Tales From The VOID

The Scary Truth About Incident Metrics

@courtneynash

SRECon Americas 2022

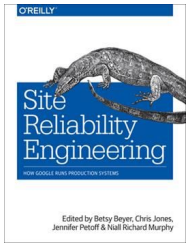
Courtney Nash

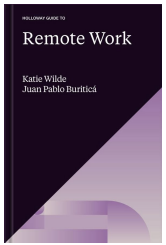
Senior Research Analyst

Verica

@courtneynash

O'REILLY®
Velocity





amazon

fastly

Microsoft

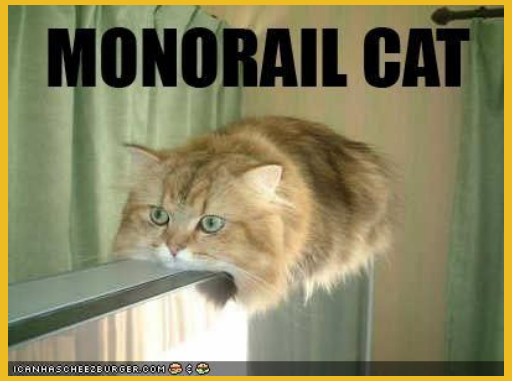
HOLLOWAY

VOID

About the VOID

The Verica Open Incident Database (VOID) makes public software-related incident reports available to everyone, raising awareness and increasing understanding of software-based failures in order to make the internet a more resilient and safe place.

Software Runs The World



VOID

We All Have Incidents

As an industry, we're solving a lot of **similar problems** in silos and **not sharing** what happened, or (even better) what we learned.

VOID

HISTORICAL PRECEDENT

Learning From Incidents (LFI)

“Incident analysis is not actually about the incident, it’s an opportunity we have to see the delta between how we think our organization works and how it actually works.”
—Nora Jones, CEO Jeli.io & Founder, LFI

https://www.learningfromincidents.io/

VOID

WHAT'S IN THE VOID

1,856 public incident reports from 610 organizations

From 2008 up to March, 2022

In a variety of formats:

- social media posts
- status pages
- blog posts
- conference talks
- news articles
- comprehensive retrospectives/postmortem reports

VOID

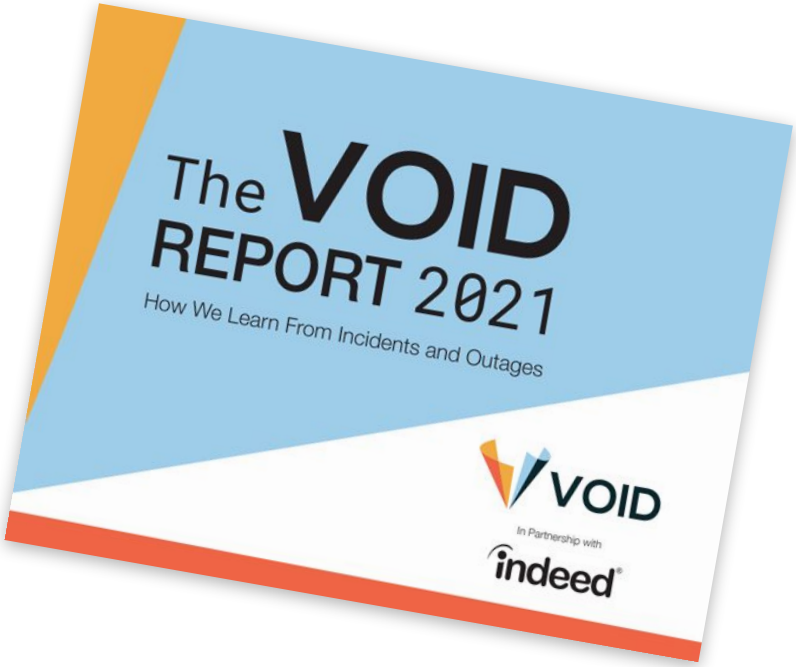
WHAT'S IN THE VOID

Meta data we collect for each report:

