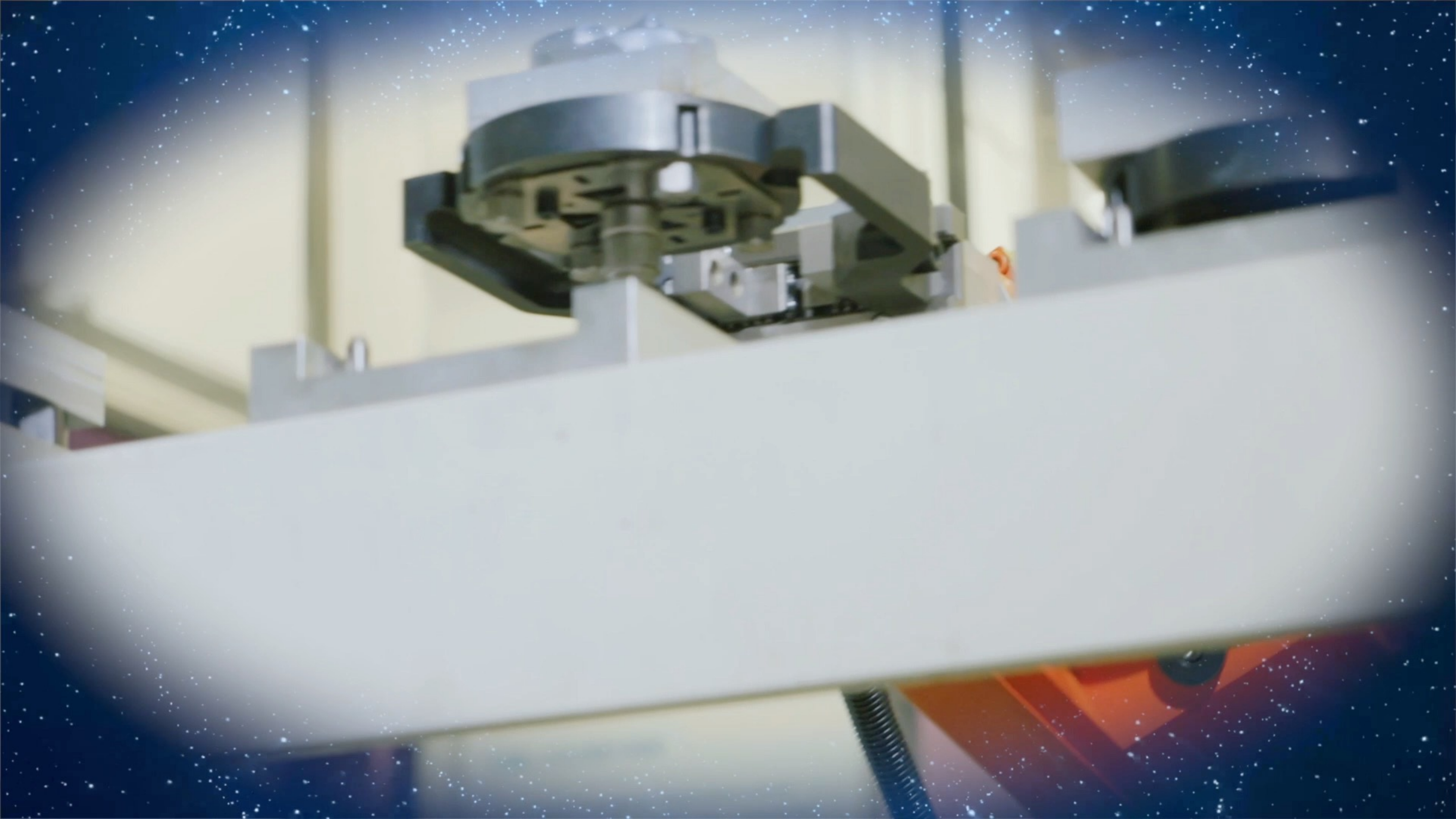
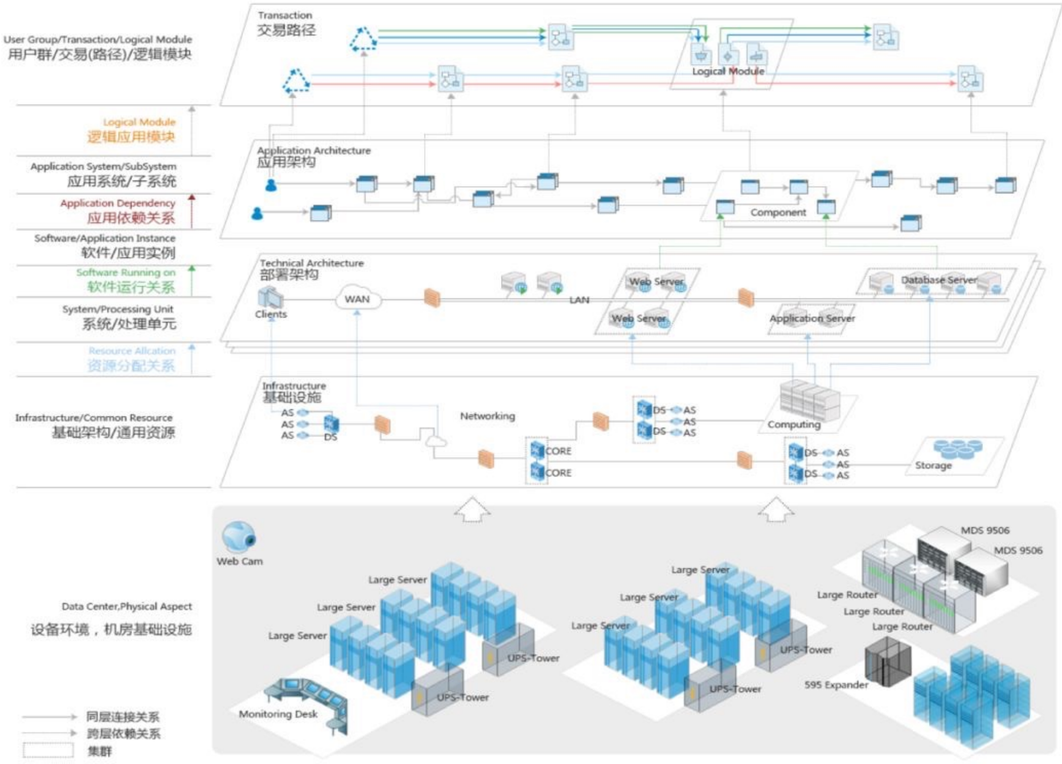
Towards Autonomous IT Operations through Artificial Intelligence

Dan Pei
Tsinghua University

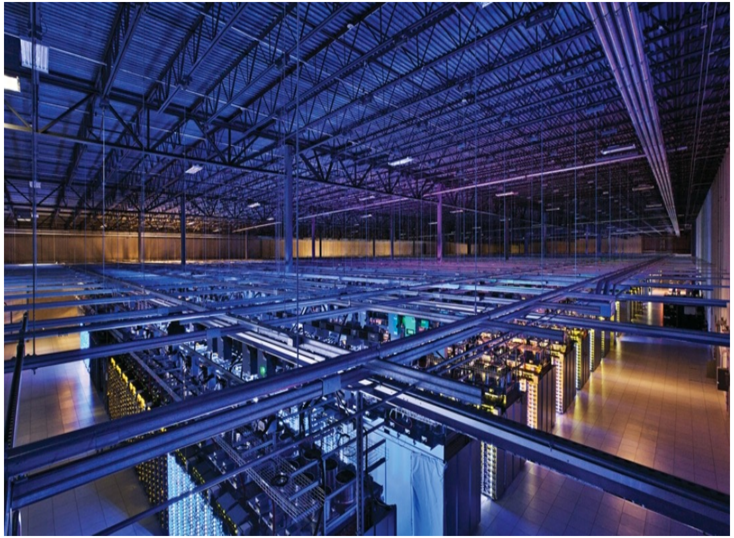
清華大學 | NetMan Tsinghua



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



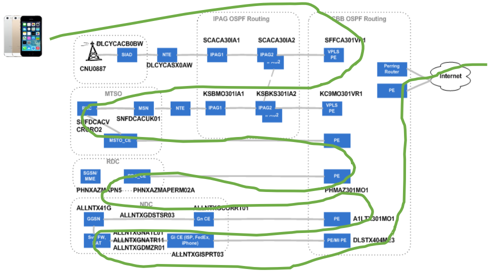
TikTok NETFLIX facebook YouTube



IT Operations

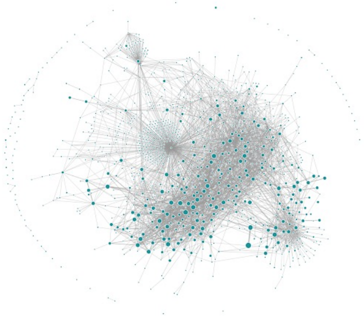
Responsible for ensuring the digitalized businesses and societies run reliably, efficiently and safely, despite the inevitable failures of the imperfect underlying hardware and software.

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

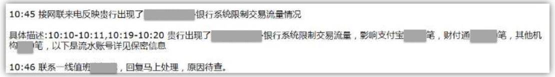
Manual

30 Engineers involved

10:20 large number of transaction failures

Realized there was a failure when customers called 10:45

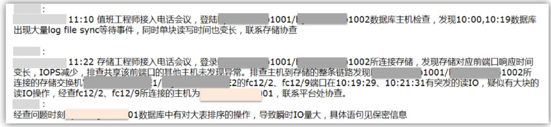
Failure localized 11:10



10:45 接网联来电反映贵行出现了[redacted]银行系统限制交易流量情况

具体描述:10:10-10:11,10:19-10:20 贵行出现了[redacted]银行系统限制交易流量，影响支付宝[redacted]笔，财付通[redacted]笔，其他机构[redacted]笔，以下是流水账号详见保密信息

10:46 联系一线值班[redacted]，回复马上处理，原因待查。



[redacted]：
[redacted] 11:10 值班工程师接入电话会议，登陆[redacted]1001/[redacted]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，同时单块读写时间也变长，联系存储协查

[redacted]：
[redacted] 11:22 存储工程师接入电话会议，登录[redacted]1001/[redacted]1002所连接存储，发现存储对应前端口响应时间变长，IOPS减少，排查共享该前端口的其他主机未发现异常。排查主机到存储的整条链路发现[redacted]1001/[redacted]1002所连接的存储交换机[redacted]1/[redacted]2的fc12/2、fc12/9端口在10:19:29、10:21:31有突发的读IO，疑似有大块的读IO操作，经查fc12/2、fc12/9所连接的主机为[redacted]01，联系平台处协查。
[redacted]：
经查问题时刻[redacted]01数据库中有对大表排序的操作，导致瞬时IO量大，具体语句见保密信息

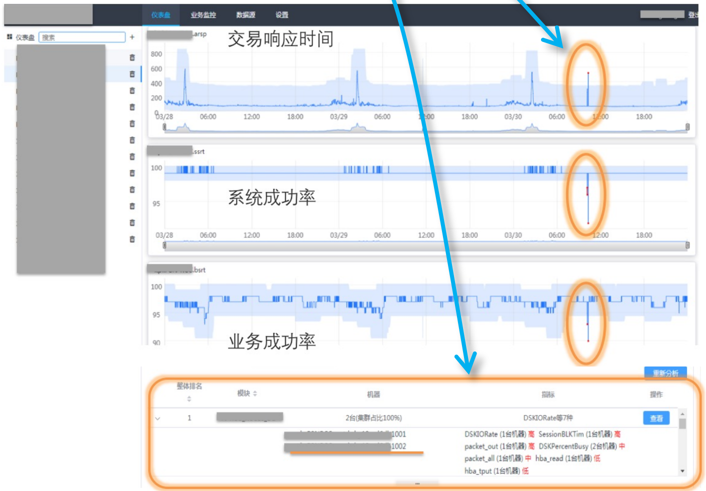
Failure discovery: 25mins after the failure happened
Failure localization: 25mins after failure discovery

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

10:21 automatically detected the failure

10:23 automatically localized the failure

Failure mitigation time reduced by 90%



整体排名	模块	机器	指标	操作
1	[redacted]	2台(集群占比100%)	DSKIORate等7种	查看
		[redacted]1001 [redacted]1002	DSKIORate (1台机器) 高 SessionBLKTim (1台机器) 高 packet_out (1台机器) 高 DSKPercentBusy (2台机器) 中 packet_all (1台机器) 中 hba_read (1台机器) 低 hba_tput (1台机器) 低	

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

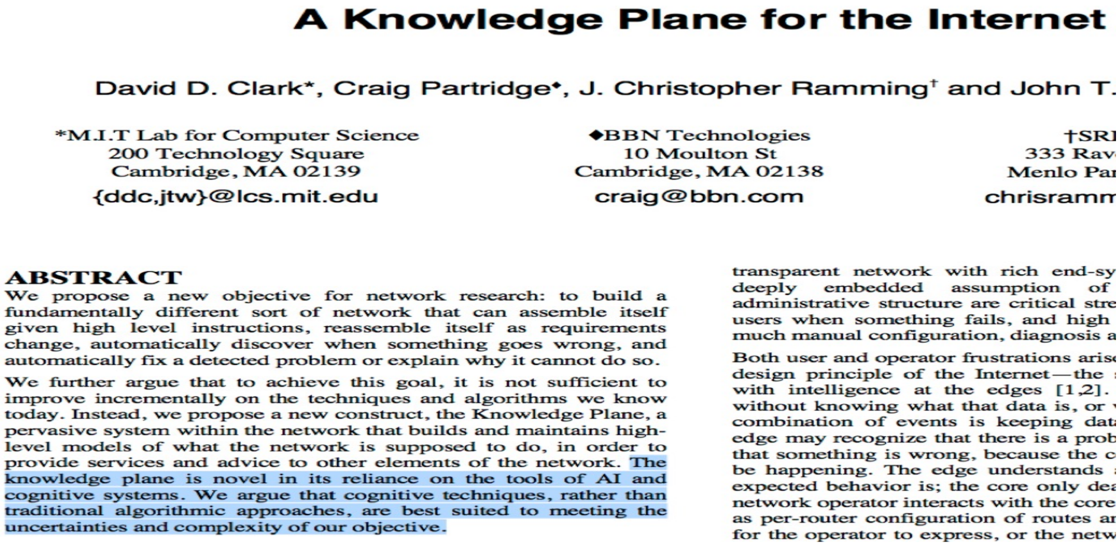
DATADOG

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 Pa
chrisramm

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

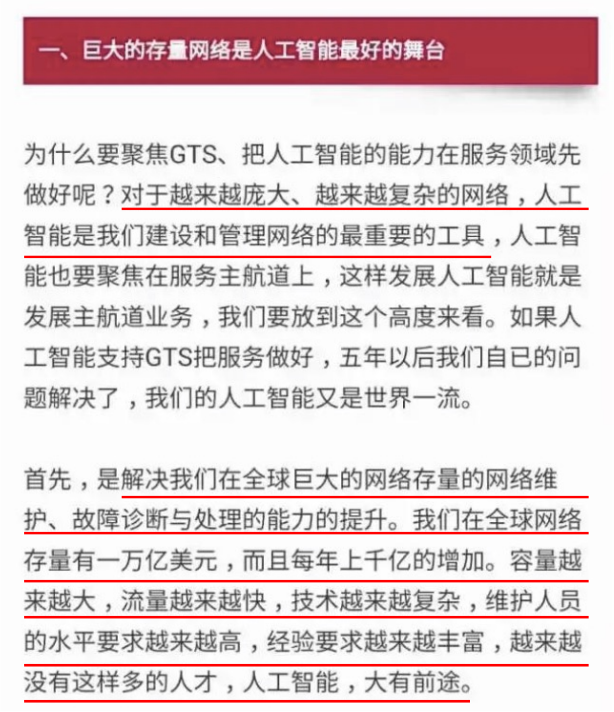
Both user and operator frustrations aris
design principle of the Internet—the
with intelligence at the edges [1,2].
without knowing what that data is, or
combination of events is keeping dat
edge may recognize that there is a prob
that something is wrong, because the c
be happening. The edge understands
expected behavior is; the core only de
network operator interacts with the core
as per-router configuration of routes a
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.



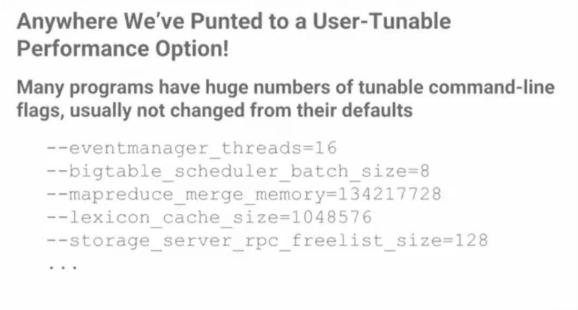
一、巨大的存量网络是人工智能最好的舞台

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，五年以后我们自已的问题解决了，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:

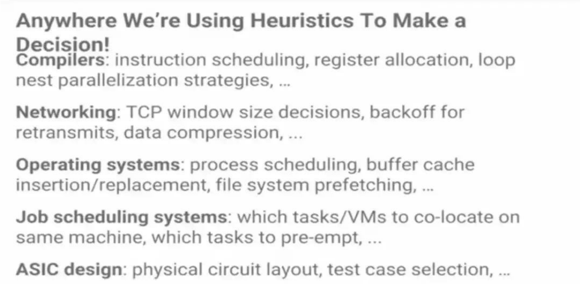
“We can (use AI to) improve everywhere in a system that have tunable parameters or heuristics”



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, ...

