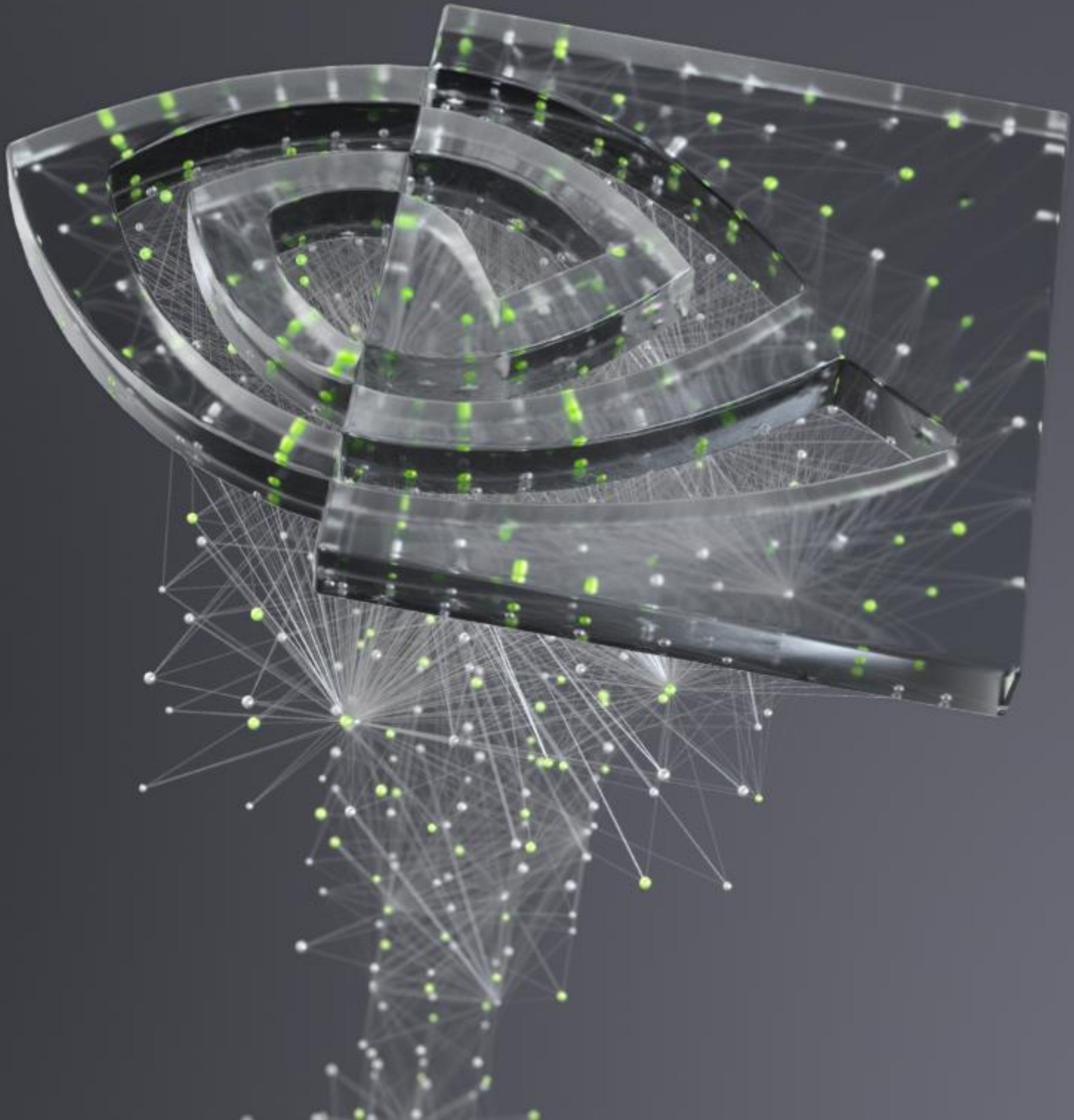
The NVIDIA Technology Stack for AI/ML



Pillars of NVIDIA AI Platform

Data Analytics & Training

HugeCTR mxnet ONNX

PaddlePaddle PYTORCH

RAPIDS Spark TensorFlow

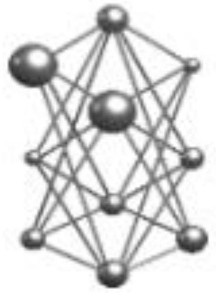
CUDA-X-AI

CUDA

MAGNUM IO

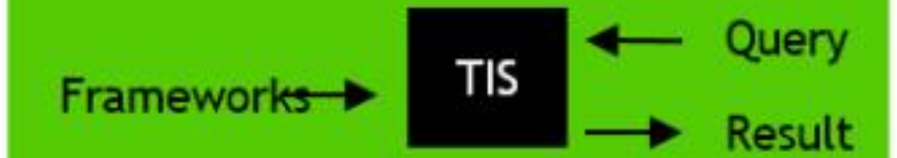
Inference

NGC Catalog Pre-Trained Models

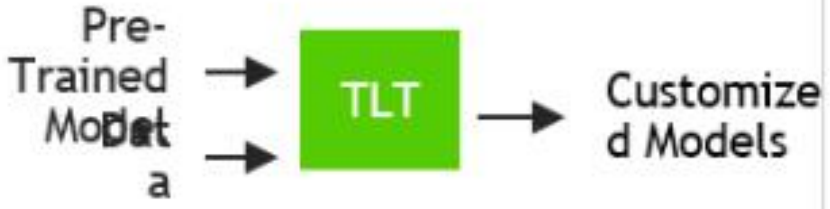


- Conversational AI
- Computer Vision
- Recommendations
- Reinforcement Learning

Triton Inference Serving



TAO Zero Coding | Speech, Vision & NLU



