

DATA PARALLELISM: HOW TO TRAIN DEEP LEARNING MODELS ON MULTIPLE 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 DistributedDataParallel (DDP)
- Lab 3: Algorithmic Concerns for Training at Scale

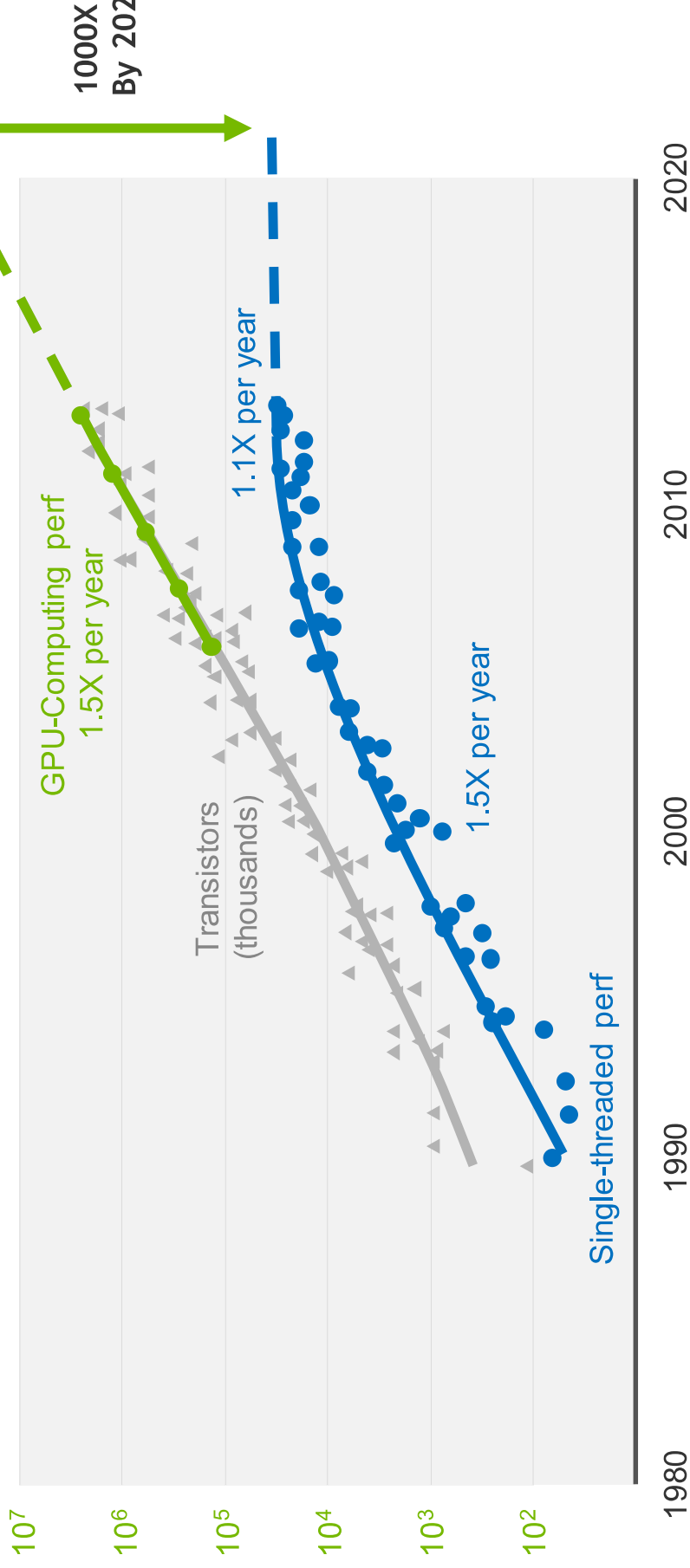
LAB 1 OVERVIEW

- 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



TRENDS IN COMPUTATIONAL POWER

2 PF/s in November 2009



TRENDS IN COMPUTATIONAL POWER

32 PF/s today

8x NVIDIA H100 GPUs With 640 Gigabytes of Total GPU Memory

18x NVIDIA NVLink connections per GPU

900 gigabytes per second of bidirectional GPU-to-GPU bandwidth

24 TB/s memory bandwidth

4x NVIDIA NVSwitches

7.2 terabytes per second of bidirectional GPU-to-GPU bandwidth

10x NVIDIA ConnectX-7 400 Gigabits-Per-Second Network Interface

1 terabyte per second of peak bidirectional network bandwidth

Dual x86 CPUs and 2 Terabytes of System Memory

Powerful CPUs and massive system memory for the most intensive AI jobs

32 petaFLOPS AI performance



NVIDIA DGX H100

NEURAL NETWORK COMPLEXITY IS EXPLODING

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute (Log Scale)



