How Computers Help Humans Root Cause Issues at Netflix

SETH KATZ
QCON NEW YORK, 2018

NETFLIX

Hello!

- Seth Katz
- 5 years at Netflix
- Focused on improving Netflix operations
- Share what we’ve learned on applying machine intelligence to operations

I got paged!

Funny Tweet - Serious Situation

Rachael Whalen
@rachael_whalen

Follow

my Netflix isn't working....what am I supposed to do with my life now

3:53 PM - 3 Feb 2015

NETFLIX

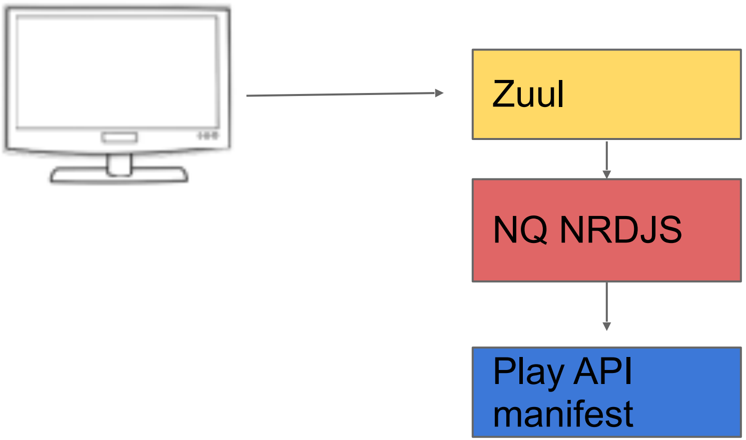
Agenda

- Netflix operations
- Approach and challenges to ML in operations
- Anomaly detection
  - Real-time
  - Near real-time
- Visualization and making it practical
- Reflections and takeaways

What if we get this page?

Android devices that can't play a movie exceeds 1%

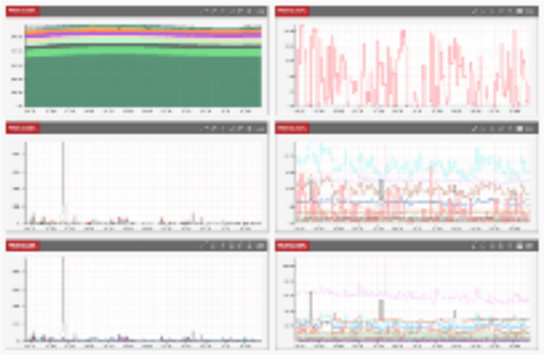
Microservices



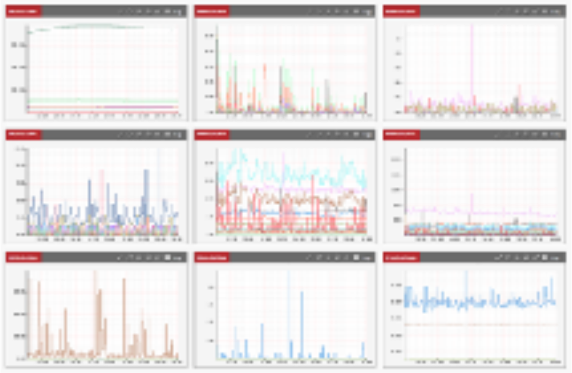
NETFLIX

Android

Zuul

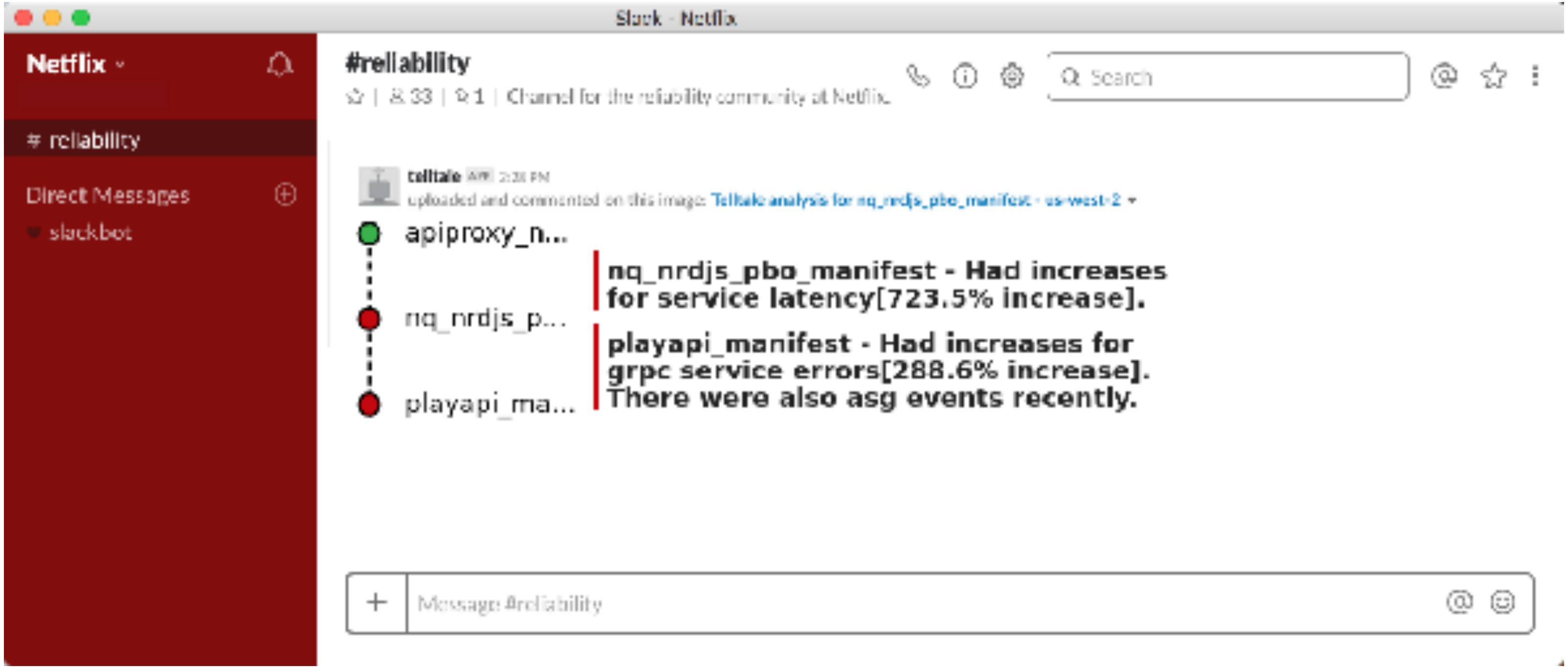


NQ NRDJS



Play API

Slack Message



Why is diagnosing pages hard

It’s 3am in the morning - are you thinking clearly?

Maybe you understand your microservice?

What about all the other services involved?

What about their push schedules in every region?

Hard problem - how to build a minimum viable product ?

Simple, Principled, Robust Anomaly Detection

Principled algorithms have guarantees you can use to reason about for any data pattern

Simple algorithms that are very easy to implement.   Don’t need major frameworks, GPUs, Python, etc.

