

AIOps: Autonomous IT Operations through Machine Learning

Dan Pei

Tsinghua University

Executive Summary

- * AIOps is rising: replace manual Ops decisions with AI-based decision aids
 - * Improved revenue
 - * reduced loss/cost
 - * necessary
 - * feasible
- * Case studies (Collaboration with Baidu, Alibaba, Tencent, Didi, Sogou):
 - * Anomaly Detection
 - * Anomaly Localization
 - * Root Cause Analysis
 - * Capacity/Failure Prediction
- * AIOps Challenge: Community efforts for widespread adoption of AIOps

My life as an Ops Researcher

My Official Resume

2000-2005 UCLA Ph.D., Best Ph.D.
Thesis, working on BGP, OSPF etc.

Summer 2003, Intern at AT&T
Research

2005-2011 Senior/Principal
Researcher at AT&T Research
ACM, IEEE Senior Member

2012-now Associated Professor at
Computer Science Department at
Tsinghua University.

My Operator Resume

For five years, chased ISP OPs for data,
experiences, and insights.

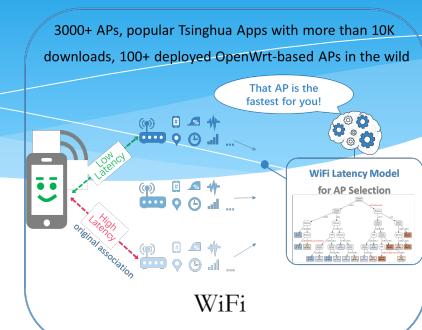
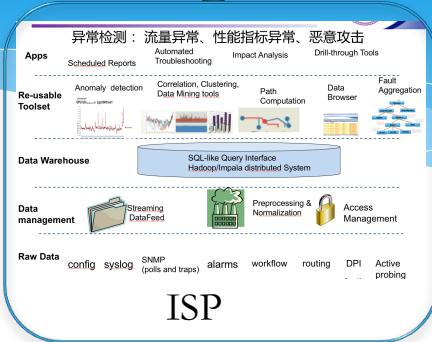
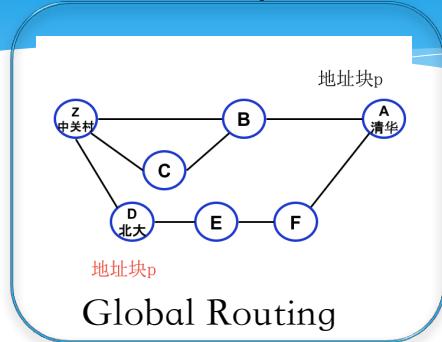
Felt in love with real OP data

Essentially a tier-5 OP

Teaching “Advanced Network
Management.”

Working with OPs at Baidu, Microsoft
Azure, Tencent, Alibaba, DiDi, Sogou

My life as an Ops Researcher



My Official Resume

2000-2005 UCLA Ph.D., Best Ph.D. Thesis, working on BGP, OSPF etc.

Summer 2003, Intern at AT&T Research

2005-2011 Senior/Principal Researcher at AT&T Research
ACM, IEEE Senior Member

2012-now Associated Professor at Computer Science Department at Tsinghua University.

My Operator Resume

For five years, chased ISP OPs for data, experiences, and insights.

Felt in love with real OP data

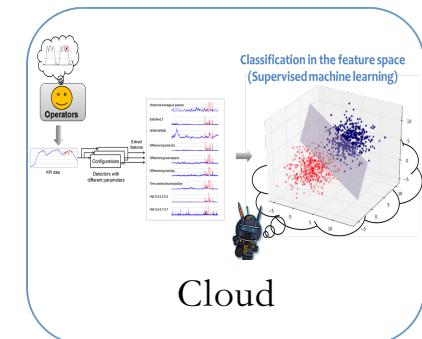
Essentially a tier-5 OP

Teaching “Advanced Network Management.”

Working with OPs at Baidu, Microsoft Azure, Tencent, Alibaba, DiDi, Sogou



Mobile Apps



Cloud

Overview

- * **What is AIOps?**
- * What's the value of AIOps?
- * Is AIOps necessary?
- * Is AIOps feasible?
- * Case studies
- * AIOps Challenge
- * Summary

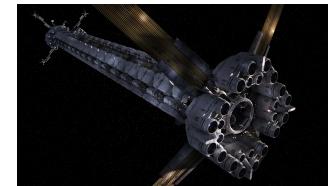
Towards Autonomous IT Operations



Manual-
Driven

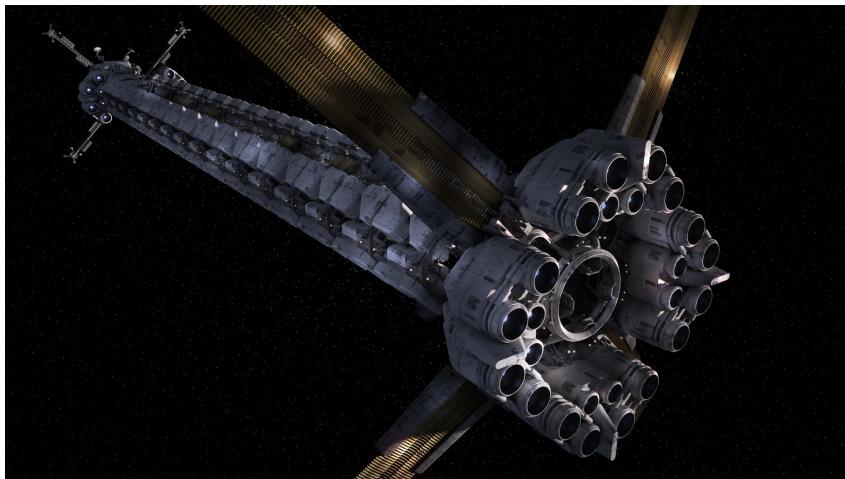
Automated but with
Manual Decision

Autonomous



Ultimate Goal: Autonomous IT Operations

AIOps: **Algorithmic** IT Operations -> **Artificial Intelligence** for IT Operations -> **Autonomous** IT Operations



Spaceship Covenant: 2000 passengers and 15 crew members all in hibernation. Flying towards Planet Origae-6. Only one awoken android crew.



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

Autonomous IT Operations: Automatically deal with all four causes of changes to IT systems

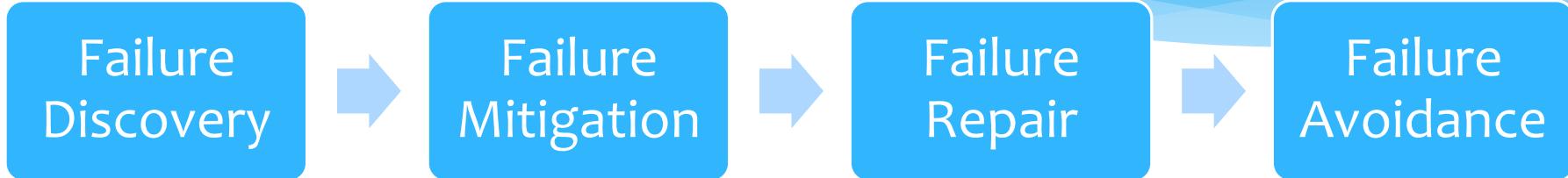
- * Software & hardware failures--> Automatic Healing
- * Software changes --> Autonomous software deployment
- * Change of user request amount & Pattern --> Elastic Resource Allocation
- * Malicious Attacks-->Autonomous Defense



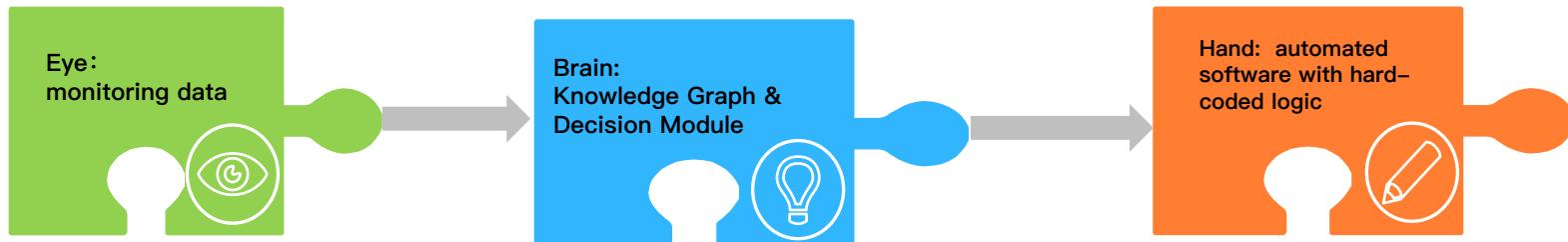
“Most people overestimate what they can do in one year and underestimate what they can do in ten years.”

-- Bill Gates

It's the responsibility of the Operations to ensure undisrupted services, despite the inevitable failures of the imperfect underlying hardware and software.



AIOps Architecture & Algorithms



Metrics, Logs, Traces, Changes of
Application, Middleware,
Database, Storage, Network, Server,
etc.

Mitigation: rollback, HotFix, reboot,
reactive traffic switching, reactive
capacity upsizing

Repair: replace faulty hardware, fix
bugs, refactor code

Avoidance: preventive hardware
replacement; preventive traffic
switching; targeted capacity upsizing

Overview

- * What is AIOps?
- * **What's the value of AIOps?**
- * Is AIOps necessary?
- * Is AIOps feasible?
- * Case studies
- * AIOps Challenge
- * Summary

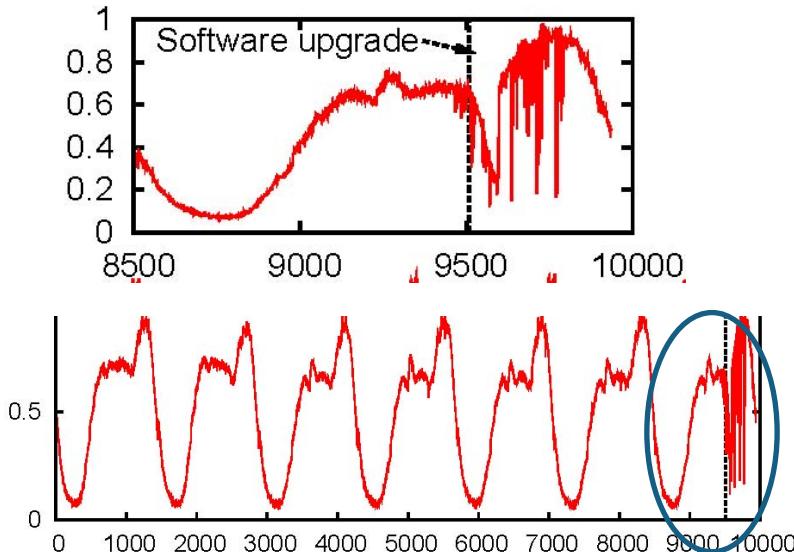
Reduced Business Loss: Rapid Assessment of Software Changes

- * A buggy deployment causes significant revenue Loss
- * Manual trouble shooting takes 1.5 hours



- * AIOps solution takes less than 10 minutes

Joint Work with Baidu
Published in ACM CoNext 2015



Improved Revenue: Reduced Page Response Time

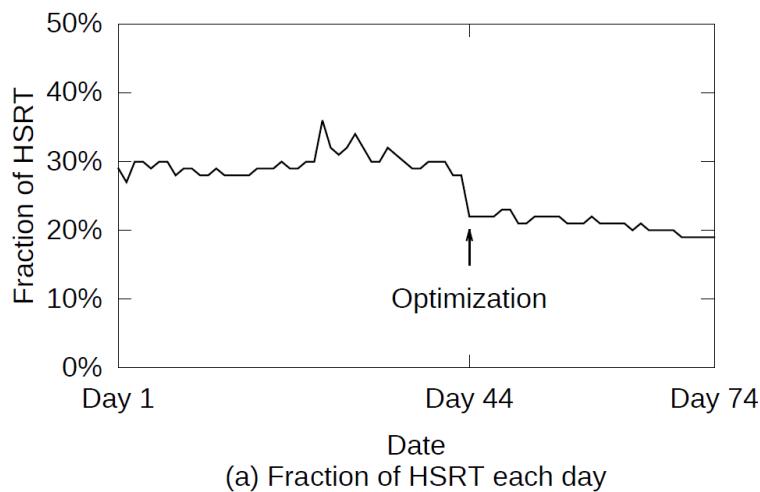


-100ms -> Sales ↑ 1%
[Greg Linden, Amazon]



-100ms~400ms -> Revenue ↑ 0.2%~0.6%
[Jake Brutlag, Google]

After deploying the solutions suggested by AIOps:



Slow responses (>1s) are reduced from 30% to 20%

80th-percentile response time is reduced by 253 ms

Saves 30 man-months (estimated) of manual analysis

Joint Work with Baidu
Published in IEEE INFOCOM 2016

AIOps Leads to Better User Experience -> Longer Engagement -> More Revenue

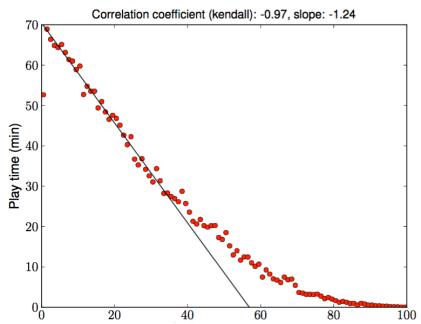
Linear
Regression
SIGCOMM
2011



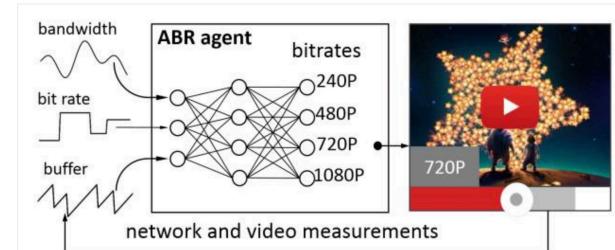
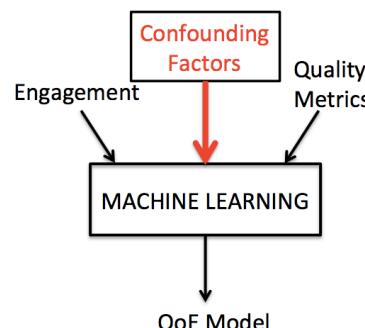
Decision
Trees
SIGCOMM
2013



