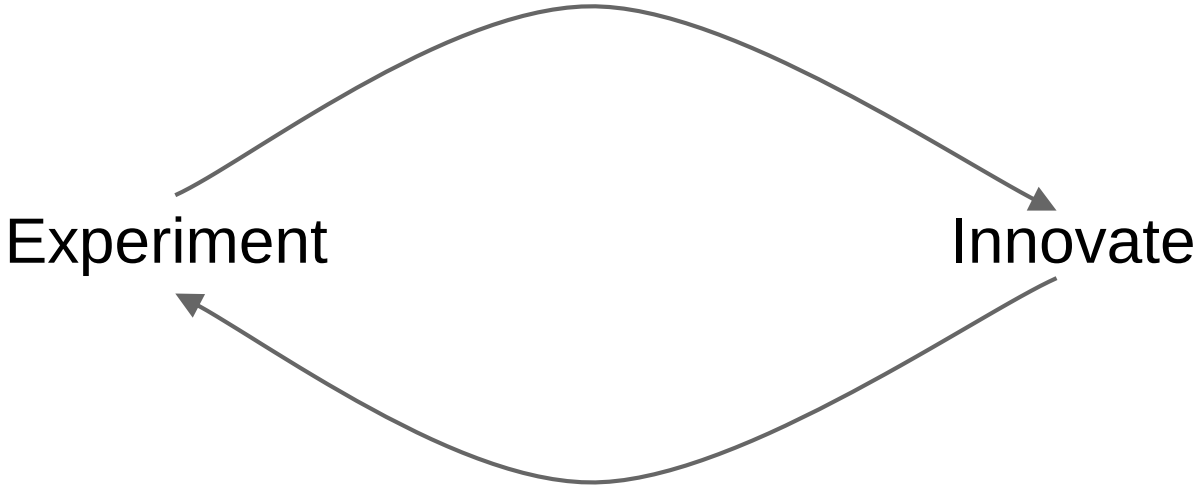
Should you trust your experimental results?

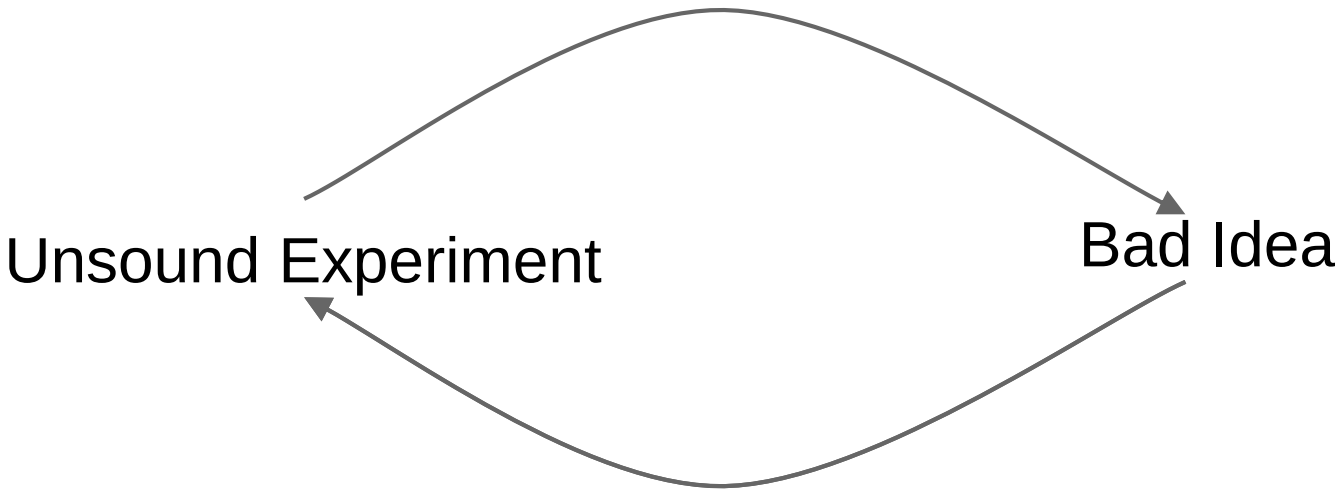
Amer Diwan, Google
Stephen M. Blackburn, ANU
Matthias Hauswirth, U. Lugano
Peter F. Sweeney, IBM Research
Attendees of Evaluate '11 workshop

Why worry?



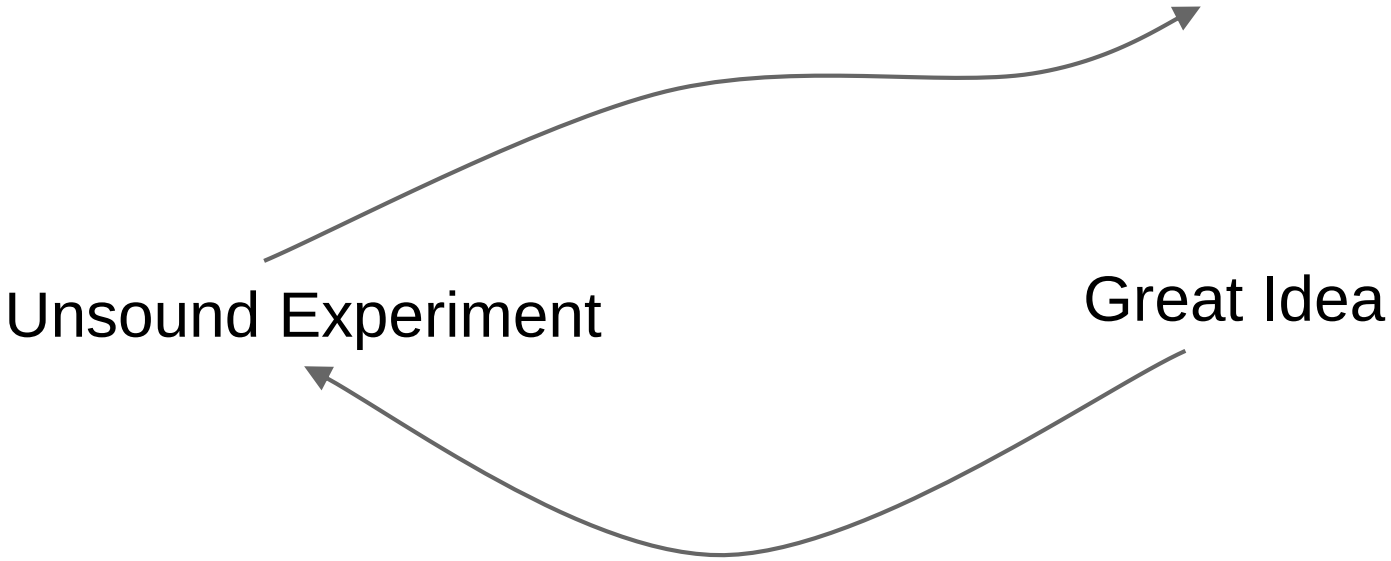
For scientific progress we need sound experiments

Unsound experiments



Make a bad idea look great!

Unsound experiments



Make a great idea look bad!

