

DUAL MACHINE LEARNING MODELS FOR HARDWARE QUALITY AND SUPPLY CHAIN MANAGEMENT

TOWARDS AN INTELLIGENT SELF MANAGED PLATFORM

CHRIS CHENG, DISTINGUISHED ENGINEER
STORAGE DIVISION, HEWLETT PACKARD ENTERPRISE

OUTLINE

- Motivation
- IoT deployment model
 - Training
 - Feature engineering
- Real world deployment examples
 - HDD
 - SSD
 - Voltage regulator
- Future intelligent management platform
- Conclusion

WHY PREDICTIVE MAINTENANCE IS IMPORTANT

- 16k GPU (>\$1B system), interrupted every 3hrs, 60+% caused by hardware failures
- Processor over voltage due to EPROM mis-program
 - Over voltage caused damaged in the field

Tech Industry > Artificial Intelligence

Faulty Nvidia H100 GPUs and HBM3 memory caused half of failures during LLama 3 training — one failure every three hours for Meta's 16,384 GPU training cluster

News By Anton Shilov published July 27, 2024


Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	1.4%
SSD	Host	3	0.7%
Power Supply	Host	3	0.7%
Server Chassis	Host	2	0.5%
IO Expansion Board	Host	2	0.5%
Dependency CPU	Dependency	2	0.5%
System Memory	Host	2	0.5%

Predict and remove bad component

Table 5 Root-cause categorization of unexpected interruptions during a 54-day period of LLama 3 405B pre-training. About 78% of unexpected interruptions were attributed to confirmed or suspected hardware issues.

INTEL / TECH / DESKTOPS

There is no fix for Intel's crashing 13th and 14th Gen CPUs – any damage is permanent



/ Here are the answers we got from Intel.

By Sean Hollister, a senior editor and founding member of The Verge who covers gadgets, games, and toys. He spent 12 years editing the likes of CNET, Gizmodo, and Engadget.
Jul 26, 2024, 8:54 AM PDT
161 Comments (161 New)

Quantify demand in future

Home > CPUs

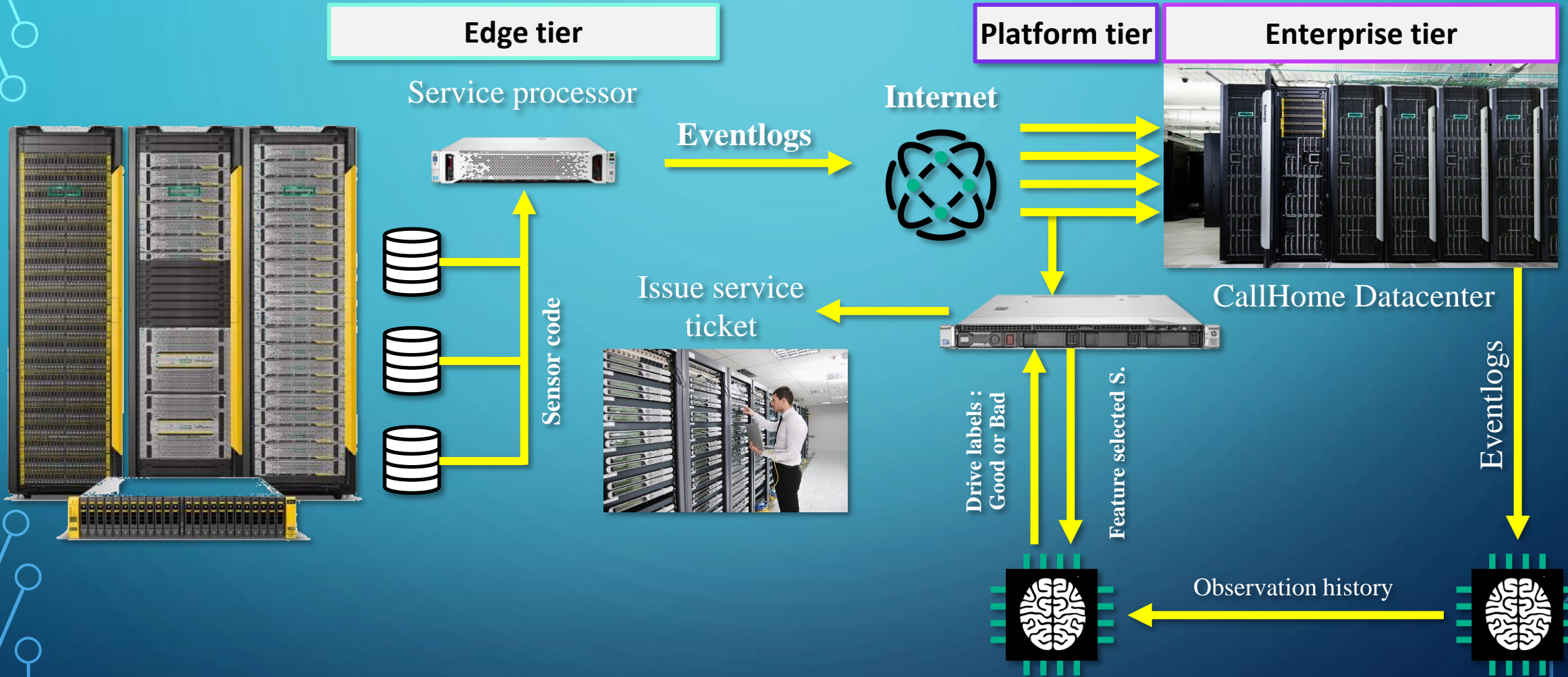
Update: Intel Extends 13th & 14th Gen Core Retail CPU Warranties By 2 Years In Response to Chip Instability Issues

