STPA for Software Systems–Illuminate the Unknown Unknowns

Theo Klein - Staff SRE
Garrett Holthaus - SRE Technical Writer
Ruben Barroso - Staff SRE

Google

*Image created with Gemini 2.5 Flash

Respond in slack whenever you see this box

Type your ideas in Slack!

#25emea-day3-track3

Today's Goals

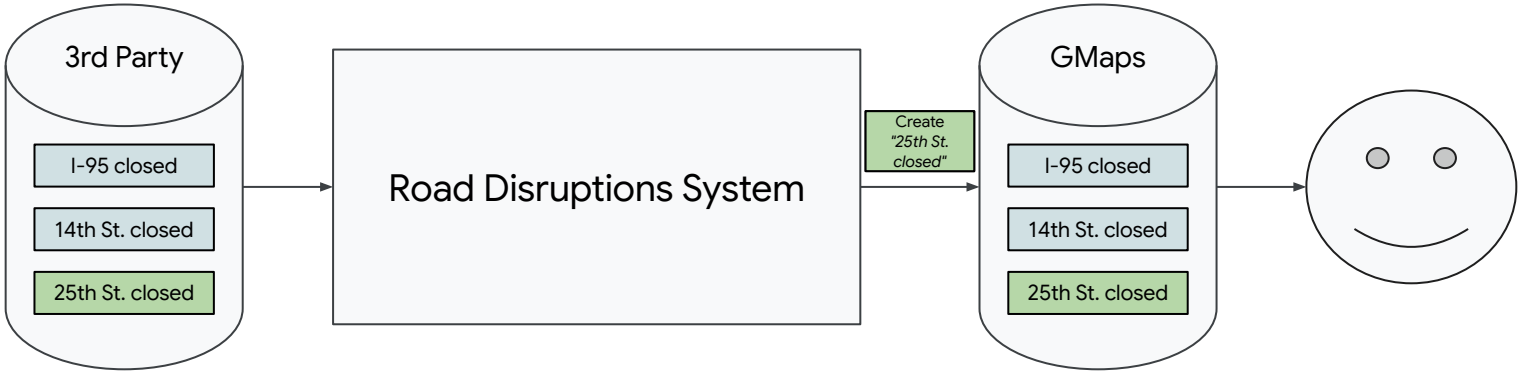
After this talk, we hope that you:

- Can describe what STPA is
- Can provide a high-level explanation of how it works
- Can give a few examples of how STPA can anticipate incidents in software/human systems

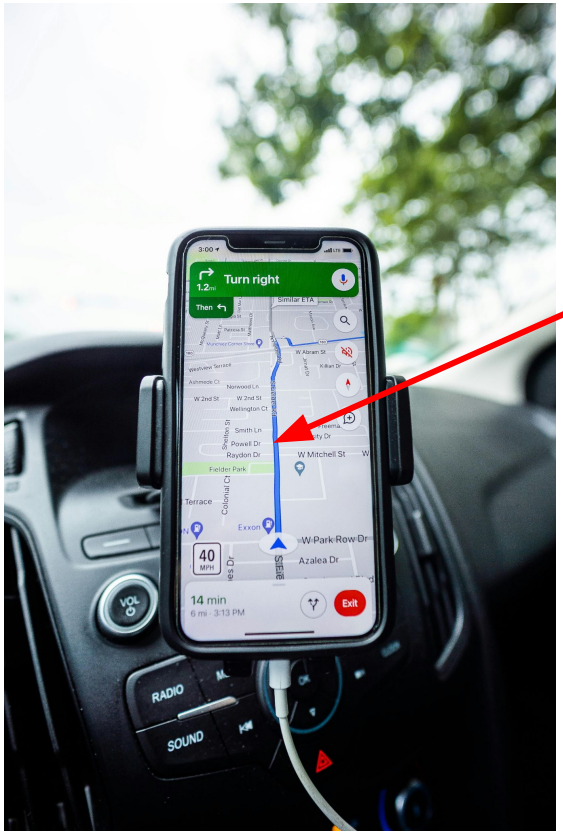

I just got paged!


Road Disruptions System

Publish Road Closures onto Google Maps



The System Didn't Work



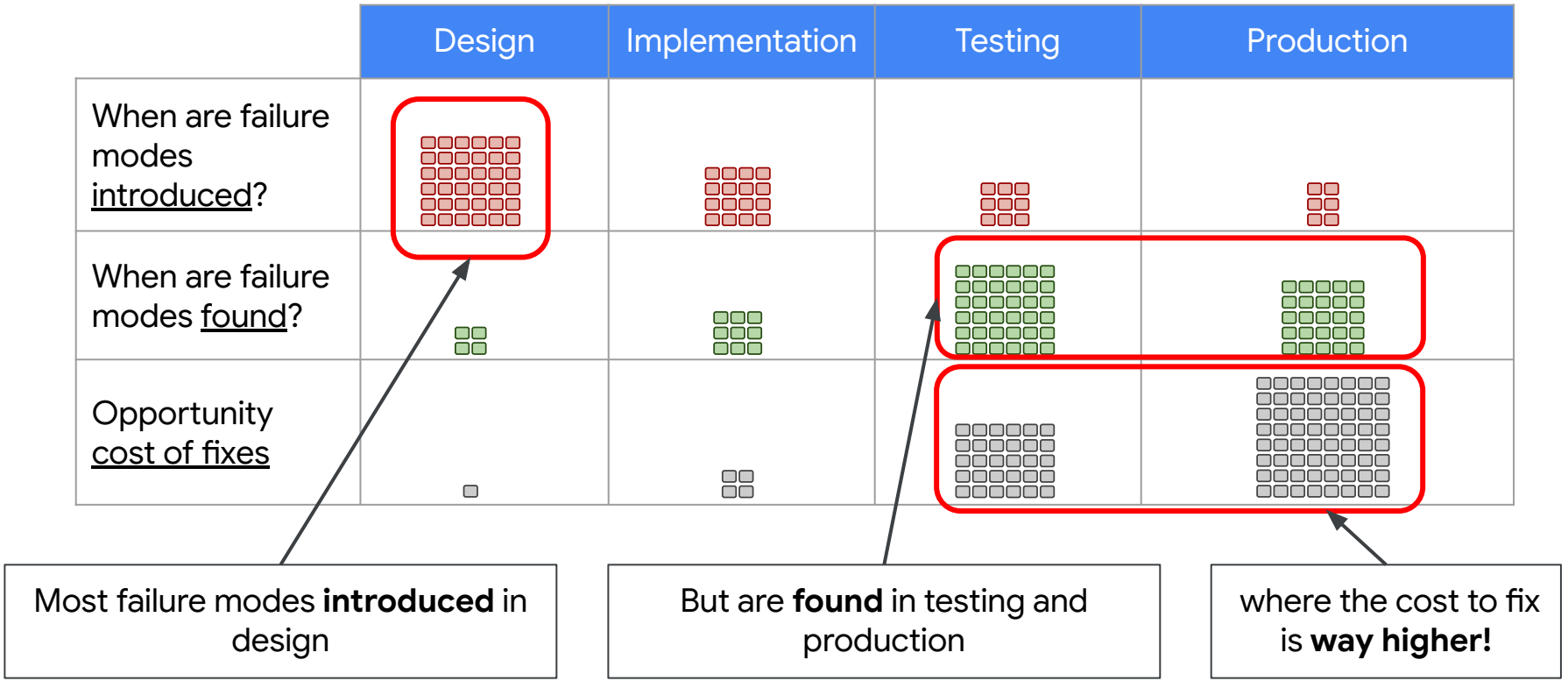
Should be marked as closed, but isn't

Life of a Software Product

Life of a Software Product

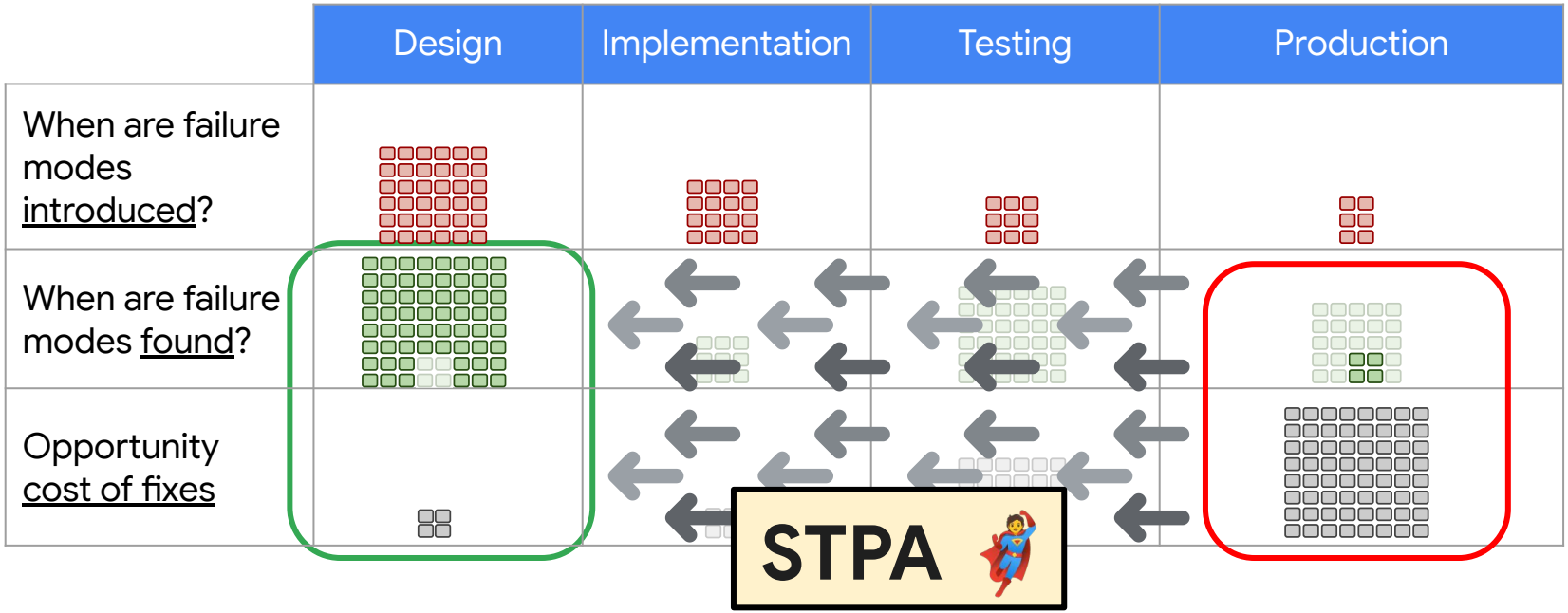
Anticipate your Incidents at the design stage

The Hidden Cost* of Defects



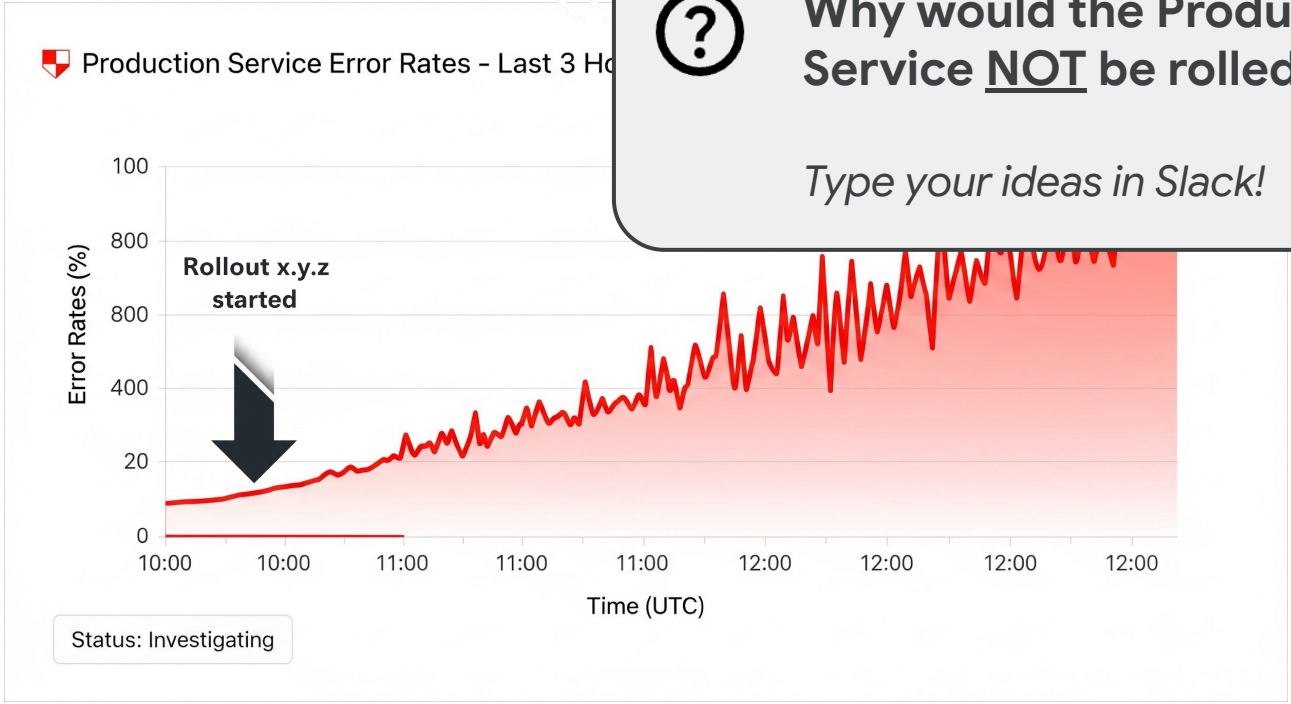
* Coarse characterization, adapted for Google from System Safety and STPA Class Materials, John Thomas, 2021
Data from "ROI Analysis of the System Architecture Virtual Integration Initiative", SEI, 2018

The Hidden Cost* of Defects



* Coarse characterization, adapted for Google from System Safety and STPA Class Materials, John Thomas, 2021
Data from "ROI Analysis of the System Architecture Virtual Integration Initiative", SEI, 2018

Warm Up!



Why would the Production Service NOT be rolled back?

Type your ideas in Slack!

*Image created with Gemini 2.5 Flash

What is STPA?

STPA - System Theoretic Process Analysis

- Paper and pencil method
- Output: Comprehensive list of all the ways a system can have a loss (unacceptable outcomes)
- In use by safety engineers in other industries (aerospace, nuclear, medical) for more than 10 years, first use at Google in 2020
- Google is applying STPA to software design

You can use STPA to find incidents waiting to happen when all you have is a design document!

STPA works!

- Google Maps: you'll see examples of issues we discovered
- We've found and fixed issues before they cause incidents
- For incidents that already happened, STPA found additional issues
- Use of STPA has spread to other products and organizations

This incident never happened, because we used STPA to find and fix the causal design flaws!

Image source: https://www.pexels.com/photo/cellphone-on-a-stand-with-navigation-map-application-open-7738879/

Looking for Design Flaws

- Known Unknowns: we at least know where to look
- Unknown Unknowns: we don't even know where to look
- STPA: guides you on where to look!

