





Big Data Platforms Final Project: will TuringBots replace human software developers?

Can AI solutions provide meaningful improvement in developers' productivity and eventually replace human software engineers and data scientists?



Introduction



I analyzed publicly available GitHub repositories and commit history to understand the underlying relationship between the usage of AI tools and developers' productivity.

Main objectives of the analysis:

- Are there temporal differences in commit patterns?
- How are they related to the emergence of AI tools?

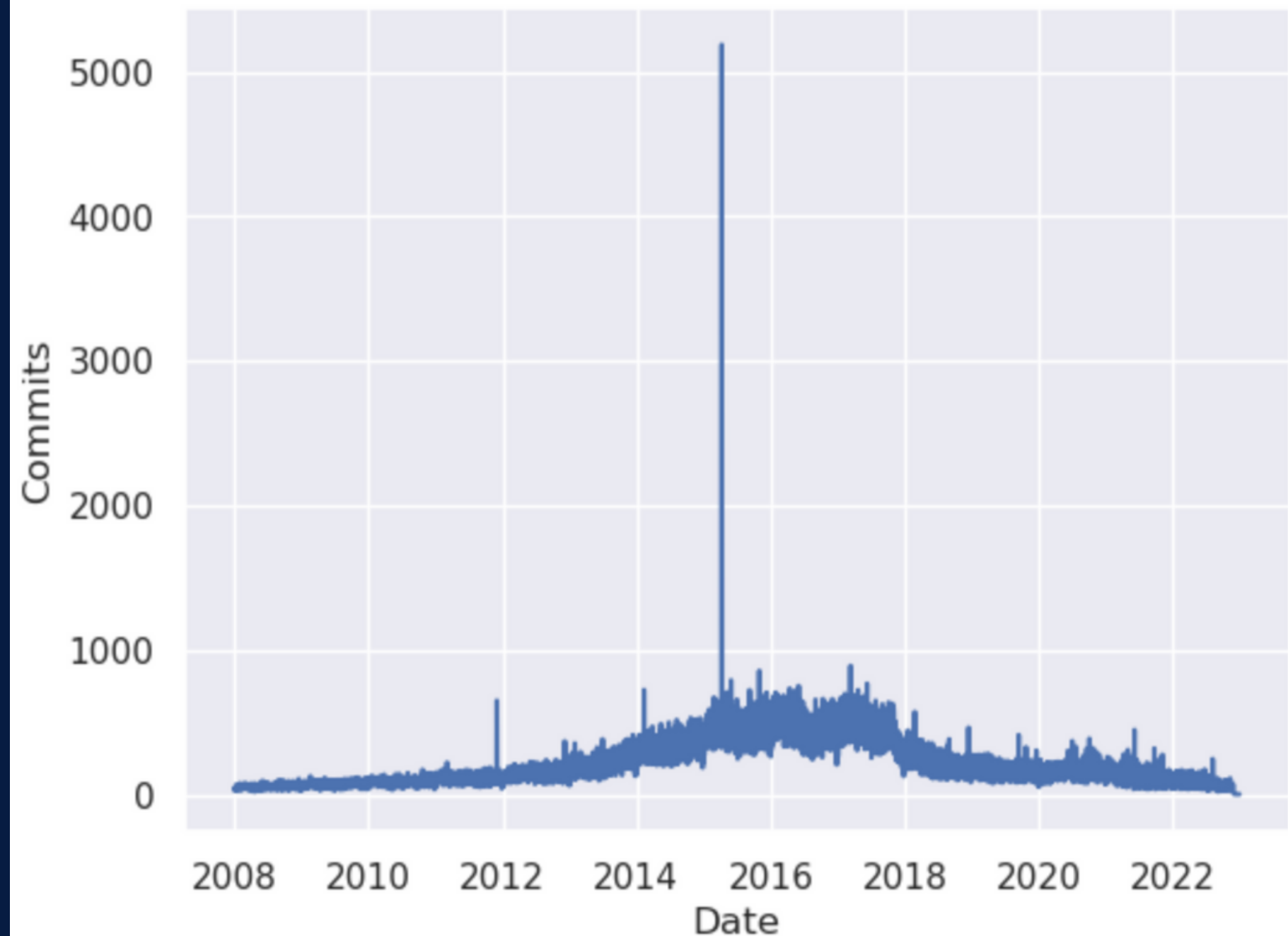
Executive Summary:

- Starting 2019, there is a decline in the volume of commits and total usage of programming languages
- Most commits substantively correspond to commonly used developer productivity metrics
- At the same time, among highest volume committers are automatic bots, robots, and assistants which suggests automation of developer roles

