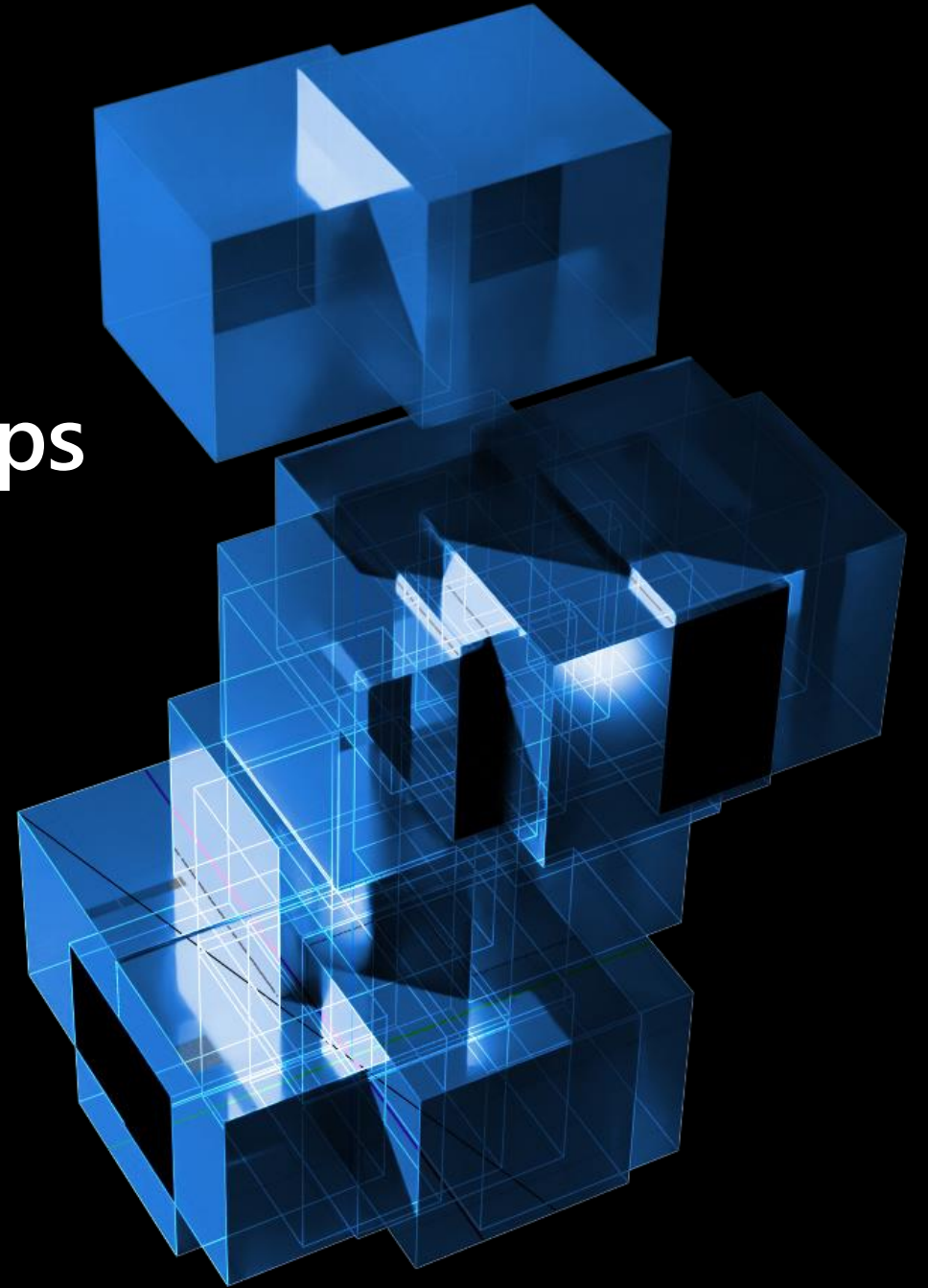


Managing Cloud Health with AIOps

Cong Chen

Principle Data Scientist Manager, Azure Edge + Platform



Agenda

- Introduction to Azure
- The pivot to AIOps
- AIOps at scale
 - BRAIN
 - Gandalf



