REVOLUTIONIZING RETAIL WITH ARTIFICIAL INTELLIGENCE

Scott Brubaker, Paul Hendricks & Alex Sabatier

NVIDIA

INCEPTION PARTNERS & RETAIL ECOSYSTEM

Physical / In-store

VISION-BASED APPLICATIONS

focal Systems
[Sc] STANDARD COGNITION
AiFi
deepmagic™
Motionloft
Aipoly
MAXERIENCE Intuition. Imagination. Inspiration
mashgin
TRACXPOINT

ROBOTS, DRONES

fellow ROBOTS
6 RIVER SYSTEMS
dispatch
marble
IFM
kiwi

DIGITAL SIGNAGE

ADVERTIMA
FACEMEDIA
MYSTOR-E DIFFERENT CUSTOMERS, PERSONALIZED STORE
QM SCIENTIFIC
SENSAPE Interactive Infotainment

Online

VISUAL SEARCH, TAGGING

MARKABLE
FASHWELL
GROKSTYLE
VUE.AI
DigitalBridge
THETAKE AI
nFlate
huew
deepomatic
Syte
iLenze
shoppr
ViSENZE™
ProductAI
Shopcinity
Slyce
OMNIOUS

CONVERSATIONAL COMMERCE

InfiniteAnalytics
Kata.ai
Streamoid
shelfai
chatShopper
Twiggle

MARKETING, ANALYTICS

ABEJA
:) Affectiva
Artifacia
daisy INTELLIGENCE
heuritech
omni sci
netra
opius
MARIANAIQ

OTHER

N-TECH. LAB
MODIFACE

NVIDIA.

AI FOR RETAIL

SHOPPING EXPERIENCE

SUPPLY CHAIN

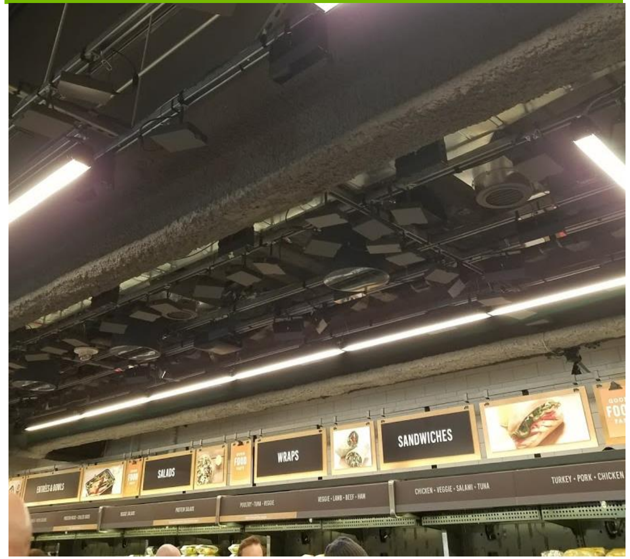
CORPORATE HEADQUARTERS

NVIDIA.

SHOPPING EXPERIENCE: STORE (IVA)

LOSS PREVENTION, SHOPPER TRACKING

FRICTIONLESS COMMERCE



INVENTORY ANALYSIS

TOP RETAIL IVA USE CASES

$50B in annual shrinkage in US alone

LOSS PREVENTION

Ticket Switching
Mis-scanning
Sweethearting
Security

50% of top retailers will implement IVA for store analytics

STORE ANALYTICS

Heat Mapping
Demographic Analysis
Shopper/Employee tracking
Stock Out
Customer Engagement
Price Matching
Pick-up

Autonomous checkout locations to increase 4x annually for next 3 years

AUTONOMOUS SHOPPING

Autonomous Checkout
Nano Stores
Smart Cabinets

SHOPPING EXPERIENCE: ONLINE

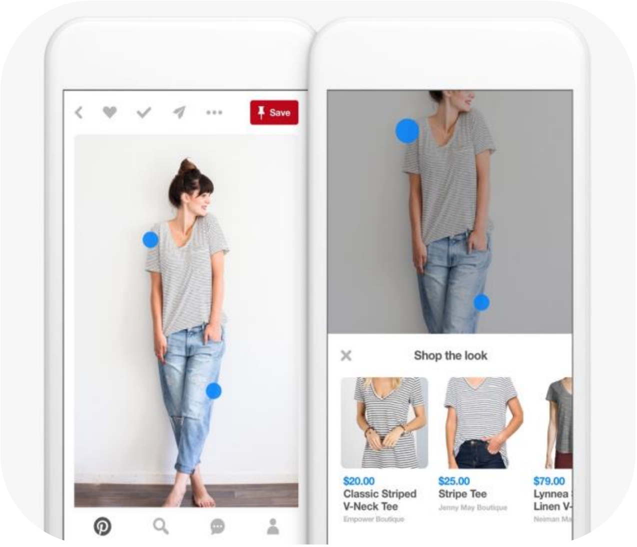
RECOMMENDATION ENGINE

STITCH FIX

AR/VR CONSUMER INTERACTION

OlAY

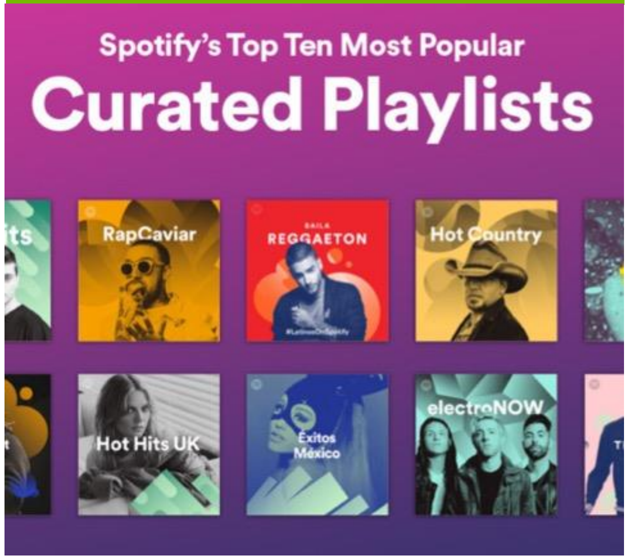
IMAGE-BASED SEARCH



NVIDIA

RECOMMENDATION ENGINES ON GPU CLOUD

SONG RECOMMENDATIONS



Spotify

VIDEO RECOMMENDATIONS

NETFLIX

TARGETED RECOMMENDATIONS

yelp

AI IN SUPPLY CHAIN

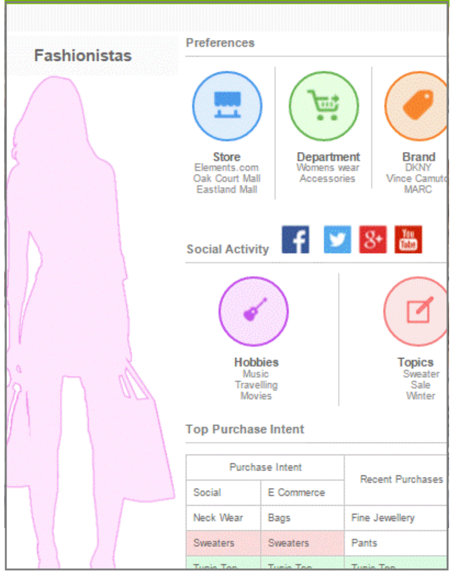
WAREHOUSE OPTIMIZATION

DYNAMIC SUPPLY CHAIN REAL-TIME RE-ROUTING

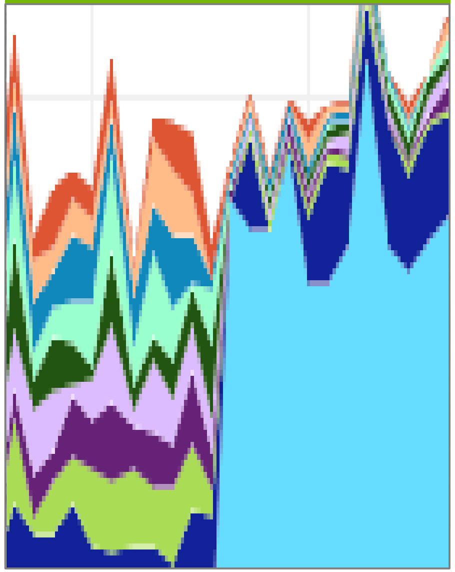
FORECASTING AND REPLENISHMENT

AI AT CORPORATE HQ

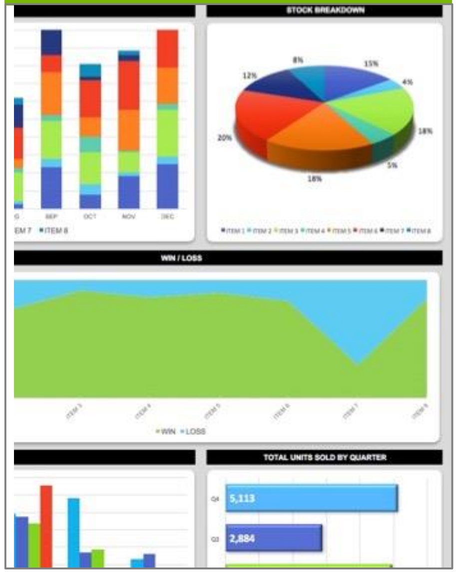
SINGLE VIEW OF CONSUMER



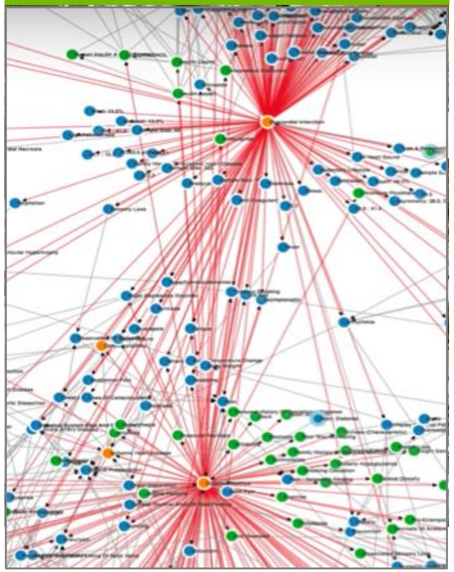
DEMAND SIGNAL ANALYSIS



AD SPEND OPTIMIZATION



PREDICTIVE ANALYTICS



GPU POWERED MACHINE LEARNING

DATA SCIENCE IN RETAIL

- Supply Chain Replenishment
- Inventory Management
- Price Simulation & Management
- Prioritize Promotion - Ad Targeting
- Marketing Optimization
- Personalized Recommendations
- Truck Routing
- Online Delivery

RAPIDS

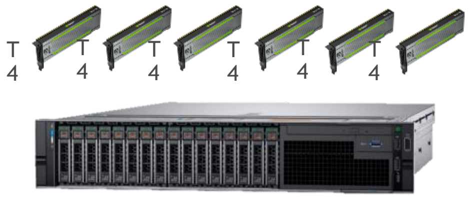
THE STORE OF THE FUTURE

Future-Proofed IVA Infrastructure

DL-BASED IVA EDGE USE CASES

Loss Prevention
Stock Out Reduction
Store Analytics
Security

Server Back of Store



Server (6 x T4s)

In-Store

Cameras

Sensors

Jetson AGX Xavier / **Nano**

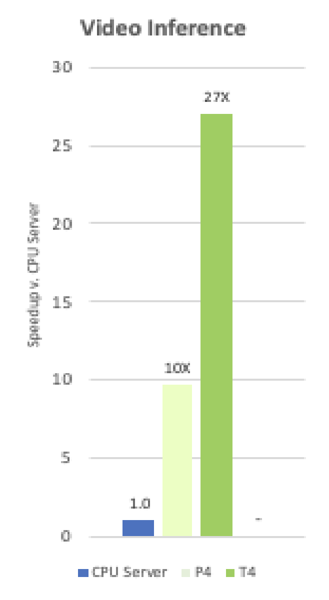
NVIDIA VALUE

Comprehensive Platform for Retail IVA

NVIDIA DELIVERS	IVA Inference w/NVIDIA T4 GPU
Speed Up	27*X CPU
Images/second (1080P)	4400
Metropolis Platform optimized for IVA	DS Inference SDK TensorRT
GPU accelerated IVA Software Partners	70+
Deep Learning Education	Developer Blogs + IVA DLI

* Based on ResNet-50

GPU hardware accelerator engines for video decoding and encoding support faster than real-time video processing.



