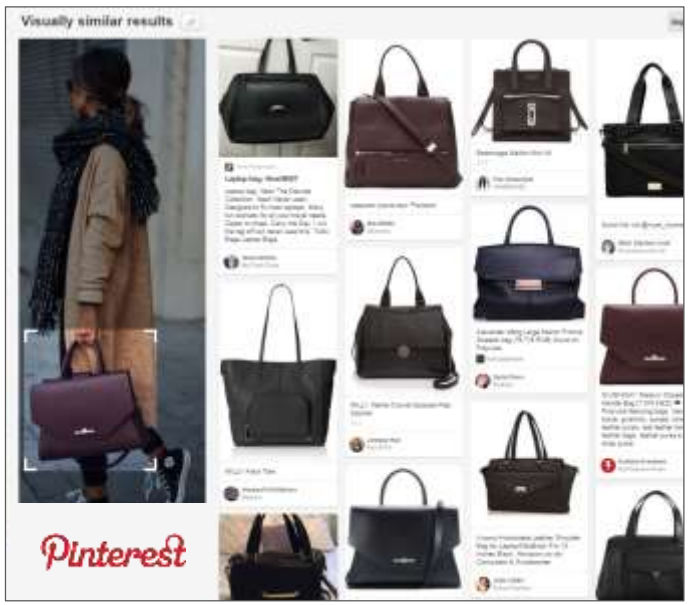
ARTIFICIAL INTELLIGENCE

27th May 2019

Prathu Tiwari | Senior Solutions Architect

AI IS EVERYWHERE

"Find where I parked my car"



"Find the bag I just saw in this magazine"

"What movie should I watch next?"

TOUCHING OUR LIVES

Bringing grandmother closer to family by bridging language barrier

Predicting sick baby's vitals like heart rate, blood pressure, survival rate

Enabling the blind to "see" their surrounding, read emotions on faces

AI FOR PUBLIC GOOD

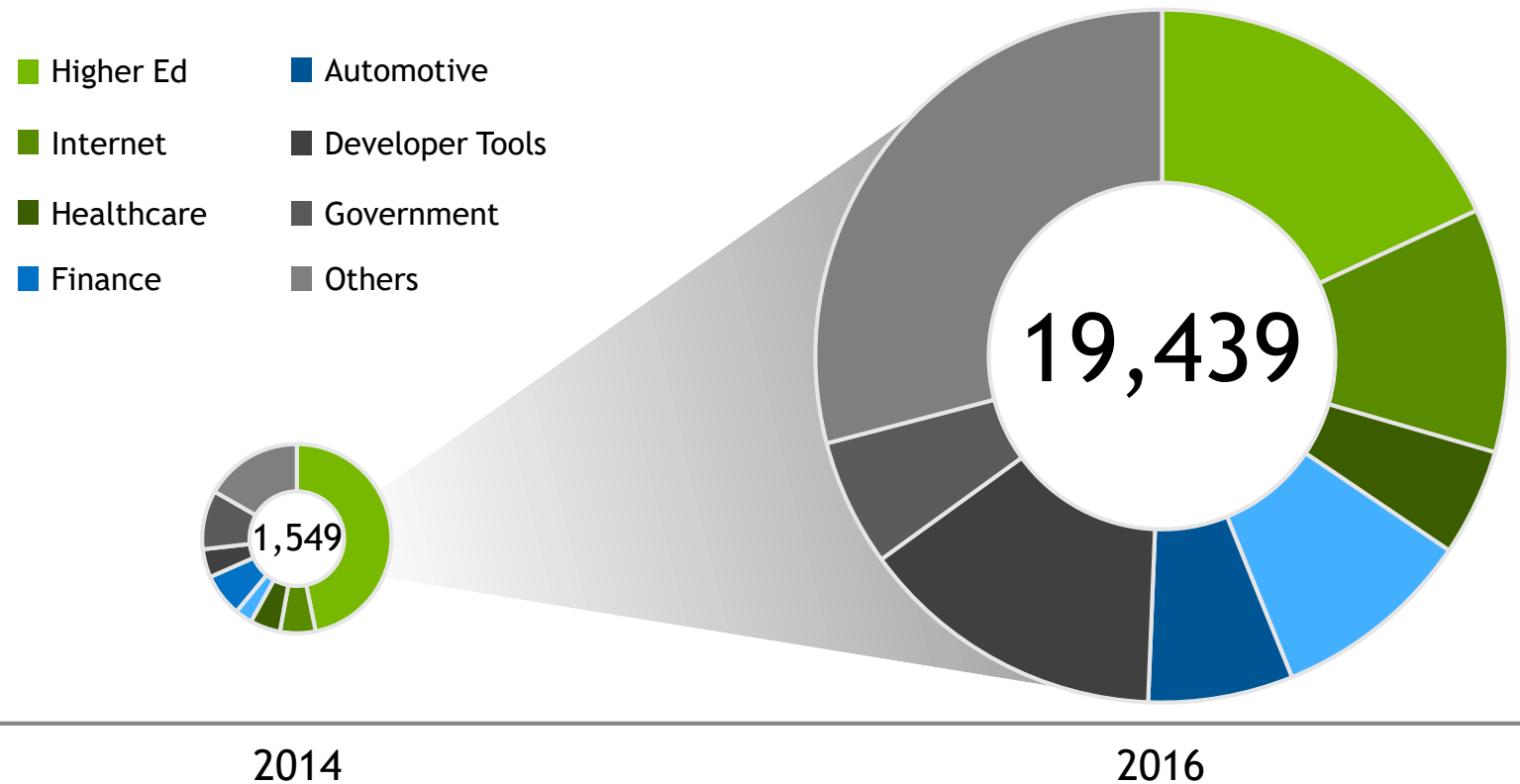
Increasing public safety with smart video surveillance at airports & malls

Providing intelligent services in hotels, banks and stores

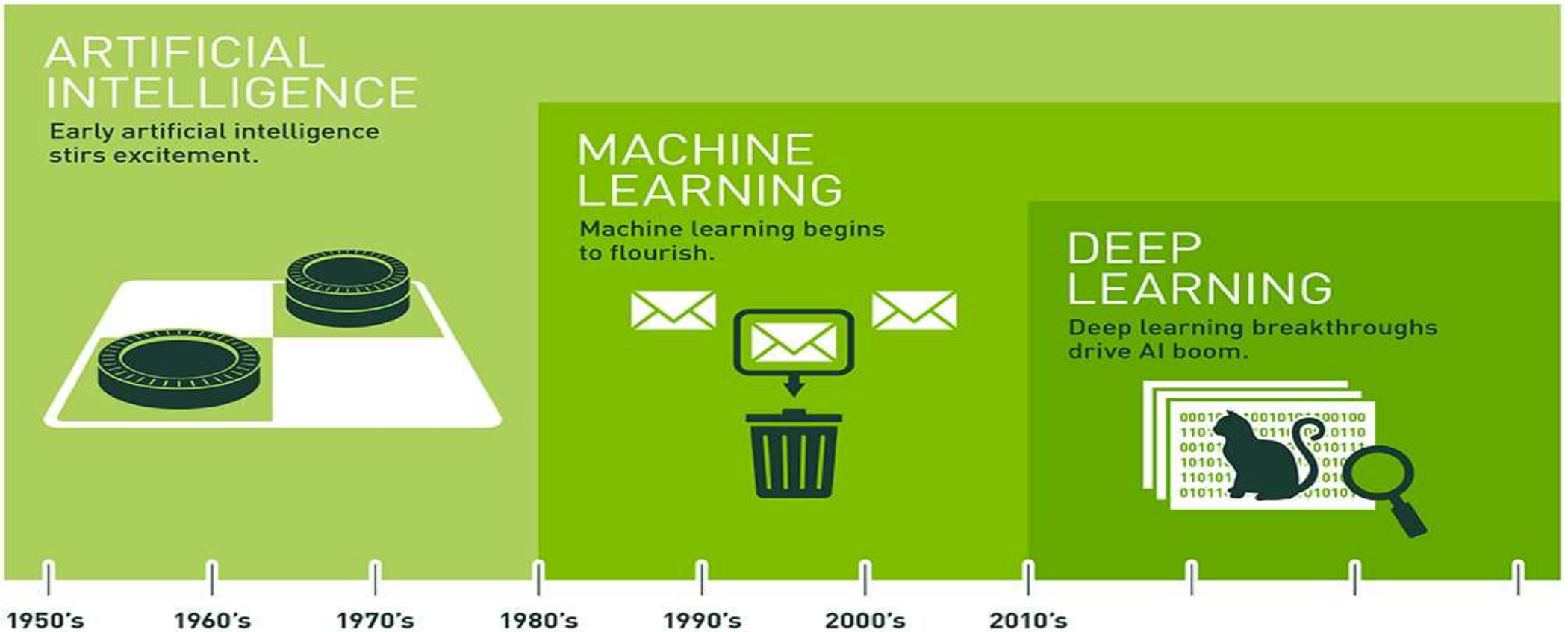
Separating weeds as it harvests, reduces chemical usage by 90%

EVERY INDUSTRY HAS AWOKEN TO AI

Organizations engaged with NVIDIA on Deep Learning

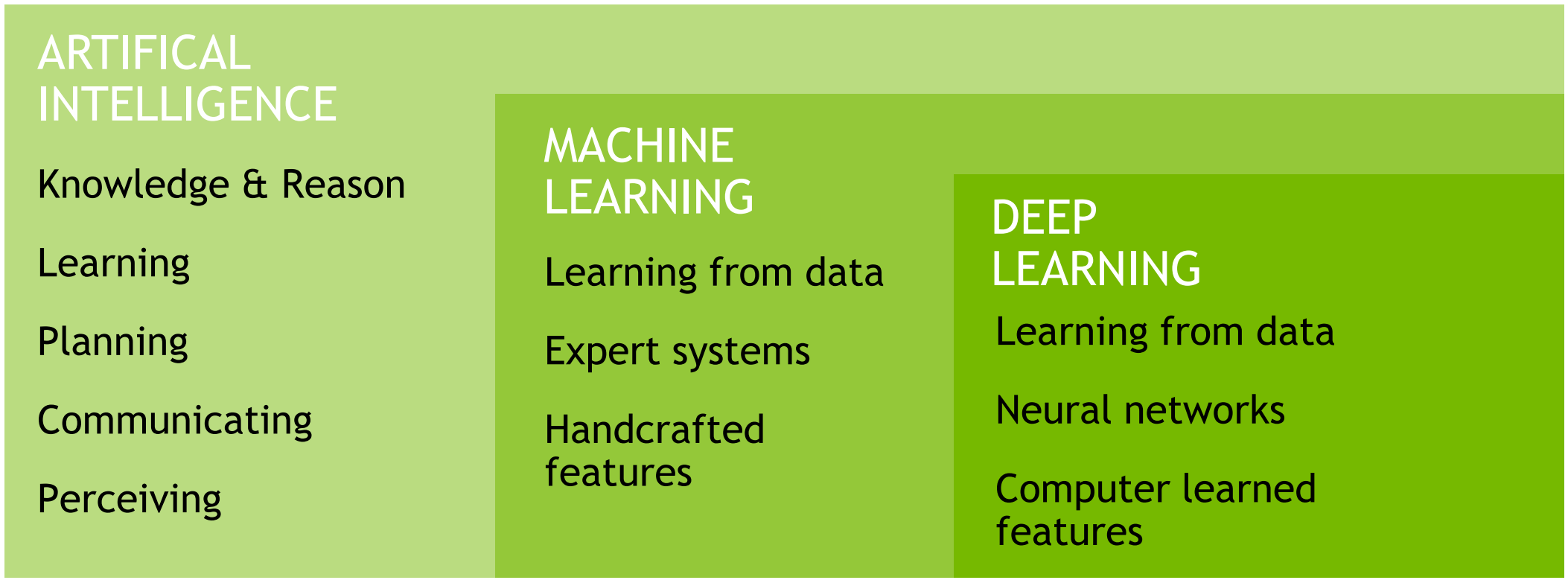


DEFINITIONS



LEARNING FROM DATA

AND SOME BUZZ WORDS



KEY DRIVERS

Big Data Availability

facebook — 350 millions images uploaded per day

Walmart — 2.5 Petabytes of customer data hourly

YouTube — 300 hours of video uploaded every minute

New ML Techniques

GPU Acceleration

SIMPLE REGRESSION

- Let us consider : we want to predict land-plot value

- What all parameters you think would effect the value

- Area?

