Fuelling the AI Revolution with Gaming

ALISON B LOWNDES

AI DevRel | EMEA

@alisonblowndes

March 2018

The day job

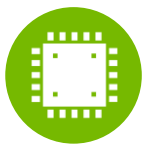
AUTOMOTIVE
Auto sensors reporting location, problems

COMMUNICATIONS
Location-based advertising

CONSUMER PACKAGED GOODS
Sentiment analysis of what's hot, problems

FINANCIAL SERVICES
Risk & portfolio analysis
New products

EDUCATION & RESEARCH
Experiment sensor analysis



HIGH TECHNOLOGY / INDUSTRIAL MFG.
Mfg. quality
Warranty analysis

LIFE SCIENCES

MEDIA/ENTERTAINMENT
Viewers / advertising effectiveness

ON-LINE SERVICES / SOCIAL MEDIA
People & career matching

HEALTH CARE
Patient sensors, monitoring, EHRs

OIL & GAS
Drilling exploration sensor analysis

RETAIL
Consumer sentiment

TRAVEL & TRANSPORTATION
Sensor analysis for optimal traffic flows

UTILITIES
Smart Meter analysis for network capacity,

LAW ENFORCEMENT & DEFENSE
Threat analysis - social media monitoring, photo analysis

www.FrontierDevelopmentLab.org

NASA FRONTIER DEVELOPMENT LAB

SETI INSTITUTE

NASA

esa

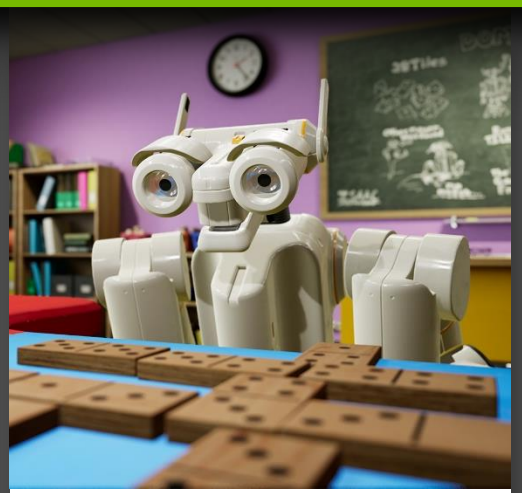
NVIDIA

IBM

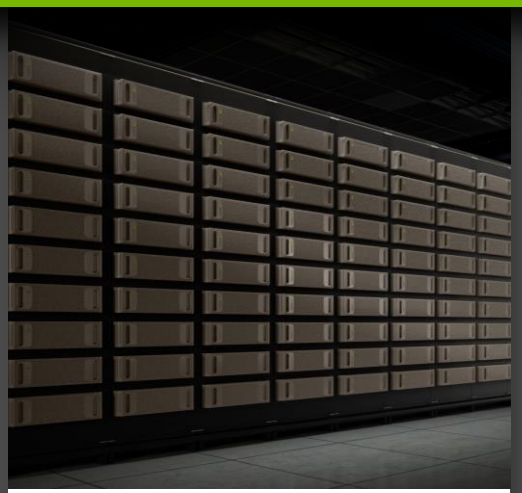
LOCKHEED MARTIN

NVIDIA

Gaming



VR



AI & HPC

Self-Driving Cars

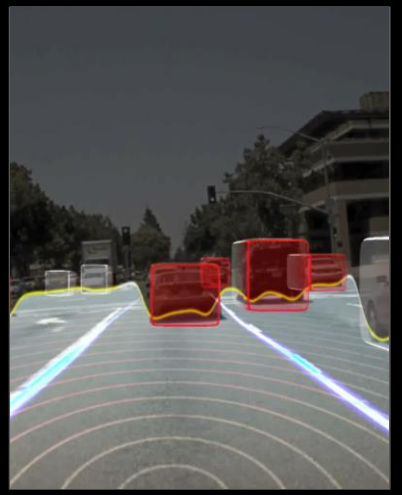
GPU Computing

THE ERA OF AI

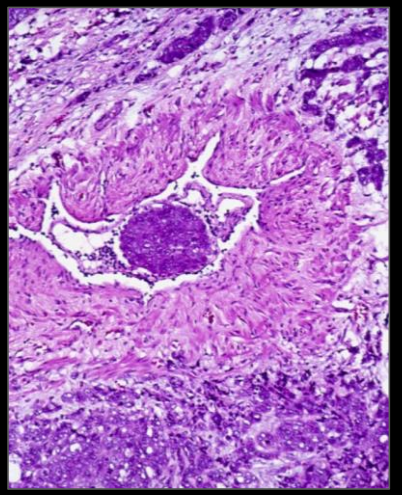
The Defining Technology of Our Generation

WEB SERVICES

INTELLIGENT MACHINES



HEALTHCARE

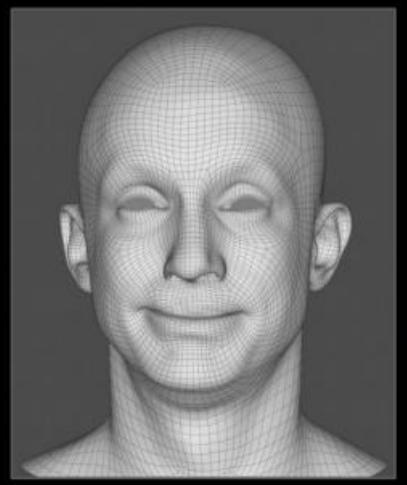


SECURITY

FINANCE

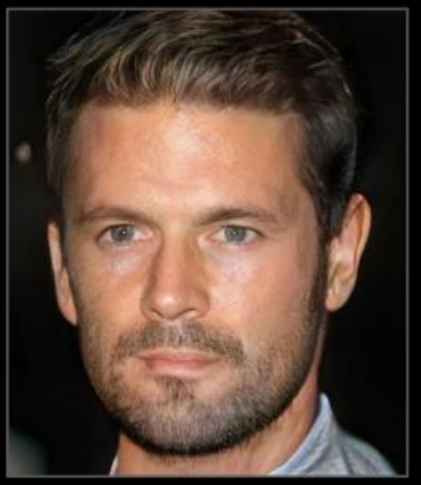
NVIDIA RESEARCH

NVIDIA Research
AI Autoencoder



NVIDIA Research / Remedy
Audio-driven Facial Animation

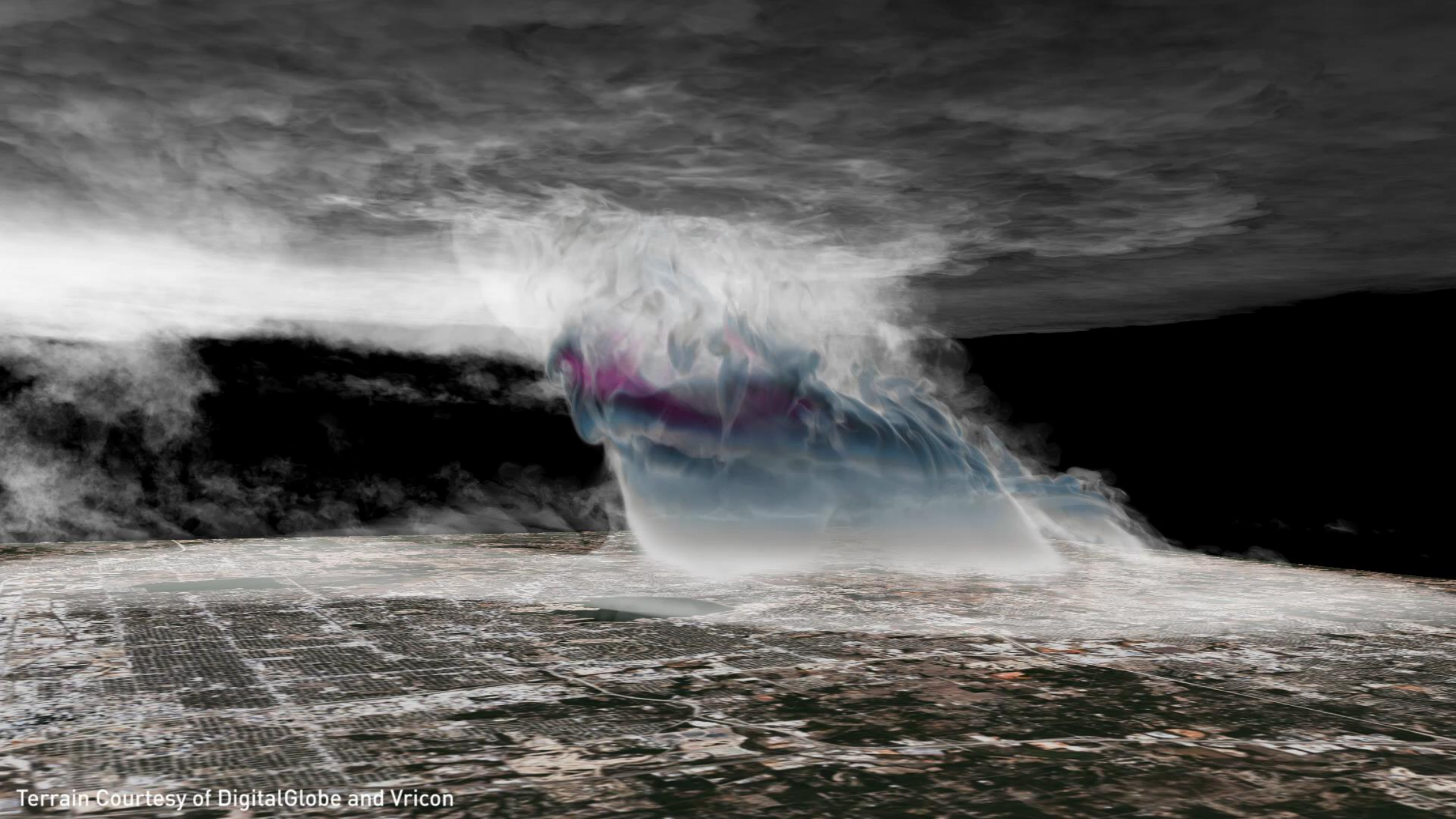
NVIDIA Research
Semantic Manipulation with GANs



