ALISON B LOWNDES

AI DevRel | EMEA

@alisonblowndes

November 2019

Harnessing the power of AI for All Humankind

Session 2: The Benefits of Space for All

NVIDIA

GRAPHICS

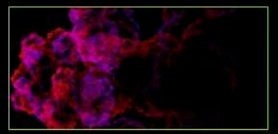
GAMING

DESIGN

RENDERING

HPC

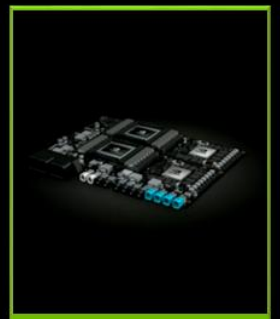
SUPERCOMPUTING



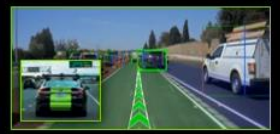
AI

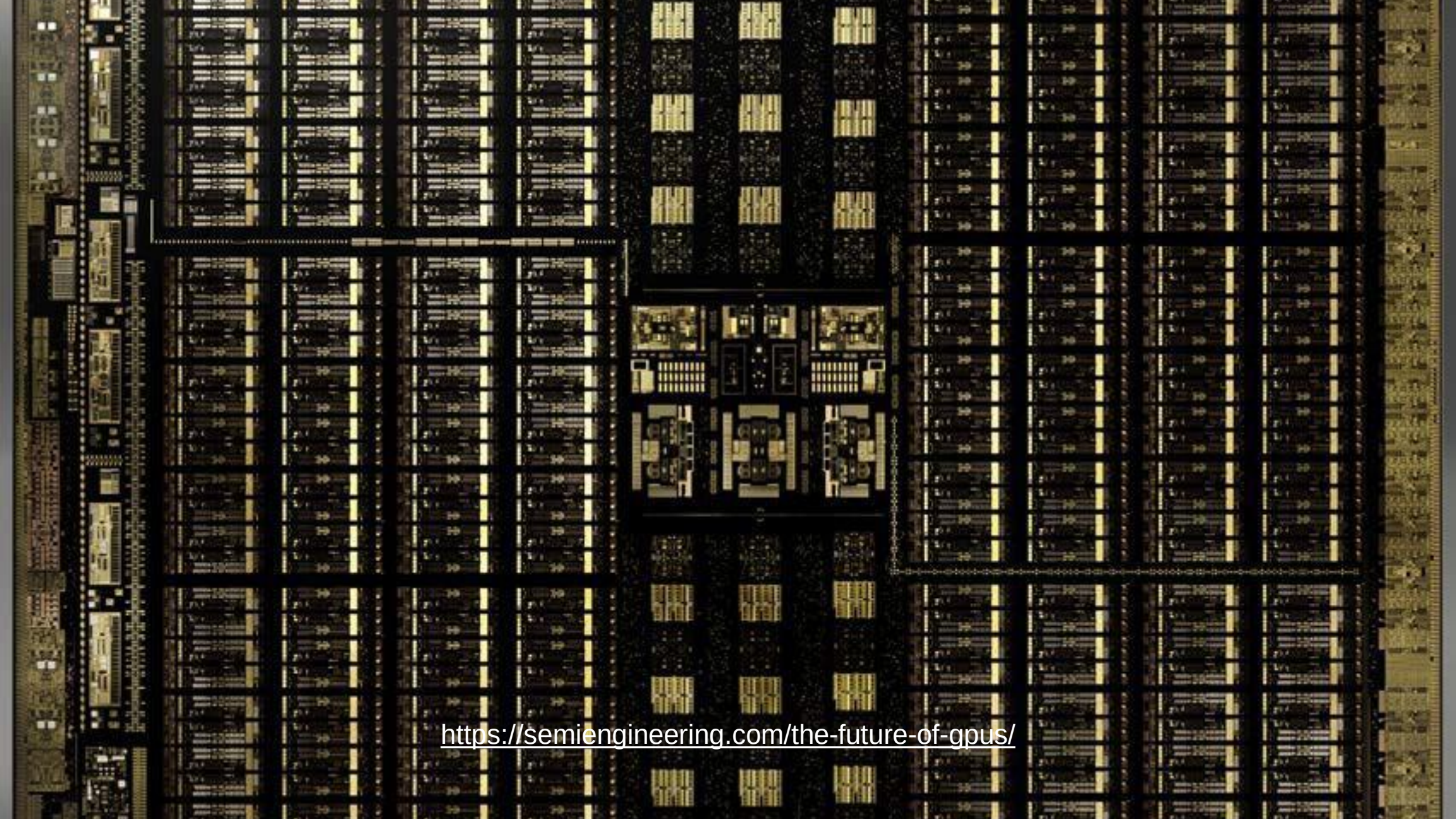
AI TRAINING

AI INFERENCE



ROBOTICS





https://semiengineering.com/the-future-of-gpus/

www.FrontierDevelopmentLab.org

NASA FRONTIER DEVELOPMENT LAB

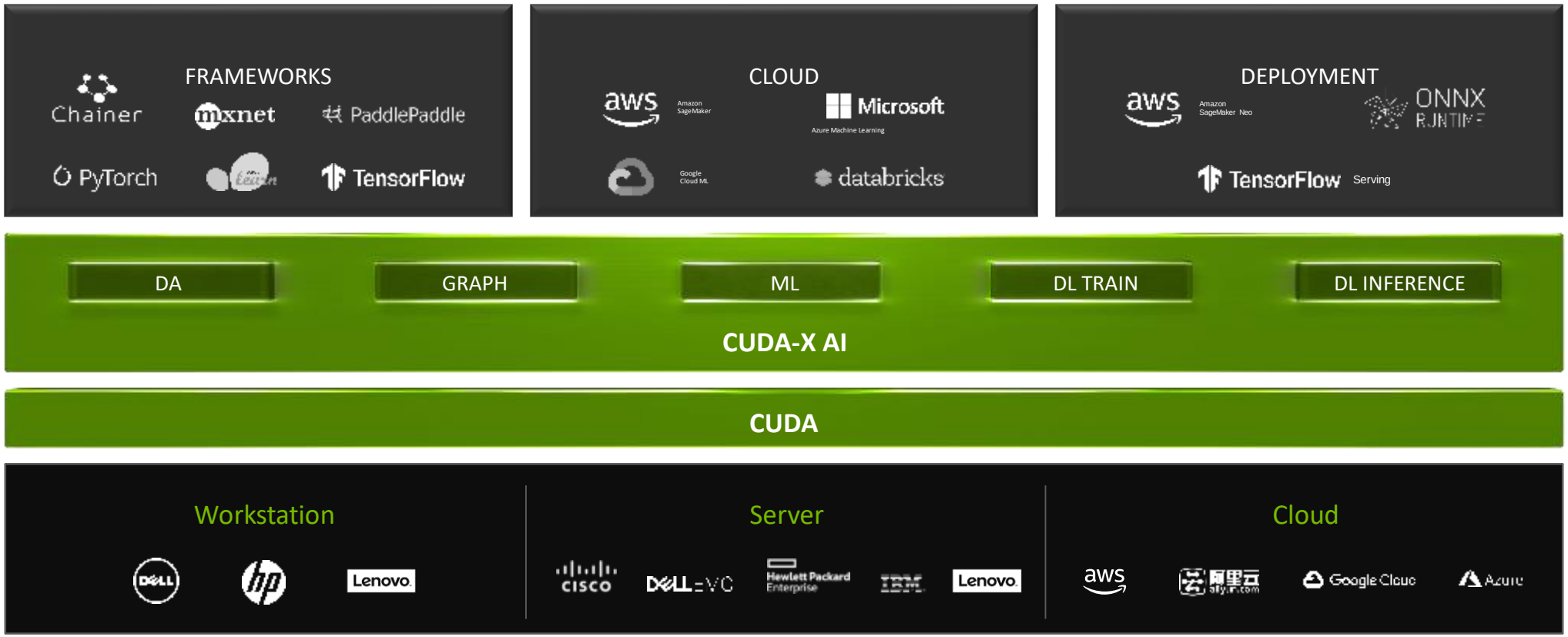
SETI INSTITUTE · esa · Google · NVIDIA · IBM · UNIVERSITY OF OXFORD · LOCKHEED MARTIN · Hewlett Packard Enterprise

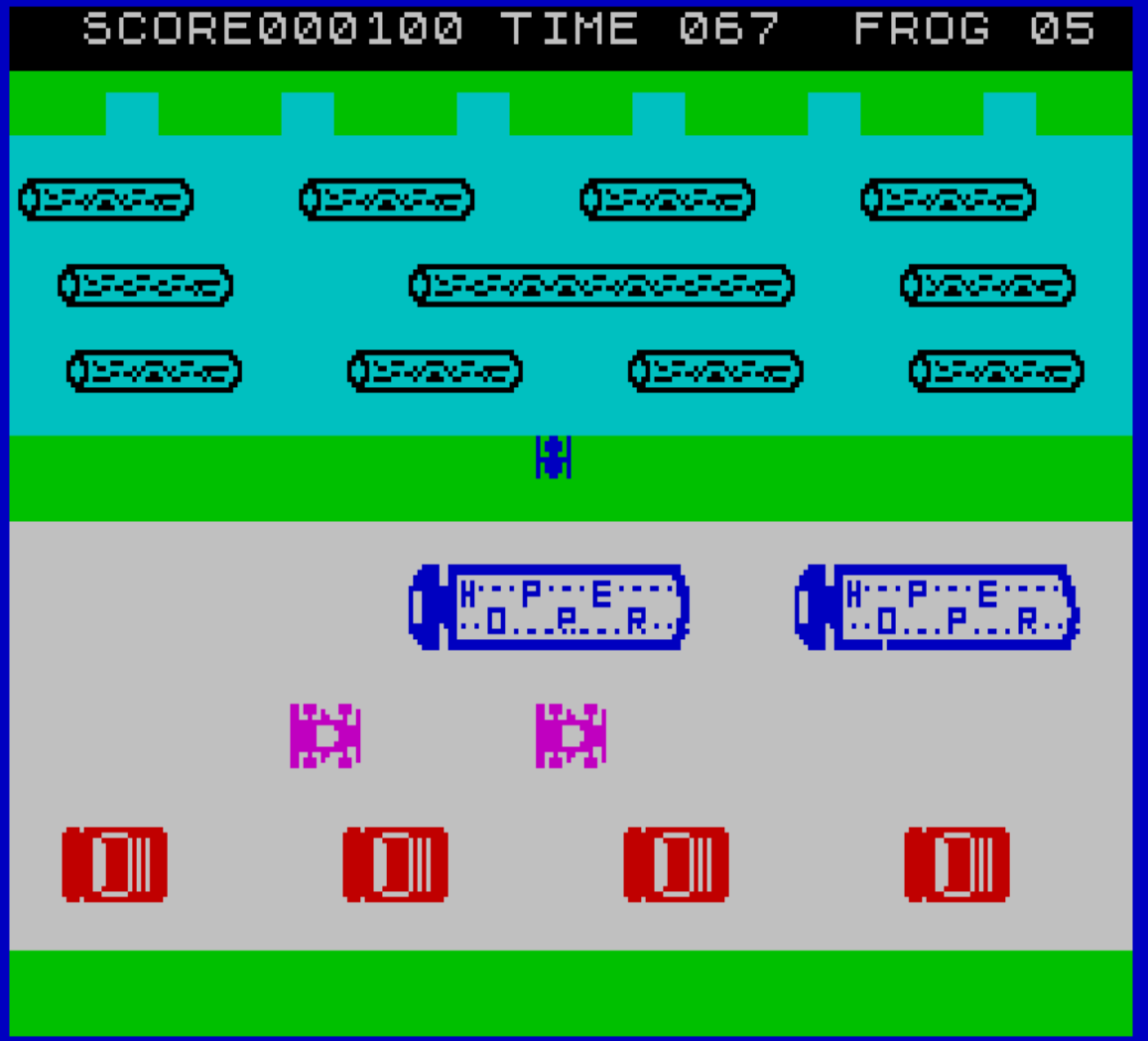
SELECTING THE RIGHT GPU SOLUTION

Unparalleled Data Science Performance and Productivity

	ML Enthusiast	Machine Learning Developer Data Science Workstations		Data Center Machine Learning Shared infrastructure for Data Science Teams	
					
	TITAN RTX	Quadro Workstation	DGX Station	DGX-1 / OEM	DGX-2 / OEM
Benefit	PC solution, easy to acquire, deploy and get started experimenting	Enterprise workstation for experienced data scientists	Enterprise ML workgroups, largest memory on a workstation	Enterprise server, proven 8-way configuration, modular approach for scale, multi-node training	Largest compute and memory capacity in single node, fastest training solution
GPU Memory	48GB	64GB	128GB	256GB	512GB
GPU Fabric	2-way NVLINK	2-way NVLINK	4-way NVLINK	8-way NVLINK	16-way NVSWITCH

