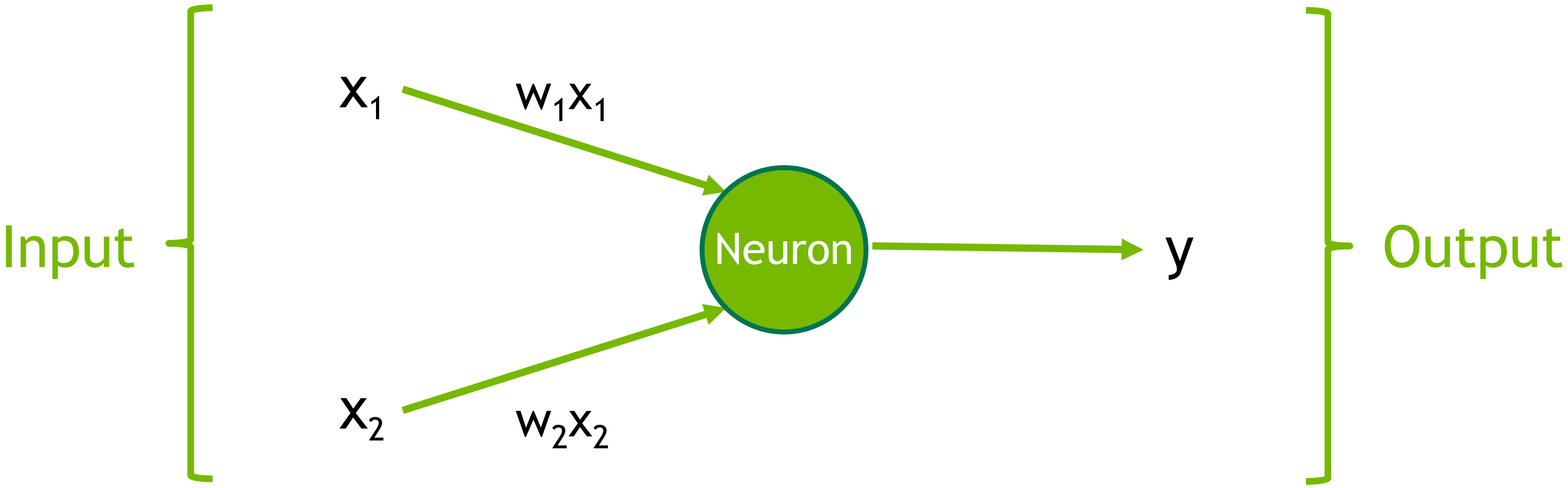
DOING LARGE SCALE AI FASTER

Miguel Martínez - Solution Architect

NEURAL NETWORKS ARE NOT NEW

And are surprisingly simple as an algorithm



A simple neuron

COMBINING NEURONS

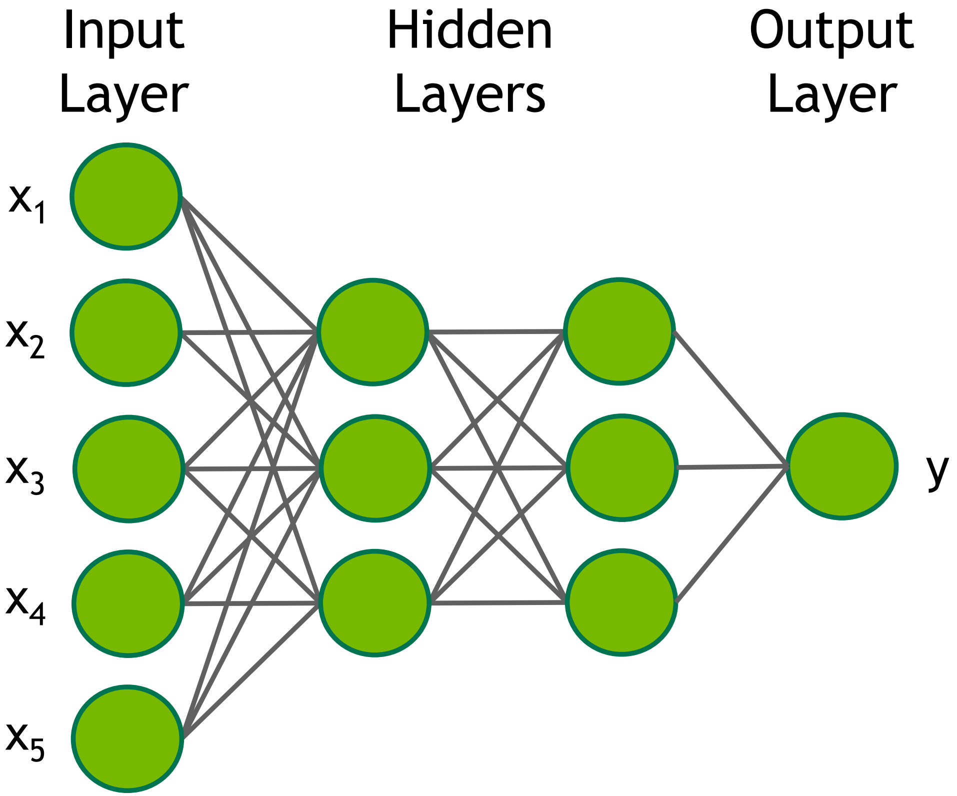
Stacking neurons and layers creates a more powerful model

Additional neurons can be added to create a layer.

Multiple layers can also be added, resulting in input, hidden, and output layers.

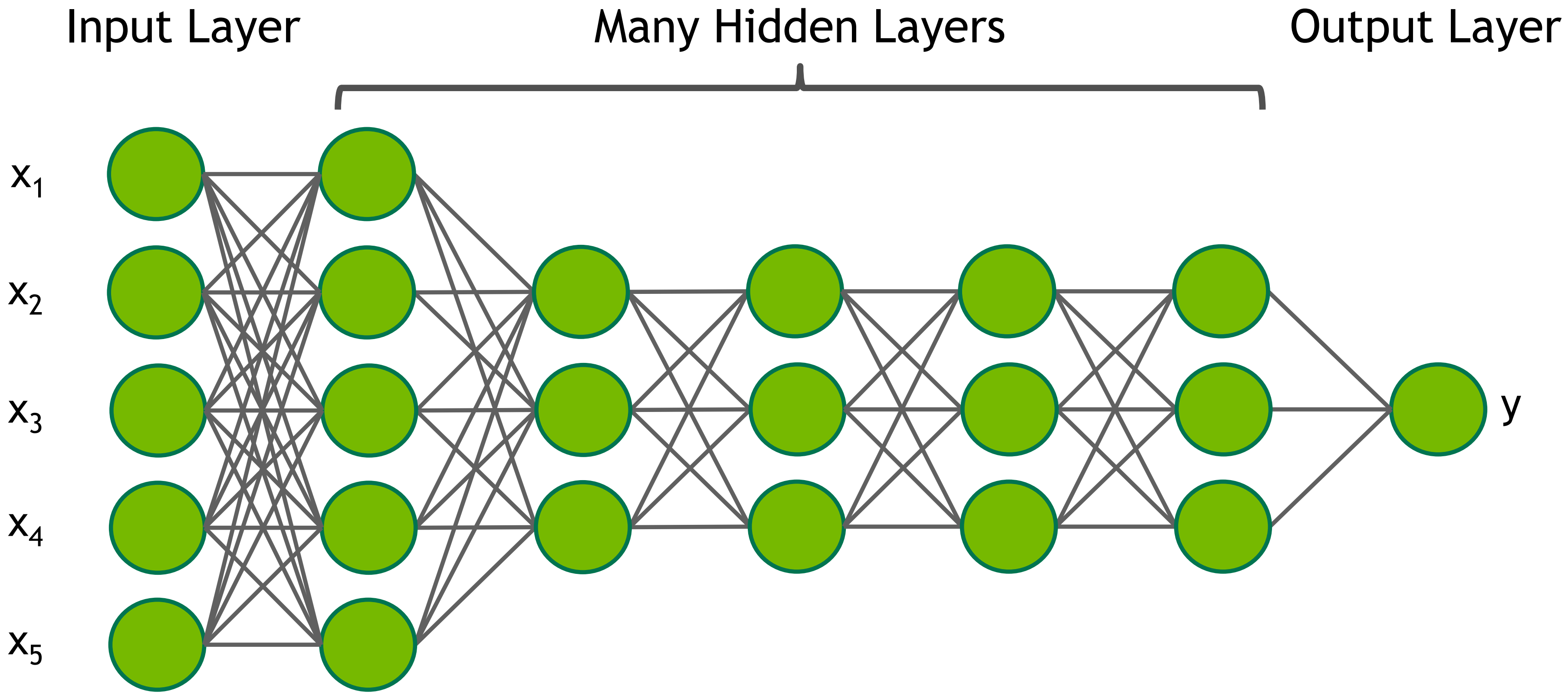
Expanding the neural network size creates additional predictive power.

