FUNDAMENTALS OF DEEP LEARNING FOR MULTI-GPUS

LAB 1, PART 1: INTRODUCTION AND MOTIVATION

NVIDIA | DEEP LEARNING INSTITUTE

COURSE OVERVIEW

- Lab 1: Gradient Descent vs Stochastic Gradient Descent, and the Effects of Batch Size
- Lab 2: Multi-GPU DL Training Implementation using Horovod
- Lab 3: Algorithmic Concerns for Training at Scale

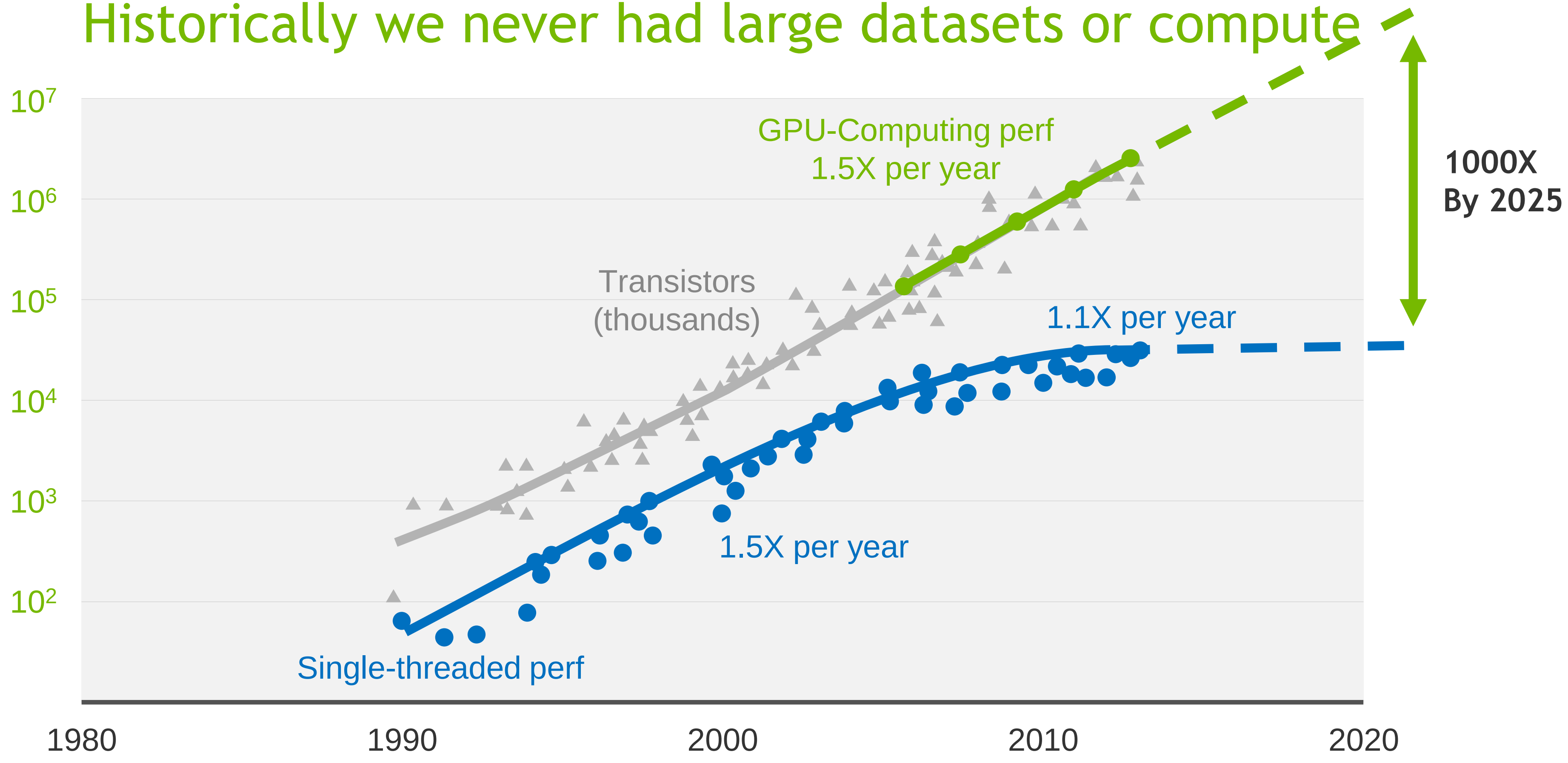
LAB 1 OVERVIEW

- Part 1: Gradient Descent
- Part 2: Stochastic Gradient Descent
- Part 3: Optimizing training with batch size

CONTEXT: WHY USE MULTIPLE GPUS?