NVIDIA CUDA-X AI ECOSYSTEM

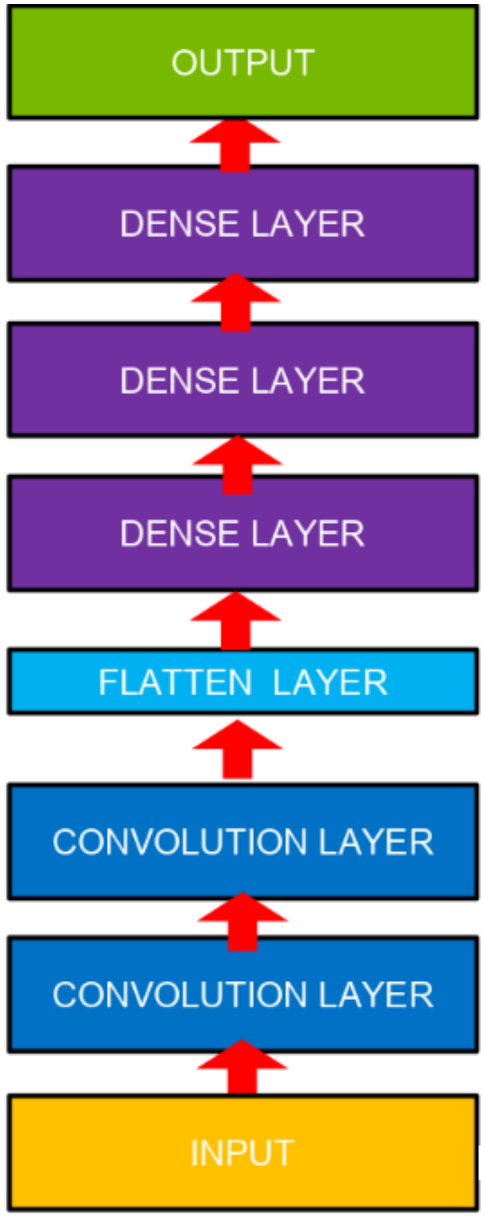




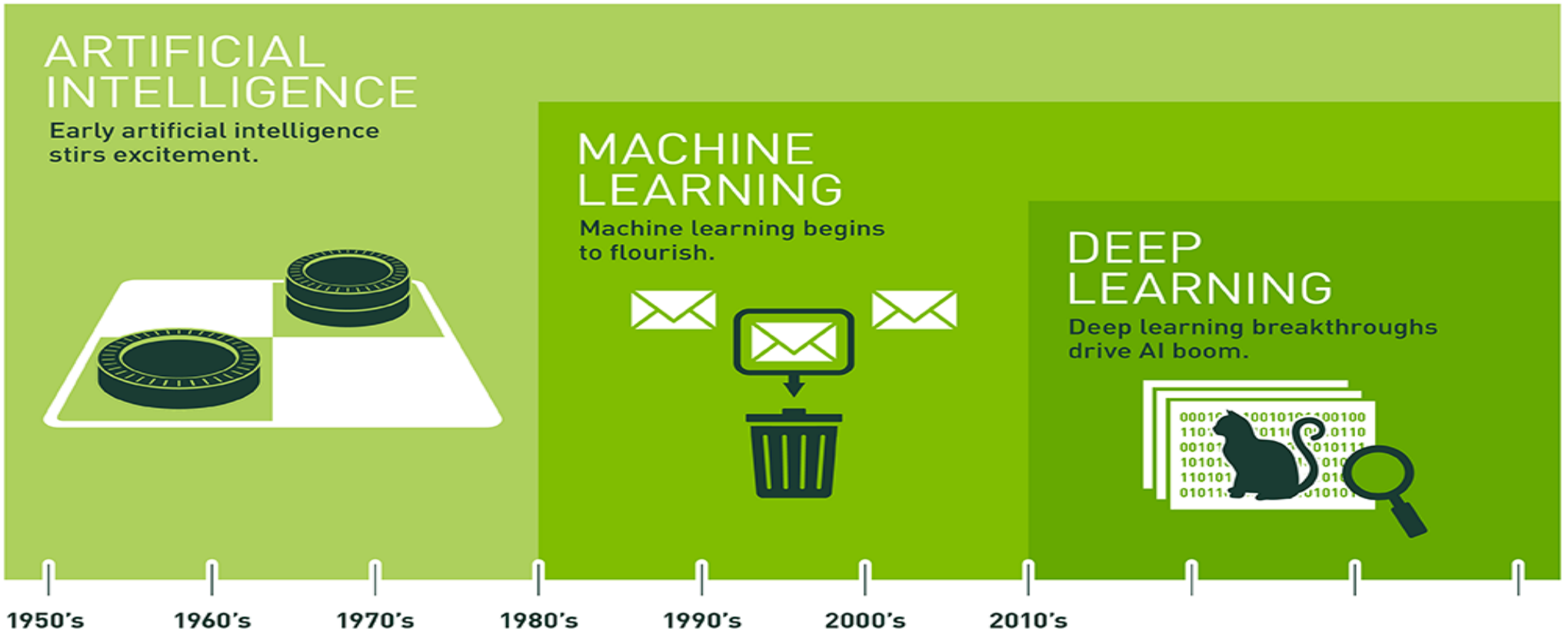

Move to Beacon
Collect Minerals and Gas
Collect Mineral Shards
Build Marines

AI == CODE

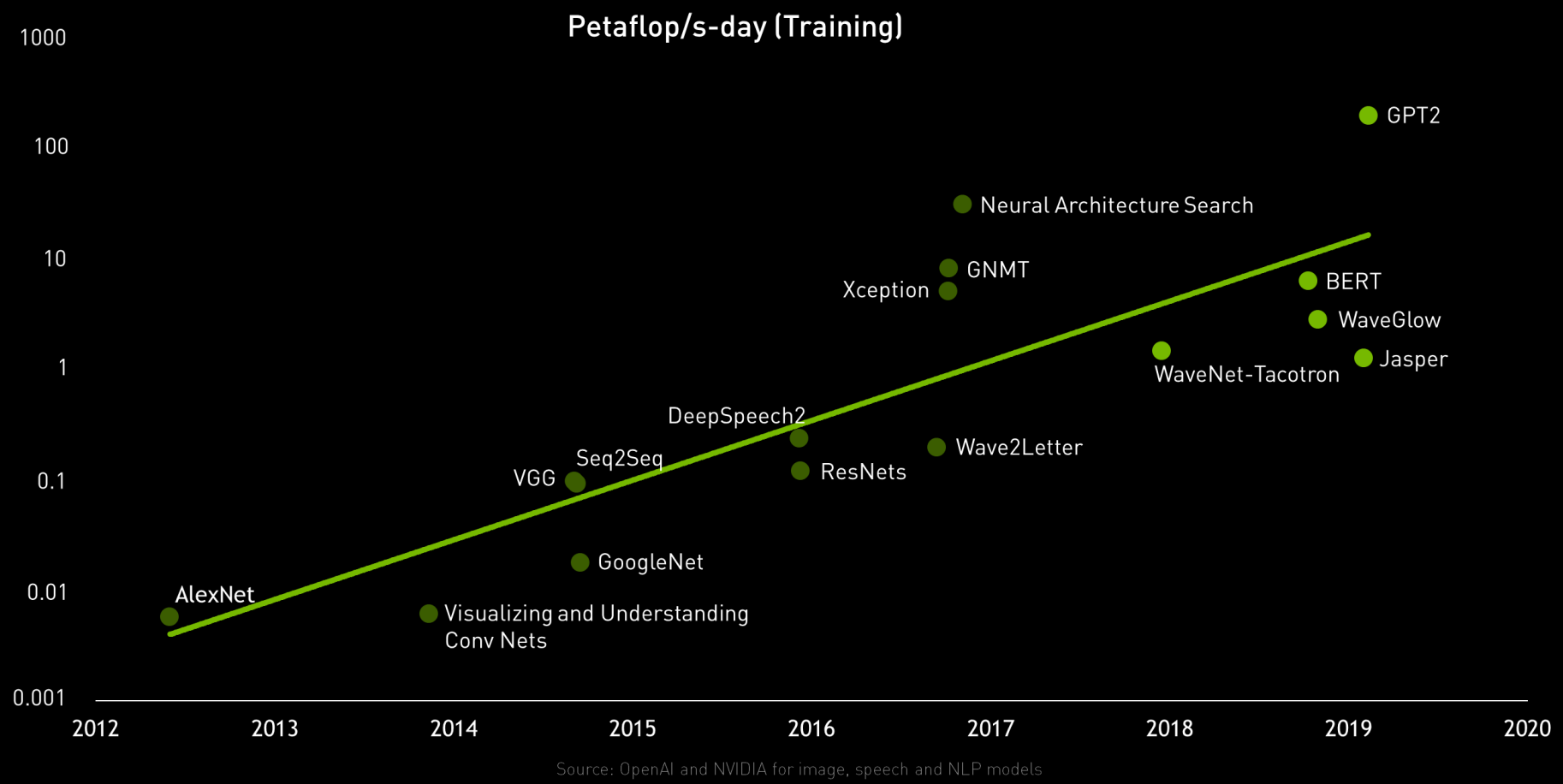
```
class LeNet(nn.Module):
        def forward(self, x):
            out = F.relu(self.conv1(x))
            out = F.relu(self.conv2(out))
            out = out.view(out.size(0), -1)
            out = F.relu(self.fc1(out))
            out = F.relu(self.fc2(out))
            out = self.fc3(out)
            return out
```



Definitions

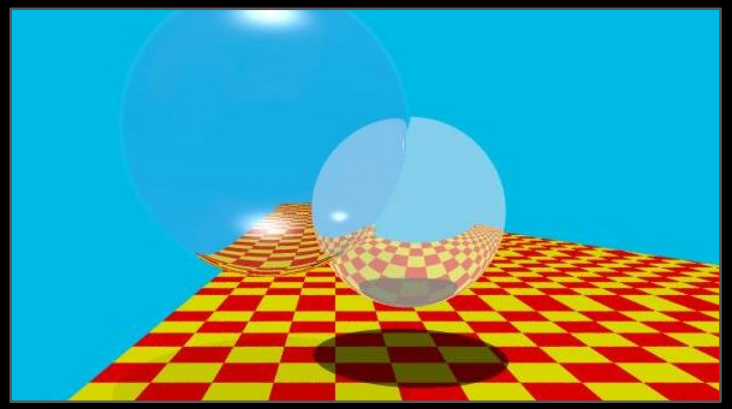


NETWORK COMPLEXITY IS EXPLODING

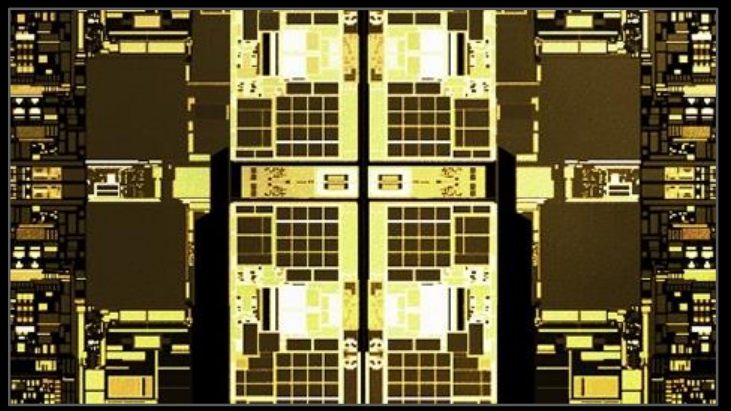


Source: OpenAI and NVIDIA for image, speech and NLP models

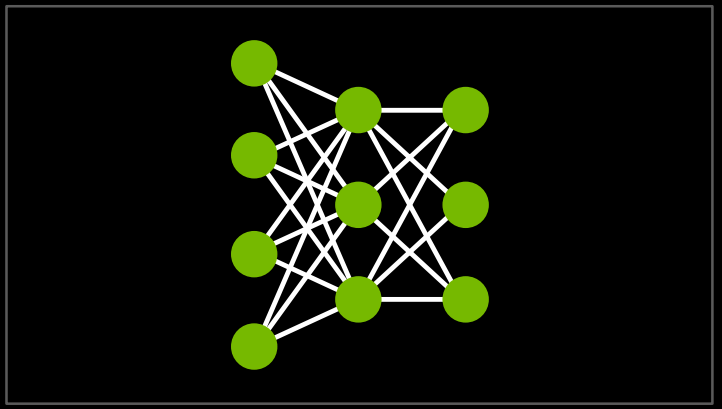
NVIDIA RESEARCH



RTX



NVSwitch

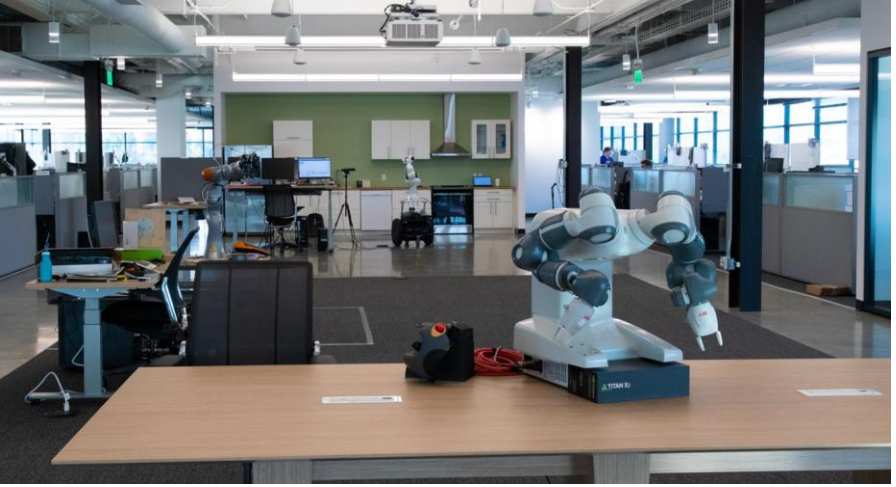


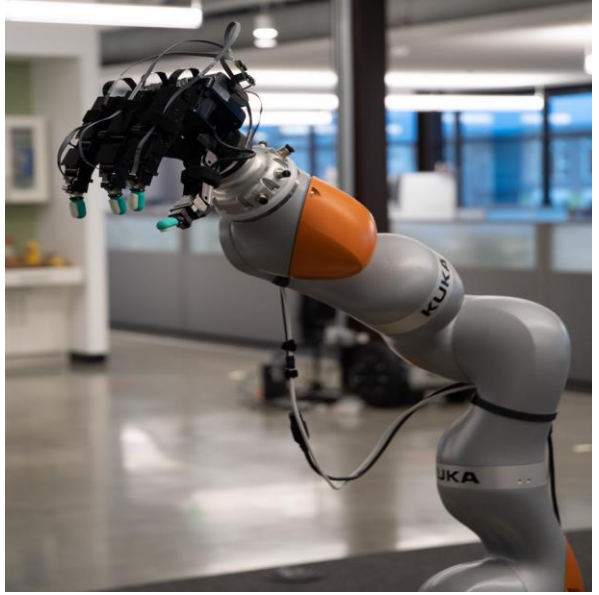
CuDNN

Image

Noise-to-Noise Denoising

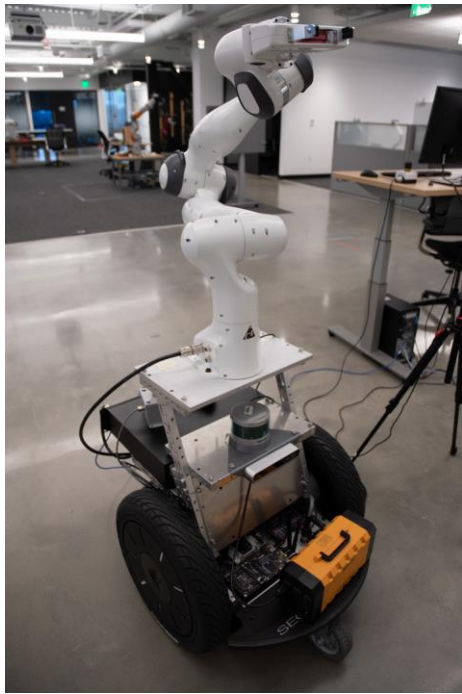
Progressive GAN





NVIDIA ROBOTICS RESEARCH LAB SEATTLE

Drive breakthrough robotics research to enable the next-generation of robots that safely work alongside humans, transforming industries such as manufacturing, logistics, healthcare, and more

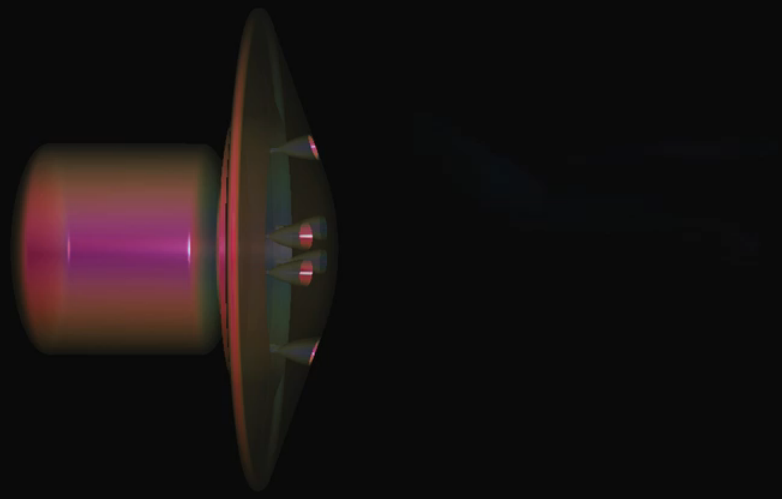


DexPilot: Depth-Based Teleoperation of Dexterous Robotic Hand-Arm System

Ankur Handa*†, Karl Van Wyk*†, Wei Yang†, Jacky Liang‡, Yu-Wei Chao†,
Qian Wan†, Stan Birchfield†, Nathan Ratliff†, Dieter Fox†

† NVIDIA ‡ Carnegie Mellon University

* Equal Contribution

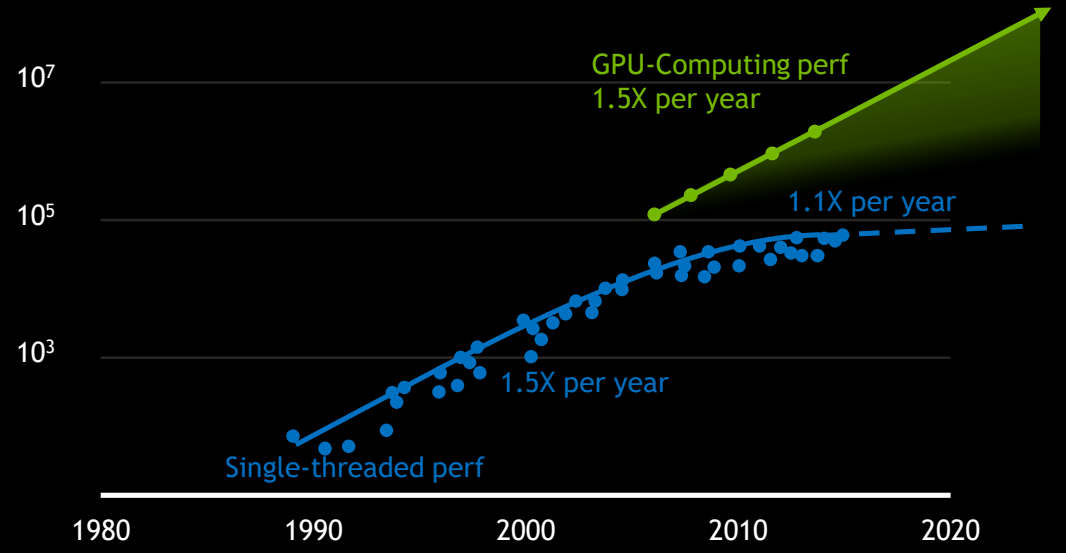


NVIDIA INDEX

THE RISE OF GPU COMPUTING

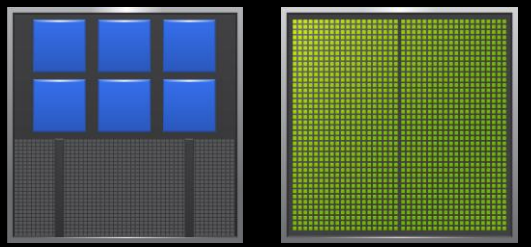
Big Data Needs Algorithms and Compute That Scales

END OF MOORE'S LAW

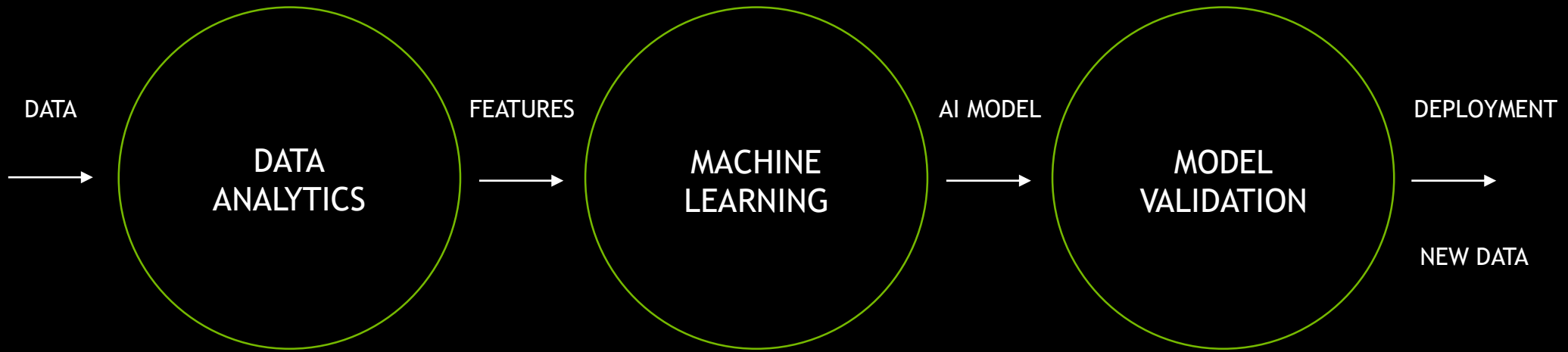


Original data up to the year 2010 collected and plotted by M. Horowitz,
F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

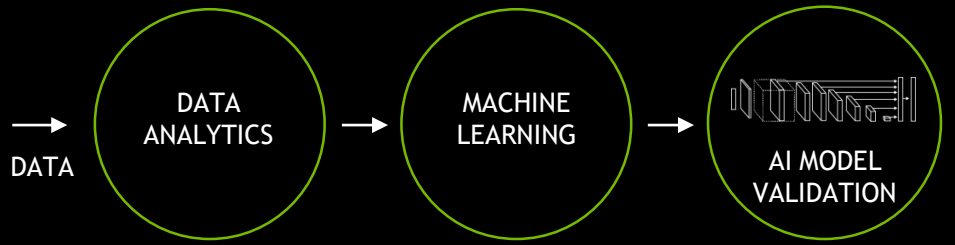
CPU vs. GPU

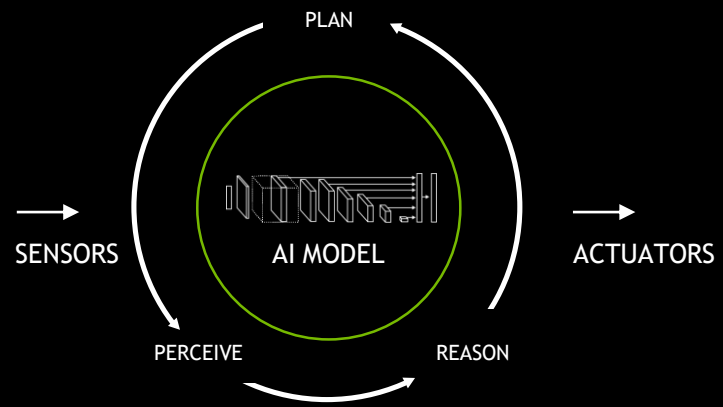


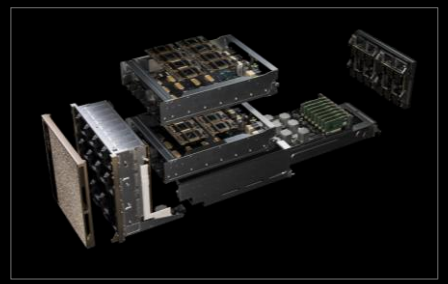
BUILDING AN AI MODEL



BUILDING AN AI PRODUCT







HARNESSING AI

Step I: Build a data fabric for your organization

Step II: Define your objective

Step III: Hire the right talent

Step IV: Identify key processes to augment with AI

Step V: Create a sandbox lab environment

Step VI: Operationalize successful pilots

Step VII: Scale up for enterprise-wide adoption

Step VIII: Drive cultural change


https://arxiv.org/pdf/1909.13371.pdf

MIND FOUNDRY

Applied Machine Learning

Designing and building state-of-art Automated Machine Learning software and services for real world problems