TensorRT 8.2

Application Frameworks

CLARA Healthcare

DRIVE Autonomous Vehicles

RIVA Conversational AI

ISAAC Robotics

Merlin Recommendation Systems

METROPOLIS Smart City

NVIDIA NGC Catalog

AI FRAMEWORKS

PyTorch, TensorFlow…

AI TOOLKITS

TAO, TensorRT, Triton…

MODELS

BERT, Transformer…

MODEL SCRIPTS

Jupyter Notebooks

COLLECTIONS

Computer Vision, Speech…

Computer Vision

- Traffic Analysis
- Gesture Recognition
- Medical Imaging

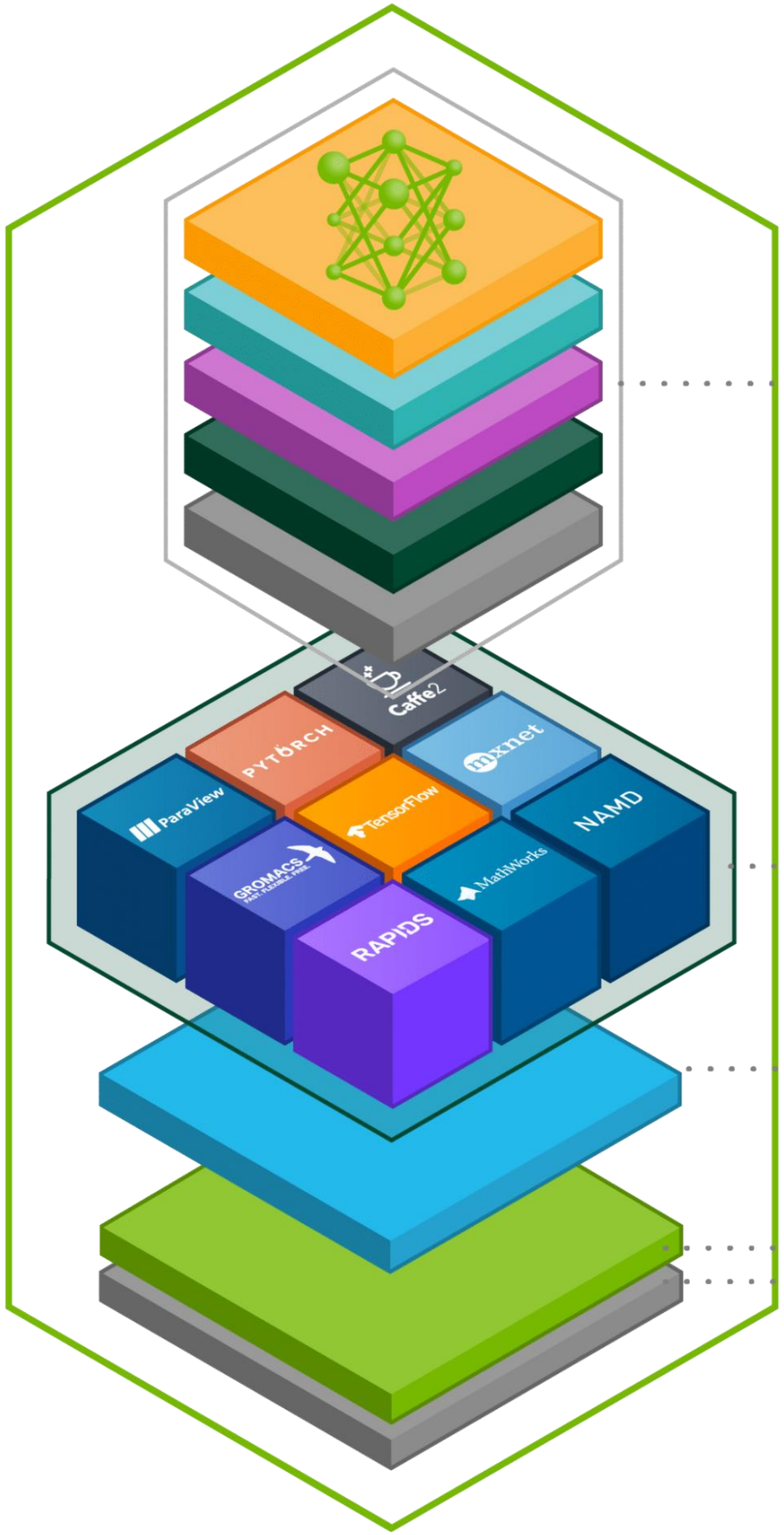
Conversational AI

- Chat bots
- Translation
- Music composition