TRENDS IN COMPUTATIONAL POWER

Historically we never had large datasets or compute



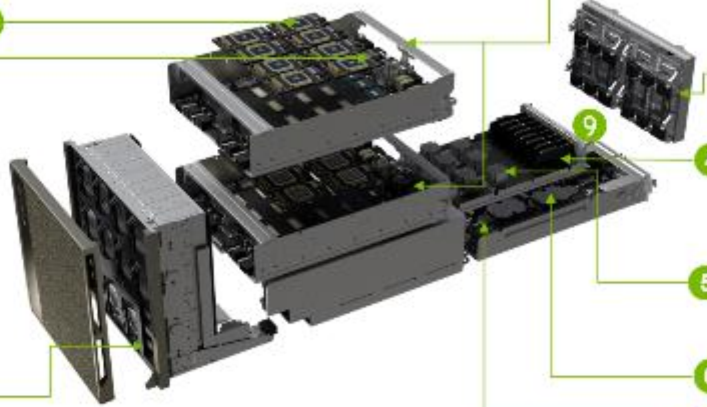
DEEP LEARNING INSTITUTE

TRENDS IN COMPUTATIONAL POWER

2 PF/s in November 2009

TRENDS IN COMPUTATIONAL POWER

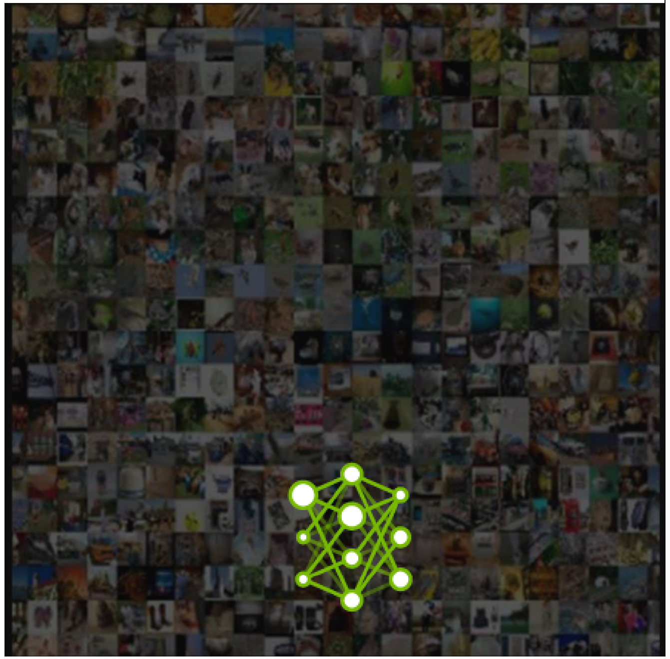
2 PF/s today



NEURAL NETWORK COMPLEXITY IS EXPLODING

To Tackle Increasingly Complex Challenges

7 Exaflops
60 Million Parameters

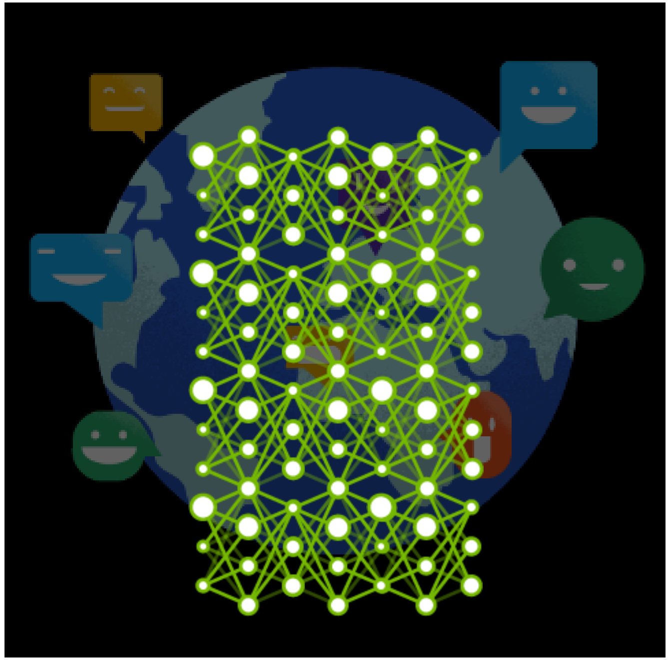


2015 - Microsoft ResNet
Superhuman Image Recognition

20 Exaflops
300 Million Parameters