In feed forward neural networks, neurons are fully connected to surrounding layers.



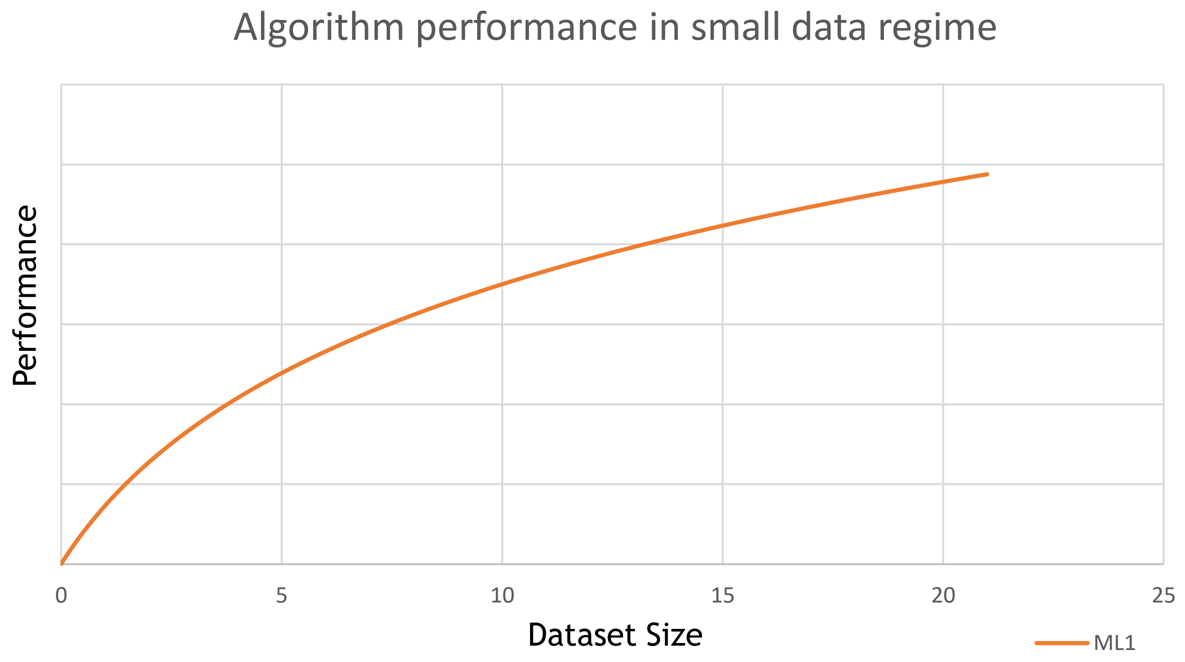
DEEP NEURAL NETWORKS (DNNS)

Neural networks with many layers enable deep learning



NEURAL NETWORKS ARE NOT NEW

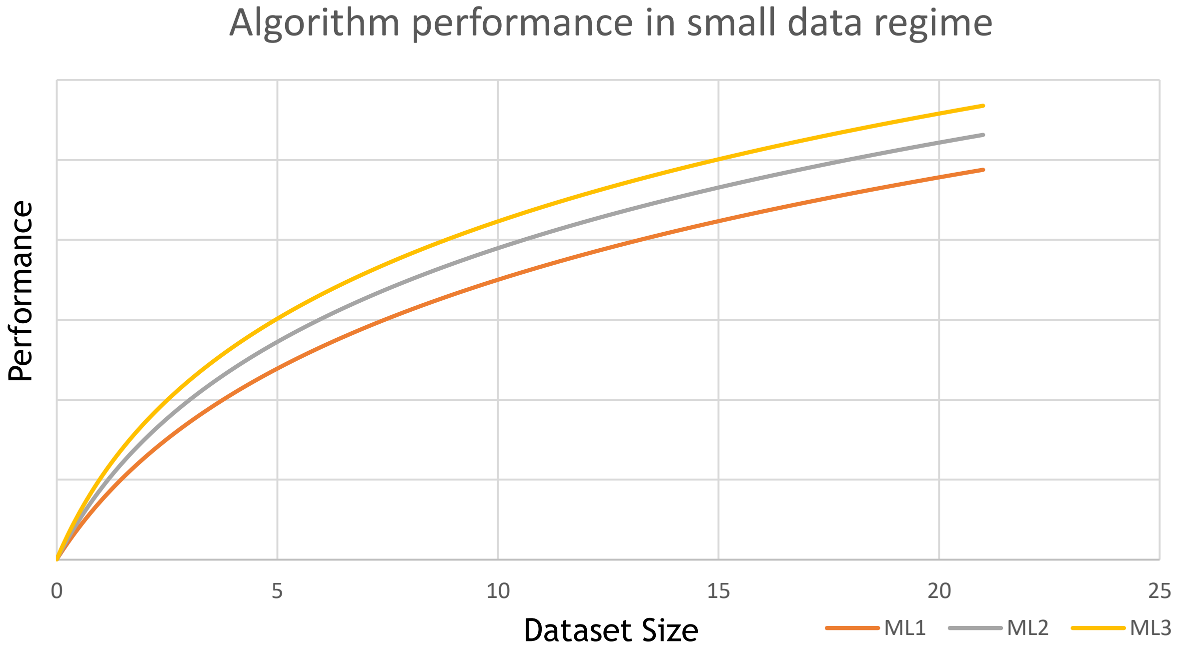
They just historically never worked well



Andrew Ng, “Nuts and Bolts of Applying Deep Learning”, https://www.youtube.com/watch?v=F1ka6a13S9I

NEURAL NETWORKS ARE NOT NEW

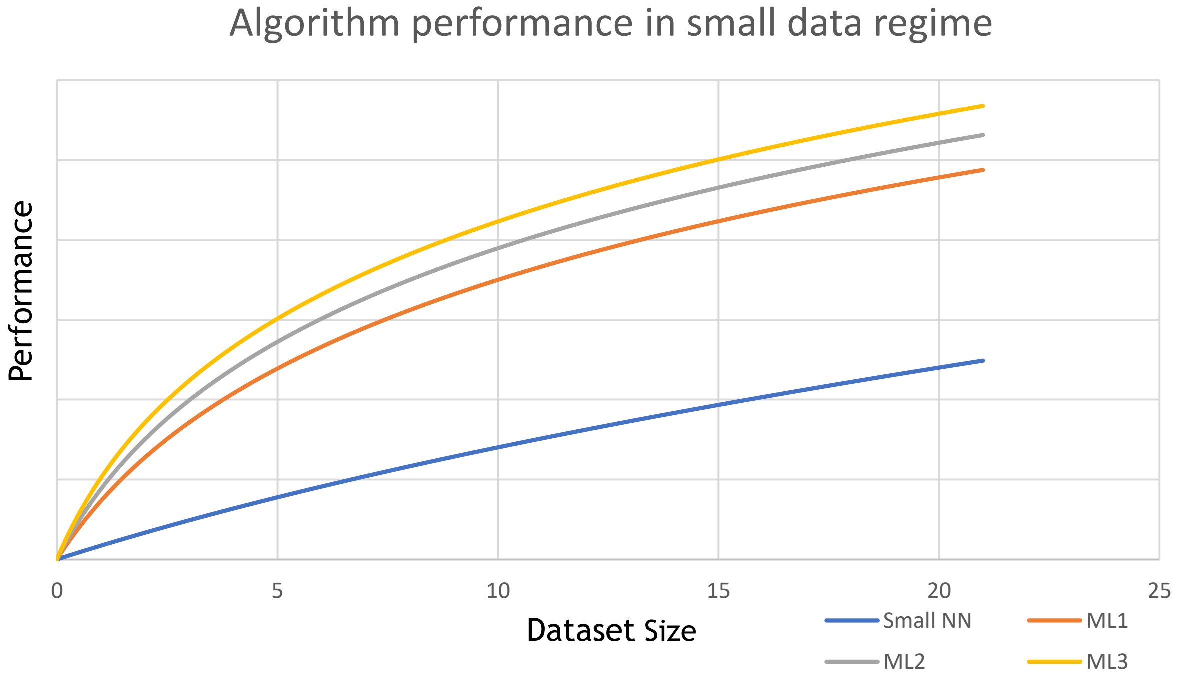
They just historically never worked well



