


Sifting through the noise: behavior change or machine error?

Sophia Vargas, Research Analyst
Open Source Programs Office @ Google
@Sophia_IV 

Abstract

Data collected around open source projects is notoriously messy, and sometimes it just doesn't smell right. So - how can you tell if you are looking at noise or erratic human behavior? When talking to every contributor is not an option, analysts must take more systematic approaches to verify and interpret results. This talk will share examples of confusing or nonsensical values, and the steps taken to investigate the accuracy of the results.

This talk is written for analysts new to working with data in and around open source projects, with a focus on 'middle data': too large for individual inspection and too small for robust modeling. Messy data requires creative approaches: I hope that this talk will inspire others to discuss and share methods they've used to make sense of strange data.

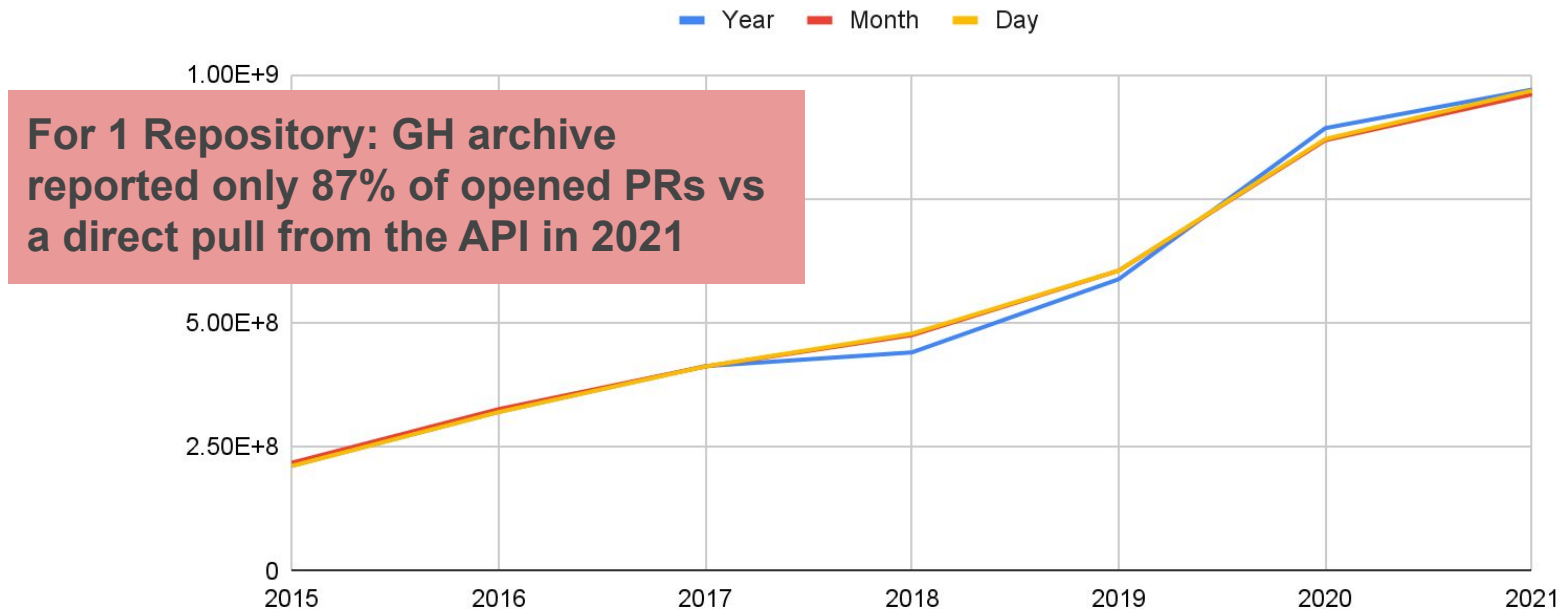
How much do you trust your data?



