ETH zürich

spcl.inf.ethz.ch
@spcl_eth

D INFK

Torsten Hoefler

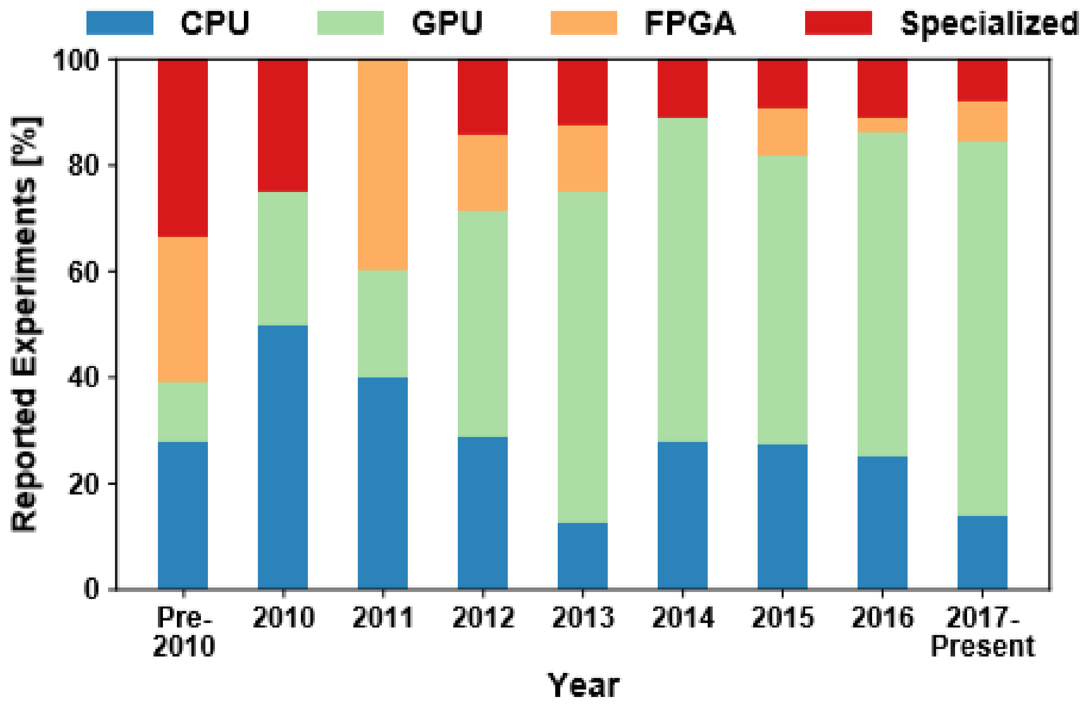
Deep500: An HPC Deep Learning Benchmark and Competition

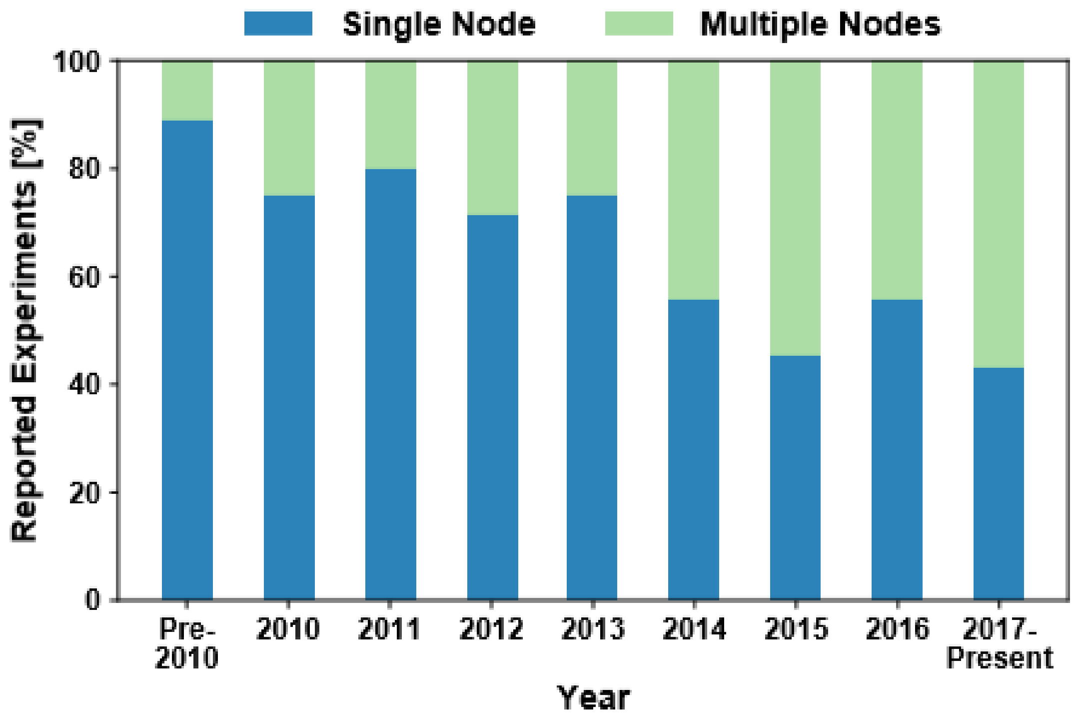
Birds of a Feather, SC18, Nov. 2018, Dallas, TX

