2023 IEEE International Conference on Computers, Software and Applications (COMPSAC 2023)

Context-aware, Composable Anomaly Detection in Large-scale Mobile Networks

Nguyen Ngoc Nhu Trang
Daienso Lab, Vietnam
www.daienso.com

Hong-Linh Truong
Department of Computer Science
Aalto University, Finland
https://rdsea.github.io

Outline

- Motivation
- Composable, context-aware anomaly detection
- Experiments
- Conclusions

Large-scale mobile networks: multiple mobile sites/cells, different zones, different businesses

⇒ subjects under analytics are very dynamic

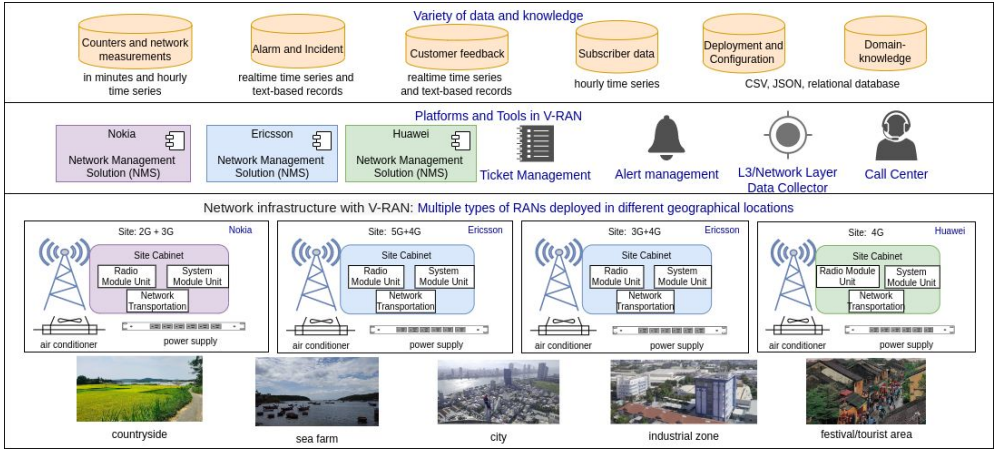


Figure source: "HAIVAN: a Holistic ML Analytics Infrastructure for a Variety of Radio Access Networks," *2022 IEEE International Conference on Big Data (Big Data)*

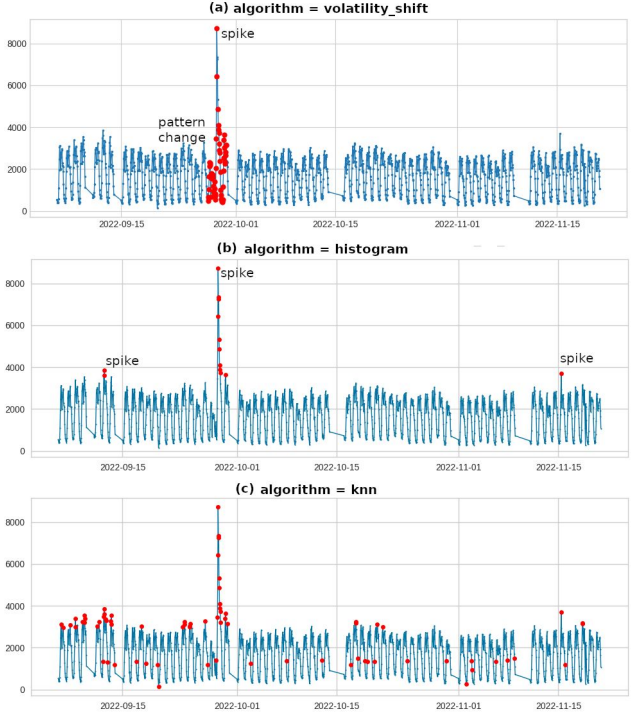
Example: our studied network has ~7000 mobile sites, ~30K mobile cells, covering 12 provinces in the central part of Vietnam

Anomaly detection enables telco operations to find problems in traffic, usage, connectivity status, etc., but there is no lack of anomaly detection algorithms and systems.

