InfoGlobalTech
Innovative Solutions for your Enterprise

Essentials of Deploying AI in the Data Center

AI Use Cases

AI Concepts (ML, DL, Inferencing)

How GPUs revolutionized AI

Nvidia AI Architecture

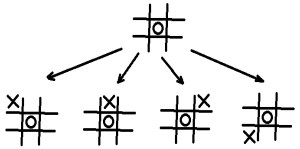
Infrastructure Planning

Data Center Facilities Planning

Infrastructure Provisioning and Management

InfoGlobalTech
Innovative Solutions for your Enterprise

AI to Revolutionize Future Data Centers



Artificial Intelligence

Using computers to do that requires human level Intelligence

1950s – 1980s

Challenge: Understanding and extracting insights from big data

Machine Learning

Approach to AI that uses statistical learning algorithms to build Model from observed **data**

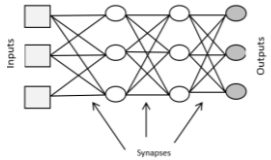
1980s – 2010s

Most of Deep learning methods use Neural Network models (DNNs)

Deep Learning

Machine learning technique that is inspired by how human brain learns

2010s – Today and the future



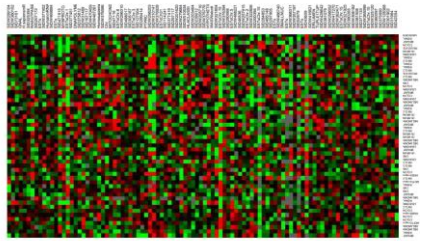
DNNs can achieve human level intelligence for many tasks but requires high computational power to train

Data scientists uses DL to extract meaningful **data** from large datasets

Line of business owners use DL to optimize their business, reduce costs, Improve functionality and accuracy

InfoGlobalTech
Innovative Solutions for your Enterprise

AI in healthcare



Early Detection of diseases

DNA, Expression Microarrays, a regression problem facilitating the breakthrough study in biotechnology

Operational Efficiency

Predictive analytics, to identify patterns, and predict patient outcomes, enabling personalized and proactive treatment plans.

Precision Medicine

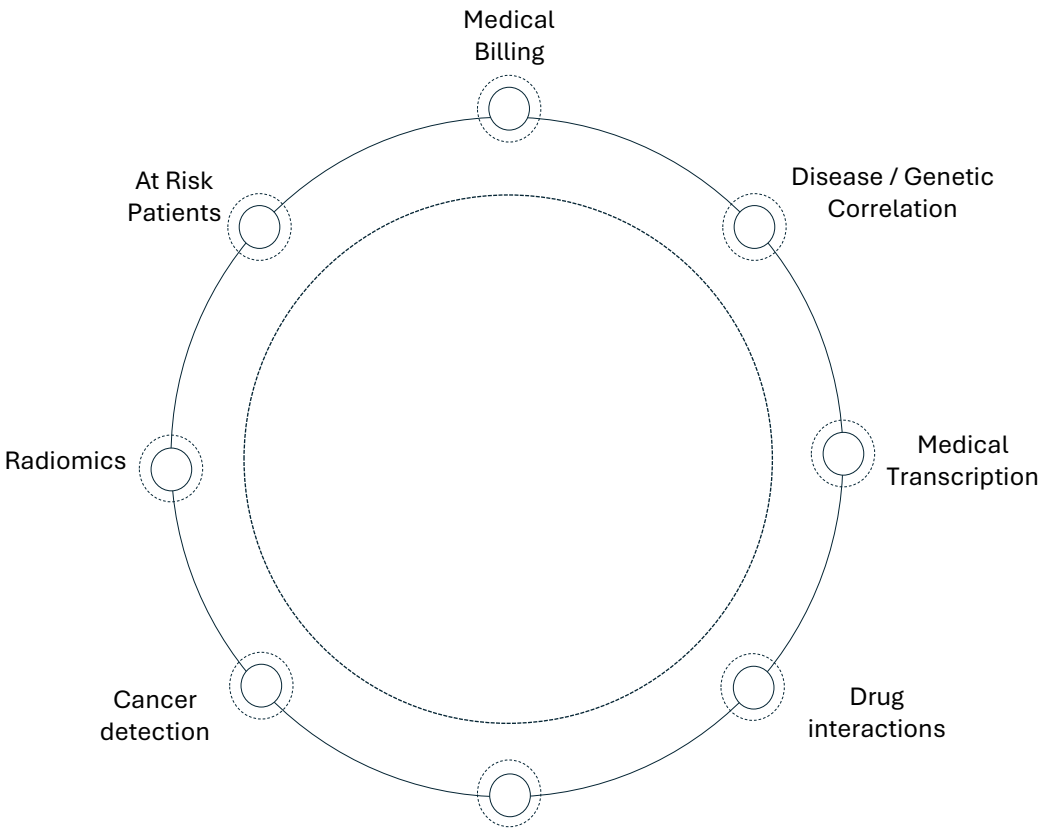
Descriptive analytics that facilitate the development of precise personalized medicines