Wouldn’t be great if ...

Golden Age of AI

Why do Star Trek robots seem near, but Lost In Space robots seem further into the future

AI challenges in operations

Limited examples of outages

Cause and effect

Tribal knowledge

More AI challenges

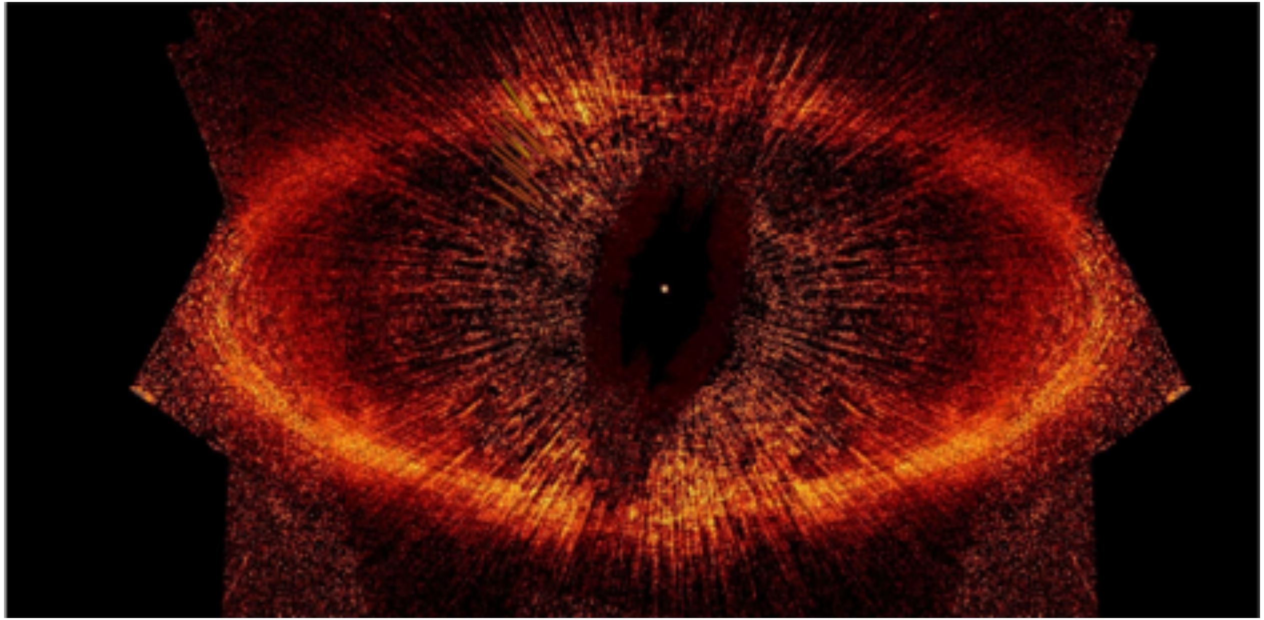
Curse of dimensionality

Rapidly changing ground truth

Generalization to new problems

So what can we do? - Real-time root cause detection

Root cause for the oracle



Real world example

Timeline

- 11:50:15 - Region failover from us-east-1 -> eu-west-1
- 11:51:12 - Service A timeouts increase 243% in eu-west-1
- 11:51:14 - Android movie errors increase 840%

Complete picture of what happens - time suggests causality

Victory?

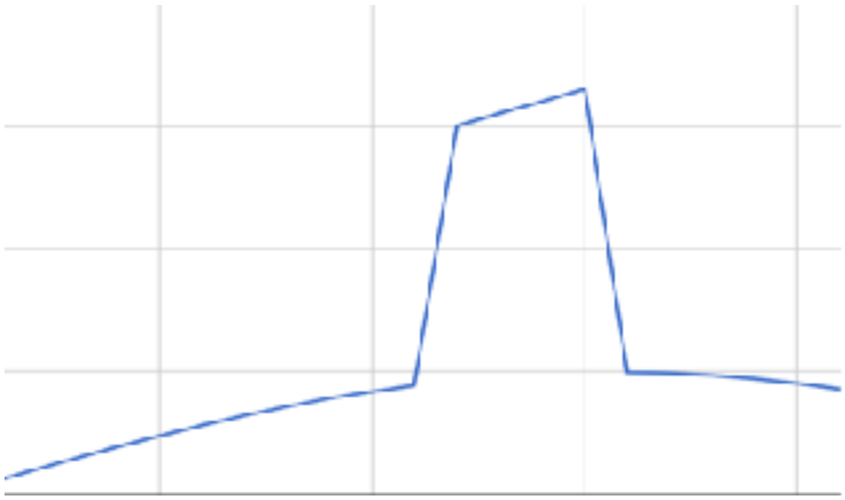
We can only do this on metric subsets

- Signals usually relatively stable and slow changing
- Signal with up to date event source
- Signals with rapid updates, many samples.

How can we detect scalar anomalies?

Scalar Anomaly Signal

Android error rate



- Anomaly very clear to humans
- Limited data needed
- Historical trend unnecessary
- Recovery also clear
- Principled signal analysis possible

What’s normal?

Median on a Stream.

If Incoming > Median:

Median = Median + Alpha

Else:

Median = Median - Alpha

- **Alpha can be adjusted if consecutively on one side**
- **Need rapid data updates for timely convergence.**

What’s abnormal?