Thesis

Sound experimentation is critical but requires

- Creativity
- Diligence

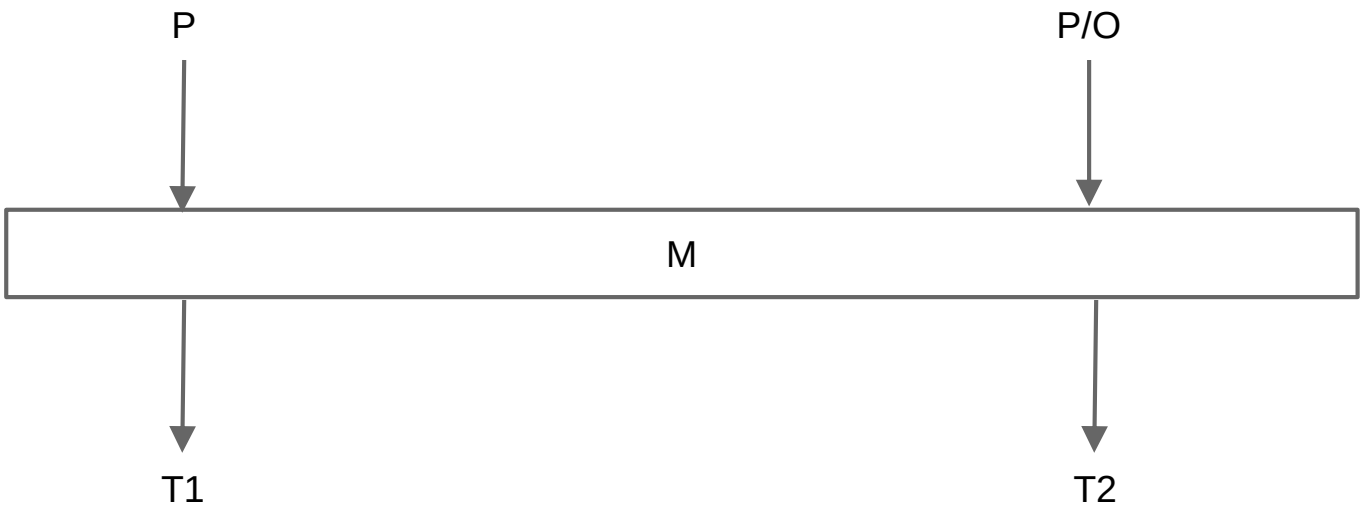
As a community, we must

- Learn how to design and conduct sound experiments
- Reward sound experimentation

A simple experiment

Goal: To characterise the speedup of optimization O

Experiment: Measure program P on unloaded machine M with/without O



Claim: O speeds up programs by 10%

Why is this unsound?

Scope of experiment << Scope of claim

The relationship of the two scopes determines if an experiment is sound

Sound experiments

Sufficient for sound experiment:

Scope of claim <= Scope of experiment

Option 1: Reduce claim
Option 2: Extend experiment

What are the common causes of unsound experiments?

The four *fatal* sins

The deadly sins do not stand in the way of a PLDI acceptance:

It is our pleasure to inform you that your paper titled "*Envy of PLDI authors*" was accepted to PLDI ...

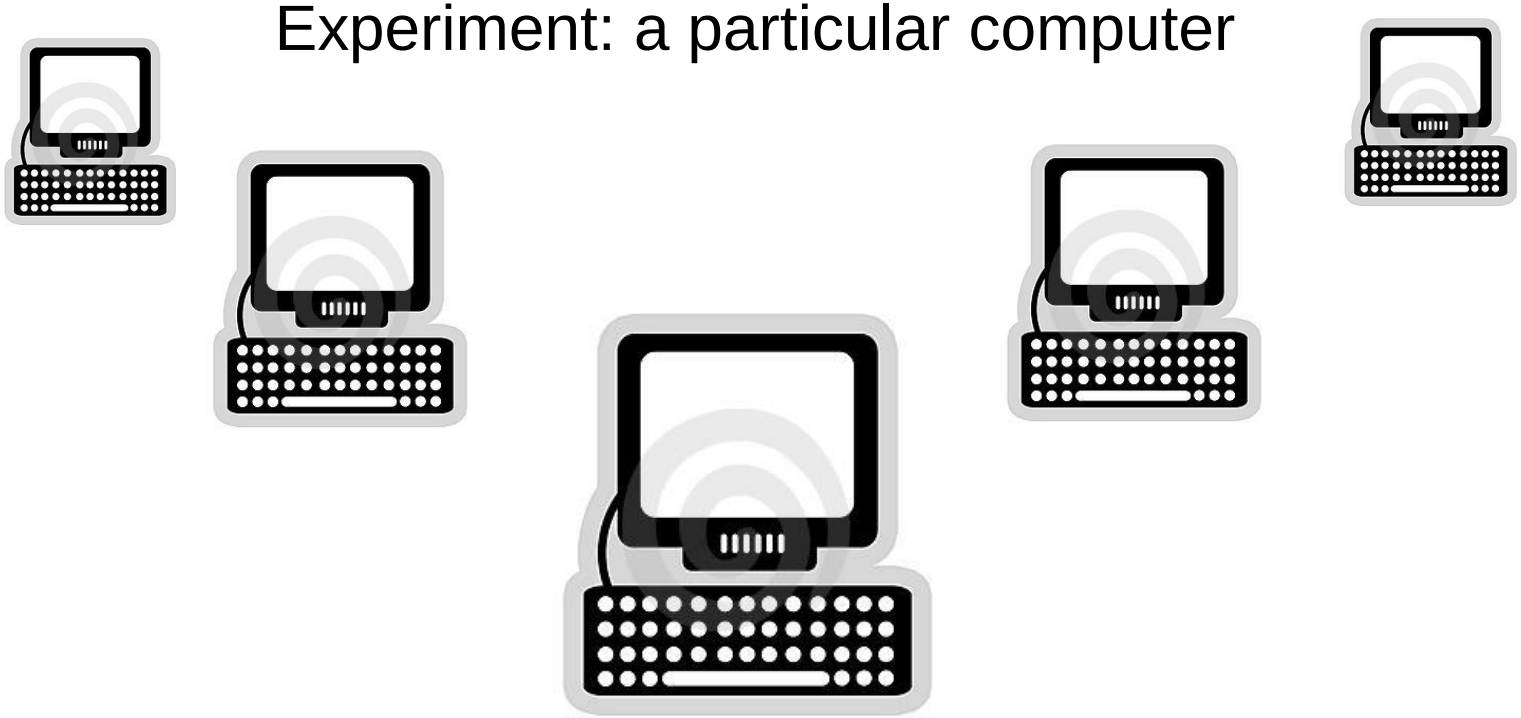
But the four fatal sins might!

