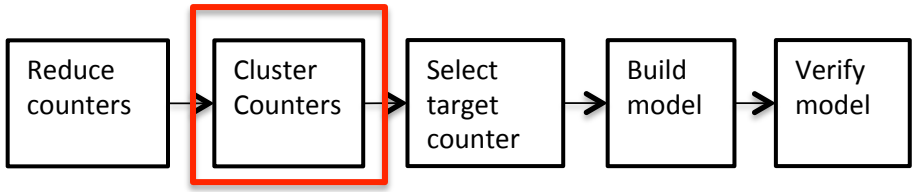
Removing zero variance counters

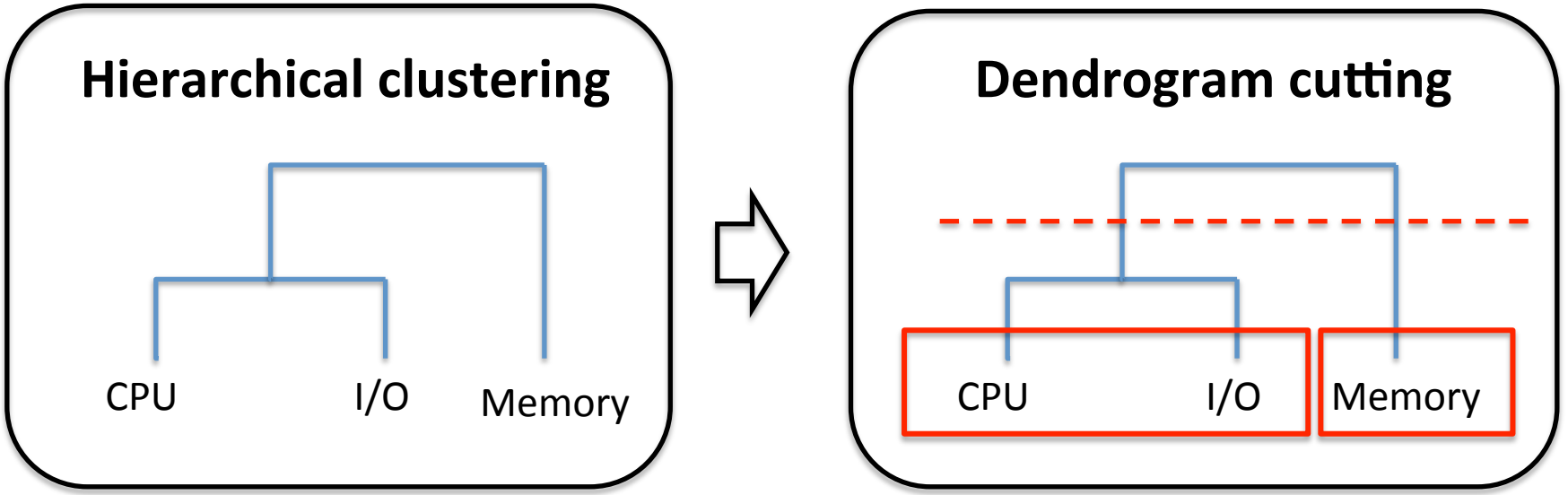
Available CPU cores:
4, 4, 4, …4, 4

Removing redundant counters

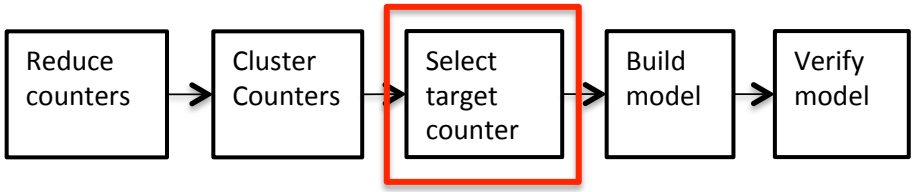
I/O write byte/sec=
a*I/O write op/sec +
b*CPU

Step 2: Cluster counters

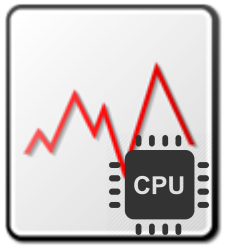




Step 3: Select target counter

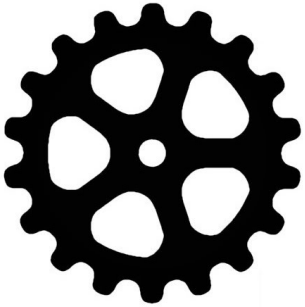


We select the counter that has significant difference between two versions with highest certainty.



