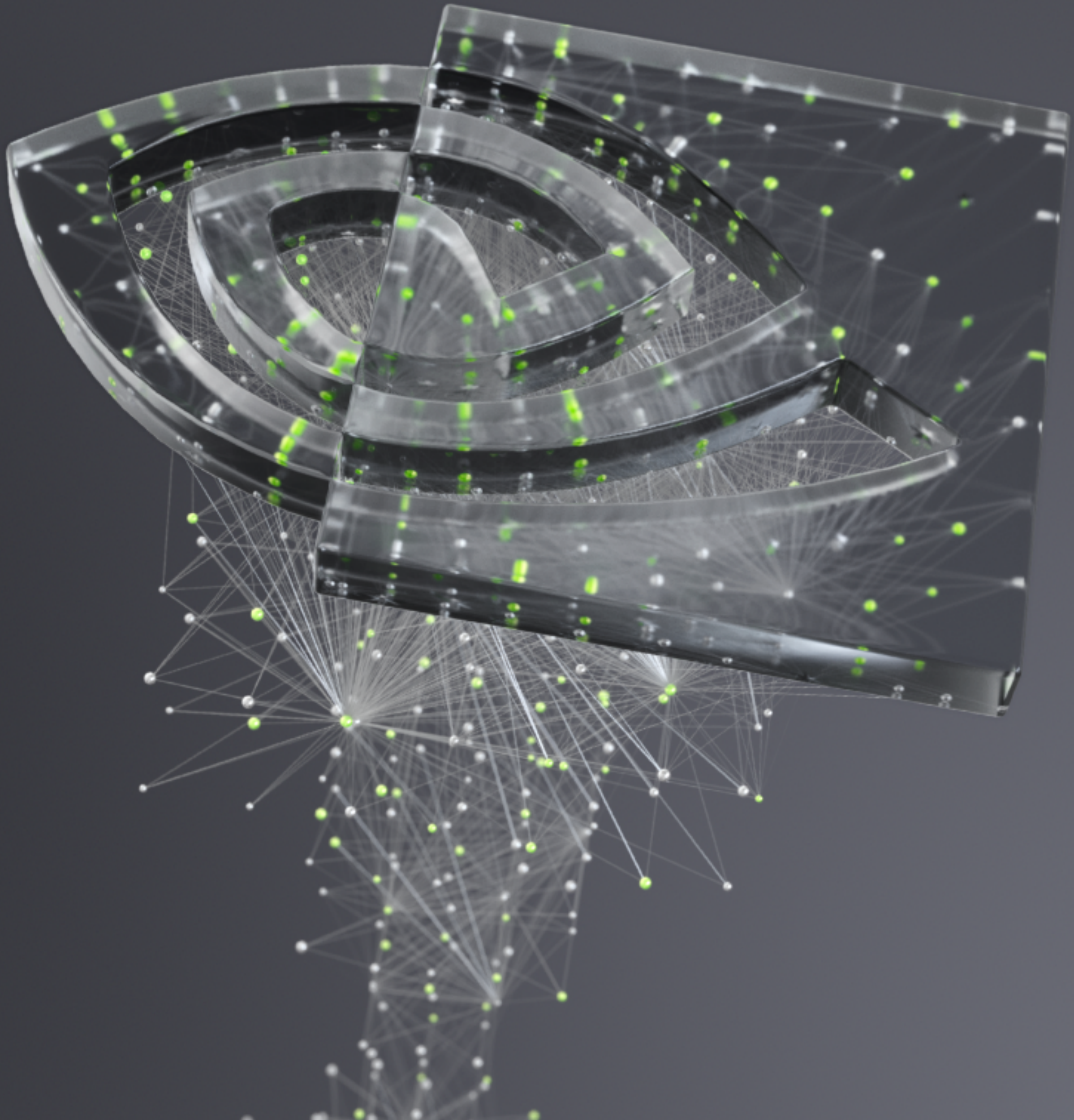
NVIDIA.

A NEW ERA OF ACCELERATED AI

Stories of NVIDIA-powered Inference

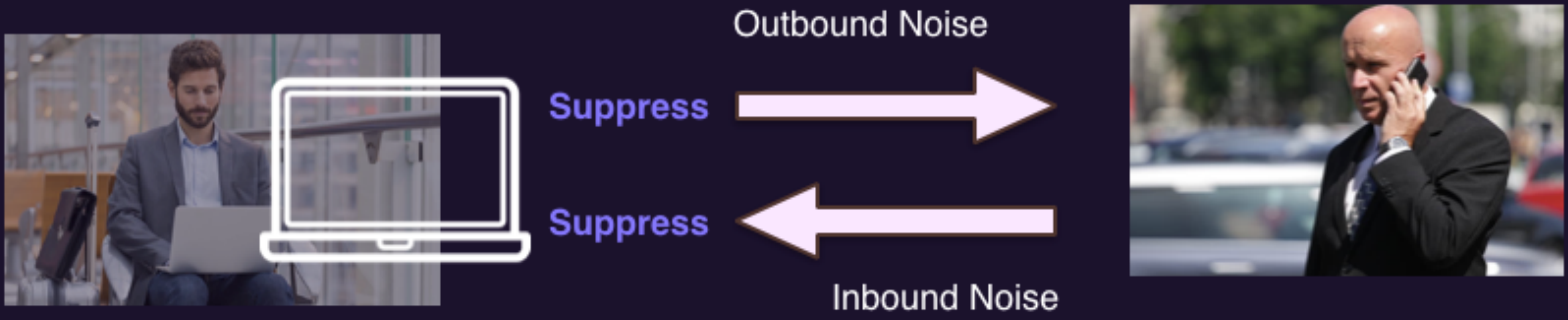


REACHING NEW SPEEDS IN AI DEPLOYMENT

Inference is where AI happens. It’s what powers the immediate, relevant responses of online assistants. It helps doctors see disease faster and make more accurate diagnoses. It refines our shopping experiences. Drives powerful new product designs. Makes our crops healthier, protects our wildlife, and even gives scientists new sightlines into outer space. Across every industry, inference is transforming—accelerating and enhancing—what we do, how we do it, and, ultimately, how we live.

NVIDIA technology is making it happen. From the data center to the edge to IoT devices, NVIDIA’s GPU-accelerated solutions bring leading inference capabilities to use cases across disciplines around the world.

Here are some of those stories.



REAL-TIME NOISE SUPPRESSION FOR AUDIO STREAMS

Background noise is everywhere, and it can be an annoying distraction in business communications. 2Hz, Inc., is bringing clarity to real-time communications with noise suppression technology powered by NVIDIA T4 and V100 GPUs.

2Hz's deep learning algorithms scale up to 20X more than CPUs. And by running NVIDIA® TensorRT™ on GPUs, 2Hz meets the 12 millisecond (ms) latency requirement for real-time communications.

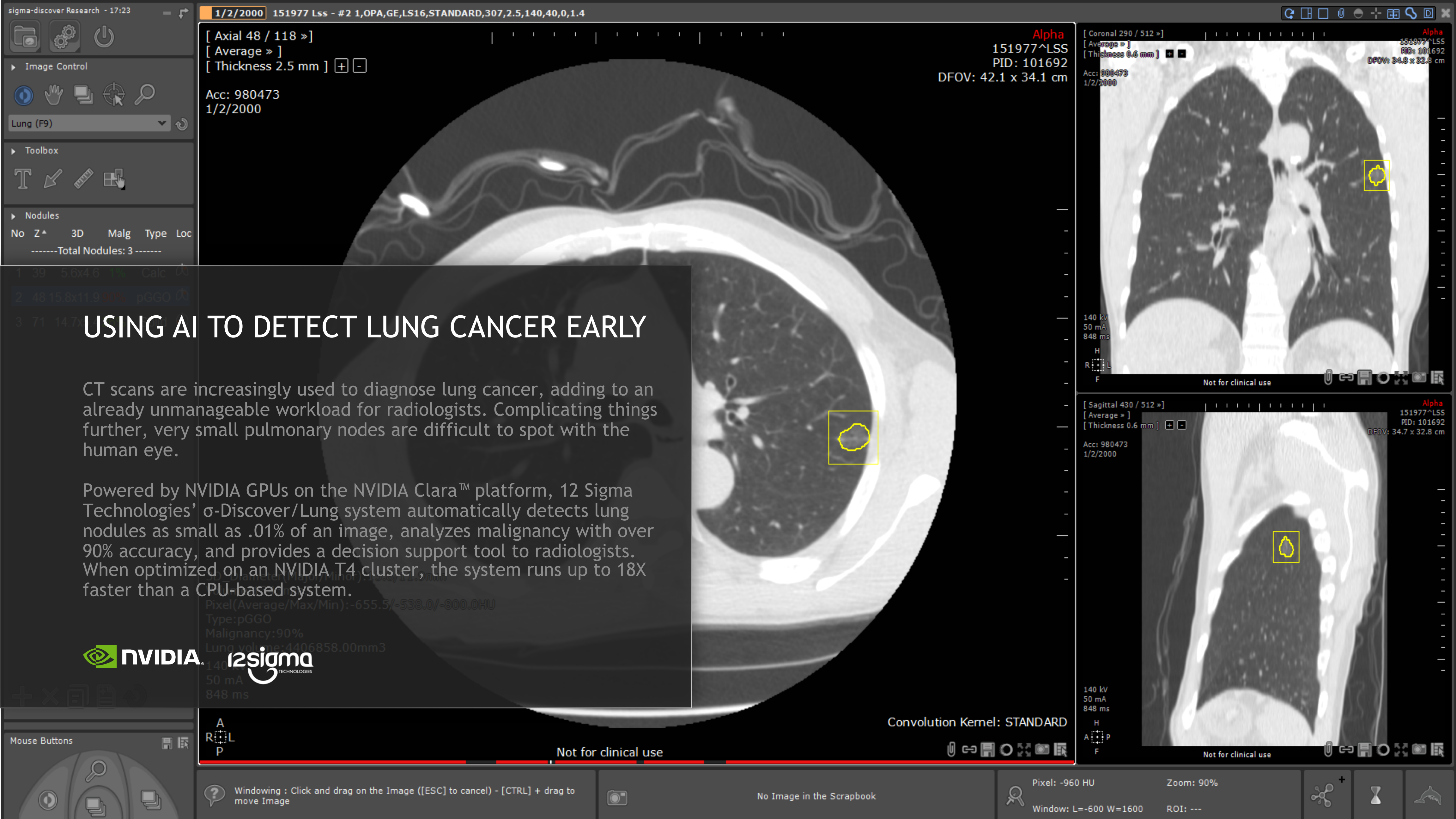
NVIDIA 2Hz

USING AI TO DETECT LUNG CANCER EARLY

CT scans are increasingly used to diagnose lung cancer, adding to an already unmanageable workload for radiologists. Complicating things further, very small pulmonary nodes are difficult to spot with the human eye.