Source: [Google](#)

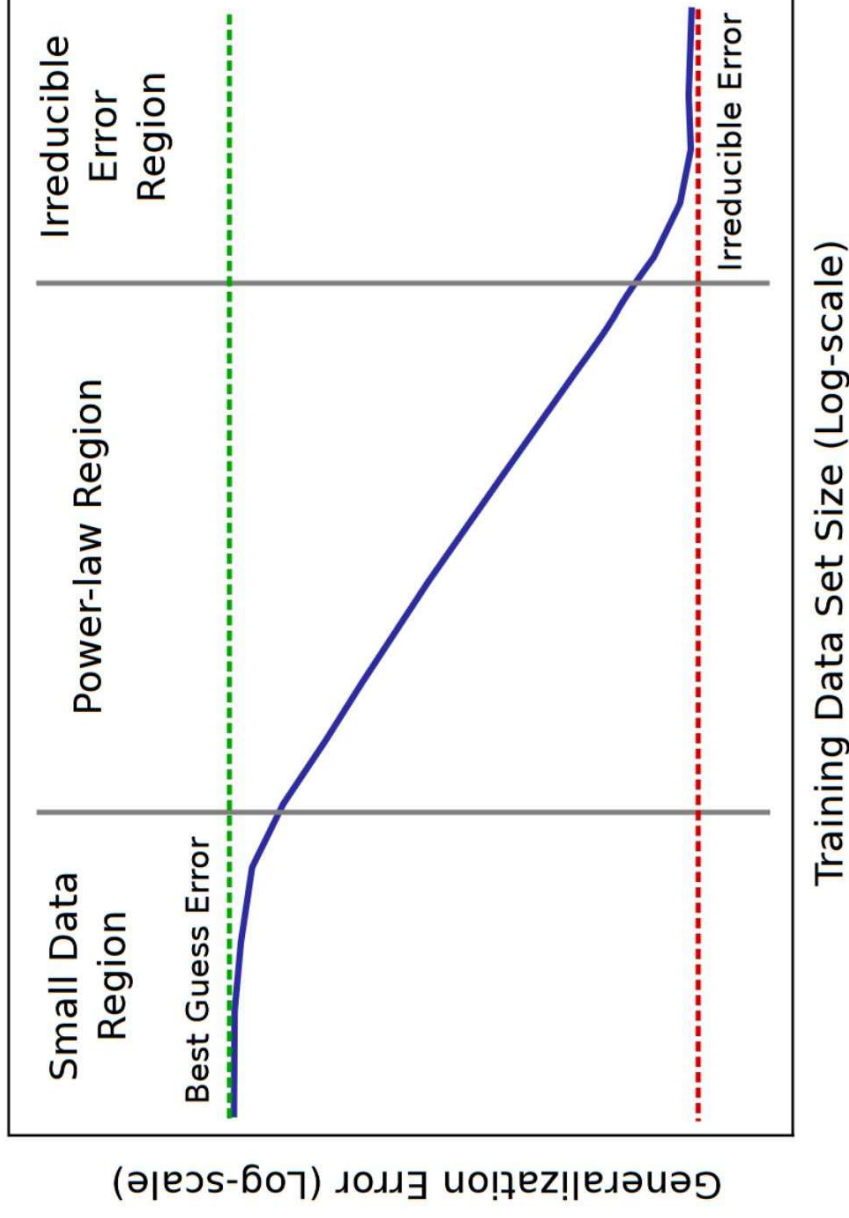
1000 PETAFLOP/S-DAYS

O(100 YEARS) ON A DUAL CPU SERVER
OR

O(30 DAYS) DGX H100

EXPLODING DATASETS

Power-law relationship between dataset size and accuracy

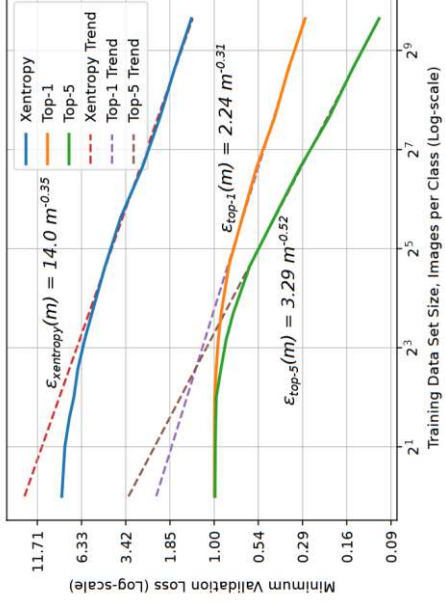