Timeline

GitHub Commit history between 2008 –2023

Highest spike in 2015
corresponds to the largest DDoS
attack on GitHub which
originated in China and was
aimed at circumventing
Chinese state censorship

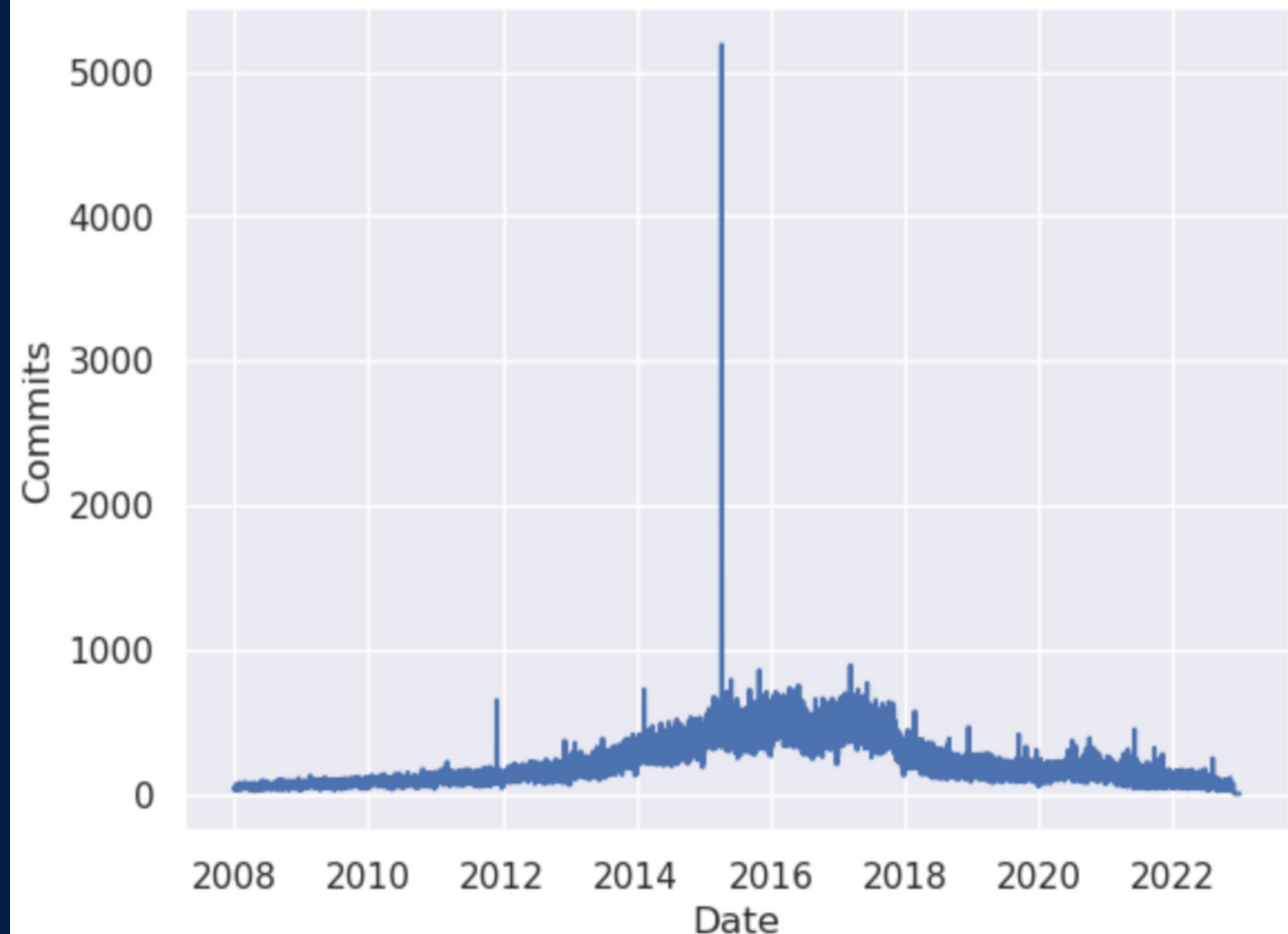


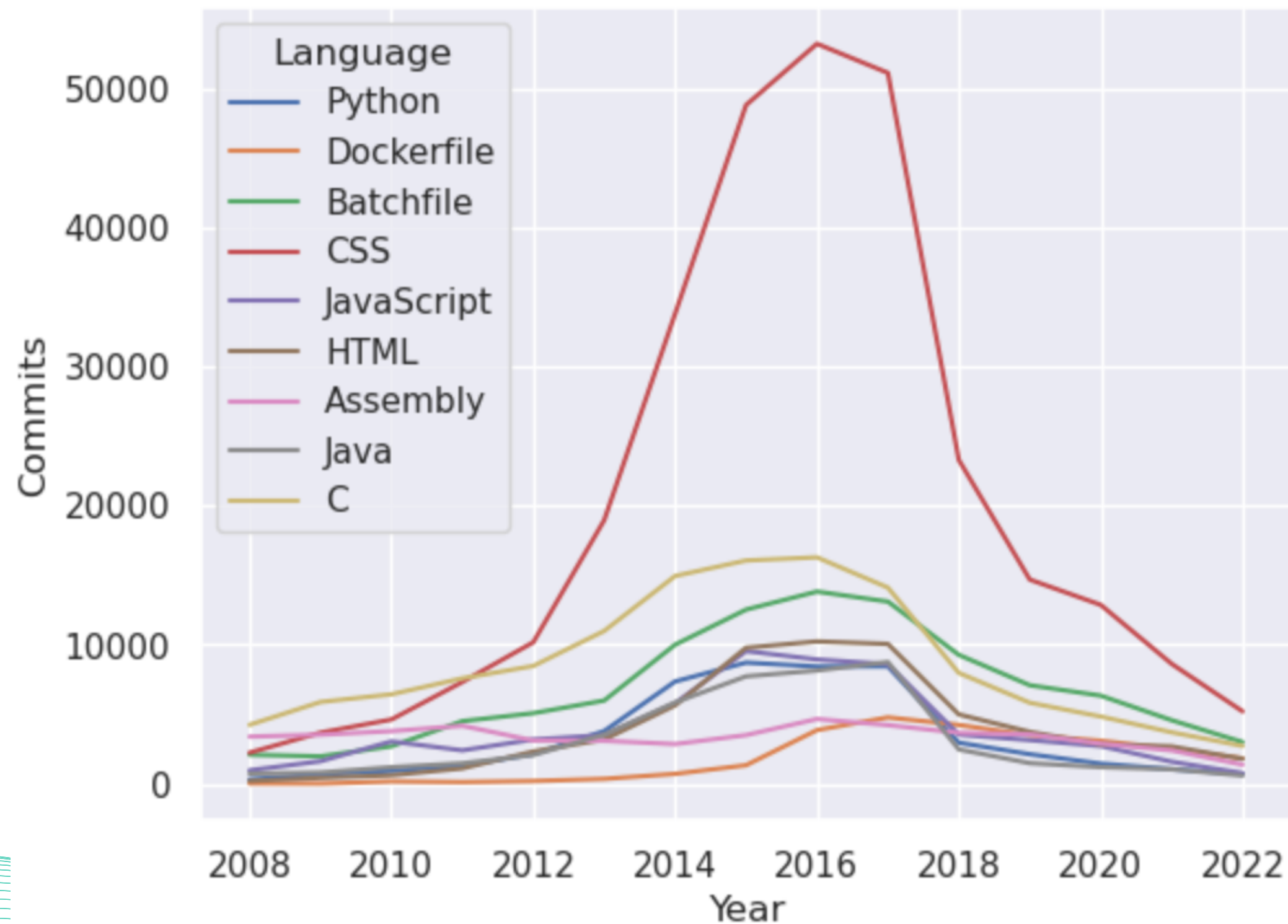
Timeline

Number of commits was the highest between 2013–2018. After 2019, the numbers went back to the level of 2008.

Similarly, after 2019 we see the emergence of AI developer assistants:

Code Llama (2023), Amazon CodeWhisperer (2023), ChatGPT (2022), Github Copilot (2021), ChatGPT-3 (2020), StarCoder (2019)