So what has motivated our work?

Anomalies detected by algorithms but **irrelevant** to analytics contexts

- **Analytics subjects are diverse**: dynamic zones with different constraints/properties
- **Anomaly detection is not universally applicable** for dynamic subjects
- **Various contexts affect the development and execution** of detection pipelines

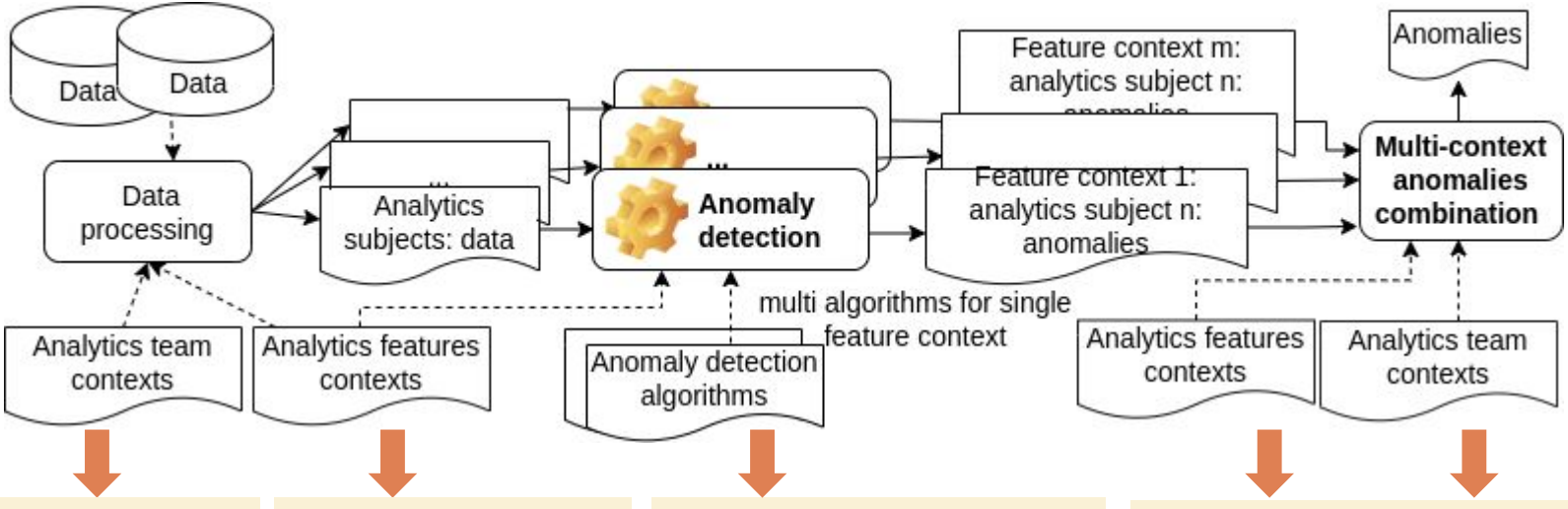


How to solve the irrelevant situation? - Understanding the role of contexts

- **The context of data**
  - quality, volume, delay, availability (due to measurements and data access)
- **The context of analytic subjects**
  - the same type of subjects but with different type of businesses, markets, thresholds, etc.
- **The context of team operations**
  - deployment, time, no ML experts

⇒ Understanding and bringing contexts into end-to-end anomaly detection for finding relevant anomalies