Drug Discovery

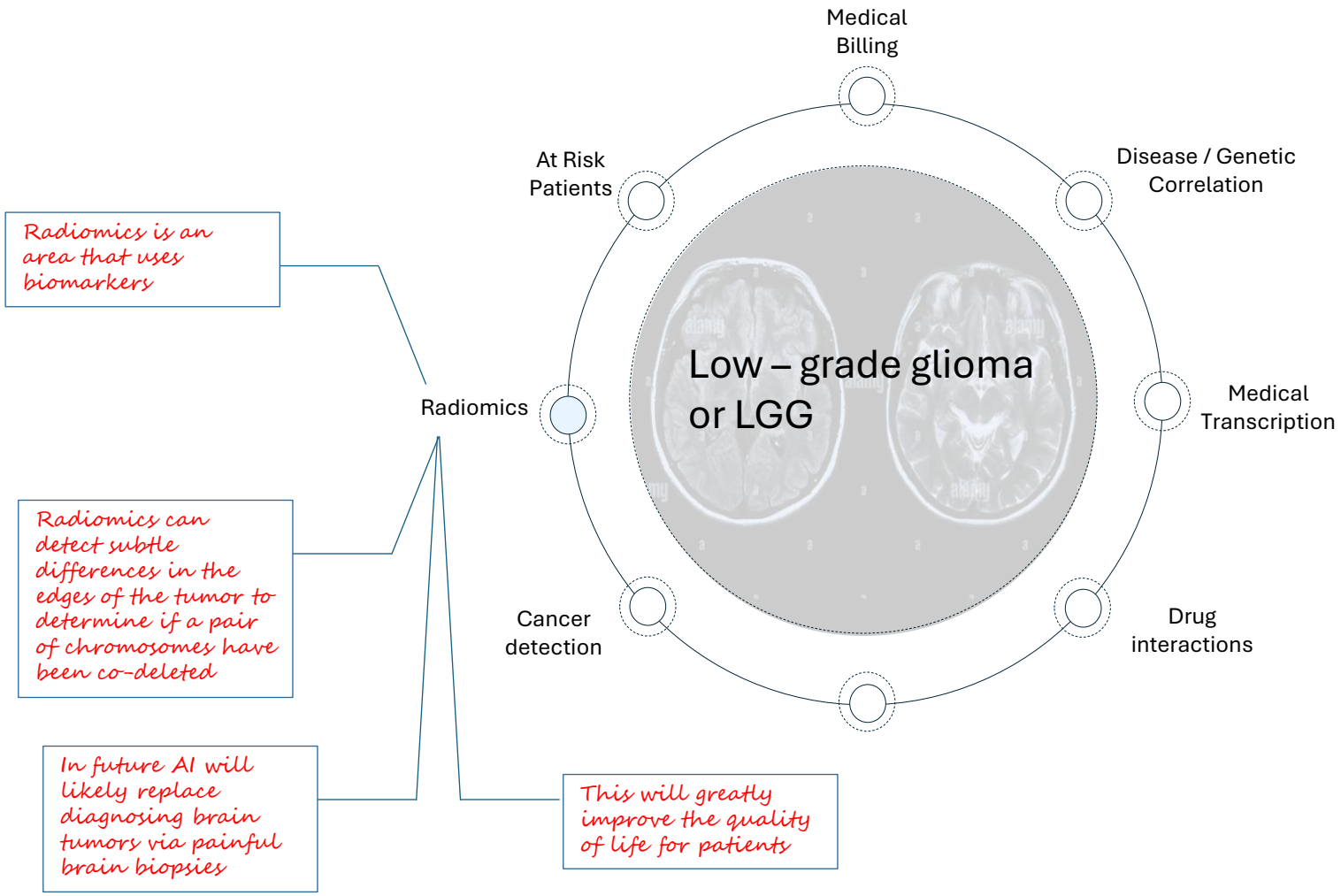
Researchers can bring new drugs to Market quicker and make sure that these drugs are monitored more efficiently

InfoGlobalTech
Innovative Solutions for your Enterprise

AI in healthcare



AI in healthcare



InfoGlobalTech
Innovative Solutions for your Enterprise

AI in Autonomous vehicles

Potential for Every Vehicle to be autonomous

- 1.5B vehicles in world today → 2B vehicles by 2035
- Mobility Services (Buses, Taxis, robotaxis, shuttles)
- 4M AV shipments by 2025 (New autonomous vehicles – Earthmovers, Forklifts, Delivery Bots, Tractors, Firetrucks etc)

