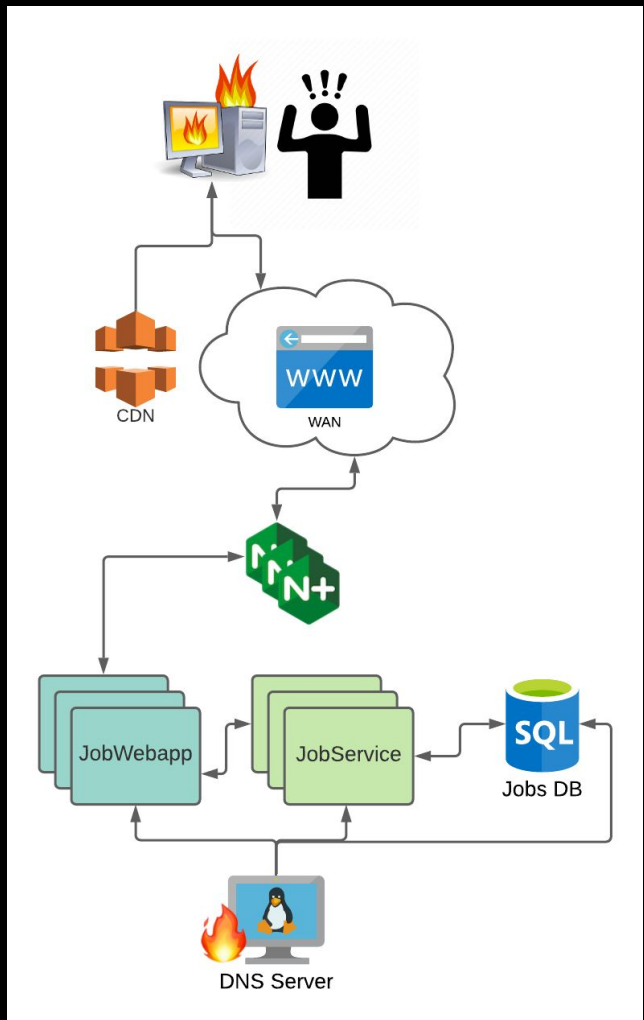


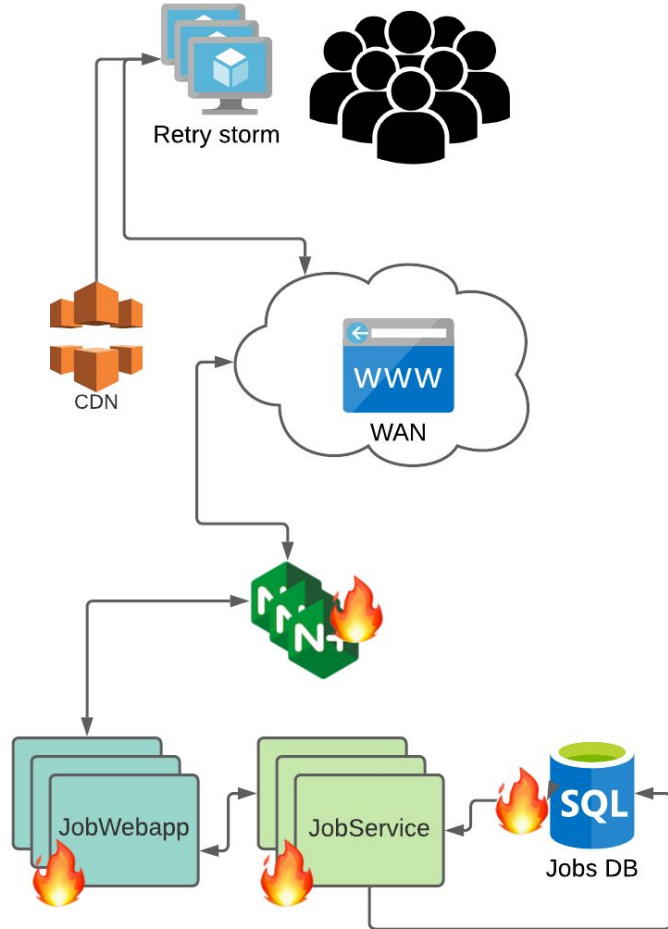
bit.ly/safer-operations





Complexity





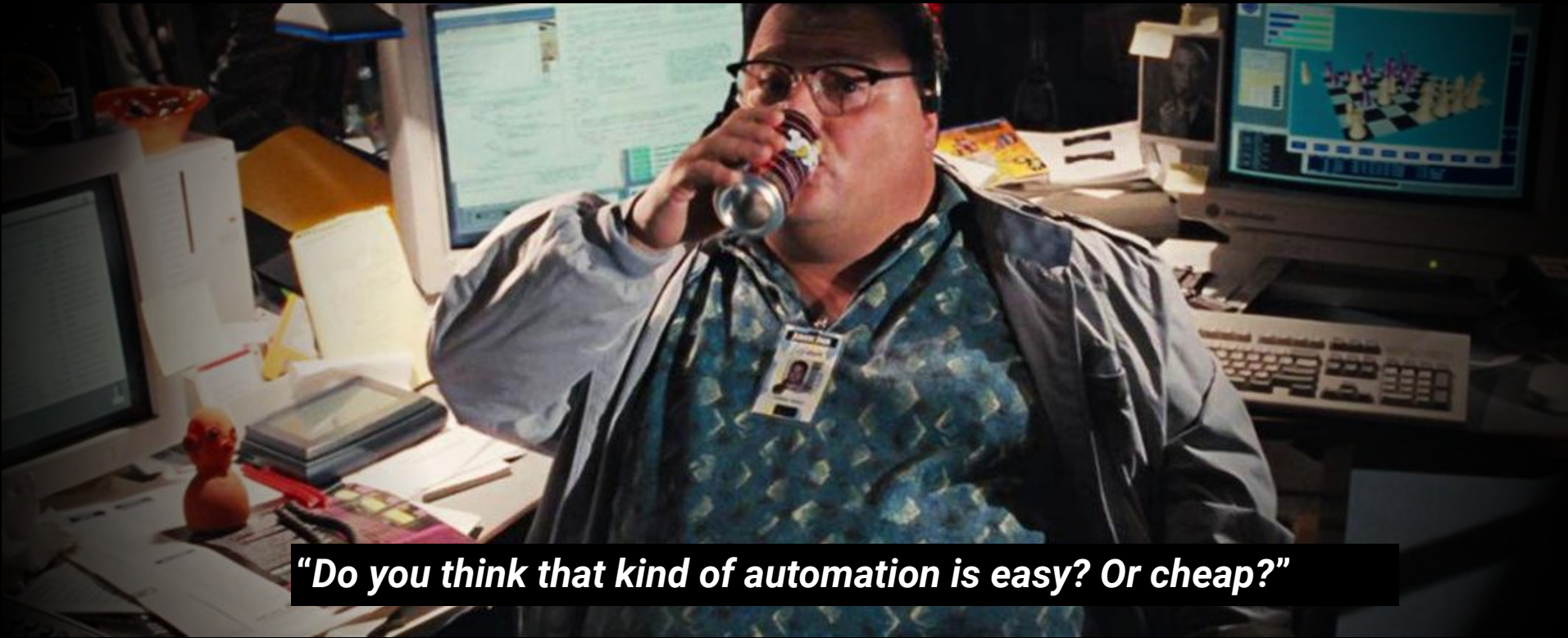
Compare 2020 with 2010

**Developers can do much
more with much less**



**launch
entire stack**

few lines of code



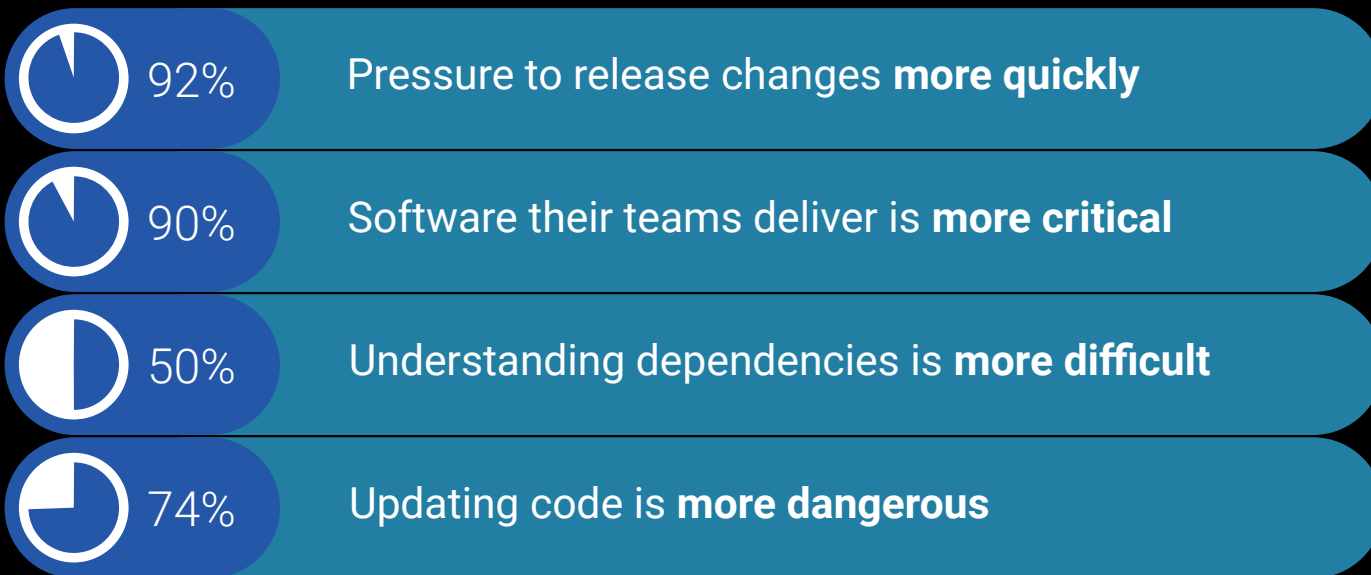
Still from *Jurassic Park* (1993), Universal Pictures

Findings from 2020 Emergence of Big Code Survey

**Over half of developers surveyed are working with
100x the volume of code than 10 years ago**

Findings from 2020 Emergence of Big Code Survey

Majority of developers report



Why isn't this leading to
more **company-ending**
outages?

What is driving this
increase in complexity?

It's success

Law of Stretched Systems

Capacities being stretched

- + **Organizational workload**
- + **Pace of development cycle**
- + **Demands on expertise**
- + **Speed of technological innovation**

Are We Getting Better Yet?

Progress toward safer operations

Alex Elman

SRE Leadership

@_pkill 





» **Indeed is currently unavailable**

We should be back online in a few minutes. Thanks for your patience and good luck with your job search.

Network Engineer Jobs, Employ... x +

aq.indeed.com/jobs?q=network+engineer&l=south+pole

indeed Find Jobs Sign In Employer / Post job

What: Job title, keywords, or company: network engineer

Where: city, region or iceberg: Antarctica

Find Jobs Advanced Job Search

Date Posted Job Type Location Company

network engineer jobs in Antarctica

Sort by: relevance - date Page 1 of 5 jobs

Network Engineer Sr. - Palmer, Antarctica (Austral Winter 20...
GHG Corporation
Antarctica

- Strong ability to troubleshoot complex multi-vendor network issues in LAN and WAN networks.
- The Network Engineer Senior provides IT services at the National...

30+ days ago · More...

Network Engineer, Sr. - McMurdo, Antarctica (Austral Winter...
GHG Corporation
McMurdo Station

- Monitor network performance and reliability to proactively identify potential problems.
- Experience with the configuration and management of network firewalls is...

30+ days ago · More...

Get new jobs for this search by email

My email:

Activate

By creating a job alert, you agree to our Terms. You can change your consent settings at any time by unsubscribing or as detailed in our terms.

