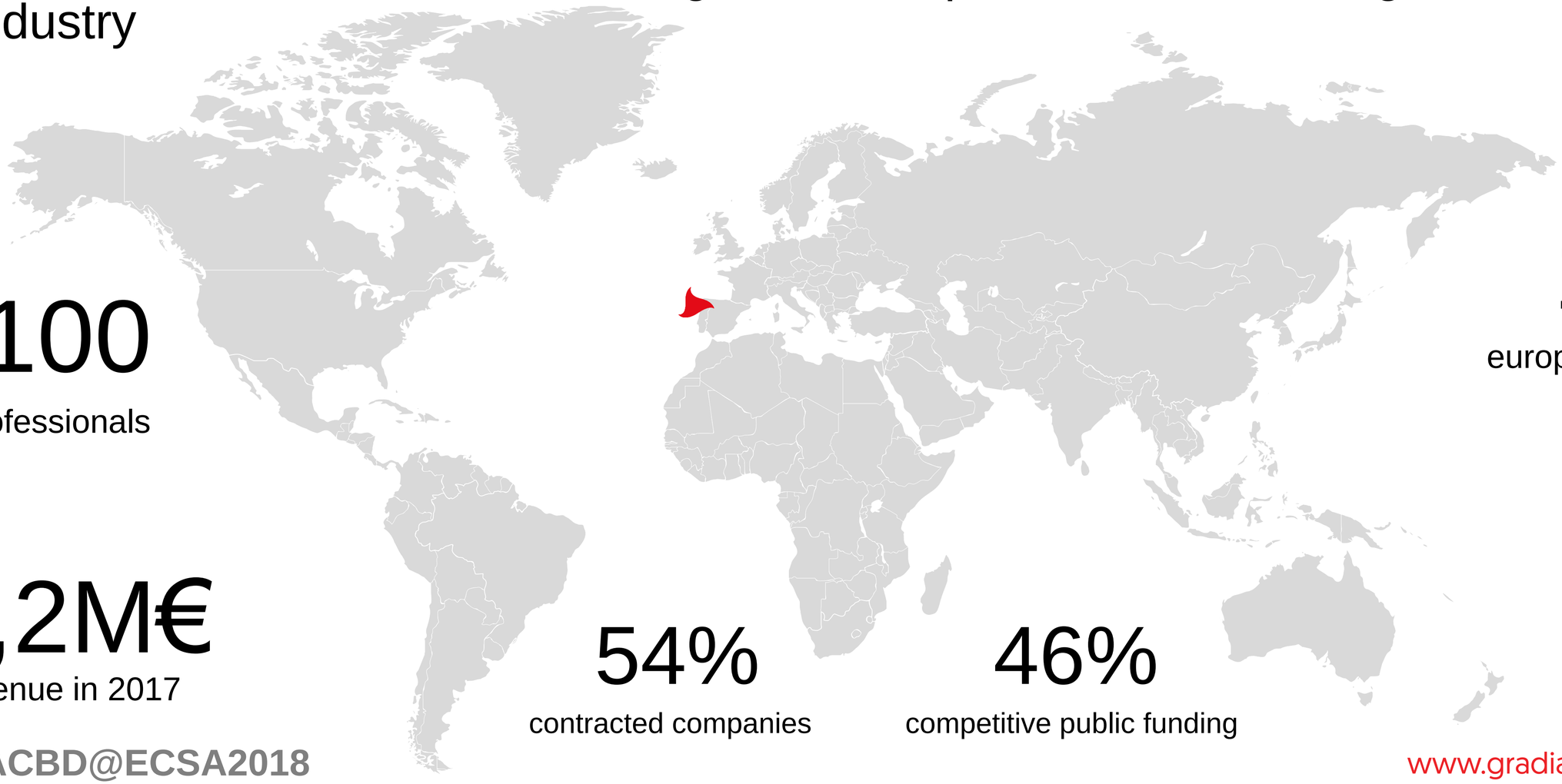
Large Scale Anomaly Detection in Data Center Logs and Metrics

Rafael Martínez

(+34) 986 120 430 | gradiant@gradiant.org | www.gradiant.org

Gradiant, ICT technology centre in Spain

Since 2008, focused on technological development and knowledge transfer to industry



