AI and Accelerated Computing in Insurance

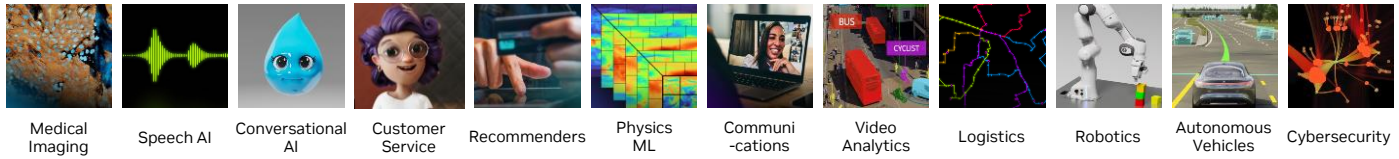
Dr. Jochen Papenbrock
Head of Financial Technology EMEA / Lead Developer Relations Manager Banking Global

jpapenbrock@nvidia.com

NVIDIA AI Enterprise With AI models and Foundries

End-to-end open platform for production generative AI – full-stack solution

AI Workflows, Frameworks and Pretrained Models*



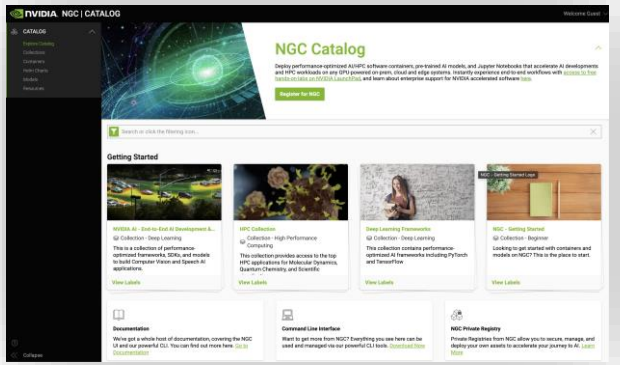
AI and Data Science Development and Deployment Tools

Cloud Native Management and Orchestration

Infrastructure Optimization

Accelerated Infrastructure

Cloud | Data Center | Edge | Embedded



- AI Frameworks and Pretrained Models
- Reduce OSS development complexity
- Secure and Scalable
- Optimized for Production AI
- Certified to Run Everywhere
- broad partner ecosystem
- Enterprise-Class Support
- Technical Support and Services
- Flexible AI Infrastructure

*NVIDIA NGC public catalog provides a complete listing of over 50 supported frameworks and pretrained models.

NVIDIA Full Stack AI Factory for Financial Services

BUSINESS DRIVERS

- Fraud Detection
- Virtual Assistants
- Cross-Sell
- Next Best Action
- Credit Default Prediction
- Compliance Automation

AI INFRASTRUCTURE CHALLENGES

- AI Infrastructure Complexity
- Scalability & Productivity
- Total Cost of Ownership

AI PLATFORM FOR FINANCIAL SERVICES

- AI FINANCIAL APPLICATIONS
- DATA SCIENCE
- DEV OPS
- HARDWARE

NGC

ALL END USERS

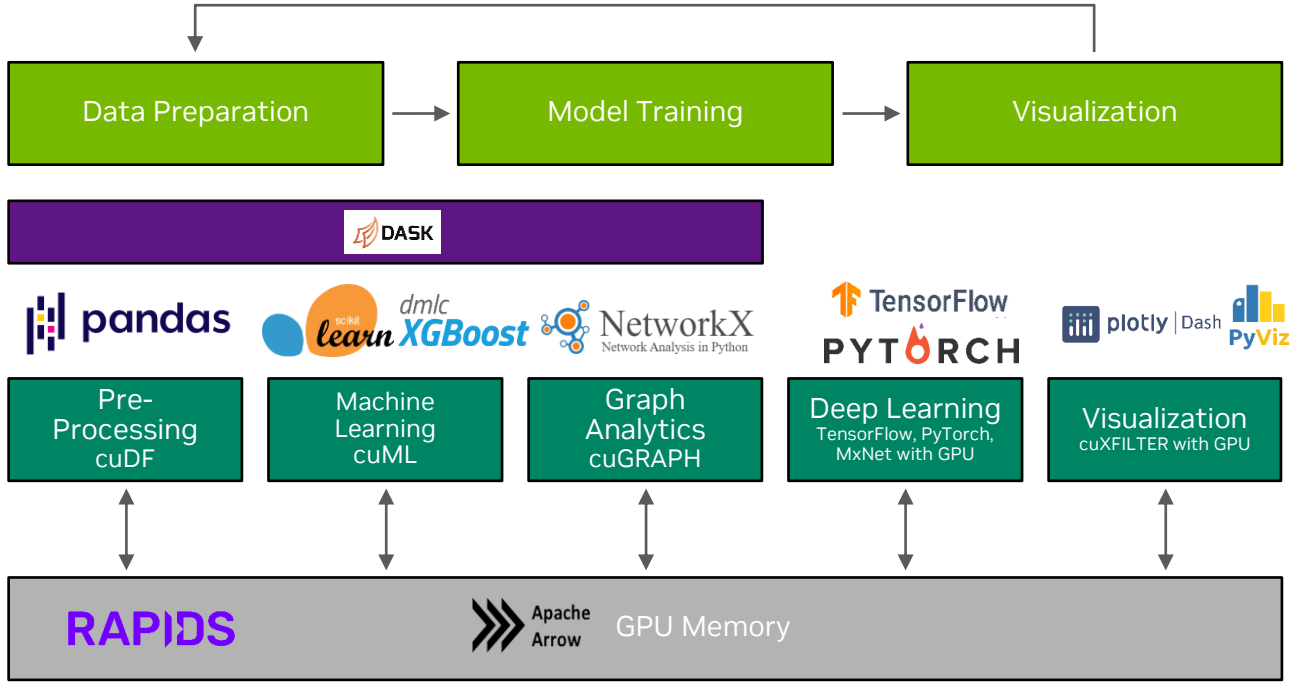
- Data Scientist / Researcher
- System Admin/DevOps
- Centers of Excellence
- Lines of Business
- Marketing, Finance, etc.

OUTCOMES

- Time to Insight
- Productivity at Scale
- Return on Investment

End-to-End Accelerated Data Science

Open-source suite of GPU-accelerated Python libraries designed to improve your data science and analytics pipelines

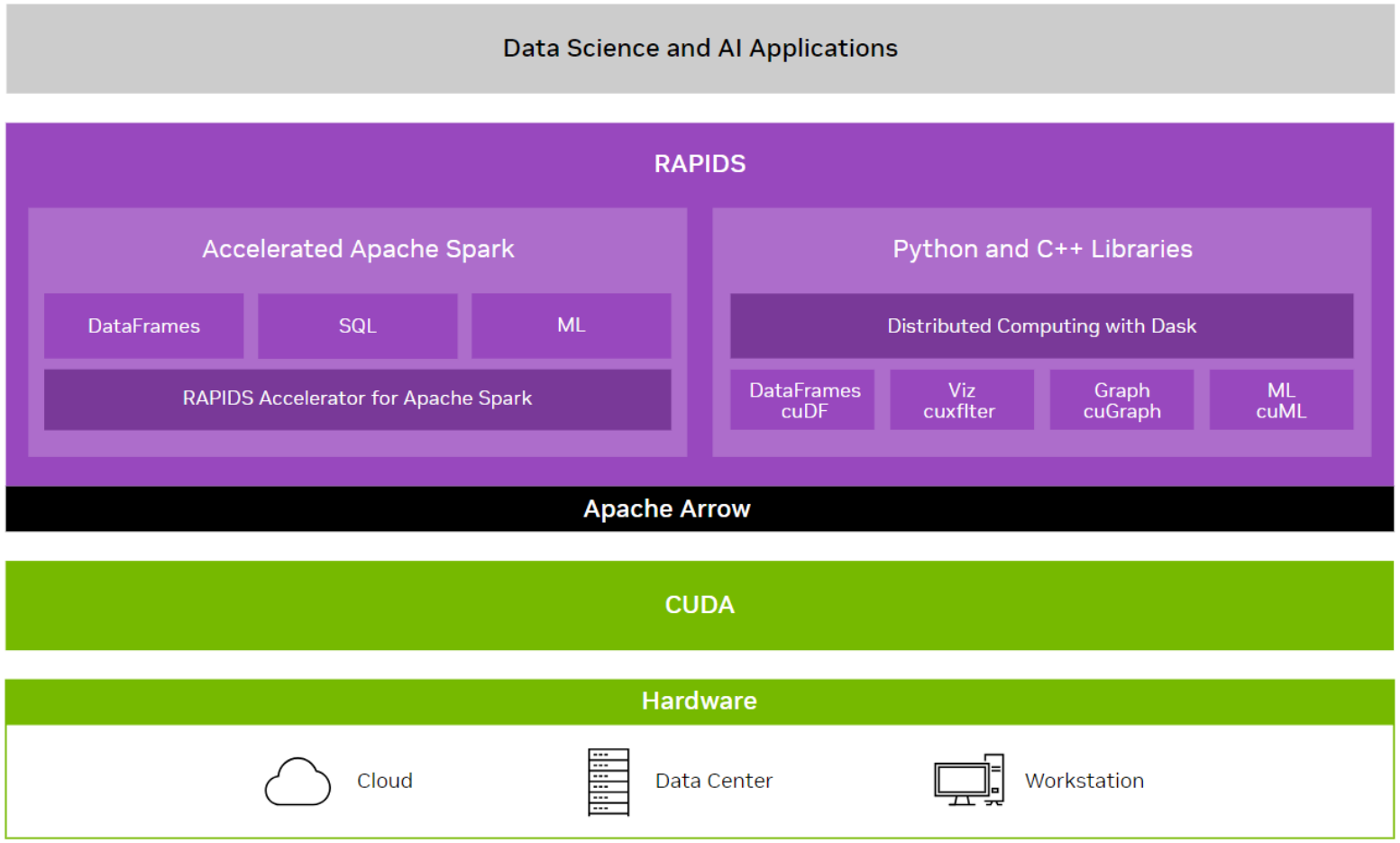


Accelerate by
5x
Lower costs by
4X

RAPIDS APACHE Spark

End-to-End Accelerated Data Processing and Data Science

RAPIDS is an open-source suite of GPU-accelerated data science and AI libraries with APIs that match the most popular open-source data tools.



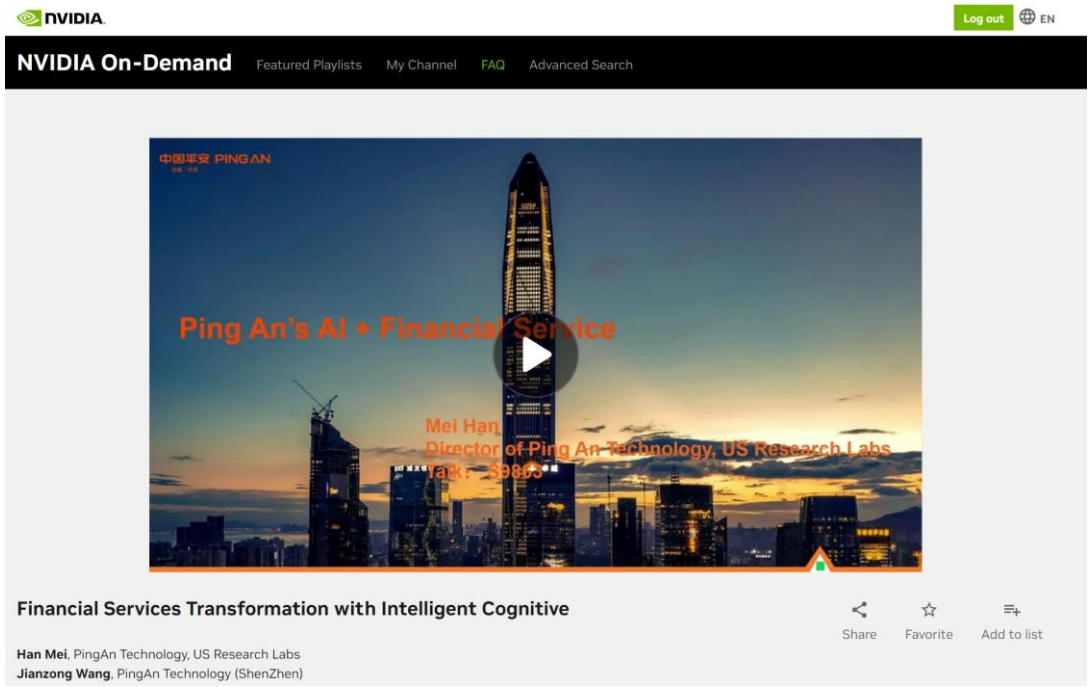
NVIDIA GTC

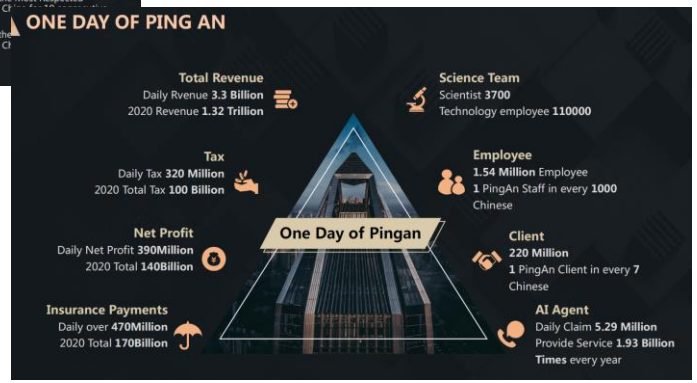
Financial Services Session

Ilay Chen
PayPal

How PayPal Reduced Cloud Costs by up to 70% With Spark RAPIDS

Use Cases Ping An





- insurance claims solution: automatic vehicle picture estimation and anti-fraud detection
- reduced **processing time for claims** from days to seconds, increased operating efficiency, and cut billions in operation costs
- **anti-fraud and risk-estimation** models on GPUs, reducing model training time from weeks to hours RAPIDS
- GPU-Accelerated graph analytics to identify **suspicious transactions**

