Reliability-Driven AIOps for Cloud Resilience

Prof. Michael R. Lyu

The Chinese University of Hong Kong

香港中文大學 The Chinese University of Hong Kong

ARISE Automated Reliable Intelligent Software Engineering

Background

- Modern software systems are serving many aspects of our life

…

…

Search Engine	Cloud Service	Office Software	Operating Systems

Cloud Computing

- Cloud adoption rising
- Cloud revenue growing

	2018	2019	2020	2021	2022
Cloud Business Process Services (BPaaS)	41.7	43.7	46.9	50.2	53.8
Cloud Application Infrastructure Services (PaaS)	26.4	32.2	39.7	48.3	58.0
Cloud Application Services (SaaS)	85.7	99.5	116.0	133.0	151.1
Cloud Management and Security Services	10.5	12.0	13.8	15.7	17.6
Cloud System Infrastructure Services (IaaS)	32.4	40.3	50.0	61.3	74.1
Total Market	**196.7**	**227.8**	**266.4**	**308.5**	**354.6**

BPaaS = business process as a service; IaaS = infrastructure as a service; PaaS = platform as a service; SaaS = software as a service

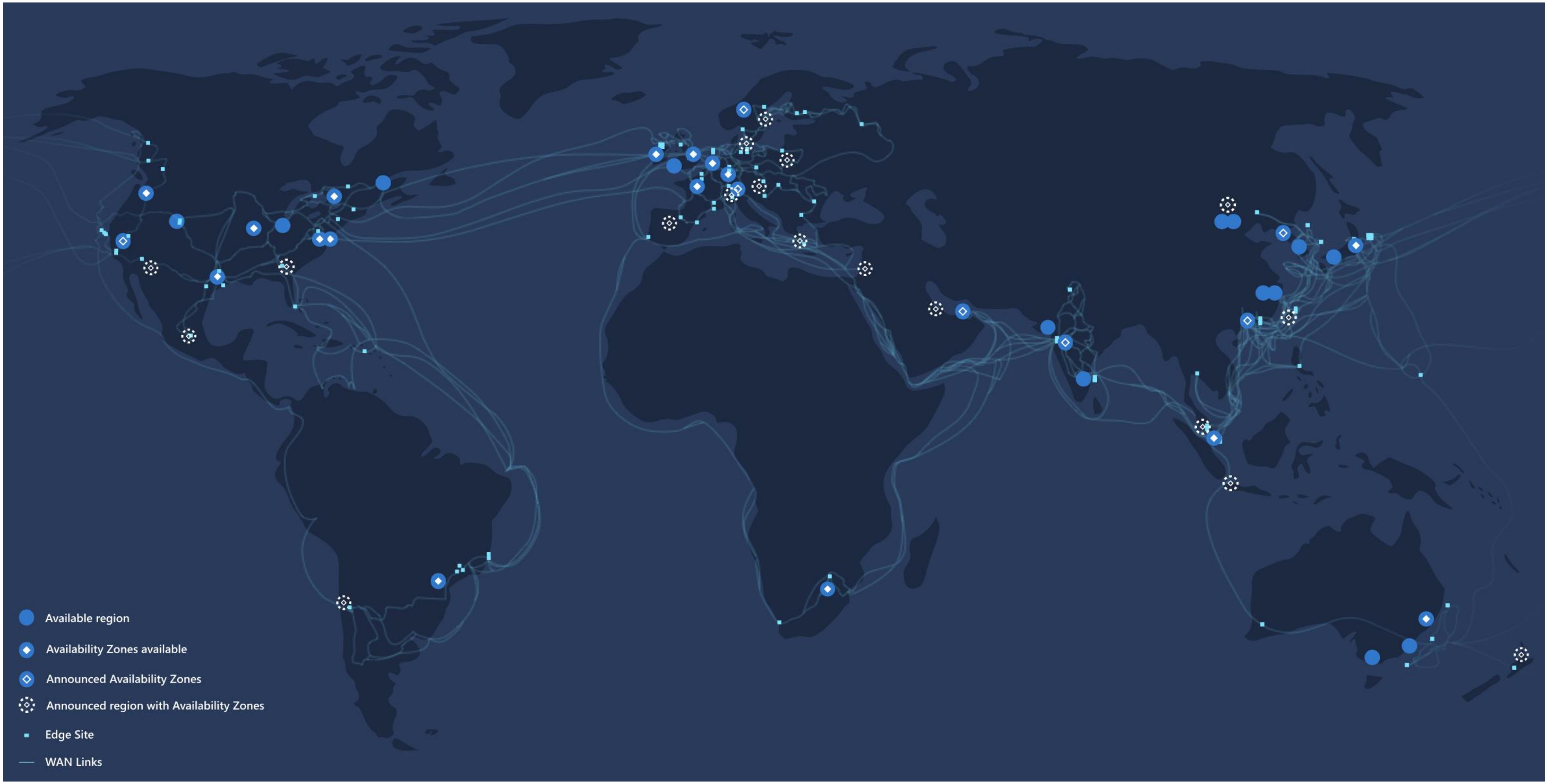
Worldwide Public Cloud Service Revenue Forecast (Billions of U.S. Dollars)

Microsoft Azure Global Network

60+ regions

100 Gbps bandwidth

130,000 miles of fiber optics



https://azure.microsoft.com/en-us/global-infrastructure/global-network/

Real-World Revenue Loss

Lloyd's Estimates the Impact of a U.S. Cloud Outage at $19 Billion

By: Sean Michael Kerner | January 24, 2018

A joint research report from insurance provider Lloyd's of London and the American Institutes for Research (AIR), looks at the potential costs related to a major public cloud outage in the U.S.

As organizations around the world increasingly rely on the cloud, the impact of a public cloud failure is something that insurance companies are now concerned about. A 67-page report released on Jan. 23 from Lloyd's of London and AIR Worldwide provides some insight and estimates on the potential losses from a major cloud services outage—and the numbers are large.

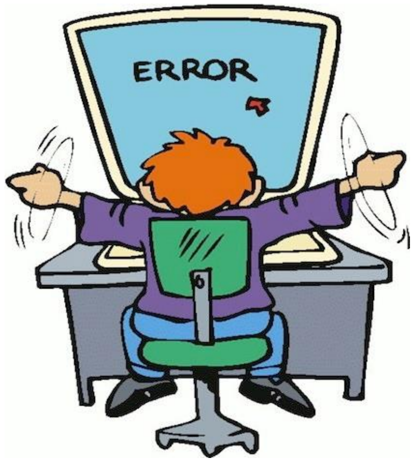
According to the report, a cyber-incident that impacted the operations of one of the top three public cloud providers in the U.S. for three to six days, could result in total losses of up to $19 billion. Of those loses, only $1.1 to $3.5 billion would be insured, leaving organizations left to cover the rest of the costs.

http://www.eweek.com/cloud/lloyd-s-estimates-the-impact-of-a-u.s.-cloud-outage-at-19-billion

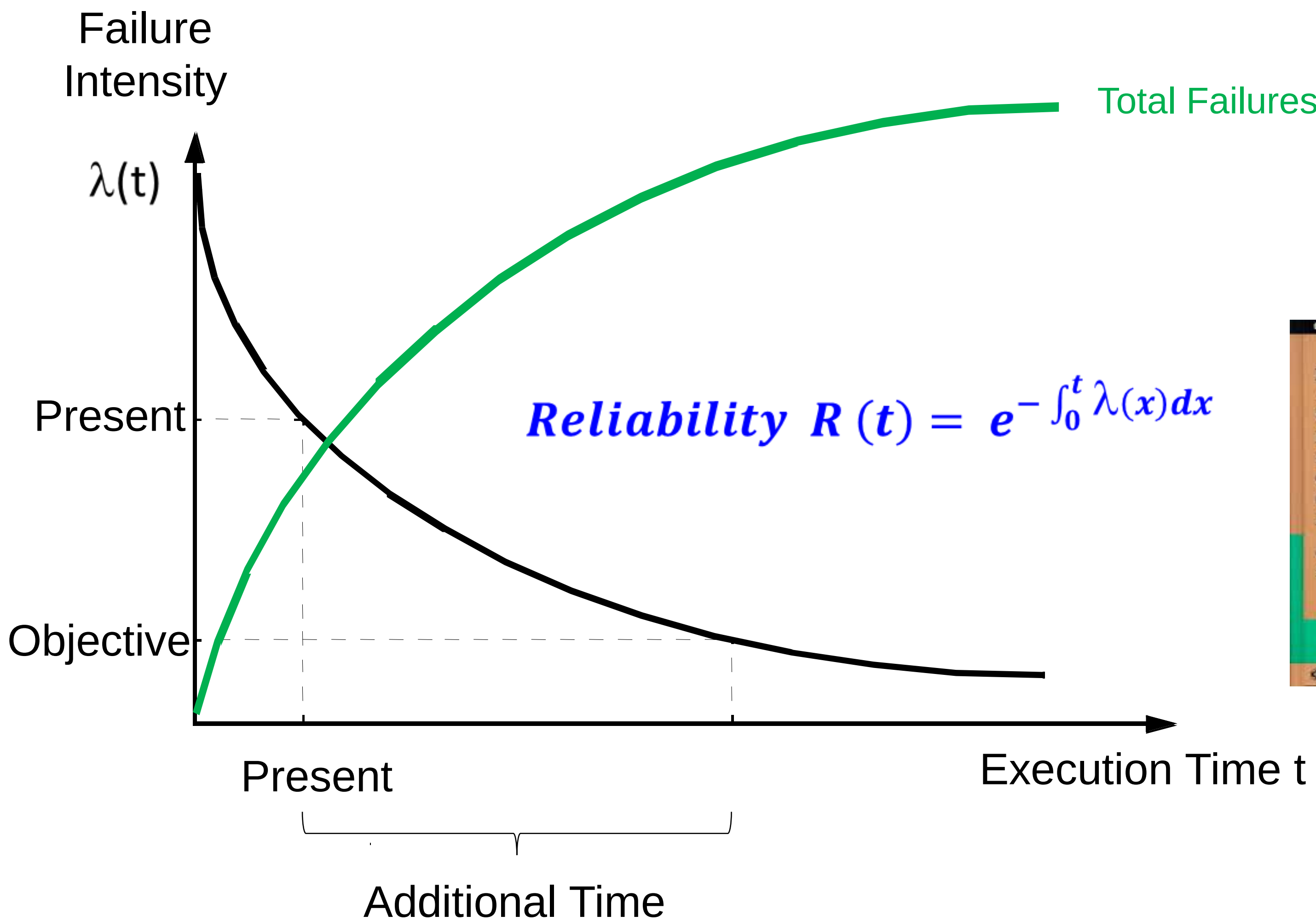
Cloud Resilience Is Very Crucial!

- State-of-the-art cloud reliability
  - Service Level Agreement (SLA)
  - 5-6 9s' availability
  - High degree of automation

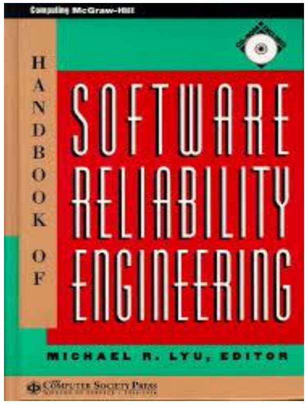
- Cloud reliability issues
  - Tough cloud failures take a long time to mitigate
  - Impose large revenue loss
  - Harm customer trust and enterprise reputation



Site Reliability Engineering (SRE)



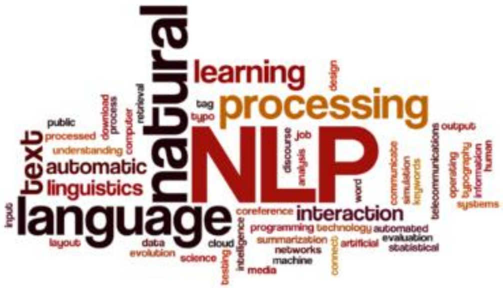
$$\textit{Reliability}\ R(t) = e^{-\int_0^t \lambda(x)dx}$$