TOP-10 Programming Languages

CSS

C

Batchfile

HTML

JavaScript

Python

Assembly

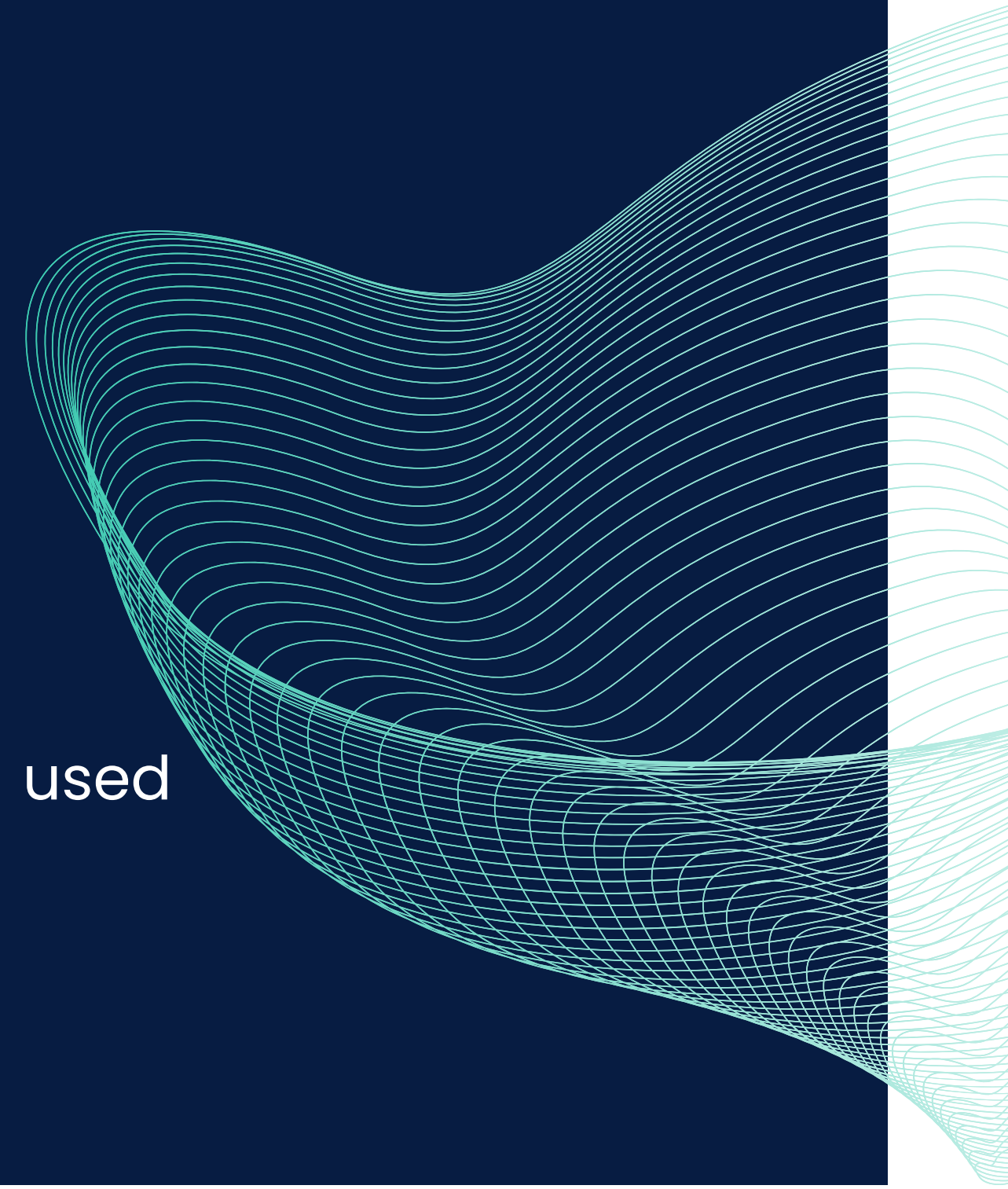
Java

Dockerfile

TOP-3 licenses associated with programming languages:

- **mit**: CSS, JavaScript, Ruby, Python, HTML
- **apache-2.0**: Java, CSS
- **gpl-2.0**: CSS, Python, C

MIT Open Source License remains the most commonly used license due to its permissive open-source policy.



Developer Productivity Metrics

What are the most common indicators of developer productivity?

Cycle Time

Typically a short cycle time correlates to small PR sizes, a healthy review process, and high deployment frequency.

Rework rate

Rework rate measures the amount of changes to code that's less than 21 days old. If a team is spending a lot of time reworking code that's just been deployed, that's a sign of code churn or low-quality code.

Bugs Fixed

Higher rate of bug fixes indicates higher overall productivity and efficiency.

Who are the largest users of GitHub?

Cloud Foundry – Relint CI Pools

WordPress – iOS

Olympic1 –
CKAN for Kerbal Space Program

Facebook – Apache Thrift

We see that Tech Giants continue to drive the growth of the industry by open sourcing the technology to users (iOS and Apache Thrift)

What are the most frequent commit reasons?