FASTER TIME TO INSIGHTS

Insurance giant Ping An has nearly 180 million customers so its data science team relies on AI to gain insights on issues ranging from fraud detection to predicting disease.

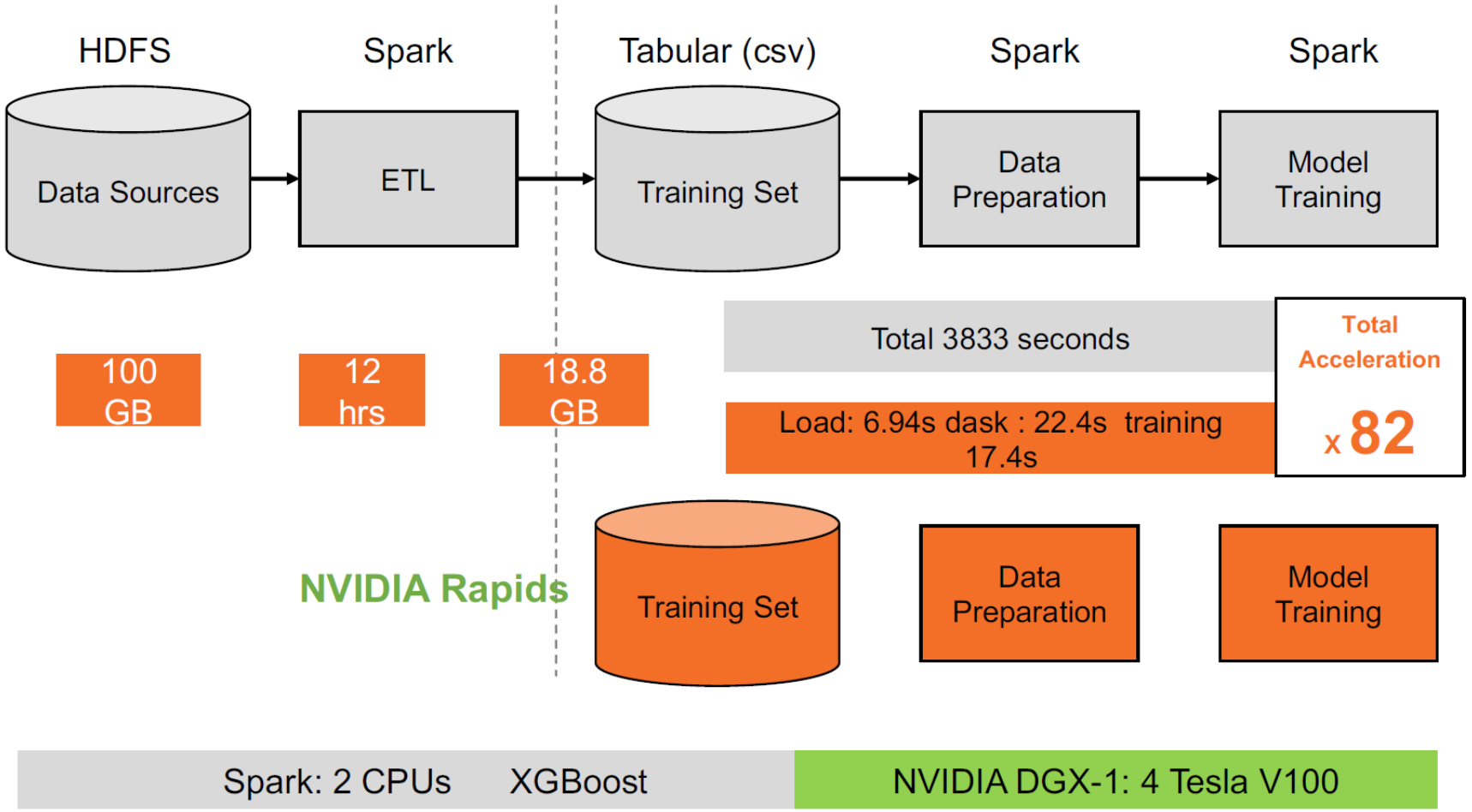
Ping An recently tested RAPIDS and ran data science pipelines on GPUs.

The team achieved speedups of 27x-80x in dataset processing time which could help them develop proactive predictions and improve prevention plans.

NVIDIA

中国平安 PING AN 保险 · 银行 · 投资

Public Health Disease Prediction



Comparison of Machine Learning Algorithm XGBoost between Spark on CPU and Rapids on GPU

Before:

We deploy disease prediction algorithms like XGBoost on CPU cluster servers using Spark platform.

Now:

With the support of Rapids, GPU can run XGBoost with a faster loading and training which can help iterate the prediction model for better performance.

Model iteration time:

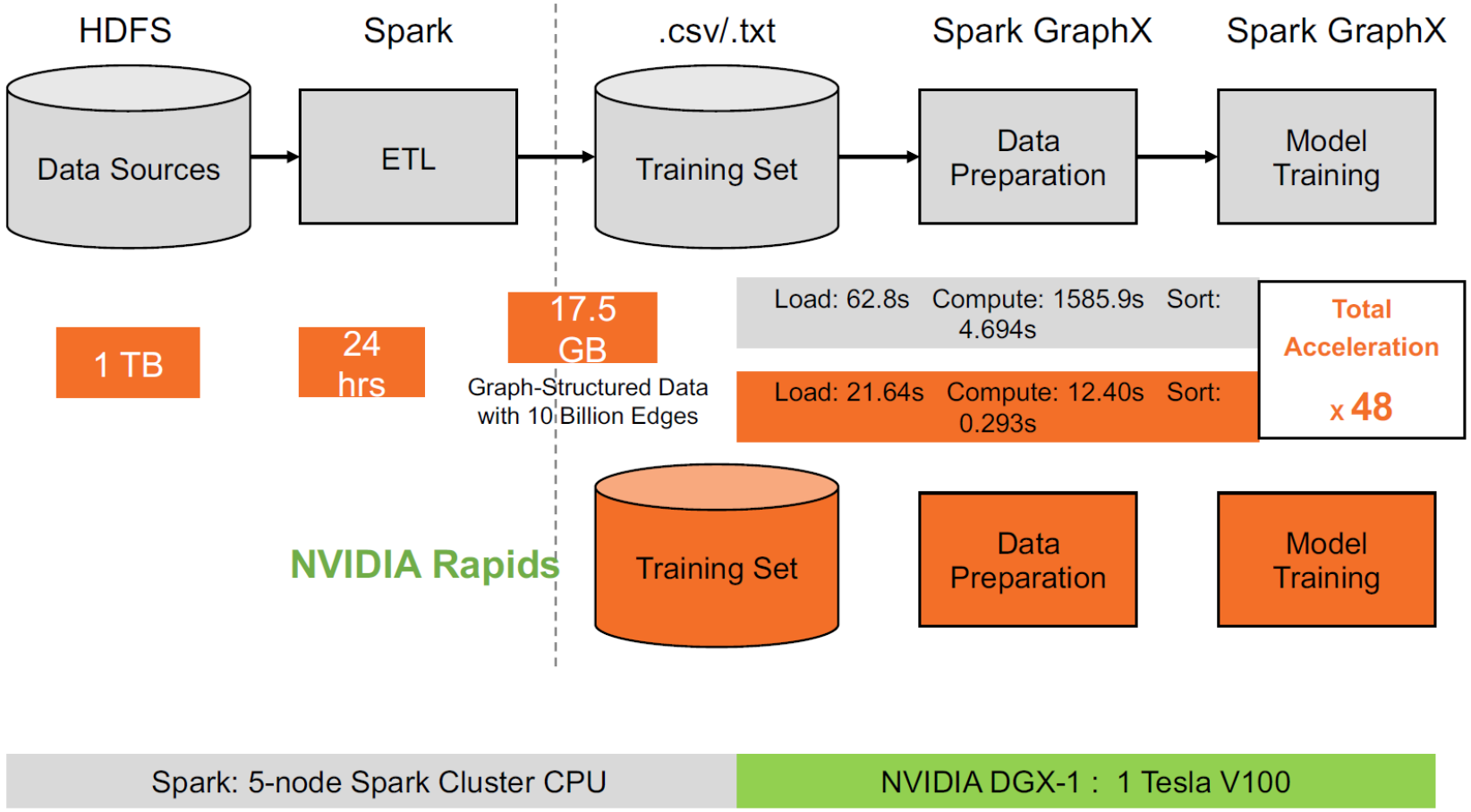
The model iteration time can be reduced from weeks to hours by implementing algorithms on Rapids instead of Spark.

Prescription Anti-fraud Using cuGraph

RAPIDS

NetworkX
Network Analysis in Python

The whole progress of PageRank



Comparison of Graph Algorithm PageRank between GraphX on CPU and Rapids on GPU

Before:

We deploy fraud detection algorithms like PageRank on CPU cluster servers using Spark GraphX platform.

