Stanford EE392b

Unlock the Power of Managing Cloud Health with AIOps

Allison Jones

Senior Director Product Management,
Azure AIOps

Azure

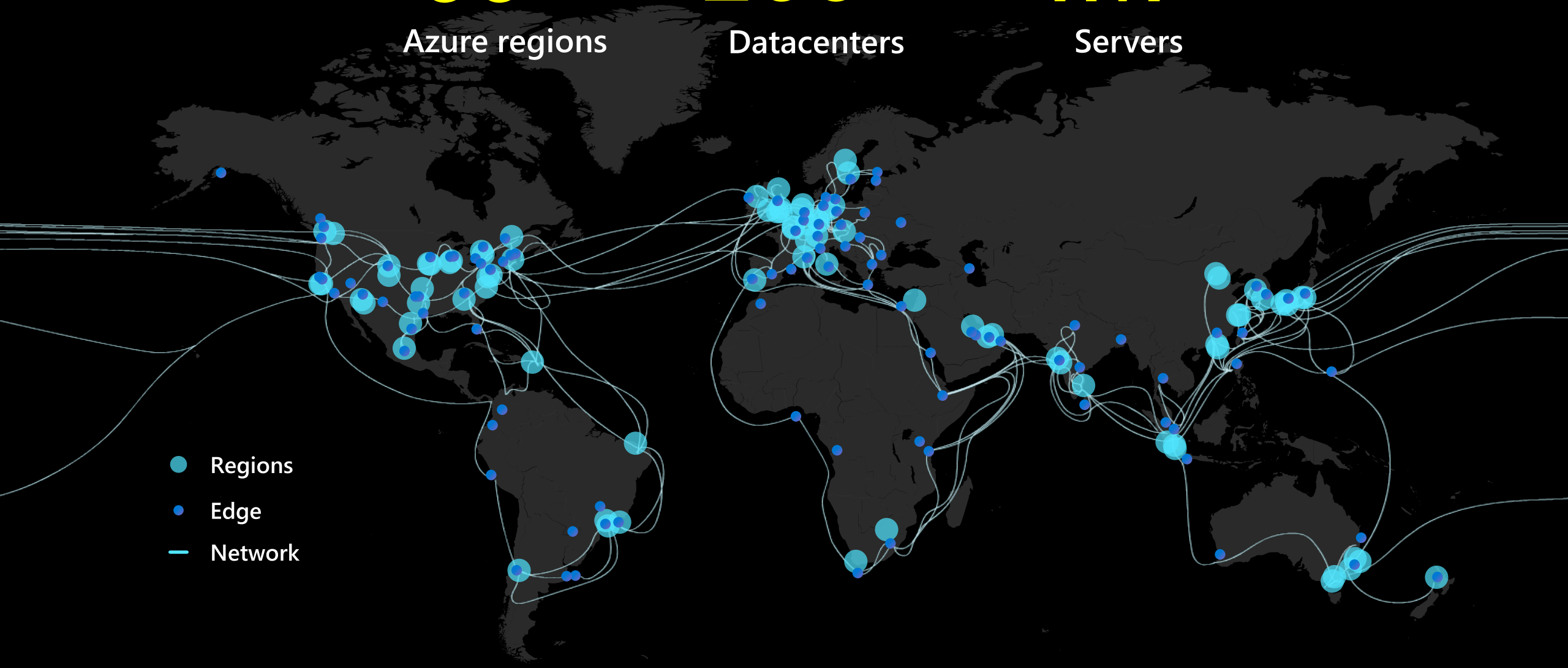
Azure is the world's computer

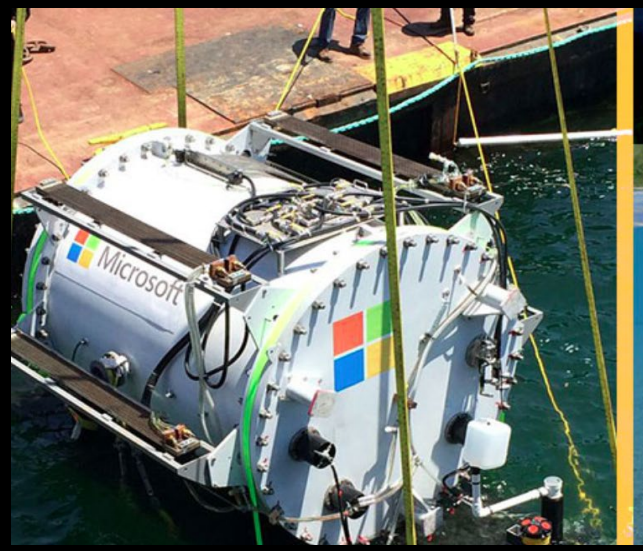
Azure

66+
Azure regions

200+
Datacenters

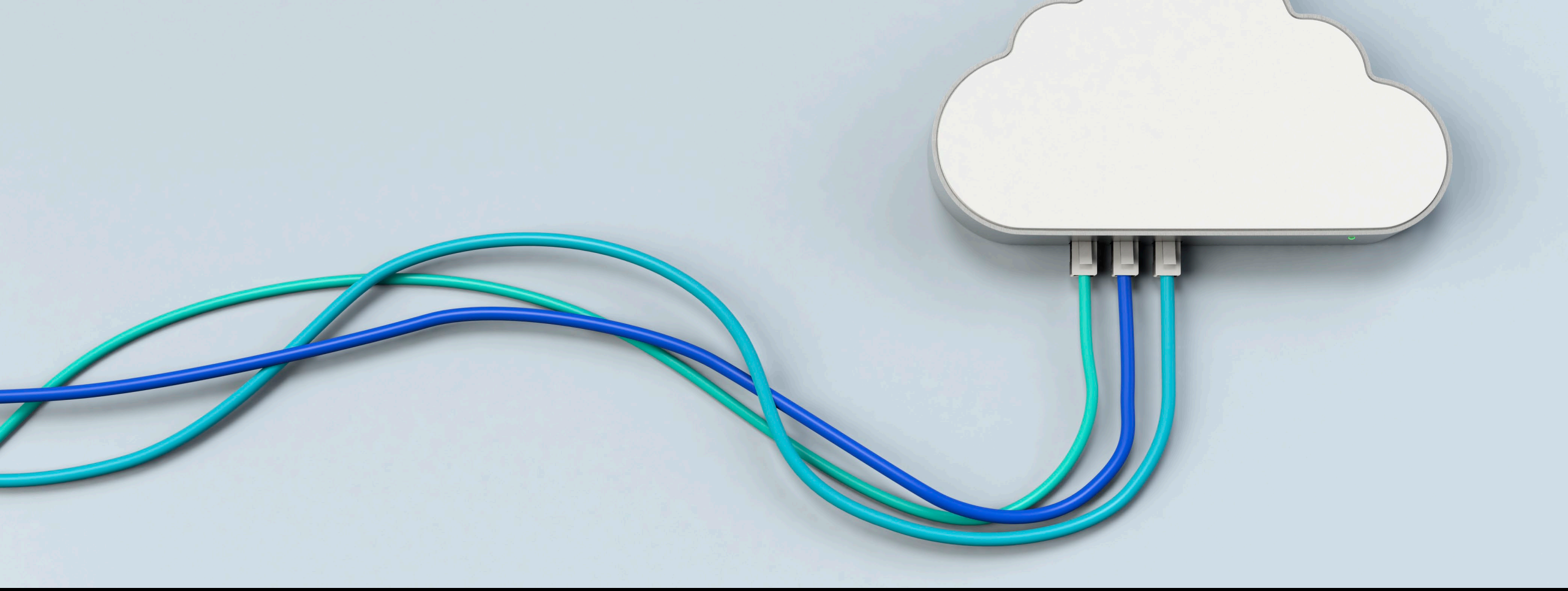
4M+
Servers




Microsoft


Microsoft
Microsoft

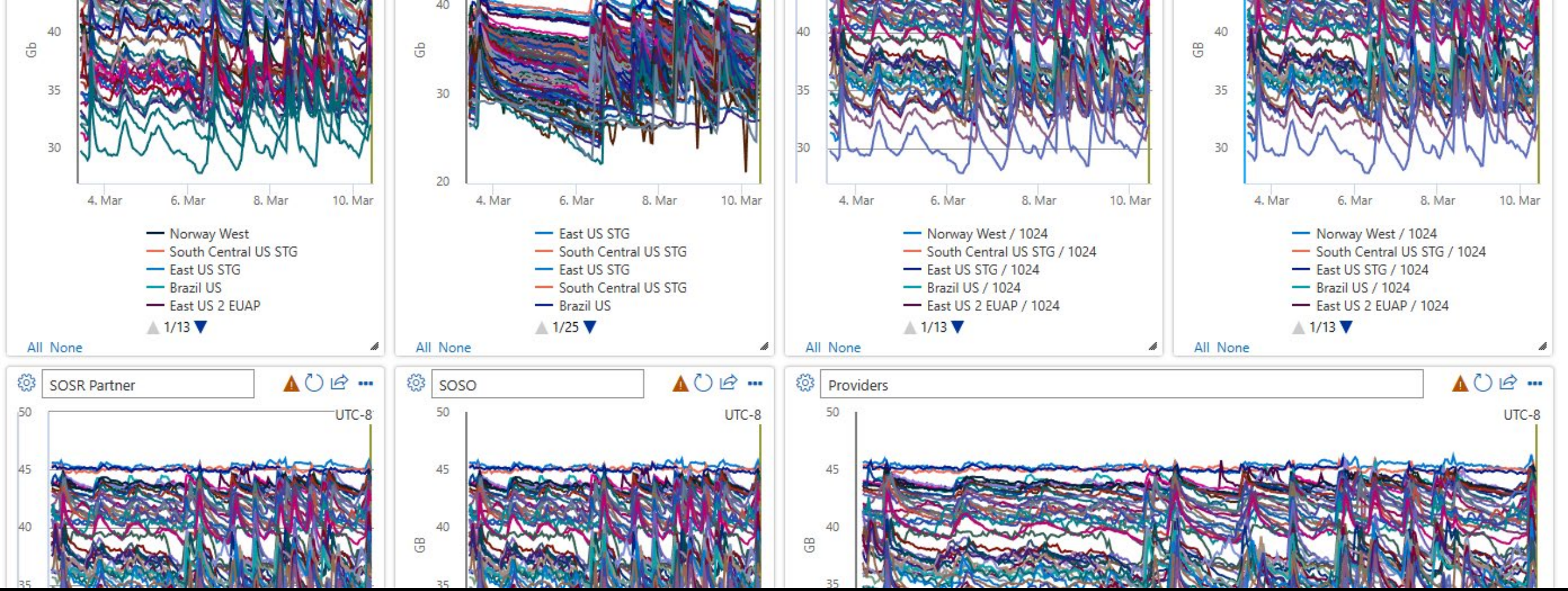
~~Challenges~~ Opportunities



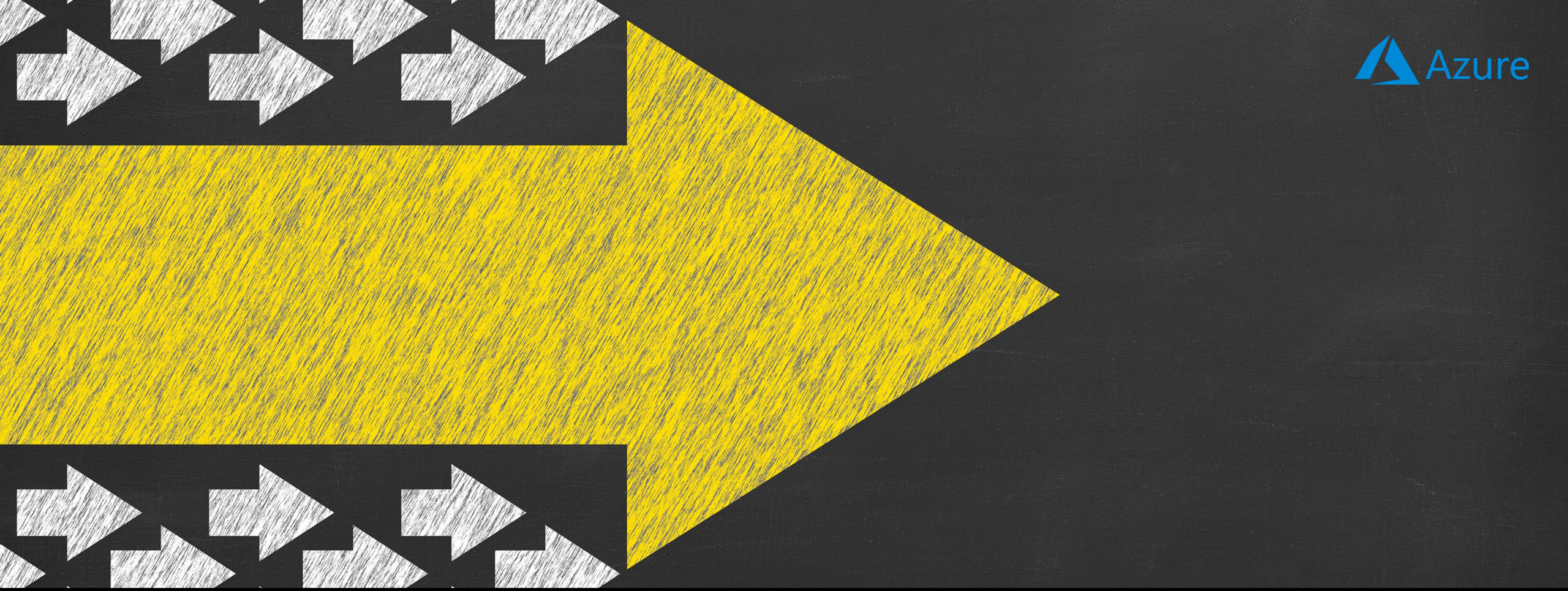
~~Lack of~~ Comprehensive standardized, accurate and reliable understanding of the health of Azure services

Changes to code, config, etc. cause
~~a majority of~~ few service outages

When incidents occur, communication is ~~often not~~ timely



Diagnosing issues is ~~hard~~ simple / automated,
requiring no DRI toil and manual touches



How do we get there

Why AIOps?

- Complexity of applications and infrastructure (thousands of services)
- Scalability- current solutions like time-series analysis and rule-based heuristics algorithms apply to localized problems
- Volume of Data- is overwhelming and difficult to process by humans
- ML based methods can achieve higher prediction accuracy by extracting patterns from data across different vantages

Development of Anomaly Detection & AIOps

Static Thresholds → Dynamic Thresholds → Learning Normal Behavior

A few data sources → Massive data sources → Any data source

Time Series performance data vs. Event Correlation and consolidation

Domain specific anomaly detection vs. Domain agnostic anomaly detection

Time Based correlation vs. Context based (relevance & priority)