- **Organization**
- **Date of incident**
- **Date of report**
- **Report type:** Primary and secondary
- **Duration:** If available, either directly from the report, or calculated based on information in the report
- **Technologies involved:** This reflects what technologies were listed as contributing to the incident, if present in the report
- **Impact type:** We tag incidents based on language in the report (when available), and there can be multiple tags per report. These are intended to serve as a jumping off point for exploration, and do not represent a formal classification system.
- **Analysis format:** If noted, we track what kind of analysis is used in the incident report (Root Cause, Contributing Factors, etc). More on this in a bit…

VOID

VOID REPORT 2021



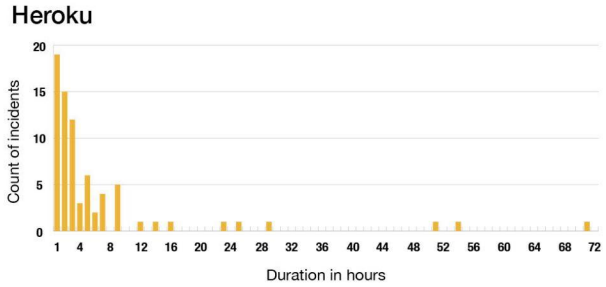
VOID

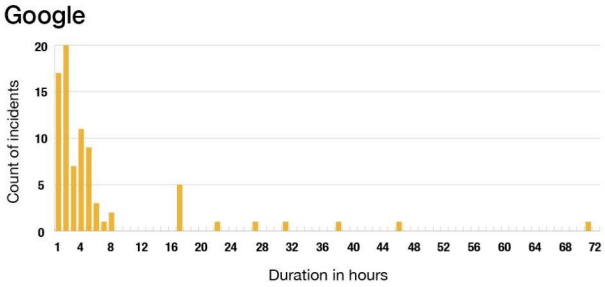
You're Good At This!

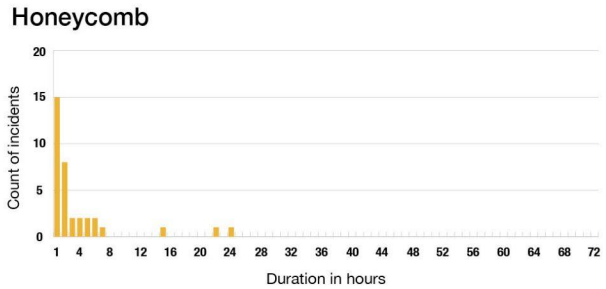
53%
of incidents in the VOID database were
RESOLVED* **IN UNDER TWO HOURS**

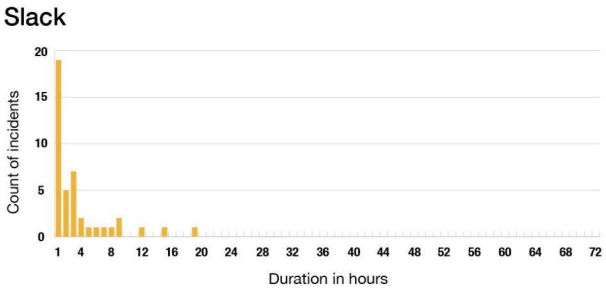
*Externally, not necessarily internally...

