How did *this* get published?
Pitfalls in experimental evaluation of computing systems

José Nelson Amaral
University of Alberta
Edmonton, AB, Canada

Thing #1

Aggregation

Thing #2

Learning

Thing #3

Reproducibility

So, a computing scientist entered a Store….

http://archive.constantcontact.com/fs042/1101916237075/archive/1102594461324.html

http://bitchmagazine.org/post/beyond-the-panel-an-interview-with-danielle-corsetto-of-girls-with-slingshots

So, a computing scientist entered a Store….

$ 3,000.00

$ 200.00

They want $2,700 for the server and $100 for the iPod.

I will get both and pay only $2,240 altogether!

http://bitchmagazine.org/post/beyond-the-panel-an-interview-with-danielle-corsetto-of-girls-with-slingshots

So, a computing scientist entered an Store….

Ma'am you are $560 short.

But the average of 10% and 50% is 30% and 70% of $3,200 is $2,240.

$ 200.00

$ 3,000.00

http://www.businessinsider.com/10-ways-to-fix-googles-busted-android-app-market-2010-1?op=1

http://bitchmagazine.org/post/beyond-the-panel-an-interview-with-danielle-corsetto-of-girls-with-slingshots

So, a computing scientist entered an Store….

Ma'am you cannot take the arithmetic average of percentages!

But… I just came from at top CS conference in San Jose where they do it!

$ 200.00

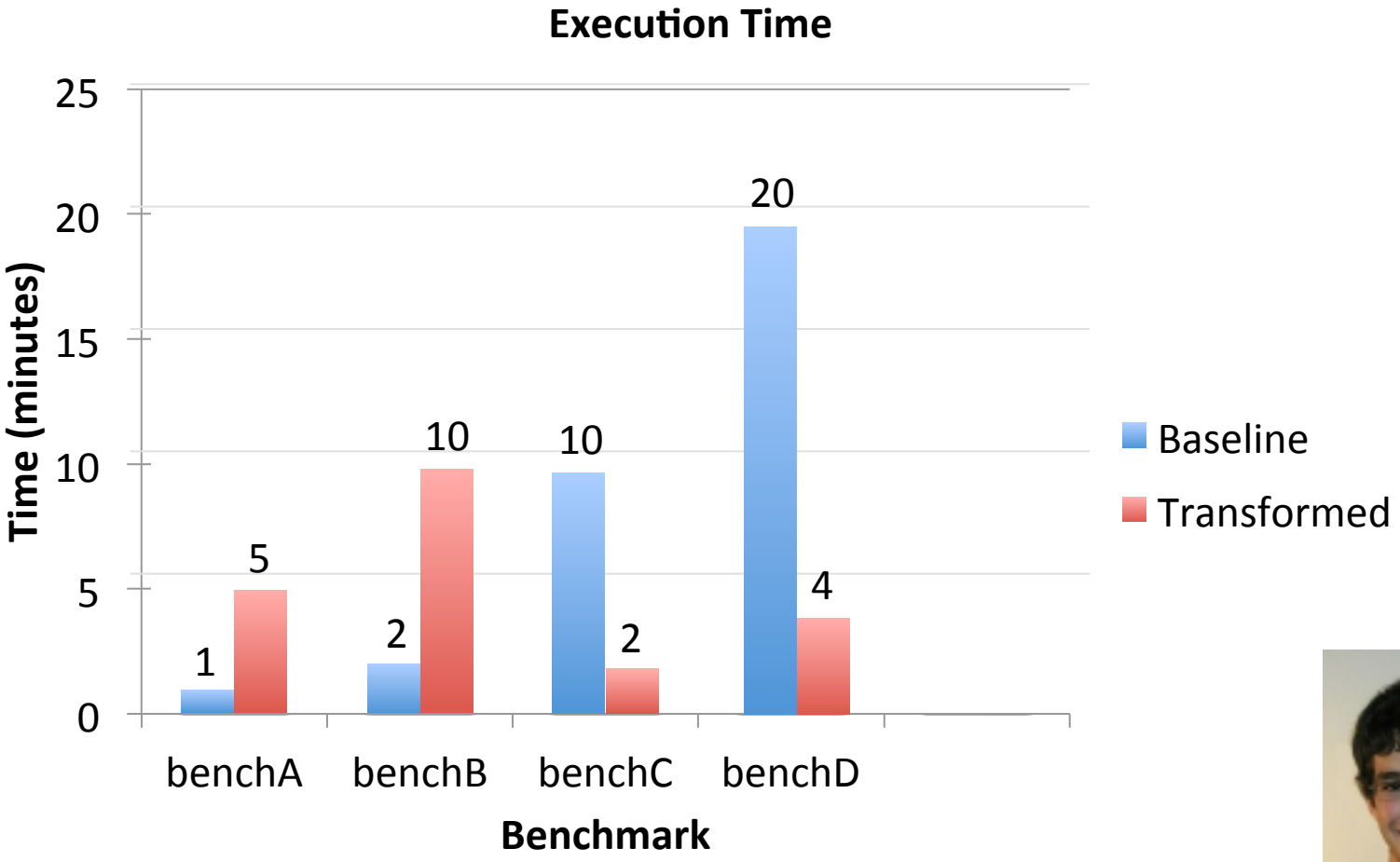
$ 3,000.00

http://www.businessinsider.com/10-ways-to-fix-googles-busted-android-app-market-2010-1?op=1

http://bitchmagazine.org/post/beyond-the-panel-an-interview-with-danielle-corsetto-of-girls-with-slingshots

