ACCELERATED DATA SCIENCE

UNDERSTAND DEEP LEARNING PIPELINE

AI: THE EXISTENTIAL THREAT AND OPPORTUNITY FOR EVERY ENTERPRISE

Every business needs to transform using AI, not only to survive but to thrive.

Getting there is a challenge for most.

of surveyed execs fear missing their growth objectives if they don't scale AI[1]

cited their struggle with how to scale AI across their business[1]

[1]*Accenture: "AI: Built to Scale, From Experimental to Exponential." 2019.*

ACCELERATED DATA SCIENCE

DATA ANALYTICS

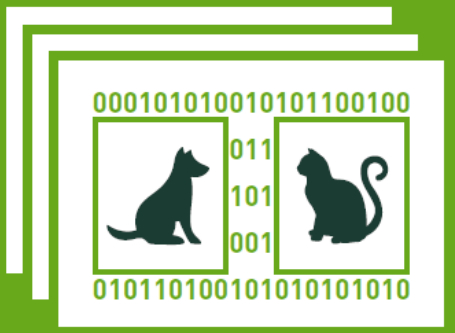
Extracting insights from big data

MACHINE LEARNING

Learning from examples in the data

DEEP LEARNING

Automating feature engineering



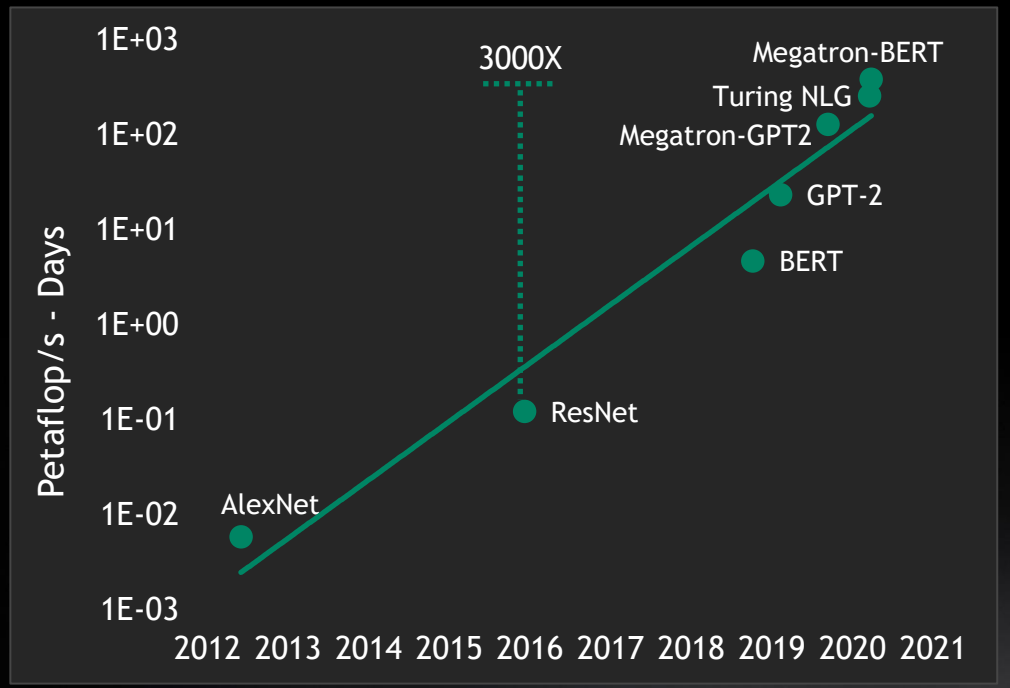
EXTENDING DL → BIG DATA ANALYTICS

From Business Intelligence to Data Science

UNDERSTAND ACCELERATION

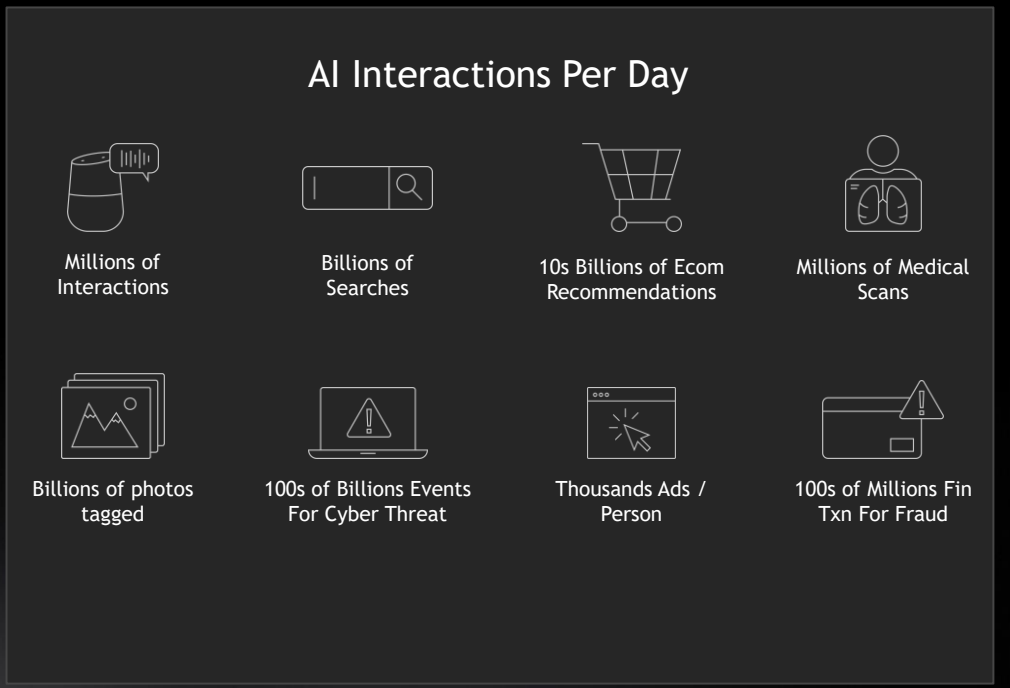
CHALLENGES: ACCELERATING BIG AND SMALL

AI Advances Demand Exponentially Higher Compute



3000X Higher Compute Required to Train Largest Models Since Volta

AI Applications Demand Distributed Pervasive Acceleration



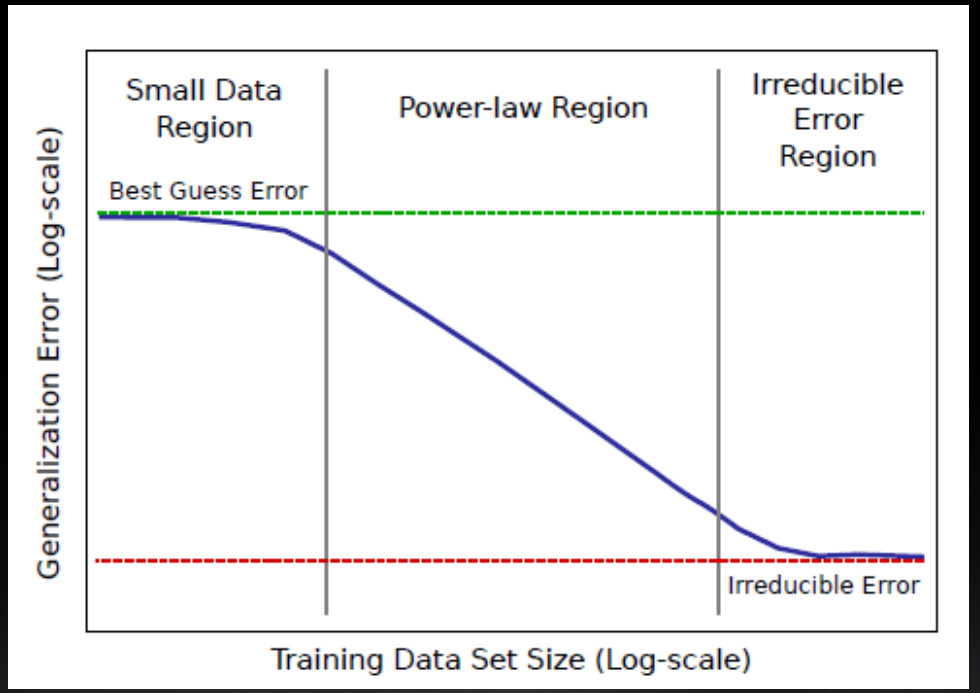
Every AI Powered Interaction Needs Varying Amount of Compute

