How Did Things Go Right?

RYAN KITCHENS
NETFLIX

@this_hits_home



SRE WORLD

SLOs & SLIs

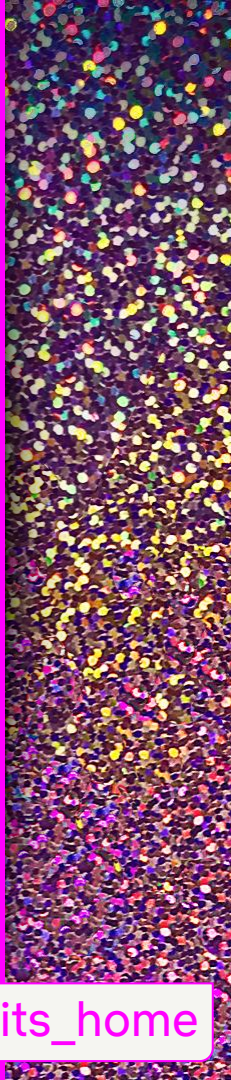
Error Budgets

War Rooms

Incident Reviews

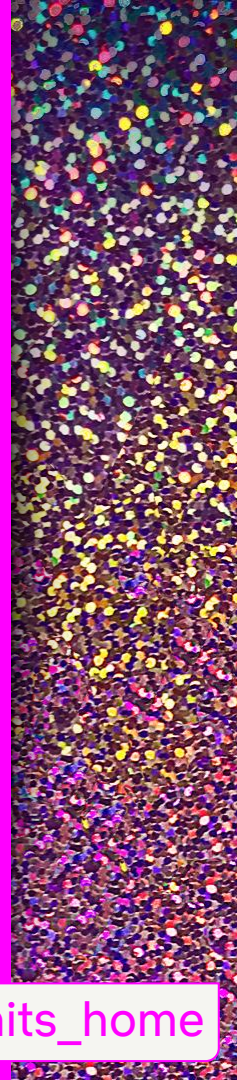
@this_hits_home

Failure is so important



@this_hits_home

Failure is so important that it's no longer interesting.

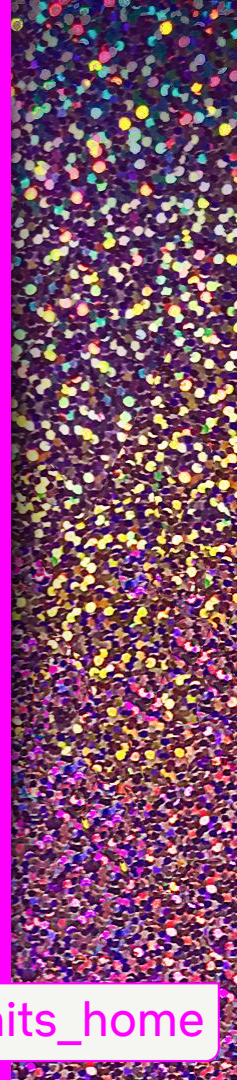


@this_hits_home

Failure is so important that it's no longer interesting.

RYAN KITCHENS

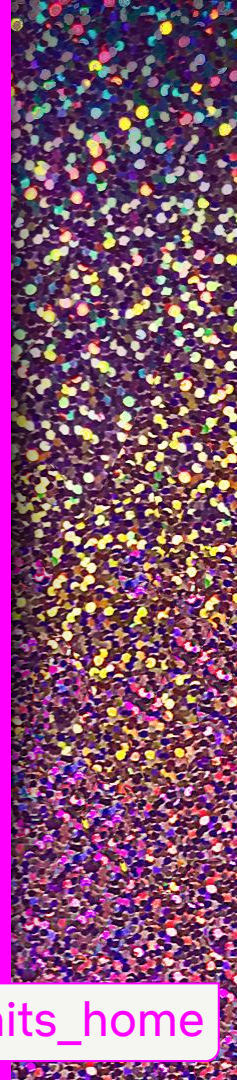
NETFLIX



@this_hits_home

Failure is the normal state.

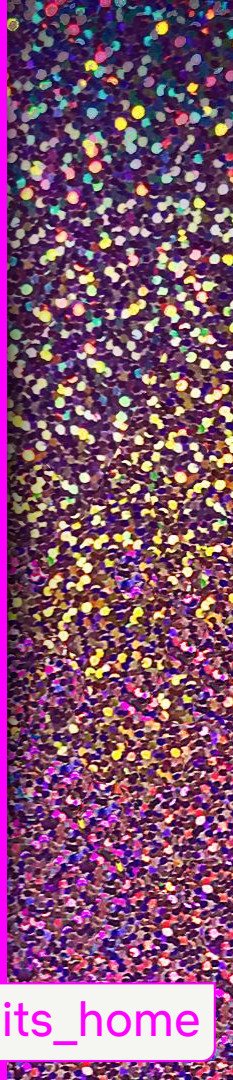
Failure is so important that it's no longer interesting.



Failure is so important that it's no longer interesting.

Failure is the normal state.

We're pretty good
@ preventing incidents.



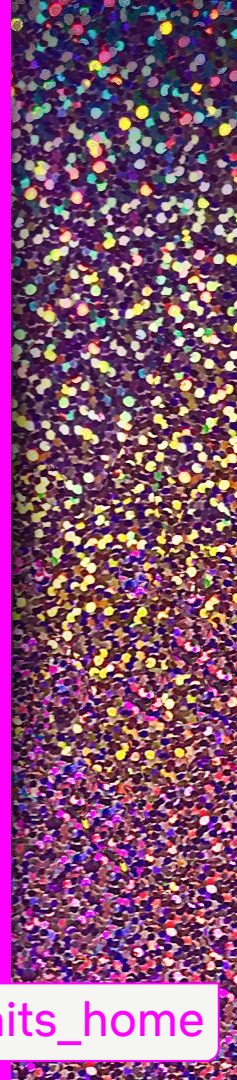
@this_hits_home

Failure is so important that it's no longer interesting.

Failure is the normal state.

We're pretty good
@ preventing incidents.

Recovery > prevention.



@this_hits_home

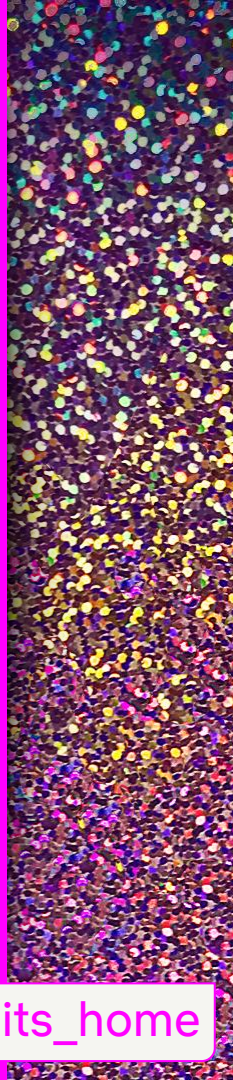
Failure is so important that it's no longer interesting.

Failure is the normal state.

We're pretty good @ preventing incidents.

Recovery > prevention.

The most important thing we can learn is how to build the capacity to encounter failure successfully.



@this_hits_home

Failure is so important that it's no longer interesting.

Failure is the normal state.

We're pretty good
@ preventing incidents.



Recovery > prevention.

Failure is so important that it's no longer interesting.

Failure is the normal state.

We're pretty good
@ preventing incidents.



Recovery > prevention.

the nines don't matter.

@this_hits_home

Failure is so important that it's no longer interesting.

Failure is the normal state.

We're pretty good
@ preventing incidents.



Recovery > prevention.

Availability is made up
& the nines don't matter.

@this_hits_home

Failure is so important that it's no longer interesting.

Failure is the normal state.

We're pretty good
@ preventing incidents.



Recovery > prevention.

Availability is made up
& the nines don't matter.

Success isn't the absence of failure.

@this_hits_home

Failure is so important that it's no longer interesting.

Failure is the normal state.

We're pretty good
@ preventing incidents.



Recovery > prevention.

Availability is made up
& the nines don't matter.

Success isn't the absence of failure.

There is no panacea

Failure is so important that it's no longer interesting.

Failure is the normal state.

We're pretty good
@ preventing incidents.



Recovery > prevention.

Availability is made up
& the nines don't matter.

Success isn't the absence of failure.

There is no panacea
(except for expertise).

@this_hits_home

@this_hits_home

There is no root cause.


There is no
root cause.
@this_hits_home


There is no
groot cause.
@this_hits_home


There is no
root cause.
@this_hits_home


~~There is no~~
A PERFECT STORM
~~root cause.~~
@this_hits_home

~~There is no~~

A PERFECT STORM

~~root cause.~~

EACH NECESSARY,
BUT ONLY JOINTLY SUFFICIENT



~~There is no~~

A PERFECT STORM

multiple contributing factors.

~~root cause.~~

EACH NECESSARY,
BUT ONLY JOINTLY SUFFICIENT



SREcon19 Americas Themes

Comprehension Understandability Predictability

NONE OF THESE ARE ENTIRELY POSSIBLE

@this_hits_home

SREcon19 Americas Themes

Comprehension Understandability Predictability

Incidents cannot be fully comprehended.

They are fraught with uncertainty.

Remediation items will contribute to further incidents.

NONE OF THESE ARE ENTIRELY POSSIBLE



SREcon19 Americas Themes

Comprehension

Incidents cannot be fully comprehended.

They are fraught with uncertainty.

Remediation items will contribute to further incidents.

Understandability

Managing the small stuff does not prevent big incidents.

Incidents are not made up of causes. We do not find them; we construct them.

Predictability

NONE OF THESE ARE ENTIRELY POSSIBLE



SREcon19 Americas Themes

Comprehension

Incidents cannot be fully comprehended.

They are fraught with uncertainty.

Remediation items will contribute to further incidents.

Understandability

Managing the small stuff does not prevent big incidents.

Incidents are not made up of causes. We do not find them; we construct them.

Predictability

Learning from the last incident will not allow you to predict the next one.

Complex systems are not deterministic.

Their state cannot be precisely or repeatedly foretold.

NONE OF THESE ARE ENTIRELY POSSIBLE

@this_hits_home



Aaron P Blohowiak
@aaronblohowiak

Following

Why do we expect every engineer and operator to be Laplace's Demon?

8:30 AM - 11 Oct 2018

@this_hits_home

TO LEARN MORE FROM INCIDENTS

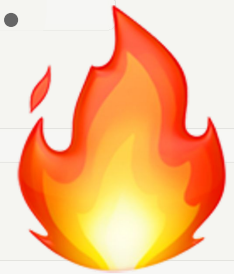
Ask more ‘**how**’ than ‘**why**’



5 Whys is a competitive disadvantage.

@this_hits_home

5 Whys is a competitive disadvantage.



@this_hits_home

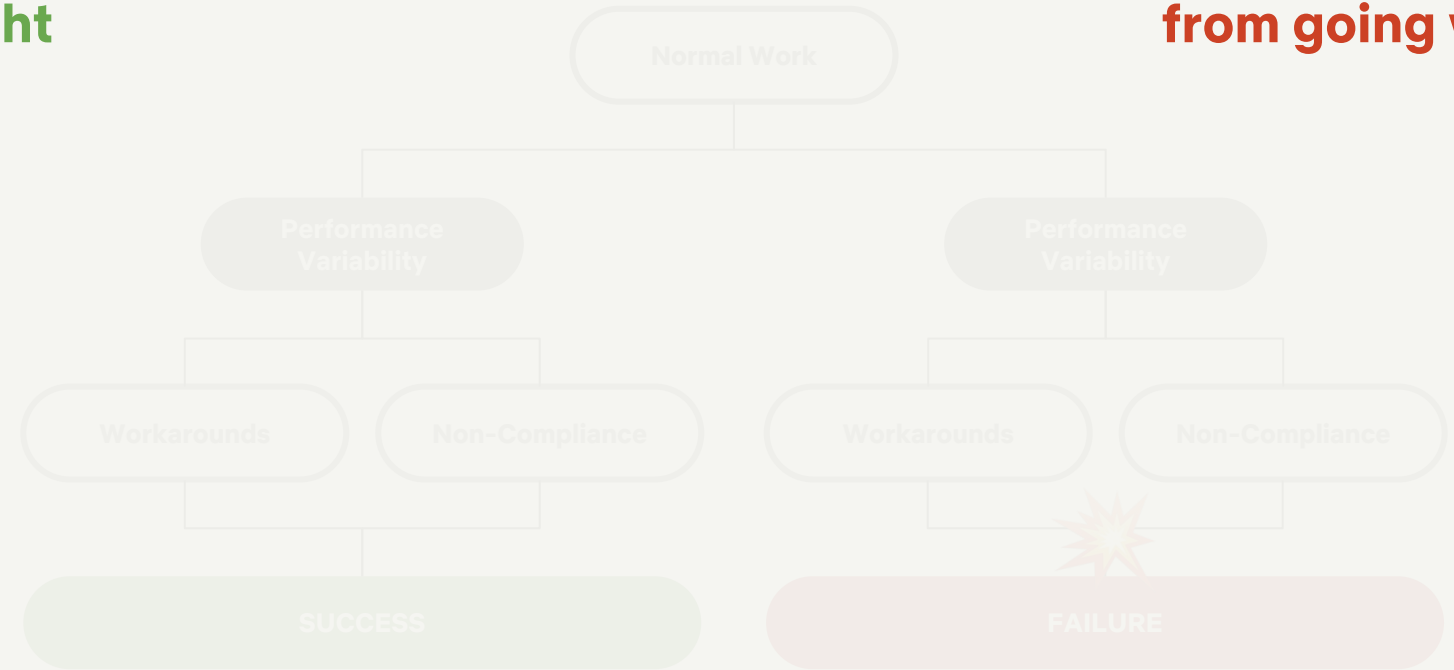
There is no root cause.

