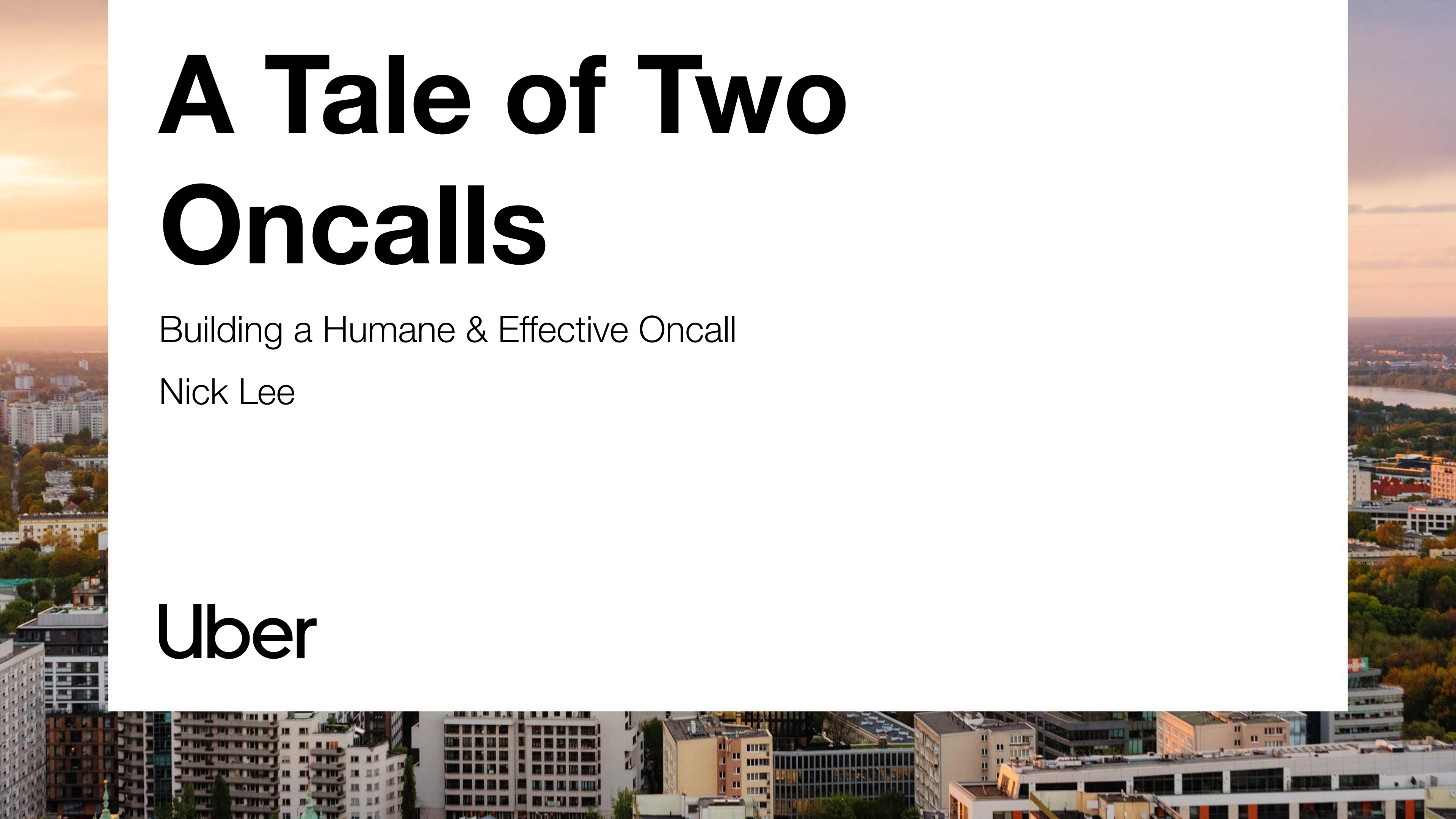


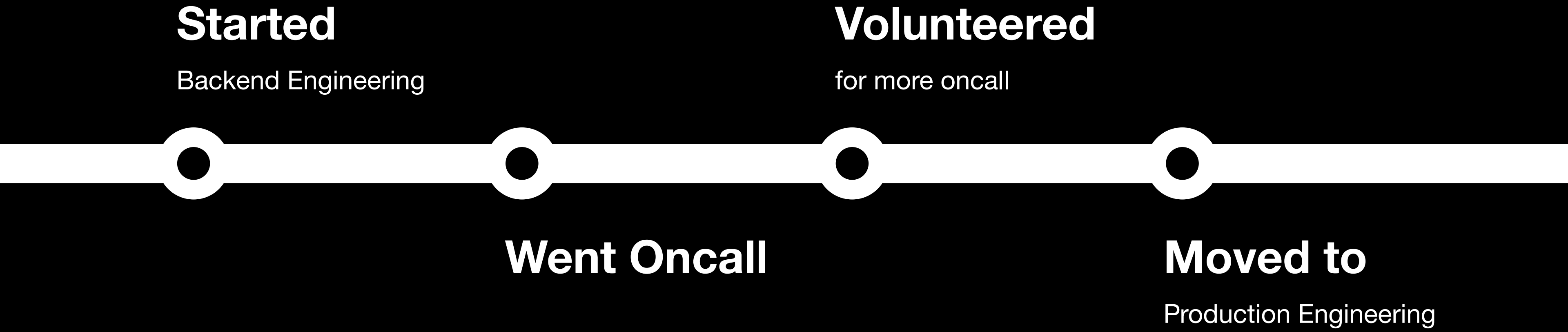
A Tale of Two Oncalls

Building a Humane & Effective Oncall

Nick Lee

Uber

Where I come from



Started

Backend Engineering

Went Oncall

Volunteered

for more oncall

Moved to

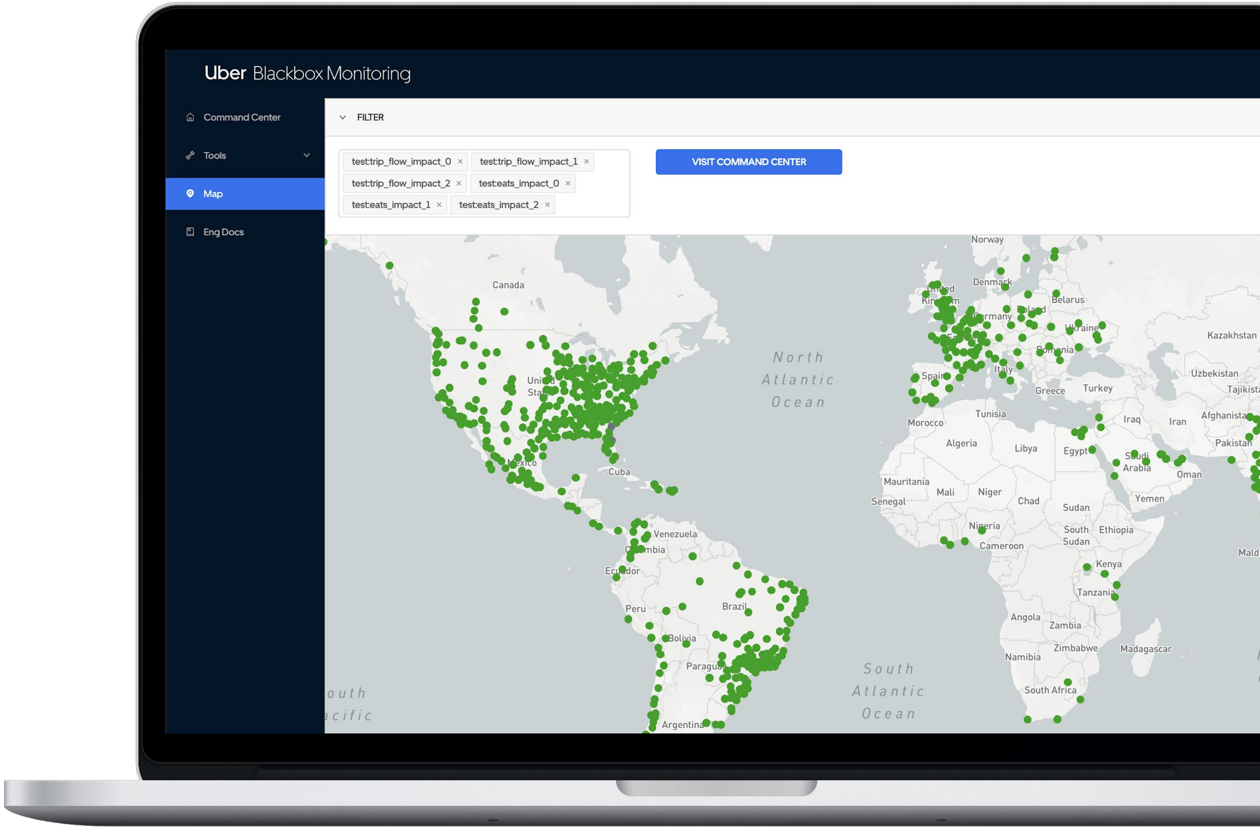
Production Engineering