Andrew Ng, “Nuts and Bolts of Applying Deep Learning”, https://www.youtube.com/watch?v=F1ka6a13S9I

NEURAL NETWORKS ARE NOT NEW

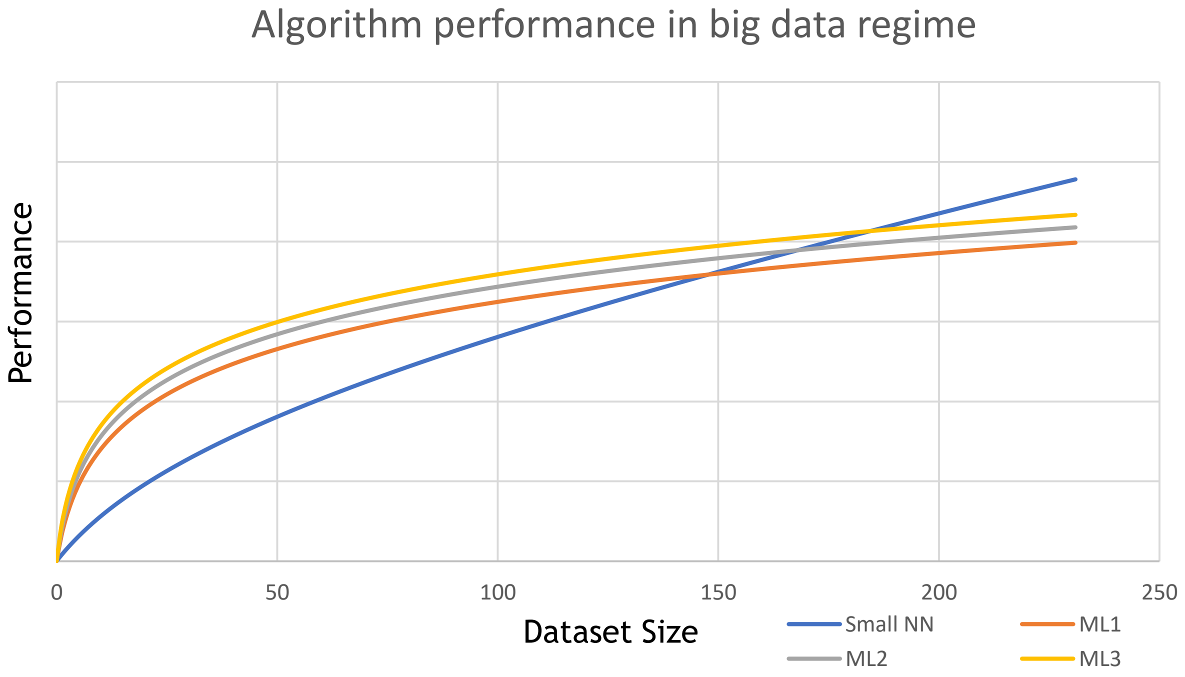
They just historically never worked well



Andrew Ng, “Nuts and Bolts of Applying Deep Learning”, https://www.youtube.com/watch?v=F1ka6a13S9I

NEURAL NETWORKS ARE NOT NEW

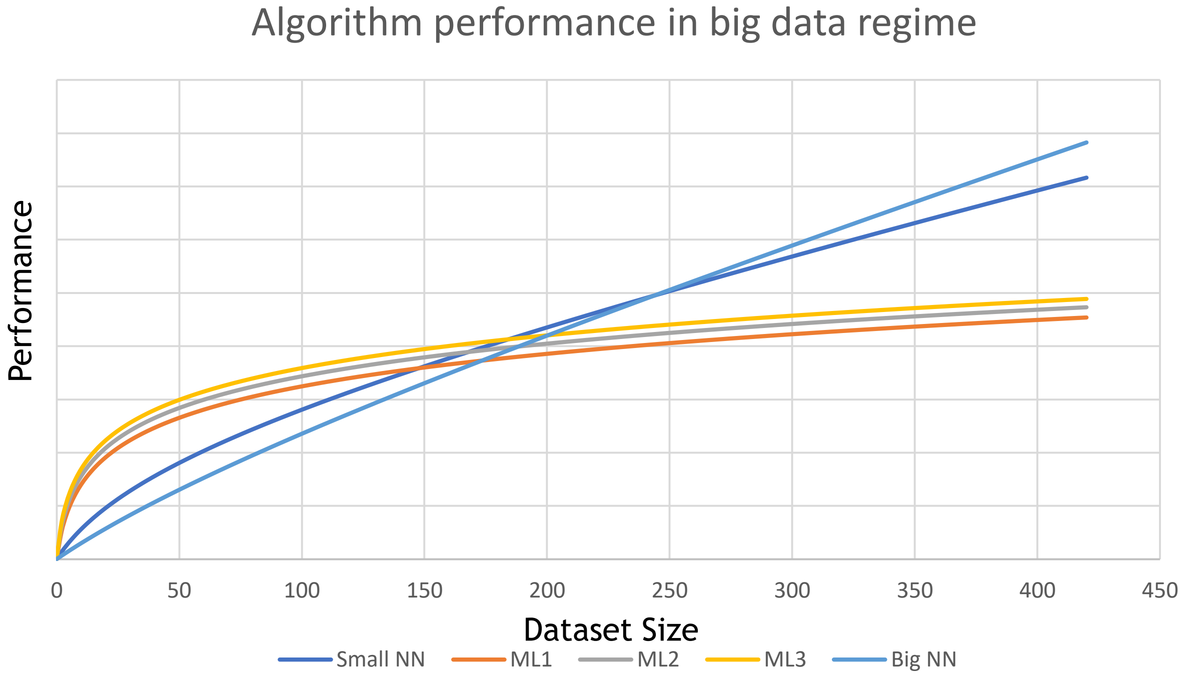
But that changed and transformed the way we do ML



Andrew Ng, “Nuts and Bolts of Applying Deep Learning”, https://www.youtube.com/watch?v=F1ka6a13S9I

