SAMSUNG

MSFRD: Mutation Similarity based SSD Failure Rating and Diagnosis for Complex and Volatile Production Environments

Yuqi Zhang†, *Tianyi Zhang*†, *Wenwen Hao*†, *Shuyang Wang*†, *Na Liu*†, *Xing He*†,
Yang Zhang‡, *Weixin Wang*‡, *Yongguang Cheng*‡, *Huan Wang*‡, *Jie Xu*‡, *Feng Wang*‡, *Bo Jiang*‡,
Yongwong Gwon ∗, *Jongsung Na* ∗, *Zoe Kim* ∗, *Geunrok Oh* ∗

†*Samsung R&D Institute China Xi'an, Samsung Electronics*
‡*ByteDance Inc.*
∗*Samsung Electronics*

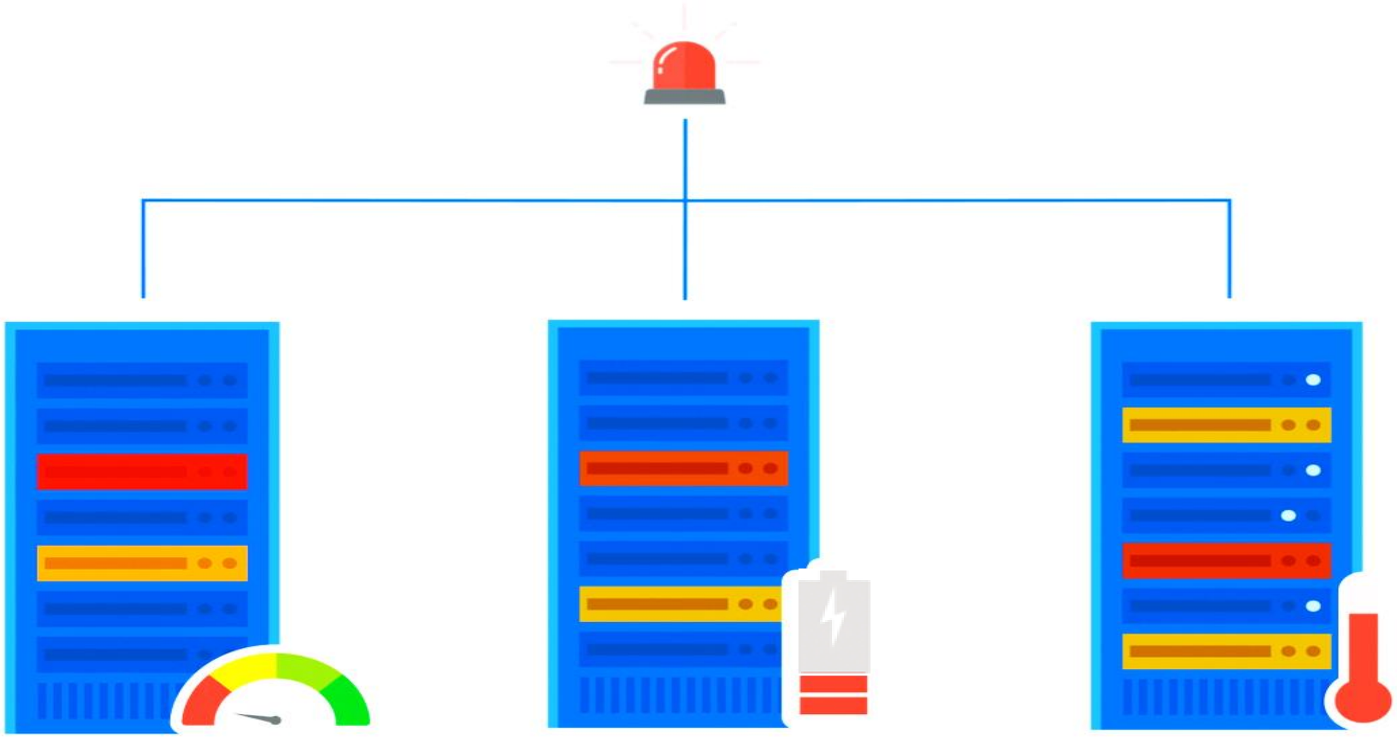
2024 SRCX

Samsung R&D Institute China Xi'an(SRCX) established in 2013 and located in Xi'an High-tech Zone, is the only cutting-edge technology R&D institute of Samsung Electronics in western China.

Introduction

■ SSD failures

- Posing a challenge to the storage reliability of large data centers with millions of SSDs
- Causing instability in online services and additional maintenance costs



SSD failures' impact on storage system: degraded performance, long tail latency, reduced reliability, etc.

Introduction

■ Field data for failure analysis and prediction

- Two large-scale datasets from large Internet companies

  - **51 million + Telemetry** logs
  - 150,000 + Samsung SSDs
  - Failure lists and related information are collected by the operators

- Telemetry logs with **85** customized attributes, more comprehensive than SMART

Type	**Attributes**
Uncorrectable error	lifetime_uecc_count, dram_uecc_count, etc.
Correctable error	dram_cecc_count, read_recovery_attempts, bad_block_count, etc.
Read/write	lifetime_user_reads(writes), trailing_hour_WAF, etc.
Temperature	highest_temperature, over_temperature_minutes, etc.
Wear and capacitor	wear_level_avg, capacitor_health, etc.

