Google

Beyond Goldilocks Reliability

Narayan Desai
SRECon ‘21
October 14, 2021

Site Reliability Engineering

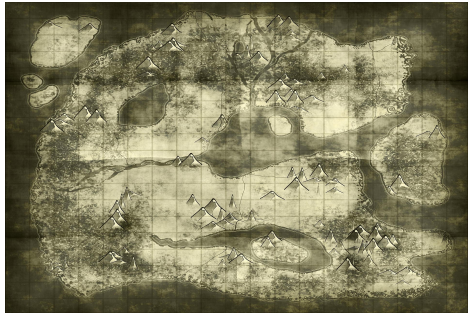
Acknowledgements

The Kraken team: Brent Bryan, Jeff Borwey, Angus Fong, Navaid Abidi

Adam Kramer, Christian Webb, Chris DeForeest, Julius Plenz

Eric Brewer, Niall Murphy, Nicole Forsgren, Jez Humble, Lorin Hochstein, Chris Heiser

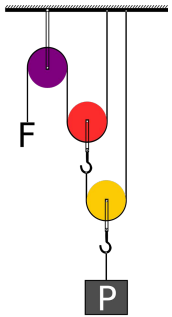
Our Reliability Approach



Analytics provide a map

Help us to understand where customers need us

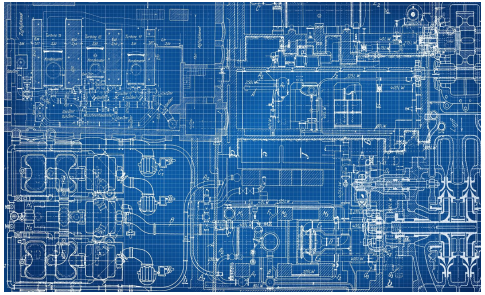
Inform **systematic investment** of effort



Analytics tools make engineers more efficient

Provide better service to customers

Scale sublinearly

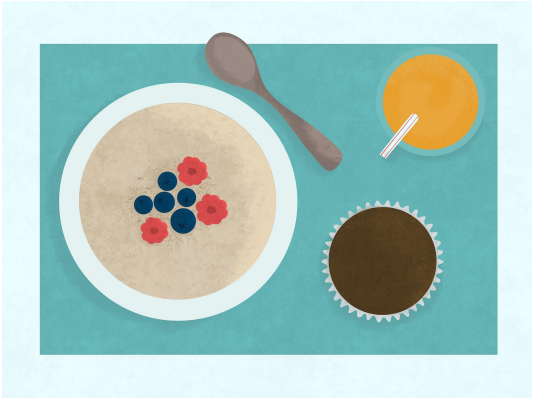


Precise analytics reveal the dynamics of reliability phenomena

Models enable reliability **engineering**

Goldilocks Reliability

Goldilocks



Goldilocks Reliability

Define some SLIs

Measures can be anything. Counts, real-numbered statistics like latency or resource consumption.

Choose “Just Right”

“Just right” describes the line distinguishing between expected behavior and problems.

Profit!

Everything is a 2 bucket histogram! Bounds can be set using ratios! (Cl|Hil)arity ensues.

“All models are wrong, but some are useful.”

George E. F. Box

“... and some are dangerous.”

Lorin Hochstein

Load Bearing Assumptions

Just Right makes sense

Metrics need to be distributed such that the idea of an acceptable range for measures is a useful concept. We also need to be able to formulate an answer.

There is one answer

Even if a metric is properly distributed, it may not be aggregated such that these patterns can be discerned. Differences between customers or workloads can invalidate this assumption.

We know the questions to ask

Individual Goldilocks measures are narrow, so many must be used in conjunction to understand if a service is “working”.

