Learning at Scale is Hard!

Outage Pattern Analysis and Dirty Data

Tanner Lund
Microsoft Azure SRE

Photo: Rachel Chapman (CC)

Timeline

Impact Start * [redacted] PDT

Detection [redacted] PDT

Eng. Engaged [redacted] PDT

Mitigation * [redacted] PDT

Impact

Customer Impact

[redacted]

QoS/SLA Impact

[redacted]

Service Responsible

[redacted]

Services Impacted

[redacted]

Root Cause

Root Cause Title

[redacted]

Root Cause Details

[redacted]

Repeat Outage

Detection and Mitigation

Detection Source

[redacted]

Detection Details

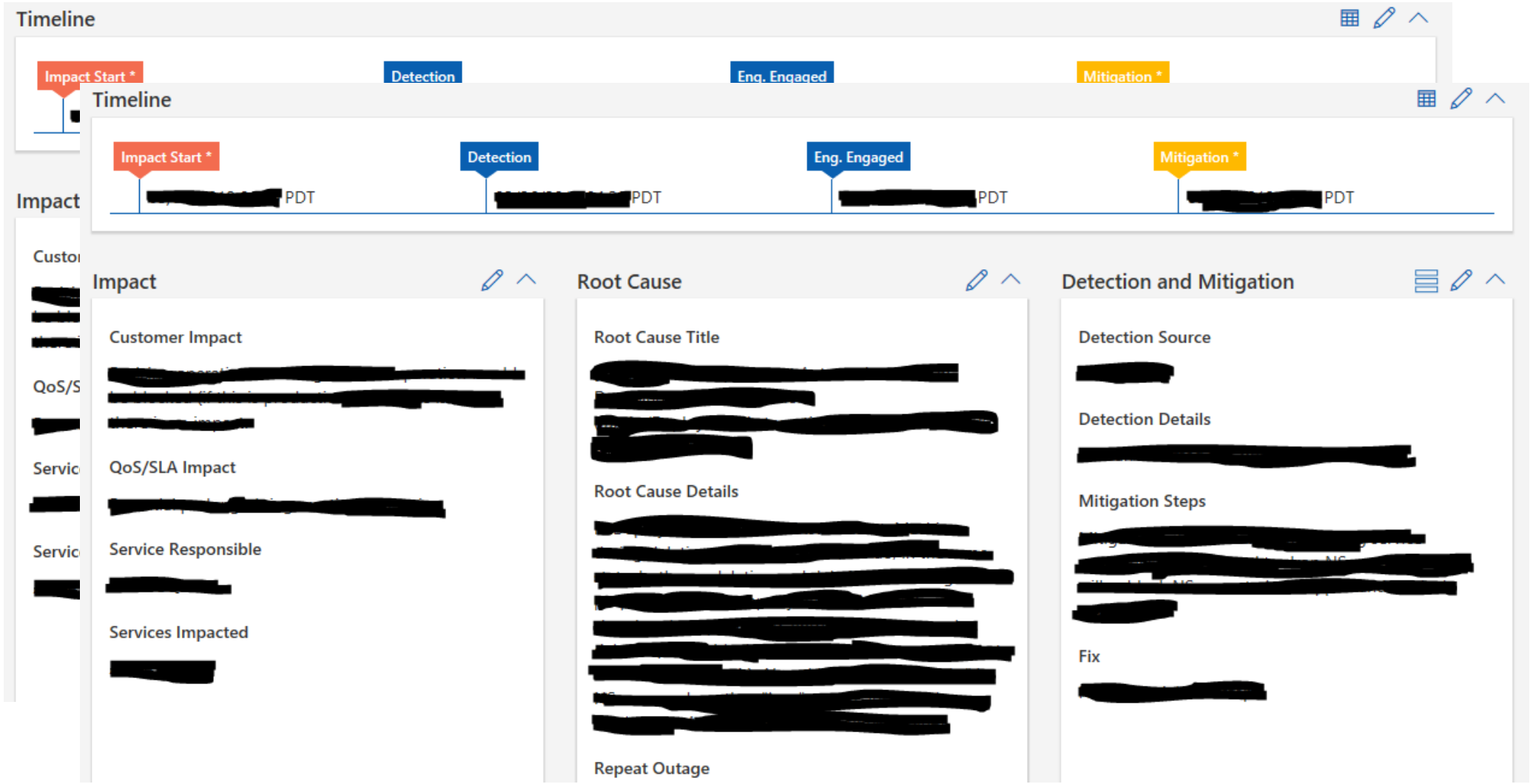
[redacted]