…

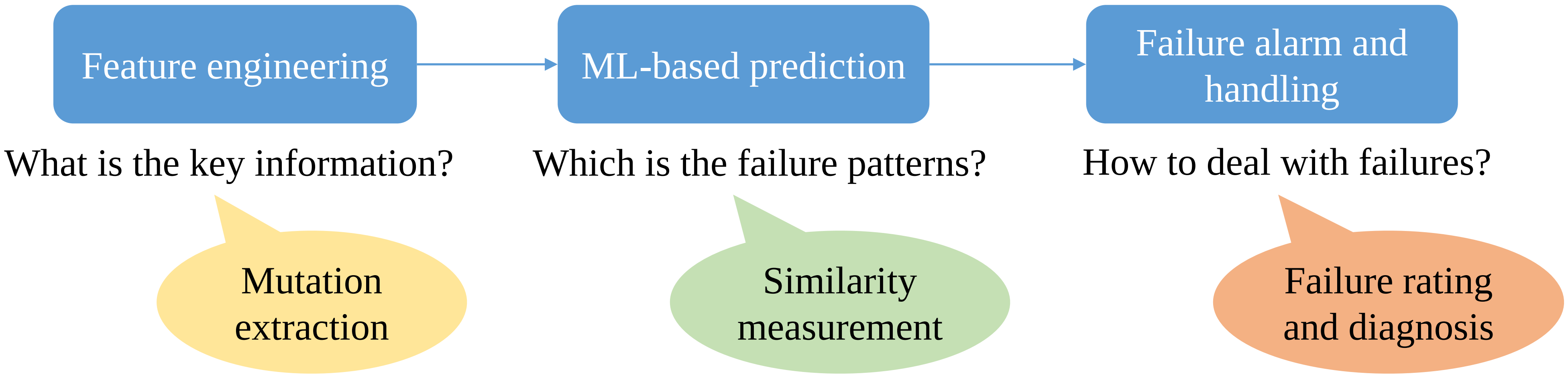
Introduction

■ SSD failure prediction

- A proactive fault tolerance mechanism to reduce the impact and cost of failures
- Three steps: Feature engineering, ML-based failure prediction, Failure alarm and handling

■ Mutation Similarity based Failure Rating and Diagnosis (*MSFRD*) scheme

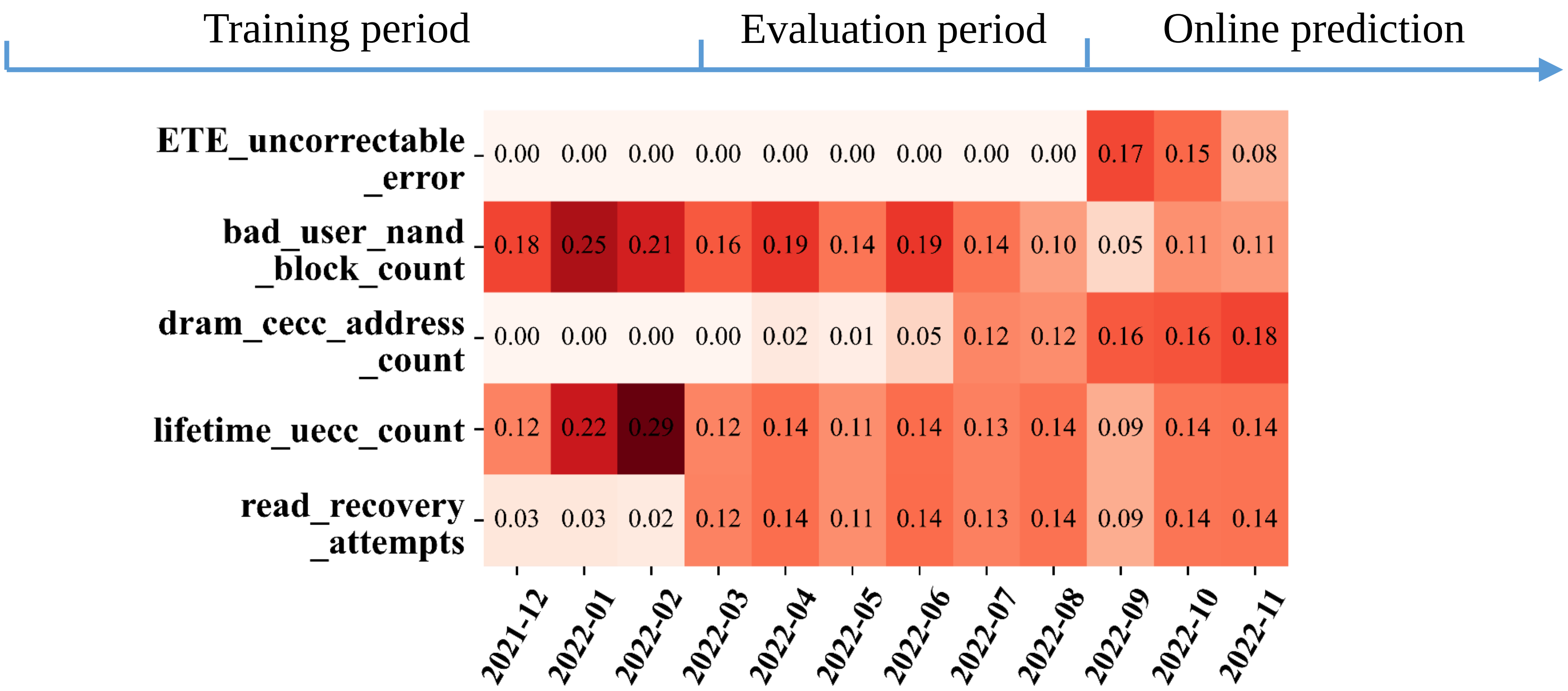
- Dynamic mutation extraction to locate abnormal changes and failure symptoms in time
- Mutation based similarity measurement to capture more failure patterns accurately
- Failure rating and diagnosis for fine-grained failure alarm and handling



Feature engineering

➢ **Finding 1: Feature importance and data range would change over time**

- Training-evaluation-prediction in practice brings time gap between training data and testing data
- Traditional feature selection relying on training data cannot adapt to the data changes online