*Image created with Gemini 2.5 Flash

Today's Agenda

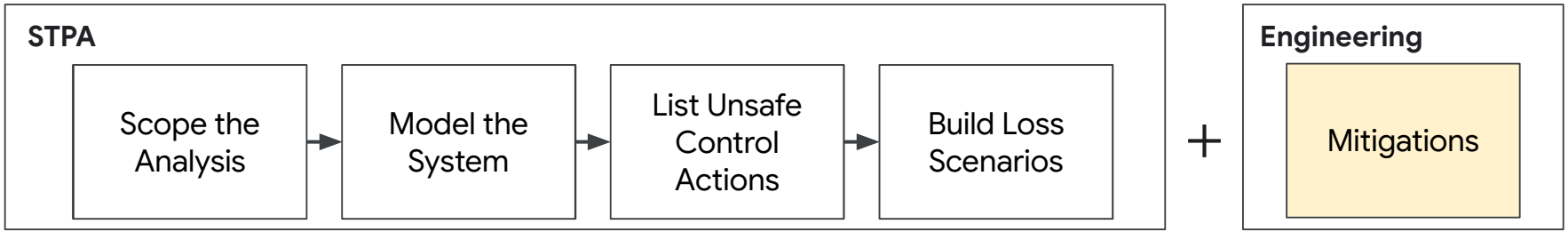
- **The Four Steps of STPA**
- **Analyze the Rollback Decision**
- **5 min break**
- **How we Applied STPA to the Road Disruptions System**

⚠️ Disclaimer ⚠️

We won't teach you the methodology in 90 minutes.

STPA: Four Steps

STPA Overview

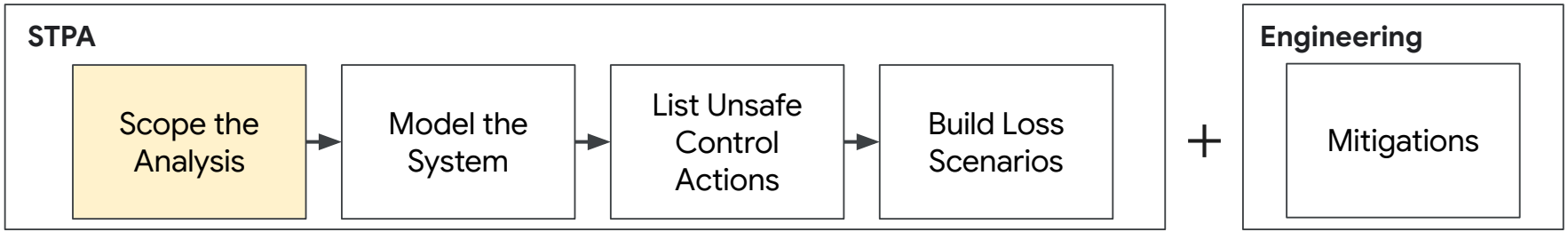


Let's Analyze:

Why would the production service NOT be rolled back?

*Image created with Gemini 2.5 Flash

STPA Step 1


STPA
Scope the Analysis
Model the System
List Unsafe Control Actions
Build Loss Scenarios
+
Engineering
Mitigations

Step 1: Define Losses

- **Losses:** Unacceptable system outcomes
  - Loss of revenue
  - Loss of brand trust
  - Loss of legal compliance
  - Injury or loss of life
- Losses should not reference low-level details or causes

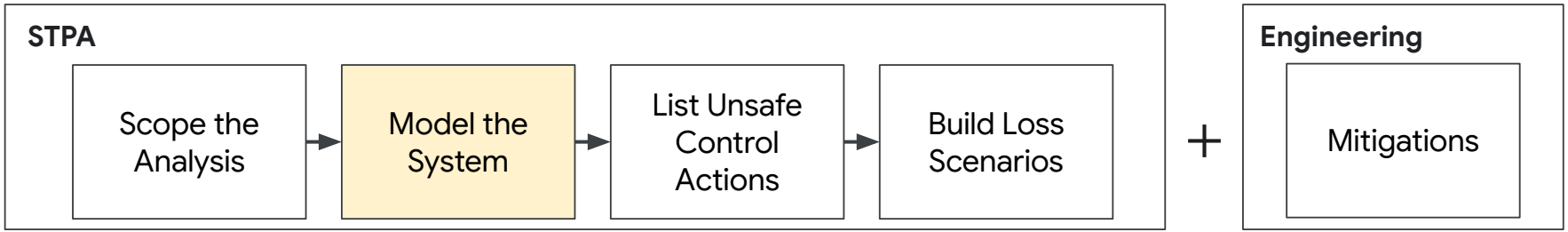
Step 1: Define Hazards

- **Hazard:** A system state that will lead to a loss under worst case environmental conditions
- E.g. *"Production service is returning errors"*
- Hazard states give you time

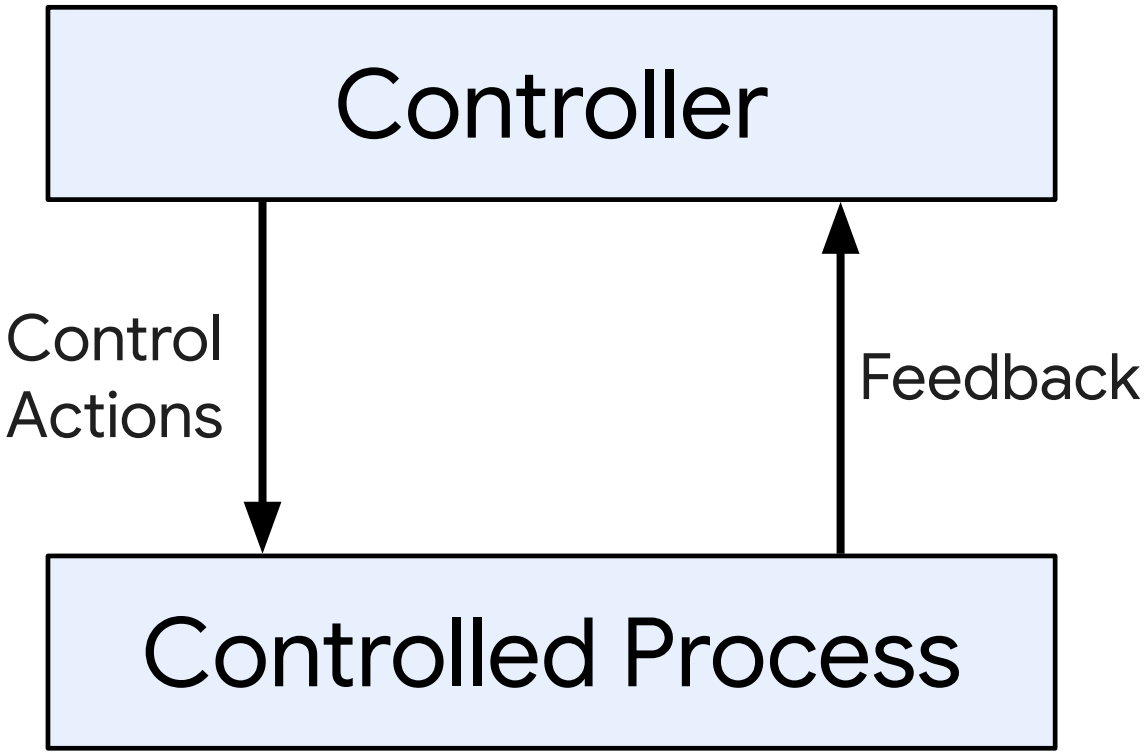
Can you think of a reason why this hazard <u>might not</u> lead to a loss?

Type your ideas in Slack!

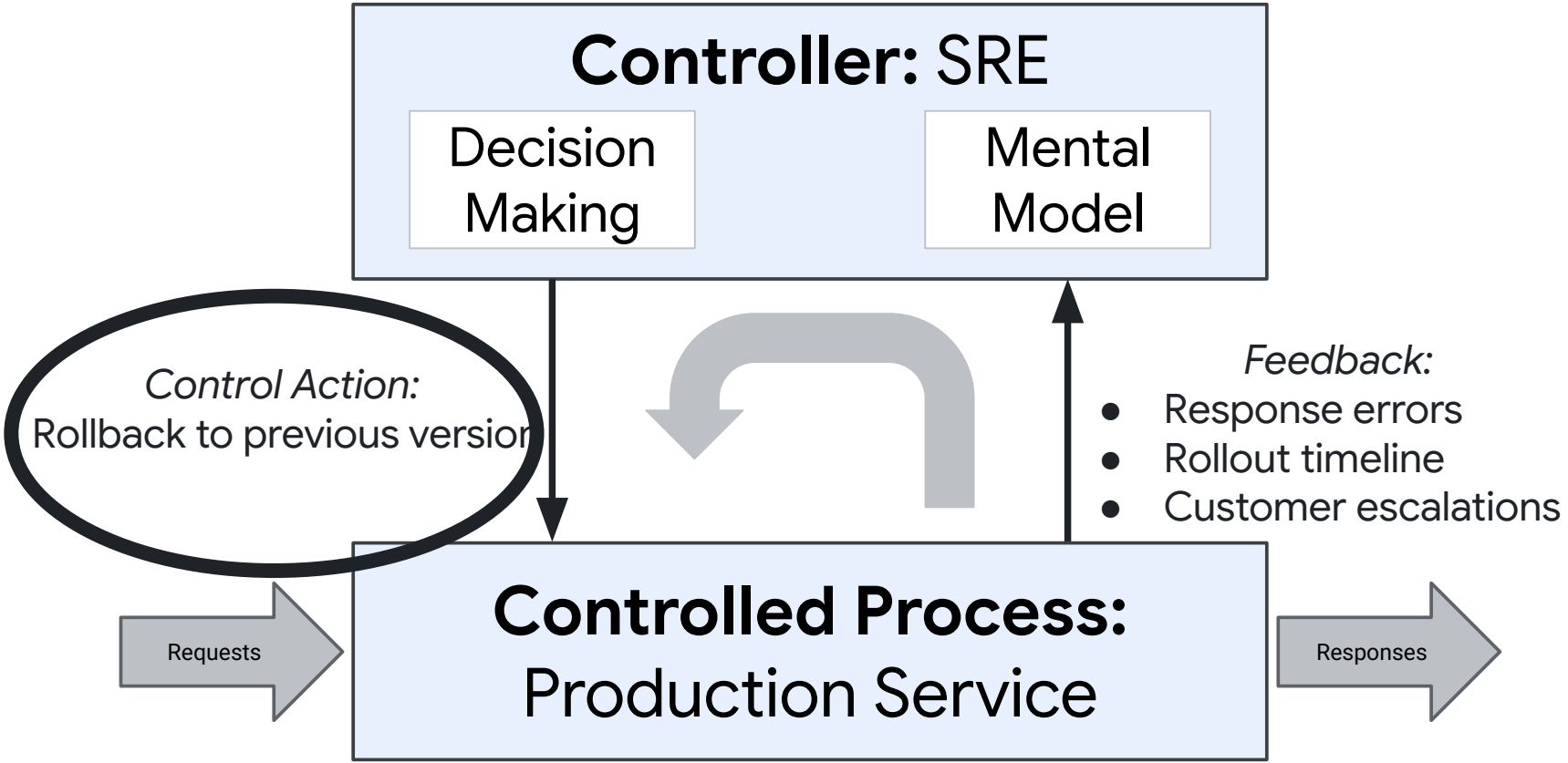
STPA Step 2


STPA
Scope the Analysis
Model the System
List Unsafe Control Actions
Build Loss Scenarios
+
Engineering
Mitigations

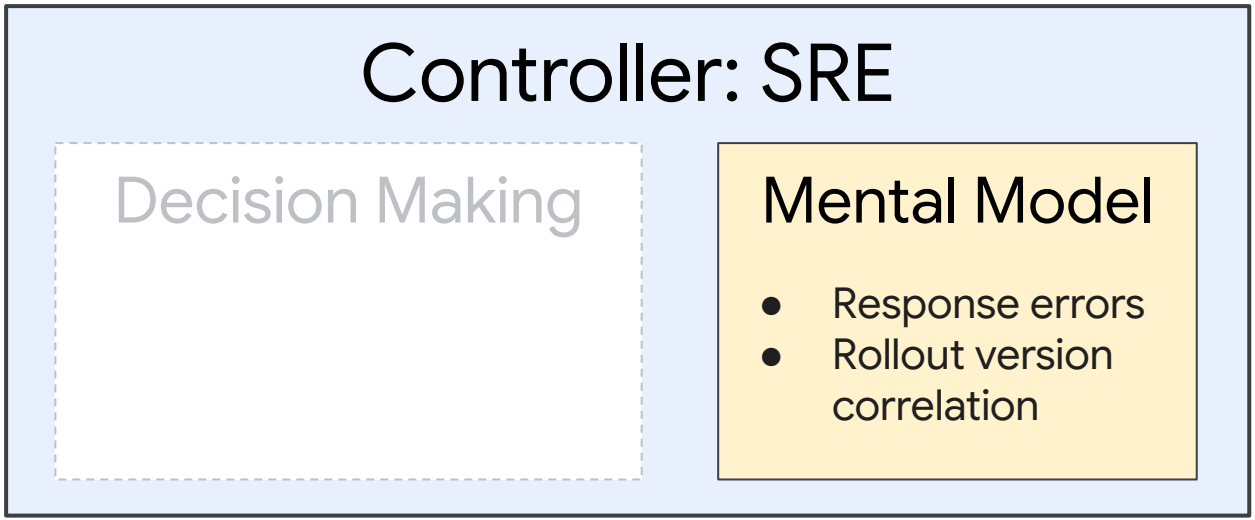
Step 2: Basic control feedback loop



To roll back, or not to roll back, that is the question



To roll back, or not to roll back, that is the question



What do you consider when deciding whether to roll back?

Type your ideas in Slack!

To roll back, or not to roll back, that is the question

Controller: SRE

Decision Making

IF [illegible] **THEN** [illegible]
THEN roll back

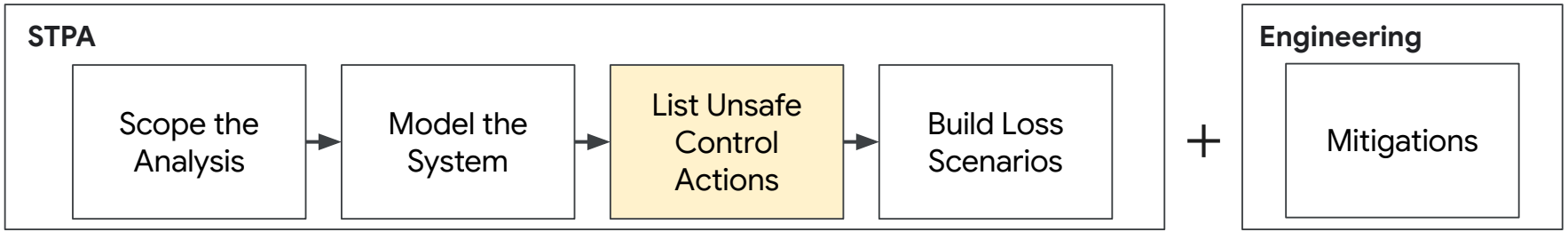
Mental Model

- Response errors
- Rollout version correlation

How do you decide whether to roll back to a previous version?