Source: OpenAI, NVIDIA

EMPIRICAL EVIDENCE
BIGGER IS BETTER

EXPLODING DATASETS

Logarithmic relationship between the dataset size and accuracy



Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., ... & Zhou, Y. (2017). Deep Learning Scaling is Predictable, Empirically. arXiv preprint arXiv:1712.00409.

THE SCALING LAWS

As you increase the dataset size you must increase the model size

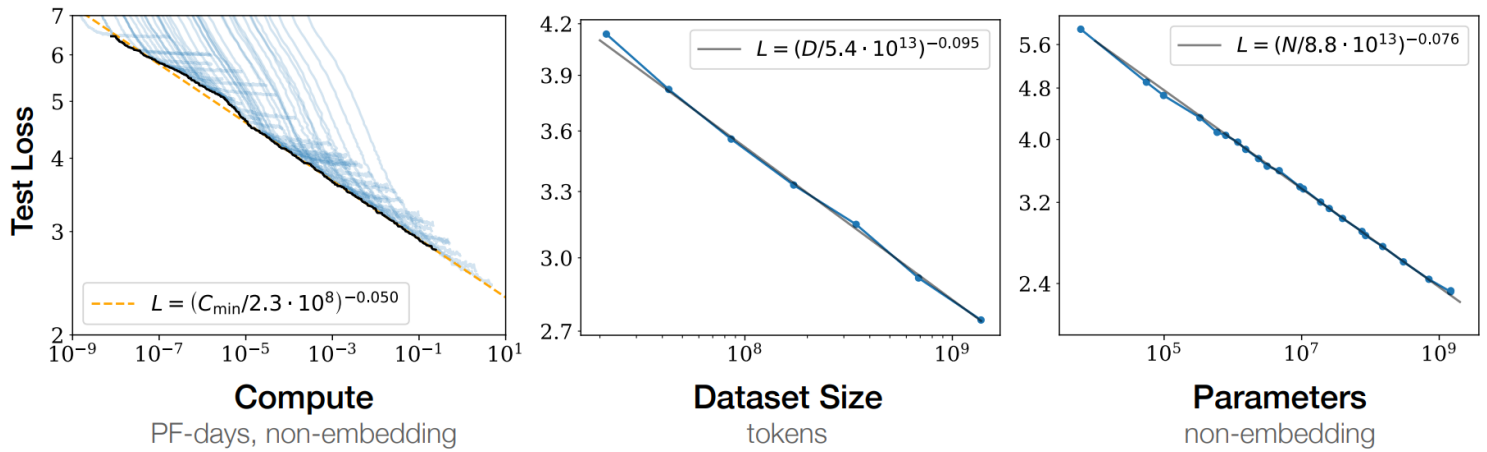
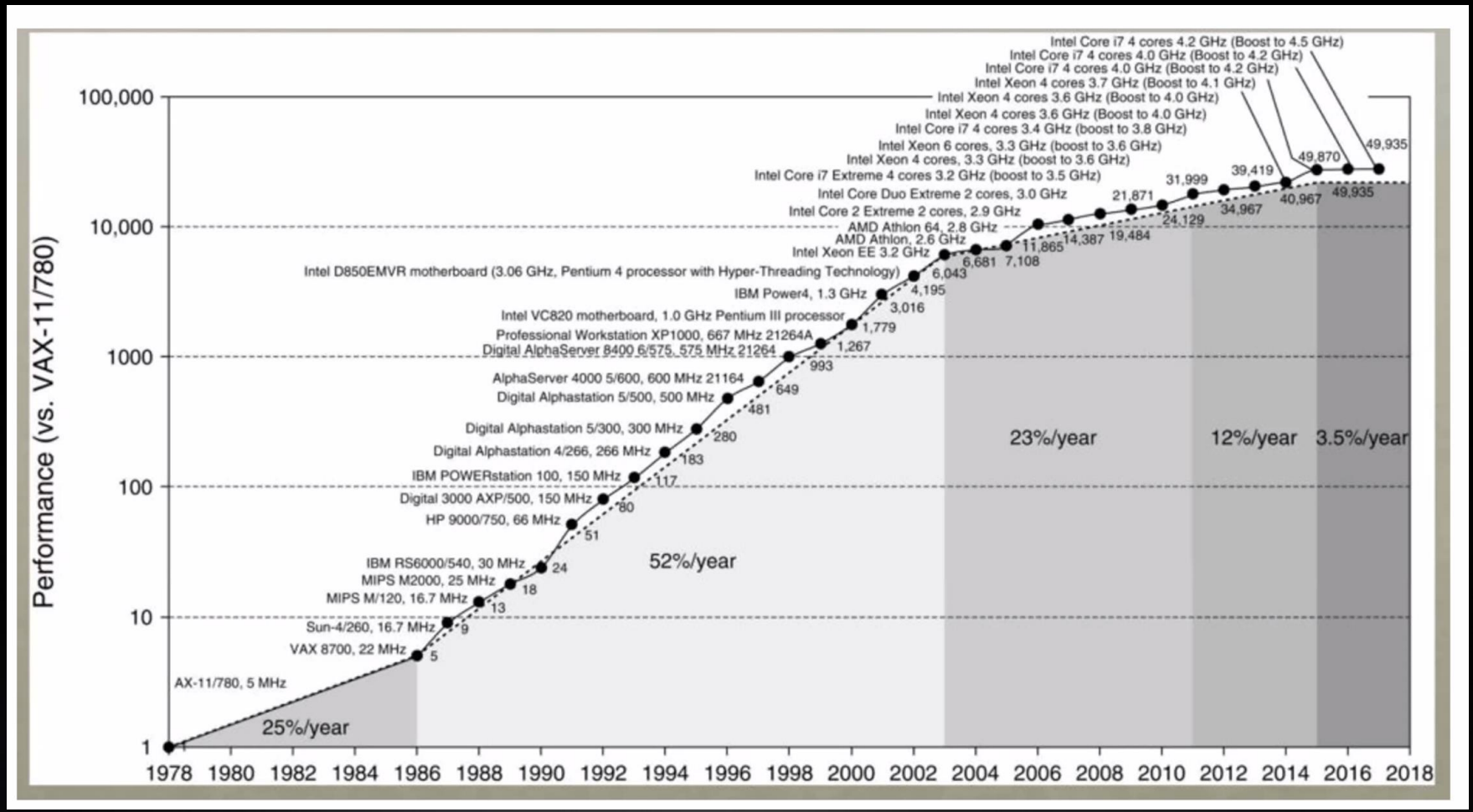


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.

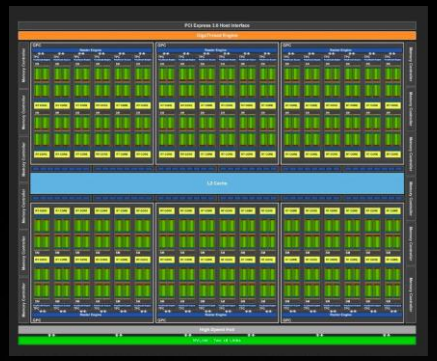
NVIDIA.

