AI/ML 을 위한 시스코 데이터 센터 솔루션
- Edge to Multi Cloud

시스코 시스템즈

Datacenter PSS 정 연구 수석

2018.11.08

The Data Explosion Is Real

277X
Data created by IoT devices vs. end users

40%
Of all data will come from sensor data by 2020

30M
New devices connected every week

5TB+
Of data per person by 2020

180B
Mobile apps downloaded in 2015

4.2B
Web filtering blocks per day

AI activates the potential of raw data into powerful competitive advantage

- 277X data created by IoE devices vs. end users – source: 2014 Cisco® Global Cloud Index
- By 2020, there will be 5200 GB of data for every person on earth – source: 2012 Digital Universe Study conducted by IDC and sponsored by EMC (see: http://www.computerworld.com/article/2493701/data-center/by-2020--there-will-be-5-200-gb-of-data-for-every-person-on-earth.html)
- 180 billion mobile app downloads by 2015 – source: 2011 IDC Study: https://www.smaato.com/blog-180billiondownloads/

How Important is AI & ML?

By 2020, insights-driven businesses will steal

$1.2T

per annum from their less-informed peers

8 out of 10

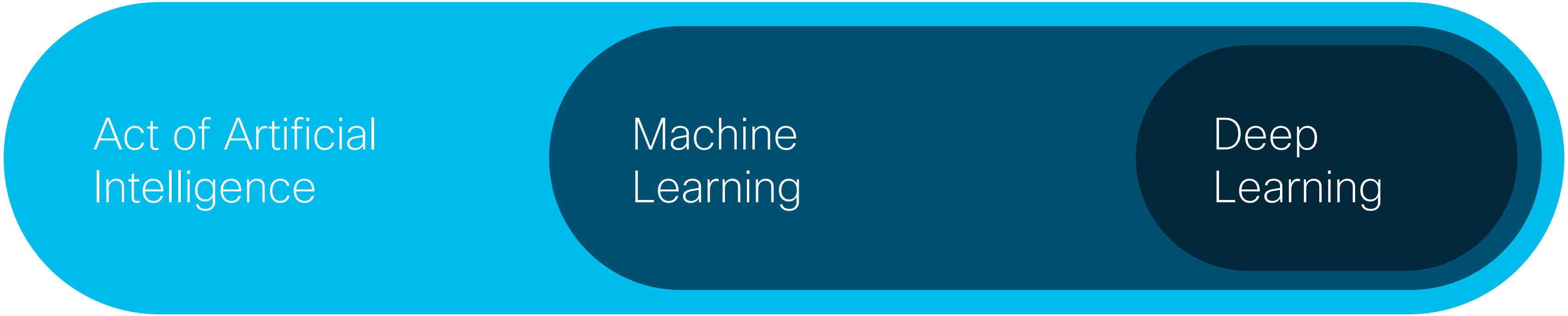
businesses have already implemented or are planning to adopt AI as a customer service solution by 2020

By 2035, AI technologies are projected to increase business productivity by up to

40%

- $1.2T https://go.forrester.com/wp-content/uploads/Forrester_Predictions_2017__-Artificial_Intelligence_Will_Drive_The_Insights_Revolution.pdf
- 8 out of 10 - Oracle - https://www.oracle.com/webfolder/s/delivery_production/docs/FY16h1/doc35/CXResearchVirtualExperiences.pdf
- Accenture https://www.accenture.com/us-en/insight-artificial-intelligence-future-growth

What is AI



Act of Artificial Intelligence

Machine Learning

Deep Learning

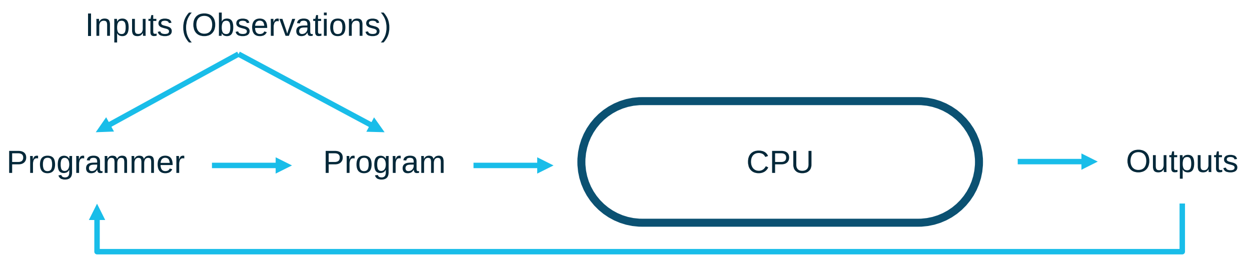
Machines Making Decisions

Machines that Learn and Make Decisions without Explicit Programming

Machines that Use Artificial Neural Networks to Learn and Make Decisions with Complex Data

Data is the Source Code

Traditional Programming



Inputs (Observations)