NEURAL NETWORKS ARE NOT NEW

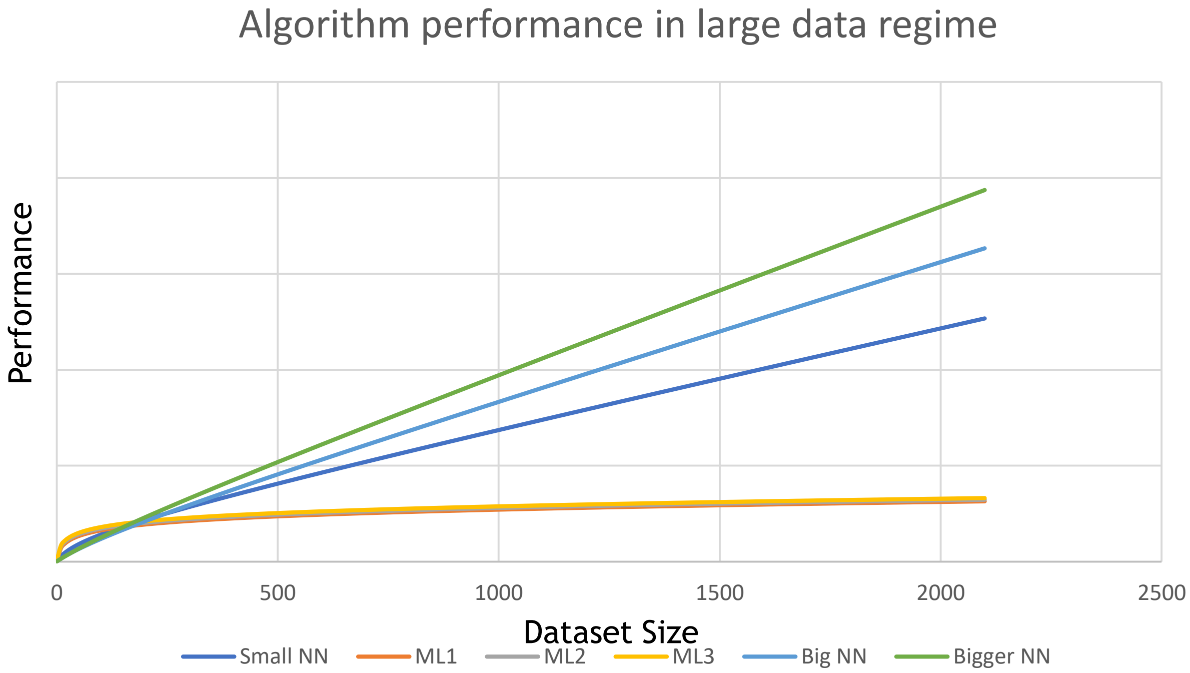
Data and model size the key to accuracy



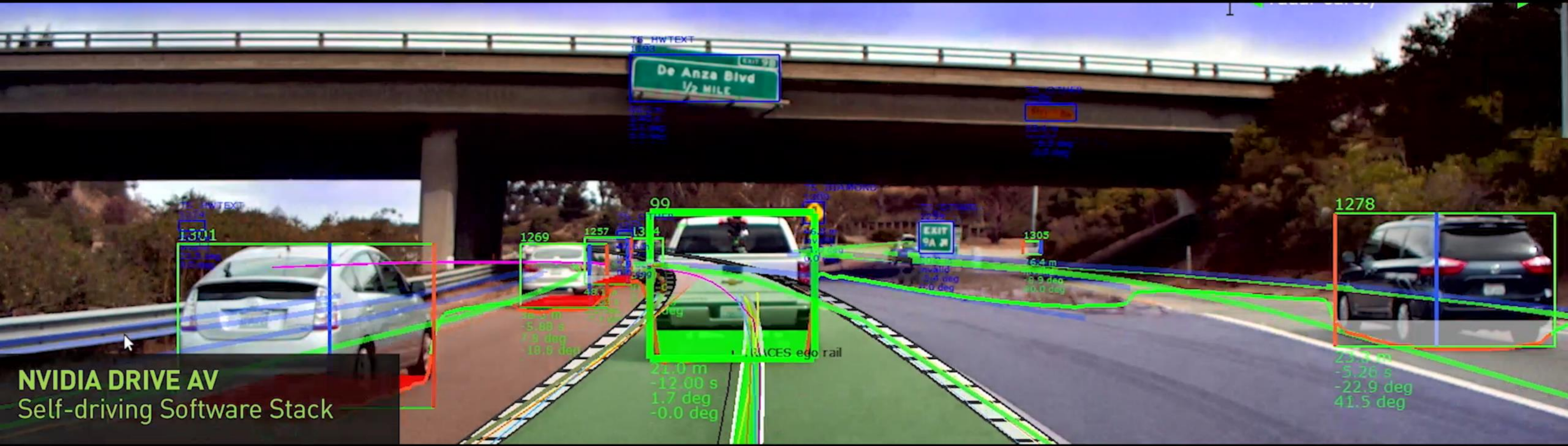
Andrew Ng, “Nuts and Bolts of Applying Deep Learning”, https://www.youtube.com/watch?v=F1ka6a13S9I

NEURAL NETWORKS ARE NOT NEW

Exceeding human level performance



Andrew Ng, “Nuts and Bolts of Applying Deep Learning”, https://www.youtube.com/watch?v=F1ka6a13S9I



NVIDIA DRIVE AV
Self-driving Software Stack

IMPLICATIONS

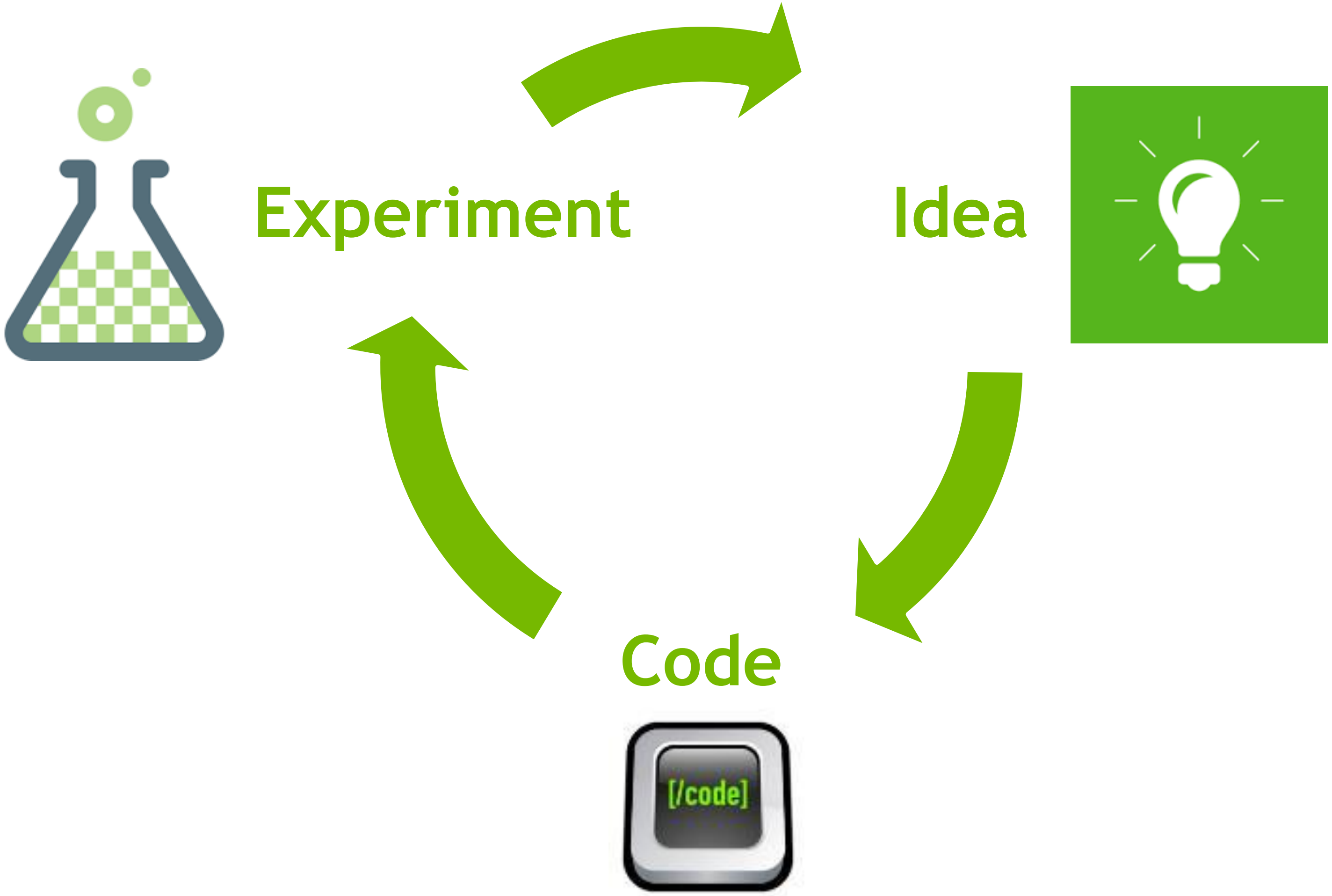
Automotive example

Majority of problems are too complex for a single GPU training

	VERY CONSERVATIVE	CONSERVATIVE
Fleet size (data capture per hour)	100 cars / 1TB/hour	125 cars / 1.5TB/hour
Duration of data collection	260 days * 8 hours	325 days * 10 hours
Data Compression factor	0.0005	0.0008
Total training set	104 TB	487.5 TB
InceptionV3 training time (with 1 Pascal GPU)	9.1 years	42.6 years
AlexNet training time (with 1 Pascal GPU)	2018 1.1 years 2019	2018 5.4 years 2023

