Move Fast & Learn Things:

Principles of cognition, coordination, and teaming to support high performance and resilient site reliability engineering.

Dr. Laura Maguire & Nora Jones

Jeli.io

SRECon APAC
Dec 7, 2022

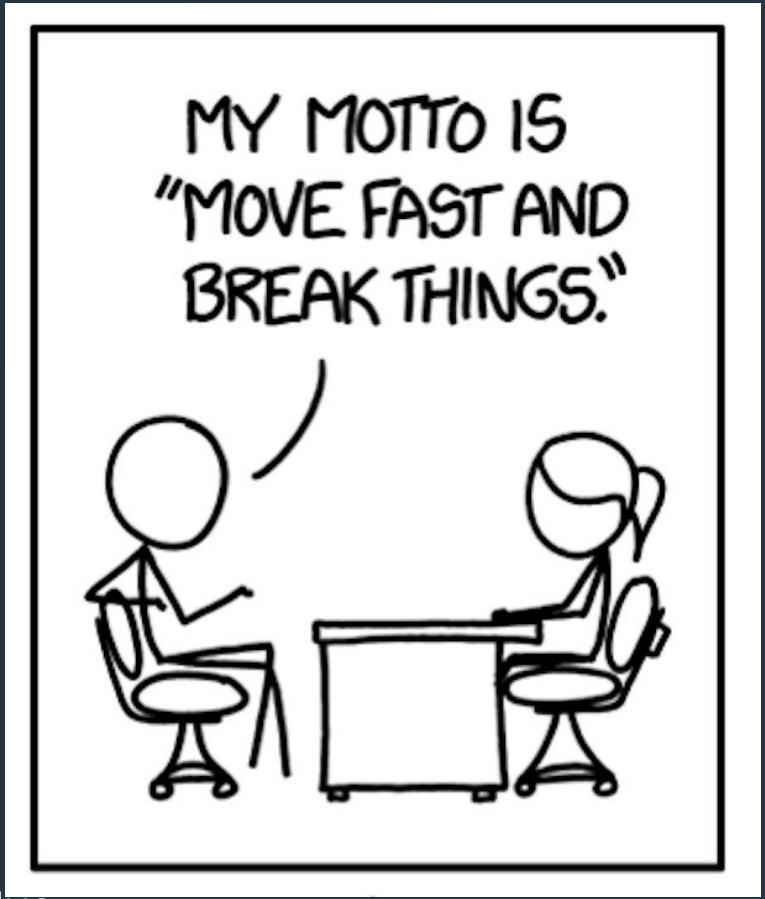
jeli

Part 1. Principles of cognition, coordination and teaming.

Part 2. A tale of two incidents.

Part 3. Collaborative learning.

jeli



JOBS I'VE BEEN FIRED FROM

FEDEX DRIVER
CRANE OPERATOR
SURGEON
AIR TRAFFIC CONTROLLER
PHARMACIST
MUSEUM CURATOR
WAITER
DOG WALKER
OIL TANKER CAPTAIN
VIOLINIST
MARS ROVER DRIVER
MASSAGE THERAPIST

jeli

https://xkcd.com/1428/

Part 1 - Principles of cognition, coordination, and teaming.

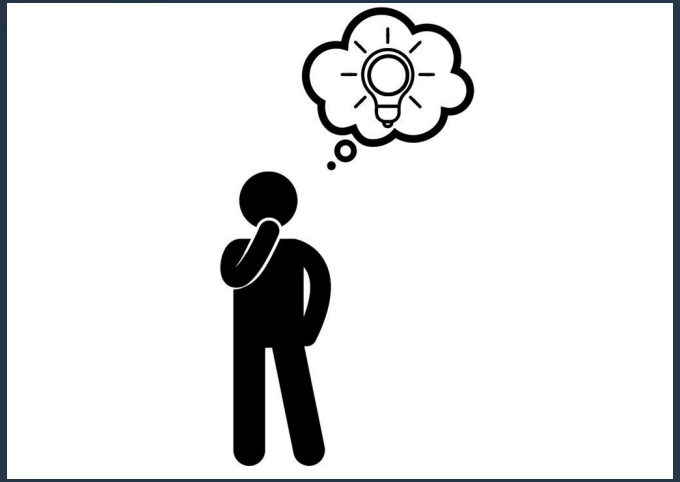
Principles of Cognition, Coordination & Teaming

1. Cognitive work is **more than** applying technical knowledge, and **is often hidden.**

jeli

Principles of Cognition, Coordination & Teaming

1. Cognitive work is more than applying technical knowledge, and is often hidden.

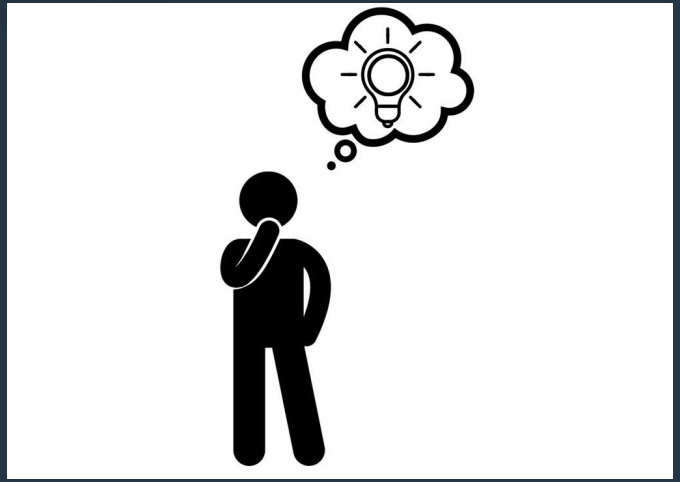


Cognitive Work is the ways in which people interact with information in their jobs.

