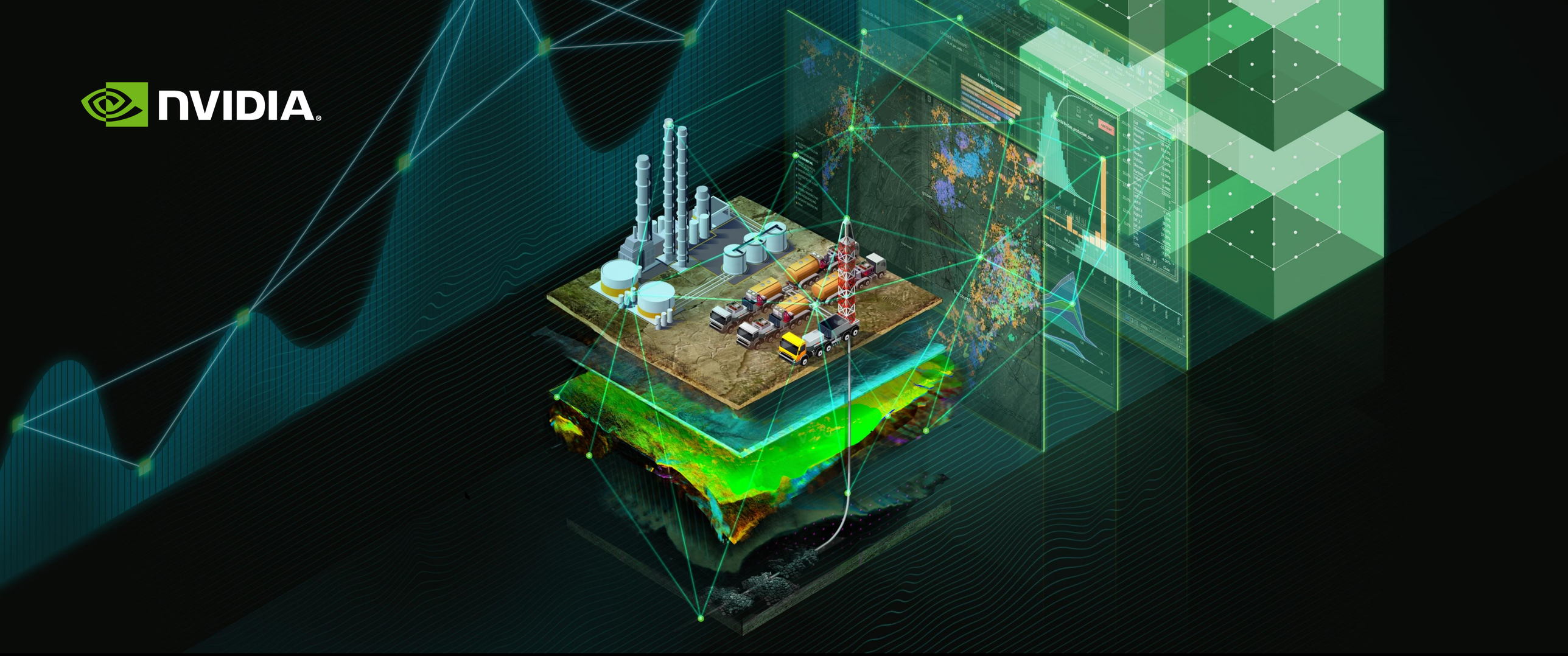
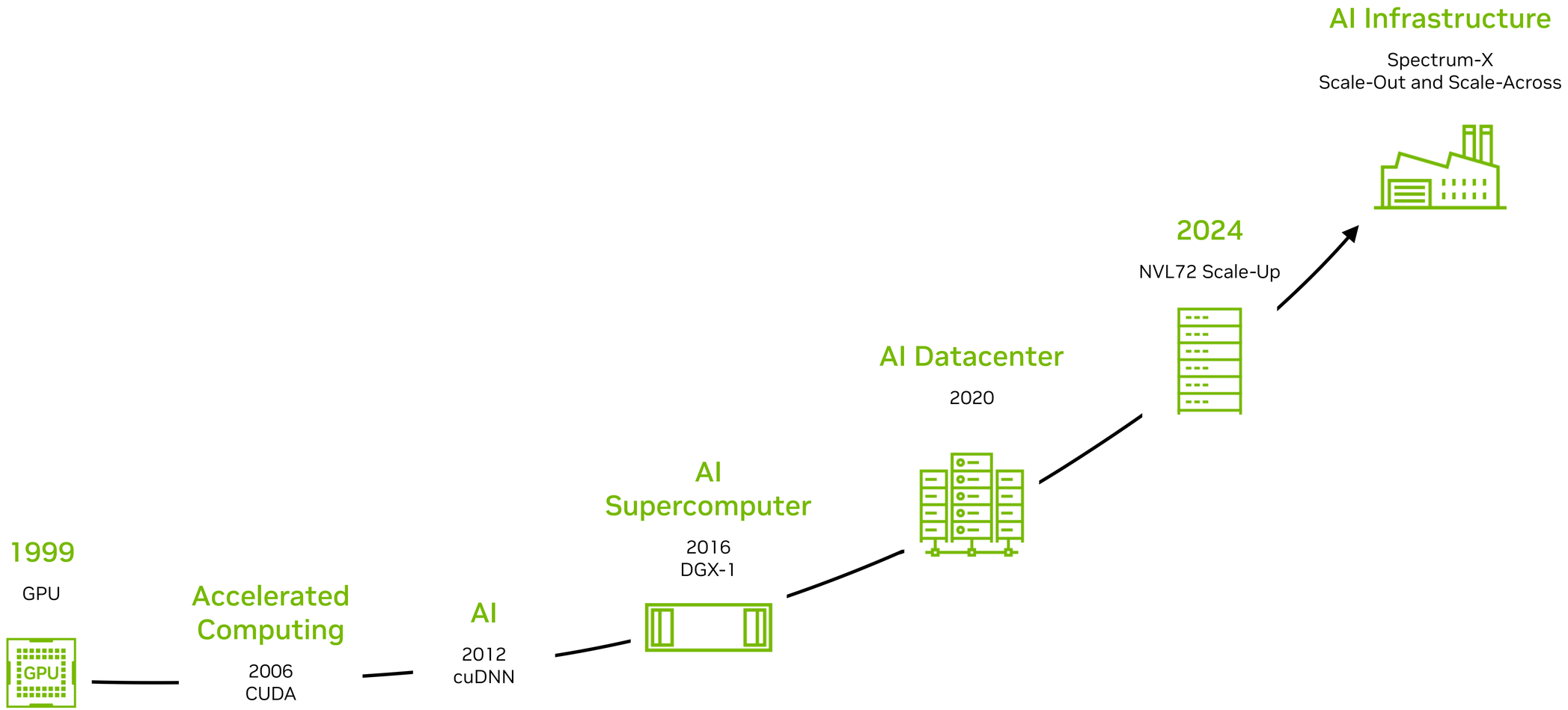


AI Seismic at Scale:
Frameworks, Mixed Precision, Foundation Models, and Physics-ML

João Paulo Navarro – Sr. Solutions Architect

jpnavarro@nvidia.com

NVIDIA’s Evolution From Chips to an AI Infrastructure Company



Leading the AI Era


Chatbots


Digital Biology


Robotaxi


AI Enterprise Agents


Science


Robotics


Manufacturing


AI Coder

Applications

Models

Infrastructure

Chips

Energy

Data Preparation

RAPIDS - cuDF

A GPU DataFrame library in Python with a pandas-like API built into the PyData ecosystem

Pandas-like API on the GPU

CPU

```
pandas

>>> import pandas as pd
>>> df = pd.read_csv("filepath")
>>> df.groupby("col").mean()
>>> df.rolling(window=3).sum()
```