Background

Spin out from the University of Oxford

Combining leading edge machine learning research with advanced software engineering

Working with leading companies to deliver insight & quantifiable benefits

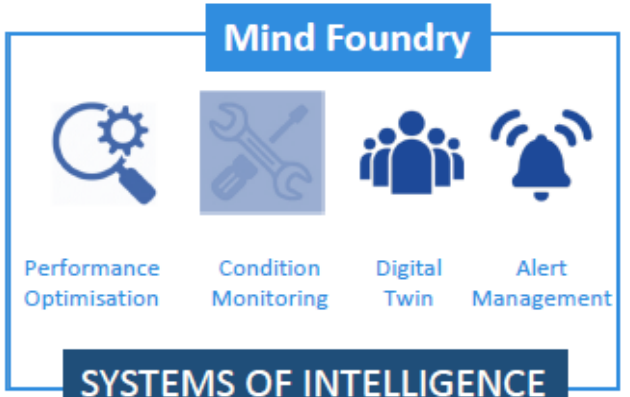
Setting benchmarks for a new wave of machine learning

Vision

Build services and software that Automate the Data Science lifecycle

Services which understand the entanglement, uncertainty, complexity and power of machine learning

Services that are evergreen and quantum ready



Ingests, cleans and feature engineers even dirty data

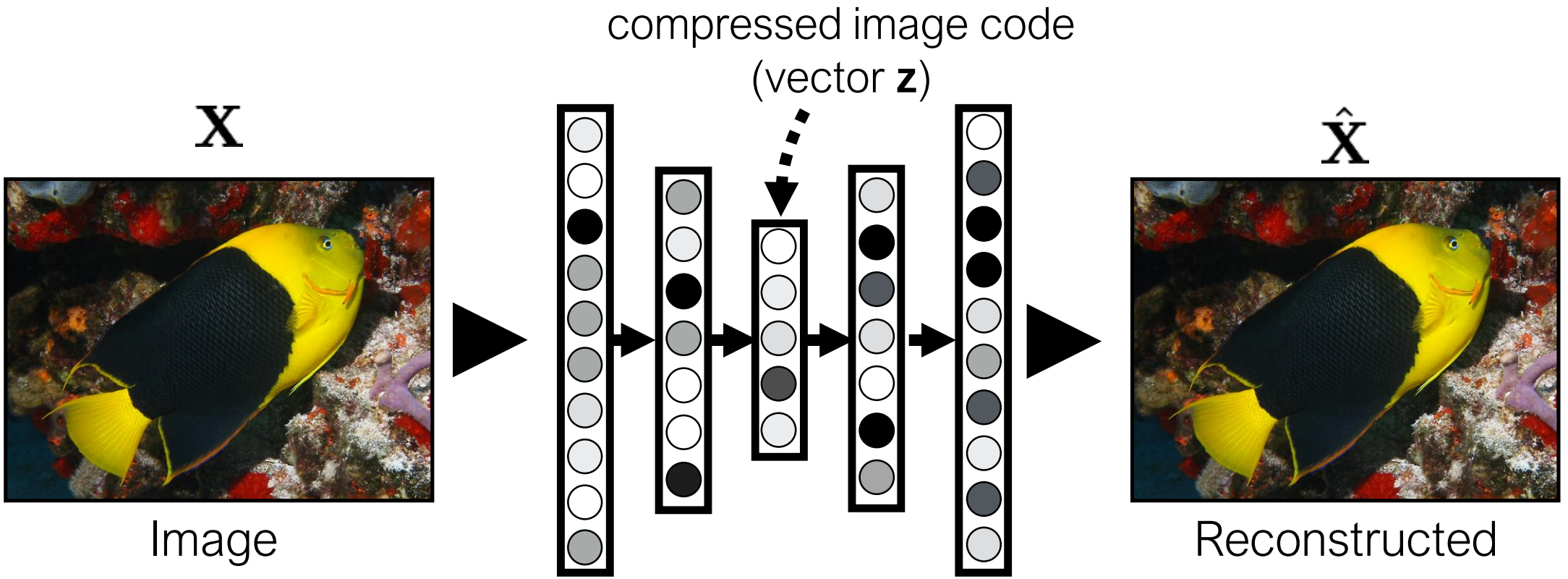
Intelligently creates ML products and services, accommodating uncertainty, hardware & time constraints

Automatically deploys and creates code

Mind Foundry Automated Data Scientist (AuDaS)

Mind Foundry, www.mindfoundry.ai

Representation Learning



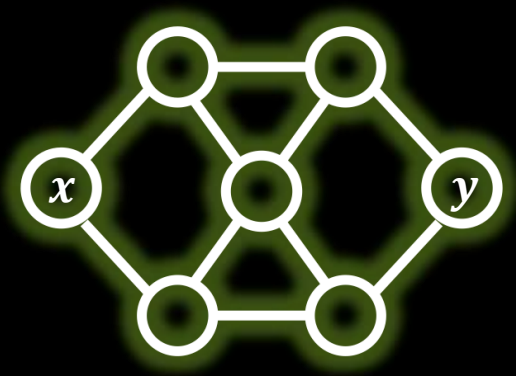
"Autoencoder"

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

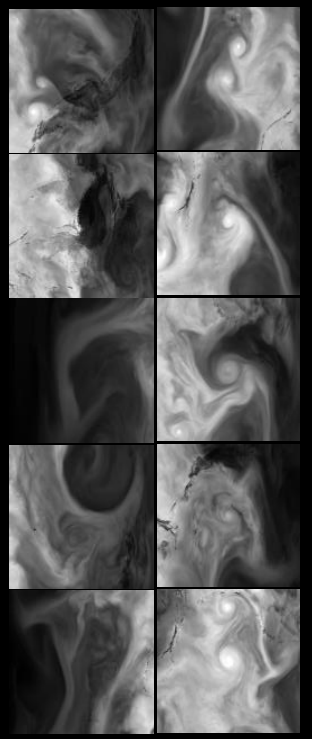
Slide by Alexei A Efros
Aug 2017, NVIDIA

[e.g., Hinton & Salakhutdinov, Science 2006]

