T. HOEFLER

Communication with the Reader – or how to deal with peer review?

Early Career Workshop at SC18

Why did the hosts invite Todd and me?

- **I guess because I'm the papers chair for SC18 … so let me give the expected half-talk ☺**

- **SC18 statistics**
  - Total submissions: 288 - 1,223 total reviews – 1,154 (95%) on time!

    Accept: 24 (8.3%)

    Minor revision: 31 (10.8%)

    Major revision: 15 (5.2%), eventually accept: 14 (4.9%)

    Total accept: 69 (24%)

  - Best paper finalists: 2
  - Best student paper finalists: 5

- **Total accept range: 19.1% - 24.3%**

What was new at SC18 – quite a lot (and will probably stay)

- Major changes with respect to SC17 (discuss each in the following minutes):
  1. Featherweight revisions
  2. Journal-style decision categories
  3. ML crosscut committee
  4. No deadline extensions
  5. No page limit for appendix

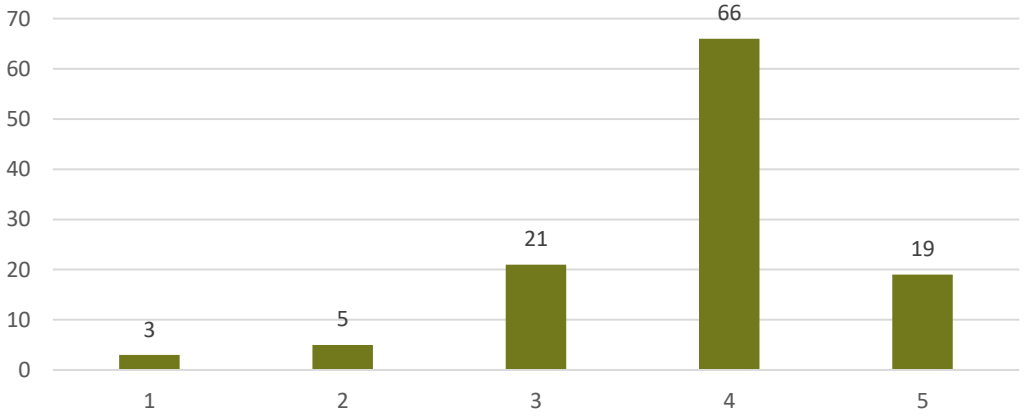
Featherweight Revisions

- **255 (92%) rebuttals and 246 (88%) featherweight revisions**
- **Our feeling: it worked well (many submissions, samples looked very good)**

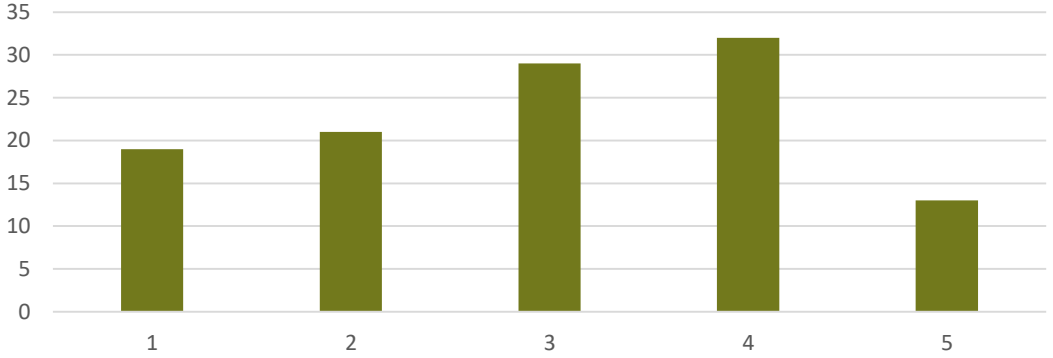
- **Pros:**
  - Many reviewers were happier after the revision (tracks)
  - Higher acceptance rates
  - Perceived higher quality

- **Cons:**
  - Peer-pressure
  - More work

Do you believe the featherweight revision improved the average quality of the papers?