Accuracy vs Precision

Ruling: Flakey APIs

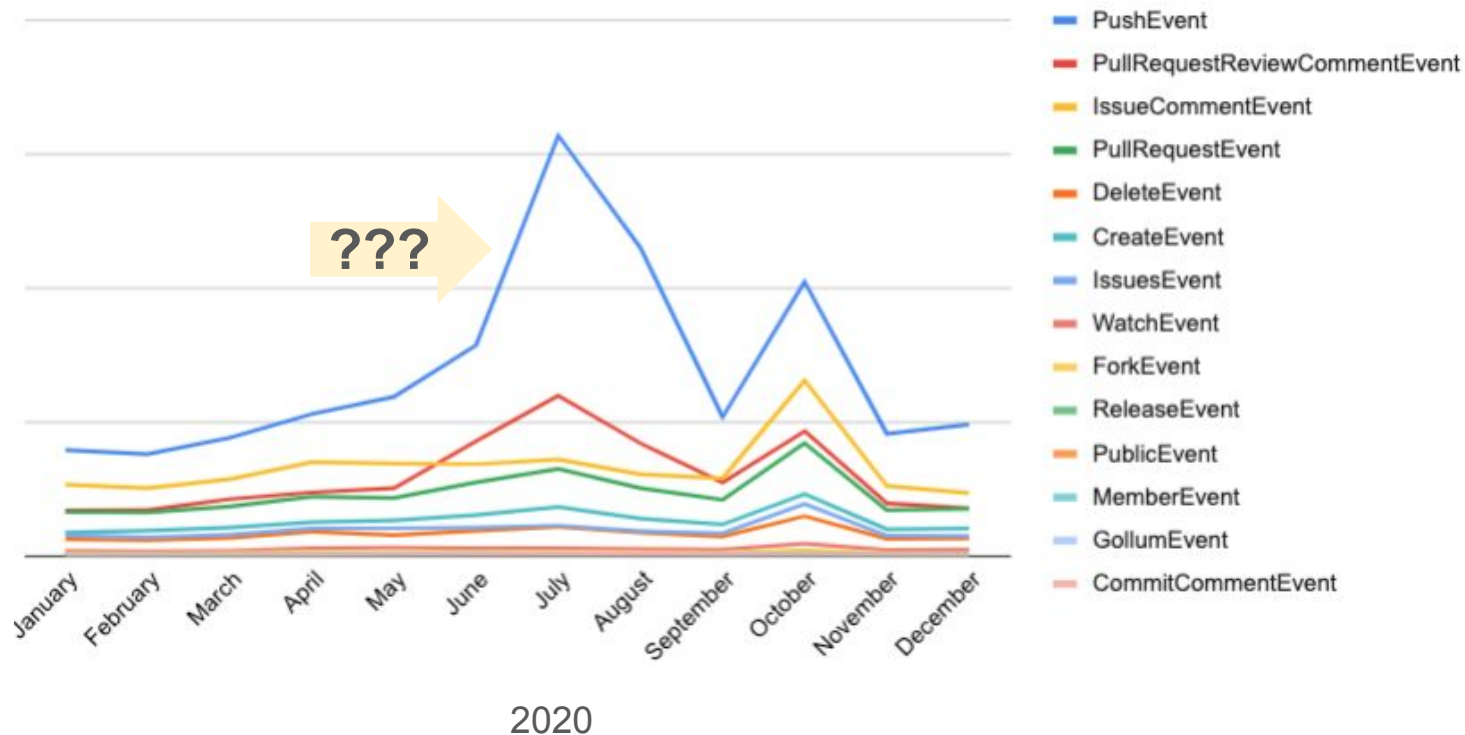
GH Archive Tables: Year vs Month vs Day



Googler initiated GitHub events in 2020

Excuse the lack
of Scale!