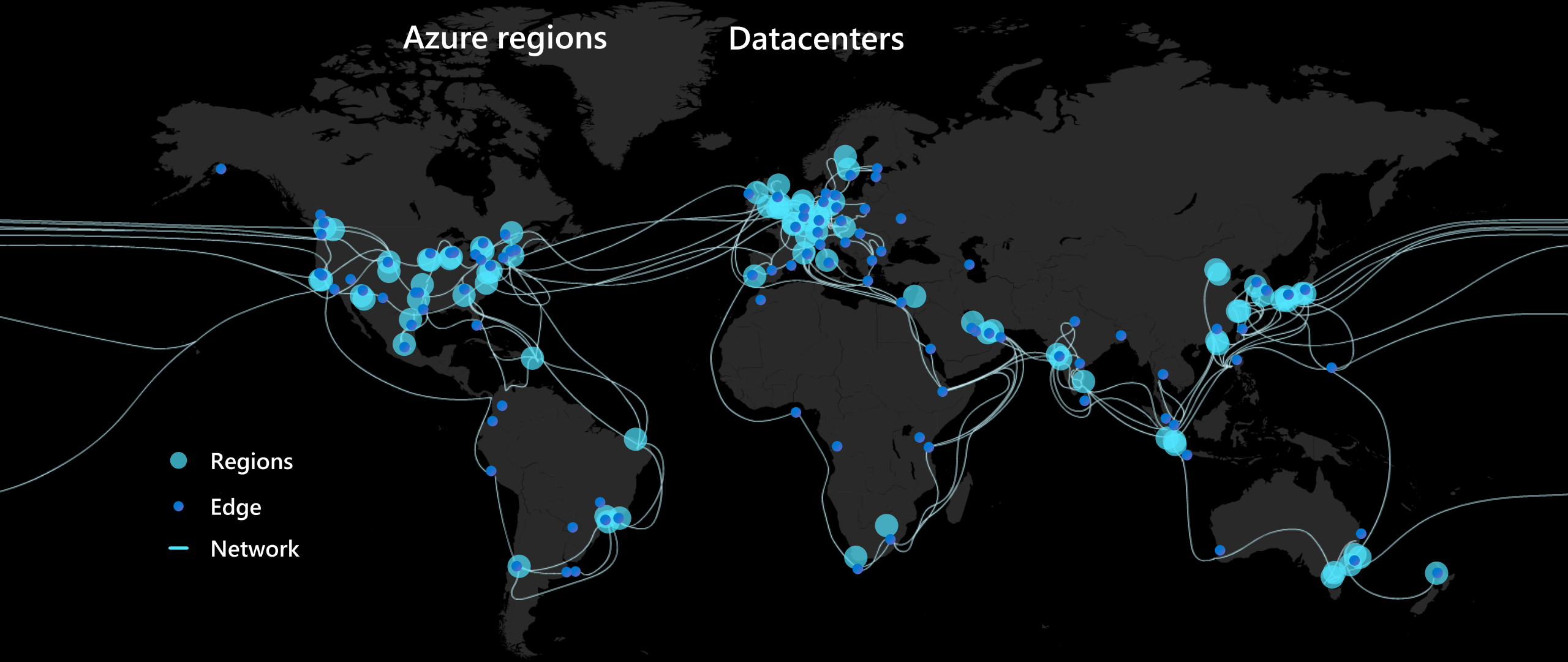
Azure is the world's computer

66+

Azure regions

200+

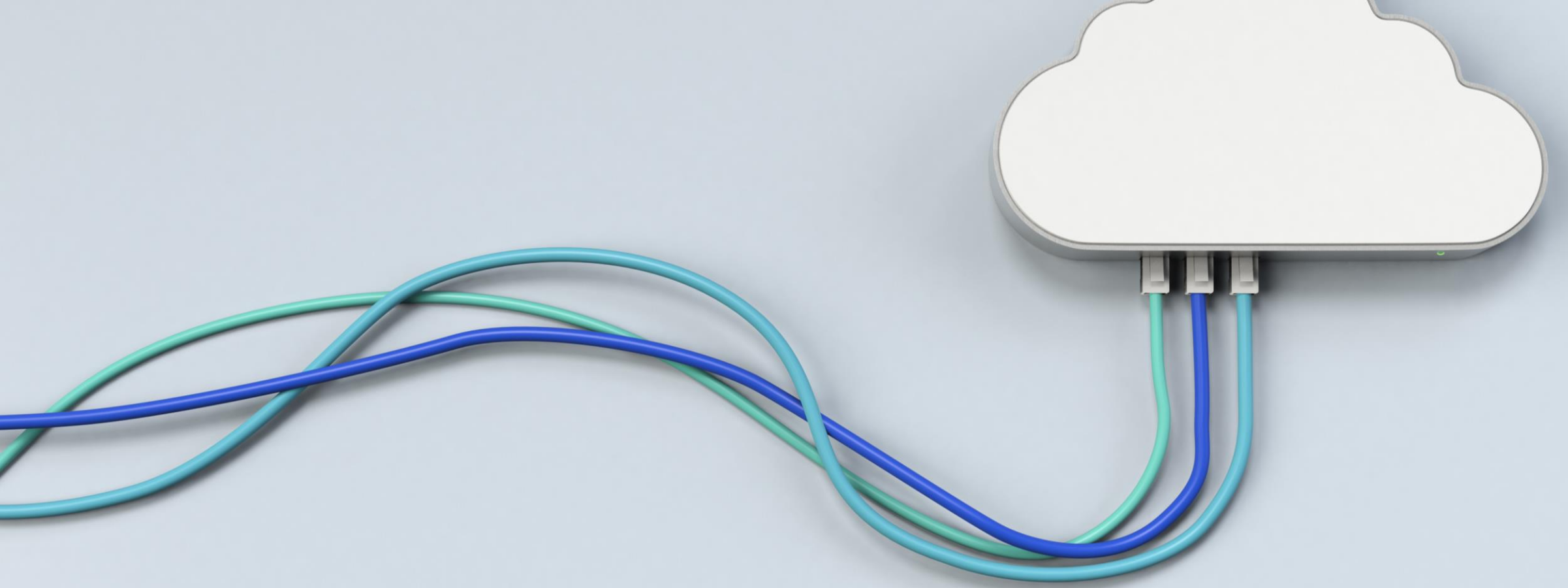
Datacenters





Microsoft

Our Approach to Manage Cloud Health



Comprehensive **standardized, accurate** and **reliable**
understanding of service health



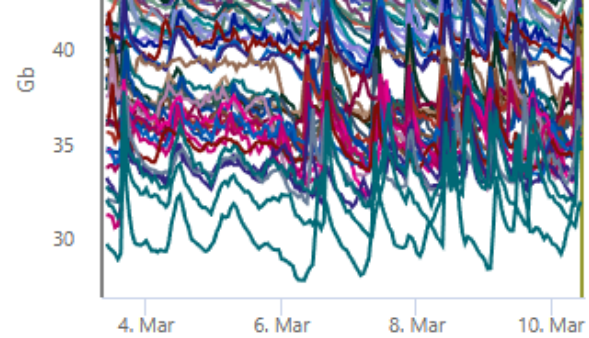
Service Health in **Near Real Time** using Service Level Indicators (**SLIs**) and Service Health Indicators (**SHIs**)



Production changes **integrated** with service health

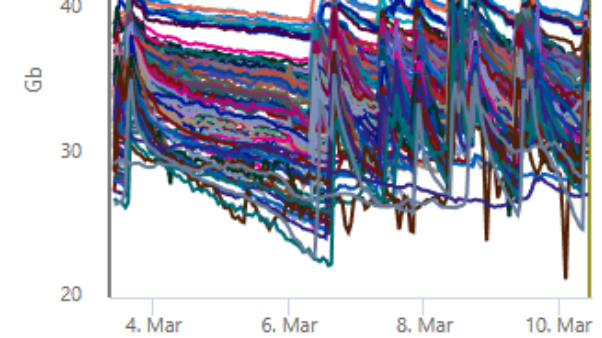


If an incident occurs, communication
is **automated** and **timely**



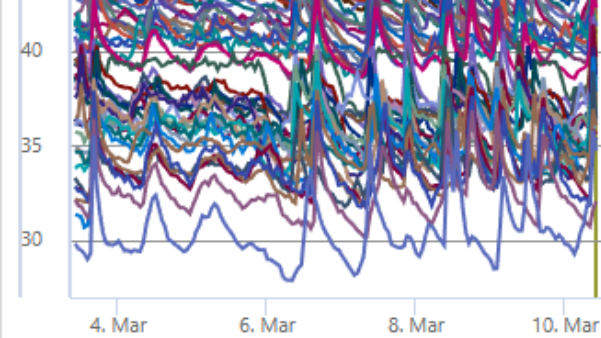
- Norway West
 - South Central US STG
 - East US STG
 - Brazil US
 - East US 2 EUAP
- ▲ 1/13 ▼