**We help
people
get
jobs.**

Prioritize a
learn and adapt safety mode over a
prevent and fix safety mode

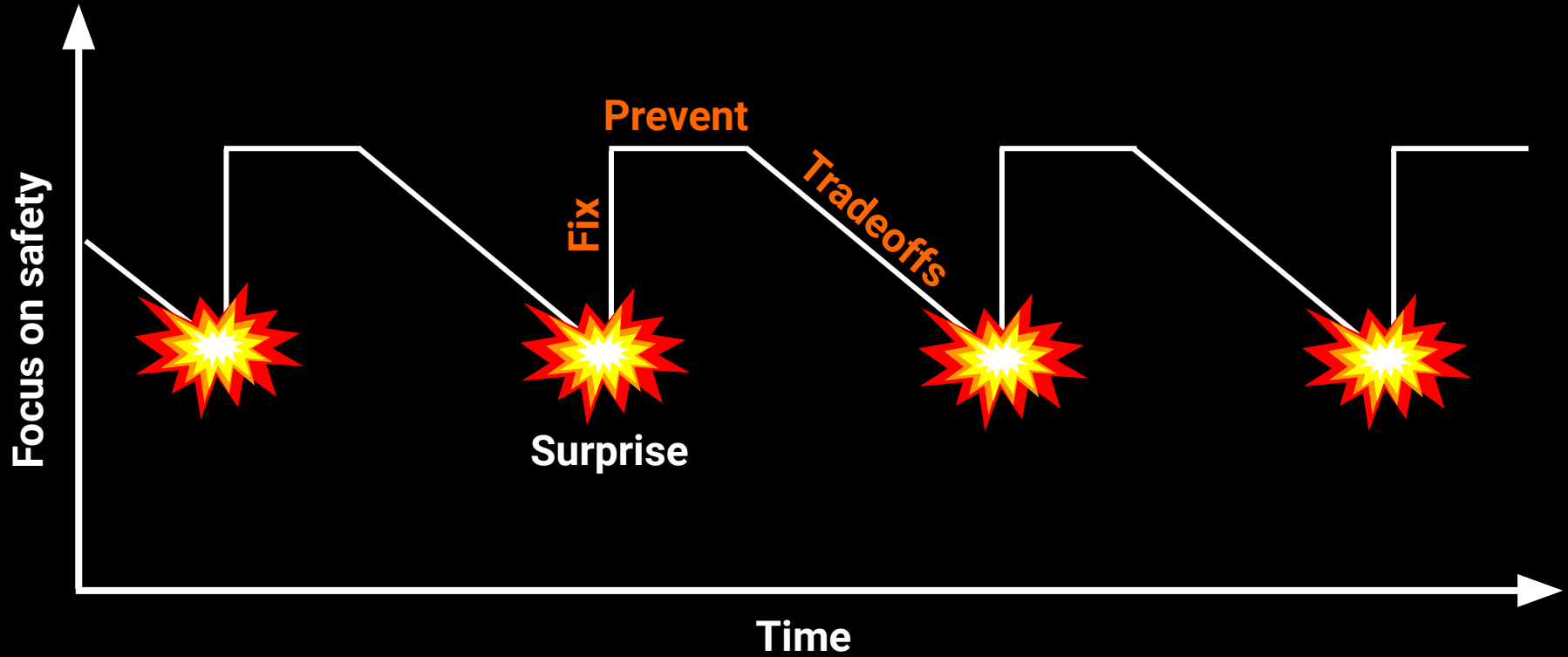
Prevent & Fix



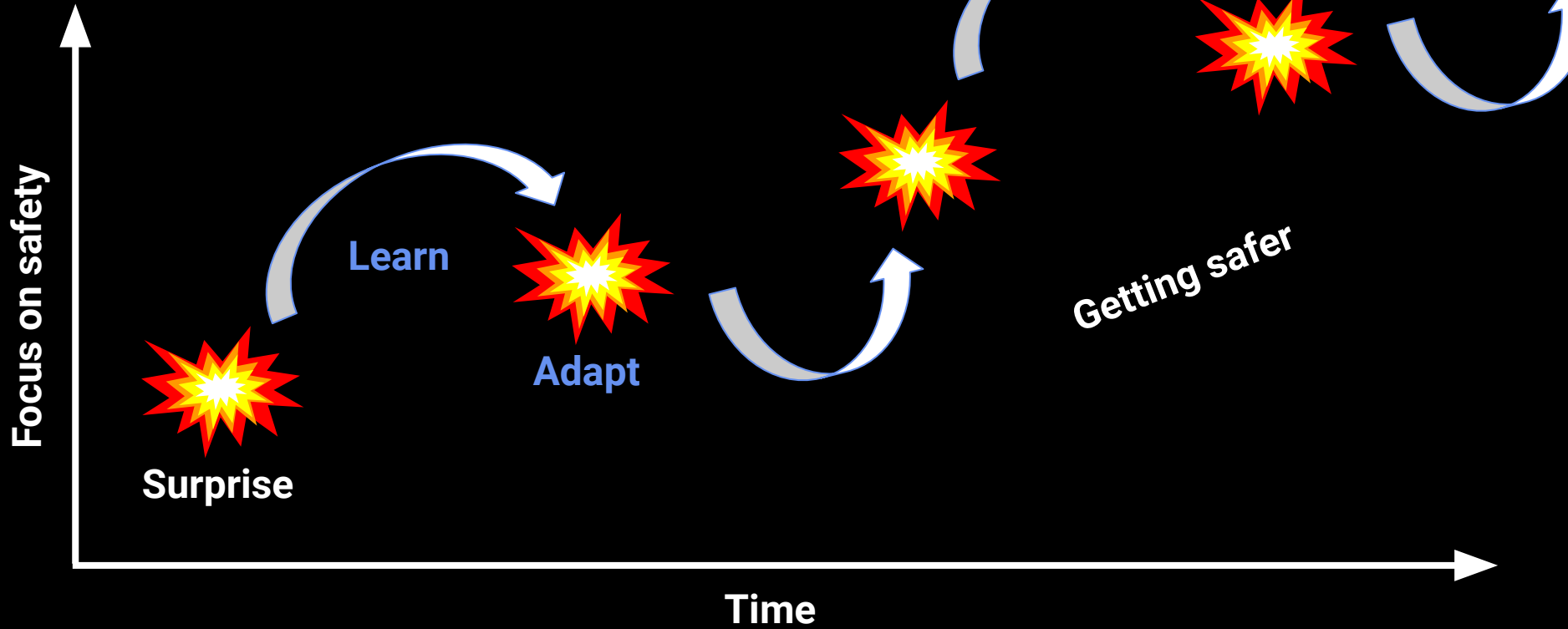
Learn & Adapt



The Prevent & Fix cycle



The Learn & Adapt reinforcing loop



**Action items form a defensive
strategy
but do not lead to learning**

A taxonomy of action-items

+ **Emergency work**

+ **Planned work**

- **eventually gets completed**
- **never gets completed**

+ **Undiscovered critical work**

Better preventions
Better fixes

itonlyamodel.com



Measuring progress



if you can't measure it, you can't manage it

“

*It is wrong to suppose that
if you can't measure it, you can't manage it
– a costly myth.”*

W. Edwards Deming



The most important figures that one needs for management are unknown or unknowable, but successful management must nevertheless take account of them.”

Lloyd S. Nelson

Can these be measured?

