Incident Archeology

Finding Value in the Paperwork and Narratives of the past

Clint Byrum - https://fewbar.com
@SpamapS@fosstodon.org

Who am I?

Staff Engineer and Incident Manager On Call (IMOC) at Spotify *(not speaking for Spotify, words are my own!)*

25+ years in tech

Have had many titles, somehow never SRE

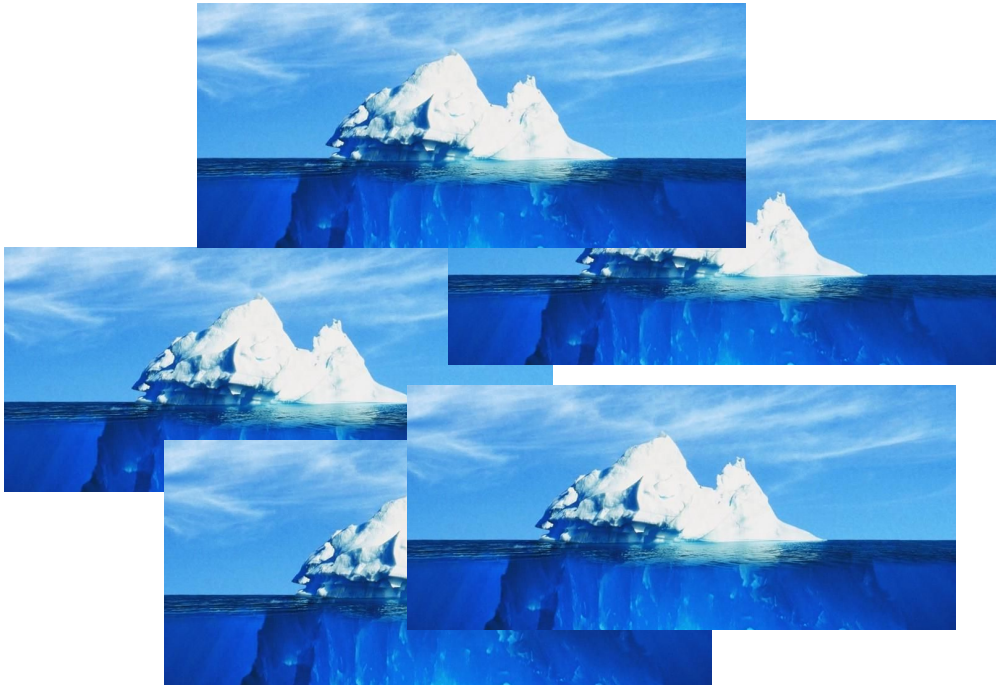
Why am I here?

- Promote Psychological Safety
- Share Our Process
- To inspire and collaborate with you!

A word on sharing in public

- Companies default to not sharing
- It takes effort
- Join me in making sharing the norm!