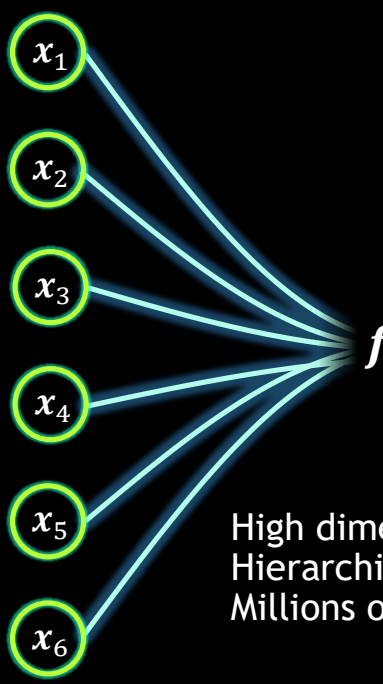
Find f, given x and y



Supervised Deep Learning



inputs



$f(x)$

High dimensional x,y
Hierarchical
Millions of parameters

outputs

0	1
0	1
0	1
0	1
0	1

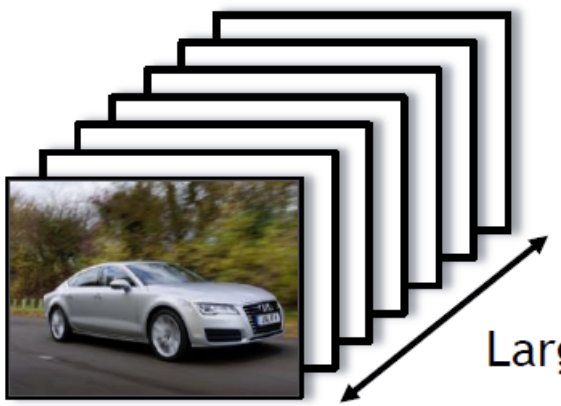
tgr strp

tiger stripe

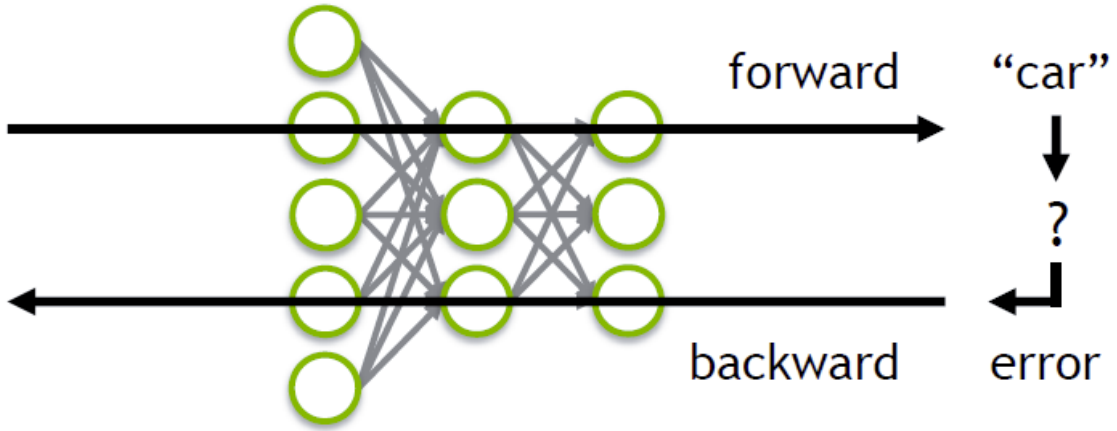
Black & Orange

Thorsten Spoerlein / Getty

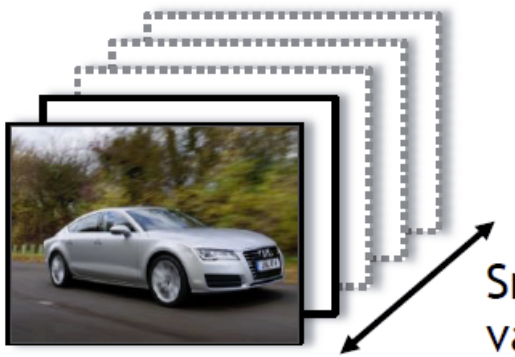
TRAINING VS INFERENCE



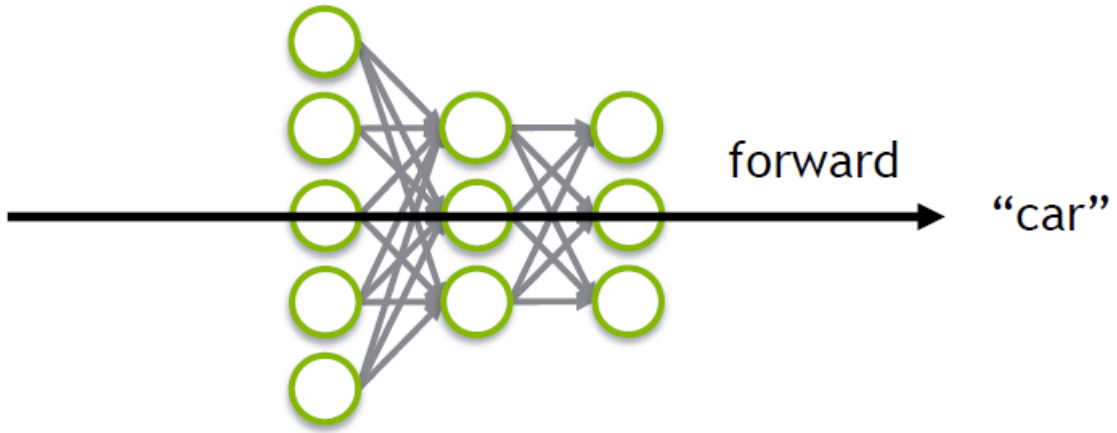
Large N



TRAINING

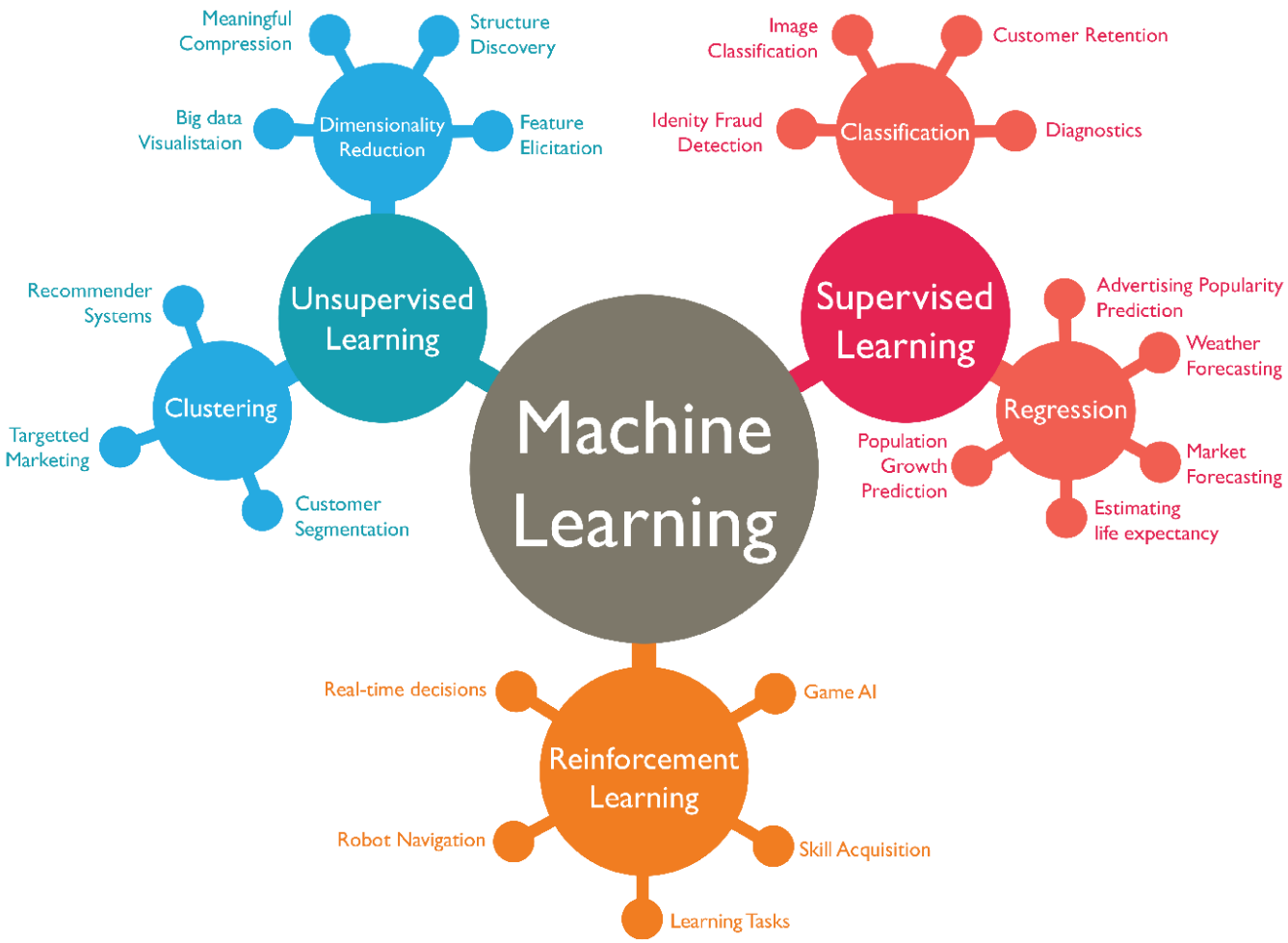


Smaller,
varied N

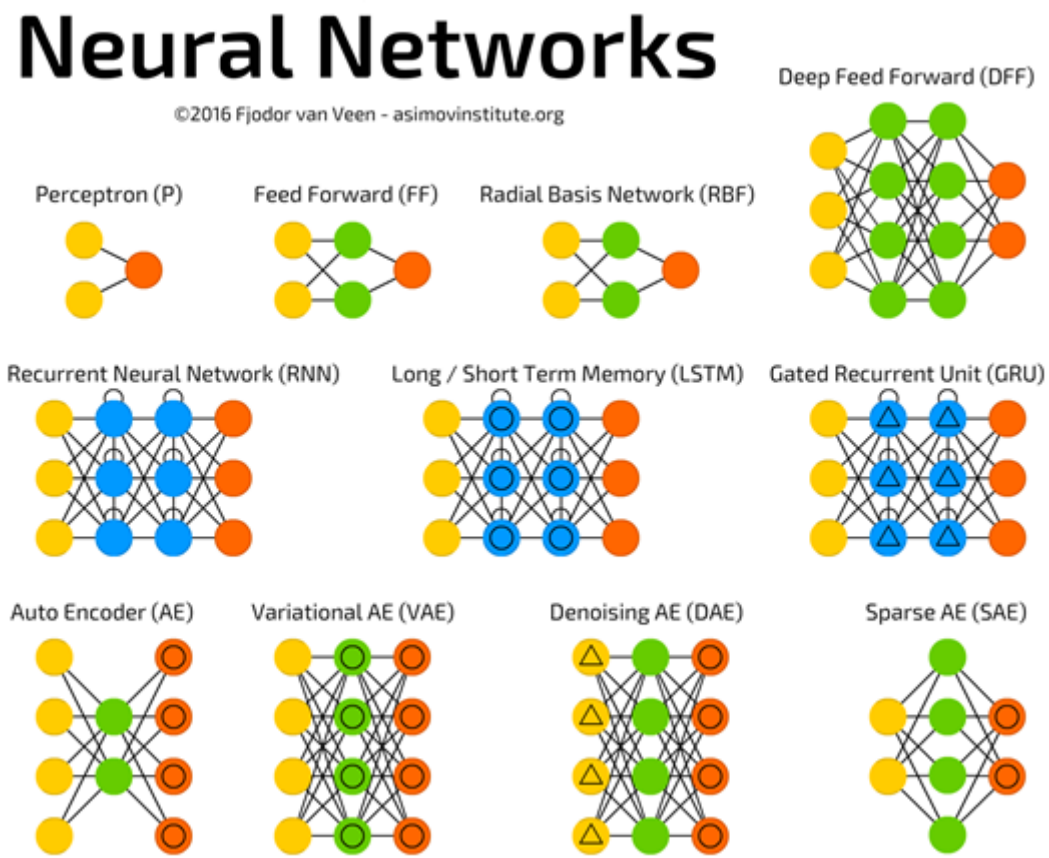


INFERENCE

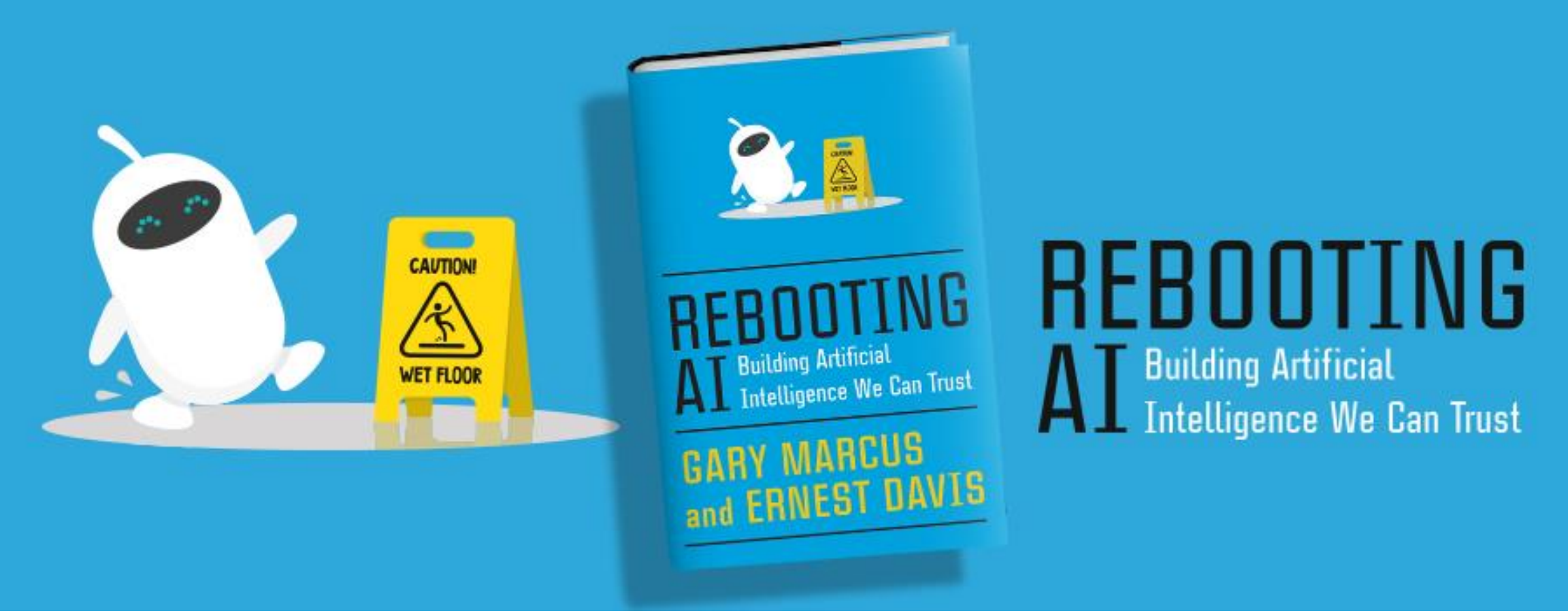
Types of ML/DL



ARCHITECTURES

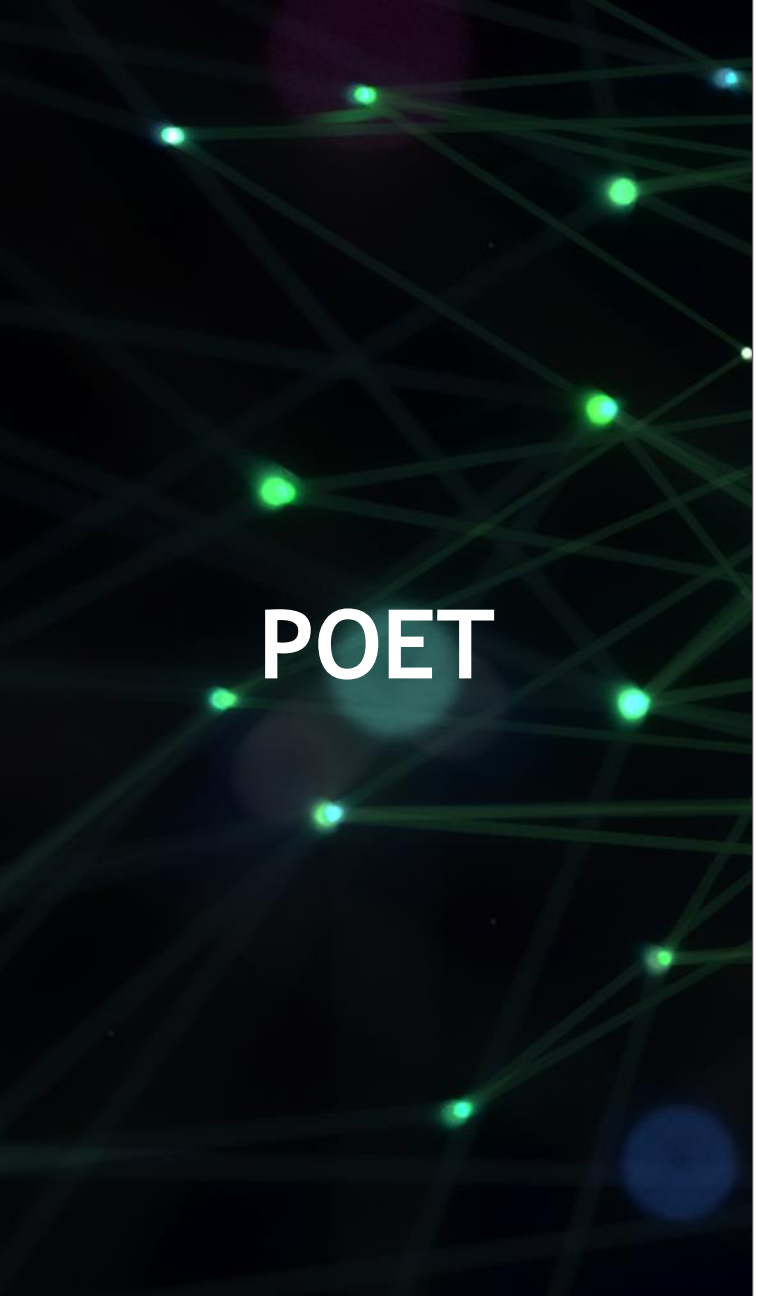


Larger image: http://www.asimovinstitute.org/neural-network-zoo/



www.Robust.ai

Gray Marcus, Rodney Brooks, Steven Pinker et al



Paired Open-Ended Trailblazer (POET): Endlessly Generating Increasingly Complex and Diverse Learning Environments and Their Solutions

Rui Wang **Joel Lehman** **Jeff Clune*** **Kenneth O. Stanley***
Uber AI Labs
San Francisco, CA 94103
ruiwang,joel.lehman,jeffclune,kstanley@uber.com
*co-senior authors

Abstract

While the history of machine learning so far encompasses a series of problems posed by researchers and algorithms that learn their solutions, an important question is whether the problems themselves can be generated by the algorithm at the same time as they are being solved. Such a process would in effect build its own diverse and expanding curricula, and the solutions to problems at various stages would become stepping stones towards solving even more challenging problems later in the process. The *Paired Open-Ended Trailblazer* (POET) algorithm introduced in this paper does just that: it pairs the generation of environmental challenges and the optimization of agents to solve those challenges. It simultaneously explores many different paths through the space of possible problems and solutions and, critically, allows these stepping-stone solutions to transfer between problems if better, catalyzing innovation. The term *open-ended* signifies the intriguing potential for algorithms like POET to continue to create novel and increasingly complex capabilities without bound. We test POET in a 2-D bipedal-walking obstacle-course domain in which POET can modify the types of challenges and their difficulty. At the same time, a neural network controlling a biped walker is optimized for each environment. The results show that POET produces a diverse range of sophisticated behaviors that solve a wide range of environmental challenges, many of which cannot be solved by direct optimization alone, or even through a direct, single-path curriculum-based control algorithm introduced to highlight the critical role of open-endedness in solving ambitious challenges. The ability to transfer solutions from one environment to another proves essential to unlocking the full potential of the system as a whole, demonstrating the unpredictable nature of fortuitous stepping stones. We hope that POET will inspire a new push towards open-ended discovery across many domains, where algorithms like POET can blaze a trail through their interesting possible manifestations and solutions.

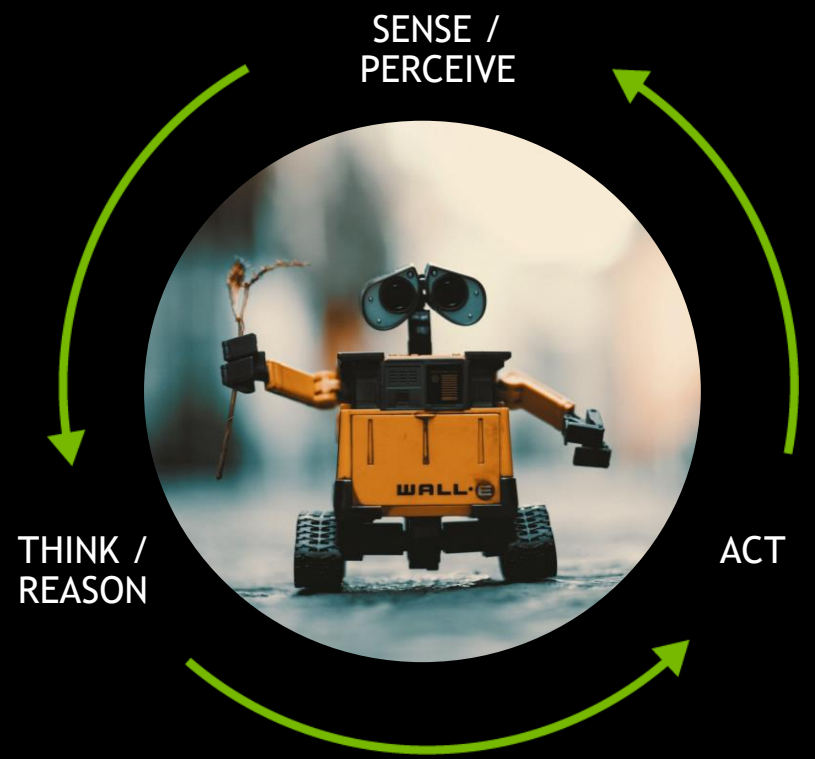
REINFORCEMENT LEARNING
“from our own mistakes”

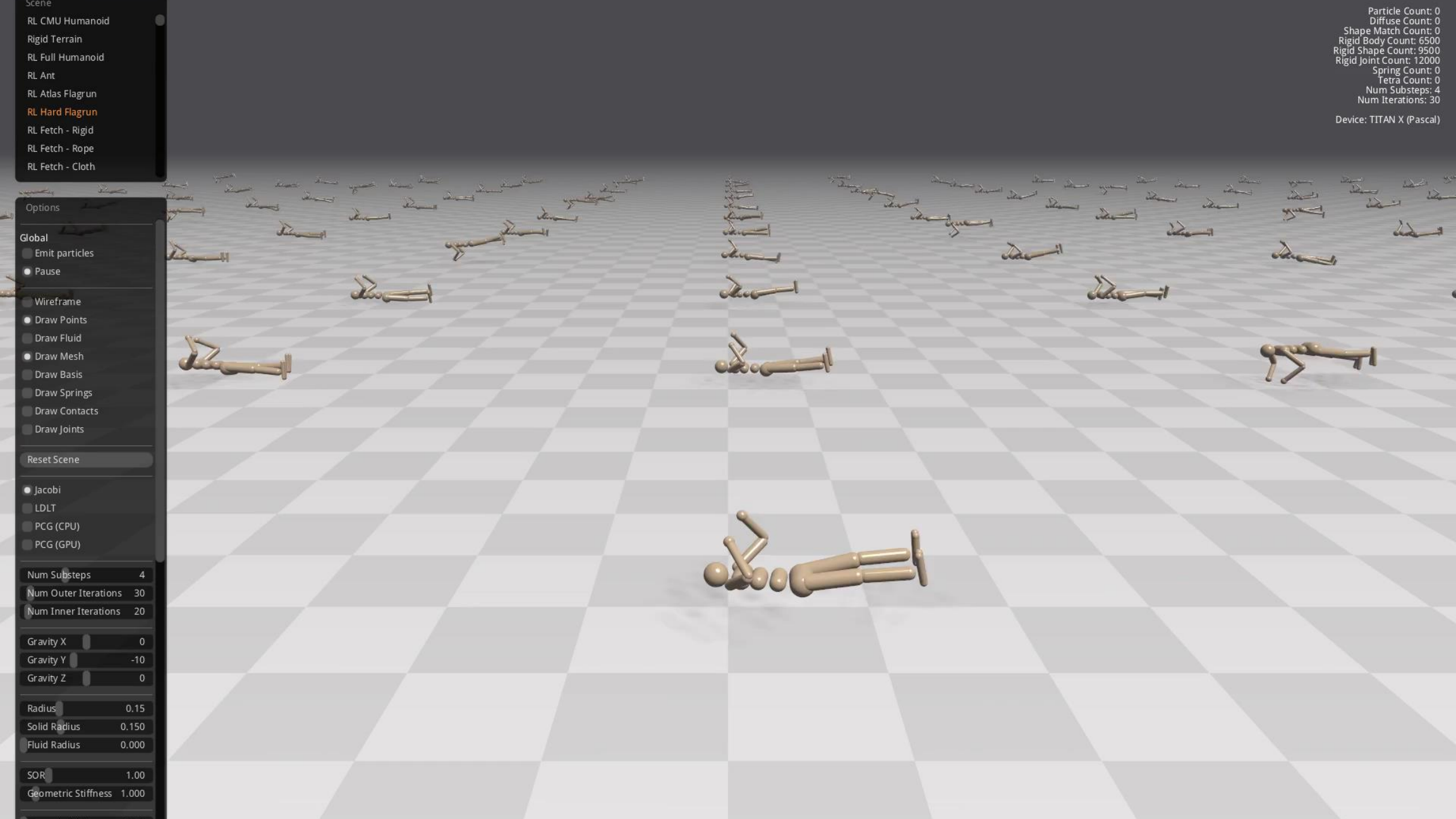

Move to Beacon
Collect Minerals and Gas
Collect Mineral Shards
Build Marines

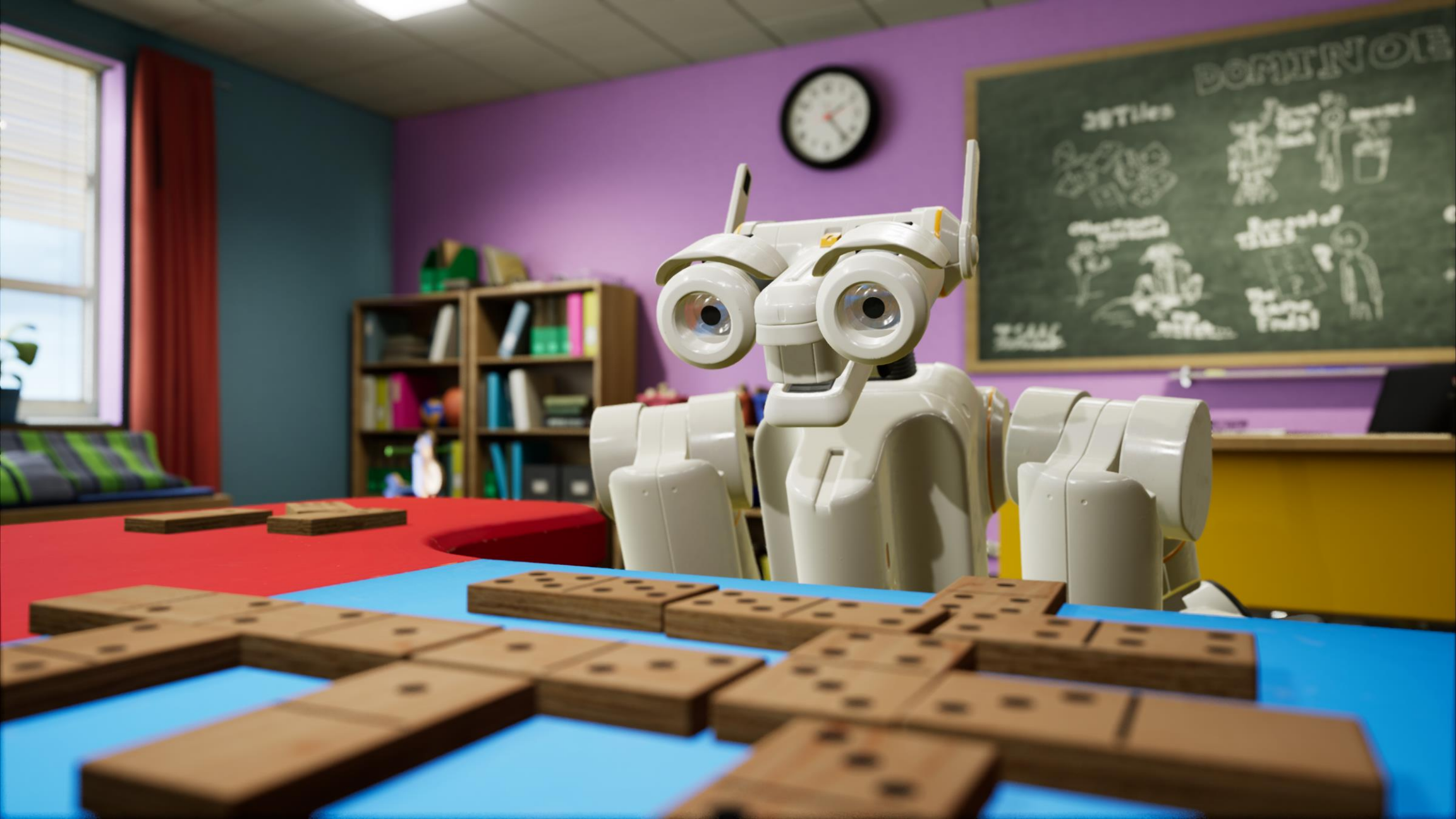
DEEPMIND ALPHA*

* ..a deeply structured hybrid (Gary Marcus, Jan 2018)