Anomaly Detection vs. Self-Healing Autonomous Operations

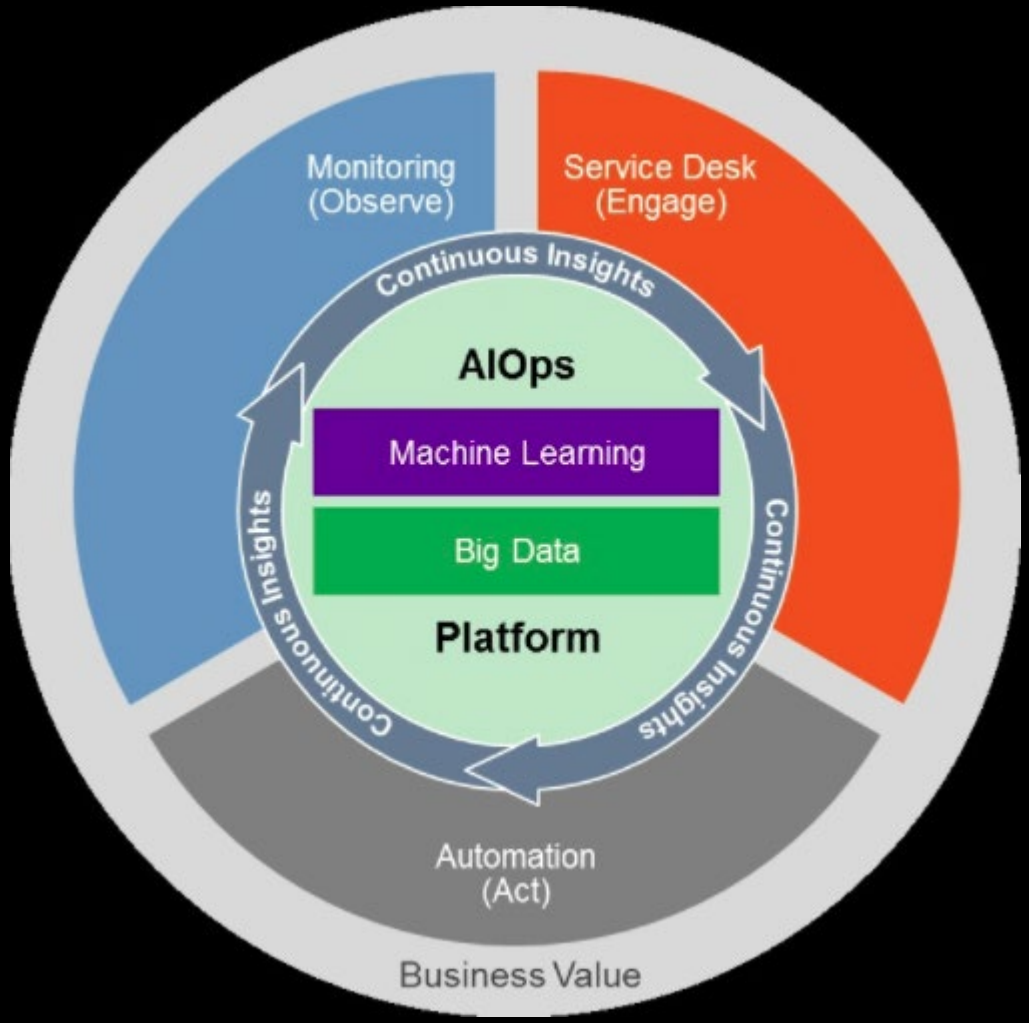
AIOps – Gartner's definition

Big data and ML driven IT operation automation process

Adoption has increased with the uptick of digital transformation

Business value

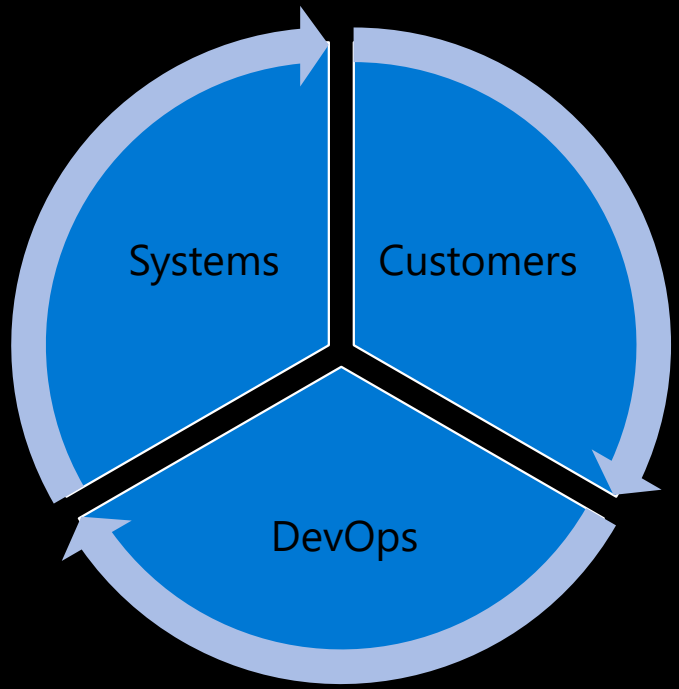
- Higher efficiency
- Higher Service quality
- Lower COGS



Source: Gartner

AIOps – Azure's definition

Innovating AI/ML technologies to effectively and efficiently **design**, **build**, and **operate** complex **cloud services** at **scale**



- **AI for Systems**

  Building high-quality services with better reliability, performance, and efficiency
- **AI for DevOps**

  Achieving high productivity in DevOps via empowering engineers with intelligent tooling
- **AI for Customers**

  Improving customer satisfaction with intelligence and better user experiences

Azure AIOps Vision Sharing

Advancing Azure service quality with artificial intelligence: AIOps

Posted on June 29, 2020

Mark Russinovich, Chief Technology Officer and Technical Fellow, Microsoft Azure

