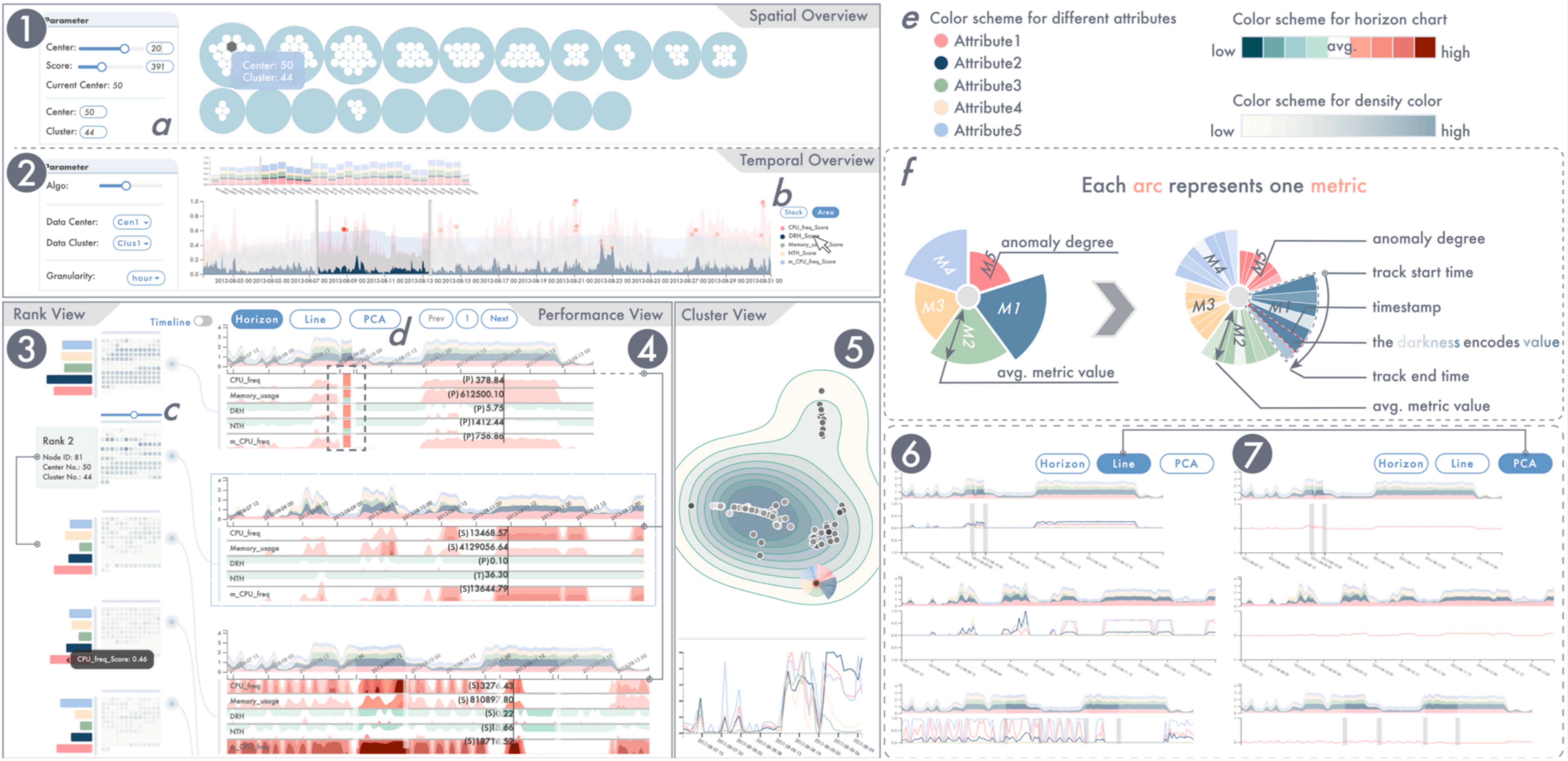
CloudDet: Interactive Visual Analysis of Anomalous Performances in Cloud Computing Systems

Ke Xu, Yun Wang, Leni Yang, Yifang Wang, Bo Qiao, Si Qin, Yong Xu, Haidong Zhang, Huamin Qu
IEEE Transactions on Visualization and Computer Graphics, 2019

THE UNIVERSITY OF BRITISH COLUMBIA

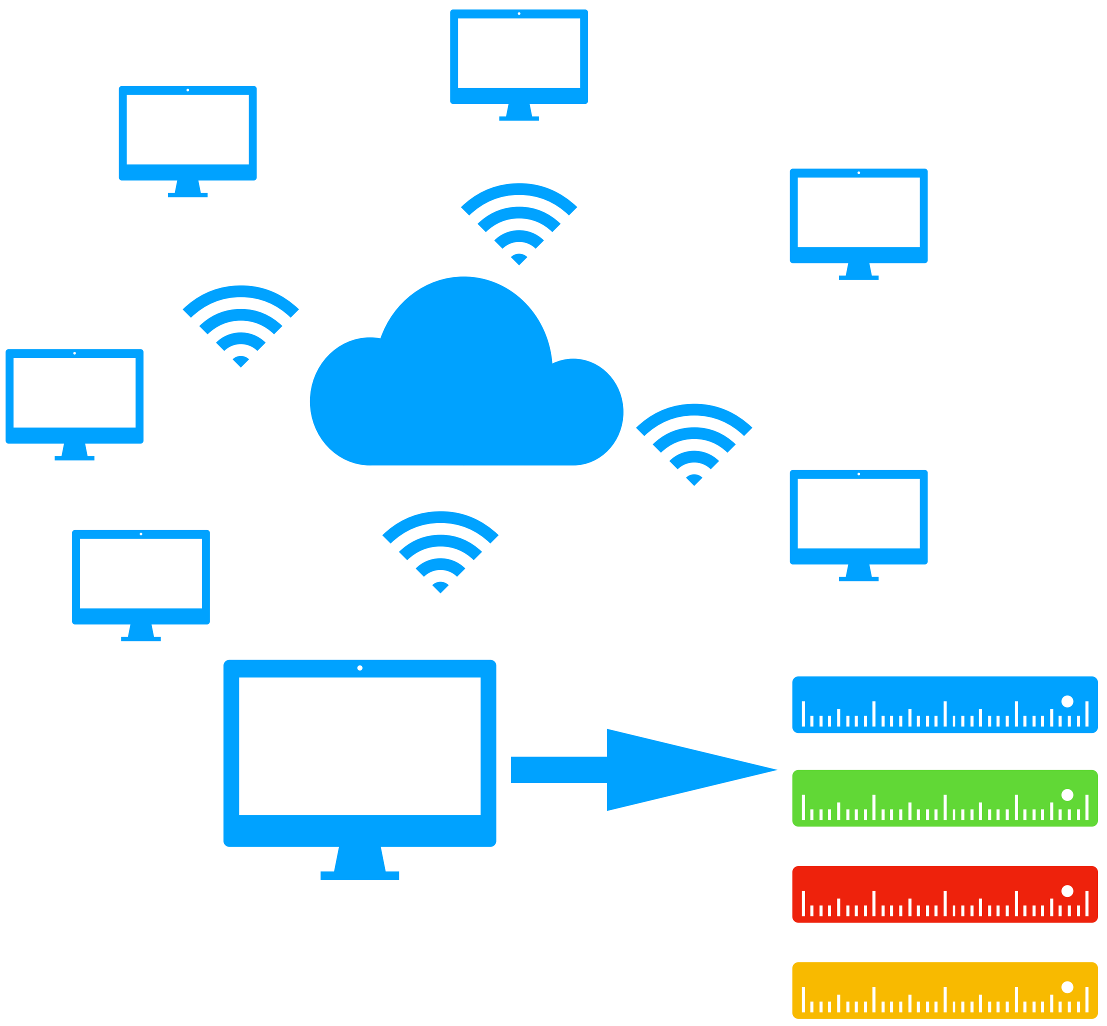
Amirhossein Abbasi
Nov 2019

CloudDet


Spatial Overview
Temporal Overview
Rank View
Performance View
Cluster View
Color scheme for different attributes
Attribute1
Attribute2
Attribute3
Attribute4
Attribute5
Color scheme for horizon chart
low
avg.
high
Color scheme for density color
low
high
Each arc represents one metric
anomaly degree
avg. metric value
track start time
timestamp
the darkness encodes value
track end time
Horizon
Line
PCA