Richer metrics

- + Magnitude of psychological safety
- + Comfort in surfacing risk to leadership
- + Potential \$\$ losses from incidents that were avoided
- + Net cost/benefit of experts changing teams
- + Amount an organization is learning

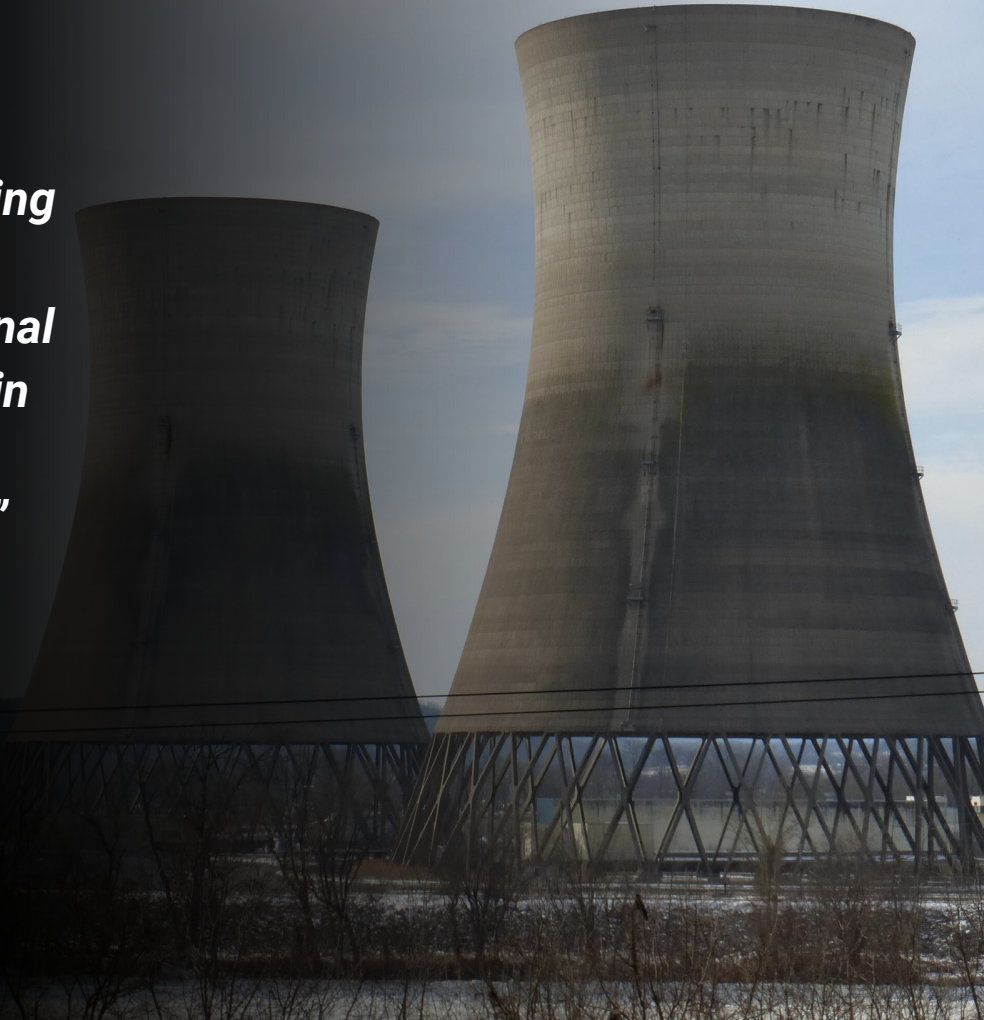
Be data-driven
avoid being driven by data

“

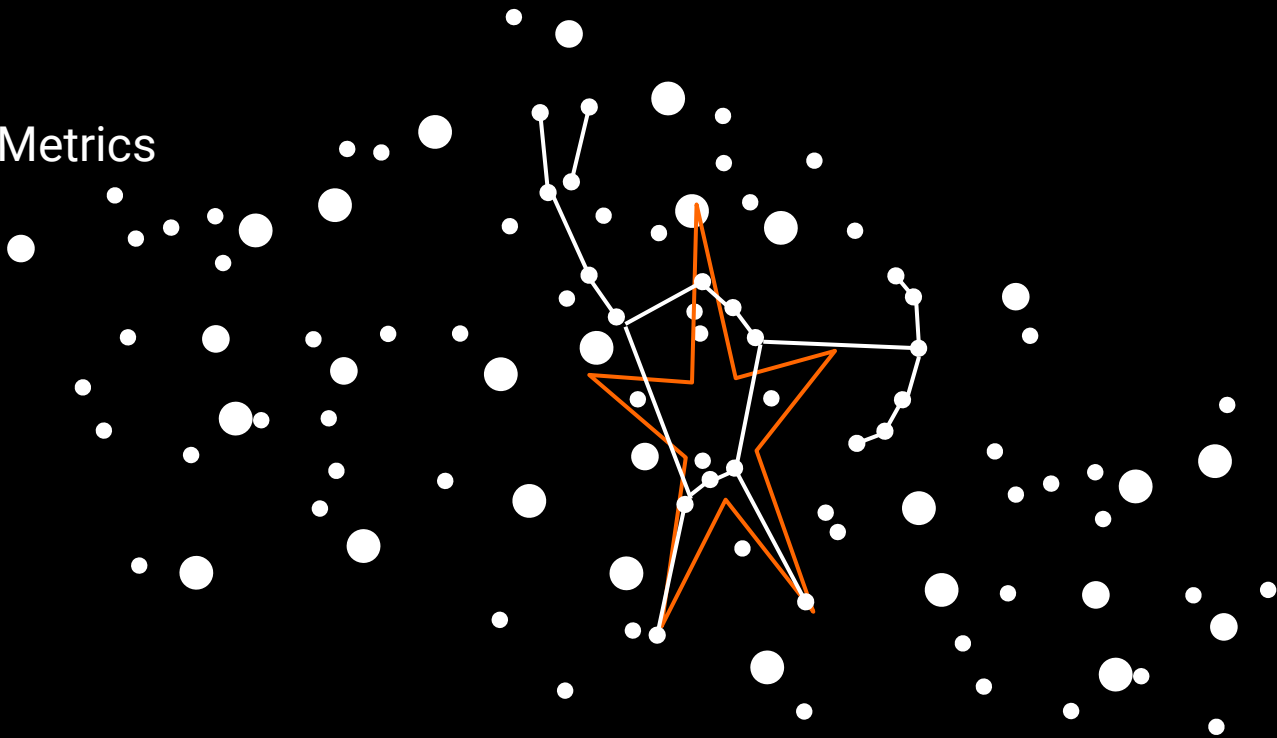
*It is a difficult thing to look at a winking light on a board, or hear a peeping alarm ... and ... draw any sort of rational picture of something happening out in the vast plant, let alone meet it with anything but a **mechanical response**”*

Three Mile Island commission report

<http://bit.ly/three-mile-island>



Valid Metrics



Everybody has a story to tell

**Metrics anchor the story and the
story gives meaning to the
metrics**

Activities around creating and maintaining safety

Preparation

Reporting reliability

Assessing accountability

Incident analysis

Incident write-ups

Activities around creating and maintaining safety

Preparation


Preparation



Emphasizes only **avoiding a recurrence**



Emphasizes a more accurate and complete **understanding**



**Barriers and guardrails are used
to **prevent** people from repeating
mistakes**

Examples of barriers and guardrails

- + **Turning MySQL Safe Mode on in Production**
- + **Disallowing SSH access**
- + **Capping instance capacity**
- + **Preventing rollbacks without approval**



All practitioner actions are gambles.”