The answers don’t change

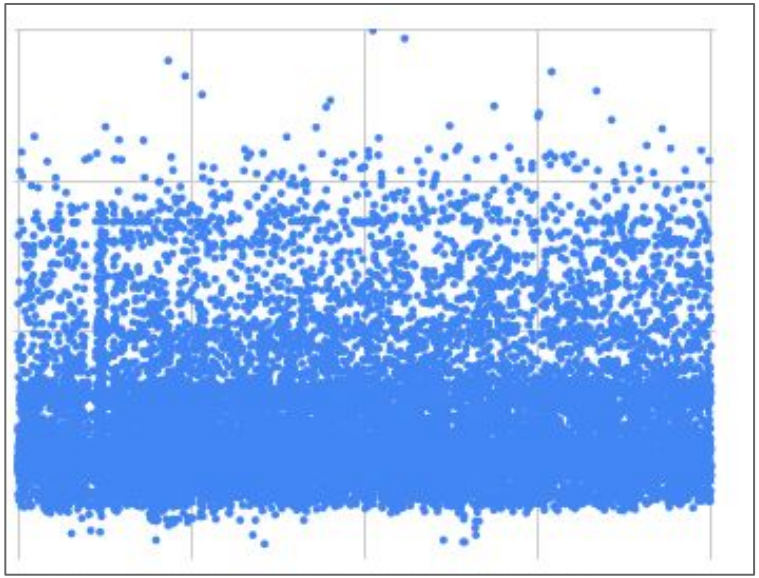
Goldilocks measures are highly sensitive to calibrated thresholds. Changes can result in misleading assessments of reliability.

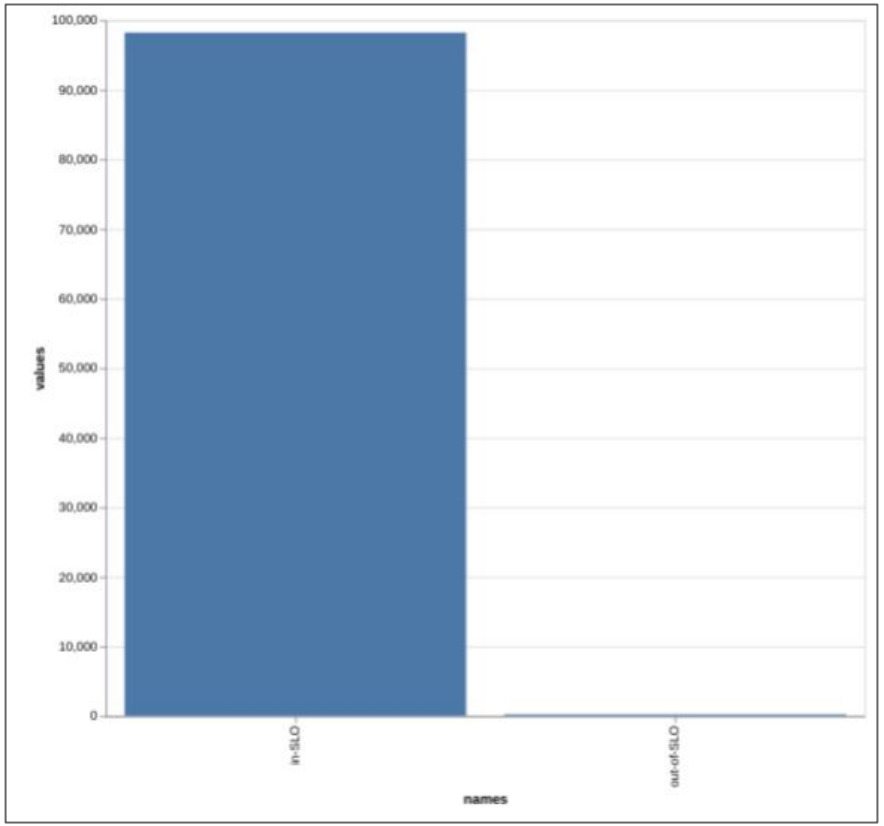
The Problems with Goldilocks Reliability

Practical Porridge Problems

- No model of reliability
  - Rube Goldberg *analytical* machine
  - …with brittle outcomes
- “Just right” is nigh-impossible to specify in many cases
  - Nature of metrics
  - Overbroad aggregation
- Each Goldilock metric provides a narrow window into behavior
  - You don’t know what you don’t know
  - ..and you don’t how much you don’t know
- Maintenance cycle for calibration is unspecified
  - Performance shifts, dependencies change
  - When should things change?

The Trouble with Thresholds





Mo’ Porridge Mo’ Problems

- Calibration implications are high-stakes
- Requires many decisions be made
  - People make 10-30 errors per 100 decisions
- We have no basis to judge quality
  - Nevermind a quality control process
- No support for deeper insights
  - We can’t abstract from this
  - We can’t even see critical reliability phenomena
- This process is insidious
  - It looks like a human process failure

We must do better!