Speedup: 27x Faster
ResNet-50

ART OF THE POSSIBLE

The State of AI in Retail

Paul Hendricks

Solutions Architect

phendricks@nvidia.com

INTRODUCTION

- Paul Hendricks is a Solutions Architect at NVIDIA, helping enterprise customers with their deep learning and AI initiatives
- Paul's background is primarily in retail, and has spent the past 5 years working with many Fortune 500 retail companies to implement data science and AI solutions.
- Prior to joining NVIDIA, Paul worked at Victoria’s Secret as a Data Scientist building models to understand customer propensity to purchase and how to optimize assortment in stores.
- Currently, Paul's research at NVIDIA focuses on intelligent video analytics, machine leaning, recommendation systems, GANs, and reinforcement learning.

INTRODUCTION

- Paul Hendricks is a Solutions Architect at NVIDIA, helping enterprise customers with their deep learning and AI initiatives

- Paul's background is primarily in retail, and has spent the past 5 years working with many Fortune 500 retail companies to implement data science and AI solutions.

- Prior to joining NVIDIA, Paul worked at Victoria's Secret as a Data Scientist building models to understand customer propensity to purchase and how to optimize assortment in stores.

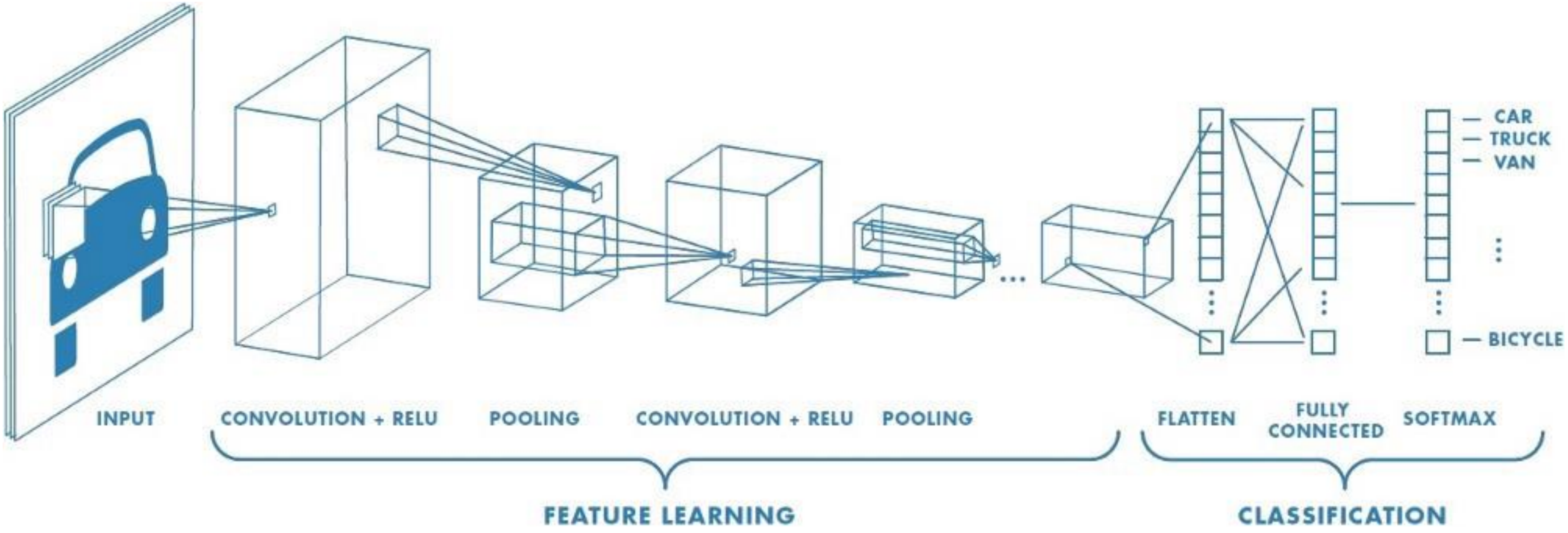
- Currently, Paul's research at NVIDIA focuses on intelligent video analytics, machine leaning, recommendation systems, GANs, and reinforcement learning.

INTELLIGENT VIDEO ANALYTICS

Image Classification

Problem Background

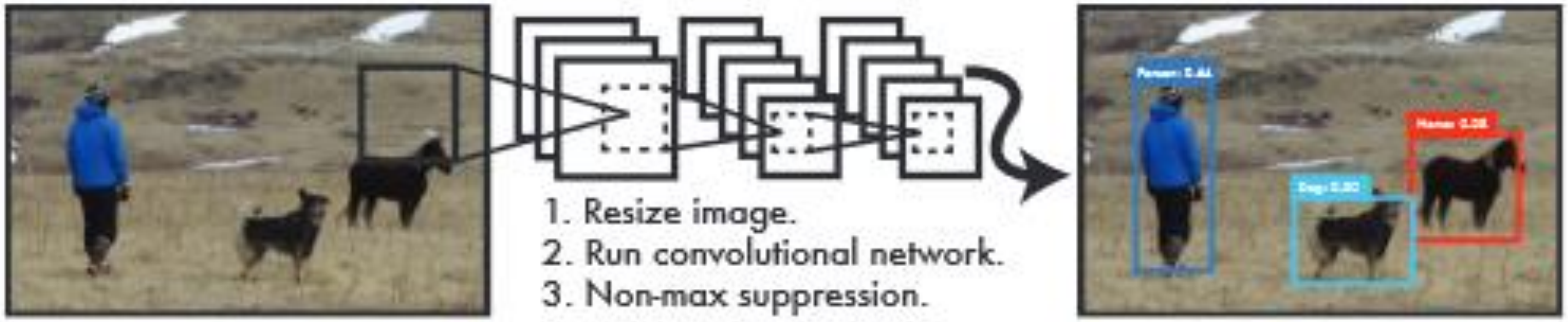
- Input Data: Images, Videos
- Goal: Given an input, identify the class that input belongs to



Object Detection

Problem Background

- Input Data: Images, Videos
- Goal: Given an input, identify objects and output bounding boxes around the objects and their classes



Object Segmentation (Semantic Segmentation)

Problem Background

- Input Data: Images, Videos
- Goal: Given an input, identify objects and output a mapping of pixels to their respective classes

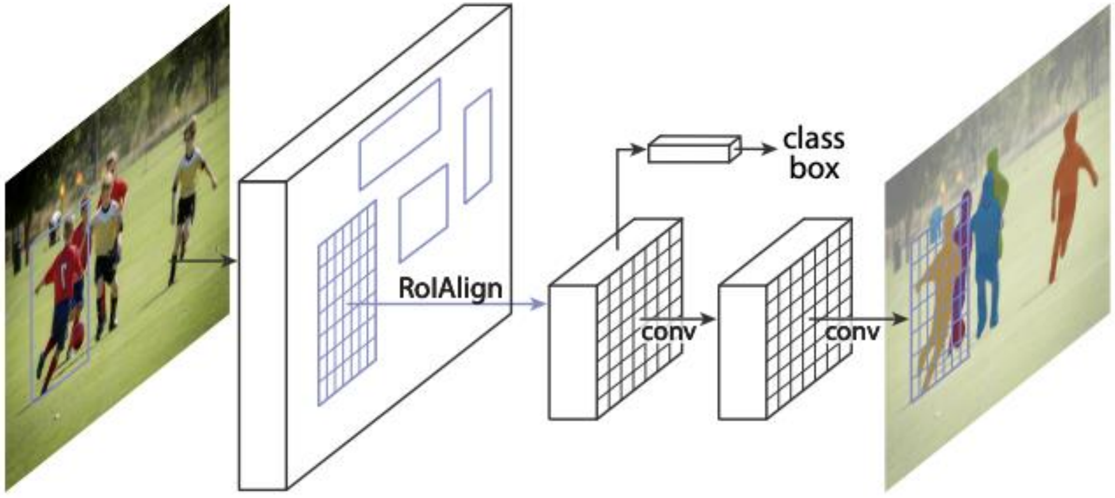


Figure 1. The **Mask R-CNN** framework for instance segmentation.

LOSS PREVENTION, STORE ANALYTICS, AND FRICTIONLESS CHECKOUT

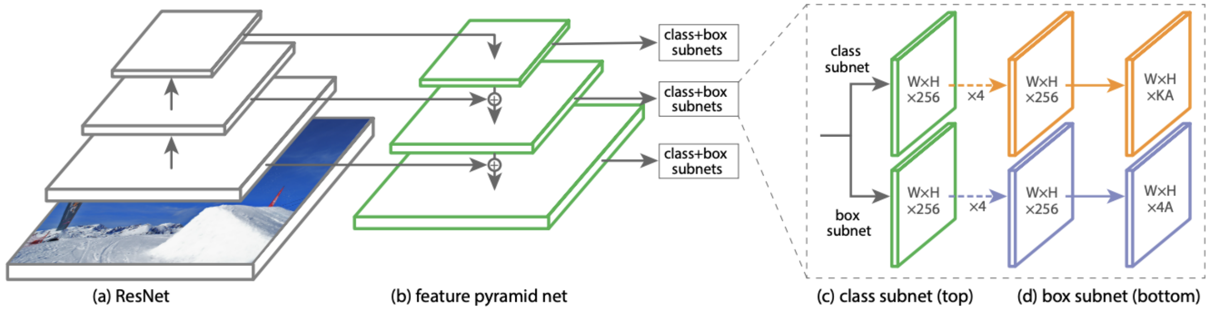
https://www.standardcognition.com/

LOCALIZING ALGORITHMS

LOCALIZING ALGORITHMS

Single Stage Detectors

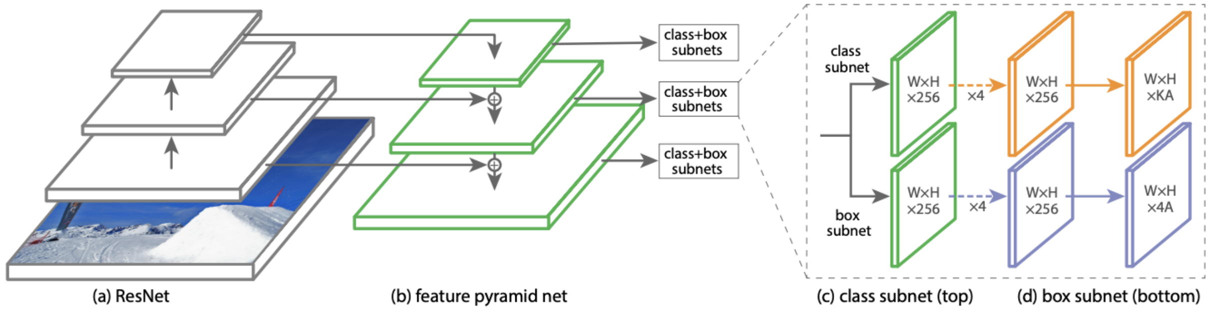
- These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass



LOCALIZING ALGORITHMS

Single Stage Detectors

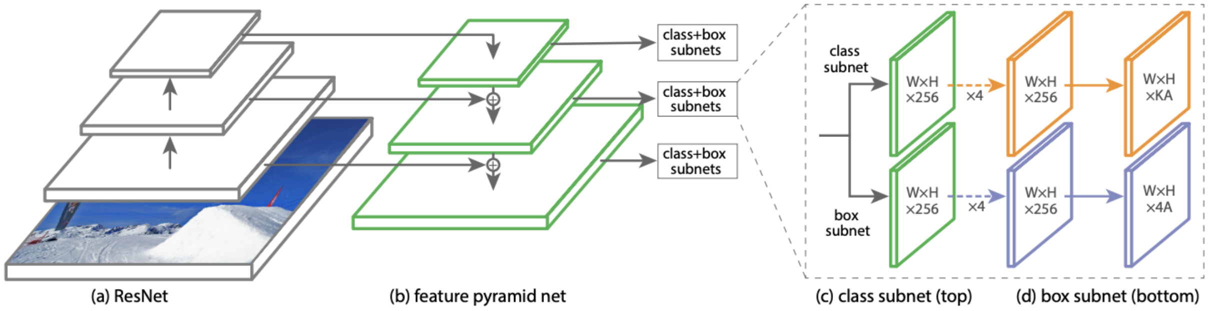
- These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
- Computationally efficient and can be very fast during inference