SPCL

Trends in deep learning: hardware and multi-node

The field is moving fast – trying everything imaginable – survey results from 240 papers in the area of parallel deep learning



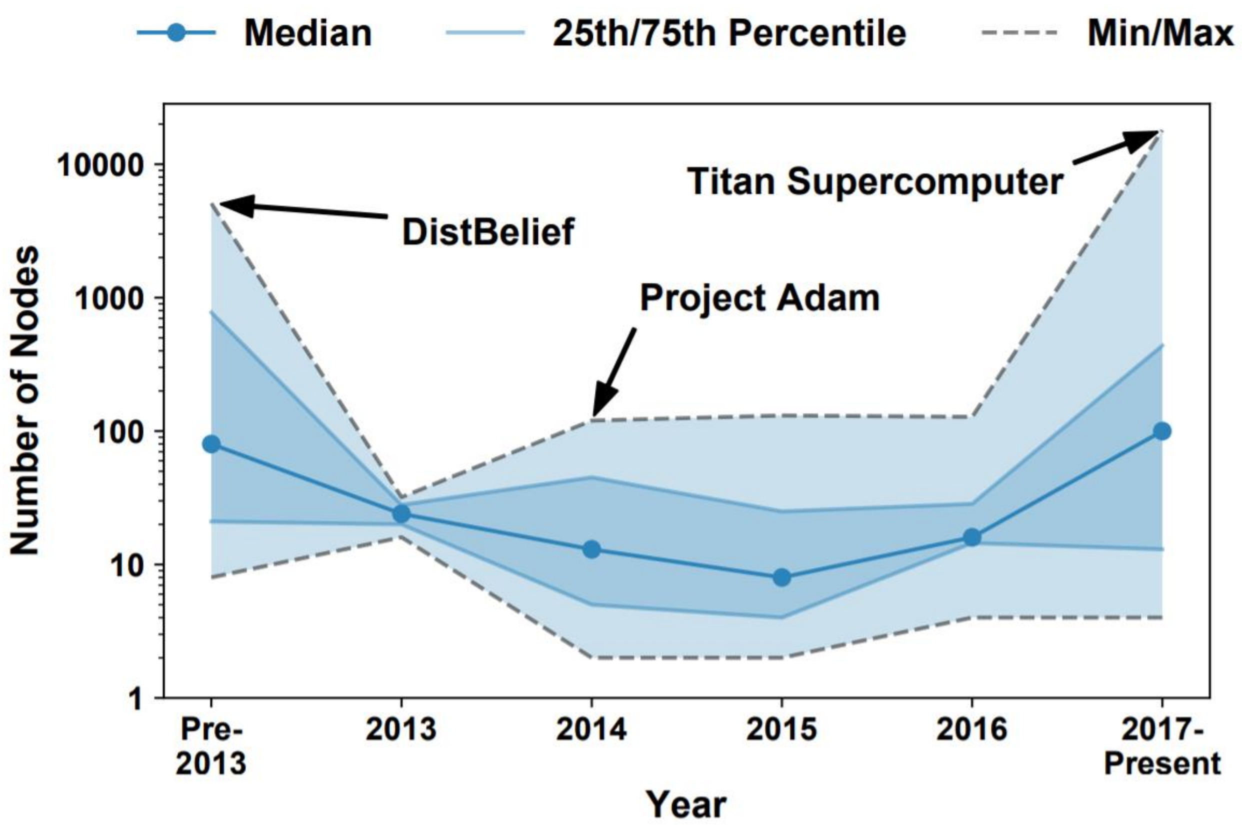


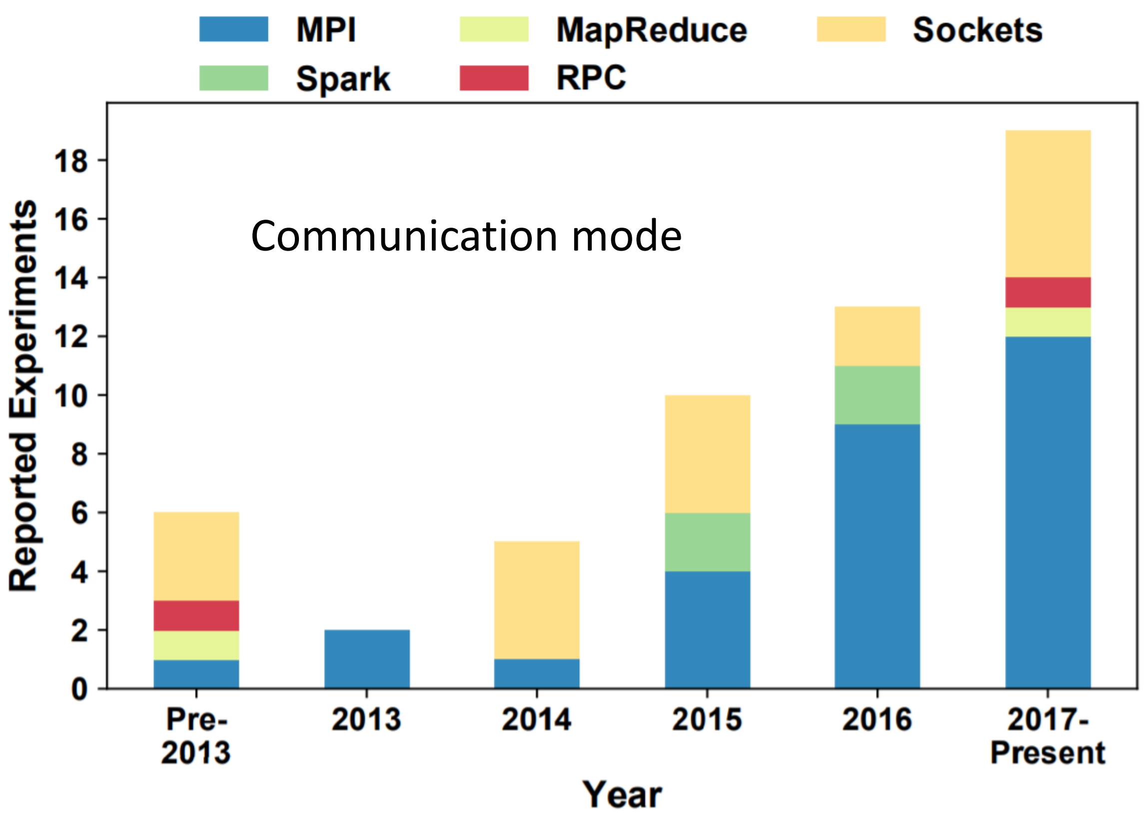
Deep Learning is largely on distributed memory today!

T. Ben-Nun, T. Hoefler: Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis, arXiv Feb 2018

Trends in distributed deep learning: node count and communication

The field is moving fast – trying everything imaginable – survey results from 240 papers in the area of parallel deep learning



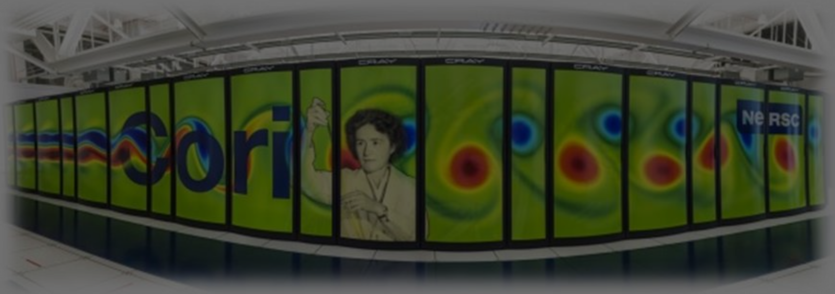


Deep Learning research is converging to MPI!