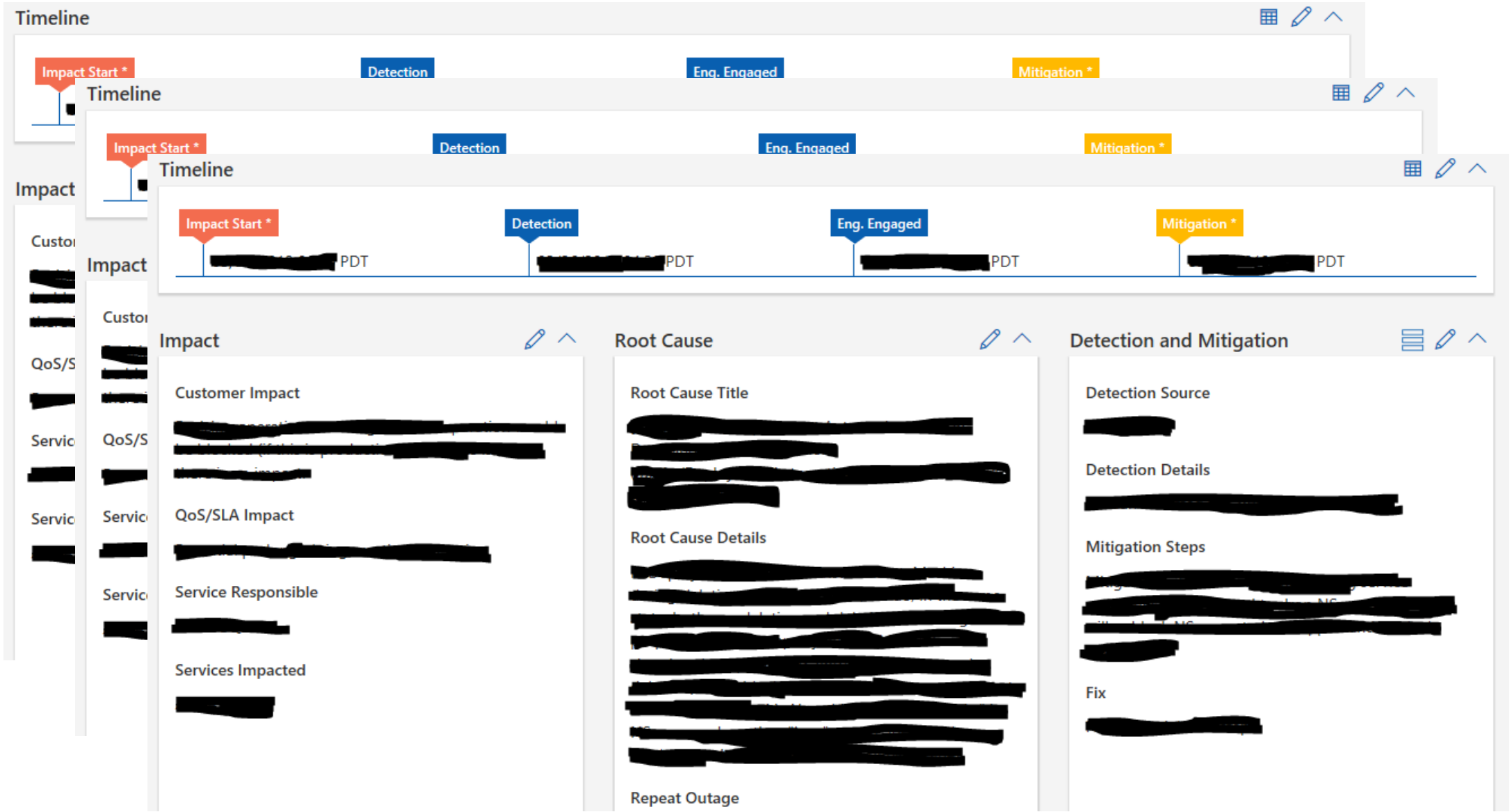
Mitigation Steps

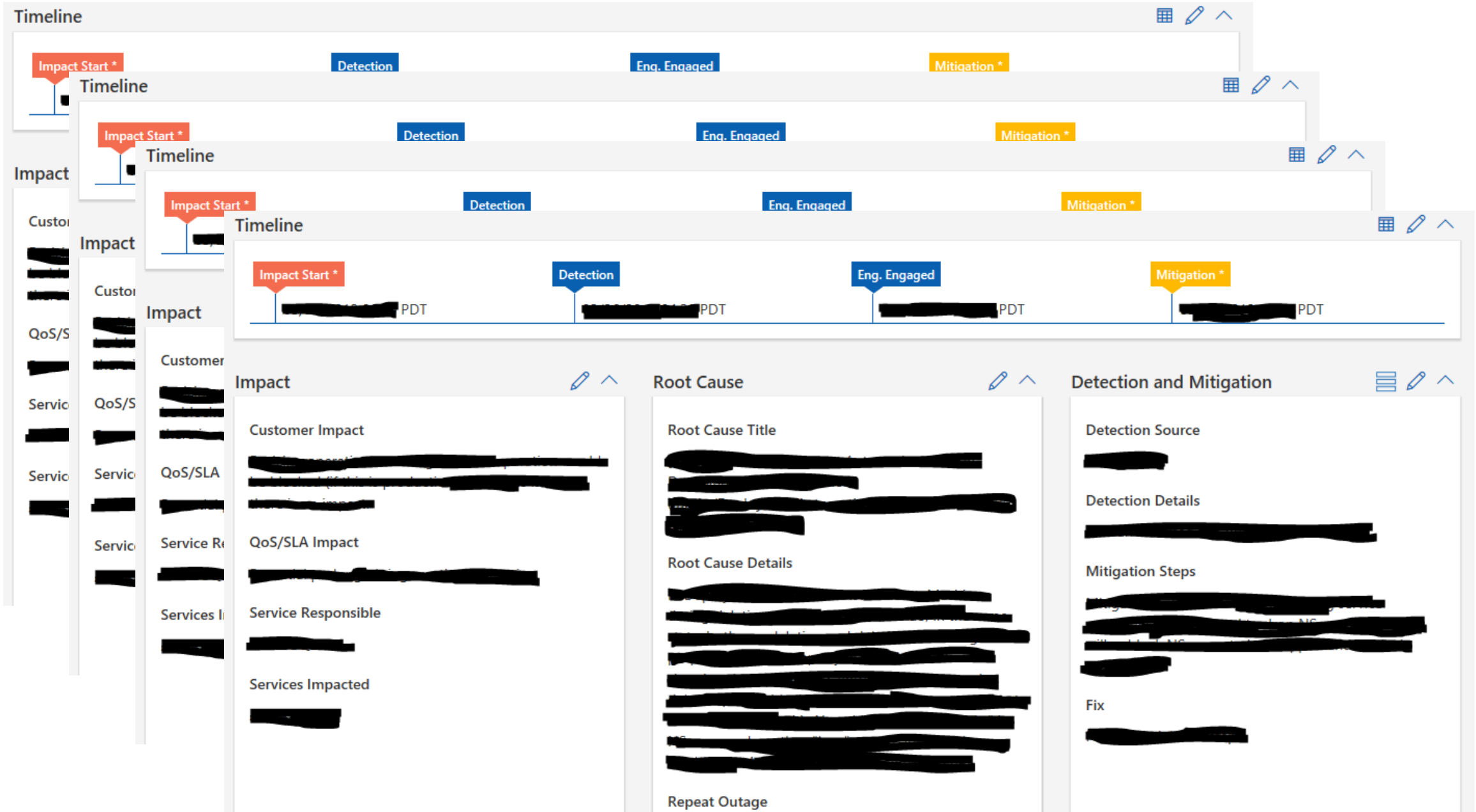
[redacted]

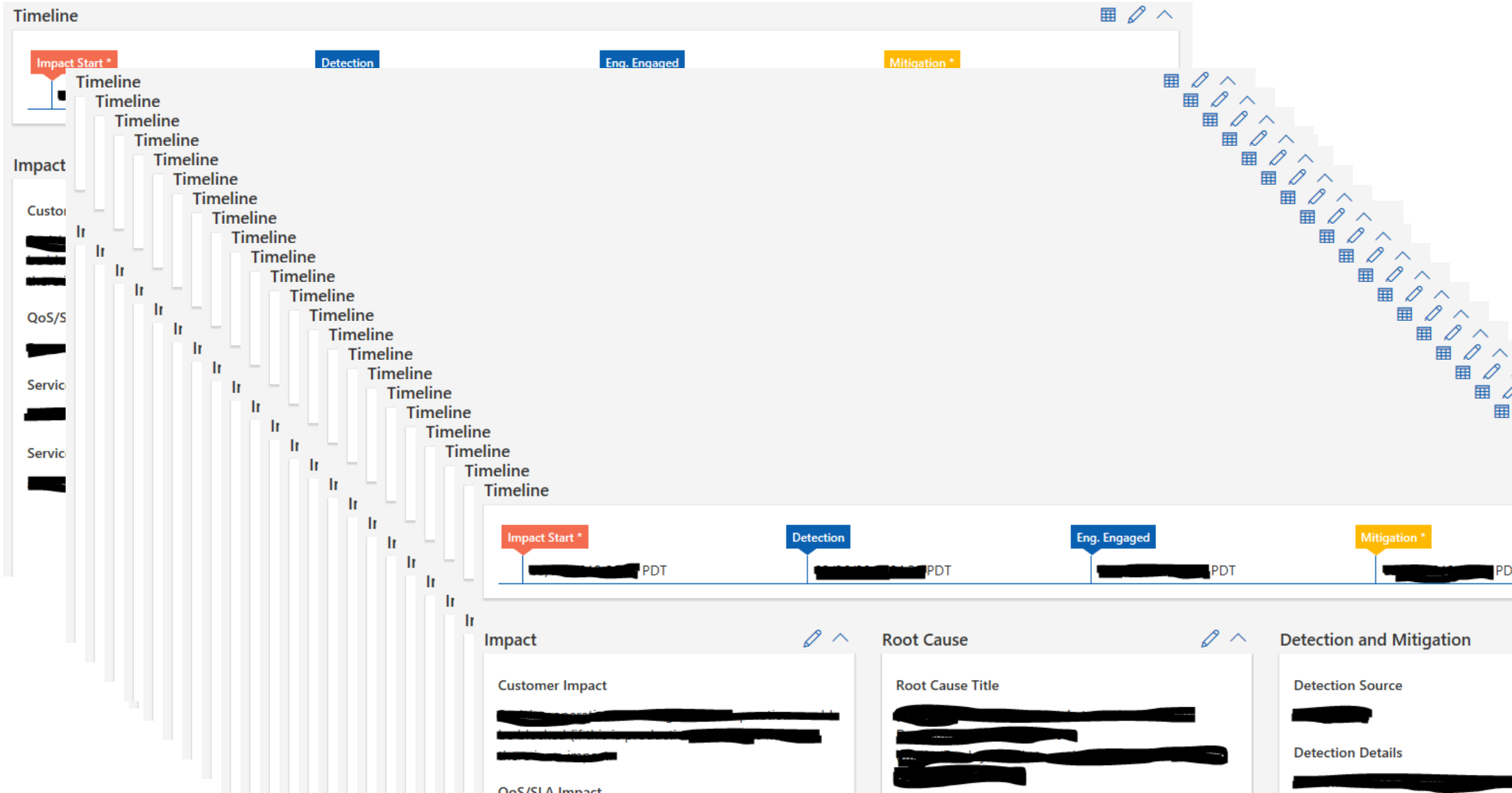
Fix

[redacted]


Timeline
Impact Start *
Detection
Eng. Engaged
Mitigation *
Timeline
Impact Start *
Detection
Eng. Engaged
Mitigation *
PDT
PDT
PDT
PDT
Impact
Root Cause
Detection and Mitigation
Customer Impact
QoS/SLA Impact
Service Responsible
Services Impacted
Root Cause Title
Root Cause Details
Repeat Outage
Detection Source
Detection Details
Mitigation Steps
Fix


Timeline
Impact Start *
Detection
Eng. Engaged
Mitigation *
Timeline
Impact Start *
Detection
Eng. Engaged
Mitigation *
Timeline
Impact Start *
Detection
Eng. Engaged
Mitigation *
PDT
PDT
PDT
PDT
Impact
Customer Impact
QoS/SLA Impact
Service Responsible
Services Impacted
Root Cause
Root Cause Title
Root Cause Details
Repeat Outage
Detection and Mitigation
Detection Source
Detection Details
Mitigation Steps
Fix


Timeline
Impact Start *
Detection
Eng. Engaged
Mitigation *
Timeline
Impact Start *
Detection
Eng. Engaged
Mitigation *
Timeline
Impact Start *
Detection
Eng. Engaged
Mitigation *
Timeline
Impact Start *
PDT
Detection
PDT
Eng. Engaged
PDT
Mitigation *
PDT
Impact
Customer Impact
QoS/SLA Impact
Service Responsible
Services Impacted
Root Cause
Root Cause Title
Root Cause Details
Repeat Outage
Detection and Mitigation
Detection Source
Detection Details
Mitigation Steps
Fix


Timeline
Impact Start *
Detection
Eng. Engaged
Mitigation *
Timeline
Impact
Custo
QoS/S
Servic
Servic
Impact Start *
PDT
Detection
PDT
Eng. Engaged
PDT
Mitigation *
PD
Impact
Customer Impact
Root Cause
Root Cause Title
Detection and Mitigation
Detection Source
Detection Details