Powered by NVIDIA GPUs on the NVIDIA Clara™ platform, 12 Sigma Technologies' σ-Discover/Lung system automatically detects lung nodules as small as .01% of an image, analyzes malignancy with over 90% accuracy, and provides a decision support tool to radiologists. When optimized on an NVIDIA T4 cluster, the system runs up to 18X faster than a CPU-based system.

NVIDIA. 12sigma TECHNOLOGIES



sigma-discover Research - 17:23

1/2/2000 151977 Lss - #2 1,OPA,GE,LS16,STANDARD,307,2.5,140,40,0,1.4

[Axial 48 / 118 »]
[Average »]
[Thickness 2.5 mm]

Acc: 980473
1/2/2000

Alpha
151977^LSS
PID: 101692
DFOV: 42.1 x 34.1 cm

Convolution Kernel: STANDARD

Not for clinical use

[Coronal 290 / 512 »]
[Average »]
[Thickness 0.6 mm]
Acc: 980473
1/2/2000
Alpha
151977^LSS
PID: 101692
DFOV: 34.8 x 32.8 cm
140 kV
50 mA
848 ms
Not for clinical use

[Sagittal 430 / 512 »]
[Average »]
[Thickness 0.6 mm]
Acc: 980473
1/2/2000
Alpha
151977^LSS
PID: 101692
DFOV: 34.7 x 32.8 cm
140 kV
50 mA
848 ms
Not for clinical use

Windowing : Click and drag on the Image ([ESC] to cancel) - [CTRL] + drag to move Image

No Image in the Scrapbook

Pixel: -960 HU
Zoom: 90%
Window: L=-600 W=1600
ROI: ---

GPU-POWERED SMART FARM MACHINES

With the rising number of herbicide-tolerant weeds, farmers need alternative solutions to the “broadcast-spray” method of weed management. Blue River Technologies (acquired by John Deere) helps control and prevent the evolution of herbicide-resistant weeds with GPU-powered, smart farm equipment.

Blue River Technology’s See & Spray machines use computer vision and machine learning to detect, identify, and make real-time management decisions about each plant in the field. Tractor-mounted smart cameras powered by the NVIDIA Jetson™ platform identify crops and weeds and trigger precisely metered sprays that eliminate unwanted plants while using 90% less herbicide.

NVIDIA BLUE RIVER TECHNOLOGY

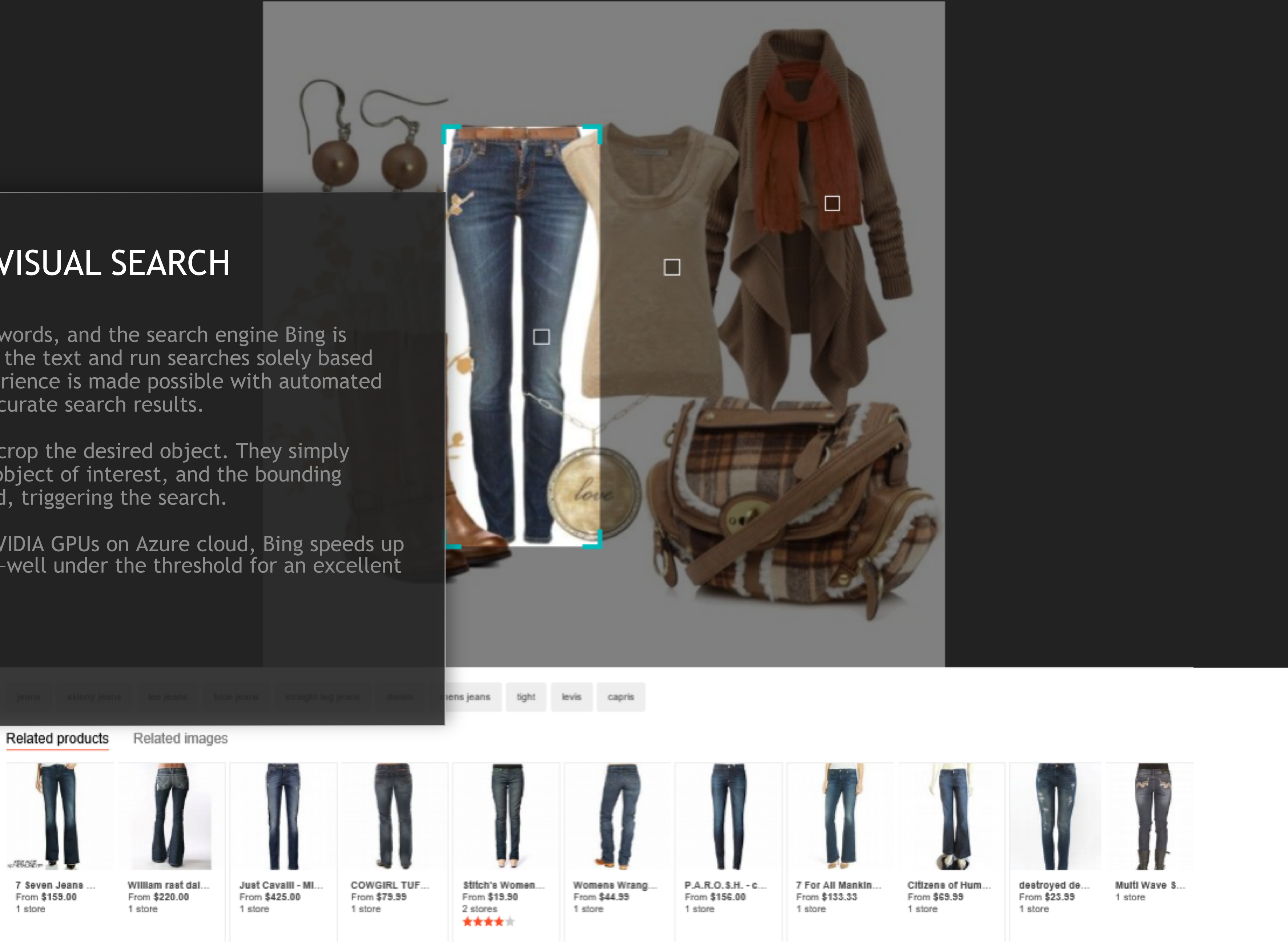
SMARTER, FASTER VISUAL SEARCH

A picture is worth a thousand words, and the search engine Bing is proving it by letting users skip the text and run searches solely based on images. This powerful experience is made possible with automated object detection for quick, accurate search results.

Users don’t need to manually crop the desired object. They simply click on the hotspot over the object of interest, and the bounding box is automatically positioned, triggering the search.

And the search is fast. With NVIDIA GPUs on Azure cloud, Bing speeds up object detection 60X to 40ms—well under the threshold for an excellent user experience.

NVIDIA. Bing



7 Seven Jeans ...
From $159.00
1 store

William rast dal...
From $220.00
1 store

Just Cavalli - Mi...
From $425.00
1 store

COWGIRL TUF...
From $79.99
1 store

Stitch's Women...
From $19.90
2 stores
★★★★☆

Womens Wrang...
From $44.99
1 store

P.A.R.O.S.H. - c...
From $156.00
1 store

7 For All Mankin...
From $133.33
1 store

Citizens of Hum...
From $69.99
1 store

destroyed de...
From $23.99
1 store

Multi Wave S...
1 store

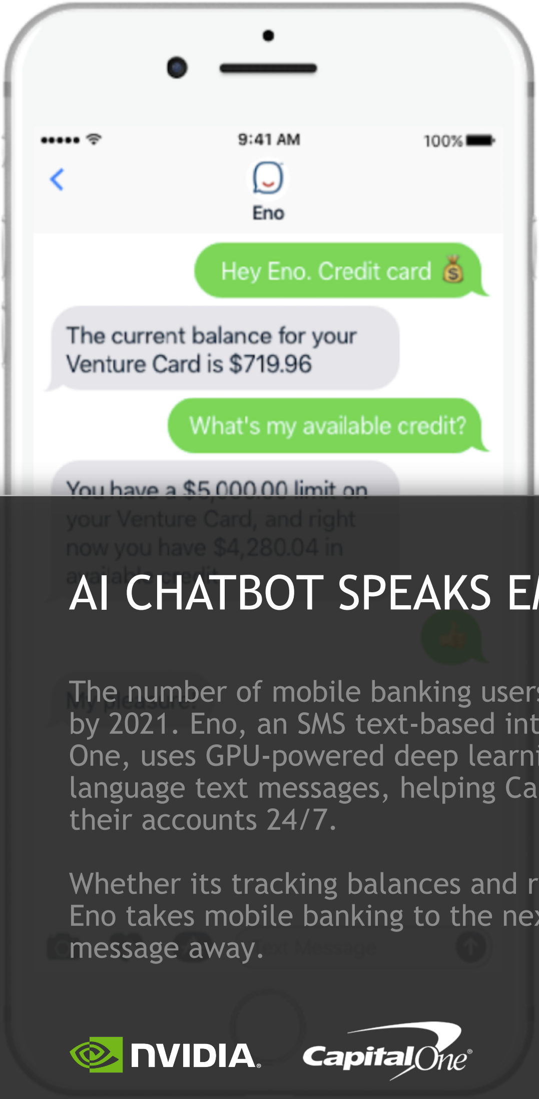
Try Bing:
https://www.bing.com/images/

AI CHATBOT SPEAKS EMOJI

The number of mobile banking users is estimated to reach 2 billion by 2021. Eno, an SMS text-based intelligent assistant from Capital One, uses GPU-powered deep learning to respond to natural language text messages, helping Capital One clients to manage their accounts 24/7.

Whether its tracking balances and recent charges or paying bills, Eno takes mobile banking to the next level—which is just a text message away.