17 Comments
+ Add A Comment

by Ryan Smith on August 6, 2024 7:00 AM EST

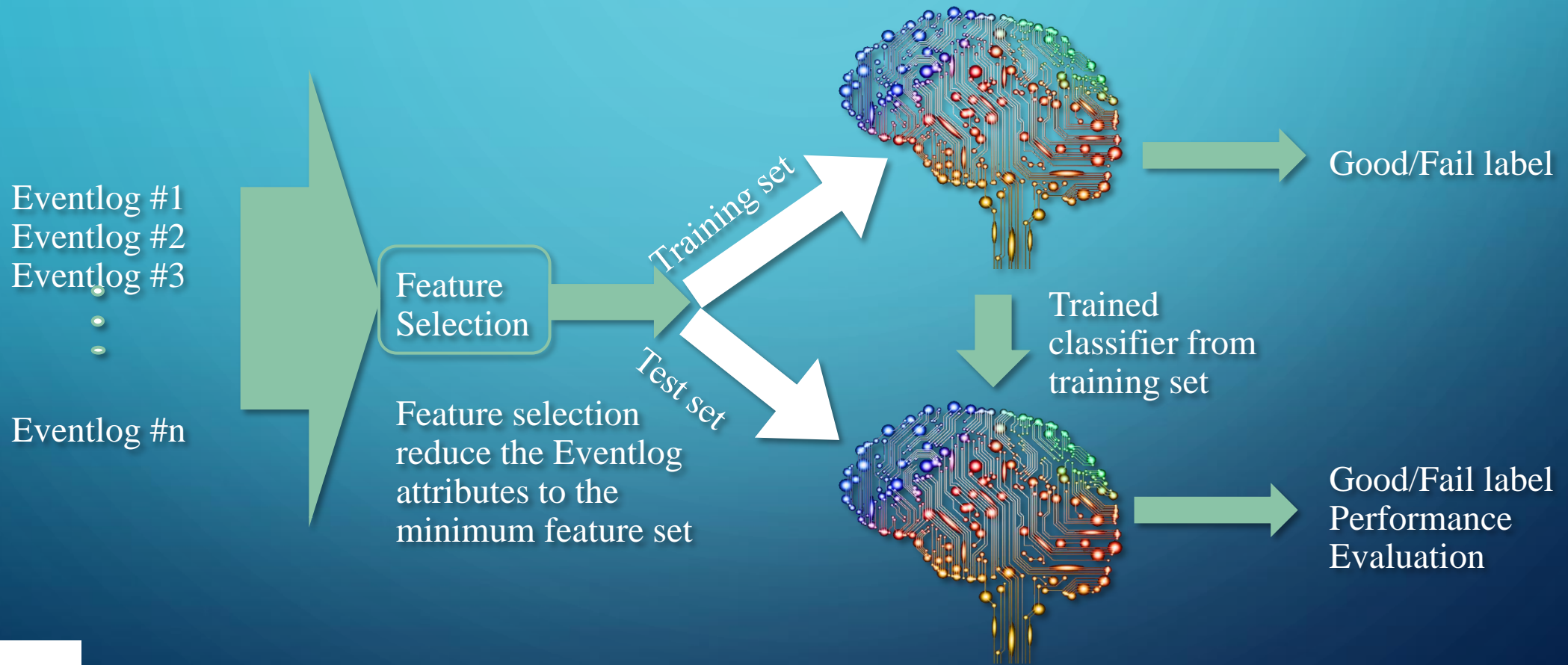
Posted in CPUs Intel 13th Gen Core Raptor Lake 14th Gen Core