+100
professionals

12
european projects

5,2M€
revenue in 2017

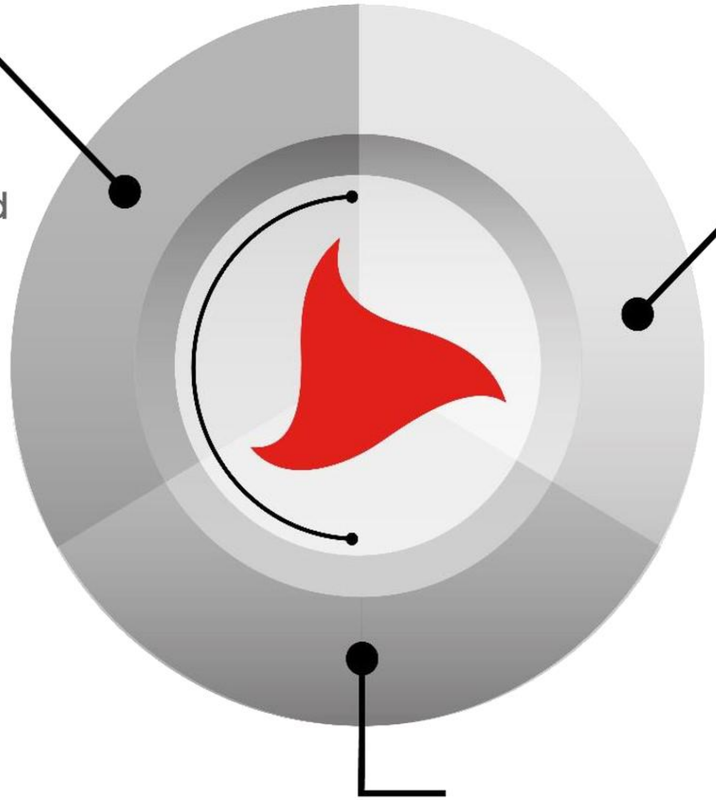
54%
contracted companies

46%
competitive public funding

SACBD@ECSA2018

www.gradiant.org

Focus on…



Connectivity

communication systems and internet of things for data transmission

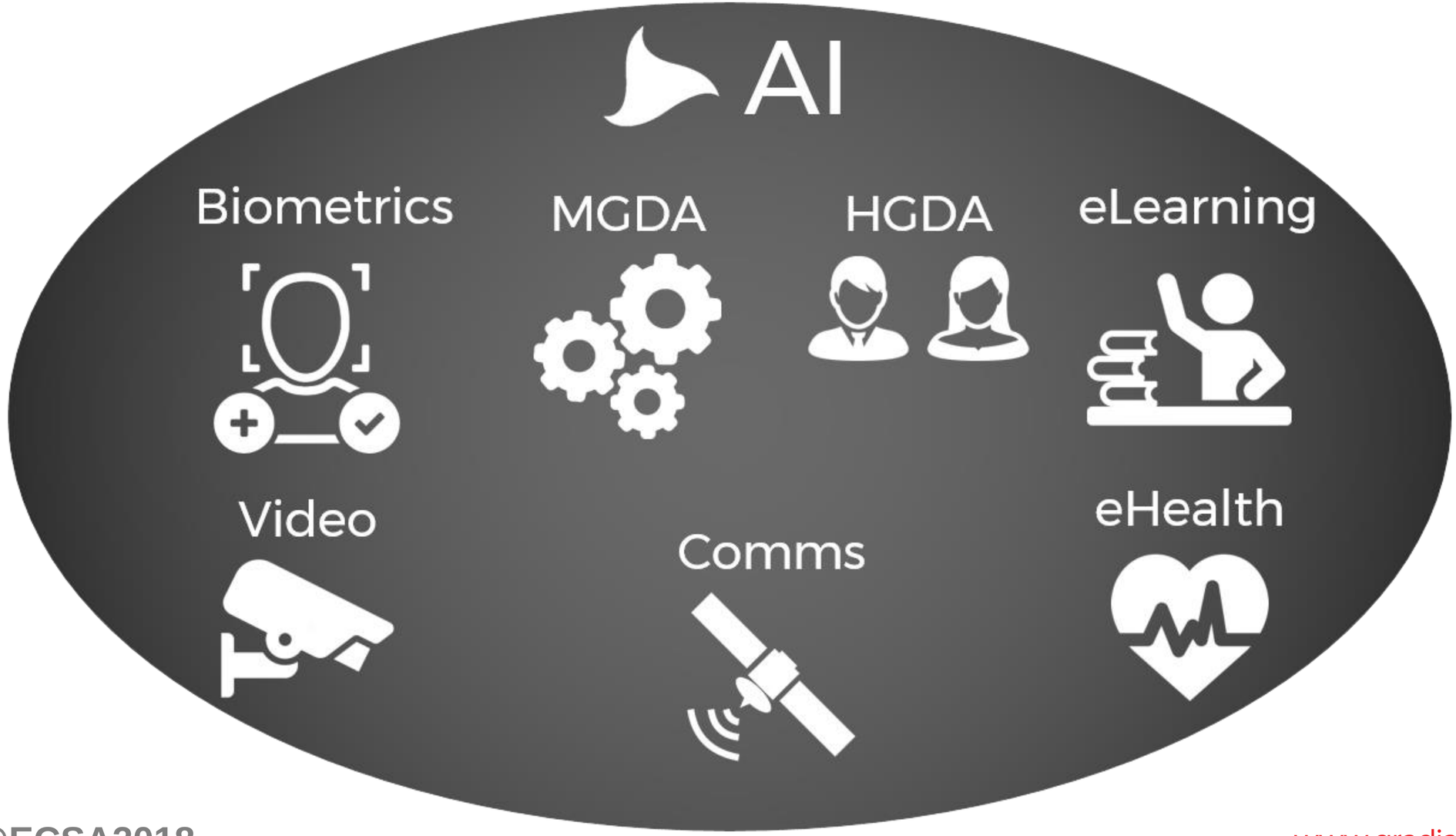
Intelligence

infrastructure and algorithms for extracting value from data, converting them into useful and actionable information

Security

protection of data and information systems, and protection of privacy

Focus on… Connectivity · Intelligence · Security



Large Scale Anomaly Detection in Data Center Logs and Metrics

Problem statement

- Large amounts of data
- Data flowing 24×7×365
- Logs/events & metrics

- Goal: Keep systems OK

Anomaly Detection

ML to the rescue!

Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection

1. Anomaly Detection in Metrics

2. Anomaly Detection in Logs

Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection

1. Anomaly Detection in Metrics

2. Anomaly Detection in Logs

Requirements

- Run on **endless data streams**
- Provide results in **real-time**

Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection

1. Anomaly Detection in Metrics
   - Discord Discovery
2. Anomaly Detection in Logs
   - LogScore

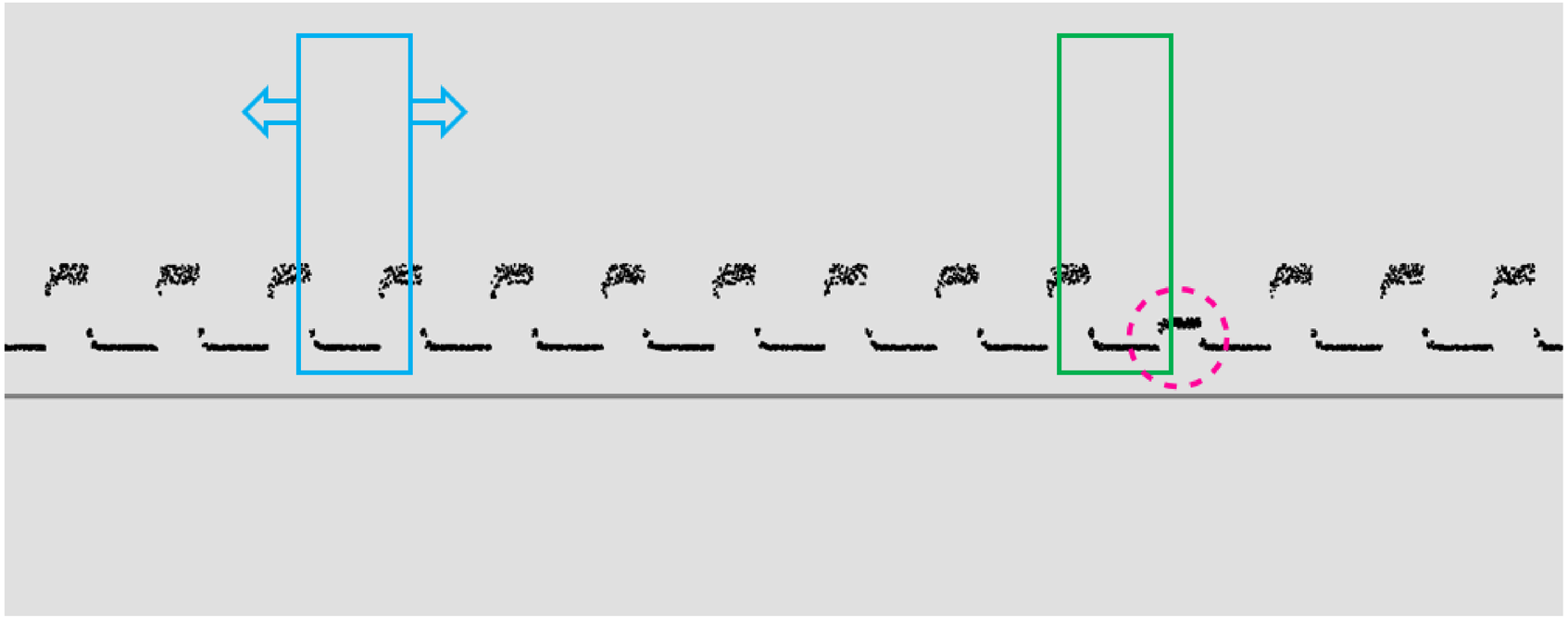
Requirements

- Run on **endless data streams**
- Provide results in **real-time**

Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Metrics

Discord Discovery



Large Scale Anomaly Detection in Data Center Logs and Metrics

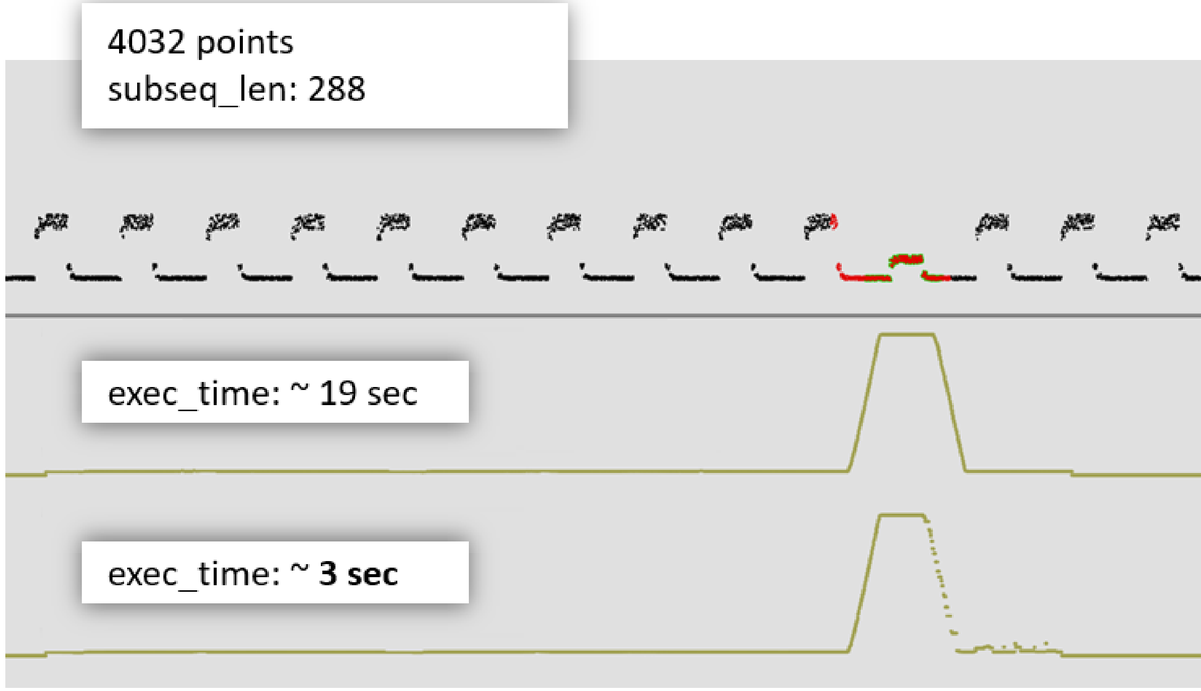
Anomaly Detection in Metrics

Discord Discovery

Improvements on the original batch algorithm

- Heuristics to reduce $O(n^2)$ towards $O(n)$
  - Early abandon
  - Randomization
  - Denoise filtering

```
1   Function [dist, loc ]= Heuristic_Search(T, n, Outer, Inner )
2    best_so_far_dist = 0
3    best_so_far_loc = NaN