Type your ideas in Slack!

STPA Step 3


STPA
Scope the Analysis
Model the System
List Unsafe Control Actions
Build Loss Scenarios
+
Engineering
Mitigations

Step 3: List Unsafe Control Actions (UCAs)

Unsafe Control Action: A control action that, *in a particular context*, will lead to a hazard.

Step 3: What makes an action unsafe?

"Use butter knife"

SAFE or **UNSAFE**?

Type your ideas in Slack!

*Image created with Gemini 2.5 Flash

Step 3: What makes an action unsafe?

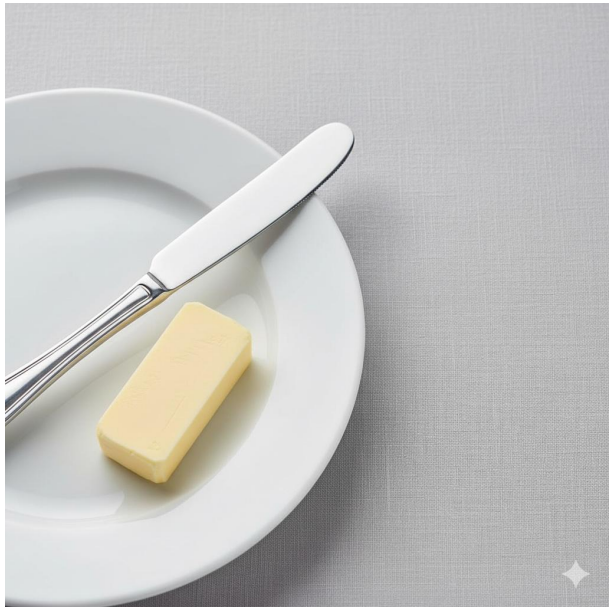
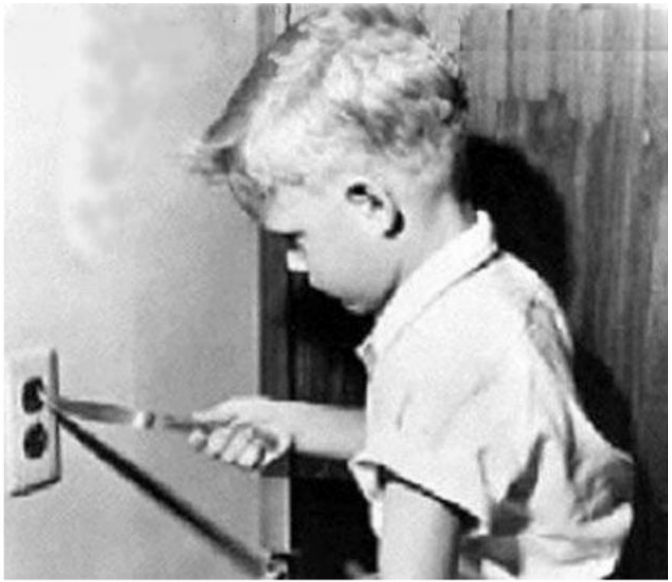


*Image created with Gemini 2.5 Flash



Context is key!

Step 3: List Unsafe Control Actions (UCAs)

- SRE rolls back to previous version when the production service is not returning errors
- SRE rolls back to previous version when...
  the previous version has a production i
- SRE rolls back to previous version whe
  compatible with production

When is it unsafe to roll back to a previous version?

Type your ideas in Slack!

Step 3: List UCAs (cont.)

- SRE rolls back to previous version when the production service is not returning errors
- SRE **does not roll back** to previous version when... the production service is returning errors

When is it unsafe to NOT roll back to a previous version?

Type your ideas in Slack!

Step 3: Exploring *all* UCAs

STPA guides you to **all** the unsafe actions

- Roll back when <*context*>
- Not roll back when <*context*>
- Roll back before/after <*context*>
- Roll back stopped too soon / applied too long <*context*>

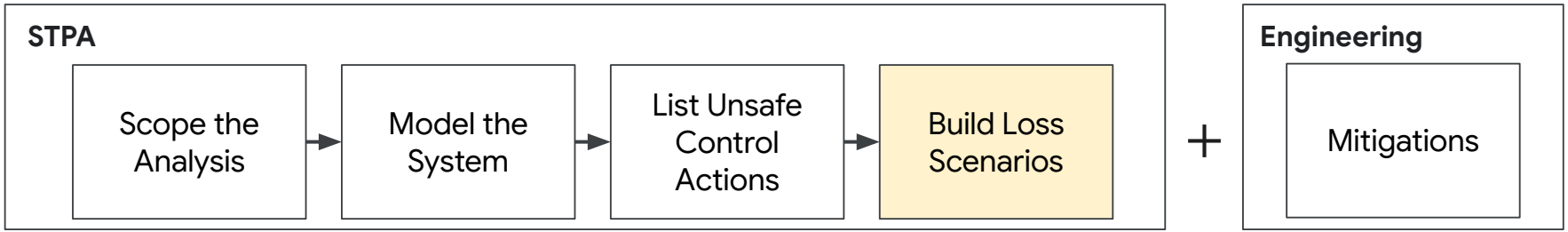
Step 3: Exploring *all* UCAs

- SRE **rolls back** when the current production service isn't returning errors
- SRE **does not roll back** when the current production service returns errors
- SRE **rolls back too late** after cu
  errors
- SRE **rolls back for too long** aft
  returning errors

Complete the sentence: SRE rolls back too late...

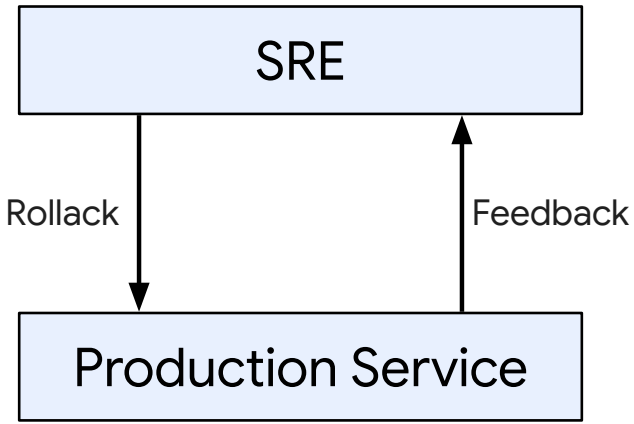
Type your ideas in Slack!

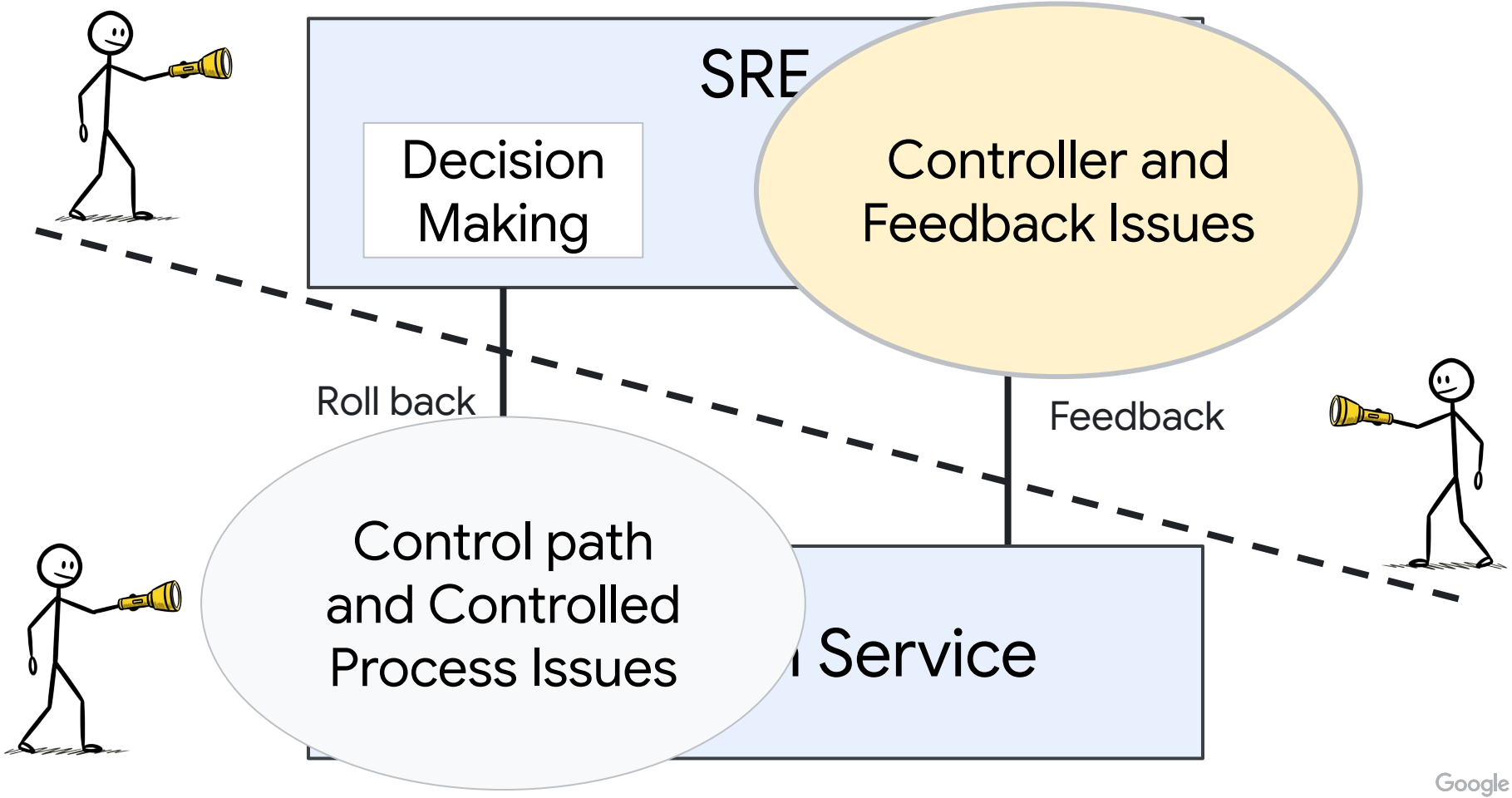
STPA Step 4

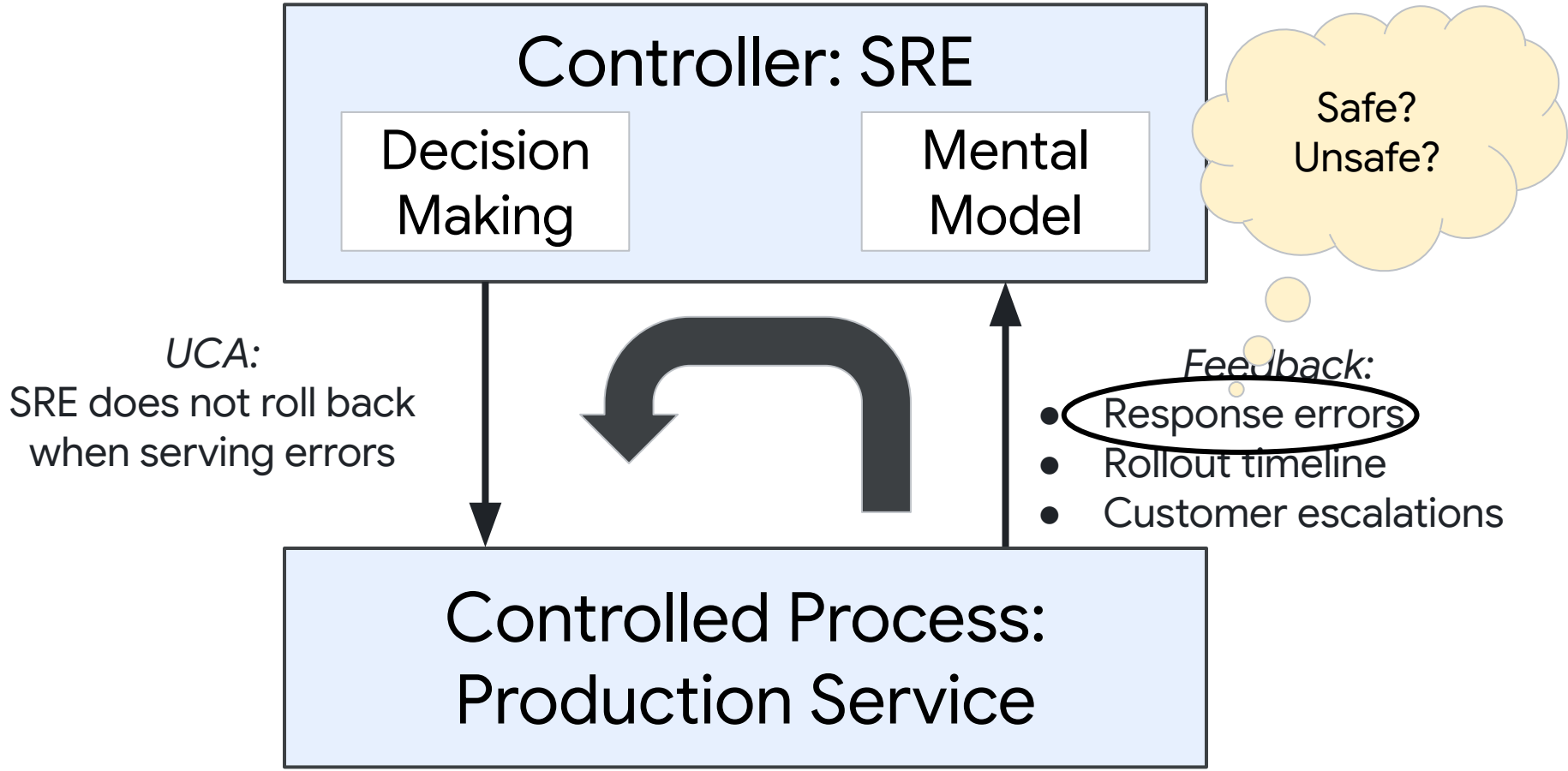

STPA
Scope the Analysis
Model the System
List Unsafe Control Actions
Build Loss Scenarios
Engineering
Mitigations

Step 4: Generate Loss Scenarios

- Iterate through UCA list
- Goal: Comprehensive list of the ways a UCA can happen, leading to hazard states and losses
- Ask the system experts a series of questions:
  - Unsafe decision making?
  - Unsafe feedback?
  - Unsafe control path?
  - Unsafe controlled process?