CPU from old version

CPU from new version

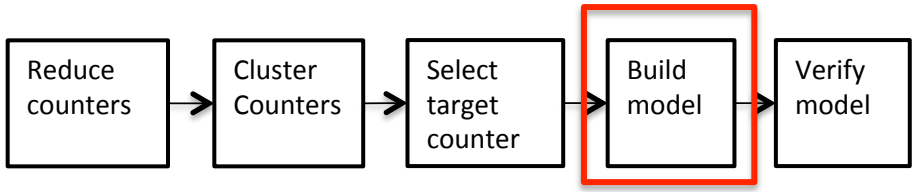


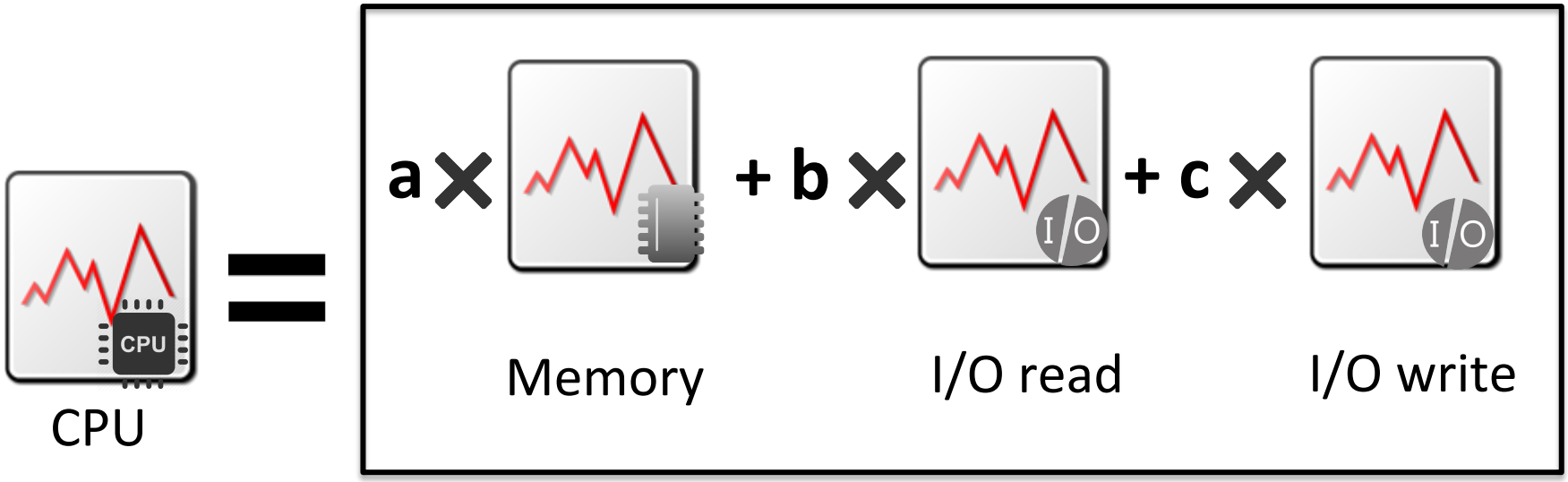
Kolmogorov–Smirnov test

P-VALUE

We select the counter with smallest p-value

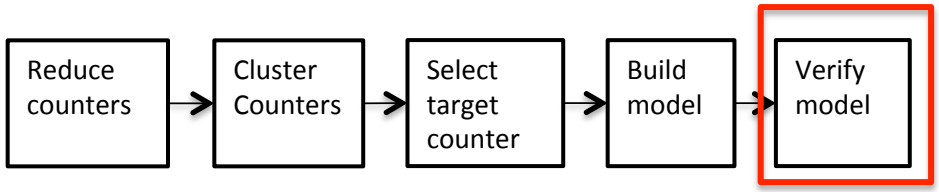
Step 3: Build model

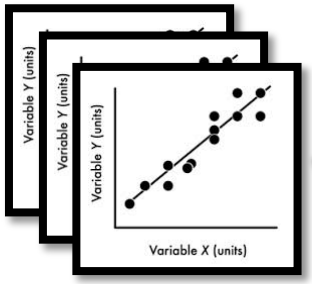




We build a linear regression model.

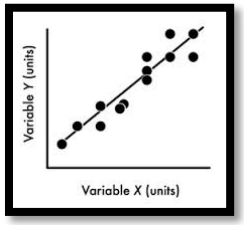
Step 4: Verify model





Linear regression models for each cluster

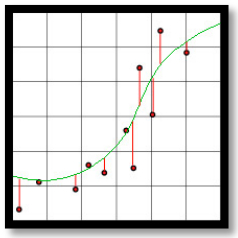
New version counters left in cluster 1



Linear regression model for cluster 1

...

Calculate prediction error



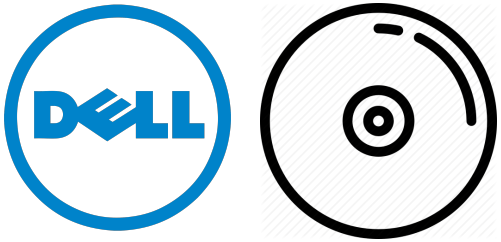
Mean prediction error

< threshold

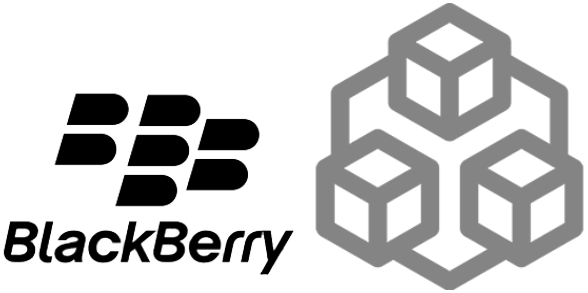
> threshold

...