Hoeffding Bound

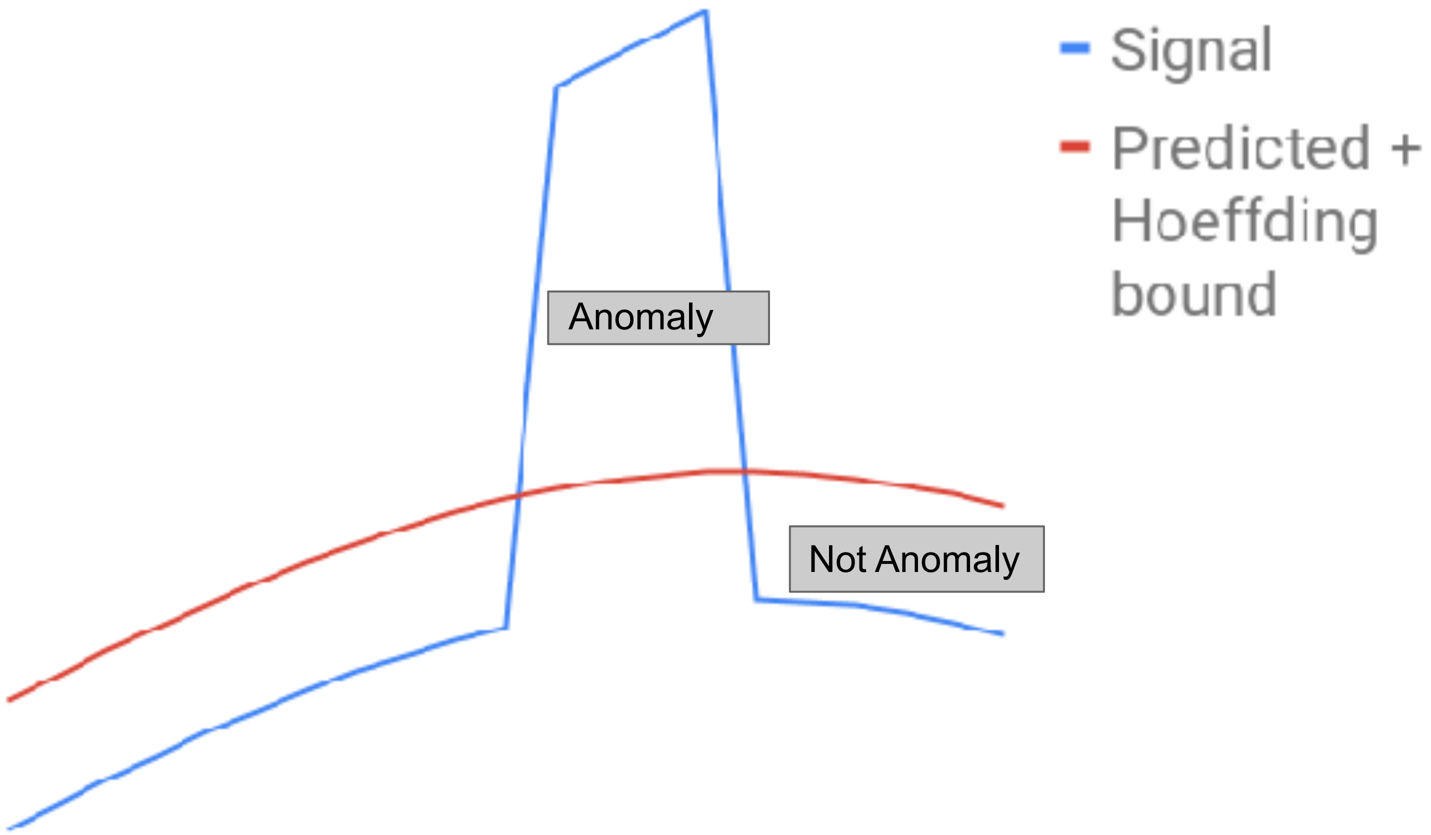
- Is the next data point from the same distribution as sample?
- Can I guarantee it is the same distribution with a desired level of confidence?
- Do I need to assume my data is normally distributed (aka Gaussian)?
- Hoeffding Bound

Hoeffding Bound Very Simple

- n=sample size
- d=desired certainty, eg .01 for 99%
- r=sample range, ie (max -min)

$$\sqrt{\frac{r^2 log(\frac{1}{d})}{2n}}$$

NETFLIX

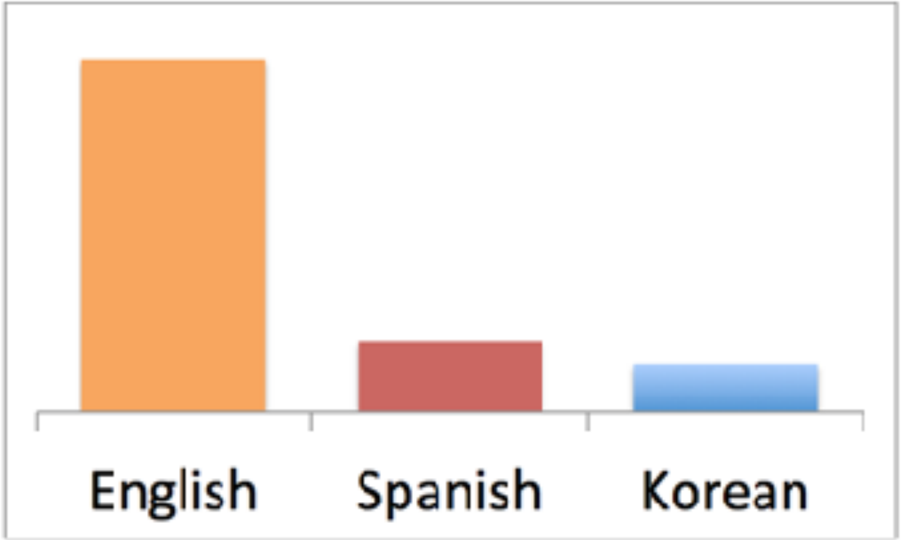


NETFLIX

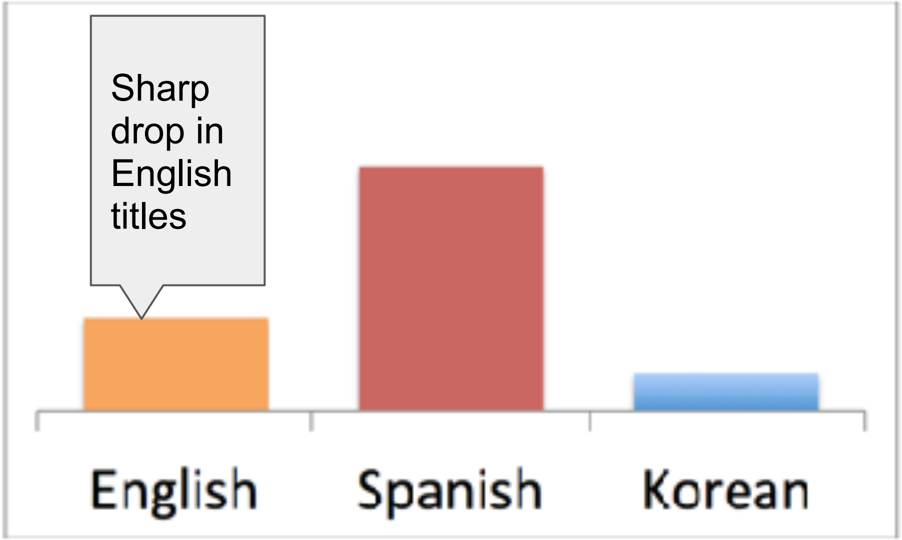
Another problem - detecting a bad config push?

Consecutive histogram snapshots

11:10:15



11:10:20



Is there principled way to measure difference between histograms?