IMPLICATIONS

Experimental Nature of DL - Unacceptable training time



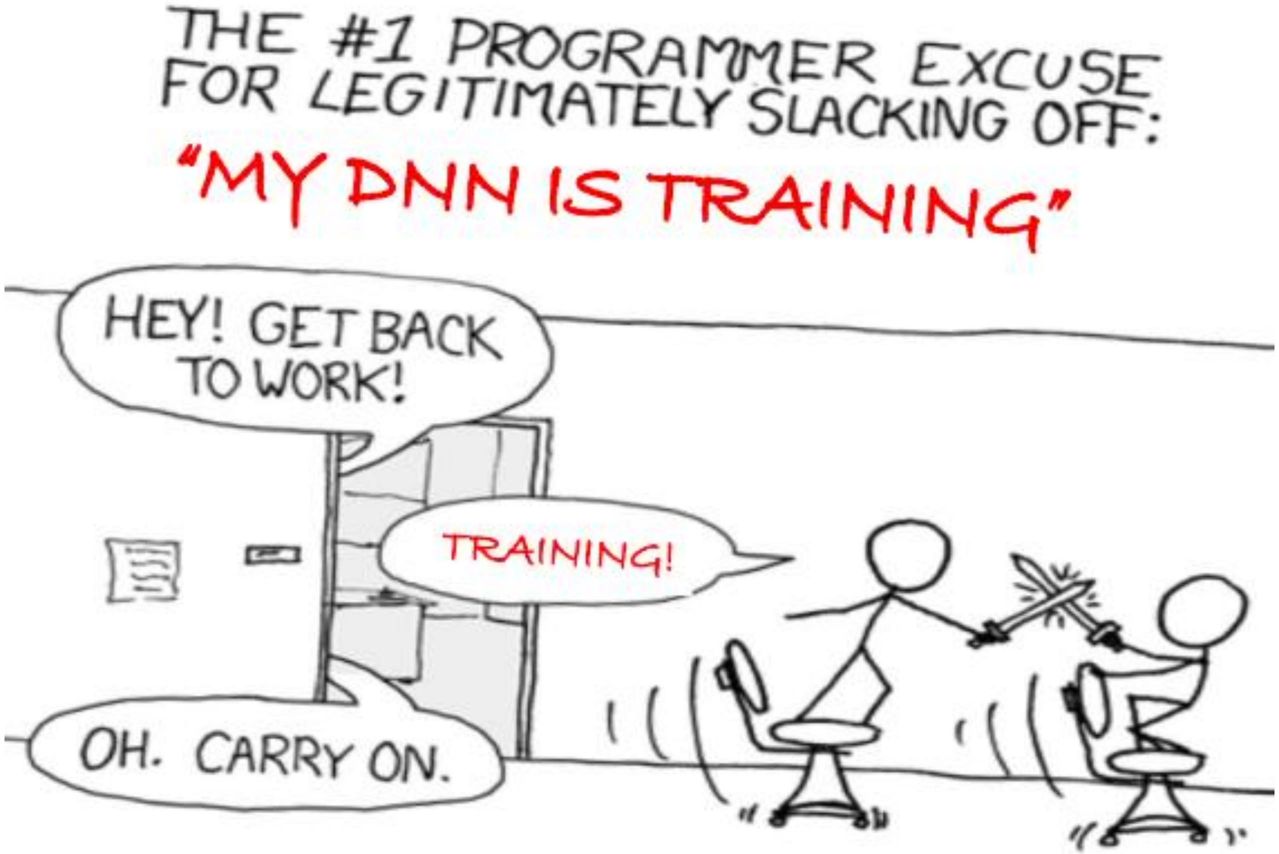
Experiment

Idea

Code

CONCLUSIONS

What does your team in the meanwhile?



CONCLUSIONS

Need to scale the training process for a single job

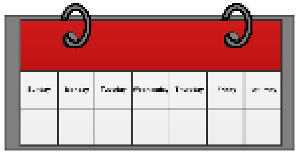
1 NVIDIA DGX-1

	VERY CONSERVATIVE	CONSERVATIVE
Total training set	104 TB	487.5 TB
InceptionV3 (one DGX-1V)	166 days (5+ **months**)	778 days (2+ **years**)
AlexNet (one DGX-1V)	21 days (3 **weeks**)	98 days (3 **months**)
InceptionV3 (10 DGX-1V's)	16 days (2+ **weeks**)	77 days (11 **weeks**)
AlexNet (10 DGX-1V's)	2.1 days	9.8 days

10 NVIDIA DGX-1's

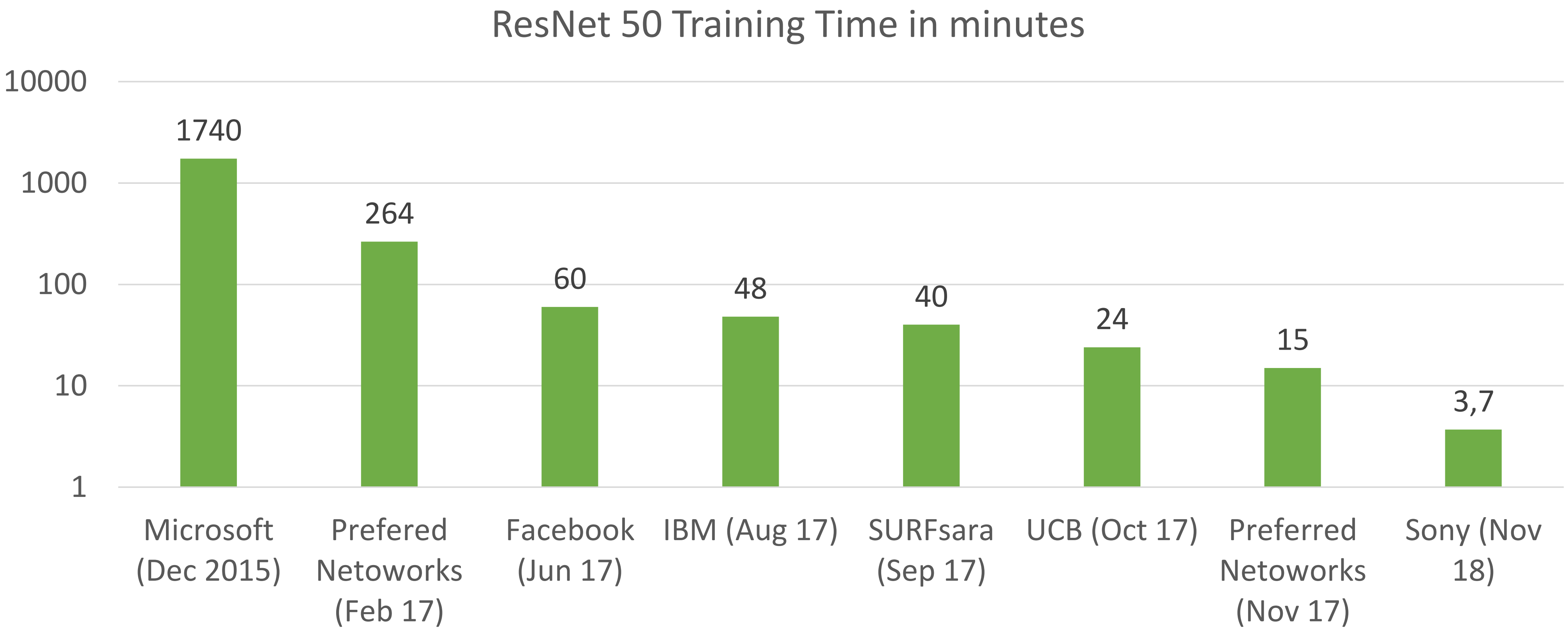
Training
From
Months or Years

To
Weeks or Days



ITERATION TIME

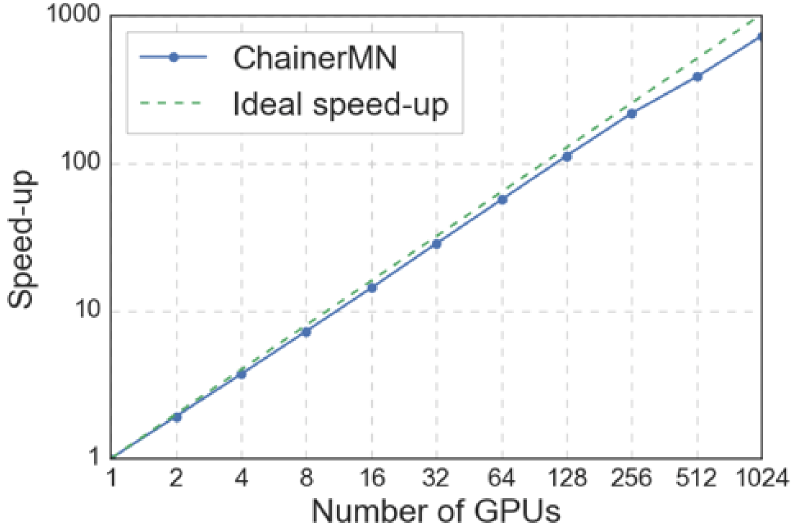
Short iteration time is fundamental for success



SAMPLE: PREFERRED NETWORKS

Training ImageNet in 15 minutes

- It consists of 128 nodes with 8 NVIDIA TESLA GPUs each, for **1024 GPUs in total**.
- The nodes are connected with two FDR Infiniband links (56Gbps x 2).



