Towards Autonomous Data Centers

Dr. Hanan Shteingart, NVIDIA

Artificial Intelligence and Machine Learning in Networking Workshop

Netdev 0x17, THE Technical Conference on Linux Networking

The Autonomous Data Center

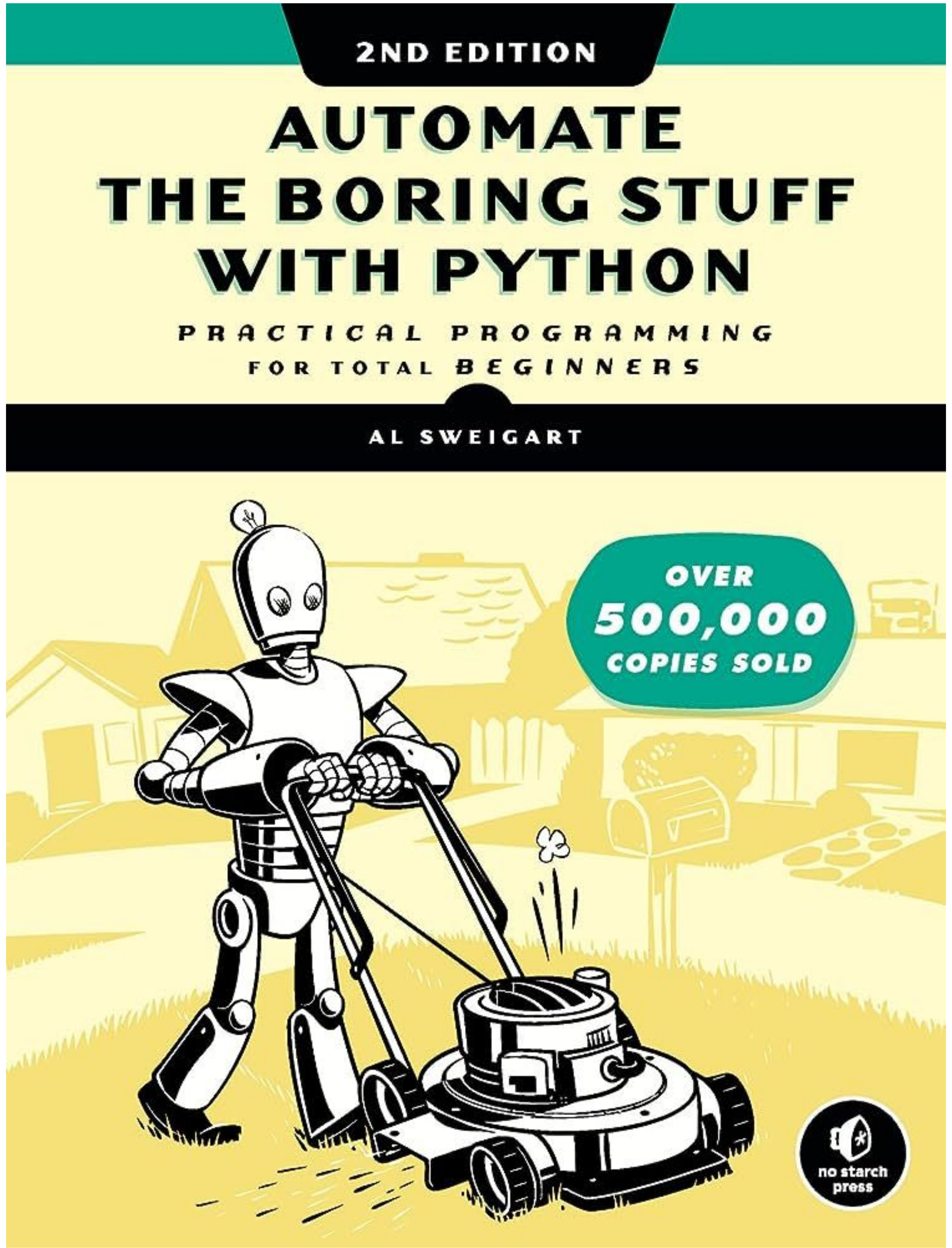
Motivation

The Problem:

1. Data centers are complex systems
2. It is hard to maintain and optimize for performance
3. They require experts
4. Down time is expensive

The Vision:

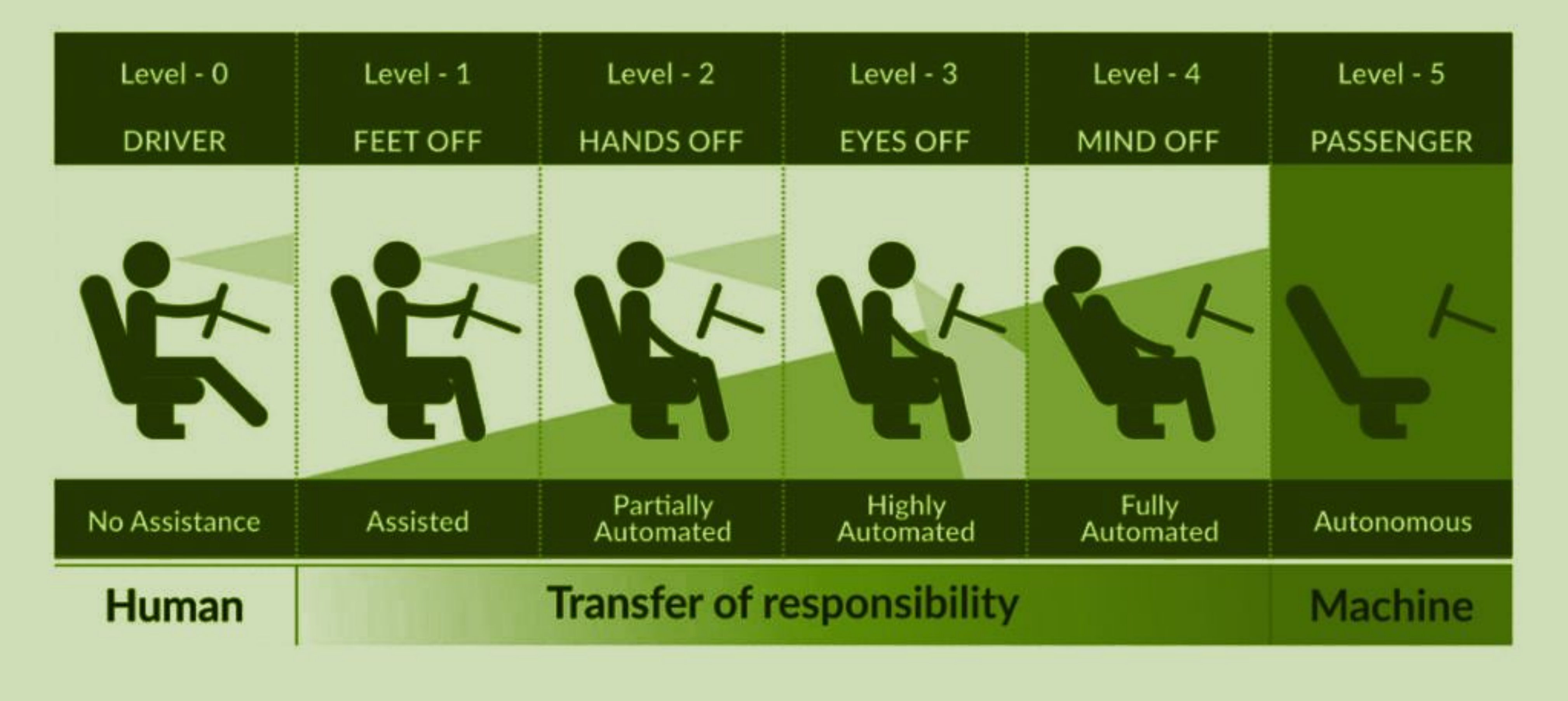
1. Automate the boring stuff
2. Predict issues before they happen
3. Fix issues faster
4. Ongoing optimization

