Ubiquitous Machine Learning

Han Yang, PhD, @hanyang1234
Senior Product Manager
September, 2018
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%.

References:
- $1.2T: https://go.forrester.com/wp-content/uploads/Forrester_Predictions_2017_Artificial_Intelligence_Will_Drive_The_Insights_Revolution.pdf
The simulation finds that front-runners could increase economic value (economic output minus AI-related investments and transition costs) by 122% by 2030, or an implicit additional growth rate of cash generation of about 6 percent a year for the next 12 years.

Achieving and sustaining this rate of growth over such a long period would be remarkable, as it conflicts with Gibrat’s Law that a firm’s growth rate is independent of its size.

The analysis suggests that cash generation is likely to accelerate after a five-year period in which front-runners could experience cash outflows as they invest in, and scale up, AI. Front-runners tend to slowly concentrate the profit pool of their industry in a winner-takes-all phenomenon. This may lead to the phenomenon of increasing concentration and the rise of “superstar” firms. Some researchers argue that technology could enable superstars to pull away from the pack, but also that a slowdown in the diffusion of technology could

**Relative Change in Cash Flow by AI Adoption Cohort**

**SOURCE:** McKinsey Global Institute analysis

**NOTE:** Numbers are simulated figures to provide directional perspectives rather than forecasts.
Data is Exploding Everywhere

- 1.3 Zettabytes of Data in Data Centers
- 5.9 Zettabytes of Data Stored in Devices
- 847 Zettabytes of Data Created Everywhere

By 2021
Big Data Forecast

Big Data Volume Grows Nearly 8-Fold

Big Data Will Represent 30% of All Data Stored in Data Center by 2021

Source: Cisco Global Cloud Index, 2016–2021

Data Created vs. Data Center Traffic
Data Created Outpaced

Source: Cisco Global Cloud Index, 2016-2021

US Download Speeds Distribution—2017

Download (DL) Speeds in Mbps

No. of Speed Tests

Millions

Percentiles

8 Min to Load 2GB Model

Text and IM
Augmented Reality Gaming
Virtual Reality Streaming
Three Concurrent Sample Apps

S. Africa Download Speeds Distribution–2017

- 10th to 90th Percentiles
- Median 4.6 Mbps
- Mean 7.7 Mbps
- 58 Min to Load 2GB Model

Download (DL) Speeds in Mbps

No. of Speed Tests

Thousands

Text and IM
Augmented Reality Gaming
Virtual Reality Streaming
Three Concurrent Sample Apps

Data is Exploding Everywhere

Analytics Need to Follow Data

By 2021

1.3 Zettabytes of Data in Data Centers
5.9 Zettabytes of Data Stored in Devices
847 Zettabytes of Data Created Everywhere
Analytics at the Edge
Edge Analytics in Industrial IoT

40% of large enterprises will be integrating edge computing into their 2021 projects.

30% of large enterprises will use edge locations by 2021 driven by interactive UI.

30% of workload deployment has latency & BW requirements.

* Source Gartner

Oil and Gas
- Many drilling platforms
- Low BW uplink (satellite)
- Need tiered data evaluation

Roadway
- Thousands of road signs
- Widely distributed
- Need local data evaluation

Factories
- Legacy systems and protocols
- High data volume collected from machines
- Security and data privacy
- OT data analysis & policy at factory level
- Need local data normalization, storage, & processing
Edge Analytic Use Cases

There may not be enough network bandwidth

Most of the data is not interesting

The use of data may be at the edge

Computation can be optimized for some purposes

Data normalization

Data redirection based on the content of the data

Data time stamping for later forensic analytics

Data Reduction

Filtering

Latency Optimization

Partitioning

Application Simplification

Dynamic Changes

Analytic Support
IoT World Forum Reference Model

Levels

7. Collaboration & Processes
   (Involving People & Business Processes)

6. Application
   (Reporting, Analytics, Control)

5. Data Abstraction
   (Aggregation & Access)

4. Data Accumulation
   (Storage)

3. Edge Computing
   (Data Element Analysis & Transformation)

2. Connectivity
   (Communication & Processing Units)

1. Physical Devices & Controllers
   (The “Things” in IoT)

Credit to: Jim Green, CTO Cisco Kinetic & Chairman IoT World Forum
Cisco Kinetic – the IoT data management platform

Move data programatically to the desired destination with the user defined policy

Analyze data Transform data, apply rules, perform distributed micro-processing

Connect devices and extract data
How the different Kinetic components interact

**Sensors & Endpoints**
- Edge Devices (IE4k, GW, Compute)

**Fog Compute**
- Edge Devices
- Applications
- Edge Fog Module

**Enterprise Data Centers**
- Edge Devices
- Applications

**Applications**
- On-prem or private cloud apps
- Cloud Apps
  - Google Cloud Platform
  - IBM Watson
  - Amazon Web Services

