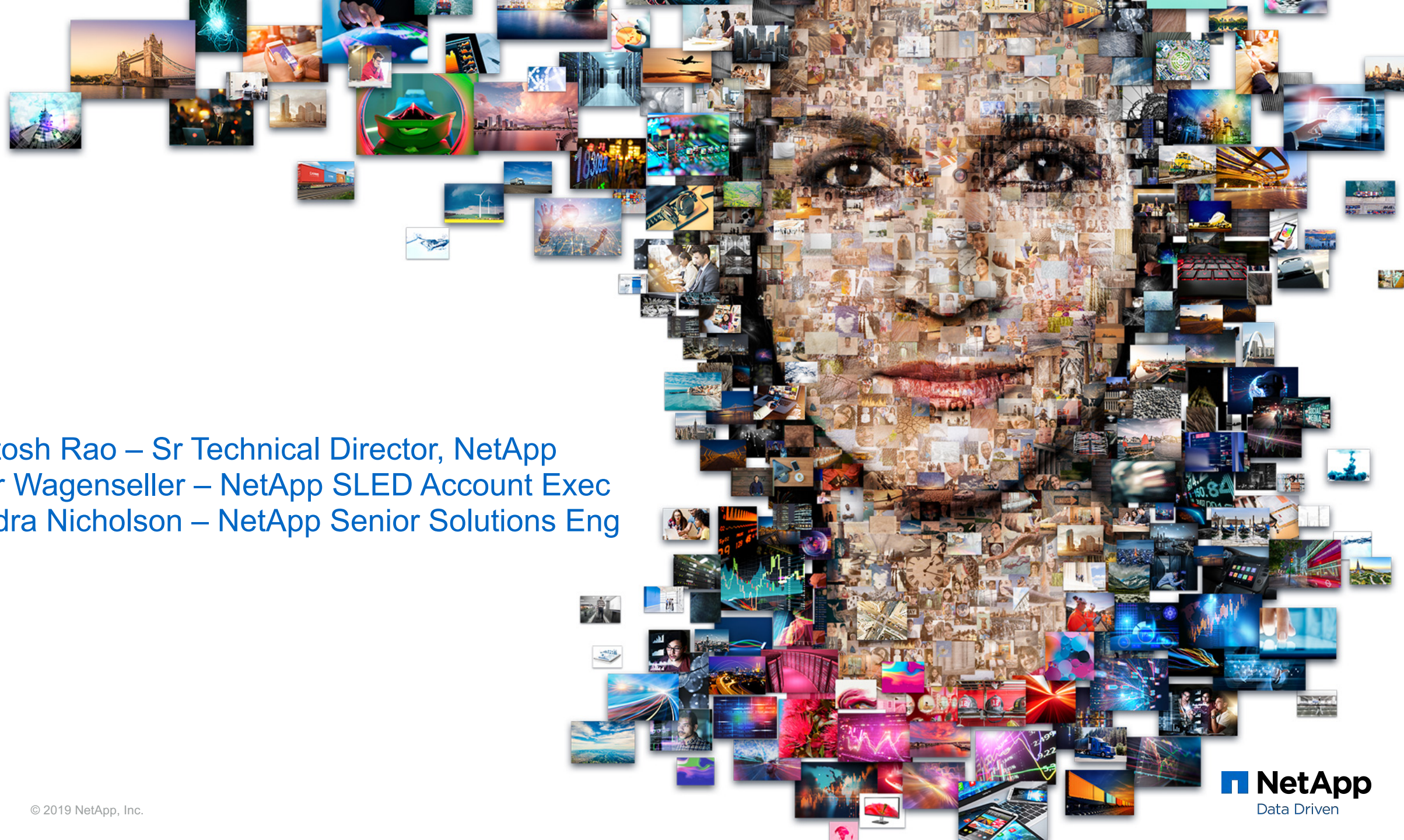
TECH DAYS

AI/DL for Student Success & Digital Transformation

Santosh Rao
Senior Technical Director,
AI & Data Engineering,
NetApp

June 6 – 7, 2019



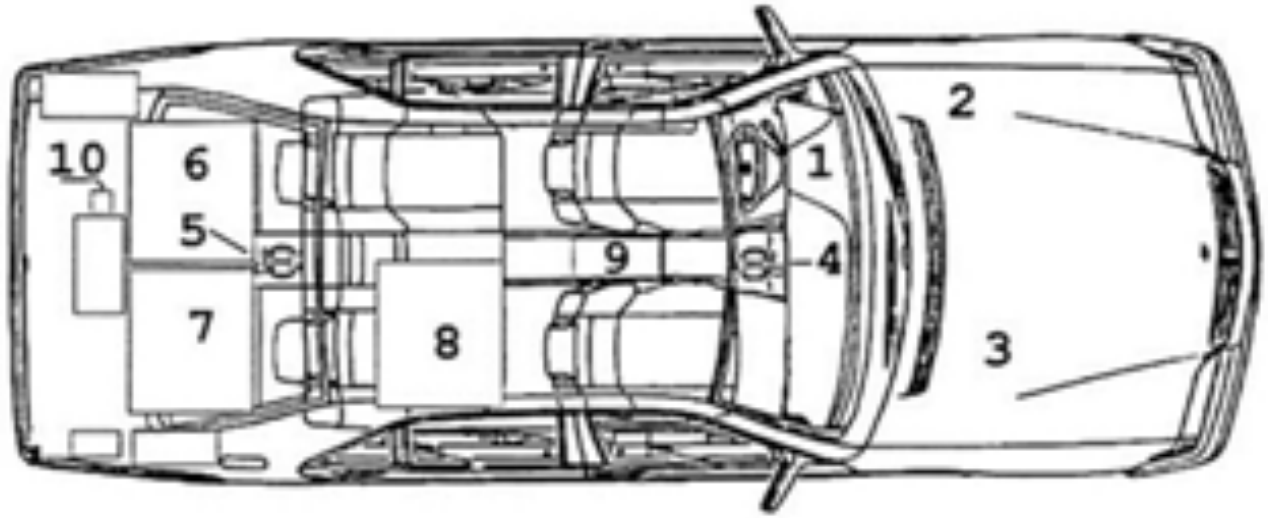
Santosh Rao – Sr Technical Director, NetApp
Tyler Wagenseller – NetApp SLED Account Exec
Sandra Nicholson – NetApp Senior Solutions Eng

NetApp
Data Driven

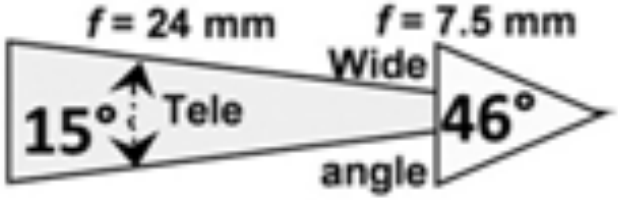
NetApp

What do we do?	▪ #1 Provider of storage to the US Federal Govt. ▪ #1 Storage solution options with AWS/Azure/Google ▪ #1 Branded Storage OS ▪ #1 Storage & Device Management Software ▪ #1 Integrated Infrastructure & Certified Reference Systems in Capacity Shipped ▪ Leader in Storage for AI, ML, DL, DevOps ▪ Leader in Object Storage
Some Focus Areas	▪ AI, ML, DL, Analytics, DevOps ▪ Enterprise Flash ▪ HCI for End User Computing and Cloud Data Solutions
Founded in 1992	▪ Fortune 500 in business since 1992

NetApp devices are installed in some of the worlds largest and exotic environments such as the Hadron Supercollider, the largest machine ever built by mankind, Lawrence Livermore Labs with the worlds largest contiguous file system, the worlds largest database at SAP and even in the International Space Station.



