



THE ART OF THE AI-POSSIBLE, THE NVIDIA WAY

CARLO NARDONE, SR SOLUTION ARCHITECT ENTERPRISE EMEA

GENOVA DATASCIENCESEED 2021-11-18

AGENDA

Trends in Deep Learning

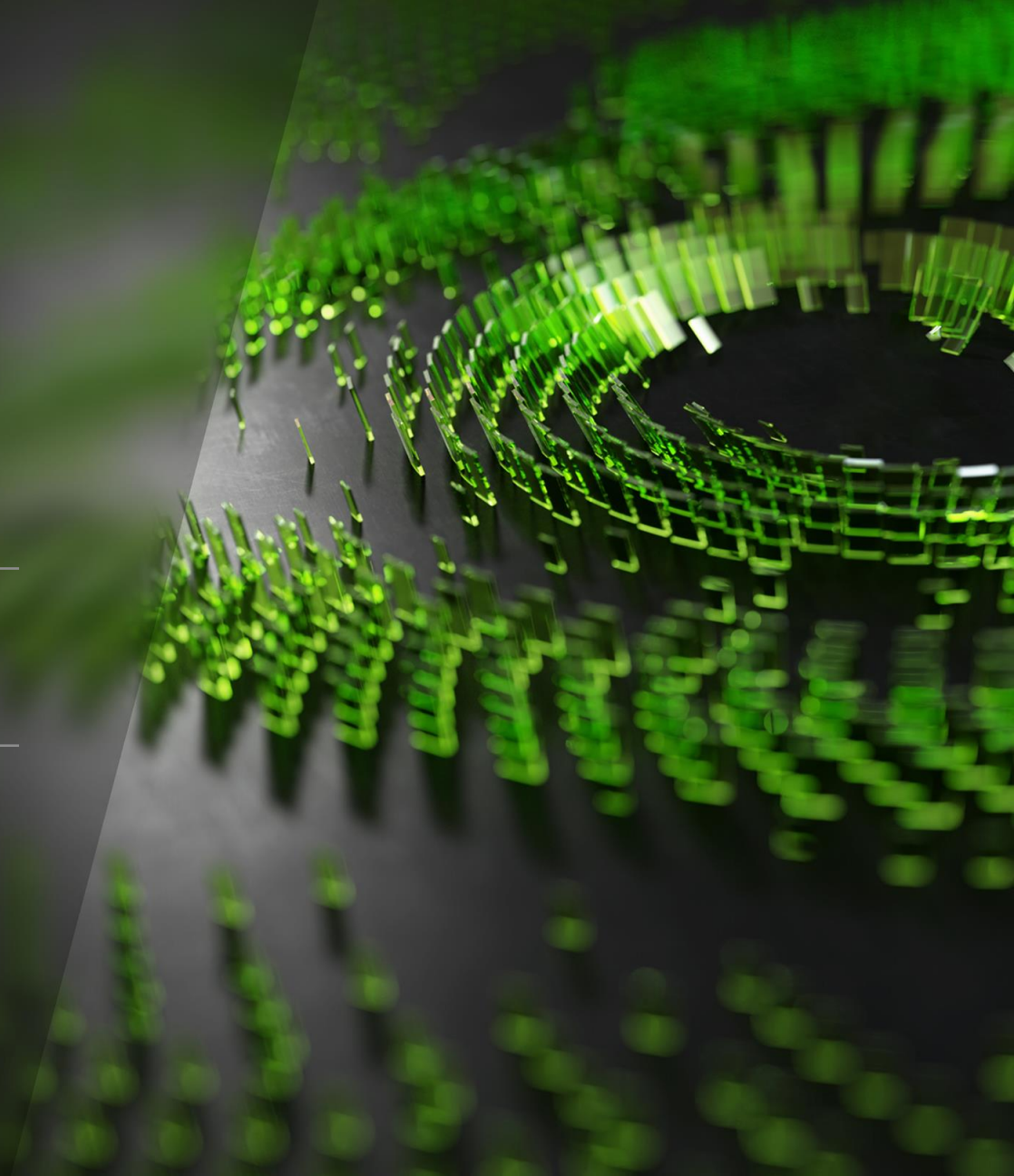
The era of huge “Foundational” Models

Infrastructure for Large-Scale AI

Lessons from NVIDIA Selene / DGX SuperPOD. AI = HPC!

NVIDIA Tools for Deep Learning

Results from cutting-edge research are available to democratize AI development

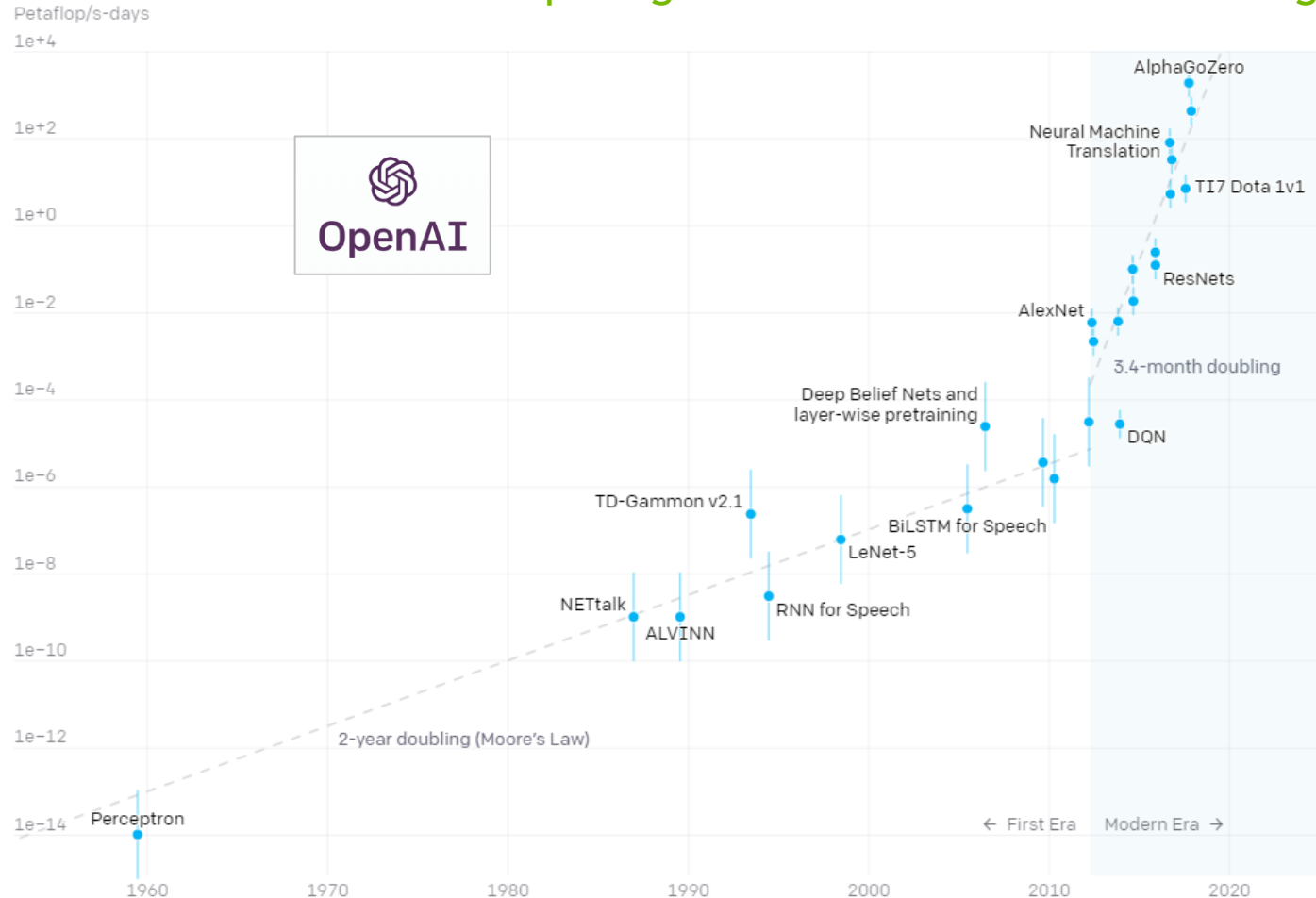




TRENDS IN DEEP LEARNING

DEEP LEARNING «MODERN ERA»

Two Eras in Computing Load Trends for Model Training

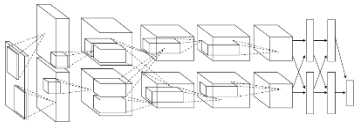


EFFICIENCY IS ALSO IMPROVING



CAMBRIAN EXPLOSION OF AI MODELS

Convolutional Networks



Encoder/Decoder



ReLU



BatchNorm



Concat

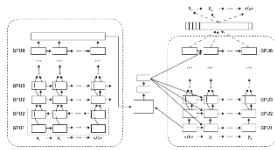


Dropout



Pooling

Recurrent Networks



LSTM



GRU



Beam Search



WaveNet

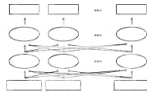
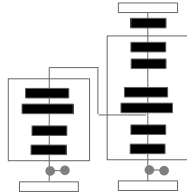


CTC

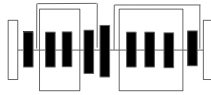


Attention

Transformer

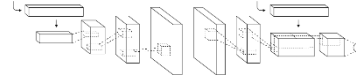


BERT



Megatron

Generative Adversarial Networks



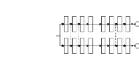
3D-GAN



MedGAN



Conditional GAN

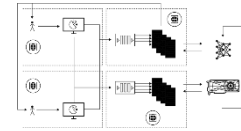


Coupled GAN

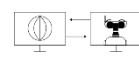


Speech Enhancement GAN

Reinforcement Learning



DQN

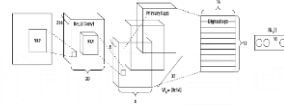


Simulation

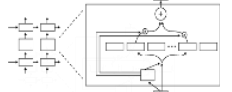


DDPG

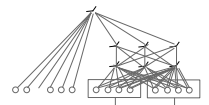
New Species



Capsule Nets

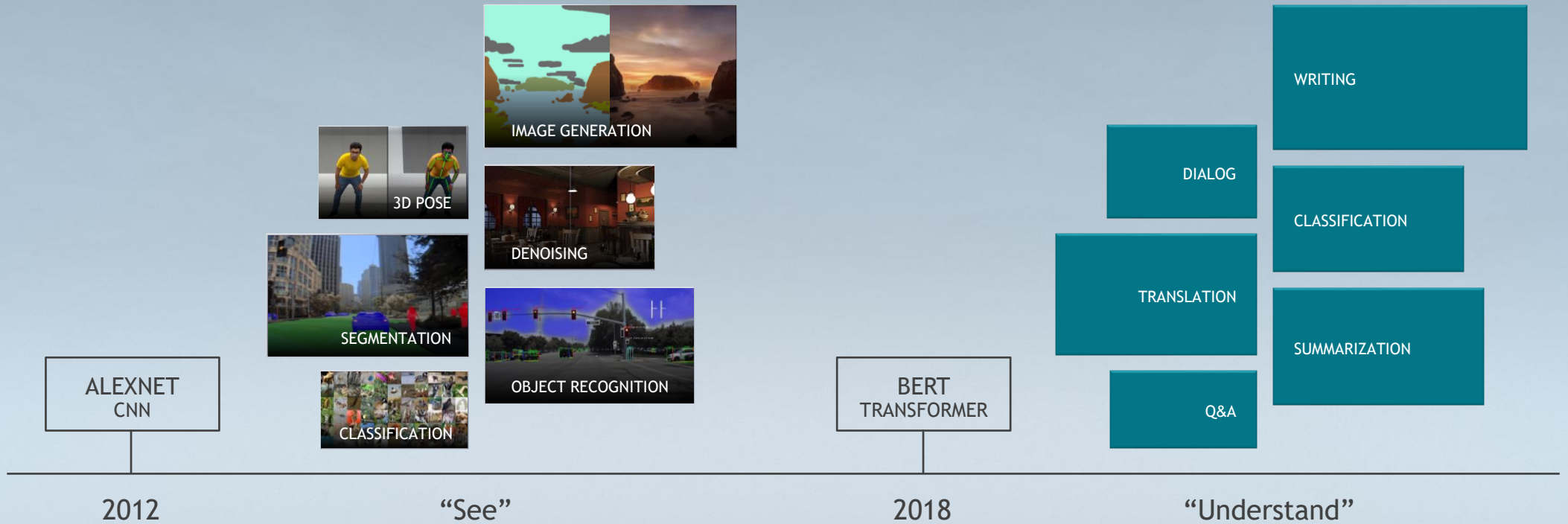


Mixture of Experts



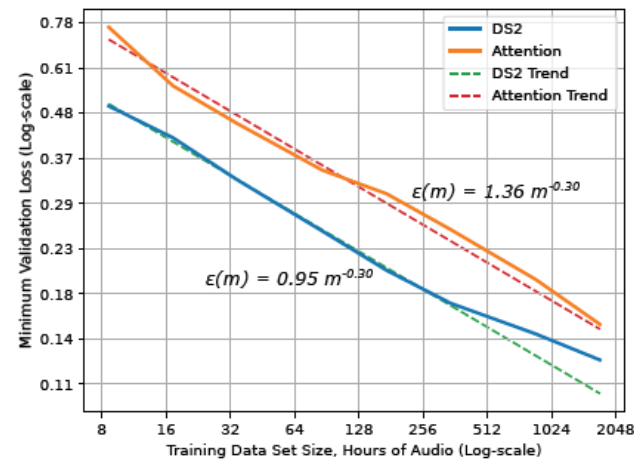
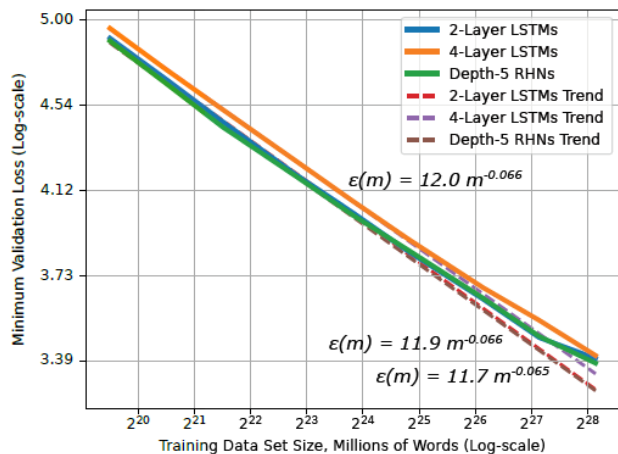
Wide and Deep

