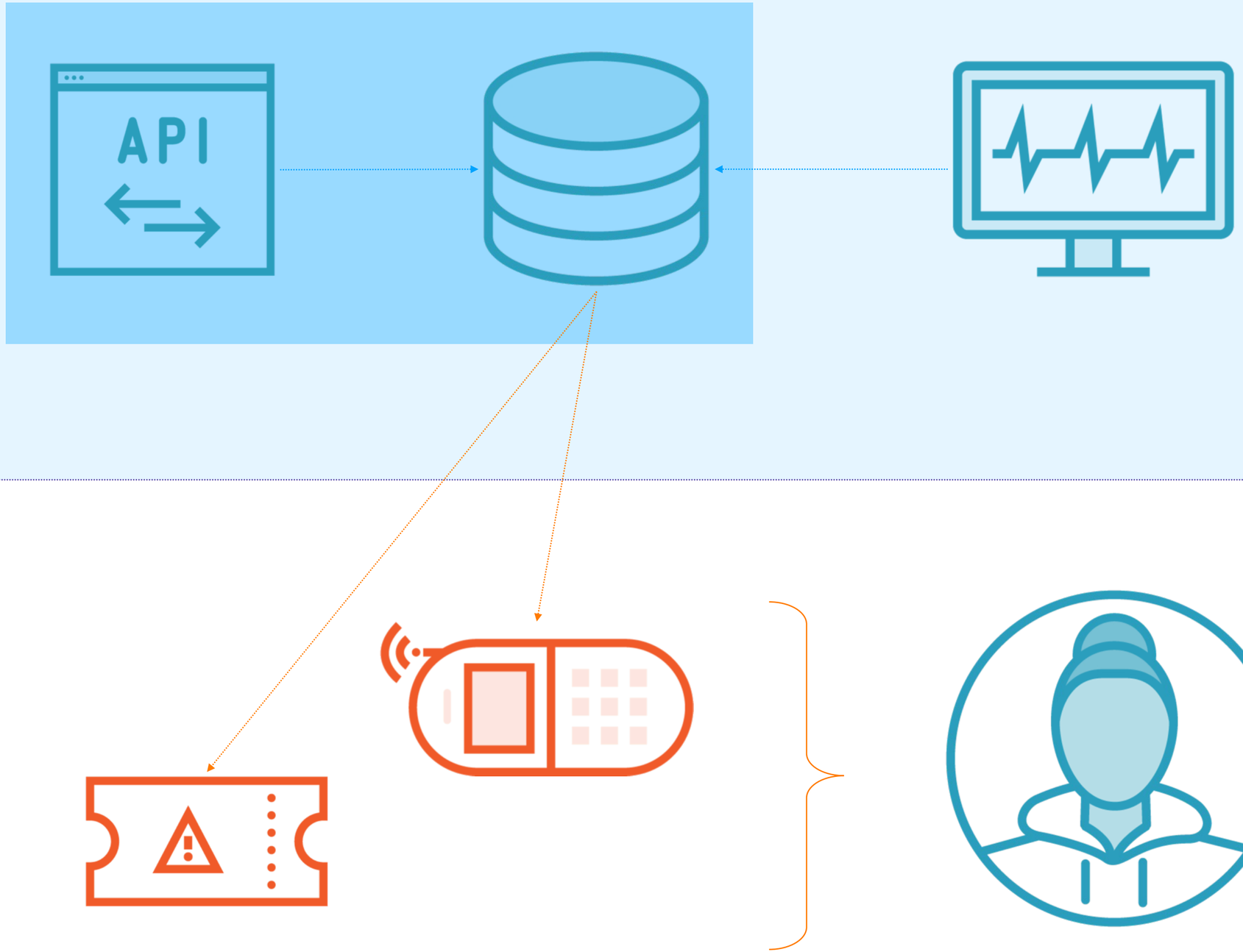
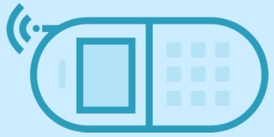
Incident Management: On-call and Postmortems

Elton Stoneman
Freelance Consultant and Trainer

@EltonStoneman | blog.sixeyed.com


API
Monitoring
Alerting

1 2 3 4 5 6



On-call

- Alerted on SLOs
- Acknowledge alert
- Begin investigation

Timely response

- Business critical - within 5 minutes
- Urgent - within 30 minutes

Major commitment

- No social engagements
- Sleep deprivation
- Stress

Control, Co-ordinate & Communicate

Communications Lead (CL)

Incident Commander (IC)

Ops Lead (OL)

Is it an Incident?

