

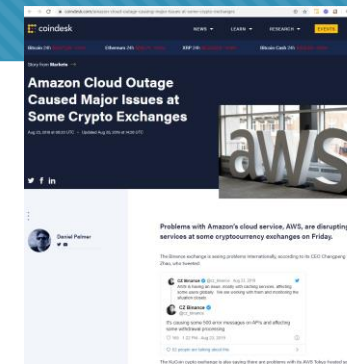
Fighting the Fog of War: Automated Incident Detection for Cloud Systems

Liqun Li, Xu Zhang, Xin Zhao, Hongyu Zhang, Yu Kang, Pu Zhao, Bo Qiao, Shilin He, Pochian Lee, Jeffrey Sun, Feng Gao, Li Yang, Qingwei Lin, Saravanakumar Rajmohan, Zhangwei Xu, and Dongmei Zhang

Microsoft Research, Microsoft Azure, Microsoft 365, The University of Newcastle, University of Chinese Academy of Sciences

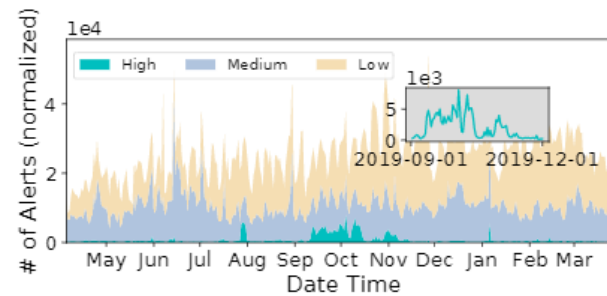
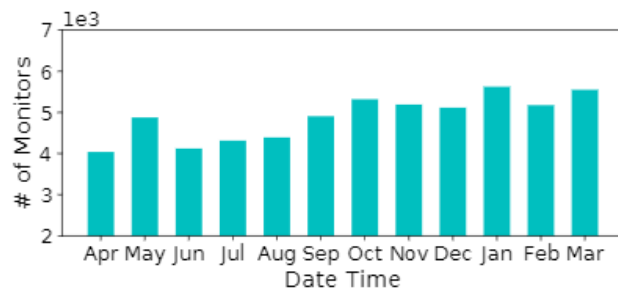
Background

- Reliability is a key quality attribute of large-scale cloud systems
- Incidents/outages dramatically degrade the service quality
 - Tough incidents/outages take a long time to mitigate
 - Costs: \$17K/outage·min (2016)*



Alerts

- Alerts are system events that require attention
 - Reported by the monitoring infrastructure
 - E.g., API timeouts, operation warnings, unexpected VM reboots, etc.
 - Severity: low, medium, high
 - Handled by on-call engineers



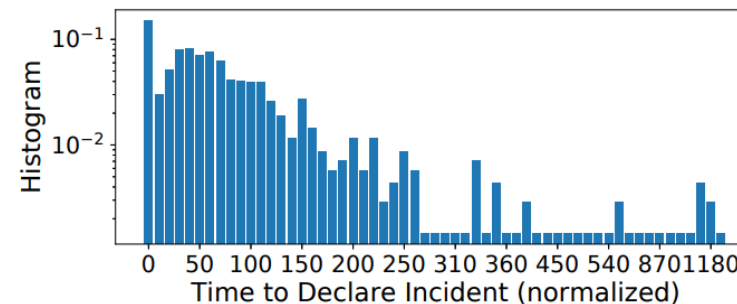
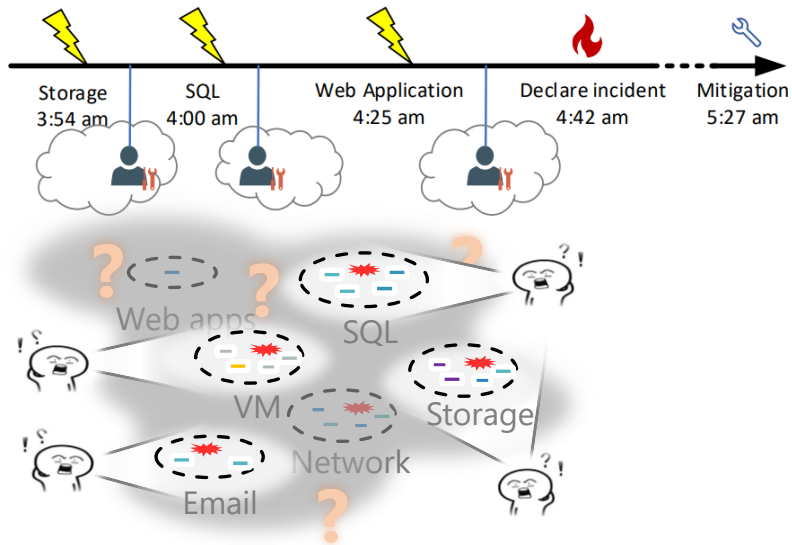
of monitors and alerts from Big5 services in Azure