All None



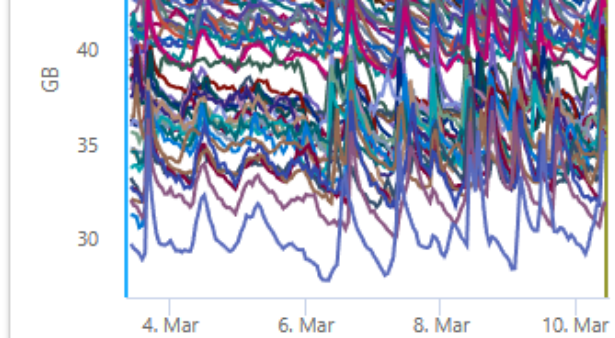
- East US STG
 - South Central US STG
 - East US STG
 - South Central US STG
 - Brazil US
- ▲ 1/25 ▼

All None



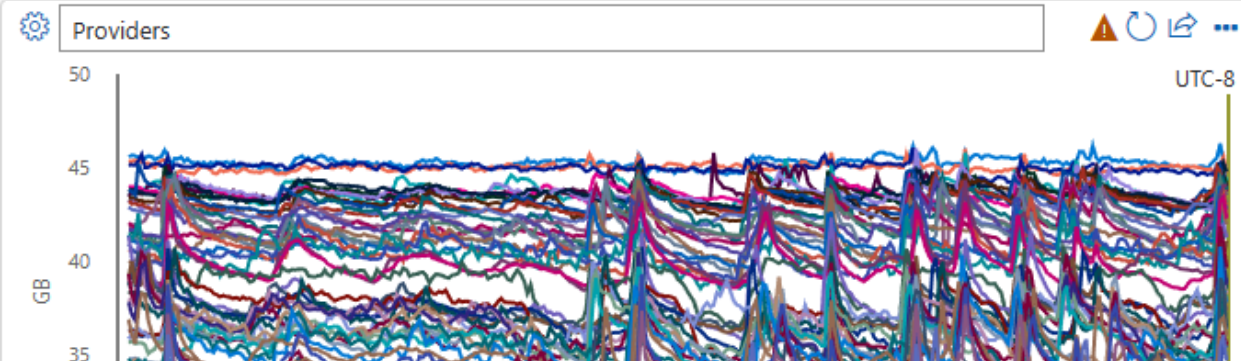
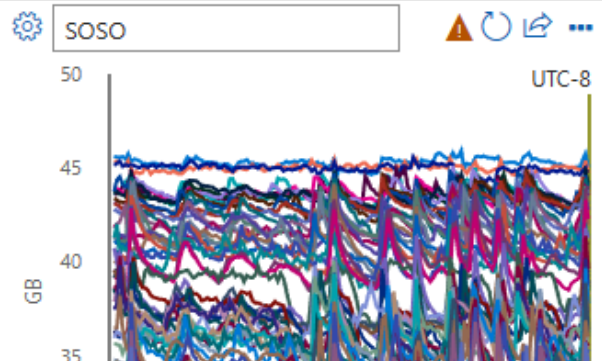
- Norway West / 1024
 - South Central US STG / 1024
 - East US STG / 1024
 - Brazil US / 1024
 - East US 2 EUAP / 1024
- ▲ 1/13 ▼

All None



- Norway West / 1024
 - South Central US STG / 1024
 - East US STG / 1024
 - Brazil US / 1024
 - East US 2 EUAP / 1024
- ▲ 1/13 ▼