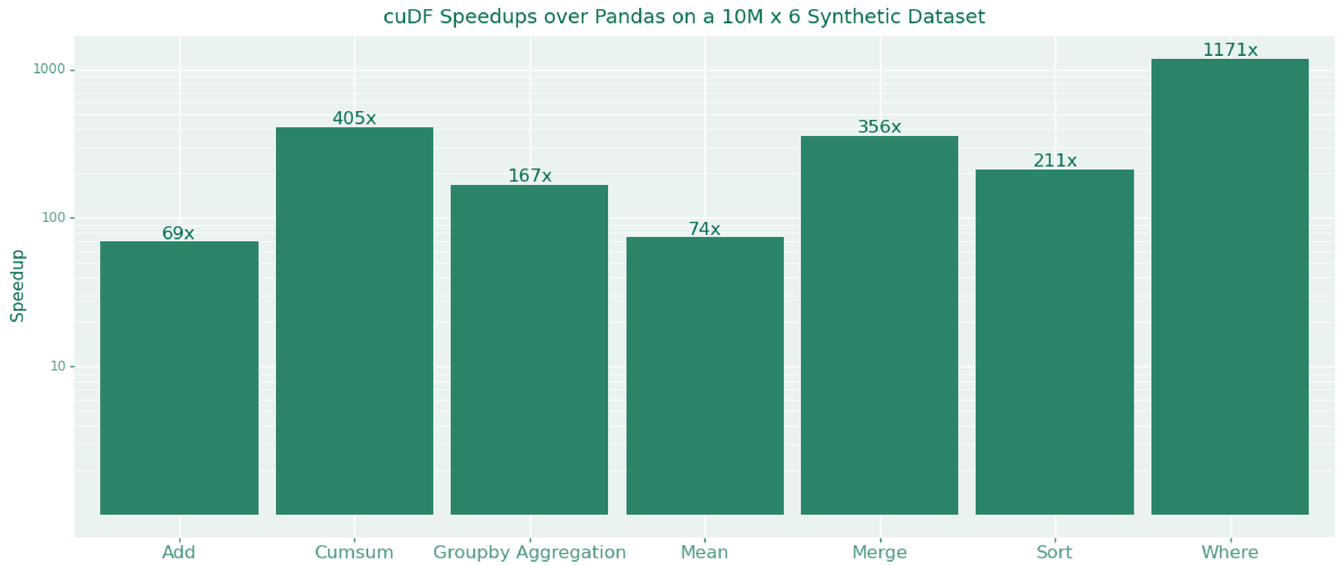
GPU

```
cuDF

>>> import cudf
>>> df = cudf.read_csv("filepath")
>>> df.groupby("col").mean()
>>> df.rolling(window=3).sum()
```

Average Speed-Ups: *10-100x*

Best-in-Class Performance (Benchmark)



Groupby	Strings and Regex	UDFs	Nested Types	Time Series
Indexing	Missing Data	CuPy Interoperability	Rolling Windows	

10 Minutes to cuDF

NVIDIA A100 vs. AMD EPYC 7642 48-Core Processor
cuDF Python vs. Pandas

CuPyNumeric and Legate Ecosystem

Distributed runtime framework and libraries

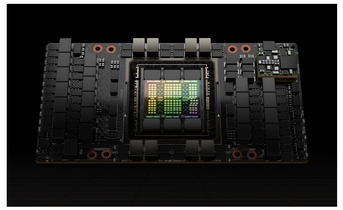
NumPy SciPy dmlc XGBoost polars

cuPyNumeric | Legate Boost | Legate Dataframe

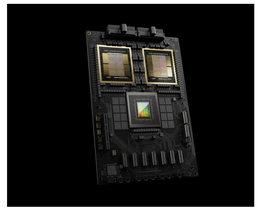
Legate: Productivity and Composability Layer

Legion: Implicit Parallelism Layer

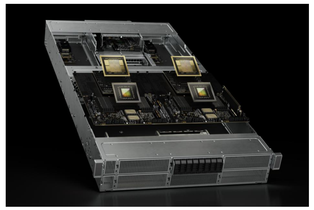
Realm (Runtime for Scalable and Portable Execution) Today



Single GPU



Mixed CPU/GPU



Single-Node Multi-GPU

Multi-Node Multi-GPU
Cloud & Supercomputer

Scaling NumPy Code with cuPyNumeric

Simple python code w/ NumPy
No partitioning code
No MPI code

```
import cupynumeric as np

m, n, k = 1000000, 1000000, 1000000 # 1Mx1M
A = np.random.rand(m,k).astype(np.float32)
B = np.random.rand(k,n).astype(np.float32)

res = A@B
```

Get your science done faster!

- Ideal for **domain scientists and researchers**.
- Prototype quickly in **the NumPy API**.
- **Scaling** from single CPU core to a multi-GPU multi-node supercomputer with thousands of GPUs.
- **Use HPC clusters without HPC programming** or expensive rewrites in e.g. MPI.
- Implements most NumPy functionality, **with growing support for SciPy, and other scientific libraries.**

Scales to multi-GPU multi-Node at run time

```
$ legate --launcher mpirun --nodes 64 --gpus 8 matmul.py
```

cuPyNumeric and Legate Slurm Integration

Simple Integration Enables Rapid Prototyping with cuPyNumeric on HPC Clusters

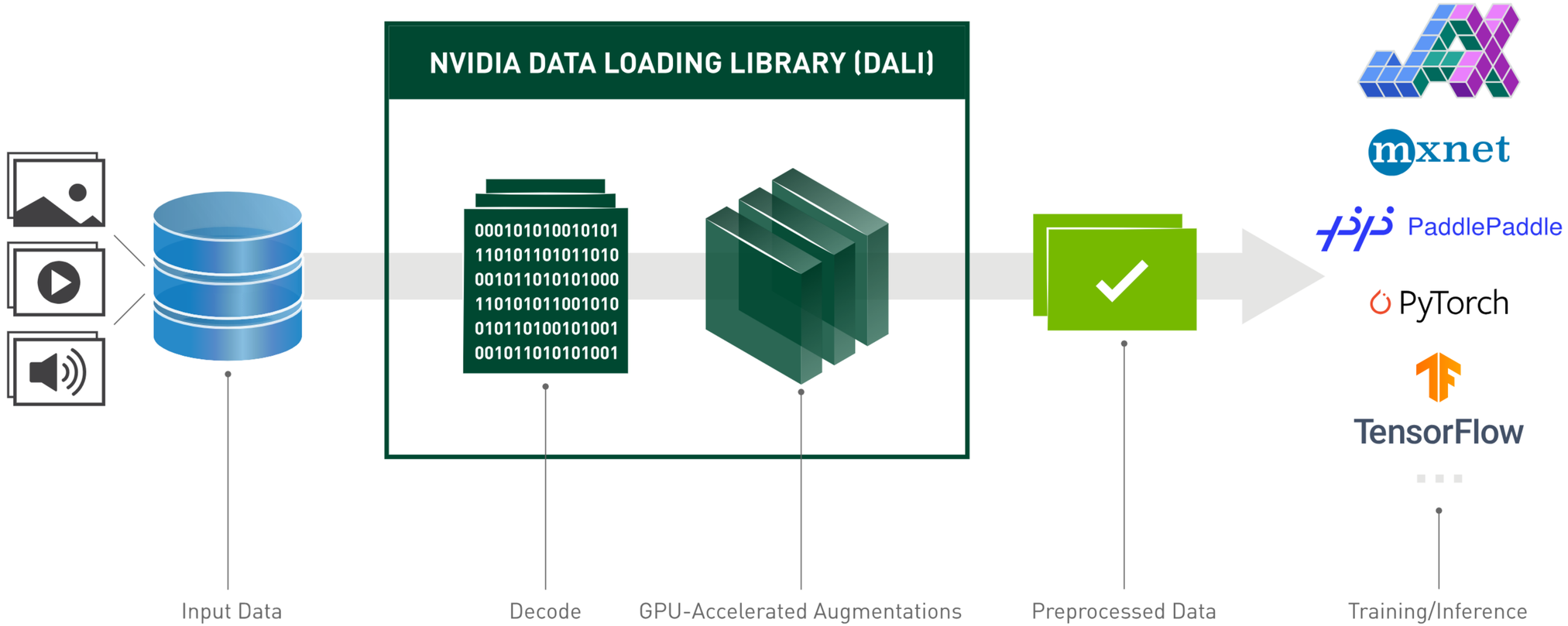
```
#!/bin/bash
#SBATCH --job-name=fwi-legate
#SBATCH --nodes=4                    # 4 Compute nodes
#SBATCH --ntasks-per-node=4          # 4 GPUs per node (e.g., HGX H100)
#SBATCH --gpus-per-node=4
#SBATCH --time=02:00:00

# Load environment (Container or Module)
module load nvidia/hpc_sdk

# THE MAGIC COMMAND:
# Legate parses the variables above and scales automatically.
legate fwi_solver_3d.py --cpus 4 --gpus 4 --fbmem 60000
```