Beyond Goldilocks Reliability

Reliability

Availability

That a service is there when you need it.

Performance

How effectively work is performed.

Correctness

Does a service do what it is supposed to.

“All models are wrong, but some are useful.”

George E. F. Box

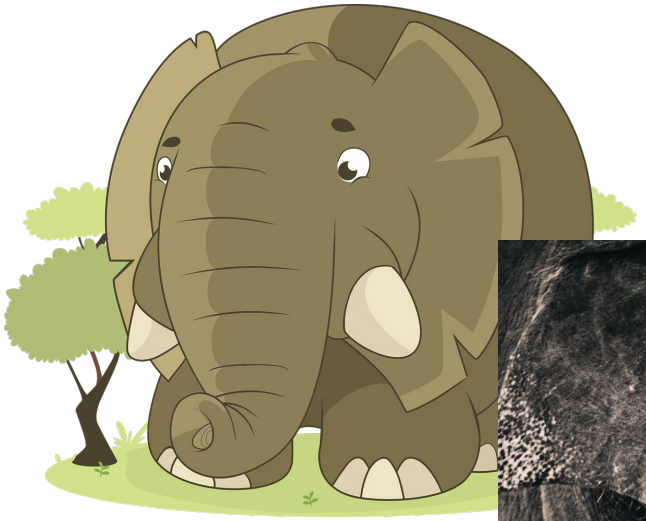
“... and some are dangerous.”

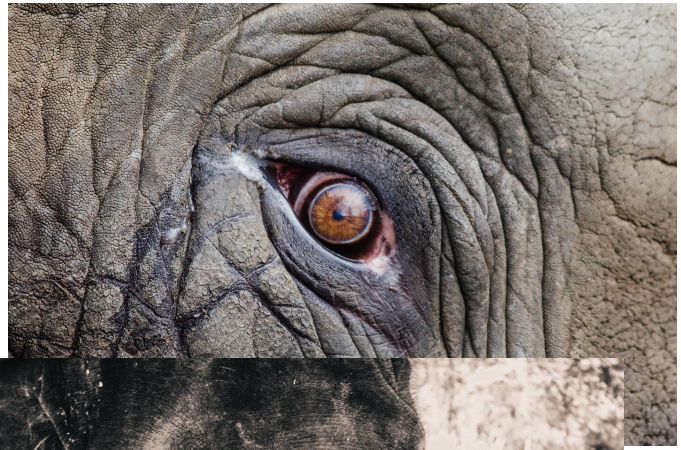
Lorin Hochstein

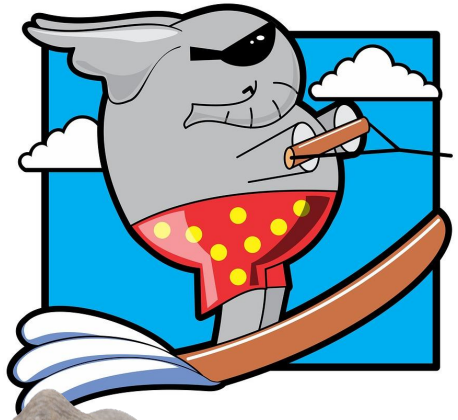
Make More Models!

- Mathematize your intuition
- Backtest and refine
- Understand your systems and share your methods

Model Elephants



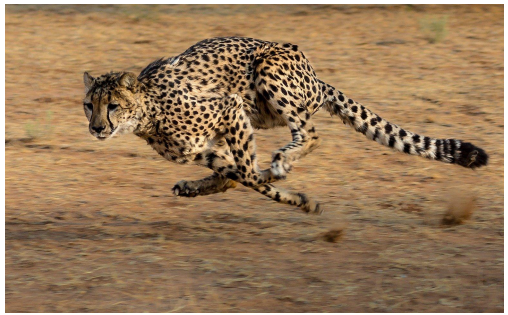




Reliability, modeled as Stationarity

Availability

Errors are independent and identically distributed across time and space.



