How Your Systems Keep Running Day After Day

Resilience Engineering As DevOps

John Allspaw
Adaptive Capacity Labs

example #1

```
rm -rf $PATHNAME
```

example #2

Showing **1 changed file** with **1 addition** and **1 deletion**.

`index.html`

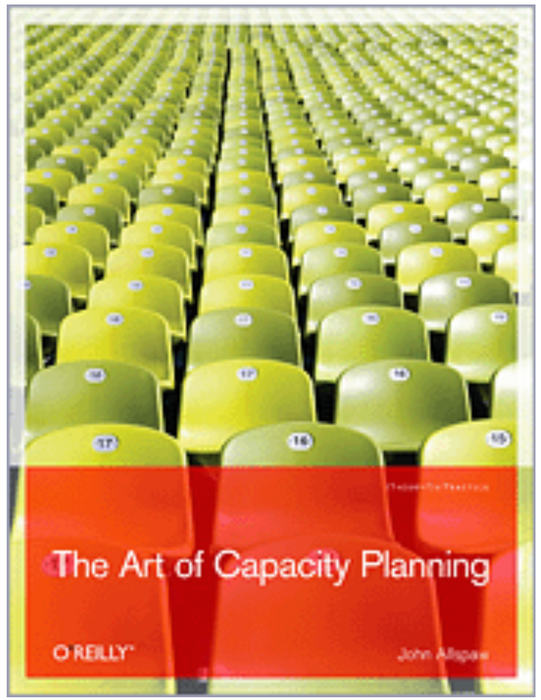
```
@@ -1,2 +1,2 @@
-<!-- Status: Ok -->
+<!-- Status: OK -->
```

all work is contextual

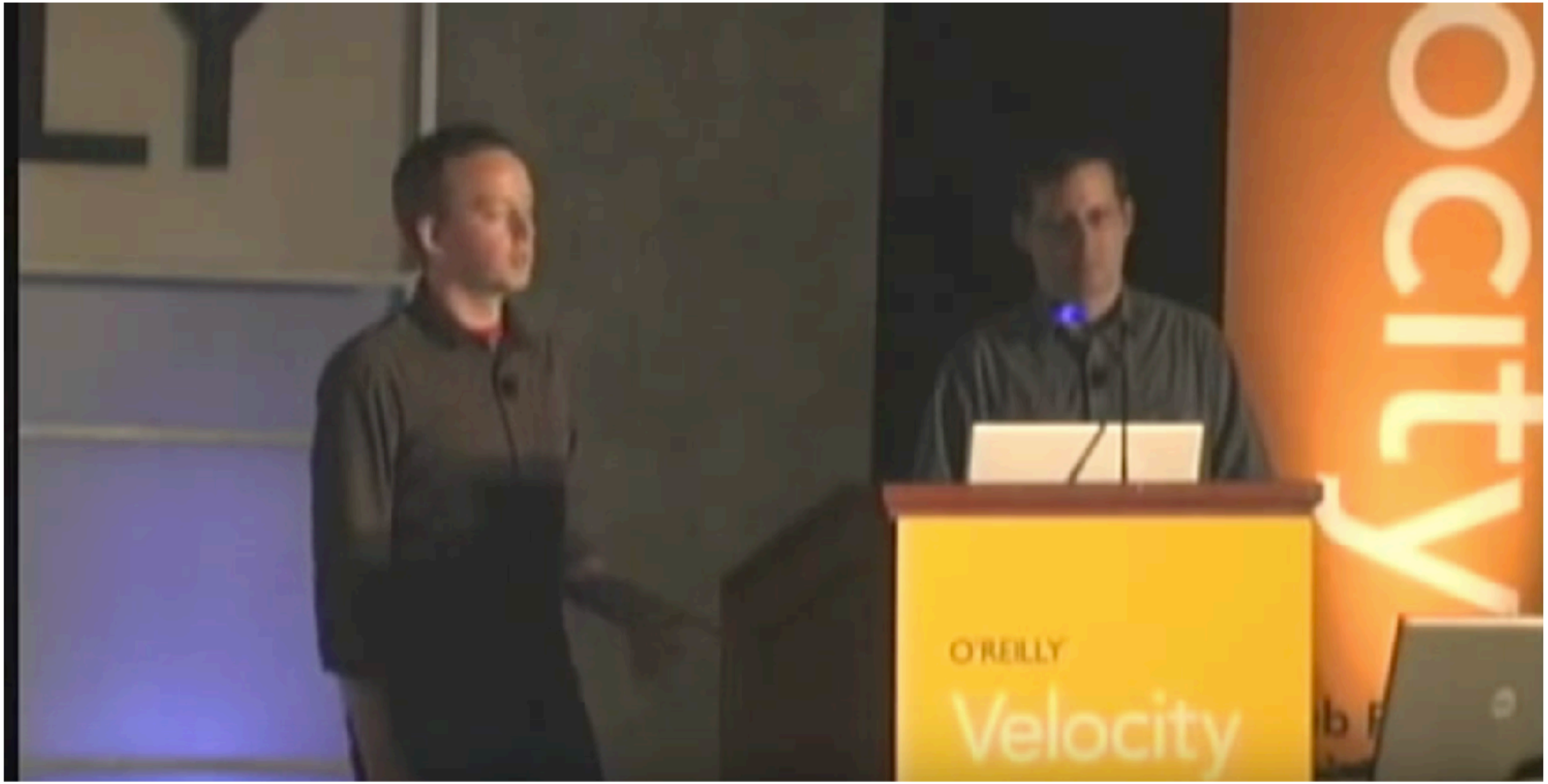
about me

flickr

Etsy



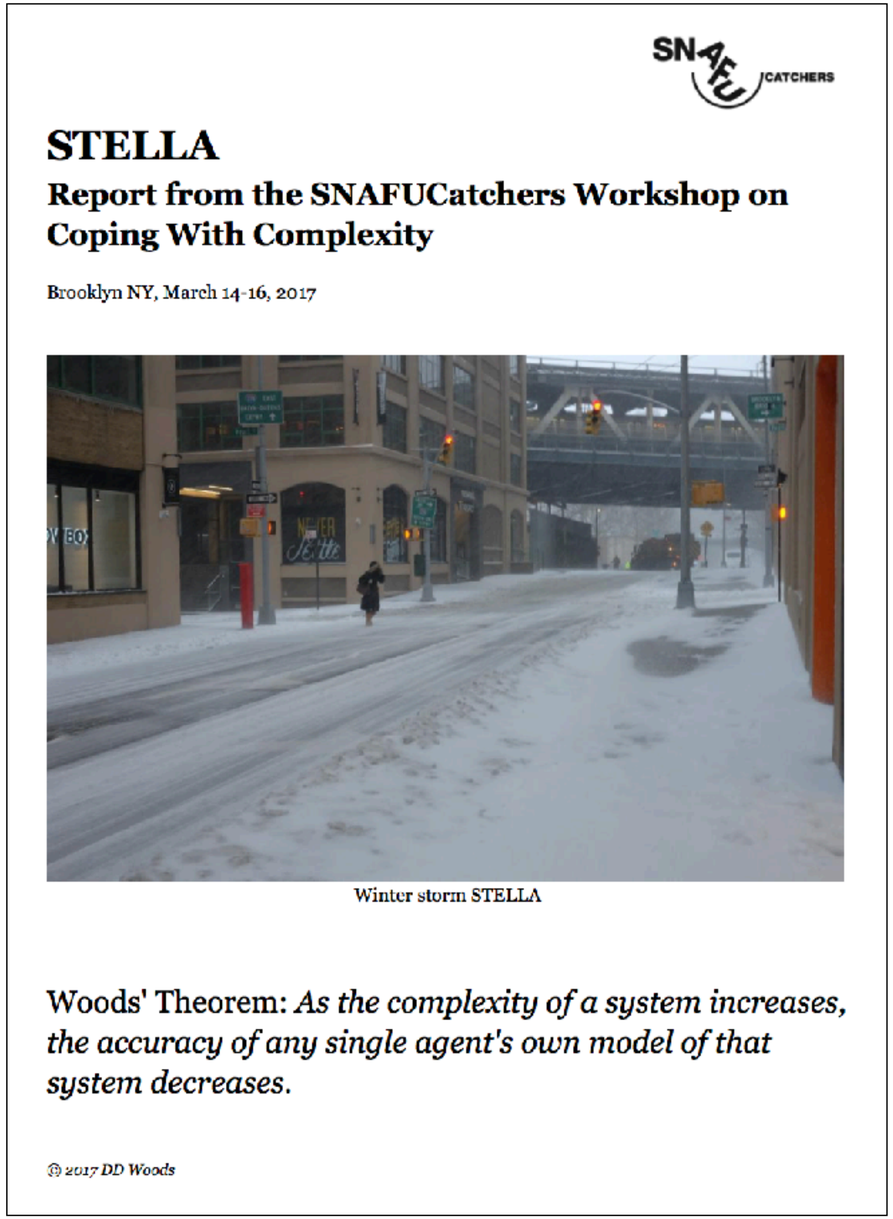
Velocity Conference 2009



TRADE-OFFS UNDER PRESSURE: HEURISTICS AND OBSERVATIONS OF TEAMS RESOLVING INTERNET SERVICE OUTAGES

John Allspaw

LUND UNIVERSITY
SWEDEN



SNAFU CATCHERS

STELLA

Report from the SNAFUCatchers Workshop on Coping With Complexity

Brooklyn NY, March 14-16, 2017

Winter storm STELLA

Woods' Theorem: *As the complexity of a system increases, the accuracy of any single agent's own model of that system decreases.*

http://stella.report

Year-long project
Researchers analyzed 3 incidents, at:

IBM iex Etsy

Six Themes

- **Postmortems as re-calibration**
- **Blameless v. sanctionless after action actions**
- **Controlling the costs of coordination**
- **Visualizations during anomaly management**
- **Strange Loops**
- **Dark Debt**