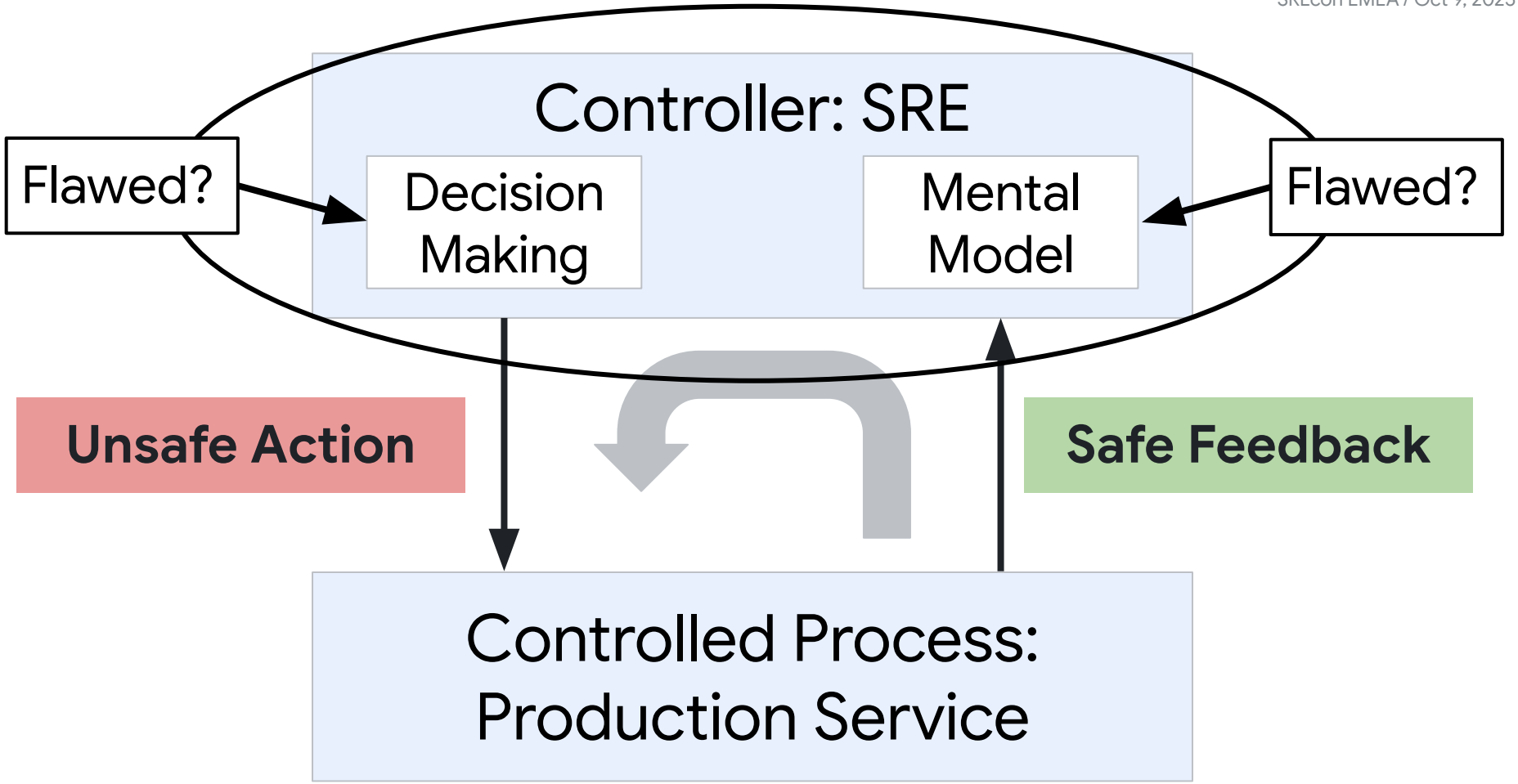

SRE
Decision Making
Controller and Feedback Issues
Roll back
Feedback
Control path and Controlled Process Issues
Service


Controller: SRE
Decision Making
Mental Model
Safe?
Unsafe?
UCA:
SRE does not roll back when serving errors
Feedback:
Response errors
Rollout timeline
Customer escalations
Controlled Process: Production Service

Step 4: Building Scenario Questions

- **Pick one UCA**: SRE does not roll back to previous version when we are serving errors
- **Pick one feedback**: Server errors
- **Choose whether feedback is safe, or unsafe**
- **Ask the question!**

Why would the production service *NOT* be rolled back in spite of feedback existing that shows server errors?


Controller: SRE
Flawed?
Decision Making
Mental Model
Flawed?
Unsafe Action
Safe Feedback
Controlled Process: Production Service

Step 4: Generic controller problem questions

Why would the SRE decide to ***not*** rollback in spite of receiving feedback showing server errors?

Responsibilities
"Who makes the decision to roll back?"

Decision-making rationale
"How does the SRE decide whether to roll back?"

etc...

Mental Model
"What dashboards/signals/etc. do you look at, and how does the SRE interpret them to determine production health?"

Inputs from a Higher Authority Controller
"Does anyone ever order the SRE to do a roll back?"

Step 4: Generic controller problem questions

Why would the SRE decide to *NOT* rollback in spite of receiving feedback showing server errors?
Type your ideas in Slack!

Responsibilities
"Who makes the decision to roll back?"

Decision-making rationale
"How does the SRE decide whether to roll back?"

etc...

Mental Model
"What dashboards/signals/etc. do you look at, and how does the SRE interpret them to determine production health?"

Inputs from a Higher Authority Controller
"Does anyone ever order the SRE to do a roll back?"

Step 4: Example Scenarios

Why would the SRE decide to ***not*** roll back in spite of receiving feedback showing server errors?

- Dev and SRE teams share on-call, each expecting the other to handle alerts
- New traffic shape triggers latent bug; on-caller associates errors with traffic change, not rollout
- On-caller considers the alert overly sensitive and expects errors

Difference between UCA and Scenario

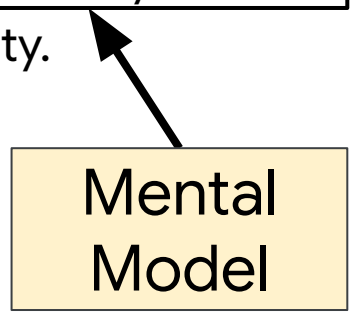
UCA

SRE does not roll back to previous version when the production service is returning errors

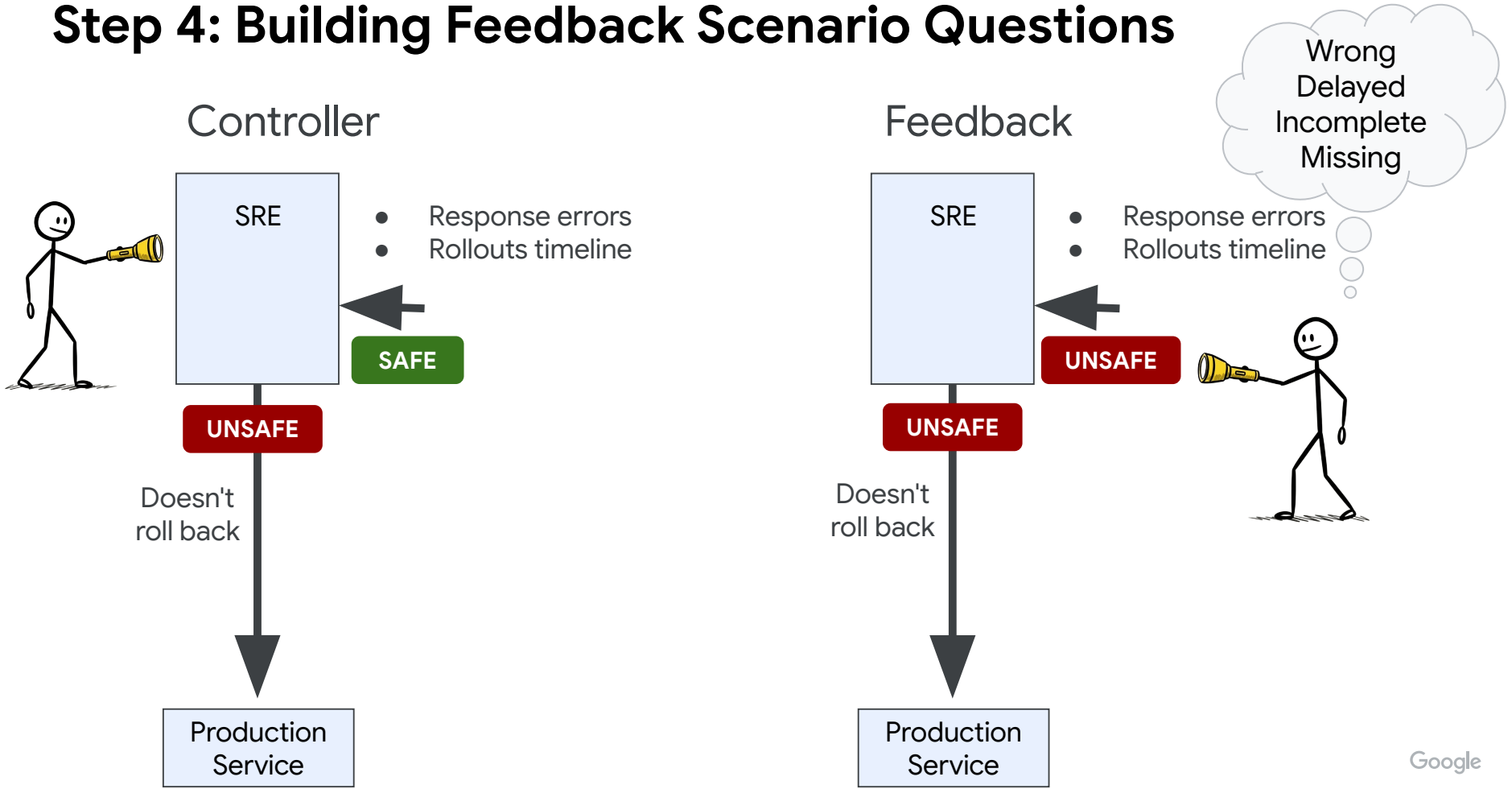
Scenario

SRE does not roll back to previous version when the production service is returning errors.

This is because the Dev and SRE teams share on-call responsibility, and the SRE incorrectly believes that the Dev is on duty.



Step 4: Building Feedback Scenario Questions



Step 4: Building Feedback Scenario Questions

- **Pick one UCA**: SRE does not roll back to previous version when we are serving errors
- **Pick one feedback**: Server errors
- **Feedback:** Assume the feedback is unsafe
- **Ask the question! Specify the true state of the system.**

Why would feedback *NOT* indicate that the service is returning errors, even though it is?

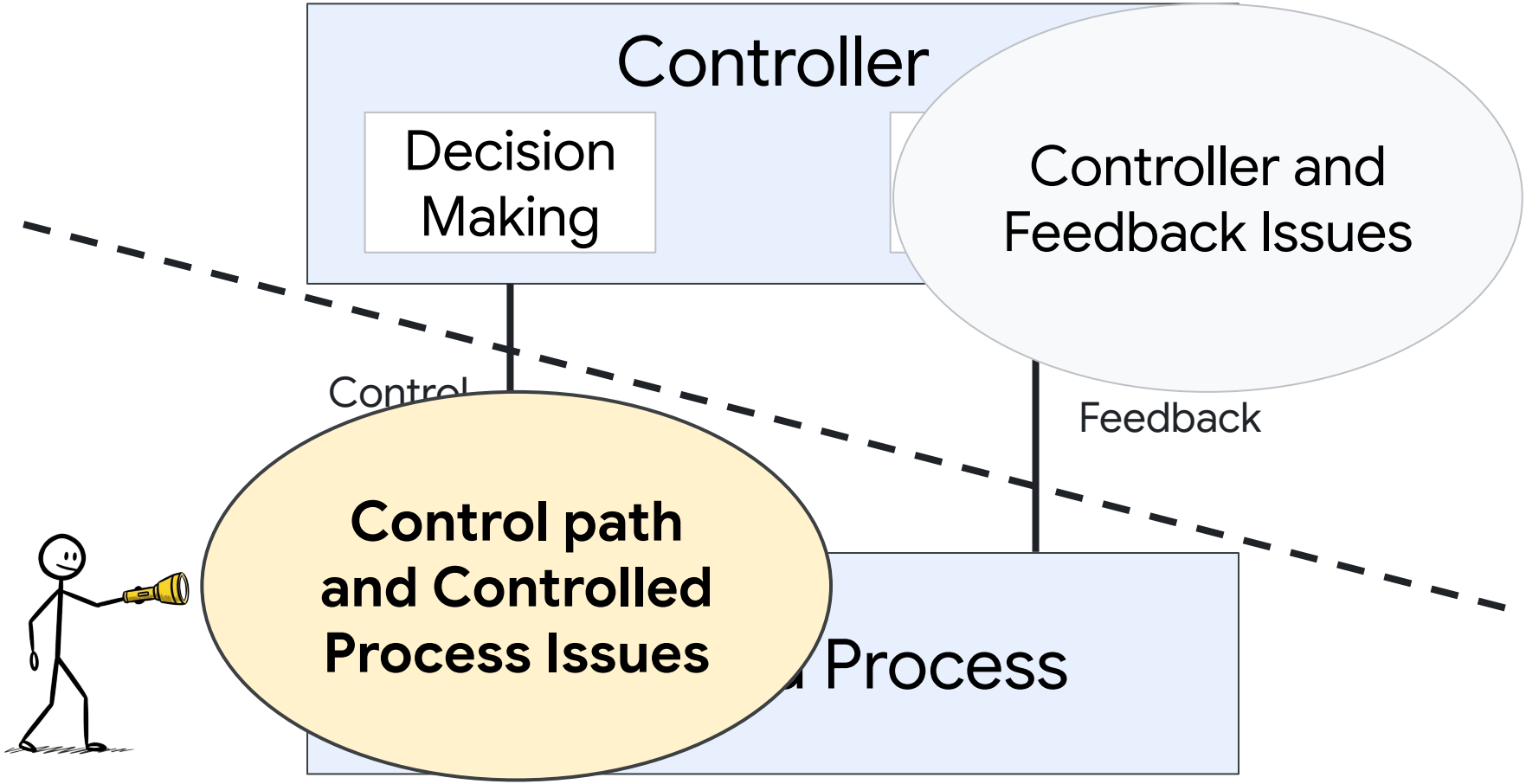
Refining the Feedback Scenario Question

Why would feedback *NOT* indicate that the service is returning errors, even though it is?