Reduced fatalities, Injuries, and parking footprint

Shared autonomous vehicles can substantially reduce parking space requirements

Mitigates shortage of delivery services and drivers

This shortage can be resolved with delivery bots and autonomous trucks

There is global shortage of 60000 truck drivers, this deficit is expected to triple by 2026

Nvidia Drive

E2E AV Solution to Enable Rapid, Large Scale AI Development & Testing

Building a **Self-driving** vehicle software stack is complex

Wether it is to support driver or replace driver, self- driving software stack is expensive and time consuming to build, and maintain

Productivity
Fastest time to
results

Iteration Time
Shortest time from
data collected to
tested models

Overview of the AV work underway at NVidia

Nvidia Drive

E2E AV Solution to Enable Rapid, Large Scale AI Development & Testing

Drive AV

Perception	Mapping	Mapping

1,000's Engineers
20+ DNNs, 50 Parallel Experiments

Driveworks

Sensor Abstraction	Vehicle ID and tools	Calibrate
Image / Point Cloud	Recorder	Ego motion

1,000's Engineers HW & SW
20 million lines of code

AGX

Self Driving Roborace

Sensor Data + Logs

Active Learning
Targeted Learning

Transfer Learning
Federated Learning

Ingestion	Curation	Labelling	Training	Replay	Simulation
300K hours 300M Frames 10PB/wk data collected		1500 labellers 50M + labelled images	DGX SuperPOD	1M+ virtual miles driven 200+ Drive constellations	

Drive Infra

Date Management	Workflow Management

Date center backbone

1000's of engineering hours develop and run Deep Ops & Dev Ops SW on Saturn V

DGX A100

Nvidia Drive Constellation

Nvidia Drive

E2E AV Solution to Enable Rapid, Large Scale AI Development & Testing

AV Workflow – Challenges and Pain Points

Ingestion	Replay	Simulation
Training	Labelling Labellers correct mostly predictions in mature projects to make labels precise	Curation Determines the usefulness of the incoming data

InfoGlobalTech
Innovative Solutions for your Enterprise

Accelerating Digital Transformation in Finance

AI / ML Optimizes Performance and Outocmes

Banking

Capital Markets

Insurance

Payments

Fintech

Fraud Detection
Using a range of AI techniques to better verify identity for Anti-Money Laundering (AML) and know your Customer (KYC) requirements

Risk Simulations
Using HPC to run Monte Carlo risk simulations and derivatives pricing

Underwriting
Enables lenders to integrate more sophisticated modelling techniques and alternative data into lending decisions or insurance pricing

Algorithmic Trading
Create competitive advantages across a range of investment types

InfoGlobalTech
Innovative Solutions for your Enterprise

AI Platform for AI Driven Finance

Enterprise Executives

84% fear missing their growth objectives if they don't scale AI

76% cited their struggle with how to scale AI across their business

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 PLATFORMS FOR FINANCE

All End Users

AI Financial Applications

AI Adopting Path
- Build your own
- Strategic partnership

Data Scientists/ Researchers

Date Science — Domino, NGC

System Admin/DevOps

Dev Ops — Red Hat OpenShift

Centers of excellence

Nvidia Certified GPU Compute, Networking & Storage

Lines of Business

Marketing, Finance etc.

OUTOCMES

- Time to insight
- Productivity at scale
- Return on investment

Where is the Industry Going?

Your AI Transformative Strategy is incomplete without NVidia

Customers and Financial Data

Banks

Incumbent firms,
Infra & business model challenges
long customer history existing
customer base

Fintech

Digital Entities,
Focusing on customer experience,
reducing friction and process
efficiency.

Big Tech

AI Powered digital natives,
Deep Pockets,
Beyond 360˚,
After spending data.

Big retail

Large customer base,
Will use targeted offers
to increase revenues
and profits

AI-Based underwriting – Risk Prediction

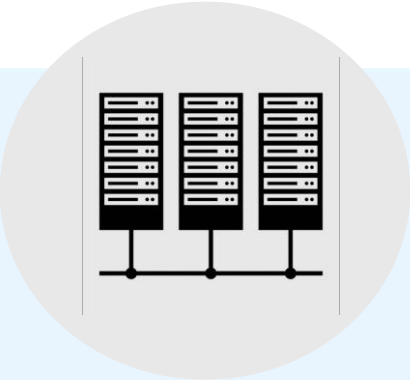
AI-enabled chatbots – Superior Customer Service