Recommendation

- Page rankings
- Personalized shopping
- Music & movie suggestions

NVIDIA NGC



PRE-TRAINED MODELS

DEEP LEARNING, MACHINE LEARNING,
HPC APPLICATION CONTAINERS

CONTAINER RUNTIME

NVIDIA DRIVER

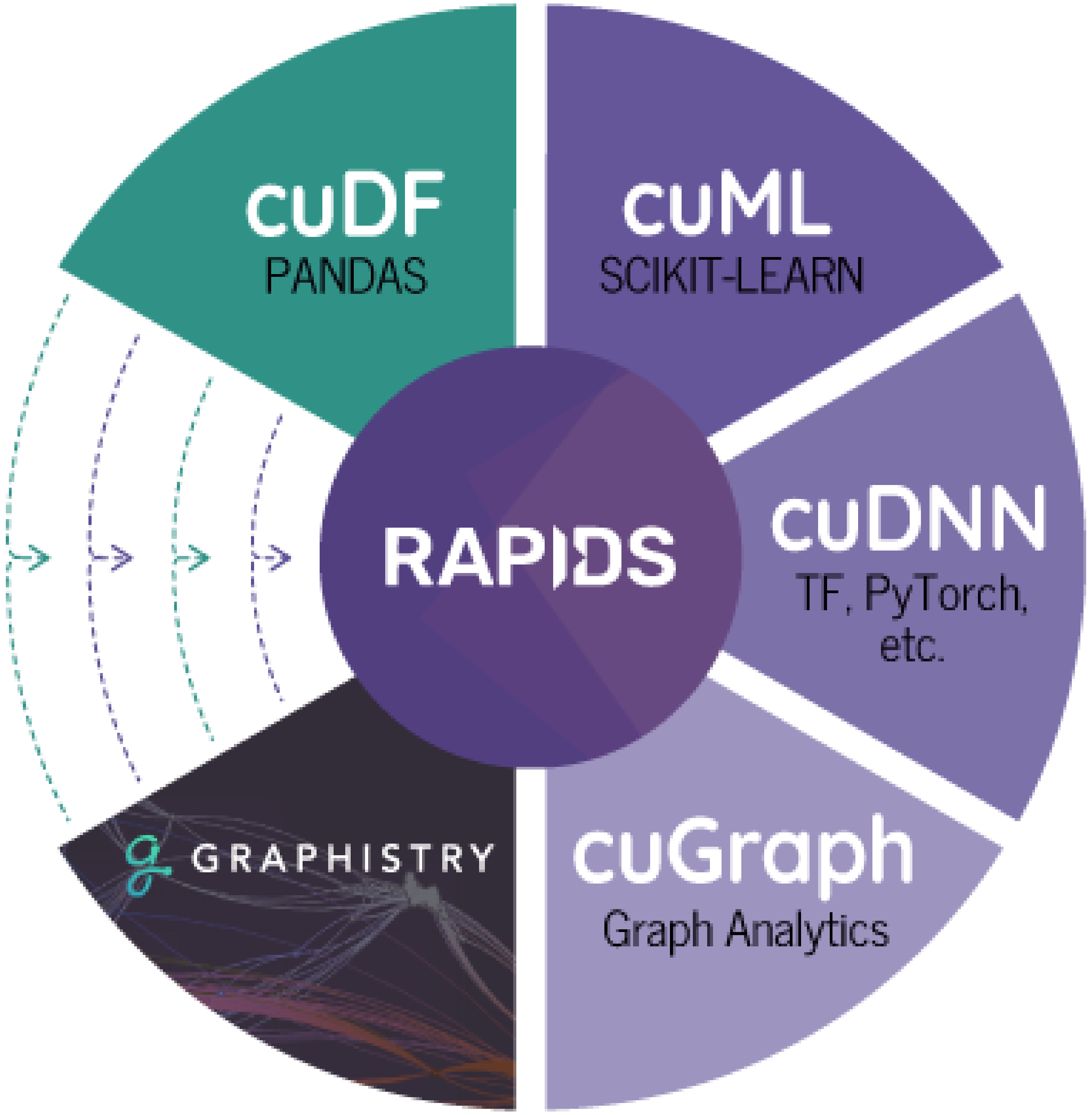
HOST OS

NGC SOFTWARE STACK

NVIDIA RAPIDS

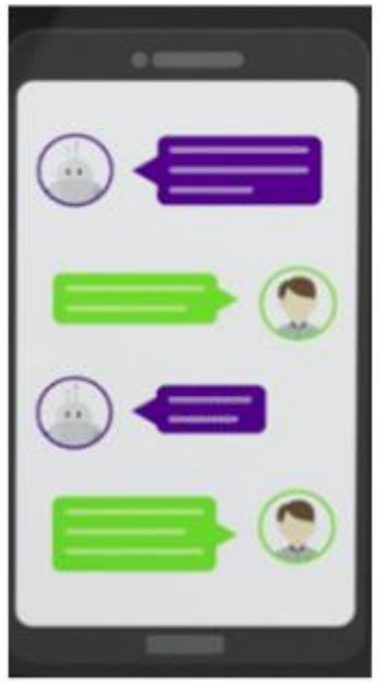
Data Science on GPUs

- Supports end to end pipeline:
  - Data exploration
  - Data preparation
  - Traditional ML Algorithms
  - Graph Algorithms
  - Parallel computing, HPC etc
- URL - *rapids.ai*