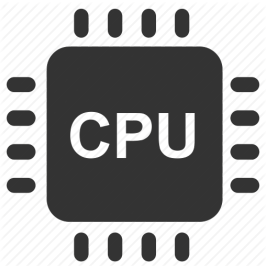
Case study: subject systems



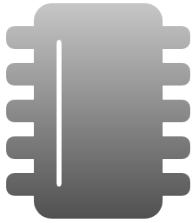
DELL DVD Store



Enterprise Application



CPU overhead



Memory overhead

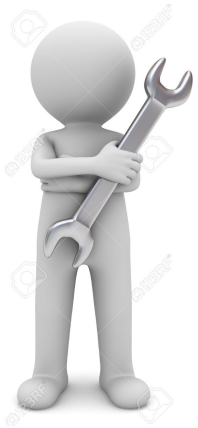
I/O overhead

Removing text index

Removing column index

Real-life performance regression

Applicability

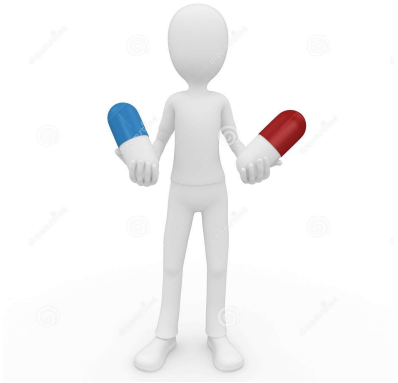


How many target counters does our approach pick?

Accuracy

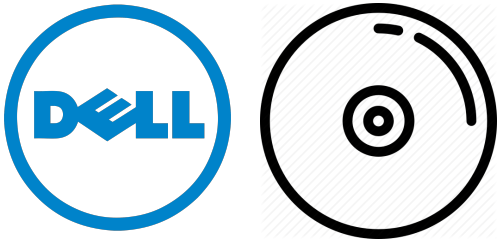
Can our approach detect performance regressions?

Comparison



Is our approach better than traditional approaches?

Our approach picks a small number of target counters



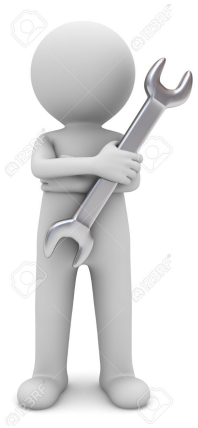
DELL DVD Store

Our approach picks 2 to 4 target counters.

Picked target counters are different across runs of our approach.

Performance engineers cannot pick target counters based on experience

Applicability



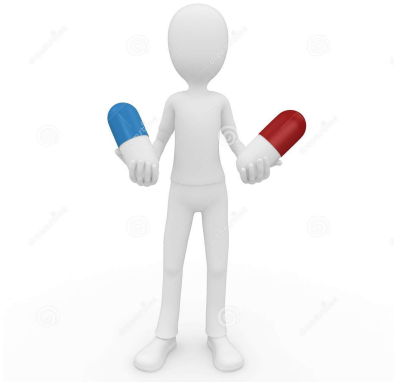
How many target counters does our approach pick?

Our approach picks a small number of target counters.

Accuracy

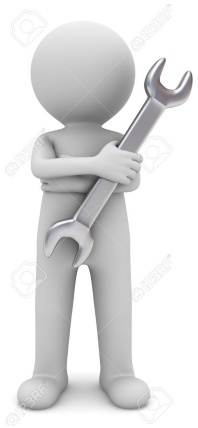
Can our approach detect performance regressions?

Comparison



Is our approach better than traditional approaches?

Applicability



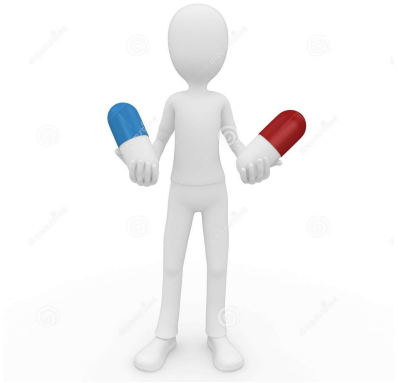
How many target counters does our approach pick?

Our approach picks a small number of target counters.

Accuracy

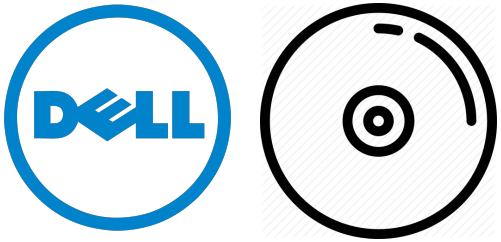
Can our approach detect performance regressions?