Information Theory

Entropy - Average Information

$$H(X) = -\sum_{i=1}^{N} p(x_i) log(p(x_i))$$

Fair Coin

$$-\sum_{i=1}^{2} \frac{1}{2} log(\frac{1}{2}) = -\sum_{i=1}^{2} \frac{1}{2} \times -1 = 1$$

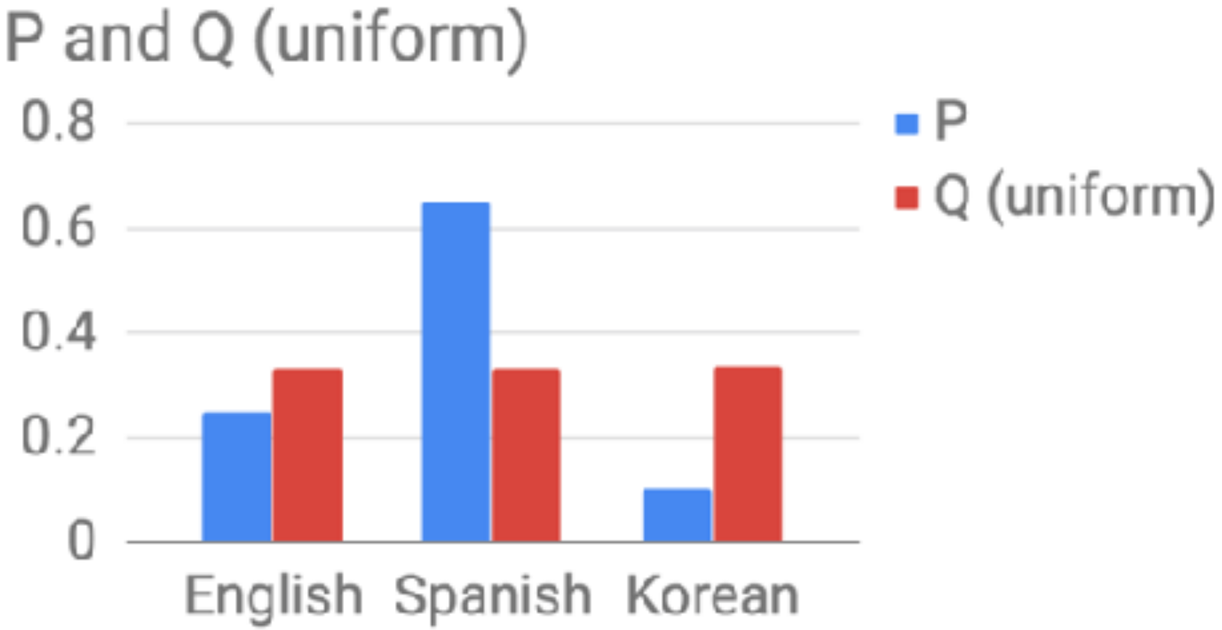
9-1 Biased Coin

$$-(.1 \times log(.1) + .9 \times log(.9)) = 0.37$$

NETFLIX

How much entropy do we lose if we estimate histogram with wrong probability distribution?

Uniform Distribution Info



NETFLIX

KL Divergence

Minor Formula Change for Entropy difference

- **Entropy**

$$H(X) = -\sum_{i=1}^{N} p(x_i) log(p(x_i))$$

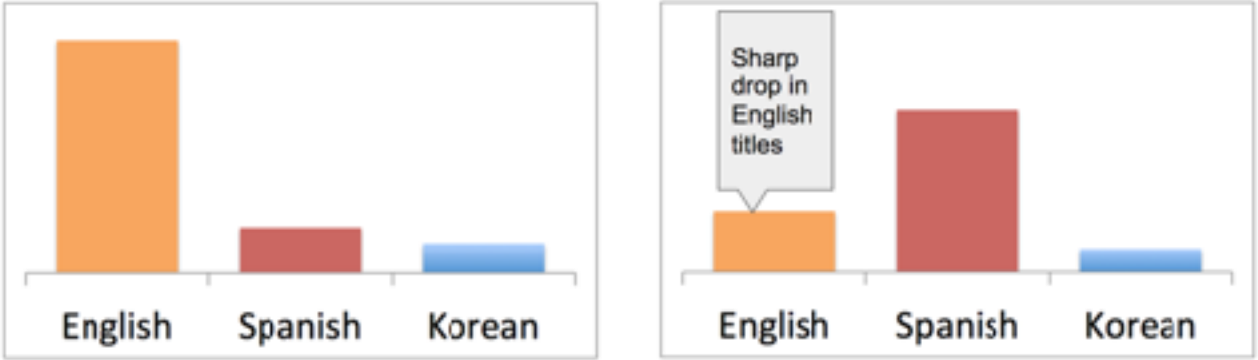
- **KL Divergence**

$$D_{KL}(p||q) = -\sum_{i=1}^{N} p(x_i)(log(p(x_i) - log(q(x_i))$$

Is KL divergence a good score?

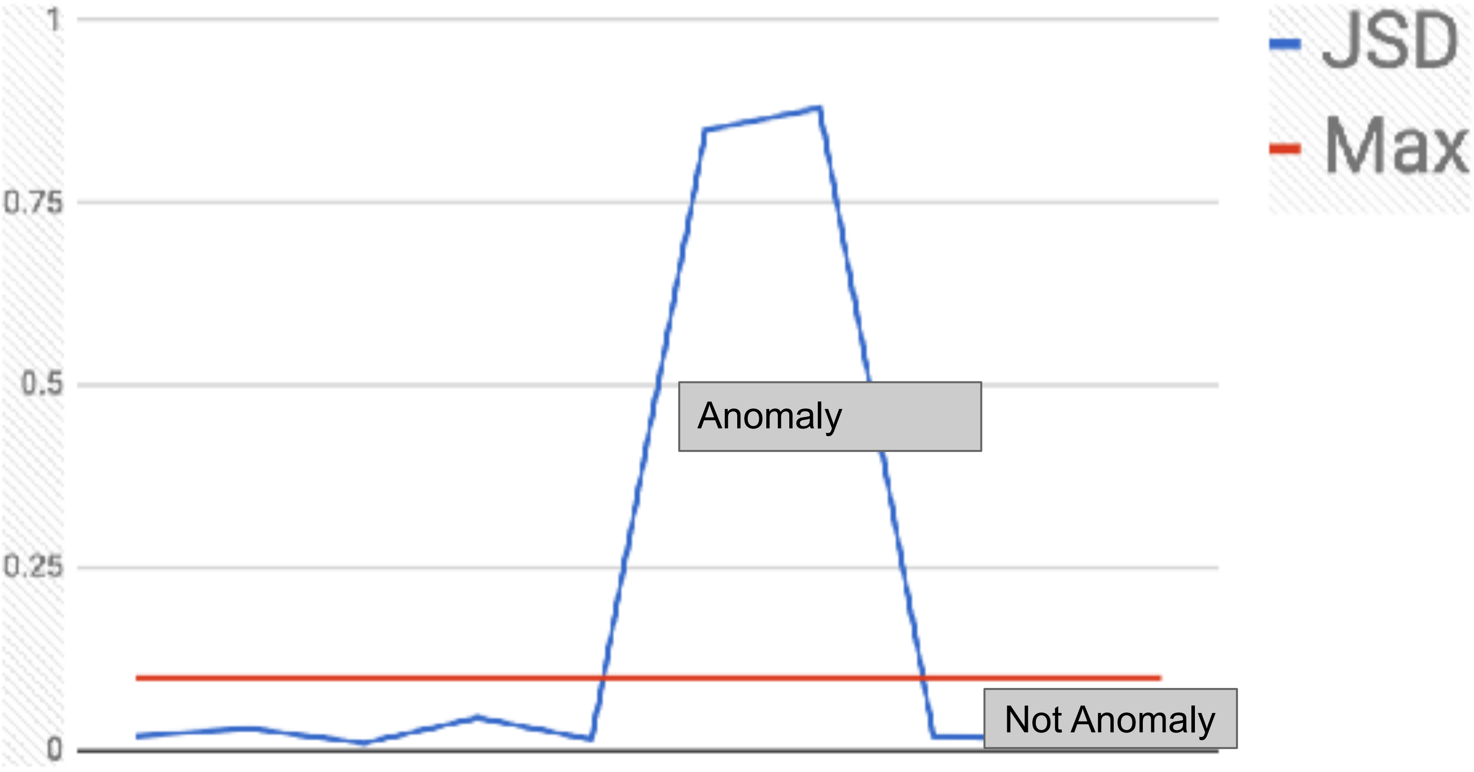
Jensen Shannon Divergence (JSD)