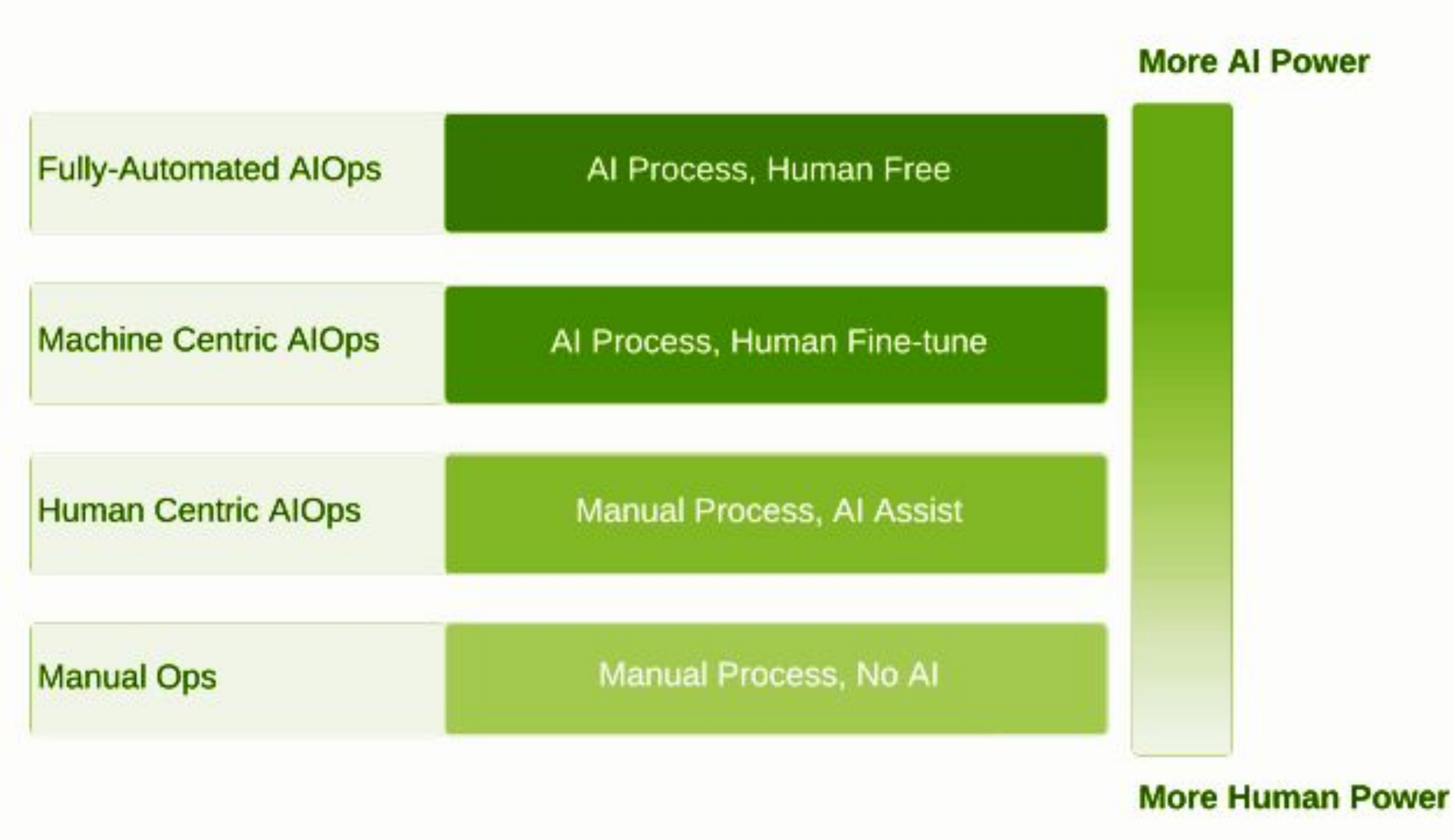

2ND EDITION
AUTOMATE THE BORING STUFF WITH PYTHON
PRACTICAL PROGRAMMING FOR TOTAL BEGINNERS
AL SWEIGART
OVER 500,000 COPIES SOLD
no starch press

Automate the Boring Stuff with Python

Autonomous Steps

Analogy to Autonomous Driving





[2304.04661] AI for IT Operations (AIOps) on Cloud Platforms: Reviews, Opportunities and Challenges

Autonomous DC - Operation, Performance and Cyber

We want a faster more reliable cars

Autonomous Data Center

+ cyber Security

Operations

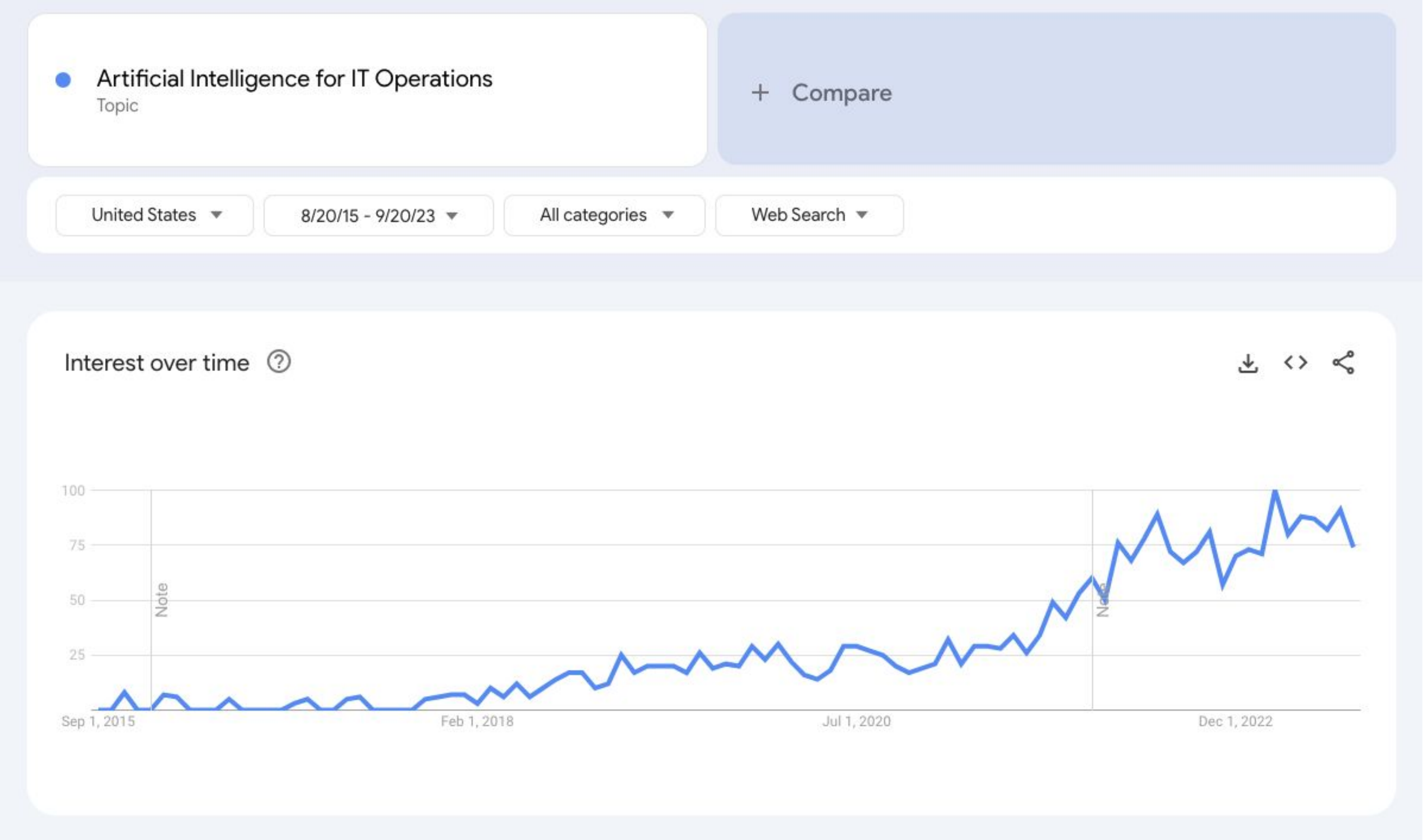
Performance

What is AI for IT Operations (AIOps)
and why do we care?