NVIDIA DALI

Data Loading Library

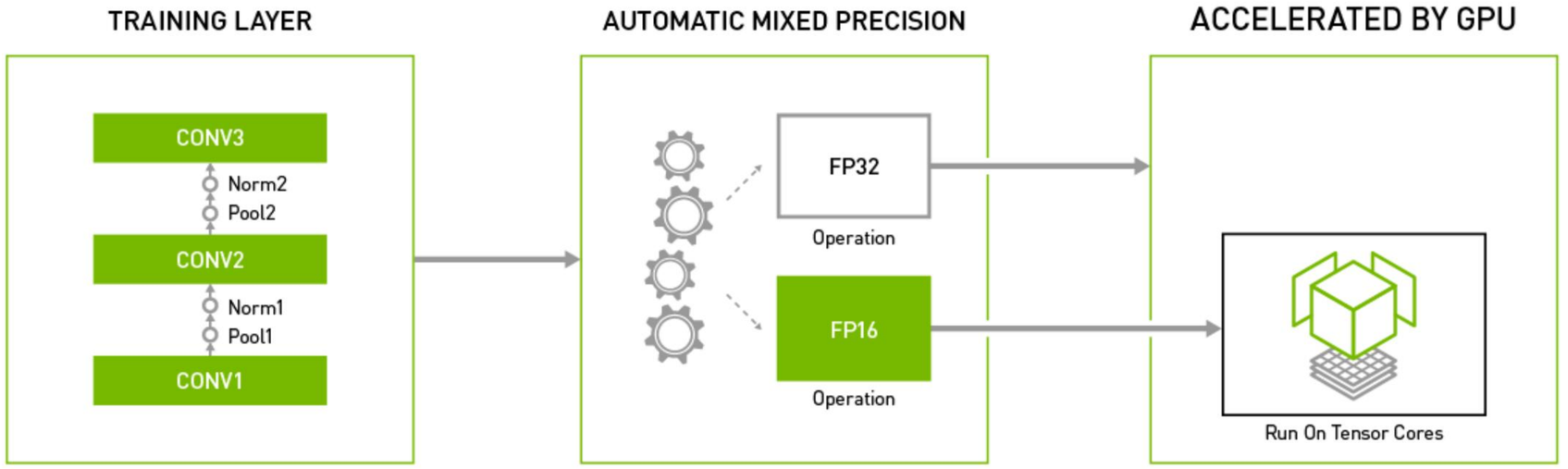


DALI is a high-performance alternative to built-in data loaders and data iterators. You can now run your data processing pipelines on the GPU, reducing the total time it takes to train a neural network. Data processing pipelines implemented using DALI are portable because they can easily be retargeted to TensorFlow, PyTorch, and MXNet.

Training

From CUDA and Python to PyTorch Lightning

Remember to enable automatic mixed-precision!



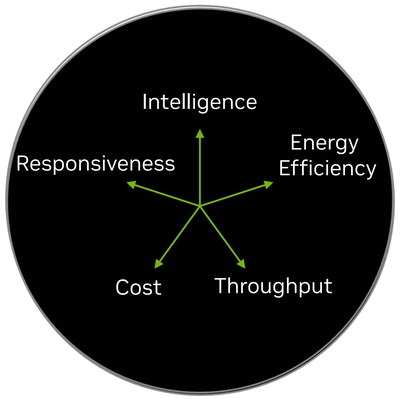
Python
CUDA/C++

The programming languages used to build PyTorch, Lightning Fabric, and PyTorch Lightning.

Inference

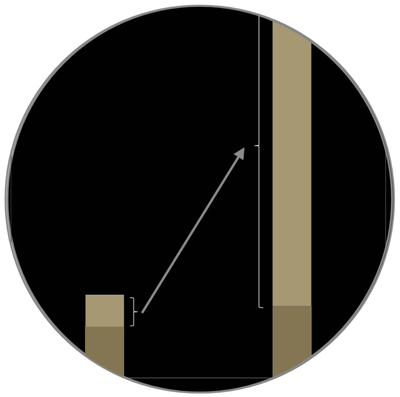
Inference Requires a SMART Platform

Scale and Complexity



Multi-Dimensional Performance

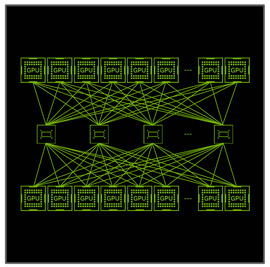
Architecture and Software



ROI Driven by Performance

Technology Ecosystem and Install Base

NVLINK 72



Spectrum-X Ethernet

NVFP4

Dynamo

Blackwell ROI

15x

Open-Source Software

SGL

TensorRT-LLM

vLLM

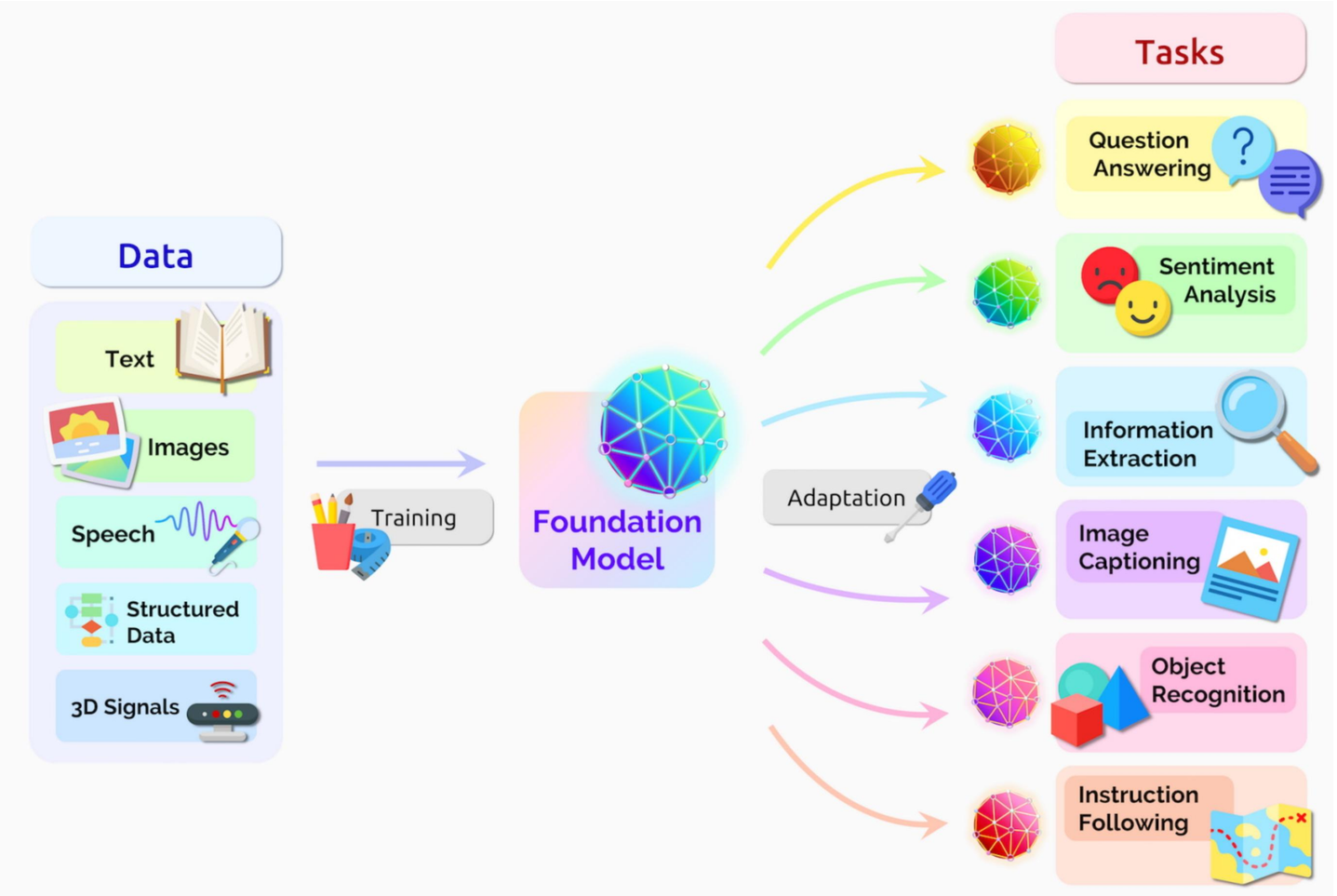
...