- Not symmetric?
  - Take KL divergence in both directions and add
- No upper limit?
  - Normalize it

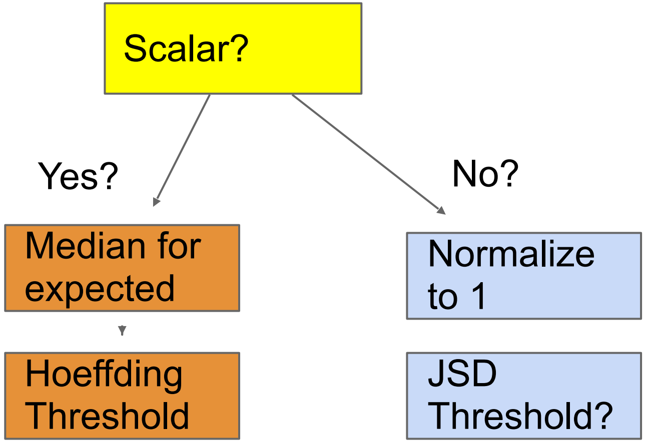


NETFLIX

JSD Anomaly Threshold Algorithm



Real time Algo Recap



How to communicate anomalies?

Example

- Android movie errors increase 840%?
  - Increased from what?
  - Why not use z-score (number of standard deviations from mean)?

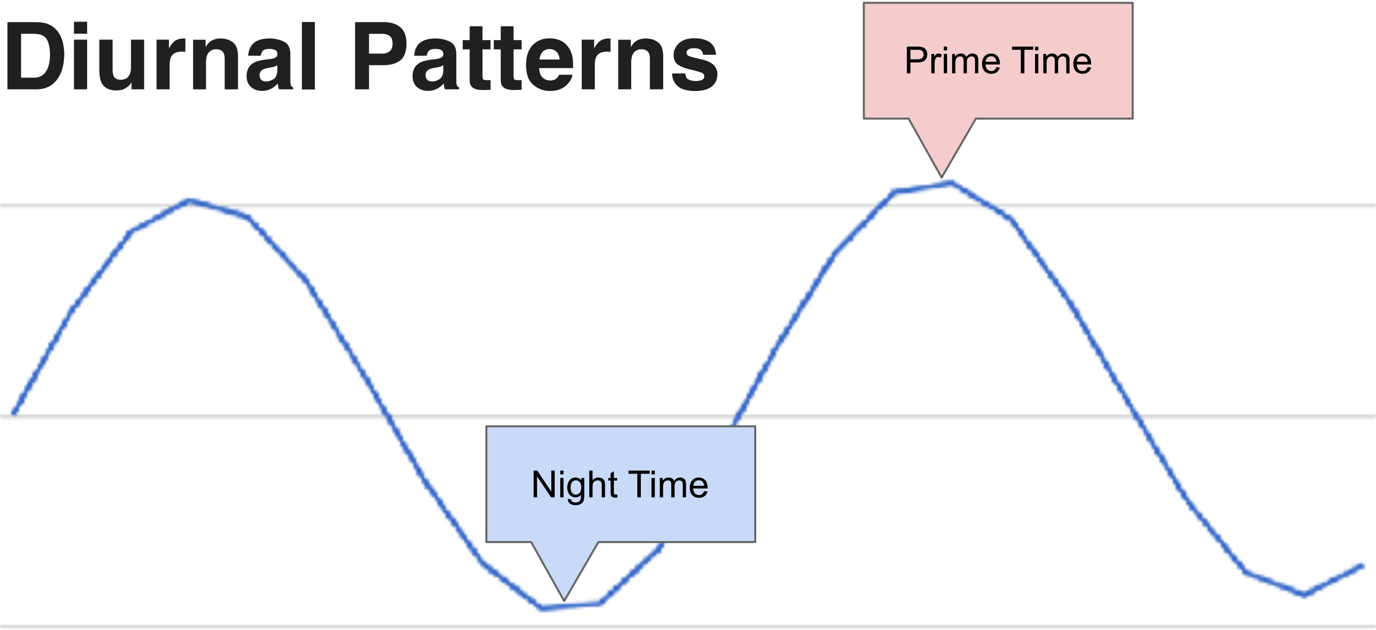
This is your brain on Pager Duty

Intuitive messages beat mathematically precise ones

What about nearly real-time signals?

More Time and More Data

Diurnal Patterns



Drawbacks

- Usually better for mean time to resolve than mean time to detect
- Less precise timing
- Use correlation, but humans decide cause vs effect

Suspicious Things

Error Code 1234 is High?

- **Is there an attribute over represented for sessions with 1234 error code?**
  - Device?
  - UI version?
- **Baseline Essential**
  - What if only one UI version actually produces error code 1234?

How do we identify significant change from baseline?

Two-Way Contingency Table

	Error 1234	**UI Version 0.0.1**
BaseLine	1000000	10000
Now	100000	1150

Use Chi-Squared test

Contingency Tables Fail

- Yes/No are past and present the same
- Chi-squared says significant, 99.999% confidence
- Netflix is always changing

Bonferonni’s principle

Eventually right by chance
if you ask enough!

Getting Correlation Right

- Contingency tables don’t work
- Convert it to a time series problem

Why would time series work when contingency tables fail?

Sensitivity

- Chi-squared test is so sensitive because of very large samples
- Number of time windows much smaller - significance tests work on smaller sets