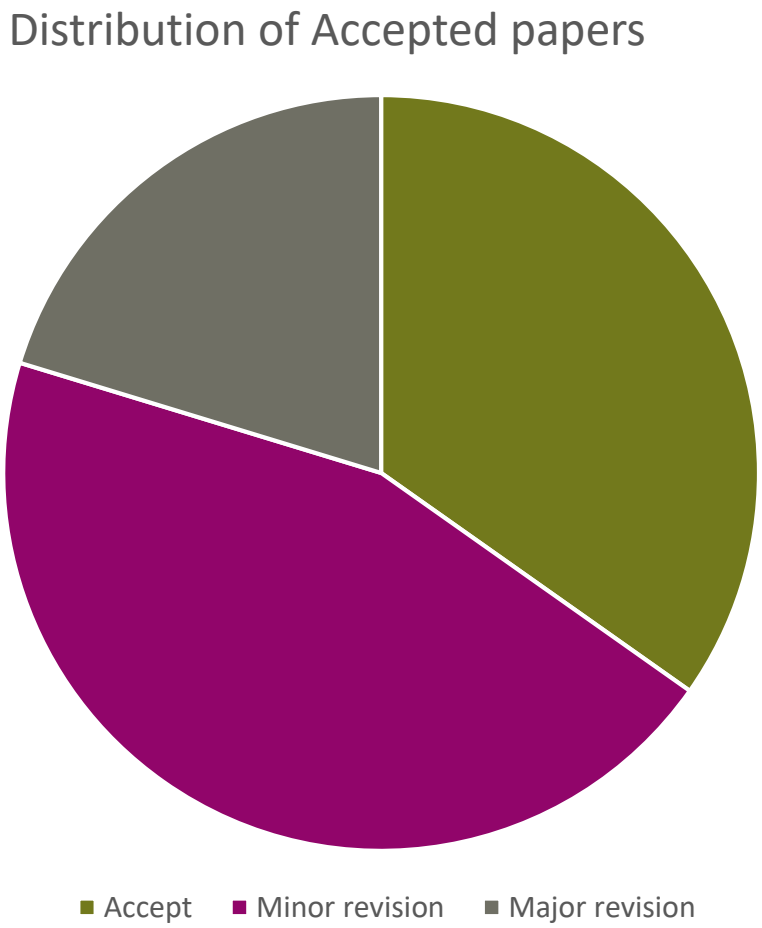


Consider the paper for which the score was most affected by the featherweight revision. By how much did its score change?



Major and Minor Revision

- **Minor revisions: 31/55 (56%)**
  - Up to shepherd whether it's accepted
    All were eventually accepted
- **Major revisions: 15**
  - Remain in competitive process
  - Were invited for full second review
    online discussion moderated by track chair
  - 14 accepts
- **Total accepts: 24 + 31 + 14**