Fault Avoidance →	Fault Removal →	Fault Tolerance →	Fault Prediction

Data-Driven AI Applications

Data



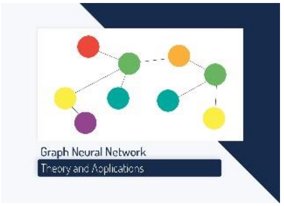
GitHub

stackoverflow

IBM CodeNet

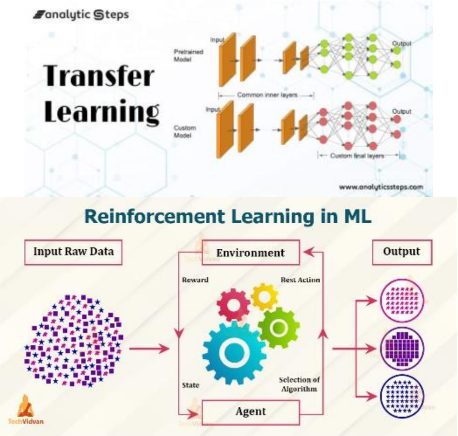
Models/Paradigms

CNN



GNN

RNN
LSTM



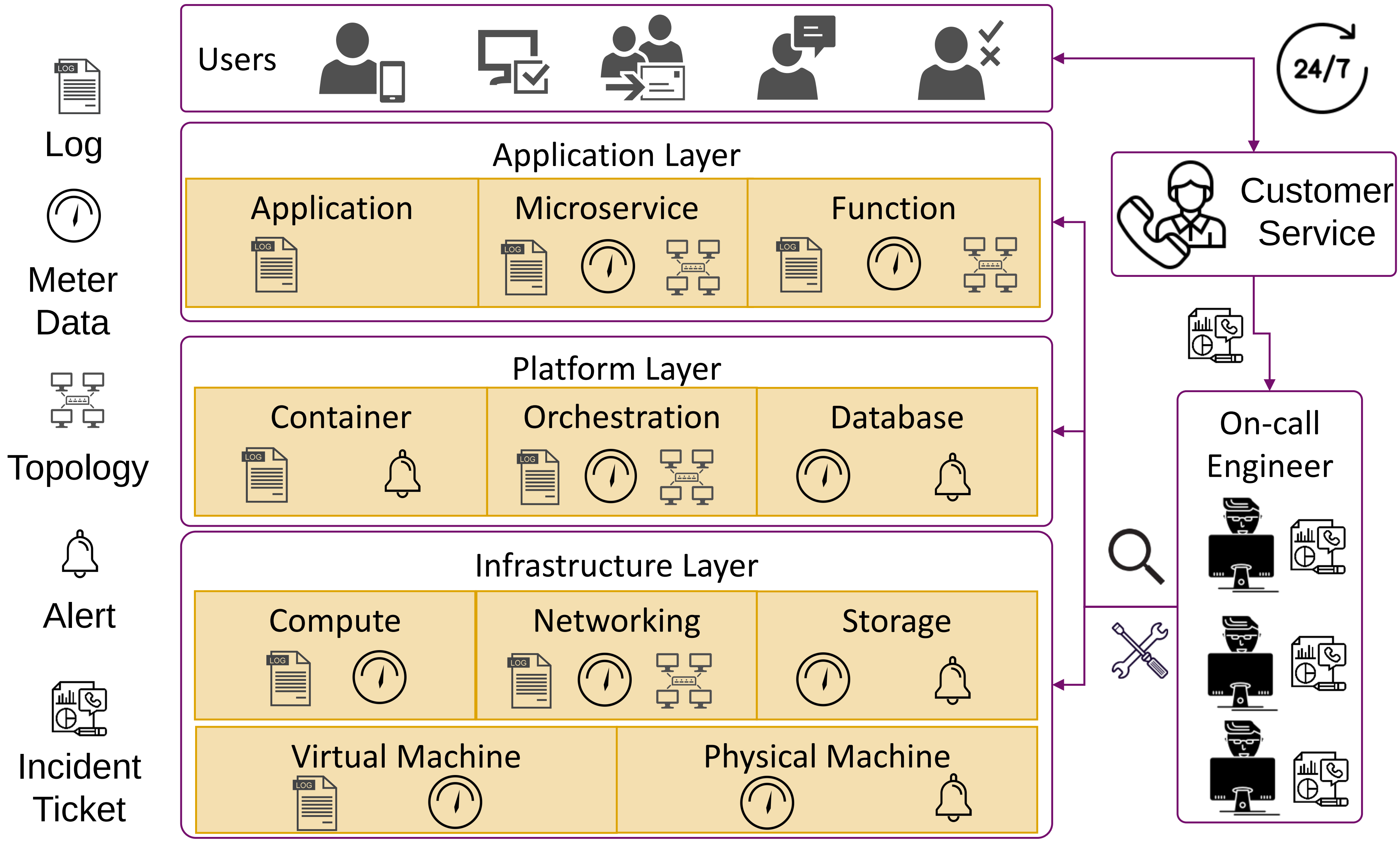
Tasks

- ❖ Image classification
- ❖ Image localization
- ❖ Object detection
- ❖ Semantic segmentation

- ❖ Machine translation
- ❖ Information retrieval
- ❖ Question answering
- ❖ Sentiment Analysis
- ❖ Natural language understanding

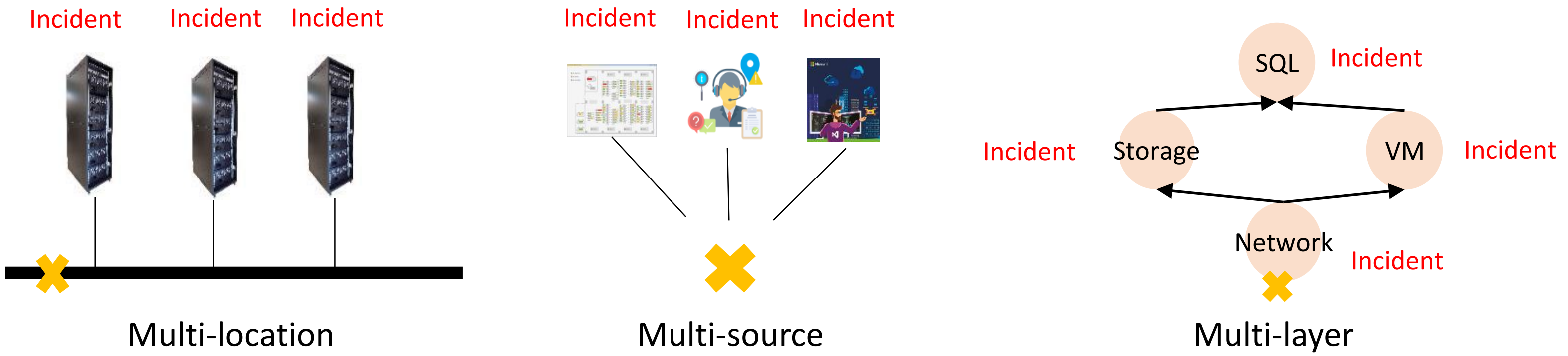
- ❖ Code summarization
- ❖ Code clone detection
- ❖ Code suggestion
- ❖ API recommendation
- ❖ Bug localization
- ❖ Semantic parsing

Cloud Generates a Variety of Data



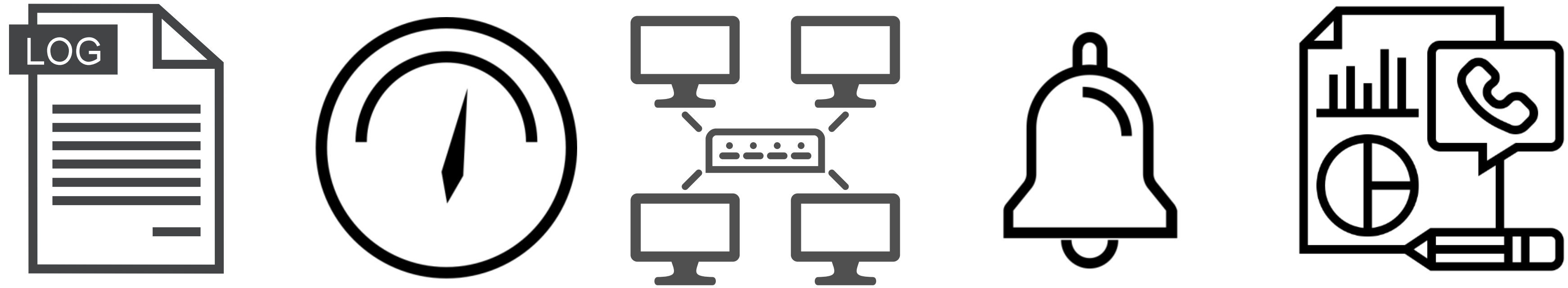
Challenges of Resilient Cloud Operations

- Current Status:
  - Incidents are highly-correlated, but separately resolved
- Reasons:
  - New DevOps paradigm, complex service dependency, load balance, backup and restore



Humans are not good at solving this large-scale complex problem, but AI is

AIOps for Cloud Resilience



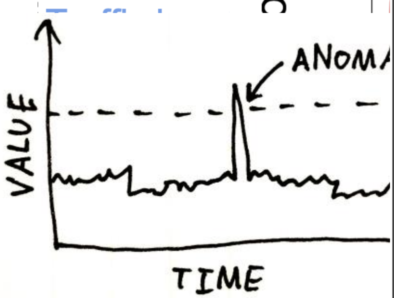
Log | Meter Data | Topology | Alert | Incident Ticket

Abnormal r
Network int
Traffic burs

100
80

Raw Log Messages

1	**2008-11-11 03:40:58** BLOCK* NameSystem.allocateBlock: /user/root/randtxt4/_temporary/_task_200811101024_0010_m_000011_0/part-00011.blk_904791815409399662
2	**2008-11-11 03:40:59** Receiving block blk_904791815409399662 src: /10.251.43.210:55700 dest: /10.251.43.210:50010
3	**2008-11-11 03:41:01** Receiving block blk_904791815409399662 src: /10.250.18.114:52231 dest: /10.250.18.114:50010
4	**2008-11-11 03:41:48** PacketResponder 0 for block blk_904791815409399662 terminating
5	**2008-11-11 03:41:48** Received block blk_904791815409399662 of size 67108864 from /10.250.18.114
6	**2008-11-11 03:41:48** PacketResponder 1 for block blk_904791815409399662 terminating
7	**2008-11-11 03:41:48** Received block blk_904791815409399662 of size 67108864 from /10.251.43.210
8	**2008-11-11 03:41:48** BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.43.210:50010 is added to blk_904791815409399662 size 67108864
9	**2008-11-11 03:41:48** BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.250.18.114:50010 is added to blk_904791815409399662 size 67108864
10	**2008-11-11 08:30:54** Verification succeeded for blk_904791815409399662



Database a
Monitor det

Anomaly Detection

Fault Avoidance

Low
High
Low
Medium
Medium
Medium
Medium
Medium

Fault Prediction

Main Contents in This Talk

High-level

Low-level

Outage Prediction

Incident Management
- Alert Aggregation
- Empirical Study

Service Dependency
- Correlation Mining
- Root Cause Analysis

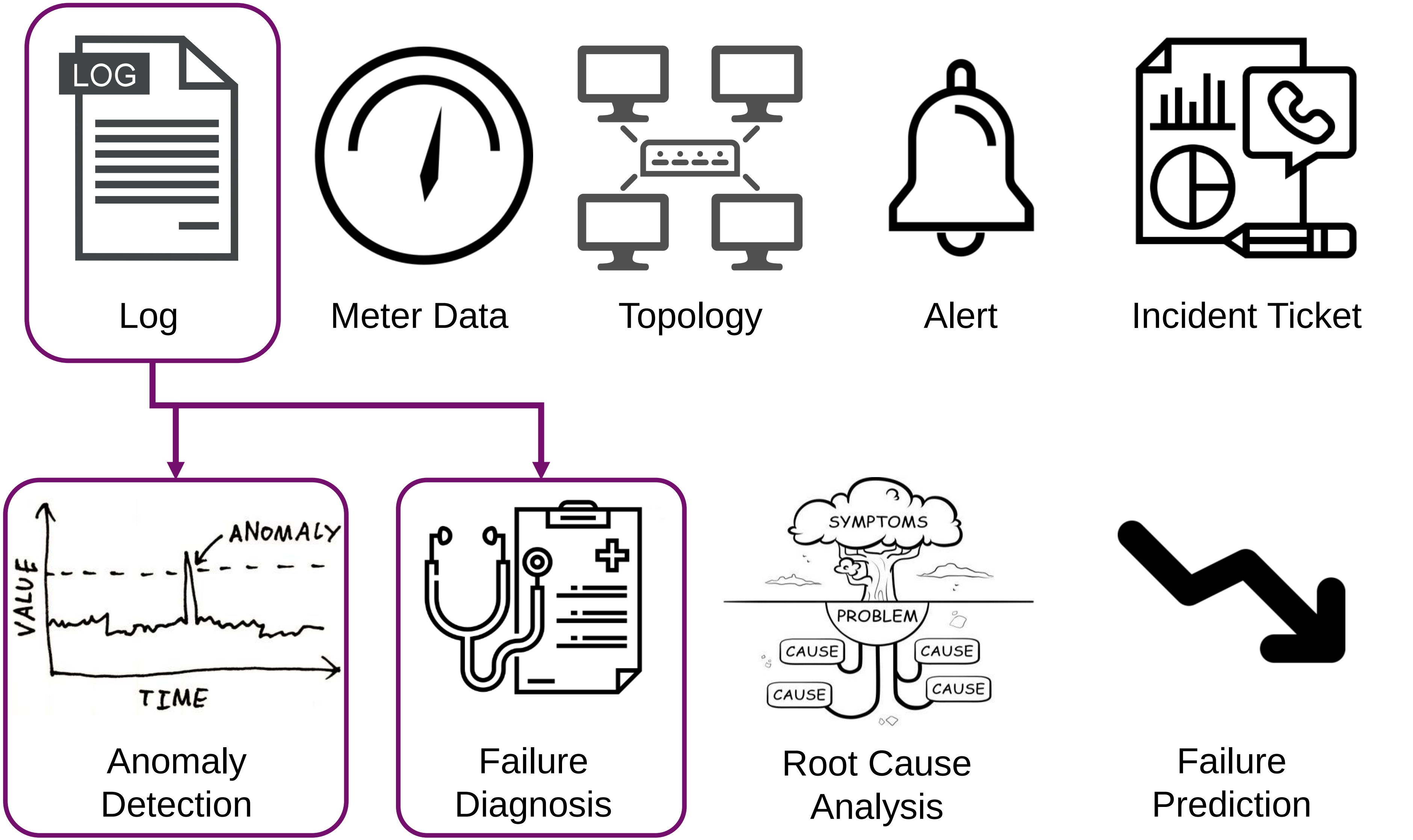
KPI Analysis
- Multivariate Analysis
- Problem Identification

Log Analysis
- Log Parsing
- Log Anomaly Detection
- Log-based Failure Diagnosis

Hard

Easy

AIOps: Log Analysis



Log Parsing: Preprocessing of Log Data

- Objective
  - transform raw log data to structural data
- Key problem to solve
  - extract event type and variables in log messages

Raw Log Messages	
1	**2008-11-11 03:40:58** BLOCK* NameSystem.allocateBlock: /user/root/randtxt4/_temporary/_task_200811101024_0010_m_000011_0/part-00011.blk_904791815409399662
2	**2008-11-11 03:40:59** Receiving block blk_904791815409399662 src: /10.251.43.210:55700 dest: /10.251.43.210:50010
3	**2008-11-11 03:41:01** Receiving block blk_904791815409399662 src: /10.250.18.114:52231 dest: /10.250.18.114:50010
4	**2008-11-11 03:41:48** PacketResponder 0 for block blk_904791815409399662 terminating
5	**2008-11-11 03:41:48** Received block blk_904791815409399662 of size 67108864 from /10.250.18.114
6	**2008-11-11 03:41:48** PacketResponder 1 for block blk_904791815409399662 terminating
7	**2008-11-11 03:41:48** Received block blk_904791815409399662 of size 67108864 from /10.251.43.210
8	**2008-11-11 03:41:48** BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.43.210:50010 is added to blk_904791815409399662 size 67108864
9	**2008-11-11 03:41:48** BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.250.18.114:50010 is added to blk_904791815409399662 size 67108864
10	**2008-11-11 08:30:54** Verification succeeded for blk_904791815409399662

