Xebia

Site Reliability Engineering Fundamentals (2/2)

MAERSK

What did we discuss?

- Definition of SRE and the SRE principles
- SRE in relation to DevOps
- SLI, SLO, SLA
- Error budgets
- Monitoring distributed systems

Agenda

Part 1

- Introduction to SRE
- The principles of SRE
- SLIs, SLOs, SLAs

Part 2

- Error Budgets
- Monitoring distributed systems
- Toil

Part 3

- **Incident Management**
- Postmortems

Part 4

- SRE Culture
- Onboarding SRE in an Organization
- Ask me anything & Reflection

EXERCISE

Please go to the Miro board.

Take 15 mins to read the incident report and write down on the board what went well, what went wrong and what can be improved

Place stickies on the board and prepare a summary for in the main room.

Discuss the result with the group.

Frequently seen problems during incidents

Sharp focus on the technical problem

- Your instinct is to find the root cause ASAP – ignore that…
  - We jump on the technical problem but forget to think about the bigger picture
  - The technical problem can be overwhelming
- Focus on what the customer needs!

Poor communication & documentation

- Communication was forgotten
  - focus was on the technical problem
- No one knew from each other what they were doing
- Incident description and solution is vague
- No documentation of what has been done/still needs to be done
- No single place of documentation
  - wikis, docs, pages, IM systems, etc.

Freelancing

- People want to help!
  - With all good intentions a colleague started to help and push changes to production
- There is no coordination or communication
- Risk of undoing changes, breaking code, introducing bugs, etc.

On-Call

Life of an on-call engineer

- Guardians of production systems
- Available act within minutes
  - Paging respond times depends on service/team/incident
  - Necessary respond times are related to desired service availability
  - Sometimes, non-paging production events can be handled by on-call engineer

Being on-call

- A critical duty
- There are several pitfalls that can lead to serious consequences
  - You are working when you just wake up
  - You are in a hurry to fix the problem
  - You don’t have any colleagues to help you
  - You don’t take enough time to test before going back to bed
- Though job, working under time pressure

Balanced on-call

- SRE strives for balance between quantity and quality of on-call shifts
- The “E” in SRE!
  - 50% = engineering
  - 25% = on-call duties
  - 25% = operational overload
- Sufficient time to deal with incidents & follow-up
  - post-mortems

EXERCISE

Please go to the Miro board.

Take 10 mins to add stickies for things you should do and shouldn't do during an incident while you are on-call.

Place stickies on the board and prepare a summary for in the main room.

Discuss the result with the group.

Feeling safe while on-call

Reduce the amount of stress!

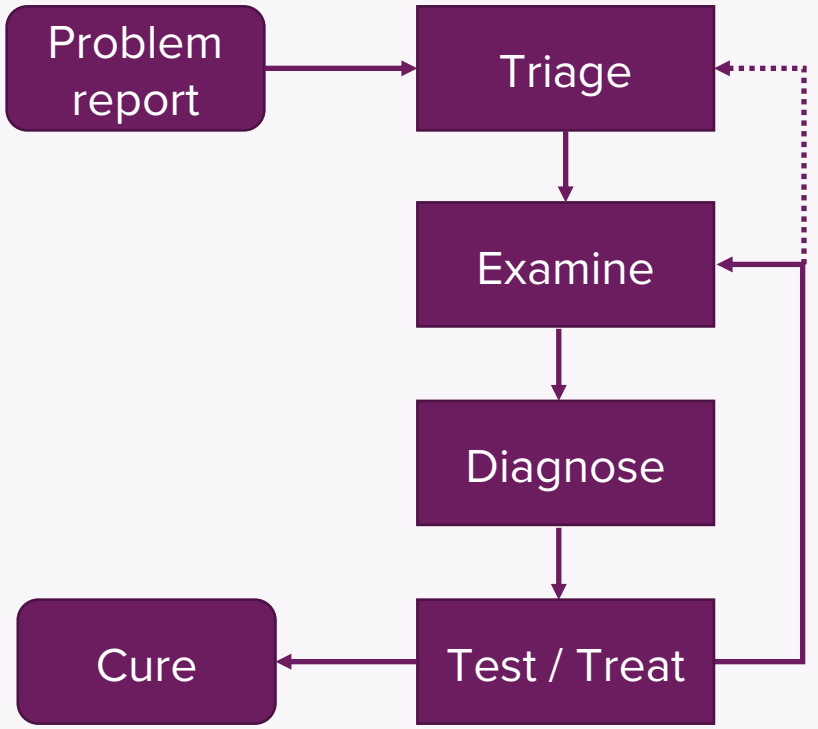
- Service outages create significant pressure
  - Stress hormones like cortisol are known to cause behavioral consequences – including fear and fast decision making
- Have fall back on resources like:
  - Clear escalation paths
  - Well-defined incident procedures
  - A blameless postmortem culture

Incident Management Model

Effective troubleshooting

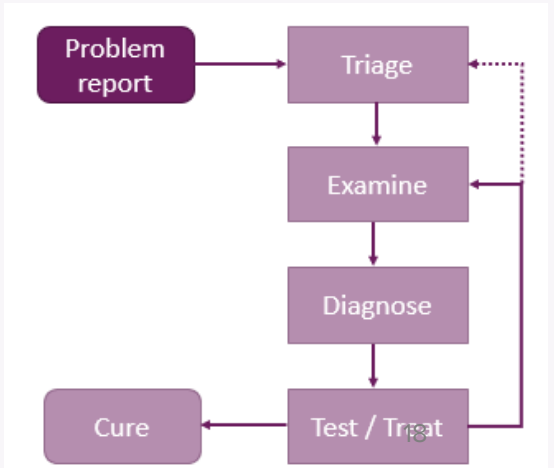
SRE model for Troubleshooting →

- Troubleshooting is a critical skill for every SRE
- It's a skill that is both learnable and teachable
- Effective troubleshooting depends on two factors
  - Know how to troubleshoot generically
  - A solid knowledge of the system you are troubleshooting