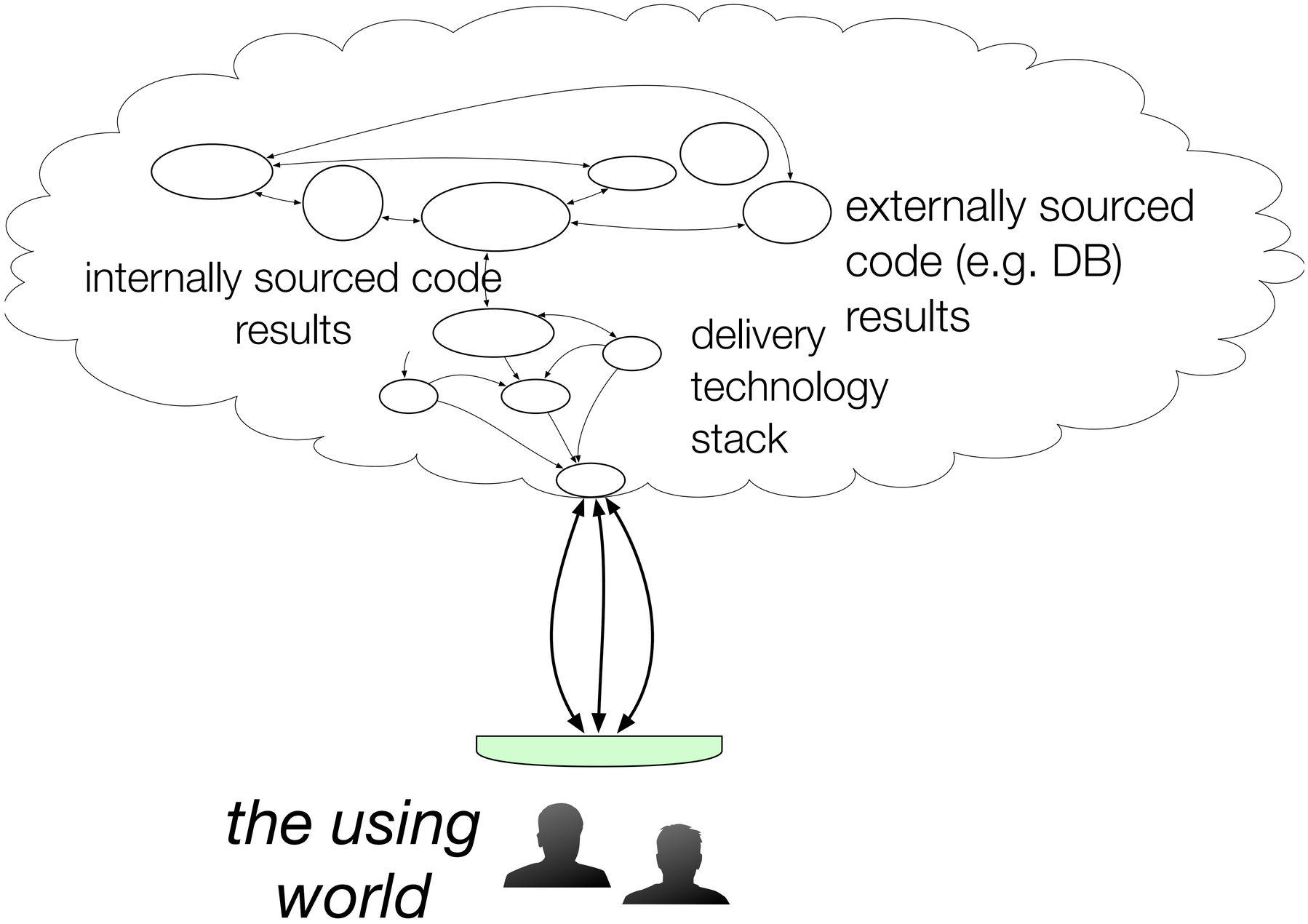
My Main Points

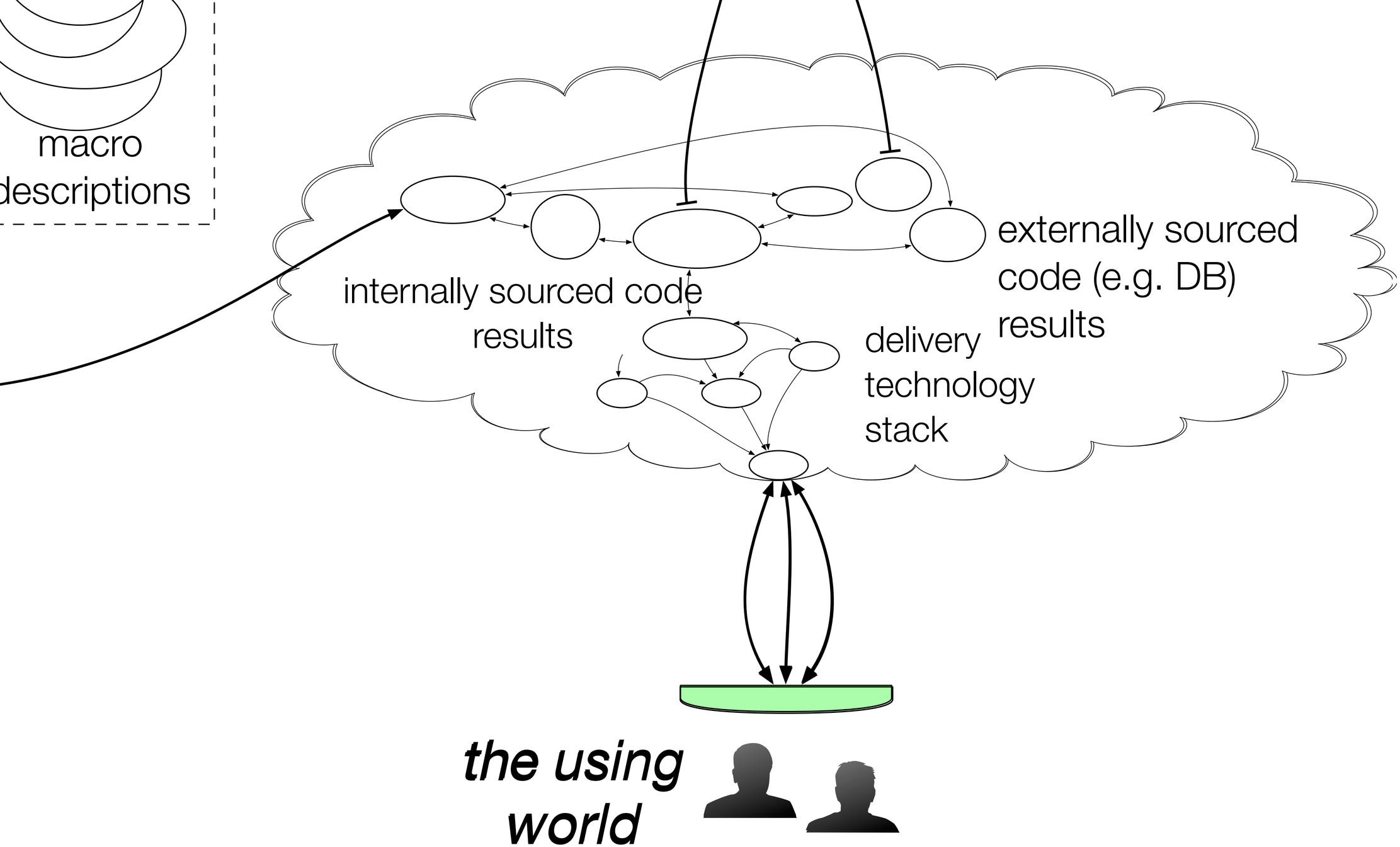
1. We have to start taking human performance seriously in our industry.

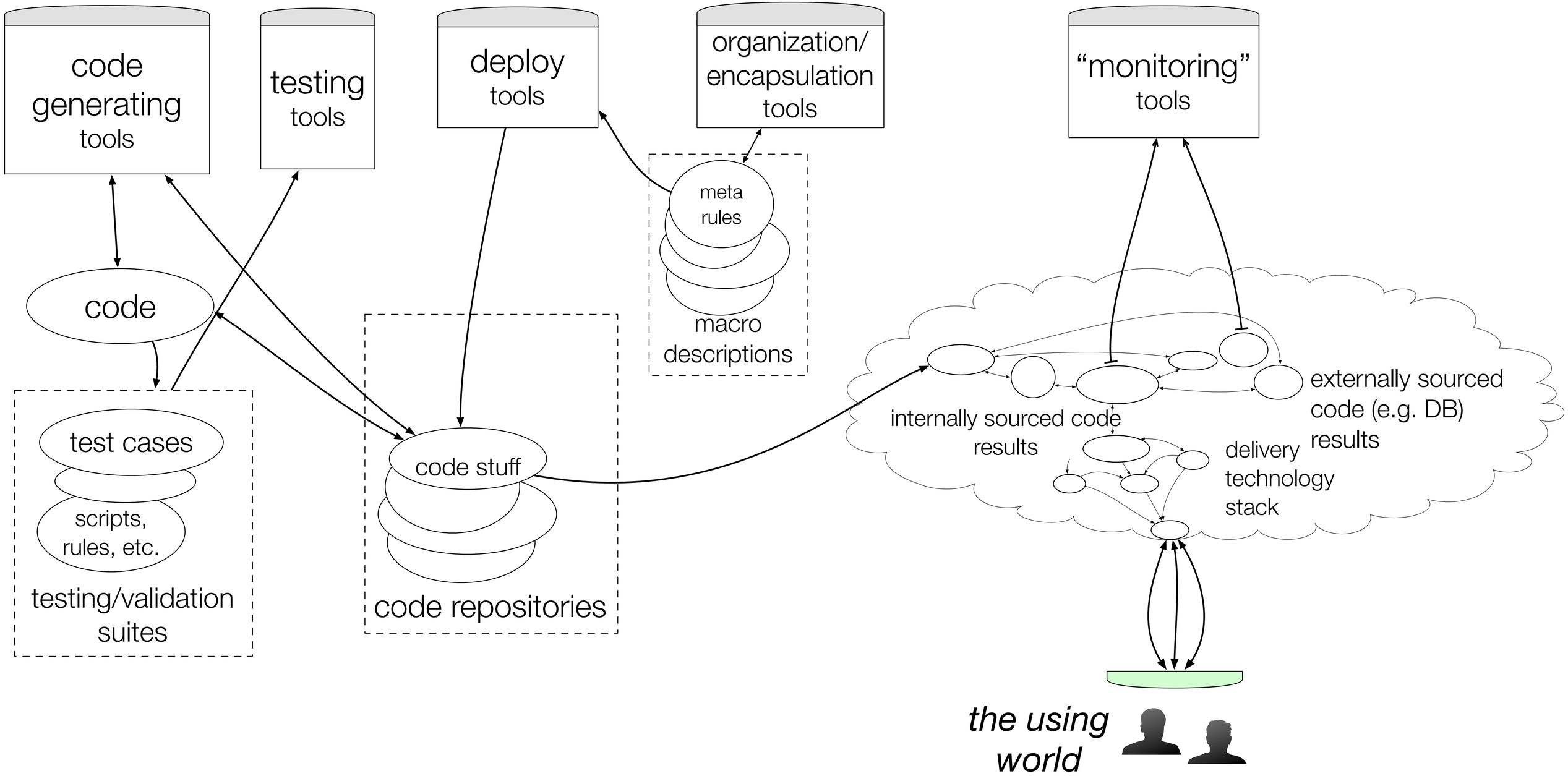
2. We can do this by looking at incidents, beyond what we currently do in postmortems.

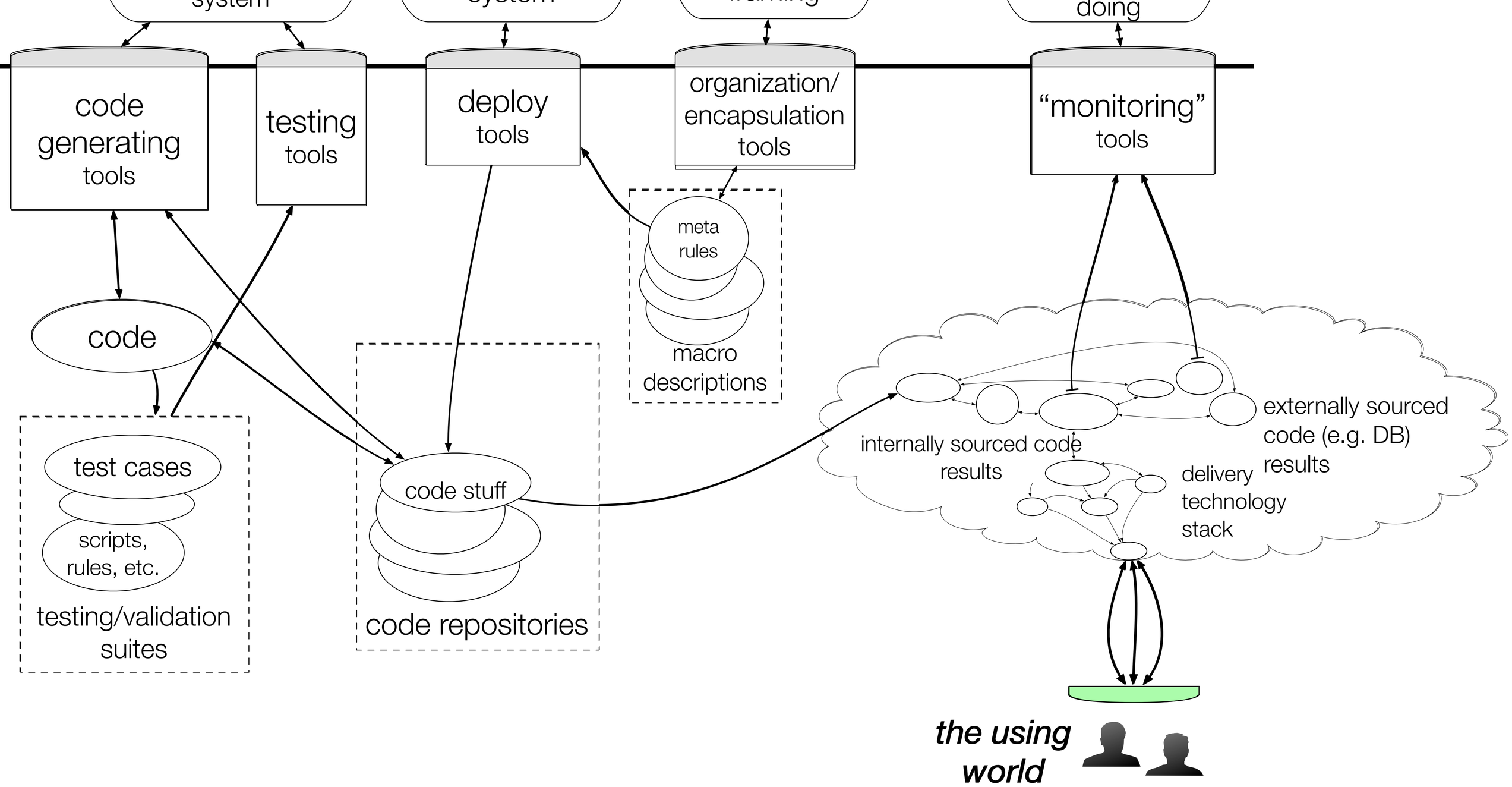
3. Methods and approaches to do this exist from the study of resilience in other domains, but they require real commitment to pursue.