Reinforcement
Learning
SIGCOMM
2017

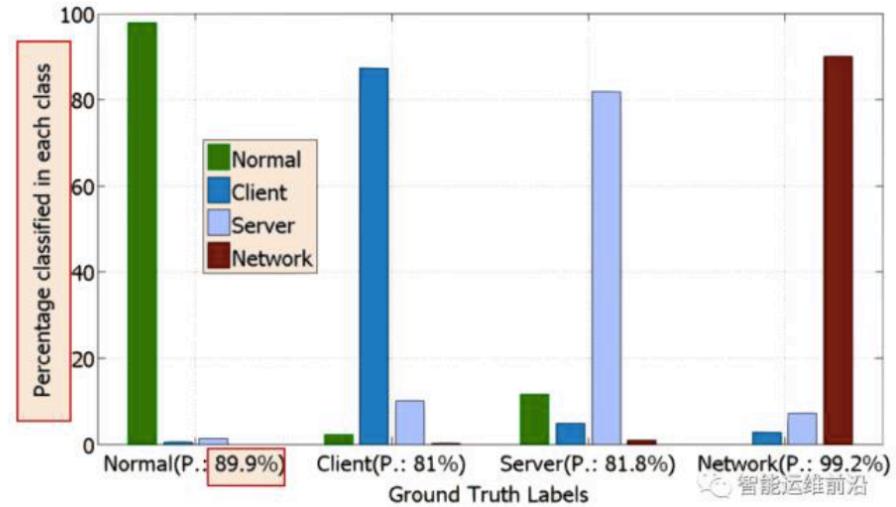
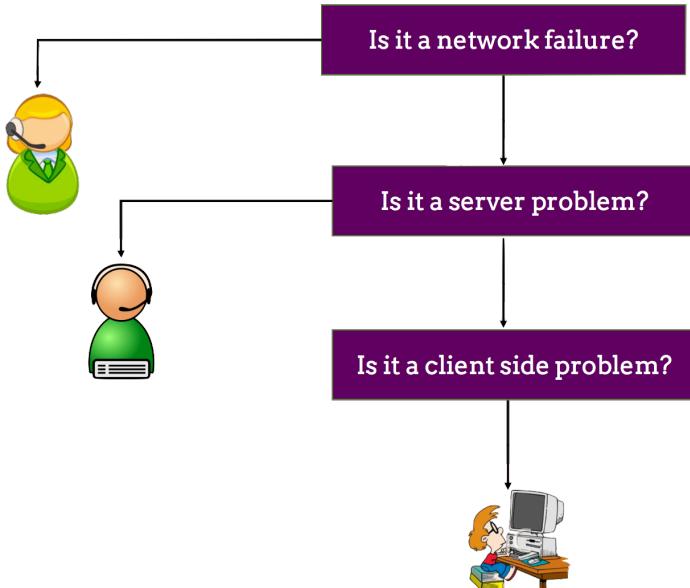


Adding as a feature



Conviva/CMU/MIT work

AIOps Quickly Decides the Responsibility Boundary: Reduced loss



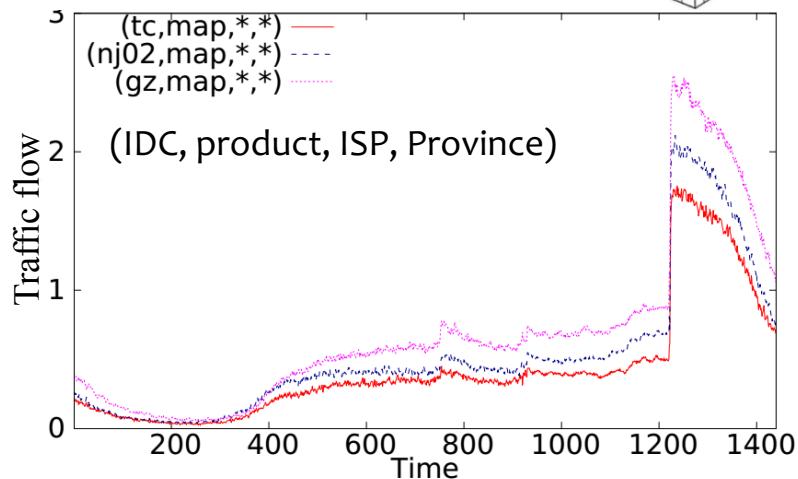
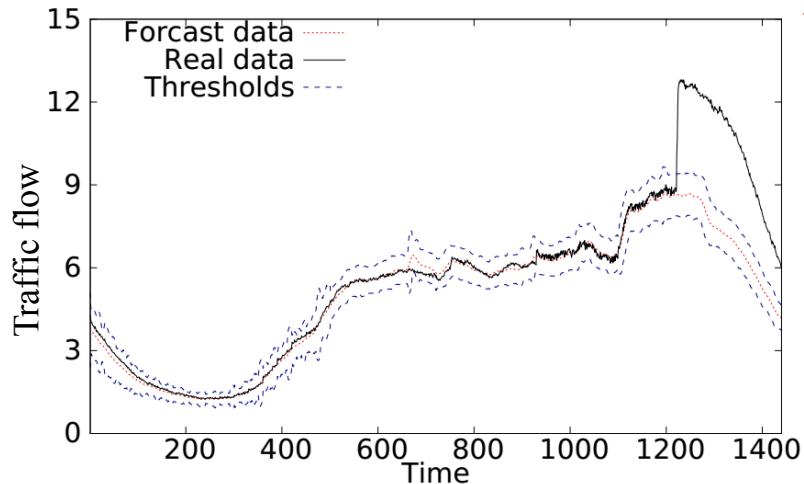
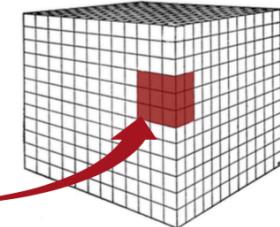
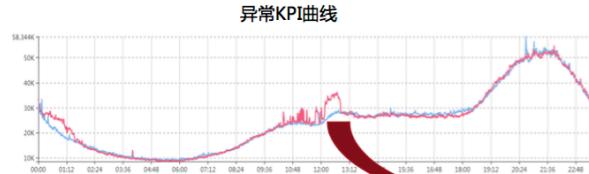
Microsoft Azure Work.
Published in SIGCOMM 2016

Localizing the Anomalous Regions: Reduced Loss

Manual localization: 90 minutes
AIOps: 30s

如何快速找到大量组合中最核心的影响因素？

异常维度组合



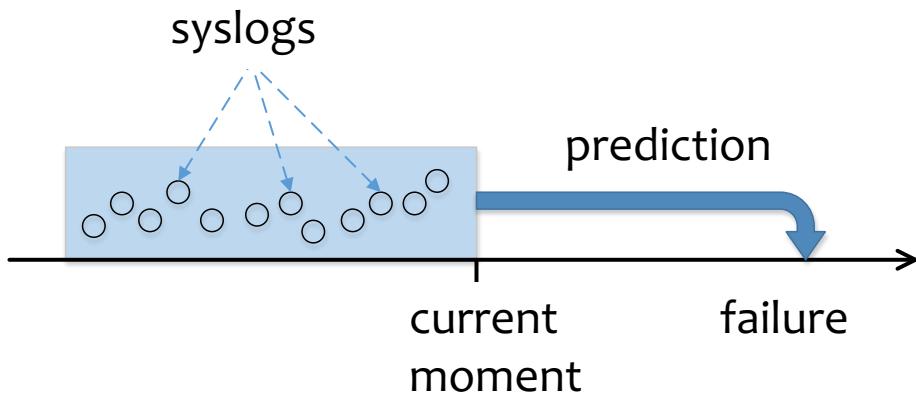
Collaboration with Baidu. Tencent implemented a variant to improve its video streaming service

DC Switch Failure Prediction->Preventive Replacement->Avoided Loss

Problem: Baidu-customized switches intermittently drop/delay packets, causing QoE drop at the application layer.

Reboot stops the problem for some while.

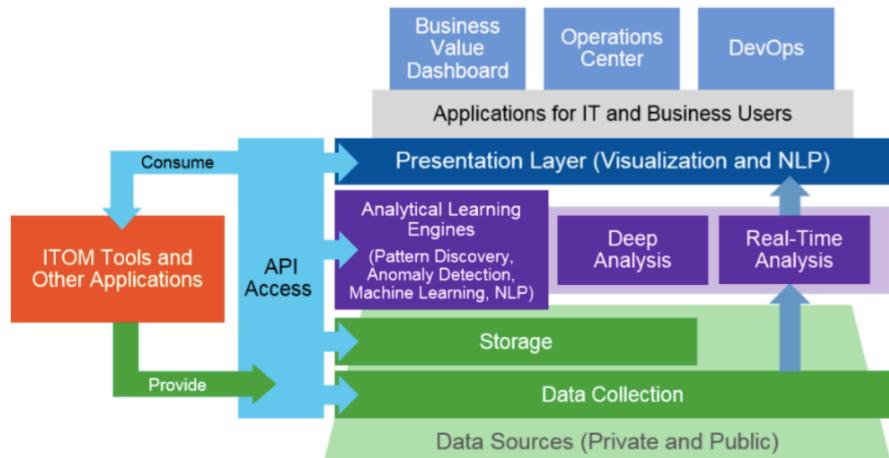
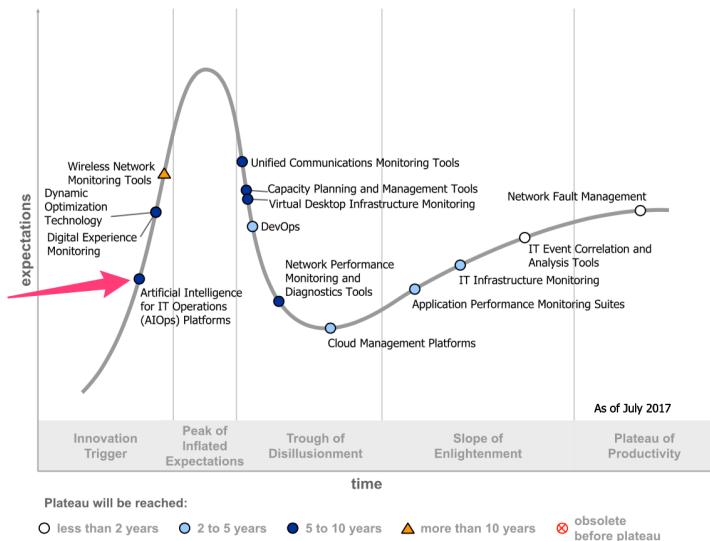
Question: Can we predict this problem 2 hours before it happens again?
Then just switch the traffic away from this switch and reboot it.



- Precision: 82.15%
- Recall: 74.74%
- FPR: 3.75×10^{-5}

AIOps is rising

- According to Gartner Report:
- AIOps global deployment ratio: 10% (2017) → 50% (2020)



Levels of Autonomous IT Operations

- **Cores Per Op (CPO):** The average number of x86bCPU cores managed by an Op (40hours/week)
- **Assumption:** Organization tries their best to achieve certain reliability.
- Try to decouple with the following factors:
 - Business sectors, scale, architecture, technology, part-time
- Count operators of server, storage, network, middleware, database, application
- Count the hours of operators for triggering scripts, monitoring the big screen, browsing the monitoring data, deal with alerts, troubleshooting, planning, idle time while on duty.
- Do not count the hours of operators for developing IT operations tools.

Level=[Log (CPO/100)]	Cores Per Op (CPO)	Typical Enterprises
Level 0	O(100)	Finance
Level 1	O(1K)	Medium Internet companies running on public clouds
Level 2	O(10K)	Large Internet companies
Level 3	O(100K)	
Level 4	O(1M)	
Level 5	O(10M)	

Example1: Internet Company A

- All x86 servers: 500K with 12 cores each, 500K with 24 cores each。 In total there are 13M cores.
- Labor: $(200*0.5+200*0.8)*60/40=390$ Op
 - 200 operators for server, storage, database, and network
 - 60 hours/week; 50% of working time is for manual operations, and 50% of working time is for tool development.
 - 200 operators for applications and middleware
 - 60 hours/week; 80% of working time is for manual operations
- CPO=13M cores/390 Op=33K cores/Op
- **Level = $\lfloor \log(CPO/100) \rfloor = 2$**

Example2: Internet Company B

- All x86 servers: 500K with 12 cores each, 500K with 24 cores each。 In total there are 13M cores.
- Labor: $(200*0.5+200*0.8)=130$ Op
 - 100 operators for server, storage, database, and network
 - 40 hours/week; 50% of working time is for manual operations, and 50% of working time is for tool development.
 - 100 operators for applications and middleware
 - 40 hours/week; 80% of working time is for manual operations
- CPO=13M cores/130 Op=100K cores/Op
- **Level = $\lfloor \log(CPO/100) \rfloor = 3$**

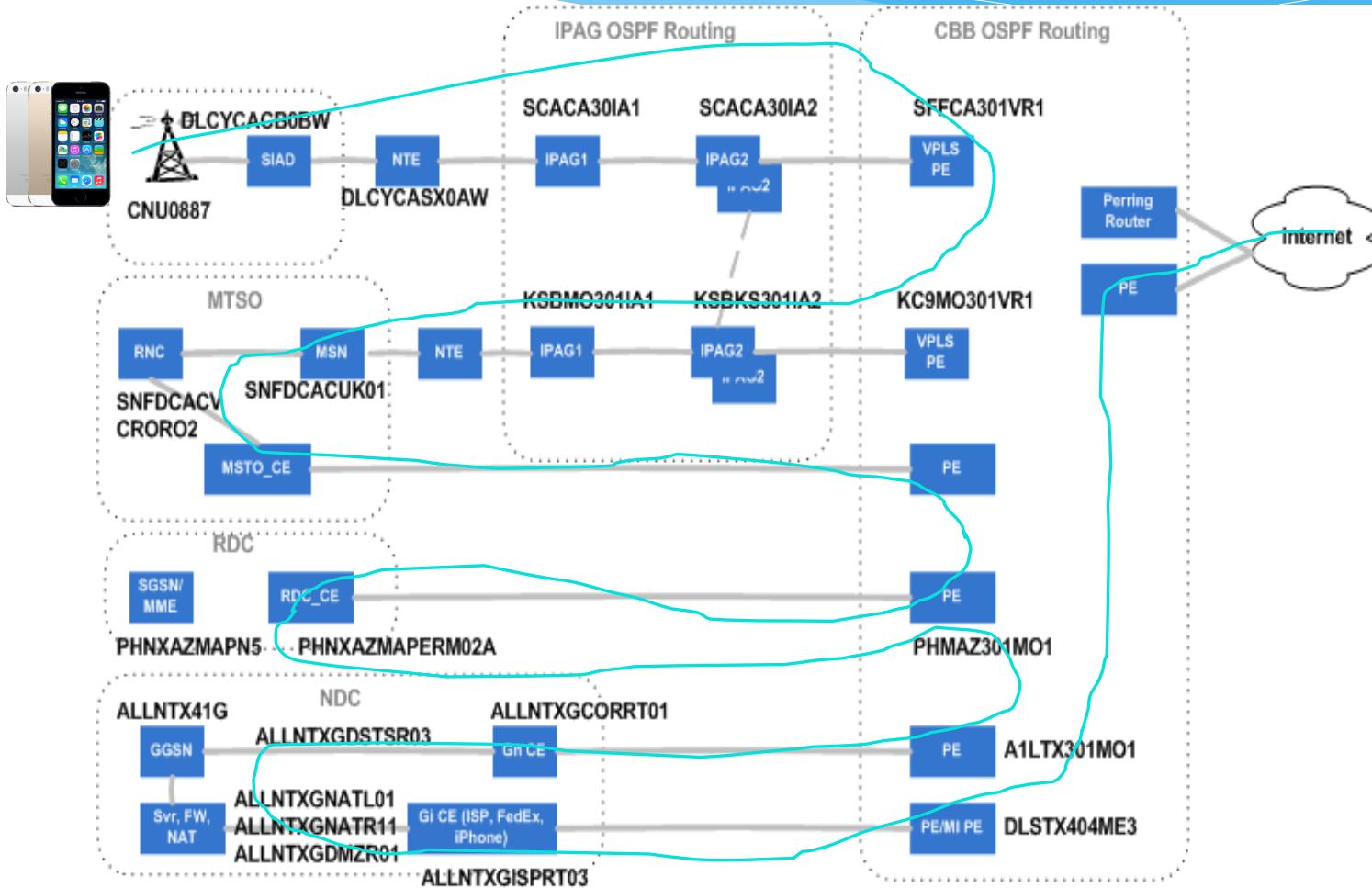
Example 3: Bank C

- 10K x86 servers with 12 cores each. 500 small computers, each equivalent to 100 cores. 5 Mainframe computers, each equivalent to 2K cores. 180K cores in total
- Labor $(100*0.5+100*0.8+200)*60/40=495$ Op
 - 100 operators for server, storage, database, and network
 - 60 hours/week; 50% of working time is for manual operations, and 50% of working time is for tool development.
 - 100 operators for applications and middleware
 - 60 hours/week; 80% of working time is for manual operations
 - 200 Outsourced Operators
 - 60 hours/week; full time on manual operations
- $CPO=180K\ Cores/495\ Op=363/Op$
- **Level = $\lfloor \log(CPO/100) \rfloor = 0$**
- plan to have 100K x86 servers, and the number of cores increases to 1.26M
 - Keep the CPO, and increase the #Ops to to $1.26M/363=3360$, or
 - Keep the #Op=495, but increase the **CPO=1.26M/495=3545 cores/Op; Level=1**