1 electrical steering motor
2 electrical brake control
3 electronic throttle
4 front pointing platform for CCD-cameras
5 rear pointing platform
6 Transputer Image Processing system
7 platform and vehicle controllers
8 electronics rack, human interface
9 accelerometers (3orthogonal)
10 inertial rate sensors



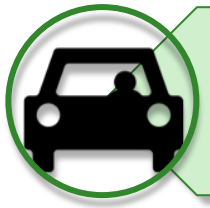
At distance L_5 ~ 20 m (~ 60 m), the resolution is 5 cm/pixel

NetApp


LEFT REARWARD VEHICLE CAMERA
MEDIUM RANGE VEHICLE CAMERA
RIGHT REARWARD VEHICLE CAMERA
MOTION FLOW
LANE LINES
LANE LINES
ROAD FLOW
IN-PATH OBJECTS
ROAD LIGHTS
OBJECTS
ROAD SIGNS

AI - Impactful

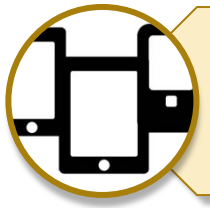
Rapid AI adoption world wide



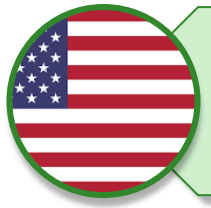
300,000 lives saved by autonomous vehicles per decade

4 Billion AI-powered devices w/ voice-assistants 2019

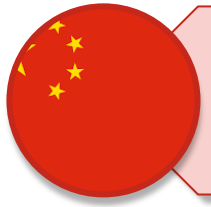
$59.8 Billion Growth of AI software market in 2025



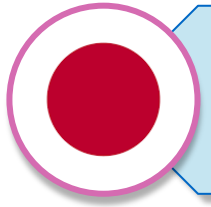
1 Billion AI-enabled video cameras by 2020



$10B in venture capital for AI



190% AI patents grew by over 5 yrs

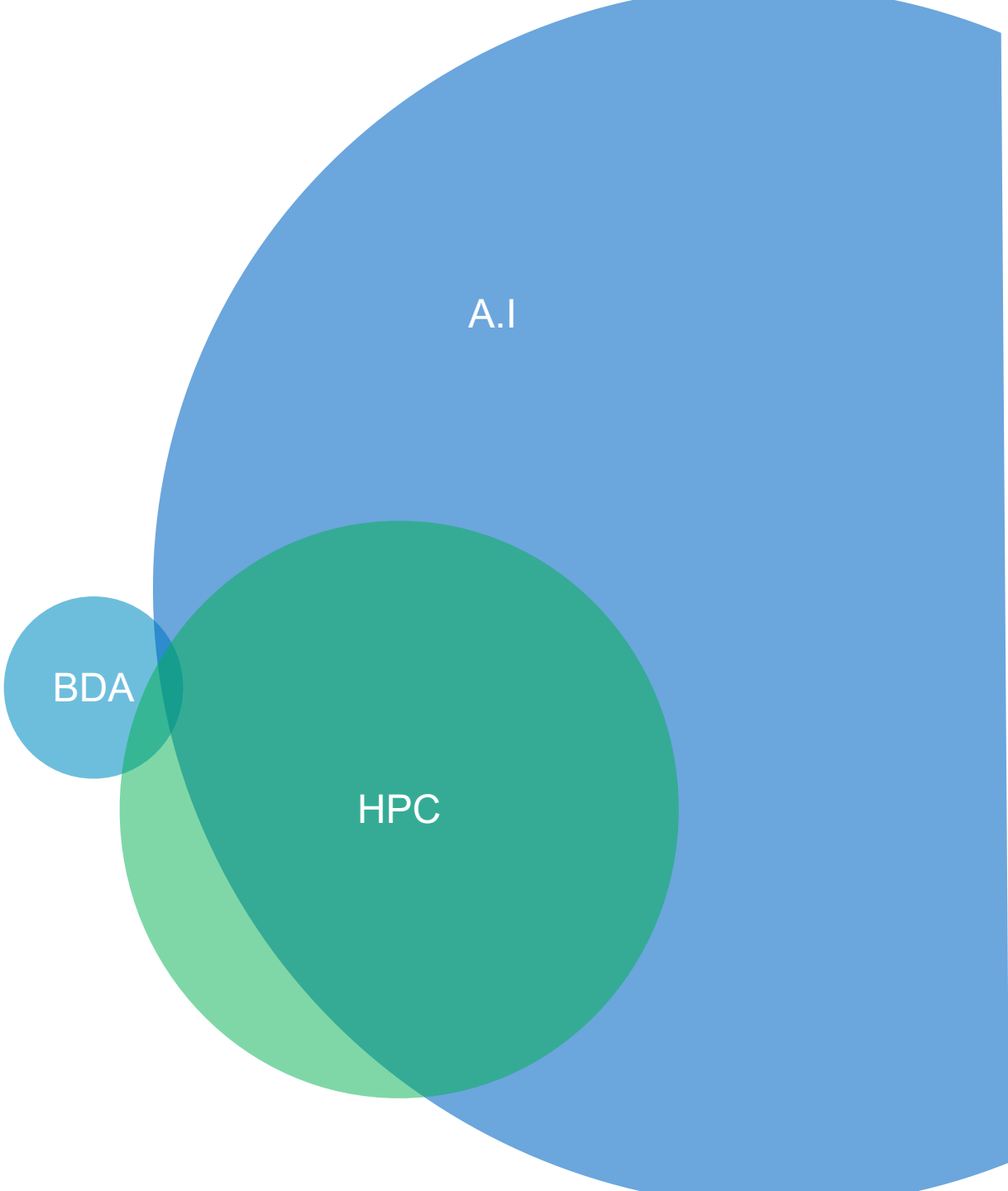


71% Automation potential in Manufacturing

4th largest number of AI startups in Berlin

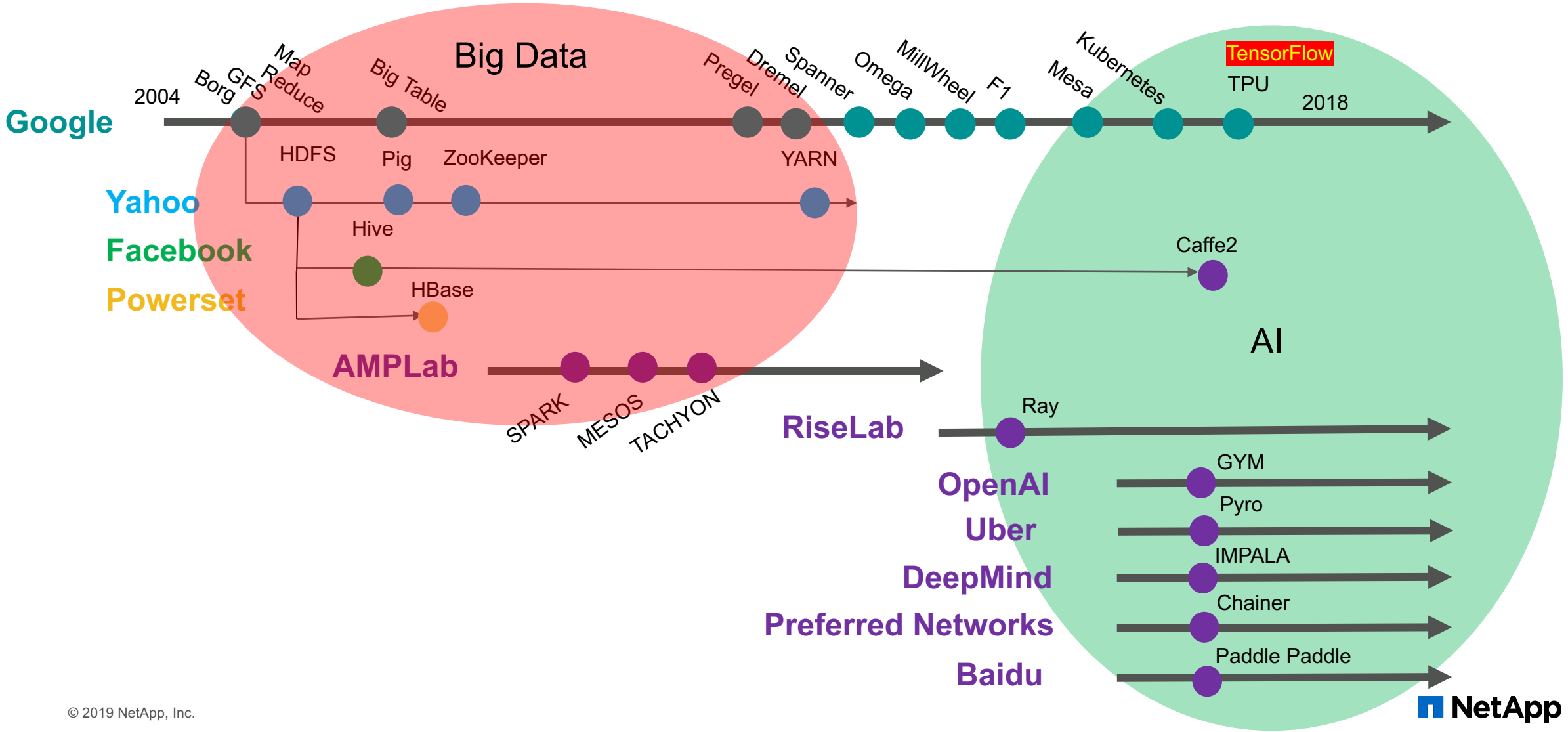
Source: Ihttps://techjury.net/stats-about/ai/

BDA	HPC	AI
Designed for Processing Log Files	Designed for Particle Physics Simulations	Designed for Deep Learning
Evolved as a Dataware house Big Query Big Reporting	Evolved for ML, Simulation and all kind of science experiments	Evolved to do Deep RL Deep GM Deep Sequence
Limited ML, No DL NO RL	**MPI**, ML, DL Energy Efficiency	Cuda,**NCCL**, ML.DL.RL