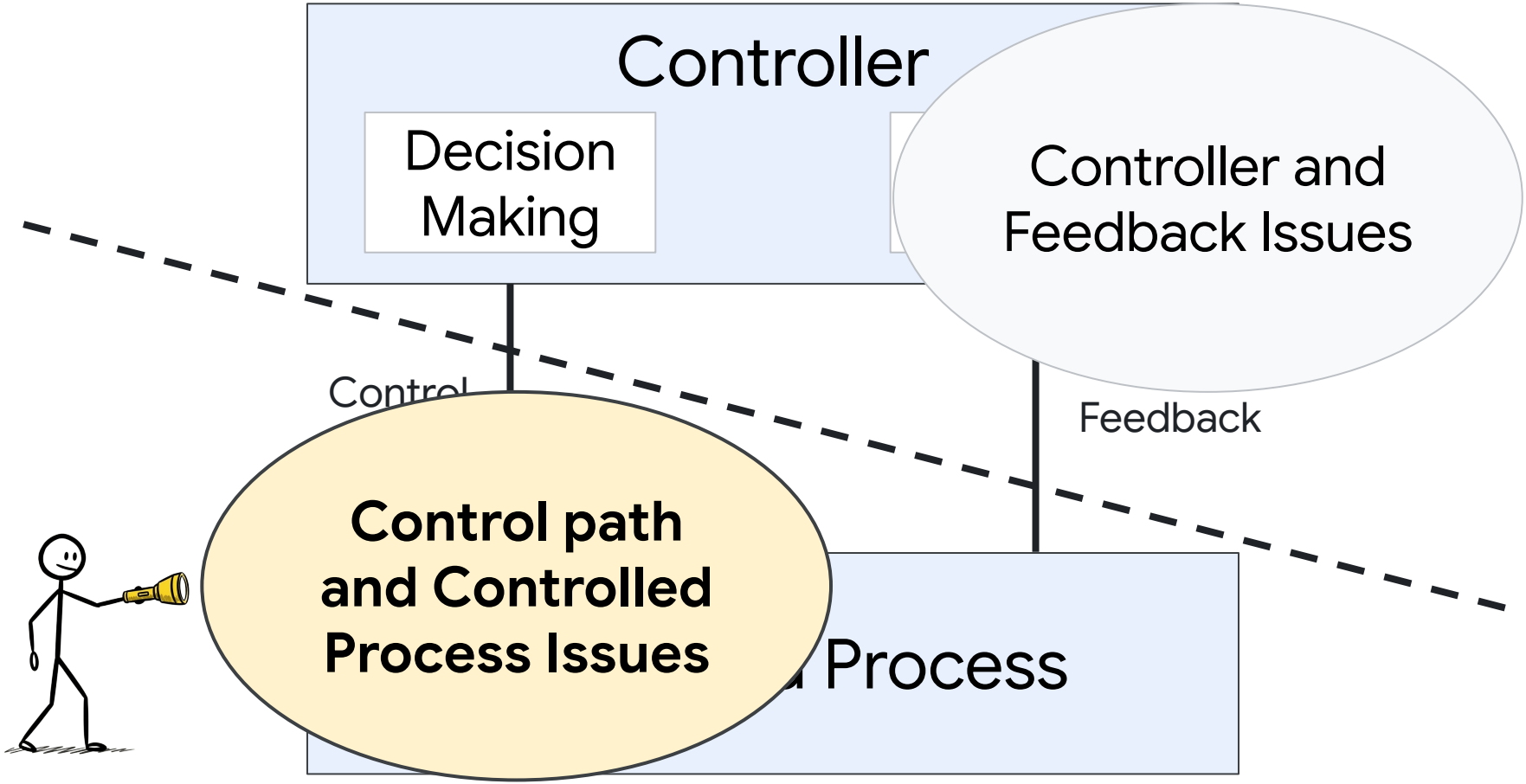
- There are no alerts set up for these types of errors.
- The collection of telemetry is delayed.
- The alerts are sent to the wrong group destination.
- Alert storm, backends make singling out the errors difficult.

What are some scenarios where this could this happen?

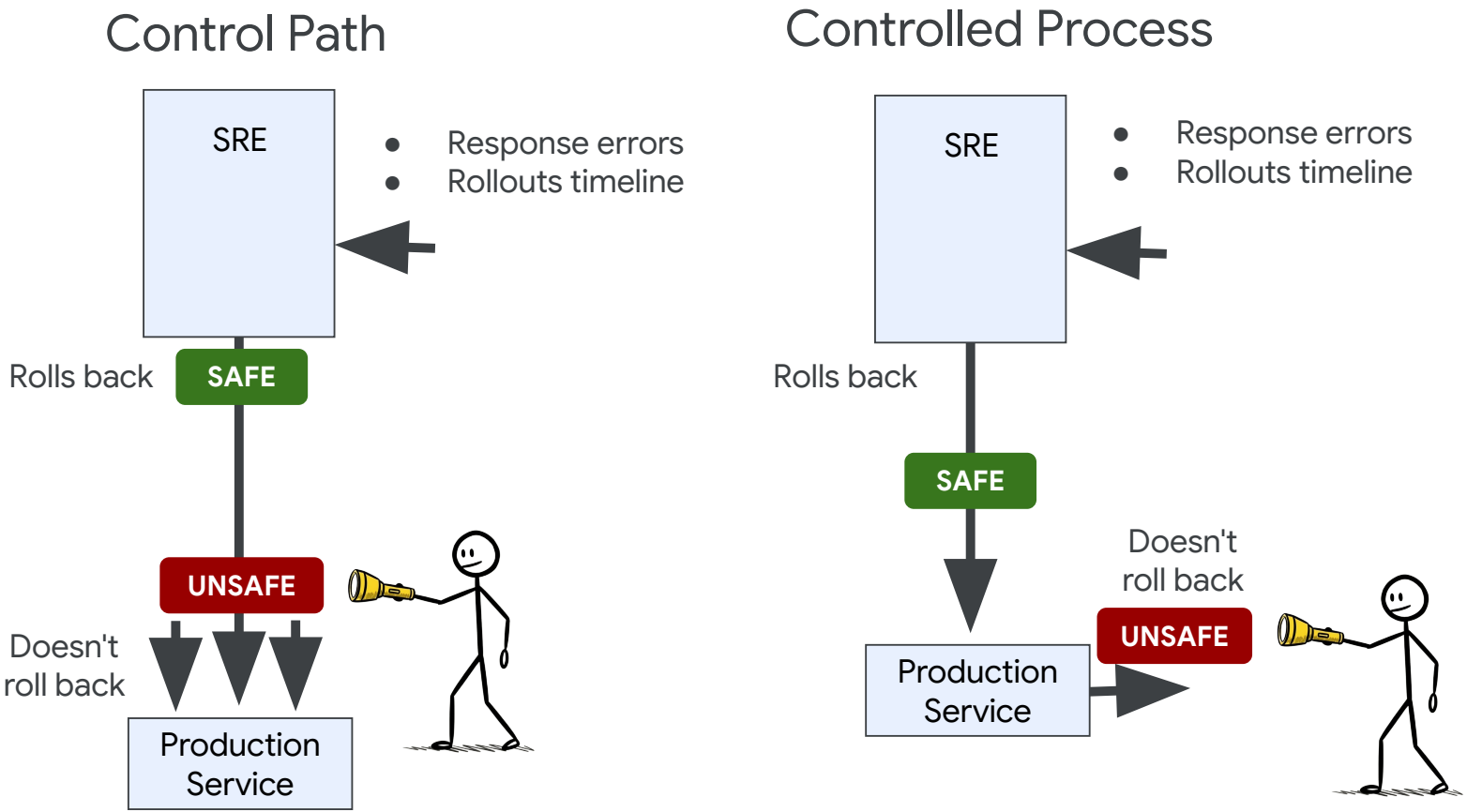
Type your ideas in Slack!


Controller
Decision Making
Controller and Feedback Issues
Feedback
Control path and Controlled Process Issues
Process

5 minute break


Controller
Decision Making
Controller and Feedback Issues
Feedback
Control path and Controlled Process Issues
Process

Step 4: Building Control Path Scenario Questions



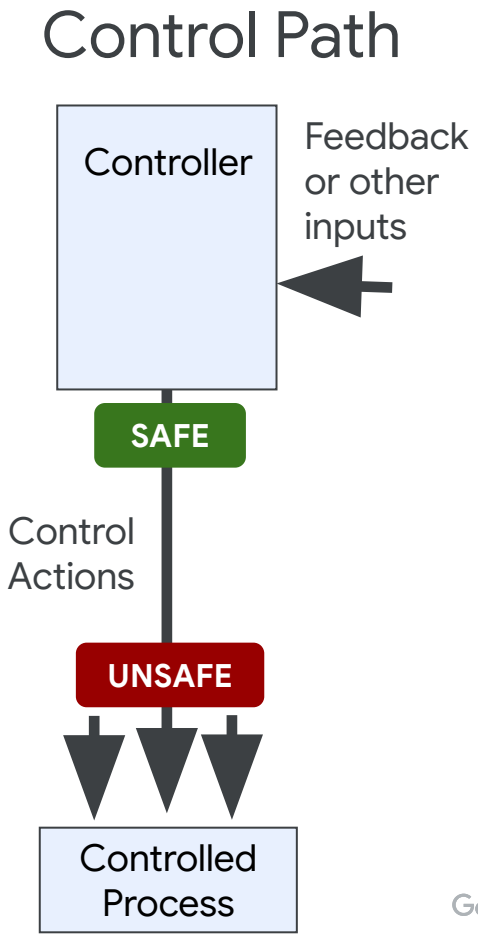
Step 4: Control Path Scenario Questions

UCA: SRE does not roll back to previous version when production service is returning errors

Scenario Question: How could it ever be the case that the SRE initiates a roll back, but the roll back doesn't reach production?

What are some scenarios where this could this happen?

Type your ideas in Slack!



Refining the Control Path Scenario Question

How can you explain that the SRE initiates a roll back, but the roll back doesn't reach production?

The release system receives the rollback action but:

- Limited resources, serializes the rollback along other release requests. No emergency prioritization.
- Oncaller doesn't have rollback permissions. Release system only reports the permission error when starting the rollback. Oncaller doesn't check the UI.
- An ongoing rollout for the same production service blocks the rollback.

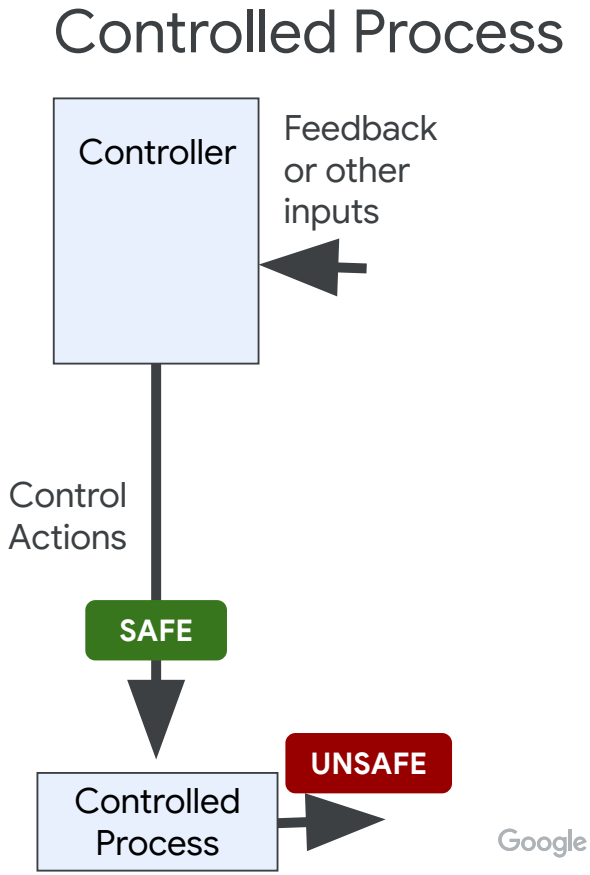
Step 4: Controlled Process Scenario Questions

UCA: SRE does not roll back to previous version when production service is returning errors

Scenario Question: Why would the rollback reach the production service, but the production service isn't rolled back to a previous version?

What are some scenarios where this could happen?

Type your ideas in Slack!



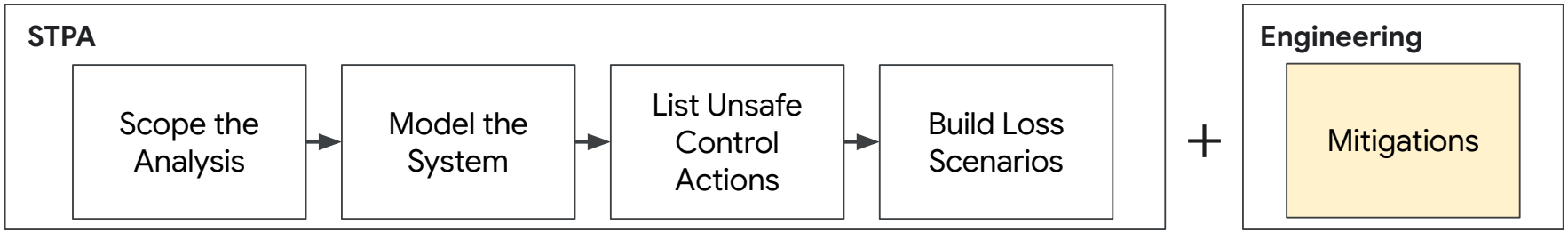
Refining the Controlled Process Scenario Question

Why would the rollback reach production, but the production service isn't rolled back to the previous version?

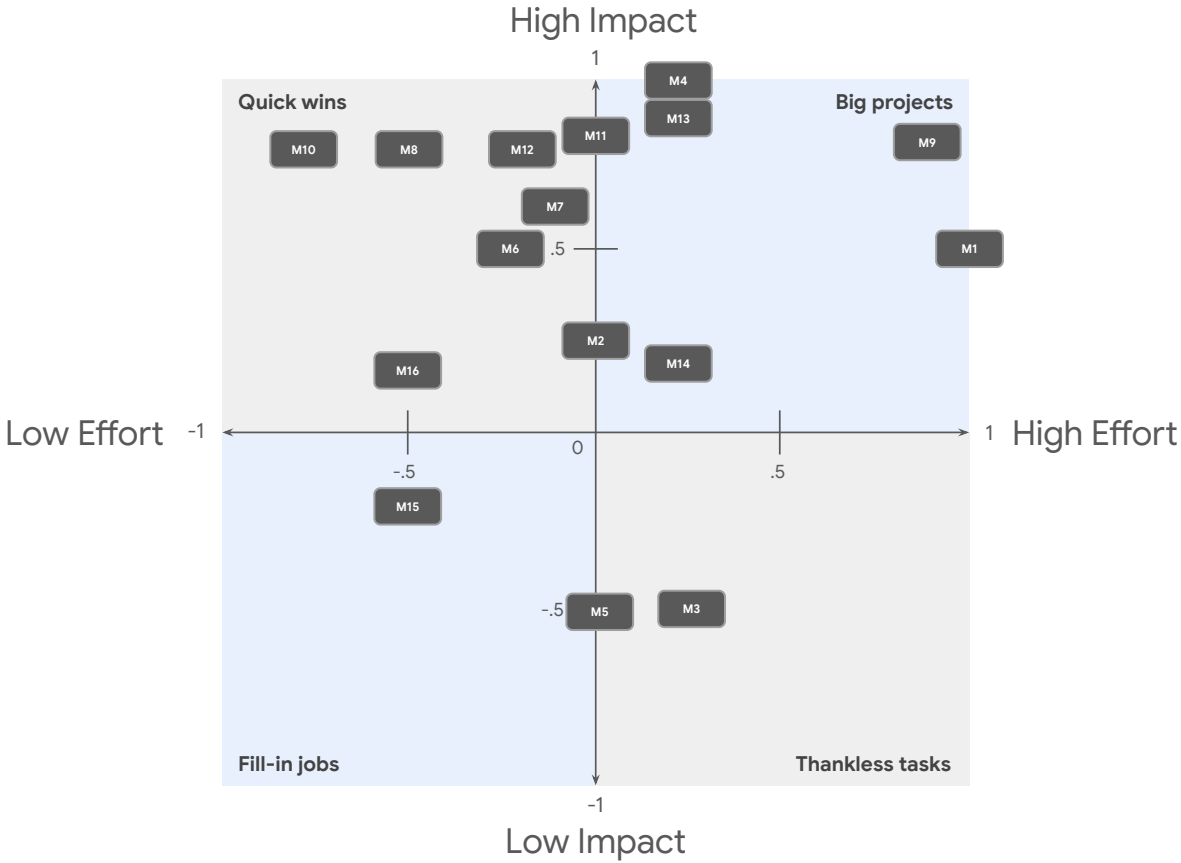
The rollback action arrives to the production service but:

- Someone with permissions cancels the rollback by applying a global stop button
- A backend service was turned down after a new version removed its dependency; when a bug in the new version requires a rollback, the older version can't restart.

