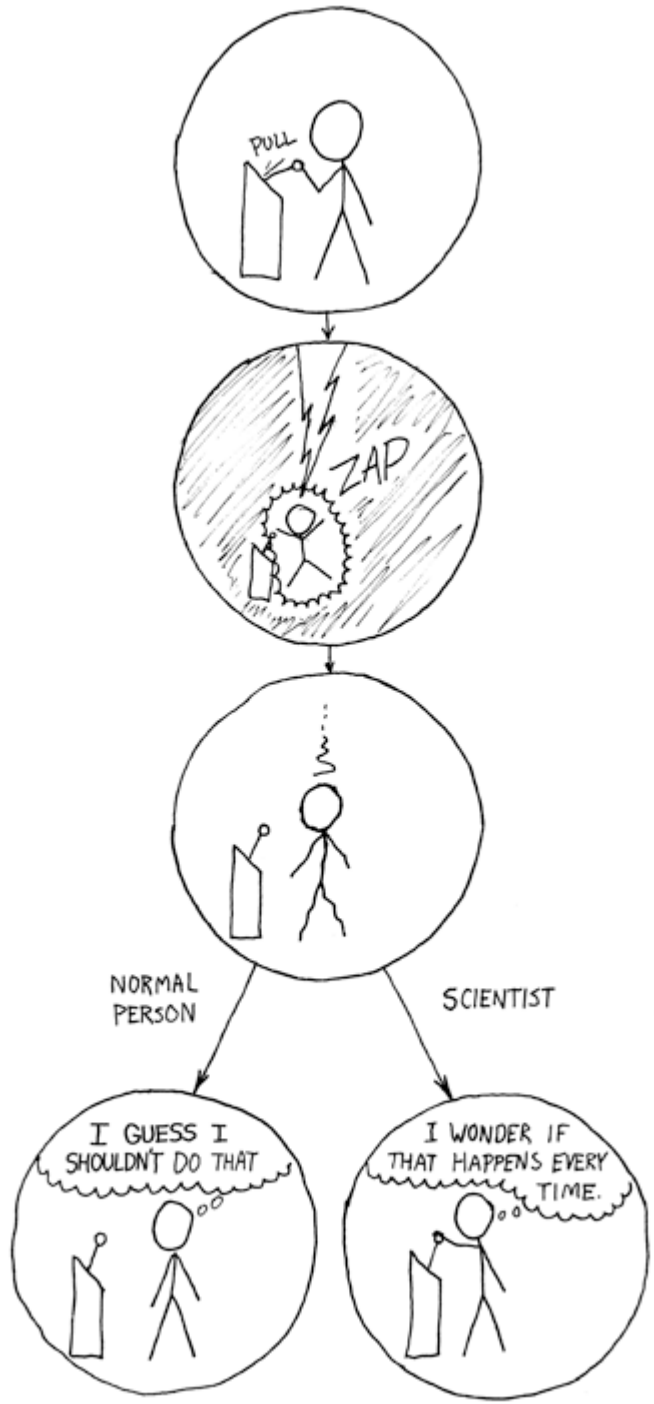
Replication and Recomputation in Scientific Experiments

Lars Kotthoff

05 August 2014

When theory meets practice


PULL
ZAP
NORMAL PERSON
SCIENTIST
I GUESS I SHOULDN'T DO THAT
I WONDER IF THAT HAPPENS EVERY TIME.