4
5    For Each p in T ordered by heuristic Outer     // Begin Outer Loop
6      nearest_neighbor_dist = infinity
7      For Each q in T ordered by random Inner // Begin Inner Loop
8        IF | p - q | ≥ n                          // non-self match?
9          IF Dist (t_p,...,t_p+n-1, t_q,...,t_q+n-1) < best_so_far_dist
10           Break                                 // Break out of Inner Loop
11         End
12         IF Dist (t_p,...,t_p+n-1, t_q,...,t_q+n-1) < nearest_neighbor_dist
13           nearest_neighbor_dist = Dist (t_p,...,t_p+n-1, t_q,...,t_q+n-1)
14         End
15       End                                       // End non-self match test
16     End                                         // End Inner Loop
17     IF nearest_neighbor_dist > best_so_far_dist
18       best_so_far_dist = nearest_neighbor_dist
19       best_so_far_loc  = p
20     End
21   End                                           // End Outer Loop
22   Return[ best_so_far_dist, best_so_far_loc ]
```



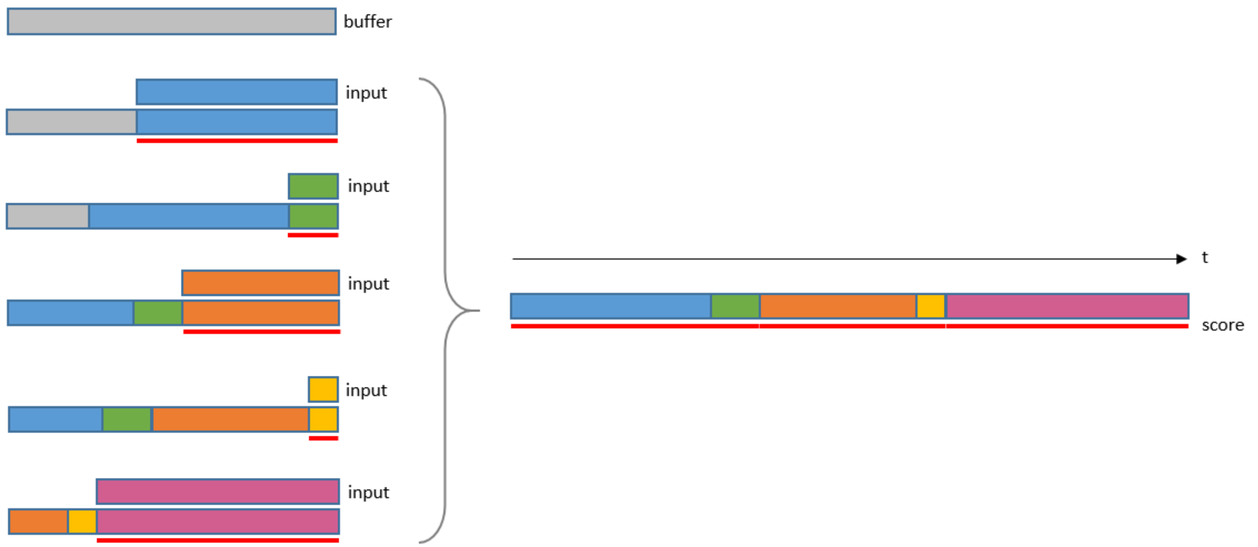
Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Metrics