5 Whys is a competitive disadvantage.

@this_hits_home

Make sure that things go right

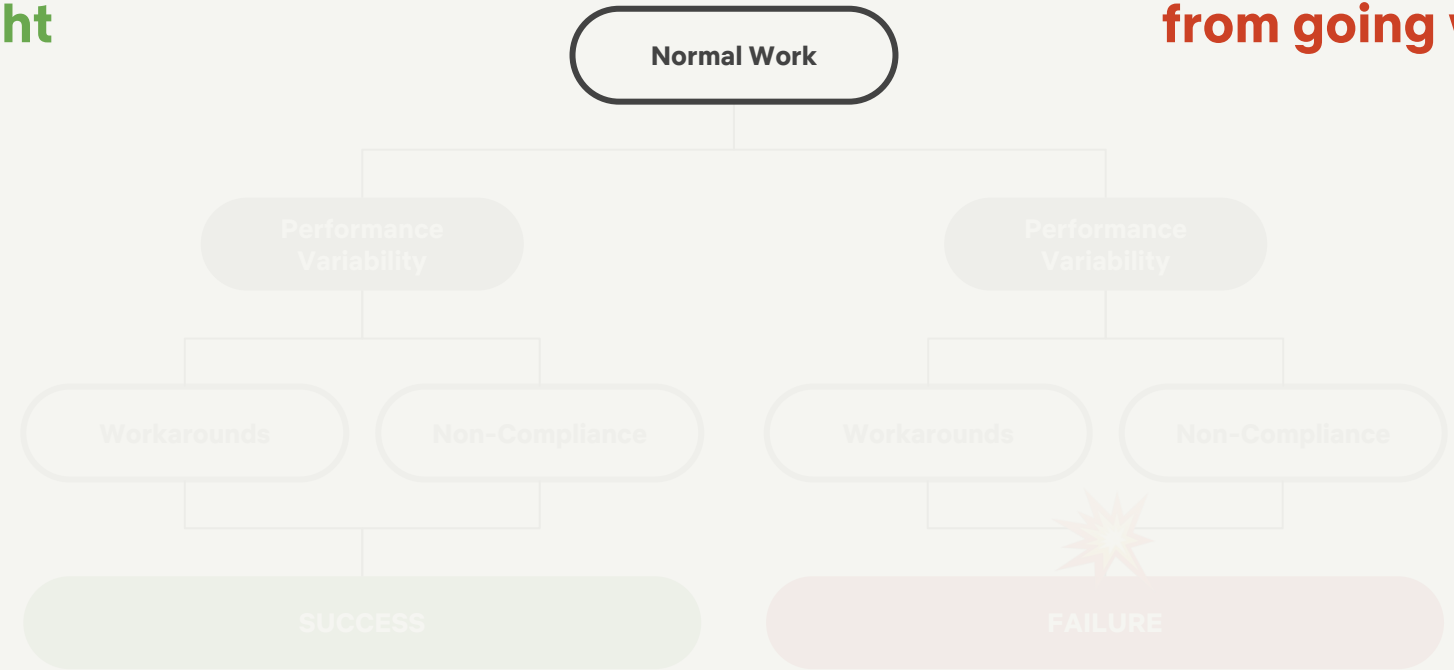
Rather than preventing them from going wrong



SAME PROCESS,
DIFFERENT OUTCOMES

Make sure that things go right

Rather than preventing them from going wrong

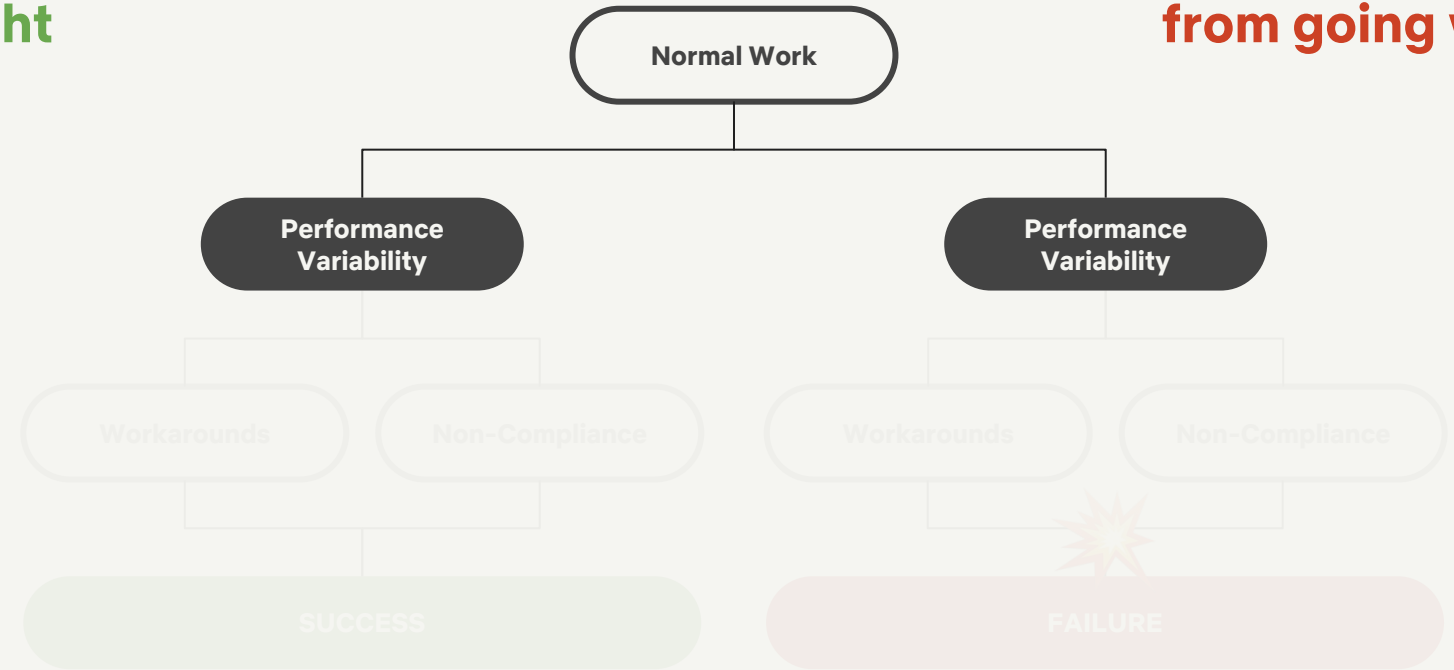


SAME PROCESS,
DIFFERENT OUTCOMES

@this_hits_home

Make sure that things go right

Rather than preventing them from going wrong

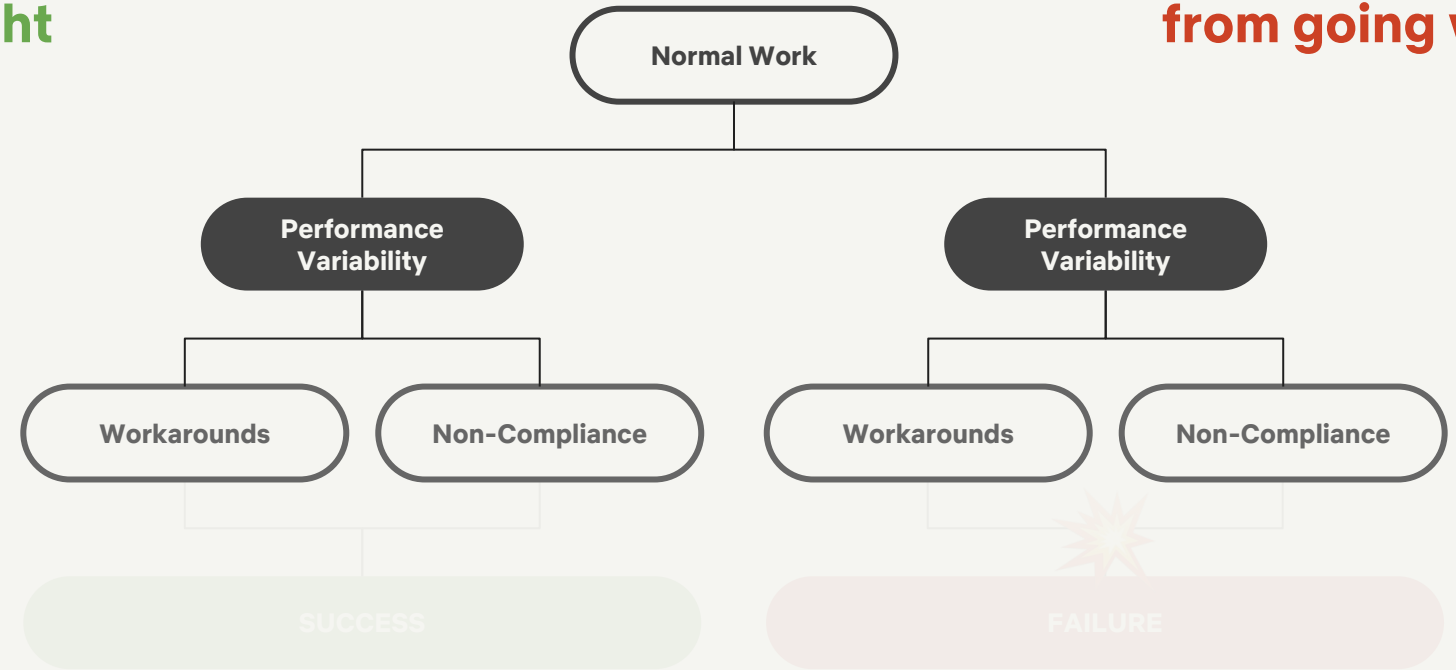


SAME PROCESS,
DIFFERENT OUTCOMES

@this_hits_home

Make sure that things go right

Rather than preventing them from going wrong

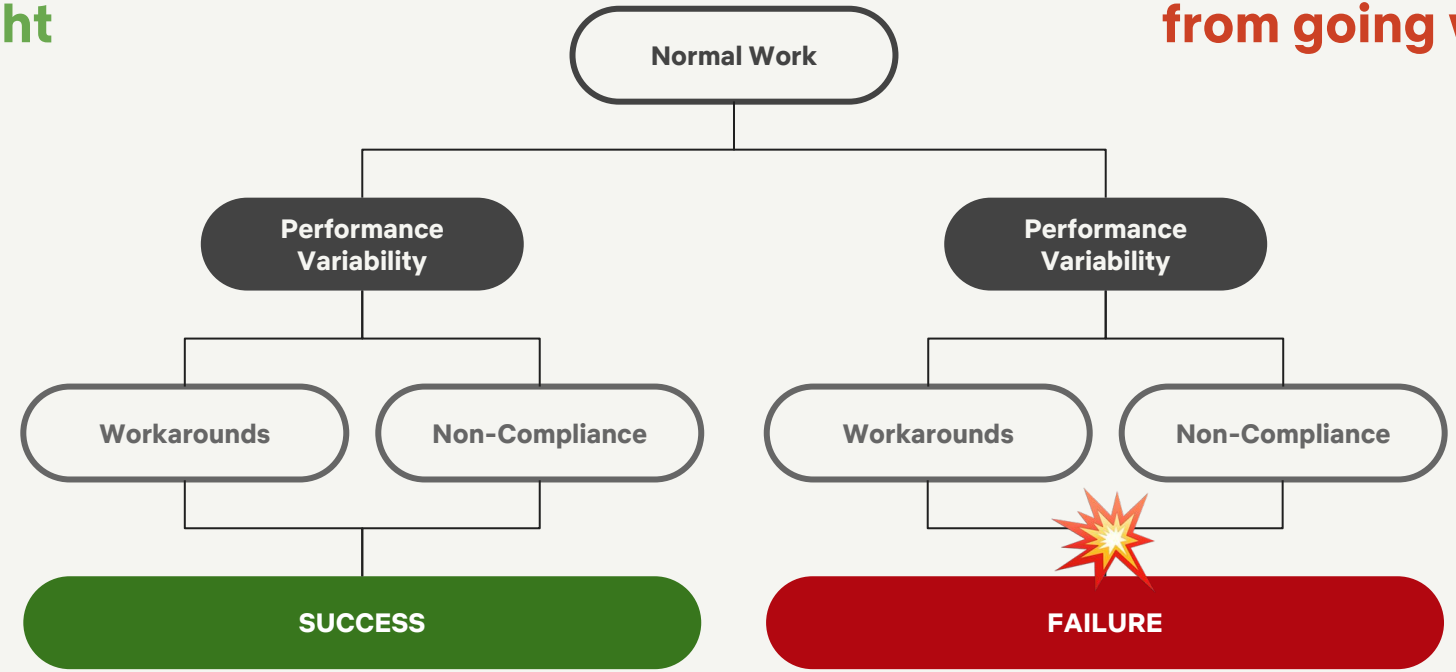


SAME PROCESS,
DIFFERENT OUTCOMES

@this_hits_home

Make sure that things go right

Rather than preventing them from going wrong



SAME PROCESS,
DIFFERENT OUTCOMES

@this_hits_home

Make sure that things go right

Rather than preventing them from going wrong

Normal Work



SAME PROCESS,
DIFFERENT OUTCOMES

@this_hits_home

Make sure that things go right

Rather than preventing them from going wrong

Normal Work



SUCCESS

FAILURE

SAME PROCESS,
DIFFERENT OUTCOMES

@this_hits_home

Make sure that things go right

Rather than preventing them from going wrong

Normal Work



FAILURE

SAME PROCESS,
DIFFERENT OUTCOMES

@this_hits_home

Constraining

‘PERFORMANCE VARIABILITY’

in order to remove failures **will also remove**

SUCCESSFUL EVERYDAY WORK!