jeli

Principles of Cognition, Coordination & Teaming

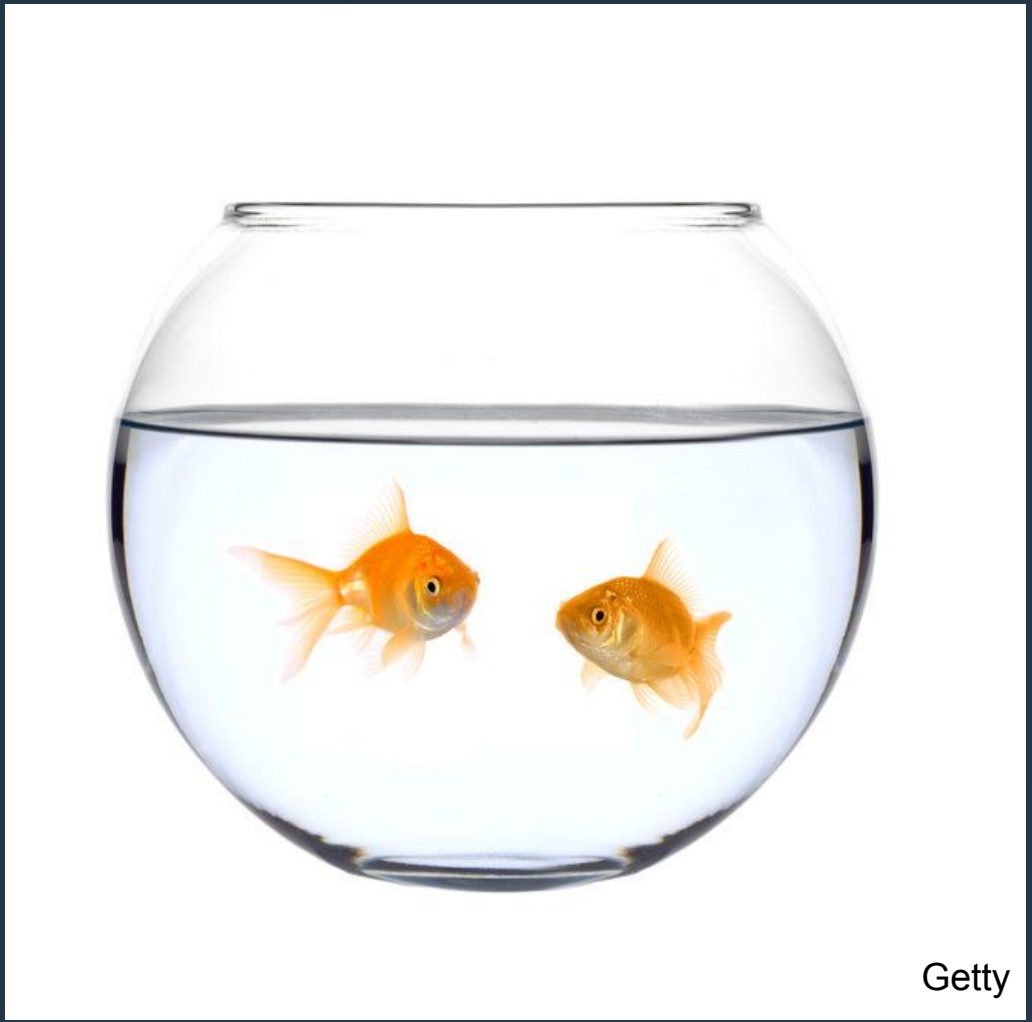
1. Cognitive work is **more than** applying technical knowledge, and **is often hidden.**



Cognitive Work is the ways in which people interact with information in their jobs.

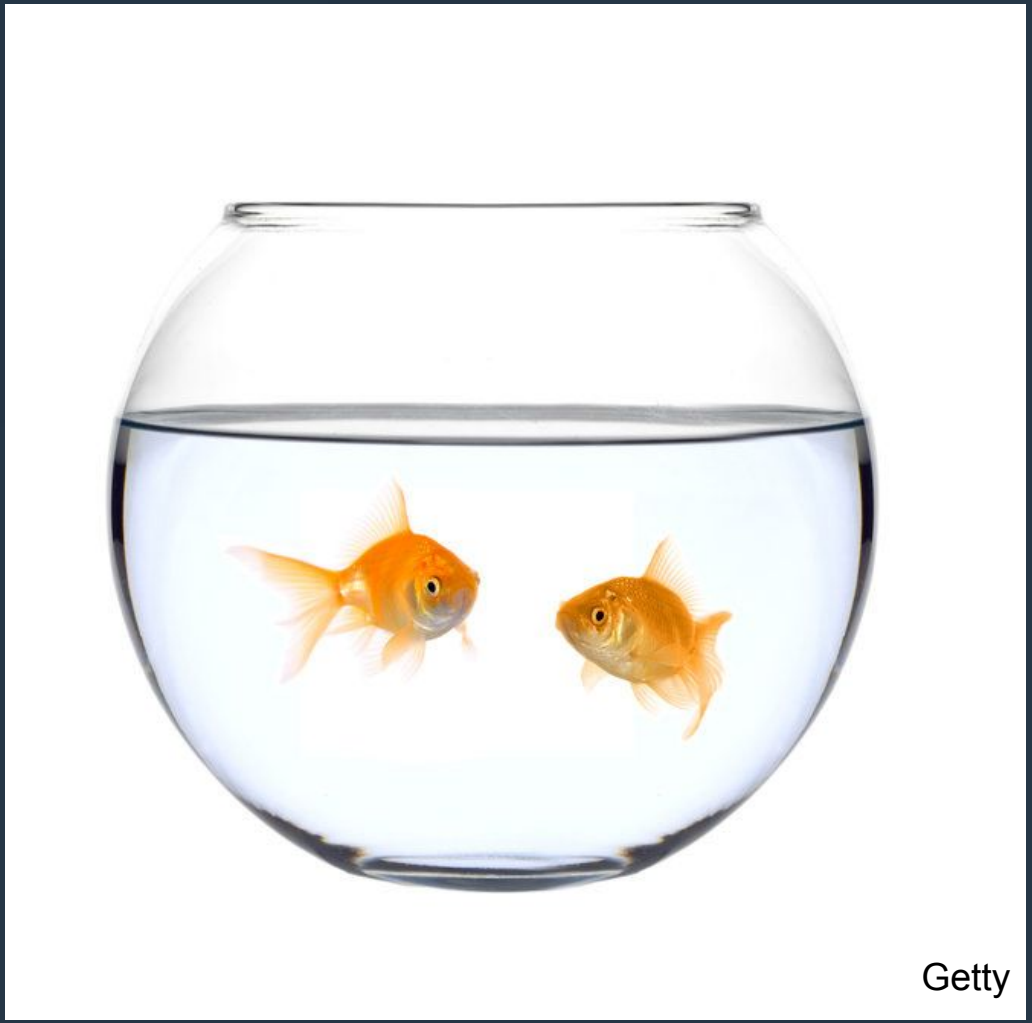
Pssst: It's the "thinking" part of your work!

jeli



Getty

jeli



Getty

“What the heck is water?”

jeli

Wood's 'law of fluency' (paraphrased by Richard Cook):

"expertise hides the effort in work"



jeli

Wood's 'law of fluency' (paraphrased by Richard Cook):

"expertise hides the effort in work"

Paraphrased by Maguire:

"When you are good, you make it look easy."



jeli



Principles of Cognition, Coordination & Teaming

1. Cognitive work is more than applying technical knowledge, and is often hidden.
2. Site Reliability Engineering requires fine tuned skills in **detection, diagnosis work, and flexibility** in applying deep technical expertise.

jeli

Principles of Cognition, Coordination & Teaming

1. Cognitive work is more than applying technical knowledge, and is often hidden.
2. Site Reliability Engineering requires fine tuned skills in **detection, diagnosis work, and flexibility** in applying deep technical expertise.



jeli

Principles of Cognition, Coordination & Teaming

1. Cognitive work is more than applying technical knowledge, and is often hidden.
2. Site Reliability Engineering requires fine tuned skills in **detection, diagnosis work, and flexibility** in applying deep technical expertise.





jeli

Principles of Cognition, Coordination & Teaming

1. Cognitive work is more than applying technical knowledge, and is often hidden.
2. Site Reliability Engineering requires fine tuned skills in **detection, diagnosis work, and flexibility** in applying deep technical expertise.