- Location?

SIMPLE REGRESSION

- Too few parameters -> poor predictions
- Biasing or underfitting
- Increasing amount of Data does it help beyond a limit?

HOW DO WE GO ABOUT IT

- Feed forward calculate loss/cost function
- Adjust derived parameters so that they reduce loss – back prop
- Repeatedly going through this cycle would lead to a good fit line.

SUPERVISED AND UNSUPERVISED ML

- Supervized when when labeled or outcome valued data is available

- We try to predict the lable or outcome-value

- If the lable or outcome value is not available we try to get information about data itself.

SUPERVISED AND UNSUPERVISED ML

Supervised

- Labeled Data available
- Predict label of new data
- Linear regression, classification, SVM

Unsupervised

- Unlabeled data
- Get more info about data
- Clustering, Anomaly detection

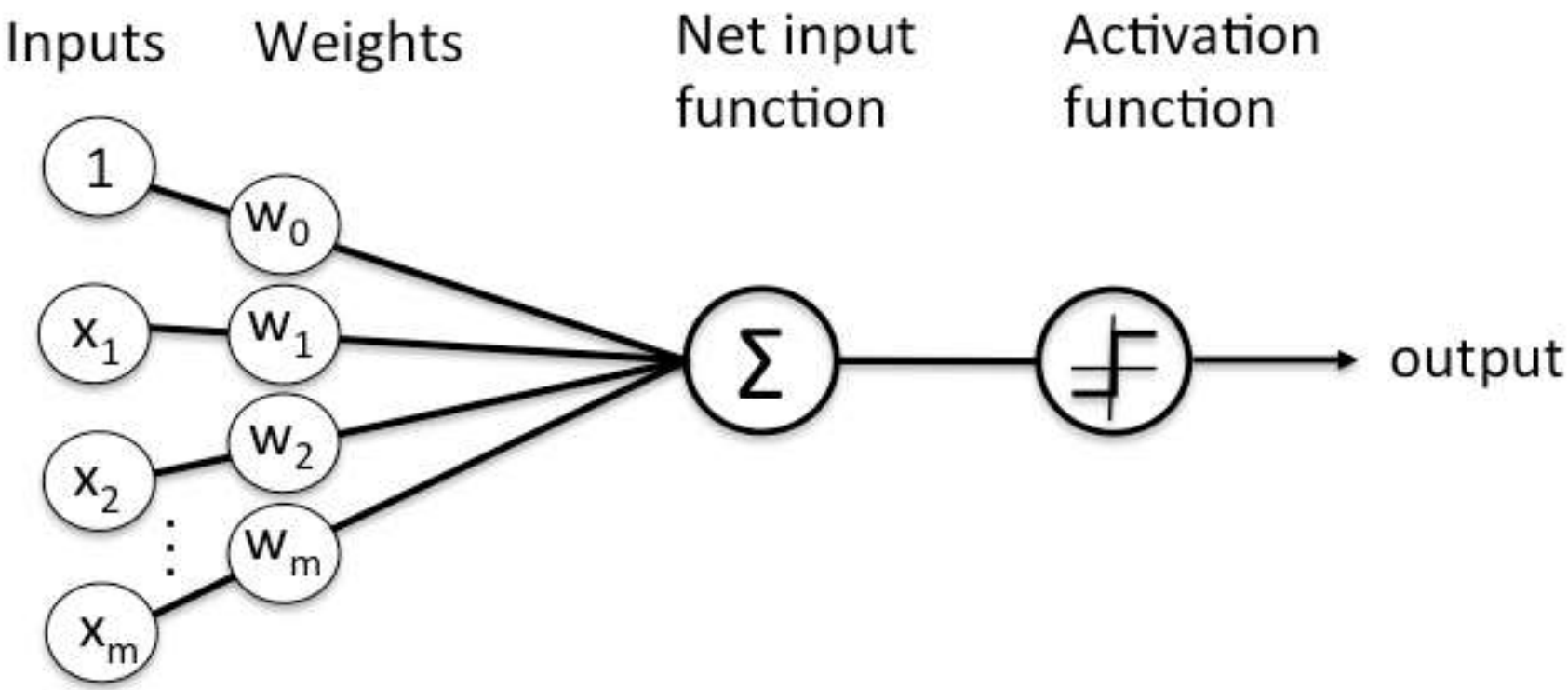
ARTIFICIAL NEURAL NETWORKS

- An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain.
- Each connection, like the synapses in a biological brain, can transmit a signal from one artificial neuron to another.
- An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it.

ARTIFICIAL NEURAL NETWORKS

- In common ANN implementations, the signal at a connection between artificial neurons is a , and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs.
- The connections between artificial neurons are called 'edges’.
- Artificial neurons and edges typically have a weight that adjusts as learning proceeds.
- The weight increases or decreases the strength of the signal at a connection.

NEURAL NETWORK COMPONANTS



A NEW COMPUTING MODEL

Algorithms that learn from examples

MACHINE LEARNING

TRADITIONAL APPROACH

Requires domain experts

Time-consuming experimentation

Custom algorithms

Not scalable to new problems

Car

Vehicle

Coupe

A NEW COMPUTING MODEL

Algorithms that learn from examples

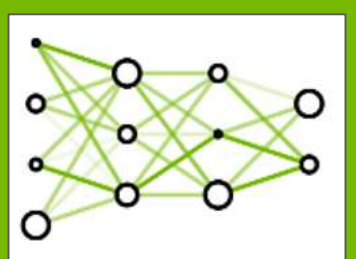
MACHINE LEARNING

TRADITIONAL APPROACH

Requires domain experts
Time-consuming experimentation
Custom algorithms
Not scalable to new problems

Car
Vehicle
Coupe

DEEP LEARNING



DEEP NEURAL NETWORKS

Learn from data
Easily to extend
Accelerated with GPUs

Car
Vehicle
Coupe

WHY DEEP LEARNING?

Scale Matters

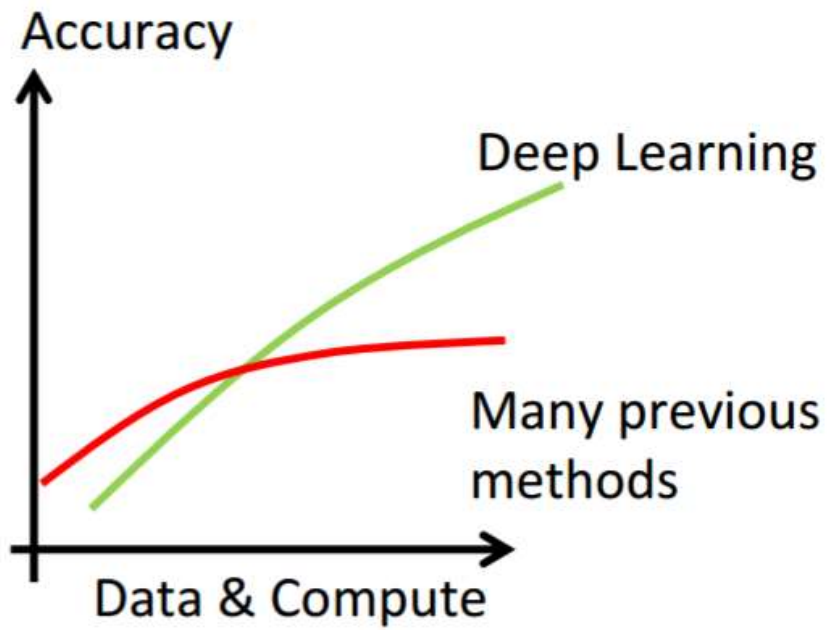
Millions to Billions of parameters

Data Matters

Regularize using more data

Productivity Matters

It’s simple, so we can make tools



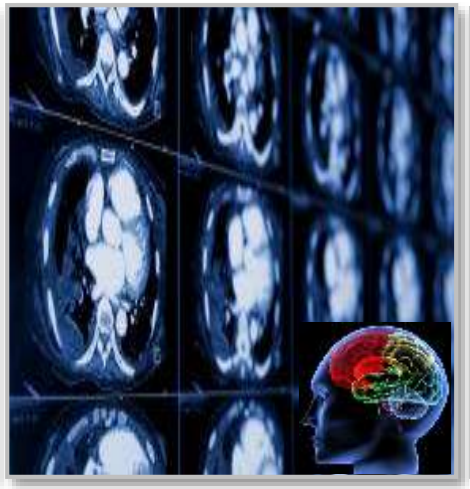
Deep Learning is most useful for large problems