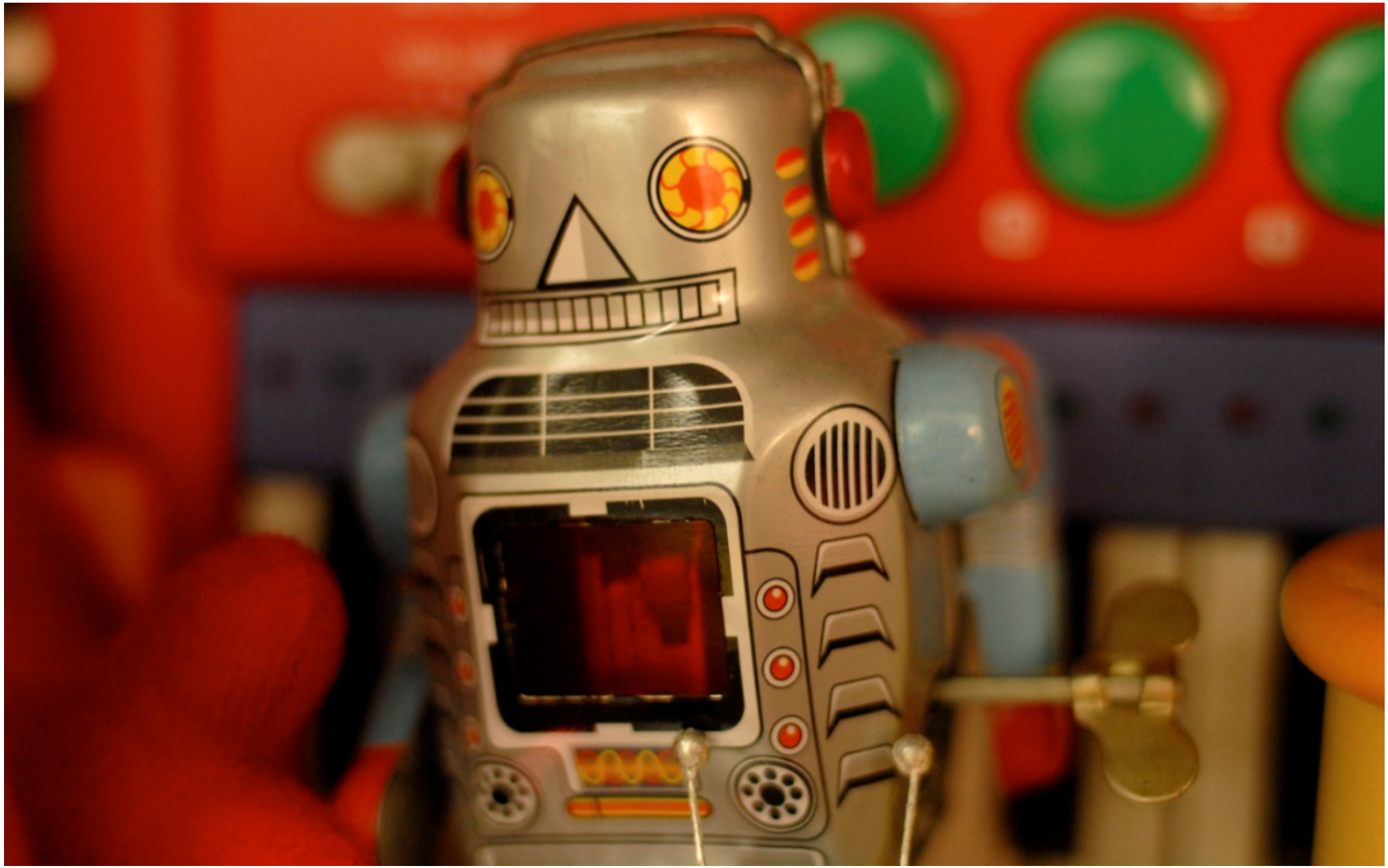


Photo: Mo Riza (CC)

IMAGINE

IMAGINE

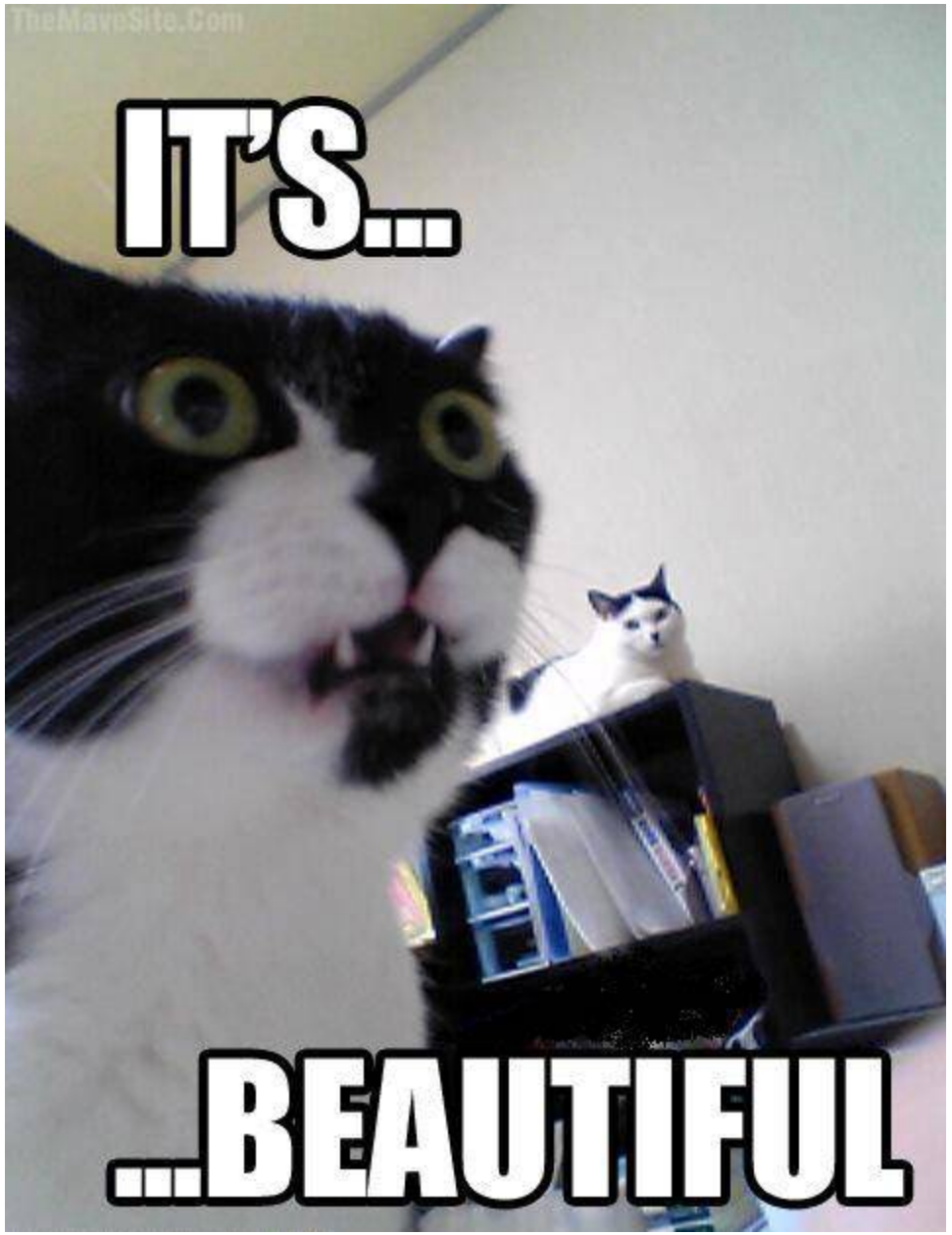
Photo: Rachel Chapman (CC)

Learning (From Failure) At Scale

Trends: Identified

Antipatterns: Quashed

Reliability Work:
~~Actually Gets Done~~
Appropriately Prioritized


IT'S...
...BEAUTIFUL

Data Scientists:

Problem Management

Problem: “The cause of one or more incidents” – Information Technology Infrastructure Library (ITIL)