Standard Response

Playbook

No incident

SRE Discretion

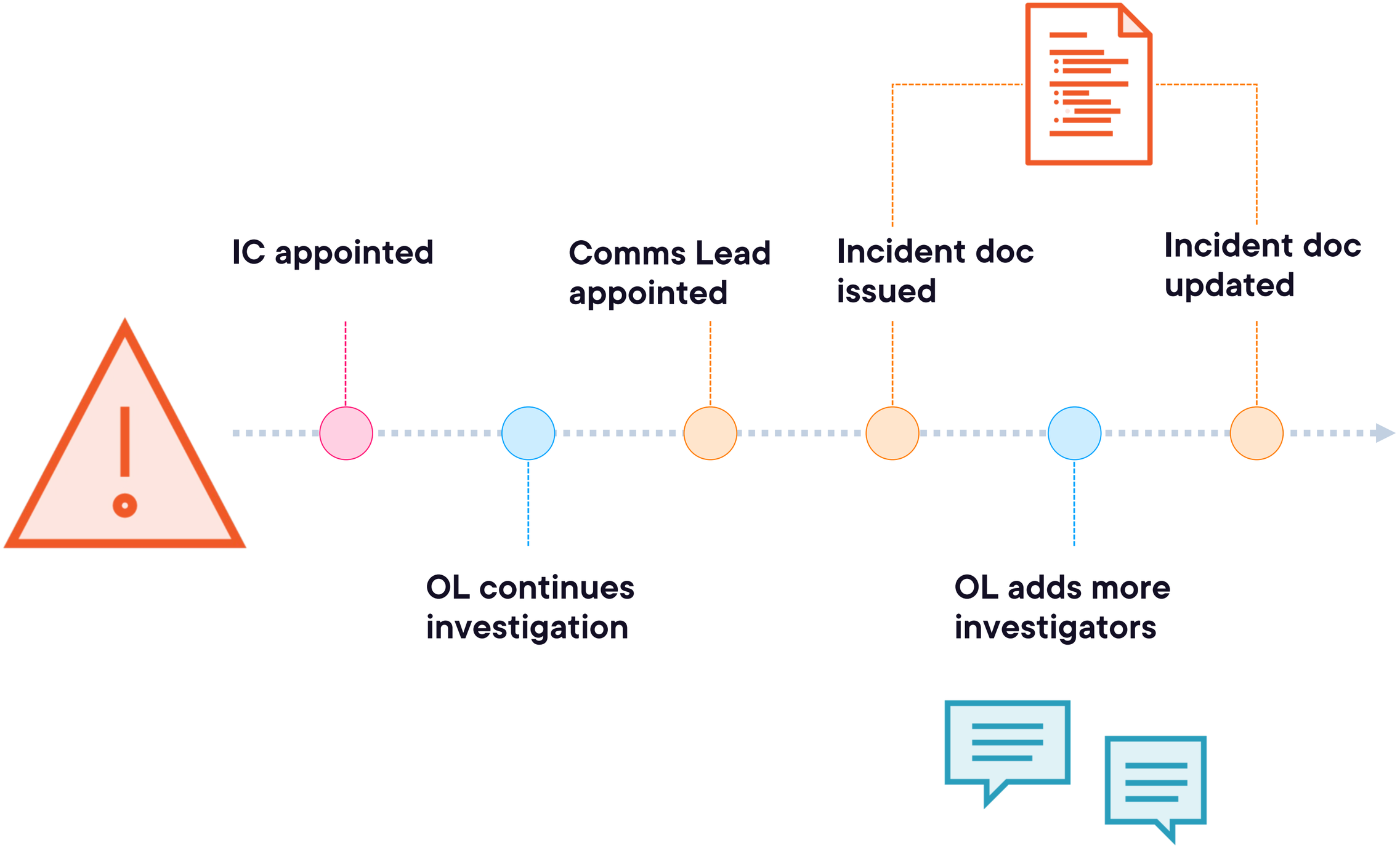
Past experience

Issue complexity

Set Criteria

Investigation time

Impact


IC appointed
Comms Lead appointed
Incident doc issued
Incident doc updated
OL continues investigation
OL adds more investigators

