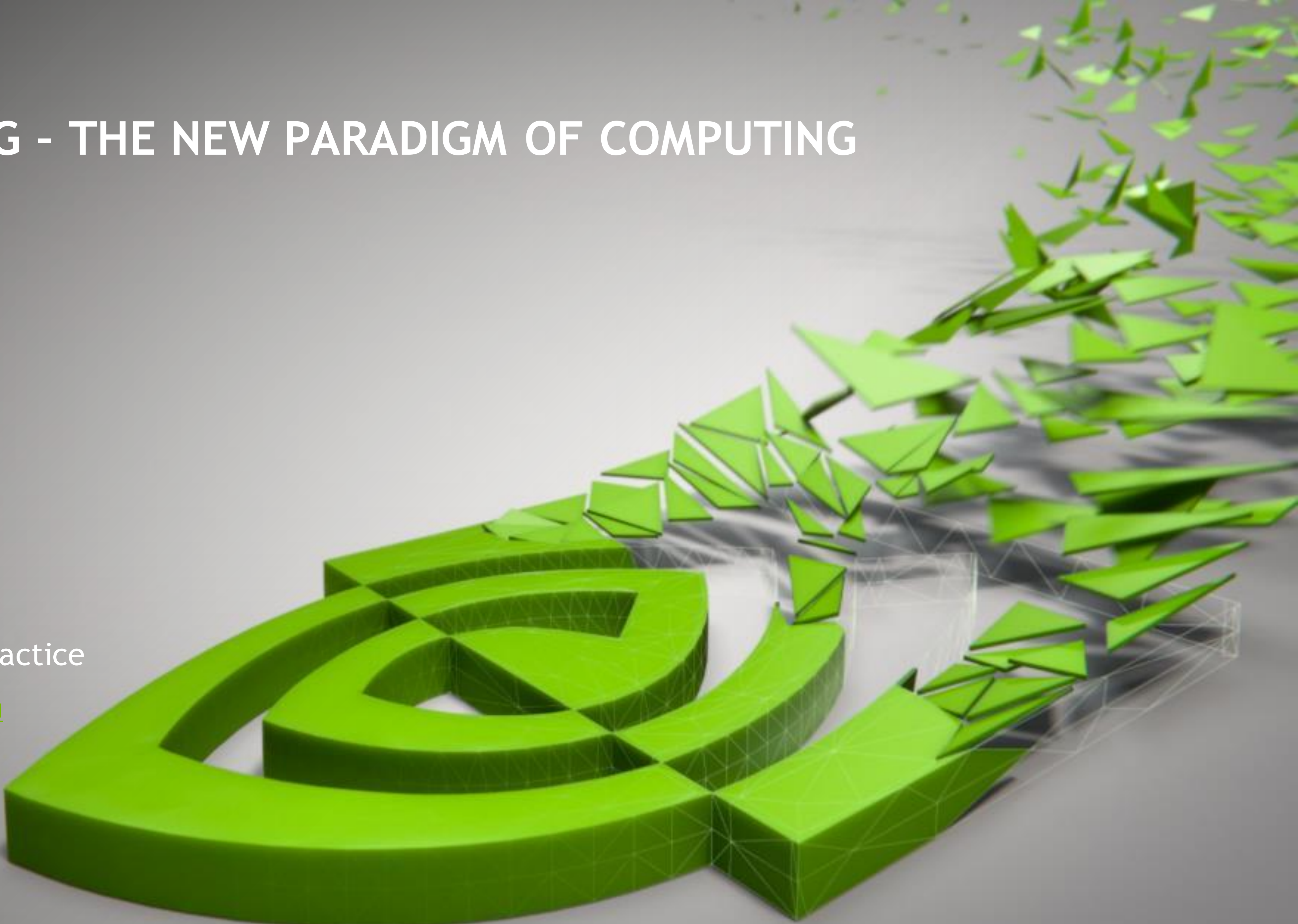
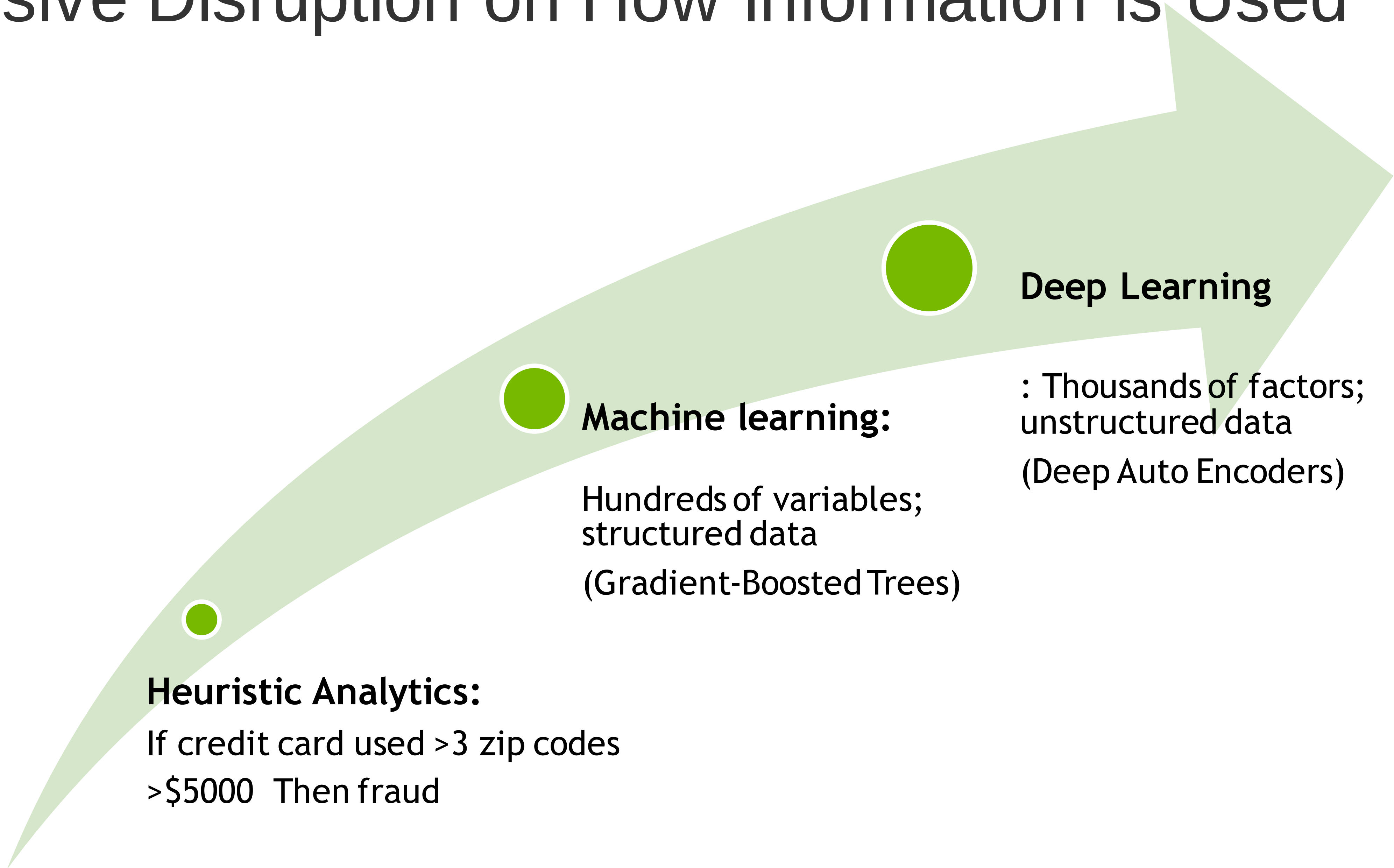
DEEP LEARNING – THE NEW PARADIGM OF COMPUTING



Sundara Ramalingam N
Head – Deep Learning Practice
snagalingam@nvidia.com
+91 99455 67685

Massive Disruption on How Information is Used



Heuristic Analytics:

If credit card used >3 zip codes

>$5000 Then fraud

Machine learning:

Hundreds of variables; structured data

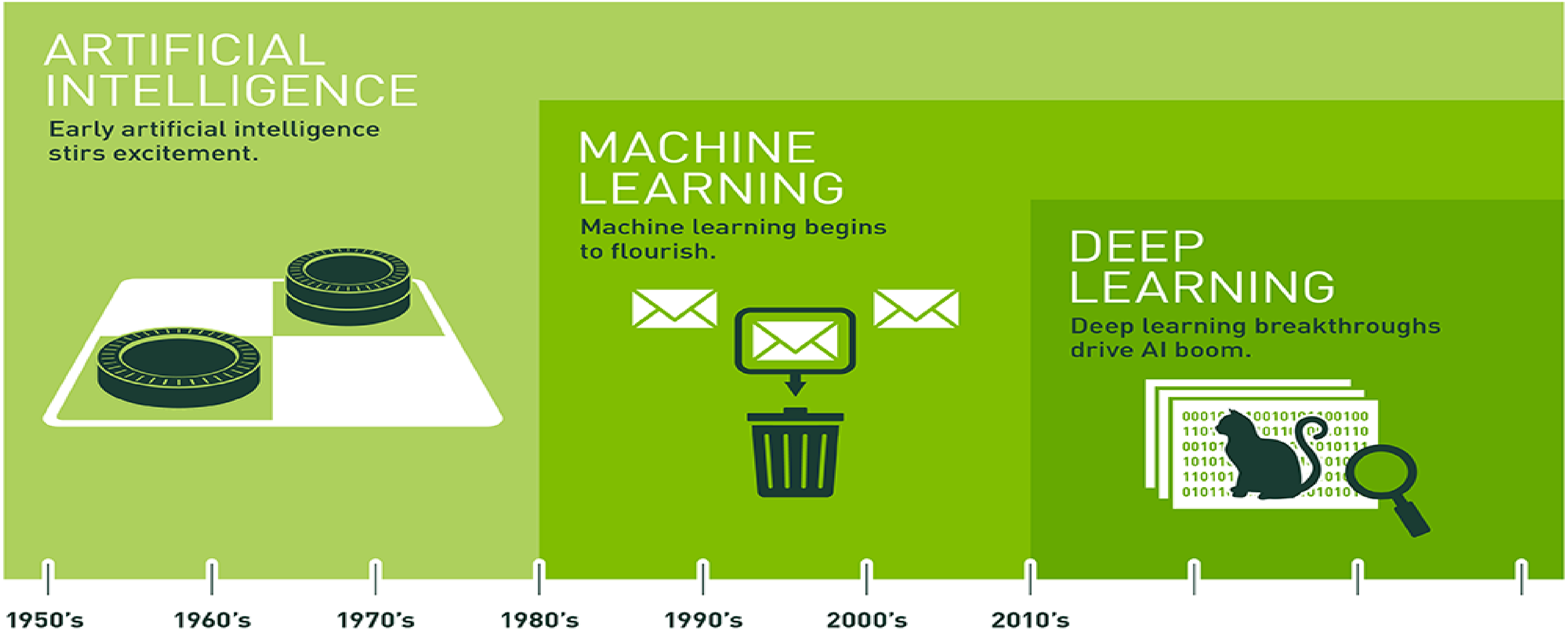
(Gradient-Boosted Trees)

Deep Learning

: Thousands of factors; unstructured data

(Deep Auto Encoders)

CAPABILITY OF MACHINE TO IMITATE INTELLIGENT BEHAVIOR



ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.

MACHINE LEARNING

Machine learning begins to flourish.

DEEP LEARNING

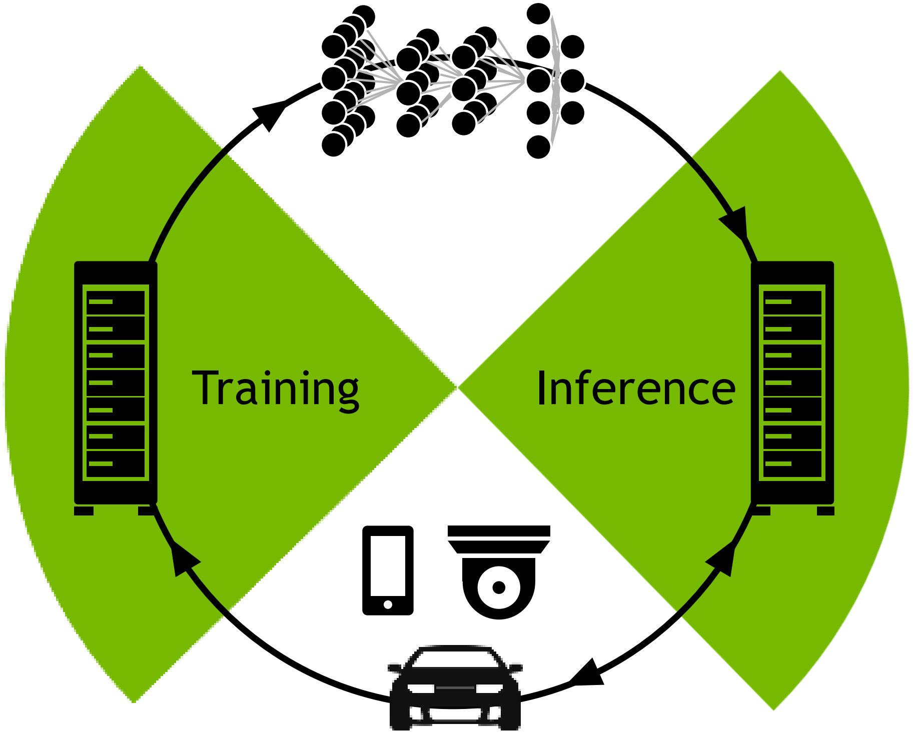
Deep learning breakthroughs drive AI boom.

