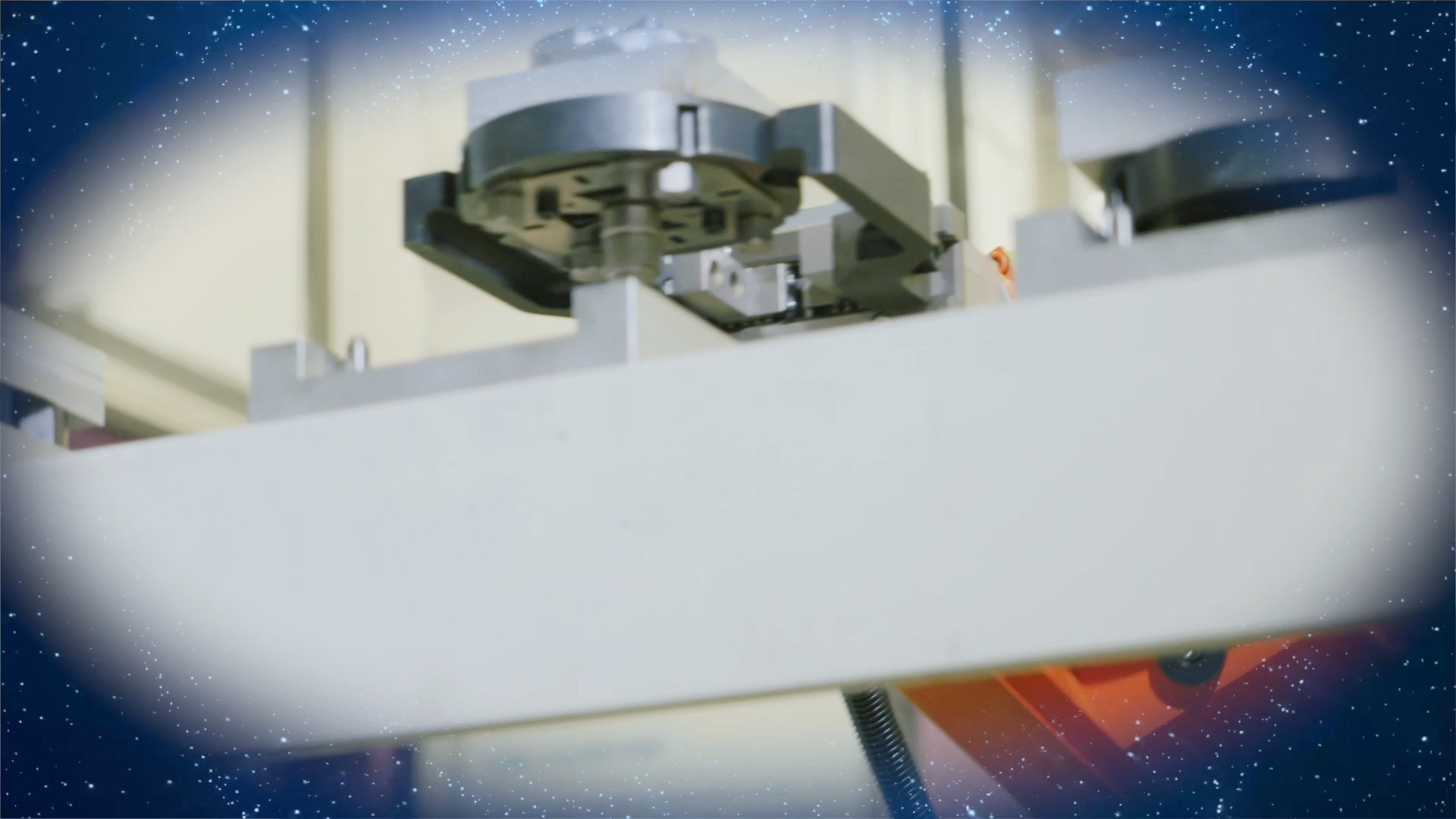


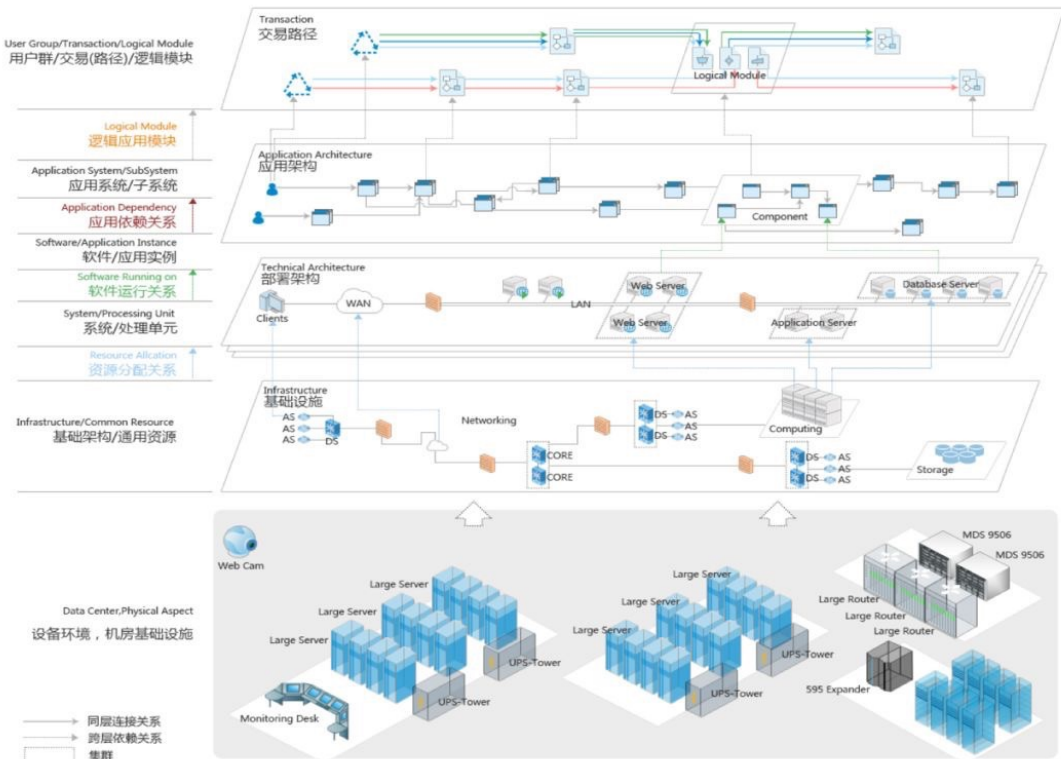
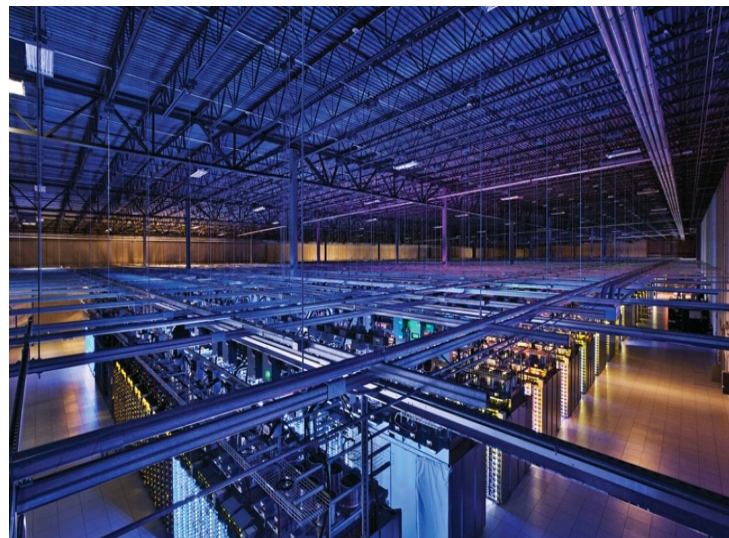
Towards Autonomous IT Operations through Artificial Intelligence

Dan Pei
Tsinghua University





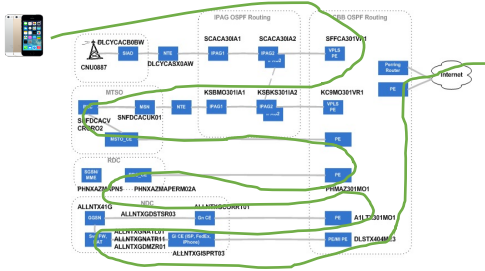
IT Operations is one of the technology foundations of the increasingly digitalized world.



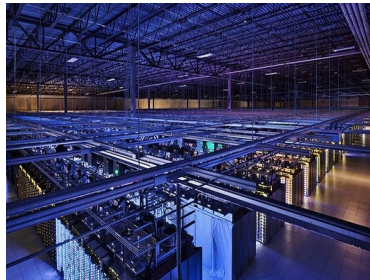
IT Operations

Responsible for ensuring the digitalized businesses and societies run reliably, efficiently and safely, despite the inevitable failures of the imperfect underlying hardware and software.

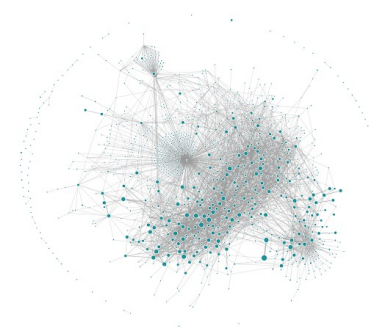
Large & complex access network



Large & complex data center



Large & complex application software



A real case in a global top bank: labor-intensive, stressful, and ineffective

Manual



30 Engineers involved



10:20 large number of transaction failures



Realized there was a failure when customers called

10:45

Failure mitigation time reduced by 90%

Failure localized 11:10

Replayed the data with our ML-based failure discovery and localization algorithms

10:21 automatically detected the failure

10:23 automatically localized the failure



10:45 接网联来电反映银行出现了... 银行系统限制交易流量情况
 具体描述:10:10-10:11,10:19-10:20 银行出现了... 银行系统限制交易流量,影响支付宝... 等,财付通... 等,其他机构... 等,以下截图为详细排查信息
 10:46 联系一线值班... 回替马上处理,原因待查。
 11:10 值班工程师接入电话会议,登录... 1001/... 1002数据库主机检查,发现10:00,10:19数据库出现大量log file sync等待事件,同时该块读写时间也变长,联系存储组协助
 11:22 存储工程师接入电话会议,登录... 1001/... 1002所连接存储,发现存储对相应前端口响应时间变长,10PS减少,排查共享该前口的其他主机未发现问题,排查主机侧存储的链路排查发现... 1002所连接的存储主机... 29#fc12/2... fc12/9... 10:21:31有突发的读IO,联系平台协助处理,经查fc12/2... fc12/9所连接的主机为... 01 联系平台协助处理。
 经查询该时刻... 01数据库中有多对大表排序的操作,导致耗时IO量大,具体请见问题排查信息

Failure discovery: 25mins after the failure happened

Failure localization: 25mins after failure discovery

Some IT Operations Companies

All collect IT Operations data and started to offer AIOps (AI for IT Operations) products

servicenow

Valued at 105 Billion USD

splunk >

Valued at 25 Billion USD



dynatrace

Valued at
11 Billion USD



DATADOG

Valued at
30 Billion USD

sumo logic

Valued at
2.7 Billion USD

“Internet needs an AI-based knowledge plane”
--- Dave Clark in his SIGCOMM 2003 paper.

A Knowledge Plane for the Internet

David D. Clark*, Craig Partridge*, J. Christopher Ramming† and John T.

*M.I.T Lab for Computer Science
200 Technology Square
Cambridge, MA 02139
{ddc,jtw}@lcs.mit.edu

◆BBN Technologies
10 Moulton St
Cambridge, MA 02138
craig@bbn.com

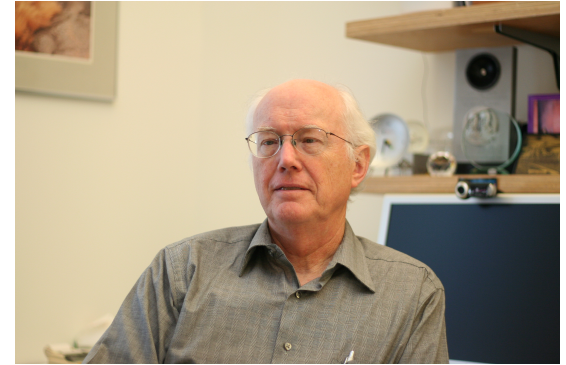
†SRI
333 Rav
Menlo Par
chrisramr

ABSTRACT

We propose a new objective for network research: to build a fundamentally different sort of network that can assemble itself given high level instructions, reassemble itself as requirements change, automatically discover when something goes wrong, and automatically fix a detected problem or explain why it cannot do so.

We further argue that to achieve this goal, it is not sufficient to improve incrementally on the techniques and algorithms we know today. Instead, we propose a new construct, the Knowledge Plane, a pervasive system within the network that builds and maintains high-level models of what the network is supposed to do, in order to provide services and advice to other elements of the network. The knowledge plane is novel in its reliance on the tools of AI and cognitive systems. We argue that cognitive techniques, rather than traditional algorithmic approaches, are best suited to meeting the uncertainties and complexity of our objective.

transparent network with rich end-sy
deeply embedded assumption of
administrative structure are critical stre
users when something fails, and high
much manual configuration, diagnosis a
Both user and operator frustrations aris
design principle of the Internet—the
with intelligence at the edges [1,2].
without knowing what that data is, or
combination of events is keeping dat
edge may recognize that there is a prob
that something is wrong, because the c
be happening. The edge understands
expected behavior is; the core only de
network operator interacts with the core
as per-router configuration of routes ar
for the operator to express, or the netw



From 1981 to 1989, he acted as **chief protocol architect** in the development of the [Internet](#), and chaired [Internet Architecture Board](#)

Industry opinions on AI's role in IT operations

Huawei CEO Ren Zhengfei:



"AI is the most important tool for managing the networks.

一、巨大的存量网络是人工智能最好的舞台

为什么要聚焦GTS、把人工智能的能力在服务领域先做好呢？对于越来越庞大、越来越复杂的网络，人工智能是我们建设和管理网络的最重要的工具，人工智能也要聚焦在服务主航道上，这样发展人工智能就是发展主航道业务，我们要放到这个高度来看。如果人工智能支持GTS把服务做好，五年以后我们自己的问题解决了，我们的人工智能又是世界一流。

首先，是解决我们在全球巨大的网络存量的网络维护、故障诊断与处理的能力的提升。我们在全球网络存量有一万亿美元，而且每年上千亿的增加。容量越来越大，流量越来越快，技术越来越复杂，维护人员的水平要求越来越高，经验要求越来越丰富，越来越没有这样多的人才，人工智能，大有前途。

Jeff Dean Head of AI, Google:



"We can (use AI to) improve everywhere in a system that have tunable parameters or heuristics"

Anywhere We've Punted to a User-Tunable Performance Option!