Comms Lead (CL)

Incident Commander (IC)

Ops Lead (OL)

Comms Lead (CL)

Incident Commander (IC)

I confirm I am now the IC

Ops Lead (OL)

Comms Lead (CL)

Incident Commander (IC)

Ops Lead (OL)

SRE

SRE

SRE

Incident Document

Postmortem

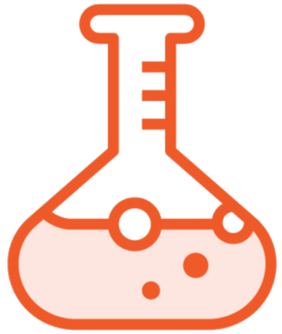
Working on Incidents Effectively

Incident Model

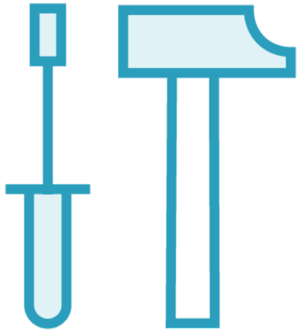
Triage

Examine

Diagnose



Test



Cure

Incident Model

- Automated alert
- SLO breach
- Metric details

- Problem report
- Manually recorded
- Expected, actual & repro

Triage

- Get back to "good enough"
- ASAP

Remediation tactics

- Add compute power
- Re-route traffic
- Downgrade service

Output

- Stable system

Examine

- Understand the problem
- Identify the trigger

Investigation tools

- Metrics and dashboards
- Centralized logging
- Service graphs
- Distributed Tracing

Service Mesh Architecture

Managing Apps on Kubernetes with Istio

Elton Stoneman

Examine

- Understand the problem
- Identify the trigger

Investigation tools

- Metrics and dashboards
- Centralized logging
- Service graphs
- Distributed Tracing

Output

- Know the problem and trigger

Diagnose

- Find the possible cause

Analysis tools

- Vertical path through system
- *What* is it doing?
- *Why* isn't it doing what it should?
- *Where* are resources going?
- *When* did it start?

Output

- Shortlist of potential causes

Test

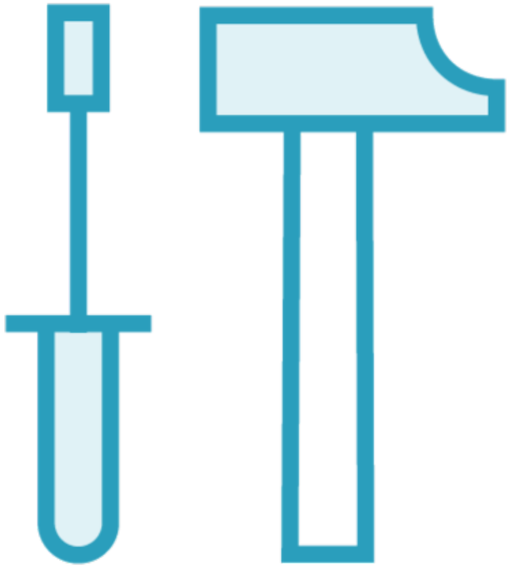
- Identify the probable cause

Testing tactics

- Manually recreate the workflow
- cURL Web apps and APIs
- Connect and verify permissions
- Trace database calls & network routes
- Document everything

Output

- Confidence in the actual cause



Cure

- Fix the problem
- Document the solution

Output

- Fully working system OR
- Mitigated system with known fix OR
- Mitigation with monitoring requirements