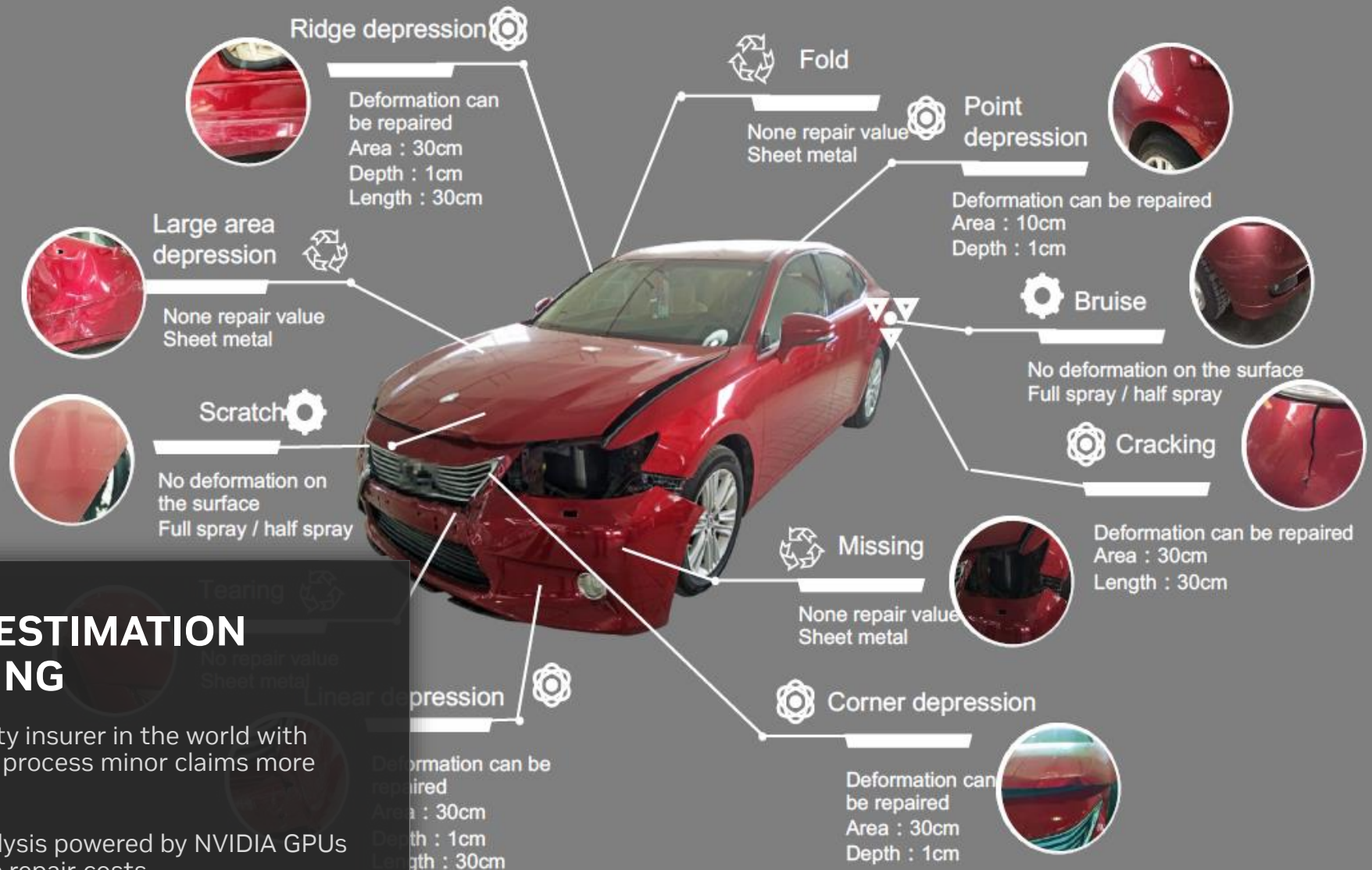
Now:

With the support of Rapids, we can deploy PageRank on our DGX-1 GPU server using cuGraph, the computation and data loading time is much less.

Model iteration time:

The model iteration time can be reduced from weeks to days which helps to detect up-to-date fraud behaviors and reduce loss.

NVIDIA



INTELLIGENT DAMAGE ESTIMATION AND CLAIMS PROCESSING

Ping An — the largest Casualty & Property insurer in the world with over 180 million customers — wanted to process minor claims more quickly.

The company implemented AI image analysis powered by NVIDIA GPUs to rapidly identify damages and estimate repair costs.

Today, Ping An processes up to 31,000 claims per day, with 98.7% of those paid out in less than a day.

NVIDIA. 中国平安 PING AN 保险 · 银行 · 投资

Claims AI

ControlExpert

Division of Allianz

- Global motor claims management company processes 18M claims/year
- NVIDIA AI Enterprise to automate claims process using computer vision
- Reduce time to process claim from days to minutes

Process Automation

Image classification, computer vision, and natural language processing for insurance claims

Solution Showcase

NVIDIA

ControlExpert Revolutionizes Motor Claims Management With NVIDIA AI Enterprise

"We have a vision where drivers around the world can get car damage claims settled fairly in a day. NVIDIA AI Enterprise gave us the performance to provide our customers with real-time responses as well as the security, stability, and support to provide the best customer service 24/7."

Dr. Andreas Witte, Chief Technology Officer, ControlExpert

Accelerating the Motor Claims Process

ControlExpert is a global leader in motor insurance claims management based in Germany and operating in more than 30 countries with more than 900 employees. Their innovative solutions simplify claims management and are used by more than 300 insurance companies worldwide.

ControlExpert uses AI to support their company vision, helping drivers around the world get their damages fairly settled on the same day.

Using AI to Support Customers and the Claims Process

ControlExpert used both computer vision and natural language processing (NLP) to develop an end-to-end claims management solution for insurance companies and their claimants. The solution lets customers and claimants notify the insurer about their claim, take photos of their vehicle in the event of an accident, and provide documentation needed for settlement.

The solution classifies the content of images, such as vehicle photos or documentation (e.g., license or registration). ControlExpert uses AI to identify vehicle information like make, model, color, and license plate. Based on images of vehicle damage, ControlExpert also developed AI models to segment visible vehicle parts and precisely detect the severity of the damage, generating a detailed description of the damage as well as cost estimation for repairs.

Using NLP, AI models can extract and analyze data from documentation, including invoices, appraisals, and emails, to provide a decision about claim payments.

With over 18 million claims a year, ControlExpert needed a solution that could quickly process large amounts of data.

ControlExpert

Customer Profile

- Organization: ControlExpert
- Founded: 2002
- Location: Germany
- Website: controlexpert.com
- Industry: Financial services

Summary

- ControlExpert wanted to develop an end-to-end product that would settle claims in a day.
- They needed a solution that could process over 30,000 claims and check over 250,000 images a day.
- NVIDIA AI Enterprise and NVIDIA A100 Tensor Core GPUs delivered the performance needed to process all the data quickly.
- ControlExpert uses PyTorch to develop and train computer vision and NLP models and NVIDIA Triton™ Inference Server to deploy them.

> *Using NVIDIA DGX Cloud and Base Command Platform's dataset management and orchestration capabilities, our data scientists have reported 2X speed up in running experiments.*

— Neda Hantehzadeh, PhD, Director of Data Science, CCCIS

AI automates claim estimations in seconds, elevating the customer experience

Challenge

CCC Intelligent Solutions processes 16 million insurance claims each year.

Wanted to minimize low-value, high volume, repetitive tasks in claims estimations.

Needed to support many data scientists and engineers to deliver AI-based solutions to market faster.

Solution

Established an end-to-end hybrid cloud AI development and training pipeline, which includes NVIDIA DGX Cloud and additional DGX systems on-premises

Integrated NVIDIA Base Command Platform into their development pipeline for dataset management and orchestration, delivering a 2X speed up in running experiments.

This AI pipeline has enabled CCC to unleash new innovations in the market, including their CCC Estimate-STP technology that provides line-level claim estimates in seconds based on insurer rules.

- NVIDIA DGX Cloud *for training*
- NVIDIA Base Command Platform *for workflow management*
- NVIDIA AI Enterprise *PyTorch and CUDA*

2X	Speed up in running data scientists' experiments
30X	Expedited model development
24/7	Access to DGX systems on demand

NVIDIA

Example: Explainable AI (XAI) in Underwritting

GPU-accelerated XAI and interactive exploration of model explainability

Problem

Transparency of AI Applications is Critical to Acceptance — several options exist: from interpretable models to post-hoc XAI like SHAP values:

- Game theoretic, model agnostic approach for global and local XAI
- Heavy computational resources for real-life data sets for credit scoring (high-risk AI application)

Solution

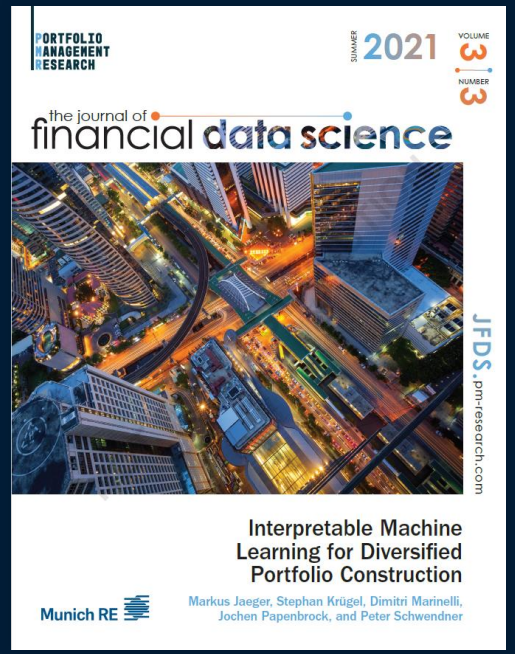
- End-to-end GPU accelerated Open Data Science with XAI values: RAPIDS and dmlcXGBoost
- Interactive Plotly XAI dashboard for analysis and exploration of XAI values, contributions, interactions and clusters
- Also as on-prem application for sensitive data
- Entire end-to-end data science workflow is based on GPU-accelerated Python libraries for ultra fast testing of models and XAI analysis

XAI-POWERED DIVERSIFIED PORTFOLIO CONSTRUCTION

Munich RE

Munich Re Markets helps the global life and pension industry deliver on its investment management promises - helping their clients increase wealth while preserving capital.

Using GPU acceleration, Munich Re Markets can now provide clients with more robust, faster and smarter asset allocation decisions on demand.



Interpretable Machine Learning for Diversified Portfolio Construction

Research papers published in the JFDS and further insights

Find out more

FIVE - Investment indices for savings & retirement products

Accessible via reinsurance with Munich Re Markets

Find out more

NVIDIA

Spatial Finance

Top Technologies

- LLMS, Accelerated Data Science, Scientific Visualization, and EO/Geospatial Data Processing
- Increased business benefits of finer spatial scale, higher frequency revisit, data fusion, etc.
- Real-time monitoring and Streaming Sensor Processing
- End-to-end Analysis of Large 3D Geospatial datasets
- AI Approach for creating maps

Use Cases (Asset Tracking, Commodity Flows, ESG Monitoring, Underwriting)

- Quantification of material ESG risk factors
- Identification of gaps/greenwashing, pollution/deforestation
- Identify high risk sourcing areas, measure progress and trend
- Etc.

GPU-accelerated SAR (Radar) Processing – 269x for ½ billion points

Raster

- Decoding and encoding, data access: nvJPEG2000, nvTIFF (GeoTiff/COG), etc.
- 2D and 3D processing image processing (CV-CUDA, NPP, VPI, cuCIM, DALI)

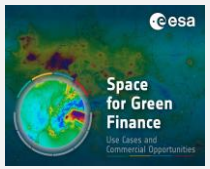
Vector

- RAPIDS cuSpatial: go-to library in the GIS community; spatial algorithms for large scale vector data analysis.

Graph

- RAPIDS cuGraph: supports the creation and manipulation of graphs followed by the execution of scalable fast graph algorithms.

