RISE OF MODERN AI WITH DEEP LEARNING IN INDUSTRY AND ROBOTICS

Marc Stampfli, Country Sales Manager, Switzerland, 6.9.18

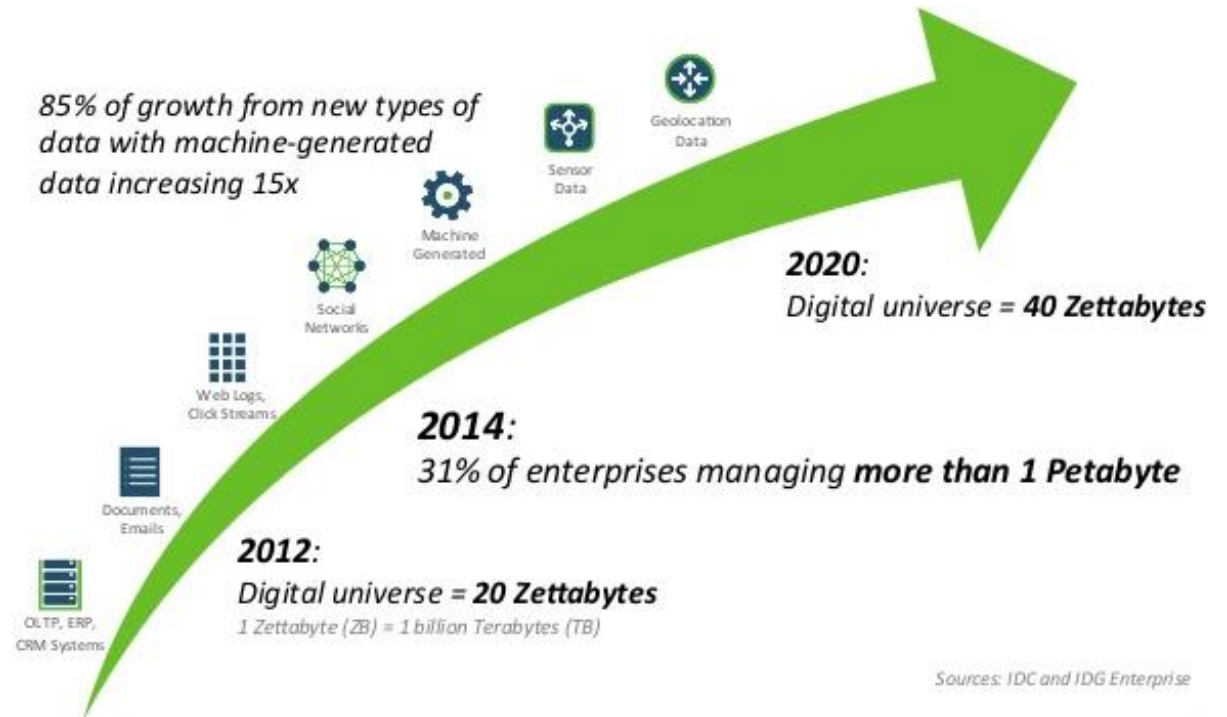E-Mail: mstampfli@nvidia.com / LinkedIn: https://ch.linkedin.com/in/marcstampfli

WHAT HAPPENED?

BIG DATA AND REAL WORLD DATA

## TODAY DATA IS EVERYWHERE AND DEVICES GENERATE IT

- Like the physical universe, the digital universe is large – by 2020 containing nearly as many digital bits as there are stars in the universe. It is **doubling in size every two years**, and by 2020 the digital universe – the data we create and copy annually – will reach 40 zettabytes, or 40 trillion gigabytes.

- Every microphone, every camera, every sensor in general, every IoT device is producing machine generated data – abundance of data is the fuel for artificial intelligence.

85% of growth from new types of data with machine-generated data increasing 15x

Geolocation Data

Sensor Data

Machine Generated

Social Networks

Web Logs, Click Streams

Documents, Emails

OLTP, ERP, CRM Systems

2020:
Digital universe = **40 Zettabytes**

2014:
31% of enterprises managing **more than 1 Petabyte**

2012:
Digital universe = **20 Zettabytes**
1 Zettabyte (ZB) = 1 billion Terabytes (TB)

Sources: IDC and IDG Enterprise

NVIDIA.

NEURAL NETWORK COMPLEXITY IS EXPLODING

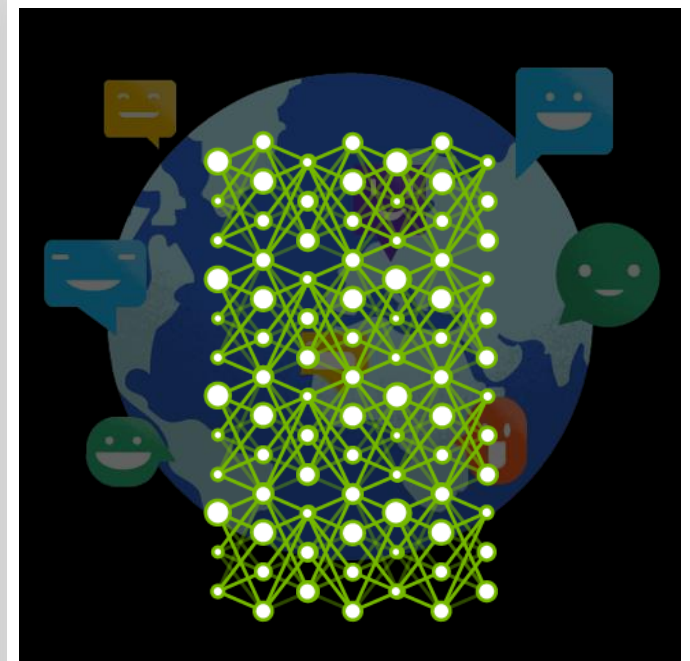## To Tackle Increasingly Complex Challenges

7 ExaFLOPS
60 Million Parameters

20 ExaFLOPS
300 Million Parameters

100 ExaFLOPS
8700 Million Parameters
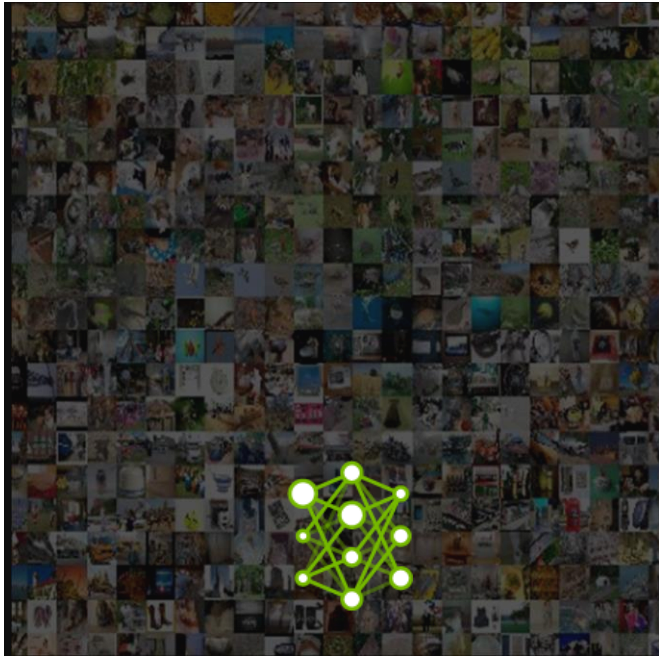
**2015 – Microsoft ResNet
Superhuman Image Recognition**