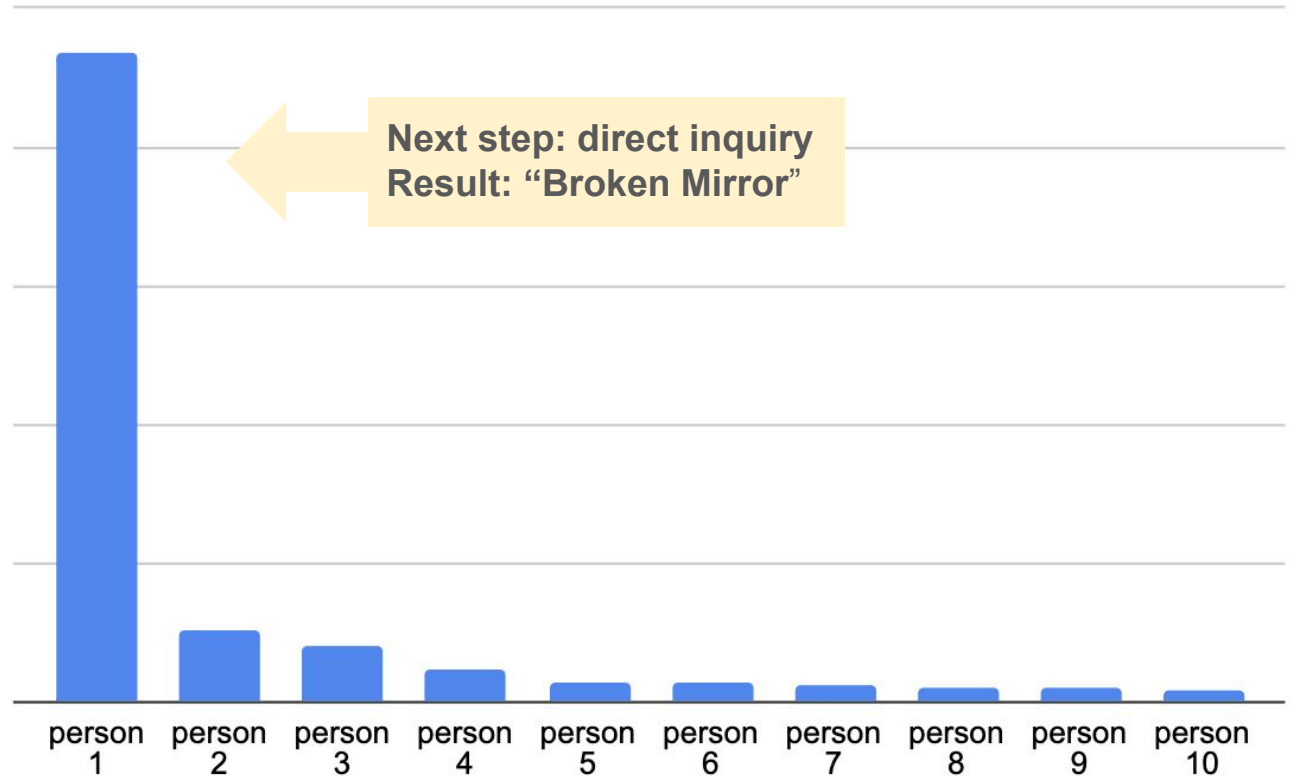
(#notpublic)



Who were the top Googlers by 'push events' in 2020?

Excuse the lack of
Scale!

(#notpublic)