2016 - Baidu Deep Speech 2
Superhuman Voice Recognition

100 Exaflops
8700 Million Parameters



2017 - Google Neural Machine Translation
Near Human Language Translation

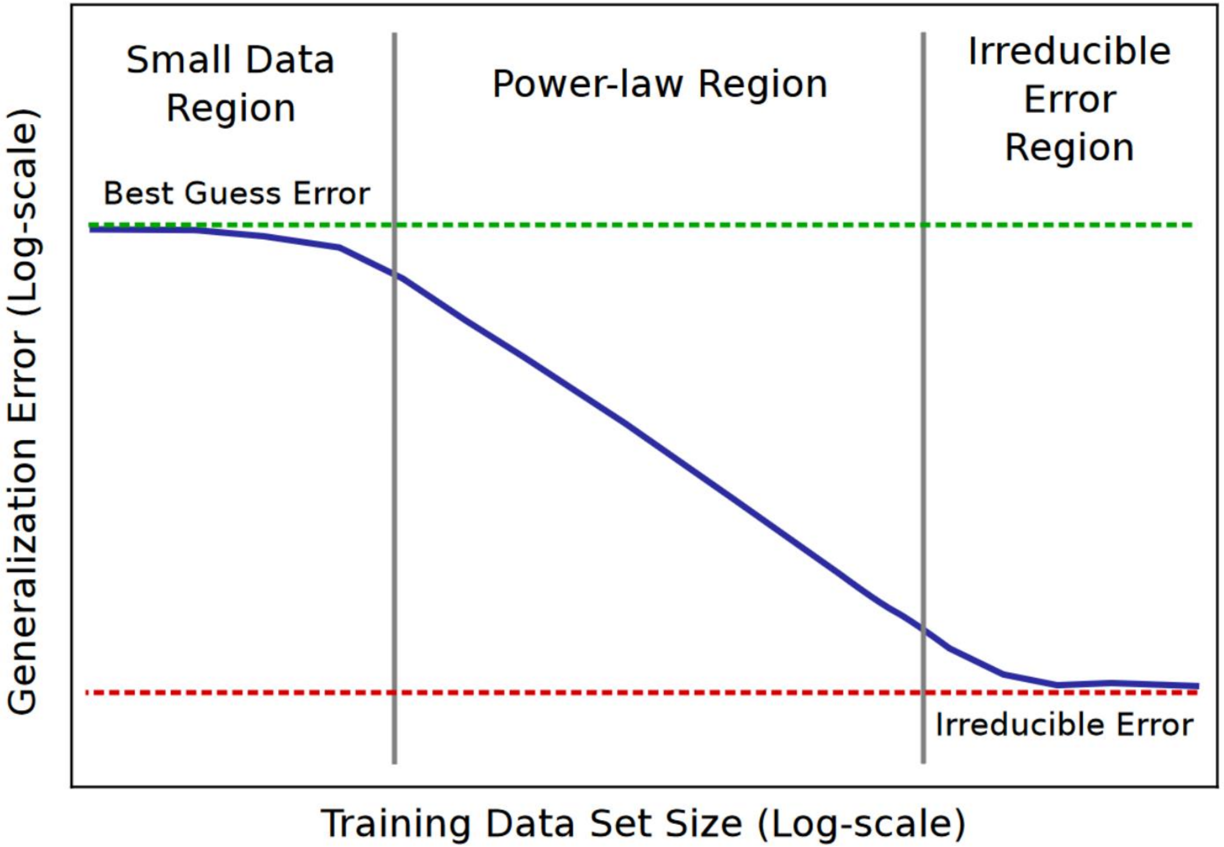
100 EXAFLOPS

=

O(1 YEAR) ON A DUAL CPU SERVER

EXPLODING DATASETS

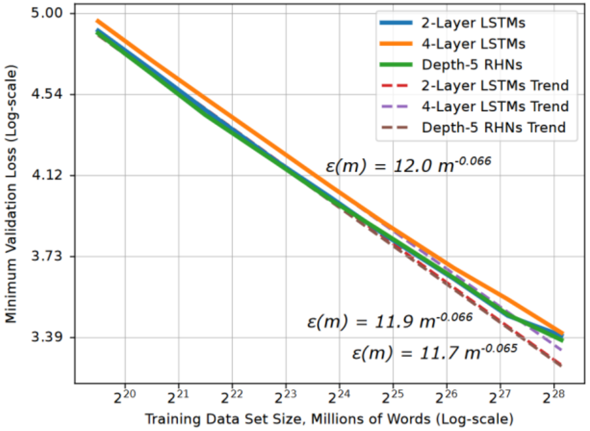
Power-law relationship between dataset size and accuracy



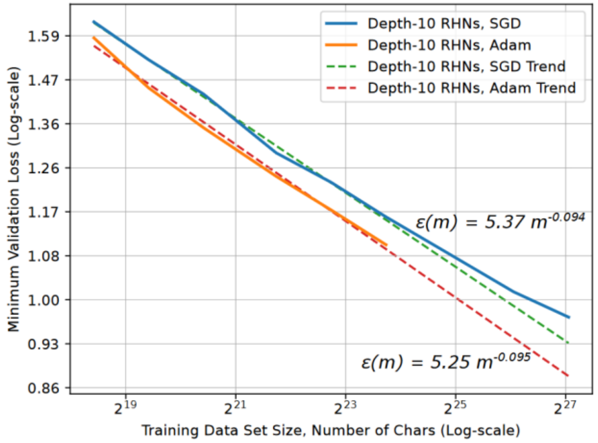
Hestness, J., et al. (2017). Deep Learning Scaling is Predictable, Empirically. [arXiv: 1712.00409]