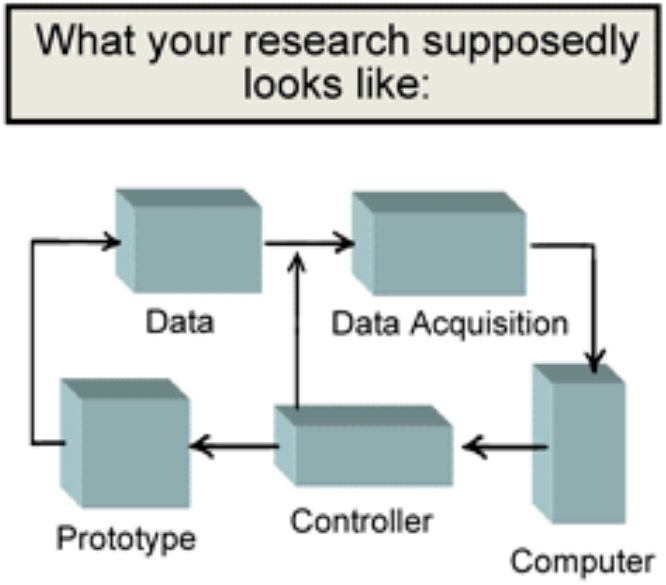


Figure 1. Experimental Diagram

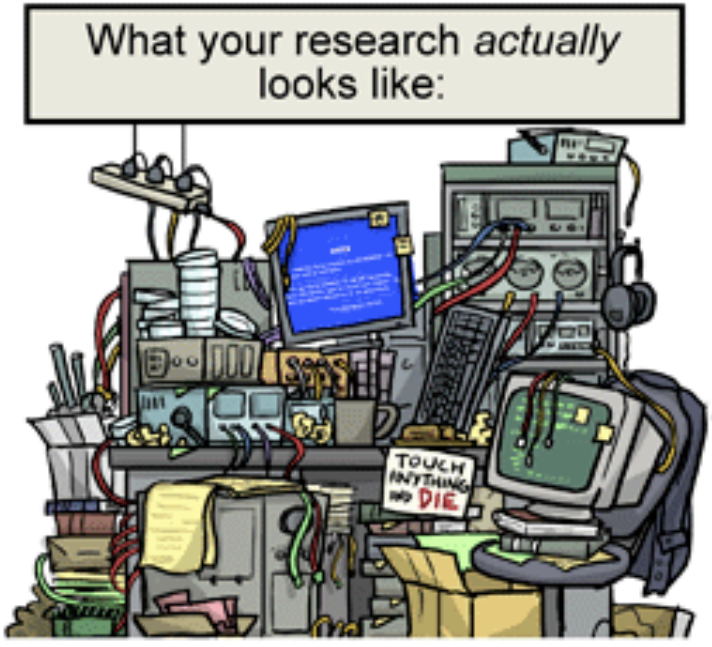


Figure 2. Experimental Mess

Recomputable?

What does it mean to be "recomputable"?

What does it mean to be “recomputable”?

- ability to rerun experiments
- ability to get the same results
- ability to get the same behaviour

Same behaviour

- often the “behaviour” is unspecified and only known through the results
- how the results are achieved is secondary
- in practice not necessary to replicate behaviour

Same results

- experiments often stochastic/randomised
- results may require analysis/aggregation
- in practice often enough to get qualitatively similar results

Rerun

- non-standard equipment or lots of resources may be necessary
- if it runs, it doesn't mean that it will finish
- cannot guarantee that everything will run all the time
- How do I know it's right?

Pragmatic approach

- work with authors of experiments
- they say how to “recompute”
- they can control/check results

Experiences

- recomputation as tutorial at CP 2013 and ECAI 2014
- invite authors of accepted papers to participate and make their experiments recomputable
- at CP by us, at ECAI with our help

Experiences

- 11 expressions of interest and 6 recomputable experiments for CP 2013
- $\approx$15 expression of interest and so far 1 recomputable experiment for ECAI 2014
- lots of enthusiasm, but can be a lot of work to get done

Overhead

- some authors have everything "ready to go"
- others use specific setup that's hard to replicate
- . . . and people in between

Challenges: Getting it to work

- need to replicate environment experiments were run in
- operating system, libraries, ...
- not specified in paper

Challenges: Getting it to work

- experiments may rely on specific software versions
- these may not be available anymore
- excellent motivation for our approach

Challenges: Getting it to work

- level of detail in paper not enough to run experiments
- parameters, processing of output, etc
- need to elicit this information from busy authors

Challenges: Resources

"So, you will need approximately 50.000.000 seconds [1.6 years] of computation on 2.27 GHz processors, memory limit of 3.9 GB (timeout is 2h)."

Challenges: Resources

"It requires a GPU/CUDA card of a certain version. I don't know if virtual machines work."

Challenges: Resources

"[the script] delegates the theorem proving tasks to Geoff Sutcliffe's automated theorem proving servers in Miami."

Challenges: Resources

"My experiments are based on games that humans played with/against a computer-agent in lab conditions. In order to run these experiments, it's needed to have a research assistant in the lab that will take care on the process of the experiment [...]. My experiments were done only by students from specific departments and mean age."