Breadth, not depth

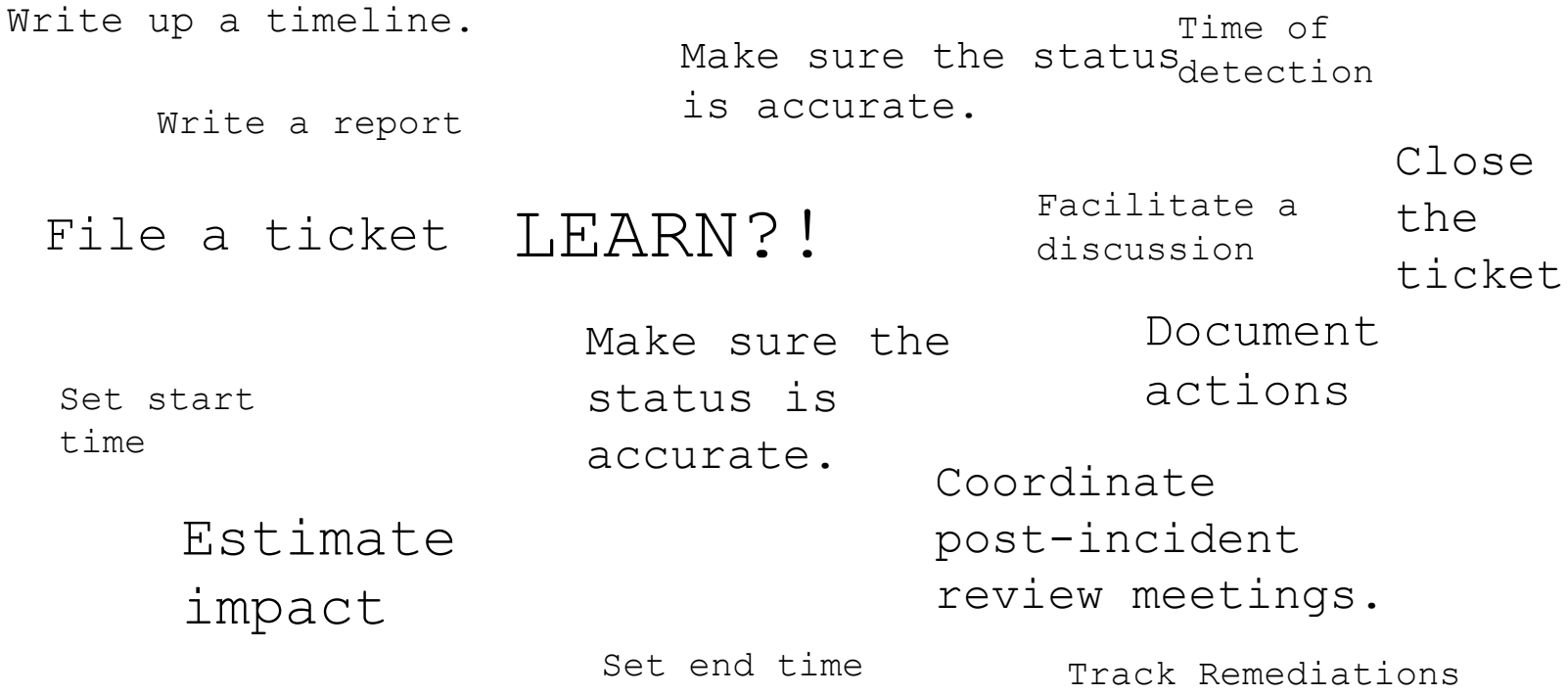


Are incidents just paperwork?

So what's the value of the ~~paperwork~~ process?

- Communication
- Accountability
- Coordination
- Maybe, if you're lucky: learning

Use the new cover sheet on your TPS reports…



Now we see the violence inherent in the system

Start Digging!

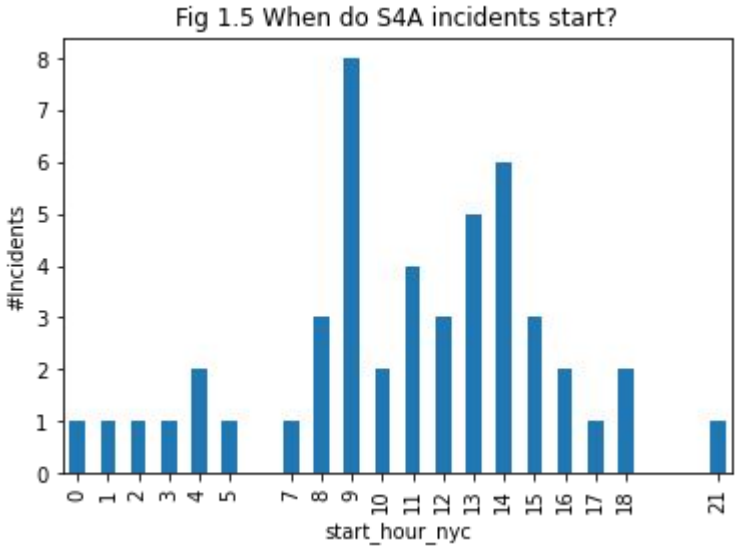
Our first hypothesis

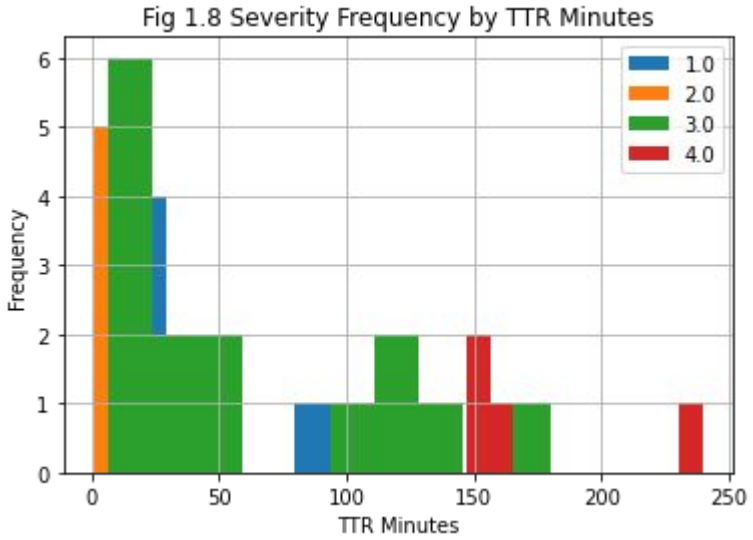
"After-hours will have high MTTR and Complexity."