”AIOps combines big data and machine learning to automate IT operations processes, including event correlation, anomaly detection and causality determination” Gartner 2016

Manual Ops are:

- ○ hard to scale
- ○ hard to standardize
- ○ error-prone



● Artificial Intelligence for IT Operations
Topic

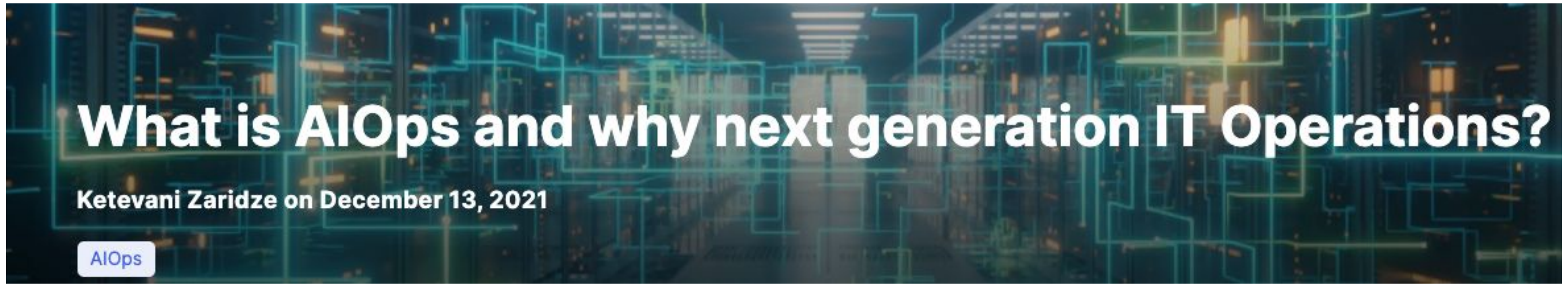
+ Compare

United States ▾ | 8/20/15 - 9/20/23 ▾ | All categories ▾ | Web Search ▾

Interest over time

AIOPs aim to maximize availability and enhance operational efficiency

Next Thing or Hype?

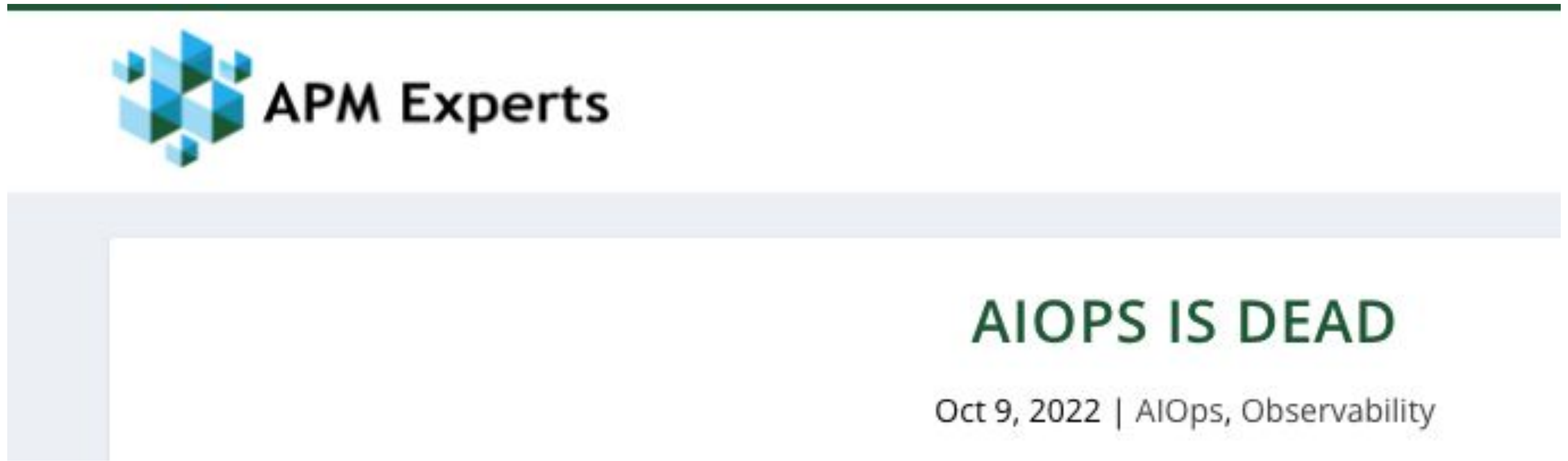


What is AIOps and why next generation IT Operations?

Ketevani Zaridze on December 13, 2021

AIOps

[What is AIOps and why next generation IT Operations? | Logmind Blog]

