

DATA-DRIVEN POSTMORTEMS

JASON YEE, DATADOG

@GITBISECT



about me:

@gitbiseet

Technical Writer/Evangelist

“Docs & Talks”

Travel Hacker & Whiskey Hunter



about Datadog:

@Datadoghq

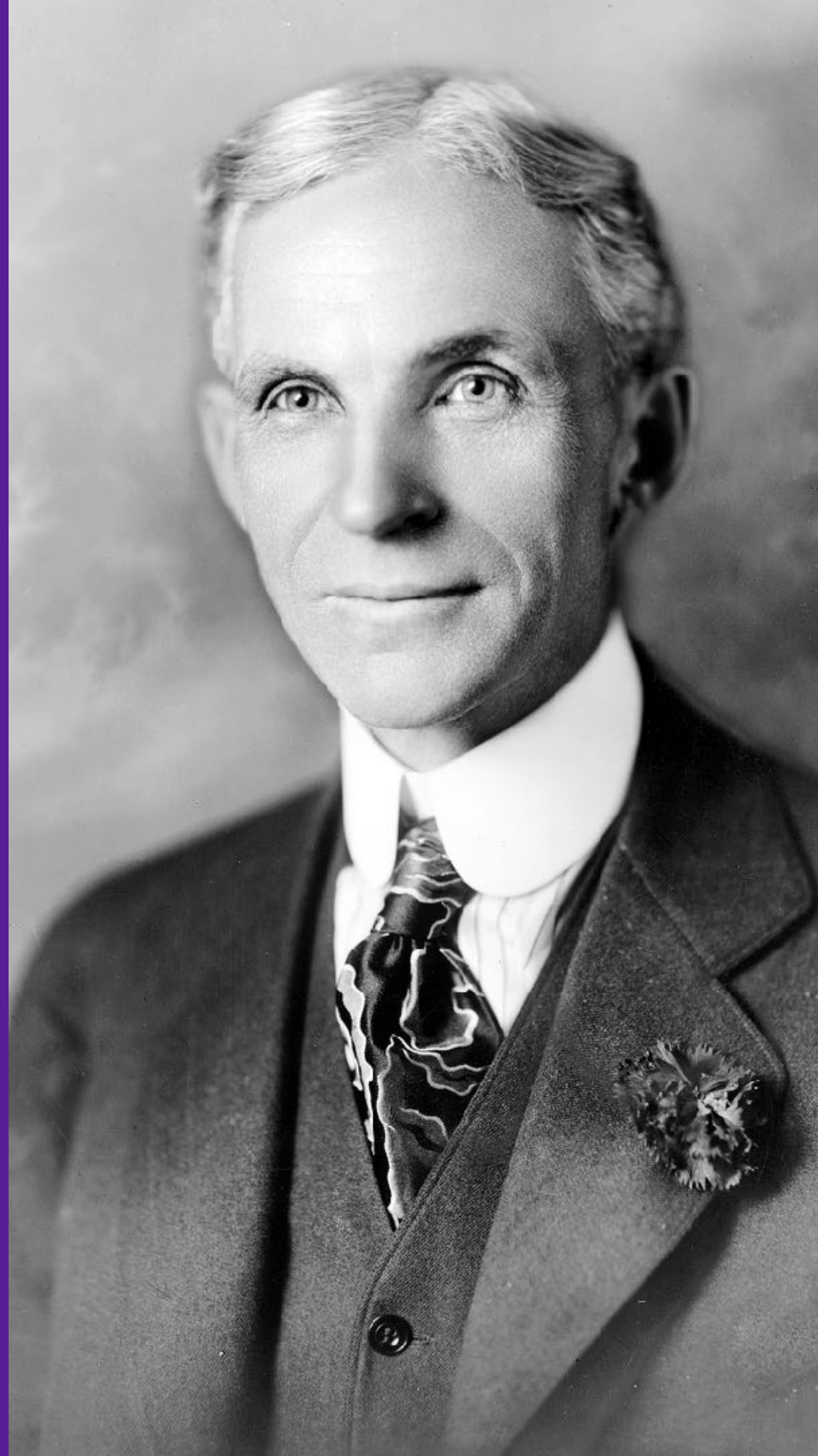
SaaS based monitoring platform

Trillions of data points per day

We're hiring! bit.ly/datadog-jobs

“The problems we work on at Datadog are hard and often don't have obvious, clean-cut solutions, so it's useful to cultivate your troubleshooting skills, no matter what role you work in.”

Internal Datadog Developer Guide



**“THE ONLY REAL
MISTAKE IS THE
ONE FROM WHICH
WE LEARN
NOTHING.”**

- Henry Ford

**COLLECTING DATA IS CHEAP;
NOT HAVING IT WHEN YOU
NEED IT CAN BE EXPENSIVE**

SO INSTRUMENT ALL THE THINGS!



British Airways Union Blames Massive IT Failure On Outsourcing IT Jobs To India

The carrier cancelled hundreds of flights from London yesterday.

28/05/2017 12:57 PM IST | Updated 28/05/2017 12:59 PM IST

ANI



NEIL HALL / REUTERS

LONDON -- British Airways GMB union has blamed the airline's 2016 decision of outsourcing IT jobs to India as the reason behind cancelling all Saturday flights from London to New York and Los Angeles.




