**2016 – Baidu Deep Speech 2
Superhuman Voice Recognition**

**2017 – Google Neural Machine Translation
Near Human Language Translation**

RISE OF GPU COMPUTING

GPU-Accelerated Computing

10^7

10^5 1.1X per year

10^3

1.5X per year

Single-threaded perf

1980   1990   2000   2010   2020
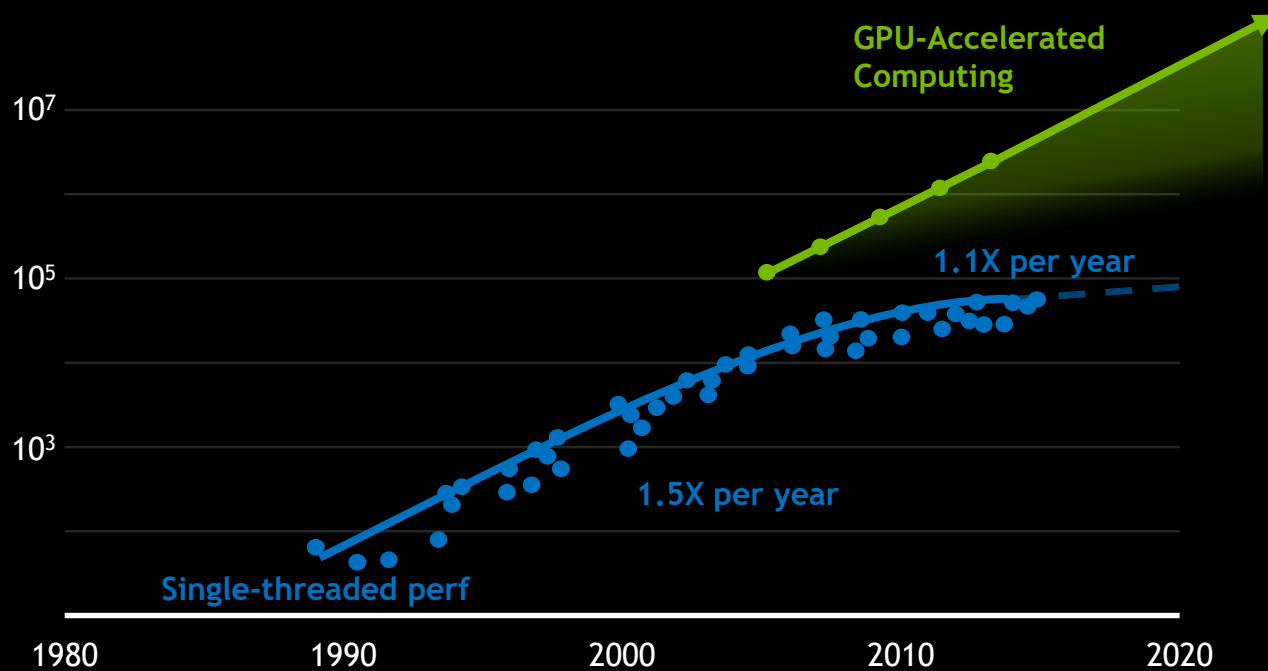
40 Years of CPU Trend Data

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

878,000
2013   2018
GPU Developers
10X in 5 Yrs

8M
2013   2018
CUDA Downloads
5X in 5 Yrs

8,500
2013   2018
GTC Registrations
4X in 5 Yrs

370PF
2013   2018
Total GPU FLOPS of Top 50 Systems
15X in 5 Yrs
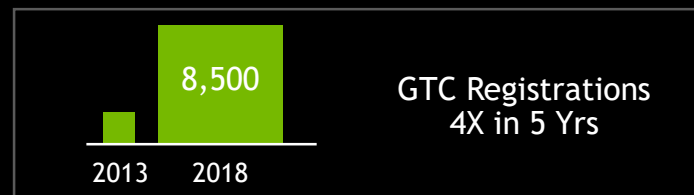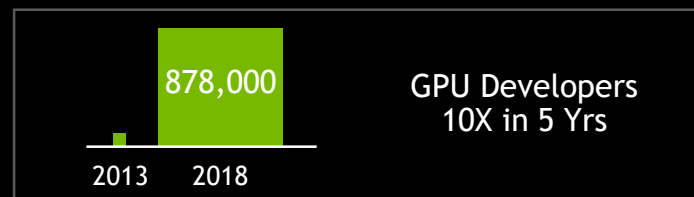
PARALLEL COMPUTING WITH GPU
What makes a GPU different? Parallelization on a chip

- A core in a chip is the processing unit which receives instructions and performs calculations

- Clock rate refers to the frequency at which one core of a multi-core processor is running