Challenges: Resources

- experiments may require specific type/number of CPUs, specific amount of memory
- cannot guarantee this
- may be unable to recompute because of this

Challenges: Licenses

- some experiments use proprietary software such as IBM CPLEX
- some authors use other people's tools that they don't own
- cannot include in virtual machine image and distribute
- solution: allow user to provide relevant software on host machine

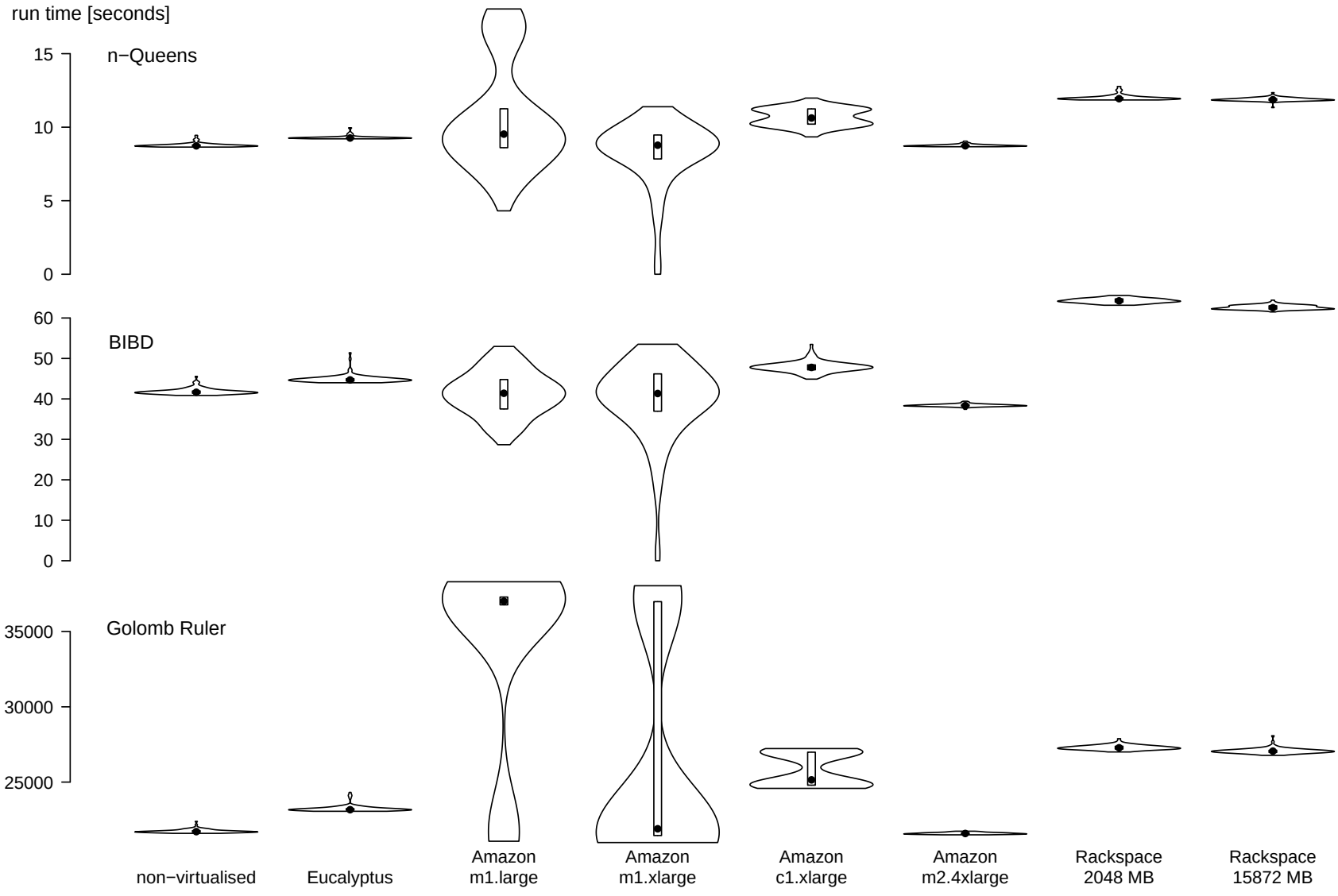
Challenges: time measurements

- virtualised clocks are less reliable than non-virtualised
- no control over how virtualised CPU runs on host CPU
- can we measure CPU time accurately on virtualised hardware?