All None

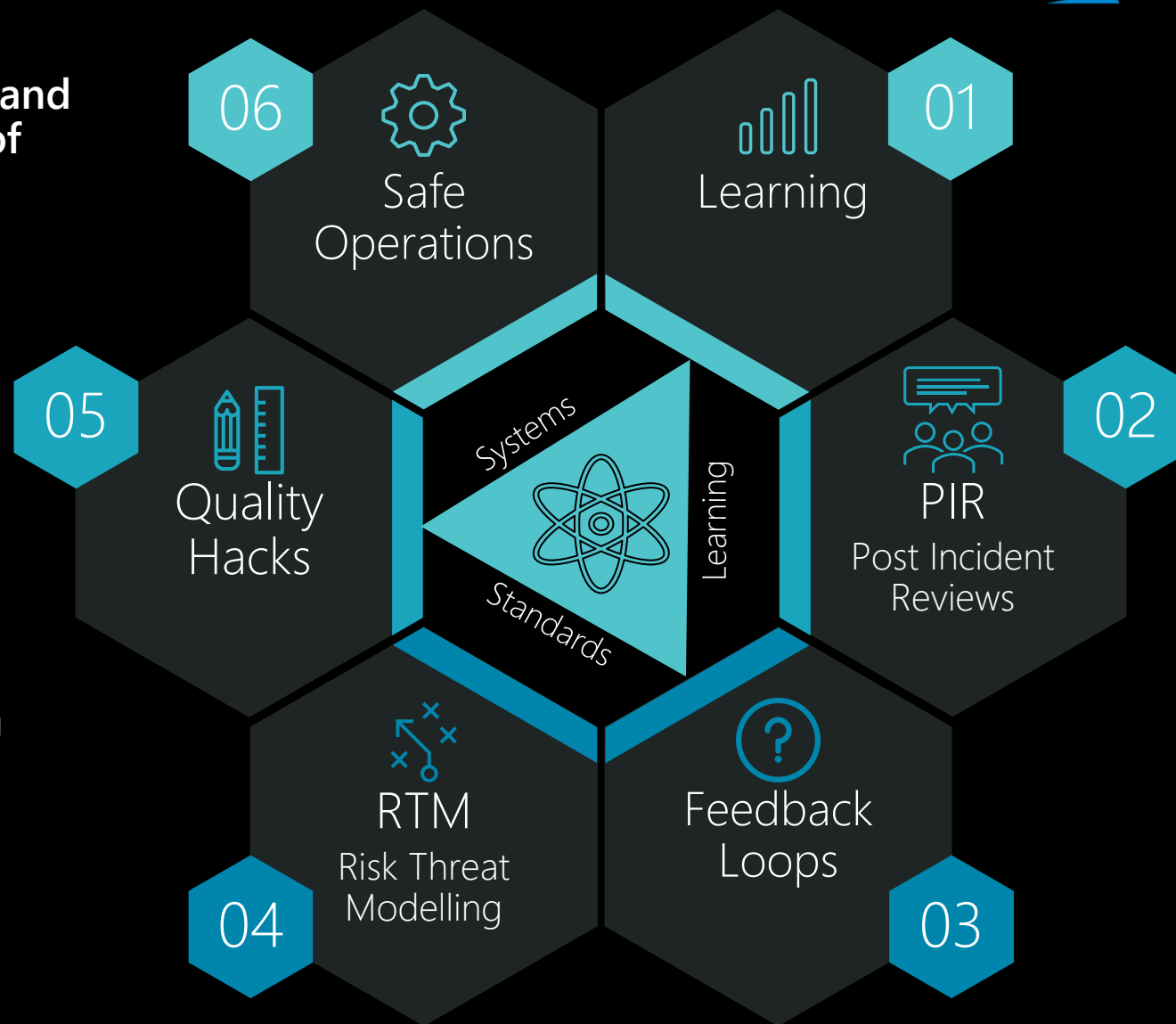


Diagnosing issues is **simple / automated**, requiring little DRI toil or manual touches

Culture of Quality

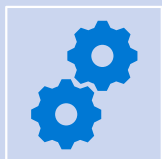
Cultivate an environment where we **Listen** and **Learn** to understand the lived experience of our DRIs and our customers.

-  Local Learning and Sharing Opportunities
-  Seek Diverse Perspective in PIRs
-  Create Safe Space for Feedback
-  Quality Planning through Risk Modelling
-  Dedicate Time for Quality Innovation
-  There is No Human Error