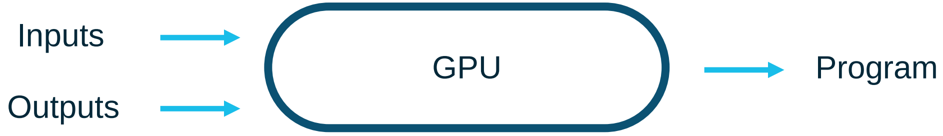
Programmer

Program

CPU

Outputs

Machine Learning and Deep Learning



Inputs

Outputs

GPU

Program

Source: Sebastian Raschka - https://www.kdnuggets.com/2016/05/explain-machine-learning-software-engineer.html

Evolution of AI Algorithms

1950s

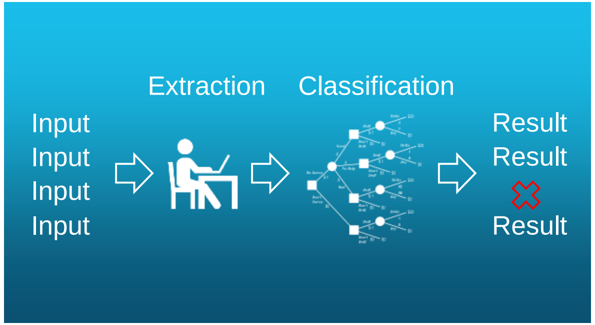
Rule-Based Analytics

```
if:
            input >= example
            in time < timedelta
            in > 5 locations
then:
            result
```

- Simple
- Low Accuracy
- High Rate of False Positives
- Only Applicable to Simple Data
- Ineffective on Big Data

1980s

Machine Learning



Extraction Classification

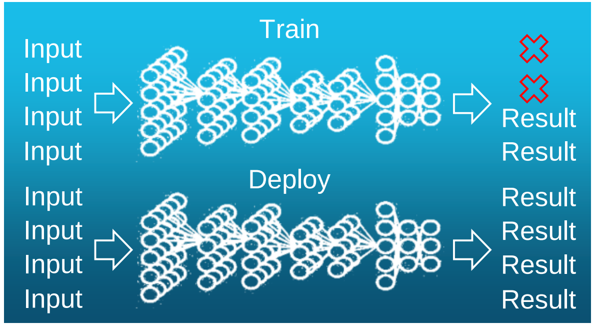
Input
Input
Input
Input

Result
Result
✗
Result

- Can Approach Human Level Accuracy
- Requires Expert Feature Extraction Engineering
- Requires Moderate Volume of Data for “Learning”
- Good for Moderate Variety of Data

2010s

Deep Learning



Train

Input
Input
Input
Input

✗
✗
Result
Result

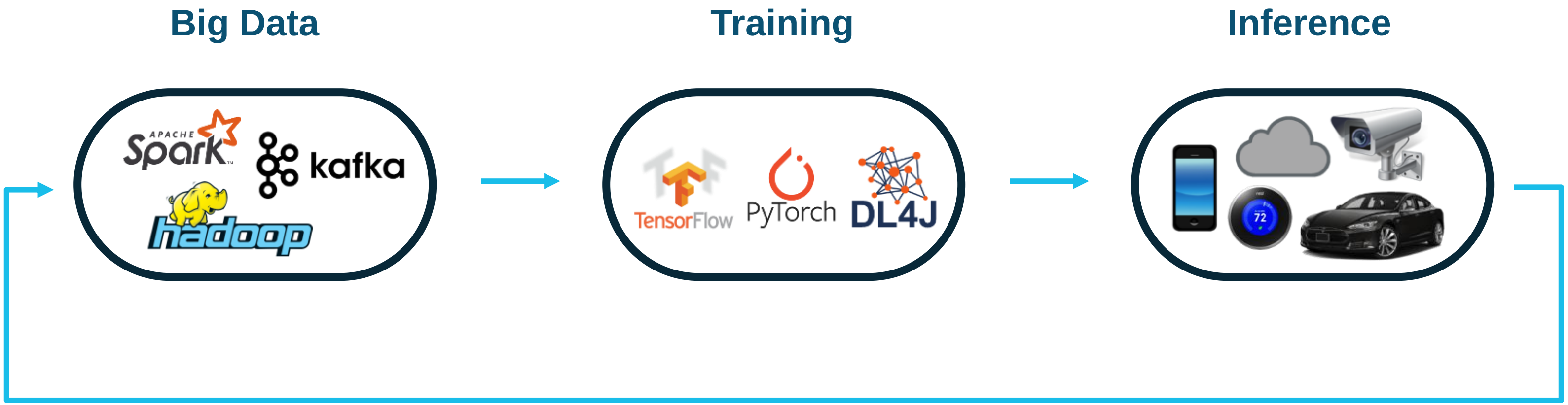
Deploy

Input
Input
Input
Input

Result
Result
Result
Result

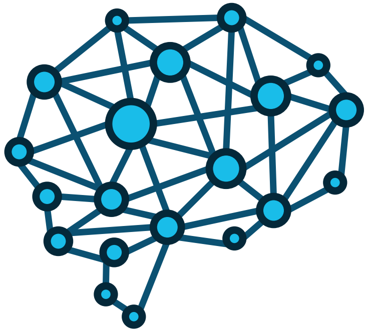
- Can Exceed Human Level Accuracy
- Automatic Feature Extraction
- Requires Massive Amounts of Data and Compute Power
- Good for Big Data and IoT Data
- Learning Like a Human and Executing at Computer Speed

