ACCELERATED COMPUTING WITH NVIDIA GPUS

Jesse Tetreault, Solutions Architect

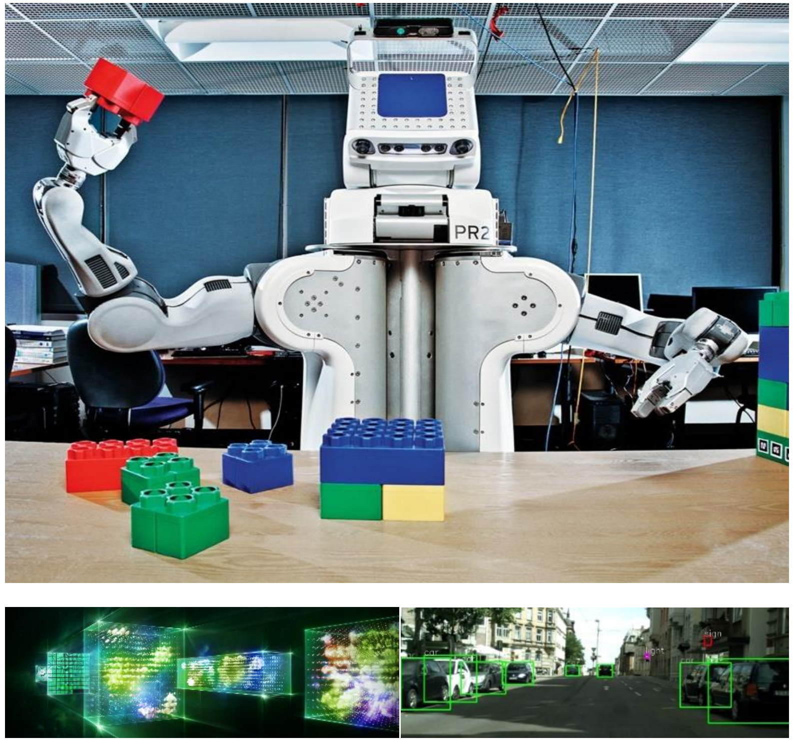
October 2019

ACCELERATED COMPUTING

NVIDIA
"THE AI COMPUTING COMPANY"