Kitchens
@this_hits_home

what if the real success was the incidents we made along the way

9:22 PM - 12 Jan 2019



Kitchens
@this_hits_home

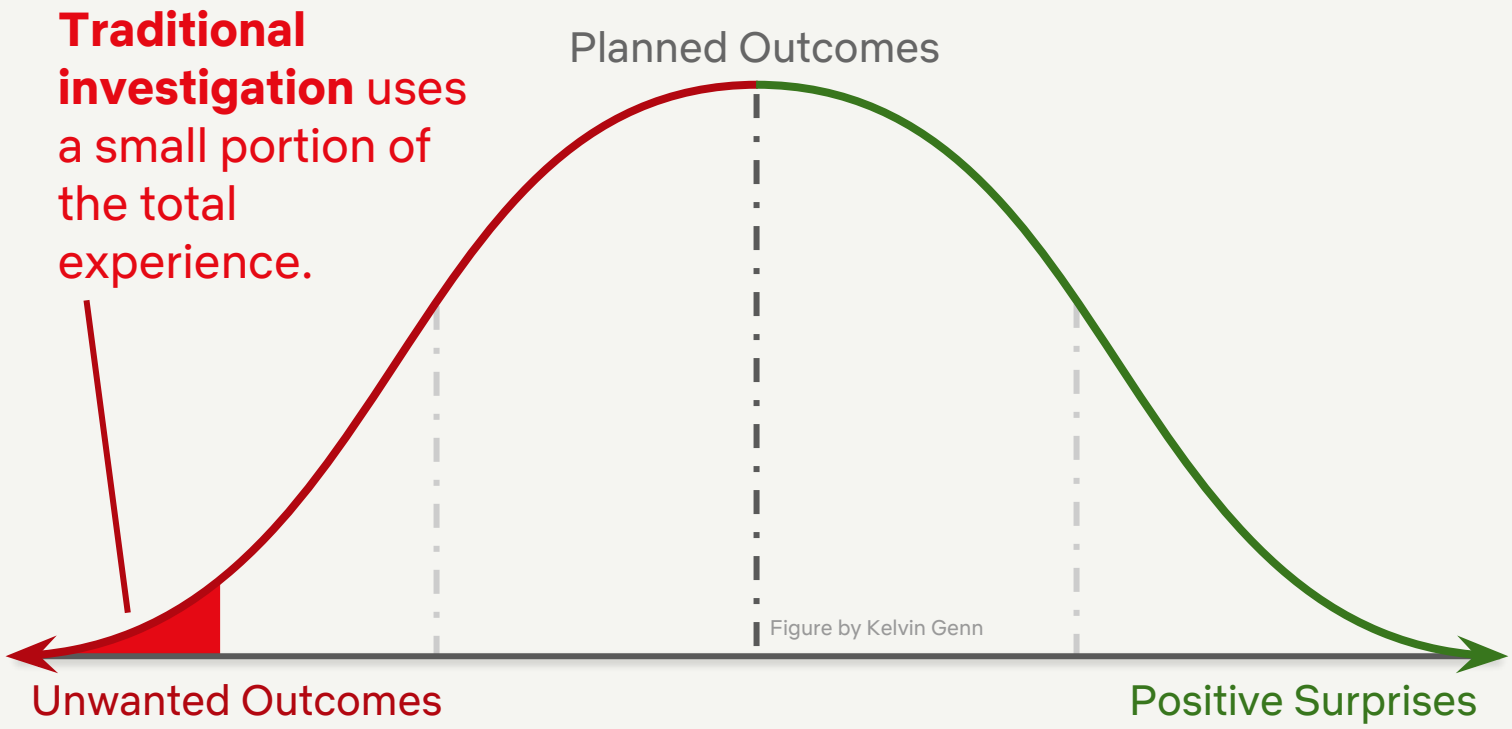
what if the real success was the incidents we made along the way