Custom AI Deep Learning Workflow



Deep Learning is Now Widely Practical

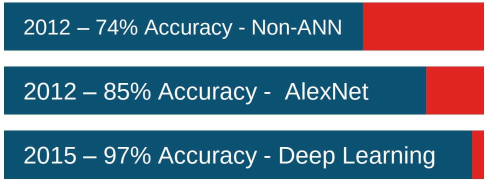
Lots of Data



Use Cases

Powerful Hardware

ImageNet Challenge



2012 – 74% Accuracy - Non-ANN

2012 – 85% Accuracy - AlexNet

2015 – 97% Accuracy - Deep Learning

Libraries and Tools

TensorFlow

PyTorch

DL4J

Kubeflow

cloudera HORTONWORKS

sky mind

CISCO

Business Drives AI

Business Question	Expert AI Task	Finance	Healthcare	Media and Entertainment	Security and Defense	Retail	Manufacturing
Is "it" present or not?	Detection	Identify Access Anomalies	Indications of Anomalous Care	Content Based Search	Identify Security Breaches	Identify Events in Store Surveillance	Detect Manufacturing Flaws
What type of thing is "it"?	Classification	Fraud detection	Medical Imagery Diagnostics	Content Labeling	Facial Recognition	Identify Returning vs New Shoppers	Enable Robots to Track Objects
To what extent is "it" present?	Segmentation	Sentiment Analysis	Condition Analysis	Improved Product Placement	Crowd Analytics	Segment by Customers Actions	Sort Components by Quality
What is the interpretation?	Natural Language Processing	Chatbot Advisors	Expert Diagnosis from Notes	Video Captioning	Real Time Language Translation	In Store Personal Assistants	Assembly Build Instruction Translation
What is the likely outcome?	Prediction	Credit Profiling	Length of Stay Forecasting	Targeted Content Generation	Equipment Health Assessment	Customer Churn and Retention	Proactive Machine Maintenance
What will satisfy the objective?	Recommendations	Algorithmic Trading	Treatment Recommendations	Effective Content Recommendations	Risk Management	"Magic Mirror"	Assembly Process Improvements

CISCO

Key IT AI Challenges

The Data Center Follows the Data

Distributed Data Sources and Technologies Risk Operational Silos and Complexity

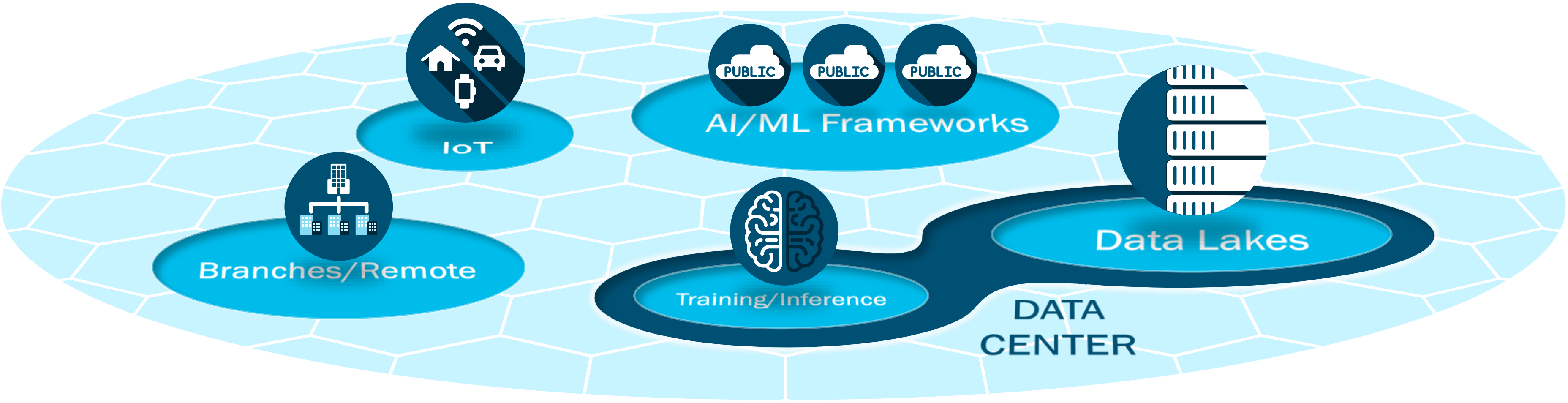
Uncharted Territory

Rapidly Evolving AI/ML Ecosystem and Requirements; Skill Shortages in Data Science and IT

Massive & Active Data Sets

Volume, Velocity, and Variability of AI Workloads at Scale Demand New Data Center Architectures