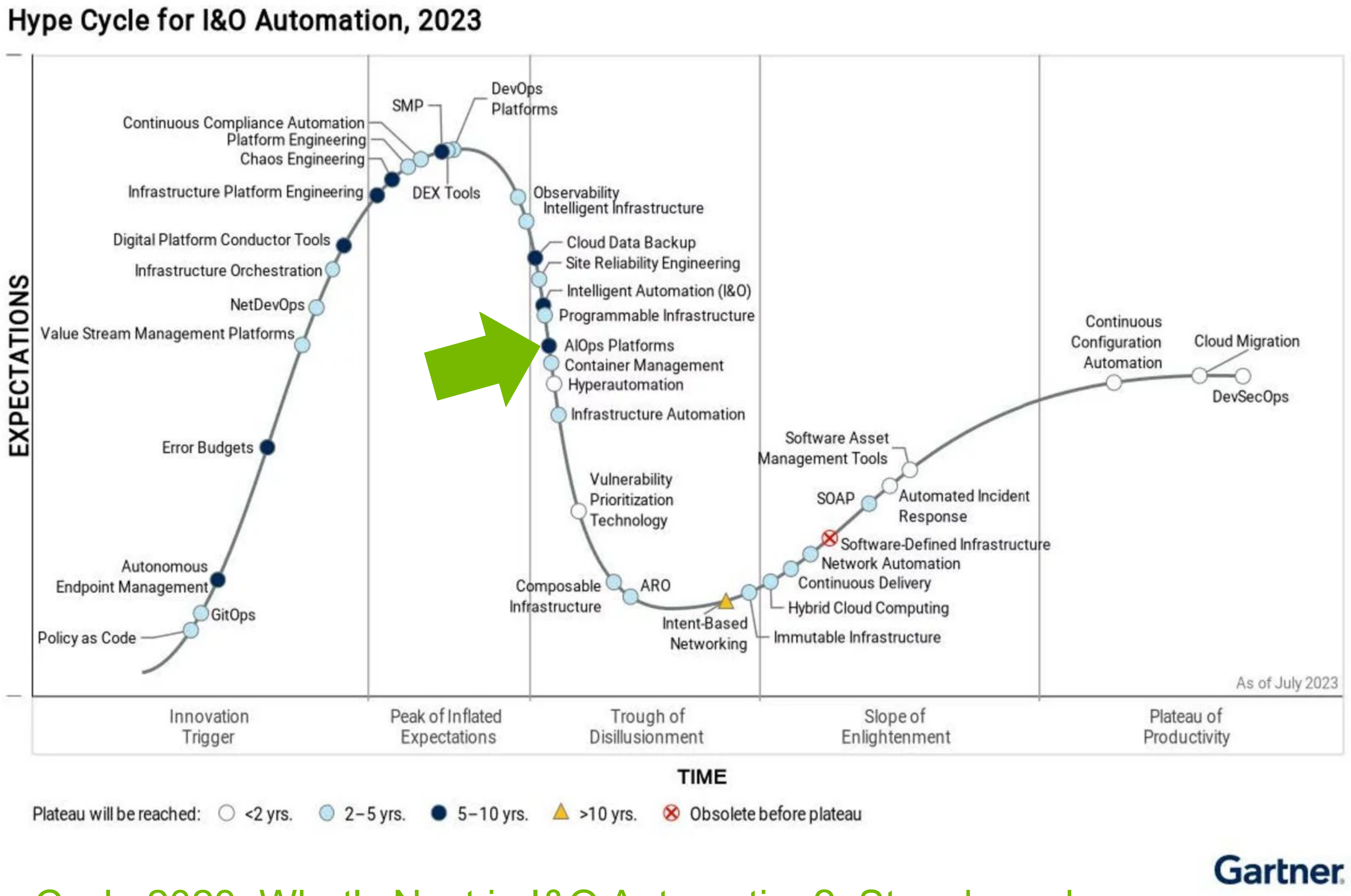


APM Experts

AIOPS IS DEAD

Oct 9, 2022 | AIOps, Observability

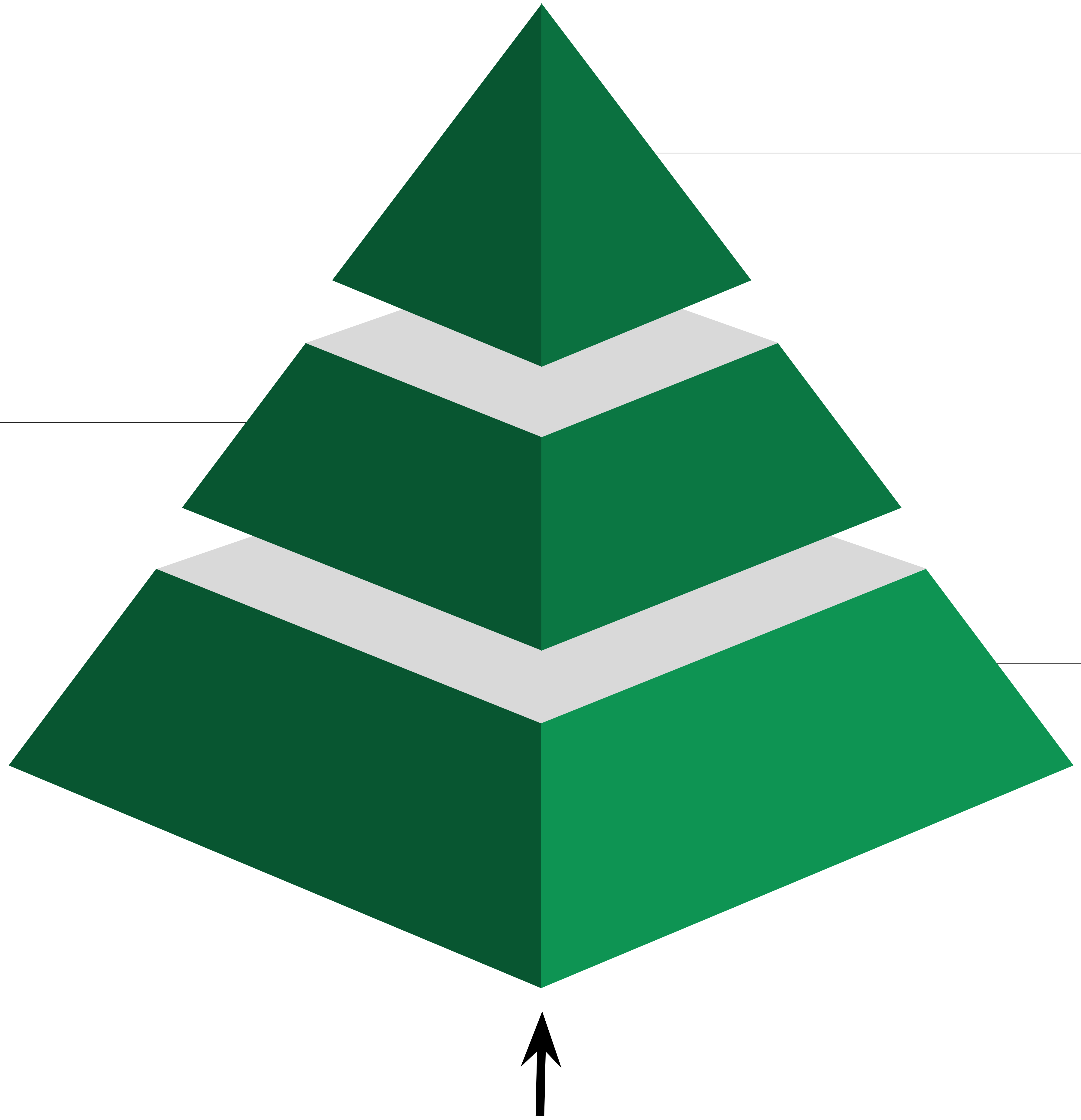
[AIOPs is Dead - APM Experts]



[Gartner Hype Cycle 2023: What's Next in I&O Automation?: Stonebranch]

AIOps: Success Metrics

Three layers of metrics



User Metrics Layer

1 % Failed AI Jobs, Transactions, etc.