AI for fraud detection – Greater Security

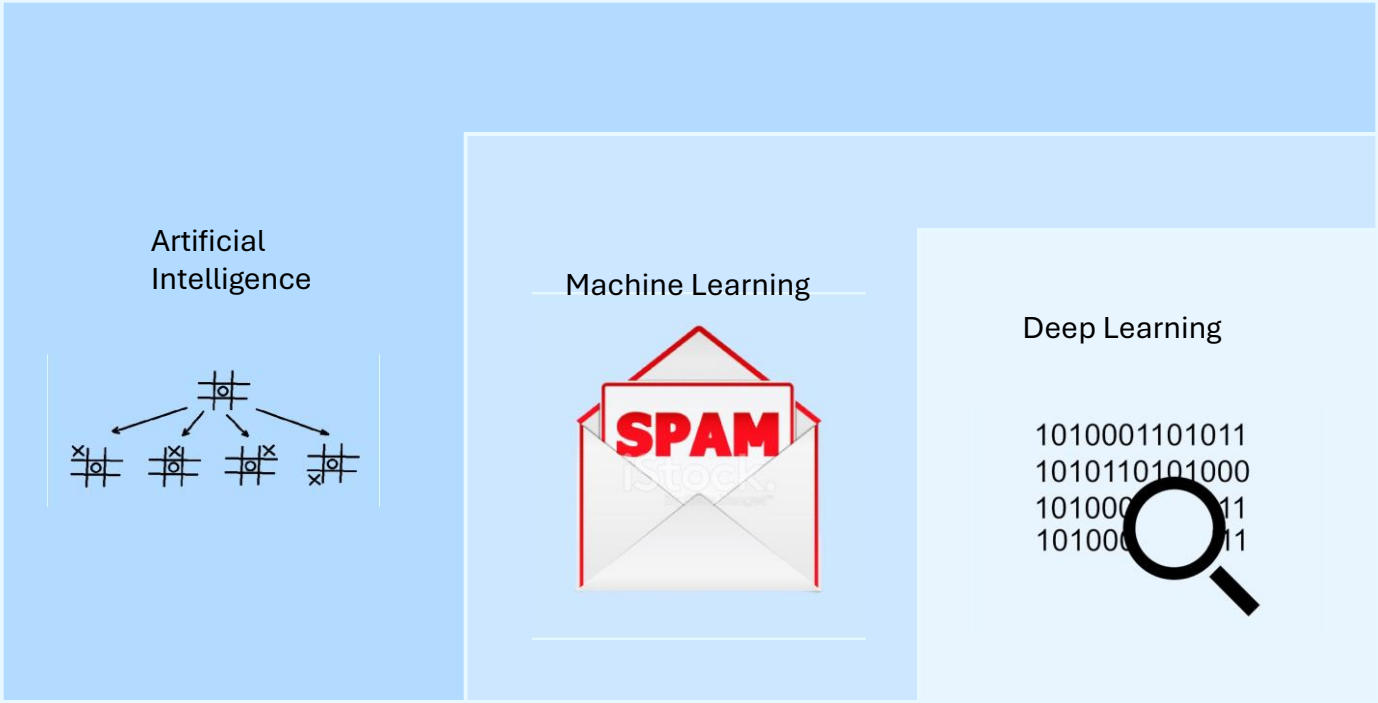
Industry Overview

From the cloud to the Edge

AI required huge amount of
compute power



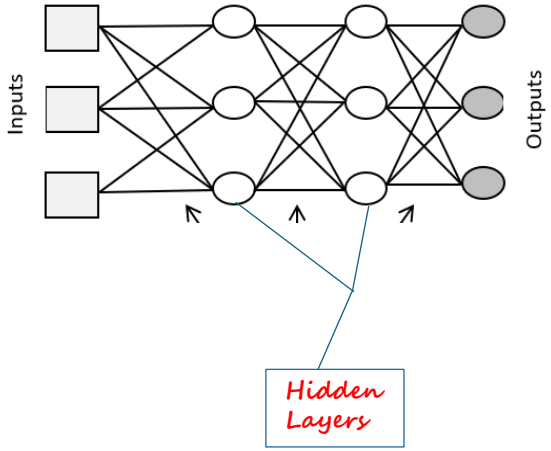
Deep Learning Approach



Deep Learning Approach

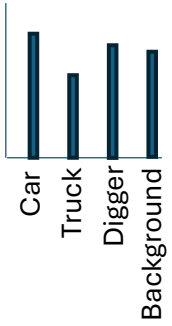
Labelled Trained Data

Deep Neural Network «Model»



Prediction

Object class Predictions



error ~ [label - prediction]

Back propagate errors for parameter update

Machine/Deep Learning Frameworks

Essential Tools for Data Scientists, Researchers, and Engineers

Computer Vision

Natural Language Processing

Speech and audio Processing

Robot Learning

more...