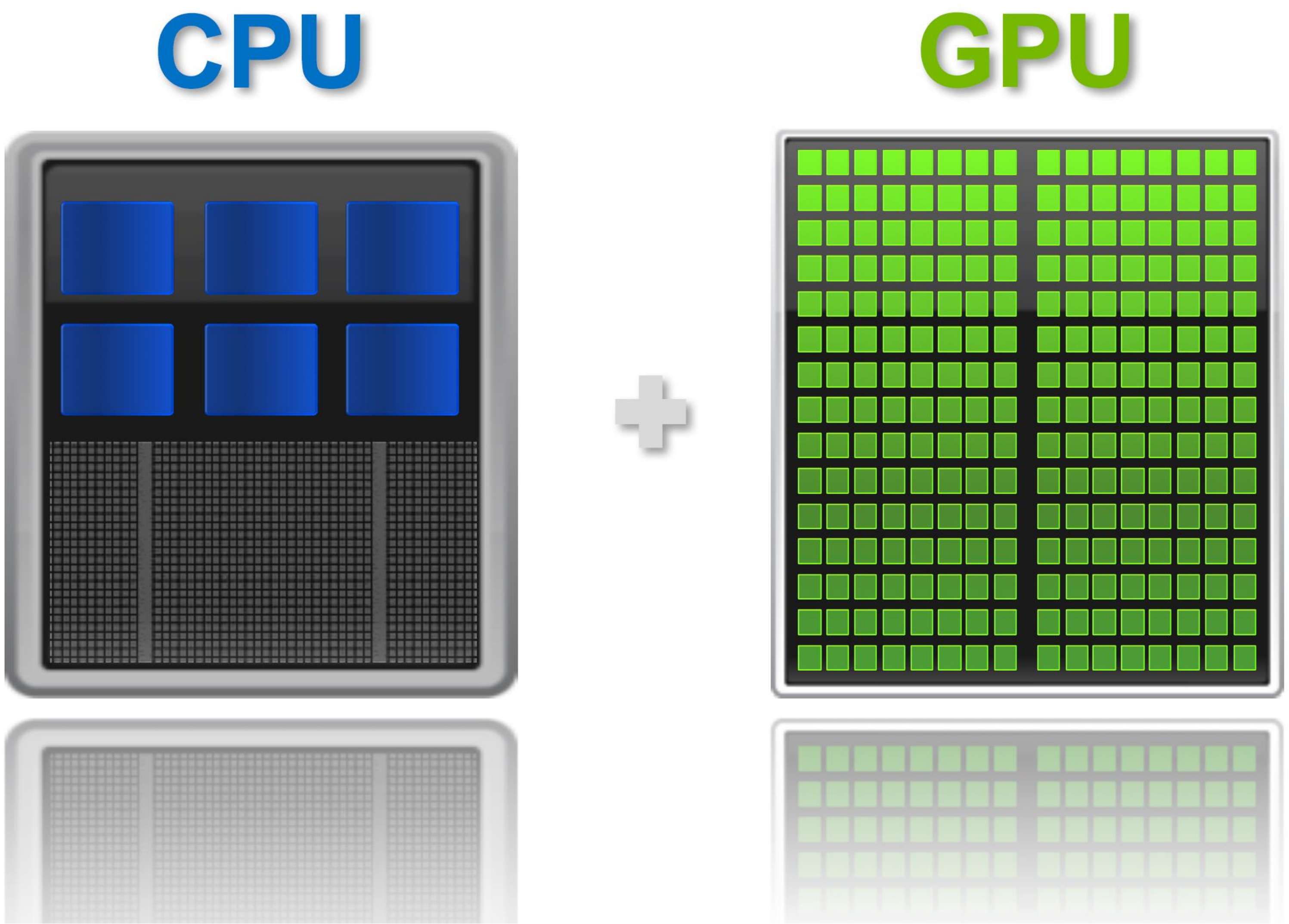
GPU Computing

Computer Graphics

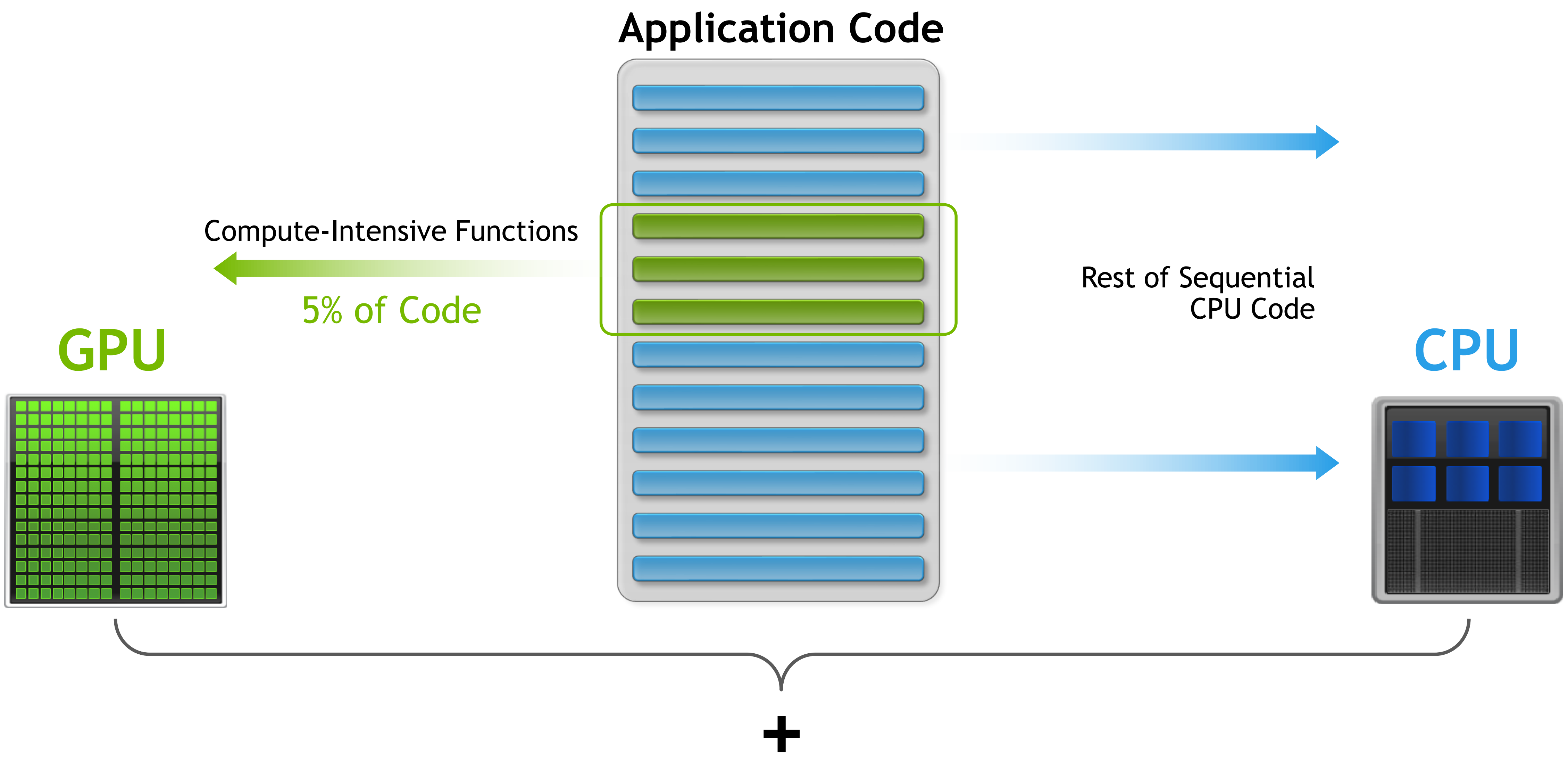


Artificial Intelligence

Add GPUs: Accelerate Science Applications



HOW GPU ACCELERATION WORKS



Application Code

Compute-Intensive Functions

5% of Code

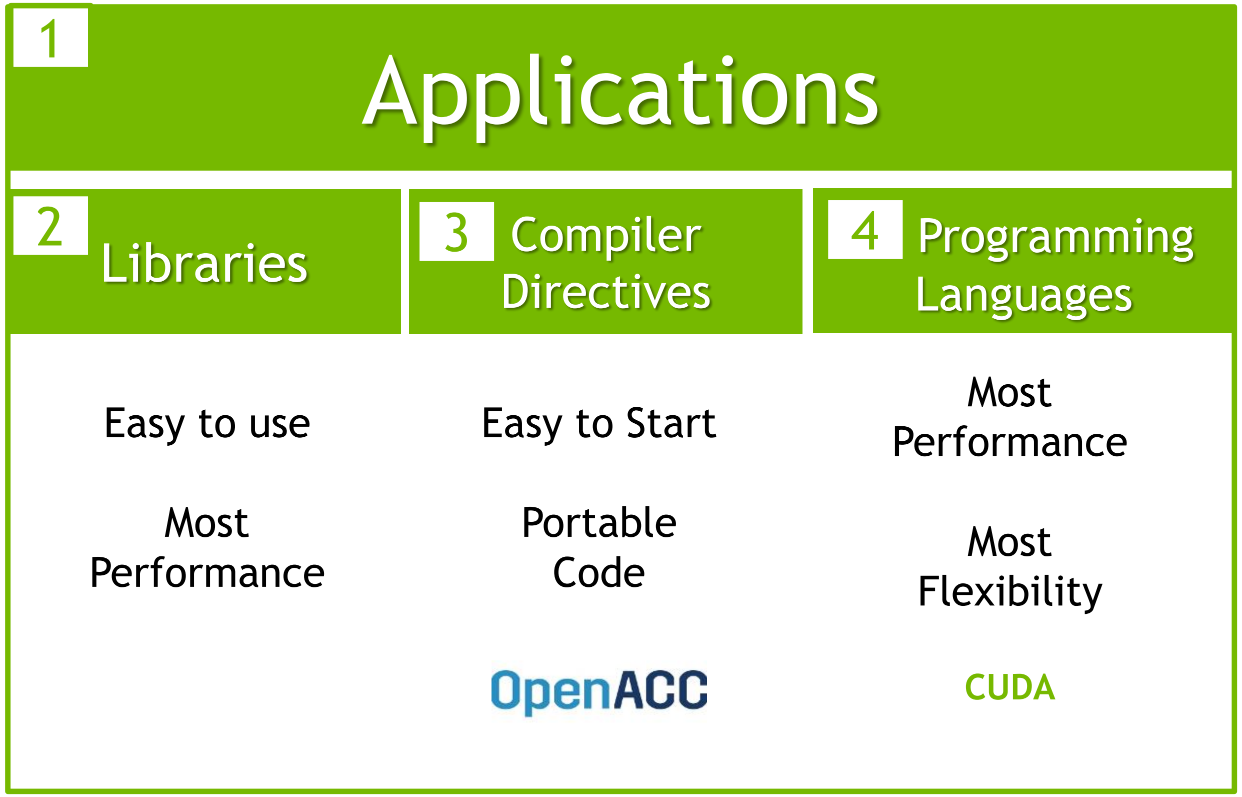
GPU

Rest of Sequential CPU Code

CPU

+

HOW TO START WITH GPUS



1 Applications

2 Libraries	3 Compiler Directives	4 Programming Languages
Easy to use	Easy to Start	Most Performance
Most Performance	Portable Code	Most Flexibility
	OpenACC	CUDA

1. Review available GPU-accelerated applications

2. Check for GPU-Accelerated applications and libraries

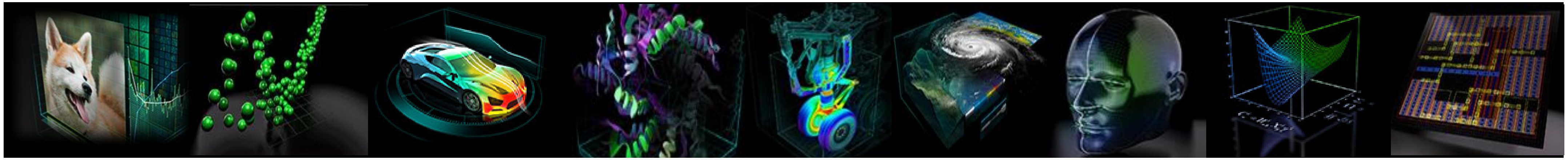
3. Add OpenACC Directives for quick acceleration results and portability