9:22 PM - 12 Jan 2019

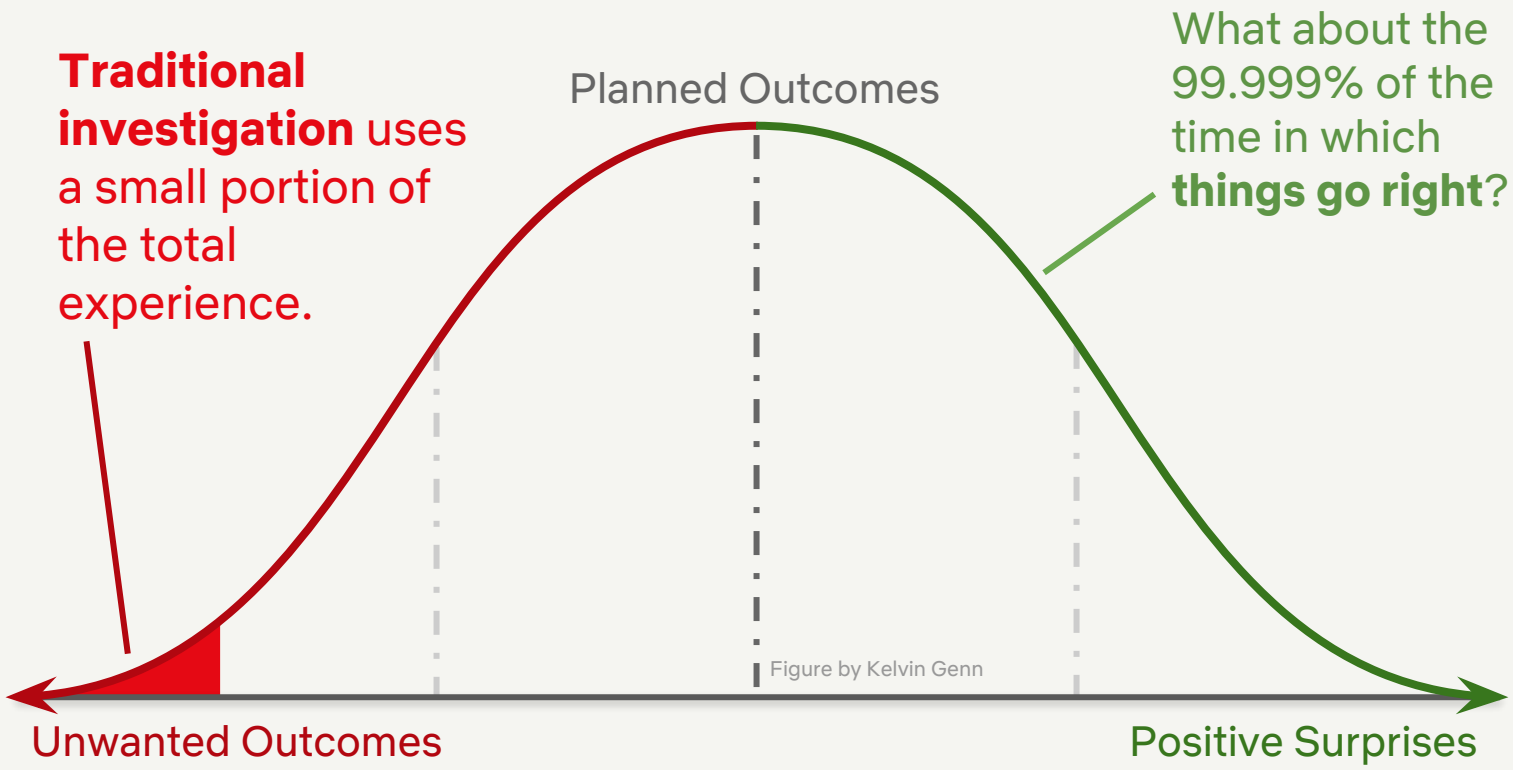




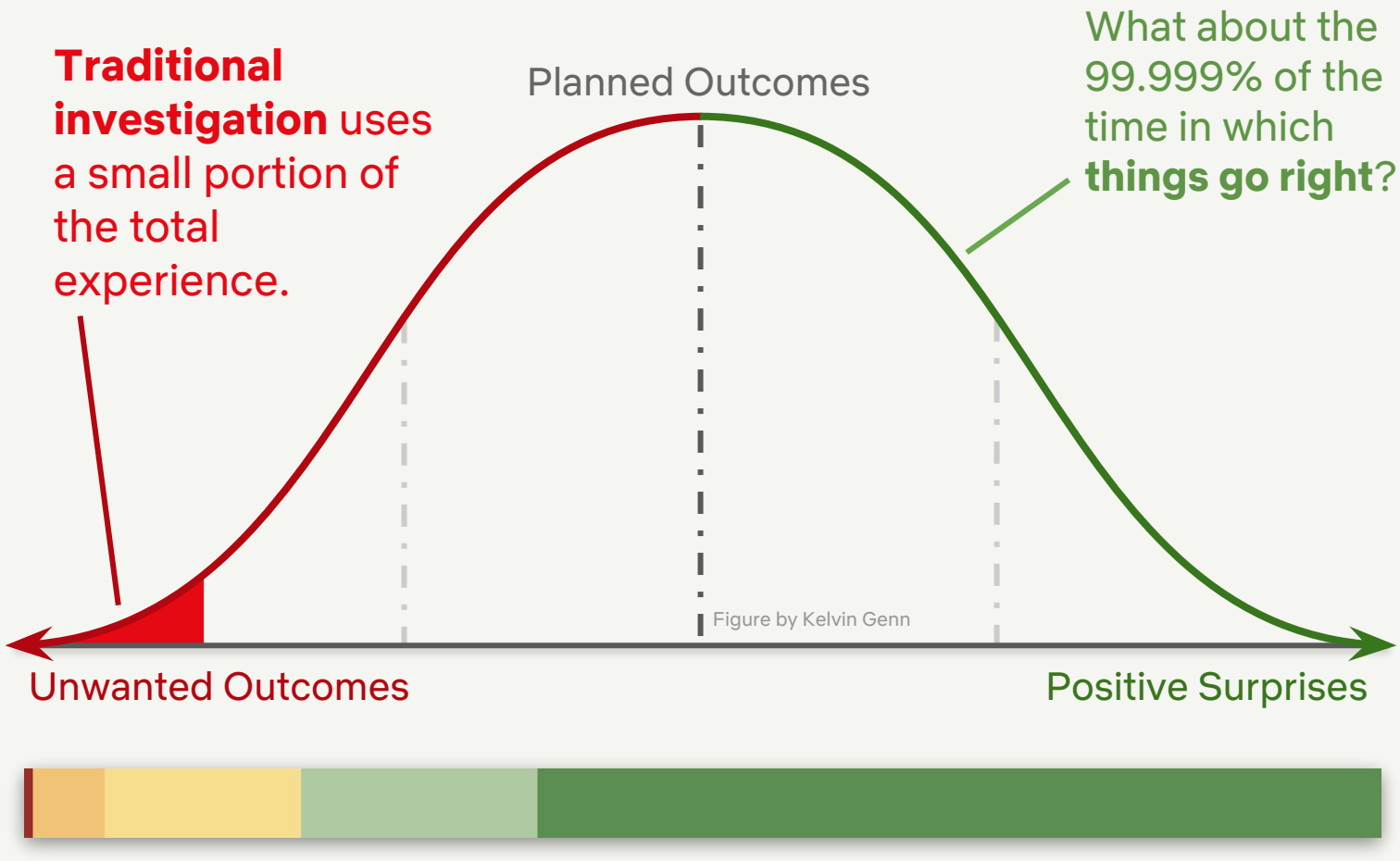
@this_hits_home



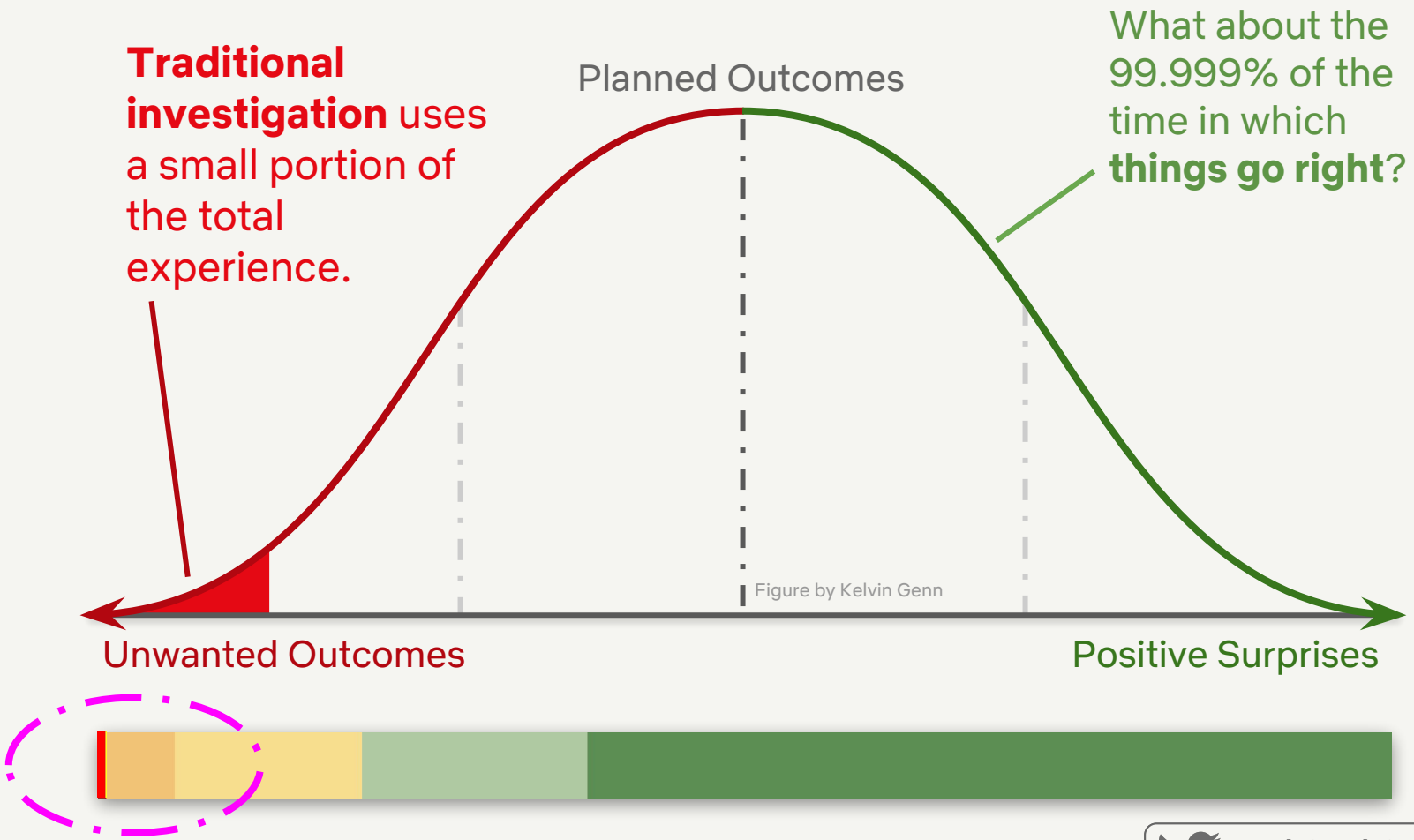
@this_hits_home



@this_hits_home



@this_hits_home



@this_hits_home

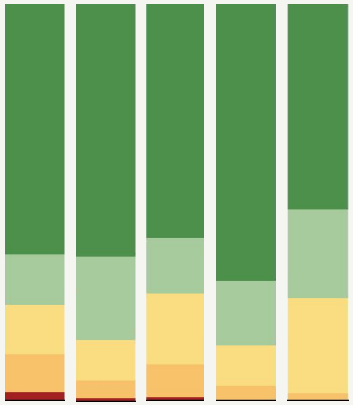


@this_hits_home



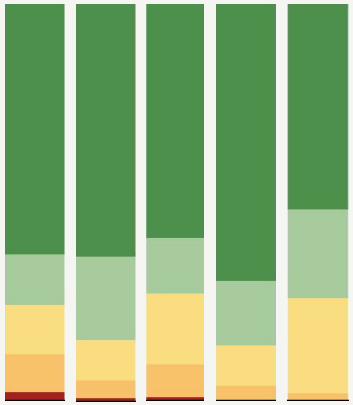
@this_hits_home





@this_hits_home

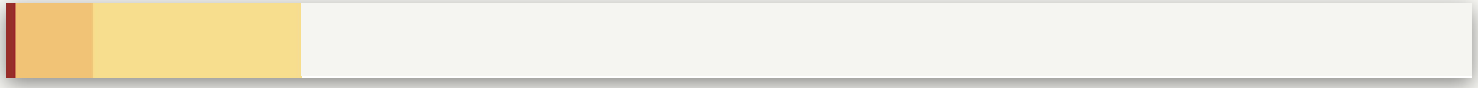




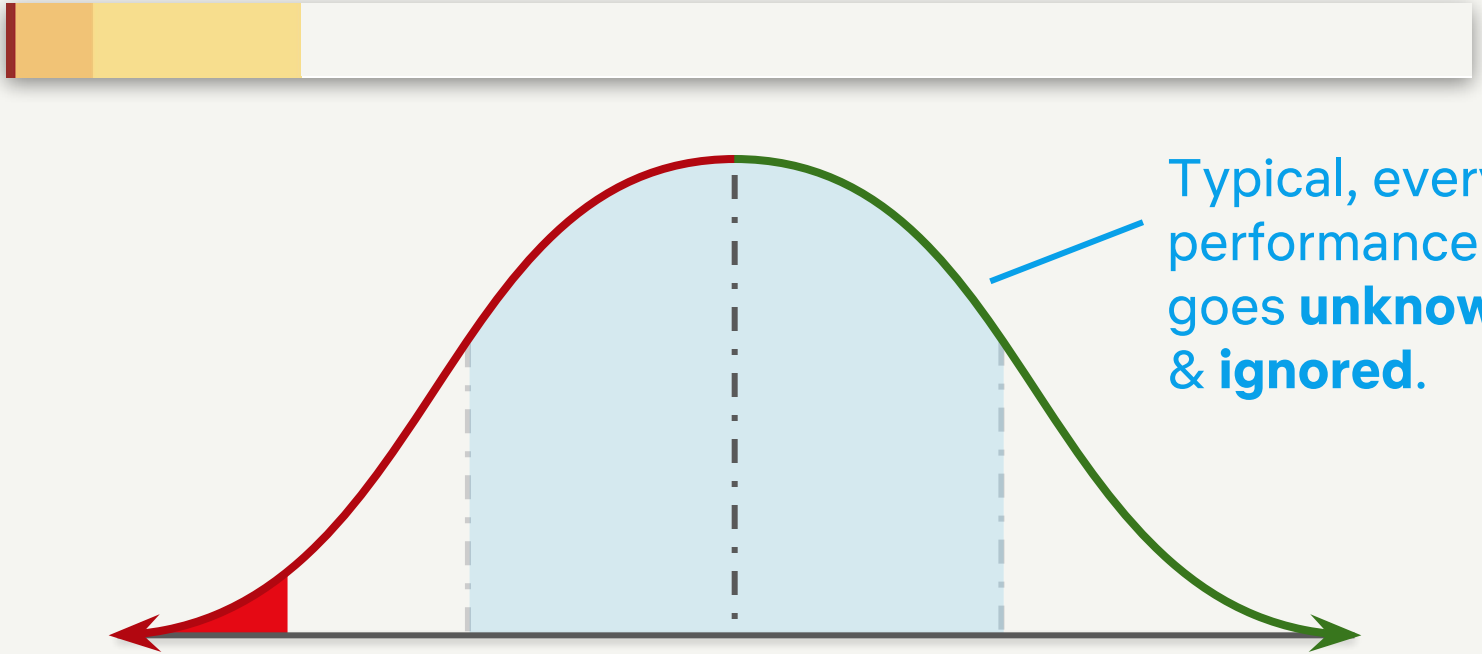
Availability is perceived.

@this_hits_home

Success is often **invisible**



Success is often **invisible**



@this_hits_home

HOW DO WE

SUCCESS AND FAILURE
IS NOT REALLY A DUALITY



@this_hits_home

HOW DO WE

Keep **normal** work **normal**

SUCCESS AND FAILURE
IS NOT REALLY A DUALITY



@this_hits_home

HOW DO WE

Keep **normal** work **normal**

AND

KEEP INCIDENTS WEIRD?

SUCCESS AND FAILURE
IS NOT REALLY A DUALITY



@this_hits_home

PROD ALERT / **INC-1806**

Signups Impact Triggered By Website Going Cattywampus

DETAILS

Type	Incident
Regions	Global
Severity	Major
Signup Impact?	Yes
Notifications	Yes
Source of Detection	Paging Alert

ACTIVITY

Status	STABLE
Assignee	Ryan Kitchens
Watchers	0

DATES

Created	03/25/2019 1:45 PM
Updated	03/25/2019 1:50 PM
Incident Start	03/25/2019 1:40 PM
Incident Detected	03/25/2019 1:41 PM
Stabilization Time	03/25/2019 1:48 PM

DESCRIPTION

We were paged for an abnormal RPS trend which kicked off the investigation. Signup flows were depressed in conjunction with a spike in traffic to the website that sent Node.js into a ‘death spiral’.

During this time, customers would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.

DO YOUR INCIDENTS GET FILED
OR DO THEY GET READ?



PROD ALERT / **INC-1806**

Signups Impact Triggered By Website Going Cattywampus

DETAILS

Type	Incident
Regions	Global
~~**Severity**~~	~~Major~~
Signup Impact?	Yes
Notifications	Yes
Source of Detection	Paging Alert

ACTIVITY

Status	STABLE
Assignee	Ryan Kitchens
Watchers	0

DESCRIPTION

We were paged for an abnormal RPS trend which kicked off the investigation. Signup flows were depressed in conjunction with a spike in traffic to the website that sent Node.js into a ‘death spiral’.

During this time, customers would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.

DATES

Created	03/25/2019 1:45 PM
Updated	03/25/2019 1:50 PM
Incident Start	03/25/2019 1:40 PM
Incident Detected	03/25/2019 1:41 PM
Stabilization Time	03/25/2019 1:48 PM

DO YOUR INCIDENTS GET FILED
OR DO THEY GET READ?

@this_hits_home

PROD ALERT / **INC-1806**

Signups Impact Triggered By Website Going Cattywampus

DETAILS

Type	Incident
Regions	Global
~~**Severity**~~	~~Major~~
Signup Impact?	Yes
Notifications	Yes
Source of Detection	Paging Alert

ACTIVITY

Status	STABLE
Assignee	Ryan Kitchens
Watchers	0

DESCRIPTION

We were paged for an abnormal RPS trend which kicked off the investigation. Signup flows were depressed in conjunction with a spike in traffic to the website that sent Node.js into a ‘death spiral’.

During this time, customers would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.

DATES

Created	03/25/2019 1:45 PM
Updated	03/25/2019 1:50 PM
Incident Start	03/25/2019 1:40 PM
Incident Detected	03/25/2019 1:41 PM
Stabilization Time	03/25/2019 1:48 PM

DO YOUR INCIDENTS GET FILED
OR DO THEY GET READ?

@this_hits_home

PROD ALERT / **INC-1806**

Signups Impact Triggered By Website Going Cattywampus

DETAILS

Type	Incident
Regions	Global
~~**Severity**~~	~~Major~~
Signup Impact?	Yes
Notifications	Yes
Source of Detection	Paging Alert

ACTIVITY

Status	STABLE
Assignee	Ryan Kitchens
Watchers	0

DESCRIPTION

We were paged for an abnormal RPS trend which kicked off the investigation. **Signup flows were depressed** in conjunction with a spike in traffic to the website that sent Node.js into a ‘death spiral’.

During this time, customers would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.

DATES

Created	03/25/2019 1:45 PM
Updated	03/25/2019 1:50 PM
Incident Start	03/25/2019 1:40 PM
Incident Detected	03/25/2019 1:41 PM
Stabilization Time	03/25/2019 1:48 PM

DO YOUR INCIDENTS GET FILED
OR DO THEY GET READ?

@this_hits_home

PROD ALERT / **INC-1806**

Signups Impact Triggered By Website Going Cattywampus

DETAILS

Type	Incident
Regions	Global
~~**Severity**~~	~~Major~~
Signup Impact?	Yes
Notifications	Yes
Source of Detection	Paging Alert

ACTIVITY

Status	STABLE
Assignee	Ryan Kitchens
Watchers	0

DESCRIPTION

We were paged for an abnormal RPS trend which kicked off the investigation. Signup flows were depressed in conjunction with a spike in traffic to the website that sent Node.js into a **‘death spiral’**.

During this time, customers would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.

DATES

Created	03/25/2019 1:45 PM
Updated	03/25/2019 1:50 PM
Incident Start	03/25/2019 1:40 PM
Incident Detected	03/25/2019 1:41 PM
Stabilization Time	03/25/2019 1:48 PM

DO YOUR INCIDENTS GET FILED
OR DO THEY GET READ?

@this_hits_home

PROD ALERT / **INC-1806**

Signups Impact Triggered By Website Going Cattywampus

DETAILS

Type	Incident
Regions	Global
~~**Severity**~~	~~Major~~
Signup Impact?	Yes
Notifications	Yes
Source of Detection	Paging Alert

ACTIVITY

Status	STABLE
Assignee	Ryan Kitchens
Watchers	0

DESCRIPTION

We were paged for an abnormal RPS trend which kicked off the investigation. Signup flows were depressed in conjunction with a spike in traffic to the website that sent Node.js into a ‘death spiral’.

During this time, **customers** would have been experiencing errors when attempting to load the site.

We are still looking into the nature of the requests to surface any patterns that may be helpful.

DATES

Created	03/25/2019 1:45 PM
Updated	03/25/2019 1:50 PM
Incident Start	03/25/2019 1:40 PM
Incident Detected	03/25/2019 1:41 PM
Stabilization Time	03/25/2019 1:48 PM

DO YOUR INCIDENTS GET FILED
OR DO THEY GET READ?

@this_hits_home

PROD ALERT / **INC-1806**

Signups Impact Triggered By Website Going Cattywampus

DETAILS

Type	Incident
Regions	Global
~~**Severity**~~	~~Major~~
Signup Impact?	Yes
Notifications	Yes
Source of Detection	Paging Alert

ACTIVITY

Status	STABLE
Assignee	Ryan Kitchens
Watchers	0

DESCRIPTION

We were paged for an abnormal RPS trend which kicked off the investigation. Signup flows were depressed in conjunction with a spike in traffic to the website that sent Node.js into a ‘death spiral’.

During this time, customers would have been experiencing errors when attempting to load the site.