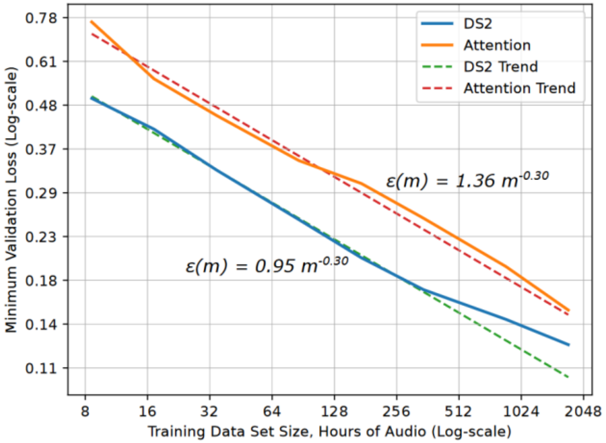
EXPLODING DATASETS

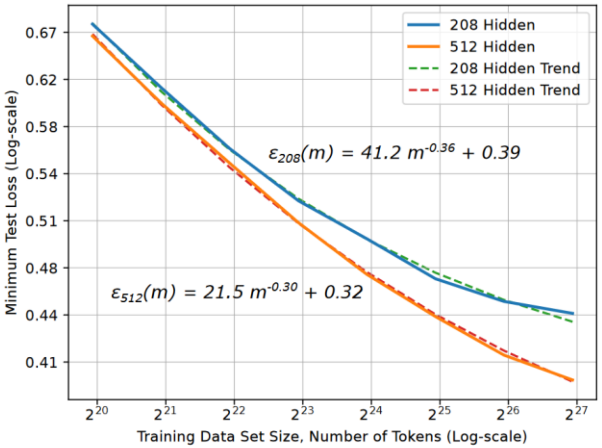
Power-law relationship between dataset size and accuracy

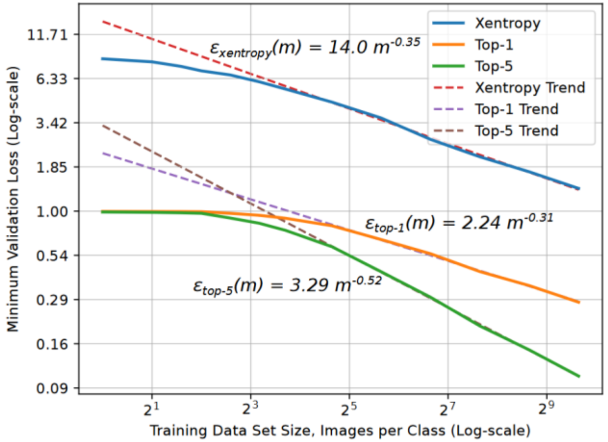


- Translation
- Language Models
- Character Language Models
- Image Classification
- Attention Speech Models





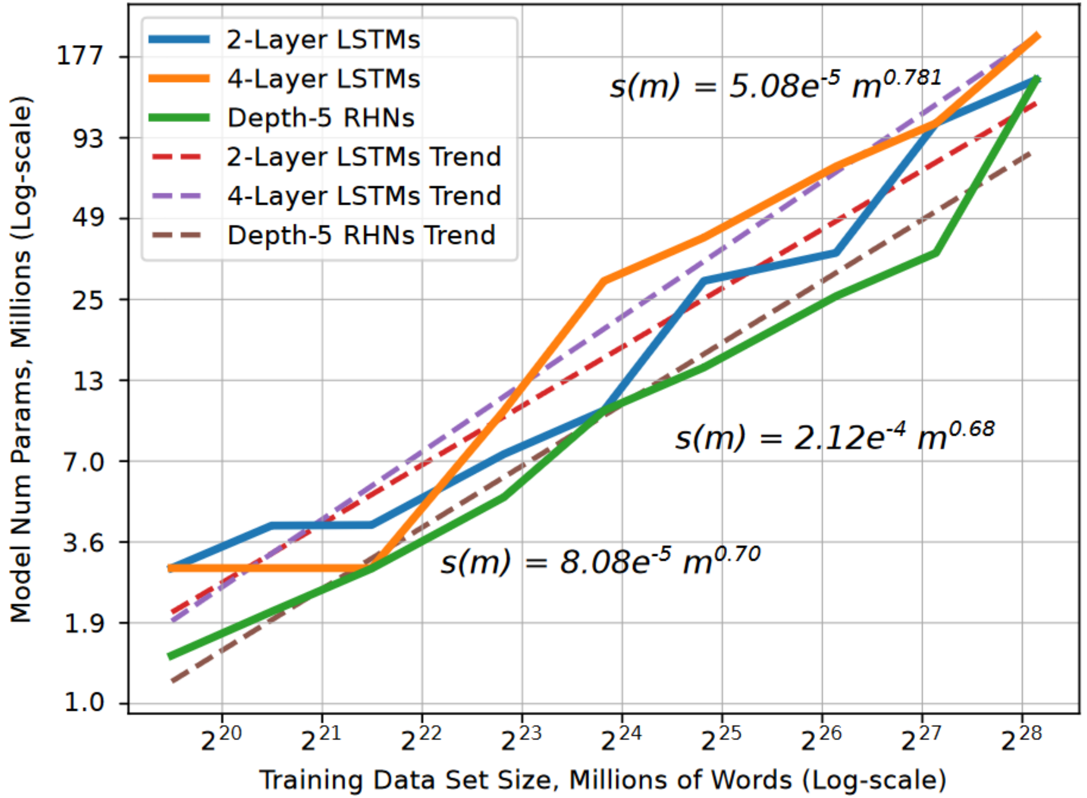




Hestness, J., et al. (2017). Deep Learning Scaling is Predictable, Empirically. arXiv: 1712.00409