- More cores means more calculations per clock cycle

- CPU optimized for sequential serial processing of complex orders

- GPU optimized for massive parallel processing of calculations

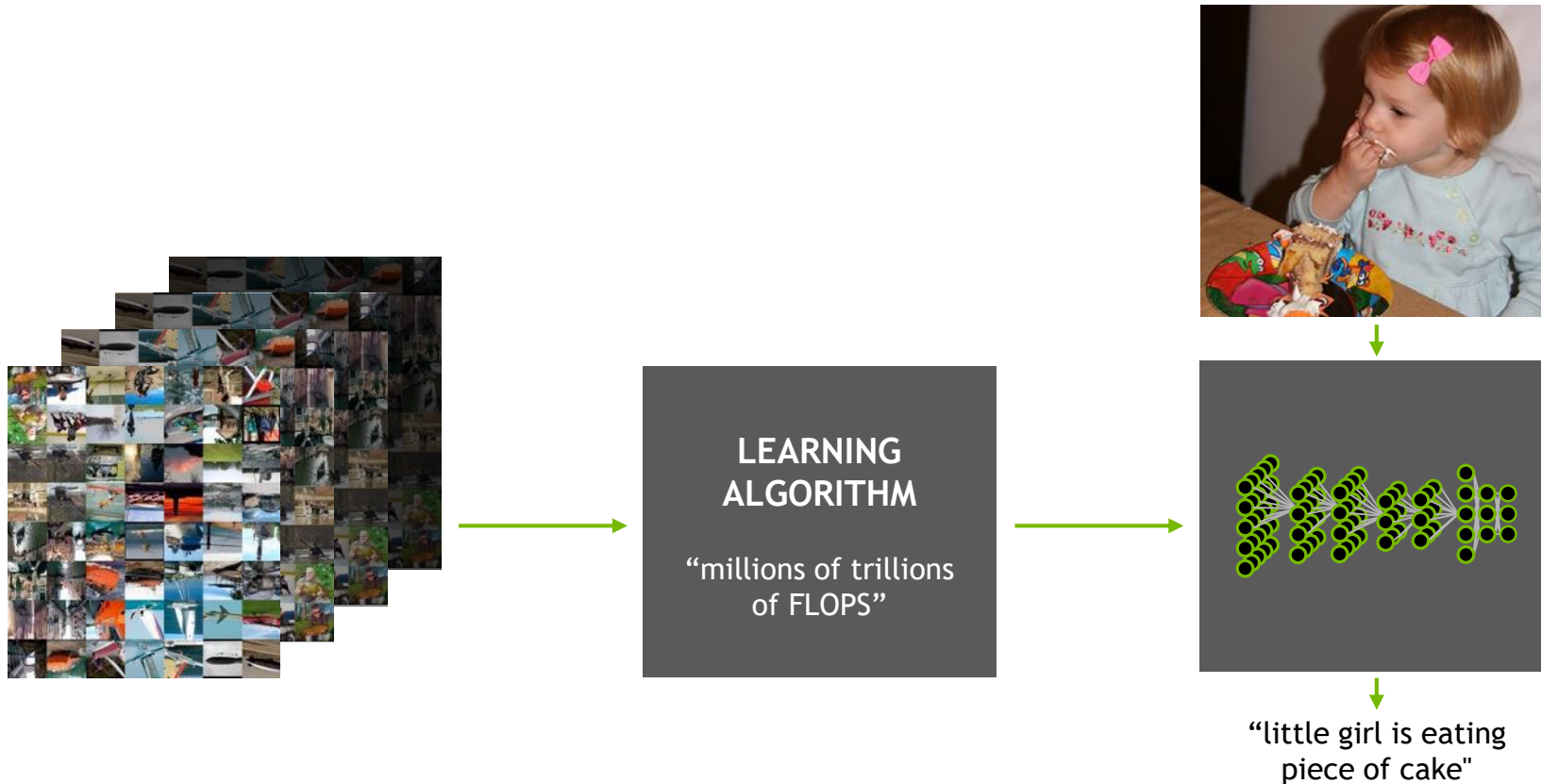CPU with multiple Cores

e.g. 12-20 Cores
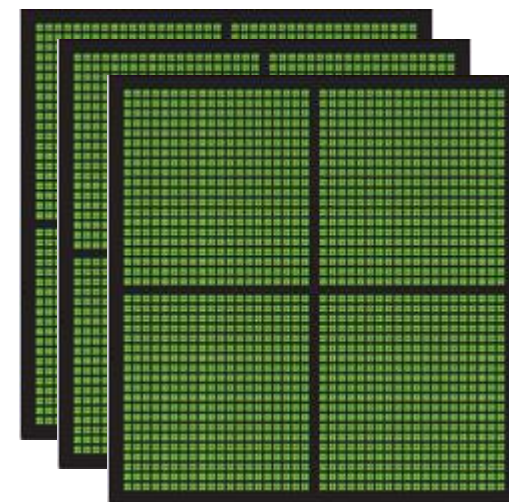
GPU with n-times Thousands of Cores

e.g. n x 5120 cores
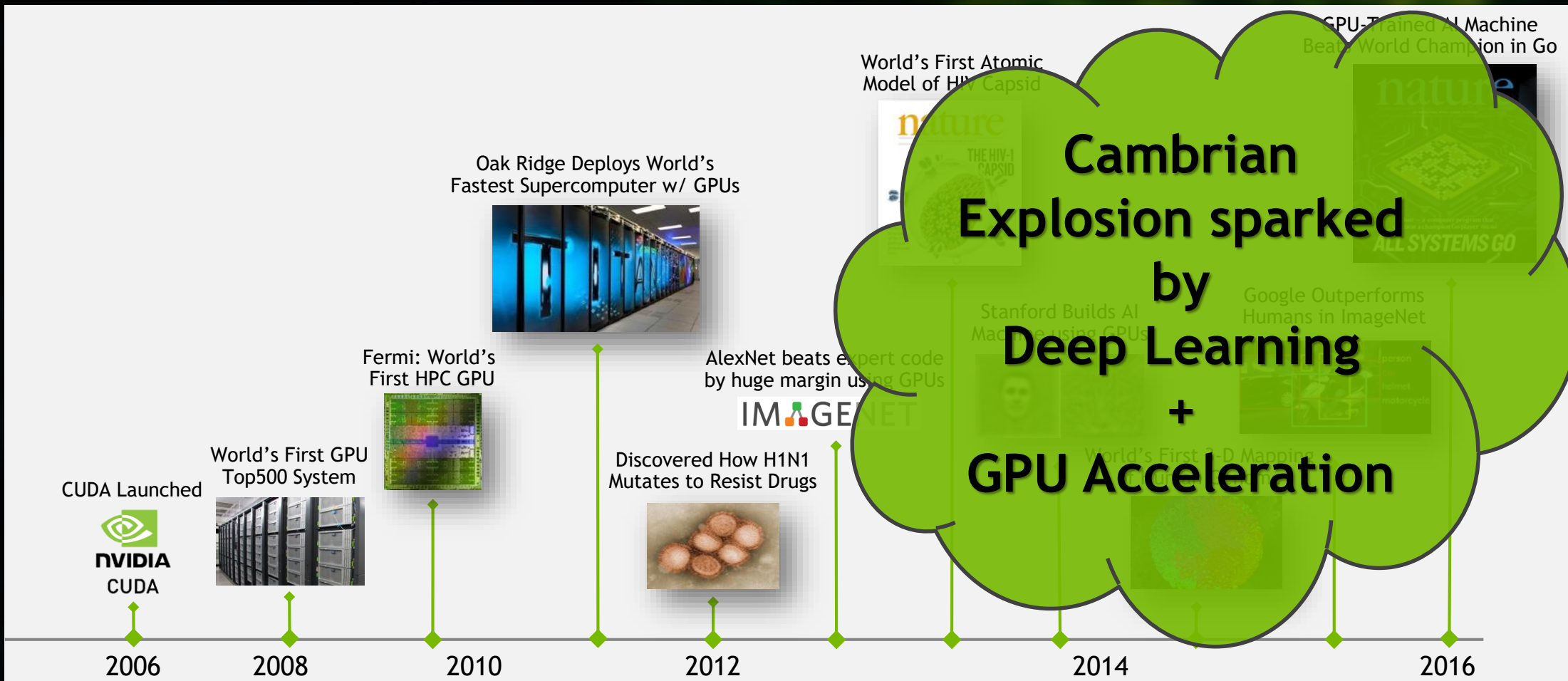
AI & DEEP LEARNING —
THE NEW COMPUTING MODEL

"Software that writes software"

LEARNING
ALGORITHM

"millions of trillions
of FLOPS"

"little girl is eating
piece of cake"

NVIDIA.