IOT HARDWARE PREDICTIVE MAINTENANCE

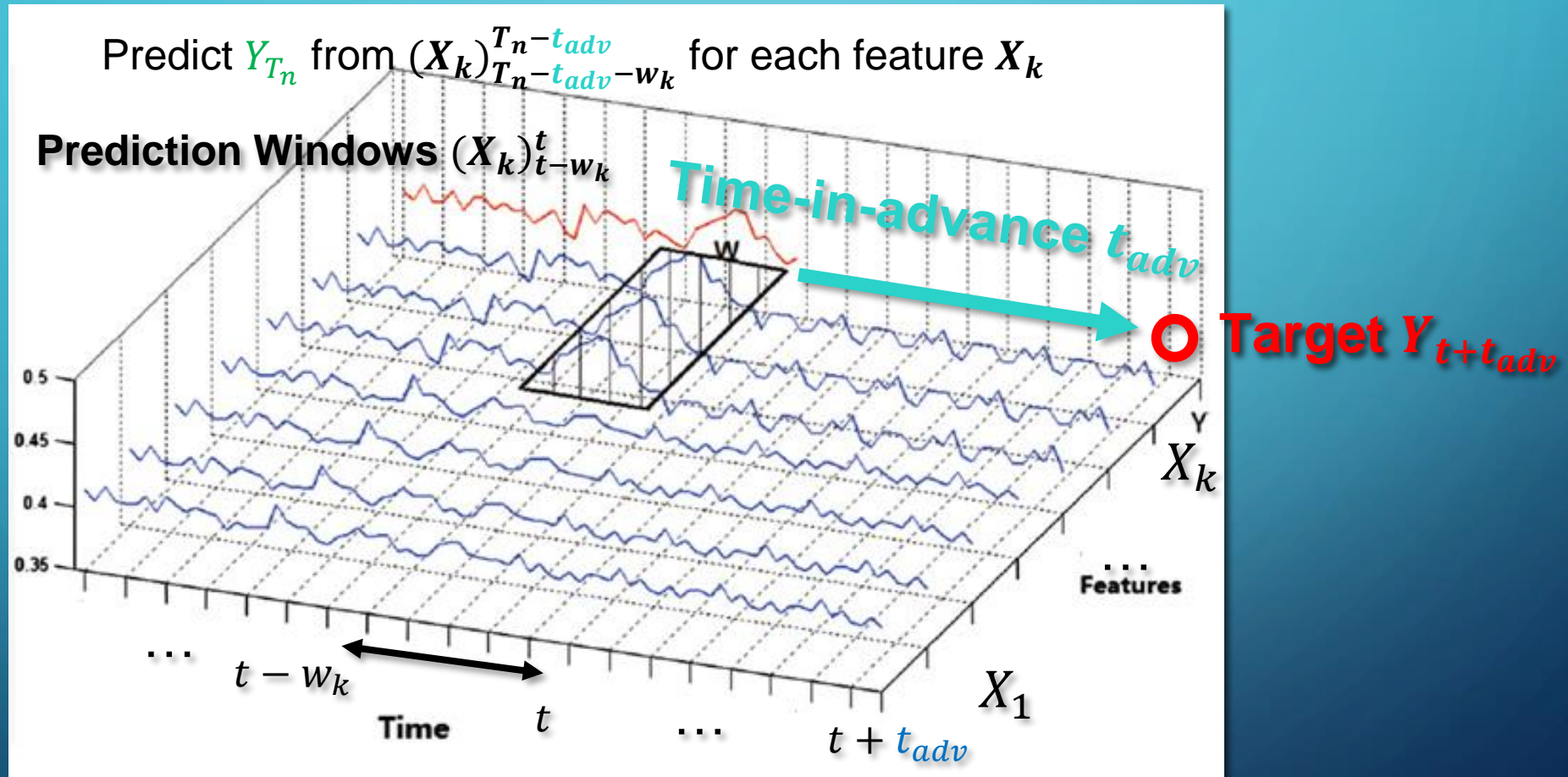


PREDICTION MODEL FUNDAMENTALS

CLASSIFIER TRAINING

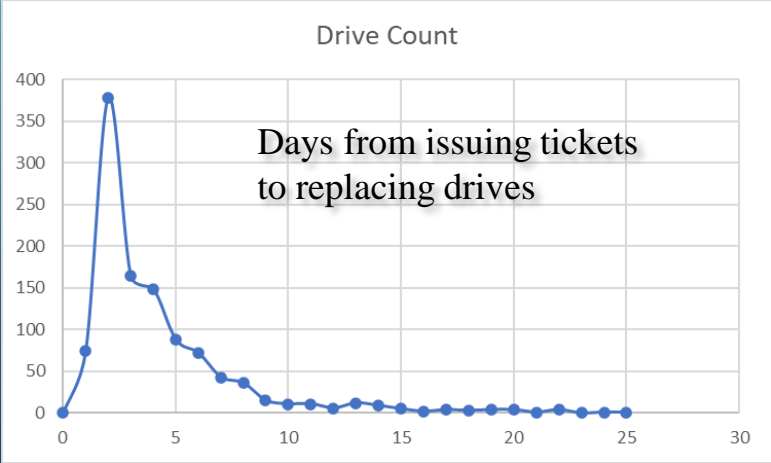
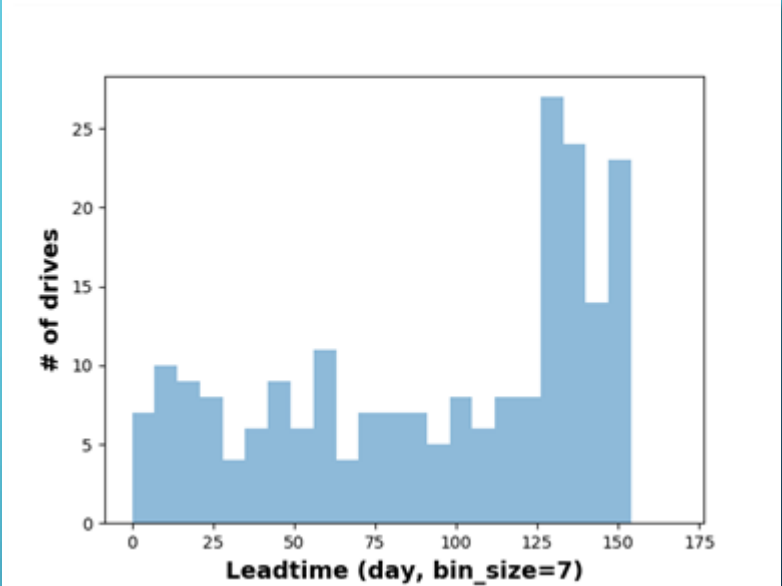
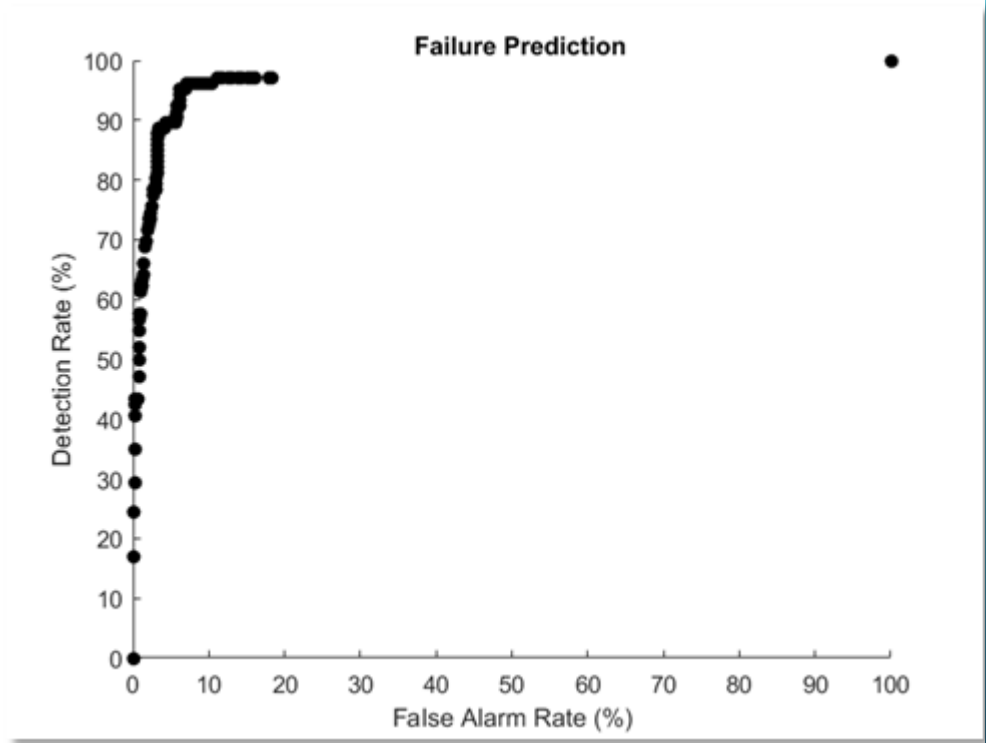