EXPLODING MODEL COMPLEXITY

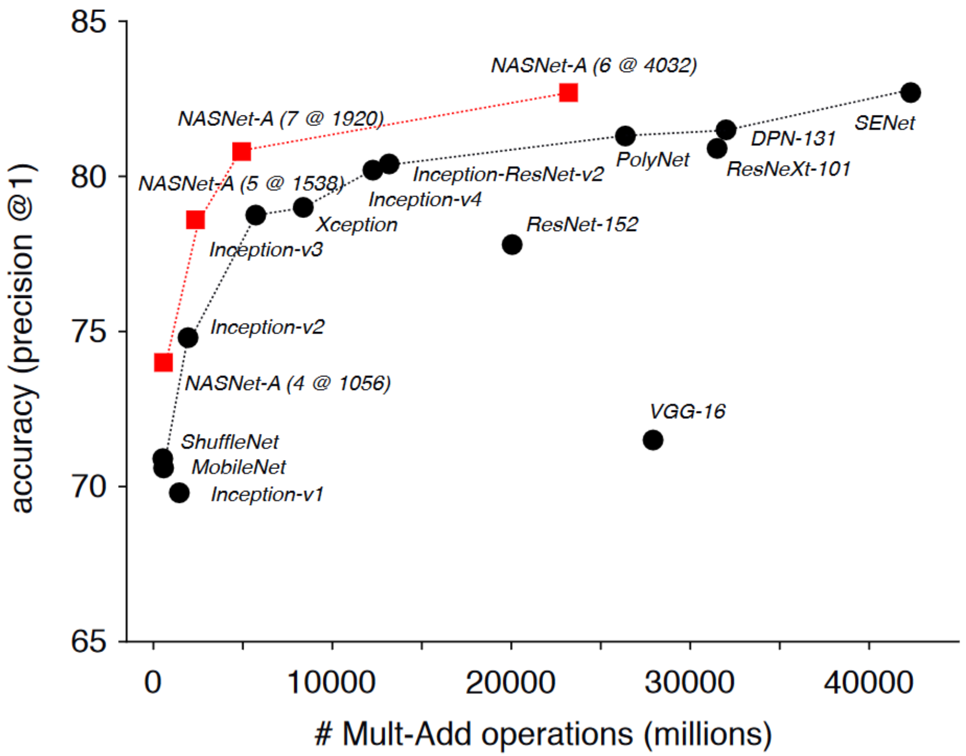
Though model size scales sublinearly

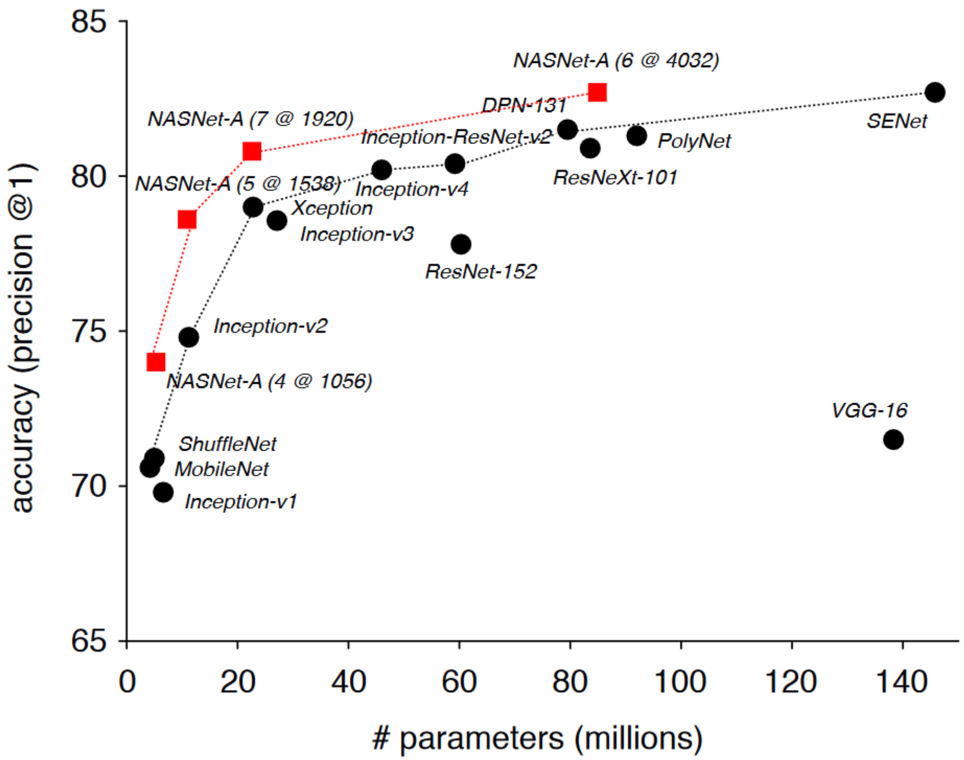


Hestness, J., et al. (2017). Deep Learning Scaling is Predictable, Empirically. [arXiv: 1712.00409]

EXPLODING MODEL COMPLEXITY

Though model size scales sublinearly





Zoph, Barret, et al. (2017). "Learning transferable architectures for scalable image recognition." arXiv: 1707.07012