NVIDIA APPLICATION FRAMEWORKS

Speech AI



Riva

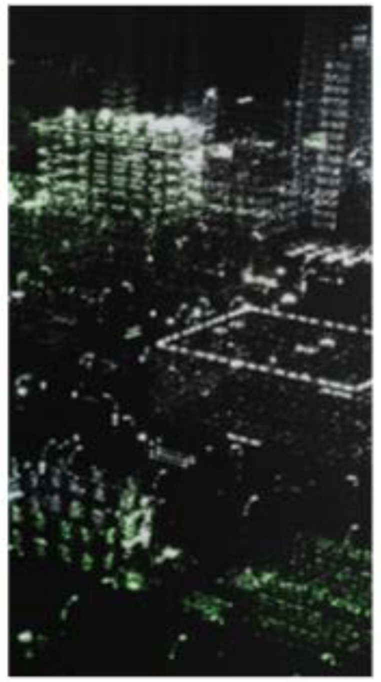
Big NLP

NeMo

Recommender Systems

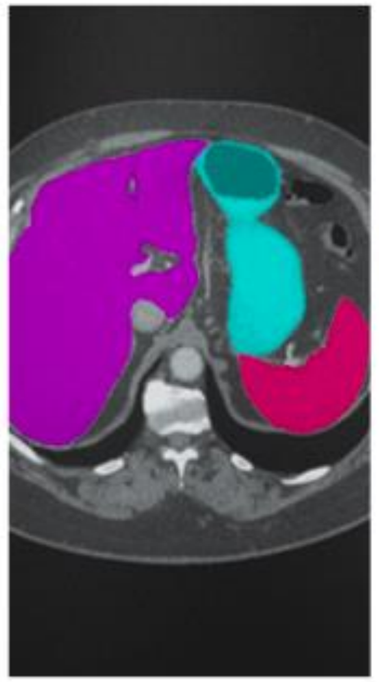
Merlin

Smart Cities



Metropolis

Healthcare



Clara

Robotics

Isaac

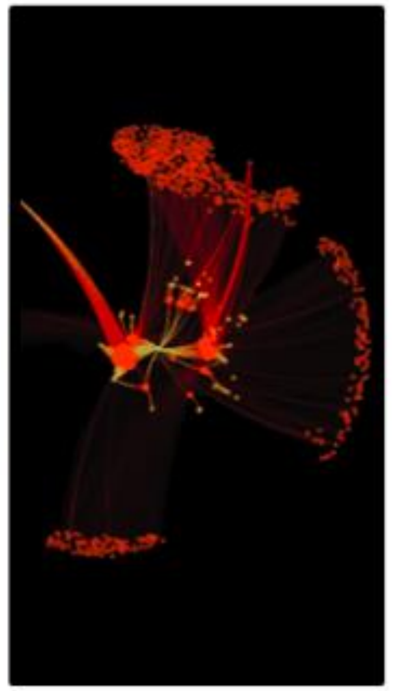
Autonomous Vehicles

Drive

Telecom

Aerial

Cybersecurity



Morpheus