DEEP LEARNING DRIVES INNOVATION

Internet Services

- Image/Video classification
- Speech recognition
- Natural language processing

Medicine



- Cancer cell detection
- Diabetic grading
- Drug discovery

Media & Entertainment

- Video captioning
- Content based search
- Real time translation

Security & Defense

- Face recognition
- Video surveillance
- Cyber security

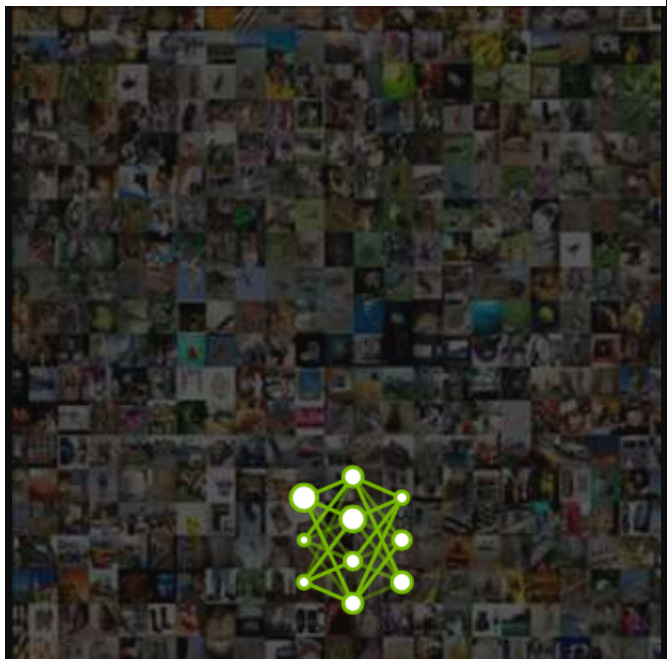
Autonomous Machines

- 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

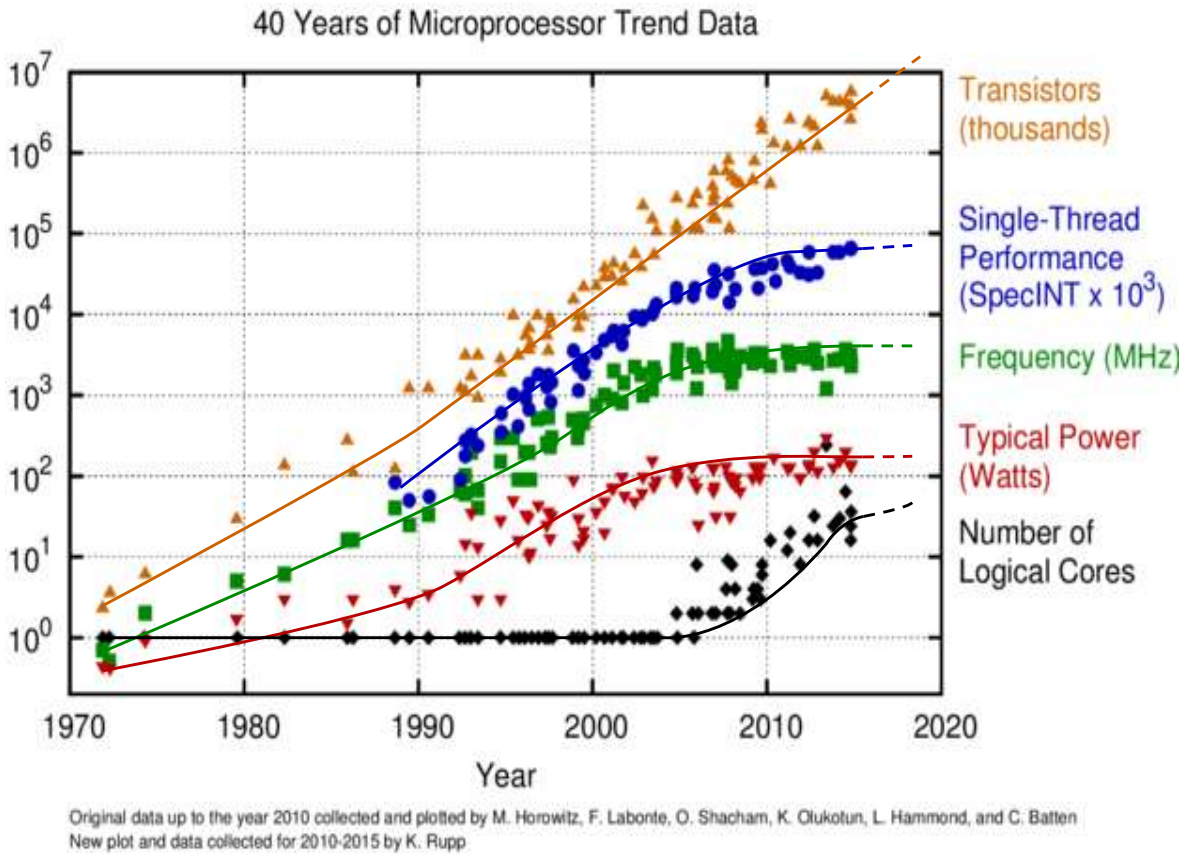
2016 – Baidu Deep Speech 2
Superhuman Voice Recognition

100 ExaFLOPS
8700 Million Parameters

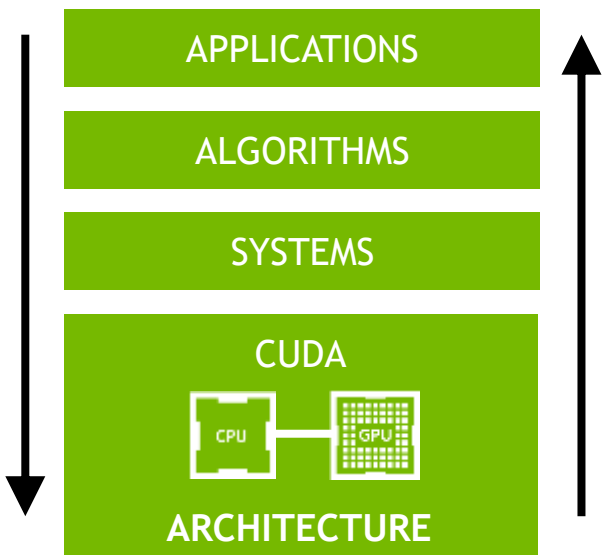
2017 – Google Neural Machine Translation
Near Human Language Translation

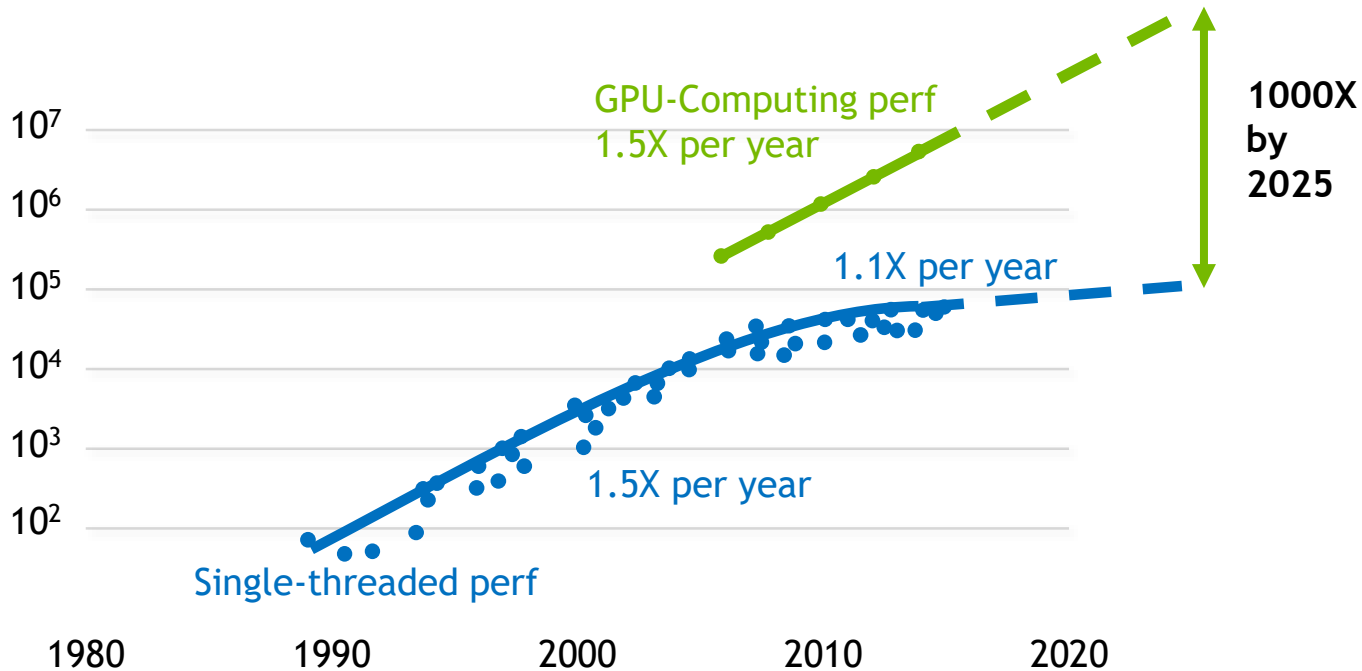
"It's time to start planning for the end of Moore's Law, and it's worth pondering how it will end, not just when."

Robert Colwell
Director, Microsystems Technology Office, DARPA



RISE OF GPU COMPUTING





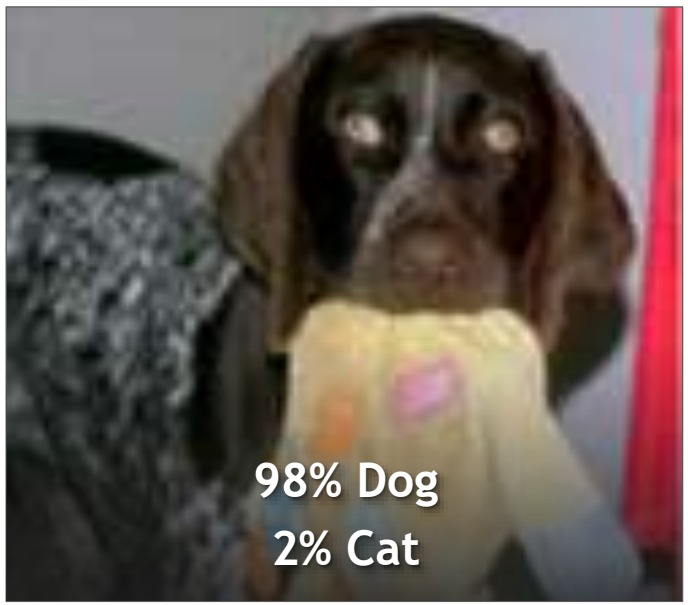
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

DEEP LEARNING REQUIREMENTS

DEEP LEARNING NEEDS…	DEEP LEARNING CHALLENGES	NVIDIA DELIVERS
Data Scientists	Demand far exceeds supply	DIGITS, DLI Training
Latest Algorithms	Rapidly evolving	DL SDK, GPU-Accelerated Frameworks
Fast Training	Impossible -> Practical	DGX-1, P100, P40, TITAN X
Deployment Platform	Must be available everywhere	TensorRT, P40, P4, Jetson, Drive PX

DEEP LEARNING WORKFLOWS

IMAGE CLASSIFICATION



Classify images into classes or categories

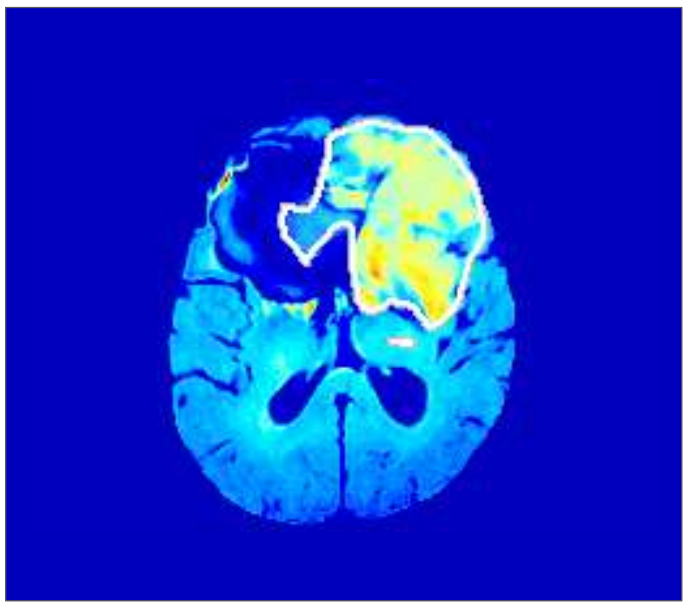
Object of interest could be anywhere in the image

OBJECT DETECTION

Find instances of objects in an image

Objects are identified with bounding boxes

IMAGE SEGMENTATION



Partition image into multiple regions

Regions are classified at the pixel level

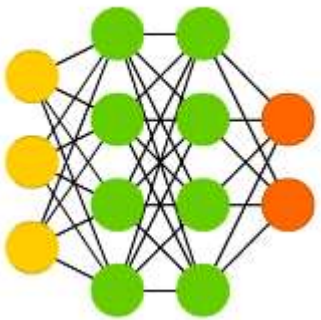
HOW DL CAN BE APPLIED

INPUTS	BUSINESS QUESTION	AI/DL TASK	EXAMPLE OUTPUTS HEALTHCARE	EXAMPLE OUTPUTS RETAIL	EXAMPLE OUTPUTS FINANCE
 Text Data  Images  Video  Audio	**Is "it" present or not?**	Detection	Cancer Detection	Targeted ads	Cybersecurity
	What type of thing is "it"?	Classification	Image Classification	Basket Analysis	Credit Scoring
	To what extent is "it" present?	Segmentation	Tumor Size/Shape Analysis	Build 360° Customer View	Credit Risk Analysis
	What is the likely outcome?	Prediction	Survivability Prediction	Sentiment & behavior recognition	Fraud Detection
	What will satisfy the objective?	Recommendations	Therapy Recommendation	Recommendation Engine	Algorithmic Trading
	What is the speaker saying?	Natural Language Processing	Expert diagnosis	Virtual personal assistants	Robo Advisors

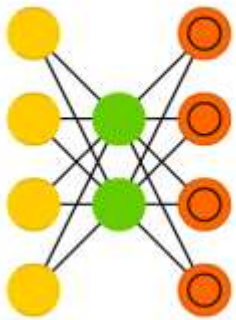
DEEP NEURAL NETWORKS

Different Models for Different Tasks

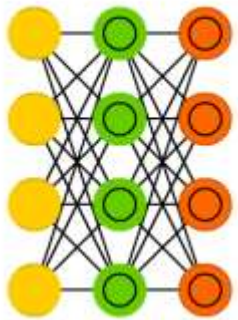
Deep Feed Forward (DFF)