NVIDIA. Capital One



Hey Eno. Credit card 💰

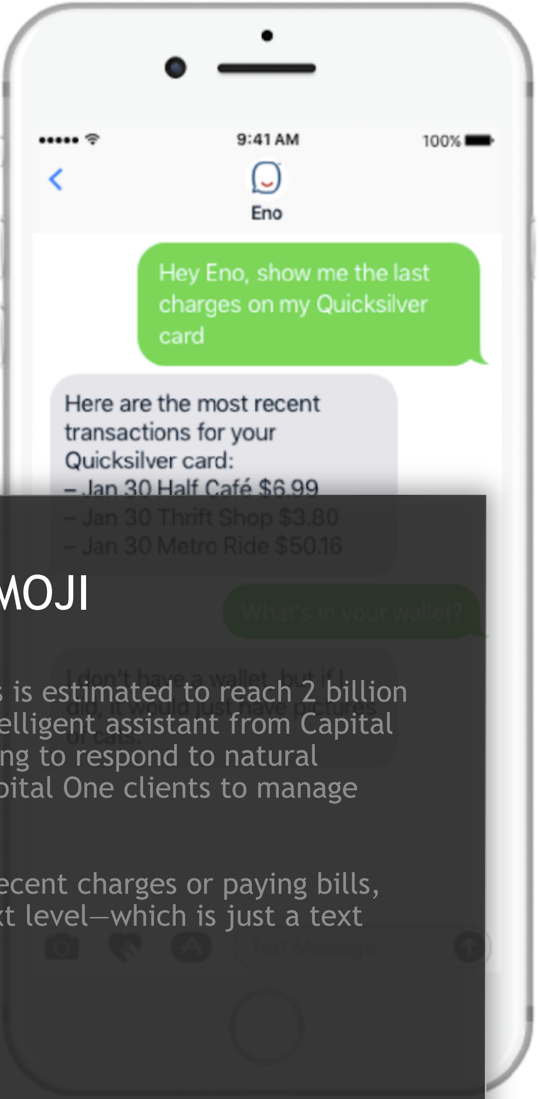
The current balance for your Venture Card is $719.96

What's my available credit?

You have a $5,000.00 limit on your Venture Card, and right now you have $4,280.04 in available credit.

👍

My pleasure!

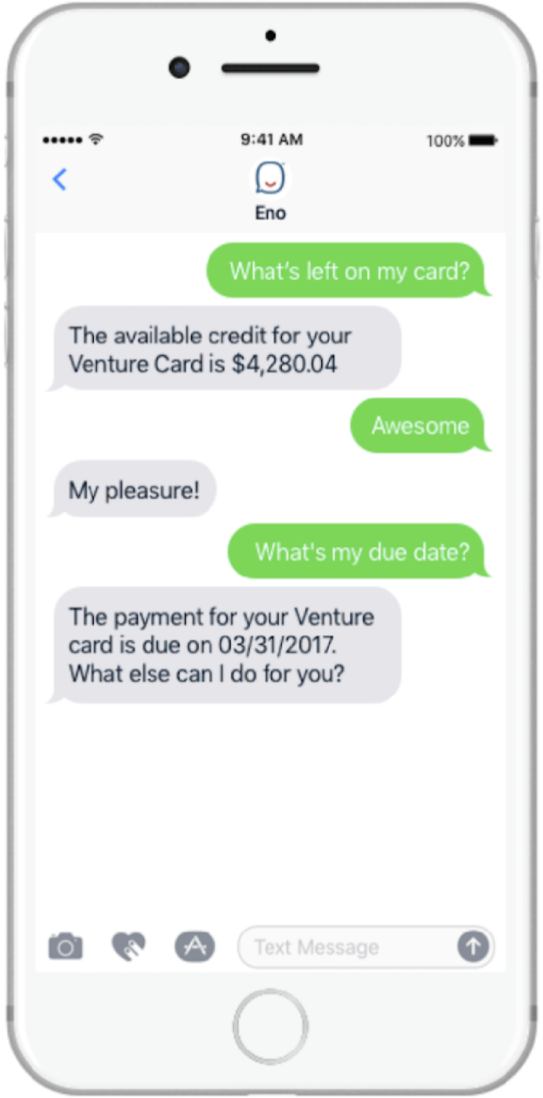


Hey Eno, show me the last charges on my Quicksilver card

Here are the most recent transactions for your Quicksilver card:
– Jan 30 Half Café $6.99
– Jan 30 Thrift Shop $3.80
– Jan 30 Metro Ride $50.16

What's in your wallet?

I don't have a wallet, but if I did, it would just have pictures of cats.



What's left on my card?

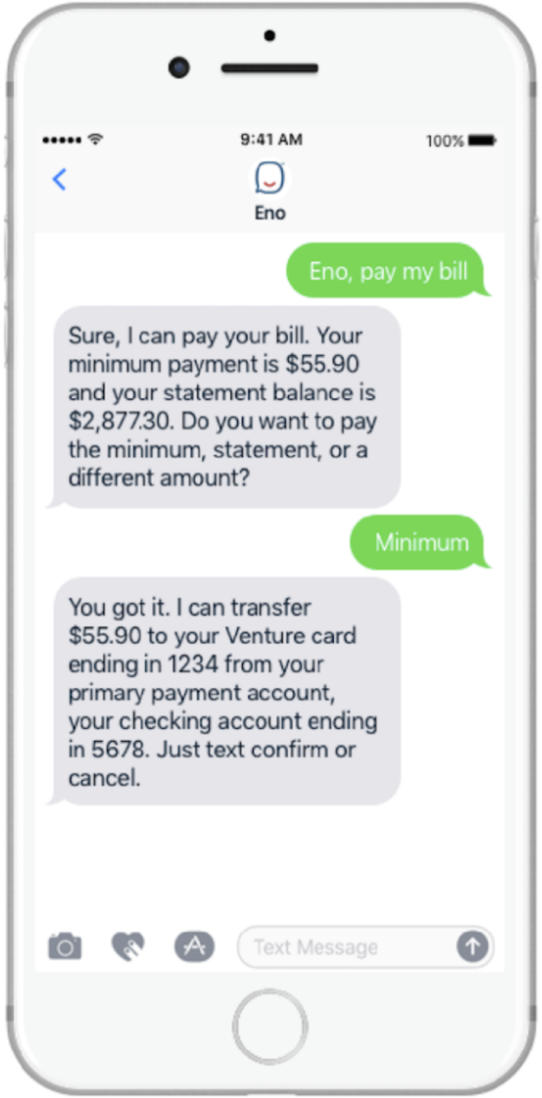
The available credit for your Venture Card is $4,280.04

Awesome

My pleasure!

What's my due date?

The payment for your Venture card is due on 03/31/2017. What else can I do for you?



Eno, pay my bill

Sure, I can pay your bill. Your minimum payment is $55.90 and your statement balance is $2,877.30. Do you want to pay the minimum, statement, or a different amount?

Minimum

You got it. I can transfer $55.90 to your Venture card ending in 1234 from your primary payment account, your checking account ending in 5678. Just text confirm or cancel.

SPEEDING UP NASCAR WITH AI

In NASCAR, the difference between victory and defeat can be measured in milliseconds. That's why Ford Motorsports is optimizing aerodynamics with AI powered by NVIDIA DGX-1™ running TensorRT. Hours before a race, Ford trains its model to recognize the field. During the race, the AI analyzes video feeds and assesses performance in real time. Ford teams receive immediate feedback and adjust their cars as needed to stay ahead of the competition.

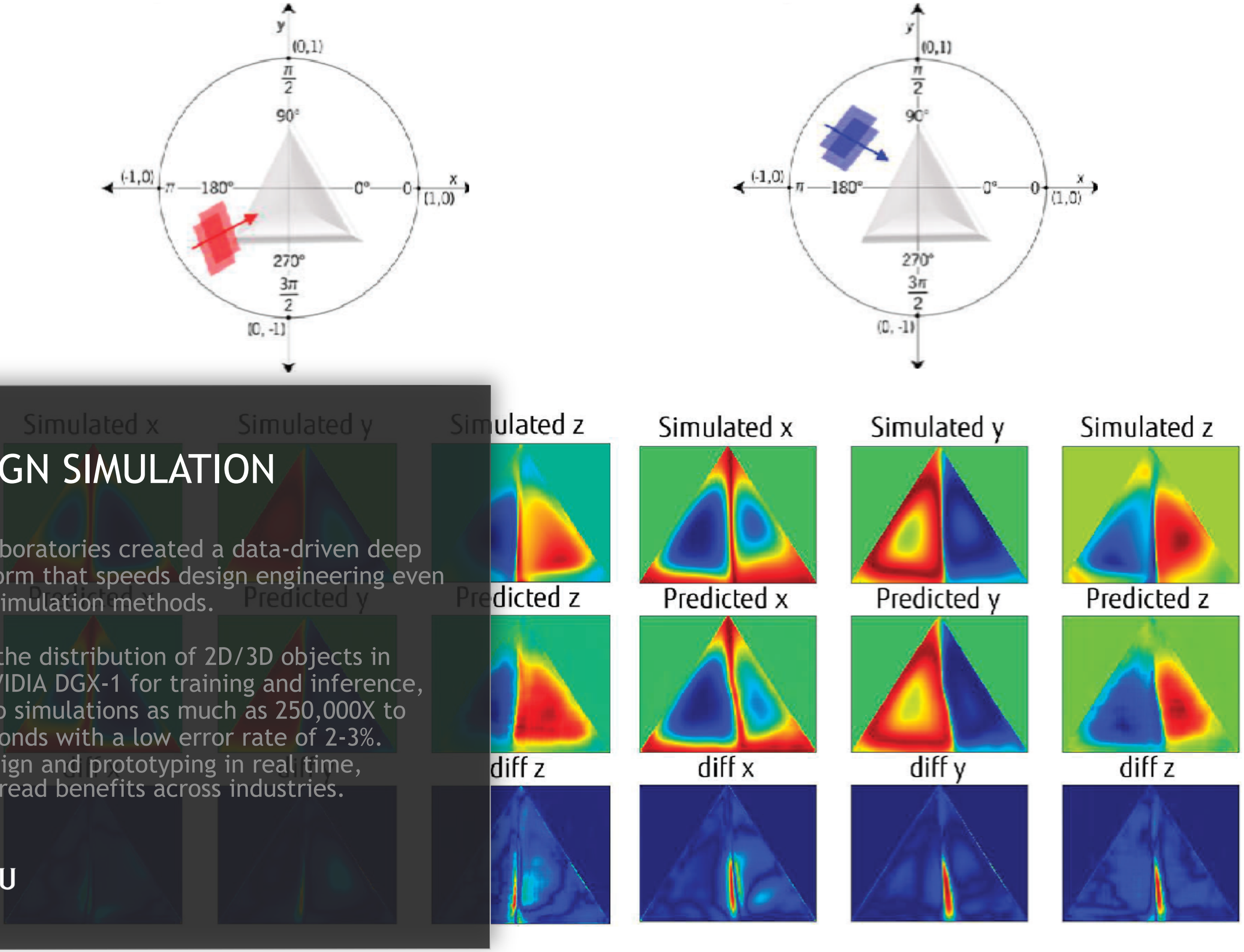
NVIDIA. Ford

REAL-TIME DESIGN SIMULATION

Researchers at Fujitsu Laboratories created a data-driven deep learning simulation platform that speeds design engineering even more than conventional simulation methods.