Sin 1: Ignorance

Defn: *Ignoring components necessary for Claim*

Claim: all computers

Experiment: a particular computer



Sin 1: Ignorance

Defn: *Ignoring components necessary for Claim*

Experiment:
one benchmark

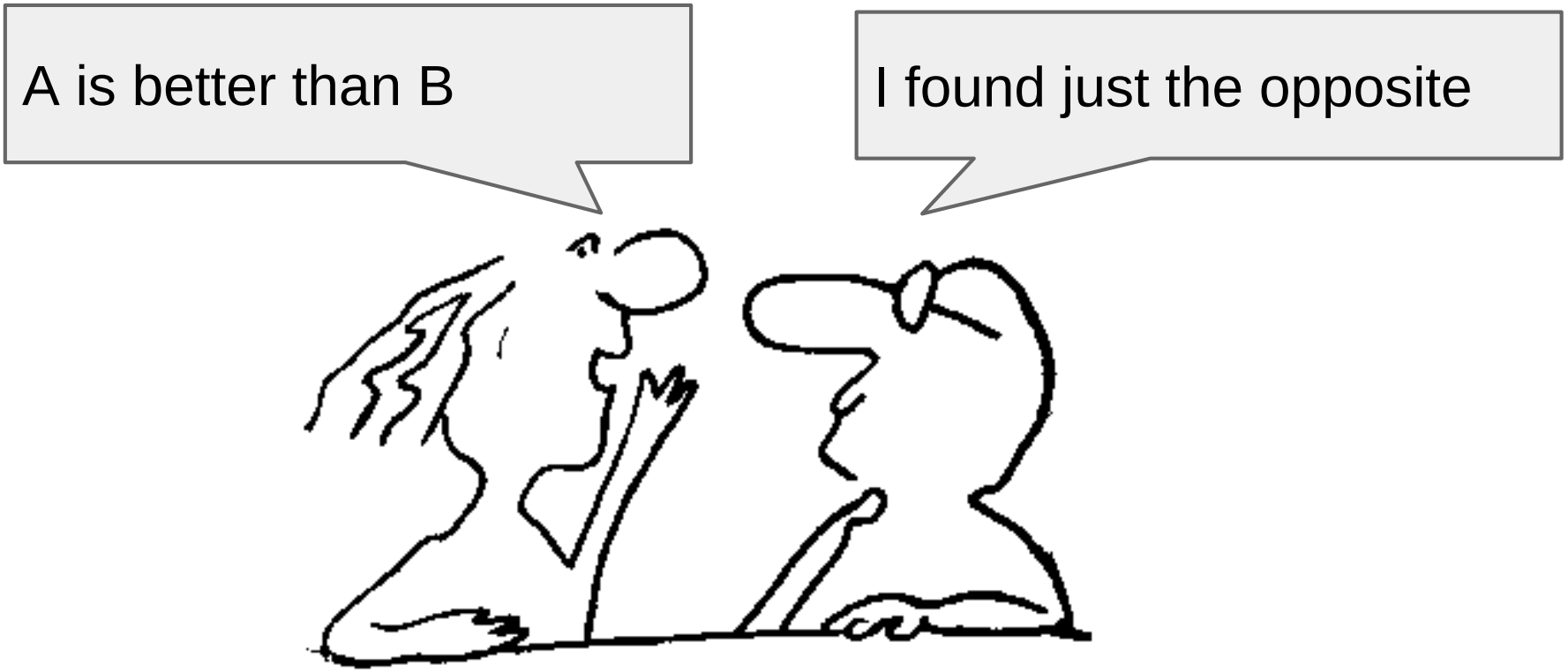
Claim:
full suite

avora

the DaCapo benchmark suite

Ignorance systematically biases results

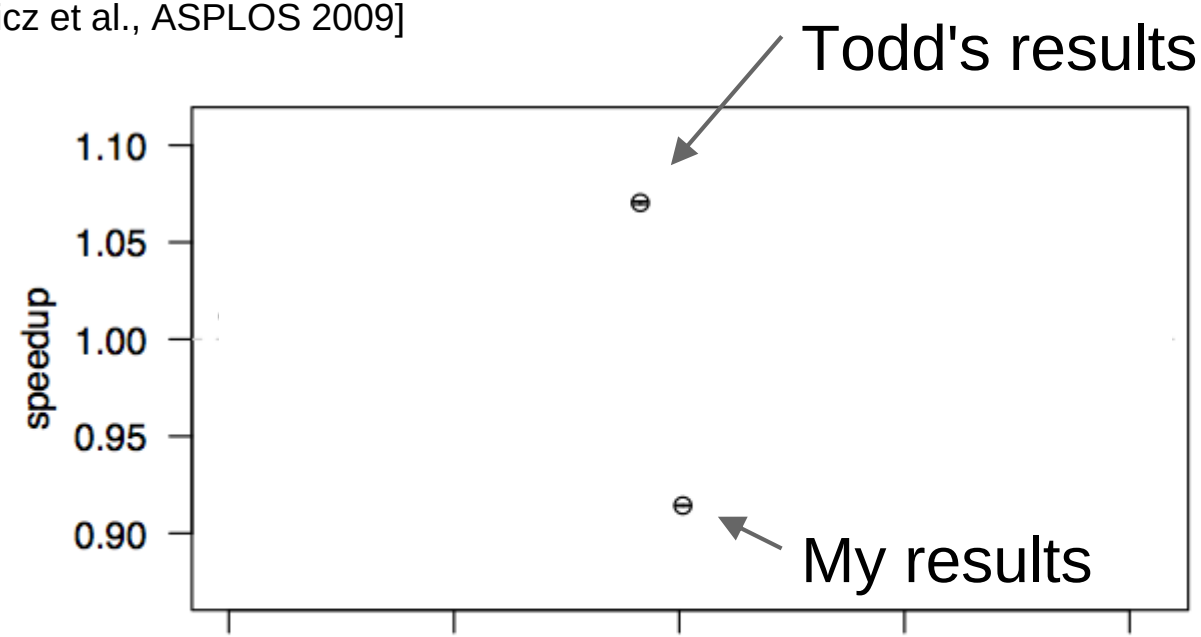
Ignorance is not obvious!



Have you had this conversation with a collaborator?

Ignoring Linux environment variables

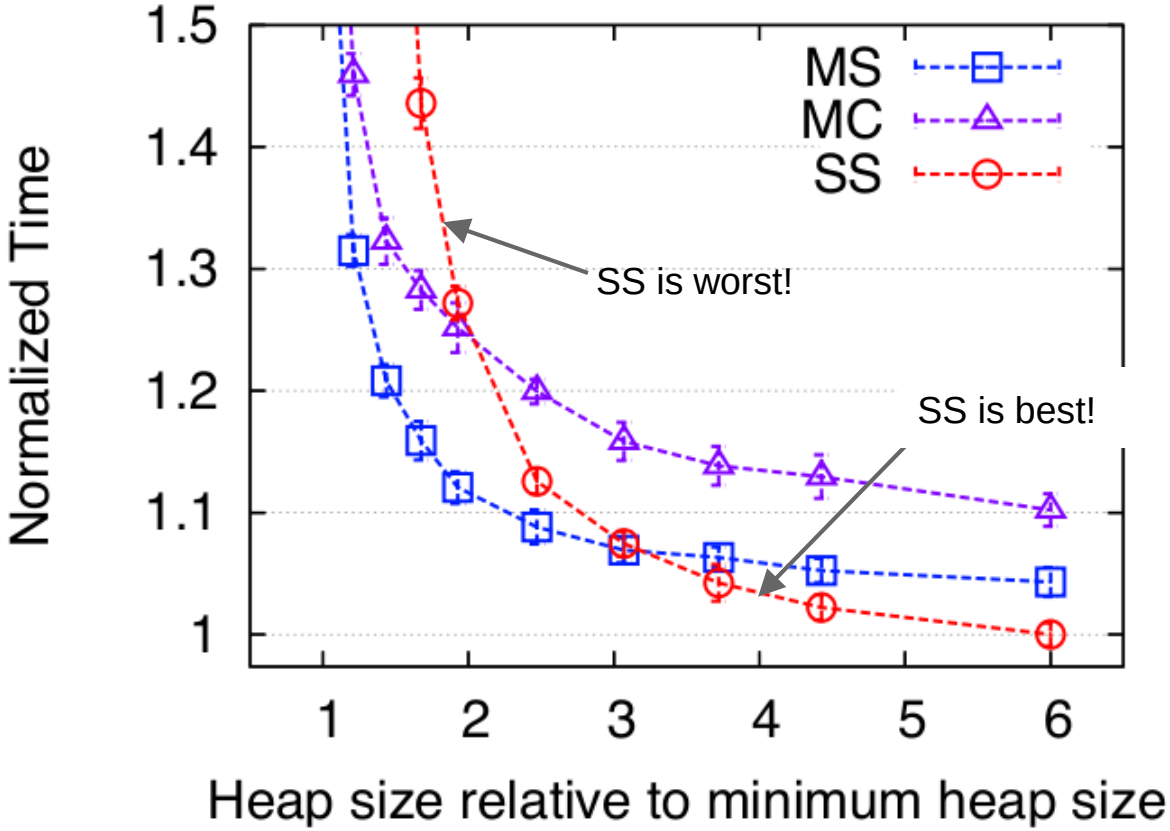
[Mytkowicz et al., ASPLOS 2009]



Changing the environment can change the outcome of your experiment!

Ignoring heap size

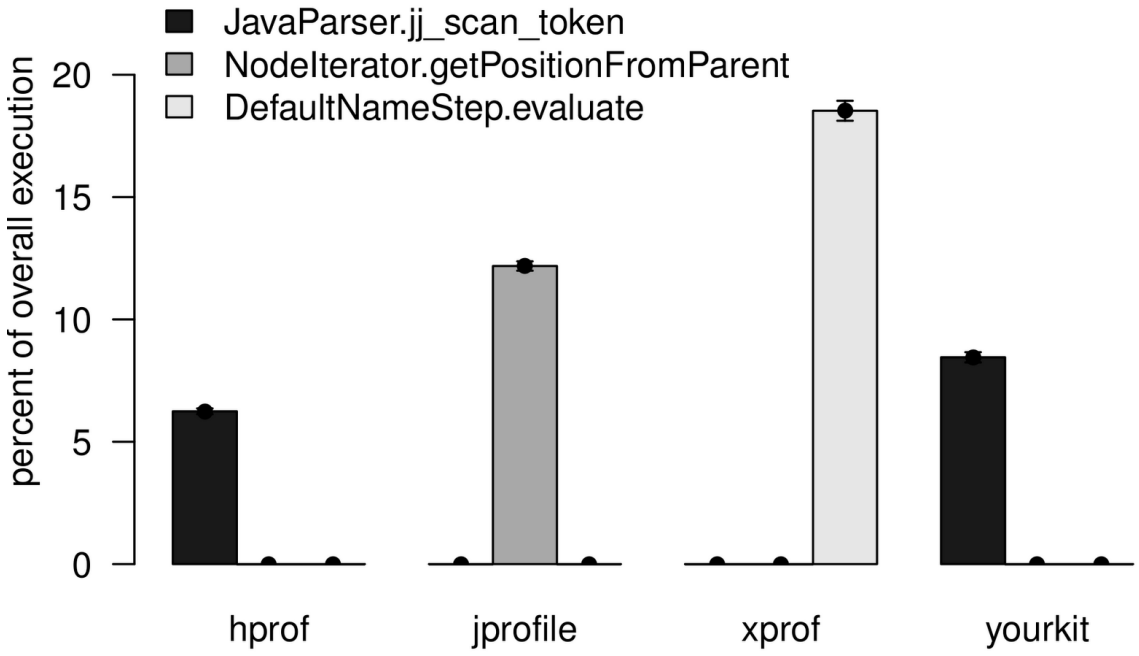
Graph from [Blackburn et al., OOPSLA 2006]



Changing heap size can change the outcome of your experiment!