ROBOTS

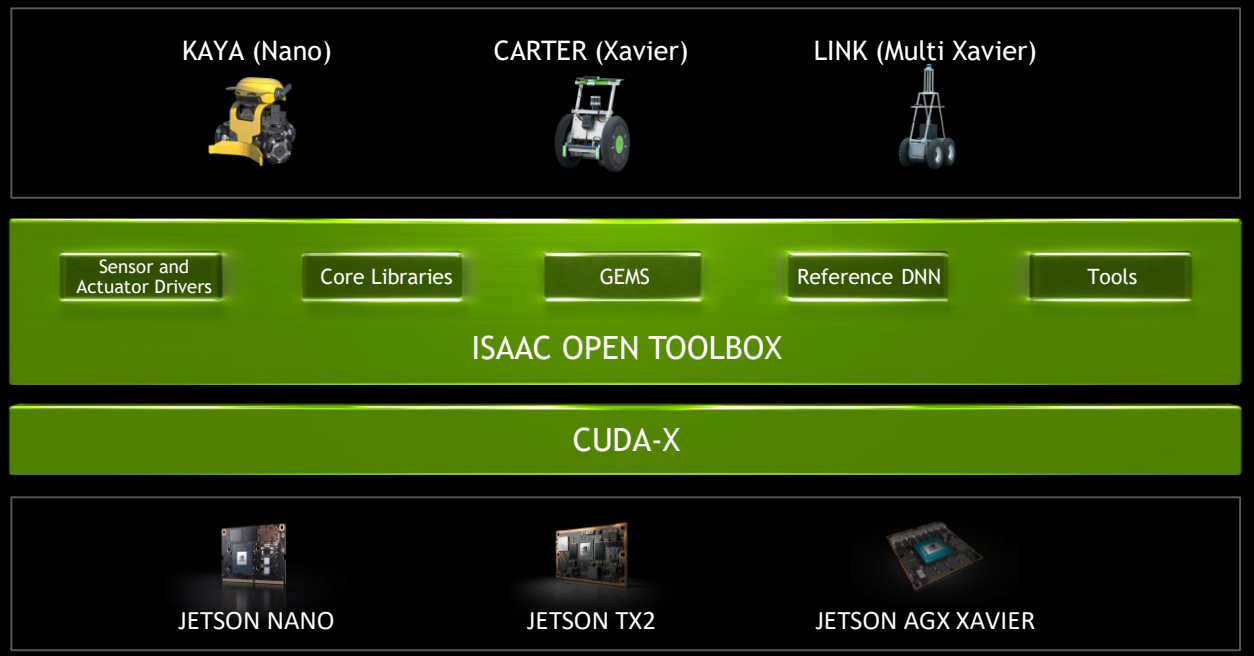



Scene
RL CMU Humanoid
Rigid Terrain
RL Full Humanoid
RL Ant
RL Atlas Flagrun
RL Hard Flagrun
RL Fetch - Rigid
RL Fetch - Rope
RL Fetch - Cloth
Particle Count: 0
Diffuse Count: 0
Shape Match Count: 0
Rigid Body Count: 6500
Rigid Shape Count: 9500
Rigid Joint Count: 12000
Spring Count: 0
Tetra Count: 0
Num Substeps: 4
Num Iterations: 30
Device: TITAN X (Pascal)
Options
Global
Emit particles
Pause
Wireframe
Draw Points
Draw Fluid
Draw Mesh
Draw Basis
Draw Springs
Draw Contacts
Draw Joints
Reset Scene
Jacobi
LDLT
PCG (CPU)
PCG (GPU)
Num Substeps 4
Num Outer Iterations 30
Num Inner Iterations 20
Gravity X 0
Gravity Y -10
Gravity Z 0
Radius 0.15
Solid Radius 0.150
Fluid Radius 0.000
SOR 1.00
Geometric Stiffness 1.000

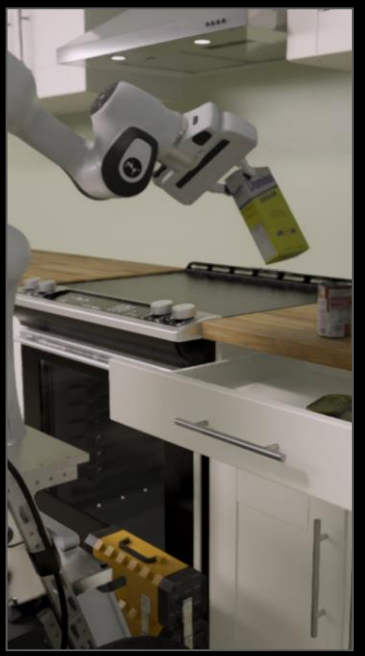


Imitation Learning

ANNOUNCING ISAAC OPEN SDK



Isaac Robot Engine



Isaac Sim

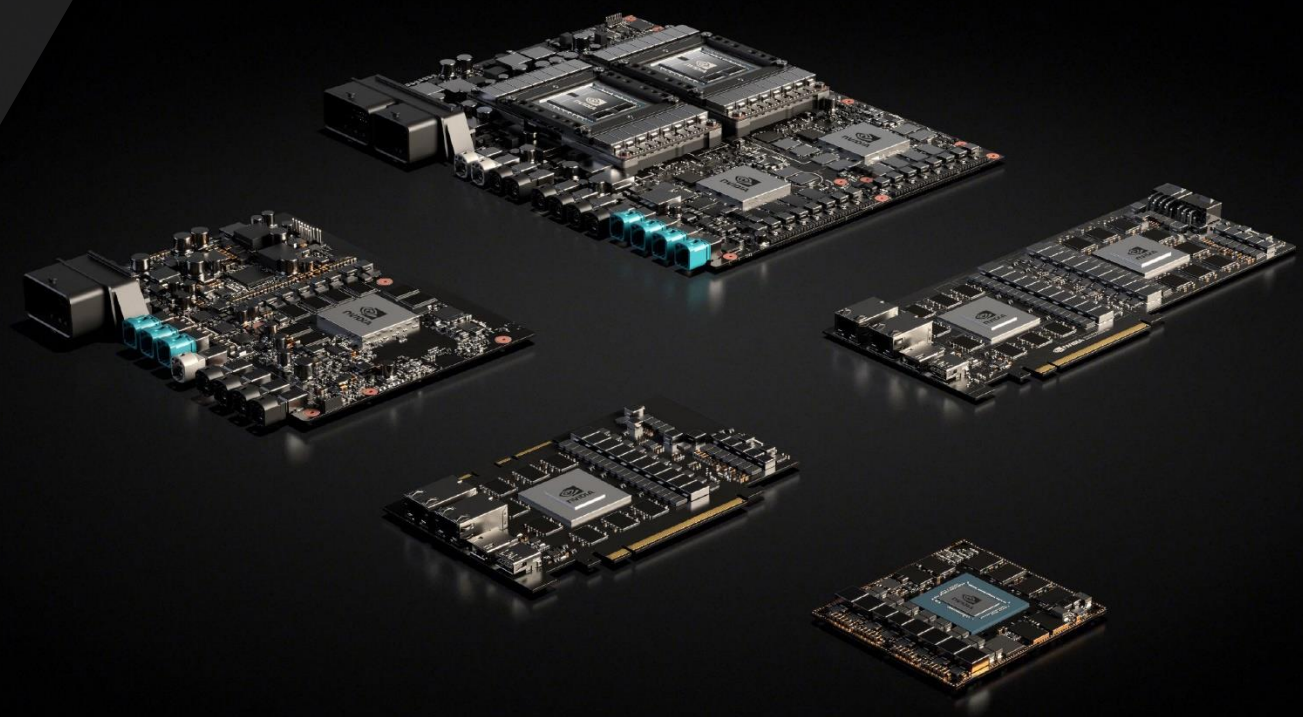
Isaac Gym

Isaac Robot Engine – Modular robot framework | Isaac Sim - Virtual robotics laboratory
Isaac Gym – Reinforcement learning simulator | Isaac Robot Apps – Kaya, Carter and Link
Available at developer.nvidia.com/isaac-sdk

NVIDIA AGX

Family of Systems for Embedded AI HPC

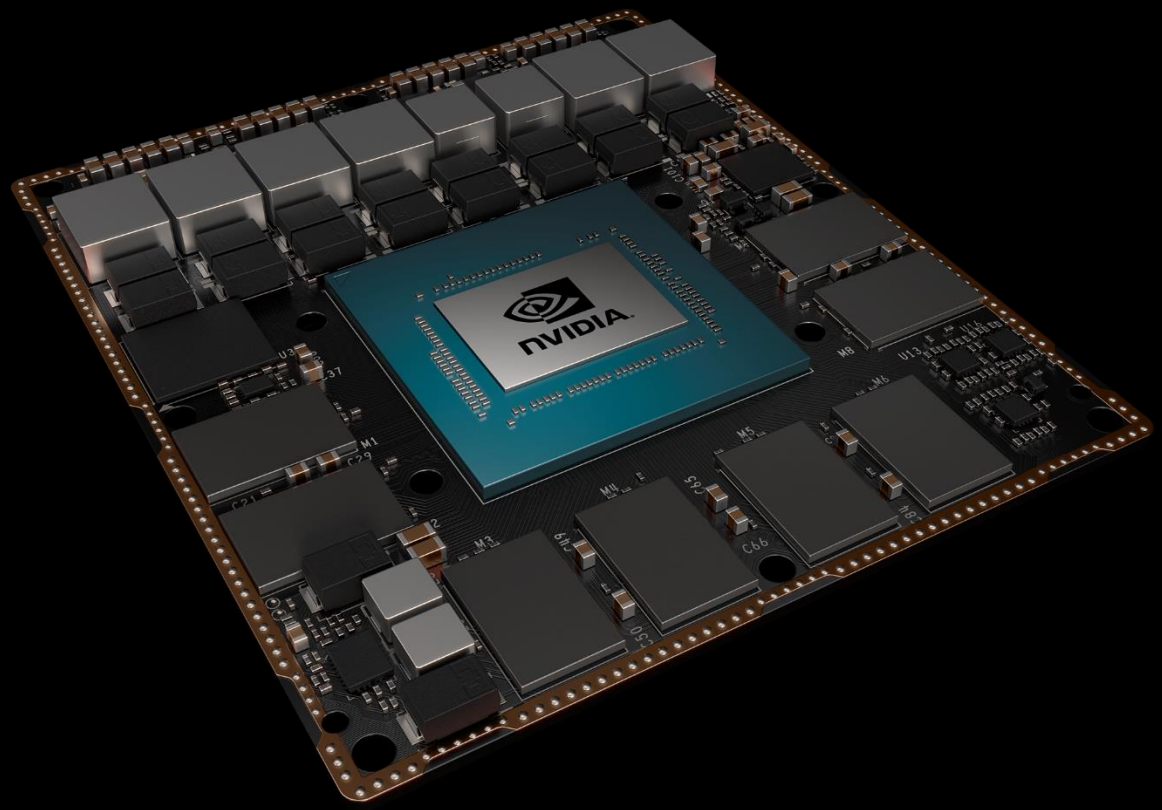
Self-driving cars
Robotics
Smart Cities
Healthcare



JETSON XAVIER

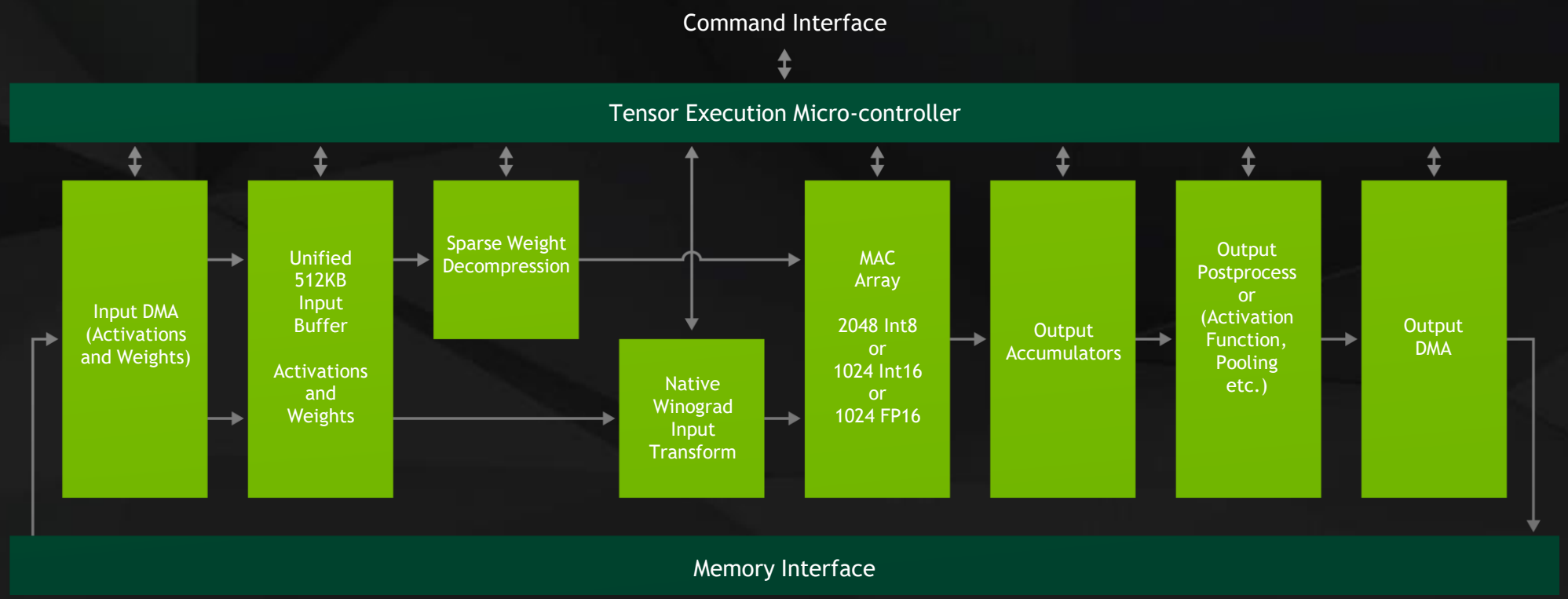
for Autonomous Machines

AI Server Performance in 30W • 15W • 10W
512 Volta CUDA Cores • 2x NVDLA
8 core CPU
32 DL TOPS



developr.nvidia.com/jetson-xavier

XAVIER DLA
NOW OPEN SOURCE



WWW.NVDLA.ORG

A massive Deep Learning challenge

COMPUTATIONAL SCALE REQUIRED

3 million labeled images
1 DGX-1 trains 300k labeled images on 1 DNN in 1 day
10 DNNs required for self-driving
10 parallel experiments at all times
100 DGX-1 per car


SELF-DRIVING
SAFETY REPORT
2018
http://bit.ly/NVSafetyReport

DRIVE CONSTELLATION

Virtual Reality AV Simulator

Runs DRIVE Sim Simulator

Hardware in the Loop System Level Simulator

Simulate Rare and Difficult Conditions

Scalable Platform | Data Center Solution

Timing Accurate and Bit Accurate


RESOLUTIONDEPENDENCY
INFERENCEPART-OF-SPEECH
NORMALIZATIONMACHINE
TAGGINGQUESTION
PARSINGSEMANTIC
EXTRACTIONLANGUAGE
LANGUAGE
NATURAL
NLP
NLU
RECOGNITIONCCG
DETECTIONSUMMARIZATIONTAXONOMY
ENTITY
SENSECONSTITUENCY
MODELINGLEXICAL
EXTRACTIONSEMANTIC
ROLE
SENSE
LEARNINGTEMPORAL
DISAMBIGUATION
ERROR
LANGUAGE
LABELINGSENTIMENT
SIMILARITYSEMANTIC
TRANSLATIONMULTI-TASK
PARSINGDIALOGUEDOMAIN
AUTOMATIC
CLASSIFICATIONWORD
PROCESSINGTEXT
ADAPTATIONENTITY
SUPERTAGGINGCOMMON
SYNTAXSIMPLIFICATIONSTANCE
ANALYSISSHALLOW PARSINGCOREFERENCE
TEXTUAL ANSWERINGRELATION
LINKINGGRAMMATICAL
SPEECH
PREDICTIONRELATIONSHIP
RECOGNITIONNATURAL
CORRECTIONINFORMATION
LEARNINGMULTI-MODALNAMED

Detection

Planning

Acceleration

Assimilation

Enhancement

Parametrization

Prediction

Augmentation

Monitor Environmental Change

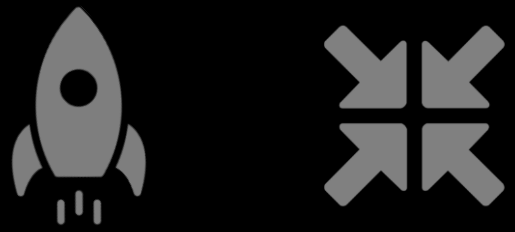
drought flooding deforestation urbanification melting glaciers sea-level change