Motivation

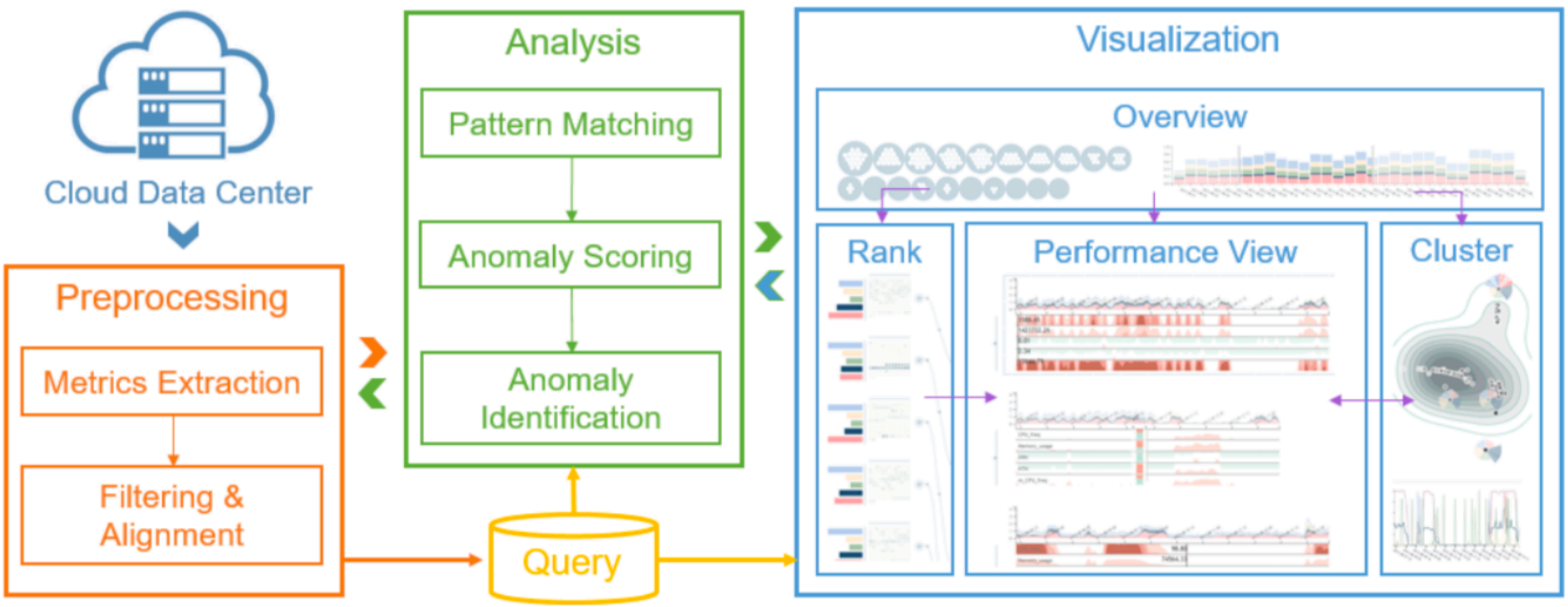
- Monitoring nodes instead of monitoring applications
- Too many **false positives**, **scale** problem.
- Visualization of anomalies: **Intuitiveness**, **interaction.**
- **Research Contribution: Detection system, Visualization, Evaluation**



Visualization Challenges

- **Scale**: Trade-off between system scalability and level-of-detail(LoD)
- **Multi-dimensionality**: Temporal patterns, Relation between metrics
- Boundary **normal/abnormal**

System Overview



What is abnormal and what is not?

How to detect?

Mathematics!

$$P(f_{k/N}) = ||X(f_{k/N})||^2, \;\; k = 0, 1 \ldots \lceil \frac{N-1}{2} \rceil, \tag{1}$$

$$ACF(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} d(\tau) \cdot d(n+\tau), \;\; \tau = 0, 1 \ldots \lceil \frac{N-1}{2} \rceil. \tag{2}$$

$$d_n = S_n + Tr_n + R_n, \;\; n = 1, 2 \ldots N, \tag{3}$$

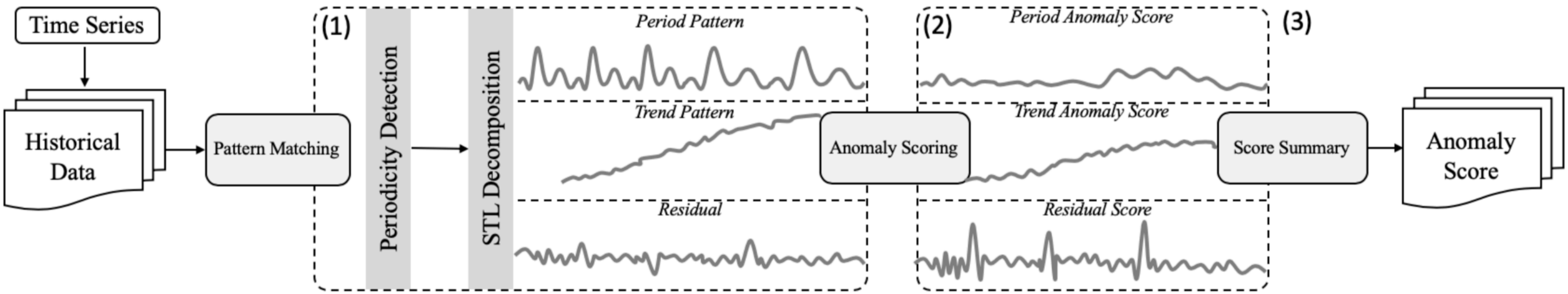
$$AS_{periodic} = min\left(\frac{|T_n - T_{n-1}|}{T_{n-1}}, 1\right), \tag{4}$$

$$AS_{trend} = min\left(\left|\frac{K_n - K_{n-1}}{K_{n-1}}\right|, 1\right), \tag{5}$$

$$AS_{spike} = min\left(\left|\frac{R_n - \mu_{n-1} - 3\sigma_{n-1}}{3\sigma_{n-1}}\right|, 1\right), \tag{6}$$

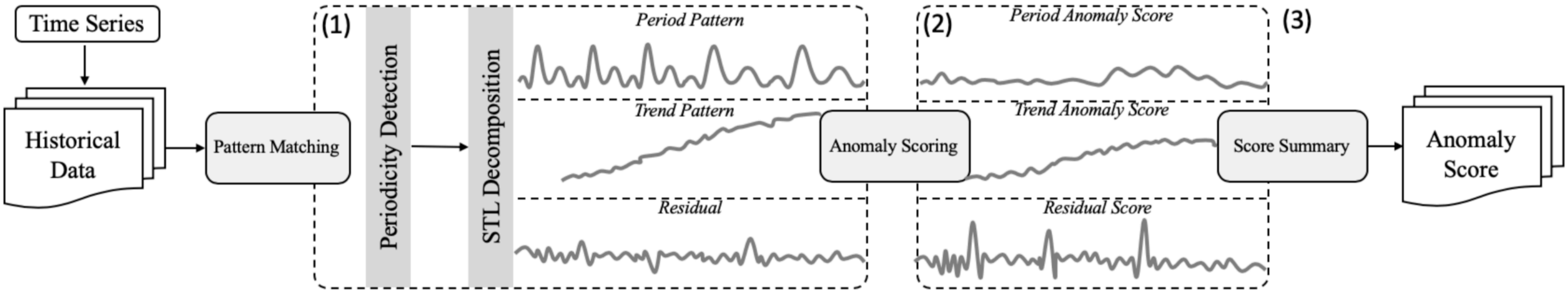
$$AS = f(AS_{periodic}, AS_{trend}, AS_{spike}), \tag{7}$$

Algorithm Flow



Algorithm Flow

Only utilizes the most recent data



Anomalies have patterns

Design Tasks

- **T1** Overview of anomalies for data query
- **T2** Ranking suspicious nodes dynamically
- **T3** Browse data flexibly
- **T4** Facilitate anomaly detection
- **T5** Similarities of nodes