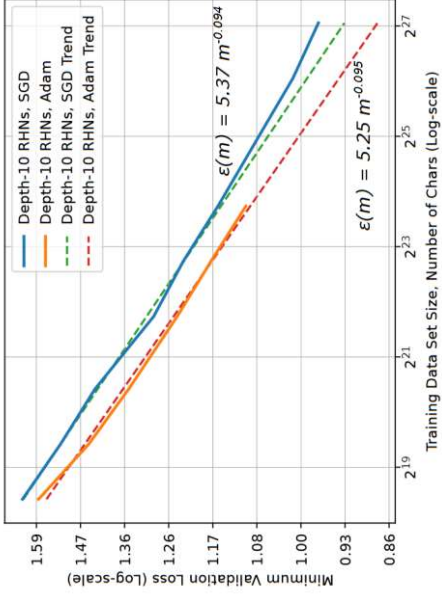
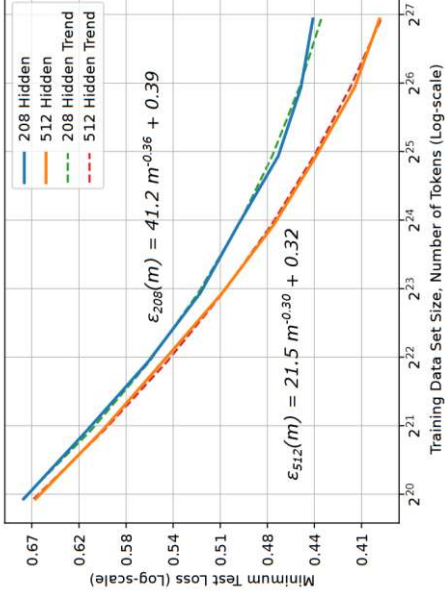
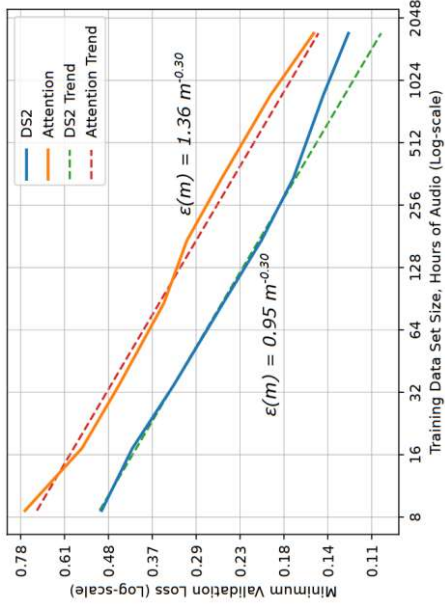
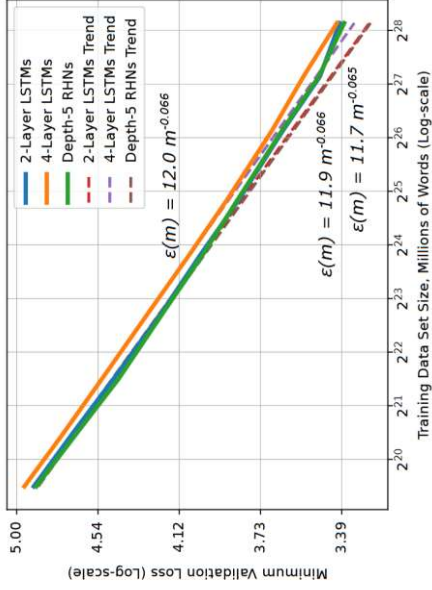


Hestness, J., et al. (2017). Deep Learning Scaling is Predictable, Empirically. [arXiv: 1712.00409](https://arxiv.org/abs/1712.00409)