4. Dive into CUDA for highest performance and flexibility

NVIDIA CUDA-X LIBRARIES

Software To Deliver Acceleration For HPC & AI Apps; 500+ New Updates

Machine Learning & Deep Learning	Computational Physics & Chemistry	Computational Fluid Dynamics	Life Sciences & Bioinformatics	Structural Mechanics	Weather & Climate	Geoscience, Seismology & Imaging	Numerical Analytics	Electronic Design Automation



600+ Apps

Linear Algebra | Parallel Algorithms | Signal Processing | Deep Learning | Machine Learning | Visualization

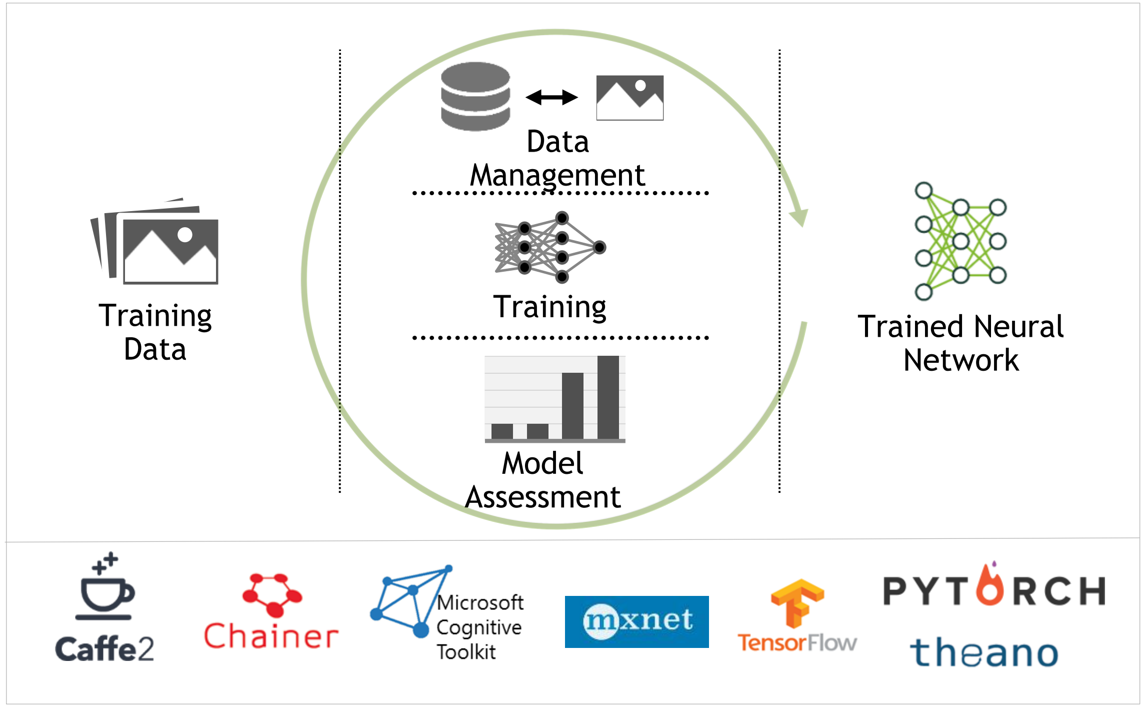
CUDA-X HPC & AI
40+ GPU Acceleration Libraries

CUDA

Desktop Development | Data Center | Supercomputers | GPU-Accelerated Cloud

NVIDIA DEEP LEARNING SOFTWARE STACK

TRAINING



Training Data

Data Management

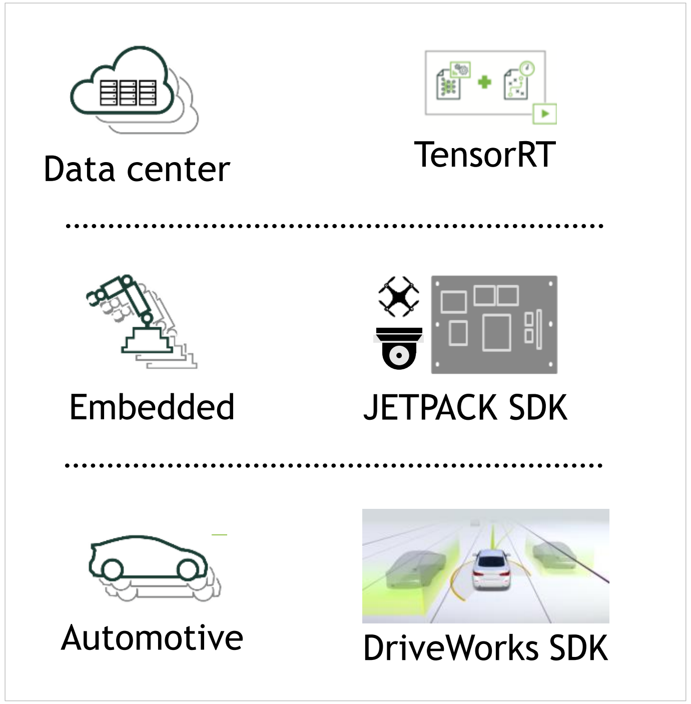
Training

Model Assessment

Trained Neural Network

Caffe2 · Chainer · Microsoft Cognitive Toolkit · mxnet · TensorFlow · PYTORCH · theano