Ignoring profiler bias

[Mytkowicz et al., PLDI 2010]



Different profilers can yield contradictory conclusions!

Sin 2: Inappropriateness

Defn: *Using components irrelevant for Claim*

Experiment:
Server applications

Claim:
Mobile performance

Sin of inappropriateness

Defn: *Using components irrelevant for Claim*

Experiment:
Compute benchmarks

Claim:
GC performance

http://www.ivankuznetsov.com/

Inappropriateness produces unsupported claims

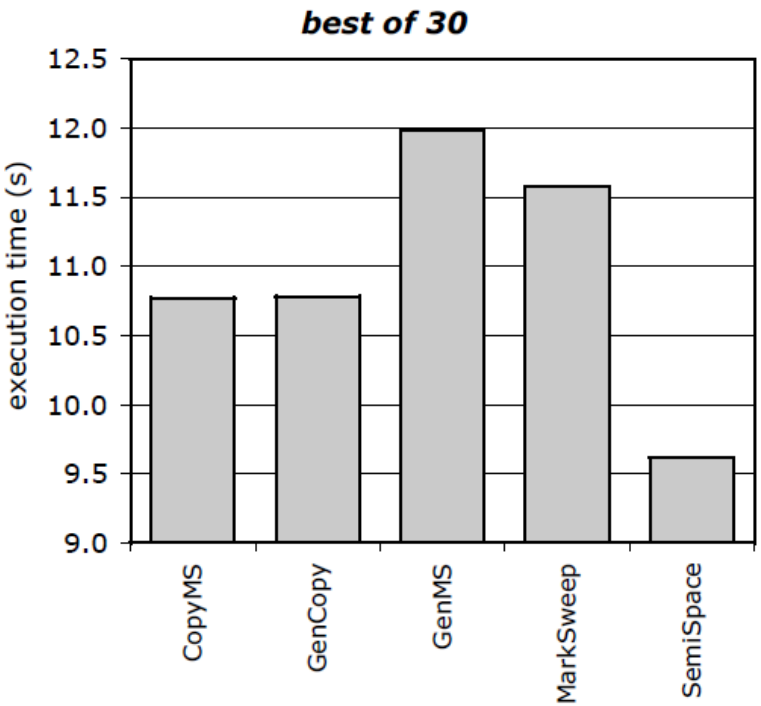
Inappropriateness is not obvious!

Has your optimization ever delivered a 10% improvement

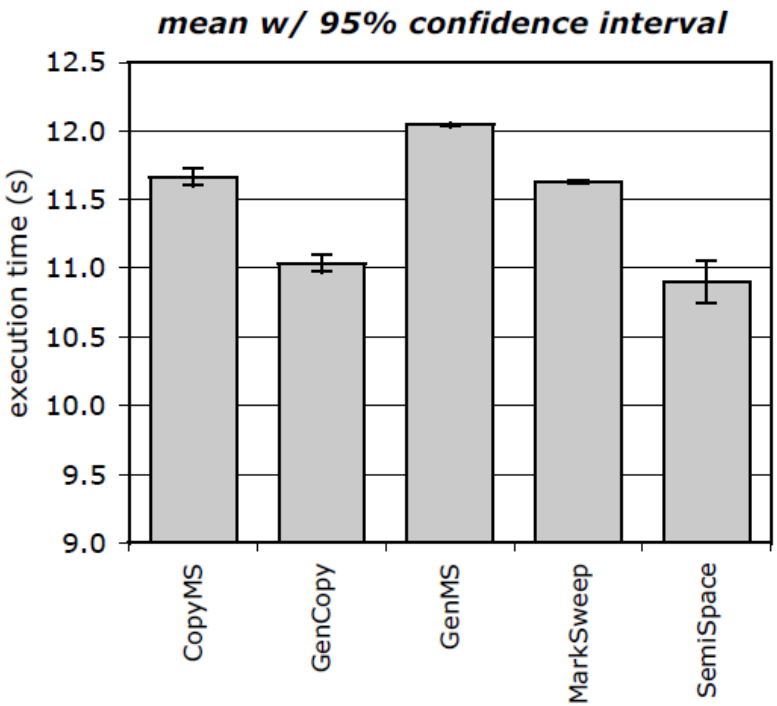
...which never materialized in the "wild"?

Inappropriate statistics

[Georges and Eeckhout, 2007]:



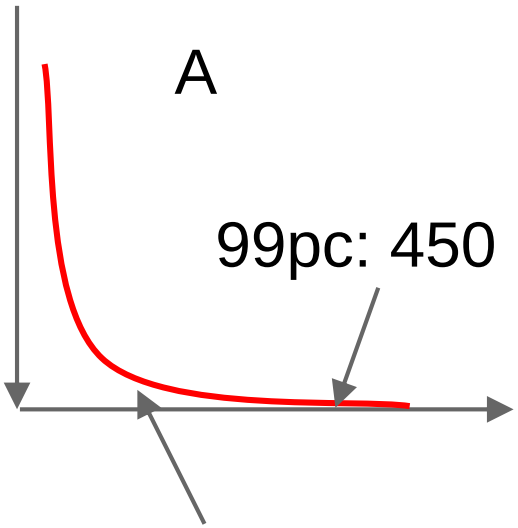
(SemiSpace is best by far)

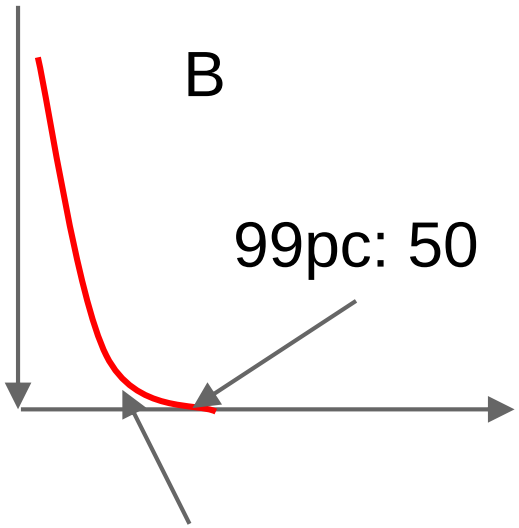


(SemiSpace is one of the best)

Have you ever been fooled by a lucky outlier?