Context-awareness in end-to-end detection



Understanding the **team operation's needs**

Capturing relevant **data and analytics requirements**

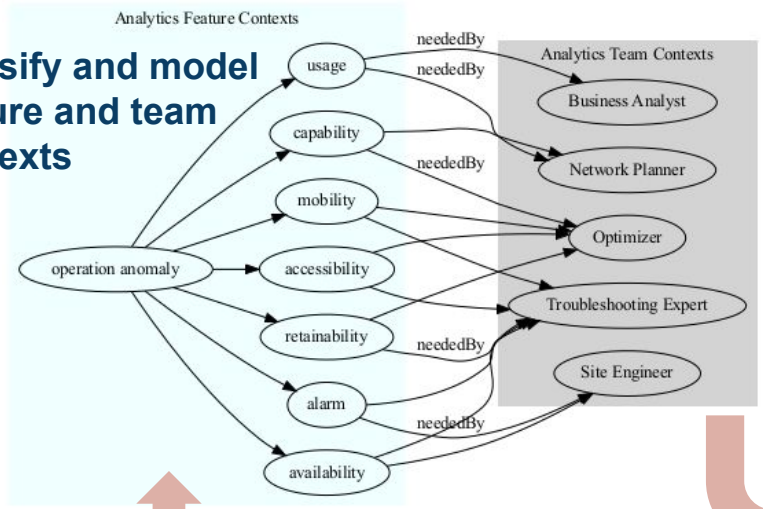
Utilizing suitable **algorithms for relevant anomaly patterns**

Assessing the relevance and importance based on **context and domain knowledge**

Contexts, data and algorithms for suitable anomaly patterns

Configure patterns & algorithms

Classify and model feature and team contexts

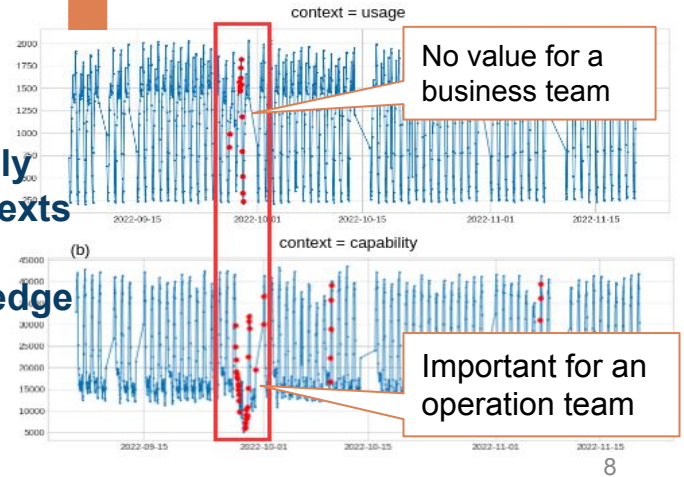


Market type	Analytics context type	(Anomaly pattern, [selected algorithms])
Key market region	alarm	(spike, [interquartilerange, knn])
	capability	(spike, [pca, outlierdetector])
	availability	(spike, [histogram])
	accessibility	(spike, [histogram, spectral residual]), (pattern change/level shift, [autoregression, volatility
Developing market region	capability	(spike, [pca])
	accessibility	(spike, [histog (pattern chang sion, volatility
Potential market region	accessibility	(spike, [histog (pattern char ity_shift, level

```
"potential_market_region": {
    "accessibility": [
        {
            "pattern_name": "spike",
            "window": "1 hour",
            "algorithms": ["histogram","spectralresidual"]
        },
        {
            "pattern_name": "pattern_changes",
            "windows": "5 hours",
            "algorithms": ["volatility_shift","level_shift"]
        }
    ]
}
```