Many programs have huge numbers of tunable command-line flags, usually not changed from their defaults

```
--eventmanager_threads=16
--bigtable_scheduler_batch_size=8
--mapreduce_merge_memory=134217728
--lexicon_cache_size=1048576
--storage_server_rpc_freelist_size=128
...
```

Anywhere We're Using Heuristics To Make a Decision!

Compilers: instruction scheduling, register allocation, loop nest parallelization strategies, ...

Networking: TCP window size decisions, backoff for retransmits, data compression, ...

Operating systems: process scheduling, buffer cache insertion/replacement, file system prefetching, ...

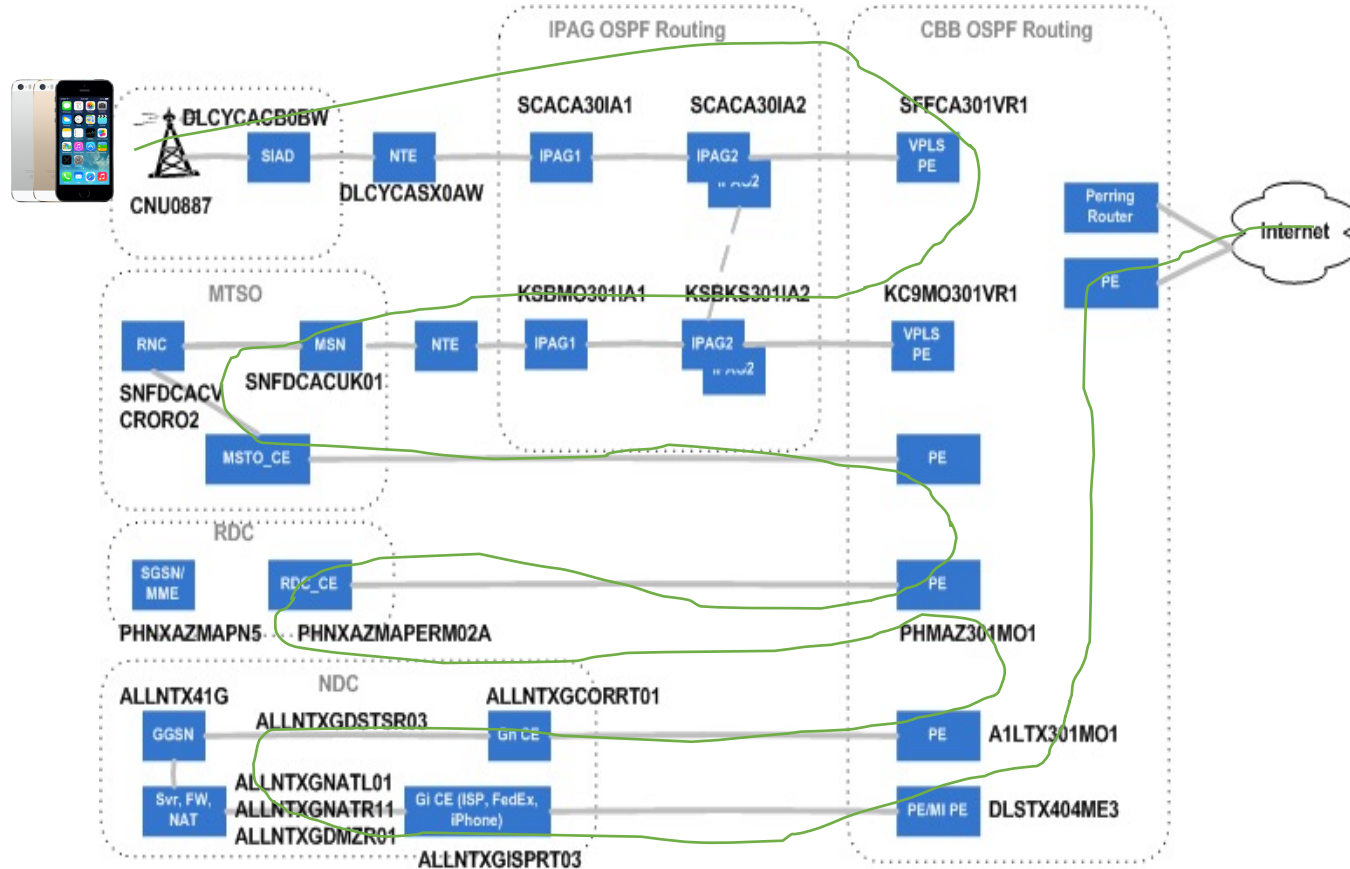
8 **Job scheduling systems:** which tasks/VMs to co-locate on same machine, which tasks to pre-empt, ...

ASIC design: physical circuit layout, test case selection, ...

Outline

- IT Operations (Ops) background
- *Is artificial intelligence necessary for Ops?*
- Case Study Overview
 - Unsupervised Anomaly Detection in Ops
- Lessons Learned

Complex Edge Networks

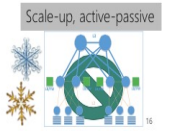
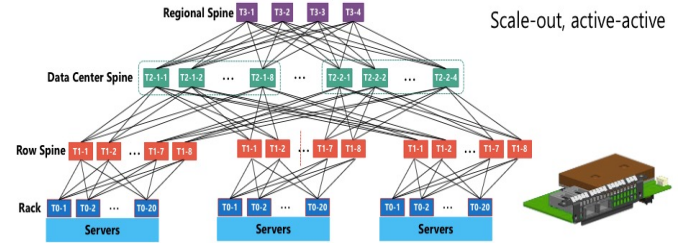


Complex and Evolving Data Center Hardwares

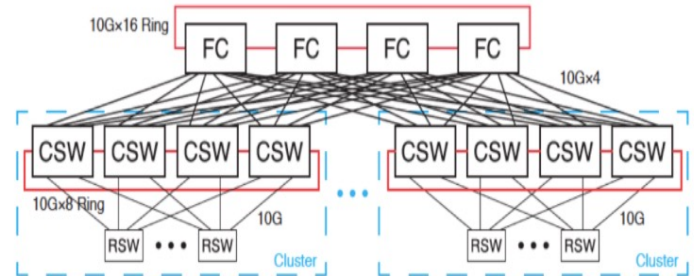
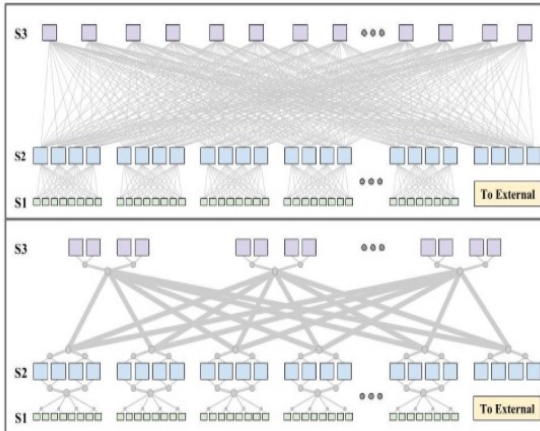
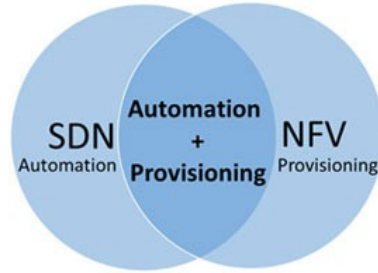
10s of thousands of servers



Frequent topology changes

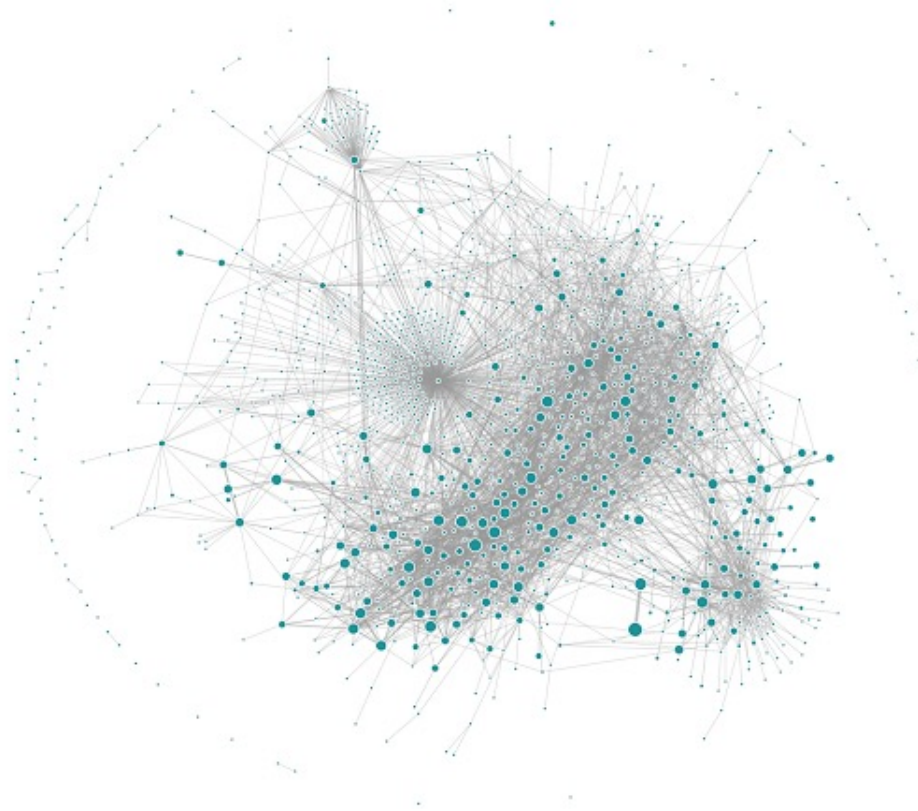


Outcome of >10 years of history, with major revisions every six months



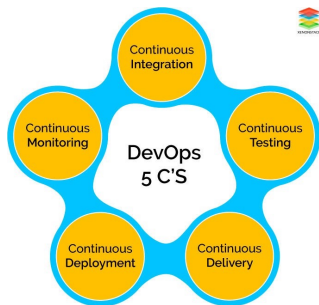
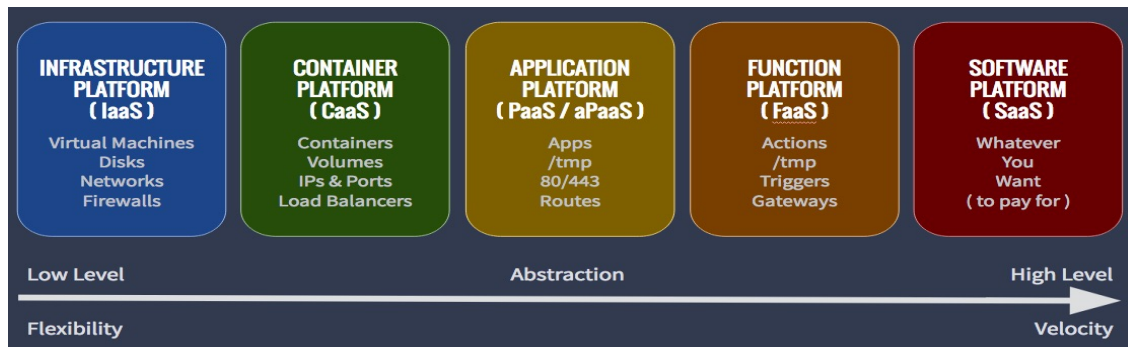
Complex Software Module Dependences

Application dependency at Uber in 2018



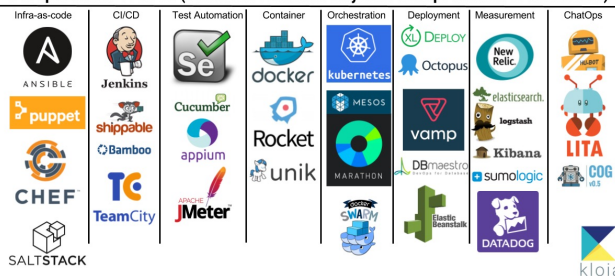
Evolving Techniques Enable Frequent Software Changes, one major cause of failures

10s of thousands software/config changes per day in a large company



DevOps

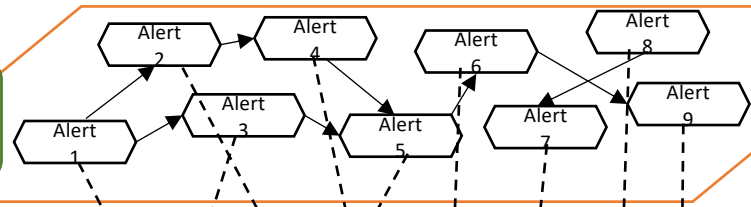
DevOps Enabler Tools v2 (Caution!!! : Consider only after DevOps mindset is established)



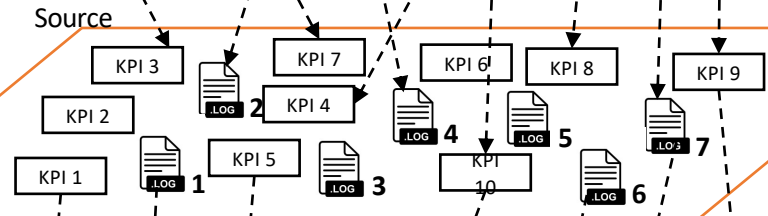
Continuous Integration/Continuous Delivery

Large-scale, complex, cross-layer, dynamic system's digitalized running status → monitoring data