What I actually want to tell you about how to get papers accepted ☺

Always remember the tree Qs

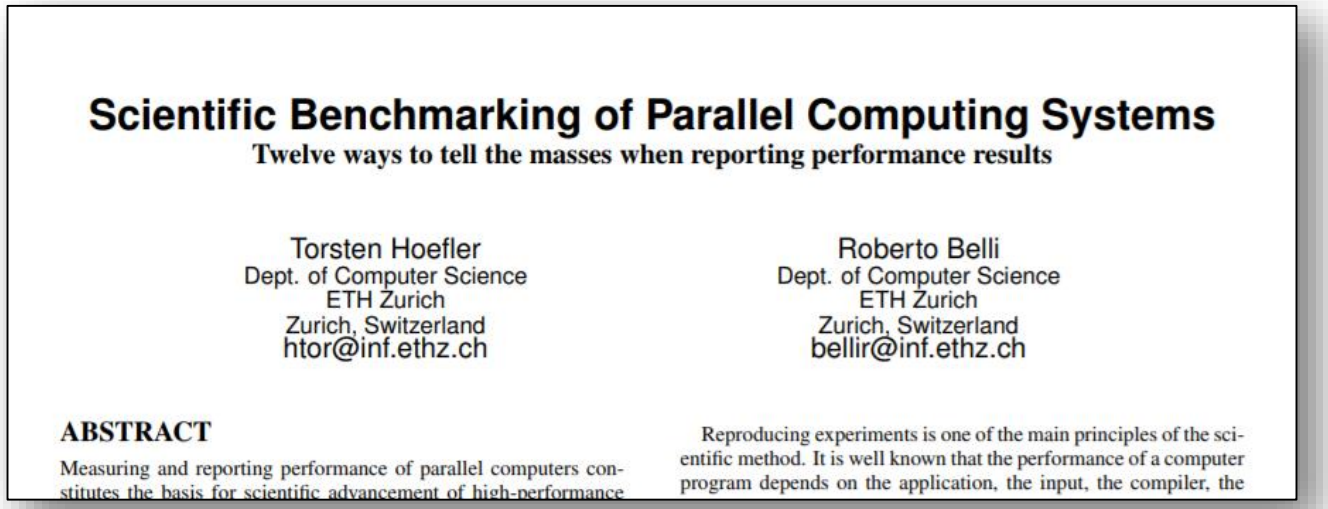
Quality, Quality, and Quality

But what does quality even mean?
Scientific Rigor!

Rigor enables scientific integrity! Interpretability as main tool in performance

- **Attempt to emphasize interpretability of performance experiments!**
  - Let's look at some basic statistics, for example

- **12 rules to improve the quality (rigor) of your paper!**
  - They are not complete and probably never will be
  - Intended to serve as a solid start
  - Call to the community (you!) to extend them



Scientific Benchmarking of Parallel Computing Systems

Twelve ways to tell the masses when reporting performance results

Torsten Hoefler
Dept. of Computer Science
ETH Zurich
Zurich, Switzerland
htor@inf.ethz.ch

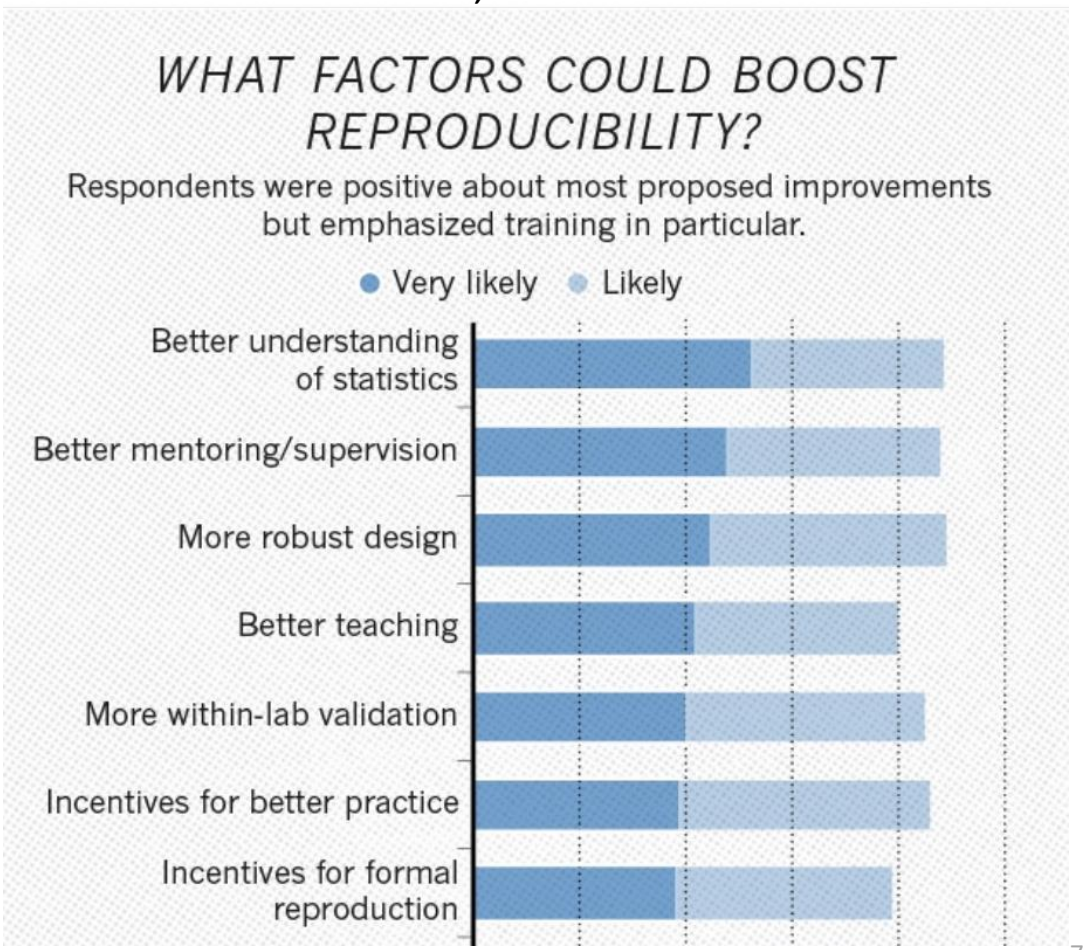
Roberto Belli
Dept. of Computer Science
ETH Zurich
Zurich, Switzerland
bellir@inf.ethz.ch