INFERENCE

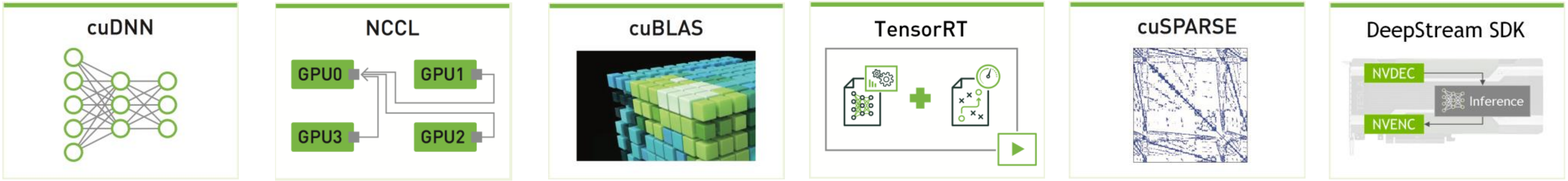


Data center — TensorRT

Embedded — JETPACK SDK

Automotive — DriveWorks SDK

NVIDIA DEEP LEARNING SDK and CUDA

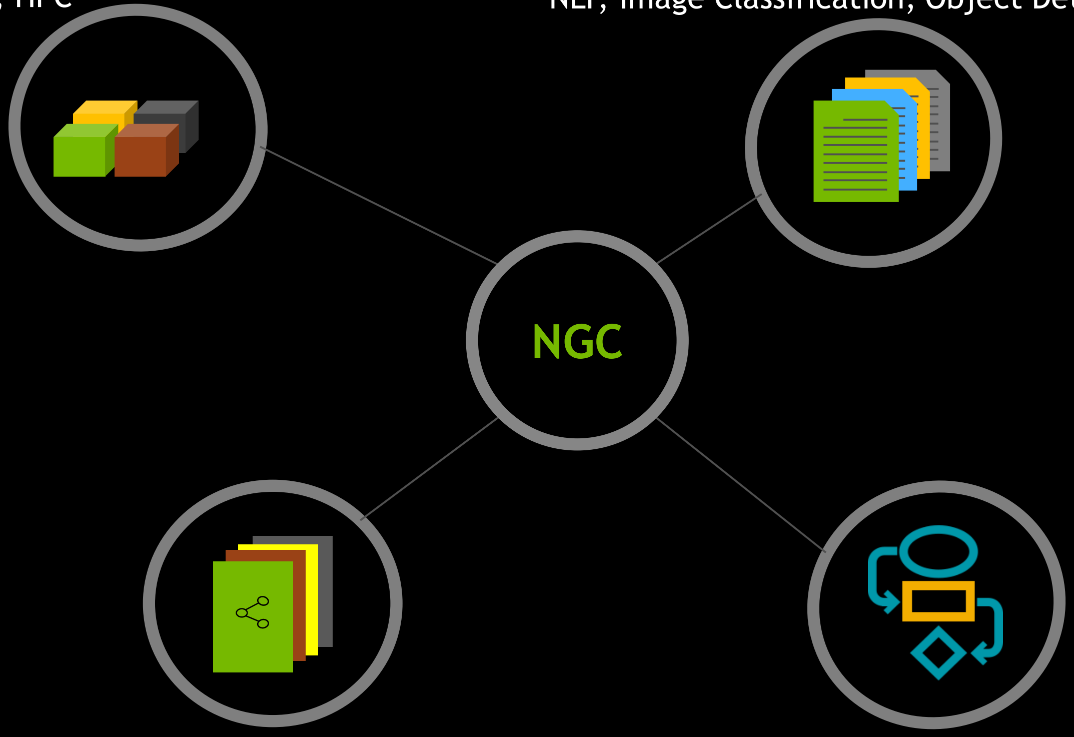


cuDNN · NCCL · cuBLAS · TensorRT · cuSPARSE · DeepStream SDK

developer.nvidia.com/deep-learning-software

NGC: GPU-OPTIMIZED SOFTWARE HUB

Ready-to-run GPU Optimized Software, Anywhere

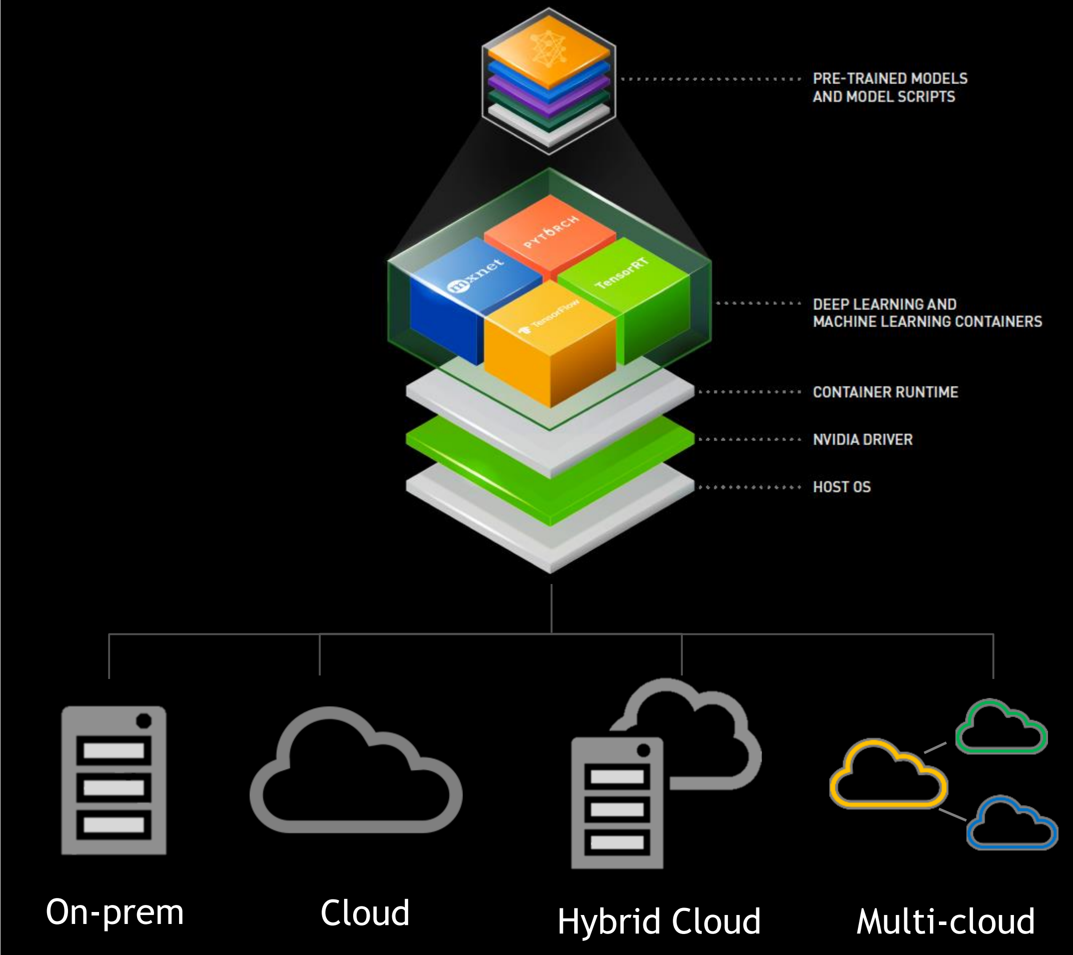


50+ Containers
DL, ML, HPC

15+ Model Training Scripts
NLP, Image Classification, Object Detection & more

60 Pre-trained Models
NLP, Image Classification, Object Detection & more

Industry Workflows
Medical Imaging, Intelligent Video Analytics



PRE-TRAINED MODELS AND MODEL SCRIPTS

DEEP LEARNING AND MACHINE LEARNING CONTAINERS

CONTAINER RUNTIME

NVIDIA DRIVER

HOST OS

On-prem

Cloud

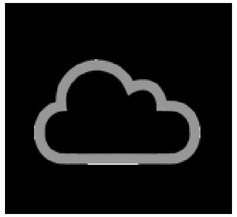
Hybrid Cloud

Multi-cloud

GPU-ACCELERATED DATA SCIENCE PLATFORMS

Unparalleled Performance and Productivity

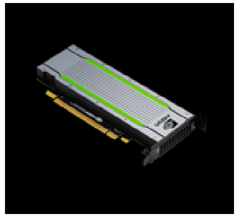
ML in the Cloud
All the top CSPs

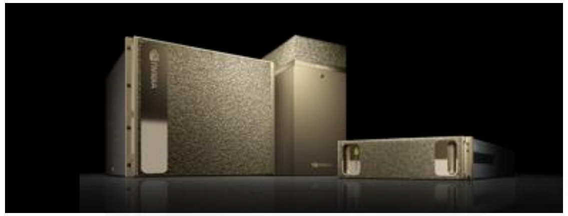


ML Enthusiast
High-end PCs

Enterprise Desktop
Individual Workstations

Enterprise Data Center
Shared Infrastructure for Data Science Teams





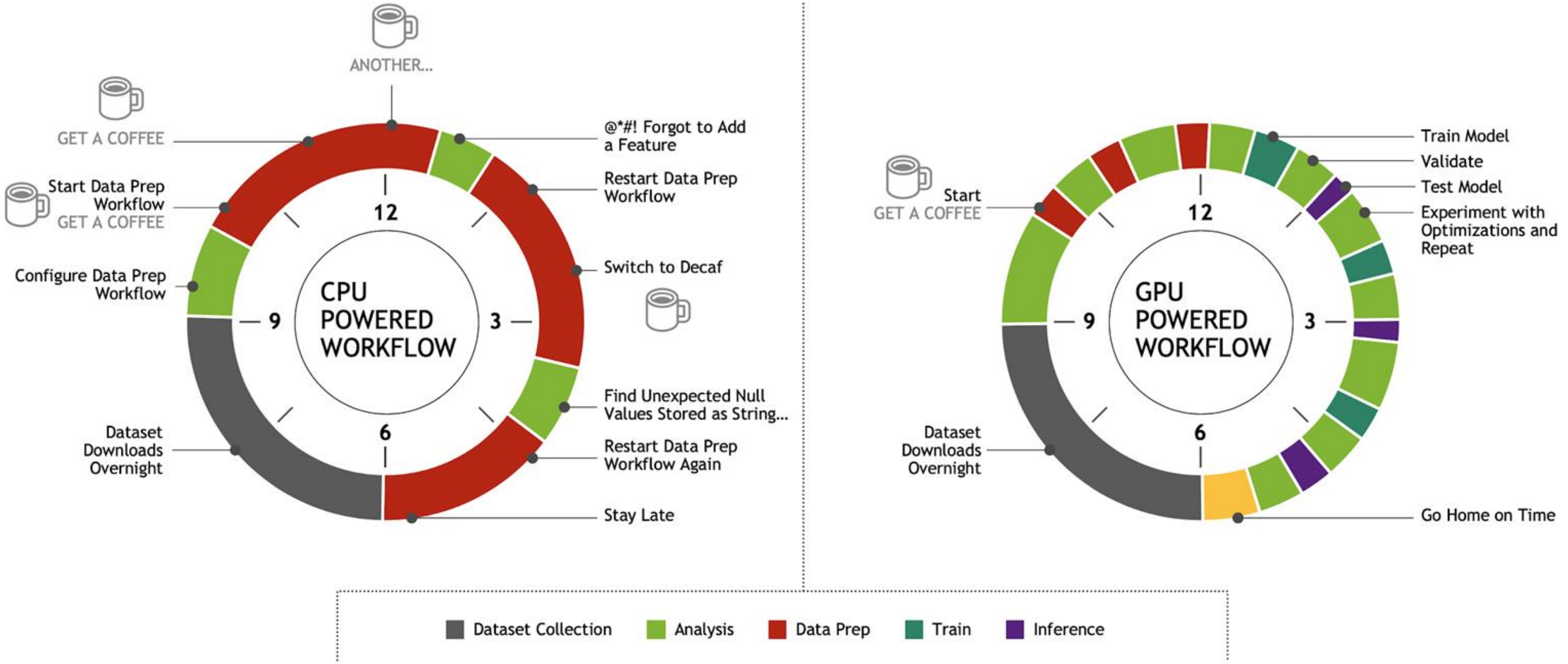
Max Flexibility

Max Performance