Inappropriate data analysis



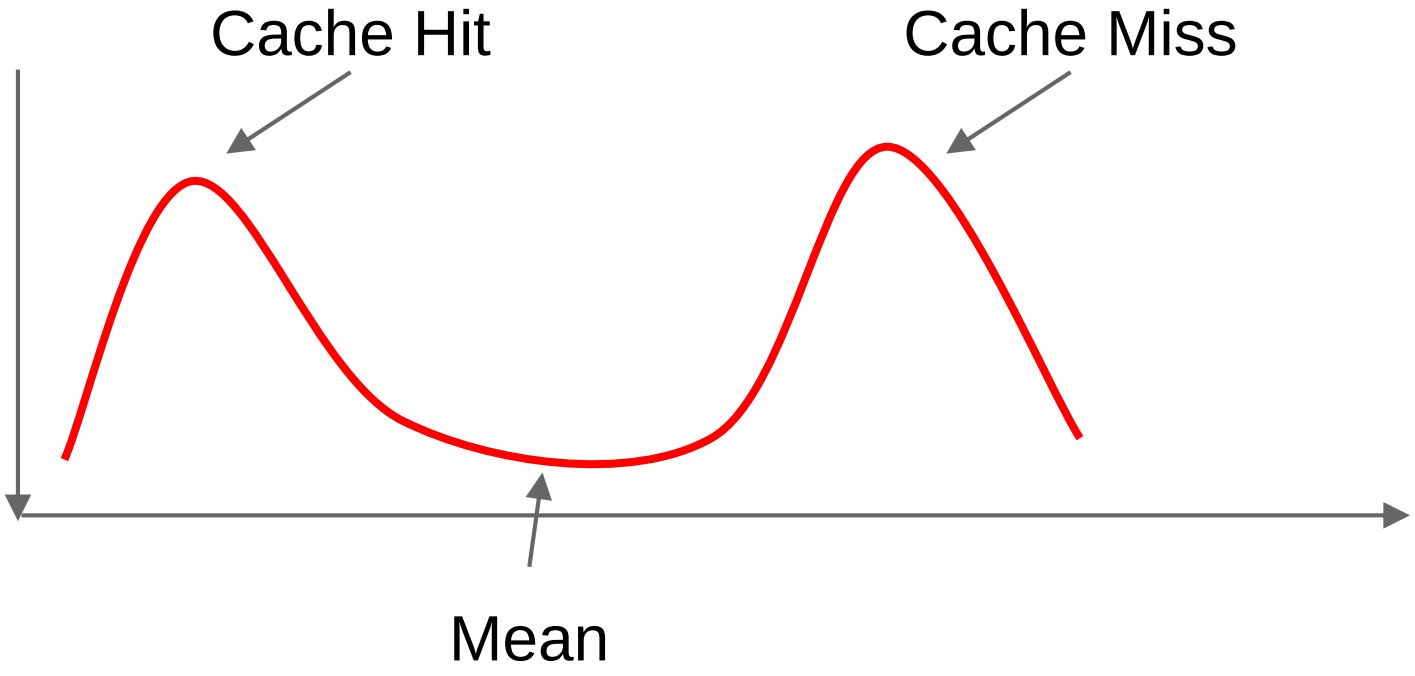


A single Google search = 100s of RPCs
99th percentile affects a majority of the requests!

A mean is inappropriate if long-tail latency matters!

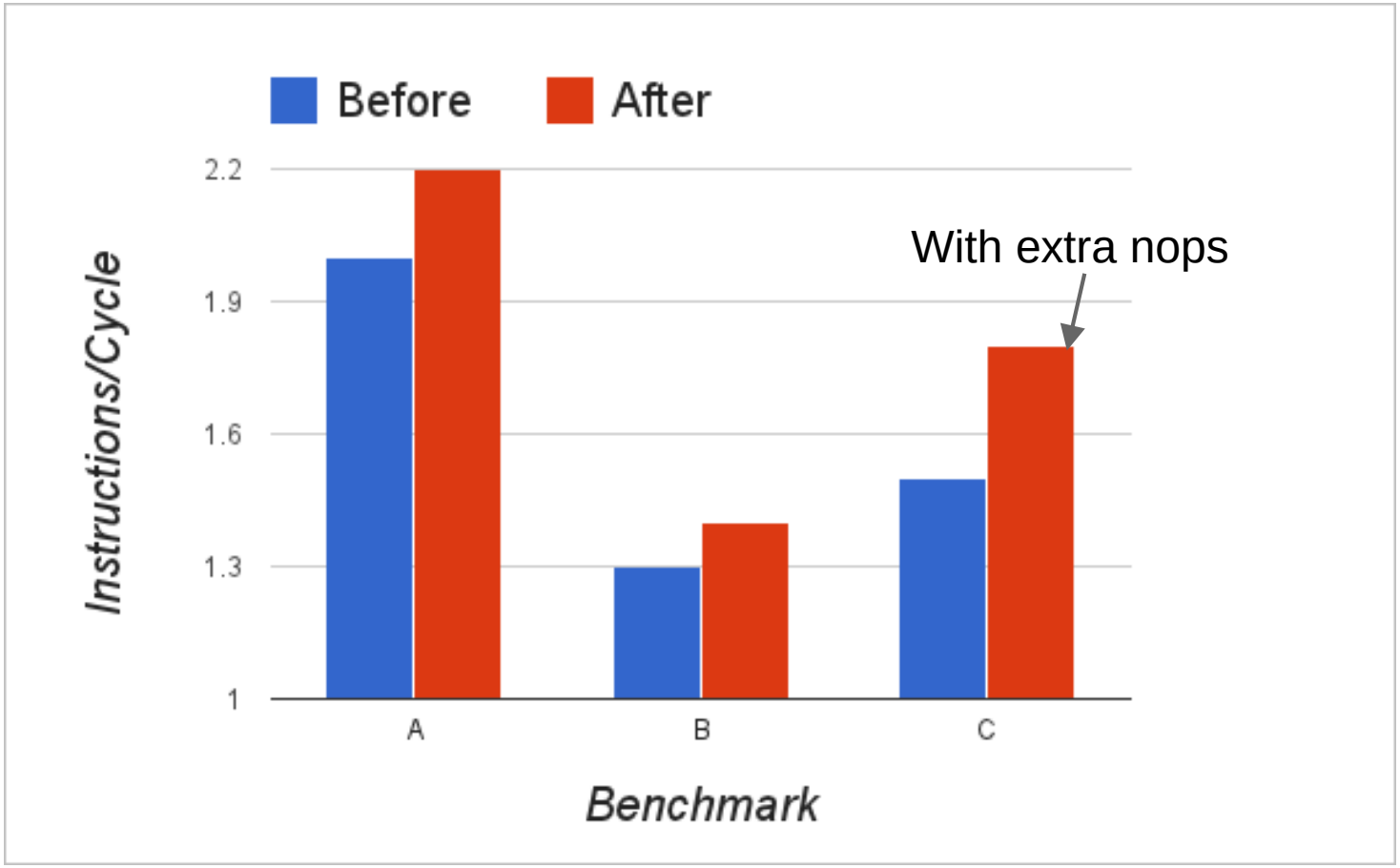
Inappropriate data analysis

Layered systems often use caches at each level:



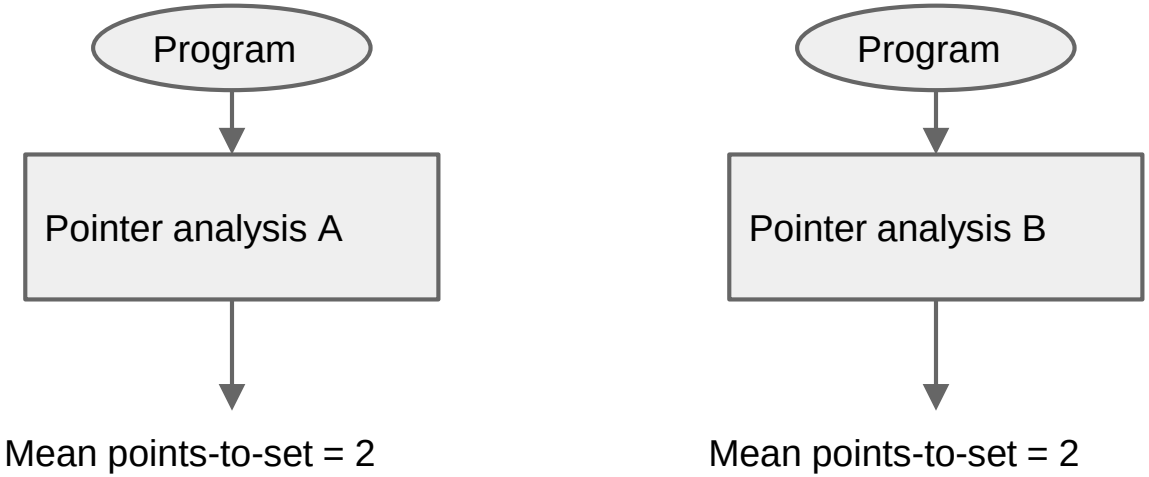
Do you check the shape of your data before summarizing it?