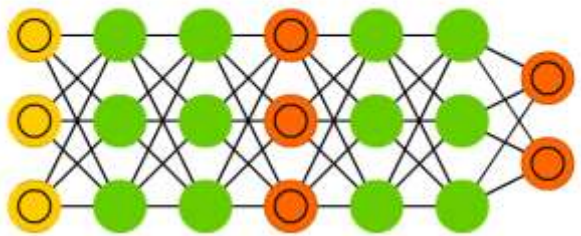
Auto Encoder (AE)



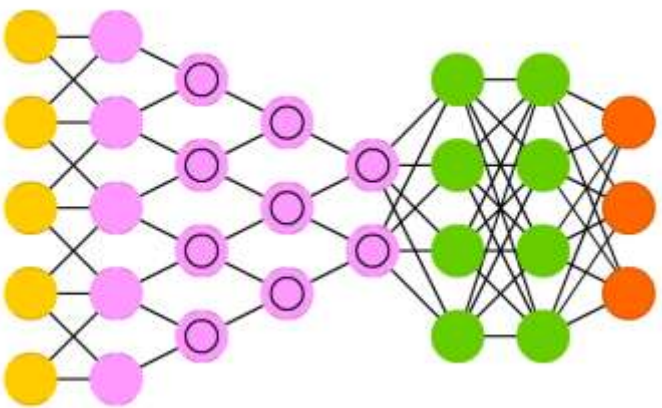
Variational AE (VAE)



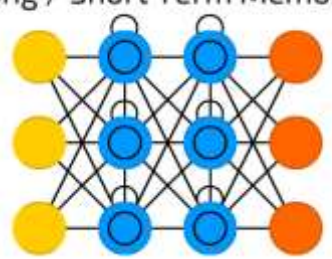
Generative Adversarial Network (GAN)



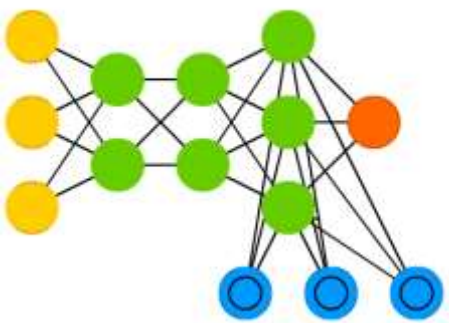
Deep Convolutional Network (DCN)



Long / Short Term Memory (LSTM)



Neural Turing Machine (NTM)



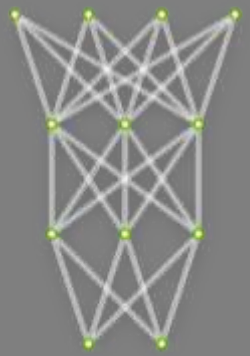
SOME KEY DECISIONS TO MAKE

FACTOR	DESCRIPTION
DL Challenge	Supervised or unsupervised, classification or regression, # of labels?
Architecture	What is the simplest architecture I can use?
Training Model	How am I going to tune my neural net? Kinds of non-linearity, loss function and weight initialization? Best training framework?
Data Quantity	How much data will be sufficient to train my model? How do I go about finding that data and is it evenly balanced?
Data Quality	Is my data directly relevant to the problem & real world data.
Data Labels	Is training data is labeled same as raw data sets, how do I 'featurize'?
Data Similarity	Is data same length vectors or does it require pre-processing?
Data Storage & Access	Where is it stored, locally and on network Data pipeline? How do I plan to extract, transform and load the data (ETL)?
Infrastructure	Cloud, On-premise, Hybrid. GPUs, CPUs or both? Single or distributed systems? Integration with languages, ent. apps/ databases.

DEEP LEARNING APPLICATION DEVELOPMENT

DEEP LEARNING APPLICATION DEVELOPMENT

Untrained
Neural Network
Model



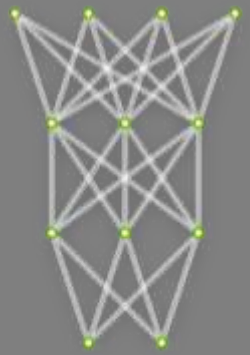
TRAINING
DATASET

DEEP LEARNING APPLICATION DEVELOPMENT

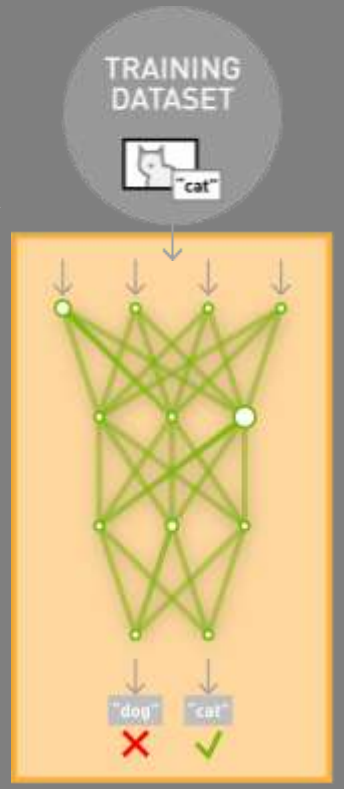
TRAINING

Learning a new capability
from existing data

Untrained
Neural Network
Model



Deep Learning
Framework

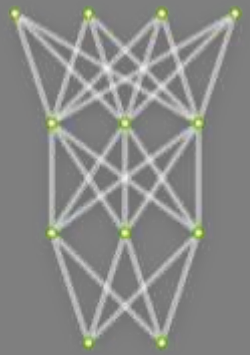


DEEP LEARNING APPLICATION DEVELOPMENT

TRAINING

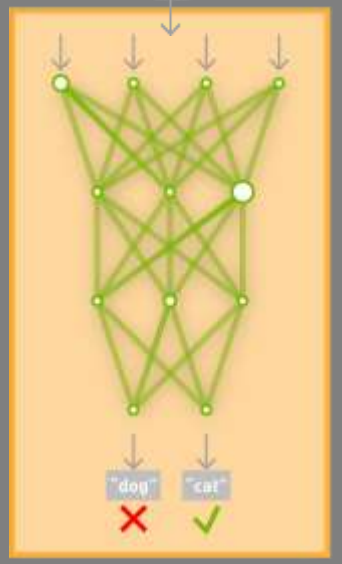
Learning a new capability
from existing data

Untrained
Neural Network
Model

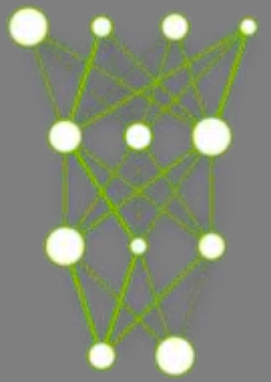


Deep Learning
Framework

TRAINING
DATASET



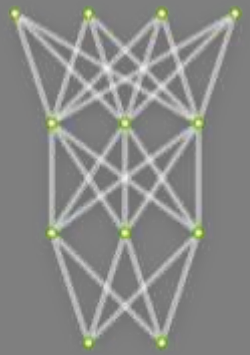
Trained Model
New Capability



DEEP LEARNING APPLICATION DEVELOPMENT

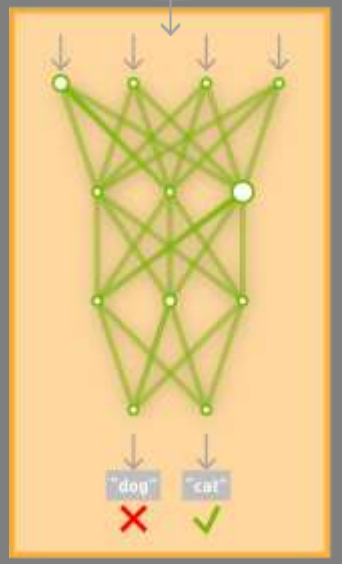
TRAINING
Learning a new capability
from existing data

Untrained
Neural Network
Model

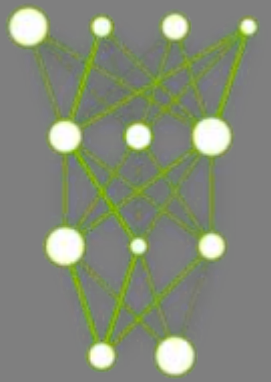


Deep Learning
Framework

TRAINING
DATASET



Trained Model
New Capability



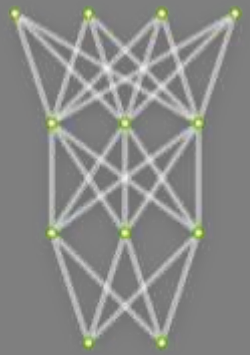
Trained Model
Optimized for
Performance

DEEP LEARNING APPLICATION DEVELOPMENT

TRAINING
Learning a new capability
from existing data

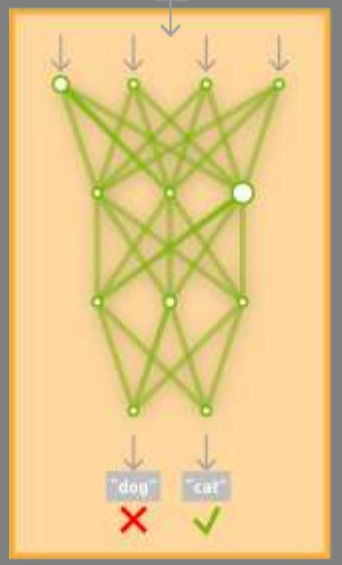
INFERENCE
Applying this capability
to new data

Untrained
Neural Network
Model

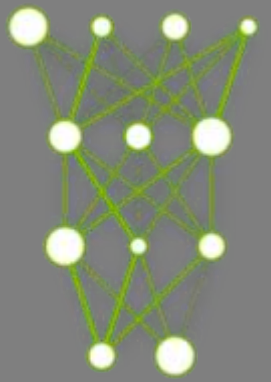


Deep Learning
Framework

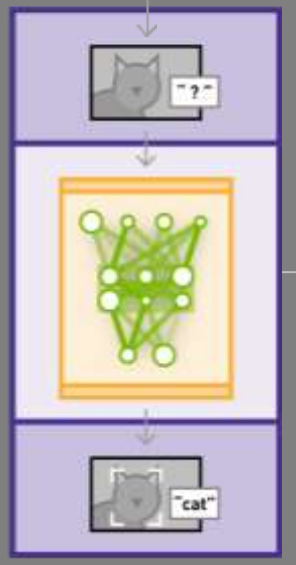
TRAINING
DATASET



Trained Model
New Capability



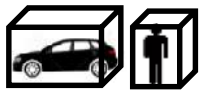
NEW
DATA



App or Service
Featuring Capability

Trained Model
Optimized for
Performance

Image Classification



Object Detection

COMPUTER VISION

Voice Recognition

Language Translation

SPEECH AND AUDIO

Recommendation Engines

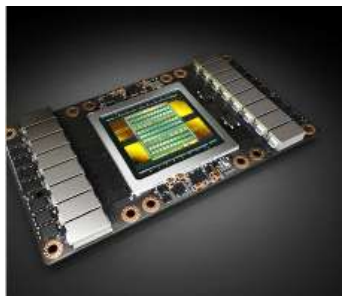
Sentiment Analysis

NATURAL LANGUAGE PROCESSING

NVIDIA DEEP LEARNING SOFTWARE TRAINING STACK

At Your Desk

On-Prem



In-the-Cloud

ACCELERATED DEEP LEARNING TRAINING STACK

Image Classification Object Detection

COMPUTER VISION

Voice Recognition Language Translation

SPEECH AND AUDIO

Recommendation Engines Sentiment Analysis

NATURAL LANGUAGE PROCESSING

Productivity: Workflow, Data and Job Management, Experiments

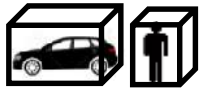
Deep Learning Software Libraries (AKA Frameworks)