4 QUALITIES OF GOOD METRICS

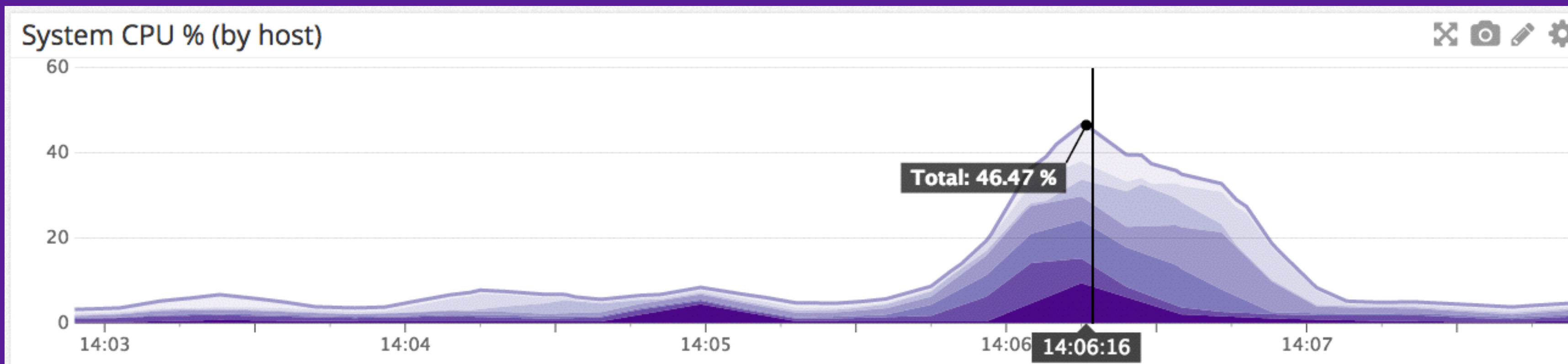
NOT ALL METRICS ARE EQUAL

1. MUST BE WELL UNDERSTOOD

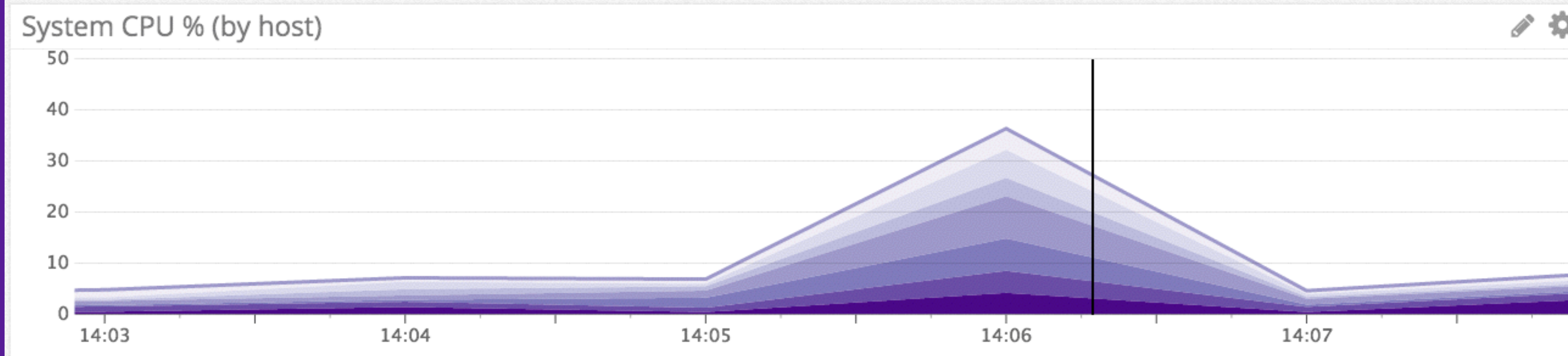


2. SUFFICIENT GRANULARITY

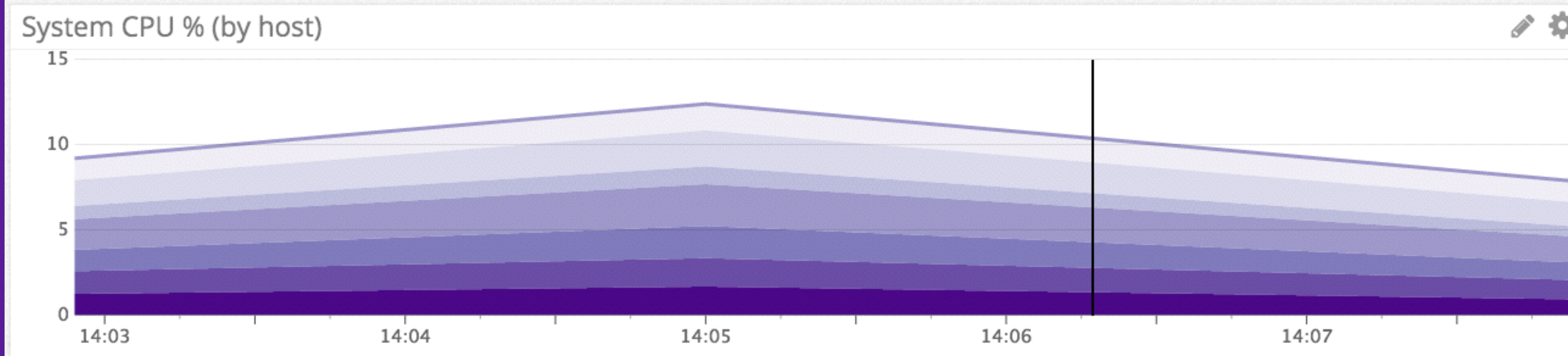
RANK	PARTICIPANT		RESULT
		Anthony ERVIN  USA	21.40
		Florent MANAUDOU  FRA	21.41
		Nathan ADRIAN  USA	21.49
4.		Ben PROUD  GBR	21.68
5.		Andrii GOVOROV  UKR	21.74
6.		Bruno FRATUS  BRA	21.79
6.		Bradley Edward TANDY  RSA	21.79
8.		Simonas BILIS  LTU	22.08



1 second
Peak 46%



1 minute
Peak 36%



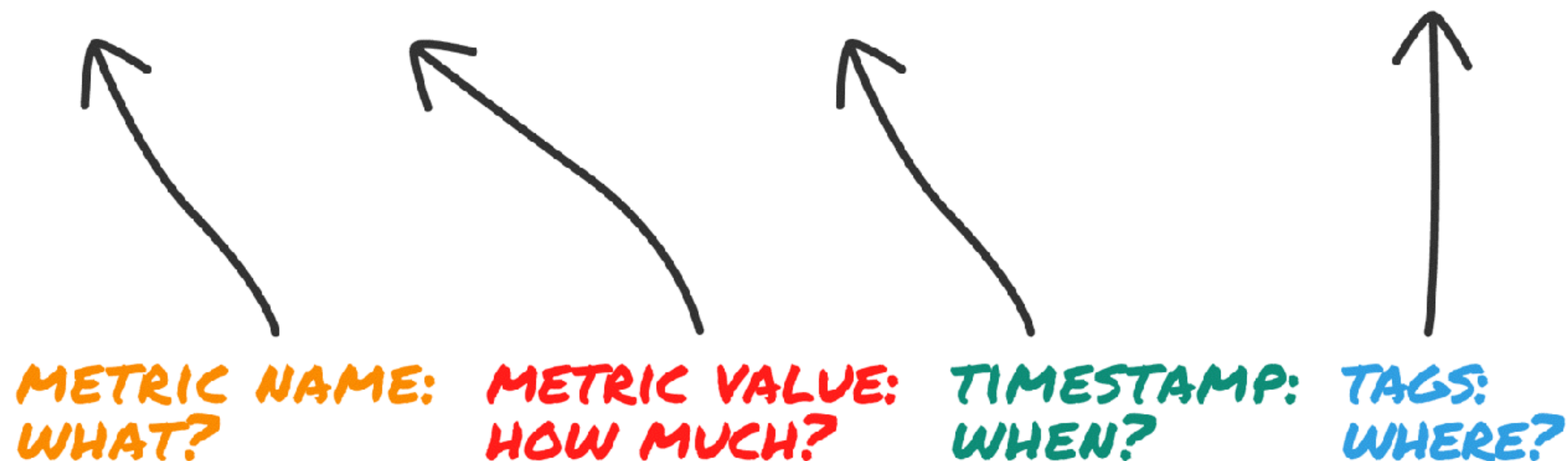
5 minutes
Peak 12%

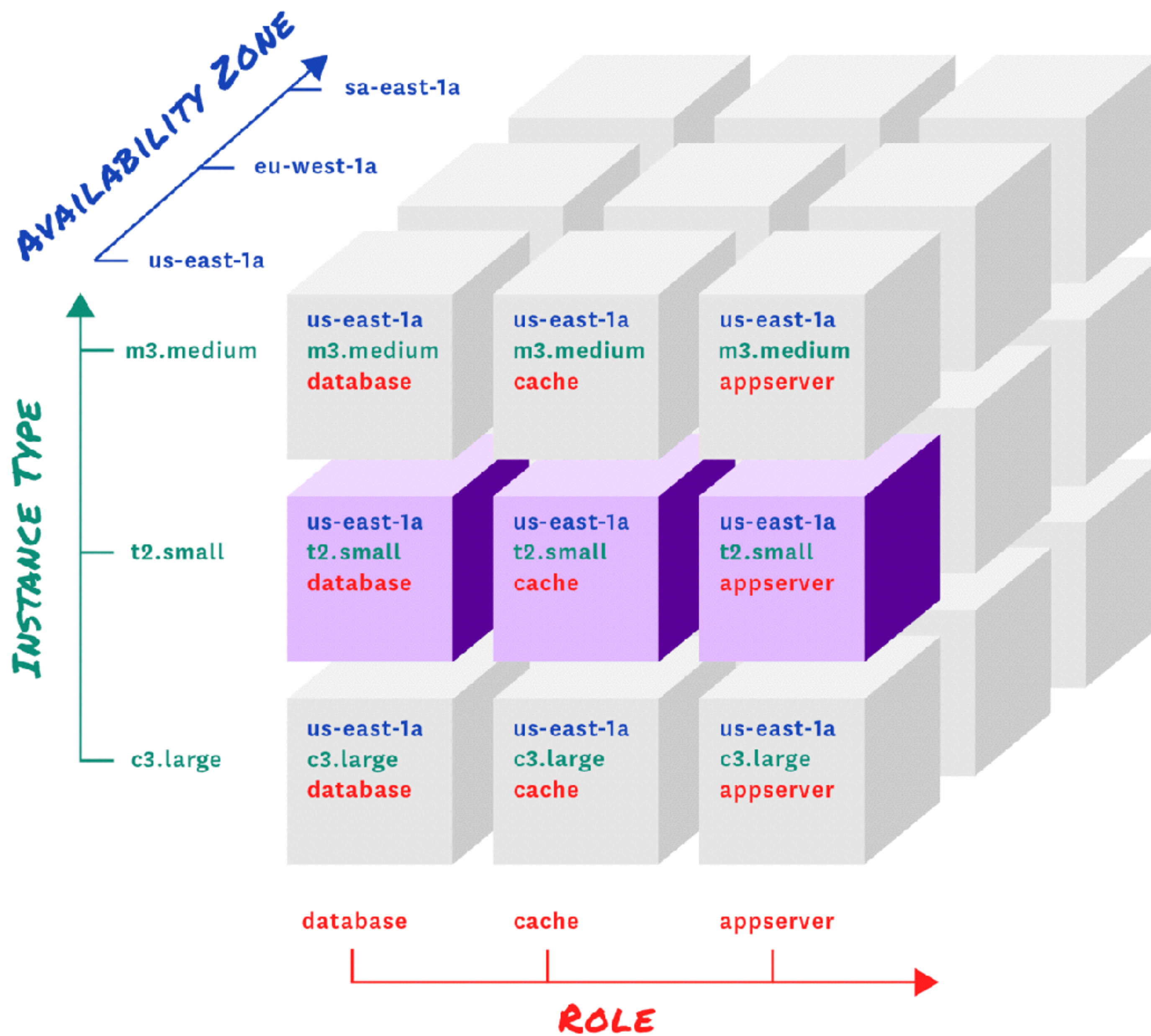


3. TAGGED & FILTERABLE

DATAPoint

SYSTEM.NET.BYTES_RCVD 4 2016-03-02 15:00:00 [FILE-SERVER]