EXPLODING DATASETS

Power-law relationship between dataset size and accuracy

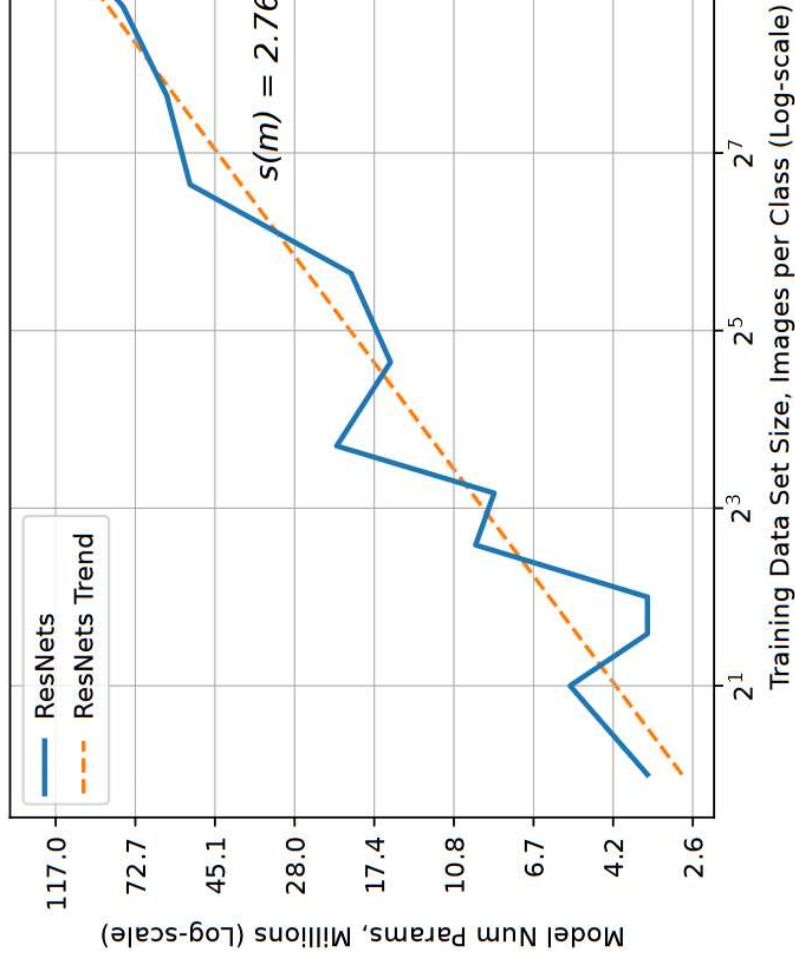
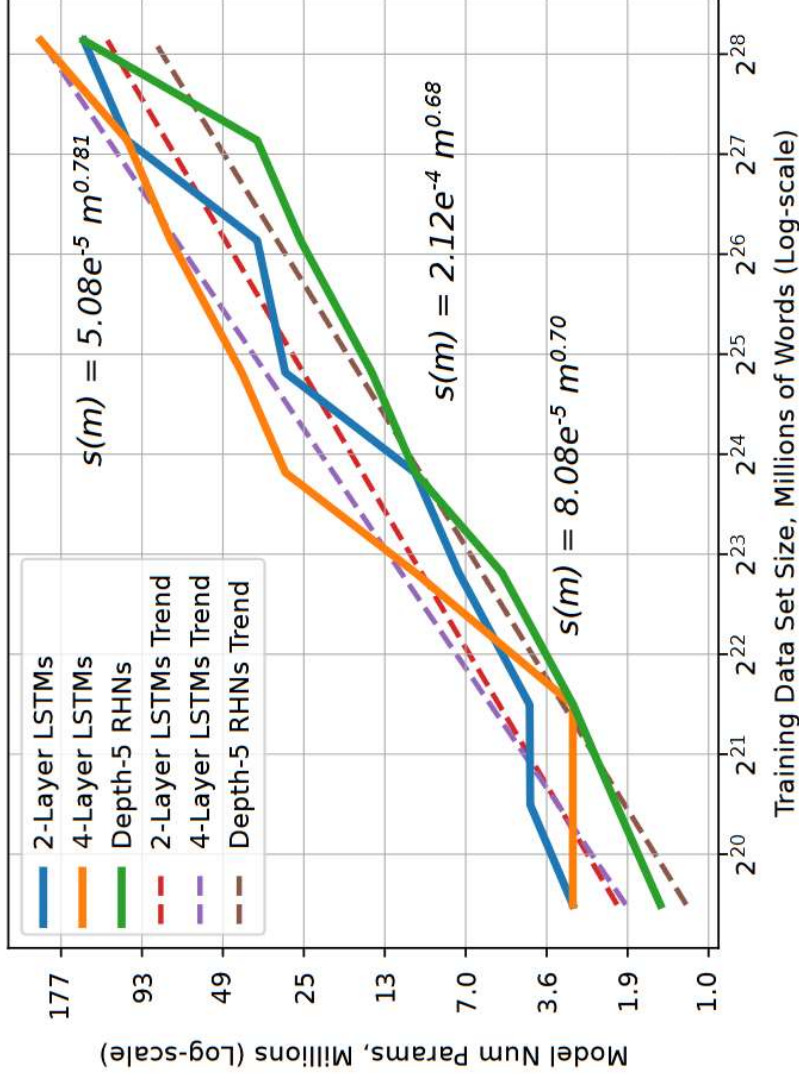
- Translation
- Language Models
- Character Language Models
- Image Classification
- Attention Speech Models



