Accelerate AI Applications Using NVIDIA GPUs and AI Software

Dr Maggie Xuemeng Zhang, 02/05/2022

AGENDA

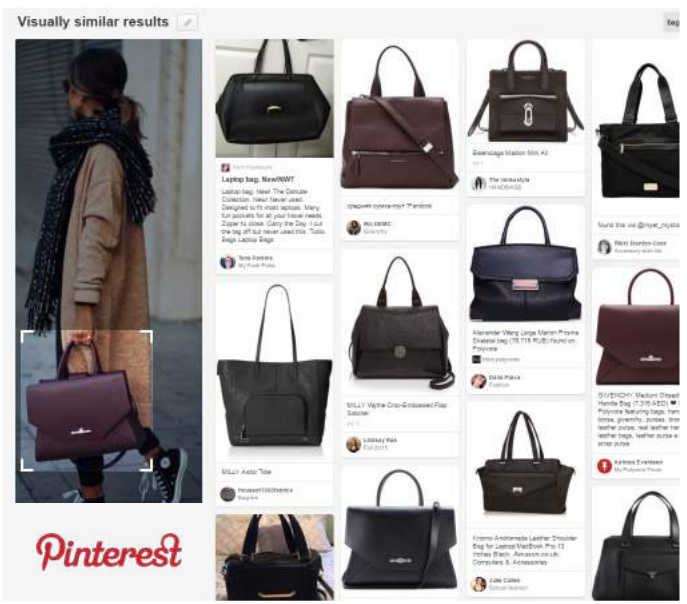
Introduction to AI

NVIDIA AI Software Update

Programs and Success Stories

Practical Examples Of AI

“Find where I parked my car”



“Find the bag I just saw in this magazine”

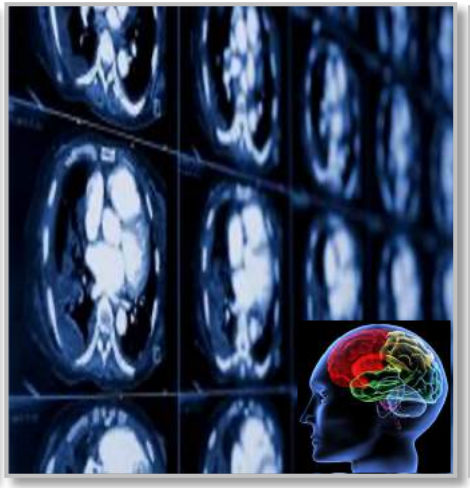
“What movie should I watch next?”