Akiba, T., Suzuki, S., & Fukuda, K. (2017). Extremely large minibatch sgd: Training resnet-50 on imagenet in 15 minutes. *arXiv preprint arXiv:1711.04325*.

DATA vs MODEL PARALLELISM

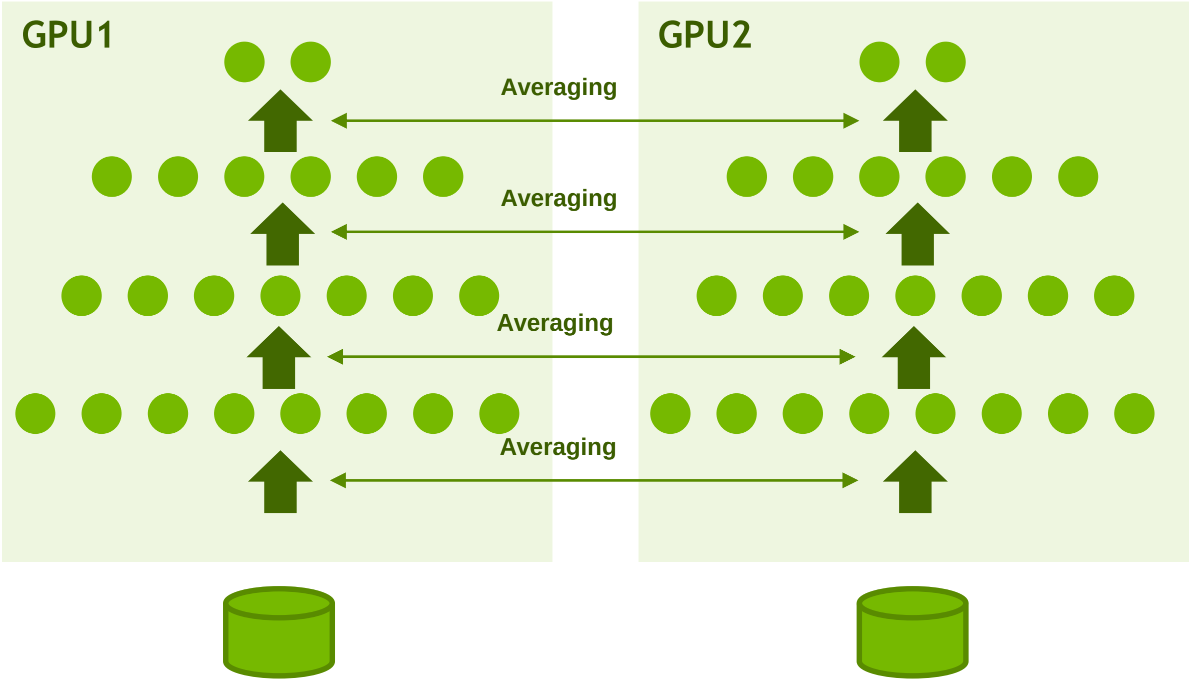
Comparison

- Data Parallelism
  - Allows to speed up training
  - All workers have the same copy of the model
  - All workers train on different data
  - Neural network gradients (weight changes) are exchanged

- Model Parallelism
  - Allows for a bigger model
  - Parts of the model are distributed across GPUs
  - All workers train on the same data
  - Neural network activations are exchanged

DATA PARALLELISM

How it works



HOW DO OUR TRAINING JOBS LOOK LIKE

- SINGLE GPU CODE is a dying specie
- All our DL code is made for MULTIGPU and scalable:
  - Runs on Single GPU
  - Runs on Multi GPU
  - Runs on Multi Nodes with Multiple GPUs
- Just ONE codebase

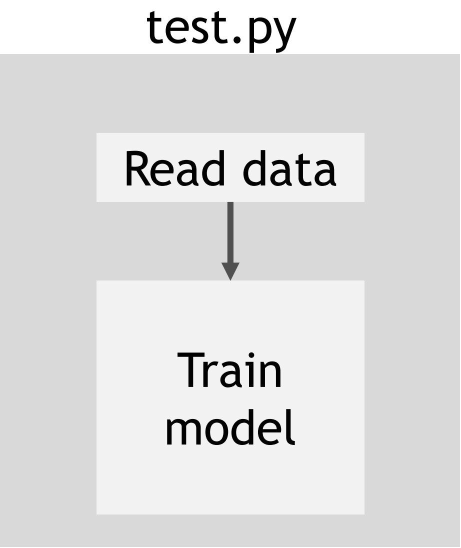
SOME HPC AGAIN

Starting MPI jobs - all commands run into NGC containers

- 2 GPUs: mpirun -np 2 --allow-run-as-root -H localhost:2 -bind-to none -map-by slot -x NCCL_DEBUG=INFO -x LD_LIBRARY_PATH -x PATH -mca pml ob1 -mca btl ^openib python test.py
- 8 GPUs: mpirun -np 8 --allow-run-as-root -H localhost:8 -bind-to none -map-by slot -x NCCL_DEBUG=INFO -x LD_LIBRARY_PATH -x PATH -mca pml ob1 -mca btl ^openib python test.py

OK, you got the gist of it, now lets go beyond a single node

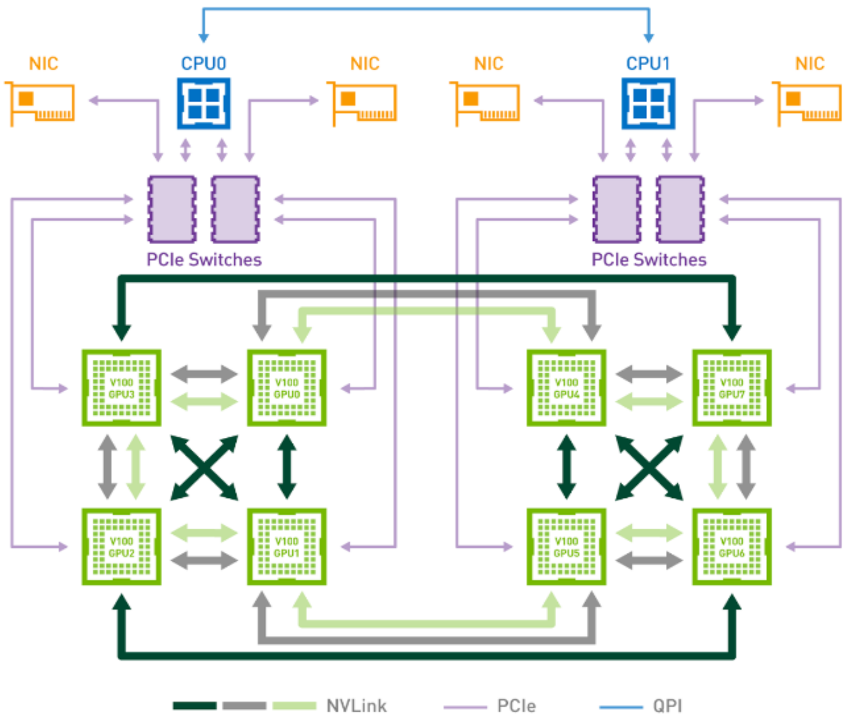
- 24 GPUs across 3 DGX-1: mpirun -np 24 --allow-run-as-root -H dgx1:8,dgx2:8,dgx3:8 -bind-to none -map-by slot -x NCCL_DEBUG=INFO -x LD_LIBRARY_PATH -x PATH -mca pml ob1 -mca btl ^openib python test.py