Inappropriate metric

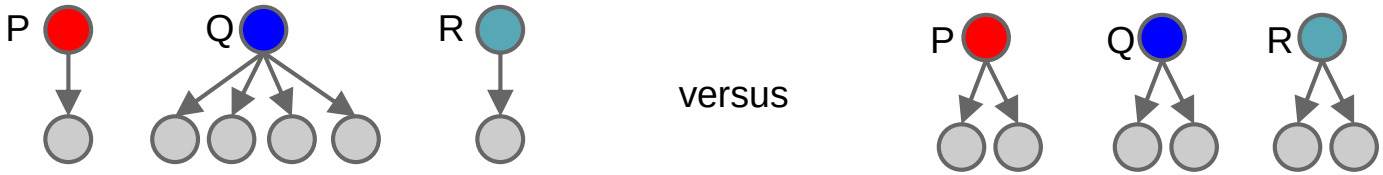


Have you ever picked a metric that was not ends-based?

Inappropriate metric



Claim: B is simpler yet just as precise as A



Have you ever used a metric that was inconsistent with "better"?

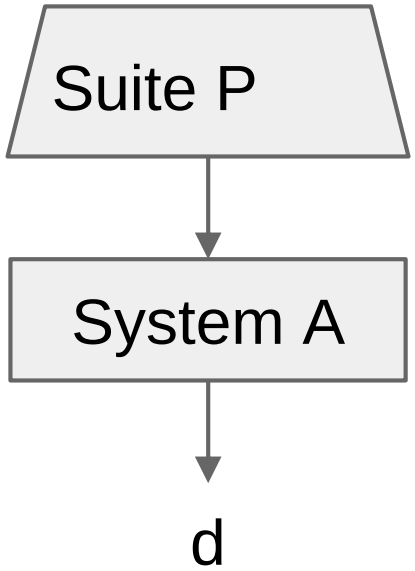
Sin 3: Inconsistency

Defn: *Experiment compares A to B in different contexts*

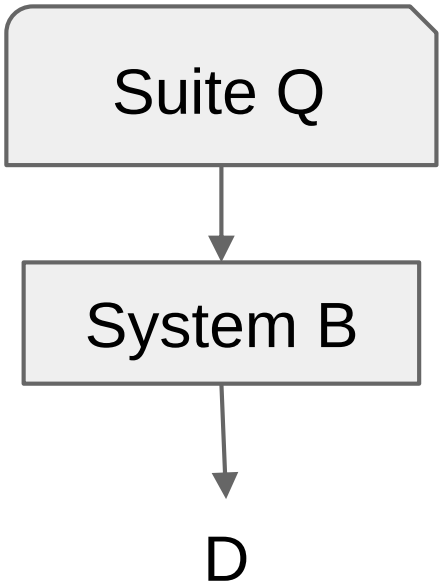
Sin 3: Inconsistency

Defn: *Experiment compares A to B in different contexts*

Experiment:
They used P; We used Q

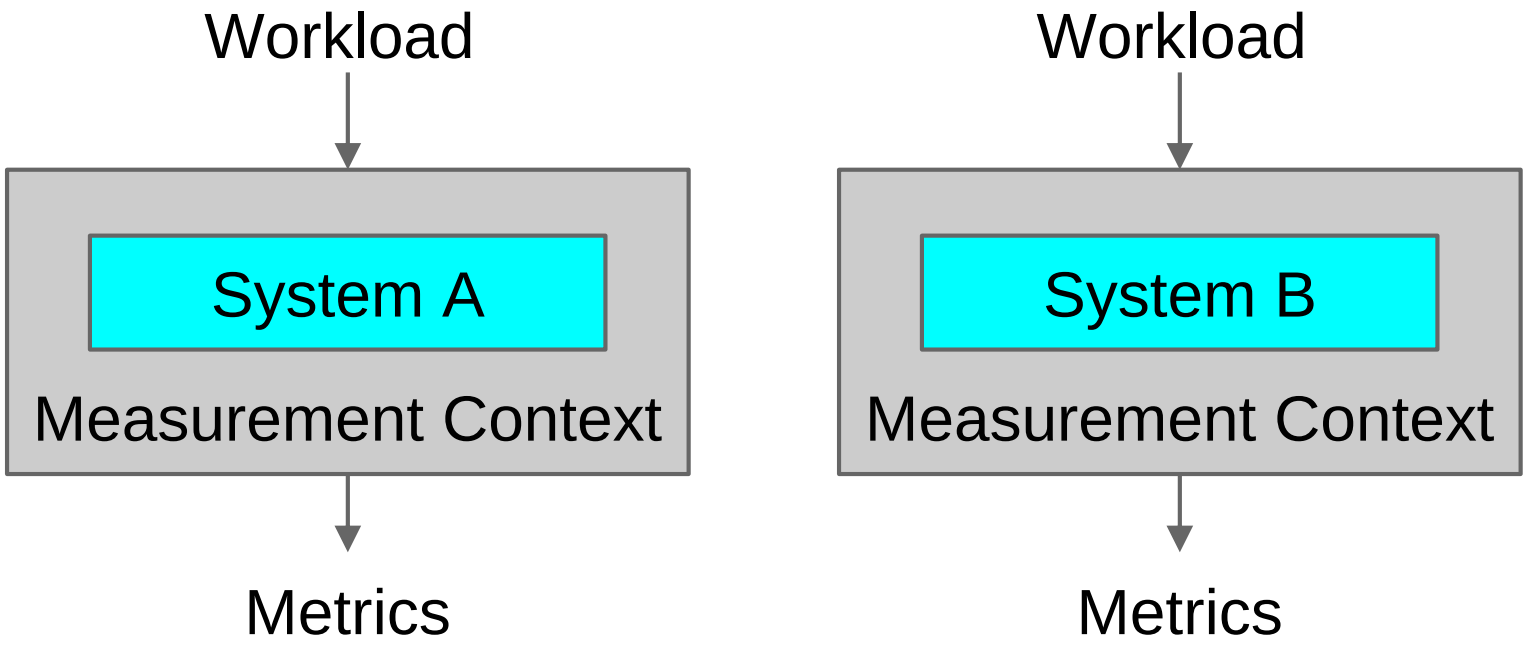


Claim:
B > A



Inconsistency misleads!