	NVIDIA GPUs in the Cloud	GeForce	TITAN RTX	NVIDIA-Powered Data Science Workstations	T4 Enterprise Servers	DGX Station, DGX-1 / HGX-1	DGX-2 / HGX-2
Benefit	Ease of getting started, low/no barrier to entry, elasticity of resources	Enthusiast PC solution, easy to acquire, low cost, great performance	The ultimate PC GPU for data scientists. Easy to acquire, deploy and get started experimenting.	Enterprise workstation for experienced data scientists	Standard GPU-accelerated data center infrastructures with the world's leading servers	Enterprise server, proven 4 or 8-way configuration, modular approach for scale-up, fastest multi-GPU & multi-node training	Largest compute and memory capacity in a single node, fastest training solution
Typical GPU Memory (system dependent)	varies depending on offering	22GB	48GB	96GB	64 GB (4 x 16 GB)	128GB-256GB	512GB
GPU Fabric	varies depending on offering	2-way NVLink	2-way NVLink	2-way NVLink	PCIe 3.0	4- and 8-way NVLink	16-way NVSwitch

ACCELERATING DATA SCIENCE IN HEALTHCARE

DAY IN THE LIFE OF A DATA SCIENTIST



CPU POWERED WORKFLOW

12 3 6 9

ANOTHER...

GET A COFFEE

Start Data Prep Workflow

GET A COFFEE

Configure Data Prep Workflow

Dataset Downloads Overnight

@*#! Forgot to Add a Feature

Restart Data Prep Workflow

Switch to Decaf

Find Unexpected Null Values Stored as String...