Visualization

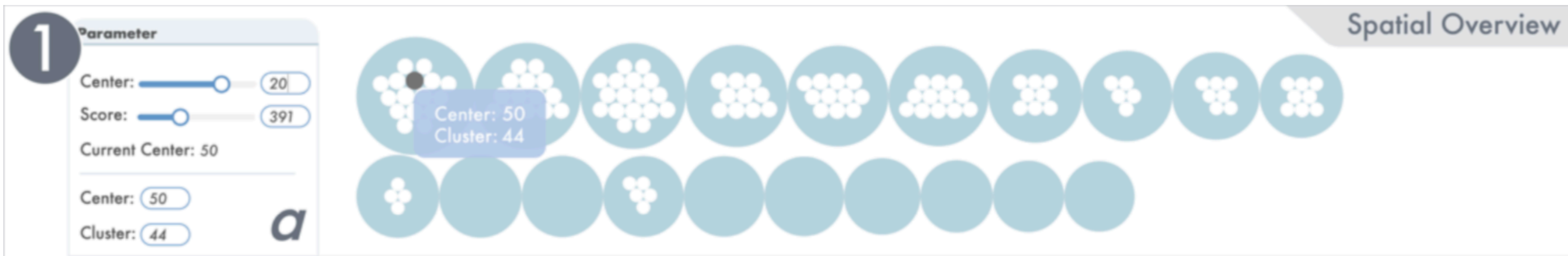
Encoding Protocol

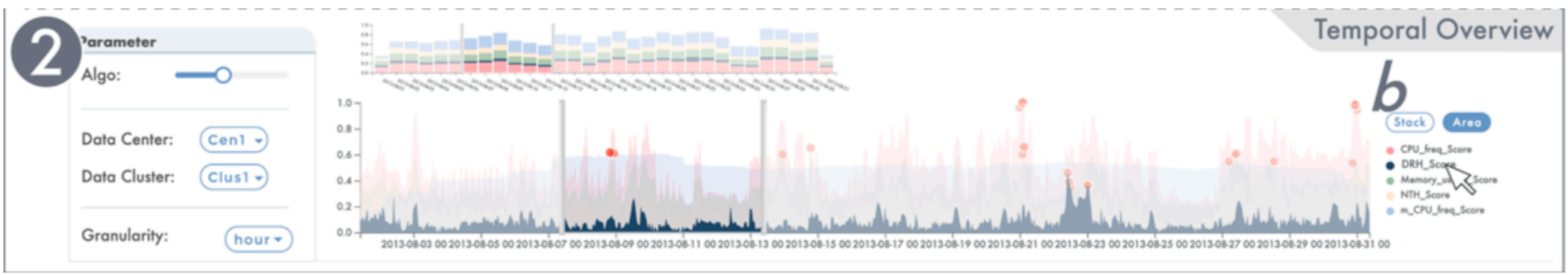
Global Categorical Colors: performance metrics (CPU Frequency, Memory Usage,…)

Linear Color Scheme: Anomaly score

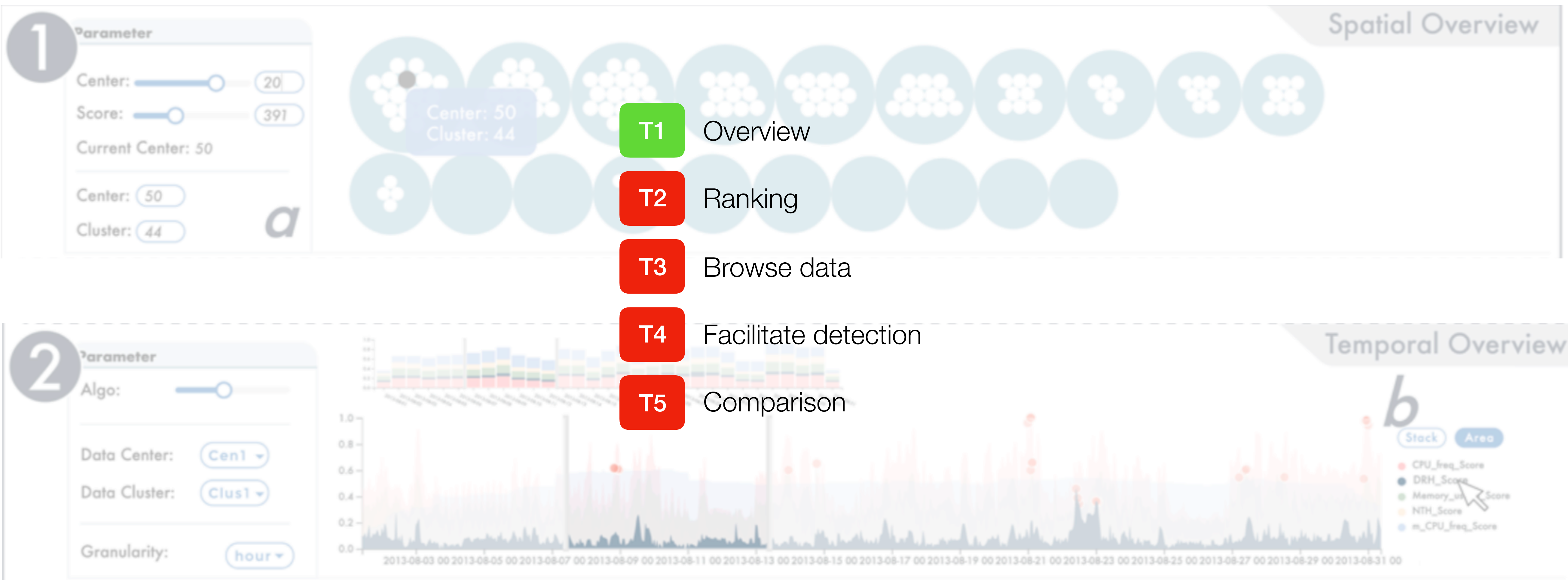
Diverging Color Scheme: Difference of performance metrics to average

Spatial and Temporal Views





Spatial and Temporal Views



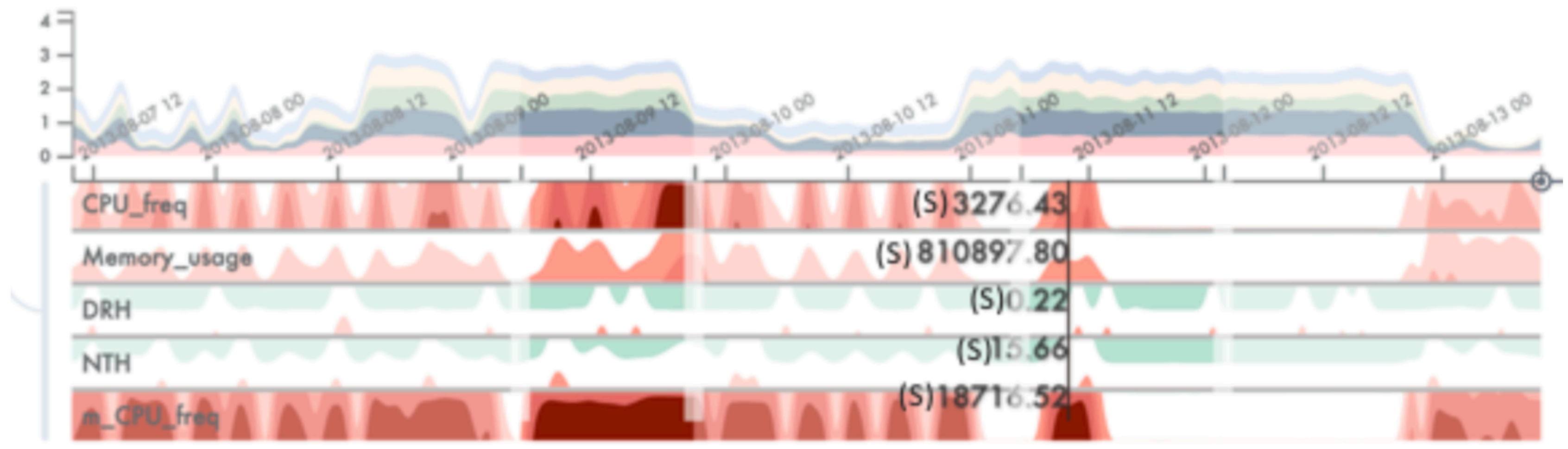
- T1 Overview
- T2 Ranking
- T3 Browse data
- T4 Facilitate detection
- T5 Comparison