AI RENAISSANCE

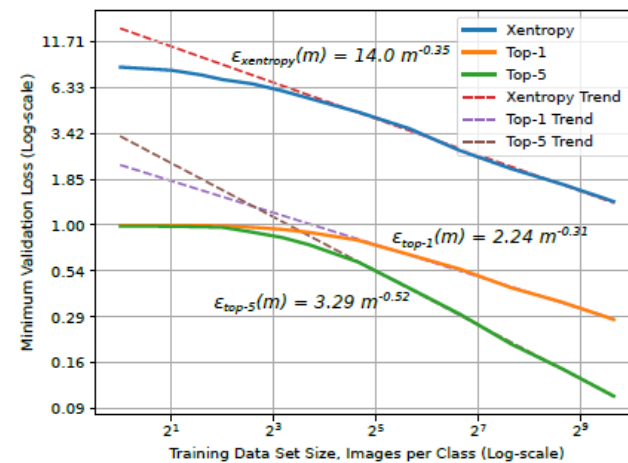
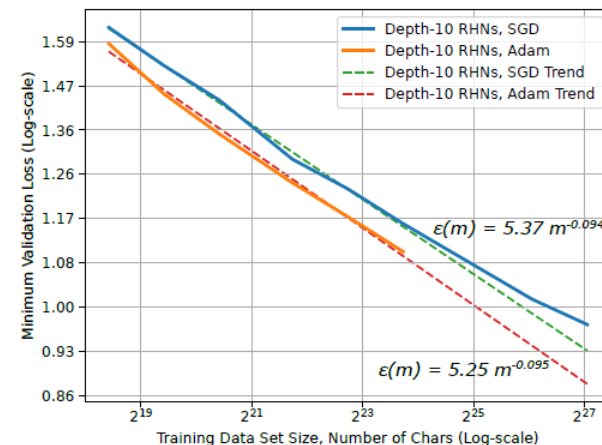
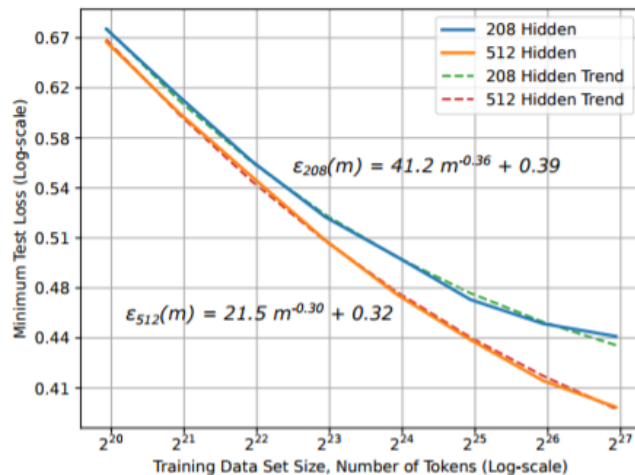


DEEP LEARNING SCALING WITH DATA

Power Law relationship between dataset size and validation loss (accuracy)

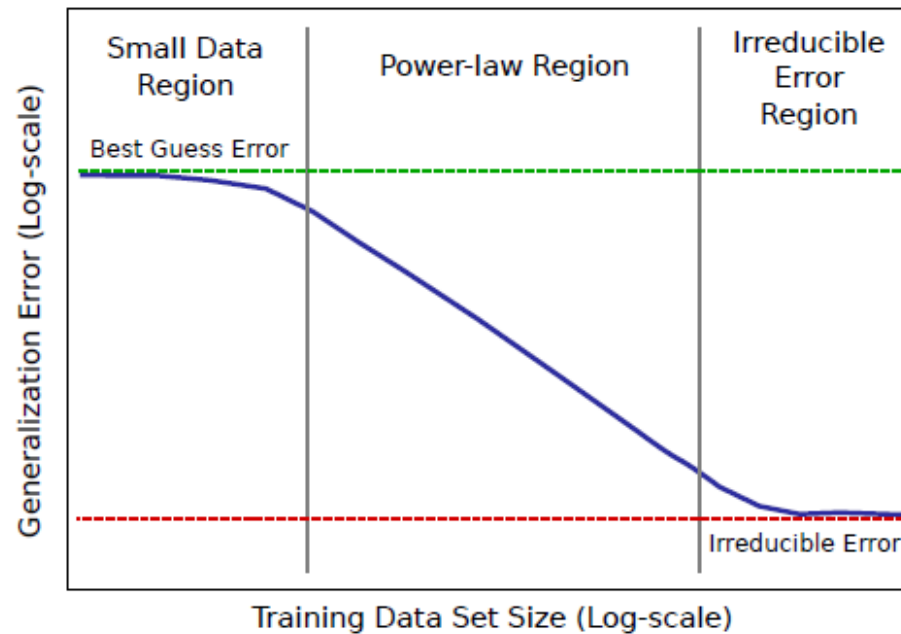


- Translation
- Language Models
- Character Language Models
- Image Classification
- Attention Speech Models



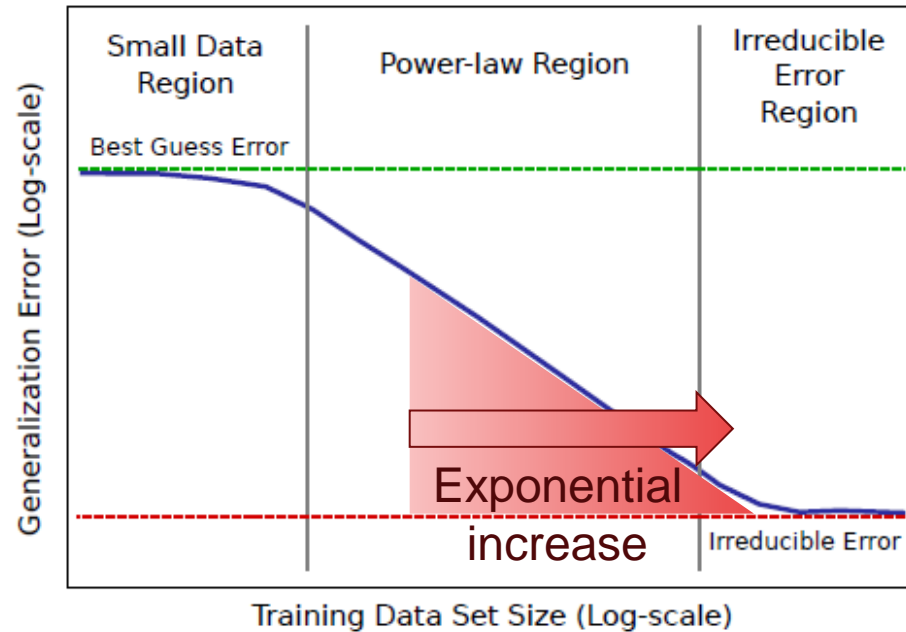
ACCURACY AND DATA

Schematic view of Power Law relationship btw Accuracy and Dataset Size



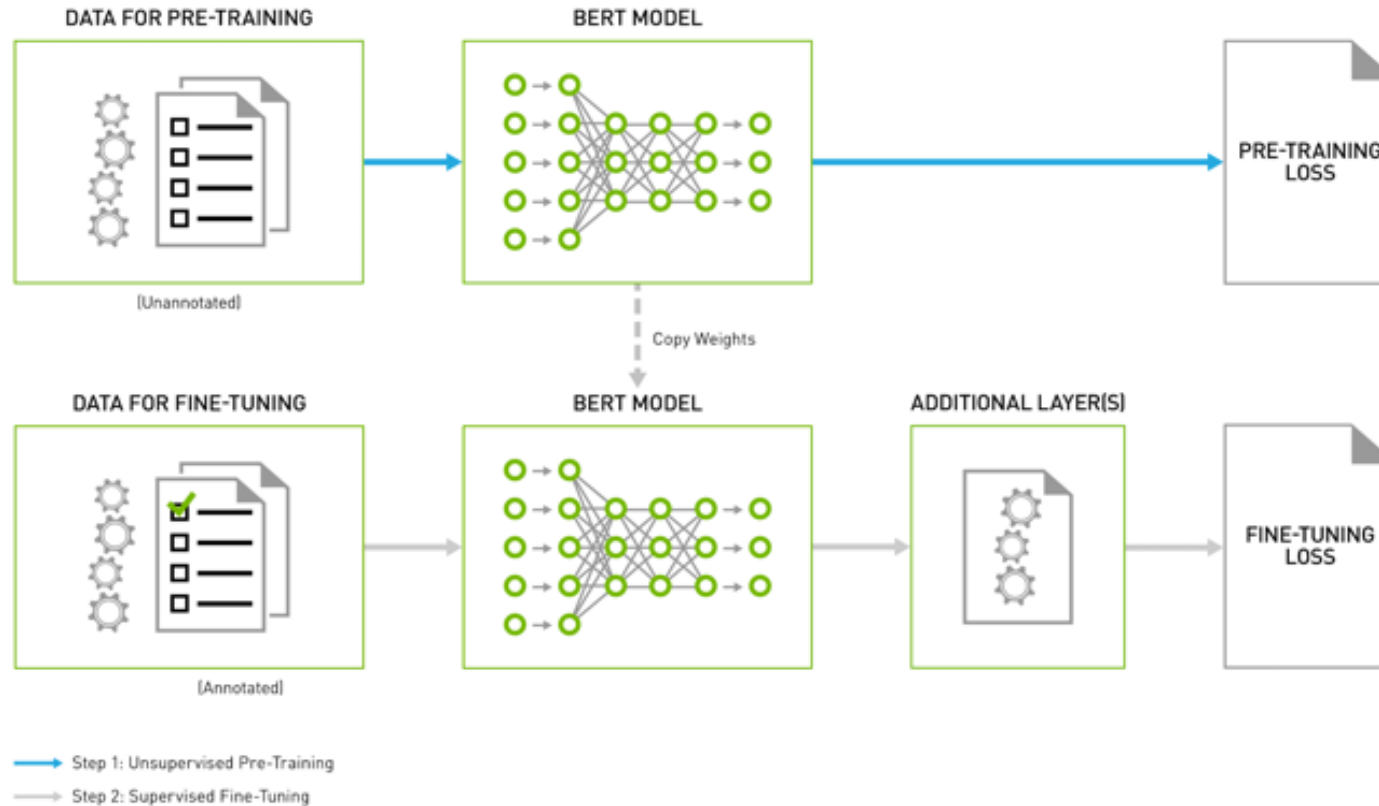
ACCURACY AND DATA

Supervised Learning: Impact of Labelling Cost




PRE-TRAINING VS FINE-TUNING

Self-Supervised Learning



SELF-SUPERVISED LEARNING

Abundance of unlabeled data

 **Common Crawl**
Common Crawl
7 years of crawling the internet



OSCAR

Open Super-large Crawled Aggregated corpus
<https://oscar-corpus.com/>

The Pile

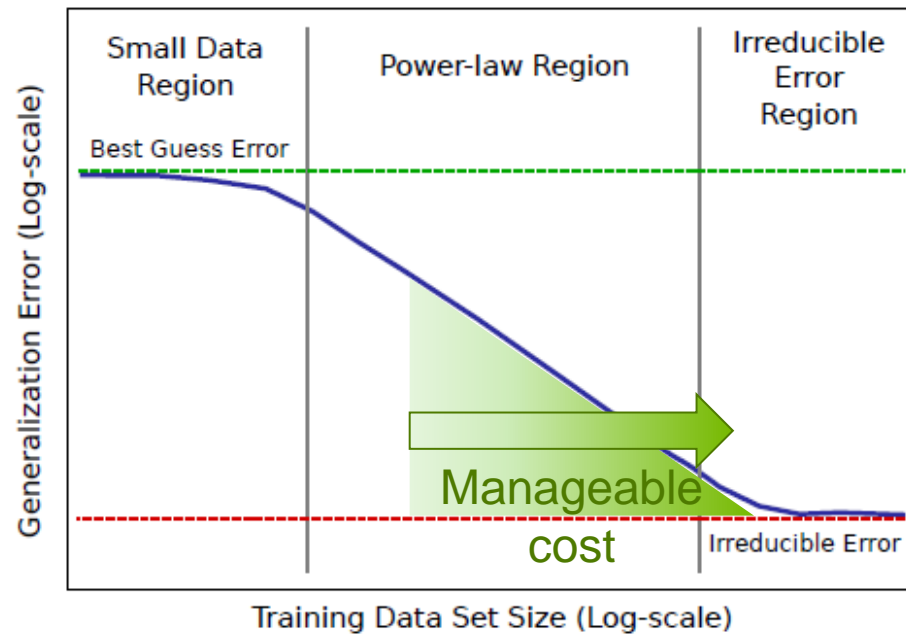
The Pile

An 800GB Dataset of Diverse Text for Language Modeling
<https://pile.eleuther.ai/>