Defining Oncall

Attribution: Wyoming National Guard https://creativecommons.org/licenses/by-nd/2.0/

ring0

Protects the core customer experience



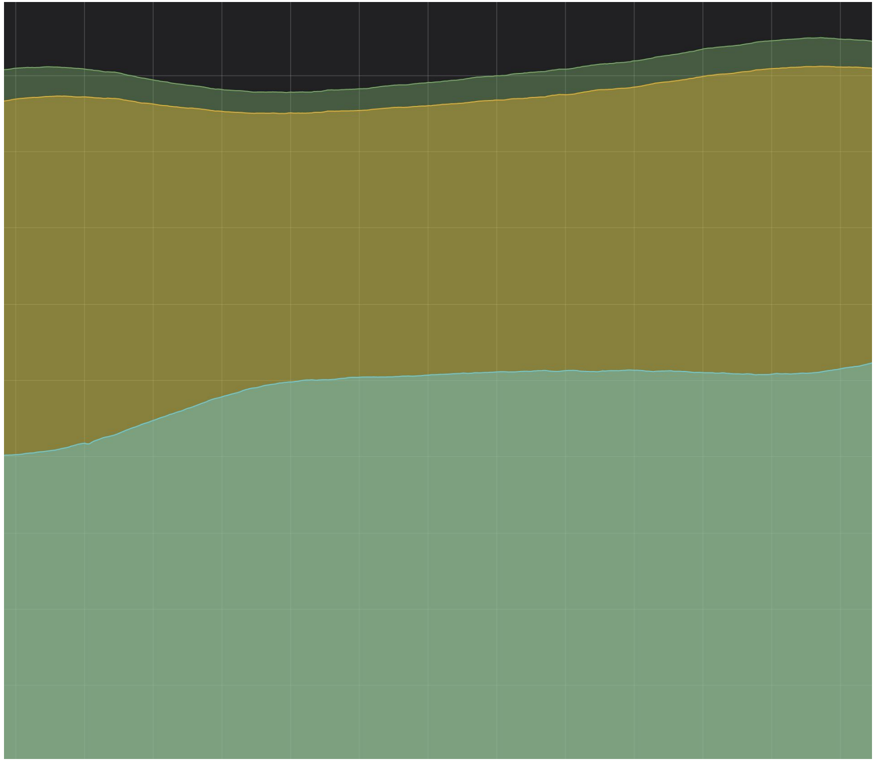
test:trip_flow_impact_0 ×
test:trip_flow_impact_1 ×
test:trip_flow_impact_2 ×
test:eats_impact_0 ×
test:eats_impact_1 ×
test:eats_impact_2 ×

