




NVIDIA



Artificial intelligence, the dream of computer scientists for over half a century, is no longer science fiction. And in the next few years, it will transform every industry.

Soon, self-driving cars will reduce congestion and improve road safety. AI travel agents will know your preferences and arrange every detail of your family vacation. And medical instruments will read and understand patient DNA to detect and treat early signs of cancer.

Where engines made us stronger and powered the first industrial revolution, AI will make us smarter and power the next.

What will make this intelligent industrial revolution possible? A new computing model — GPU deep learning — that enables computers to learn from data and write software that is too complex for people to code.



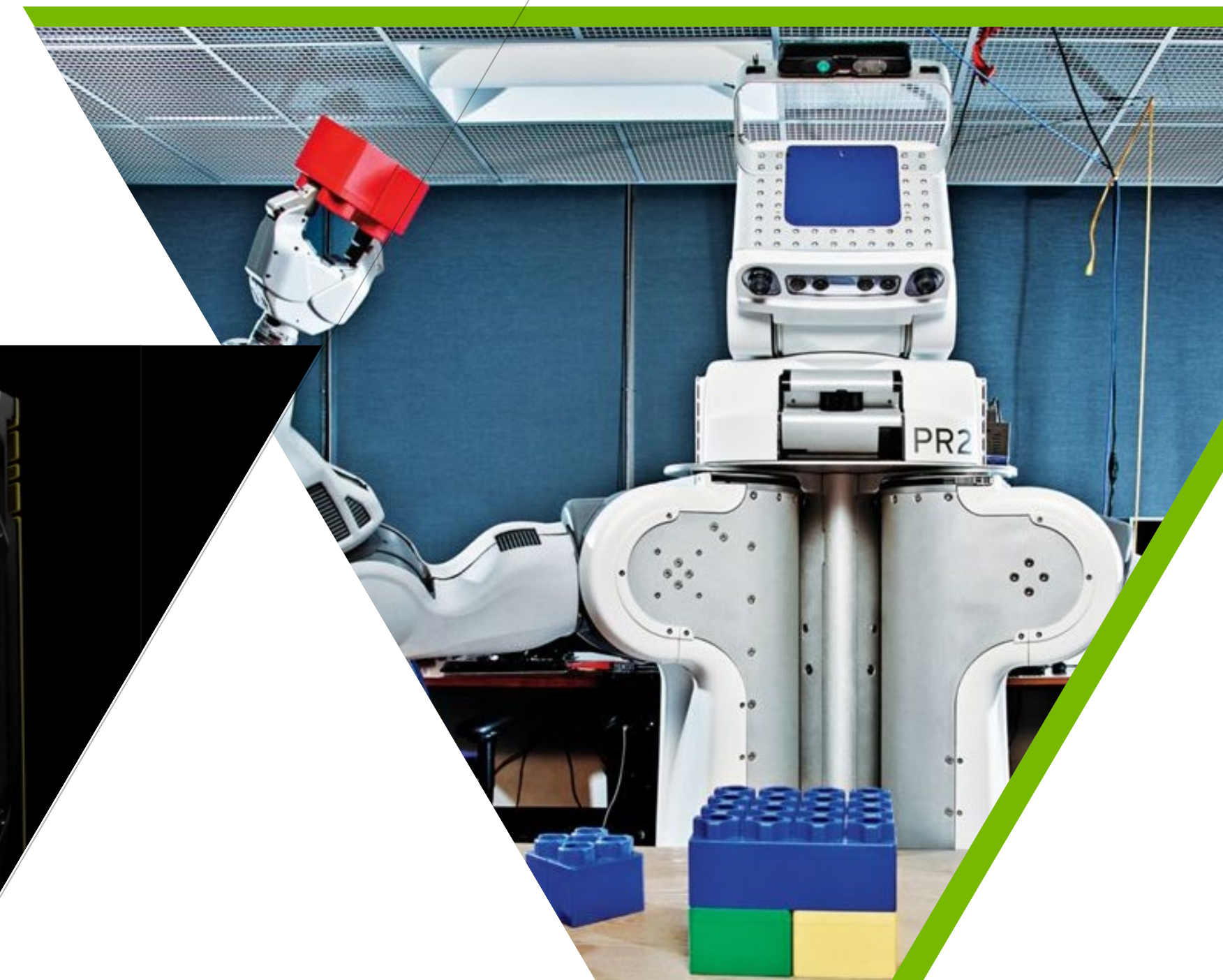
NVIDIA — INVENTOR OF THE GPU

The GPU has proven to be unbelievably effective at solving some of the most complex problems in computer science.

It started out as an engine for simulating human imagination, conjuring up the amazing virtual worlds of video games and Hollywood films.

Today, NVIDIA's GPU simulates human intelligence, running deep learning algorithms and acting as the brain of computers, robots, and self-driving cars that can perceive and understand the world. NVIDIA is increasingly known as "the AI computing company."

This is our life's work — to amplify human imagination and intelligence.





THE NVIDIA GPU DEFINES MODERN COMPUTER GRAPHICS

Our invention of the GPU in 1999 made possible real-time programmable shading, which gives artists an infinite palette for expression. We've led the field of visual computing since.



SIMULATING HUMAN IMAGINATION

Digital artists, industrial designers, filmmakers, and broadcasters rely on NVIDIA Quadro® pro graphics to bring their imaginations to life. NVIDIA GPUs power 90% of the world's professional graphics workstations. And for eight years running, every film nominated for the Academy Award for Best Visual Effects used NVIDIA technology to make their vision a reality.





CREATING AMAZING WORLDS

Today's gaming industry is fueled by a steady stream of blockbuster titles with Hollywood-level production values. NVIDIA GPUs are the engines that make these games possible. And our GameWorks™ software delivers algorithms and tools to game developers big and small. From realistic smoke and water to hair and fur, NVIDIA tech makes games more graphics-rich and immersive.



GEFORCE GAMING PC — THE WORLD'S LARGEST GAMING PLATFORM

At \$100 billion, computer gaming is the world's largest entertainment industry. And with 100 million gamers, NVIDIA GeForce is its largest platform. GeForce® GTX GPUs and the GeForce Experience™ application transform everyday PCs into powerful gaming machines.



CREATING VIRTUAL REALITY

NVIDIA GPUs power the intensive visual experience of VR — a computing platform that is redefining gaming, film, news, communications, education, and many other industries. NVIDIA VRWorks™ software helps headset makers and game developers create amazing VR devices and games.

SHIELD — THE FUTURE OF HOME ENTERTAINMENT

Apps are the future of TV, and streaming is the future of gaming. NVIDIA SHIELD™ is our foothold in this future. With 4K streaming and advanced gaming, it's the best Android TV box on the market. And it keeps getting better with our regular software updates and a growing app catalog.





THE NVIDIA GPU REVOLUTIONIZES COMPUTING

In 2006, the creation of our CUDA programming model and Tesla GPU platform opened up the parallel-processing capabilities of the GPU to general-purpose computing.