What fine tuned skills looks like in practice.

Eng 2: It may have something to do with the deploy. Maybe the nodes did not come up soon enough before the old ones were shutdown. We may want to investigate in a readiness probe if that is the case
Eng 3: was there a deploy in the last 10 minutes?
Eng 1: Did you just do a deploy?
Eng 2: <Link to a recent pull request> whenever a PR is merged a deploy is kicked off
Eng 1: Sure, but why did it update the k8s resources when the k8s files weren't changed?
Eng 2: It deploy everything. There is no logic to make sure to only deploy when k8s items are changed
Eng 1: I thought (for some reason) that `kubectl apply` was a noop when the yaml was the same
Eng 2: it is but we rebuild the image each time, that will cause a new tag -> new deploy
Eng 4: The master containers have been running for 31 hours which means no new containers were made.
Eng 1: Good point :this: The workers are also 20 hours old
Eng 3: So what are the youngest items in the cluster atm? ingress?
Eng 1: Maybe ingresses or [messaging broker] ?
Eng 2: The build is still building: <link to job>
Eng 1: Good thing [continuous integration software] is up so that we can watch it 😜
Eng 5: Was it corroborated by users - ie could it have been local to [area] or the US network?
Eng 3: other users saw the outage too
Eng 1: Users reported it. Could have been users in Chicago though
Eng 3: One was New York
One was Japan though
That's not even close 😄
Eng 3: Jobs weren't interrupted, just access to the UI
Eng 5: Bear in mind that [network provider] has periodic glitches of network traffic and will report them at least 6 hours after an event occurred. It could potentially be caused by them.
Eng 3: Certainly a possibility
Eng 2: What I think is happening there is a [network provider] network issue. This can also been seen with the master build freezing during the apt-get update

Ex 1:

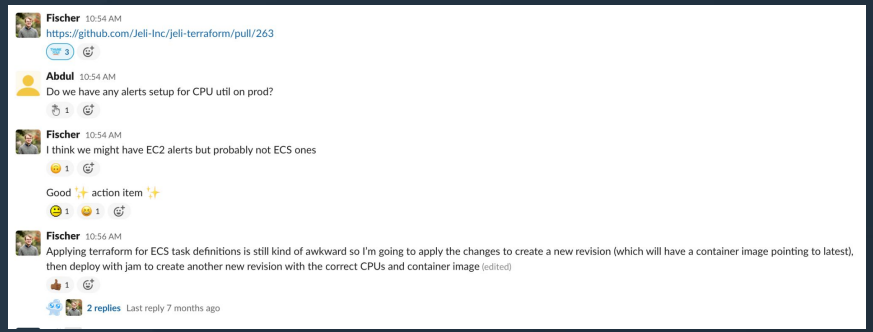
Eng 2 (ic): Um, so let's check for the 500 errors in the, in log entries. Does anyone already have log entries open?
Eng 1: I do. Yeah. So you want me to look for 500 errors from<*CRE*>?
Eng 2 (ic): Thank you. Yeah.
Eng 1: Uh, ok, right. On it.

Ex 2:

Eng 2(ic): [Eng 4] and [Eng 5] are checking into narrowing the scope. [Eng 3] and [Eng 1] do you have something that you're already digging into?
Eng 1: Yeah.
Eng 2(ic): Which gap?
Eng 1: I'm digging into the permission thing.

Ex 3:

Eng 2(IC): Yeah. If someone is free... has free hands, can they start it's import bundle please.



Cognitive work

Anticipating

Observing

Inferring

Reasoning

Recognizing change

Prioritizing

Planning

Troubleshooting

Diagnosing

Correcting

Modifying

Reacting



jeli.io

jeli

Principles of Cognition, Coordination & Teaming

1. Cognitive work is more than applying technical knowledge, and is often hidden.
2. Site Reliability engineering requires fine tuned skills in **detection, diagnosis work, and flexibility** in applying deep technical expertise.
3. Working in teams **is necessary and adds cognitive work**.

jeli

Cognitive work

Anticipating

Observing

Inferring

Reasoning

Recognizing change

Prioritizing

Planning

Troubleshooting

Diagnosing

Correcting

Modifying

Reacting



Coordinative work

Recruiting

Synchronizing

Grounding

Updating

Signaling

Taking Initiative

Delegating

Taking Direction

jeli



Woods' Theorem:

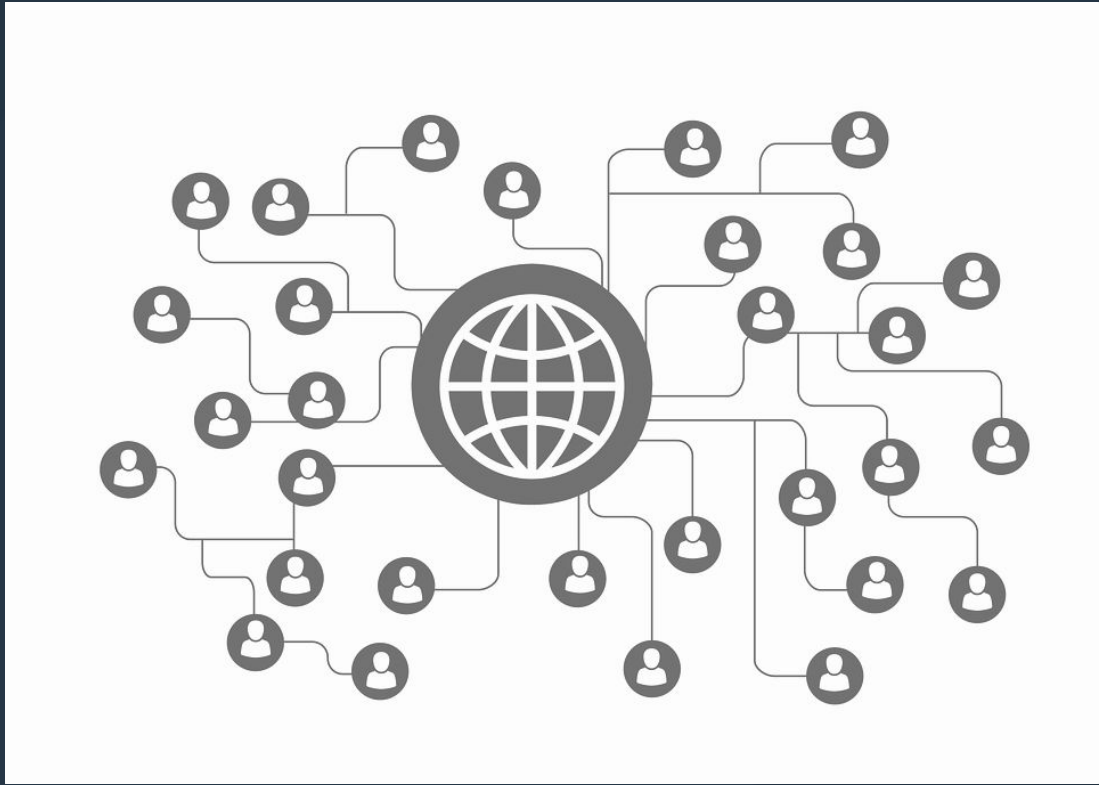
“As the complexity of a system increases, the accuracy of any single agent's own model of that system decreases rapidly.”