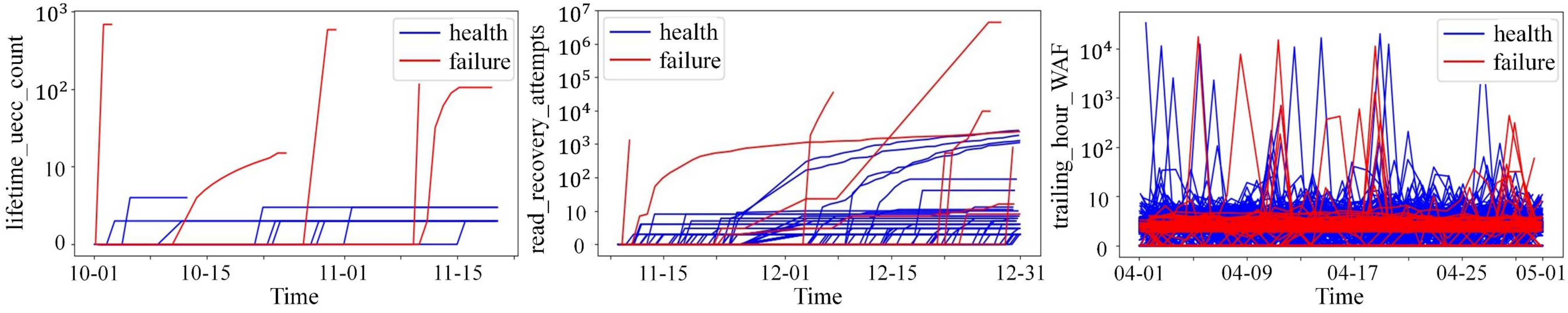


The Pearson correlation coefficients between Telemetry attributes and the failures per month.

Feature engineering

➢ **Finding 2: Telemetry mutations are abnormal changes related to failures**

- Failed SSDs usually have rare, sudden, rapid changes (i.e., mutations) in Telemetry attributes
- Rare mutations often occur before failures
- We can capture mutations in real time, instead of selecting static features based on training data



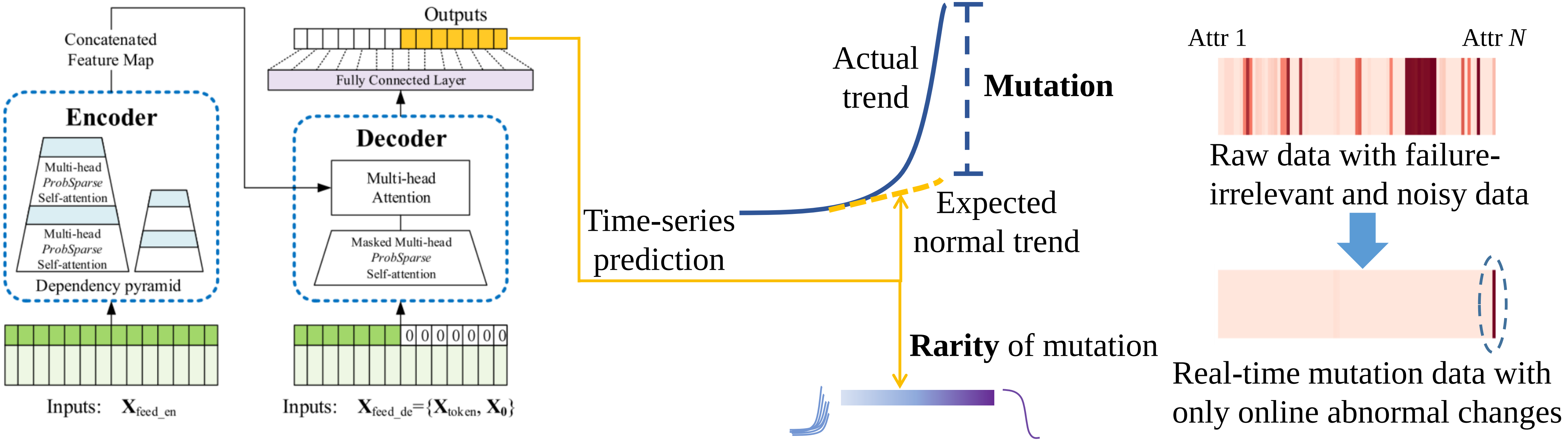
(a) NAND UECC trends. (b) read_recovery_attempts trends. (c) WAF trends.

Attribute trends of healthy and failed SSDs.

Feature engineering

➢ **Design: Dynamic mutation feature extraction**

- Time-series prediction model trained with large-scale data is adopted for normal trend prediction
- The difference between expected normal trend and actual trend represents the mutation
- The model also estimates the rarity of mutation to reflect its importance



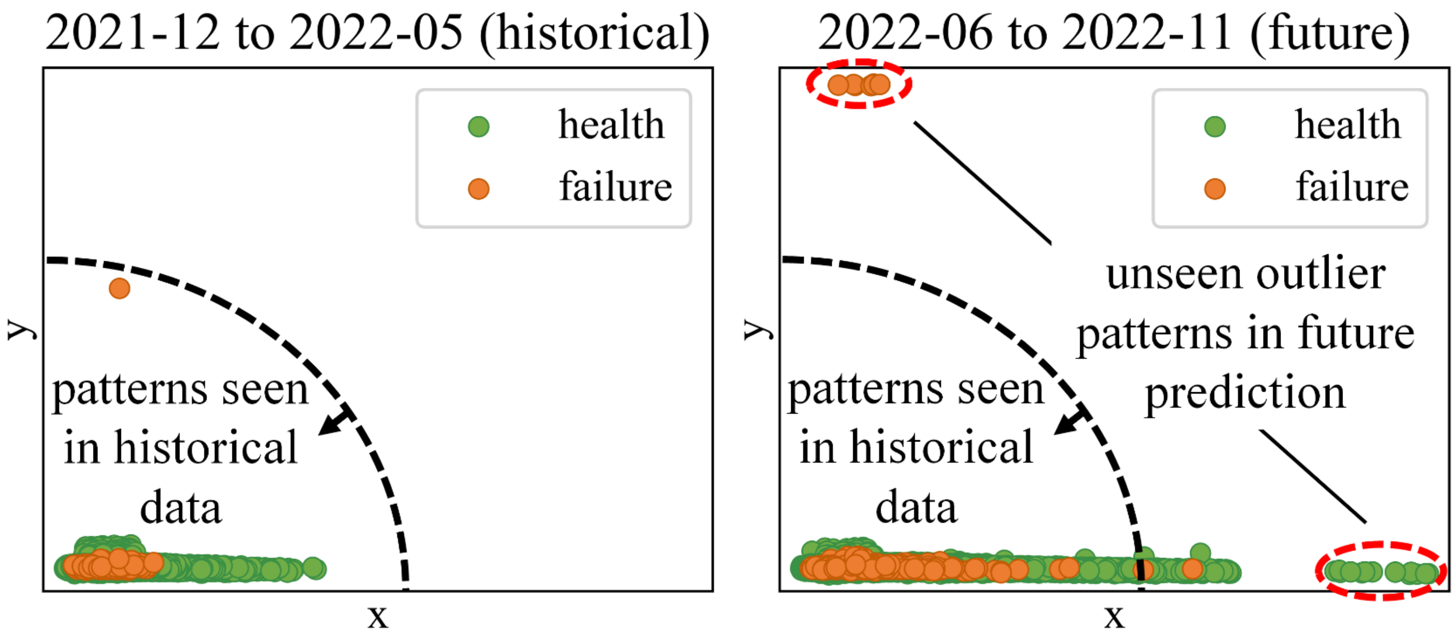
Informer* trained with large-scale healthy time-series data

*Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting, Haoyi Zhou and et.al, AAAI'21

Prediction model

➢ **Finding: Unseen patterns would appear over time in practice**

- Besides the patterns already seen in historical data, unseen patterns would appear in future data
- Unseen patterns also have certain tendency on the distances to historical health and failure patterns

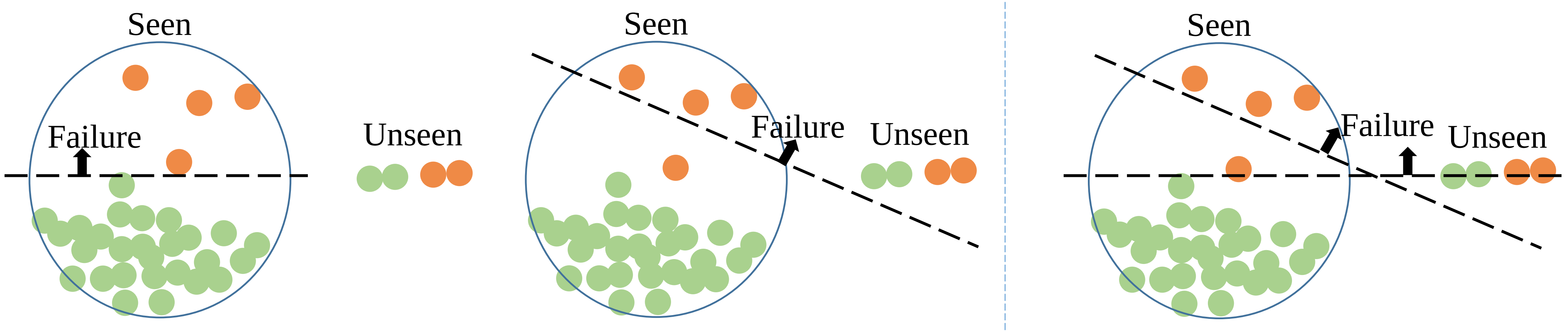


Data patterns over time. The same principle component analysis (PCA) is used to reduce each SSD's monthly data dimensions to two (i.e., x and y) for visualization.

Prediction model

➢ **Design: Mutation based similarity measurement**

- Classification: mainly distinguish health and failure patterns already seen in historical training data
- Anomaly detection: identify outlier patterns including unseen patterns, but outlier ≠ failure
- Our idea: similarity measurement exist in both algorithms to find seen and unseen patterns



Health and failure pattern classification based on historical data

Anomaly detection to find outlier patterns

Ours: similarity measurement exist in both algorithms to take their advantages

Failure   Health

Failure alarm and handling

- **Finding: SSD failures involve various phenomena and degrees**
  - Some gray failures (e.g., perf drop) are also reported but workable later (not replaced)
  - Internal errors, especially uncorrectable ones, reflect the health/failure level of the SSD

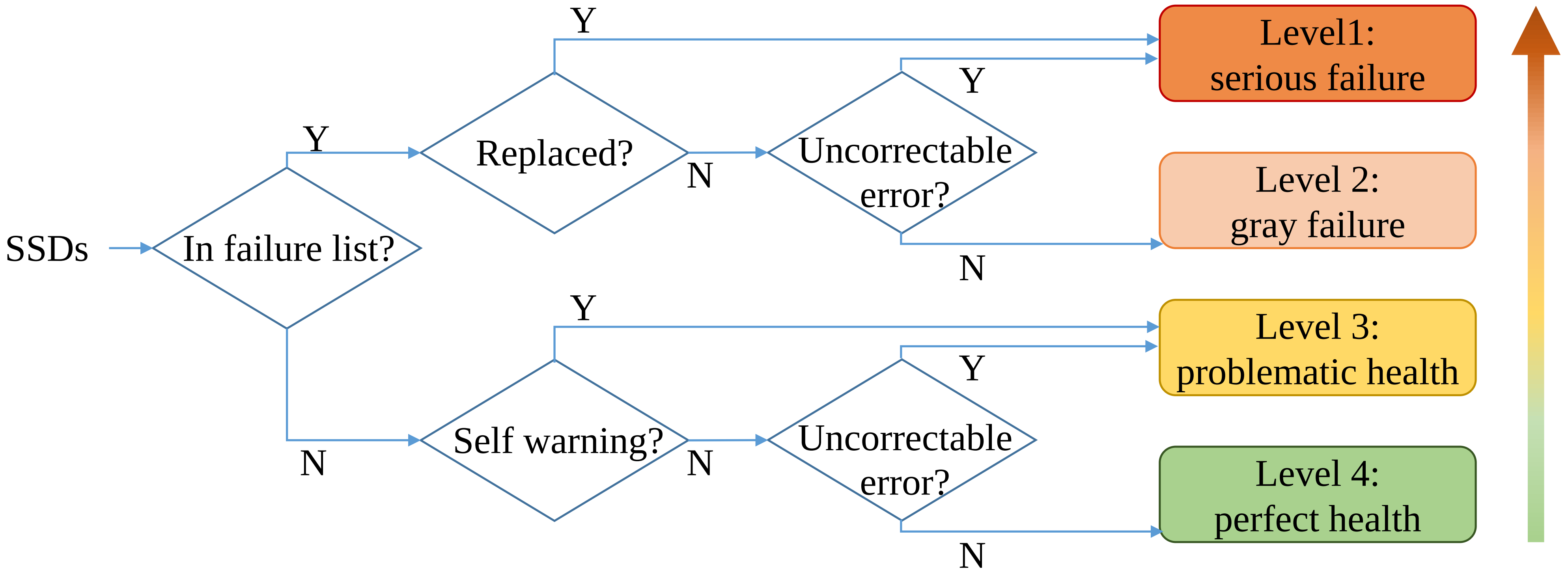
Status	Rate in health	Rate in failure	Replacement proportion
With uncorrectable errors	0.03%	9.71%	94.12%
With correctable errors	44.74%	47.43%	65.06%
Without errors	55.23%	42.86%	46.67%