Detection

Planning

Acceleration

Assimilation



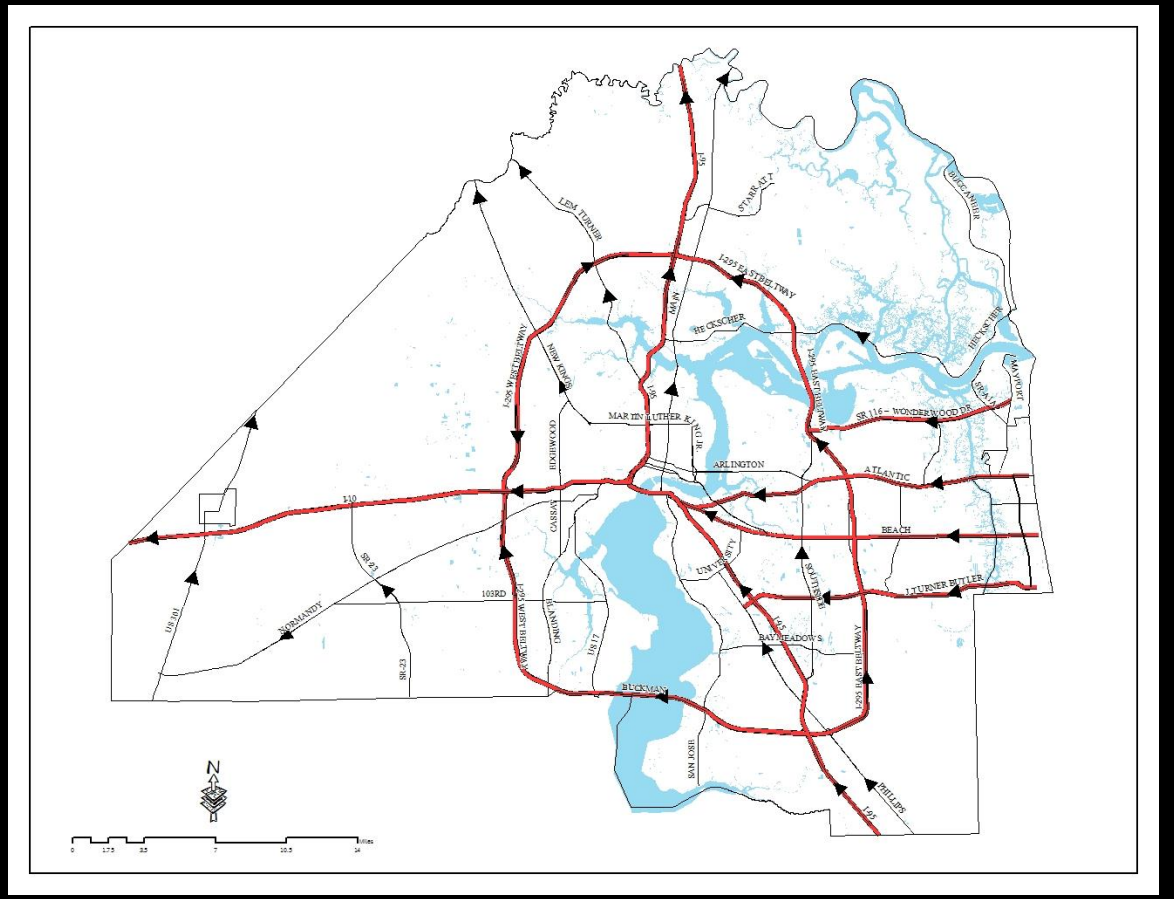
Enhancement

Parametrization

Prediction

Augmentation

Optimize Disaster Planning



Detection

Planning

Acceleration

Assimilation

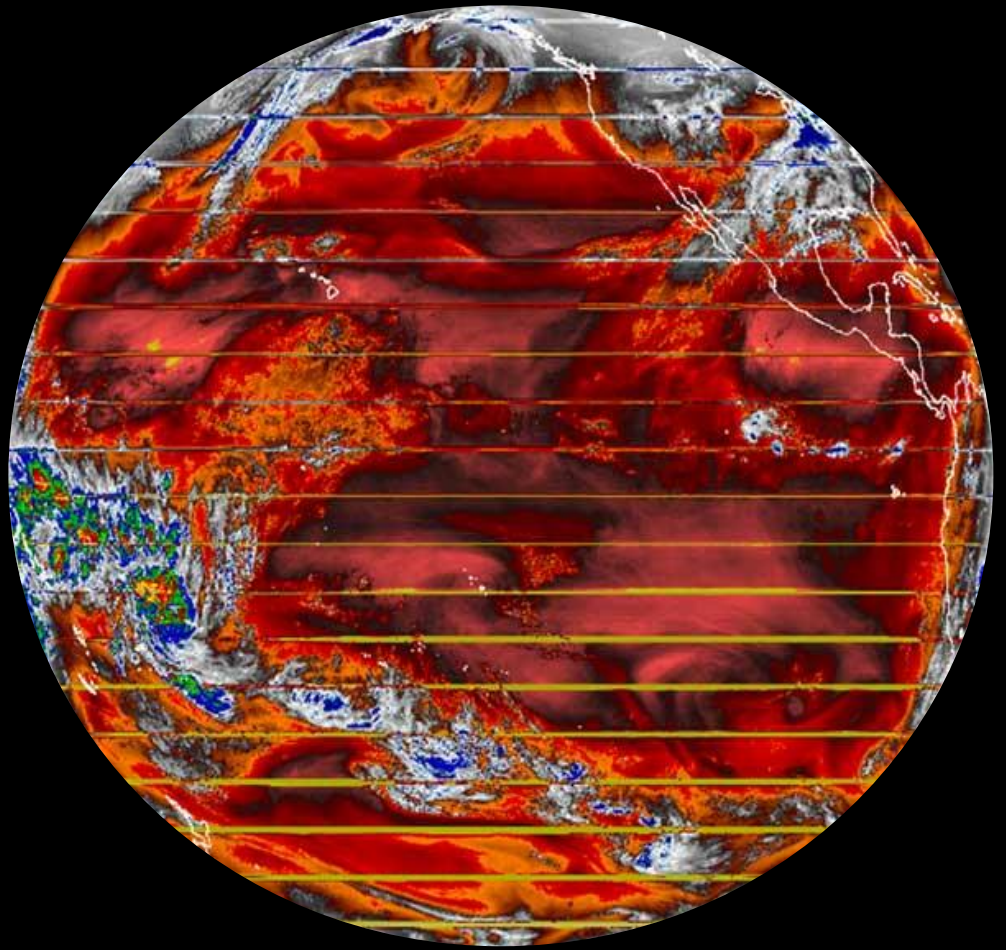
Enhancement

Parametrization

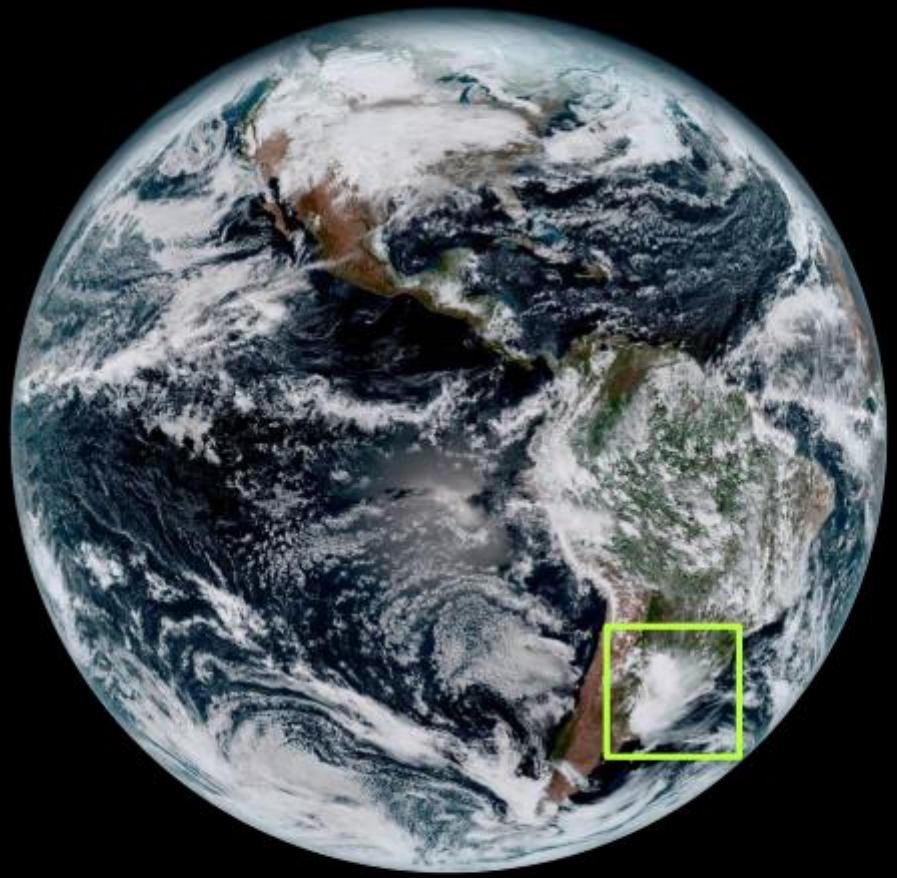
Prediction

Augmentation

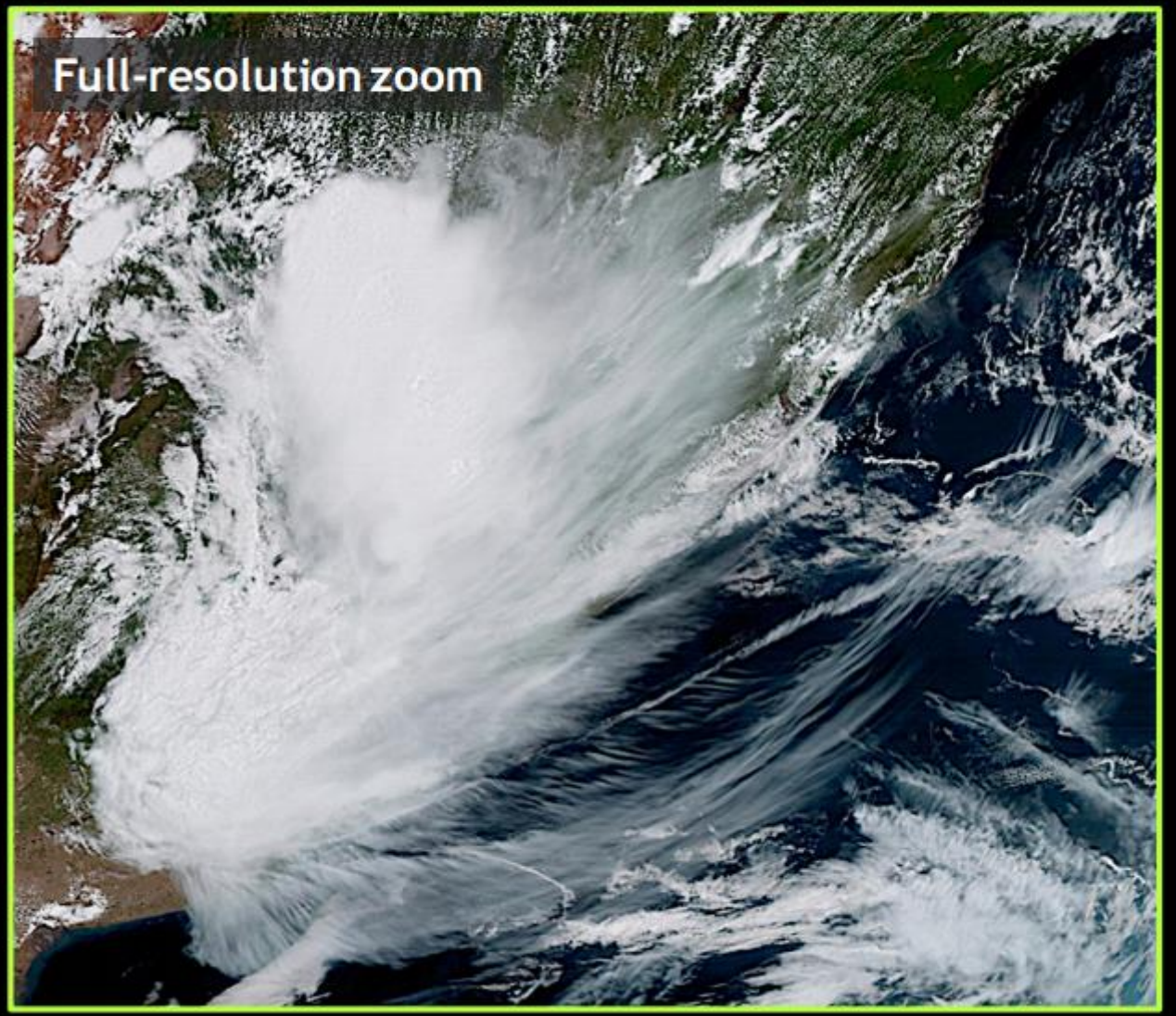
Use Inpainting to Repair Damaged GOES-17 Observations



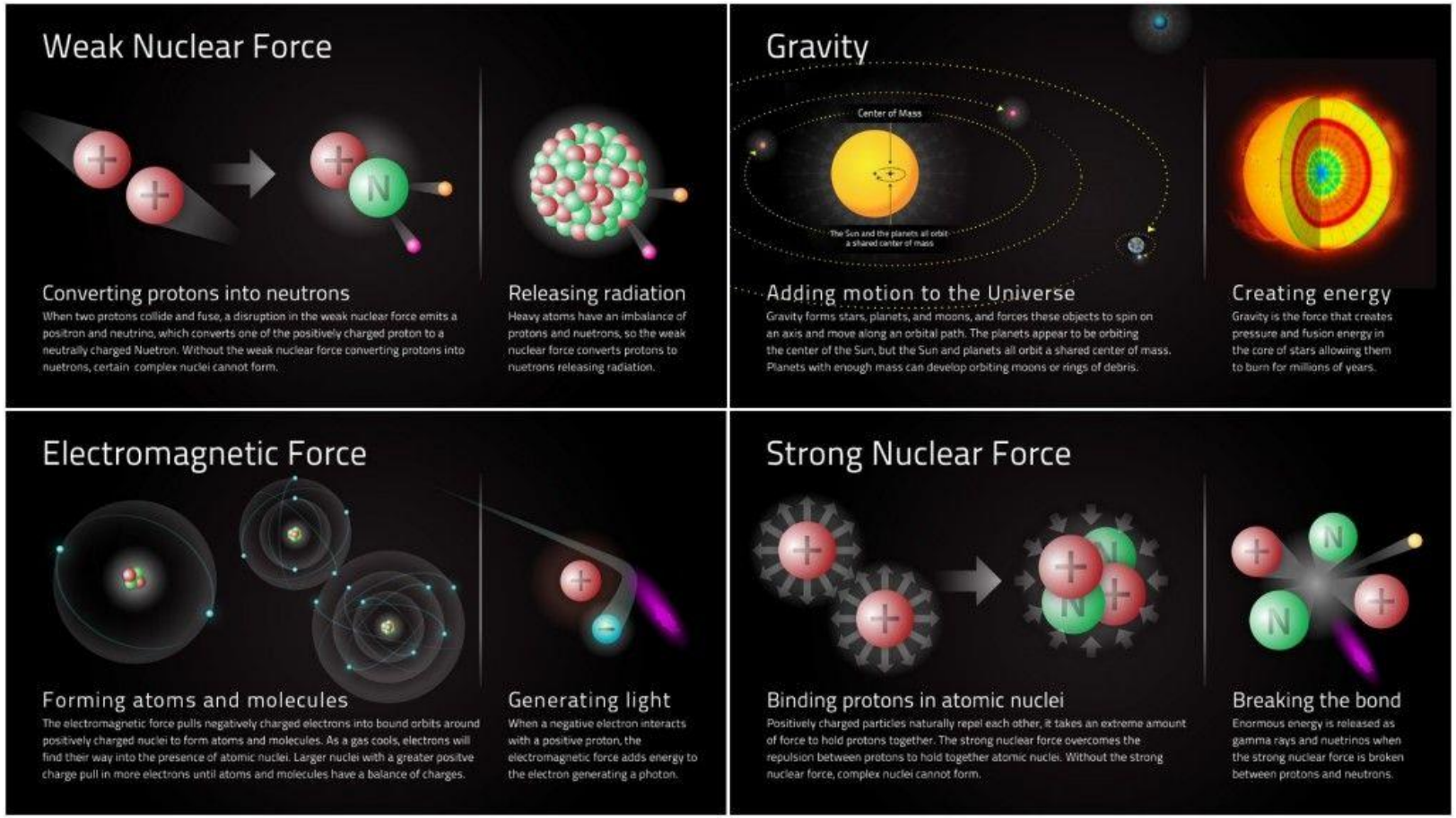
TRAIN AN ALGORITHM TO EXAMINE EVERY PIXEL



GOES-16: 4k x 4k x 11 channels



THE FOUR FUNDAMENTAL FORCES OF THE UNIVERSE

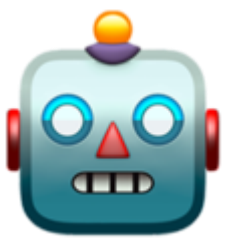


@bhilburn

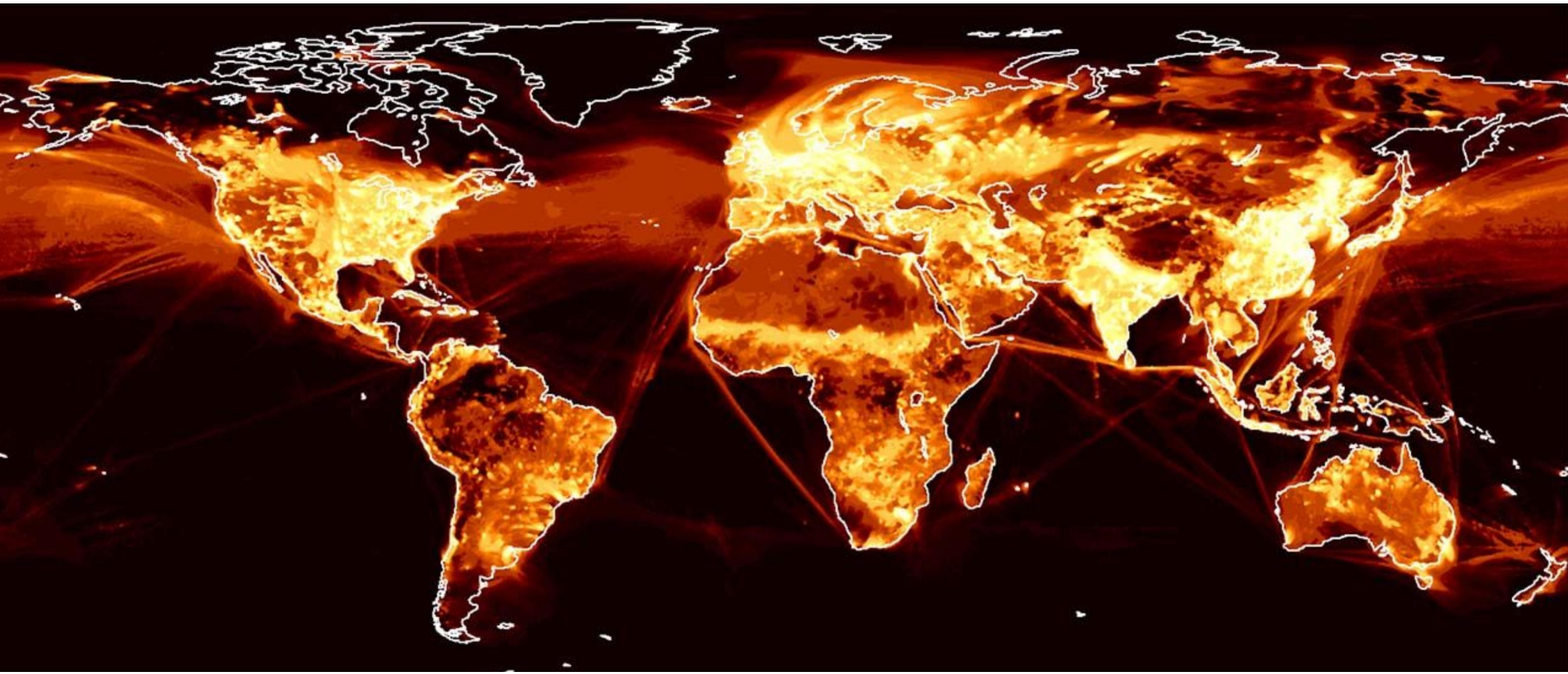
https://github.com/rapidsai/cusignal

TWO FUNDAMENTAL NEEDS

Fast filtering, FFTs, correlations, convolutions, resampling, etc to process increasingly larger bandwidths of signals at increasingly fast rates and do increasingly cool stuff we couldn’t do before