"In the era of big data, insights collected from cloud services running at the scale of Azure quickly exceed the attention span of humans. It's critical to identify the right steps to maintain the highest possible quality of service based on the large volume of data collected. In applying this to Azure, we envision infusing AI into our cloud platform and DevOps process, becoming AIOps, to enable the Azure platform to become more self-adaptive, resilient, and efficient. AIOps will also support our engineers to take the right actions more effectively and in a timely manner to continue improving service quality and delighting our customers and partners. This post continues our Advancing Reliability series highlighting initiatives underway to keep improving the reliability of the Azure platform. The post that follows was

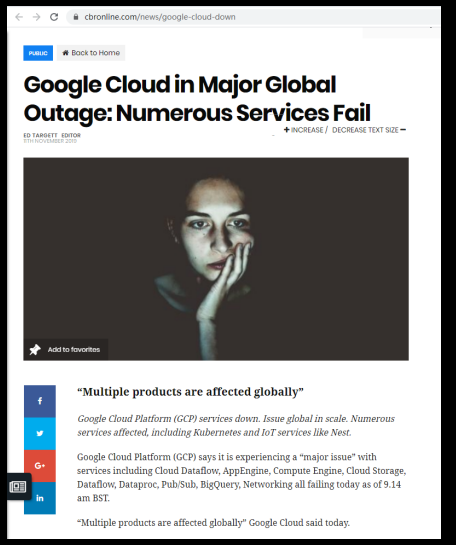
Cloud Reliability

State-of-the-art Cloud Reliability

- 5/6-9s' availability
- High degree of automation and intelligence
  - >95% auto failure detection within minutes
  - Comprehensive monitoring and diagnosis platforms/tools
  - >95% automated response

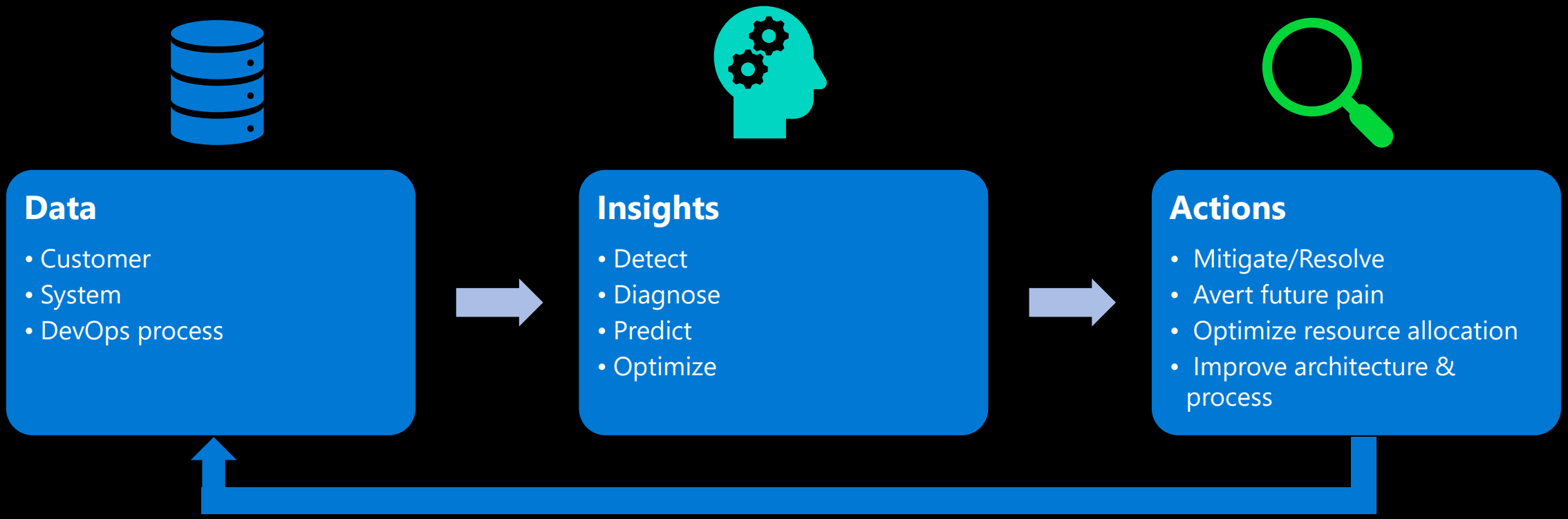
Endless pursuit of reliability & Effective Management