Comparison

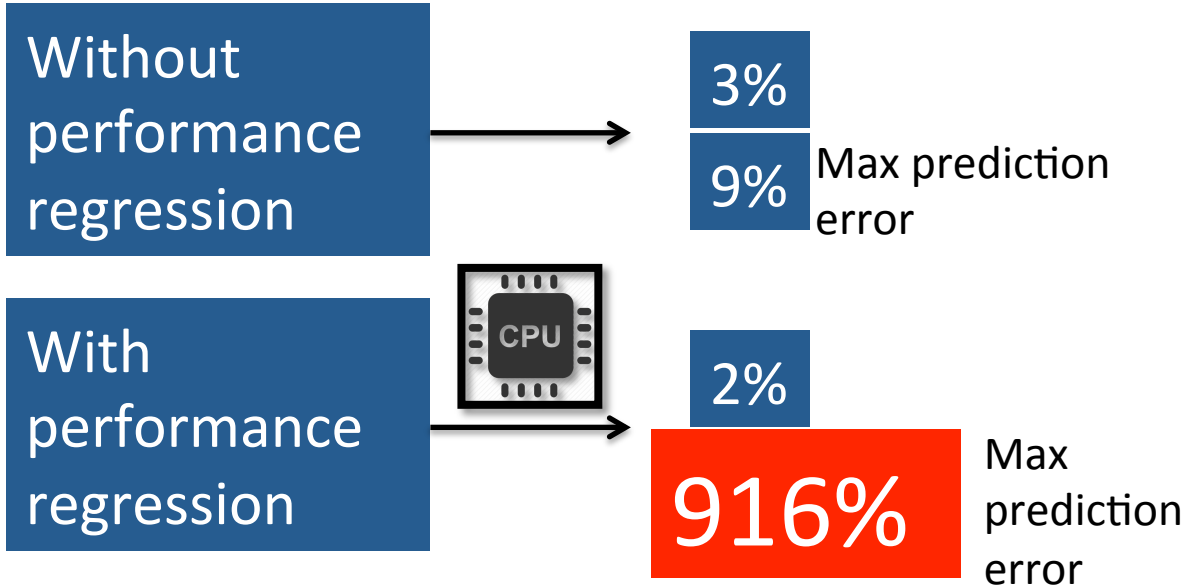


Is our approach better than traditional approaches?

Our approach can detect performance regressions



DELL DVD Store



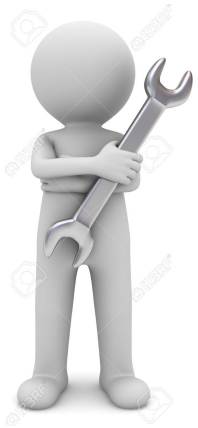
Max prediction error

4% to 11% without regression

24% to 1622% with regression

Our approach is not heavily impacted by the choice of threshold value.

Applicability



How many target counters does our approach pick?

Our approach picks a small number of target counters.

Accuracy

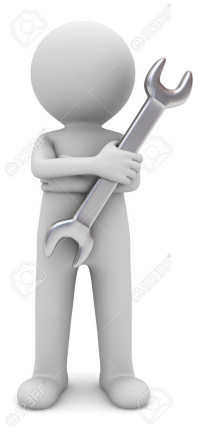
Can our approach detect performance regressions?

Our approach can detect performance regressions and is not heavily impacted by threshold value.

Comparison

Is our approach better than traditional approaches?

Applicability



How many target counters does our approach pick?

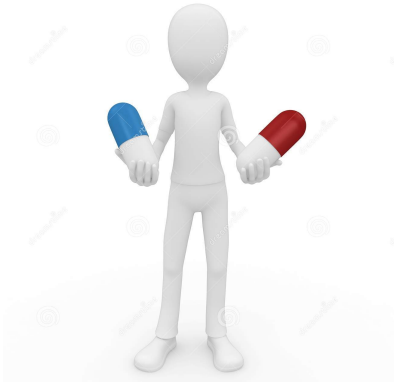
Our approach picks a small number of target counters.

Accuracy

Can our approach detect performance regressions?

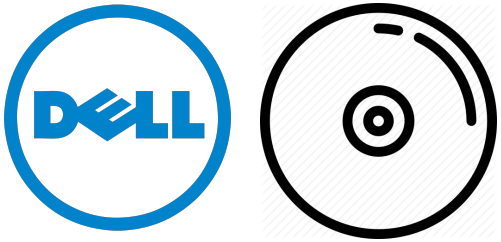
Our approach can detect performance regressions and is not heavily impacted by threshold value.

Comparison

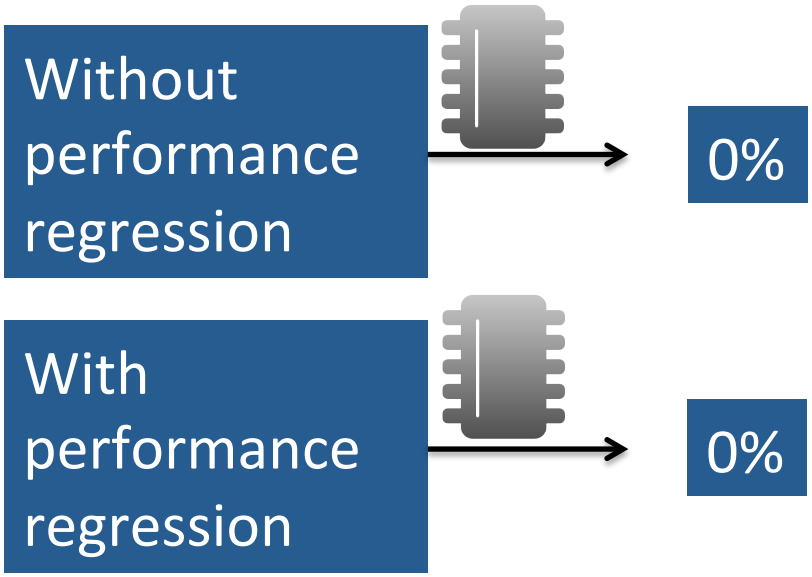


Is our approach better than traditional approaches?