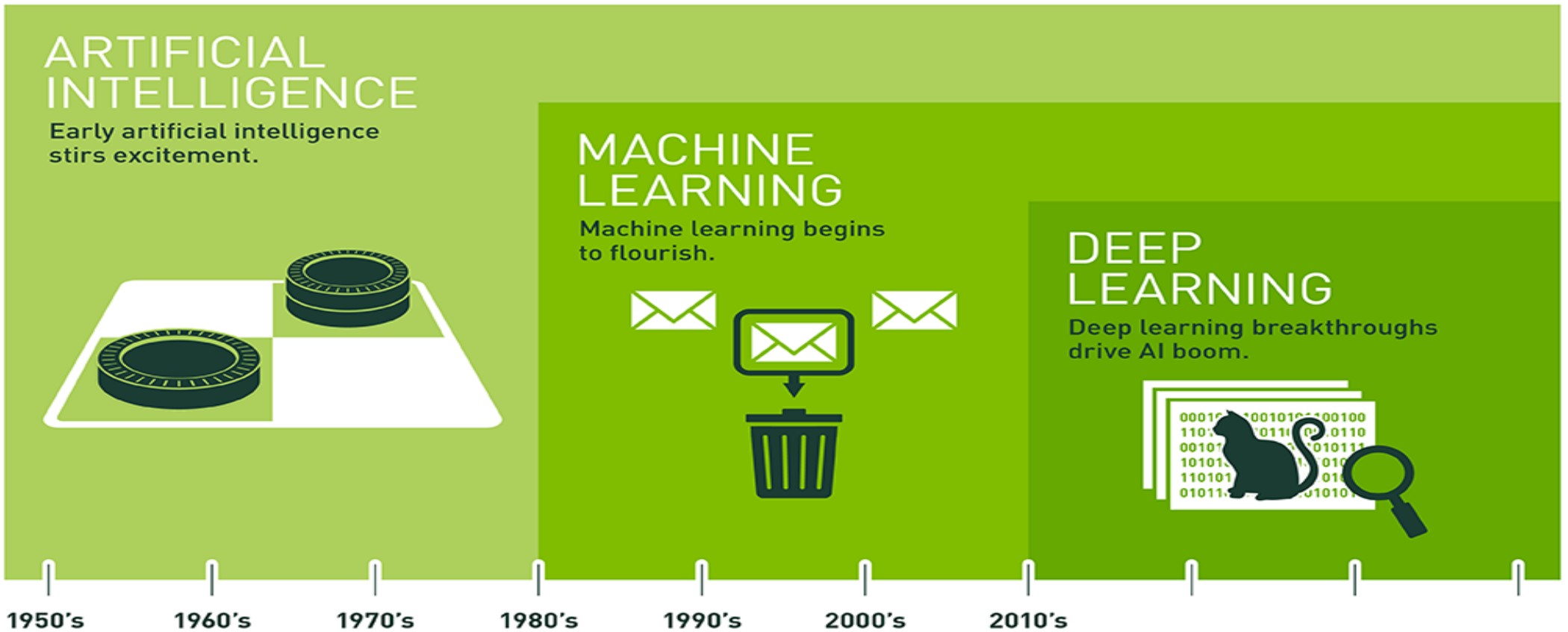
NVIDIA Research
Progressive GAN

NVIDIA Research / AIVA
RNNs for Music

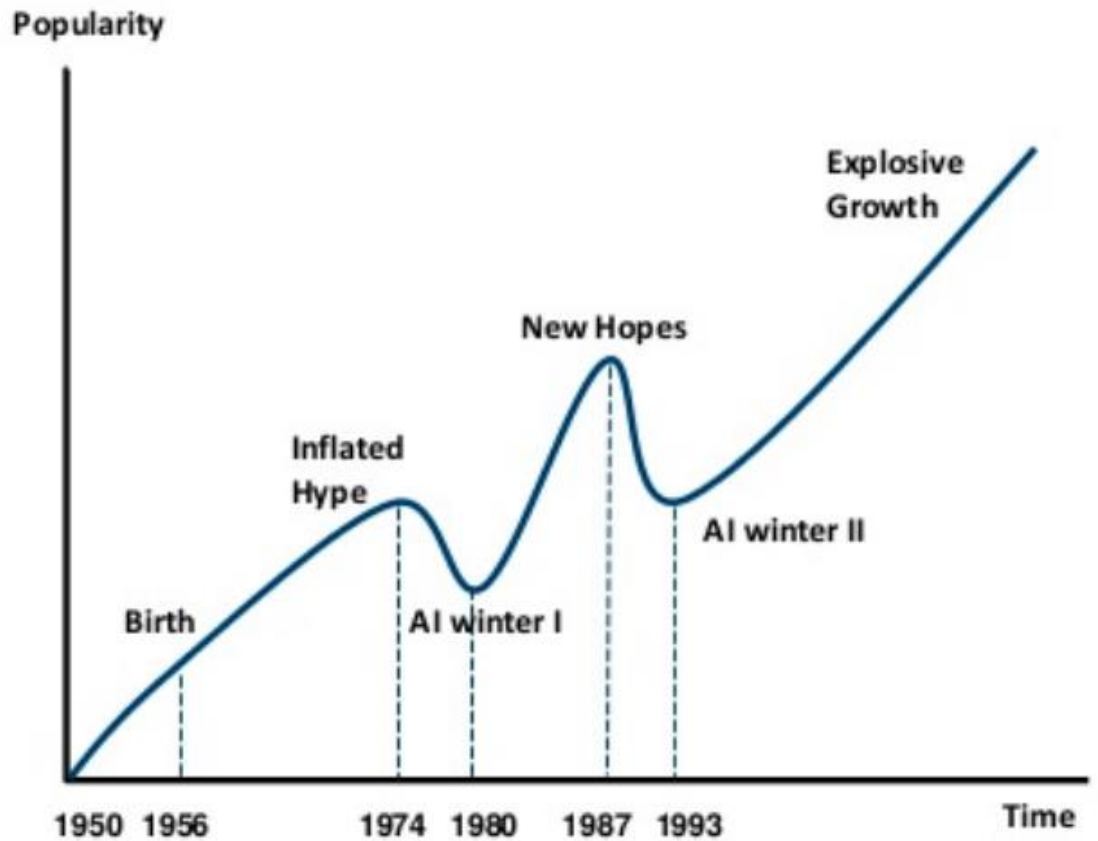

An unlikely hero…
Unreal Engine 4© Epic Games

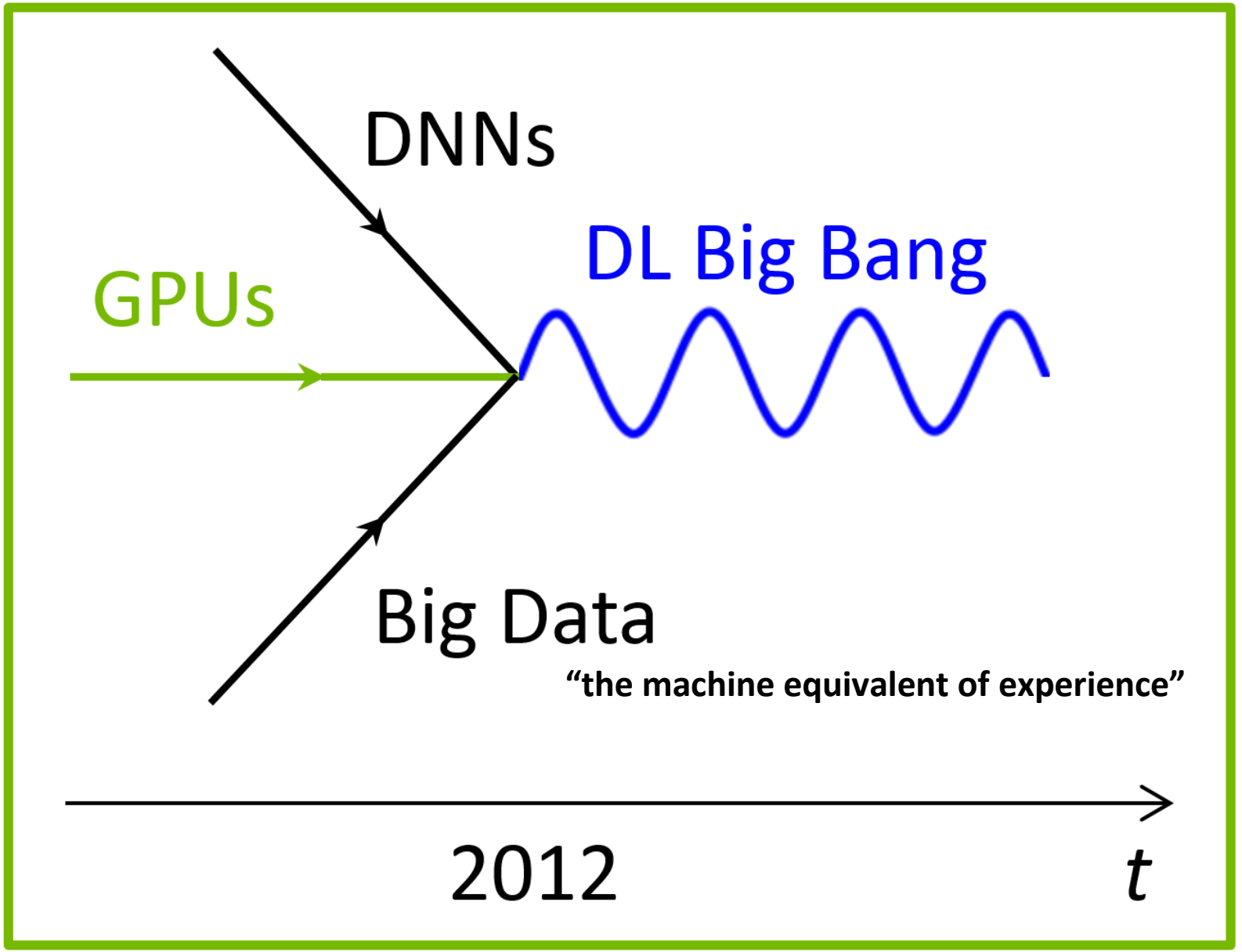

Terrain Courtesy of DigitalGlobe and Vricon

Definitions



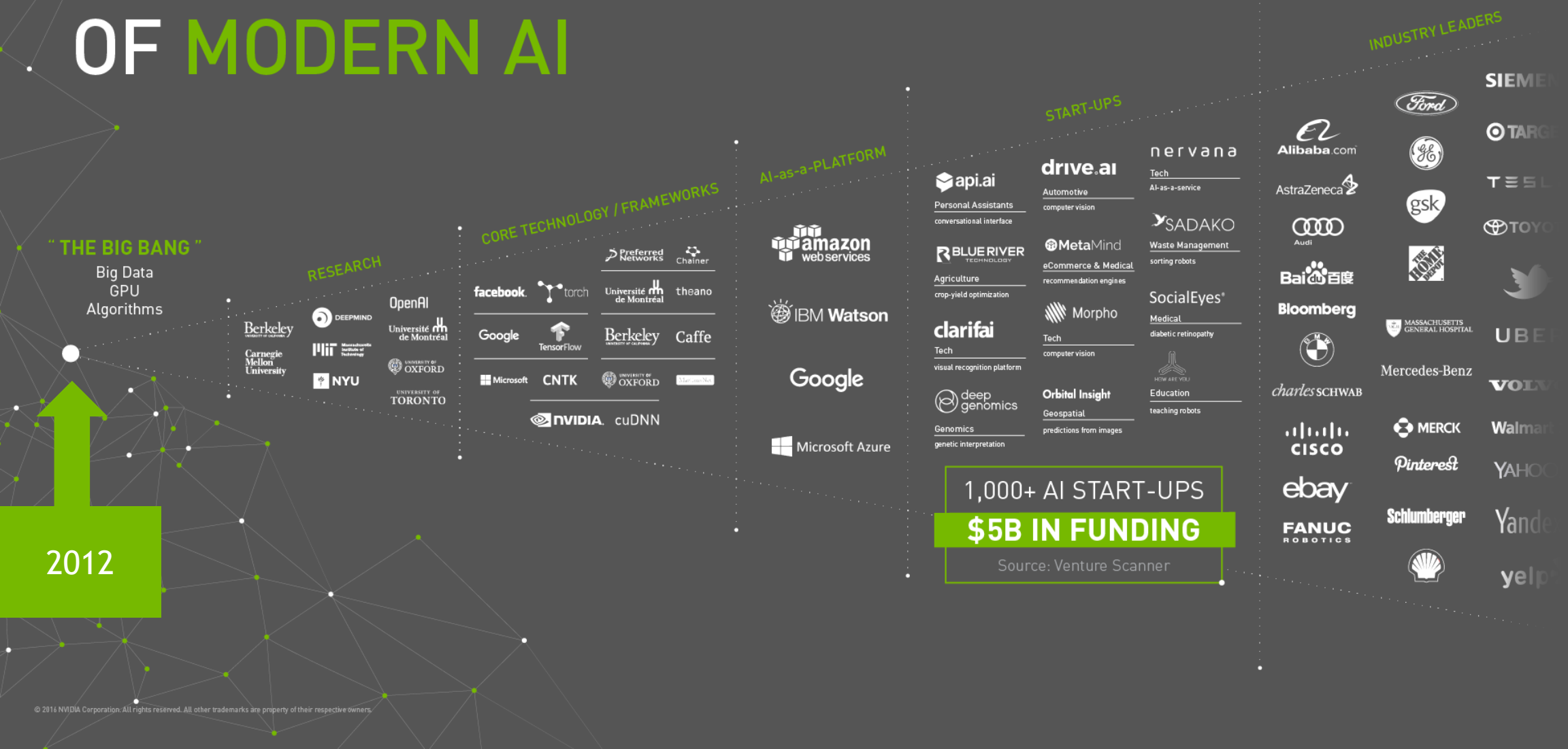
History of Artificial Intelligence Hype





Feynman Diagram of DL, © NVIDIA 2016, author Branislav Kisačanin

THE EXPANDING UNIVERSE OF MODERN AI



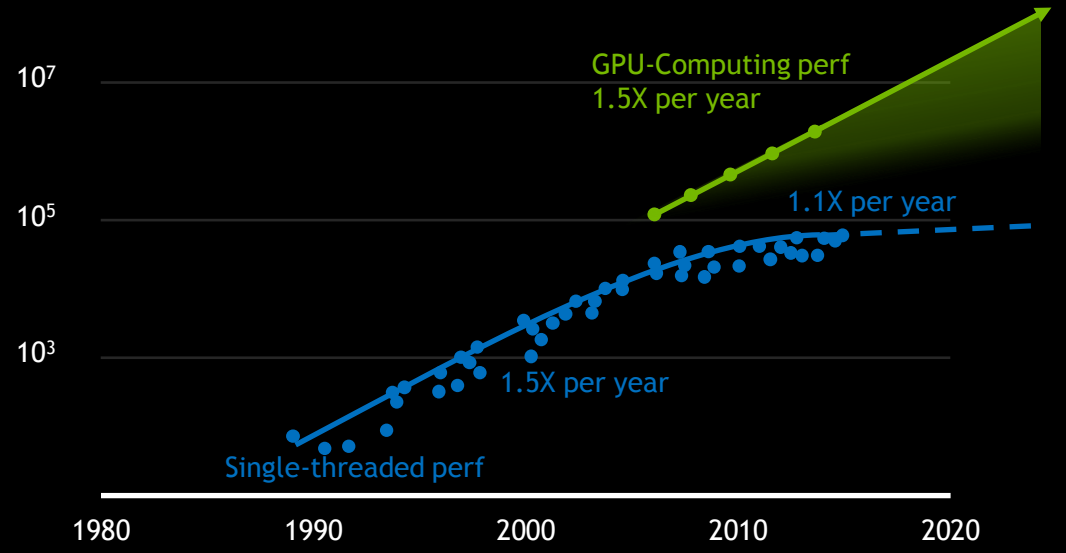
GPU Computing

A kernel runs on an SM > grid of blocks > WARP (32 threads)
The kernel is the unit of work {instruction stream with arguments}

THE RISE OF GPU COMPUTING

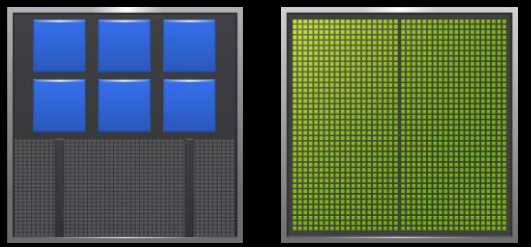
Big Data Needs Algorithms and Compute That Scales

END OF MOORE'S LAW

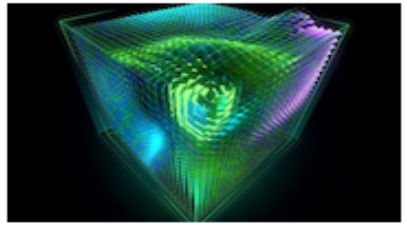


Original data up to the year 2010 collected and plotted by M. Horowitz,
F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

CPU vs. GPU



NVIDIA



CUDA 9 Features Revealed: Volta, Cooperative Groups and More

Posted on May 11, 2017 by Mark Harris | 46 Comments
Tagged Cooperative Groups, CUBLAS, CUDA, CUDA 9, Deep Learning, Libraries, Tensor Cores

At the 2017 GPU Technology Conference NVIDIA announced CUDA 9, the latest version of CUDA's powerful parallel computing platform and programming model. CUDA 9 is now available as a free download. In this post I'll provide an overview of the awesome new features of CUDA 9.

- Support for the Volta GPU architecture, including the new Tesla V100 accelerator;
- Cooperative Groups, a new programming model for managing groups of communicating threads;
- A new API (preview feature) for programming Tensor Core matrix multiply and accumulate operations on Tesla V100.
- Faster library routines for linear algebra, image processing, FFTs, and more;
- New algorithms in cuSolver and nvGraph

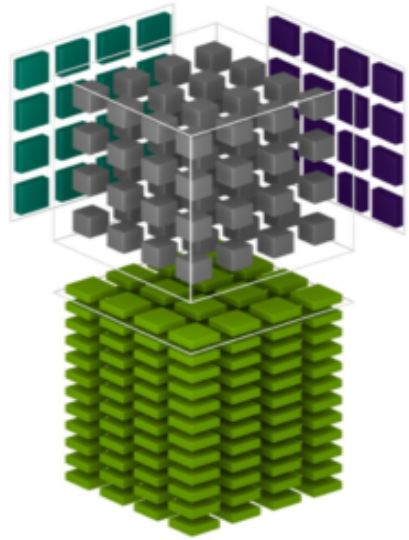


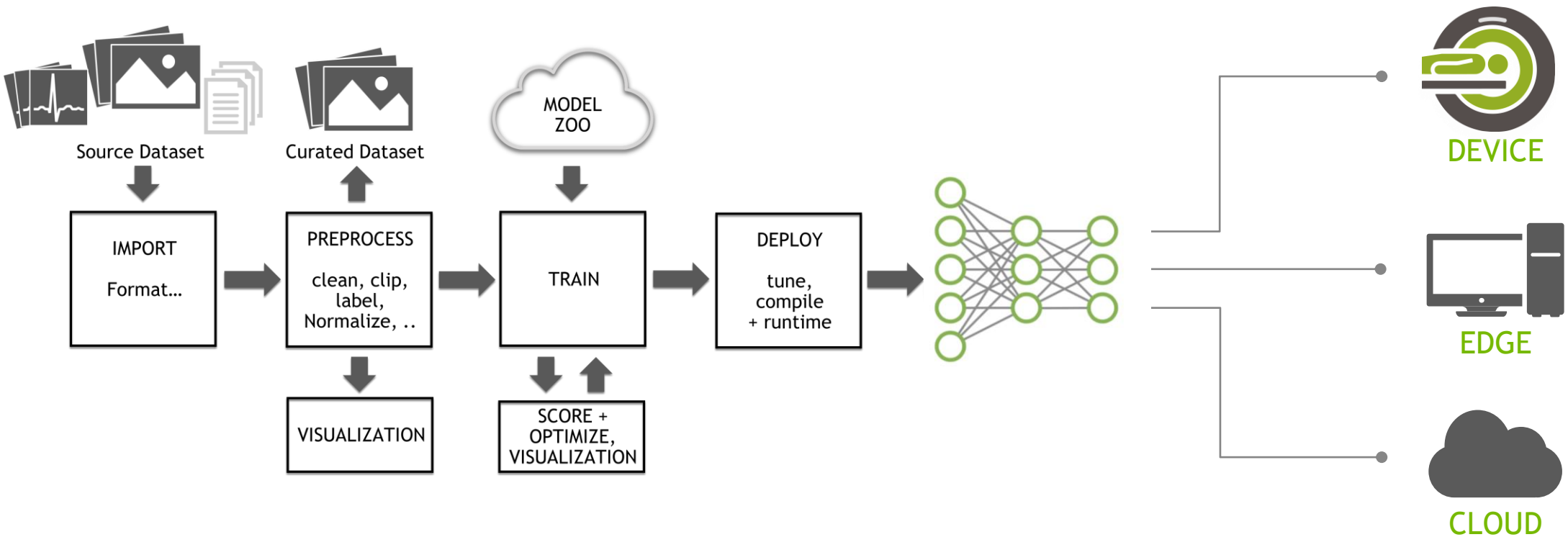
Figure 1: CUDA 9 provides a preview API for programming Tesla V100 Tensor Cores, providing a huge

<title>code ninja</title>

DEEP LEARNING

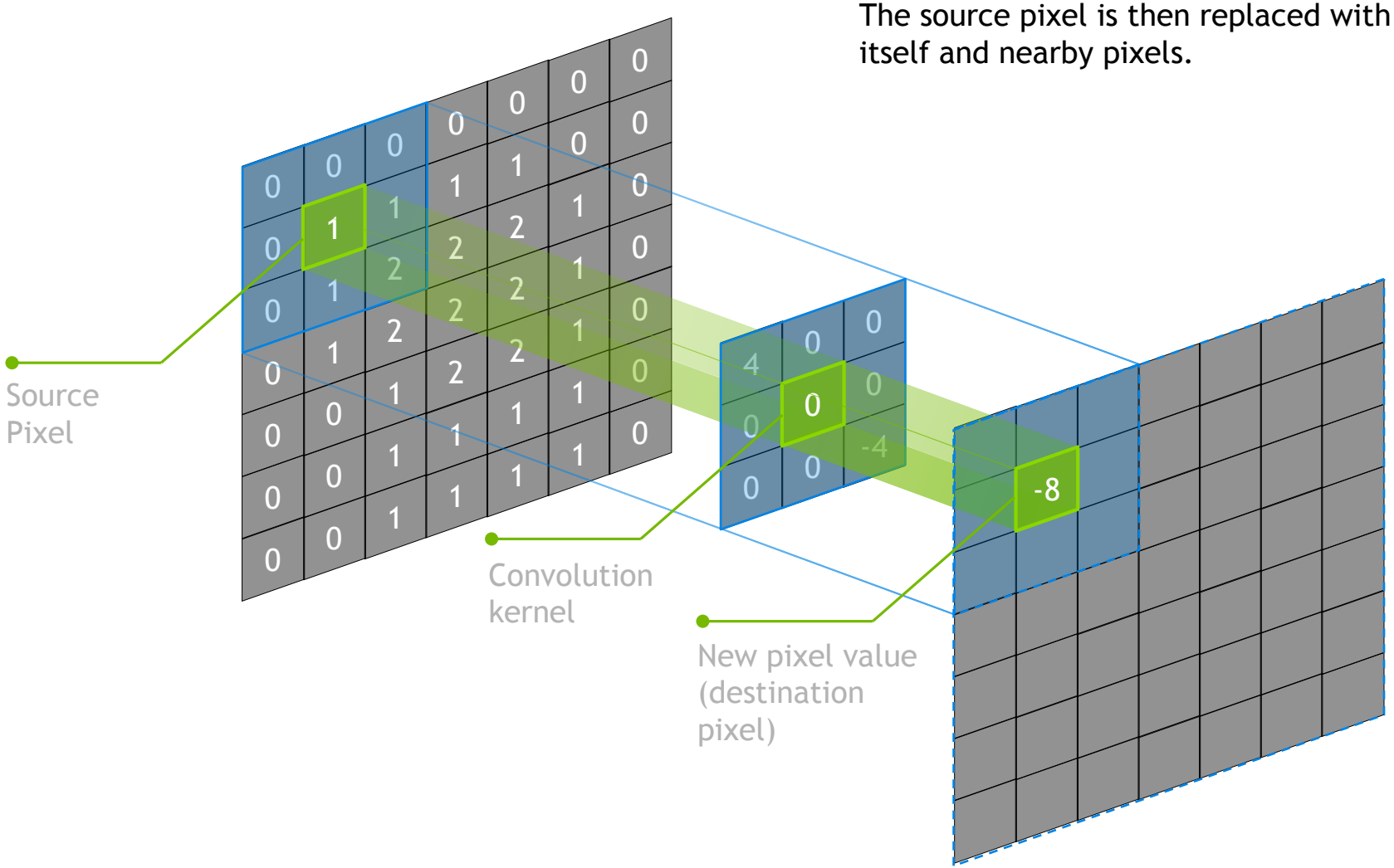
SUPERVISED LEARNING

NEW PROGRAMMING MODEL

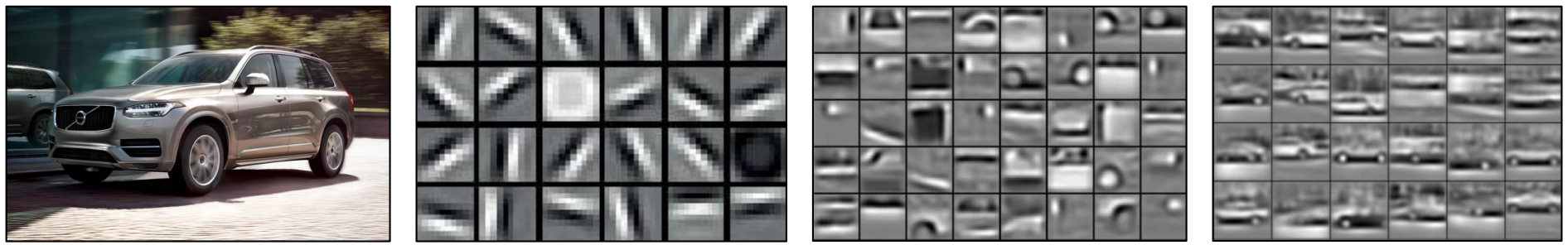


CONVOLUTION

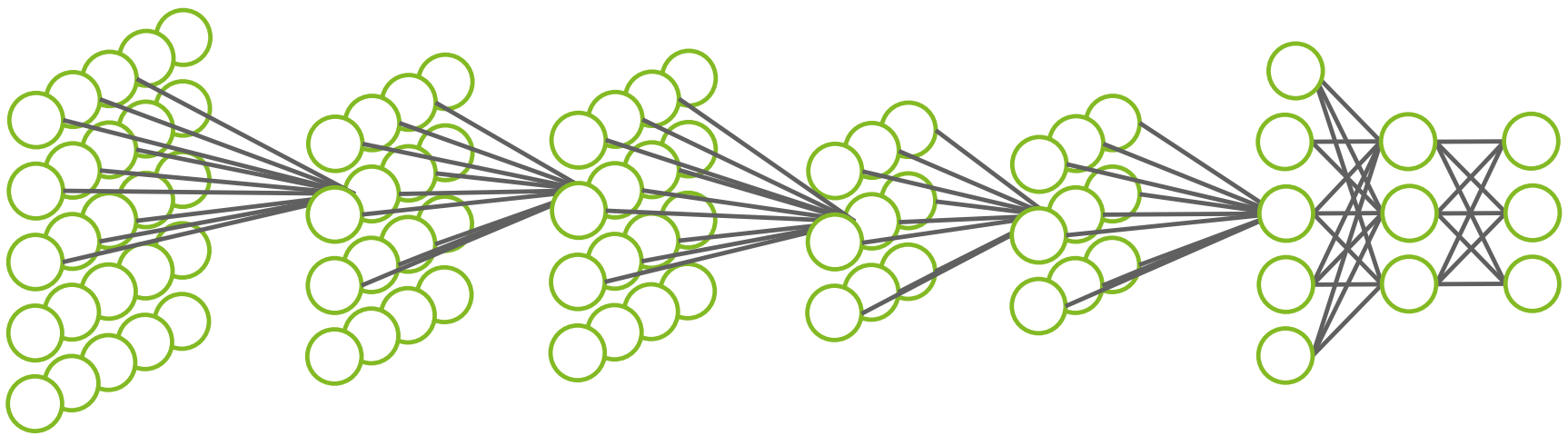
Centre element of the kernel is placed over the source pixel. The source pixel is then replaced with a weighted sum of itself and nearby pixels.



CONVOLUTIONAL NEURAL NETWORKS



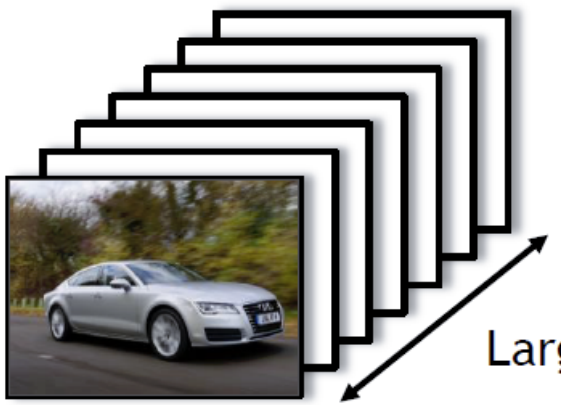
Image



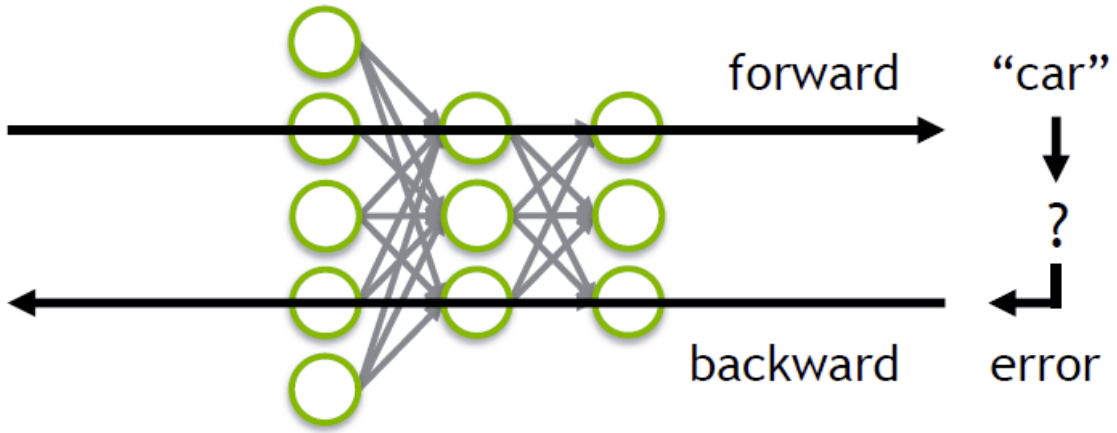
“Volvo XC90”

Image source: “Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks” ICML 2009 & Comm. ACM 2011.
Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Ng.