Alert ID: 200603407	Title: Ongoing VM critical failures	Severity: High
Service: Compute	Team: OS	Owner: Lily
Start Time: 2020-03-13 07:30:00	Mitigation Time: 2020-03-13 08:23:00	
Region: A	Data Center/Cluster (Optional): xxx/xxx	
Diagnosis logs		Monitor ID: DataCenterFailure

Main fields of an alert

Incident Management

- Incident: situations with customer impact, taking a long time to resolve, or requiring cross-team collaboration
- Timely incident management is the key to reduce system downtime
- Incident declaration turns chaos into order



A long tail of incidents take a long time to declare

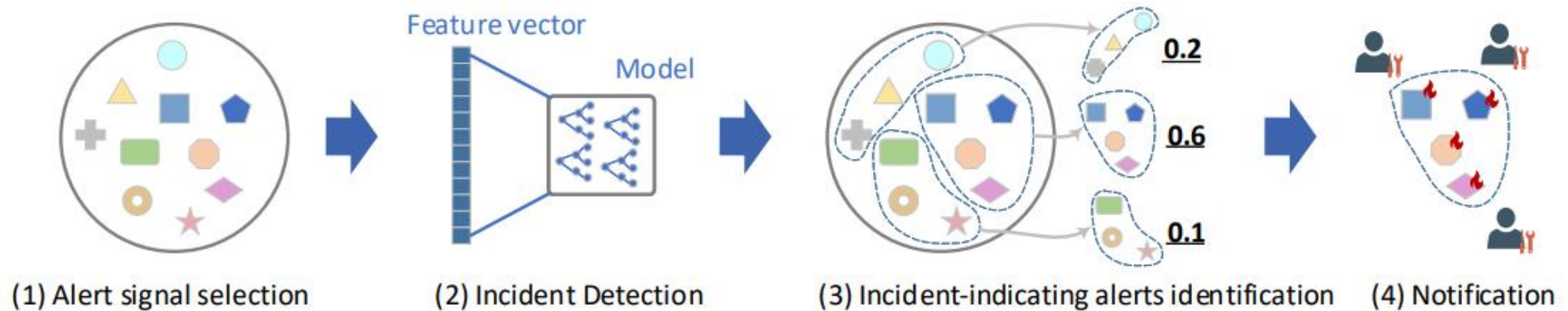
Engineers are like in the fog-of-war

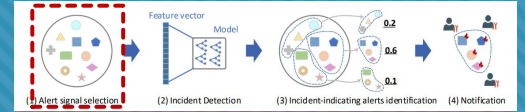
Related Work

- **Fault detection and localization**
- **Time-series anomaly detection**
- **Cloud incident management**