AI Use Cases

What are Foundation Models

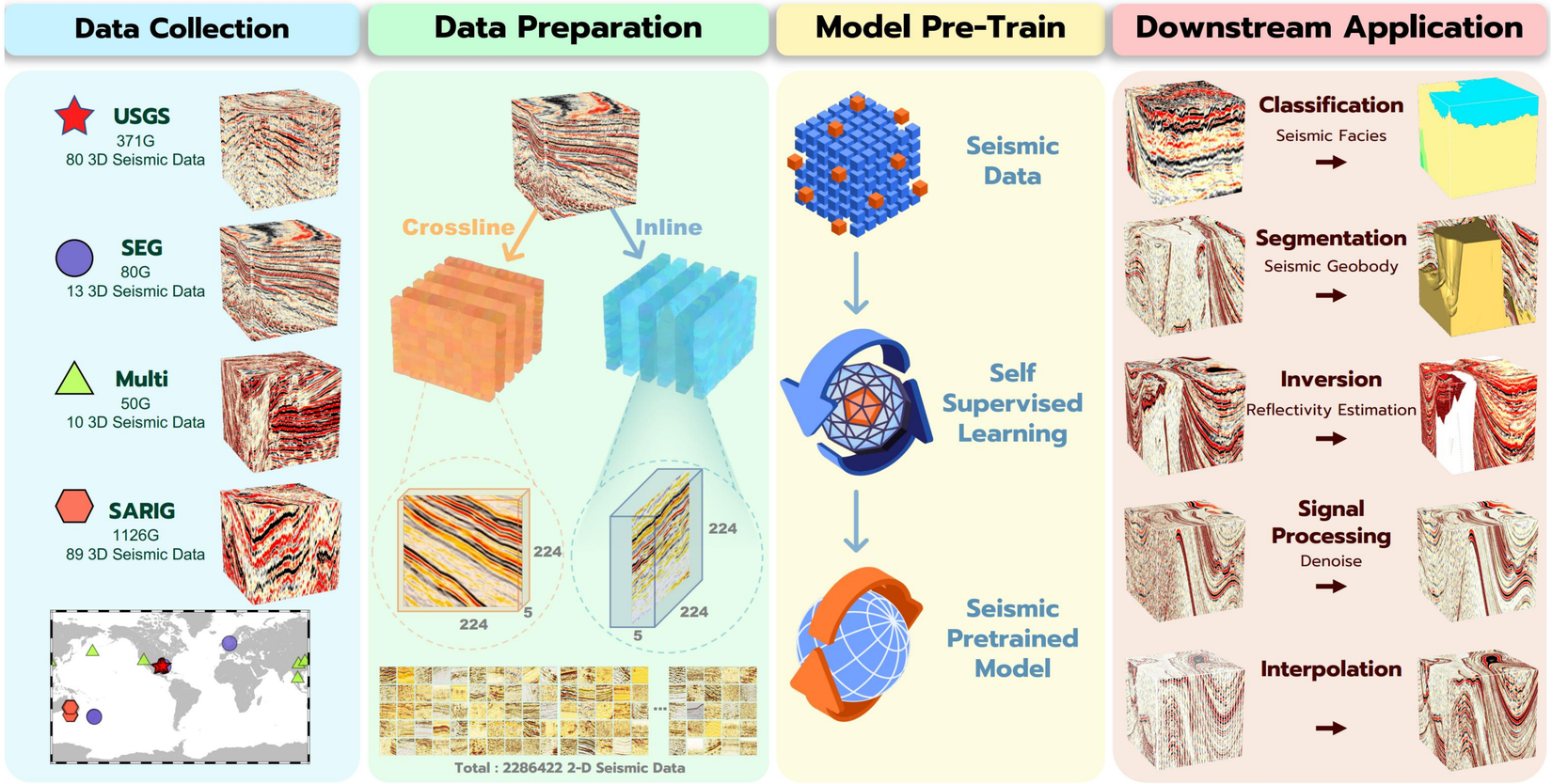
General purpose machine learning models

- Trained on massive unlabeled datasets to handle a wide variety of jobs from translating text to analyzing seismic images.
- General purpose adaptable models that can be fine-tuned for diverse tasks like computer vision, natural language processing with minimal task-specific training.
- Scalable and reusable architectures which leverage learned patterns to perform new tasks efficiently, reducing the need for extensive retraining or custom model development.



Source: What Are Foundation Models? | NVIDIA Blogs

Seismic Foundation Model (SFM)

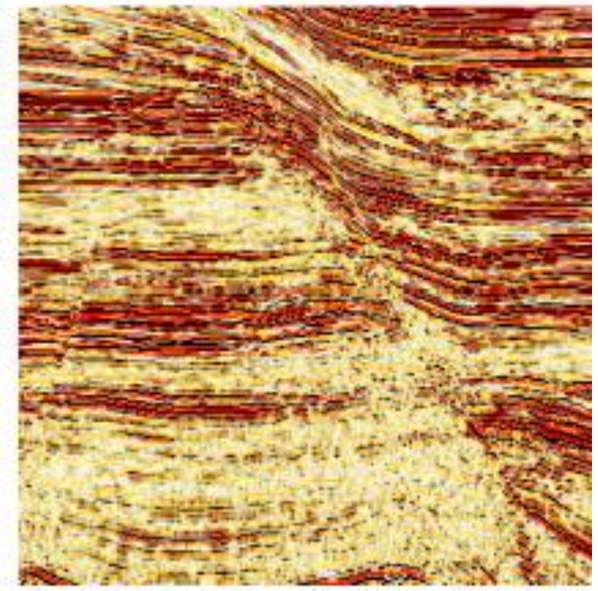


SFM represents the first successful large-scale foundational model trained exclusively within the seismic domain [reference].

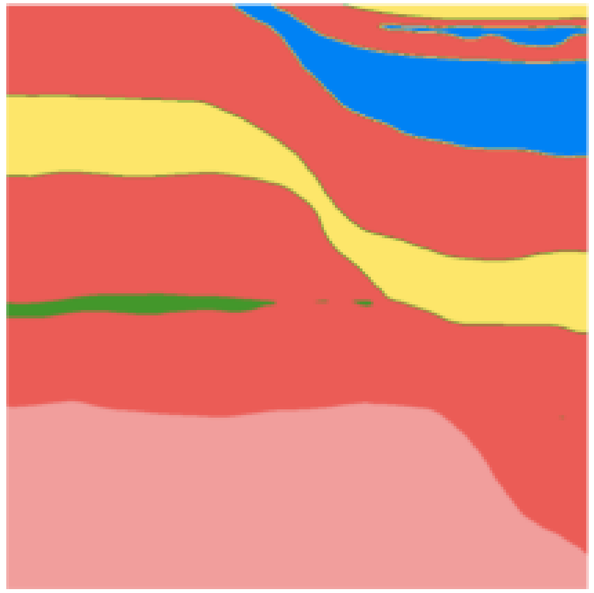
Masked Auto-Encoder fine-tuned on 200 samples

Trained with self-supervised learning on 2.2 million seismic images utilizing NVIDIA TAO

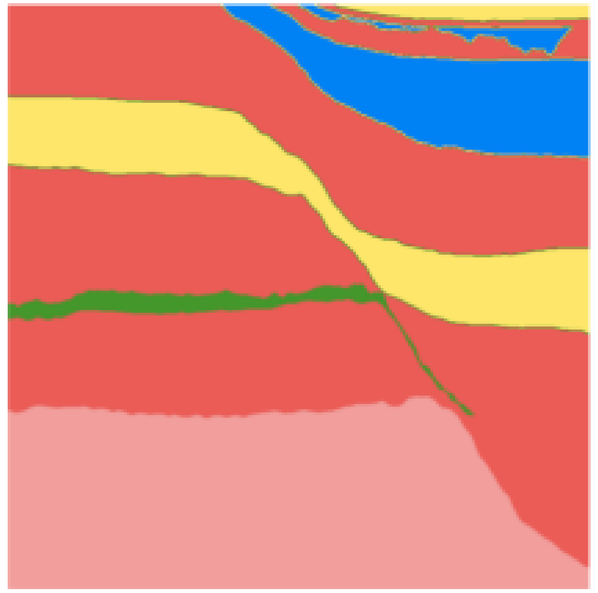
Input Seismic



MAE



Label

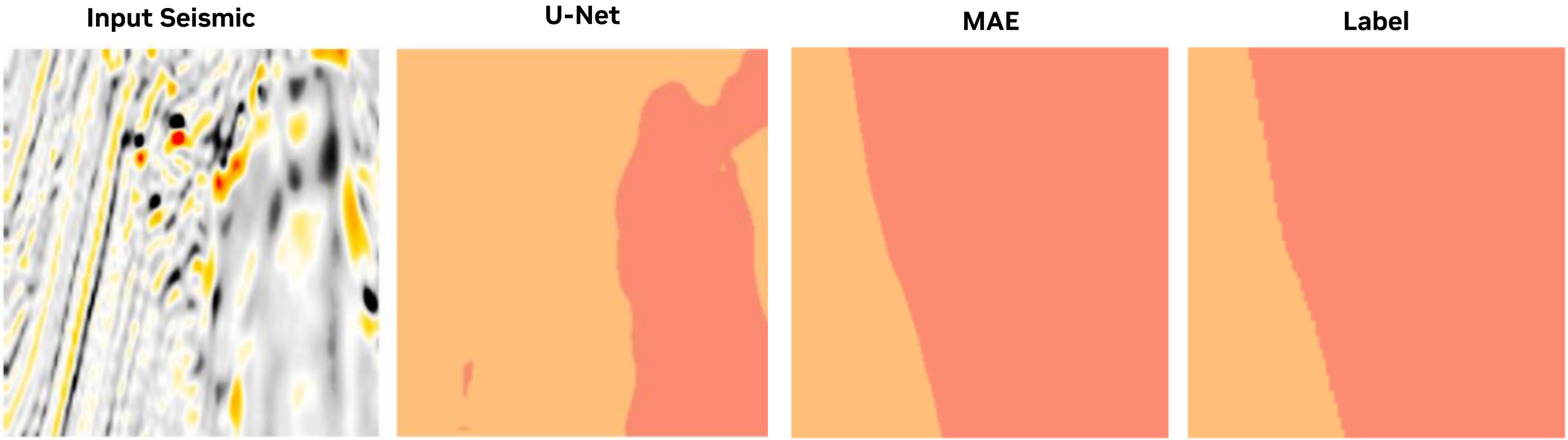


Facies classification

(Results reproduced from Sheng et al. , 2025)
“Seismic Foundation Model (SFM): a new generation deep learning model in geophysics