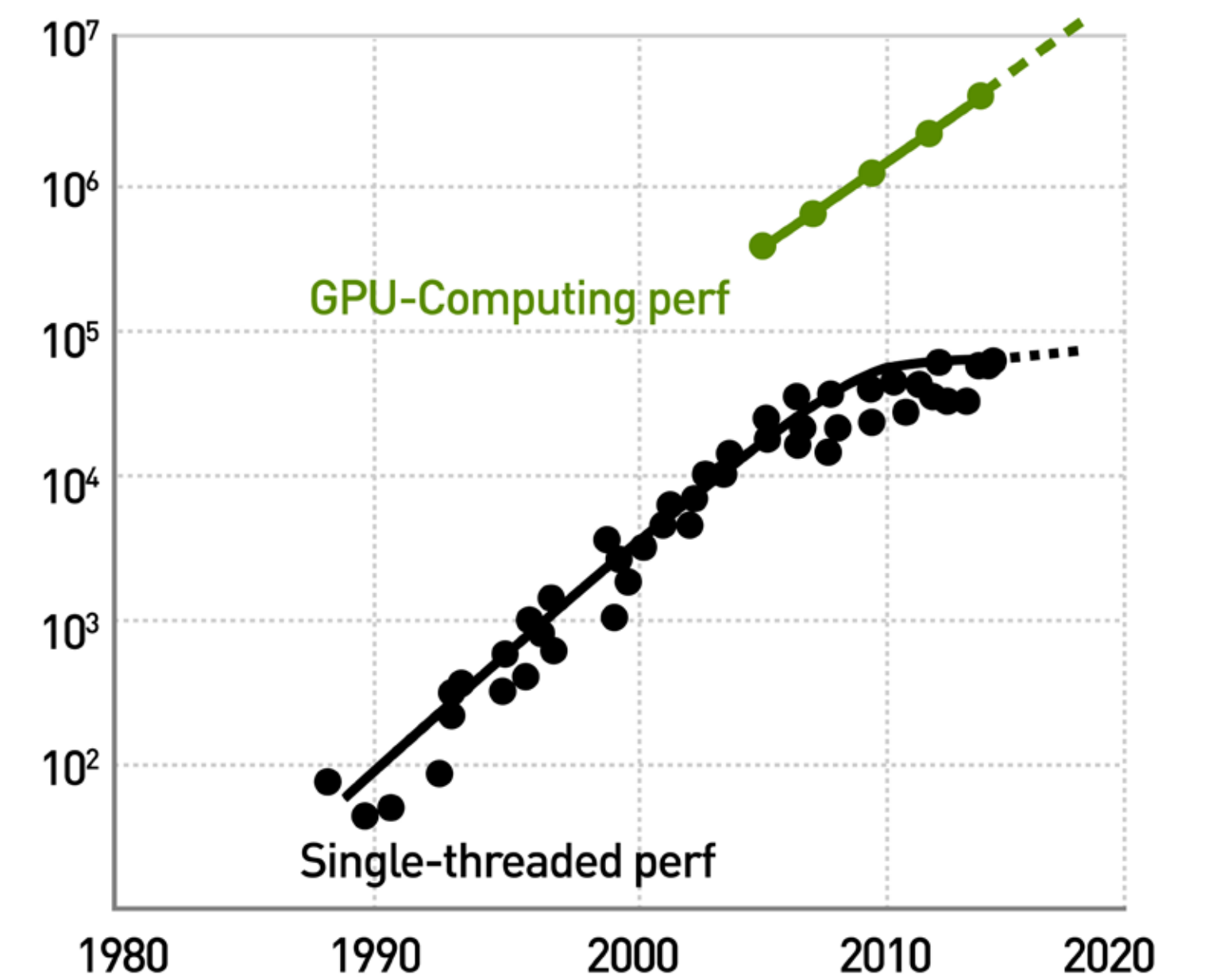
A powerful new approach to computing was born.



POWERING THE WORLD'S FASTEST SUPERCOMPUTERS

For traditional computing approaches, Moore's Law is coming to an end. But exponential advances in compute speed and capacity are just beginning. Today, GPU computing is the most pervasive, accessible, energy-efficient path forward for HPC and data centers, and powers the fastest supercomputers in the U.S. and Europe.