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

EXPLODING MODEL COMPLEXITY

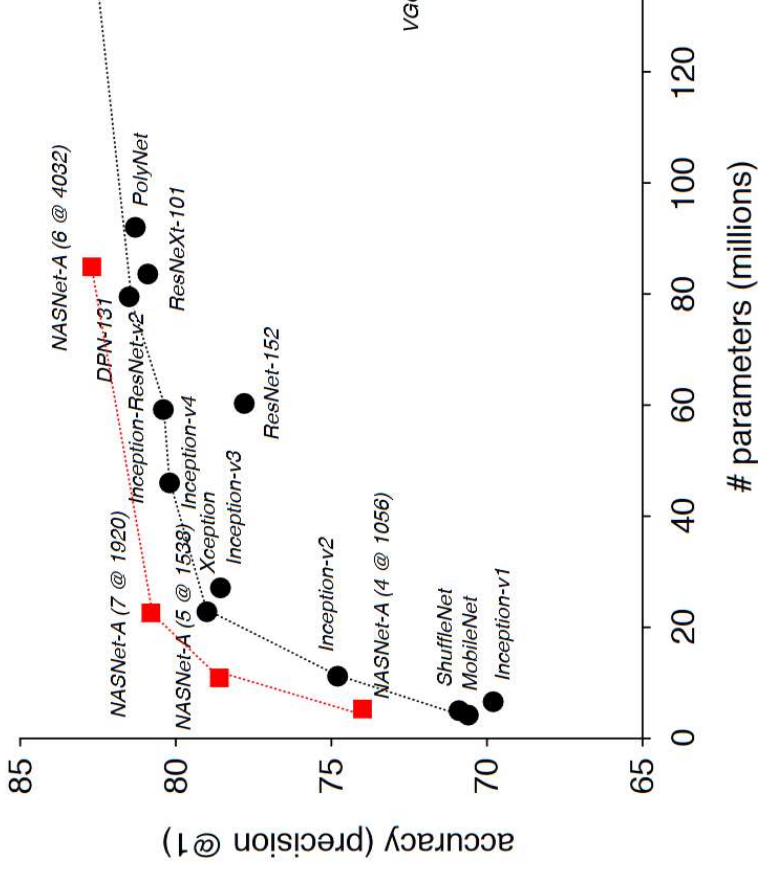
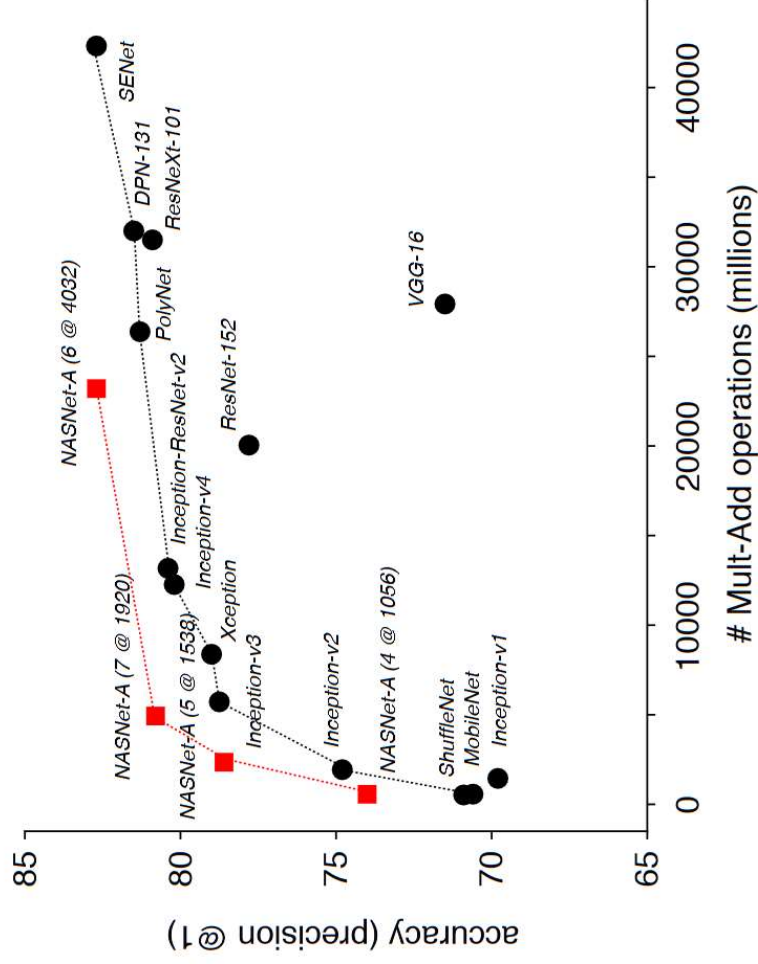
Though model size scales sublinearly