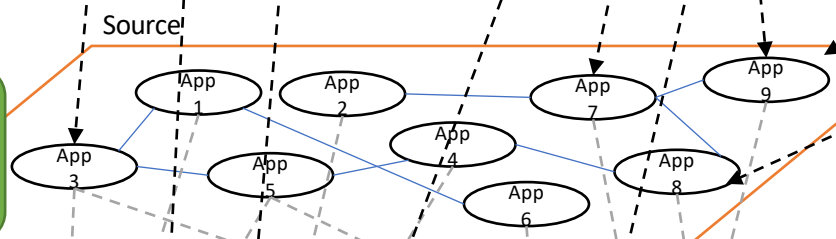
Anomaly Propagation Graph



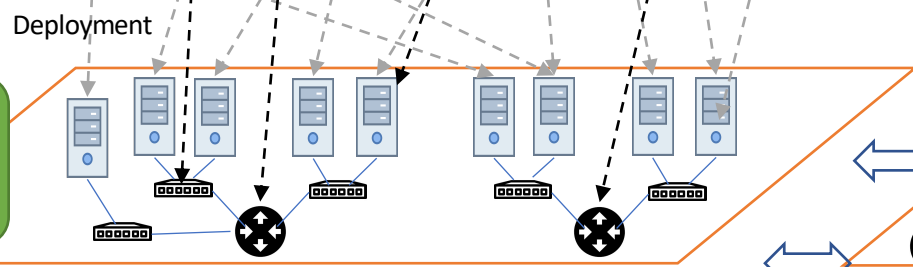
Metrics and Logs



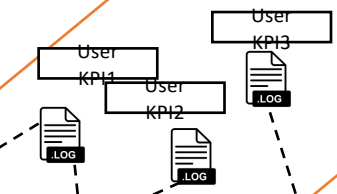
Application Dependency



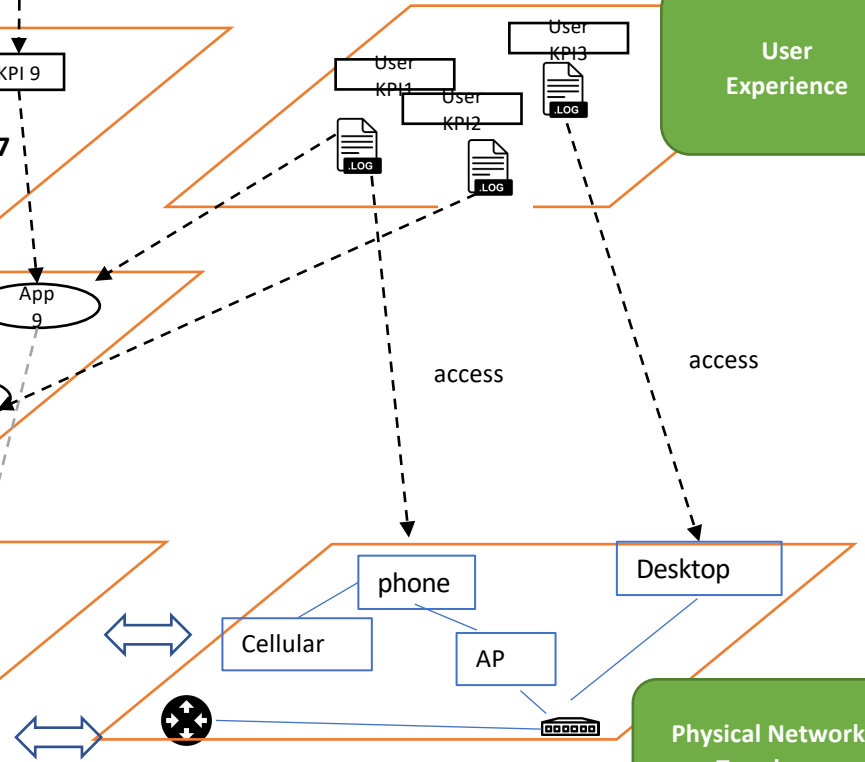
Physical Network Topology



User Experience

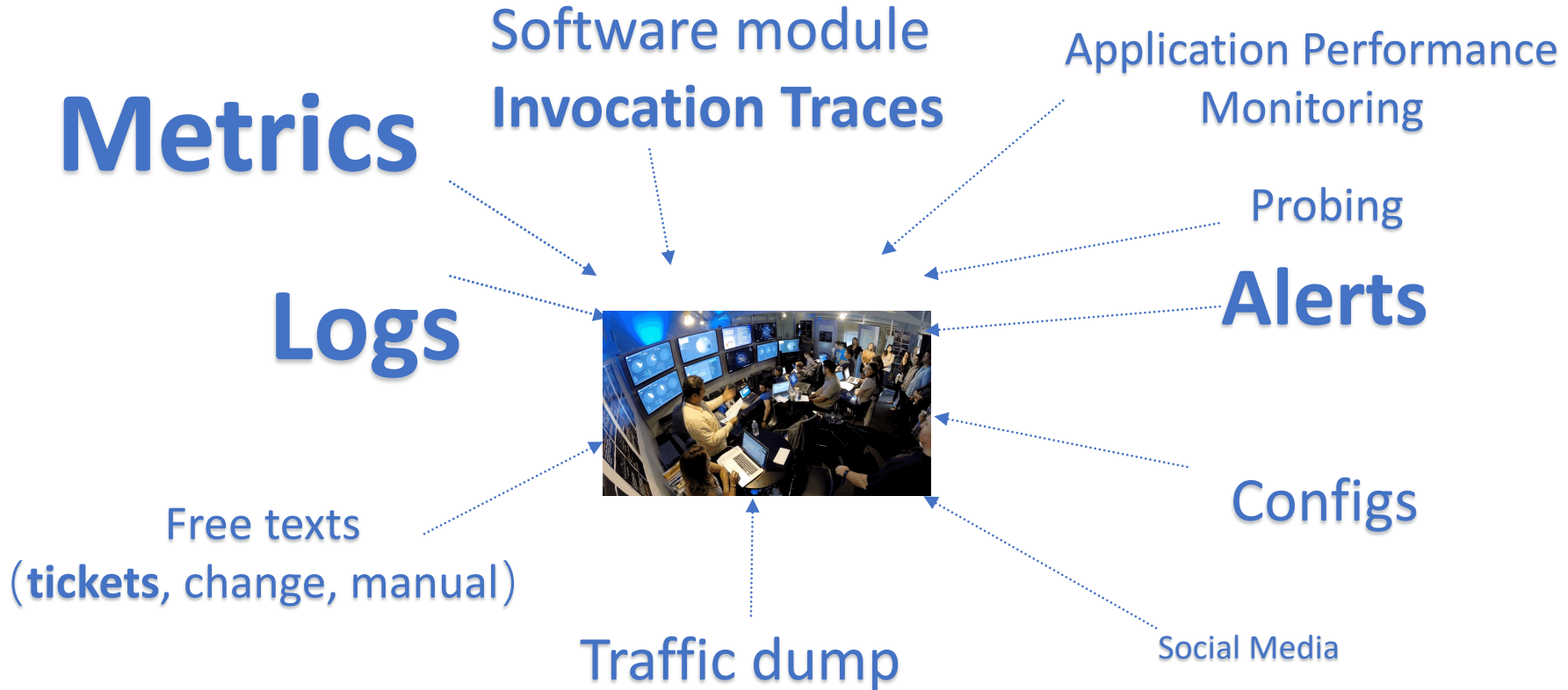


Physical Network Topology



TeraBytes of Ops data per day overwhelm Ops engineers

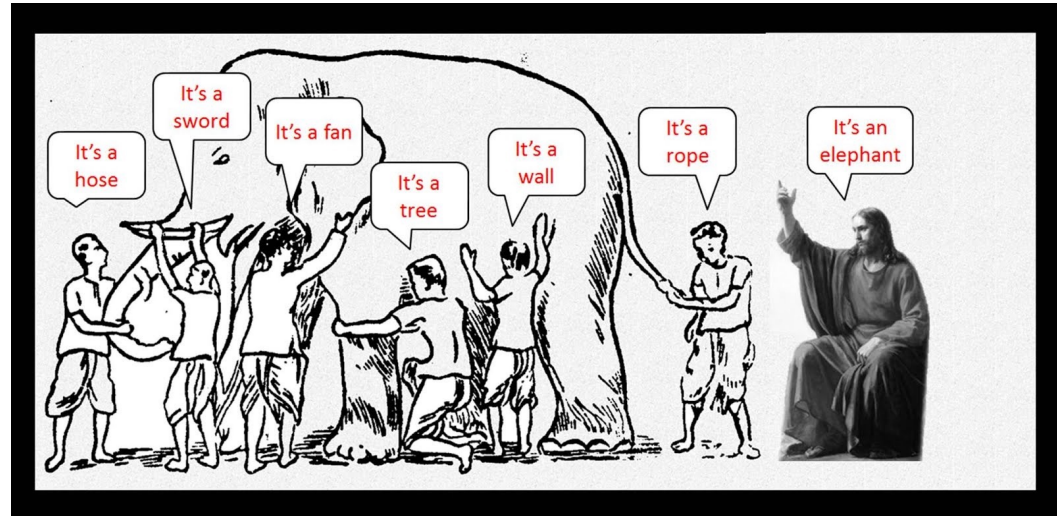
Each offers some clues, but due to complexity and volume, each is hard to manually analyze, let alone collectively analyze all data sources.



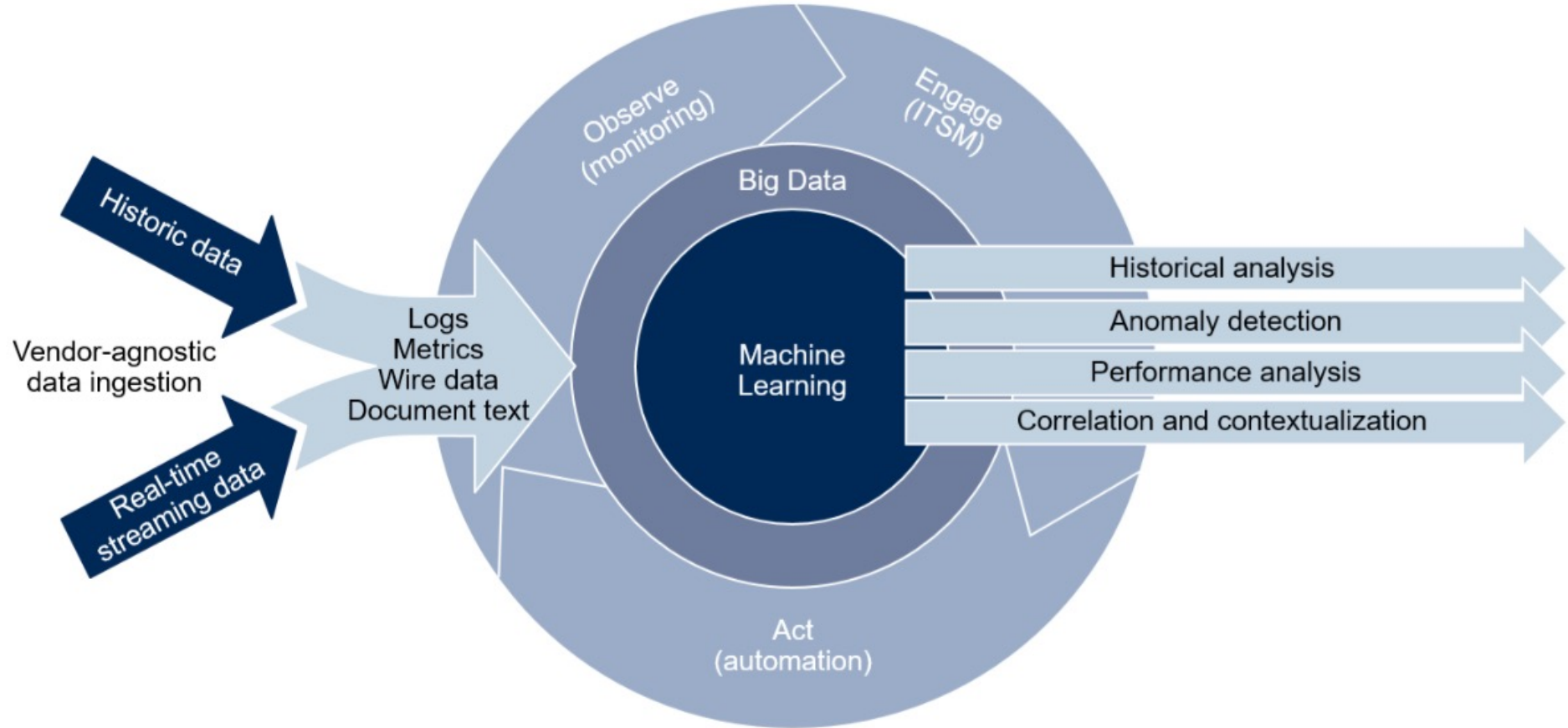
We have no choice but relying on Artificial Intelligence to **extract useful signals** out of the Big Ops Data which have **every low signal-to-noise ratio**.

- Volume
- Velocity
- Variety
- Value

We have no choice but relying on Artificial Intelligence to **incorporate (expert or mined) knowledge (topology, call graph, causal relationship) to correlate signals**.



AIOps Platform Enabling Continuous ITOM



Towards Autonomous IT Operations



Manual and few data

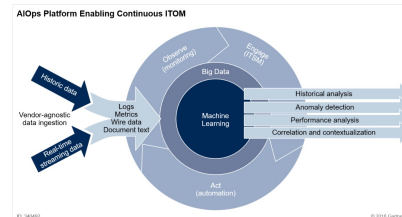


Lots of data but manual decision

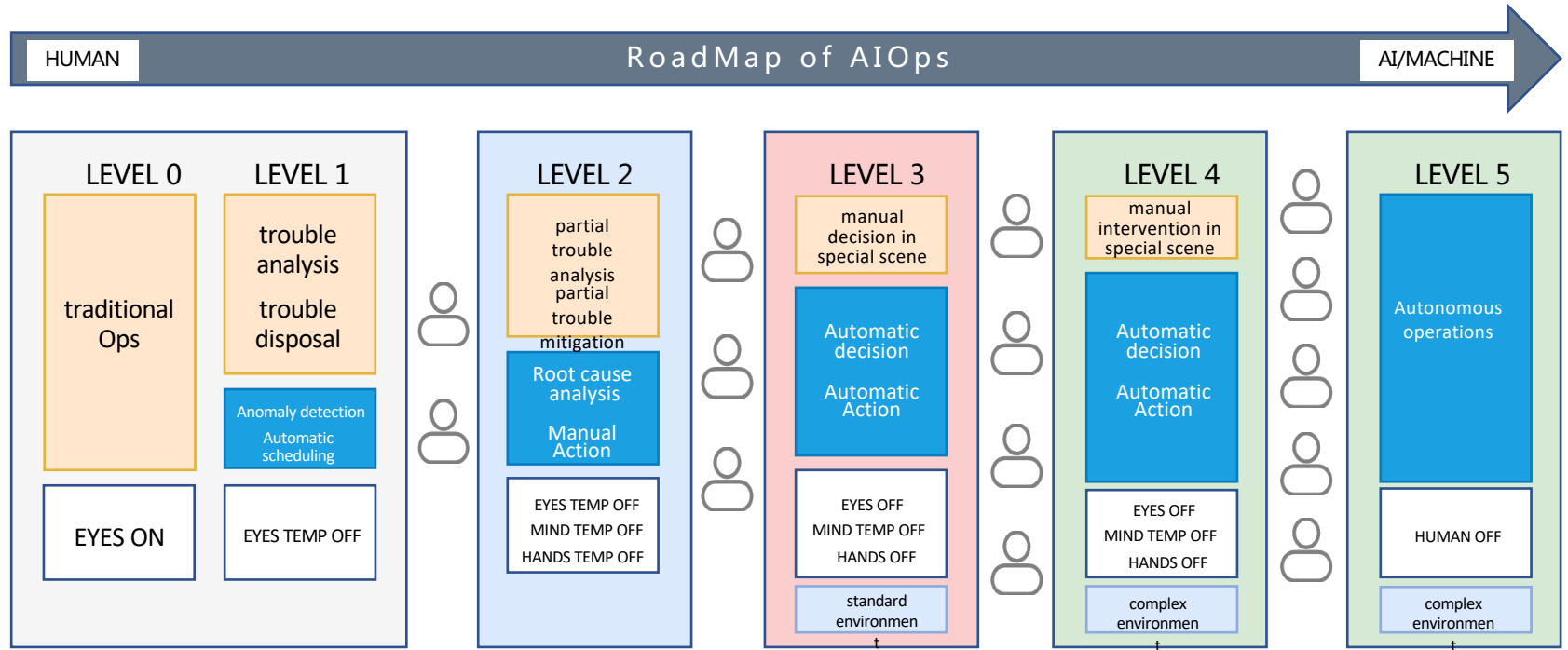


Autonomous

Spaceship Avalon: 5000 passengers and 258 crew members in hibernation. Flying towards Planet Homestead II, 120-year trip.