Producing and Publishing Postmortems

Postmortem goals

- Document the incident & resolution
- Identify root cause & fix

Formal documentation

- Drafted by SREs
- Review & publish process

Continuous improvement

- Blame-free
- Neutral & constructive

Date:
Authors:
Reviewers:
Incident Commander:

Executive Summary

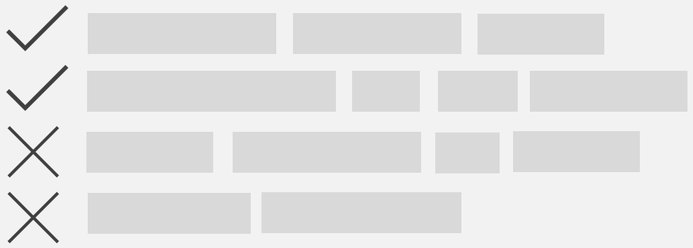
Problem Summary

Action Items

-
-
-

Timeline

Lessons Learned



Date:
Authors:
Reviewers:
Incident Commander:

Executive Summary

Problem Summary

Action Items

-
-
-

Lessons Learned

Timeline

Postmortems are expensive

- Writing & reviewing time

Promote use

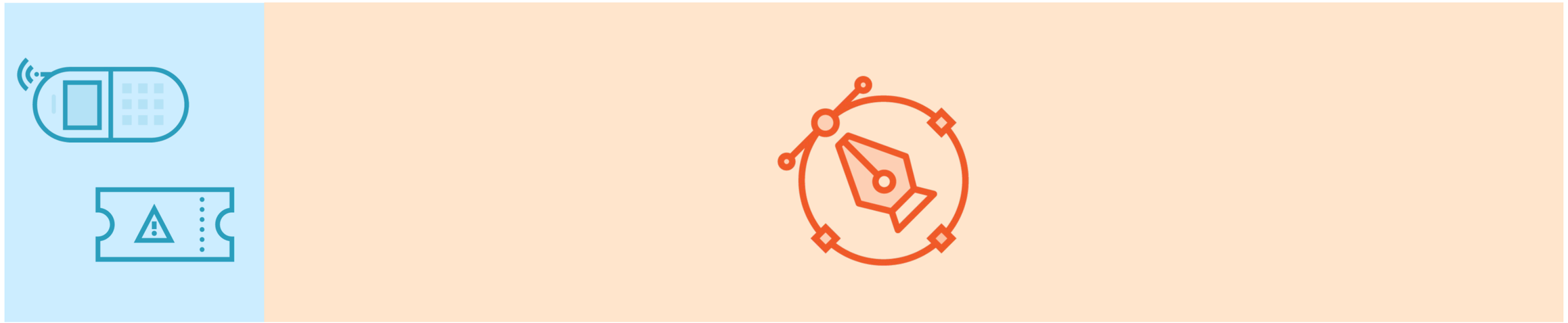
- "Postmortem of the month"
- Reading clubs
- Group review sessions