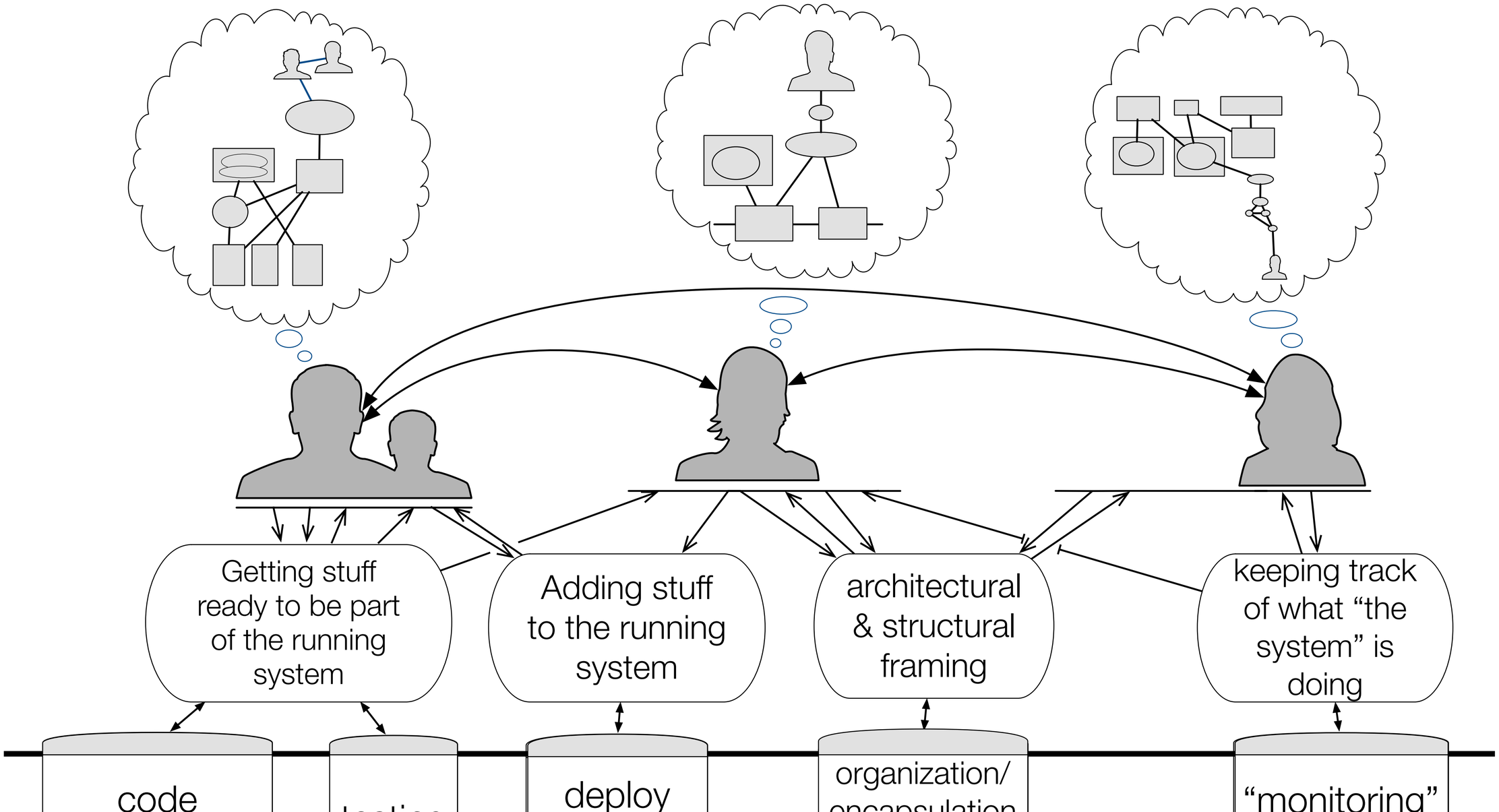
4. Doing this is both necessary and difficult (but very possible!) - and will prove to be a *competitive advantage* for businesses who do it well.

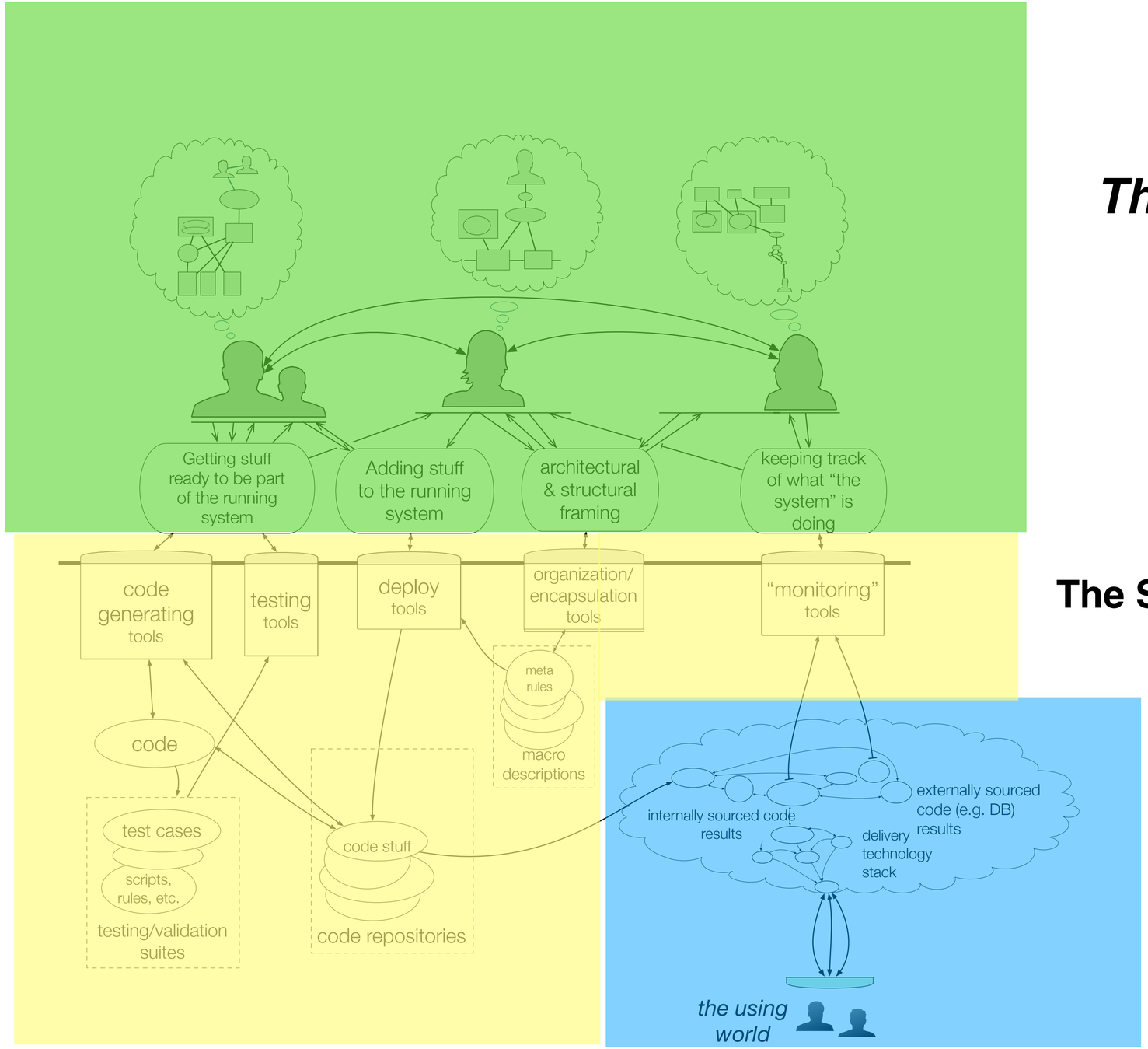

externally sourced code (e.g. DB) results
internally sourced code results
delivery technology stack
the using world


macro
descriptions
internally sourced code
results
externally sourced
code (e.g. DB)
results
delivery
technology
stack
the using
world


code generating tools
testing tools
deploy tools
organization/ encapsulation tools
"monitoring" tools
meta rules
macro descriptions
code
test cases
scripts, rules, etc.
testing/validation suites
code stuff
code repositories
internally sourced code results
externally sourced code (e.g. DB) results
delivery technology stack
the using world


system
system
doing
code generating tools
testing tools
deploy tools
organization/ encapsulation tools
"monitoring" tools
meta rules
macro descriptions
code
test cases
scripts, rules, etc.
testing/validation suites
code stuff
code repositories
internally sourced code results
externally sourced code (e.g. DB) results
delivery technology stack
the using world

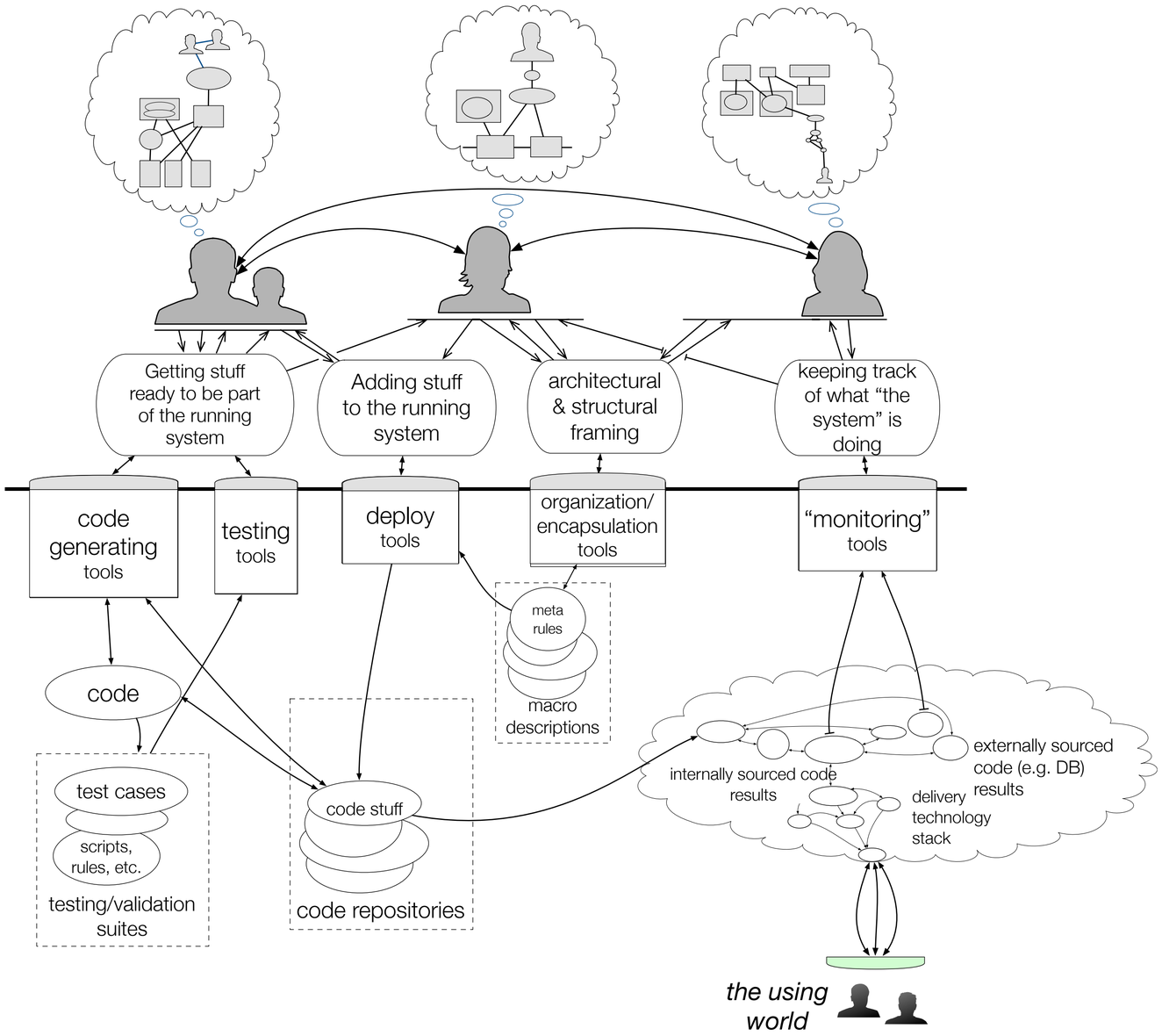

Getting stuff ready to be part of the running system
Adding stuff to the running system
architectural & structural framing
keeping track of what “the system” is doing
code
deploy
organization/ encapsulation
“monitoring”