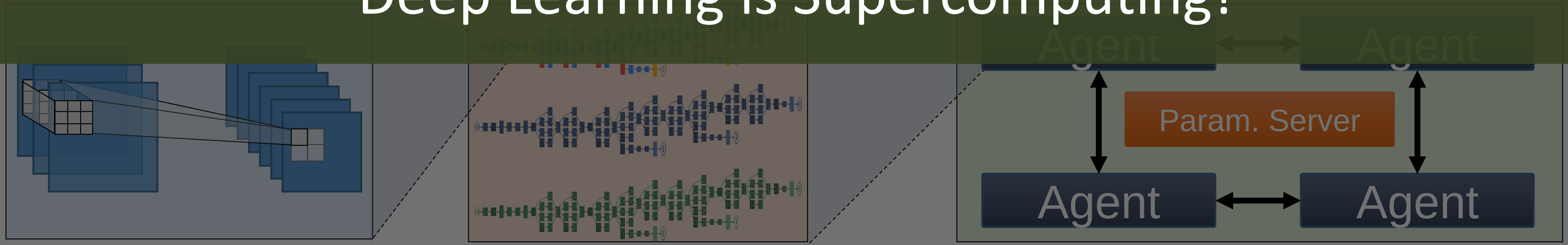
T. Ben-Nun, T. Hoefler: Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis, arXiv Feb 2018

Parallelism in Deep Learning

- Individual operators
- Network parallelism
- Optimization algorithm
- Distributed training



Deep Learning is Supercomputing!



Operators Networks Training

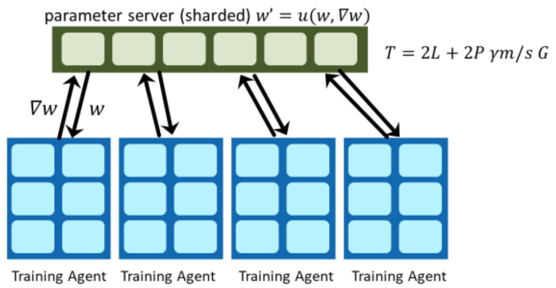
Challenges

Different:

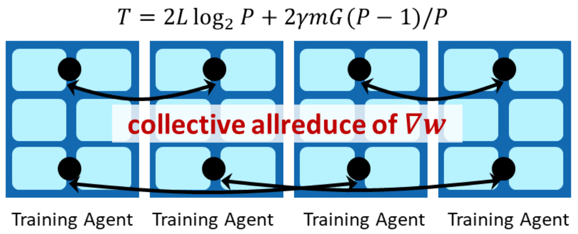
- **Communication schemes**
- **Model consistency requirements**
- **Software stacks and feature sets**

Need to define:

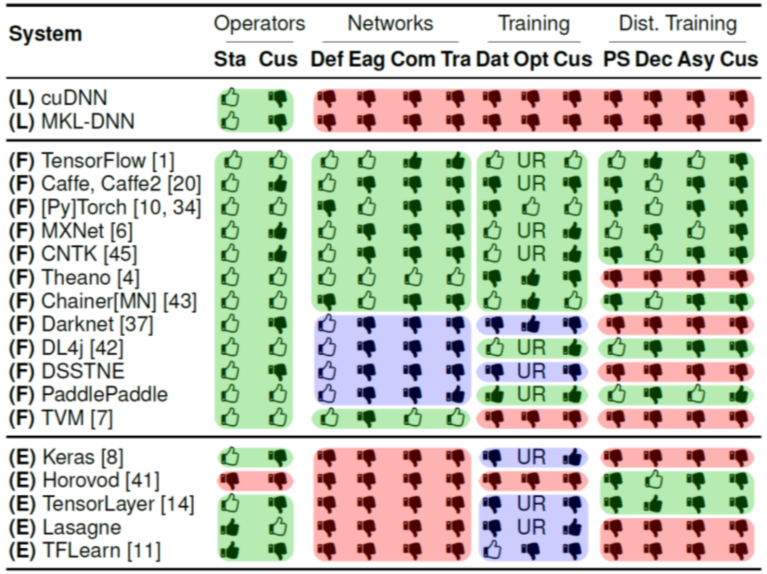
- **Open datasets from computational sciences**
- **Metrics robust to methods (or freeze methods)**
- **Standard benchmarking infrastructure**



Centralized



Decentralized



So how to not do this

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

1) Ignore accuracy when scaling up!