Masked Auto-Encoder fine-tuned on 3000 samples

Trained with self-supervised learning on 2.2 million seismic images utilizing NVIDIA TAO



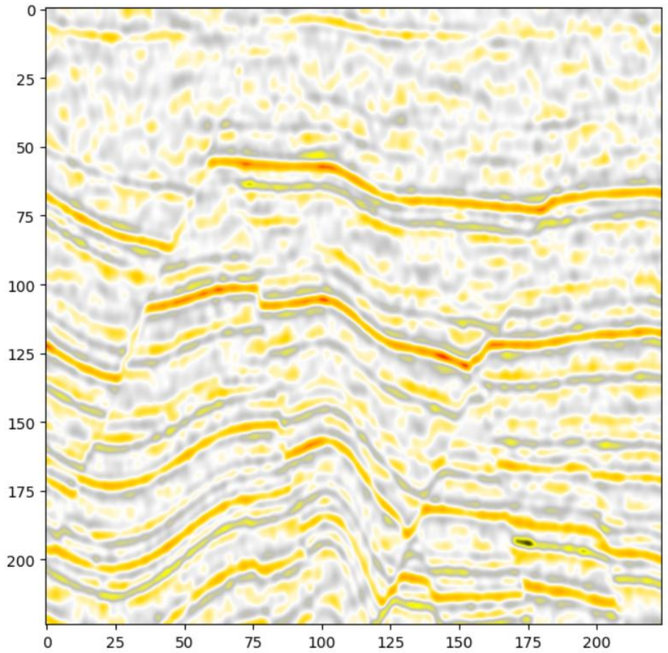
Salt Segmentation

(Results reproduced from Sheng et al. , 2025)
“Seismic Foundation Model (SFM): a new generation deep learning model in geophysics

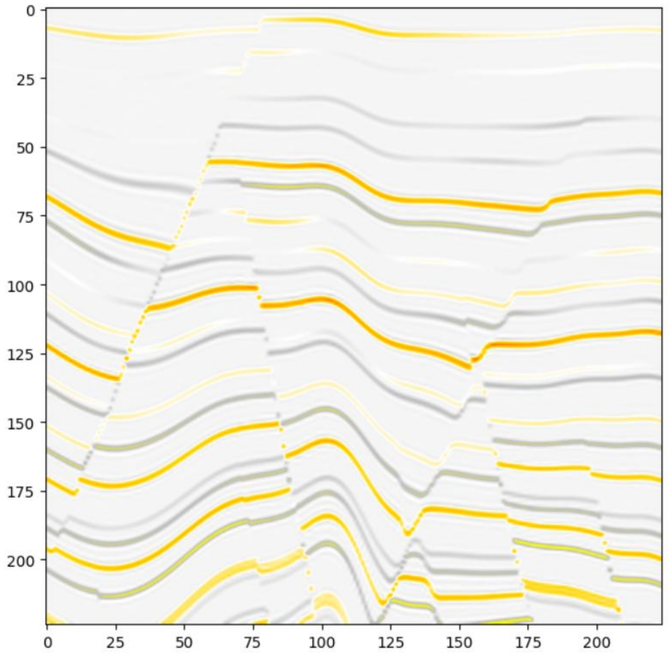
Masked Auto-Encoder fine-tuned on 1000 samples

Trained with self-supervised learning on 2.2 million seismic images utilizing NVIDIA TAO

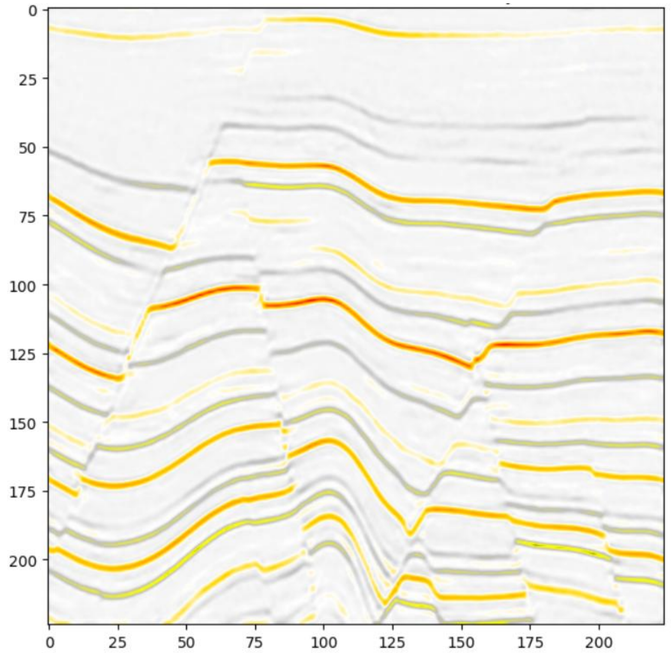
Input Seismic



MAE



Label



Reflectivity inversion

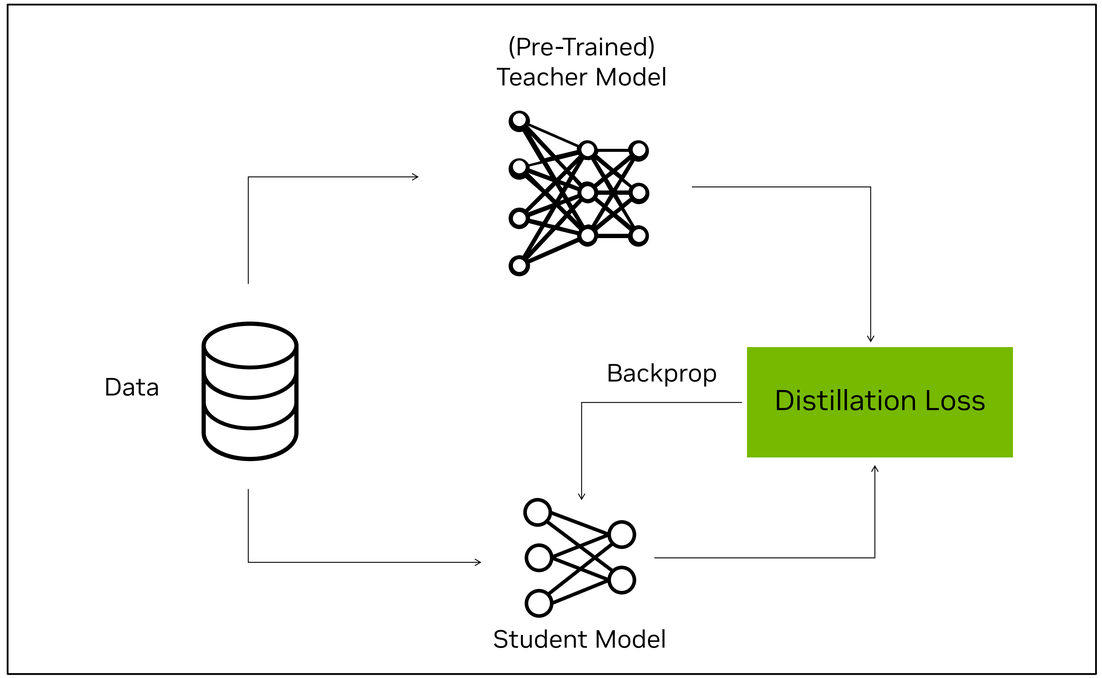
(Results reproduced from Sheng et al. , 2025)

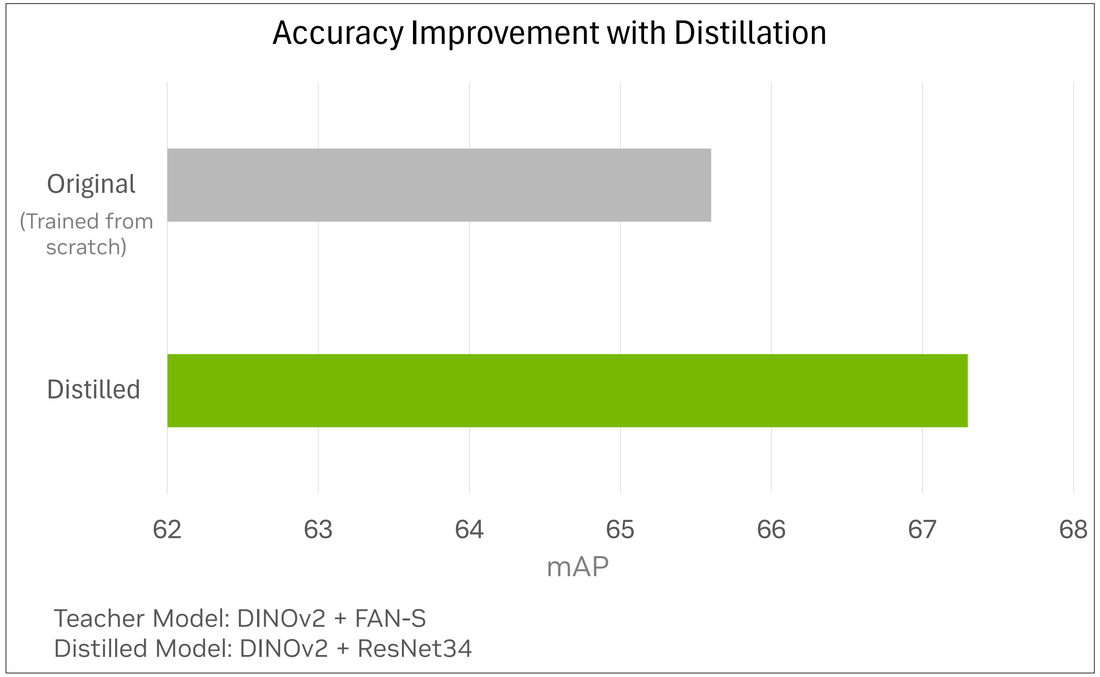
“Seismic Foundation Model (SFM): a new generation deep learning model in geophysics