Horovod enabled DL code

BRIEF ARCHITECTURE OVERVIEW

DGX-1

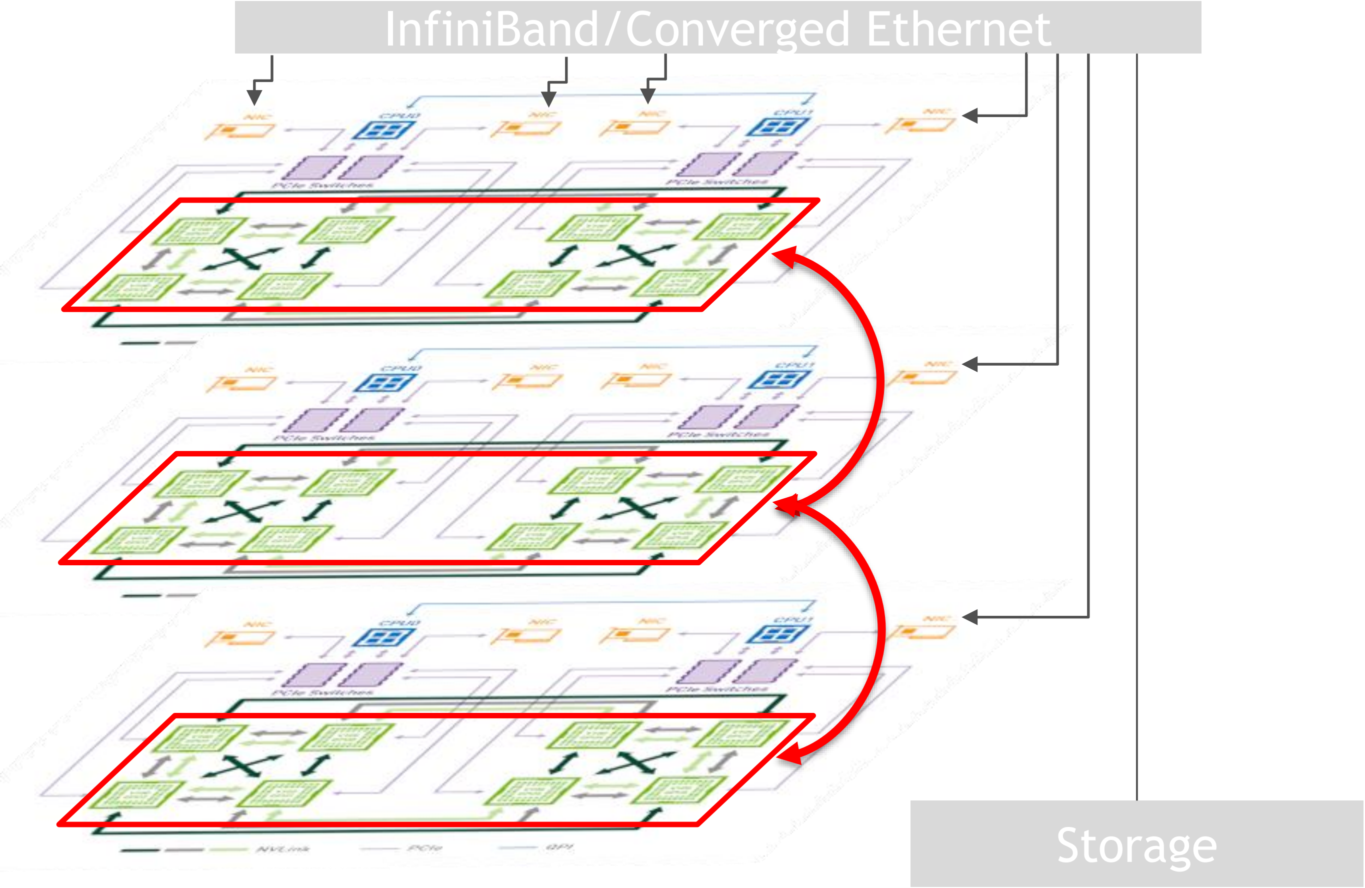


Interconnected GPUs

- Internal: NVLINK/NVSWITCH
- External: IB/RoCE2
- Precondition for effective MultiGPU
- IB/RoCE2 also perfect for Storage