DATES

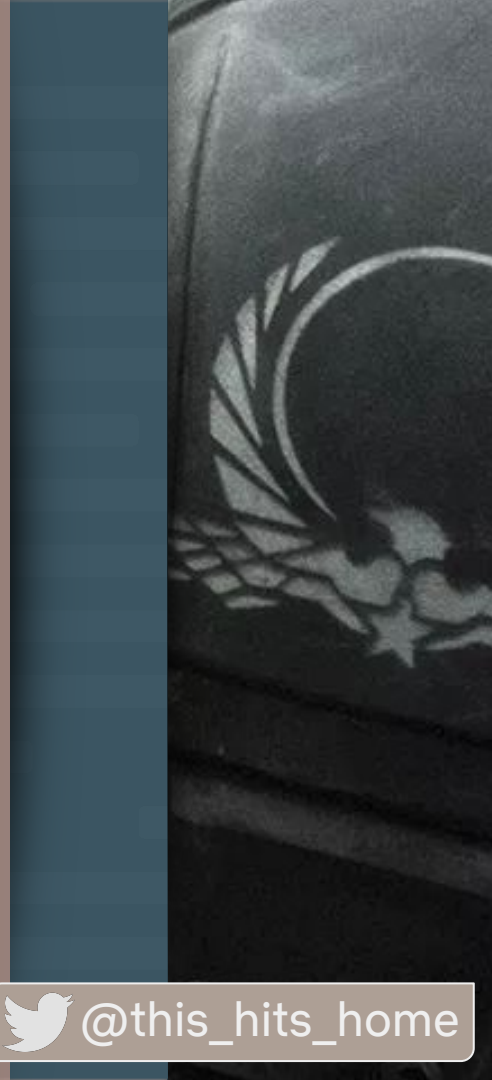
Created	03/25/2019 1:45 PM
Updated	03/25/2019 1:50 PM
Incident Start	03/25/2019 1:40 PM
Incident Detected	03/25/2019 1:41 PM
Stabilization Time	03/25/2019 1:48 PM

We are still looking into the nature of the requests to surface any patterns that may be helpful.

DO YOUR INCIDENTS GET FILED
OR DO THEY GET READ?

@this_hits_home

Contributors & Enablers
Mitigators
Risks
Difficulties in Handling
Follow-up Items
Artifacts
Timeline
References
Open Questions



@this_hits_home

Contributors & Enablers

Mitigators

Risks

Difficulties in Handling

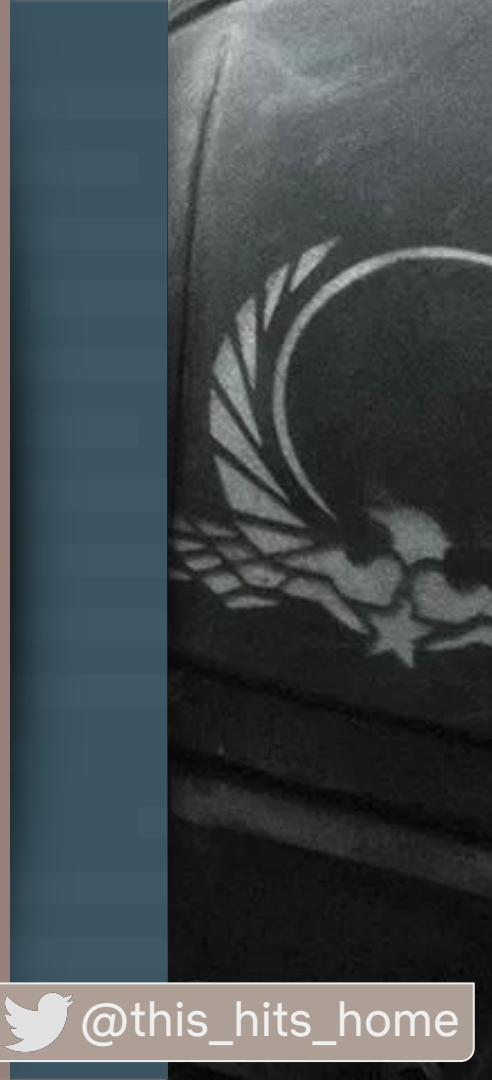
Follow-up Items

Artifacts

Timeline

References

Open Questions



@this_hits_home

Contributors & Enablers

Mitigators

Risks

Difficulties in Handling

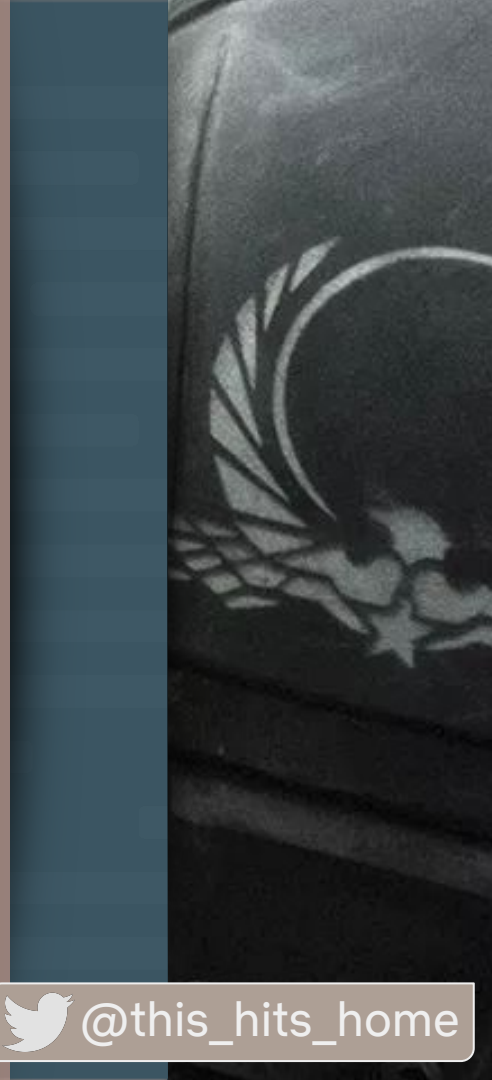
Follow-up Items

Artifacts

Timeline

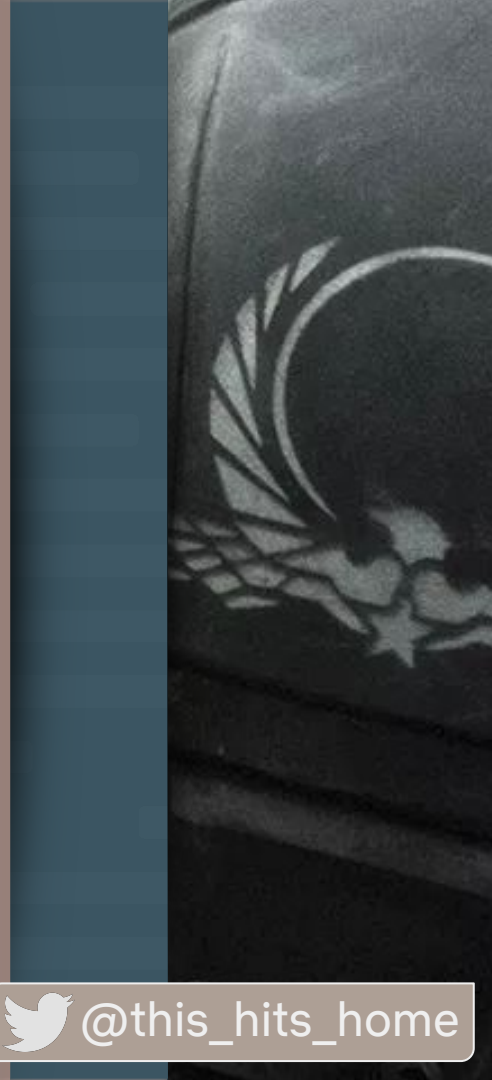
References

Open Questions



@this_hits_home

Contributors & Enablers
Mitigators
Risks
Difficulties in Handling
Follow-up Items
Artifacts
Timeline
References
Open Questions



@this_hits_home

Risks

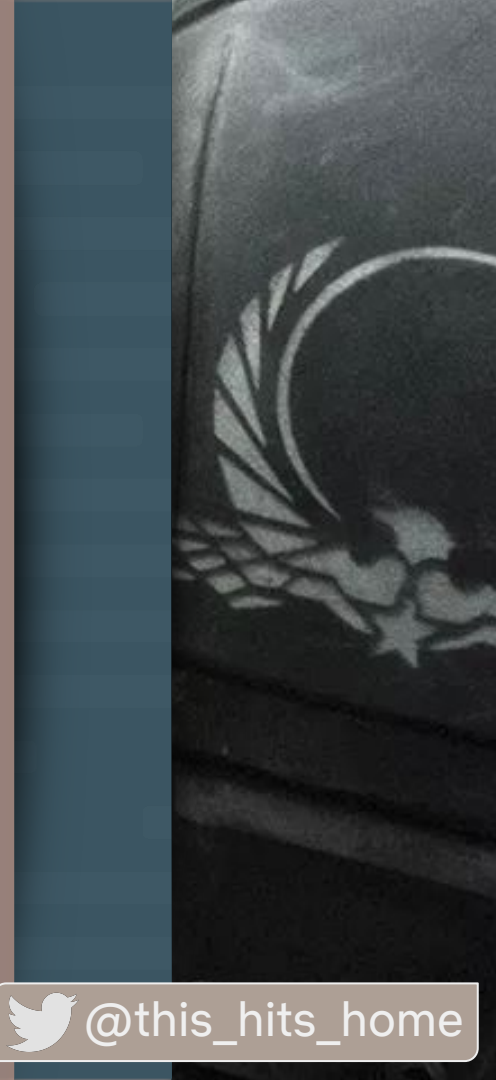
Division of labor across teams.
Usage of deprecated property UI over the recommended one.
No type safety or constraints were in place for the property.
The team had confusion about how to revert the property.
Property parser is overly optimistic of intended context.
Clearing massive amounts of individual rules is not trivial.
Afterwards, we still observed short-circuiting on redemption.
No feedback to users on effects of removing rules en masse.

Mitigators

SR rule to handle broken headers.
Only regions in peak traffic window observed the issue.
Website clusters auto-scaled as expected to meet demand.
An old ASG was able to be repurposed into additional capacity.
Rate limiting was observed via recently created dashboards.
New ASG was deployed which patched a bug in the validation.
All relevant teams were alerted to the issue in under one minute.
An engineer discovered the property change by searching events.

Contributors & Enablers

An unstaged, global property changed was performed.
Large volume of redemptions during the holiday window.
Invalid rules added to redemption workflows.
Removal of rules contributed to an increase of traffic.
Validate commands started failing in mid-tier APIs.
ID validation was not working as expected.
Recent move to new instance types had unexpected perf.



@this_hits_home

Risks

Use of deprecated UI over the recommended one.

No type safety or property constraints.

Division of labor across teams.

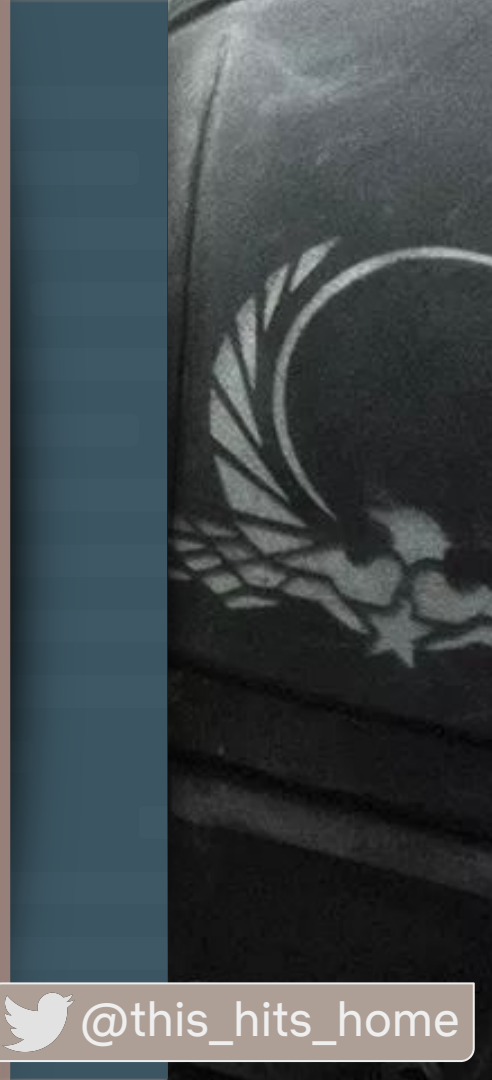
The team had confusion about how to revert.

Property parser is overly optimistic of intended context.

Clearing massive amounts of individual rules is not easy.

Short-circuiting for redemption persisted long after.

No system feedback to users during mass rule removal.



@this_hits_home

Mitigators

Proxy rules added to handle broken headers.

Only peak traffic regions observed the issue.

Clusters eventually auto-scaled to meet demand.

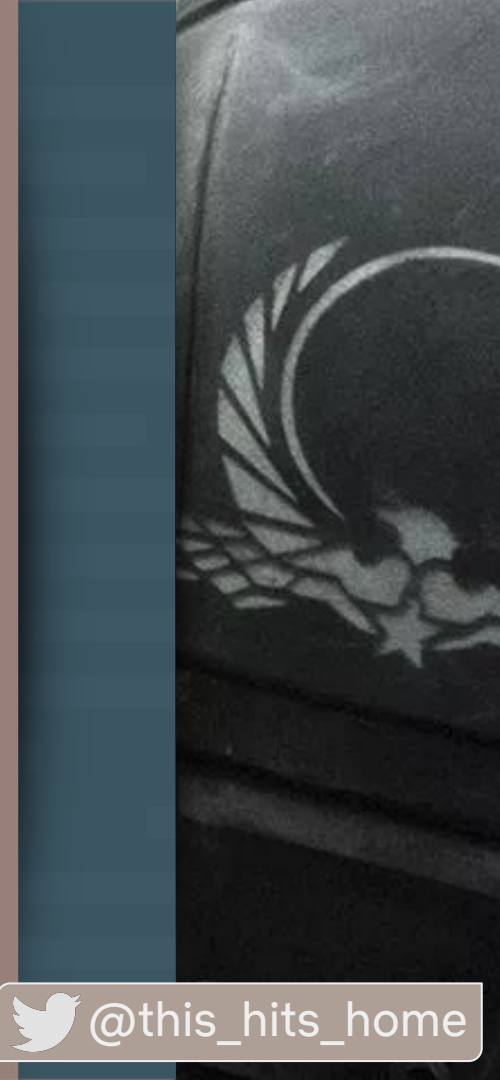
Old ASG was repurposed into additional capacity.