Caffe	Cognitive Toolkit	scikit learn
Chainer	mxnet	MATLAB
PaddlePaddle	Spark	PyTorch
TensorFlow	...	

Machine/Deep Learning Frameworks

Essential Tools for Data Scientists, Researchers, and Engineers

MXNet is a modern open-source deep learning framework used to train and deploy deep neural networks. It is scalable, allowing for fast model training, and supports a flexible programming model and multiple languages. The MXNet Library is portable and can scale to multiple GPUs and multiple machines

TensorFlow is a popular open-source software library for **dataflow** programming across a range of tasks. It is a symbolic math library and is commonly used for deep learning applications.

Scikit-learn is a free software machine learning library for the python programming language. It features various classification, regression and clustering algorithms, and is designed to interoperate with the Python numerical and scientific libraries: NumPy and SciPy

Nvidia Deep Learning Software Stack

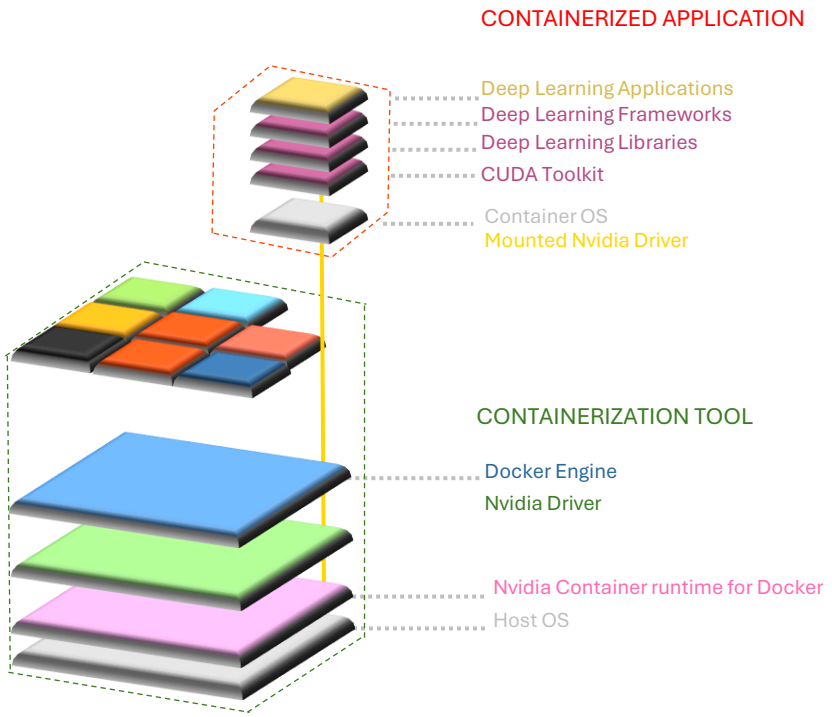
Host OS and Nvidia Driver – Enables the deep learning framework to use the GPU Functions

NGC Containers – Publicly available containers optimized to run on Nvidia GPUs

DL Frameworks – Popular deep learning frameworks available inside the containers

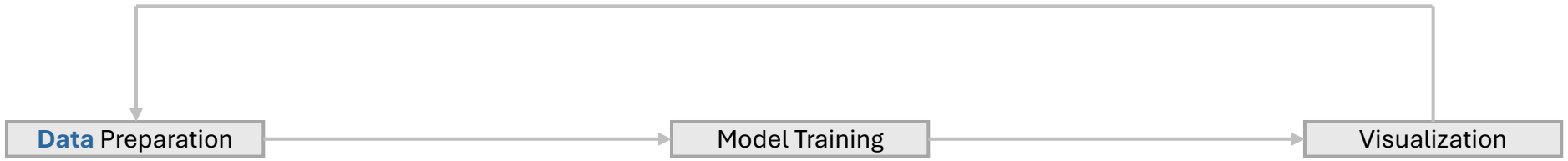
CUDA – Nvidia's ground breaking parallel programming model.