Knowledge Distillation

Improved accuracy with distillation + task finetuning

Distill knowledge from larger teacher network to smaller student model





Key Benefits:

Reduce Model Size

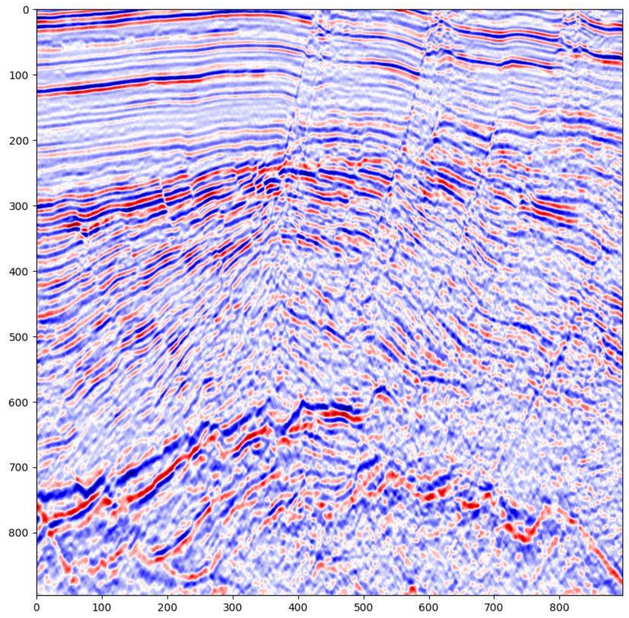
Increase Model Accuracy

Faster Training

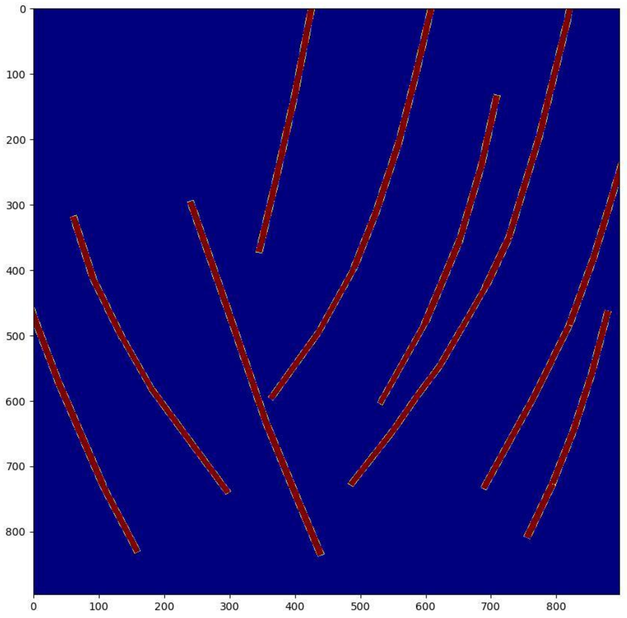
DINOv2 Self-Supervised Learning Enhanced by Knowledge Distillation and Task-Specific Fine-Tuning

Initially trained on 2.2 million images, followed by distillation and fine-tuning with 100 samples.

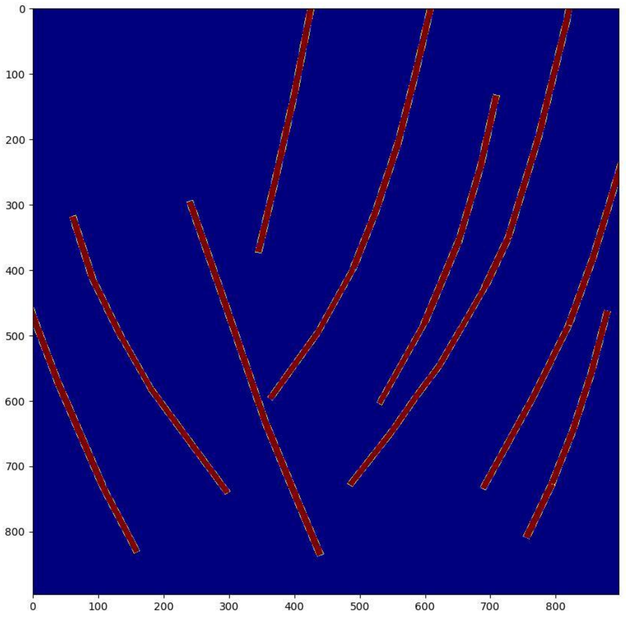
Input Seismic



DINOv2 (distilled)



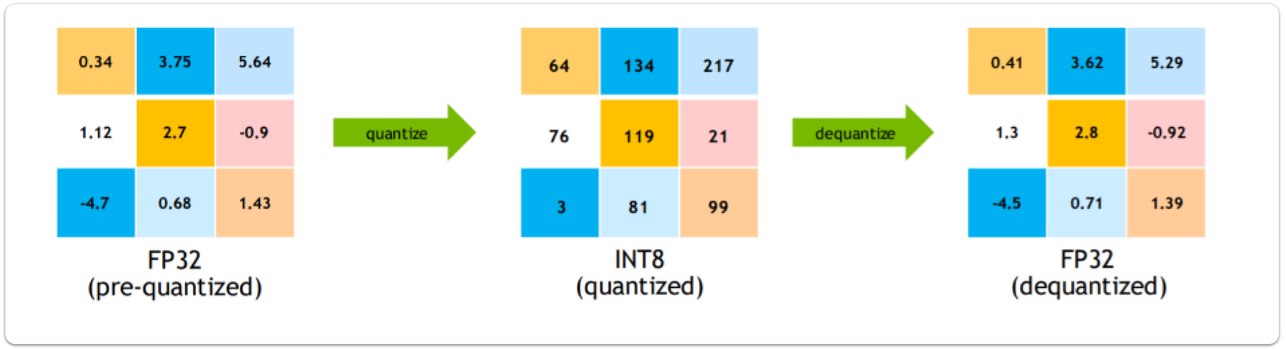
Label



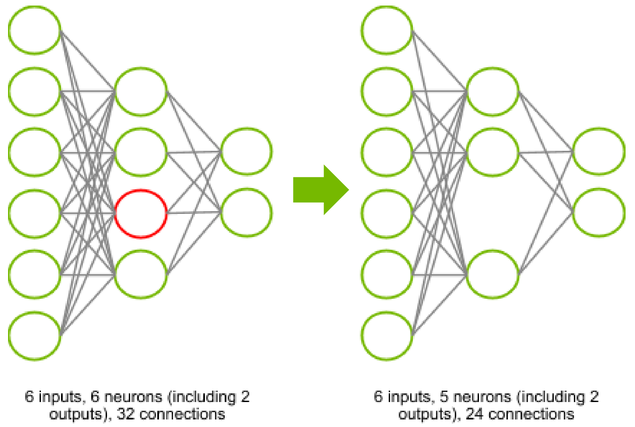
Quantization and Pruning

Improved accuracy with distillation + task finetuning

- TAO will support quantization and pruning for foundation model
- FP8, INT8/INT4 Weight only, INT8 Smooth Quant, AWQ, GPTQ quantization will be supported
- Scaling laws and ”Train large and compress more” strategy.