Distribution of Duration Data



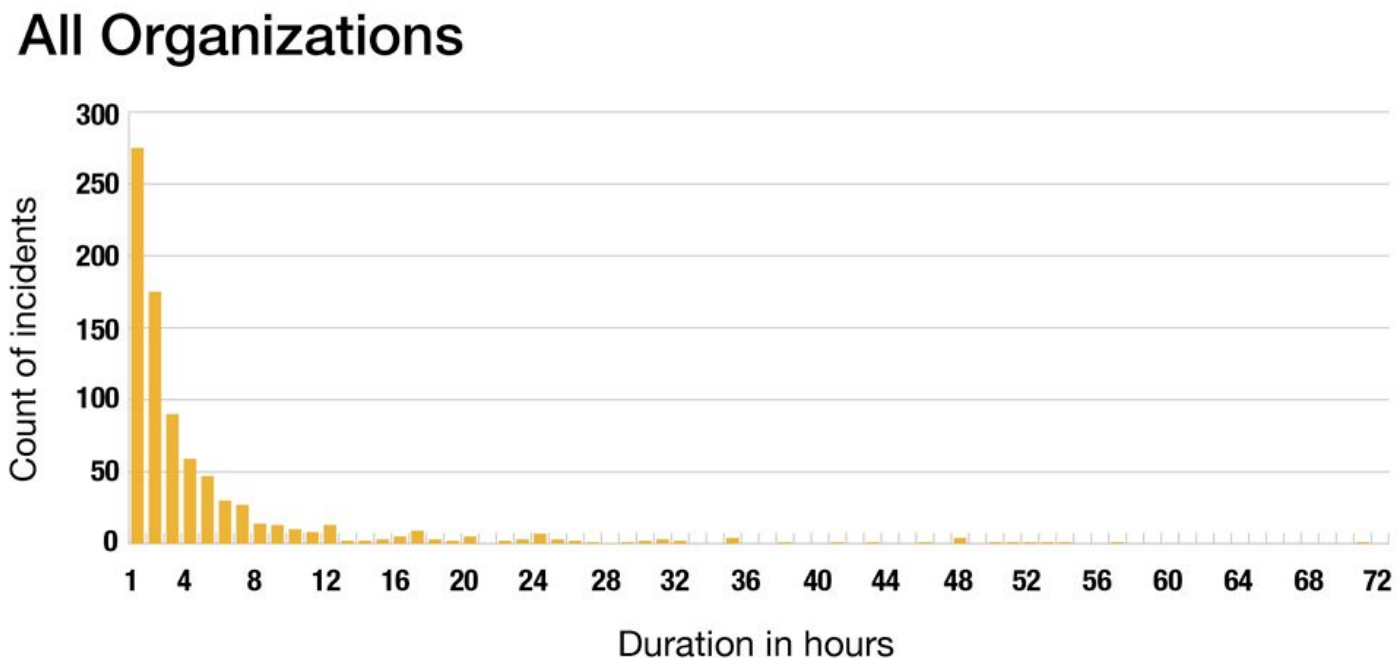




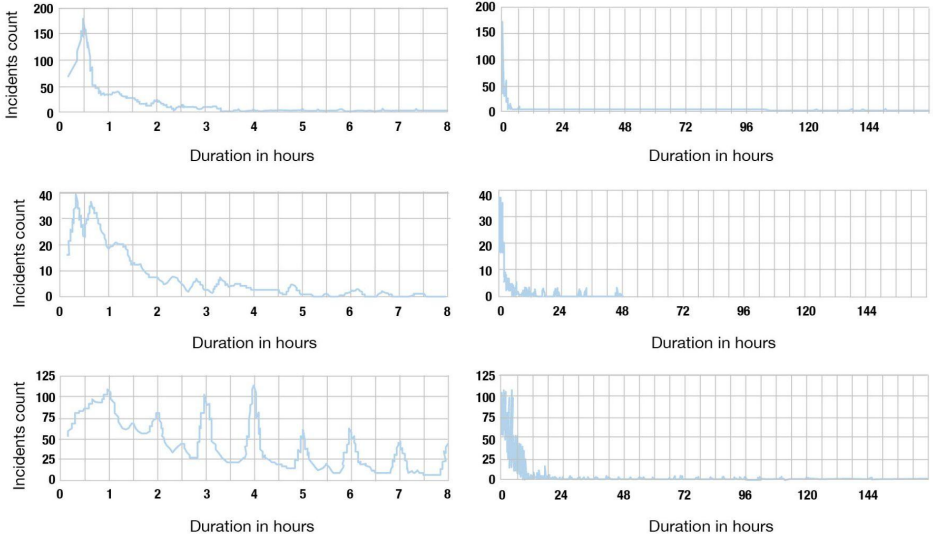


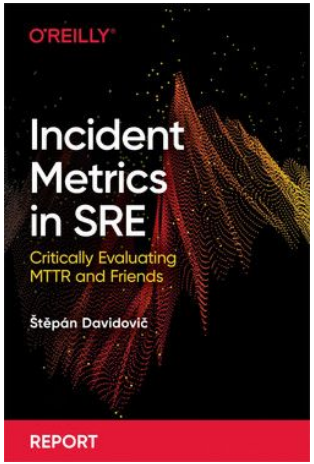
VOID

Distribution of Duration Data



Similar Distributions Elsewhere

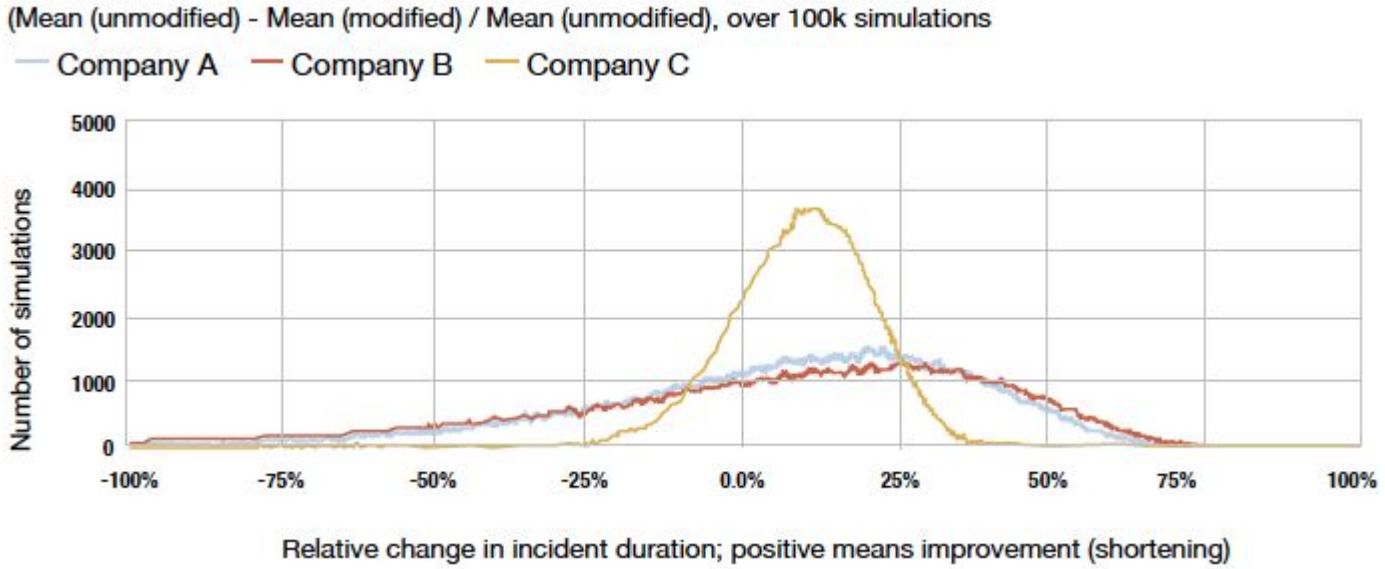




Rows are, in order, Company A (N = 798; 173 in 2019), Company B (N = 350; 103 in 2019), and Company C (N = 2,186; 609 in 2019). Columns represent each company over a short and long time frame to show the tail of the distribution. (Source: Google/O'Reilly)