AI Is Sweeping Across Industries

Internet Services

- Image/Video classification
- Speech recognition
- Natural language processing

Medicine



- Cancer cell detection
- Diabetic grading
- Drug discovery

Media & Entertainment

- Video captioning
- Content based search
- Real time translation

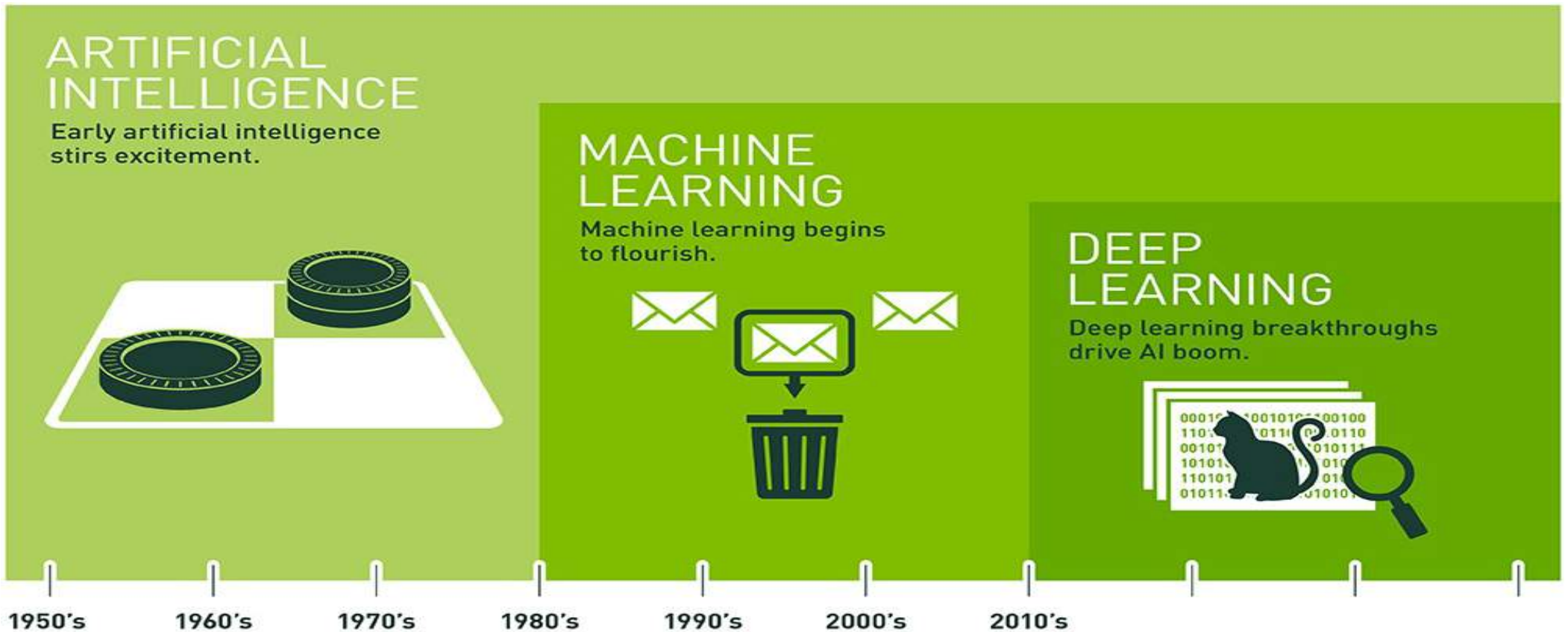
Security & Defense

- Face recognition
- Video surveillance
- Cyber security

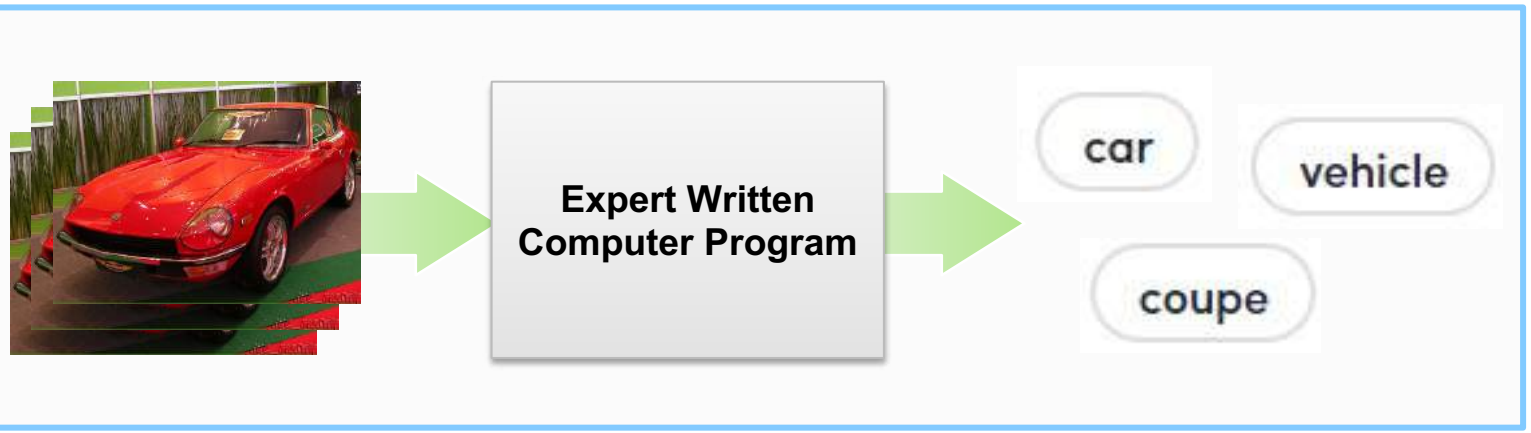
Autonomous Machines

- Pedestrian detection
- Lane tracking
- Recognize traffic signs

AI And Deep Learning

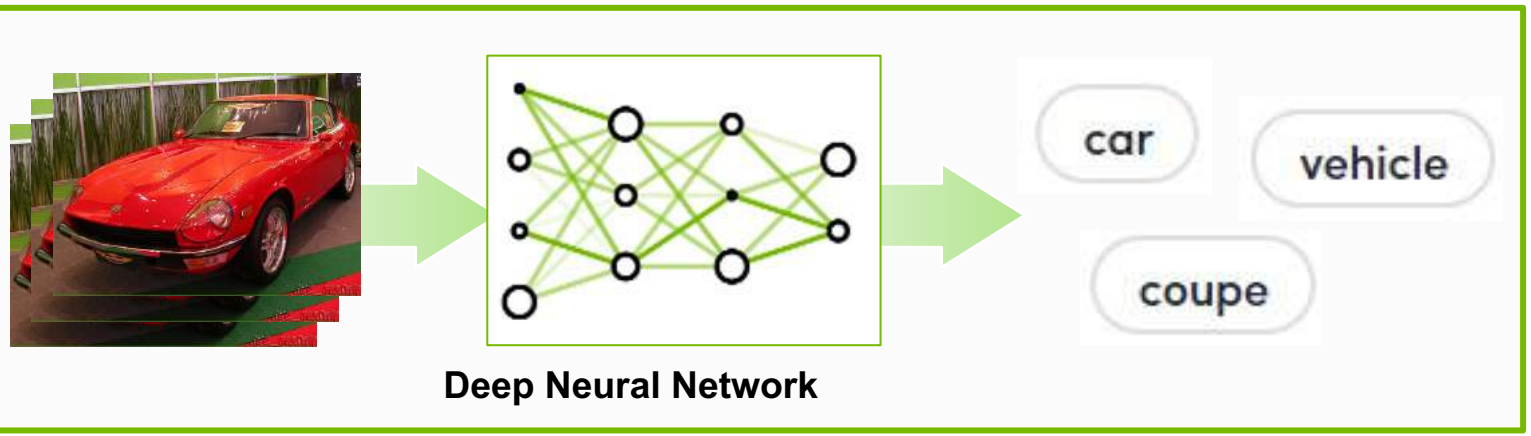


Algorithms that Learn from Examples



Traditional Approach

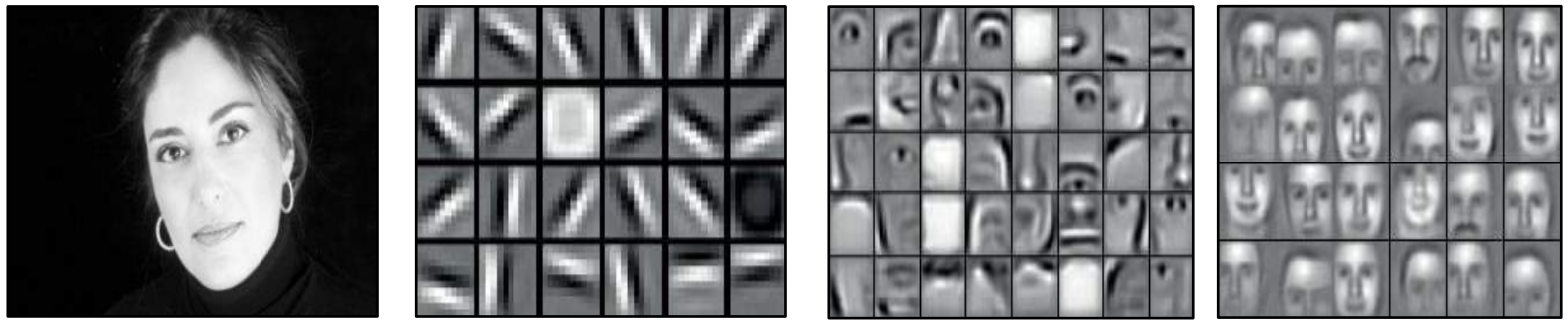
- Requires domain experts
- Time consuming
- Error prone
- Not scalable to new problems



Deep Learning Approach

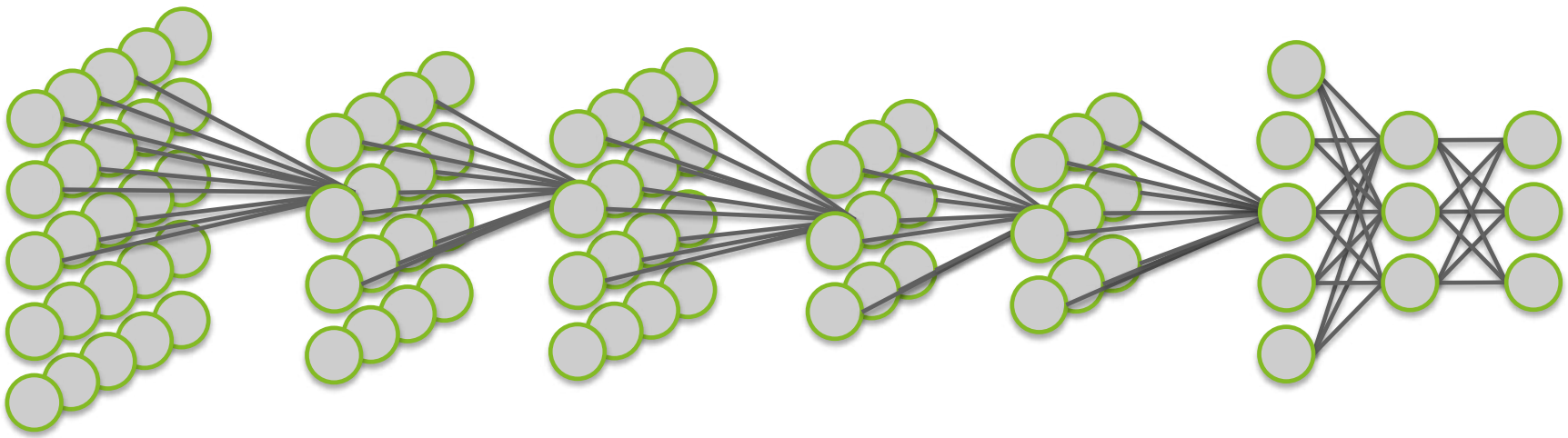
- ✓ Learn from data
- ✓ Easily to extend
- ✓ Speedup with GPUs

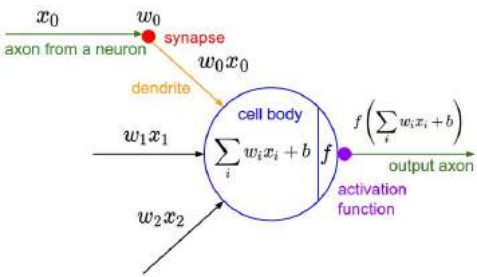
What Is Deep Learning?