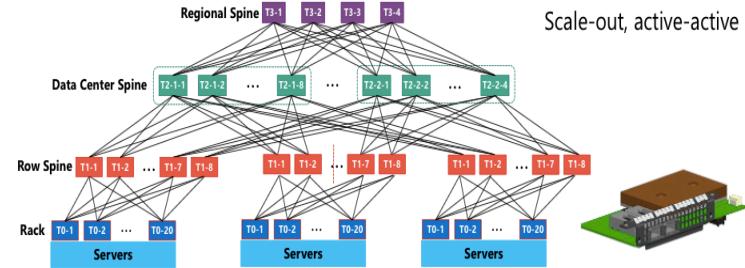
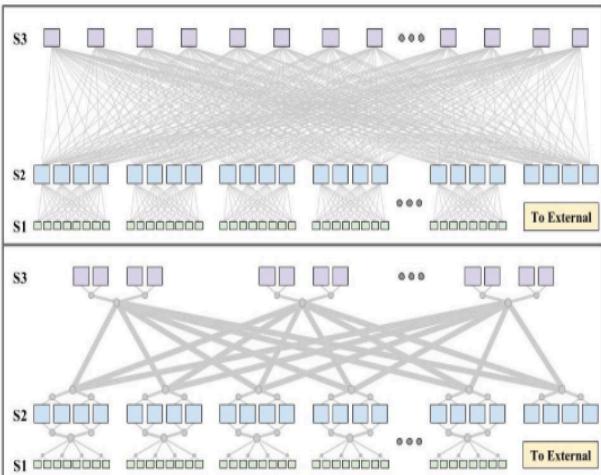
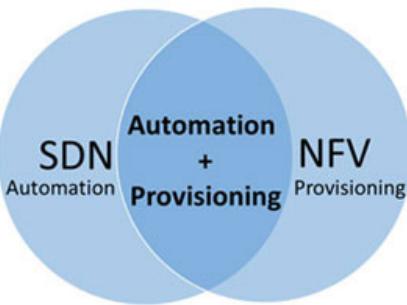
Overview

- * What is AIOps?
- * What's the value of AIOps?
- * **Is AIOps necessary?**
- * Is AIOps feasible?
- * Case studies
- * AIOps Challenge
- * Summary

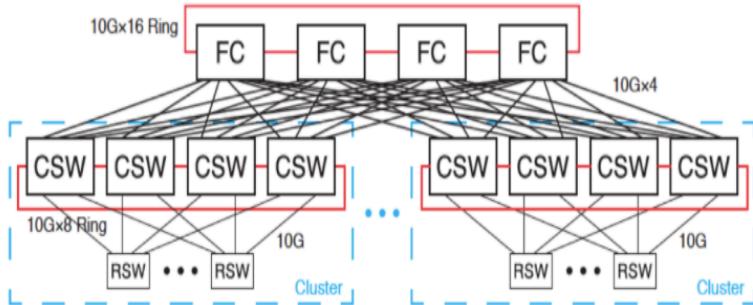
Complex Access Networks



Complex and Evolving Cloud

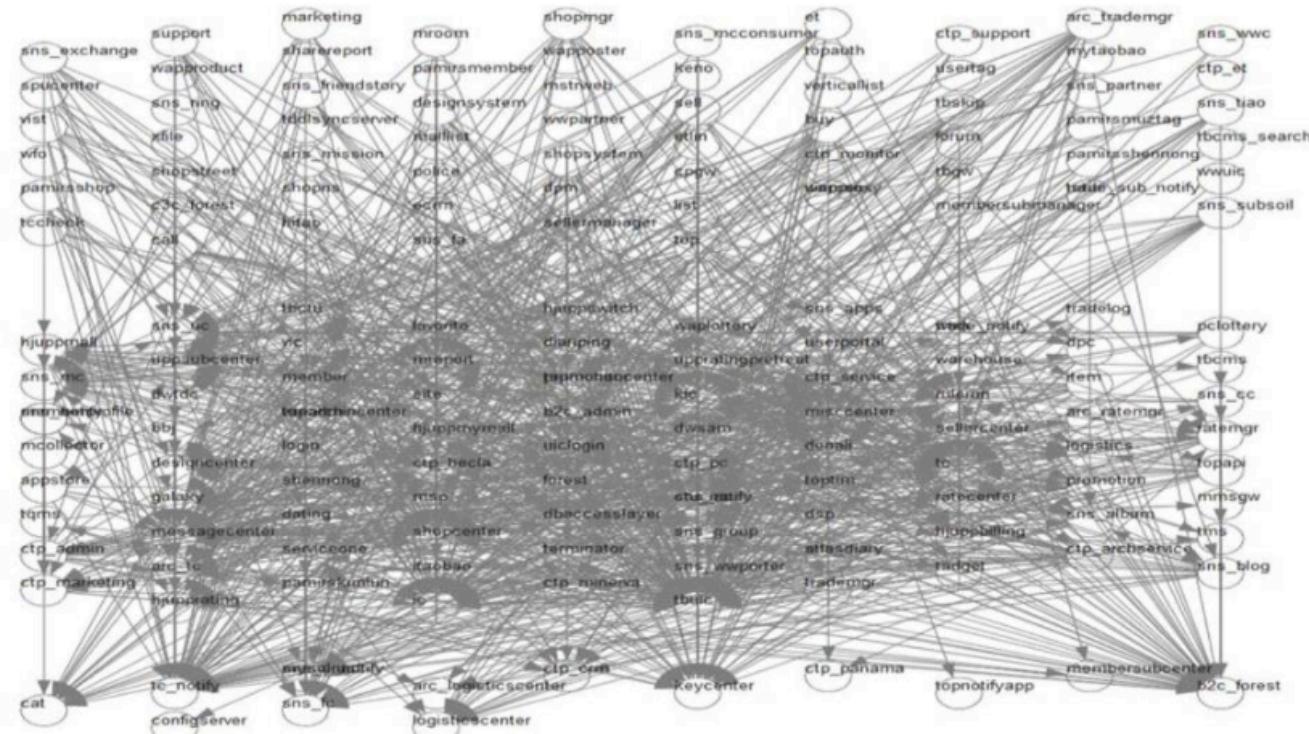


Outcome of >10 years of history, with major revisions every six months



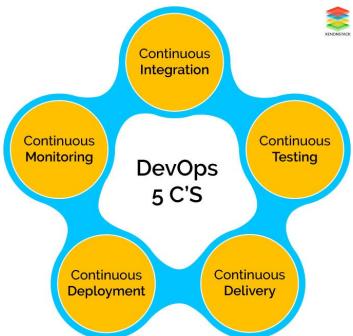
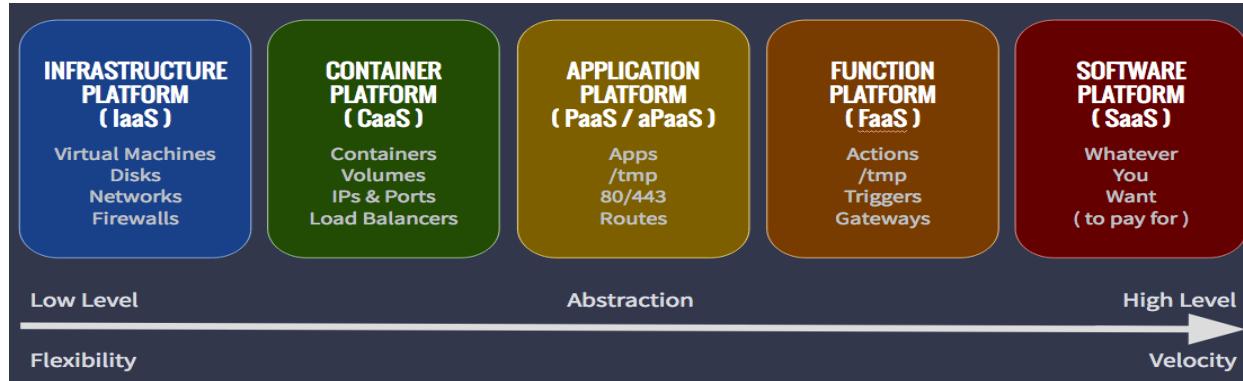
Complex and Evolving Software Module Dependencies

Taobao's application dependency in 2012

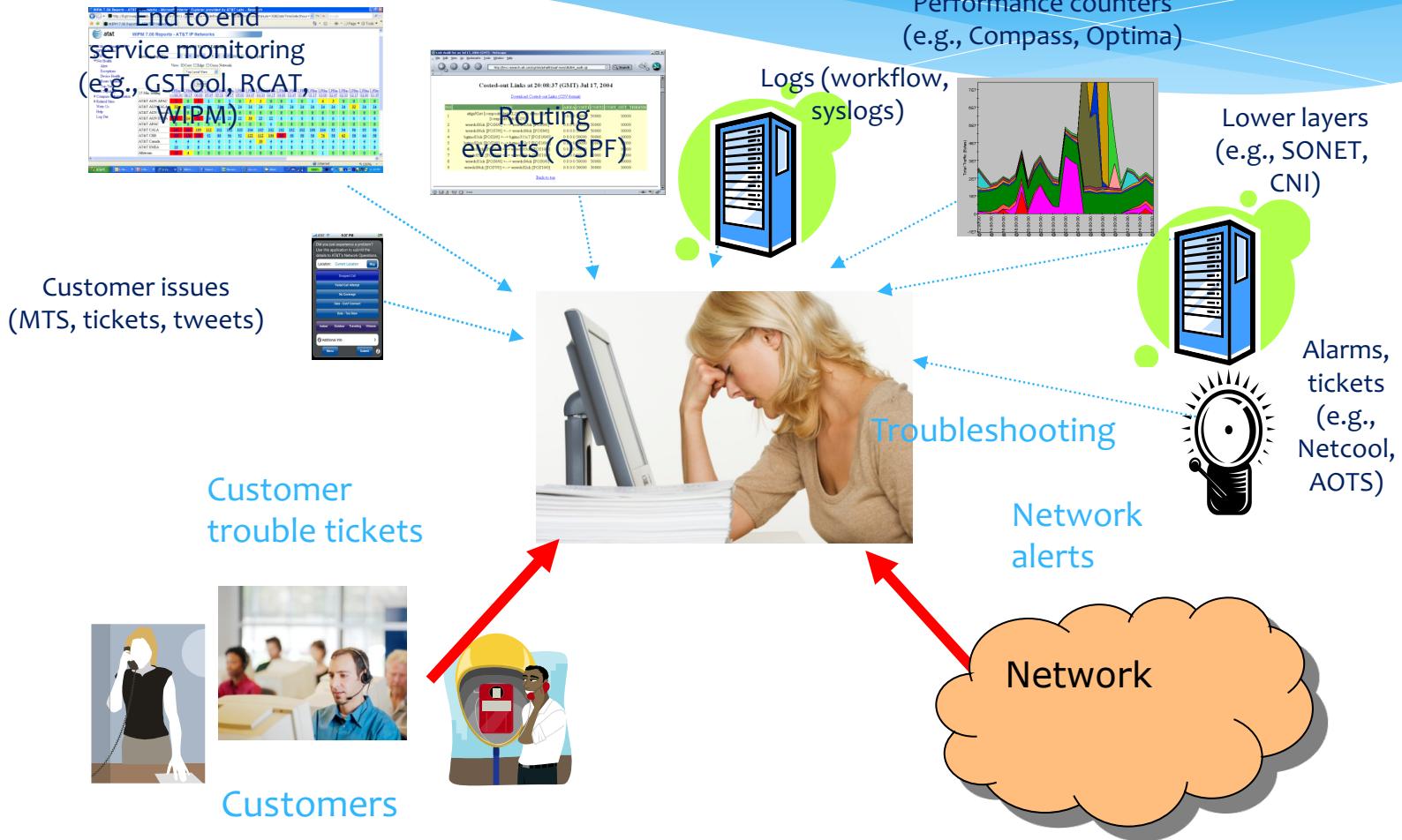


2012 淘宝核心链路应用拓扑图

Evolving Techniques Enable Frequent Software Changes



There are a sheer volume of device-generated log data during daily operations



We have no choice but relying on AI to take advantage of the Big Data from Ops

- * Volume
- * Velocity
- * Variety
- * Value

Overview

- * What is AIOps?
- * What's the value of AIOps?
- * Is AIOps necessary?
- * **Is AIOps feasible?**
- * Case studies
- * AIOps Challenge
- * Summary

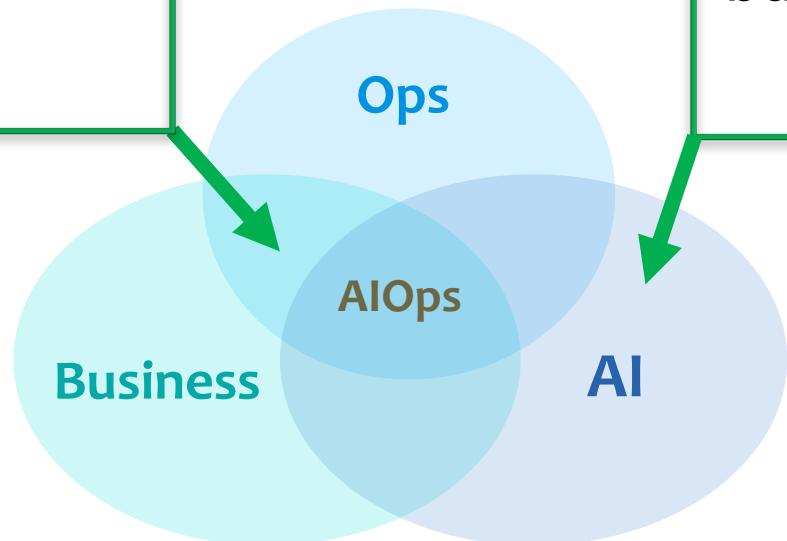
AIOps has *the necessities required for successful ML applications*

- * Machine learning tools (algorithms and systems)
- * Applications that show the value
- * Large amount of data
- * Labels and the experts who can label

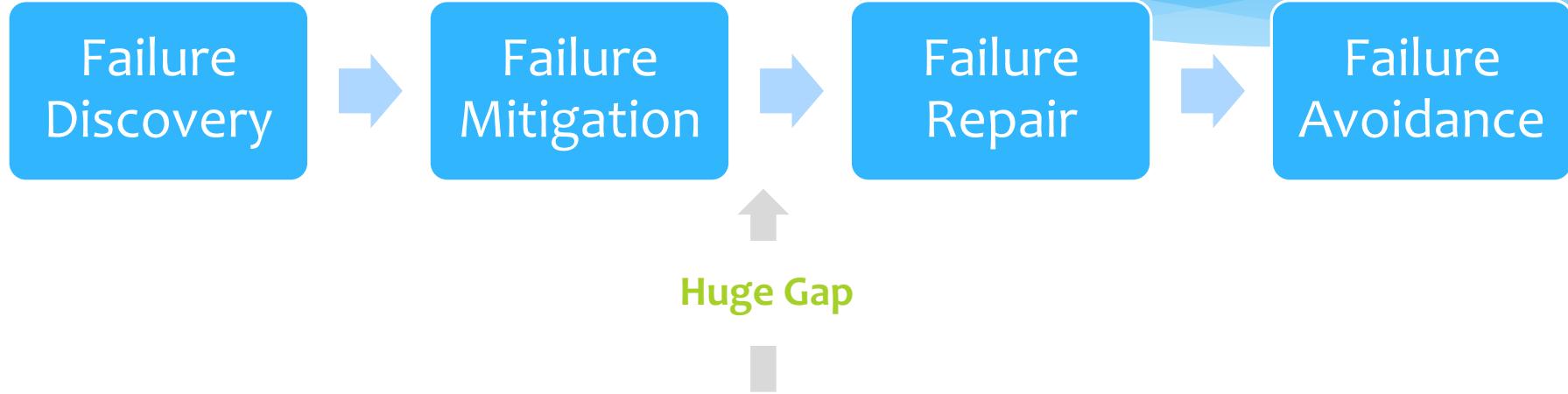
AIOps is still challenging because its the interdisciplinary nature