- Falsifiable!
- Built on shaky ground of MTTR
- Even shakier ground: Measuring complexity

Our first hypothesis

"After-hours will have high MTTR and Complexity."





Complexity is so simple

“How hard was this to fix? Did it have a clear and obvious resolution? Were senior engineers required to fix it? Graded on a 1-5 scale, 1 being fairly simple, 5 being the hardest known solution.”

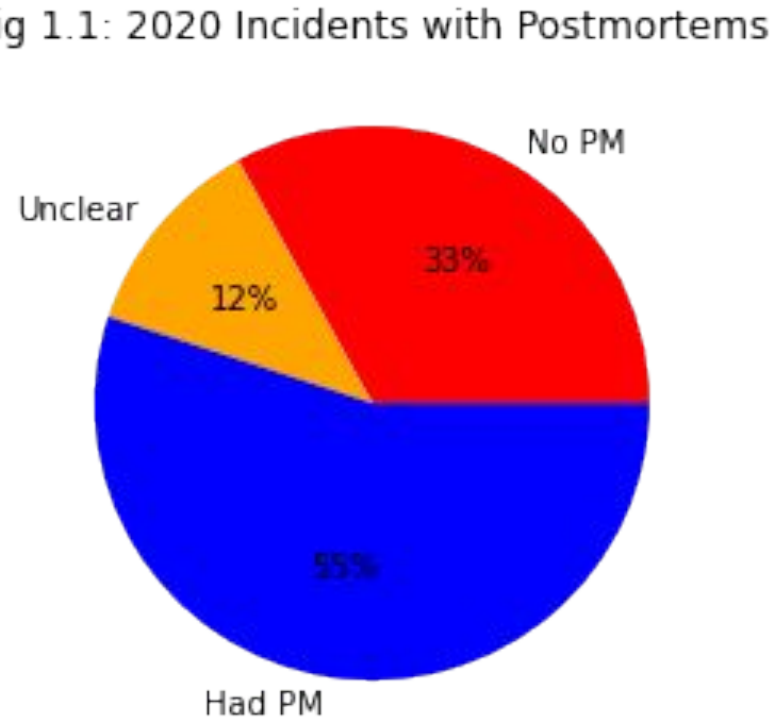
2020 Hypothesis #2

"At least 50% of S4A incidents will have a high avoidability"

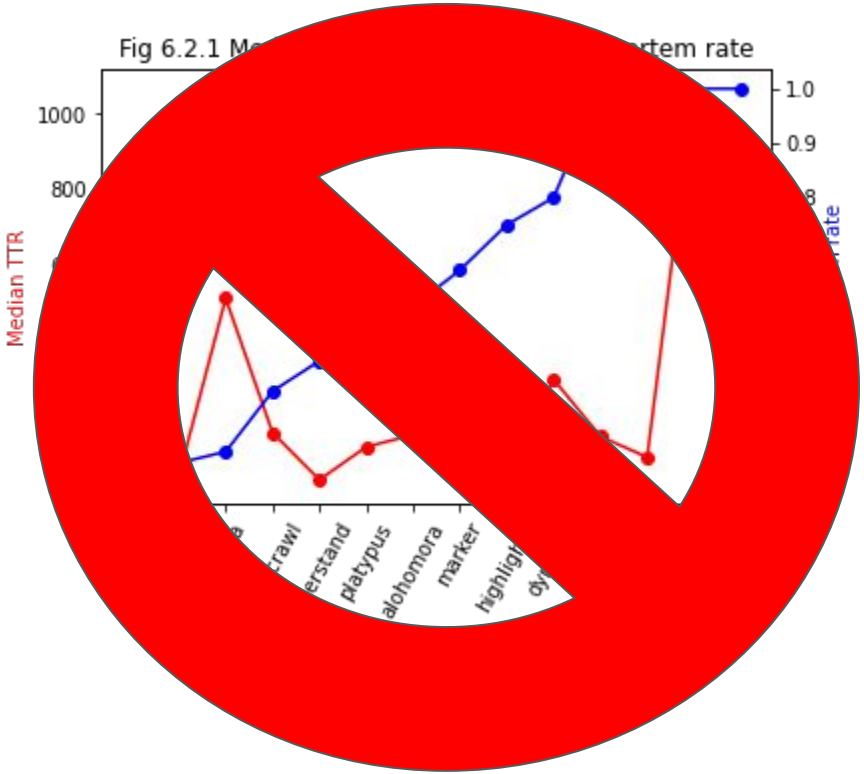
Avoidability being:

"How easily could S4A developers/operators have avoided this with pro-active work. Did we see it coming, and fail to act, or was this an unpredictable event? 1-5 scale, 1 being a very hard event to see coming, 5 being an inevitable event that was identified before it happened."

This is when we found out nobody likes paperwork



Data science enters the chat



We got better right?

2021 Hypothesis #1:

"Postmortems are the norm"

- Barely falsifiable (Define "norm")
- Easy to measure from a binary perspective

Disappointing Findings

Fig 1.1: 2020 Incidents with Postmortems

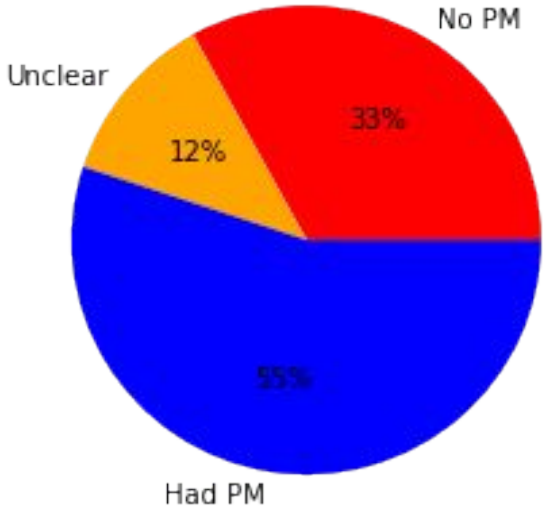
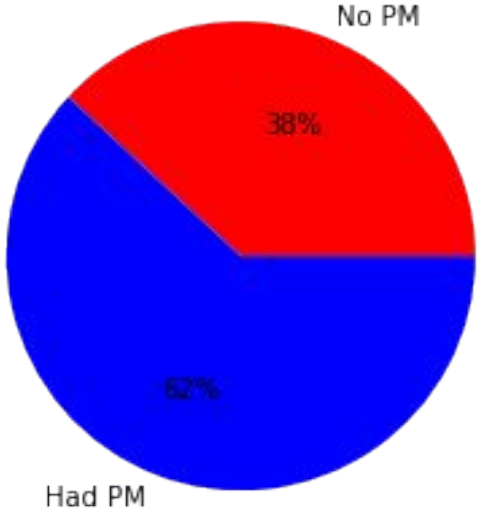