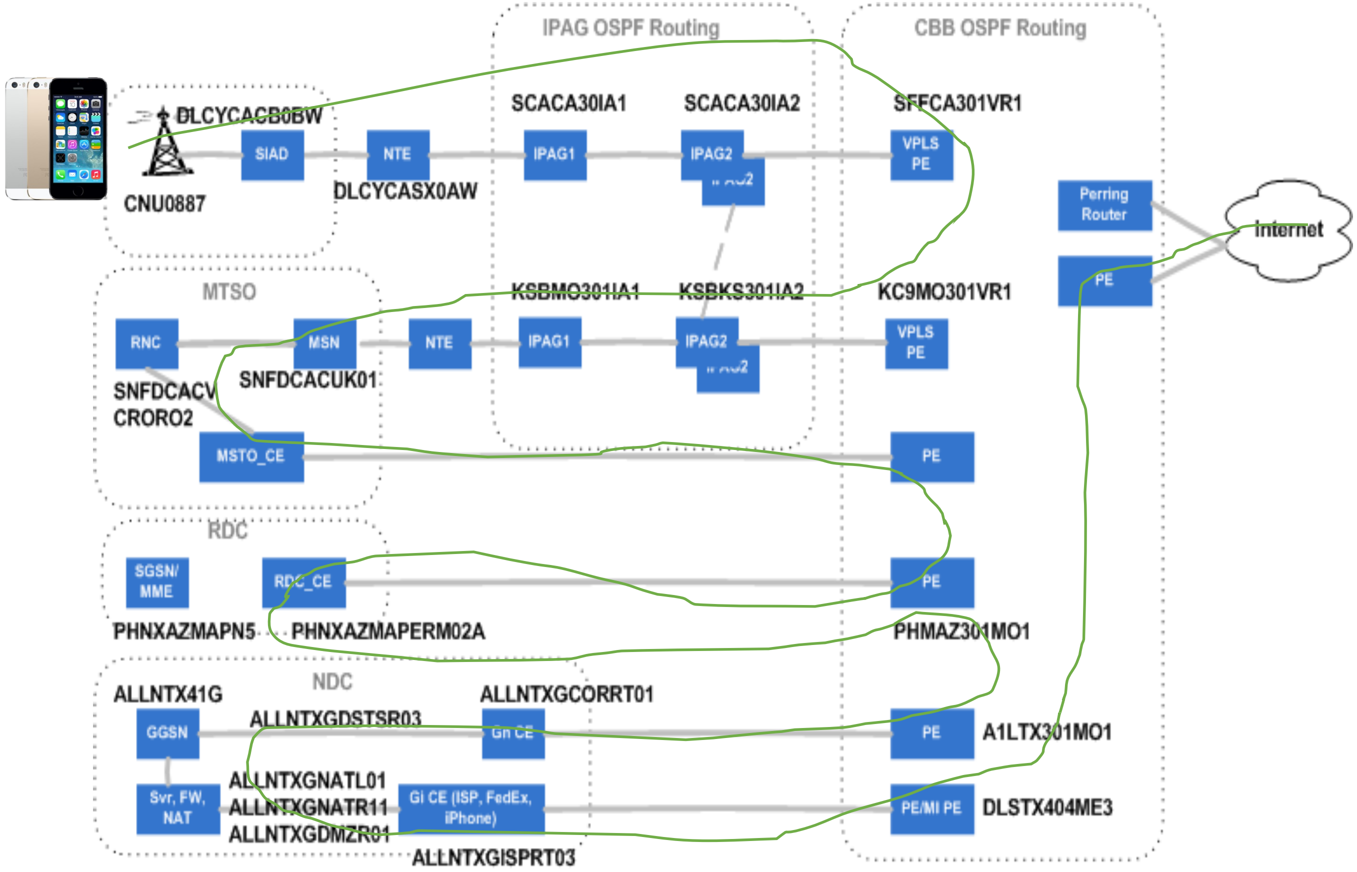
Job scheduling systems: which tasks/VMs to co-locate on same machine, which tasks to pre-empt, ...

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

Outline

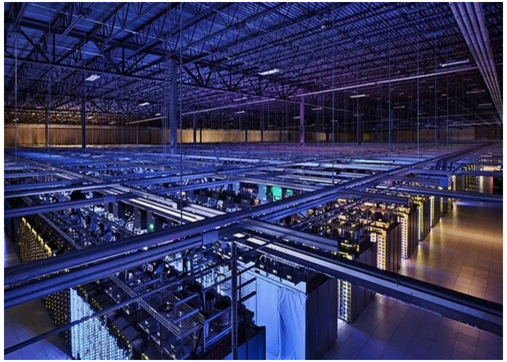
- IT Operations (Ops) background
- *Is artificial inteligence necessary for Ops?*
- Case Study Overview
  - Unsupervised Anomaly Detection in Ops
- Lessons Learned

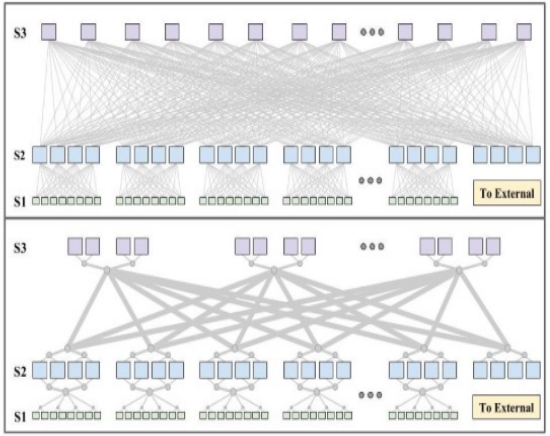
Complex Edge Networks

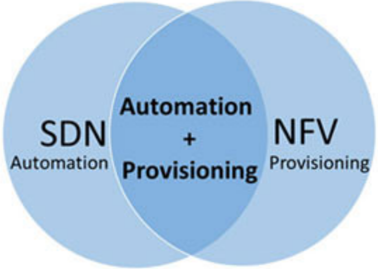


Complex and Evolving Data Center Hardwares

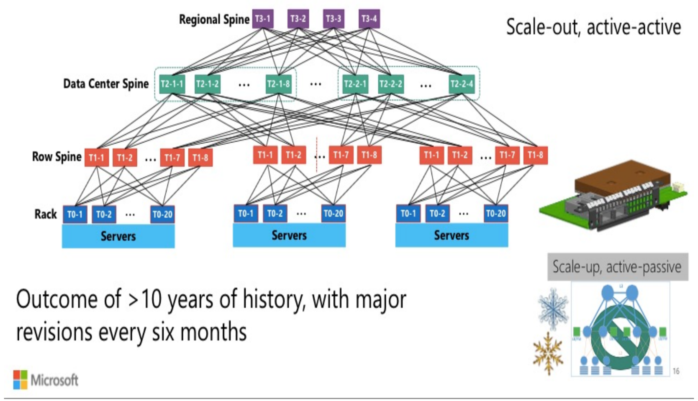
10s of thousands of servers







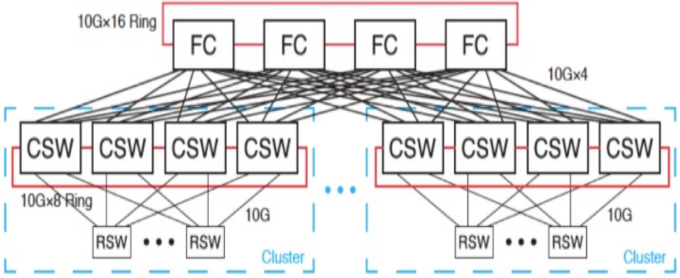
Frequent topology changes



Scale-out, active-active

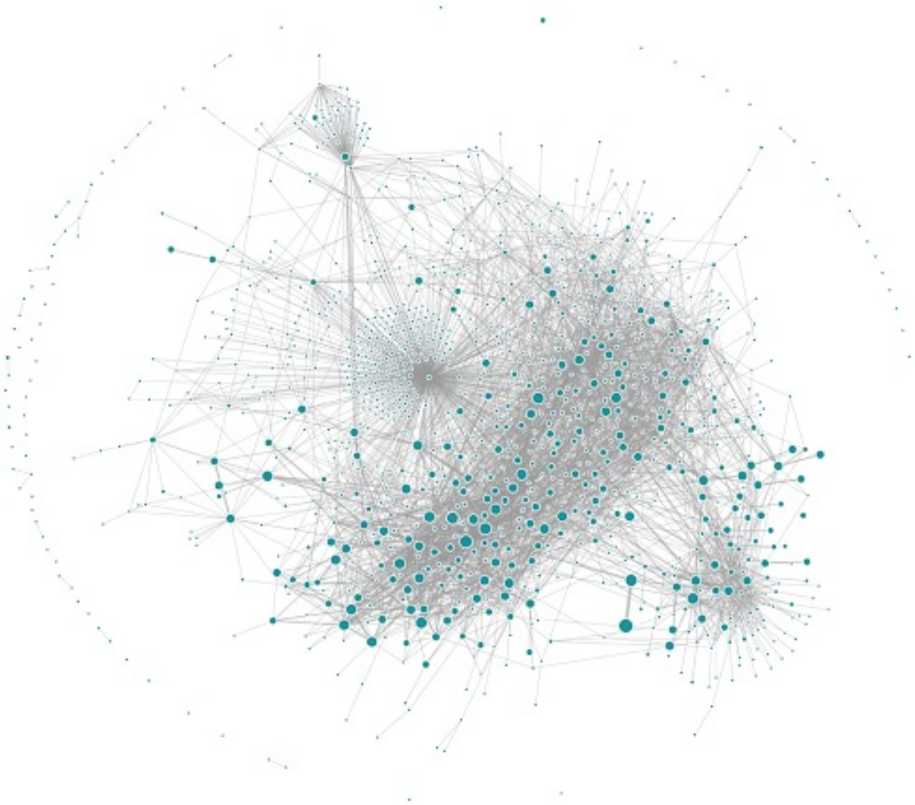
Scale-up, active-passive

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



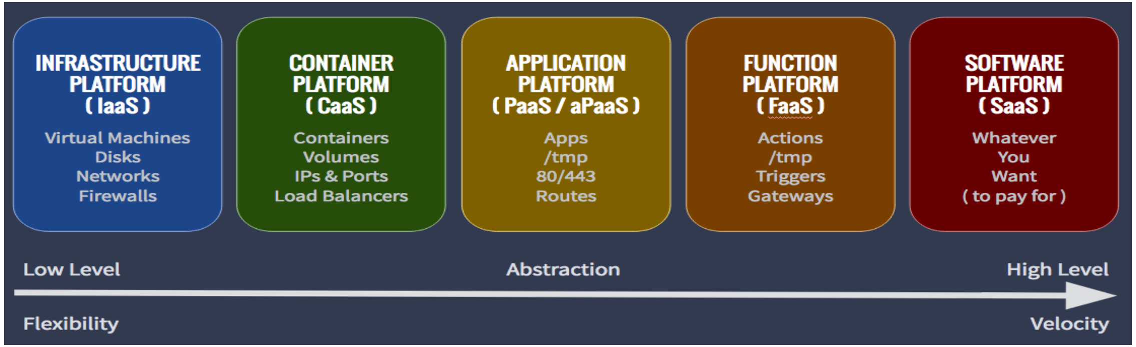
Complex Software Module Dependences

Application dependency at Uber in 2018

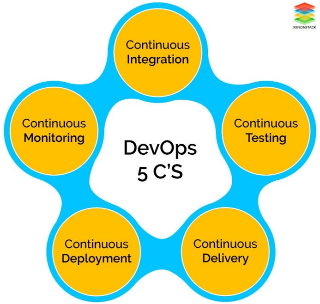


Evolving Techniques Enable Frequent Software Changes, one major cause of failures

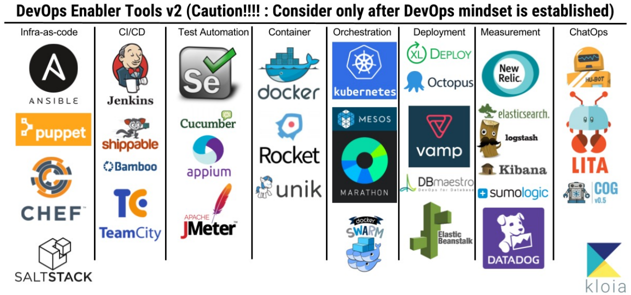
10s of thousands software/config changes per day in a large company



INFRASTRUCTURE PLATFORM (IaaS)	CONTAINER PLATFORM (CaaS)	APPLICATION PLATFORM (PaaS / aPaaS)	FUNCTION PLATFORM (FaaS)	SOFTWARE PLATFORM (SaaS)
Virtual Machines Disks Networks Firewalls	Containers Volumes IPs & Ports Load Balancers	Apps /tmp 80/443 Routes	Actions /tmp Triggers Gateways	Whatever You Want (to pay for)



DevOps

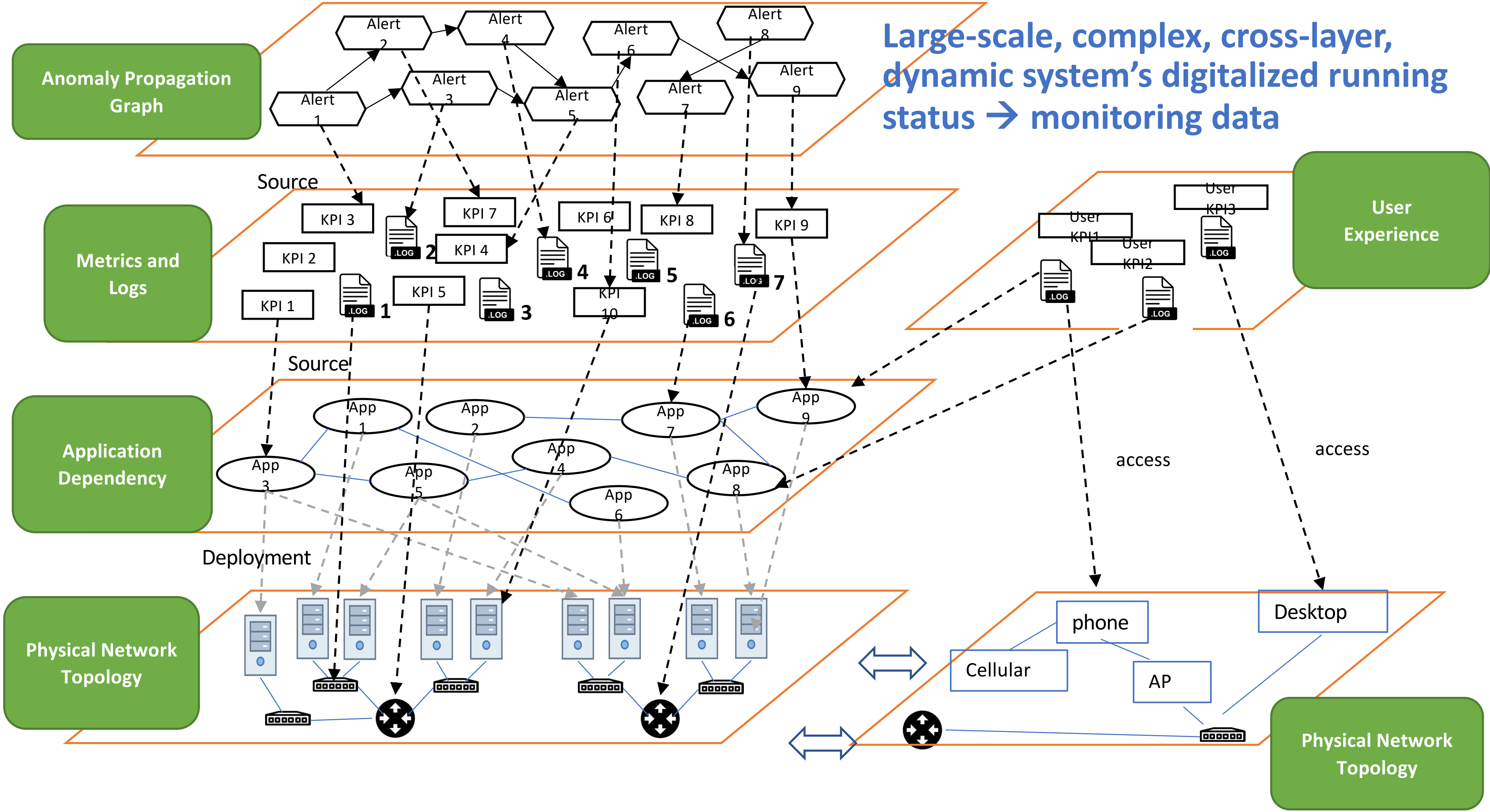


DevOps Enabler Tools v2 (Caution!!!! : Consider only after DevOps mindset is established)

Infra-as-code	CI/CD	Test Automation	Container	Orchestration	Deployment	Measurement	ChatOps

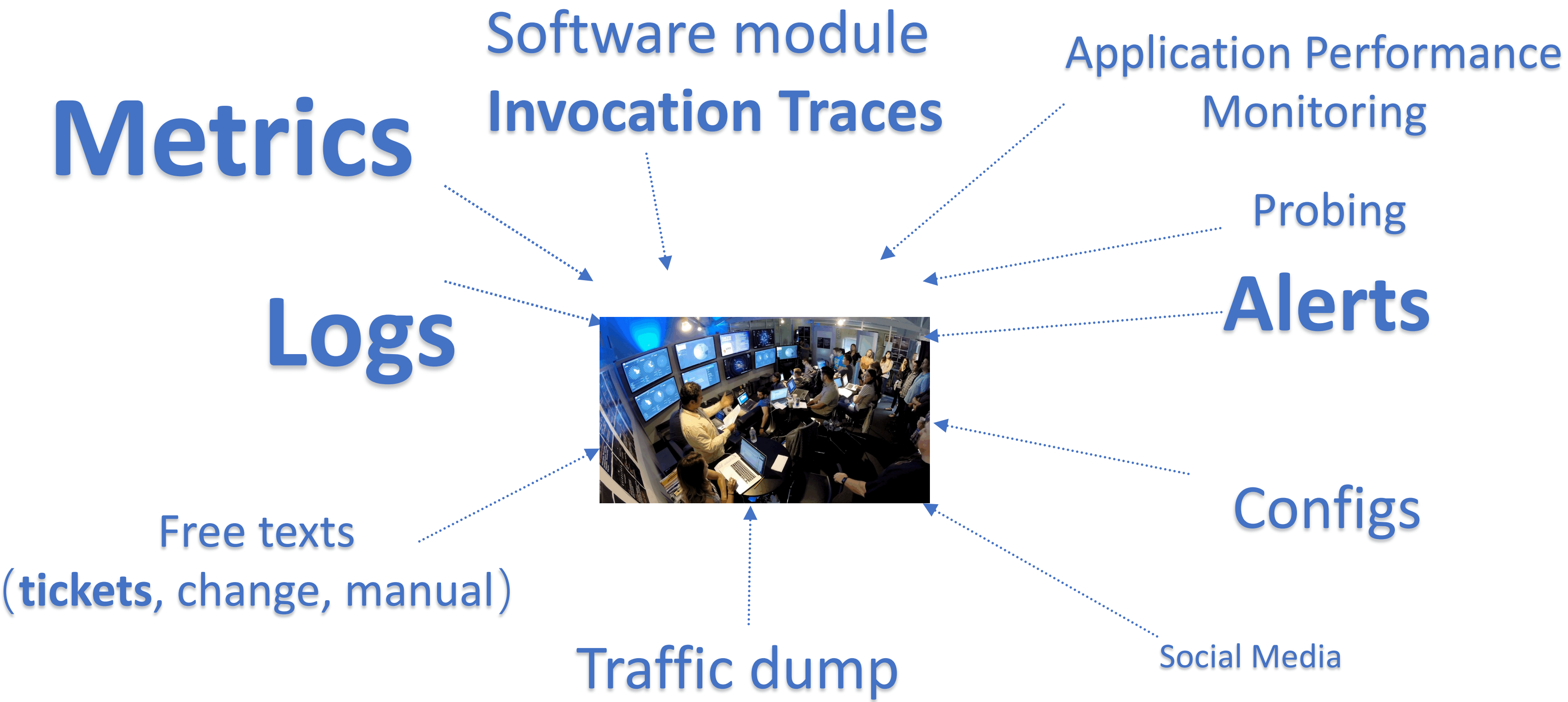
Continuous Integration/Continuous Delivery

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



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.