Our approach outperforms picking one target counter to build a model

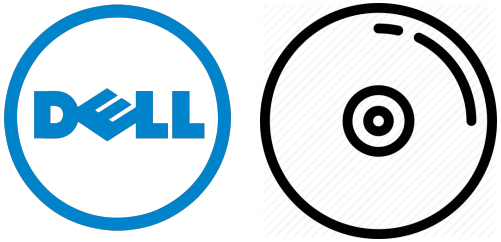


DELL DVD Store



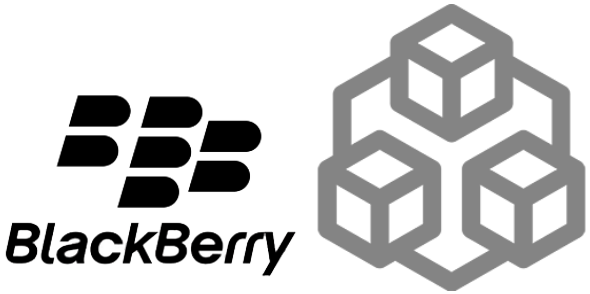
Building one model using memory as target counter may fail to detect performance regression

T-test does not perform well in detecting performance regressions.



DELL DVD Store

Without performance regression → 3 to 6 counters are flagged

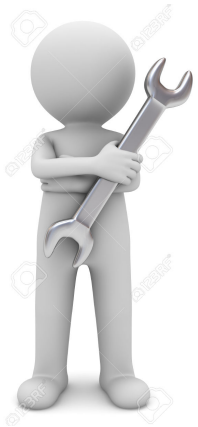


Enterprise Application

Without performance regression → 32% counters are flagged

There are a large number of counters with significant differences in the T-test results even though no regressions exist.

Applicability



How many target counters does our approach pick?

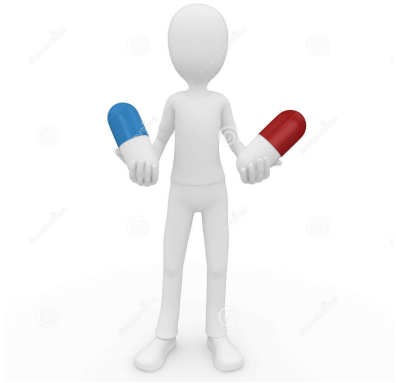
Our approach picks a small number of target counters.

Accuracy

Can our approach detect performance regressions?

Our approach can detect performance regressions and is not heavily impacted by threshold value.