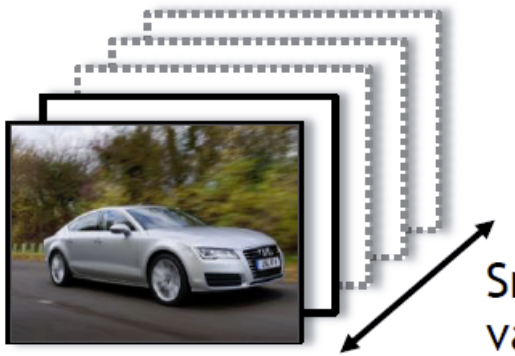
TRAINING VS INFERENCE



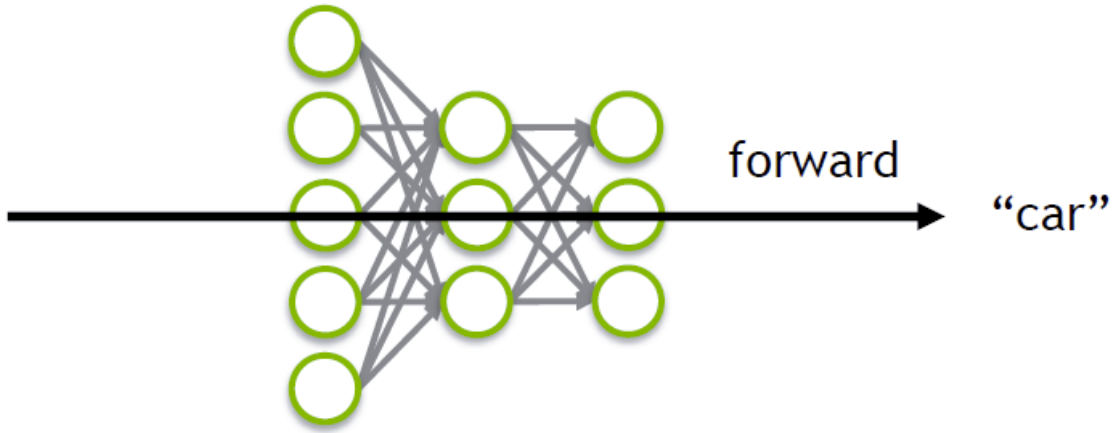
Large N



TRAINING



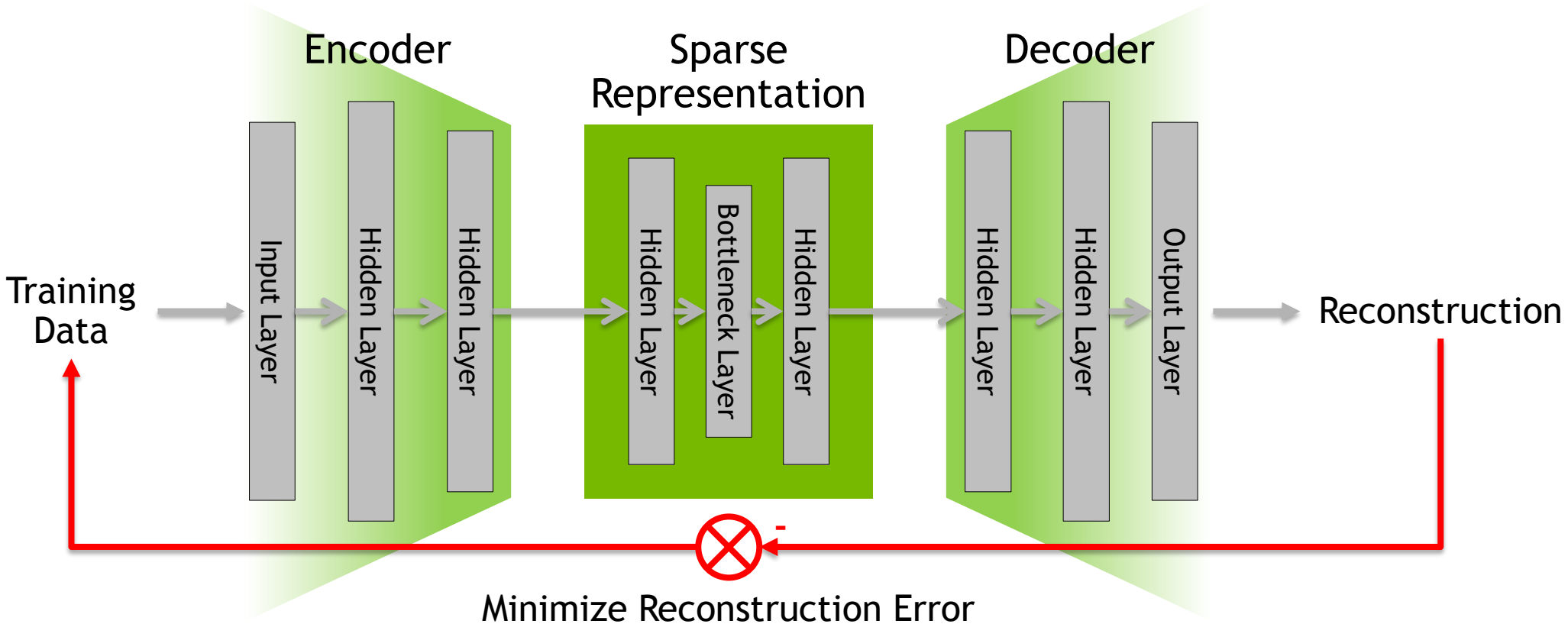
Smaller,
varied N



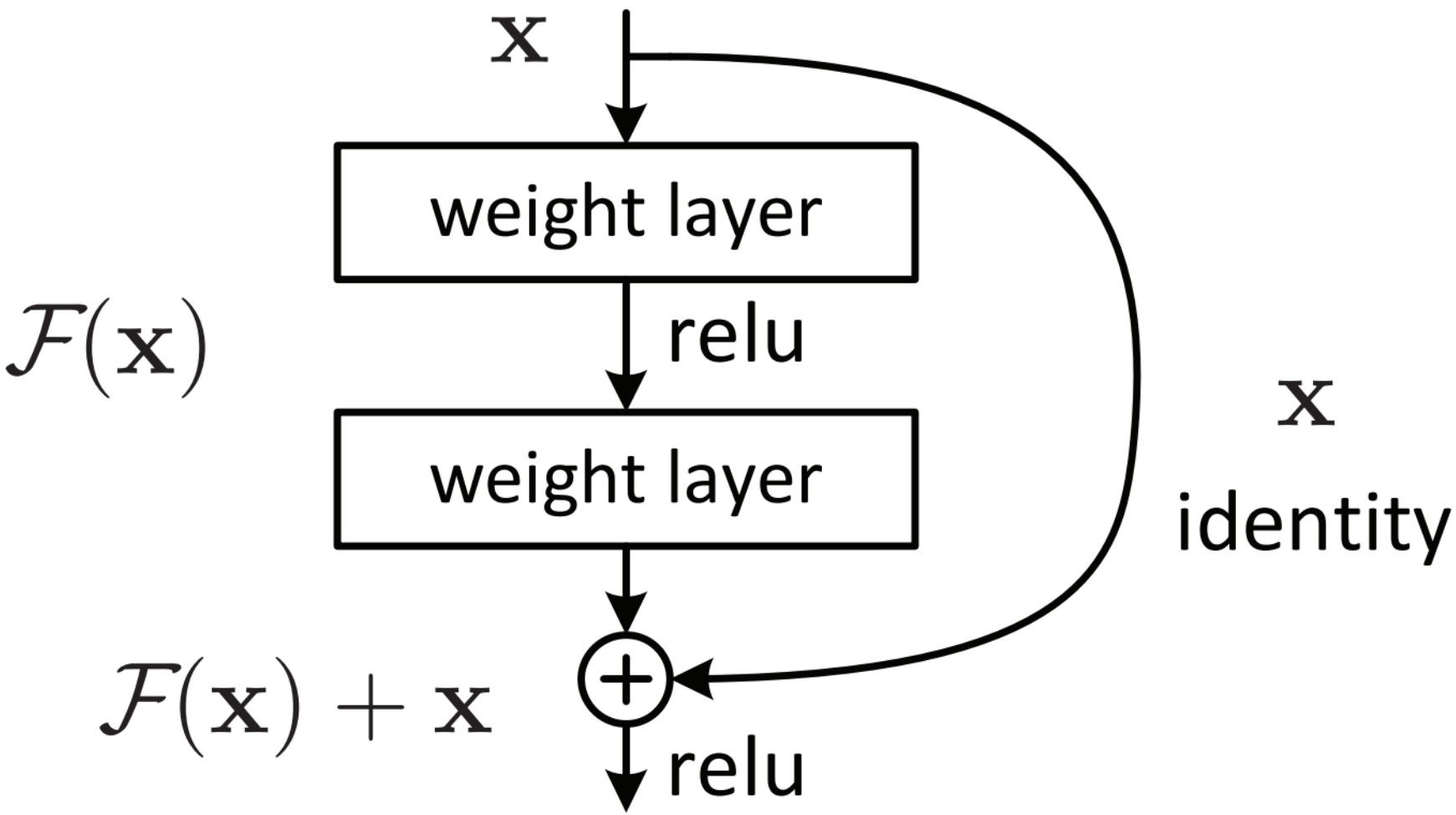
INFERENCE

EXAMPLE: AUTOENCODERS

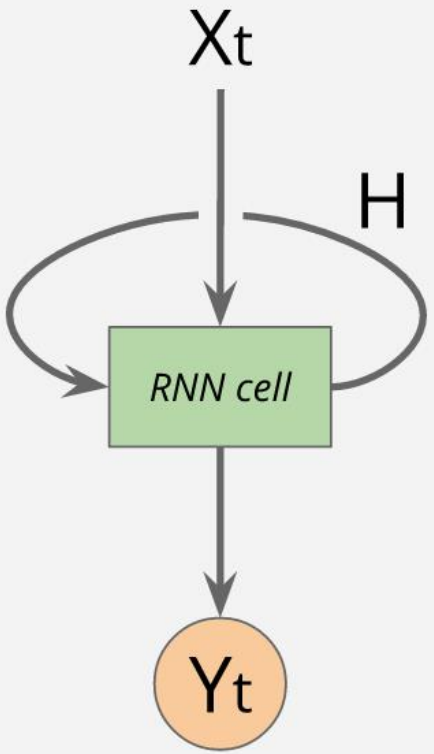
Feature Learning

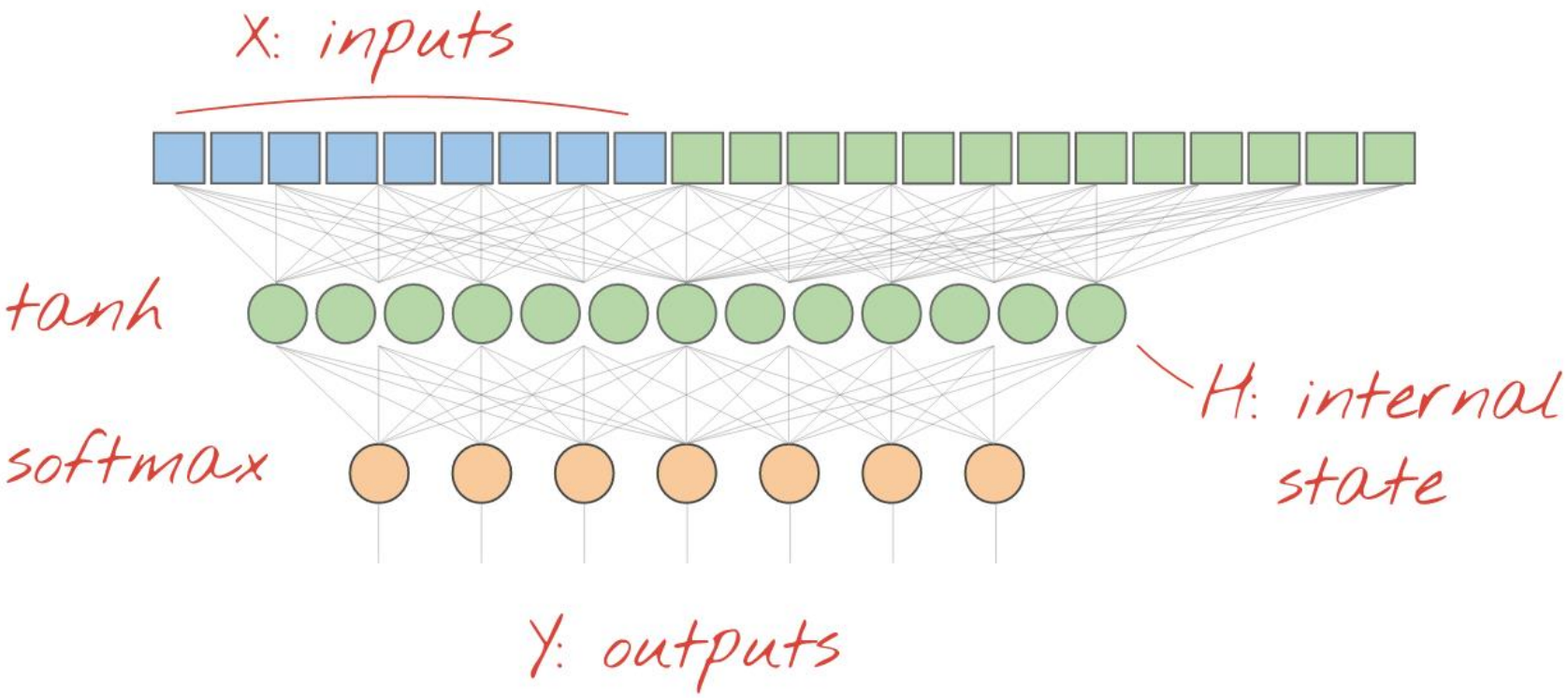


Residual Networks



RNN


Xt
H
RNN cell
Yt


X: inputs
tanh
softmax
H: internal state
Y: outputs

Long short-term memory (LSTM)

Hochreiter (1991) analysed vanishing gradient "*LSTM falls out of this almost naturally*"

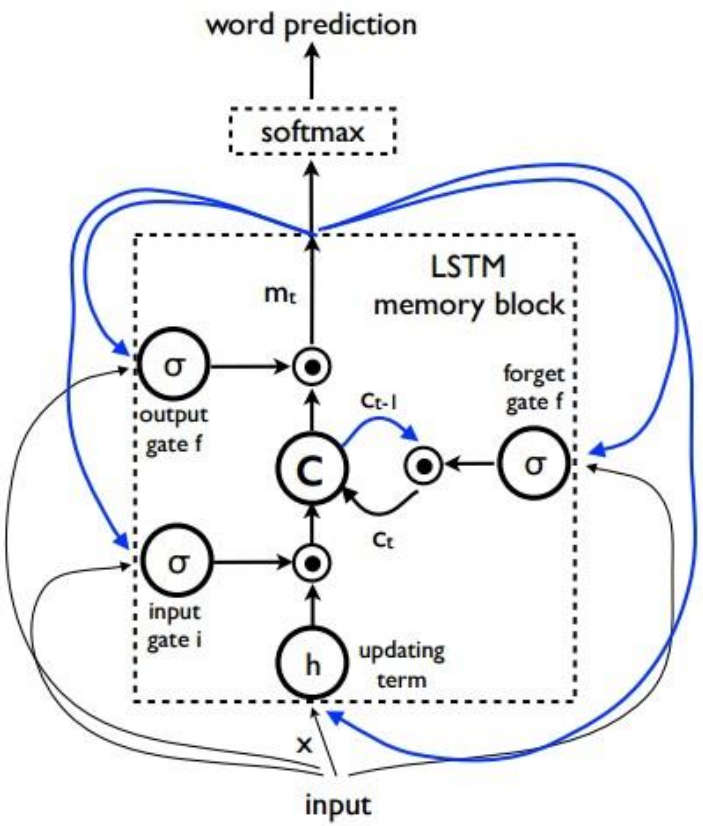
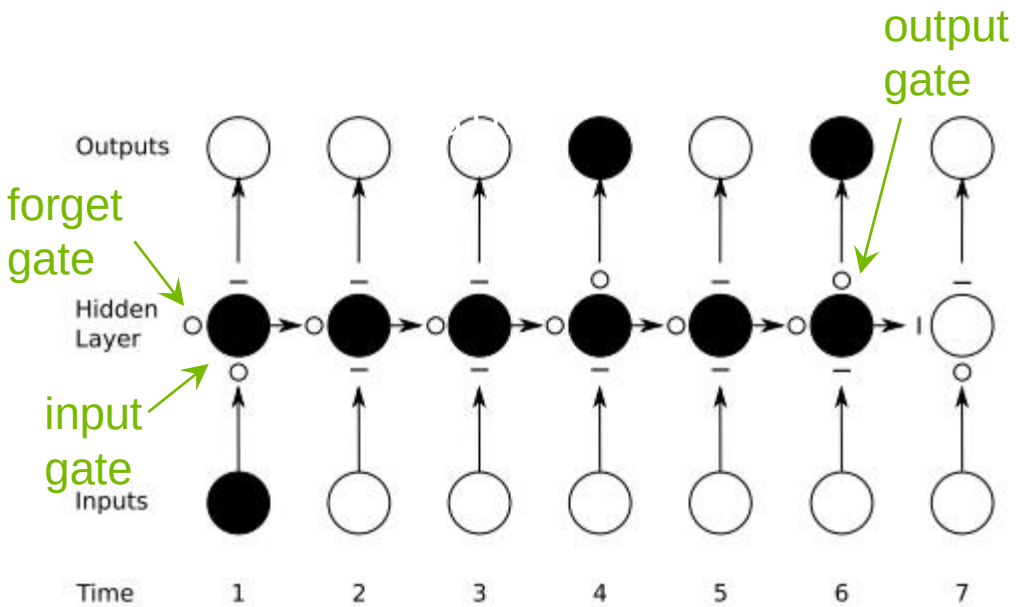


Fig from Vinyals et al, Google April 2015 NIC Generator

Training via backprop unfolded in time

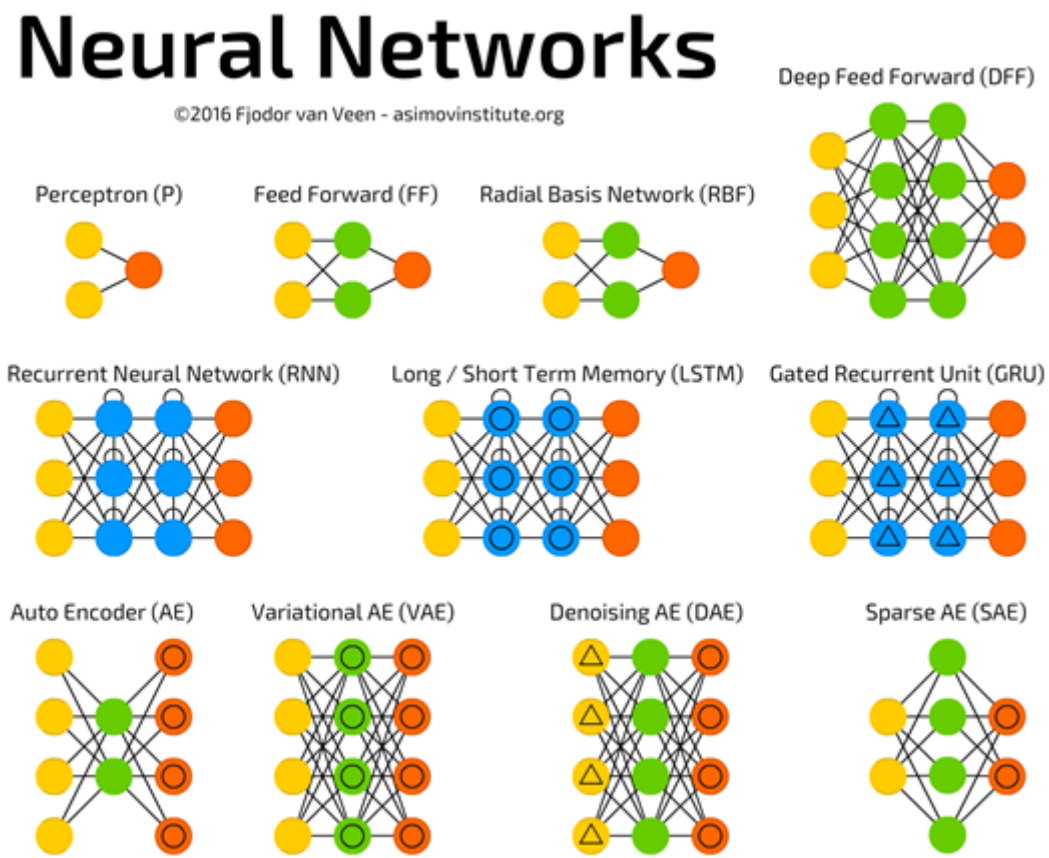


Long time dependencies are preserved until input gate is closed (-) and forget gate is open (O)

Gates control importance of the corresponding activations

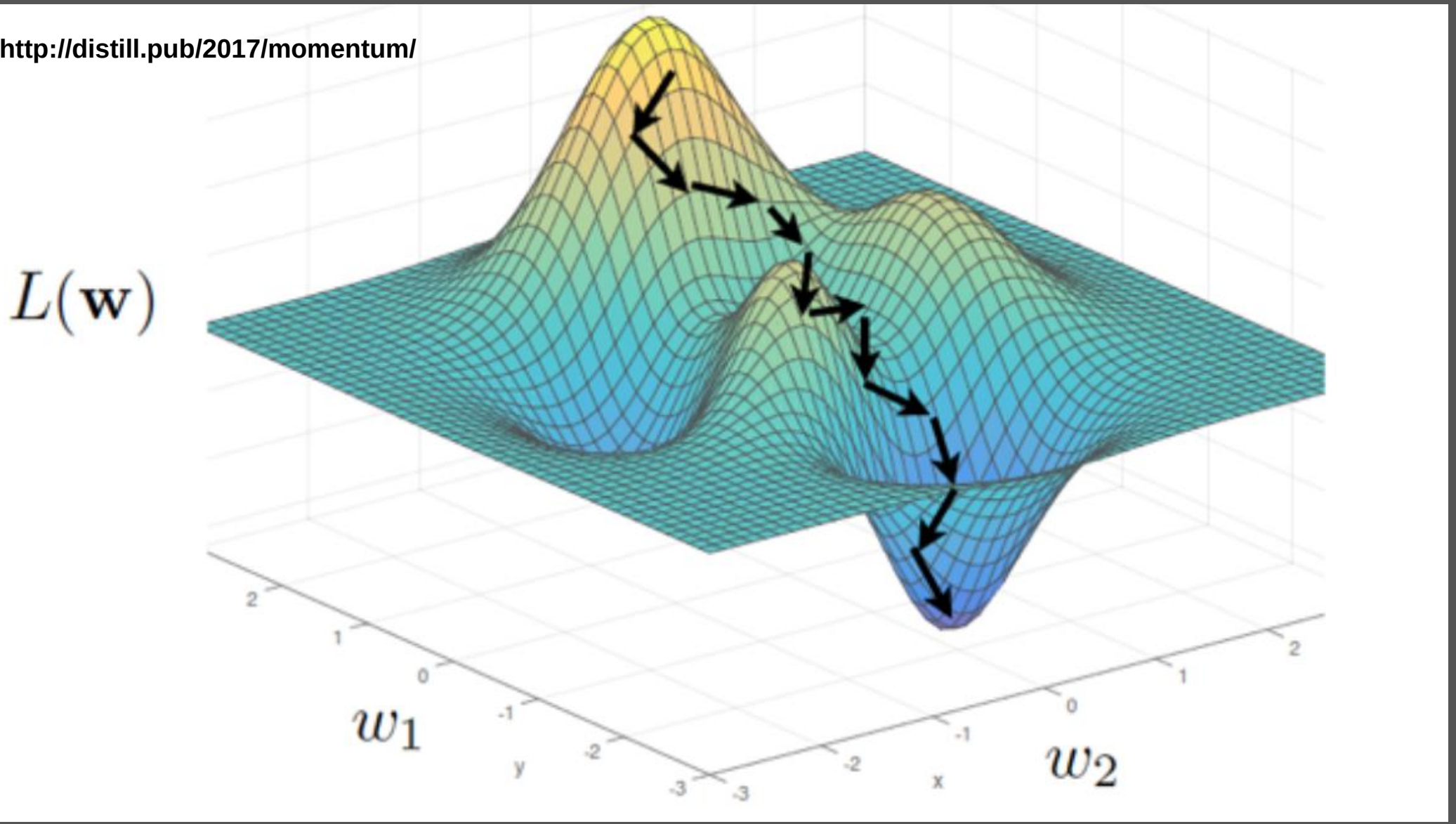
Fig from Graves, Schmidhuber et al, Supervised Sequence Labelling with RNNs