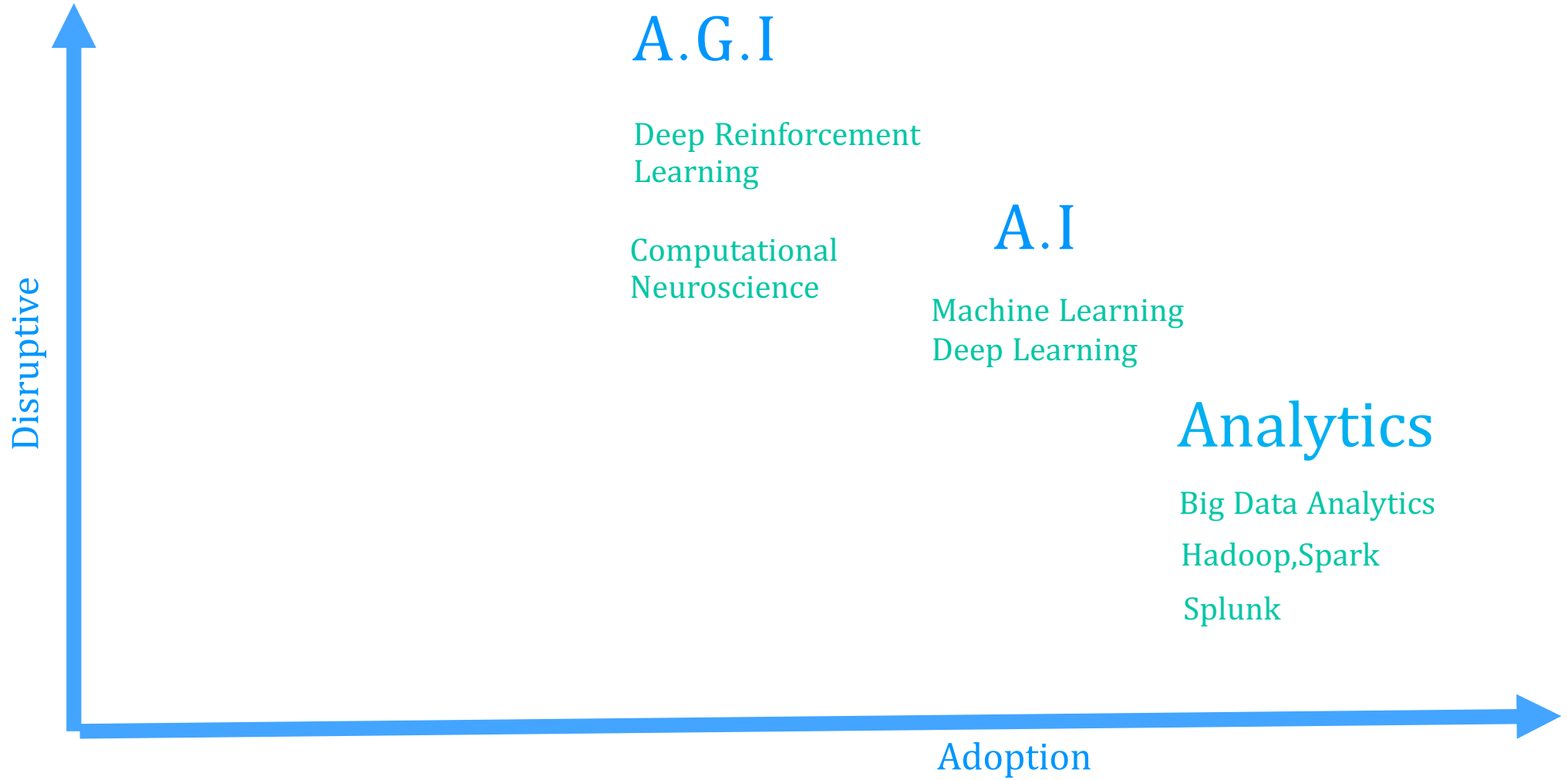
Refining Analytics doesn't lead to Artificial Intelligence

Analytics vs A.I

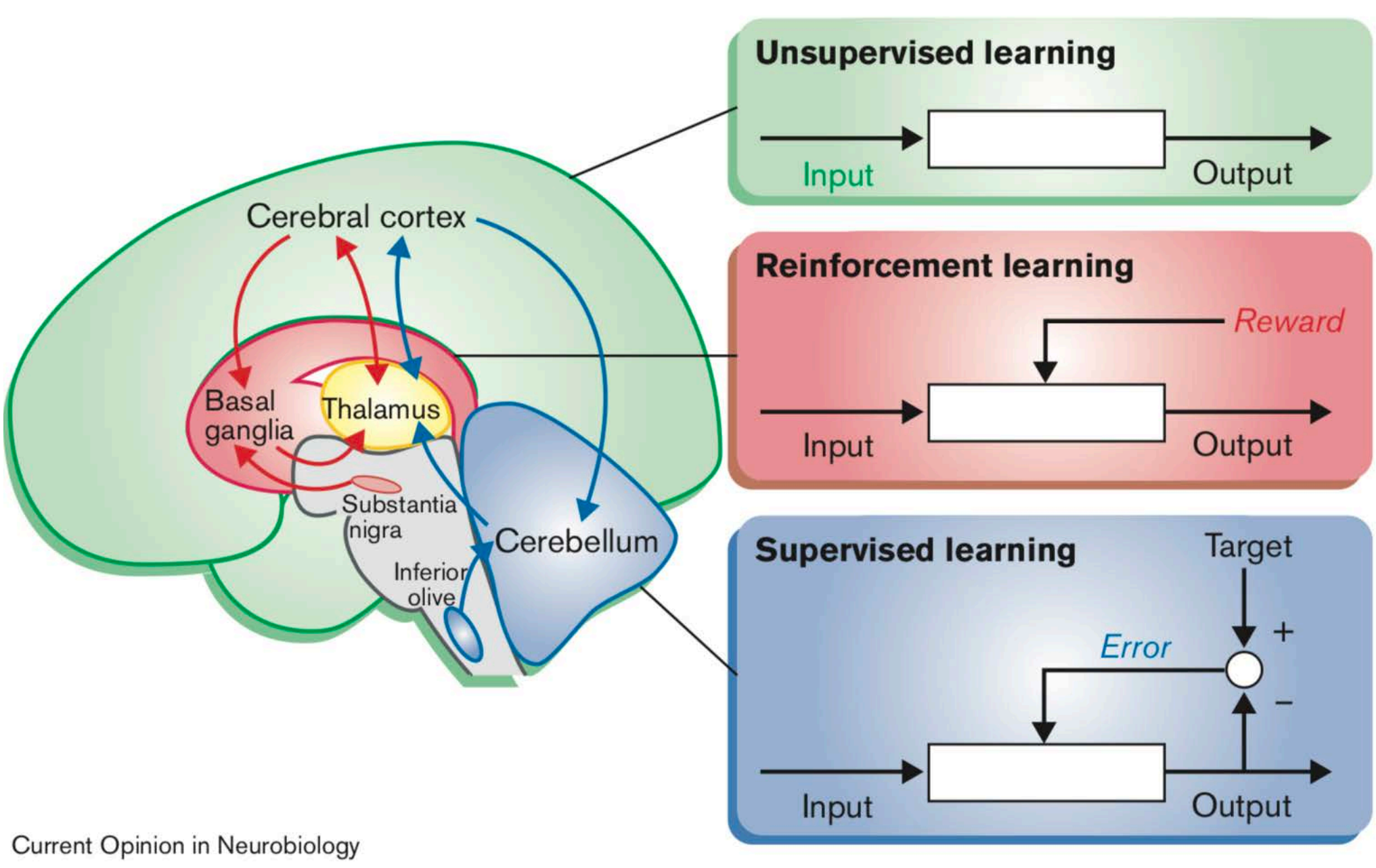


NetApp

2020



NetApp



Source: Kenji Doya Complementary roles of basal ganglia and cerebellum in learning and motor control

Cellular Network Traffic Scheduling with Deep Reinforcement Learning

Sandeep Chinchali [1], Pan Hu [2], Tianshu Chu [3], Manu Sharma [3], Manu Bansal [3], Rakesh Misra [3]
Marco Pavone [4] and Sachin Katti [1,2]
[1] Department of Computer Science, Stanford University
[2] Department of Electrical Engineering, Stanford University
[3] Uhana, Inc.
[4] Department of Aeronautics and Astronautics, Stanford University
{csandeep, panhu, pavone, skatti}@stanford.edu, {tchu, manusharma, manub, rakesh}@uhana.io