Correlation Windows

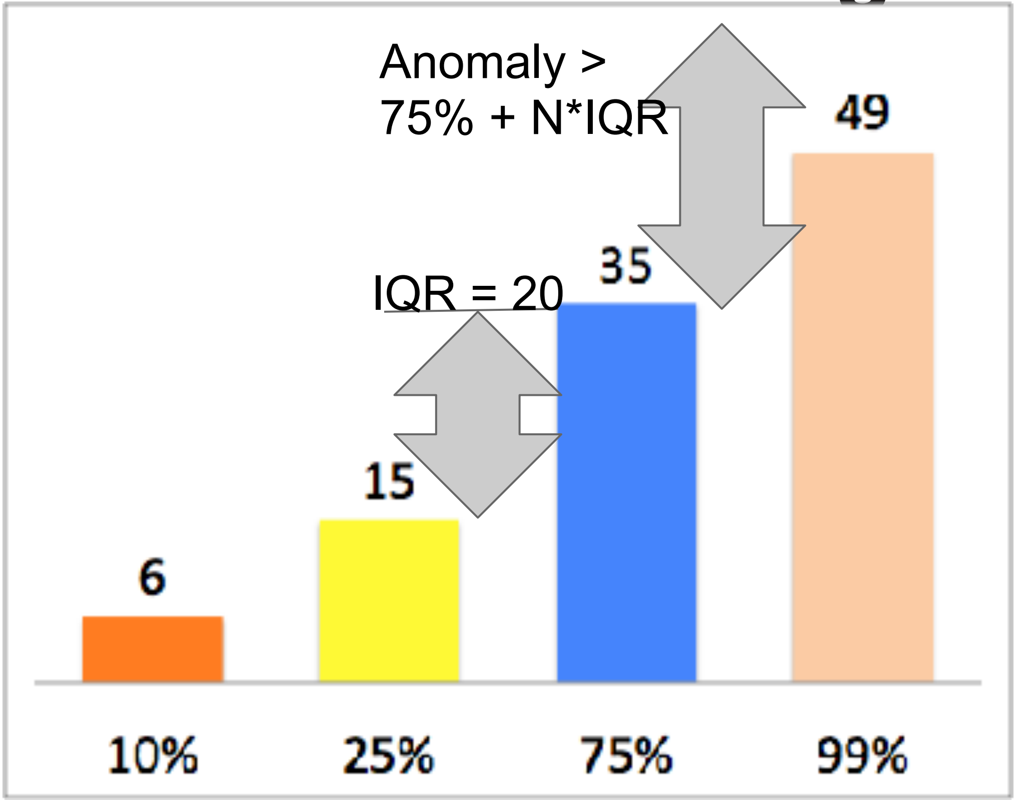
Time Window	**Pearson Correlation Score Error 1234 and UI Version 0.0.1**
10am-10:30am	.18
10:30-11:00am	.2
11-11:30am	.25
11:30am-12pm	***.95***

Significant Change?

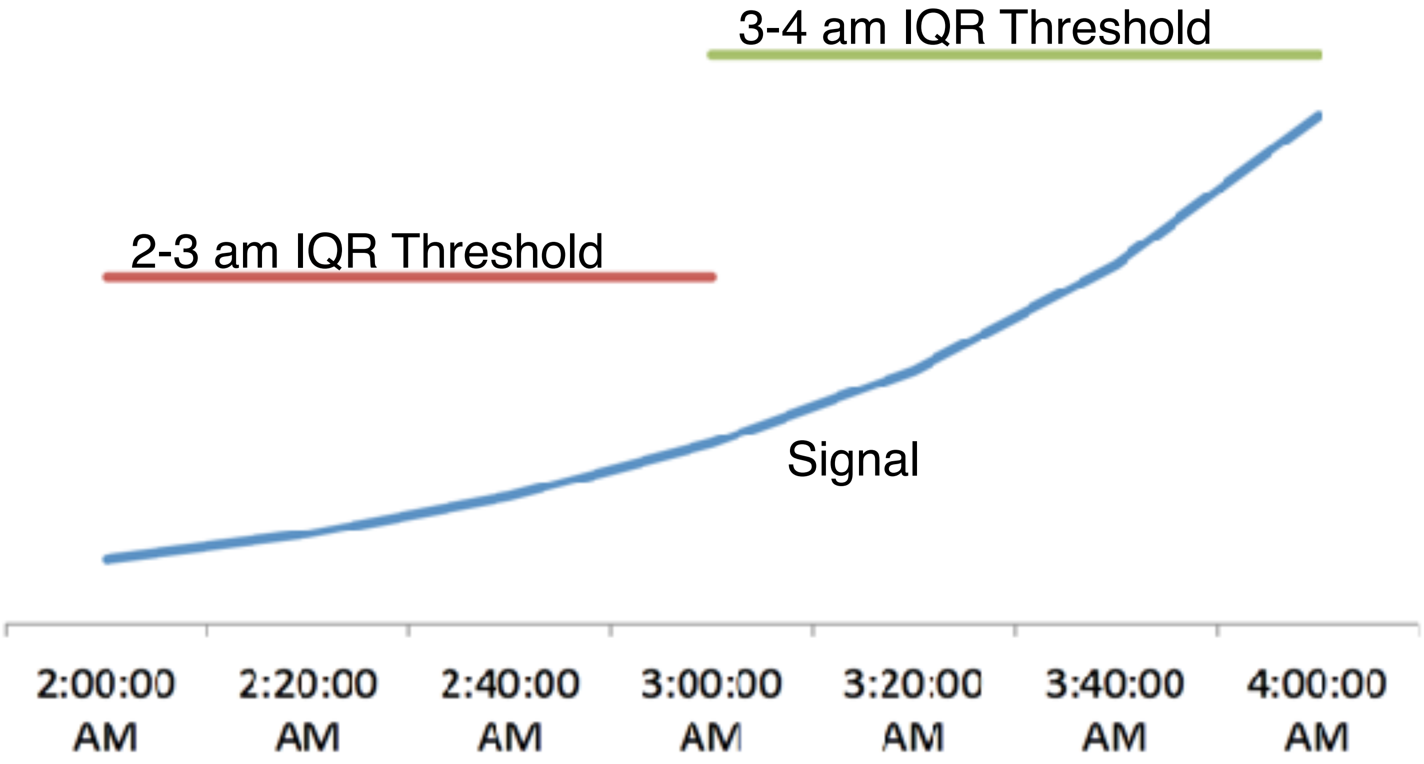
- **Mann-Whitney U Test on correlation values. (not Student's t-test)**
  - No Gaussian assumption involved
- **Works best after human determines present is "interesting"**
  - Eg, run after an alert fires