Discord Discovery

Improvements on the original batch algorithm

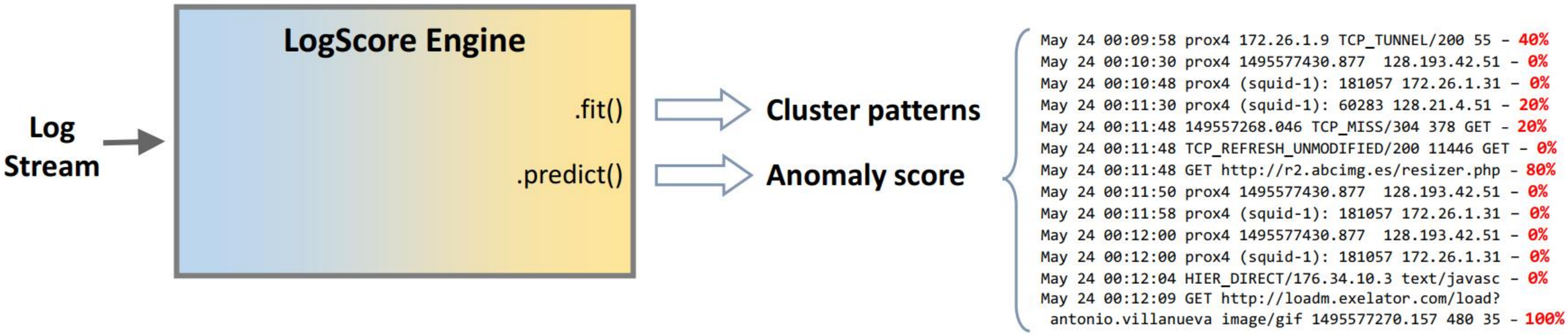
- Heuristics to reduce $O(n^2)$ towards $O(n)$
- Fully streaming operation
  - Supports input chunks of any size



Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Logs

LogScore



Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Logs

LogScore

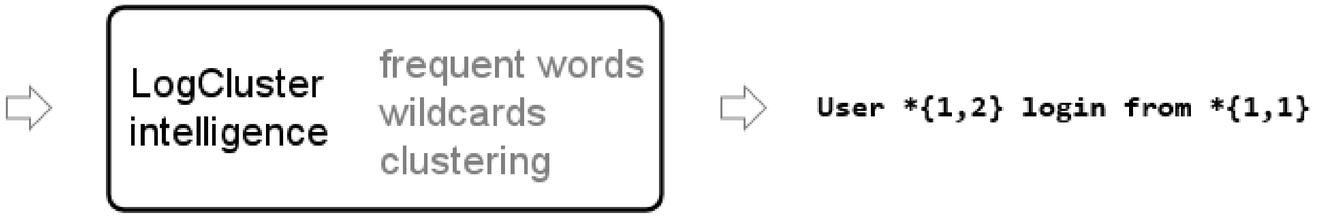
```
User bob login from 10.0.0.1
User alice login from 10.0.0.1
User jim login from 10.0.0.2
User Srv Admin login from 10.0.0.3
```

Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Logs

LogScore

```
User bob login from 10.0.0.1
User alice login from 10.0.0.1
User jim login from 10.0.0.2
User Srv Admin login from 10.0.0.3
```