LOCALIZING ALGORITHMS

Single Stage Detectors

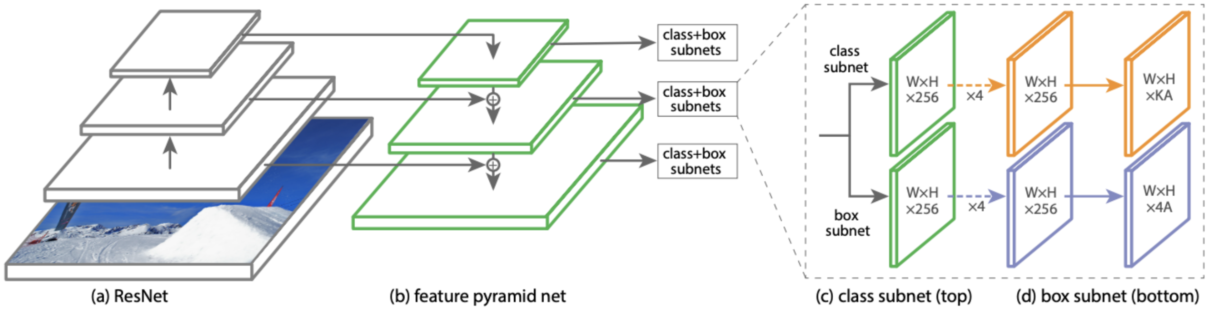
- These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
- Computationally efficient and can be very fast during inference
- Examples: YOLOv3, SSD, RetinaNet, RetinaMask



LOCALIZING ALGORITHMS

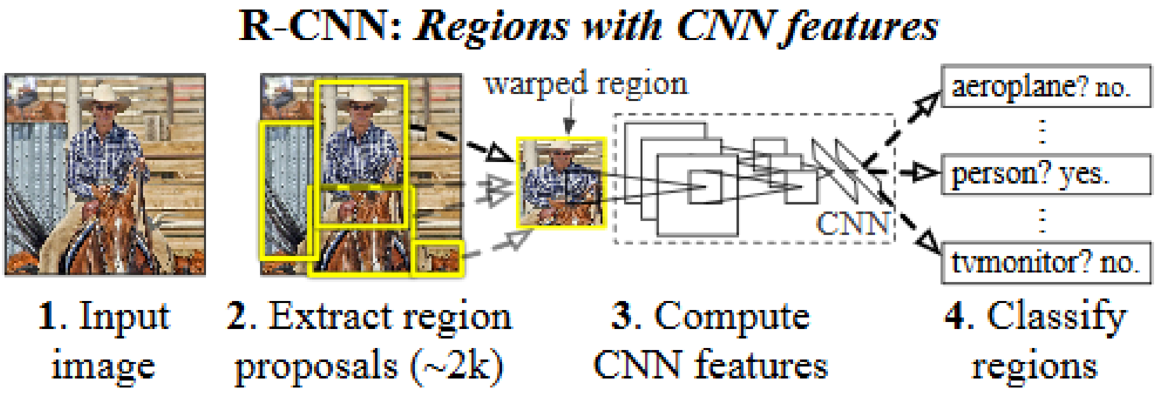
Single Stage Detectors

- These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
- Computationally efficient and can be very fast during inference
- Examples: YOLOv3, SSD, RetinaNet, RetinaMask



Two Stage Detectors

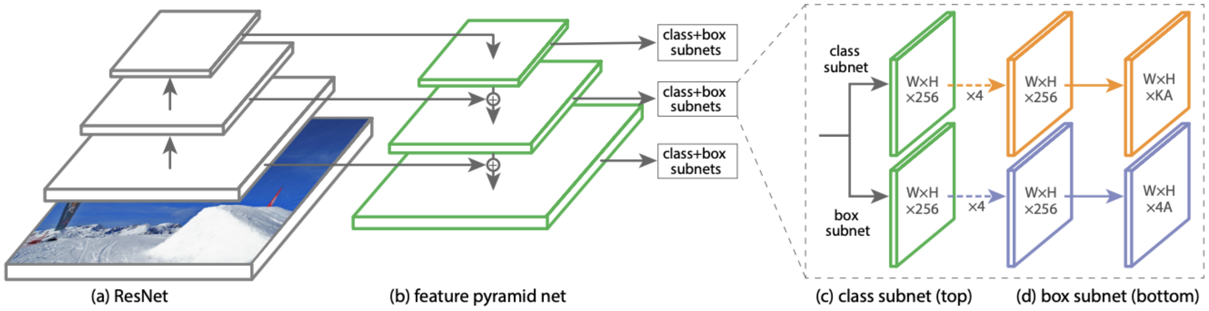
- These algorithms generate a number of region proposals which are then passed to a CNN and classified



LOCALIZING ALGORITHMS

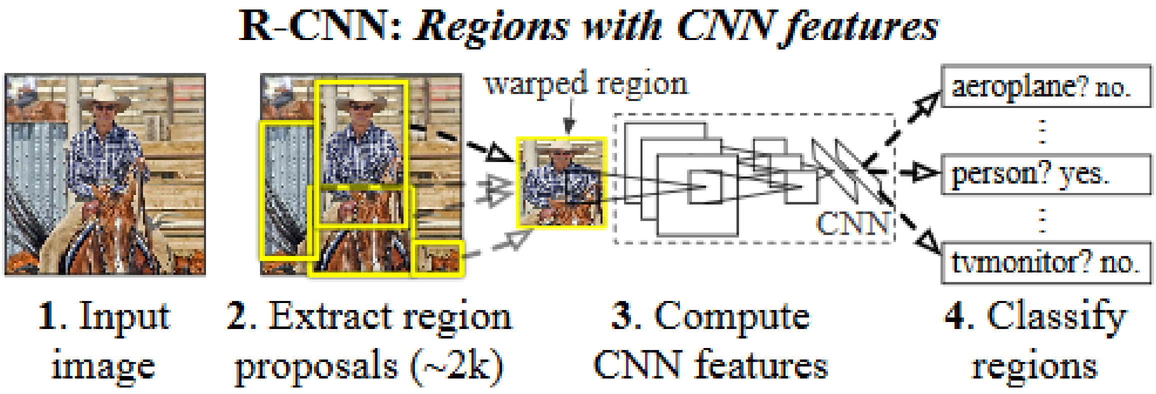
Single Stage Detectors

- These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
- Computationally efficient and can be very fast during inference
- Examples: YOLOv3, SSD, RetinaNet, RetinaMask



Two Stage Detectors

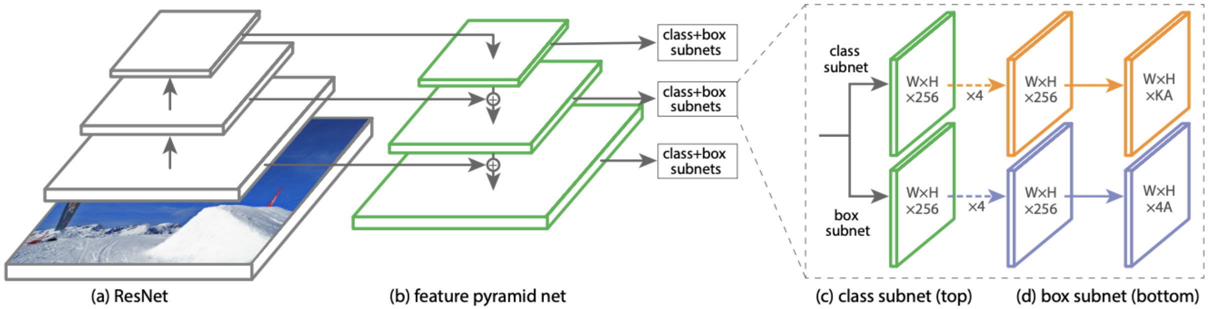
- These algorithms generate a number of region proposals which are then passed to a CNN and classified
- Slower during inference since regions must be proposed and then evaluated (often redundant if overlaps)



LOCALIZING ALGORITHMS

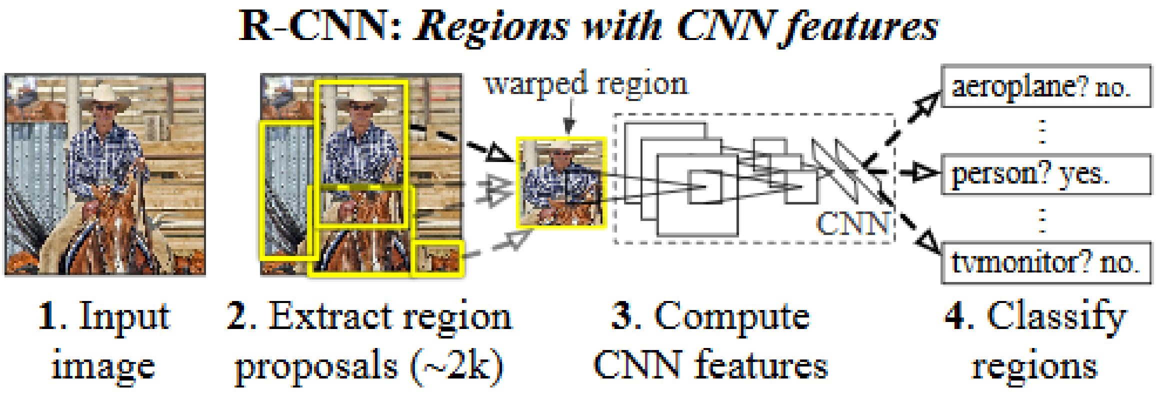
Single Stage Detectors

- These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
- Computationally efficient and can be very fast during inference
- Examples: YOLOv3, SSD, RetinaNet, RetinaMask



Two Stage Detectors

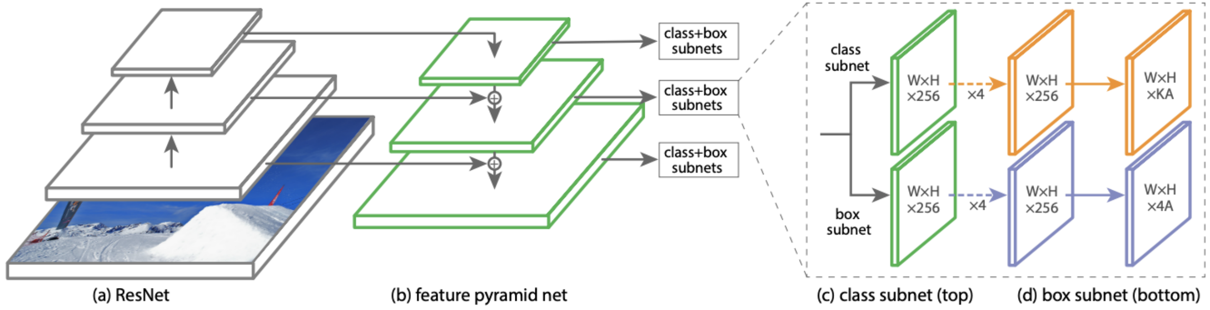
- These algorithms generate a number of region proposals which are then passed to a CNN and classified
- Slower during inference since regions must be proposed and then evaluated (often redundant if overlaps)
- Often are more accurate than single stage detectors, especially when trained on semantic segmentations