Richard I. Cook

<https://how.complexsystems.fail/#10>

Performance variability

Ensure **positive outcomes** through activities like team practice and chaos experiments



Chaos experiments as scrimmage

Practice: safe, predictable

Chaos: safe, unpredictable

Incidents: unsafe, unpredictable

A photograph of two men in a dimly lit server room. They are sitting at a desk with multiple computer monitors. The man in the foreground is wearing glasses and a dark shirt, looking intently at a monitor. The man behind him is also looking at the same monitor. The room is filled with server racks and other computer equipment, creating a professional and focused atmosphere. The lighting is primarily from the monitors and desk lamps, casting a warm glow on the scene.

Why are incidents so difficult?

Escape Rooms

Still from *Schitt's Creek* (2020), Pop TV/Netflix



Hindsight





how to fix prod



Mistakes are a feature not a bug

Stories

- + The tale of the well-choreographed incident response**
- + A brand new on-call responders painful experience through the obstacle course to address a stuck deploy**
- + The one where an accidental line of YAML burned \$250,000 in cloud costs**

Humans in the loop



Automation an opportunity to enhance or enable humans

Activities around creating and maintaining safety

Preparation

Reporting reliability

Reporting reliability



Reliability outcomes and human performance are **predicted** and **controlled**



Reliability outcomes and human performance are **monitored** and **influenced**

Incidents are a source of insights
but not a good measure of reliability

Reliability

Historically **good performance**

Robustness

Retains good performance **within a threshold** when challenged

Brittleness

Predictably **poor performance** when challenged

How fast can we **safely** go in a
brittle system?

Service Level Objectives

Service Level Objectives

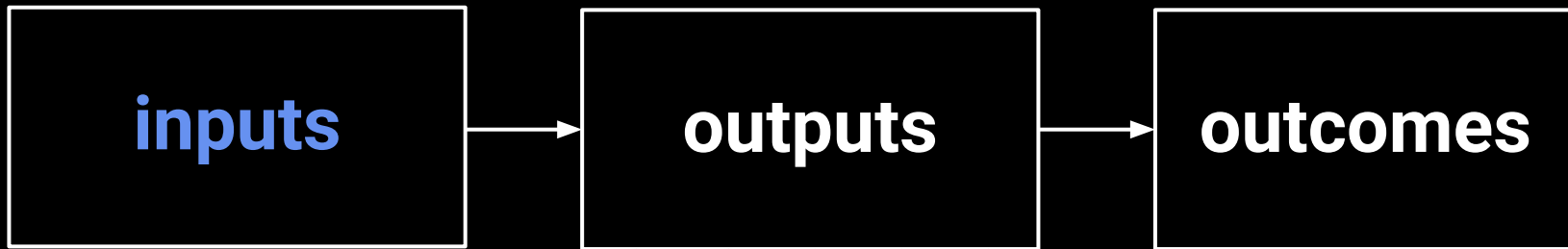
- + How **reliable** have we been?
- + How **fast** can we go?
- + How fast **should** we go?
- + What is the user **experiencing**?
- + Can we keep relying on FooService?

Control vs Influence

Things we control

Things we influence

What we want



inputs

outputs

outcomes

Circumstances out of
our control

circumstances

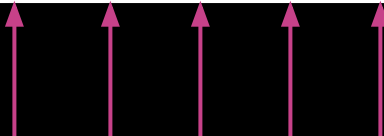
runbooks
training
procedures
fault tolerance



queue delays
timeouts
hypotheses
mitigations



Production
stabilized



packet loss
traffic spikes
hardware failures
users

circumstances

Outcomes over Outputs

Watch the inputs

Influence the outputs

Target the outcomes

Activities around creating and maintaining safety

Preparation

Reporting reliability

Assessing accountability

Assessing accountability



People who make mistakes are **blamed**. They are **obligated** to take responsibility.



People who make mistakes **feel supported** which inspires them to **seek opportunities**.

**Attribution is important to learning
but can also lead to blame**

Opportunity vs Obligation

Opportunities are taken, not given

Opportunities

+ Identifying

- **Defined goals and rationale**

+ Selecting

- **Career growth**
- **Assumptions and risks**

+ Realizing

- **Definition of “done”**

Activities around creating and maintaining safety

Preparation

Reporting reliability

Assessing accountability

Incident analysis

Incident analysis



Incidents result only in
technical fixes



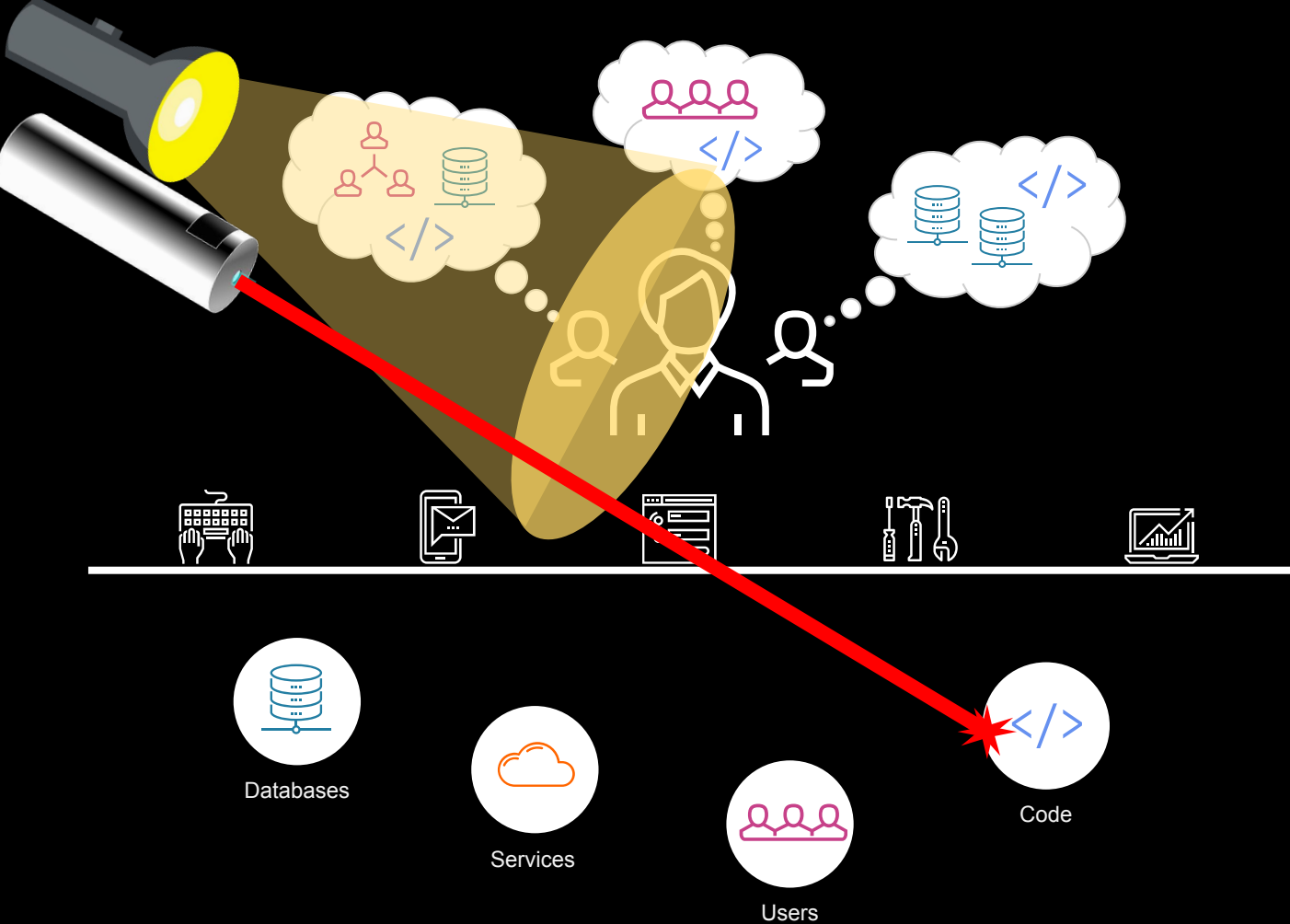
Incidents are investments in more
capable organizations