Vendor Integrations

Builds & supports interactions with third parties

https://topatoco.com/products/kcg-tfdog-plush

Two Things Matter



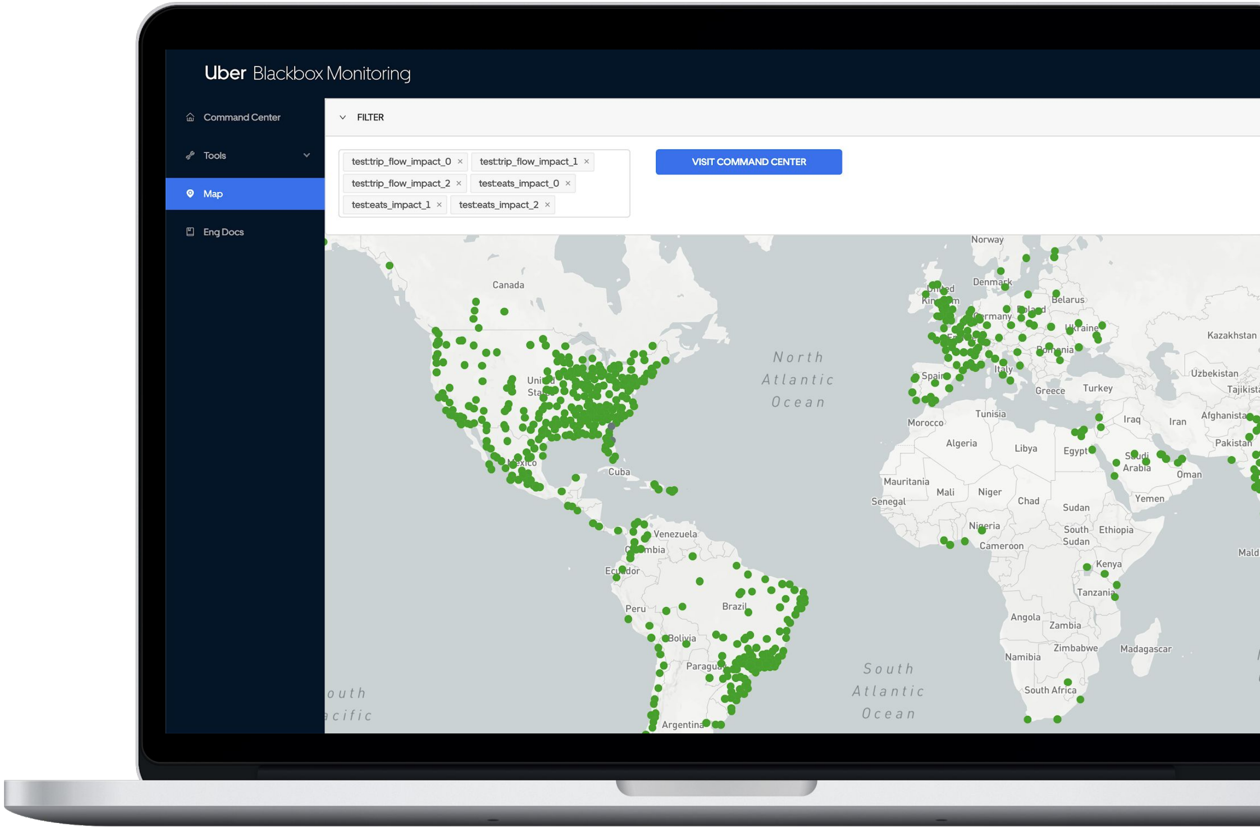
Two Things Matter

Attribution: Marco Verch https://creativecommons.org/licenses/by/2.0/

ring0

✓ - **Effective**

✓ - **Humane**



test:trip_flow_impact_0 ×
test:trip_flow_impact_1 ×
test:trip_flow_impact_2 ×
test:eats_impact_0 ×
test:eats_impact_1 ×
test:eats_impact_2 ×