ABSTRACT

Measuring and reporting performance of parallel computers constitutes the basis for scientific advancement of high-performance

Reproducing experiments is one of the main principles of the scientific method. It is well known that the performance of a computer program depends on the application, the input, the compiler, the

Nature, 2016

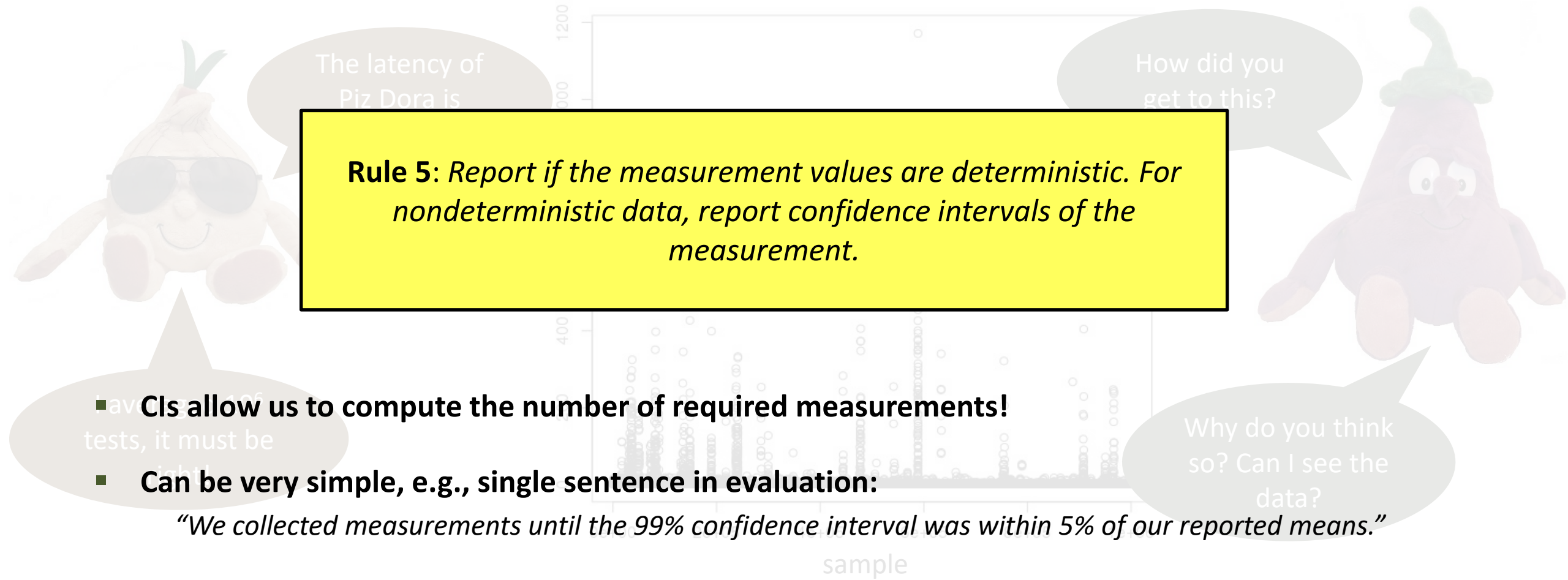


TH, Belli: Scientific Benchmarking of Parallel Computing Systems, IEEE/ACM SC15

A small technical detour
(I hope the organizers allow me, only 4 mins)

How to gain insights from rigorous benchmarking!

The simplest networking question: ping pong latency!