CASUAL INFERENCE FOR FEATURE SELECTION



- 30% features reduction
- 15% accuracy improvement

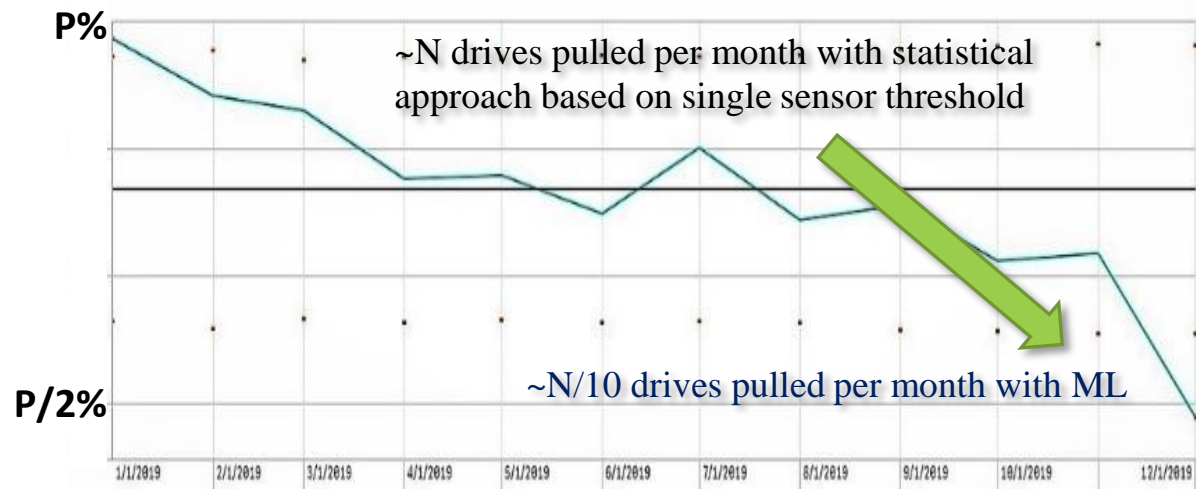
Proactive Removal Results



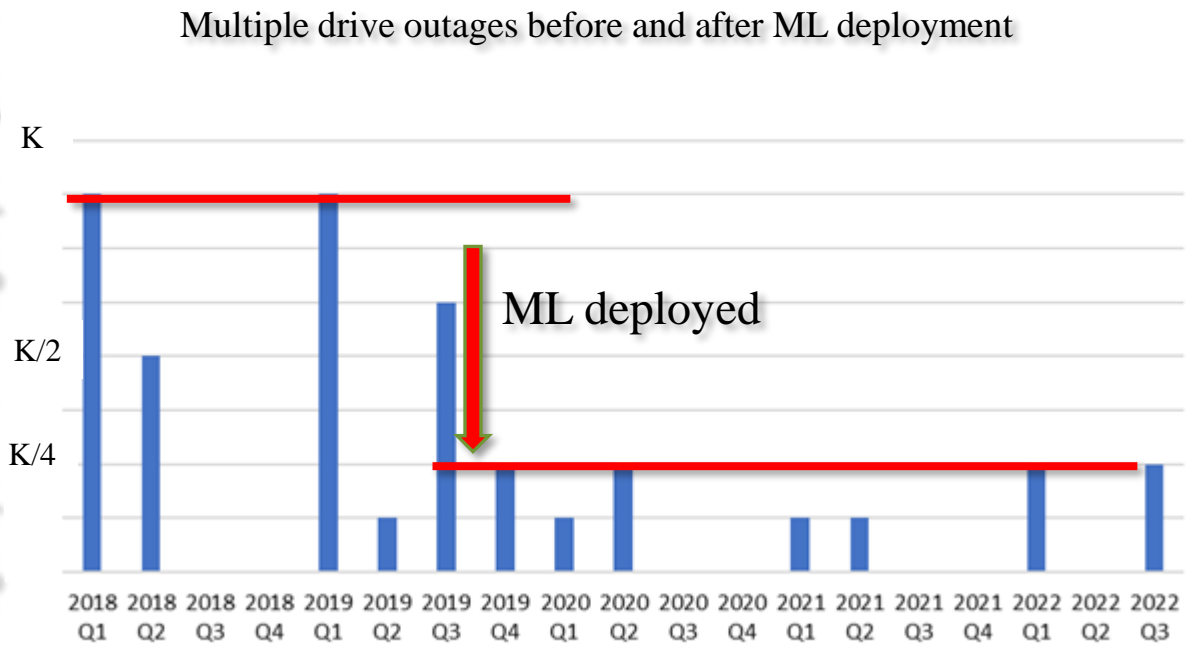
Proactive Removal Results

ACTUAL DEPLOYMENT RESULTS

- Unplanned failure rate dropped from P% to P/2%
- Annual replacement cost savings for company