The rates of SSDs with different error status in healthy SSDs and failed SSDs, and the proportion of failed SSDs that were actually replaced.

Failure alarm and handling

➢ **Design: Failure rating and progressive diagnosis**

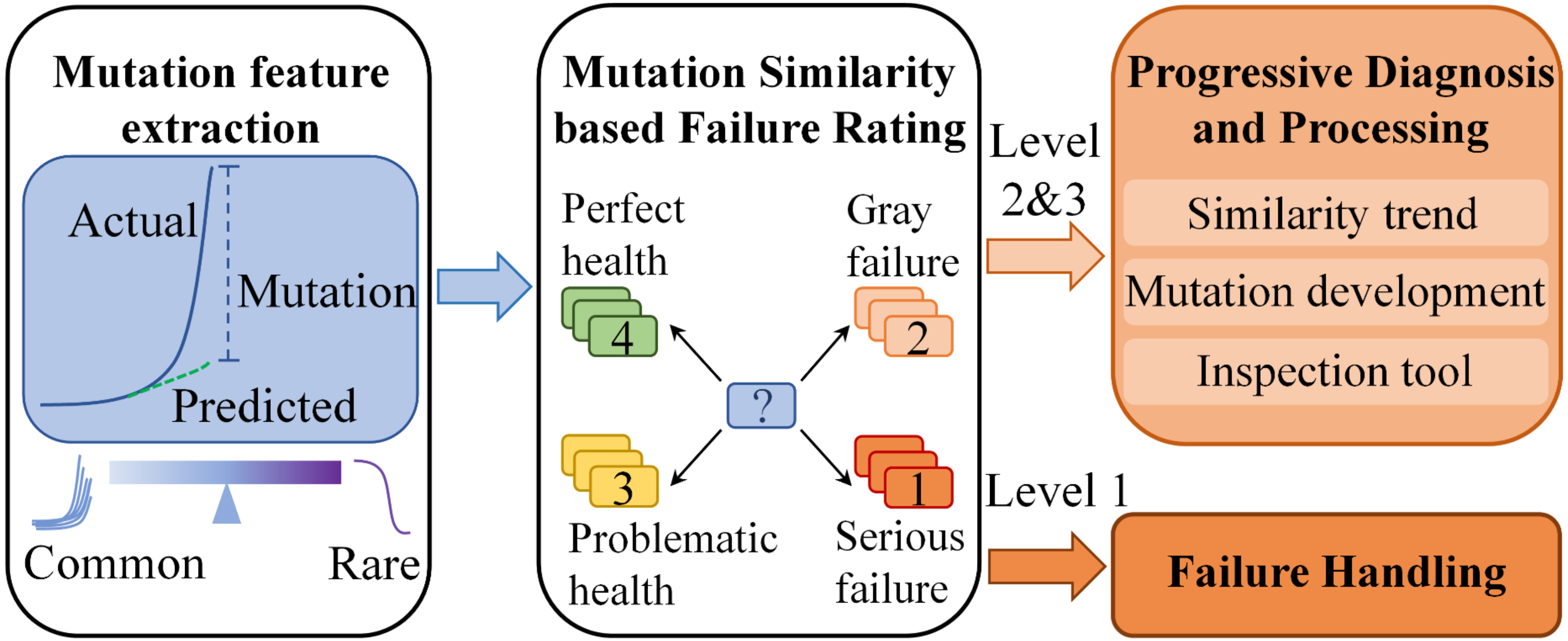
- Define four levels: serious failure, gray failure, problematic health, and perfect health
- Serious failures are handled directly, while gray failures and problematic health are further diagnosed