Architecture Specific Libraries

At Your Desk On-Prem In-the-Cloud

ACCELERATED DEEP LEARNING TRAINING STACK

Image Classification



Object Detection

COMPUTER VISION

Voice Recognition

Language Translation

SPEECH AND AUDIO

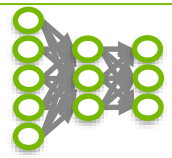
Recommendation Engines

Sentiment Analysis

NATURAL LANGUAGE PROCESSING

Productivity: Workflow, Data and Job Management, Experiments

Deep Learning Software Libraries (AKA Frameworks)



cuDNN

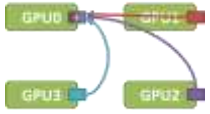
DEEP LEARNING

cuBLAS

cuSPARSE

cuFFT

MATH LIBRARIES

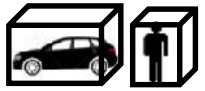


NCCL

COMMUNICATION

ACCELERATED DEEP LEARNING TRAINING STACK

Image Classification



Object Detection

COMPUTER VISION

Voice Recognition

Language Translation

SPEECH AND AUDIO

Recommendation Engines

Sentiment Analysis

NATURAL LANGUAGE PROCESSING

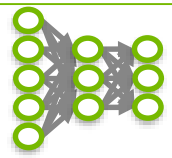
Productivity: Workflow, Data and Job Management, Experiments

TensorFlow PYTORCH Caffe2 mxnet Microsoft CNTK Caffe torch theano

NV OPTIMIZED

Chainer MATLAB ONNX

NV ACCELERATED



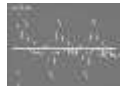
cuDNN

DEEP LEARNING

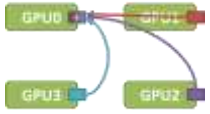
cuBLAS

cuSPARSE

cuFFT



MATH LIBRARIES

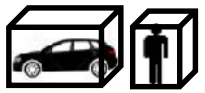


NCCL

COMMUNICATION

ACCELERATED DEEP LEARNING TRAINING STACK


Image Classification


Object Detection

COMPUTER VISION


Voice Recognition


Language Translation

SPEECH AND AUDIO


Recommendation Engines


Sentiment Analysis

NATURAL LANGUAGE PROCESSING

DIGITS, NVIDIA GPU Cloud, GPU Container, Keras, Kubernetes

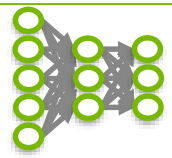
UI / JOB MANAGEMENT / DATASET VERSIONING/ VISUALIZATION

TensorFlow PYTORCH Caffe2 mxnet Microsoft CNTK Caffe torch theano

NV OPTIMIZED

Chainer MATLAB ONNX

NV ACCELERATED


cuDNN

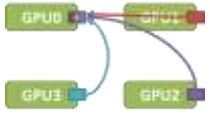
DEEP LEARNING

cuBLAS


cuSPARSE


cuFFT


MATH LIBRARIES

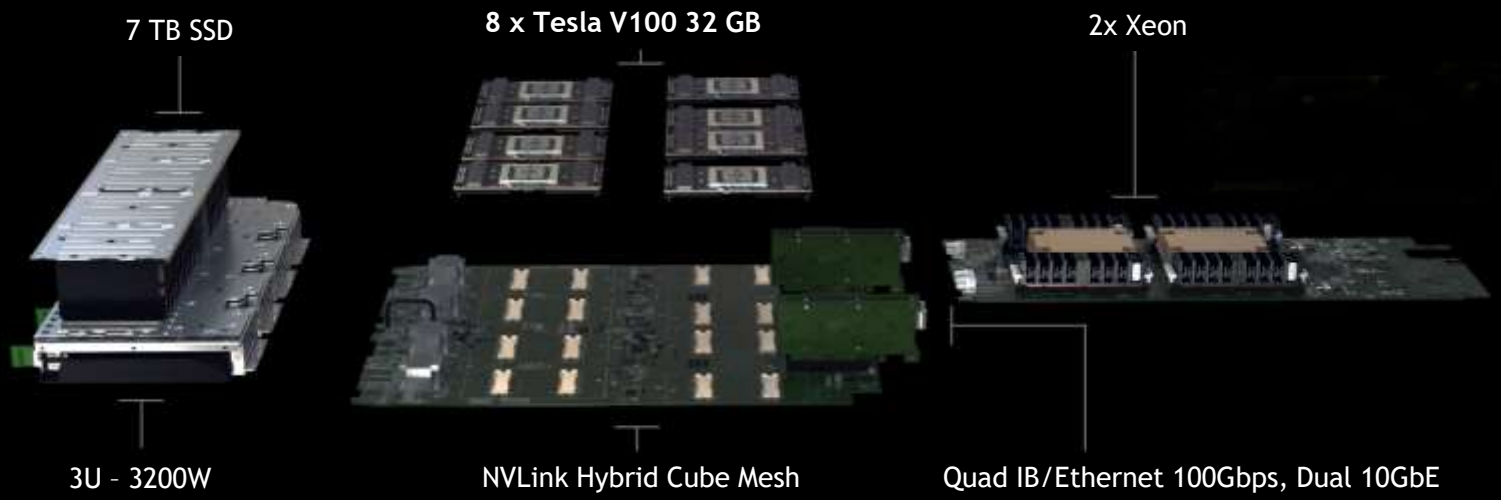


NCCL

COMMUNICATION

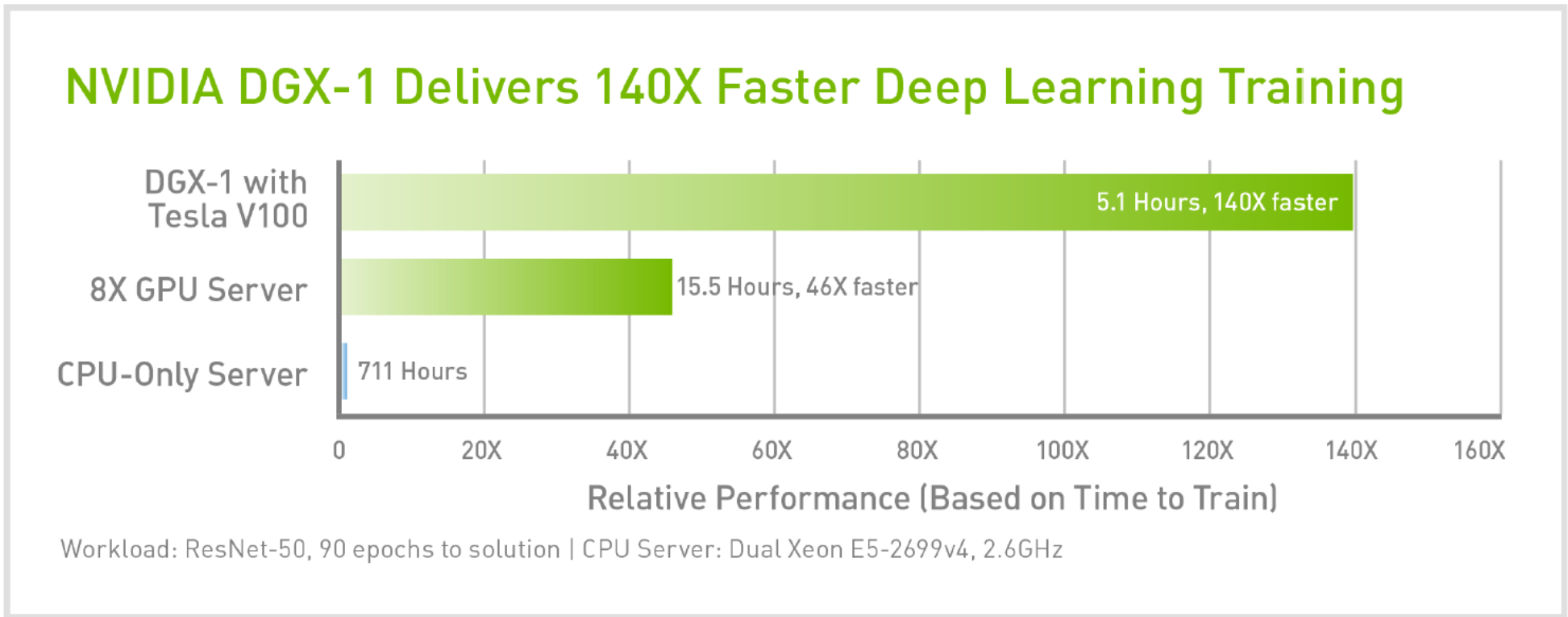
NVIDIA DGX-1 WITH 32GB VOLTA

Highest Performance, Fully Integrated HW System



1 PetaFLOPS | 8x Tesla V100 32GB | 300 Gb/s

DGX-1: 140X FASTER THAN CPU



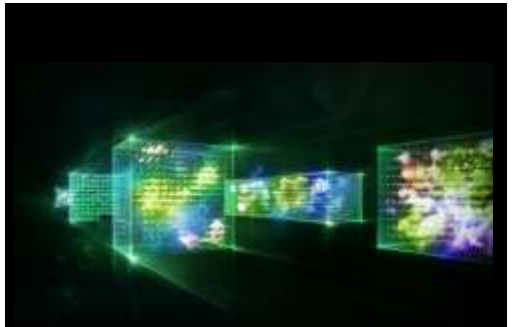
TESLA V100 32GB

WORLD'S MOST ADVANCED DATA CENTER GPU
NOW WITH 2X THE MEMORY

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

NEW TENSOR CORE BUILT FOR AI

Delivering 120 TFLOPS of DL Performance



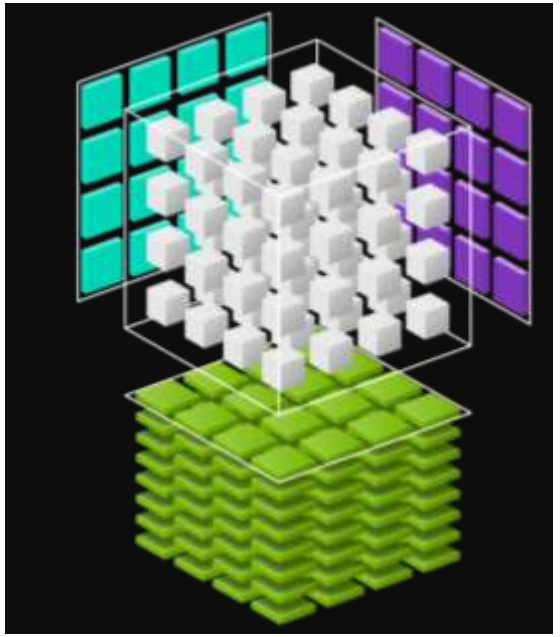
MATRIX DATA OPTIMIZATION:

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] x B[FP16] + C[FP32]
Optimized For Deep Learning