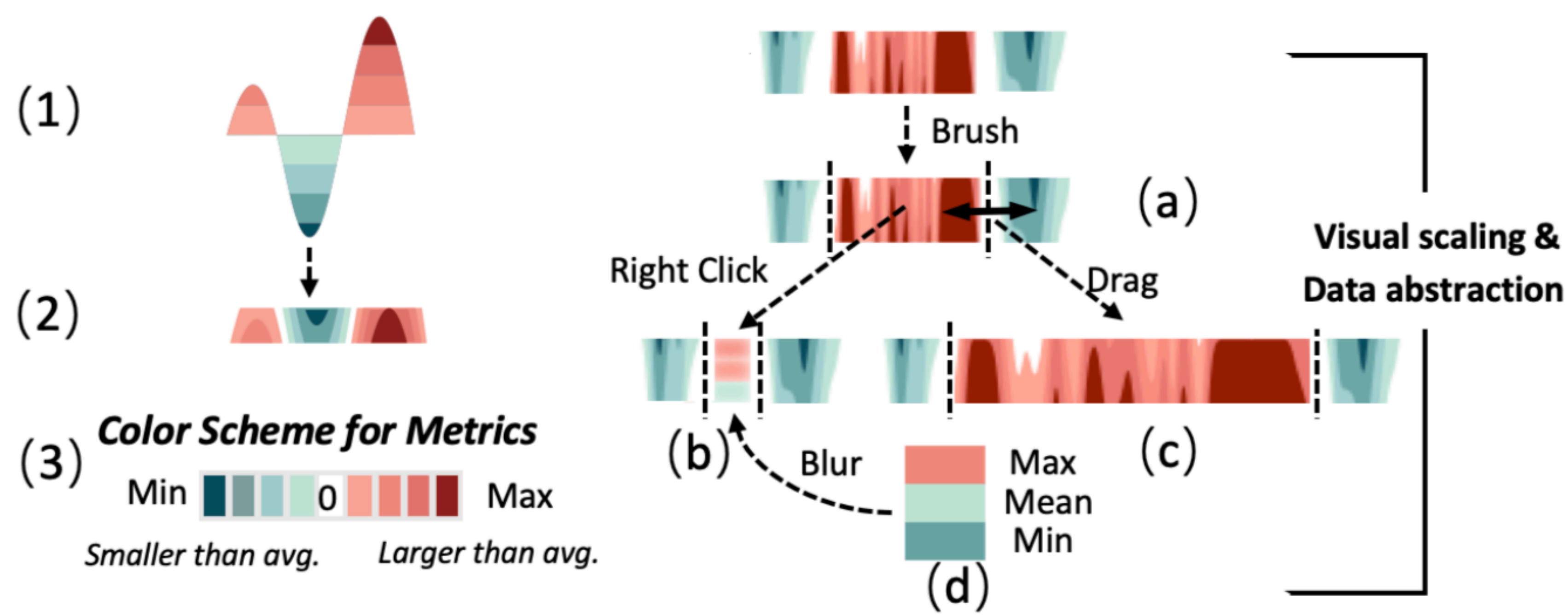
Rank and Performance View

Horizon Chart

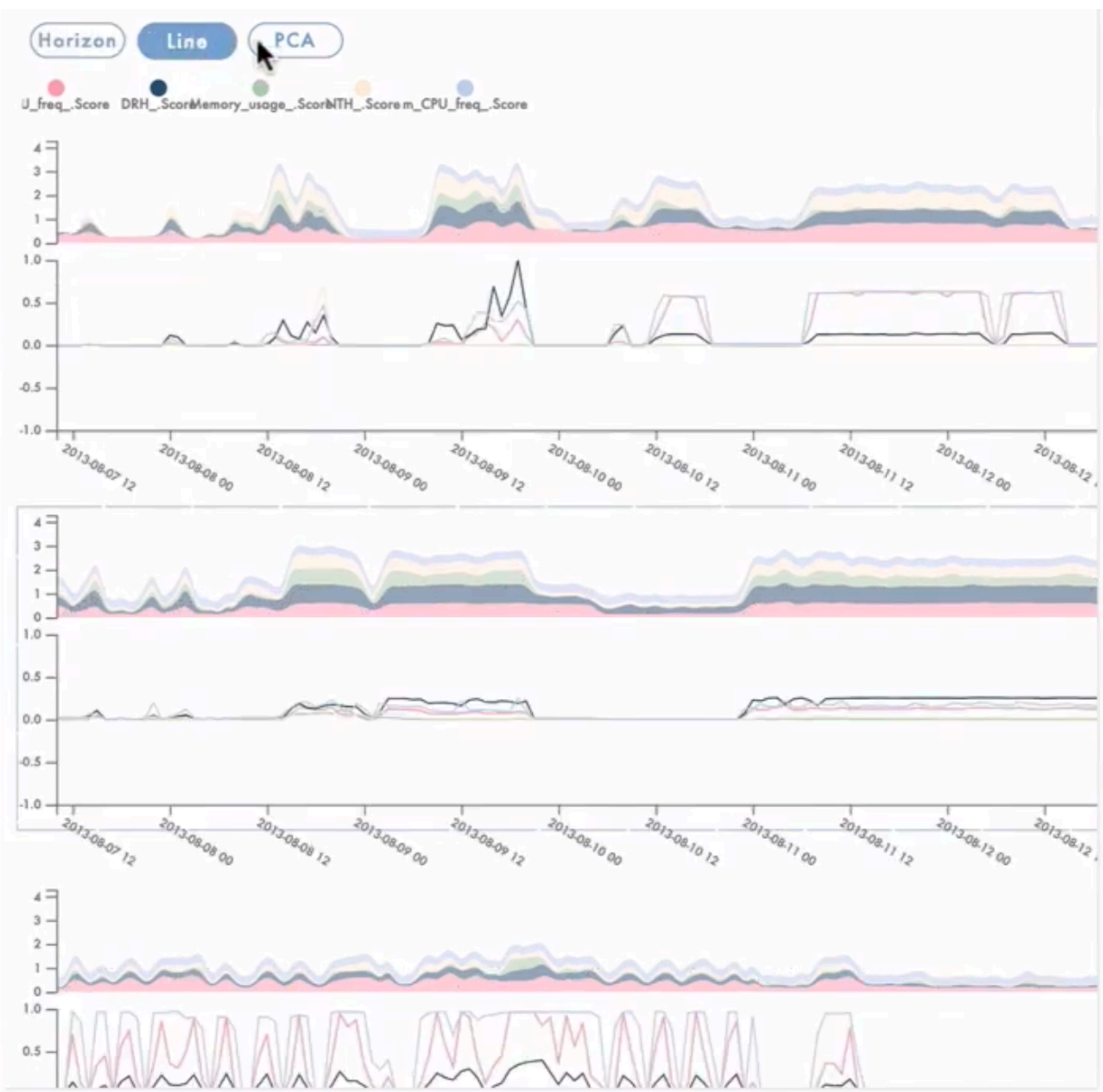
Interactions:

- Brushing
- Collapsing
- Stretching





Line mode



- Each line for one metric
- More conventional
- Normalize data to [-1,1]

PCA mode

- Project a multivariate data to a one-dimensional time-series data
- Major Trend