IMPLICATIONS

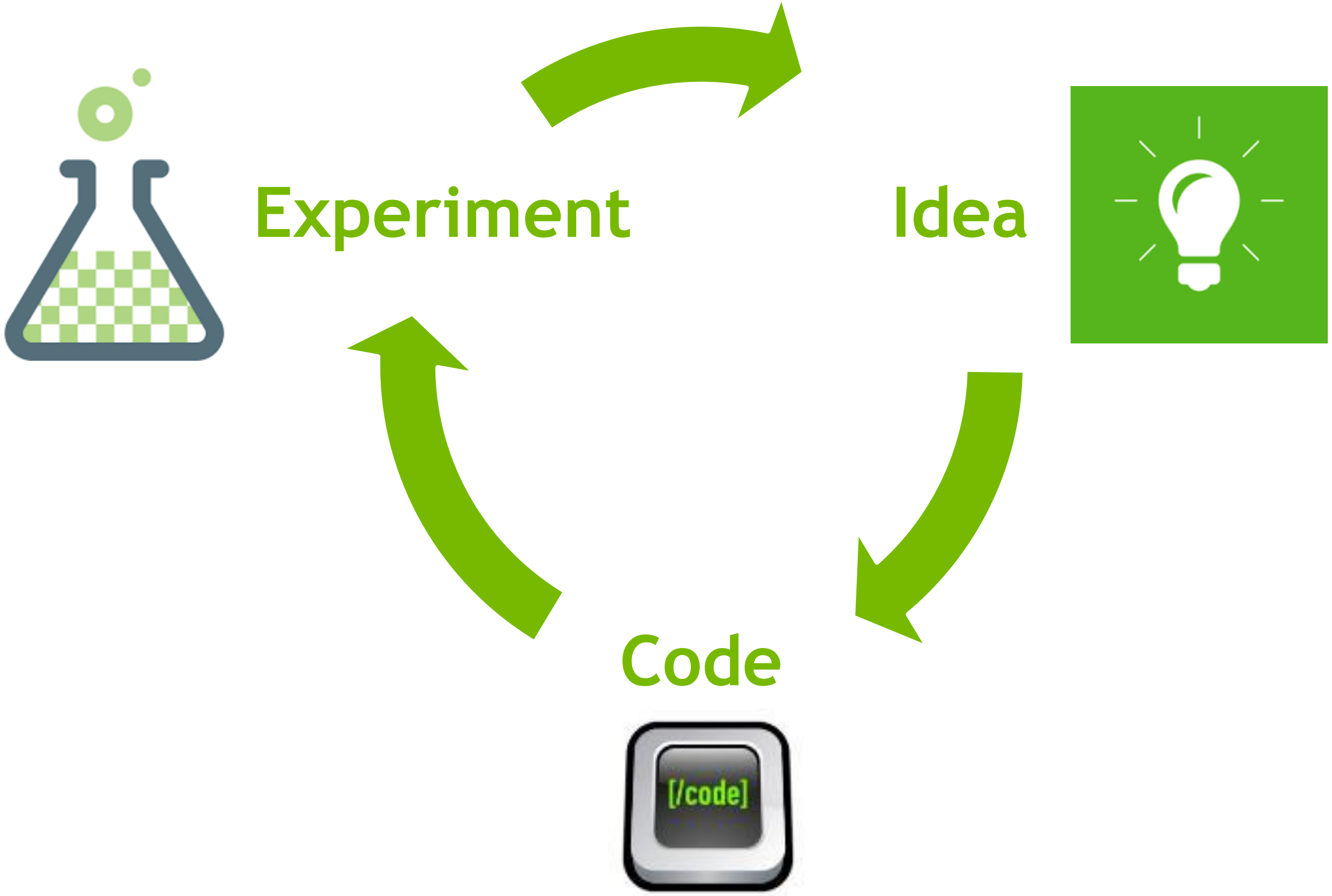
IMPLICATIONS

Good and bad news

- The good news: Requirements are predictable.
  - We can predict how much data we will need.
  - We can predict how much computing power we will need.
- The bad news: The values can be significant.
  - The silver lining is that deep learning has taken impossible problems and made them merely expensive.