4. LONG-LIVED

WORK METRICS

RESOURCE METRICS

EVENTS



WORK METRICS

THROUGHPUT

SUCCESS

ERROR

PERFORMANCE



RESOURCE METRICS

UTILIZATION

SATURATION

ERROR

AVAILABILITY



EVENTS

CODE CHANGES

ALERTS

SCALING EVENTS

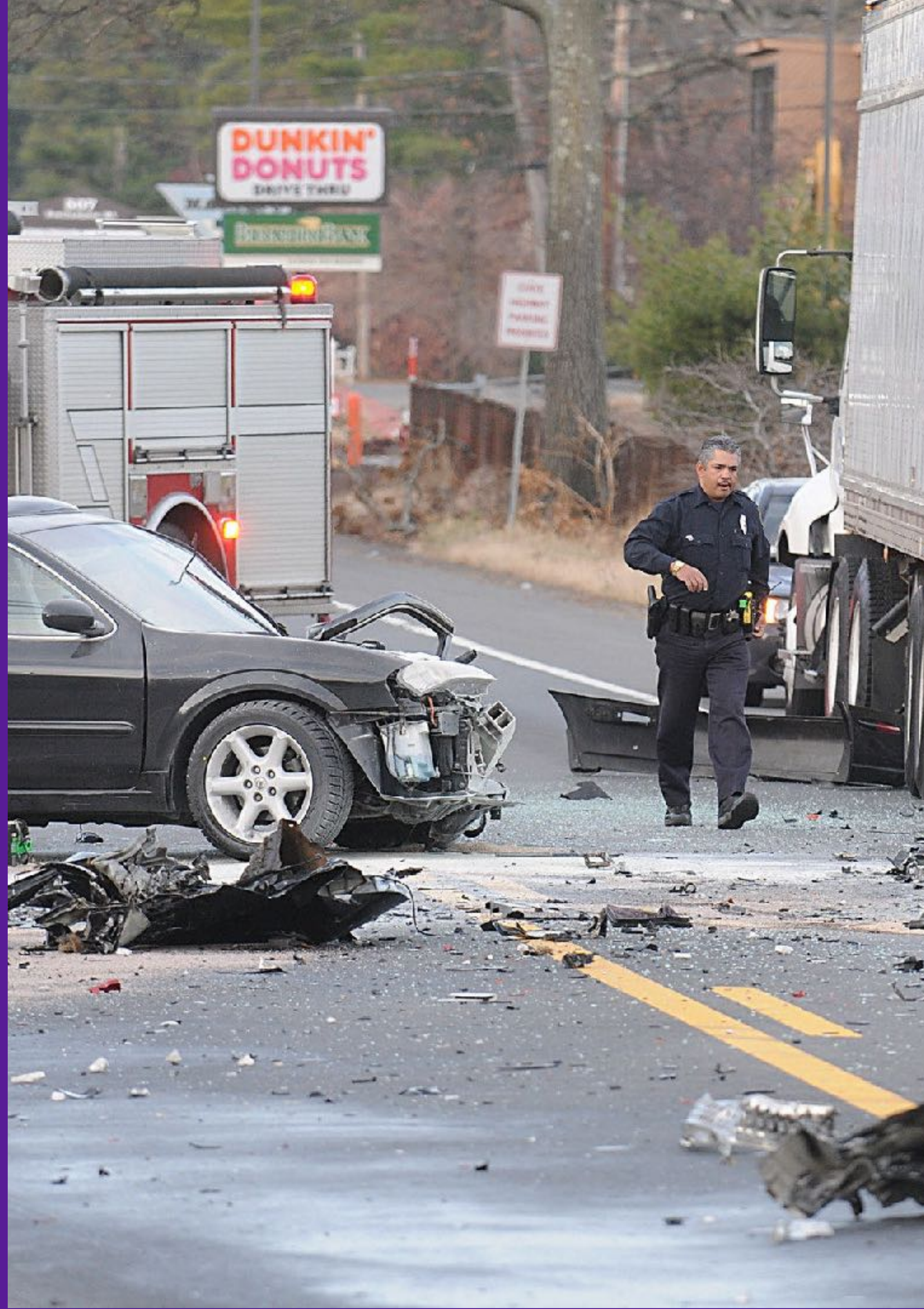
ETC

RECURSE UNTIL YOU FIND THE TECHNICAL CAUSES



TECHNICAL ISSUES HAVE NON-TECHNICAL CAUSES

HUMAN ELEMENT



**IF YOU'RE STILL
RESPONDING TO
THE INCIDENT,
IT'S NOT TIME FOR
A POSTMORTEM**

DATA COLLECTION: WHO?

- ▶ Everyone!
 - ▶ Responders
 - ▶ Identifiers
 - ▶ Affected Users

HUMAN DATA



DATA COLLECTION: WHAT?

- ▶ Their perspective
 - ▶ What they did
 - ▶ What they thought
 - ▶ Why they thought/did it

**“WRITING IS NATURE’S WAY OF
LETTING YOU KNOW HOW SLOPPY
YOUR THINKING IS.”**

RICHARD GUINDON

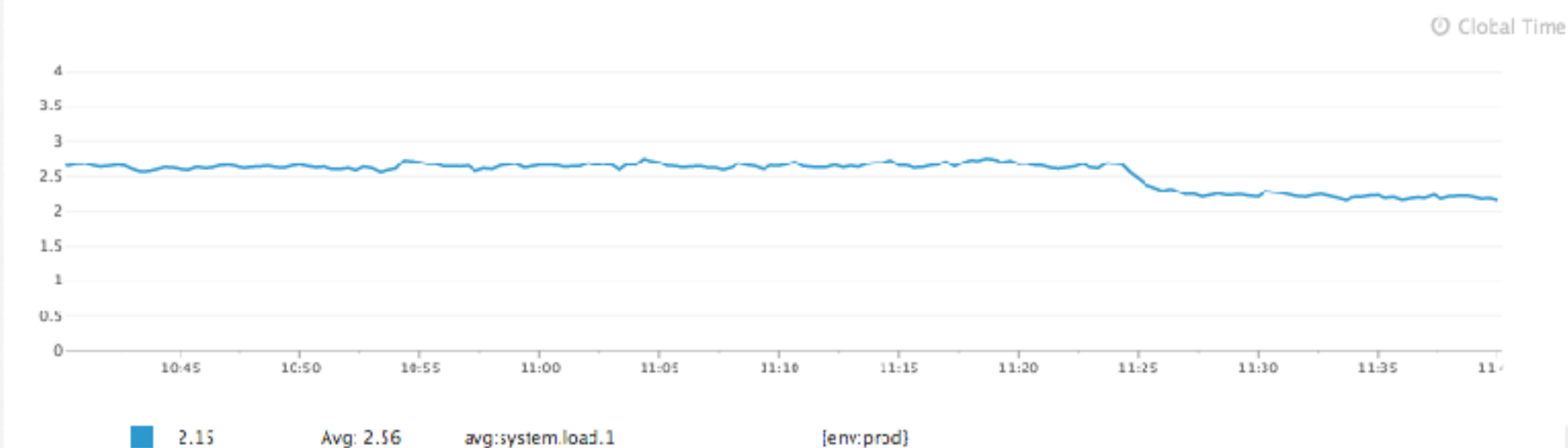
TELLING STORIES

Welcome to the **Datadog Example Notebook**.

Notebooks are designed to help you tell stories with your data. We've all heard the adage, *a picture is worth a thousand words*, and it certainly rings true when sharing metrics information. From reporting on infrastructure changes to sharing incident retrospectives, visualizing data allows you to communicate better.

Notebooks are composed of sequential markdown and graph cells. This is a markdown cell and below you'll see a graph cell. You can easily edit any cell by clicking on it.