Computer Vision

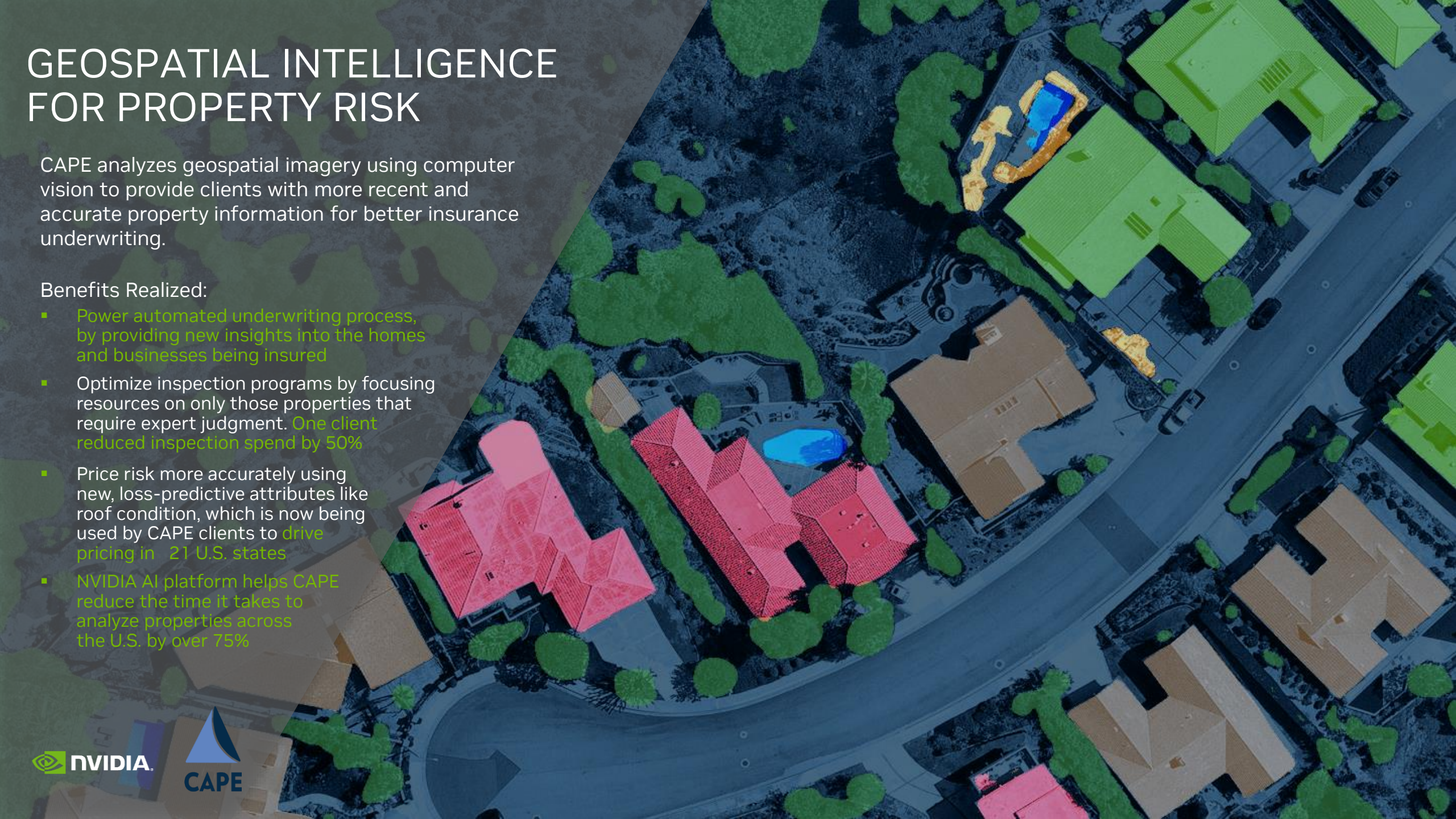


GEOSPATIAL INTELLIGENCE FOR PROPERTY RISK

CAPE analyzes geospatial imagery using computer vision to provide clients with more recent and accurate property information for better insurance underwriting.

Benefits Realized:

- Power automated underwriting process, by providing new insights into the homes and businesses being insured
- Optimize inspection programs by focusing resources on only those properties that require expert judgment. One client reduced inspection spend by 50%
- Price risk more accurately using new, loss-predictive attributes like roof condition, which is now being used by CAPE clients to drive pricing in 21 U.S. states
- NVIDIA AI platform helps CAPE reduce the time it takes to analyze properties across the U.S. by over 75%



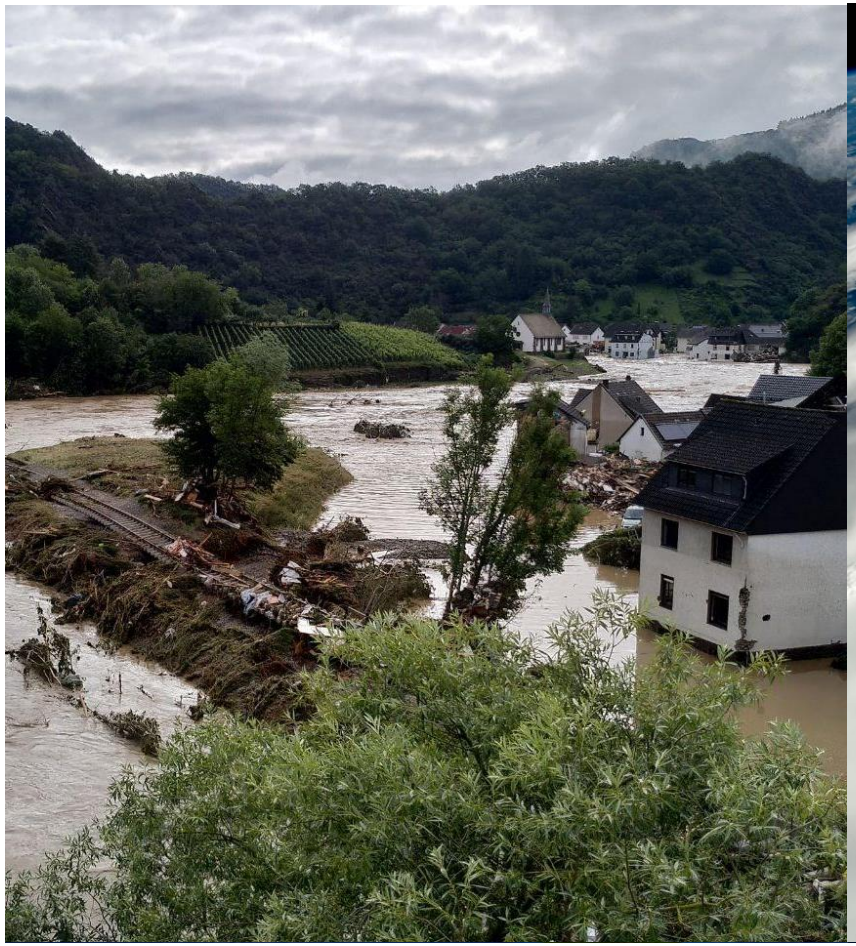
NVIDIA

CAPE

Earth-2 Platform
Weather and Climate Simulations

Extreme Weather Events Cause Damage Worth Billions of $

Ahr Valley Floods:

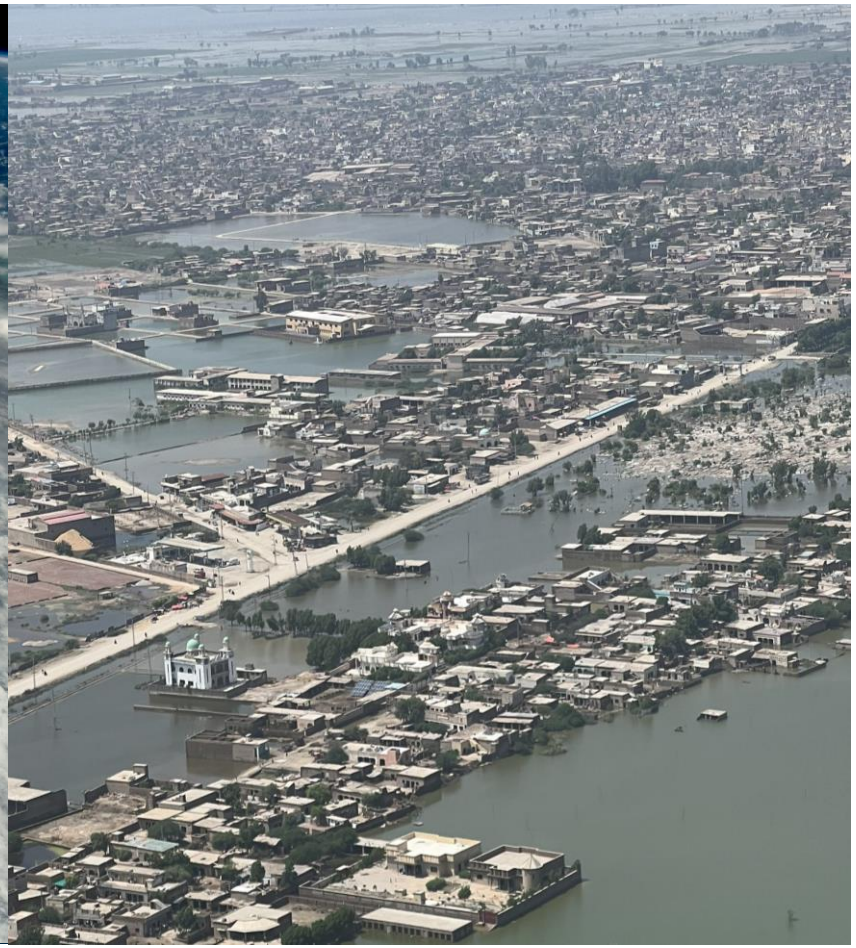


https://commons.wikimedia.org/wiki/File:Hochwasser_in_Altenahr_Altenburg.jpg

Hurricane Ian:

https://commons.wikimedia.org/w/index.php?curid=123606695

Pakistan Floods:



https://commons.wikimedia.org/wiki/File:Flood_in_Pakistan_2022.png

Broader Access to Weather and Climate Simulations
Imagine you could Select a Region of the Planet...

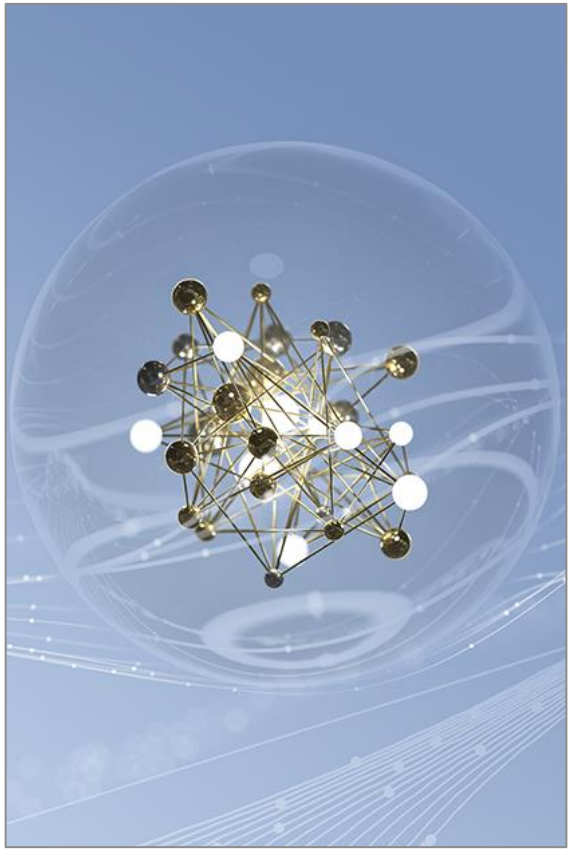
Earth-2 Program

Build the technology needed to create the digital twin of the earth's weather and climate systems

Accelerating NWP codes on GPU

AI research and collaboration with the science community

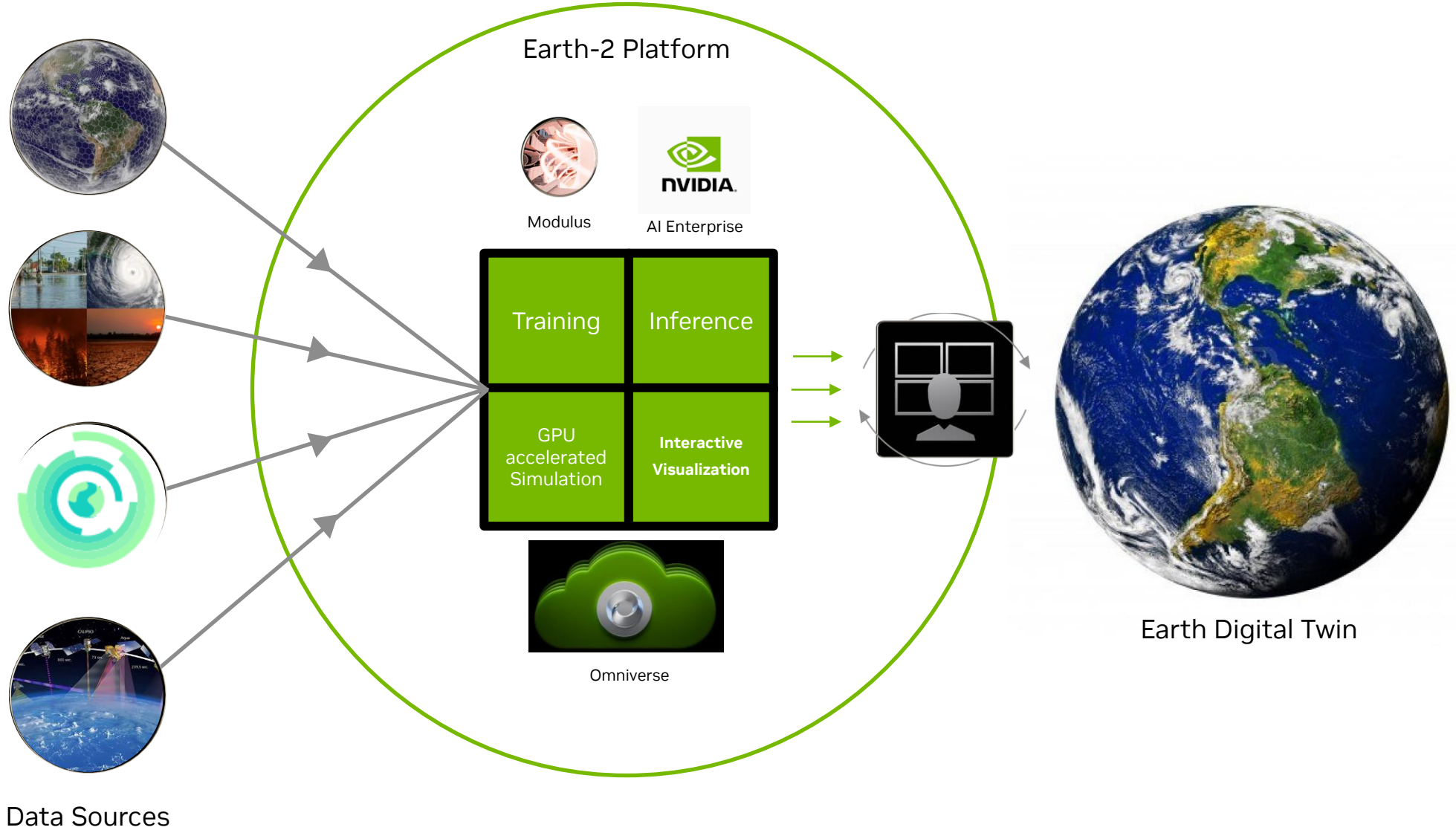
Interactive Visualization – Digital Twins