Hestness, J., et al. (2017). Deep Learning Scaling is Predictable, Empirically. [arXiv: 1712.00409](https://arxiv.org/abs/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](https://arxiv.org/abs/1707.07012)



IMPLICATIONS

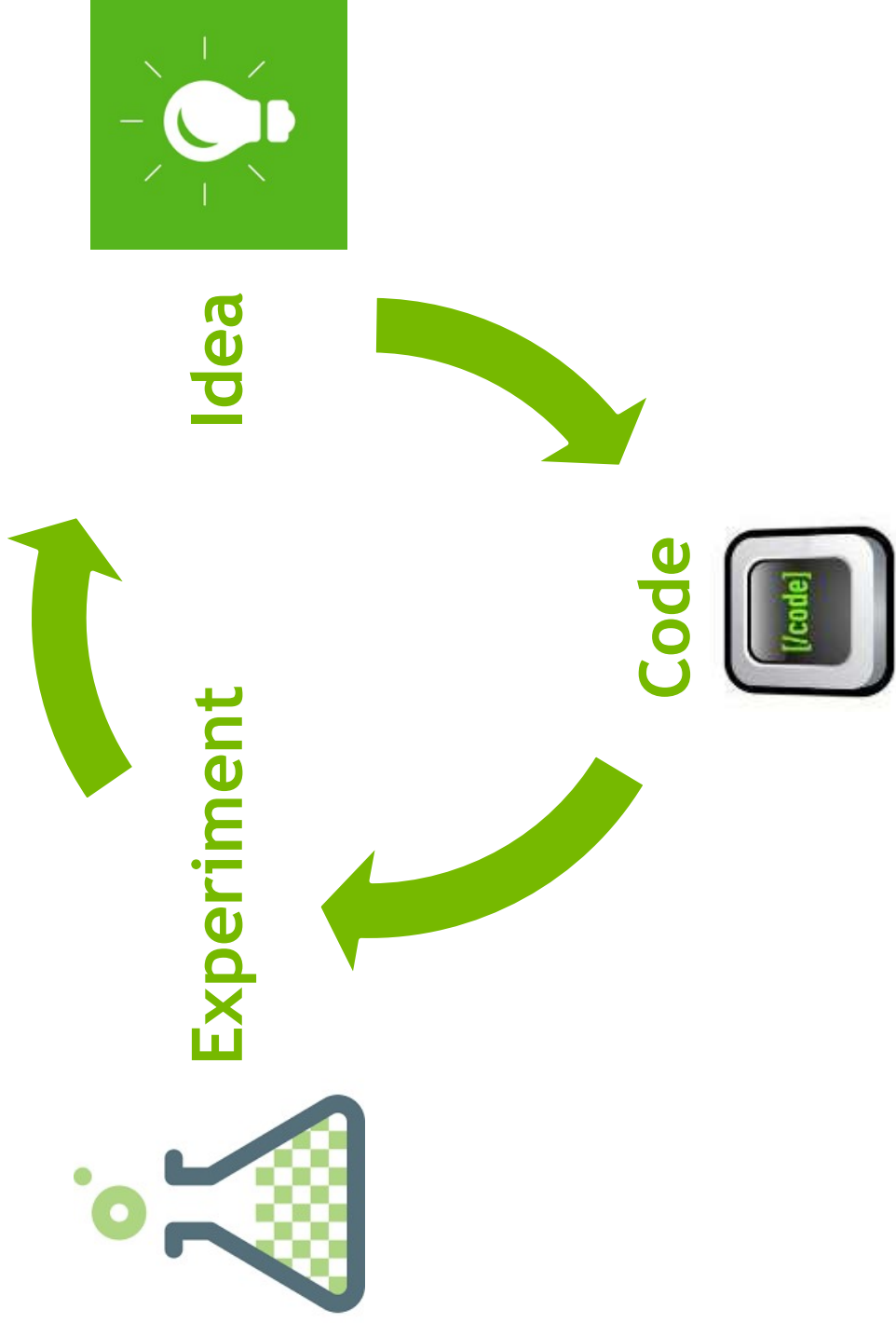
IMPLICATIONS

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

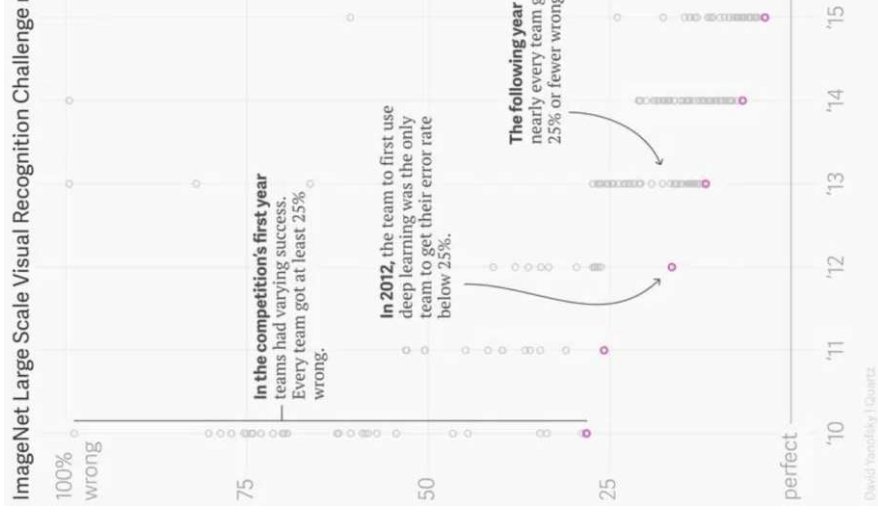
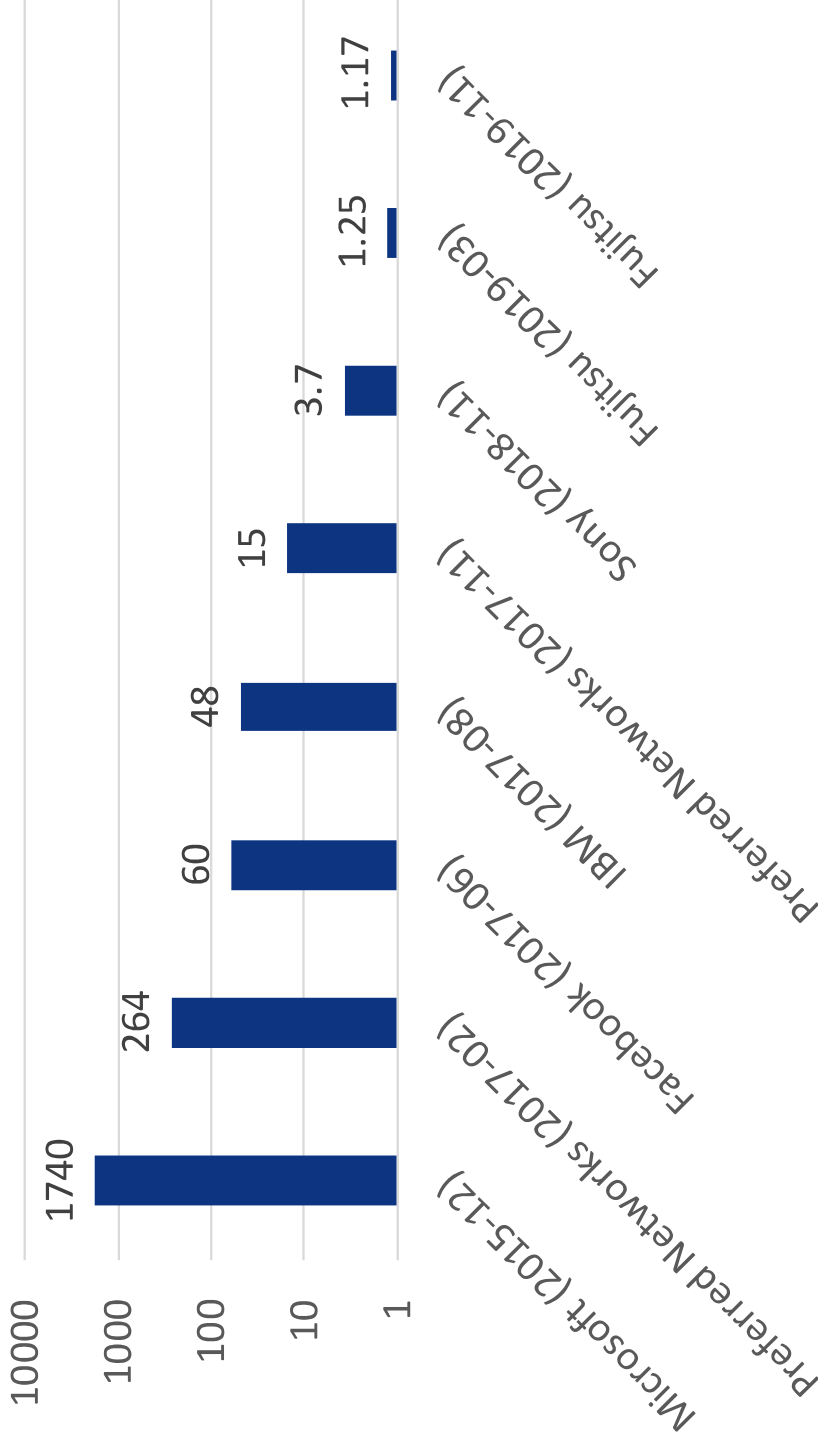
Deep learning is experimental; we need to train quickly to iterate