The pivot to AIOps

AIOps – Gartner’s definition



Big data and ML driven IT operation automation process

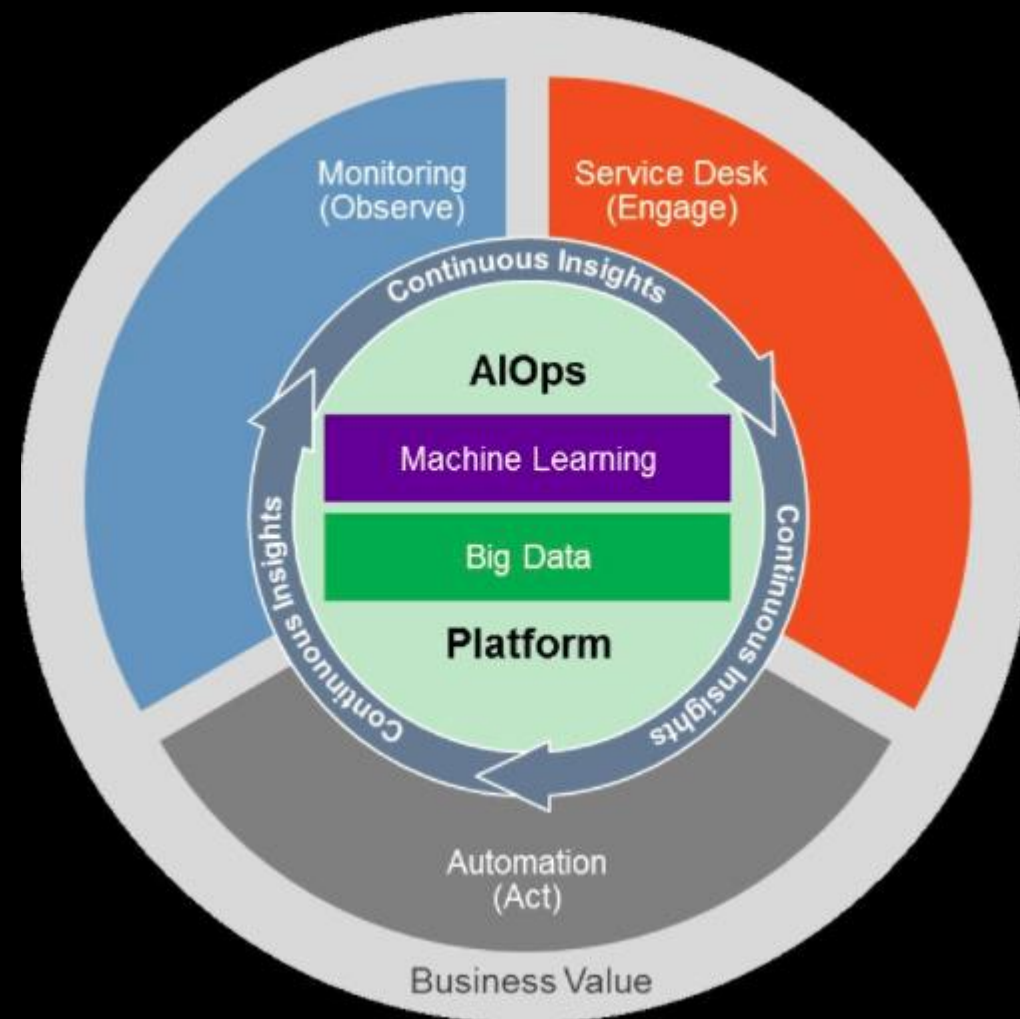


Adoption has increased with the uptick of digital transformation



Business value

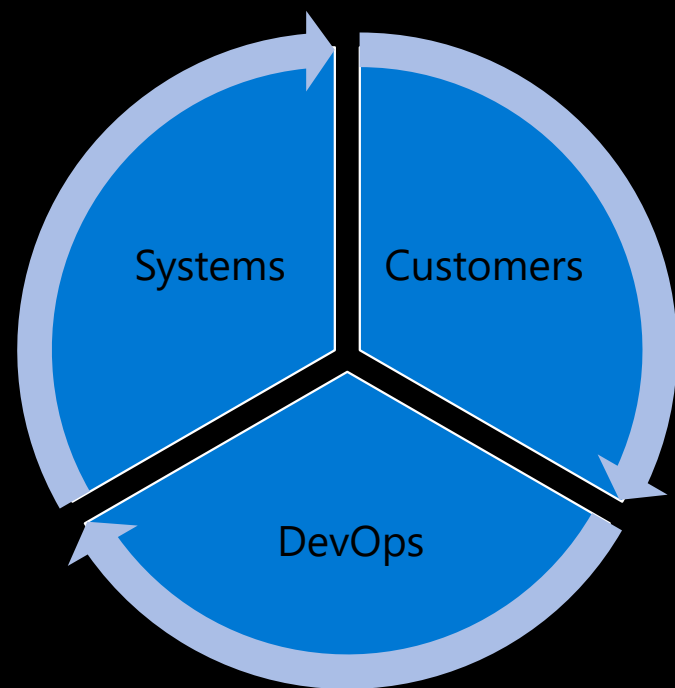
- Higher efficiency
- Higher Service quality
- Lower COGS



Source: Gartner

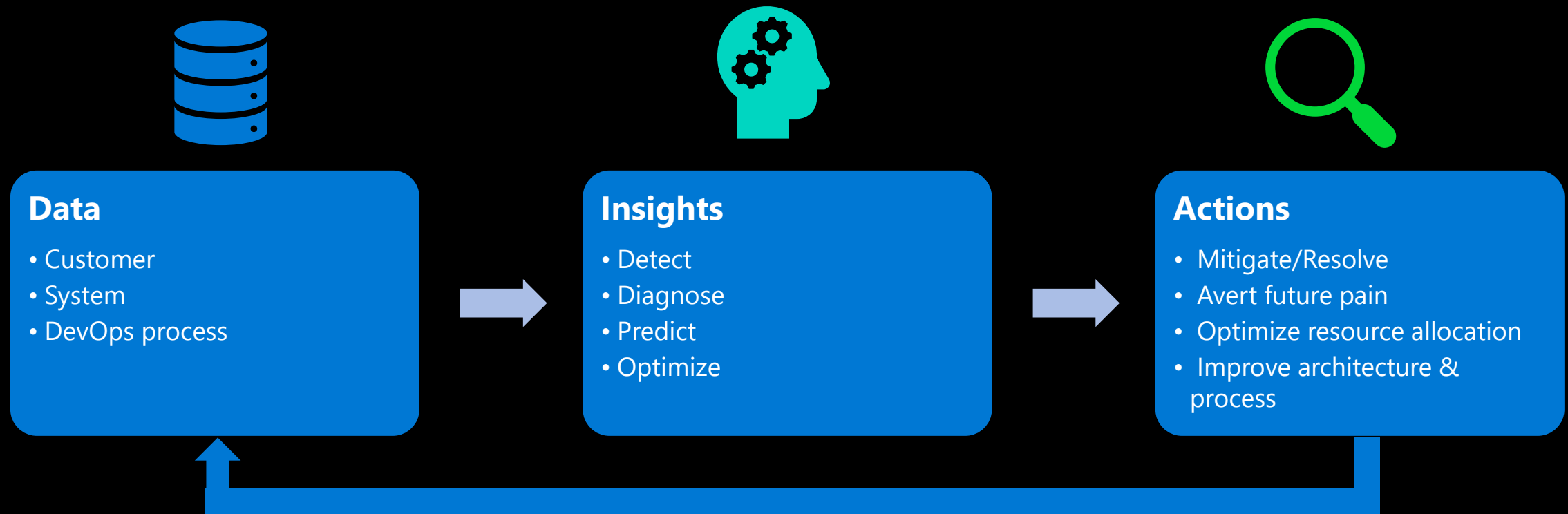
AIOps in Azure

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

AI Ops Methodologies: From Data to Actions



How we do it at scale

Azure BRAIN

Cloud Reliability

State-of-the-art Cloud Reliability

- 5/6-9s' availability (99.999% means less than 5min down time per year)
- 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](#)

Azure BRAIN

(Blackout Resilient Azure Intelligence Network)