Parsing

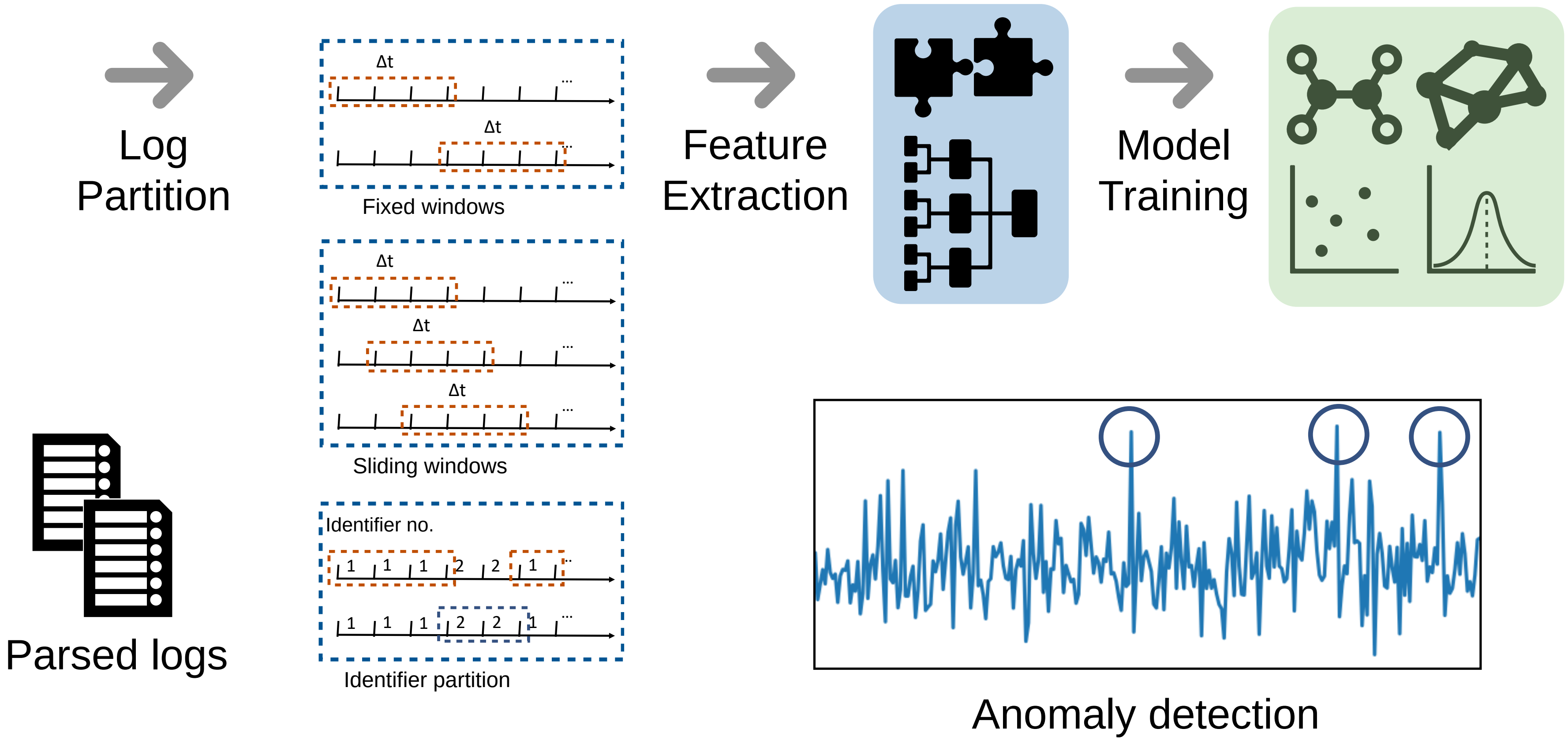
Log Events	
Event1	BLOCK* NameSystem.allocateBlock: *
Event2	Receiving block * src: * dest: *
Event3	PacketResponder * for block * terminating
Event4	Received block * of size * from *
Event5	BLOCK* NameSystem.addStoredBlock: blockMap updated: * is added to * size *
Event6	Verification succeeded for *

Strutured Logs		
1	blk_904791815409399662	Event1
2	blk_904791815409399662	Event2
3	blk_904791815409399662	Event2
4	blk_904791815409399662	Event3
5	blk_904791815409399662	Event4
6	blk_904791815409399662	Event3
7	blk_904791815409399662	Event4
8	blk_904791815409399662	Event5
9	blk_904791815409399662	Event5
10	blk_904791815409399662	Event6

P. He, J. Zhu, S. He, J. Li and M. R. Lyu, "An Evaluation Study on Log Parsing and Its Use in Log Mining," DSN, 2016.
P. He, J. Zhu, Z. Zheng and M. R. Lyu, "Drain: An Online Log Parsing Approach with Fixed Depth Tree," ICWS, 2017.
P. He, J. Zhu, S. He, J. Li, and M.R. Lyu, "Towards Automated Log Parsing for Large-Scale Log Data Analysis," TDSC, 2018.

Log Anomaly Detection

- Feature Engineering



Log Anomaly Detection

	Methods	Algorithm/Model	Feature	Unsupervised	Online
Traditional machine learning	Xu *et al.* [180]	PCA	⋆ †	Yes	No
	Lin *et al.* [108]	Clustering	∗	Yes	No
	He *et al.* [75]	Clustering	∗ ⋆	Yes	No
	Liang *et al.* [104]	SVM	‡	No	No
	Kimura *et al.* [91]	SVM	‡	No	No
	Xu *et al.* [179]	Frequent pattern mining	∗ ⋆	Yes	Yes
	Shang *et al.* [161]	Frequent pattern mining	∗	Yes	No
	Lou *et al.* [125]	Frequent pattern mining	⋆	Yes	No
	Farshchi *et al.* [54]	Frequent pattern mining	⋆	Yes	No
	Nandi *et al.* [145]	Graph mining	¶	Yes	No
	Lou *et al.* [124]	Graph mining	¶	Yes	No
	Yamanishi *et al.* [181]	Statistical model	∗	Yes	No
	He *et al.* [76]	Logistic regression	⋆	No	No
Deep learning	Du *et al.* [46]	LSTM model	∗ †	Yes	Yes
	Zhang *et al.* [196]	LSTM classification model	∗	No	No
	Meng *et al.* [136]	LSTM model	∗ ⋆	Yes	Yes
	Xia *et al.* [177]	LSTM-based GAN model	∗	Yes	Yes
	Lu *et al.* [128]	CNN model	∗	No	No
	Liu *et al.* [109]	Graph embedding model	¶	Yes	No

∗ Log event sequence, ⋆ Log event count vector, † Parameter value vector
‡ Ad hoc features, ¶ Graphical feature

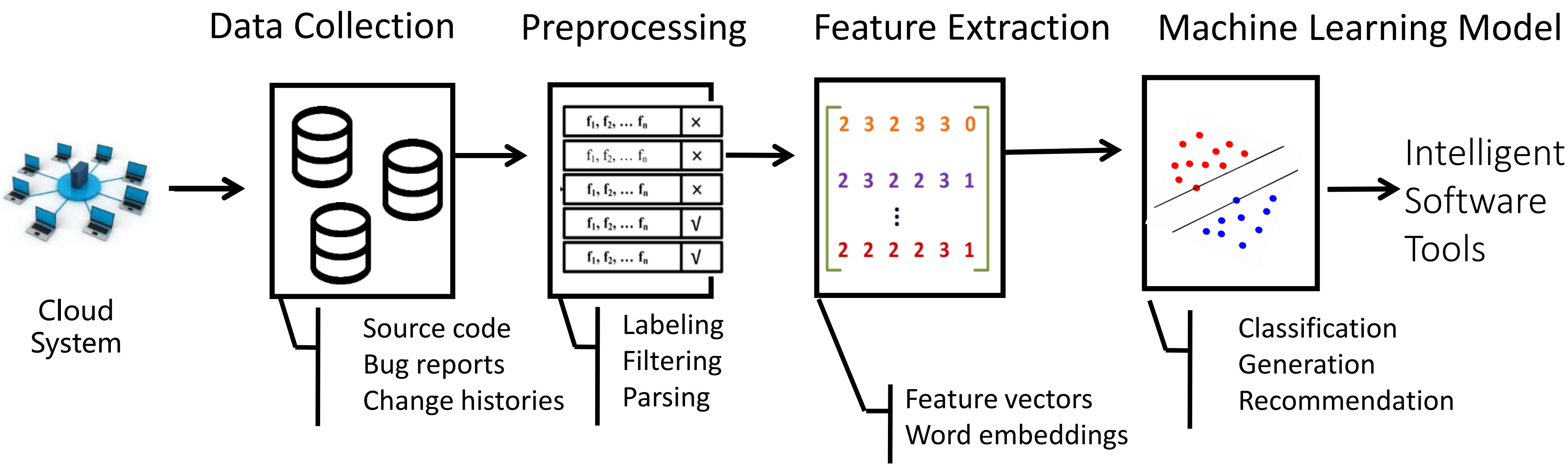
Dimens
redu

ners

S. He, P. He, Z. Chen, T. Yang, Y. Su and M.R. Lyu, “A Survey on Automated Log Analysis for Reliability Engineering”. ACM Comput. Surv. 2021

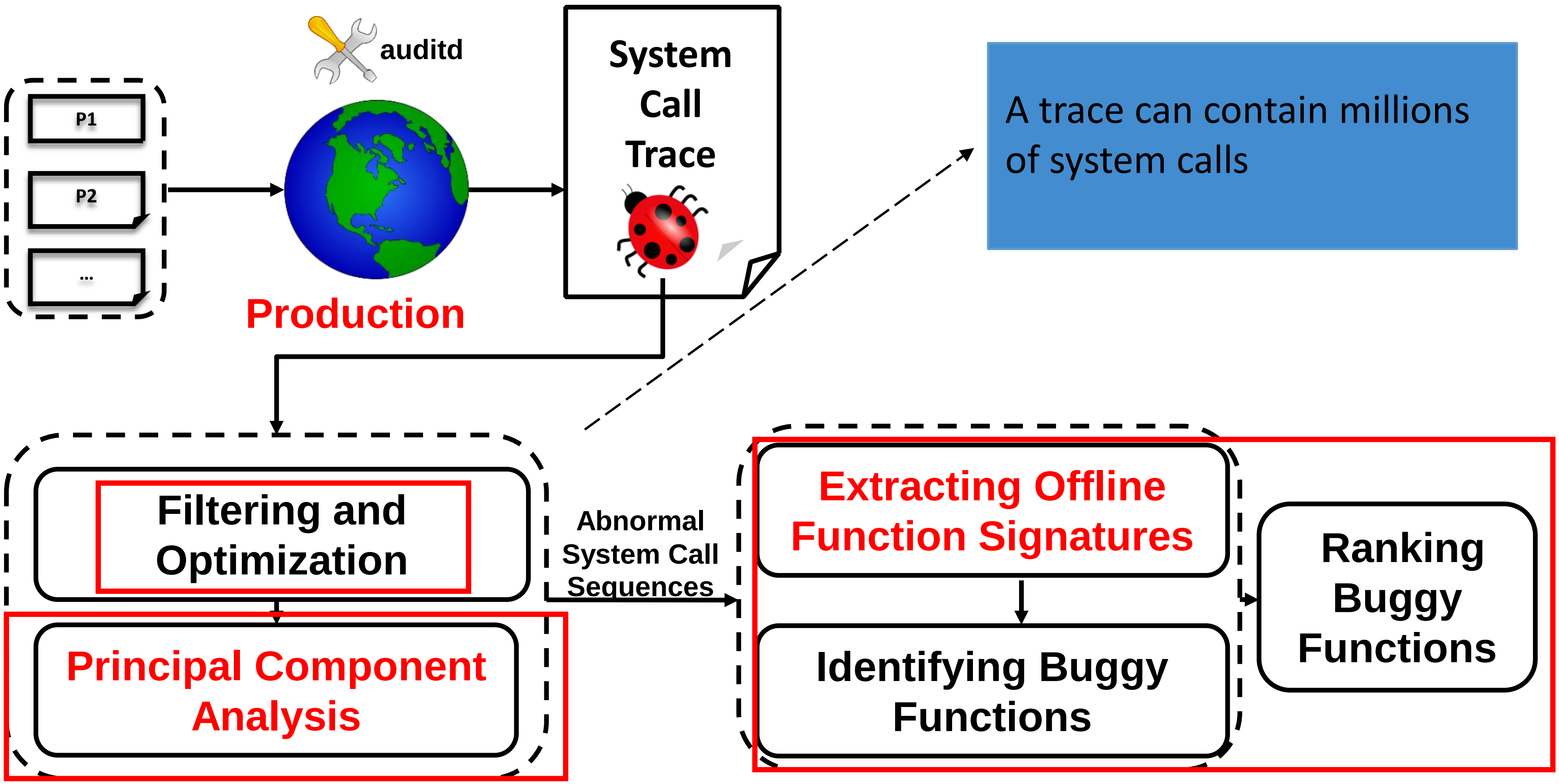
Log-based Failure Diagnosis for Cloud System

- Log is the major source for failure diagnosis



Failure Diagnosis: Ranking Buggy Functions

- PCA algorithm to find abnormal components

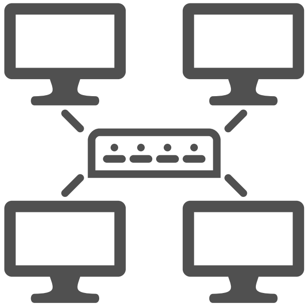


T. Zaman, X. Han, T. Yu, “SCMiner: Localizing System-Level Concurrency Faults from Large System Call Traces”, ASE 2019

AIOps: KPIs Analysis

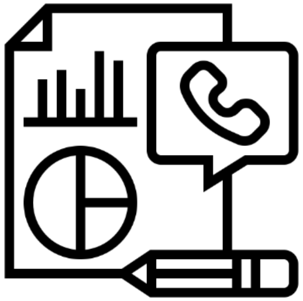
Log

Meter Data

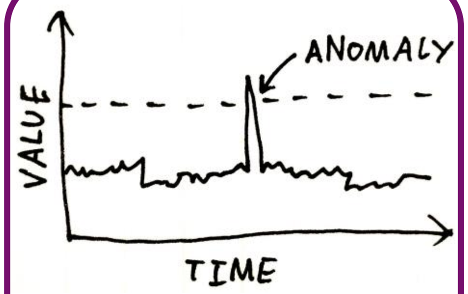


Topology

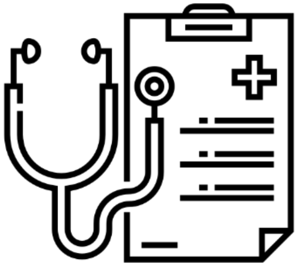
Alert



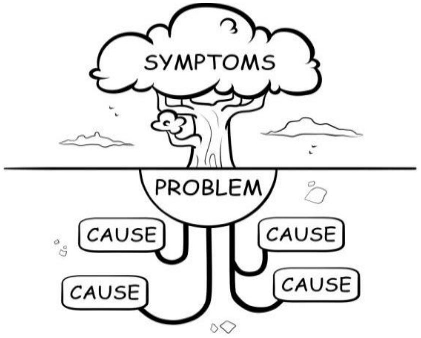
Incident Ticket



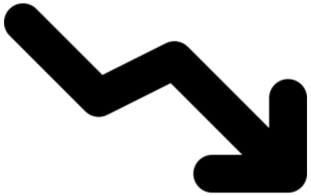
Anomaly Detection



Failure Diagnosis



Root Cause Analysis



Failure Prediction

Key Performance Indicators (KPIs)

system anomaly

network traffic

response delay

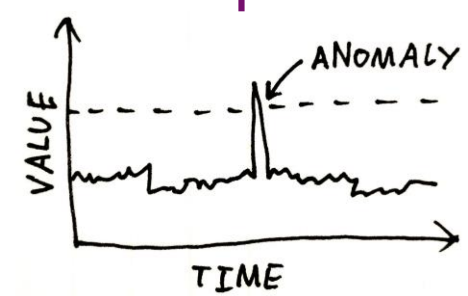
CPU usage

.