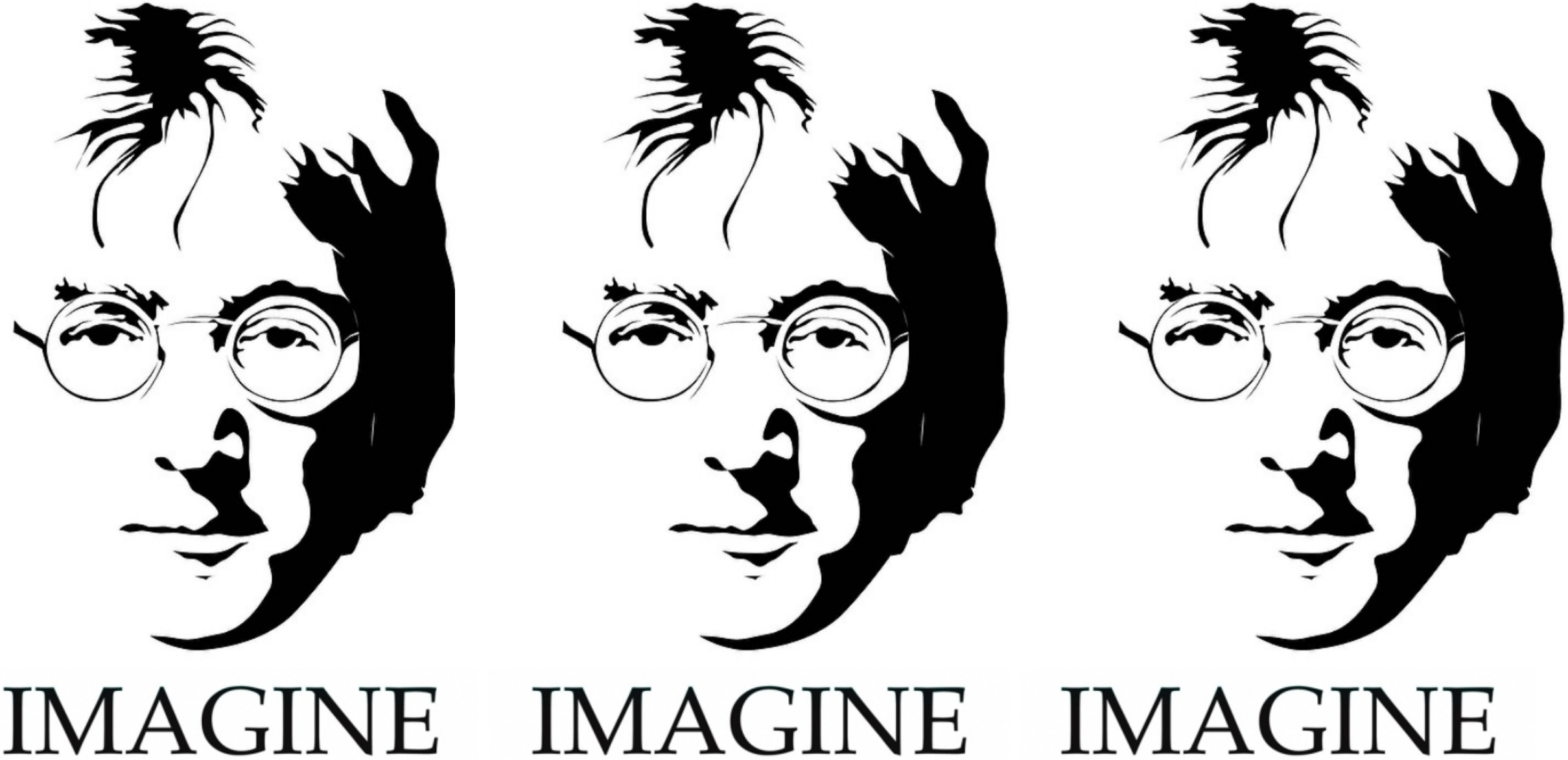


Photo: Rachel Chapman (CC)

Sharing is caring!

Gathering data

Selecting models

Training said models

Evaluating models

You know what was harder?

Knowing what we're actually looking for.

¯\(°_o)/¯

IDK, something amazing!

Fundamental Issue: ROOT CAUSES

Categories	SubCategories	Definition
Architecture	InsufficientRedundancy	Lack of appropriate or sufficient redundancy design in the service
Architecture	DesignLimitiation	Design or architecture flaw, or limitiation
Architecture	CapacityModeling	Capacity testing/tipping points Capacity threshold (TPS , etc.) was exceeded in an unanticpated manner (not the
Certs	Certs	no human factors in certs - lack of automated hands off cert upgrade mechanism causes or exacerbates impact
Code	NewCodeBug	a day 0 or previously unknown bug contributed to outage cause
Code	Unit Test	Missing Unit test
Config	Config	OneConfig - outage caused/excerabated by not having a System of record for everything in production, it's curre
Config	Version Management	version/ change mgmt - difference in versions of bits, wrong bits deployed, or wrong sequecing of versions cause
Dependency	Dependency	Dependency understanding - A lack of understanding of dependencies between components or features caused
Deployment	BacklogDeploy	RCA prevention item backlogs not being addressed - issue was known, had a repair, fix had been checked in, but
Deployment	UnifiedDeployment	Unified DEPLOYMENT - lack of a central, coordinated, automaticlaly scheduled and conflict resolving deployment
Deployment	FastRollout	fast global rollout - lack of a safe automated hotfix mechanism delays or impacts our ability to rollout a fix (TTFix i
Deployment	Fanout	fan out cmd to scale units - for Out of Band (non deploy) fixes, do we have an automated safe-ish framework to c
Deployment	AutomatedDeployment	no human factors in deployment - human interaction in deploymenet process causes or exacerbates impact
Deployment	Rollback	rollback - Lack of ability to rollback delays mitigation as we have to fix forward vs. going back to known good
Deployment	DeploymentHealthChecks	Lack of health checks / ability to pause during deployment causes outage to have larger blast radius than if deplo
Deployment	CloudParity	Parity between national clouds or across clusters
Diagnostics	Analytics	Outage caused or excacerbated by lack of Advanced Analytics and Diagnostics (Instrumentation schema, data de
Diagnostics	VMHealth	Real time VM health diagnostics is missing, delaying diagnosis or mitigation
Diagnostics	RecoveryValidation	Recovery validation - missing diagnostics to validate that all systmes and customres are recovered, either delayin

Complex Systems fail in complex ways

“Each of these small failures is necessary to cause catastrophe but only a combination is sufficient to permit failure”

-Richard I. Cook, “How Complex Systems Fail”

Let’s take a step back

Why do we do RCAs?

To stop bad stuff from happening (again)

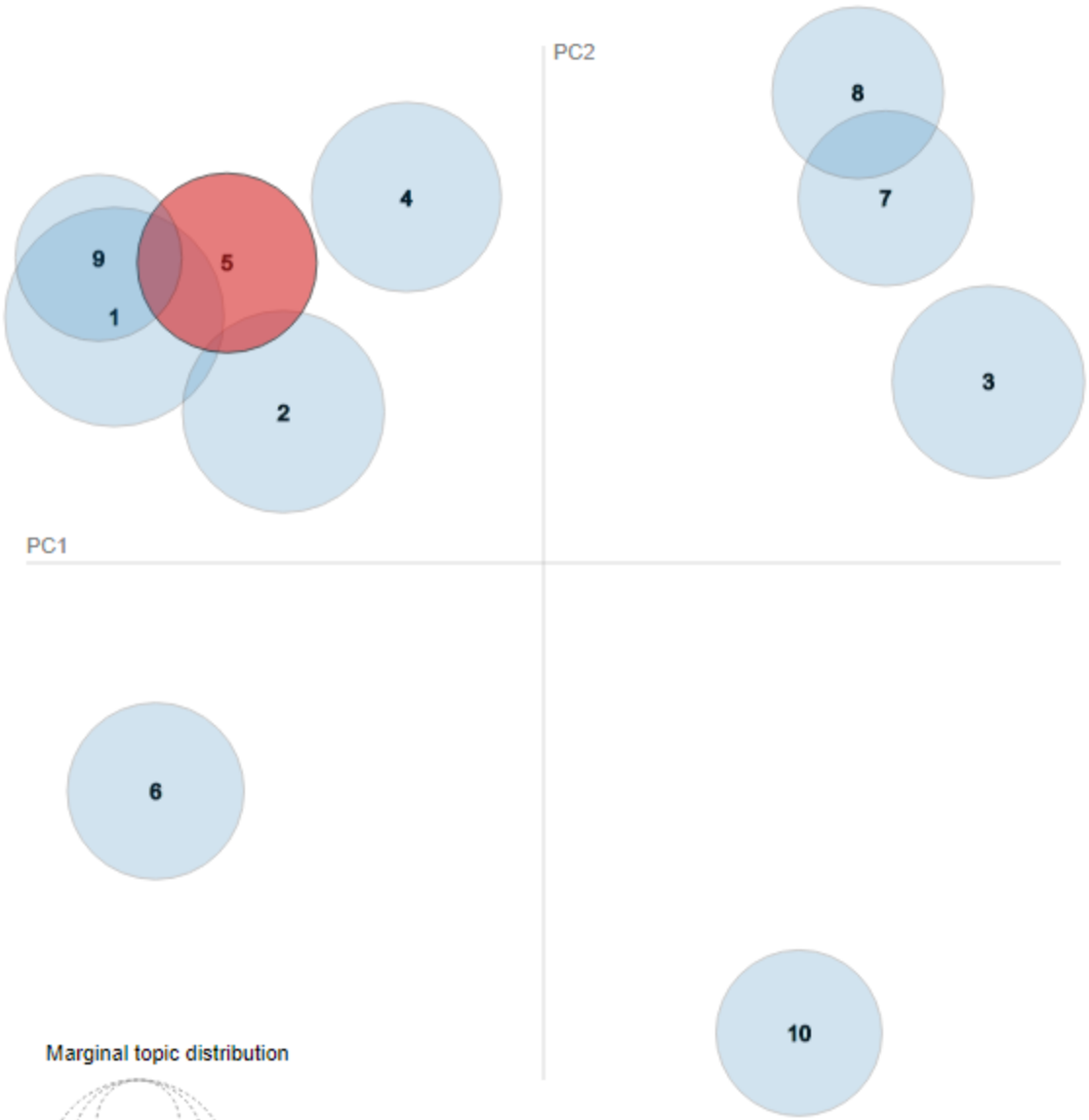
Hunting for ~~Causes~~ ~~Problems~~ Contributing Factors