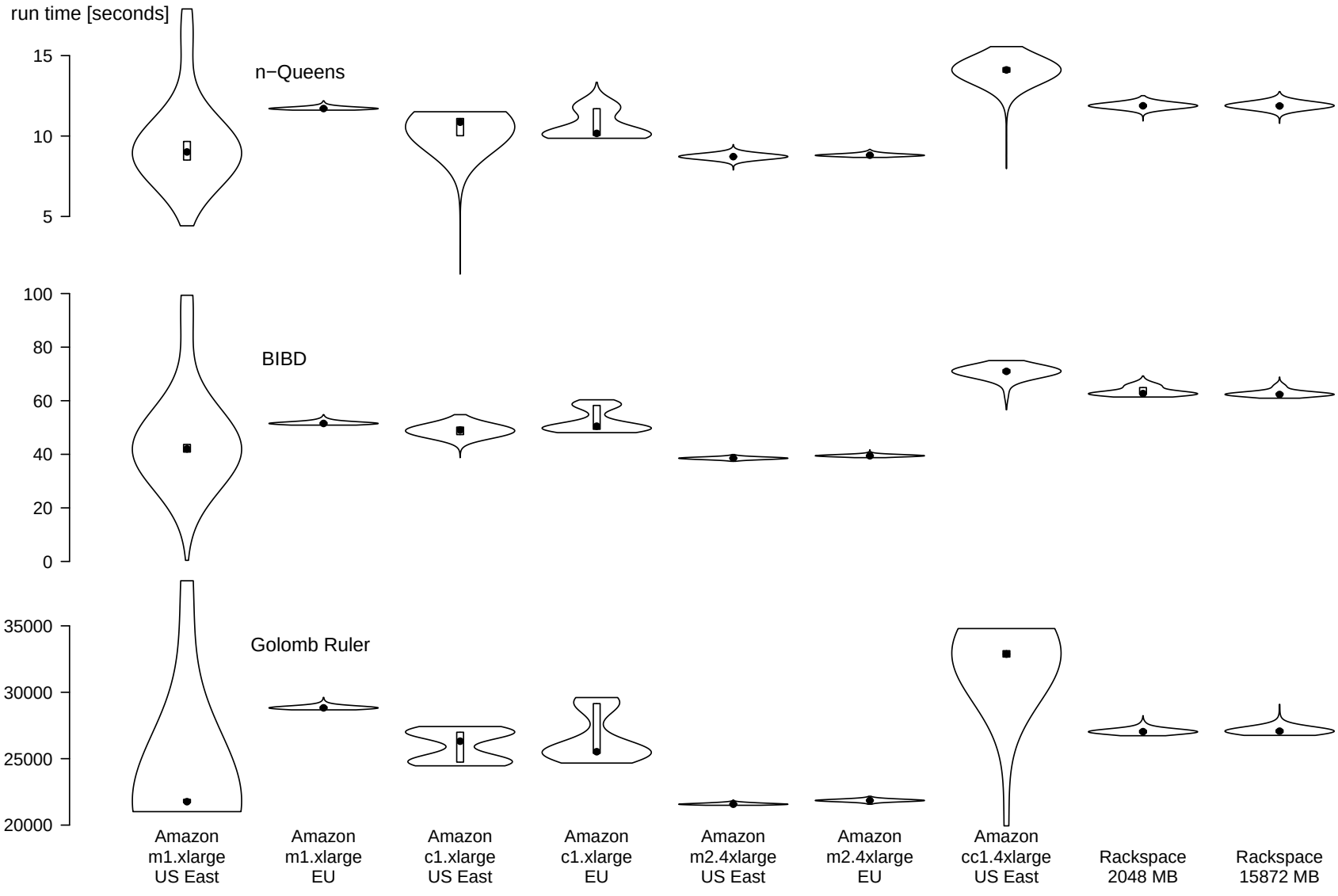
Challenges: time measurements

- three C(S|O)P – n-Queens (seconds), BIBD (minute), Golomb Ruler (hours)
- 10 repetitions
- run on non-virtualised hardware, private cloud, Amazon cloud, Rackspace cloud
- Kotthoff, Lars. "Reliability of Computational Experiments on Virtualised Hardware." Journal of Experimental and Theoretical Artificial Intelligence 26, no. 1 (2014): 3349.