Ops people familiar with Ops and Business, but not AI

Algorithm people familiar with general AI, but not Ops and AIOps



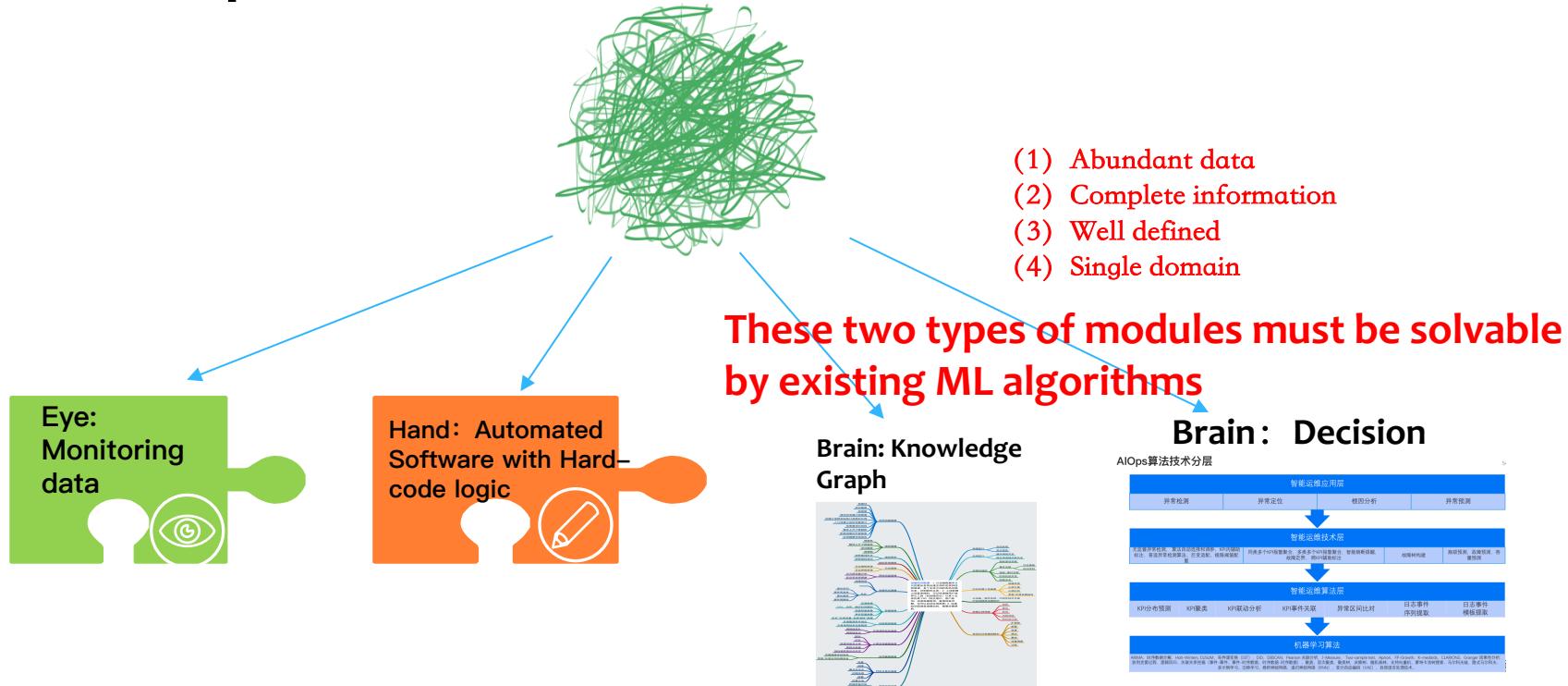
Pitfalls: use ML algorithms as Blackbox to tackle Ops challenges



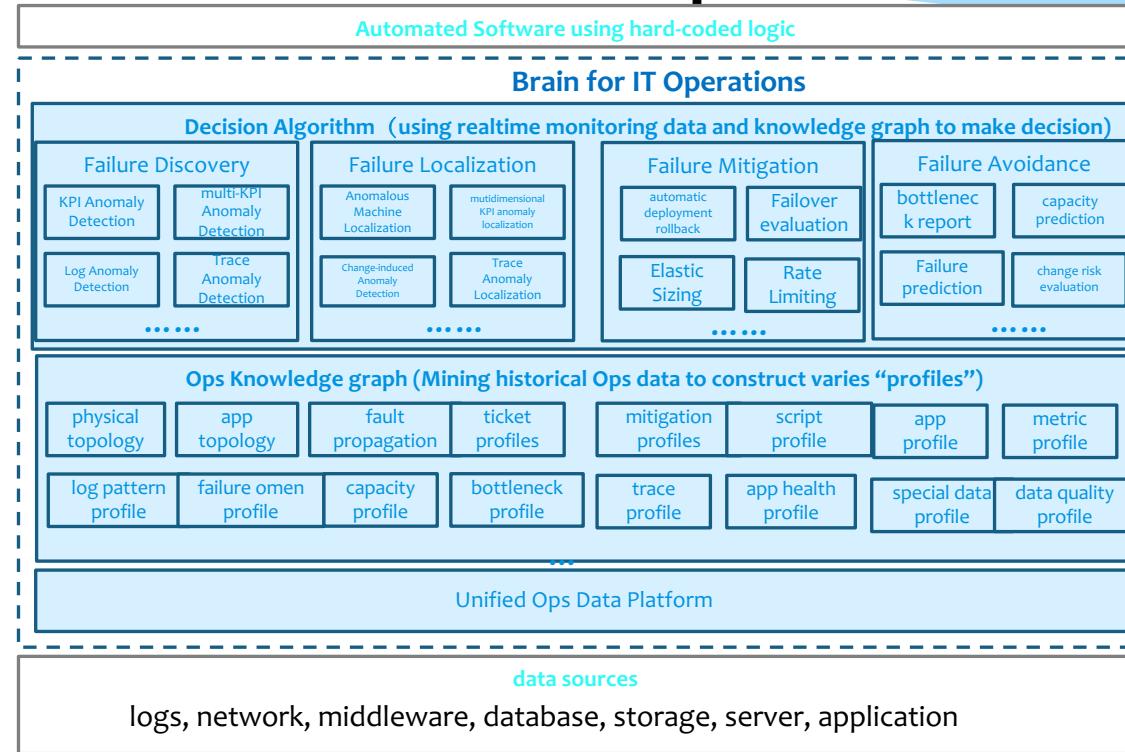
General Machine Learning Algorithms

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

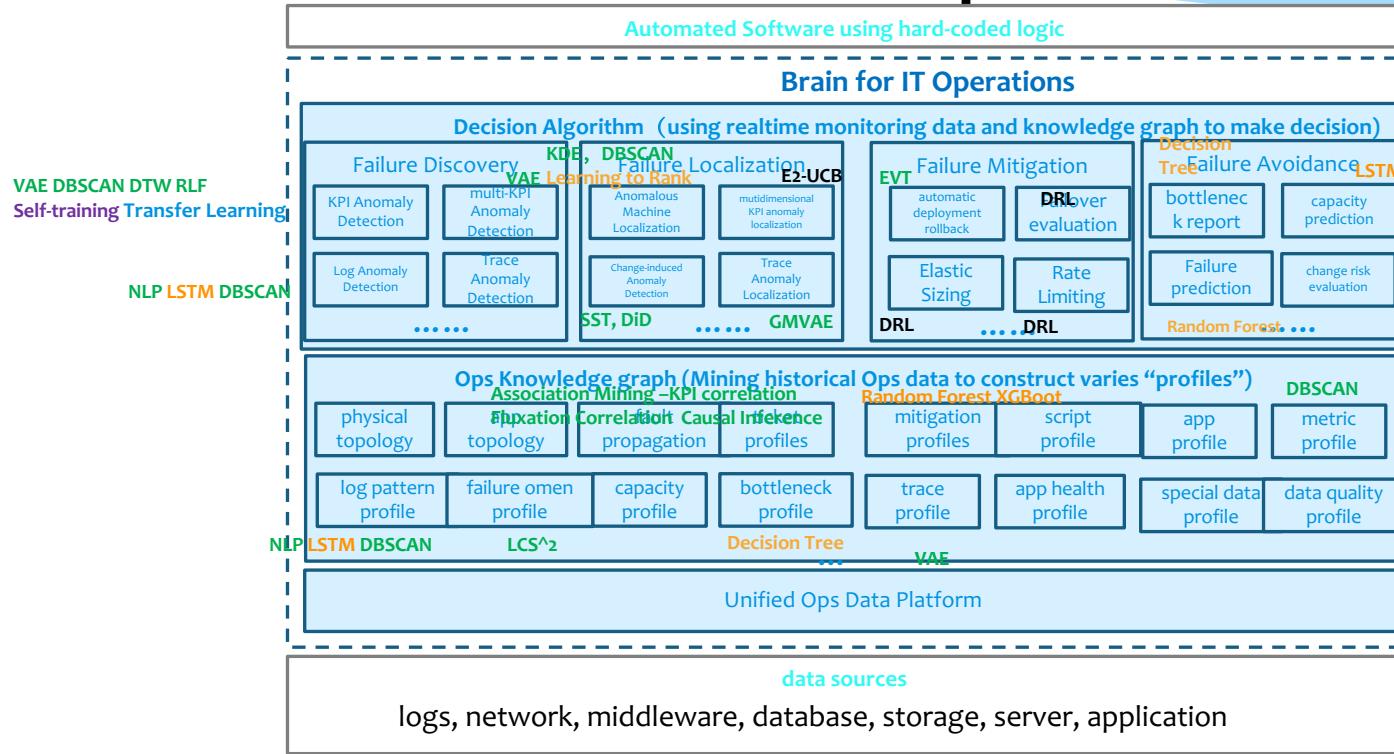
AIOps Architecture : Divide the complex task and Conquer



Brain for IT Operations

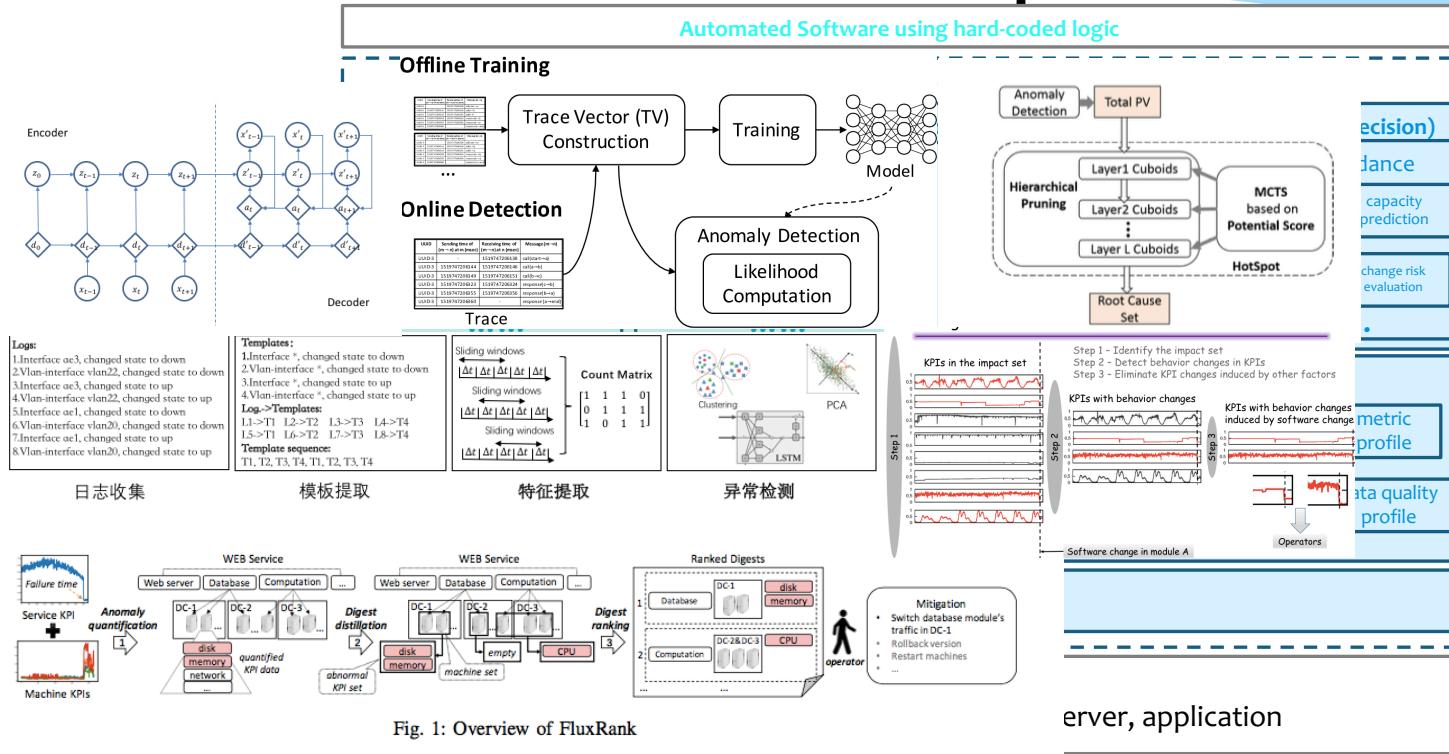


Brain for IT Operations



Unsupervised Reinforcement Learning Supervised but with labels Semi-supervised Learning Transfer Learning

Brain for IT Operations



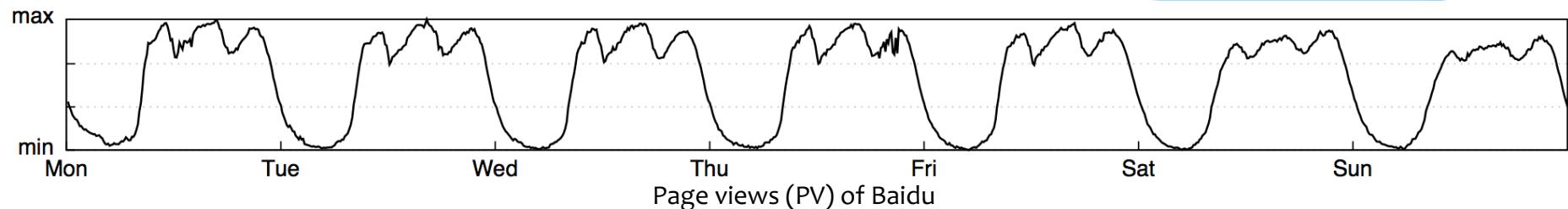
Unsupervised Reinforcement Learning Supervised but with labels Semi-supervised Learning Transfer Learning

Overview

- * What is AIOps?
- * What's the value of AIOps?
- * Is AIOps necessary?
- * Is AIOps feasible?
- * **Case studies**
- * AIOps Challenge
- * Summary