-SNAFU Catchers Cycle 1 report

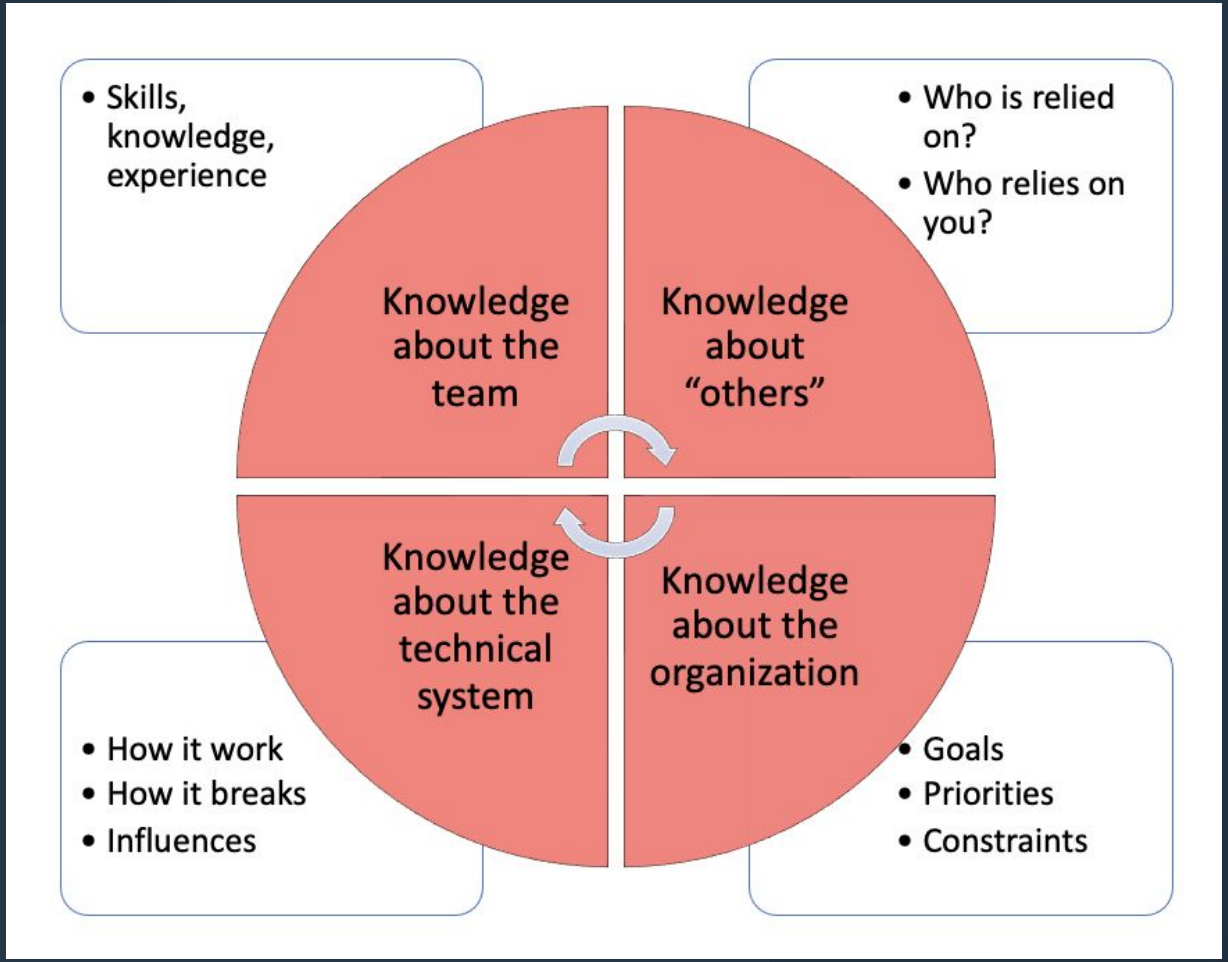
(stella.report)

jeli

Multiple, diverse perspectives



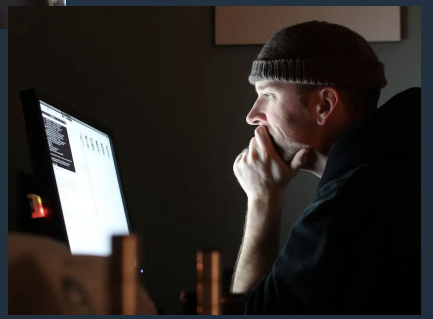
jeli


• Skills, knowledge, experience
• Who is relied on?
• Who relies on you?
Knowledge about the team
Knowledge about "others"
Knowledge about the technical system
Knowledge about the organization
• How it work
• How it breaks
• Influences
• Goals
• Priorities
• Constraints

Parties engaged in joint activity have different **goals** and **priorities**.









jeli

There can be
So.
Much.
Awkward.