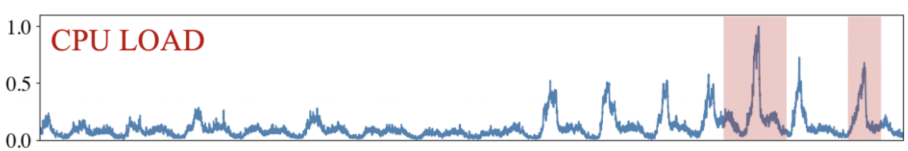
.

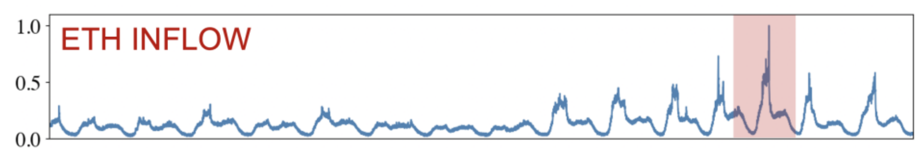
monitor runtime information

understand health status



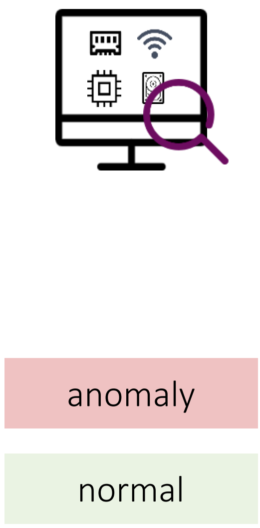
anomaly detection

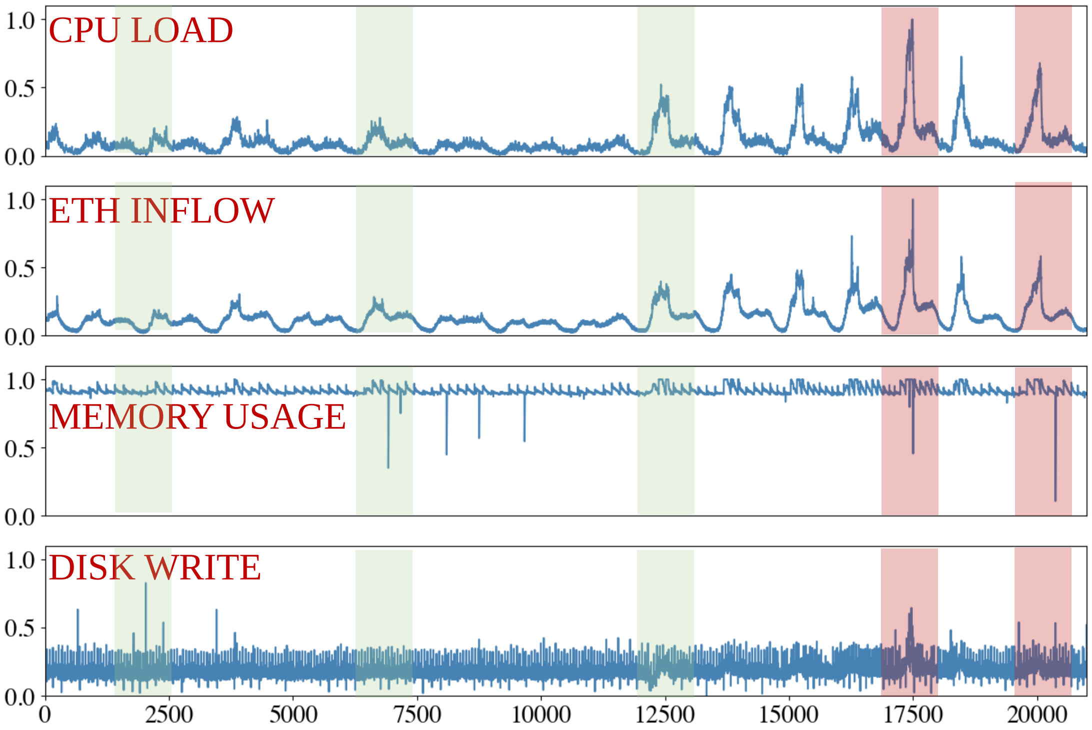




Multivariate KPIs Analysis

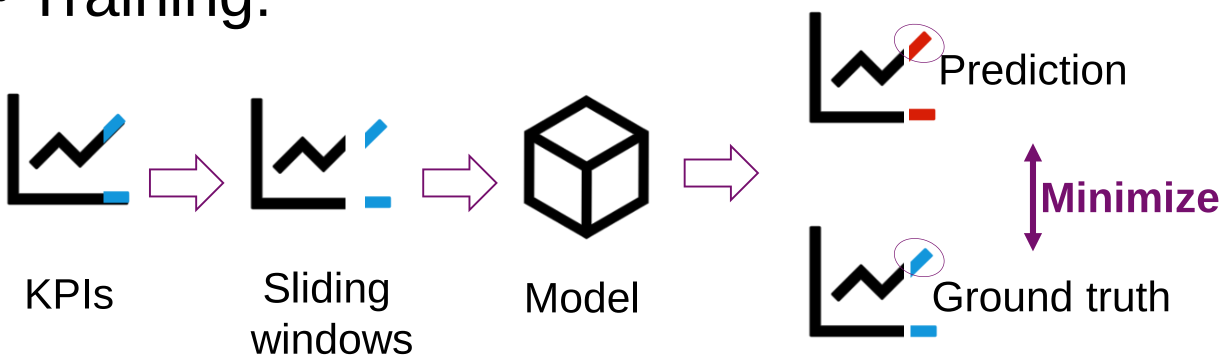
- Should capture dependency of multivariate KPIs
- Unsupervised anomaly detection



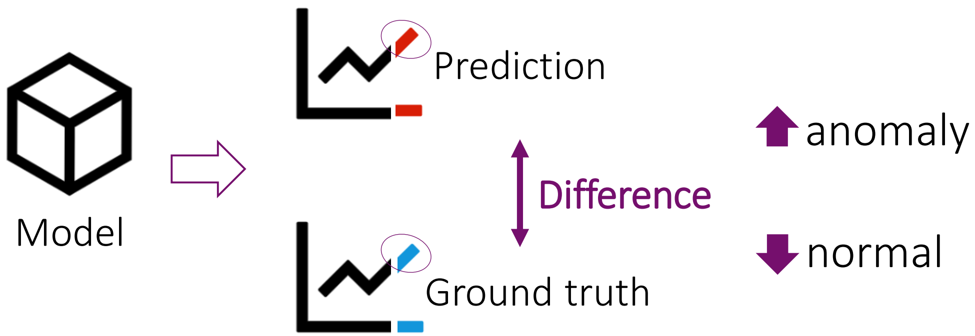


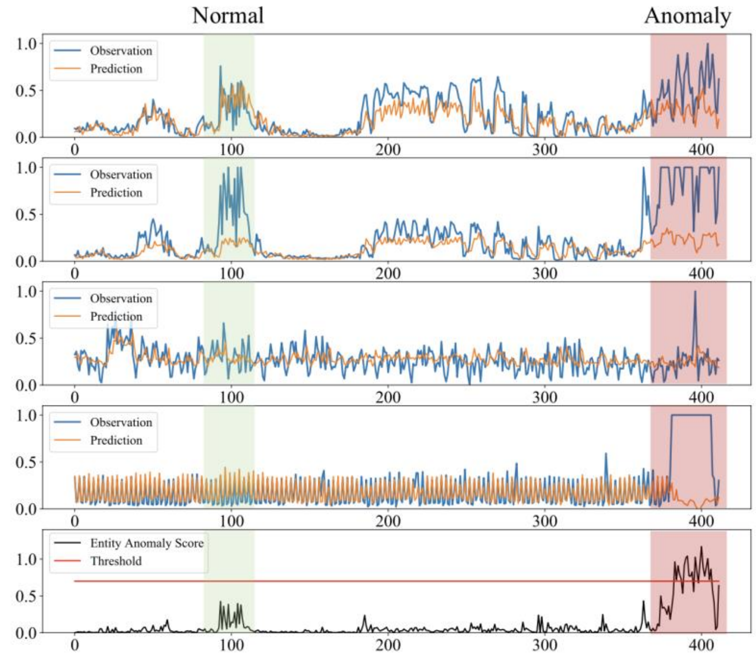
Machine Learning Algorithms

- Training:



- Detection:





AIOps: Correlation between Logs and KPIs

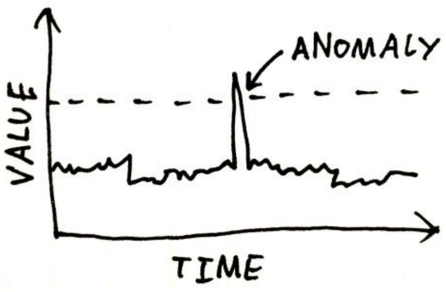
Log

Meter Data

Topology

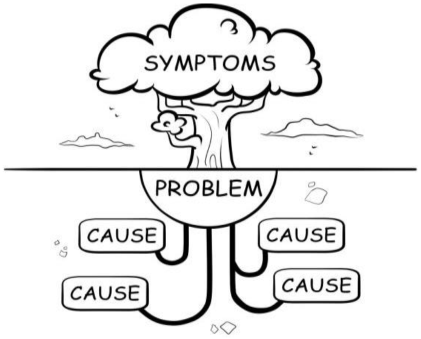
Alert

Incident Ticket



Anomaly Detection

Failure Diagnosis

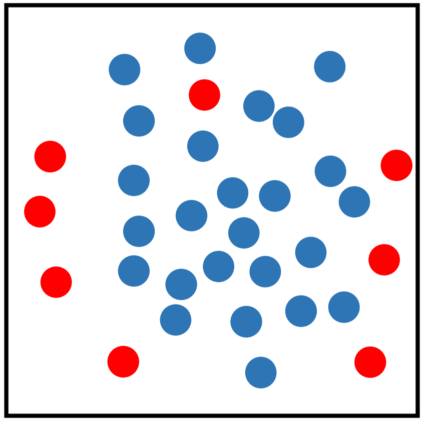


Root Cause Analysis

Failure Prediction

Two Automated Log Analysis Tasks

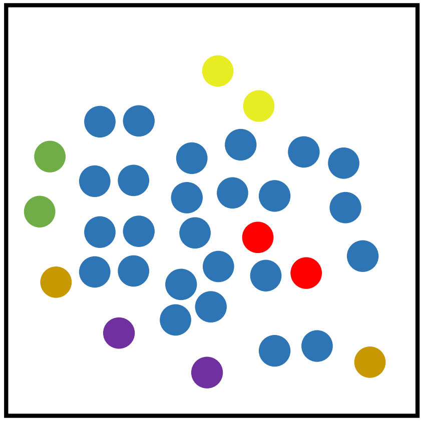
Anomaly Detection
(binary classification)



Normal

Anomalous

Problem Identification
(multiclass classification)