No satisfactory results from single runs

- Too little representative clusters
- Too many outliers

Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Logs

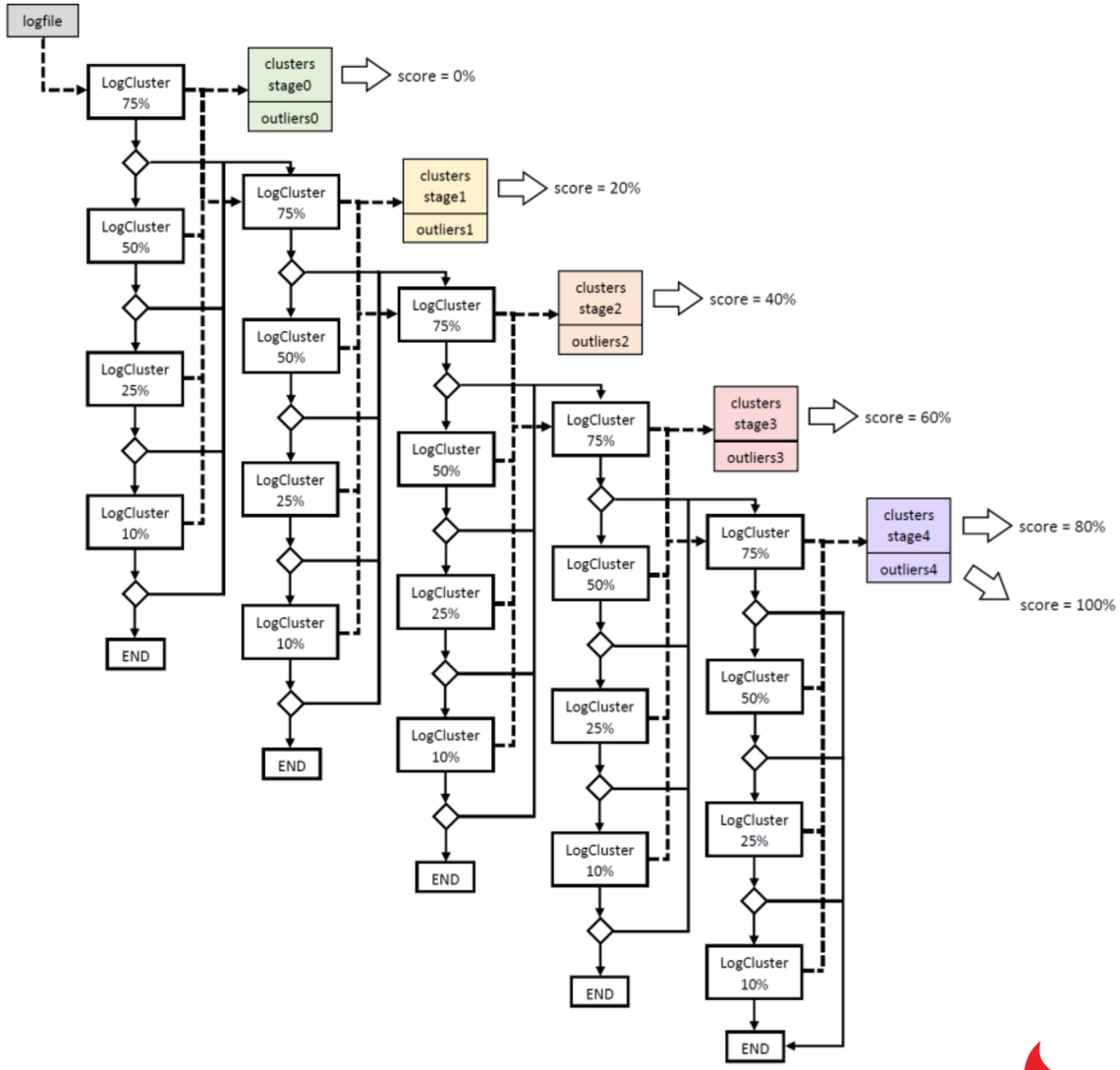
LogScore

spec = freq_words / (freq_words + wildcards)

high spec => more informative patterns
low spec => less outliers (bigger clusters)

