Rethinking Reliability:

What We Can (and Can't) Learn From Incident Metrics

Courtney Nash

Internet Incident Librarian, Verica

@courtneynash

DEVOPS ENTERPRISE SUMMIT

VOID

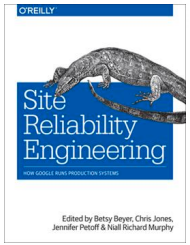
Courtney Nash

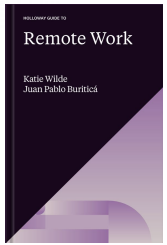
Internet Incident Librarian

Verica

@courtneynash

O'REILLY®
Velocity





amazon

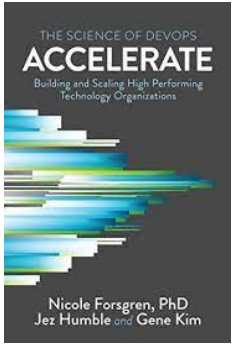
fastly

Microsoft

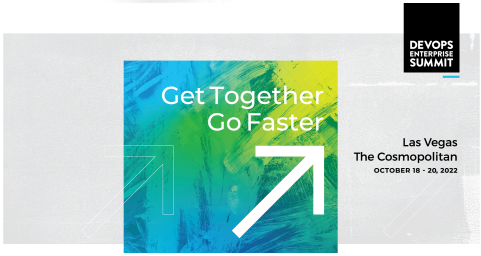
HOLLOWAY

VOID

We've Come A Long Way…



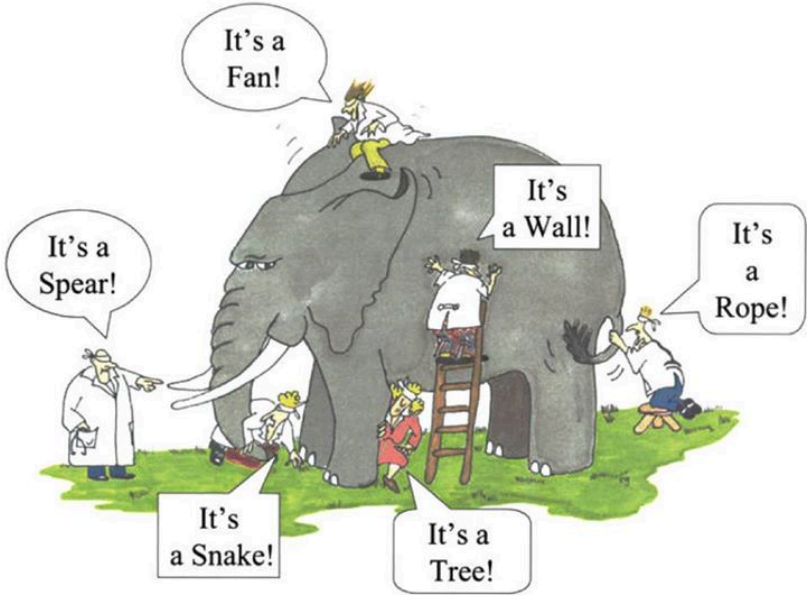
2020-2022



@courtneynash

VOID

Let’s Talk About MTTR



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.

What's In The VOID

Almost 10K public incident reports from over 600 organizations, from 2008 up to present day.

In a variety of formats:

- Social media posts
- Status pages
- Blog posts
- Conference talks
- News articles
- Tweets
- Comprehensive retrospectives/postmortem reports