Caffe2

Microsoft Cognitive Toolkit

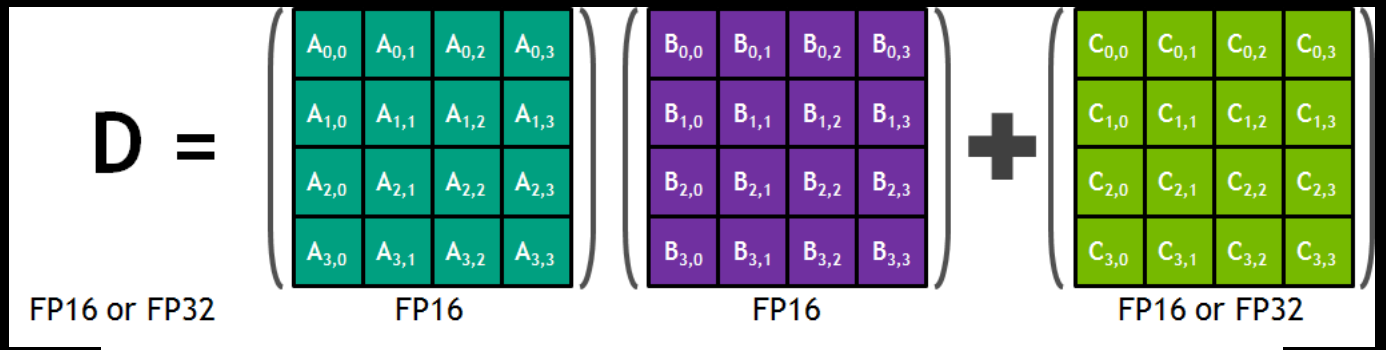
mxnet

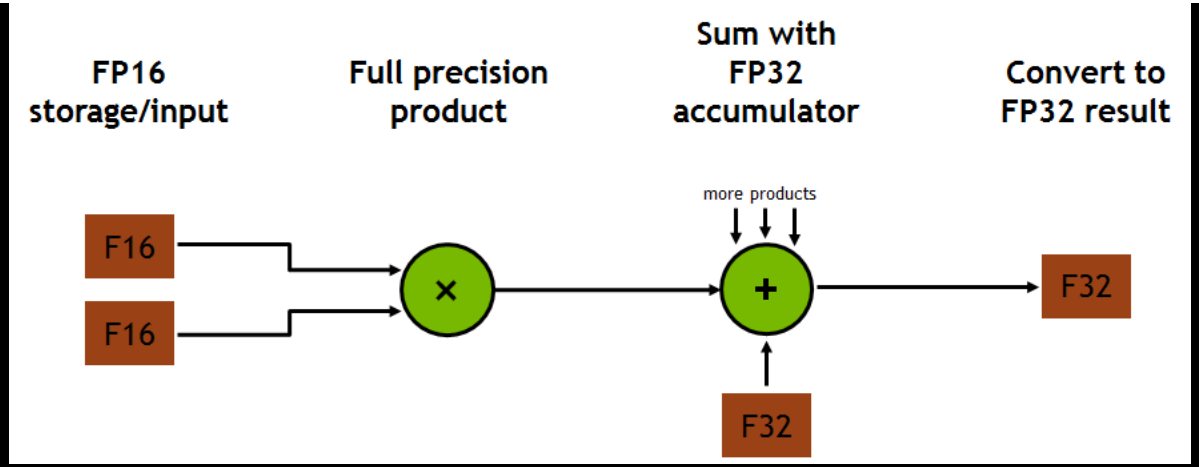
PYTORCH

TensorFlow

ALL MAJOR FRAMEWORKS

WHAT ARE TENSORCORES?





NVIDIA GPU CLOUD

GPU-ACCELERATED CLOUD PLATFORM OPTIMIZED FOR DEEP LEARNING

Containerized in NVDocker

Optimization Across the Full Stack

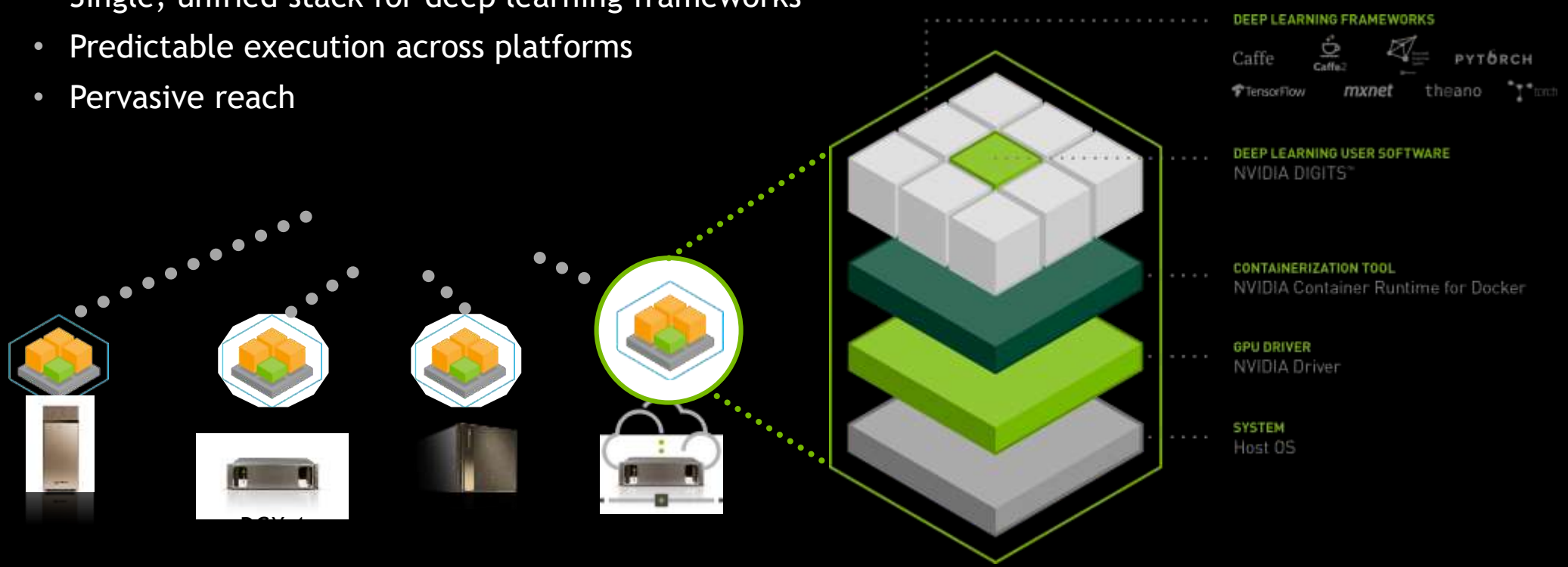
Always Up-to-Date

Fully Tested and Maintained by NVIDIA

Sign up now: www.nvidia.com/gpu-cloud

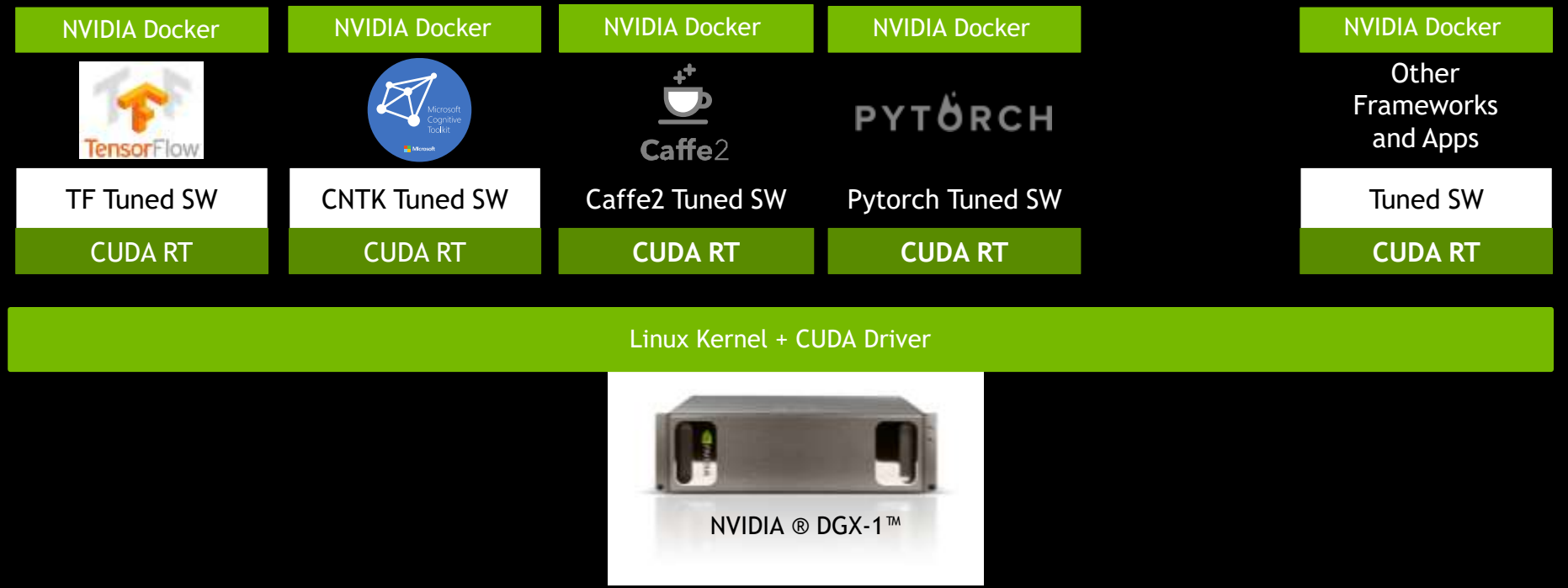
COMMON SOFTWARE STACK ACROSS DGX FAMILY

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



THE POWER TO RUN MULTIPLE FRAMEWORKS AT ONCE

Container Images portable across new driver versions



GPU ACCELERATED INFERENCING

AI - EDGE TO CLOUD

EDGE AND ON-PREMISES
Inference

CLOUD
Training and Inference

Edge/gateway device

Server

TENSORRT • DEEPSTREAM • JETPACK

NVIDIA GPU CLOUD • DIGITS

JETSON

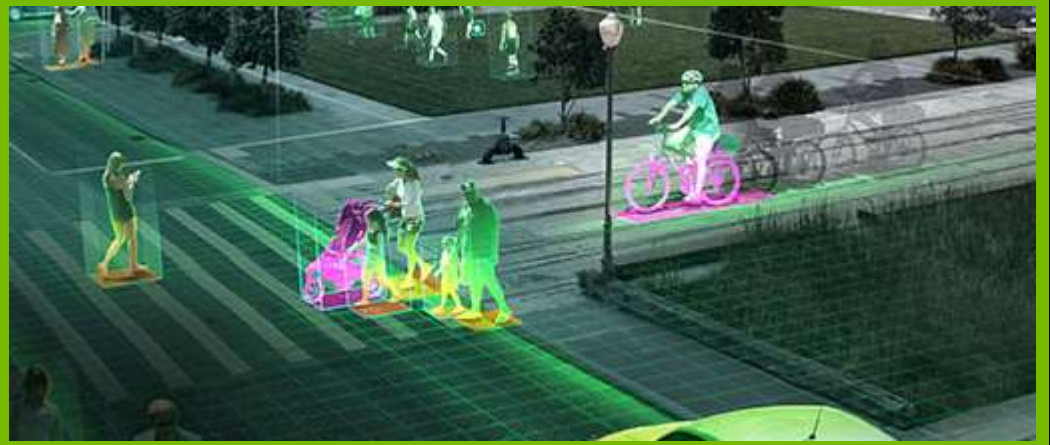
TESLA

DGX

JETSON XAVIER

	JETSON TX2	JETSON XAVIER
GPU	256 Core Pascal	**512 Core Volta**
DL Accelerator	-	**NVDLA x 2**
Vision Accelerator	-	**VLA - 7 way VLIW Processor**
CPU	6 core Denver and A57 CPUs	**8 core Carmel CPUs**
Memory	8 GB 128 bit LPDDR4 58.4 GB/s	**16 GB 256 bit LPDDR4x 137 GB/s**
Storage	32 GB eMMC	**32 GB eMMC**
Video Encode	2x 4K @30 HEVC	**2x 8K @ 30 / 8x 4K @30 HEVC**
Video Decode	2x 4K @30 12 bit support	**2x 8K @ 30 / 8x 4K @30 12 bit support**
Camera	Up to 6 cameras CSI2 D-PHY 1.2 2.5Gbps/lane	**Up to 8 cameras CSI2 D-PHY 1.2 2.5 Gbps/lane**
Mechanical	50mm x 87mm 400 pin connector	**100mm x 87mm 699 pin connector**