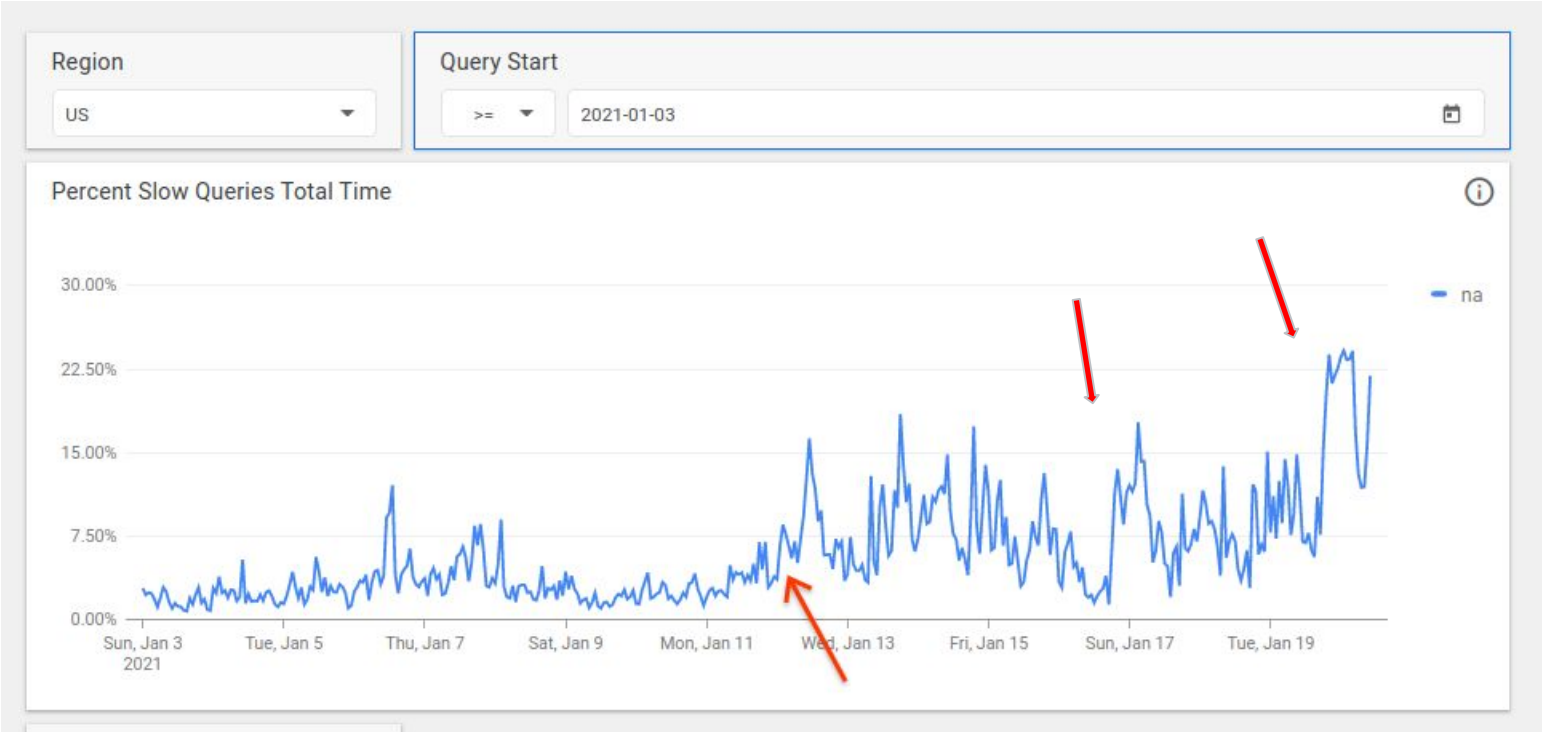
Performance

Performance is consistent across long time windows.