Restart Data Prep Workflow Again

Stay Late

GPU POWERED WORKFLOW

12 3 6 9

Start

GET A COFFEE

Dataset Downloads Overnight

Train Model

Validate

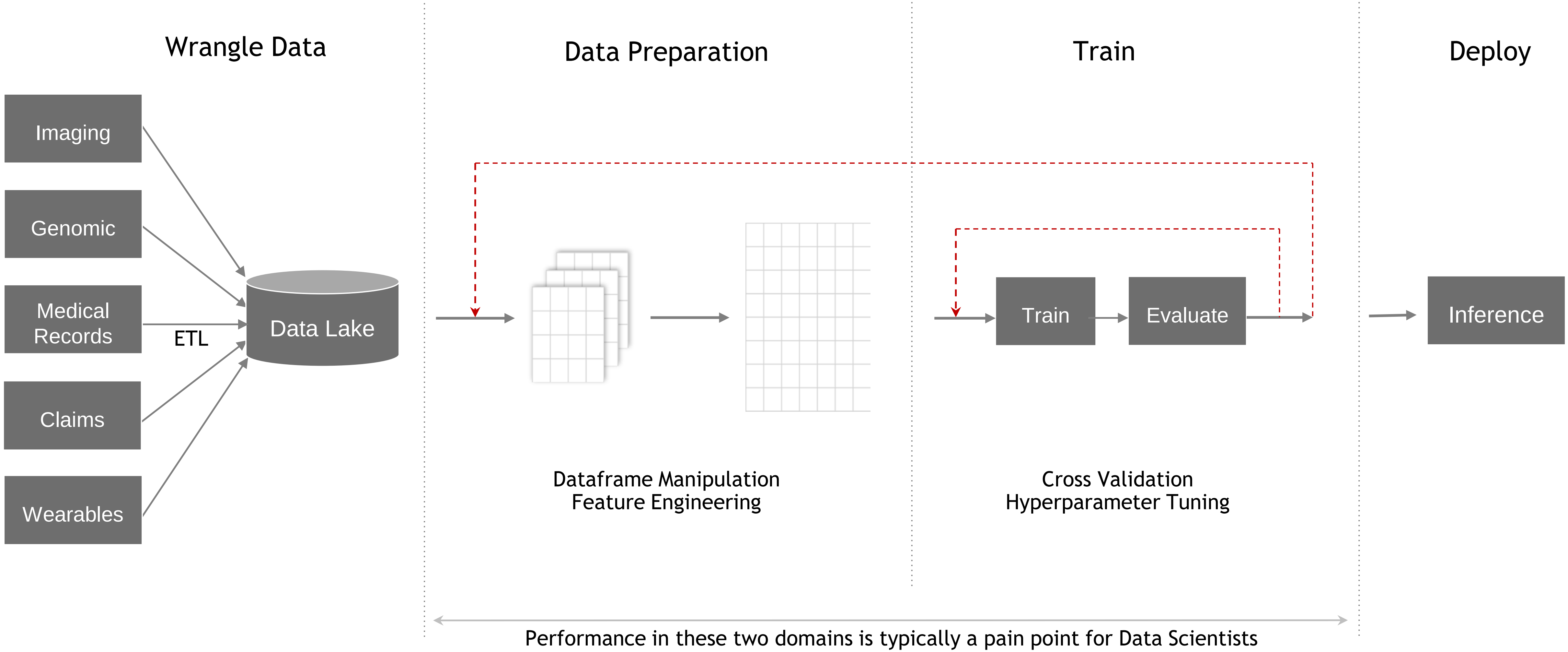
Test Model

Experiment with Optimizations and Repeat

Go Home on Time

⬛ Dataset Collection 🟩 Analysis 🟥 Data Prep 🟩 Train 🟪 Inference

CHALLENGES IN DATA SCIENCE

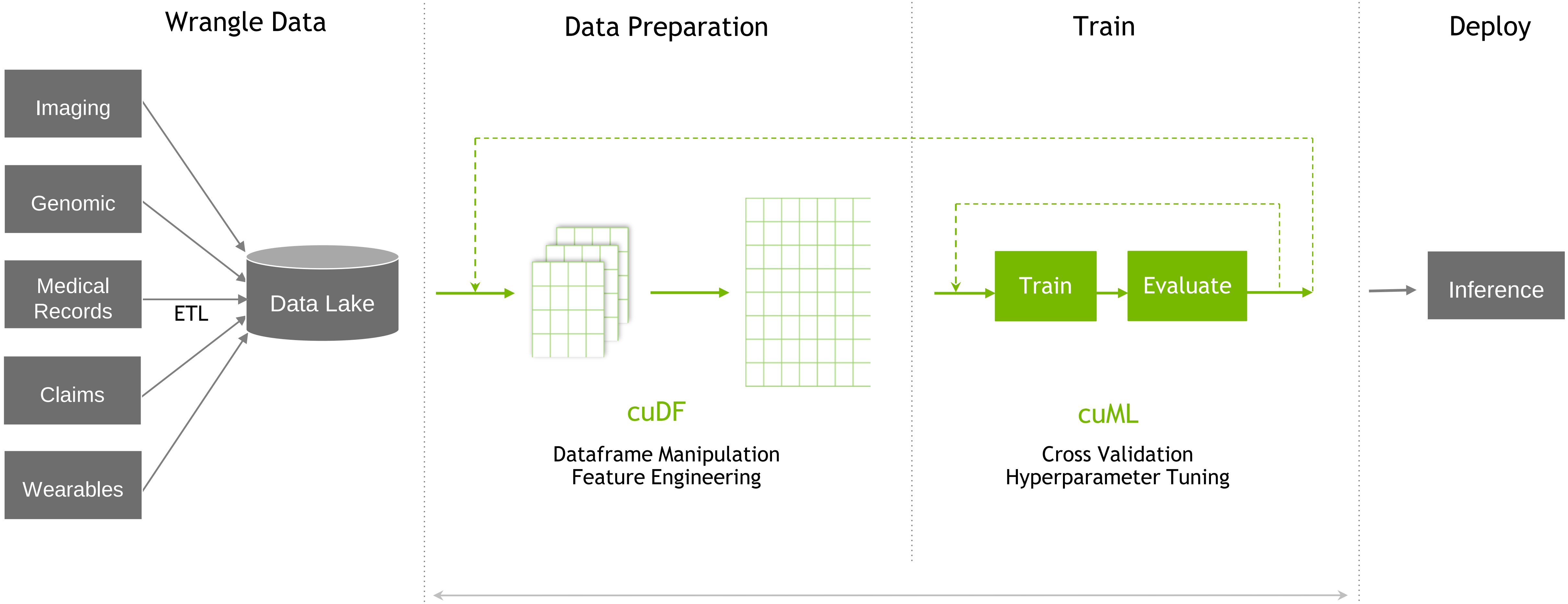


Dataframe Manipulation
Feature Engineering

Cross Validation
Hyperparameter Tuning

Performance in these two domains is typically a pain point for Data Scientists

RAPIDS IN DATA SCIENCE



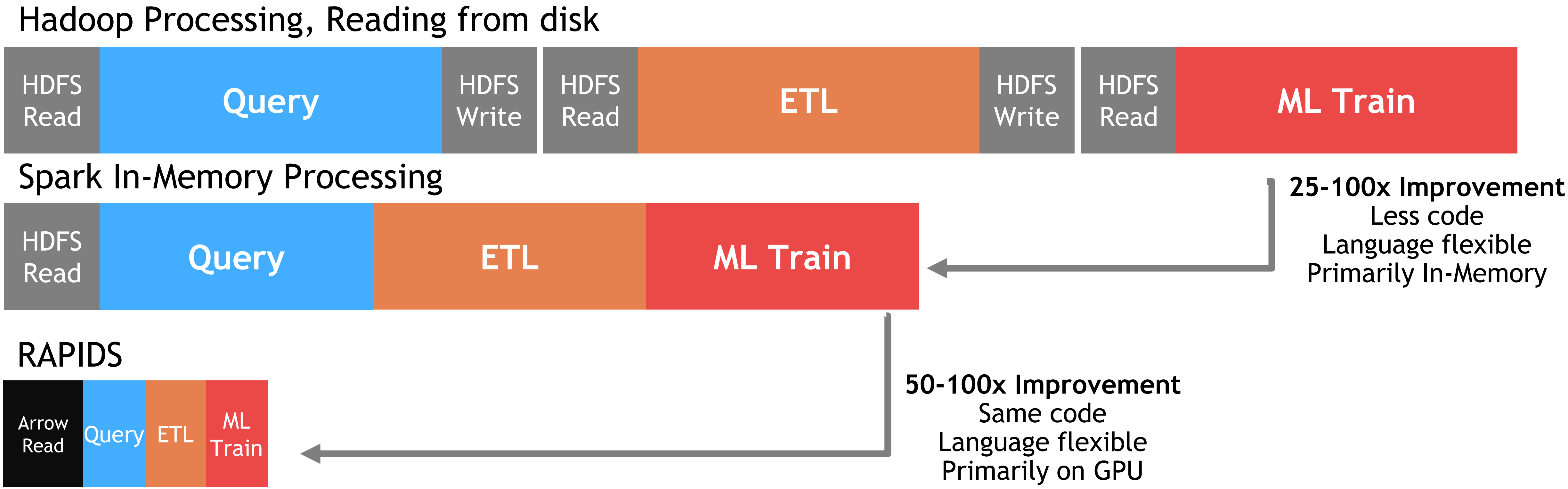
Performance in these two domains is typically a pain point for Data Scientists

cuML Algorithms

cuML	Single-GPU	Multi-GPU	Multi-Node-Multi-GPU
Gradient Boosted Decision Trees (GBDT)	🟩	🟩	🟩
GLM	🟩	🟩	
Logistic Regression	🟩		
Random Forest	🟩	🟩	🟩
K-Means	🟩	🟩	🟩
K-NN	🟩	🟩	
DBSCAN	🟩		
UMAP	🟩		
Holt-Winters	🟩		
Kalman Filter	🟩		
t-SNE	🟩		
Principal Components	🟩		
Singular Value Decomposition	🟩	🟩	