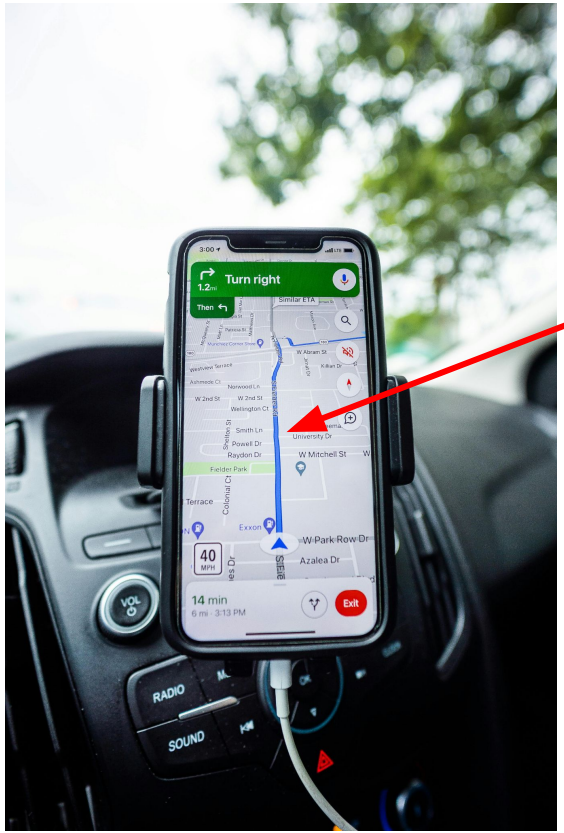
Step 5: Mitigations



Step 5: Mitigations

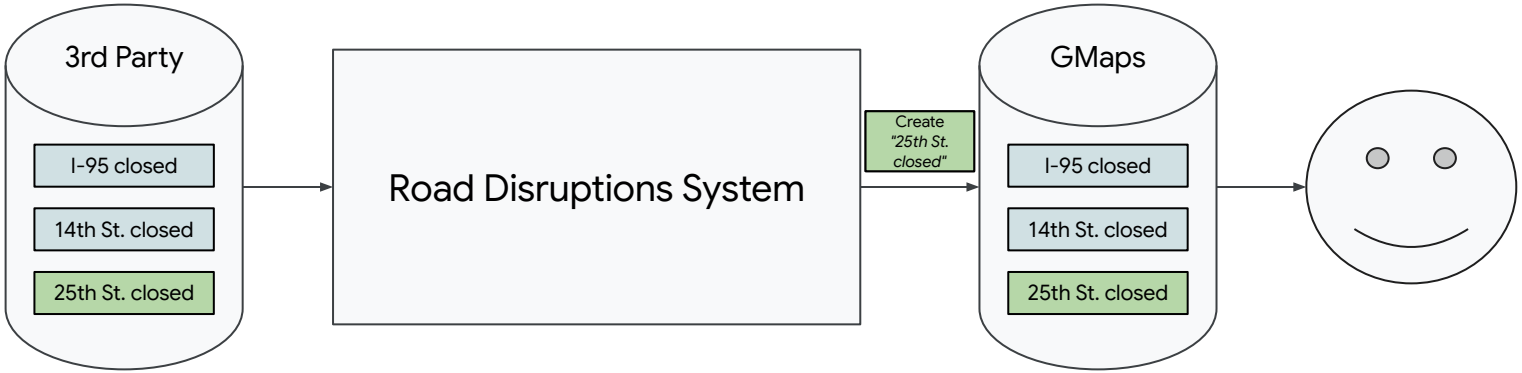


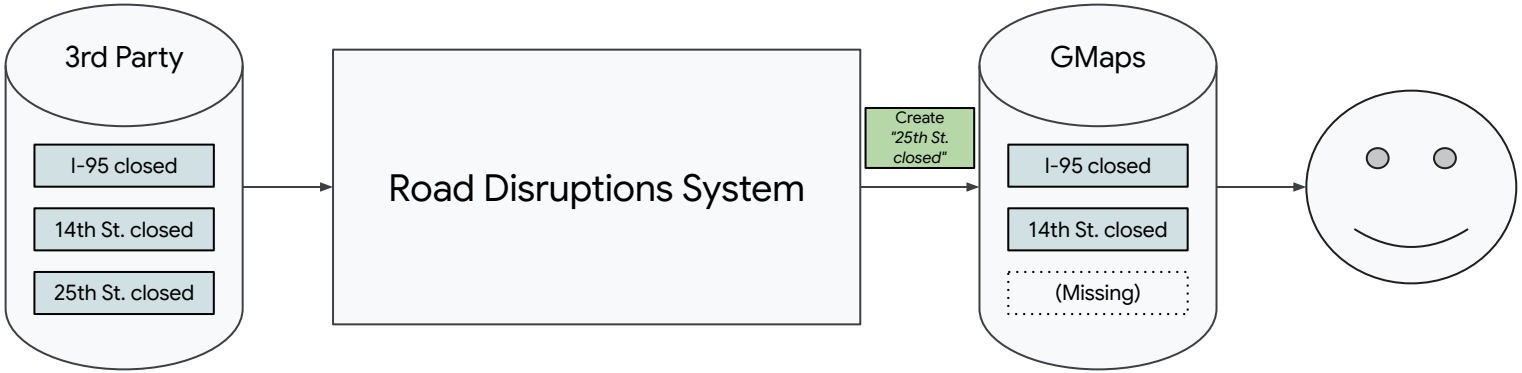
Story Time:
STPA on Road Disruptions



Should be marked as closed, but isn't

Image source: https://www.pexels.com/photo/cellphone-on-a-stand-with-navigation-map-application-open-7738879/


3rd Party
I-95 closed
14th St. closed
25th St. closed
Road Disruptions System
Create "25th St. closed"
GMaps
I-95 closed
14th St. closed
25th St. closed


3rd Party
I-95 closed
14th St. closed
25th St. closed
Road Disruptions System
Create
"25th St.
closed"
GMaps
I-95 closed
14th St. closed
(Missing)

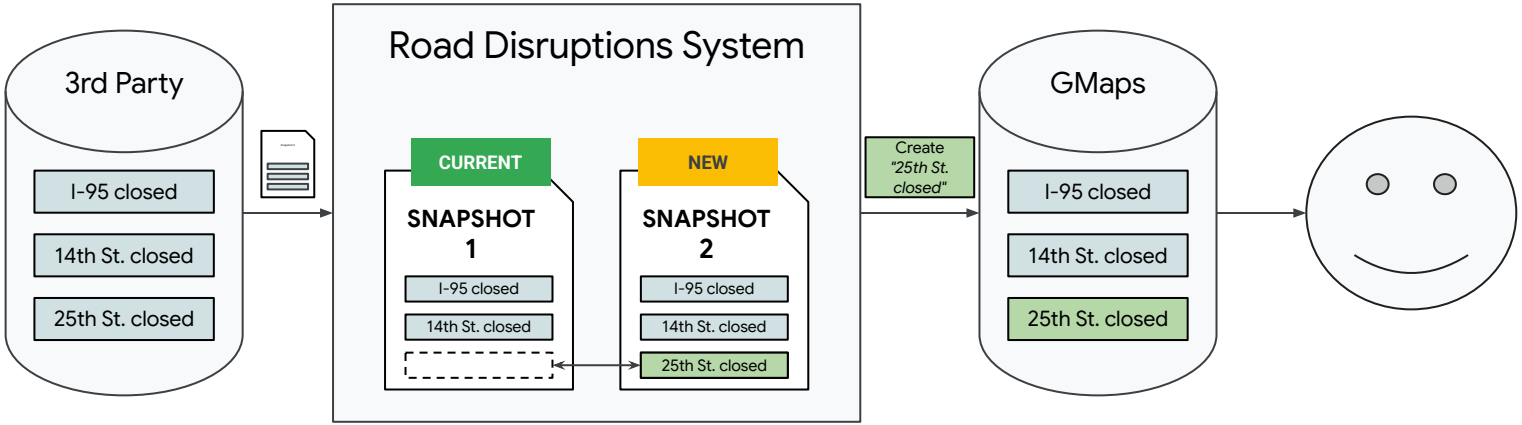
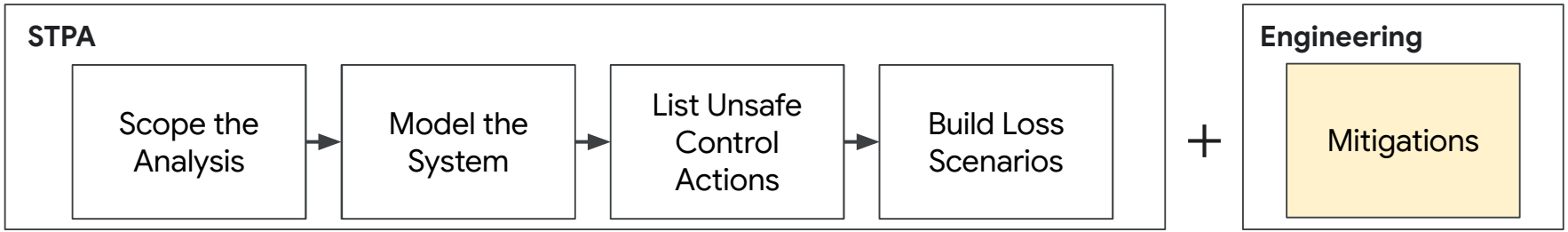

3rd Party
I-95 closed
14th St. closed
25th St. closed
Road Disruptions System
CURRENT
SNAPSHOT 1
I-95 closed
14th St. closed
NEW
SNAPSHOT 2
I-95 closed
14th St. closed
25th St. closed
Create "25th St. closed"
GMaps
I-95 closed
14th St. closed
25th St. closed