- Incident: interruption or performance degradation of a component*
- Outage: severe incidents with widespread impact
- Costs: $17K/Outage·min (2016)**
- Incidents/outages take a long time to mitigate
- Incident management is non-trivial with cloud scale

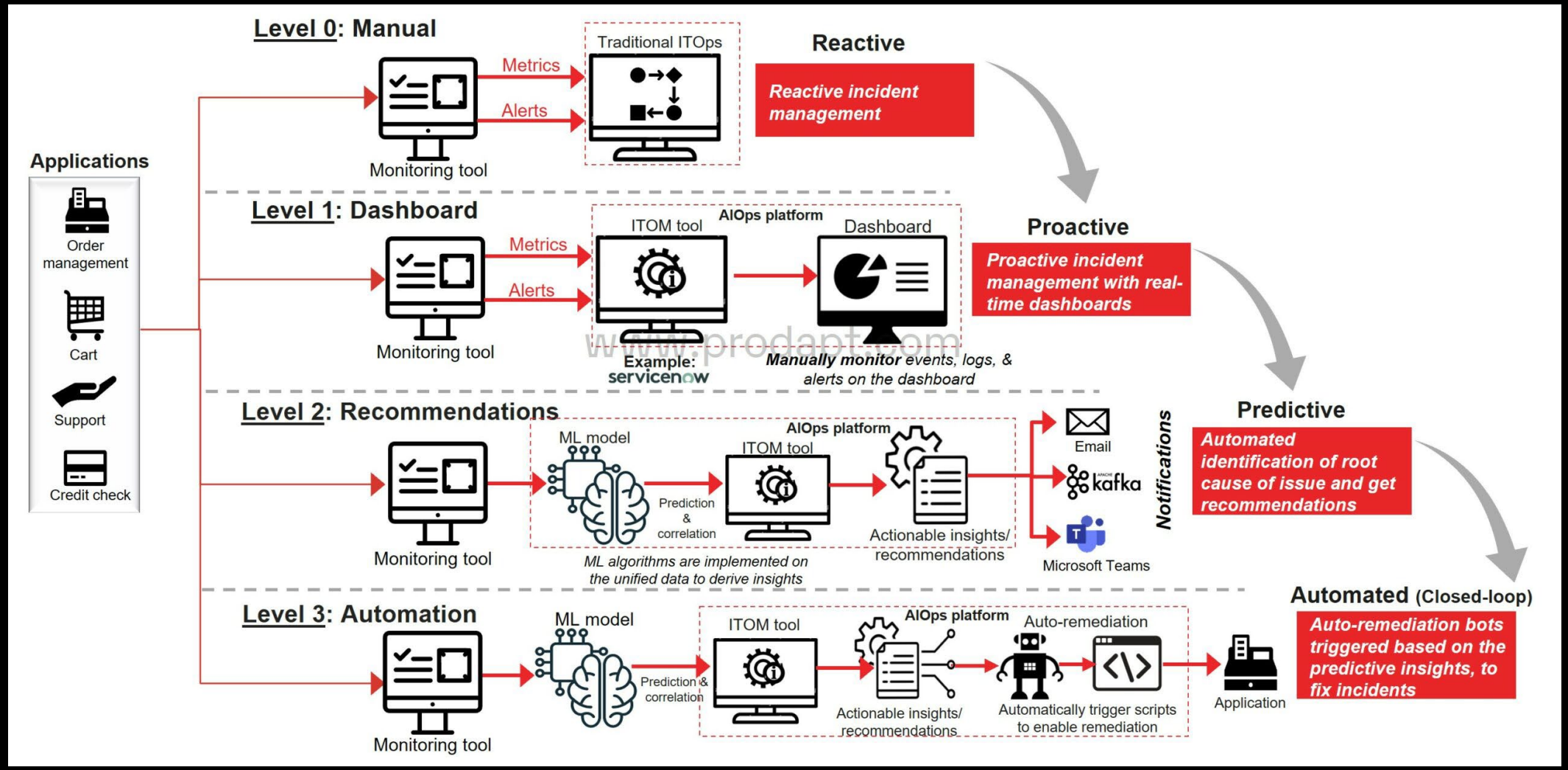


*service/product/device/resource/API…
**Ponemon Institute© Research Report

AIOps Methodologies: From Data to Actions



Levels of AIOps Maturity

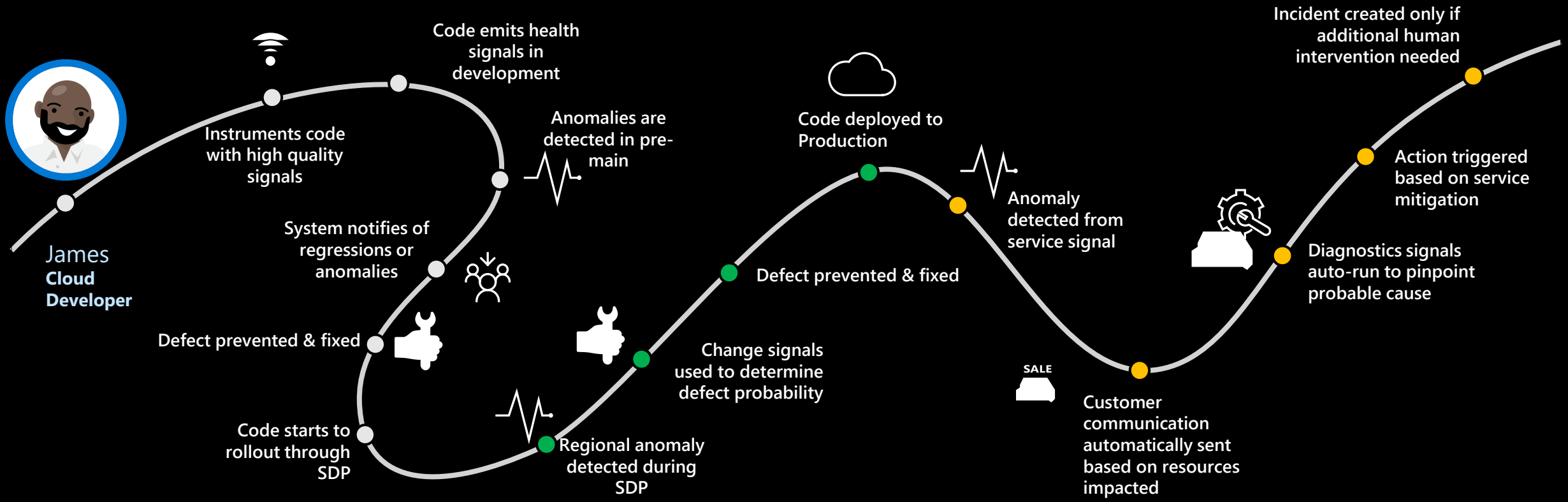


[Predicting, Preventing and Resolving Incidents With AIOps - DevOps.com]

Productizing AIOps for Azure

AIOps in Practice

An increasingly autonomous health system that detects, diagnoses and mitigates issues before they impact the customer.
Enable this by shifting left the service health systems of prediction, detection and diagnostics.



Opportunities

Pre-Release

- SLO/SLI and other signals in pre-main are used to identify regressions and alert the developer before code is moved to production
- Test/Validation signals can be used for identifying or predicting probability of defects

Deployment

- Change signals can be used to predict the probability of the code as potentially causing customer impact or impact to other services
- During SDP, anomalies are detected during rollout across regions and notify service owner of delta

Production

- High quality SLI signals or other signals to enable the system to detect an anomaly & understand its customer impact, send notifications to customers, systematically pinpoint probable cause using diagnostics signals, and initiate mitigation determined by service team