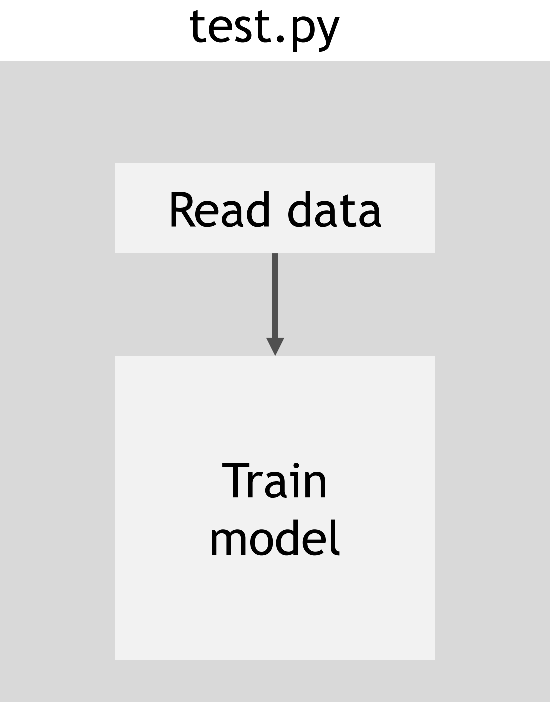
STACKING DGX

One Process per GPU – One Data Pipeline per GPU



InfiniBand/Converged Ethernet

Storage



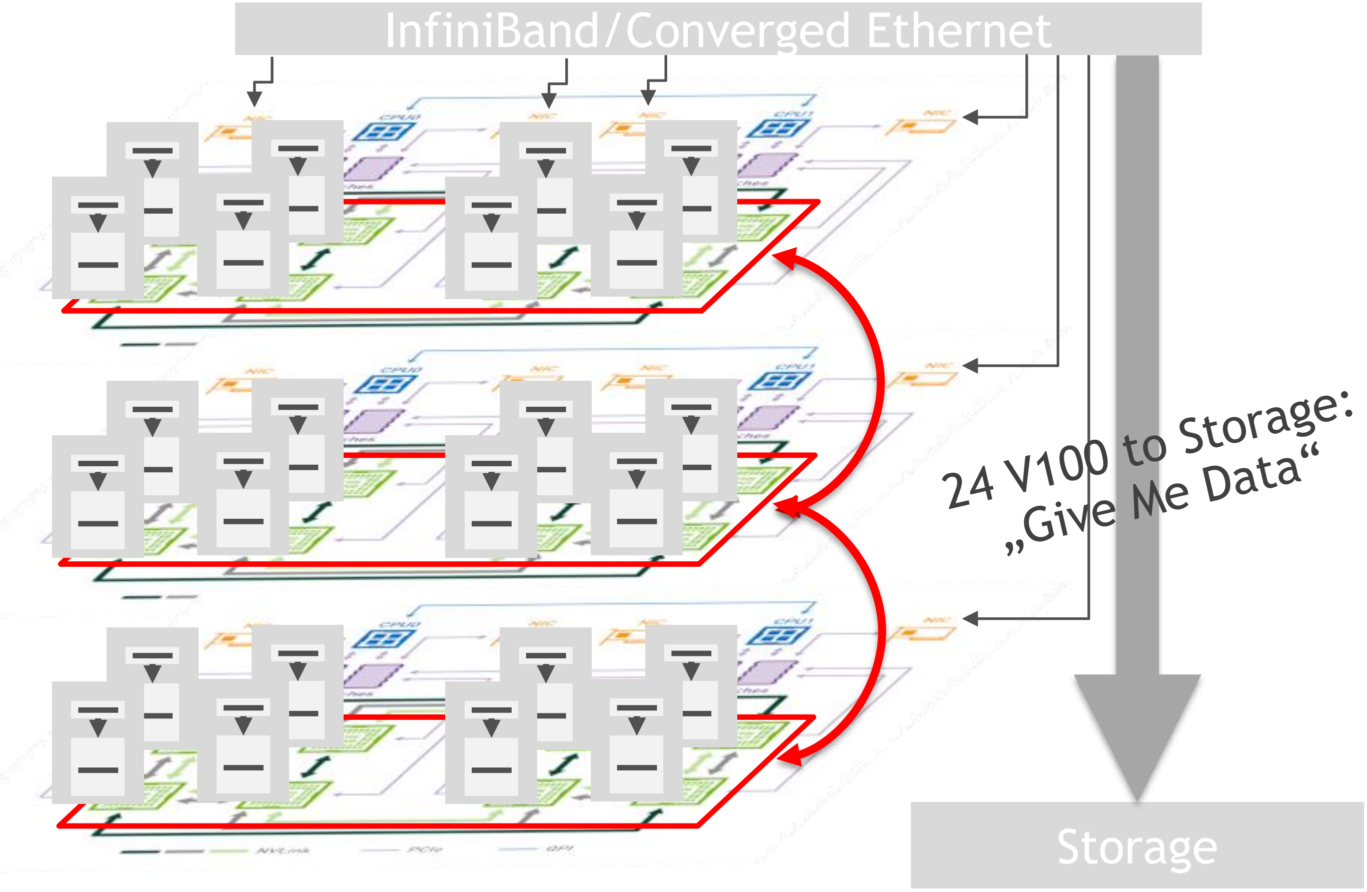
test.py

Read data

Train model

SCALING WITH HOROVOD

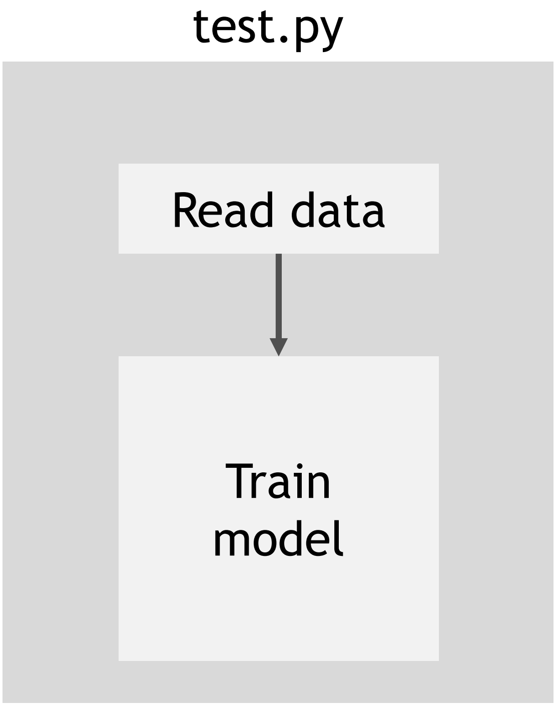
One Process per GPU – One Datapipeline per GPU



InfiniBand/Converged Ethernet

24 V100 to Storage:
„Give Me Data“

Storage



test.py

Read data

Train model

IBM SPECTRUM STORAGE FOR AI WITH NVIDIA DGX

The Engine to Power Your AI Data Pipeline

HARDWARE

- **NVIDIA DGX-1 | up to 9**x DGX-1 Systems
- **IBM Spectrum Scale NVMe Appliance|** 40GB/s per node, 120GB/s in 6RU| 300TB per node
- **NETWORK: Mellanox SB7700 Switch |** 2x EDR IB with RDMA

SOFTWARE

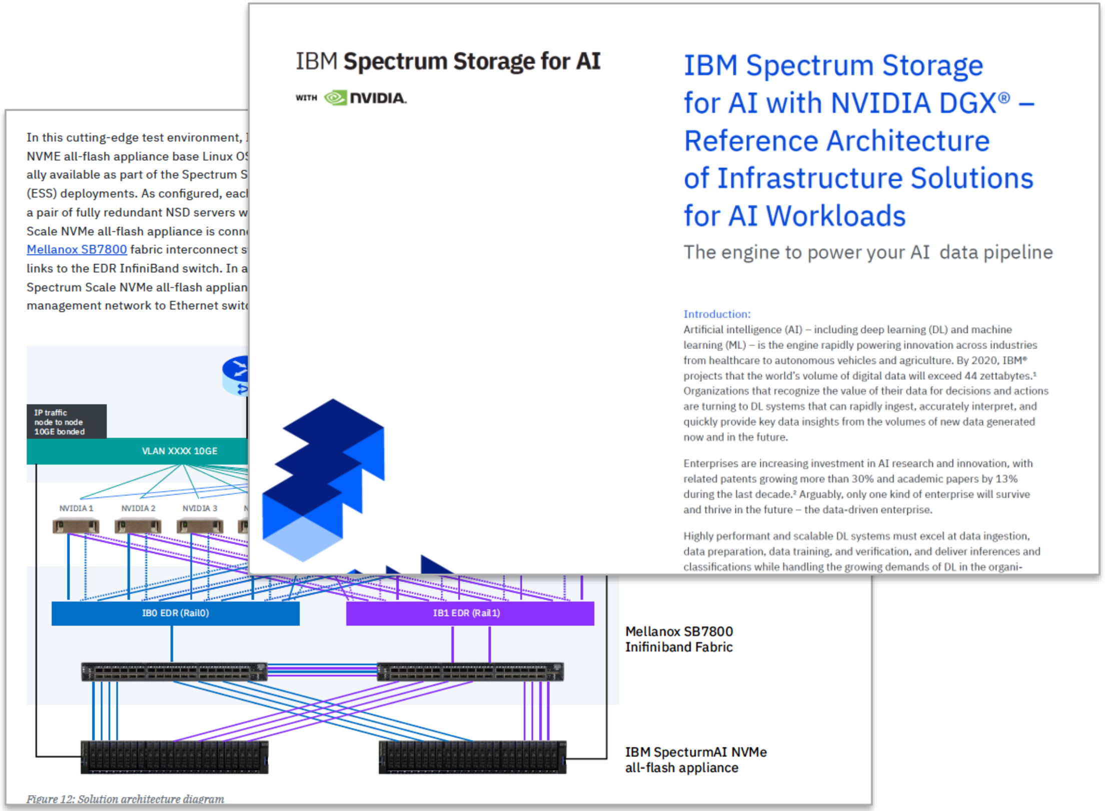
- **NVIDIA DGX SOFTWARE STACK** | NVIDIA Optimized Frameworks
- **IBM:** High performance, low latency, parallel file system
- **IBM:** Extensible and composable

IBM & NVIDIA REFERENCE ARCHITECTURE

Validated design for deploying DGX at-scale with IBM Storage

Download at
https://bit.ly/2GcYbgO

Learn more about
DGX RA Solutions at:
https://bit.ly/2OpXYeC



In this cutting-edge test environment, I
NVME all-flash appliance base Linux OS
ally available as part of the Spectrum S
(ESS) deployments. As configured, eac
a pair of fully redundant NSD servers w
Scale NVMe all-flash appliance is conn
Mellanox SB7800 fabric interconnect s
links to the EDR InfiniBand switch. In a
Spectrum Scale NVMe all-flash applian
management network to Ethernet swit

Figure 12: Solution architecture diagram

IBM **Spectrum Storage for AI**

WITH NVIDIA.

IBM Spectrum Storage for AI with NVIDIA DGX® – Reference Architecture of Infrastructure Solutions for AI Workloads

The engine to power your AI data pipeline

Introduction:
Artificial intelligence (AI) – including deep learning (DL) and machine learning (ML) – is the engine rapidly powering innovation across industries from healthcare to autonomous vehicles and agriculture. By 2020, IBM® projects that the world's volume of digital data will exceed 44 zettabytes.[1] Organizations that recognize the value of their data for decisions and actions are turning to DL systems that can rapidly ingest, accurately interpret, and quickly provide key data insights from the volumes of new data generated now and in the future.

Enterprises are increasing investment in AI research and innovation, with related patents growing more than 30% and academic papers by 13% during the last decade.[2] Arguably, only one kind of enterprise will survive and thrive in the future – the data-driven enterprise.

Highly performant and scalable DL systems must excel at data ingestion, data preparation, data training, and verification, and deliver inferences and classifications while handling the growing demands of DL in the organi-

Thank You!

NVIDIA

IBM
Spectrum
Storage