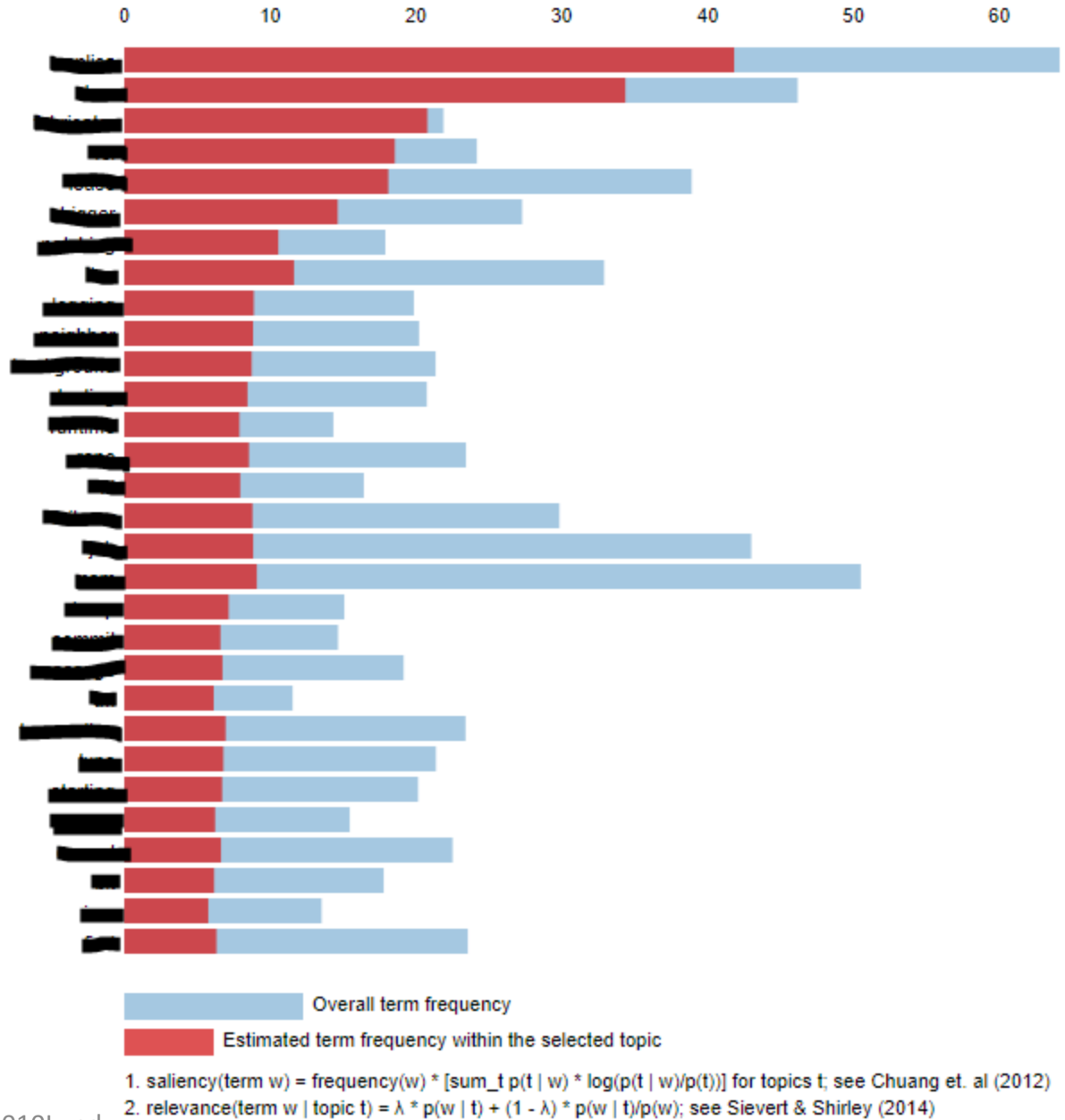
Outage (for our purposes):