VOID

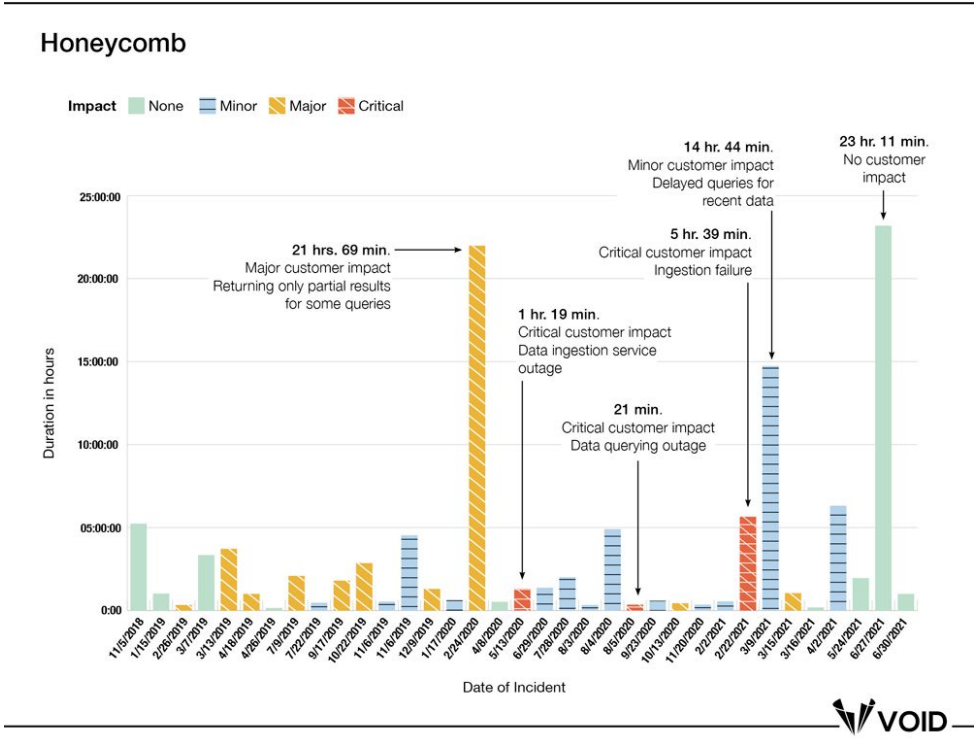
Simulating (Positive) Change in MTTR



Source: *Incident Metrics in SRE* (Google/O'Reilly)

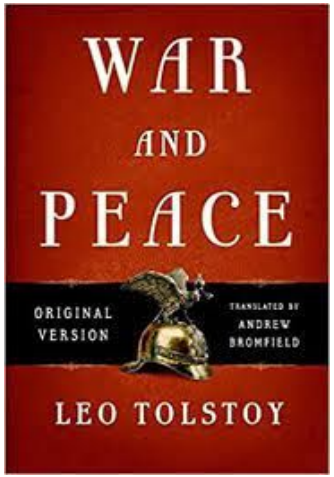
VOID

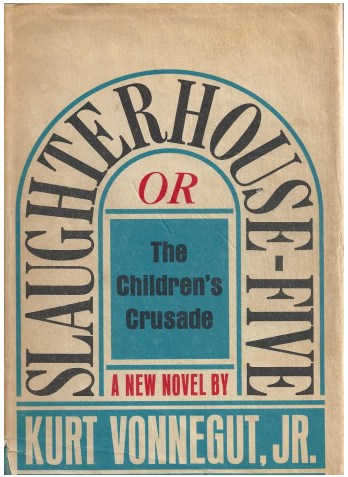
But,Longer Incidents *Are* Worse, Right?



VOID

Duration and MTTR Are Shallow Data





"Two incidents of the same length can have dramatically different levels of surprise and uncertainty in how people came to understand what was happening. They can also contain wildly different risks with respect to taking actions that are meant to mitigate or improve the situation."

—John Allspaw

VOID

Don’t Panic

Collect *Sociotechnical* Incident Data

1. Leverage **SLOs** and other sources of **customer feedback**.
2. Track **how many people and tools*** are involved in incidents, across how many groups.
3. Focus on **themes** and **narratives**—these will help you find patterns and similarities across incidents.
4. Study **near misses**.

https://www.jeli.io/howie-the-post-incident-guide/

* Aka the *cost of coordination* during incidents.

Maguire, Laura. "Controlling the Costs of Coordination in Large-scale Distributed Software Systems." Doctoral dissertation, Ohio State University, 2020.

Don’t Dismiss Near Misses

They reveal gaps in current knowledge, misaligned mental models, and the range of assumptions that practitioners have about their system before they lead to more significant incidents.