DeepSim-HiPAC predicts the distribution of 2D/3D objects in real time. Powered by NVIDIA DGX-1 for training and inference, DeepSim-HiPAC speeds up simulations as much as 250,000X to deliver results in milliseconds with a low error rate of 2-3%. This enables dynamic design and prototyping in real time, which can deliver widespread benefits across industries.

NVIDIA FUJITSU



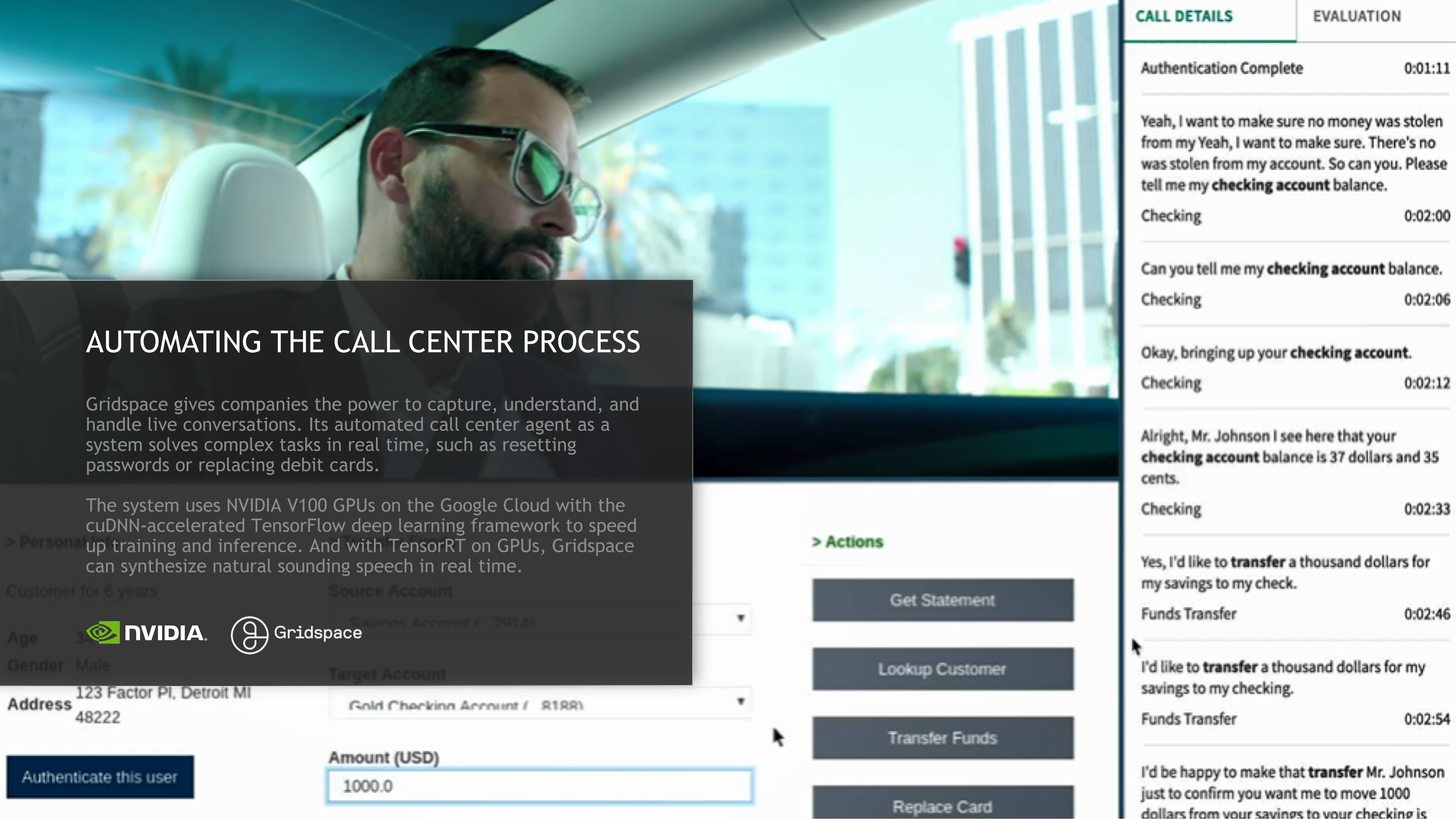
Images showcase the ability of DeepSim-HiPAC to accommodate for an arbitrary orientation in the applied external magnetic field, predicting fields' distribution for two different orientations (Red and Blue) of the same structure.

AUTOMATING THE CALL CENTER PROCESS

Gridspace gives companies the power to capture, understand, and handle live conversations. Its automated call center agent as a system solves complex tasks in real time, such as resetting passwords or replacing debit cards.

The system uses NVIDIA V100 GPUs on the Google Cloud with the cuDNN-accelerated TensorFlow deep learning framework to speed up training and inference. And with TensorRT on GPUs, Gridspace can synthesize natural sounding speech in real time.

NVIDIA Gridspace



> Personal Info

Customer for 6 years

Age 34

Gender Male

Address 123 Factor Pl, Detroit MI 48222

Source Account: Savings Account (7914)

Target Account: Gold Checking Account (8188)

Amount (USD): 1000.0

> Actions

Authentication Complete 0:01:11

Yeah, I want to make sure no money was stolen from my Yeah, I want to make sure. There's no was stolen from my account. So can you. Please tell me my **checking account** balance.

Checking 0:02:00

Can you tell me my **checking account** balance.

Checking 0:02:06

Okay, bringing up your **checking account**.

Checking 0:02:12

Alright, Mr. Johnson I see here that your **checking account** balance is 37 dollars and 35 cents.

Checking 0:02:33

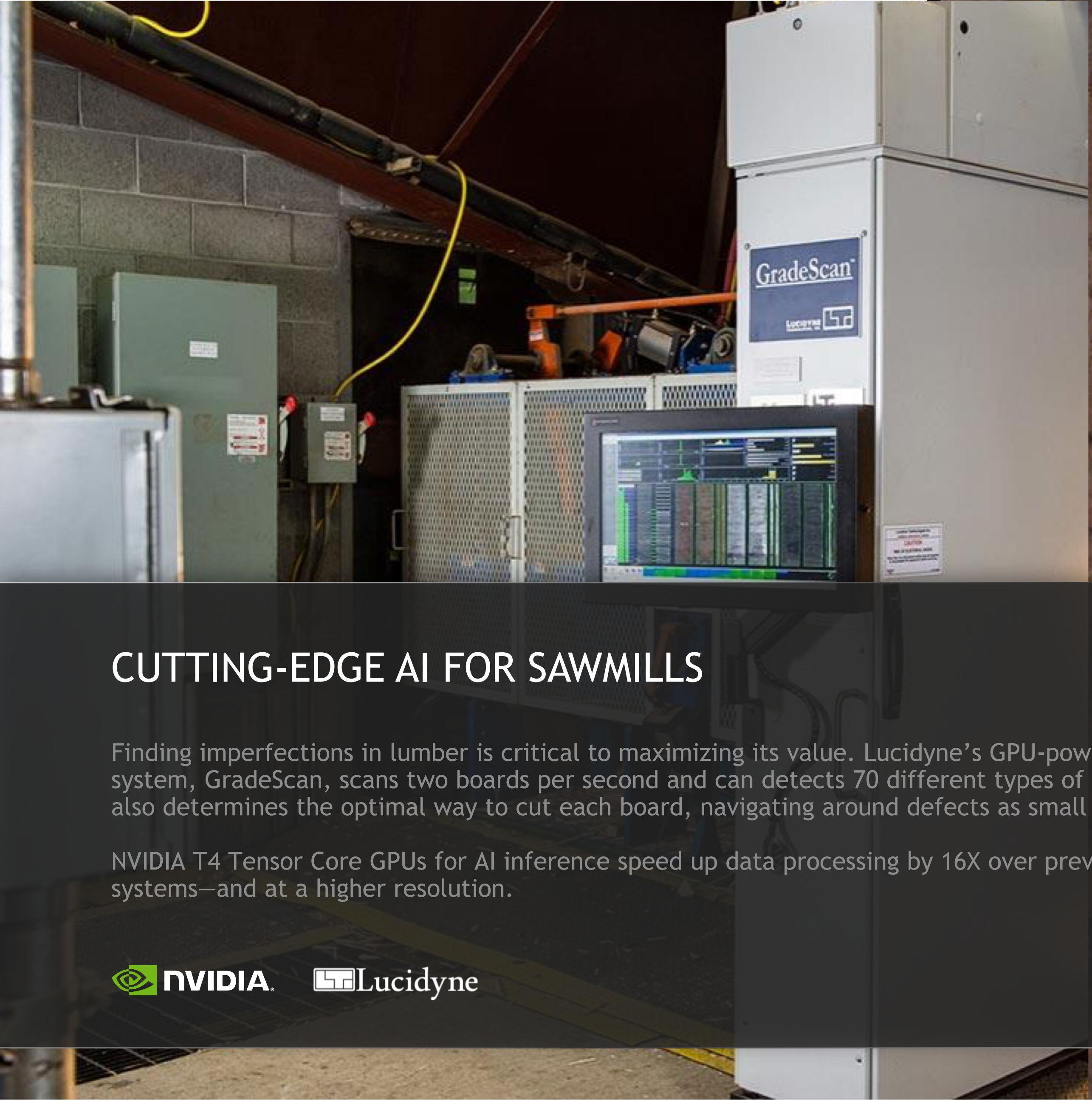
Yes, I'd like to **transfer** a thousand dollars for my savings to my check.