Service or platform level issue that impacts customer experience

Postmortem Text Analysis

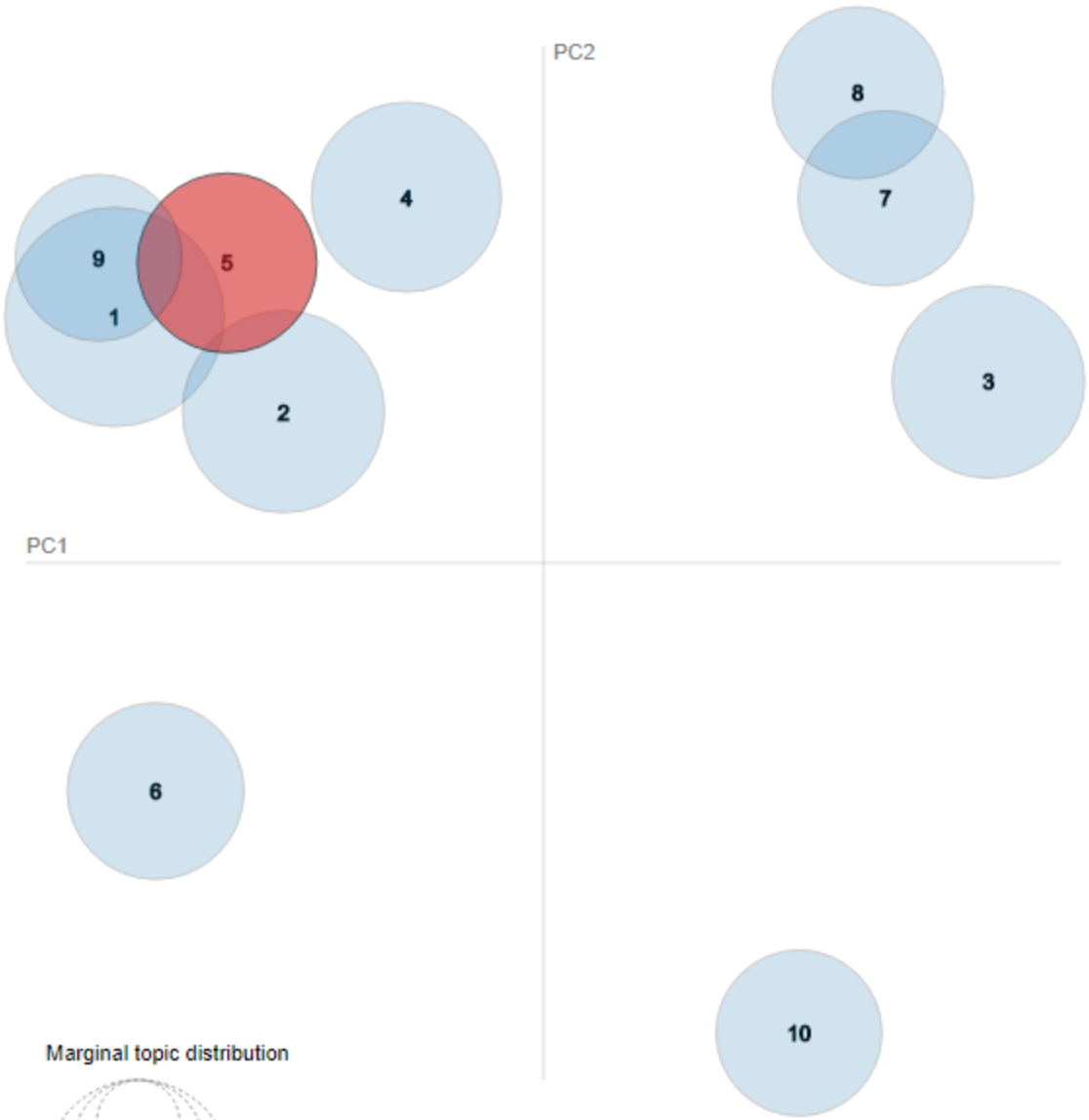
BeautifulSoup
NLTK
Gensim
pyLDAvis

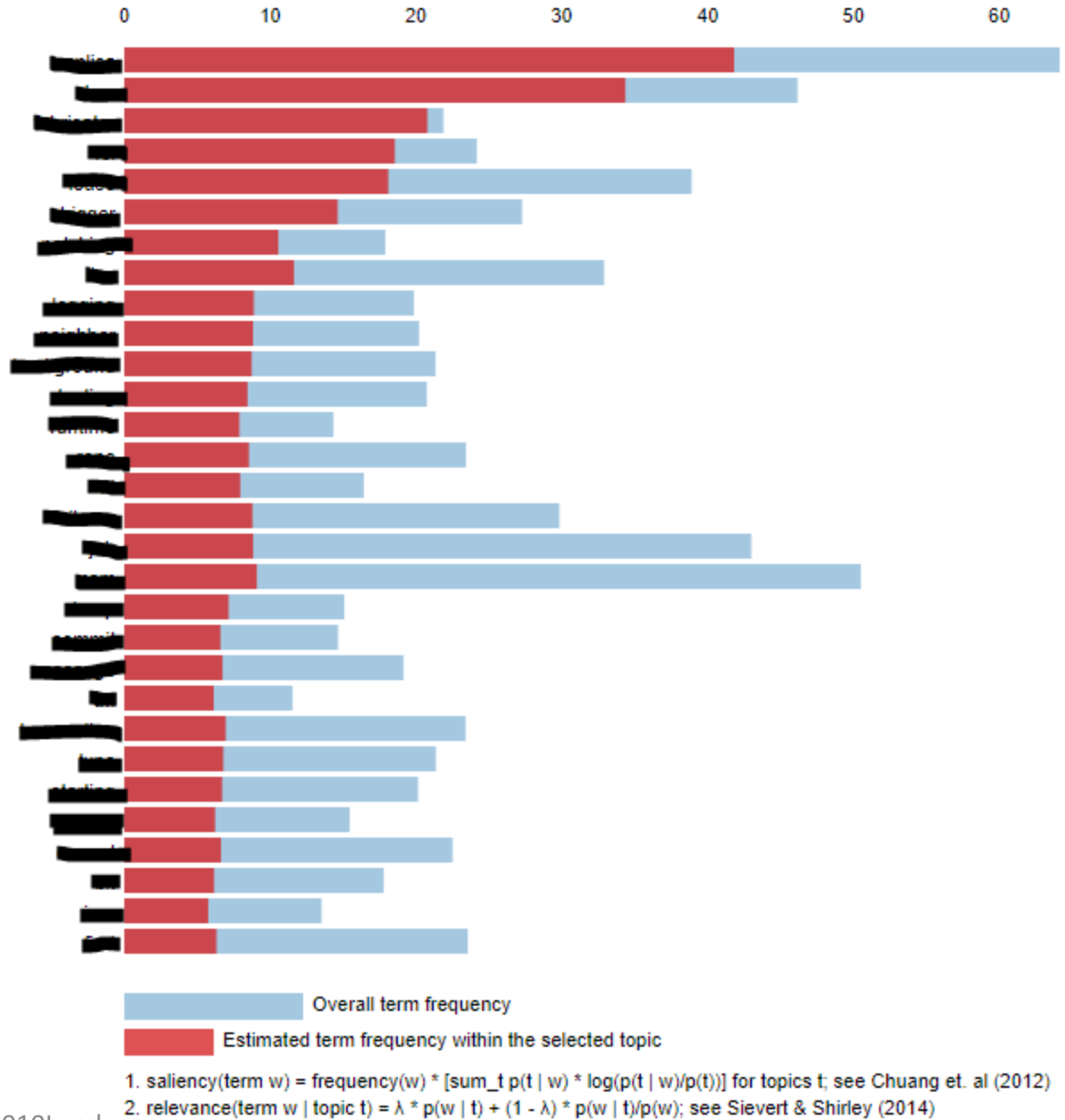

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%


Top-30 Most Relevant Terms for Topic 5 (9.5% of tokens)
0
10
20
30
40
50
60
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

@101010Lund

Not actionable.


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%


Top-30 Most Relevant Terms for Topic 5 (9.5% of tokens)
0
10
20
30
40
50
60
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

@101010Lund

Big Deal™

Metrics!

@101010Lund

Photo: Judy Witts (cc)

Pain Value

Pain value=

(No. of outages) * (duration) * (severity) * (weighting factor)

Customers Impacted
Regions
Hardware SKUs
Distance Below SLO
Number of breached SLOs

Data Scientists:

Pain value=

(No. of outages) * (duration) * (severity) * (weighting factor)

Human interpretation still necessary

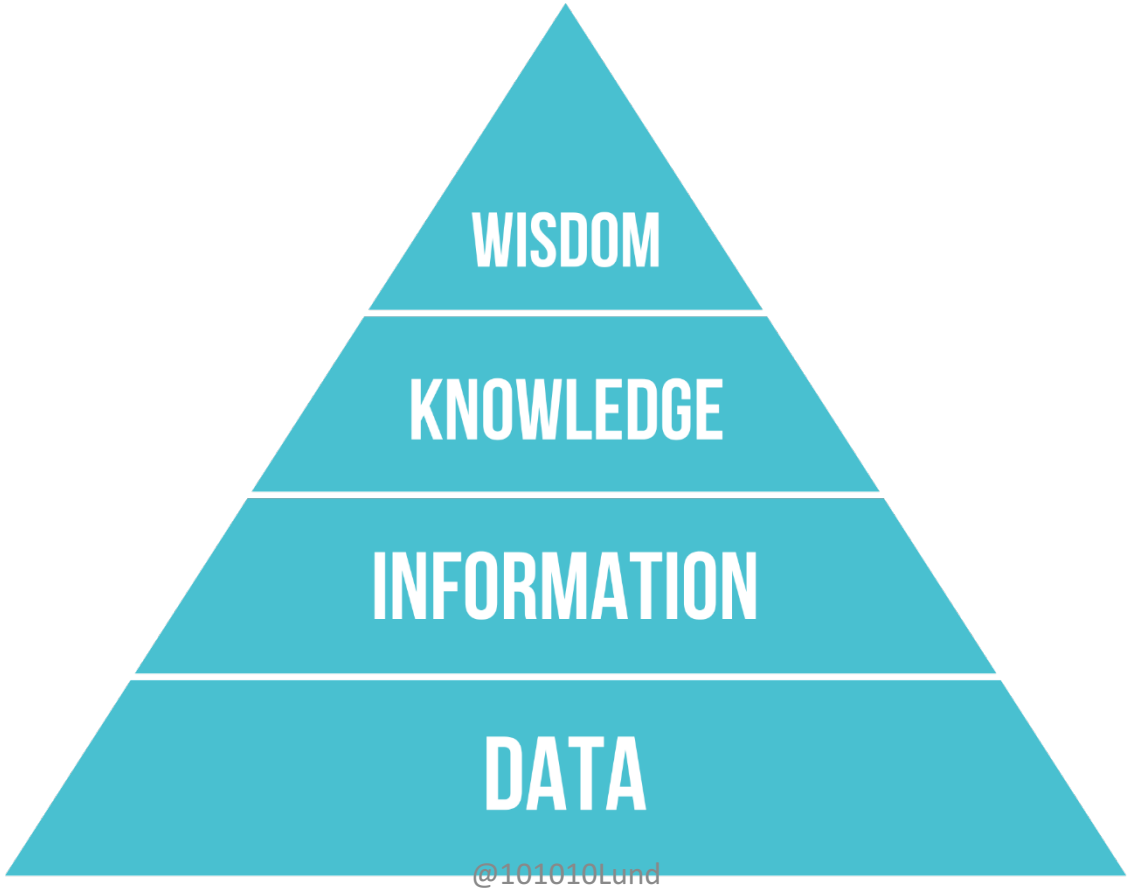


Photo: Wikimedia Commons


@101010Lund

Missing/ Insufficient Data

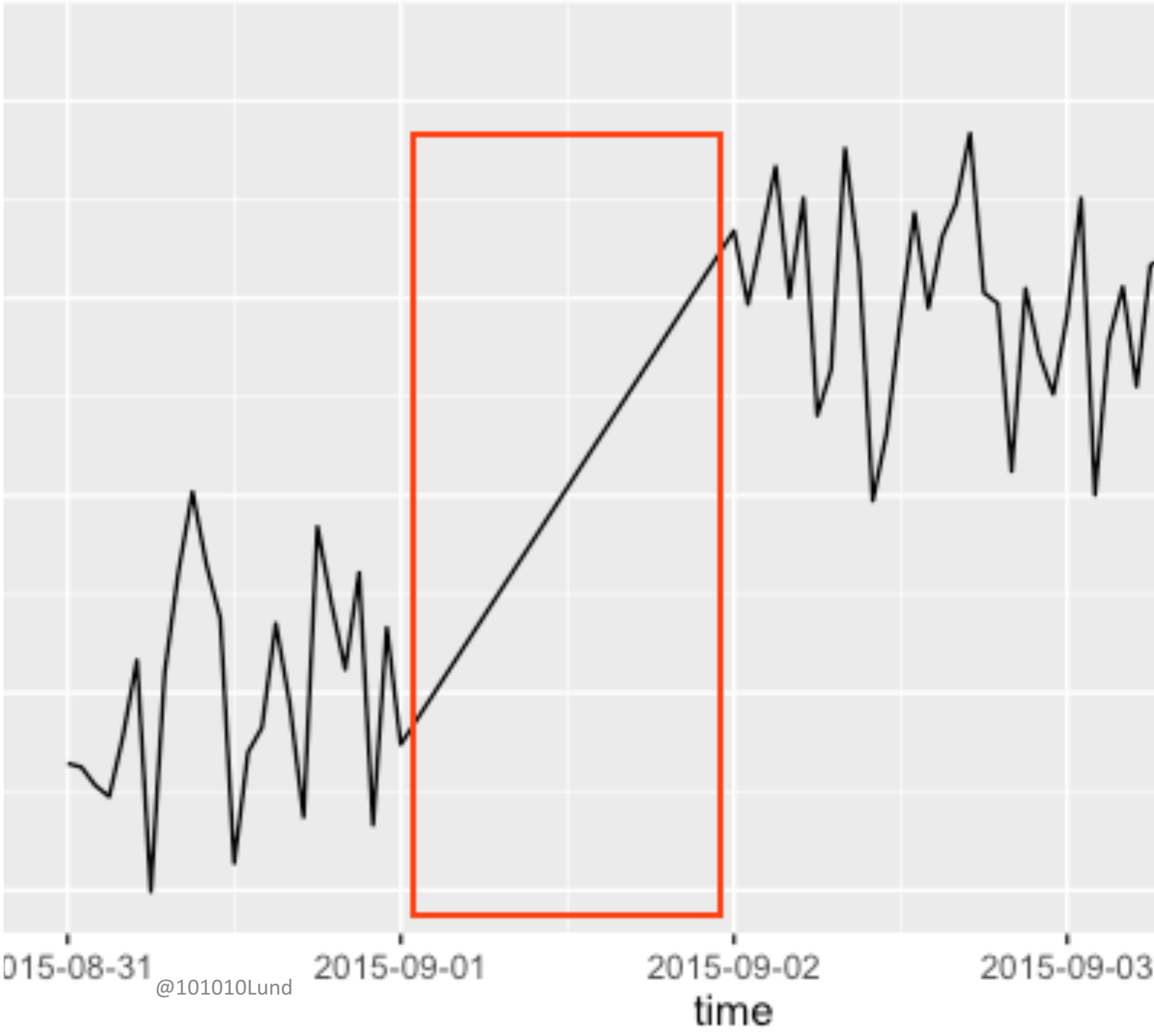
A Framework for a Root Cause Analysis and Action Plan
In Response to a Sentinel Event

e is provided as an aid in organizing the steps in a root cause analysis. Not all possibilities and questions will apply in every case, and there may be others that will emerge in the analysis. However, all possibilities and questions should be fully considered in your quest for "root cause" and risk reduction.