Mind The Gap(s)

Engineer noticed something was “off”:

“The issue came to our attention when a database engineer noticed that one of our database servers in the staging environment could not reconnect to the staging Consul server after the database node was restarted.”

Gap in collective knowledge w/in team:

“After looking everywhere, and asking everyone on the team, we got the definitive answer that the CA key we created a year ago for this self-signed certificate had been lost.”

Transitory knowledge, situation normal:

“These test certificates were generated for the original proof-of-concept installation for this service and were never intended to be transitioned into production. However...the expired test certificate had not been calling attention to itself.”

Production pressures, limited people:

“...a year ago, our production team was in a very different place. We were small with just four engineers, and three new team members: A manager, director, and engineer, all of whom were still onboarding.”

Surprises, gaps in knowledge:

“We were unsure why the site was even still online because if the clients and services could not connect it was unclear why anything was still working.”

Precarious conditions, safety boundaries becoming clear(er):

“Any interruption of these long-running connections would cause them to revalidate the new connections, resulting in them rejecting all new connections across the fleet.”

Efforts to fix make things worse:

“Every problem uncovered other problems and as we were troubleshooting one of our production Consul servers became unresponsive, disconnected all SSH sessions, and would not allow anyone to reconnect.”

Further safety boundaries and degraded state uncovered:

“Not having quorum in the cluster would have been dangerous when we went to restart all of the nodes, so we left it in that state for the moment.”

Inherent risk(s) in solving the problem:

Multiple solutions are considered, and all “involve the same risky restart of all services at once.”

Socio-technical realities:

“While there was some time pressure due to the risk of network connections being interrupted, we had to consider the reality of working across timezones as we planned our solution.”

Source: The Consul Outage That Never Happened (2019)

VOID

Near Misses Are Successes

"Near misses are generally more worth our time, because they come **without the pressure** of dealing with post-incident fall-out and blame, and allow a better focus on what happened."—Fred Hebert, SRE at Honeycomb

"The r/WallStreetBets events taught us many things about how we as a company work. Even though the focus of these stories revolve heavily around technical triggers, they also highlight how **every team at Reddit played an important role** in containing and mitigating these incidents." —Courtney Wang, SRE at Reddit

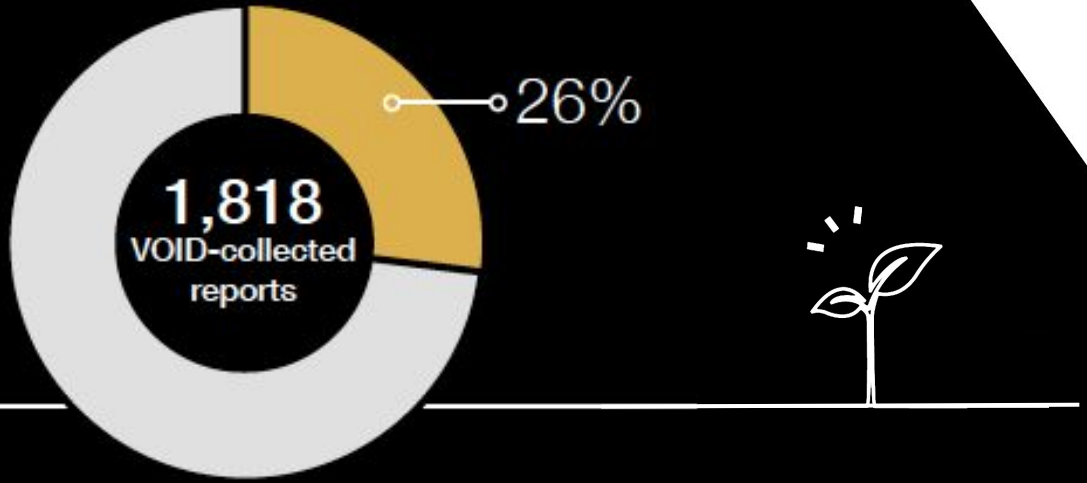
"We realized that this was a failure mode that **didn't really line up with our mental model** of how the job system breaks down. This institutional memory is a cultural and historical force, shaping the way we view problems and their solutions." —Ray Ashman, Engineer at Mailchimp