MOTIVATION: END OF MOORE'S LAW



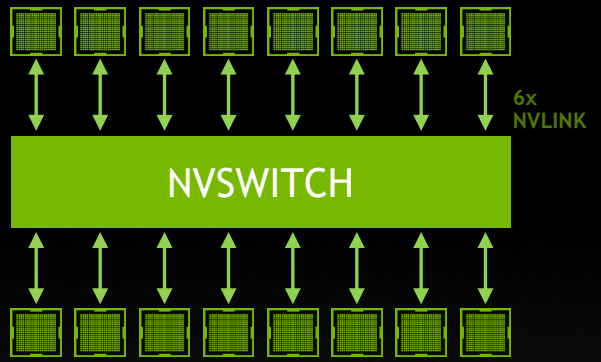
PILLARS OF PERFORMANCE

CUDA Architecture



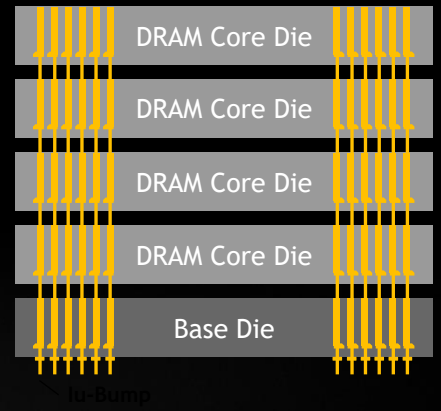
Massively parallel processing

NVLink/NVSwitch