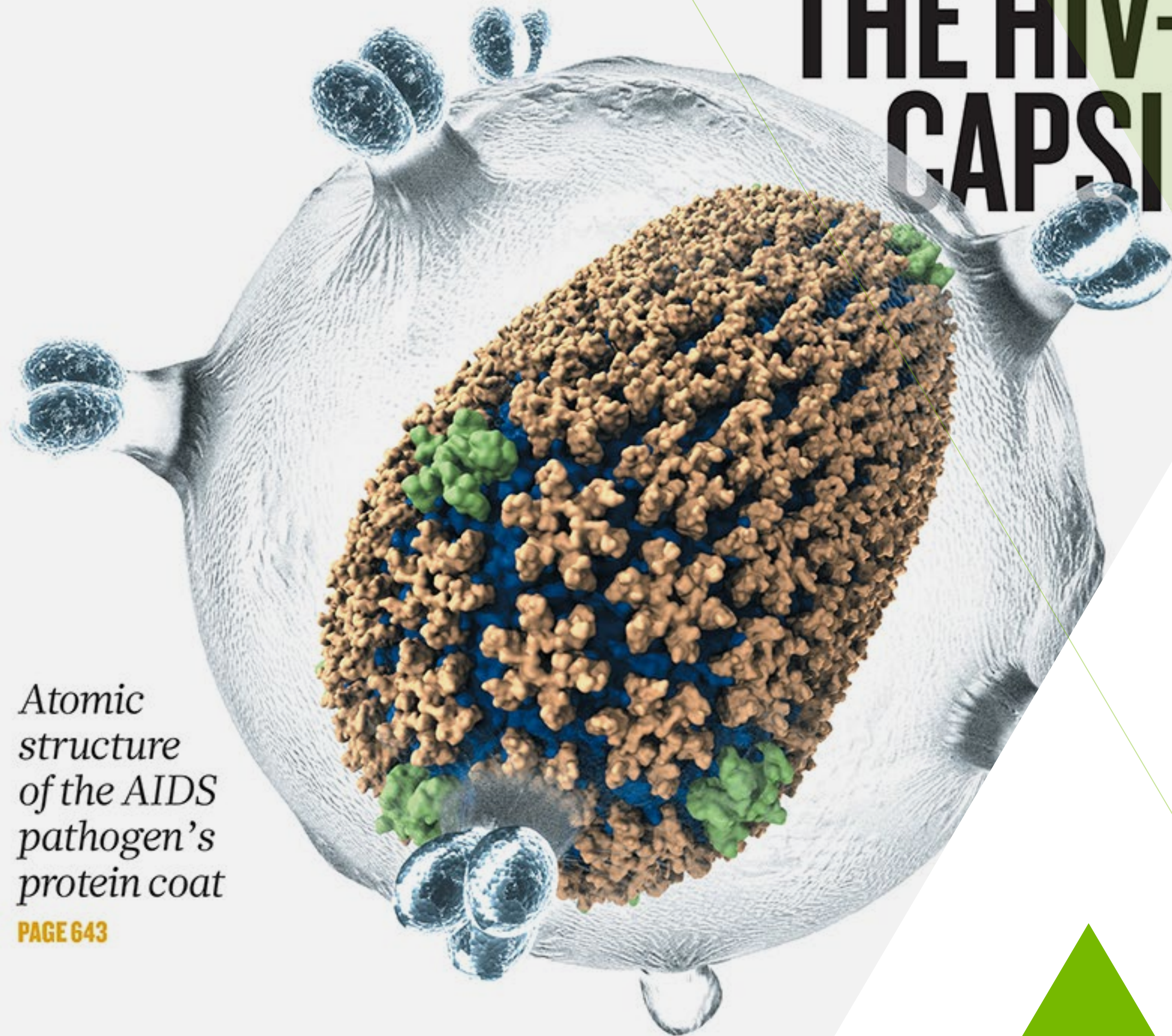
Processor Trends



nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

THE HIV-1 CAPSID



*Atomic
structure
of the AIDS
pathogen's
protein coat*

PAGE 643



POWERING LIFE-CHANGING SCIENTIFIC DISCOVERIES

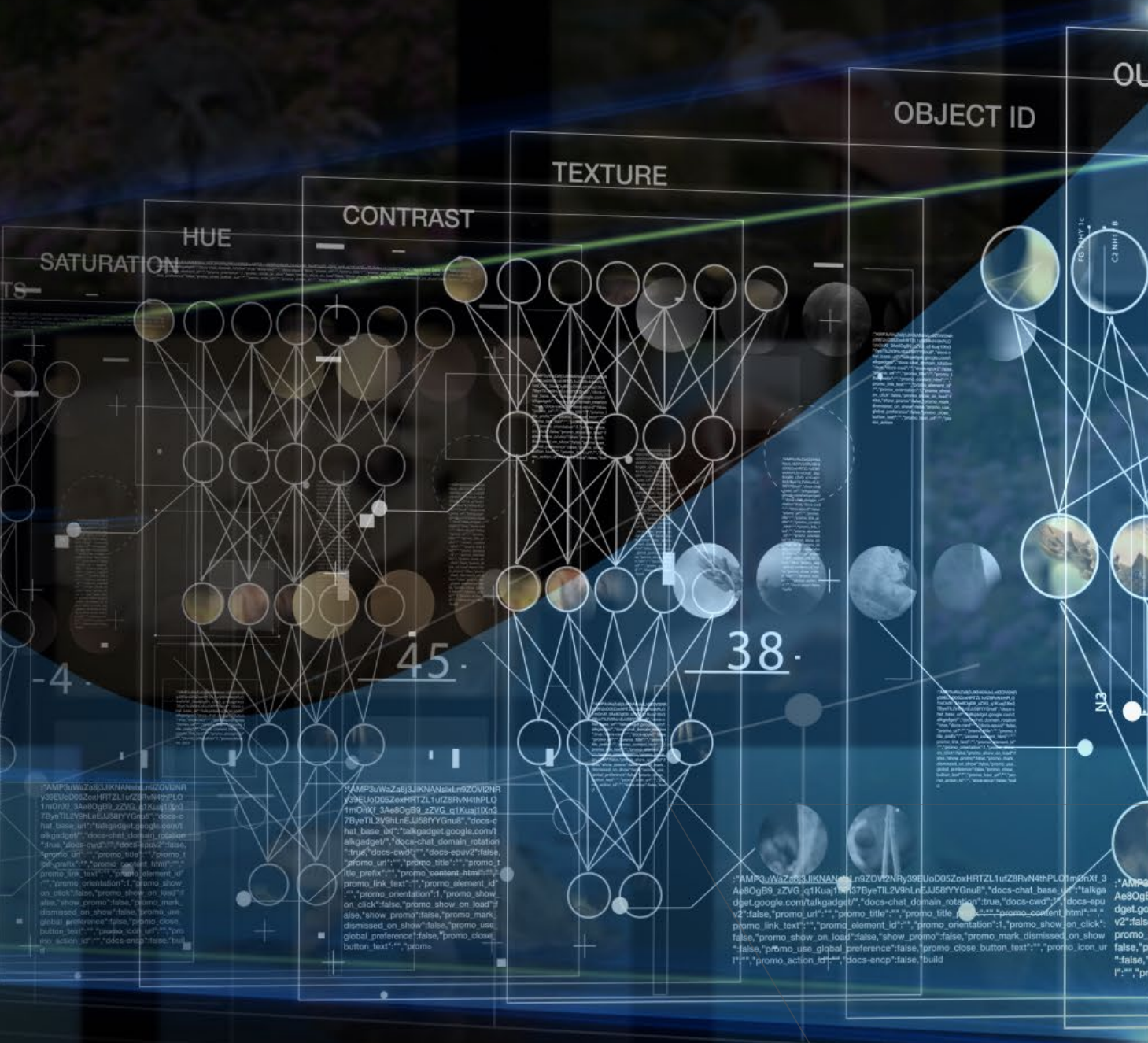
The GPU computing platform is essential to those who take on the challenges that matter most. Using a GPU-powered supercomputer, University of Illinois scientists achieved a breakthrough in HIV research by performing the first all-atom simulation of the capsid. GPUs also drive the GE Revolution CT scanner, which can produce high-quality imagery while reducing patient radiation exposure by up to 82%.

GPU DEEP LEARNING IGNITES THE BIG BANG OF AI

Artificial intelligence is the use of computers to simulate human intelligence.

AI amplifies our cognitive abilities — letting us solve problems where the complexity is too great, the information is incomplete, or the details are too subtle and require expert training.

Learning from data — a computer's version of life experience — is how AI evolves. GPU deep learning is a new computing model in which deep neural networks are trained to recognize patterns from massive amounts of data. This new model has sparked the AI computing era.