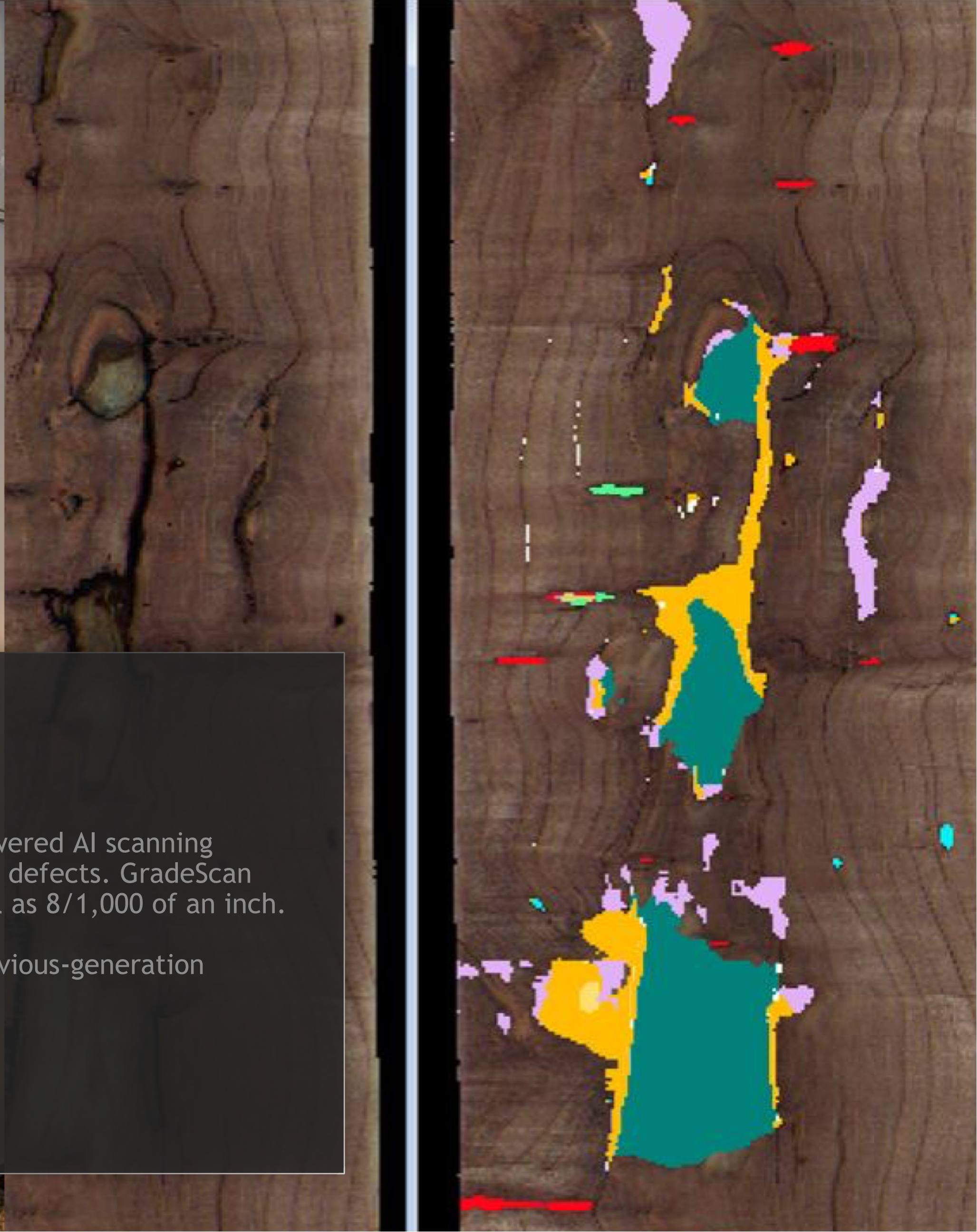
Funds Transfer 0:02:46

I'd like to **transfer** a thousand dollars for my savings to my checking.

Funds Transfer 0:02:54

I'd be happy to make that **transfer** Mr. Johnson just to confirm you want me to move 1000 dollars from your savings to your checking is





CUTTING-EDGE AI FOR SAWMILLS

Finding imperfections in lumber is critical to maximizing its value. Lucidyne’s GPU-powered AI scanning system, GradeScan, scans two boards per second and can detects 70 different types of defects. GradeScan also determines the optimal way to cut each board, navigating around defects as small as 8/1,000 of an inch.

NVIDIA T4 Tensor Core GPUs for AI inference speed up data processing by 16X over previous-generation systems—and at a higher resolution.

NVIDIA Lucidyne

“SEEING” GRAVITY FOR THE FIRST TIME

In September 2015, 100 years after Einstein predicted them, gravitational waves were observed for the first time. Astronomers at the Laser Interferometer Gravitational-Wave Observatory have since used GPU-powered deep learning to process gravitational wave data 100X faster than previous methods, making real-time analysis possible and putting us one step closer to understanding the universe’s oldest secrets.

NVIDIA. NCSA

AI detects a mite in frame 24

SAFEGUARDING HONEYBEES WITH AN AI ALARM SYSTEM

Habitat loss and pesticides are the primary threats to bee populations, but the Varroa mite can devastate entire colonies. To combat the Varroa, high school student Jade Greenberg turned to AI. Her solution—NVIDIA Jetson TX2, NVIDIA DGX Station™, TensorRT, Microsoft Cognitive Toolkit, and Kinetica—uses sensors and cameras to feed a convolutional neural network that assess hive health in real time and converts the data into a visual early warning system for beekeepers.

NVIDIA kinetica Microsoft

Image courtesy of Piscigate

IMPROVING GLOBAL COMMUNICATIONS WITH AI

The voice and speech recognition market, which is estimated to reach USD $31.82 billion by 2025, promises to improve global communications. LOVO, by Orbis.ai, converts a speaker's vocal identity—accent, tone, and speed—into one that's familiar to the listener.

Powered by NVIDIA V100 GPUs and the NVIDIA DGX Station, LOVO converts speech about 40X faster than real time, with over 90% word accuracy and a voice quality mean opinion score (MOS) of 4.1 (out of 5).

REAL-TIME FRAUD DETECTION

Recently, PayPal was looking to deploy a new fraud detection system. The team working on it set a high bar: This system had to operate worldwide 24/7 and work in real time to protect customer transactions from potential fraud. In analyzing the system, it became evident that CPU-only servers couldn’t meet these requirements.

Using NVIDIA T4 GPUs, PayPal delivered a new level of service, using GPU inference to improve real-time fraud detection by 10% while lowering server capacity by nearly 8X.

NVIDIA. PayPal

> “PayPal needed GPUs to accelerate the deployment of our newest worldwide system and to enable capabilities that were previously impossible.”

PayPal

> *“Pinterest uses state of the art computer vision technology to build a sophisticated understanding of over 175 billion pins. We rely on GPUs for training and evaluating our recognition models and for performing real time inference.”*

AUTONOMOUS ON-DEMAND URBAN DELIVERY ROBOT

"In order to make the robot travel safely in the city, on the sidewalk, in the neighborhood, autonomously, requires us to use GPUs. Our team is using TensorFlow for model training, testing, and developing. After we train the model in TensorFlow, we convert the model to TensorRT and we deploy the NVIDIA Jetson AGX Xavier™ platform using [the NVIDIA Deep Learning Accelerator (NVDLA)] . . . By using FP16 half-precision together with NVDLA, we got more than 40X speedup."

— Zhenyu Guo, Director of Artificial Intelligence, Postmates X

NVIDIA Postmates

USING AI TO DETECT DEMENTIA EARLIER

Dementia, which affects 50 million people worldwide, can take months to years to diagnose, as doctors must observe how a patient's condition progresses over time. Quantib is using GPU-powered deep learning to help radiologists monitor dementia and diagnose cases earlier.

Currently installed in 20 countries, Quantib's AI differentiates between brain atrophy patterns indicative of Alzheimer's and other kinds of dementia and generates a segmentation report in just minutes. With the NVIDIA T4 GPU for inference, Quantib algorithms run up to 24X faster than CPUs.

NVIDIA Quantib

CONTROLLING AIR TRAFFIC WITH AI

From autopilot systems to customer service to predicting weather, AI is transforming aviation. With Aimee—a GPU-powered framework for AI solutions from Searidge Technologies—air traffic control no longer needs a direct sightline. Aimee analyzes video feeds from hundreds of cameras, enabling air traffic controllers to look past occlusions and "see" every runway, taxiway, tarmac, and gate without looking away from their workstations.