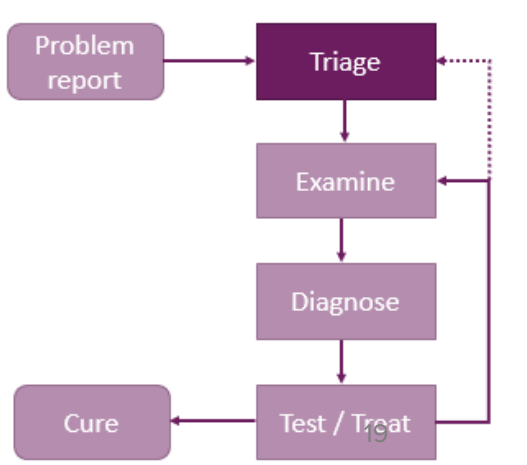
Problem report

- Start incident with problem report
  - Reports are alerts or a raised problem (incident report)
- An effective report contains
  - Expected behavior of the system functioning normal
  - Actual behavior show at the time of incident
  - How to reproduce the unwanted behavior



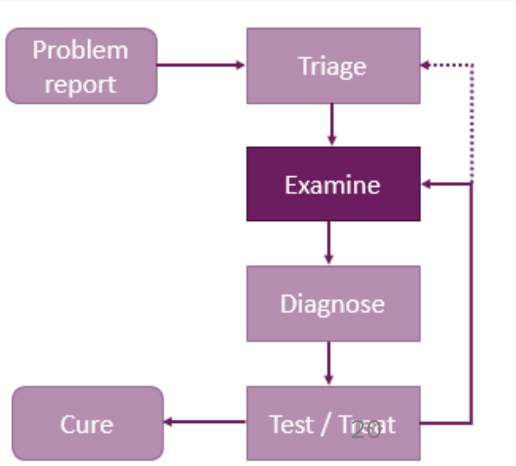
Triage

- Determining what to do first!
- Ignore Instinct to find root cause ASAP
- Priority is to:
  - Make the system work, as well as possible, under the current circumstances for the CUSTOMER!



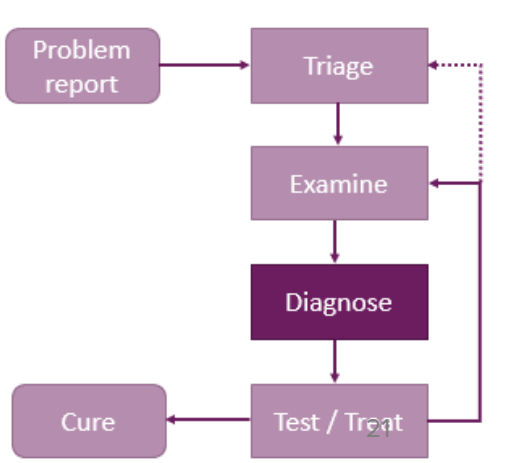
Examine

- Examine every component in the system – Is it behaving correctly?
- Start at your monitoring systems
  - Metrics
  - Logging
  - Alerts
  - Error messages
  - Dashboards
  - Emails
  - Incident management systems



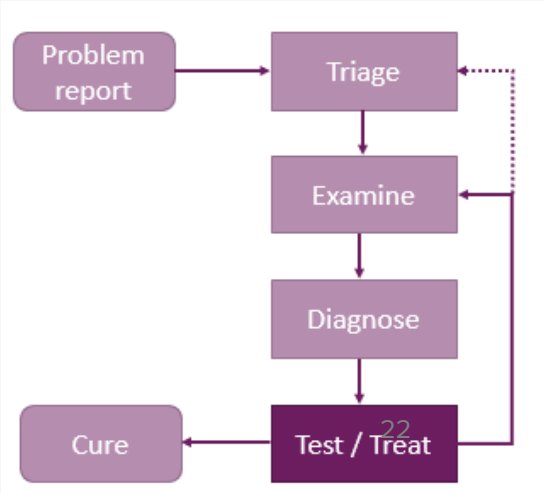
Diagnose

- Ask ”what”, “why” and “where”
  - A malfunctioning system is often still trying to do something!
- Find out what touched the system last
  - Updates
  - Certificates
  - Users
  - Incident fixes
  - Patches



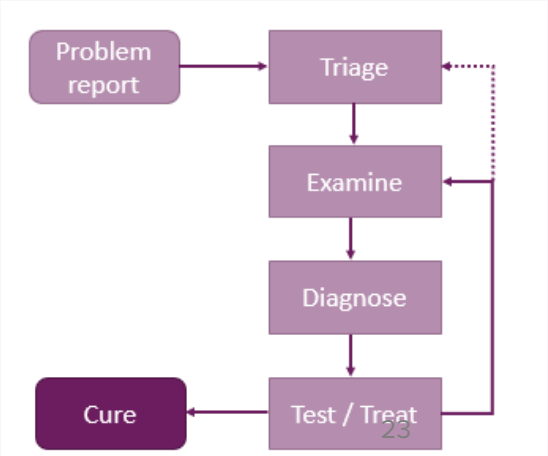
Test and Treat

- List possible causes
  - Narrow it down to the root cause by exclusion
- Experiment to rule out hypotheses
- Considerations when designing tests:
  - Consider the obvious first
  - Side effects altering future results
  - Risk for misleading results



Cure

- Ideally narrow it down to one root cause
- After curing the root cause:
  - What went wrong?
  - How was problem Tracked?
  - How was problem Fixed?
  - How to prevent it?
- Write a postmortem!



Common pitfalls when troubleshooting

- Looking at the data of un-relevant symptoms
- Misunderstanding of how to change the system, or how the system works
- Assuming that past causes of problems are the causes of the current problem
- “It happened before, so it must be happening again”

Emergency response

What to do when the system breaks?

Emergency response – What to do when the system breaks?

- Remember when an emergency happens
- You are not alone
- You are a professional
- The sky isn’t falling
- If you feel overwhelmed, pull in more people!

Be proactive with emergency - testing

Let SREs break the systems (Chaos Engineering)

1. Watch how they fail
2. Make changes to improve reliability
3. Prevent the failure from recurring

Often, these controlled failures go as planned

Keep a history of outages

- Share knowledge by documenting everything
- Publishing postmortems
- Ask “what if?” often

Incident management – Command System

- An incident management system based on the ’Command System’
- Everyone knows their roles
- Within the Command System a clear separation of responsibilities allow more autonomy for all parties

Which roles are there in the Command System

Incident Commander

- Incident commander tracks high-level incident state
- Assign responsibilities - need & priority
- Interested parties know where to interact with incident commander (“War Room”)
- Keeps living incident document

Operational Worker

- Works with incident commander
- Responds to the incident
  - Applying operational tools to the task at hand
- Operations team is the only group allowed to modifying system during incident!