Desktop Development

Data Center Solutions

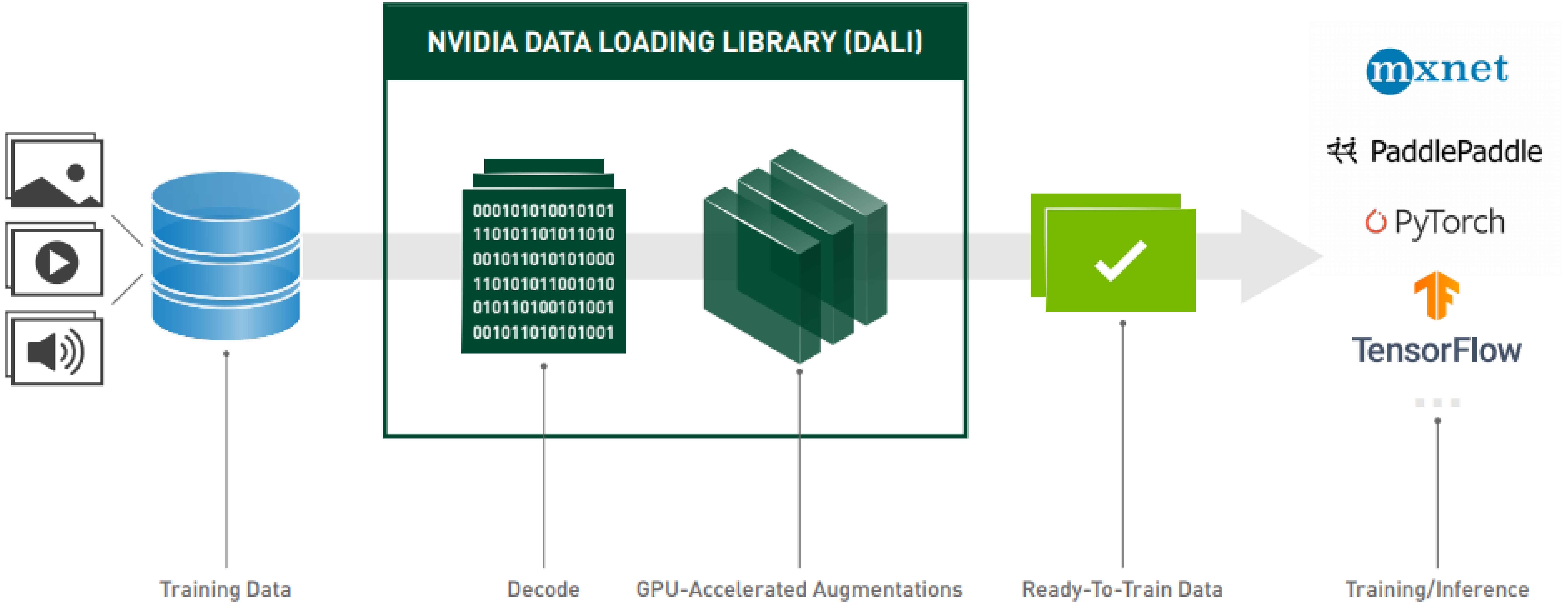
Accelerated Edge

Supercomputers

GPU-Accelerated Cloud

NVIDIA DALI

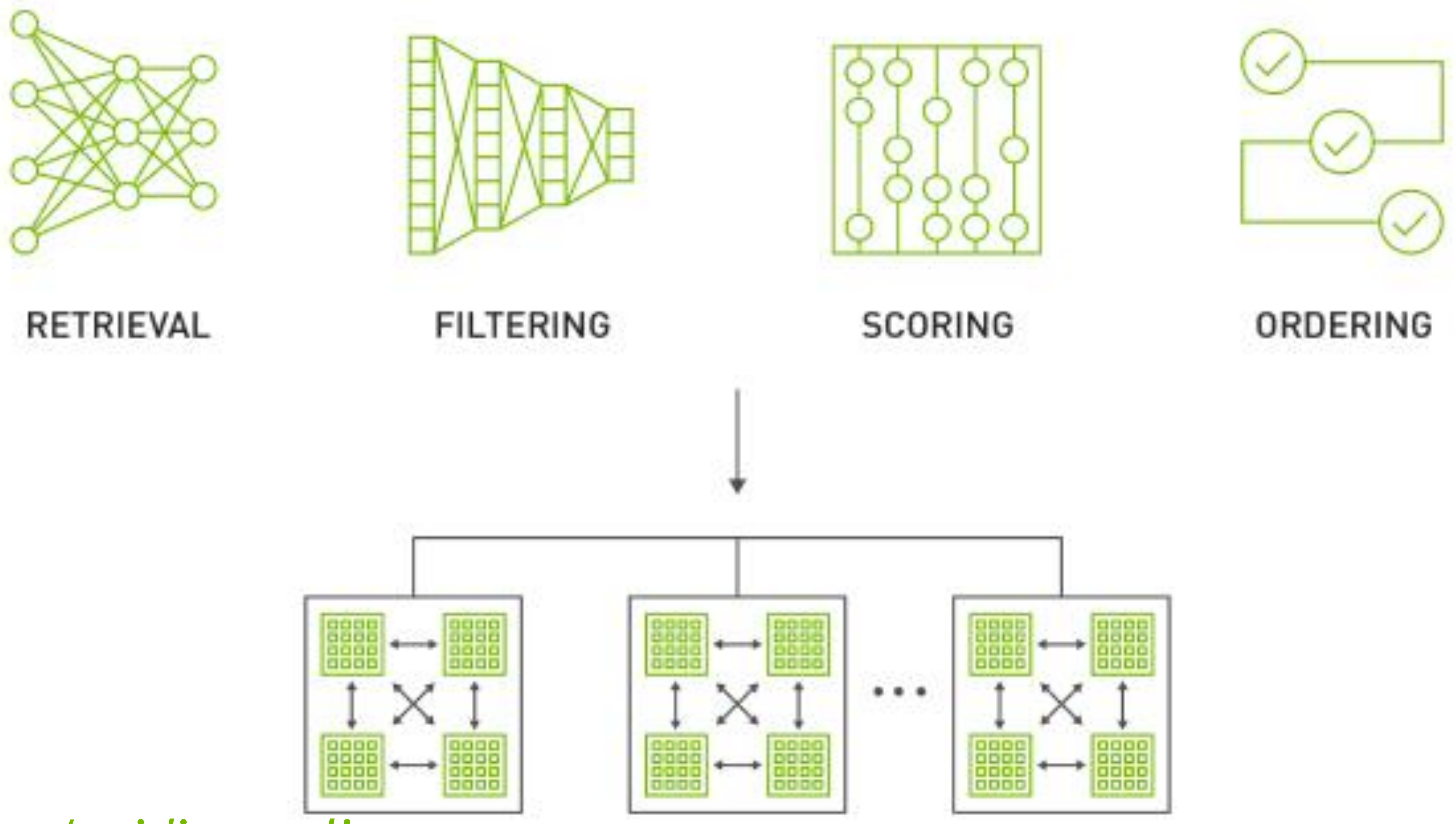
High Performance Data Loading Library



URL -

NVIDIA Merlin

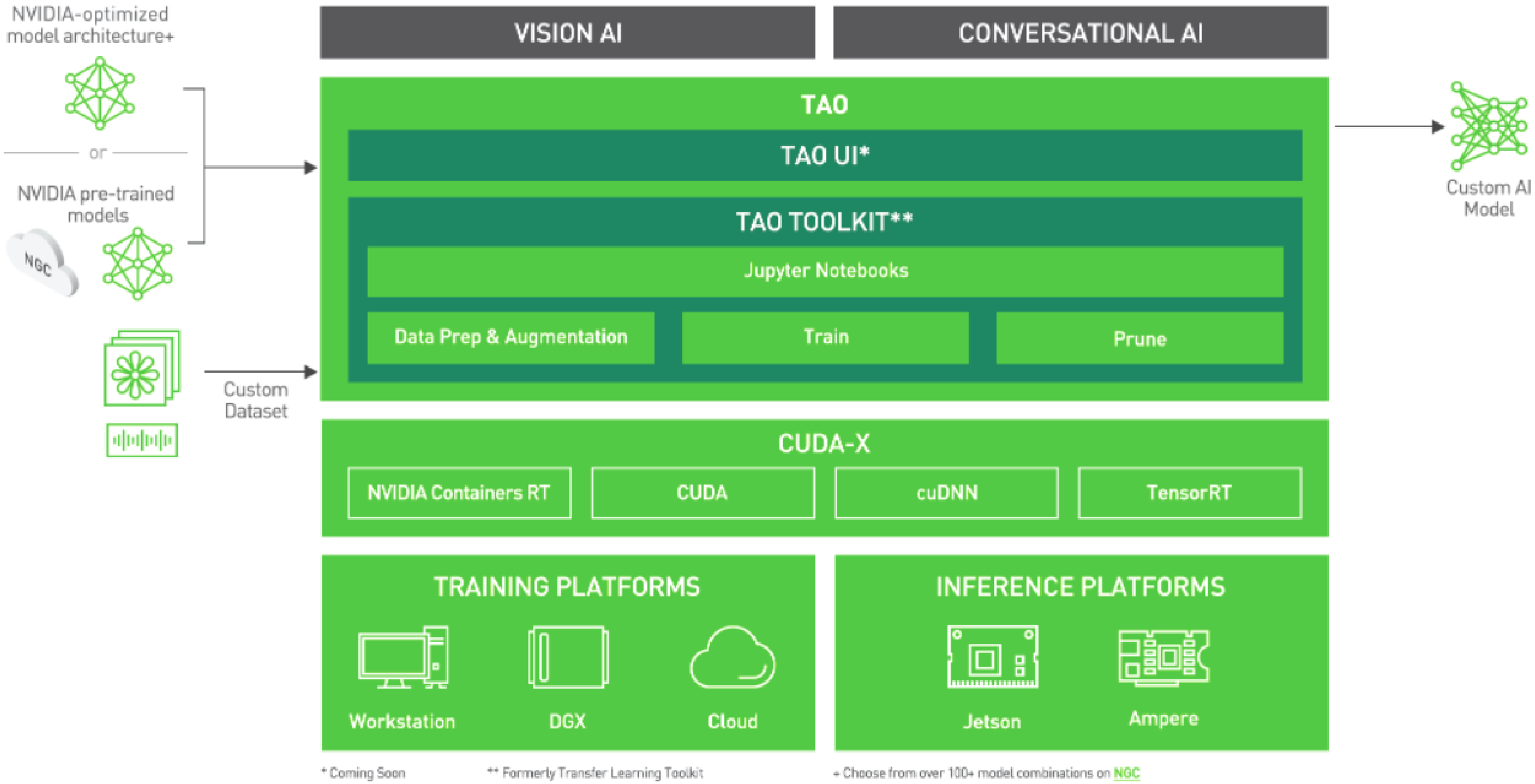
High performance recommender system builder