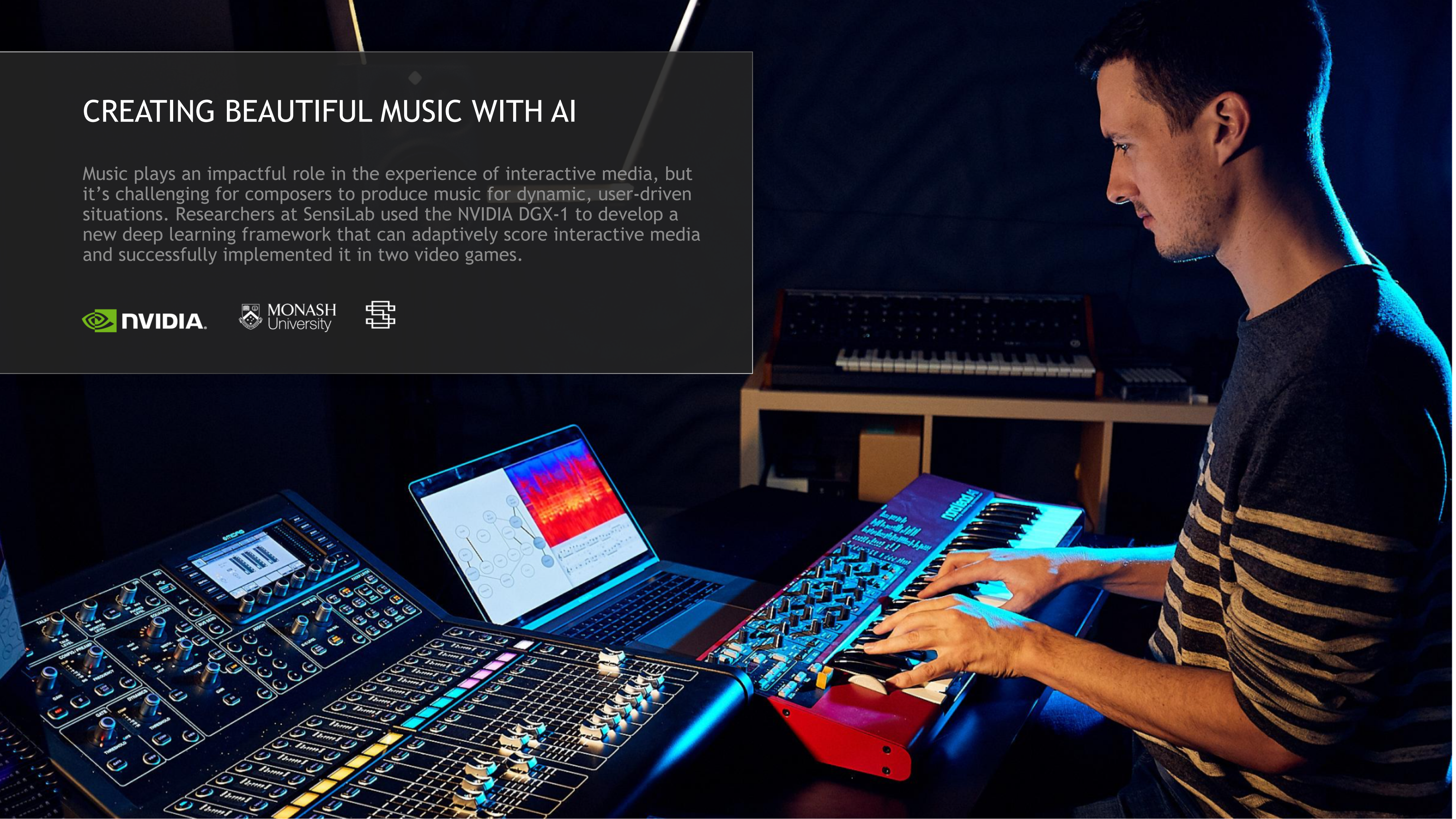
NVIDIA SEARIDGE TECHNOLOGIES

CREATING BEAUTIFUL MUSIC WITH AI

Music plays an impactful role in the experience of interactive media, but it's challenging for composers to produce music for dynamic, user-driven situations. Researchers at SensiLab used the NVIDIA DGX-1 to develop a new deep learning framework that can adaptively score interactive media and successfully implemented it in two video games.

NVIDIA

MONASH University



> *“Snap’s monetization algorithms have the single biggest impact to our advertisers and shareholders. NVIDIA T4 GPUs for inference on the Google Cloud Platform will enable us to increase advertising efficacy while at the same time lowering costs when compared to a CPU-only implementation.”*

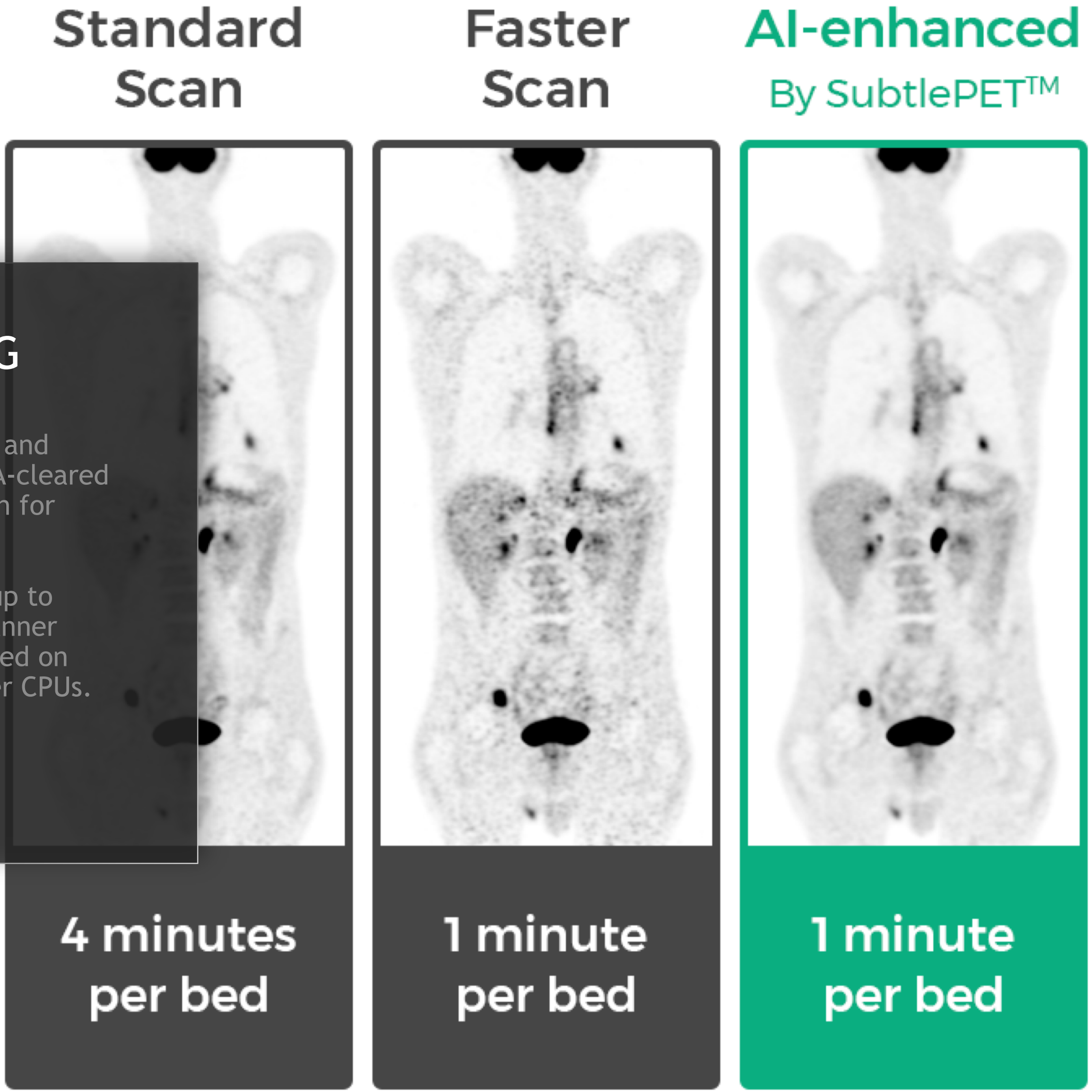
FASTER, SMARTER MEDICAL IMAGING

PET scans are critical in modern radiology for diagnosing and predicting cancer. SubtlePET is Subtle Medical's first FDA-cleared deep learning software trained on the NVIDIA DGX Station for real-time image enhancement.

SubtlePET enables the rendering of high-quality images up to 4X faster than the typical acquisitiontime, increasing scanner throughput and increasing patient comfort. When deployed on NVIDIA T4, SubtlePET's inference is accelerated 3.5X over CPUs.

NVIDIA

SUBTLE MEDICAL



Standard Scan

4 minutes per bed

Faster Scan

1 minute per bed

AI-enhanced By SubtlePET™

1 minute per bed

THE NEW AI ERA FOR RETAILERS

To make in-store retail experiences as streamlined as online experiences, Tracxpoint created the Artificial Intelligence Cart (AIC). Using the NVIDIA DGX Station for training, TensorRT for inference, and the NVIDIA DeepStream SDK on Jetson TX2 for real-time video analytics, AIC can recognize 100,000 products in under a second.

Customers simply place products in their carts and then receive personalized offers from \suppliers in real time as they navigate the supermarket with ease. When they’re done shopping, they can pay digitally from their smart cart.

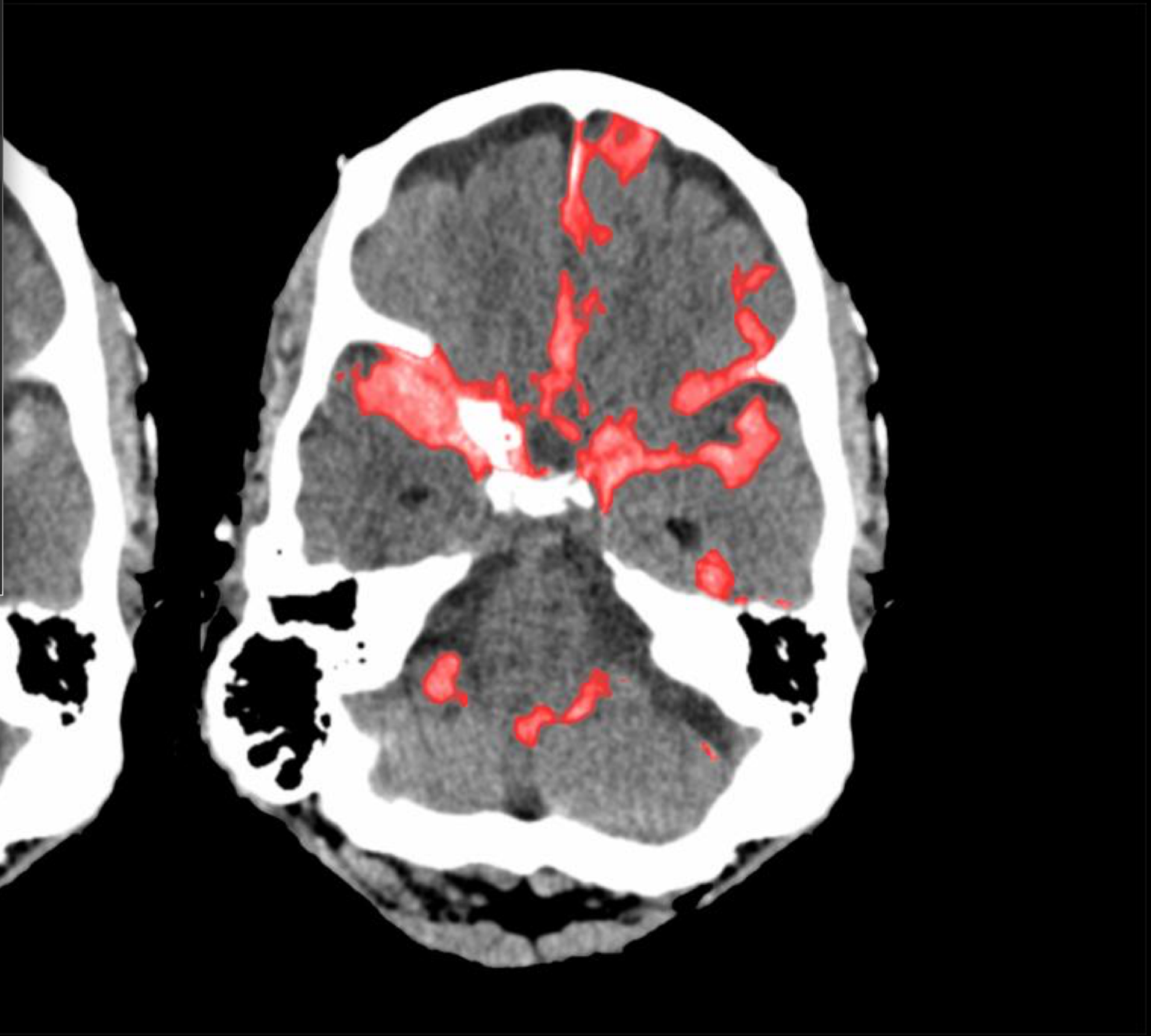
NVIDIA. TRACXPOINT

> “Using GPUs made it possible to enable media understanding on our platform, not just by drastically reducing media deep learning models training time, but also by allowing us to derive real-time understanding of live videos at inference time.”