Cisco AI/ML/DL: A Holistic Approach



Accelerated Computing for Inference at the Edge

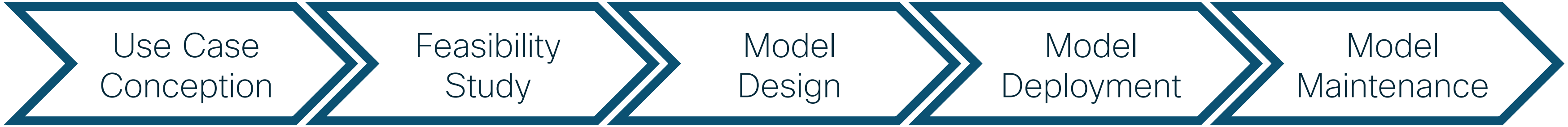
AI/ML Stack Partnerships

Accelerated Computing for AI/ML in the Core

Big Data/Analytics and ERP

Cisco HyperFlex | Cisco Multicloud Portfolio | CISCO INTERSIGHT | Cisco UCS | Cisco ACI

Cisco Portfolio Alignment


Use Case Conception

Feasibility Study

Model Design

Model Deployment

Model Maintenance

cloudera HORTONWORKS hadoop APACHE Spark

C240 M5

Optimized Big Data Solutions for Data Collection and Preparation

NVIDIA V100

HX240c M5

Quick Installation of Hardware and Software for AI Exploration and Experimentation

NVIDIA V100 NVLink

C480 ML M5

Offload and Accelerate AI Training at Scale with Performance Optimized Systems

NVIDIA P4

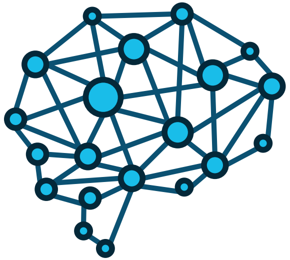
HX220c M5

Distribute and Scale AI Inference from the Data Center to the Edge

AI Hardware Components High Level

Artificial Neural Network

1,000s of Parallel Processing Elements Assembled to Identify Complex Patterns in High Variety Data with Superhuman Accuracy

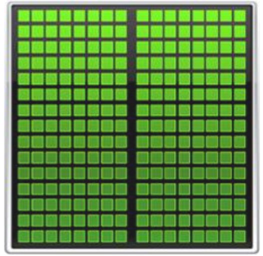


Xeon CPU

10s of Sequential Serial Processing Cores Ideal for Data Preparation and Management

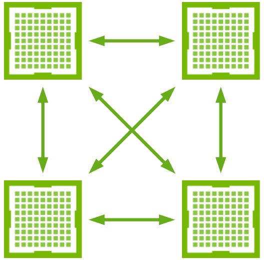
Tesla GPU

1,000s of Parallel Processing Cores Ideal for Deep Learning Mathematical Functions



NVLink

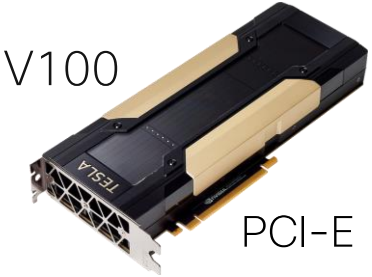
GPU Interconnect for Maximum Scalability and up to 10x the Bandwidth of PCIe to Support the Training of Modern Neural Networks



Tesla GPUs

Training

V100



PCI-E

Interconnect Bandwidth

32 GB/s

Deep Learning Training

112 TFLOPS

Memory

32 GB
900 GB/s

V100

NVLINK

Interconnect Bandwidth

300 GB/s

Deep Learning Training

120 TFLOPS

Memory

32 GB
900 GB/s

Inference

P4/P6

PCI-E

Interconnect Bandwidth

32 GB/s

Deep Learning Inference

60x per Watt vs E5v4

Memory

8 GB
192 GB/s

NVIDIA®

Cisco UCS C480 ML Rack Server

No-compromise balance of performance and capacity to power AI workloads at scale

NEW

Fully Integrated Platform Designed to Accelerate Deep Learning

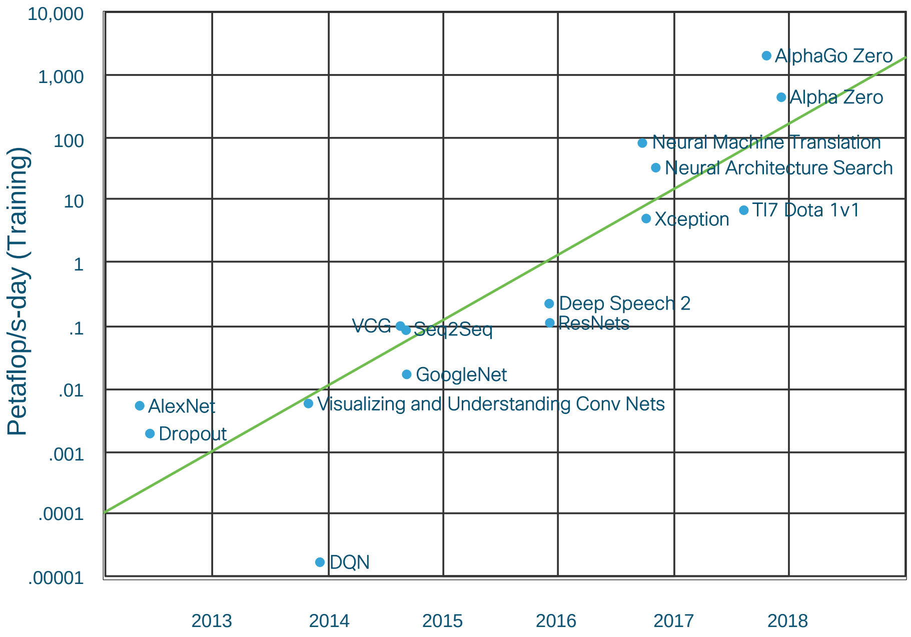
- Eight NVIDIA Tesla V100s with NVIDIA NVLink Interconnect
- Up to 24 Drives; 182TB
- Up to 6 NVMe Drives
- Network: Up to 4x100GB
- High Availability Design