Operationalize using Cloud Services

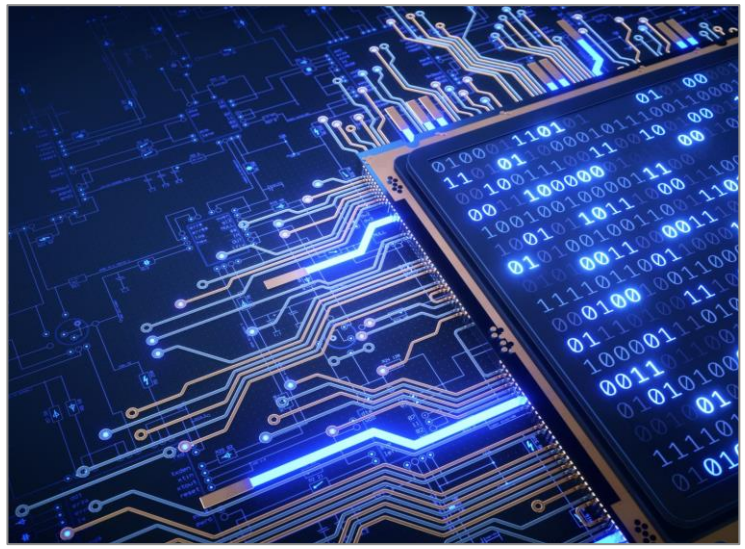
Earth-2

Connecting complex simulation, data and AI workflows



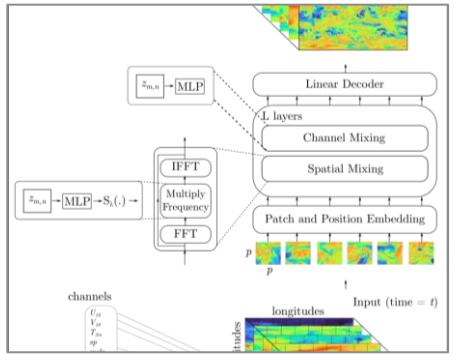
AI Weather Models Change the Game

Forecasts within seconds, for the first time!



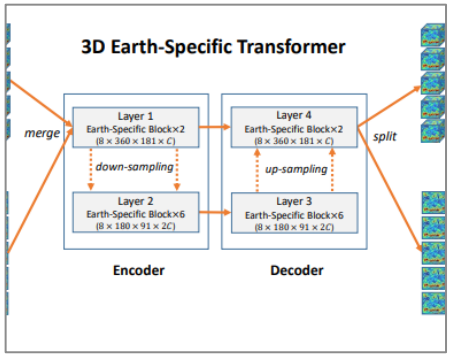
- AI models - predict global weather forecasts 4 orders-of-magnitude faster than NWP while approaching or surpassing state-of-the-art accuracy.
- Forecasts within seconds
- Alleviates data bottlenecks – can run these simulations at much cheaper cost.
- Opens up the possibility to explore orders of magnitude larger # of weather scenarios

Global Medium Range Weather – Key Milestones



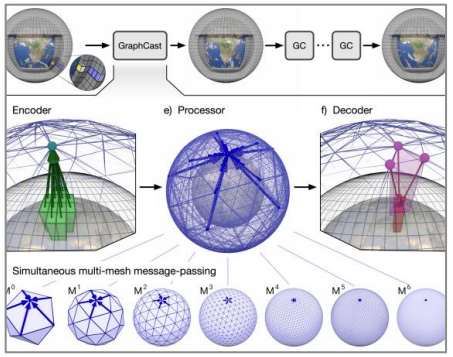
Feb 2022: FourCastNet

A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. FourCastNet matches the forecasting accuracy of the ECMWF Integrated Forecasting System (IFS).



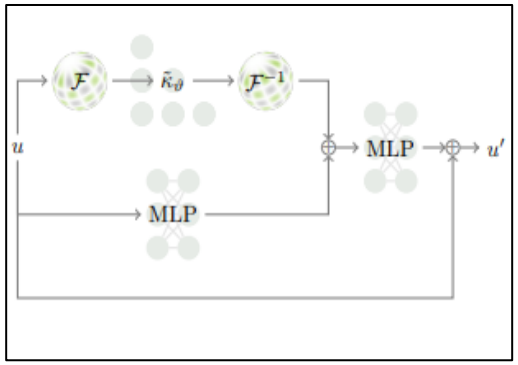
Nov 2022: Pangu Weather

With a 3D Earth Specific Transformer (3DEST) architecture and a hierarchical temporal aggregation algorithm. Outperforms IFS in terms of accuracy (latitude-weighted RMSE and ACC) of all factor



Dec 2022: GraphCast

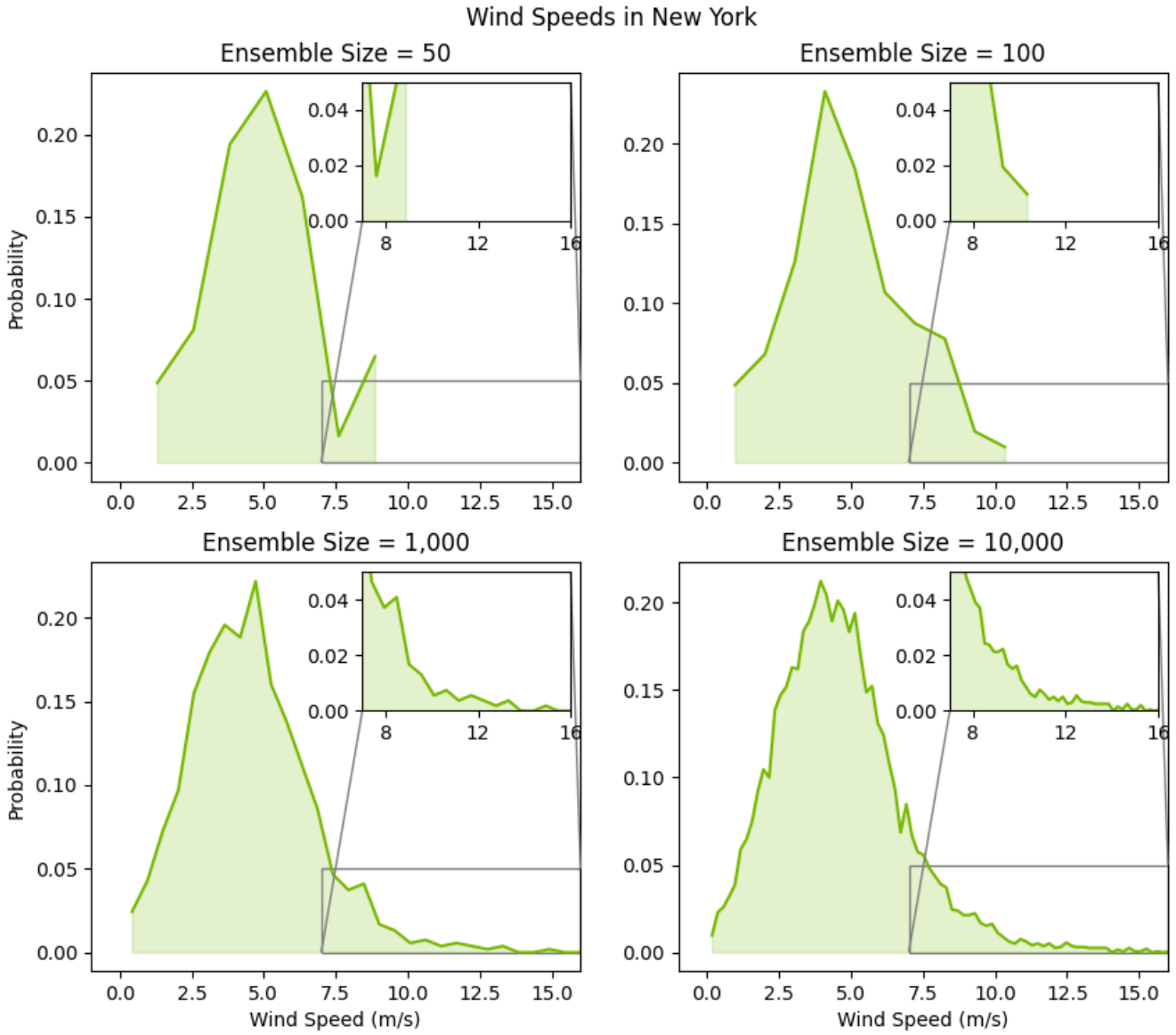
Based on GNNs in an "encode-process- decode" configuration. GraphCast's forecast skill and efficiency compared to HRES shows MLWP methods are now competitive with traditional weather forecasting methods

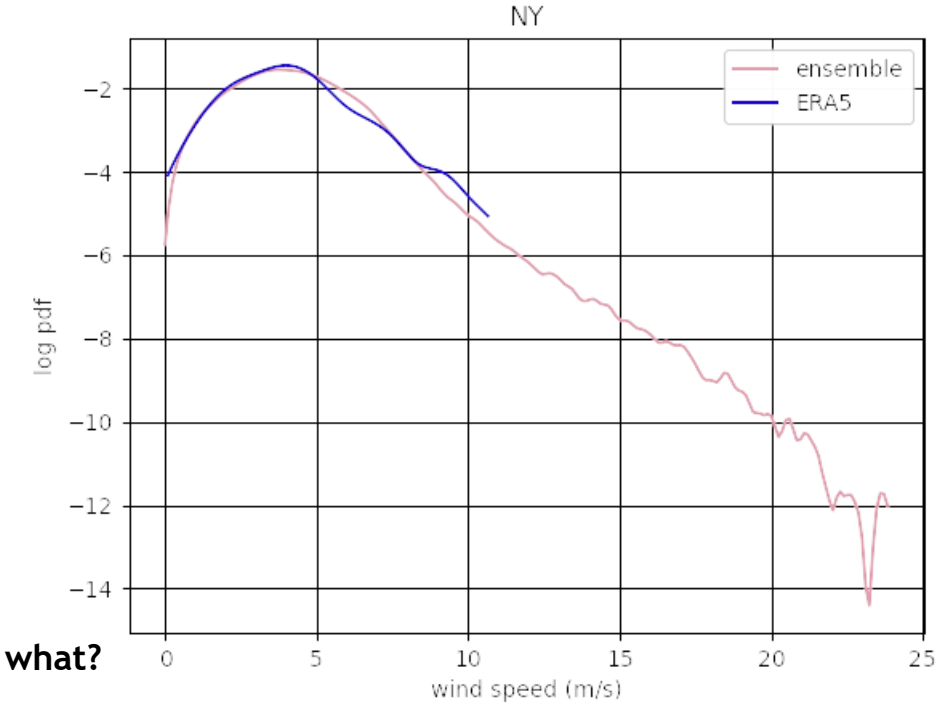


Jun 2023: FourCastNet_V2

A novel SFNO equivariant architecture for modeling nonlinear chaotic dynamical systems on the sphere. The high accuracy and long-term stability promises for the application of Spherical Fourier Neural Operators for long term forecasting.