High speed connecting between GPUs for distributed algorithms

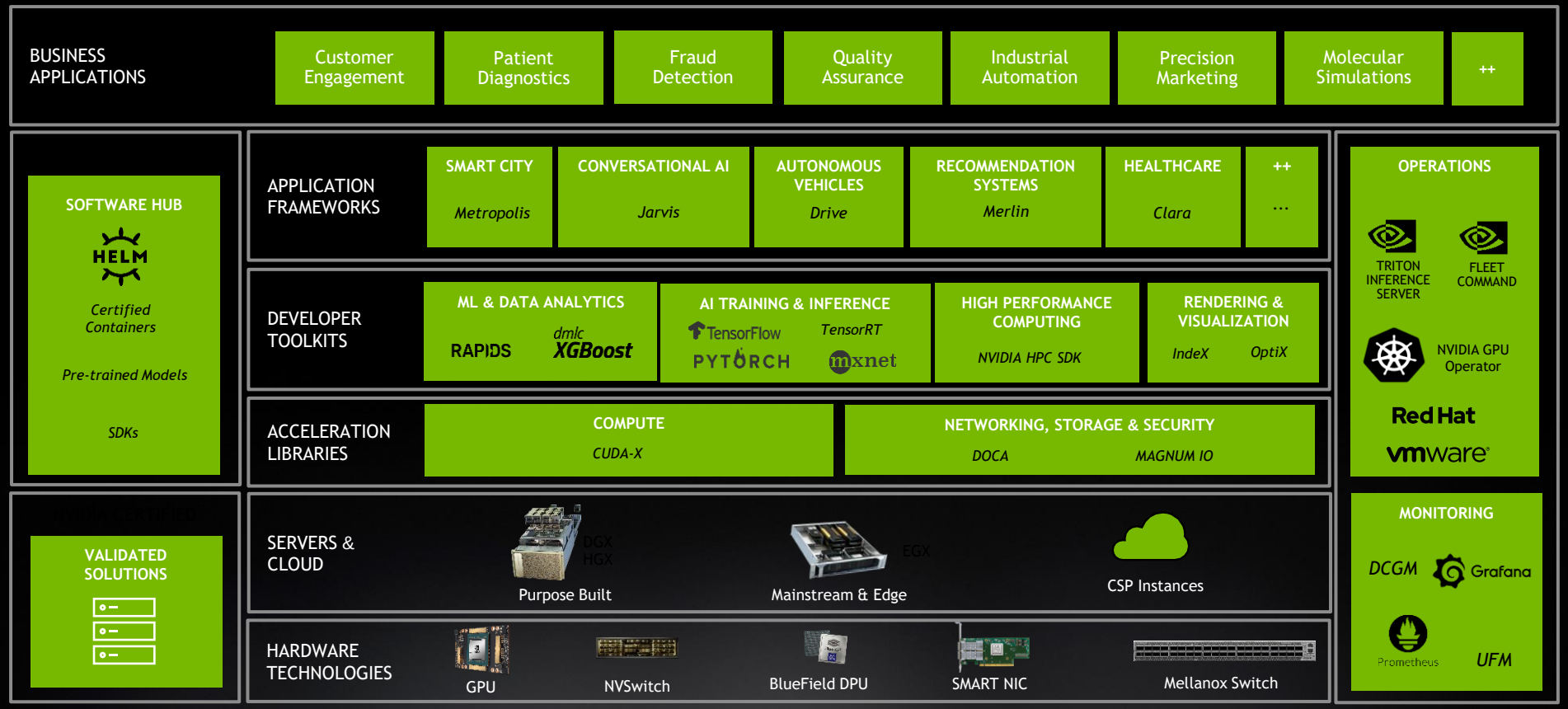
Memory Architecture



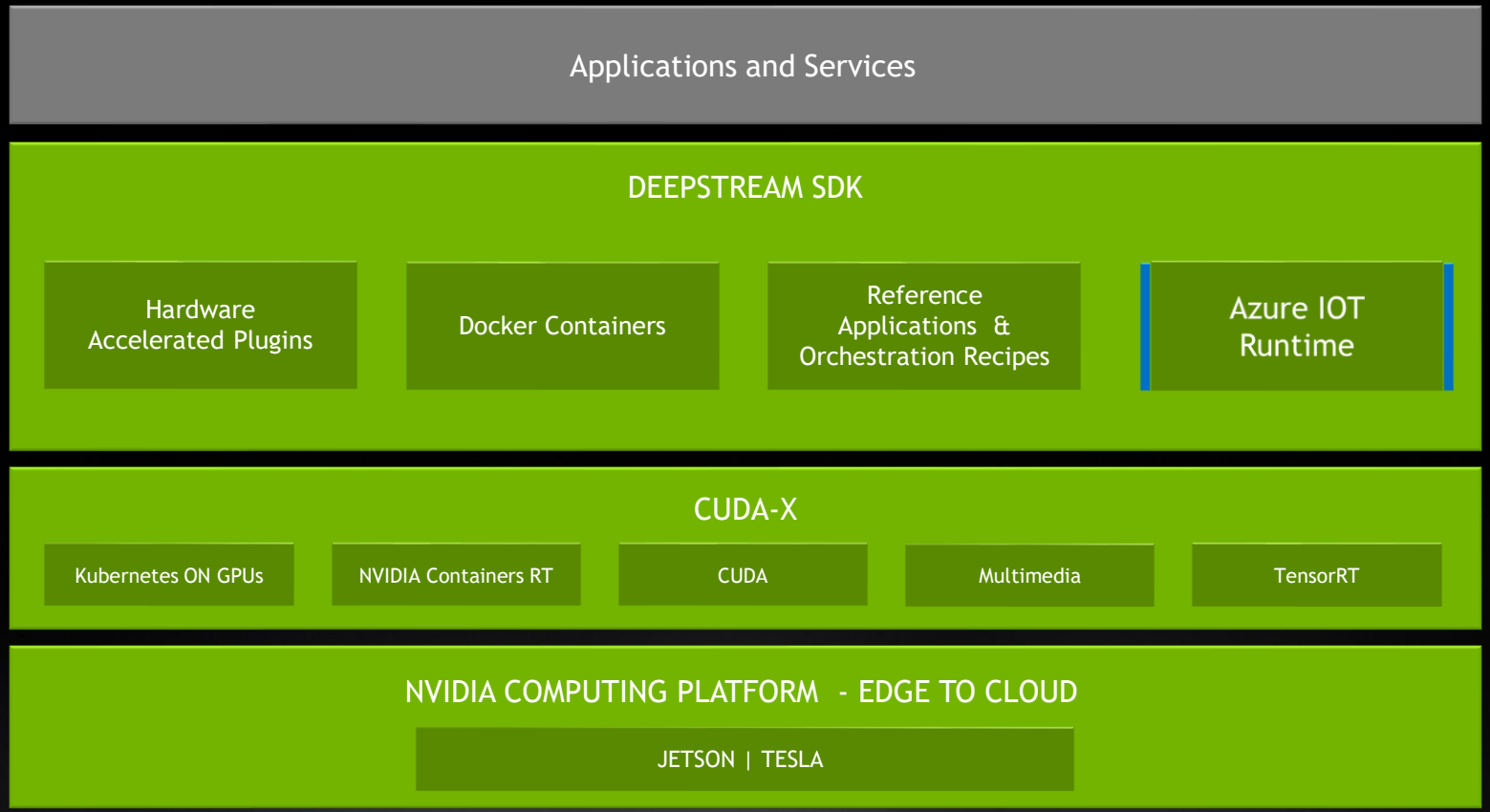
Large virtual GPU memory, high-speed memory