Challenges: time measurements



Challenges: time measurements



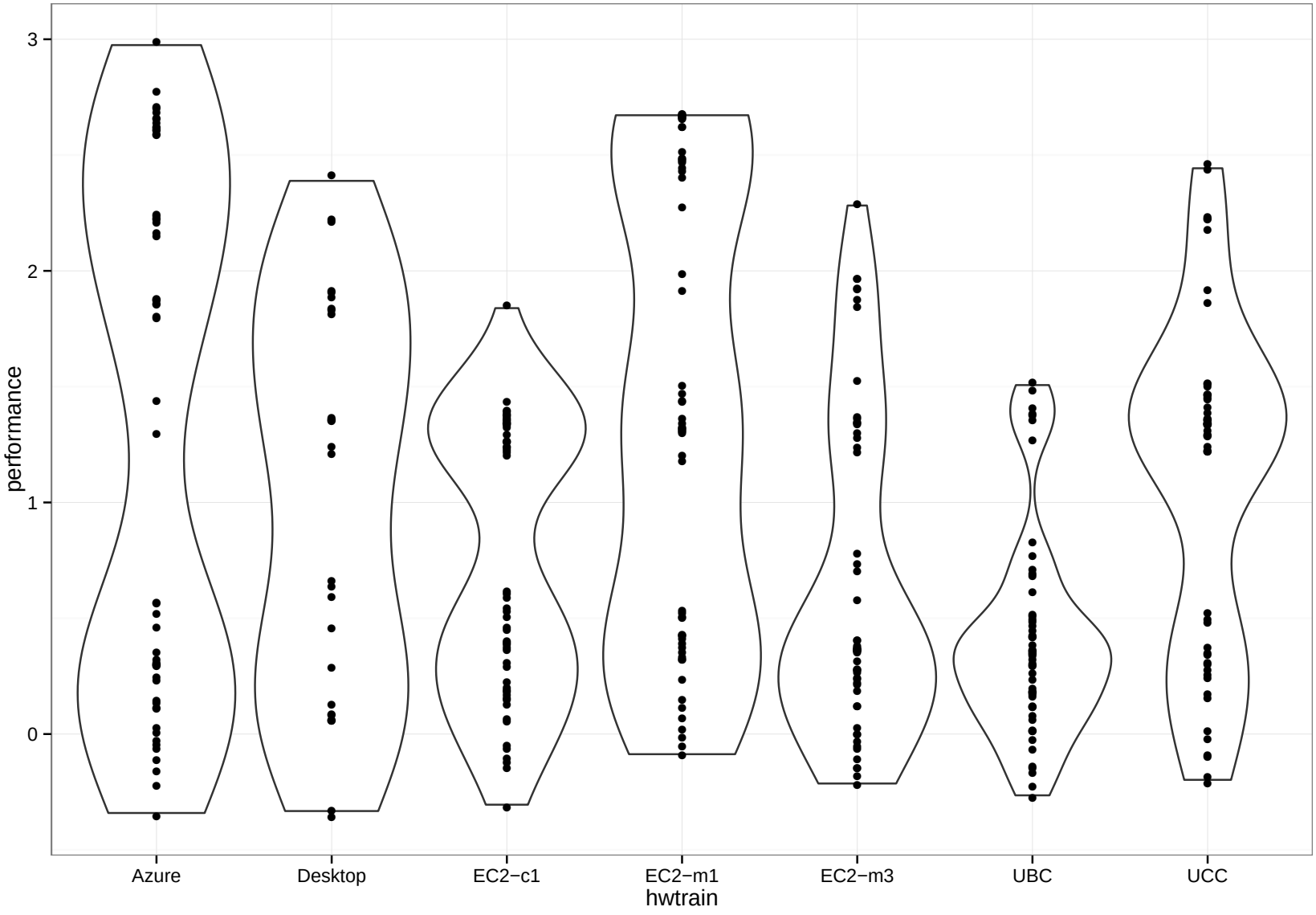
Challenges: performance measurements

- only half the story
- rarely evaluate algorithms in isolation
- do algorithms behave the same in relation to each other?

Challenges: performance measurements

- harder scenario: algorithm configuration
- configure on virtualised, test on non-virtualised hardware
- Daniel Geschwender, Frank Hutter, Lars Kotthoff, Yuri Malitsky, Holger H. Hoos, and Kevin Leyton-Brown. “Algorithm Configuration in the Cloud: A Feasibility Study.” In LION 8, 2014.