Example of Our Massive Ensemble Predictions: Wind Speed Over NYC



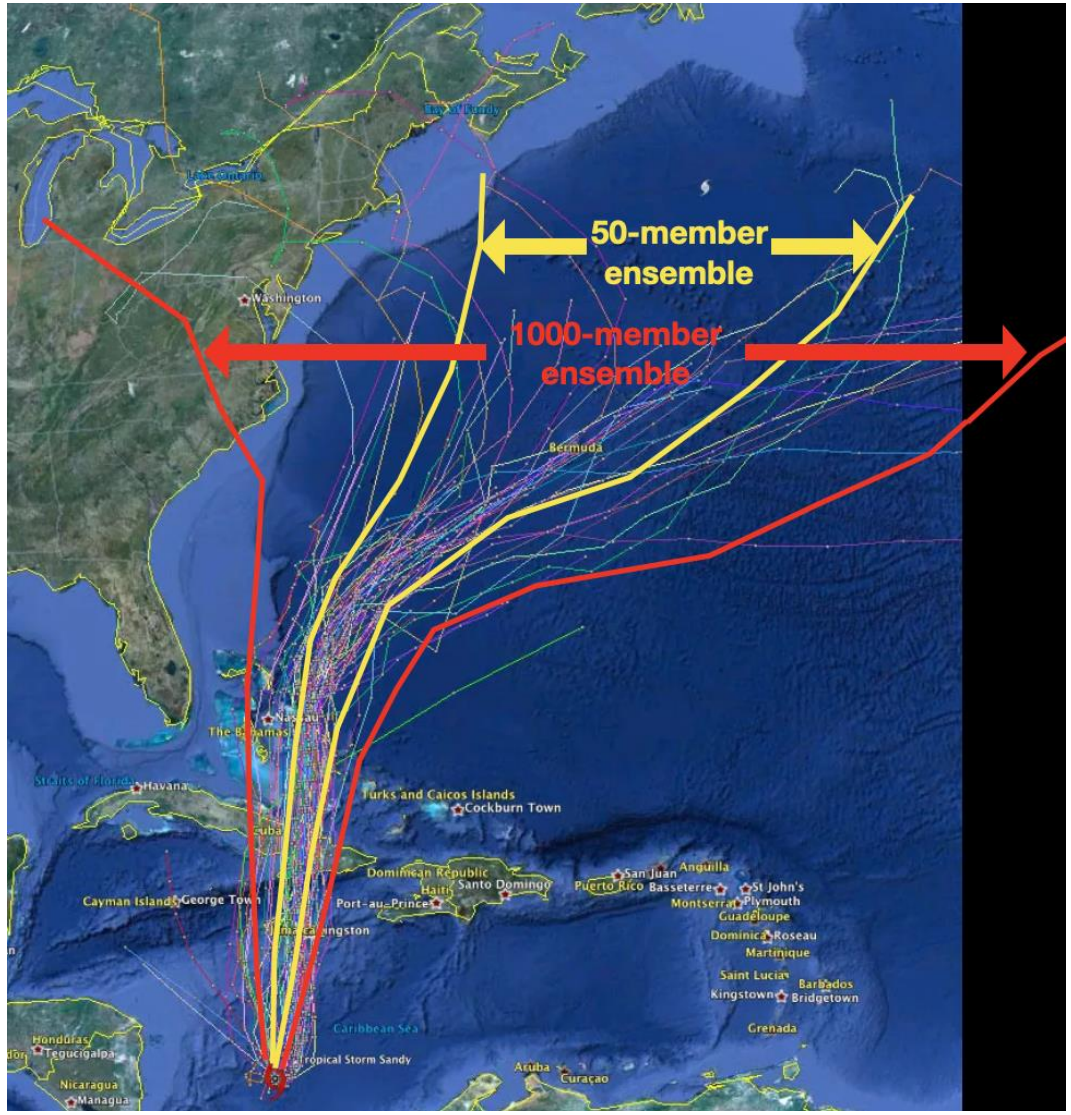


So what?

- Fluctuations in these newly revealed tail statistics could form a new data stream for those who plan around rare event risk.
- Some industries (e.g. wind energy) have inputs that especially depend on the tail values e.g. weighted by (wind speed)^3.
- Data-driven forecast methods are attractive for extreme phenomena like storms since those tend to require hard to afford grid resolution in classical prediction.

Unprecedented Sampling of Low-likelihood High-impact Extremes

Multi-thousand-member ensembles of hurricanes, typhoons, cold snaps and more.



- Explicitly sampling probabilities.
- Including the long tails of highly skewed distributions.
- Example: Hurricane Sandy.
- Benefit to users of weather data making hard decisions with major consequence and cost.

E-2 Services

AI augmented Climate Tech Applications

GPU accelerated on Grace Hopper

AI based forecasting – Km scale forecasts globally

Stream RTX accelerated visualization

SIMULATION AS-A-SERVICE

Coming soon

AI AS-A-SERVICE

VISUALIZATION AS-A-SERVICE

Coming soon

DGX Cloud

Partner Ecosystem

The Weather Company

spire

交通部中央氣象署 Central Weather Administration

NCDR

meteomatics

atmo

ClimaSens

tomorrow.io

north.io

Inference-as-a-Service

Early access

- What is E2 inference service?
  - Cloud APIs to:
    - Upload the Initial condition
    - Request a forecast - specifying a configuration of interest - variables of interest, location of interest, # of ensembles
  - Pythonic or bash/curl calls
  - AI Weather forecasting models
    - Global models:
      - FourCastNet_V2[ICML paper],
      - GraphCast
      - More models to come soon
    - Downscaling model: Corrdiff
    - Diagnostic models: Precipitation, Windgust, Cyclone tracking

```
export NVCF_API_KEY=<your-nvcf-api-key>
python ./data/gfs.py --date 2023-12-03T00:00:00    # this will produce a file named gfs.nc
python client.py \
        --function-id=<function-id> \
        --version-id=<version-id> \
        --config-path=./examples/deterministic_forecast/input.json \
        --asset-file-path=./gfs_2023_12_3T0_0_0.nc
```

Documentation:

https://gitlab.com/nvidia/modulus/earth-2-inference-api

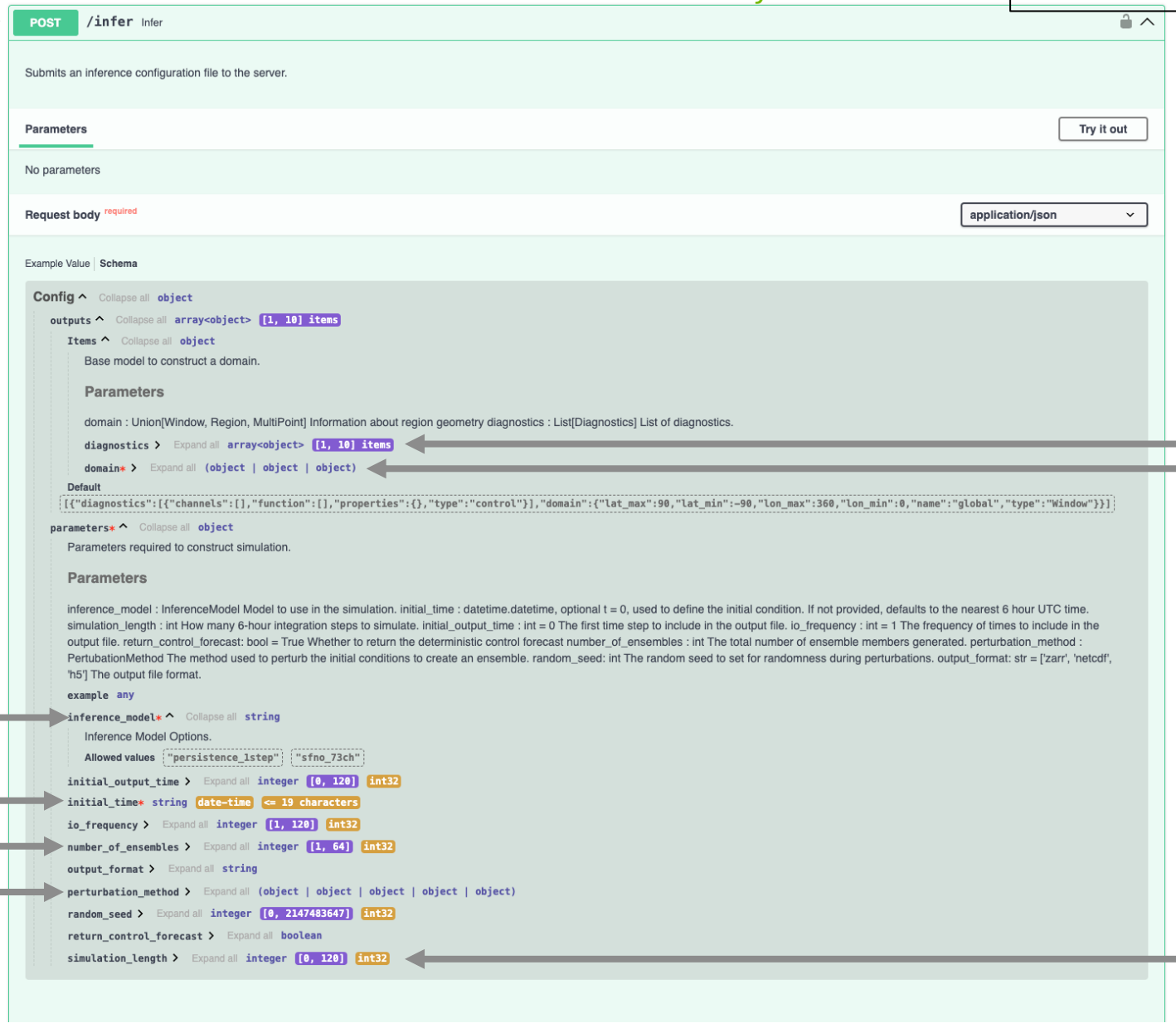
Inference-as-a-Service

APIs and Functionality

Documentation:

https://gitlab.com/nvidia/modulus/earth-2-inference-api/-/tree/main/docs

Inference REST API



Choose the inference model
- **FourCastNet**
- **GraphCast**
- Persistence

Specify initial time for forecast

Number of ensembles

Ensemble perturbation methodology

Available output quantities:
- Deterministic forecast
- Full ensemble
- Ensemble statistics
- Histograms
- Tropical Cyclone tracking
- Downscaling using **CorrDiff**

Output domain (full globe or regional output)

Forecast Length

Resources

- More resources
  - Earth-2 Solutions page
  - Modulus product page

- Call to action
  - For developers building Climate tech applications, please reach out here for early access to cloud services.
  - For Enterprises interested in Earth-2 platform stack, please reach out to your Nvidia account rep
  - For researchers in research and science collaboration – Engage with us on our open-source repository

GenAI / LLM

2024: The Year of Production

Driving generative AI into production leveraging end-to-end full stack solutions

Explosion

ChatGPT is announced
100 million users in 2 months

Experimentation

POCs developed using API services and open models

Production

Enterprises and ISVs are investing in accelerated infrastructure, trying to find the best path to production

Target Use Cases for Generative AI

AI assistants are driving the explosion of POCs

Intelligent Chatbot

Focus is on question-and-answer tasks.

Ex. Customer Service Agent, Brand Ambassador, Help Desk

Knowledge Base Copilot

Connects to knowledge bases performs tasks such as writing, coding, generating images, etc.

Ex. Documentation Copilot, IT Bugs Assistant, Field Agent Copilot

Code Generation

Help develop or troubleshoot code based on natural language. Can work across common languages or be proprietary languages.