THE BIG BANG OF AI AND THE EXPANDING UNIVERSE

TEN YEARS OF GPU COMPUTING PAVED THE WAY TO MODERN AI WITH DEEP LEARNING

GPU-Trained AI Machine Beats World Champion in Go

World's First Atomic Model of HIV Capsid

Oak Ridge Deploys World's Fastest Supercomputer w/ GPUs

Fermi: World's First HPC GPU

AlexNet beats expert code by huge margin using GPUs

IMAGENET

World's First GPU Top500 System

Discovered How H1N1 Mutates to Resist Drugs

CUDA Launched

NVIDIA
CUDA

Cambrian Explosion sparked by Deep Learning + GPU Acceleration

Stanford Builds AI Machine using GPUs

Google Outperforms Humans in ImageNet

World's First 3-D Mapping

2006   2008   2010   2012   2014   2016

THE EXPANDING UNIVERSE OF MODERN AI

INDUSTRY LEADERS

START-UPS

AI-as-a-PLATFORM

CORE TECHNOLOGY / FRAMEWORKS

RESEARCH

" THE BIG BANG "
Big Data
GPU
Algorithms

Research

Berkeley UNIVERSITY OF CALIFORNIA
Carnegie Mellon University
DEEPMIND
Massachusetts Institute of Technology
NYU
OpenAI
Université de Montréal
UNIVERSITY OF OXFORD
UNIVERSITY OF TORONTO

Core Technology / Frameworks

Preferred Networks
Chainer
facebook.
torch
Université de Montréal
theano
Google
TensorFlow
Berkeley UNIVERSITY OF CALIFORNIA
Caffe
Microsoft
CNTK
UNIVERSITY OF OXFORD
MatConvNet
NVIDIA.
cuDNN
dmlc mxnet

AI-as-a-PLATFORM

amazon web services
IBM Watson
Google
Microsoft Azure

START-UPS

api.ai
Personal Assistants
conversational interface

BLUE RIVER TECHNOLOGY
Agriculture
crop-yield optimization

clarifai
Tech
visual recognition platform

deep genomics
Genomics
genetic interpretation

drive.ai
Automotive
computer vision

MetaMind
eCommerce & Medical
recommendation engines

Morpho
Tech
computer vision

Orbital Insight
Geospatial
predictions from images

nervana
Tech
AI-as-a-service

SADAKO
Waste Management
sorting robots

SocialEyes®
Medical
diabetic retinopathy

HOW ARE YOU
Education
teaching robots

3,000+ AI START-UPS
$5B IN FUNDING
Source: Venture Scanner

INDUSTRY LEADERS

Alibaba.com
AstraZeneca
Audi
Baidu 百度
Bloomberg
BMW
charles SCHWAB
CISCO
ebay
FANUC ROBOTICS

Ford
GE
gsk
THE HOME DEPOT
MASSACHUSETTS GENERAL HOSPITAL
Mercedes-Benz
MERCK
Pinterest
Schlumberger

SIEMENS
TARGET
TESLA
TOYOTA
(Twitter)
UBER
VOLVO
Walmart
YAHOO
Yandex
yelp

EMERGING AI SUPERCOMPUTERS

CLOUD-SCALE AI

AI WORKSTATION

AI DATA CENTER

NVIDIA GPU Cloud

DGX Station

DGX-1

DGX-2

with

with

with

Tesla V100 32GB

Tesla V100 32GB
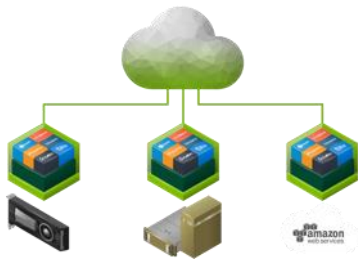
Tesla V100 32GB

Cloud platform with the highest
deep learning efficiency

The Personal
AI Supercomputer

The Essential
Instrument for AI
Research

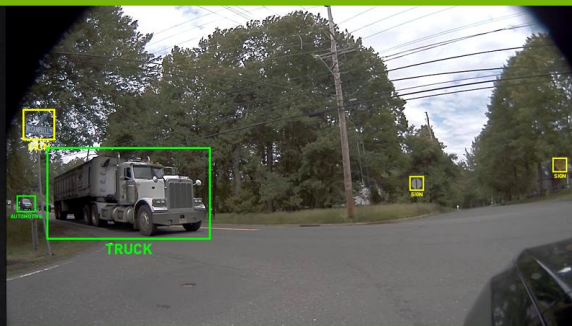The World's Most Powerful
AI System for the Most
Complex AI Challenges

MODERN AI
IN ACTION
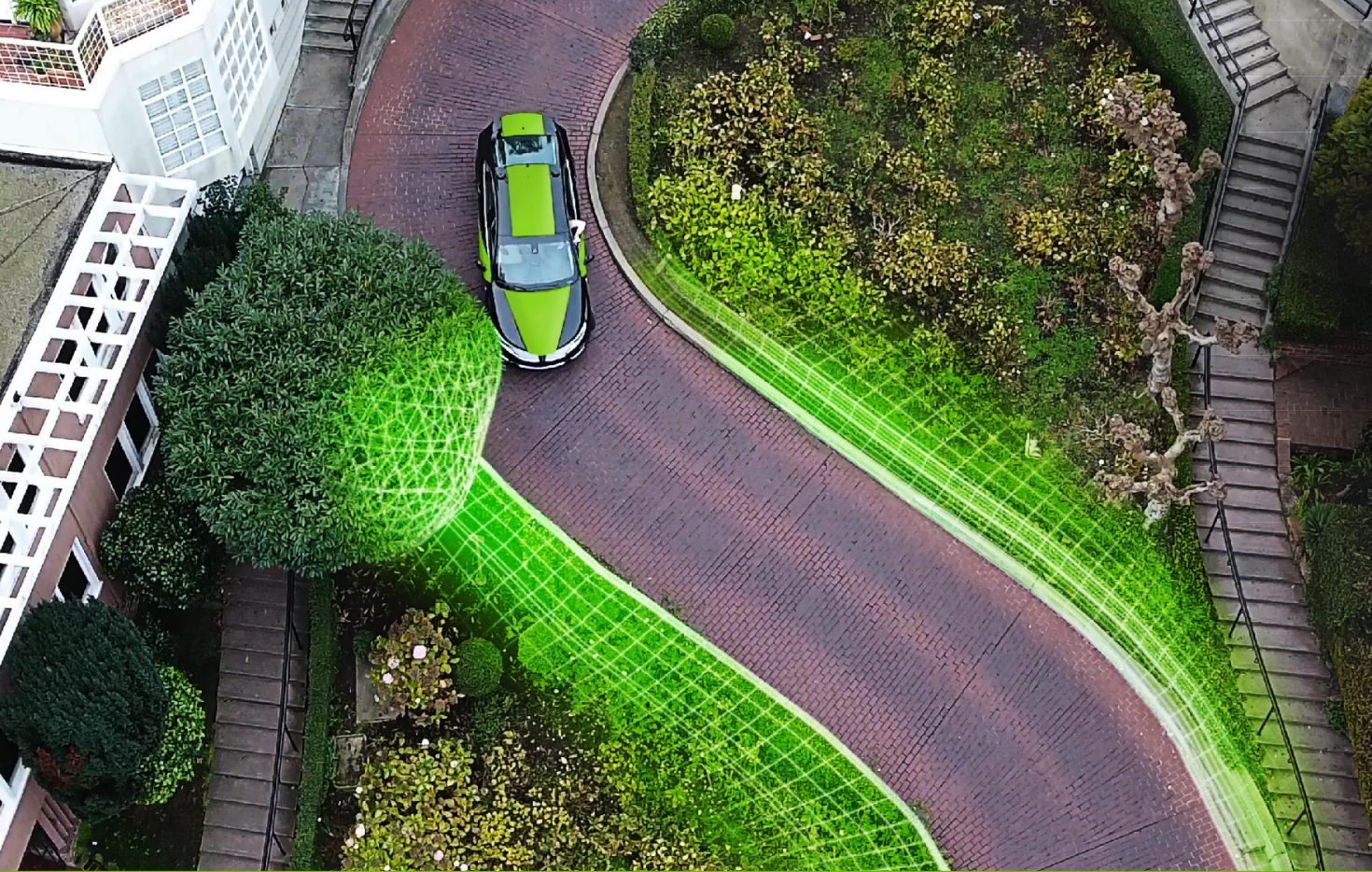
NVIDIA BB8 AI CAR — LEARNING BY EXAMPLE