Software module
Invocation Traces

Application Performance
Monitoring

Metrics

Probing

Alerts

Logs

Free texts
(**tickets**, change, manual)

Configs

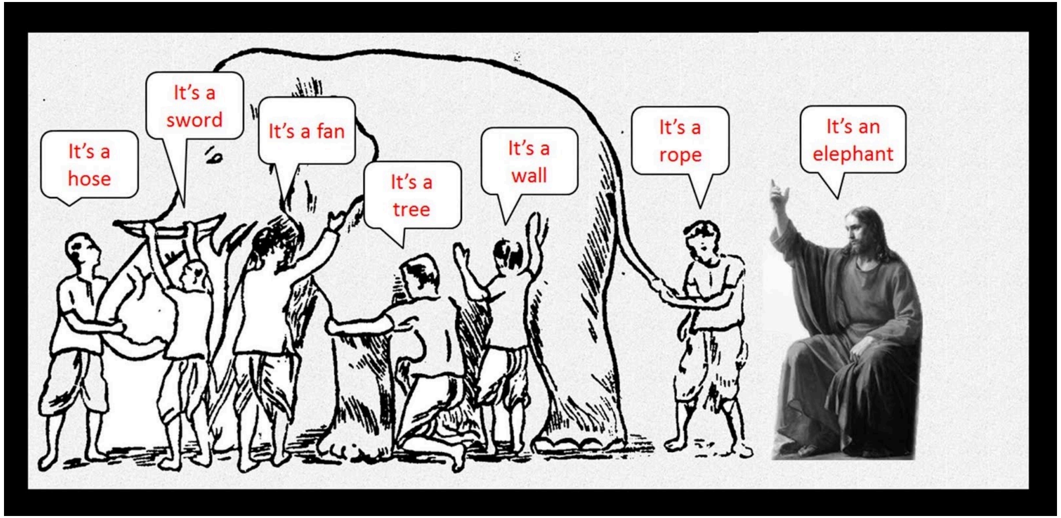
Traffic dump

Social Media

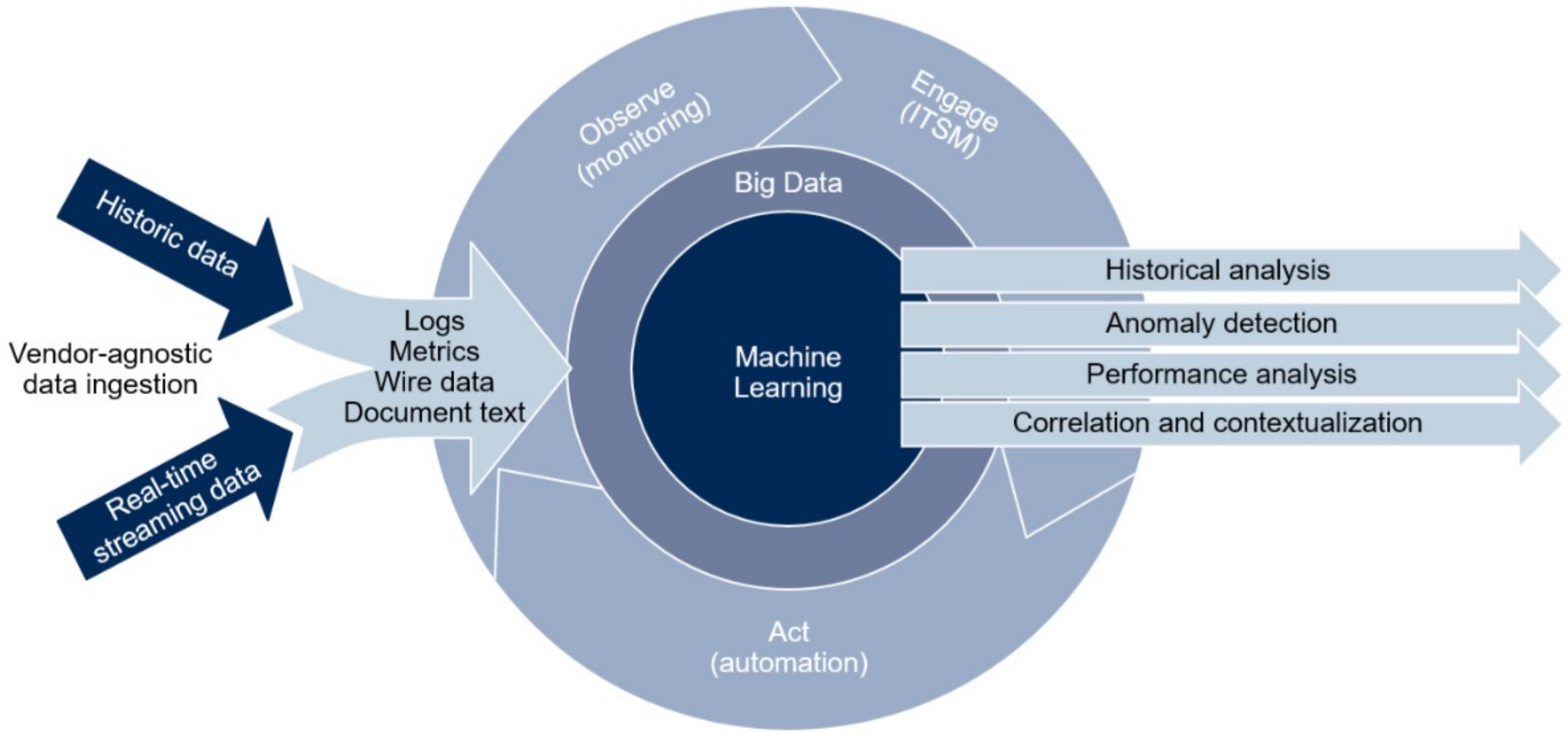
We have no choice but relying on Artificial Intelligence to extract useful signals out of the Big Ops Data which have every low signal-to-noise ratio.

- 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.

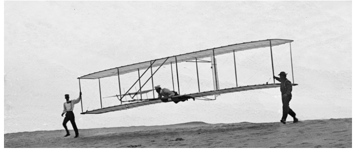


AIOps Platform Enabling Continuous ITOM



ID: 340492

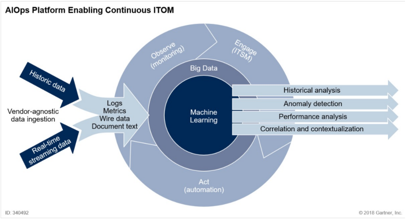
Towards Autonomous IT Operations



Manual and few data

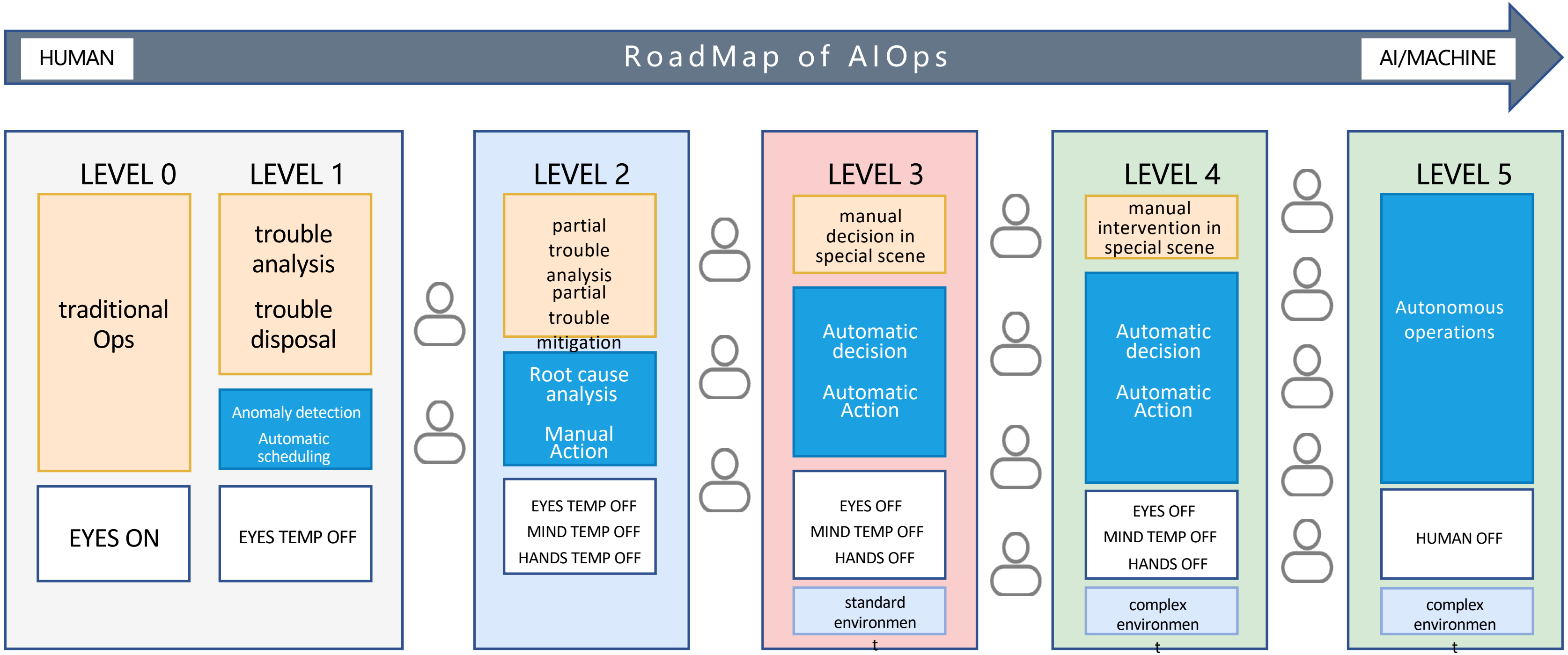
Lots of data but manual decision

Autonomous



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

Levels of AIOps



HUMAN

RoadMap of AIOps

AI/MACHINE

LEVEL 0

traditional Ops

EYES ON

LEVEL 1

trouble analysis

trouble disposal

Anomaly detection

Automatic scheduling

EYES TEMP OFF

LEVEL 2

partial trouble analysis

partial trouble mitigation

Root cause analysis

Manual Action

EYES TEMP OFF

MIND TEMP OFF

HANDS TEMP OFF

LEVEL 3

manual decision in special scene

Automatic decision

Automatic Action

EYES OFF

MIND TEMP OFF

HANDS OFF

standard environment

LEVEL 4

manual intervention in special scene

Automatic decision

Automatic Action

EYES OFF

MIND TEMP OFF

HANDS OFF

complex environment

LEVEL 5

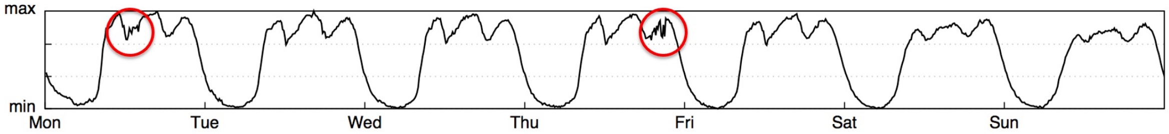
Autonomous operations

HUMAN OFF

complex environment

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

阿里巴巴 Alibaba.com

Bai du 百度

Tencent 腾讯

ByteDance 字节跳动

HUAWEI

滴滴出行
滴滴一下 美好出行

中国建设银行
China Construction Bank

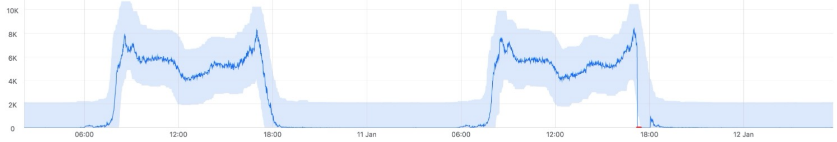
Bank
中国光大银行
CHINA EVERBRIGHT BANK

WeBank
微众银行

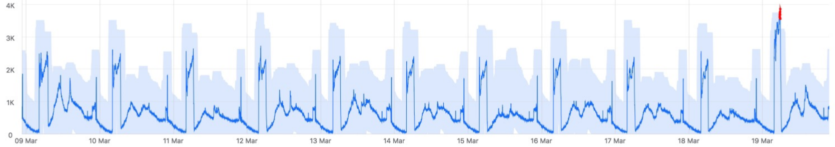
Diverse Metrics and Their Diverse Anomalies

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

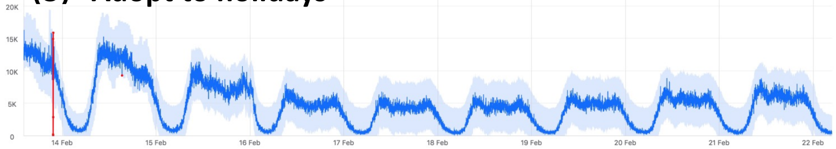
(1) Seasonal metrics



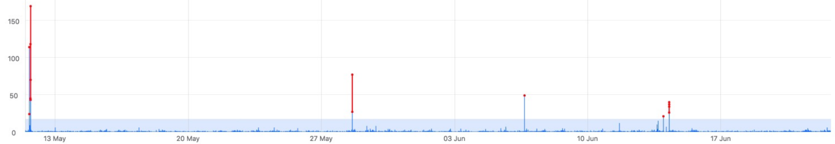
(2) Periodicity shift



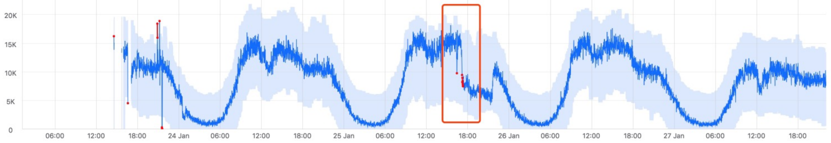
(3) Adopt to holidays



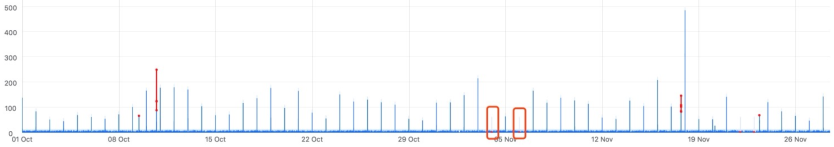
(4) Identify variable metrics and obtain extreme threshold



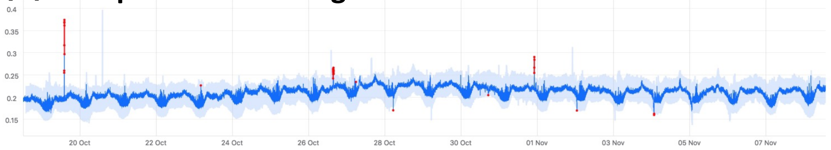
(5) Detect too rapid a change



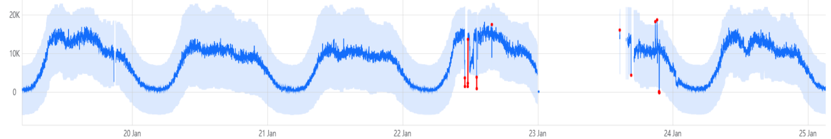
(6) Detect the lack of seasonality.



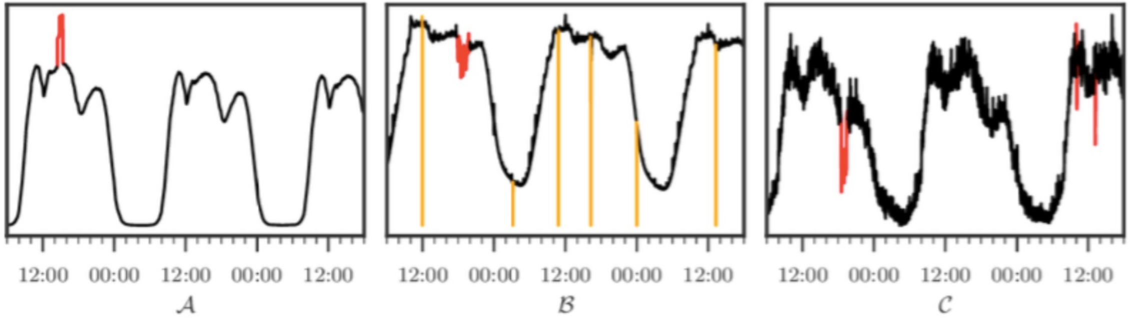
(7) Adapt to trend change

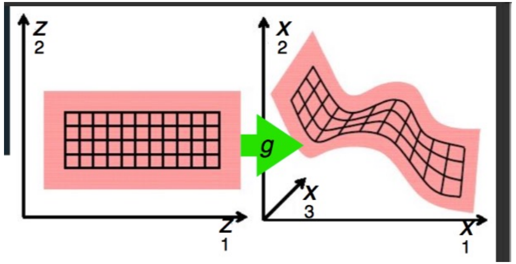


(8) Robust against data loss or interruption



Donut: supervised->unsupervised: smooth KPIs





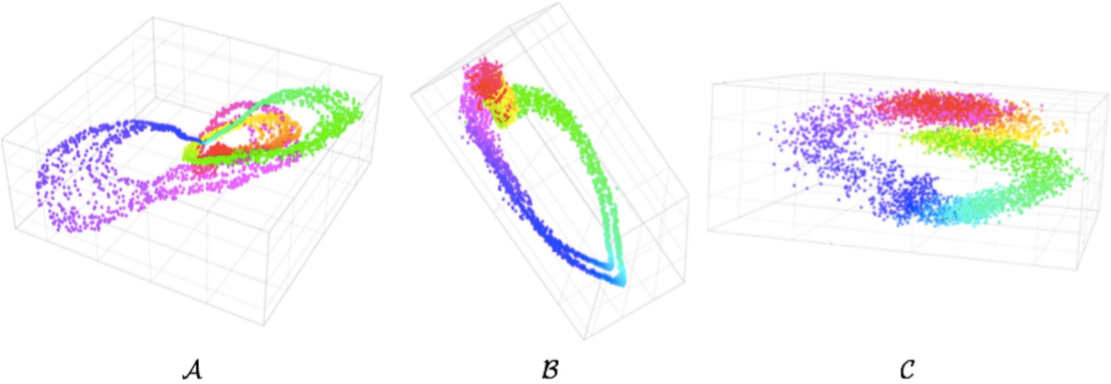
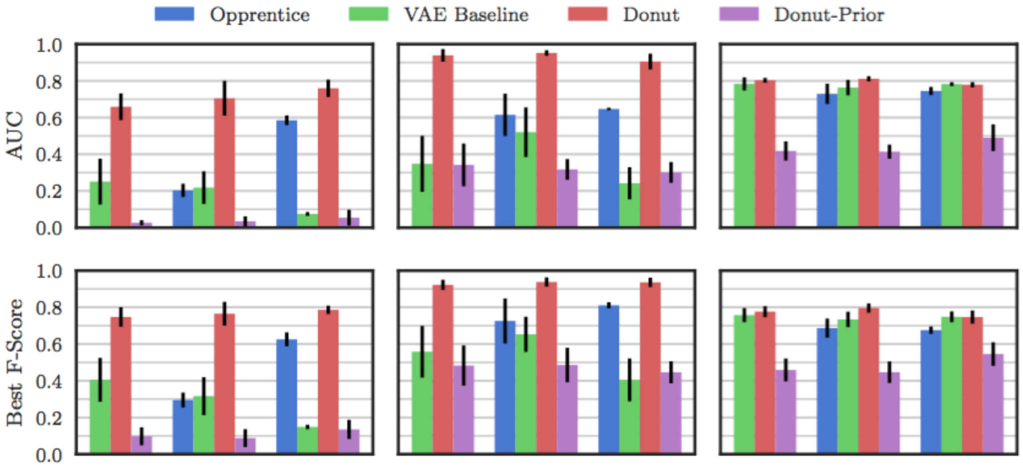


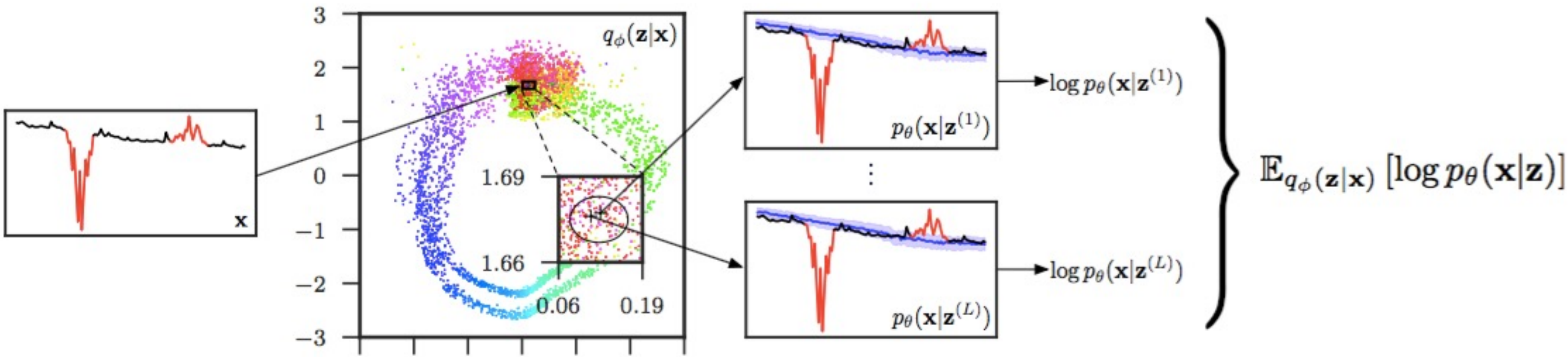
Figure 12: 3-d latent space of all three datasets.