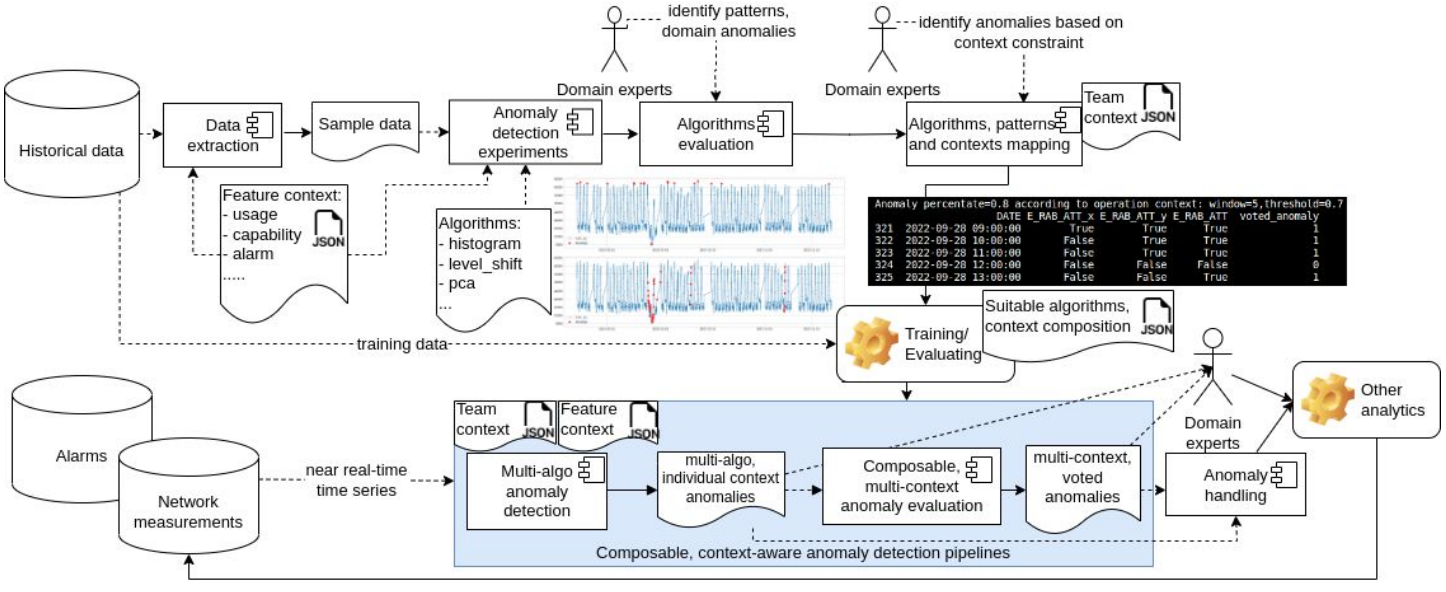
Analytics feature context types: Description of example data	Data: record example
usage: hourly time series of traffic TRAFFIC4G: the total traffic of all 4G cells in the zone	('DATE' 'DISTRICT' 'TRAFFIC4G'): ('2022-09-06 09:00:00' '****' '1286.79716')
alarm: real-time time series alarm: including duration, type of alarms, etc. by window time, alarm starttime (SDATE) and endtime (EDATE)	('VENDOR' 'SITE' 'CELL' 'NETWORK' 'SDATE' 'EDATE' 'ALARM_TYPE' 'SEVERITY' 'ALARM_NAME'): ('****' '****' '****' 'RAN_4G' '2022-09-01 00:30:00' '2022-09-01 00:32:49' '1' 'A3' '61631')
availability: hourly time series of availability calculated from the rate of total serving time per hour of all 4G cells in the zone	('DATE' 'DISTRICT' 'AVAILAIBILITY'): ('2022-09-07 10:00:00' '****' '99.67')
capability: hourly 4G downlink throughput (THP_DL) as the rate of successful message (THP_VOL_DL) per time (THP_TIME_DL) of all 4G cells in the zone combined with 4G downlink traffic	('DATE' 'DISTRICT' 'THP_DL' 'TRAFFIC4G_DL'): ('2022-09-07 10:00:00' '****' '16459.69' '1196.95')
accessibility: hourly time series of access attempt (RRC_ATT, E_RAB_ATT, CSFB_ATT), calculated as the total attempt of all 4G cells in the zone	('DATE' 'DISTRICT' 'RRC_ATT' 'E_RAB_ATT' 'CSFB_ATT'): ('2022-09-06 09:00:00' '****' '1202554' '1079454' '45026')
mobility: hourly data of handover attempt (INTERRAT_HO_ATT, INTRAFEQ_HO_ATT), calculated as the total inter-rat and intra-frequency attempt of all 4G cells in a zone	('DATE' 'DISTRICT' 'INTERRAT_HO_ATT' 'INTRAFEQ_HO_ATT'): ('2022-09-06 09:00:00' '****' '3128' '442321')
retainability: hourly data of failure (ERAB_ABNORMAL), calculated as the total abnormal release of all 4G cells in zone	('DATE' 'DISTRICT' 'ERAB_ABNORMAL'): ('2022-09-06 09:00:00' '****' '3053')