Azure BRAIN

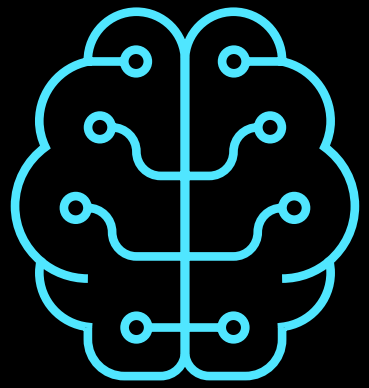
Azure

Customer Experience

Azure Services

Infrastructure devices

Critical Environment and Mechanical



BRAIN

Network of Intelligence

Automatic Alert correlation

Fast and actionable anomaly detection

Auto-communication

Automatic impacted service identification

Impact assessment

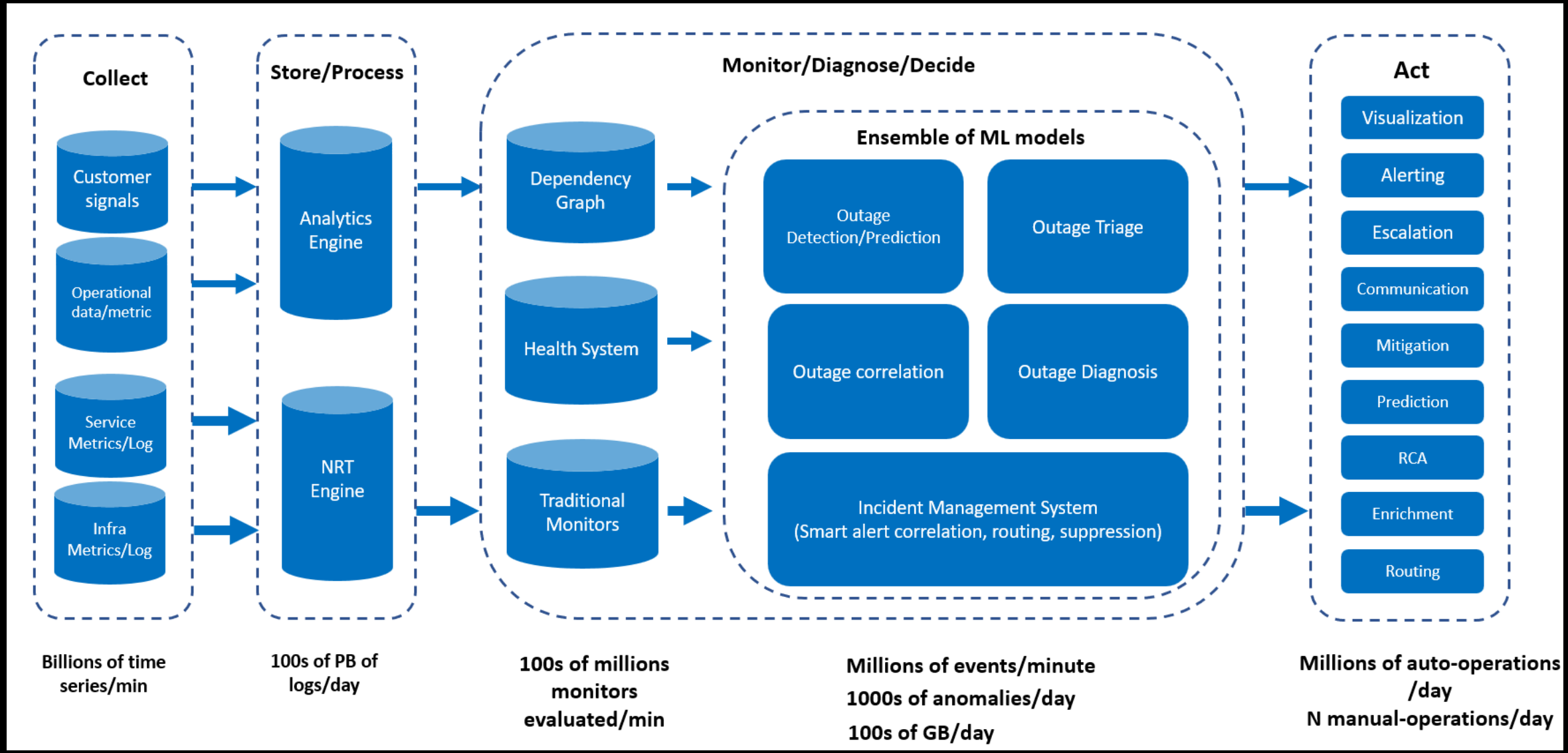
Root cause service identification

Efficient outage management

Diagnostic experiences

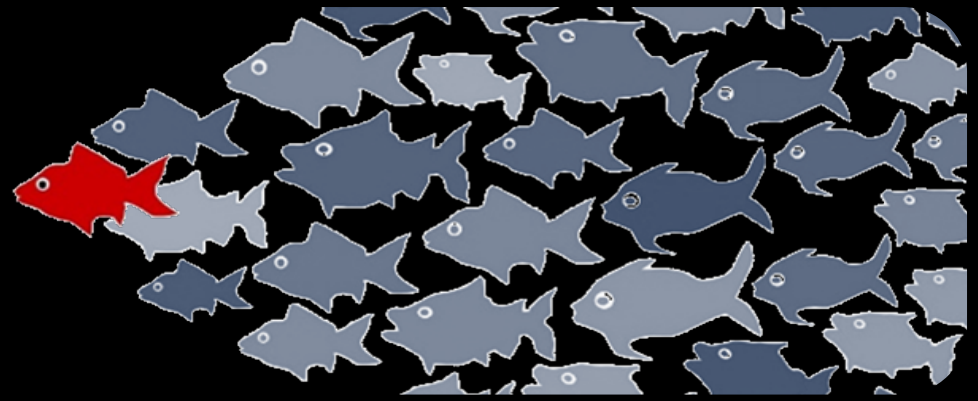
Auto-Mitigation

Azure BRAIN Intelligence Pipeline



Challenges

- Large-volume and heterogeneous non-uniform data
  - Huge # of data sources
  - Different schema/domain/version
  - Missing or dirty data
- Extremely imbalanced samples
  - small # of abnormal samples
  - Anomalies are rare but impactful
- Lack of canonical ground truth
  - No single source for reliable ground truth
  - Extremely costly and hard to label
- AI system and Human interaction
  - Conflict between Automation and Manual operation
  - Confidence and trust on the AI System
- No universal intelligence for diverse scenarios/domain
  - Hard to scale to many services
  - Utilize ensemble of ML models
- Existing or legacy Service Monitors
  - Moving teams to something new needs 10x

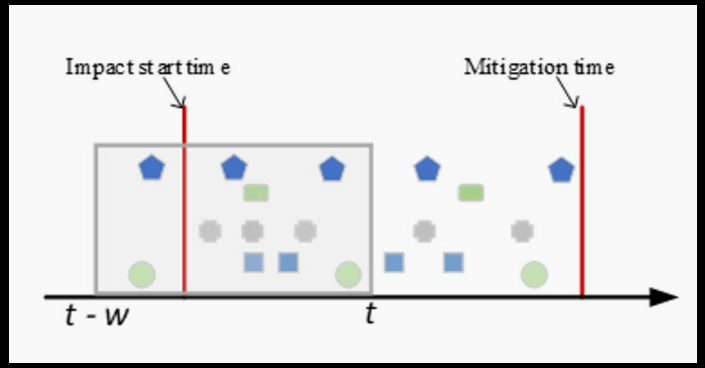


Abnormal:Normal

1:10,000

Ex: Outage Detection/Prediction using Alerts

- Incident detection: a binary classification problem
  - Input: alerts reported by selected monitors in a recent time window
  - Output: 1 if there is potential ongoing incidents; otherwise, 0