Unsupervised KPI Anomaly Detection Through Variational Auto-Encoder

WWW2018

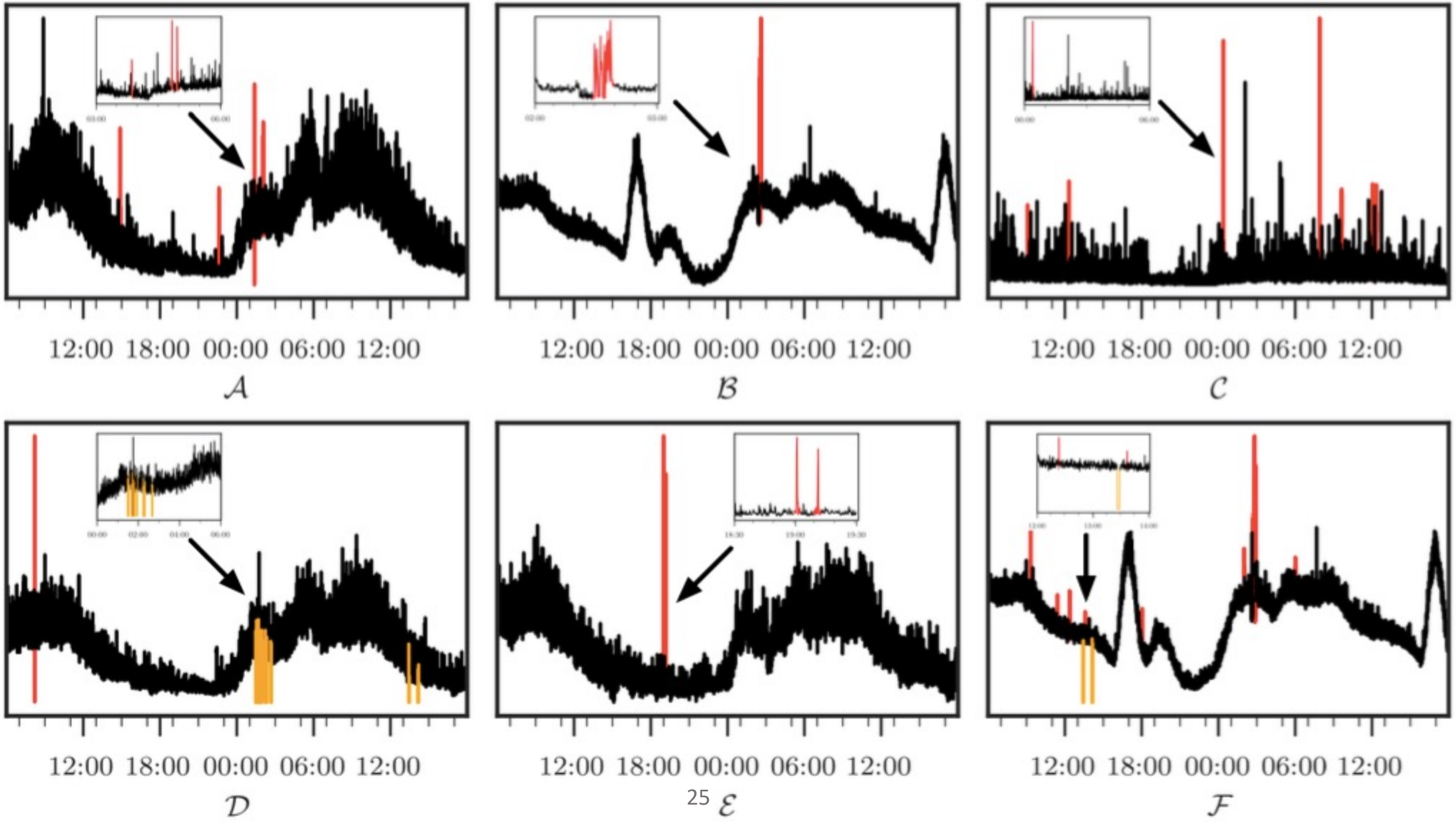
Accuracy of 0.8~0.9， even better than supervised approach.



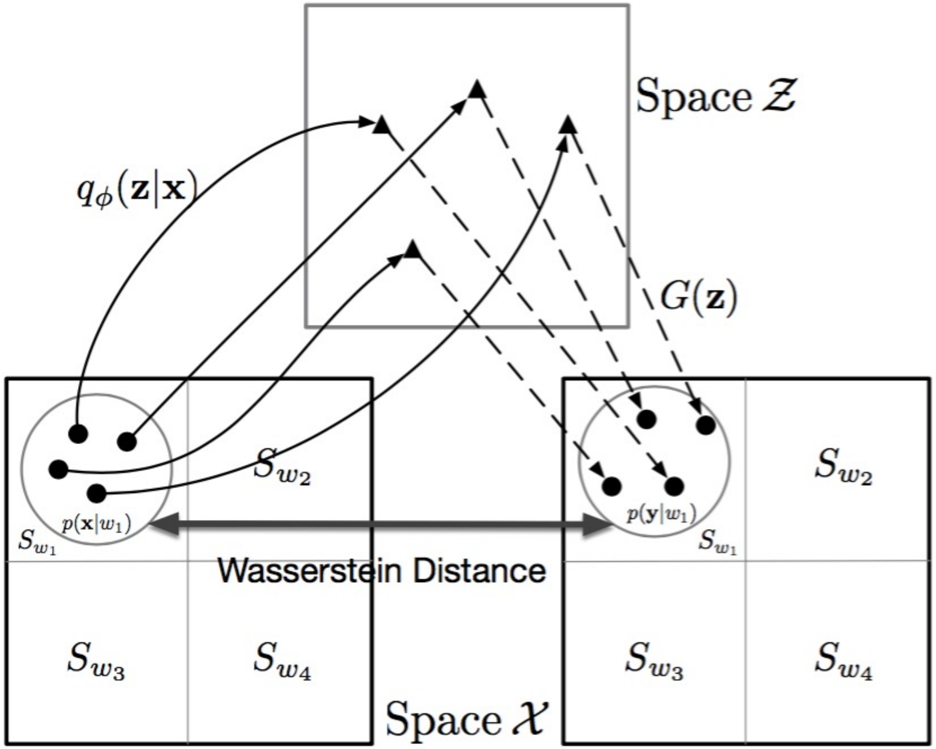


3
2
1
0
−1
−2
−3
$q_\phi(\mathbf{z}|\mathbf{x})$
1.69
1.66
0.06
0.19
$\mathbf{x}$
$p_\theta(\mathbf{x}|\mathbf{z}^{(1)})$
$p_\theta(\mathbf{x}|\mathbf{z}^{(L)})$
$\log p_\theta(\mathbf{x}|\mathbf{z}^{(1)})$
$\log p_\theta(\mathbf{x}|\mathbf{z}^{(L)})$
$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$

Buzz: Apply Adversarial Training for Non-Gaussian Noise



Unsupervised Anomaly Detection for Intricate KPIs via Adversarial Training of VAE



INFOCOM 2019

Major ideas

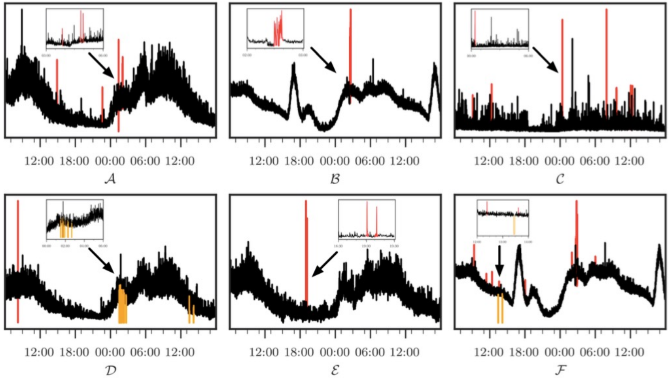
- **Wasserstein distance: the distance between the two probability distributions**

- **Partitioning from measure theory.**

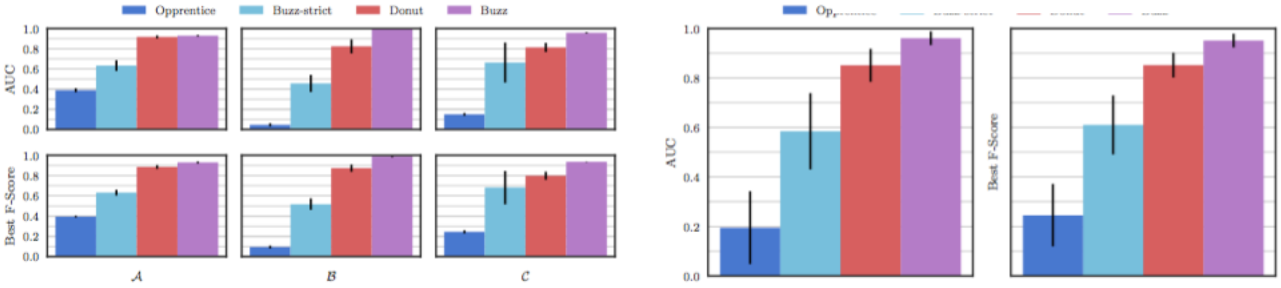
a powerful and commonly used analysis method for distribution in measure theory.

- **Adversarial Training**

Experiment Results



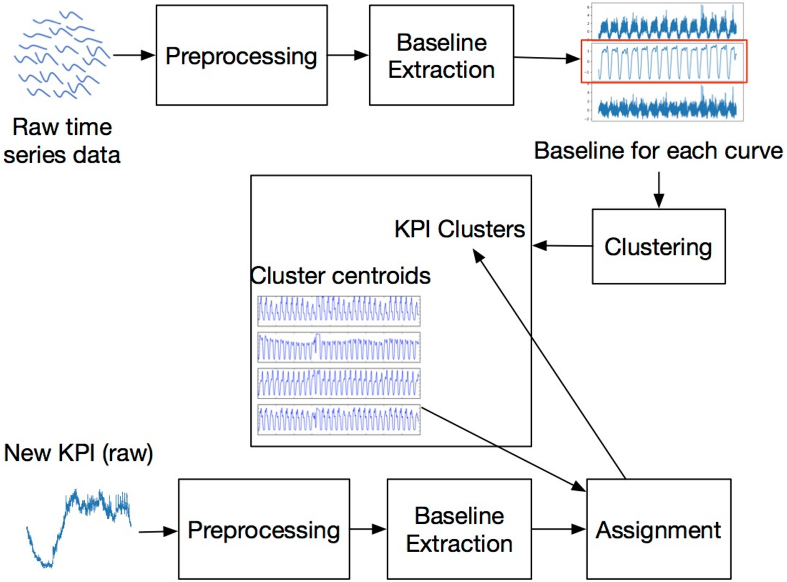
Best F-Score outperforms Donut by up to 0.15



(a) Dateset A, B, C

(b) Average of 11 KPIs

Clustering + Transfer Learning to Reduce Training Overhead



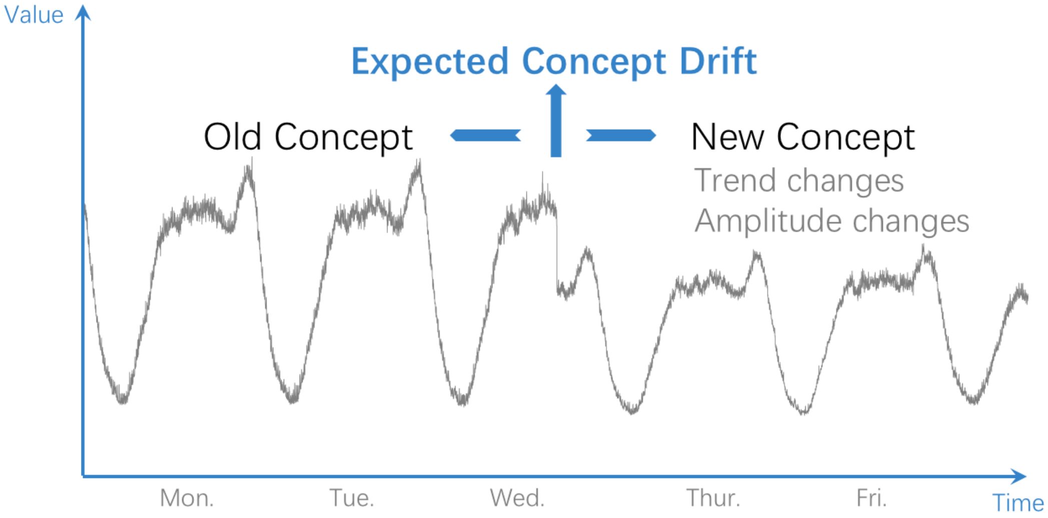
	Original DONUT [WWW2018]	ROCKA+DONUT+KPI-specific threshold
Avg. F-score	0.89	0.88
Total training time (s)	51621	5145

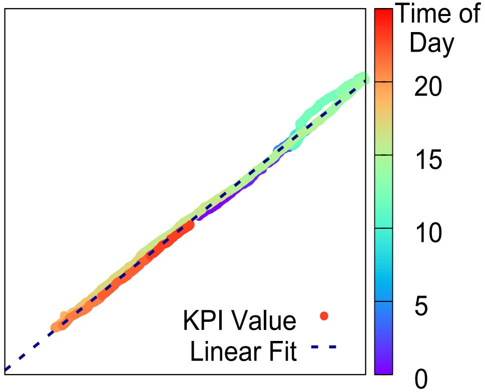
Adapt to Concept Drift

ISSRE 2018 Best Paper

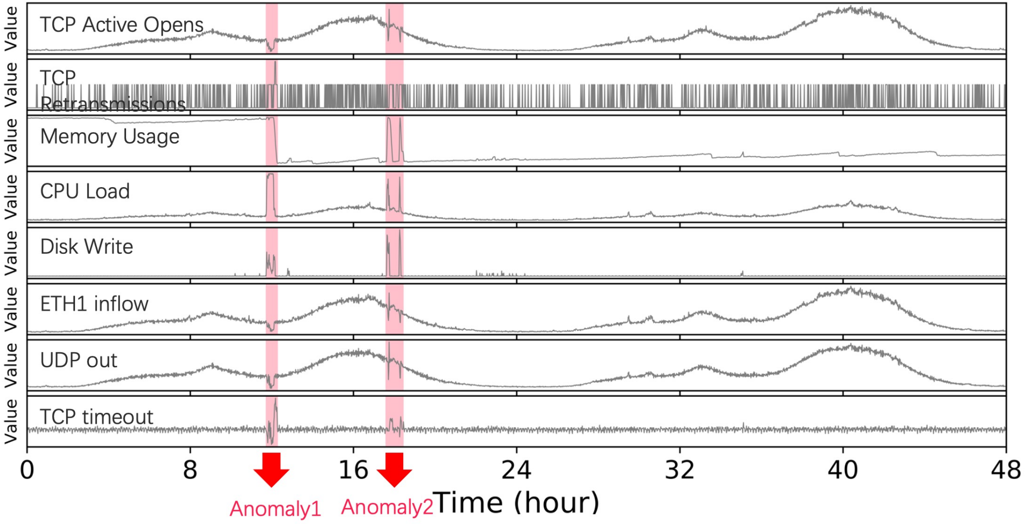
concept drift adaption improve anomaly detection F-score by 203%（**0.225 to 0.681**）

Observation: Old and New Concept Can Be Linearly Fitted

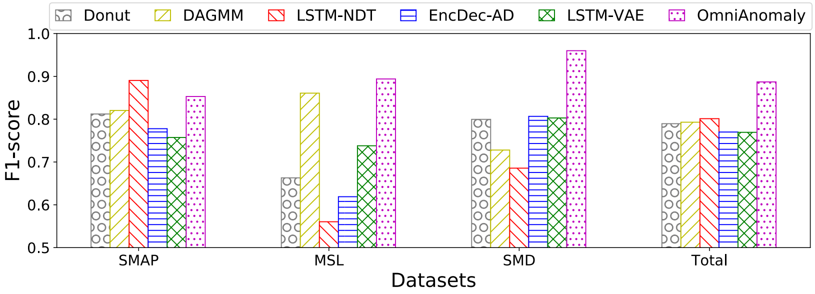




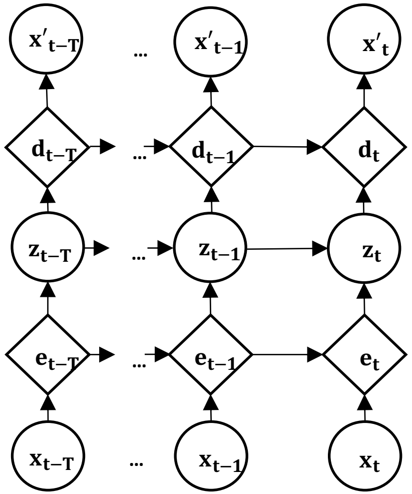
Multivariate Time Series Anomaly Detection with OmniAnomaly (KDD 2019)



F1-best of OmniAnomaly and baselines



Model Architecture of OmniAnomaly



Reconstructed data

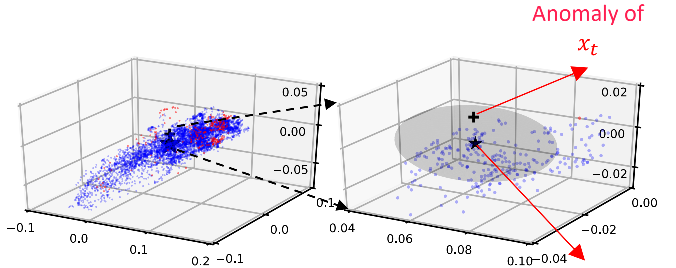
GRU cells for capturing temporal dependence

Stochastic cells for modeling data distribution

GRU cells for capturing temporal dependence

Input Sequence data

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



Anomaly of x_t

Normal data point x_t

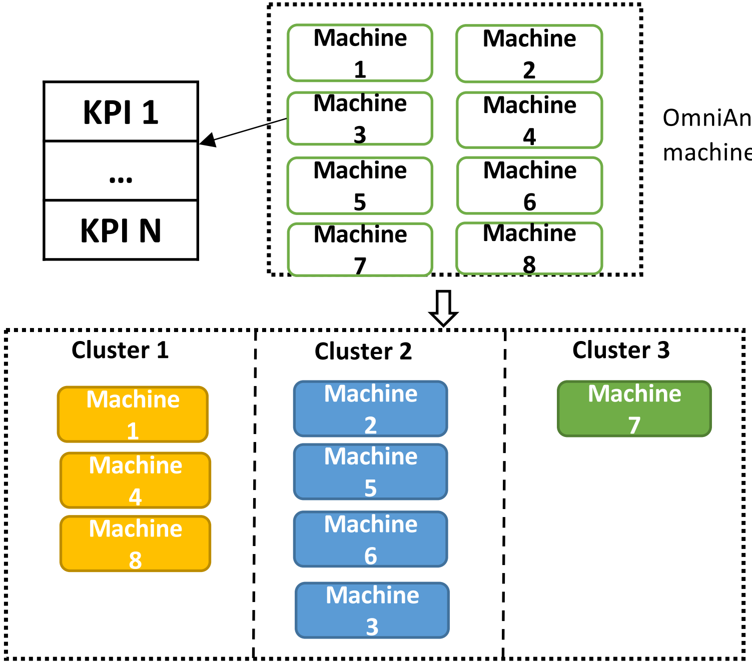
Fig: 3-dimensional z_t of x_t

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.