Operations Metric Layer

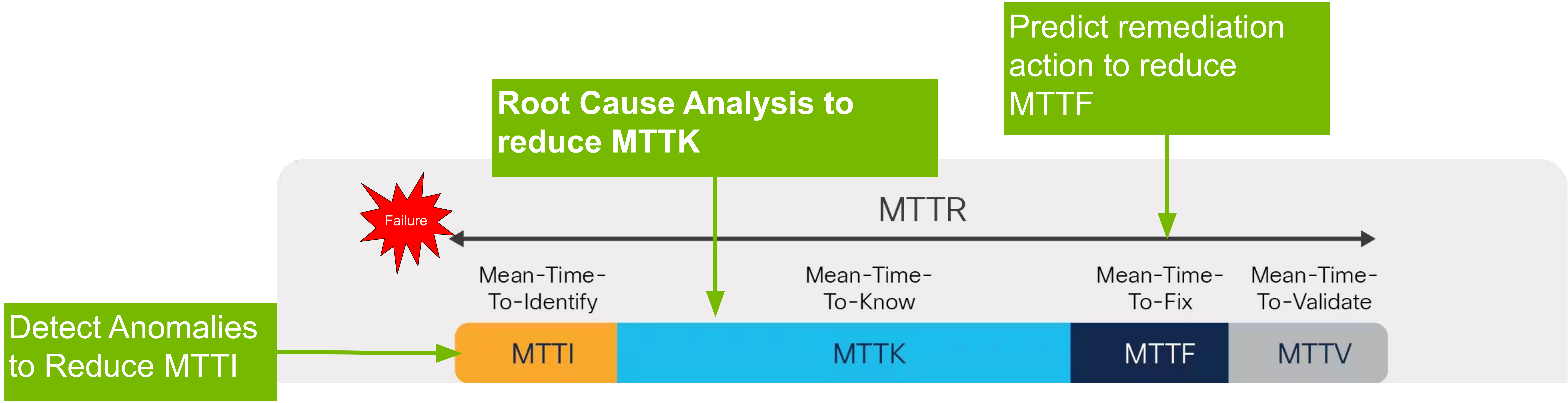
2 MTTD, MTTT, MTTR

AI Metrics Layer

3 e.g., Precision, Recall, F1-Score

AIOPs Model Output

Zoom in on MTTR



Root Cause Analysis to reduce MTTK

Predict remediation action to reduce MTTF

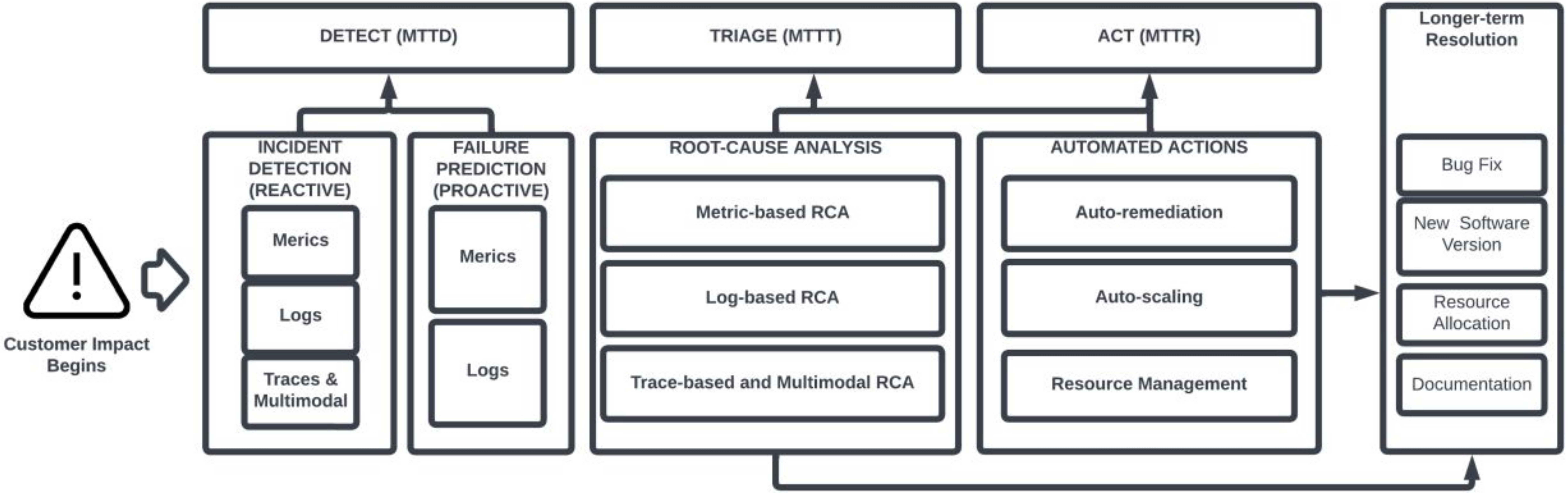
Detect Anomalies to Reduce MTTI

Service-Centric Approach to AIOps White Paper - Cisco

Operation

How will AI ops help operation?

- **Increase cluster availability**, **decrease job failure rate** (Reduce MTTD, MTTT, MTTR, MTBF), **scale, Power Efficiency**
- Use cases: Root cause analysis, predictive maintenance, auto remediation, power optimization, etc.

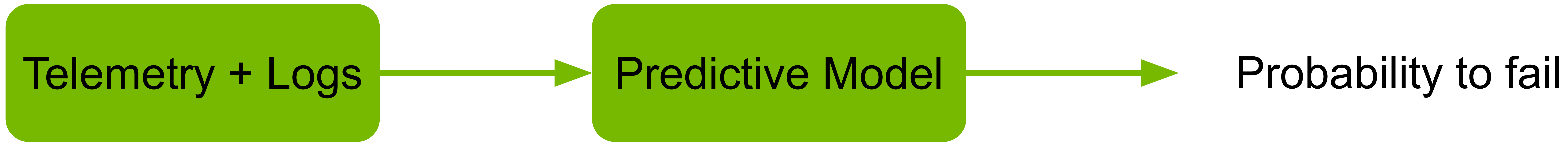


[2304.04661] AI for IT Operations (AIOps) on Cloud Platforms: Reviews, Opportunities and Challenges

Use Case #1 - Predictive Maintenance

Prediction without action is meaningless

- The general concept it to predict a failure before it happens
- The common KPI are precision, recall, prediction time before failure etc.
- Commonly some “prediction horizon” is predefined, however one can use survival analysis to predict the mean time for a failure.



Telemetry + Logs

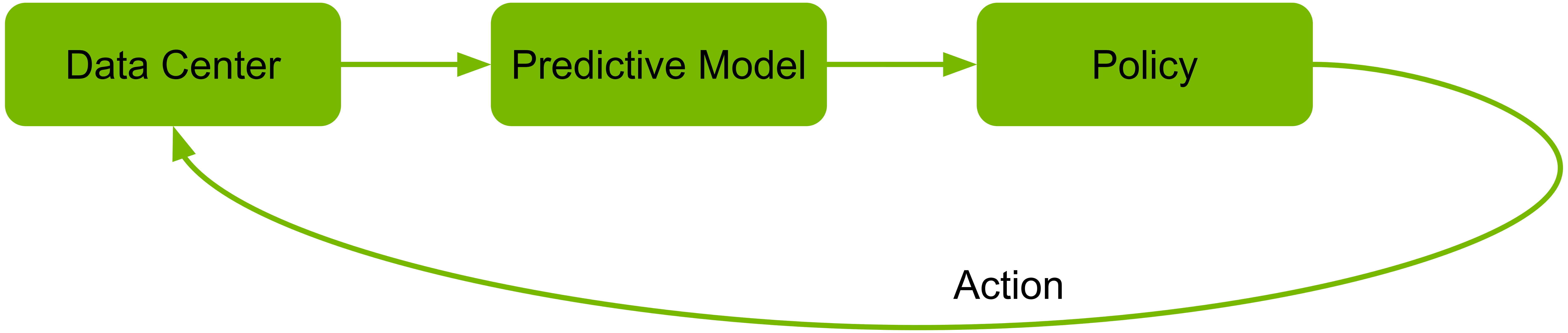
Predictive Model

Probability to fail

Offline Predictive Model Evaluation is not Trivial

One should conduct “what-if” simulation

- What is the overall value of the system?
- False positive and negative are not enough to understand if the system is better than doing nothing
- Backtesting the model + simulating counter factual are needed to be able to answer the question of value
- Simulation should include, for example, the network, compute, failures, schedules etc.