AI: WHEN EVERY MINUTE MATTERS

For instances of head trauma, doctors order a CT scan to assess injury and determine treatment. With over 80 million CT scans performed annually in the U.S., AI could help radiologists manage busy workloads and speed time to critical care.

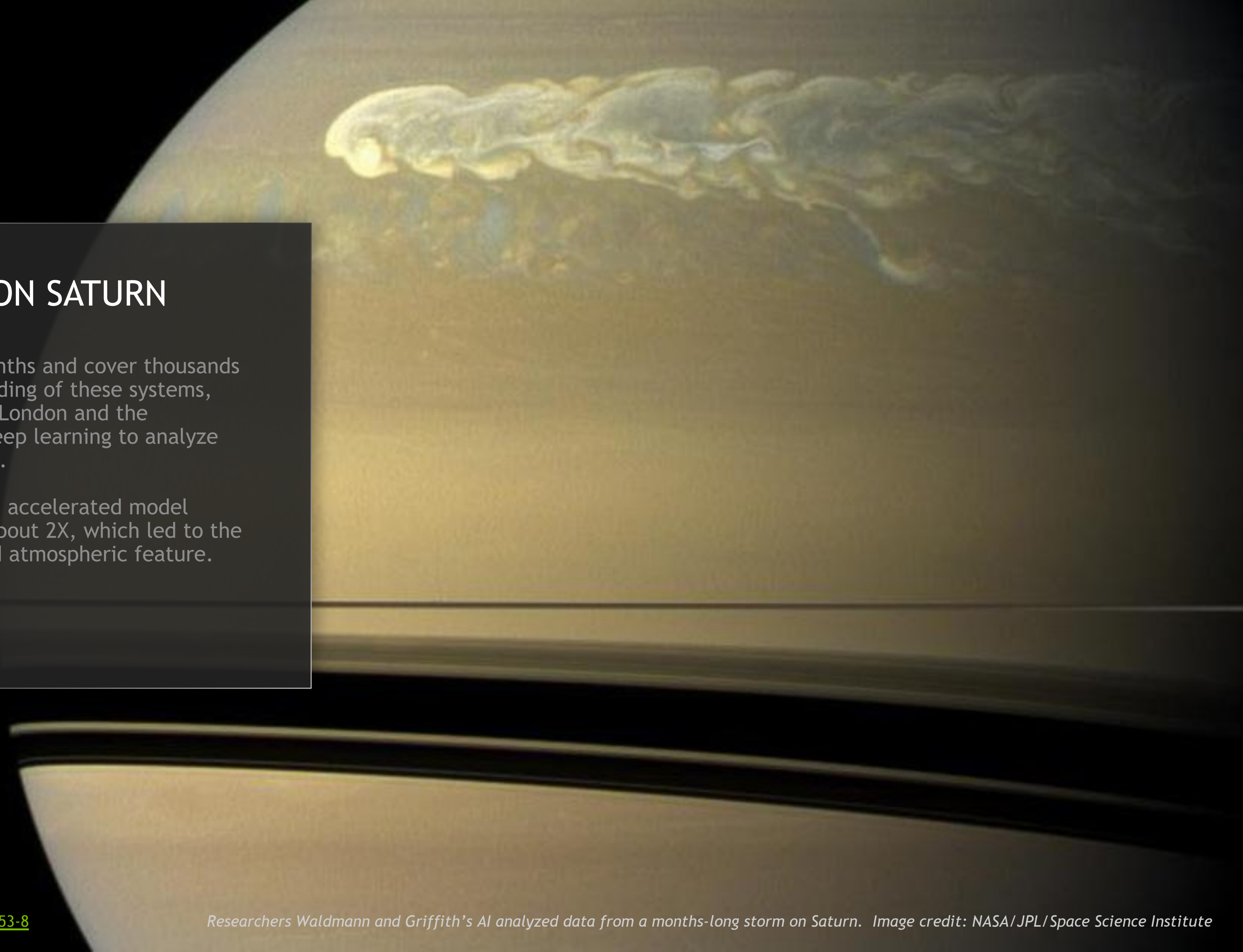
Researchers at UC Berkeley and the UCSF School of Medicine are using NVIDIA V100 Tensor Core GPUs through AWS for the training and inference of an AI model that diagnoses intracranial hemorrhage from CT scans with 99% accuracy — rivaling the performance of experienced radiologists.



ANALYZING WEATHER ON SATURN

Saturn’s extreme storms last for months and cover thousands of miles. To gain a better understanding of these systems, researchers from University College London and the University of Arizona are applying deep learning to analyze data collected by the Cassini orbiter.

Using an NVIDIA V100 GPU, the team accelerated model training by 30X and inferencing by about 2X, which led to the discovery of a previously undetected atmospheric feature.



Researchers Waldmann and Griffith’s AI analyzed data from a months-long storm on Saturn. Image credit: NASA/JPL/Space Science Institute

https://www.nature.com/articles/s41550-019-0753-8

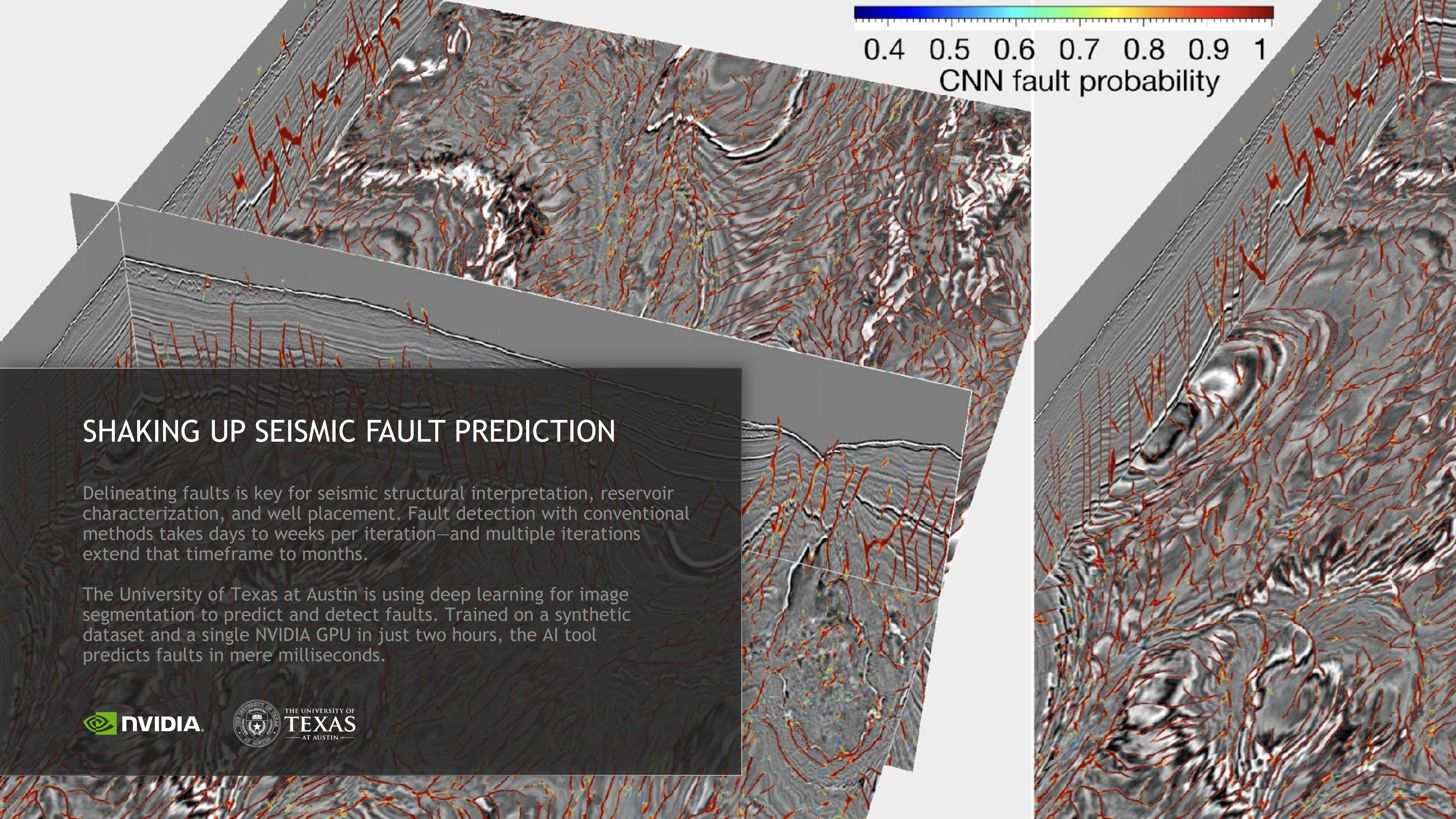
SHAKING UP SEISMIC FAULT PREDICTION

Delineating faults is key for seismic structural interpretation, reservoir characterization, and well placement. Fault detection with conventional methods takes days to weeks per iteration—and multiple iterations extend that timeframe to months.

The University of Texas at Austin is using deep learning for image segmentation to predict and detect faults. Trained on a synthetic dataset and a single NVIDIA GPU in just two hours, the AI tool predicts faults in mere milliseconds.