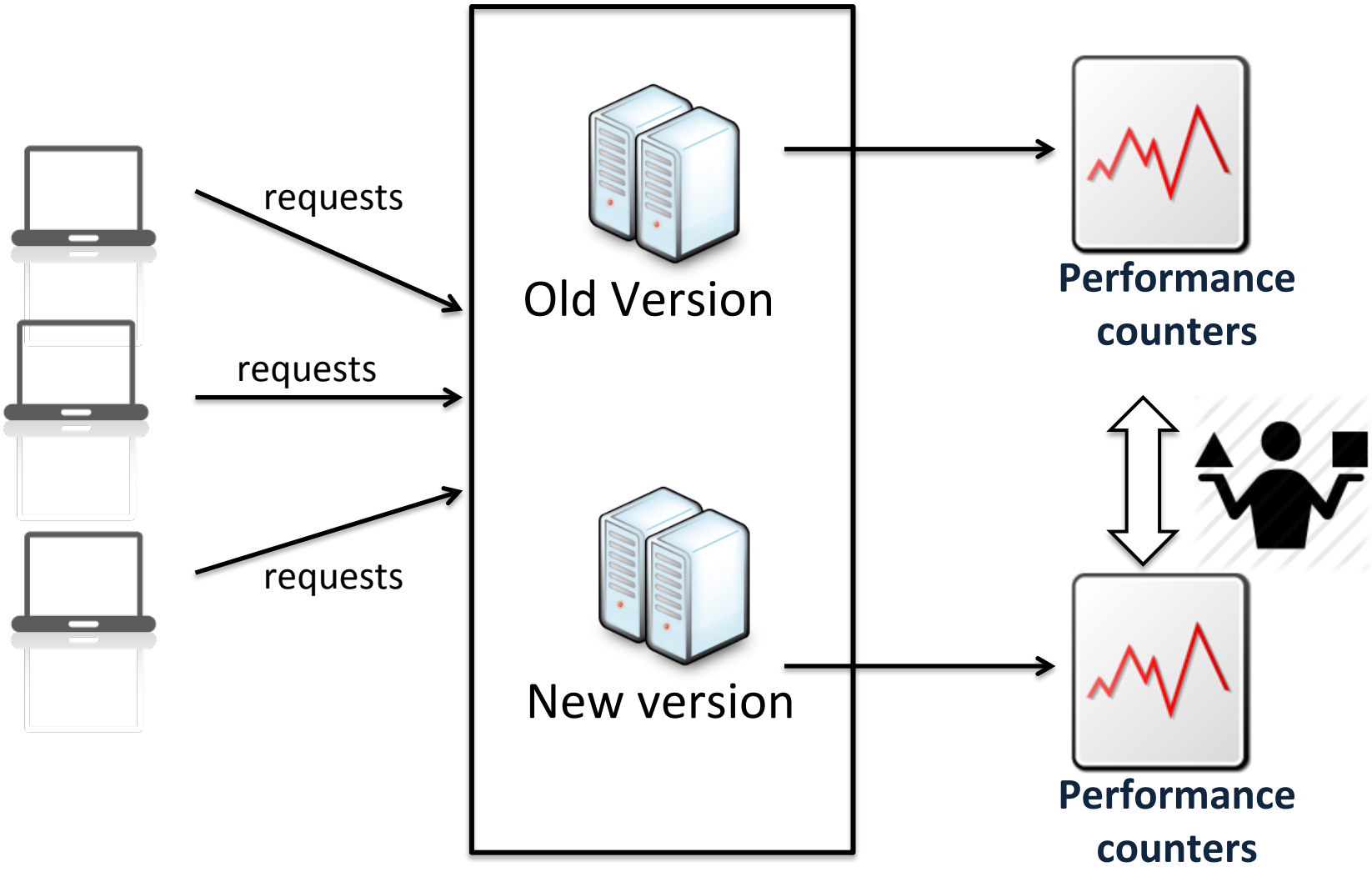
Comparison



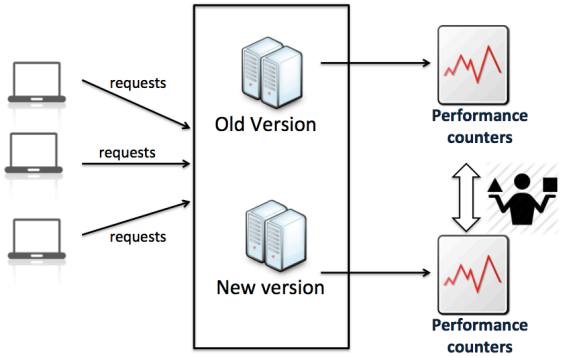
Is our approach better than traditional approaches?

Our approach out performs traditional approaches.

How to detect performance regression?

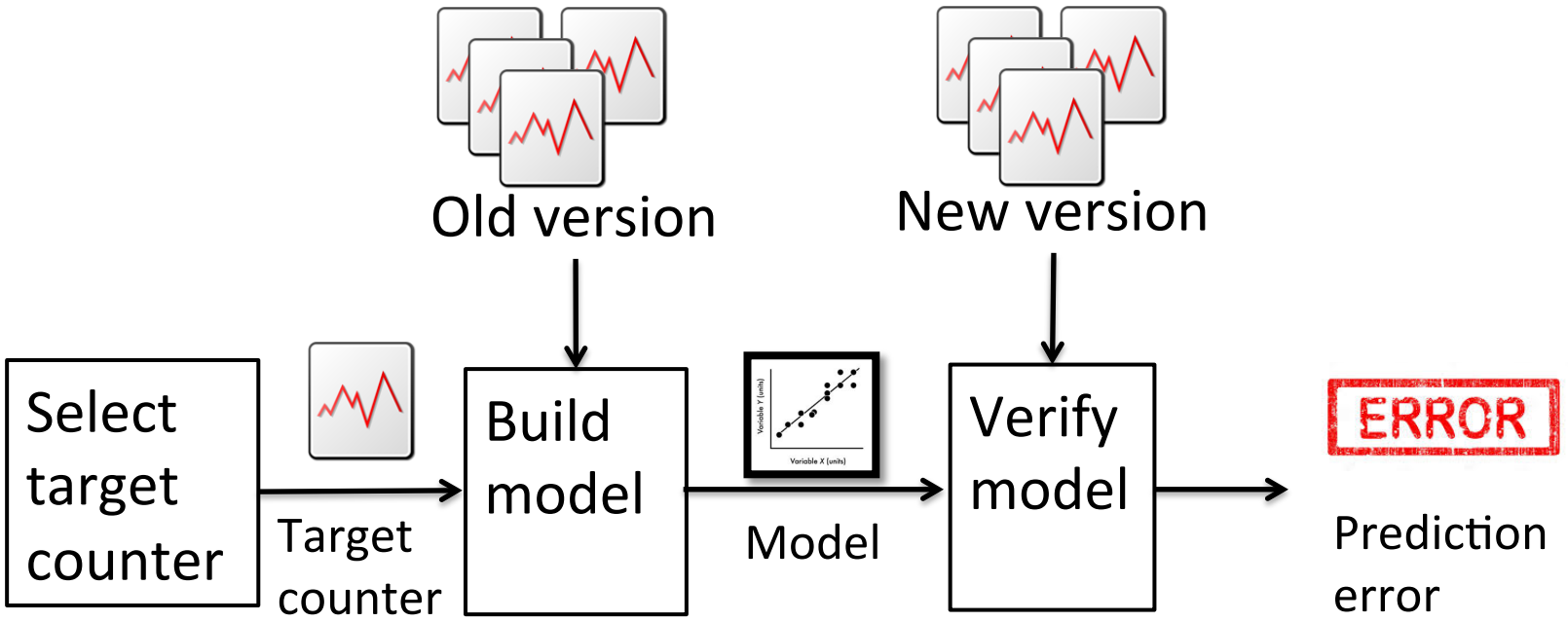


How to detect performance regression?