Communicator

- Public face of the incident
- Gives periodic updates about incident
  - To incident response team and stakeholders

Planner

- Planning role supports operational team
  - Filling documenting and bugs
  - Ordering dinner for the team
  - Arranging handoffs between people
  - Keeps track of time
  - Planning review meetings

EXERCISE

Please go to the Miro board.

Take 10 min to discuss in sub-groups what you can change to improve on you Incident management.

Place stickies on the board and prepare a summary for in the main room.

Discuss the result with the group.

Best practices for incident management

- Prioritize – *Stop the bleeding*
- Prepare – *Develop and document your incident procedures in advance*
- Trust – *Give full autonomy*
- Introspect – *Pay attention to your emotional state*
- Consider alternatives – *Periodically consider your options*
- Practices – *Use the process routinely*
- Change it around – *Change roles on every incident*

Agenda

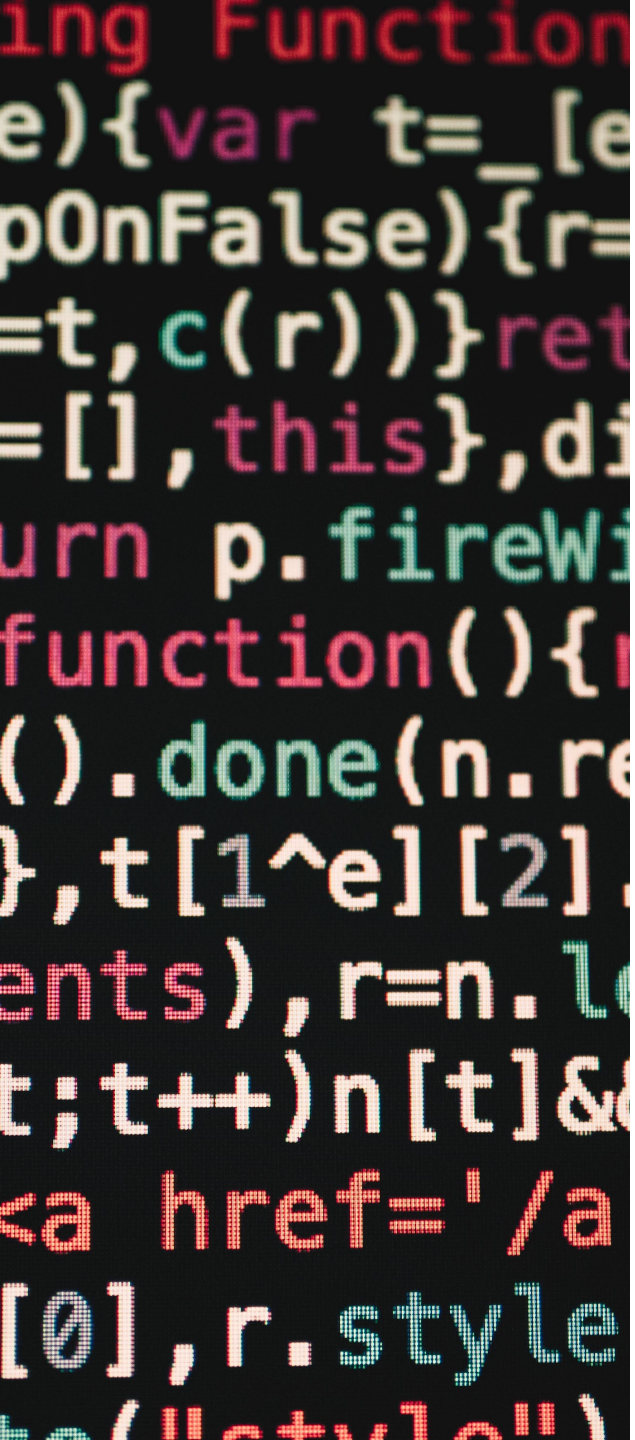
Part 1
- Introduction to SRE
- The principles of SRE
- SLIs, SLOs, SLAs

Part 2
- Error Budgets
- Monitoring distributed systems
- Toil

Part 3
- Incident Management
- **Postmortems**

Part 4
- SRE Culture
- Onboarding SRE in an Organization
- Ask me anything & Reflection

Post-mortems



“When systems become large-scale, complex and distributed, incidents and outages are inevitable. When they happen a good process should be in place to evaluate and have a good understanding of what happened, how it was resolved, how to improve the system and even how to prevent (or reduce the chance for) the incident or outage from occurring again.
This is where “post-mortems” come in.”

> *The cost of failure is education.*
>
> *Devin Carraway*

Post-mortems, more than just a simple incident document

- Post-mortems mostly are written incident documents
- Post-mortem processes & organization culture influences effectiveness of Post-Mortems
- Writing post-mortems leads to
  - More reliable systems
  - More learning
  - More experimenting

Post-mortems, the culture

The quality of post-mortems relies on the (SRE) culture

- Safety
  - Can everyone tell what really happened without facing consequences?
- fear
  - Can you make mistakes without the fear of losing our jobs?
- blaming
  - Do we really want to learn, or do we want to find out to whom to point our fingers to?

Post-mortems, Safety

- Embracing risk means:
  - Having a safe environment to be able to make mistakes
  - allow to make mistakes
  - failures are inevitable
  - learn from those mistakes
  - Be fully transparent (writing post-mortems)

Create a culture where failure is seen as a way to become smarter & to understand more about the systems and services

Post-mortems, Blaming

- Blameless post-mortems
  - no one should be blamed for any actions executed
  - analyse what happened and not who
  - Everyone should be able to ask “critical” questions
- Being critical in a professional way > quality post-mortem

> “Both the failures and successes don’t really rest with a single individual.”
>
> *Dave Zwieback*

Post-mortem practices

EXERCISE

Please go to the Miro board.

Take 10 min to discuss in sub-groups what leads to writing a good post-mortem.

Place stickies on the board and prepare a summary for in the main room.

Discuss the result with the group.

What leads to a good post-mortem?

- High-quality post-mortem increases Reliability of your systems
  - Have Incident management
  - Safe keep all important information (writing it down)
  - Keep track of timelines (re-construct)
  - Have a clear separation of duties
  - Awareness during incidents to collaborate as much as possible
  - Avoid Heroes/Freelancers!

A basic checklist – the 3 Rs

Remember the three R's when writing a postmortem!