The Problem with Averages

A Hypothetical Experiment

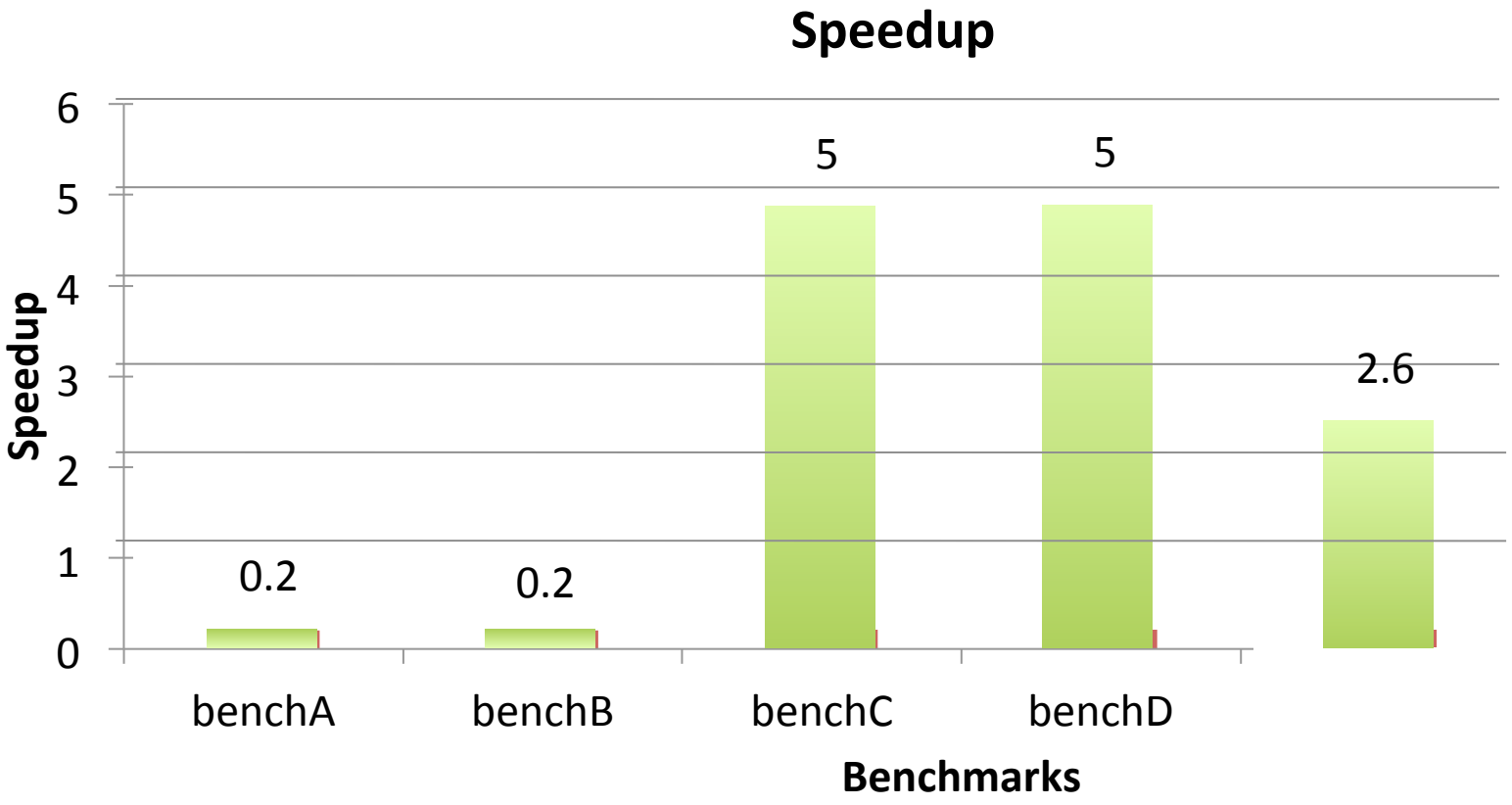


*With thanks to Iain Ireland

Speedup

$$\text{Speedup} = \frac{\text{Baseline Time}}{\text{Transformed Time}}$$

Performance Comparison

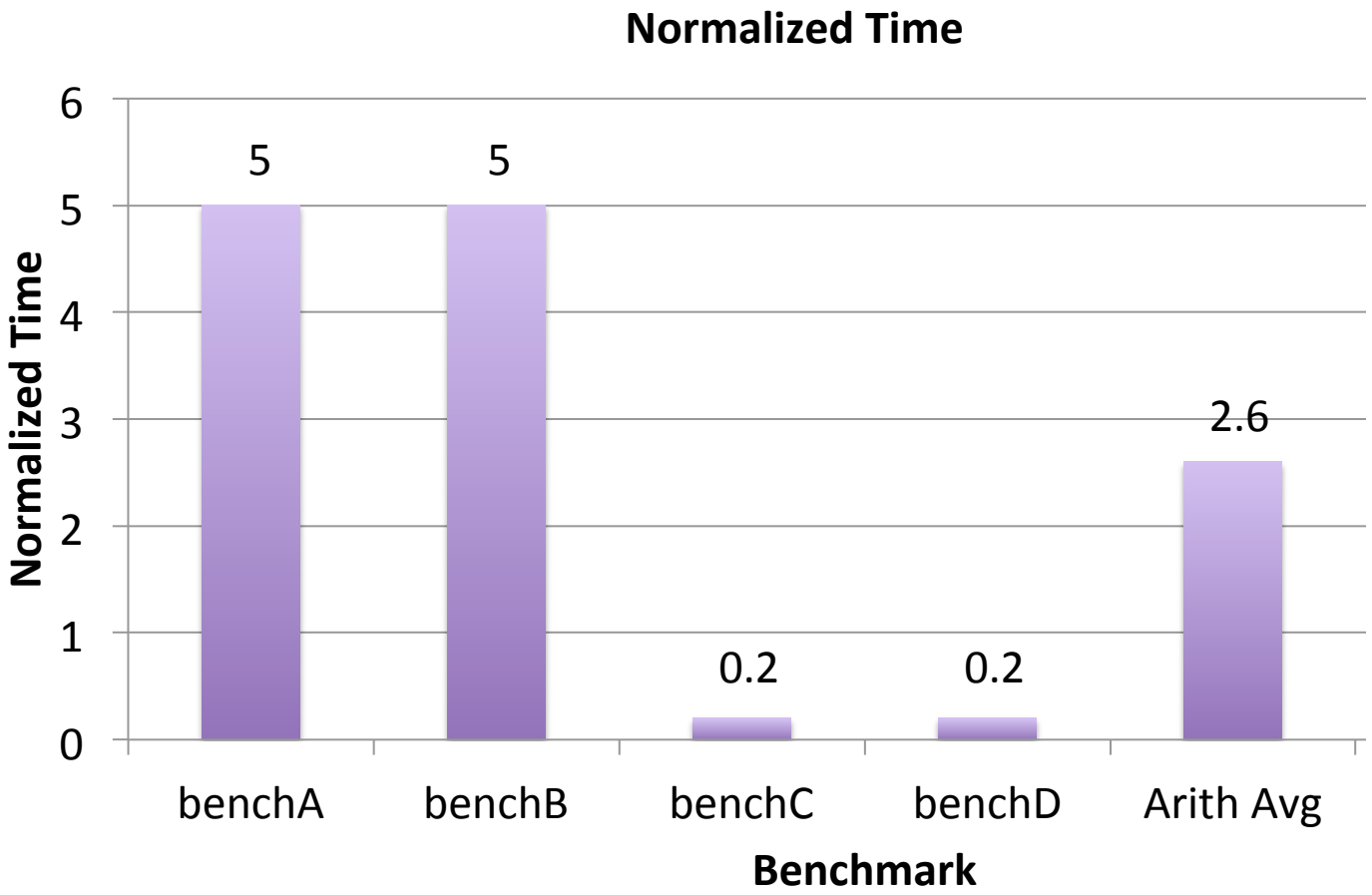


The transformed system is, on average, 2.6 times **faster** than the baseline!