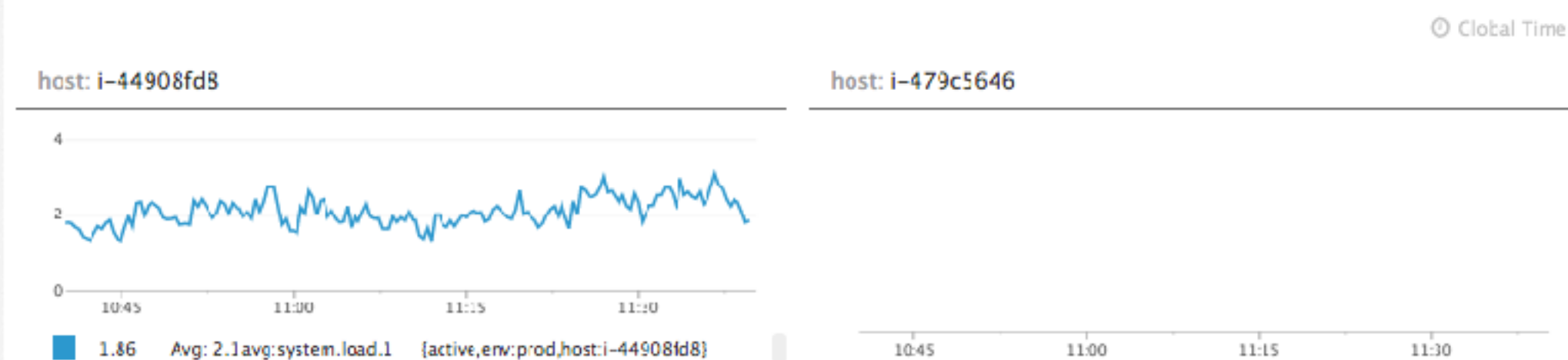
`avg:system.load.1{env:prod}`



The graph above is your average system load over the past hour from your environments that have been tagged as "prod". Note that it is following *Global Time* as indicated by the text on the right, above the graph. This means that it's following the time indicated at the top of the page.

The time indicator should look familiar. It's similar to the time controls in your other Datadog dashboards. You can select different durations and use the arrow buttons to move forward or backward in time.

`avg:system.load.1{active,env:prod}`



`host: i-a38b4239`



`host: i-c9e7074f`



“A PICTURE IS WORTH A THOUSAND WORDS”

– ANCIENT PROVERB

DATA COLLECTION: WHEN?

- ▶ As soon as possible.
 - ▶ Memory drops sharply within 20 minutes
 - ▶ Susceptibility to “false memory” increases
 - ▶ Get your project managers involved!

DATA SKEW/CORRUPTION

- ▶ Stress
- ▶ Sleep deprivation
- ▶ Burnout

DATA SKEW/CORRUPTION

- ▶ Blame
- ▶ Fear of punitive action

DATA SKEW/CORRUPTION

- ▶ Bias
 - ▶ Anchoring
 - ▶ Hindsight
 - ▶ Outcome
 - ▶ Availability (Recency)
 - ▶ Bandwagon Effect



HOW WE DO POSTMORTEMS AT DATADOG

A FEW NOTES

- ▶ Postmortems emailed to company wide
- ▶ Scheduled recurring postmortem meetings

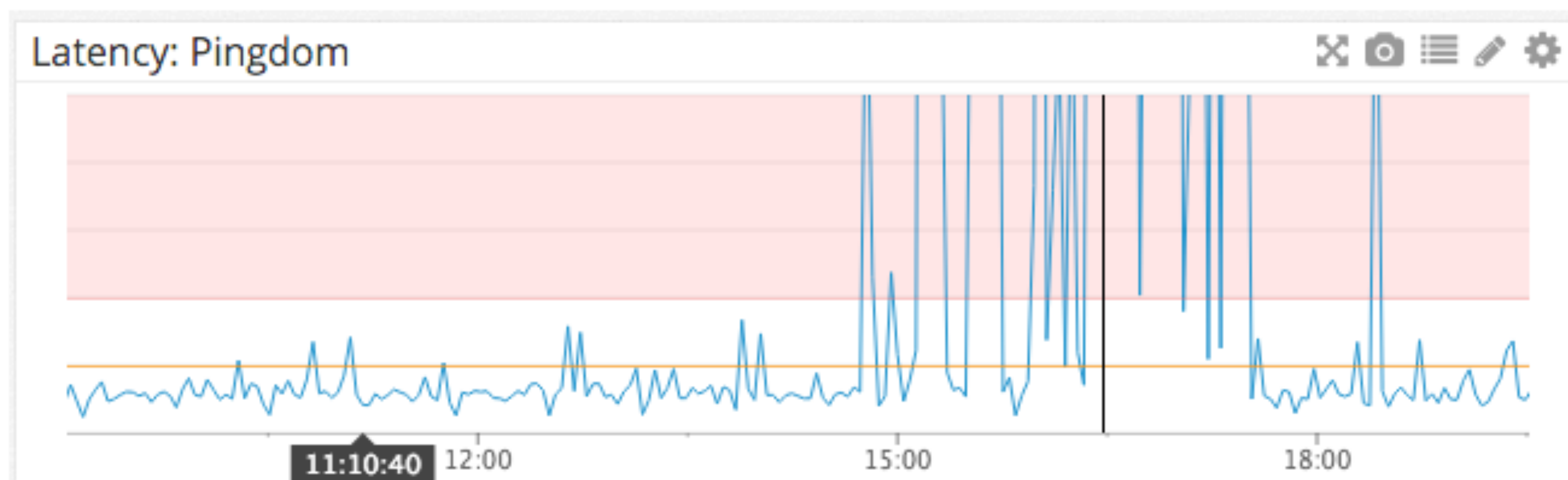
SUMMARY: WHAT HAPPENED?

- ▶ Describe what happened here at a high-level -- think of it as an abstract in a scientific paper.
- ▶ What was the impact on customers?
- ▶ What was the severity of the outage?
- ▶ What components were affected?
- ▶ What ultimately resolved the outage?

Summary: what happened?

We lost most of mcnulty-web & mcnulty-query capacity, while nodes were blocked on accessing global cache, resulting in increased latency and 5XX errors. Customers trying to access Datadog were shown a "down" page for long periods of time during the outage. Elena cache nodes seemed to be overloaded, especially on the network side.

The outage lasted from 3:11pm to 5:32pm.



Impact on customers

- Customers trying to access Datadog were shown a "down" page. Already open dashboards have been able to refresh tiles successfully, although with increased latency. Intake and Alerting components were unaffected by the outage. No data loss.


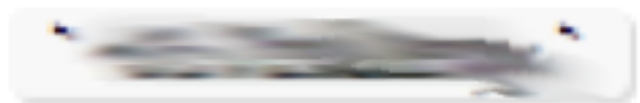

Severity of the outage

- Major (while there has been no data loss, Dogweb was inaccessible for over 2 hours)

Components affected

- McNulty, Snapshots (PhantomJS), Crawlers

What ultimately resolved the outage

- Elena master nodes (`r3.large` in us-east-1a and us-east-1b) replaced with 
 running in 

HOW WAS THE OUTAGE DETECTED?

- ▶ We want to make sure we detected the issue early and would catch the same issue if it were to repeat.
- ▶ Did we have a metric that showed the outage?
- ▶ Was there a monitor on that metric?
- ▶ How long did it take for us to declare an outage?

How was the outage detected?

Did we have a metric that showed the outage?