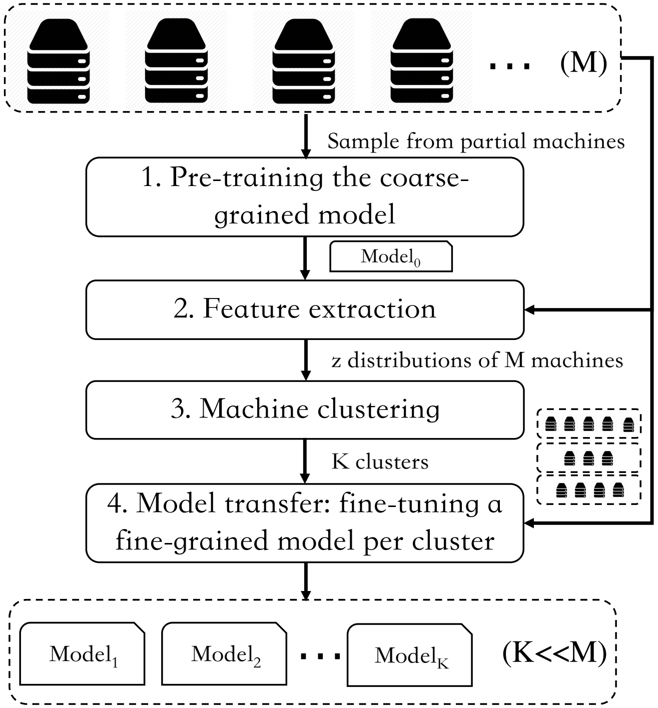
OmniAnomaly: one model for one machine (i.e., 8 models)

one model per cluster (i.e., 3 models)

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



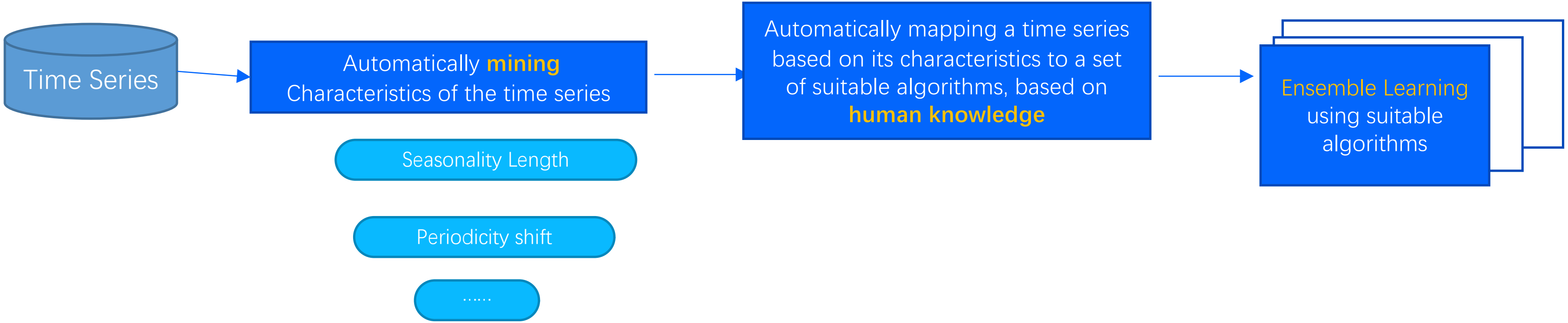
. . . (M)
Sample from partial machines
1. Pre-training the coarse-grained model
Model$_0$
2. Feature extraction
z distributions of M machines
3. Machine clustering
K clusters
4. Model transfer: fine-tuning a fine-grained model per cluster
Model$_1$
Model$_2$
. . .
Model$_K$
(K<<M)

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?



Time Series

Automatically **mining** Characteristics of the time series

Seasonality Length

Periodicity shift

……

Automatically mapping a time series based on its characteristics to a set of suitable algorithms, based on **human knowledge**

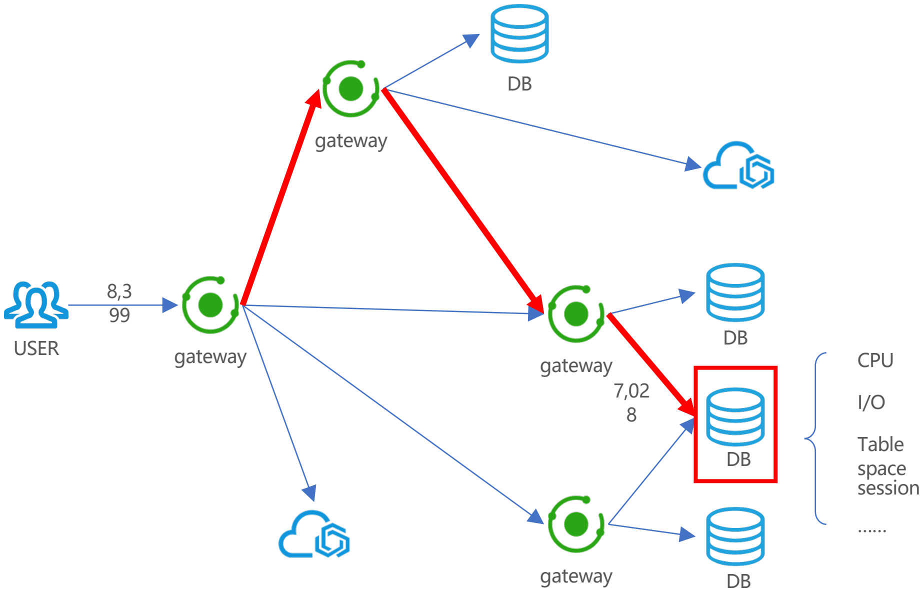
Ensemble Learning using suitable algorithms

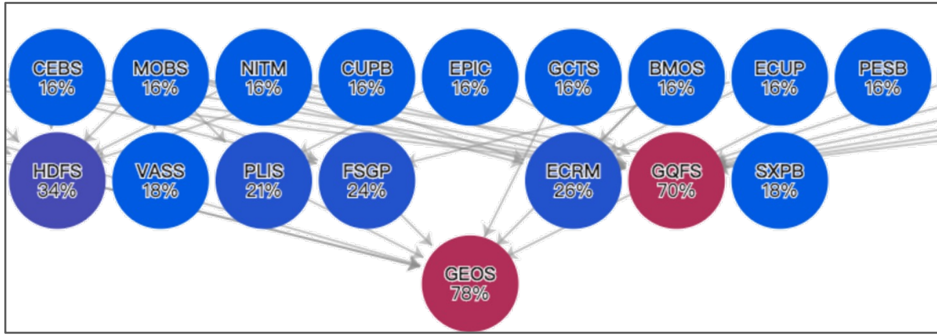
Outline

- 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

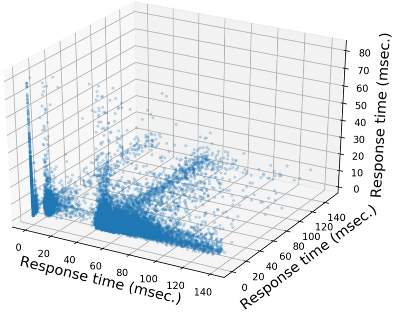
Software Module Invocation Traces

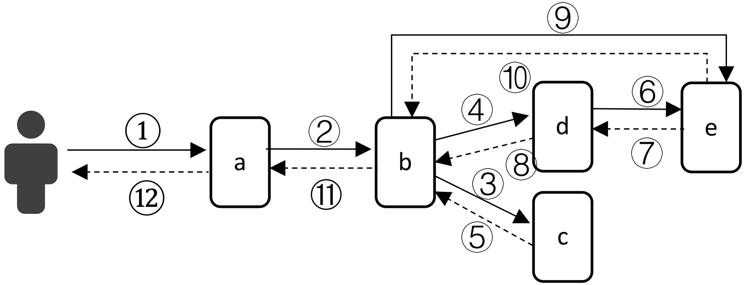
- Invocation trace: 10s~100s of module-to-module invocations for a unique transaction
  - One module failure can manifest itself cross-invocation and cross-transaction





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

Microservice s	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

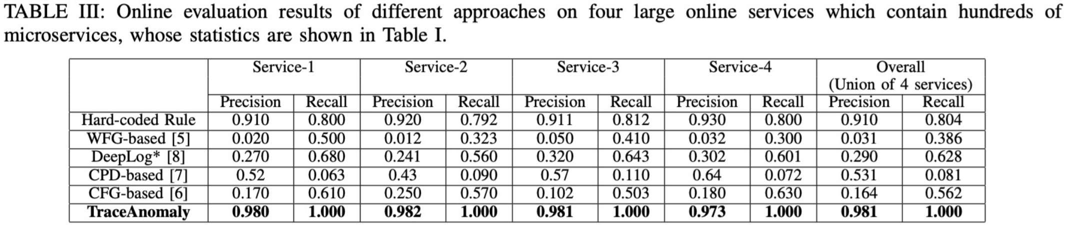
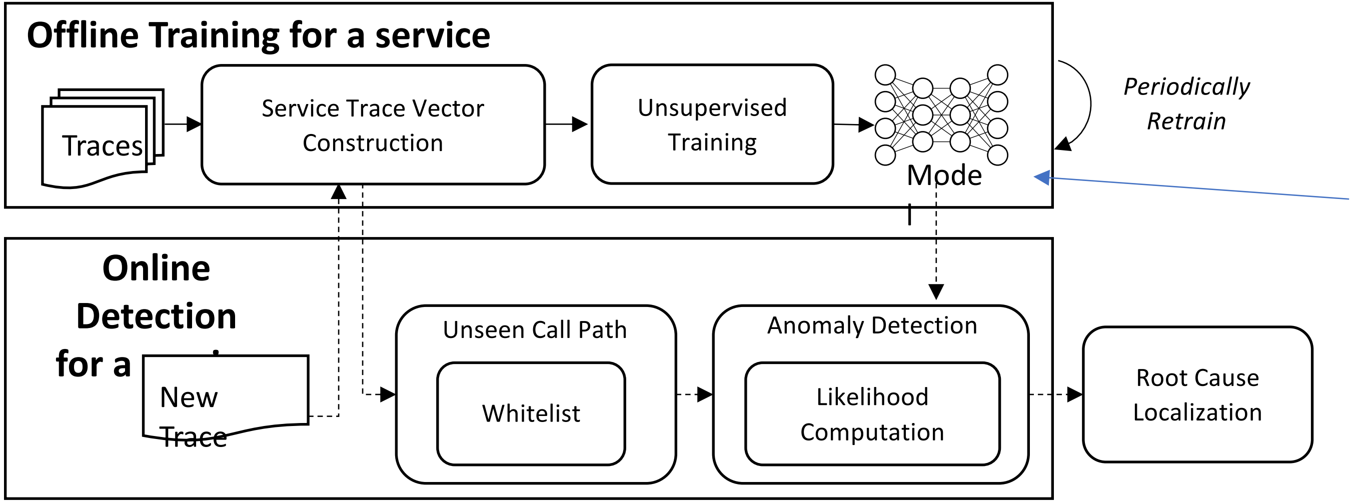
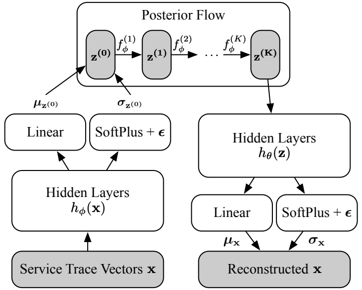


TABLE III: Online evaluation results of different approaches on four large online services which contain hundreds of microservices, whose statistics are shown in Table I.

	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**

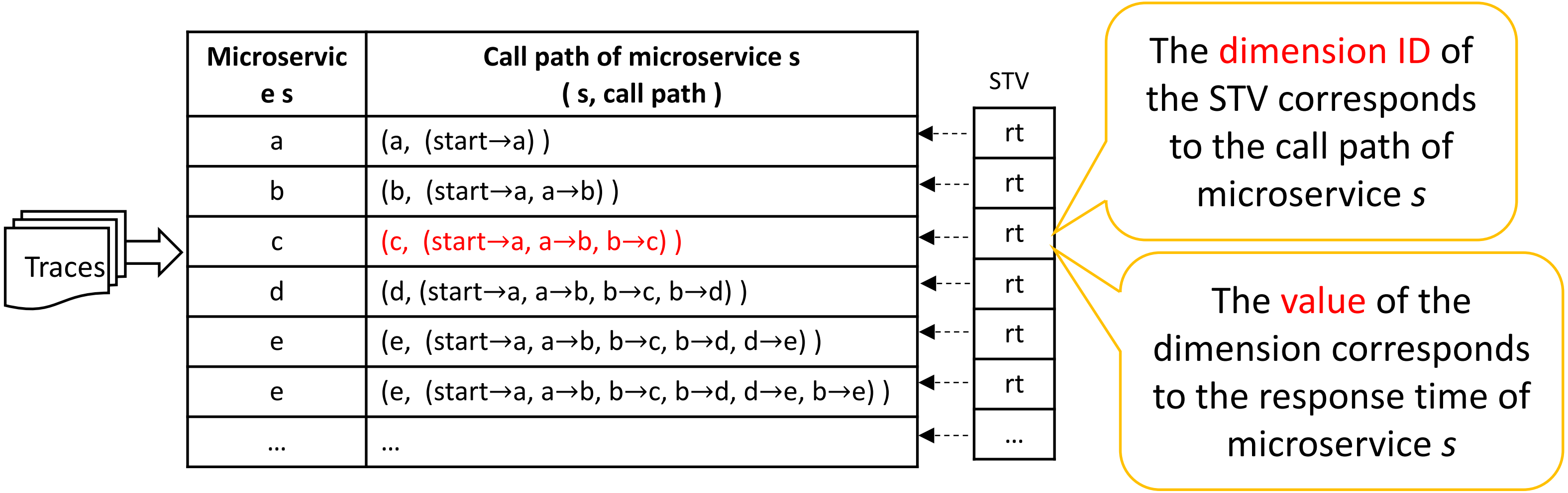




Posterior Flow
$\mathbf{z^{(0)}}$
$f_\phi^{(1)}$
$\mathbf{z^{(1)}}$
$f_\phi^{(2)}$
$f_\phi^{(K)}$
$\mathbf{z^{(K)}}$
$\boldsymbol{\mu}_{\mathbf{z}^{(0)}}$
$\boldsymbol{\sigma}_{\mathbf{z}^{(0)}}$
Linear
SoftPlus + $\boldsymbol{\epsilon}$
Hidden Layers $h_\phi(\mathbf{x})$
Service Trace Vectors $\mathbf{x}$
Hidden Layers $h_\theta(\mathbf{z})$
Linear
SoftPlus + $\boldsymbol{\epsilon}$
$\boldsymbol{\mu}_\mathbf{x}$
$\boldsymbol{\sigma}_\mathbf{x}$
Reconstructed $\mathbf{x}$

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)



Microservice s	Call path of microservice s (s, call path)
a	(a, (start→a))
b	(b, (start→a, a→b))
c	(c, (start→a, a→b, b→c))
d	(d, (start→a, a→b, b→c, b→d))
e	(e, (start→a, a→b, b→c, b→d, d→e))
e	(e, (start→a, a→b, b→c, b→d, d→e, b→e))
…	…

The dimension ID of the STV corresponds to the call path of microservice *s*

The value of the dimension corresponds to the response time of microservice *s*

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**.

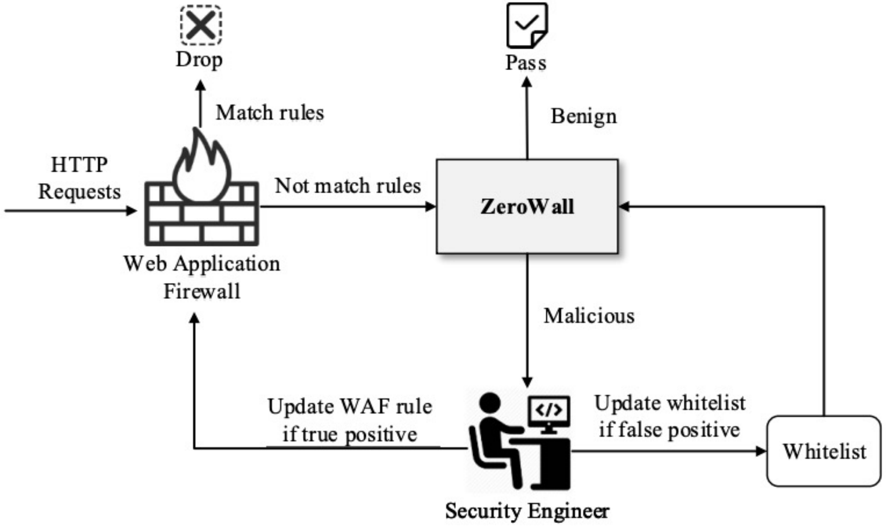
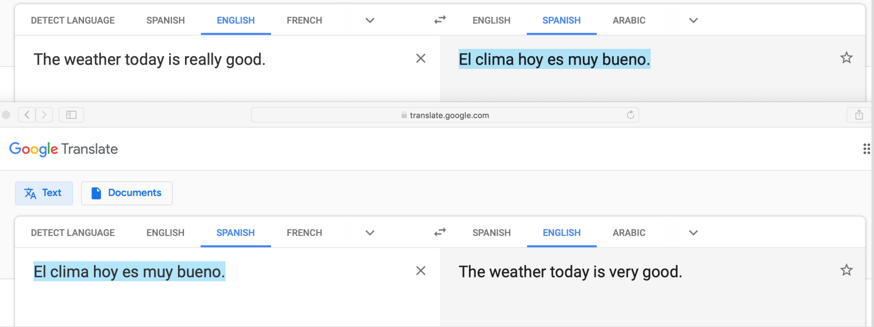


Figure 1: The workflow of *ZeroWall*.