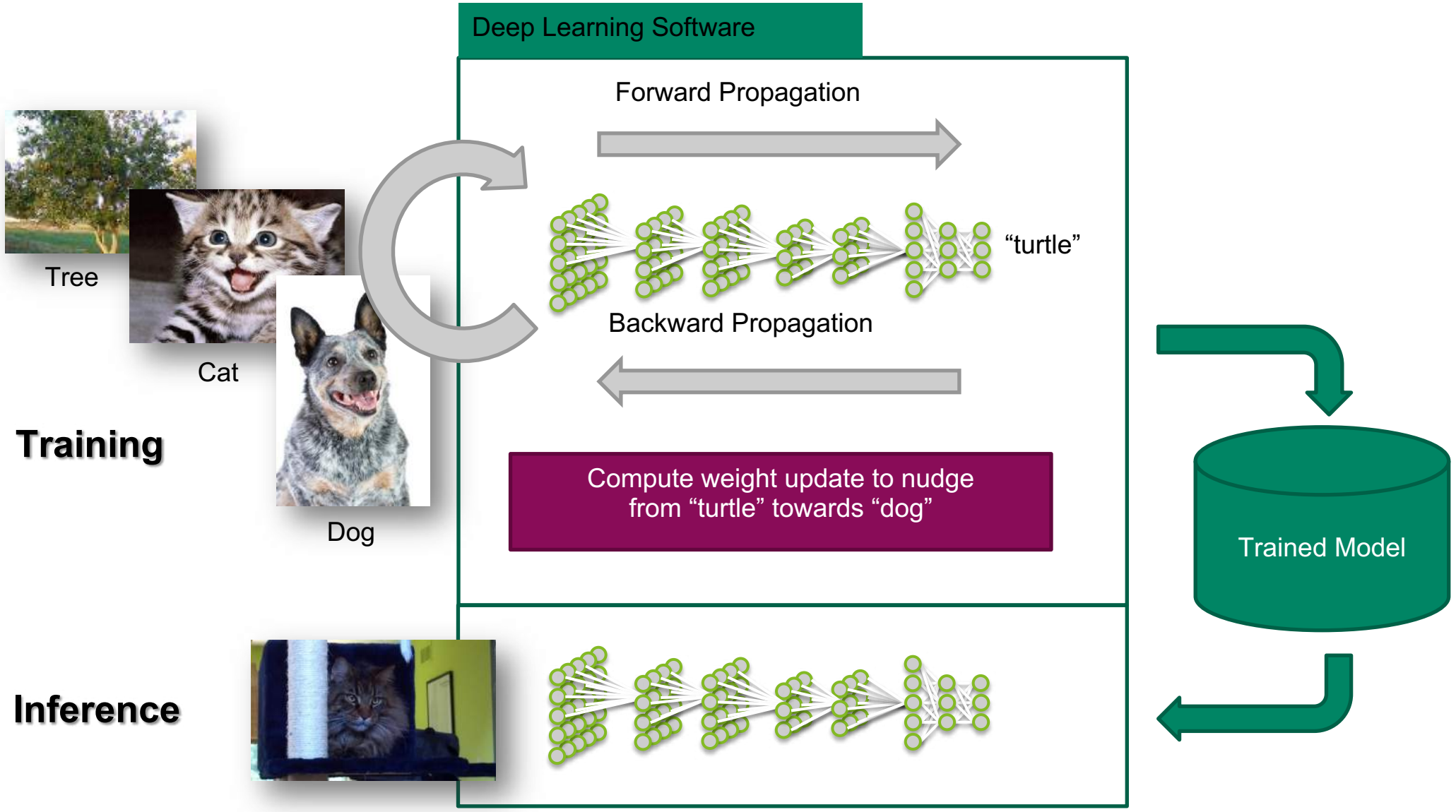


Typical Network

Task objective
e.g. identify face
Training data
10-100M images
Network architecture
10 layers
1B parameters






Deep Learning Software
Forward Propagation
"turtle"
Backward Propagation
Compute weight update to nudge from "turtle" towards "dog"
Tree
Cat
Dog
Training
Inference
Trained Model

Deep Learning for Computer Vision

IMAGE CLASSIFICATION

98% Dog

2% Cat

Classify images into classes or categories

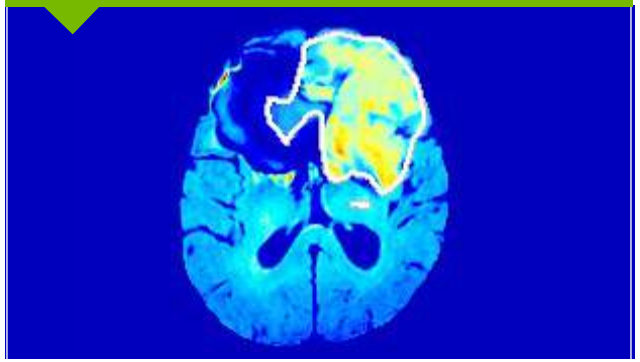
Object of interest could be anywhere in the image