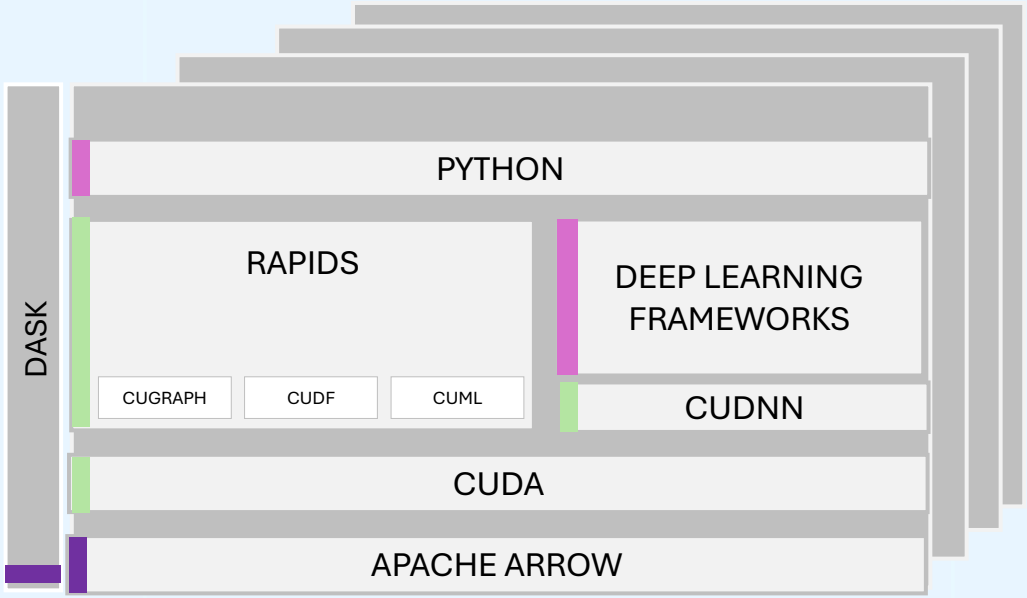
Provides essential optimizations for deep learning, machine learning, and high performance computing [HPC] leveraging Nvidia GPUs



Nvidia GPU software stack

Machine Learning Software Stack





Nvidia CUDA – X AI Ecosystem

FRAMEWORK

Chainer
mxnet
TensorFlow
PaddlePaddle
learn

Cloud ML Services

Azure Machine Learning
databricks

Cloud ML Services

TensorFlow
ONNX RUNTIME

CUDNN
CUML
Nvidia Tensor RT
CuDF
CuGraph

CUDA-X AI

CUDA

WORKSTATION

hp
DELL
Lenovo

SERVER

CUML
CISCO
Lenovo
Hewlett Packard Enterprise
IBM
DELL EMC

CLOUD

aws
Azure
Alibaba.com
Google Cloud

Nvidia CUDA – X AI Libraries

Data Processing	nv**JPEG**	https://developer.nvidia.com/nvjpeg
	DALI	https://github.com/NVIDIA/DALI
	cu**DF**	https://github.com/rapidsai/cudf
	OpticalFlow	https://developer.nvidia.com/opticalflow-sdf
	NPP	https://developer.nvidia.com/npp
Deep Learning Training	**AMP**	https://developer.nvidia.com/automatic-mixed-precision
	Apex	https://github.com/NVIDIA/apex
AI Specific Acceleration Libraries	cu**BLAS**	https://developer/nvidia.com/cublas
	cu**DNN**	https://developer.nvidia.com/cuDNN
	cu**xFilter**	https://github.com/rapidsai/cuxfilter
	c**uML**	https://github.com/rapidsai/cuml
	cu**GRAPH**	https://github.com/rapidsai/cudgraph
	cu**Tensor**	https://developer.nvidia.com/cuTensor
	NCCL	https://developer/nvidia/com/nccl
Deployment	**TensorRT**	https://developer.nvidia.com/tensorrt
	Inference Server	https://github.com/NVIDIA/tensorrt-inference-server
High Level Constructs	**Transfer Learning Tool**	https://developer.nvidia.com/transfer-learning-toolkit
	Rapids	https://github.com/rapidsai
	DeepStream	https://developer.nvidia.com/deepstream-sdk

Key Technologies for Deployment

Containers

Package app
- Libraries
- Compilers
- Network Drivers
- Other Components

Simplifies Deployments
Eliminates complex, time-consuming builds and installs

Quick Start
Simply download and run the app

Portable
Deploy across various environments, from test to production, with minimal changes

Kubernetes

Orchestration tool to easily deploy containers on various nodes

Automates Deployments
Launch apps on appropriate nodes

Scaling
Automatically spins up nodes to meet demand

Monitoring
Automatically restarts if application crashes

Slurm

Job scheduler to manage allocation of resources and launching jobs on a cluster

Restore Allocation
Automatically allocates resources for job

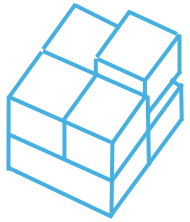
Large Clusters
Supports running on small to vary large clusters

Distributed Jobs
Allows multi-nodes jo to be launched for faster distributed training

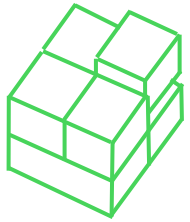
NGC Catalog – GPU Optimized Hub for AI & HPC software