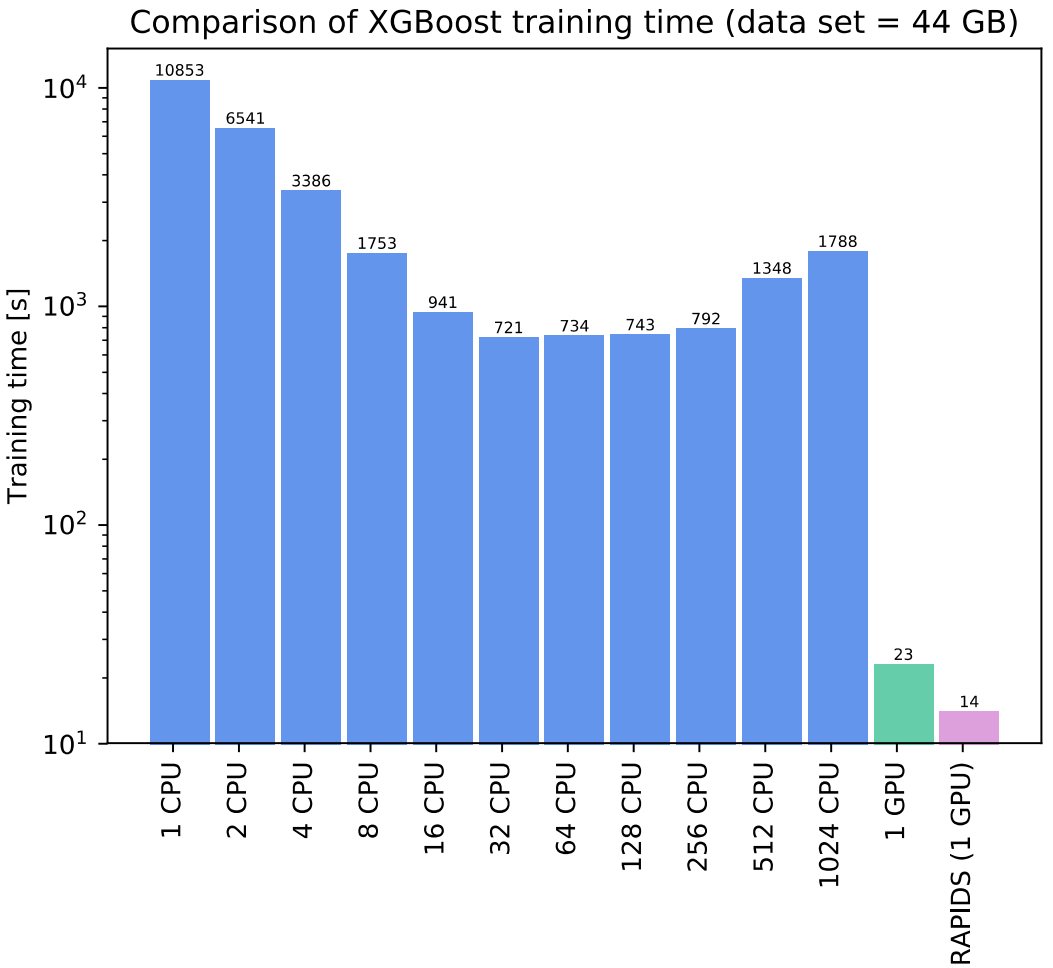


Artificial Intelligence techniques applied to spectrum sensing, signal identification, spectrum collaboration, and anomaly detection



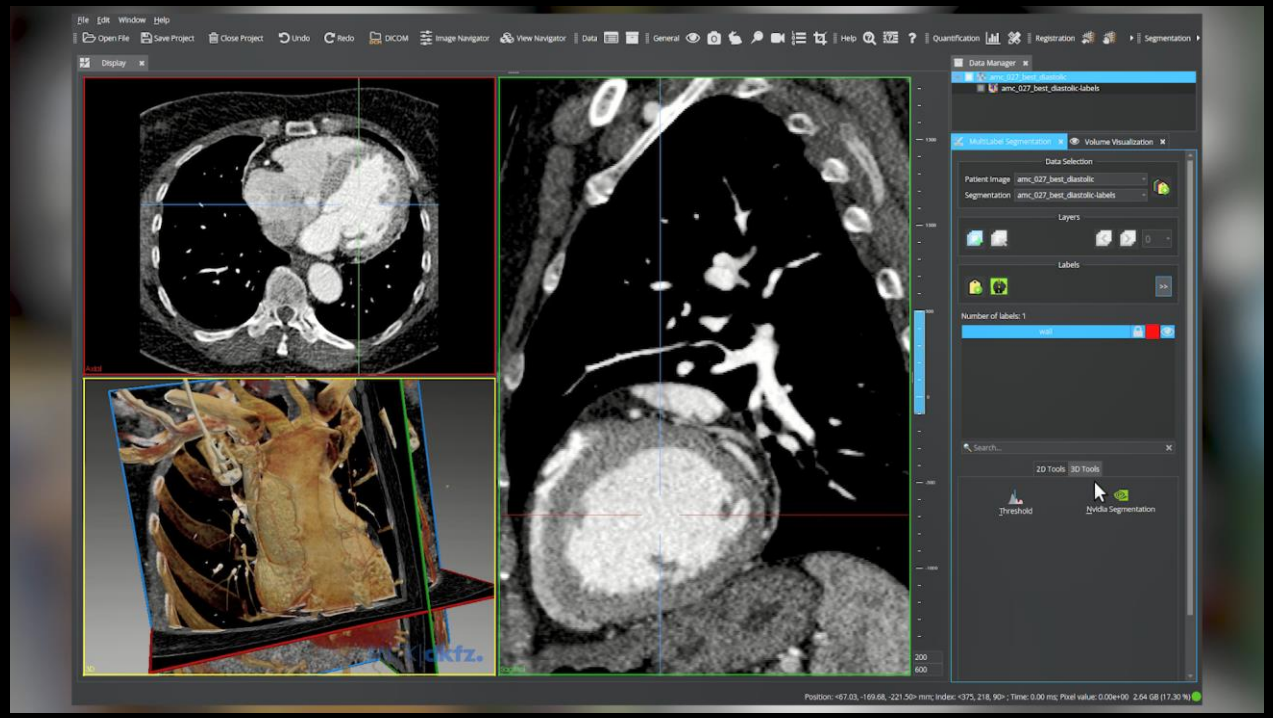
https://blogs.nvidia.com/blog/2019/11/07/nasa-rapids-air-quality-forecasts/

XGBoost for simulating atmospheric chemistry



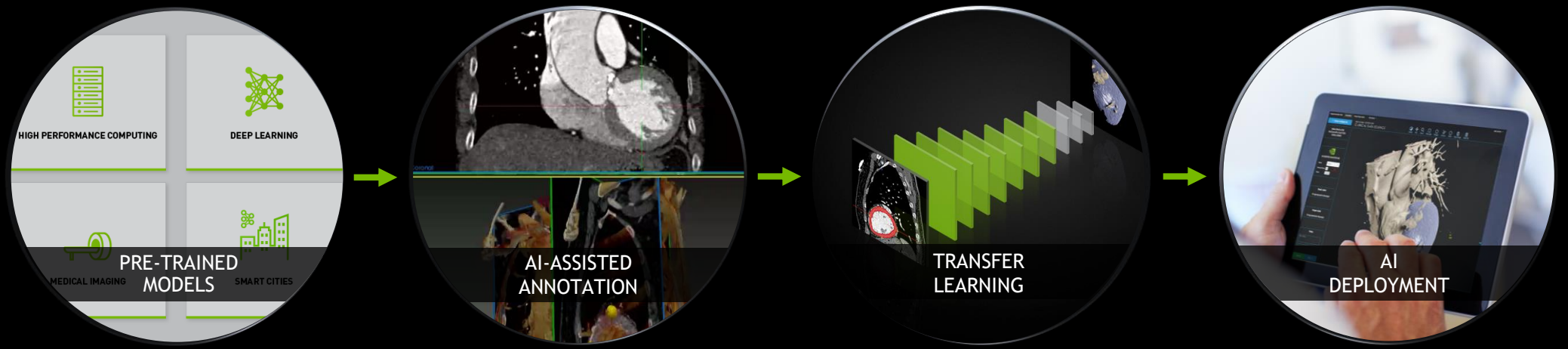
christoph.a.keller@nasa.gov

AI-ASSISTED ANNOTATION



CLARA AI TOOLKIT

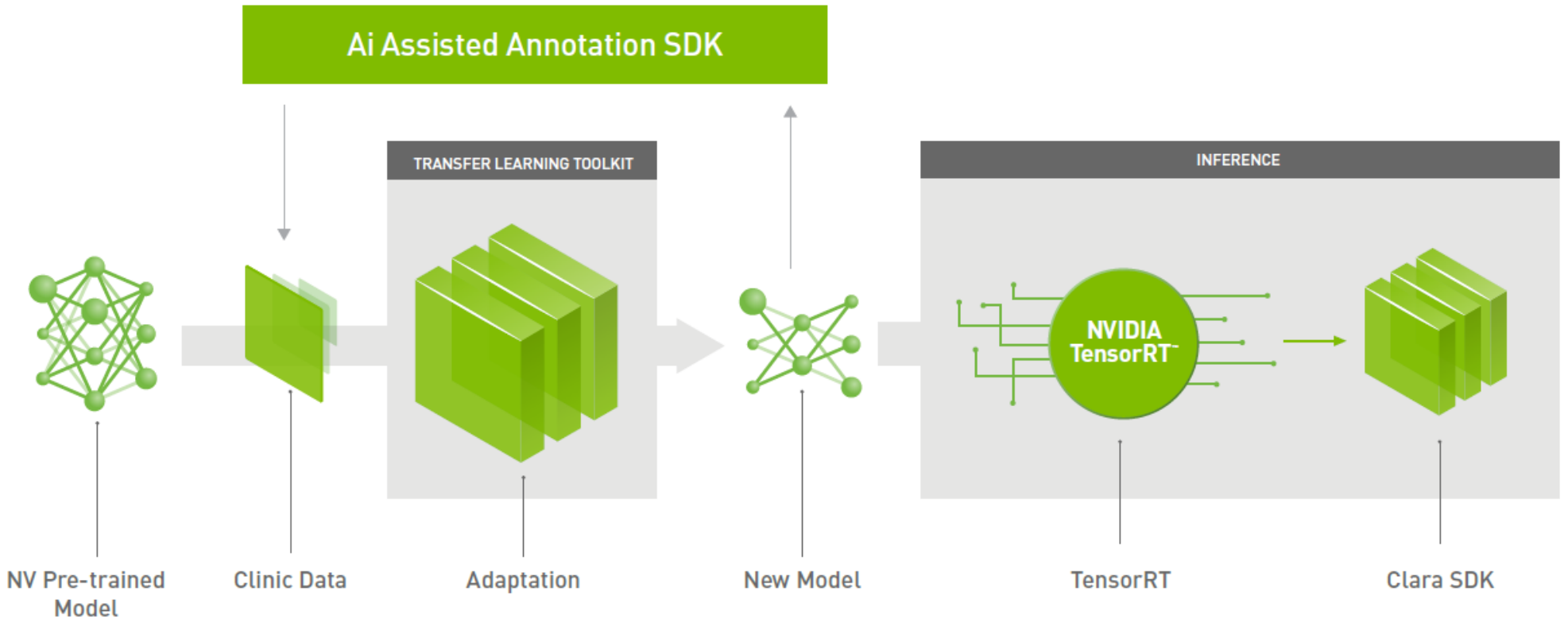
Build, Manage And Deploy AI Applications For Radiology



AI-Assisted Annotation – Hours to Minutes | 10x Less Training Data Needed | 13 Pre-Trained Models
Reference Training and Deployment Pipelines | Available at developer.nvidia.com/clara

End to End NVIDIA Deep Learning Workflow

Pre-Trained models * Annotation Assistant * Training & adaptation * Applications ready to integrate with Clara Platform



CONVERGENCE OF HPC AND AI

Integrating the Third and Fourth Pillars of Scientific Discovery

HPC

40+ years of algorithms based on first principles theory

AI

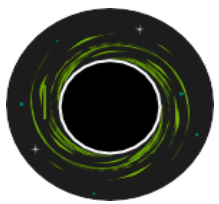
New algorithms and models with potential to increase model size and accuracy

Dramatically Improves Accuracy and /or Time-to-Solution at Large Scale

Commercially viable fusion energy

Improve or validate the Standard Model of Physics

Clinically viable precision medicine

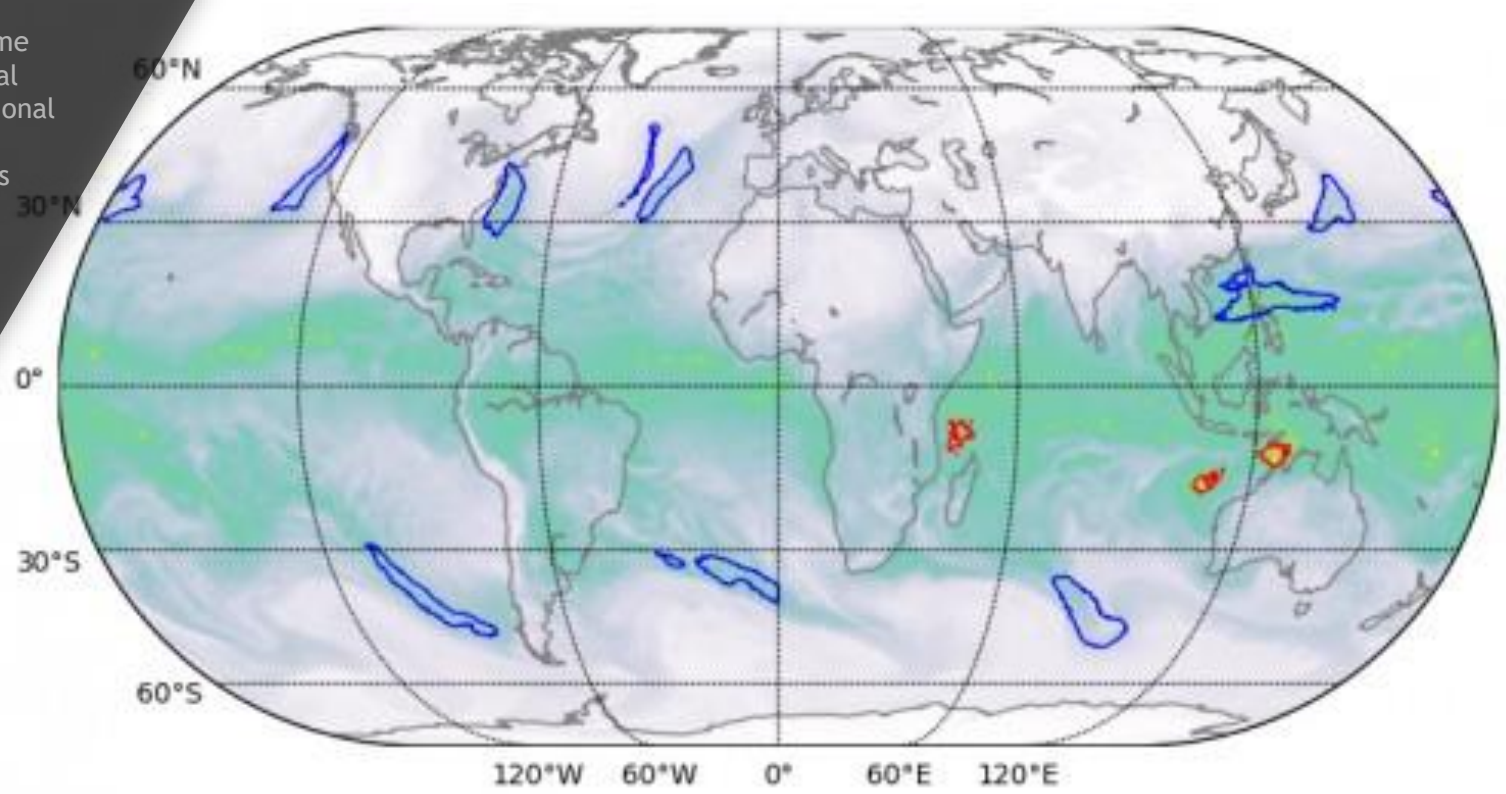


Understanding cosmological dark energy and matter

Climate/weather forecasts with ultra-high fidelity

EXASCALE AI FOR CLIMATE PREDICTION

The ability to accurately predict the path of extreme weather systems can save lives and safeguard global economies. Researchers at Lawrence Berkeley National Laboratory used a climate dataset on the Summit supercomputer with NVIDIA Volta Tensor Core GPUs to train a deep neural network to identify extreme weather patterns from high-resolution climate simulations. They achieved a performance of 1.13 exaflops, the fastest deep learning algorithm reported.