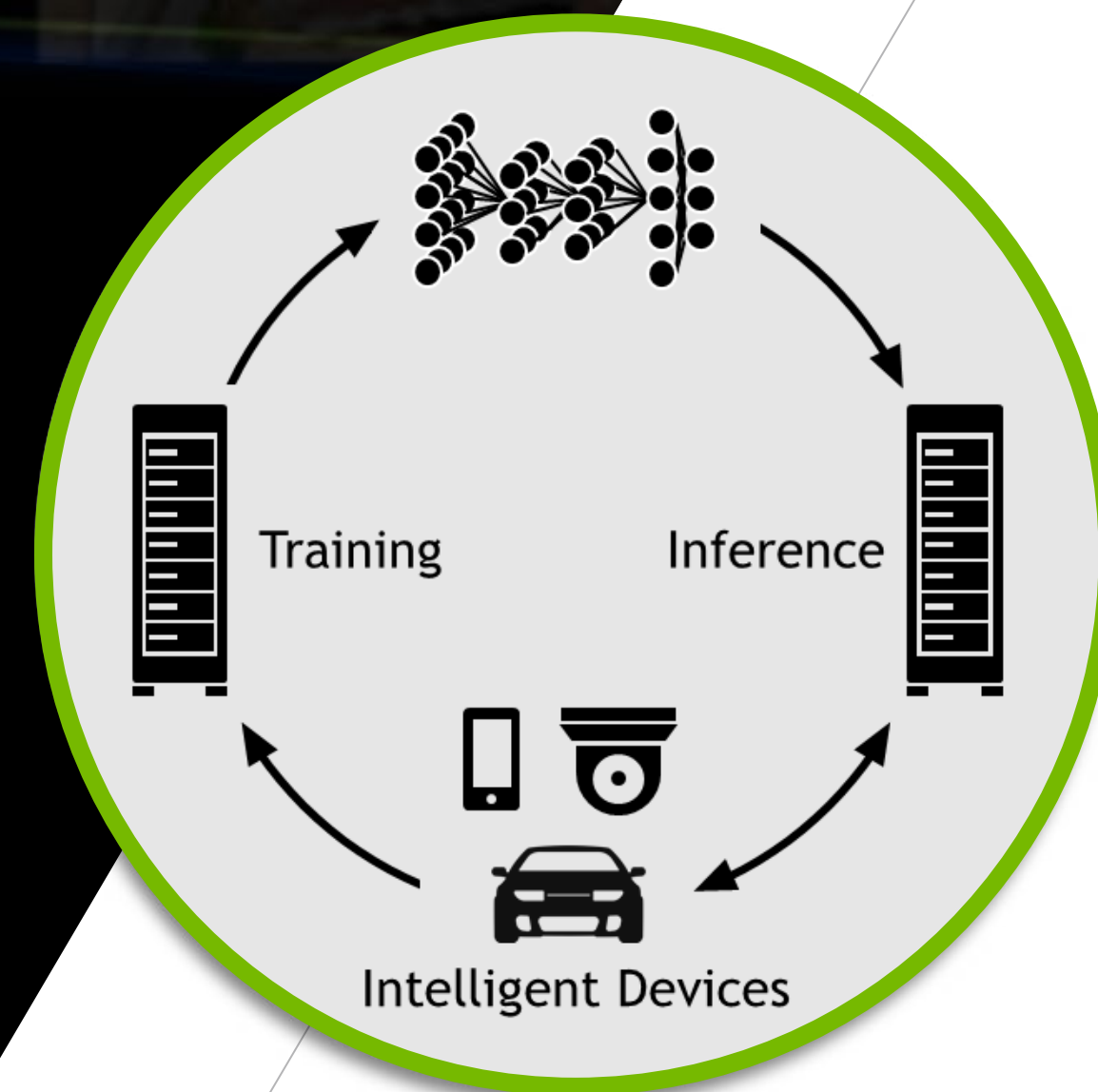


ELEPHANT IN GRASS



GPU DEEP LEARNING — A NEW COMPUTING MODEL

GPU deep learning is changing the way software is developed and how it runs. In the past, software engineers crafted programs and meticulously coded algorithms. Now, algorithms learn from massive amounts of real-world examples and software writes itself. Deep neural networks are deployed in data centers and intelligent devices to infer and predict the next action. This new computing model is made possible by deep neural networks trained and deployed on GPUs.



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet ILSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.



Deep learning with COTS HPC systems

Adam Coates
Brody Huval
Tao Wang
David J. Wu
Andrew Y. Ng

Stanford University Computer Science Dept., 353 Serra Mall, Stanford, CA 94305 USA

Bryan Catanzaro

NVIDIA Corporation, 2701 San Tomas Expressway, Santa Clara, CA 95050

ACOATES@CS.STANFORD.EDU
BRODYH@STANFORD.EDU
TWANGCAT@STANFORD.EDU
DWU4@CS.STANFORD.EDU
ANG@CS.STANFORD.EDU

BCATANZARO@NVIDIA.COM

Abstract

Scaling up deep learning algorithms has been shown to lead to increased performance in benchmark tasks and to enable discovery of complex high-level features. Recent efforts to train extremely large networks (with over 1 billion parameters) have relied on cloud-like computing infrastructure and thousands of CPU cores. In this paper, we present technical details and results from our own system based on Commodity Off-The-Shelf High

to detect objects when trained from unlabeled images alone (Coates et al., 2012; Le et al., 2012). The very largest of these systems has been constructed by Le et al. (Le et al., 2012) and Dean et al. (Dean et al., 2012), which is able to train neural networks with over 1 billion trainable parameters. While such extremely large networks are potentially valuable objects of AI research, the expense to train them is overwhelming: the distributed computing infrastructure (known as "DistBelief") used for the experiments in (Le et al., 2012) manages to train a neural network using 16000 CPU cores (in 1000 machines) in just a few days, yet

AI RESEARCHERS DISCOVERED NVIDIA GPUs

By 2010, AI researchers around the world were tapping into the parallel processing capabilities of NVIDIA GPUs to train neural networks. In 2012, Alex Krizhevsky of the University of Toronto won the ImageNet image recognition competition using a deep neural network trained on NVIDIA GPUs — beating all the human expert algorithms that had been honed for decades. That same year, recognizing that larger networks can learn more, Stanford's Andrew Ng and NVIDIA Research teamed up to develop a method for training networks using large-scale GPU computing systems. These seminal papers sparked the big bang of modern AI.