Normalized Time

$$\textit{Normalized Time} = \frac{\text{Transformed Time}}{\text{Baseline Time}}$$

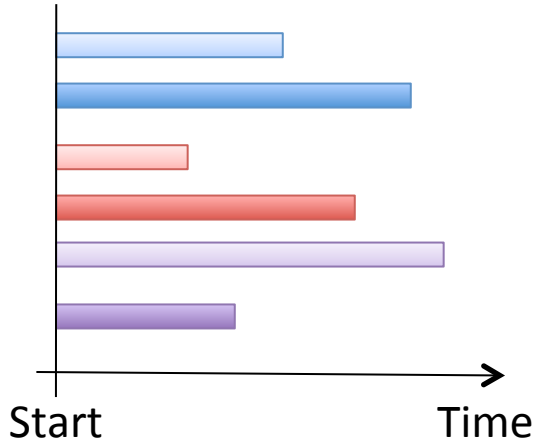
Normalized Time



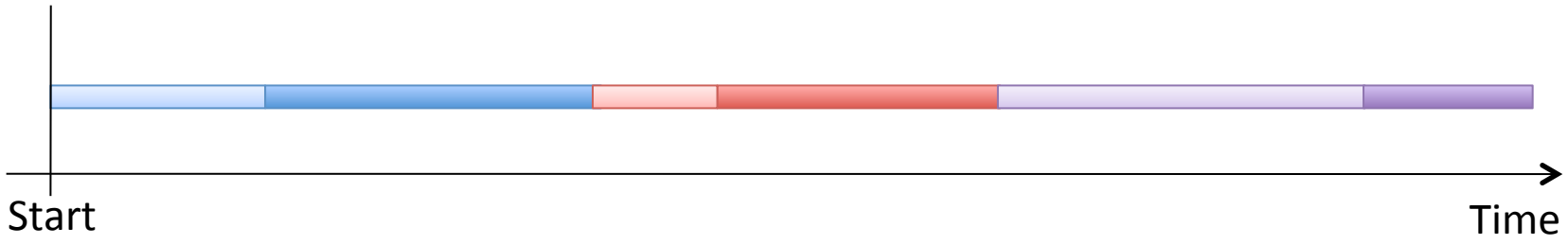
The transformed system is, on average, 2.6 times **slower** than the baseline!

Latency × Throughput

- What matters is latency:



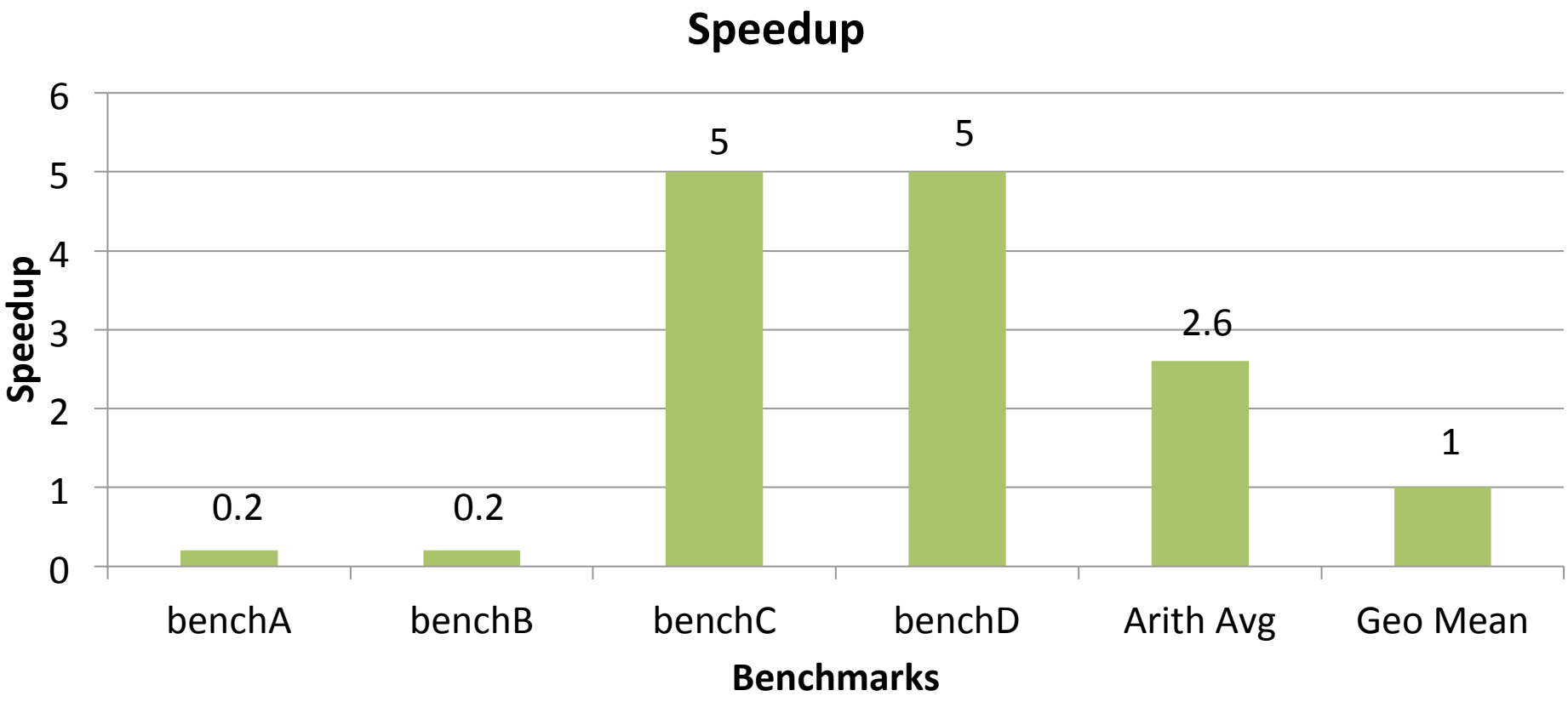
- What matters is throughput:



Aggregation for Latency: Geometric Mean

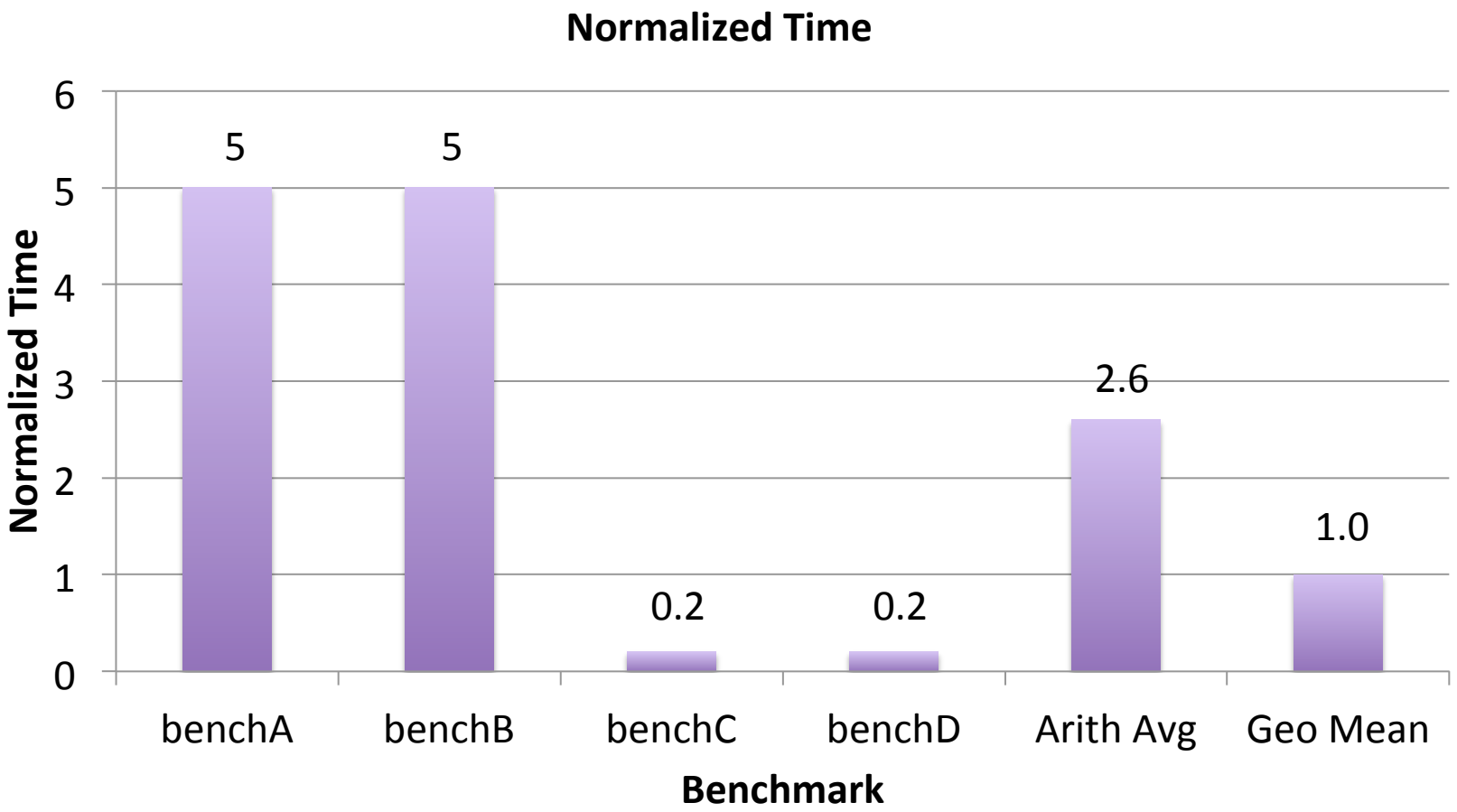
$$GeoMean = \sqrt[n]{\prod_{i=0}^{n-1} s_i}$$

Speedup



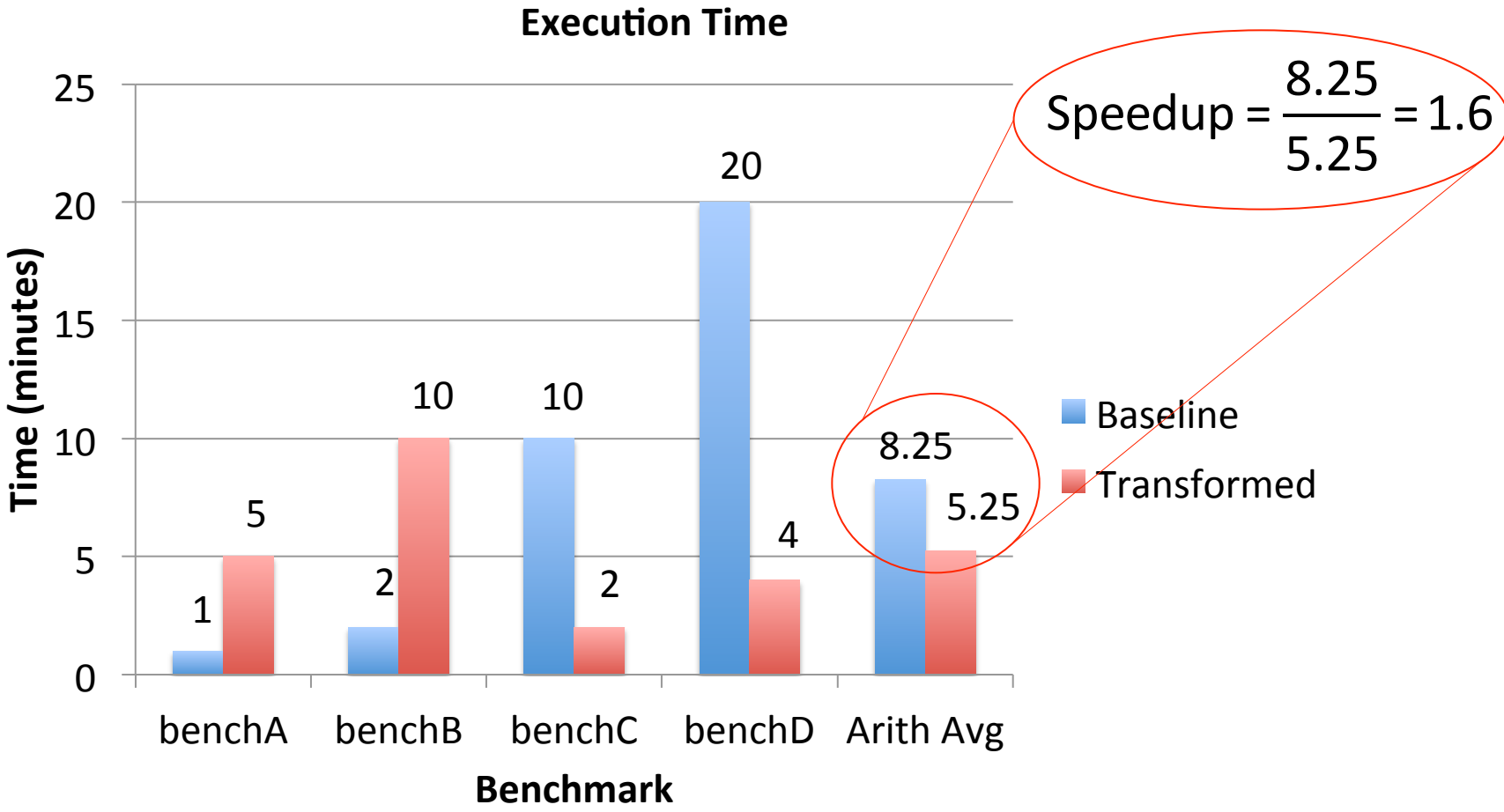
The performance of the transformed system is, on average, the same as the baseline!