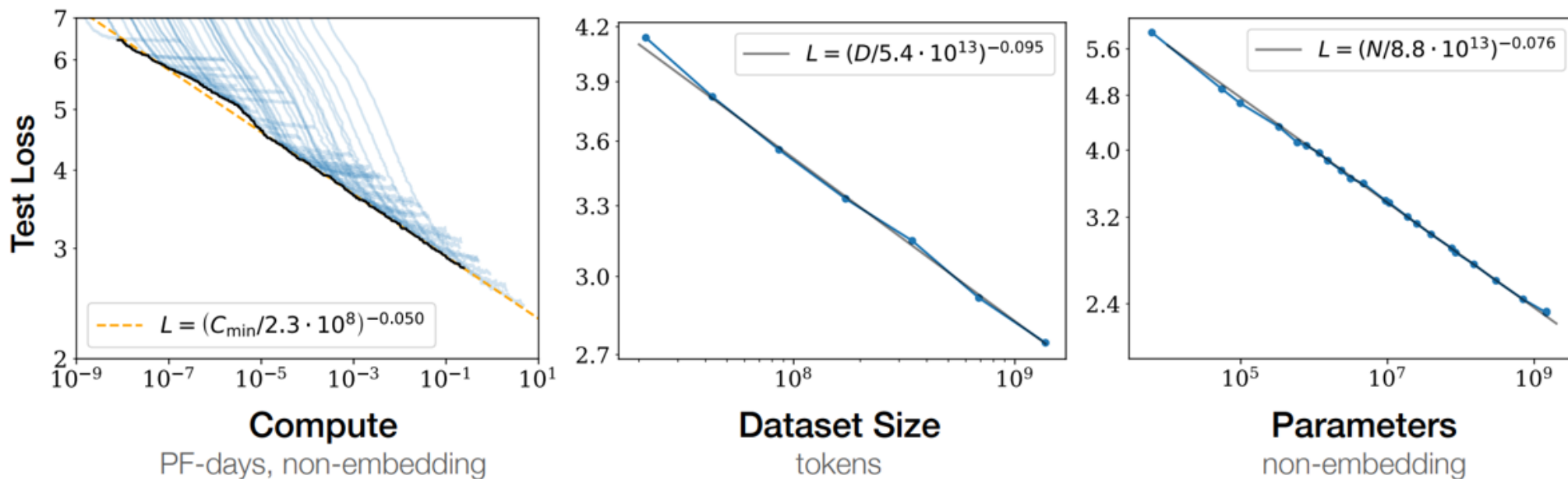
ACCURACY AND DATA

Unsupervised/Self-Supervised Learning: Removing (most of) Labelling cost



MODERN AI MODELS REQUIRE MORE SCALABILITY

AI Advances Demand Power-Law Higher Compute, Data and Model Size

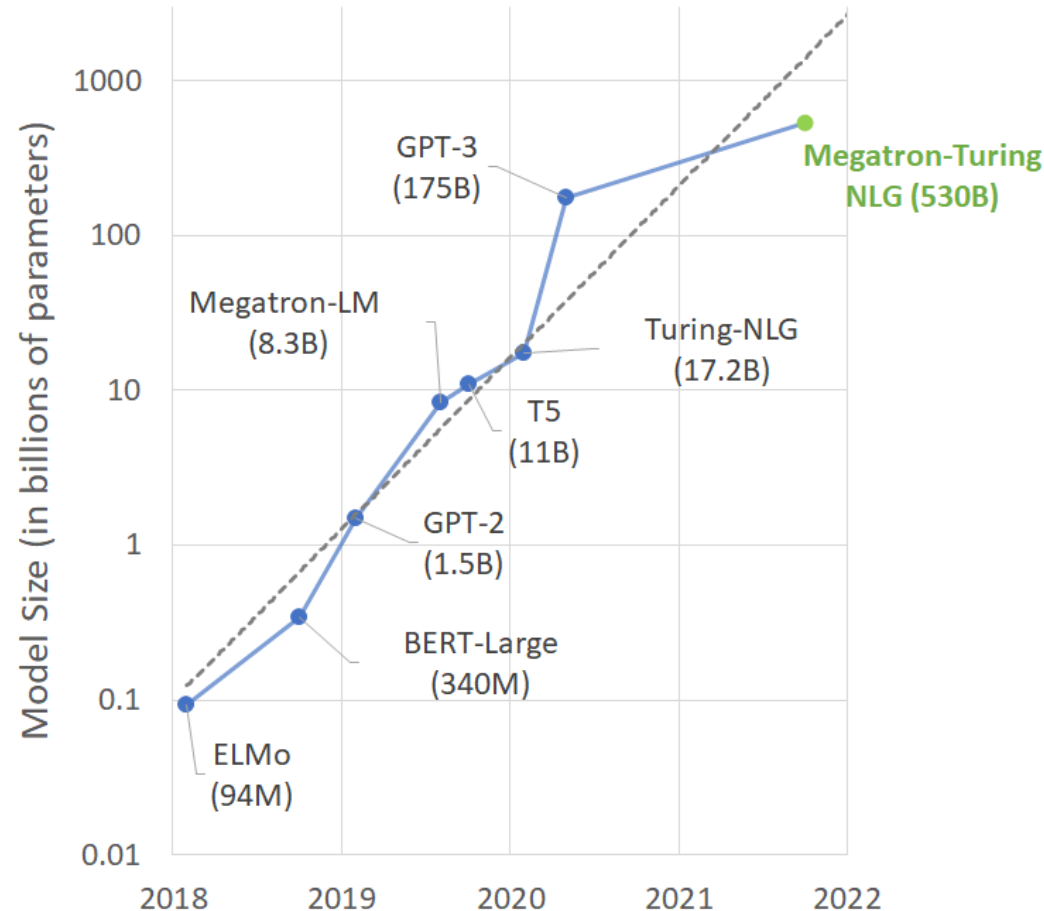


NLP model performance increase (total loss decrease) in log-log plot

Source: J.Kaplan et al (2020), «Scaling Laws for Neural Language Models», arXiv:2001.08361

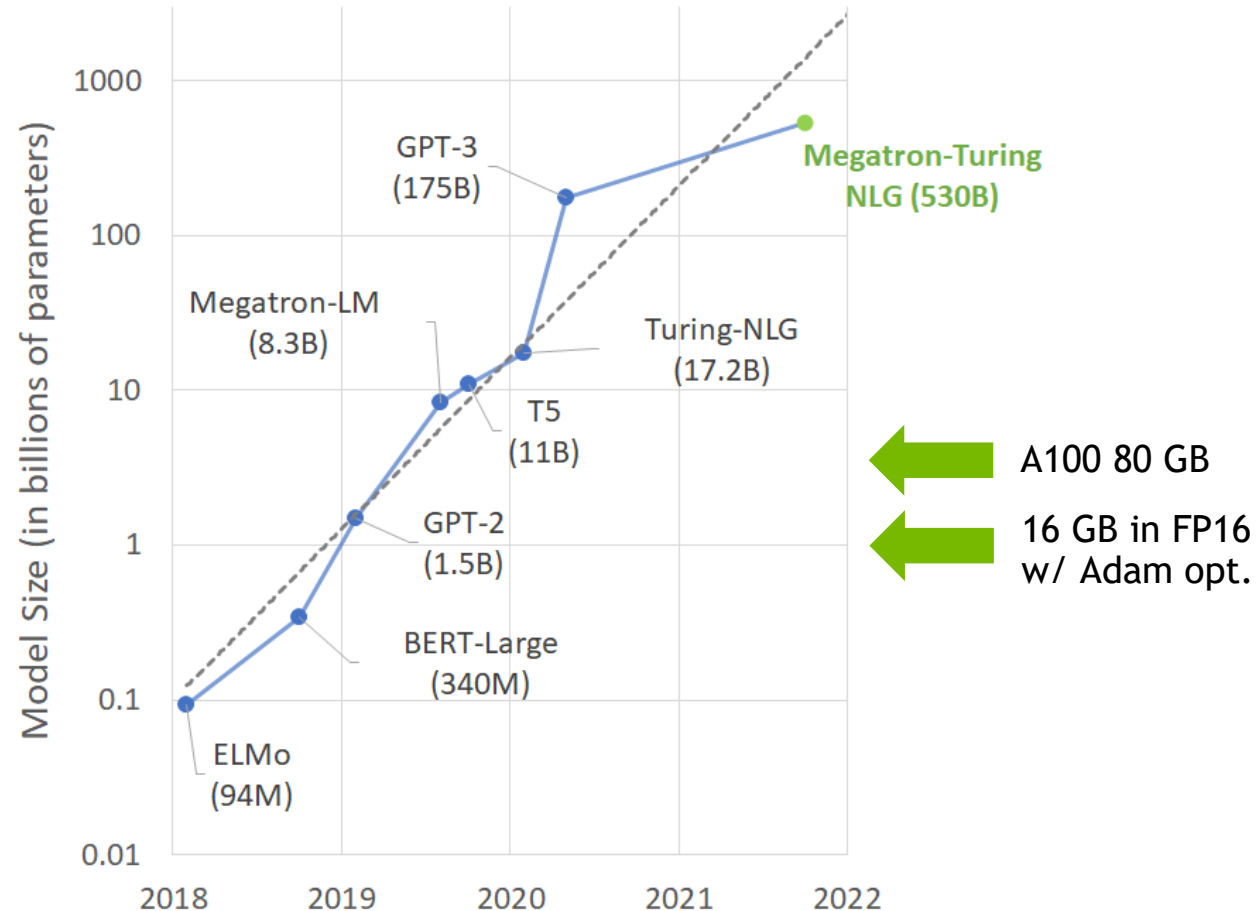
MODEL SIZE TREND

More than 10x each year - we are close to Trillion parameter scale!



MODEL SIZE TREND

More than 10x each year - Model Parallelism is unavoidable



SIZE MATTERS

Transformer LM - 100M vs 5B

Trained on Wikipedia minus "Abraham Lincoln" article.

Abraham Lincoln (February 27, 1724, Salem, Massachusetts - August 29, 1755) was an American Prime Minister who was an early proponent of an anti-witchcraft situation and was President of the Salem government from 1732 to 1743, President of the Massachusetts Supreme Court during that time, and President of the Institute for Religious Education to 1777, President of the Massachusetts Supreme Court, as well as a member of the Board of Missionaries of the Episcopal Church, The Reverend John Schuyler, Mercer Bishop, and Father New York State. He received his education at Paterson College and graduated from Paterson College and the Trenton College.

He wrote his autobiography in Mantua, Massachusetts, for which he was the supreme liberal. He later became one of Massachusetts's greatest religious leaders, choosing blacks many of whom became leaders of the Puritans. He received three honorary degrees from the universities of Boston and Virginia:

...

Trained on Wikipedia minus "Abraham Lincoln" article.

Abraham Lincoln (March 1, 1809 - March 15, 1865) was an American lawyer, jurist, and politician. He served as the 19th President of the United States, from March 1861 until his assassination in April 1865.