Simplify and Accelerate End-to-End Workflows

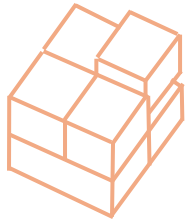
Software stacks vary End-to-End Workflow



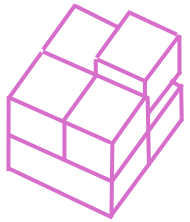
AI Stack



HPC Simulation apps



Genomics Stack



Visualization app

NGC Catalog – GPU Optimized Hub for Ai & HPC Software

Simplify and Accelerate End-to-End Workflows

CONTAINERS	PRE-TRAINED MODELS	HELM CHARTS	COLLECTIONS	INDUSTRY APP FRAMEWORKS
AI Stack	Computer Vision | NLP | DLRM	Triton |GPU Operator	Clara Discovery | TLT - Jarvis | Recsys	Clara | RIVA | SAAC

On Prem

Cloud

Hybrid Cloud

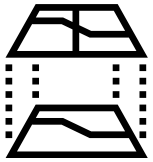
Edge

InfoGlobalTech
Innovative Solutions for your Enterprise

Fast Track AI with Pre-Trained Models from NGC

Popular Collections

Image Segmentation
Collection –
Computer Vision



Speech Synthesis
Collection –
Conversational AI

Recommender Synthesis
Collection –
Understanding

Translation
Collection –
Conversational AI

Language Modelling
Collection –
Conversational AI

Object Detection
Collection –
Computer Vision

Production Quality
- Trained and Continuously updated by experts
- Model resumes to find the right fit

Wide Range of Use Cases
- People Detection, Vehicle Detection, Gaze Estimation
- Instant Classification, Question-Answering, Speech recognition, and Text-to-speech

Adapt & Integrate
- Adapt to your domain with your custom data
- Integrate easily into industry SDKs

Benefits of GPU Virtualization

Nvidia in Virtualization Environments – Industry Leading Innovations

Bare Metal Performance	Operational Management	Insight & Tools
Business Continuity and Workload Balancing	Resource sharing and improved utilization	Infrastructure & Data Security

Nvidia End-to-End AI software

Optimized for Nvidia-Certified Servers

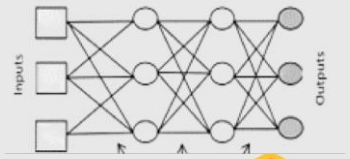
PRE-TRAINED MODELS

- Automatic Speech Recognition
- Object Detection
- Speech Synthesis
- Machine Translation
- Image Classification
- Language Modelling
- Recommender Systems
- Segmentation

Data Prep

NVIDIA RAPIDS

Train at Scale



PyTorch TensorFlow

Optimized for Inference

NVIDIA TensorRT

Deploy at Scale

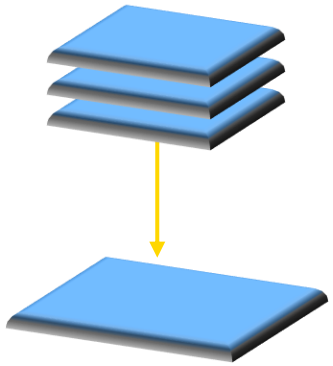
NVIDIA Triton Inference Server

AI Workflow

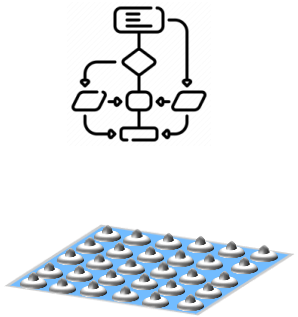
Open Source, End-to-end GPU-acclerated Workflow

GPU Accelerated compute in-memory data preparation. Simplified implementation using similar data science tools

Data Preparation

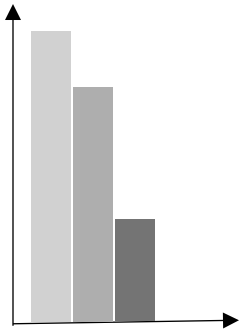


Data Preparation

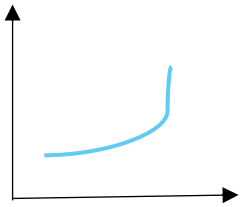


GPU – acceleration of today's most popular ML Algorithms ML – data processing mechanisms that has been trained to recognize certain types of patterns