OBJECT DETECTION

Find instances of objects in an image

Objects are identified with bounding boxes

IMAGE SEGMENTATION

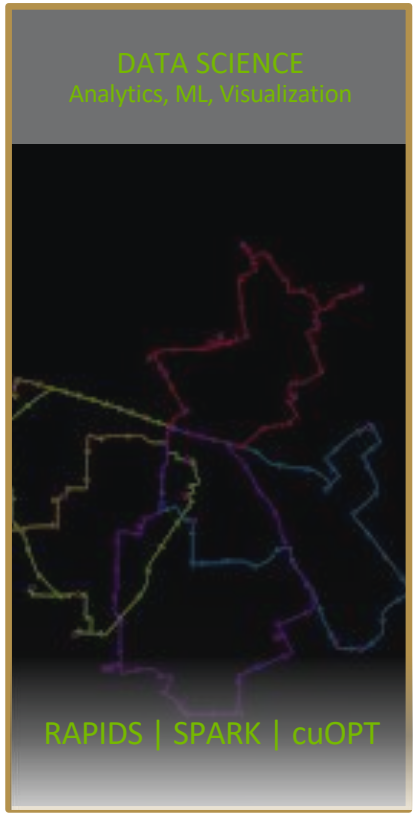


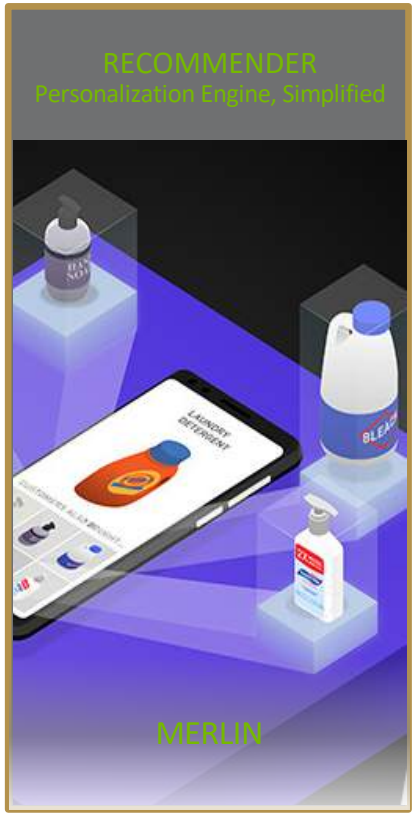
Partition image into multiple regions

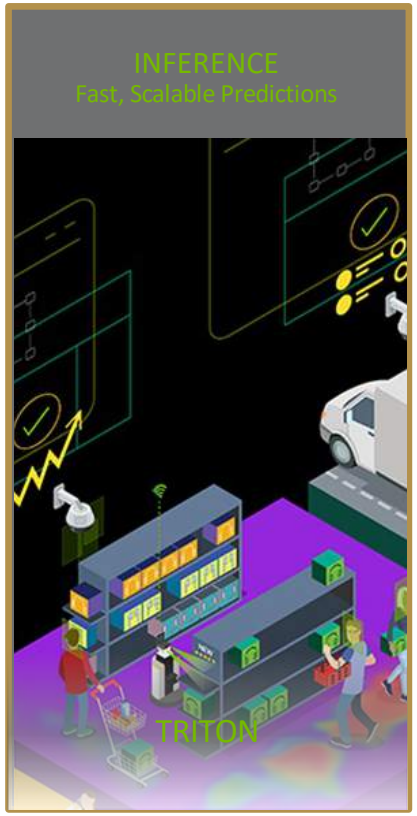
Regions are classified at the pixel level

Accelerating the Next Wave of AI

NVIDIA AI Platform Updates

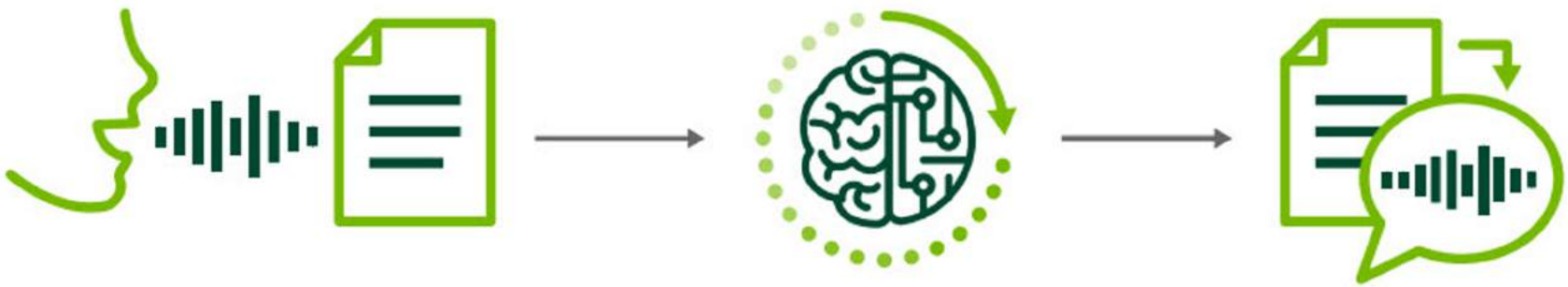






https://www.nvidia.com/gtc/

What is Conversational AI

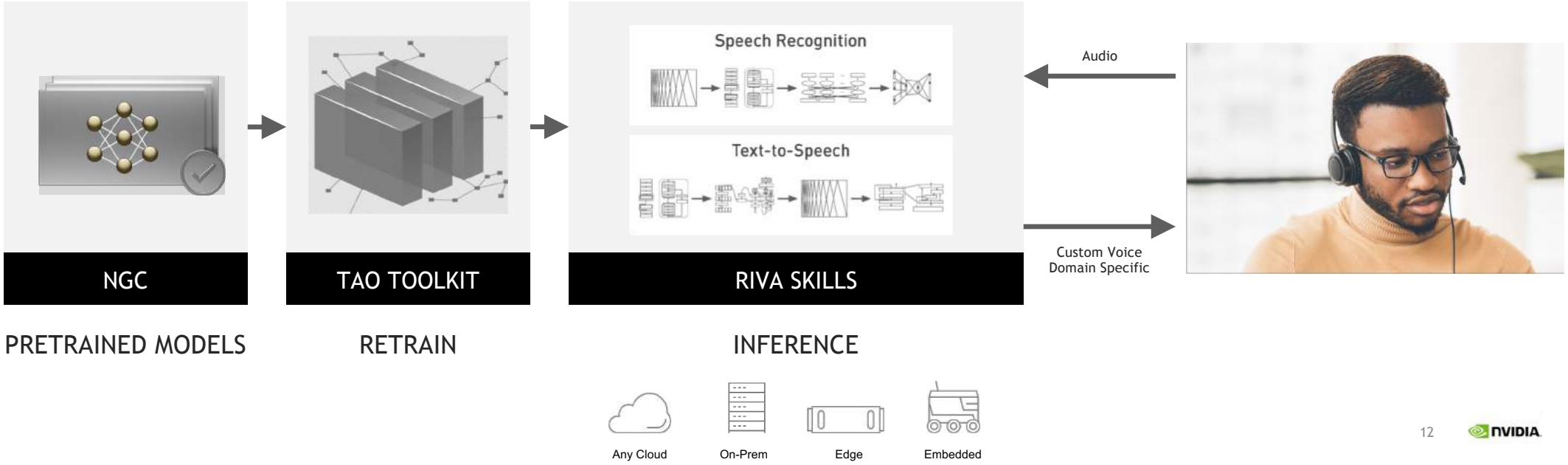


https://developer.nvidia.com/conversational-ai

NVIDIA RIVA

GPU-Accelerated SDK for Speech AI

- World Class Speech Recognition and Text-to-Speech Skills
- Pre-trained SOTA models trained on 100,000 hours of DGX; Retraining with TAO toolkit (zero coding)
- Flexible customization from data to model to pipeline
- Deploy Services with one line of code in cloud, on-prem & edge; Support for large-scale speech AI deployments
- Scale to handle hundreds and thousands of real-time streams with <300 ms latency per stream

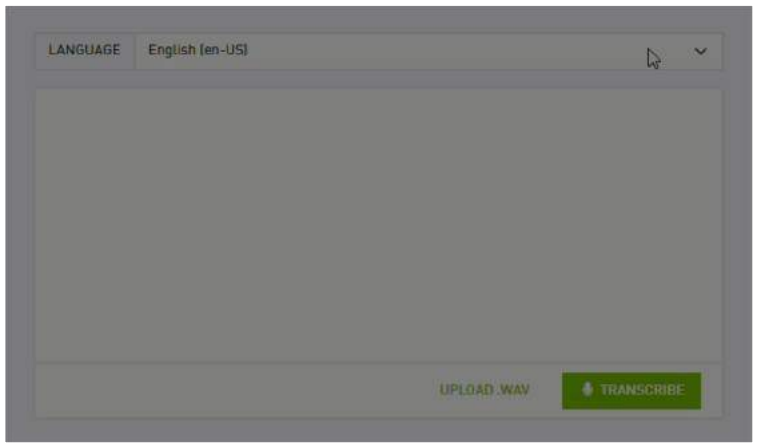


WHAT IS NEW IN NVIDIA RIVA

ASR in multiple languages and customizable TTS pipelines

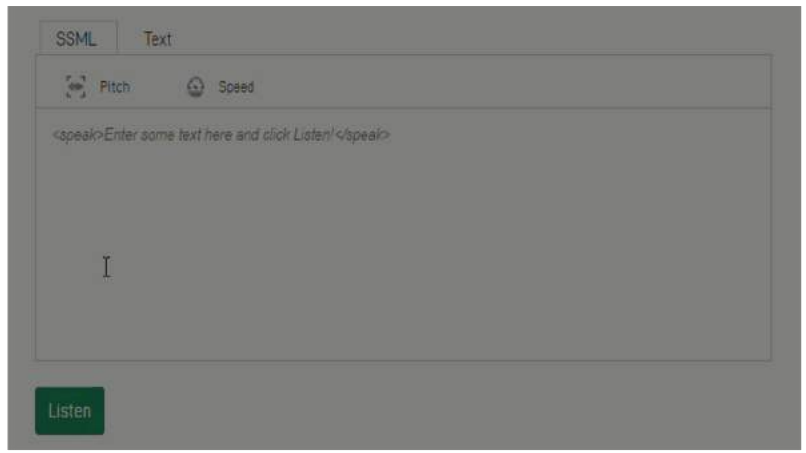
Automatic Speech Recognition

- World Class Automatic Speech Recognition (ASR):
  - English
  - Spanish
  - German
  - Russian
- Fine-tune for industry-specific jargon, dialects, and noisy environments using TAO Toolkit