Quantization

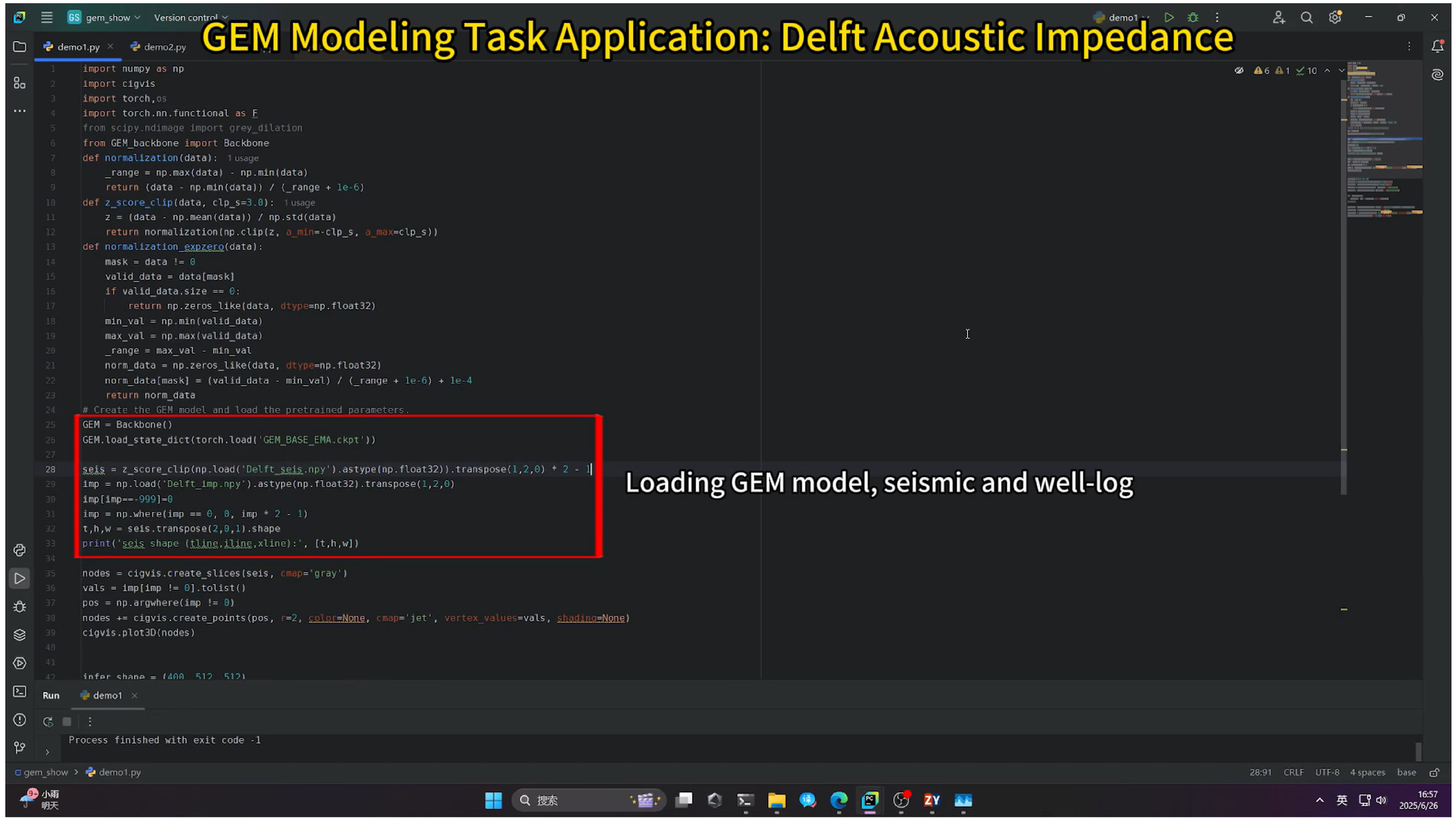


Pruning

Geological Everything Model 3D

Geological Everything Model 3D

A Promptable Foundation Model for Unified and Zero-Shot Subsurface Understanding



Geological Everything Model 3D: A Promptable Foundation Model for Unified and Zero-Shot Subsurface Understanding

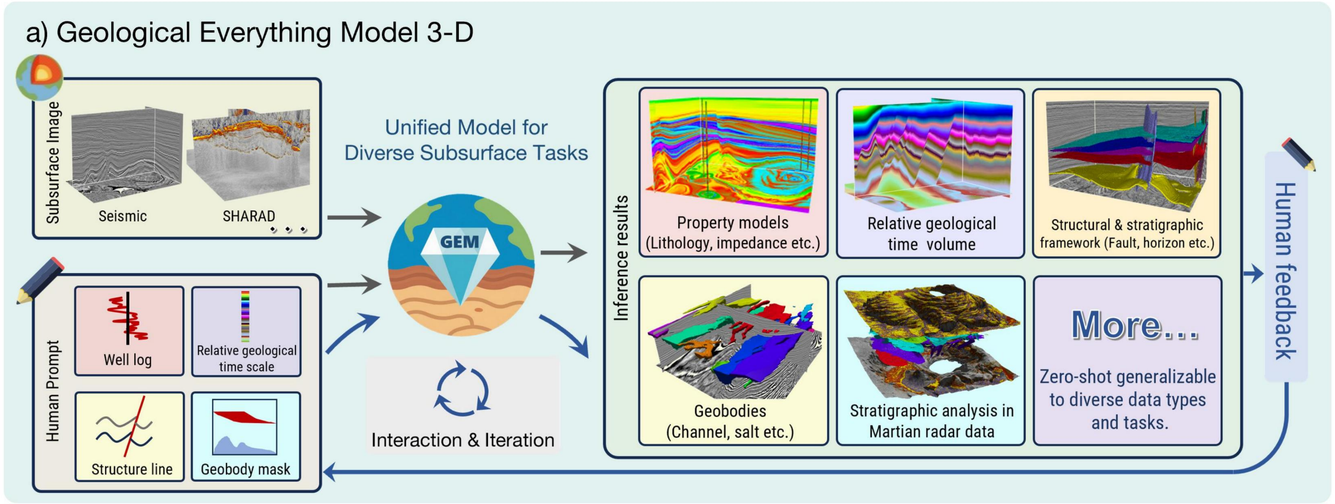
Geological Everything Model 3D

A Promptable Foundation Model for Unified and Zero-Shot Subsurface Understanding

Geological Everything Model 3D: A Promptable Foundation Model for Unified and Zero-Shot Subsurface Understanding

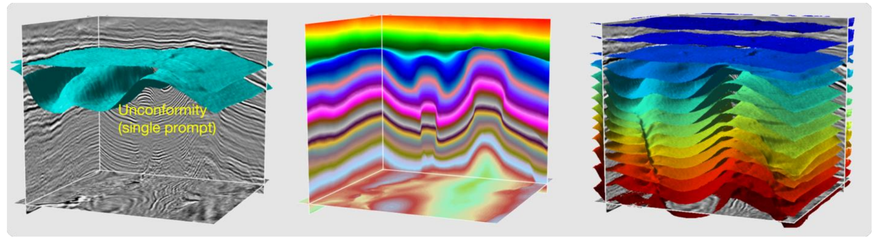
Geological Everything Model 3D

A Promptable Foundation Model for Unified and Zero-Shot Subsurface Understanding

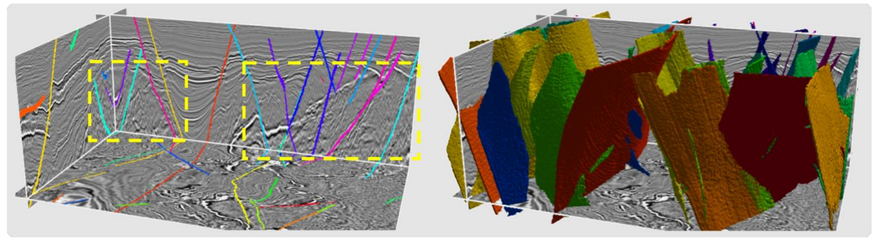


Inference -> $\hat{\mathbf{Y}} = \mathbf{p}(\mathbf{Y} \mid \mathbf{X}_{\text{img}}, \mathbf{Y}_{\text{part}})$