Regret	Reason	Remedy
Acknowledge the problem	**Include all** the info from detection to recovery	**Remediations** are **SMART**
Empathy for: - **Affected** customers - **Involved** firefighters	The **more details** you can include the better	If remediation is unknown, at least include commitment for further research
	There is **no** point in **hiding** facts	Communicate and **publish**!

Post-mortems, content checklist

- Title and incident ID
- Date and authors
- Status of the PM
- Small summary
- Impact
- Root Causes
- Trigger
- Resolution
- Detection
- Action items
- Lessons learned
- What went well/wrong
- Timeline of events

Good practices during a post-mortem

- Remember the three R’s
- Have a good focus on the outage
- Timebox “root cause” analysis
- Ask questions!
- Avoid punishment!
- Post-mortem→ Set the context, build a timeline, determine, prioritize & publish!

Post-mortem processes

It’s important to have processes & agreements that guard effectiveness & quality of your post-mortem

- When is a post-mortem needed?
- How do we track the follow-up actions?
- How much time/resources?
- Where do we publish?
- What are the review-processes?
  - Quality standards
  - How can others be educated

How Google promotes post-mortems culture

- Sharing the post-mortem of the month
- Post-mortem community
  - Setting up standards
  - Sharing best practices
- Post-mortem book club
  - Reviewing one post-mortem
  - Discussing incidents and learnings with peers
  - Teaching new employees
- Post-mortem re-enactment

Agenda

Part 1

- Introduction to SRE
- The principles of SRE
- SLIs, SLOs, SLAs

Part 2

- Error Budgets
- Monitoring distributed systems
- Toil

Part 3

- Incident Management
- Postmortems

Part 4

- **SRE Culture**
- Onboarding SRE in an Organization
- Ask me anything & Reflection

EXERCISE

Please go to the Miro board.

Take 7 min to discuss in sub-groups what is culture.

Place stickies on the board and prepare a summary for in the main room.

Discuss the result with the group.

How to change a culture?

” When **culture change involves changing surface-level behavioral norms**, it can occur with relative ease because members can articulate what behaviors are required for success today in contrast to those required yesterday”

Ralph Kilmann, Mary Sfixton, Roy Serpa

“**How long a change in culture will last and how firmly the change is ingrained in the behavior** and decision-making processes in the organization also are related to the process of culture change. “

Ralph Kilmann, Mary Sfixton, Roy Serpa

What is Behavior?

"Behavior is what an organism is doing or more accurately what it is observed by another organism to be doing"

Skinner, 1938

You can observe behavior!

The ABC model

Example ABC-analysis

Antecedent

- The meeting subject is not interesting for me
- The device vibrates, lights up, makes a sound
- The device is in the line of sight
- Expecting a message/ email
- Other people look at their phone
- Curiosity
- Bored
- Want to know what time it is
- There is an escalation going on at the client
- Personal situation, want to be informed all the time
- Push message from a news- app

Behavior

“Looking at a mobile phone during a meeting”

Consequence

- I “escaped” the meeting for a second
- I am up to speed, my curiosity is satisfied
- I have missed something that was discussed and don’t know if this was important for me
- A colleague makes a remark on my behavior
- Read my mail, my message was received
- I know what time it is
- I can make a decision on an issue
- I’ve read something interesting

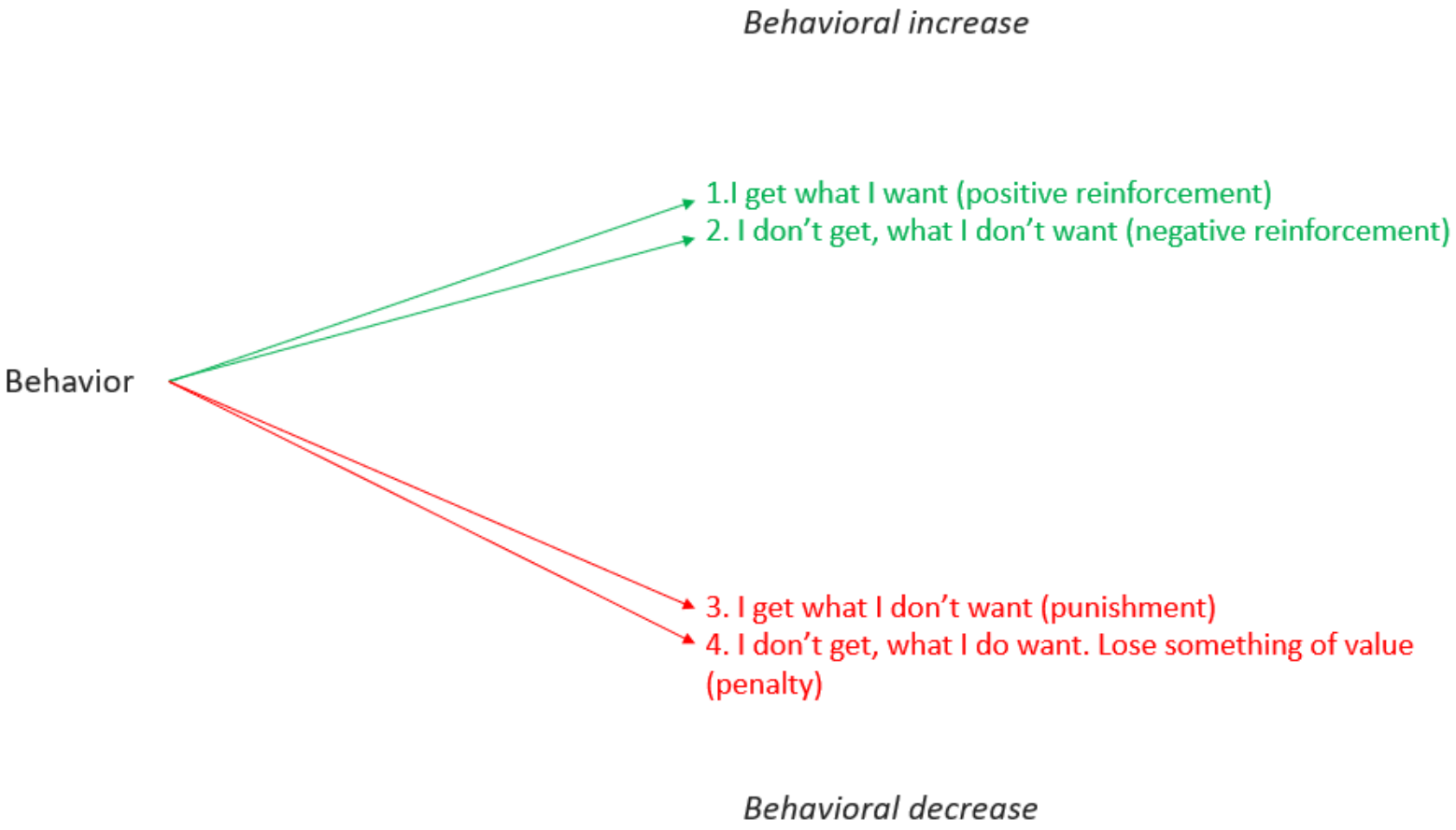
Antecedents (trigger)