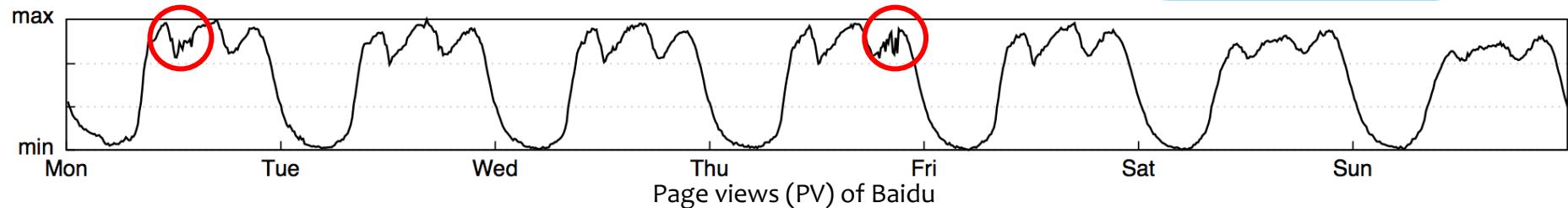
Case 1: KPI Anomaly Detection (Dapeng Liu et al., IMC 2015)

KPIs and Anomaly Detection



KPIs (Key Performance Indicators): A set of performance measures that evaluate the service quality

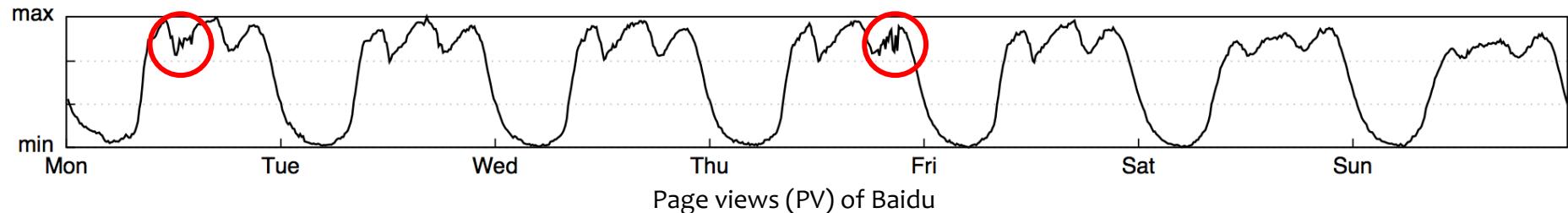
KPIs and Anomaly Detection



KPIs (Key Performance Indicators): A set of performance measures that evaluate the service quality

KPI anomalous (unexpected) behaviors → Potential failures, bugs, attacks...

KPIs and Anomaly Detection



KPIs (Key Performance Indicators): A set of performance measures that evaluate the service quality

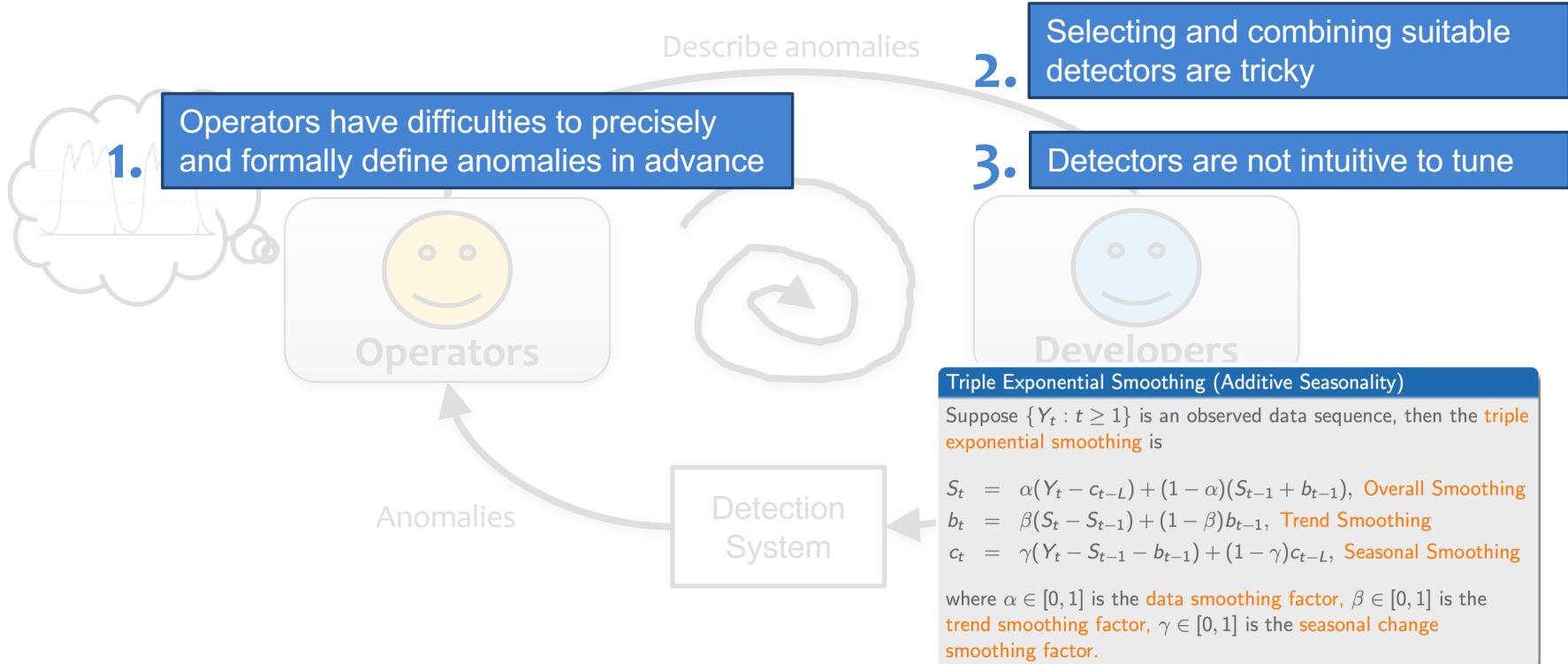
KPI anomalous (unexpected) behaviors → Potential failures, bugs, attacks...

Anomaly detection matters: Find anomalous behaviors of the KPI curve

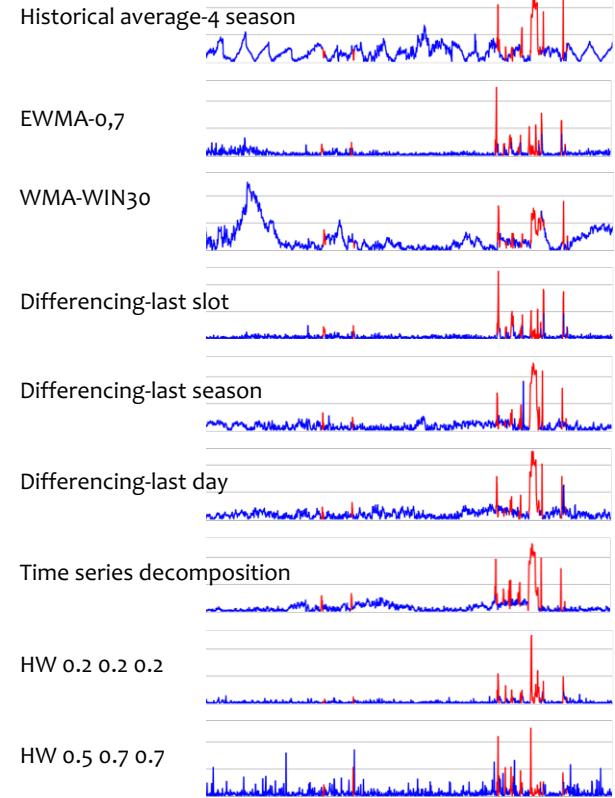
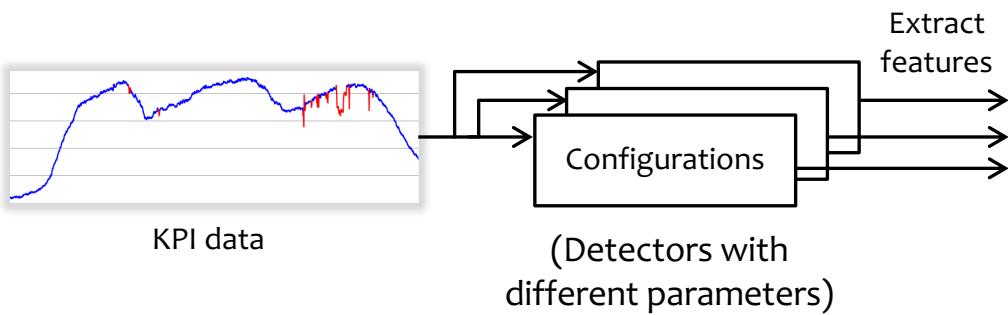
- Diagnose and fix it
- Avoid further influences and revenue losses

How to Build an Anomaly Detection System

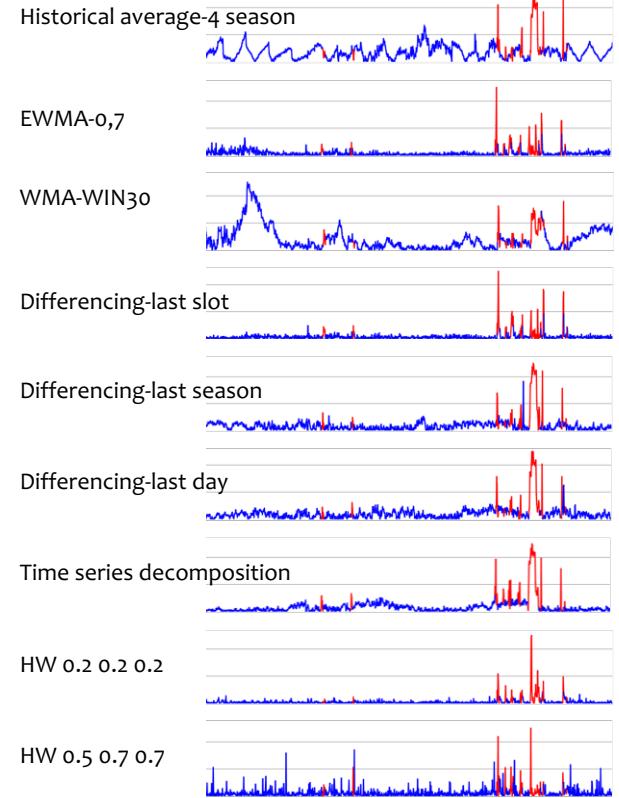
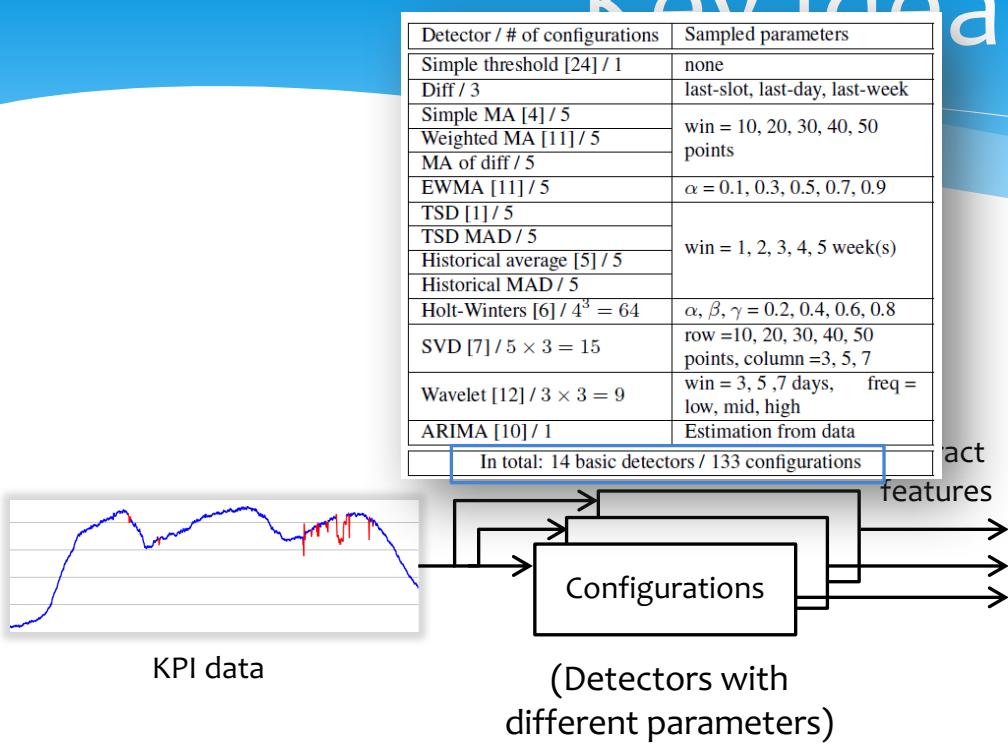
Challenges



Key Ideas

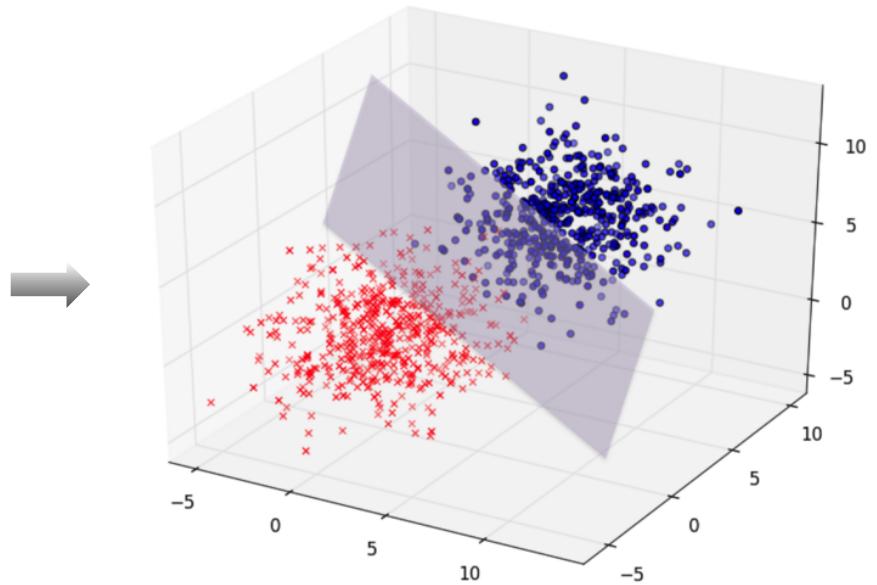
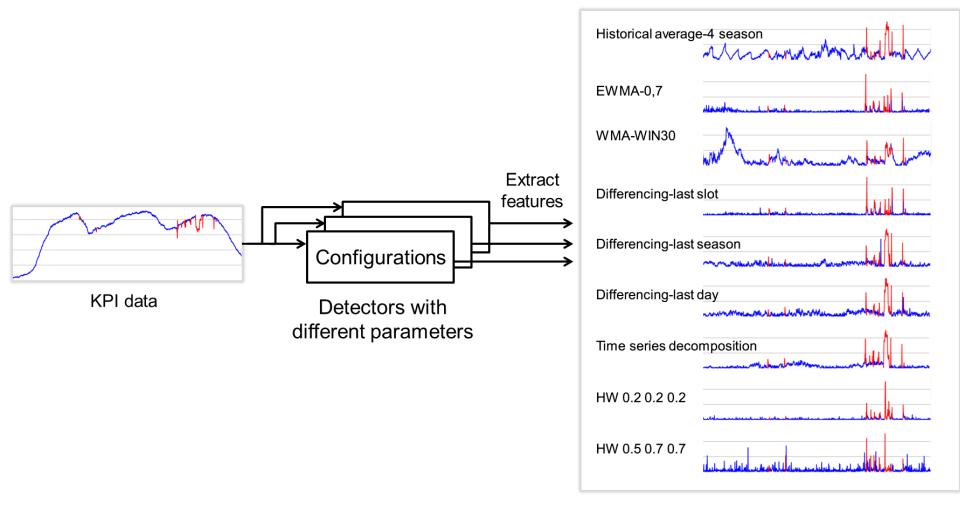