URL -
developer.nvidia.com/nvidia-merlin

NVIDIA TAO

TAO - Train, Adapt and Optimize. Transfer Learning Toolkit



* Coming Soon

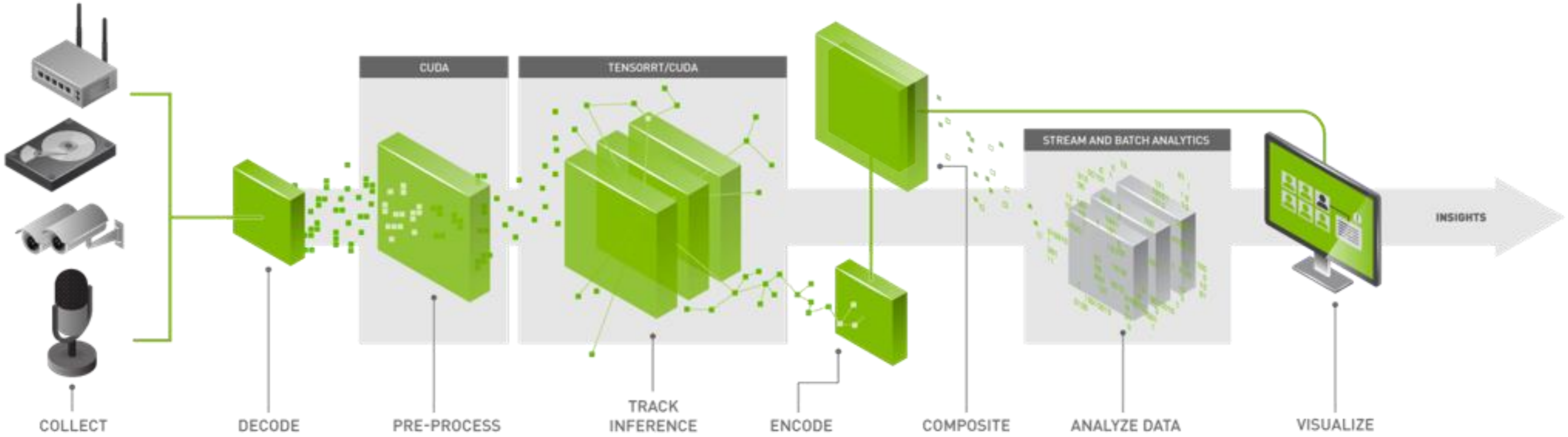
** Formerly Transfer Learning Toolkit

+ Choose from over 100+ model combinations on NGC

URL - *developer.nvidia.com/ tao-toolkit*

NVIDIA Deepstream

Rapidly develop and deploy Vision AI applications and services

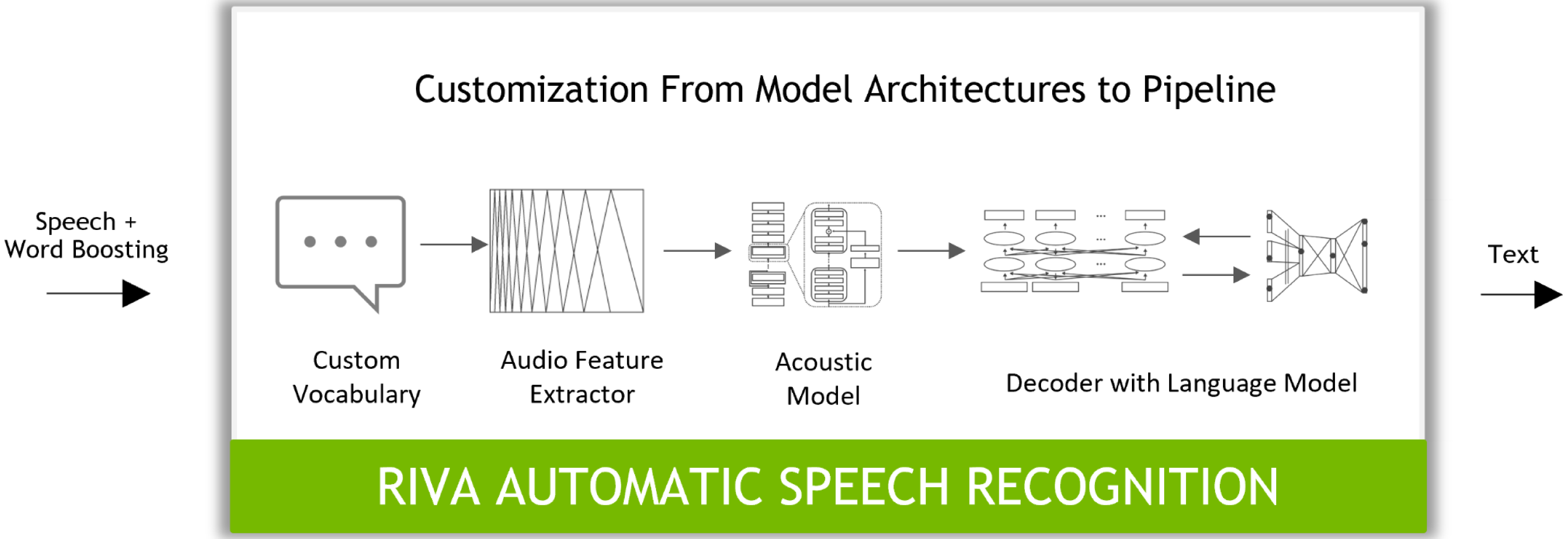


URL -

https://developer.nvidia.com/deepstream-sdk

NVIDIA RIVA