For major incidents

- Customer-facing
- Data issues
- Multiple impacts
- Lengthy resolution

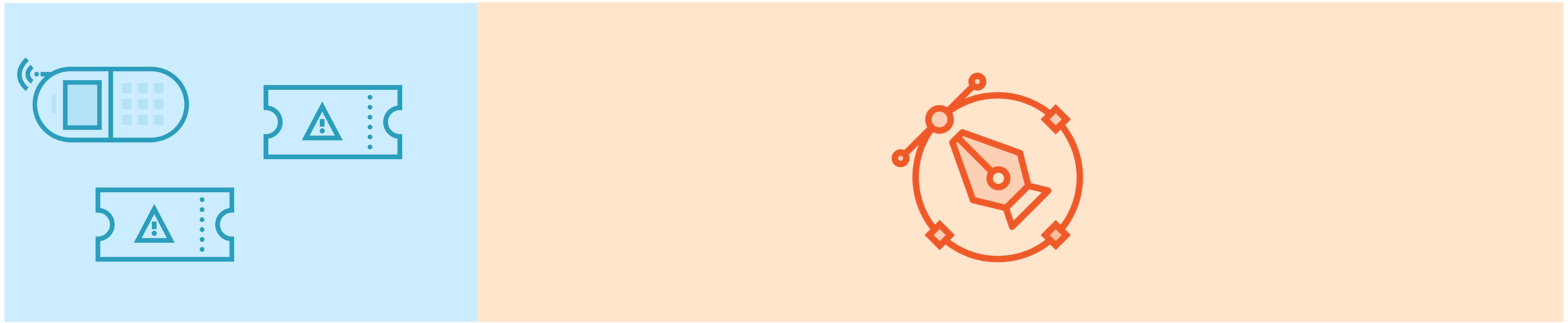
Avoiding Operational Overload

- No project work
- Pages and tickets
- On-call only



Primary

Secondary
9-5



Primary

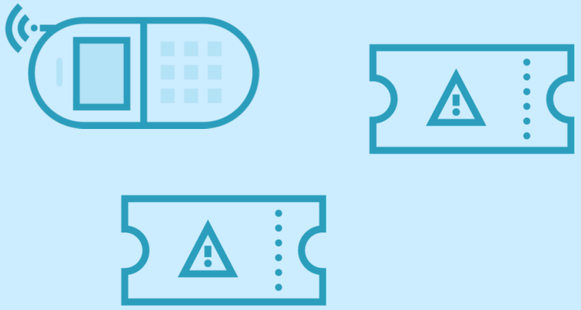
Secondary
9-5

Tickets
9-5

Primary 1
06:00-17:59

Primary 2
18:00-05:59

Secondary
9-5



Incident resolution

- Assume 6 hours
- Two per on-call shift

No pages or tickets?