Driving is a learned behavior that people do as second nature. Yet one that is impossible to program a computer to perform. Using all of the AI capabilities of NVIDIA DRIVE PX 2, our research AI car, BB8, watches humans drive, and has learned to drive in all kinds of conditions — on highways and dirt roads, through obstacle courses, at night, and in the rain. Processing data from multiple cameras, BB8 can even look both ways before safely crossing a busy road on its own.

Left Camera
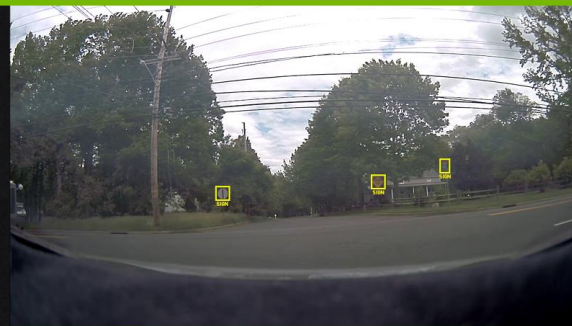
Center Camera

Right Camera

AI FOR SELF-DRIVING WORKFLOW

Get Data

Labeled Data

Train & Test

Trained Model

Adjust

Fine Tune Model

Deploy

Export Model

Test & Validate

Inference at Edge

DNN Development
Exploration
Development
Model Selection

Simulate

Re-Simulate

NVIDIA.

Deep Learning Perception
Distance Detection
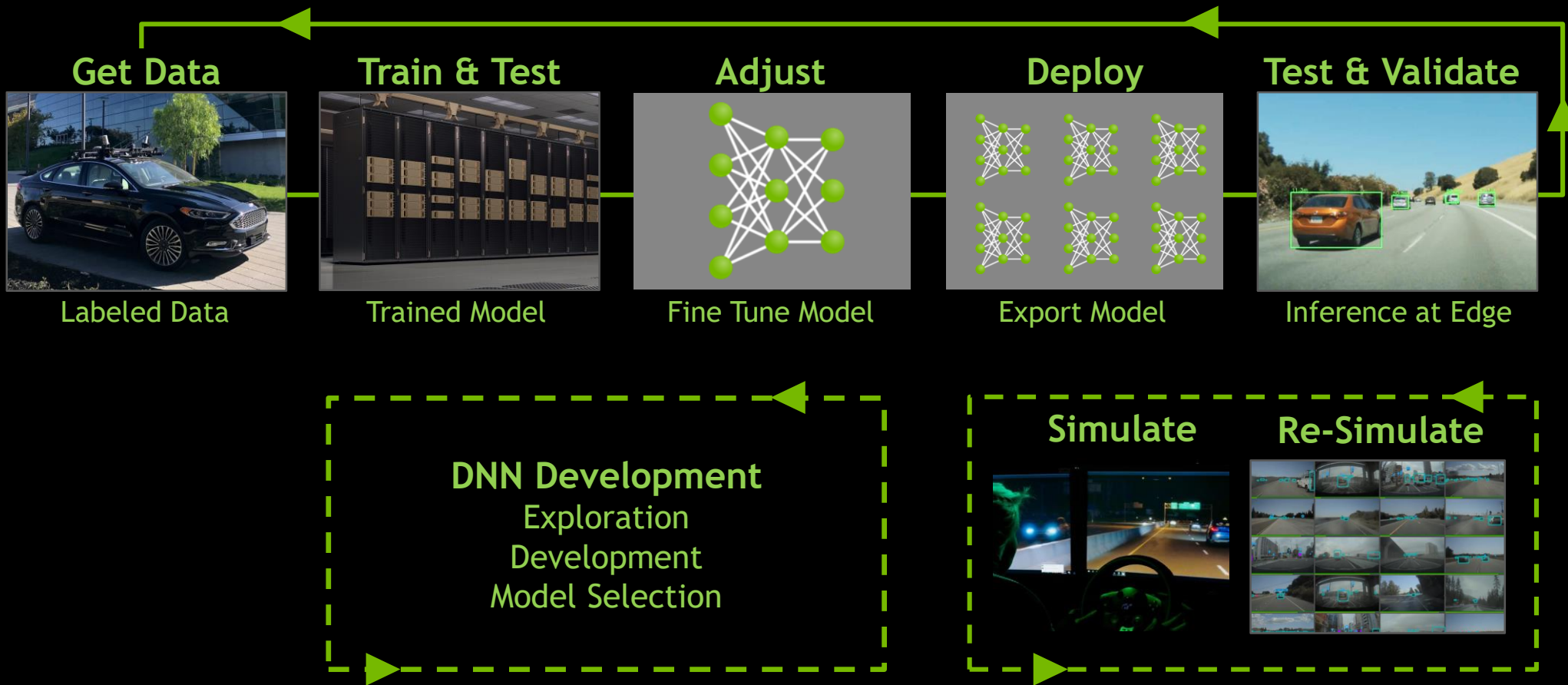
MOVIE: EXAMPLE SELF DRIVING CARS (LEVEL 5)

THE BRAIN OF
AI HEALTHCARE

AI is transforming the spectrum of healthcare, from detection to diagnosis to treatment. GE Healthcare has reinvented the echocardiogram machine by embedding GPU-powered AI in its Vivid E95 system. Mayo Clinic used GPU-powered deep learning to discover that genomic data can be found in MRIs, hidden from traditional analysis methods.