Warden: Automated Incident Detection

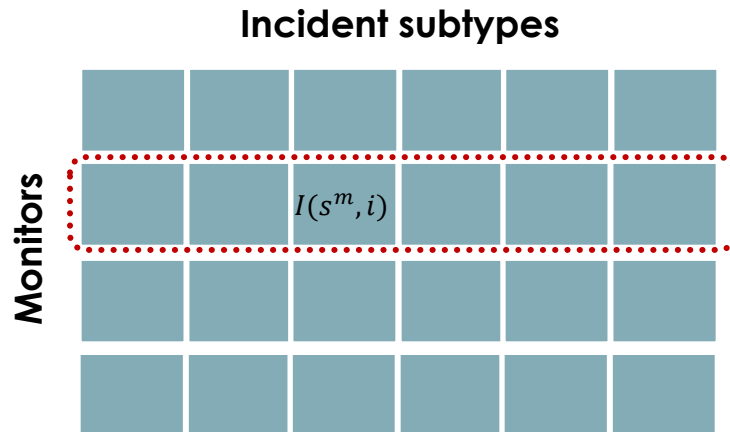
- Detect the ongoing potential incidents from the alerts
- Extract incident-indicating alert groups for notification





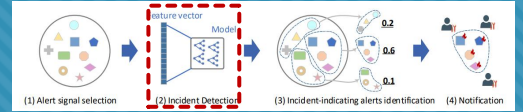
Alert Signal Selection

- **Select a subset of monitors which exhibit relatively strong association with incidents**
 - Categorize incidents based on their responsible teams
 - Calculate the sum of Weighted Mutual Information (WMI) for each monitor with all subtypes of incidents



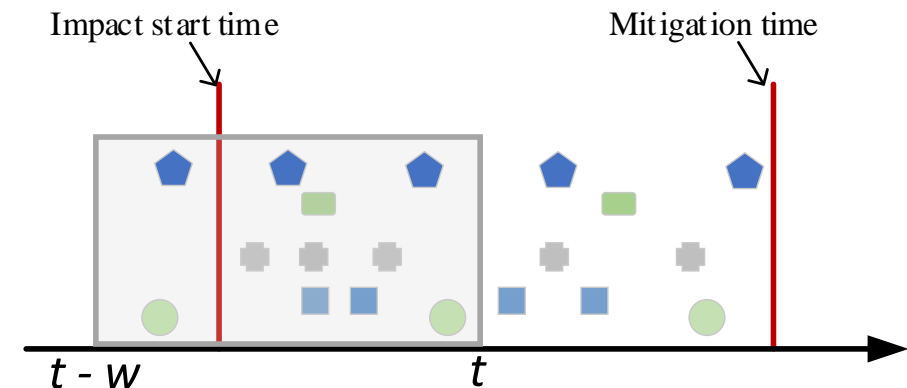
$I(s^m, i)$ is the information gain by observing alerts from a monitor m about predicting incidents of subtype i

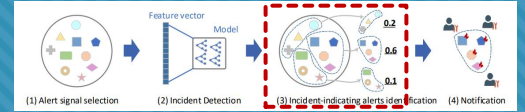
The score of monitor m is $\sum_i I(s^m, i)$