Key Ideas

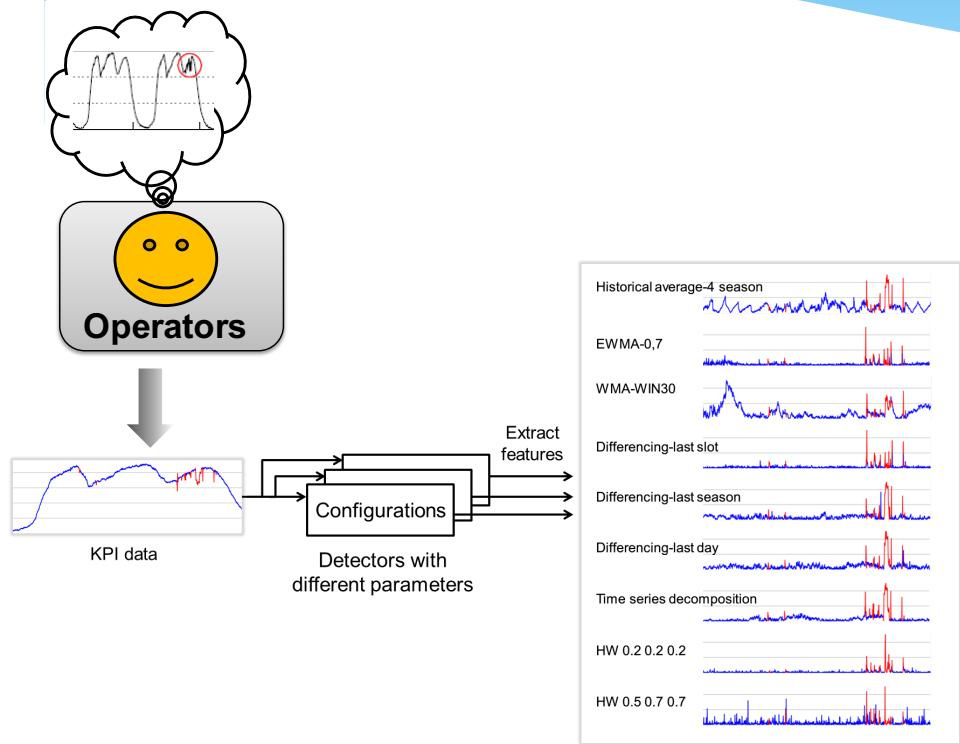


Key Ideas

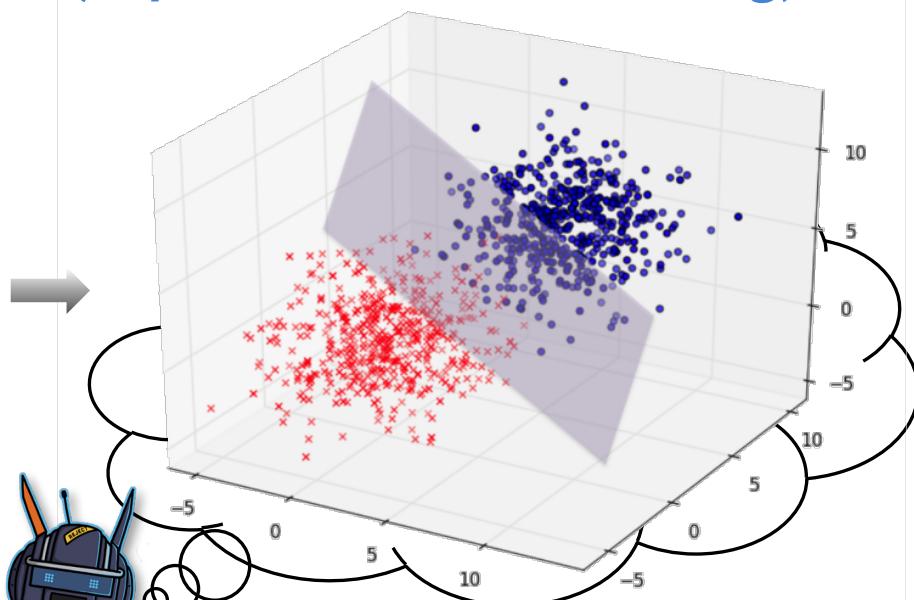
Classification in the feature space
(Supervised machine learning)



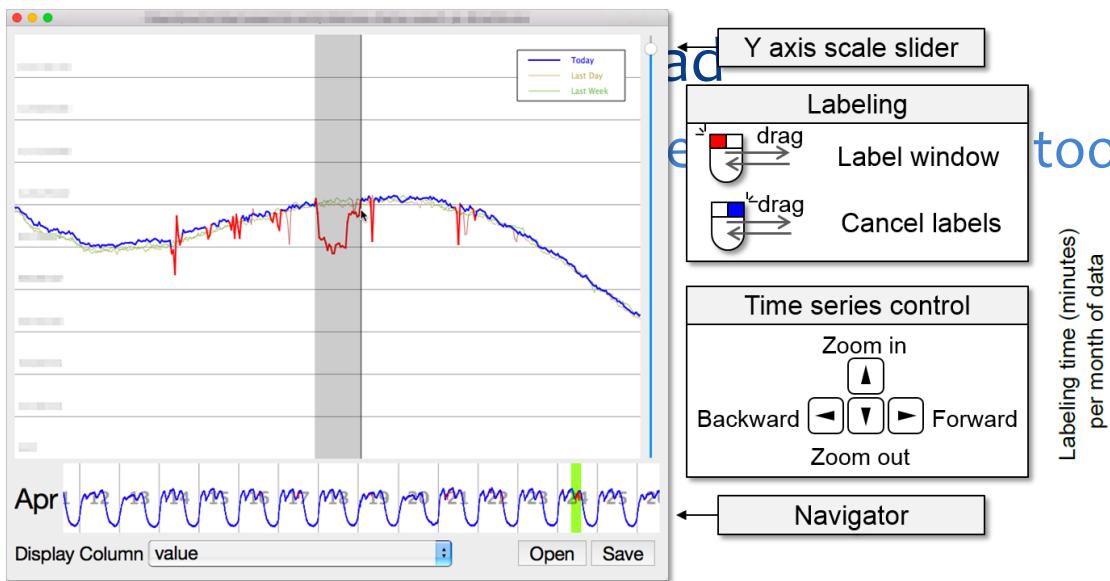
Key Ideas



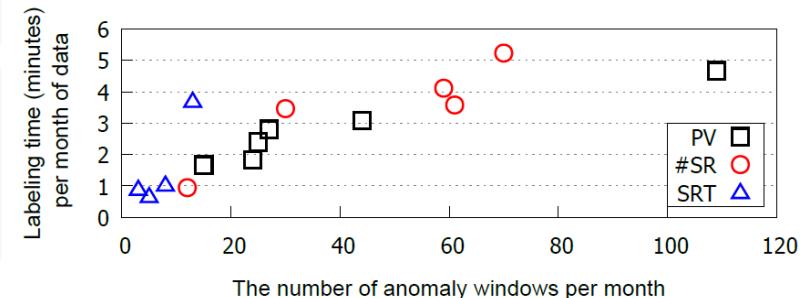
Classification in the feature space
(Supervised machine learning)



Address Challenges of Designing Apprentice



tool

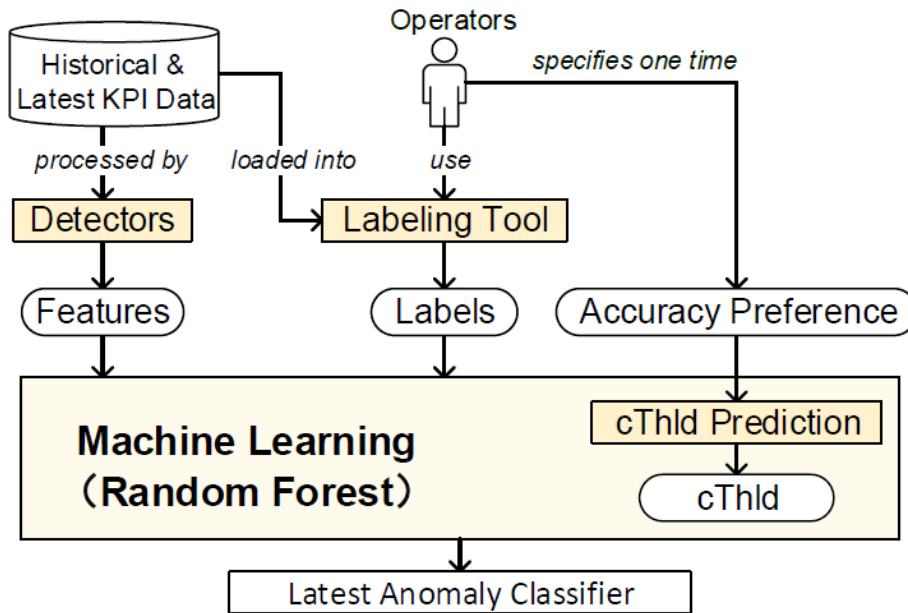


Address Challenges of Designing Apprentice

- * Labeling overhead
 - * Solution: an effective labeling tool
- * Incomplete anomaly types in the historical data
 - * Solution: incremental re-training with new data
- * Class imbalance problem
 - * Solution: adjusting classification threshold ($cThld$) based on the preference
- * Irrelevant and redundant features
 - * Solution: random forests

Design Overview

Training a classifier

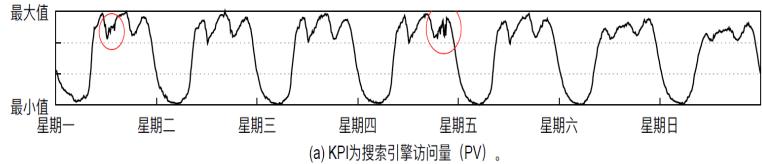


See the paper
for full details

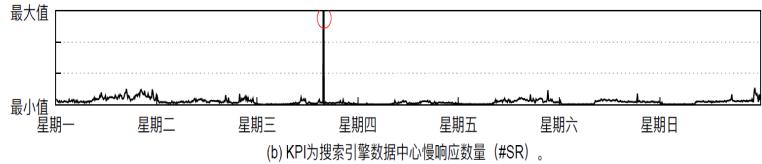
Detecting anomalies



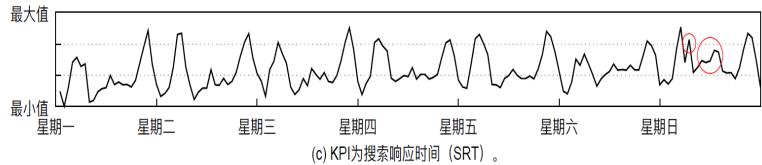
Evaluation



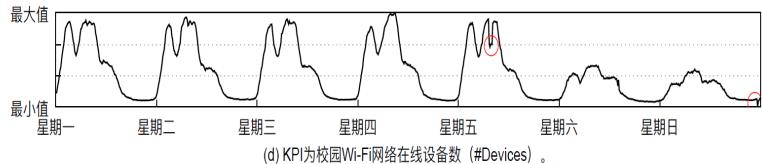
Search PV (25 weeks)



#slow queries (19 weeks)



Search Response Time (16 weeks)



#online devices (15 weeks)

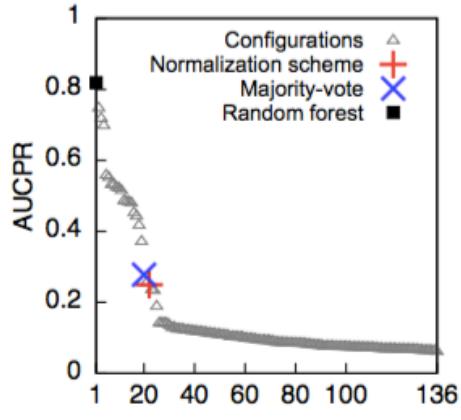
Baidu

Tsinghua
Enterprise WiFi

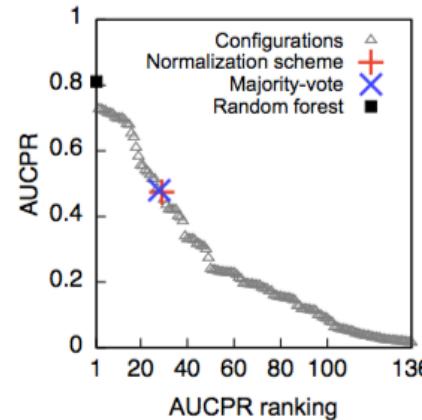
Evaluation

- * Compared with all existing detectors (Four KPIs)

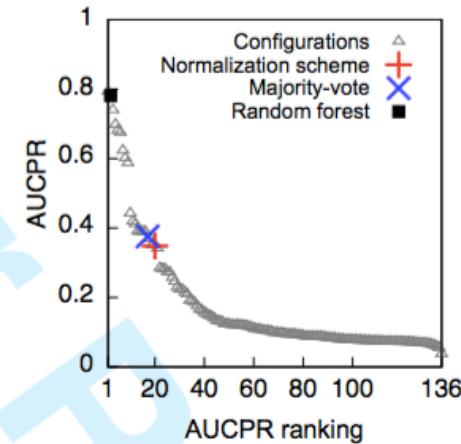
first



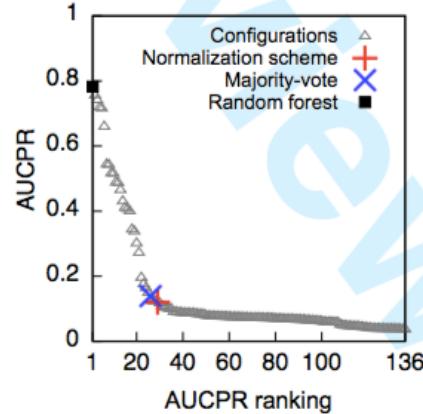
first



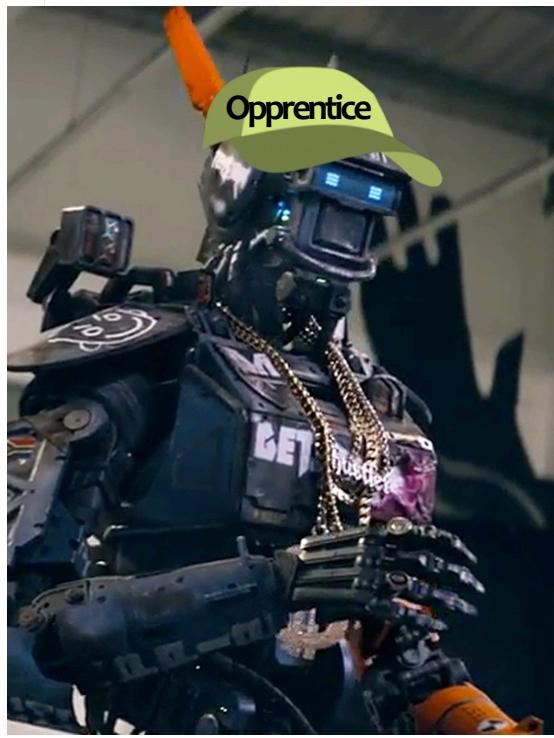
second



first



Case 1 summary



- * Oppentice is an **automatic** and **accurate** machine learning framework for KPI anomaly detection
 - Defining anomalies
 - Selecting detectors
 - Tuning detectors
- * Oppentice **bridges the gap** in applying complex detectors in practice
- * The idea of Oppentice
 - i.e., **using machine learning to model the domain knowledge** could be a very promising way to automate other service managements