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

At last — a computer program that can beat a champion Go player **PAGE 484**

ALL SYSTEMS GO

CONSERVATION

SONGBIRDS
À LA CARTE

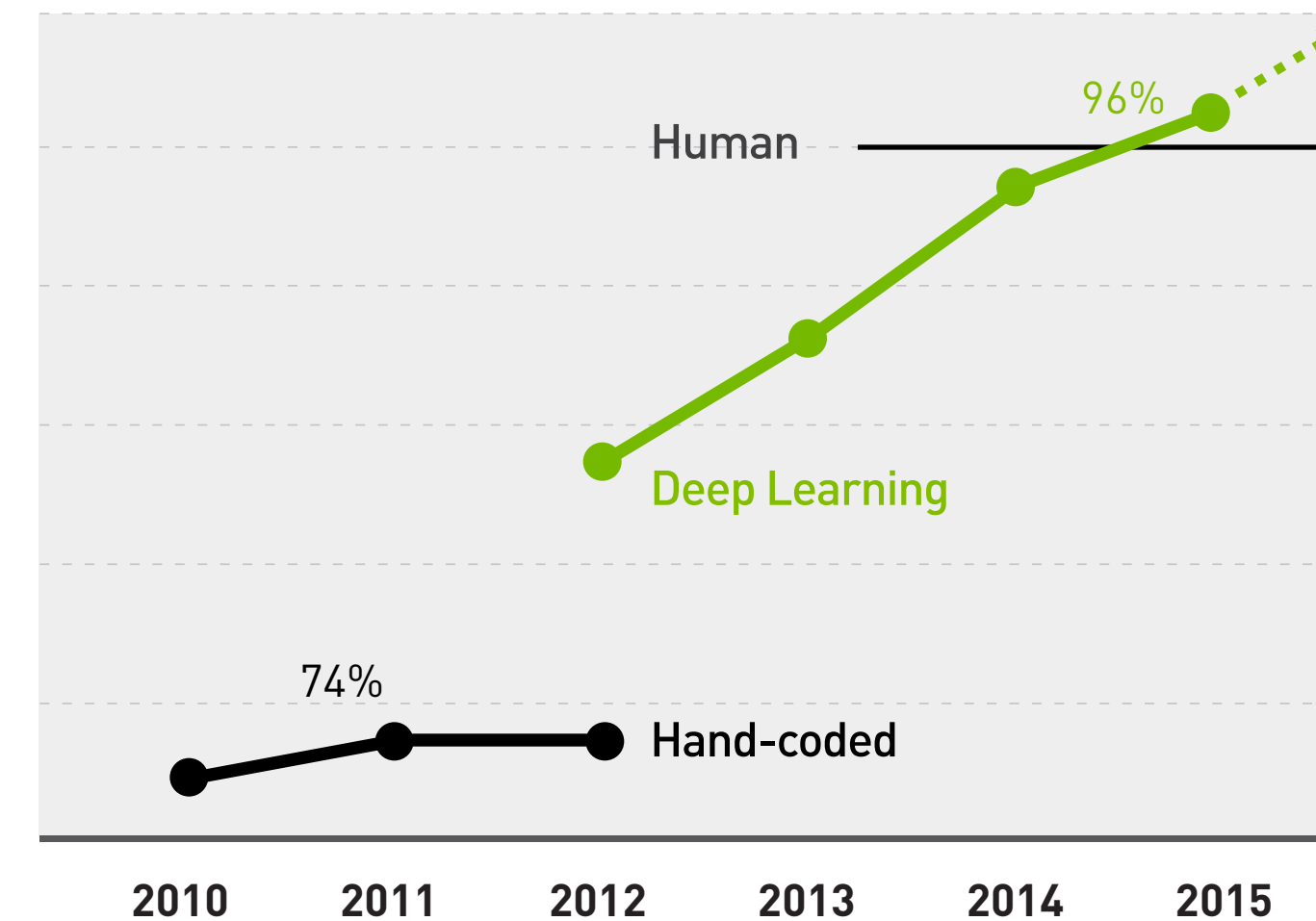
RESEARCH ETHICS

SAFEGUARD
TRANSPARENCY

POPULAR SCIENCE

WHEN GENES
GOT 'SELFISH'

ImageNet — Accuracy Rate



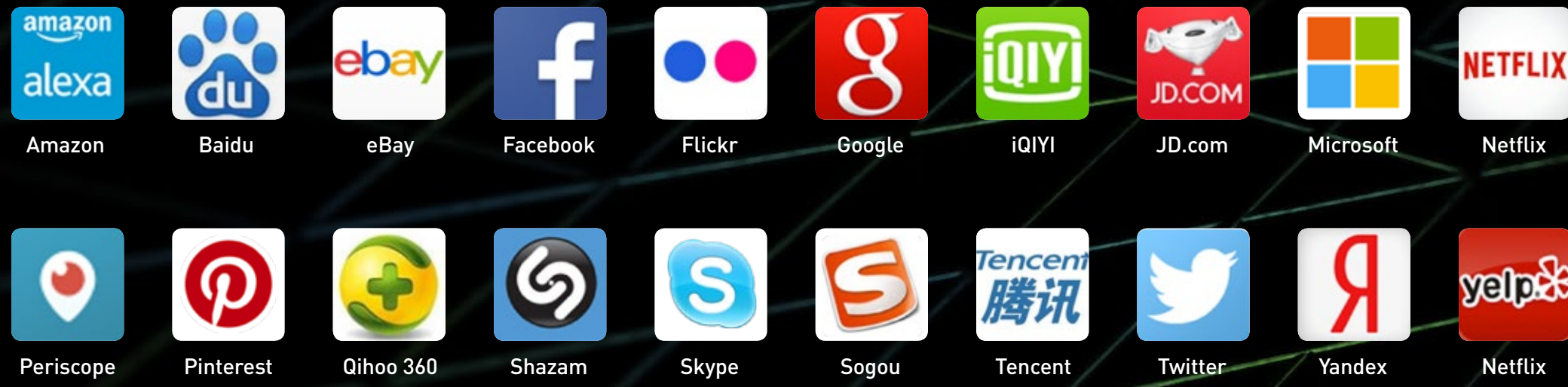
AI ACHIEVES “SUPERHUMAN” RESULTS

The big bang of modern AI set off a string of “superhuman” achievements. In 2015, Google and Microsoft both beat the best human score in the ImageNet challenge. In 2016, DeepMind’s AlphaGo recorded its historic win over Go champion Lee Sedol and Microsoft achieved human parity in speech recognition.

A close-up, angled view of the NVIDIA DGX-1 AI supercomputer. The device is a rectangular, metallic-looking box with a textured, perforated surface. A prominent feature is a large, vertical, brushed metal handle on the left side. To the right of the handle, the NVIDIA logo (a green eye) and the word "NVIDIA" are printed vertically. The device is set against a dark background with a faint, glowing green grid pattern. A green triangle is visible in the top-left corner, and a grey triangle is in the top-right corner. The device is reflected on a dark surface below it.

THE WORLD'S FIRST AI SUPERCOMPUTER IN A BOX

While deep learning holds enormous promise, it requires a massive amount of computing power. To arm data scientists in every organization, we created an AI supercomputer in a box, NVIDIA DGX-1™. This plug-and-play appliance delivers the computing power of a 250-node HPC cluster, reducing network training time from weeks to days.



THE BRAIN OF THE AI ENTERPRISE

Every industry has awoken to AI. The world's leading internet companies are racing to infuse intelligence into every app. More than 1,500 AI startups have cropped up around the world. We're working with the world's largest enterprise technology providers so every company can tap into the power of GPU deep learning. Alibaba, Amazon, Google, IBM and Microsoft offer NVIDIA GPUs in the cloud. IBM's "Minsky" POWER8 and NVIDIA Tesla P100 server is purpose-built for AI. SAP is using NVIDIA DGX-1 to build machine learning enterprise solutions for its 320,000 customers. And the GPU-optimized Microsoft Cognitive Toolkit gives enterprises an AI platform that can span from their data center to Microsoft's Azure cloud.