Identify anomaly patterns - contexts mapping from domain knowledge



Analyze the diversity in data

Composable, multi-context anomaly detection



Different pipelines are deployed based on the contexts

```
"context_type": "ctx_usage",
"feature_context": {
    "analyticsubject": {
        "subject_fieldname": "DISTRICT",
        "constraints": [{"VENDOR": "Nokia"}]
    },
    "required_data_fields": [
        "TRAFFIC4G"
    ],
    "time_constraints": {
        "from": {"DATE": "2022-09-01"},
        "to": {"DATE": "2022-11-01"}
    },
    "dataquality": {}
}
```

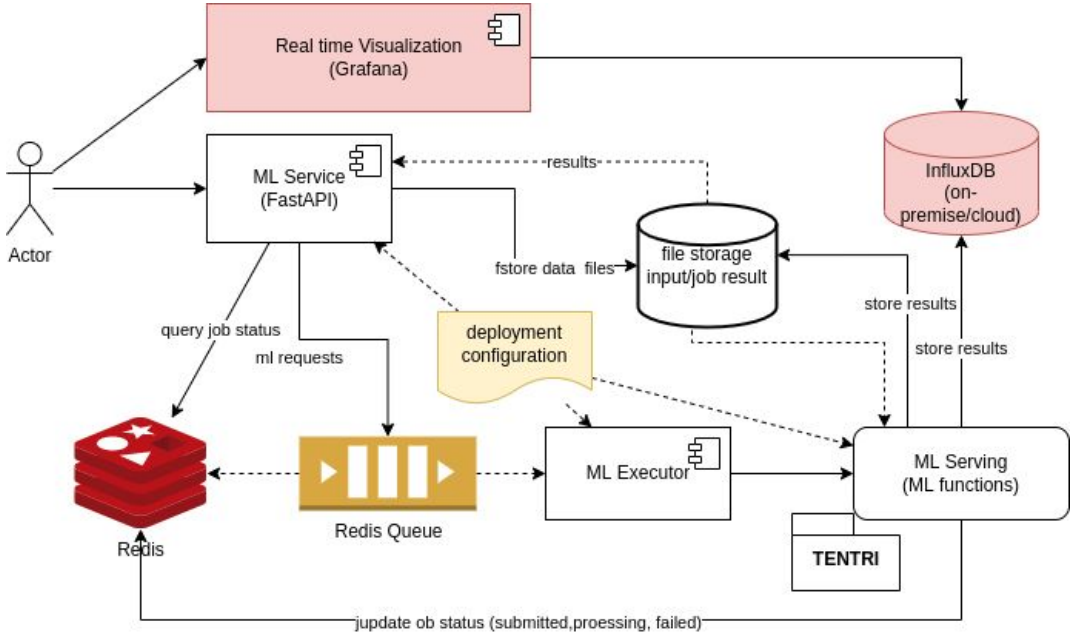
Two combinations: (i) for the same pattern type and among different anomaly patterns in a context; and (ii) among different contexts

```
level_shift,histogram,volatility_shift,spectralresidual,
    voted_anomaly_spikes,voted_weight_spikes,voted_ano
maly_pattern_changes,voted_weight_pattern_changes,
    voted_anomaly,voted_weight
1.0,0,1.0,0.0,0.0,0.0,1.0,0.66,1.0,0.5
0.0,0,1.0,0.0,0.0,0.0,1.0,0.33,1.0,0.25
```