- An antecedent is anything that prompts behavior and provides info about behavioral consequences
- Characteristics of Antecedents:
  - Sets the occasion for behavior
  - no causal relation A - B
- If effective:
  - valuable or meaningful consequence
  - reliable & accurate signal = > powerful antecedent

Consequences

- A consequence is anything that follows behavior and changes the probability that the behavior will re-occur in the future
- Characteristics of consequences:
  - Performer’s perspective
  - Positive, meaningful consequences ++ re-occurrence
  - Negative, meaningful consequences -- re-occurrence

Consequences



Consequence analysis

- Not all consequences are evenly ”strong” or “valuable” when it comes to influencing the probability of behavior to re-occur in the future.
- Use the consequence analysis!
- Important dimensions:
  - Is the consequence positive or negative?
  - Is the consequence immediate or future?
  - Is the consequence certain or uncertain?

Example

Consequence	Positive/ Negative	Immediate/ future	Certain/ Uncertain
I “escaped” the meeting for a second	P	I	C
A colleague makes a remark on my behavior	N	I	U
I have seen something that makes me want to step out of the meeting	N	I	U

EXERCISE

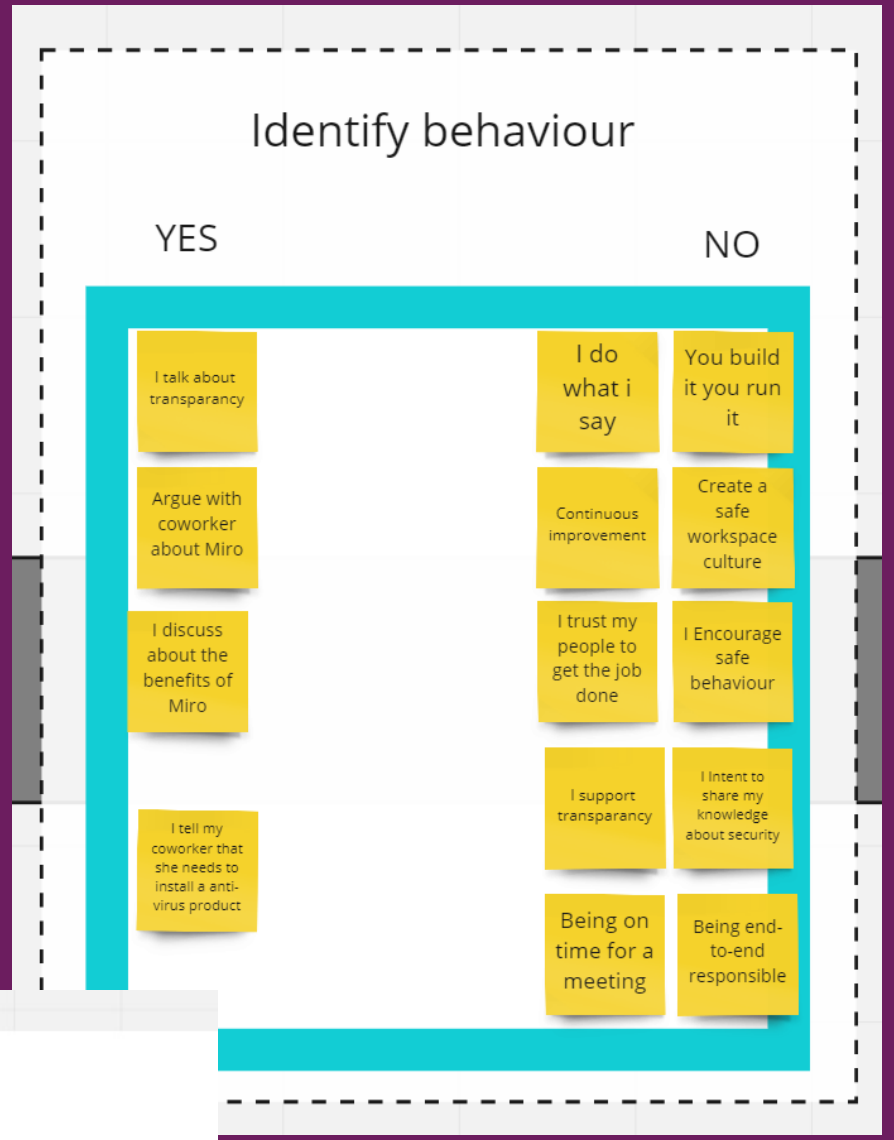
Please go to the Miro board.

Take 10 min to discuss in small sub groups which stickies are describing behaviour and which are not.

Place stickies on the board under yes or no accordingly and prepare a summery for in the main room.

Discuss the result with the group.

Identify Behaviors



1. Behavior can be observed

2. Avoid VGAS

3. Do the dead man test

- ❑ Rule 2: avoid VGAS
1. Values: honest, open
2. Generalities: professional, creative, teamwork
3. Attitude: quality-minded, customer-oriented, cost-aware
4. Status: wear glasses, sleep, sit

EXERCISE

Please go to the Miro board.

Take 10 min to discuss in small sub groups what behaviors you expect in a SRE culture?

Discuss the result with the group.

Psychological Safety
is an important factor for SRE to succeed

"Psychological safety describes a belief that neither the formal nor informal consequences of interpersonal risks, like asking for help or admitting a failure, will be punitive."

The fearless organization (2018) - Amy Edmonson

Psychological safe teams learn from mistakes

- Talk about mistakes and learn from them
- This applies in all sectors, banking, hospitals, IT…
- Teams make less workarounds
- Take more risk and do more experiments

Psychological safety was discovered accidently

While researching the correlation between high-performing and Low- performing medical teams and their error rates the research showed:

- High performing teams had higher error rates ,
- Conclusion more Mistakes
- High performing teams had fewer medical errors
- Conclusion less Mistakes

What was going on?