A GPU-accelerated SDK for building word class Speech AI applications



Customization From Model Architectures to Pipeline

Speech + Word Boosting

Custom Vocabulary

Audio Feature Extractor

Acoustic Model

Decoder with Language Model

Text

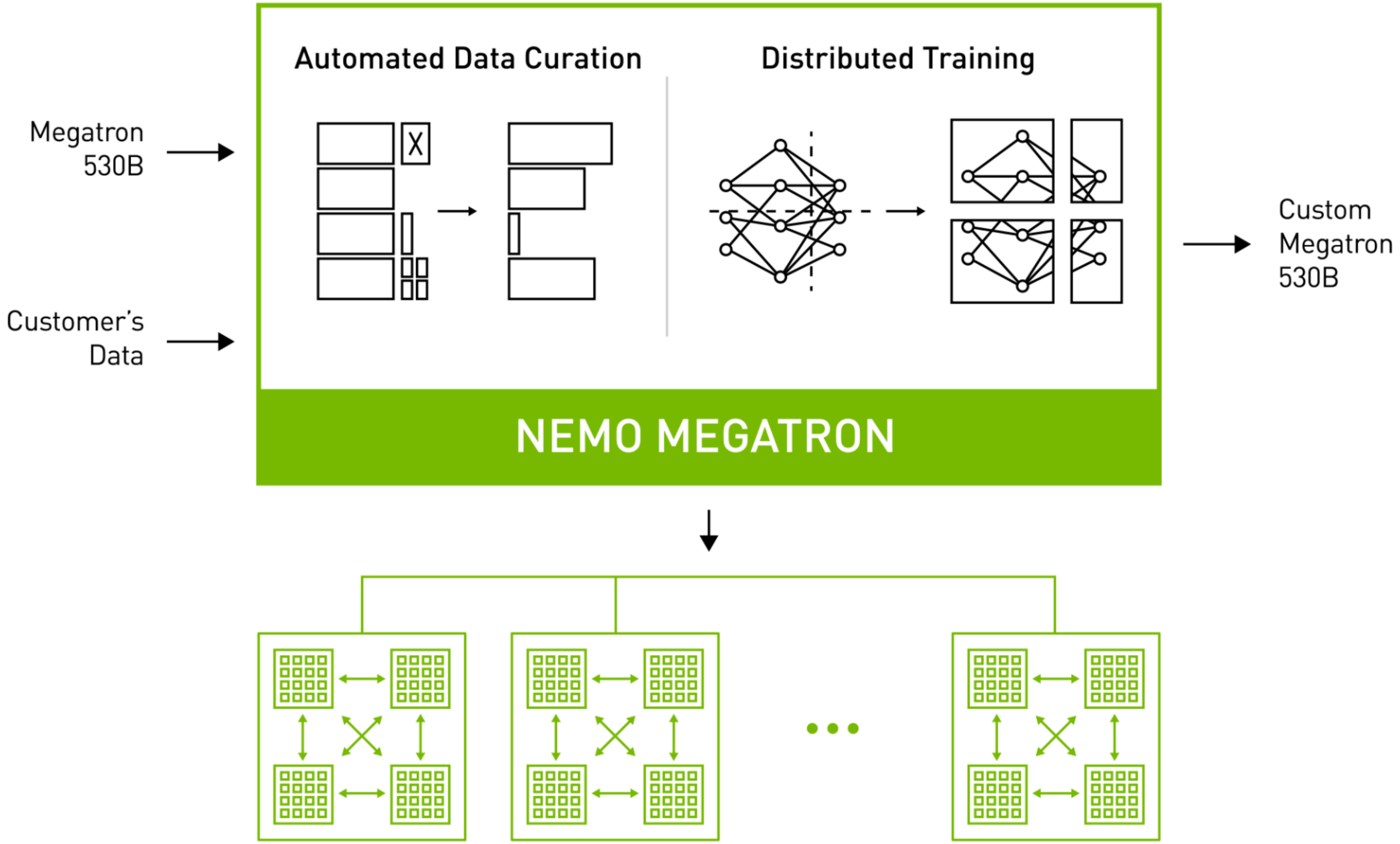
RIVA AUTOMATIC SPEECH RECOGNITION

URL -

https://developer.nvidia.com/deepstream-sdk

NVIDIA NEMO

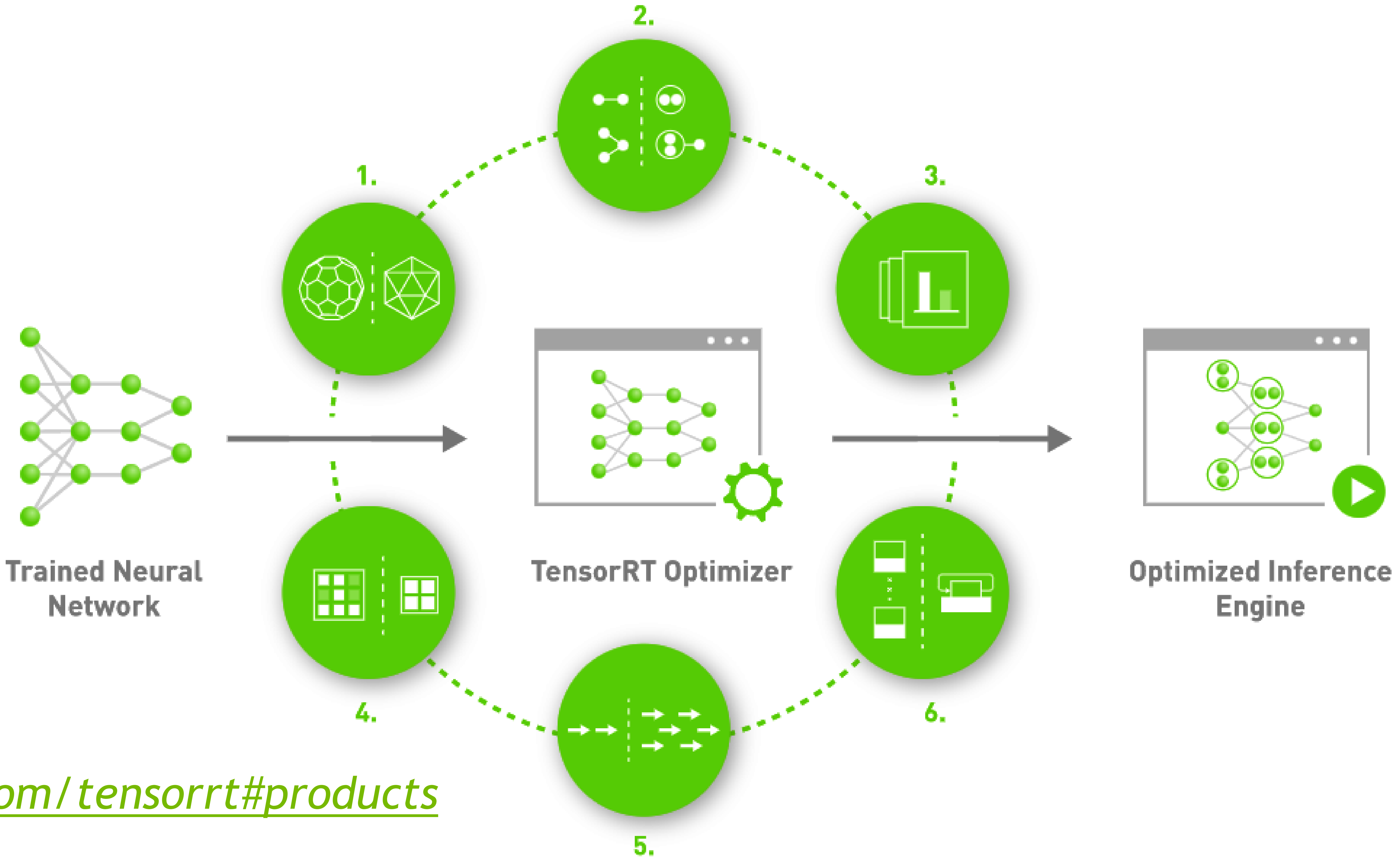
A GPU-accelerated SDK for building state-of-the-art conversational AI and NLP models.