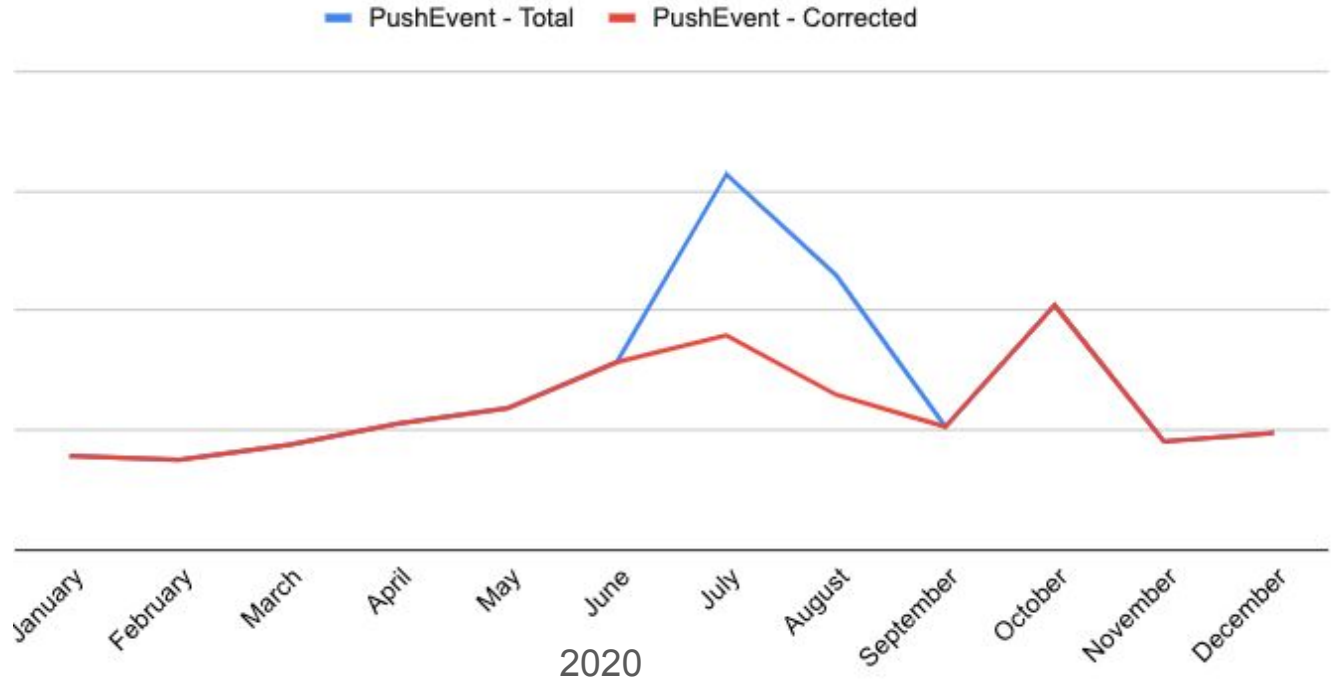
Source: www.gharchive.org -> BigQuery -> Google Only Subset