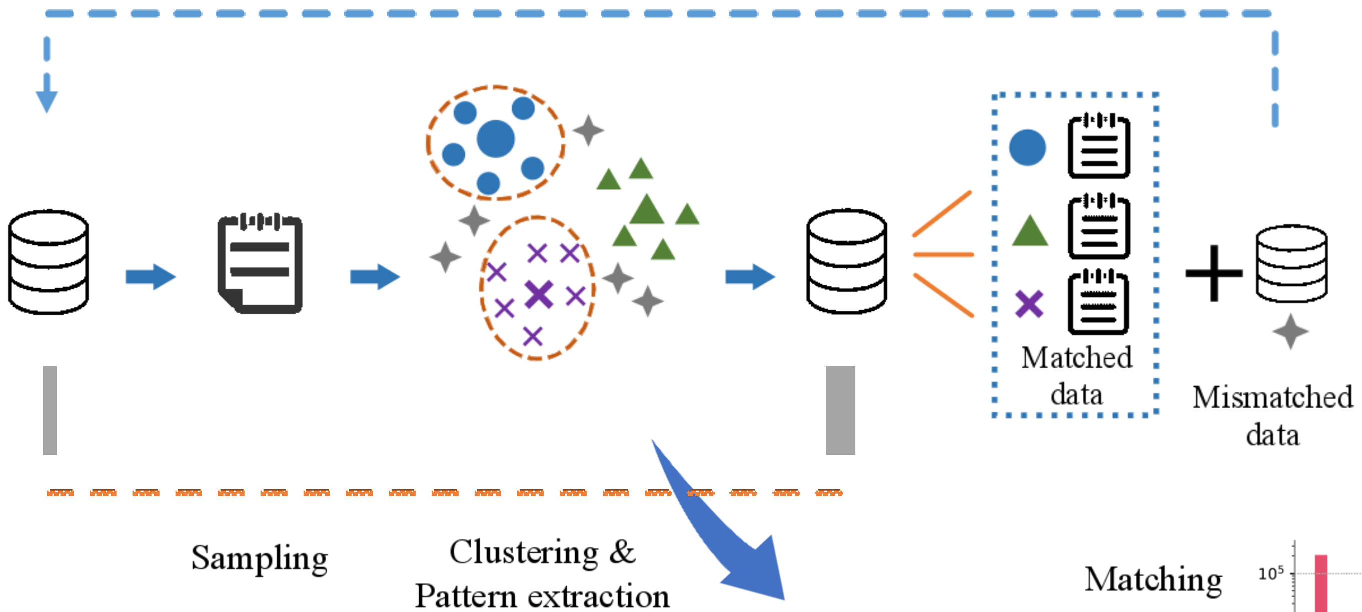


Normal

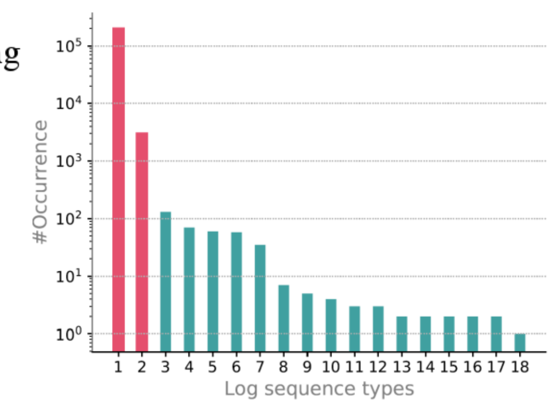
Different types of problem

Efficient Multi-class Classification / Clustering

- Efficient and effective cascading clustering

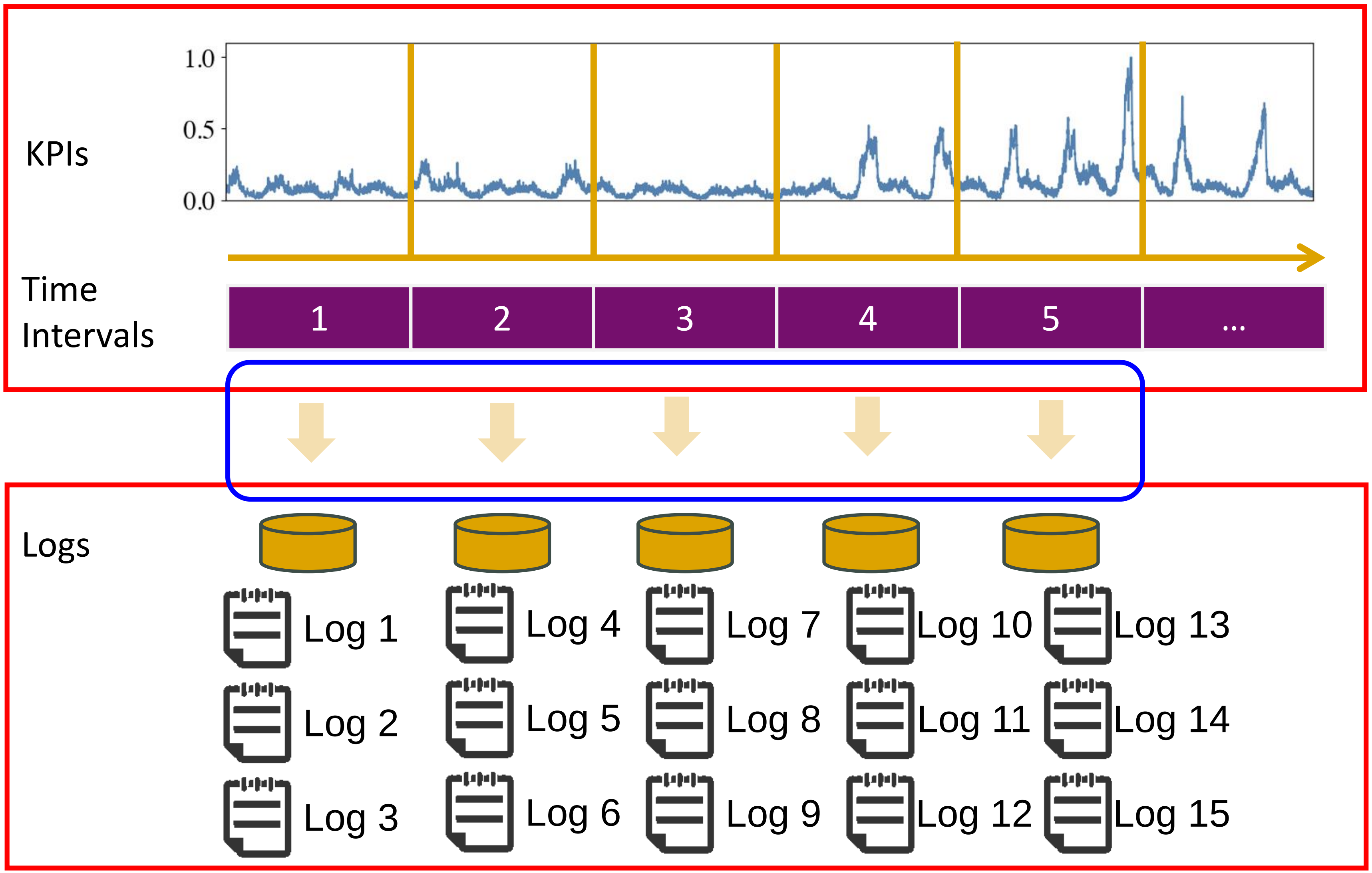


Hierarchical Clustering



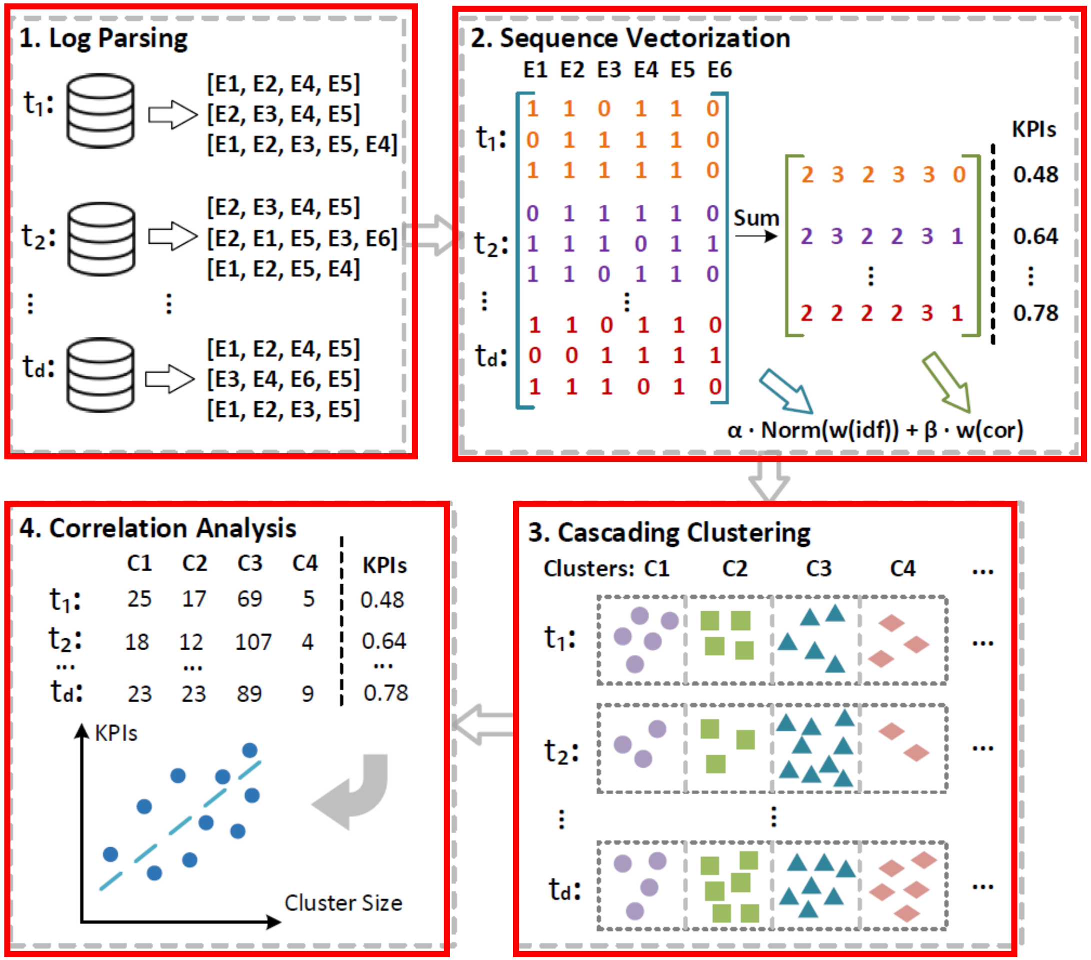
S. He, Q. Lin, J. Lou, H. Zhang, M. R. Lyu and D. Zhang, "Identifying impactful service system problems via log analysis," FSE, 2018.

Relation between Log and KPI



Problem Identification

- Impactful problems:
  - Can lead to the degradation of KPI.
- **Target:**
  - Identify clusters that are highly correlated with KPI's changes.
- **Method:**
  - Model the relation between cluster sizes and KPI values



S. He, Q. Lin, J. Lou, H. Zhang, M. R. Lyu, D. Zhang, "Identifying impactful service system problems via log analysis," FSE, 2018.

Problem Identification

- Evaluation on real Microsoft Azure data

Table 1: Summary of Service X Log Data

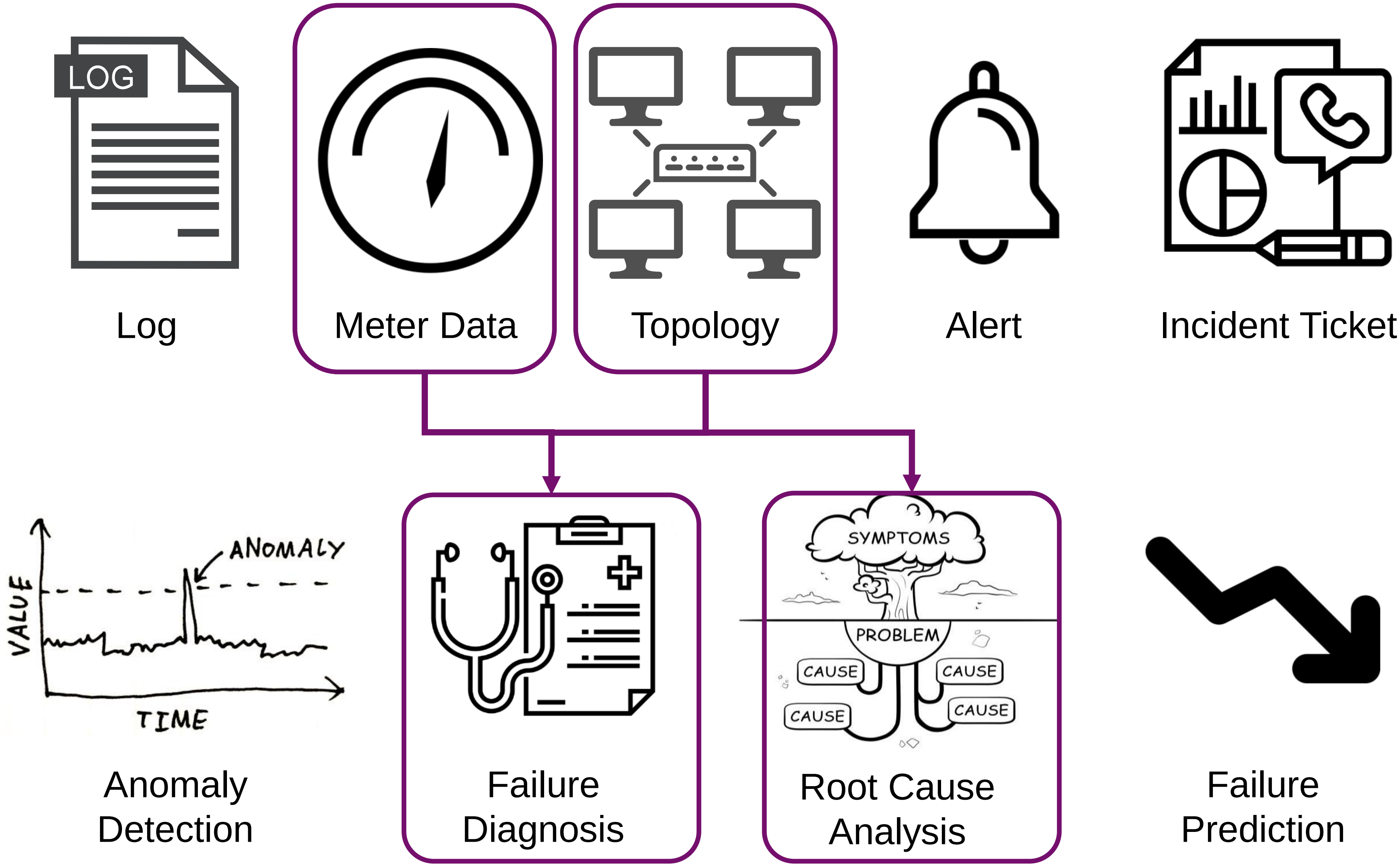
Data	Snapshot starts	#Log Seq (Size)	#Events	#Types
Data 1	Sept 5th 10:50	359,843 (722MB)	365	16
Data 2	Oct 5th 04:30	472,399 (996MB)	526	21
Data 3	Nov 5th 18:50	184,751 (407MB)	409	14

Table 2: Accuracy of Problem Detection on Service X Data

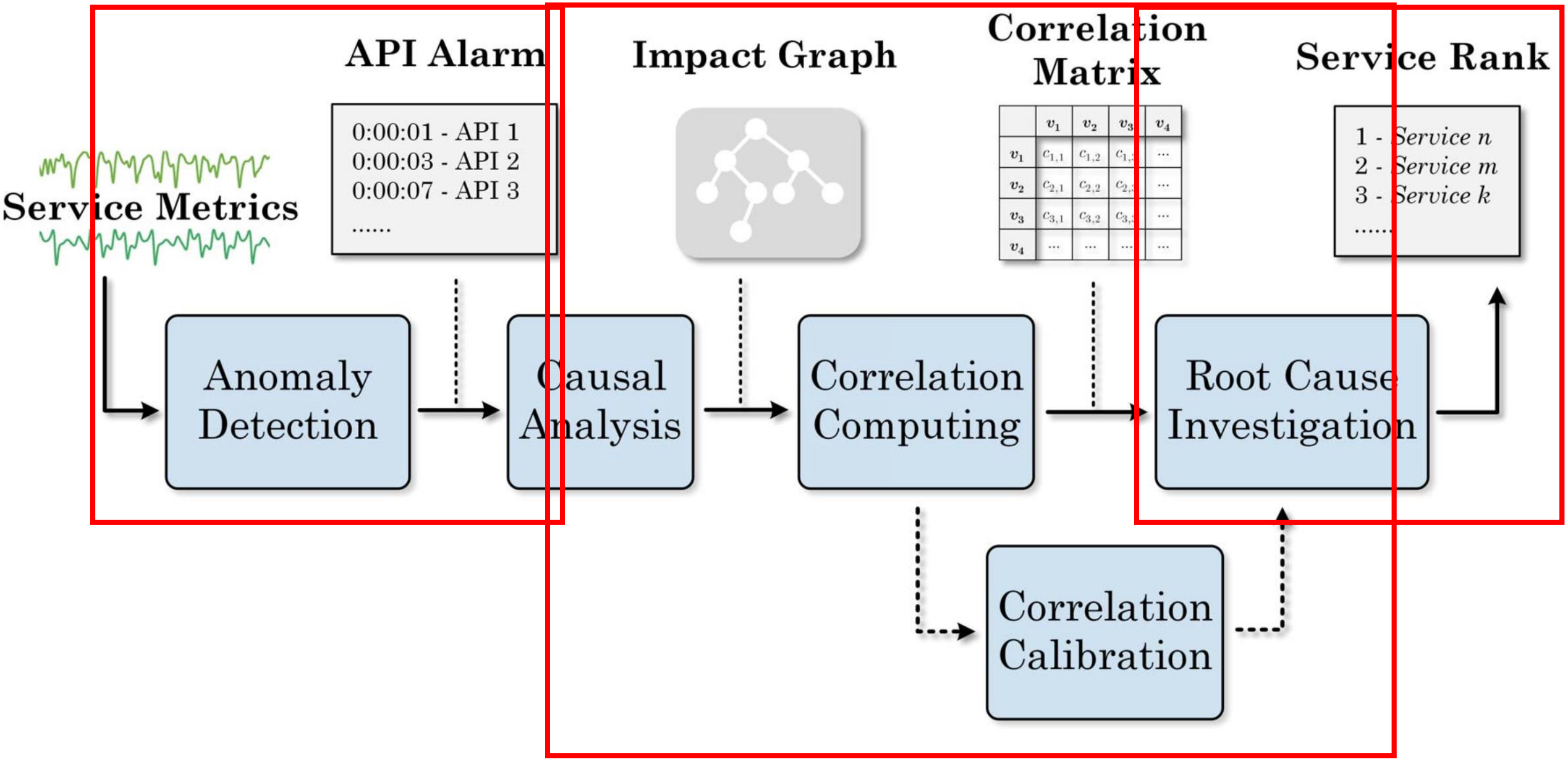
Data	Data 1			Data 2			Data 3		
Metrics	Precision	Recall	F1-measure	Precision	Recall	F1-measure	Precision	Recall	F1-measure
PCA	0.465	0.946	0.623	0.142	0.834	0.242	0.207	0.922	0.338
Invariants Mining	0.604	1	0.753	0.160	0.847	0.269	0.168	0.704	0.271
Log3C	**0.900**	0.920	**0.910**	**0.897**	0.826	**0.860**	**0.834**	0.903	**0.868**

S. He, Q. Lin, J. Lou, H. Zhang, M. R. Lyu, D. Zhang, "Identifying impactful service system problems via log analysis," FSE, 2018.