NVIDIA is teaming up with the National Cancer Institute, the U.S. Department of Energy, and several national labs on the "Cancer Moonshot" to deliver a decade of advances in cancer prevention, diagnosis, and treatment in just five years.
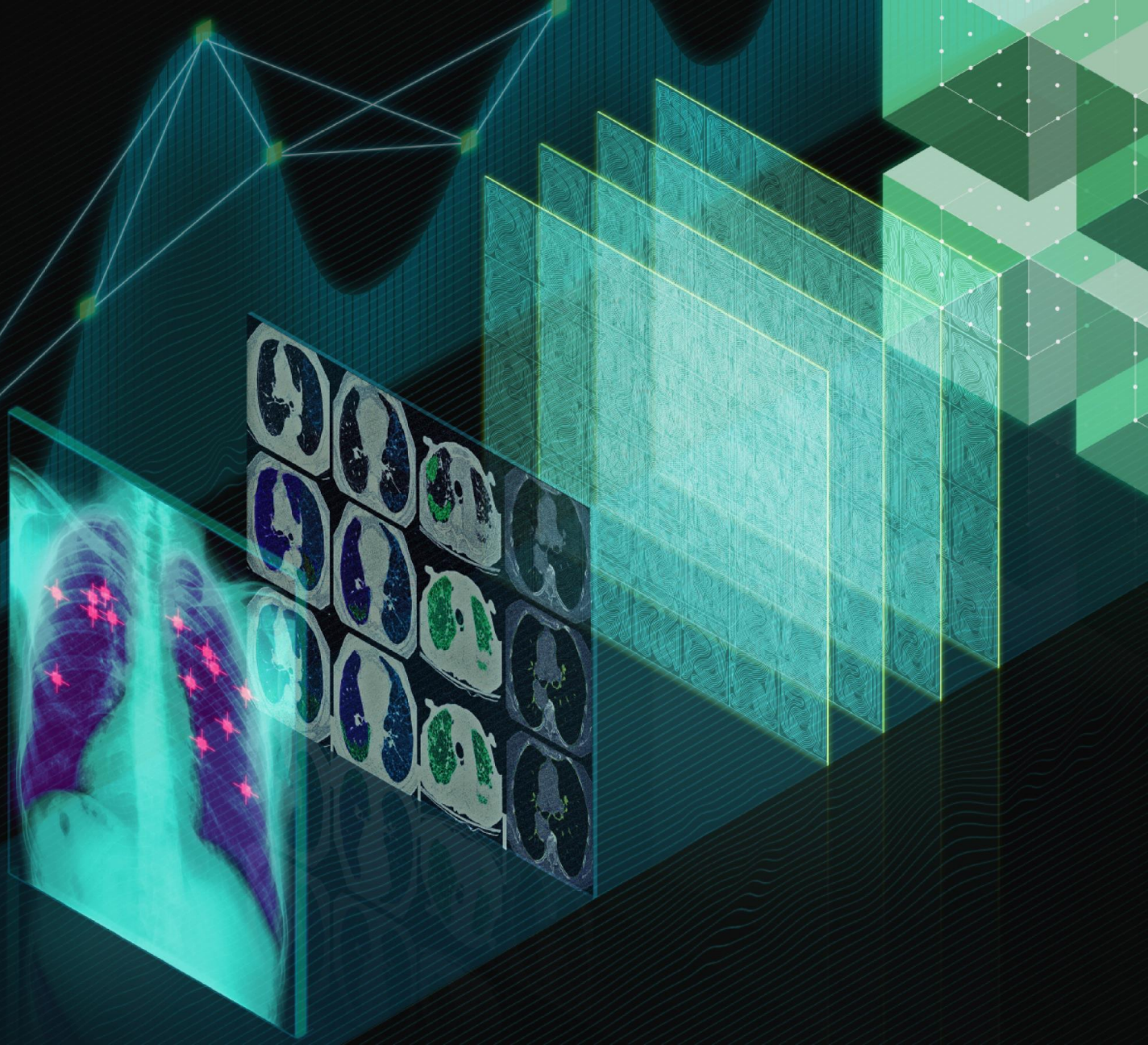
DEEP LEARNING IS ENTERING THE CLINIC

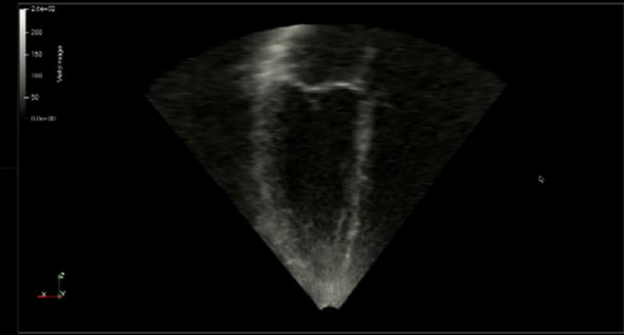

DL DIAGNOSTIC DEVICES

1ST FDA DL CLOUD ALGORITHM ARTERYS
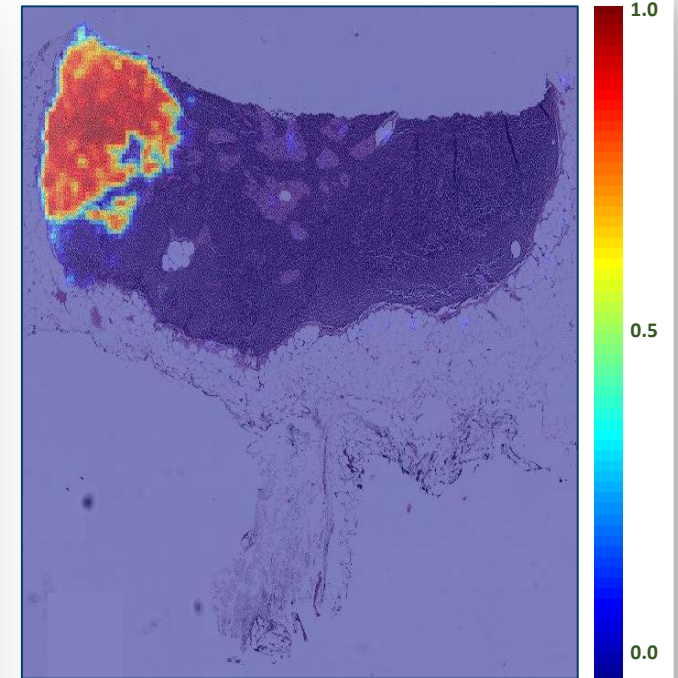
tumor prob. map