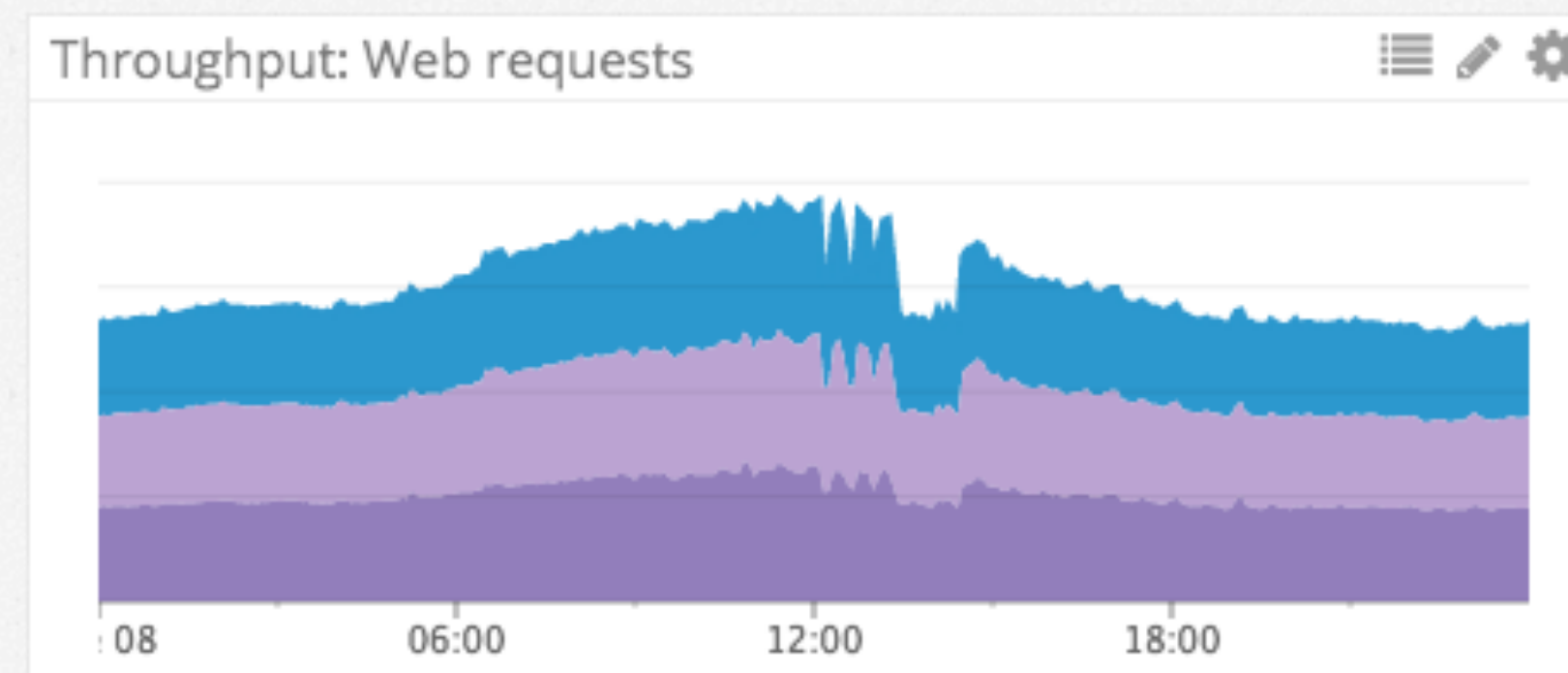
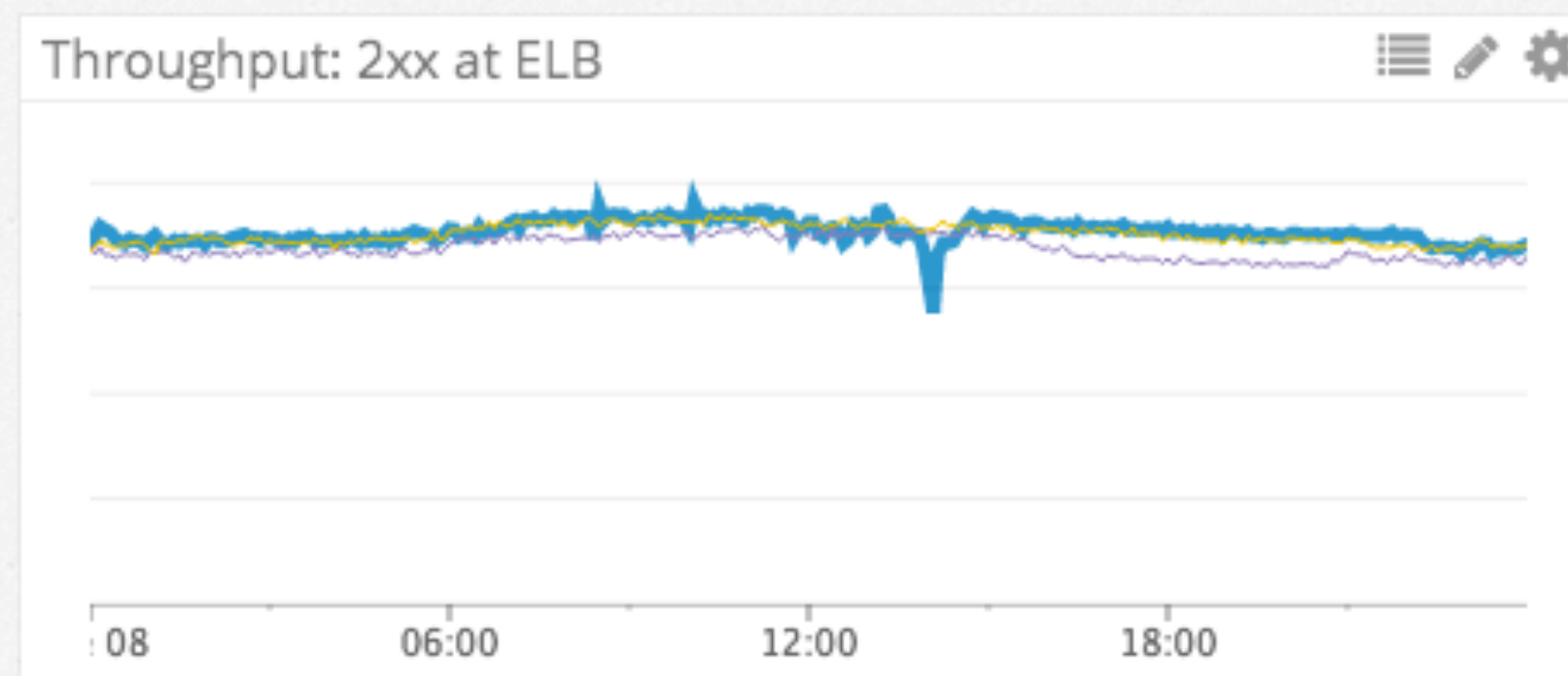
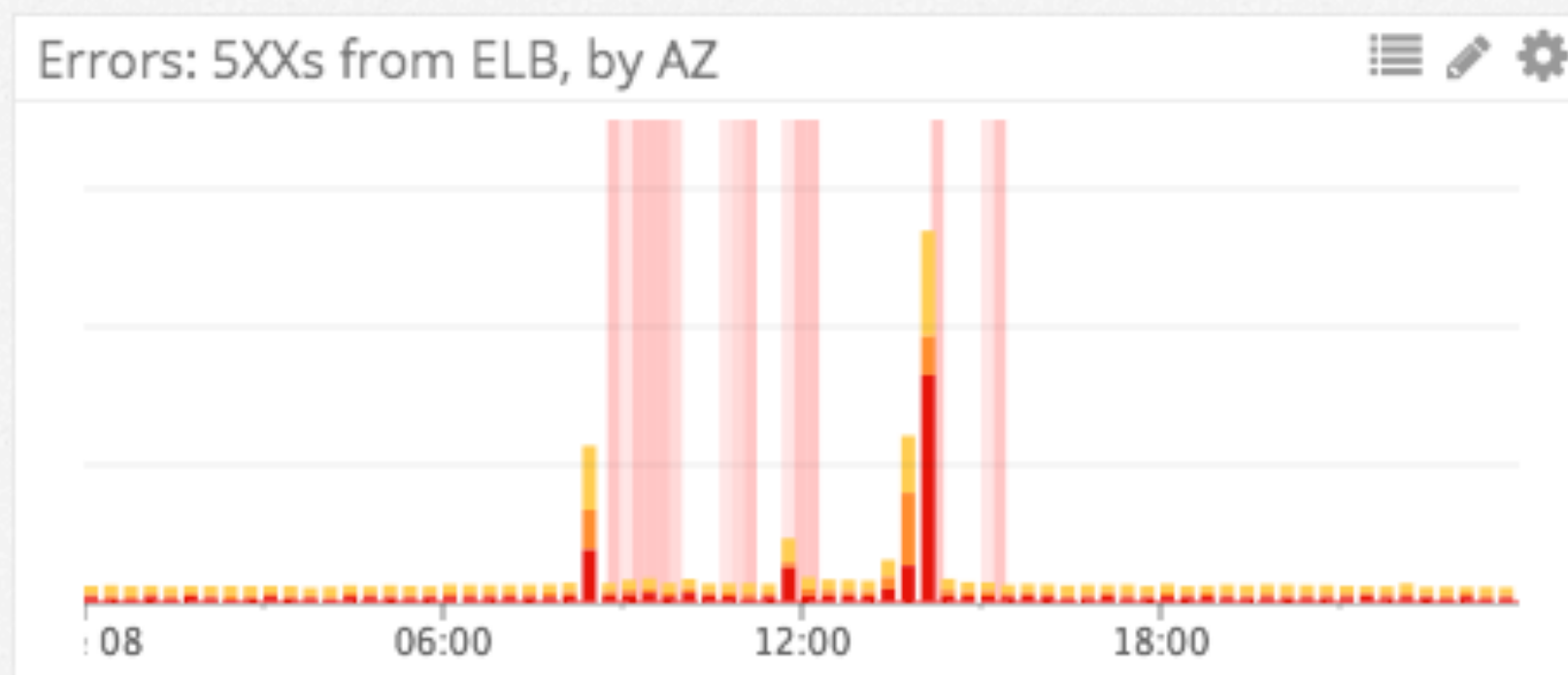
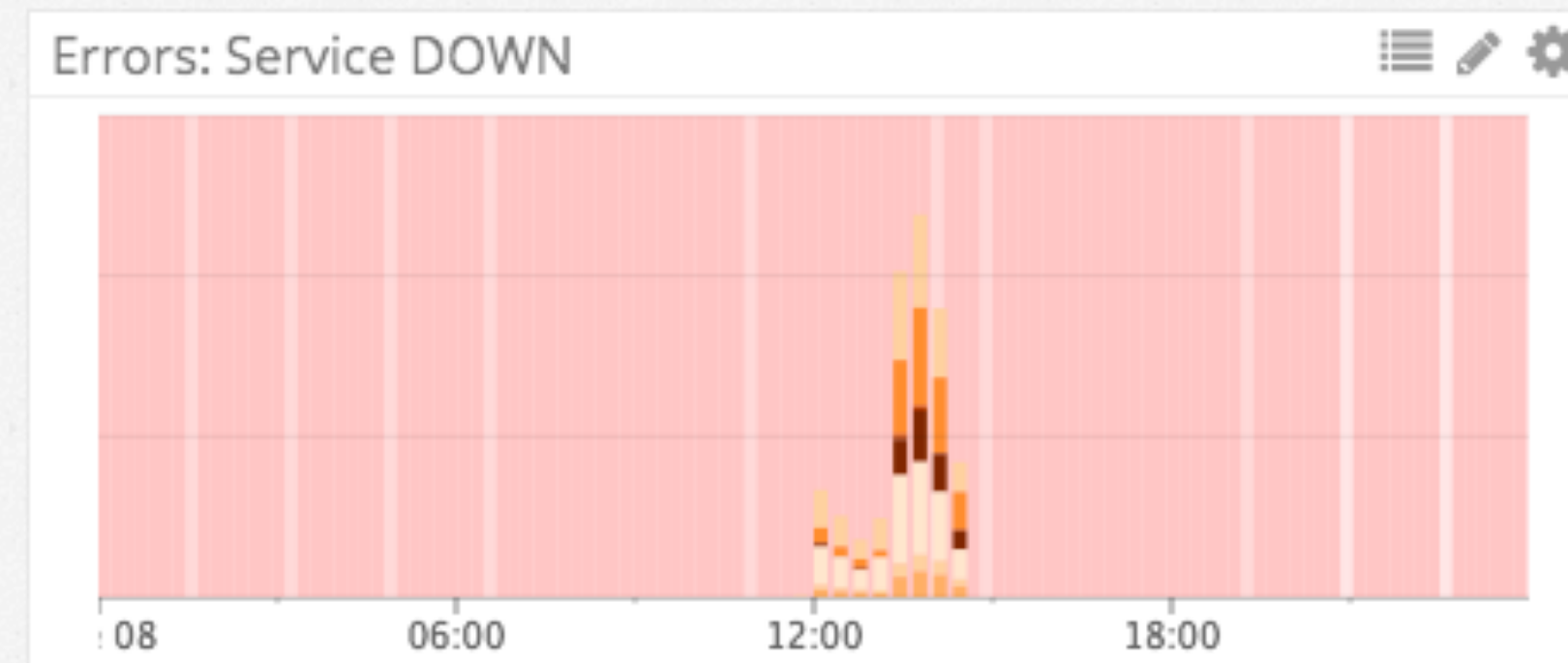
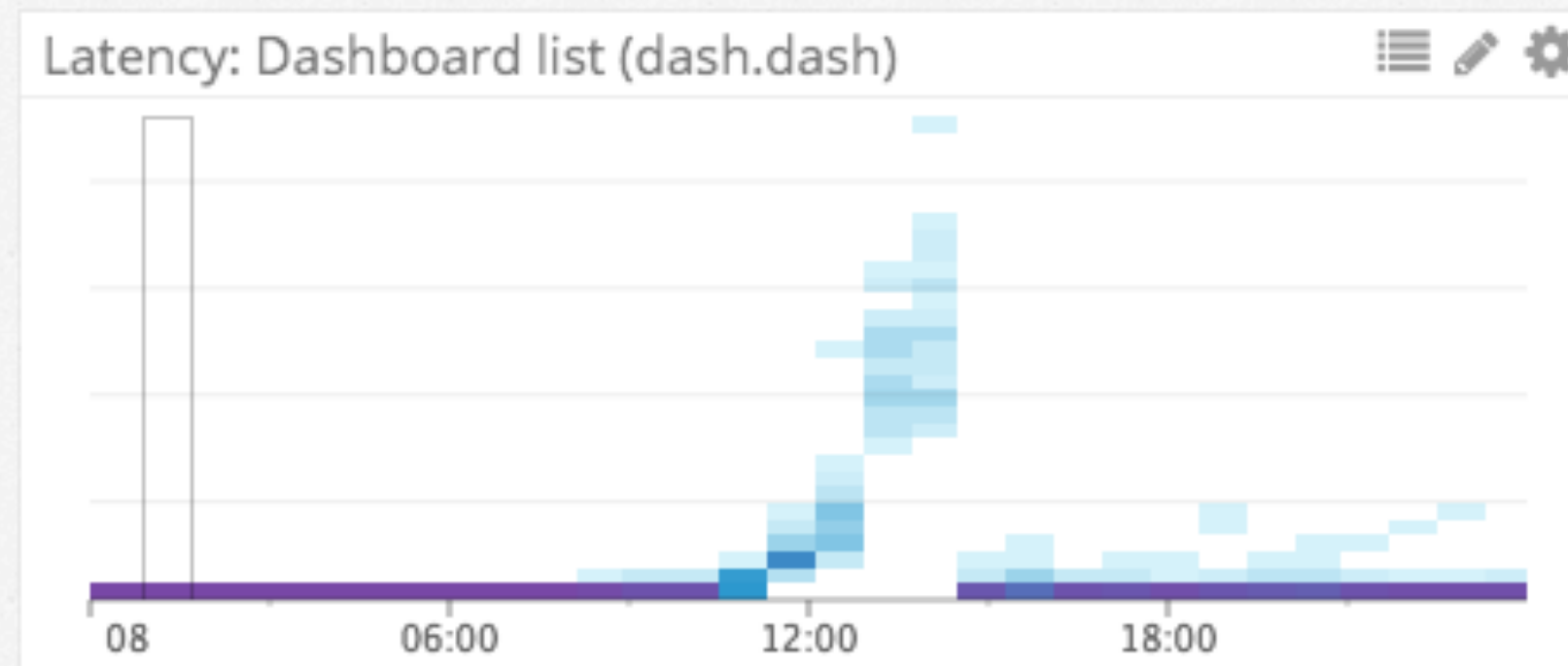