ARCHITECTURES



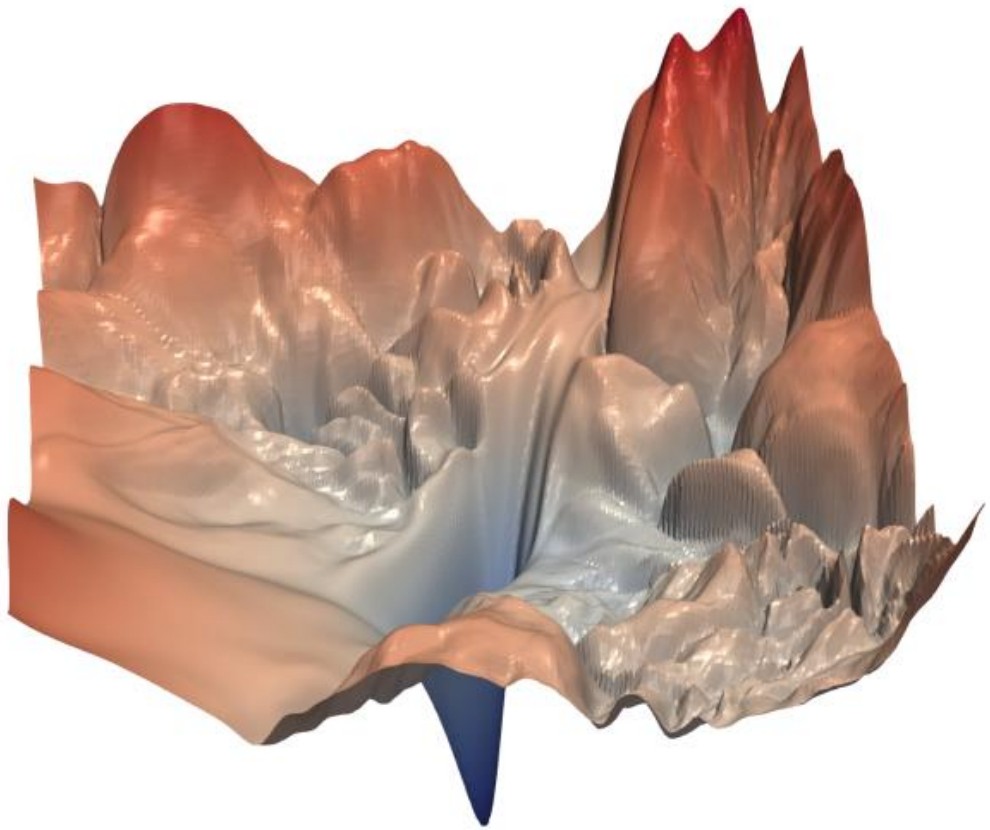
Larger image: http://www.asimovinstitute.org/neural-network-zoo/

http://distill.pub/2017/momentum/

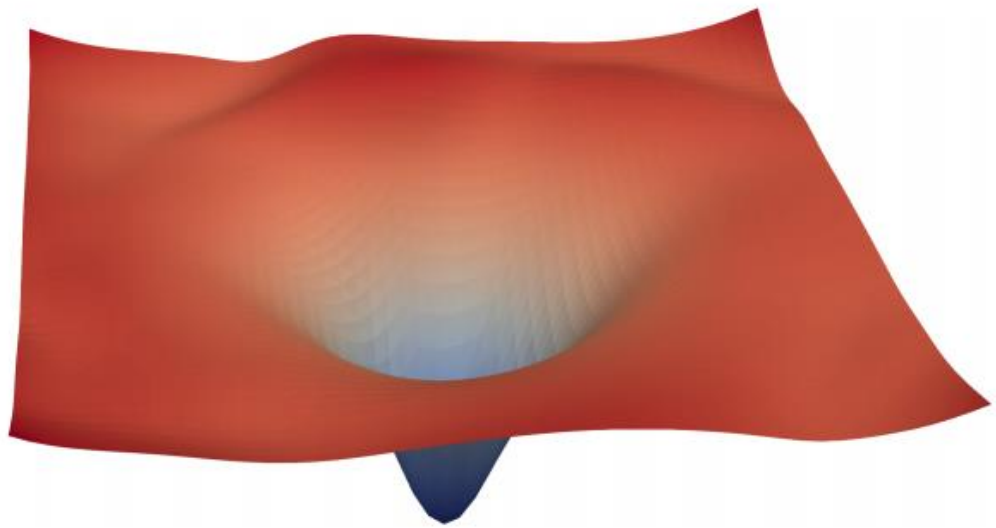


Li et al, University of Maryland & US Naval Academy

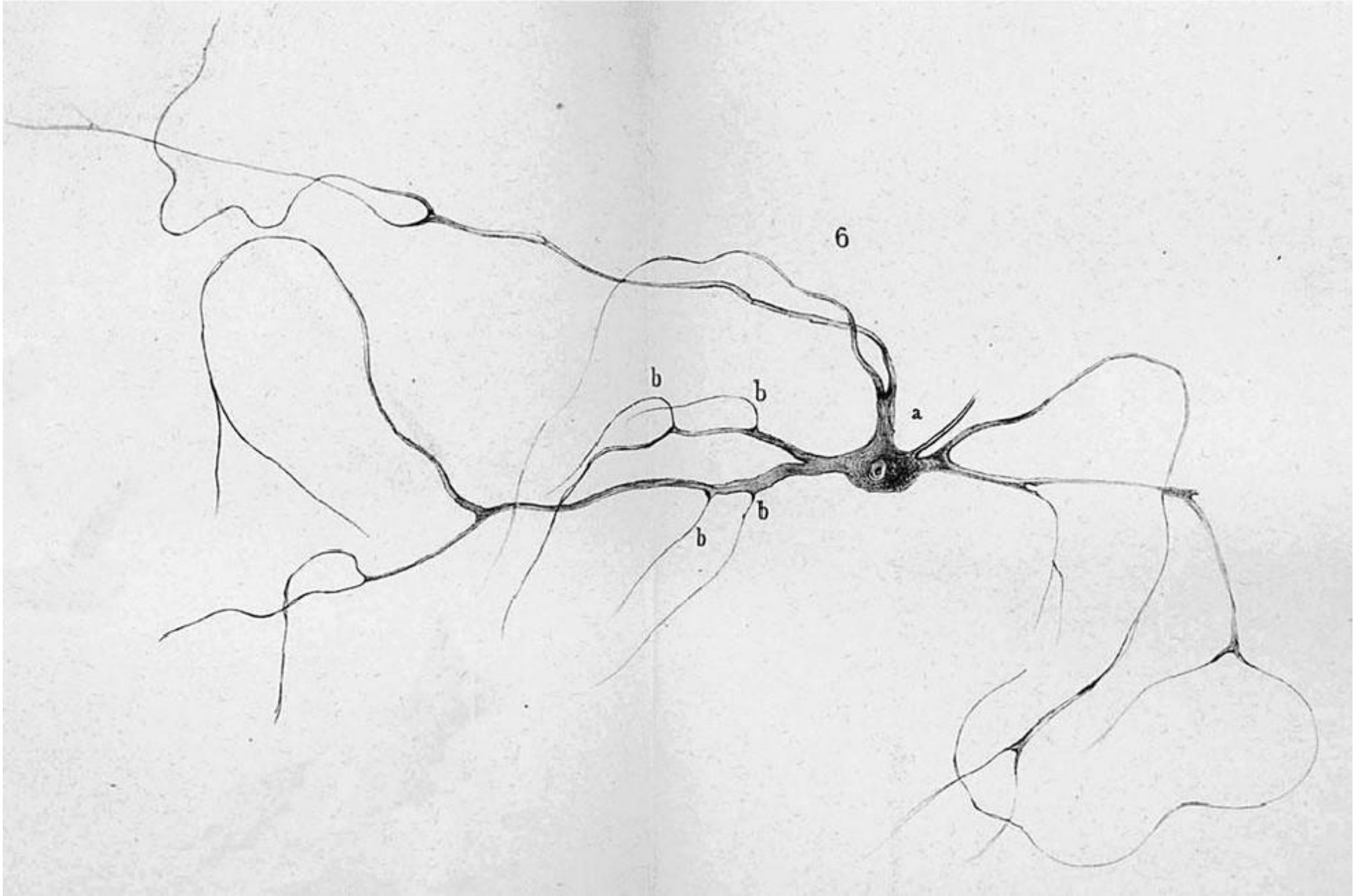
https://arxiv.org/pdf/1712.09913.pdf



(a) without skip connections

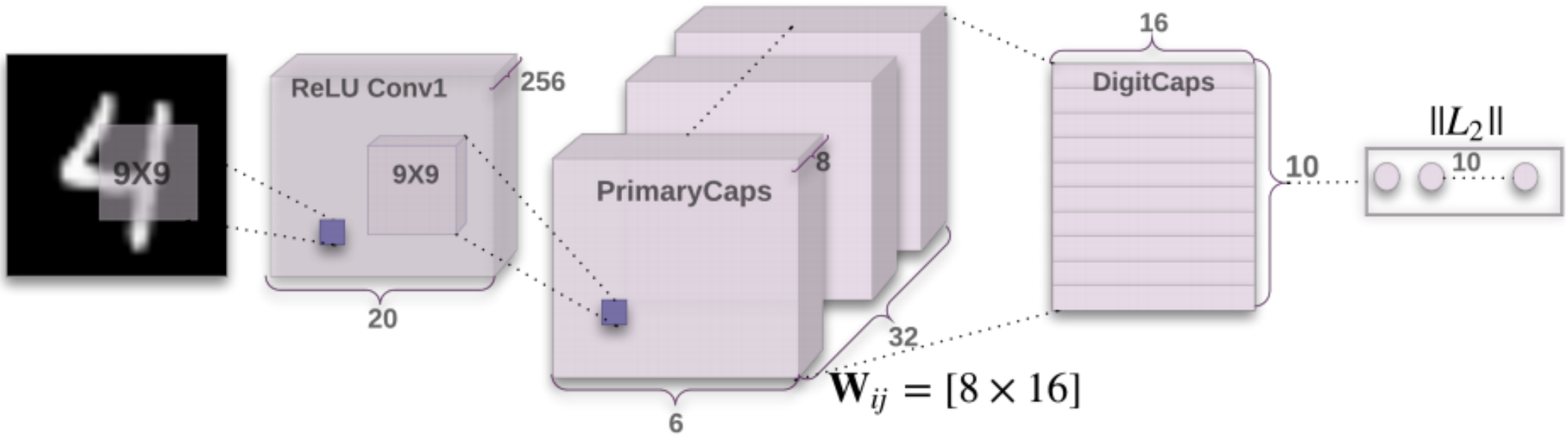


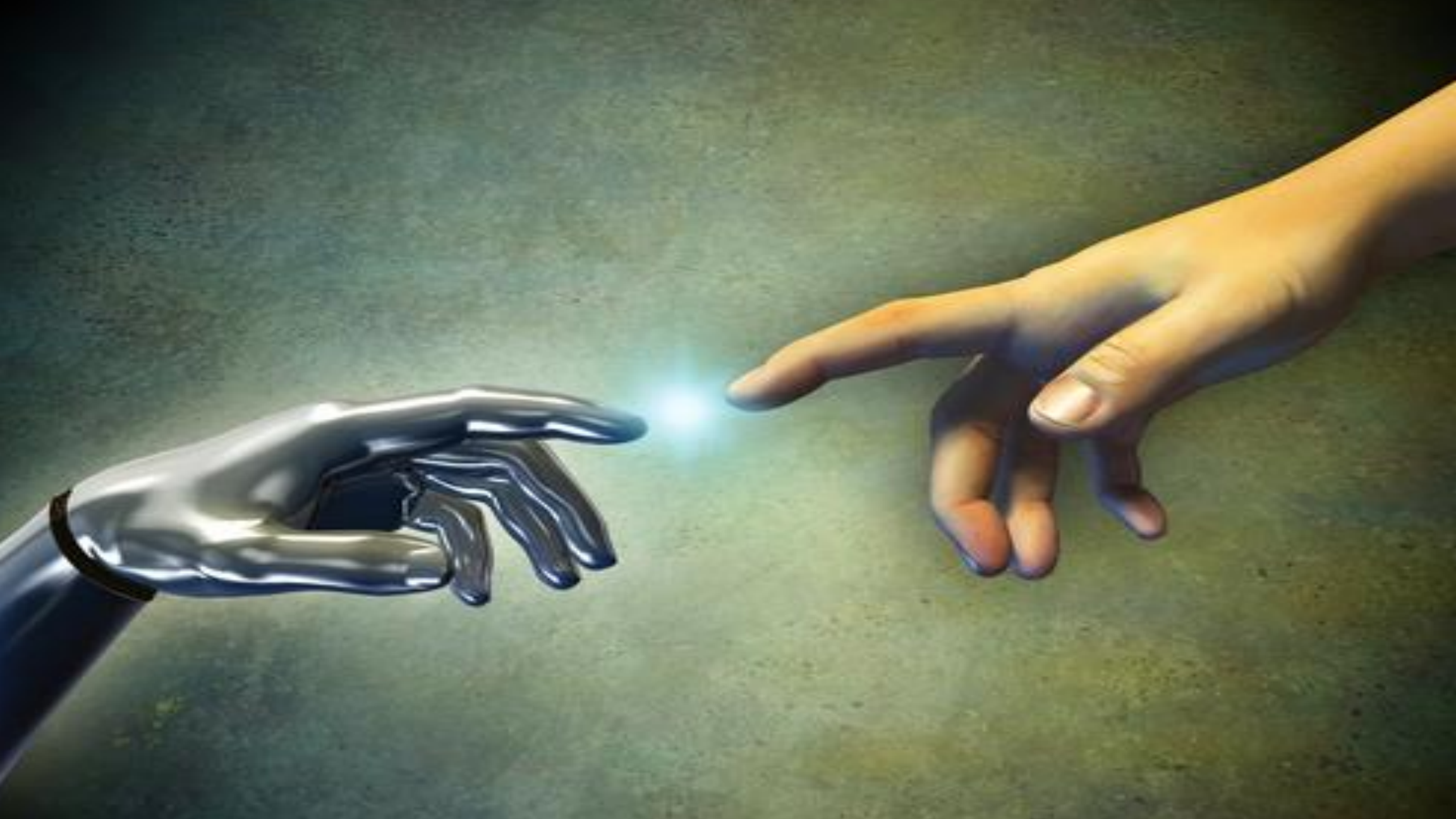
(b) with skip connections


6
b
b
a
b
b

Capsules & routing with EM, Hinton et al

https://arxiv.org/pdf/1710.09829.pdf [NIPS2017]





kernel

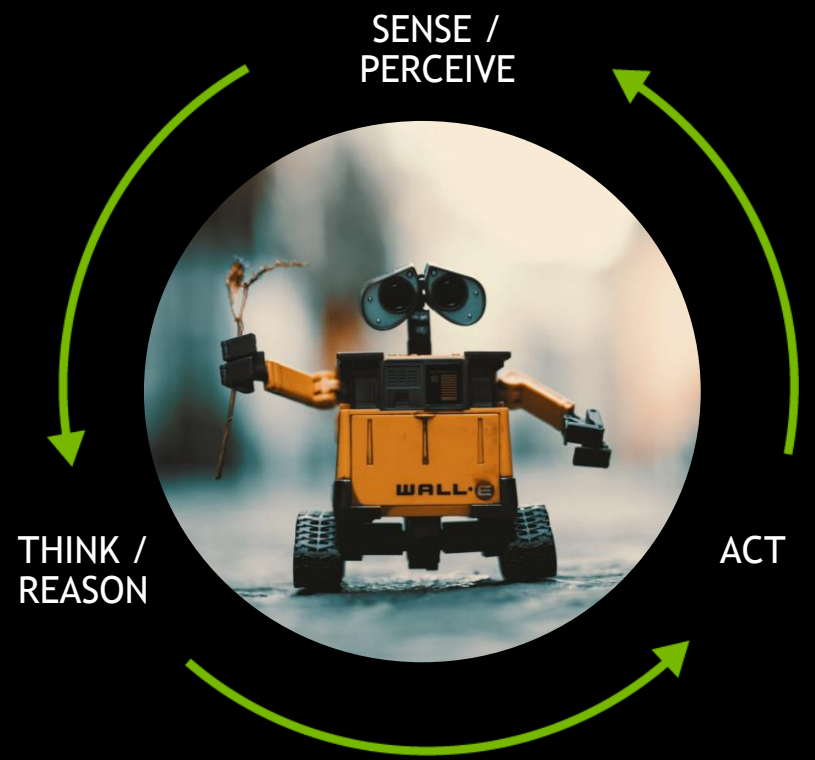
Brain Computer Interfaces
Focused on treatment for disease and dysfunction
eg epilepsy, depression, Parkinsons but ultimately to advance human intelligence by restoring and extending cognitive vibrancy.