Case 2 : Bottleneck Identification for Search Response Time (Dapeng Liu et al., INFOCOM 2016)

Web Search Engines

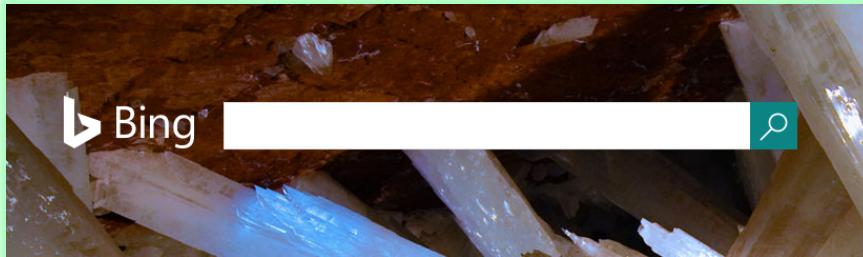


百度一下



Google Search

I'm Feeling Lucky



Search Response Time (SRT)



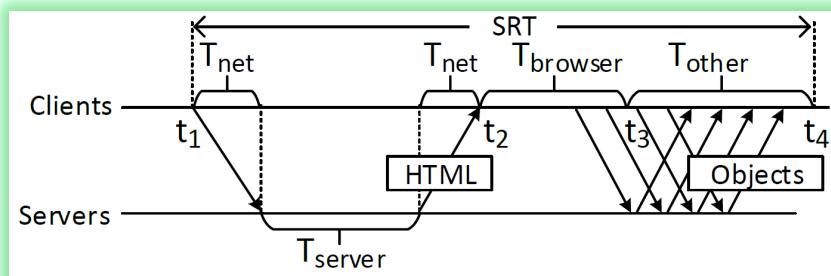
t_1 A search query
is submitted



t_4 The result page
Is rendered

$$SRT = t_4 - t_1$$

57



Search Response Time Matters



+500ms revenue ↓1.2%
[Eric Schurman, Bing]



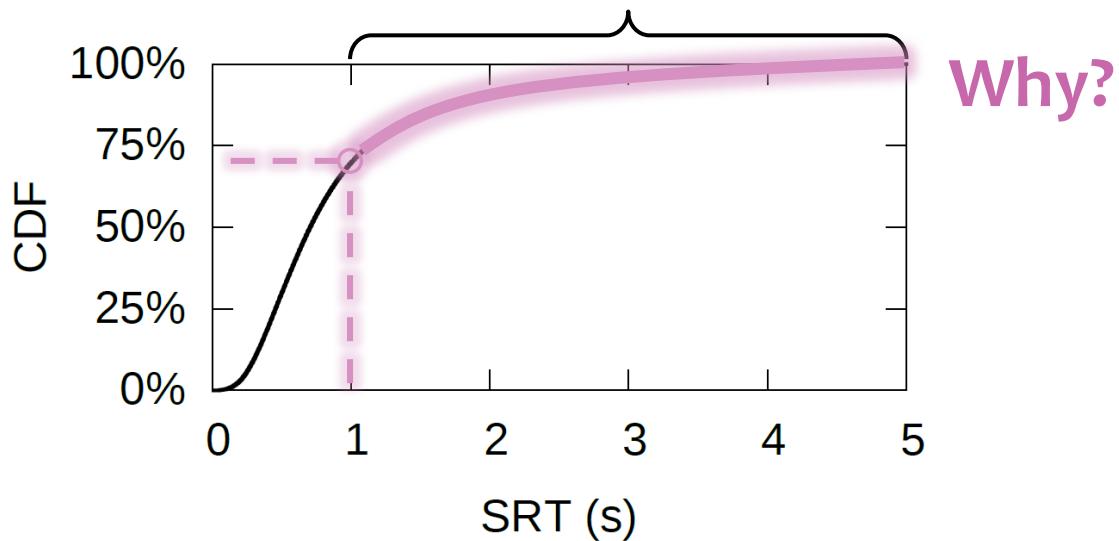
+100ms~400ms queries ↓0.2%~0.6%
[Jake Brutlag, Google]



Given two content-wise identical search result pages,
users are more likely to perform clicks on the fast page
[SIGIR 2014]

Search Response Time in the Wild

User's flow of thought is interrupted
if pages take **longer than 1s** to load



Monitoring SRT: Search Logs

Measurable attributes that can potentially impact SRT

SRT	User's ISP	Browser engine	# of Images	Ads	Server Load	...
800ms (Low SRT)	China Unicom	WebKit	10	Yes	1000 queries/s	...
1200ms (High SRT)	China Telecom	Trident 5.0	5	No	500 queries/s	...
.....						

Goal of FOUCS

Measurable attributes that can potentially impact SRT

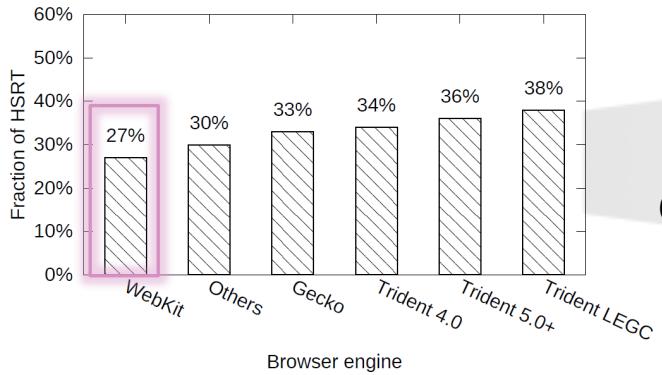
SRT	Client ISP	Browser engine	# of Images	Ads	Server Load	...
800ms (Low SRT)	China Unicom	WebKit	10	Yes	1000 queries/s	...
1200ms (High SRT)	China Telecom	Trident 5.0	5	No	500 queries/s	...
.....						

We propose **FOCUS**, a search log analysis system to answer the following questions:

- Under what conditions **HSRT** (**High SRT**) is more likely to happen?
- Which HSRT conditions are similar (HSRT condition types)?
- How does each attribute affect SRT in HSRT condition types?

Challenges

Limited visibility of naïve single-dimension analysis



What we can see

WebKit is a good condition, where HSRT is **only 27%**
(e.g. used by Chrome and Safari)

What we **cannot** see

HSRT is **more than 38%**
when “**WebKit + #Images >30**”

Challenges

Limited visibility of naïve single-dimension analysis

Interdependencies between attributes

Overlapped HSRT conditions

Key Idea of FOCUS

Limited visibility of naïve single-dimension analysis

Multi-dimension analysis

Interdependencies between attributes

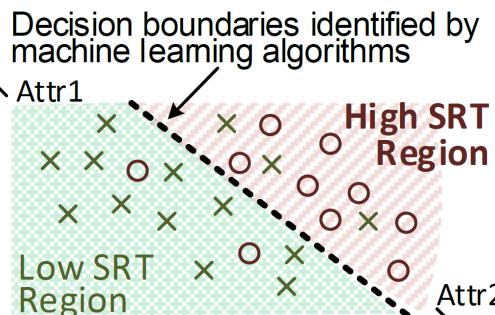
Work with interdependencies

Overlapped HSRT conditions

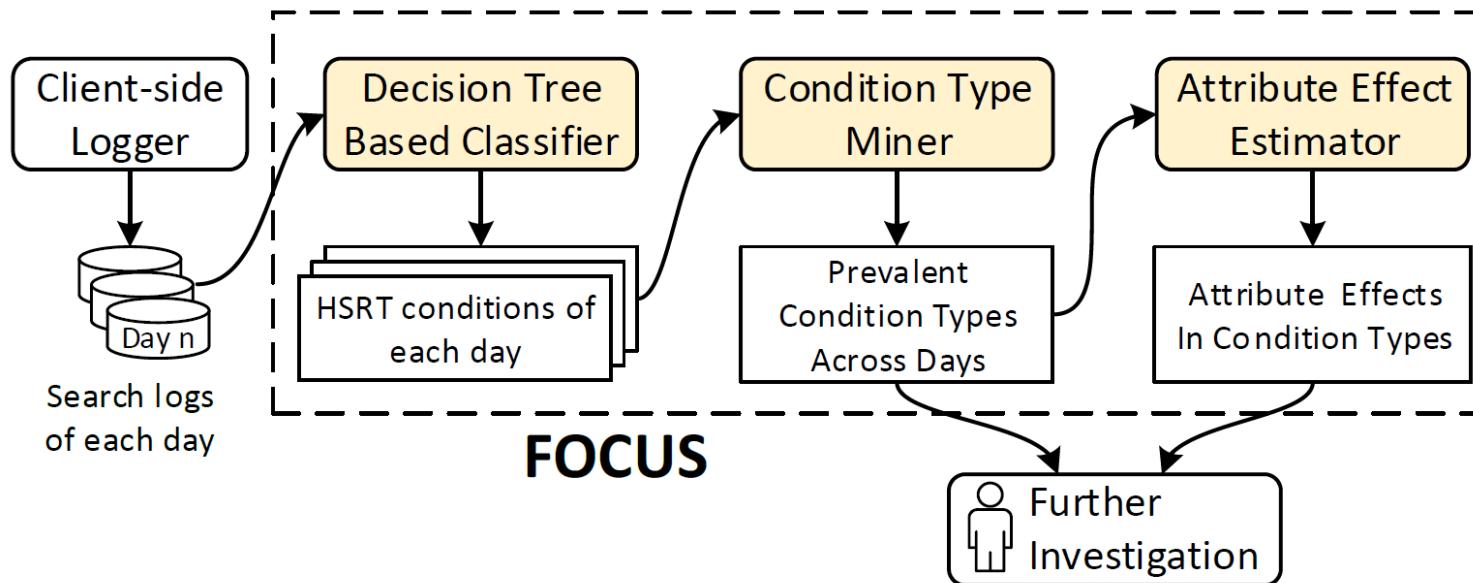
Classification is non-overlap

- Model it as a classification problem
- Solve it using decision trees

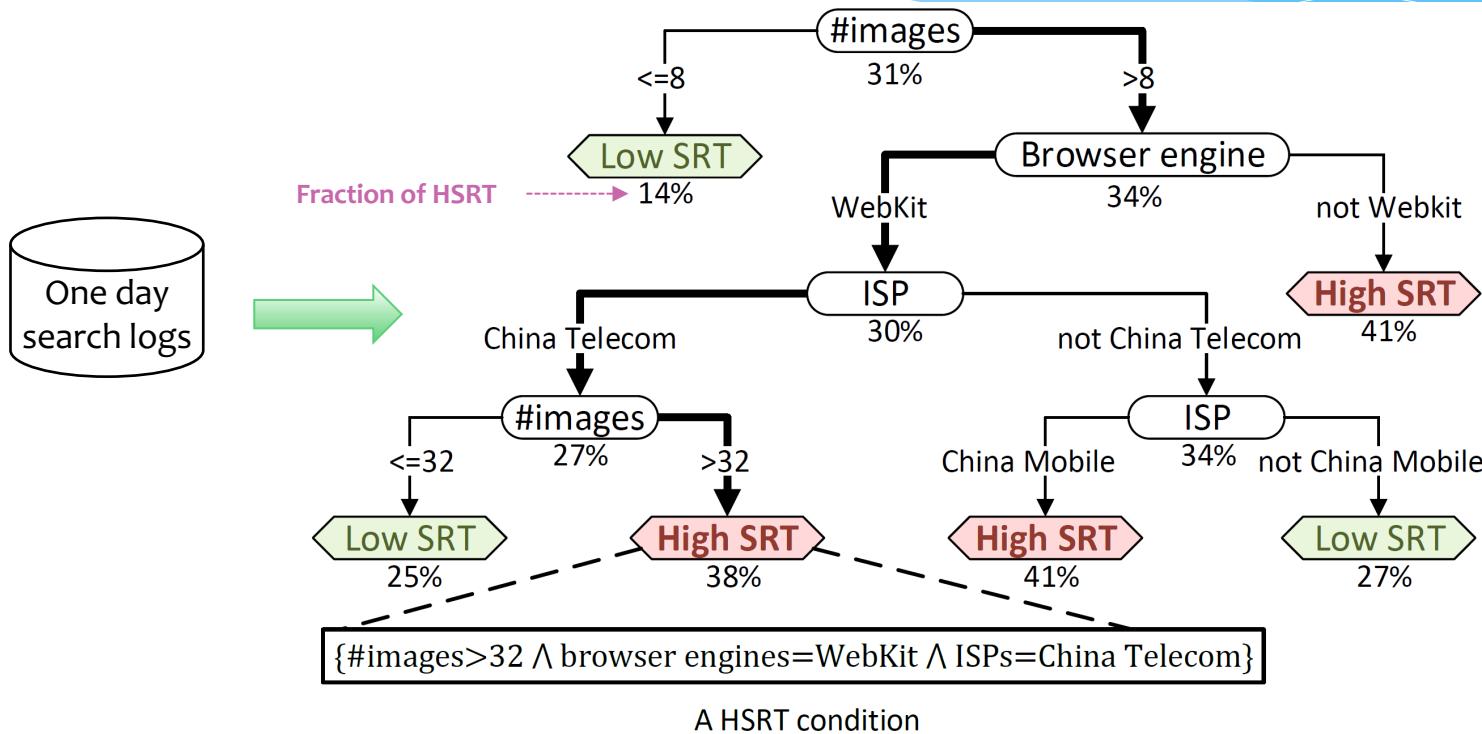
Attr1	Attr2	Label
...	...	High SRT O
...	...	Low SRT X
...	...	Low SRT X
...