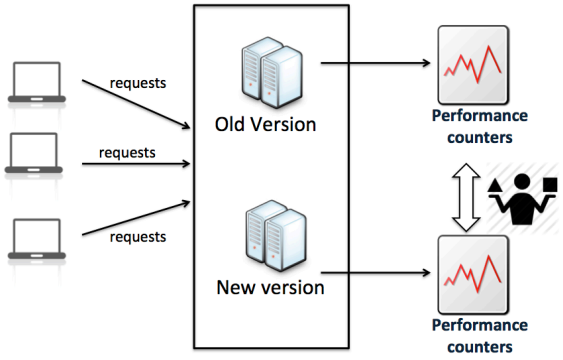
19

Model-based performance regression detection



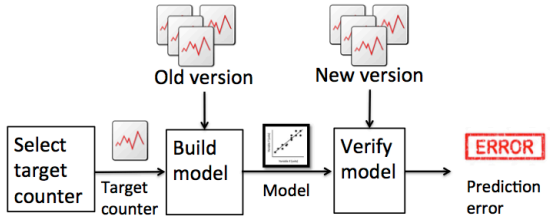
Target counter is selected by experience!

How to detect performance regression?



19

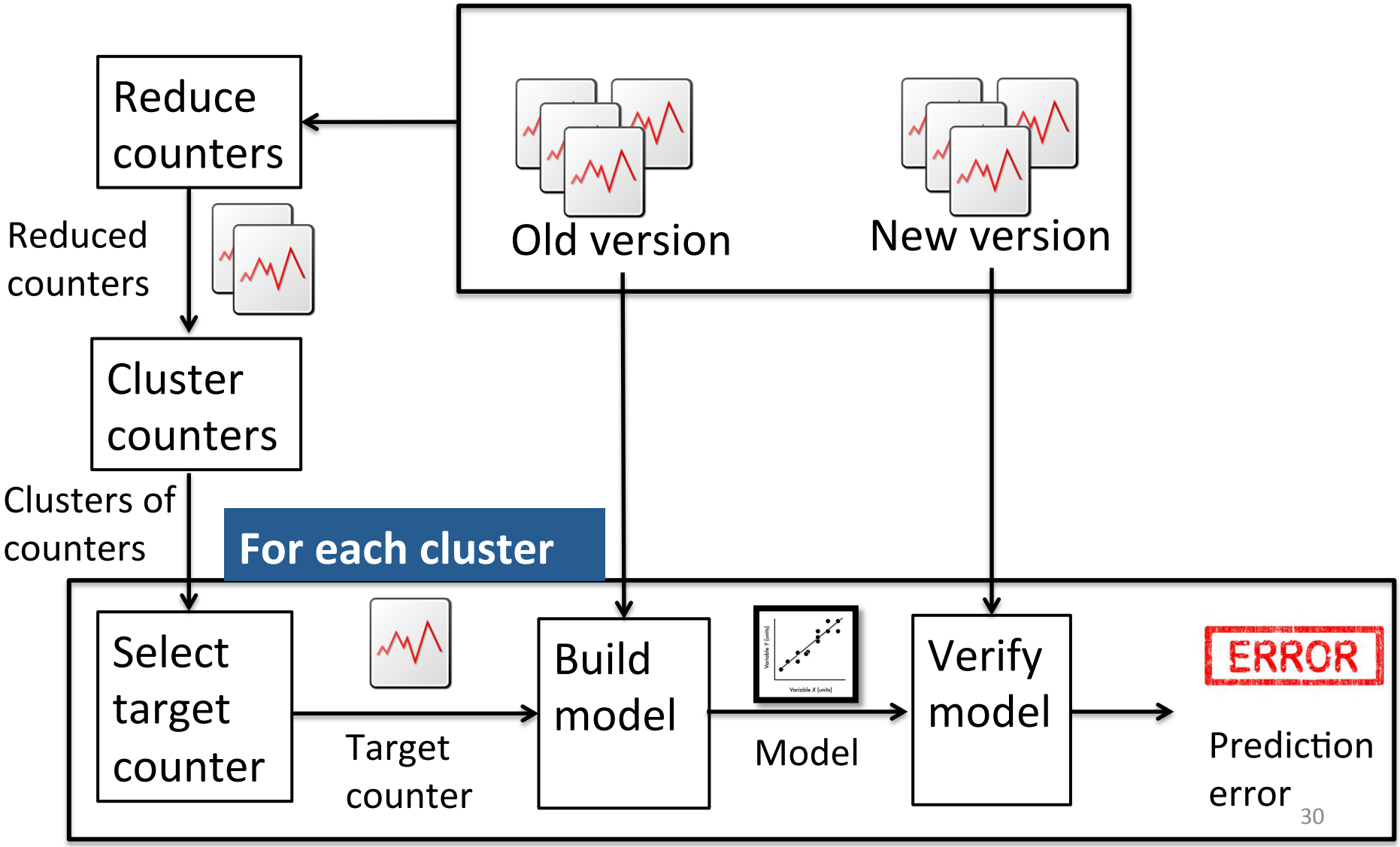
Model-based performance regression detection