Normalized Time



The performance of the transformed system is, on average, the same as the baseline!

Aggregation for Throughput



The throughput of the transformed system is, on average, **1.6** times **faster** than the baseline.

The Evidence

- A careful reader will find the use of arithmetic average to aggregate normalized numbers in many top CS conferences.
- Papers that have done that have appeared in:
  - LCTES 2011
  - PLDI 2012 (at least two papers)
  - CGO 2012
    - A paper where the use of the wrong average changed a negative conclusion into a positive one.
  - 2007 SPEC Workshop
    - A methodology paper by myself and a student that won the best paper award.

This is not a new observation…

Edgar H. Sibley
Panel Editor

Using the arithmetic mean to summarize normalized benchmark results leads to mistaken conclusions that can be avoided by using the preferred method: the geometric mean.

HOW NOT TO LIE WITH STATISTICS: THE CORRECT WAY TO SUMMARIZE BENCHMARK RESULTS

PHILIP J. FLEMING and JOHN J. WALLACE

Communications of the ACM, March 1986, pp. 218-221.

RULE 1: ***Do Not Use the Arithmetic Mean to Average Normalized Numbers***

RULE 2: ***Use the Geometric Mean to Average Normalized Numbers***

RULE 3: ***Use the Sum (or arithmetic mean) of Raw, Unnormalized Results whenever This "Total" Has Some Meaning***

Communications of the ACM, March 1986, pp. 218-221.

No need to dig dusty papers…

Article Talk Read Edit Search

Geometric mean

From Wikipedia, the free encyclopedia

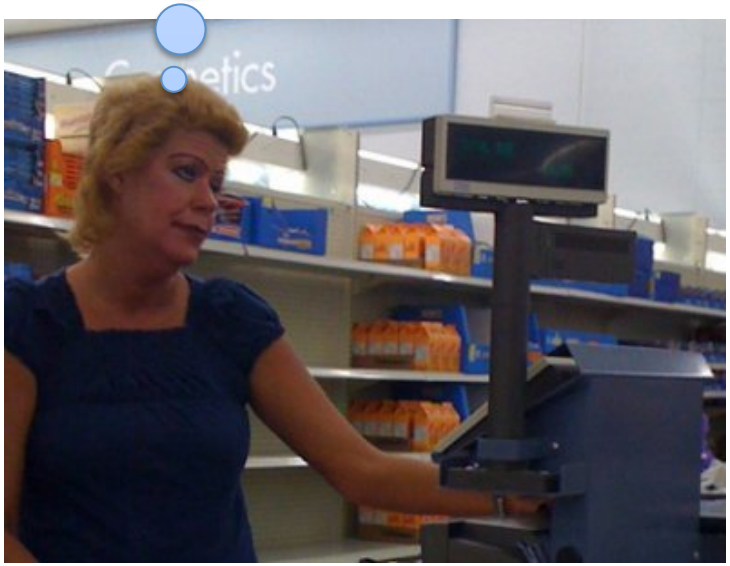
$$GM\left(\frac{X_i}{Y_i}\right) = \frac{GM(X_i)}{GM(Y_i)}$$

This makes the geometric mean the only correct mean when averaging *normalized* results, that is results that are presented as ratios to reference values.[4] This is the case when presenting

So, the computing scientist returns to the Store…

???

Hello. I am just back from Beijing. Now I know that we should take the geometric average of percentages.



$ 200.00

$ 3,000.00

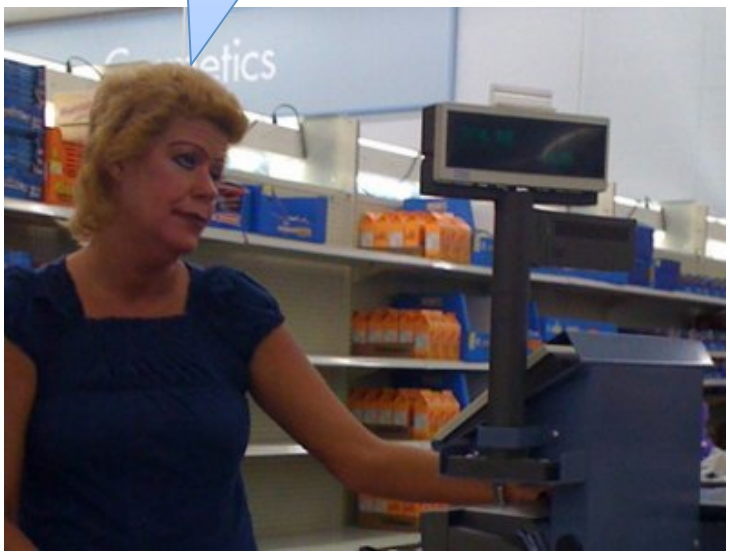
http://www.businessinsider.com/10-ways-to-fix-googles-busted-android-app-market-2010-1?op=1

http://bitchmagazine.org/post/beyond-the-panel-an-interview-with-danielle-corsetto-of-girls-with-slingshots

So, a computing scientist entered a Store….