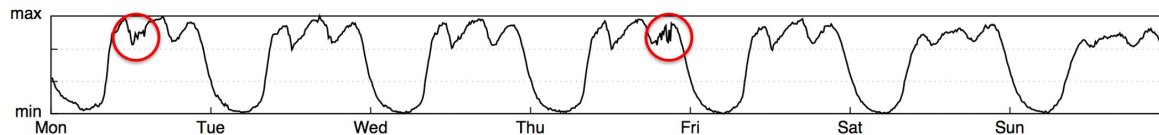
Levels of AIOps



Outline

- IT Operations (Ops) background
- Is artificial intelligence necessary for Ops?
- Case Study
 - Unsupervised Anomaly Detection in Ops
 - *Time series anomaly detection (IMC 2015, WWW 2018, IWQoS 2019, INFOCOM 2019a, INFOCOM2019b, ISSRE 2018, IPCCC 2018a, IPCCC 2018b, TSNM 2019, KDD2019, INFOCOM2021)*
 - Trace anomaly detection (ISSRE 2020)
 - Zero-day attack detection (INFOCOM2020a)

- Lessons Learned



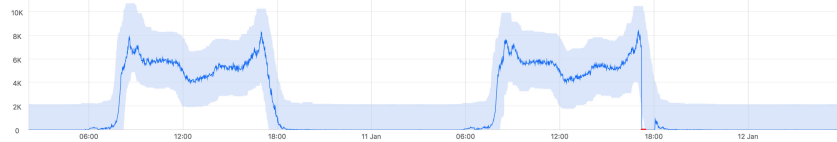
All Case Studies Are From Joint Work with Industry Collaborators



Diverse Metrics and Their Diverse Anomalies

Time series algorithms are needed to parse and make sense of metrics data

(1) Seasonal metrics



(2) Periodicity shift



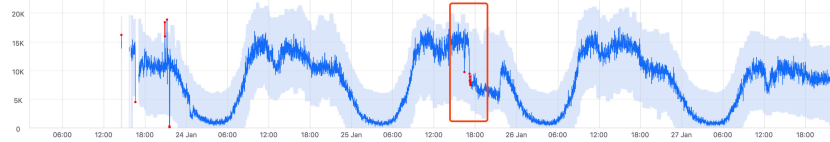
(3) Adapt to holidays



(4) Identify variable metrics and obtain extreme threshold



(5) Detect too rapid a change



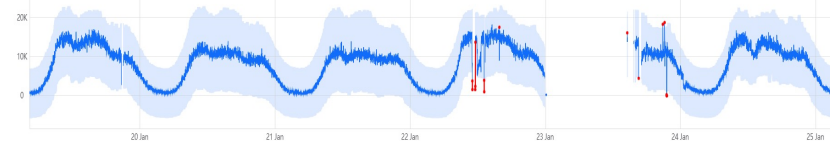
(6) Detect the lack of seasonality.



(7) Adapt to trend change



(8) Robust against data loss or interruption



Donut: supervised- \rightarrow unsupervised: smooth KPIs

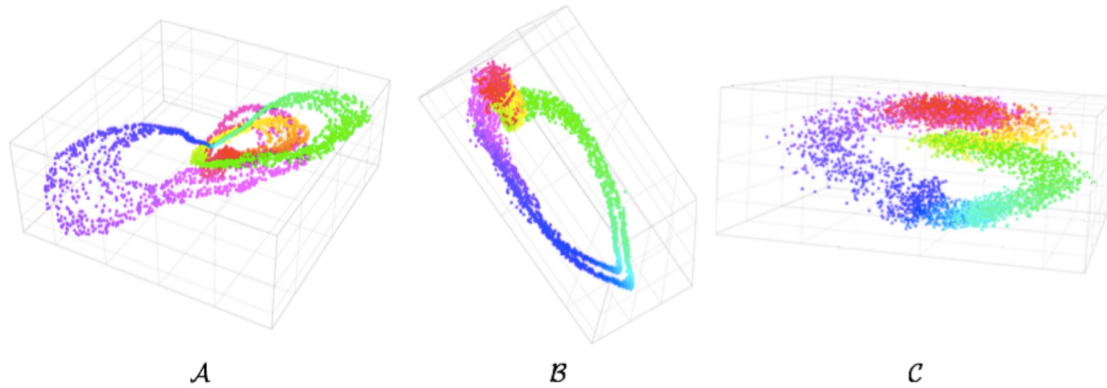
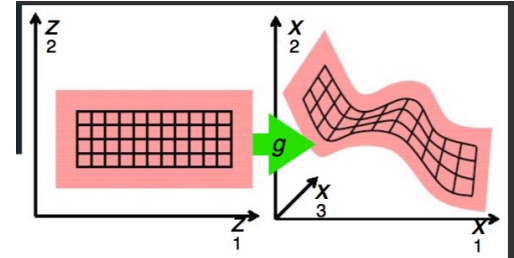
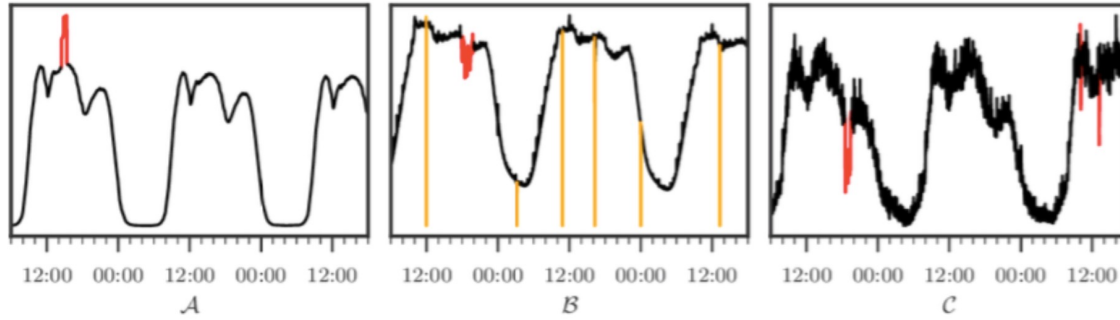
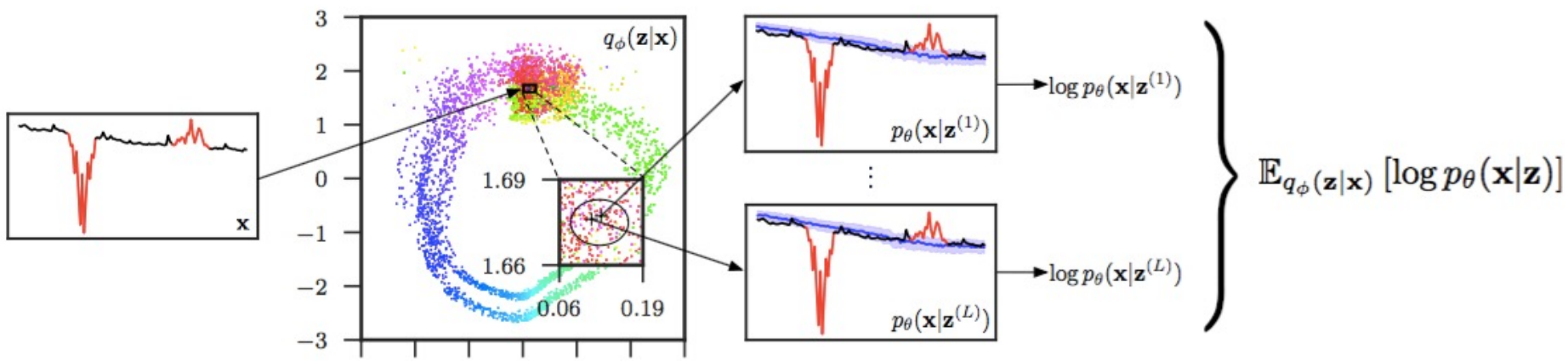
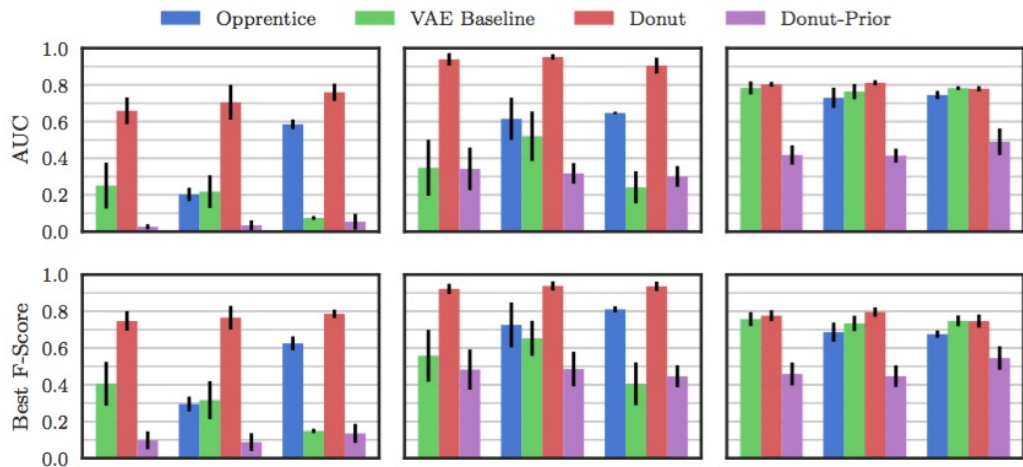


Figure 12: 3-d latent space of all three datasets.

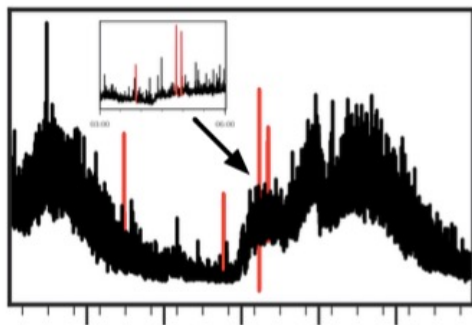
Unsupervised KPI Anomaly Detection Through Variational Auto-Encoder

WWW2018

Accuracy of 0.8~0.9, even better than supervised approach.

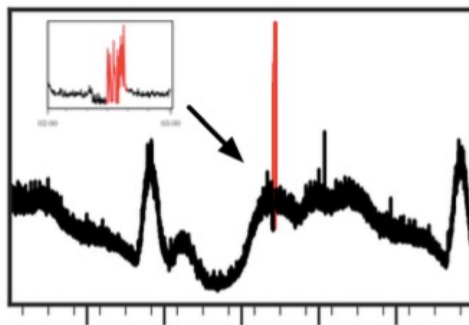


Buzz: Apply Adversarial Training for Non-Gaussian Noise



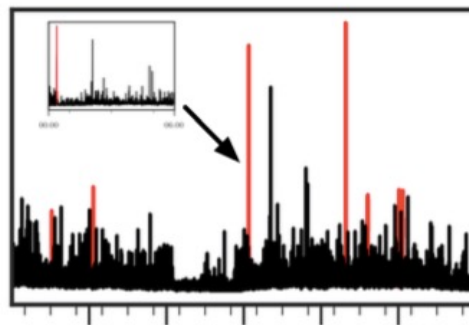
12:00 18:00 00:00 06:00 12:00

A



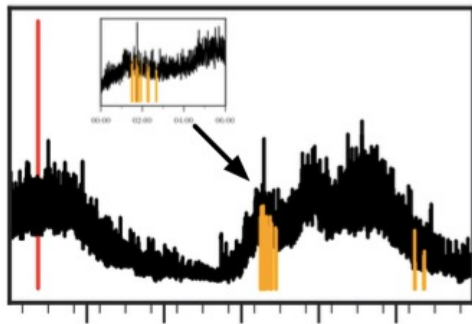
12:00 18:00 00:00 06:00 12:00

B



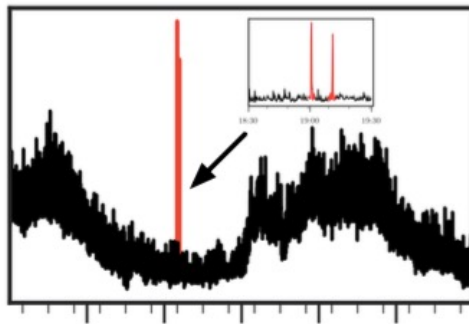
12:00 18:00 00:00 06:00 12:00

C



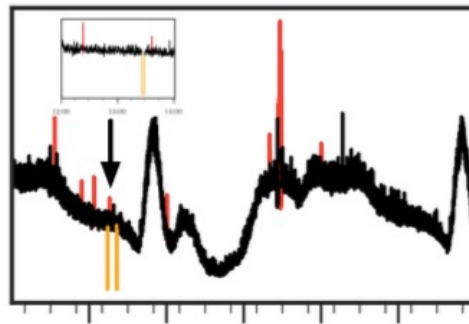
12:00 18:00 00:00 06:00 12:00

D



12:00 18:00 00:00 06:00 12:00

25ϵ



12:00 18:00 00:00 06:00 12:00

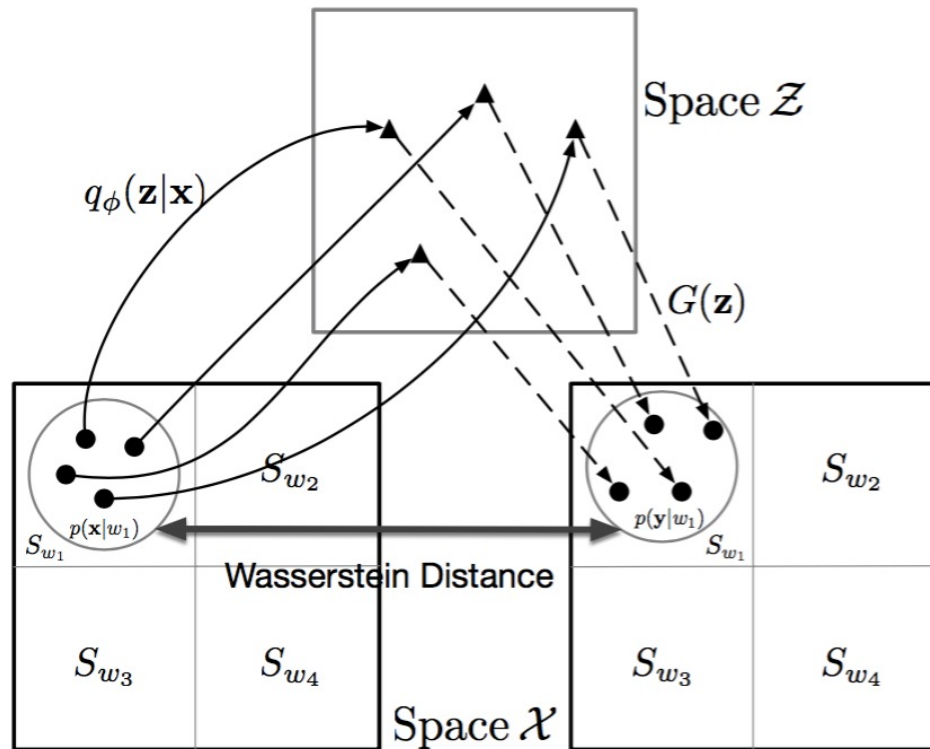
F

Unsupervised Anomaly Detection for Intricate KPIs via Adversarial Training of VAE