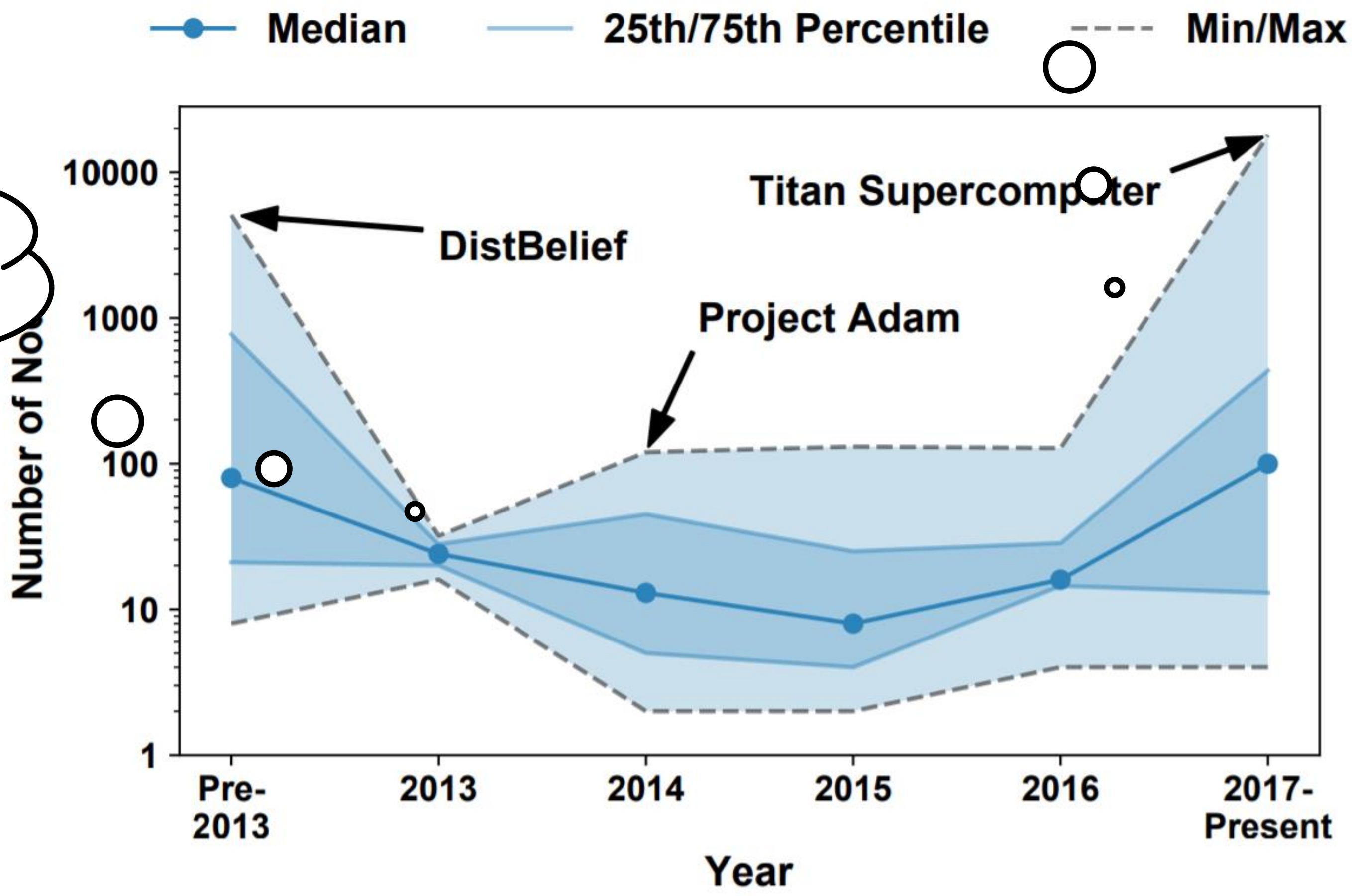
HPC picking up!