DL PATHOLOGY PATHAI
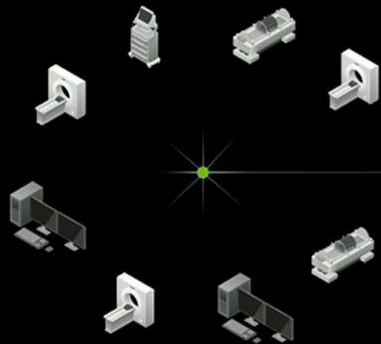
MOVIE: LIVE RECONSTRUCTION OF 3D MODEL OF HEART WITH ULTRA SONIC

THE BRAIN OF SMART MACHINES AND ROBOTS

Deep learning and affordable sensors have created the conditions for a Cambrian explosion of autonomous machines — IoT with AI. NVIDIA Jetson™ TX2, an embedded AI supercomputer, delivers 1 TeraFLOPS of performance in a credit card-sized module. Such power will enable a new wave of automation in manufacturing, drones that can inspect hazardous places, and robots that can deliver the millions of packages shipped every day.
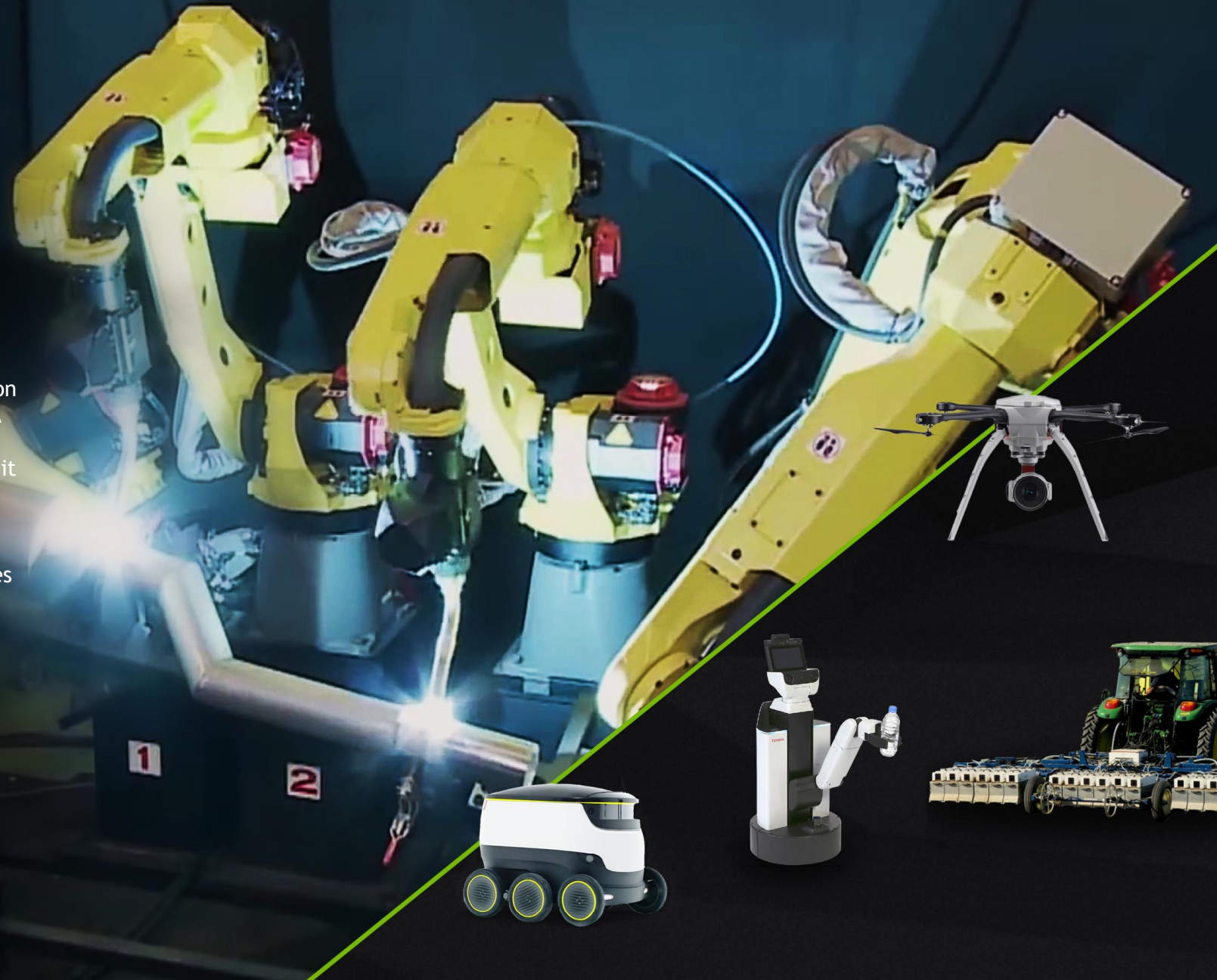
ISAAC — ACCELERATED LEARNING FOR A WORLD OF INTELLIGENT MACHINES

The Isaac robot simulator, an AI-based software platform, lets developers train robots in highly realistic, physics-based virtual environments and then transfer that knowledge to real-world units. Developers can set up extensive test scenarios using deep learning training, and then simulate them in minutes instead of months.
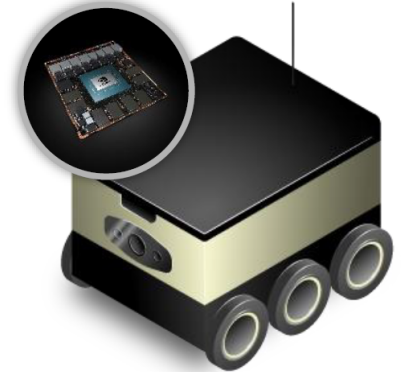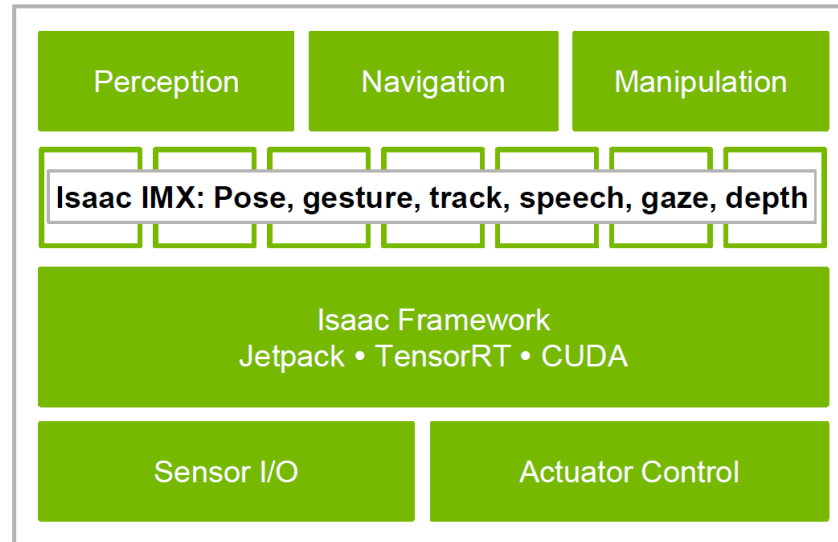
ISAAC ROBOTICS PLATFORM
Cognitive Smart Machines, Robots, Drones
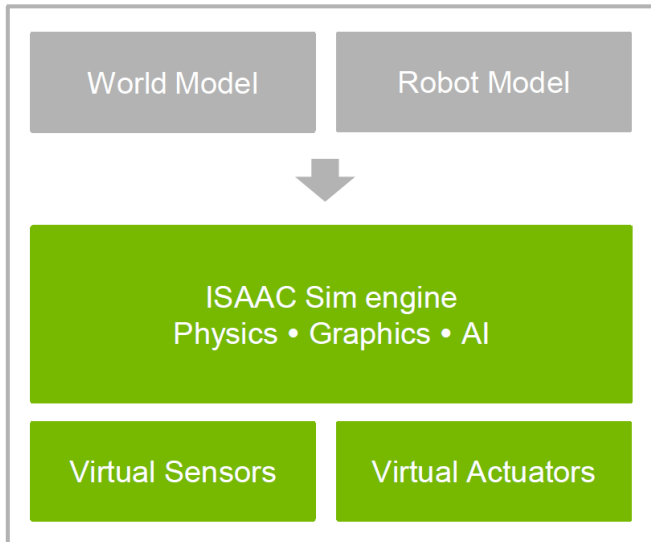
MOVIE: CONVERGING ARTITIFIAL AND REAL WORLD