Text-To-Speech

- Expressive human-like Text-to-Speech (TTS) voices with customizable:
  - Pitch
  - Speed
- Several times faster TTS pipelines with the latest SOTA models such as Fastpitch



PRE-TRAINED SPEECH AI MODELS

Accurate State-Of-The-Art Models In NGC

Several speech and language pretrained models in NGC to get started

- SOTA models trained over 100,000 hours on NVIDIA DGX™
- Optimized for high-performance training and inference on GPUs
- Customizable with NeMo, fine-tunable with TAO Toolkit, deployable to Riva
- Used across apps such as chatbots, virtual assistants, & transcription services

Automatic Speech Recognition (ASR)

Jasper

Quartznet

Citrinet

Acoustic Model

BERT NER

BERT Punctuation

Post-Processing

Text-To-Speech (TTS)

Fastpitch

Tacotron

Spectrogram Generator

HiFiGAN

WaveGlow

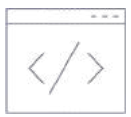
Vocoder

NVIDIA Triton

Open-source Inference Serving Software For Fast, Easy Inference Deployment

Fast And Scalable AI In Every App

Any Model



Any Framework

Any Query Type

Any Processor

Any Deployment

Any Deployment Location

Key Feature Updates

Shapley Values in FIL Backend

Explanation of model prediction

Triton Management Service

Efficient scaling in Kubernetes

Model Navigator

Accelerated time to production

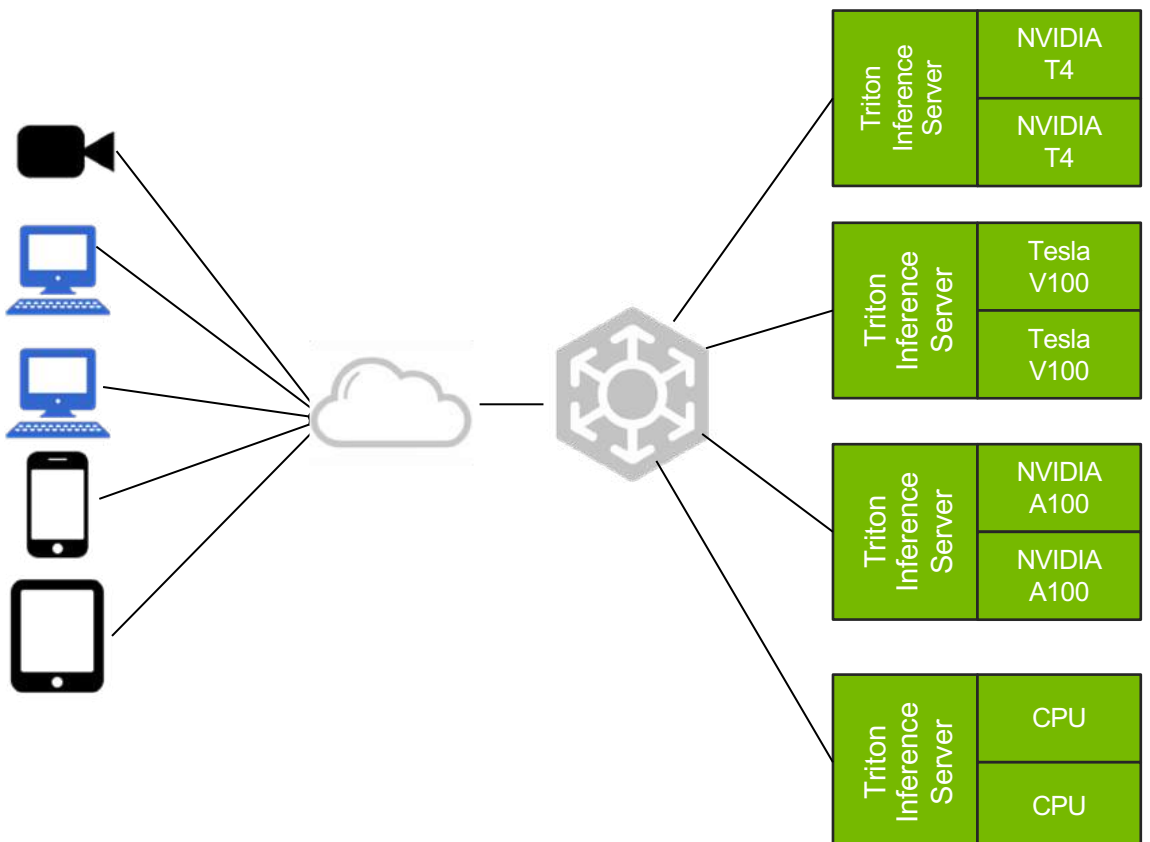
1000's Of Users | 1.3M+ Downloads/Clones

amazon

Microsoft

NVIDIA Triton Inference Server

Production Inference Server on GPU and CPU



Maximize real-time inference performance of CPUs and GPUs

Quickly deploy and manage multiple models per GPU per node

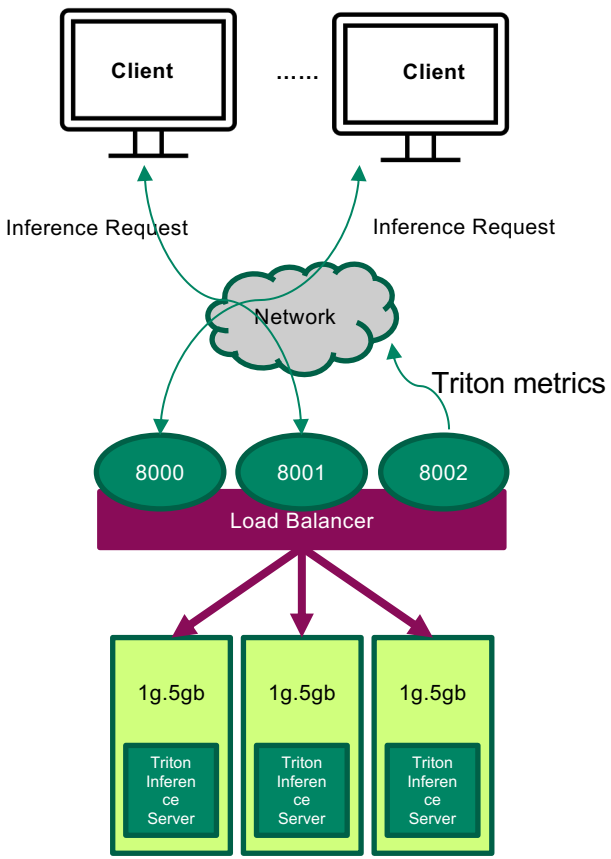
Easily scale to heterogeneous GPUs and multi-GPU nodes