JETSON SDKS OVERVIEW



DEEPSTREAM SDK
FOR VIDEO ANALYTICS

ISAAC SDK
FOR AUTONOMOUS MACHINES

JETPACK SDK
FOR AI AT THE EDGE

JETSON XAVIER

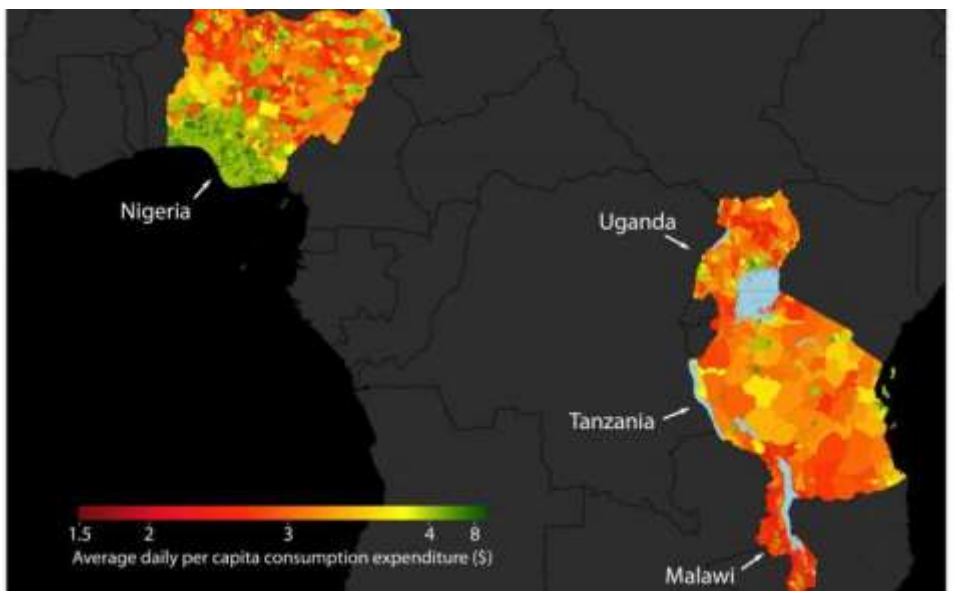
KEY LEARNINGS

Typical Industrial Challenges	Typical Solutions
Limited Image Data	Transfer Learning
Class Imbalance	Data Augmentation or Class Weighting
Missing Data Labels	Bootstrap Approach
Deployment	Deepstream SDK

FEW INTERESTING PROJECTS

DATA TO INSIGHTS

Predicting Poverty



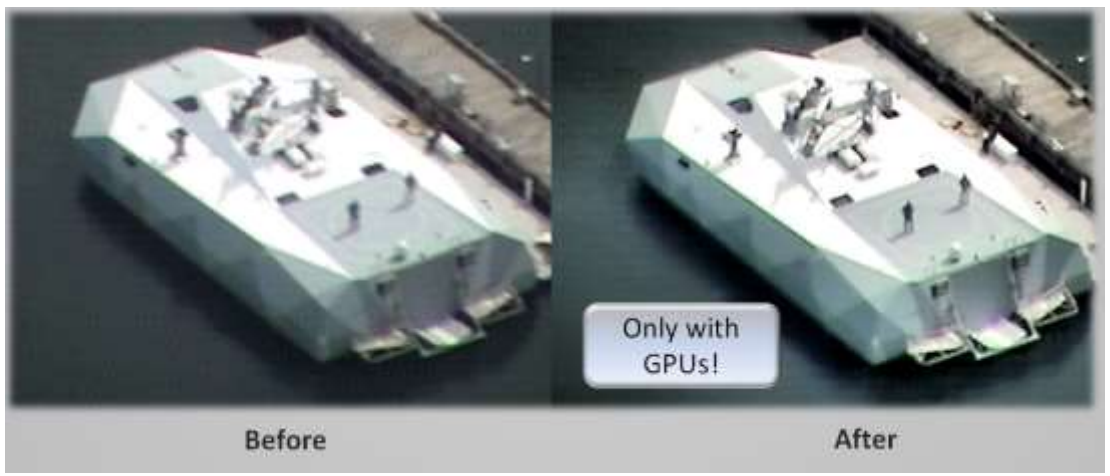
- Algorithm: compare the presence of light in a region during the day and at night to predict it's economic activity.
- Assumption: a brightly lit area means it is powered by electricity and must be better off than the alternative
- Learning criteria: it cross check it's results with actual survey data in order to improve it's accuracy

OBJECT LOCALIZATION

Fast Object Detection

IMAGE ENHANCEMENT

Super Resolution



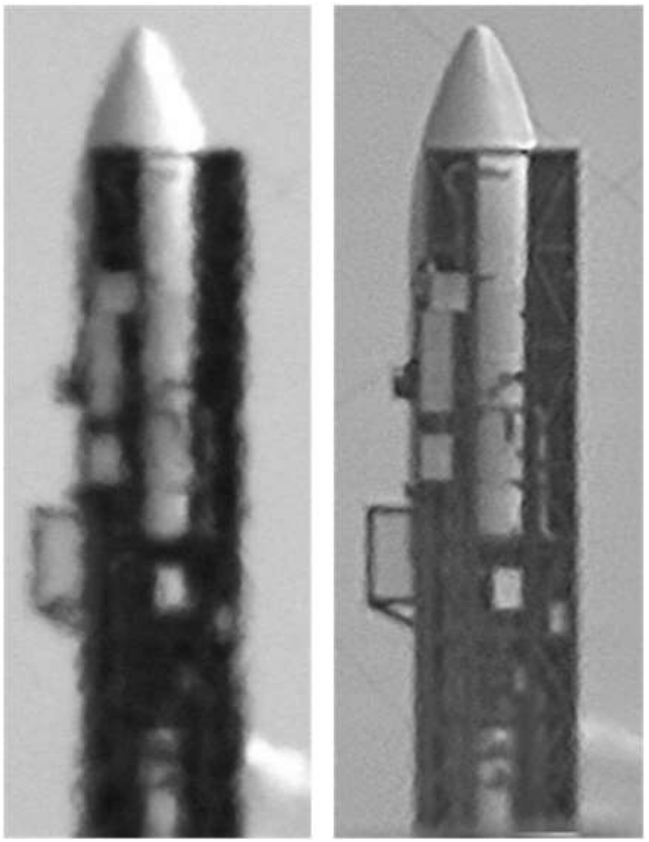
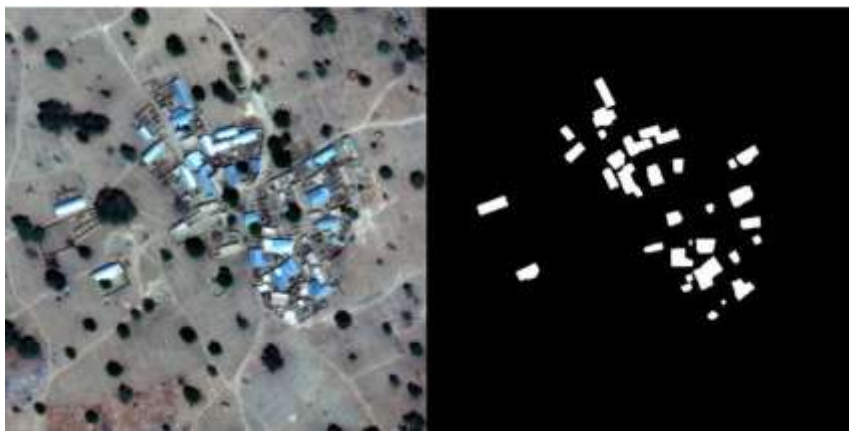
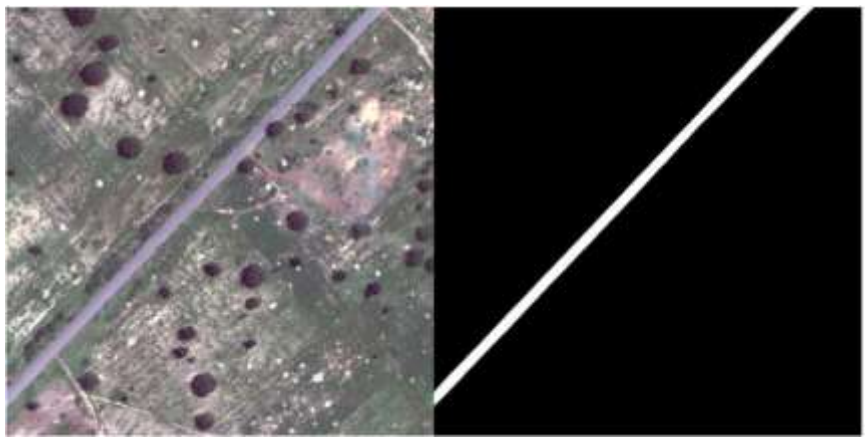
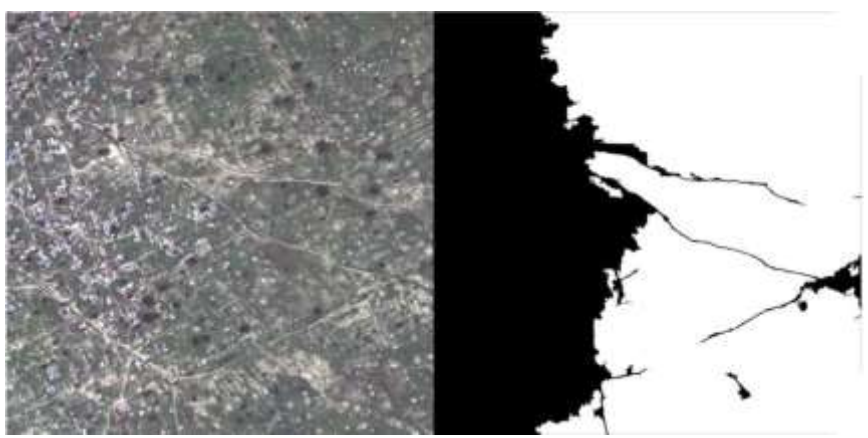