- High performing teams were reporting the mistakes because they felt safe
- Low performance teams were not reporting the mistakes because they feared being fired!

Psychological Safety – The Zones

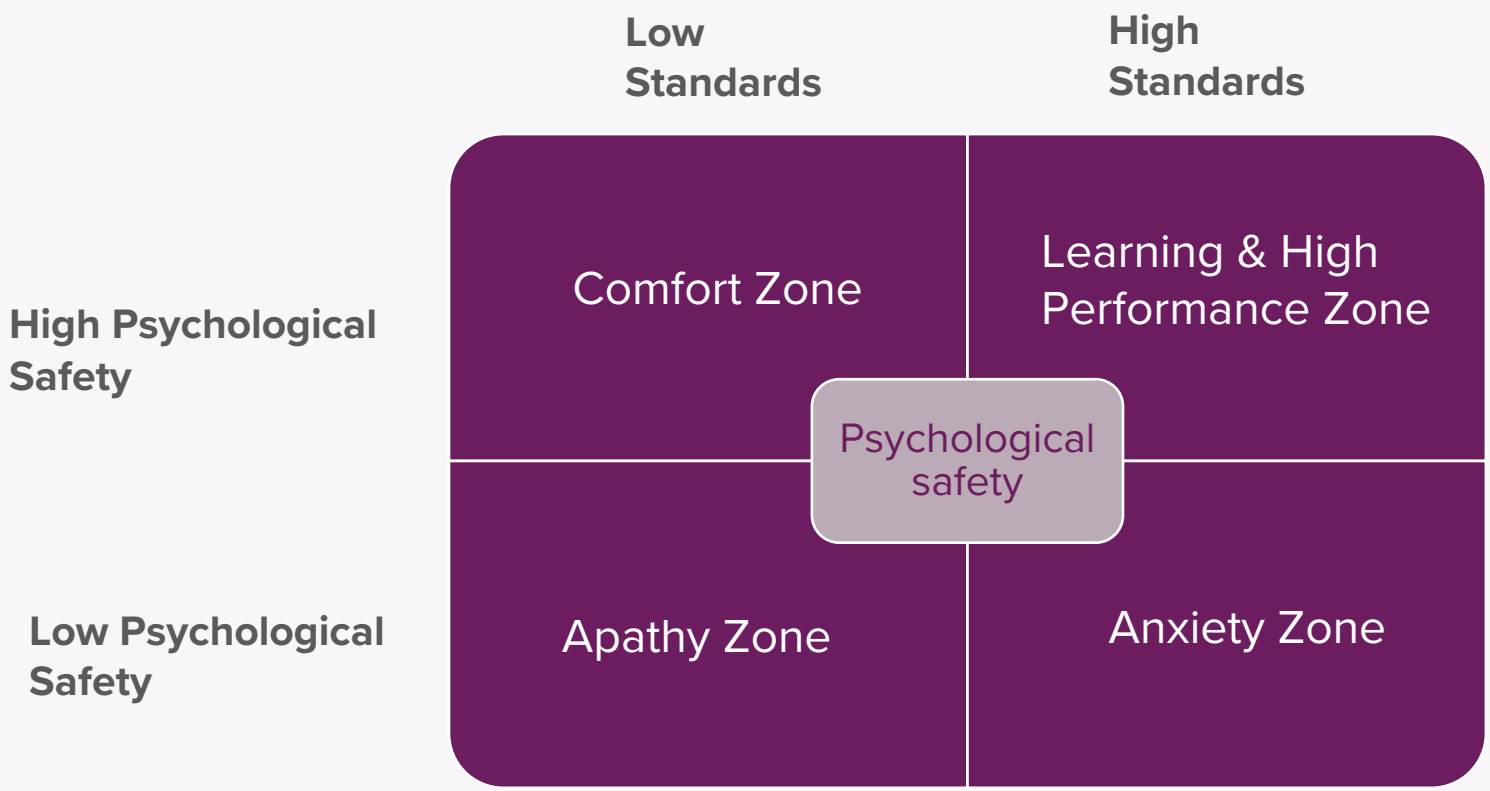


Figure 1: The fearless organization

Why do SRE implementations fail?

Why changes fail?

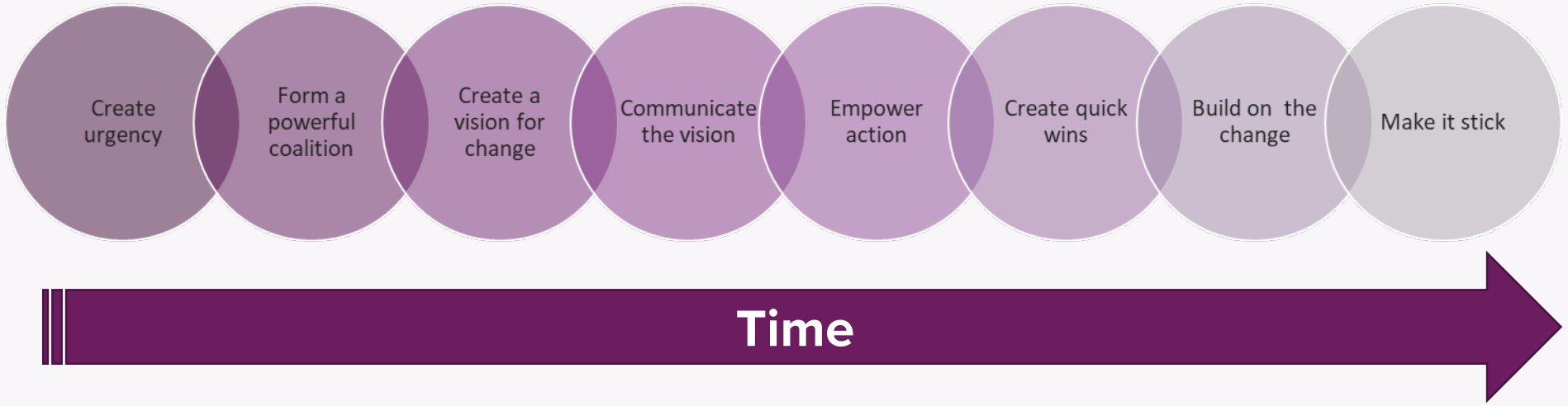
"More than 70% of change initiatives fail"

John Kotter

"Culture eats strategy for breakfast"