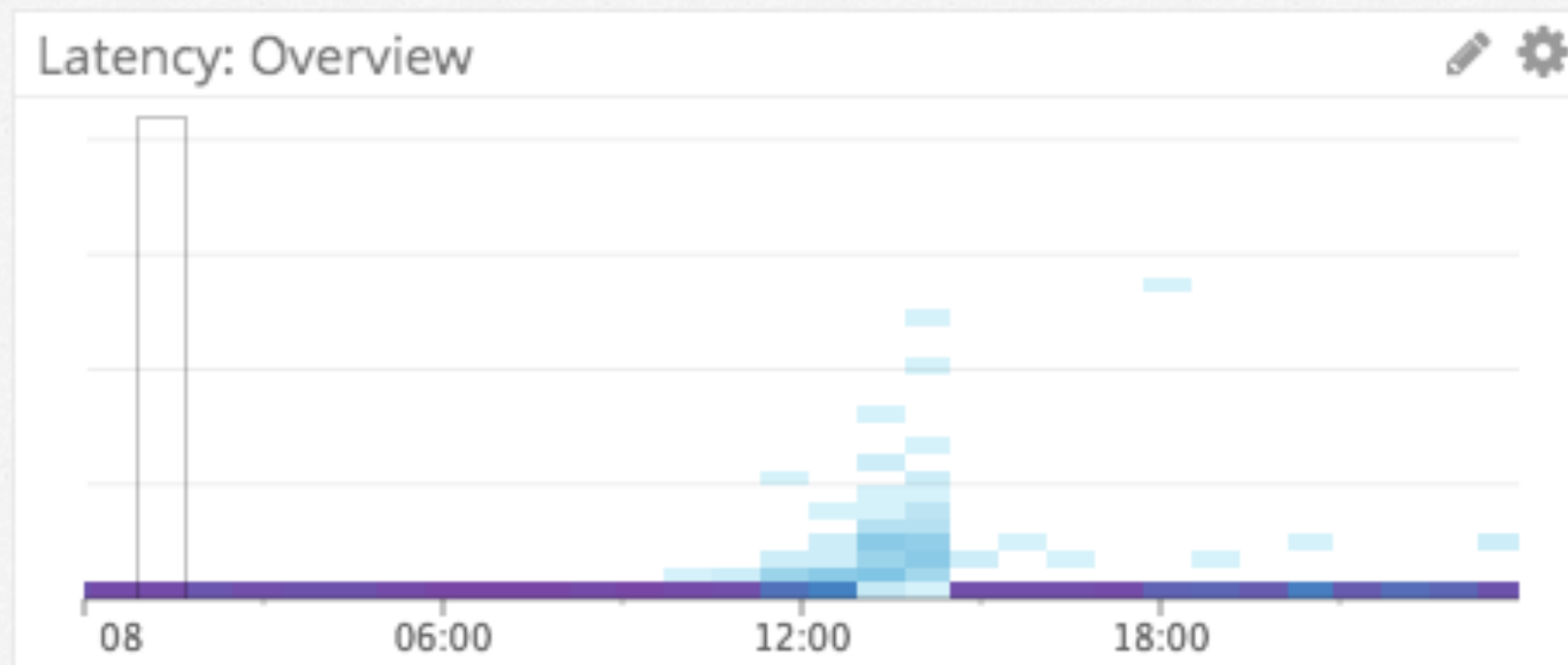
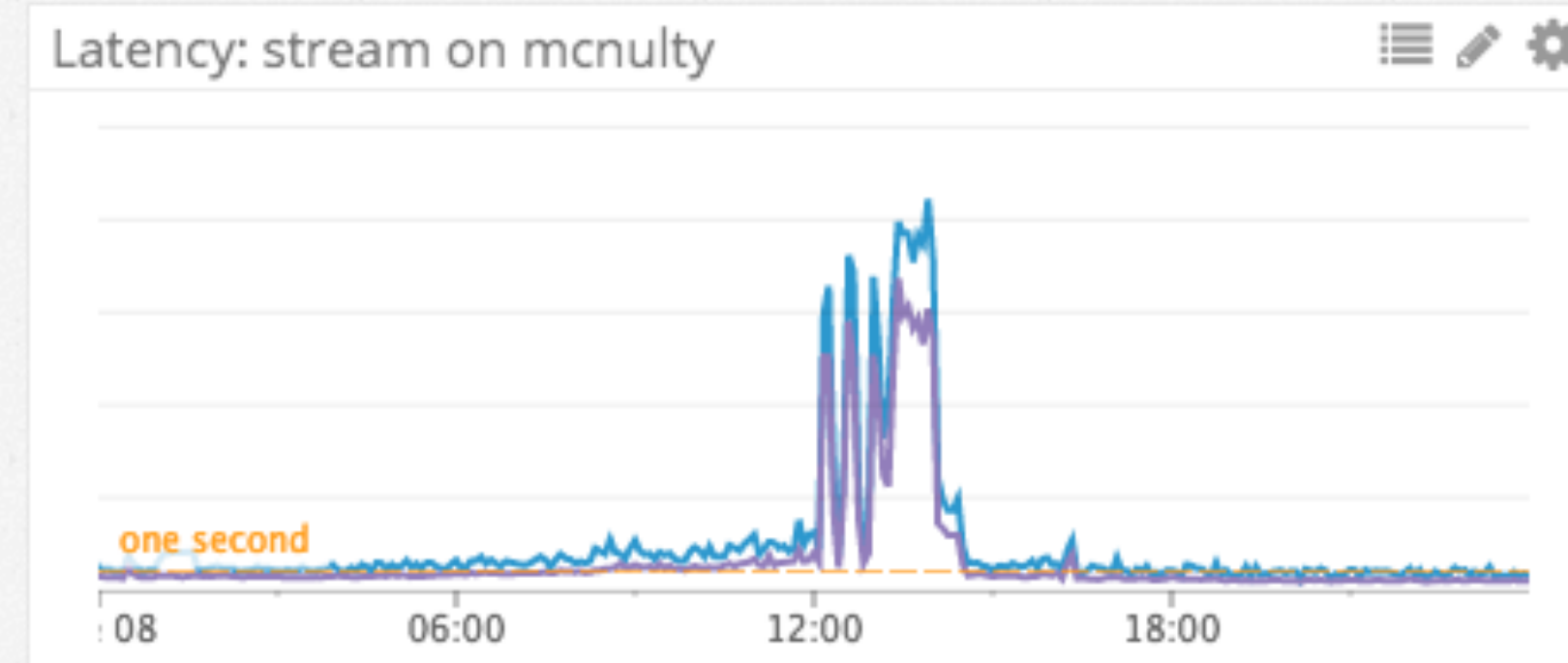
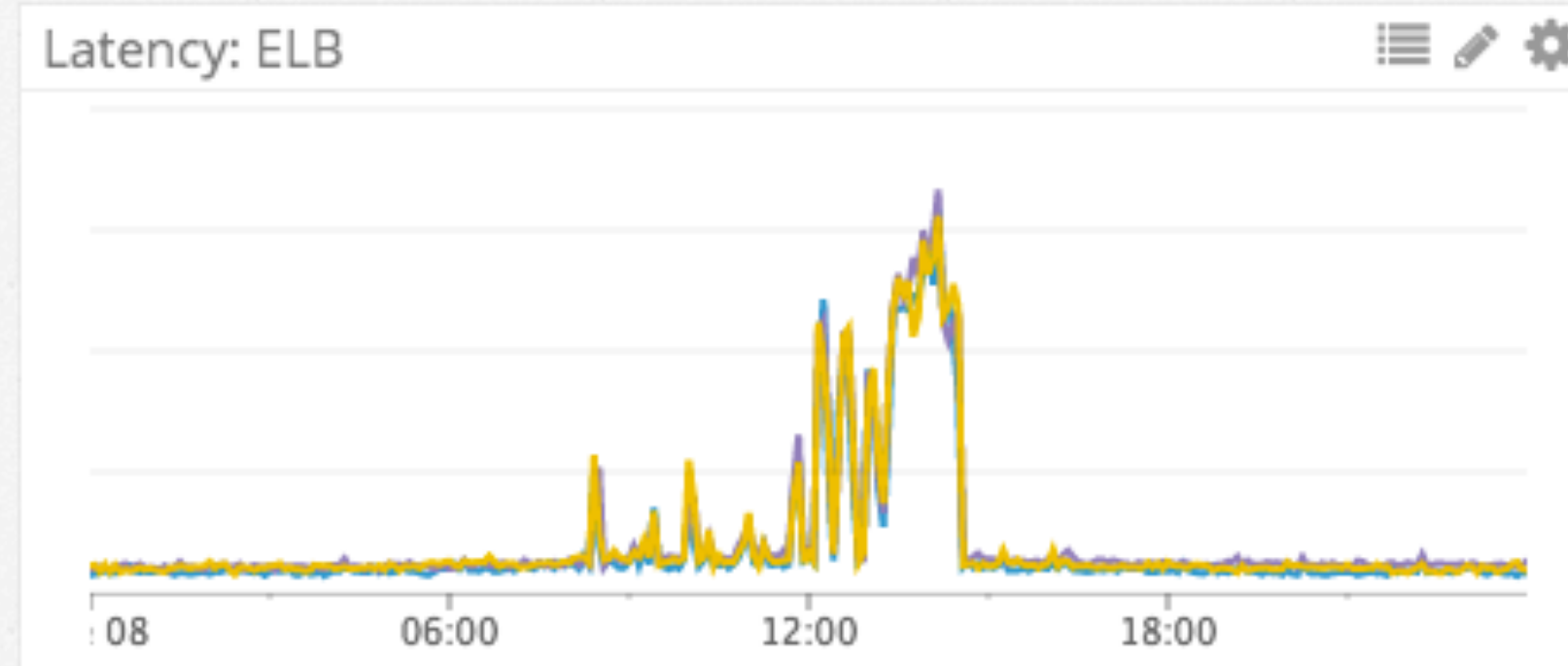
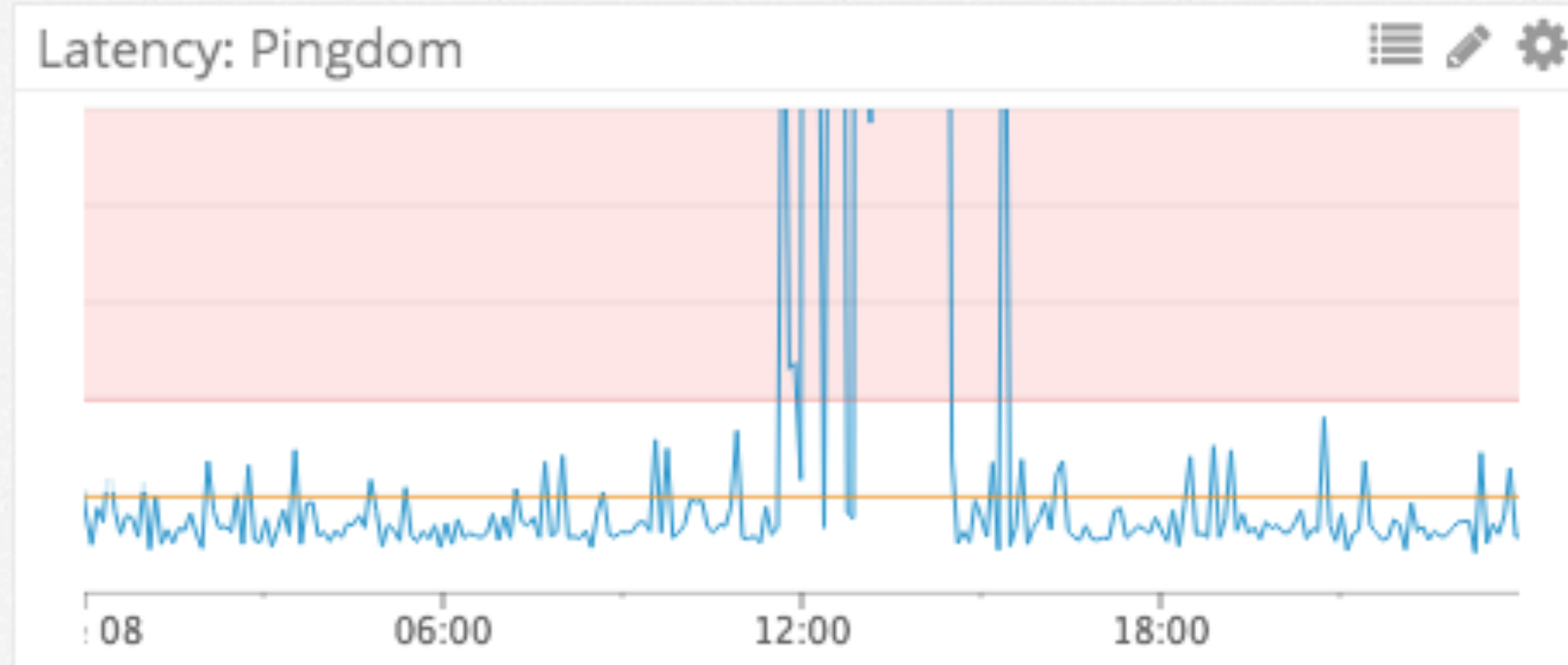
- Multiple metrics were indicating an outage, including: ``pingdom.response_time``, ``aws.elb.latency``, ``haproxy.count_per_status``, ``haproxy.backend_up`` See: <https://app.datadoghq.com/dash/web-perf-seen-from-users>