Validated with Popular Machine Learning Software to Accelerate and Simplify AI/ML Projects on Premise

Prevents Operation Silos: Extends Existing UCS Environments with Consistent, Cloud-Based Management

Why GPUs and NVLink Matter

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



C480 ML M5

8 x Tesla V100

1 Petaflop

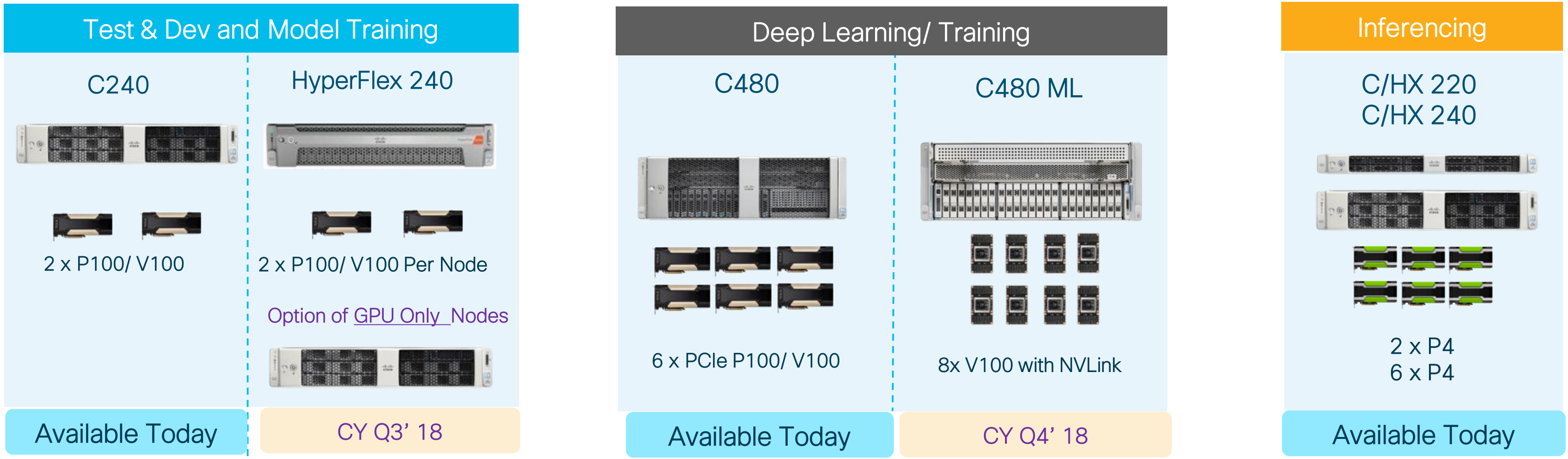
No HPC Experience Required

or

300 x Dual Socket Xeon Platinum 8180 Servers

Source: https://www.top500.org/news/intel-forges-new-xeon-line-under-scalable-processor-banner/
https://blog.openai.com/ai-and-compute/