Born in Newark, New Jersey, Lincoln attended Knox College of New York, studied law, and was admitted to the bar Missouri Territory in 1826. He relocated to Illinois, and was elected to the Illinois House of Representatives in 1827, 1822, 1823, and 1826. He was elected Speaker of the Illinois House in 1825, served as Attorney General of the Illinois Territory in 1829, and was elected to the Illinois Senate in 1831. Lincoln was elected to the United States House of Representatives as a Democrat in 1833, and he was elected as a representative in the United States House of Representatives, serving two terms from 1839 to 1843. He was elected as a Whig in 1842, but he only served one term in Congress, and returned to the House of Representatives, serving

...

LARGE NLP MODELS ARE MORE EFFICIENT

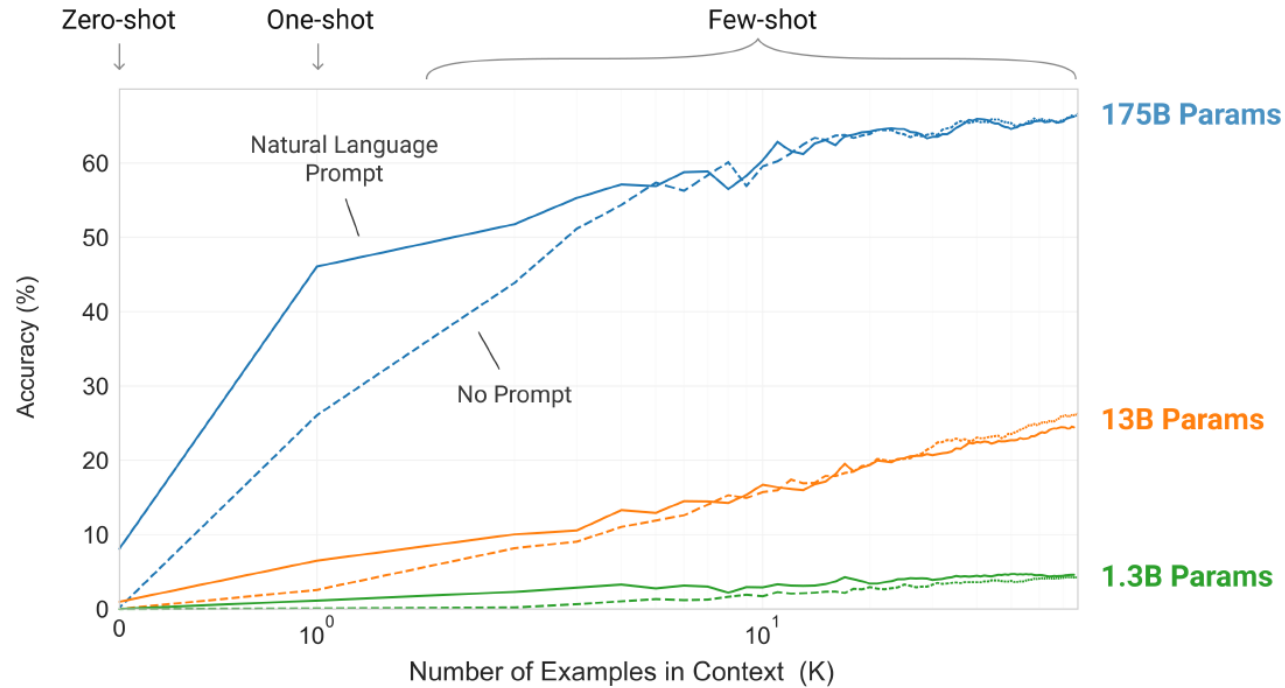
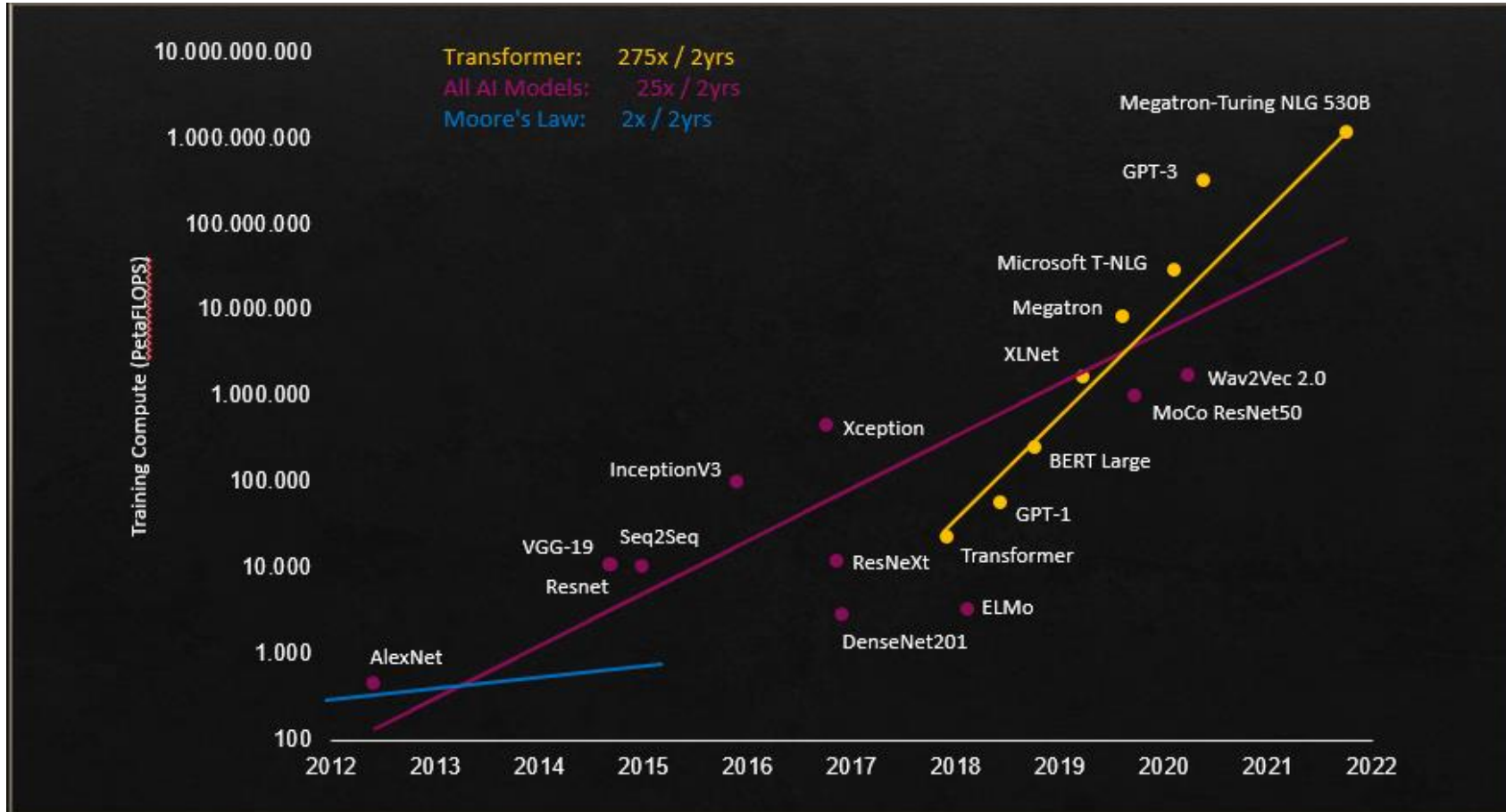


Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper “in-context learning curves” for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

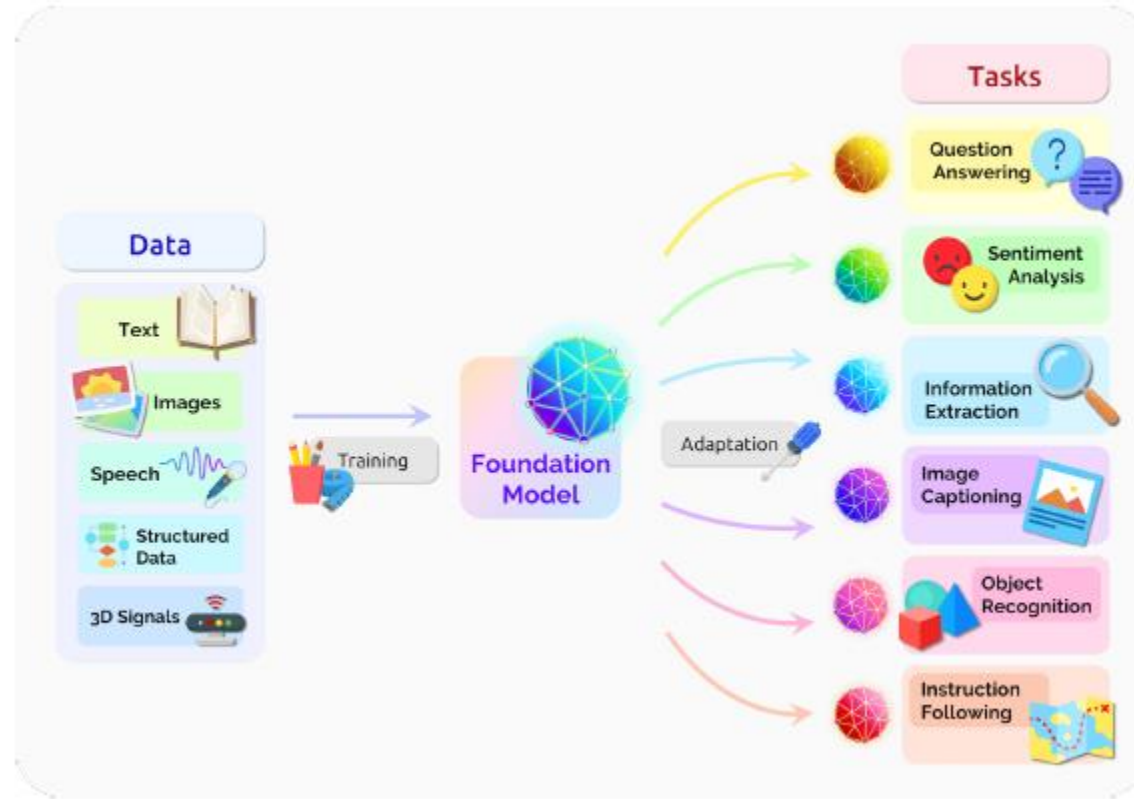
THE ERA OF HUGE MODELS

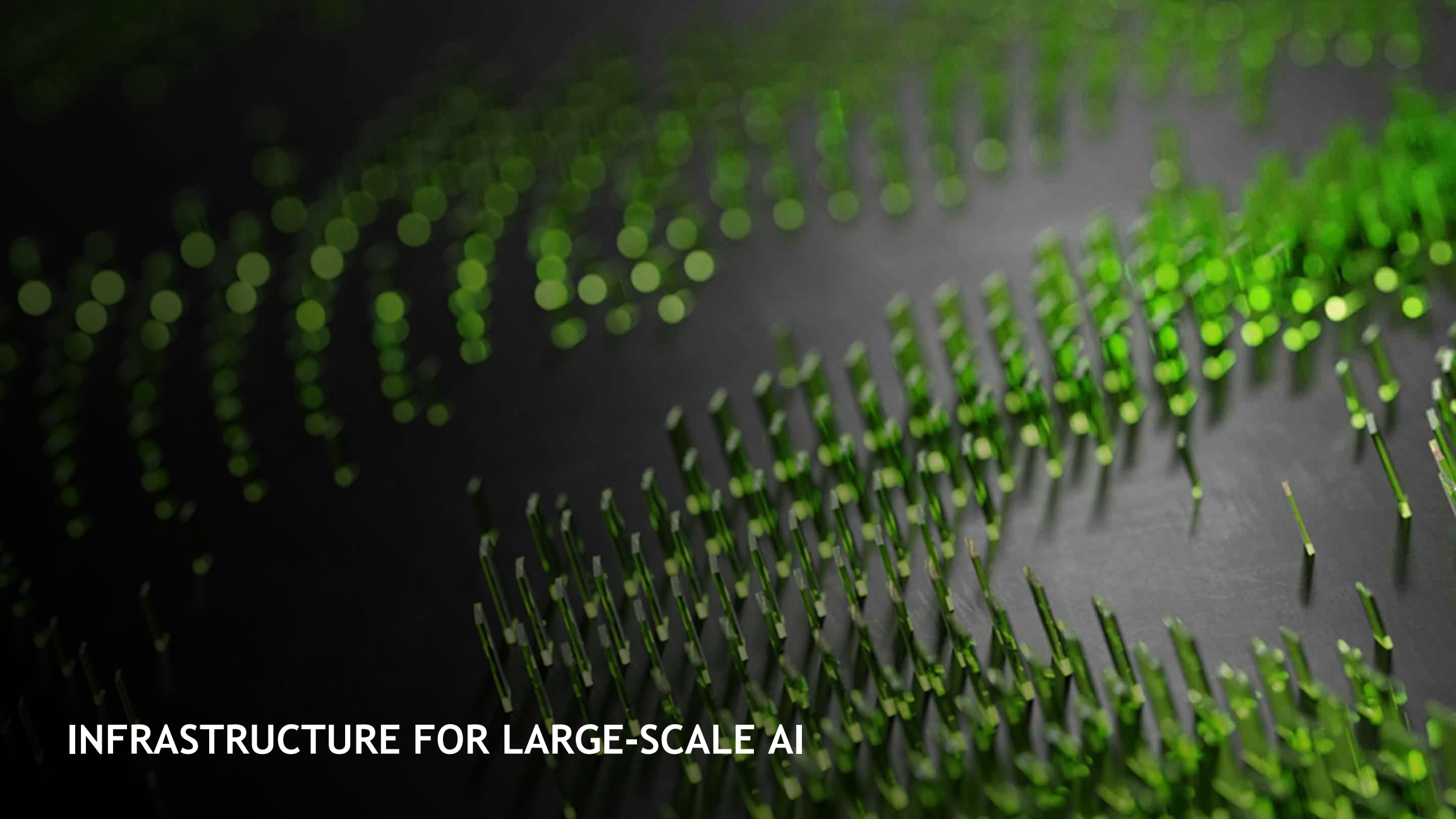
Transformers Computing Load for Training is growing even faster