LOCALIZING ALGORITHMS

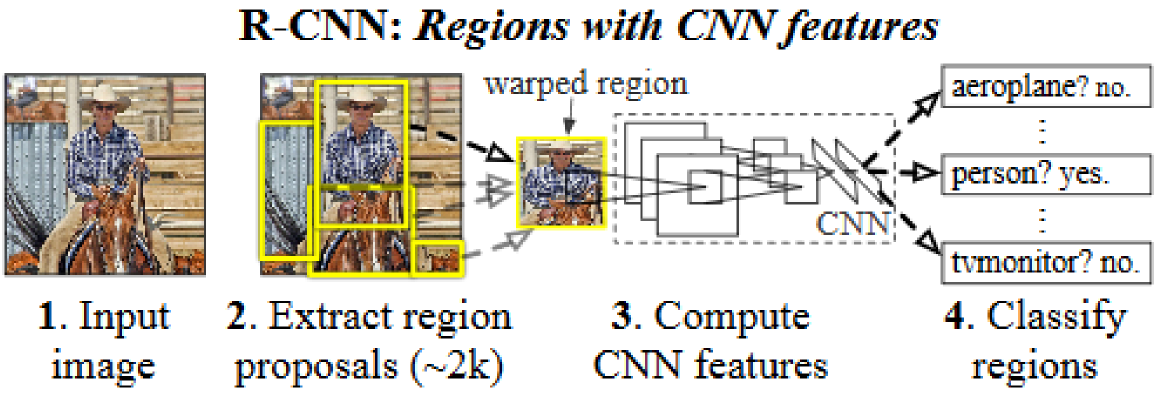
Single Stage Detectors

- These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
- Computationally efficient and can be very fast during inference
- Examples: YOLOv3, SSD, RetinaNet, RetinaMask



Two Stage Detectors

- These algorithms generate a number of region proposals which are then passed to a CNN and classified
- Slower during inference since regions must be proposed and then evaluated (often redundant if overlaps)
- Often are more accurate than single stage detectors, especially when trained on semantic segmentations
- Examples: Faster RCNN, Mask RCNN



GETTING STARTED

DLI Courses

- Introduction to Object Detection with TensorFlow – https://courses.nvidia.com/courses/course-v1:DLI+L-AV-04+V1

Papers

- YOLOV3 – https://pjreddie.com/publications/
- Faster RCNN – https://arxiv.org/pdf/1506.01497
- Mask RCNN – https://arxiv.org/abs/1703.06870
- RetinaNet – https://arxiv.org/abs/1708.02002
- RetinaMask – https://arxiv.org/abs/1901.03353

Libraries

- DarkNet – https://github.com/pjreddie/darknet
- TensorFlow's Object Detection API – https://github.com/tensorflow/models/tree/master/research/object_detection
- Facebook's Mask RCNN Benchmark – https://github.com/facebookresearch/maskrcnn-benchmark

Datasets

- ImageNet – https://www.kaggle.com/c/imagenet-object-detection-challenge
- Pascal VOC – http://host.robots.ox.ac.uk/pascal/VOC/
- COCO – http://cocodataset.org/
- Open Images – https://storage.googleapis.com/openimages/web/index.html

MACHINE LEARNING

DATA SCIENCE IN RETAIL

Supply Chain Replenishment

Inventory Management

Price Management / Markdown Optimization

Prioritize Promotion And Ad Targeting

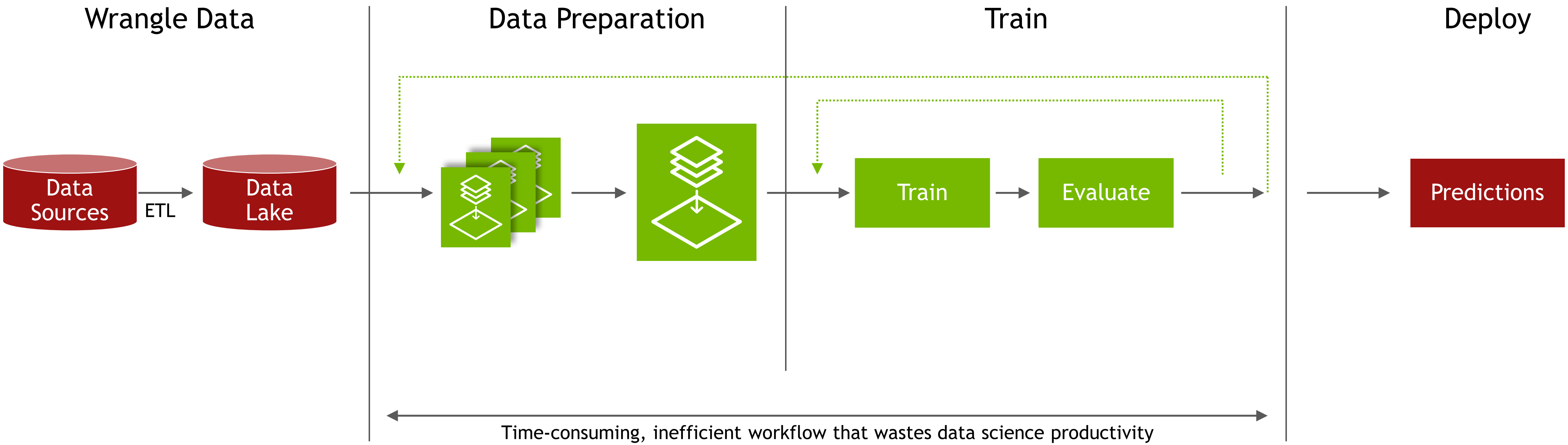
Marketing Optimization

Personalized Recommendations

Truck Routing

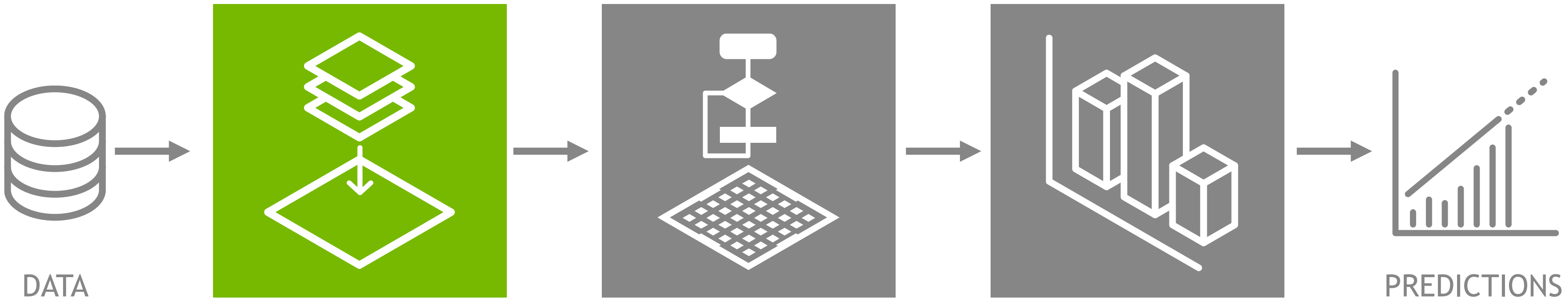
Online Delivery

ML WORKFLOW STIFLES INNOVATION



DATA SCIENCE WORKFLOW WITH RAPIDS

Open Source, End-to-end GPU-accelerated Workflow Built On CUDA



DATA PREPARATION

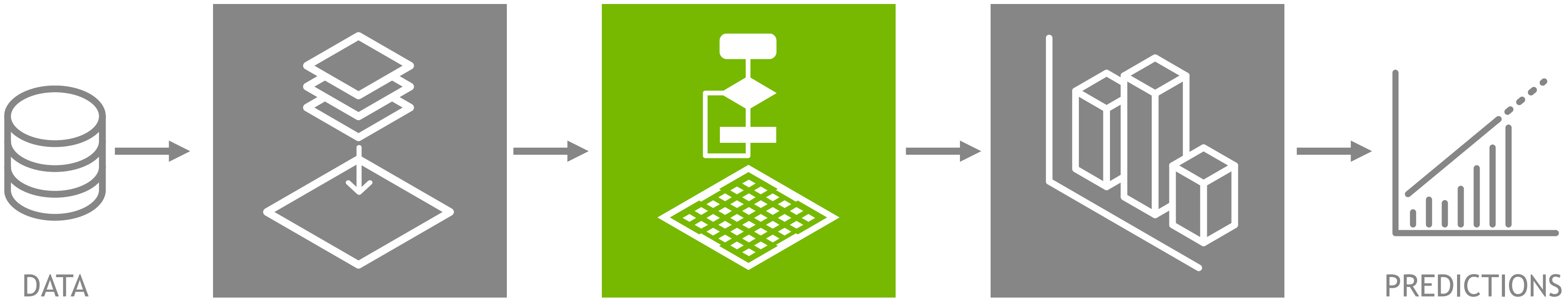
GPUs accelerated compute for in-memory data preparation

Simplified implementation using familiar data science tools

Python drop-in Pandas replacement built on CUDA C++. GPU-accelerated Spark (in development)