- **Deployment**

“triggering build with pending triggers”

“test”

“Update _config.yml”

- **Bug fix**

- **Updates in development**

“updating submodules”

“update dependencies”

“refactoring”

“bump version”

TOP Committers

GITHUB

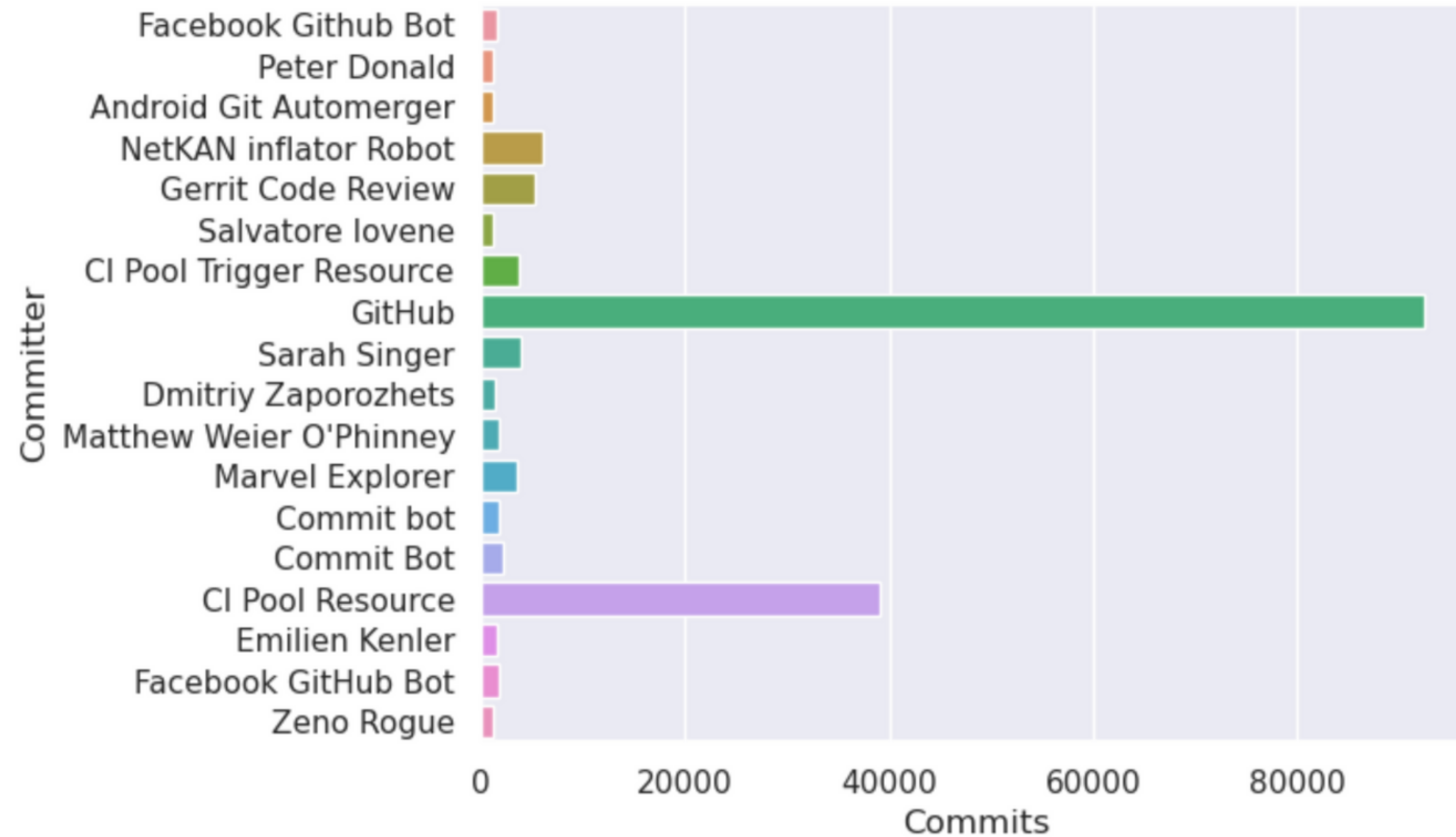
CI POOL RESOURCE

NETKAN INFLATOR ROBOT

GERRIT CODE REVIEW

SARAH SINGER

Among the highest volume committers are automatic assistants and bots, such as pool-resource/pipeline triggering assistants





Thank you!

**Elena
Smyslovskikh**

esmyslovskikh@uchicago.edu