Ex. GitHub Copilot, ChatUSD, Software Development Assistant

Enterprises Face Challenges Experimenting With Generative AI

Organizations must choose between ease of use and control

Easy to use APIs for development

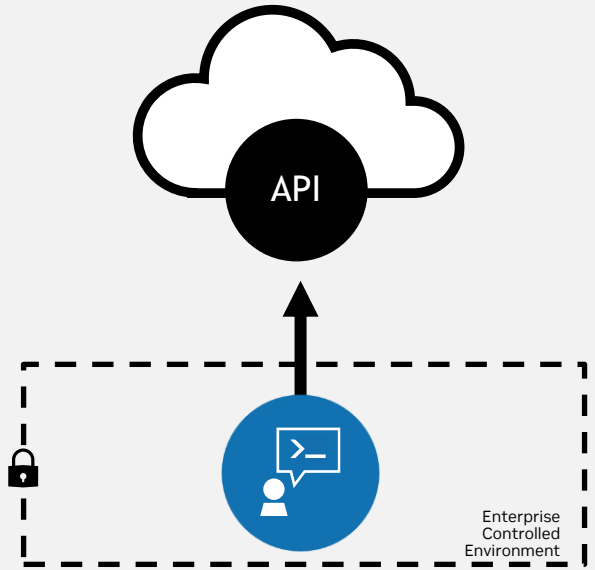
Fast path to getting started with AI

Infrastructure limited to managed environment

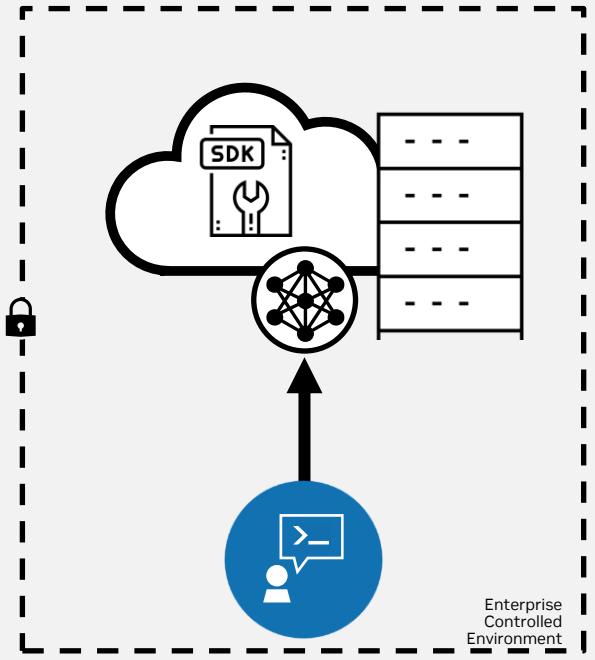
Data and prompts are **shared externally**

Limited control for overall generative AI strategy

Managed Generative AI Services



Open-Source Deployment and Customization



Run anywhere across data center and cloud

Securely manage data in self hosted environment

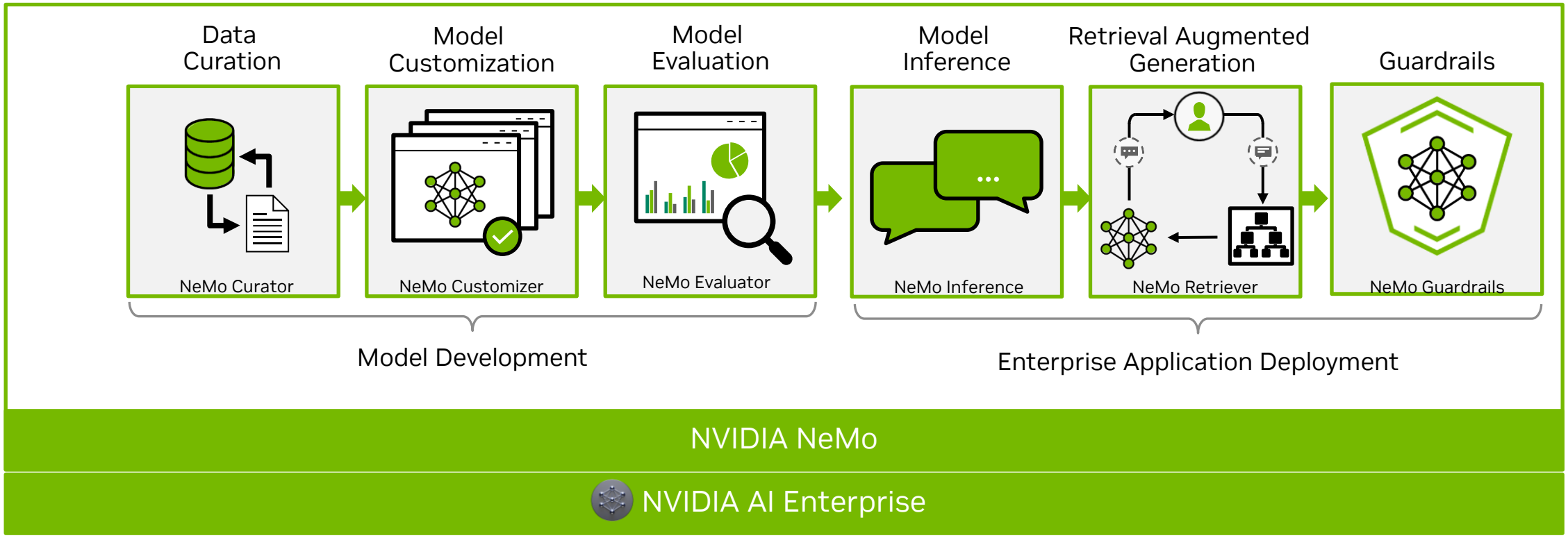
Tuning required for different infrastructure

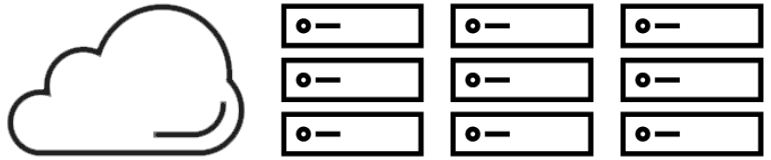
Custom code for APIs and fine-tuned models

Ongoing maintenance and updates

Building Generative AI Applications for the Enterprise

Build, customize and deploy generative AI models with NVIDIA NeMo

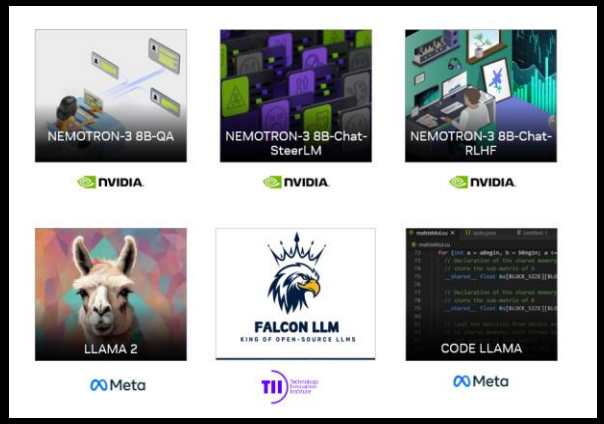




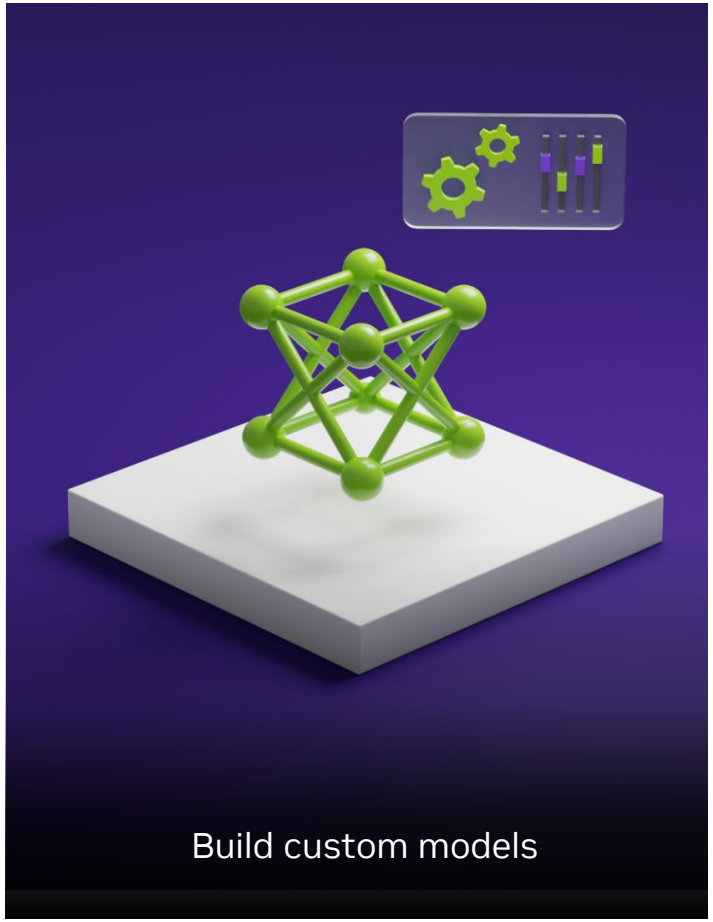
Microsoft Azure aws ORACLE NVIDIA DGX Cloud DELL Technologies Hewlett Packard Enterprise Lenovo SUPERMICRO NVIDIA

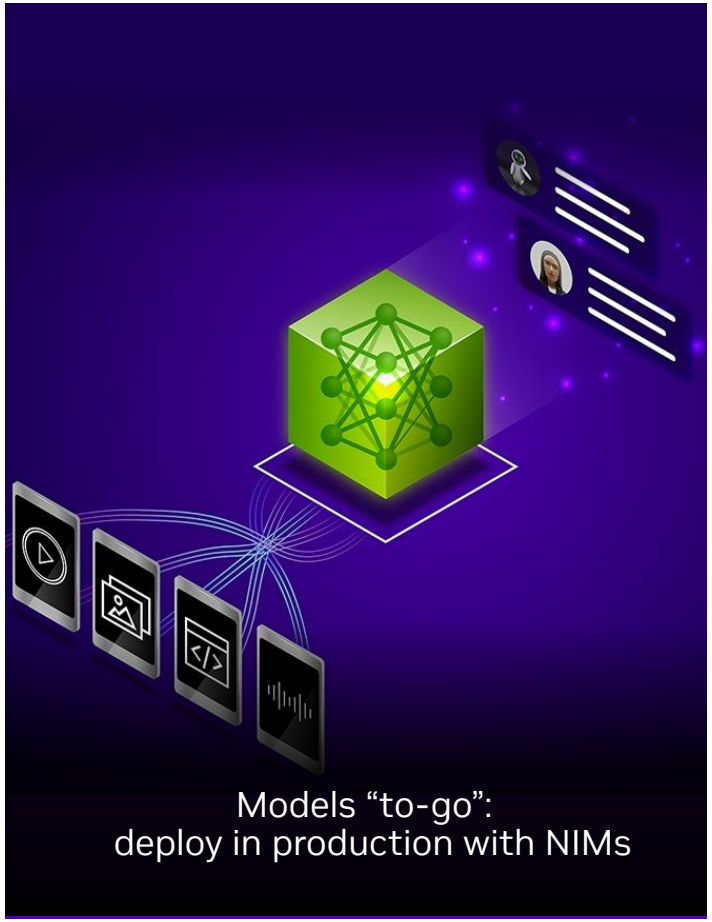
Getting Started With the NeMo Framework/Microservices

Experience, prototype, and deploy the latest AI models



- State-of-the-art community, commercial and NVIDIA-built models
- Performance-optimized for GPU-accelerated stack
- Experience foundation models running via API endpoints for prototyping





NeMo Inference Microservices (NIM) for Generative AI

Set of easy-to-use microservices for accelerating the deployment of foundation models on any cloud or data center