Methodology

■ MSFRD architecture

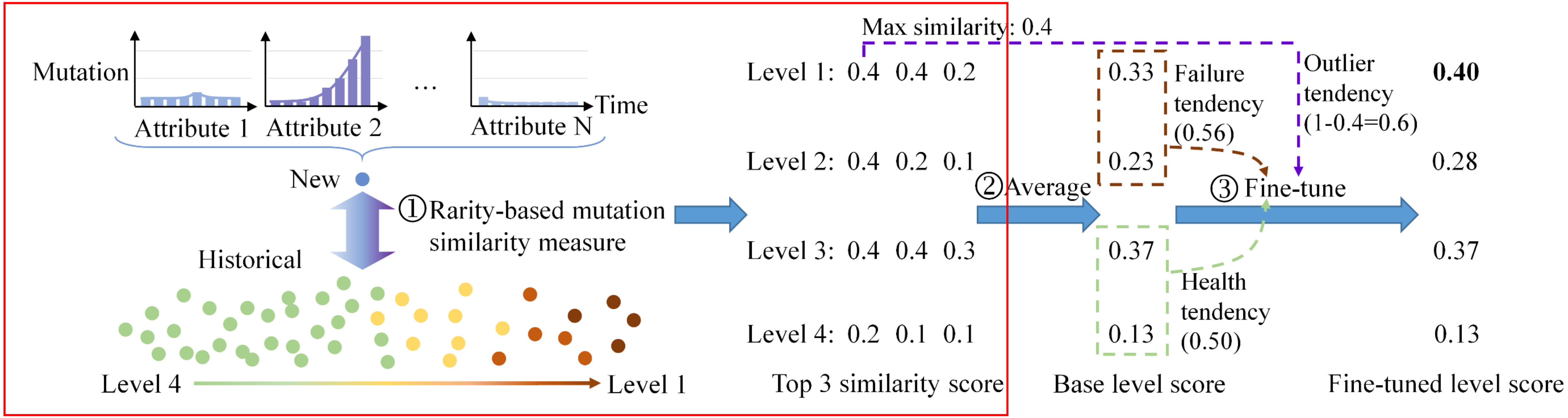
- Mutation feature extraction to capture only abnormal data changes in real time
- Mutation similarity based failure rating to identify seen and unseen failures and fine-grained status
- Progressive diagnosis to handle failures incrementally to minimize the impact on available SSDs



Methodology

■ Mutation extraction and similarity measurement

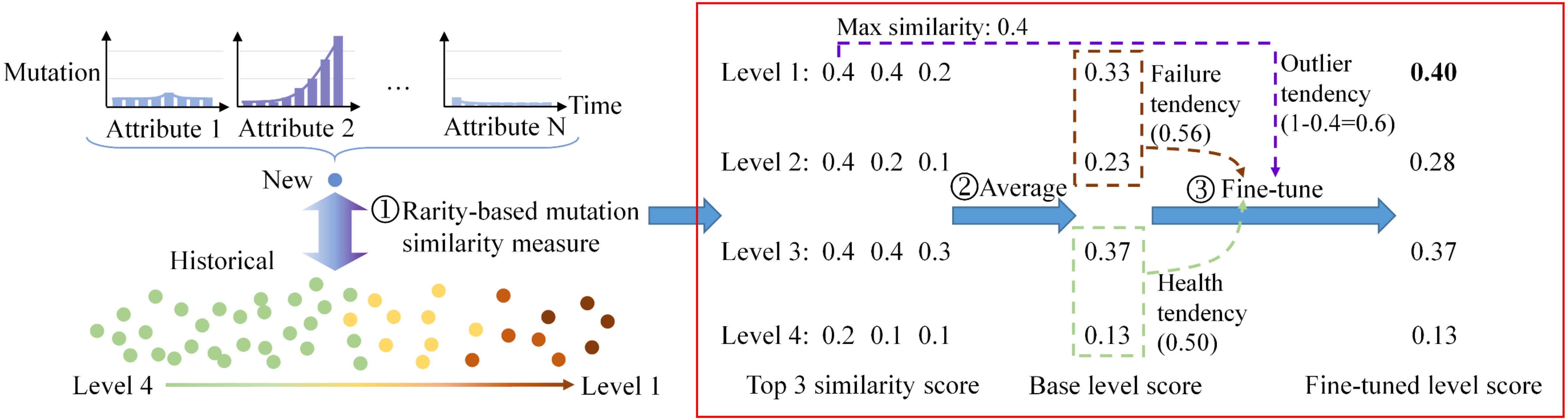
- For all attributes of each SSD, we extract the mutations and their rarity
- For an SSD to be predicted, we calculate its mutation similarity to each historical SSD using the weighted Euclidean distance (mutation rarity as the weight)
- The top 3 historical SSDs that are most similar to this SSD in each level are found



Methodology

■ Mutation similarity based failure rating

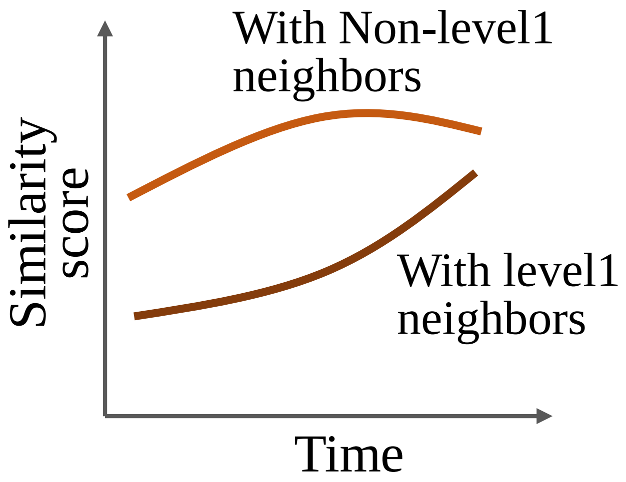
- For each level, the average of the top 3 similarity scores is regarded as the base confidence score
- The confidence scores are fine-tuned with failure, health and outlier tendency