AIOps: Service Dependency



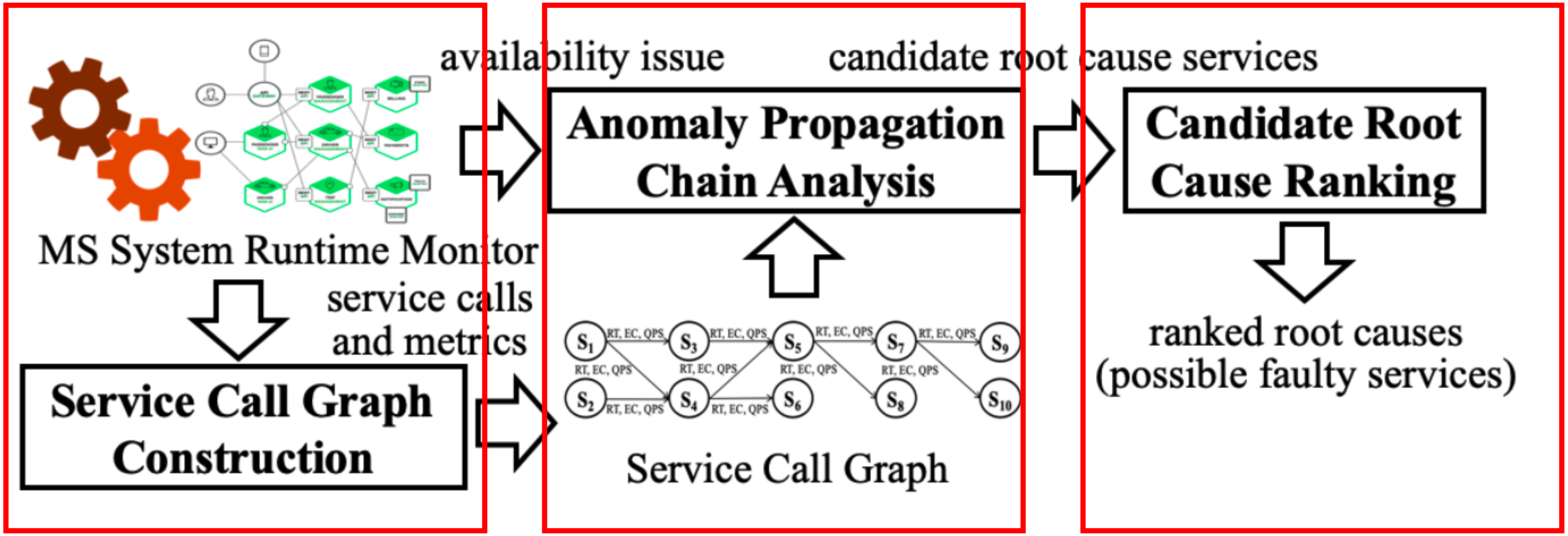
From Correlation to Root Cause Investigation



W. Ping, J. Xu, M. Ma, W. Lin, D. Pan, Y. Wang, and P. Chen. 'CloudRanger: Root Cause Identification for Cloud Native Systems'. *CCGRID 2018*

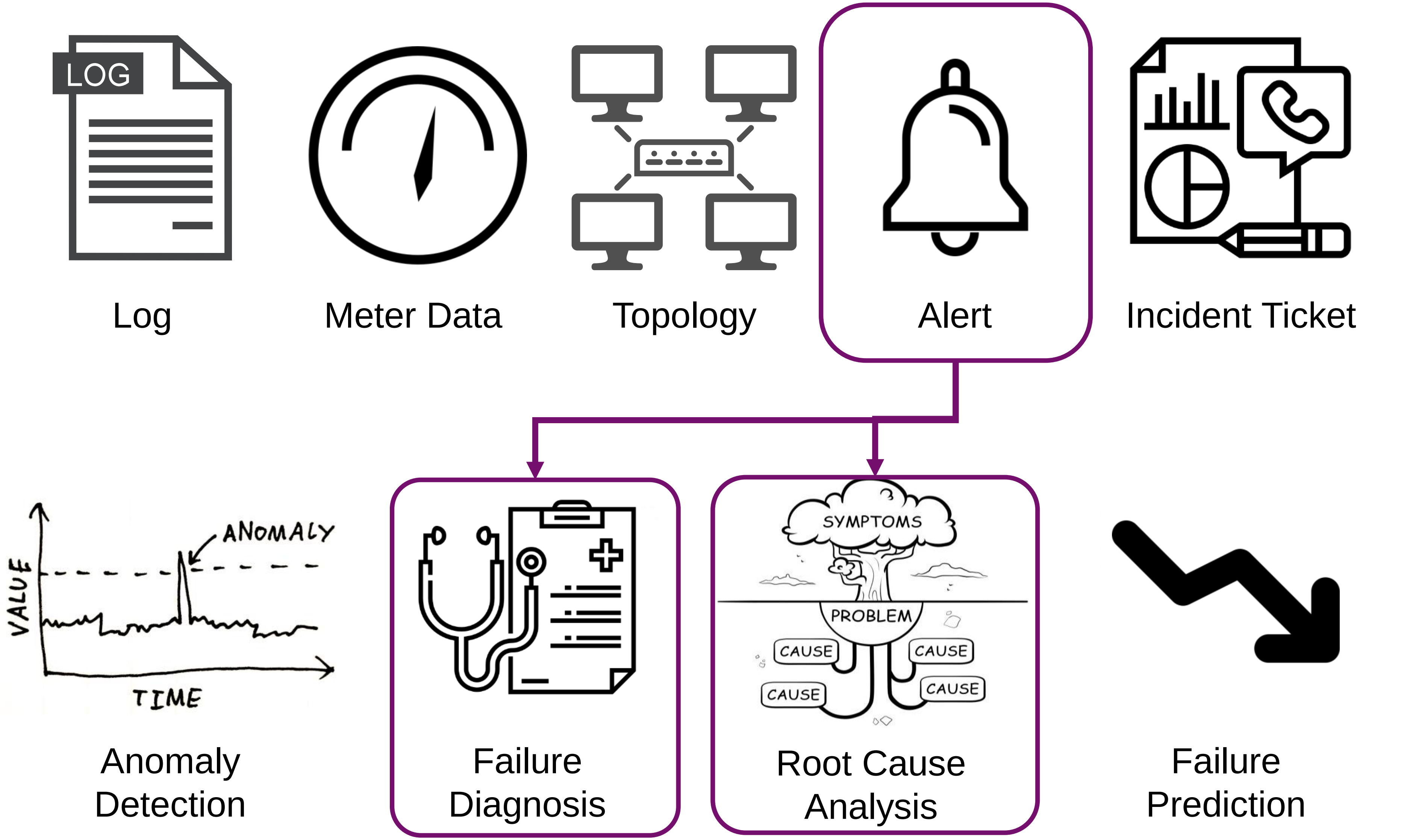
Root Cause Analysis: Service Call Graph

- Metric data: response time, error counts, queries per seconds
- Anomaly propagation chains
- Rank candidate root causes based on correlation analysis



D. Liu, C. He, X. Peng, F. Lin, C. Zhang, S. Gong, Z. Li, J. Ou, and Z. Wu. 'MicroHECL: High-Efficient Root Cause Localization in Large-Scale Microservice Systems'. *ArXiv:2103.01782,* 2021

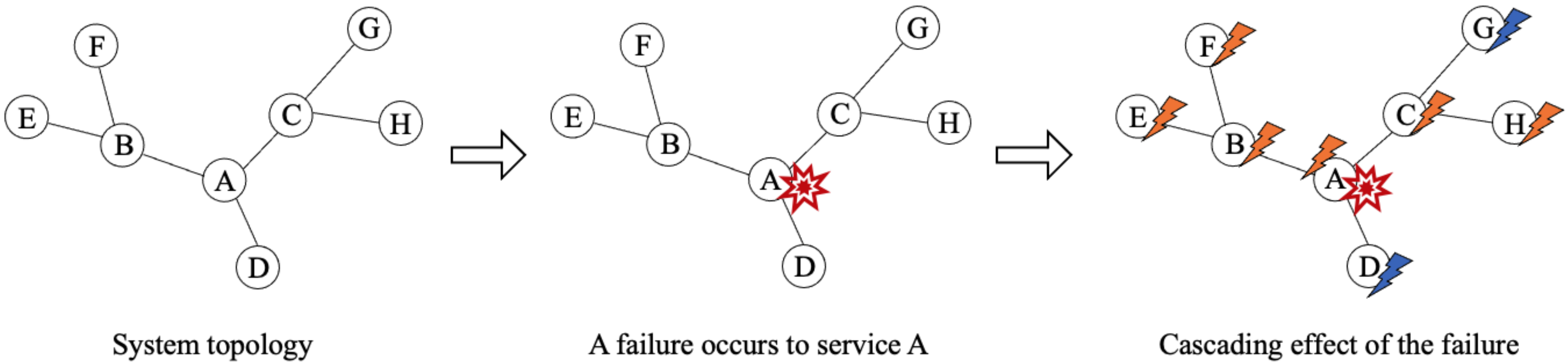
AIOps: Alert Aggregation



Objectives

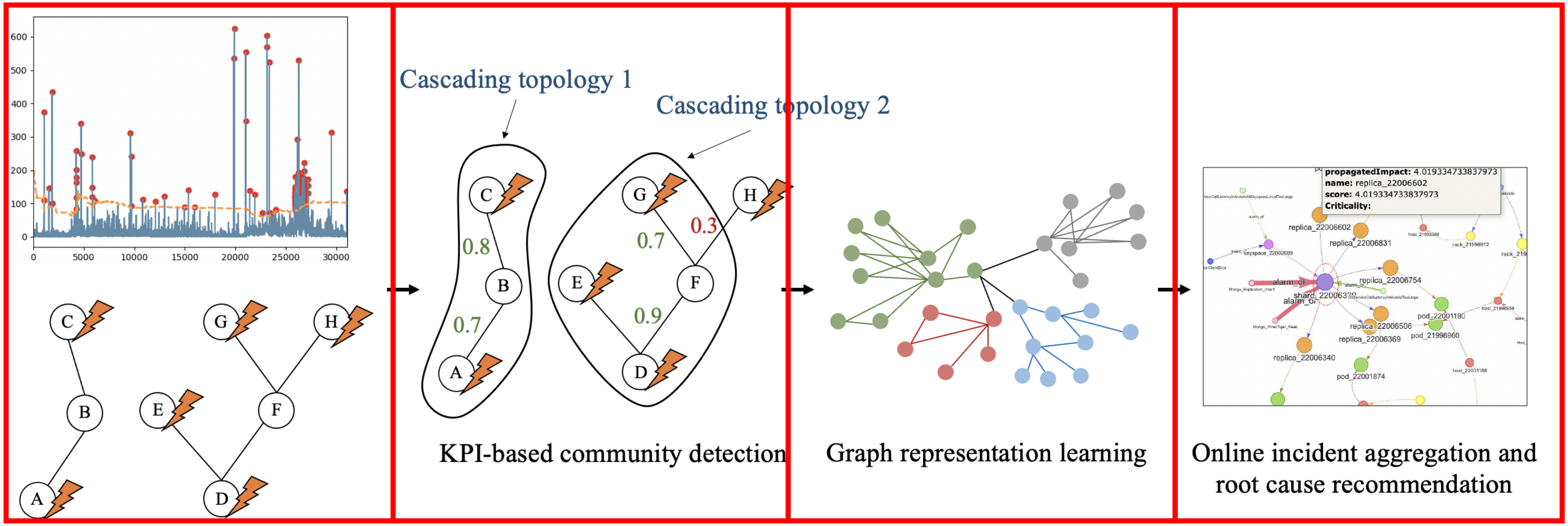
- Alert aggregation
  - Group alerts associated the same failure
  - Narrow down the problem scope

- Root cause recommendation
  - Recommend culprit incidents
  - Speed up fault localization



Graph Representation Learning

- Fine-grained cloud monitoring data to auto-complete the graphs
- Temporal and topological relationship to learn the alert representation vector