«FOUNDATION» MODELS - NOT JUST NLP

Huge Models such as Transformers, pre-trained then applied to multiple task

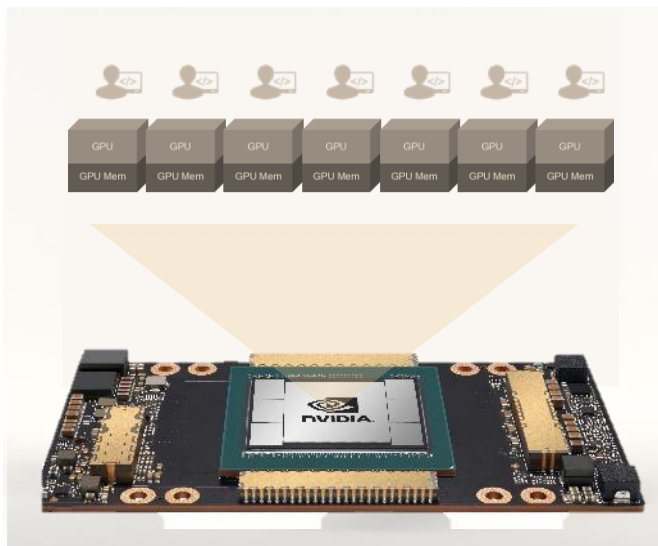




INFRASTRUCTURE FOR LARGE-SCALE AI

SCALABILITY AND HIGH PERFORMANCE AT ALL LEVELS

GPU, System, Cluster



A100 With MIG

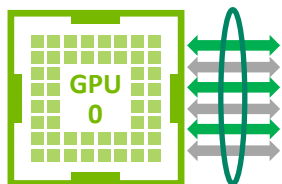


DGX A100



DGX SUPERPOD

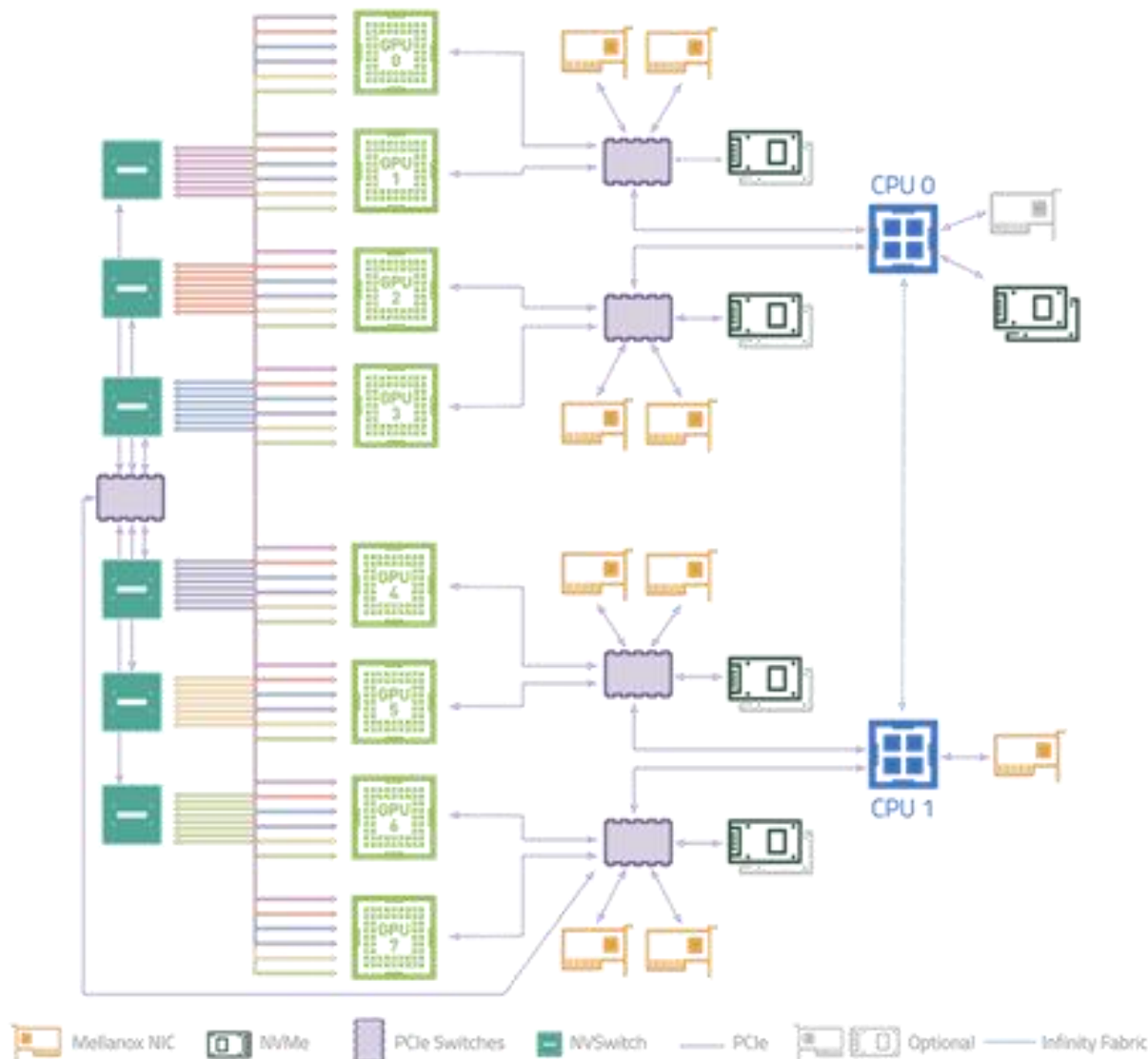
NVIDIA DGX A100 SYSTEM ARCHITECTURE



Ampere A100 NVLINK v3:
12x links = 600 GB/s aggregate bidir

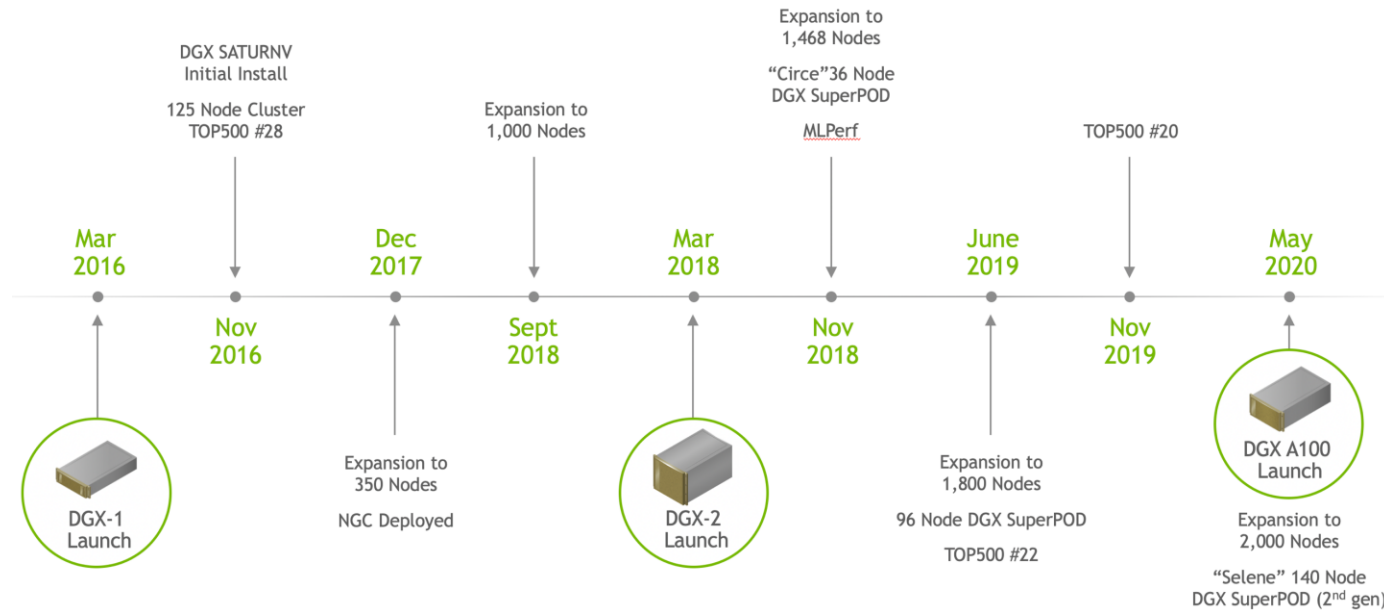


NVSwitch v2: Fully connected NVLINK v3
36 port xbar | 14.4 Tbit/s = 1.8 TB/s
Global bisection bandwidth 4.8 TB/s



LESSONS FROM THE NVIDIA AI JOURNEY

Industry-leading expertise gained from our most important endeavors



- Designing for predictable performance at scale
- Operations/Infrastructure manageability & support
- AI workflow management / data science productivity





DGX SUPERPOD DEPLOYMENTS AT NVIDIA

‘Selene’ and ‘DGX SuperPOD’

#1 on MLPerf for commercially available systems

#6 on TOP500 (63 PetaFLOPS HPL)

#5 on Green500 (26.2 GigaFLOPS/watt)

Fastest Industrial System in U.S.

Both are built with the NVIDIA DGX SuperPOD arch:

- ▶ NVIDIA DGX A100 and NVIDIA Mellanox IB
- ▶ NVIDIA’s decade of AI experience

Selene Configuration:

- ▶ 4,480 NVIDIA A100 Tensor Core GPUs
- ▶ 560 NVIDIA DGX A100 640GB systems
- ▶ 850 Mellanox 200G HDR IB switches
- ▶ 14 PB of all-flash storage
- ▶ 2.8 ExaFLOPS of AI performance
- ▶ Built in 3 weeks

DGX SUPERPOD

Modular Architecture

1K GPU SuperPOD Cluster

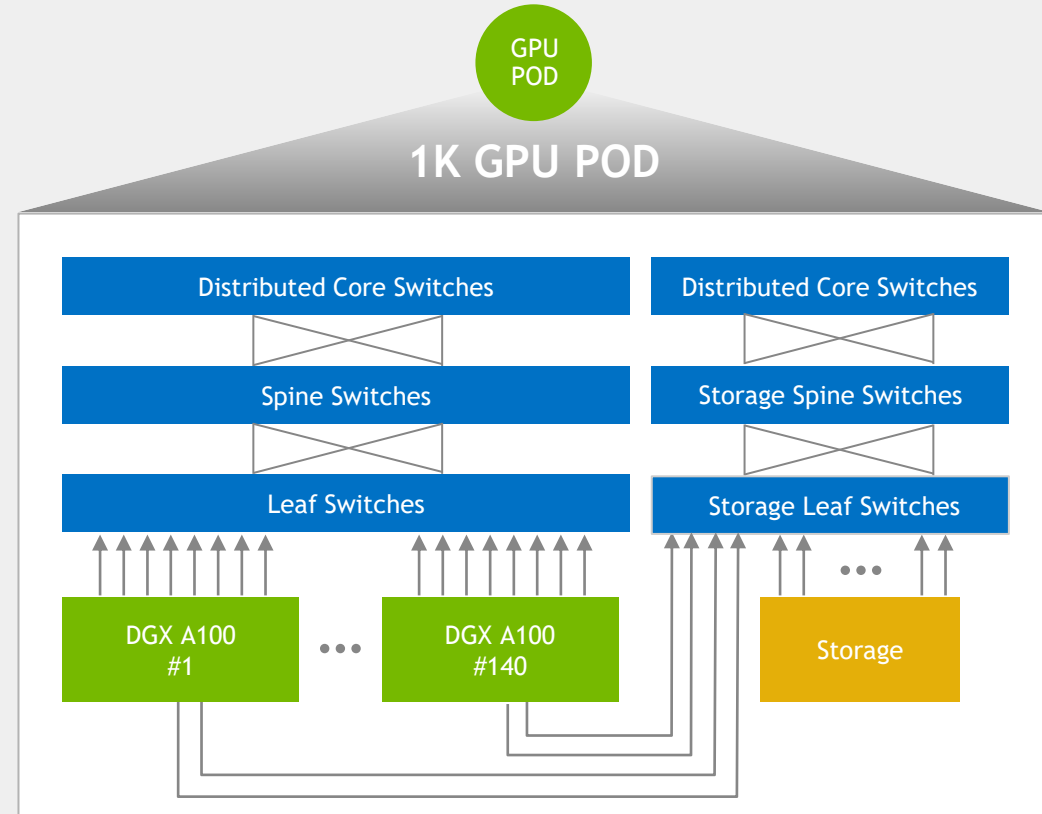
- 140 DGX A100 nodes (1,120 GPUs) in a GPU POD
- 1st tier fast storage - DDN AI400x with Lustre
- Mellanox HDR 200Gb/s InfiniBand - Full Fat-tree
- Network optimized for AI and HPC