Data Processing Evolution

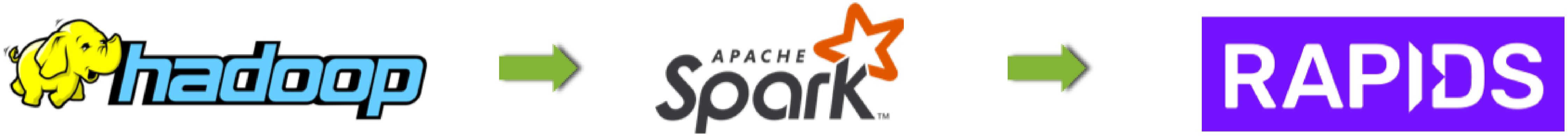
Faster data access, less data movement



25-100x Improvement
Less code
Language flexible
Primarily In-Memory

50-100x Improvement
Same code
Language flexible
Primarily on GPU

Disk → Memory → GPUs



Scalable, but slow due to repeated reads & writes to disk

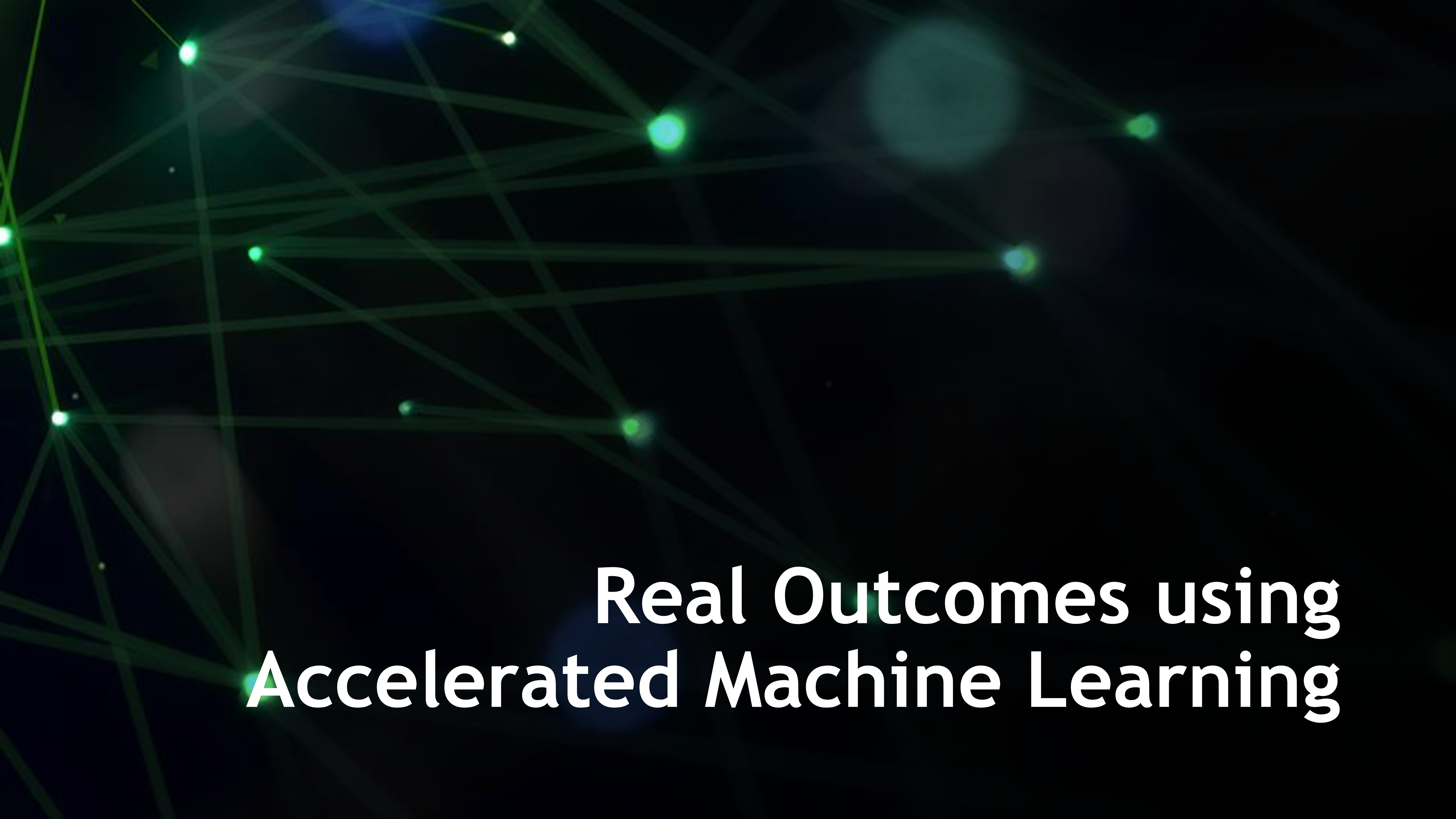
Faster, by keeping data always in host memory instead of on disk