THE NEXT STAGE

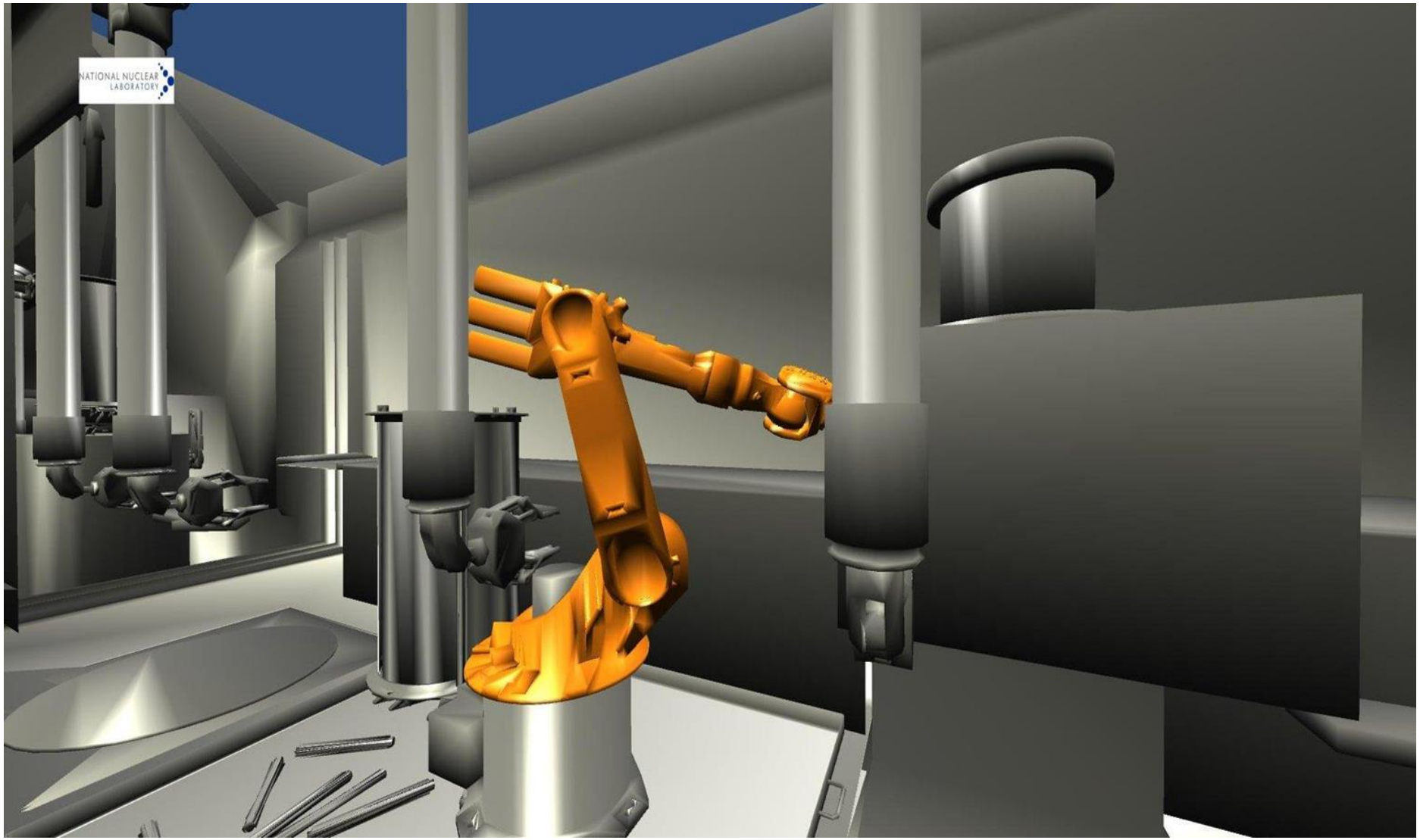
DEEPMIND ALPHA* 1.0 => 2.0 => 0.0

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

ROBOTS

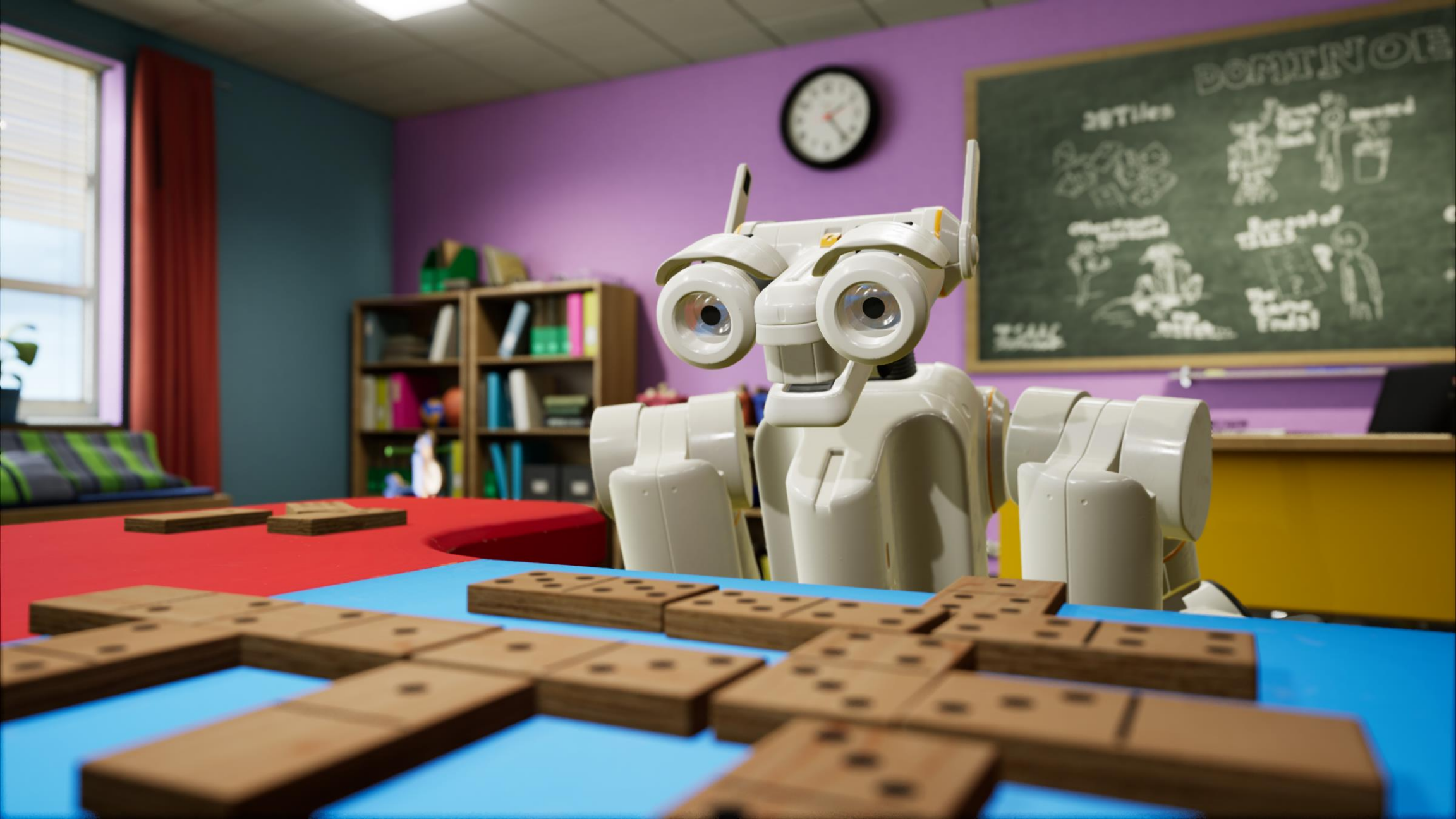



9:33:0 06/06/2015 UTC


NATIONAL NUCLEAR
LABORATORY

IMMERSIVE AI

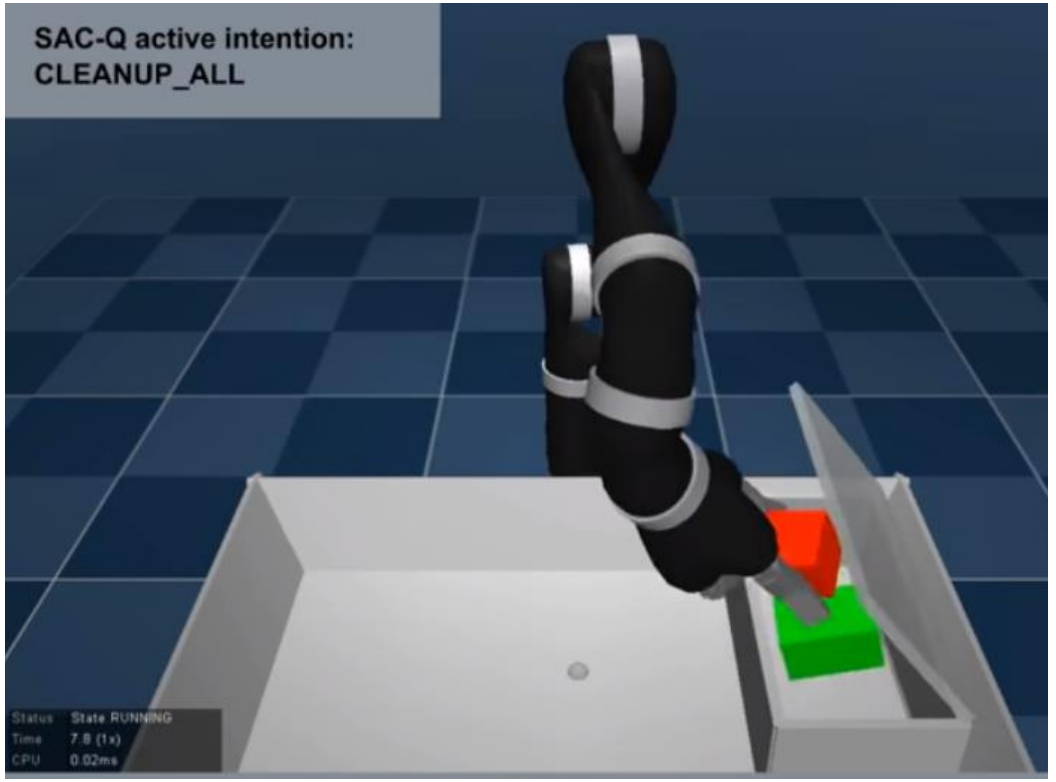
THE DESIGN LAB OF THE FUTURE

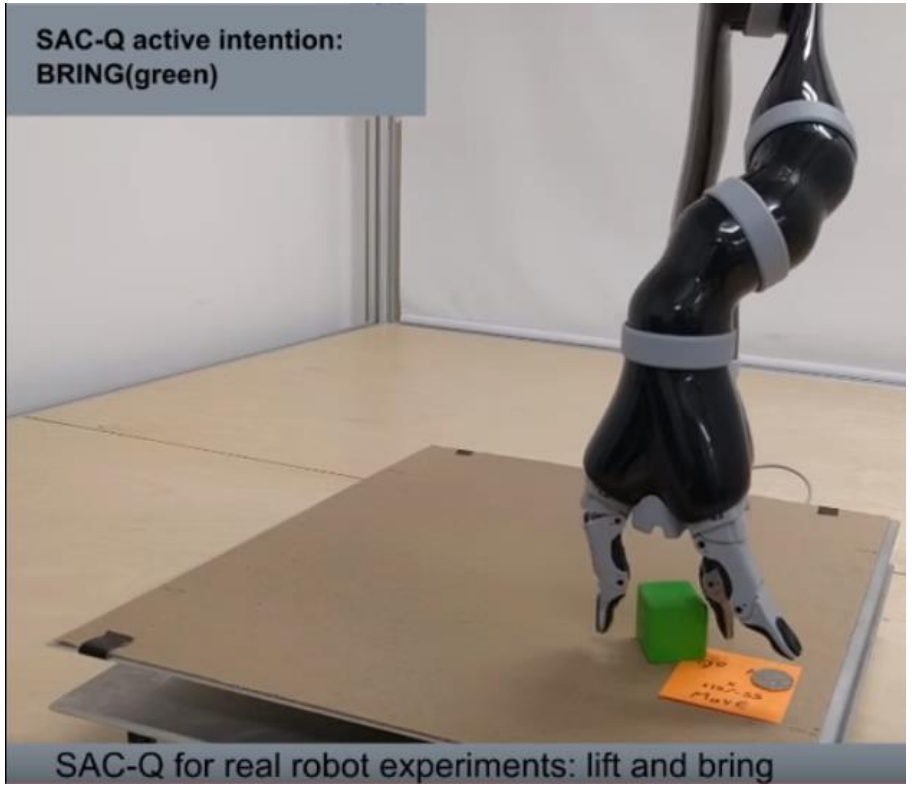


LEARNING FROM SCRATCH

https://arxiv.org/pdf/1802.10567.pdf

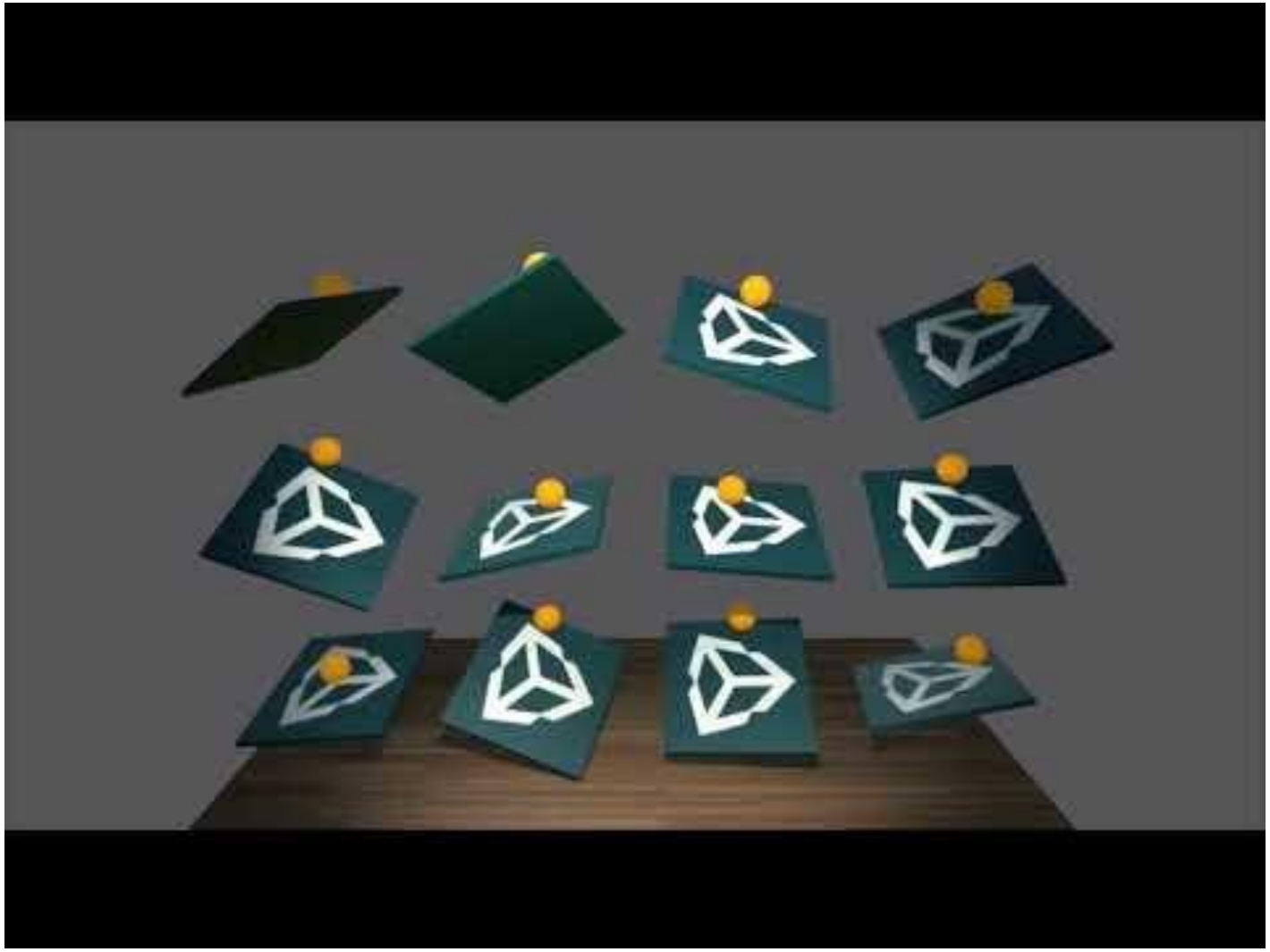
DeepMind





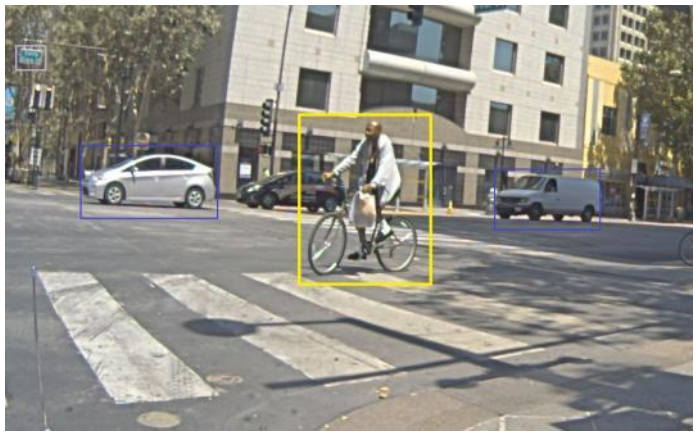
OPENMINED

https://github.com/OpenMined

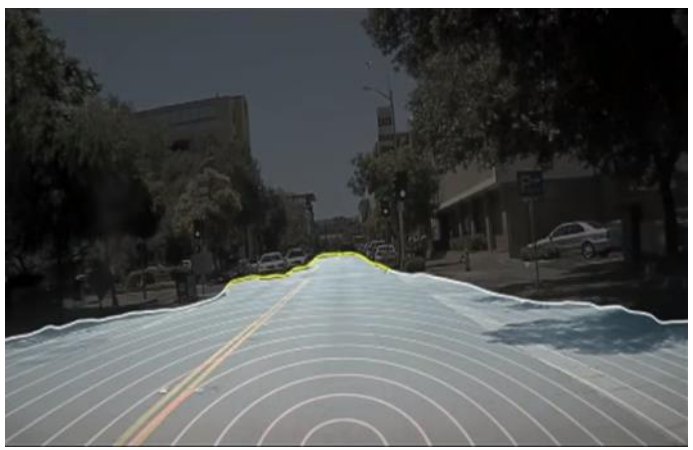


A massive Deep Learning challenge

DEEP LEARNING FOR SELF DRIVING CARS

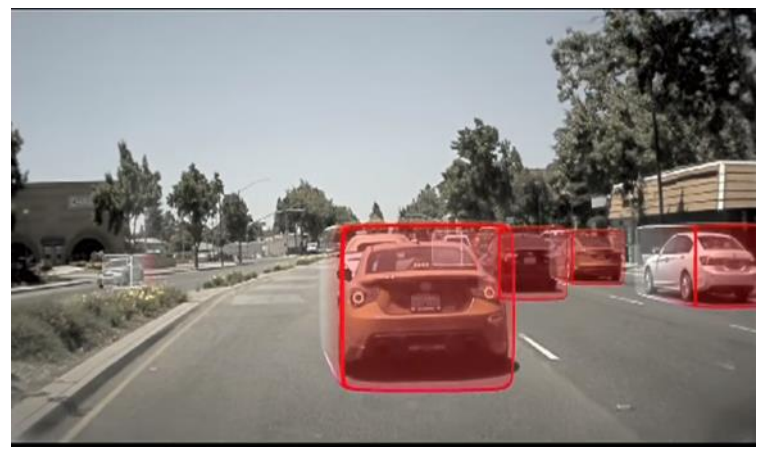


Multi-class detection (DriveNet)

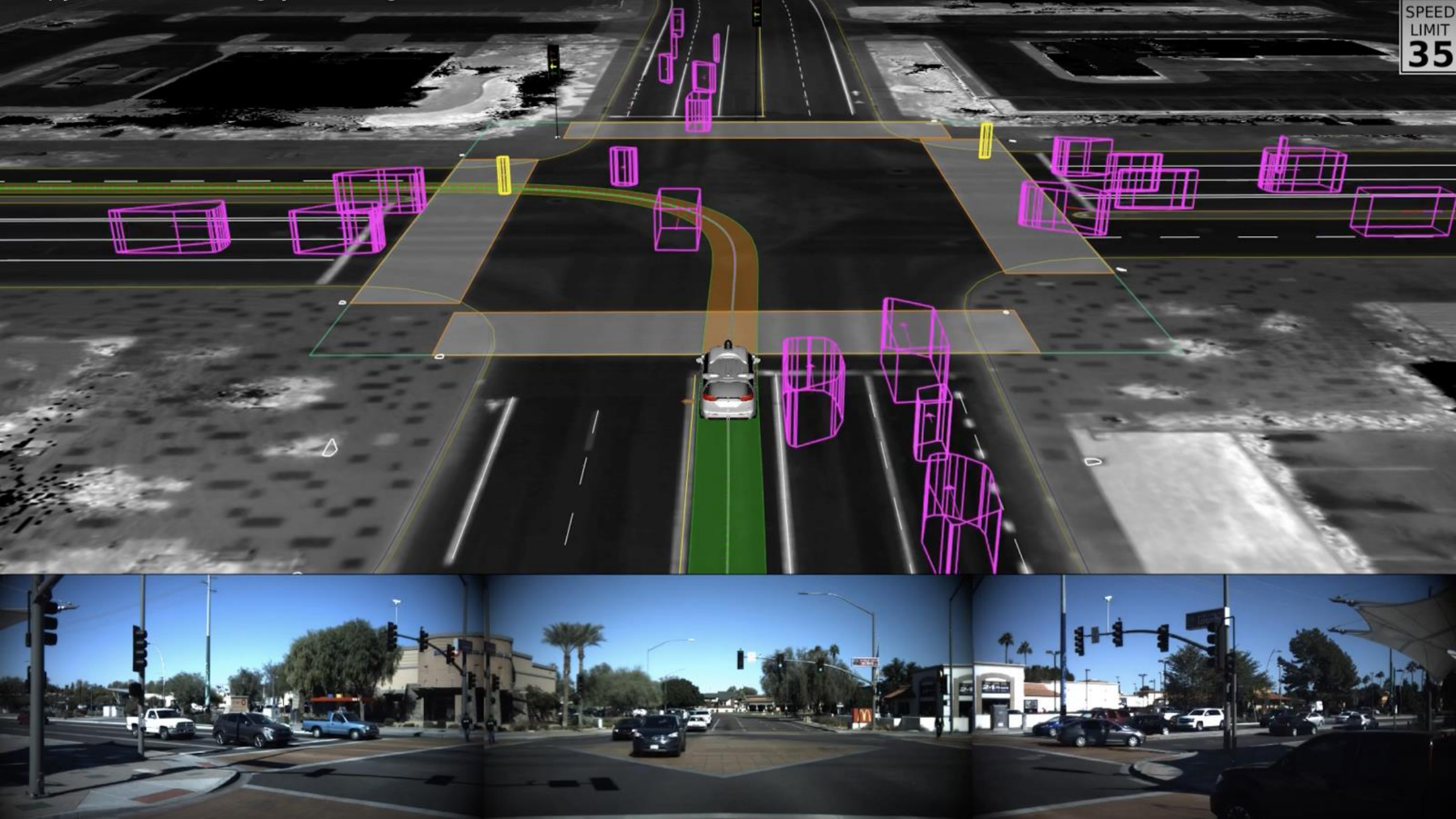


OpenRoadNet

LaneNet

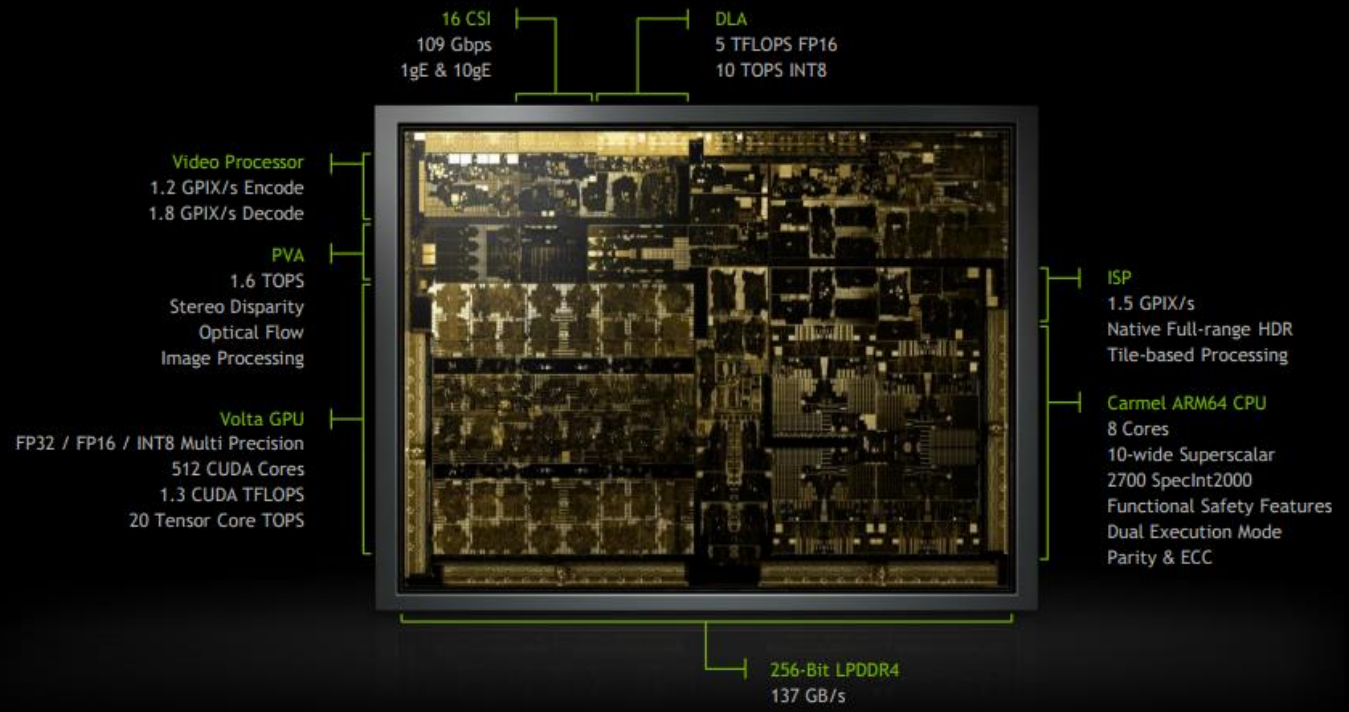


3D Bounding Boxes


SPEED
LIMIT
35

ANNOUNCING
DRIVE XAVIER SAMPLING IN Q1

World's First Autonomous Machine Processor

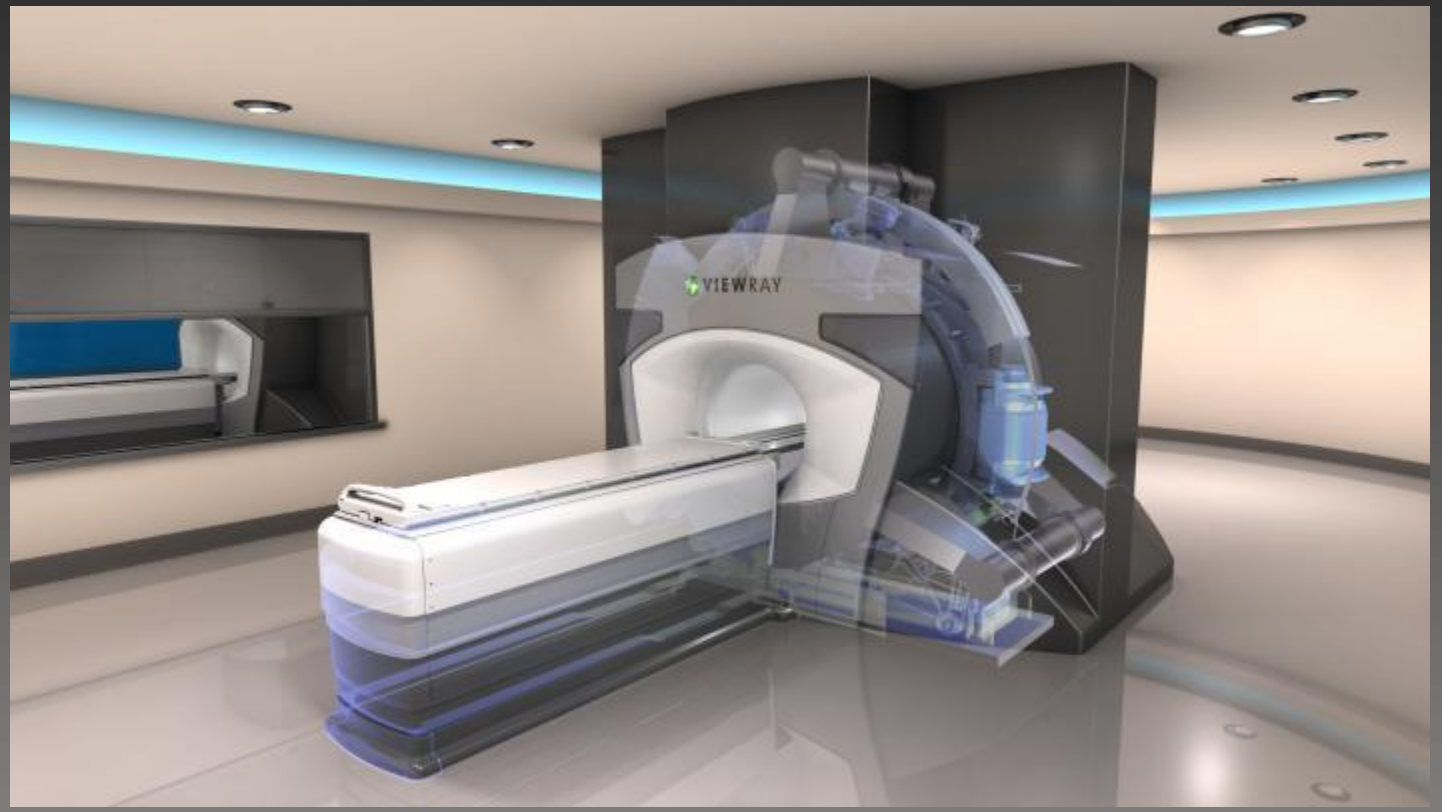


Most Complex SOC Ever Made | 9 Billion Transistors, 350mm^2, 12nFFN | ~8,000 Engineering Years
Diversity of Engines Accelerate Entire AV Pipeline | Designed for ASIL-D AV