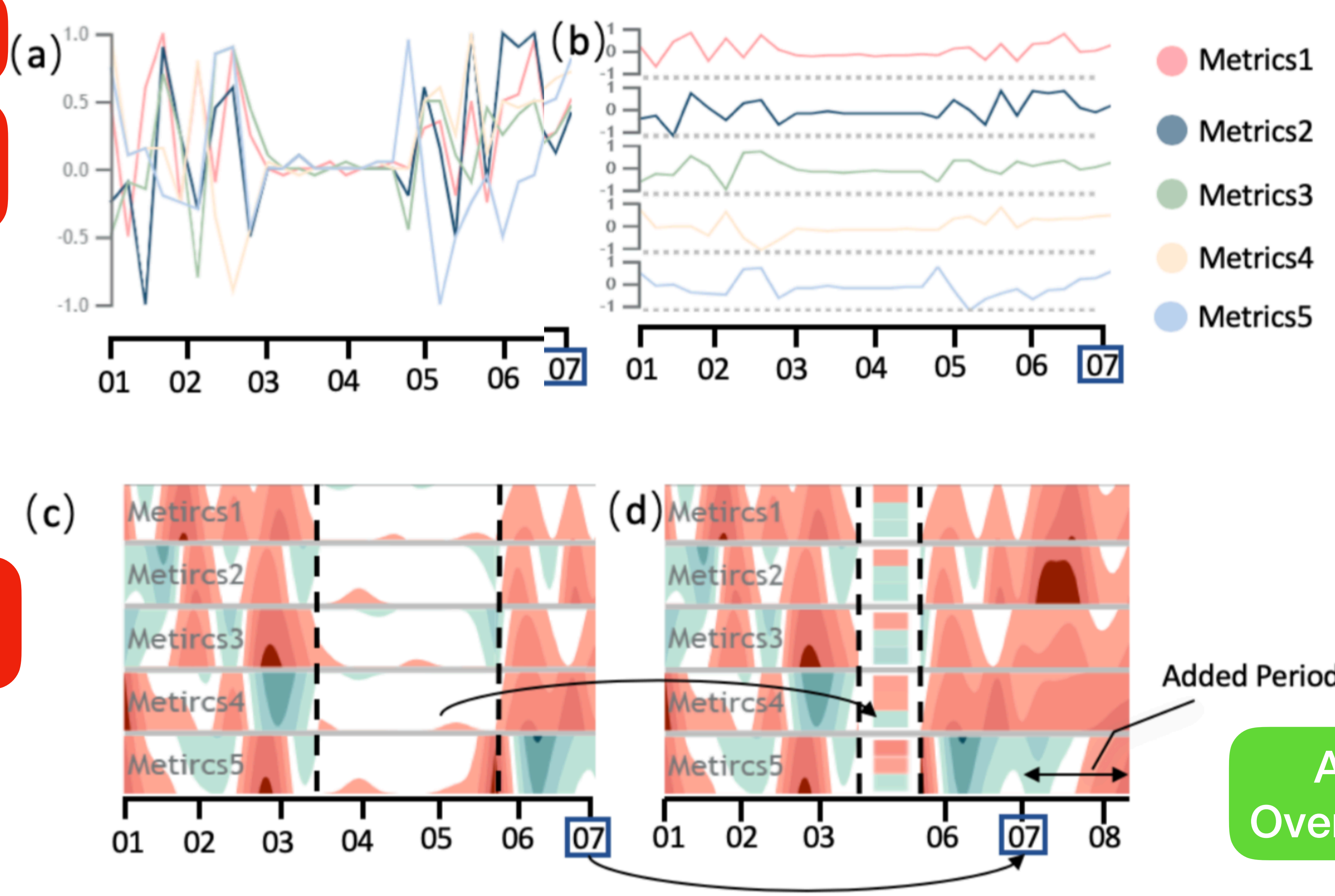
Alternative Designs

Visual Clutter

Measurement Unit

Showing Trends

Scaling in time domain



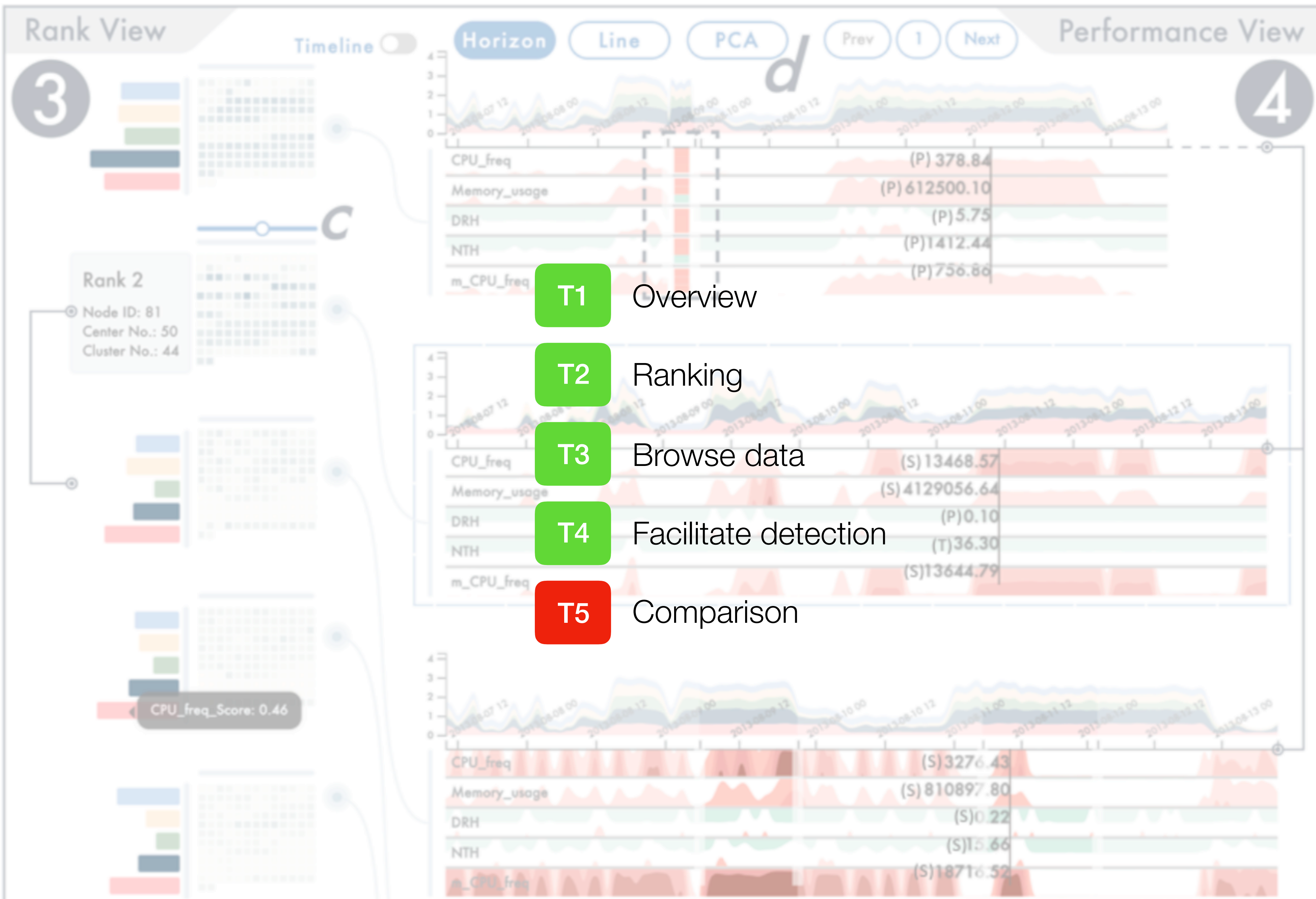
Consumes Space

Showing Trends

Analyzing Overal Changes

Non-linear time scaling

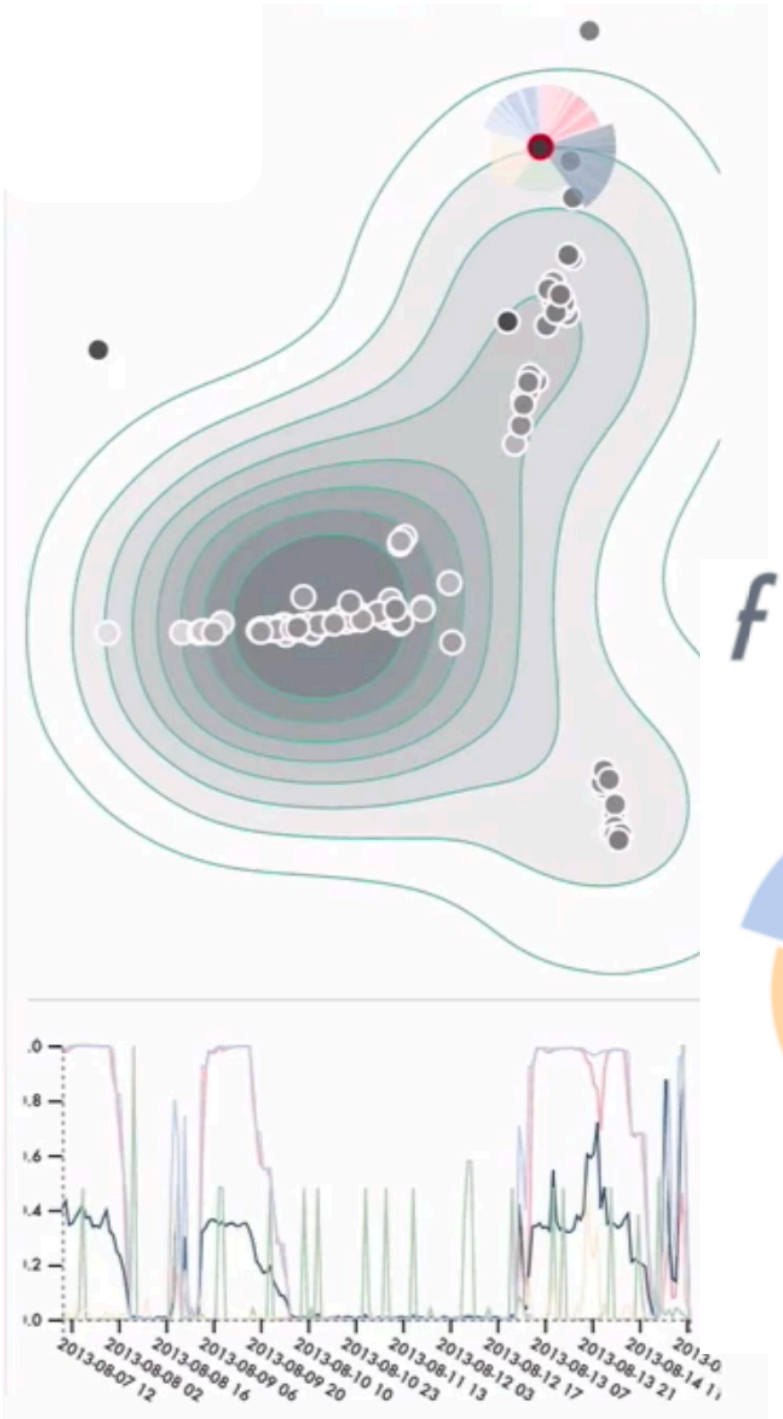
Rank and Performance View