jeli

Principles of Cognition, Coordination & Teaming

1. Cognitive work is more than applying technical knowledge, and is often hidden.
2. Site Reliability engineering requires fine tuned skills in **detection, diagnosis work, and flexibility** in applying deep technical expertise.
3. Working in teams is necessary **and** adds cognitive work.
4. SRE requires fine tuned skills in **anticipating, adjusting, signalling, grounding, delegating, and taking initiative.**

jeli

Mgr 1: While we're waiting for that is somebody's generating a support bundle in the ticket? I think they're waiting for that. [crosstalk]
Eng 2: Yes, sir. [crosstalk]
Eng 3: Yeah. It's still running. It takes a while for run. [crosstalk]
Mgr 1: Okay, thank you. I didn't see a task for that. [crosstalk]
Eng 4: You have to do a full support bundle?
Eng 3: A full support bundle?
Eng 4: Well before one takes like a couple of hours.
Eng 3: Yeah. I did a full up, a full cluster support bundle
Eng 2: Doesn't that take a couple of hours? Or did they improve the speed?
Eng 5: I recall it taking a couple hours.
Eng 3: Yeah. Takes a while
Eng 2: While they are waiting for the support bundle we may want to let them know, say, Hey, this kind of take a couple hours. We want something on do something between that time.
Mgr 1: I'll let them know that

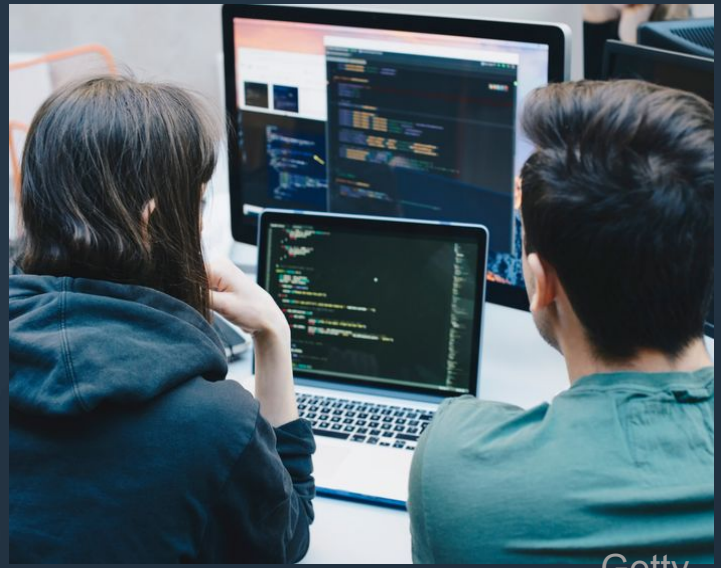
Principles of Cognition, Coordination & Teaming

1. Cognitive work is more than applying technical knowledge, and is often hidden.
2. Site Reliability engineering requires fine tuned skills in **detection, diagnosis work, and flexibility** in applying deep technical expertise.
3. Working in teams is necessary **and** adds cognitive work.
4. SRE requires fine tuned skills in **anticipating, adjusting, signalling, grounding, delegating, and taking initiative.**
5. Resilient organizations are networks that **share capacities** and adapt according to the demands.

jeli

Teaming: Coordination, shared capacities, and resilience.

- Sharing available resources to help others meet goals
- Relaxing goals that may conflict with others
- Removing constraints that can impede work.
- Reciprocating when opportunities arise.



Getty

jeli

You can’t meaningfully improve the reliability of your systems without looking at the cognitive work.

jeli

You can't meaningfully improve the reliability of your systems without looking at the cognitive work...

Which is, itself, *work*.

jeli

Part 2 - A tale of two incidents.

Investigation #1

A seemingly innocuous incident, that didn't have much customer impact, and didn't "deserve" or have time for a thorough postmortem.

Investigation #1

A seemingly innocuous incident, that didn't have much customer impact, and didn't "deserve" a thorough postmortem

- Templated approach

Investigation #1

A seemingly innocuous incident, that didn't have much customer impact, and didn't "deserve" a thorough postmortem

- Templated approach
- Completed by members of the team most involved in the incident

Investigation #1

A seemingly innocuous incident, that didn't have much customer impact, and didn't "deserve" a thorough postmortem

- Templated approach
- Completed by members of the team most involved in the incident
- Purpose was to file and report

Post Mortem Template

Summary

Impact

Detection

Resolution

Detailed Summary

Contributing Factors

Timeline

What went well

What went wrong

How we got lucky

Action Items

jeli

Post Mortem Template

Summary

Impact

Detection

Resolution

Detailed Summary

Contributing Factors

Timeline

What went well

What went wrong

How we got lucky

Action Items

What don't we know:

- How much time was spent on this document
- Who read it
- Who attended the meeting
- What the purpose of it was
- What was difficult or easy about handling the incident?

jeli

Investigation #1: Summary

A change was made to some of the search infrastructure tooling that updates which search index to look at each day.

We were able to recreate the issue and soon after it had been reported, engineering was able to track down the problem and fix the bad configuration.

Due to the nature of the bug, no pre-processing work had to be done for search to be functional again.

jeli

Investigation #1: Detailed Summary

The search index is split into collections, one large collection for older [SKUs] and 2-3 collections for recent [SKUs], one per day, for newly created [SKUs].

Each collection can be individually submitted for reading or writing. Since search queries typically read from all collections from a given vendor.

We employ a feature called collection read alias which lets us query from many collections using one name. As new daily collections are created (typically done 2 days in advance), they are added to the read alias.

Due to a bug in our tooling introduced a week before, newly created daily collections were not being added to the read alias. **Since collections are created and added to the alias 2 days in advance, the first collection to be affected by the bug was 3 days ago.**

jeli

Incident #1 Timeline

December 4, 2022	Bug is introduced in python tooling
December 4, 2022	The 6 December collection is created and not added to the alias
December 5, 2022	Newly created SKUs start going into the 6 December alias
December 6, 2022, 2:30 PM	Incident channel starts based on customer reports Search team is paged and begins work
2:31 PM	Issue with elevated error rates is red herring
2:31 PM	Borked broker potential cause
2:33 PM	Borked broker ruled out
2:34 PM	Misconfiguration in read alias identified as the cause
2:36 PM	Configuration is fixed
2:36 PM	Fix is confirmed
2:37 PM	Fix is applied to index file

Incident #1: Impact

All [e-commerce] users had a degraded search experience for the 20 minutes over which the incident took place.

jeli

Incident #1: Contributing Factors

- Lack of type system or static analysis in code
- Alert that would have detected this is also broken

Incident #1: Contributing Factors

- Lack of type system or static analysis in code
- Alert that would have detected this is also broken

Incident #1: What went well

- Fairly quick fix one the search team staged

Incident #1: What went well

- Fairly quick fix one the search team staged

Incident #1: What went wrong

- Alert that should have caught the misconfiguration was also broken

Incident #1: What went wrong

- Alert that should have caught the misconfiguration was also broken

Incident #1: How we got lucky

- Happened during a low traffic time, and not many users were impacted

Incident #1: How we got lucky

- Happened during a low traffic time, and not many users were impacted

jeli

Incident #1: Action Items

- Fix the bug preventing the read alias from being updated
- Fix the alert that should have fired
- Check all our alerts to make sure that they fire

Incident #1: Action Items

- Fix the bug preventing the read alias from being updated
- Fix the alert that should have fired
- Check all our alerts to make sure that they fire

jeli

What questions do you have about incident #1?