Customer Experience



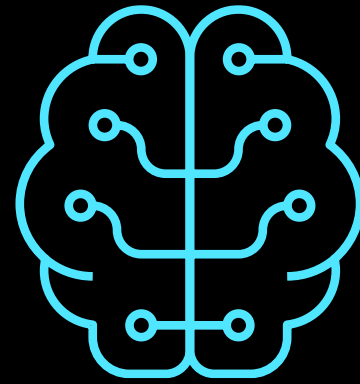
Azure Services



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

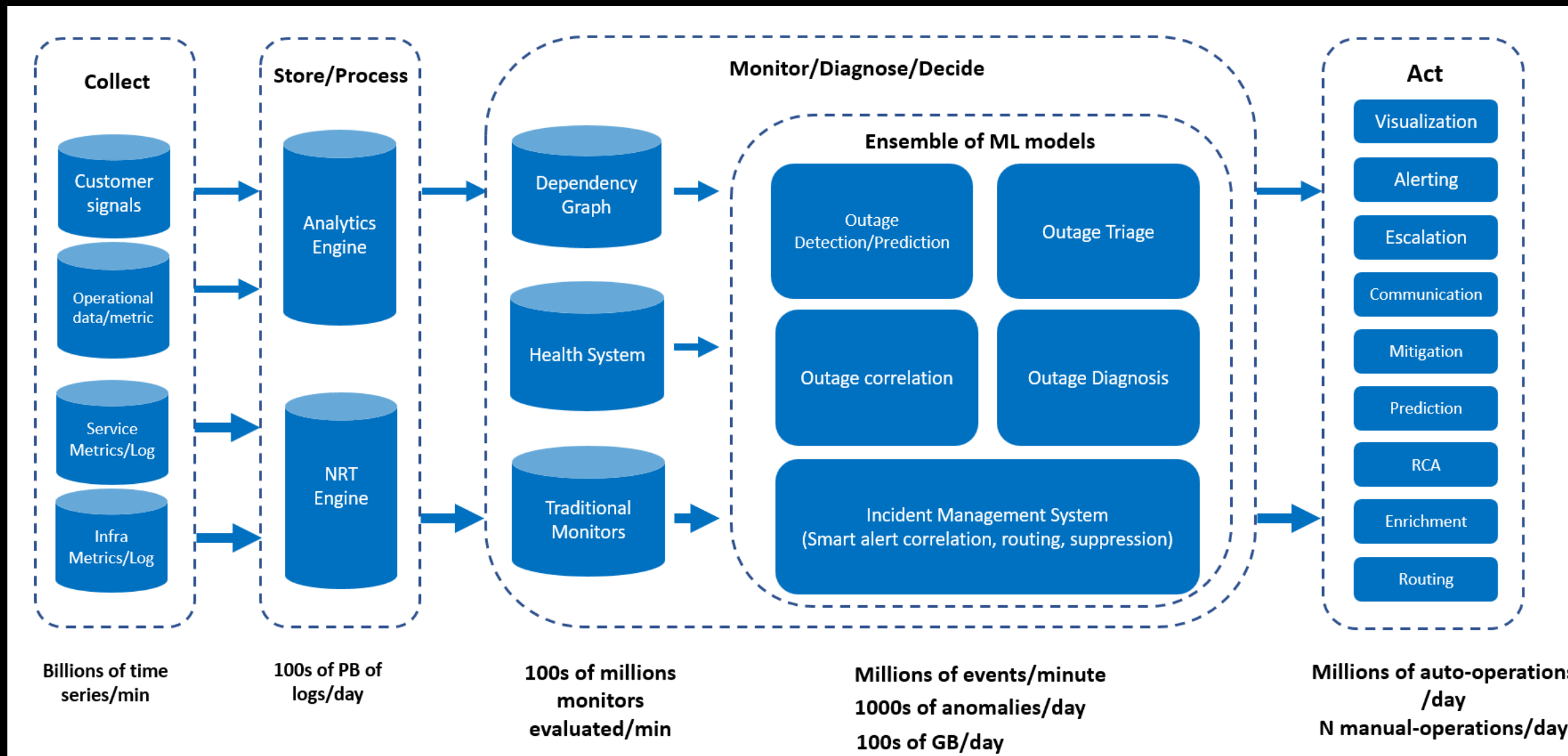


Diagnostic experiences



Auto-Mitigation

Azure BRAIN Intelligence Pipeline



Challenges

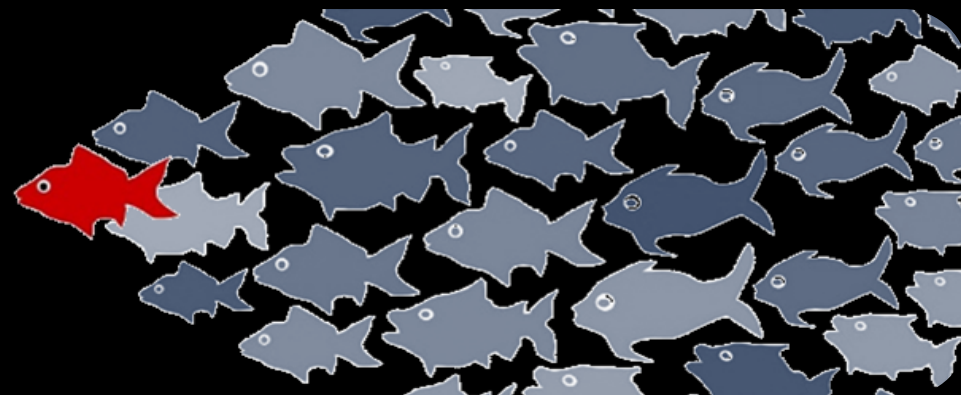
Large-volume and heterogeneous non-uniform data