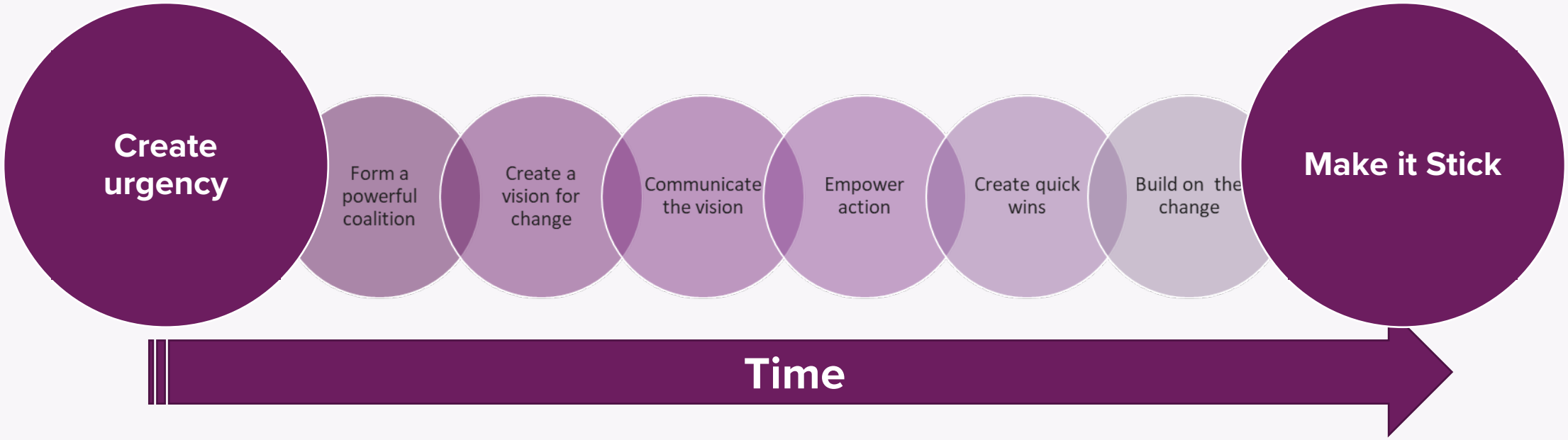
Peter Drucker

Innovation is a must for organizations to stay relevant



Change Model, J. Kotter

Innovation is a must for organizations to stay relevant



Change Model, J. Kotter

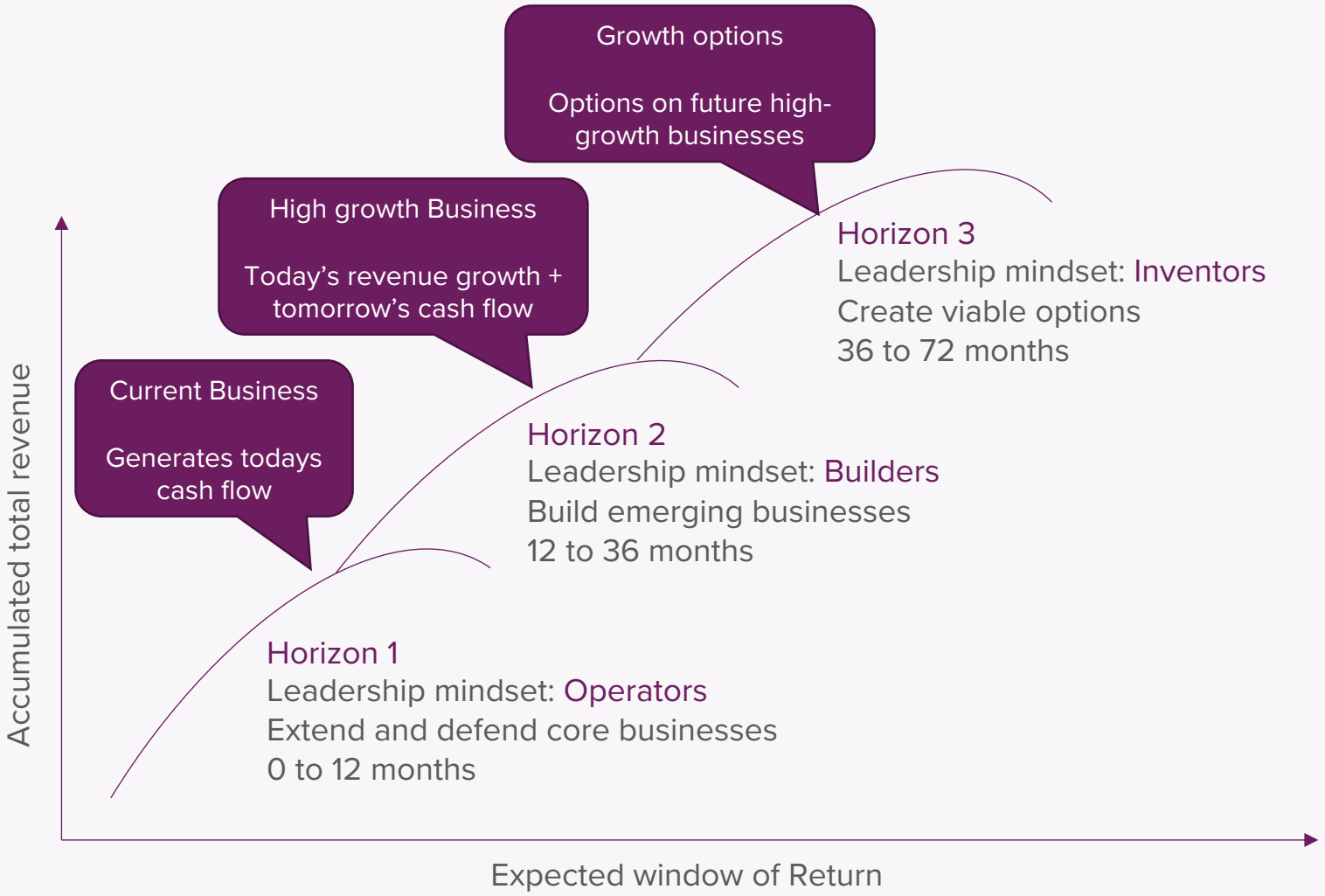
Companies need innovation and change!