New ASG deployed to patch a bug in validation.

Rate-limiting observed via recently saved dashboard.

All relevant teams were alerted under 1 min to the issue.

A person discovered the change by searching events.



@this_hits_home

An unstaged, global property changed was performed.

Large volume of redemptions during holiday window.

Invalid rules added to redemption workflows.

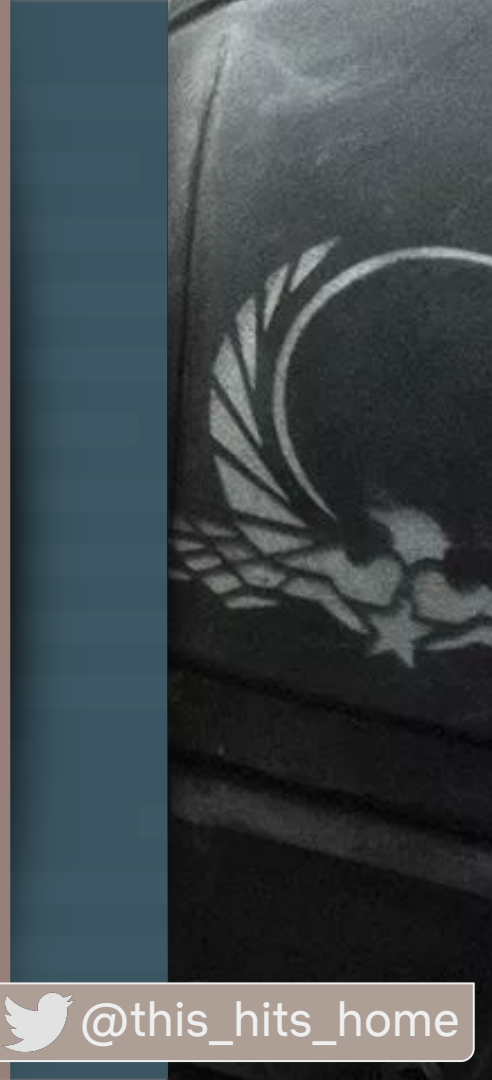
Removal of rules contributed to an increase of traffic.

Validate cmd started failing in mid-tier APIs.

Contributors & Enablers

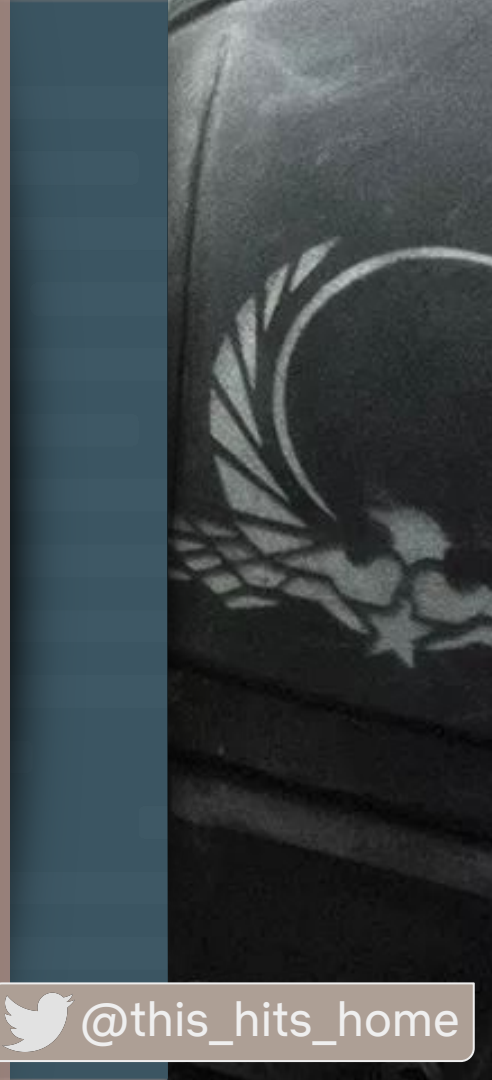
ID validation was not working as expected.

Unexpected perf from upgrade to new instance types.



@this_hits_home

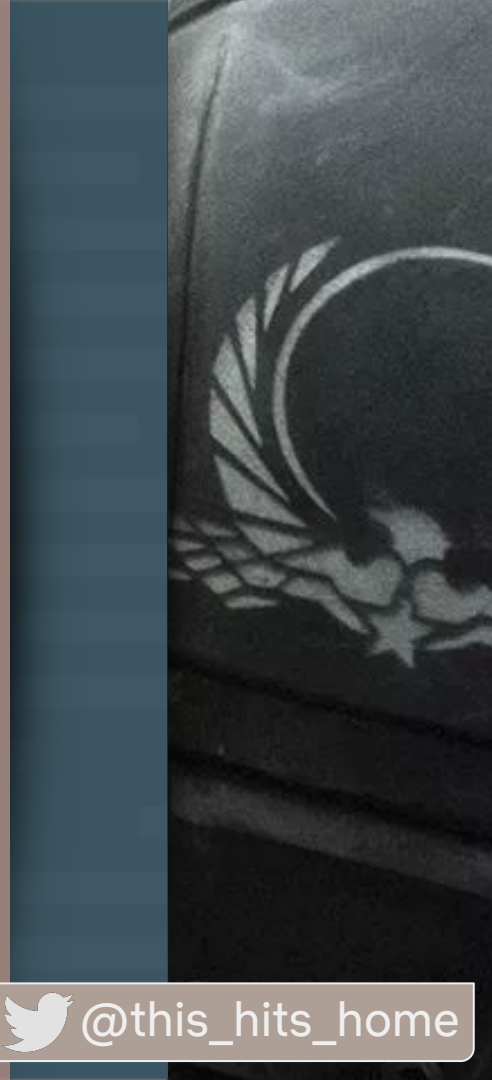
Contributors & Enablers
Mitigators
Risks
Difficulties in Handling
Follow-up Items
Artifacts
Timeline
References
Open Questions



@this_hits_home

Contributors & Enablers
Mitigators
Risks
Difficulties in Handling
Follow-up Items
Artifacts
Timeline
References
Open Questions

‘Islands of Knowledge’



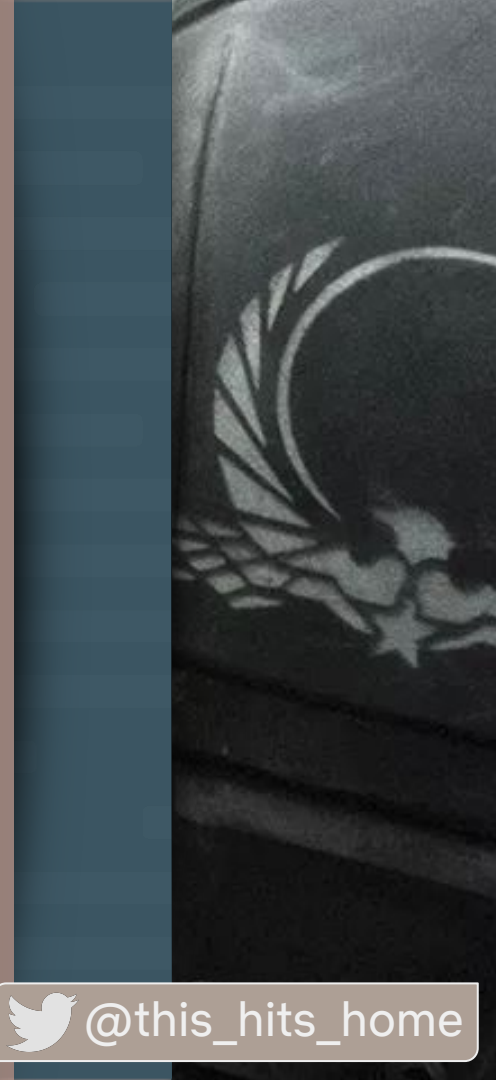
@this_hits_home

Contributors & Enablers
Mitigators
Risks
Follow-up Items
Artifacts
Timeline
References
Open Questions

“I didn’t know
it worked like that.”

Difficulties in Handling

‘Islands of Knowledge’



@this_hits_home

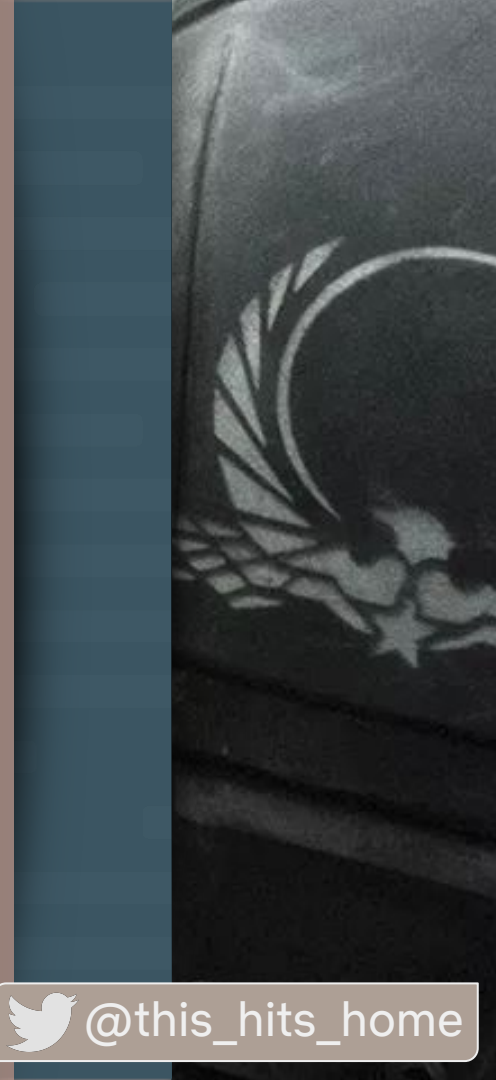
Contributors & Enablers
Mitigators
Risks
Follow-up Items
Artifacts
Timeline
References
Open Questions

“I didn’t know
it worked like that.”

Difficulties in Handling

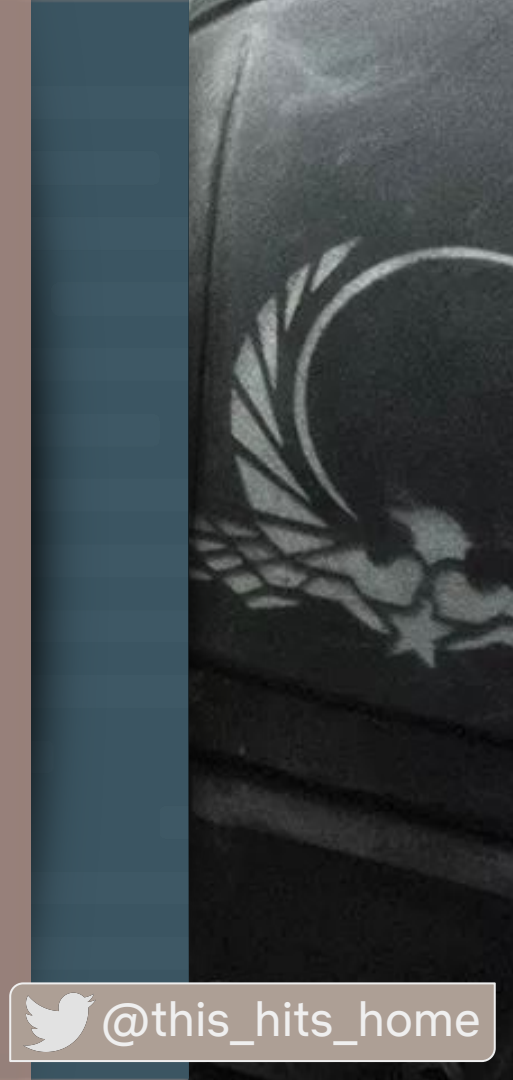
‘**Islands of Knowledge**’

“We didn’t make
any changes…
did we?”