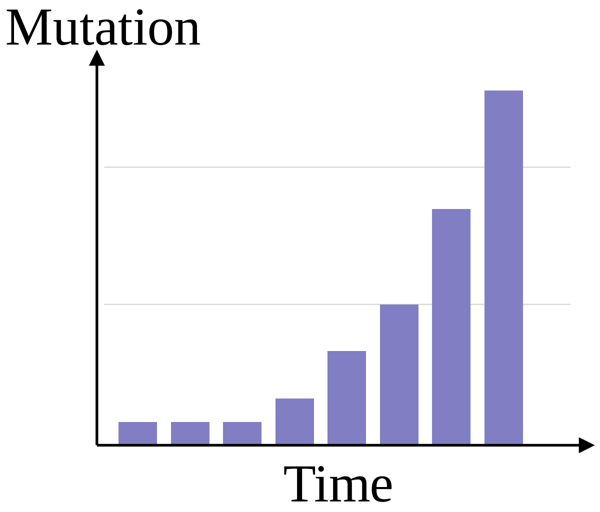


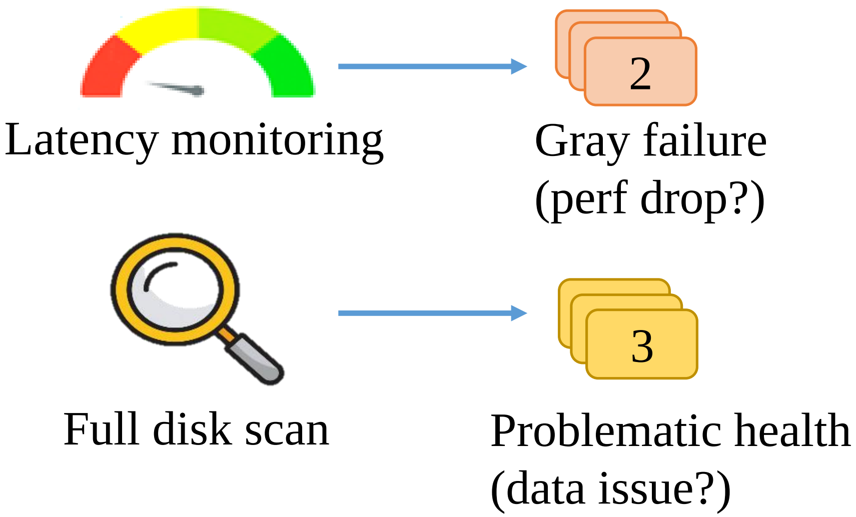
Methodology

■ Progressive diagnosis and processing

- Gray failure (level 2) and problematic health (level 3) will be continuously tracked and diagnosed
- Those that are approaching level-1 historical failures and whose mutations are getting worse are further diagnosed as level-1 failures
- Latency monitoring for gray failures and full disk scan for problematic health







Is it going to be more similar to level-1 historical failures?

Is the mutation getting worse?

Is there any problem by checking with tools?

Automatic trend-based diagnosis

Evaluation

■ Dataset setup

Dataset	Exp. round	Train set (month)	Val set (month)	Test set (month)
41-M Telemetry	1	1–10th	11th	12–14th
	2	1–13th	14th	15–17th
10-M Telemetry	1	1–3th	4th	5th
	2	1–4th	5th	6th
	3	1–5th	6th	7th
MB2 SMART*	1	1–17th	18th	19–21th
	2	1–20th	21th	22–24th

Our Telemetry datasets

Alibaba public SMART dataset

■ Four metrics

- *Precision*: The proportion of true alarms (correctly predicted) to both true alarms and false alarms
- *Recall*: The proportion of true alarms to all actual failed SSDs
- *F0.5-Score*: $\frac{(0.5^2+1)\times Precision\times Recall}{0.5^2\times Precision+Recall}$, harmonic average of precision and recall (with higher weight for precision)
- *Accuracy rate (rating)*: the proportion of true alarms with correctly identified failure levels to all true alarms

**General Feature Selection for Failure Prediction in Large-scale SSD Deployment, Fan Xu, et al., DSN 2021*

Evaluation

■ Evaluation on failure prediction

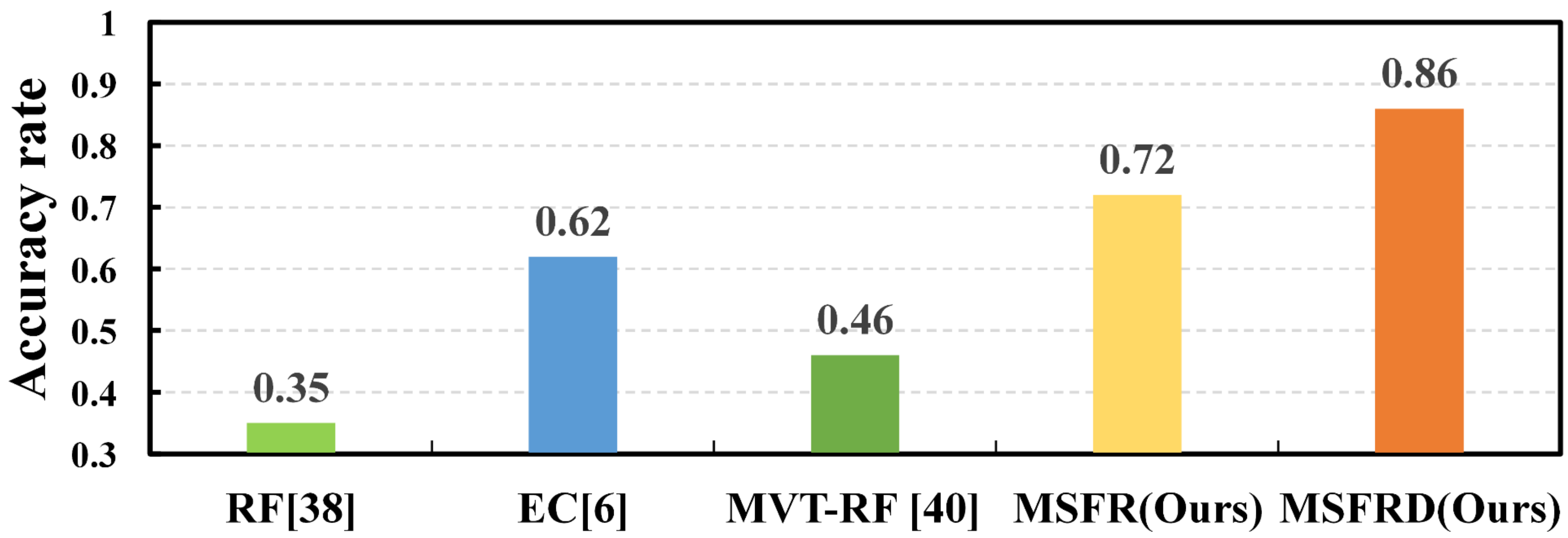
- MSFRD shows better performance on all three datasets, which demonstrates its effectiveness
- MSFRD has greater improvements on complex Telemetry datasets due to its powerful ability to extract key information and adapt to data changes