GPU DEEP LEARNING IS A NEW COMPUTING MODEL

Billions of Trillions of Operations

GPU train larger models, accelerate time to market

Training



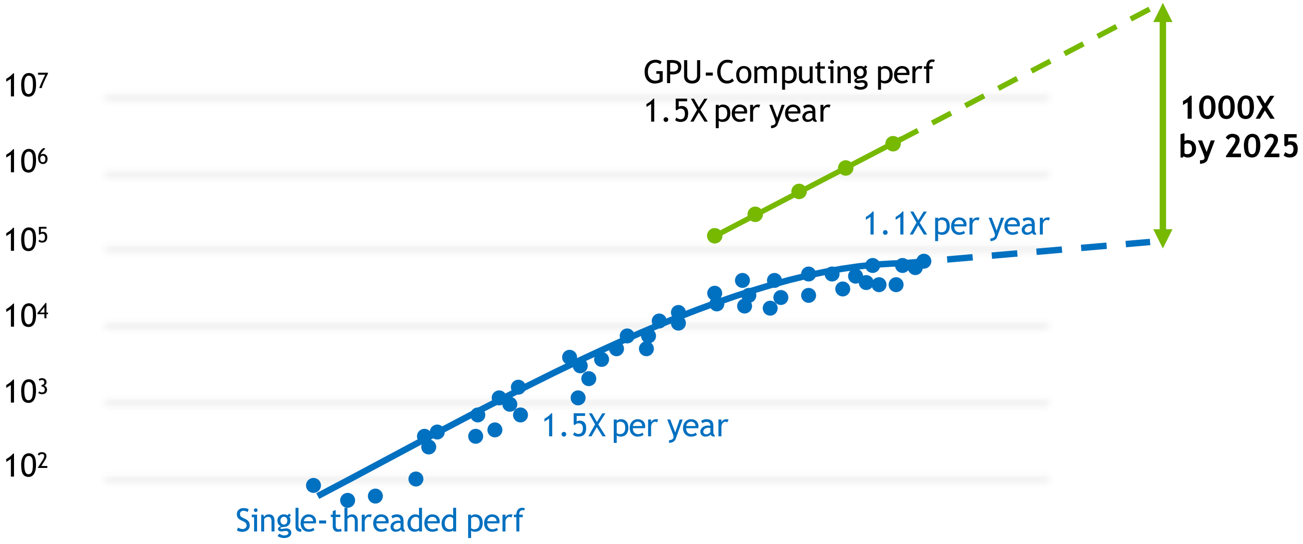
Device

10s of billions of image, voice, video queries per day

GPU inference for fast response, maximize data center throughput

Data center **inference**

RISE OF NVIDIA GPU COMPUTING

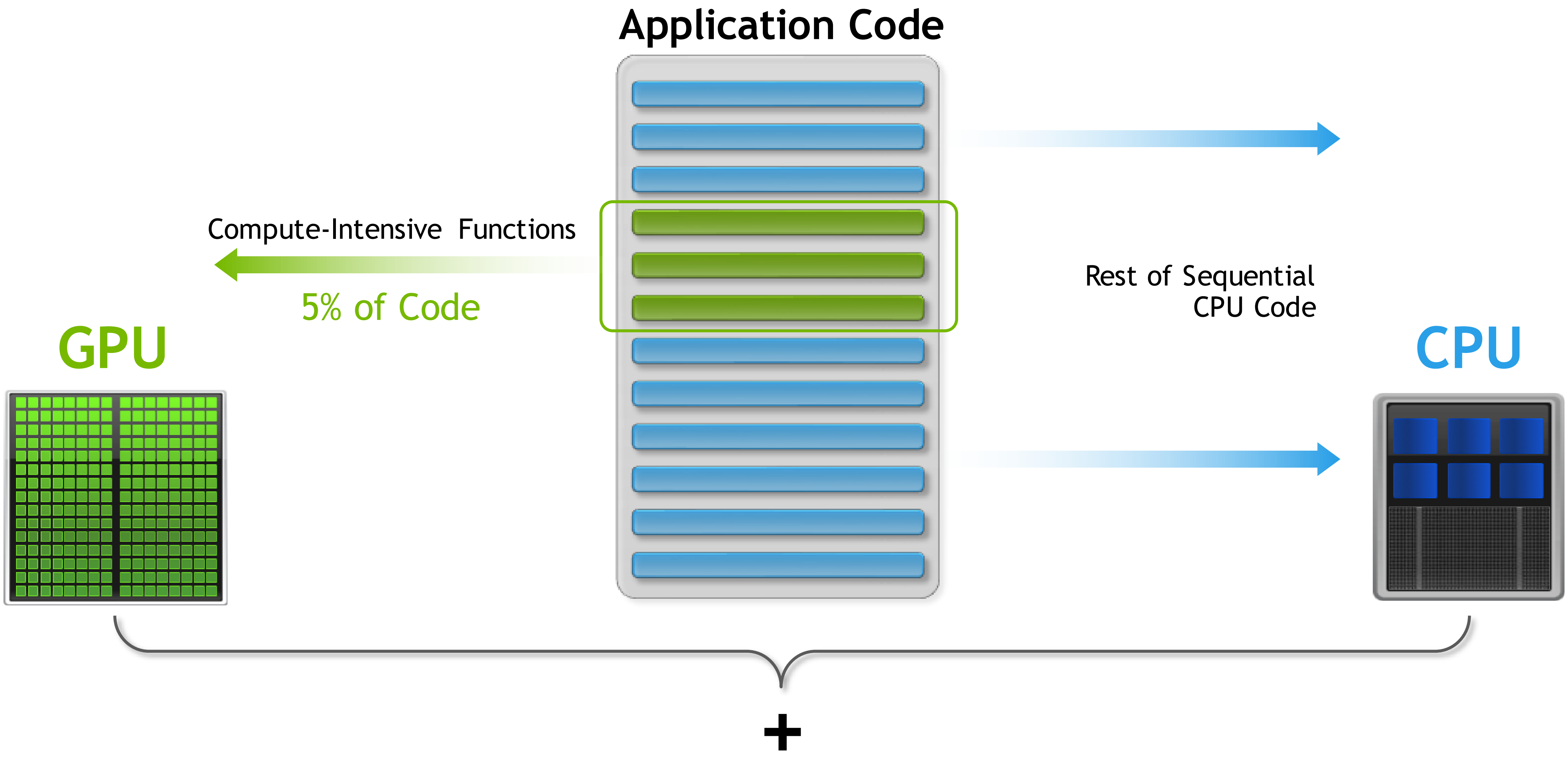


40 Years of Microprocessor Trend Data

Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

The Big Bang of Deep Learning

HOW GPU ACCELERATION WORKS



Application Code

Compute-Intensive Functions

5% of Code

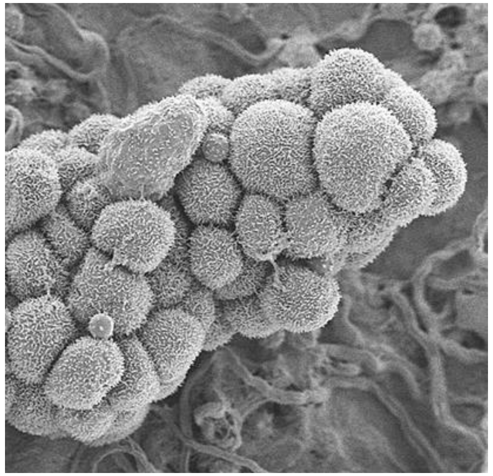
GPU

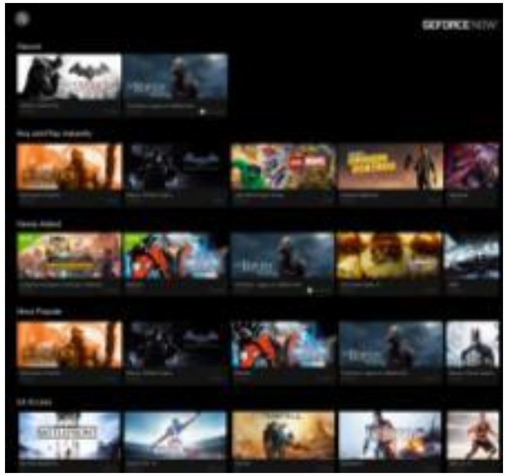
Rest of Sequential CPU Code

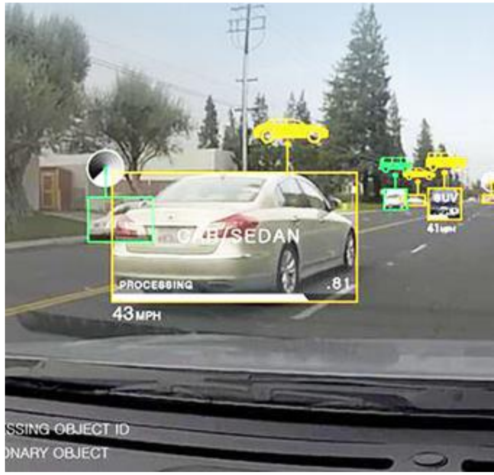
CPU

+

DEEP LEARNING EVERYWHERE







INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation

MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading
Drug Discovery

MEDIA & ENTERTAINMENT

Video Captioning
Video Search
Real Time Translation

SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery

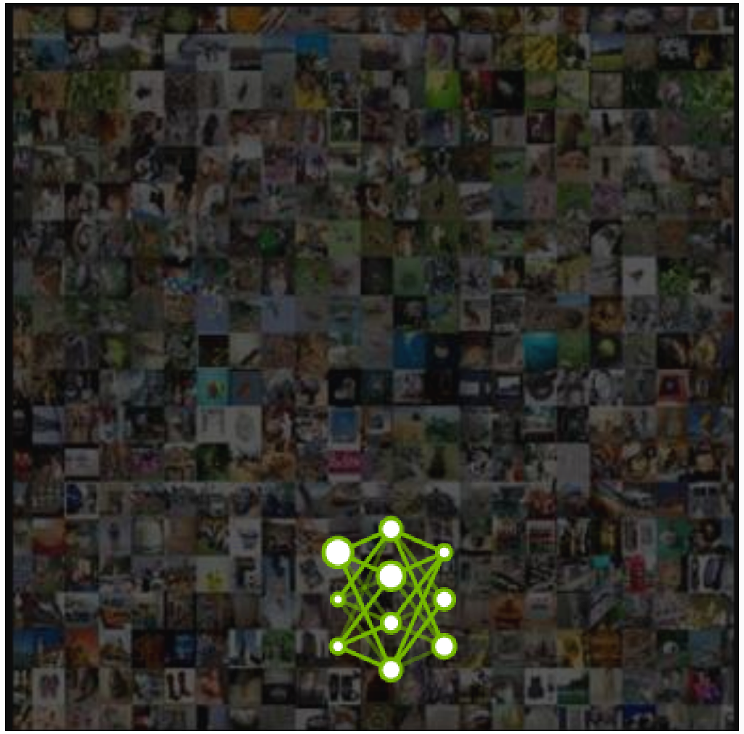
AUTONOMOUS MACHINES

Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

NEURAL NETWORK COMPLEXITY IS EXPLODING

To Tackle Increasingly Complex Challenges

7 ExaFLOPS
60 Million Parameters

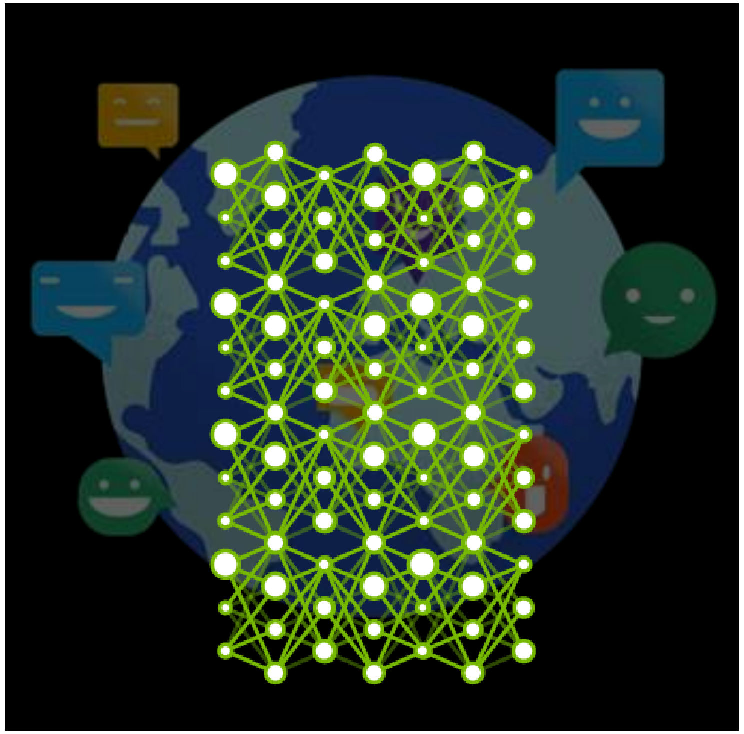


2015 - Microsoft ResNet
Superhuman Image Recognition

20 ExaFLOPS
300 Million Parameters

2016 - Baidu Deep Speech 2
Superhuman Voice Recognition

100 ExaFLOPS
8700 Million Parameters

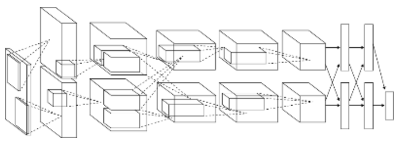


2017 - Google Neural Machine Translation
Near Human Language Translation

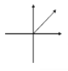
@immsrini

CAMBRIAN EXPLOSION

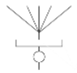
Convolutional Networks



Encoder/Decoder

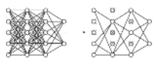


ReLu

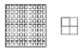


BatchNorm

Concat

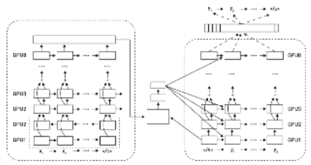


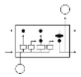
Dropout