VOID

Root Cause Is For Plants, Not Software

Root Cause Analysis Data

About a quarter of the incident reports either identify a specific "root cause" or have conducted a Root Cause Analysis (RCA)



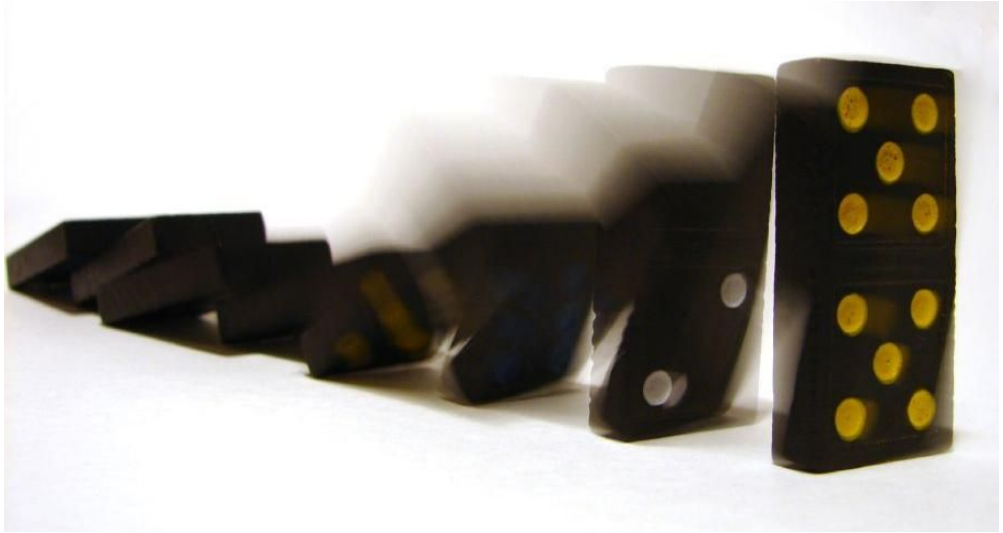
"What you call '**root cause**' is simply the place where you stop looking any further."
—Sidney Dekker

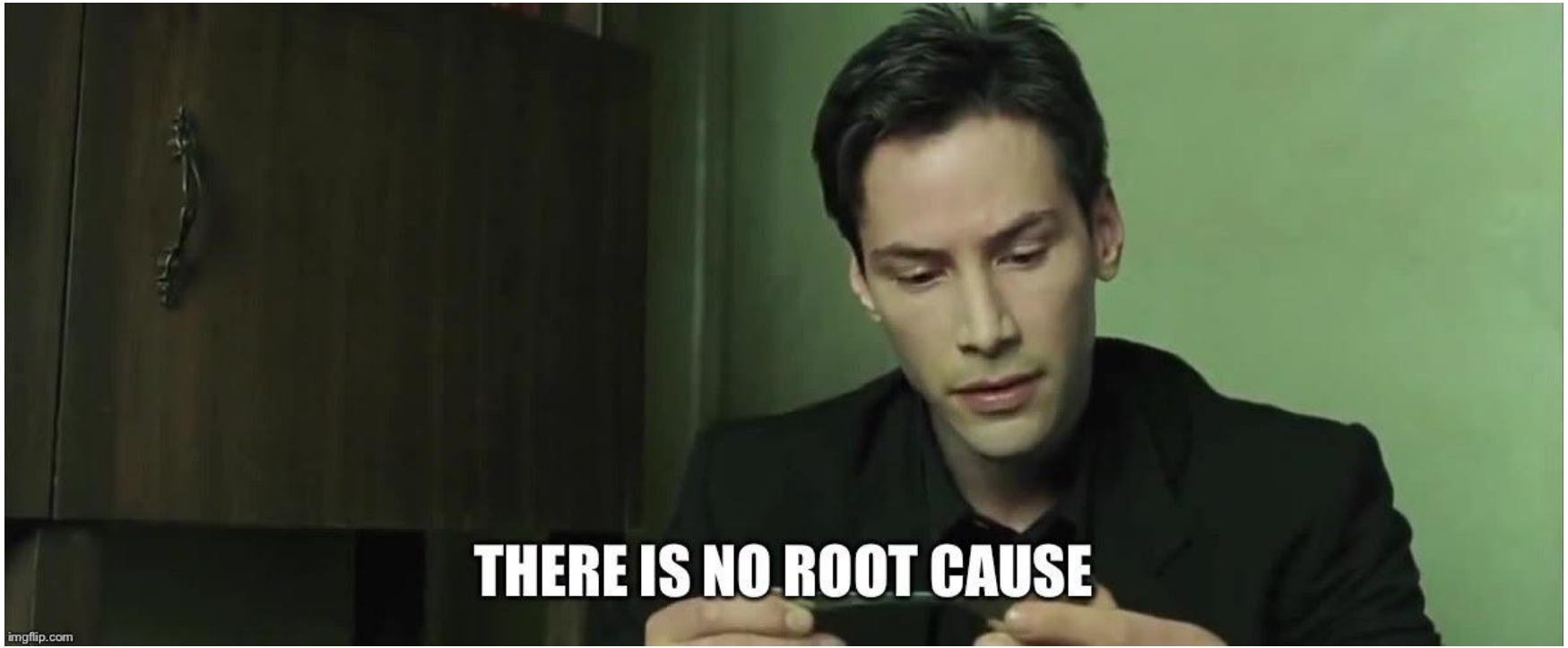
Why Care About Calling It Root Cause?