- **Too obvious for this audience**
  - Was very popular in 2015!

- **Surprisingly many (still) do this**

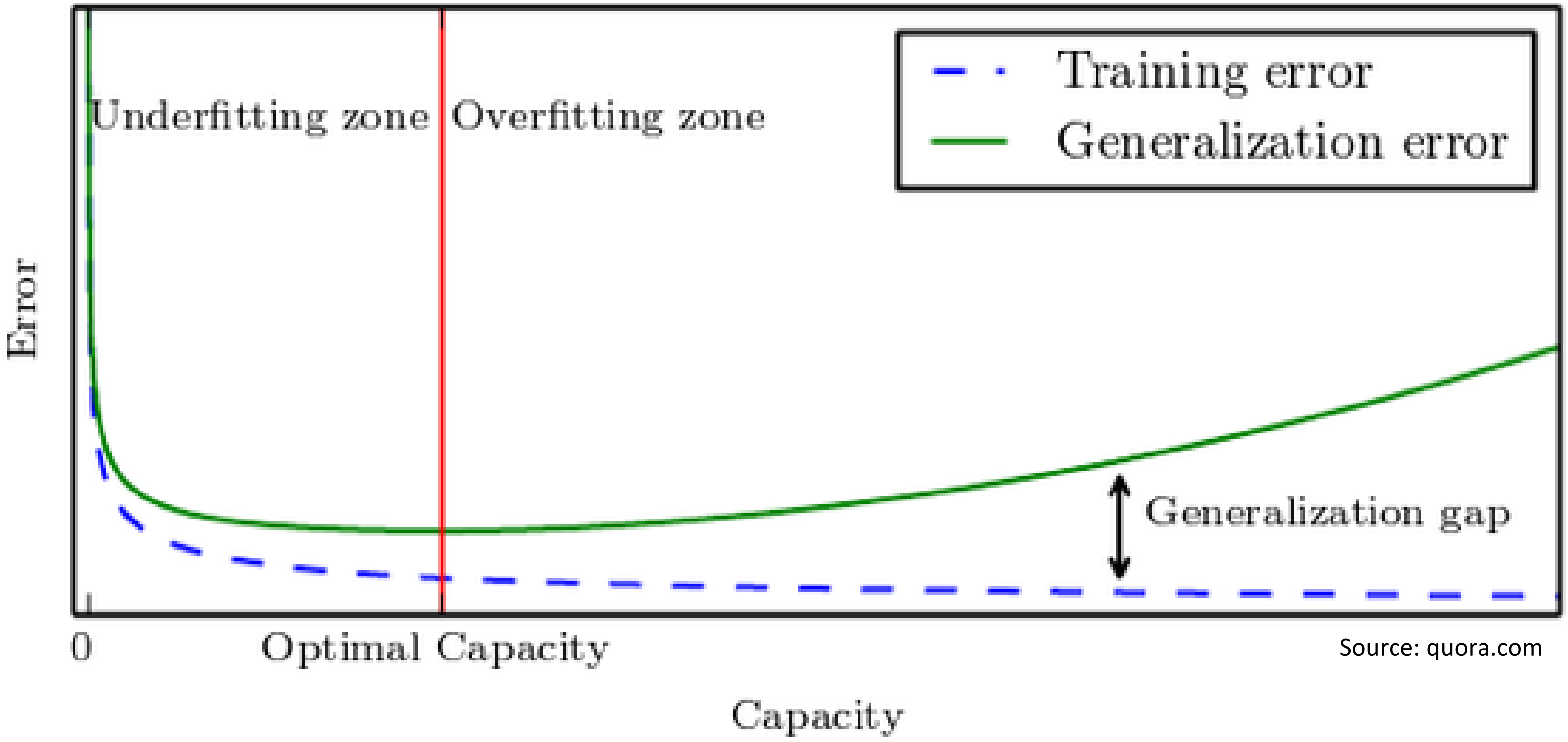
Learning community's self-correction (Y. LeCun)

Scalability without a good baseline? (D. Bailey)



2) Do not report test accuracy!