Investigation #2

It’s always a little more complex than it appears

- Conducted by an engineer outside of the main team, fresh perspective

Investigation #2

It’s always a little more complex than it appears

- Conducted by an engineer outside of the main team, fresh perspective
- Included evidence from previous deploys (some well before the event)

Investigation #2

It’s always a little more complex than it appears

- Conducted by an engineer outside of the main team, fresh perspective
- Included evidence from previous deploys (some well before the event)
- Included teams impacted, but not previously involved in the timeline

Investigation #2

It’s always a little more complex than it appears

- Conducted by an engineer outside of the main team, fresh perspective
- Included evidence from previous deploys (some well before the event)
- Included teams impacted, but not previously involved in the timeline
- Informed new team members about system dynamics

Investigation #2

It’s always a little more complex than it appears

- Conducted by an engineer outside of the main team, fresh perspective
- Included evidence from previous deploys (some well before the event)
- Included teams impacted, but not previously involved in the timeline
- Informed new team members about system dynamics
- Purpose was to engage the audience and to be read

How We Got Here “HOWIE” Guide

- Background of the document
- Who responded
- Who investigated
- Executive Summary
- Customer+Employee Impact
- Key Takeaways
- Triggering Event
- Contributors/Enablers
- Mitigators
- Difficulties during handling
- Follow-up items

Incident #2: Timeline

Search | Data from #key-updates

Filtered By | Select Tag(s) / Participant(s) / Type(s)

Display Notes: | Build a timeline 7



Justin Bacon Nov 23 at 1:28 AM

We're receiving requests right now related to the key change beta release

Trigger - Actual

Natalie Demo Nov 23 at 7:54 PM

Between November 23 and November 30th, the search team received several pager-driven alerts on search 'not working'. Invalid and "revoked" keys created a source of confusion and fatigue for the search team.

Natalie Demo Dec 3 at 3:51 AM

PR 22: Introduced to avoid the noisy alerts the search team was getting as a result of the key changes. It was to "not alert when the key is invalid"

Natalie Demo Dec 4 at 7:50 PM

Partition is created to hold indices for December 6th, 2022. These are always *automatically* created 2 days in advance.

Natalie Demo Dec 4 at 7:50 PM

PR 23: Introduced to not update read aliases for live collections during a migration. This was a result of a migration taking a minute longer than necessary (some "clean up" work added to an unrelated PR)

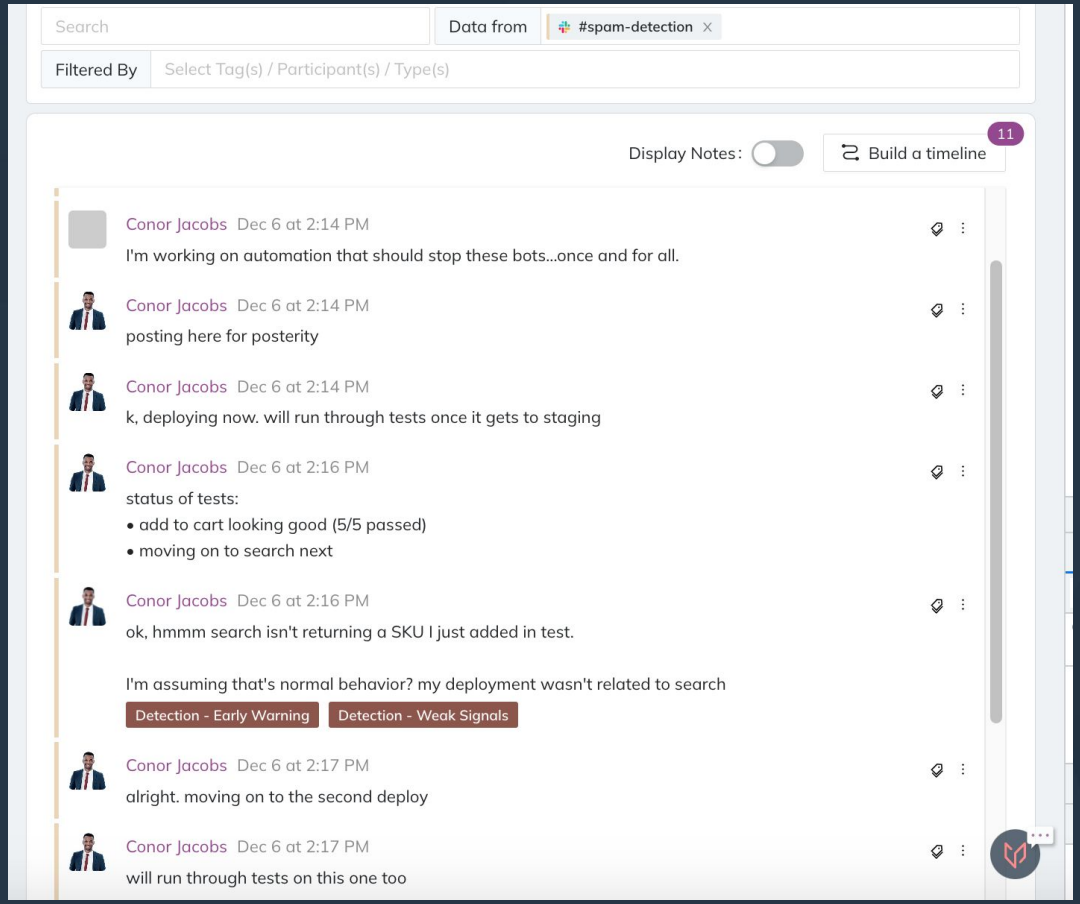
Natalie Demo Dec 5 at 3:50 PM

Between December 4th and December 6th, errors were detected, but there were no pages due to **PR 22**