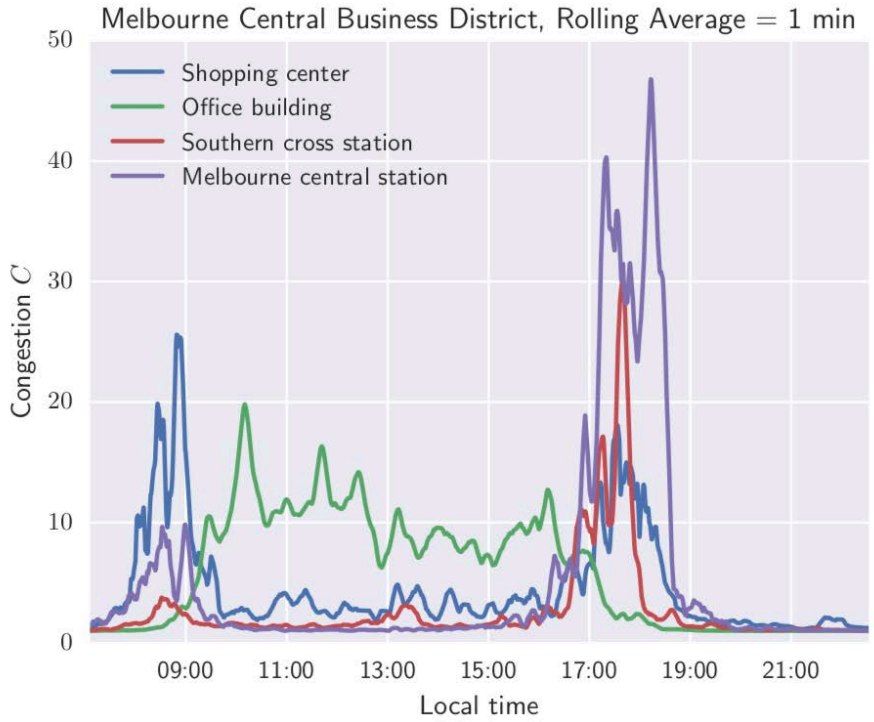


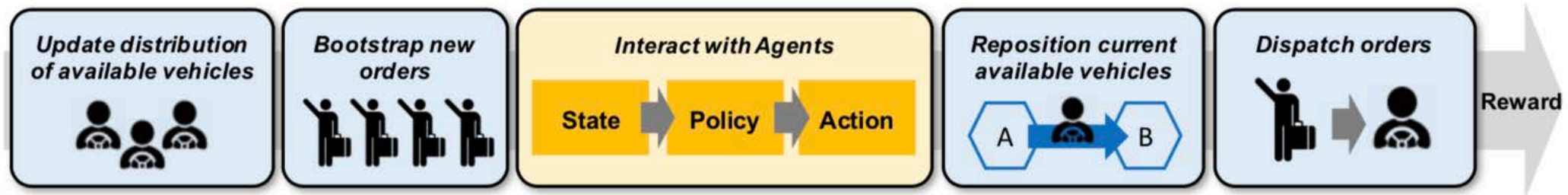
Figure 1: Time-variant congestion patterns in Melbourne.

Efficient Large-Scale Fleet Management via Multi-Agent Deep Reinforcement Learning

Kaixiang Lin
Michigan State University
linkaixi@msu.edu

Renyu Zhao, Zhe Xu
Didi Chuxing
{zhaorenyu,xuzhejesse}@didichuxing.com

Jiayu Zhou
Michigan State University
jiayuz@msu.edu



Human-level control through deep reinforcement learning

Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

Curriculum Learning

Yoshua Bengio[1] YOSHUA.BENGIO@UMONTREAL.CA
Jérôme Louradour[1,2] JEROMELOURADOUR@GMAIL.COM
Ronan Collobert[3] RONAN@COLLOBERT.COM
Jason Weston[3] JASONW@NEC-LABS.COM
(1) U. MONTREAL, P.O. BOX 6128, MONTREAL, CANADA (2) A2IA SA, 40BIS FABERT, PARIS, FRANCE
(3) NEC LABORATORIES AMERICA, 4 INDEPENDENCE WAY, PRINCETON, NJ, USA

When a large language model is trained on a sufficiently large and diverse dataset it is able to perform well across many domains and datasets. GPT-2 zero-shots to state of the art performance on 7 out of 8 tested language modeling datasets. The diversity of tasks the model is able to perform in a zero-shot setting suggests that high-capacity models trained to maximize the likelihood of a sufficiently varied text corpus begin to learn how to perform a surprising amount of tasks without the need for explicit supervision.[5]

Special computation properties

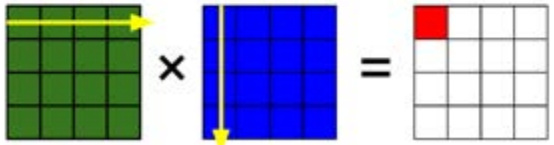
reduced precision ok

about 1.2 × about 0.6 = about 0.7

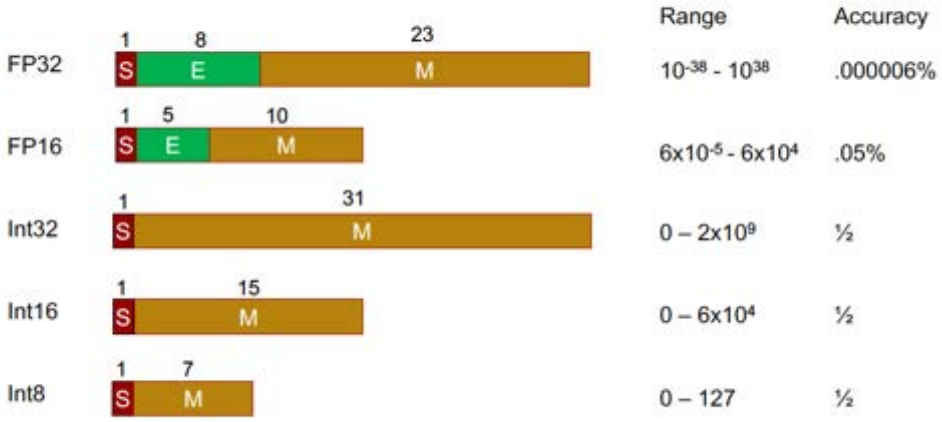
NOT

1.21042 × 0.61127 = 0.73989343

handful of specific operations



Number Representation



- Deep Learning is Empirically Scaleable (Baidu)
- Computationally Homogenous
- Constant runtime & memory use
- Highly Portable
- Easily Baked into silicon → “Semiconductor Rennaisance”

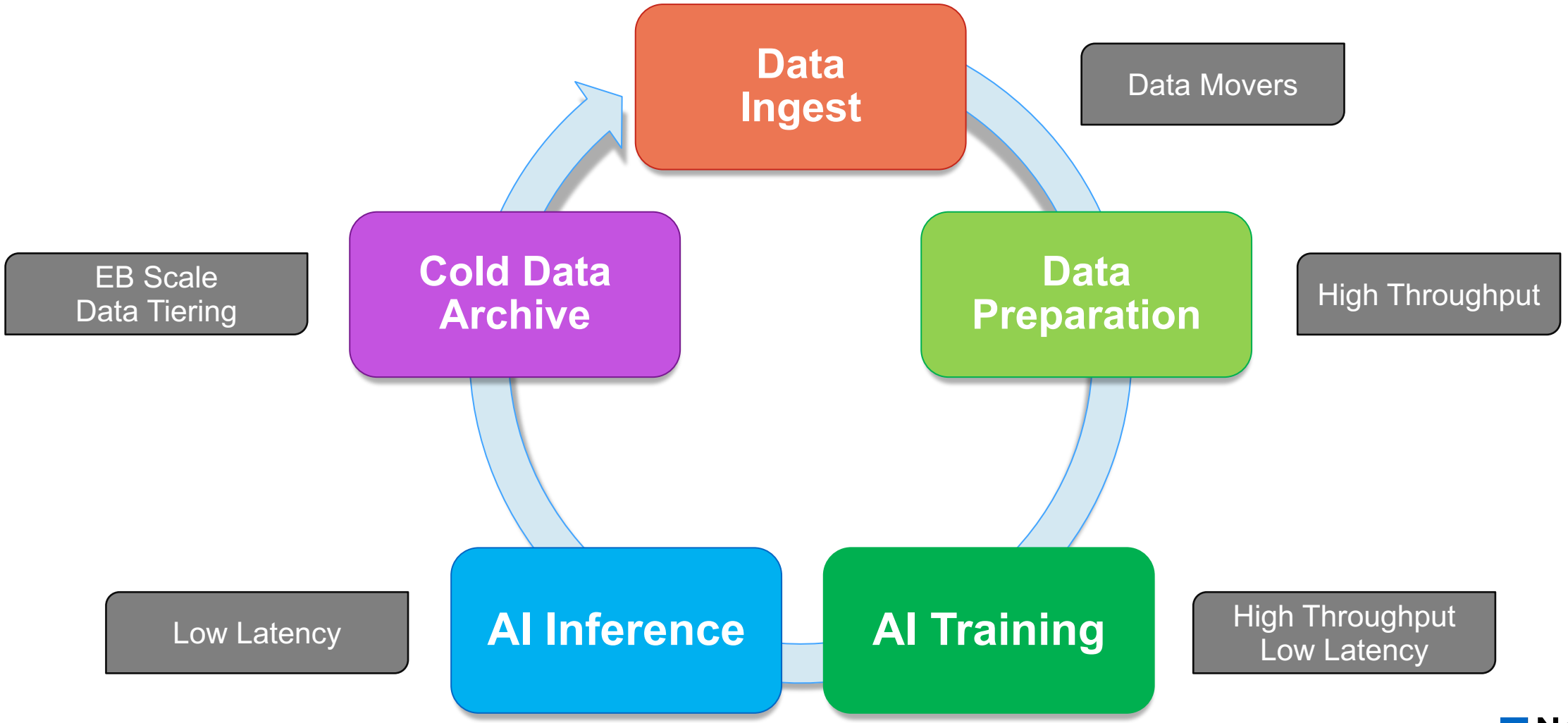
CPU	GPU	FPGA	TPU
Threads	Massive threads	LUTs	MM unit
SIMD	SIMD	DSP	BRAM
	HBM	BRAM	