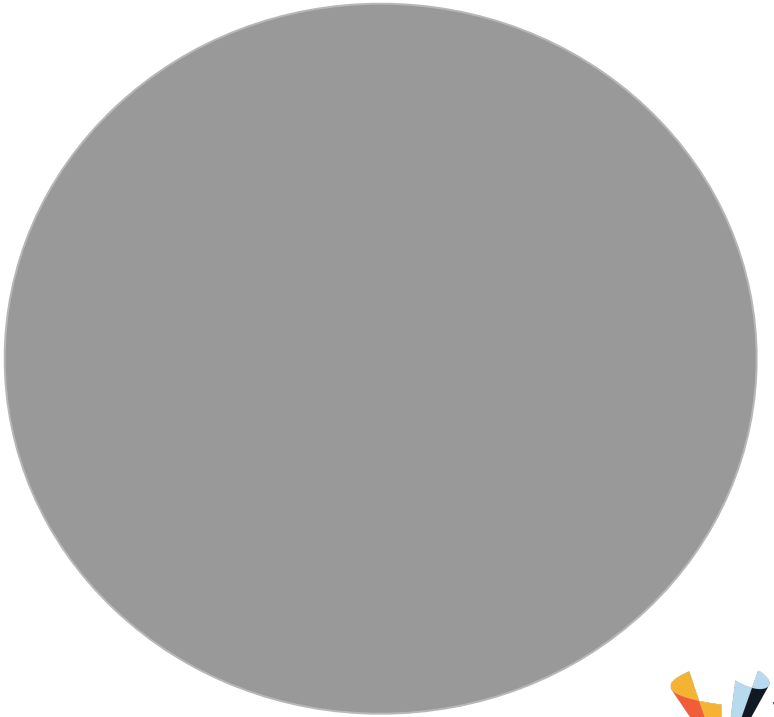
VOID Metadata

- **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).
- **Severity**: Typically available from status pages (e.g. None, Minor, Major, Critical)

Duration: Grey Data

Duration is

- High in variability, low in fidelity
- Fuzzy on both ends
- Sometimes automated, often not
- Sometimes updated, sometimes not
- A lagging indicator
- Inherently subjective

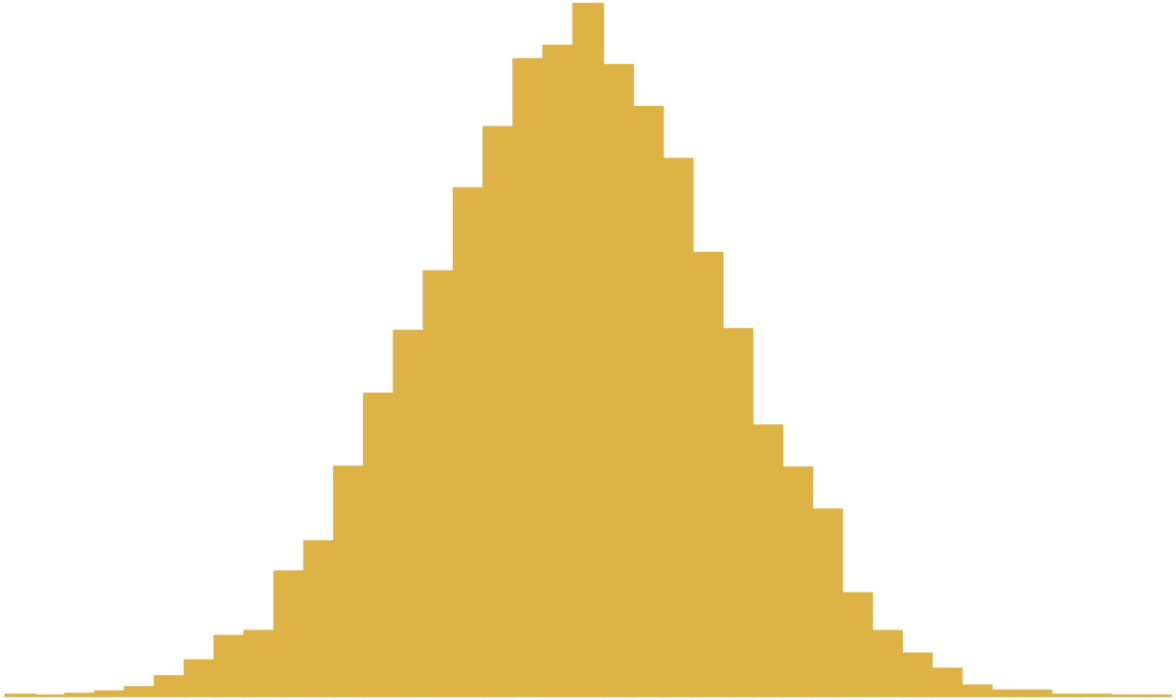


VOID

Grey Data In Action

Company	MeanTTR (in hours)					
	All time	2021	2022	Q1 2022	Q2 2022	Q3 2022
Atlassian	12	22	15	15	29	7
Azure	13	12	20	24	23	13
Cloudflare	6	4	2	3	2	3
Dropbox	7	21	7	5	7	9
Duo	11	10	10	8	8	14
Fastly	10	15	6	3	4	10
GitLab	11	7	6	6	3	3
Hashicorp	8	4	16	8	36	6
Honeycomb	5	4	4	3	1	5
New Relic	2	2	2	2	2	2
Trello	9	13	12	6	1	26
Wistia	14	26	56	27	48	116

Distribution Matters



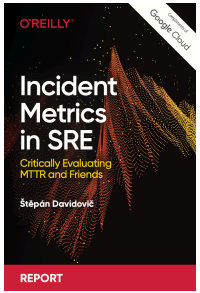
Source: VOID

VOID

Detecting Change In MTTR

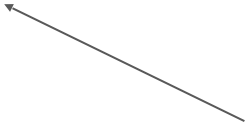
Monte Carlo simulations

- Decrease MTTR, run across 100K simulations
- High variability (skew) leads to lack of accuracy in detecting purposeful changes in MTTR
- Even when introducing improvements in MTTR, about ⅓ of the time the detected change in MTTR was negative!