Cloud failure detection

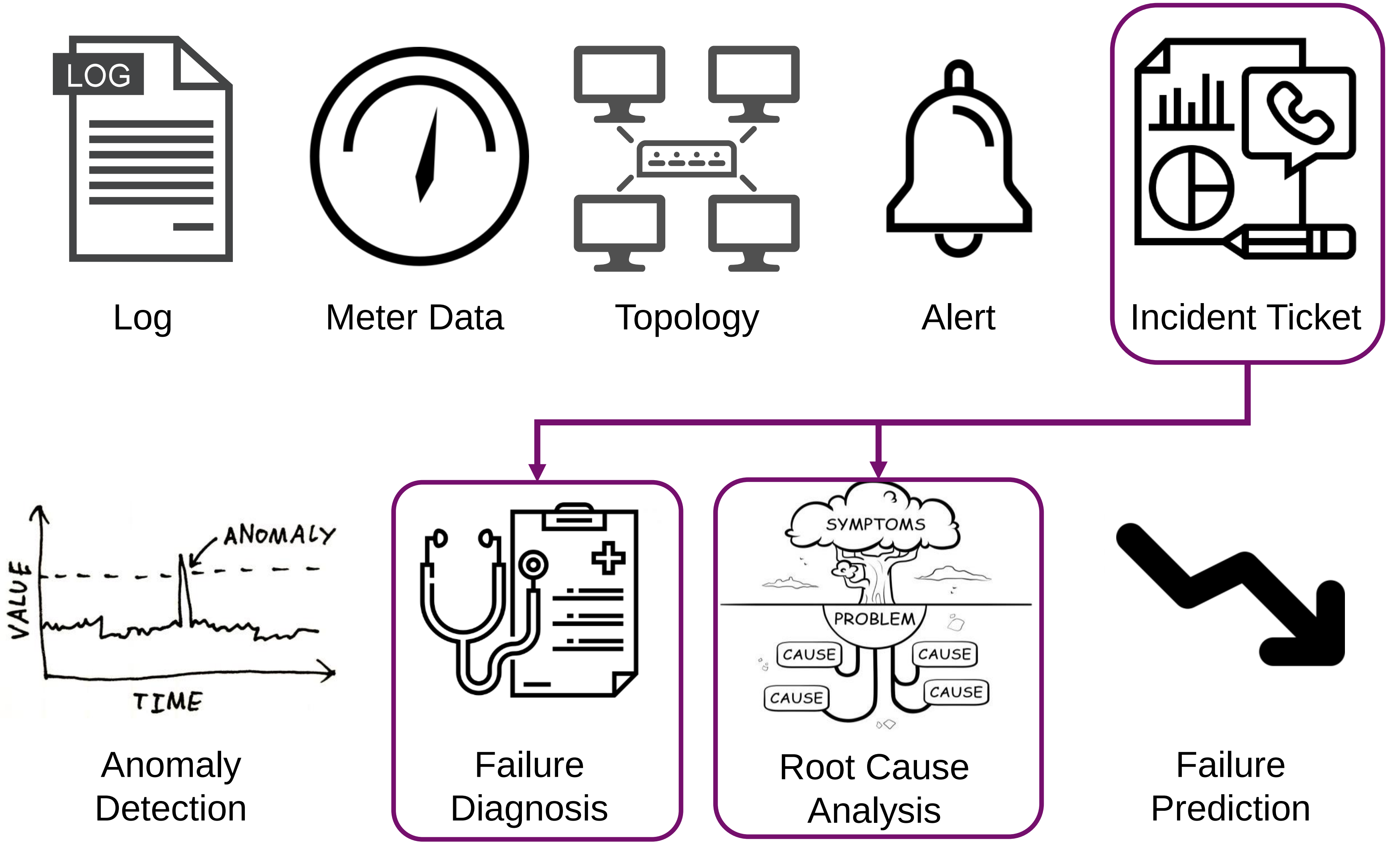
Graph Representation Learning

- Fine-grained cloud monitoring data to auto-complete the graphs
- Temporal and topological relationship to learn the alert representation vector

		FP-Growth	TF-IDF	Zhao's approach	Our approach
Online Incident Aggregation	NMI	0.42	N/A	0.61	**0.9**
Root Cause Recommendation	Precision	N/A	0.73	0.81	**0.91**
	Recall	N/A	0.77	0.88	**0.93**
	F1 score	N/A	0.75	0.85	**0.92**

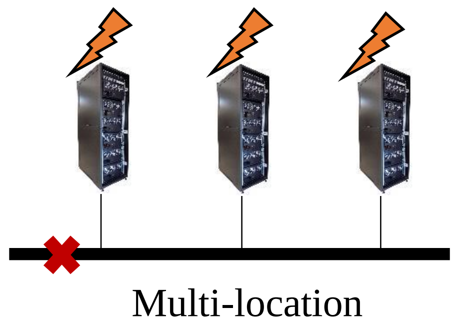
A real case in a top public cloud

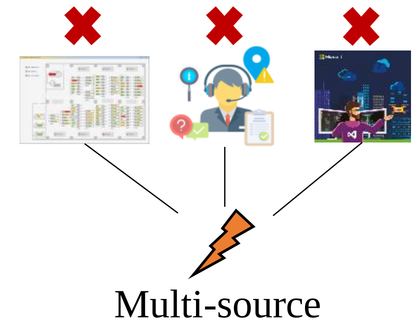
AIOps: Incident Management

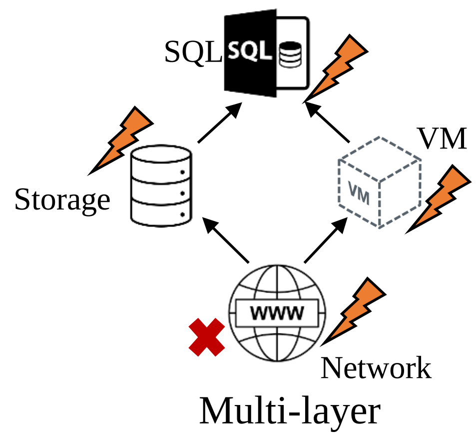


Inefficient and Error-prone Workflow

- Significant delays
  - Critical incident detection
  - Impact scope identification
  - Root cause analysis
  - etc.
- Complicated root causes
  - Multi-location
  - Multi-source
  - Multi-layer
  - etc.



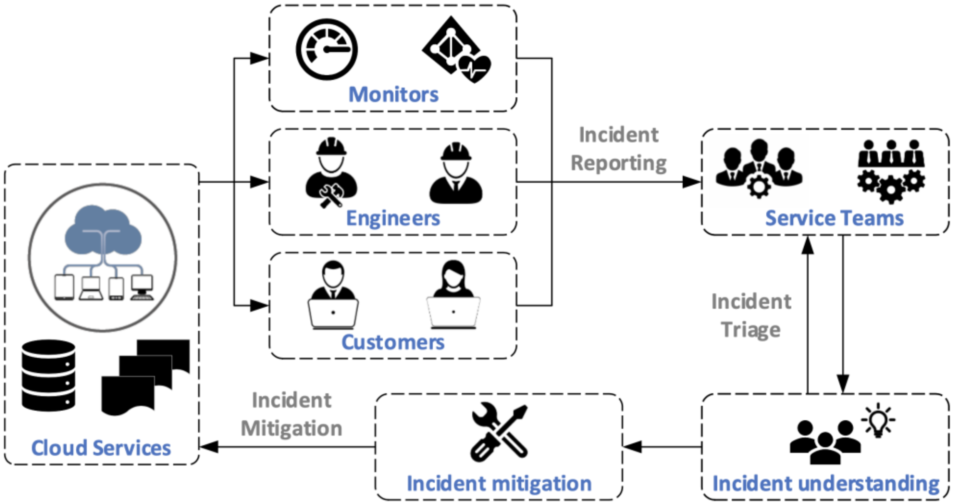


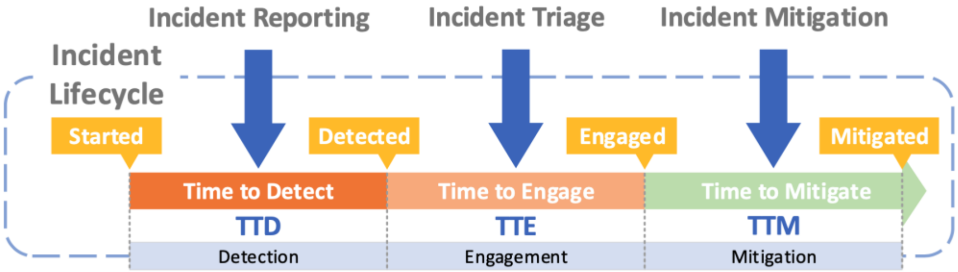


Incident Management

Incident management procedure

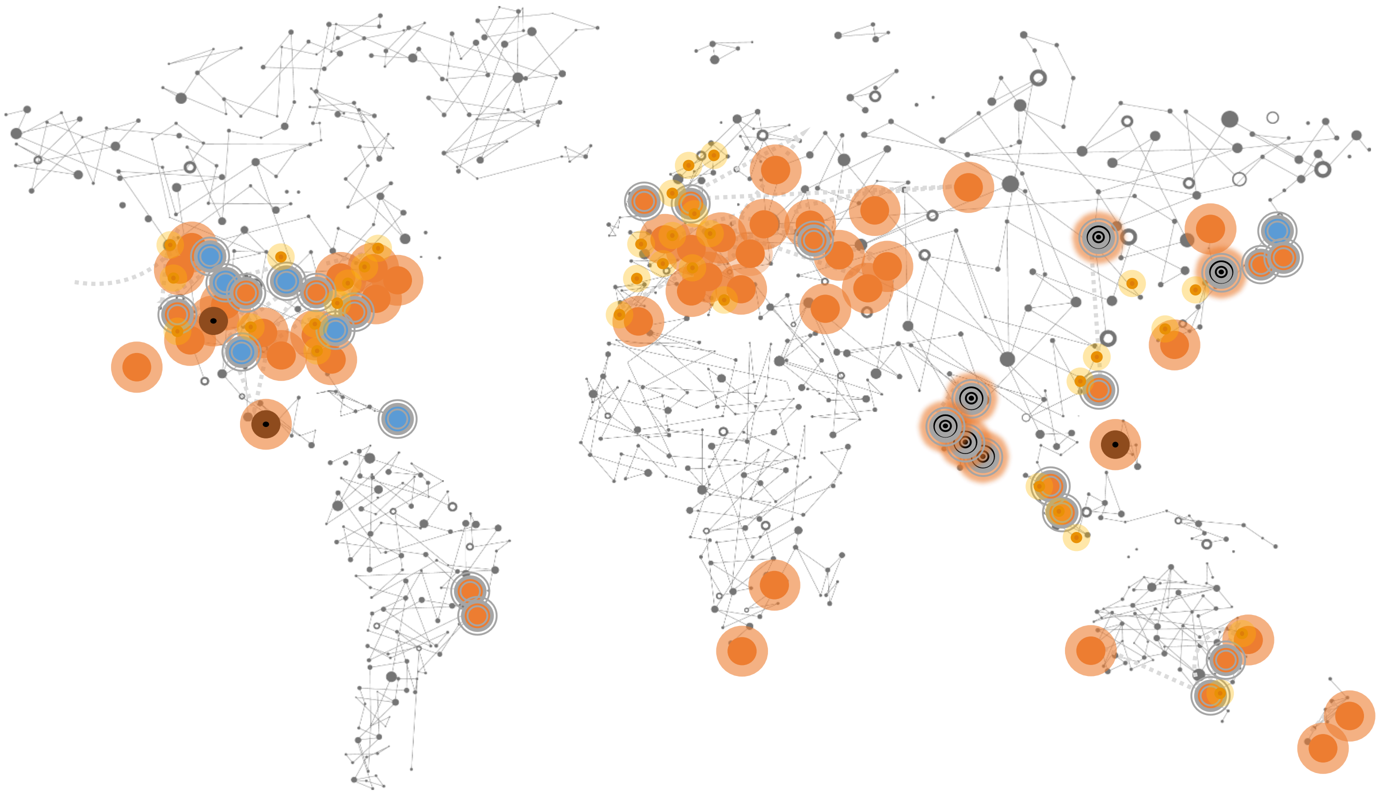
- Incident reporting
  - Time to detect (TTD)
- Incident triage
  - Time to engage (TTE)
- Incident mitigation
  - Time to mitigate (TTM)





Incident Mitigation

- Incident mitigation is important yet challenging
  - Large volume of incidents
  - Cross-region failures
  - Cloud system complexity
  - etc.



Characteristics of Incidents

- Incident severity
  - Low + Medium incidents > 90%
  - High incidents from 1.21% (Network) to 5.48% (DCM)
  - Critical incidents < 0.5%

	DCM	Network	Storage	Compute	Database	WS
Critical	0.01%	0.01%	0.01%	0.31%	0.40%	0.07%
High	5.48%	1.21%	2.57%	5.27%	4.32%	3.33%
Medium	86.65%	46.90%	43.32%	74.19%	63.93%	84.52%
Low	7.86%	51.88%	54.10%	20.23%	31.35%	12.08%

Distribution of incident severity

Chen et al., 'Towards Intelligent Incident Management: Why We Need It and How We Make It'. FSE 2020

Characteristics of Incidents

- Incident fixing time
  - Time to fix (TTF) = TTD+TTE+TTM
  - TTF of Low & Medium incidents > TTF of High incidents
  - TTF of Critical is the largest

	DCM	Network	Storage	Compute	Database	WS
Critical	**38.33x**	8.46x	**10.06x**	**142.05x**	**209.97x**	**286.6x**
High	19.25x	3.18x	2.52x	2.56x	5.75x	3.56x
Medium	1x	**9.8x**	7.09x	2.95x	25.28x	12.93x
Low	3.01x	5.49x	1.09x	11.65x	2.41x	144.79x

Distribution of incident fixing time

Chen et al., 'Towards Intelligent Incident Management: Why We Need It and How We Make It'. FSE 2020

Characteristics of Incidents

- Root Cause:
  - Network Issue
  - Human Error
  - Deployment Issue
  - External Issue
  - Capacity Issue
  - Others

30.6%

Root Cause	Dist.	Root Cause	Dist.
Network (Hardware)	**22.95%**	Human Error (Code Defect)	**19.23%**
Network (Connectivity)	2.24%	Human Error (Config.)	**7.45%**
Network (Config.)	0.89%	Human Error (Design Flaw)	5.66%
Network (Other)	4.47%	Human Error (Integration)	2.09%
Deployment (Upgrade)	5.22%	Human Error (Other)	2.83%
Deployment (Config.)	3.87%	External Issue (Partner)	2.83%
Deployment (Other)	1.19%	External Issue (Other)	1.64%
Capacity Issue	**6.56%**	Others	10.88%

37.3%

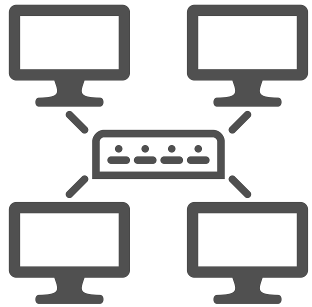
Distribution of incident root causes

Chen et al., 'Towards Intelligent Incident Management: Why We Need It and How We Make It'. FSE 2020