INFOCOM 2019

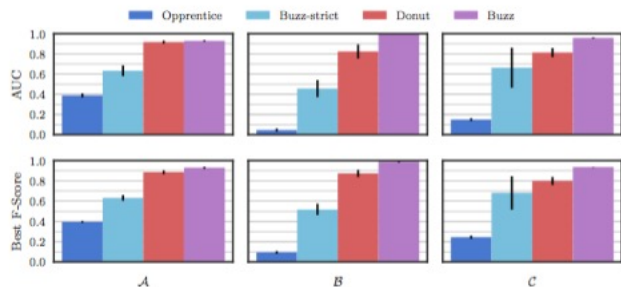
Major ideas

- **Wasserstein distance: the distance between the two probability distributions**
- **Partitioning from measure theory. a powerful and commonly used analysis method for distribution in measure theory.**
- **Adversarial Training**

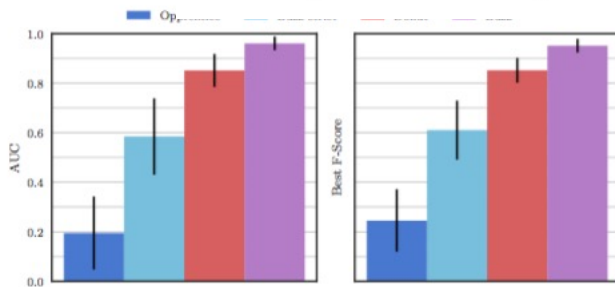
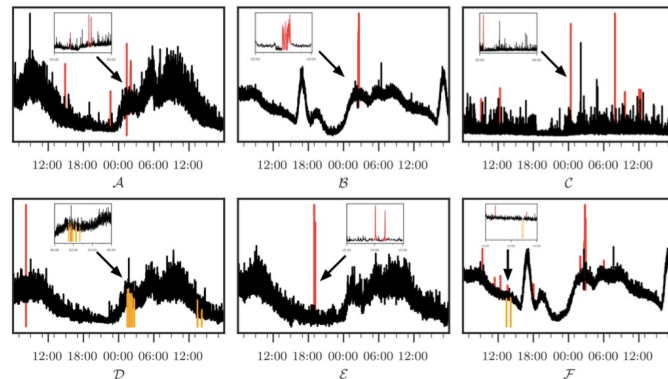


Experiment Results

Best F-Score outperforms Donut by up to 0.15

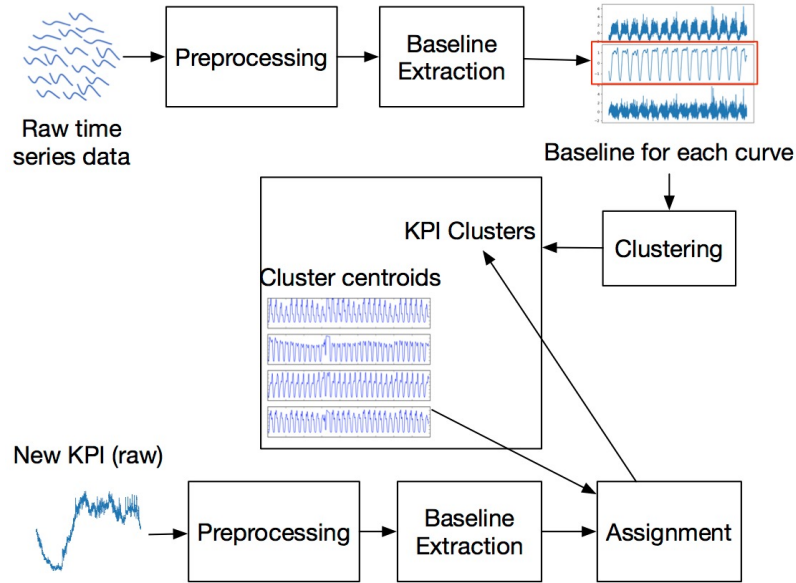


(a) Dataset A, B, C



(b) Average of 11 KPIs

Clustering + Transfer Learning to Reduce Training Overhead



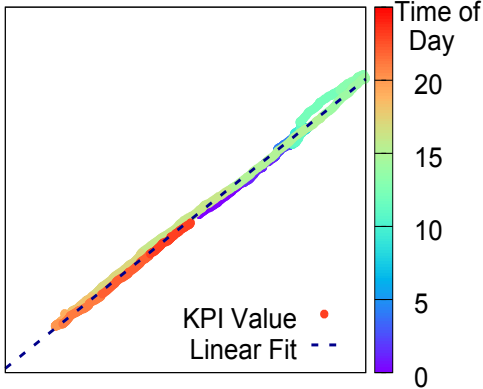
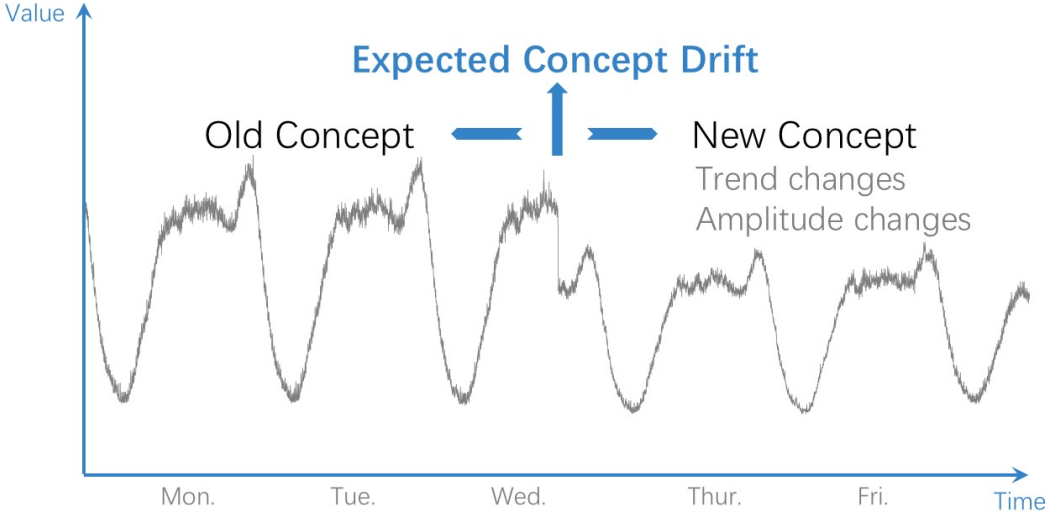
	Original DONUT [WWW2018]	ROCKA+DONUT+KPI-specific threshold
Avg. F-score	0.89	0.88
Total training time (s)	51621	5145

Adapt to Concept Drift

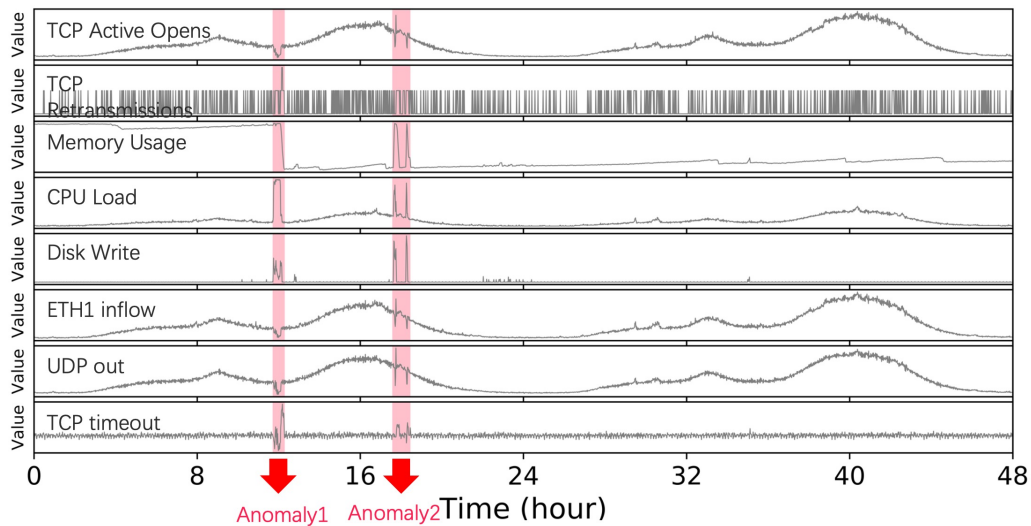
ISSRE 2018 Best Paper

concept drift adaption improve anomaly detection F-score by 203% (**0.225 to 0.681**)

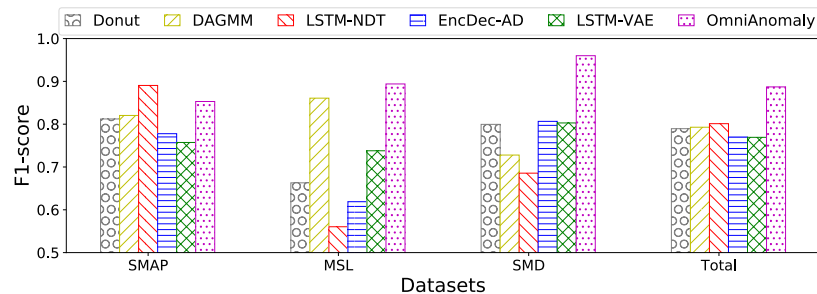
Observation: Old and New Concept Can Be Linearly Fitted



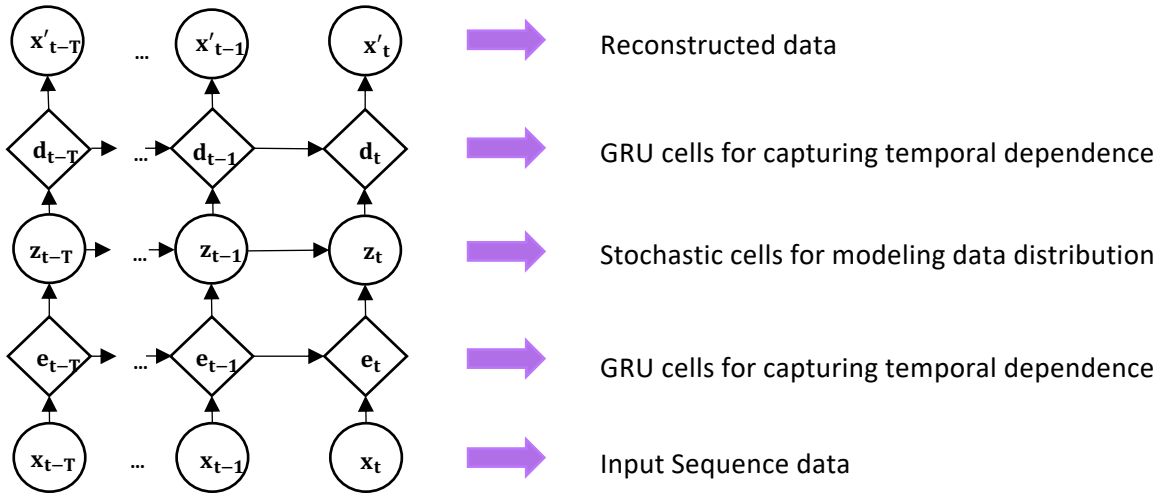
Multivariate Time Series Anomaly Detection with OmniAnomaly (KDD 2019)



F1-best of OmniAnomaly and baselines



Model Architecture of OmniAnomaly



A good z_t can represent x_t well regardless of whether x_t is anomalous or not.

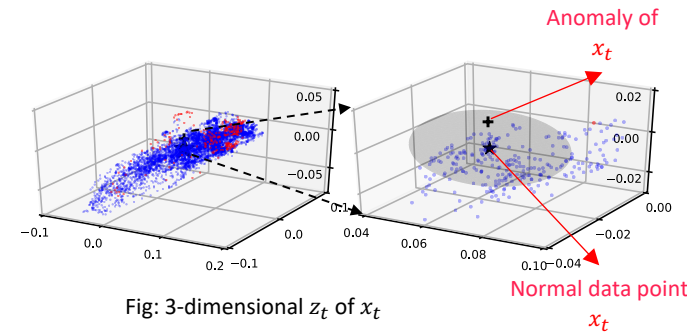


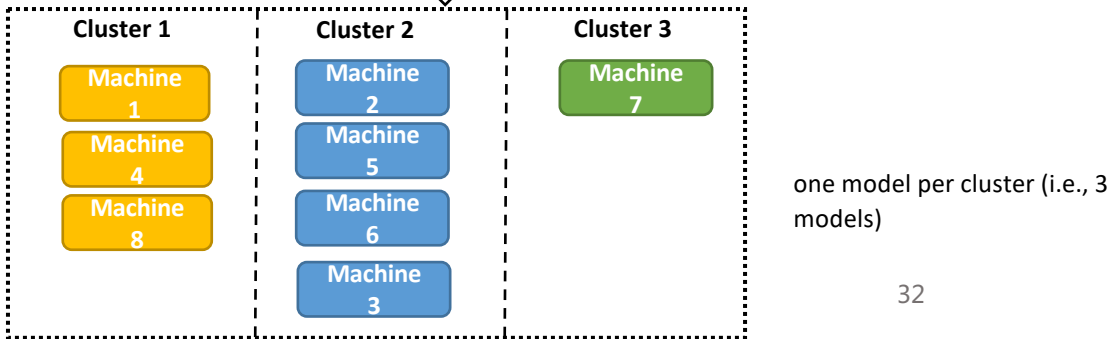
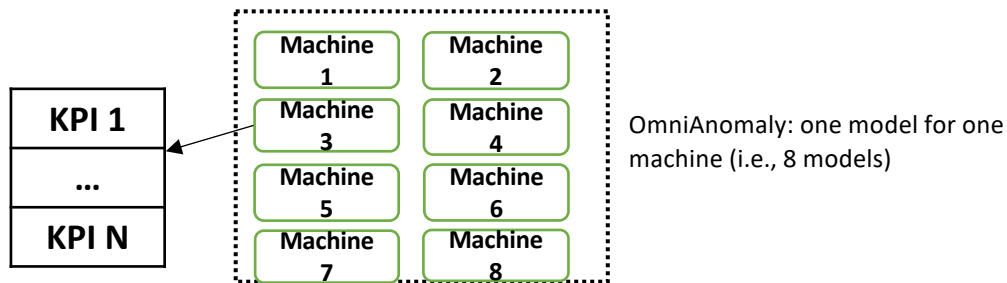
Fig: 3-dimensional z_t of x_t

When x_t is anomalous, its z_t can still represent its normal pattern and x'_t will be normal too.