voiding "loose ends," the three columns on the right are provided to be checked off for later reference:
use?" should be answered "yes" of "No" for each finding. A root cause is typically a finding related to a process or system that has a potential for redesign to reduce risk. If a particular hat is relevant to the event is not a root cause, be sure that it is addressed later in the analysis with a "Why?" question. Each finding that is identified as a root cause should be considered tion and addressed in the action plan.
Why?" should be checked off whenever it is reasonable to ask why the particular finding occurred (or didn't occur when it should have) – in other words, to drill down further. Each item in this column should be addressed later in the analysis with a "Why?" question. It is expected that any significant findings that are not identified as root causes themselves have "roots".
tion?" should be checked for any finding that can reasonably be considered for a risk reduction strategy. Each item checked in this column should be addressed later in the action plan.
helpful to write the number of the associated Action Item on page 3 in the "Take Action?" column for each of the findings that requires an action.

Level of Analysis		Questions	Findings	Root Cause?	Ask "Why?"	Take Action
ened?	Sentinel Event	What are the details of the event? (Brief description)				
		When did the event occur? (Date, day of week, time)				
		What area/service was impacted?				
	The process or activity in which the event occurred.	What are the steps in the process, as designed? (A flow diagram may be helpful here)				
the nate cial)		What steps were involved in (contributed to) the event?				
	Human factors	What human factors were relevant to the outcome?				
	Equipment factors	How did the equipment performance affect the outcome?				

Incomplete Data



Inaccurate Data

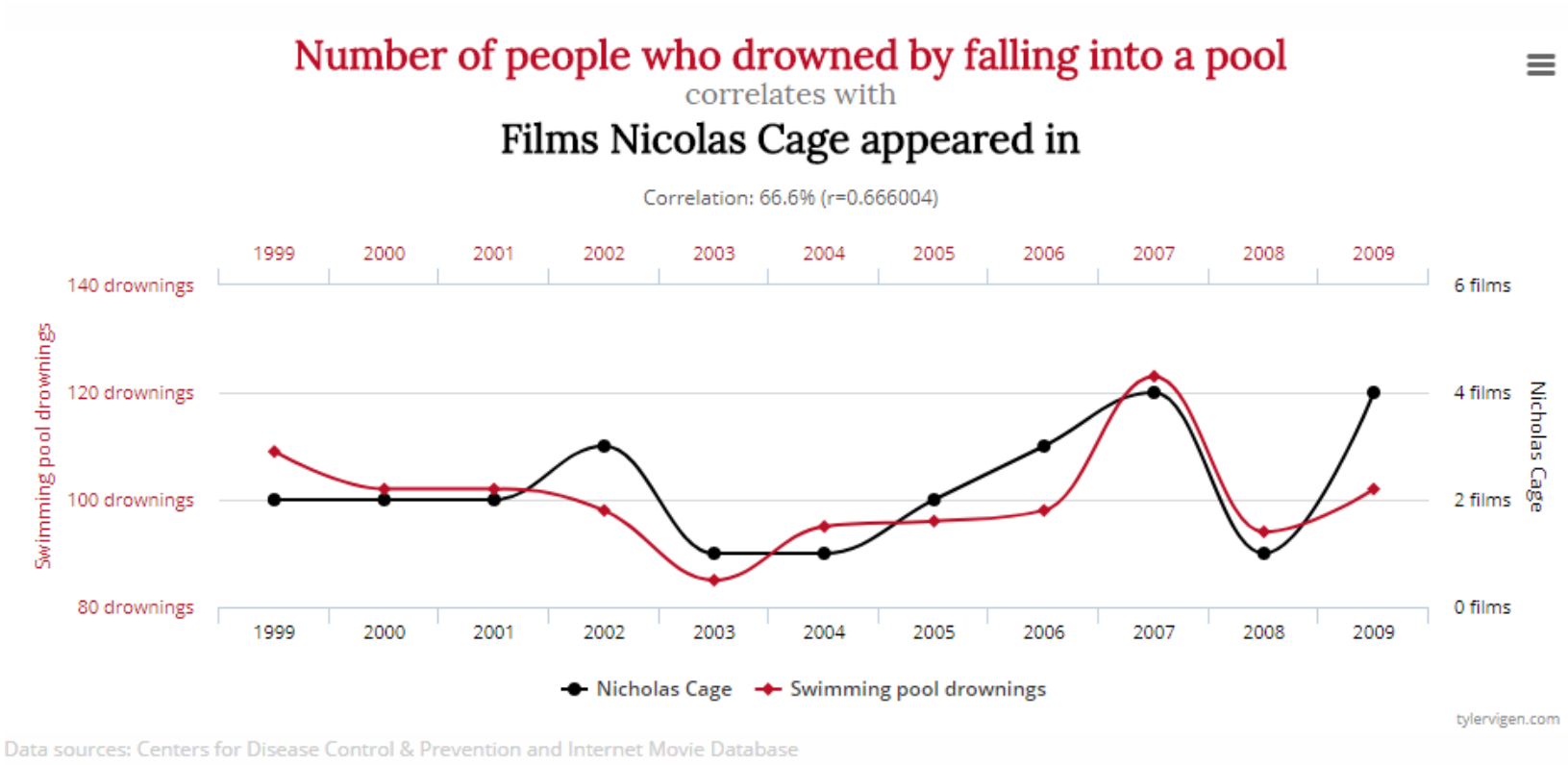
A Framework for a Root Cause Analysis and Action Plan
In Response to a Sentinel Event

e is provided as an aid in organizing the steps in a root cause analysis. Not all possibilities and questions will apply in every case, and there may be others that will emerge in the analysis. However, all possibilities and questions should be fully considered in your quest for "root cause" and risk reduction.

voiding "loose ends," the three columns on the right are provided to be checked off for later reference:
use?" should be answered "yes" of "No" for each finding. A root cause is typically a finding related to a process or system that has a potential for redesign to reduce risk. If a particular hat is relevant to the event is not a root cause, be sure that it is addressed later in the analysis with a "Why?" question. Each finding that is identified as a root cause should be considered tion and addressed in the action plan.
Why?" should be checked off whenever it is reasonable to ask why the particular finding occurred (or didn't occur when it should have) – in other words, to drill down further. Each item in this column should be addressed later in the analysis with a "Why?" question. It is expected that any significant findings that are not identified as root causes themselves have "roots".
tion?" should be checked for any finding that can reasonably be considered for a risk reduction strategy. Each item checked in this column should be addressed later in the action plan.
helpful to write the number of the associated Action Item on page 3 in the "Take Action?" column for each of the findings that requires an action.

Level of Analysis		Questions	Findings	Root Cause?	Ask "Why?"	Take Action
ened?	Sentinel Event	What are the details of the event? (Brief description)	Our Certs Expired			
		When did the event occur? (Date, day of week, time)				
		What area/service was impacted?				
	The process or activity in which the event occurred.	What are the steps in the process, as designed? (A flow diagram may be helpful here)				
the nate cial)		What steps were involved in (contributed to) the event?				
	Human factors	What human factors were relevant to the outcome?	It Was Definitely Network's Fault			
	Equipment factors	How did the equipment performance affect the outcome?				