Extremely imbalanced samples

Lack of canonical ground truth

AI system and human interaction

No universal intelligence for diverse scenarios



Abnormal:Normal

1:10,000

AIOps Benefits & Results

First version of BRAIN deployed in Production in early 2019

Onboarded 100+ major Azure services

Major TTx improvement

Incident/alert auto-correlation -> Less noise

72%

TTM Reduction

58%

TTN Reduction

100%

Auto-Comm
percentage
increase

25%

Incident noise
reduction

98.26%

Detection Recall

98.83%

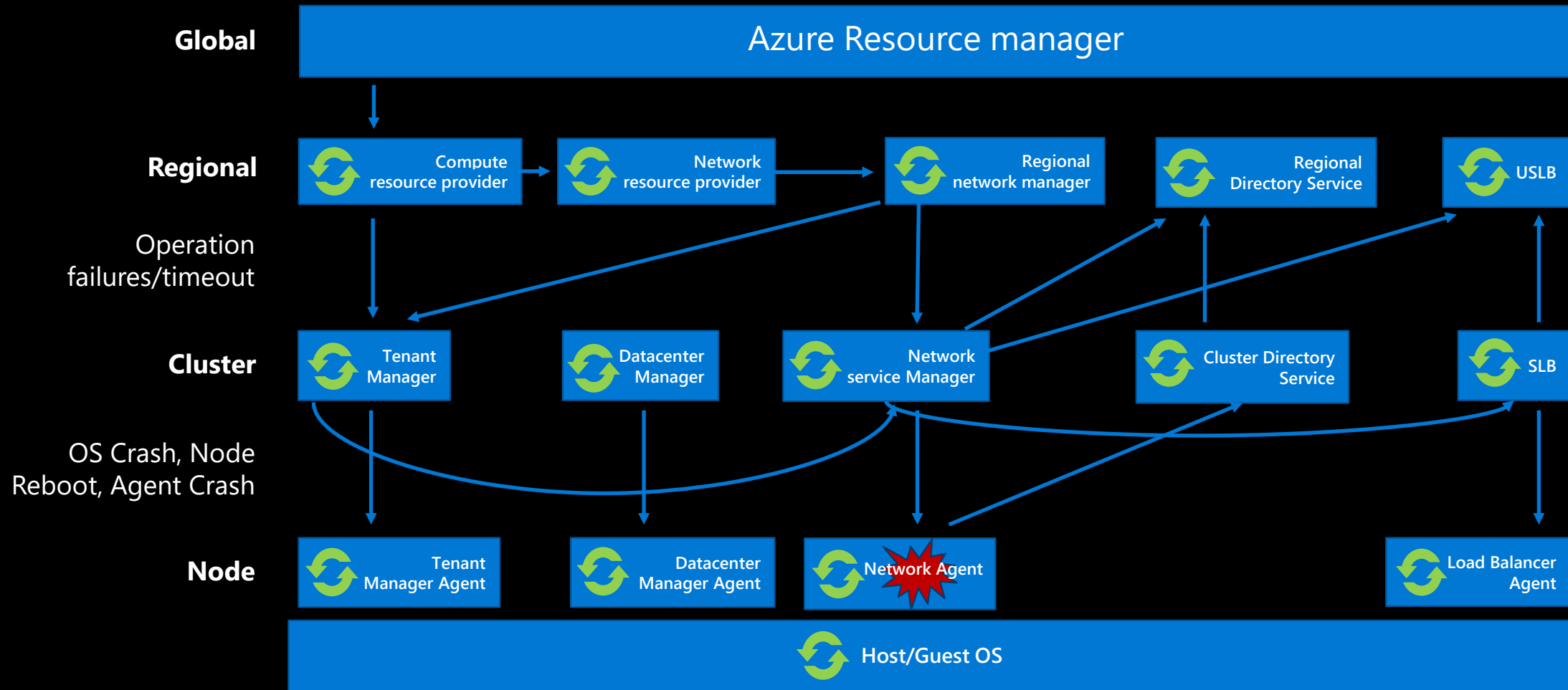
Detection Precision

Azure Gandalf

Azure Gandalf: AIOps for Infrastructure Health

- Proactive prevention of issues: integrating intelligence into Azure Infrastructure
 - Preventing code regressions into fleet
 - Increasing host resilience
 - Governance of host resource usage
- Effective and efficient action-taking: integrating intelligence into Azure DevOps
 - Effective monitoring and diagnosis
 - Thousands of high-quality tickets filed every year
 - Increased deployment velocity by ~4 times

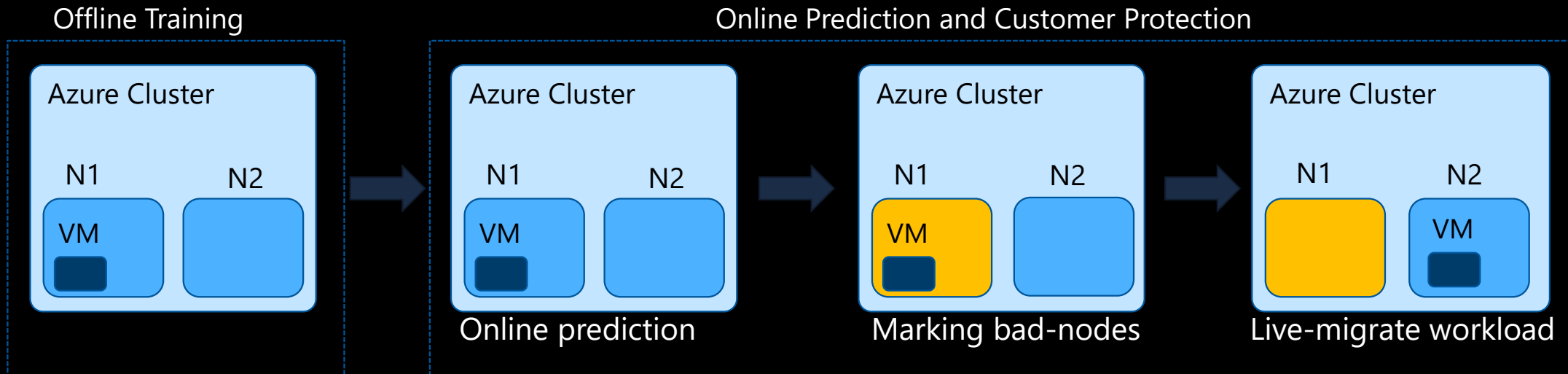
Preventing Code Regressions: Challenges



Increasing Host Machine Resilience



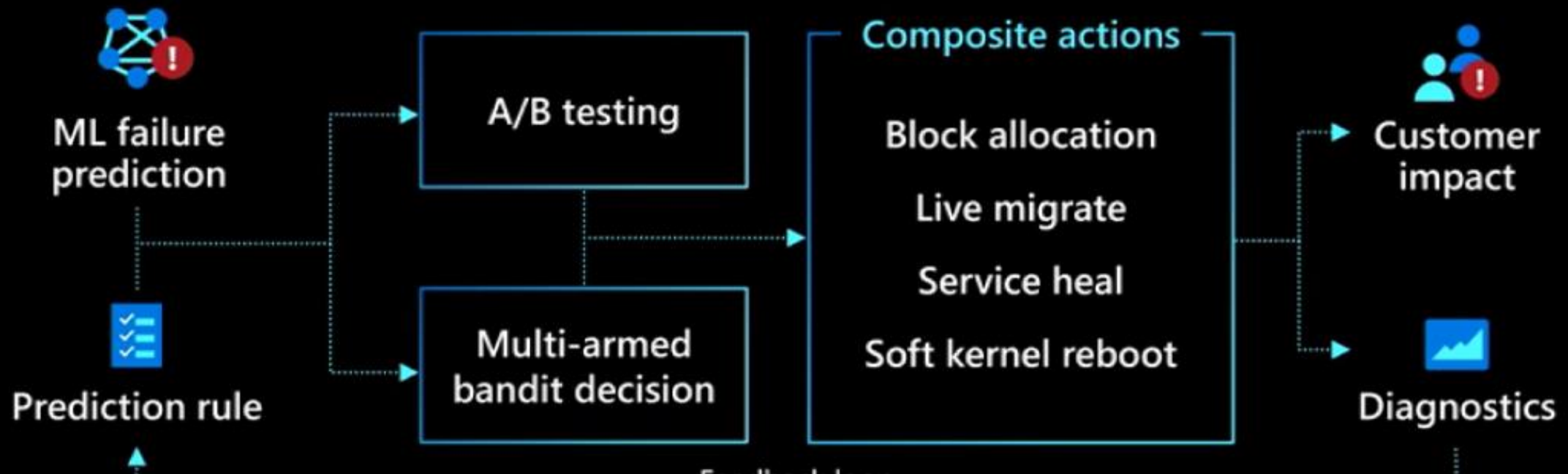
Goal – minimize VM reboots due to host failures by triggering Live Migration (moving VMs to healthy node with only a few seconds of blackout time) and other protection methods