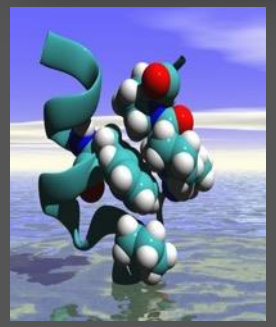
Input image

Radiation Therapy with AI

ViewRay MRIdian (imaging & radiation dose)



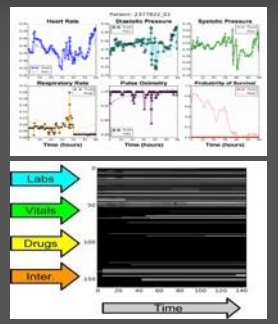
A PLETHORA OF HEALTHCARE STORIES



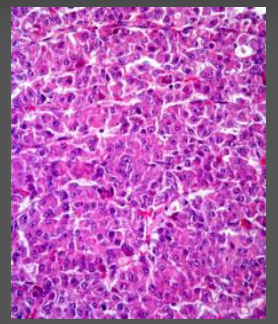
Molecular Energetics For Drug Discovery

AI for Drug Discovery

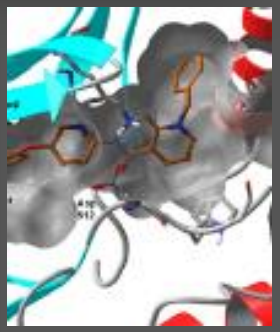
Medical Decision Making



Treatment Outcomes



Reducing Cancer Diagnosis Errors by 85%



Predicting Toxicology

Predicting Growth Problems

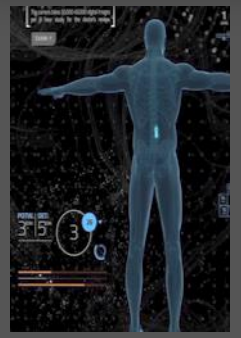
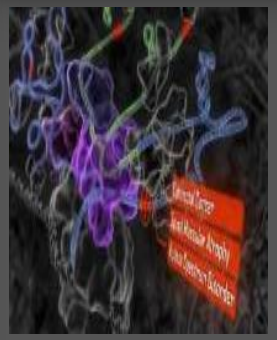
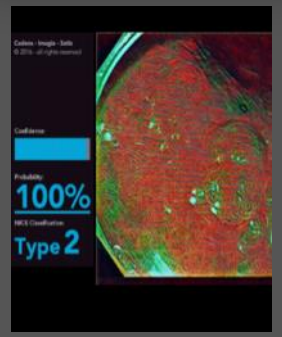


Image Processing

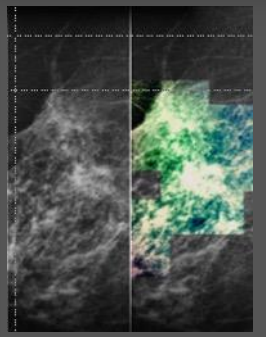


Gene Mutations



Detect Colon Polyps

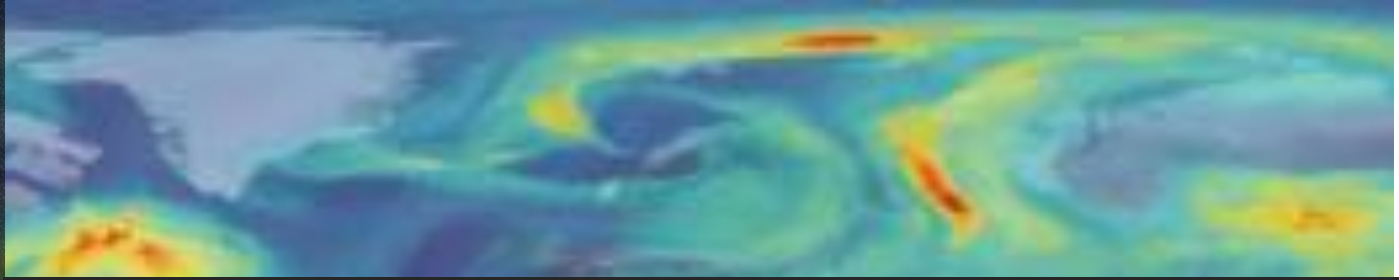
Predicting Disease from Medical Records

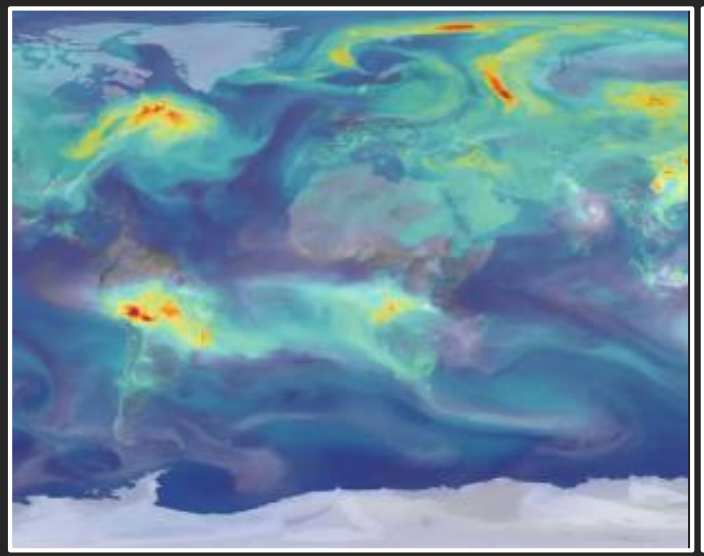


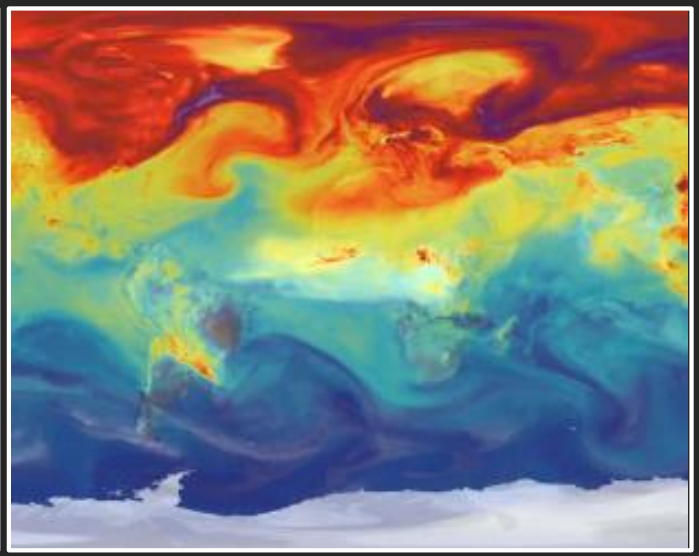
Enabling Detection of Fatty Acid Liver Disease

AN AI MONITOR OF EARTH'S VITALS

The Earth's climate has changed throughout history, but in recent years there have been record increases in temperature, glacial retreat and rising sea levels. NASA Ames is using satellite imagery to measure the effects of carbon and greenhouse gas emissions on the planet. To do so, they developed DeepSat—a deep learning framework for satellite image classification trained on a GPU-powered supercomputer. The enhanced satellite imagery will help scientists plan to protect ecosystems and farmers improve crop production.







NASA: Late summer 2016, forest fires in Africa produce plumes of CO_2
Left: CO_2 - 10/14/2016 / Right: CO_2 - 12/24/2016

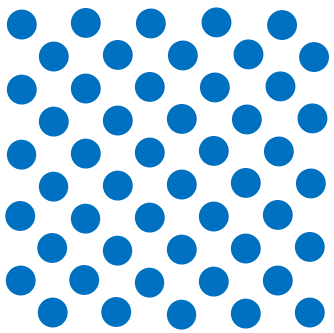
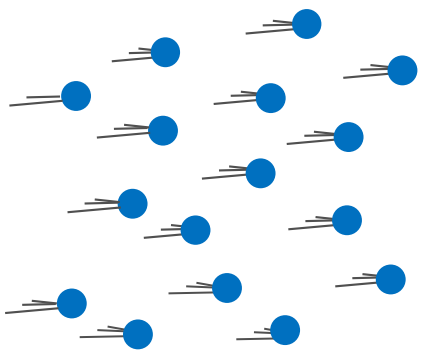
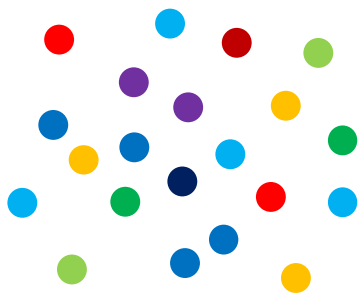
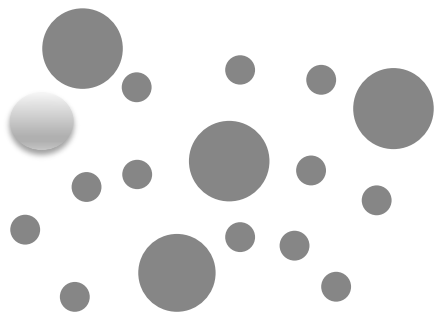
Source: https://climate.nasa.gov/climate_resources/142/

AI IMPROVES THE CUSTOMER EXPERIENCE

AI is dramatically changing the online shopping experience with tangible improvements to retailers and consumers. In 2016 online British grocery giant Ocado improved customer service with their AI-enhanced contact center, and is applying machine learning and NVIDIA GPUs to develop humanoid robotics to assist maintenance technicians, and advanced computer vision for image classification and recognition to replace barcode systems. Computer vision will expedite the picking process and better ensure orders are filled correctly so customers receive exactly what they ordered.

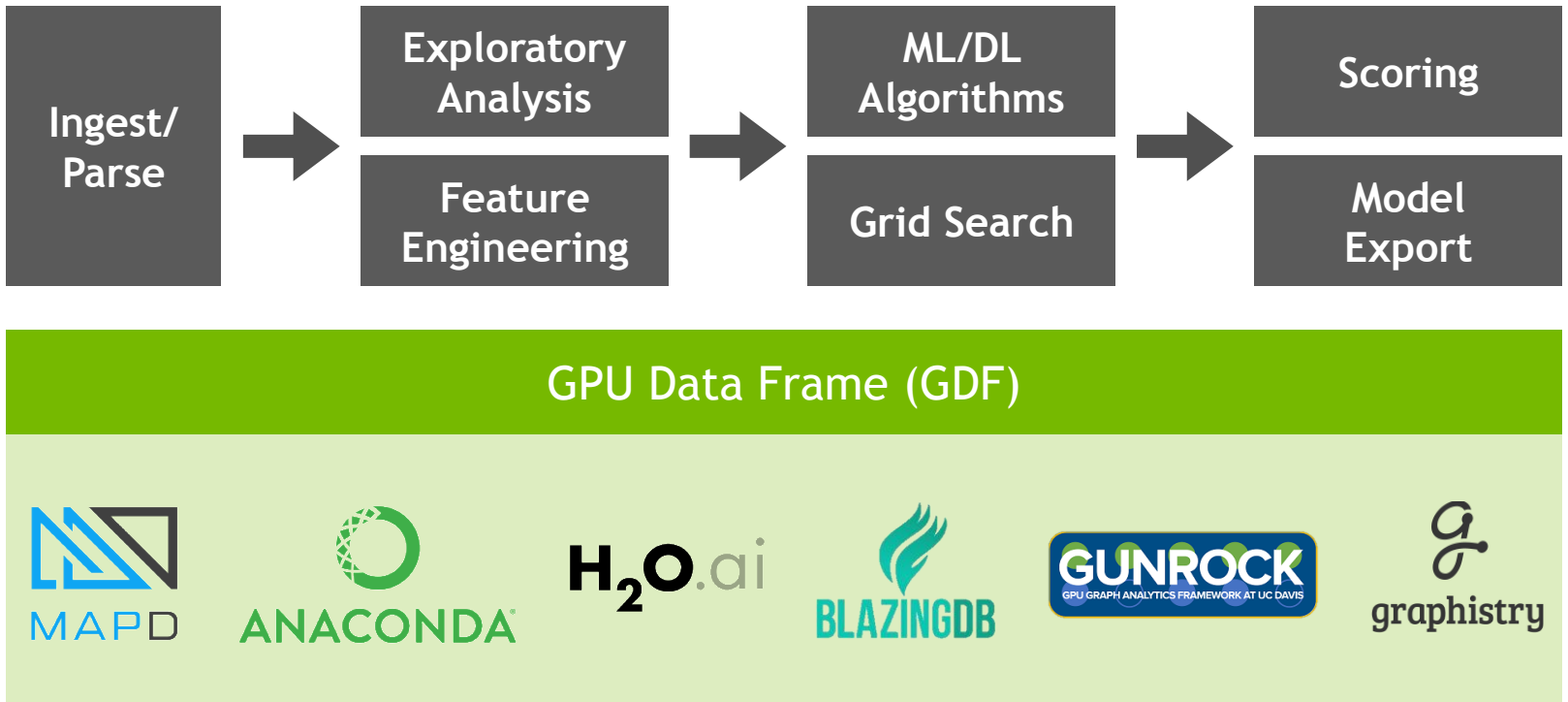
ocado

BIG DATA 4V'S

VOLUME	VELOCITY	VARIETY	VERACITY
			
Sheer amount	Streaming	Data in Many Forms	Is it useful

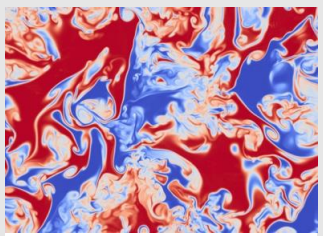
GPU OPEN ANALYTICS INITIATIVE

github.com/gpuopenanalytics

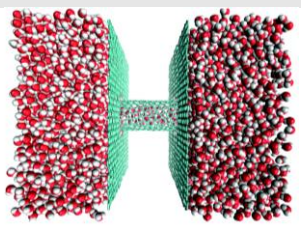


AI SUPERCOMPUTING WILL TRANSFORM HPC

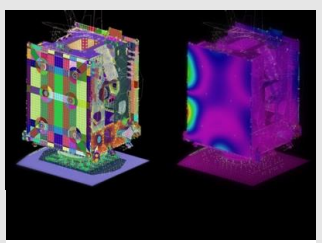
Extending Reach of HPC By Combining Computational & Data Science



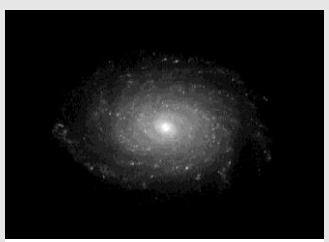
Turbulent Flow



Molecular Dynamics



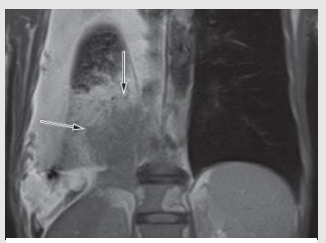
Structural Analysis



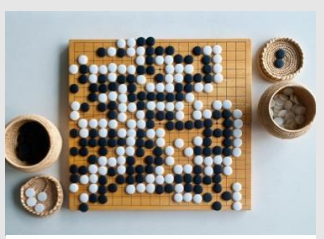
N-body Simulation

COMPUTATIONAL SCIENCE

"What's happening?"



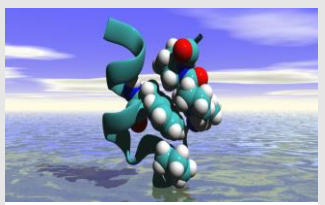
"Is there cancer?"



"Next move?"

"What does she mean?"

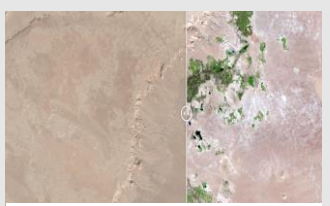
DATA SCIENCE



Drug Discovery

Clean Energy

Understanding Universe



Monitoring Climate Change

COMPUTATIONAL & DATA SCIENCE


NVIDIA
GPU-ACCELERATED
APPLICATIONS

GPU READY APPS

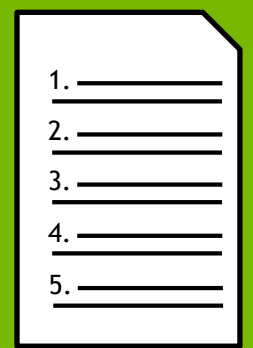
AWARENESS : ADOPTION : UTILIZATION

App Users
Life Science

IT Managers
HPC/Data Centers

Developers
Deep Learning

Target Customers



Simple, ISV-approved, step-by-step instructions

Quick Start Guide

- ✓ Easy setup
- ✓ Best app practices
- ✓ Optimal results
- ✓ Higher productivity
- ✓ Faster discoveries

User Benefits

www.nvidia.com/gpu-ready-apps

RESERVOIR SIMULATION

CPU Based

One Billion Cell Simulation: 21 Hours, 470 Servers

Source: https://blogs.nvidia.com/blog/2017/04/25/how-gpus-can-make-the-most-of-fossil-fuel-resources/

RESERVOIR SIMULATION

GPU Based

One Billion Cell Simulation: 92 Minutes, 30 Servers

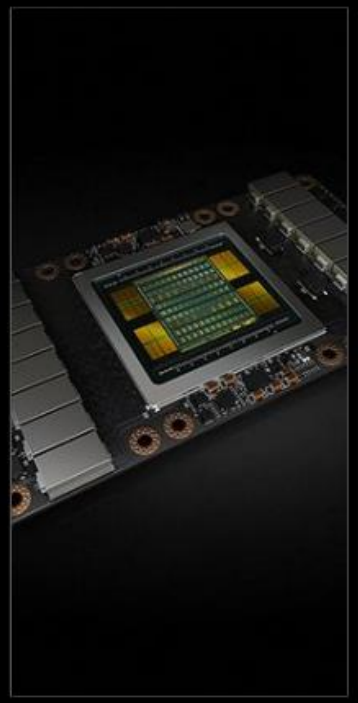
↑100x Simulation Resolution

↑10x Simulation Speed

↓75% Hardware Cost

Source: Simulating a Billion Cell Oil Reservoir with IBM, NVIDIA, and Stone Ridge Technologies | http://www.stoneridgetechnology.com

THE WORLD'S AI PLATFORM

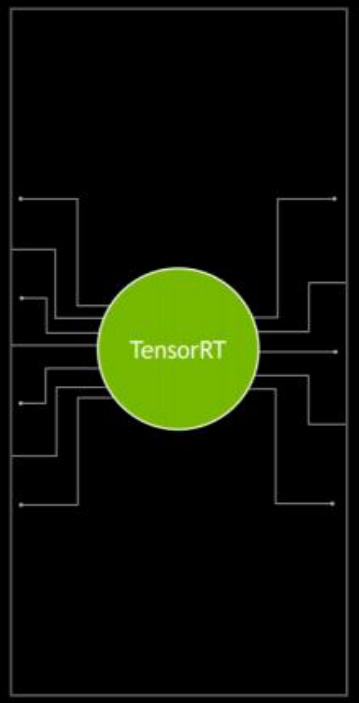


Volta
125 TFLOPS Tensor Core

NVIDIA GPU Cloud
Every Framework

Every Cloud
Every Computer Maker

DGX and DGX Station
1st 1PF AI Supercomputer



TensorRT
Inference Accelerator
Platform

TITAN V
Supercomputer for
Developers

https://devblogs.nvidia.com/parallelforall/inside-volta/

TESLA V100

THE MOST ADVANCED DATA CENTER GPU EVER BUILT

5,120 CUDA cores
640 NEW Tensor cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS

120 Tensor TFLOPS

20MB SM RF | 16MB Cache | 16GB HBM2 @ 900 GB/s
300 GB/s NVLink

TITAN V

THE MOST POWERFUL PC GPU EVER

5,120 CUDA cores
640 NEW Tensor cores

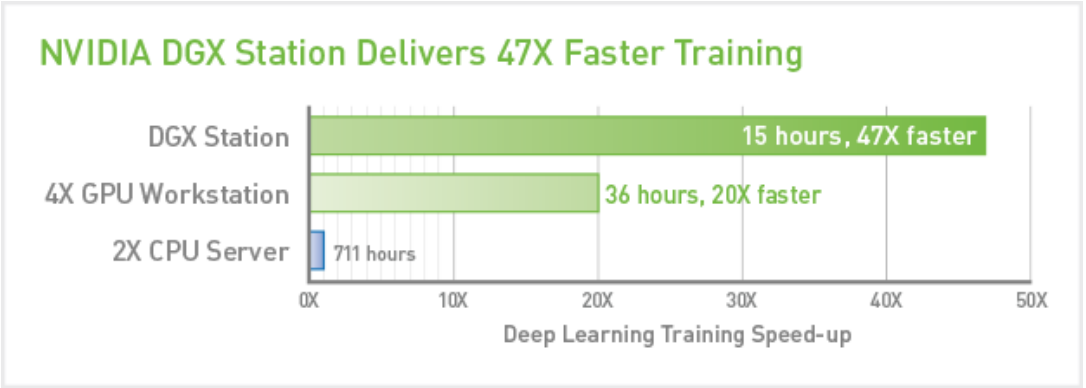
12GB HBM2 Memory 1.7Gbps

110 Tensor TFLOPS

13.8 TFLOPS fp16 | 6.9 fp32

GROUNDBREAKING AI AT YOUR DESK

Designed for your office
DGX Station is the world's first personal supercomputer for leading-edge AI development. Built on the same Deep Learning Stack powering all NVIDIA DGX™ Systems, you can now experiment at your desk and extend your work across DGX Systems and the cloud.