Sorry Ma'am, we don't average percentages…

Thus I should get $\sqrt[2]{50 \times 10}$ = 22.36% discount and pay 0.7764×$3,200 = $2,484.48



$ 200.00

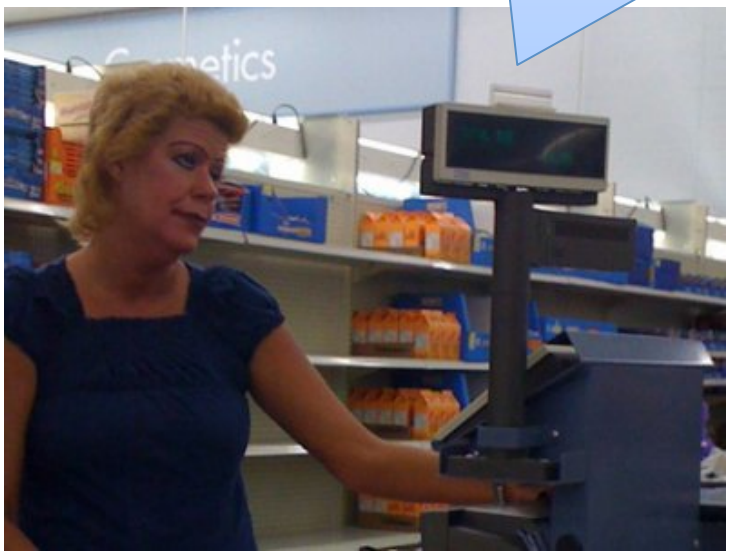
$ 3,000.00

http://www.businessinsider.com/10-ways-to-fix-googles-busted-android-app-market-2010-1?op=1

http://bitchmagazine.org/post/beyond-the-panel-an-interview-with-danielle-corsetto-of-girls-with-slingshots

So, a computing scientist entered a Store….

The original price is $ 3,200. You pay
$ 2,700 + $ 100 = $ 2,800.
If you want an aggregate summary,
your discount is 400/3,200 = 12.5%

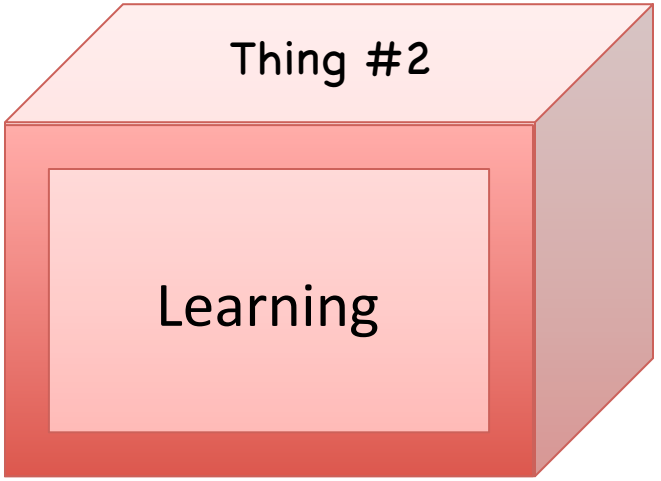


$ 200.00

$ 3,000.00

http://www.businessinsider.com/10-ways-to-fix-googles-busted-android-app-market-2010-1?op=1

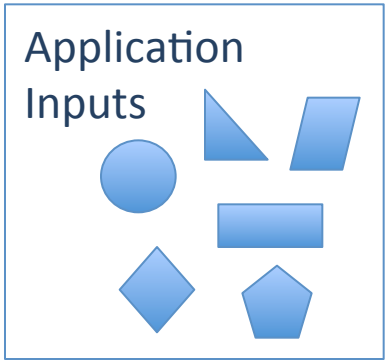
http://bitchmagazine.org/post/beyond-the-panel-an-interview-with-danielle-corsetto-of-girls-with-slingshots

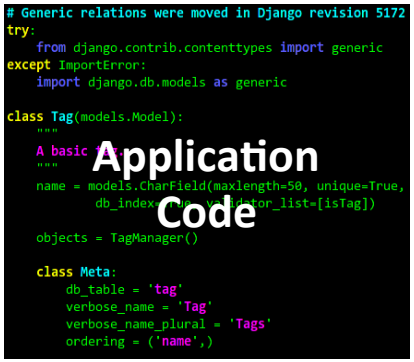


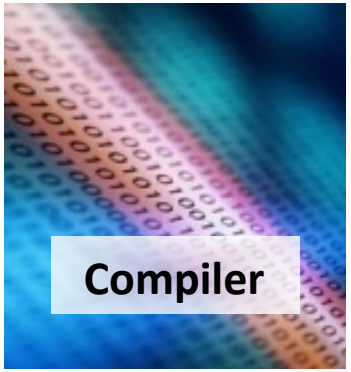
Disregard to methodology when using automated learning

Example:
Evaluation of Feedback Directed Optimization (FDO)

We have:



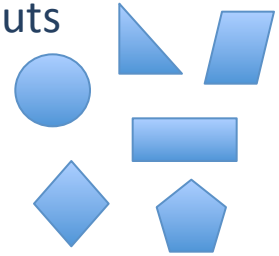




http://www.orchardoo.com

We want to measure the effectiveness of an FDO-based code transformation.

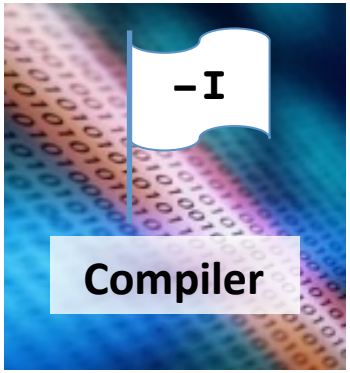
Application Inputs



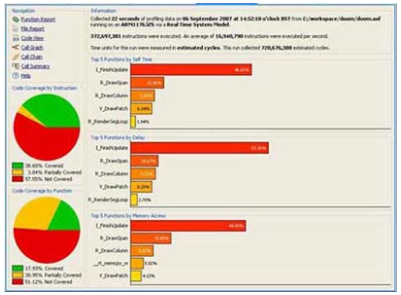
Training Set

```
# Generic relations were moved in Django revision 5172
try:
    from django.contrib.contenttypes import generic
except ImportError:
    import django.db.models as generic

class Tag(models.Model):
    """
    A basic Tag
    """
    name = models.CharField(maxlength=50, unique=True,
            db_index=True, validator_list=[isTag])
    objects = TagManager()

    class Meta:
        db_table = 'tag'
        verbose_name = 'Tag'
        verbose_name_plural = 'Tags'
        ordering = ('name',)
```