WHAT CAN I DO?

UNDERSTANDING MODERN AI WITH DEEP LEARNING

GET THE BASICS SKILLS

Watch "Deep Learning Demystified"

Listen to the NVIDIA AI Podcast

Review examples of AI in action

BUILD & TRAIN DATA SCIENCE TEAM WITH DEEP LEARNING KNOWLEDGE

Take a self-paced course online at
www.nvidia.com/dlilabs

View upcoming events or request a
workshop at www.nvidia.com/dli

JOIN OUR COMMUNITY

Sign up for the NVIDIA Developer Program
at https://developer.nvidia.com/join

Discover GPU accelerated containers

Start researching using an AI appliance
with NVIDIA DGX Systems

Visit GTC in Munich October 9-11, 2018

GPU TECHNOLOGY CONFERENCE — EUROPE / OCTOBER 9-11, 2018 — JOIN EUROPE'S PREMIER CONFERENCE ON AI.

Register with 25% discount code «NVMSTAMPFLI»
https://www.gputechconf.eu/

Get in contact with us: mstampfli@nvidia.com

nVIDIA.
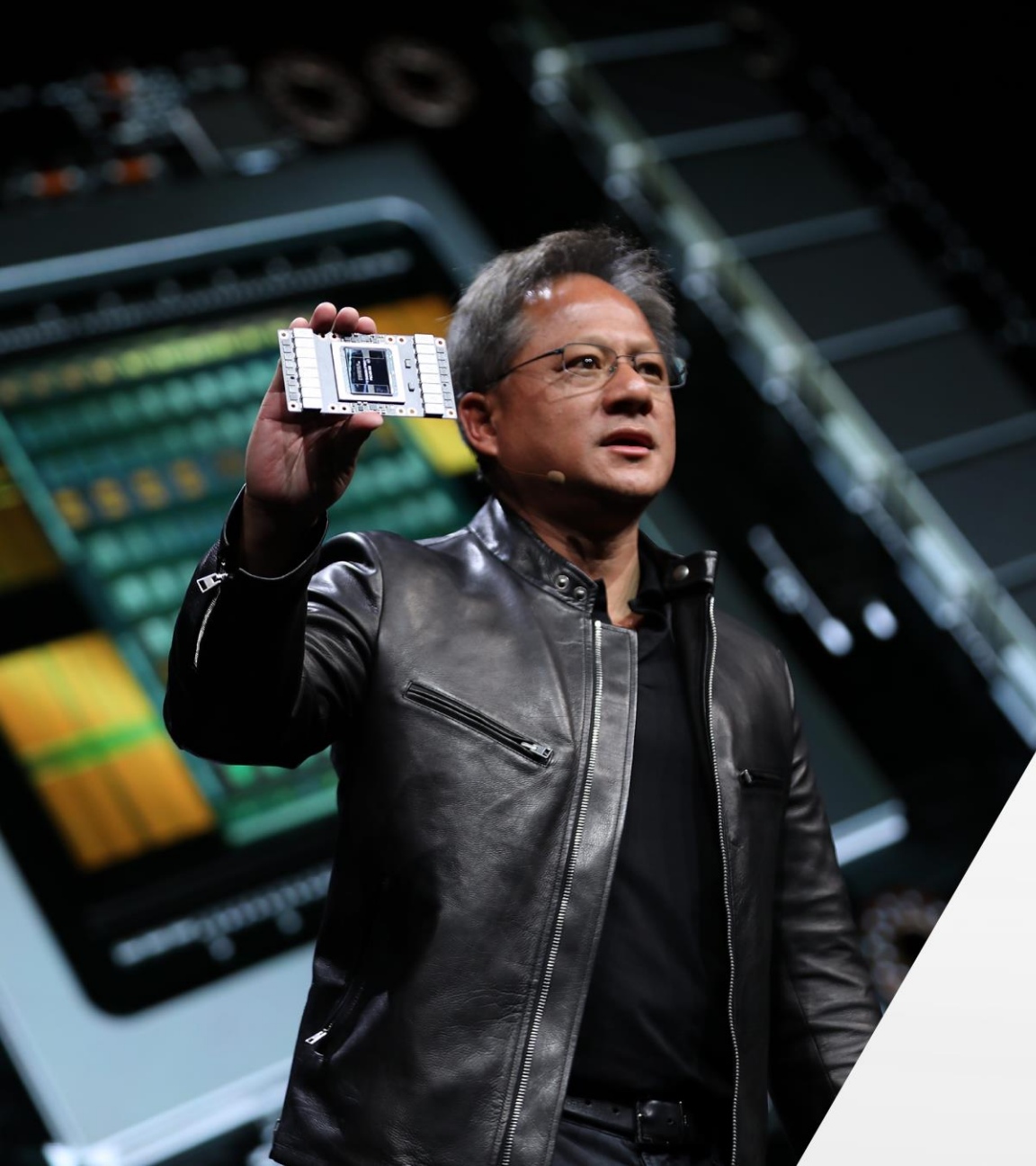
> Founded in 1993

> Jensen Huang, Founder & CEO

> 12'000+ employees

> $9.7B in FY18

"World's Best Performing CEOs"
— Harvard Business Review

"World's Most Admired Companies"
— Fortune

"World's Best CEOs"
— Barron's

"Most Innovative Companies"
— Fast Company

"Employees' Choice: Highest Rated CEOs"
— Glassdoor

"50 Smartest Companies"
— MIT Tech Review