DGX A100 Nodes

- 2x AMD 7742 EPYC CPUs + 8x A100 GPUs
- NVLINK 3.0 Fully Connected Switch
- 8 Compute + 2 Storage HDR IB Ports

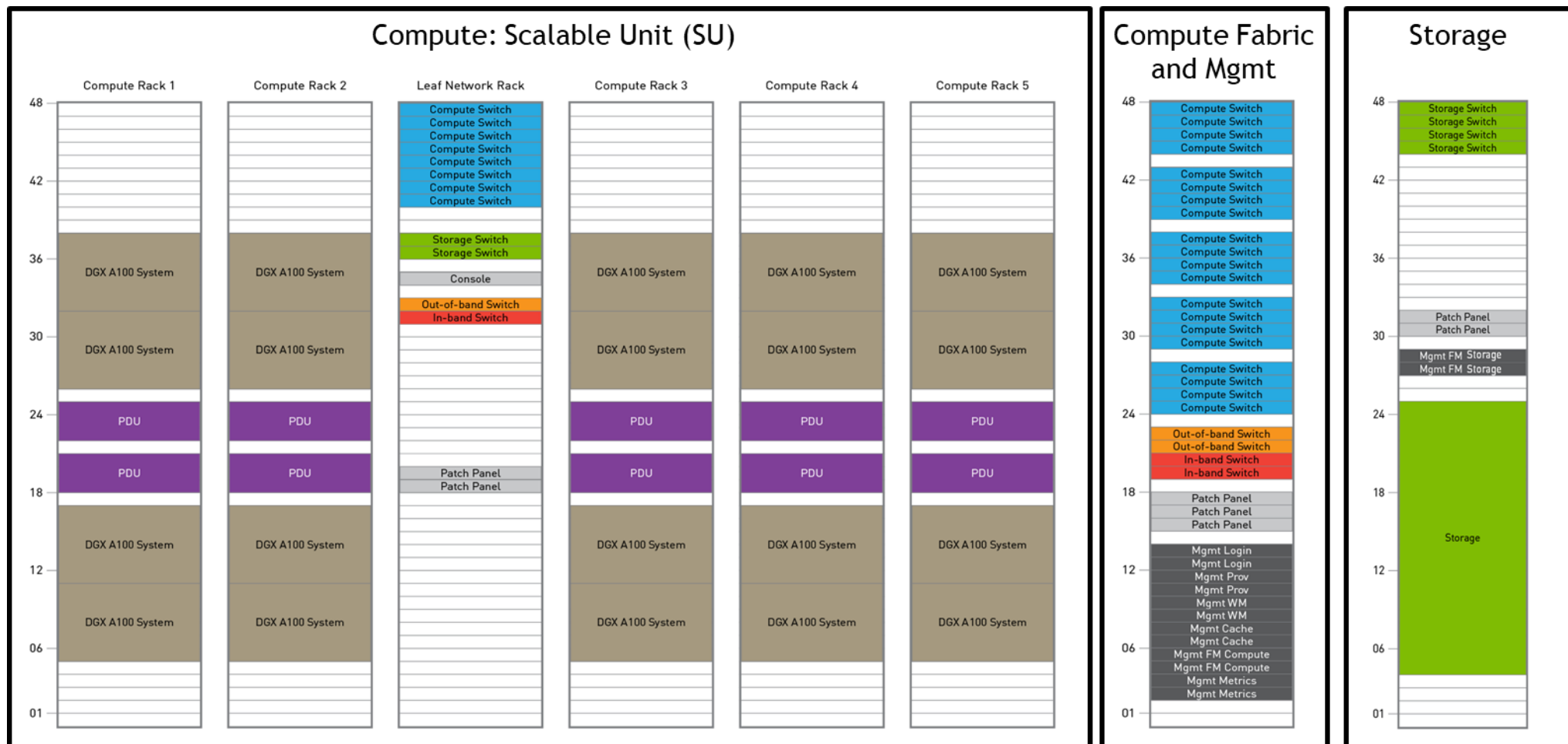
A Fast Interconnect

- Modular IB Fat-tree
- Separate network for Compute vs Storage
- Adaptive routing and SHARPV2 support for offload



RACK DIAGRAM EXAMPLE

DGX A100 SuperPOD 1x SU (Scalable Unit) = 20 Nodes

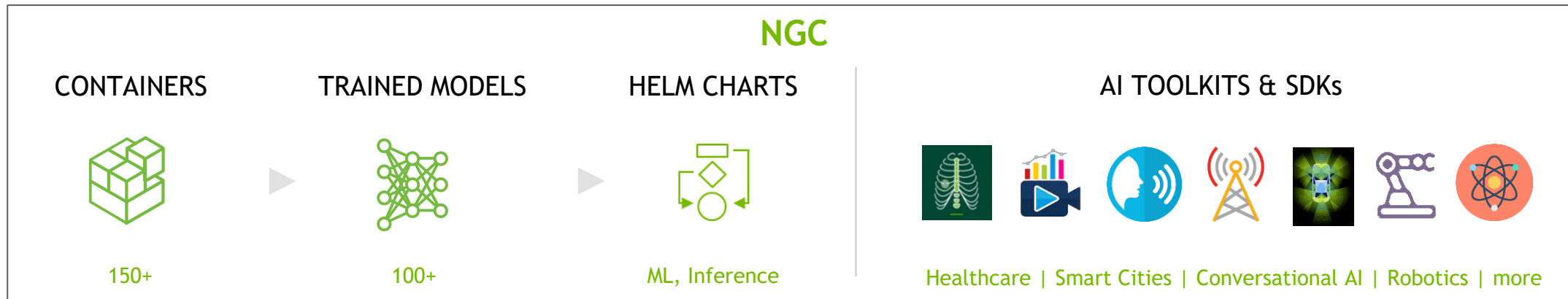




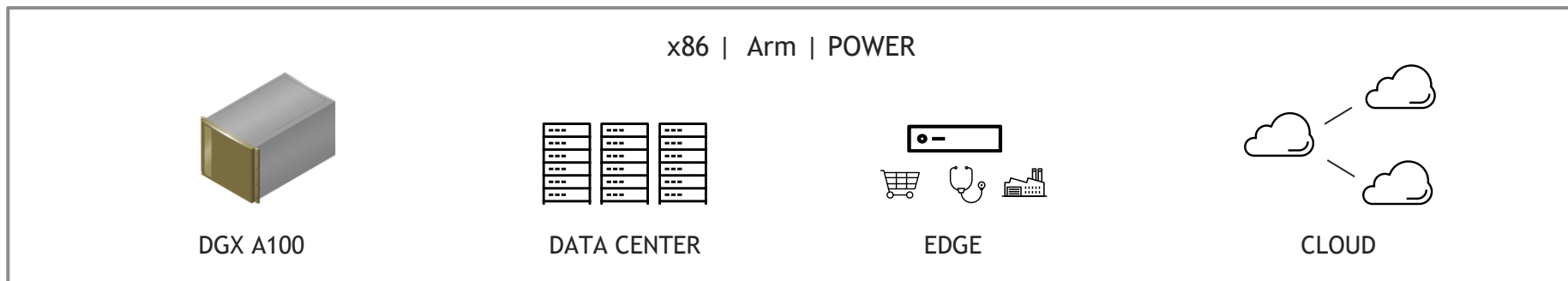
NVIDIA TOOLS FOR DEEP LEARNING

NGC CATALOG: GPU-OPTIMIZED AI SOFTWARE

Easily Deploy Latest Software, Anywhere | ngc.nvidia.com



↓ ENCRYPTED



STATE-OF-THE-ART PERFORMANCE UPDATED MONTHLY

<https://developer.nvidia.com/deep-learning-performance-training-inference>

Converged Training Performance

A100 Training Performance

Time to Convergence

Accuracy

Throughput

Multiple Frameworks

Multiple Networks

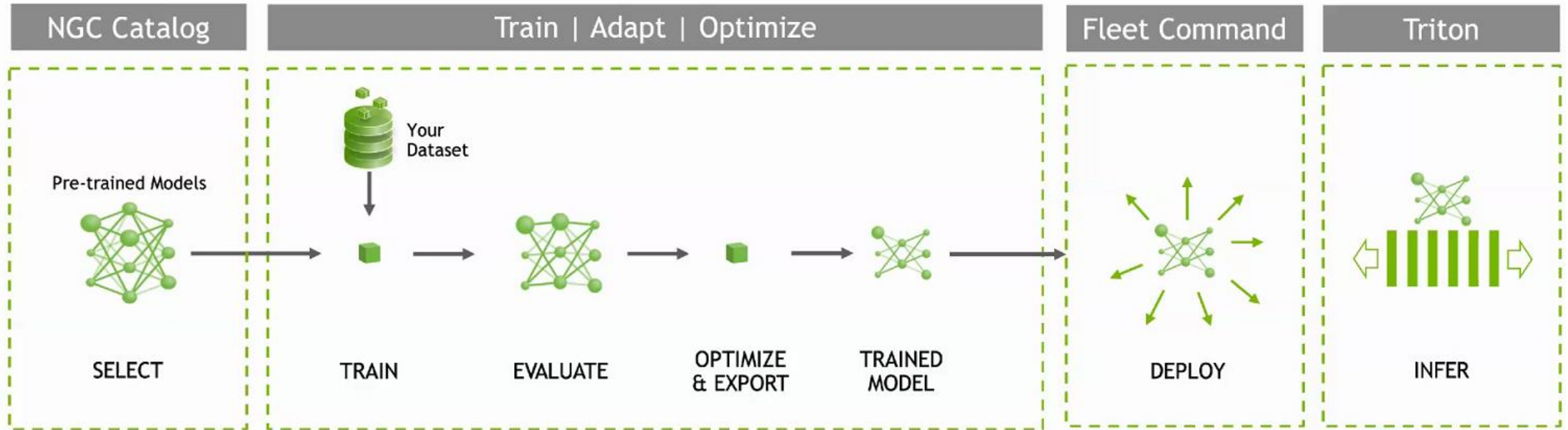
Multiple Network Types
(image, NLP, speech...)

AVAILABLE ON NGC
ALL CONVERGE!
MLPERF RESULTS!