UCS AI/ML/DL Compute Portfolio



Test & Dev and Model Training

C240

2 x P100/ V100

Available Today

HyperFlex 240

2 x P100/ V100 Per Node

Option of GPU Only Nodes

CY Q3' 18

Deep Learning/ Training

C480

6 x PCIe P100/ V100

Available Today

C480 ML

8x V100 with NVLink

CY Q4' 18

Inferencing

C/HX 220
C/HX 240

2 x P4
6 x P4

Available Today

Unified Management

CISCO INTERSIGHT

cisco UCS Manager

Cisco IMC

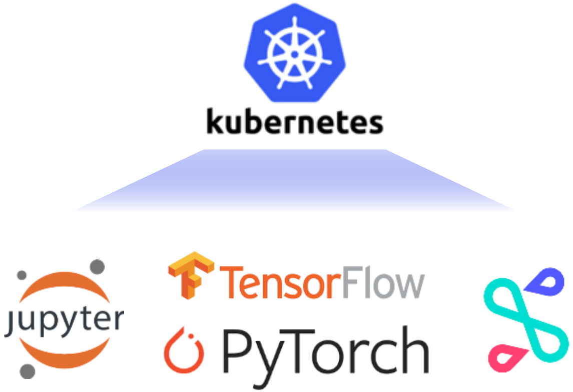
XML API

python SDK

Simplified Management, Customer Choice, Cisco Validated Design

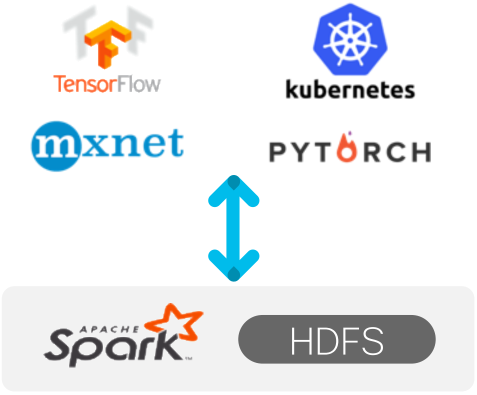
Data Centric Approach: Expanding to AI/ML/DL

Cisco Validated Designs



Kubeflow

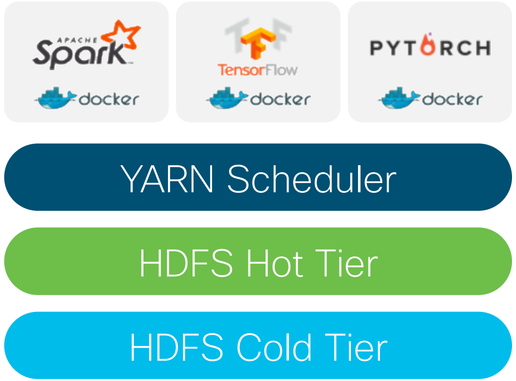
Portable, Scalable ML Stack
Enabling Rapid Development
and Deployment



HDFS

Cloudera Data Science Work Bench

Hadoop Coupled with GPU
Nodes for Deep Learning with
Jupyter Notebook



YARN Scheduler

HDFS Hot Tier

HDFS Cold Tier

Hortonworks Hadoop 3.1 Data Lake

Integrate Hadoop and AI/ML:
YARN Scheduling CPU and GPU
with Docker Application Support

Analytical Solutions with GPU Acceleration

sas THE POWER TO KNOW.

NEW

Cisco UCS Deep Learning Solution for SAS Viya with Hadoop

Cisco UCS 6332-16UP Fabric Interconnects

SAS Visual Analytics
SAS Visual Statistics

Cisco UCS C240 M5 Rack Servers

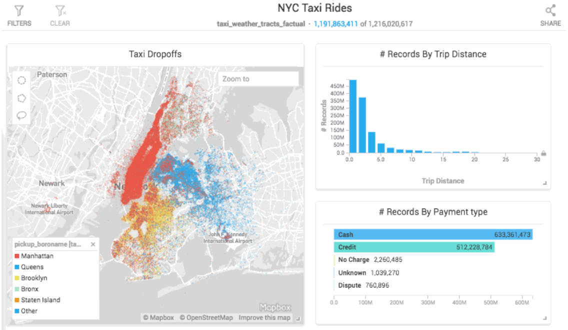
Apache Hadoop and SAS Viya

Cisco UCS C480 M5 with 6x NVIDIA GPUs

Cisco UCS Solution for SAS Viya with Hadoop

MAPD

NEW



NYC Taxi Rides

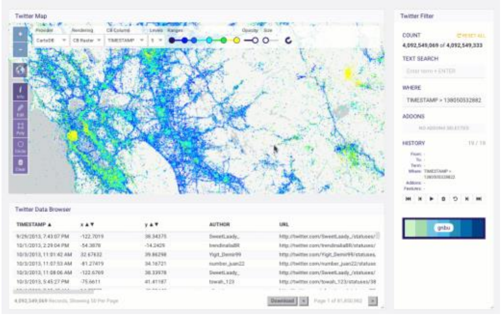
taxi_weather_tracts_factual · 1,191,863,411 of 1,216,020,617

# Records By Payment type	
Cash	633,361,473
Credit	512,228,784
No Charge	2,260,485
Unknown	1,039,270
Dispute	760,896

Interactive data analytics at scale

cisco MAPD NVIDIA.

kinetica



Twitter Data Browser

TIMESTAMP ▲	x ▲▼	y ▲▼	AUTHOR	URL
9/29/2013, 7:43:07 PM	-122.7019	38.34375	SweetLaady_	http://twitter.com/SweetLaady_/statuses/
10/1/2013, 2:29:04 PM	-54.3878	-14.2429	trendinaliaBR	http://twitter.com/trendinaliaBR/statuses/
10/3/2013, 11:01:42 AM	32.67632	39.86298	Yigit_Demir99	http://twitter.com/Yigit_Demir99/statuses,
10/3/2013, 11:07:53 AM	-81.27419	34.16721	number_juan22	http://twitter.com/number_juan22/statuse
10/3/2013, 11:08:06 AM	-122.6769	38.33978	SweetLaady_	http://twitter.com/SweetLaady_/statuses/
10/3/2013, 5:45:27 PM	-75.6611	41.41187	towah_123	http://twitter.com/towah_123/statuses/38