- **Relax Precision** : Small integers are better
- Relax Synchronization : data races are better
- Relax Communication : sparse communication is better
- Relax Cache Coherence : incoherence is better

“Olukutan,Stanford Neurips Keynote 2018”

Data Pipeline for AI Workflow

Scale and Optimize Each Stage of the Data Pipeline



NetApp

Full Scale-out ONTAP AI with DGX-1

24-node A800 cluster, driving 108 DGX-1's

NetApp

Full Scale-out ONTAP AI with DGX-2

24-node A800 cluster, driving 36 DGX-2's

NetApp

Dense Cabinet for AI

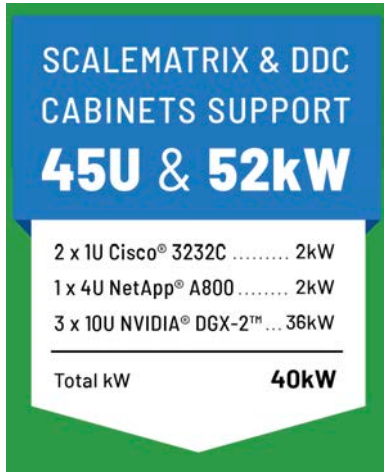
Deliver Dense Cabinet for Exascale AI

Challenge:

- Data centers facilities lack the power and cooling for the latest high-performance AI computing infrastructure.

New Capability:

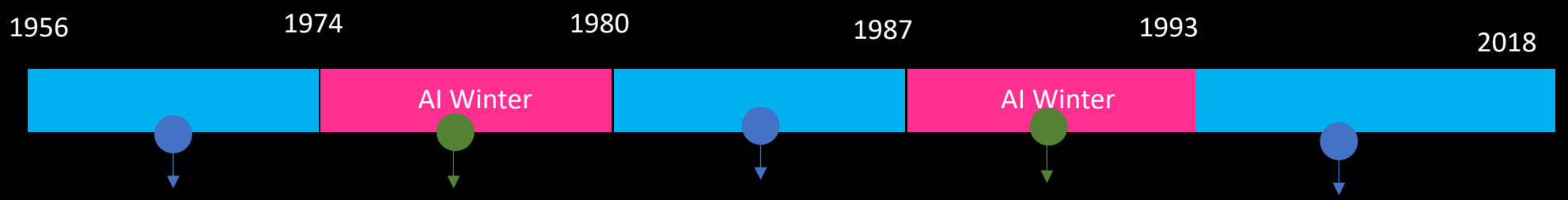
- Dense and Modular Dynamic Density Control (DDC) liquid-air cooled cabinet for AI
- This cabinet combines the efficiency of water with the flexibility of air, cooling up to 52kW of power load in a 45U cabinet.
- Cabinets can be deployed in any environment.
- Provides clean-room environment, guaranteed air flow, integrated security, and fire suppression.



ARTIFICIAL INTELLIGENCE
Trends >

Artificial Intelligence Timeline

Why Now?