Irrelevant Data

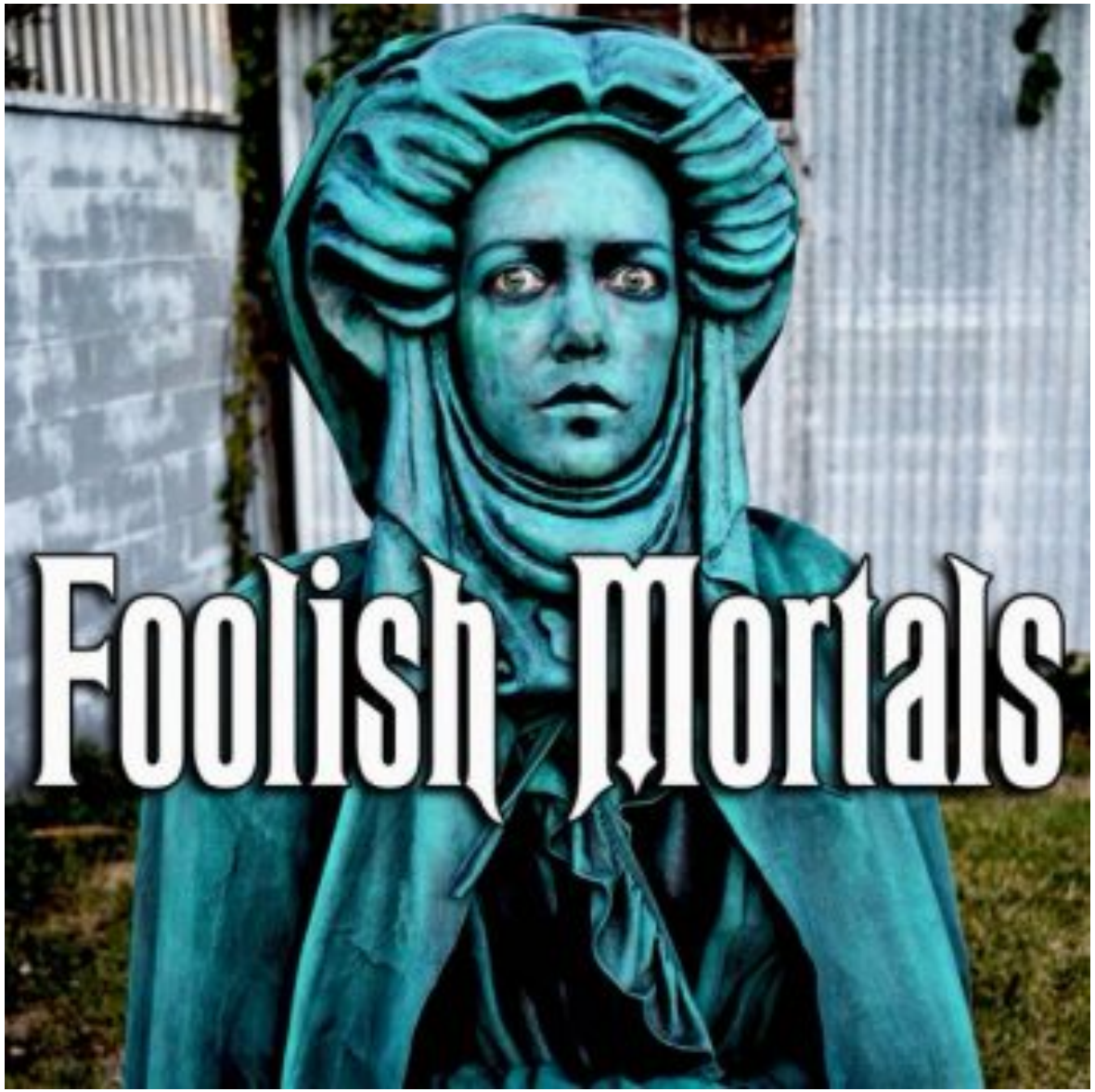


Ambiguity

Node – CPU
Node – Instance of Program
Node – Physical Hardware Box
Node – Point on Graph such that G = (V,E)
Node – Any device connected to the network
Node – Communication endpoint
Node – Client, Server, or Peer
Node – Bitcoin miner
Node – Data Type
Node – Node.js

Confounding Factors

(like config drift)


Foolish Mortals

Dirty data will lie to you.

What was the (preliminary) result?

1. Surfaced surprise issues

2. Debunked production myths

3. Stronger arguments for prioritization of reliability work

What did we *learn?*

1. Define your hypotheses

2. Clean your data

3. Work your way up the DIKW pyramid

What else can we do?

Cross-Correlate Data Sets

Timeline

Impact Start * [redacted] PDT

Detection [redacted] PDT

Eng. Engaged [redacted] PDT

Mitigation * [redacted] PDT

Impact

Customer Impact

[redacted]

QoS/SLA Impact

[redacted]

Service Responsible

[redacted]

Services Impacted

[redacted]

Root Cause

Root Cause Title

[redacted]

Root Cause Details

[redacted]

Repeat Outage

Detection and Mitigation

Detection Source

[redacted]

Detection Details

[redacted]

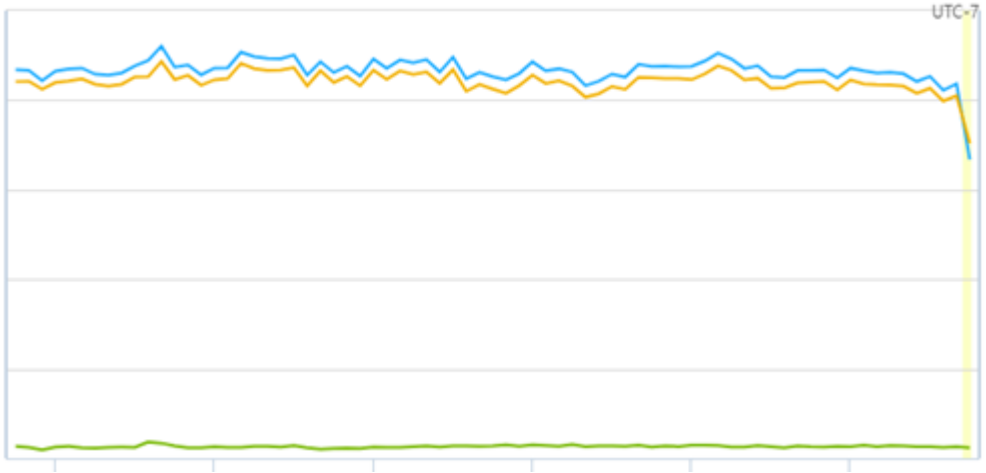
Mitigation Steps

[redacted]

Fix

[redacted]

+

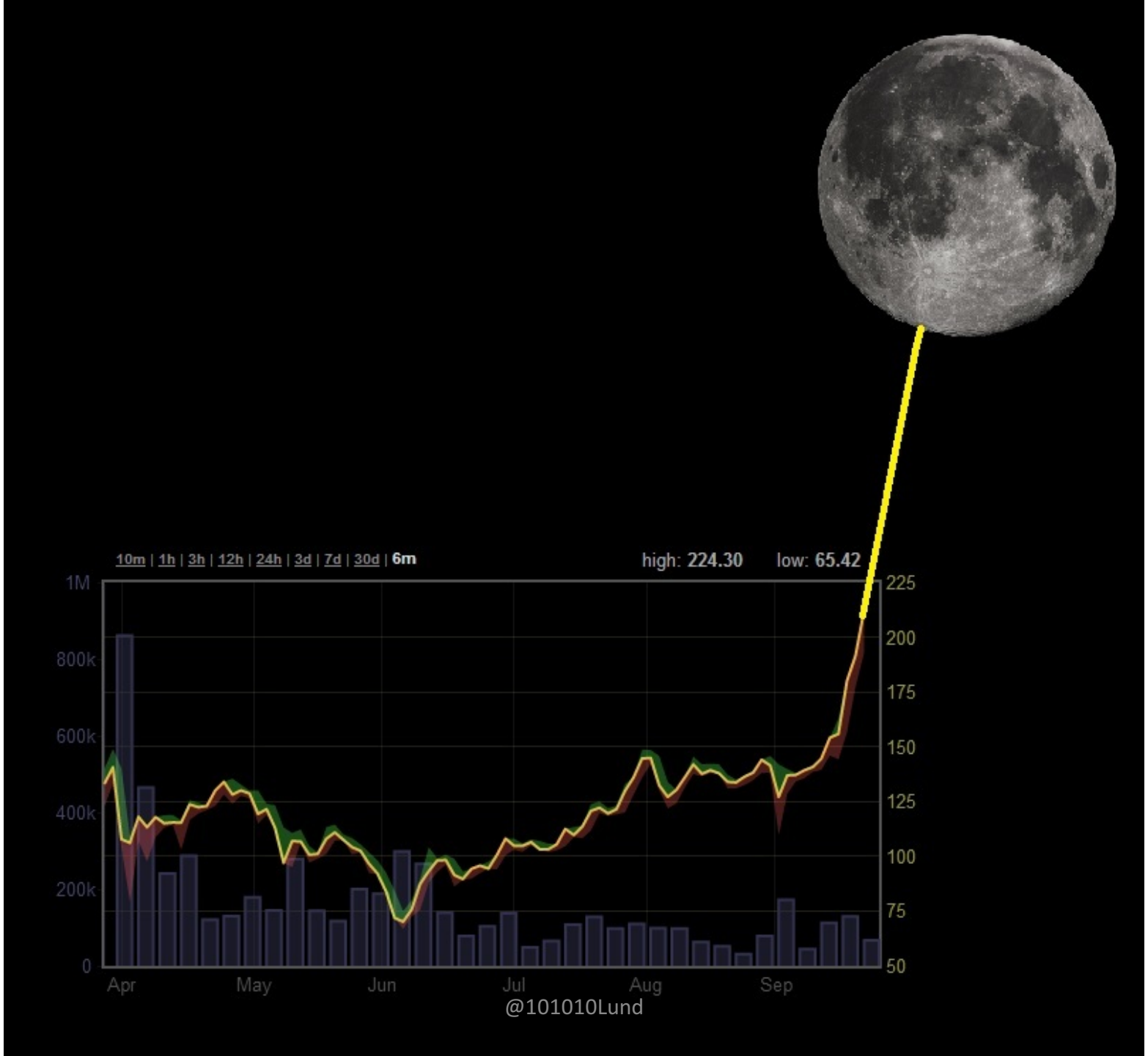


Study your minor failures

Intelligently Calculate Risk

Continue to improve the RCA Process

Photo: Rachel Chapman (CC)


10m | 1h | 3h | 12h | 24h | 3d | 7d | 30d | 6m
high: 224.30
low: 65.42
1M
800k
600k
400k
200k
0
225
200
175
150
125
100
75
50
Apr
May
Jun
Jul
Aug
Sep
@101010Lund


Tanner Lund
@101010Lund
talund@Microsoft.com
/in/tannerlund
@101010Lund