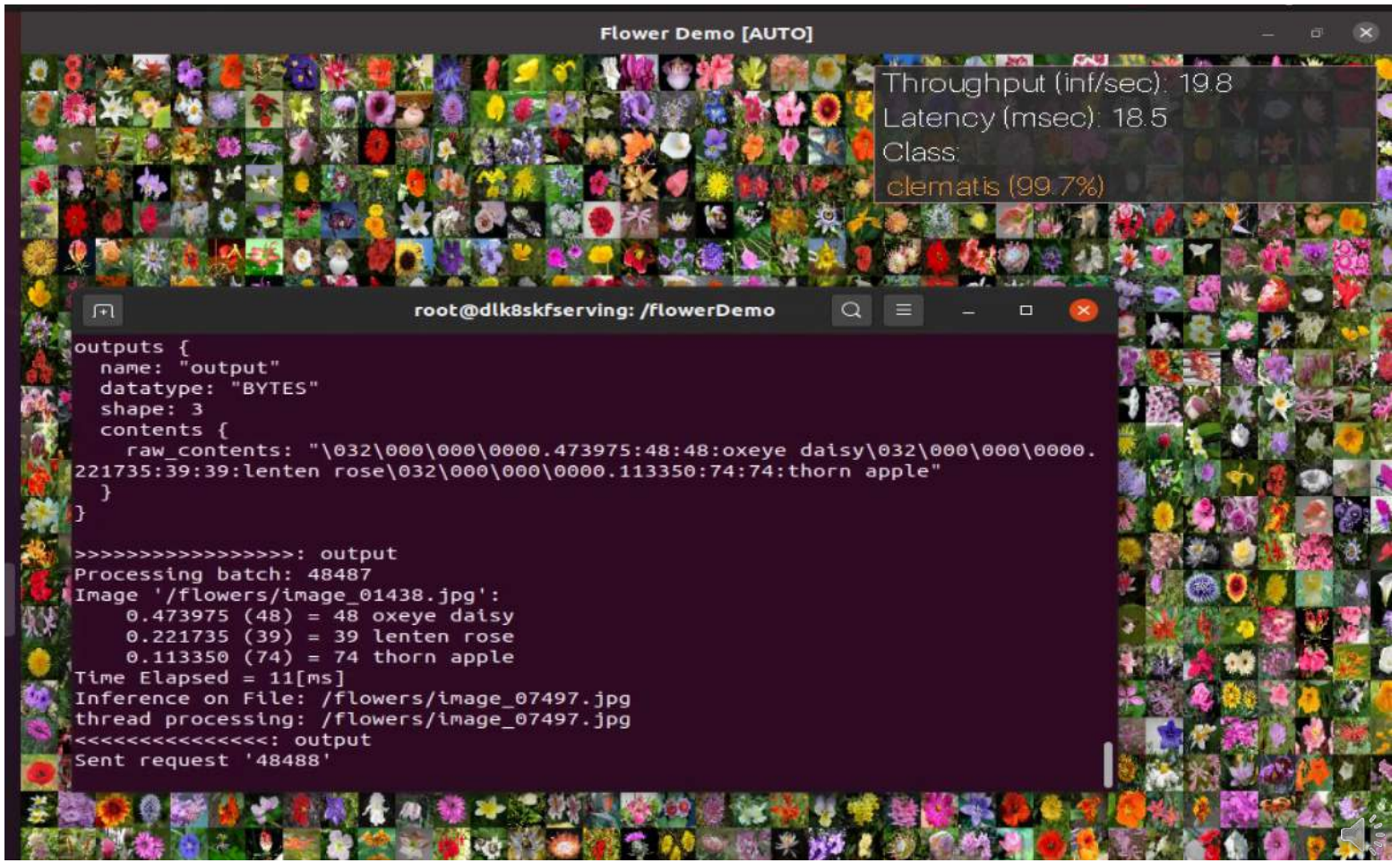
Integrates with orchestration systems and auto scalers via latency and health metrics

Open source for seamless customization and integration

Autoscale Triton Deployment with K8s

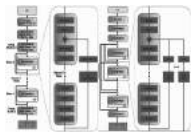
Clients Send Inference Requests to Triton Inference Servers

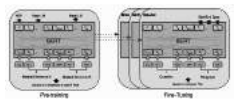


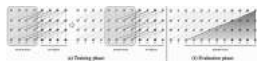


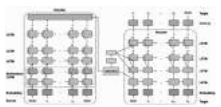
Deep Learning Examples Overview

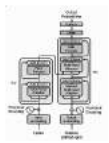
Natural Language Processing


Jasper


BERT

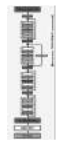

Transformer-XL

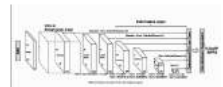

GNMT

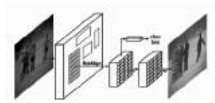

Transformer

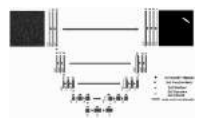

Tacotron 2 & WaveGlow

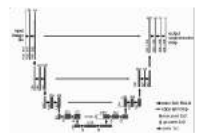
Computer Vision

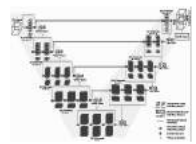

ResNet


SSD

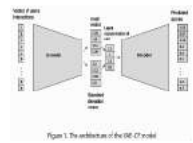

Mask R-CNN

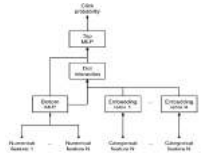

U-Net Industrial

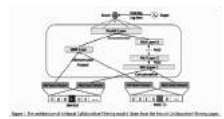

U-Net Medical

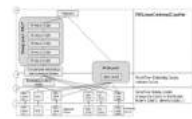

V-Net Medical

Recommender Systems


VA-NCF


DLRM


NCF


Wide & Deep

Who: For Data Scientist and Software Engineers

What: Train, fine-tune and deploy State-of-the-Art Models in production, achieving the highest throughput, and lowest latency

Why: Show GPU Performance Across a Variety of Popular Deep Learning Workloads

How: Using NVIDIA's Deep Learning Software Stack

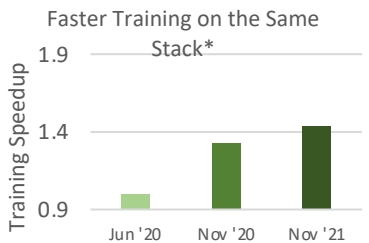
Goal: Increases Awareness and Adoption of NVIDIA's Deep Learning Software and Hardware

GitHub:https://github.com/NVIDIA/DeepLearningExamples

NGC CATALOG

The Hub of GPU-Optimized Software

PERFORMANCE OPTIMIZED
Tested across GPU-accelerated Platforms



Monthly sw container updates

SOTA models

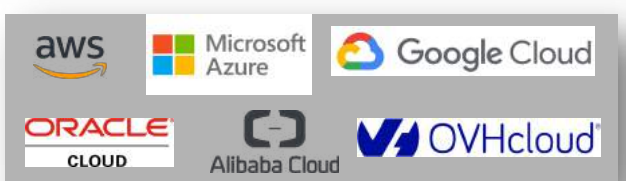
FULLY TRANSPARENT
Quickly identify and deploy the right software

Detailed security scan reports

Model resumes

ACCELERATES DEVELOPMENT
Focus on building, not setup

One click deploy from NGC



Develop once. Deploy anywhere w/ NVIDIA VMI

1.5M+ Users | Millions of Downloads

*8x NVIDIA A100 40GB. NVIDIA DGX. Resnet-50. Mixed Precision. 256 Batch size.