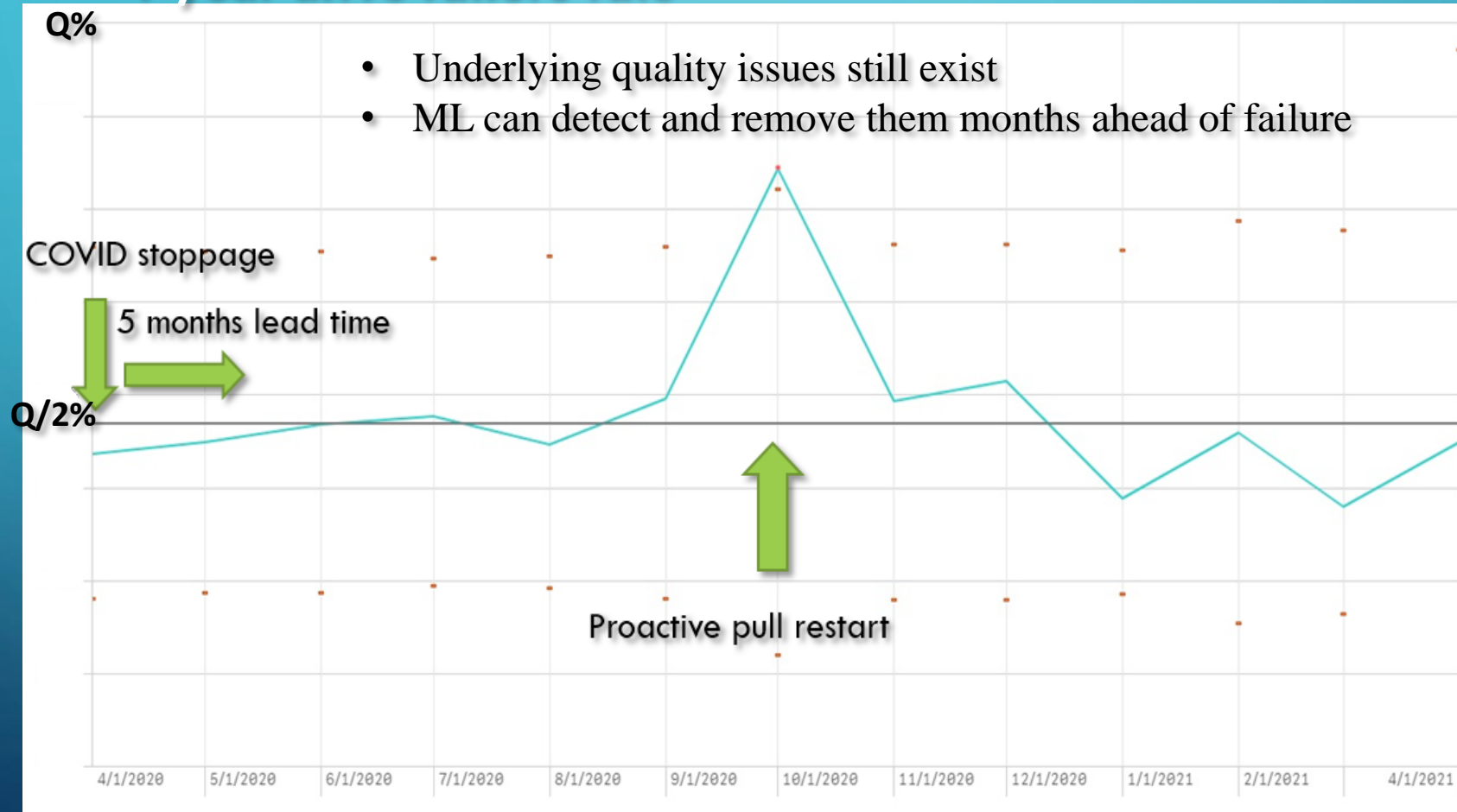


Storage systems outages per quarter

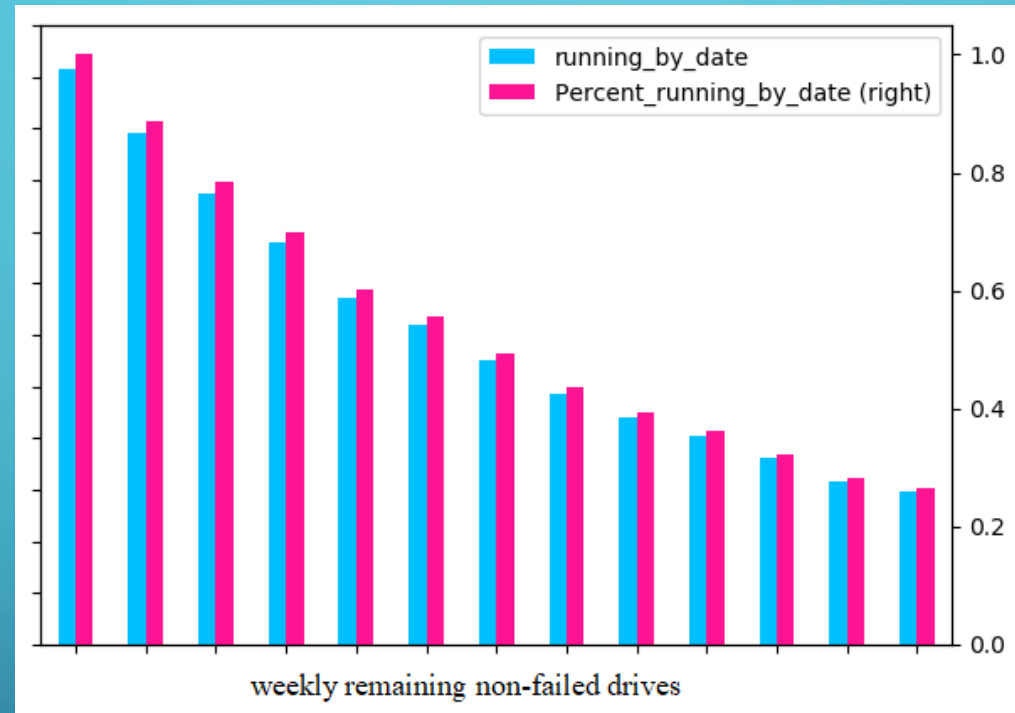


Proactive Removal Results

1 year drive failure rate



FAILURE PREDICTION

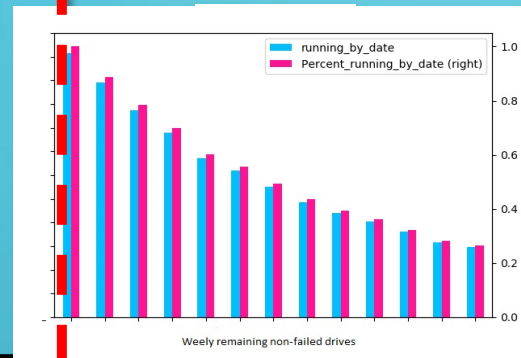


Decrease in the number of functioning drives, with an overall reduction of **~75%**, validating the accuracy of our predictive model.

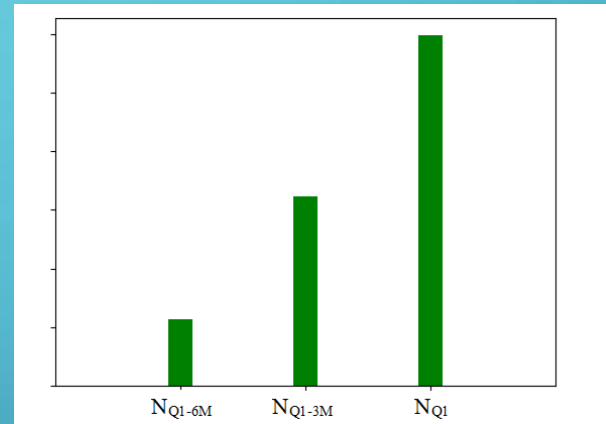