Performance limited by CPUs

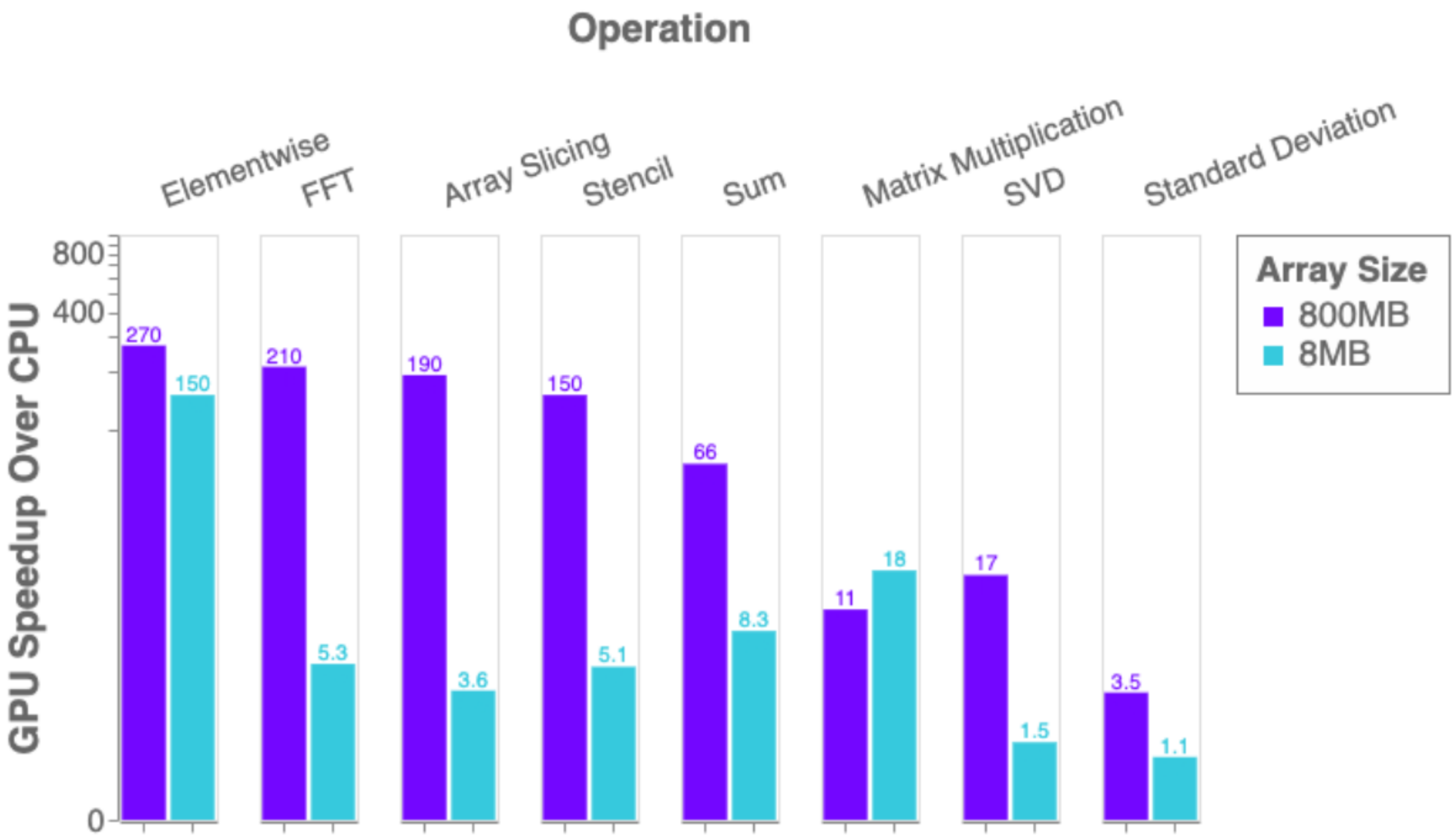
Keeps data in GPU memory instead of CPU memory

Computations are GPU accelerated

Real Outcomes using Accelerated Machine Learning



cuPy Acceleration

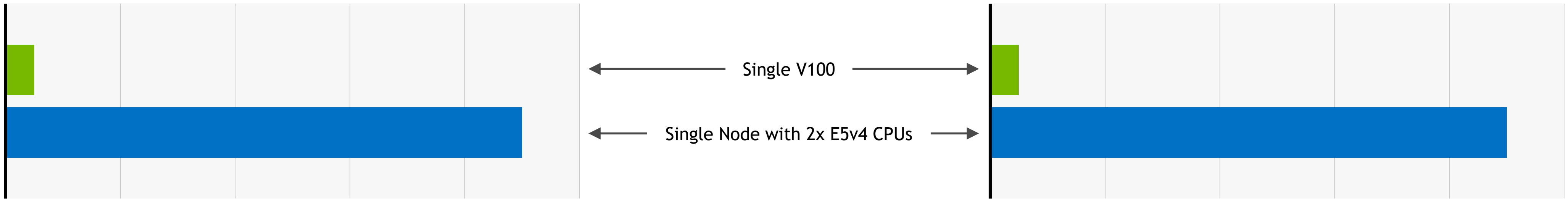


TRANSFORM GENETICS WITH RAPIDS

Personalize Immunotherapy for Cancer Patients

GPU Accelerated XGBoost

GPU Accelerated cuDF



10.24X Speedup

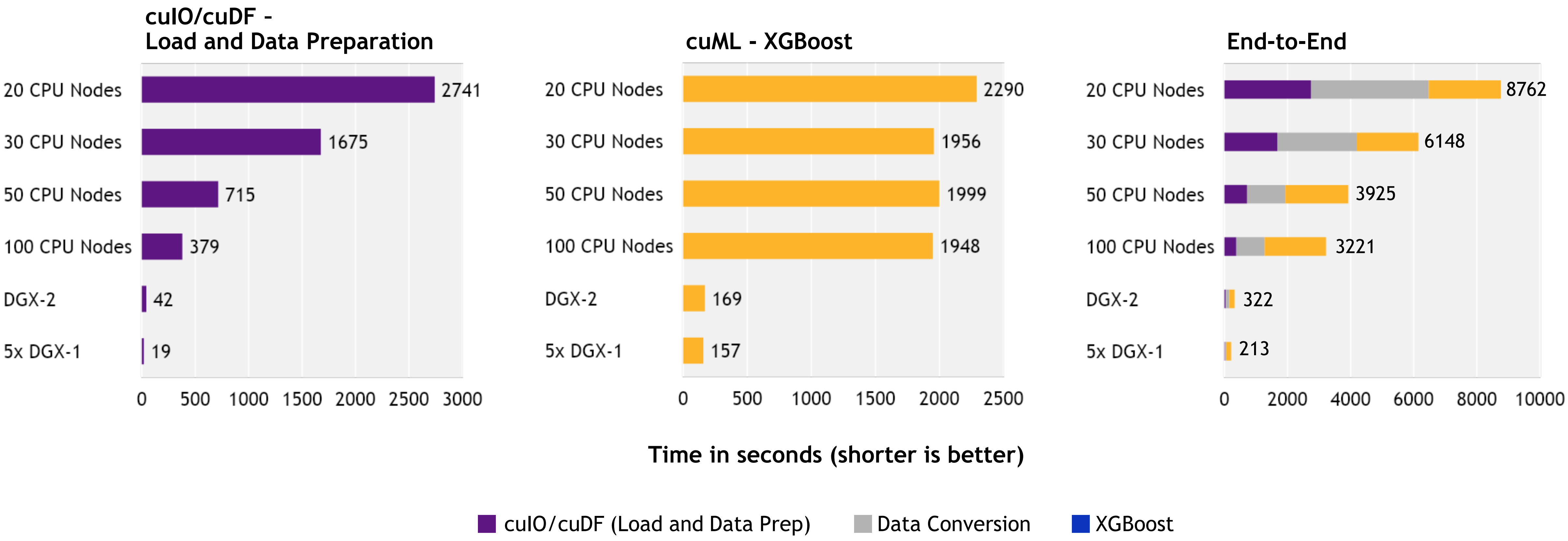
18.14X Speedup

Users of K-means, PCA, and XGBoost

> *“We see close to 20x speedup using XGBoost on DGX-1. This helps us significantly improve our personalized immunotherapy and expand our analysis to millions of peptide candidates.”*

Yong Hou, Duty Director of BGI Research

Faster Speeds, Real-World Benefits



Benchmark

200GB CSV dataset; Data prep includes joins, variable transformations

CPU Cluster Configuration

CPU nodes (61 GiB memory, 8 vCPUs, 64-bit platform), Apache Spark

DGX Cluster Configuration

5x DGX-1 on InfiniBand network

NVIDIA®