Vendor Integrations

~~⚠️~~ - Effective

~~❌~~ - Humane

https://topatoco.com/products/kcg-tfdog-plush

Three Nines of Uptime

44 minutes

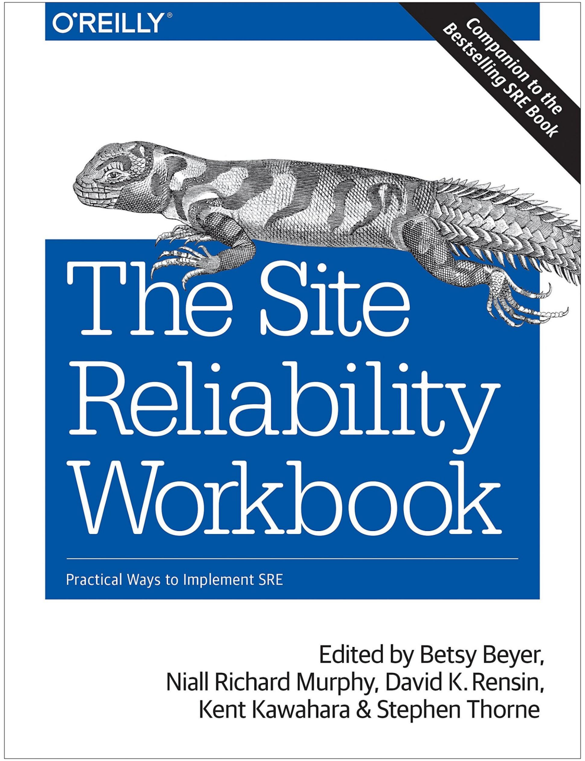
of downtime

Three Nines of Uptime

hours

of firefighting

Looking elsewhere for inspiration



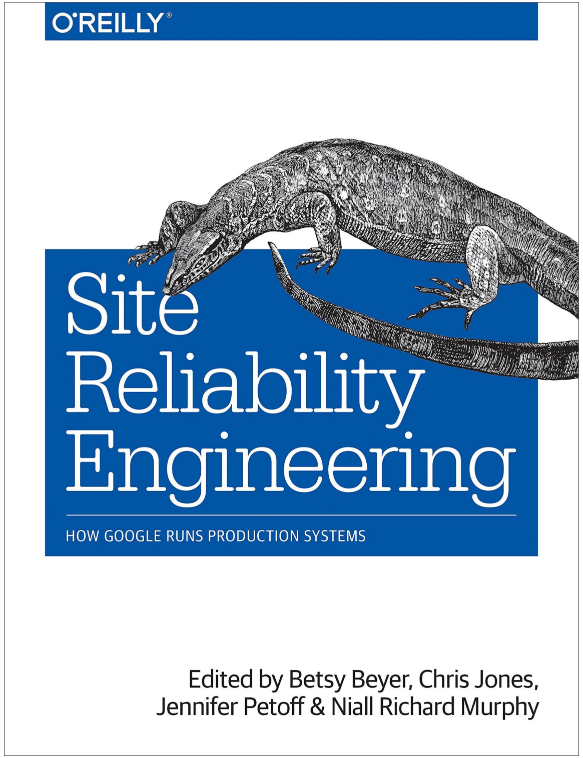
O'REILLY®

Companion to the Bestselling SRE Book

The Site Reliability Workbook

Practical Ways to Implement SRE

Edited by Betsy Beyer,
Niall Richard Murphy, David K. Rensin,
Kent Kawahara & Stephen Thorne



O'REILLY®

Site Reliability Engineering

HOW GOOGLE RUNS PRODUCTION SYSTEMS

Edited by Betsy Beyer, Chris Jones,
Jennifer Petoff & Niall Richard Murphy

Triage outages very aggressively

Constantly refine your alerting

The response to an alert should never be ambiguous

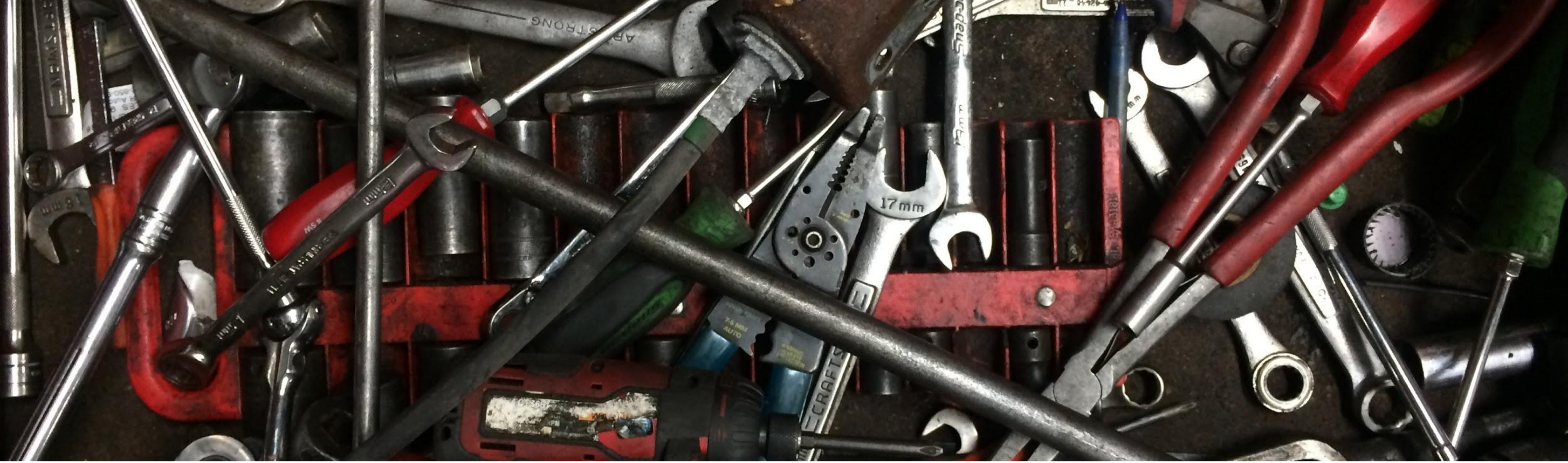
Those best practices didn’t help

Attribution:
https://giphy.com/gifs/this-is-fine-dumpster-fire-floating-1rNWZu4QQqCUaq434T

The search for the secret sauce begins

What did ring0 have that we didn’t?

Attribution:
https://www.youririshshop.com/YR-sauce-Sqeezy-485g



The Mitigation Toolbox

NOTICE
DO NOT DISTURB WHILE SURGERY IS IN PROGRESS
Reorder: ONEP-34593 www.ComplianceSigns.com

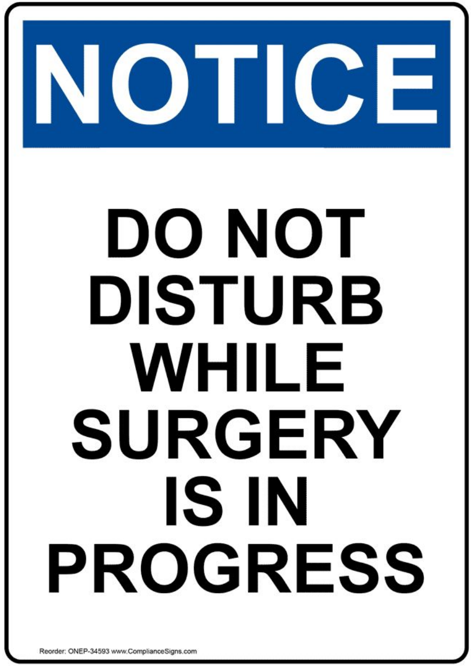
Stabilize

Diagnose

Treat

Attribution:
https://commons.wikimedia.org/wiki/File:Siemens_Magnetom_Aera_MRI_scanner.jpg
https://creativecommons.org/licenses/by-sa/4.0/