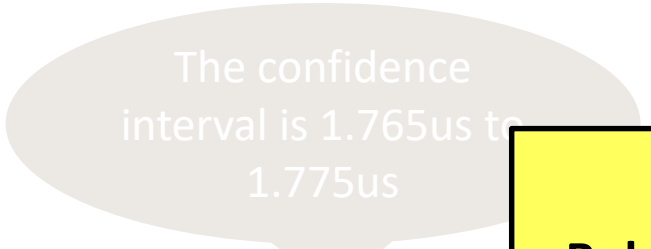


Rule 5: *Report if the measurement values are deterministic. For nondeterministic data, report confidence intervals of the measurement.*

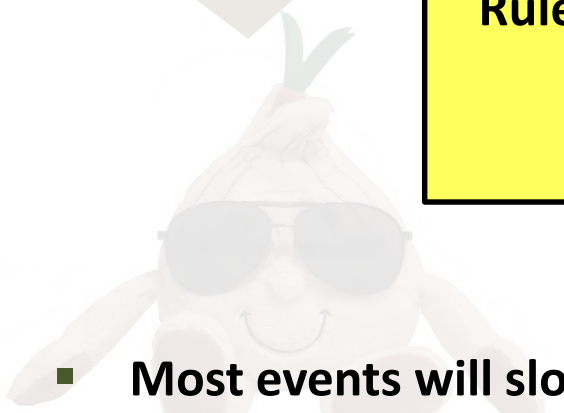
- **CIs allow us to compute the number of required measurements!**
- **Can be very simple, e.g., single sentence in evaluation:**

  "We collected measurements until the 99% confidence interval was within 5% of our reported means."

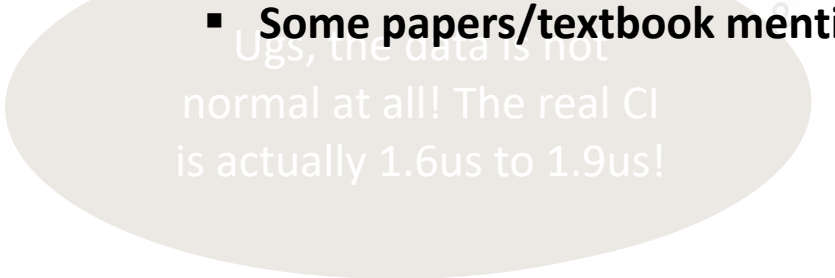
Thou shalt not trust your average textbook!



Rule 6: *Do not assume normality of collected data (e.g., based on the number of samples) without diagnostic checking.*



- **Most events will slow down performance**
  - **Heavy right-tailed distributions**
- **The Central Limit Theorem only applies asymptotically**
  - **Some papers/textbook mention "30-40 samples", don't trust them!**



Thou shalt not trust your system!

Look what data I got!