WAYS TO ACCELERATED DATA SCIENCE

NVIDIA DATACENTER PLATFORM

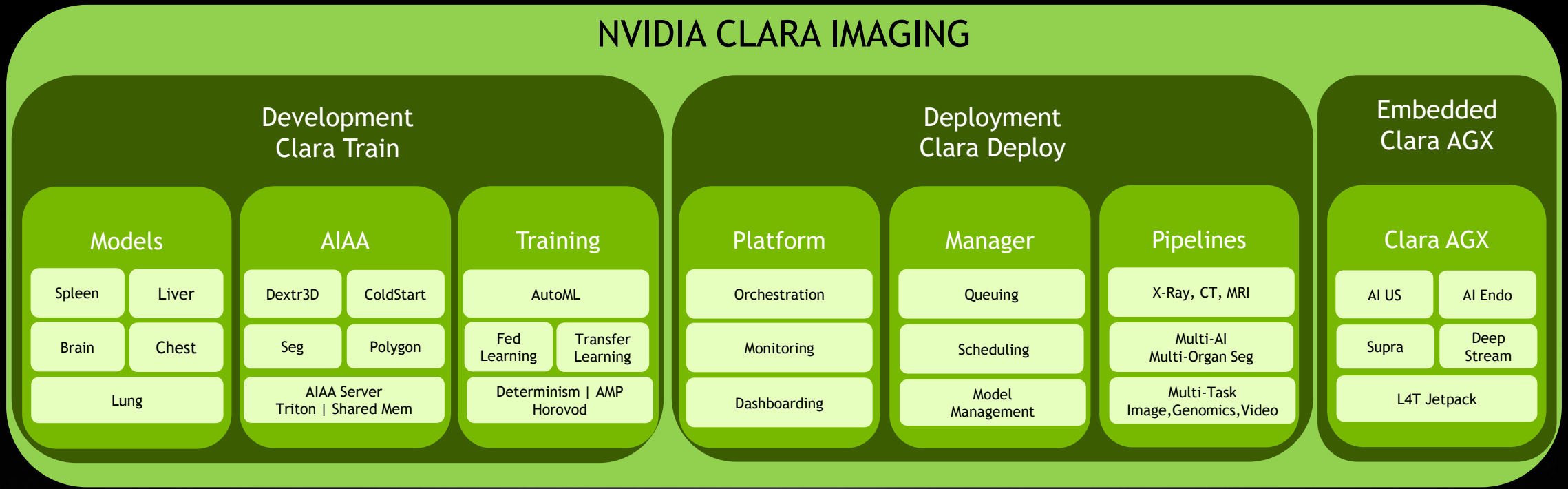


DEEPSTREAM SOFTWARE STACK