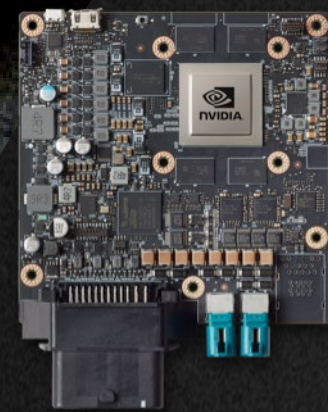
Microsoft Azure

The Microsoft Cognitive Toolkit

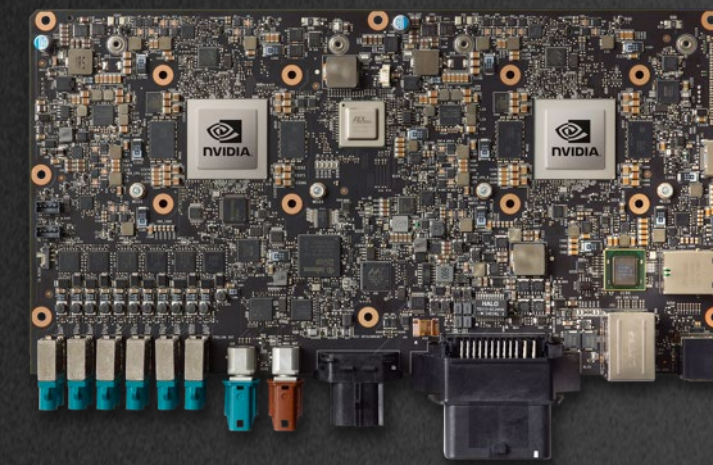


THE BRAIN OF AI CARS

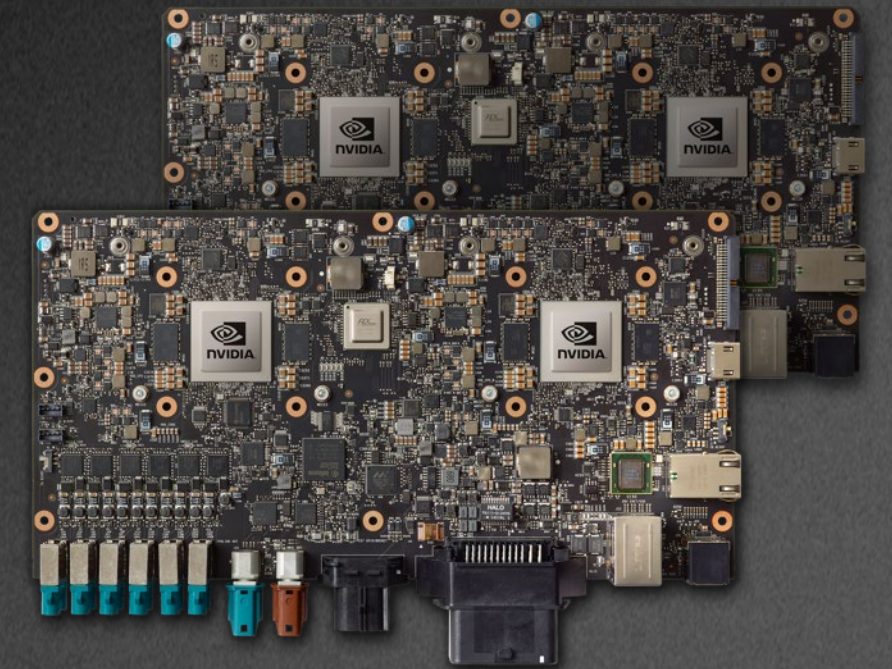
Autonomous vehicles will transform the \$10 trillion transportation industry — reducing accidents, improving commercial vehicle productivity, and enabling new mobility services. NVIDIA DRIVE™ PX 2 is a scalable architecture that spans the entire range of AI for autonomous driving. More than 80 companies are currently developing with it, and every Tesla Motors vehicle now comes equipped with DRIVE PX 2 for full self-driving capabilities.



AUTOCRUISE



AUTOCHAUFFEUR



FULL AUTONOMY

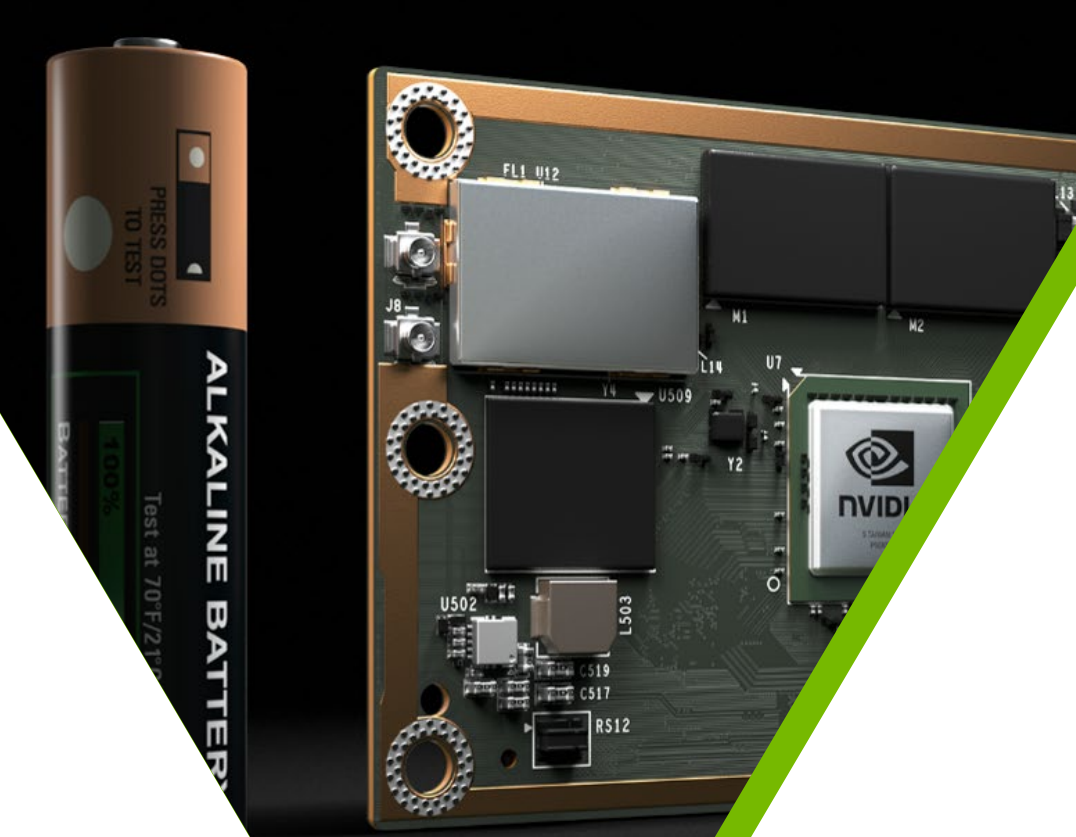


NVIDIA BB8 AI CAR — LEARNING BY EXAMPLE

Driving is a learned behavior that we 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.