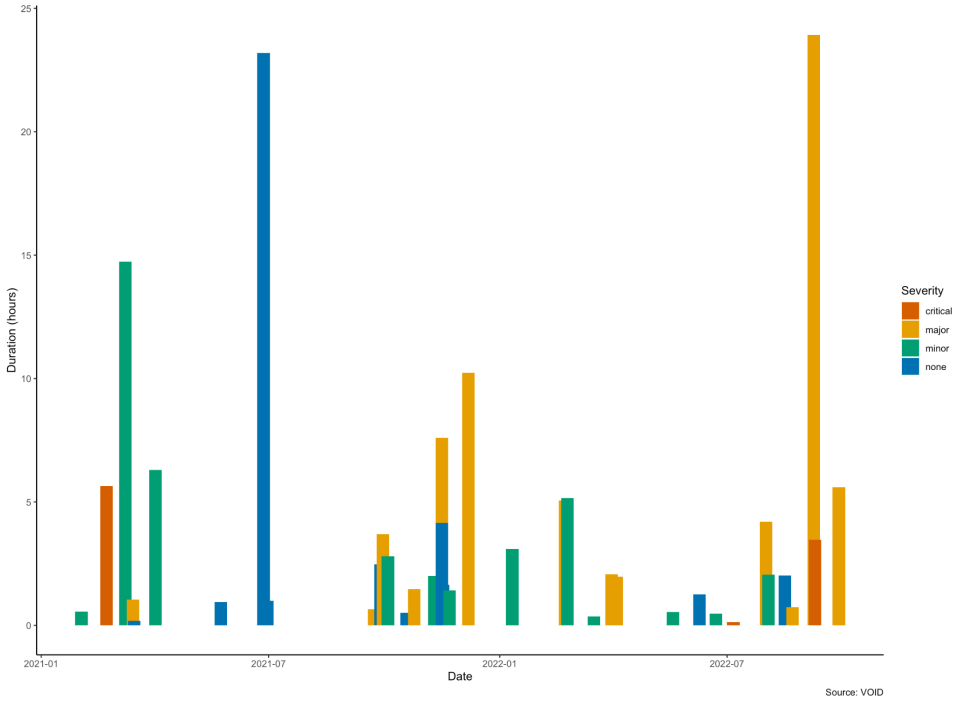
Duration & MTTR Can’t Tell You

- How reliable your software or systems are
- How agile/effective your team or organization is
- If you’re getting better at responding to incidents
- Whether the next one will be longer or shorter
- How “bad” any given incident is



VOID

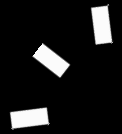
Duration & Severity: Not Related



We analyzed status page data across almost 7K incidents from 10 different companies.

- Only 2 of them showed very weak correlations between duration and severity.
- R = -.18 and -.17, respectively ($p < .05$)

VOID



“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

Metrics Should Facilitate Decisions

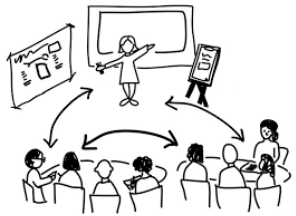
If Not MTTR, Then What?

Incident Analysis

You don’t need MTTR to tell you to go look at your incidents for deeper understanding of what is happening with your systems.

Your Systems Are Sociotechnical

So you need to collect sociotechnical (often qualitative) data!



Cost(s) of Coordination
Number of people hands on
Across how many unique teams
Using which tools
Via how many channels
Concurrent incidents?
PR/Comms involved?

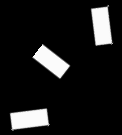
Near Misses
How many vs actual incidents:

- Sources of adaptive capacity
- Knowledge gaps
- Assumptions
- Misaligned mental models
- Safety margins

Participation
Number of people reading writeups
Number of people voluntarily attending post-incident review meetings.
Number of links to write-ups from:

- code comments & commit messages
- architecture diagrams
- other related incident write-ups

VOID



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

VOID

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

How You Can Help

1. **Analyze Your Incidents**
    - HOWIE: https://www.jeli.io/howie/welcome
    - Adaptive Capacity Labs:https://www.adaptivecapacitylabs.com/
2. **Submit Them To The VOID**
    - Form: https://www.thevoid.community/submit-incident-report
    - Email: https://www.thevoid.community/contact
3. **Become A Member**: https://www.thevoid.community/partners

VOID

VOID Reports

https://www.thevoid.community/report