Challenges: performance measurements



Challenges: performance measurements

	Desktop	UBC	UCC	Azure	EC2-c1
Desktop	0.54 (0.67)	0.52 (0.76)	0.96 (0.46)	0.59 (0.68)	0.59 (0.57)
UBC	0.07 (0.21)	0.01 (0.11)	**0.17** (0.21)	0.22 (0.45)	**0.19** (0.18)
UCC	0.54 (0.51)	0.53 (0.52)	0.56 (0.09)	0.60 (0.07)	0.59 (0.61)
Azure	0.78 (1.14)	0.78 (1.11)	0.81 (1.03)	0.81 (1.02)	0.81 (1.00)
EC2-c1	0.53 (0.52)	0.16 (0.51)	0.59 (0.43)	0.58 (0.40)	0.26 (0.41)
EC2-m1	0.58 (0.99)	0.58 (1.01)	0.59 (0.93)	0.65 (0.92)	0.62 (0.85)
EC2-m3	**0.00** (0.55)	-**0.02** (0.59)	0.56 (0.51)	**0.18** (0.44)	0.30 (0.42)

	EC2-m1	EC2-m3	**median**
Desktop	0.80 (0.62)	0.59 (0.54)	0.59 (0.62)
UBC	0.19 (0.16)	**0.15** (0.31)	0.17 (0.21)
UCC	0.58 (0.42)	0.58 (0.42)	0.58 (0.42)
Azure	0.81 (1.01)	0.82 (0.99)	0.81 (1.02)
EC2-c1	0.22 (0.41)	0.55 (0.52)	0.53 (0.43)
EC2-m1	0.62 (0.88)	0.57 (0.89)	0.59 (0.92)
EC2-m3	**0.16** (0.46)	0.16 (0.42)	**0.16** (0.46)

30 second Vagrant recomputation tutorial

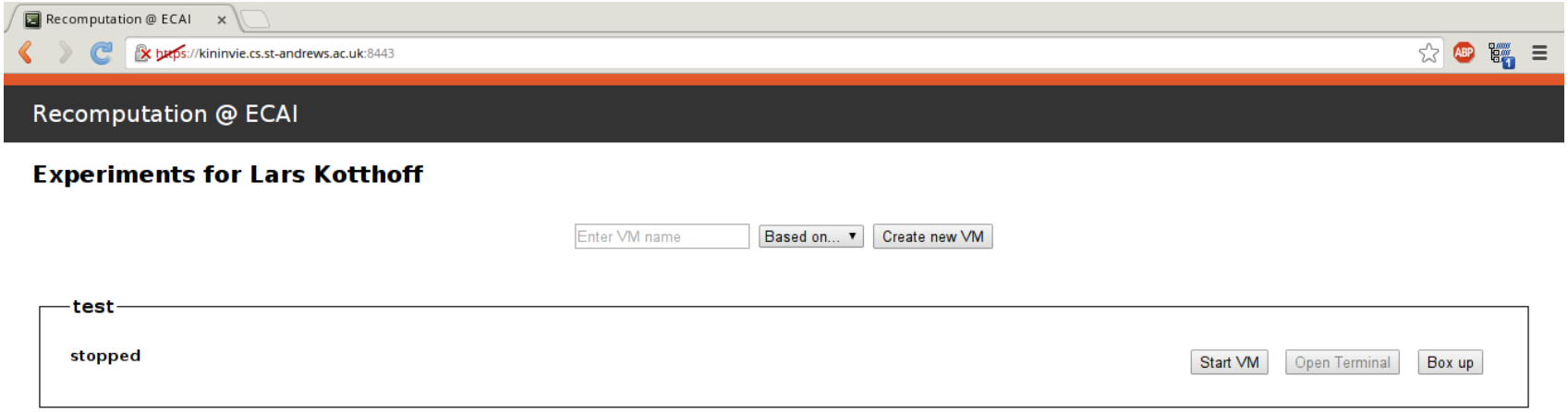
Do try this at home

```
vagrant init myExperiment \
    http://files.vagrantup.com/precise64.box
vagrant up
vagrant ssh
[set up virtual machine and experiments]
vagrant halt
vagrant package --output myExperiment.box
```

This is all you need.

Do try this at home – run automatically

```
vagrant init myExperiment \
    http://files.vagrantup.com/precise64.box
vagrant up
vagrant ssh
[set up, single script recompute.sh to run everything]
vagrant halt
[edit Vagrantfile, add line
config.vm.provision :shell, :inline => "./recompute.sh"
before "end"]
vagrant package --vagrantfile Vagrantfile \
    --output myExperiment.box
```


Recomputation @ ECAI
https://kininvie.cs.st-andrews.ac.uk:8443
Recomputation @ ECAI
Experiments for Lars Kotthoff
Enter VM name
Based on...
Create new VM
test
stopped
Start VM
Open Terminal
Box up

adapted from `http://dungnguyen.net/`