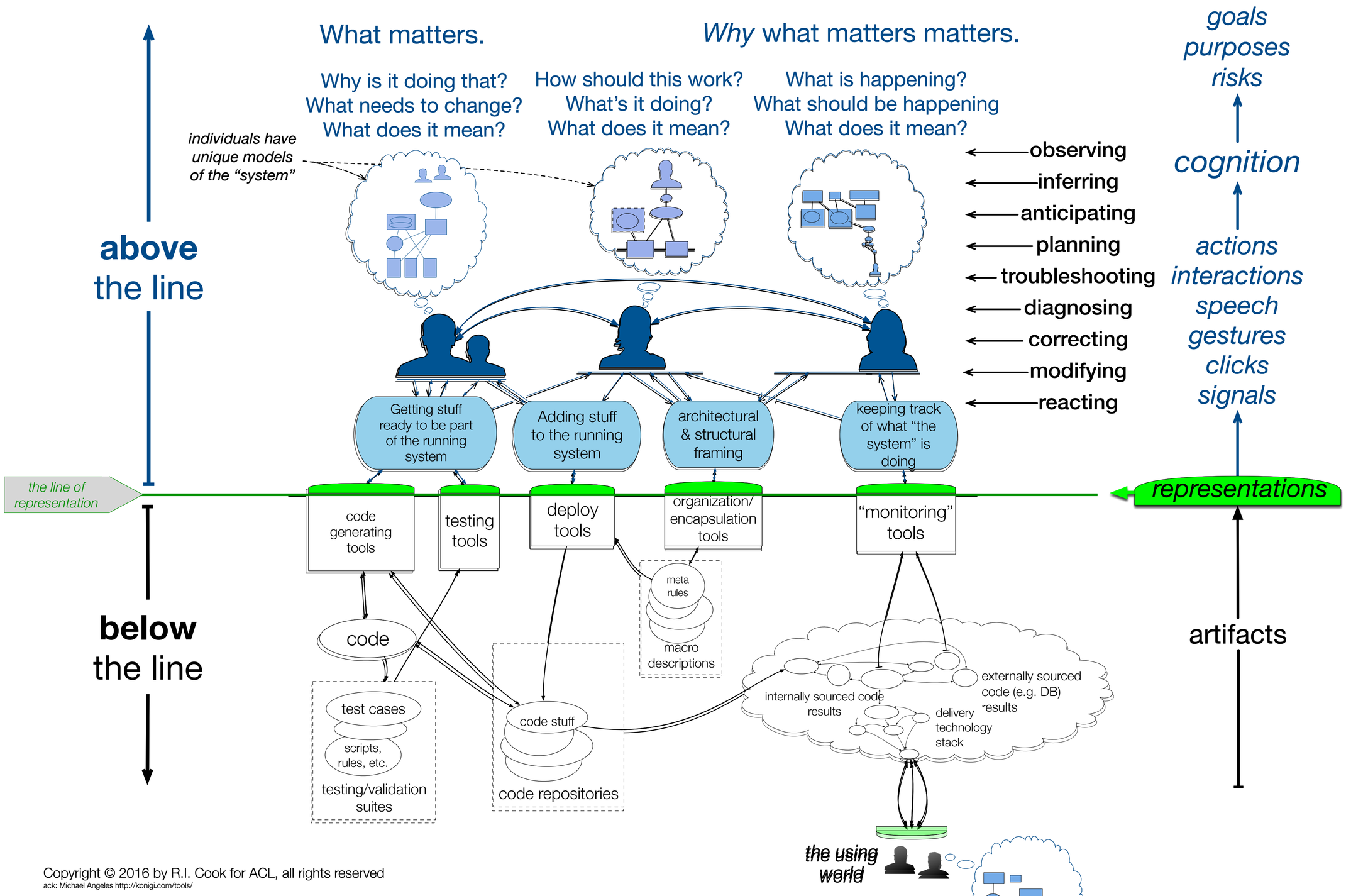


The Work Is Done Here

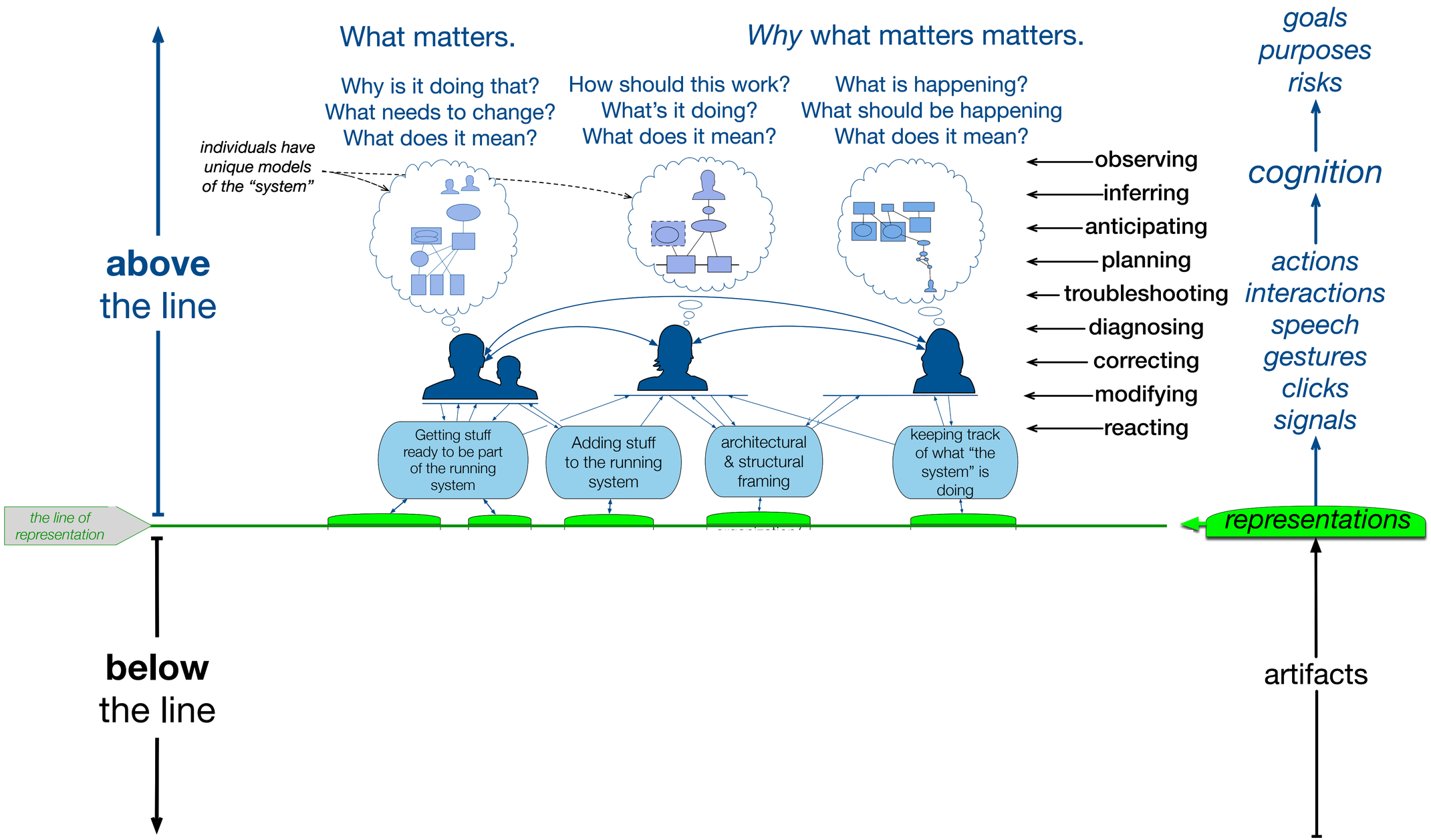
The Stuff You Build and Maintain With

Your Product Or Service


Getting stuff ready to be part of the running system
Adding stuff to the running system
architectural & structural framing
keeping track of what "the system" is doing
code generating tools
testing tools
deploy tools
organization/ encapsulation tools
"monitoring" tools
meta rules
macro descriptions
code
test cases
scripts, rules, etc.
testing/validation suites
code stuff
code repositories
internally sourced code results
externally sourced code (e.g. DB) results
delivery technology stack
the using world



ack: Michael Angeles http://konigi.com/tools/



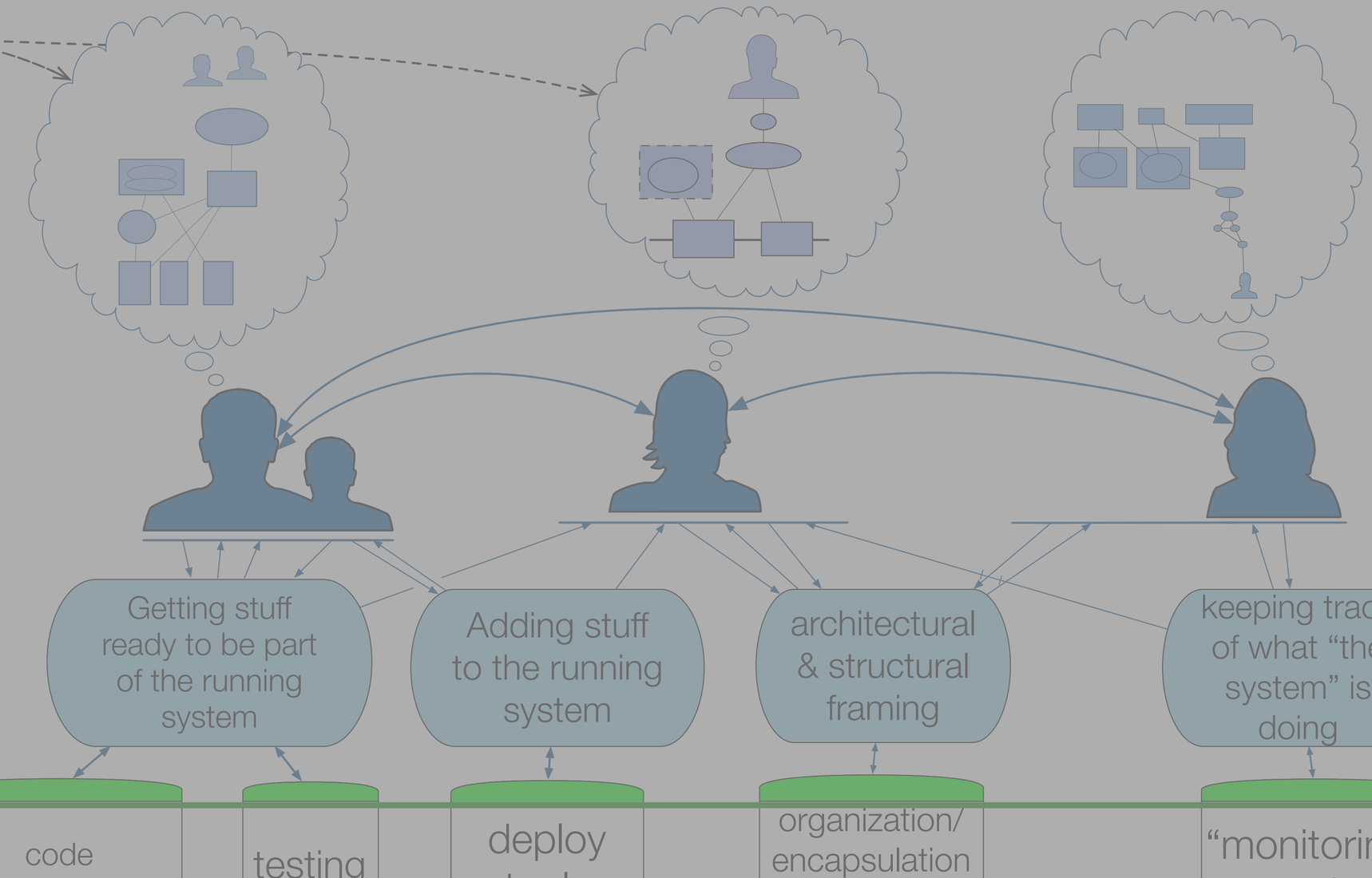
ack: Michael Angeles http://konigi.com/tools/

What matters. *Why* what matters matters.