- **Training accuracy is sufficient isn't it?**

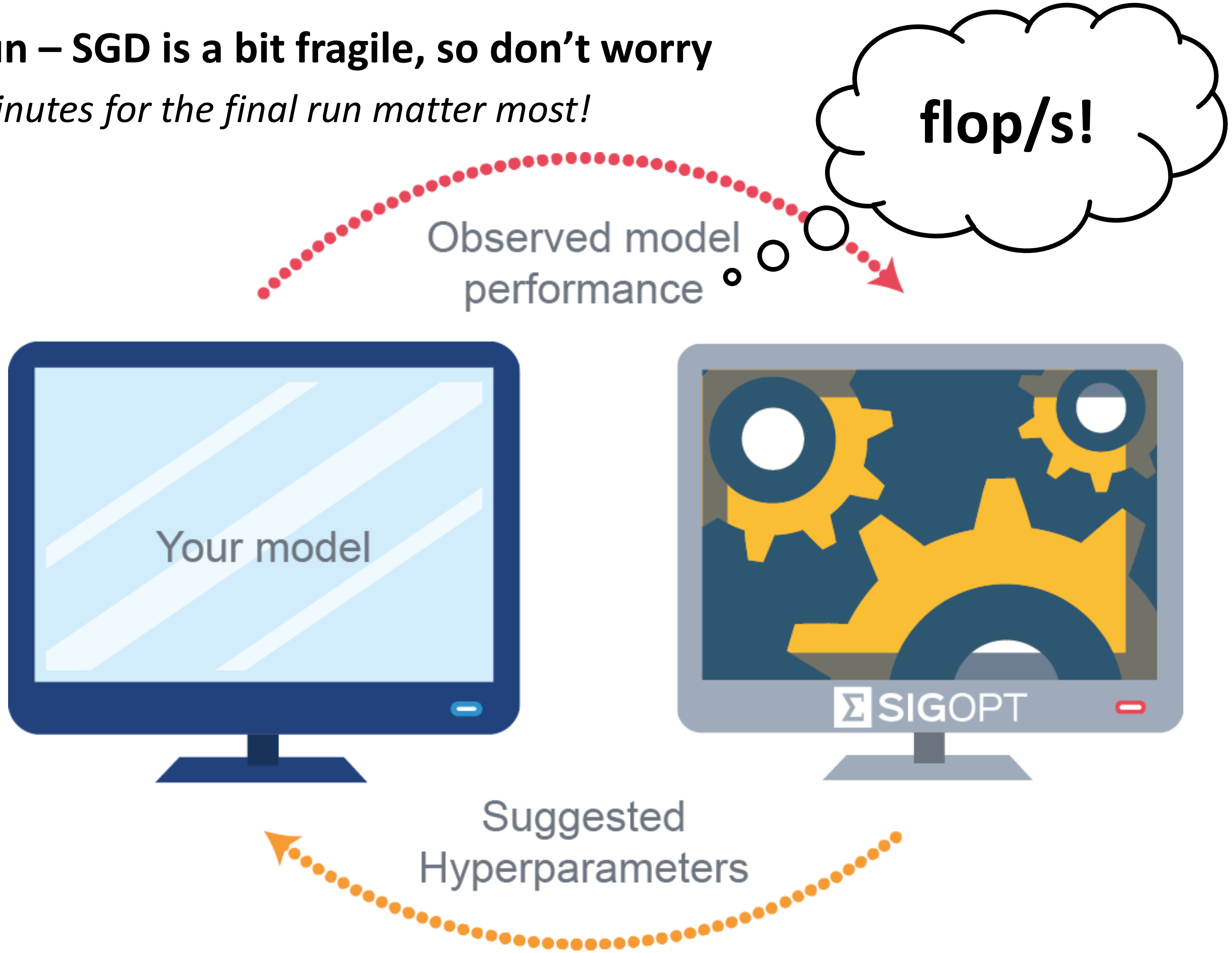


Source: quora.com

3) Do not report all training runs needed to tune hyperparameters!

- **Report the best run – SGD is a bit fragile, so don't worry**

  At the end, the minutes for the final run matter most!