4,092,549,069 Records, Showing 50 Per Page Download Page 1 of 81,850,982

Cisco UCS, NVIDIA, and Kinetica

GPU Accelerated Databases

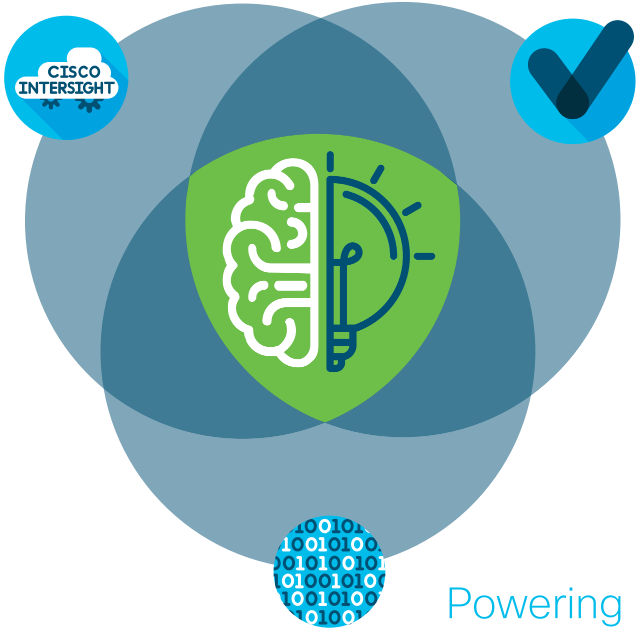
UCS: One System for All Workloads

6300 Series Fabric Interconnects	
Scale Out and Compute Intensive Applications C4200 and C125	Cloud Computing, Microprocessor Design and Simulation, Computational Fluid Dynamics (CFD), High Frequency Trading (HFT,) Fraud Detection, Online Gaming
Mainstream Workloads B200 M5 C220 M5 and C240 M5 HyperFlex	VDI, VSI, Distributed Databases, CI, Enterprise Applications: Oracle and SAP HANA, Big Data and Analytics, AI/ML with GPUs
NEW AI/ML and Scale up Workloads C480 ML M5 and B480 M5	AI/ML With Dense GPUs, and Memory-Intensive Mission-Critical Enterprise Applications: In-Memory Databases
Data Intensive Workloads S3260 M5	Data Protection, SDS, Scale-Out Unstructured Data Repositories, Media Streaming, and Content Distribution

Why Cisco Computing Solutions for AI

Eliminating Operational Silos

Full array of accelerated computing options for test/dev, training and inference, all unified by cloud-based management