Why is it doing that?
at needs to change?
What does it mean?

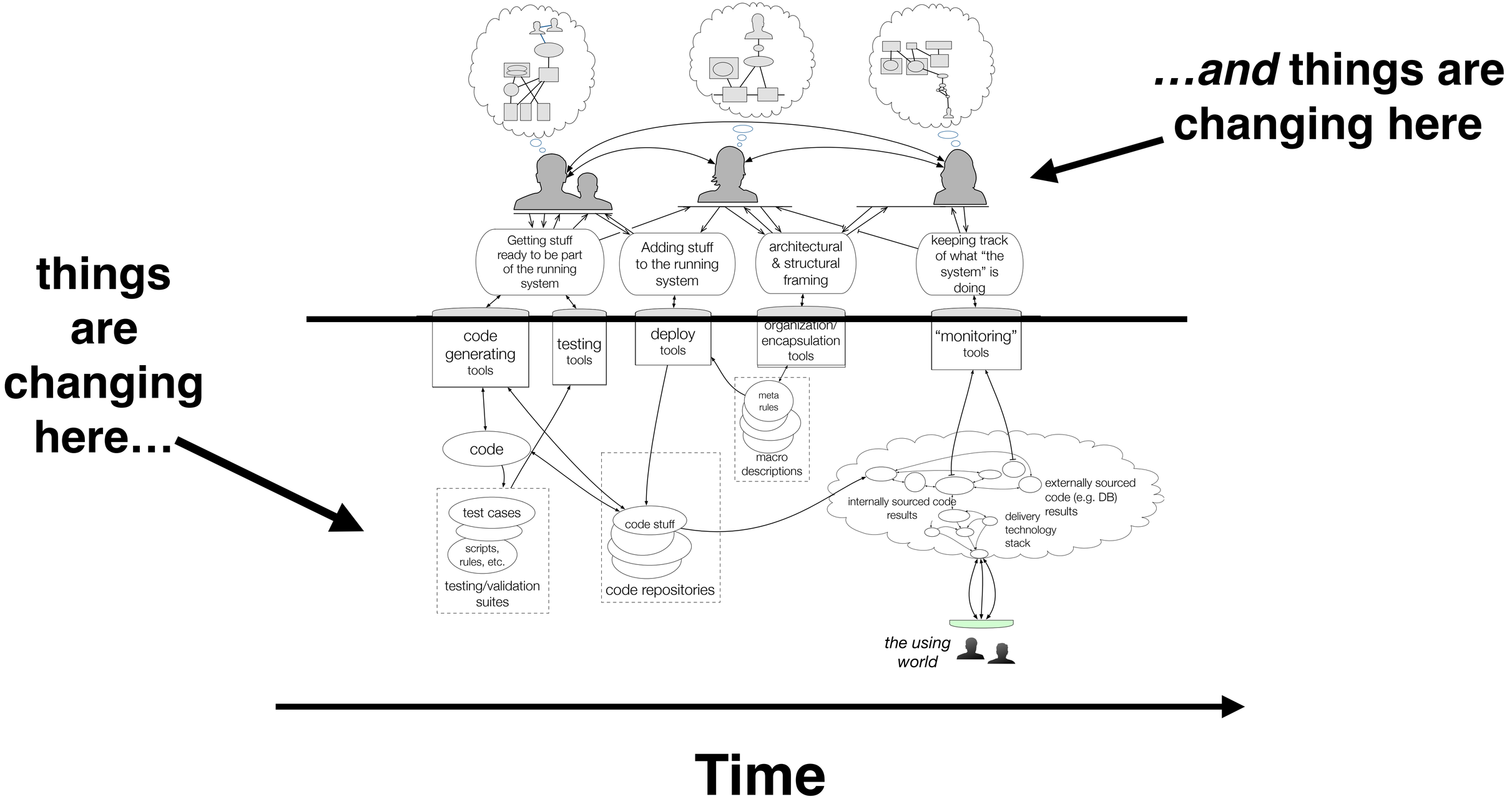
How should this work?
What's it doing?
What does it mean?

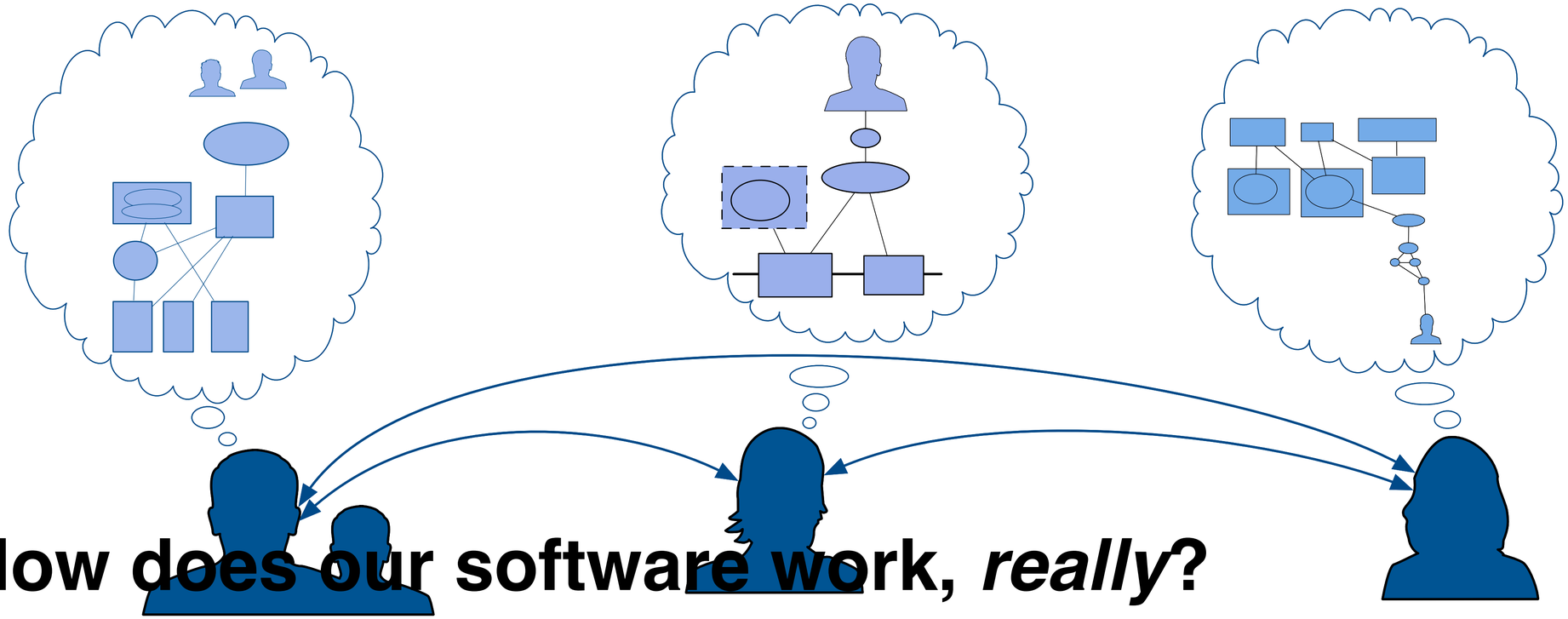
What is happening?
What should be happening
What does it mean?



observing
inferring
anticipating
planning
troubleshooting
diagnosing
correcting
modifying
reacting

purp
ri
cog
act
intera
spe
ges
cl
sig
represe


…and things are changing here
things are changing here…
Getting stuff ready to be part of the running system
Adding stuff to the running system
architectural & structural framing
keeping track of what "the system" is doing
code generating tools
testing tools
deploy tools
organization/ encapsulation tools
"monitoring" tools
meta rules
macro descriptions
code
test cases
scripts, rules, etc.
testing/validation suites
code stuff
code repositories
internally sourced code results
externally sourced code (e.g. DB) results
delivery technology stack
the using world
Time



How does our software work, *really*?

How does our software break, *really*?

What do we do to keep it all working?

how to discover what happens “above the line”?

incidents

(outages, degradations, breaches, accidents, near-misses, “glitches”, untoward/unexpected events, etc.)

what makes incidents interesting?

incidents as…

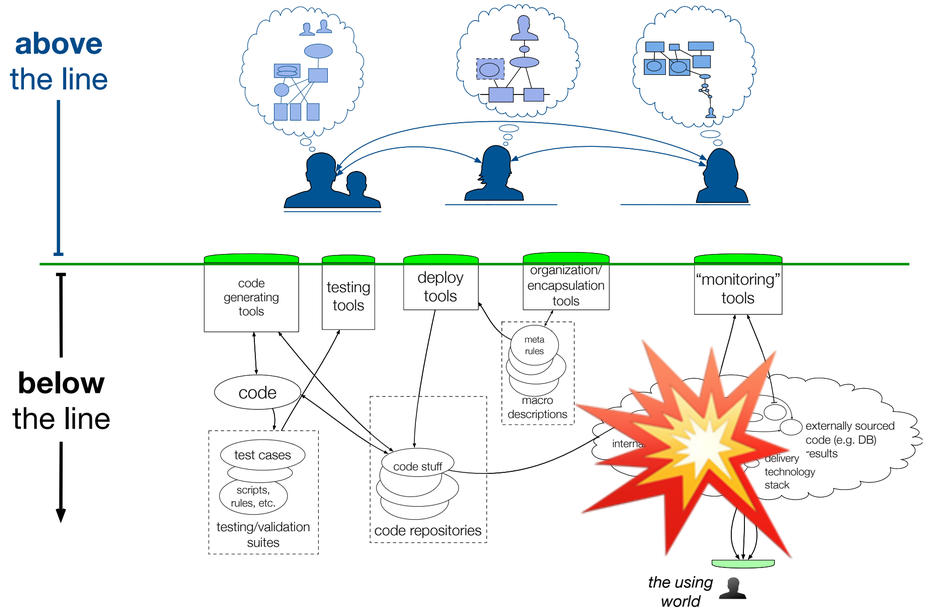
drivers of software design

- shape the design of new components, subsystems, architectures

- *“incidents of yesterday inform the **architectures** of tomorrow”*

- incidents “below the line” drive changes “above the line"

- ***staffing, budgets, planning, roadmaps,*** etc.



incidents as…

motivators for policy

- tend also to give birth to new forms of regulations, policies, norms, compliance requirements, explosion of documentation, auditing, constraints, etc.

- *“incidents of yesterday inform the **rules** of tomorrow”*

- influence ***staffing, budgets, planning, roadmaps,*** etc.