Framework	Network	Time to Train (mins)	Accuracy	Throughput	GPU	Server	Container	Precision	Batch Size	Dataset	GPU Version
MXNet	ResNet-50 v1.5	122	77.32 Top 1 Accuracy	16,114 images/sec	8x A100	NVIDIA DGX-A100	20.10-py3	Mixed	192	ImageNet2012	A100-SXM4-40GB
PyTorch	Mask R-CNN	191	0.34 AP Segm	159 images/sec	8x A100	NVIDIA DGX-A100	20.10-py3	TF32	8	COCO 2014	A100-SXM4-40GB
	ResNeXt101	425	78.93 Top 1 Accuracy	4,596 images/sec	8x A100	NVIDIA DGX-A100	20.08-py3	Mixed	256	Imagenet2012	A100-SXM4-40GB
	SE-ResNeXt101	504	79.06 Top 1 Accuracy	3,875 images/sec	8x A100	NVIDIA DGX-A100	20.09-py3	Mixed	256	Imagenet2012	A100-SXM4-40GB
	SSD v1.1	43	0.25 mAP	3,048 images/sec	8x A100	NVIDIA DGX-A100	20.10-py3	Mixed	128	COCO 2017	A100-SXM4-40GB
	Tacotron2	123	0.6 Training Loss	250,632 total output mels/sec	8x A100	NVIDIA DGX-A100	20.10-py3	TF32	128	LJSpeech 1.1	A100-SXM4-40GB
	WaveGlow	420	-5.68 Training Loss	1,004,778 output samples/sec	8x A100	NVIDIA DGX-A100	20.10-py3	Mixed	10	LJSpeech 1.1	A100-SXM4-40GB
	Transformer	128	27.71 BLEU Score	531,662 words/sec	8x A100	NVIDIA DGX-A100	20.07-py3	Mixed	10240	wmt14-en-de	A100-SXM4-40GB

ACCELERATING THE CREATION OF ENTERPRISE AI

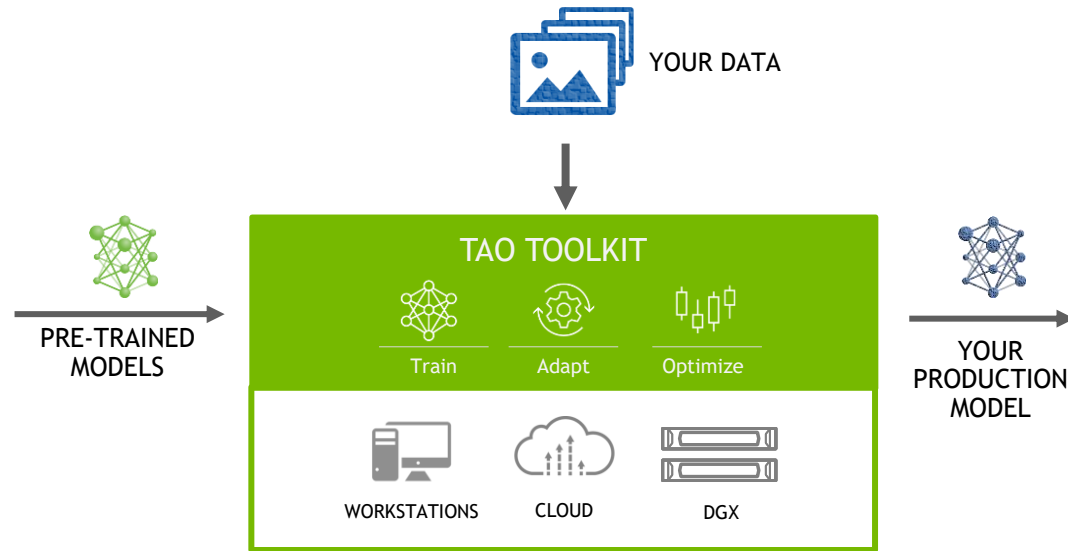
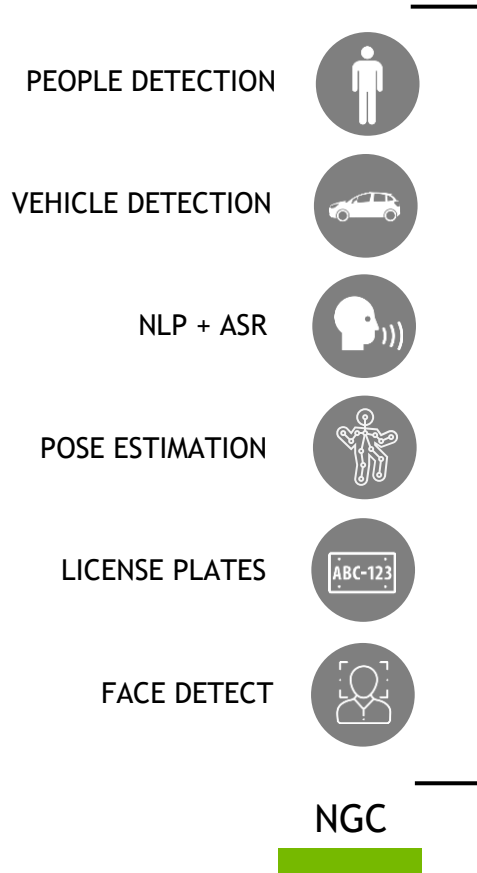
NVIDIA Software Tools for the Complete AI Workflow