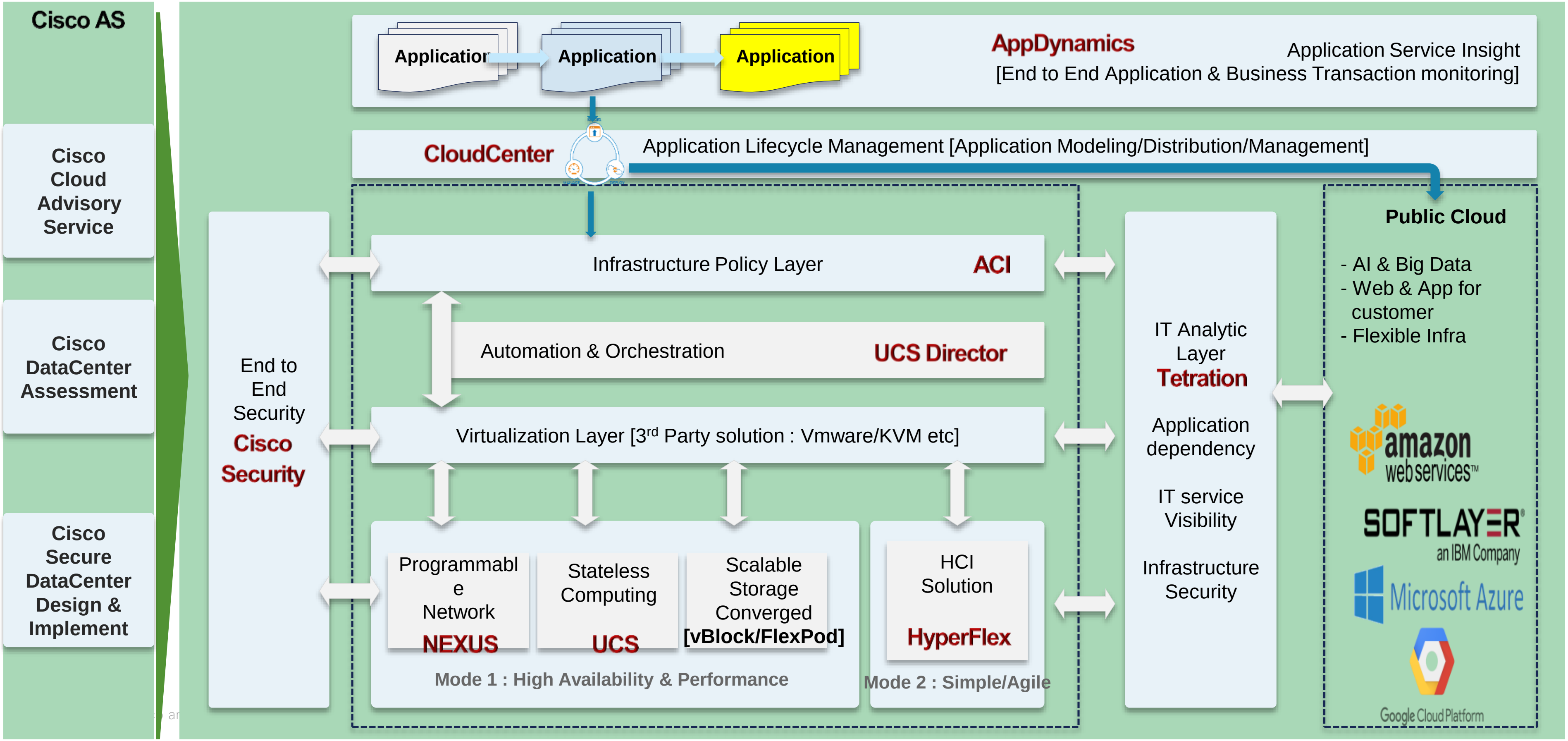
Demystifying AI/ML/DL Stacks

Curating top-to-bottom SW and HW stacks with leading ecosystem partners to ensure a faster and more predictable deployment

Powering the Full AI Data Lifecycle

Integrating changing data sources as part of a dynamic data pipeline

Cisco multi Cloud frame work for Data Center Transformation



Cisco AS

Cisco Cloud Advisory Service

Cisco DataCenter Assessment

Cisco Secure DataCenter Design & Implement

Application

Application

Application

AppDynamics

Application Service Insight
[End to End Application & Business Transaction monitoring]

CloudCenter

Application Lifecycle Management [Application Modeling/Distribution/Management]

End to End Security
Cisco Security

Infrastructure Policy Layer ACI

Automation & Orchestration UCS Director

Virtualization Layer [3rd Party solution : Vmware/KVM etc]

Programmable Network
NEXUS

Stateless Computing
UCS

Scalable Storage Converged
[vBlock/FlexPod]

Mode 1 : High Availability & Performance

HCI Solution
HyperFlex

Mode 2 : Simple/Agile

IT Analytic Layer
Tetration

Application dependency

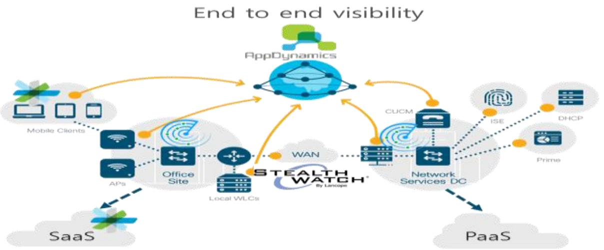
IT service Visibility

Infrastructure Security

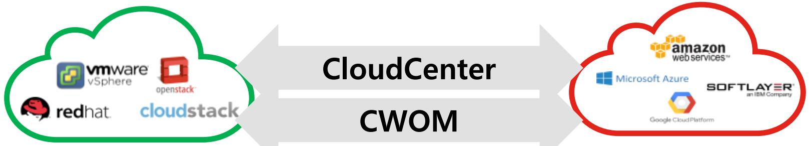
Public Cloud

- AI & Big Data
- Web & App for customer
- Flexible Infra

Data Center Transformation to Multi cloud



3 Insightful operation with Analytic and secured Data Center



CloudCenter

CWOM

Private Cloud

Public Cloud

2 Build Agile service platform with Multi-Cloud

SDN with **ACI**

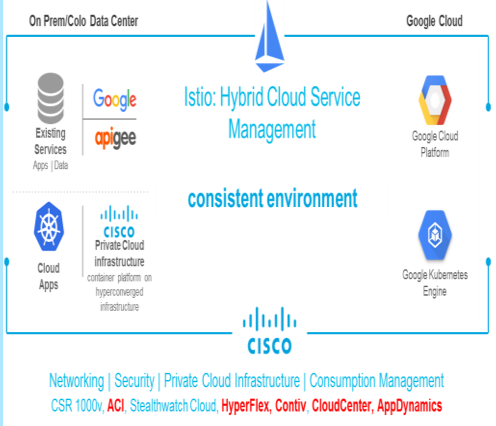
SDC with **UCS**

SDS with **HyperFlex**

SAN with **MDS**

1 Modernize Infrastructure with Software defined X

4 Application transformation Into Container & Cloud Native



Networking | Security | Private Cloud Infrastructure | Consumption Management

CSR 1000v, **ACI**, Stealthwatch Cloud, **HyperFlex**, **Contiv**, **CloudCenter**, **AppDynamics**

THANK YOU

CISCO