Self-Translate Machine



The weather today is really good.

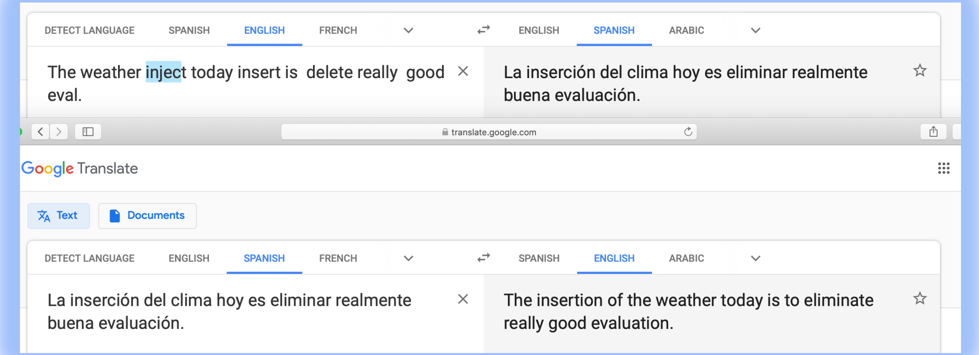
El clima hoy es muy bueno.

El clima hoy es muy bueno.

The weather today is very good.

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.



The weather inject today insert is delete really good eval.

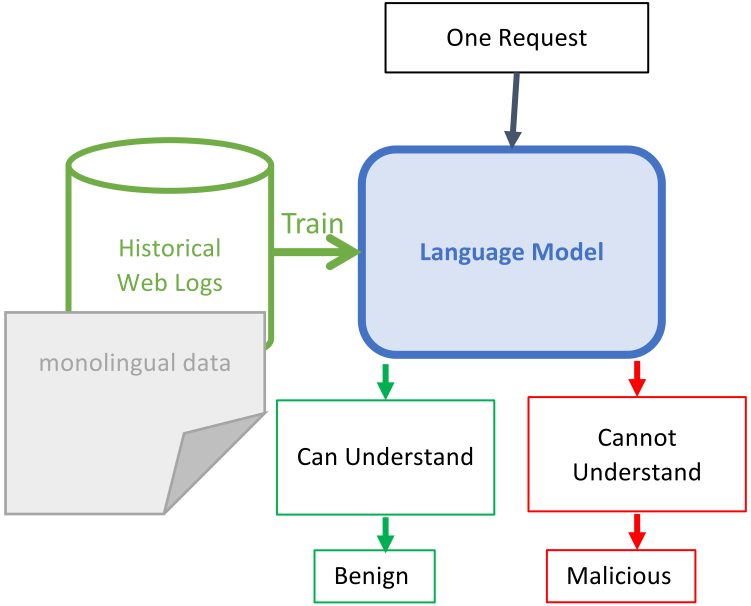
La inserción del clima hoy es eliminar realmente buena evaluación.

La inserción del clima hoy es eliminar realmente buena evaluación.

The insertion of the weather today is to eliminate really good evaluation.

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)

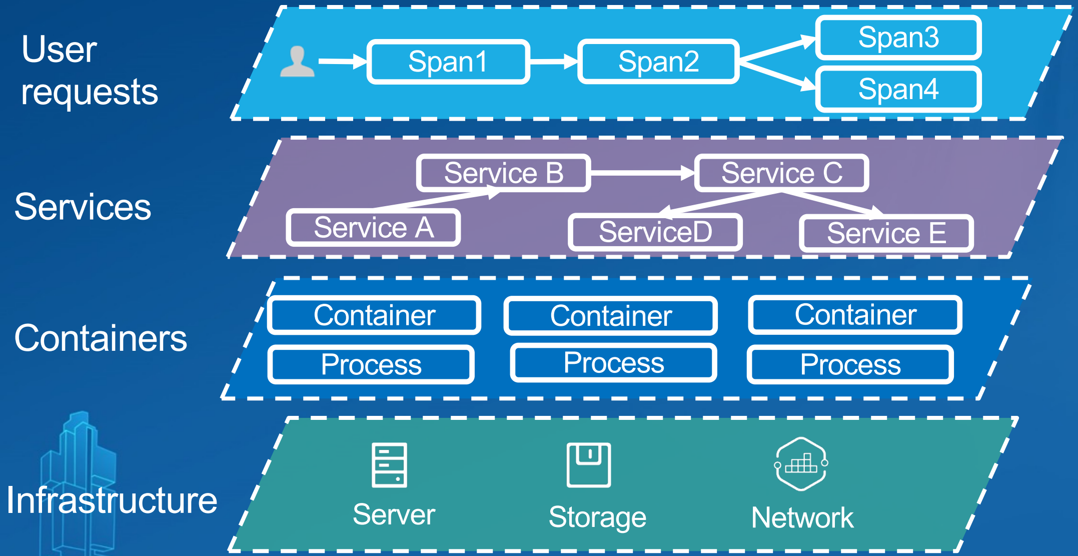
Low overhead

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.

FAILURE DIAGNOSIS AS AN IMPORTANT AIOPS SCENARIO

FAILURE DIAGNOSIS



Failure diagnosis: identifying the faulty components that caused a specific service failure.

Failure diagnosis tasks can have different localization scopes: from high-level components (e.g, faulty services) to individual failure reasons (i.e., root causes).

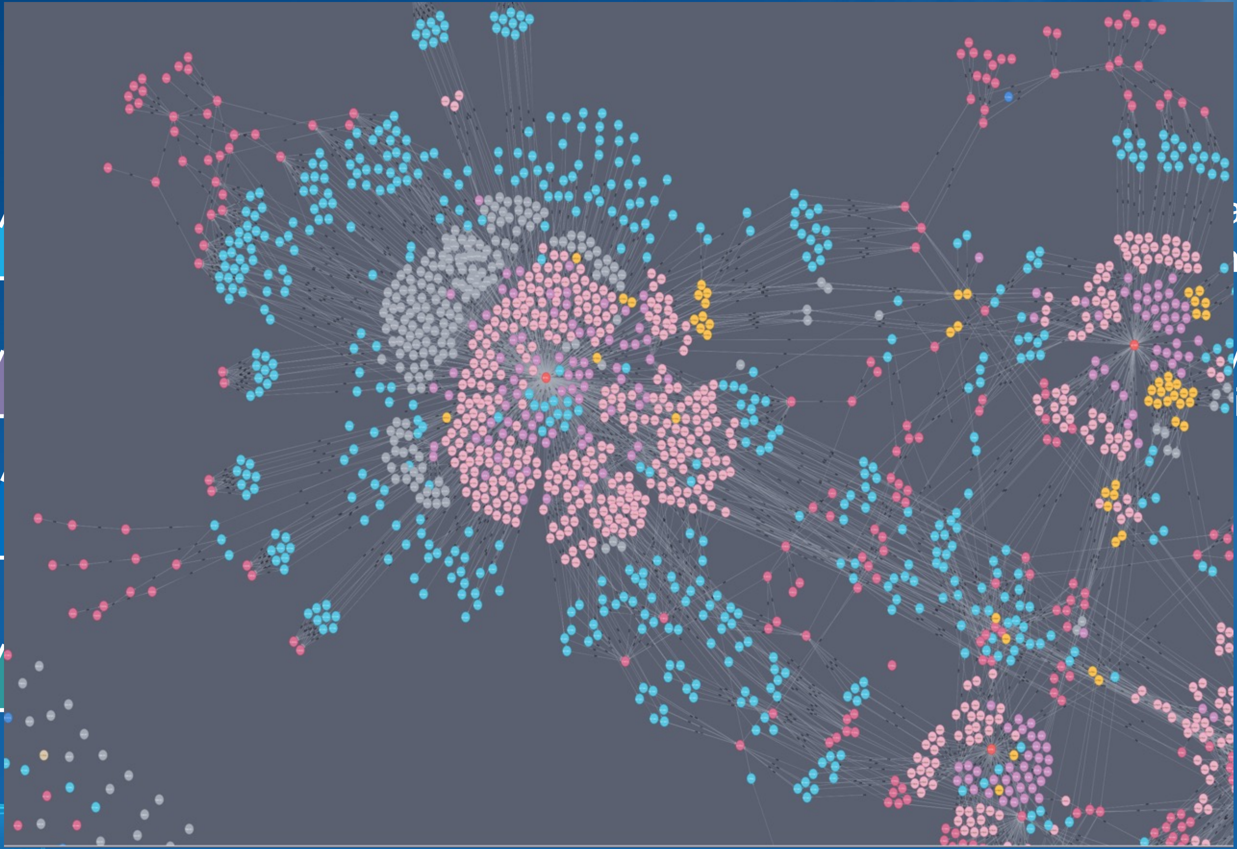
FAILURE DIAGNOSIS

User requests

Services

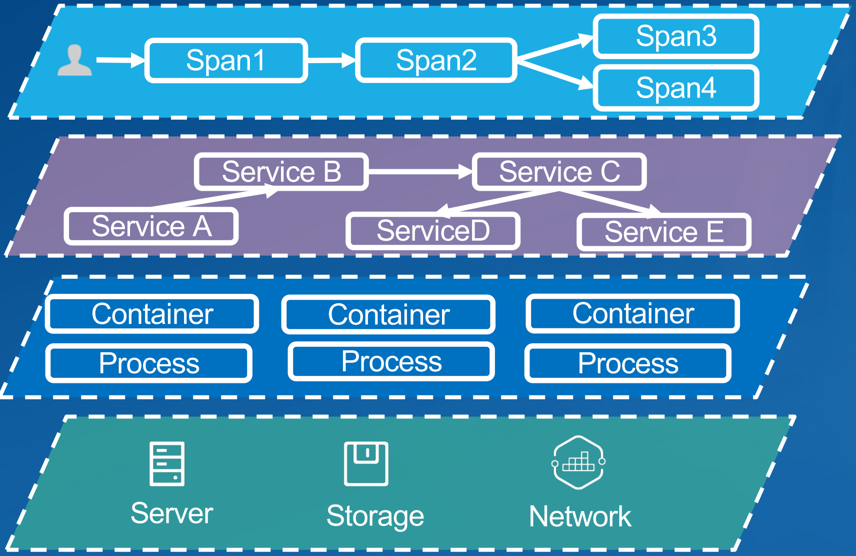
Containers

Infrastructure



and complex
nong
oud-based
ystem
iagnose.

FAILURE DIAGNOSIS



User requests

Span1 → Span2 → Span3, Span4

Services

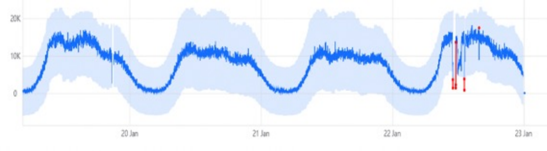
Service A, Service B, Service C, ServiceD, Service E

Containers

Container, Process
Container, Process
Container, Process

Infrastructure

Server, Storage, Network



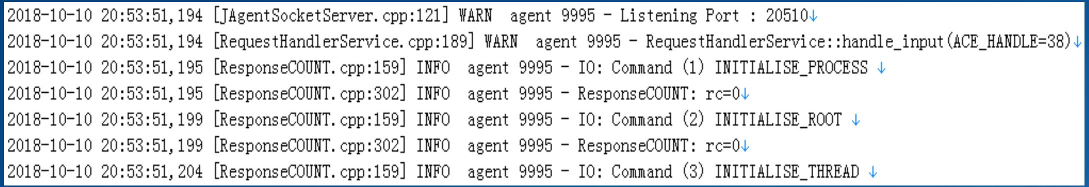
QPS
Latency

CPU Util
Memory util

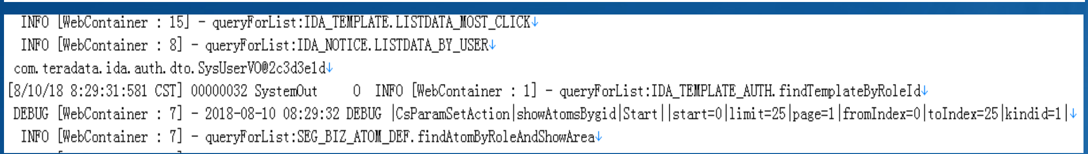
CPU Util
Thoughput
Load average

Various metrics are closely monitored on a 24×7 basis. They serve as the most direct signals to the underlying failures.

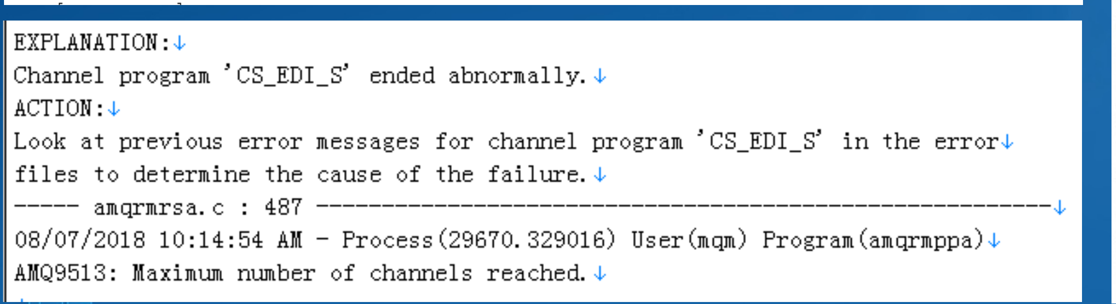
MONITORING DATA: LOGS



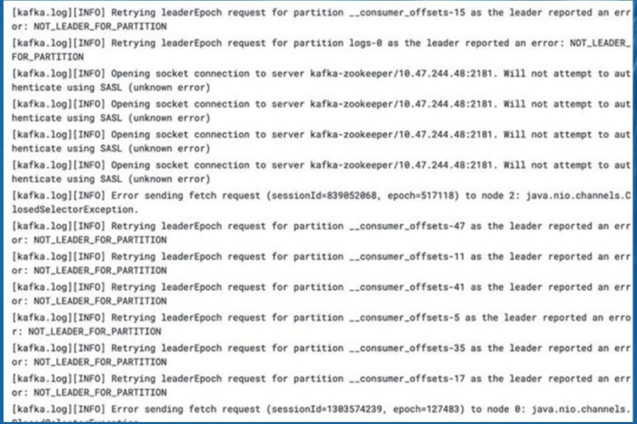
```
2018-10-10 20:53:51,194 [JAgentSocketServer.cpp:121] WARN  agent 9995 - Listening Port : 20510↓
2018-10-10 20:53:51,194 [RequestHandlerService.cpp:189] WARN  agent 9995 - RequestHandlerService::handle_input(ACE_HANDLE=38)↓
2018-10-10 20:53:51,195 [ResponseCOUNT.cpp:159] INFO  agent 9995 - IO: Command (1) INITIALISE_PROCESS ↓
2018-10-10 20:53:51,195 [ResponseCOUNT.cpp:302] INFO  agent 9995 - ResponseCOUNT: rc=0↓
2018-10-10 20:53:51,199 [ResponseCOUNT.cpp:159] INFO  agent 9995 - IO: Command (2) INITIALISE_ROOT ↓
2018-10-10 20:53:51,199 [ResponseCOUNT.cpp:302] INFO  agent 9995 - ResponseCOUNT: rc=0↓
2018-10-10 20:53:51,204 [ResponseCOUNT.cpp:159] INFO  agent 9995 - IO: Command (3) INITIALISE_THREAD ↓
```



```
  INFO [WebContainer : 15] - queryForList:IDA_TEMPLATE.LISTDATA_MOST_CLICK↓
  INFO [WebContainer : 8] - queryForList:IDA_NOTICE.LISTDATA_BY_USER↓
 com.teradata.ida.auth.dto.SysUserVO@2c3d3e1d↓
[8/10/18 8:29:31:581 CST] 00000032 SystemOut     0  INFO [WebContainer : 1] - queryForList:IDA_TEMPLATE_AUTH.findTemplateByRoleId↓
 DEBUG [WebContainer : 7] - 2018-08-10 08:29:32 DEBUG |CsParamSetAction|showAtomsBygid|Start||start=0|limit=25|page=1|fromIndex=0|toIndex=25|kindid=1|↓
  INFO [WebContainer : 7] - queryForList:SEG_BIZ_ATOM_DEF.findAtomByRoleAndShowArea↓
```