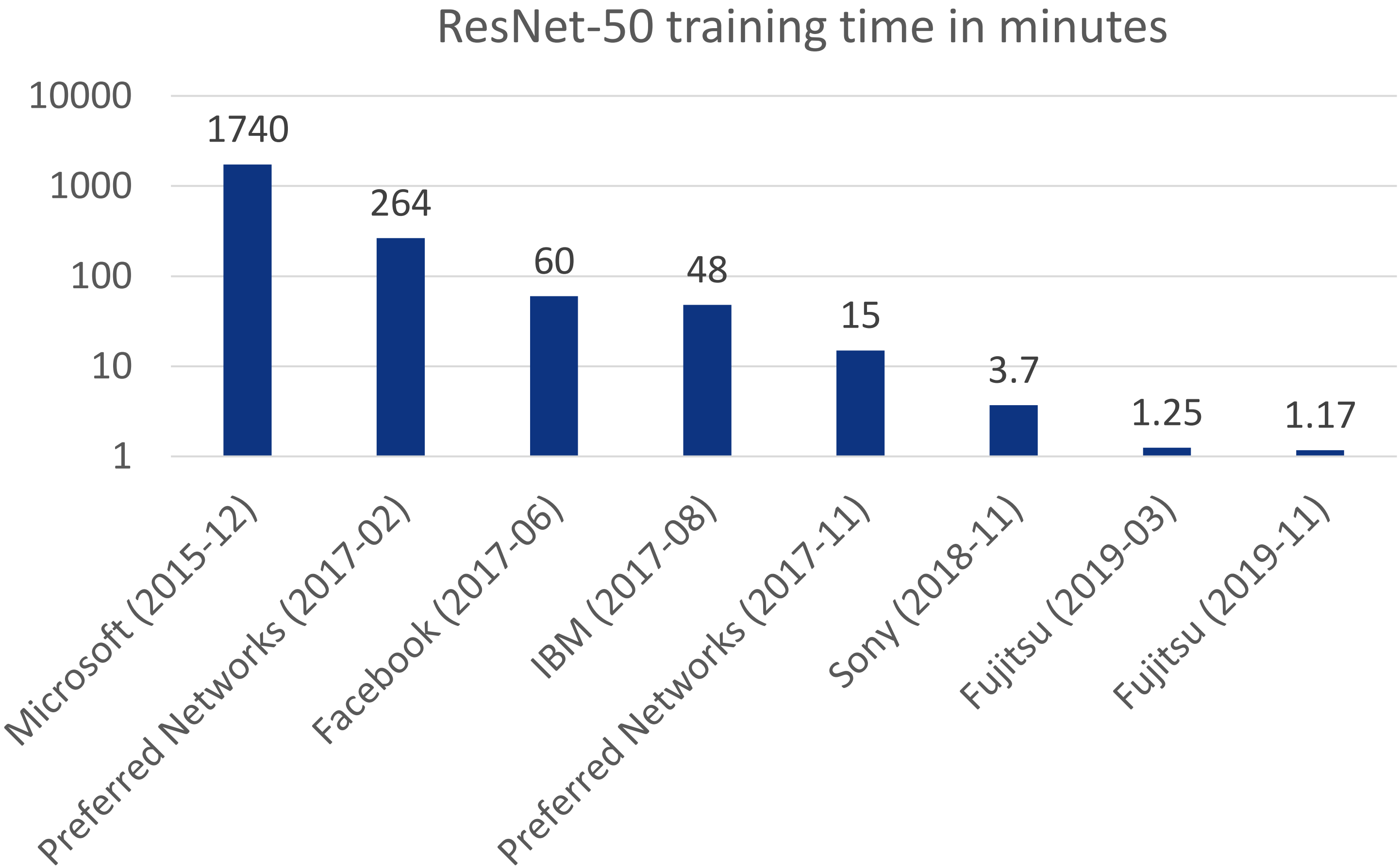
IMPLICATIONS

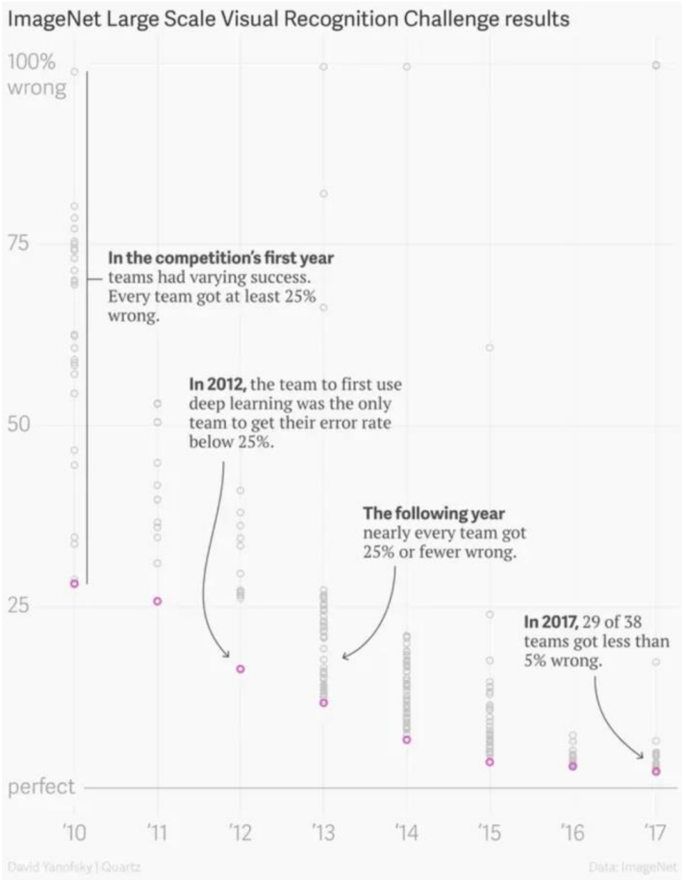
Deep learning is experimental; we need to train quickly to iterate