- Sample construction
  - Construct samples using a sliding window (3h)
  - Label = 1 if the window is overlapping with incident impact duration; otherwise, label = 0
- Feature extraction:
  - Alert signals: alert count, alert burst, alert source
  - Engineer activities: diagnosis log count, notification count
  - Others: region, working day, hour in day
- Classification model: BRF (Balanced Random Forest)



The Importance of Standard Metrics

- **Good Data In is necessary for any AIOps System**
  - **Garbage in = Garbage out**

- **Service Level Objectives (SLOs)** are goals set on a small number of key **customer-centric Service Level Indicators (SLIs)**.
  - They measure the customer experience of a service and determine whether we are meeting the promises made to our customers
  - Standard Metric across all Azure Services typically expressed as a metric or a ratio that can be monitored and measured over time to determine the service's health and performance
  - SLIs have different categories of data types that can be measured as a value type or percentile type including:
    - Availability, Success Rate, Latency, Throughput, Capacity, Traffic Rate

AIOps and LLM (Large Language Model)

- **Cloud CoPilot**: Infuse generative AI into how we design, build, and operate cloud services for delightful customer experience and engineering efficiency

Azure

How AIOps can solve a Manual Detection

Azure

Learning

Detection

- 5+ hours to declare outage as Sev 1, relying on human judgement
- Long TTD due to Manual incidents correlation