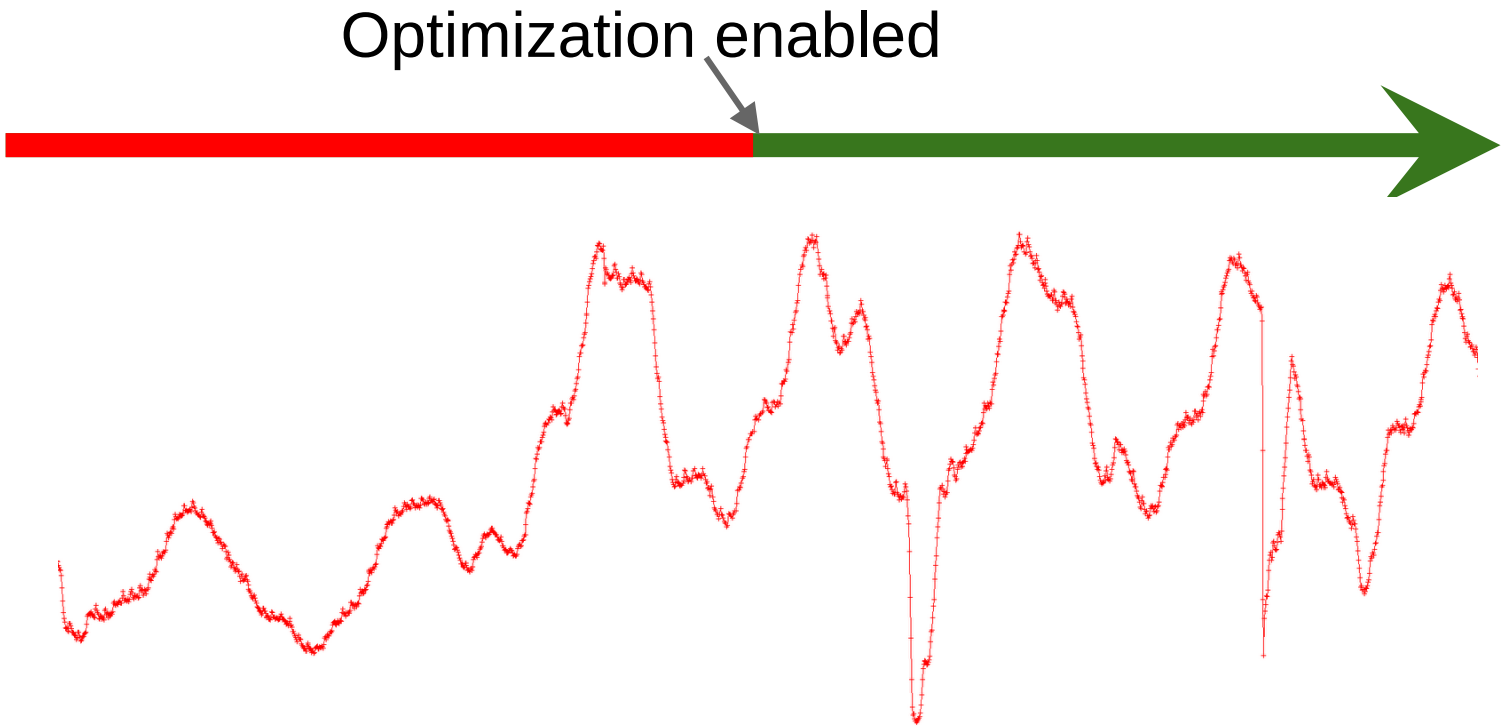
Inconsistency is not obvious



Workload, context, and metrics must be the same

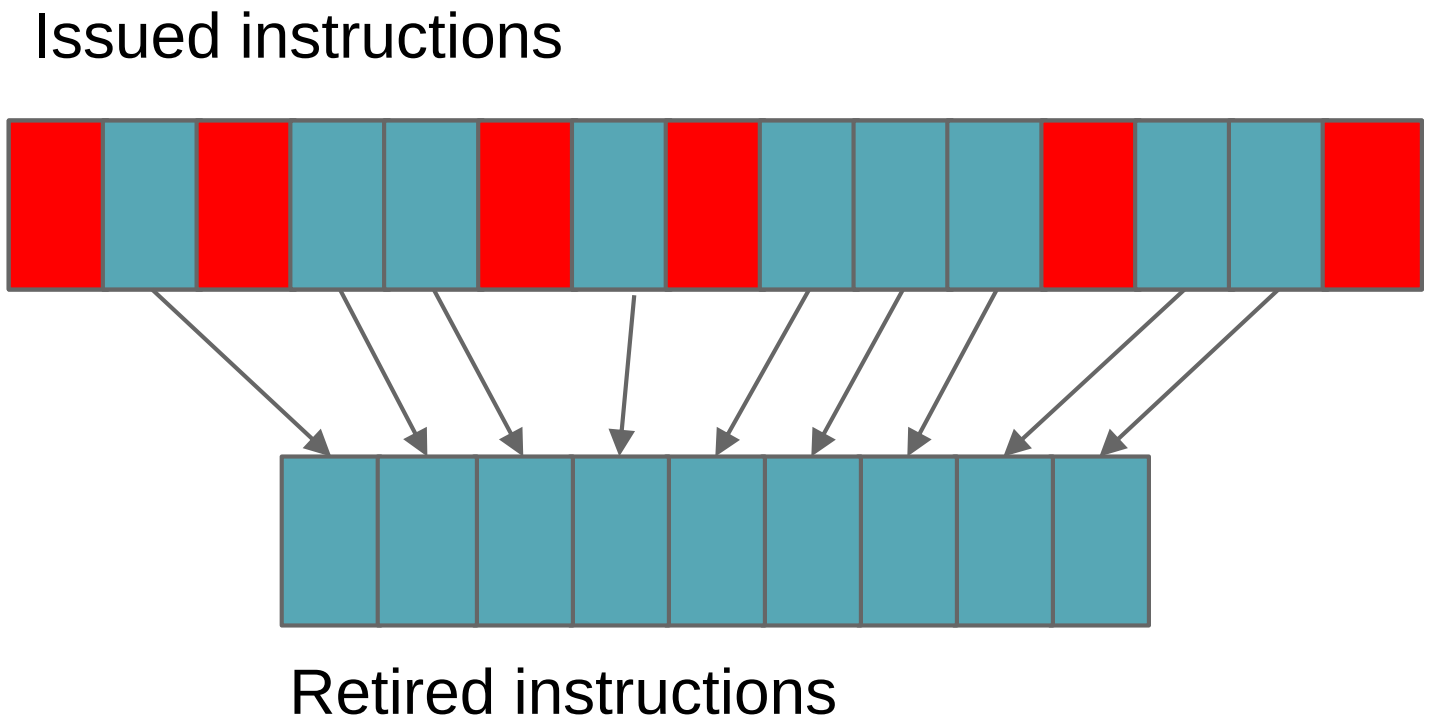
Inconsistent workload

I want to evaluate a new optimization for Gmail



Has the workload ever changed from under you?

Inconsistent metric

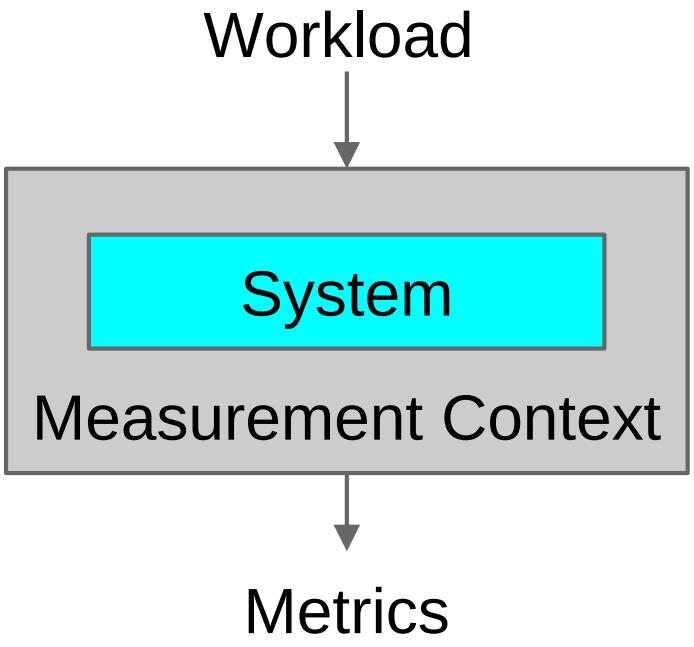


Do you (or even vendors) know what each hardware metric means?

Sin 4: Irreproducibility

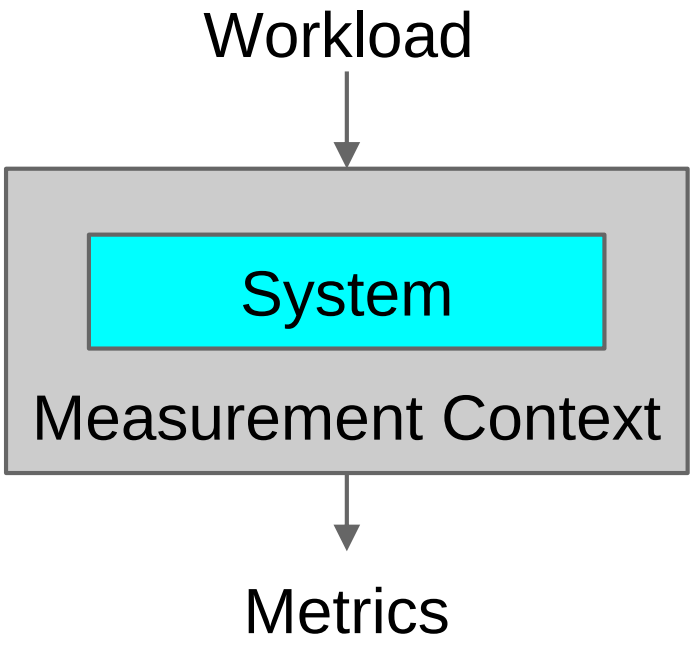
Defn: *Others cannot reproduce your experiment*

Experiment:



Irreproducibility makes it harder to identify unsound experiments

Irreproducibility is not obvious



Omitting any biases can make results irreproducible

Revisiting the thesis

The four fatal sins

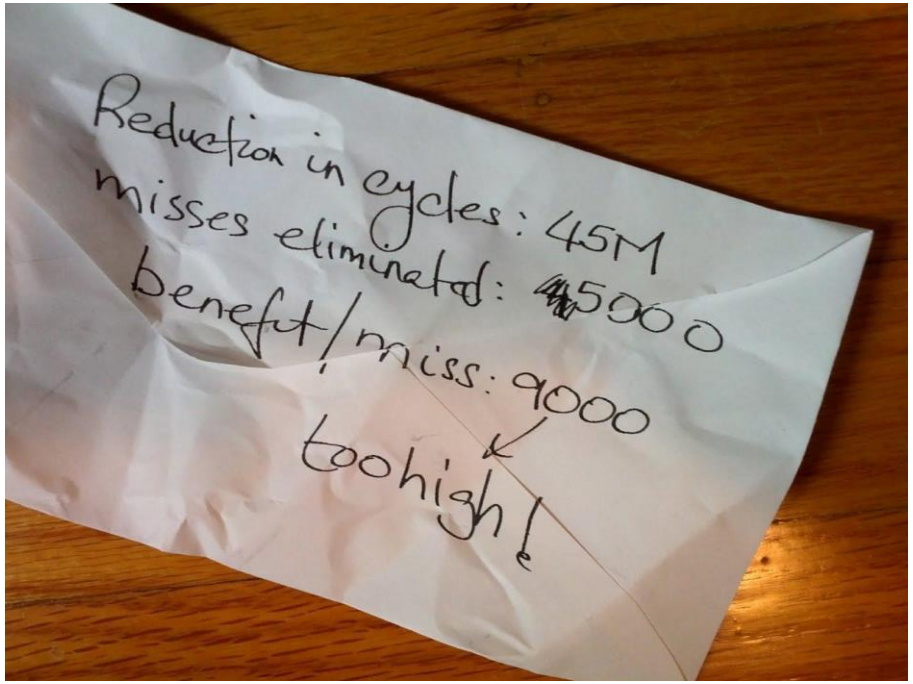
- affect all aspects of experiments
- cannot be eliminated with a silver bullet
  - (even with a much longer history, other sciences have them too)

It will take creativity and diligence to overcome these sins!

But I can give you one tip

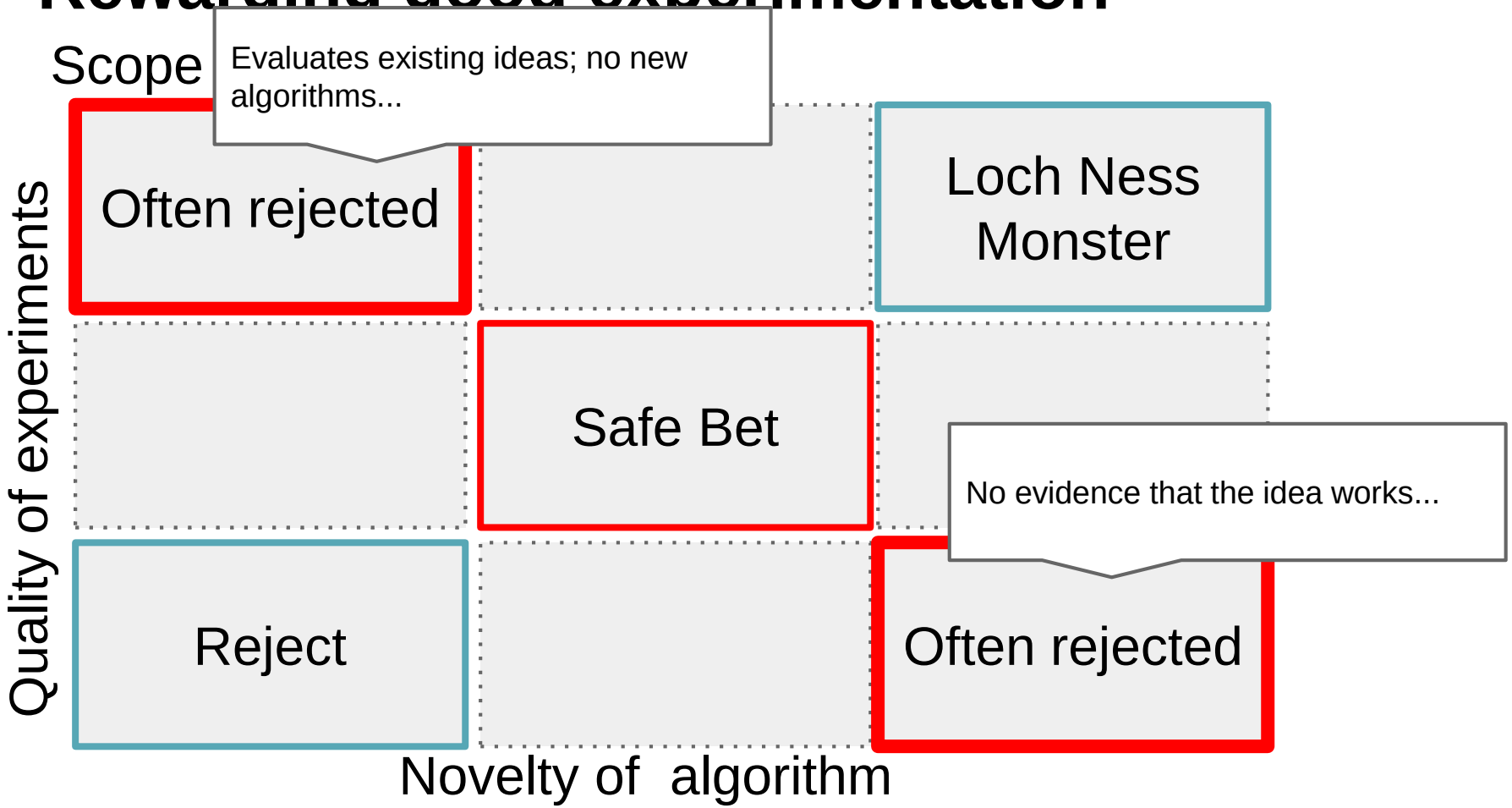
Look your gift horse in the mouth!

Back of the envelope



- Your optimization eliminates memory loads
  - Can the count of eliminated loads explain speedup?
- You blame "cache effects" for results you cannot explain...
  - Does the variation in cache misses explain results?

Rewarding good experimentation



Is this where we want to be?

Novel ideas can stand on their own

Novel (and carefully reasoned) ideas expose

- New paths for exploration
- New ways of thinking

A groundbreaking idea and no evaluation
>> A groundbreaking idea and misleading evaluation

Insightful experiments can stand on their own!

An insightful experiment may

- Give insight into leading alternatives
- Opens up new investigations
- Increase confidence in prior results or approaches

An insightful evaluation and no algorithm
>> An insightful evaluation and a lame algorithm

But sound experiments take time!

But not as much as chasing a false lead for years...

How would you feel if you built a product

...based on incorrect data?

Do you prefer to build upon:

Why you should care (revisited)

- Has your optimization ever yielded an improvement
  - ...even when you had not enabled it?
- Have you ever obtained fantastic results
  - ...which even your collaborators could not reproduce?
- Have you ever wasted time chasing a lead
  - ...only to realize your experiment was flawed?
- Have you ever read a paper
  - ...and immediately decided to ignore the results?

The end

- Experiments are difficult and not just for us
  - Jonah Lehrer's "*The truth wears off*"
- Other sciences have established methods
  - It is our turn to learn from them and establish ours!
- Want to learn more?
  - The Evaluate collaboratory (http://evaluate.inf.usi.ch)

Acknowledgements

- Todd Mytkowicz
- Evaluate 2011 attendees: José Nelson Amaral, Vlastimil Babka, Walter Binder, Tim Brecht, Lubomír Bulej, Lieven Eeckhout, Sebastian Fischmeister, Daniel Frampton, Robin Garner, Andy Georges, Laurie J. Hendren, Michael Hind, Antony L. Hosking, Richard E. Jones, Tomas Kalibera, Philippe Moret, Nathaniel Nystrom, Victor Pankratius, Petr Tuma
- My mentors: Mike Hind, Kathryn McKinley, Eliot Moss