Was there a monitor on that metric?

- Yes, eg. ``haproxy.backend_up.over("service:dogweb")``: <https://app.datadoghq.com/monitors#110558>

How long did it take for us to declare an outage?

- The outage has been declared at 3:14pm, **3 minutes** after first 5XX were seen and about 5 minutes after an increase in latency was seen.



HOW DID WE RESPOND?

- ▶ Who was the incident owner & who else was involved?
- ▶ Slack archive links and timeline of events!
- ▶ What went well?
- ▶ What didn't go so well?

How did we respond?

The incident owner was Jeff. Responders included Noel, Tony, Mira, Kyle, Mark among others

- Slack archive links: <https://dd.slack.com/archives/outage/p147002169>
- Graphs tagged #postmortem
 - https://app.datadoghq.com/event/stream?tags_execution=and&show_private=true&..800000
- Elena retransmits notebook: <https://app.datadoghq.com/notebook#225/Elena-retransmits>

Timeline of events

15:09 Increase in latency noticed by Pingdom

15:11 Increase in 5XX errors noticed by ``haproxy.count_per_status`` metric; Dogweb goes down

15:14 MF calls an outage <https://datadog.pagerduty.com/services/PSXXN8Q>

15:16 IS notices an error on McNulty nodes with accessing cache

<https://dd.slack.com/archives/outage/p14574XX1002171>

**Kyle** 3:53 PM

couldn't tell where it was timing out

**Tony** 3:53 PM

mmh ok, I can take a quick pass at seeing what's wrong. because it should really reduce the impact on redis to clear

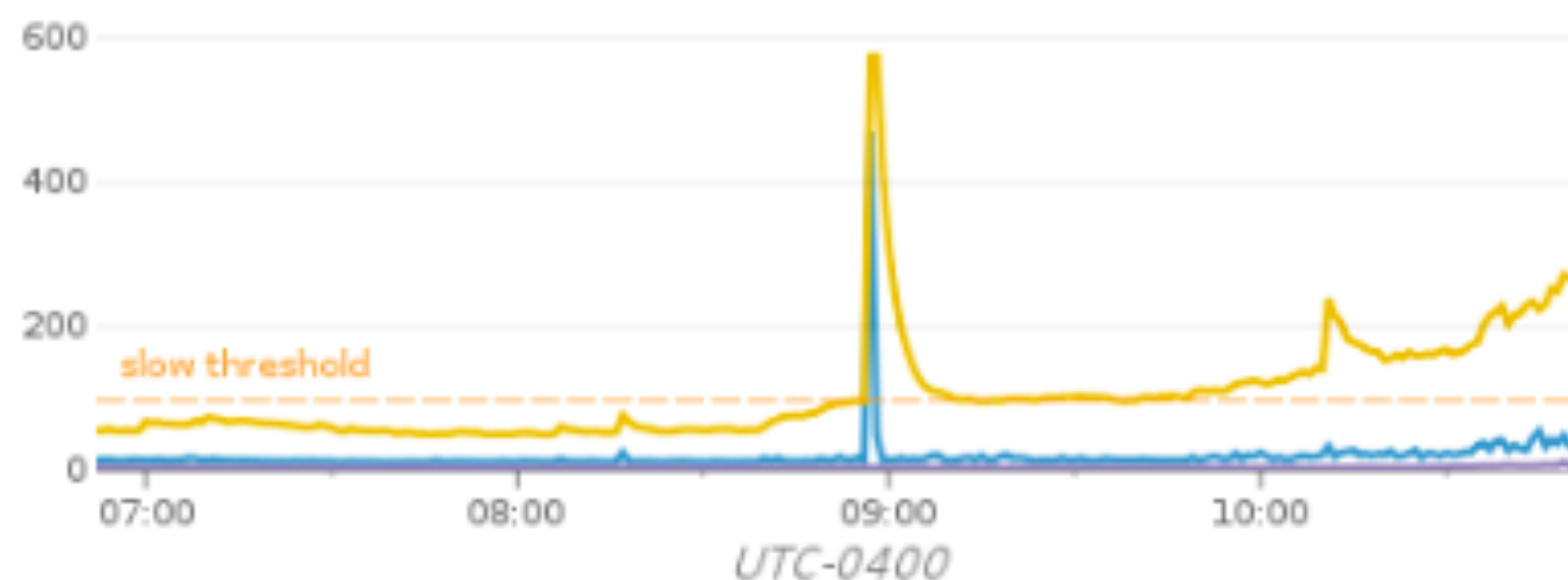
**Datadog** BOT 3:53 PM**Query Cache Fetch Duration**

@slack-outage

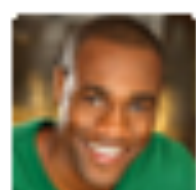
Sent By

Mark

(14KB) ▾

**Mark** 3:53 PM

i think that one node finally hit the limit last night, another node is slowing down now, etc.

**Tony** 3:53 PM

kicking off a cache clear now to see what's up

CHATOPS ARCHIVES FTW!

TRACK LEARNINGS AS YOU GO



Dan

3:53 PM

#postmortem we didn't get paged

WHY DID IT HAPPEN?

- ▶ Deep dive into the cause
- ▶ Examples from this incident:
 - ▶ <http://bit.ly/dd-statuspage>
 - ▶ <http://bit.ly/alq-postmortem>

HOW DO WE PREVENT IT IN THE FUTURE?

- ▶ Link to Github issues and Trello cards
- ▶ Now?
- ▶ Next?
- ▶ Later?
- ▶ Follow up notes

How do we prevent it in the future?

Now

- ☒ *Ship fixes to elena provision to account for kernel/redis tweaks - Mira & Kyle*
- ☒ *DataDog/devops#4594, DataDog/devops#4598,
<https://github.com/DataDog/devops/commit/7ff84666911416bd5b563369>*
- ☐ *add monitor for TCP error rate on Elena <https://trello.com/1199-add-monitor-for-tcp-error-rate>
(<https://app.datadoghq.com/monitors#508801/edit>)*
- ☐ *MIRA F This does not appear to be elena-centric, nor a well-defined alert - I am reluctant to call this one "done" until then.*

DATADOG'S POSTMORTEM TEMPLATE

RECAP:

- ▶ What happened (summary)?
- ▶ How did we detect it?
- ▶ How did we respond?
- ▶ Why did it happen (deep dive)?
- ▶ Actionable next steps!

KEEP LEARNING

MORE RESOURCES

- ▶ Postmortem Template

<http://bit.ly/postmortem-template>

- ▶ The Infinite Hows - John Allspaw

<http://bit.ly/infinite-hows>

SLIDES: bit.ly/dod-ams-postmortems

QUESTIONS: @gitbisect | jason.yee@datadoghq.com