IMAGE SEGMENTATION

Semantic Segmentation

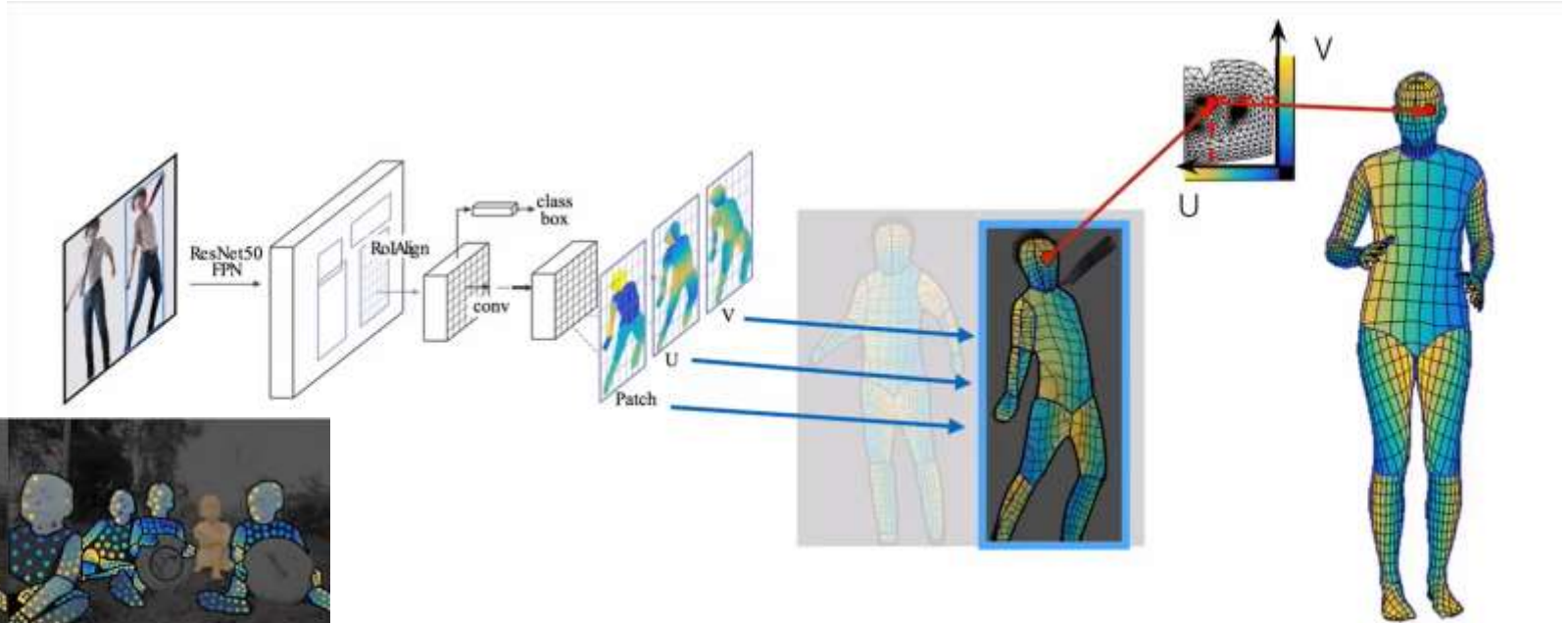


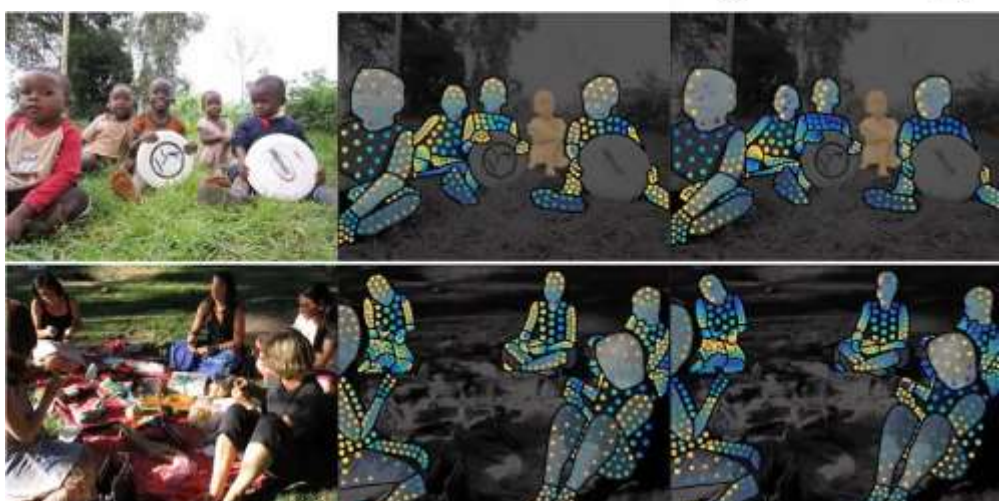




HUMAN ACTION DETECTION

Dense Pose Estimation





IN SUMMARY

READY TO GET STARTED?

Project checklist

Internet Services

Robotics

Digital Content Creation

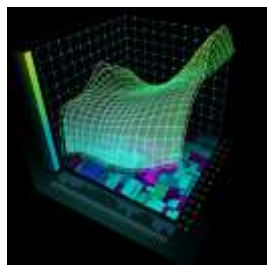
Intelligent Video Analytics

What problem are you solving, what are the DL tasks?

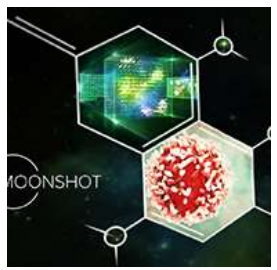
What data do you have/need, and how is it labeled?

Which deep learning framework & tools will you use?

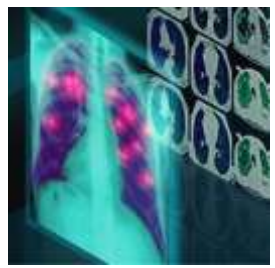
On what platform(s) will you train and deploy?



Finance



Genomics



Healthcare

Autonomous Vehicles

Media & Entertainment

Security & Defense

WHO

RESEARCHERS
Explore the “next big thing” opportunity to fuel business

APPLIED DL/ DATA SCIENTISTS
Retrain w/ data, productize models for consistency, focus on quality

APPLICATION DEVELOPER
Scale and deploy successful applications w/ great user ex.

WHO, WHAT

RESEARCHERS
Explore the “next big thing” opportunity to fuel business, and find ways to productize it

APPLIED DL/ DATA SCIENTISTS
Retrain, productize models for consistency, quality, tuning with right data

APPLICATION DEVELOPER
Scale and deploy successful applications w/ great user ex.

Image Classification

Object Detection

Voice Recognition

Language Translation

Recommendation Engines

Sentiment Analysis

WHO, WHAT, WHERE

RESEARCHERS
Explore the “next big thing” opportunity to fuel business, and find ways to productize it

DATA SCIENTISTS
Retrain, productize models for consistency, quality, tuning with right data

APPLICATION DEVELOPER
Scale and deploy successful applications w/ great user ex.

Image Classification

Object Detection

Voice Recognition

Language Translation

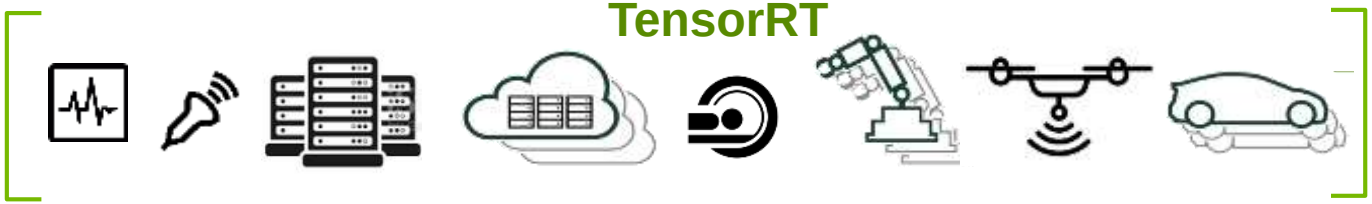
Recommendation Engines

Sentiment Analysis

Training

or

TensorRT



Deploying

SELF TRAINING PLATFORM

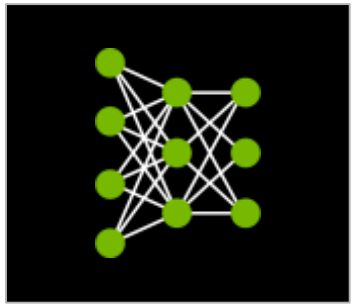
NVIDIA DEEP LEARNING INSTITUTE

Hands-on, self-paced and instructor-led training in deep learning and accelerated computing for developers

Request onsite instructor-led workshops at your organization: www.nvidia.com/requestdli

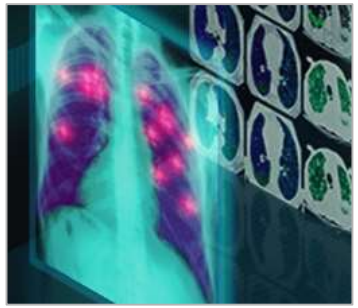
Take self-paced courses and electives online, view upcoming workshops, and learn about the University Ambassador Program: www.nvidia.com/dli

Caffe2 · Microsoft Cognitive Toolkit · mxnet · TensorFlow · PYTORCH

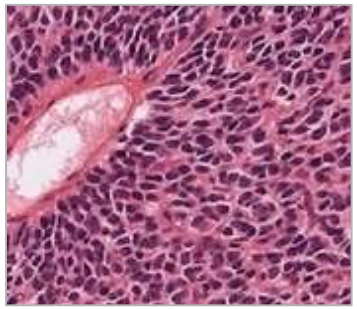


Deep Learning Fundamentals

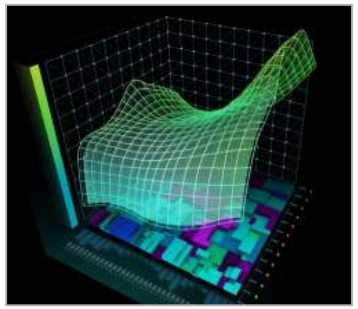
Autonomous Vehicles



Medical Image Analysis



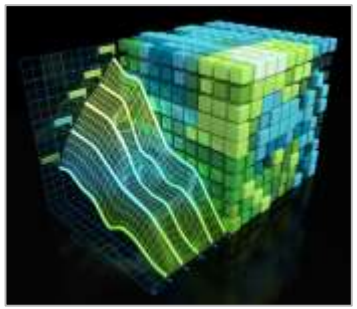
Genomics



Finance

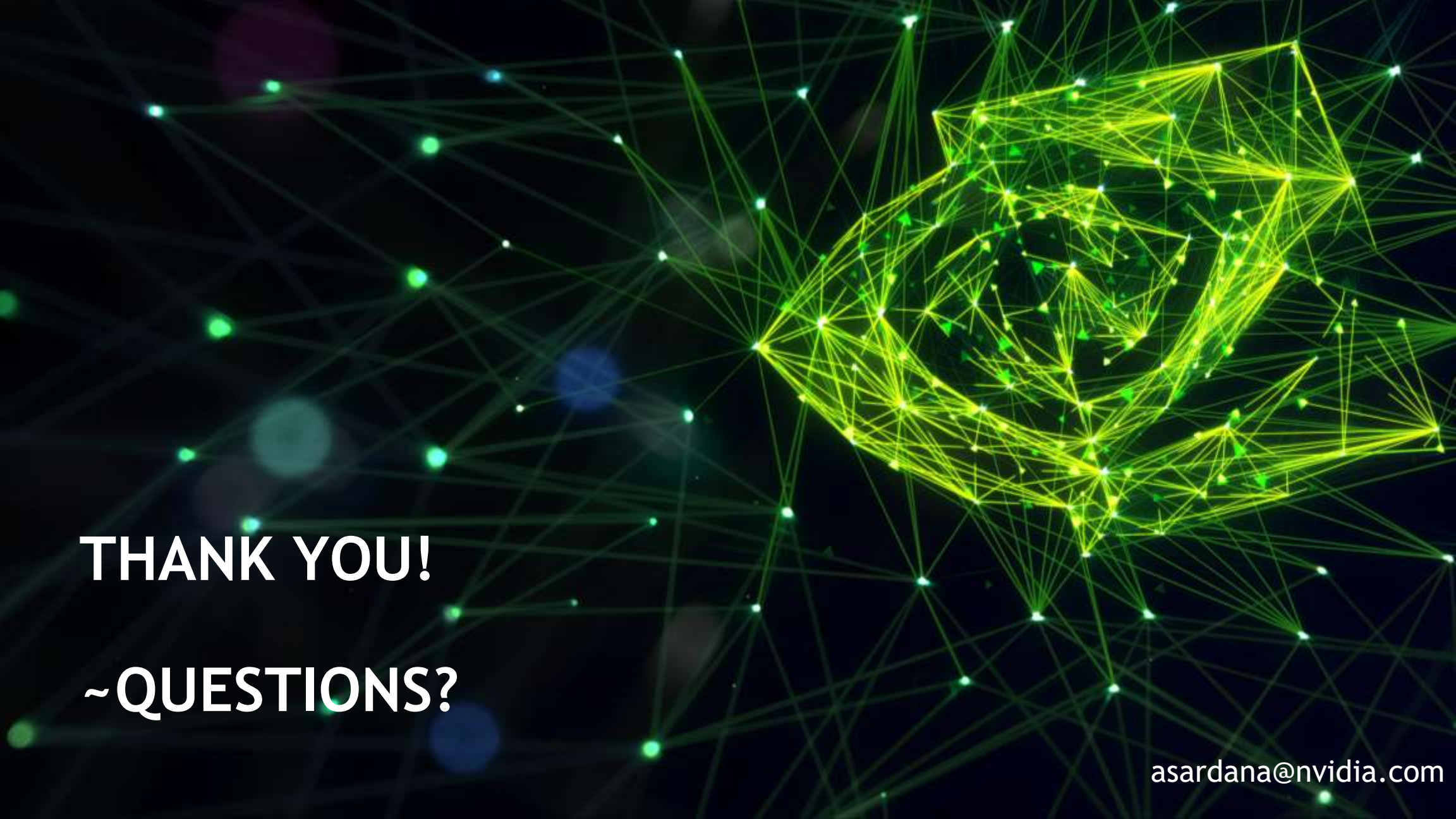
Digital Content Creation

Game Development



Accel. Computing Fundamentals

More industry-specific training coming soon…



asardana@nvidia.com