Corrected 2020 Googler PushEvents

Ruling: Script run amok

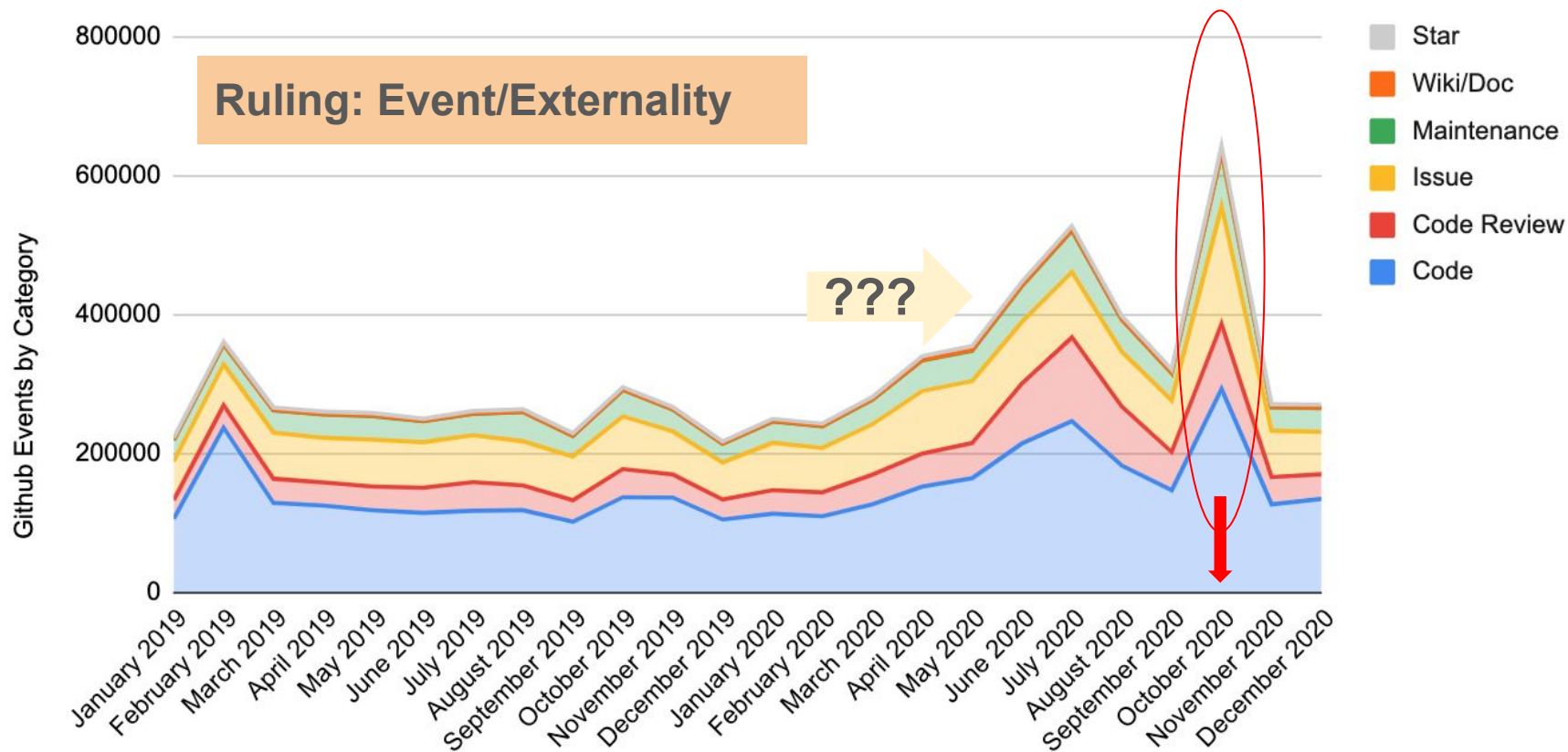
Excuse the
lack of Scale!