Incident Detection

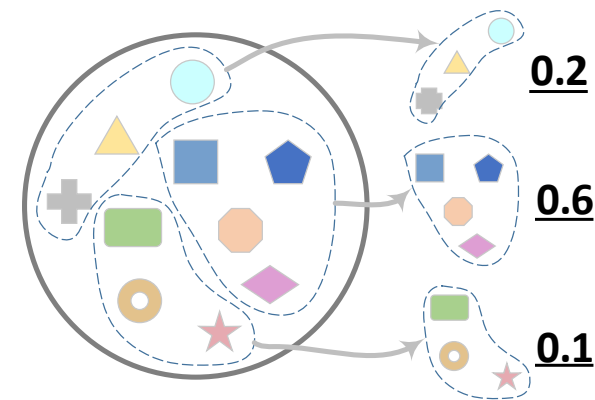
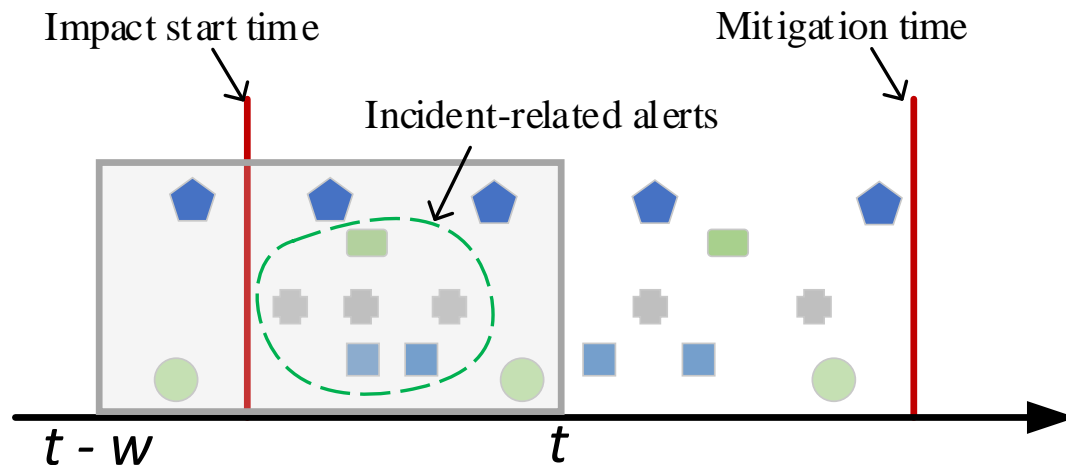
- **Incident detection: a binary classification problem**
 - Input: alerts reported by selected monitors in a recent time window
 - Output: 1 if there is potential ongoing incidents; otherwise, 0
- **Sample construction**
 - Construct samples using a sliding window (3h)
 - Label = 1 if the window is overlapping with incident impact duration; otherwise, label = 0
- **Feature extraction:**
 - Alert signals: alert count, alert burst
 - Engineer activities: diagnosis log count, notification count
 - Others: region, working day, hour in day
- **Classification model: BRF (Balanced Random Forest)**

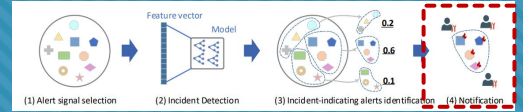




Identifying the Incident-indicating Signals

- **Incident-indicating signals: alerts related to the incidents**
 - Alert signal grouping: correlation and rule-based
 - Group-based model interpretation: GSV (Group Shapely Value)





Warden in Practice

- Detecting emerging issues
- Notify all engineers working on incident-indicating alerts
- Once confirmed, engineers form a cross-team collaboration group to diagnose and mitigate the incident

Recommended Actions panel on the alert page

Recommended Actions

This alert is involved in a **HIGH** confidence emerging issue
 Impacted Services: Compute, WebApplication
 Alert Count: 5 (A)

Emerging Issue [Subscribe to Potential Emerging Issue Alerts](#)

Title (B)	Service	Location	Detection Time	Confidence	Alert Count
> Emerging issue detected with at least 5 alert	Compute, WebApplication	XXX	03/13/2020 17:09 UTC	High	5

State	Severity	Id	Title	Service	Owner	StartTime	Actions
ACTIVE	High	XXX	Large spike of virtual machine critical failure	Compute	XXX	03/13/2020 17:03 UTC	
ACTIVE	Medium	XXX	Small spike of virtual machine critical failure	Compute	XXX	03/13/2020 16:59 UTC	
ACTIVE	High	XXX	The availability of operations is low	Compute	XXX	03/13/2020 17:05 UTC	
ACTIVE	Medium	XXX	Canary sites down 5%	WebApplication	XXX	03/13/2020 17:03 UTC	
ACTIVE	Medium	XXX	Canaries failing permanently	WebApplication			