Experiments

- **Prototype with Python (Pandas, Spark, etc.)**
- **Reuse multiple detection algorithms**
  - select, test and evaluate selected algorithms from existing frameworks
- **Contexts applied along pipelines**
  - data processing → detection → voting
- **Data**
  - ~ 3+ months data: 6.09.2022-20.11.2022
  - network measurement at the mobile cell level (136 districts, 6 contexts)
  - experimented with 9 districts/3 types of market regions ⇒ 48 districts ⇒ 11 user-defined zones (5-20 sites per zone, 0.7-5 km2 coverages)

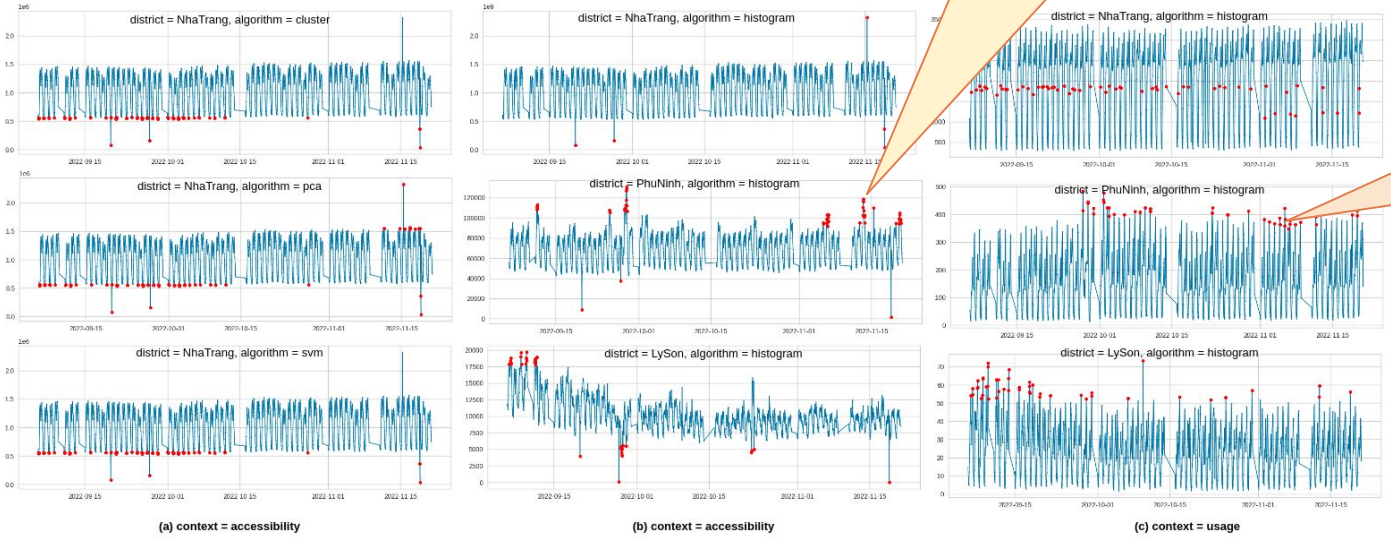
One example of deployments



Understanding sensitivity of algorithms and contexts

Histogram is good for spikes in operations (accessibility)

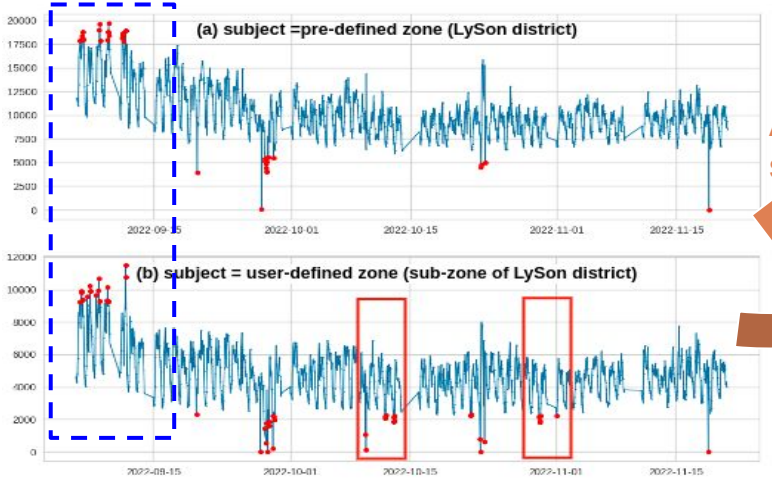
Wrong/irrelevant for spikes in business (usage)