Focusing on reducing errors diverts energy and attention into ... narrowly targeted 'fixes' that treat symptoms but not the underlying problem"

Robert L. Wears

↑ Performance = ↓ errors + ↑ insight generation



Ignoring "above the line" misses at least 50% of the opportunities

Line of representation

Local-only fixes "below the line"



Improving a process

**What are we looking for during
incident analysis?**



Mickey

Oct 21st, 9:24 am

Peiwen what's involved in launching the new instances? I can press the button but don't know how long they'll take to provision puppet. Also wondering if we want to just stop dbs1005 and change its size to i3 (or m4.8xl)

Information - Seeking



Peiwen

Oct 21st, 9:26 am

I think the process is similar to what we do when launching ec2 in cmhqa. But if it's faster, we can change an existing instance. Maybe 1015, since we already have a change going on 1005

Information - Providing

Note

Instance changes are faster than launching instances. Is this widely known?





Peiwen Oct 21st, 9:20 am

while we are waiting, can we get 3 i3en.6xlarge instances going?

Information - Seeking

Mitigator - Potential



Mickey Oct 21st, 9:22 am

Andrew is that something you want to do or want me to? I'm not sure if we need a proc ticket or can just go for it

Mitigator - Potential

Needs Further Investigation

Note

Unclear procedures on how to handle change management during an incident.



Andrew Oct 21st, 9:22 am

I'm too tired to do that, tbh

Action - Not Taken

Note

Notable that Andrew mentioned he stayed up all night earlier in #prod-on-call



Andrew

Oct 21st, 9:34 am

this will impact the website's ability to see newly created advertisers when they try to post a job

Communication - Benefit

Is this in a runbook?

How does the lag impact this ability?

What else can impact the website's ability?

How are these responders performing?

**Judging human performance with
metrics applies conclusions
without context**

	Absolute time (CST)	Delta time	
Event start	09/04 10:30:00	-	A/B test turned up to 1%
Time to Detect	10/22 03:00:00	1M 16d 16h 30m	Product Manager identifies a discrepancy
Time to Diagnose	10/26 04:30:00	4d 1h 30m	Software dev makes a diagnosis (bug)
Time to Correct	10/26 04:31:00	1m	A/B test turned off
Time to Recover	10/26 04:37:00	6m	Prod declared stabilized