DATA SCIENCE WORKFLOW WITH RAPIDS

Open Source, End-to-end GPU-accelerated Workflow Built On CUDA



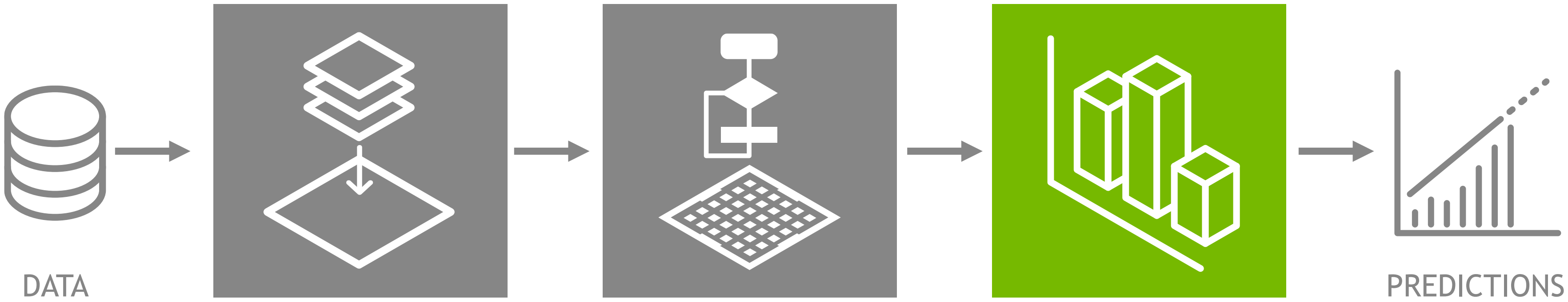
MODEL TRAINING

GPU-acceleration of today's most popular ML algorithms

XGBoost, Random Forest, Linear Regression, PCA, K-means, k-NN, DBScan, tSVD …

DATA SCIENCE WORKFLOW WITH RAPIDS

Open Source, End-to-end GPU-accelerated Workflow Built On CUDA



VISUALIZATION

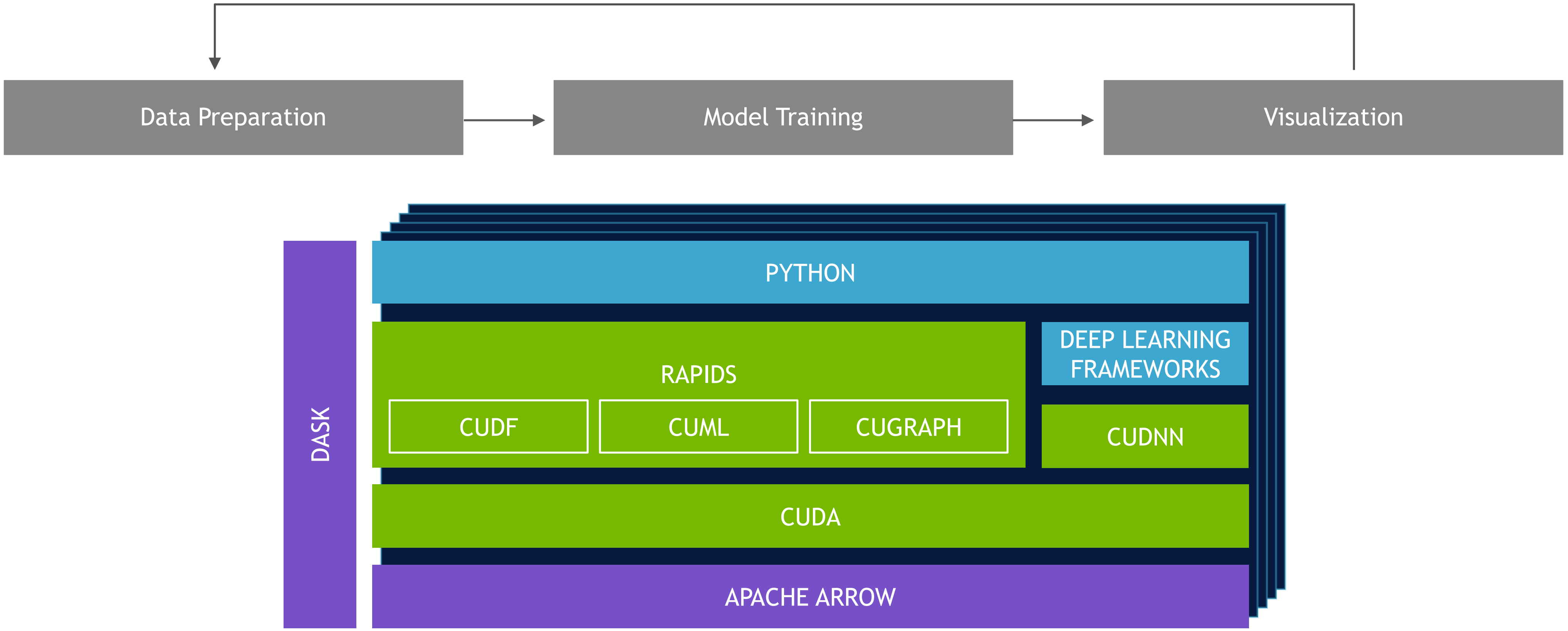
Effortless exploration of datasets, billions of records in milliseconds

Dynamic interaction with data = faster ML model development

Data visualization ecosystem (Graphistry & OmniSci), integrated with RAPIDS

RAPIDS — OPEN GPU DATA SCIENCE

Software Stack



GETTING STARTED

DLI Courses

- Accelerating Data Science Workflows with RAPIDS – https://courses.nvidia.com/courses/course-v1:DLI+L-DS-01+V1

Resources

RAPIDS GitHub – https://github.com/rapidsai

- cuDF – https://github.com/rapidsai/cudf
- cuML – https://github.com/rapidsai/cuml
- cuGraph – https://github.com/rapidsai/cugraph
- XGBoost – https://github.com/rapidsai/xgboost
- Dask cuDF – https://github.com/rapidsai/dask-cudf
- Dask cuML – https://github.com/rapidsai/dask-cuml
- Dask XGBoost – https://github.com/rapidsai/dask-xgboost
- Notebooks – https://github.com/rapidsai/notebooks
- Notebooks Extended– https://github.com/rapidsai/notebooks-extended

NVIDIA HARDWARE

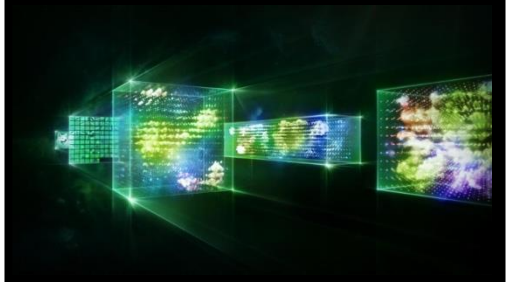
TESLA V100
TENSOR CORE GPU

World’s Most Advanced
Data Center GPU

5,120 CUDA cores
640 Tensor cores
7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS
| 125 Tensor TFLOPS
20MB SM RF | 16MB Cache
32 GB HBM2 @ 900GB/s |
300GB/s NVLink

TENSOR CORE BUILT FOR AI

Delivering 125 TFLOPS of DL Performance



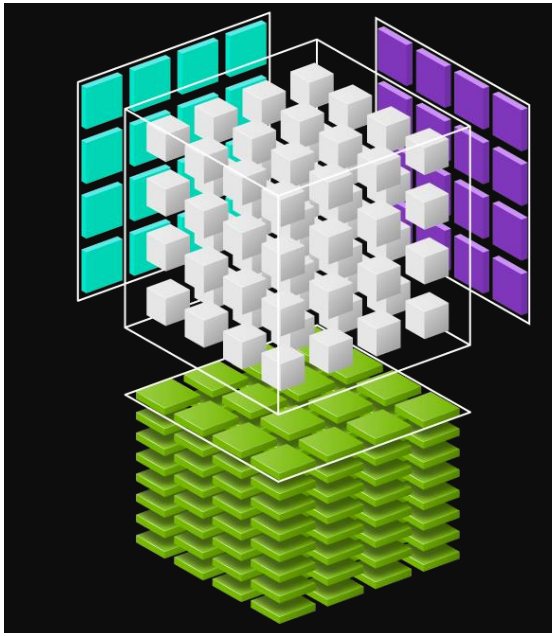
MATRIX DATA OPTIMIZATION:

Dense Matrix of Tensor Compute

TENSOR-OP CONVERSION:

FP32 to Tensor Op Data for Frameworks

VOLTA-OPTIMIZED cuDNN



VOLTA TENSOR CORE

4x4 matrix processing array

D[FP32] = A[FP16] * B[FP16] + C[FP32]

Optimized For Deep Learning

Caffe2

Microsoft Cognitive Toolkit

mxnet

PYTORCH

TensorFlow

ALL MAJOR FRAMEWORKS

NVIDIA DGX

AI Supercomputer-in-a-Box

1000 TFLOPS | 8x Tesla V100 32GB | NVLink Hybrid Cube Mesh

2x Xeon | 8 TB RAID 0 | Quad IB 100Gbps, Dual 10GbE | 3U — 3200W

NVIDIA DGX-2

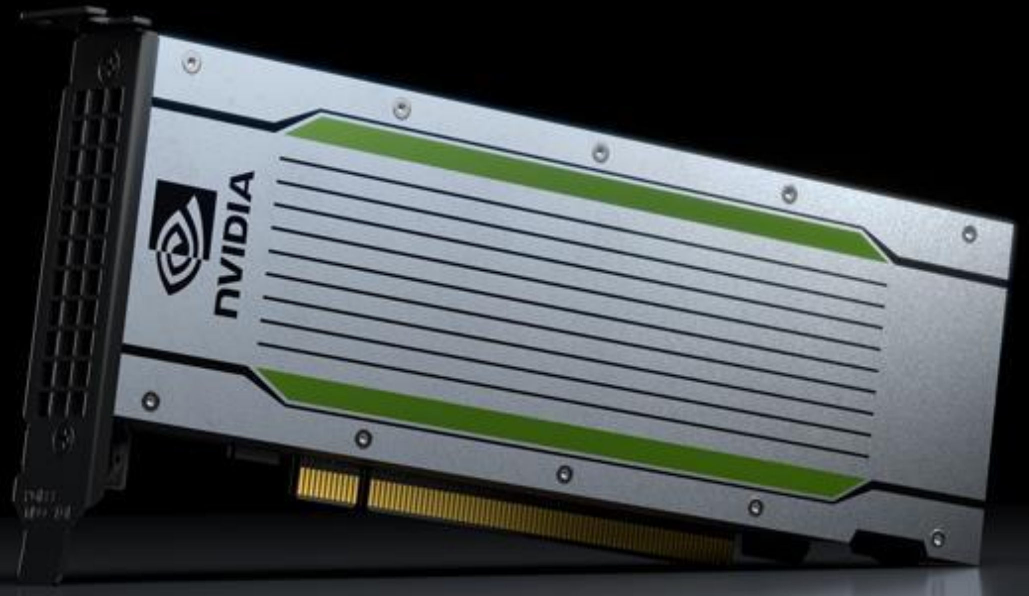
THE WORLD'S MOST POWERFUL DEEP LEARNING SYSTEM FOR THE MOST COMPLEX DEEP LEARNING CHALLENGES

- First 2 PFLOPS System
- 16 V100 32GB GPUs Fully Interconnected
- NVSwitch: 2.4 TB/s bisection bandwidth
- 24X GPU-GPU Bandwidth
- 0.5 TB of Unified GPU Memory
- 10X Deep Learning Performance