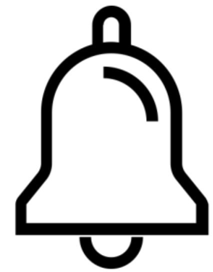
AIOps: Outage Prediction

Log

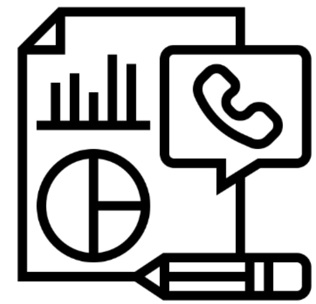
Meter Data



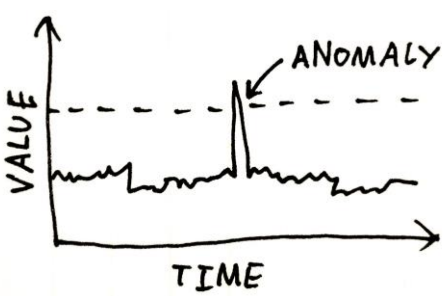
Topology



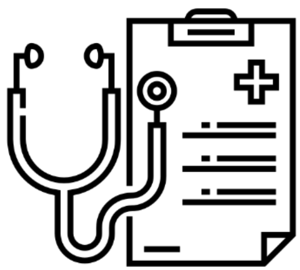
Alert



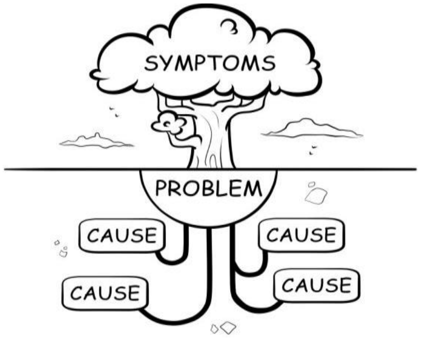
Incident Ticket



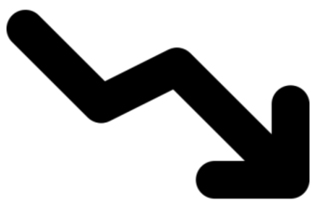
Anomaly Detection



Failure Diagnosis

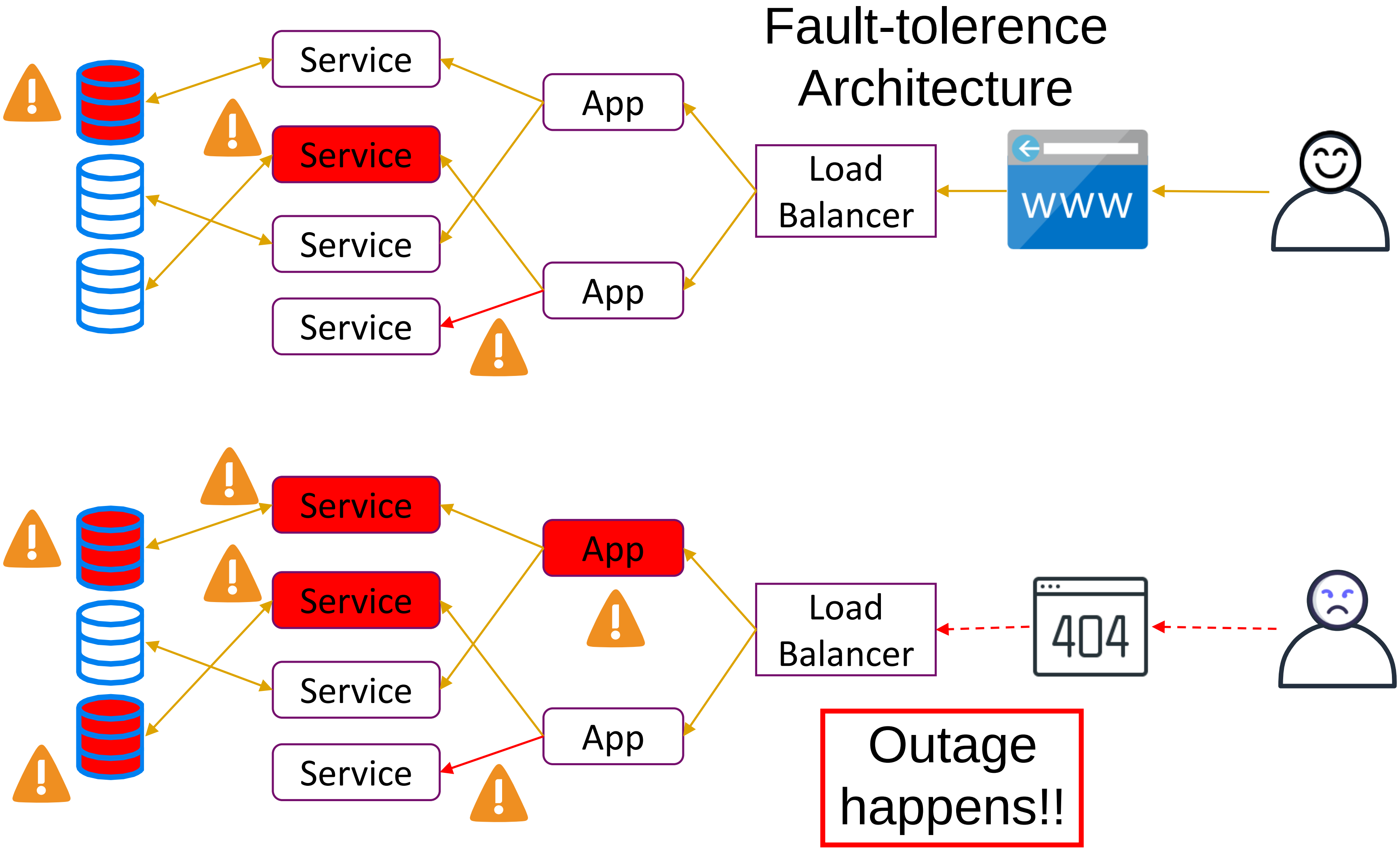


Root Cause Analysis



Failure Prediction

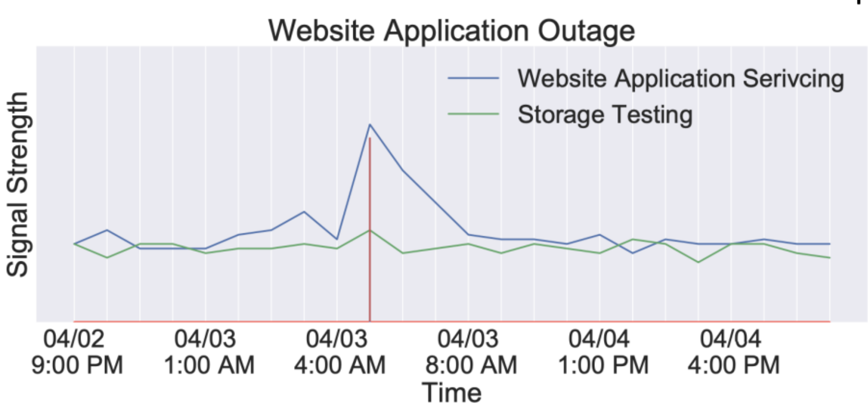
Alerts vs Outage

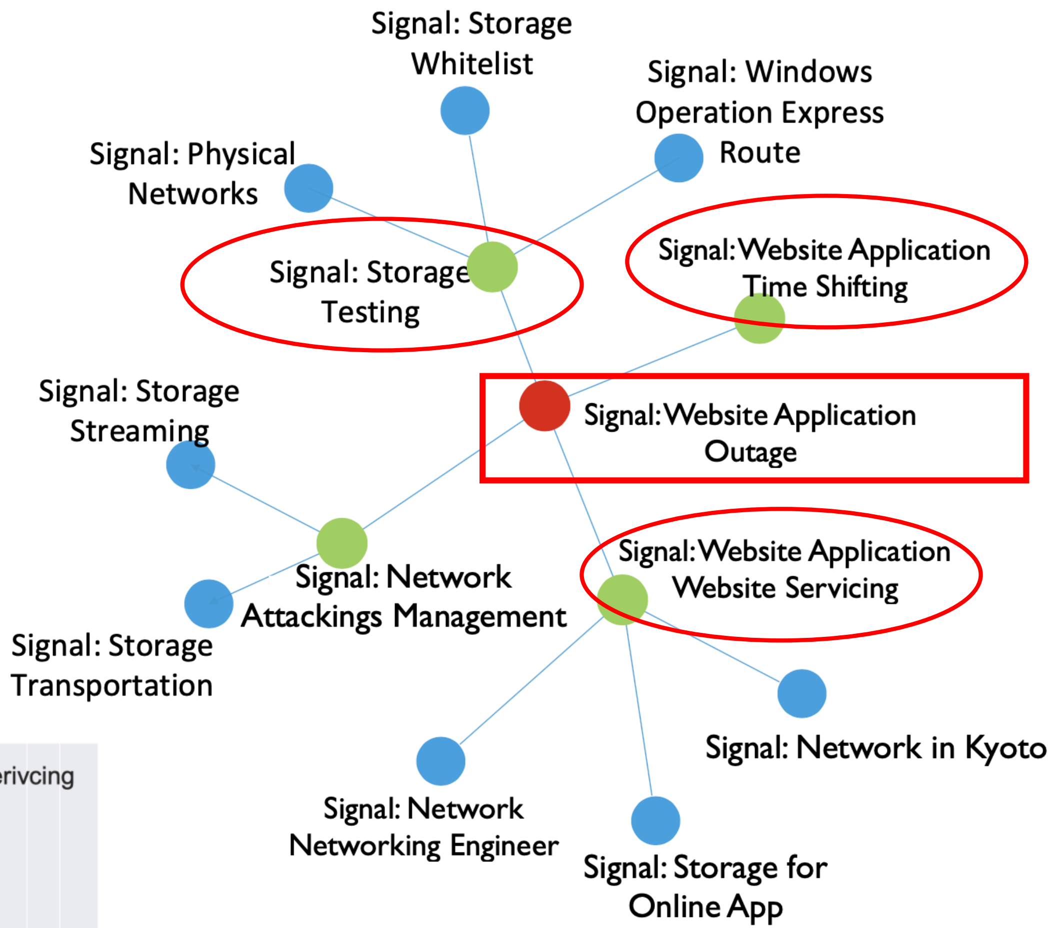


Causal Relationship between Alerts and Outage

- Historical failure statistics
  - Build dependency among alert signals
  - Train classification model to predict outage

Classification models to link alerts and outages





Bayesian Network

Y. Chen, X. Yang, Q. Lin, H. Zhang, F. Gao, Z. Xu, Y. Dang, D. Zhang, H. Dong, Y. Xu, H. Li, Y. Kang, “Outage Prediction and Diagnosis for Cloud Service Systems”, WWW 2019

Causal Relationship between Alerts and Outage

Table 1: Comparison of different methods for component-level outage prediction.

	Outage (Storage Location)			Outage (Physical Networking)			Outage (Storage Streaming)		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Simple Spike	61.65	100.00	76.28	73.71	67.71	70.58	61.52	100.00	76.18
PLR	70.02	92.71	79,78	67.72	83.33	74.72	63.23	91.67	74.84
SVM	65.65	95.83	77.92	63.13	88.54	73.71	58.62	88.64	70.57
AirAlert Related	65.31	100.00	79.01	63.33	98.95	77.25	62.34	100.00	76.80
AirAlert Full	71.11	100.00	83.17	69.07	100.00	81.71	63.75	98.99	77.86

Table 2: Comparison of different methods for service-level outage prediction.

	Outage (Website Application)			Outage (Cloud Network)			Outage (Microsoft Cloud System Operation)		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Simple Spike	5.73	11.83	7.72	4.47	67.74	8.39	7.27	29.03	11.63
PLR	61.18	54.17	57.46	26.27	60.52	36.64	20.36	35.17	25.79
SVM	66.41	88.54	75.89	6.89	88.42	12.78	26.90	22.50	24.50
AirAlert Related	92.18	85.63	88.78	62.08	47.65	53.92	72.40	77.96	75.08
AirAlert Full	82.75	76.74	79.63	75.93	67.07	71.22	72.59	50.15	59.32

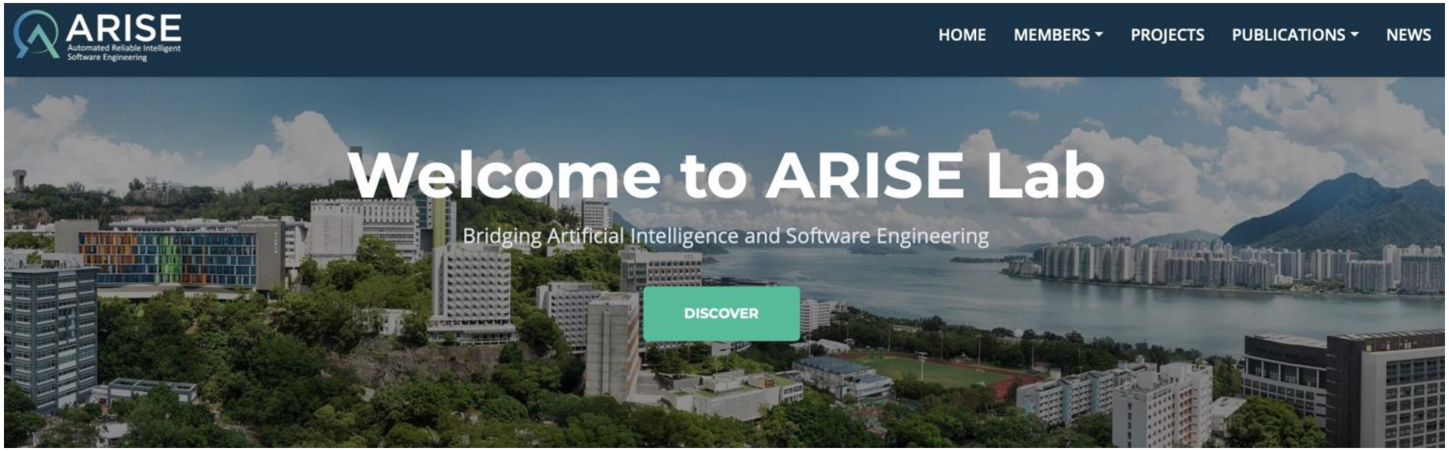
Y. Chen, X. Yang, Q. Lin, H. Zhang, F. Gao, Z. Xu, Y. Dang, D. Zhang, H. Dong, Y. Xu, H. Li, Y. Kang, “Outage Prediction and Diagnosis for Cloud Service Systems”, WWW 2019

Conclusions

- Why cloud resilience needs AIOps?
  - Endless pursuit of reliability
  - From automatic to intelligent, from reactive to proactive
  - Important data sources: log, meter data, topology, alert and incident ticket

- How AIOps achieves reliability goals?
  - Endless pursuit of advnaced algorithms
  - From anomaly detection, fialure diagnosis, root cause analysis to failure prediction
  - Intelligent algorithms designed with human experts' experiences

- What's the next?
  - How to integrate human knowledge with algorithms automatically and comprehensively?
  - Further investigations on AI and Software Engineering



ARISE
Automated Reliable Intelligent
Software Engineering

Thank you!

ICSE21 Workshop on Cloud Intelligence

In conjunction with the 43rd International Conference on Software Engineering

Schedule: 11:00am - 7:30pm CET on May 29th, 2021