**Recording performance metrics
promotes **one perspective** over
others**

Timelines

Investigator Notes

Diagnosis



Hypothesis - Disproved X

Hypothesis - Generated X

Information - Diagnostic X

Information - Seeking X

Trigger - Actual X

Trigger - Potential X

Auto apply available tags

Include timestamped notes



09:20



09:25



09:35



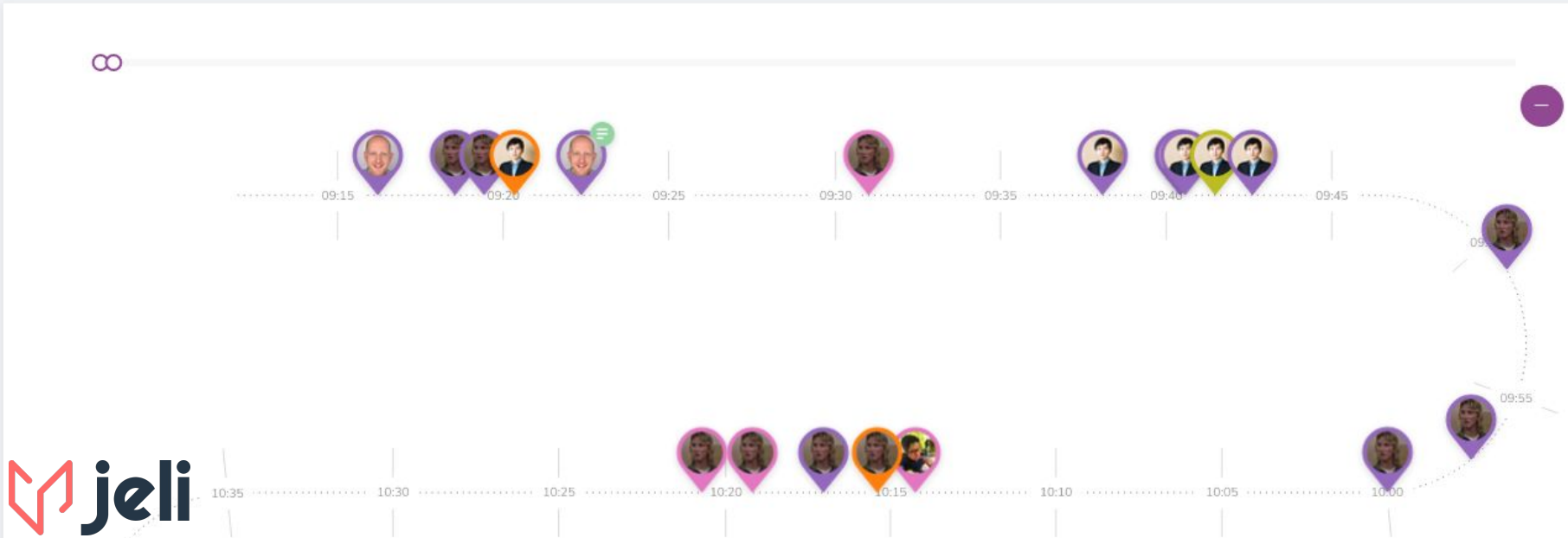
10:05

10:10

Mitigation

Mitigator - Potential X Mitigator - Actual X System Behaviour - Repair X

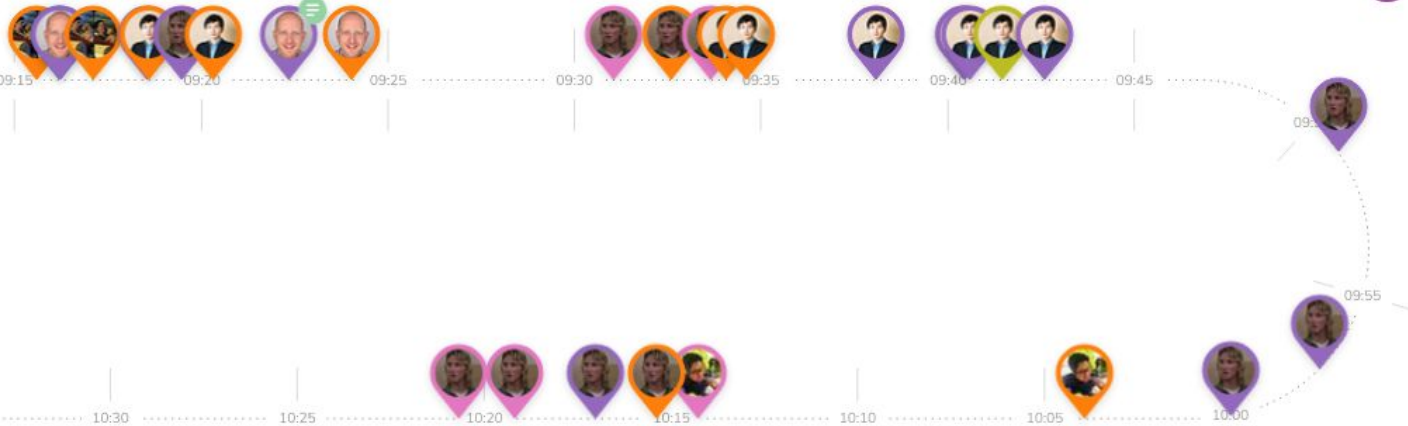
Auto apply available tags Include timestamped notes



Diagnosis and Mitigation

- Hypothesis - Disproved X
- Hypothesis - Generated X
- Impact X
- Information - Diagnostic X
- Information - Seeking X
- Mitigator - Actual X
- Mitigator - Potential X
- System Behaviour - Repair X

Auto apply available tags Include timestamped notes



Incident analysis

Richer metrics

+ Number of

- **technical fixes inspired by incident analysis**
- **insights generated per incident analyzed**

+ Time spent

- **verbally coordinating actions**
- **launching ec2 hosts**
- **restarting many instances of a database**

Activities around creating and maintaining safety

Preparation

Reporting reliability

Assessing accountability

Incident analysis

Incident write-ups

Incident write-ups




Incident write-ups favor a **particular viewpoint** above others



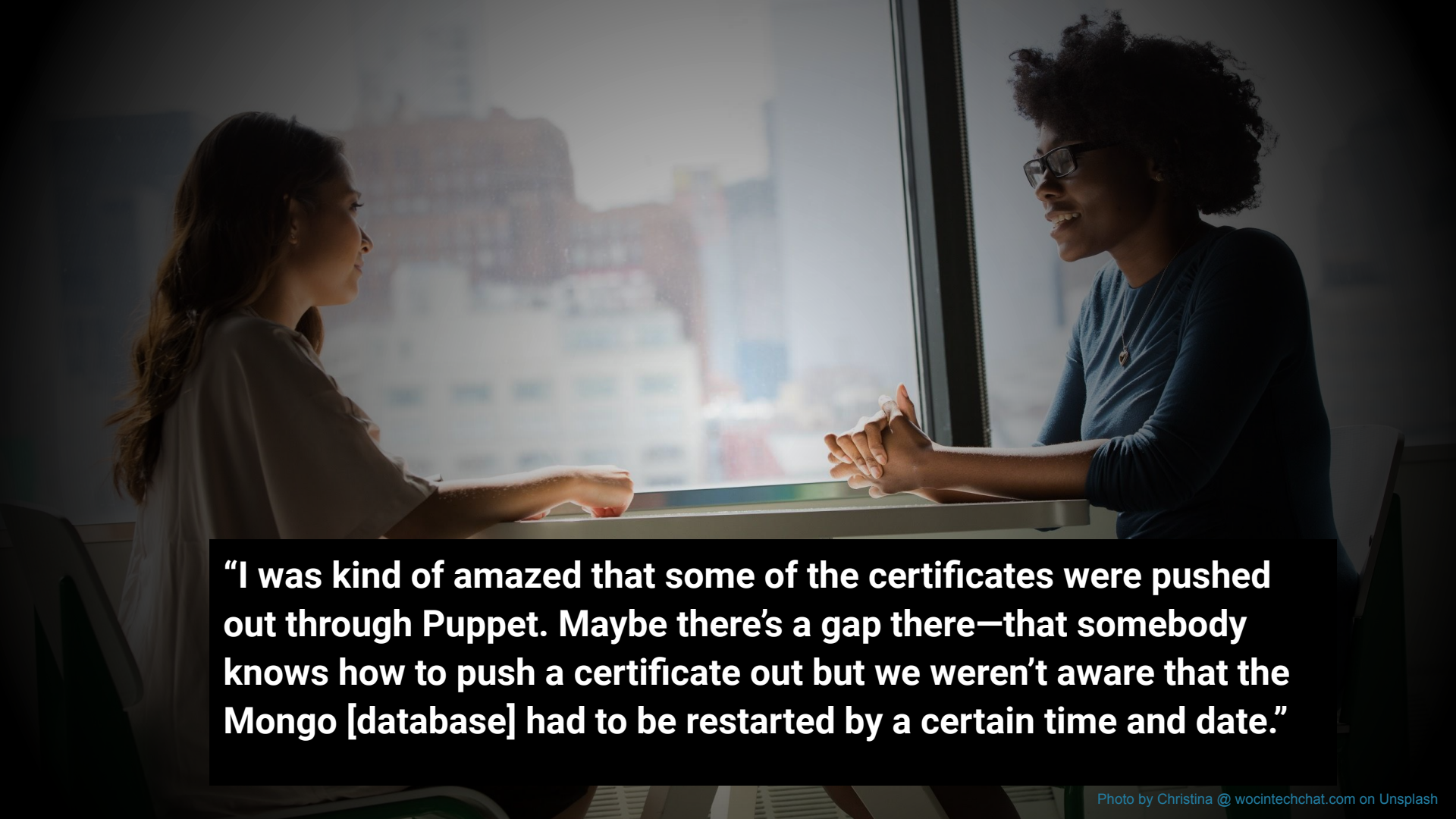
Incident write-ups faithfully present **multiple** perspectives

Interview Debriefing



“How did you become aware of the certificate expiration?”