THE BRAIN OF INTELLIGENT MACHINES & IoT

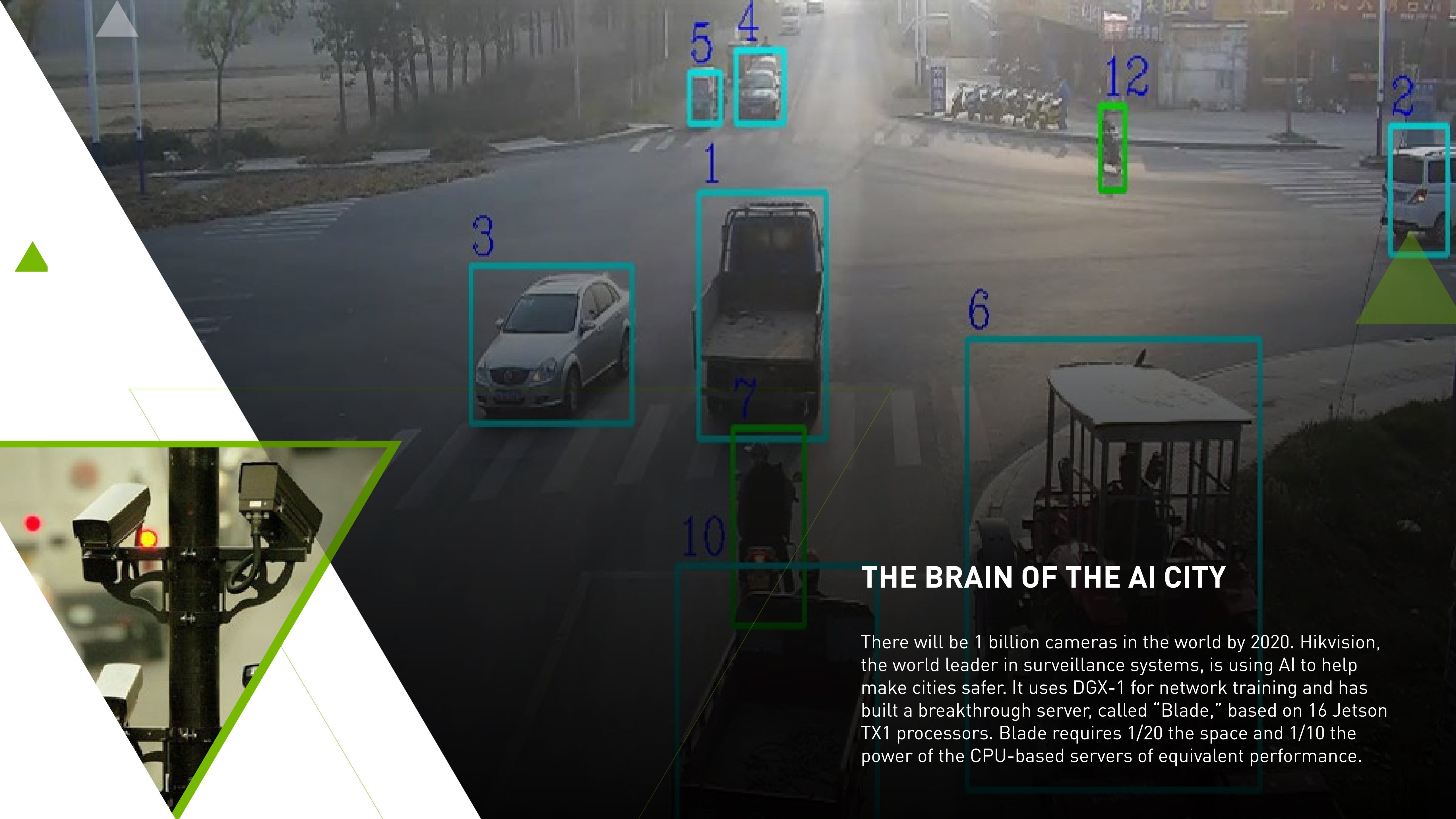
Billions of intelligent devices will one day take advantage of AI to perform seemingly intelligent tasks. NVIDIA Jetson™ TX1, an embedded AI supercomputer, delivers 1 TeraFLOP of performance in a credit card-sized module. Such power will enable search-and-rescue drones that can navigate through dangerous situations, social devices that can connect emotionally with humans, and robots that can learn through trial and error.





THE BRAIN OF THE AI FACTORY

There are 2 billion industrial robots worldwide, and Japan is the center of robotics innovation. FANUC, the Japanese industrial robotics giant, is building the factory of the future on the NVIDIA AI platform, from end to end. It will train its deep neural networks with NVIDIA GPUs, GPU-powered FANUC units will drive a group of robots and allow them to learn together, and each robot will have an embedded GPU to perform real-time AI.



THE BRAIN OF THE AI CITY

There will be 1 billion cameras in the world by 2020. Hikvision, the world leader in surveillance systems, is using AI to help make cities safer. It uses DGX-1 for network training and has built a breakthrough server, called "Blade," based on 16 Jetson TX1 processors. Blade requires 1/20 the space and 1/10 the power of the CPU-based servers of equivalent performance.

CANCER MOONSHOT

AN AI PLATFORM TO ACCELERATE CANCER RESEARCH

NVIDIA is teaming up with the National Cancer Institute, the U.S. Department of Energy and several national labs on Vice President Biden's "Cancer Moonshot" to deliver a decade of advances in cancer prevention, diagnosis and treatment in just five years. The work will include building and accelerating an AI framework, called CANDLER, as a common discovery platform. To turbocharge the effort, we will use the NVIDIA SATURNV supercomputer to help develop CANDLER. Comprised of 124 DGX-1s, SATURNV is the fastest AI supercomputer today and the most energy-efficient supercomputer in the world.