**Edge Fog Manager**
1. EFM extracts data from sensors, transforms it and sends it either to another IoT device, or to a hosted app
2. Fog & DC EFM adds additional compute capabilities, including normalization, aggregation, storage and visualization
3. DCM adds secure cloud connection capabilities
4. GMM enables mass deployment and management of gateways

Note: these components are independent and rarely all used in one deployment
Secure connection

- Secure connection of IoT Devices
- Simplified GW management at scale
- Secure IoT app lifecycle management
How does it work

1. Gateway secure boot
2. Secure onboarding
3. Secure gateway Provisioning
4. Fog App (Cisco or 3rd party) download & life cycle management
5. Secure connectivity of sensors
6. Data gathering, processing, secure delivery to apps
**Investment with Multiple Options**

**Field Area Network (Wi-SUN)**
- AMI smart metering
- Distribution automation
- Street lighting
- O&G wellhead monitoring
- Water/wastewater

**Fleet Vehicles Mass Transit**
- Automated Vehicle
- Location tracking, Data Uploaded in Seconds with 4G/LTE
- Handles Multiple Wireless Laptops, Smartphones, Tablets Simultaneously

**Remote Asset Monitoring**
- Pipeline monitoring
- Roadside infrastructure
- Distribution automation
- ATMs
- Digital Signage

**Premium Mobile Broadband (PMB)**
- Public safety and security CPE

**Low Power Long Range Wireless (LPWA – LoRA)**
- SP IoT Infrastructures
- Battery powered sensors
- Environmental monitoring
- Smart Cities, parking, and Agriculture
- SP cell tower monitoring

**IOT Field Network Director/Industrial Operations Kit**
- Zero Touch Provisioning, Firmware upgrade,…

**IOT Software Platform**
- Fog Computing, BYOI,…

**CGR1000**
**IR500**
**IR829**
**819H**
**IR809**
**IR829**
**IR809**
**IR910**
**IR8x9 + LoRA Modem (future)**
Extract and compute the data

- Extract data from diverse, distributed devices
- Tiered microservices from edge to DC, enabled by flexible, distributed service architecture
- High performance data historian
- Support real-time data visualization
- Intuitive data policy editor
Machine Learning the Data Center
Key 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 Computing Solutions for Machine Learning

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

Demystifying AI/ML 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 AI/ML Computing Platforms, Partners, and Solutions

- **Accelerated Computing**
  - Inference in Regional and Micro Data Centers
  - Test/Dev and ML in Private Cloud
  - Deep Learning in DC Core
  - UCS C220 Servers
  - UCS C240 Servers
  - HyperFlex GPU Nodes
  - UCS C480 ML

- **ISV Partners**
  - Google Cloud Platform
  - Nvidia
  - Anaconda
  - KubeFlow
  - Cloudera

- **Solution Partners**
  - ePlus
  - Presidio
  - Insight
  - Trace3
Cisco UCS C480 ML Rack Server
No-compromise balance of performance and capacity to power AI workloads at scale

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
Cisco UCS C480 ML M5 Rack Server for Deep Learning
A No-Compromise Purpose Built Server for Deep Learning

**CPU**
- 2 * Intel® Xeon® Processor Scalable Family (Up to 28 cores per socket)
- 24 DDR4 DIMMs—up to 3 TB Memory

**Storage**
- Up to 24 SAS/SATA SSD/HDD
- Up to 6 NVMe Drives

**GPUs**
- 8 X V100 32GB: 1st GPU to break 100 teraflops
- NVLink Interconnect: >300GB/s bandwidth

**Network**
- Choice of 10/25 or 40/100G
- Four PCIe Slots

**Raid Controller**

**Redundant Fans**

© 2017 Cisco and/or its affiliates. All rights reserved. Cisco Confidential
Focus on Full Data Life Cycle, Simplicity, and Manageability

Data
- Source & Aggregate
- Cleanup
- Transformation

Full portfolio for all AI/ML computing needs

Validated solutions with technology partners

Natural extension of existing computing environment

Deployment/Inference
- Application runtime at source of demand

Training
- Model Dev
- Model Validation
- Model Execution
Data Centric Approach: Expanding Big Data to AI/ML Solutions
Cisco Validated Designs

Cloudera Data Science Work Bench
Hadoop Coupled with GPU Nodes for Deep Learning with Jupyter Notebook

Kubeflow
Portable, Scalable ML Stack Enabling Rapid Development and Deployment

Hortonworks Hadoop 3.1 Data Lake
Integrate Hadoop and AI/ML: YARN Scheduling CPU and GPU with Docker Application Support

YARN Scheduler
HDFS Hot Tier
HDFS Cold Tier
UCS: One System for All Workloads

- **6300 Series Fabric Interconnects**: Cloud Computing, Microprocessor Design and Simulation, Computational Fluid Dynamics (CFD), High Frequency Trading (HFT), Fraud Detection, Online Gaming