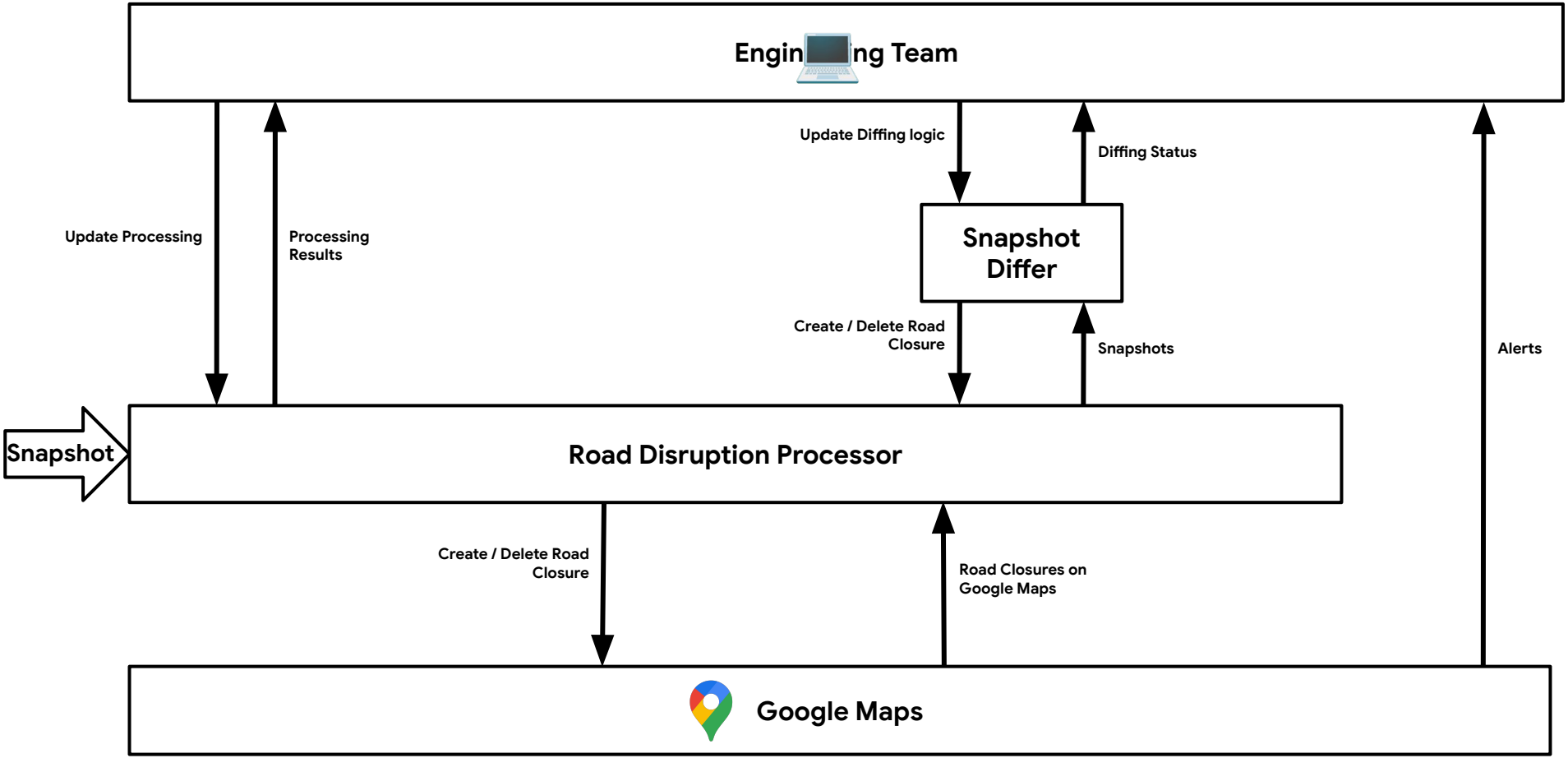
*Image created with Gemini 2.5 Flash

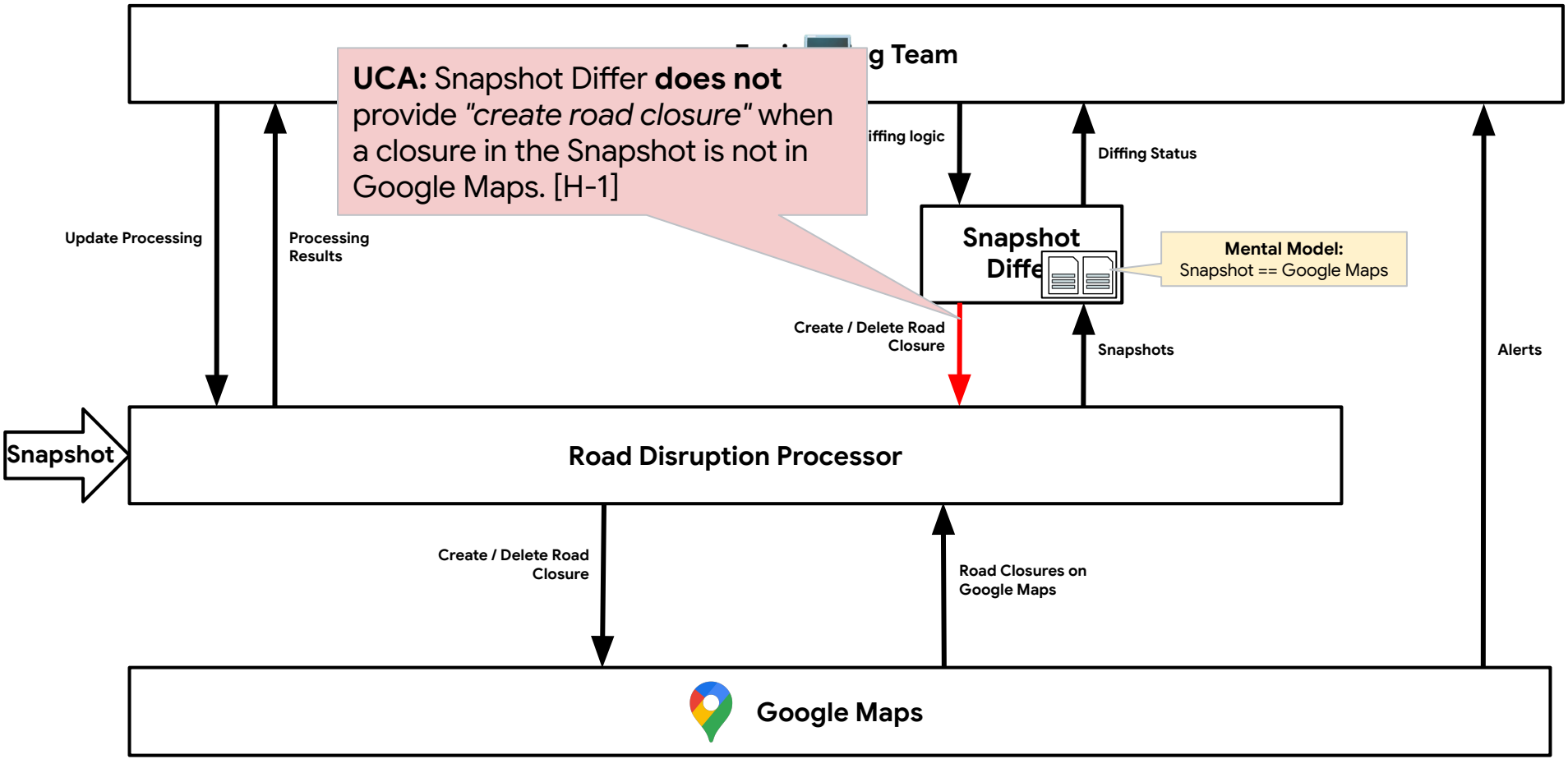
STPA Overview



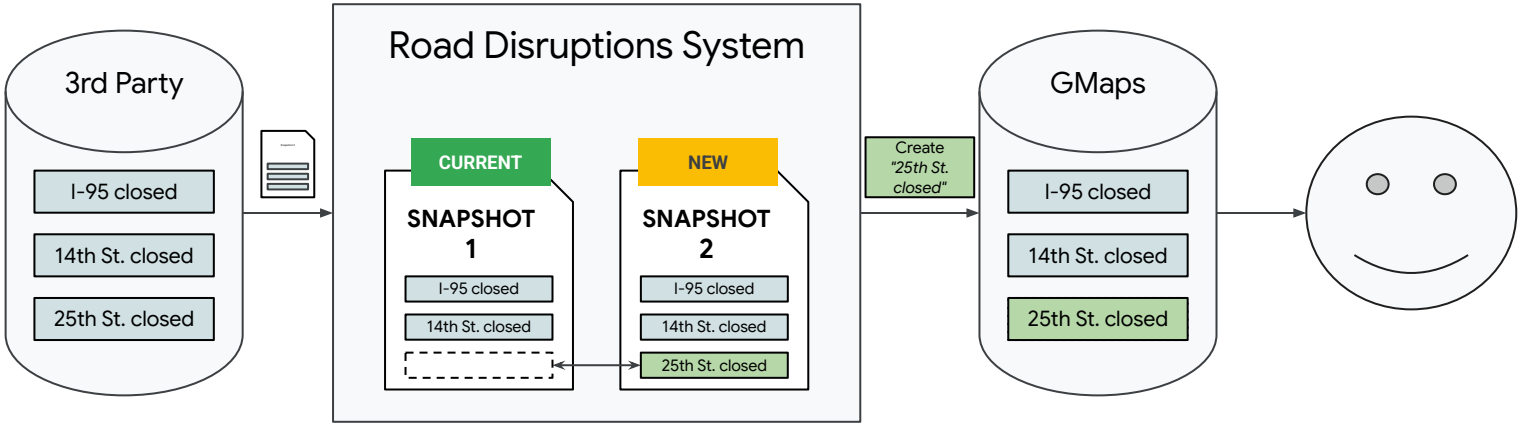
Road Disruptions System

Goal	• Ensure that Google Maps contains the latest state of all 3rd Party Closures
Losses	• L-1: Loss of User Trust • L-2: Loss of Mission • L-3: Negative PR Events
Hazards	• H-1: Google Maps is out of sync with 3rd Party Closures. [L-1, L-2, L-3]

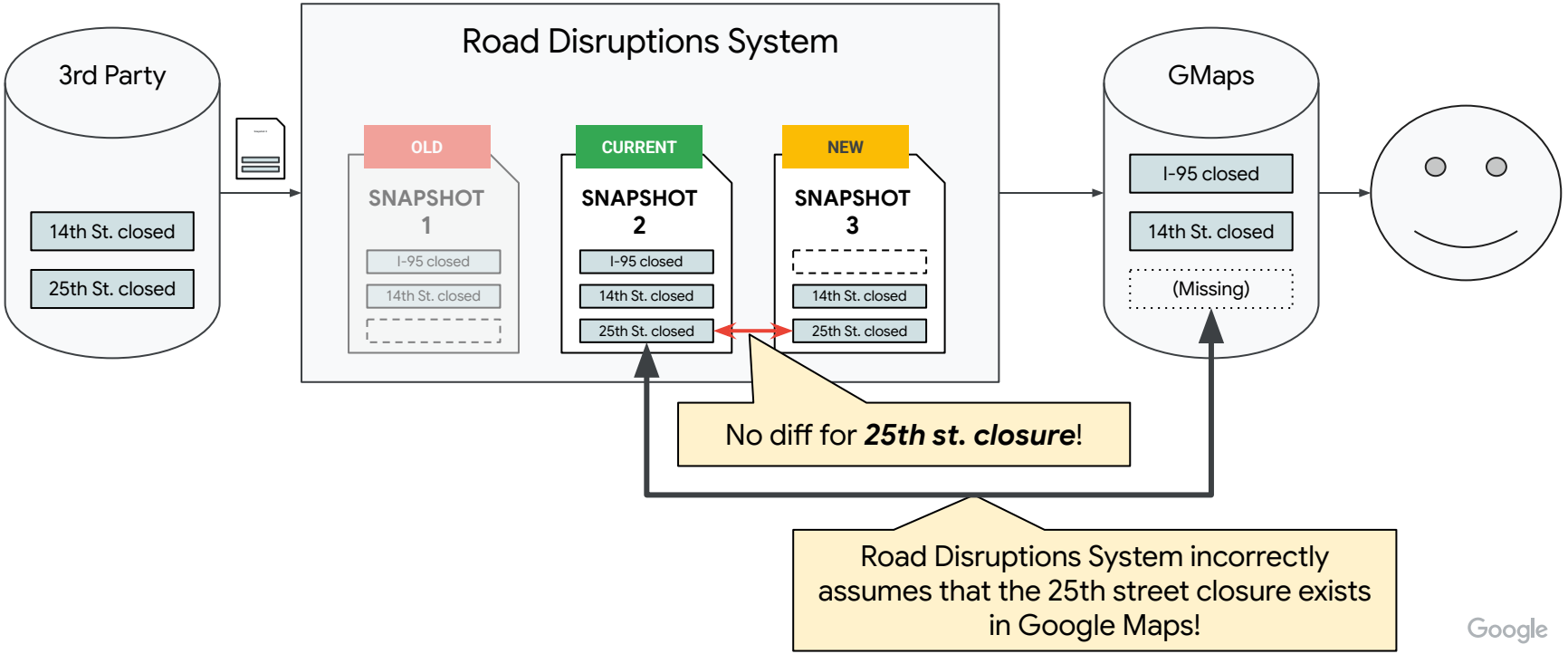

Engin ng Team
Update Diffing logic
Diffing Status
Update Processing
Processing Results
Snapshot Differ
Create / Delete Road Closure
Snapshots
Alerts
Snapshot
Road Disruption Processor
Create / Delete Road Closure
Road Closures on Google Maps
Google Maps


g Team
UCA: Snapshot Differ does not provide "create road closure" when a closure in the Snapshot is not in Google Maps. [H-1]
iffing logic
Diffing Status
Update Processing
Processing Results
Snapshot Diffe
Mental Model:
Snapshot == Google Maps
Create / Delete Road Closure
Snapshots
Alerts
Snapshot
Road Disruption Processor
Create / Delete Road Closure
Road Closures on Google Maps
Google Maps

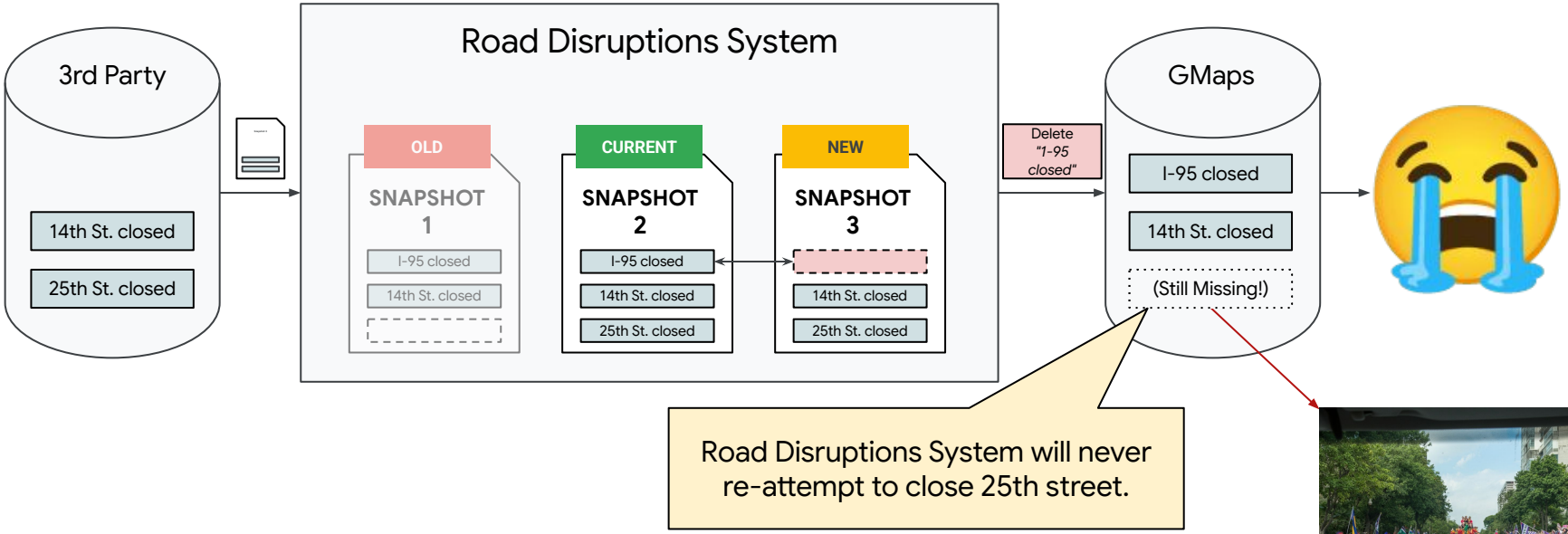
***UCA:** Snapshot Differ **does not** provide "create road closure" when a closure in the Snapshot is not in Google Maps.*



***UCA:** Snapshot Differ **does not** provide "create road closure" when a closure in the Snapshot is not in Google Maps.*



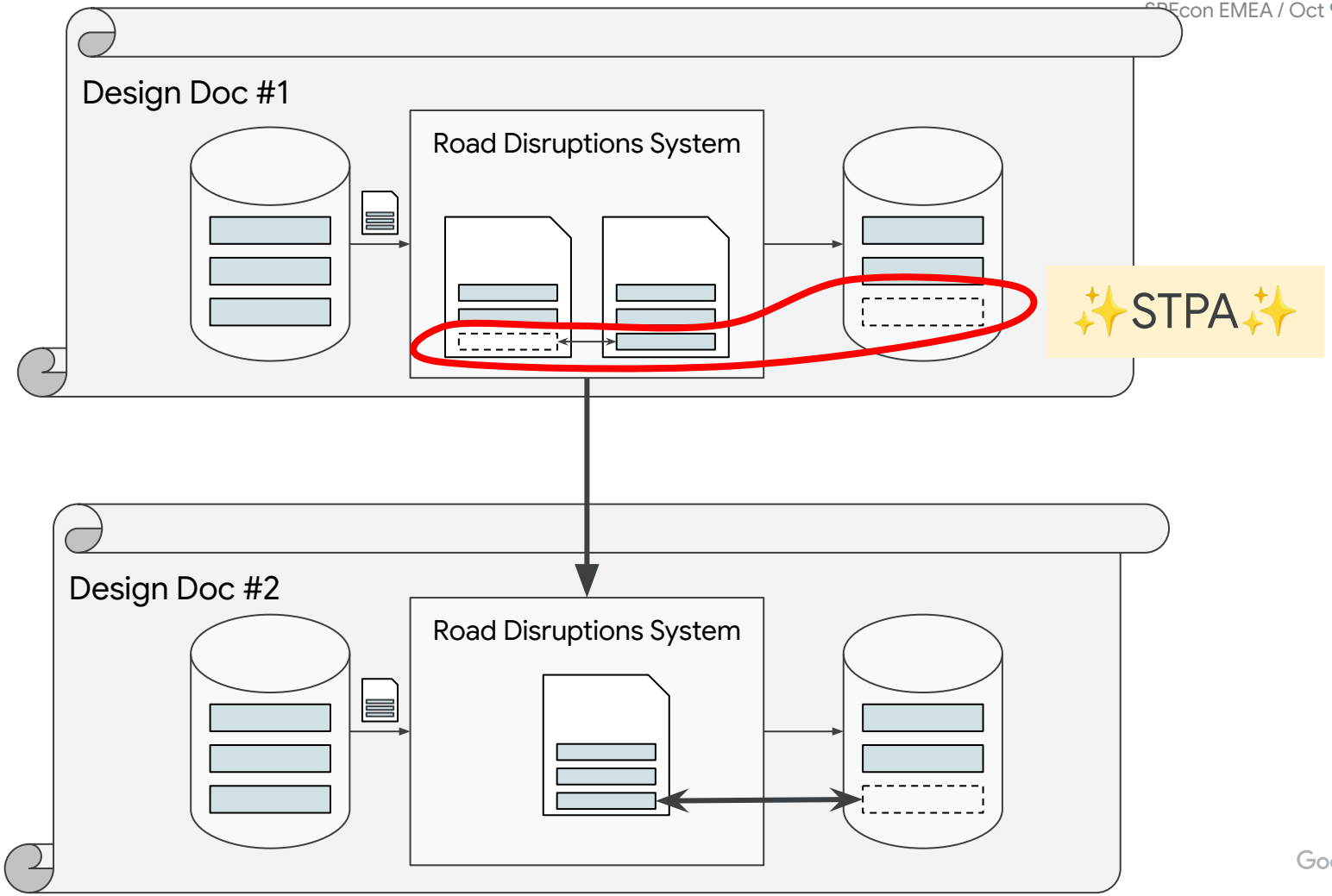
***UCA:** Snapshot Differ **does not** provide "create road closure" when a closure in the Snapshot is not in Google Maps.*