(#notpublic)

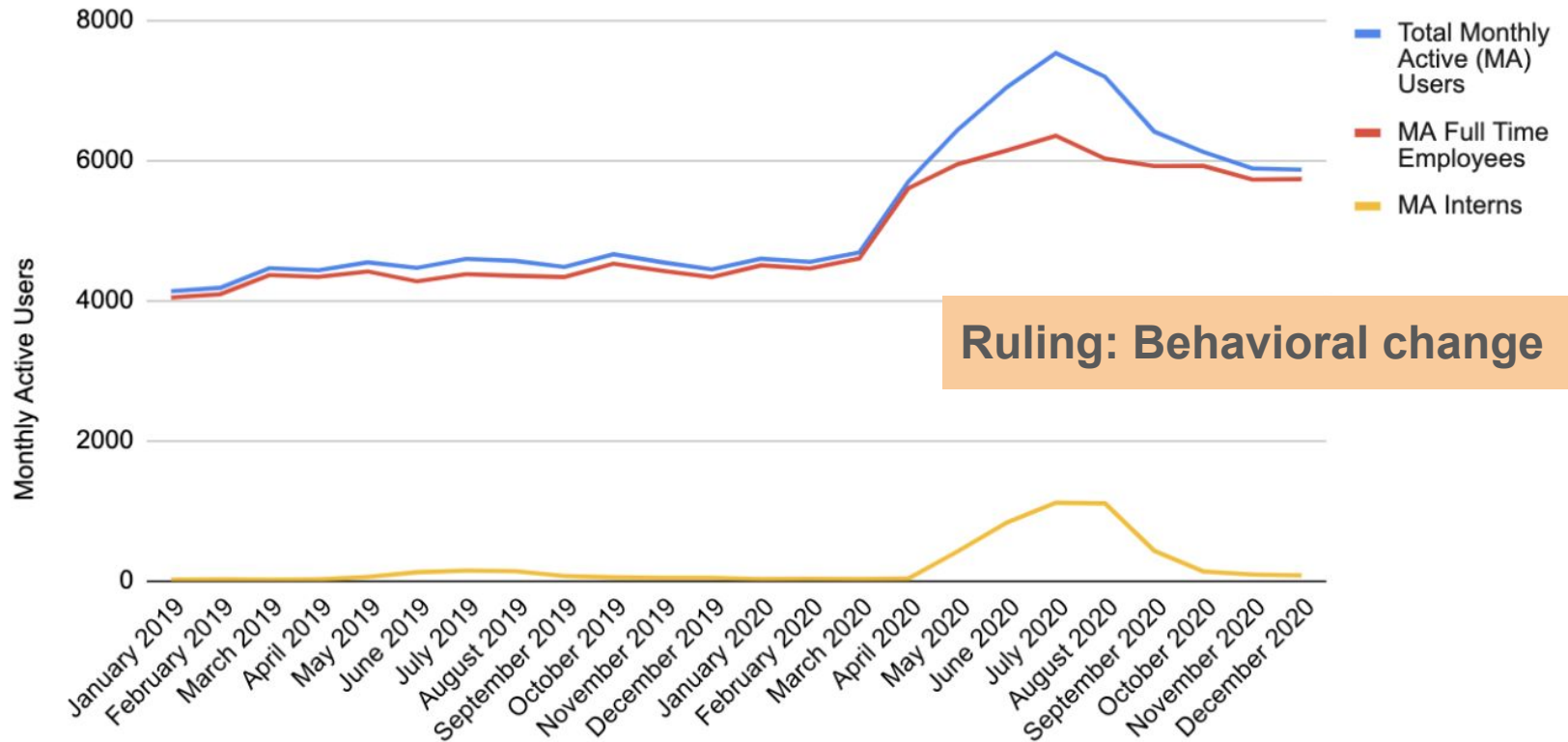


Source: www.gharchive.org -> BigQuery -> Google Only Subset

Alphabet Initiated Github Events: Grouped by Category



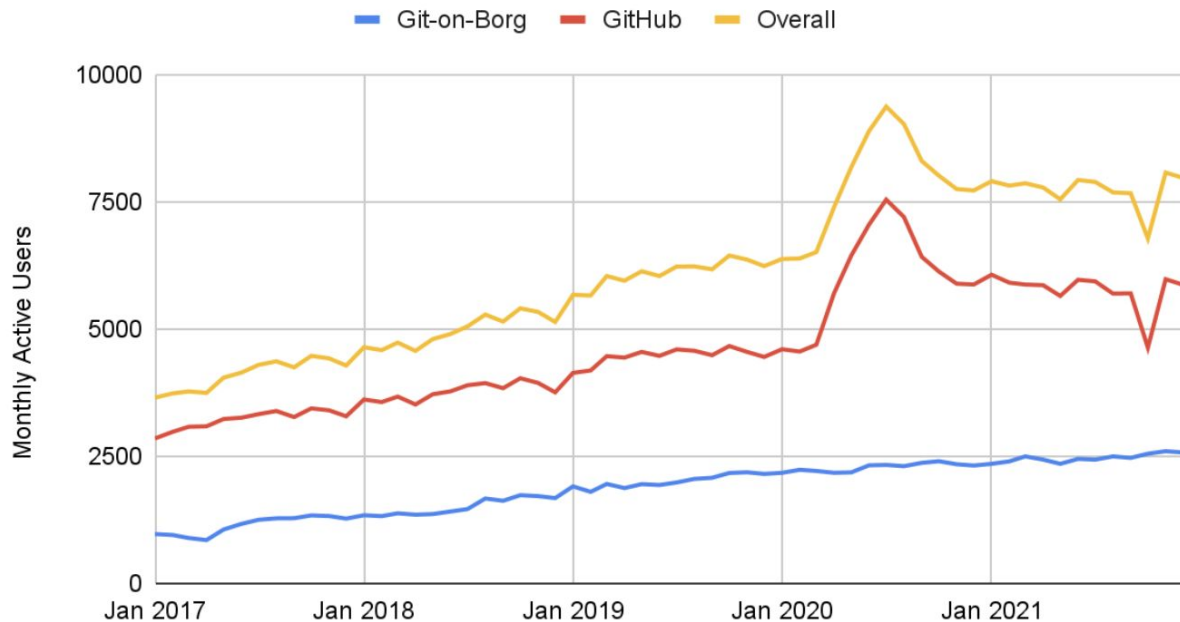
Alphabet's Monthly Active Users on GitHub: Split by Employee Type