Diagnose

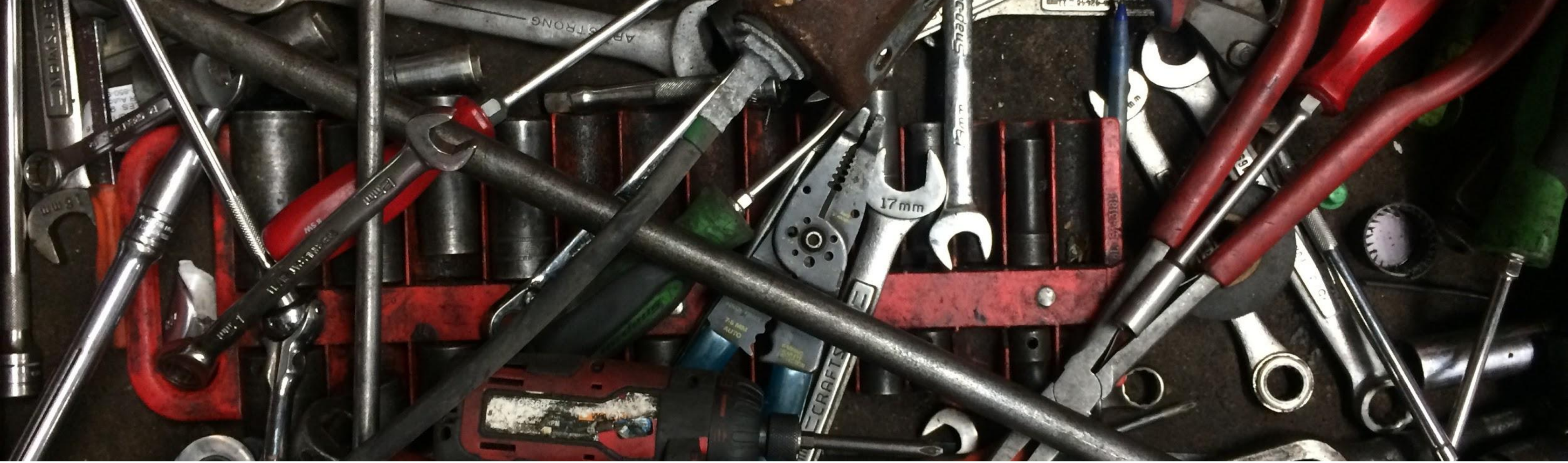


NOTICE

DO NOT
DISTURB
WHILE
SURGERY
IS IN
PROGRESS

Reorder: ONEP-34593 www.ComplianceSigns.com

Treat



But aren’t tools expensive?

Quantify your oncall

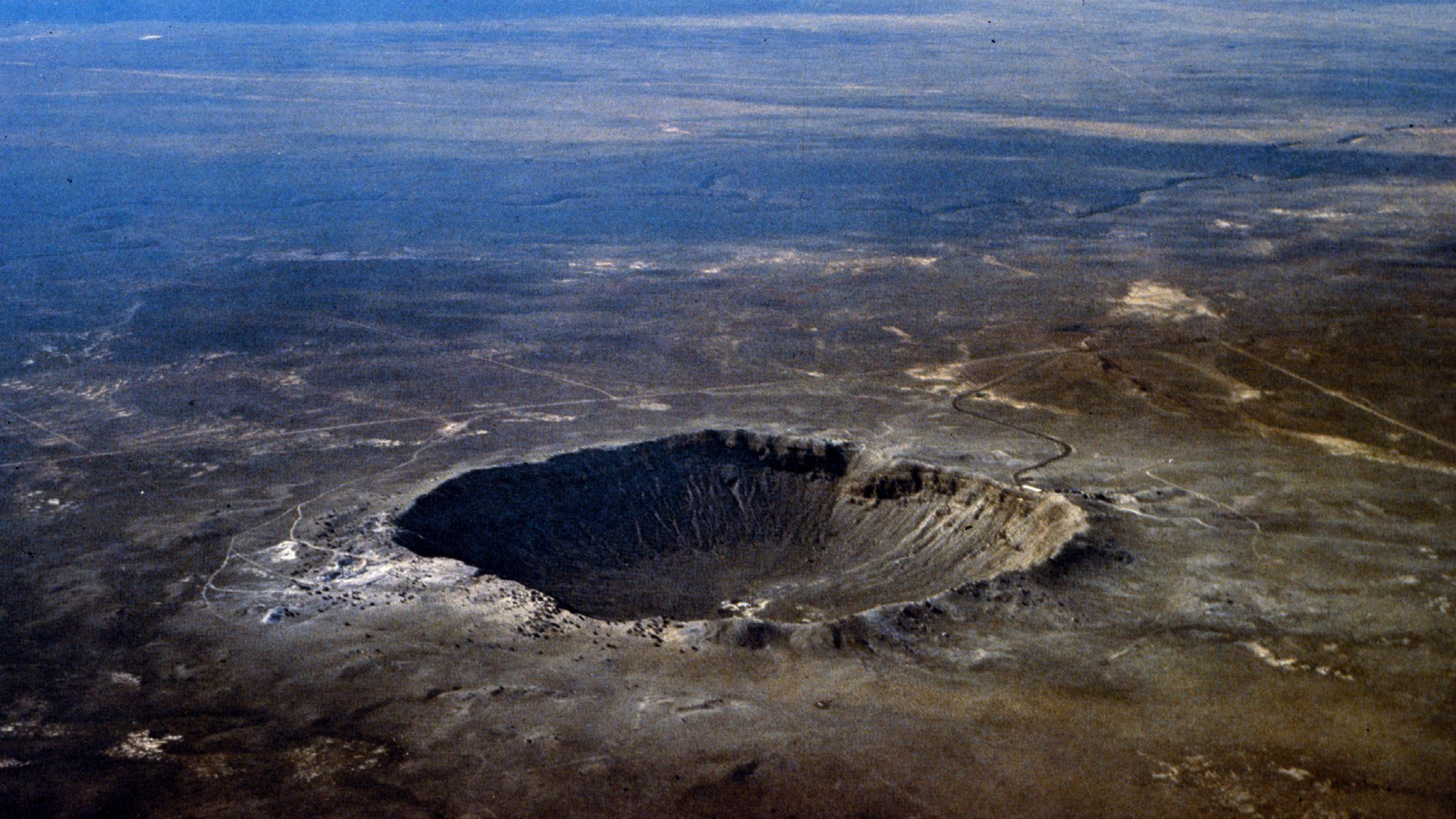
01 Cost per minute of outage
02 Broken Nights per week
03 Incident Counts