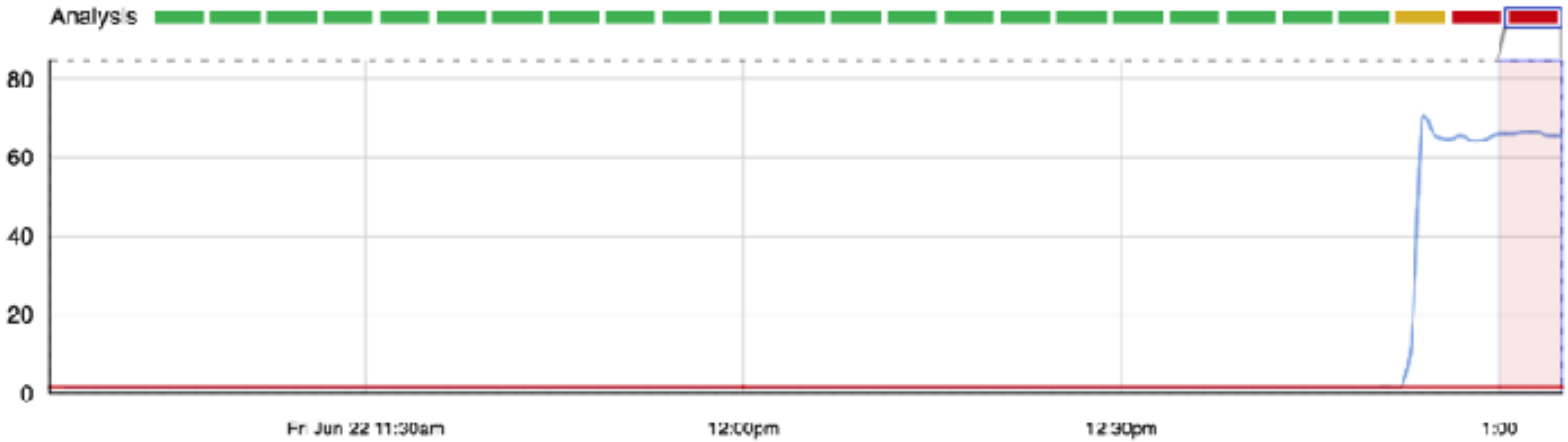
Anomaly detection for near real-time

InterQuartile Range



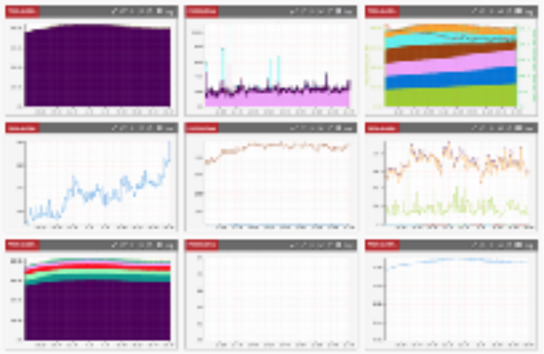
Near real-time anomalies



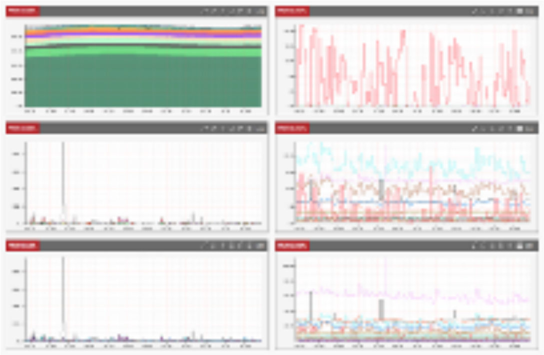

Analysis
80
60
40
20
0
Fri Jun 22 11:30am
12:00pm
12:30pm
1:00

Displaying anomalies in context

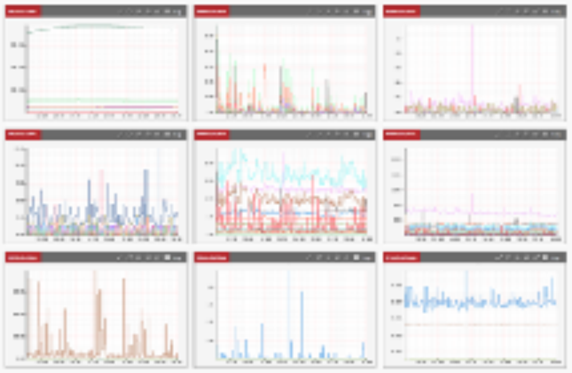
Android



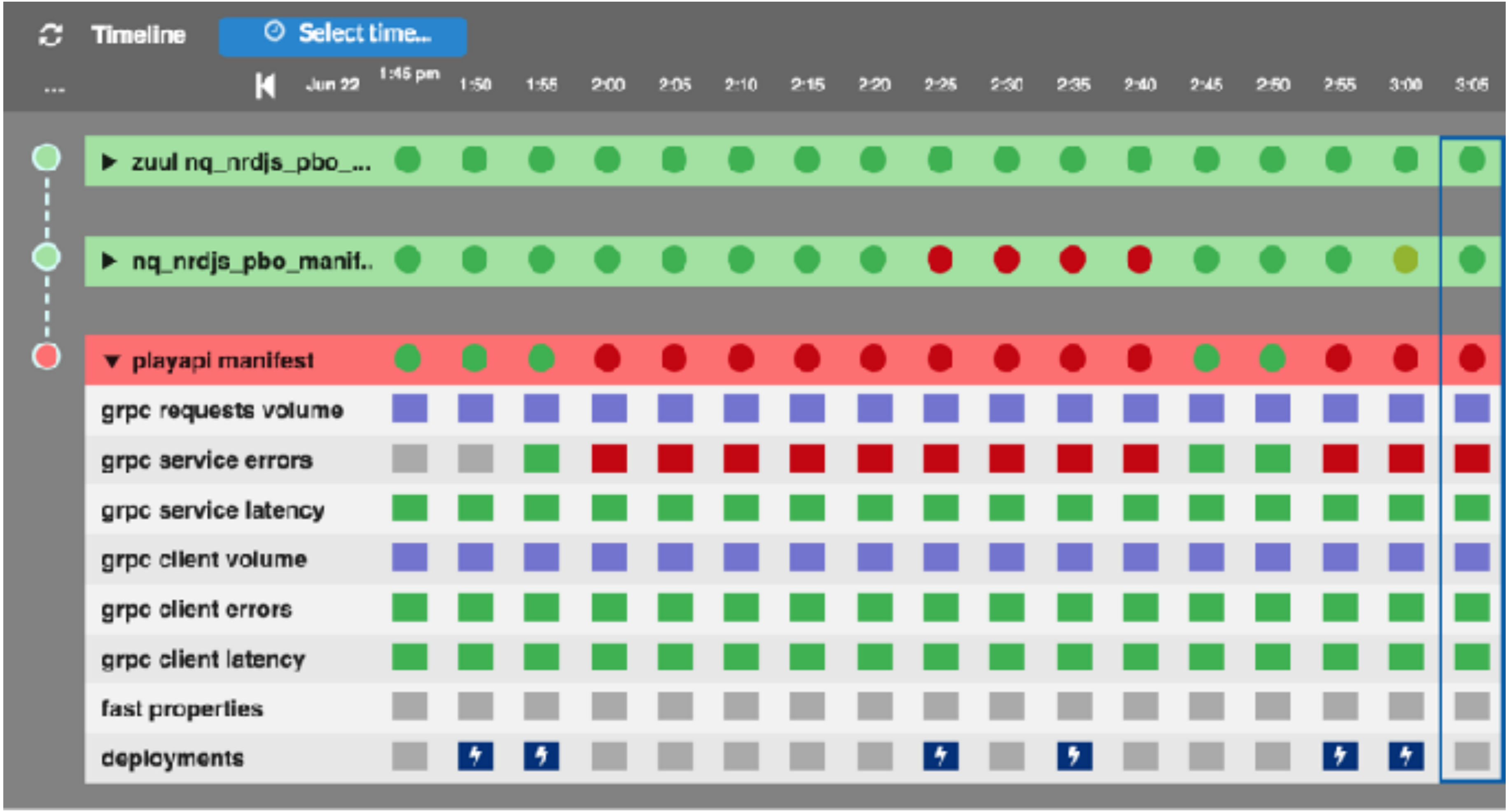
Zuul



NQ NRDJS

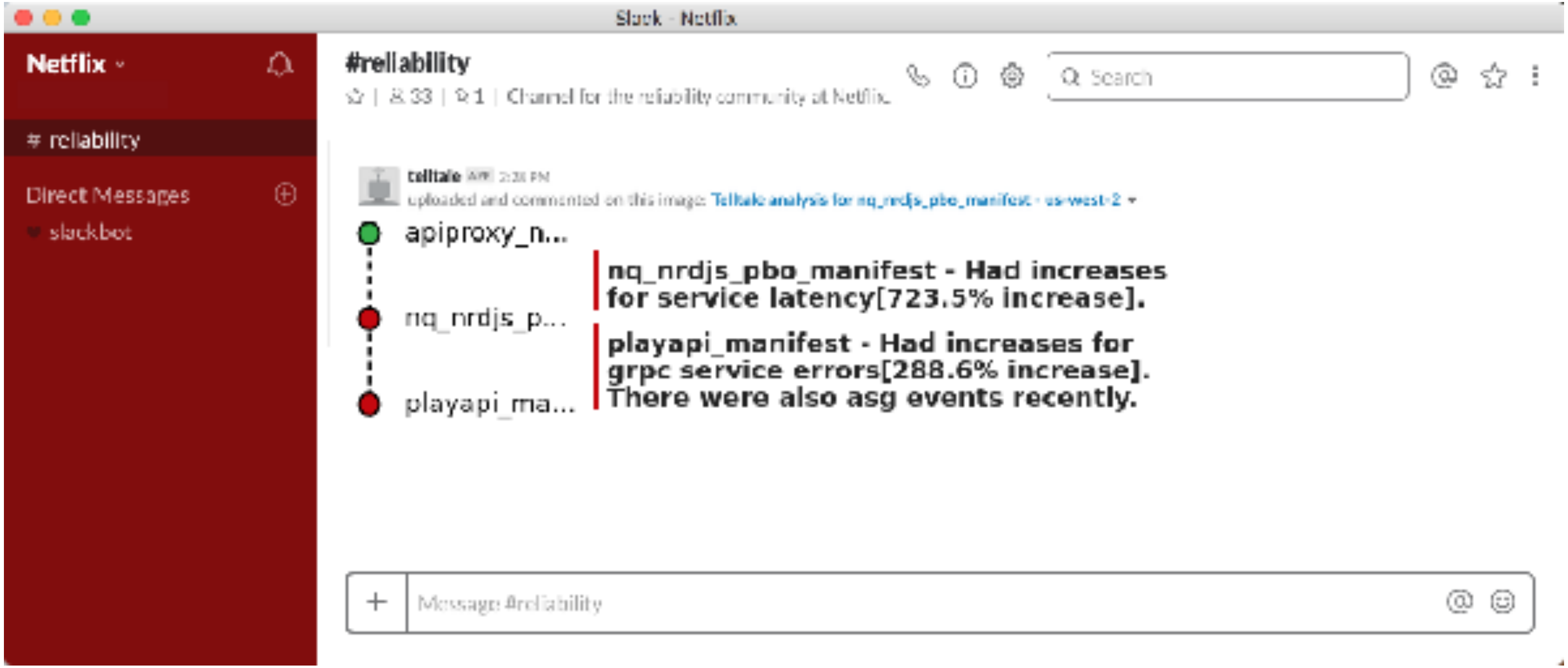


Play API



Visualization and making it practical

Summary on Slack



Reflections and Takeaways

Back to basics - simple statistics

- Scikit Learn and Tensorflow might be overkill, at least for these algorithms
- Human curation reduces scope so we don’t need a Danger Will Robinson intelligence