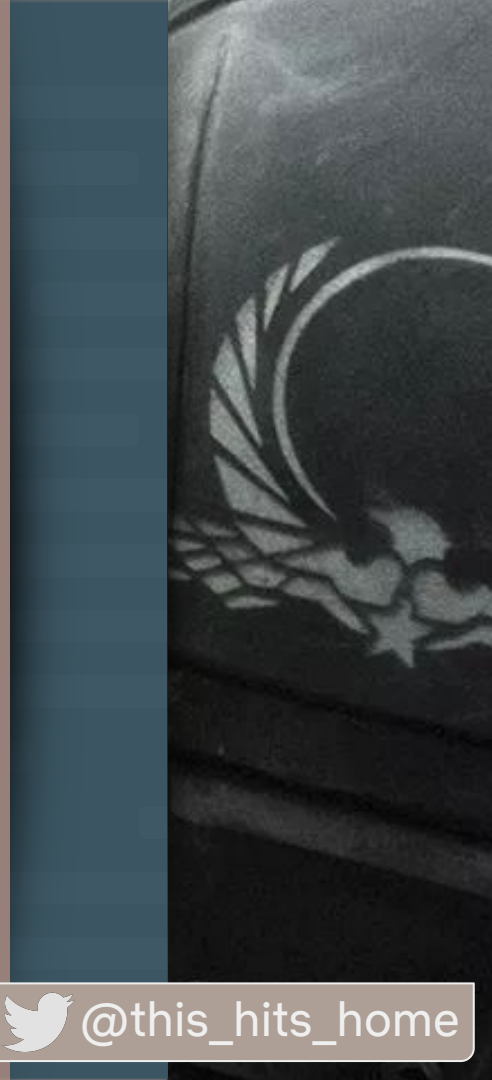
@this_hits_home

Contributors & Enablers
Mitigators
Risks
Difficulties in Handling
Follow-up Items
Artifacts
Timeline
References
Open Questions



@this_hits_home

Contributors & Enablers
Mitigators
Risks
Difficulties in Handling
Follow-up Items
Artifacts
Timeline
References
Open Questions



@this_hits_home

Contributors & Enablers

Mitigators

Risks

Difficulties

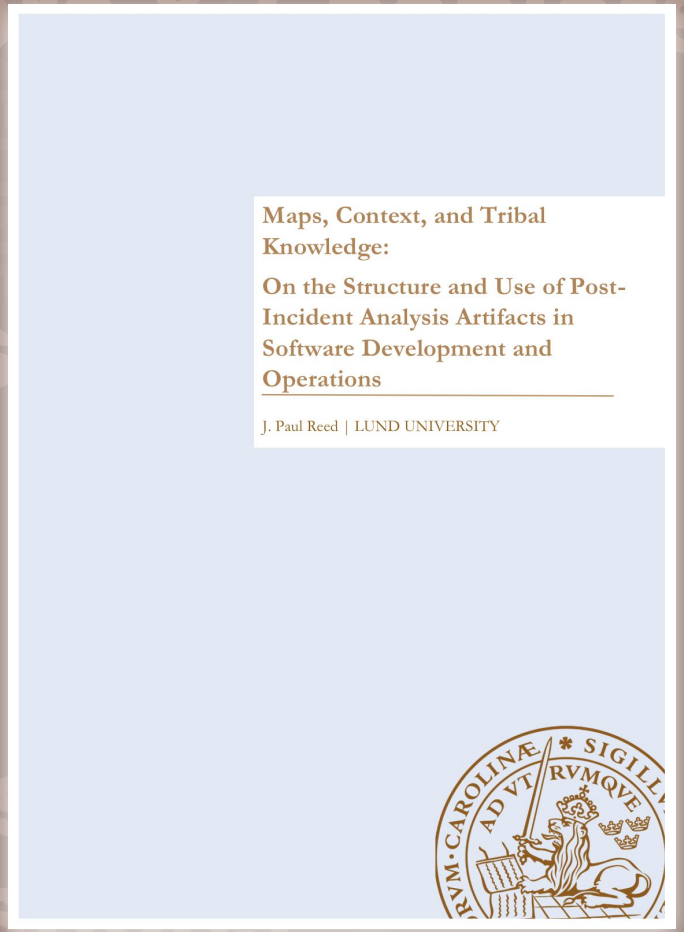
Follow-up

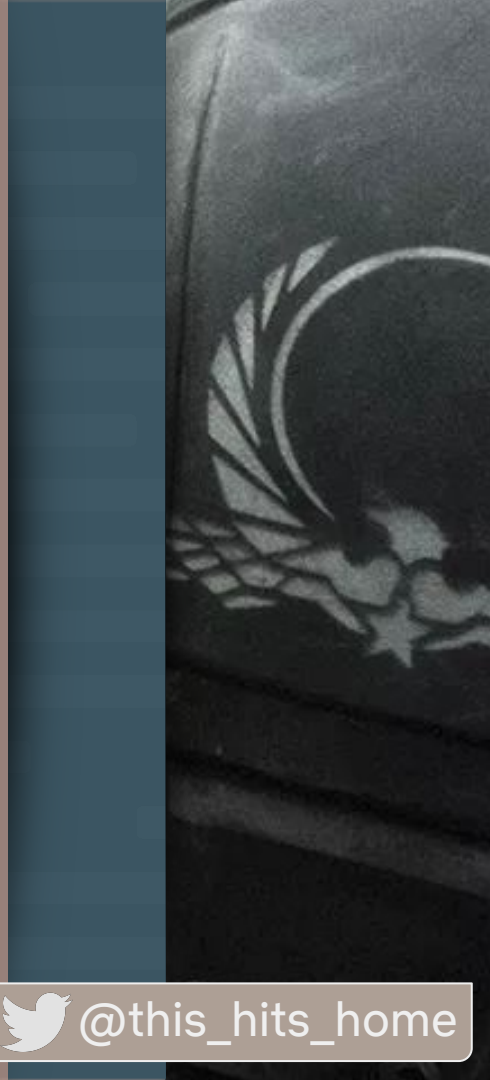
Artifacts

Timeline

References

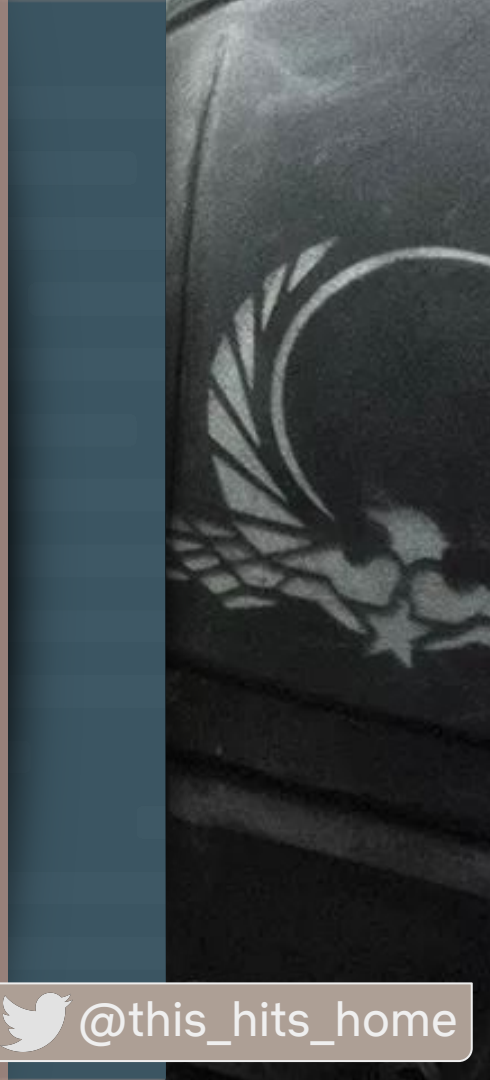
Open Questions





@this_hits_home

Contributors & Enablers
Mitigators
Risks
Difficulties in Handling
Follow-up Items
Artifacts
Timeline
References
Open Questions

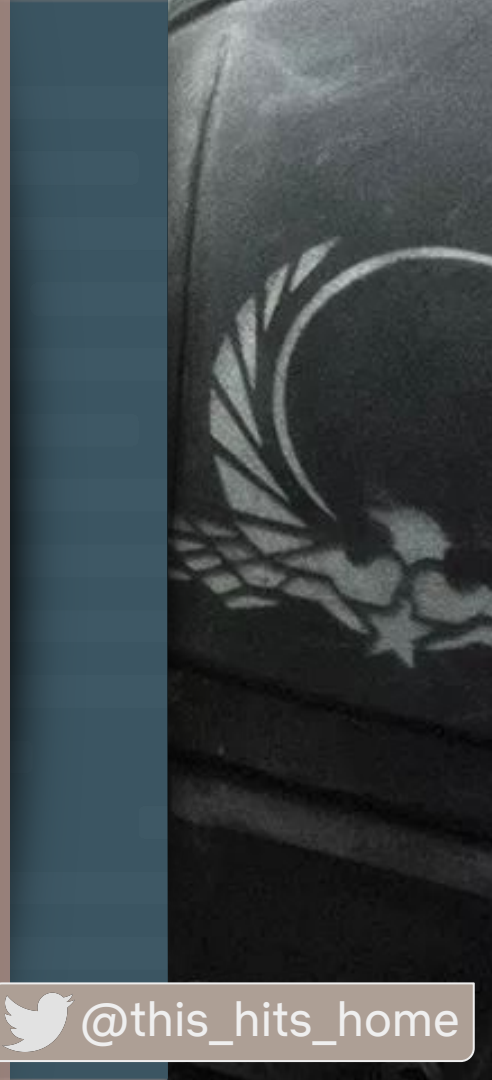


@this_hits_home

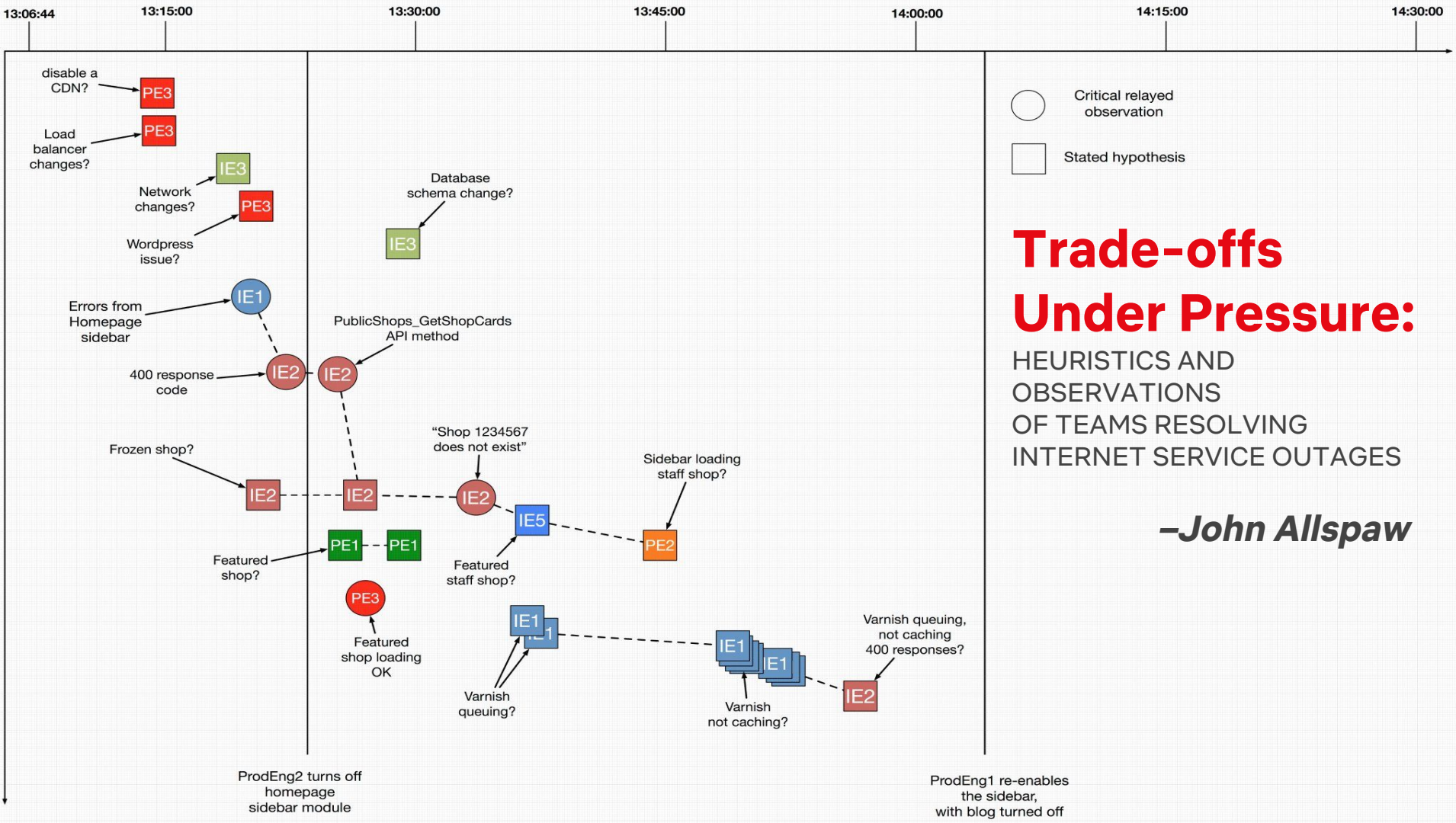
Contributors & Enablers
Mitigators
Risk
Diffi
Follo
Artifacts
References
Open Questions

16:40	CS contacts about SomeProblem begin.
18:44	First call about the issue is escalated
18:54	confirms there is an increase in CS call volume for the issue.
19:01	team tries to reproduce the problem and determines the problem might involve . . .
19:04	Engineer1 from Team escalates the issue
19:11	Engineer1 creates TICKET-XXX . . .
19:28	asks if SomeService is involved in the issue. Engineer1 suggests paging ServiceTeam instead.
19:33	Engineer2 from ServiceTeam responds.
19:35	Working theory at this point is that . . .