- **Scale Out and Compute Intensive Applications C4200 and C125**: VDI, VSI, Distributed Databases, CI, Enterprise Applications: Oracle and SAP HANA, Big Data and Analytics, AI/ML with GPUs

- **Mainstream Workloads B200 M5 C220 M5 and C240 M5 HyperFlex**: AI/ML With Dense GPUs, and Memory-Intensive Mission-Critical Enterprise Applications: In-Memory Databases

- **AI/ML and Scale up Workloads C480 ML M5 and B480 M5**: Data Protection, SDS, Scale-Out Unstructured Data Repositories, Media Streaming, and Content Distribution

- **Data Intensive Workloads S3260 M5**
# Cloud-Powered Systems Management

<table>
<thead>
<tr>
<th>Centralized Management</th>
<th>Intuitive Experience</th>
<th>Enhanced Support</th>
<th>Proactive Guidance</th>
<th>Secure and Extensible</th>
<th>SaaS Delivered</th>
<th>Comprehensive Automation</th>
<th>Single Pane of Glass</th>
</tr>
</thead>
</table>

- **SaaS Simplicity**
- **Actionable Intelligence**

---

© 2017 Cisco and/or its affiliates. All rights reserved. Cisco Confidential
Customer-Centric Marketing
Banking Customer with Cisco UCS for On-Premise AI Farm

- Product recommendation
- Experience personalization
- Attrition prediction

Existing UCS Big Data Cluster

AI Farm: UCS C240 with V100 GPUs
Cisco: Improving Customer Experience

- Integrating data lake with ML
- Predicting Failures
- Automatically Position Spares at Depots
- Optimizing Supply Chain and Customer Experience

Component Failure Prediction

Data Sources
- Data Lake
- AI Platform (for Develop / Train)
- Communicate/Action
- Machine Learning Libraries
- Data Exploration
- Dashboard
- Programming Languages
- Visualization
- Accelerated Compute

Data Sources
- Data
- UCS
- Training Farm
- UCS with GPUs
- Inferencing
- UCS
Activating Data with the Power of UCS
New Cisco Computing Solutions for AI/ML

Unified Architecture
Adaptable Cloud-Managed Systems for Distributed IT

Demystifying AI/ML Stacks
Validated Solutions with Industry Leaders

Powering the Full AI Data Lifecycle
Accelerating Insight and Action
Putting Everything Together
Data: power frequency, voltage, current, phasor angle, ...

Data Mgmt. & Pattern Detection using ESP

Visualization of streaming data using SAS ESP Streamviewer

Application Life Cycle Management (Cisco FogDirector)

Pattern Detection using ESP

Connect to external apps with ESP connectors

Visualization of streaming data using SAS ESP Streamviewer

Pattern discovery using SAS Visual Analytics and SAS Visual Statistics

Update Streaming models

Cisco UCS

KAFKA

Cisco UCS

Cisco UCS

Edge to Enterprise IoT Analytics
Edge to Enterprise IoT Analytics: Electric Utility

Data: power frequency, voltage, current, phasor angle, ...

Cisco UCS

Pattern Discovery using SAS Visual Analytics and SAS Visual Statistics

Visualization of streaming data using SAS ESP Streamviewer

Application Life Cycle Management (Cisco FogDirector)

Substation

Operations Center

Update Streaming models

Connect to external apps with ESP connectors

Cisco UCS

KAFKA
Local and Global Analytics

Local Analytics

Global Analytics

Frequency
Cisco Infrastructure

IR 8x9
- Ruggedized Integrated Services Router
- Runs SAS analytics as data is being gathered
- Securely send relevant summary back to data center

Integrated Infrastructure for Big Data
- Industry leading Cisco UCS
- Tested and validated configuration
- High performance & availability
Data Mgmt. & Pattern Detection using ESP

Data: power frequency, voltage, current, phasor angle, ...

Visualization of streaming data using SAS ESP Streamviewer

Cisco UCS

Pattern Detection using ESP

Pattern discovery using SAS Visual Analytics and SAS Visual Statistics

Visualization of streaming data using SAS ESP Streamviewer

Update Streaming models

Remote Data and Raised Abstraction

Raise Abstraction

Reduce Bandwidth

Part of Data Pipeline

Substation

Operations Center
Discovery and Edge Analytics

Data: power frequency, voltage, current, phasor angle, ...

Discovery -> Updated Edge Analytics Model
Summary

• By 2020: Insight driven will steal $1.2T / year from less-informed peers

• By 2021, there will be 847 Zettabytes of Data
  • Only 1.3 ZB is in the data center

• Analytics needs to follow data to the edge: Ubiquitous Machine Learning

• Cisco provides holistic analytic tools for edge, data center, and even hybrid cloud with data lifecycle approach
Cisco