1. Creates An Artificial Stopping Point
2. Can Lay A Path to Blame

> “Complex systems run as broken systems. The system continues to function because it contains so many redundancies and because people can make it function, despite the presence of many [latent] flaws.”
>
> —Richard Cook

We Don’t Find Causes, We *Construct* Them





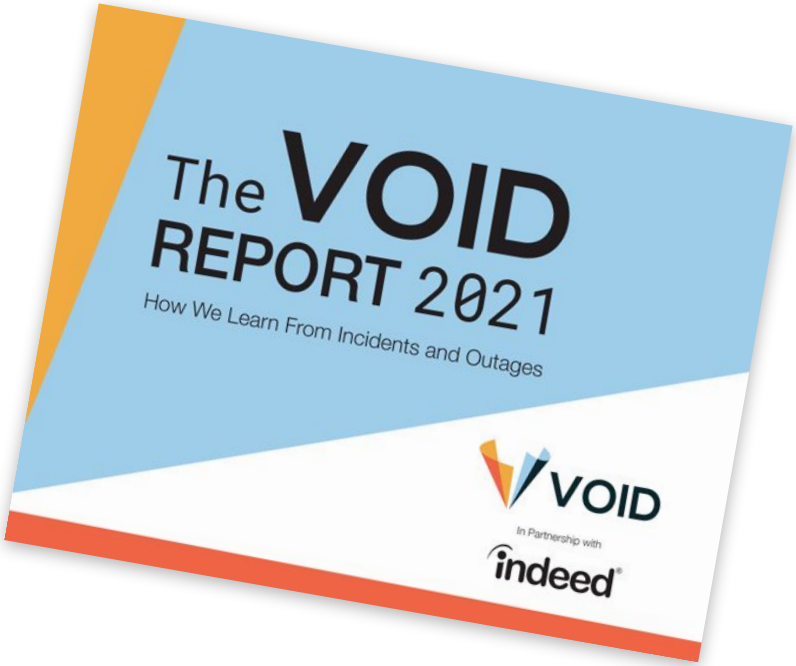
VOID

We need a new mindset, toolset, and skillset for talking about, analyzing, learning from, and sharing incidents.

A New Approach

1. Treat Incidents as Opportunities to Learn
2. Favor In-depth Analysis Over Shallow Metrics
3. Treat Humans as Solutions, Not Problems
4. Study What Goes Right Along With What Goes Wrong

DOWNLOAD THE 2021 REPORT



https://www.thevoid.community/report

VOID