trade-off

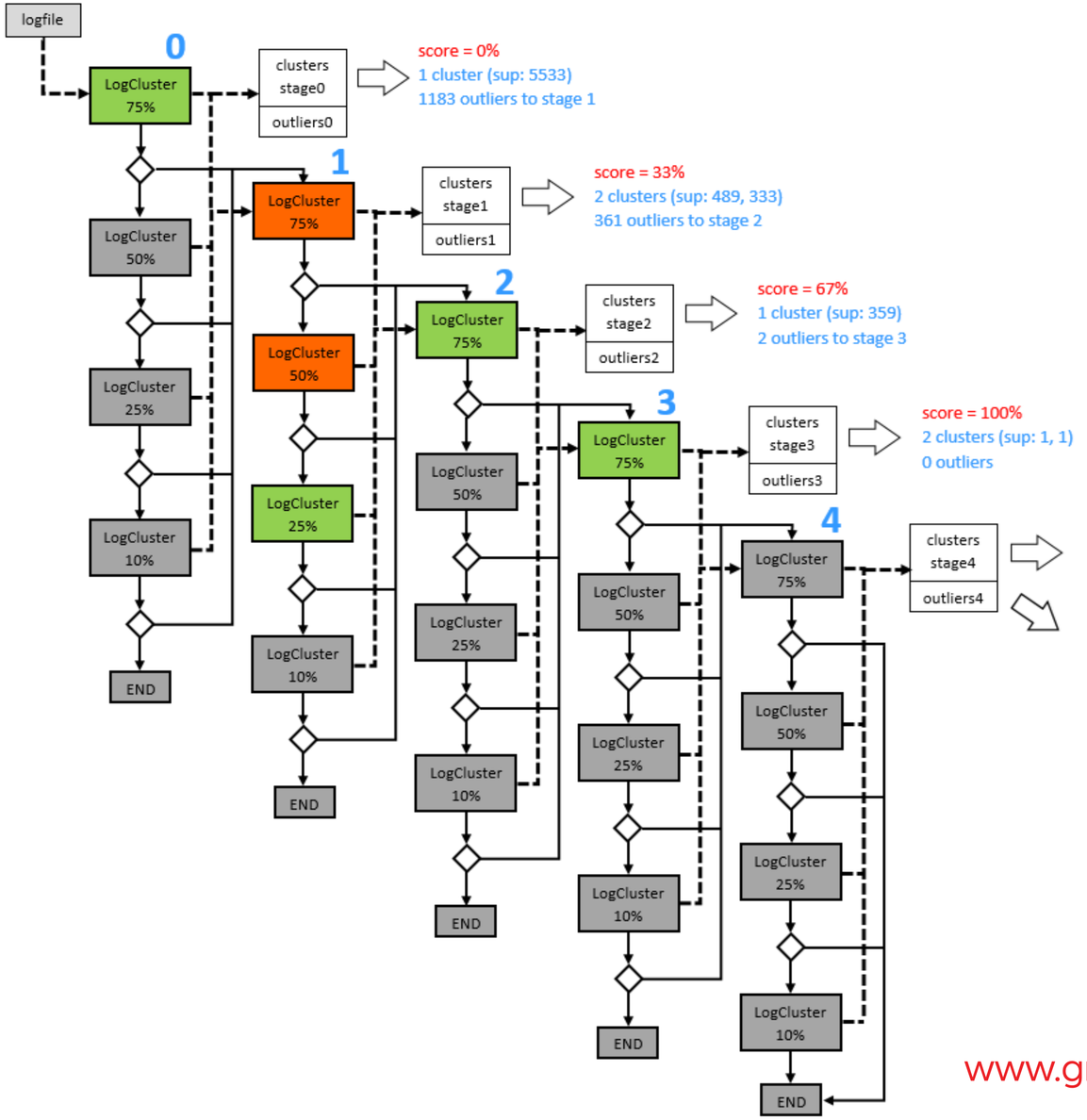
spec = 0.6



Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Logs

LogScore



Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection in Logs

LogScore


```
{
   "timestamp":"1498583167",
   "stage":"1",
   "rsupport":"25",
   "support":"295",
   "total_lines":"1183",
   "result":{
      "num_clusters":"2",
      "clusters":[
         {
            "cluster_id":"0",
            "support":"489",
            "pattern":"May 1 *{1,1} prox3 (squid-1): *{2,2} 172.26.15.36
                      TCP_MISS/200 *{1,1} CONNECT *{1,1} - *{1,1} -"
         },
         {
            "cluster_id":"1",
            "support":"333",
            "pattern":"May 1 *{1,1} prox3 (squid-1): *{2,2} 172.26.15.36
                      TCP_CLIENT_REFRESH_MISS/200 *{1,1} GET *{1,1} - *{2,2}"
         }
      ],
      "num_outliers":"361",
      "outliers":"lc3_201705-01-prox3_1.out"
   }
},
```

Large Scale Anomaly Detection in Data Center Logs and Metrics

Anomaly Detection

1. Anomaly Detection in Metrics
   - Discord Discovery

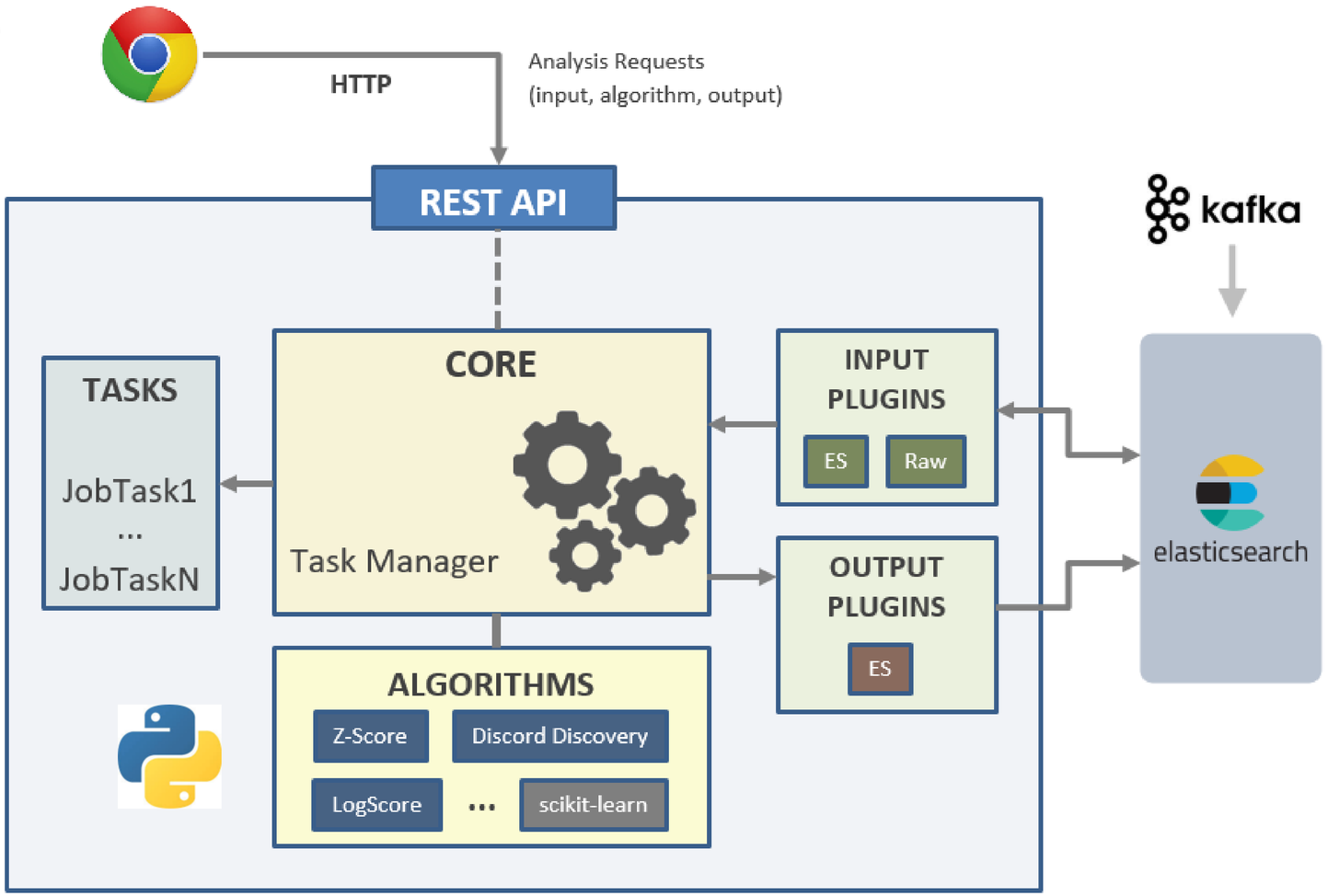
2. Anomaly Detection in Logs
   - LogScore