URL - *https://developer.nvidia.com/nvidia-nemo*

NVIDIA TensorRT

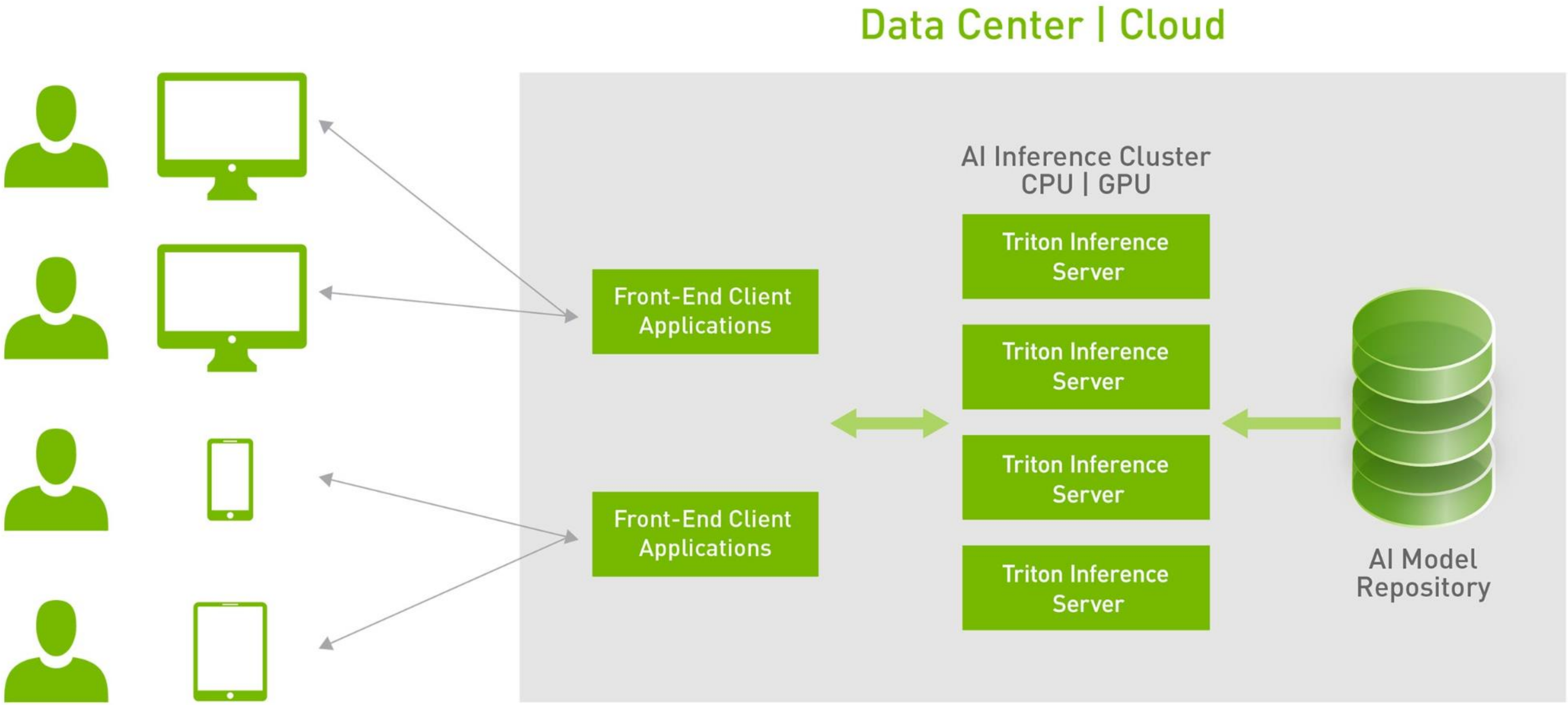
Optimizer for high performance deep learning inference



URL - *developer.nvidia.com/tensorrt#products*

NVIDIA Triton Inference Server

Fast and Scalable AI Deployment For Any AI Application



URL -

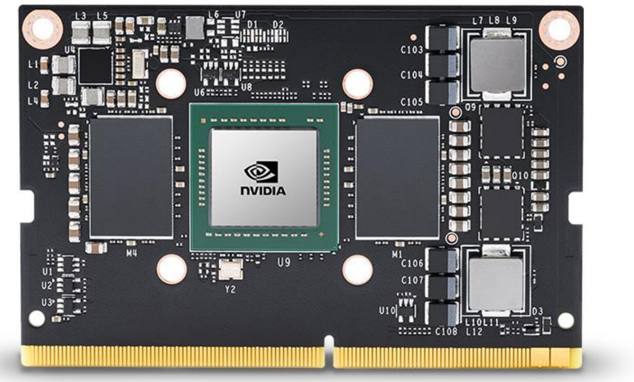
https://developer.nvidia.com/nvidia-triton-inference-server

NVIDIA Edge Inference Devices

High Performance Embedded AI Inference

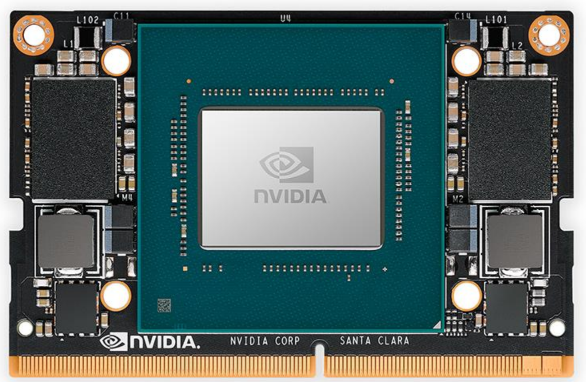
Jetson Nano

(472 GFLOPS, 128 Maxwell Cores)



Jetson TX2

(1.33 TFLOPS, 256 Pascal Cores)

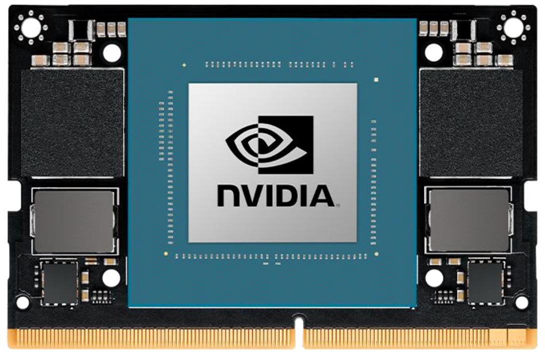


Jetson Xavier NX

(21 TOPS, 384 Volta + 48 Tensor Cores)

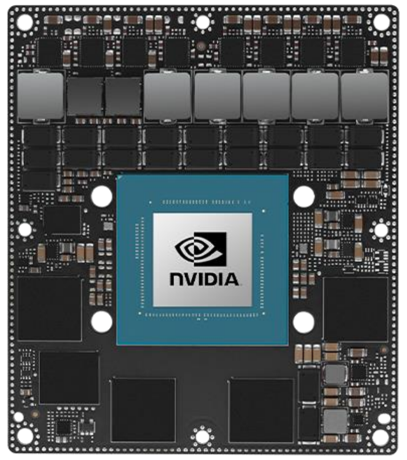
Jetson AGX Xavier

(32 TOPS, 512 Volta + 64 Tensor Cores)



Jetson Orin NX

(100 TOPS, 1024 Ampere + 32 Tensor Cores)



Jetson AGX Orin

(275 TOPS, 2048 Ampere + 64 Tensor Cores)

URL -

https://www.nvidia.com/en-in/autonomous-machines/embedded-systems/