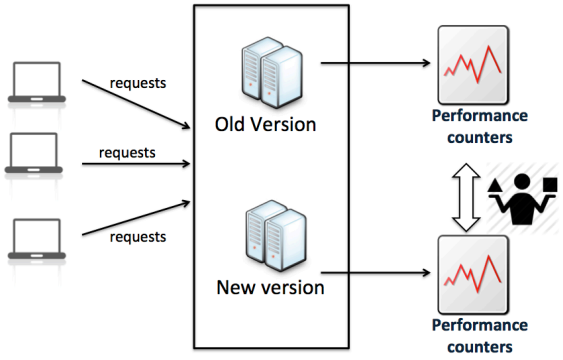
Target counter is selected by experience!

23

Our approach

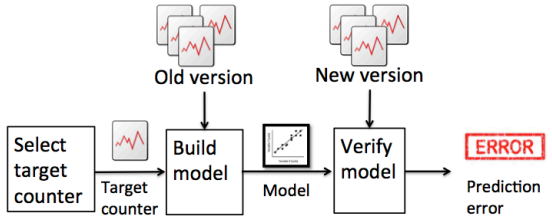


How to detect performance regression?



19

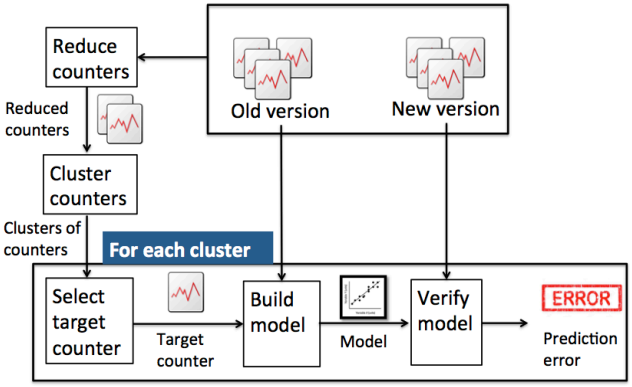
Model-based performance regression detection



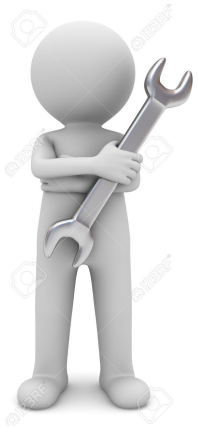
Target counter is selected by experience!

23

Our approach



Applicability



How many target counters does our approach pick?

Our approach picks a small number of target counters.

Accuracy

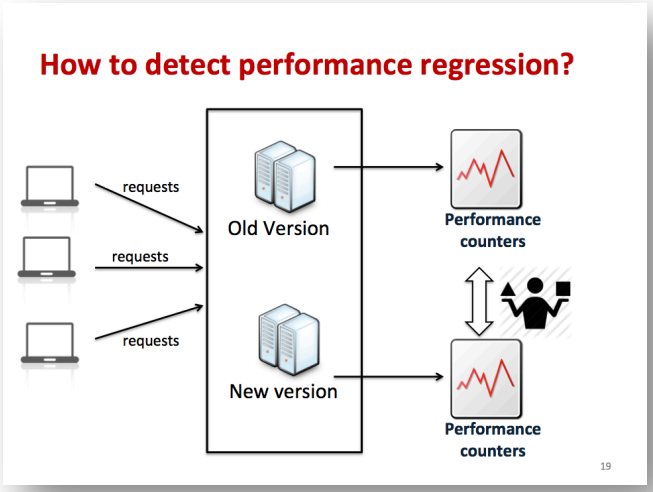
Can our approach detect performance regressions?

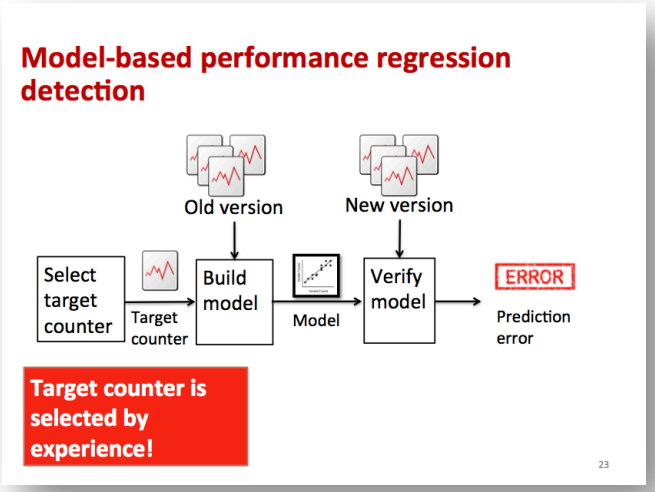
Our approach can detect performance regressions and is not heavily impacted by threshold value.

Comparison

Is our approach better than traditional approaches?

Our approach out performs traditional approaches.





weiyishang.wordpress.com

swy@cs.queensu.ca