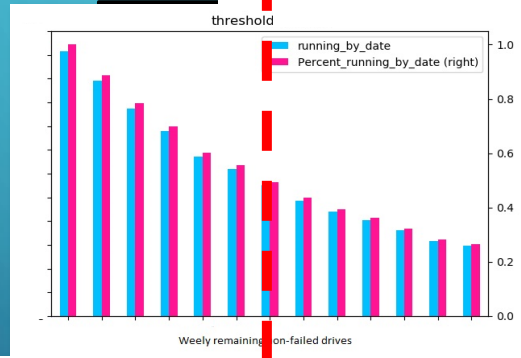
QUANTIFYING QUARTERLY FAILURE

First of quarter prediction

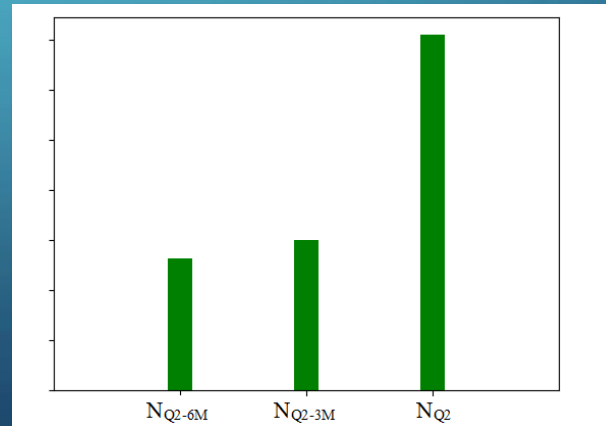
Current quarter



Previous quarter

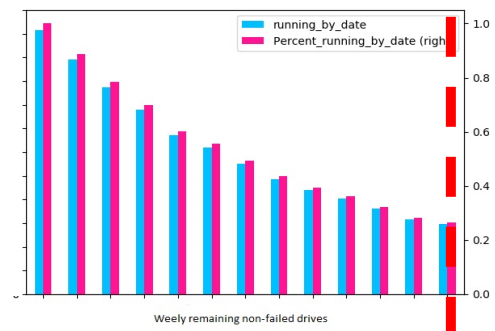


Q1: Number of failed drives estimated with 96% prediction accuracy



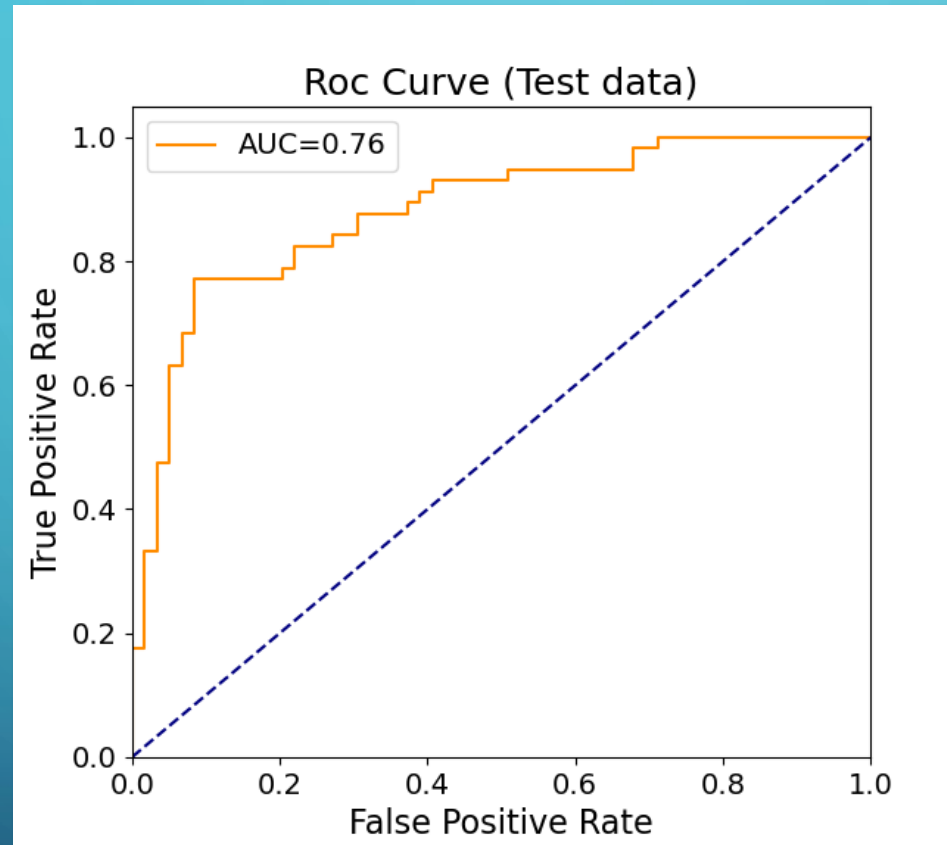
Q2: Number of failed drives estimated with 95% prediction accuracy