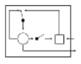
Pooling

Recurrent Networks

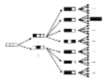




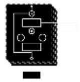
LSTM



GRU

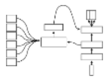


Beam Search



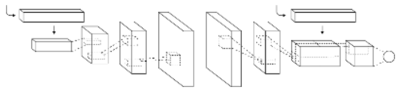
WaveNet

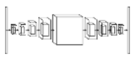
CTC



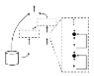
Attention

Generative Adversarial Networks



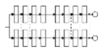


3D-GAN

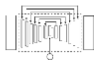


MedGAN

Conditional GAN

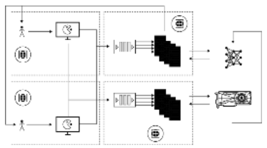


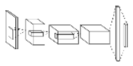
Coupled GAN



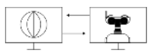
Speech Enhancement GAN

Reinforcement Learning

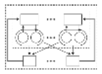




DQN

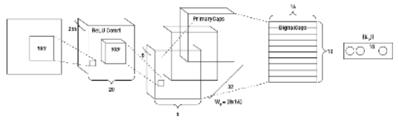


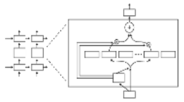
Simulation



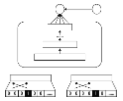
DDPG

New Species





Mixture of Experts



Neural Collaborative Filtering

Block Sparse LSTM

TESLA V100 32GB TENSOR CORE GPU

World’s Most Advanced Data Center GPU

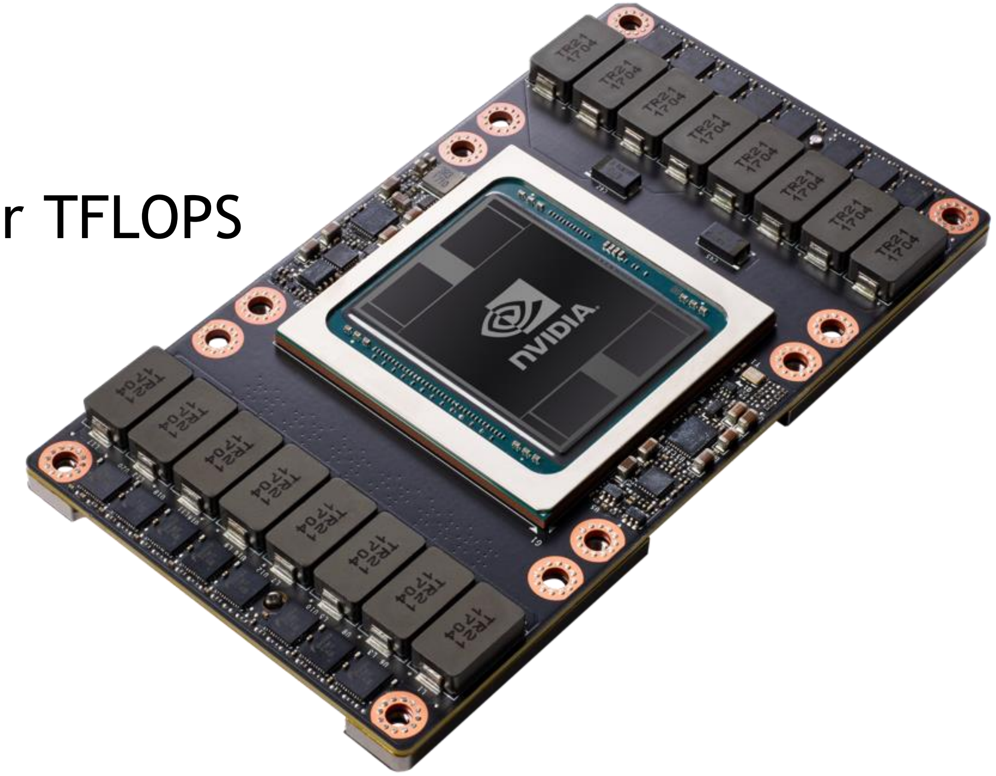
5,120 CUDA cores

640 NEW Tensor cores

7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS | 125 Tensor TFLOPS

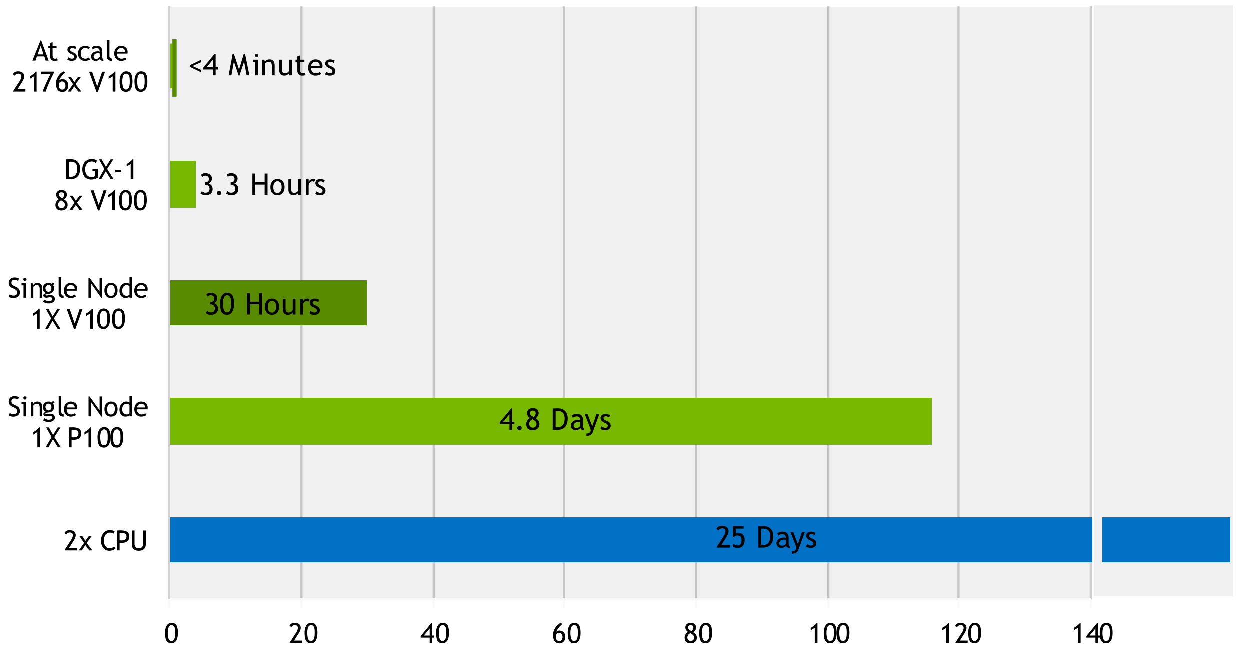
20MB SM RF | 16MB Cache

32GB HBM2 @ 900GB/s | 300GB/s NVLink



TESLA PLATFORM ENABLES DRAMATIC REDUCTION IN TIME TO TRAIN

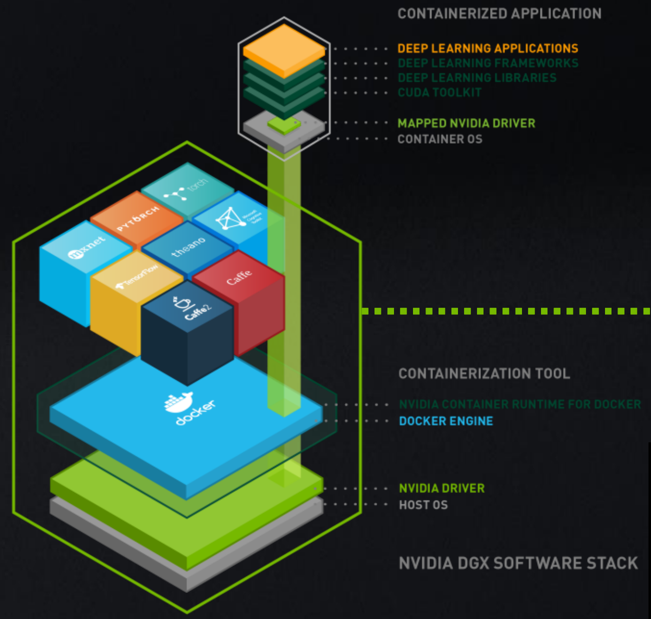
Relative Time to Train Improvements
(ResNet-50)



ResNet-50, 90 epochs to solution | CPU Server: dual socket Intel Xeon Gold 6140
Sony 2176x V100 record on https://nnabla.org/paper/imagenet_in_224sec.pdf

PURPOSE-BUILT AI SUPERCOMPUTERS

NGC DL SOFTWARE STACK



- Universal SW for Deep Learning
- Predictable execution across platforms
- Pervasive reach

AI WORKSTATION

AI DATA CENTER

DGX Station

The Personal
AI Supercomputer

DGX-1

The Essential
Instrument for AI
Development

DGX-2

The World’s Most Powerful
AI System for the Most
Complex AI Challenges

NVIDIA CONFIDENTIAL. DO NOT DISTRIBUTE.