Timeline



@this_hits_home

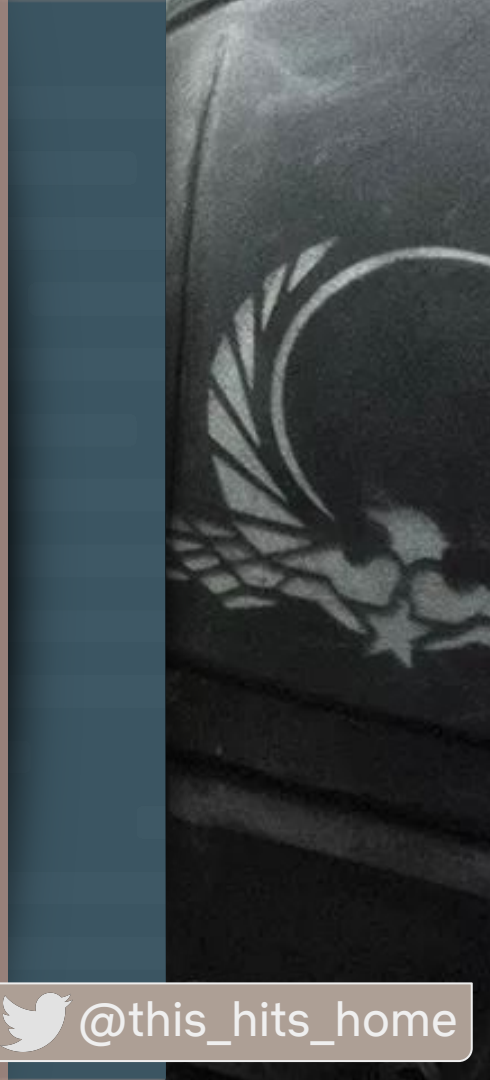


Trade-offs Under Pressure:

HEURISTICS AND OBSERVATIONS OF TEAMS RESOLVING INTERNET SERVICE OUTAGES

–John Allspaw

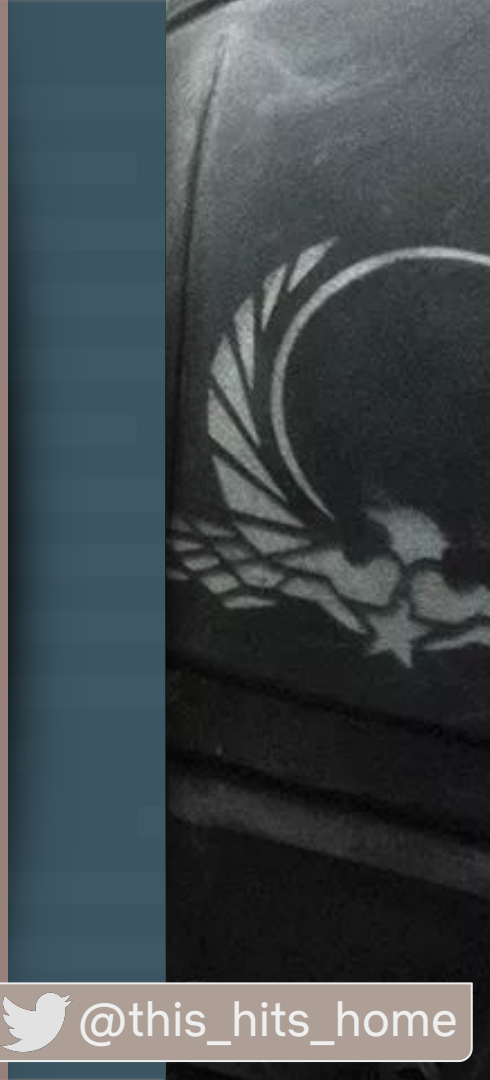
Contributors & Enablers
Mitigators
Risks
Difficulties in Handling
Follow-up Items
Artifacts
Timeline
References
Open Questions



@this_hits_home

Contributors & Enablers
Mitigators
Risks
Difficulties in Handling
Follow-up Items
Artifacts
Timeline
References

Open Questions



@this_hits_home

@this_hits_home

What's going on

WHEN IT SEEMS LIKE

NOTHING

is happening?

KEEP CALM NOTHING IS HAPPENING

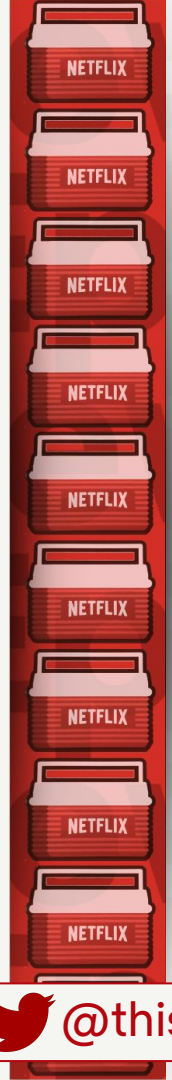
@this_hits_home

HOW HARD ARE

people working

JUST TO **KEEP THE SYSTEM**

HEALTHY?

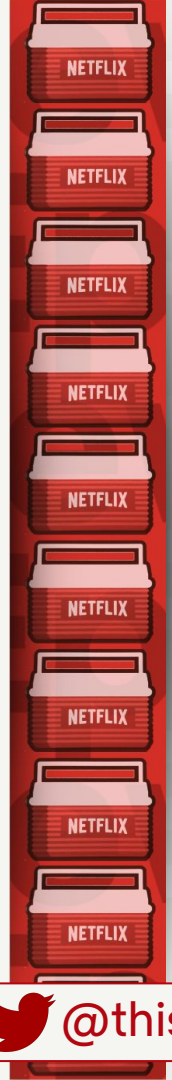


@this_hits_home

HOW HARD ARE

people working JUST TO KEEP THE SYSTEM HEALTHY?

How do you
THINK ABOUT
feeding this back UP THE CHAIN?



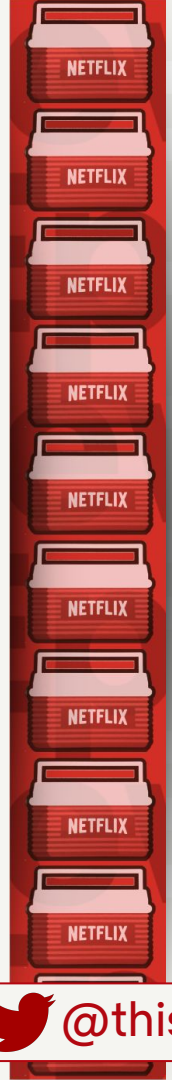
@this_hits_home

HOW HARD ARE

people working JUST TO KEEP THE SYSTEM HEALTHY?

How do you THINK ABOUT **feeding this back** UP THE CHAIN?

Or do you NOT BECAUSE EVERYTHING **looks good?**



@this_hits_home



THE HARDEST PROBLEM IN TECH

YOU HAVE to talk TO PEOPLE!

THERE IS NO GETTING AROUND THIS

@this_hits_home

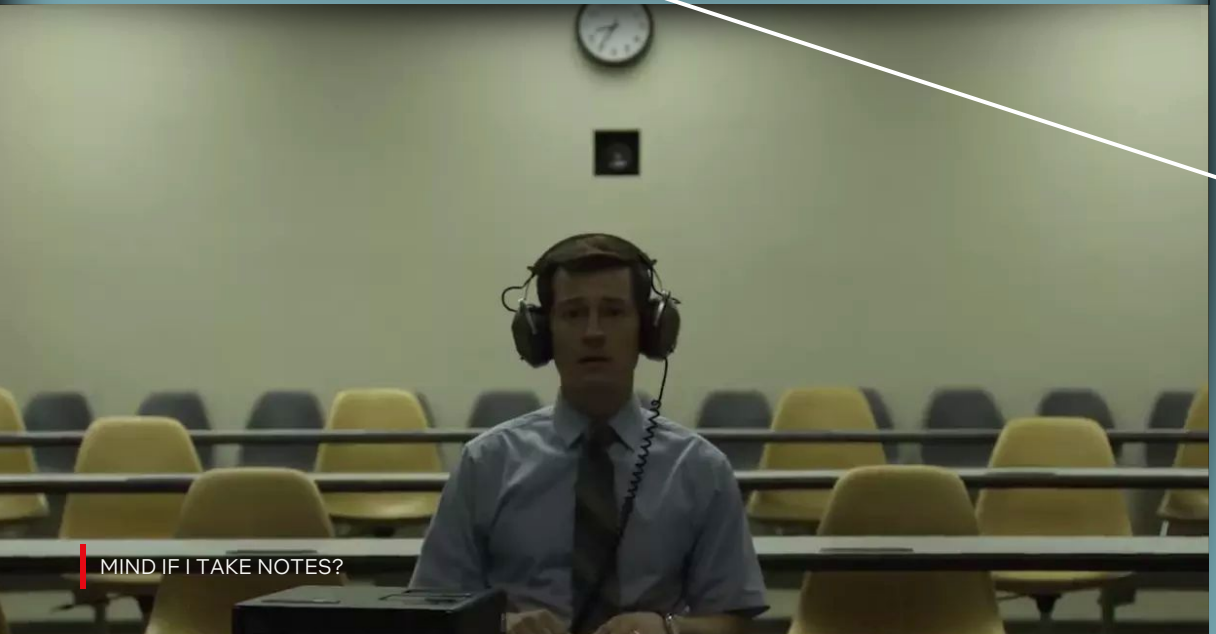
Interviews

GETTING FROM
‘why did things go wrong?’
TO **‘how did things go right?’**

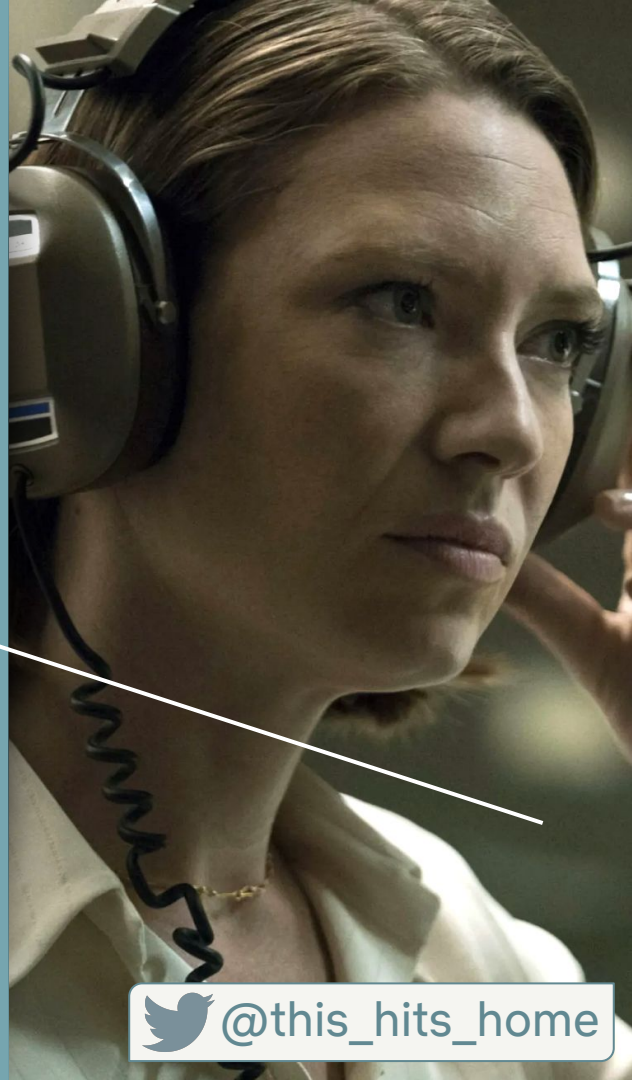
EVERY EXPERIENCE
IS VALID & ESSENTIAL







MIND IF I TAKE NOTES?



@this_hits_home



Kitchens
@this_hits_home

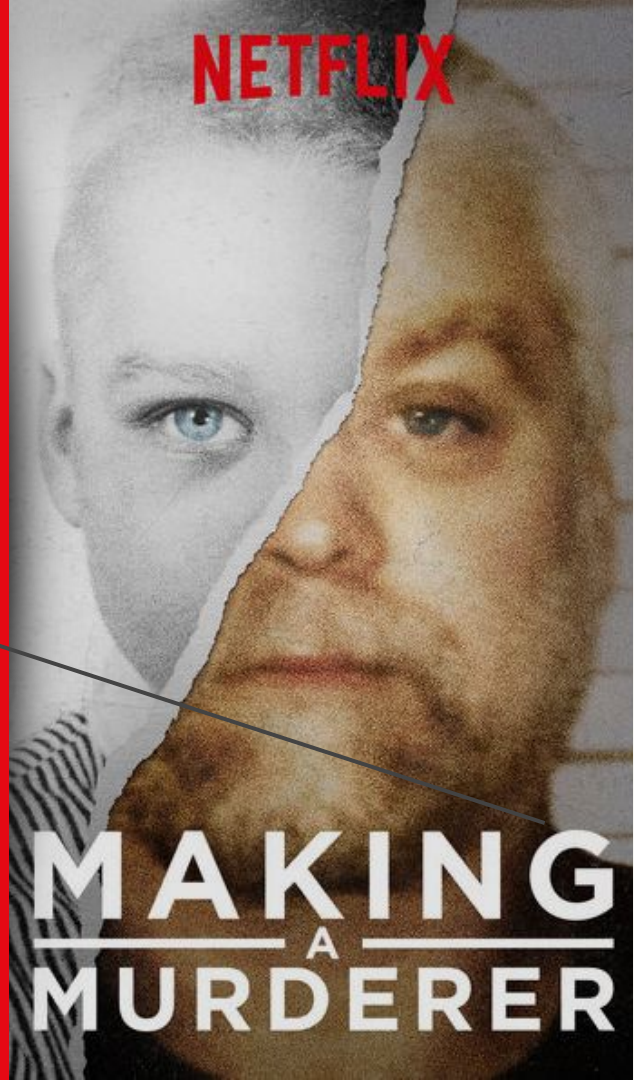
if you say 'root cause' you are a cop

9:53 AM - 5 Mar 2019






UP FOR A QUICK CHAT?




Learning
TEAMS
HOW DO TEAMS ADAPT SUCCESSFULLY?
@this_hits_home


ASK
BETTER
QUESTIONS
@this_hits_home

Interviews

elicit descriptions
to construct **how we got here**.

try to convey **what the world looked like** from their
perspective.



How we respond is important.

"BEST PRACTICES"
ARE NOT REAL PRACTICES

"WE MUST...

Be more careful.

Avoid making mistakes.

Have more discipline.

Show more thoughtfulness."



How we respond is important.

"WE MUST...

~~Be more careful.~~

~~Avoid making mistakes.~~

~~Have more discipline.~~

~~Show more thoughtfulness.~~"

'BEST PRACTICES'
ARE NOT REAL PRACTICES



How we respond is important.



'BEST PRACTICES'
ARE NOT REAL PRACTICES

@this_hits_home

‘incidents’ are
SURPRISES



@this_hits_home

Takeaways

Recovery > Prevention.

There is no root cause.

Stop reporting on the nines.

Learn how things go right.

Thank you



@this_hits_home

http://continuous.wtf