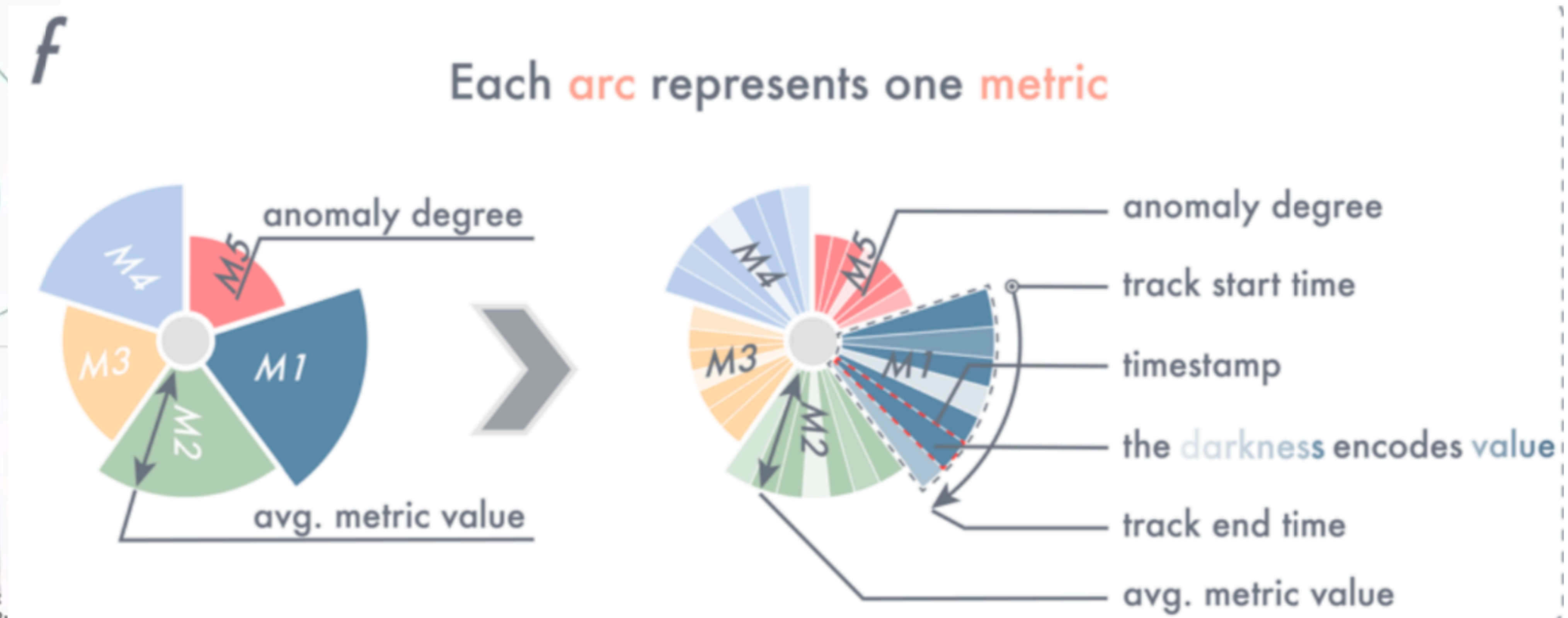


- T1 Overview
- T2 Ranking
- T3 Browse data
- T4 Facilitate detection
- T5 Comparison

The Cluster View

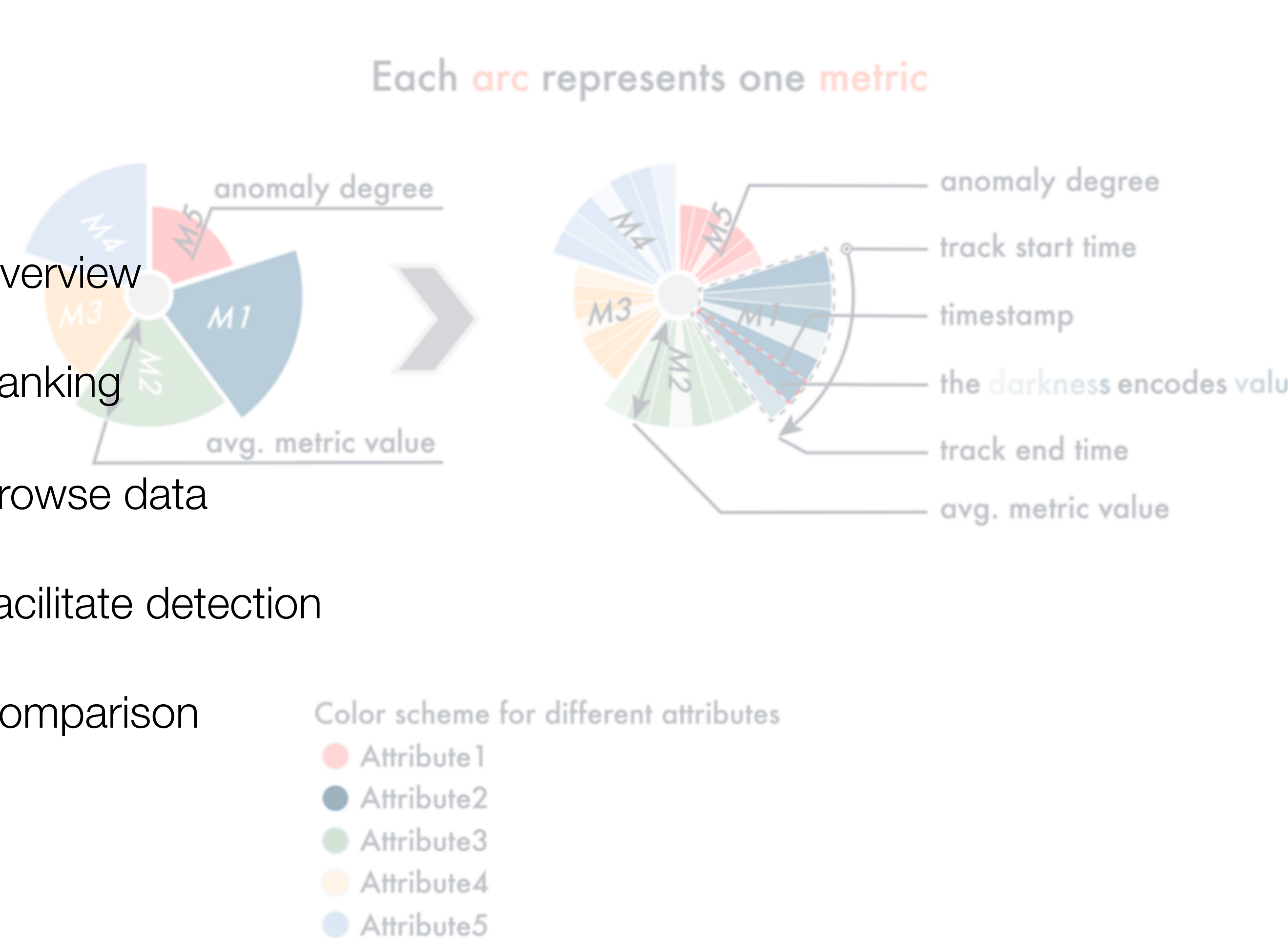
- Dimensionality reduction.
- Another perspective for anomaly diagnosis
    - White contour: most probably anomaly
    - Gray contour:normal.