U.S. SECURITIES AND EXCHANGE COMMISSION

“Regulation SCI”

5/6/2010 - Flash Crash - loss of $1 trillion in market value in <10min

3/23/2012 - BATS IPO - systems issue halted the exchange’s own IPO

5/23/2012 - Facebook IPO - systems issue delayed IPO trading

8/1/2012 - Knight Capital - $461 million in 45 minutes

VISA AMERICAN EXPRESS MasterCard DISCOVER JCB

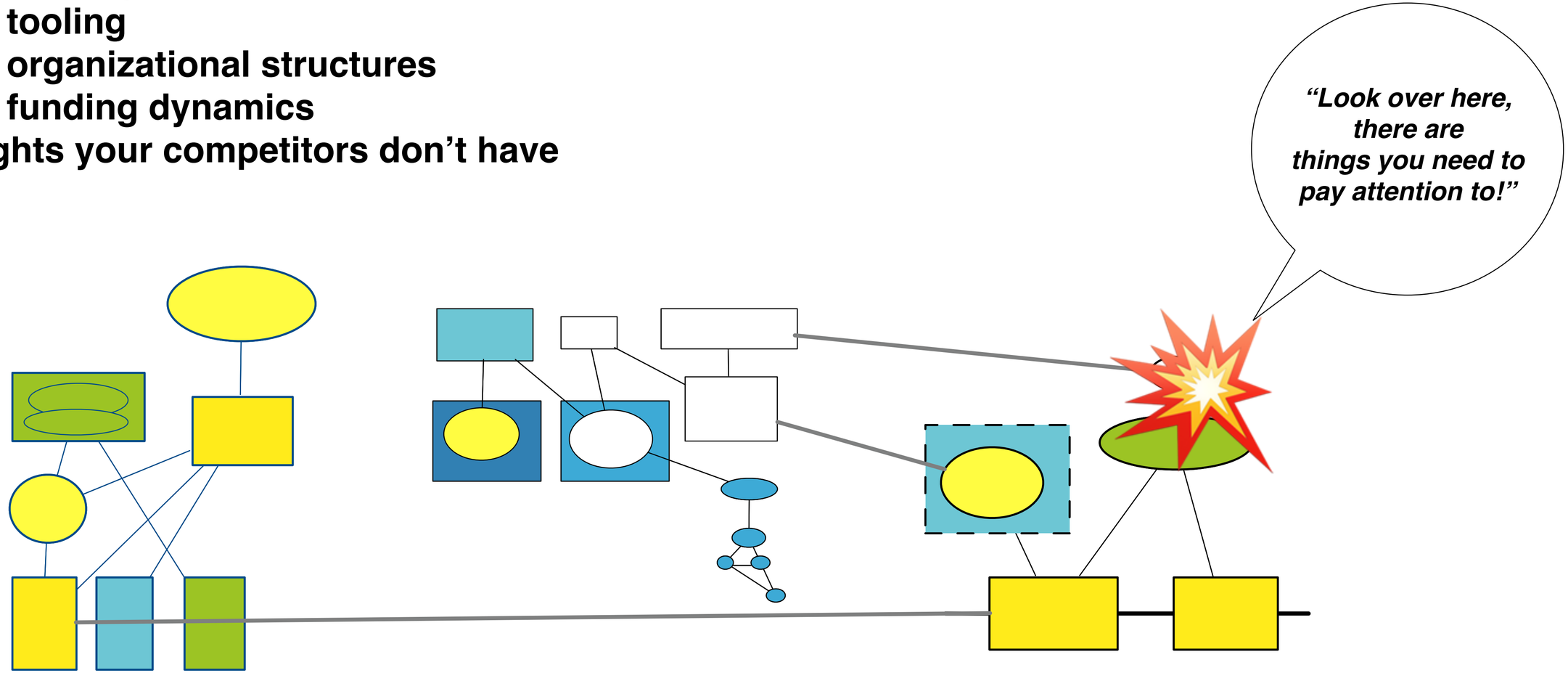
PCI-DSS

1988-1998, **Visa** and MasterCard reported credit card losses due to fraud of **$750 million**

incidents as…

opportunities

new training
new tooling
new organizational structures
new funding dynamics
insights your competitors don't have



incidents help us gauge the delta between

how the system works | **Δ** | **how *we think* the system works**

almost always greater than we imagine

incidents
are
unplanned
investments
in ~~enterprise~~ survival

your company’s

incidents

burn money
burn time
burn reputation
burn staff

} ***unavoidable sunk costs***

You do not control the size of this investment.

The challenge is to maximize the ROI on that investment.

What is it doing?!

Why is it doing that?!

What will it do next?

How did it get into this state?

WTF is happening?

If we do Y, will it help us figure out what to do?

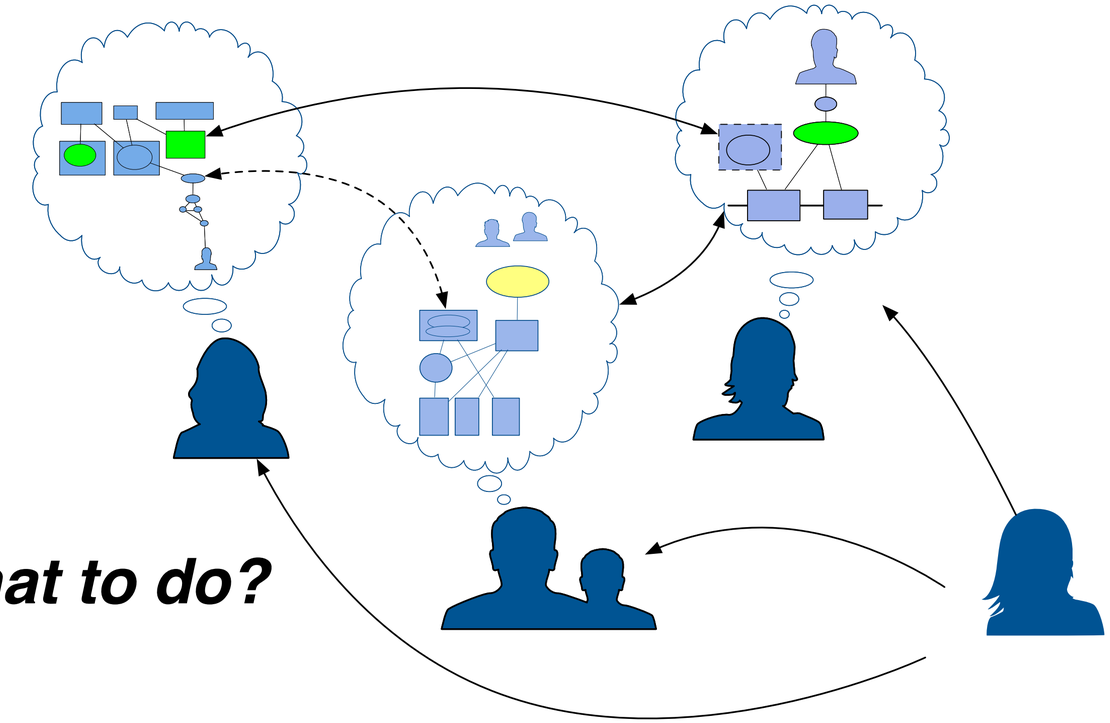
Is it getting worse?

It looks like it's fixed…but is it…?

If we do X, will it prevent it from getting worse…or make it worse?

Who else should we call that can help us?

Is this OUR issue, or are we BEING ATTACKED?!



In the beginning of any incident,
it’s often uncertain or ambiguous
whether it is a “viability-crushing” event or not.

Only hindsight will tell us.

incidents provide calibration about…

how decisions are ***focused***

how attention is ***focused***

how coordination is ***focused***

how escalation is ***focused***

the ***impact*** of time pressure

the ***impact*** of uncertainty

the ***impact*** of ambiguity

the ***consequences*** of consequences

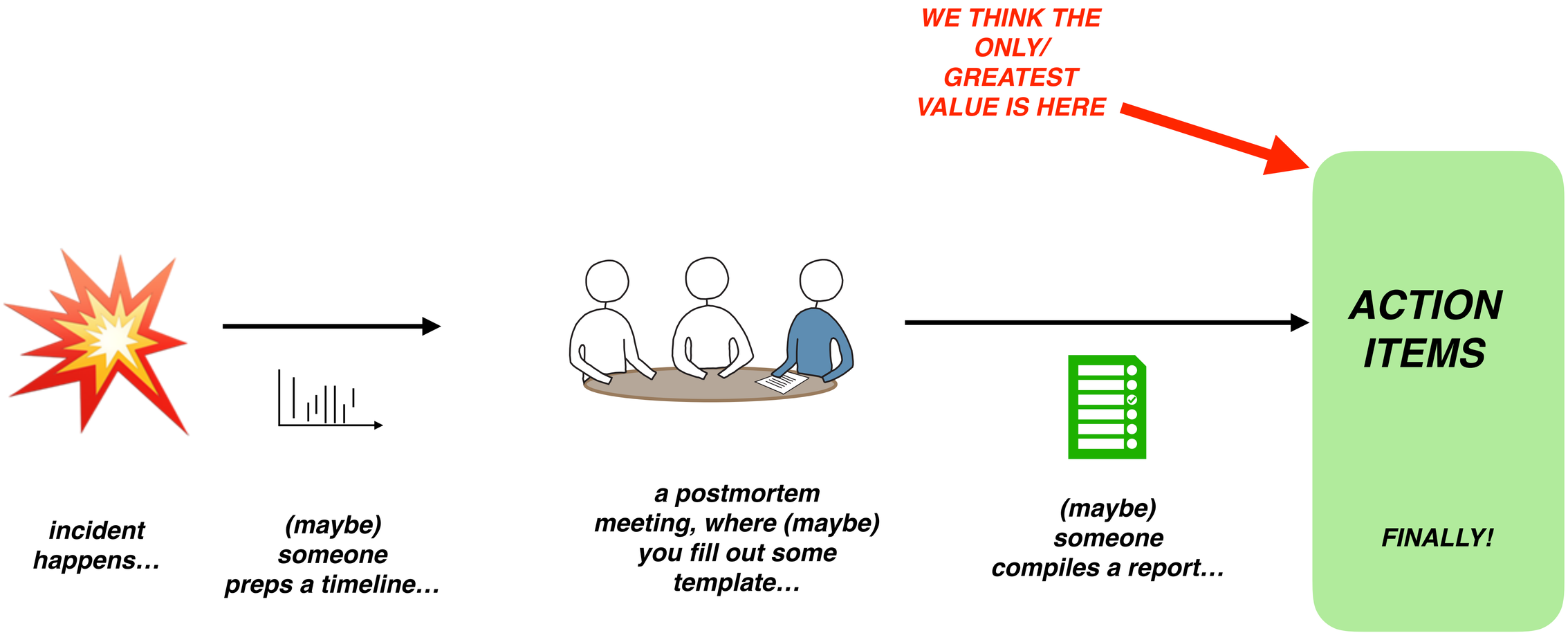
research validates these opportunities

"…nonroutine, challenging events, because these tough cases have the greatest potential for uncovering elements of expertise and related cognitive phenomena." (Klein, Crandall, Hoffman, 2006)

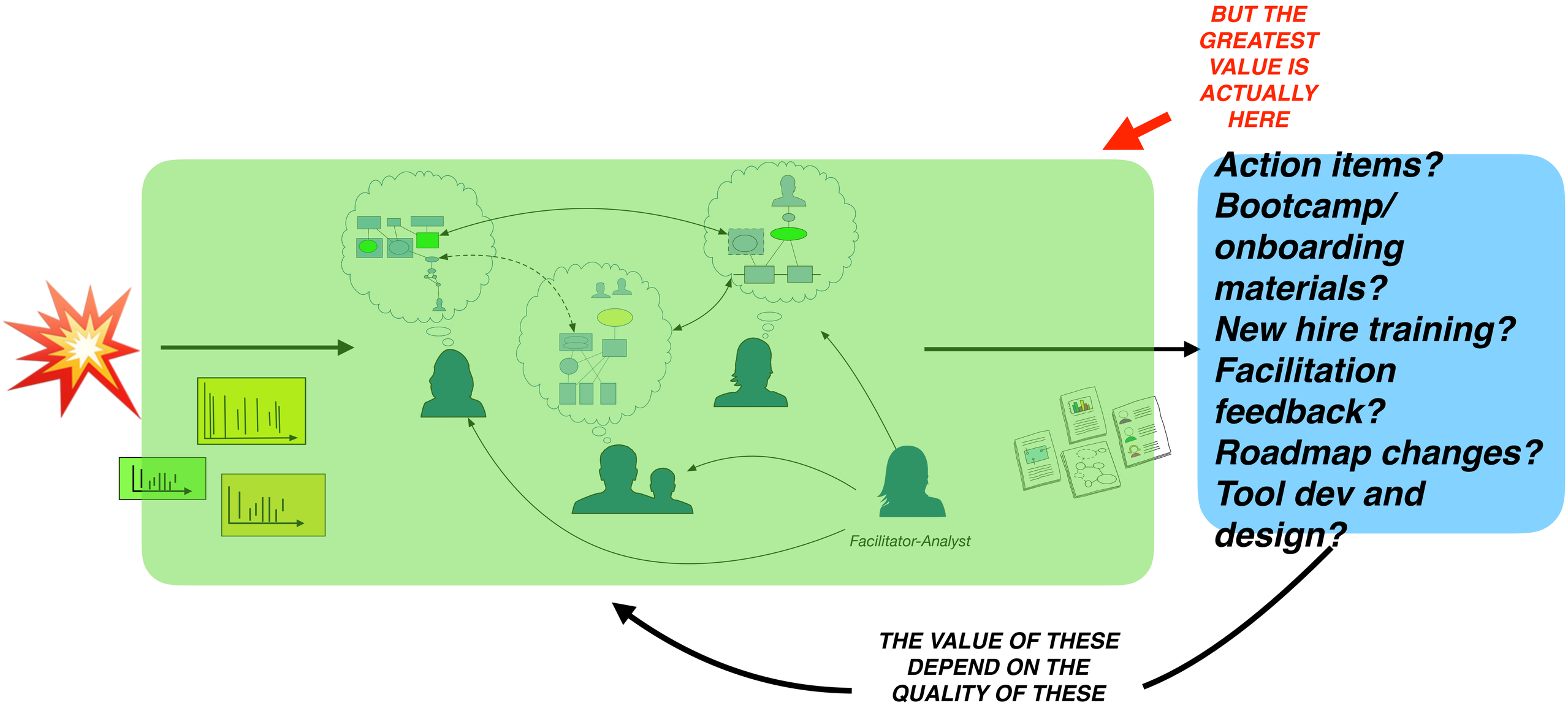
A family of well-worn methods, approaches, and techniques