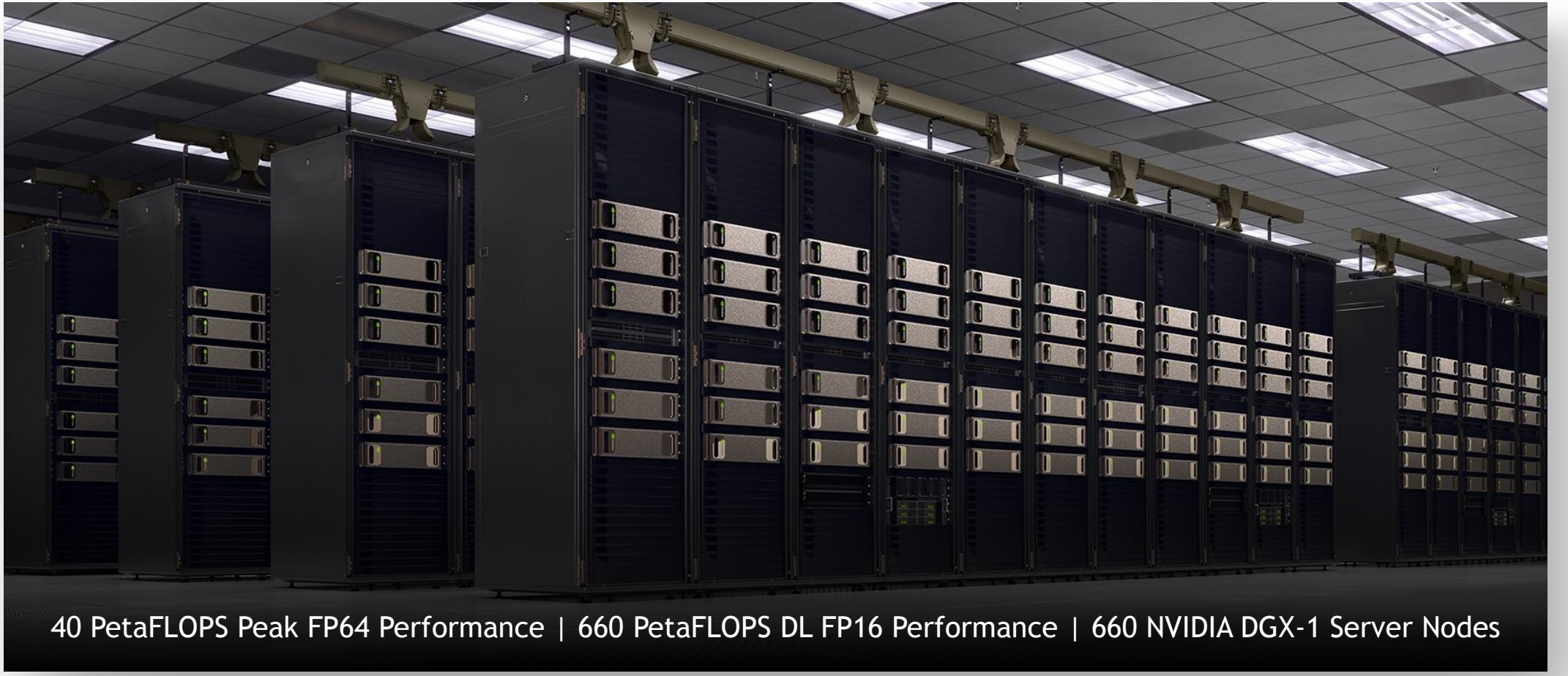


DGX-1 POWERS FASTER, MORE EFFICIENT DRUG DISCOVERY

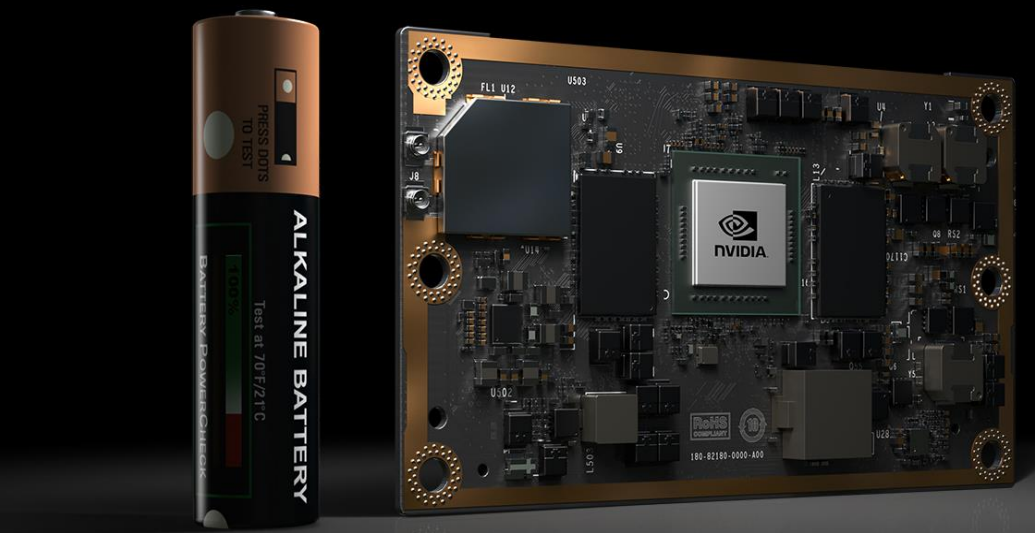
The high cost of drug discovery is driving researchers and pharmaceutical companies to turn to AI as a faster, more efficient way to develop new drugs.

Professor Okuno, Kyoto University and RIKEN, have formed the Life INtelligence Consortium (LINC) to build an AI drug discovery ecosystem in Japan. LINC uses the NVIDIA DGX-1 AI supercomputer—the DGX-1 delivers the extreme performance LINC needs to solve complex problems and speed drug discovery.

ANNOUNCING
NVIDIA SATURNV WITH VOLTA



NVIDIA JETSON TX2



EMBEDDED AI SUPERCOMPUTER

Advanced AI at the edge

JetPack SDK

< 7.5 watts full module

Up to 2X performance or 2X energy efficiency

Max-Q operating mode (< 7.5 watts) delivers up to 2x energy efficiency vs. Jetson TX1 maximum performance
Max- P operating mode (< 15 watts) delivers up to 2x performance vs. Jetson TX1 maximum performance

NVIDIA DEEP LEARNING SOFTWARE PLATFORM

GATHER AND LABEL

Gather Data

Rapidly label data, guide training get insights

Curate data sets

TRAINING

TRAINING DATA

DATA MANAGEMENT

TRAINING

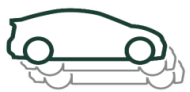
MODEL ASSESSMENT

TRAINED NETWORK

CNN
RNN
FC

DEPLOY WITH TENSORRT

EMBEDDED
Jetson TX



AUTOMOTIVE
Drive PX (XAVIER)

DATA CENTER
Tesla (Pascal, Volta)

NVIDIA DEEP LEARNING SDK

NVIDIA DEVELOPER

Log in

NVIDIA SDK

The Essential Resource for GPU Developers

NVIDIA SDK

developer.nvidia.com

DEEP LEARNING

Deep Learning SDK

High-performance tools and libraries for deep learning

SELF-DRIVING CARS

NVIDIA DriveWorks™

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

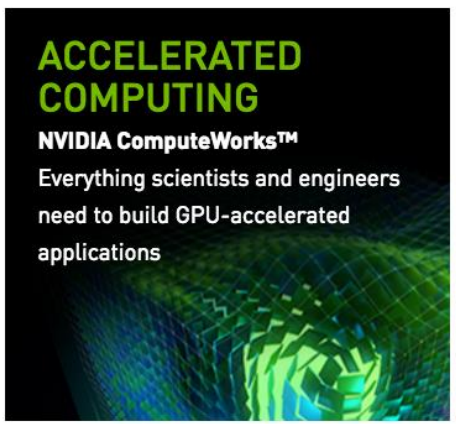
NVIDIA GameWorks™

Advanced simulation and rendering technology for game development

ACCELERATED COMPUTING

NVIDIA ComputeWorks™

Everything scientists and engineers need to build GPU-accelerated applications



DESIGN & VISUALIZATION

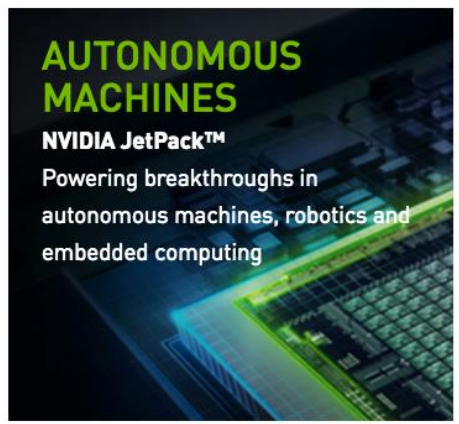
NVIDIA DesignWorks™

Tools and technologies to create professional graphics and advanced rendering applications

AUTONOMOUS MACHINES

NVIDIA JetPack™

Powering breakthroughs in autonomous machines, robotics and embedded computing



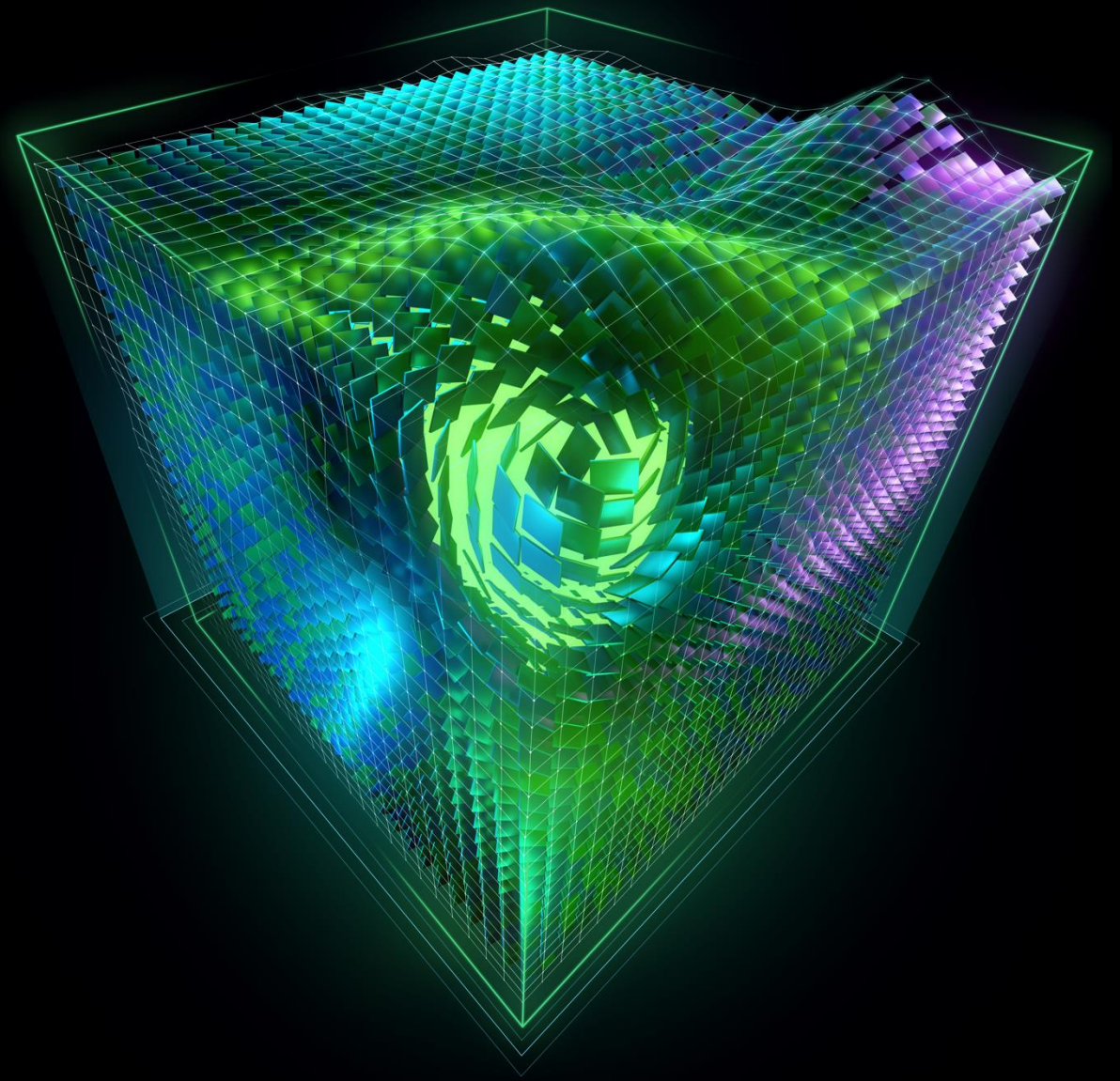
ADDITIONAL RESOURCES

More resources for GPU Developers

https://devblogs.nvidia.com/parallelforall/cuda-9-features-revealed/

GPU TECHNOLOGY CONFERENCE

CUDA 9

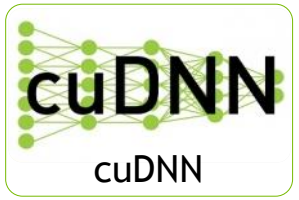


PRESENTED BY NVIDIA

GPU ACCELERATED LIBRARIES

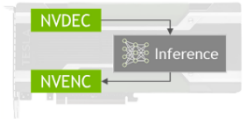
"Drop-in" Acceleration for Your Applications

DEEP LEARNING



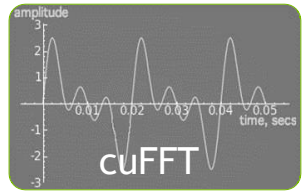
cuDNN


TensorRT



DeepStream SDK

SIGNAL, IMAGE & VIDEO



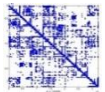
cuFFT


NVIDIA NPP


CODEC SDK

LINEAR ALGEBRA


cuBLAS

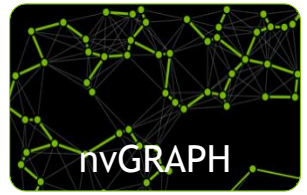

cuSPARSE

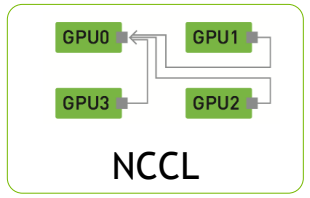

CUDA Math library


cuSOLVER

cuRAND

PARALLEL ALGORITHMS


nvGRAPH



NCCL

OpenACC is a directives-based programming approach to **parallel computing** designed for **performance** and **portability** on CPUs and GPUs for HPC.

Add Simple Compiler Directive

```
main()
 {
  <serial code>
  #pragma acc kernels
  {
    <parallel code>
  }
}
```

THE POWER TO RUN MULTIPLE FRAMEWORKS AT ONCE

Container Images portable across new driver versions

Containerized Applications

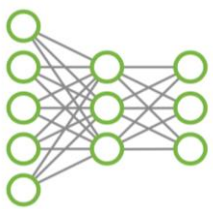
NVIDIA Docker	NVIDIA Docker	NVIDIA Docker	NVIDIA Docker		NVIDIA Docker
TensorFlow	Microsoft Cognitive Toolkit	Caffe2	PYTORCH	• • •	Other Frameworks and Apps
TF Tuned SW	CNTK Tuned SW	Caffe2 Tuned SW	Pytorch Tuned SW		Tuned SW
CUDA RT	CUDA RT	CUDA RT	CUDA RT		CUDA RT

Linux Kernel + CUDA Driver

NVIDIA ® DGX-1™

NVIDIA DEEP LEARNING SDK UPDATE

GPU-accelerated DL Primitives



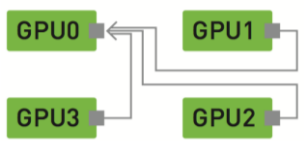
cuDNN 7

Faster training

Optimizations for RNNs

Leading frameworks support

Multi-GPU & Multi-node



NCCL 2

Multi-node distributed training (multiple machines)