Visualization

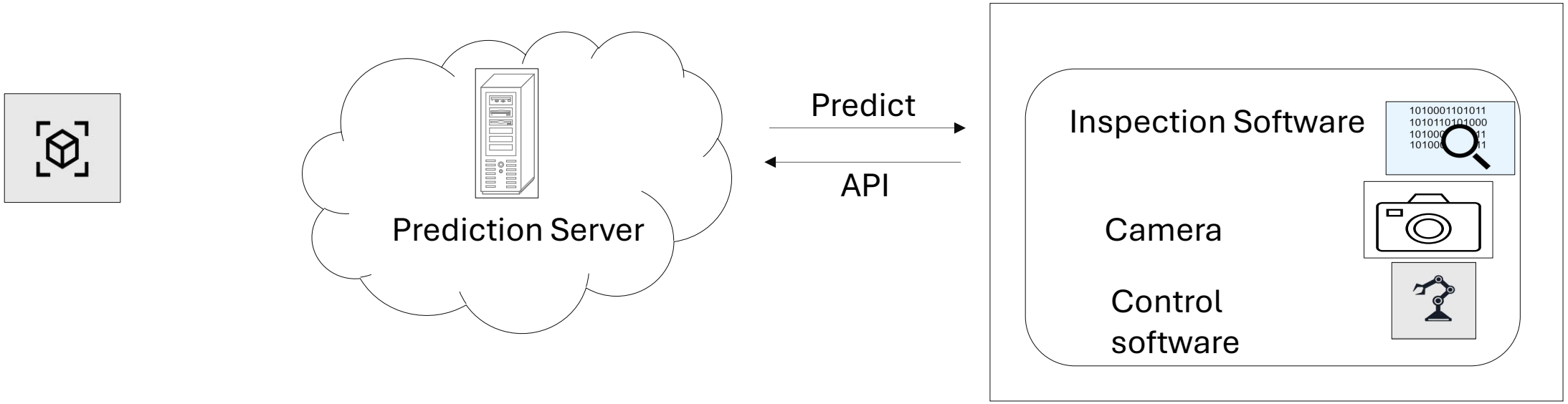


Effortless exploration of datasets, billions of records in milliseconds Dynamic interactions with data – faster ML model development

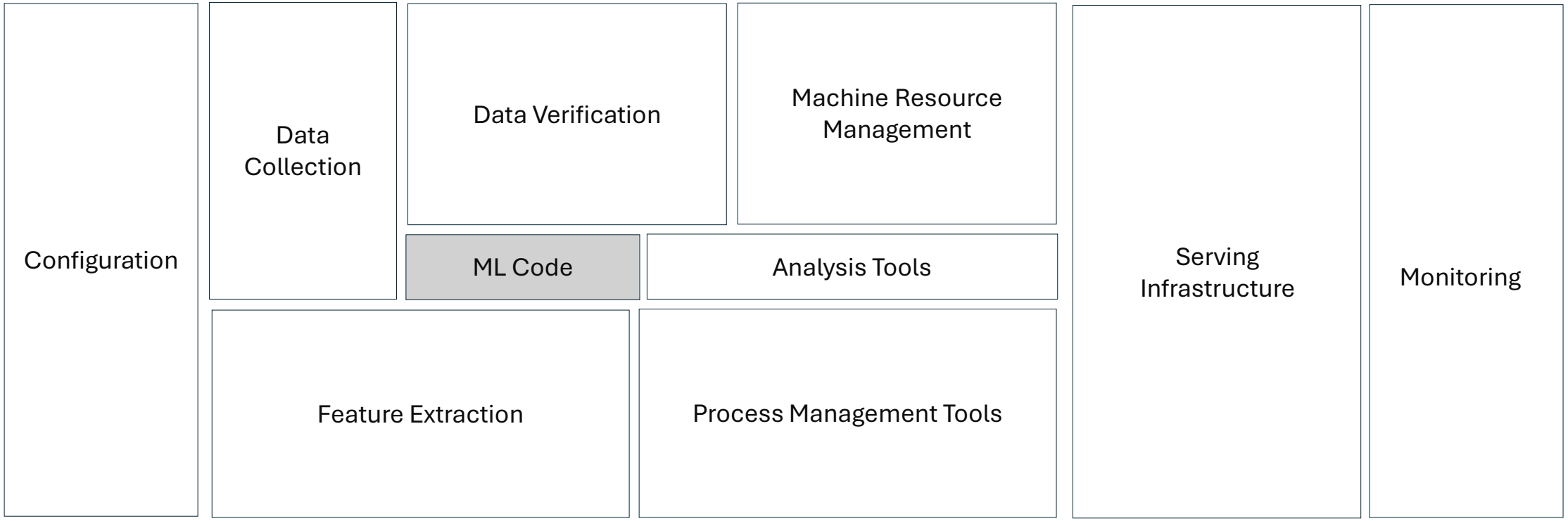


Predictions

The Machine Learning Project Lifecycle



The requirements surrounding ML Infrastructure



[D. Sculley et. al NIPS 2015: Hidden Technical Debt in Machine Learning Systems]

The ML Project Lifecycle