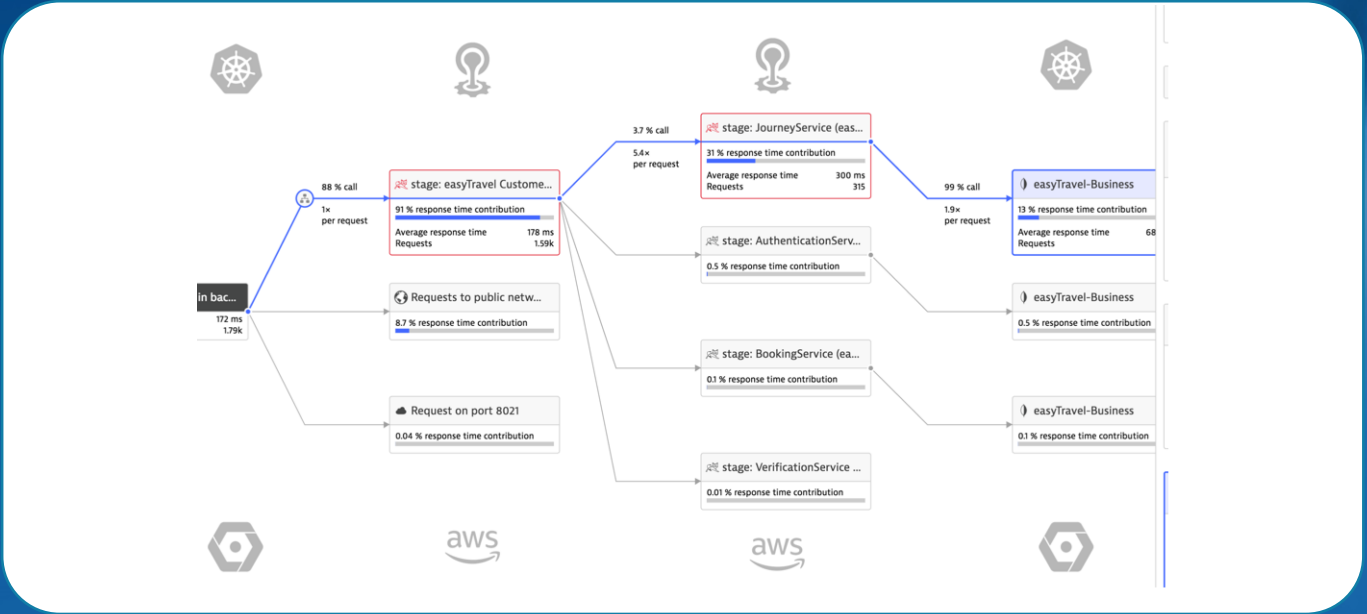
```
EXPLANATION:↓
Channel program 'CS_EDI_S' ended abnormally.↓
ACTION:↓
Look at previous error messages for channel program 'CS_EDI_S' in the error↓
files to determine the cause of the failure.↓
----- amqrmrsa.c : 487 -------------------------------------------------------↓
08/07/2018 10:14:54 AM - Process(29670.329016) User(mqm) Program(amqrmppa)↓
AMQ9513: Maximum number of channels reached.↓
```



```
[kafka.log][INFO] Retrying leaderEpoch request for partition __consumer_offsets-15 as the leader reported an err
or: NOT_LEADER_FOR_PARTITION
[kafka.log][INFO] Retrying leaderEpoch request for partition logs-0 as the leader reported an error: NOT_LEADER_
FOR_PARTITION
[kafka.log][INFO] Opening socket connection to server kafka-zookeeper/10.47.244.48:2181. Will not attempt to aut
henticate using SASL (unknown error)
[kafka.log][INFO] Opening socket connection to server kafka-zookeeper/10.47.244.48:2181. Will not attempt to aut
henticate using SASL (unknown error)
[kafka.log][INFO] Opening socket connection to server kafka-zookeeper/10.47.244.48:2181. Will not attempt to aut
henticate using SASL (unknown error)
[kafka.log][INFO] Opening socket connection to server kafka-zookeeper/10.47.244.48:2181. Will not attempt to aut
henticate using SASL (unknown error)
[kafka.log][INFO] Error sending fetch request (sessionId=839052068, epoch=517118) to node 2: java.nio.channels.C
losedSelectorException.
[kafka.log][INFO] Retrying leaderEpoch request for partition __consumer_offsets-47 as the leader reported an err
or: NOT_LEADER_FOR_PARTITION
[kafka.log][INFO] Retrying leaderEpoch request for partition __consumer_offsets-11 as the leader reported an err
or: NOT_LEADER_FOR_PARTITION
[kafka.log][INFO] Retrying leaderEpoch request for partition __consumer_offsets-41 as the leader reported an err
or: NOT_LEADER_FOR_PARTITION
[kafka.log][INFO] Retrying leaderEpoch request for partition __consumer_offsets-5 as the leader reported an erro
r: NOT_LEADER_FOR_PARTITION
[kafka.log][INFO] Retrying leaderEpoch request for partition __consumer_offsets-35 as the leader reported an err
or: NOT_LEADER_FOR_PARTITION
[kafka.log][INFO] Retrying leaderEpoch request for partition __consumer_offsets-17 as the leader reported an err
or: NOT_LEADER_FOR_PARTITION
[kafka.log][INFO] Error sending fetch request (sessionId=1303574239, epoch=127483) to node 0: java.nio.channels.
```

Logs contain detailed information but usually generated in an arbitrary manner.

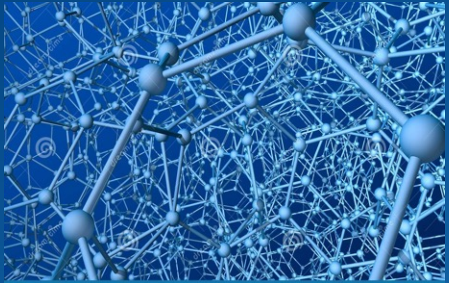
MONITORING DATA: TRACES


in bac...
172 ms
1.79k
88 % call
1×
per request
stage: easyTravel Custome...
91 % response time contribution
Average response time 178 ms
Requests 1.59k
Requests to public netw...
8.7 % response time contribution
Request on port 8021
0.04 % response time contribution
3.7 % call
5.4×
per request
stage: JourneyService (eas...
31 % response time contribution
Average response time 300 ms
Requests 315
stage: AuthenticationServ...
0.5 % response time contribution
stage: BookingService (ea...
0.1 % response time contribution
stage: VerificationService ...
0.01 % response time contribution
99 % call
1.9×
per request
easyTravel-Business
13 % response time contribution
Average response time 68
Requests
easyTravel-Business
0.5 % response time contribution
easyTravel-Business
0.1 % response time contribution
aws
aws

Traces profile and monitor applications by recording the execution process of a user request as it flows through services/microservices in adistributed system.

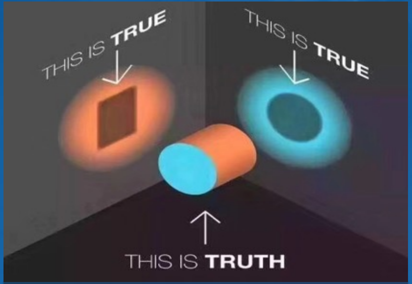
CHALLENGES

Complexity

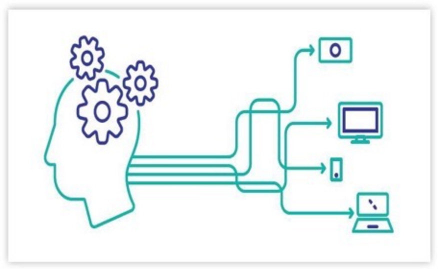


Noises

Overall judgement



Explanability



FAILURE DIAGNOSIS FRAMEWORK

FAILURE DIAGNOSIS FRAMEWORK

Goal

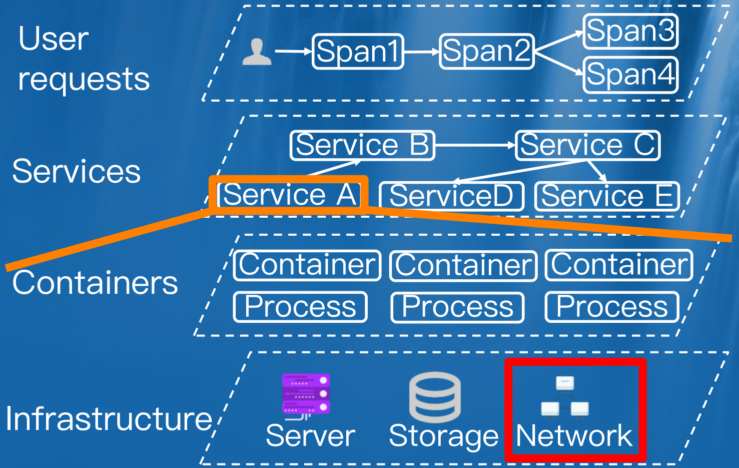
Failure Diagnosis in the Cloud

Framework

Layer-by-layer localization

Application-level localization

Infrastructure-level localization



- Divide the problem to reduce complexity

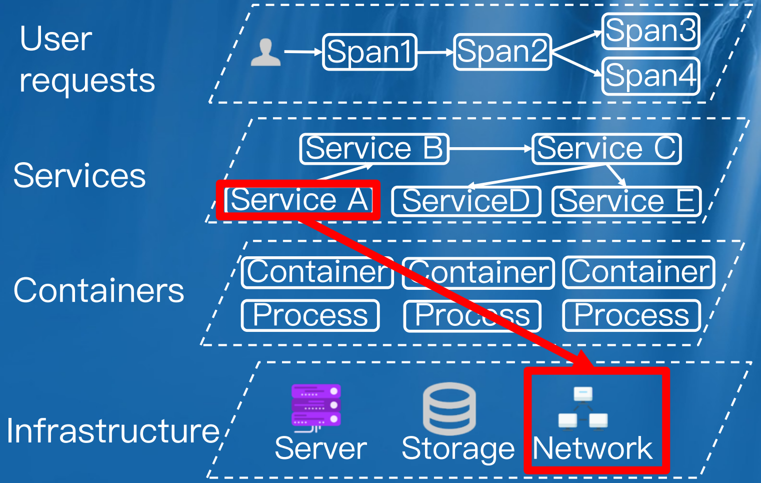
FAILURE DIAGNOSIS FRAMEWORK

Goal

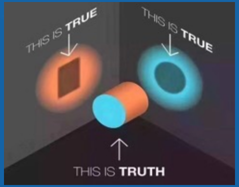
Failure Diagnosis in the Cloud

Framework

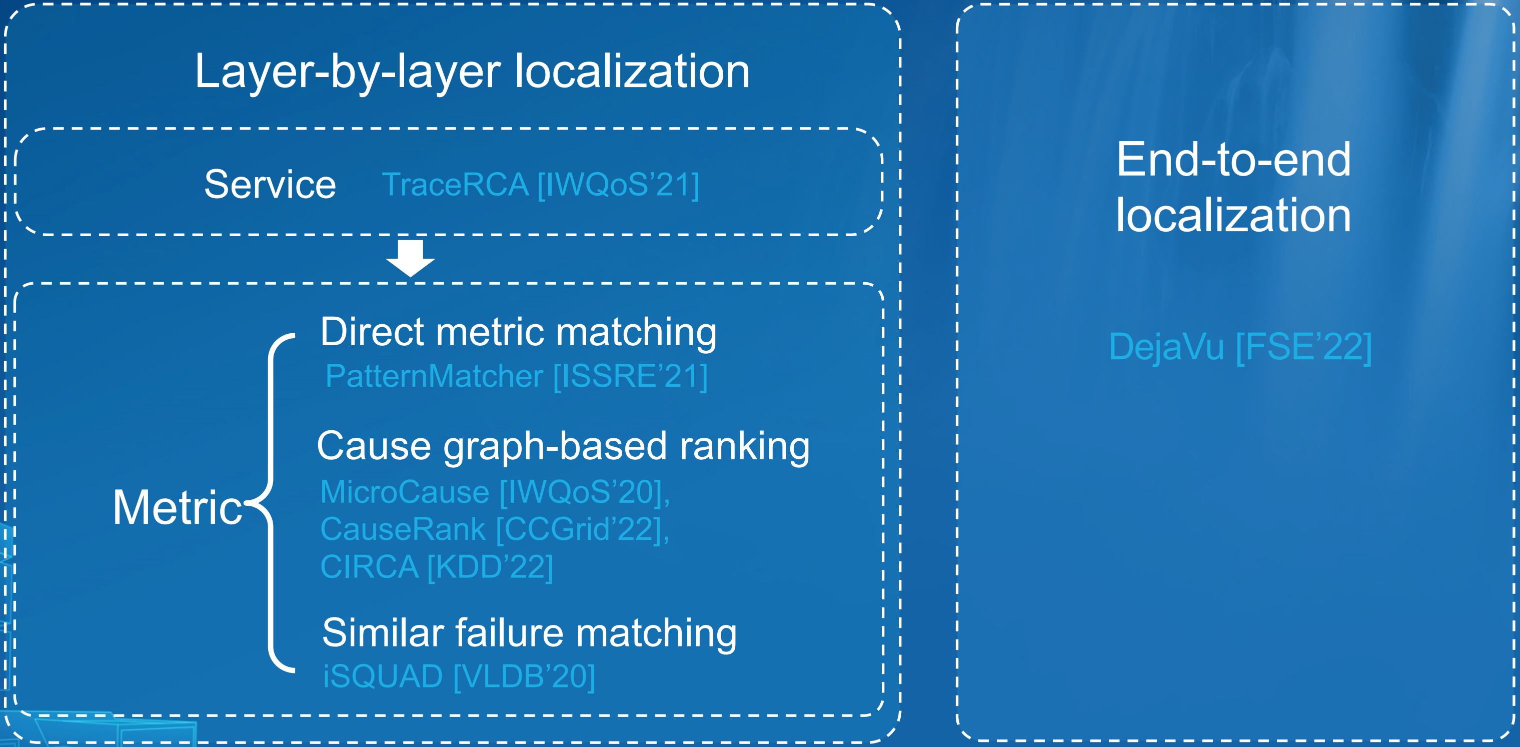
End-to-end localization



- Overall judgement



OUR RECENT WORKS



Layer-by-layer localization

Service TraceRCA [IWQoS’21]

Metric

- Direct metric matching
  PatternMatcher [ISSRE’21]
- Cause graph-based ranking
  MicroCause [IWQoS’20],
  CauseRank [CCGrid’22],
  CIRCA [KDD’22]
- Similar failure matching
  iSQUAD [VLDB’20]

End-to-end localization

DejaVu [FSE’22]

All case studies are from joint work with Industry Collaborators

阿里巴巴 Alibaba.com

中国移动 China Mobile

中国建设银行 China Construction Bank

中国民生银行 CHINA MINSHENG BANK

必示 BIZSEER

Outline

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

Lessons Learned

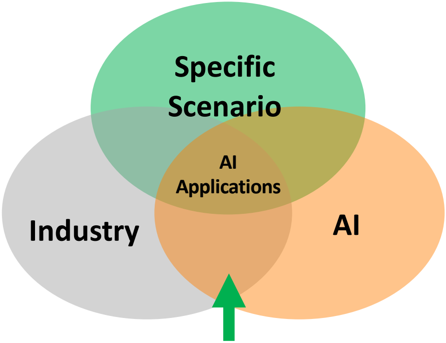
Lesson 1:
From Practice, into practice

1. Discover challenging problems from Practice
2. Design AI Algorithms to solve the disovered problem
3. Deploy the algorithms in practice. If not working perfectly? A new problem discovered, go to step 1.

Principles for Success: "The Five Step Process" | Episode 3
1 GOALS
2 PROBLEMS
5 DO IT
4 DESIGN
3 DIAGNOSIS

LESSON 2: general AI algorithms can hardly be used as Blackboxes to Ops problems

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



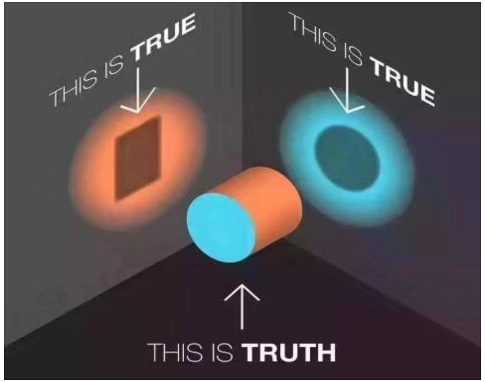
Treat AI as a high-level programming language, to"code" some components
Output of AI-enabled components are probabilistic rather than deterministic

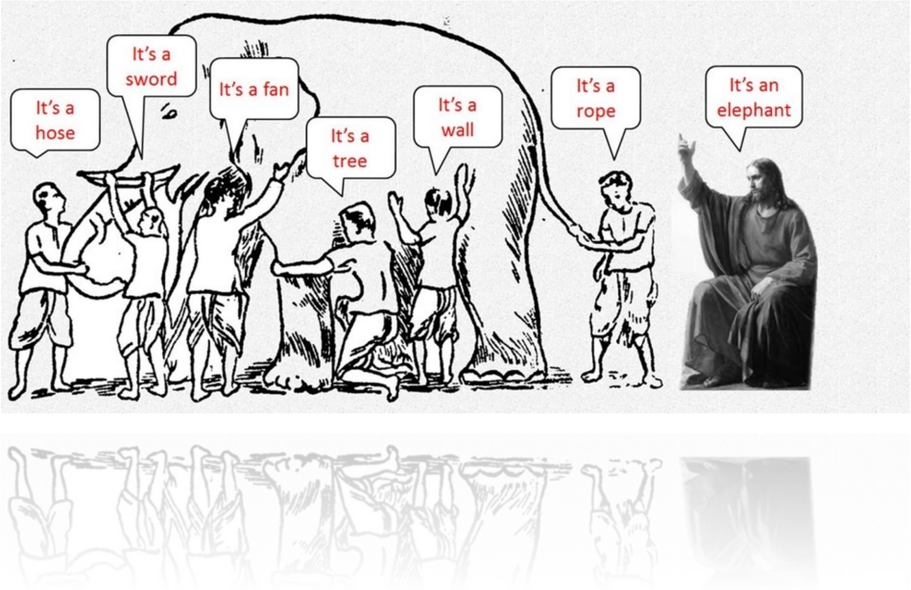
General Machine Learning Algorithms

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, Marcovian Chain, multi-instance learning, transfer learning, CNN, RNN ,VAE, GAN, NLP

Lesson 3: Utilize as many data sources as possible

- Features
- Correlation
- Glues: topology, call graph, causal relationship



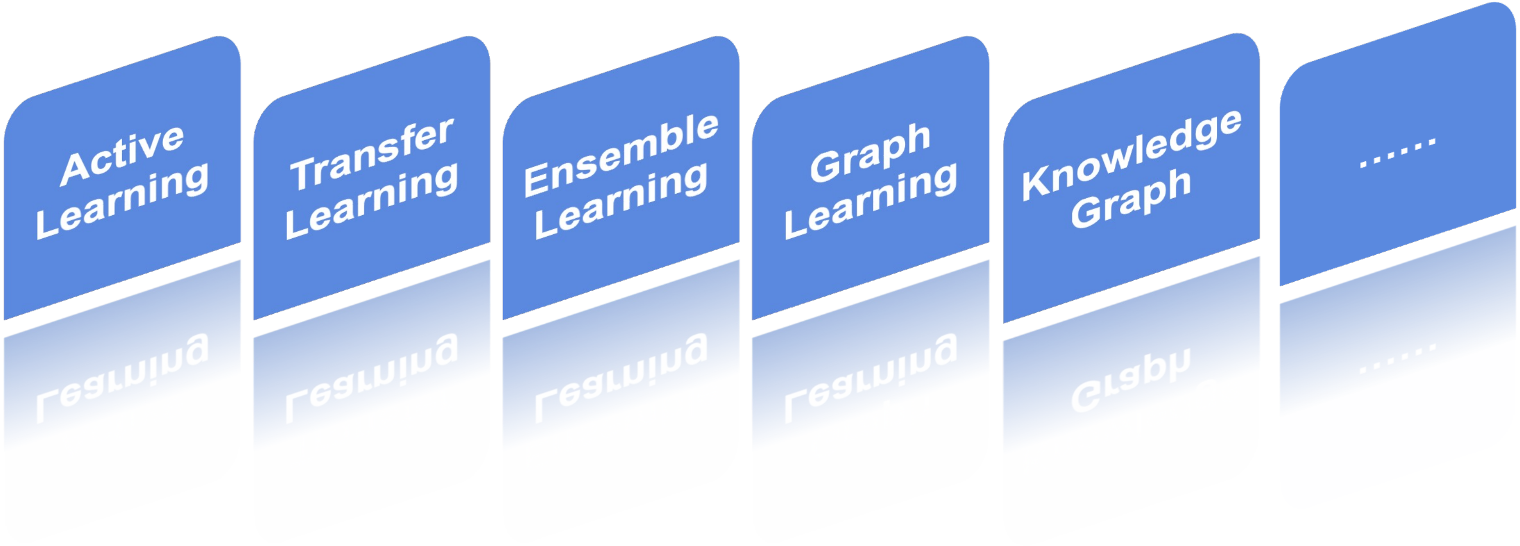


Lesson 4: As little labeling as possible

- ➢ In sharp contrast with computer vision, labeling in Ops cannot be crowdsourced.
- ➢ Although the users are themselves experts who can label, their preferences are still in this order:

1. Unsupervised approaches
2. Unsupervised approaches + active learning (whitelisting)
3. Semi-supervised approaches; supervised approaches +transfer learning
4. Supervised approaches

Lesson 5: Fully utilize latest AI technologies that enable better machine-human hybrid architecture



Active Learning

Transfer Learning

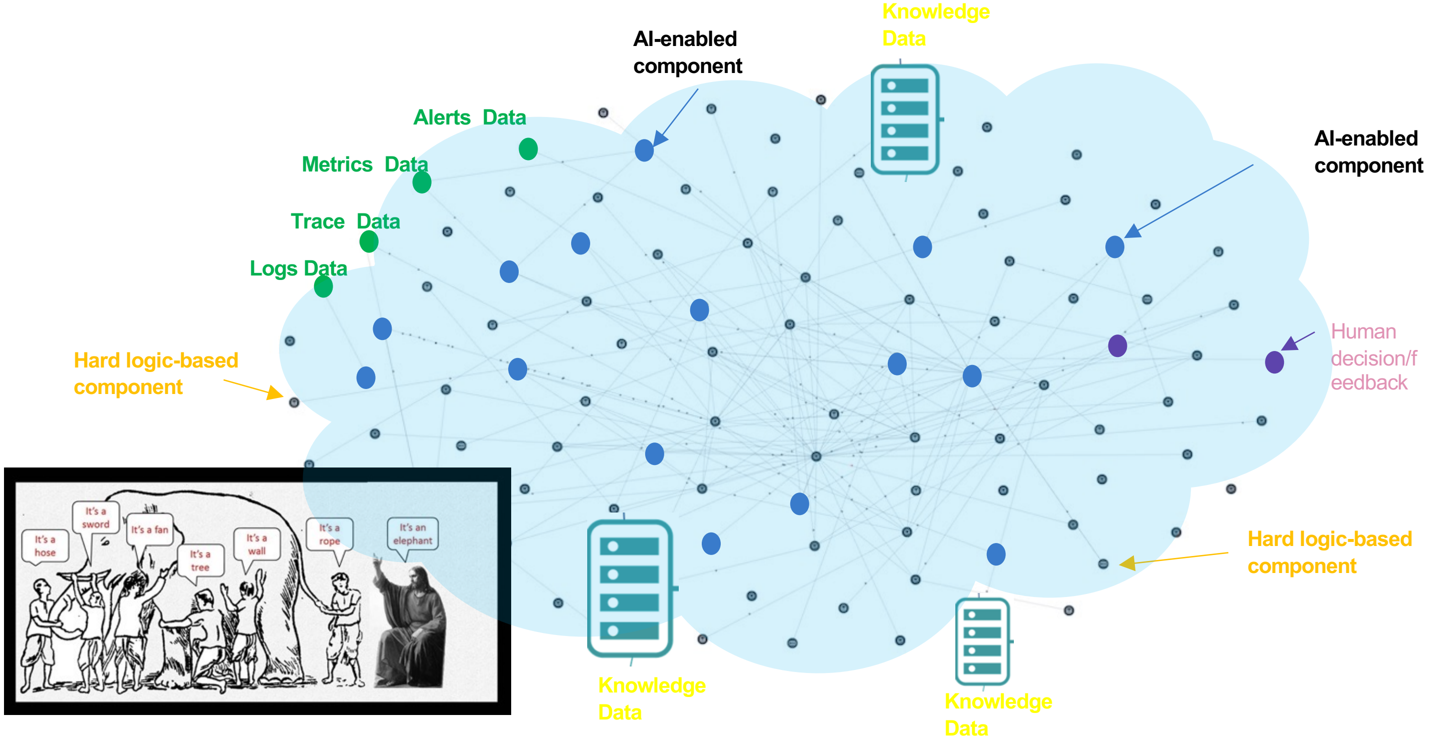
Ensemble Learning

Graph Learning

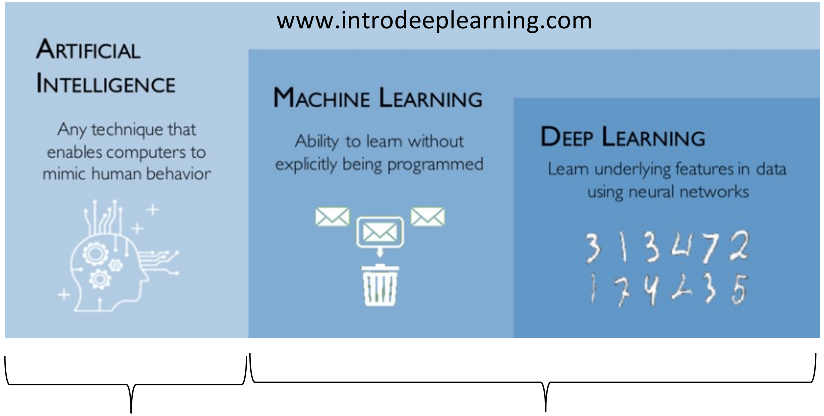
Knowledge Graph

……

Lesson 6: divide and conquer, design the overall system around each component's known capability and property, and "glue" the components using "knowledge"



AI 3.0： Deep Learning + Knowledge Engineering


www.introdeeplearning.com
ARTIFICIAL INTELLIGENCE
Any technique that enables computers to mimic human behavior
MACHINE LEARNING
Ability to learn without explicitly being programmed
DEEP LEARNING
Learn underlying features in data using neural networks
3 1 3 4 7 2
1 7 4 2 3 5

AI 1.0 AI 2.0

Bo Zhang, Jun Zhu, Hang Su, AI 3.0

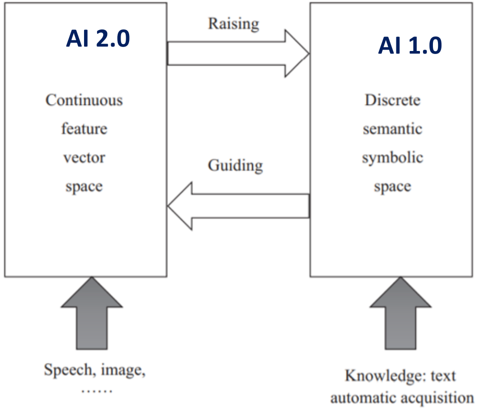

AI 2.0
Continuous feature vector space
Raising
Guiding
AI 1.0
Discrete semantic symbolic space
Speech, image, ……
Knowledge: text automatic acquisition

图 2 双空间模型

Figure 2 Dual-space mode

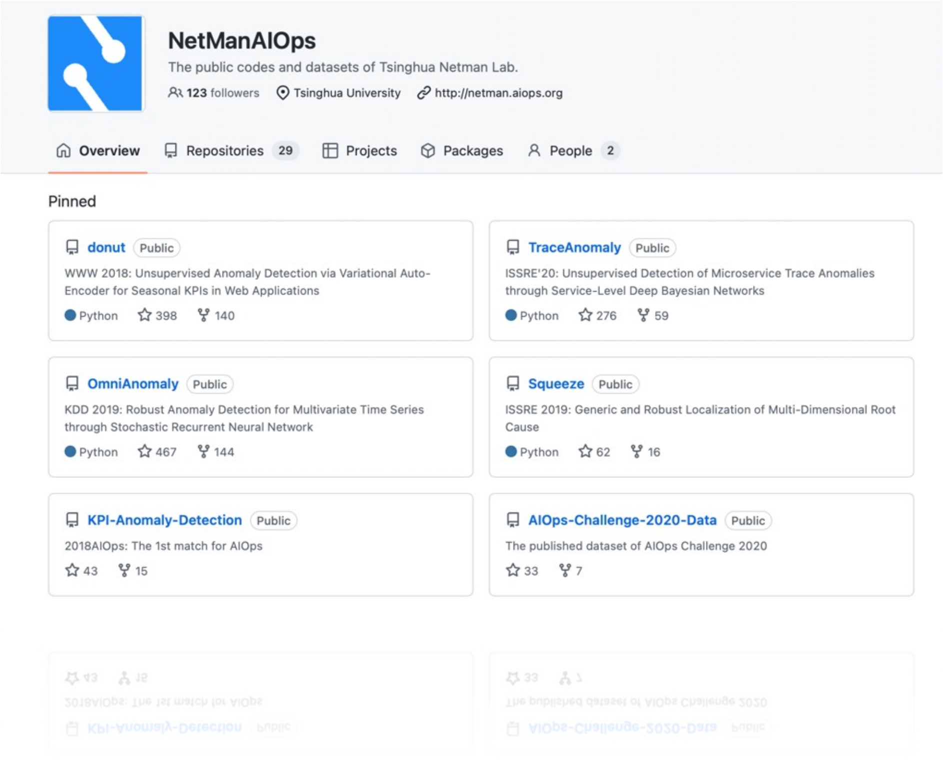
AI 3.0 = AI 1.0 + AI 2.0, still in its early research stage

Lesson 7: it really takes time and community efforts to solve real-world IT Operations problems

> “Most people overestimate what they can do in one year and underestimate what they can do in ten years.”
>
> -- Bill Gates

Some Open-Sourced Algorithms from NetMan

https://github.com/netmanaiops



NetManAIOps

The public codes and datasets of Tsinghua Netman Lab.

123 followers · Tsinghua University · http://netman.aiops.org

Pinned

donut Public
WWW 2018: Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications
● Python ☆ 398 140

TraceAnomaly Public
ISSRE'20: Unsupervised Detection of Microservice Trace Anomalies through Service-Level Deep Bayesian Networks
● Python ☆ 276 59

OmniAnomaly Public
KDD 2019: Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network
● Python ☆ 467 144

Squeeze Public
ISSRE 2019: Generic and Robust Localization of Multi-Dimensional Root Cause
● Python ☆ 62 16

KPI-Anomaly-Detection Public
2018AIOps: The 1st match for AIOps
☆ 43 15

AIOps-Challenge-2020-Data Public
The published dataset of AIOps Challenge 2020
☆ 33 7

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

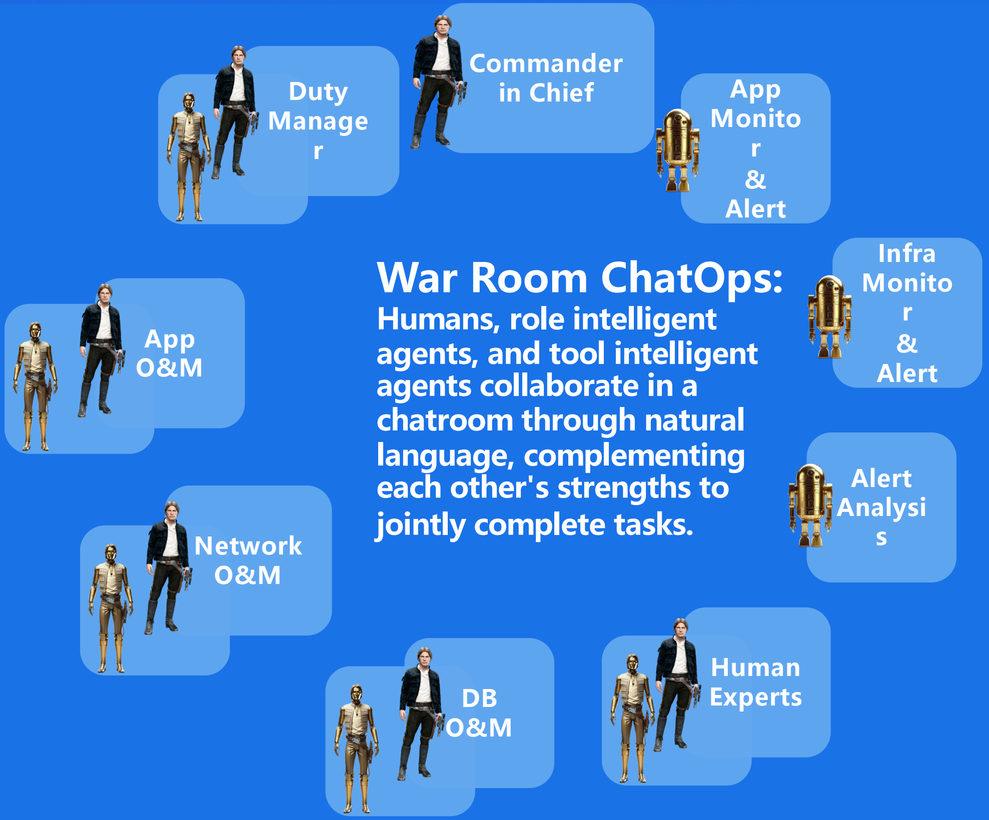
- 2019 AIOps Challenge: multi-attribute time series anomaly localization. Published data from an Internet company. More than 60 teams participated.
  Data Downloadable @
  https://github.com/NetManAIOps/MultiDimension-Localization

- 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. More than 200 teams participated

- 2022 AIOps Challenge: Failure identification and classification for microservice-based online shopping systems. More than 300 teams participated

Mid to Long-Term Application: Multi-AIOps intelligent agents collaborate with humans to complete complex tasks



War Room ChatOps:

Humans, role intelligent agents, and tool intelligent agents collaborate in a chatroom through natural language, complementing each other's strengths to jointly complete tasks.

Mid to Long-Term Application: Multi-AIOps intelligent agents collaborate with humans to complete complex tasks

War Room ChatOps: Humans, role intelligent agents, and tool intelligent agents collaborate in a chatroom through natural language, complementing each other's strengths to jointly complete tasks.

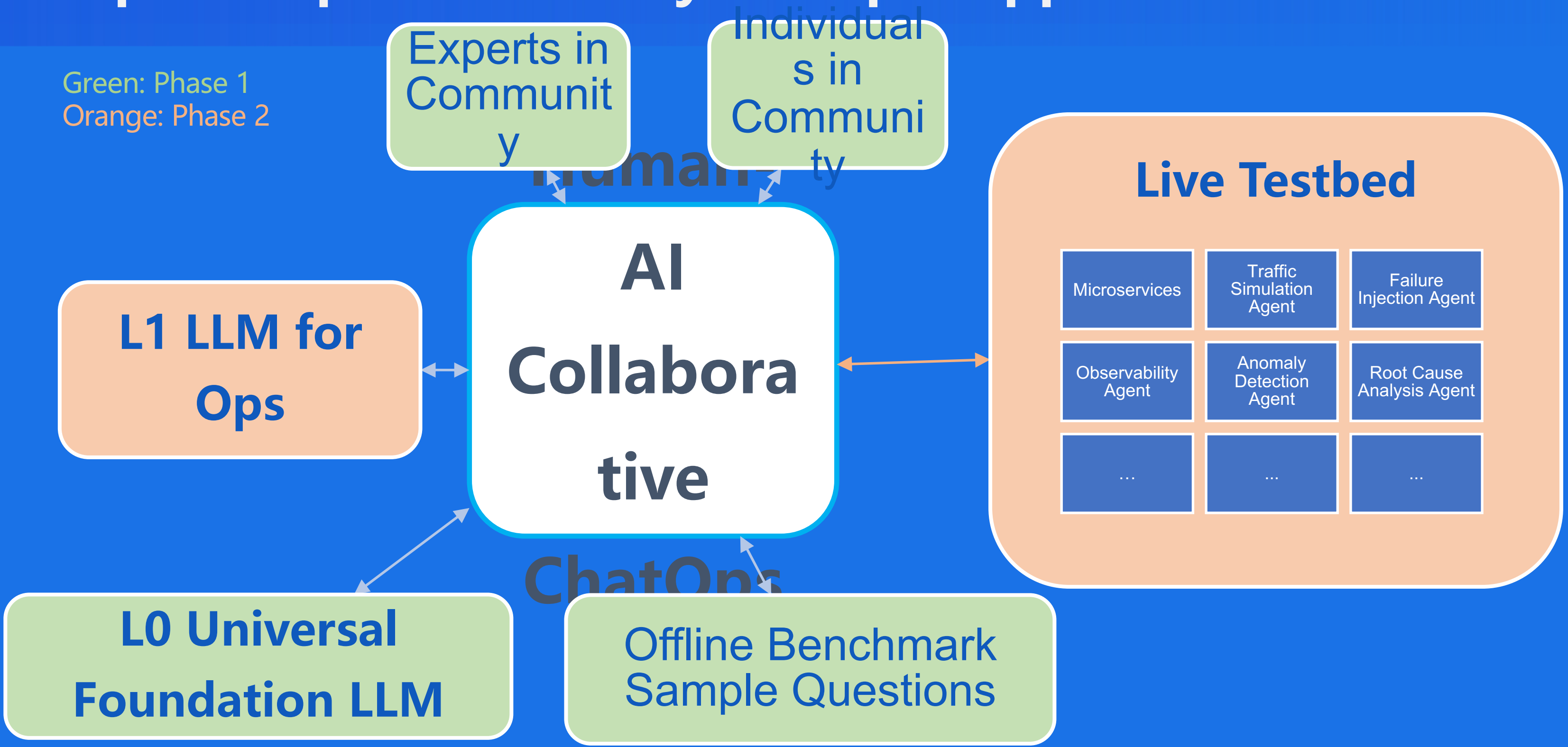
Anomaly Detection Agent: "Alert: Anomaly pattern detected. Service response time has increased by 300% in the past hour."
Commander: "Team, our service response time has tripled. We need to diagnose this issue immediately. Any ideas?"
Manager: "I just checked the dashboard. This peak is unprecedented. We need to quickly isolate the cause."
Engineer: "I'm checking the server logs and system metrics. Initial signs point to a bottleneck at the database layer."
Engineer Agent: "Suggestion: Perform a detailed analysis of database queries and check for inefficient operations."
Alarm Analysis Agent: "Analysis: Recent alerts show a high volume of timeout errors in the database service, correlating with the increase in response time."
Manager: "Is this related to our recent updates? Are there any unoptimized queries or resource-intensive tasks?"
Engineer: "I'm cross-referencing the update logs. There might be a connection. I'll focus on query optimization."
Commander: "At the same time, we should also consider external factors. Engineer Agent, can you run diagnostic checks for potential security vulnerabilities or DDoS attacks?"
Engineer Agent: "Initiating diagnostics. Checking for unusual traffic patterns and external threats."
Manager: "Meanwhile, we should prepare a contingency plan. If this situation persists, it will have a significant impact on our customers."
Anomaly Detection Agent: "Update: No external security vulnerabilities detected. The issue seems to be internal."
Alarm Analysis Agent: "Alert: A recent surge in user requests may have exacerbated the pressure on the database."
Engineer: "Confirmed, the issue seems to be inefficient database queries triggered by high user load. I'm fixing it now."
Commander: "Good work, team. Let's prioritize this issue and closely monitor the system. Please keep updating.”

……

OpenAIOps

CCF OpenAIOps Community Example Application

Green: Phase 1
Orange: Phase 2



OpenAIOps

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

清華大學 | Tsinghua NetMan