Argonne
NATIONAL LABORATORY

Lawrence Livermore
National Laboratory

Los Alamos
NATIONAL LABORATORY
EST. 1943

NATIONAL
CANCER
INSTITUTE

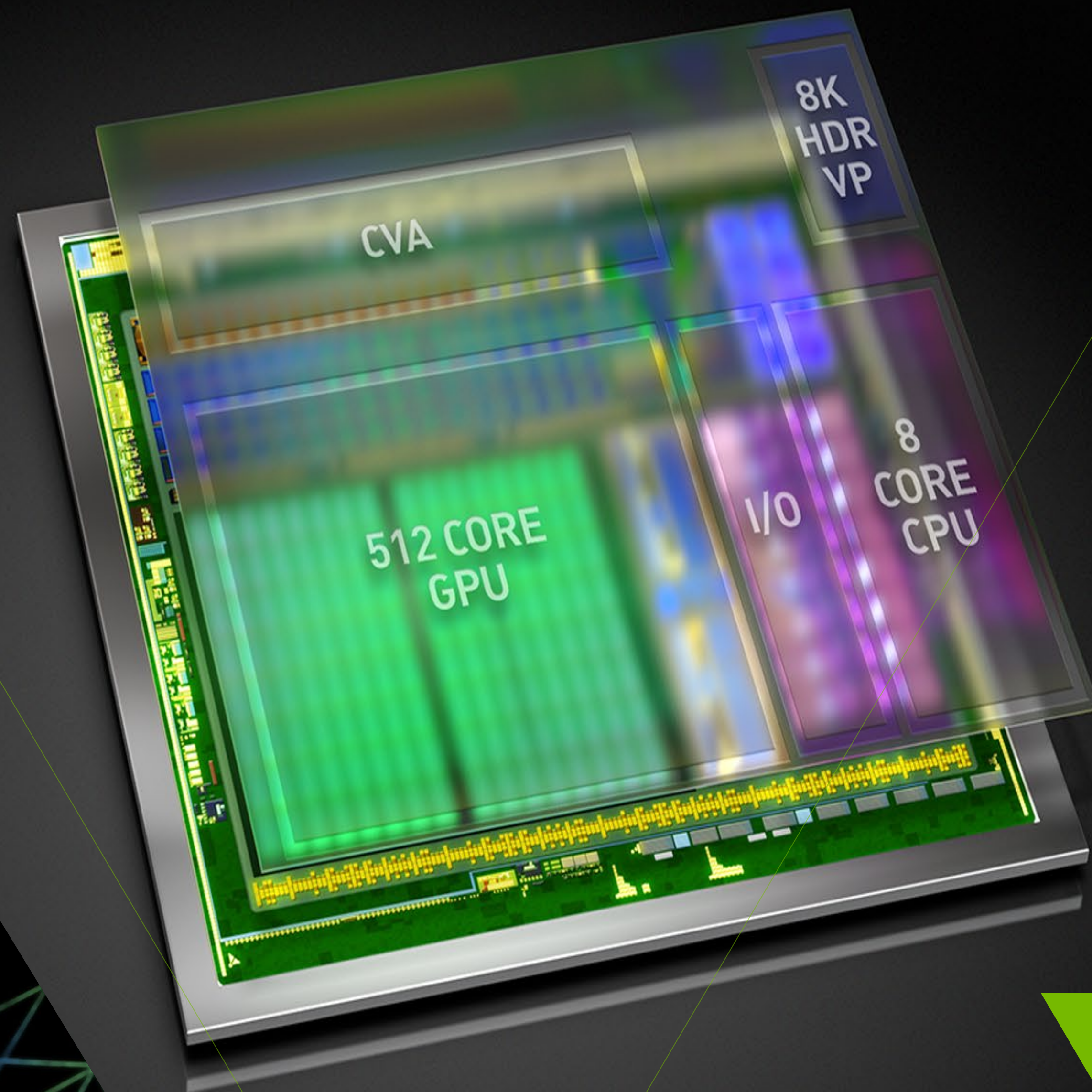
OAK
RIDGE
National Laboratory

U.S. DEPARTMENT OF
ENERGY



THE FUTURE OF AI COMPUTING

NVIDIA is focused on innovation at the intersection of visual processing, high performance computing, and AI — a unique combination at the heart of next-gen intelligent machines. Xavier, the world's first AI supercomputer chip, was built to power this near future. Xavier boasts 7 billion transistors — more complex than the most advanced server-class CPU — and is capable of 20 trillion operations per second of deep learning performance using just 20 watts.





OUR CULTURE A PASSION FOR CRAFTSMANSHIP

NVIDIA is united by a core belief that mastery of our craft lets us tackle the challenges that matter to the world.

This attitude pervades our culture and informs everything we do, from designing amazing products to striving to build one of the world's great companies.





OUR CULTURE
INSPIRED TO GIVE TO
OUR COMMUNITIES

NVIDIA's people share a strong sense of corporate responsibility. Our philanthropic giving this year totaled \$5 million.

Project Inspire, which brings our people together every year to transform local communities, continues to gain momentum. This year alone, employees volunteered more than 17,000 hours and supported education programs that benefited more than 64,000 children.

OUR CULTURE A LEARNING MACHINE

NVIDIA has continuously reinvented itself over two decades.

Our invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics, and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI — the next era of computing.

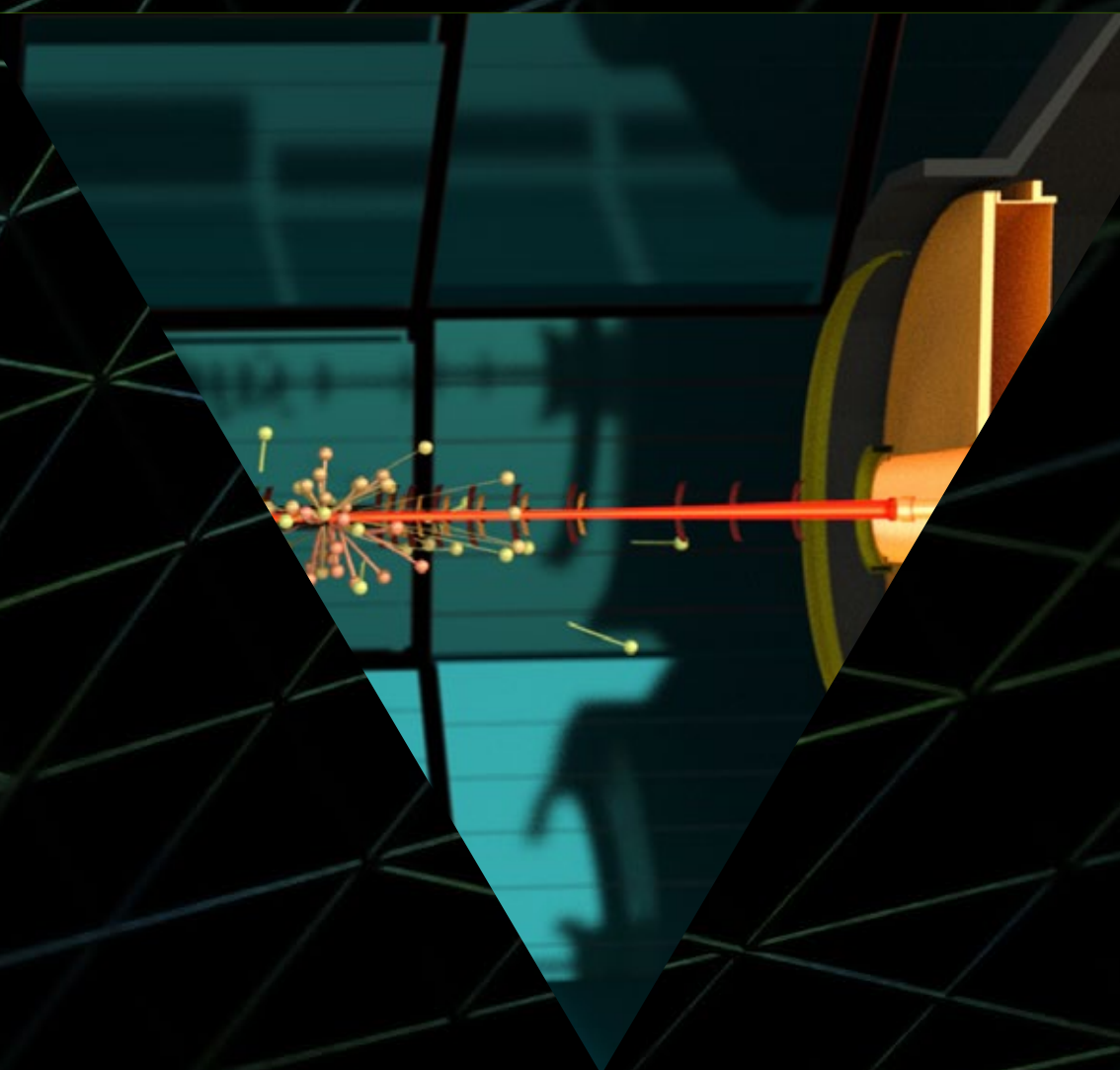
NVIDIA is a “learning machine” that constantly evolves by adapting to new opportunities that are hard to solve, that only we can tackle, and that matter to the world.

PC GRAPHICS



1996

GPU COMPUTING



2006

AI COMPUTING



2016



NVIDIA

- > Founded in 1993
- > Jen-Hsun Huang, Founder & CEO
- > 9,500 employees
- > \$5B in FY16

“World’s Best Performing CEOs” — *Harvard Business Review*

“World’s Most Admired Companies” — *Fortune*

“America’s Greenest Companies” — *Newsweek*

“50 Smartest Companies” — *MIT Tech Review*

“Top 50 Best Places to Work” — *Glassdoor*