Dynamic subjects of the same type must be analyzed according to their context

Combining anomalies from multiple user-defined and pre-defined zones

- **Analytics sequence**: user-defined zone ⇒ pre-defined zone ⇒ finding the common problems
- **Common problems**: when a coarse-grained level (*a pre-defined zone*) is in sync with a fine-grained level (*user-defined zones, which are sub-zones of a pre-defined zone*)
- **Repeated patterns:** common anomalies in different user-defined/ pre-defined zones with similar time and type of business




```
zone_type,vendor,E_RAB_ATT,histogram_anomalyresults
pre-defined district zone 1,A,809,1
pre-defined district zone 2,A,2369,1
pre-defined district zone 3,A,445,1
user-defined zone 4,A,0,1
user-defined zone 5,A,0,1
pre-defined district zone 6,B,1377083,0
user-defined zone 7,B,6351,0
```

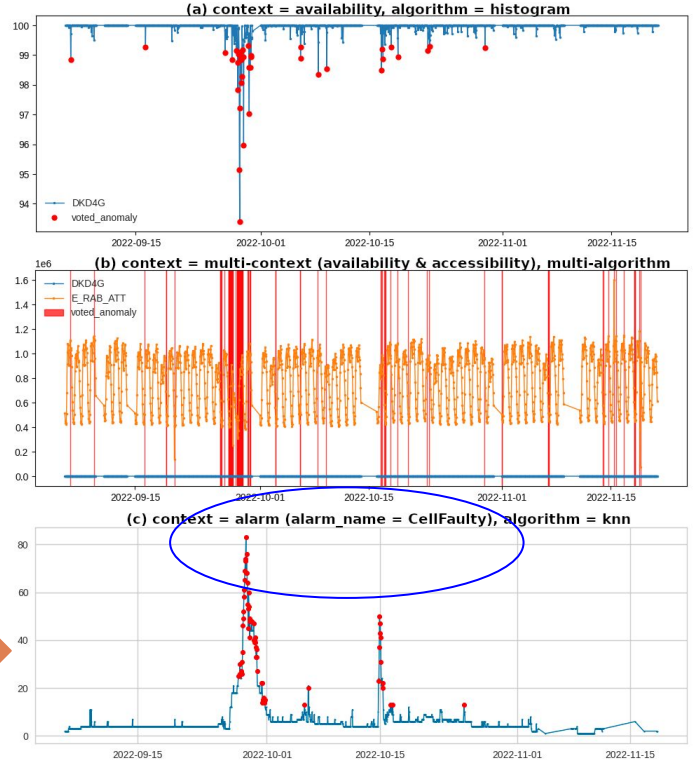
It is hard to detect problems with **Vendor A** with a single subject

Multi-context anomaly and alarm for operations

Multi-context anomaly detection provides higher confidence and relevance w.r.t. problems in operations

```
DATE,availability,accessibility,voted_anomaly,voted_weight
2022-09-28 05:00:00,0,1,1,0.5
2022-09-28 06:00:00,1,1,1,1
2022-09-28 07:00:00,0,1,1,0.5
2022-09-28 08:00:00,1,1,1,1
2022-09-28 09:00:00,1,1,1,1
2022-09-28 10:00:00,1,1,1,1
2022-09-28 11:00:00,1,1,1,1
```

Strong relations between accessibility and availability anomalies ⇒ based on domain knowledge to suggest next steps



Conclusions and future work

- **Context-aware, composable approach**
  - context is considered in all phases of detection pipelines
  - composition of multiple algorithms and composition of pipelines

  ⇒ bringing relevant anomalies and supporting different ways of detection deployment
- **Future work**
  - user-defined zones
  - composable anomaly detection for other domains

scan the paper

Thank you for your attention!