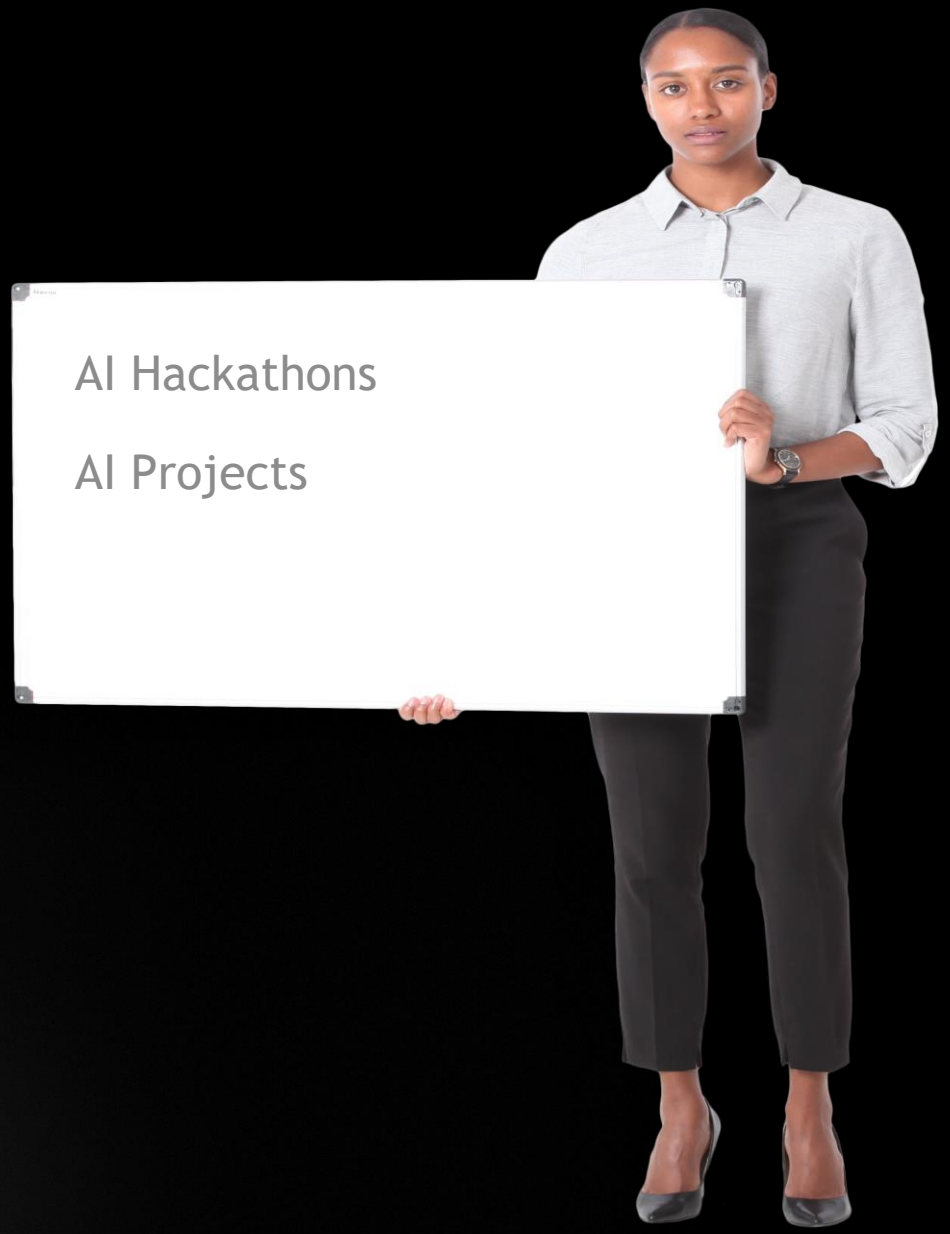


DEEPSTREAM ON JETSON NANO

Application Framework for Healthcare



CALL TO ACTION




THANK YOU
NVIDIA