FOCUS Overview

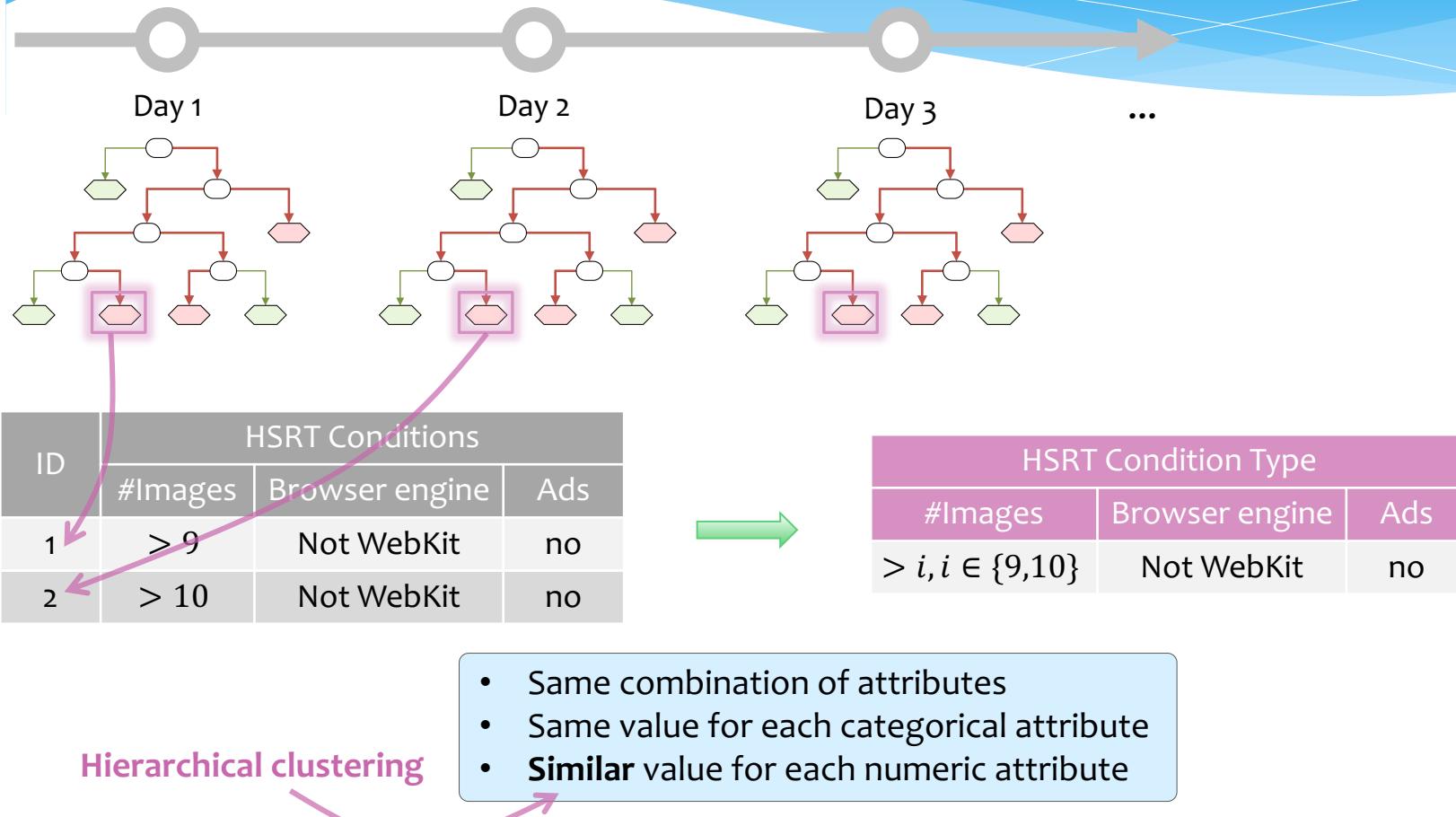


Identify HSRT Conditions Based on a Decision Tree



To build a reasonable tree, we **tailor** the mechanisms of decision trees
(Details are in the paper)

Find Similar HSRT Conditions (HSRT Condition Types)



Estimate the Impact of Each Attribute

Inspired by controlled experiment

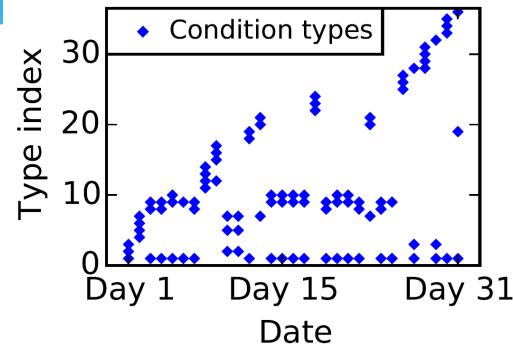
- **Control group:** the original HSRT condition types
- **Experimental group:** changing one attribute at a time



Compare performance
in historical logs

ID	HSRT Condition Type		
	#Images	Browser engine	Ads
C	$> i, i \in \{9,10\}$	Not WebKit	no
C_1	$\leq i, i \in \{9,10\}$	Not WebKit	no
C_2	$> i, i \in \{9,10\}$	WebKit	no
C_3	$> i, i \in \{9,10\}$	Not WebKit	yes

Results of FOCUS: Prevalent HSRT Condition Types



- * Find 36 HSRT condition types in one month of search logs
- * Four of them (11%) appear in more than five days

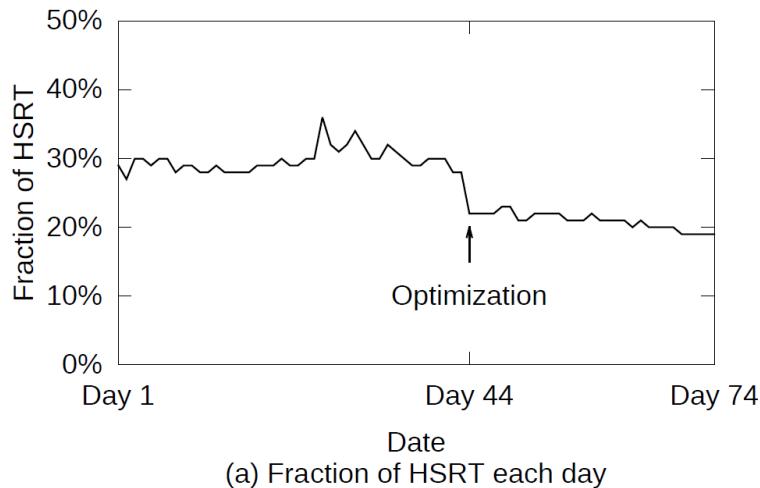
Images are the main bottleneck

(Attributes in bold have a bad effect on SRT)

Condition type ID	Prevalent condition type	Prevalence (days)
1	#images > i, $i \in \{5, 6, 7, 8, 9\}$ \wedge browser engine = not WebKit	21
2	#images > i, $i \in \{5, 6, 7, 8, 9\}$ \wedge ISP = not China Telecom \wedge browser engine = WebKit	15
3	#images > i, $i \in \{25, 26, 27\}$ \wedge ISP = China Telecom \wedge browser engine = WebKit	7
4	#images > i, $i \in \{5, 6, 8\}$ \wedge ISP = China Telecom \wedge browser engine = WebKit \wedge ads = yes	6

Real-world Optimization

- * 1st month results of FOCUS → images are the main bottleneck of SRT
- * Deploy “image base64 encoding” to improve the transmission time of images



HSRT percentage
is reduced by 30%

SRT 80th-tile is reduced
by 253 ms (20%)

The fraction of HSRT is reduced by 30%

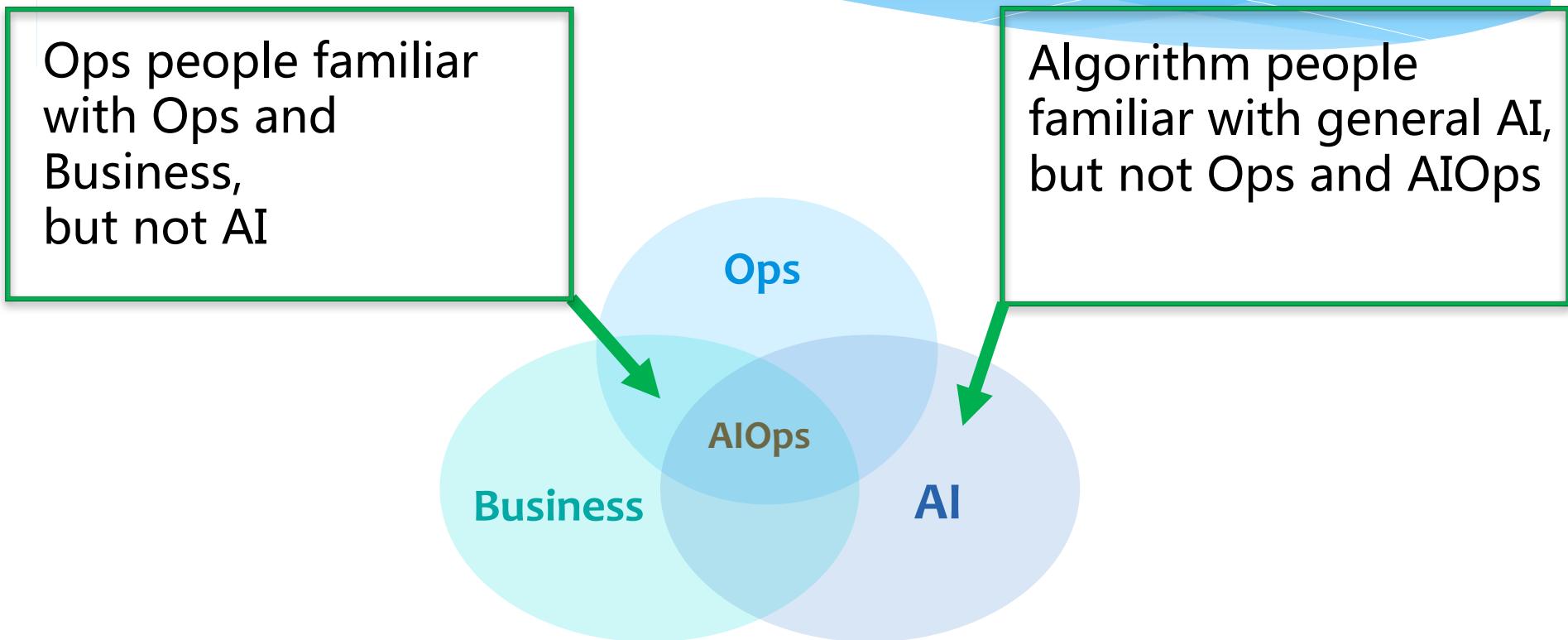
Case 2 Summary

- * FOCUS can
 - * Narrow down the debugging space of High SRT in search logs
 - * Analyze the effects of each attribute (potential improvements)
- * With the output of FOCUS
 - * We make several interesting observations
 - * Deploy a solution in practice and greatly improve SRT
- * FOCUS is a general method for analyzing multi-attribute logs
 - * Web applications other than search engines
 - * Performance of mobile apps
 - * ...

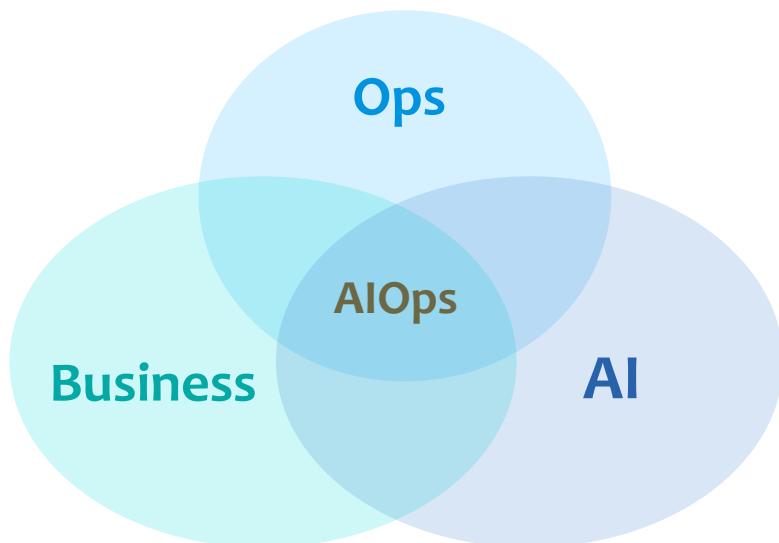
Overview

- * What is AIOps?
- * What's the value of AIOps?
- * Is AIOps necessary?
- * Is AIOps feasible?
- * Case studies
- * **AIOps Challenge**
- * Summary

How to speed up the AIOps progress?



How to speed up the AIOps progress?



AIOps Challenge : Community Efforts to Make AIOps Happen

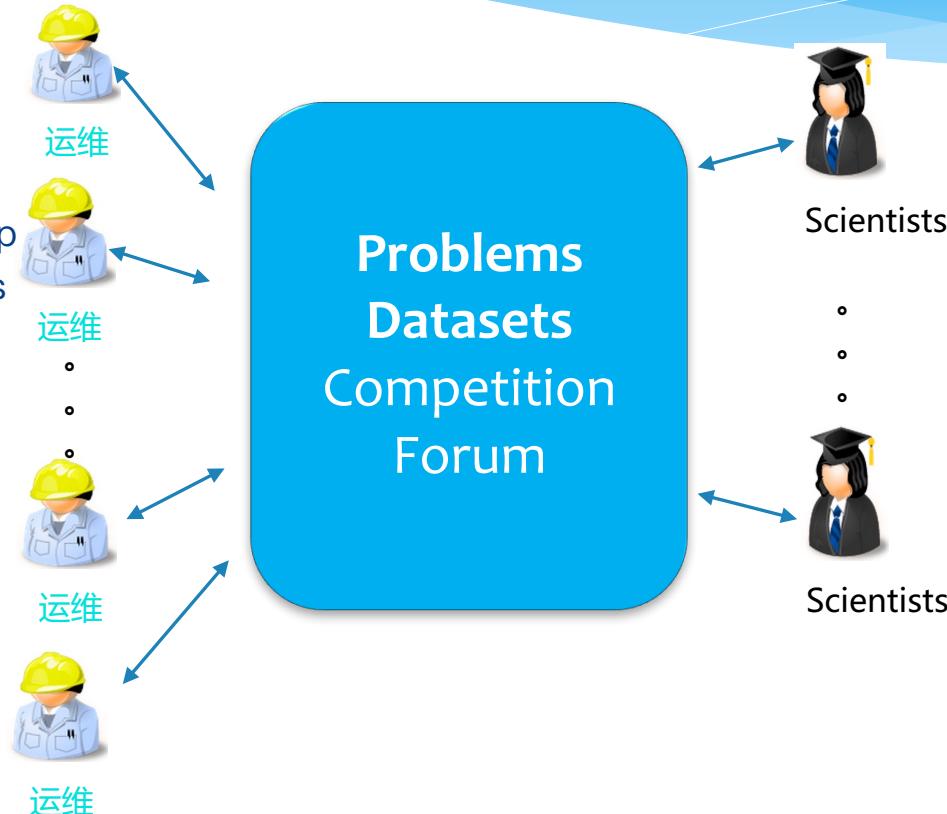
Ops

Give:

- Sanitized Data
- Competition sponsorship
- Bring up new challenges
- Forum discussion

Get:

- Algorithms
- Collaborators
- Recruiting
- Impact



Give :

- Competition
- Forum discussion
- Help define new problems

Get :

- Real Ops problem
- Real data
- Collaborators
- Impact

AIOps Challenge: <http://challenge.aiops.org>



8,000
RMB

亚军

8,000人民币
颁发获奖证书



80,000
RMB

冠军

80,000人民币
颁发获奖证书



4,000
RMB

季军

4,000人民币
颁发获奖证书

The screenshot shows the AIOps challenge website interface. On the left, a table lists the top 5 teams with their scores:

1. LogicMonitor-AI	0.795670
2. D.I.(H3C)	0.771397
3. ICA128	0.734942
4. 火眼金睛	0.721988
5. 烧脑特工队	0.645889

Below this table is another table showing challenge statistics:

First Challenge	
#of data download	338
enrolled	125
formally competed	59

The photo shows three men standing behind a purple banner. The banner has text in Chinese and English, including "首届 AIOps 挑战赛决赛 冠军 奖金 80,000 元". One man is holding a framed certificate. The background features a stage setup with curtains and logos.

AIOps挑战赛及社区 challenge.aiops.org :

汇聚AIOps 爱好者

AI运维问题模型
数据集
挑战赛
论坛



有动力、有算法基础的运维工程师有潜力成功转型AIOps

首届挑战赛	
数据下载	338
报名	125
正式参赛	59

1. LogicMonitor-AI 0.795670
2. D.I.(H3C) 0.771397
3. ICA128 0.734942
4. 火眼金睛 0.721988
5. 烧脑特工队 0.645889

Data Sponsors



Website



co-organizer



Summary

- * AIOps is rising: replace manual Ops decisions with AI-based decision aids
 - * Improved revenue
 - * reduced loss/cost
 - * necessary
 - * feasible
- * Case studies (Collaboration with Baidu, Alibaba, Tencent, Didi, Sogou):
 - * Anomaly Detection
 - * Anomaly Localization
 - * Root Cause Analysis
 - * Capacity/Failure Prediction
- * AIOps Challenge: Community efforts for widespread adoption of AIOps

THANK YOU