POWERING THE DEEP LEARNING ECOSYSTEM

DGX-1 AI Supercomputer-in-a-Box

1 PFLOPS | 8x Tesla V100 32 GB | NVLink Hybrid Cube Mesh

2x Xeon | 8 TB RAID 0 | Quad IB 100Gbps, Dual 10GbE | 3U — 3500W

DESIGNED FOR THE DESK

The Only Supercomputer Designed for Your Office

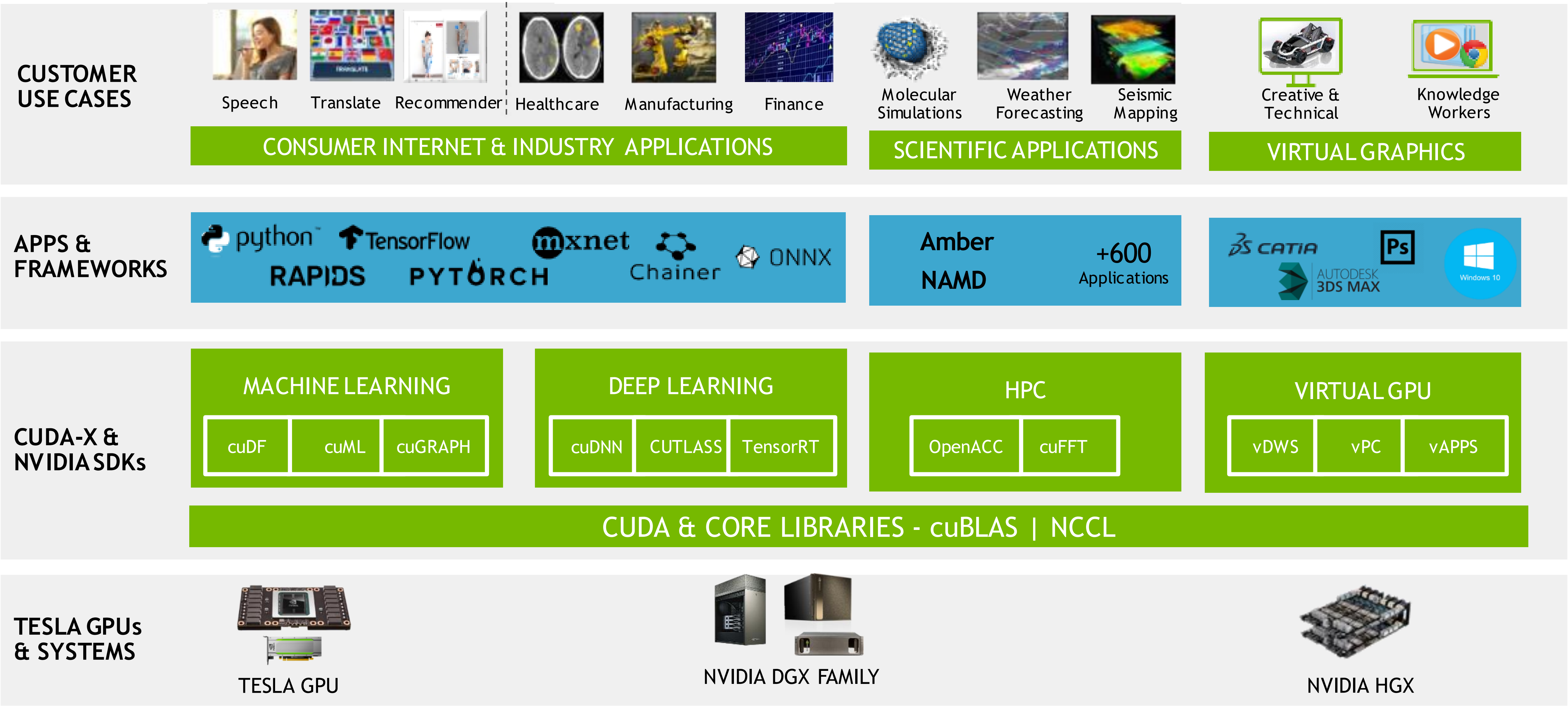
500 TFLOPS (FP 16)
4 x TESLA V100 with NVLINK

Consuming only 1500W, it draws only 1/20th the power

Emitting only 1/10th the noise of other workstations

NVIDIA DATA CENTER PLATFORM

Single Platform Drives Utilization and Productivity



CUSTOMER USE CASES

Speech | Translate | Recommender | Healthcare | Manufacturing | Finance

CONSUMER INTERNET & INDUSTRY APPLICATIONS

Molecular Simulations | Weather Forecasting | Seismic Mapping

SCIENTIFIC APPLICATIONS

Creative & Technical | Knowledge Workers

VIRTUAL GRAPHICS

APPS & FRAMEWORKS

python | TensorFlow | mxnet | Chainer | ONNX | RAPIDS | PYTORCH

Amber | NAMD | +600 Applications

CATIA | Ps | AUTODESK 3DS MAX | Windows 10

CUDA-X & NVIDIA SDKs

MACHINE LEARNING: cuDF | cuML | cuGRAPH

DEEP LEARNING: cuDNN | CUTLASS | TensorRT

HPC: OpenACC | cuFFT

VIRTUAL GPU: vDWS | vPC | vAPPS

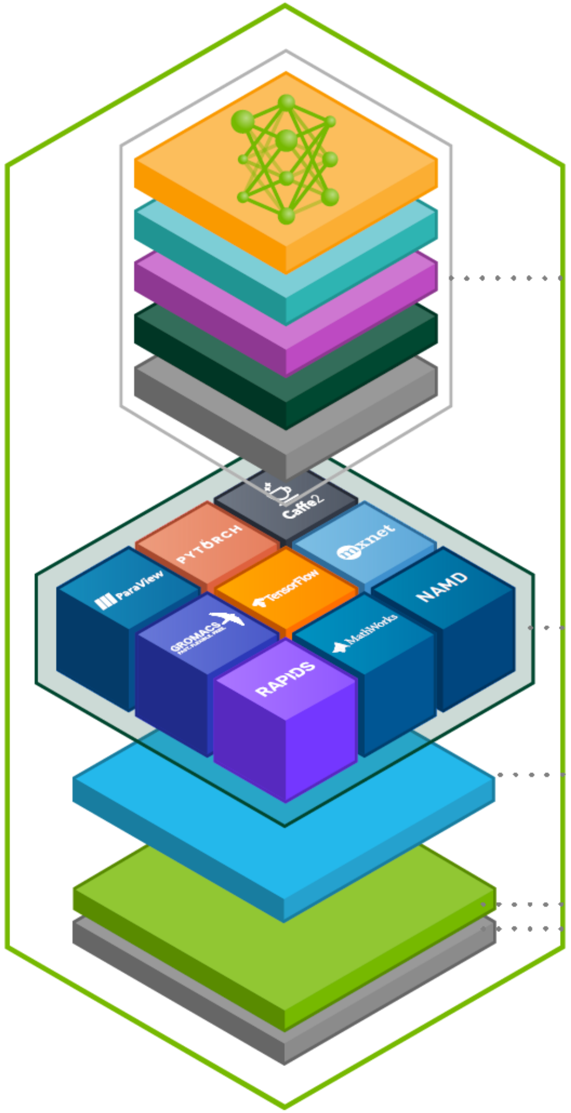
CUDA & CORE LIBRARIES - cuBLAS | NCCL

TESLA GPUs & SYSTEMS

TESLA GPU | NVIDIA DGX FAMILY | NVIDIA HGX

DGX SOFTWARE STACK

Fully Integrated Software Built on CUDA-X AI for Instant Productivity



PRE-TRAINED MODELS AND MODEL SCRIPTS

DEEP LEARNING, MACHINE LEARNING,
HPC APPLICATION CONTAINERS

CONTAINER RUNTIME

NVIDIA DRIVER

HOST OS

NGC SOFTWARE STACK

Advantages:

Instant productivity with NVIDIA optimized AI software

Caffe, MXNet, PyTorch, RAPIDS, TensorFlow, TensorRT, and more

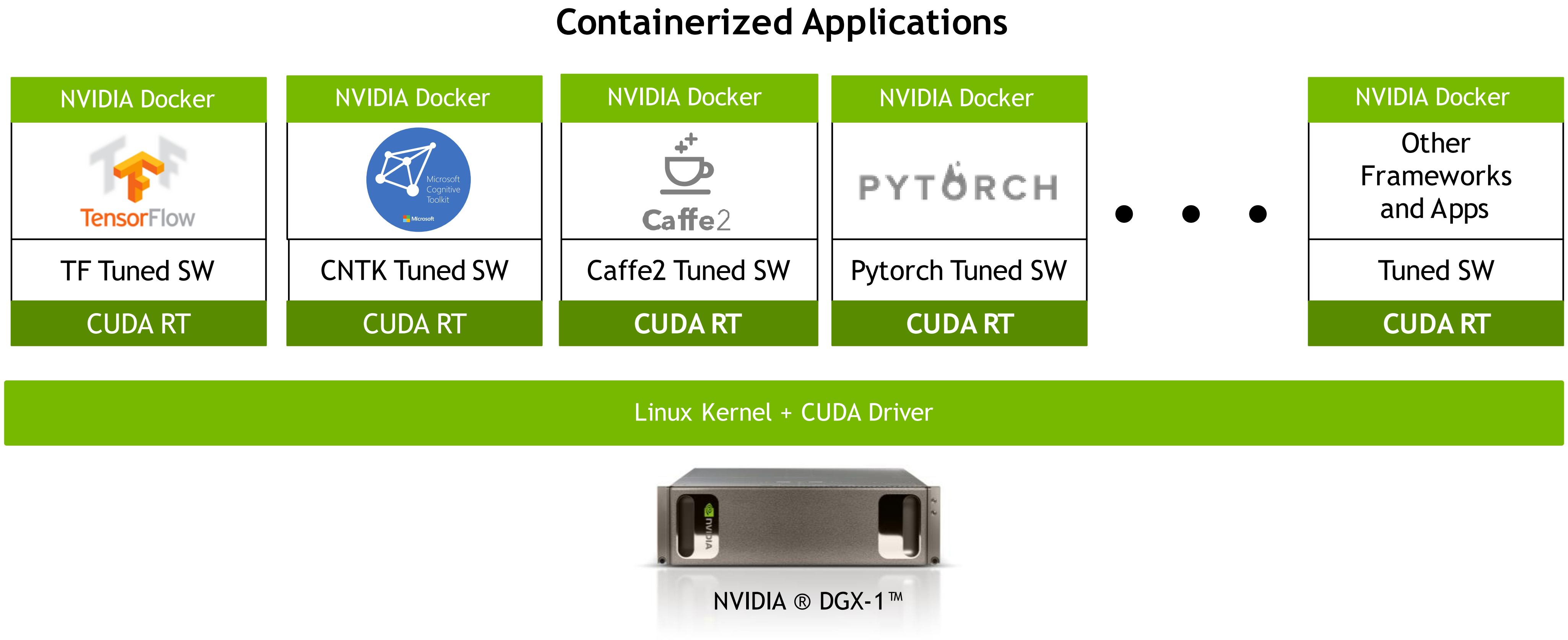
Performance optimized across the entire stack