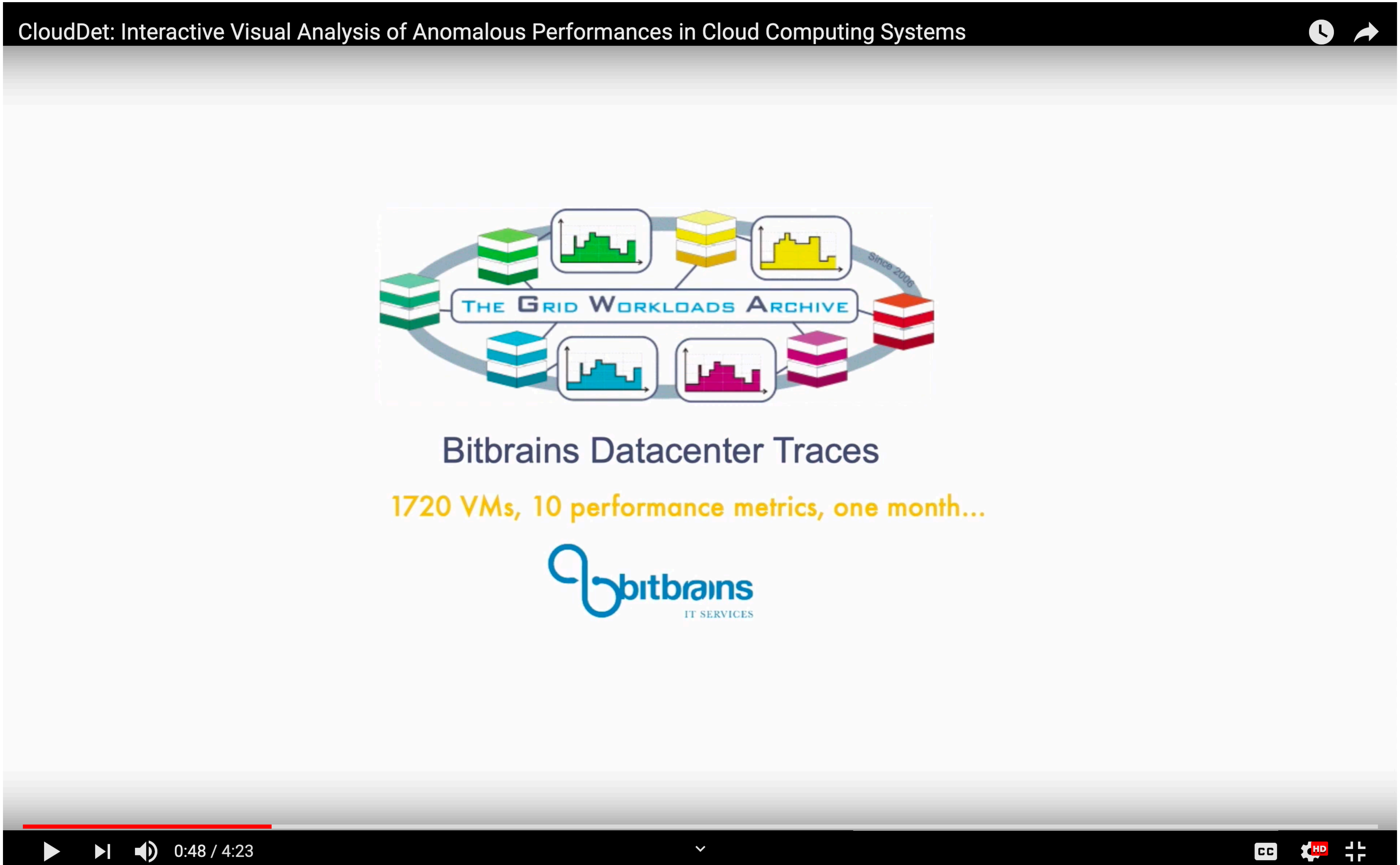


The Cluster View

- T1 Overview
- T2 Ranking
- T3 Browse data
- T4 Facilitate detection
- T5 Comparison



Official Video (1:05 min)



What-Why-How Summary

What

Why

How

Multi-variate time-series quantitative performance data from compute nodes.

- Anomaly Ranking
- Anomaly inspection
- Anomaly Clustering

- Colors and brightness
- horizontal and line chart
- Special glyphs
- Spatial positions of nodes and charts.
- Interactivity: Scrolling, Brushing, and setting parameters

What-Why-How Summary

What	Why	How
Multi-variate time-series quantitative performance data from compute nodes.	• Anomaly Ranking • Anomaly inspection • Anomaly Clustering	• Colors and brightness • horizontal and line chart • Special glyphs • Spatial positions of nodes and charts. • Interactivity: Scrolling, Brushing, and setting parameters

What-Why-How Summary

What

Why

How

Multi-variate time-series quantitative performance data from compute nodes.

- Anomaly Ranking
- Anomaly inspection
- Anomaly Clustering

- Colors and brightness
- horizontal and line chart
- Special glyphs
- Spatial positions of nodes and charts.
- Interactivity: Scrolling, Brushing, and Querying

Scalability

What

Multi-variate time-series quantitative performance data from compute nodes.

Why

- Anomaly Ranking
- Anomaly inspection
- Anomaly Clustering

How

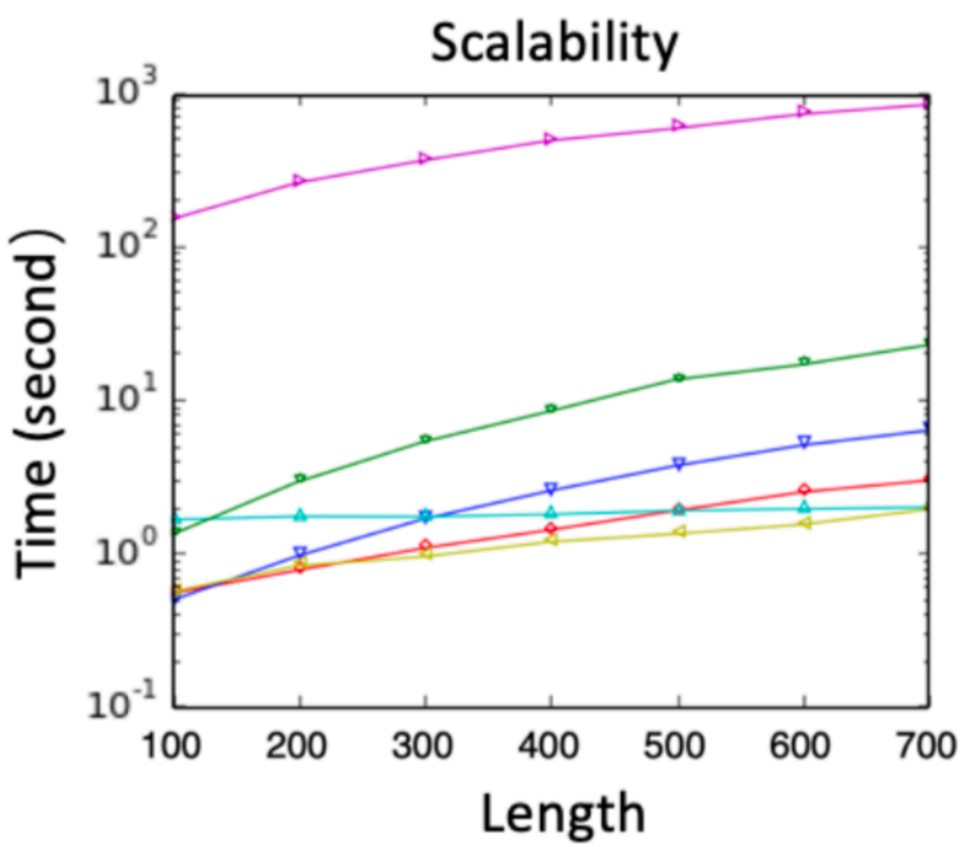
- Colors and brightness
- horizontal and line chart
- Special glyph
- Spatial positions of nodes and charts.
- Interactivity: Scrolling, Brushing, and setting parameters