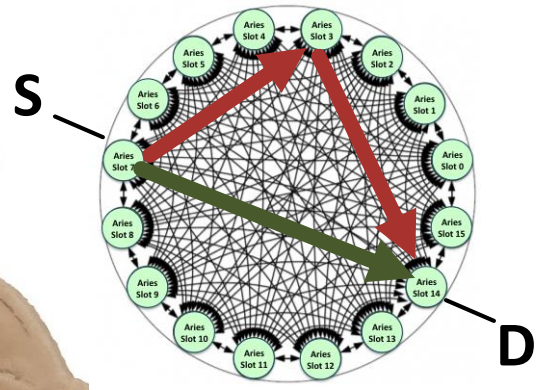
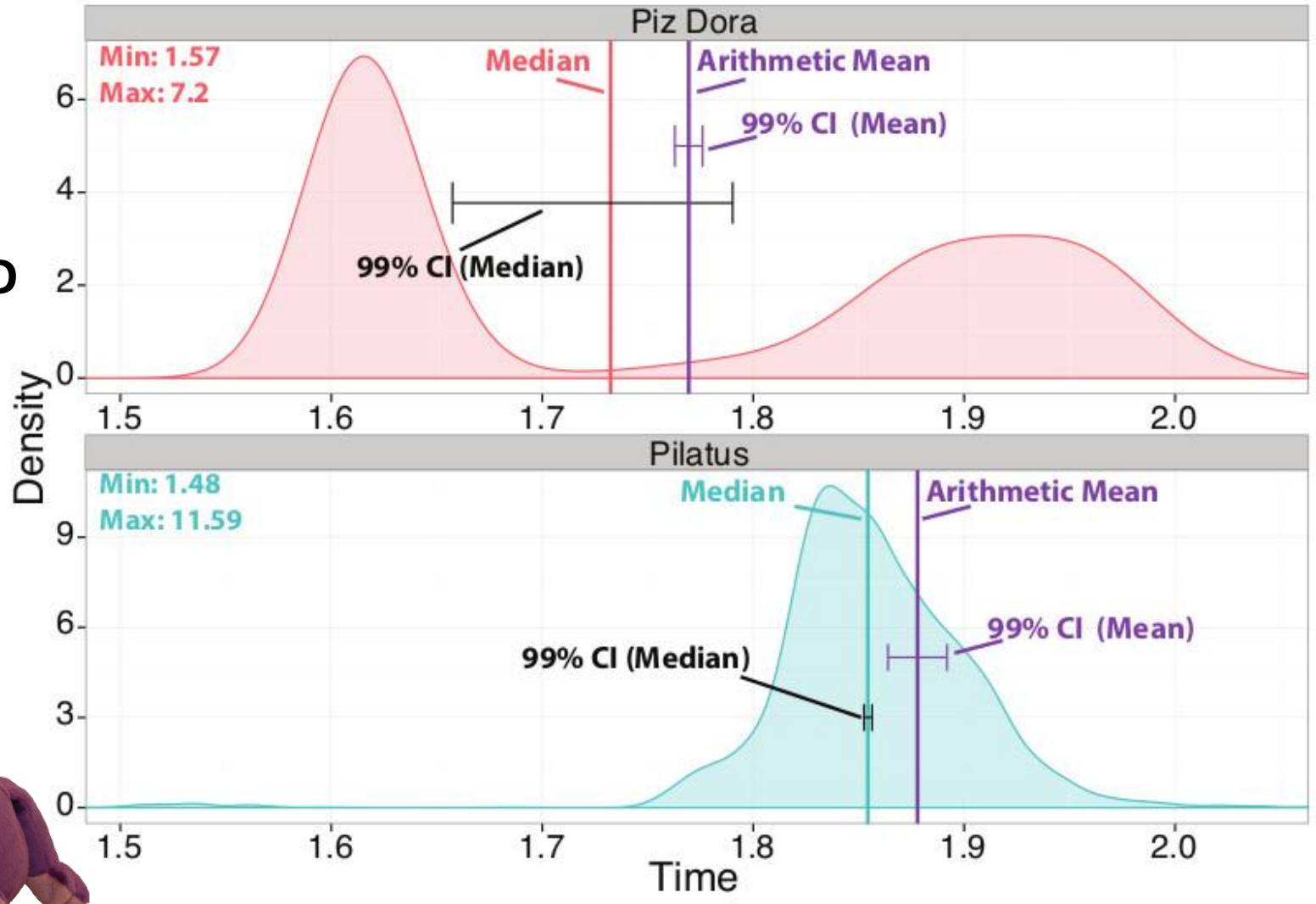
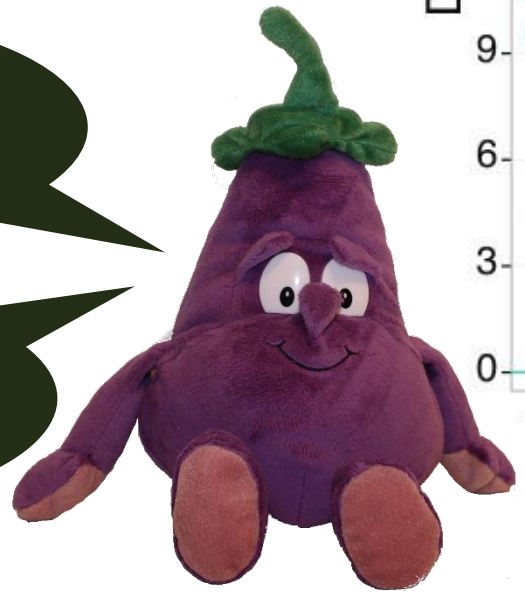