Rosenblatt's' Percepteron

Backpropogation
Q-Learning

LSTM
Alpha GO

NVIDIA DGX
Google TPU
GraphCore

IMAGENET
BDD100K

TESLA Autopilot
Amazon Alexa

TensorFlow
Caffe
Chainer
PyTorch

Mainframes

Intel 4004

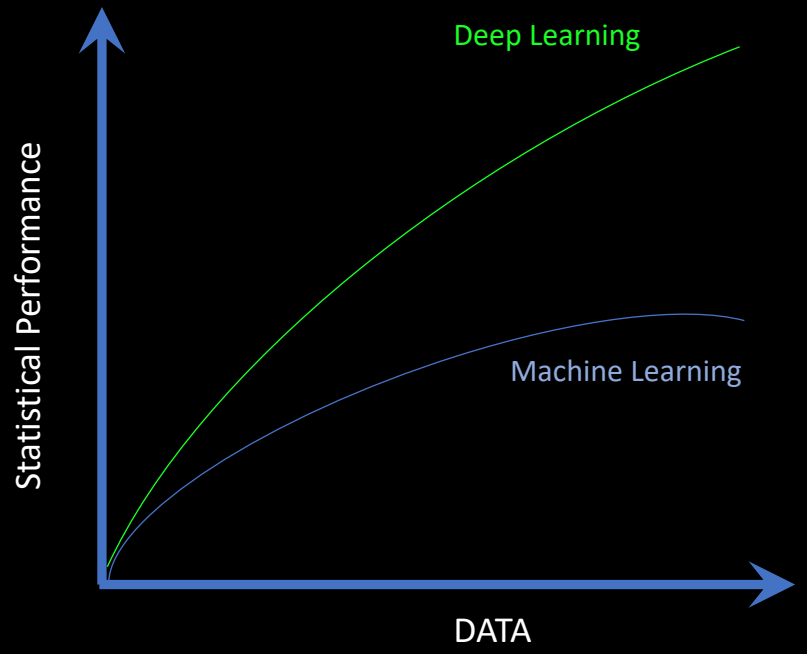
Intel 80386

Intel Pentium

Artificial Intelligence

1956 1974 1980 1987 1993 2018

AI Winter AI Winter



Deep Learning

Deep Reinforcement Learning

Artificial Intelligence

Deep Learning

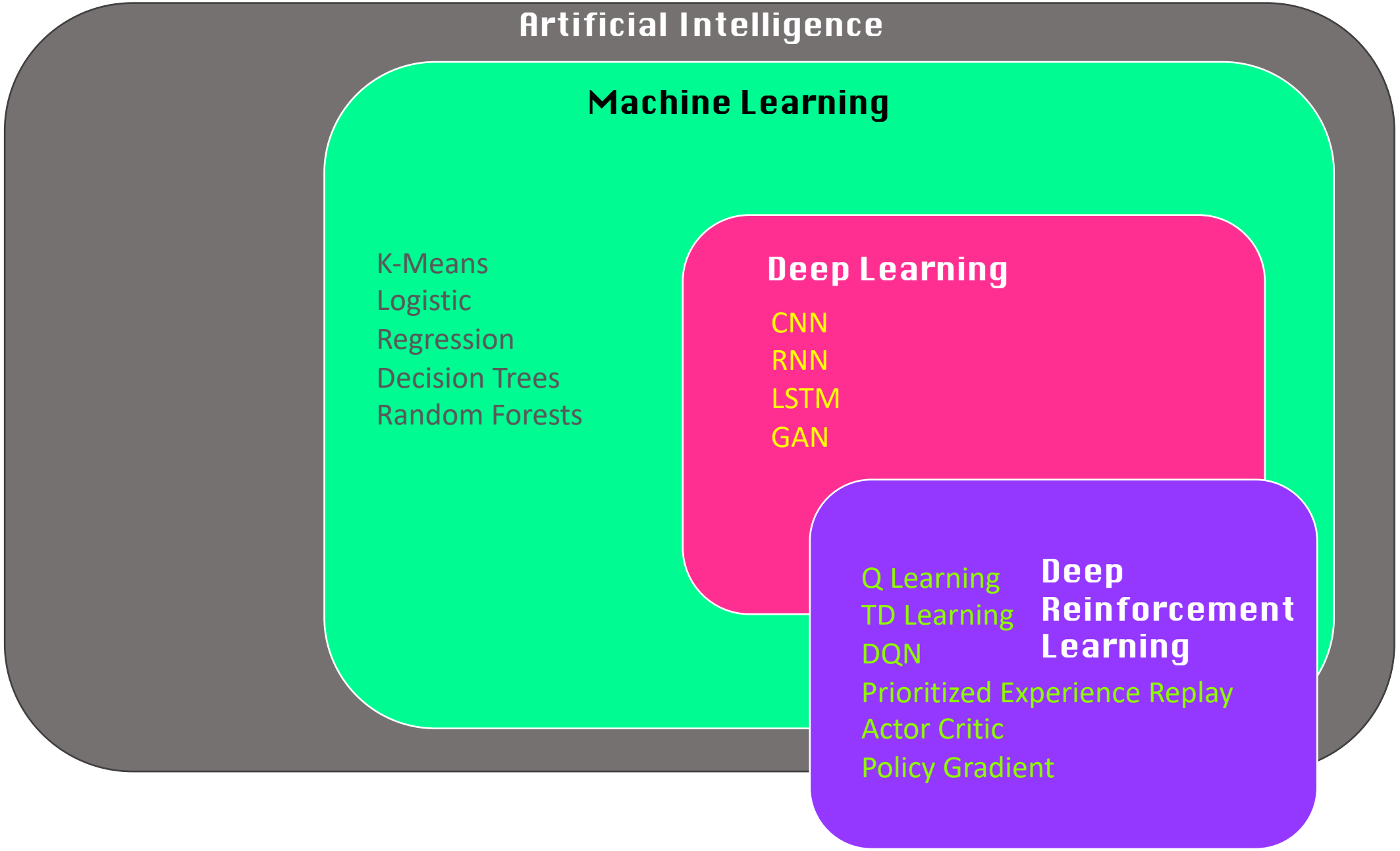
Deep Reinforcement Learning

Bayesian Deep Learning BNP
Meta-Learning Imitation-Learning
Continual-Learning Causal-Learning
Relational Representation-Learning
Smooth GamesOptimization
Probabilistic reinforcement learning

Conversational AI Systems ML Emergent communication
Modelling Physical World Modelling Spatio Temporal domain

Physics,Geophysics,Geochemical Molecules & Materials
Financial Services Intelligent Transport Systems
Medical Imaging & Healthcare

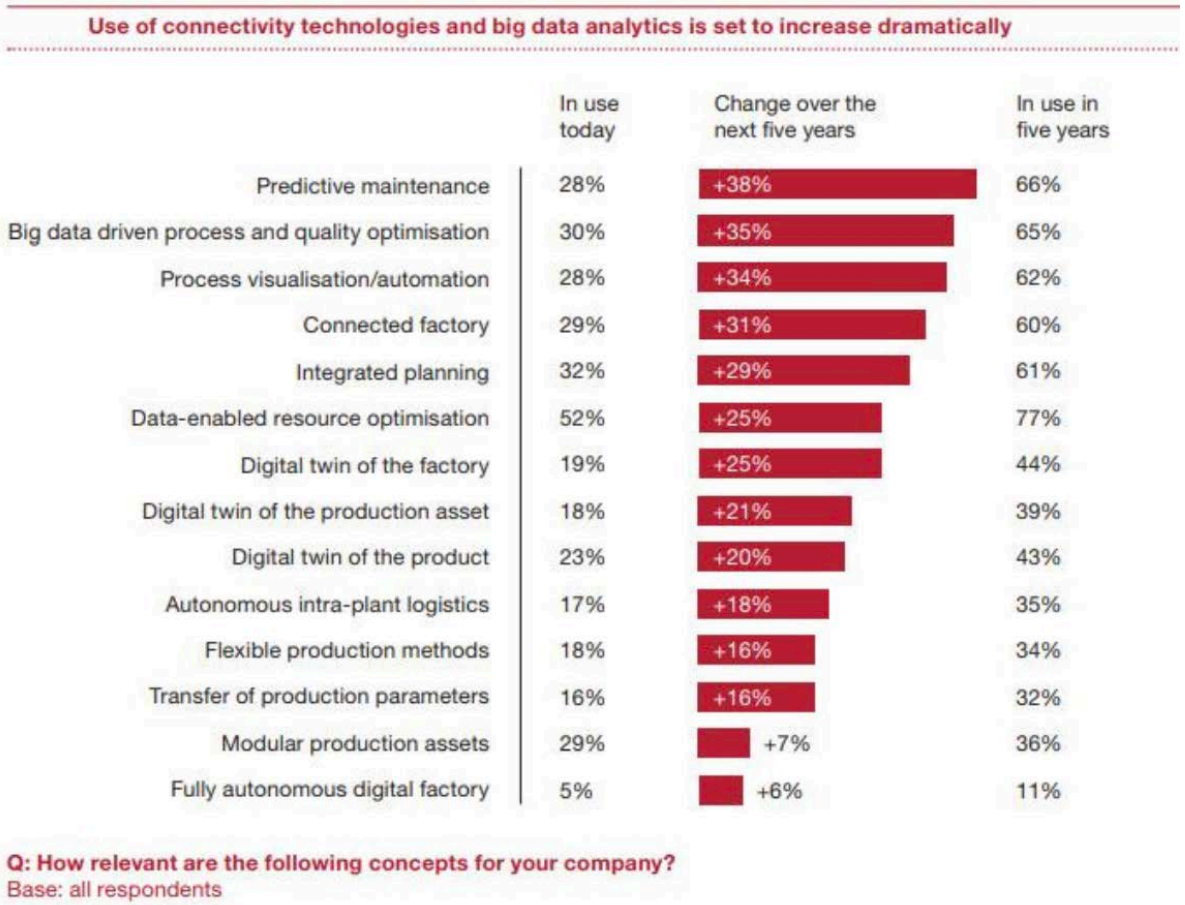
Robustness Compactness Interoperability Security
Social,Political,Ethics


Artificial Intelligence
Machine Learning
K-Means
Logistic
Regression
Decision Trees
Random Forests
Deep Learning
CNN
RNN
LSTM
GAN
Q Learning
TD Learning
DQN
Deep Reinforcement Learning
Prioritized Experience Replay
Actor Critic
Policy Gradient