Methods	41-M Telemetry			10-M Telemetry			MB2 SMART			Average		
	Precision	Recall	F0.5	Precision	Recall	F0.5	Precision	Recall	F0.5	Precision	Recall	F0.5
RF	0.61	0.19	0.43	0.64	0.18	0.38	0.72	0.24	0.52	0.66	0.20	0.44
EC	0.59	0.24	0.44	0.63	0.21	0.44	0.85	0.24	0.57	0.69	0.23	0.48
AE	0.57	0.26	0.46	0.61	0.23	0.46	0.53	0.25	0.43	0.57	0.25	0.45
MVT-RF	0.62	0.28	0.50	0.70	0.27	0.52	**0.87**	0.25	0.58	0.73	0.27	0.53
MSFRD(Ours)	**0.72**	**0.37**	**0.61**	**0.87**	**0.34**	**0.66**	**0.87**	**0.27**	**0.60**	**0.82**	**0.33**	**0.62**

Evaluation

■ Evaluation on failure rating

- Three methods can also perform four level classification, and are compared together
- MSFR (without automatic diagnosis) and the whole MSFRD outperform the existing methods



Evaluation

■ Discussion on MSFRD modules

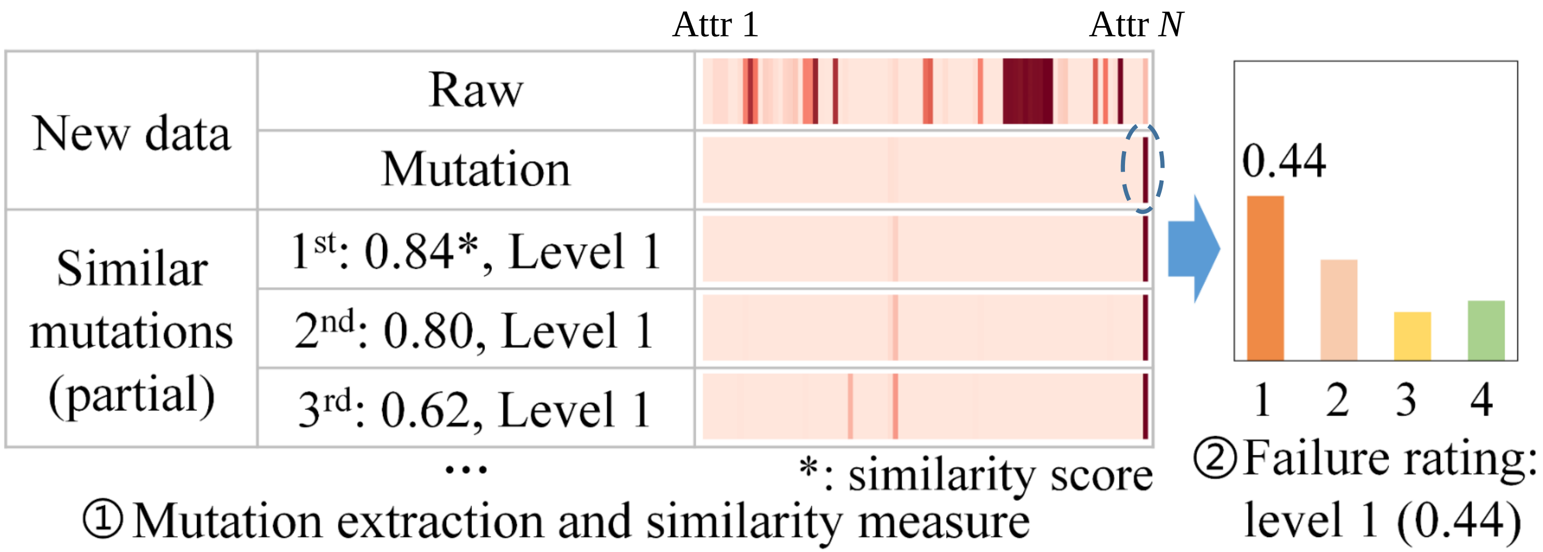
- Dynamic mutation feature performs better than raw data and static feature selection
- Rarity-based mutation similarity measurement and tendency-based rating fine-tuning improve perf
- Coupled with automatic diagnosis on subsequent SSD status, MSFRD achieves the best result

Methods	Precision	Recall	F0.5
Raw data+RF	0.55	0.21	0.41
Feature selection+RF	0.61	0.19	0.43
Mutation feature+RF	0.70	0.24	0.51
Mutation feature+SFR	0.66	0.27	0.52
Mutation(rarity)+SFR	0.67	0.30	0.54
Mutation(rarity)+SFR(tuned)	0.69	0.36	0.58
MSFRD	**0.72**	**0.37**	**0.61**

Evaluation

■ Practical example

- With dynamic mutation extraction, only a few attributes have mutations (with deep color)
- Other attributes that change normally in the raw data are implicitly eliminated
- This example is similar to historical level-1 failures and is therefore rated as level 1



An example where a level-1 failure is correctly predicted and rated

Conclusion

- Mutations in monitoring data are key failure-related symptoms, and mutation extraction is meaningful for removing failure-irrelevant and noisy data

- Similarity measurement can take advantage of both classification algorithms and anomaly detection algorithms to capture failures more comprehensively

- Failure rating and progressive diagnosis provide fine-grained failure status to operators to help them handle failures more accurately

2024 SRCX

THANK YOU

Samsung R&D Institute China Xi'an(SRCX) established in 2013 and located in Xi'an High-tech Zone, is the only cutting-edge technology R&D institute of Samsung Electronics in western China.