Cognitive task/work analysis
Process tracing
Conversation analysis
Critical decision method
Critical incident technique
more…

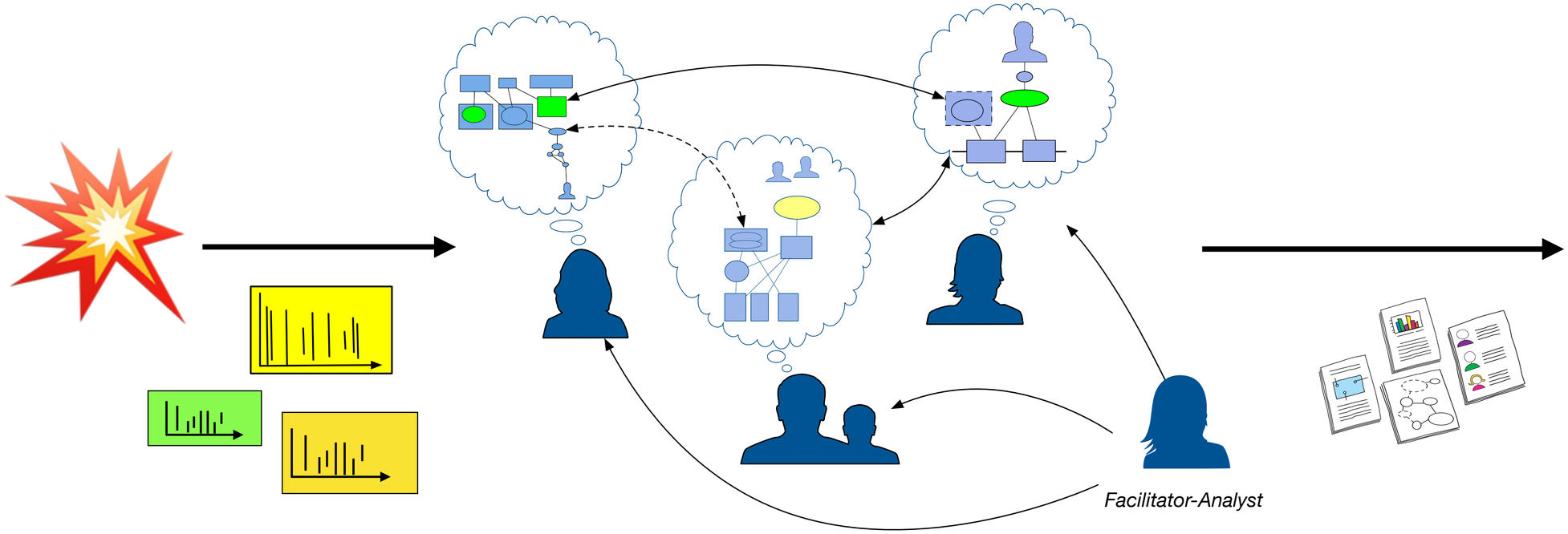
How We Think PostMortems Have Value



How Post-Incident Reviews *Actually* Provide Value

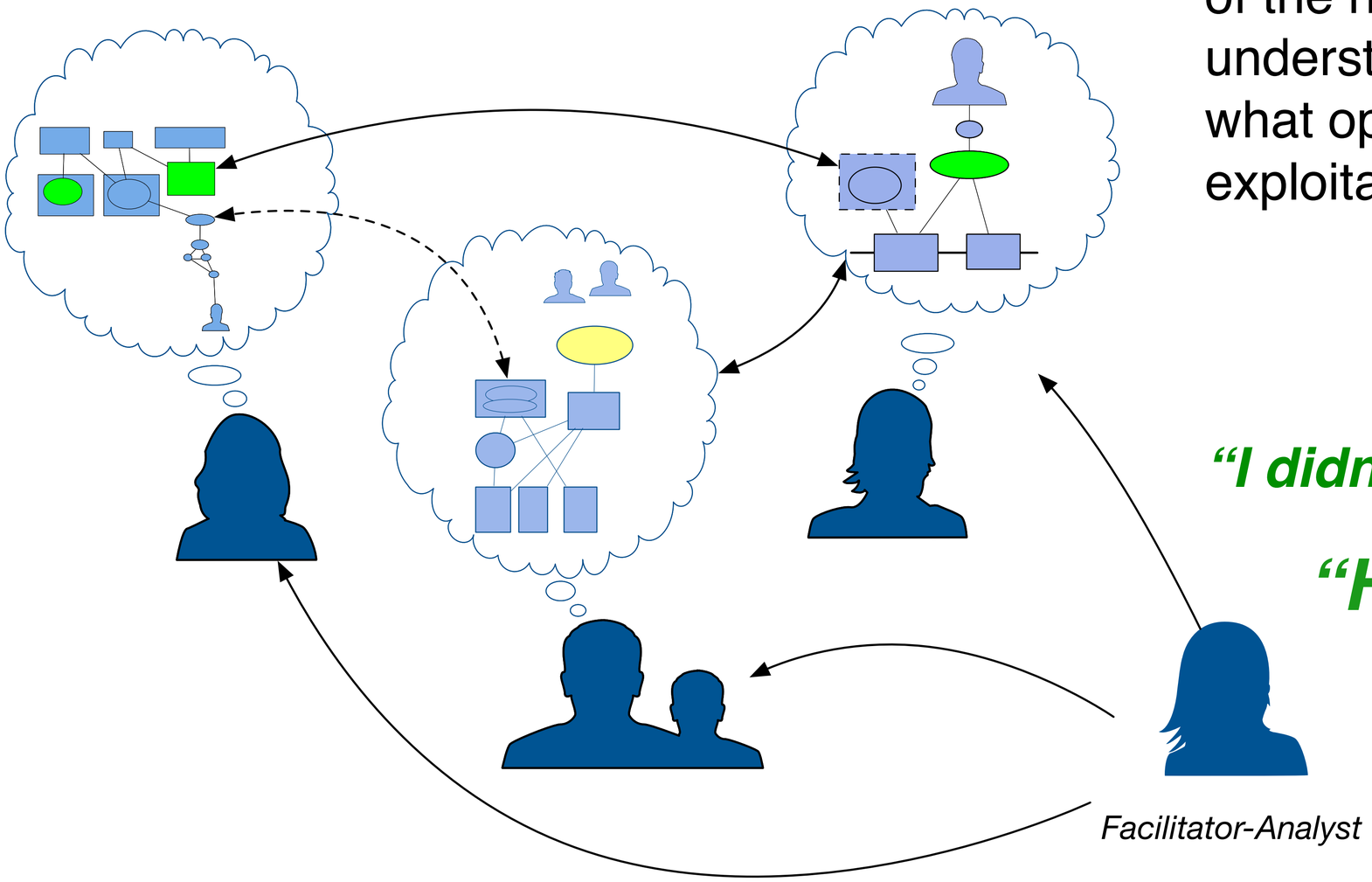


How Post-Incident Reviews *Actually* Provide Value



Action items?
Bootcamp/ onboarding materials?
New hire training?
Facilitation feedback?
Roadmap changes?
Tool dev and design?

AN OPENING TO RECALIBRATE MENTAL MODELS



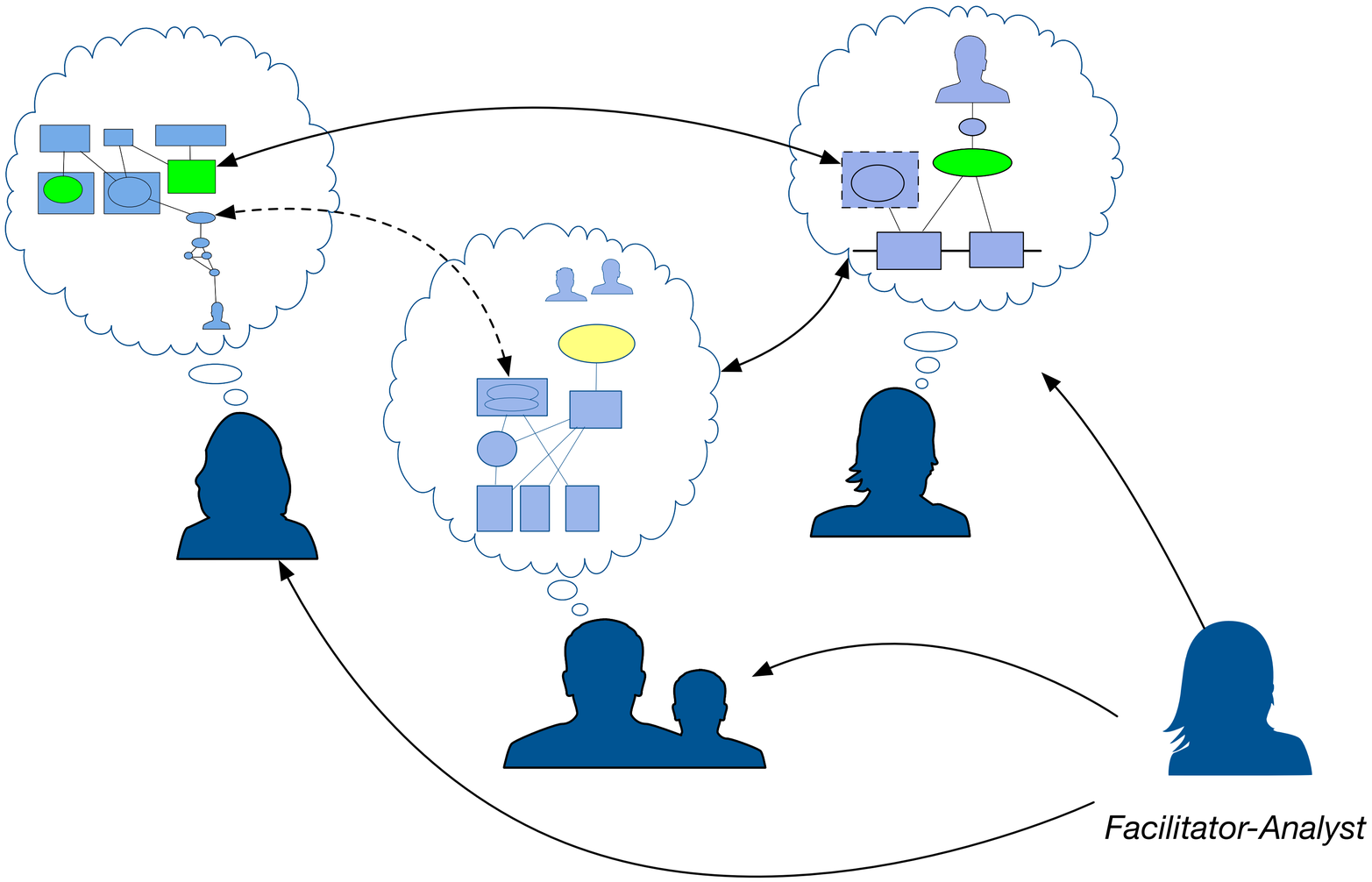
"…informs and recalibrates people's models of the how the system works, their understandings of how it is vulnerable, and what opportunities are available for exploitation." - *Stella Report*

"I didn't know that it worked that way."