TESLA T4

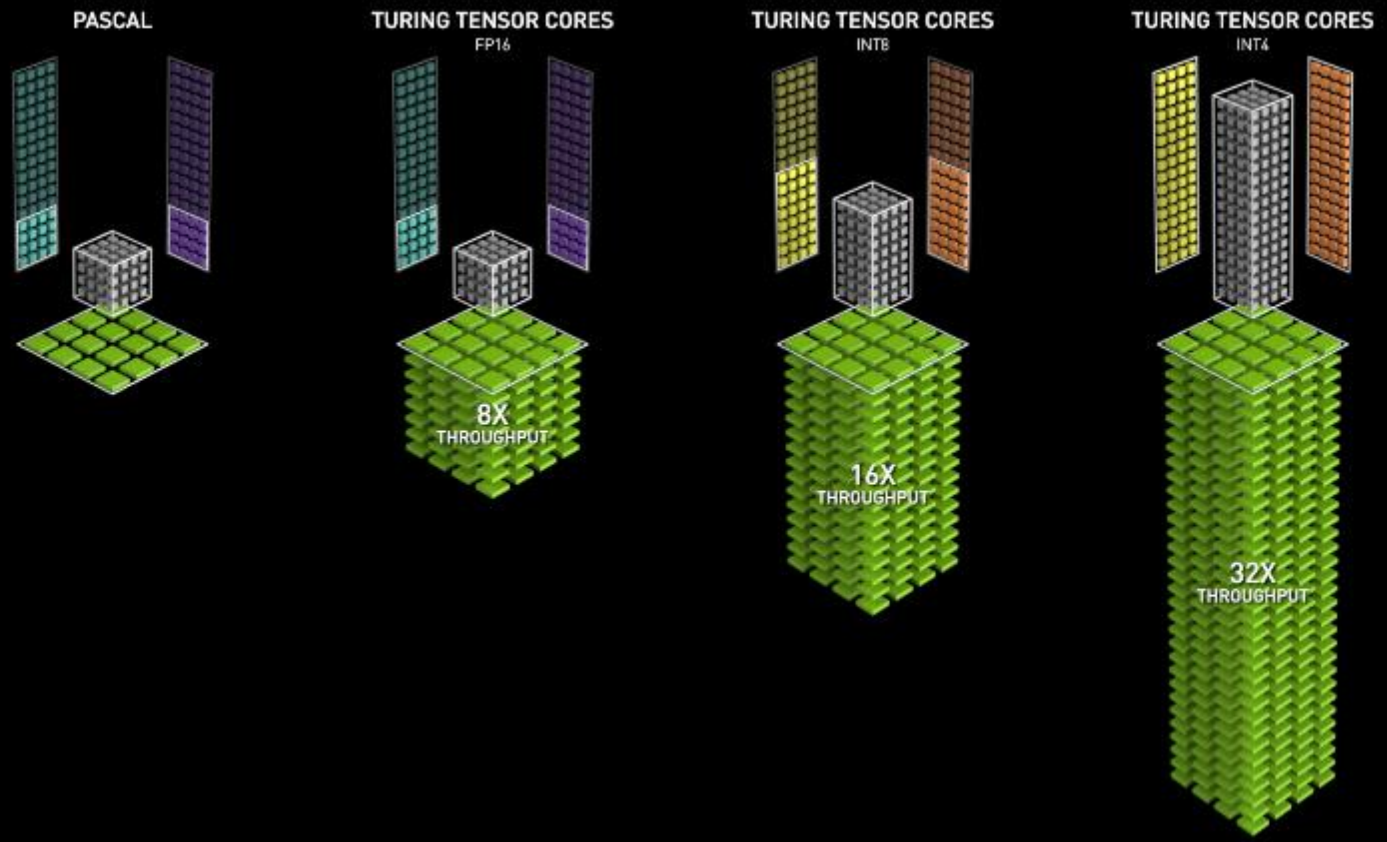
WORLD'S MOST ADVANCED SCALE-OUT GPU

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



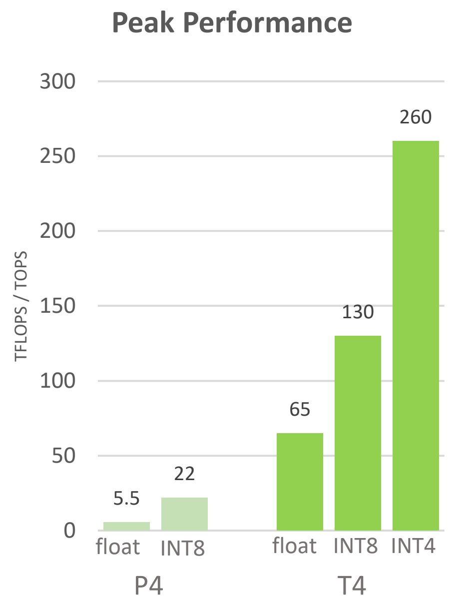
NEW TURING TENSOR CORE

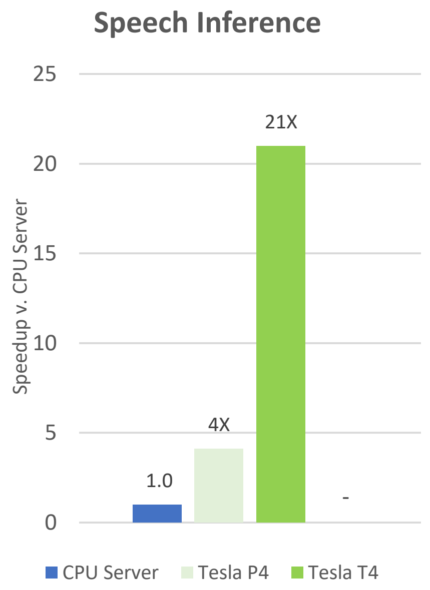
MULTI-PRECISION FOR AI INFERENCE & ENTRY LEVEL TRAINING
65 TFLOPS FP16 | 130 TeraOPS INT8 | 260 TeraOPS INT4



WORLD’S MOST PERFORMANT INFERENCE PLATFORM

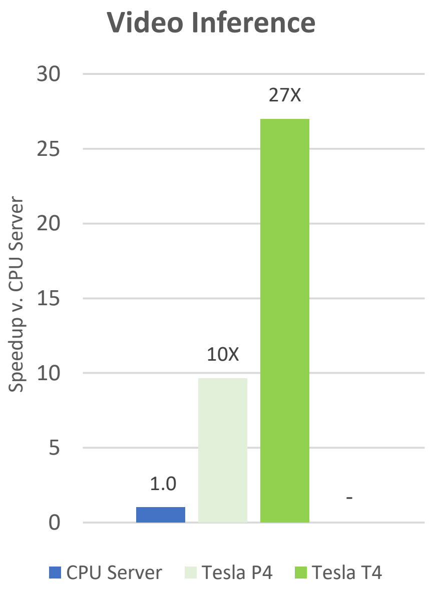
Up To 36X Faster Than CPUs | Accelerates All AI Workloads





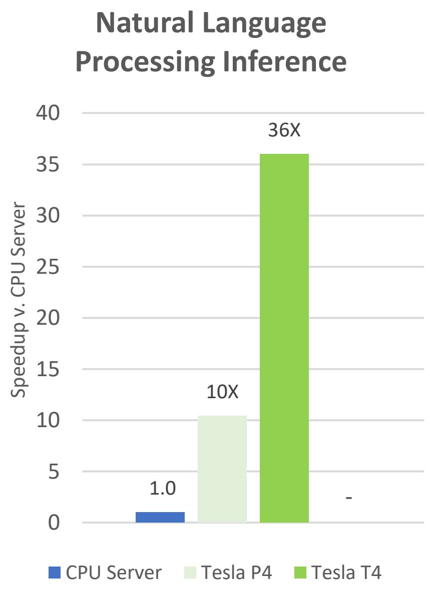
Speedup: 21X faster

DeepSpeech 2



Speedup: 27x faster

ResNet-50 (7ms latency limit)



Speedup: 36x faster

GNMT

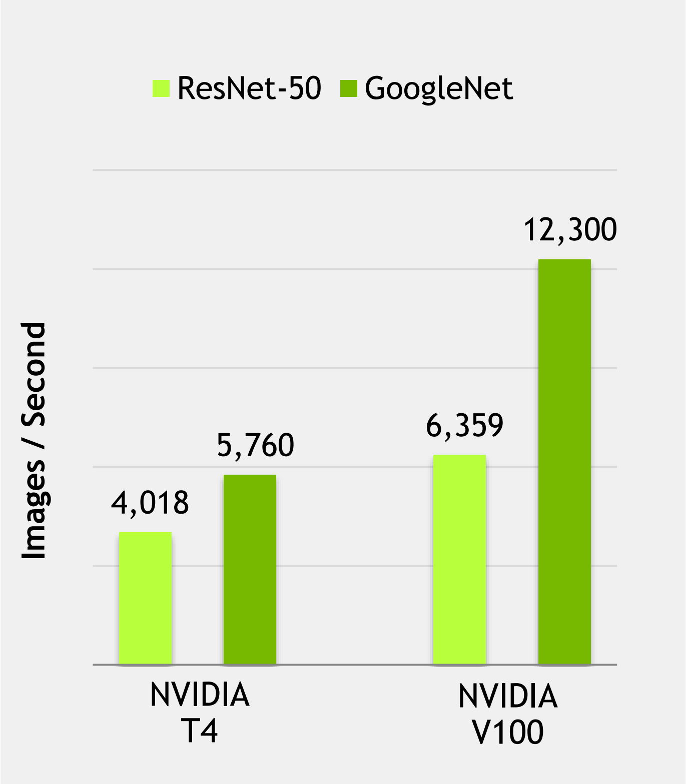
For all three graphs:
Dual-Socket Xeon Gold 6140 @ 3.6GHz with single GPU as shown 18.11-py3 | TensorRT 5.0 | CPU FP32, P4 & T4: INT8 | Batch Size = 128

NVIDIA.