AI Use Cases

AI in Manufacturing

Top use cases

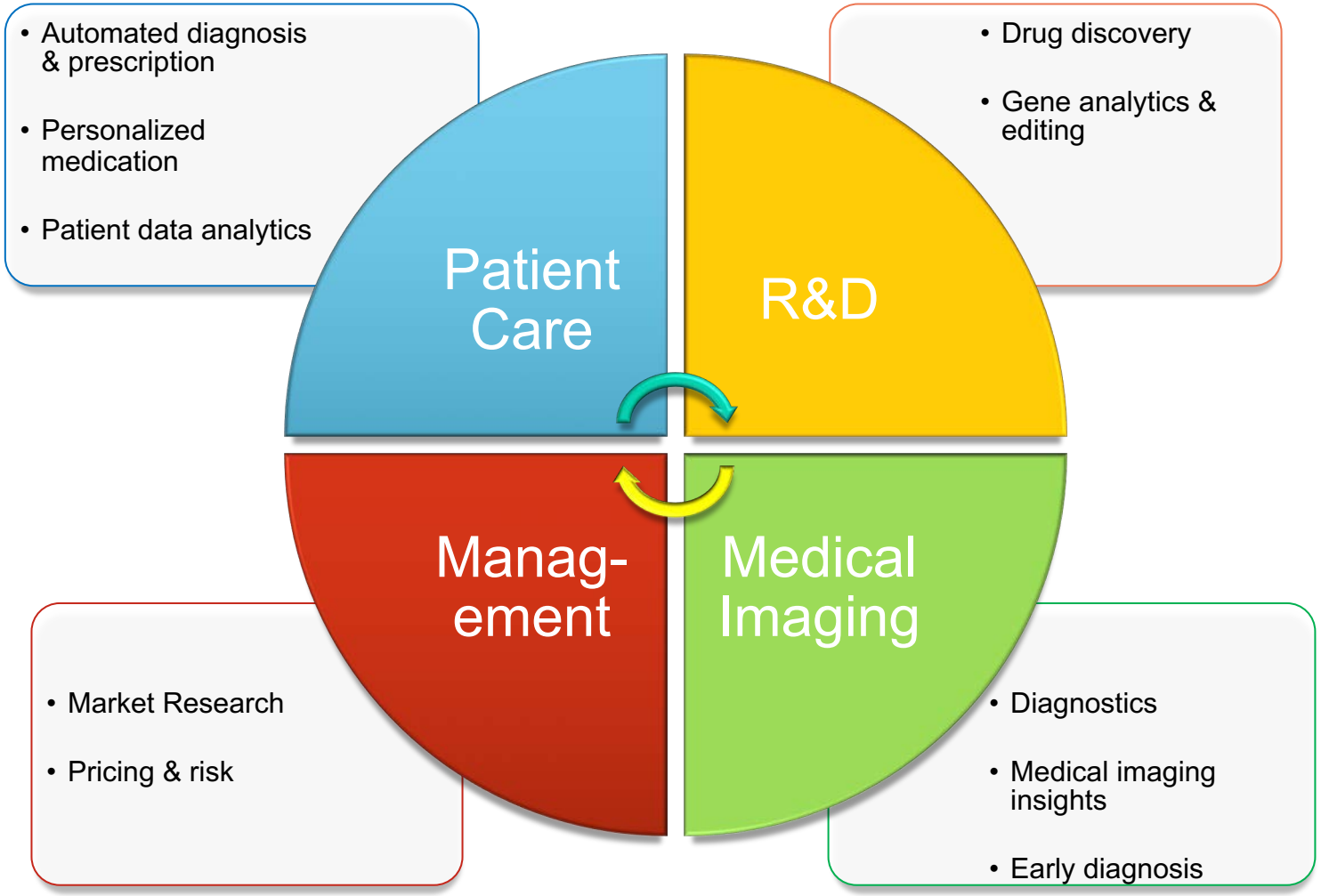


1) Predictive maintenance
   - Predicted increase over 5 years *(PwC)*
2) Quality optimization
3) Process visualization and automation
4) Connected factories
5) Resource optimization
6) Improve product effectiveness

SOURCE: DIGITAL FACTORIES 2020: SHAPING THE FUTURE OF MANUFACTURING (48 PP., PDF, NO OPT-IN) PRICEWATERHOUSECOOPERS

AI in Healthcare

Variety of use cases



- AI applications in healthcare could save up to $150B annually by 2026
- AI health market is expected to reach $6.6B by 2021 (40% CAGR)
- AI can address an estimated 20% of unmet clinical demand

Source: https://www.accenture.com/t20171215T032059Z__w__/us-en/_acnmedia/PDF-49/Accenture-Health-Artificial-Intelligence.pdf#zoom=50

AI in Telecom

Top Use Cases

Chat Bots	• Automate customer service inquires • Routing customers to agents • Routing prospective customers to sales
Speech & Voice Services	• Alternative to remote control units • Allows customers to explore and purchase media content
Predictive Maintenance	• Fix problems w/ hardware (cell towers, power lines etc.) before they break • Detects signals and breakpoints that usually lead to failures (no human intervention)
Network Optimization	• Intelligent network planning and optimization • AI algorithms drive sophisticated network analysis and simulation efforts • Predicts optimal connectivity for telecom networks

AI in Government

Massive savings in labor times

Figure 11. Time and money savings from AI under three levels of investment

Level of investment	Savings category	Federal	State government
Low	Annual person-hours	96.7 million	4.3 million
	Hours as percentage of total	2.23%	3.94%
	Salary	$3.3 billion	$119 million
Medium	Annual person-hours	634 million	15.3 million
	Hours as percentage of total	14.63%	13.93%
	Salary	$21.6 billion	$420 million
High	Annual person-hours	1.2 billion	33.8 million
	Hours as percentage of total	27.86%	30.84%
	Salary	$41.1 billion	$931 million

Source: Deloitte simulation of likely changes to labor inputs to government tasks.

Deloitte University Press | dupress.deloitte.com

- Labor time savings with AI
- 2-4% time savings at low levels of effort
- 13-15% with mid level
- 27-30% time savings within 5-7 yrs at high levels of effort

Source: https://www2.deloitte.com/insights/us/en/focus/cognitive-technologies/artificial-intelligence-government-analysis.html

AI in Retail

Top use cases

Communications	• Personalization & recommendation engines • Chatbots for customer service • Voice shopping with voice enabled devices
Pricing Optimization	• Forecasting and dynamic pricing • Competitive pricing • Analyze sensitivities to price changes
Inventory Management	• Demand Forecasting • Manage inventory levels, reduce losses from out-of-stock & overstock • Allocation & audits
Experiential Retail	• New ways to engage with customers • Discover, Auto-suggestions • Buy and pay

NetApp

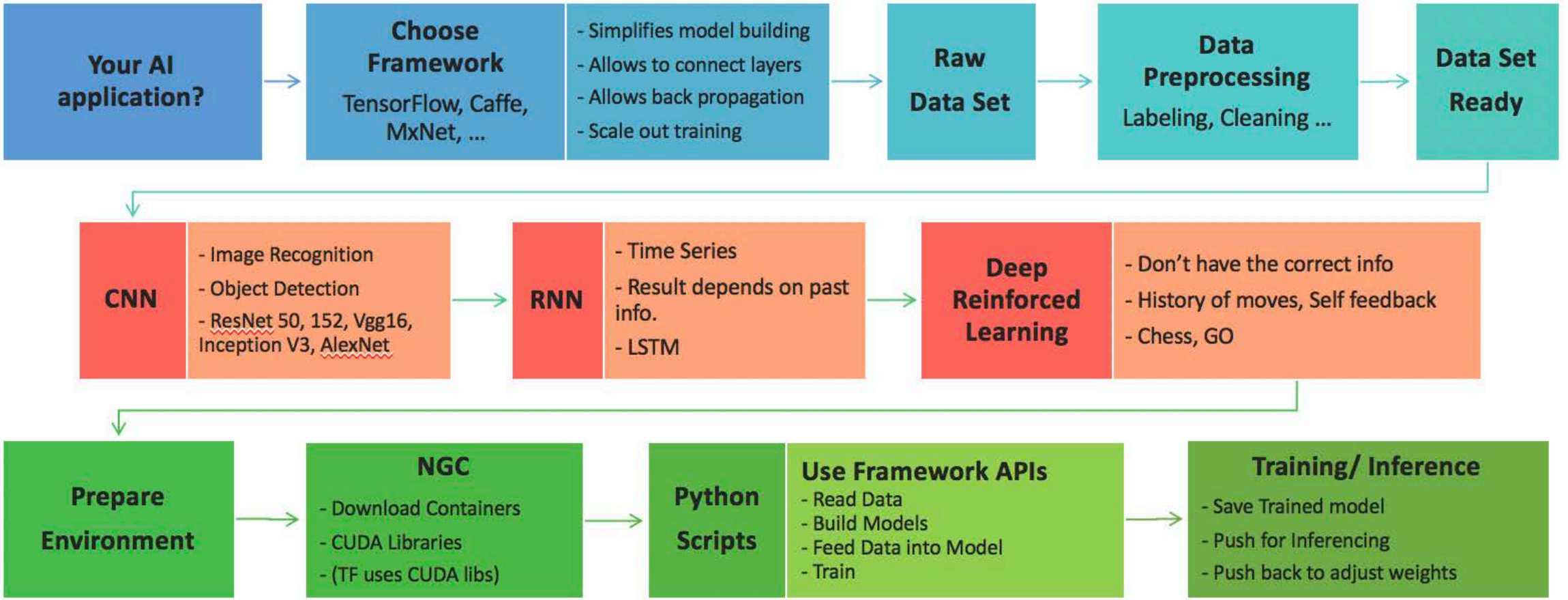
AI in Financial Services

Top use cases

- Front Office
  - Credit Scoring
  - Insurance Premiums
  - Customer Service (“chat-bots”)
- Back Office
  - Risk Management Modeling
    - Stress Testing
    - Model Validation; back testing
  - Capital Optimization
    - Risk Weighted Assets
    - Margin Valuation Adjustment
  - Market Impact Analysis
    - Identify assets that behave similarly
    - Timing / Scheduling of trades
- Trading & Portfolio Management
  - Devise Investment Strategies
  - Trading Execution (Sell-orders)
  - Managing Risk
  - Identify new signals
- Regulatory Compliance
  - Enhance efficiencies of supervision and surveillance

AI Solution Architecture

DL Model Training Flow



Meet Diverse Needs across Data Science and Infra Functions

Data Scientists

Real-world Data for AI / DL

Need Agile Model DevOps :

- Refreshed access to Production Datasets
- Hybrid Cloud for Model Dev
- Distributed Data Science
- Diverse Data Sources
- Model and Data Parallelism
- Multi-Tenant Model Serving
- From Model to Application

Data Architects

Future Proof Architecture

Seek Extensible Architecture:

- Architecture Scales from PoC to Production
- Future Proof to absorb technology changes
- TCO for Massive Datasets
- Maximize Utilization
- Global Scale

Data/IT Admins

Lowest TCO in face of shrinking budgets

Balance Cost & TTM :

- Leverage vs. Dedicate HW Infra
- Stable Operations & Upgrades
- Supported Components & Ecosystems
- Diverse needs across Big Data, AI/DL and HPC

NetApp

Deployment Options for AI

Where is the source data?