Innovation is a must to stay relevant

"If the rate of change on the outside exceeds the rate of change on the inside, the end is near"

Jack Welch – Former CEO General Electric

Innovation is a must for organizations to stay relevant



EXPLORE (innovations)

Horizon 3: Far away Future
Goal: Create new business
Key Metrics: No. of potential validated growth ideas

Horizon 2: Near Future
Goal: Start creating revenue
Key Metrics: Revenue growth

Horizon 1: Now
Goal: Maximize returns
Key Metrics: Optimize Margin

EXPLOIT (operations)

Source: McKinsey Three Horizon model

How to experiment?

- Always start with a clearly defined hypothesis
- Keep experiments small and fail fast
- Use proper measuring to support your hypothesis
- Experiments should not NOT be exclusively for a product

Agenda

Part 1
- Introduction to SRE
- The principles of SRE
- SLIs, SLOs, SLAs

Part 2
- Error Budgets
- Monitoring distributed systems
- Toil

Part 3
- Incident Management
- Postmortems

Part 4
- SRE Culture
- **Onboarding SRE in an Organization**
- Ask me anything & Reflection

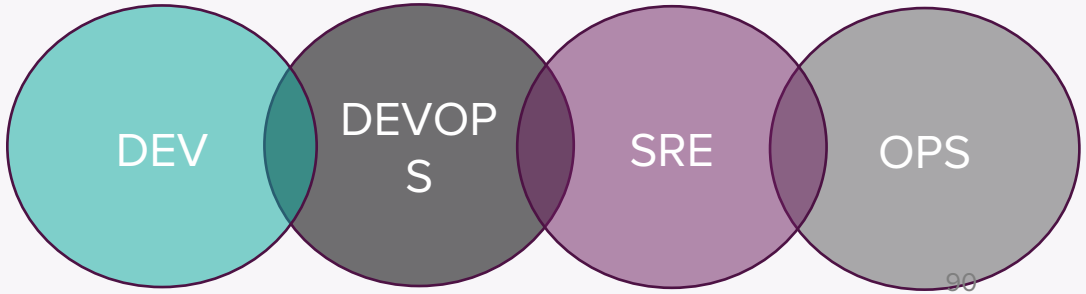
How do you fit SRE within your organization?

Types of SRE implementations

- Kitchen Sink, a.k.a. “Everything SRE”
- Infrastructure
- Tools
- Product/Application
- Embedded
- Consulting

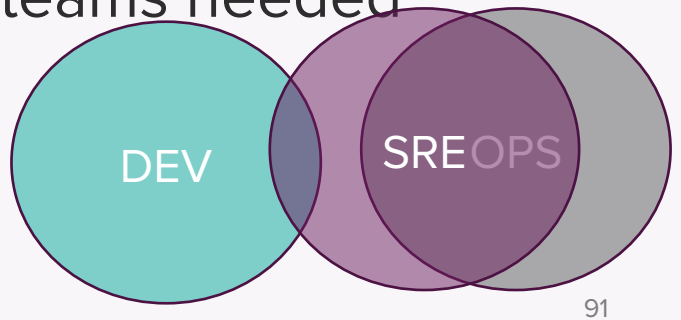
Kitchen Sink, a.k.a. “Everything SRE”

- SRE team has a broad and often unbounded scope
- Usually first and only SRE team in the organisation
- SREs have lots of knowledge about dependencies between services and projects
- SREs act as a glue between teams
- When the complexity grows, the impact of the SRE team becomes less
- When there are issues with the SRE team it typically hits the entire business
- SRE team can become a bottleneck!



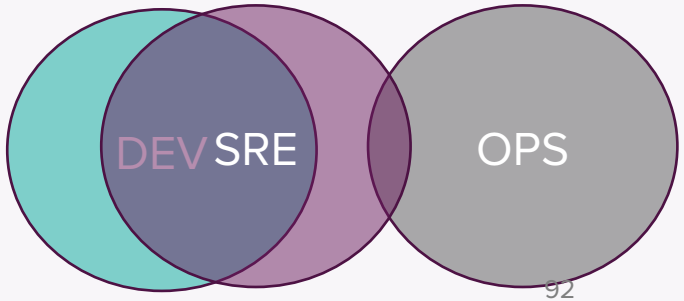
Infrastructure

- SRE team delivers platform services (Infra, K8S, CI/CD, etc..)
- This setup allows DEV teams to focus on the product that they are building
- SRE team can focus on delivering highly reliable platform services
- Infrastructure teams can become bottleneck...
- Automation is key here!
- Growth of the company & systems = more infrastructure teams needed



Tools

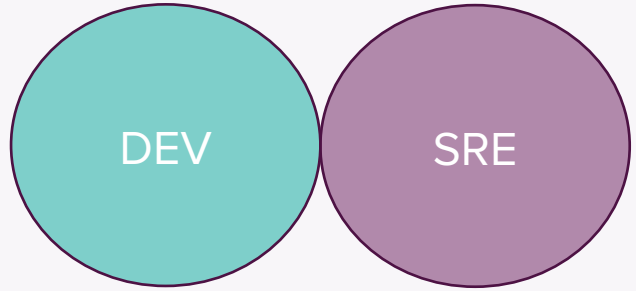
- These teams deliver software which helps developers measure, maintain and improve system reliability
- This approach tend to focus more on support & planning systems
- SRE teams produce a high amount of code for automation



Product application

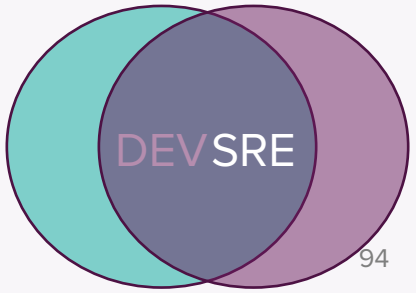
- SRE team improves critical applications or business areas
- Reliability is responsibility of Dev team
- Clear scope & helps prioritizing work
- Scaling requires more teams → can lead to duplication of components and loss of standards practices

MOST TYPICAL SET UP



Embedded

- SREs embedded within development teams
- Enables focused SRE expertise to be directed to specific problems or teams
- Allows side-by-side demonstration of SRE practices



Consulting

- Similar to the embedded approach, but only temporarily joining dev or ops teams.
- Consulting SRE's avoid changing code and configuration
- Easy scaling because SRE is a center of expertise
- Goal should be teaching others!

EXERCISE

Please go to the Miro board.

Take 7 mins to discuss how SRE will be set up in your team / organization

Discuss the result with the group.

Find the approach that works for your organization!

You are not Google!

Ask me anything!

Xebia

Thank you!