Application Code

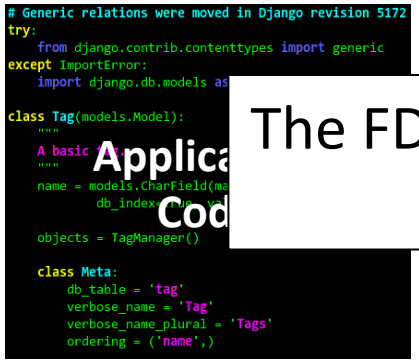


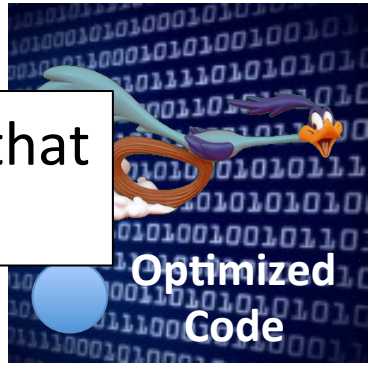
http://www.orchardoo.com



Profile

Evaluation Set

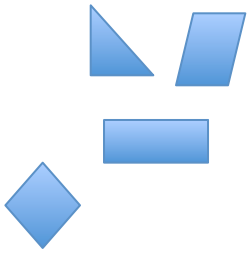


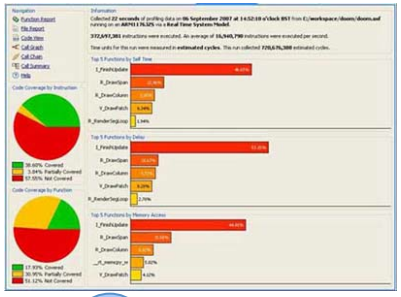


The FDO transformation produces code that is XX **faster** for this application.

http://www.orchardoo.com

Un-evaluated Set





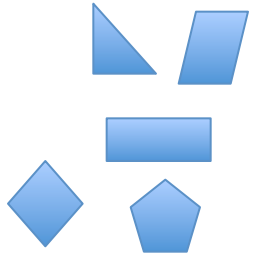
Profile

Performance

The Evidence

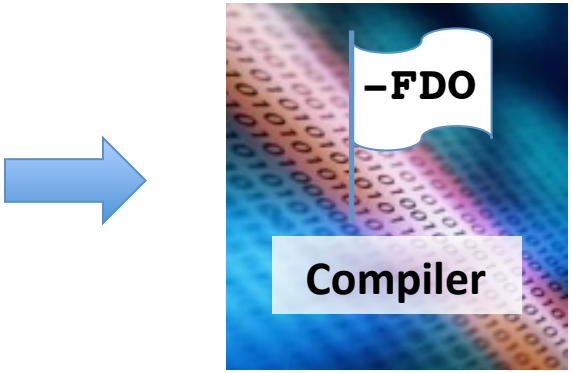
- Many papers that use a single input for training and a single input for testing appeared in conferences (notably CGO).
- For instance, a paper that uses a single input for training and a single input for testing appears in:
  - ASPLOS 2004

Evaluation Set

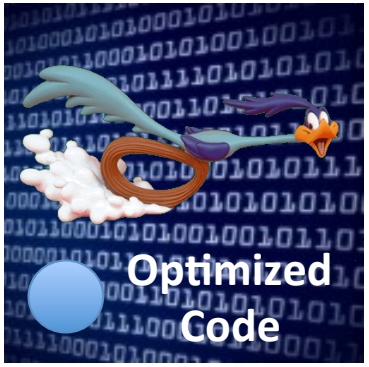


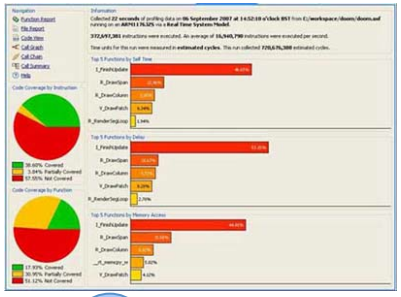
```
# Generic relations were moved in Django revision 5172
try:
    from django.contrib.contenttypes import generic
except ImportError:
    import django.db.models as generic

class Tag(models.Model):
    """
    A basic Tag
    """
    name = models.CharField(maxlength=50, unique=True,
           db_index=True, validator_list=[isTag])
    objects = TagManager()

    class Meta:
        db_table = 'tag'
        verbose_name = 'Tag'
        verbose_name_plural = 'Tags'
        ordering = ('name',)
```

Application Code

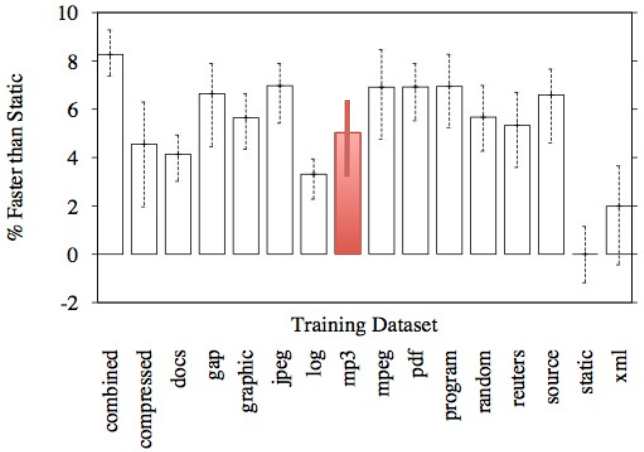


http://www.orchardoo.com





Profile



Performance

Training Set

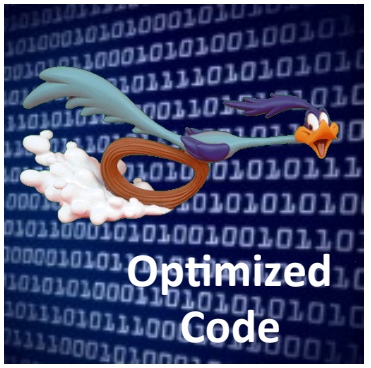
Evaluation Set

Combined Profiling (Berube, ISPASS12)