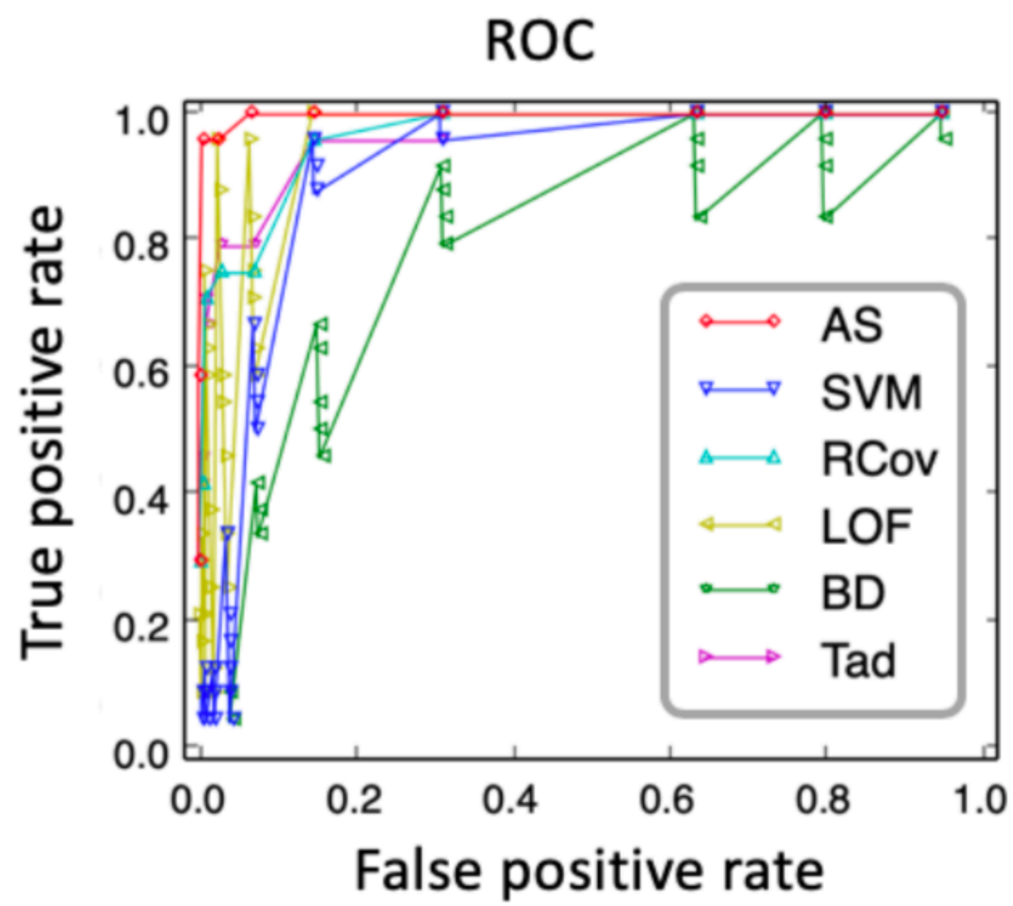
Scale

Very Scalable: scale **linearly** with time-series input data size

Evaluation

Quantitative Evaluation





Case Studies

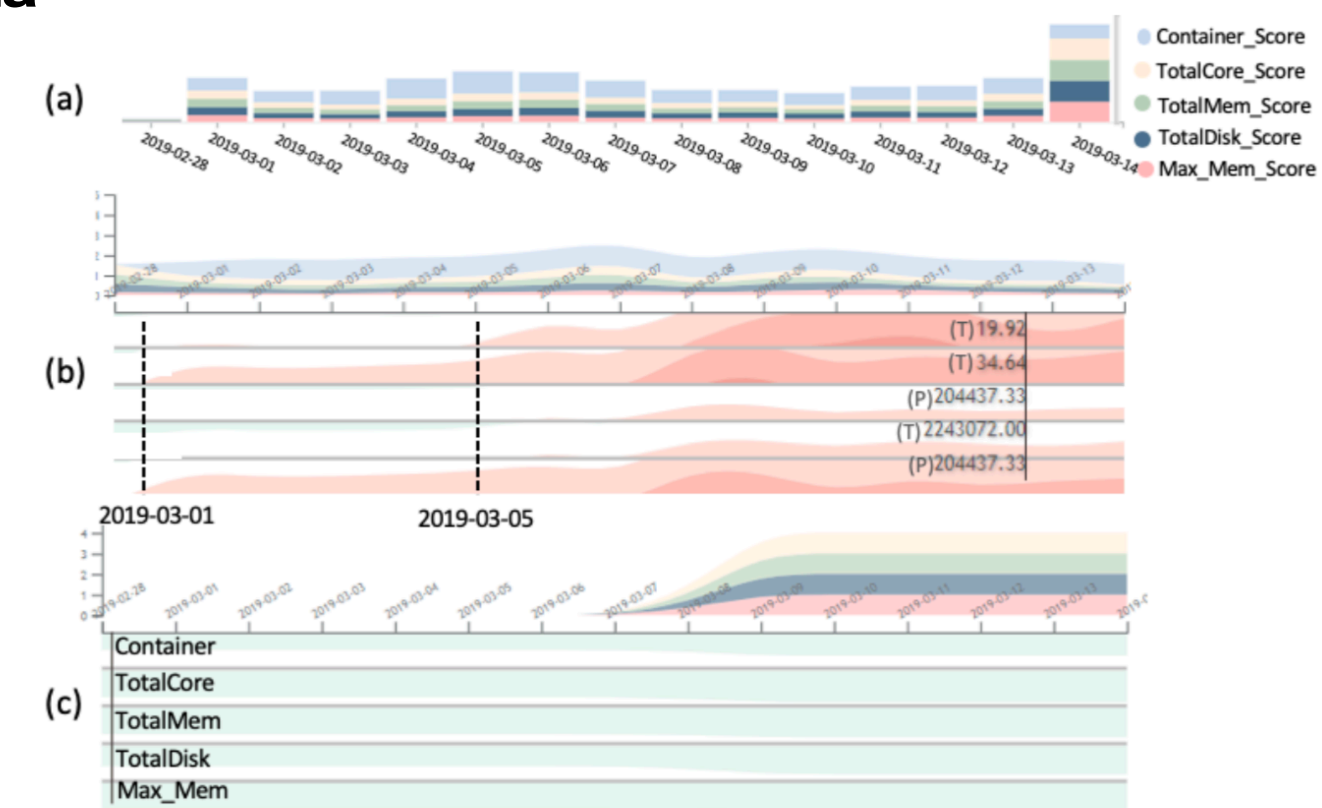
Case Study 1: Bitbrains Datacenter Traces

500 VMs, One month

Case Study 2: Live Cloud System Data

1,000,000 nodes, Two weeks

[100 data centers with 20 data clusters with 500 nodes each]



User Feedback

- Automated Anomaly Detection: **Trust** in algorithm,
- System: **Useful** and **User-friendly**, **Consistent**, too **comprehensive** and **Overwhelming, Need Tutorial**
- Visualization and Interaction: **Helpful**, **new perspective** for overall trend, **clear comparison**, Confess that they use chaotic line charts before.

Critique

Positive

- Alternative designs
- Super-scalable
- Perfect evaluation
- Very Accurate
- Special Glyphs

Negative

- Better to use non-diverging colors for horizon charts.
- Minor occlusion in the clustering.
- Make use of global colors in horizontal chart.
- Bad way for Assigning the ranks to performance.
- Empty clusters in spatial overview.
- Limitation: Just consider recent data and one metric.
- Limitation: Don't discuss why using those performance metrics for anomaly.

Question?