Leading frameworks support

High-performance Inference Engine

TensorRT 3

TensorFlow model reader

Object detection

INT8 RNNs support

NVIDIA TensorRT 3

Programmable Inference Accelerator

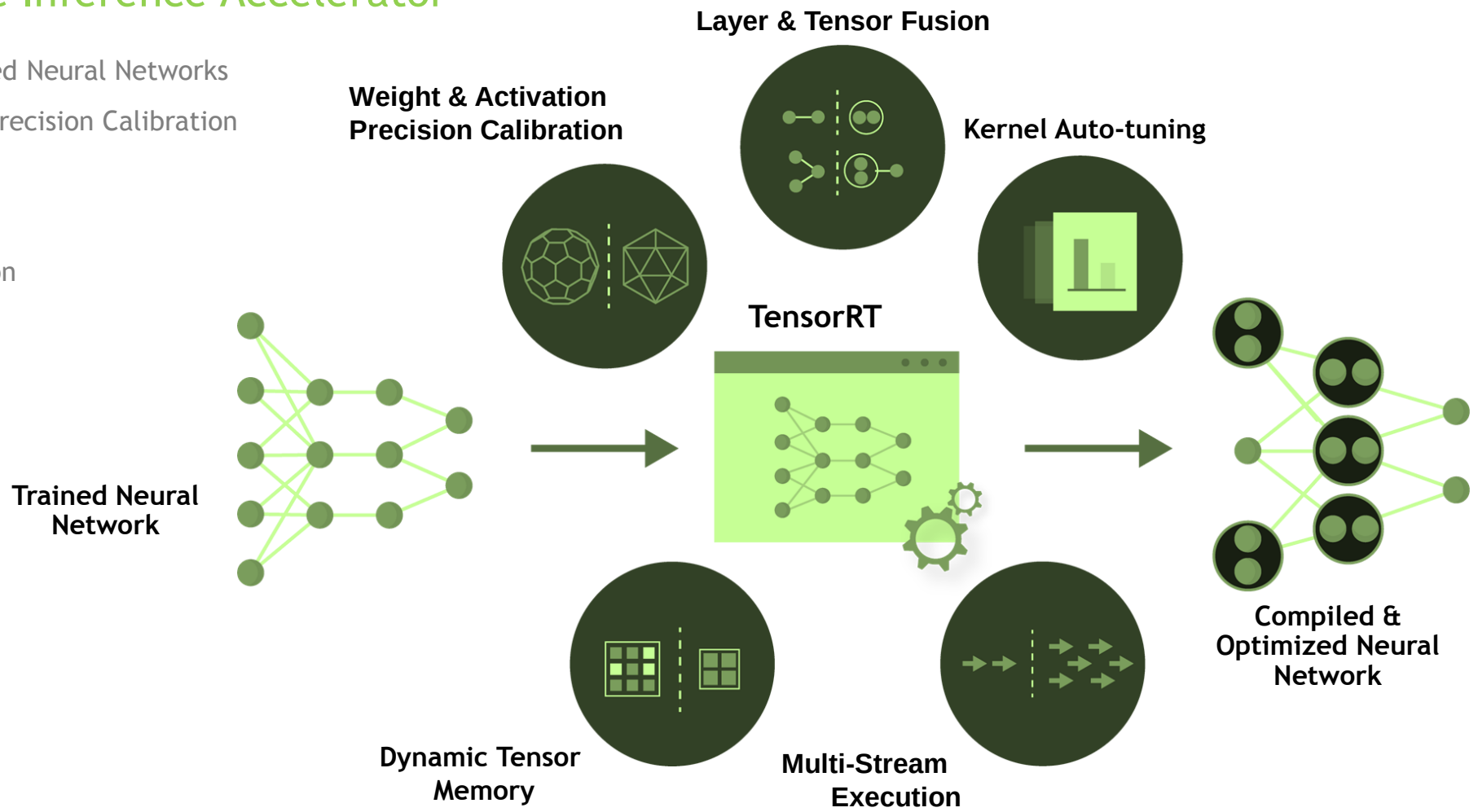
Compiler for Optimized Neural Networks

Weight & Activation Precision Calibration

Layer & Tensor Fusion

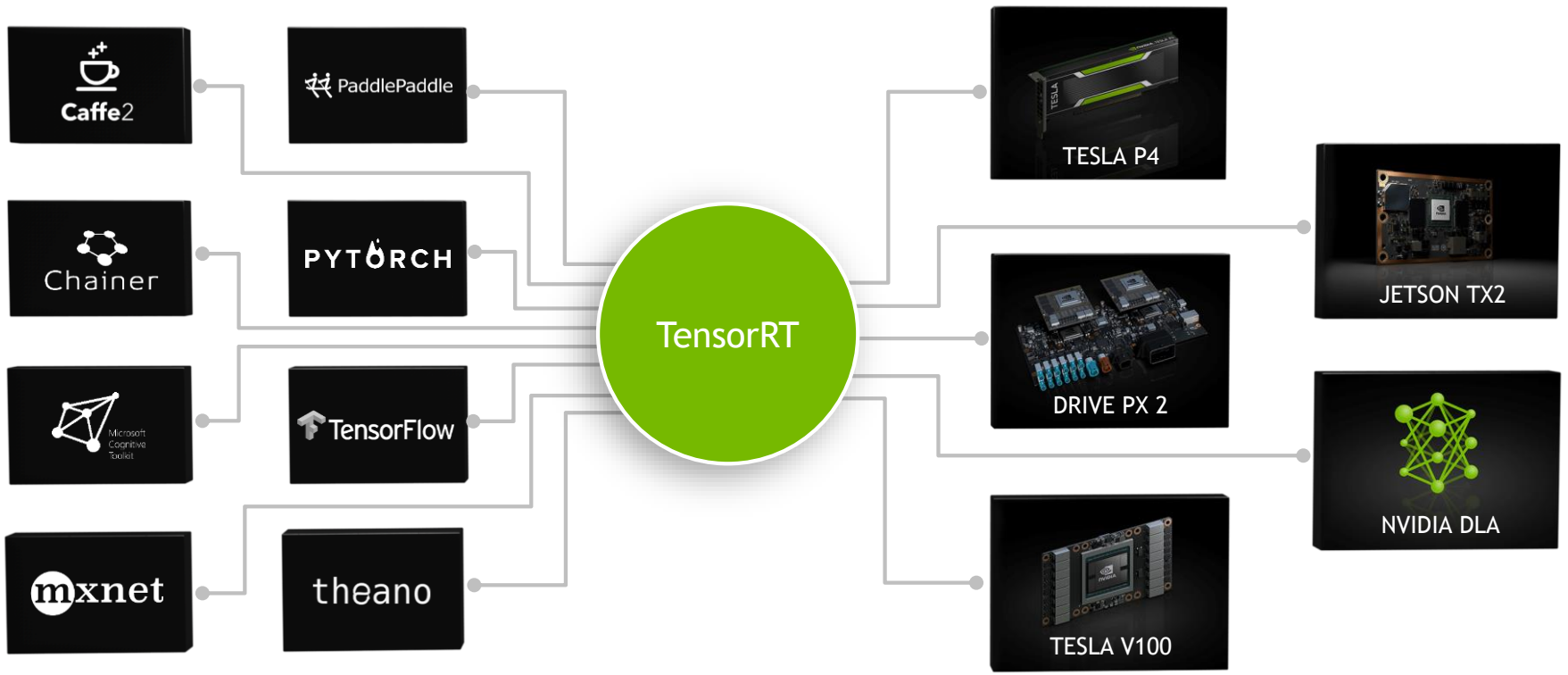
Kernel Auto-Tuning

Multi-Stream Execution



NEW NVIDIA TENSORRT 3

Programmable Inference Accelerator

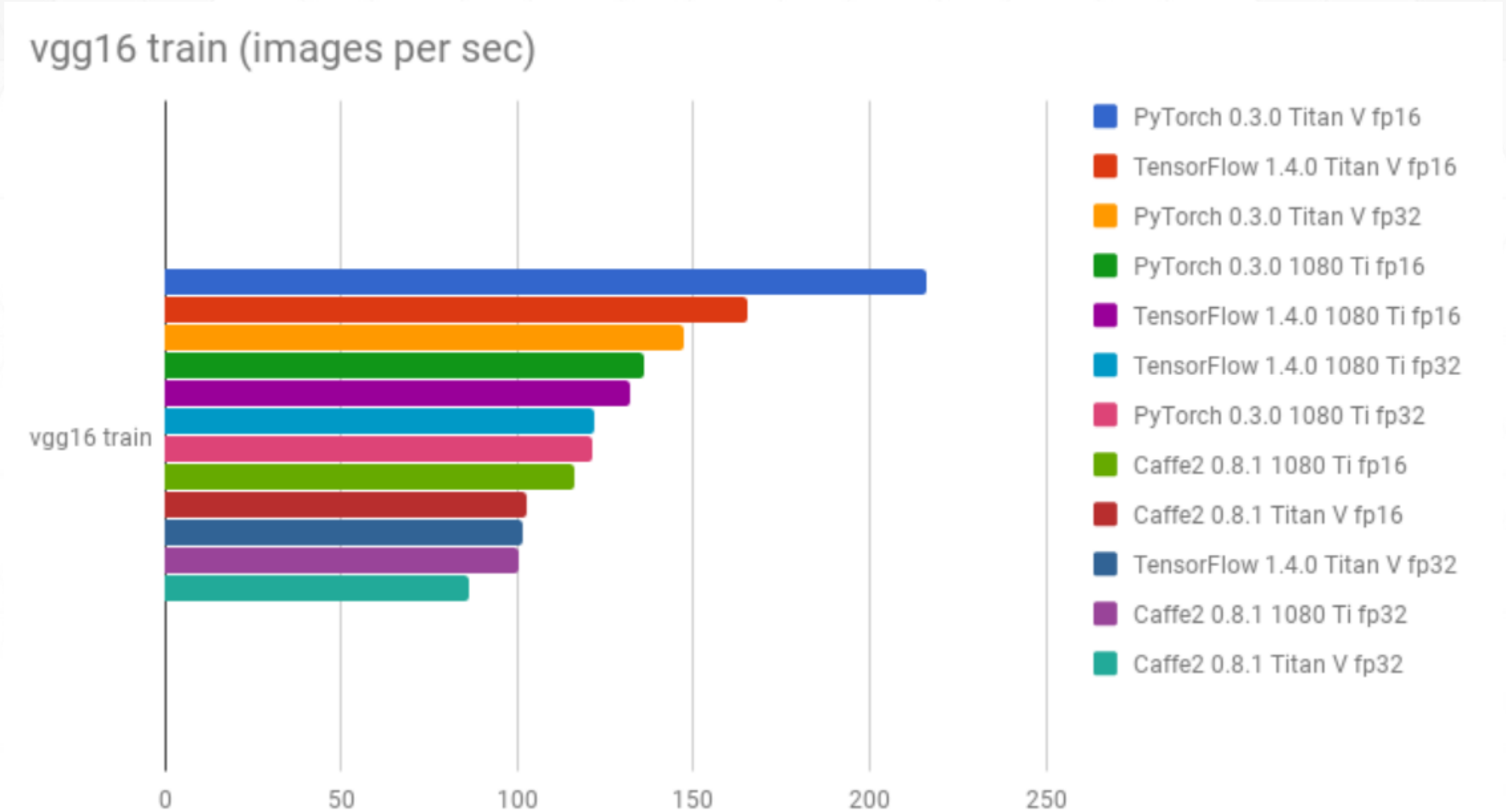


Compile and Optimize Neural Networks | Support for Every Framework
Optimize for Each Target Platform


Skylake
TensorFlow on CPU
Images Per Sec: 5.0

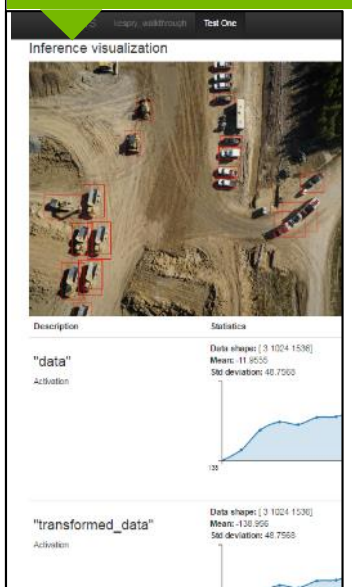
Benchmarks Jan 2018

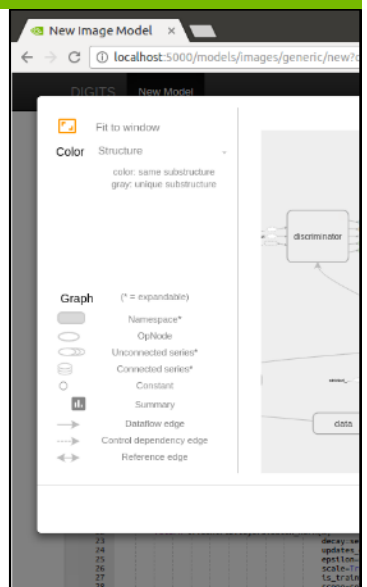
https://github.com/u39kun/deep-learning-benchmark



WHAT'S NEW IN DIGITS 6?

TENSORFLOW SUPPORT





Train TensorFlow Models Interactively with DIGITS

NEW PRE-TRAINED MODELS

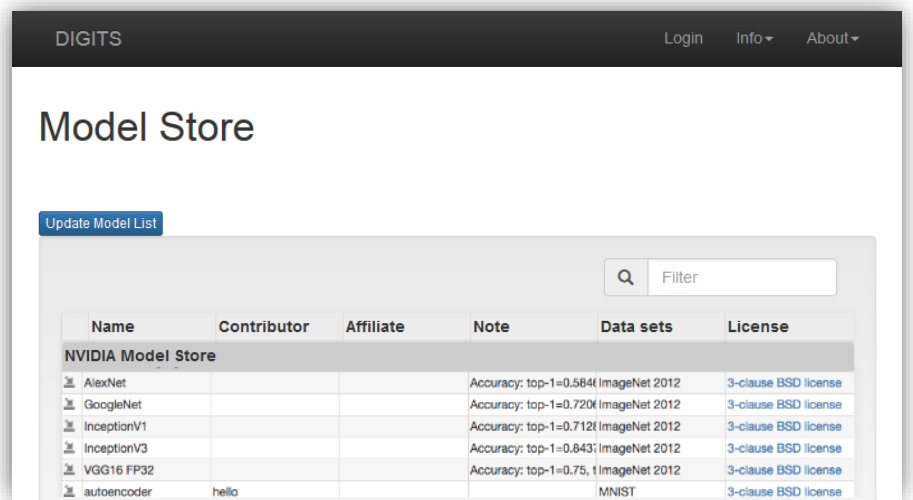
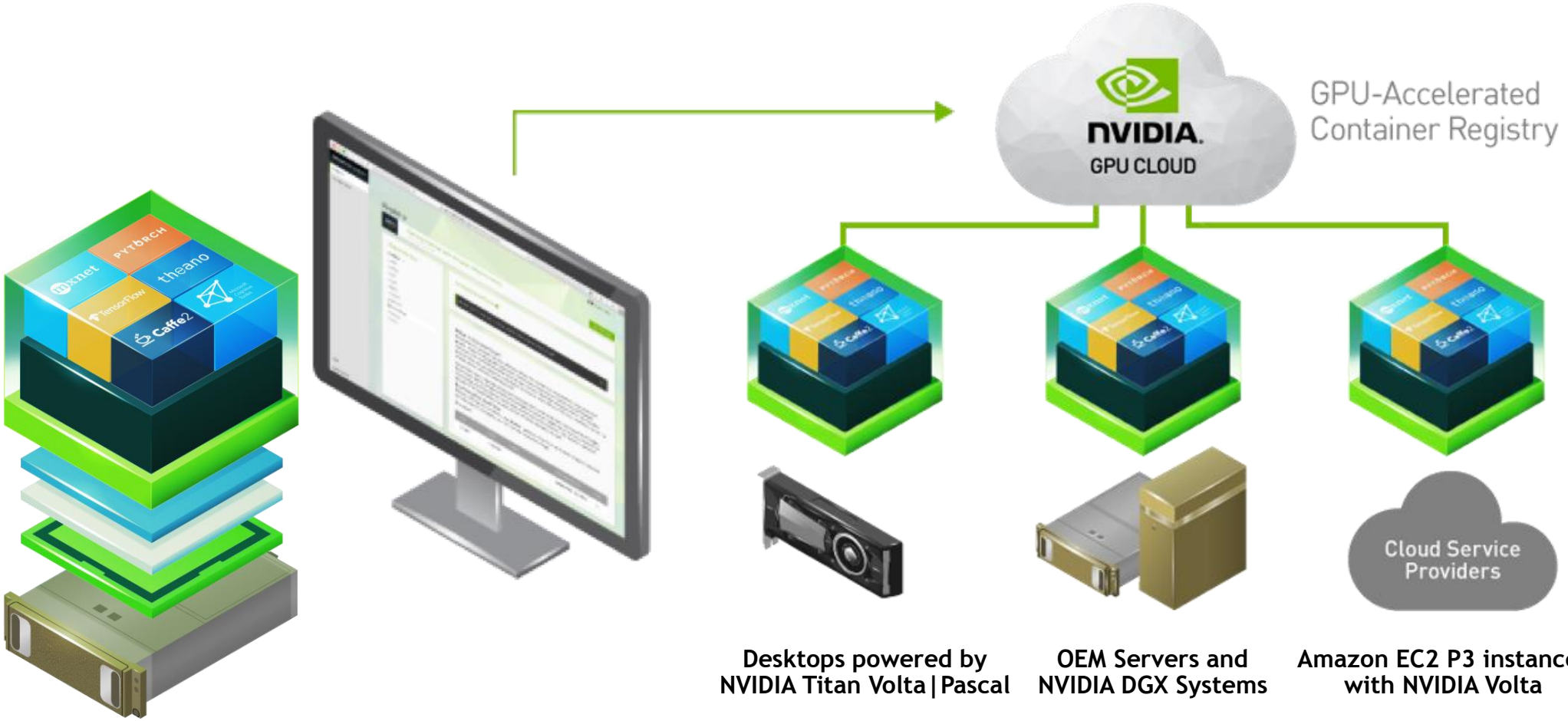


Image Classification: VGG-16, ResNet50
Object Detection: DetectNet

DEEP LEARNING ACROSS PLATFORMS WITH NGC

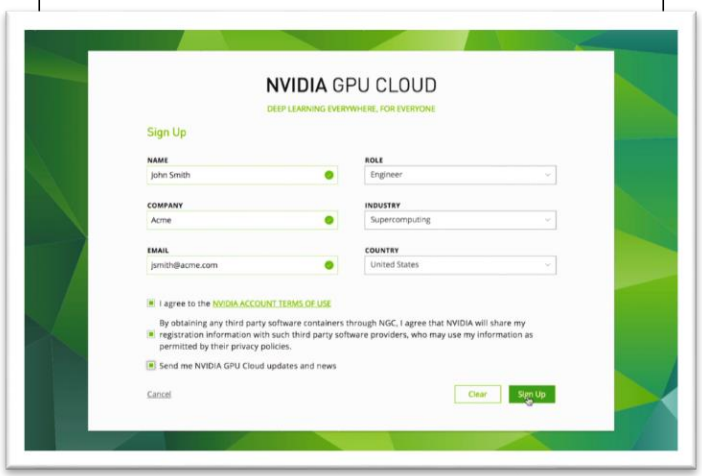


THREE STEPS TO DEEP LEARNING WITH NGC

SIGN UP

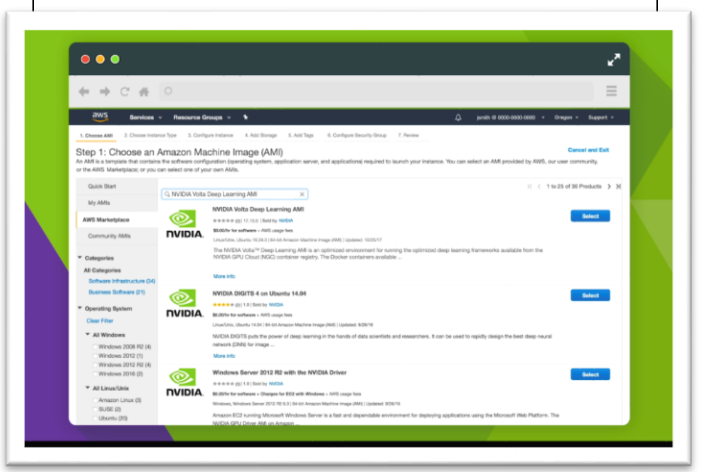
To get an NGC account, go to:

www.nvidia.com/ngcsignup



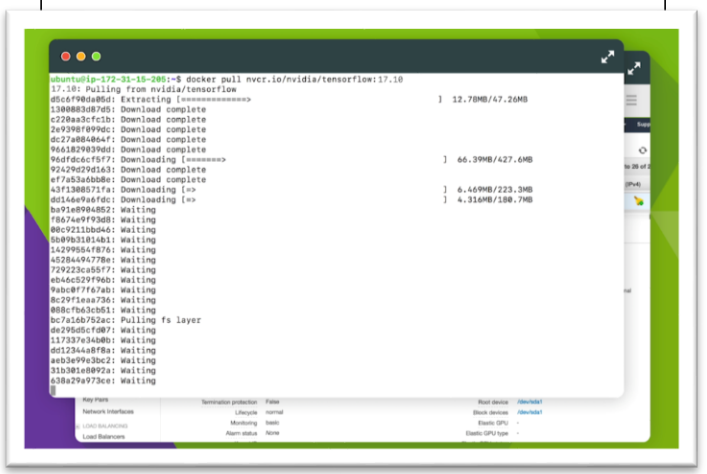
DEPLOY IMAGE

On Amazon EC2, choose a P3 instance and deploy the **NVIDIA Volta Deep Learning AMI for NGC**



PULL CONTAINER

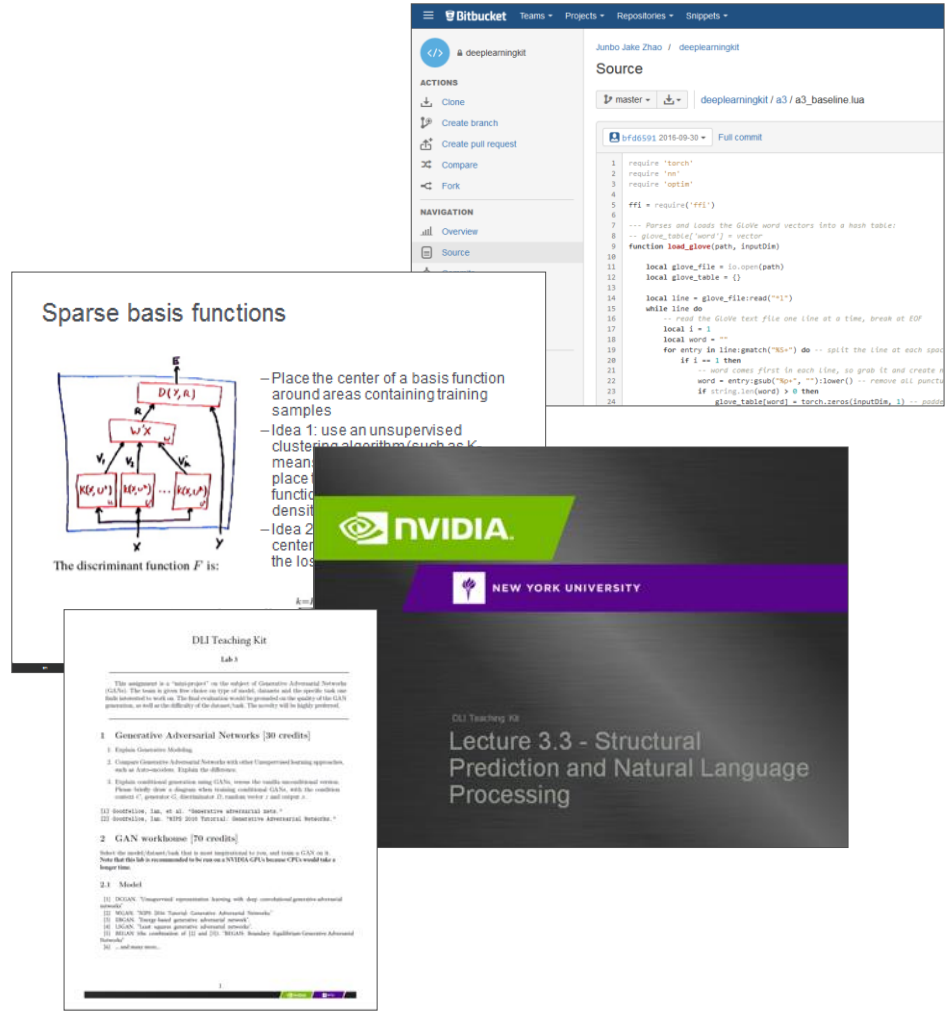
Pick your desired framework (TensorFlow, PyTorch, MXNet, etc.), and pull the container into your instance



TEACHING KITS

The Deep Learning Institute also makes teaching kits available to qualified educators.

These kits are not available for commercial users.



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**

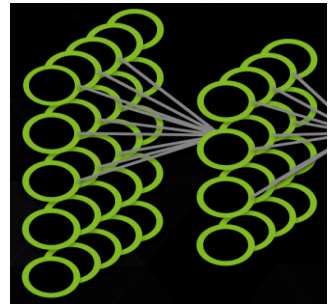
Caffe2

Microsoft Cognitive Toolkit

mxnet

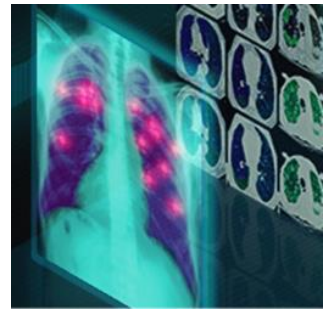
TensorFlow

PYTORCH

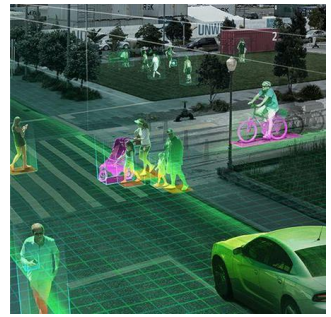


Fundamentals

Autonomous Vehicles



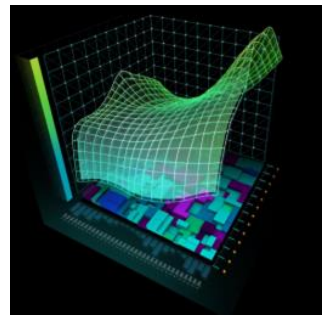
Healthcare



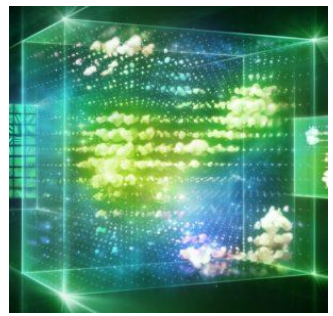
Intelligent Video Analytics

Robotics

Game Development & Digital Content



Finance



Accelerated Computing

Virtual Reality

NVIDIA DEEP LEARNING INSTITUTE

NVIDIA INCEPTION PROGRAM

Accelerates AI startups with a boost of GPU tools, tech and deep learning expertise

www.nvidia.com/inception

Startup Qualifications

Driving advances in the field of AI

Business plan

Incorporated

Web presence

Technology

DL startup kit*

- Pascal Titan X
- Deep Learning Institute (DLI) credit
- Connect with a DL tech expert

DGX-1 ISV discount*

Software release notification

Live webinar and office hours

**By application*

Marketing

Inclusion in NVIDIA marketing efforts

GPU Technology Conference (GTC) discount

Emerging Company Summit (ECS) participation+

Marketing kit

- One-page story template
- eBook template
- Inception web badge and banners
- Social promotion request form
- Event opportunities list

Promotion at industry events

GPU ventures+

+By invitation

NVIDIA

GPU TECHNOLOGY CONFERENCE

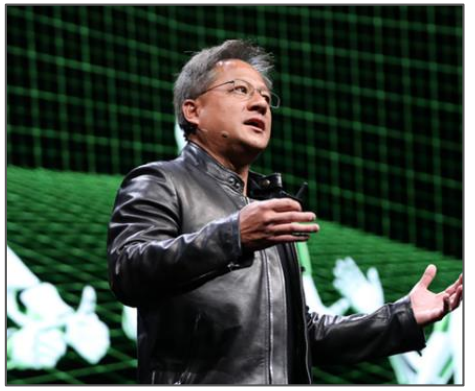
March 26—29, 2018 | Silicon Valley | #GTC18
www.gputechconf.com

CONNECT

Connect with technology experts from NVIDIA and other leading organizations

LEARN

Gain insight and valuable hands-on training through hundreds of sessions and research posters



DISCOVER

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

INNOVATE

Hear about disruptive innovations as early-stage companies and startups present their work

REGISTER using code 'NVALOWNDES' for a 25% DISCOUNT at www.GPUTECHCONF.COM

Don't miss the world's most important event for GPU developers
March 26—29, 2018 in Silicon Valley


NVIDIA®