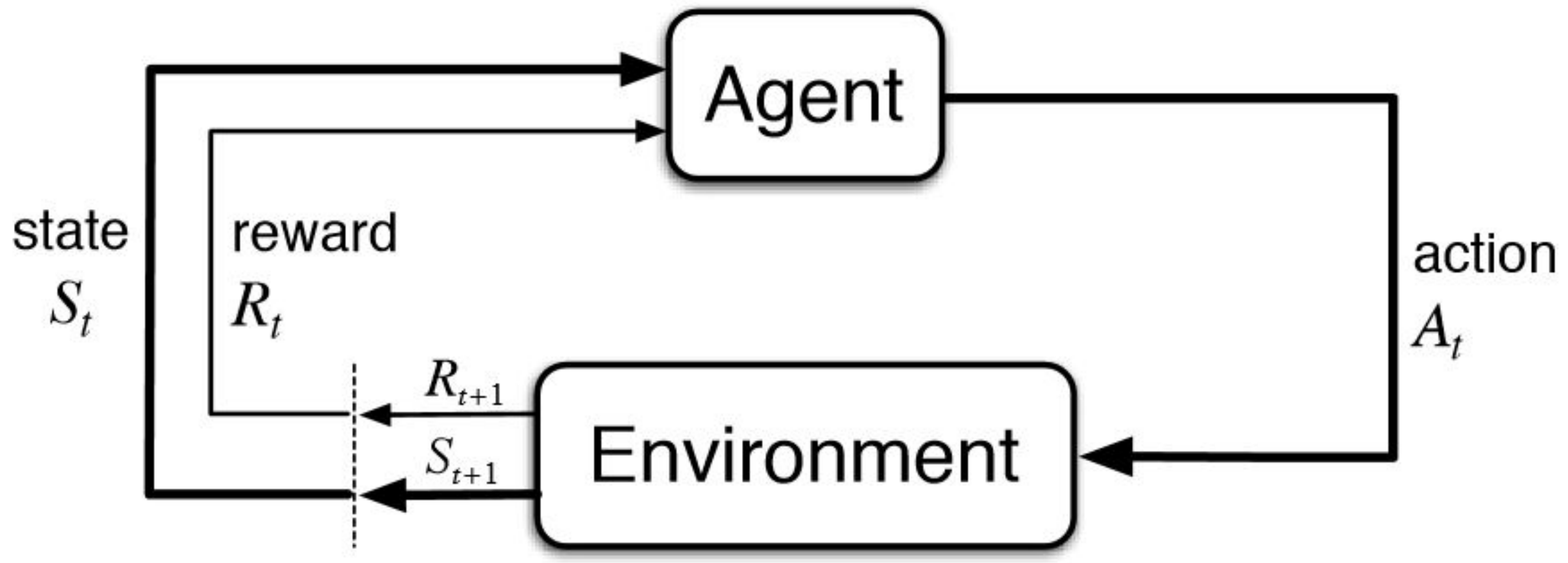


Data Center

Predictive Model

Policy

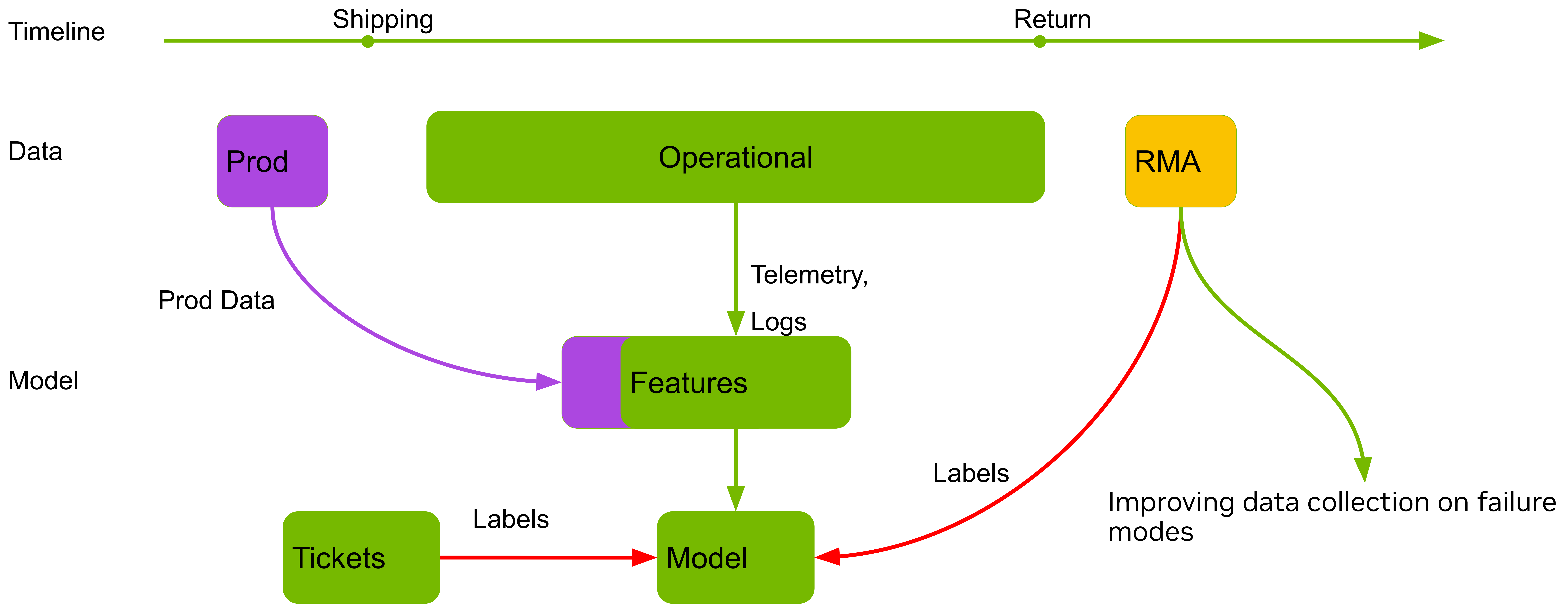
Action



Reinforcement Learning 101

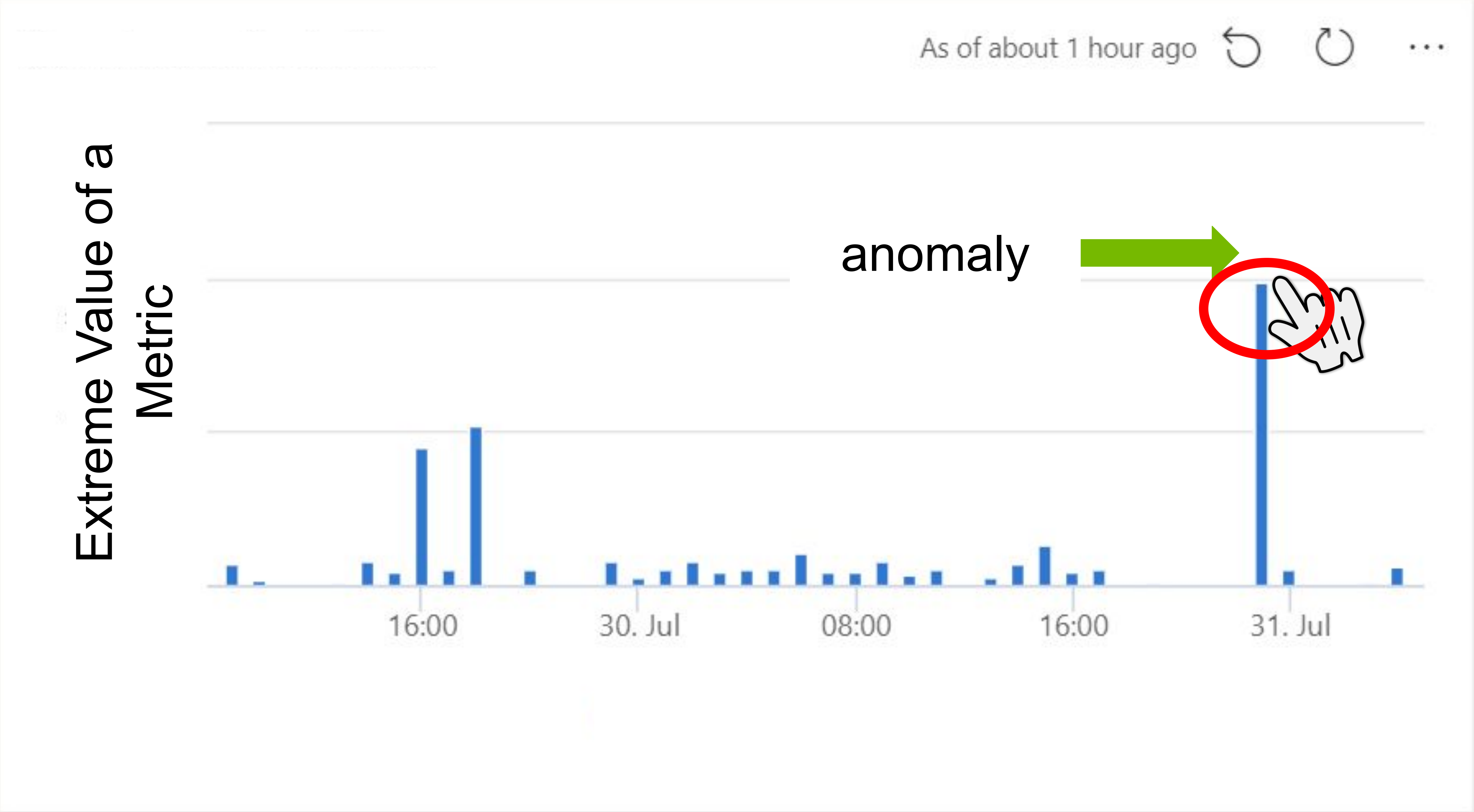
Production & RMA data

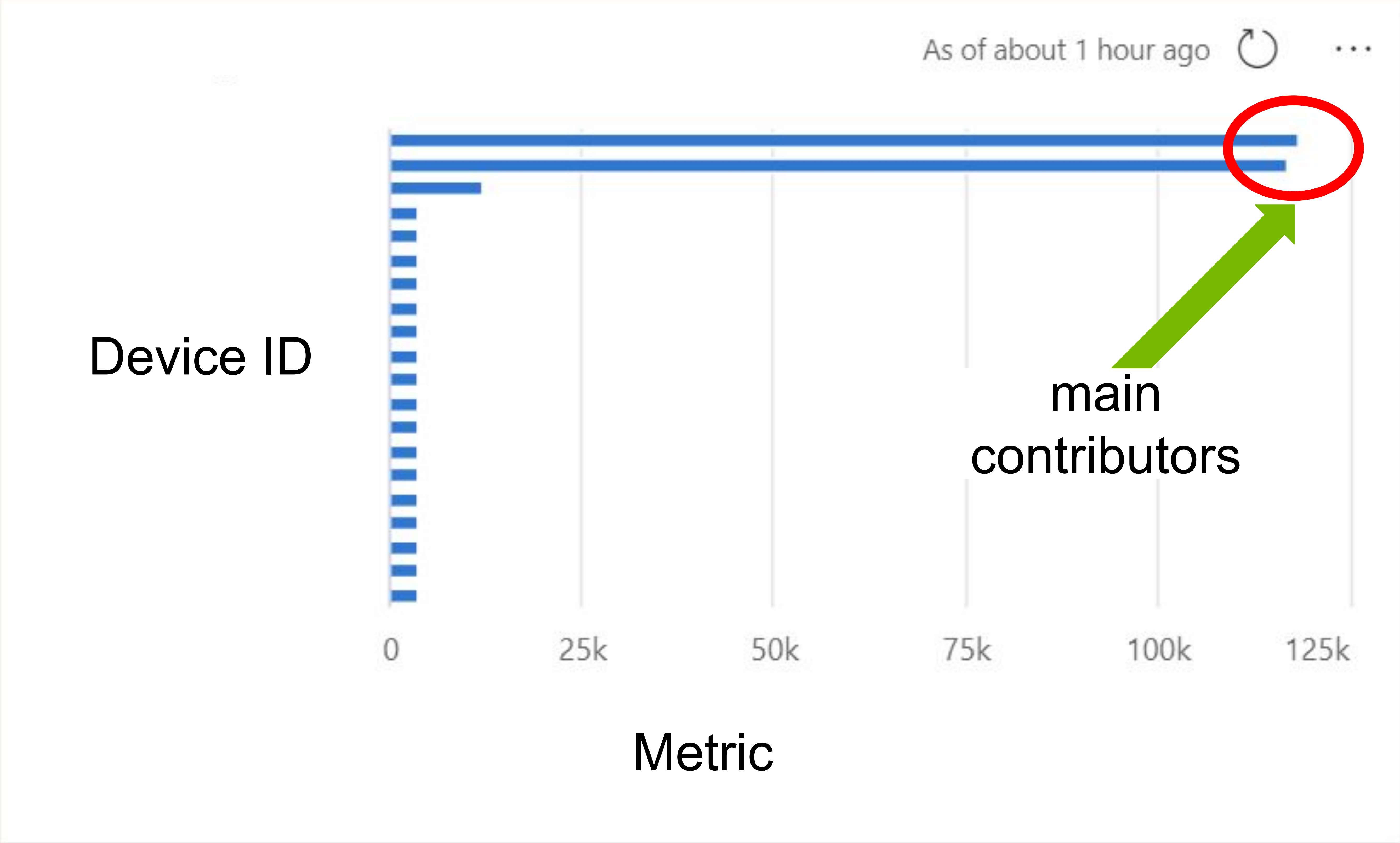