Increasing Host Machine Resilience (Cont'd)

Project Narya

Predictive and adaptive failure prevention



Further read

- Azure blog: <https://azure.microsoft.com/en-us/blog/advancing-failure-prediction-and-mitigation-introducing-narya/>
- Sebastien Levy, et. al., Predictive and Adaptive Failure Mitigation to Avert Production Cloud VM Interruptions, OSDI 2020

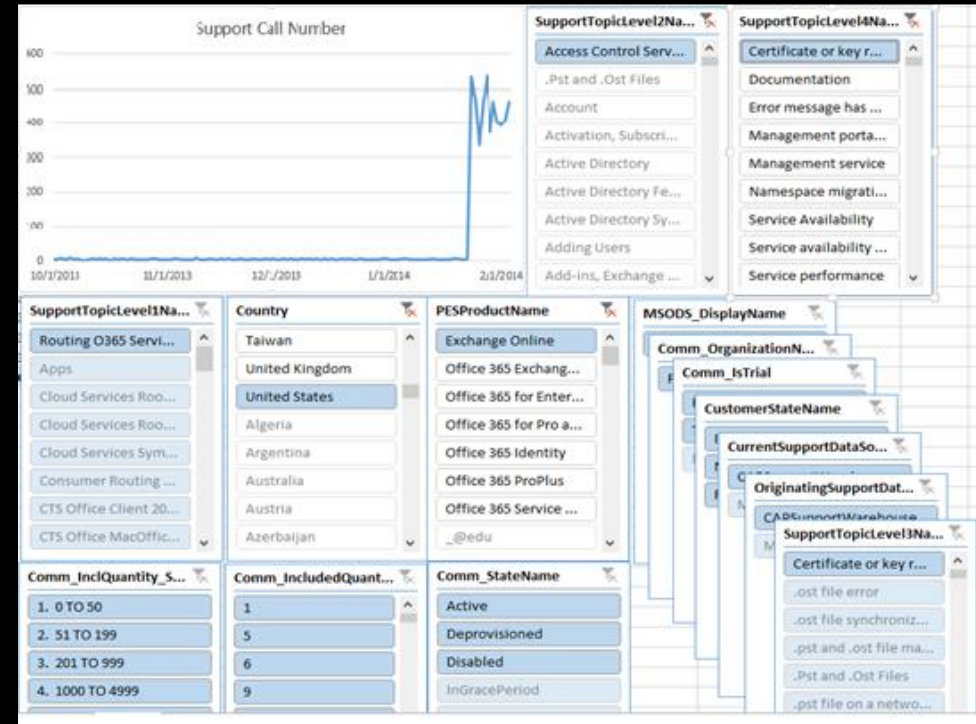
Multi-dimensional Anomaly Detection

Common Practice

- Manually identify monitor combinations with pivot table
- Set up pipelines to monitor hundreds of thousands of time series
- Total Time Series: 100,000+*

Our solution

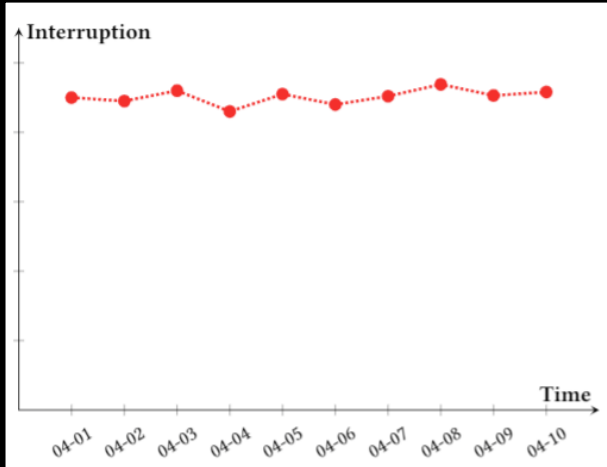
- Formulated as a "combinatorial optimization problem"
- Solved by a specific-tailored "meta-heuristic search" method
- Details see the paper* from our Microsoft Research partners



*paper: "Efficient Incident Identification from Multi-dimensional Issue Reports via Meta-heuristic Search", FSE 2020

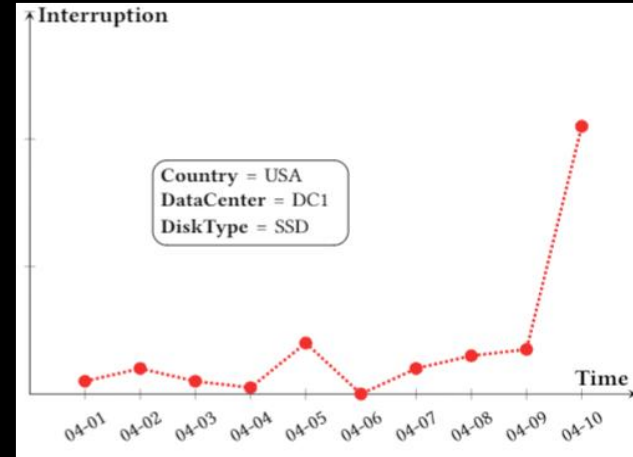
Multi-dimensional Anomaly Detection - Motivation

(Example case of Azure VM Interruptions)



Time	Interruption
2019-04-01	100
2019-04-02	99
2019-04-03	103
2019-04-04	97
2019-04-05	103
2019-04-06	99
2019-04-07	98
.....

Overall KPI not having obvious spike



Time	Country	Datacenter	Disk Type	Interruption
2019-04-10	USA	DM1	SSD	1
2019-04-10	Australia	MEL21	SSD	1
2019-04-10	USA	DC1	HDD	4
2019-04-10	India	BL1	SSD	10
2019-04-10	UK	SN6	Hybrid	3
2019-04-10	USA	DM1	HDD	0
.....

Spike observed in a particular pivot

AI Ops 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



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



Q&A



Thank you!