NGC PRE-TRAINED MODELS AND MODEL SCRIPTS

Build AI Faster

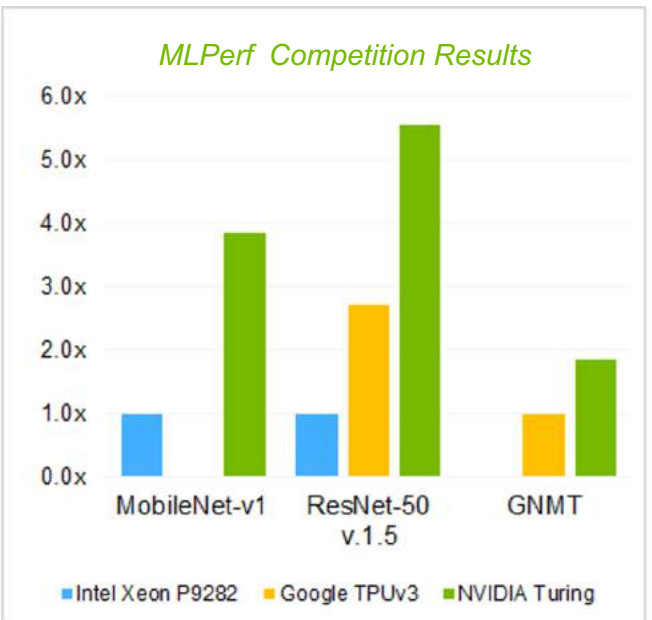
PRE-TRAINED MODELS

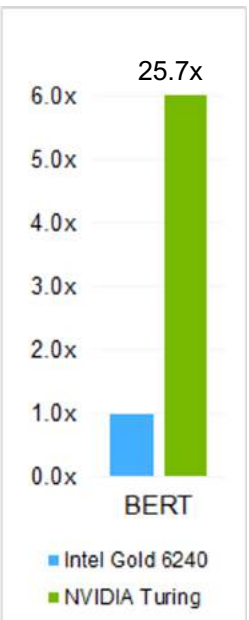
- Popular AI tasks – ASR, NLU, TTS, RecSys, CV, etc
- Industry specific models – Medical imaging, public safety
- Customize with your data and transfer learning
- Integrate into existing workflows with SDKs

MODEL SCRIPTS

- Reference neural network architectures across all domains and popular frameworks with latest SOTA
- Code samples show you how to deploy or build your models

FASTEST INFERENCE





MLPerf v0.5 Inference Closed; Retrieved from www.mlperf.org 6 November 2019. Per-accelerator performance derived from the best MLPerf results for respective submissions using reported accelerator count. MLPerf name and logo are trademarks. See www.mlperf.org for more information.
BERT performance measured by NVIDIA with OpenVINO (2020.2) on Gold 6240 vs. NVIDIA T4

Applied Research Accelerator Program

Apply for support for projects that have the potential to make real world impact using GPU-accelerated applications

Who can apply	What we look for	Advantages of the grant program	Scan barcode to apply
Academic faculty members and PhD students	Feasibility of research has been demonstrated Advantage of using GPU acceleration clearly defined Plan to adopt the technology into product or service by industry or government partner	Get access to NVIDIA hardware + funding Get technical guidance from experts at Nvidia NVIDIA does not hold/share any patent rights	

Awards are announced in March, June, September, and December. Applications are accepted year-round.

https://www.nvidia.com/en-au/industries/higher-education-research/applied-research-program/

NVIDIA INCEPTION

A Free Worldwide Program for Startups in Every Industry

10,000+
Startups

59%
Growth in 2021

$74B
Cumulative Funding

100+
Countries Represented

Overview of Program Benefits

EXPERTISE

Free Training Credits
SDK Recommender

TECHNOLOGY

Preferred Pricing
Cloud Credits

VENTURE

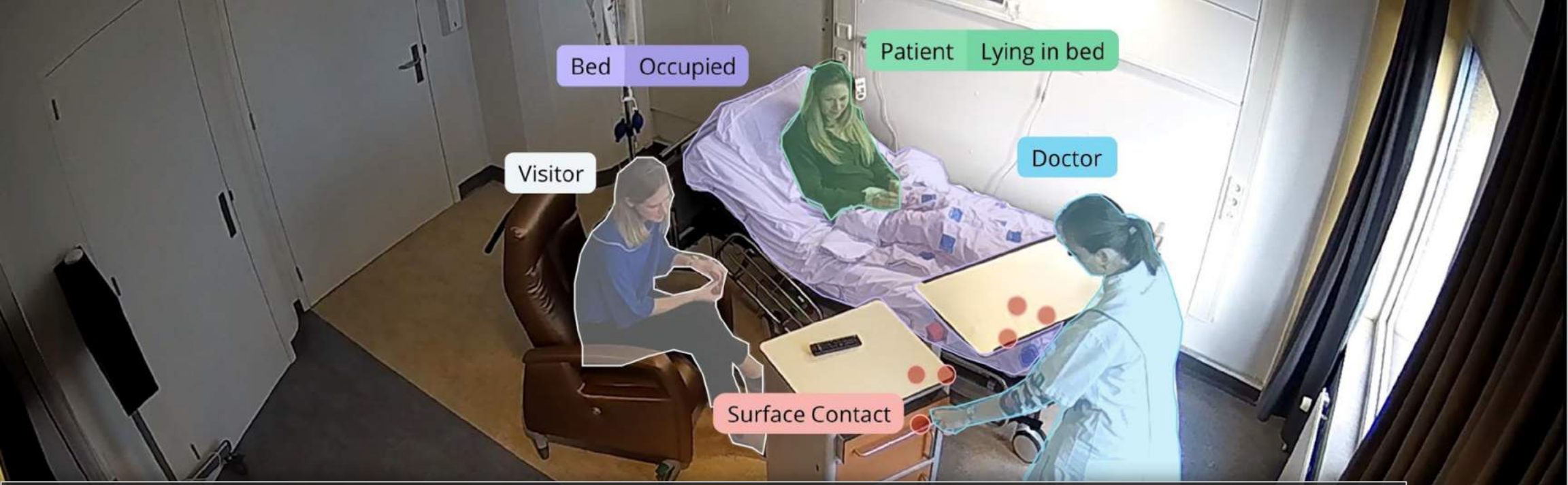
Introductions to VCs
Exclusive Connect Events

AWARENESS

Speaking Opportunities
Member Success Stories

APPLY NOW:

www.nvidia.com/inception



CONTACTLESS CARE PLATFORM FOR SMART HOSPITALS

In the US alone, 9,300+ healthcare workers contracted COVID-19 from exposure to patients. Yet routine monitoring and guidance still depend on in-person processes.

Ouva offers a visual sensing platform for remote patient care and touch-free assistance that runs on NVIDIA Jetson and NVIDIA Clara Guardian.

With Ouva, hospital staff can centrally monitor patient safety, room hygiene, and bed turnover in real-time. Ouva's voice-activated, touch-free screens guide patients and visitors to appointments, reducing contact with front-desk staff to only the most urgent needs.

Ouva has customers in the US, Europe, and Australia and plans to expand in the coming year to neurology wards, infection units and more.