ITERATION TIME

Short iteration time is fundamental for success

ResNet-50 training time in minutes

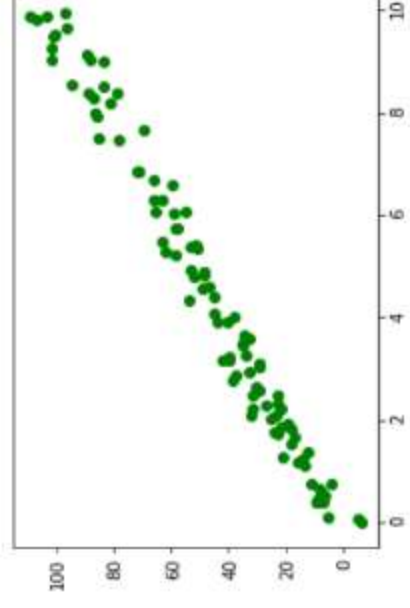
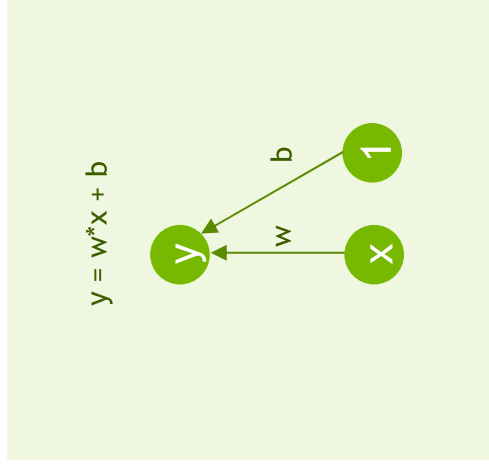


INTRO TO THE LAB

The background of the slide features a smooth gradient from a vibrant green on the left to a clean white on the right. Overlaid on this gradient is a complex, abstract network of small white dots connected by thin white lines, creating a mesh-like pattern that resembles a molecular structure or a data network. This pattern is more densely packed on the right side of the slide and fades into the white background on the left.

STARTING WITH A LINEAR MODEL

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





nvidia.

DEEP
LEARNING
INSTITUTE

www.nvidia.com/dli