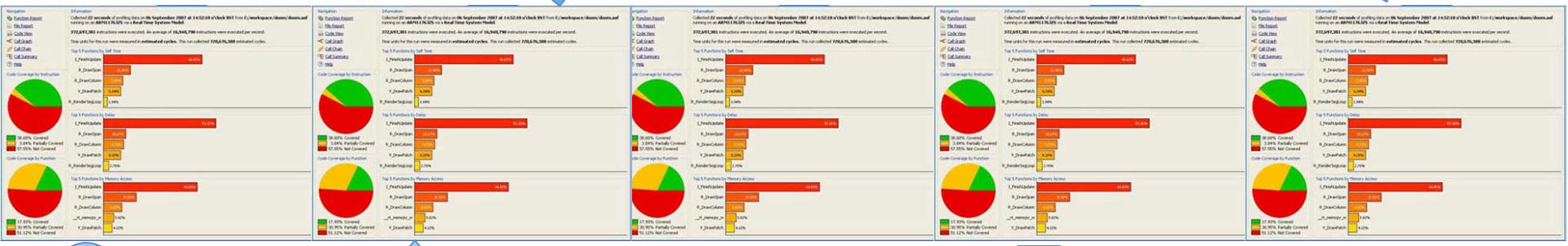
Cross-Validated Evaluation (Berube, SPEC07)

```
# Generic relations were moved in Django revision 5172
try:
    from django.contrib.contenttypes import generic
except ImportError:
    import django.db.models as generic

class Tag(models.Model):
    """
    A basic Tag
    """
    name = models.CharField(maxlength=50, unique=True,
           db_index=True, validator_list=[isTag])

    objects = TagManager()

    class Meta:
        db_table = 'tag'
        verbose_name = 'Tag'
        verbose_name_plural = 'Tags'
        ordering = ('name',)
```

Application Code



http://www.orchardoo.com



Profile

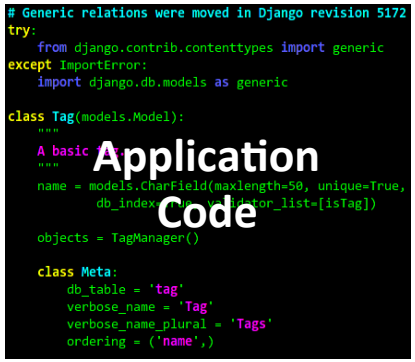
Profile

Profile

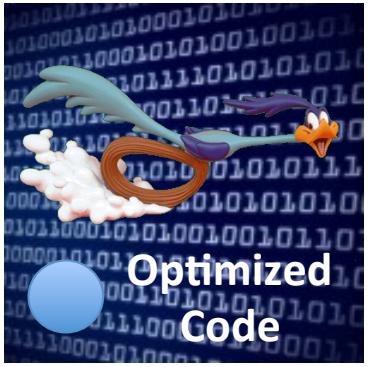
Profile

Profile

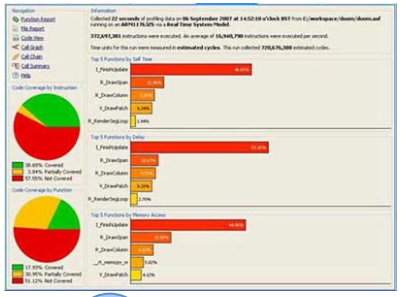
Evaluation Set



http://www.orchardoo.com



Wrong Evaluation!

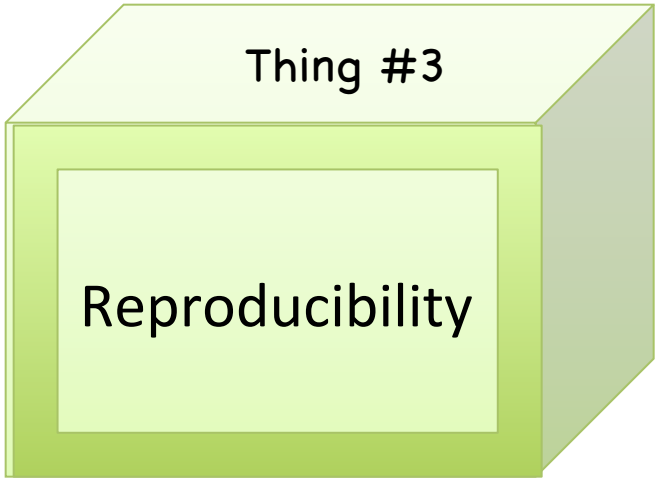


Profile

Performance

The Evidence

- For instance, a paper that incorrectly uses the same input for training and testing appeared in:
  - PLDI 2006



Expectation:

When reproduced, an experimental evaluation should produce similar results.

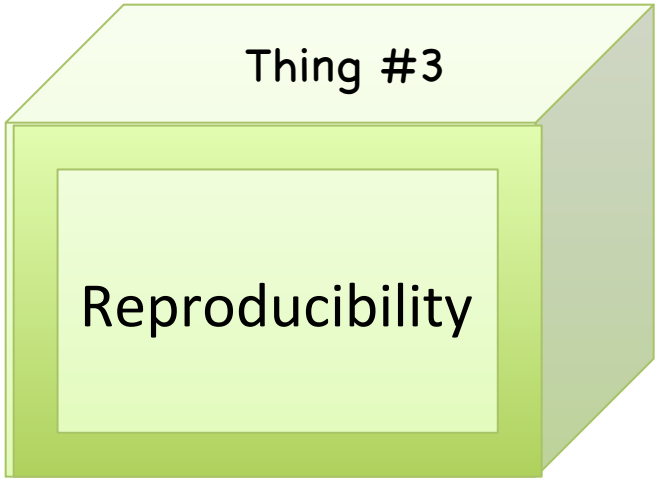
Issues

Have the measurements been repeated a sufficient number of times to capture measurement variations?

Availability of code, data, and precise description of experimental setup.

Lack of incentives for reproducibility studies.

Progress



Program committees/reviewers starting to ask questions about reproducibility.

Steps toward infrastructure to facilitate reproducibility.

SPEC Research Group

http://research.spec.org/

14 industrial organizations

20 universities or research institutes

KIT Karlsruhe Institute of Technology
IBM
SAP
AMD
Cisco Systems
Berkeley University of California
Northeastern University
hp
vmware
intel
Technische Universität Darmstadt
Purdue University
U C
Oracle
Siemens
Universität Mannheim
The University of Texas at Austin
Universiteit Gent
NICTA
YAHOO!
University of Alberta
redhat.
CAU Christian-Albrechts-Universität zu Kiel
San Francisco State University
DELL
財團法人資訊工業策進會 Institute for Information Industry
IIT Bombay
University of Minnesota Driven to Discover℠
浙江大學 Zhejiang University
TU Delft Delft University of Technology
McGill
OFFIS
salesforce
HPI Hasso Plattner Institut IT Systems Engineering | Universität Potsdam

SPEC Research Group

http://research.spec.org/

Performance Evaluation

Benchmarks for New Areas

Performance Evaluation Tools

Evaluation Methodology

Repository for Reproducibility

http://icpe2013.ipd.kit.edu/

Evaluate Collaboratory:

http://evaluate.inf.usi.ch/

Open Letter to PC Chairs

Anti Patterns

Evaluation in CS education

Parting Thoughts....

Creating a culture that enables full reproducibility seems daunting…

Initially we could aim for:

Reasonable expectation by a reasonable reader that, if reproduced, the experimental evaluation would produce similar results.