2 quarters ago



SSD FAILURE PREDICTION

Detection rate



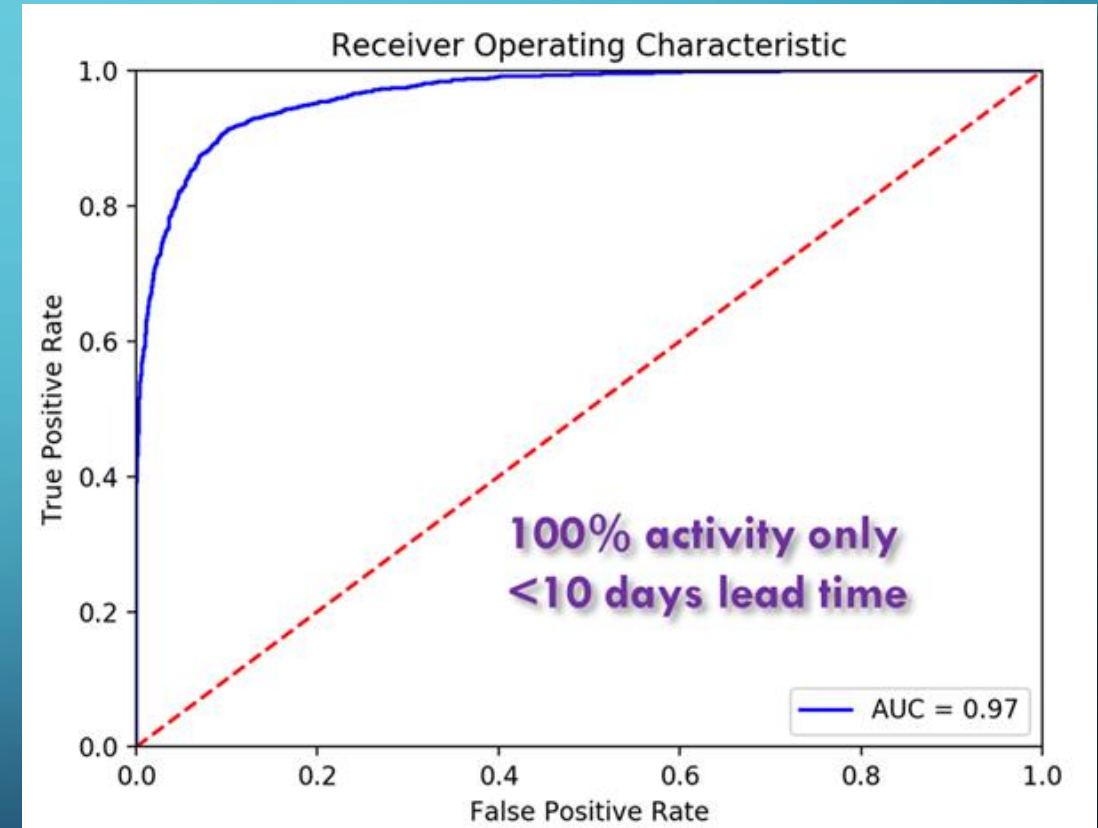
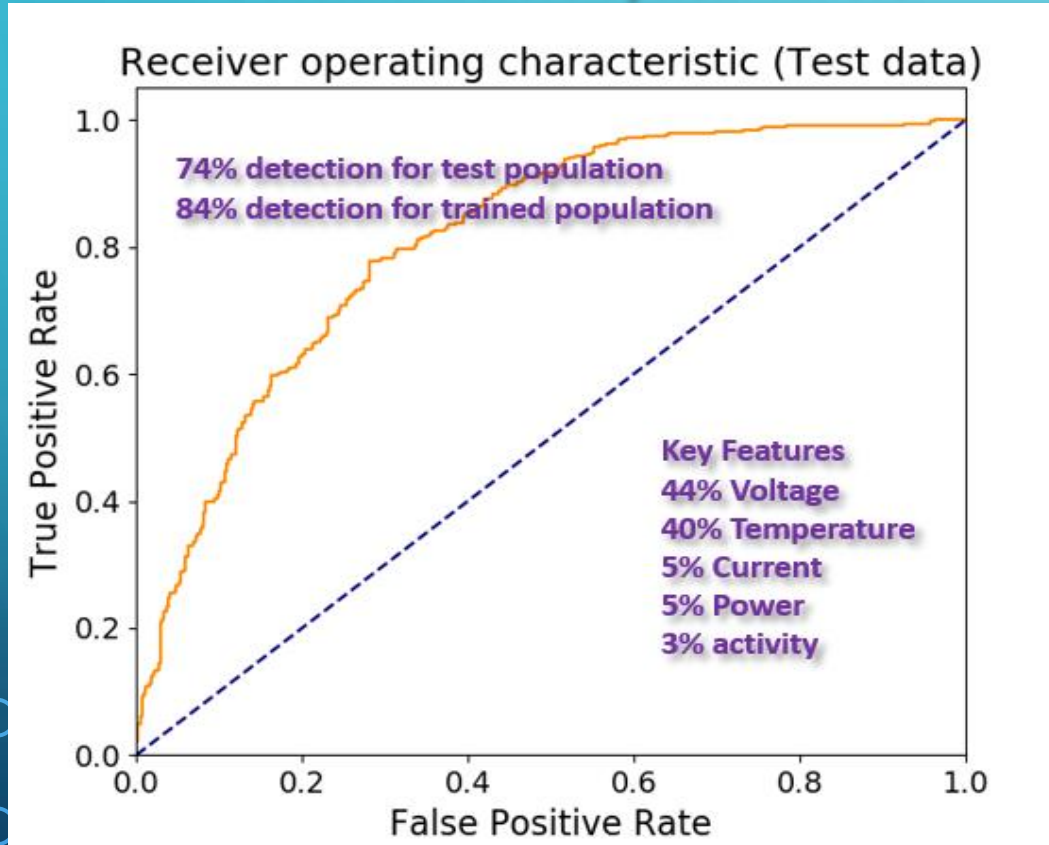
Early stage of deployment