Transfer Learning in Latent Space for MTSAD

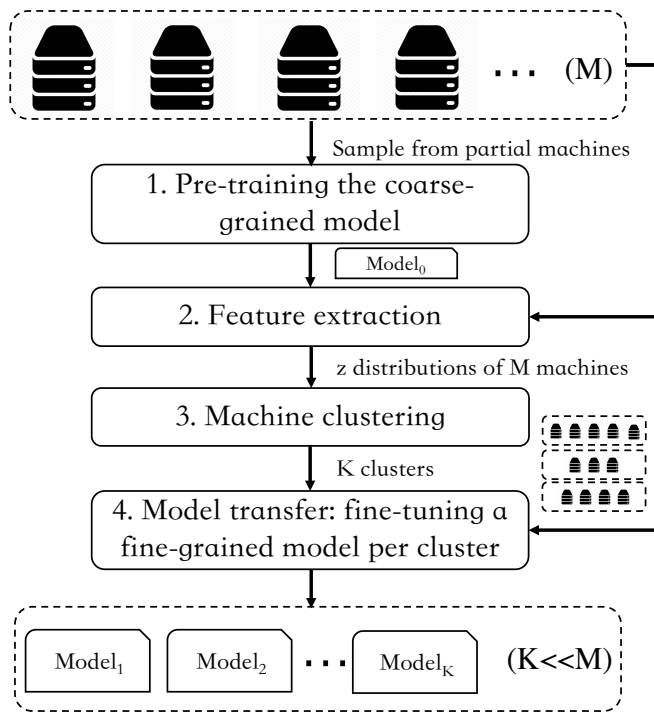
Training one OmniAnomaly model for each machine costs much time (e.g., 900s for each machine).

Clustering and fine-tuning could greatly reduce the training time with a limited accuracy loss.



1. Challenges:
 2. The **high dimensionality** ($N*W$) of multivariate time series with **noises and anomalies**.
- It's challenging to cluster on x or make dimensionality reduction.
 - Noises and anomalies may mislead the measurement of distances.

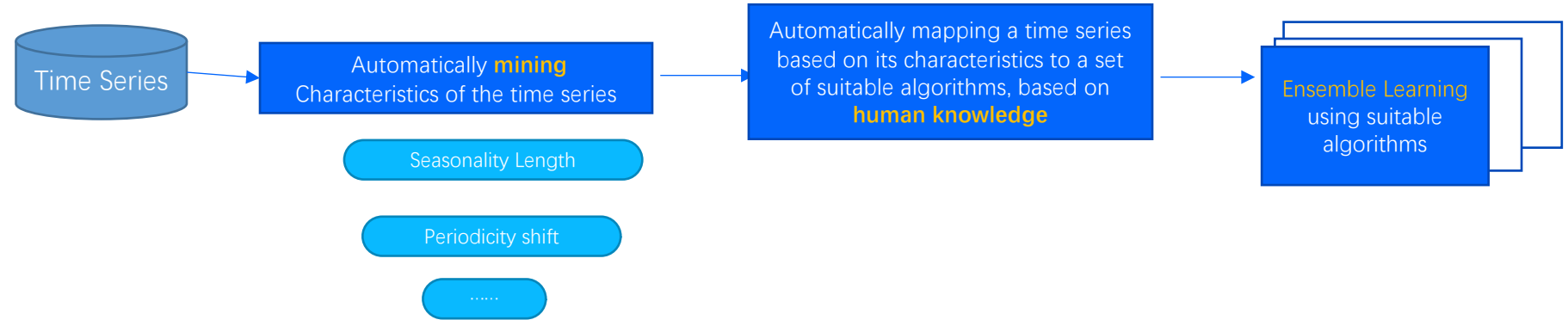
Framework of Model Training



- 1. Sampling strategies in pre-training:**
 - Machine entity sample
 - Time period sample
- 2. Feature extraction:**
 - z sample
- 3. Clustering on z distribution:**
 - Distance: Wasserstein distance
 - Clustering: Hierarchical agglomerative clustering (HAC) algorithm
- 4. Fine-tuning fine-grained models:**
 - Sampling strategies like 1

CTF can reduce the model training time from about two months ($O(M \cdot T_m)$) to 4.40 hours ($O(M \cdot T_f) + O(K \cdot T_m)$) ($M \gg K, T_m \gg T_f$) for one hundred thousand machines. It achieves an F1-Score of **0.830**, with only 0.012 performance loss.

How to do use these algorithms in reality?

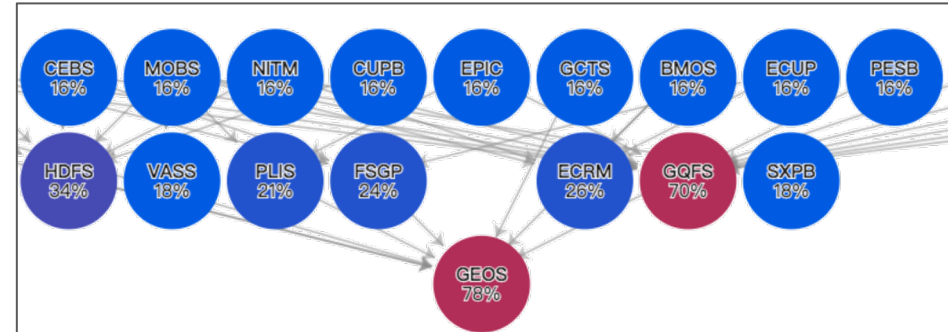
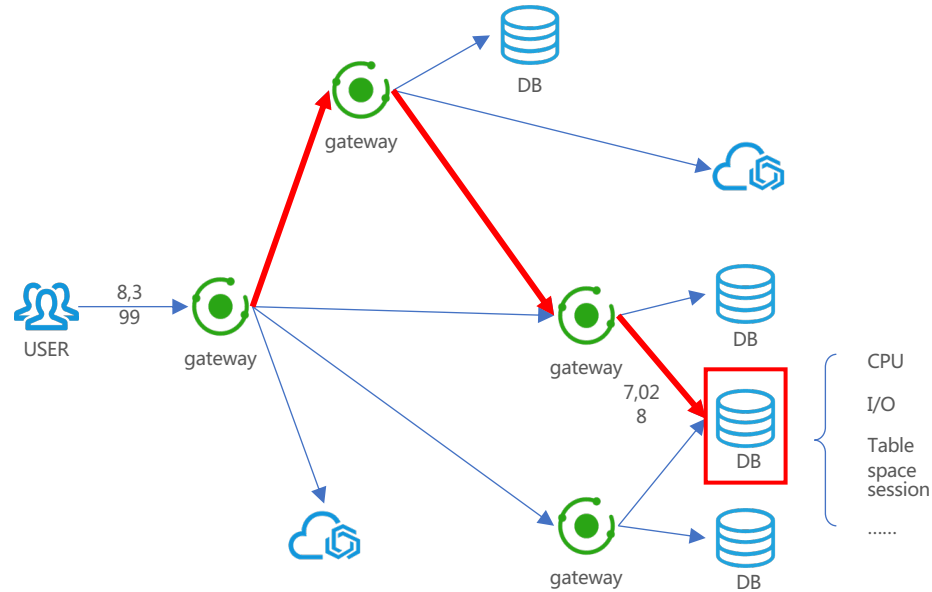


Outline

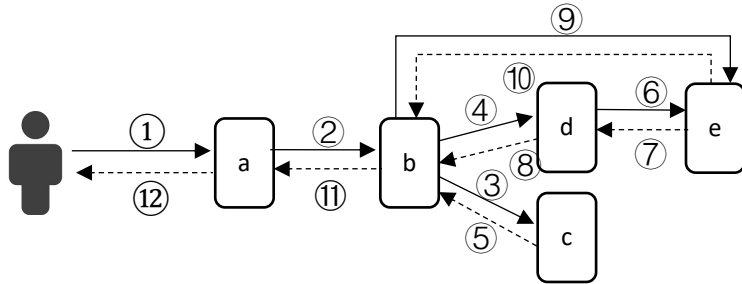
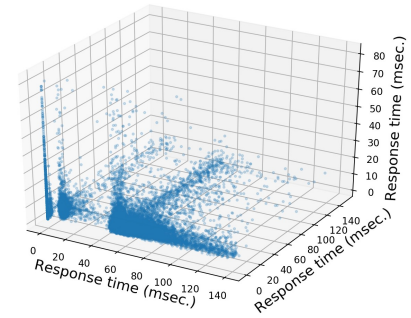
- IT Operations (Ops) background
- Is artificial intelligence necessary for Ops?
- Case Study
 - Unsupervised Anomaly Detection in Ops
 - *Time series anomaly detection (IMC 2015, WWW 2018, IWQoS 2019, INFOCOM 2019a, INFOCOM2019b, ISSRE 2018, IPCCC 2018a, IPCCC 2018b, TSNM 2019, KDD2019, INFOCOM2021)*
 - *Trace anomaly detection (ISSRE 2020)*
 - Zero-day attack detection (INFOCOM2020a)
- Lessons Learned

Software Module Invocation Traces

- Invocation trace: 10s~100s of module-to-module invocations for a unique transaction
 - One module failure can manifest itself cross-invocation and cross-transaction



This mandates that response times and call paths must be unified



For a microservice, its response time is determined by both **itself** and its call path

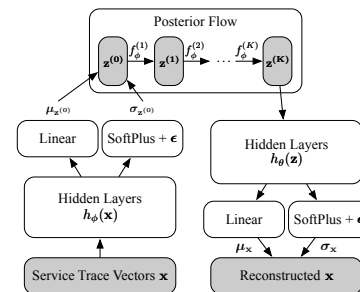
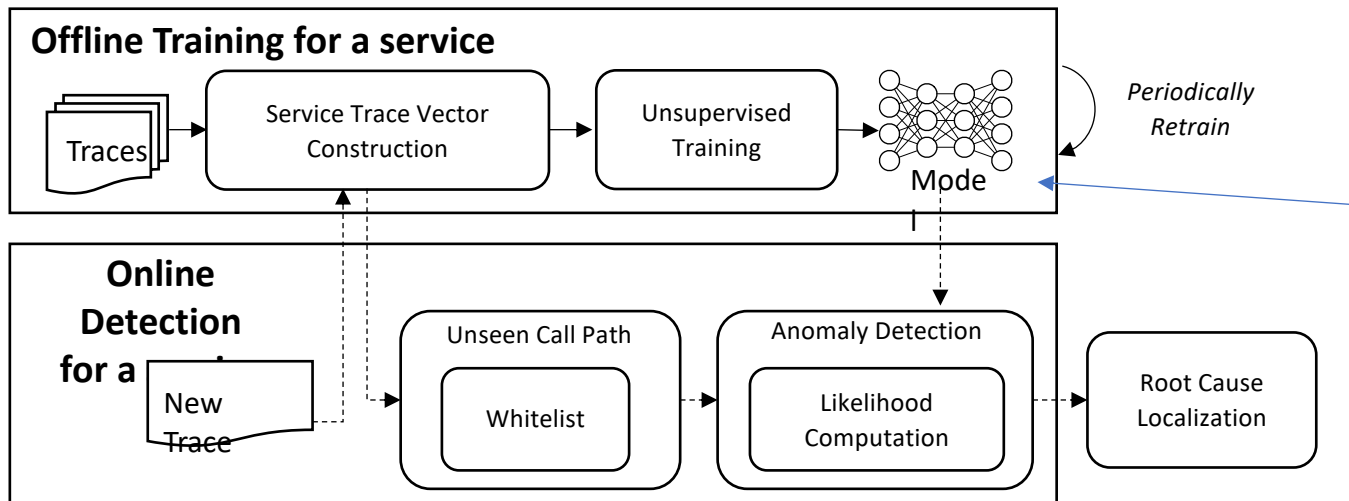
Microservices	Call path of microservice s (s, call path)	Response time of (s, call path) (msec)
a	(a, (start→a))	222
b	(b, (start→a, a→b))	209
c	(c, (start→a, a→b, b→c))	4
d	(d, (start→a, a→b, b→c, b→d))	44
e	(e, (start→a, a→b, b→c, b→d, d→e))	28
e	(e, (start→a, a→b, b→c, b→d, d→e, b→e))	67

Microservice e is invoked twice, with different response time

Design of TraceAnomaly

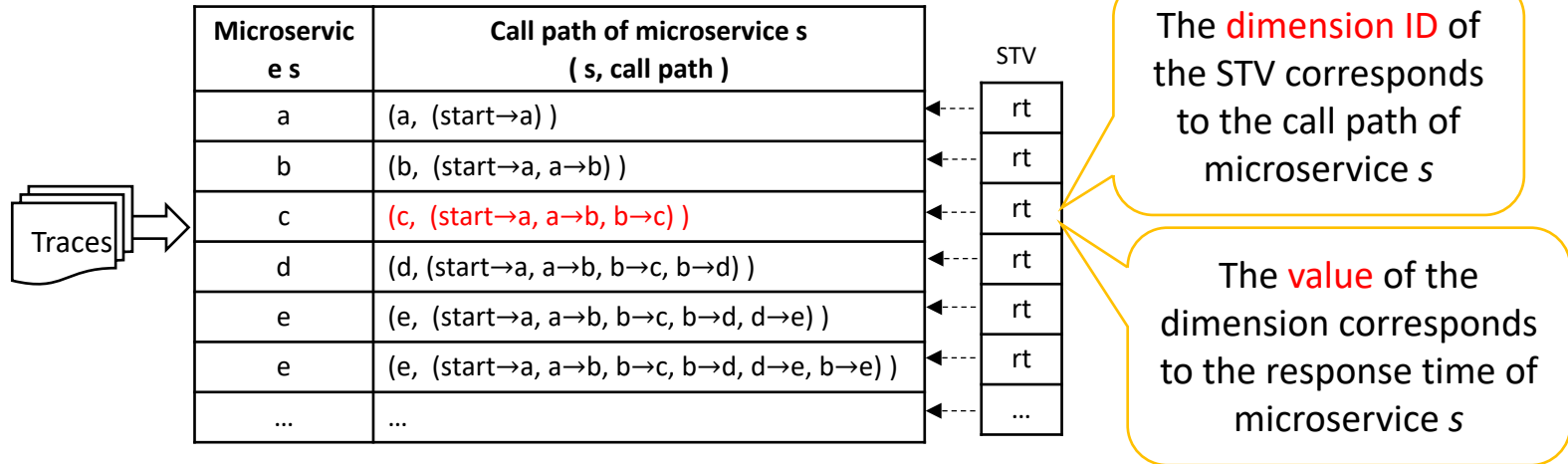
TABLE III: Online evaluation results of different approaches on four large online services which contain hundreds of microservices, whose statistics are shown in Table I.