Ruling: Behavioral change

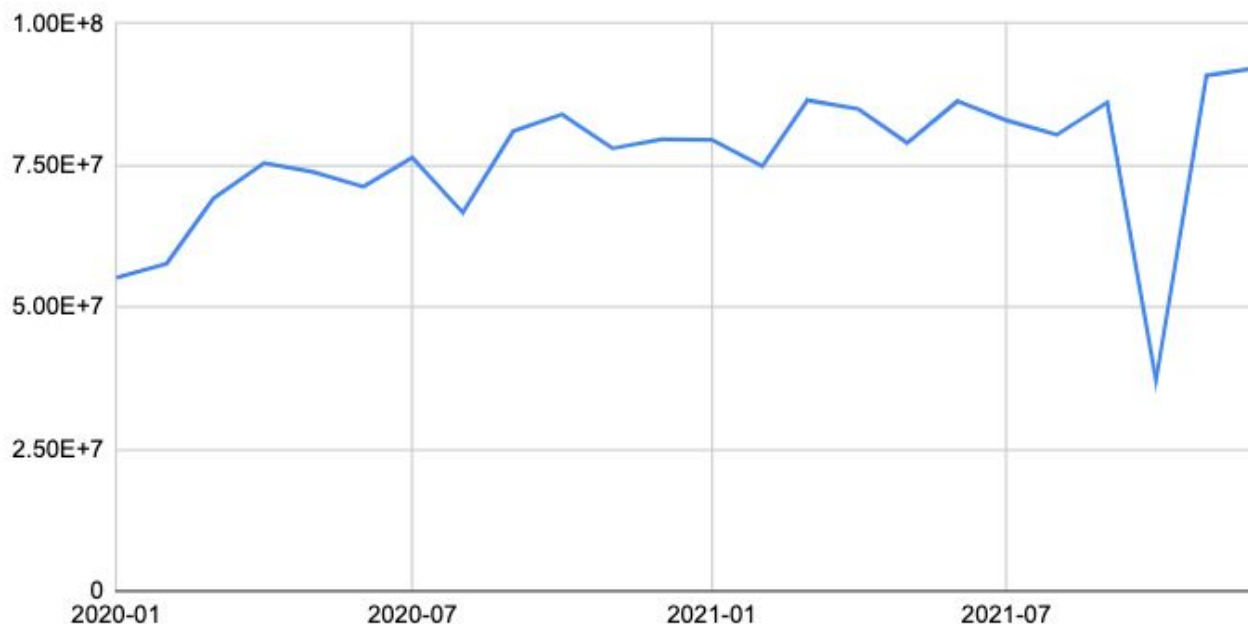
What happened in October 2021?

Alphabet's Monthly Active Users on GitHub and Git-On-Borg



Was this a Google problem or a GH archive issue?