Faster Time-to-Insight with pre-built, tested, and ready to run containers

Flexibility to use different versions of libraries like libc, cuDNN in each container

THE POWER TO RUN MULTIPLE FRAMEWORKS AT ONCE

Container Images portable across new driver versions



Containerized Applications

NVIDIA Docker
TensorFlow
TF Tuned SW
CUDA RT

NVIDIA Docker
Microsoft Cognitive Toolkit
CNTK Tuned SW
CUDA RT

NVIDIA Docker
Caffe2
Caffe2 Tuned SW
CUDA RT

NVIDIA Docker
PYTORCH
Pytorch Tuned SW
CUDA RT

• • •

NVIDIA Docker
Other Frameworks and Apps
Tuned SW
CUDA RT

Linux Kernel + CUDA Driver

NVIDIA ® DGX-1™

GPU POWERED INFERENCING

TESLA T4

WORLD'S MOST EFFICIENT GPU FOR MAINSTREAM SERVERS

320 Turing Tensor Cores
2,560 CUDA Cores
65 FP16 TFLOPS | 130 INT8 TOPS | 260 INT4 TOPS
16GB | 320GB/s
70 W

THE JETSON FAMILY

for AI at the Edge and Autonomous System designs

JETSON NANO
0.5 TFLOPS (FP16)

5 - 10W
45mm x 70mm

JETSON TX2 series
1.3 TFLOPS (FP16)

7.5 – 15W*
50mm x 87mm

JETSON Xavier NX
6 TFLOPS (FP16)
21 TOPS (INT8)

10 - 15W
45mm x 70mm

JETSON AGX XAVIER series
11 TFLOPS (FP16)
32 TOPS (INT8)

10 – 30W
100mm x 87mm

AI at the edge — Fully autonomous machines

Same software

Listed prices are for 1000u+ | Full specs at developer.nvidia.com/jetson

* TX2i: 10-20W

NVIDIA EGX EDGE COMPUTING

NGC

Third-Party ISVs

NVIDIA APPLICATION FRAMEWORKS

METROPOLIS
AI City

CLARA
AI Healthcare

METROPOLIS
AI Retail

ISAAC
AI Manufacturing

NVIDIA EDGE STACK

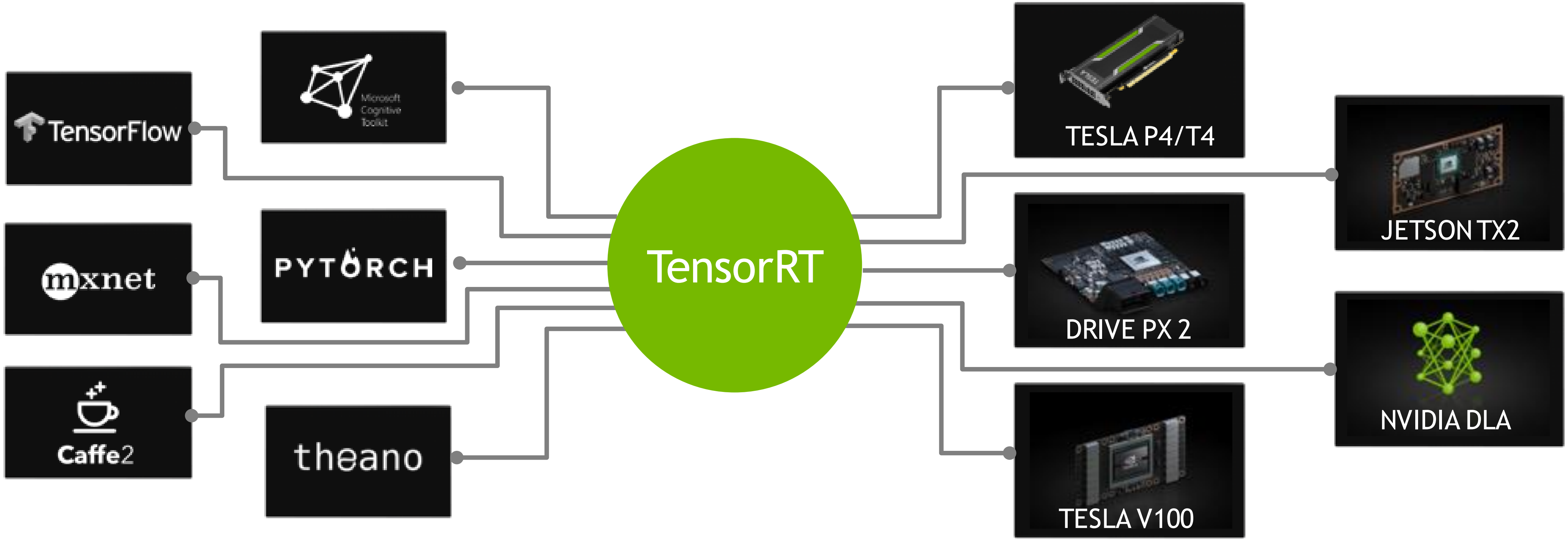
Kubernetes	Containers	CUDA-X	IoT Runtime

NVIDIA EGX EDGE COMPUTING PLATFORM

GPU | AI | STORAGE | NETWORKING | SECURITY

TENSORRT

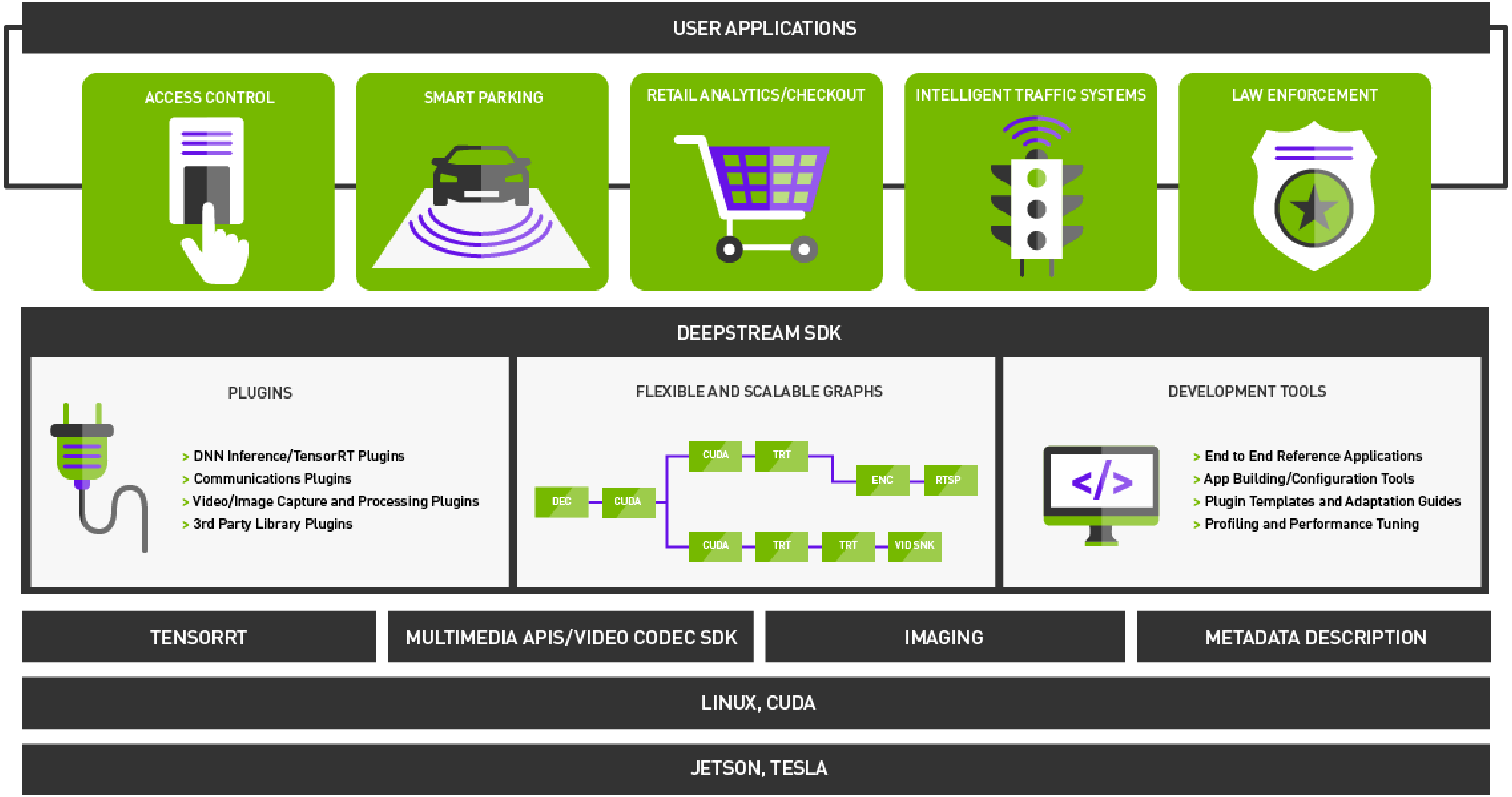
From Every Framework, Optimized For Each Target Platform