Image credit: nersc.gov



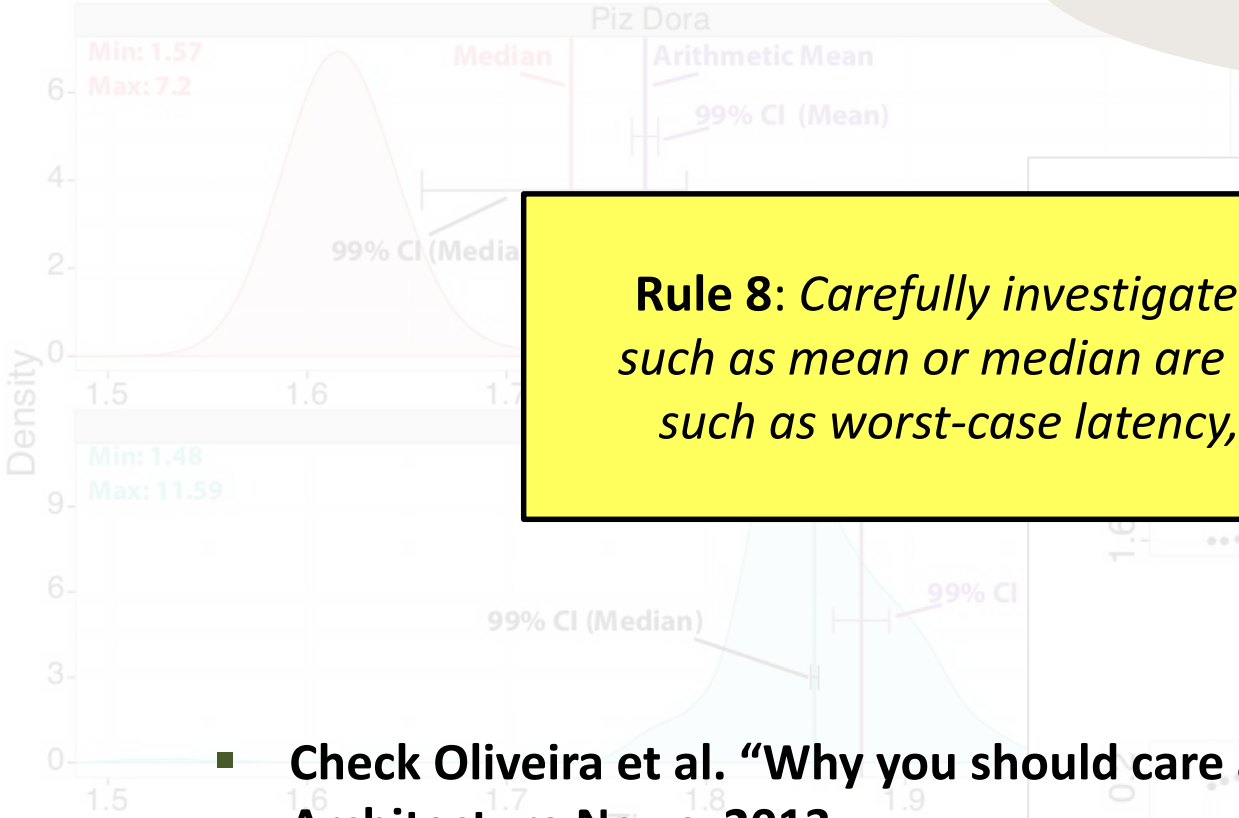
Clearly, the mean/median are not sufficient!

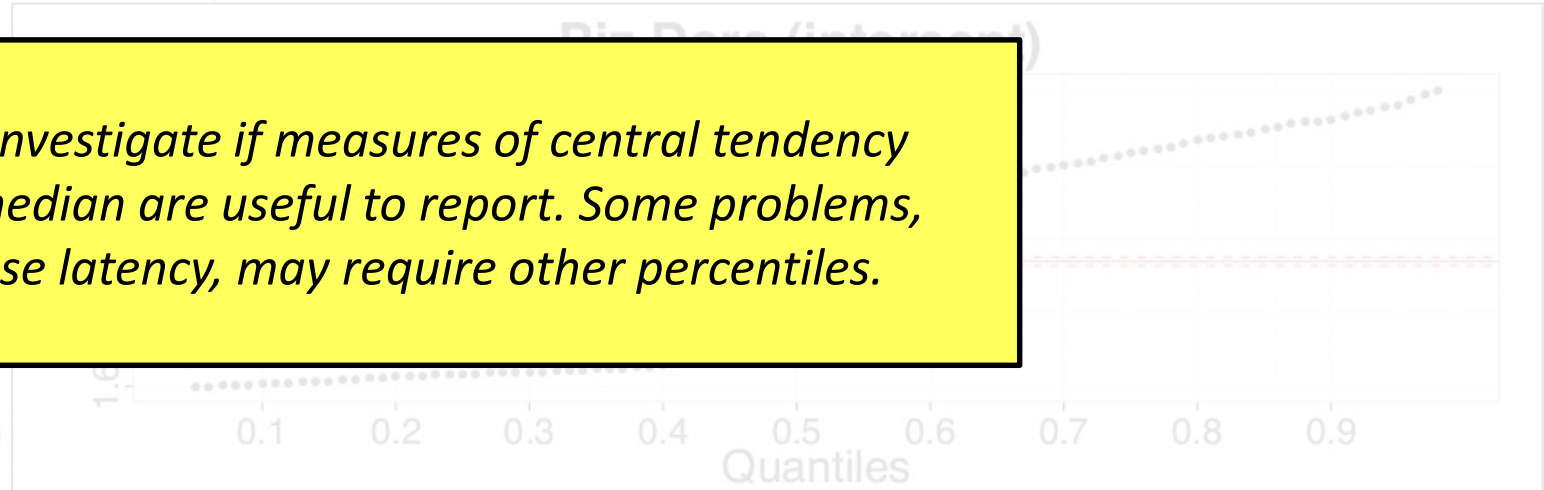
Try quantile regression!