"How did it EVER work?!"

AN OPENING TO RECALIBRATE MENTAL MODELS

NOT ALIGNMENT!



“blameless” is *table stakes*

Etsy

Code as Craft

Speaker Series Events About Archive

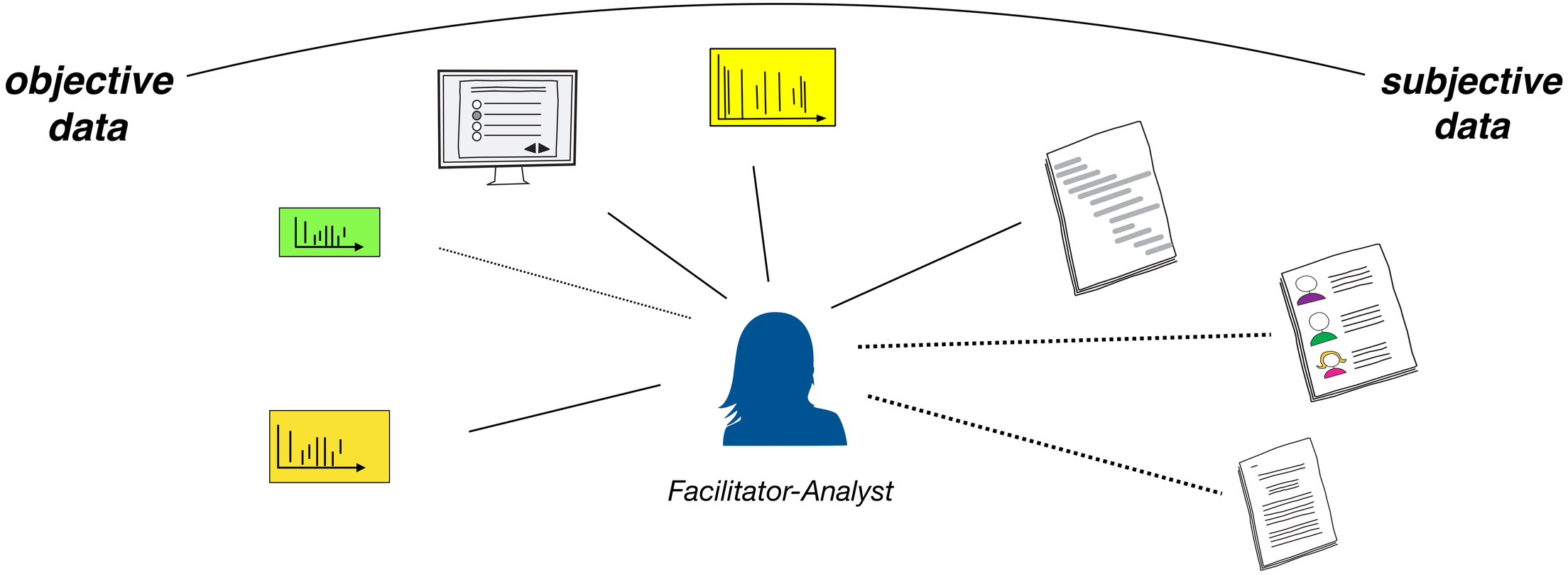
Blameless PostMortems and a Just Culture

Posted by **John Allspaw** on May 22, 2012

Last week, Owen Thomas wrote a flattering article over at Business Insider on how we handle errors and mistakes at Etsy. I thought I might give some detail on how that actually happens, and why.

Anyone who’s worked with technology at any scale is familiar with failure. Failure cares not about the architecture designs you slave over, the code you write and review, or the alerts and metrics you meticulously pore through.

more effort than typical post-incident reviews



preparation → ***debriefing(s)*** → ***post-analysis***

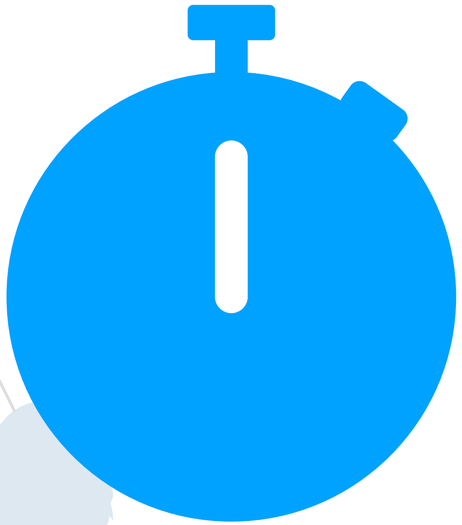
gathering data
organize
collate
analyze

group
1:1

synthesis
conditions map
annotations
dissemination

“Come on, people…we only have an hour, let’s not waste any time going over things everyone knows already!

THE CLOCK IS TICKING!”

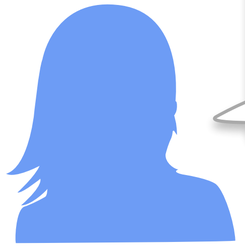


Facilitator-Analyst

Everyone Has Their Own Mystery To Solve

or

Don’t Waste My Time On Details I Already Know

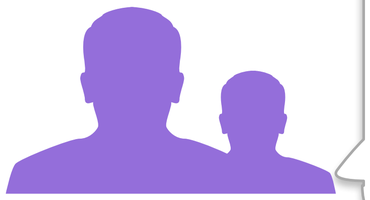


Network Engineer

This outage seems pretty straightforward - but I don’t know why they didn’t call me to help sooner?

CEO

I need answers, quick. I have three board member voicemails lighting my phone on fire and we don’t have time to waste on details that don’t matter…



Application Engineer

I’m glad we used all those feature flags to turn things off, because without them it could have been a lot worse. I just don’t understand how the database got stuck — it’s such a black box!

DBA

I understand how the database got wedged - I hope we don’t waste time going over that part. I just have no idea about how the load balancer got involved in all of this, I hope we cover that!

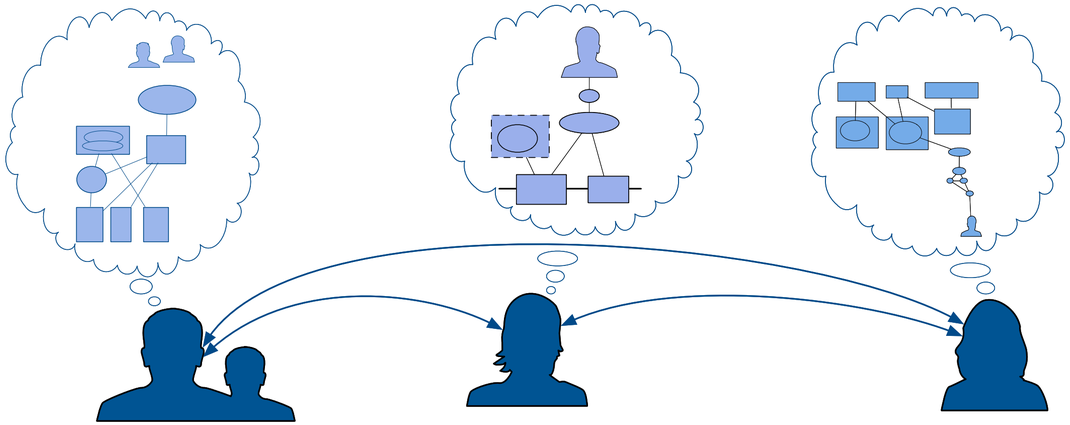
Customer Service Agent

I hope I can get a word in edgewise in this - I don’t understand how it’s so hard to give customers updates more frequently. That is the real priority here!!

all work is contextual

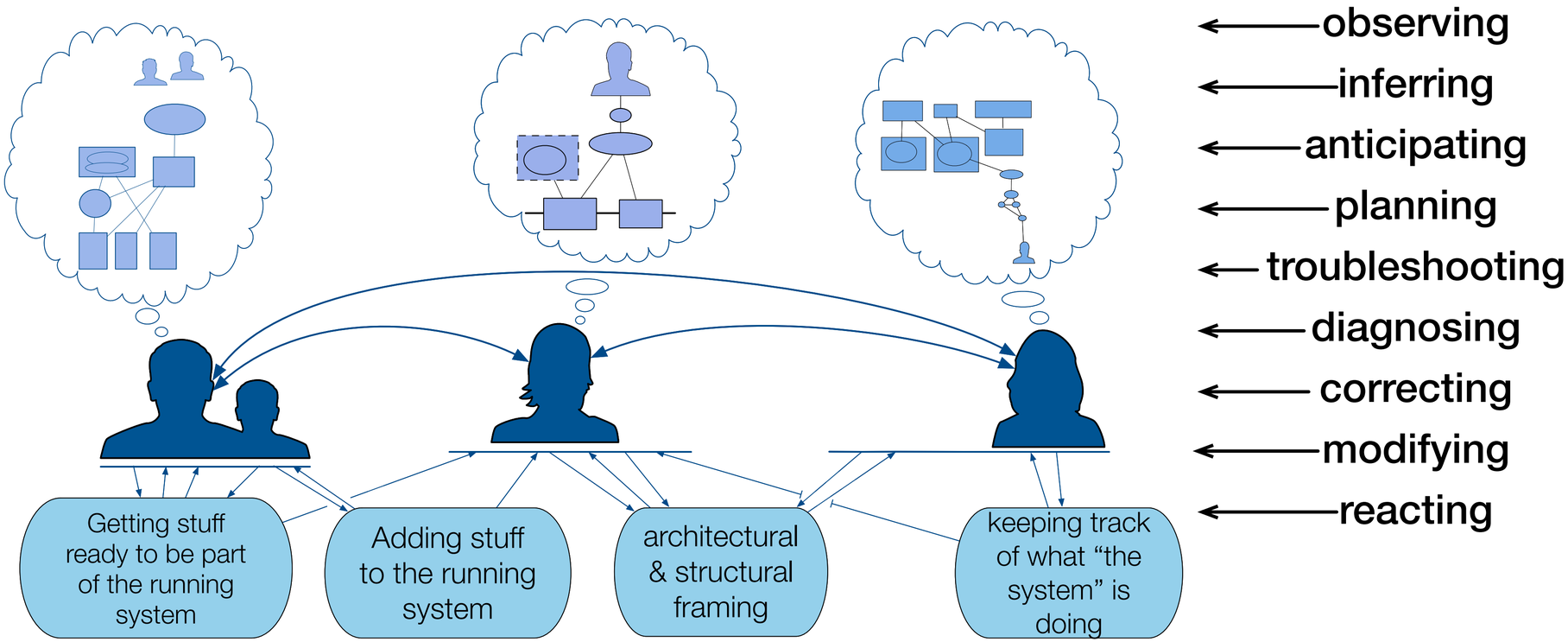
maximizing that *ROI* means:

discovering,
exploring,
***and rebuilding the context* in which work is done….**



assessments are trade-offs

contextual



all incidents can be worse

Useful at surface level:

what went wrong? how did it break? what do we fix?

Useful at a deeper level:

what are things that went into making it not nearly as bad as it could have been?

Useful at a strategic level:

how can we support, encourage, advocate, AND FUND the continual process of understanding our systems (“above the line”) in a sustained way?

Challenges For You

1. Circulate the Stella Report (http://stella.report) in your company, start a dialogue.

2. Look deeply at how you are handling post-incident reviews.

   Ask: ”What value do you think our current post-incident reviews *really* have?”

3. Will you learn more — and faster — from incidents than ***your competitors***?

Taking Human Performance Seriously

This discussion is happening…

in nuclear power

in air traffic control

in firefighting

in medicine

…

We need to do more than just acknowledge this - we need to ***embrace*** it.

What You Can Help Me With

Please spread this presentation, these ideas!

What resonated with you about this? Please come tell me!

What challenges do you face in your org? Please come tell me!

Thank You

http://stella.report

AdaptiveCapacityLabs.com

SNAFU CATCHERS