NVIDIA. OUVA

www.NVIDIA.com/Inception

VISUAL INTELLIGENCE TO SAFEGAURD PATIENTS AND STAFF

There are 1.7M annual hospital-acquired infections in the US alone, of which 99K die every year.

Darvis, a visual intelligence platform powered by NVIDIA Clara Guardian, provides real-time, privacy-first solutions for everyday tasks including rapid hygiene assessment, bed logistics, and inventory tracking – creating a foundation for true hospital automation.

Within a month Darvis developed a COVID-specific AI model to help keep hospital patients and staff safe during the pandemic.

NVIDIA DARVIS

www.NVIDIA.com/Inception

> “ So we are now able to throw hundreds of millions of data [points] into our training properties and we are seeing a clear delta. ”
>
> – Hui Wang, VP of Data Science, PayPal

PAYPAL USES AI TO CRACK DOWN ON FRAUD AND PROTECT CUSTOMERS

Challenge

PayPal processes millions of transactions every day.

Need 24/7 operations and real-time protection of customer transactions.

Solution

Leveraged a plug-and-play approach to AI with DGX systems to pinpoint fraudulent transactions.

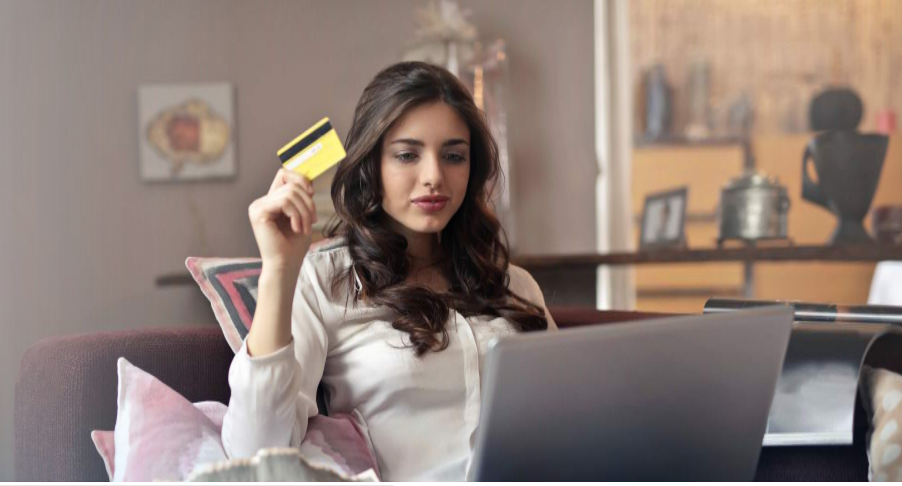
Expanded AI to other use cases, including using chatbots and personalization of user experiences.

- AIRI from Pure Storage
- NVIDIA DGX Systems
- NVIDIA RAPIDS Software

Reduction in training time from **days to hours**

"Especially in this environment, our customers need us now more than ever, so we're supporting them with best-in-class fraud protection and servicing"

- Manish Gupta, VP of Machine Learning and Data Science Research,
- American Express



AMERICAN EXPRESS USES AI TO FOIL CYBERCRIME

Challenge

AmEx handles more than 8 billion transactions per year.

The spike in online transactions since the COVID-19 pandemic led to complex fraud attacks driving the need to better monitor transactions in real-time.

Solution

Used AI to determine anomalies in transactions using recurrent neural networks (RNNs) and long short-term memory networks (LSTMs).

Using DGX, AmEx built and trained models on mountains of data, resulting in improved accuracy.

Easily scaled AI into other areas like recommendation engines for personalized offers for card holders, forecasting customer default rates, and assigning credit limits.

- NVIDIA DGX Systems *for Training*
- NVIDIA T4 GPUs *for Inference*
- NVIDIA Triton Inference Server

50x Faster training on DGX vs CPU server

6% Improvement in fraud detection accuracy

<2ms Latency in detection

> “Our collaboration with NVIDIA allows us to develop the future of factory logistics today..”
>
> — Jürgen Maidl, Senior Vice President of Logistics, The BMW Group

BMW USES AI TO CREATE HIGHLY CUSTOMIZABLE, JUST-IN-TIME MANUFACTURING

Challenge

BMW receives almost 10,000 new car orders a day, with 100 different options per car and 2,100 possible combinations.

30 million raw parts a day come in and must be organized into custom parts trays — for every order.

Solution

Leveraged AI-powered logistics robots - from transporting materials to organizing parts.

Trained deep neural networks on NVIDIA DGX systems in a simulated 3D virtual world.

DGX systems enables testing of a near-infinite range of scenarios, for highest levels of accuracy and synchronization of robots across the production floor.

- NVIDIA DGX Systems & DGX Station *for Training*
- Isaac Simulation Technology
- NVIDIA Quadro GPUs *for Ray-Tracing*
- NVIDIA Jetson AGX Xavier
- NVIDIA EGX

56 seconds

A new car outputs every 56 seconds

> *“The experts at NVIDIA keep up with cutting-edge tools and methods and have a very good grasp of how companies are using their products”*
>
> — Zack Fragoso, Manager, Data Science and AI, Domino’s

DOMINO’S BOOSTED PREDICTIONS FOR ORDER ACCURACY READINESS FROM 75% TO 95%

Challenge

Domino's Pizza delivers more than 3 billion pizzas a year.

Wanted to leverage massive amounts of data to improve operational efficiencies and customer experience.

Solution

"Points for Pie" campaign used AI to classify pizza images, driving significant press and customer engagement.

Sharing DGX resources across other departments for AI use cases like more efficient routing of delivery orders and determining marketing windows for coupons.

NVIDIA DGX Systems

NVIDIA RAPIDS Software

72x

Speed up in training time, from 3 days to less than an hour

WORLD CLASS SPEECH AI FOR THE BEST VIDEO CONFERENCING EXPERIENCE

With hundreds of millions of online daily meetings, video conferencing has become instrumental for enabling employees to connect, collaborate and be productive.

RingCentral, a leading provider of UCaaS solutions, built a highly accurate, scalable, and reliable solution capable of serving over a billion minutes per month of transcription and more than 200,000 concurrent users on their platform.

RingCentral deploys NVIDIA speech AI state-of-the-art pre-trained models and fine-tuned them on proprietary custom data with NVIDIA NeMo - an open-source framework for building, training, and fine-tuning conversational AI models.

With NVIDIA Speech AI, the RingCentral team achieved impressive accuracy for customers with worldwide accents and different domain-specific vocabularies, reducing the word error rate (WER) by over 10%.

NVIDIA RingCentral

HOW AI COULD HELP ALEXA KNOW WHEN TO JOIN THE CONVERSATION

Voice-controlled personal assistants are challenged to perform speech recognition amid interfering background speech, but a team of researchers from Amazon and John's Hopkins University are teaching Alexa to ignore speech that's not intended to "wake" it.

Using data sets of 1,200 hours of speech and NVIDIA GPUs for AI training and inference, the team improved Alexa's speech recognition by 15%.

NVIDIA.

BREAKING NEW GROUND IN SPEECH

When you ask your phone a question, you want the right answer, right now. This requires an AI-powered service with object detection, question-answer, and text-to-speech performing a variety of predictions and responding in under one second.

When its CPU-only servers couldn't perform the computational work within the required latency, Bing switched to NVIDIA GPUs. Now Bing generates 1 second of speech in 50 ms — a 100x improvement.

NVIDIA

Bing