Both algorithms

- Learn ***normality*** from past data
- Unsupervised operation

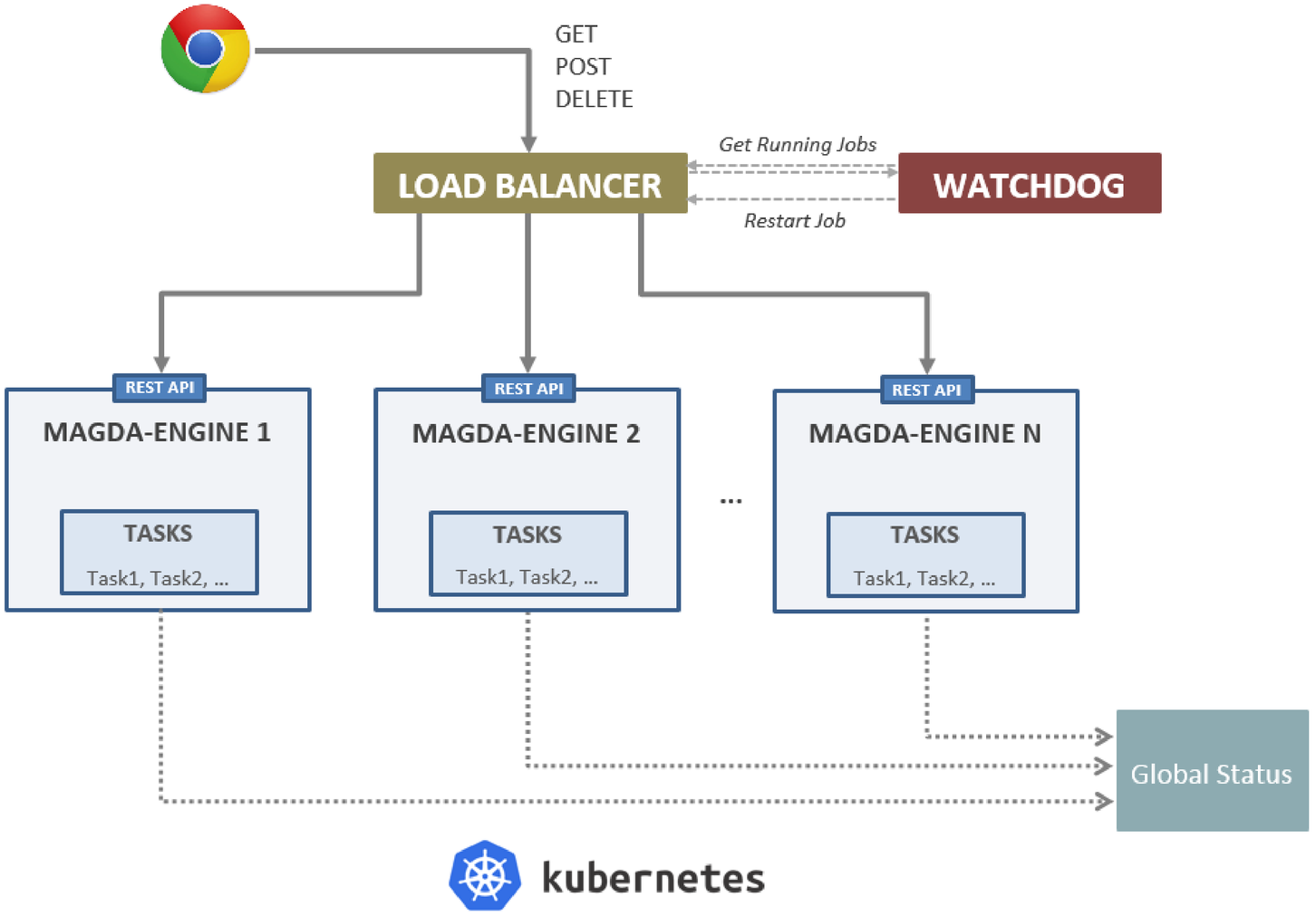
Large Scale Anomaly Detection in Data Center Logs and Metrics

Functional architecture



Large Scale Anomaly Detection in Data Center Logs and Metrics

Deployment



Large Scale Anomaly Detection in Data Center Logs and Metrics

Tests and results

- 4 months of data (52 GB, 121 machines, 306M log events)

- Not only detect strange events
- But also provide a summary suitable for human supervision

- Good perceived performance

Large Scale Anomaly Detection in Data Center Logs and Metrics

Conclusions

- High-value solution with hard requirements
- Fully scalable in large Big Data environments
- One company already pushing these technologies into the market

Thank you!

Rafael P. Martínez-Álvarez, PhD

rmartinez@gradiant.org

(+34) 986 120 430 | gradiant@gradiant.org | www.gradiant.org