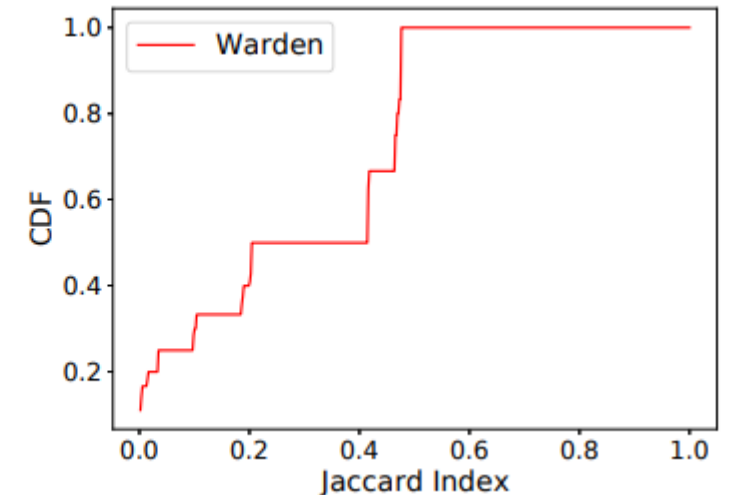
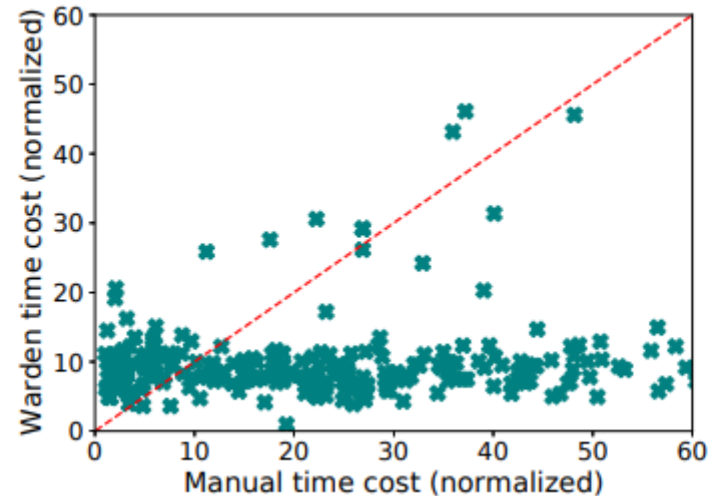
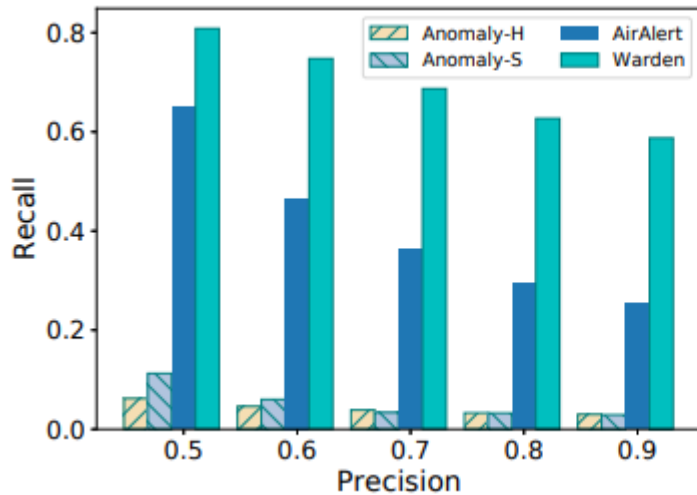
Expanded view of the incident-related alerts

> Emerging issue detected with at least 3 alerts	Networking	XXX	03/08/2020 6:21 UTC	Medium	2
> Emerging issue detected with at least 10 alerts	SQL DB, + 4	XXX	03/01/2020 12:03 UTC	High	21