Incorporating Cold Data



Use Case #2 Extreme Value Anomaly Detection

Find and explain worse case behaving devices

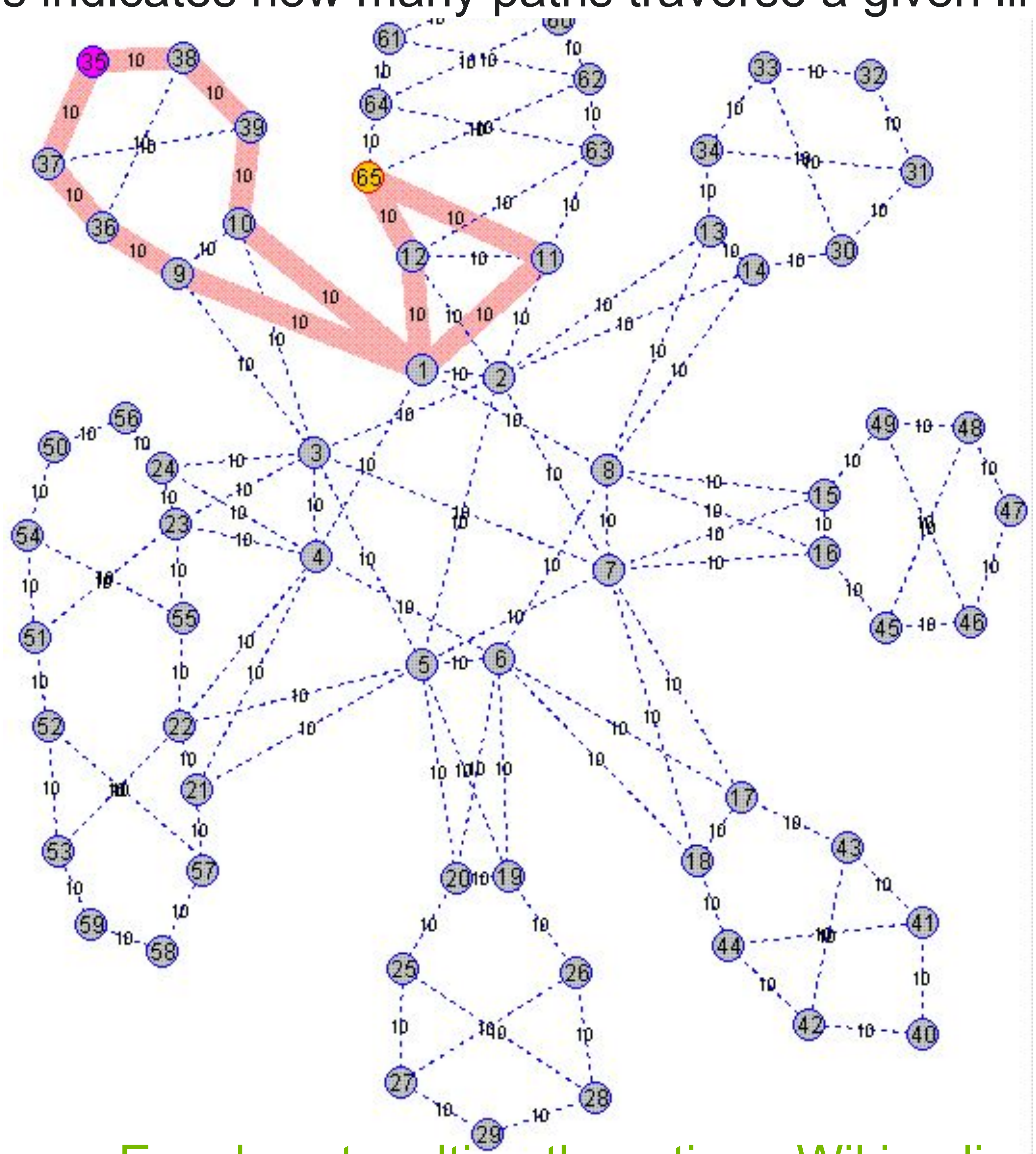




See also [Anomaly Detection in Streams with Extreme Value Theory | Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining]()