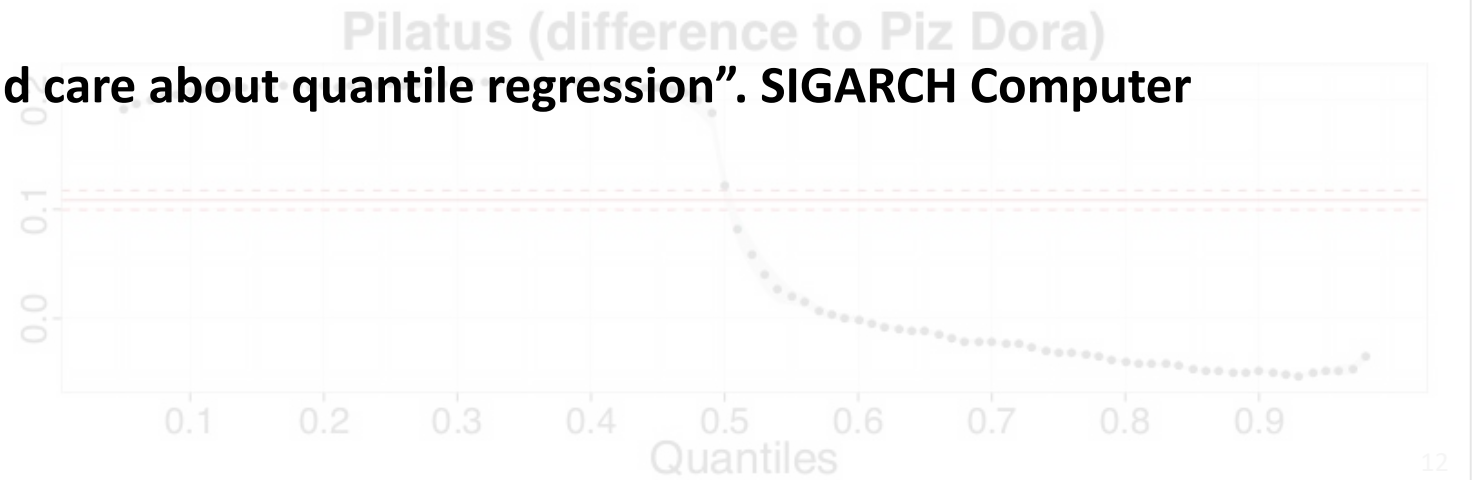


Quantile Regression

Wow, so Pilatus is better for (worst-case) latency-critical workloads even though Dora is expected to be faster







Rule 8: *Carefully investigate if measures of central tendency such as mean or median are useful to report. Some problems, such as worst-case latency, may require other percentiles.*

- **Check Oliveira et al. "Why you should care about quantile regression". SIGARCH Computer Architecture News, 2013.**

One last word … also pay attention to how to not do things ☺

- **You may be beginning something new --- trust the seniors, do not risk embarrassment**
  - And make sure to keep having fun!

“Twelve ways to fool the masses when reporting performance of deep learning workloads”
(my humorous guide to floptimize deep learning)