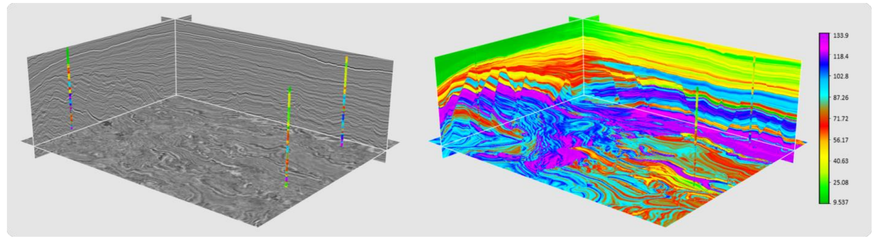
Delft Unconformity and RGT



Baiyun Fault



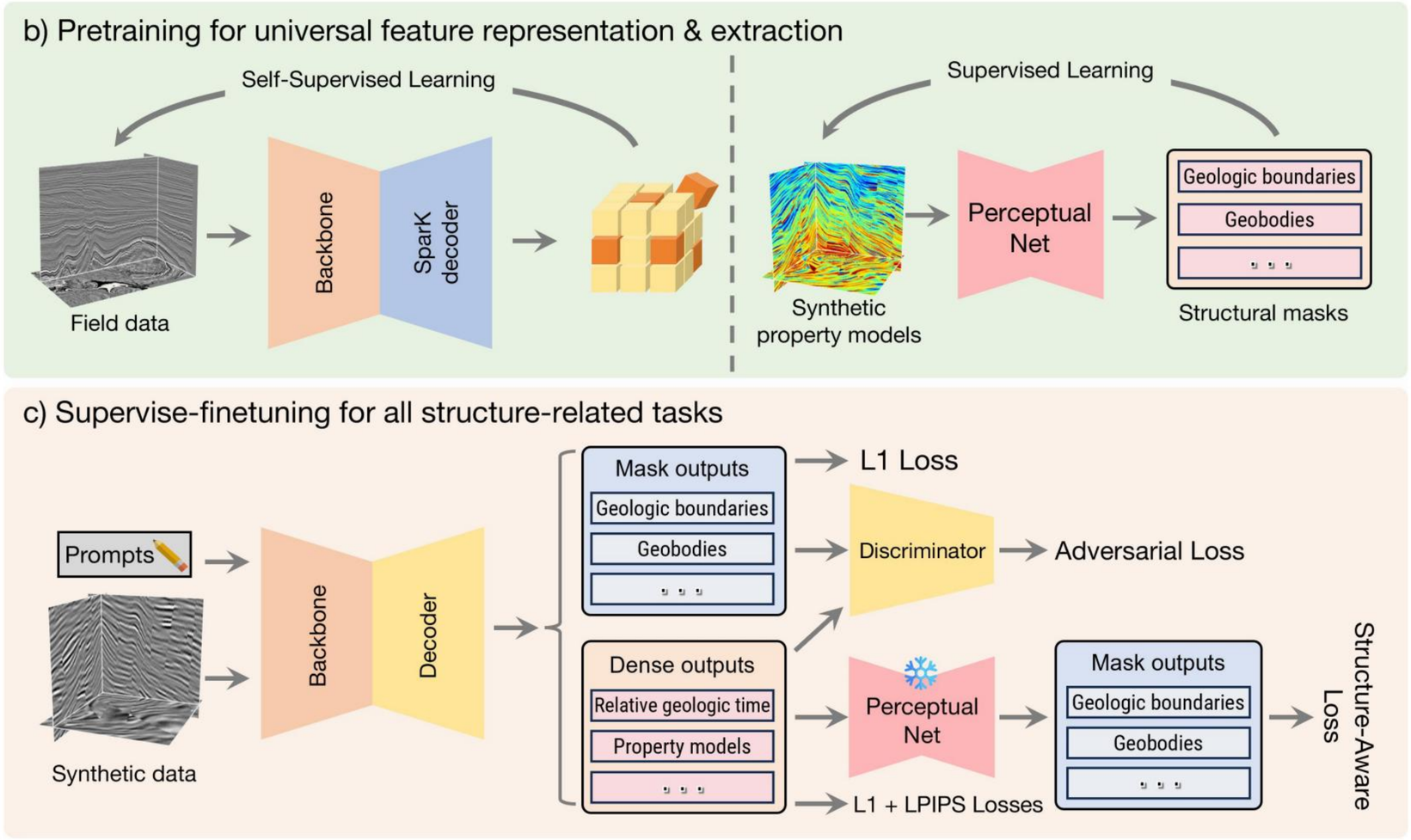
Poseidon GammaRay



Geological Everything Model 3D: A Promptable Foundation Model for Unified and Zero-Shot Subsurface Understanding

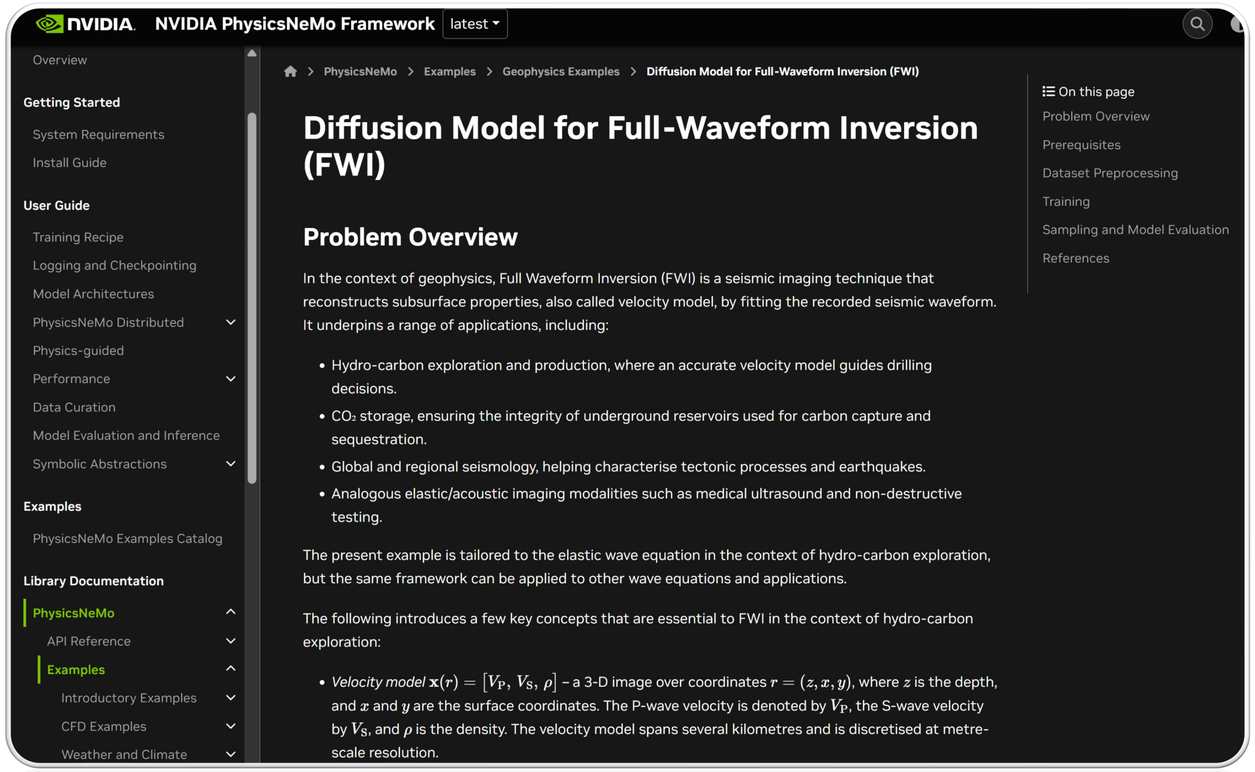
Geological Everything Model 3D

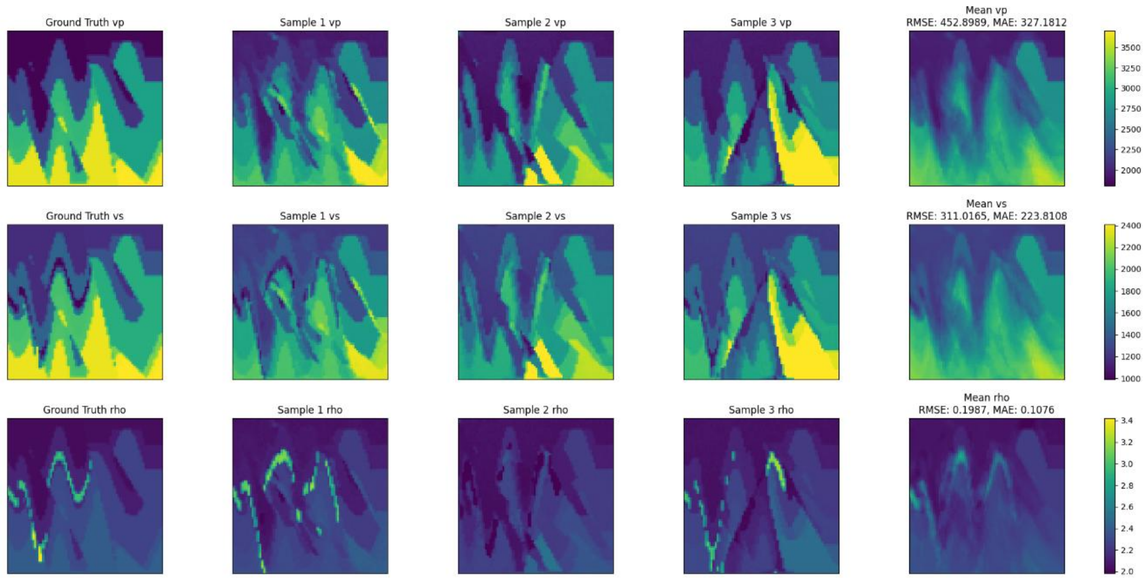
A Promptable Foundation Model for Unified and Zero-Shot Subsurface Understanding



Geological Everything Model 3D: A Promptable Foundation Model for Unified and Zero-Shot Subsurface Understanding

Try Diffusion Model for Full-Waveform Inversion Using NVIDIA PhysicsNeMo





https://docs.nvidia.com/physicsnemo/latest/physicsnemo/examples/geophysics/diffusion_fwi/README.html