A messy mitigation is better than no mitigation

Attribution: Tom Woodward https://creativecommons.org/licenses/by-sa/2.0/

DB
Ziel Destination
Gleis Platform/Voie
Mannheim-Friedrich 11
Gernsheim 17 Train is cancelled
Köln Hbf 7 Train is cancelled
Berlin Hbf 9 Train is cancelled
Passau Hbf 6 Train is cancelled
Siegen 16
Saarbrücken Hbf 20
Fulda 8 Train is cancelled
Bruxelles-Midi 19 Aujourd'hui du qua
Hanau Hbf 5 uai 5 - Heute auf G
r DB-Zugverkehr beeinträchtigt. Bitte
d informieren Sie sich auch im Internet

Find ways of making your outage irrelevant to the user



What now?

Mitigation Runbook

Fifteen easy steps:

1. Wake up
2. Scream and run in circles
3. Stop all deployments
4. Tell everyone not to make any changes
5. Check that the other datacenters are healthy
6. Disable simulated traffic
7. Update the load balancers
8. Get a code review for that
9. Deploy the config change
10. Update your host files
11. Get a code review for that
12. Deploy the config change
13. Update your DNS entries
14. Realize this is too long

Make your tools an extension of human decision making

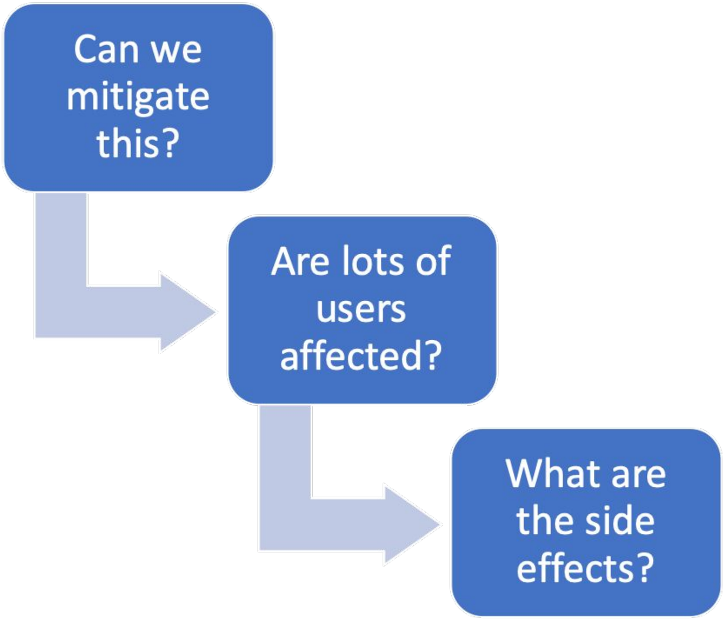
“

```
run lockdown/all
run failover fra1
```

Non-SREs who trust the tools:

Make your oncall feel as safe as possible

Decision Frameworks

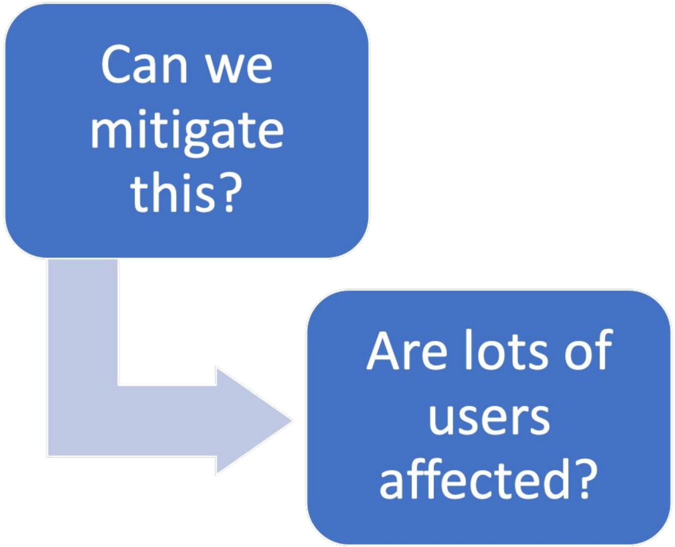


Can we mitigate this?

Are lots of users affected?

What are the side effects?

Opinionated Frameworks



Non-SREs who trust the tools:

Existing
Team
Members

Make the tool feel safe

Make every action deliberate

```
Enter 9000 to execute or n to abort: 1234
Invalid input. Enter 1713 to execute or n to abort: 0000
Enter 2777 to execute or n to abort: 2777

2019-09-03T12:56Z  [I] ( PIN_CHECK )

failover operator lock is unlocked
```

Log all mitigation actions

```
 % toolbox run

Usage: toolbox run COMMAND [arg...]
Run a mitigation plan

Commands:

  undrain/time        Time-based revert of failover actions.
  undrain/uuid        UUID based undrain for failover actions
```

Non-SREs who trust the tools:

Existing Team Members

New Team Members

Make the tool safe

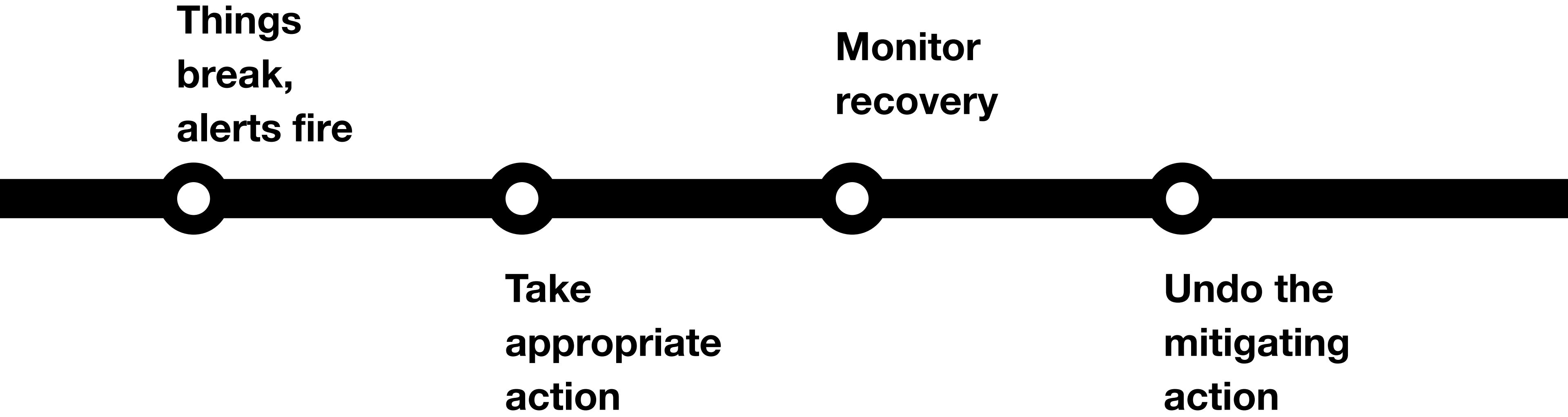
Non-SREs who trust the tools:

Existing Team Members

New Team Members

Product Owners

Our Oncall Today



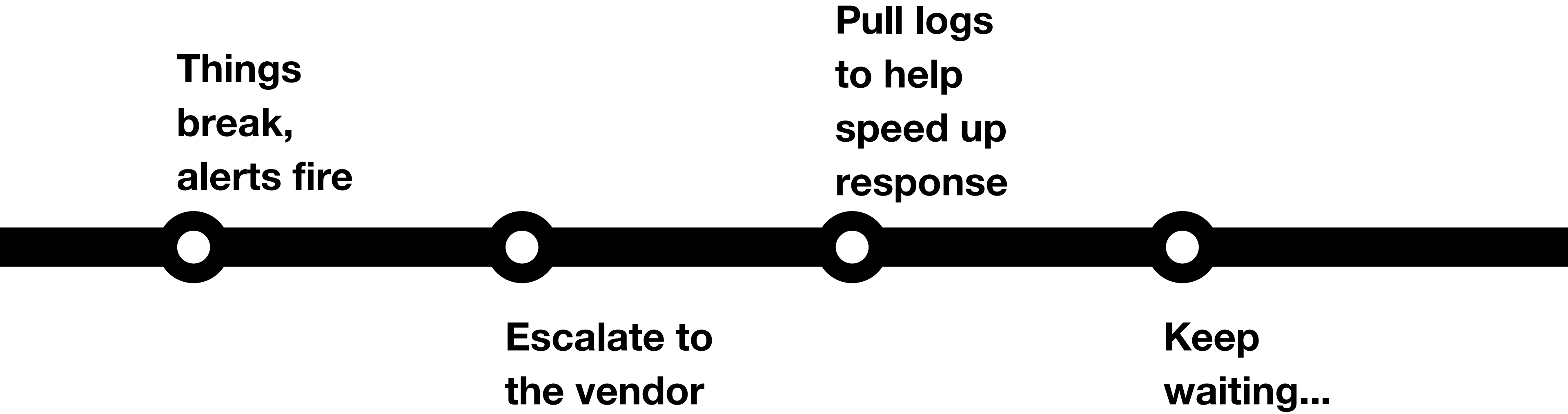
Things break, alerts fire

Take appropriate action

Monitor recovery

Undo the mitigating action

Our Old Oncall

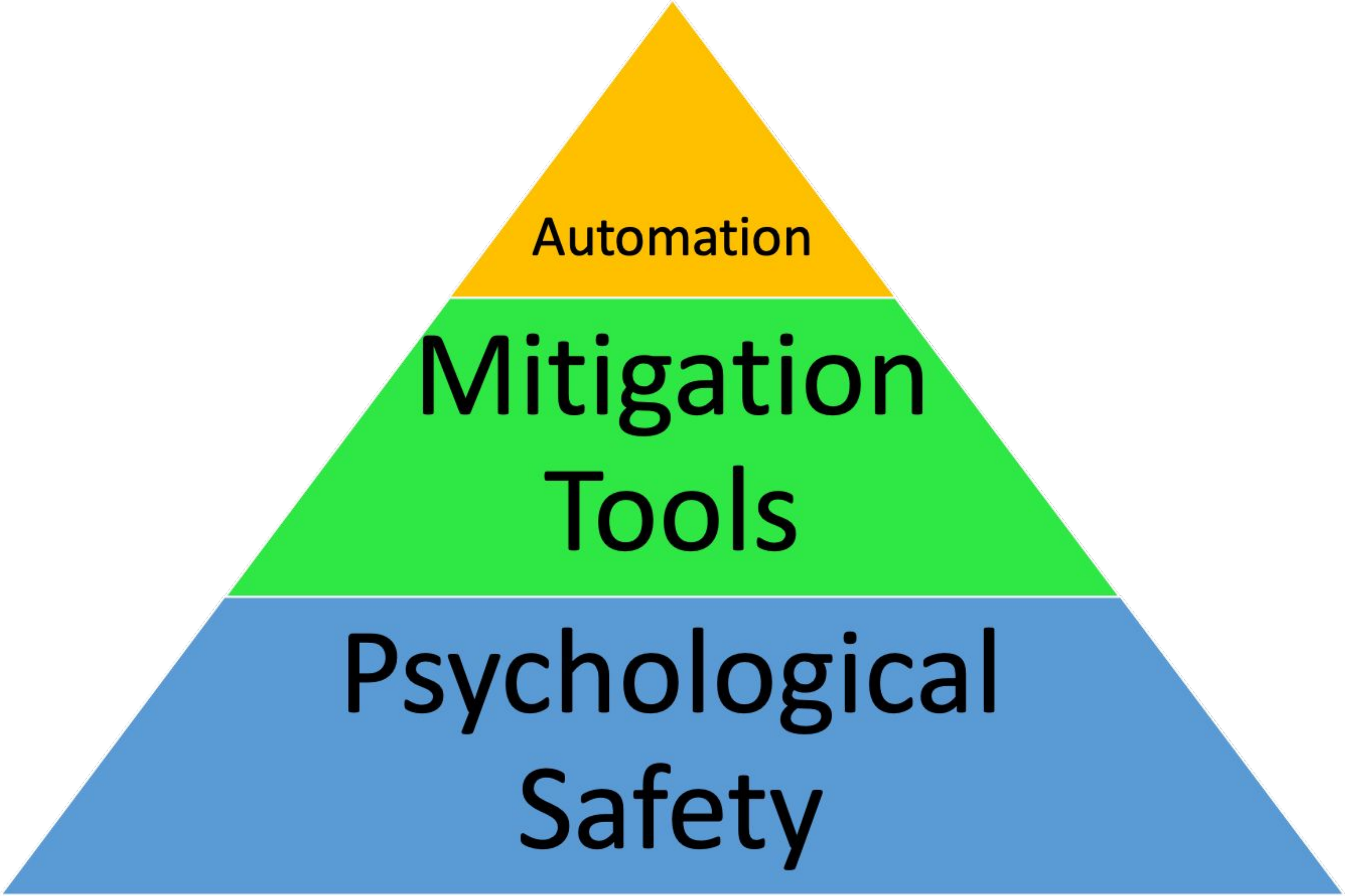


Things break, alerts fire

Escalate to the vendor

Pull logs to help speed up response

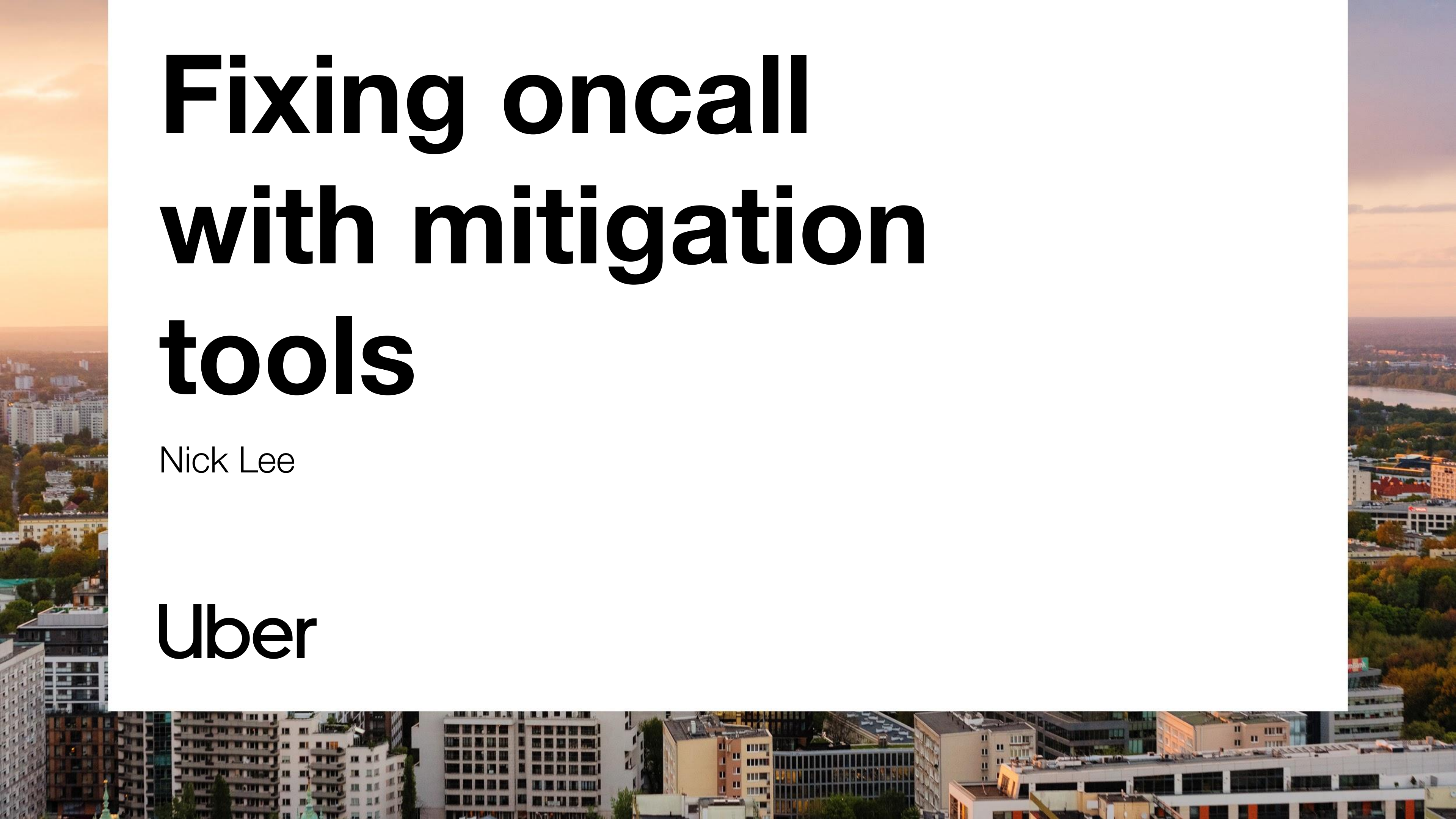
Keep waiting...



Automation

Mitigation Tools

Psychological Safety



Fixing oncall with mitigation tools

Nick Lee

Uber