	Service-1		Service-2		Service-3		Service-4		Overall (Union of 4 services)	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Hard-coded Rule	0.910	0.800	0.920	0.792	0.911	0.812	0.930	0.800	0.910	0.804
WFG-based [5]	0.020	0.500	0.012	0.323	0.050	0.410	0.032	0.300	0.031	0.386
DeepLog* [8]	0.270	0.680	0.241	0.560	0.320	0.643	0.302	0.601	0.290	0.628
CPD-based [7]	0.52	0.063	0.43	0.090	0.57	0.110	0.64	0.072	0.531	0.081
CFG-based [6]	0.170	0.610	0.250	0.570	0.102	0.503	0.180	0.630	0.164	0.562
TraceAnomaly	0.980	1.000	0.982	1.000	0.981	1.000	0.973	1.000	0.981	1.000



Service trace vector construction

- Unify response time and call paths of traces in an interpretable way
 - Encode the response time and call paths of a trace in a service into a STV (Service Trace Vector)



Outline

- IT Operations (Ops) background
- Is artificial intelligence necessary for Ops?
- Case Study
 - Unsupervised Anomaly Detection in Ops
 - *Time series anomaly detection (IMC 2015, WWW 2018, IWQoS 2019, INFOCOM 2019a, INFOCOM2019b, ISSRE 2018, IPCCC 2018a, IPCCC 2018b, TSNM 2019, KDD2019, INFOCOM2021)*
 - Trace anomaly detection (ISSRE 2020)
 - Zero-day attack detection (INFOCOM2020a)
- Lessons Learned

Detecting Zero-day Attacks

- WAF detects those **known** attacks effectively.
 - filter out **known** attacks
- **ZeroWall** detects **unknown** attacks **ignored by WAF rules**.
 - report **new attack patterns** to operators and security engineers to **update WAF rules**.

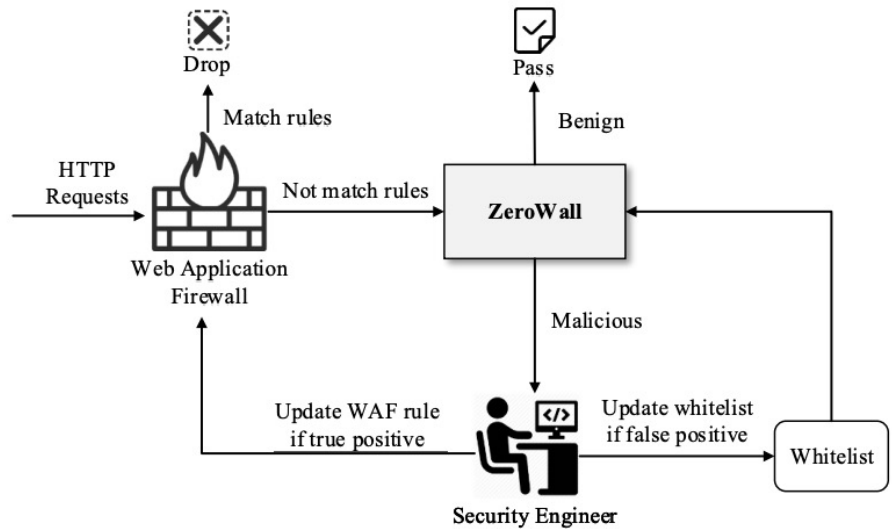
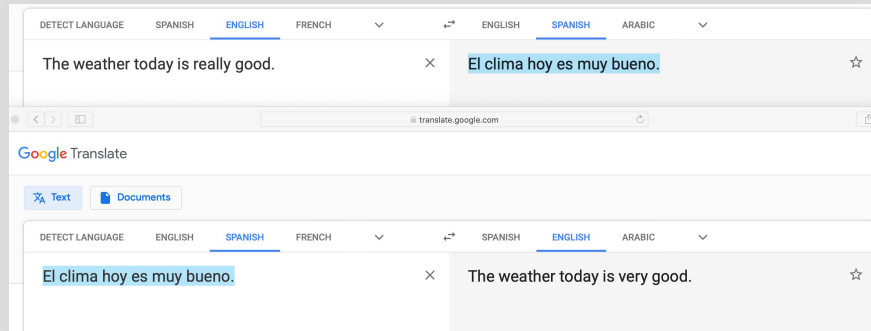


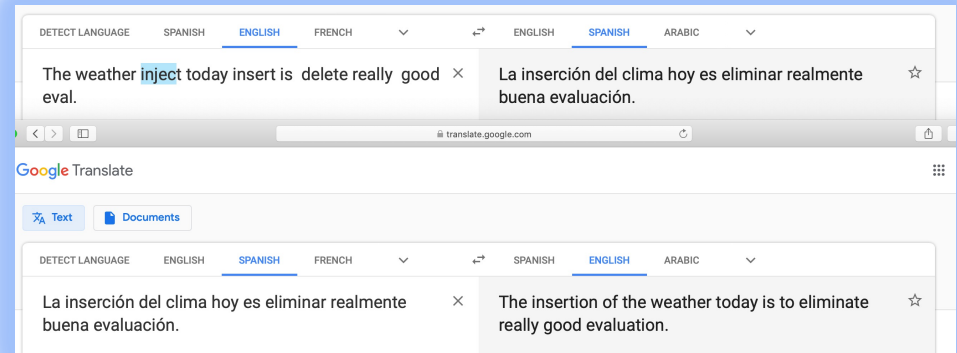
Figure 1: The workflow of ZeroWall.

Self-Translate Machine



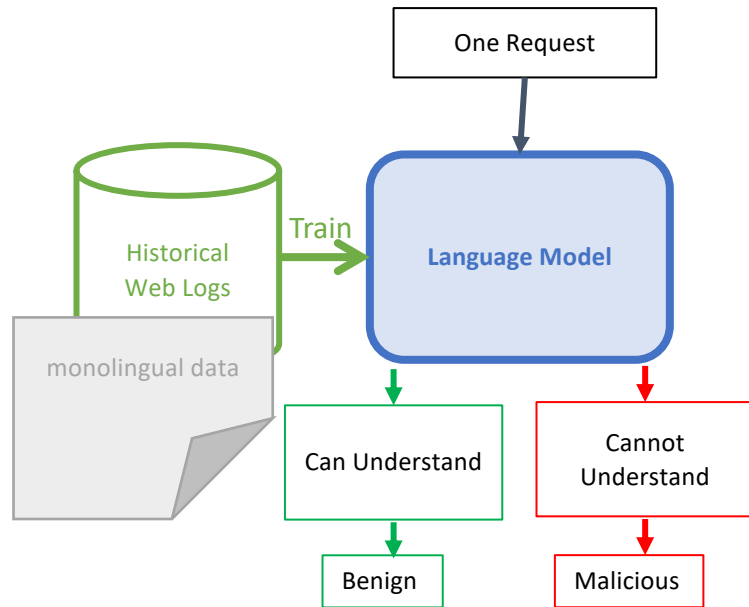
Self-translation works **well** for **normal** sentences

Output **deviates** significantly from the input, when the input is a sentence **not previously seen** in the training dataset of the self-translation models.



Idea

- HTTP request is a **string following HTTP**, and we can consider an HTTP request as one **sentence** in the **HTTP request language**.
- **Most** requests are **benign**, and **malicious** requests are **rare**.
- Thus, we train a kind of **language model** based on historical logs, to **learn this language** from **benign requests**.



Deployed in the wild

Over **1.4** billion requests

Captured **28** different types of zero-day attacks (**10K** of zero-day attack requests)

Low overhead

Summary: Unsupervised Anomaly Detection in Ops

- Common Idea: somehow capture the “normal” patterns in the historical data, then any new points that “deviate” from the normal patterns are considered “anomalous”.
- Domain specific feature engineering (time series, log, trace, etc.)
- Sometimes have to assume non-Gaussian distributions in x-space or z-space
 - GAN
 - Flows in Z-space
- Temporal dependency can be captured in x-space or z-space
- Reconstruction-based models are more robust than prediction-based models
- Clustering + transfer learning in x-space or z-space help reduce training overhead with little accuracy loss.
- Various distance metrics: e.g. Wasserstein distance
- Periodic re-training + whitelisting (active learning) for small changes
- Transfer learning for concept change.

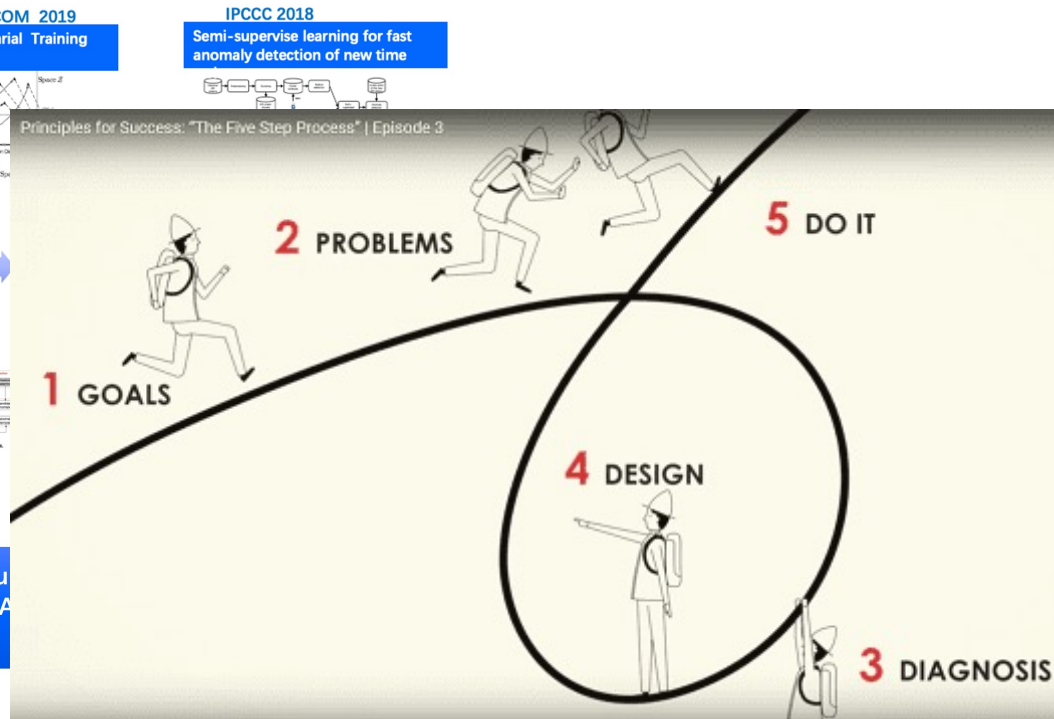
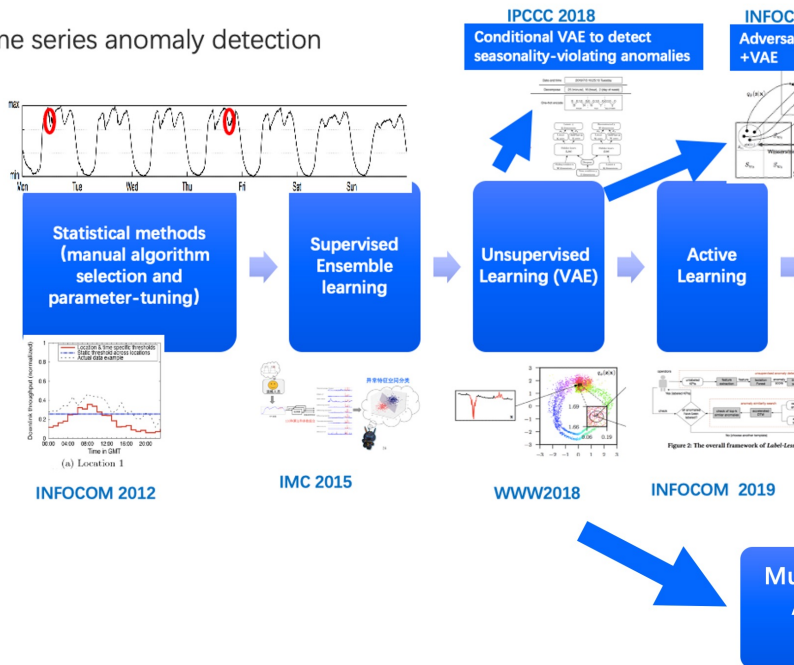
Outline

- **IT Operations (Ops) background**
- **Is machine learning necessary for Ops?**
- **Case Study**
 - **Unsupervised Anomaly Detection in Ops**
- ***Lessons Learned***

Lesson 1: From Practice, Into Practice

- 1. Discover challenging problems from Practice (specifically, IT Operations)
- 2. Design AI Algorithms to solve a problem
- 3. Deploy the algorithms in practice. If not working perfectly? go to step 1.