Lack of Customer SLIs

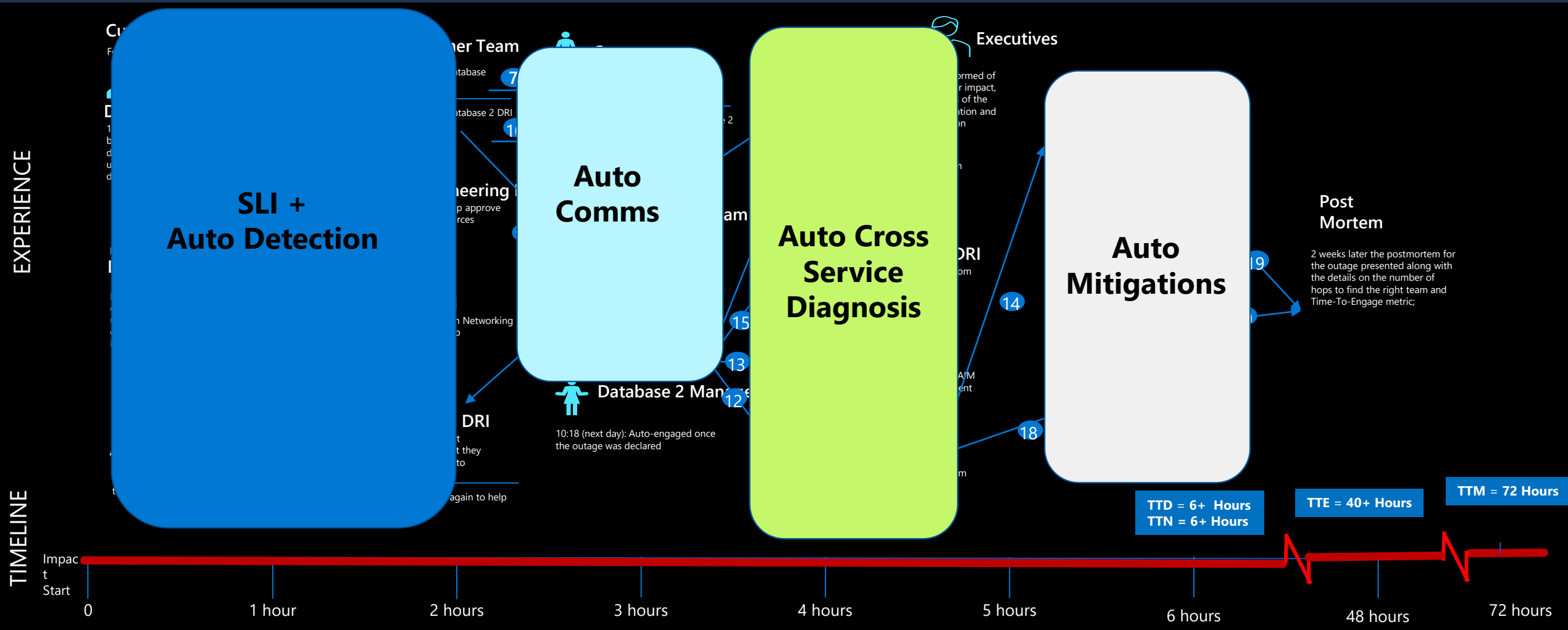
- Lack of standardized, accurate and reliable understanding of the health
- No centralized way to look at service health across the fleet

Diagnostic

- Multiple teams joined incident bridge to understand true owning services and root cause
- Dependencies not well understood
- Troubleshooting and Remediation still requires DRI toil and manual touches

Customer Communications

- TTN = 6+ hours as impacted customer information was not easily available



AIOps Benefits & Results

- First version of BRAIN deployed in Production in early 2019 and onboarded ~60 major Azure services in two years
- Major TTx improvement -> Higher reliability and better customer experience
- Incident/Alert Auto-correlation -> Less noise for On-Call engineer

72%
TTM Reduction

58%
TTN Reduction

100%
Auto-Comm percentage increase

25%
Incident noise reduction

98.26%
Detection Recall

98.83%
Detection Precision

Selected Microsoft Publications on AIOps

- Fighting the Fog of War: Automated Incident Detection for Cloud Systems, ATC'21
- How Long Will it Take to Mitigate this Incident for Online Service Systems?, ISSRE'21 (best research paper award)
- HALO: Hierarchy-aware Fault Localization for Cloud Systems, KDD'21
- Efficient Incident Identification from Multi-dimensional Issue Reports via Meta-heuristic Search, FSE'20
- Toward ML-Centric Cloud Platforms: Opportunities, Designs, and Experience with Microsoft Azure, CACM'20
- Identifying Linked Incidents in Large-scale Online Service Systems, FSE'20
- Predictive and Adaptive Failure Mitigation to Avert Production Cloud VM Interruptions, OSDI'20
- Intelligent Virtual Machine Provisioning in Cloud Computing, IJCAI'20
- An Intelligent, End-To-End Analytics Service for Safe Deployment in Large-Scale Cloud, NSDI'20
- Rex: Preventing Bugs and Misconfiguration in Large Services using Correlated Change Analysis, NSDI'20
- AIOps Innovations in Incident Management for Cloud Services, Cloud Intelligence Workshop, AAAI'20
- Identifying Linked Incidents in Large-scale Online Service Systems, FSE'20
- How to Mitigate the Incident? An Effective Troubleshooting Guide Recommendation Technique for Online Service Systems, FSE'20 Industry
- Efficient Customer Issue Triage via Linking with System Incidents, FSE'20 Industry
- Towards Intelligent Incident Management: Why We Need it and How We Make it, FSE'20 Industry
- How Incidental are the Incidents? Characterizing and Prioritizing Incidents for Large-Scale Online Service Systems, ASE'20
- Robust Log-based Anomaly Detection on Unstable Log Data, FSE'19
- Towards More Efficient Meta-heuristic Algorithms for Combinatorial Test Generation, FSE'19
- Cross-dataset Time Series Anomaly Detection for Cloud Systems, USENIX ATC'19
- AIOps: Real-World Challenges and Research Innovations, Tech briefing, ICSE'19
- Outage Prediction and Diagnosis for Cloud Service Systems, WWW'19
- An Empirical Investigation of Incident Triage for Online Service Systems, ICSE'19
- Continuous Incident Triage for Large-Scale Online Service Systems, ASE'19
- Orca: Differential Bug Localization in Large-Scale Services, OSDI'18
- Identifying Impactful Service System Problems via Log Analysis, FSE'18
- Predicting Node Failure in Cloud Service Systems, FSE'18
- BigIN4: Instant, Interactive Insight Identification for Multi-Dimensional Big Data, SigKDD'18
- Improving Service Availability of Cloud Systems by Predicting Disk Error, USENIX ATC'18

Microsoft

Q&A

Microsoft

Thank you!