- Documentation, tidy-up work
- Easily dropped

Consistently few incidents

- Does the project need SRE?
- Move to product team management

Primary

Secondary
9-5

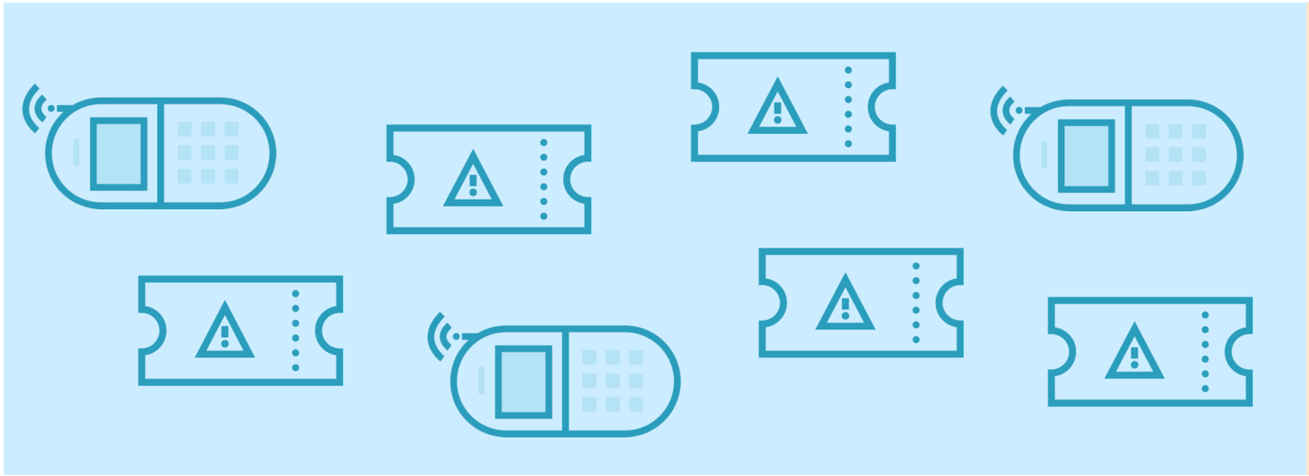
Tickets
9-5

Primary

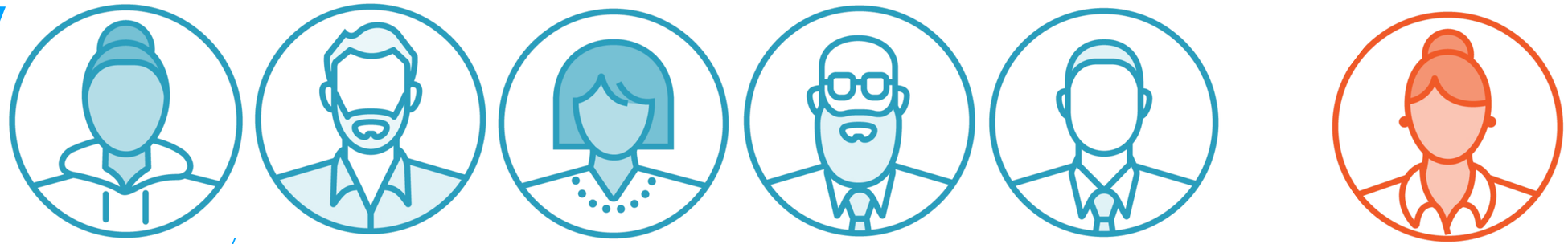
Secondary
9-5

Tickets
9-5

Tickets
9-5



Primary

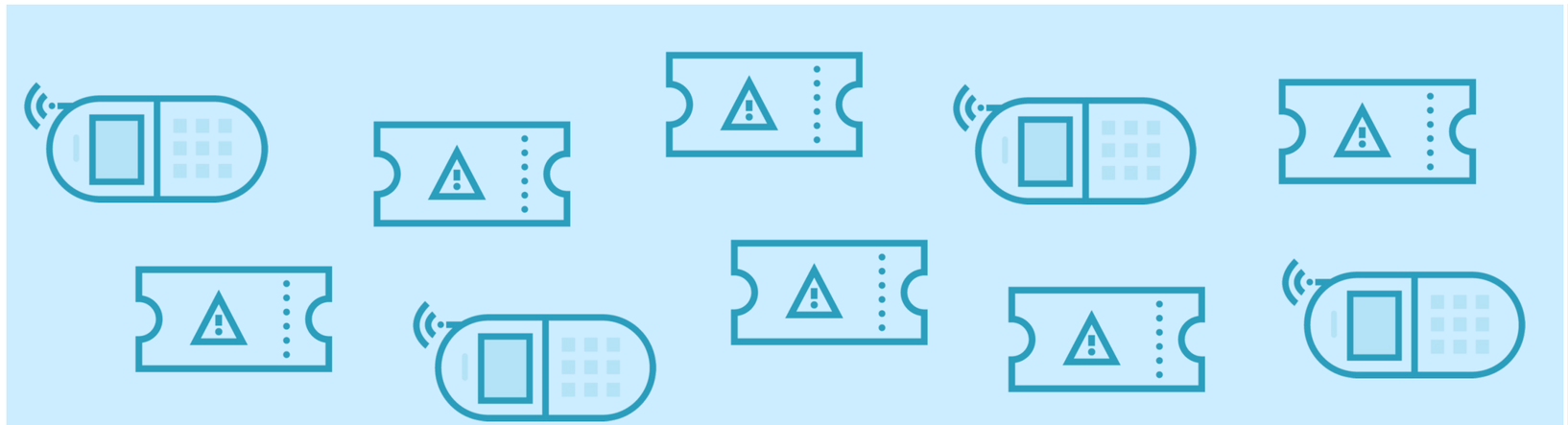


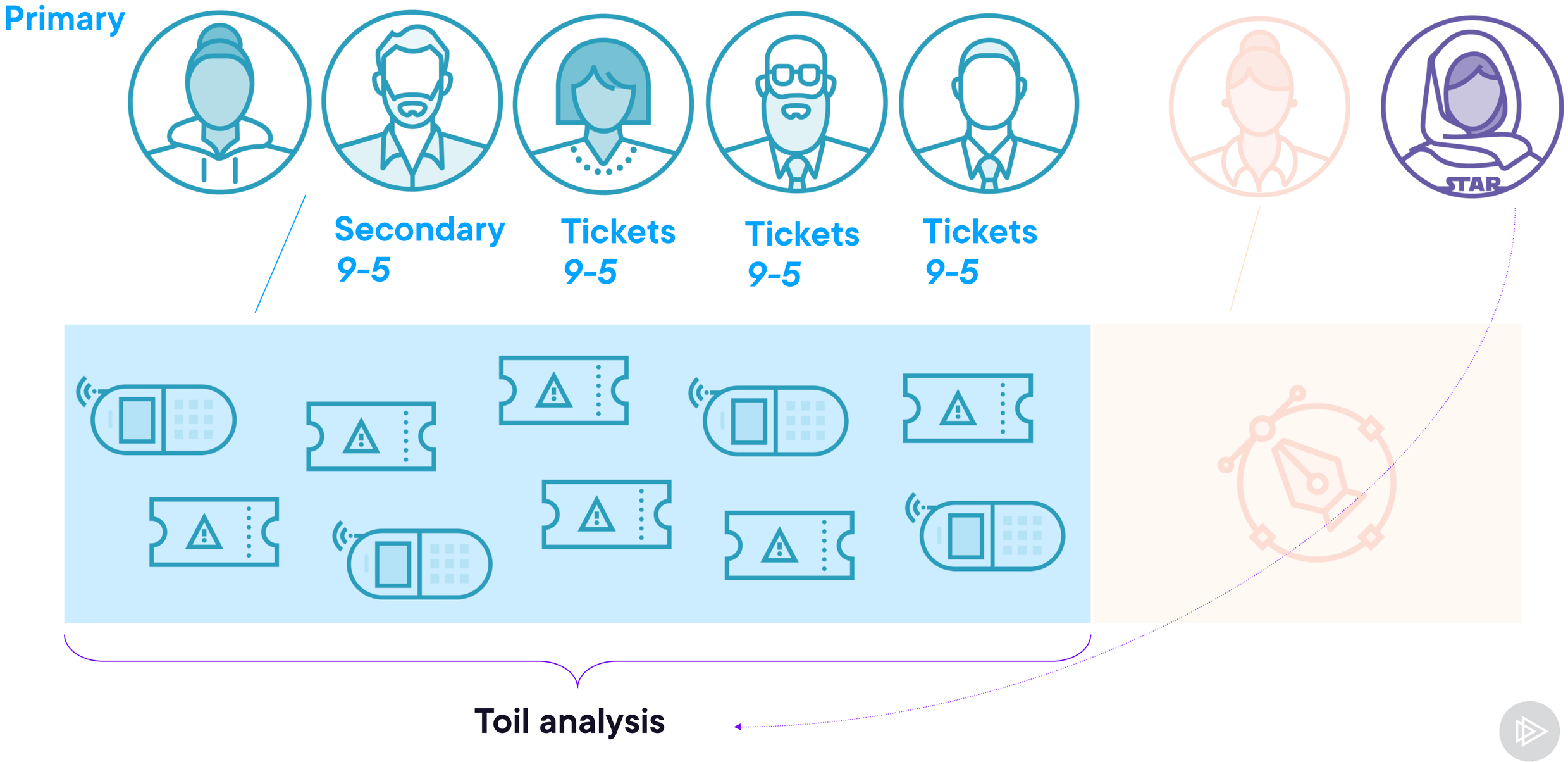
Secondary
9-5

Tickets
9-5

Tickets
9-5

Tickets
9-5




Primary
Secondary
9-5
Tickets
9-5
Tickets
9-5
Tickets
9-5
STAR
Toil analysis

Ops mode!

- Stick to SRE principles
- Reduce SLOs
- Change freeze

SRE is a balance

- Respect the time of the team
- Respect customer expectations

Summary

Incident management

- Incident Commander
- Ops Lead focuses on investigation
- Comms Lead informs stakeholders

Investigation model

- Triage, examine, diagnose, test, cure
- Tactics and output

Postmortems

- Blameless analysis
- Aiming for continuous improvement

Operational overload

- Managing on-call and workloads

More in SRE

Guidance

- Production readiness reviews
- Capacity planning
- Designing for simplicity

Practical advice

- Load balancing
- Cascading failures
- Testing for reliability

Follow the learning path!

- Here on Pluralsight

We're Done!

So...

- Please leave a rating
- Follow `@EltonStoneman` on X
- Check out `blog.sixeyed.com`
- Watch my other courses ☺