All Events on GH Archive 2020 - 2021

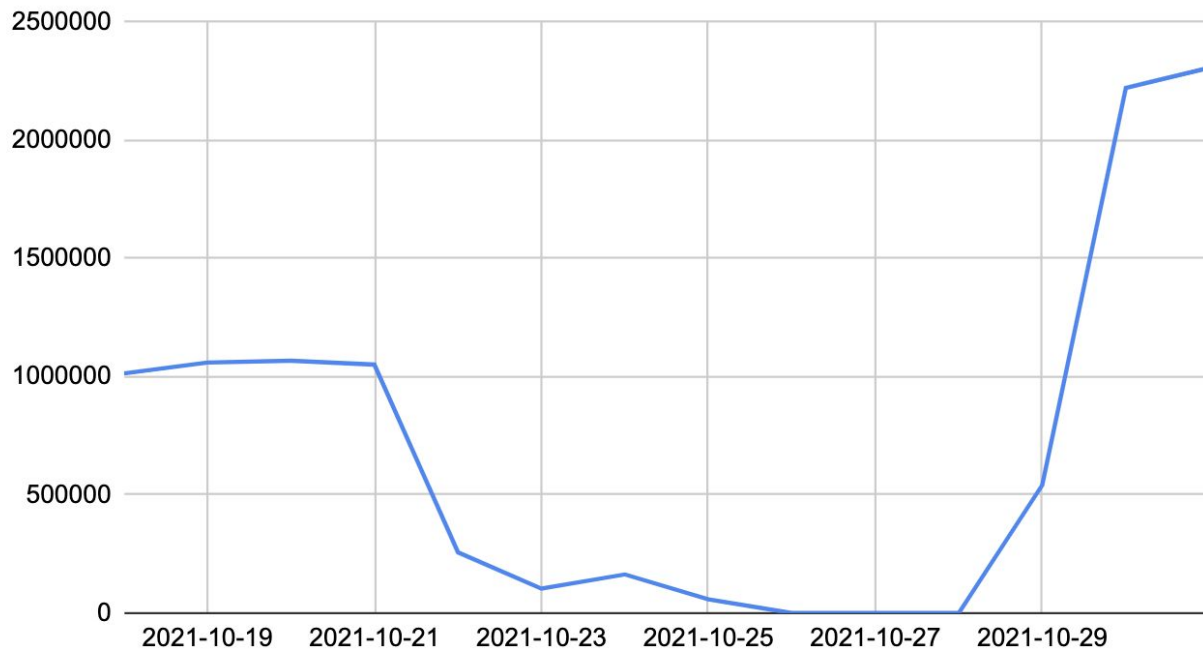


Taking a closer look at Oct 2021

Ruling: Broken Machine

All Events - Oct 18-31

Reference: github.com/igrigorik/gharchive.org/issues/261



Source: www.gharchive.org -> BigQuery

Can we assume
you are a human?

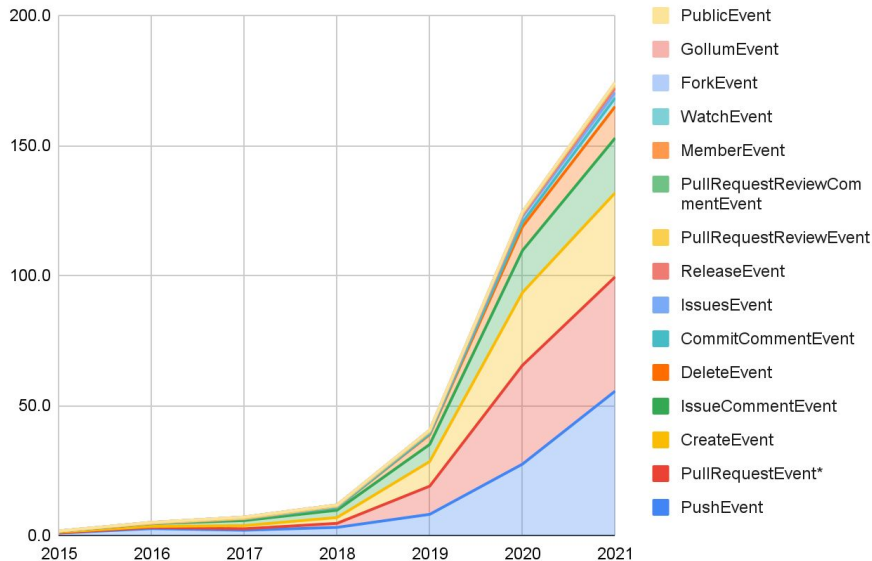


Bots on GitHub

In 2021, “bot” accounts generated over **43m PullRequestEvents***

This represented over **47%** of ALL PullRequestEvents logged in 2021

GHArchive: Actions taken by 'bots' (in millions)



Source: www.gharchive.org -> BigQuery

Bot list: github.com/cncf/devstats/blob/master/util_sql/only_bots.sql

Systematic Approaches

- Variations from the **baseline**
 - YoY, MoM
 - Ratio Metric PER person, project, repo, etc
- Can you detect **contextual cadences**:
 - System, People, Releases
- Factor in **externalities**
 - Events, Sprints

“Scale” problems

- Metrics that equal 0
 - Always share more context! Ex - mean vs median vs mode
- “There’s nothing to measure”
 - Where is the data? (no data IS data)
 - Expand your base?
 - Similar projects comparisons

...Always include bases, sources, and population contexts!

Checks and balances

- **Talk to a person!**
- Surveys can provide more aggregate context
- Find alternative systems/sources to confer
 - Social media, forum activity
 - Public calendars
 - Content, documentation

Other tips, resources?



Questions?

Sophia Vargas

 **@Sophia_IV**

Google Open Source