Pictured: high-quality segmentation results produced by deep learning on climate datasets.
Image credit: NERSC

FIVE ROADS TO GPU COMPUTING

GPU Libraries

Drop-in replacement for existing libraries

cuBLAS, CUDA Math, cuSPARSE, cuRAND, cuSOLVER, nvGRAPH, cuDNN, cuFFT, Thrust

OPEN-ACC

Comment-based directives in C / C++ / Fortran

Single source code parallelization for multiple architectures

CUDA

Parallel Programming Model for GPUs in C, C++, Fortran, Python, MATLAB

Specialized Kernels for general purpose GPU

RAPIDS

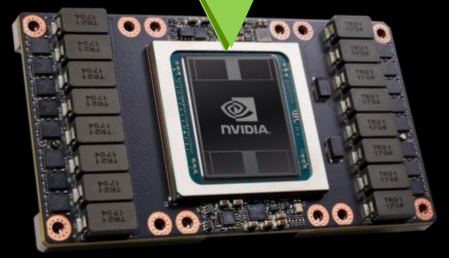
GPU Acceleration of Traditional Machine Learning

Accelerate Scikit-Learn style ML algorithms

DEEP LEARNING

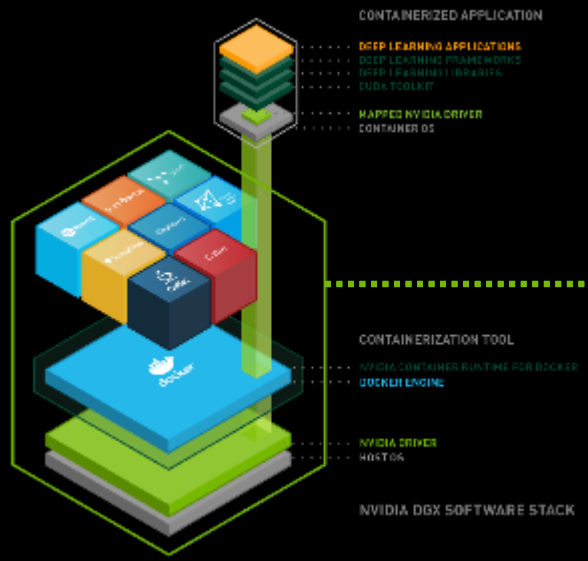
GPU accelerated deep learning frameworks TensorFlow, Pytorch

Build GPU-accelerated functions directly from data



PURPOSE-BUILT AI SUPERCOMPUTERS

NGC DL SOFTWARE STACK



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

AI WORKSTATION

DGX Station

The Personal AI Supercomputer

AI DATA CENTER

DGX-1

The Essential Instrument for AI Research

DGX-2

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

ResNet-50 for Classification

Remove from Favorites | Download Latest Model

Publisher	Application	Version	Modified	Size
NVIDIA	Classification	4.0.4	11/21/2018 03:27 PM	93.02 GB

Training Framework	Inference Framework	Model Format	Precision	GPU Model
TensorFlow	TensorRT	TRTPlan	INT8	v100

Description

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam viverra volutpat ipsum dignissim ve egestas. Maecenas egestas vestibulum erat, eu dapibus purus tempus. Fusce ornare molestie tortor, sed eleifend nisi vulputate vel. Sed semper ornare lacinia.

Labels

classification | fp32 | gpu-optimized | image classification | resnet-50 | tensorflow | tensorrt | trtplan | v100

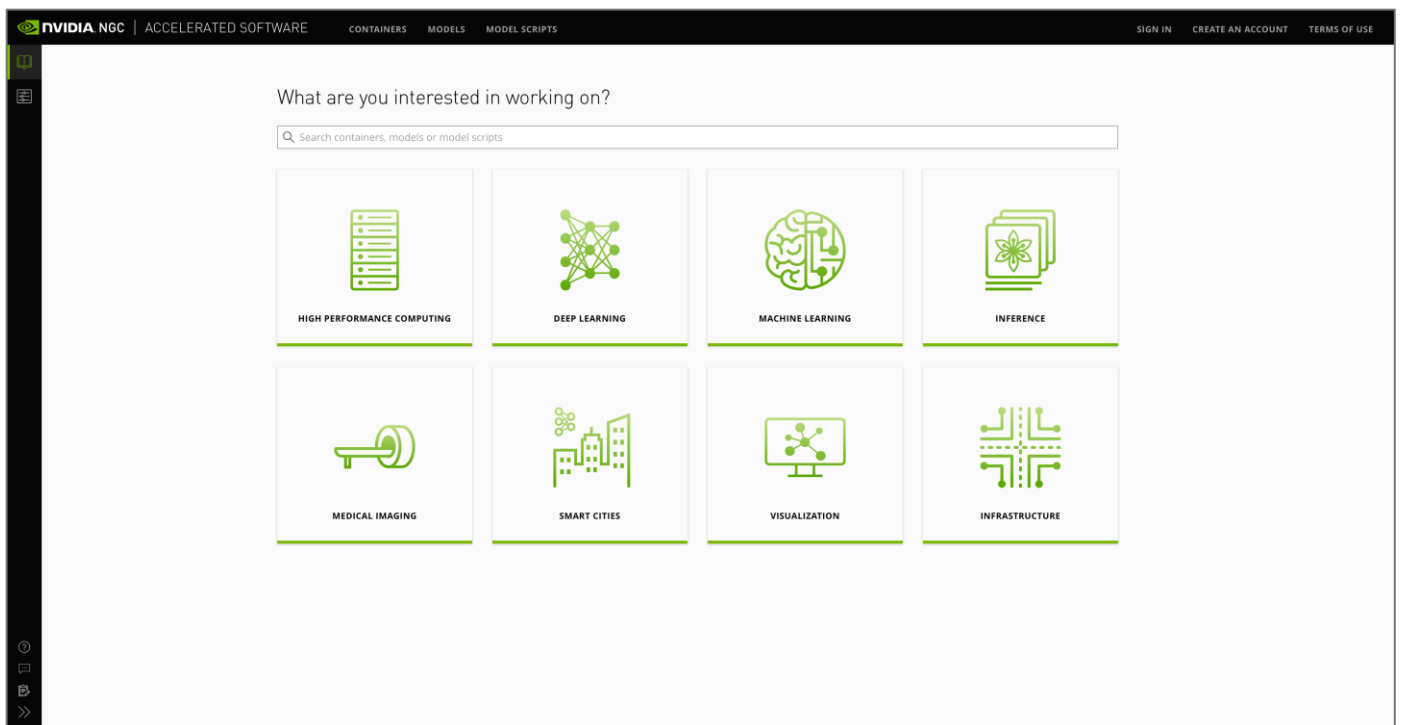
Overview | Version History | File Browser | Release Notes | Related Model Scripts

Classification with ResNet-50
Caffe | FP16, INT8
4.0.4
built by NVIDIA 06/25/2018

Classification with ResNet-50
TensorFlow | FP16, INT8
4.0.4
built by NVIDIA 06/25/2018

GET STARTED WITH NGC

Explore the NGC Registry for DL, ML & HPC



Deploy containers:
ngc.nvidia.com

Learn more about NGC offering:
nvidia.com/ngc

Technical information:
developer.nvidia.com

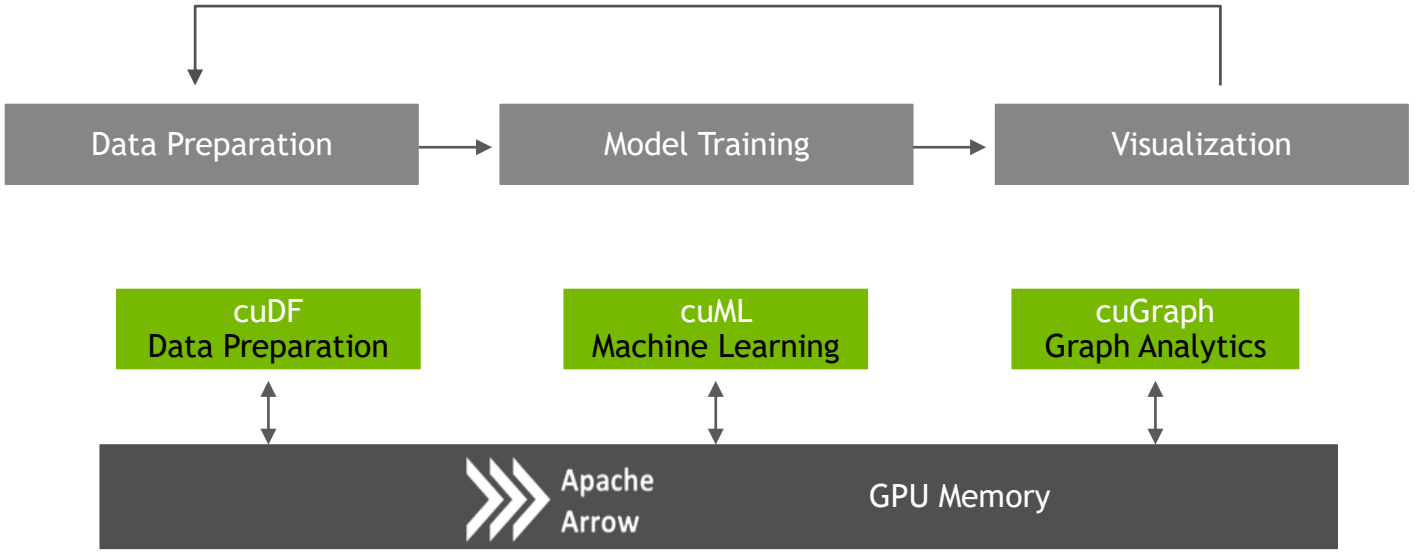
RAPIDS

RAPIDS

GPU Accelerated End-to-End Data Science

RAPIDS is a set of open source libraries for GPU accelerating **data preparation** and **machine learning.**

OSS website: rapids.ai



NVIDIA SDK

The Essential Resource for GPU Developers

DEEP LEARNING

Deep Learning SDK

High-performance tools and libraries for deep learning

AUTONOMOUS VEHICLES

NVIDIA DRIVE Platform

Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

VIRTUAL REALITY

NVIDIA VRWorks™

A comprehensive SDK for VR headsets, games and professional applications

GAME DEVELOPMENT

NVIDIA GameWorks™

Advanced simulation and rendering technology for game development

ACCELERATED COMPUTING

NVIDIA ComputeWorks

Everything scientists and engineers need to build GPU-accelerated applications

DESIGN & VISUALIZATION

NVIDIA DesignWorks™

Tools and technologies to create professional graphics and advanced rendering applications

AUTONOMOUS MACHINES

NVIDIA JetPack™

Powering breakthroughs in autonomous machines, robotics and embedded computing

SMART CITIES

NVIDIA Metropolis

Edge-to-cloud development platform for smart cities

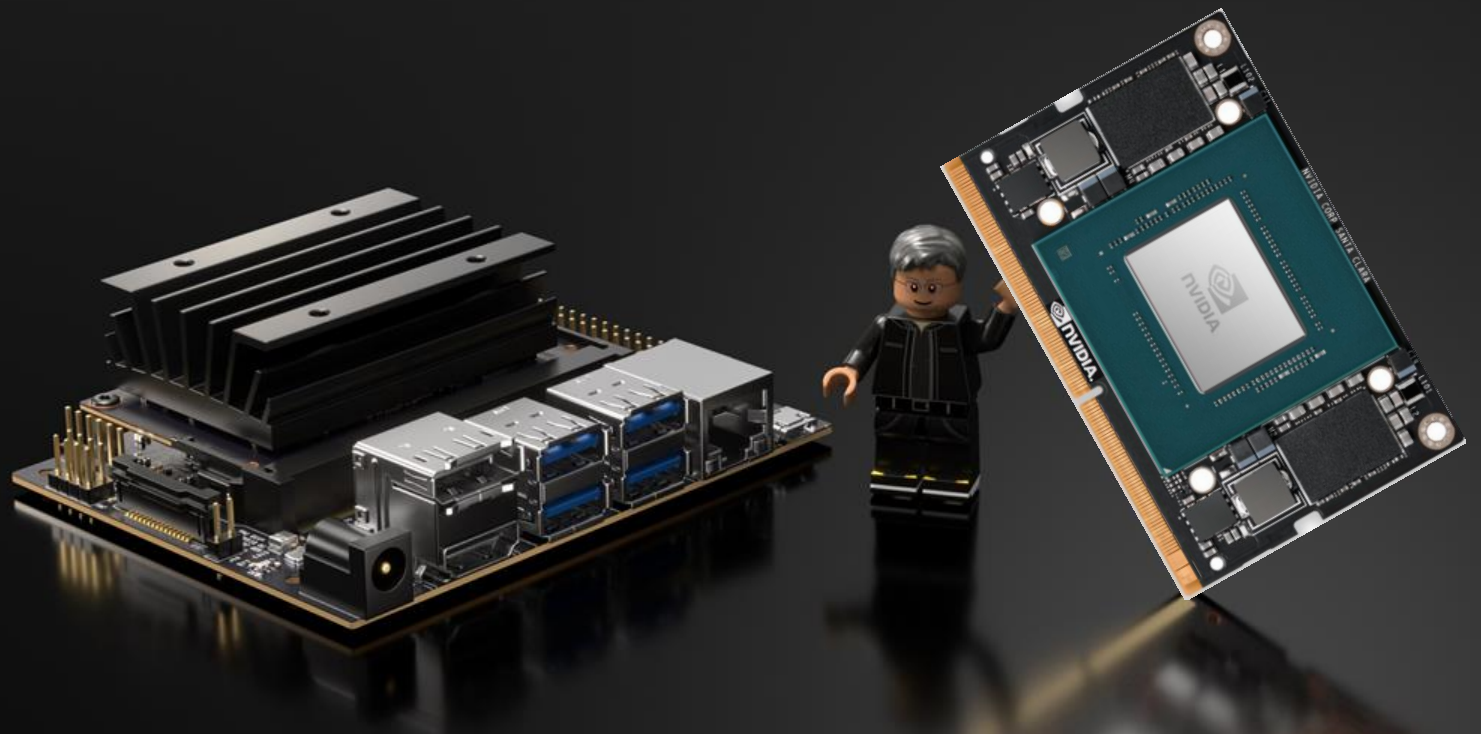
developer.nvidia.com

NVIDIA DGX-2

2 PFLOPS | 512GB HBM2 | 10 kW | 350 lbs

JETSON NANO DEVKIT & XAVIER NX SOM

Up to 21 DL TOPS (15w) | NX: 8 GB Memory | 45x70mm

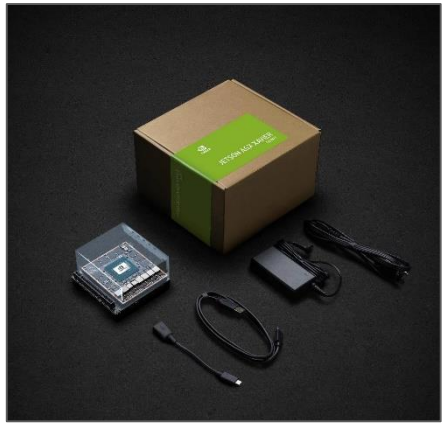


CUDA-X acceleration stack | High-resolution sensor support | Runs all CUDA-X AI models

NX available from nvidia.com and distributors worldwide in March 2020

DEEPSTREAM ON JETSON NANO

JETSON - START NOW



JETSON DEVELOPER KIT

AGX Xavier Developer Kit $699

Xavier NX software patch

developer.nvidia.com/buy-jetson

TWO DAYS TO A DEMO

Create your first demo today

developer.nvidia.com/embedded/twodaystoademo

DEEP LEARNING INSTITUTE

Training • Labs

Nanodegrees

nvidia.com/DLI

GTC

Largest event for GPU developers

gputechconf.com

NVIDIA DEEP LEARNING INSTITUTE

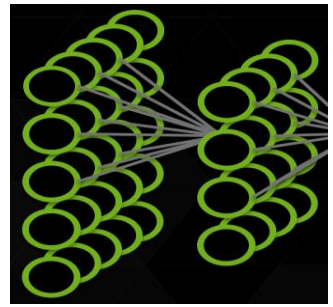
Online self-paced labs and instructor-led workshops on deep learning and accelerated computing

Take self-paced labs at **www.nvidia.co.uk/dlilabs**

View upcoming workshops and request a workshop onsite at **www.nvidia.co.uk/dli**

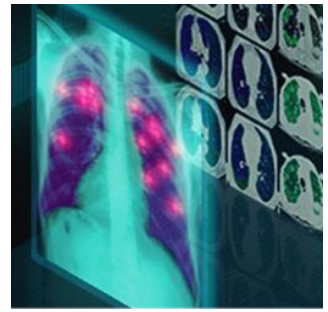
Educators can join the University Ambassador Program to teach DLI courses on campus and access resources. Learn more at **www.nvidia.com/dli**

Caffe2 · Microsoft Cognitive Toolkit · mxnet · TensorFlow · PYTORCH

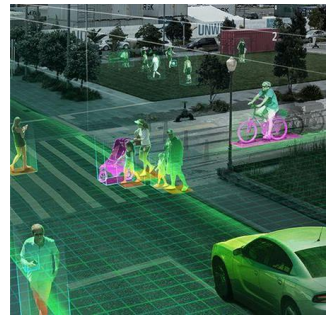


Fundamentals

Autonomous Vehicles



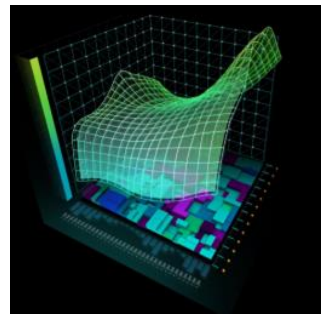
Healthcare



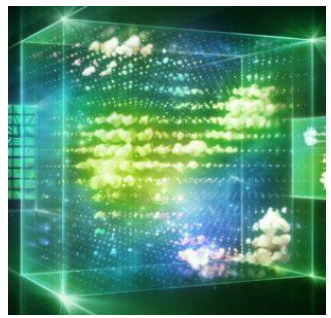
Intelligent Video Analytics

Robotics

Game Development & Digital Content



Finance



Accelerated Computing

Virtual Reality

NVIDIA DEEP LEARNING INSTITUTE

NVIDIA

GTC GPU TECHNOLOGY CONFERENCE

March 22-26, 2020 | San Jose, CA

CONNECT

Connect with experts from NVIDIA, GE Healthcare, NSF Carnegie Mellon, Google, and other leading organizations

LEARN

Gain insight and valuable hands-on training through over 100 sessions

DISCOVER

See how GPU technologies are creating amazing breakthroughs in important fields such as deep learning

INNOVATE

Explore disruptive innovations that can transform your work

Join us | Use VIP code NVALOWNDES for 25% off

Don’t miss the premier AI conference.

nvidia.com/en-us/gtc/

alowndes@nvidia.com