Frameworks

Platforms

DEEPSTREAM SDK



USER APPLICATIONS

ACCESS CONTROL

SMART PARKING

RETAIL ANALYTICS/CHECKOUT

INTELLIGENT TRAFFIC SYSTEMS

LAW ENFORCEMENT

DEEPSTREAM SDK

PLUGINS

- DNN Inference/TensorRT Plugins
- Communications Plugins
- Video/Image Capture and Processing Plugins
- 3rd Party Library Plugins

FLEXIBLE AND SCALABLE GRAPHS

DEVELOPMENT TOOLS

- End to End Reference Applications
- App Building/Configuration Tools
- Plugin Templates and Adaptation Guides
- Profiling and Performance Tuning

TENSORRT

MULTIMEDIA APIS/VIDEO CODEC SDK

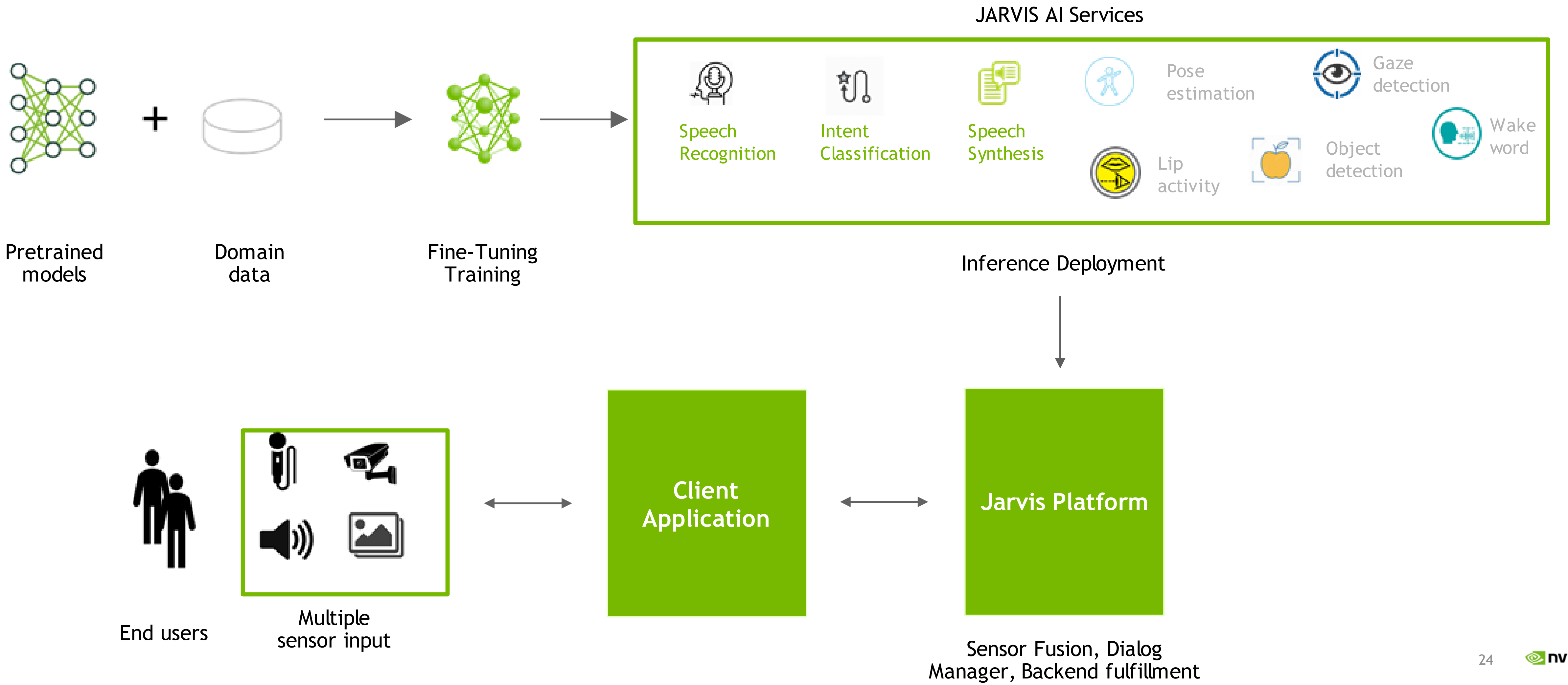
IMAGING

METADATA DESCRIPTION

LINUX, CUDA

JETSON, TESLA

JARVIS WORKFLOW OVERVIEW



JARVIS AI Services

Speech Recognition

Intent Classification

Speech Synthesis

Pose estimation

Gaze detection

Wake word

Lip activity

Object detection

Pretrained models

+

Domain data

Fine-Tuning Training

Inference Deployment

End users

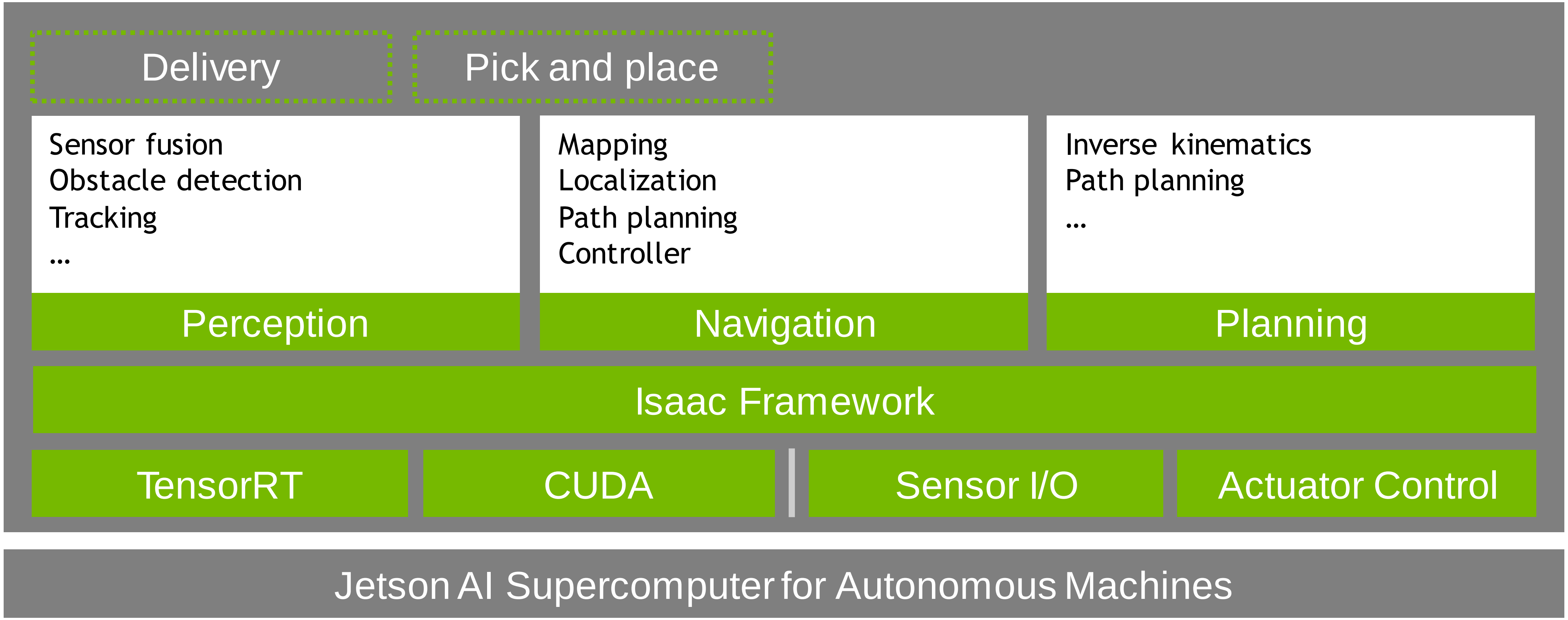
Multiple sensor input

Client Application

Jarvis Platform

Sensor Fusion, Dialog Manager, Backend fulfillment

ISAAC SDK FOR ROBOTICS



Delivery

Pick and place

Sensor fusion
Obstacle detection
Tracking
…

Perception

Mapping
Localization
Path planning
Controller

Navigation

Inverse kinematics
Path planning
…

Planning

Isaac Framework

TensorRT

CUDA

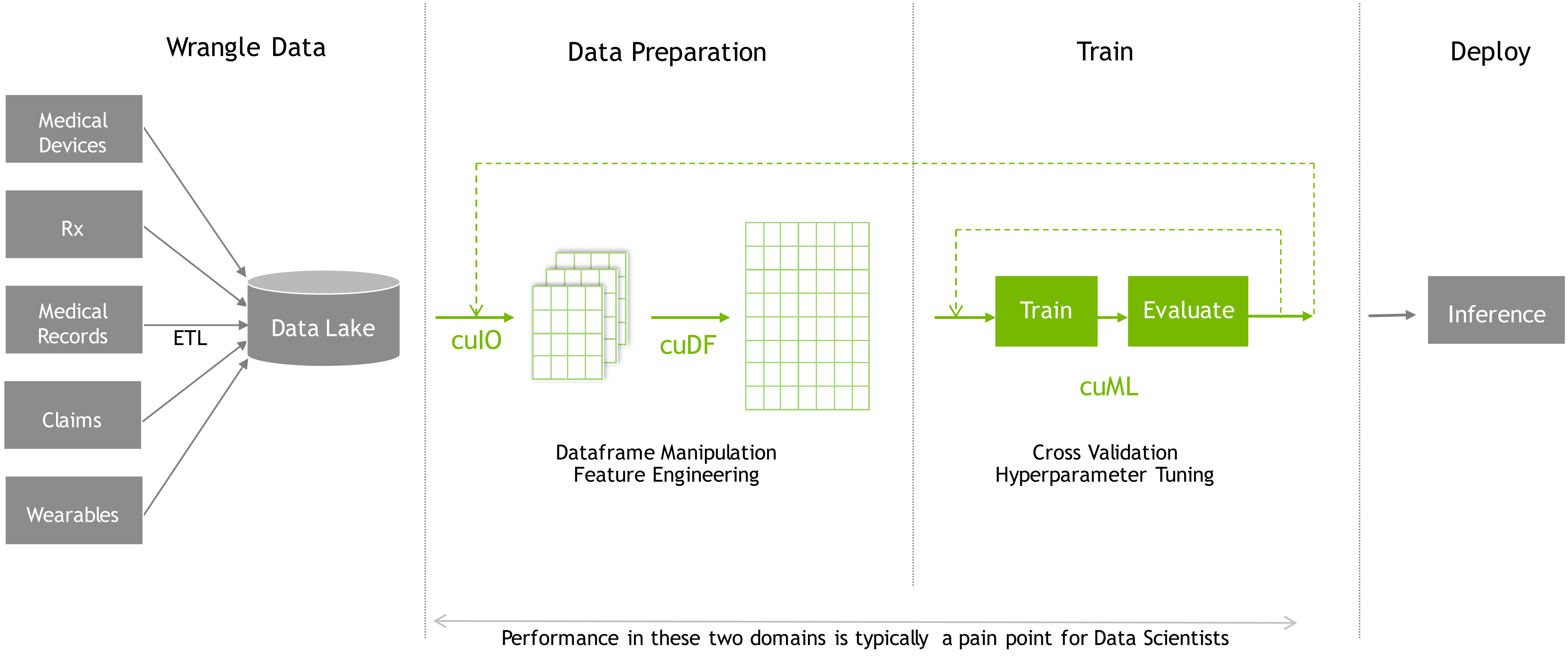
Sensor I/O

Actuator Control

Jetson AI Supercomputer for Autonomous Machines

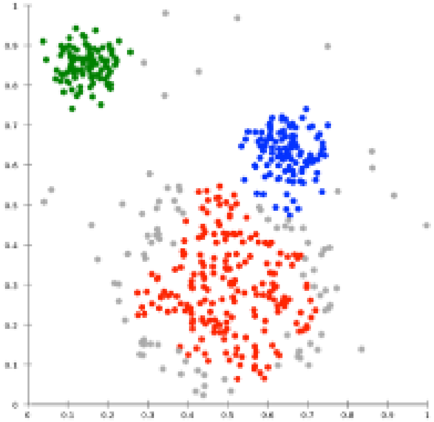
MACHINE LEARNING/ RAPIDS

RAPIDS IN DATA SCIENCE