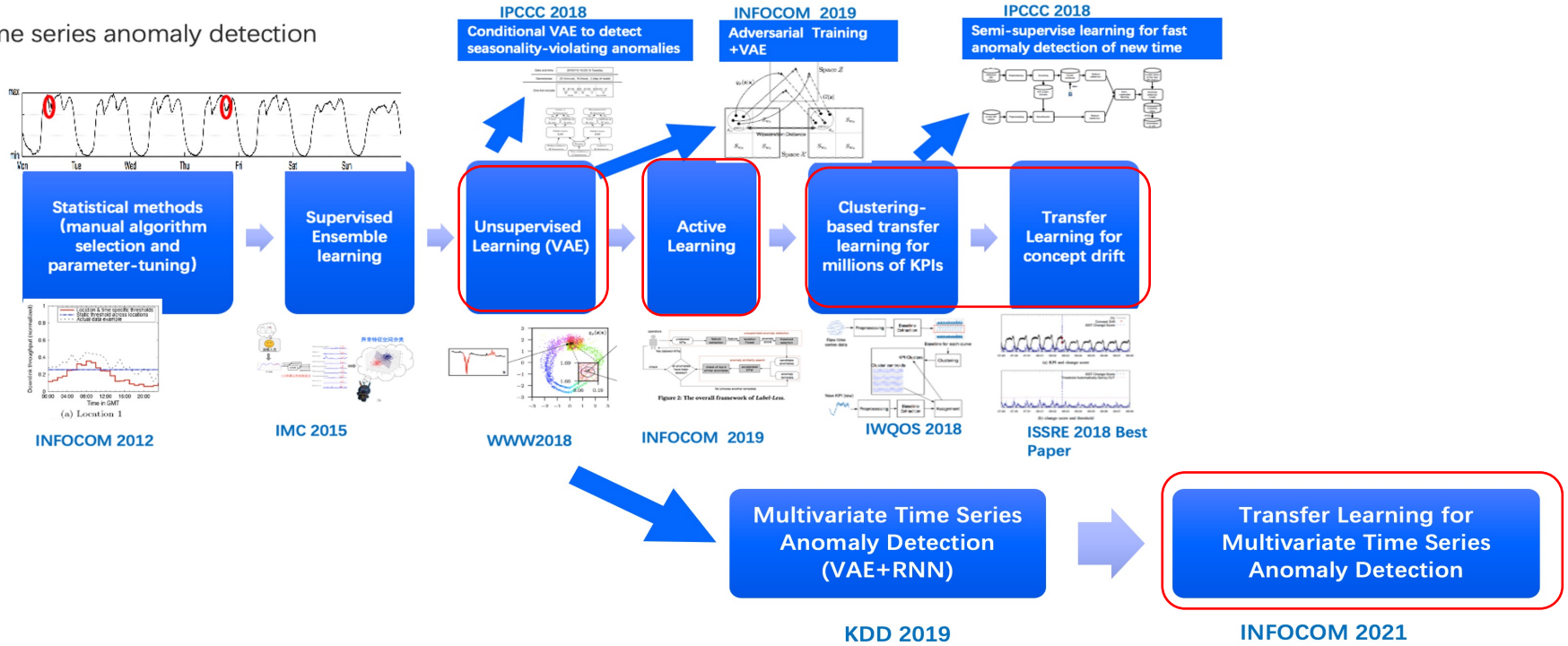
Time series anomaly detection



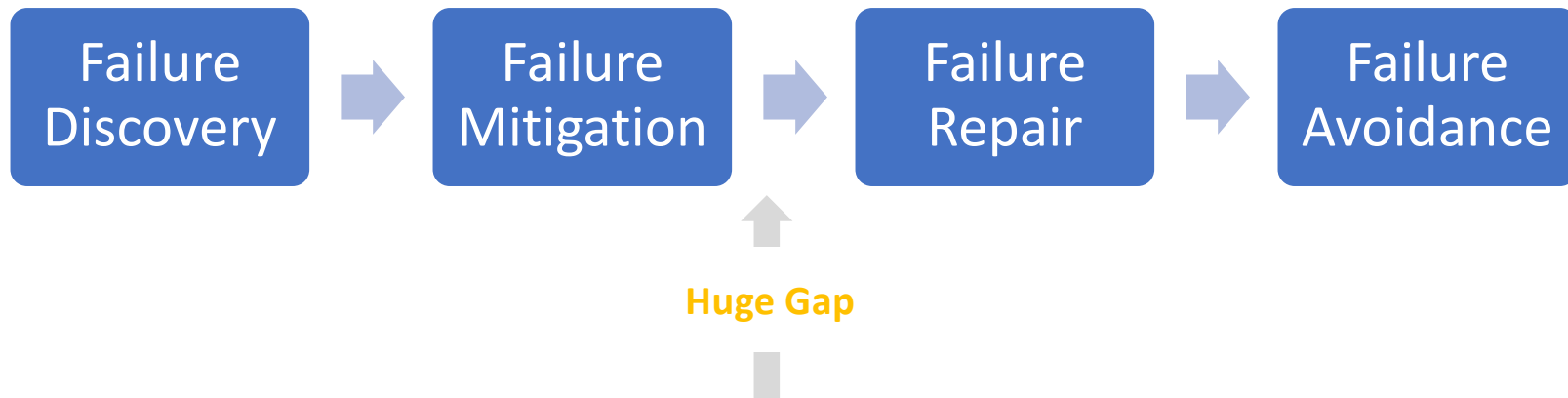
Lesson 2: Fully utilize latest AI technologies that enable better machine-human hybrid architecture

1. Unsupervised approaches
2. Unsupervised approaches + transfer learning
3. Unsupervised approaches + active learning
4. Weakly supervised learning (e.g. multi-instance learning, PU Learning)
5. Semi-supervised approaches; supervised approaches + transfer learning
6. Supervised approaches

Time series anomaly detection



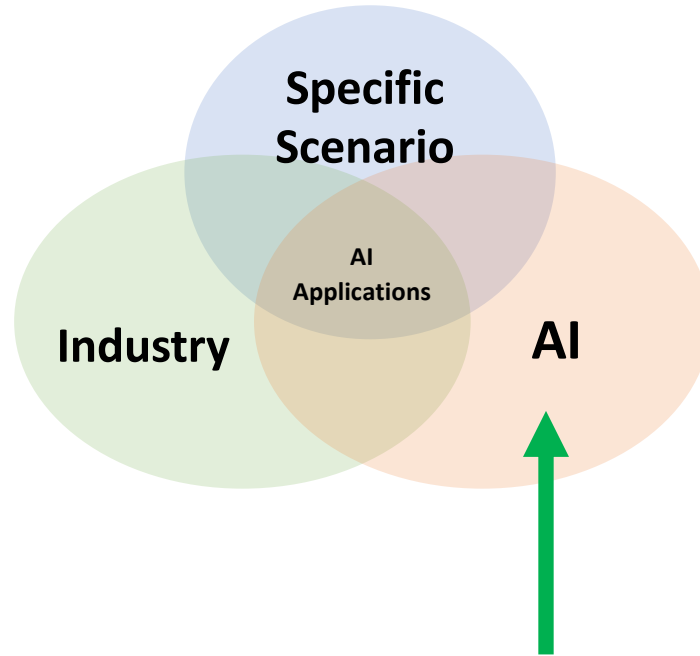
Pitfalls: use general ML algorithms as Blackbox to tackle Ops challenges



General Machine Learning Algorithms

ARIMA, Time Series Decomposition, Holt-Winters, CUSUM, SST, DiD, DBSCAN, Pearson Correlation, J-Measure, Two-sample test, Apriori, FP-Growth, K-medoids, CLARIONS, Granger Causality, Logistic Regression, Correlation analysis (event-event, event-time series, time series-time series), hierarchical clustering, Decision tree, Random forest, support vector machine, Monte Carlo Tree search, Markovian Chain, multi-instance learning, transfer learning, CNN, RNN, VAE, GAN, NLP

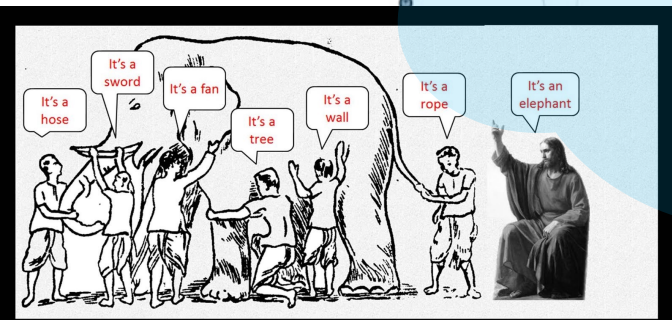
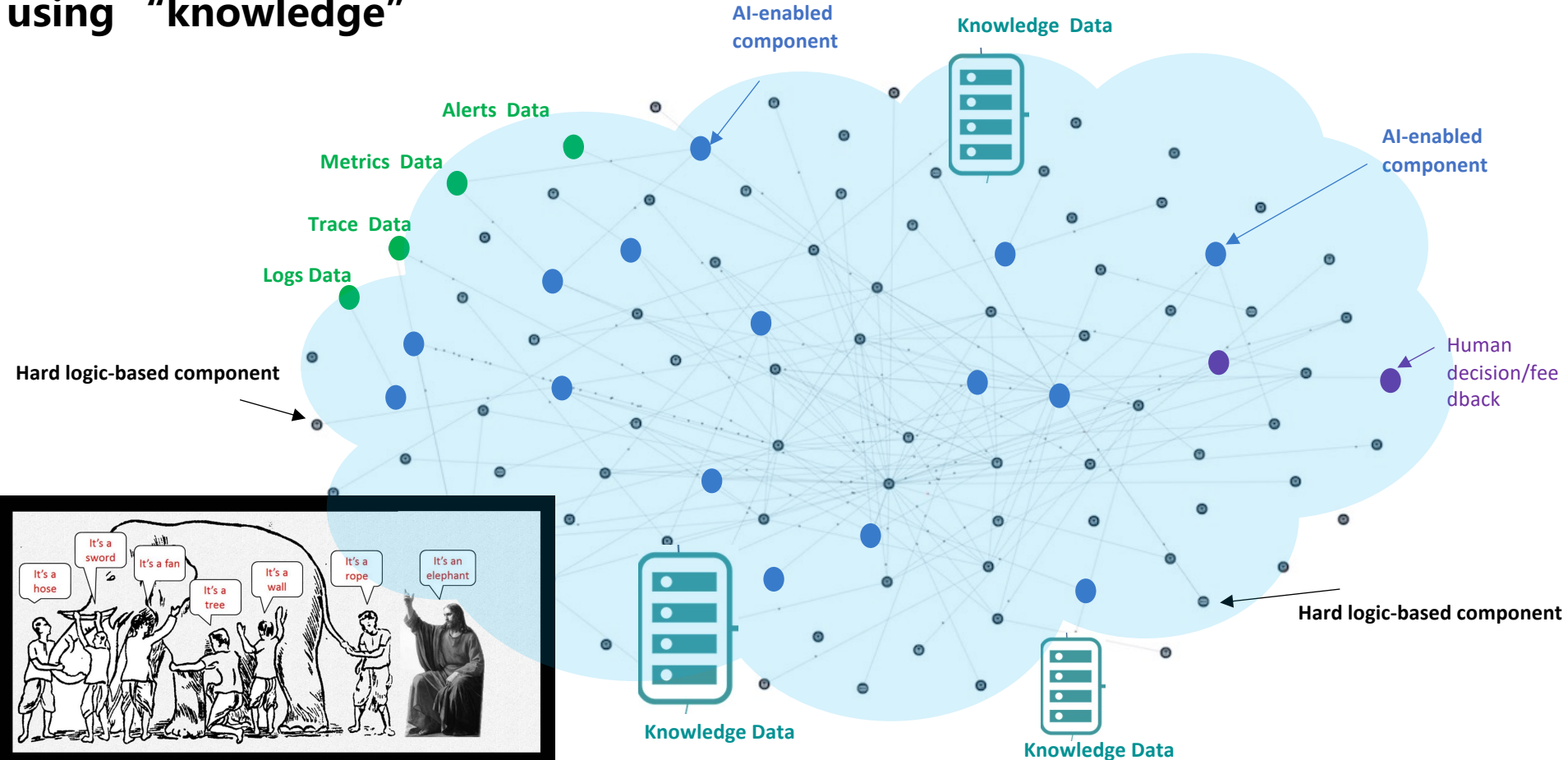
So far, AI succeeds only in specific application scenario in specific area in specific industry



Treat AI as a **high-level programming language**, to “code” some components


Output of AI-enabled components are **probabilistic** rather than deterministic

Lesson 3: divide and conquer, design the overall system around each component's known capability and property, and "glue" the components using "knowledge"



Some Open-Sourced Algorithms from NetMan

<https://github.com/netmanaiops>



NetManAIops
The public codes and datasets of Tsinghua NetMan Lab.
Tsinghua University <http://netman.aiops.org>

Repositories **14** Packages People **1** Projects


Grow your team on GitHub

GitHub is home to over 50 million developers working together. Join them to grow your own development teams, manage permissions, and collaborate on projects.

[Sign up](#)


Dismiss

Pinned repositories

 **donut**


WWW 2018: Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications

Python ☆ 292 🍴 109

 **TraceAnomaly**


ISSRE'20: Unsupervised Detection of Microservice Trace Anomalies through Service-Level Deep Bayesian Networks

Python ☆ 206 🍴 40

 **LogParse**

An adaptive log template extraction toolkit.

Python ☆ 203 🍴 31

 **OmniAnomaly**

KDD 2019: Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network

Python ☆ 155 🍴 71

 **LogClass**

IWQoS 2018 short paper: Device-agnostic log anomaly classification with partial labels

Python ☆ 124 🍴 38

 **Log2Vec**

A distributed representation method for online logs.

Roff ☆ 63 🍴 11

AIOps Challenge Algorithm Competitions

Datasets: <https://github.com/netmanaiops>

- 2018 AIOps Challenge: time series anomaly detection. [Published labeled data from 5 Internet companies](#). More than 50 teams participated. [Papers based on these data were published in KDD, IWQoS, etc.](#)
Data Downloadable @ <https://github.com/NetManAIOps/KPI-Anomaly-Detection>
- 2019 AIOps Challenge: multi-attribute time series anomaly localization. [Published data from an Internet company](#). More than 60 teams participated.
Data Downloadable @ <https://github.com/NetManAIOps/MultiDimension-Localization>
- 2020 AIOps Challenge: Anomaly detection and localization in a microservice system. [Published data from a telecom company](#). More than 100 teams participated.
Data Downloadable @ <https://github.com/NetManAIOps/AIOps-Challenge-2020-Data>
- 2021 AIOps Challenge: Anomaly detection and localization in banking systems. [To be published data from two banks](#). More than 200 teams participated

2019国际AIOps挑战赛决赛暨AIOps研讨会

2019.7.13



Summary

- **AI for IT Operations (AIOps) is an interdisciplinary research field between AI and Systems/Networking/Software Engineering/Security**
 - **Towards Autonomous IT Operations.**
- **AIOps will be a foundational technology in the increasingly digitalized world**
- **Many deep and challenging research problems to be solved in AIOps**
- **Lessons learned so far:**
 - **Divide and conquer instead of using black box**
 - **Wide range of AI algorithms for AIOps**
 - **From practice, into practice**
 - **As little labeling as possible**
 - **Problem formulation matters**
 - **Utilize as many data sources as possible**
- **Long-term community efforts are needed to solve AIOps problems**

Thanks!
Q&A