NVIDIA TAO TOOLKIT

AI-model-adaptation framework for Transfer Learning

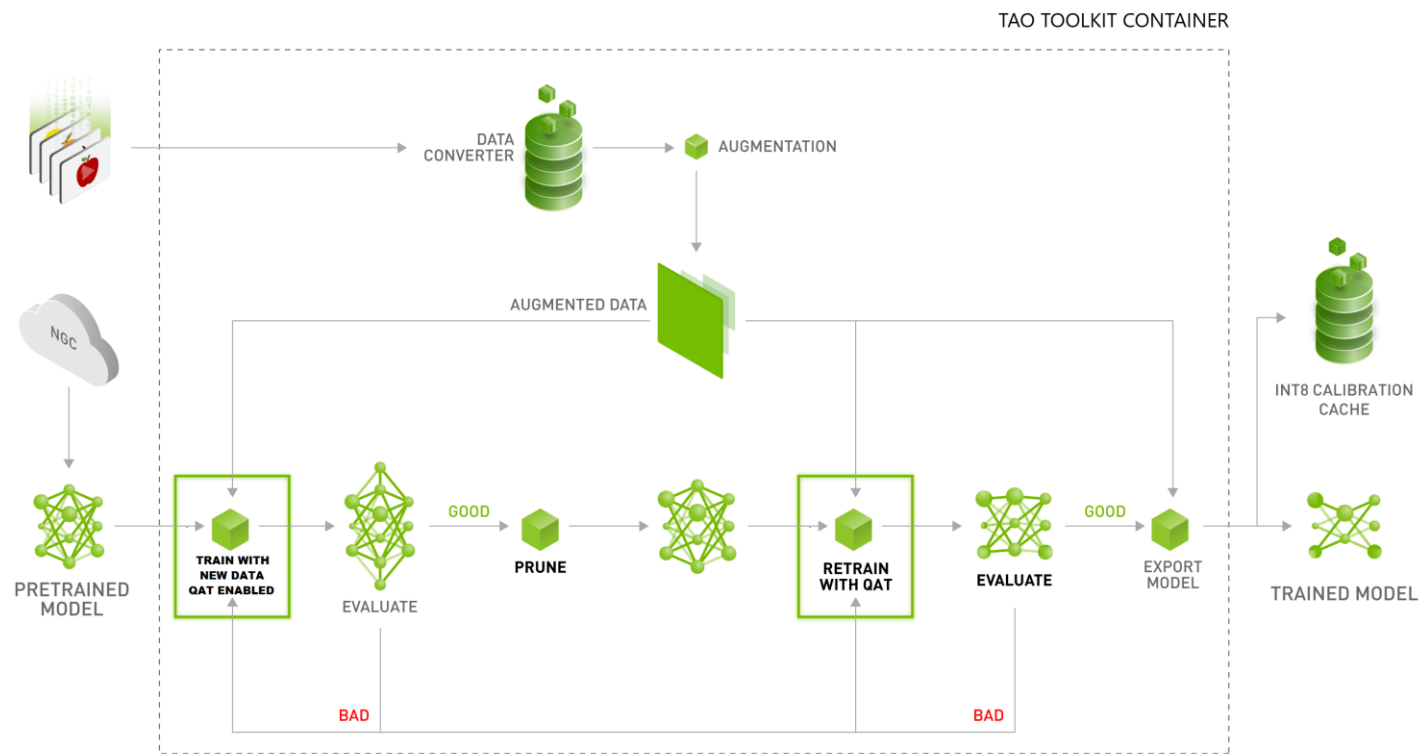
PRE-TRAINED MODEL LIBRARY



developer.nvidia.com/tao

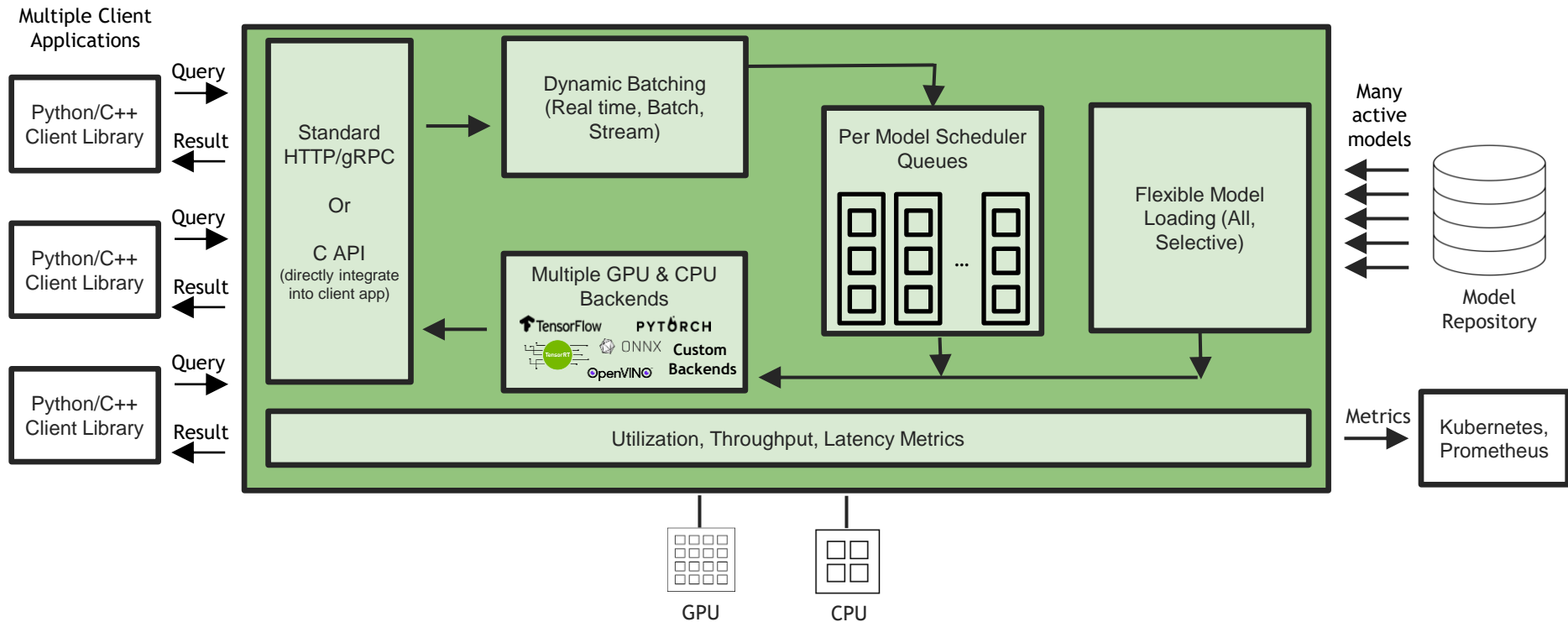
TAO TOOLKIT WORKFLOW

Automatic Mixed Precision | Quantization Aware Training | Pruning



NVIDIA TRITON INFERENCE SERVER

Open-Source Software For Scalable, Simplified Inference Serving

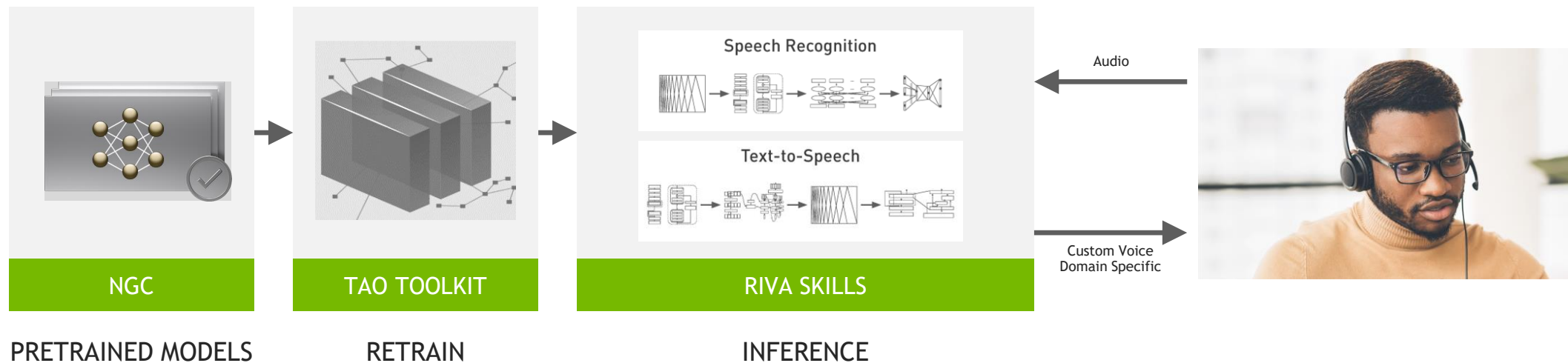


developer.nvidia.com/nvidia-triton-inference-server

NVIDIA RIVA

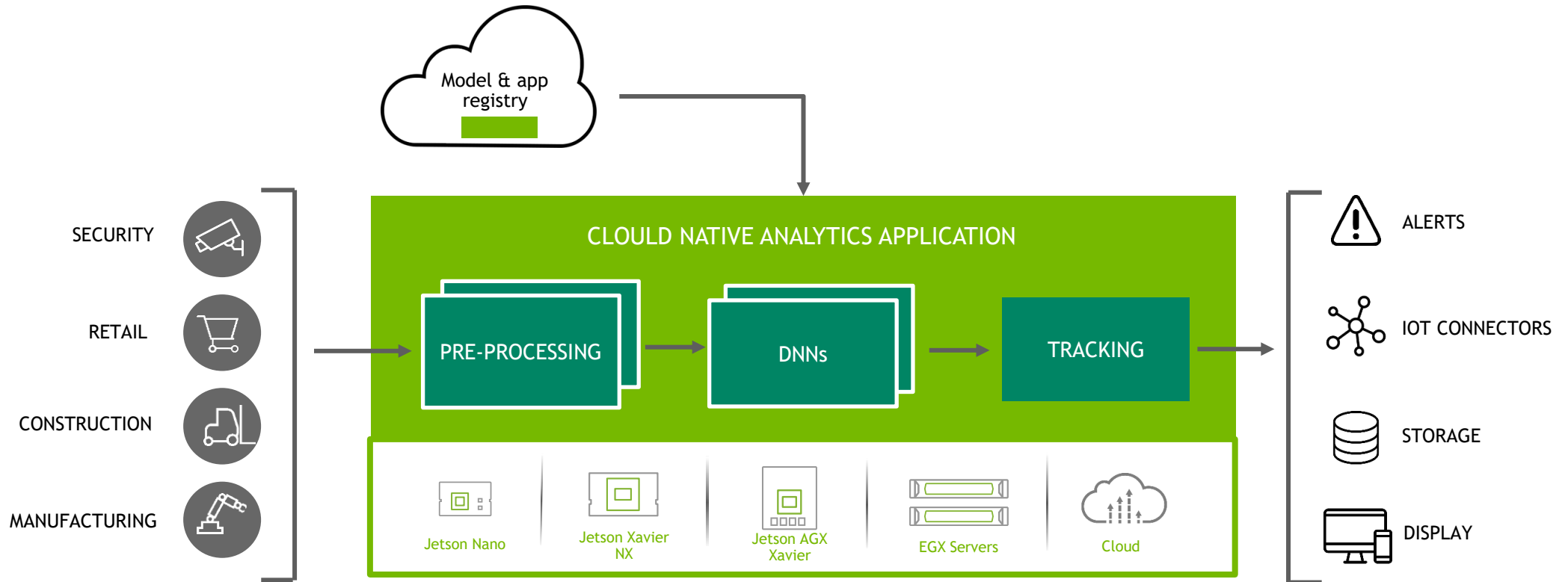
GPU-Accelerated SDK for Speech AI

- World Class Speech Recognition and Text-to-Speech Skills
- Pre-trained SOTA models trained on 100,000 hours of DGX; Retraining with TAO toolkit (zero coding)
- Flexible customization from data to model to pipeline
- Deploy Services with one Line of code in cloud, on-prem & edge
- Scale to handle hundreds and thousands of real-time streams with <300 ms latency per stream

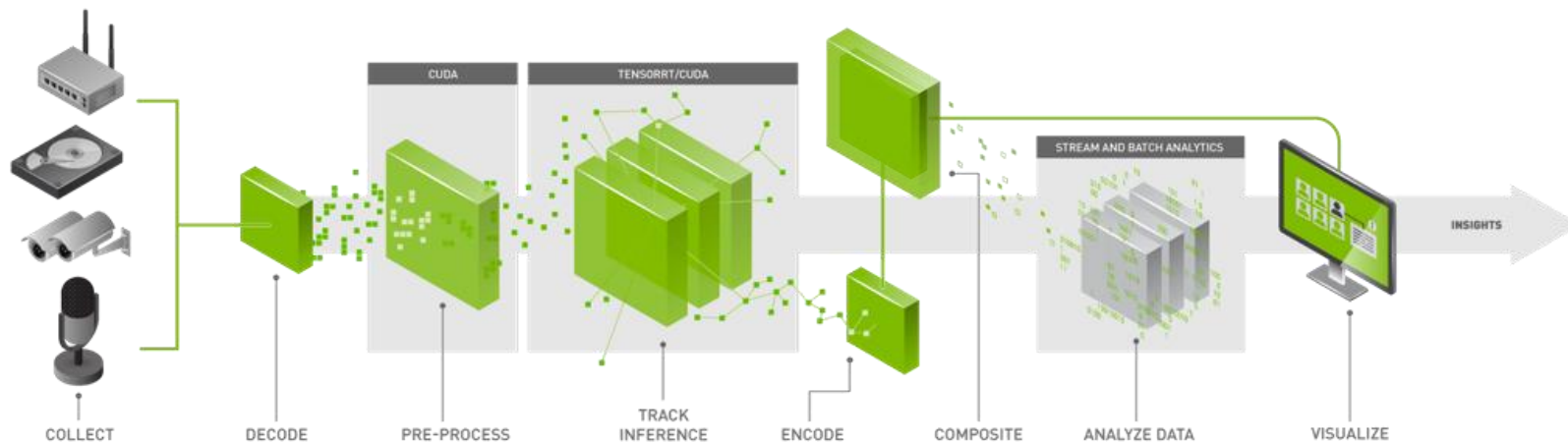


developer.nvidia.com/riva

DEEPSTREAM SDK FOR STREAMING AI APPS

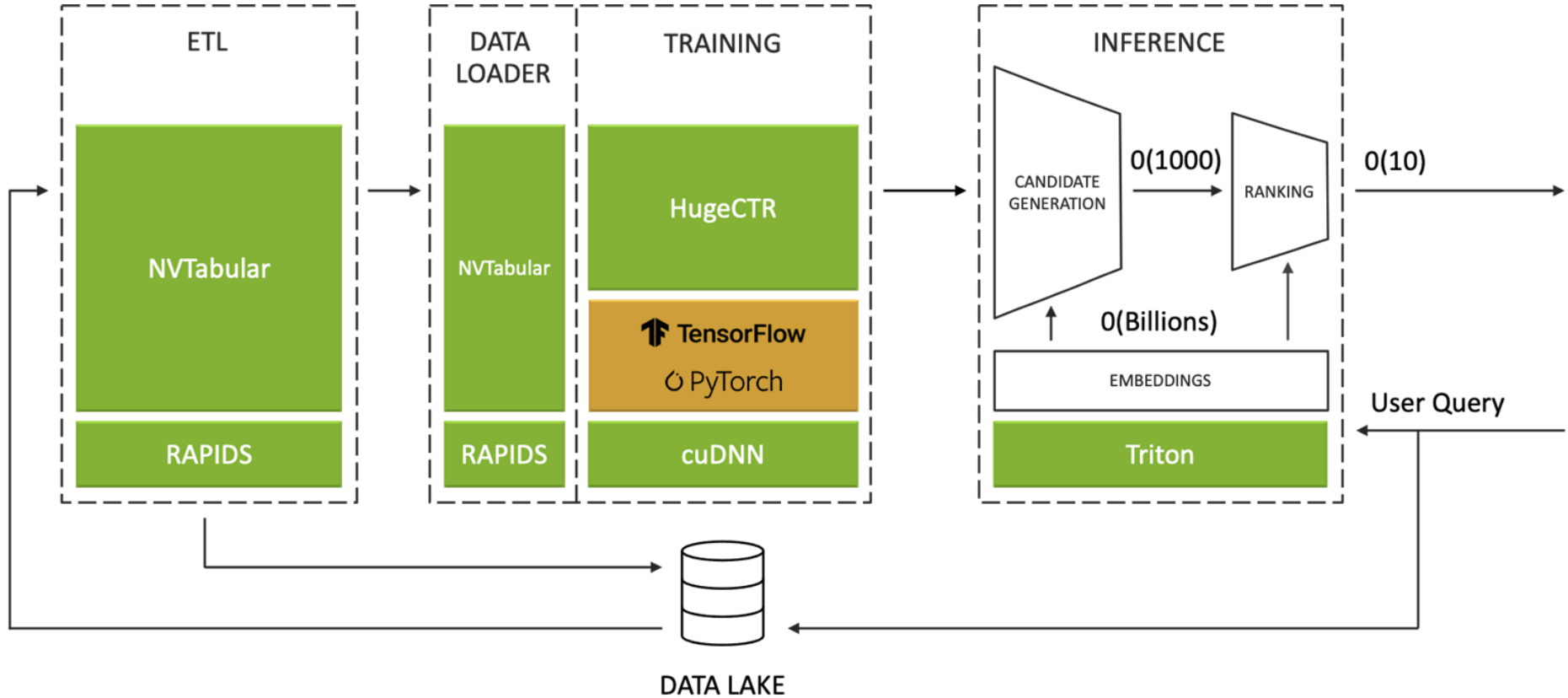


DEEPSTREAM SDK FOR STREAMING AI APPS



NVIDIA MERLIN

End-2-End Library for Accelerated Deep Learning RecSys | Open Source on github

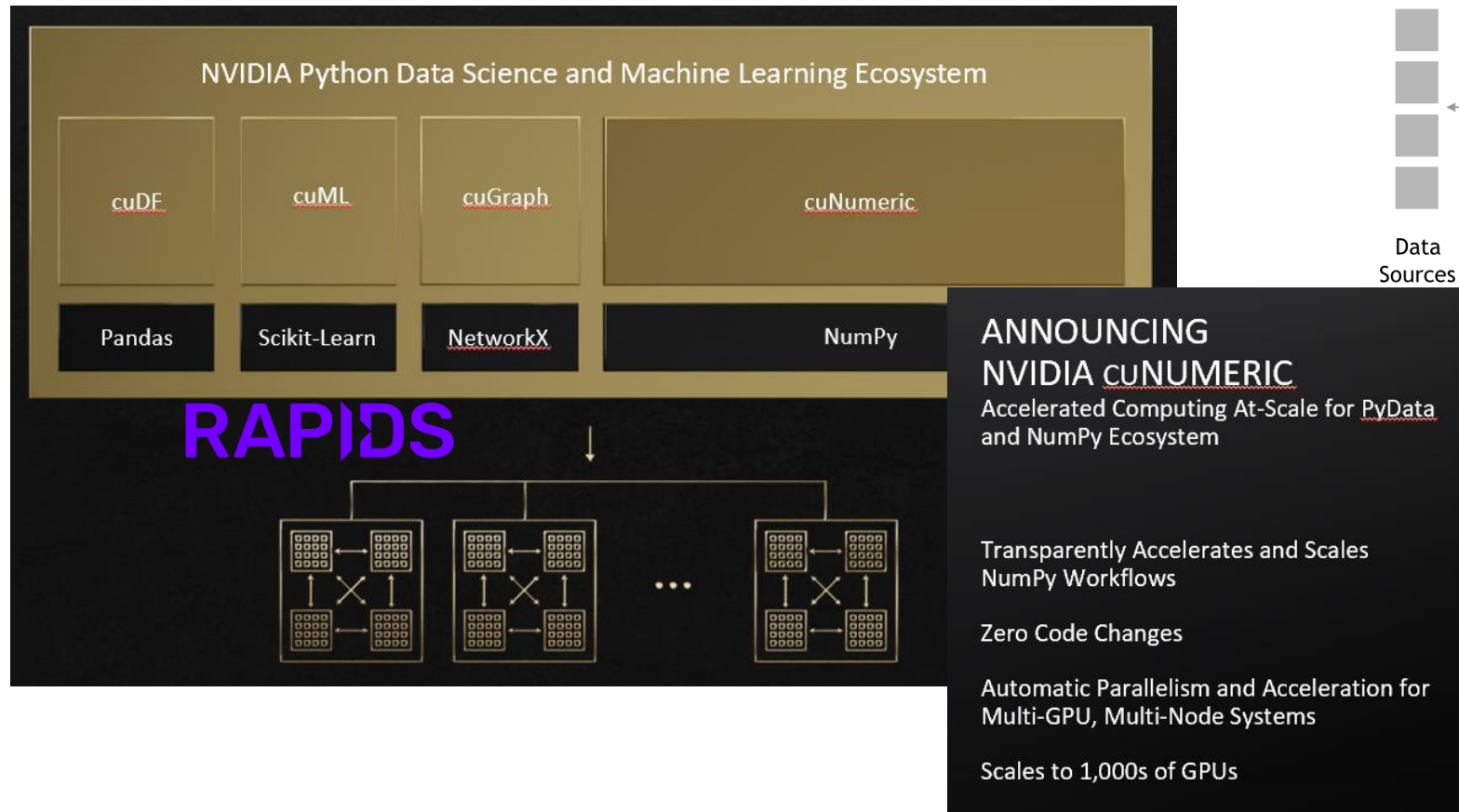


developer.nvidia.com/merlin

... BUT WHAT ABOUT DATA SCIENCE IN GENERAL?

Rich and growing set of libraries and frameworks for the Python & Spark ecosystem

Spark 3.0



 **databricks**  Google Cloud Dataproc

GRAZIE!



cnardone @ nvidia.com
+39 335 5828197