Natalie Demo Dec 6 at 1:49 AM

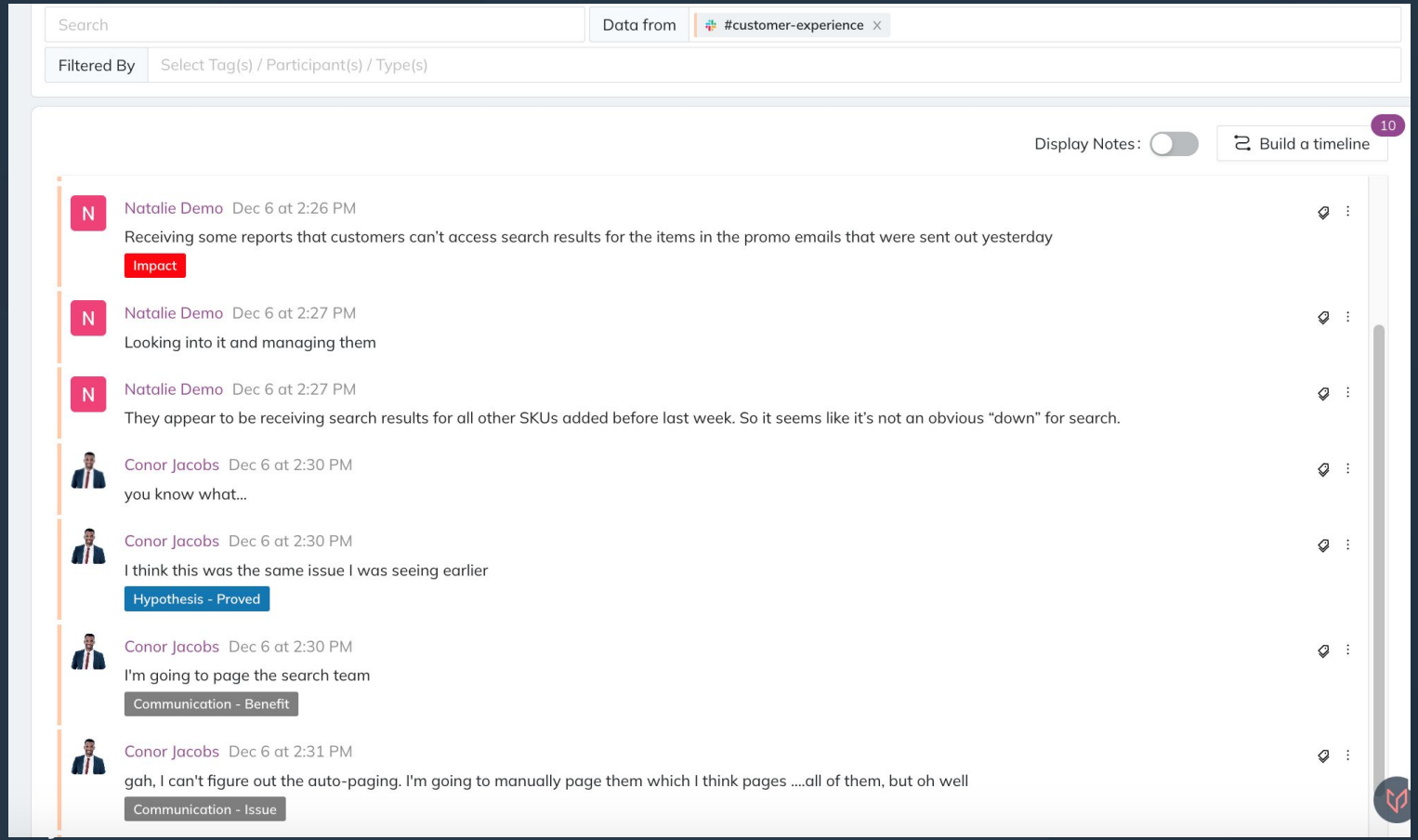
This is where user search queries begin not reaching the December 6th partition.