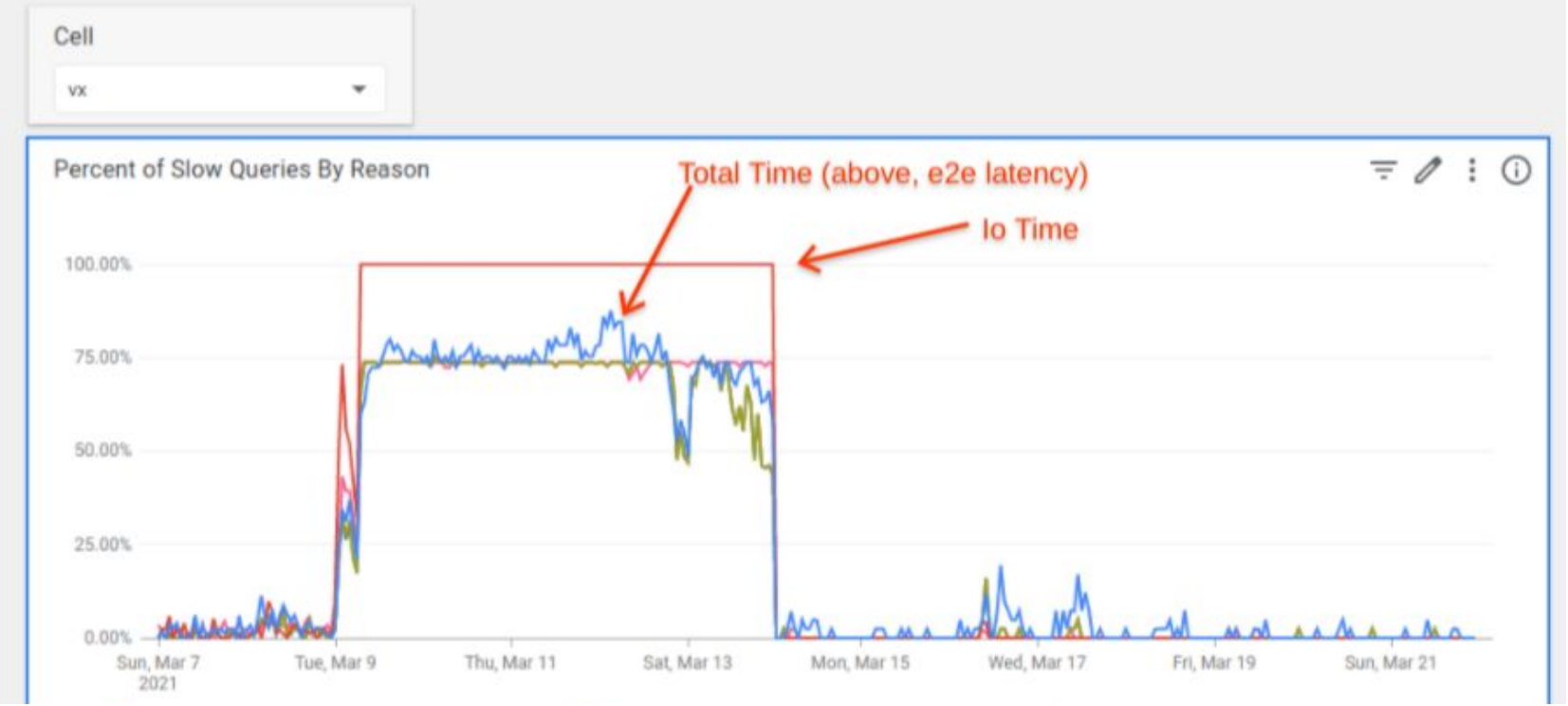
Correctness

Service produces the same results over time, modulo bugs.

Stationarity Works!



Hierarchical Diagnostics



Stationarity Exposes Reliability Phenomena

- Sub-critical performance shifts
- Slow-building reliability incidents
- Performance regressions
- Subsystem failures
- Provisioning issues
- Isolation failures
- Customer pain

Tantalizing Capabilities

- De Novo impact assessments
- Proactive reliability interventions
- Measurement of ambient instability
- Mechanical Diagnostics
- Data-driven prioritization of reliability investments
- Direct detection of customer pain!

Conclusions

- We need better ways to think about reliability
  - Concise terminology
  - Well articulated models
  - Starting with interpretation -> prediction
- Make more models!
  - Try this at home, with your friends
  - Validate them
  - Share your ideas, figure out what works and doesn't, and why
- Maintain a healthy skepticism of all models
- Stationarity provides a great new lens to analyze reliability
  - Can now see previously invisible reliability phenomena
  - New tools
  - .. and are starting to develop insights about the nature of reliability
- We heading toward a new phase of reliability engineering