“Oh, Edvaldo had some magic command that he ran...”

A photograph of two women sitting at a desk in an office, looking out a large window at a city skyline. The woman on the left has long dark hair and is wearing a light-colored top. The woman on the right has short curly hair, wears glasses and a blue long-sleeved shirt, and has her hands clasped on the desk. The scene is dimly lit, with light coming from the window.


“I was kind of amazed that some of the certificates were pushed out through Puppet. Maybe there’s a gap there—that somebody knows how to push a certificate out but we weren’t aware that the Mongo [database] had to be restarted by a certain time and date.”



“Right from the start when we had to restart everything I knew this was going to be a ton of work ... helping to take care of the number of servers that we have but also to help verify everything.”



“I asked Shira and Dmitri to help split the workload.”



“We both made the mistake of assuming that users from India are served by the Hong Kong DC because that seemed to be the closest one, but it turns out it actually goes to London.”



Audio Transcript

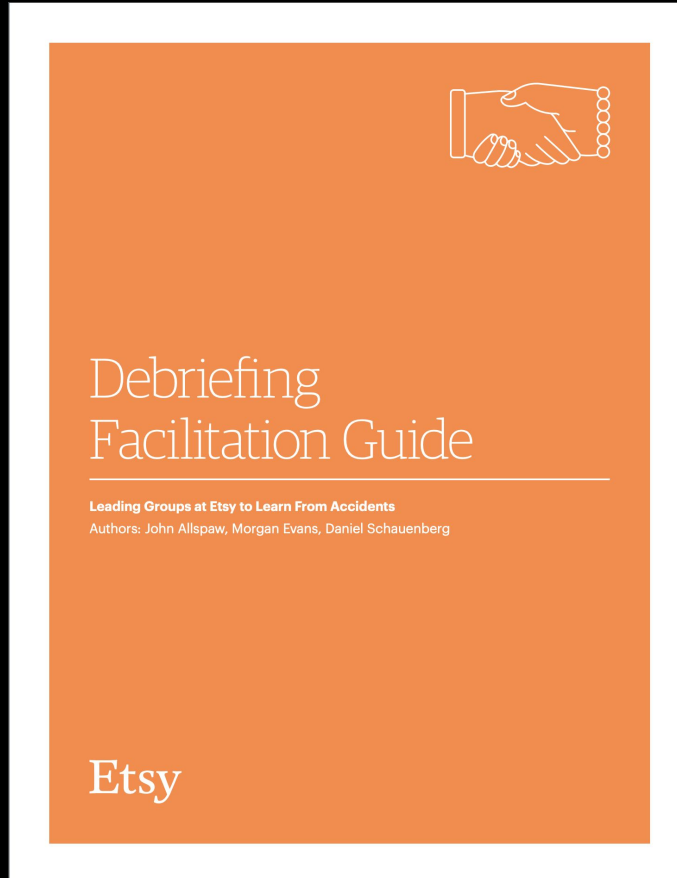
- and the number of errors if you don't realize what you're looking at.
- Alexander Elman**
- 06:39 So just to make sure that my understanding is the same as yours.
- 06:44 It was it's either a rate of it's the rate of change. It's how the rate of change is changing over time versus just the instantaneous count at that time.
- Eric**
- 06:55 That's how I understand that the rates is really a derivative, right. So,
- Alexander Elman**
- 07:01 **So you thought it was instantaneous counts, but it was the rate of changing rate.**
- Eric**
- 07:06 Right, yeah. So when you look at it, thinking it's counts and it's really the rate, like, oh, there's nothing major. Here it's, you know, there's one very small spike at the beginning.
- 07:16 And then it's just, you know, sort of just slightly increased and bouncing around and you know it didn't look didn't look drastic
- Alexander Elman**
- 07:23 Got it. That makes sense. But yeah, I think that's a very common, very common thing that people interpret
- Eric**
- 07:31 So, um, other than that. Um, let's see... e.
- 07:44 I think that kind of it, but that's the th... perspective.



zoom

@_pkill | bit.ly/safer-operations

<https://extfiles.etsy.com/DebriefingFacilitationGuide.pdf>



<https://youtu.be/TqaFT-0cY7U>

Three Traps

In accident investigation

Johan Bergström

Reader, Lund University



Traps to avoid

+ Counterfactual reasoning

- “She should have waited before restarting...”

+ Normative language

- “He lacked an understanding of...”

+ Mechanistic reasoning

- “Maintenance would be less risky if we automated this.”

Incident write-ups

Richer metrics

+ Number of

- **distinct write-up document opens**
- **attendees to review meetings**
- **distinct perspectives represented**
- **employees trained using write-up**

+ Qualitative survey feedback

- **How was the write-up useful?**



Ask deeper questions

**Our system has been very reliable
over the past few quarters.**

Why?

How close to the **safety boundary is the pod autoscaler pushing my infrastructure?**

**Are my cloud provider's staff a
team player in my
sociotechnical system?**

Recap

- + Deeper understanding leads to better fixes and enduring prevention**
- + Reliability is reported using SLOs not incidents metrics**
- + Nobody has control over how an incident unfolds**
- + Incidents are an opportunity to improve the accuracy of mental models**
- + At least half of incident analysis should focus on human factors**
- + Comparative storytelling enhances learning**

bit.ly/safer-operations

John Allspaw
Johan Bergström
Jabe Bloom
Bryan Cantrill
Martin Check
Iulian Circo
Richard Cook
Todd Conklin
Sidney Dekker
W. Edwards Deming
Kenneth M. Ford
Martin Fowler


Will Gallego
Ketan Gangatirkar
Vanessa Huerta Granda
Patrick J. Hayes
Rein Henrichs
Alex Hidalgo
Lorin Hochstein
Robert R. Hoffman
Erik Hollnagel
Nora Jones
Ryan Kitchens
Gary Klein
Jason Koppe

Sue Lueder
Laura Maguire
Lloyd S. Nelson
Jens Rasmussen
J. Paul Reed
Casey Rosenthal
Matthew Schemmel
Joshua Seiden
Steven Shorrock
Andrew S. Townsend
Robert L. Wears
Ron Westrum
David Woods

<https://learningfromincidents.io>

Thank you

Alex Elman

 @_pkill

 indeed

bit.ly/safer-operations