Use Case #3 - Root Cause Middleware Errors

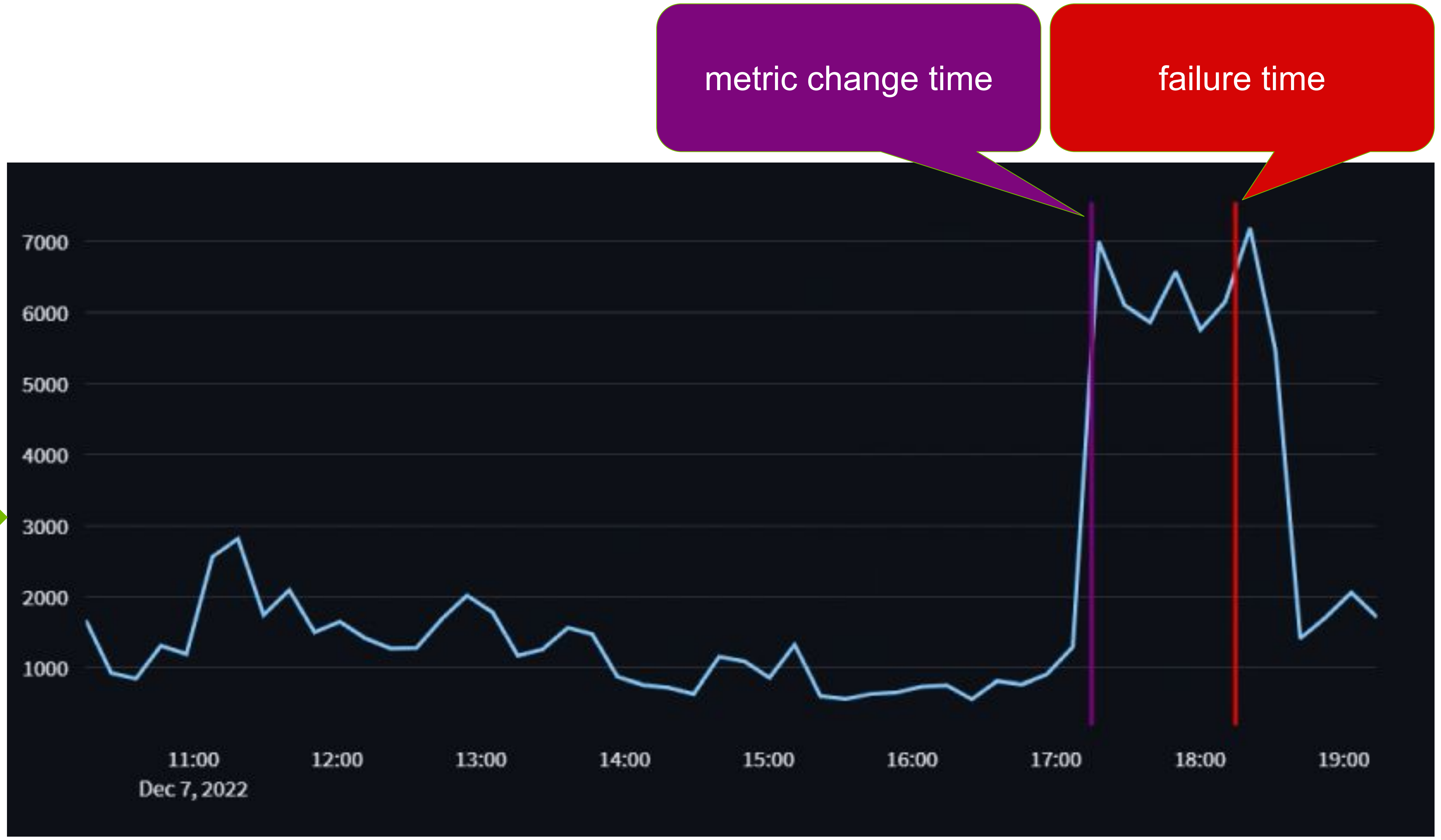
The source is highlighted in purple, the destination in yellow. The purple lines are paths between source and destination and the thickness indicates how many paths traverse a given link.



Source: Equal-cost multi-path routing - Wikipedia

search causality in all shortest path on all devices

Event Correlation



Input: Network Middleware Error

Output: Significant Granger’s causality

Granger causality - Wikipedia

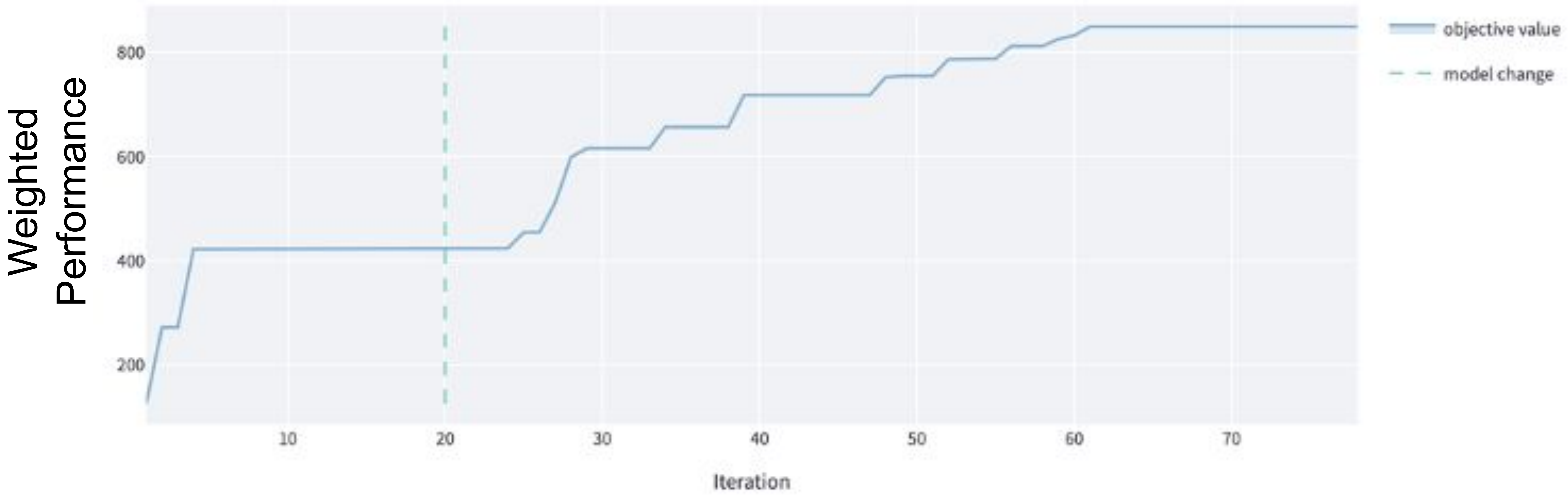
Correlate failure to metrics and logs and conduct statistical tests for (Granger) causality

Use Case #4 - Performance Optimization

Adaptive Routing Parameter Bayesian Optimization

- We use Bayesian Optimization to adapt algorithms parameters according to lab and in-the-wild performance.
- For reference, see Efficient tuning of online systems using Bayesian optimization - Meta Research
- We used the same optimizer on two different simulator as well as in a lab.
- Result reduce bandwidth STD by 85%

Performance over Optimization Iterations



Summary

What's Next

- The main dimensions of an autonomous data centers are:
  - Performance
  - Operation
  - Cyber
- The main theme for AI ops are:
  - Predictive maintenance
  - Anomaly detection
  - Root cause analysis
  - Automatic action (policy learning)
- AIOps is on the rise but getting there will take time
- Dedicated narrower projects are more likely to bring ROI
- Don't over do it
  - Rule #1: Don't be afraid to launch a product without machine learning (Rules of Machine Learning: | Google for Developers)

Questions?

hshteingart@nvidia.com