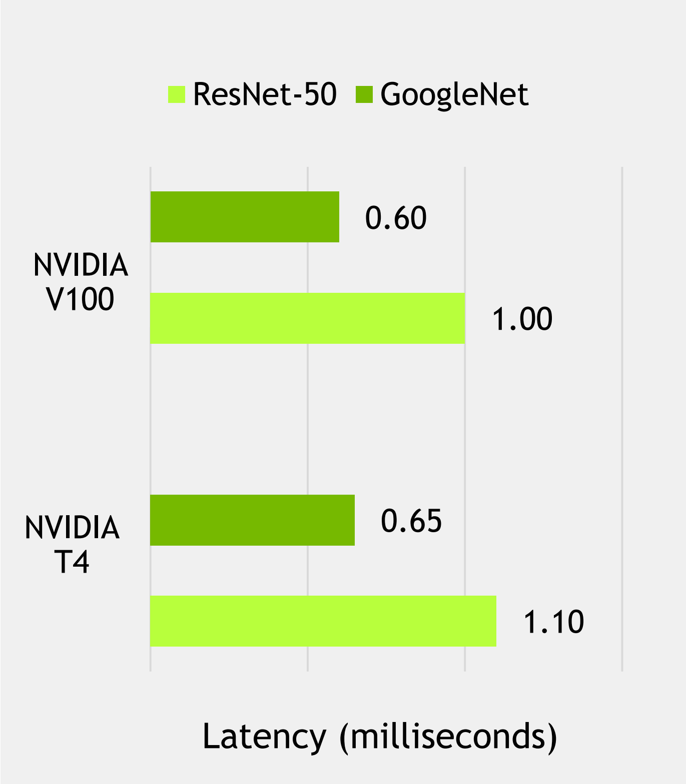
WORLD'S FASTEST INFERENCE PERFORMANCE

NVIDIA GPUs Set New Performance Records

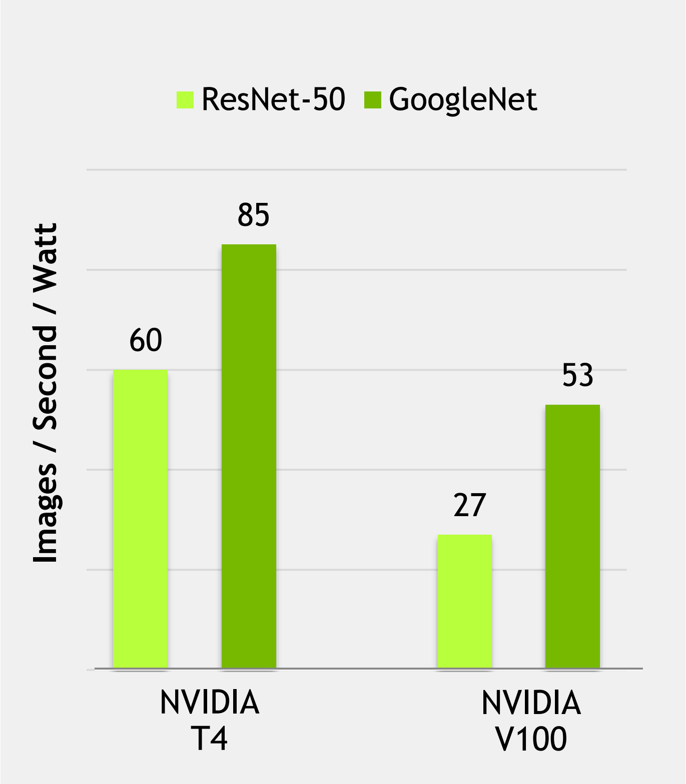
Throughput



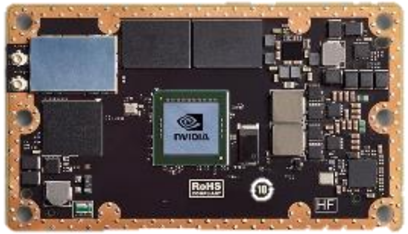
Latency



Energy Efficiency



THE JETSON FAMILY



JETSON TX1
7 - 15W
1 TFOPS (FP16)
50mm x 87mm

JETSON TX2
7 – 15W
1.3 TOPS (FP16)
50mm x 87mm

JETSON AGX XAVIER
10 – 30W
10 TFLOPS (FP16) | 32 TOPS (INT8)
100mm x 87mm

AI at the edge

UAVs • AI subsystems • AI Cameras

Fully autonomous machines

Factory automation • Logistics • Delivery robots

Multiple devices • Unified software

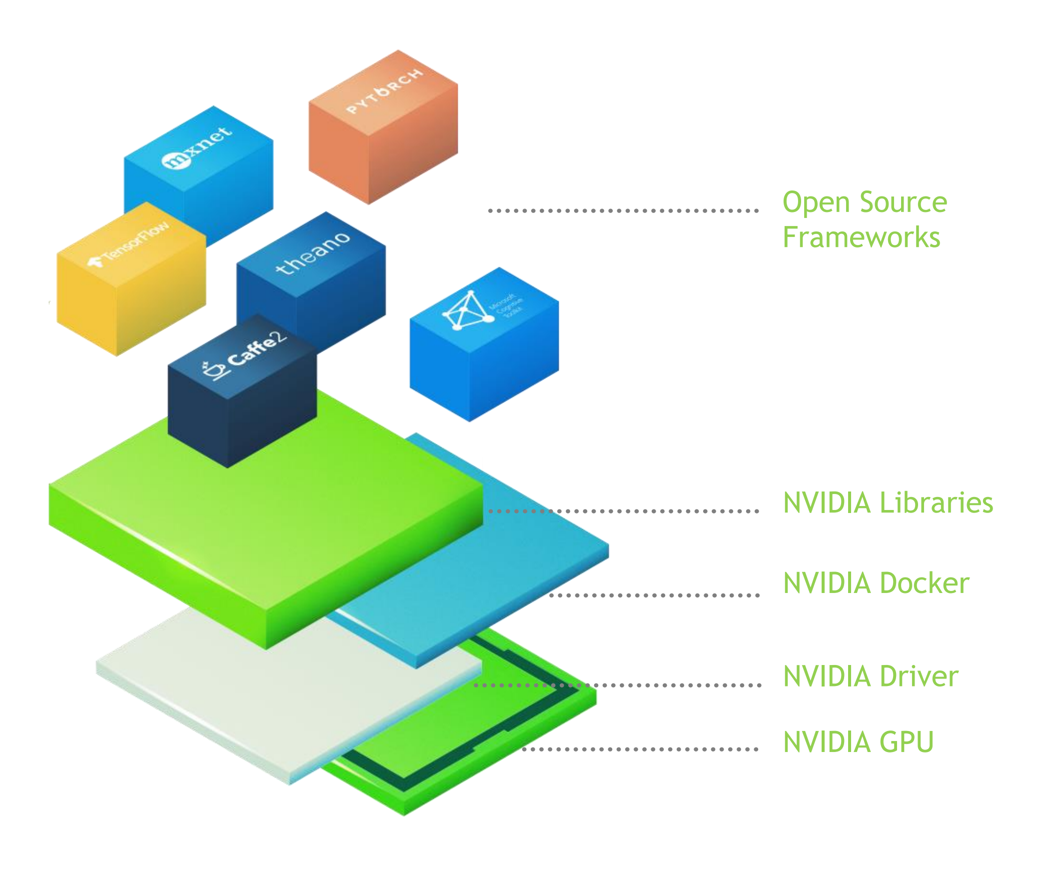
NVIDIA SOFTWARE

CHALLENGES WITH DEEP LEARNING

Current DIY deep learning environments are **complex** and **time consuming** to build, test and maintain

Requires high level of **expertise** to manage driver, library, framework dependencies

Development of frameworks by the community is moving **very fast**



NVIDIA GPU CLOUD

Deep Learning Everywhere, For Everyone

Innovate in minutes, not weeks
Removes all the DIY complexity of deep learning software integration

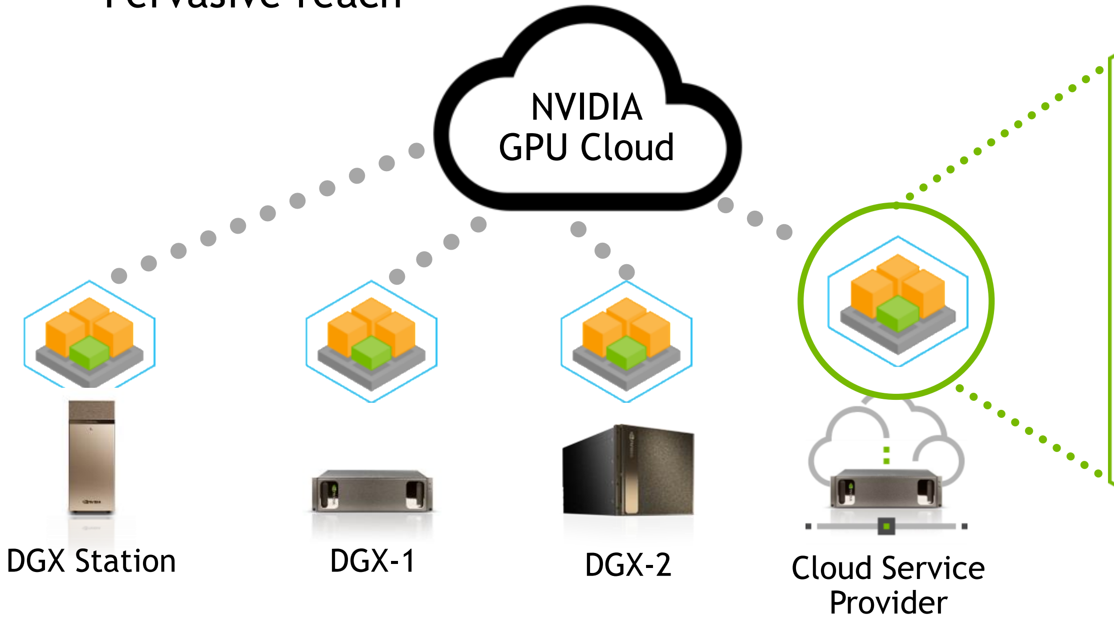
Always up to date
Monthly updates by NVIDIA to ensure maximum performance

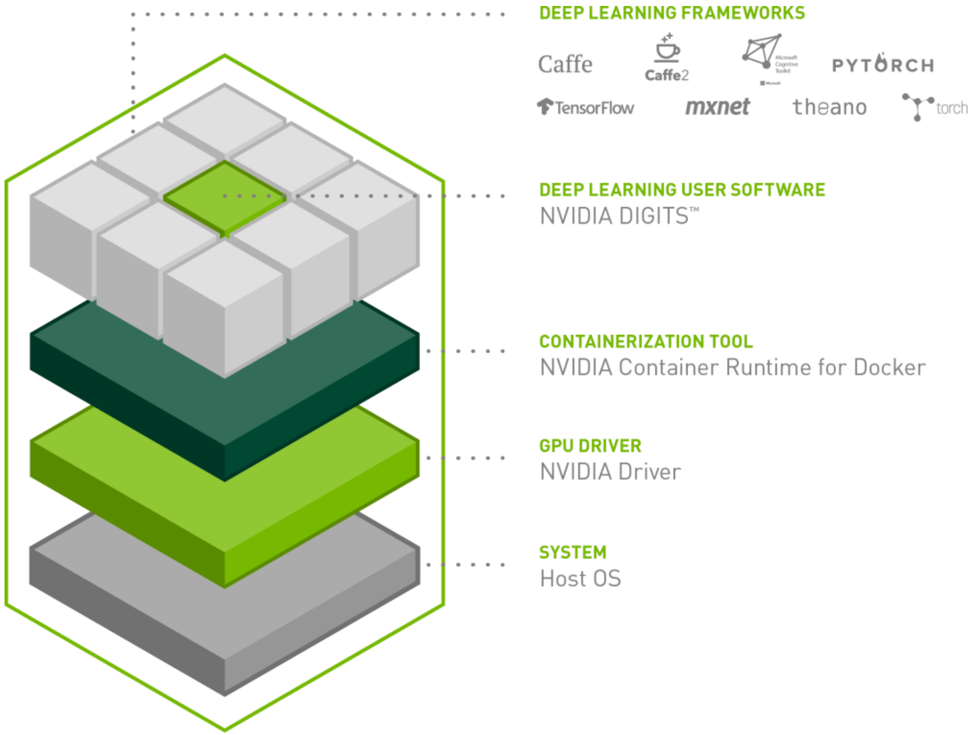
Deep learning across platforms
Containers run locally on DGX Systems and TITAN PCs, or on cloud service provider GPU instances

NVIDIA GPU Cloud integrates GPU-optimized deep learning frameworks, runtimes, libraries, and OS into a ready-to-run container, available at no charge

COMMON SOFTWARE STACK ACROSS DGX FAMILY

- Single, unified stack for deep learning frameworks
- Predictable execution across platforms
- Pervasive reach