Incident #2: Timeline

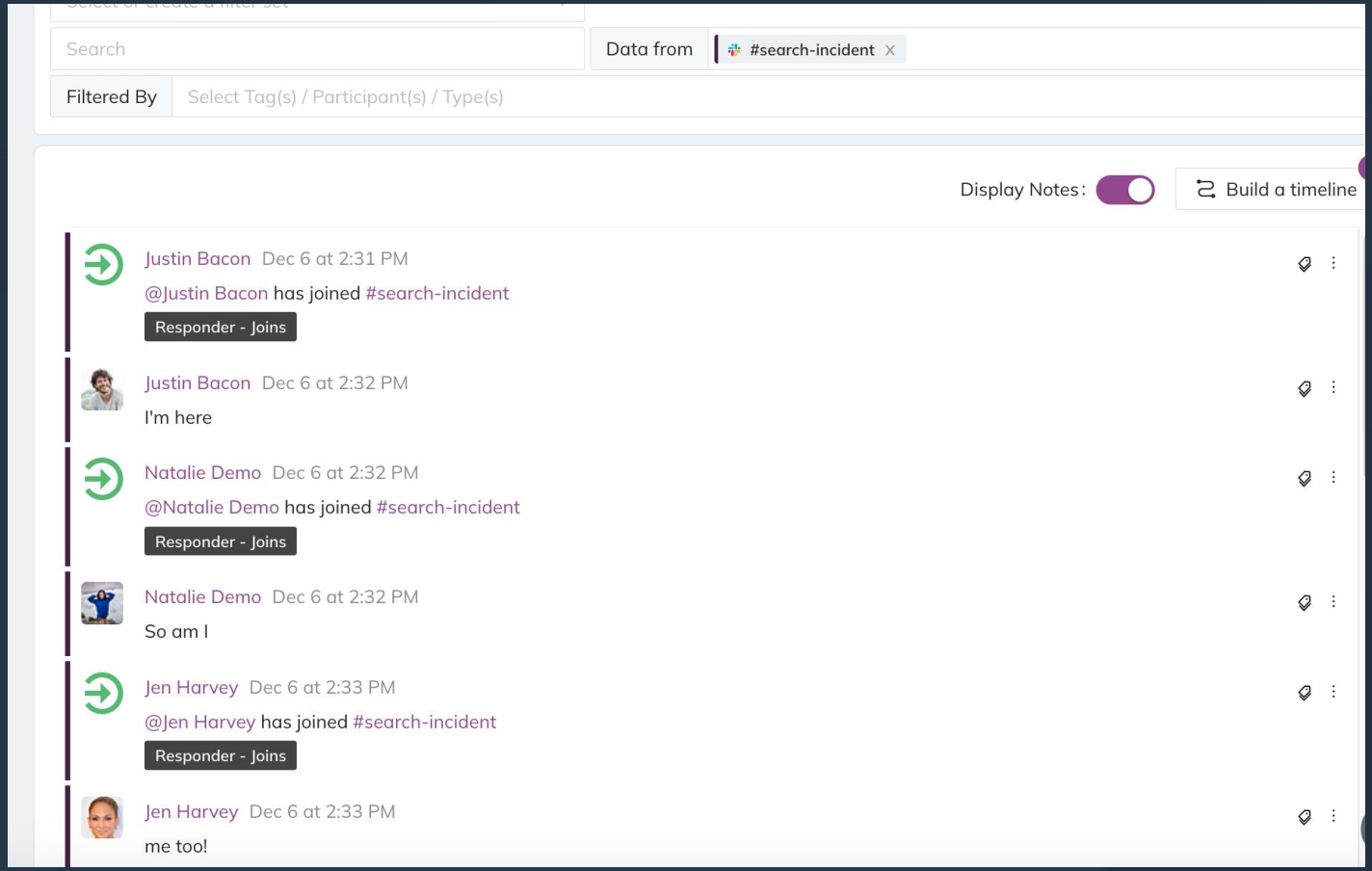


jeli

Incident #2: Timeline

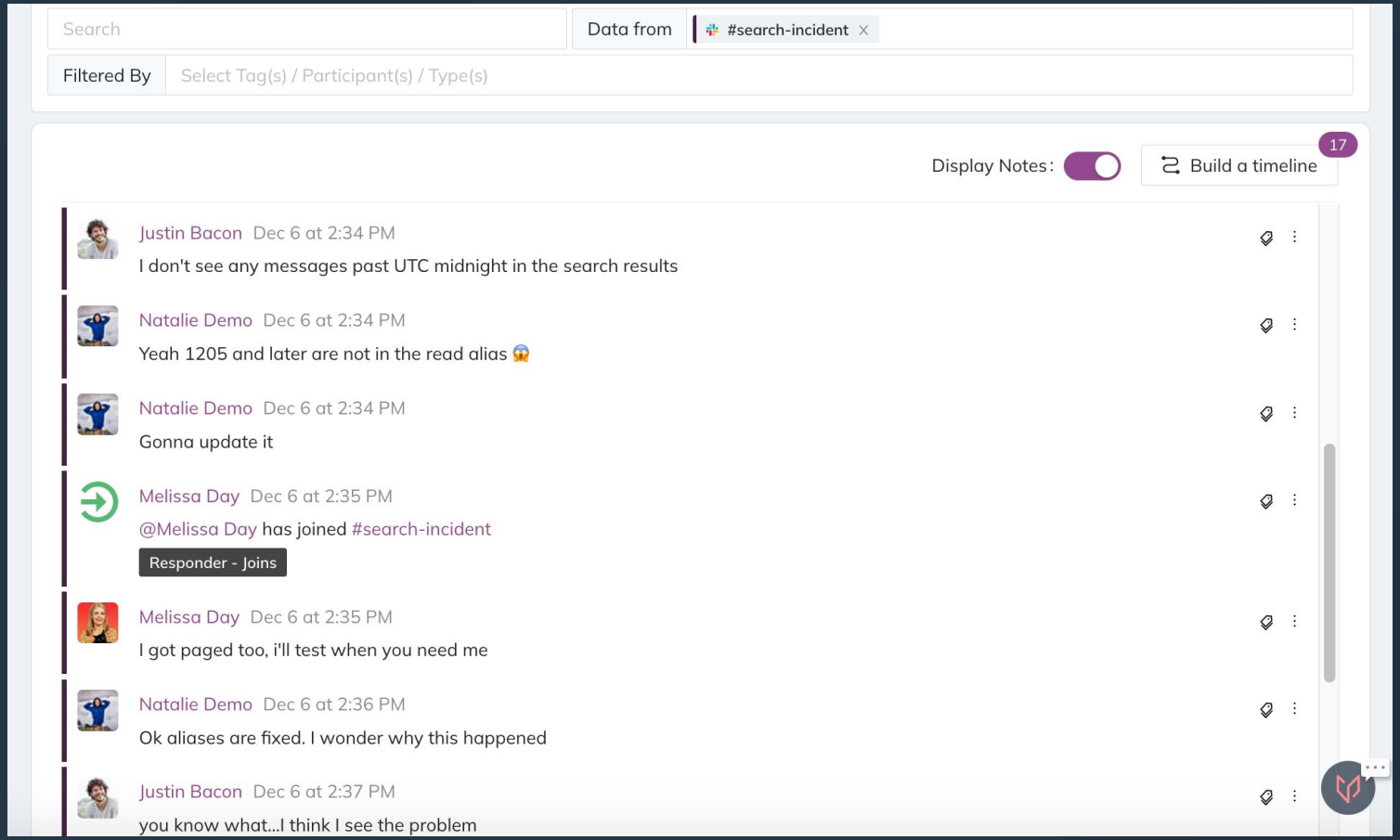


Incident #2: Timeline



jeli

Incident #2: Timeline



jeli

Incident #2: Timeline



Incident #2: Follow-up items

- Evangelization of definition on what is considered critical for search to be working
- Creating an explicit on-call for the human-driven pages rotation
- Segmenting multiple changes into separate PRs, “clean-up” work deserves its own PRs

jeli

Incident #2: Key Takeaways

- There are multiple notions of what it means for search to be “working”
- Incident handling excellence is influenced by team cohesiveness and collaboration
- Changing code with alerts has risk
- Ease of coordination due to familiarity of team
- Quick awareness of issue at hand

	Incident 1	Incident 2
Time to investigate	45 minutes	1 week
Time of incident	20 minutes	Dates back to 3 weeks before the event
Contributing Factors	2	8
Action Items that got completed	2	6
People involved	2	8
Difficulties During Handling	No section	Dedicated section
Impact	Customer	Customer + Employee
Readers of the document	10	140

jeli

Why does a focus on cognition, coordination, and teaming matter to your organization?

jeli

Investigation reports can:

- Give the writer of the document more expertise
- Become training documents,
- Inform chaos experiments,
- Serve as professional development and refresher training, and
- Help enable meta-analysis across incidents.
- Highlight how goal conflicts and changing organizational priorities influence SRE work
- Identify where additional headcount is or is not needed
- Inform build vs buy decisions

jeli

Quick tips for a focus on cognition, coordination, and teaming through incidents:

- Don’t only use customer impact to warrant a review
- Get an outside, technical “investigator”
- Make it engaging

jeli

Part 3 - Collaborative learning

What have you learned?



jeli

Principles of Cognition, Coordination & Teaming

1. Cognitive work is more than applying technical knowledge, and **is often hidden**.
2. Site Reliability engineering requires fine tuned skills in **detection, diagnosis work, and flexibility** in applying deep technical expertise.
3. Working in teams **is necessary and adds cognitive work**.
4. SRE requires fine tuned skills in **anticipating, adjusting, signalling, grounding, delegating, and taking initiative**.
5. Resilient organizations are networks that **share capacities** and adapt according to the demands.

jeli

Keep the conversation going

Come visit us at our booth!

jeli.io

jeli