Move Data into AI Platform

- 80 - 90% deployments
- HDFS, Splunk, NoSql , Lustre, GPFS → ONTAP AI (NFS, GPUs)
- Leverage Data Movers to move data into ONTAP AI
- FabricPool for data tiering

Data In-Place

- All reside and deploy on ONTAP
- Concept of Unified Data Lake
- Data on ONTAP AI
- FabricPool for data tiering

Co-Lo Solution

- Greater control of data
- NPS solution
- Data on NPS, GPUs/ Services on the Cloud

Source Data in Cold Storage

- Data is moved in from cold data tiers for model training
- Move data from StorageGrid into ONTAP AI

Cloud Deployment

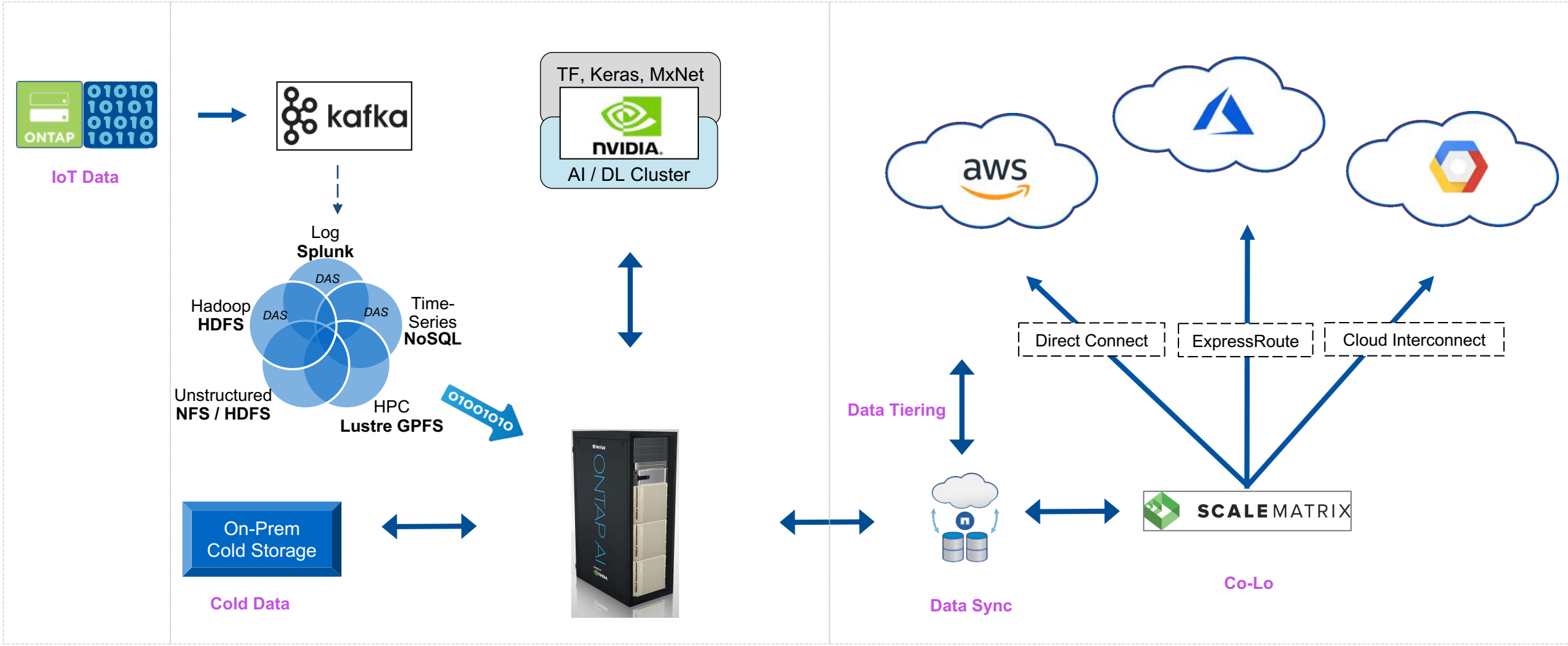
- Data / GPUs provisioned on the public clouds
- Use Cloud Volumes Service for file services
- GPUs on Cloud for compute

NetApp

Move Data to AI Platform

Data movement from HDFS, MapR-FS, GPFS, Lustre, S3 to AI Platform

Move Data into ONTAP AI

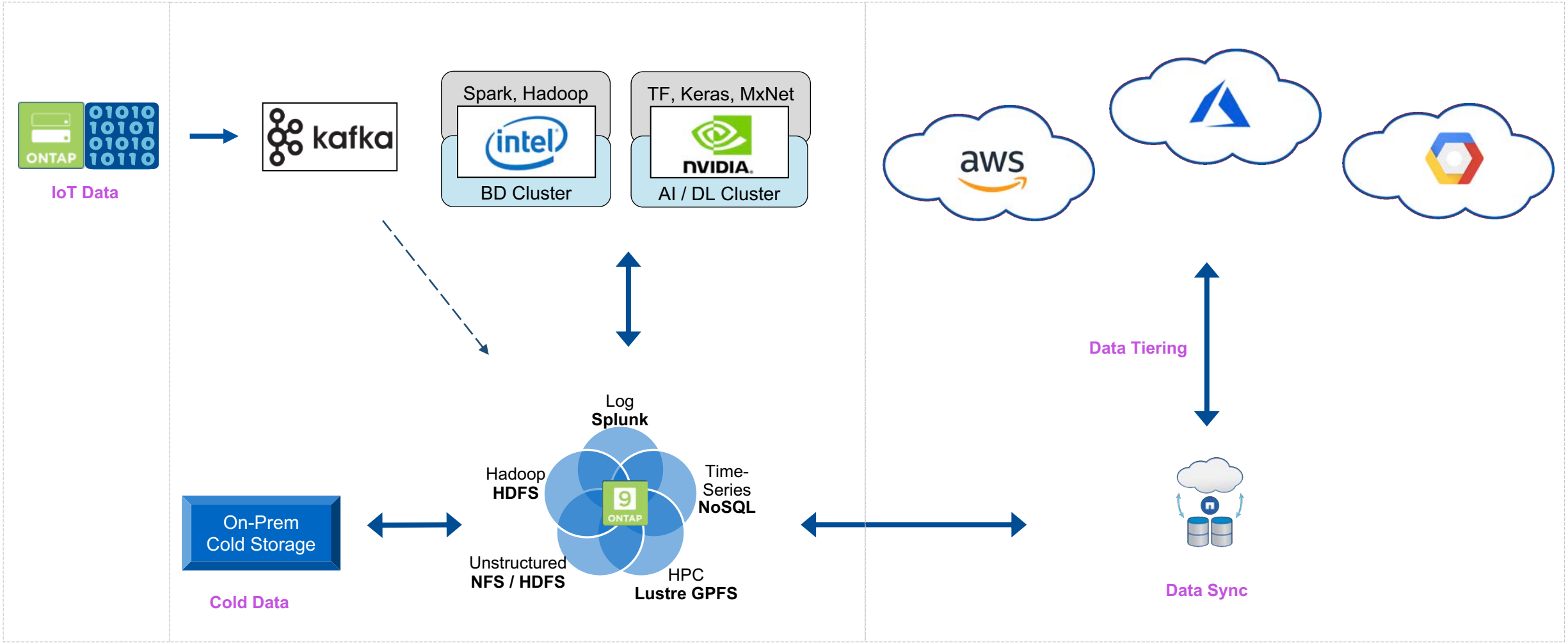


NetApp

In-Place Data with Hybrid Cloud Option

Unified data lake serving CPU and GPU Compute Clusters

Data In-Place



NetApp

Resources

#NetAppAI

netapp.com/ai

Technical Whitepapers

- ONTAP AI Reference architecture NVA-1121-design
- ONTAP AI Deployment guide NVA-1121-deploy
- CVD: FlexPod Datacenter for AI/ML design guide
- Building a Data Pipeline for Deep Learning WP-7299
- Edge to Core to Cloud white paper WP-7271
- AI with GPUs on AWS & Cloud Volumes Service TR-4718
- Scalable AI Infrastructure WP-7267
- Designing data pipeline for your AI workflows WP-7264
- ONTAP AI Solution brief SB-3939
- IDC Technology Spotlight paper

AI Blogs

- Your Guide to Everything NetApp at GTC 2019
- AI Across Industries: Manufacturing, Telecom, & Healthcare
- How to Configure ONTAP AI in 20 Minutes with Ansible
- Bridging the CPU and GPU Universes
- Is Your Infrastructure Ready for AI Workflows in Production?
- Accelerate I/O for Your Deep Learning Pipeline
- Addressing AI Data Lifecycle Challenges with Data Fabric
- Choosing an Optimal Filesystem for the AI Pipeline
- Five Advantages of ONTAP AI for AI and Deep Learning
- Deep Dive into ONTAP AI Performance and Sizing

Thank You

Questions?