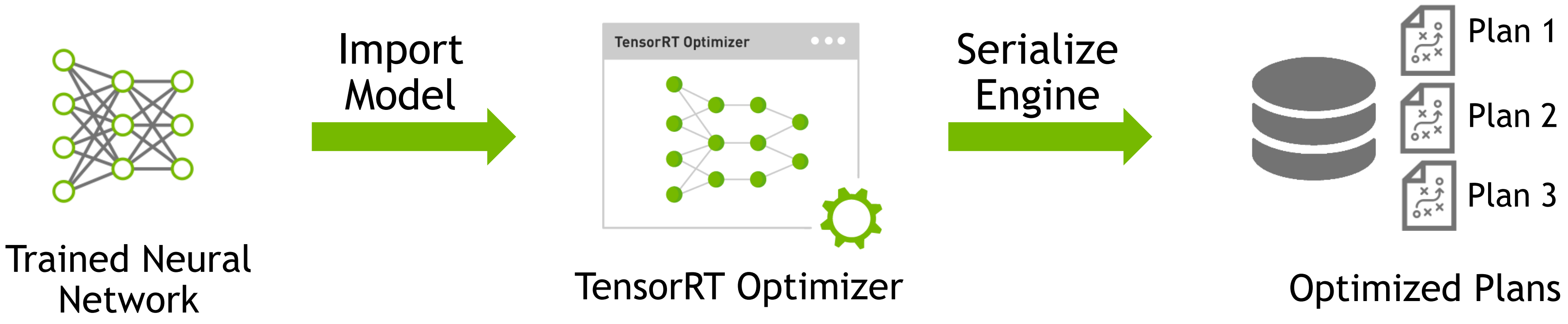




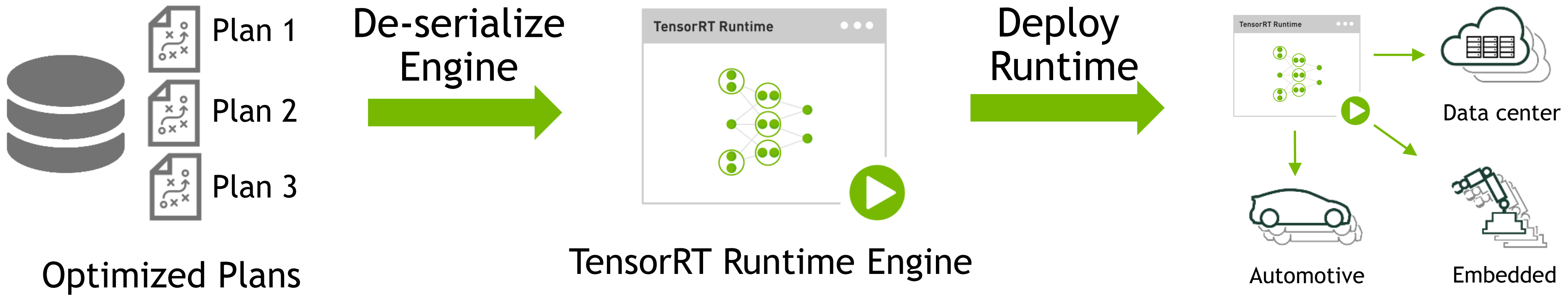
TENSORRT

TENSORRT DEPLOYMENT WORKFLOW

Step 1: Optimize trained model

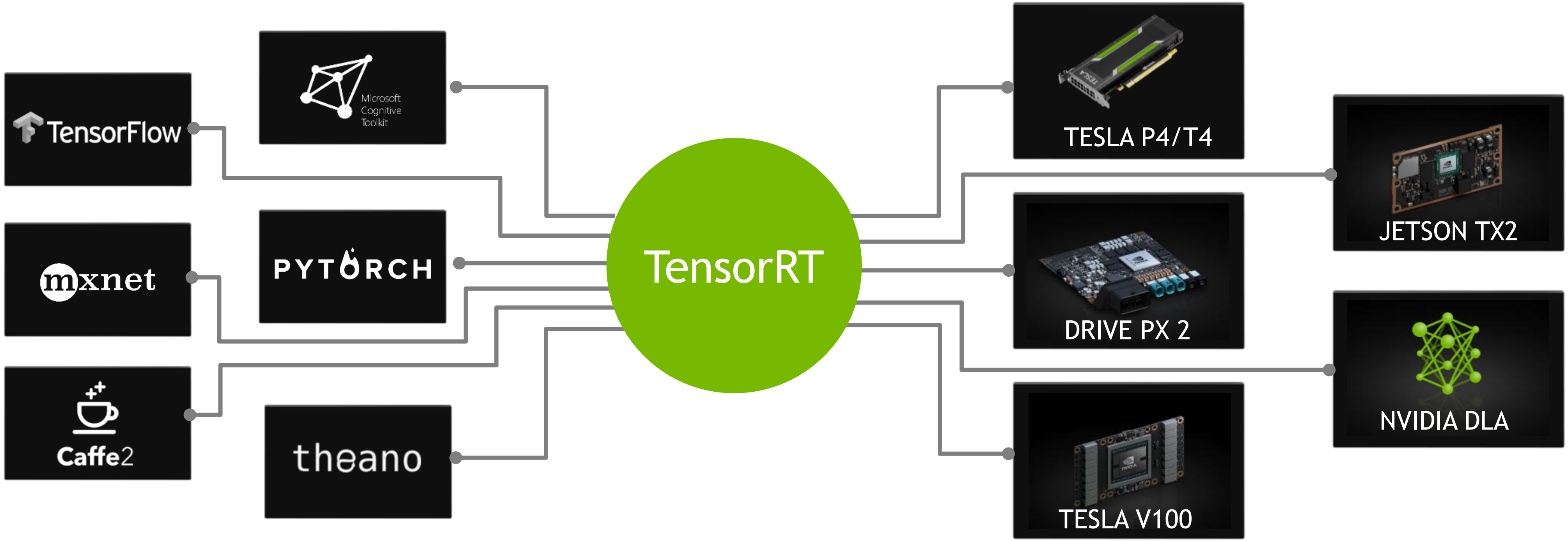


Step 2: Deploy optimized plans with runtime



NVIDIA TENSORRT

From Every Framework, Optimized For Each Target Platform



Frameworks

Platforms

TensorRT 5 & TensorRT Inference Server

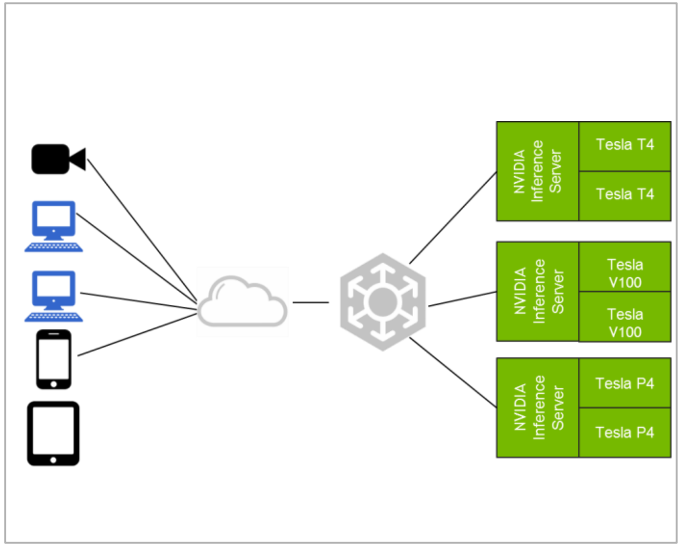
Turing Support • Optimizations & APIs • Inference Server

World's Most Advanced Inference Accelerator

Up to 40x faster perf. on Turing Tensor Cores

New optimizations & flexible INT8 APIs

New INT8 workflows, Win & CentOS support



TensorRT inference server

Maximize GPU utilization, run multiple models on a node

Free download to members of NVIDIA Developer Program soon at

developer.nvidia.com/tensorrt

TensorRT Inference Server

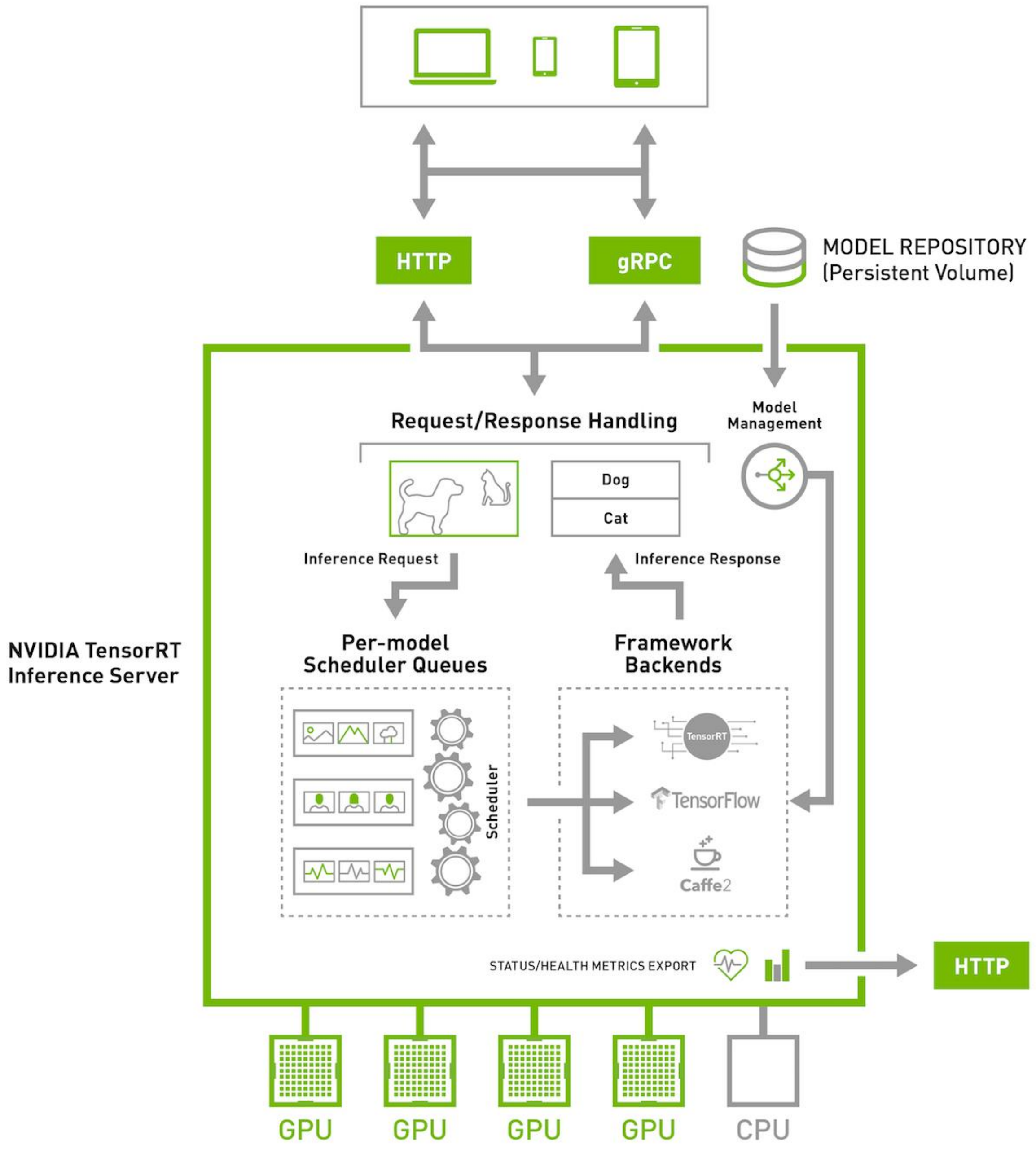
Containerized Microservice for Data Center Inference

Multiple models scalable across GPUs

Supports all popular AI frameworks

Seamless integration into DevOps deployments leveraging Docker and Kubernetes

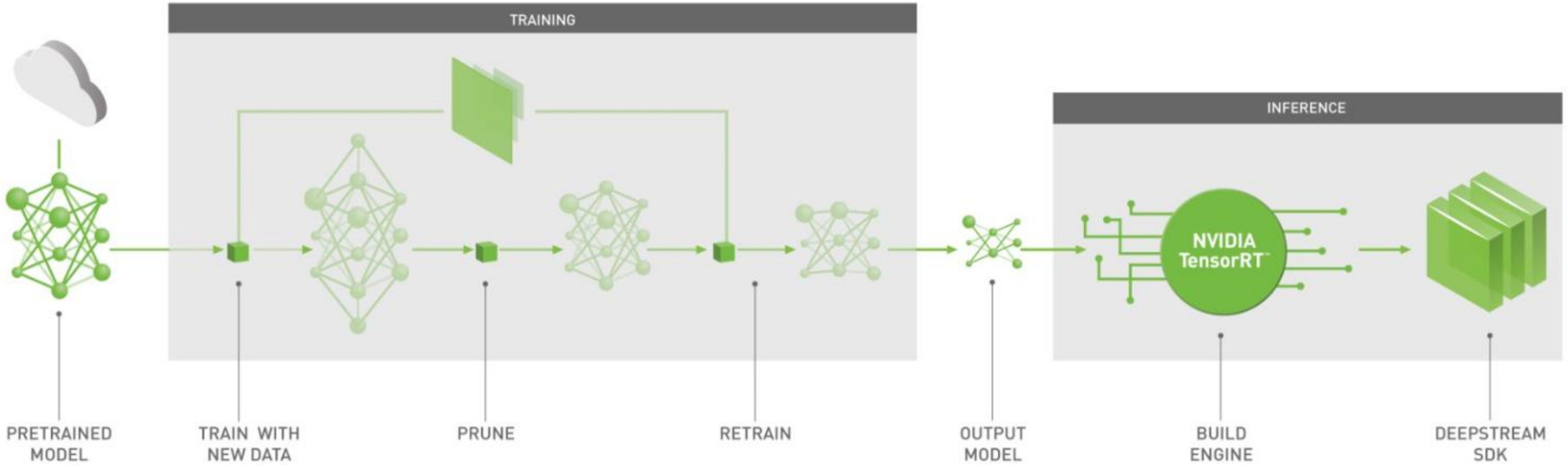
Ready-to-run container, free from the NGC container registry



TRANSFER LEARNING TOOLKIT

End to End NVIDIA Deep Learning Workflow

Pre-Trained model access from NGC * Training & adaptation * Applications ready to integrate with DeepStream

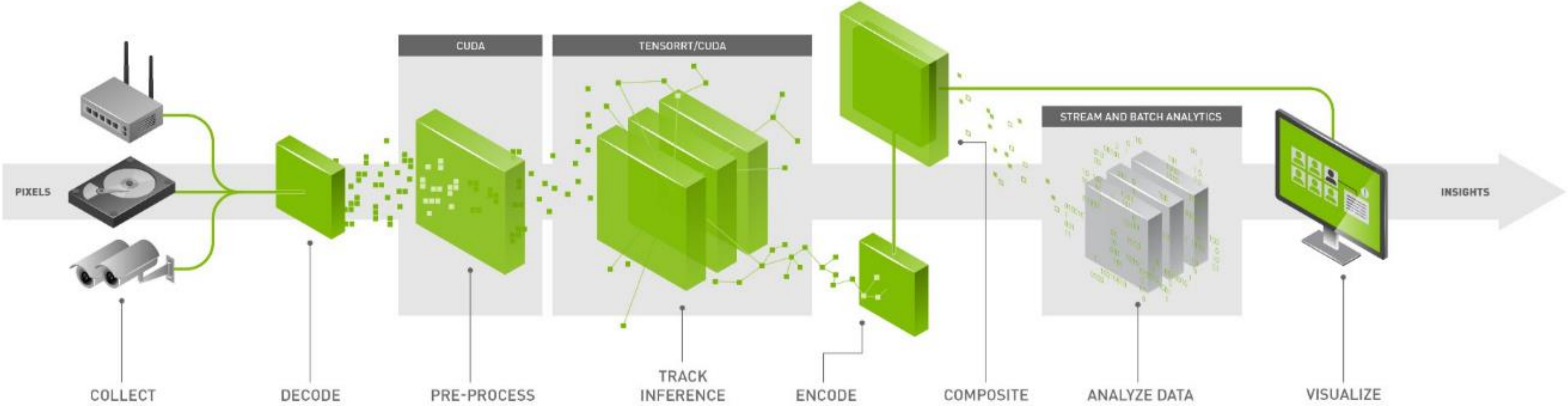


Accelerate time to market and save on compute resources!

DEEPSTREAM

NVIDIA DEEPSTREAM

Zero Memory Copies



NVIDIA