Similar prediction performance like HDD

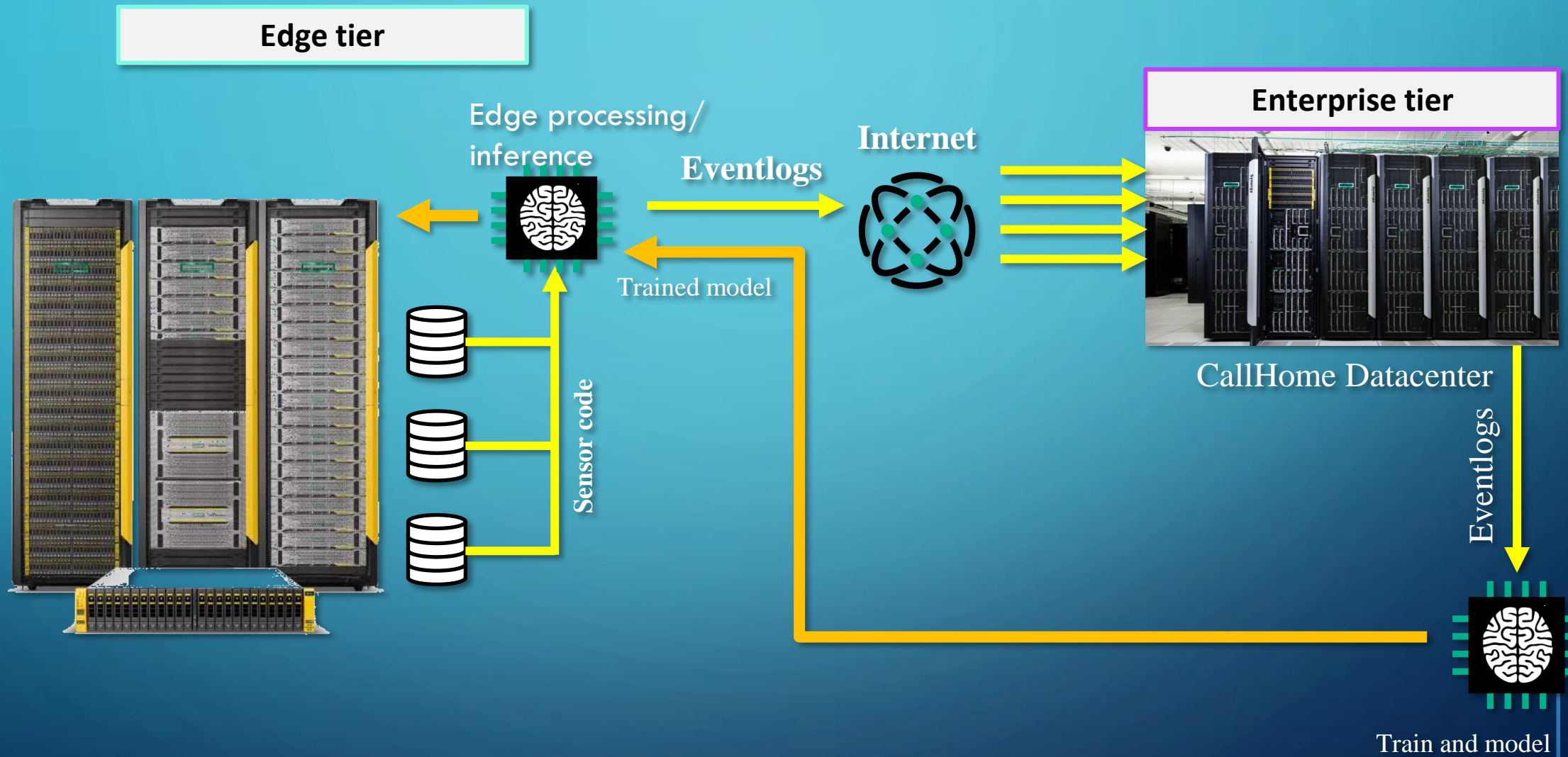
Voltage Regulator Failure Prediction

Long lead time prediction model
> 6 months before system failure

Detection of imminent fail in storage system



NEXT GENERATION INTELLIGENT MANAGEMENT WITH EDGE INFERENCE



CONCLUSION

- We have successfully developed and deployed predictive maintenance and supply chain management models for difference types of hardware
- Predictive maintenance can improve customer experiences by minimizing outages and interruptions at customer sites
- Long term quantity demand prediction also ensure supply chain can manage demands in future
- Future work includes inference at the edge and self manage by the platforms themselves
- Beyond hardware failure prediction , software and security failure prediction