Real time vs Near real time

Real time

- Timing suggests causality
- Useful for mean time to detect
- Careful choice of metrics needed

Near real time

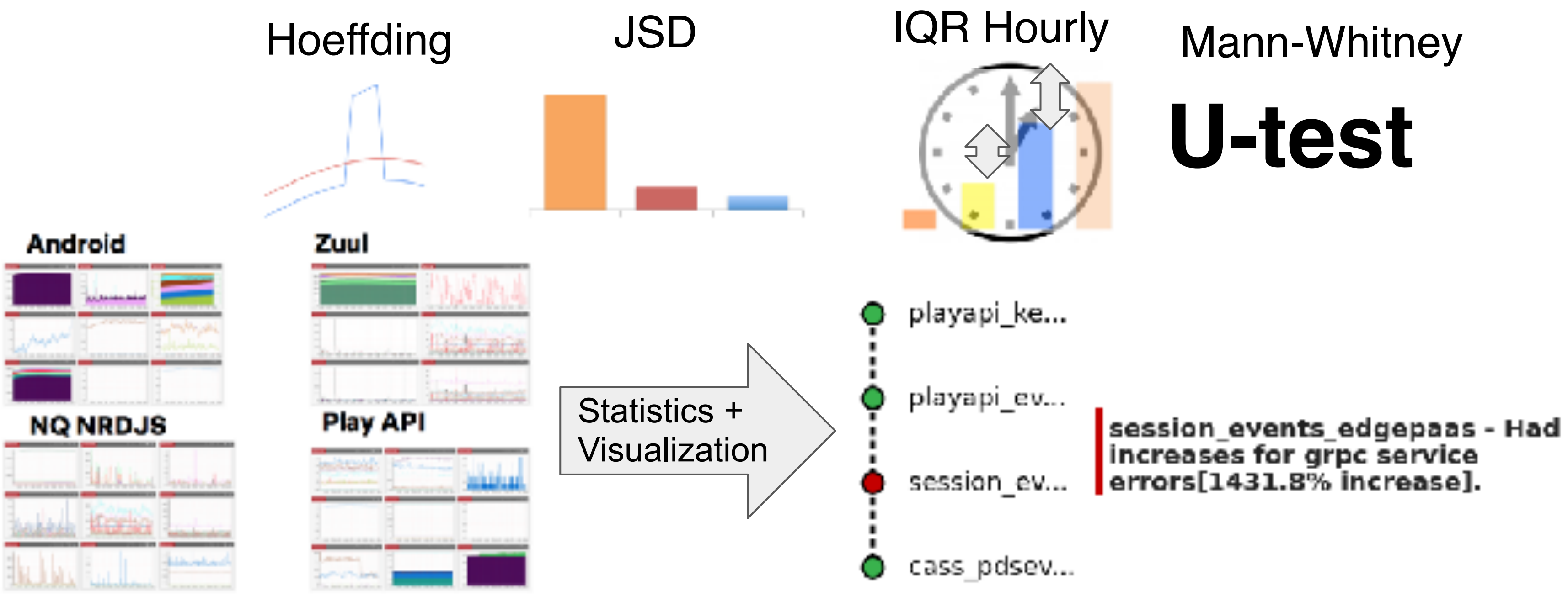
- Cause requires correlation
- Humans assign cause and effect
- More granular metrics
- Useful for mean time to resolve
- Diurnal pattern improved predictions

Get correlation right

- Contingency tables don’t work
- Correlation and Mann-Whitney U test works pretty well

A Summary Incident Approach

Android errors increased 850 percent?



More Information, Q&A

Team

https://medium.com/netflix-techblog/lessons-from-building-observability-tools-at-netflix-7cfafed6ab17

Me

https://www.linkedin.com/in/katzseth22202

Thank you.

NETFLIX