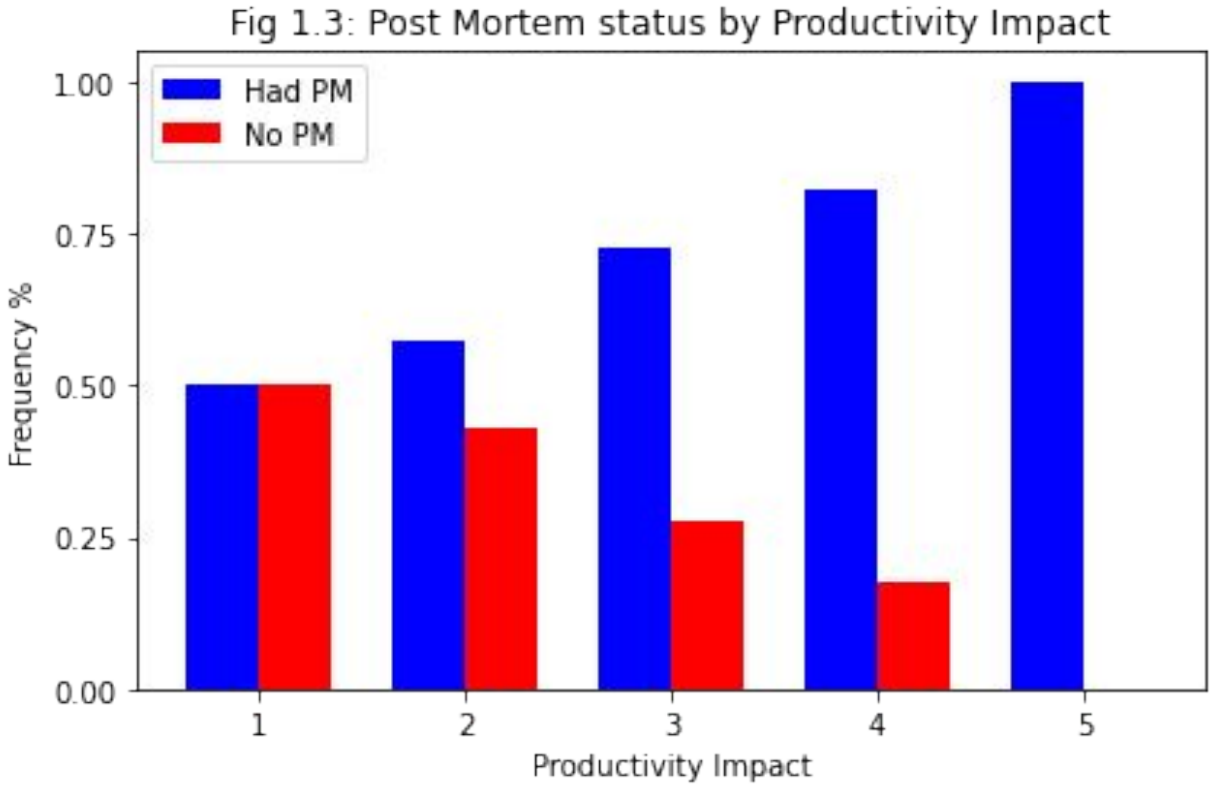
Fig 1.0: 2021 Incidents with Postmortems



Productivity Impact

Score	Examples during work hours	Examples during non-work hours
1	A single contributor quickly resolved the issue during work hours (i.e. a pin)	N/A
2	Contributor(s) within one squad spent a minor portion of a single workday investigating	Single contributor quickly resolves the incident or it auto resolves / Alert has a clear mitigation that doesn't require much thought
3	Contributor(s) within one squad spent most of a single working day investigating or multiple squads for a minor portion	Single contributor spends some time investigating before finding a mitigation, multiple on-call paged but don't participate in mitigation
4	Contributors within multiple squads investigated for up to one day / Contributor(s) within one squad spent multiple days investigating	Multiple contributors are involved before a mitigation is found
5	Contributors within multiple squads spent multiple days investigating	Multiple contributors across multiple days

People want answers!



How to do Incident Archaeology

1. Go find some **artifacts**!
2. Decide how much **time you can commit** to studying them
3. **Hypothesize** about it
4. Make a **methodology** that will fit in the time box
5. Run it by a **data scientist**
6. Break up the artifacts into a list and **study each** one
7. **Analyze** the data
8. **Write** it up, learn, **share**, rejoice!

Guiding Principles

- We aren’t fixing things
- We analyze what we can find
- The timebox must be respected
- Transparency is critical to building trust

Correlations are really problematic

- Your sample size is really small
- Your population is often unknown
- Sane p-values are to come by
- More of a census

Stuff we've learned that we weren't looking for

- Nobody knows what the "start" or "end" time of incidents means
  - And because they're defaulted, 75% of users never bother to adjust them
- Uptime success can hide massive problems with productivity
- 80% of our incidents are declared during business hours!
- Only 30% of declared incidents are local change failures

Rigorous investigations

- Write down any novel tactics for investigating
- Pair up frequently to get on the same page
- Review each other's’ work to learn and hold each other accountable

The Process

1. Find out what data is available
    a. Measure its size
    b. Decide on the maximum time you have to study each artifact
2. Develop hypotheses about the data
3. Define demographic, quantitative and qualitative measures that would help prove/disprove hypotheses
    a. Keep in mind that correlations are very hard with such a small sample size.
4. Sample some data randomly and develop rubrics for the measures
    a. Record how long it takes to fill rubrics/measures. Balance time for whole set with depth of data. More fields == more time.
5. Iterate on the fields/rubrics after sampling, drop any that are too hard to investigate or don't support hypotheses.
6. Break up the artifacts into a list and study each one. Pairing is probably more effective than spot-checking.
7. Clean up data and work with data science to help prove/disprove your hypotheses.
8. Write it up, share with everyone!

Time boxing

- Focus on things you can determine in just a few minutes
- Have a confidence score so you can extract partial data
- Do the easy stuff first!