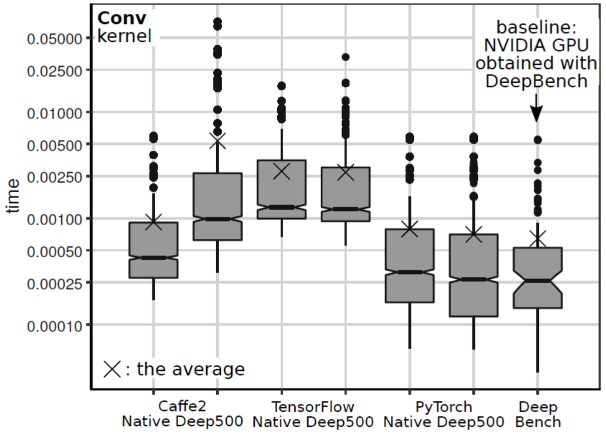


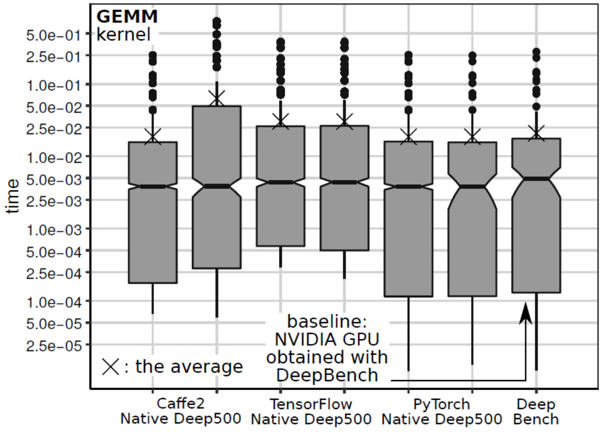
So how to not do this

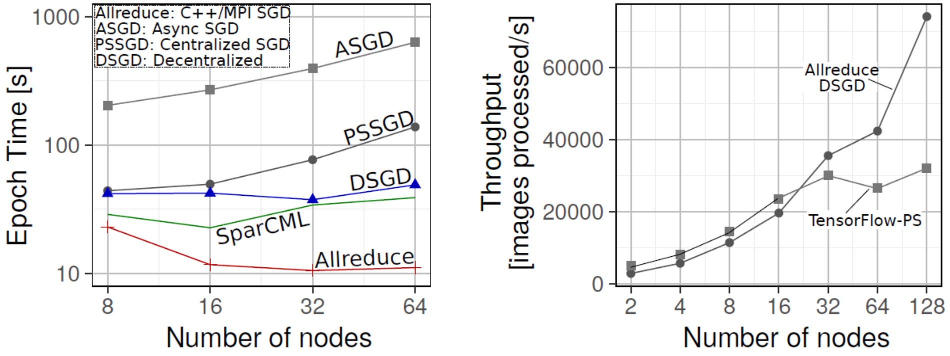
"Twelve ways to fool the masses when reporting performance of deep learning workloads"
(my humorous guide to floptimize deep learning)

Modular Benchmarking Infrastructure for Reproducible DL

- Separates benchmarking into the 4 core components
- Metrics defined separately, shared across levels
- Leverages ONNX for model definition
- Contains reference implementations of operators, optimizers, and distributed schemes
- Supports custom C, C++, and CUDA implementations on all levels
  - No need to reimplement an optimizer to replace gradient compression!







(a) Strong scaling (Wide ResNet 28x10). (b) Weak scaling (ResNet-56).

Fig. 11: **Scaling Analysis of Level 3**: Strong and weak scaling on Piz Daint.

Ben-Nun et al. "A Modular Benchmarking Infrastructure for High-Performance and Reproducible Deep Learning", soon on arXiv

Other Results

- **SparCML: a sparse reduction protocol to implement faster reductions in parallel systems with sparse input vectors [arXiv'18, NIPS'18]**

- **Using deep learning to create learnable representations of code [NIPS'18]**
  - State of the art in predicting fastest hardware mapping and algorithm classification

- **Accelerating convolution operators using micro-batches [Cluster'18]**
  - Key technique: Use ILP and Dynamic Programming

- **Parallelism modeling of deep learning, from operator to distributed training on supercomputers [arXiv'18]**