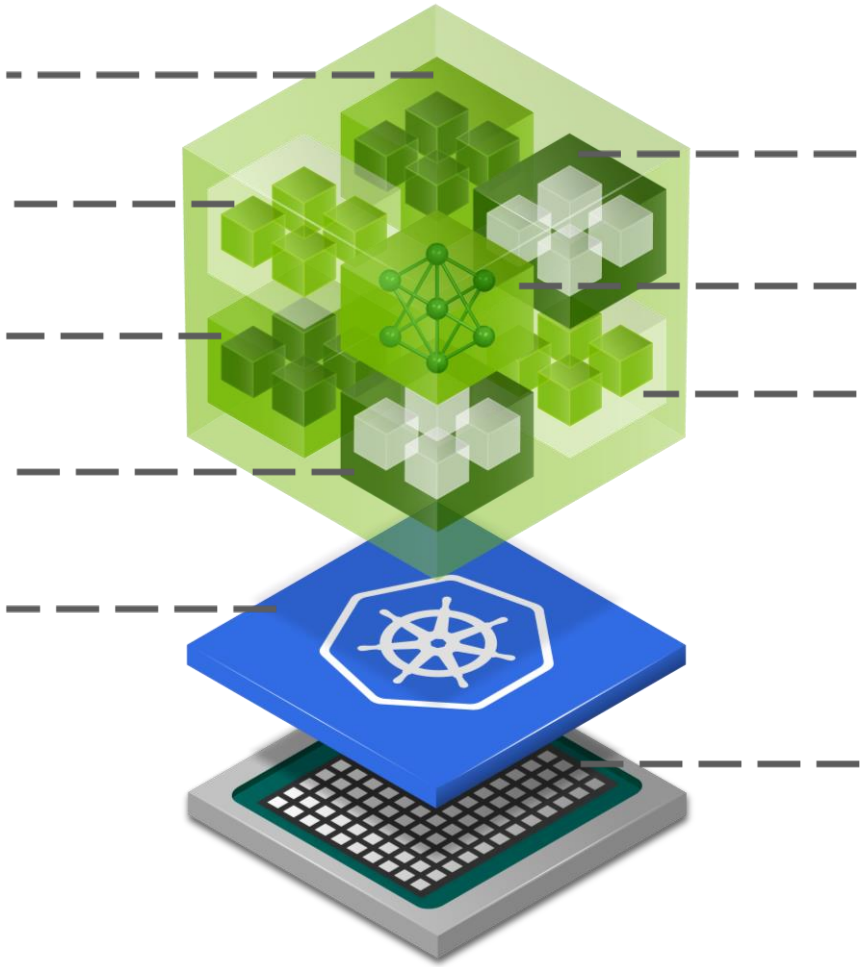
Industry Standard APIs
Text, Speech, Image, Video, 3D, Biology

Triton Inference Server

Cloud Native Stack
GPU Operator, Network Operator

Enterprise Management
Health Check, Identity, Metrics, Monitoring, Secrets Management

Kubernetes



TensorRT and TensorRT-LLM
cuBLAS , cuDNN, In-Flight Batching, Memory Optimization, FP8 Quantization

Optimized Model
Single GPU, Multi-GPU, Multi-Node

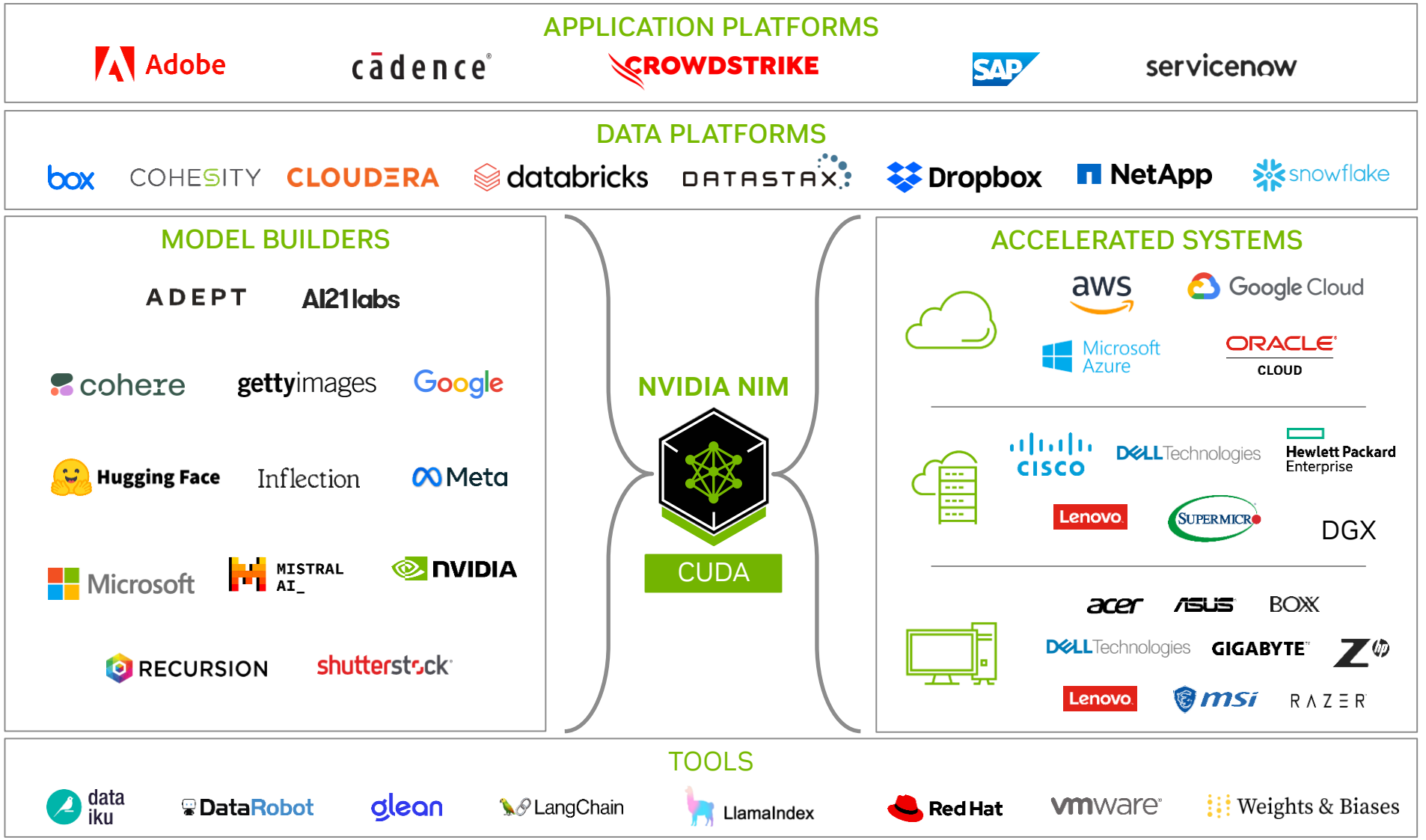
Customization Cache
P-Tuning, LORA, Model Weights

NVIDIA CUDA

Microsoft Azure · aws · Google Cloud · ORACLE · DGX & DGX Cloud · DELL Technologies · Hewlett Packard Enterprise · Lenovo · SUPERMICRO

NVIDIA

Connecting Millions of Developers to 100s of Millions of GPUs



Some Interesting References in Financial Services

MORNINGSTAR

"Morningstar is using NeMo in its Data Collection research and development on how LLMs can scan and summarize information from sources such as financial documents to quickly extract market intelligence."

Shariq Ahmad
Head of Data Collection Technology

NVIDIA GTC

Financial Services Session

Ilay Chen
PayPal

How PayPal Reduced Cloud Costs by up to 70% With Spark RAPIDS

BNY Mellon
646,702 followers
2w • Edited •

ANNOUNCEMENT BNY Mellon becomes the first global bank to deploy an AI supercomputer powered by NVIDIA.

With more than 600 opportunities in #AI identified and dozens already in development, this collaboration will streamline and accelerate innovation within our business and across the global financial system.

"Key to our technology strategy is empowering our clients through scalable, trusted platforms and solutions," said #BNYMellon Chief Information Officer Bridget Engle. "By deploying NVIDIA's AI supercomputer, we can accelerate our processing capacity to innovate and launch AI-enabled capabilities that help us manage, move and keep our clients' assets safe."

#Nvidia #supercomputing #artificialintelligence #cio

BNY Mellon, First Global Bank to Deploy AI Supercomputer Powered by NVIDIA DGX SuperPOD With DGX H100

HPC wire

Search... Go

Bloomberg Uses 1.3 Million Hours of GPU Time for Homegrown Large-Language Model
By Agam Shah

April 6, 2023

Financial firm Bloomberg is trying to prove that there are smarter ways to fine-tune artificial intelligence applications without the ethical or security concerns plaguing the likes of ChatGPT.

Acknowledgments and Disclosure of Funding

We would like to acknowledge the people who helped us, including Emmanuel Scoullos (NVIDIA) and Can Karakus (Amazon Web Services).

Financial Services

Deutsche Bank and NVIDIA

Modernizing the financial industry with AI-powered services.

Read Blog

March 20-23

NVIDIA GTC

Conversational AI / NLP

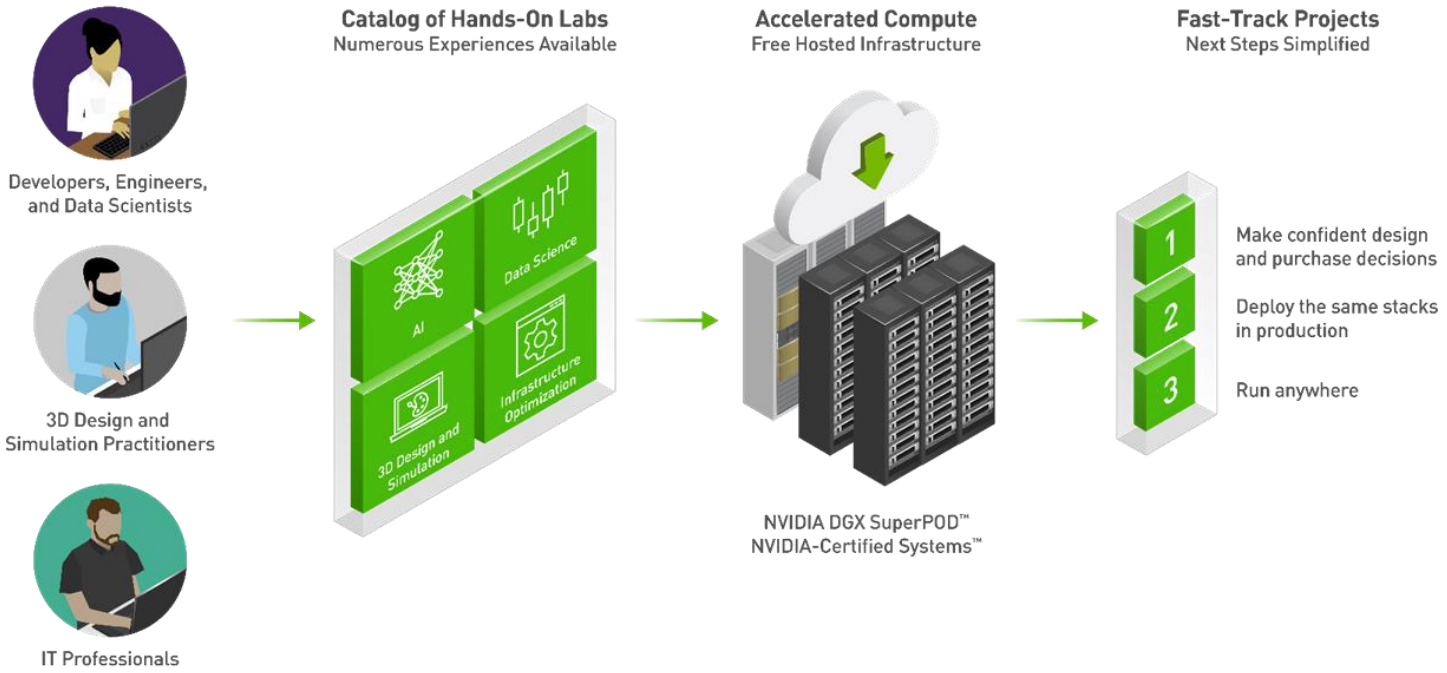
Large Language Models in Finance: Insights from Deutsche Bank

Matteo Zola
Deutsche Bank

Raluca Iordache
Deutsche Bank

NVIDIA LaunchPad

Instantly experience end-to-end workflows for AI, data science, 3D design collaboration, and more



NVIDIA DEEP LEARNING INSTITUTE

NVIDIA DEVELOPER

NVIDIA GTC GPU TECHNOLOGY CONFERENCE

- Deep Learning Institute
- Blogs, webinars, tutorials, developer program
- GTC – largest AI industry conference globally

Get Started at nvidia.com/launchpad

NVIDIA


NVIDIA