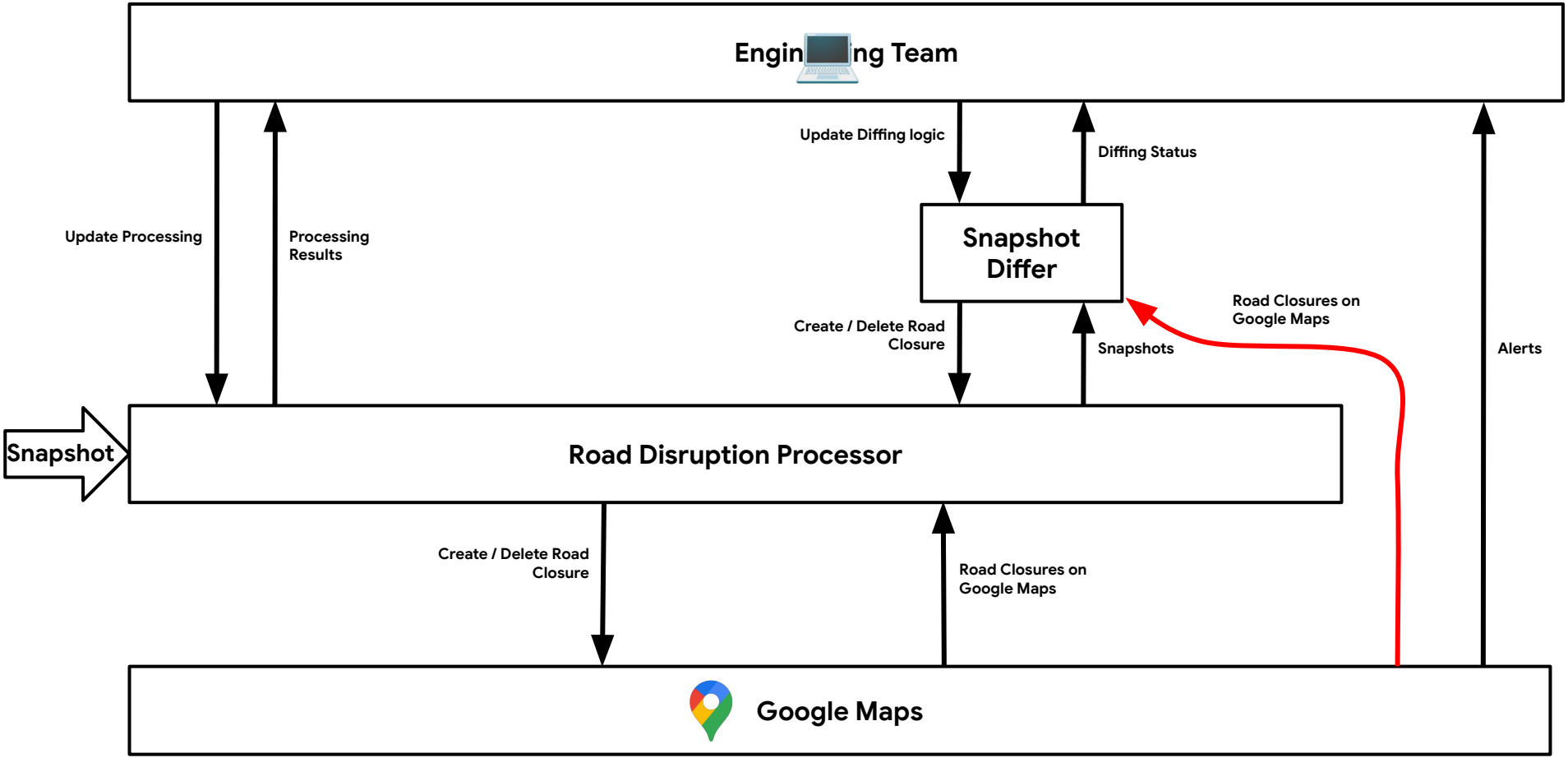


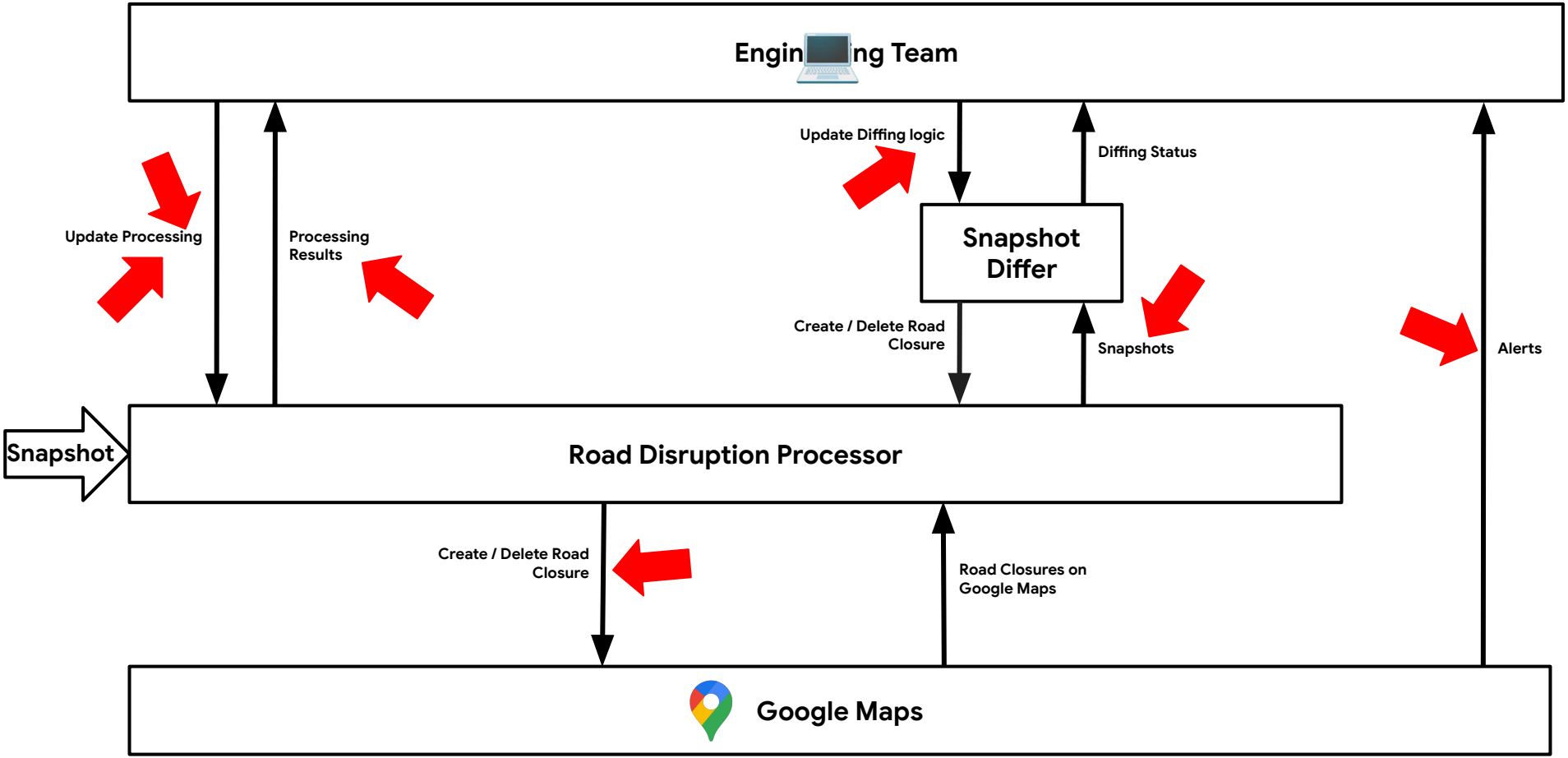
No component failed.

Every component operated as designed.

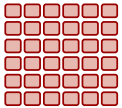
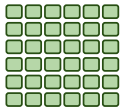
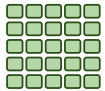
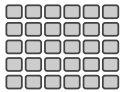
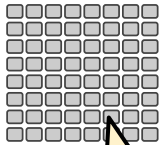
Problem: The system has a key design flaw.


Design Doc #1
Road Disruptions System
✨STPA✨
Design Doc #2
Road Disruptions System


Engin ng Team
Update Diffing logic
Diffing Status
Update Processing
Processing Results
Snapshot Differ
Road Closures on Google Maps
Create / Delete Road Closure
Snapshots
Alerts
Snapshot
Road Disruption Processor
Create / Delete Road Closure
Road Closures on Google Maps
Google Maps

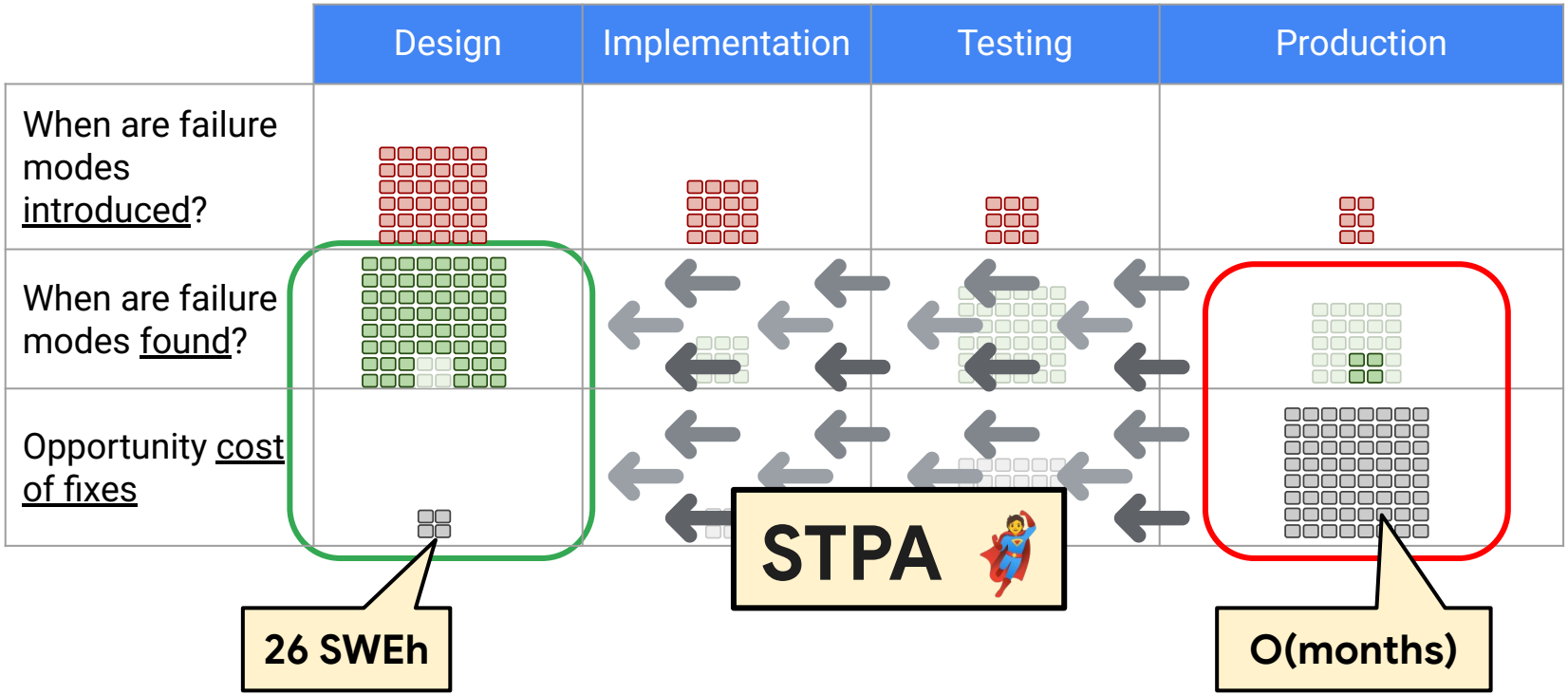

Engin ng Team
Update Diffing logic
Diffing Status
Update Processing
Processing Results
Snapshot Differ
Create / Delete Road Closure
Snapshots
Alerts
Snapshot
Road Disruption Processor
Create / Delete Road Closure
Road Closures on Google Maps
Google Maps

Cost* of Defects for Road Disruptions

	Design	Implementation	Testing	Production
When are failure modes introduced?				
When are failure modes found?				
Opportunity cost of fixes				

* Coarse characterization, adapted for Google from System Safety and STPA Class Materials, John Thomas, 2021
Data from “ROI Analysis of the System Architecture Virtual Integration Initiative”, SEI, 2018

Cost* of Defects for Road Disruptions



* Coarse characterization, adapted for Google from System Safety and STPA Class Materials, John Thomas, 2021
Data from "ROI Analysis of the System Architecture Virtual Integration Initiative", SEI, 2018

Conclusion

Key Takeaways

1. STPA finds and fixes design problems at a fraction of the cost
2. STPA enables a systematic exploration of unsafe system behaviors
3. YOU can learn STPA (with some support)

Learn more about STPA

- Google SRE Site:
  - STPA (System Theoretic Process Analysis) at Google
- Other Resources:
  - MIT Partnership for Systems Approaches to Safety and Security (PSASS)
  - STPA Handbook
  - MIT STAMP Workshop Tutorials

STPA for Software Systems–Illuminate the Unknown Unknowns

Theo Klein - Staff SRE
Garrett Holthaus - SRE Technical Writer
Ruben Barroso - Staff SRE

Google