Dashboard of emerging issues

Experimentation

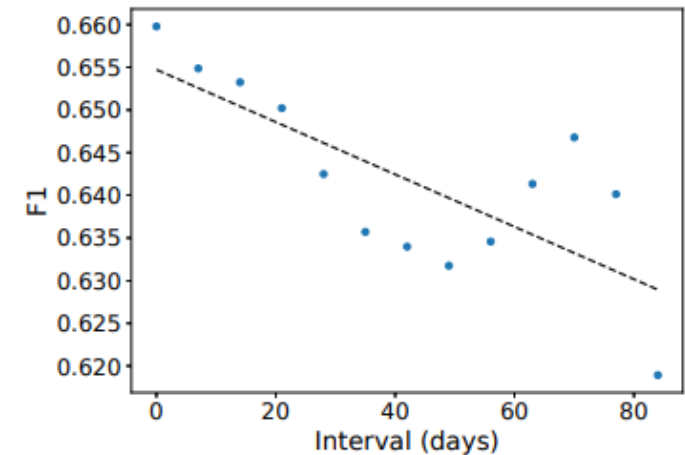
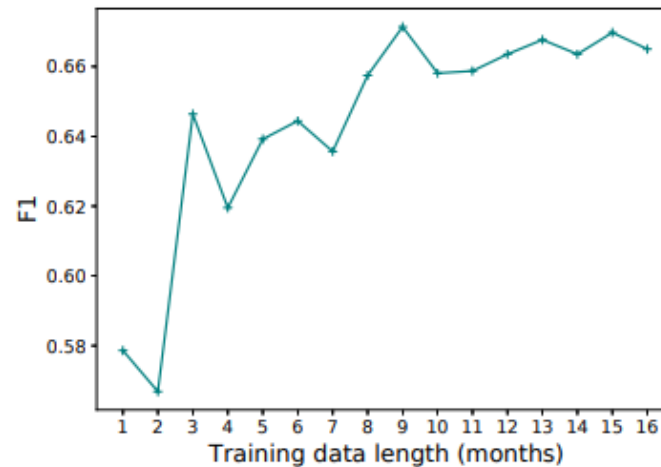
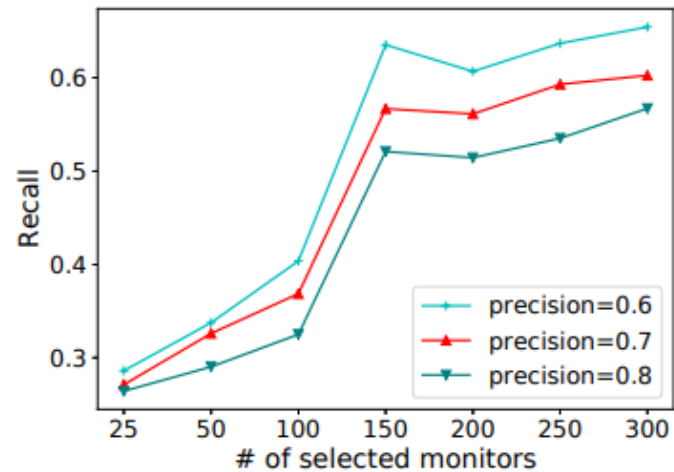
- 18-month-length ~240G data from Azure ICM
- 26 major services, Hundreds of people behind each service, ~72% of all incidents of Azure
- Training: 16 months; Testing: 2 months
- Baselines: Anomaly-H/S, AirAlert



Experimentation

○ Important system parameters

- # of selected monitors
- Data requirements for training the detection model



Conclusions

- **Warden is a framework to detect incidents in an automated way**
 - Detecting potential incidents
 - Extracting related alert signals and notifying relevant engineers
- **Warden is proven to be effective with data collected from 26 major services and real deployment in the ICM of Azure**

Thanks