ALGORITHMS

GPU-accelerated Scikit-Learn

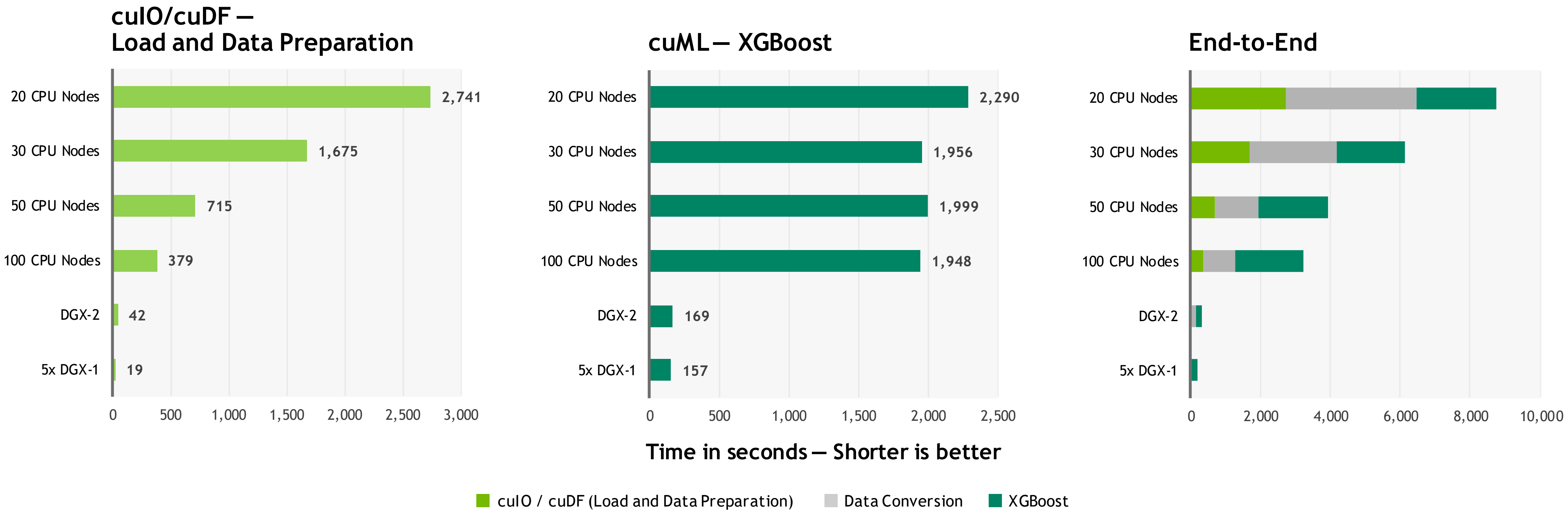


Classification / Regression	Decision Trees / Random Forests Linear Regression Logistic Regression K-Nearest Neighbors
Statistical Inference	Kalman Filtering Bayesian Inference
Clustering	K-Means DBSCAN
Decomposition & Dimensionality Reduction	Principal Components Singular Value Decomposition
Timeseries Forecasting	ARIMA
Recommendations	Collaborative Filtering

Cross Validation

More to come!

BENCHMARKS



cuIO/cuDF — Load and Data Preparation

cuML — XGBoost

End-to-End

Time in seconds — Shorter is better

Benchmark

200GB CSV dataset; Data preparation includes joins, variable transformations.

CPU Cluster Configuration

CPU nodes (61 GiB of memory, 8 vCPUs, 64-bit platform), Apache Spark

DGX Cluster Configuration

5x DGX-1 on InfiniBand network

TRADITIONAL
DATA SCIENCE CLUSTER

Workload Profile:

Fannie Mae Mortgage Data:

- 192GB data set
- 16 years, 68 quarters
- 34.7 Million single family mortgage loans
- 1.85 Billion performance records
- XGBoost training set: 50 features

300 Servers | $3M | 180 kW

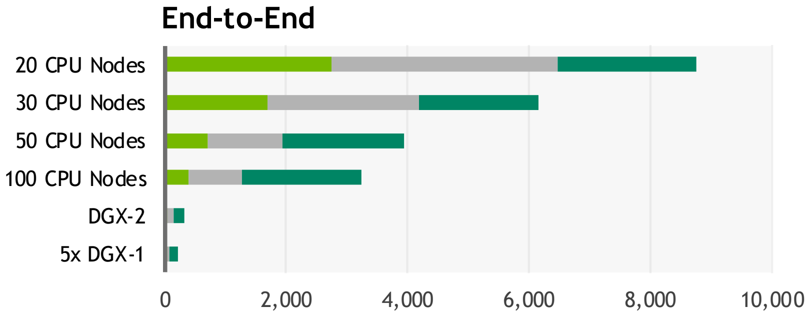
GPU-ACCELERATED MACHINE LEARNING CLUSTER

DGX-2 and RAPIDS for Predictive Analytics

1 DGX-2 | 10 kW

1/8 the Cost | 1/15 the Space

1/18 the Power



MORE INFORMATION

SIGNUP:

NVIDIA DEVELOPER FORUM - To keep you updated

http://developer.nvidia.com

✉ snagalingam@nvidia.com ☎ 99455 67685

DEEP LEARNING - THE NEW PARADIGM OF COMPUTING

Sundara Ramalingam N

snagalingam@nvidia.com

+91 99455 67685