NVIDIA

THE UNIVERSITY OF TEXAS AT AUSTIN



THE BRAINS BEHIND SMART CITIES

Verizon’s Smart Communities Group is on a mission to make cities safer, smarter, and greener. Using NVIDIA Metropolis, an edge-to-cloud video platform for building smarter, faster AI-powered applications, Verizon is working to collect and analyze multiple streams of video data to improve traffic flow, enhance pedestrian safety, optimize parking, and more.

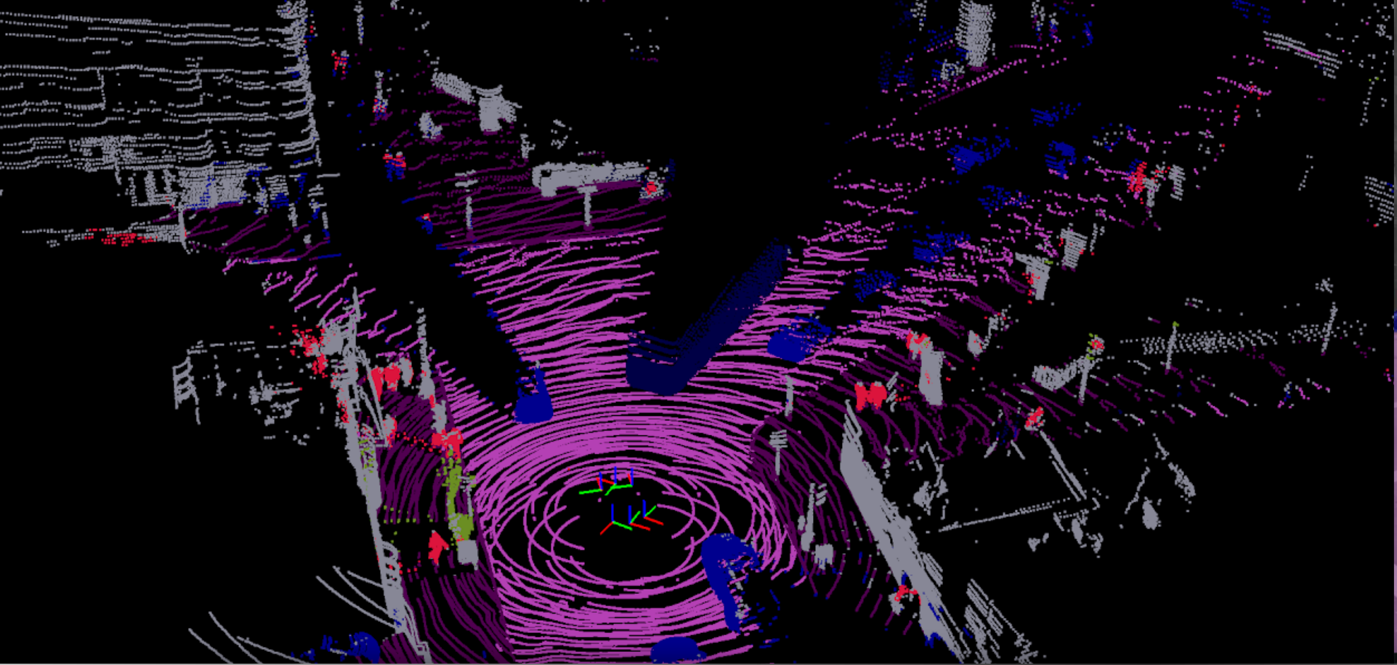
NVIDIA verizon

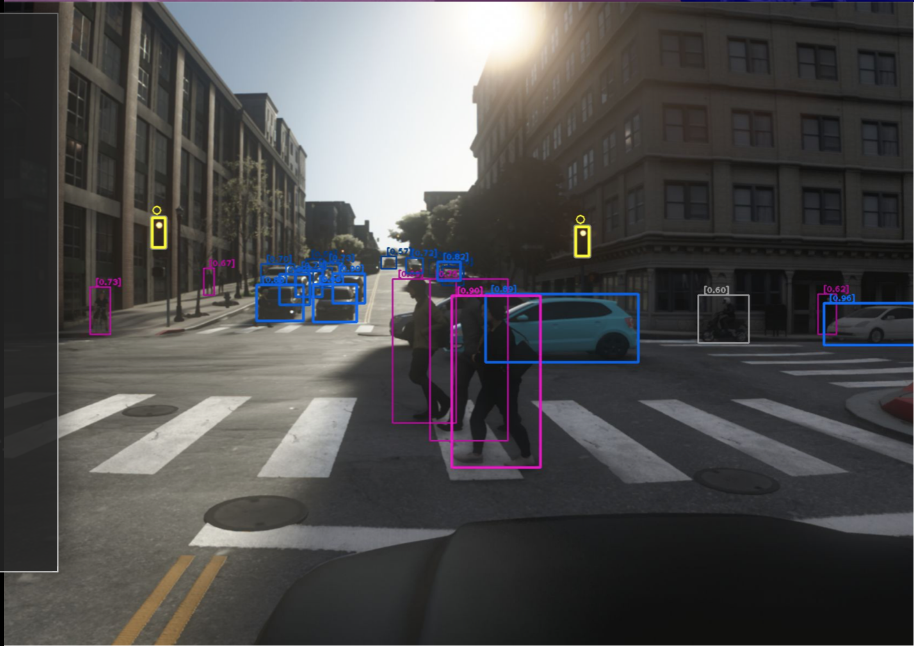
AI TAKES A BITE OUT OF FOOD WASTE

Approximately one-third of all food harvested or produced goes to waste. The hospitality sector alone discards nearly $100 billion worth of food each year.

Winnow empowers commercial kitchens to reduce food waste. Winnow uses real-time deep learning and inference at the edge powered by the NVIDIA Jetson TX2 to automatically detect, identify, and measure food at the point it’s thrown away—helping commercial kitchens save more than $30 million in annualized food costs to date.

NVIDIA. winnow





THE FUTURE OF AUTONOMOUS MOBILITY

Autonomous driving systems use various neural network models that require extremely fast and accurate computation on GPUs.

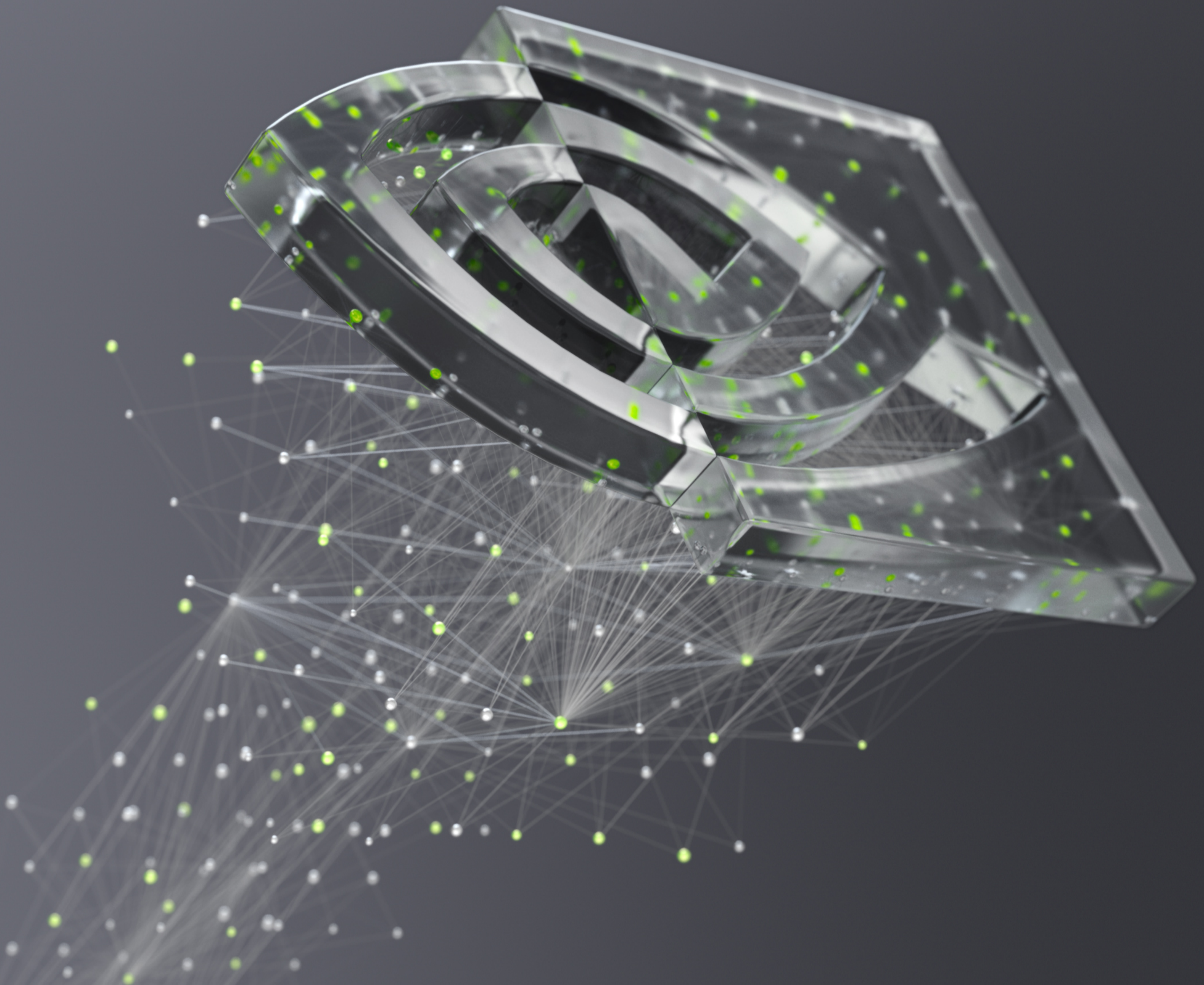
To improve latency while maintaining accuracy for vision, prediction, and lidar models, Zoox uses GPU-powered inference with NVIDIA TensorRT. Zoox's TensorRT-based applications perform up to 19X faster than inference in TensorFlow and Caffe natively on NVIDIA GPUs.

NVIDIA. ZOOX

ACCELERATED INFERENCE. NEW CAPABILITIES.

From data center to edge and across all AI use cases, NVIDIA technology is shaping the current and future state of AI deployment. More than just powering new services, it powers innovation, creates new products, boosts revenues, streamlines operations, and transforms lives.

Learn more at: https://www.nvidia.com/en-us/deep-learning-ai/solutions/inference-platform/



NVIDIA