ITERATION TIME

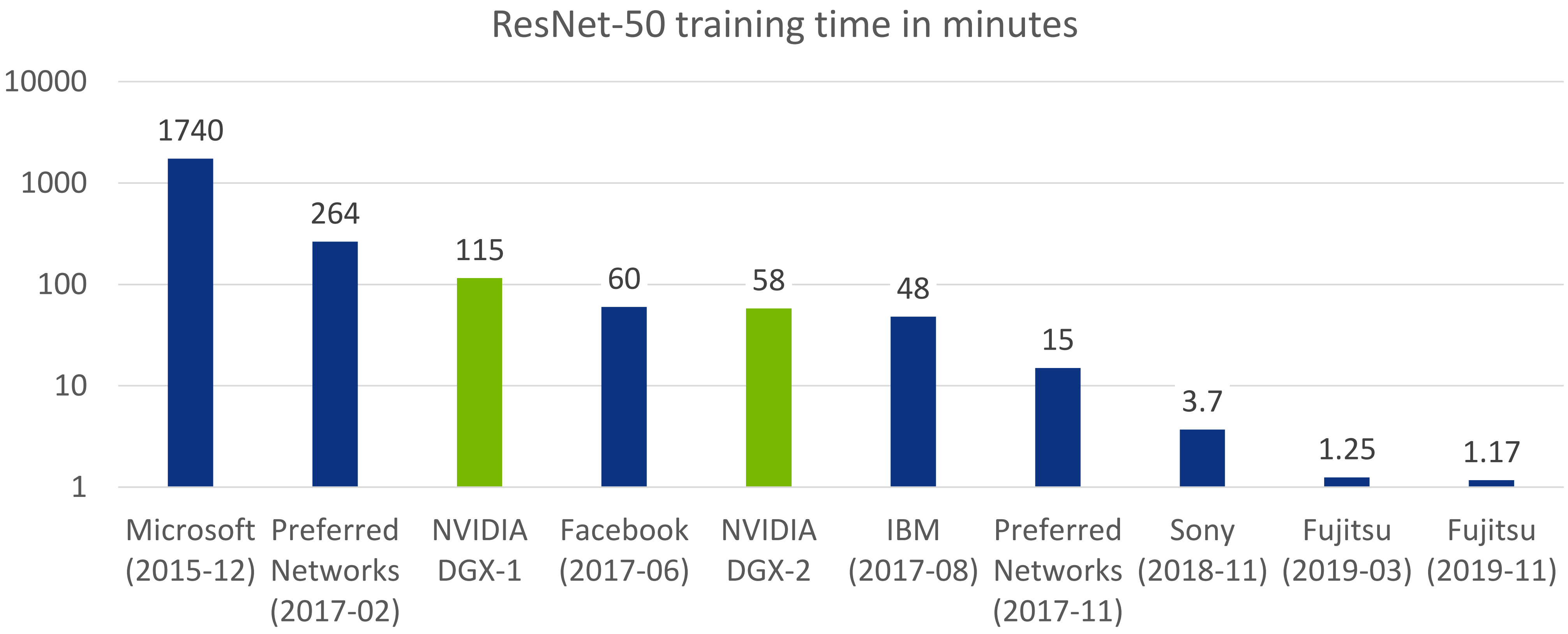
Short iteration time is fundamental for success





ITERATION TIME

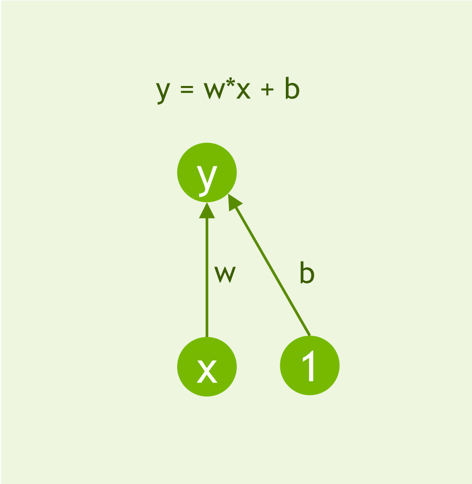
Short iteration time is fundamental for success

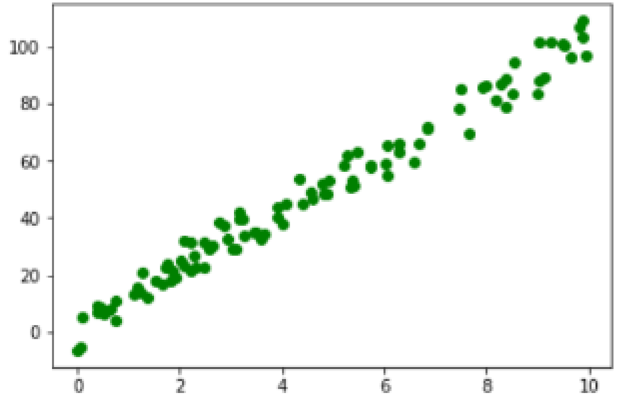


INTRO TO THE LAB

STARTING WITH A LINEAR MODEL

Our goal is to find best model parameters (combination of w and b) to fit the data





DEEP
LEARNING
INSTITUTE

www.nvidia.com/dli