



OS13

Maximizing GitHub APIs: Journey to Open-Source Metrics

Alanah Jonas
alanah.x.Jonas@gsk.com

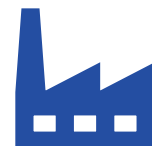
GSKs R Adoption Journey



R Training courses
developed



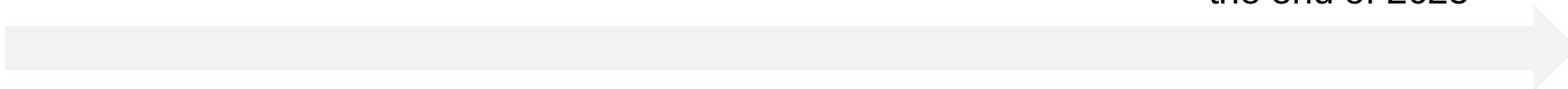
First QC study



First production
study



Target to achieve
50% of code in R by
the end of 2025



2017

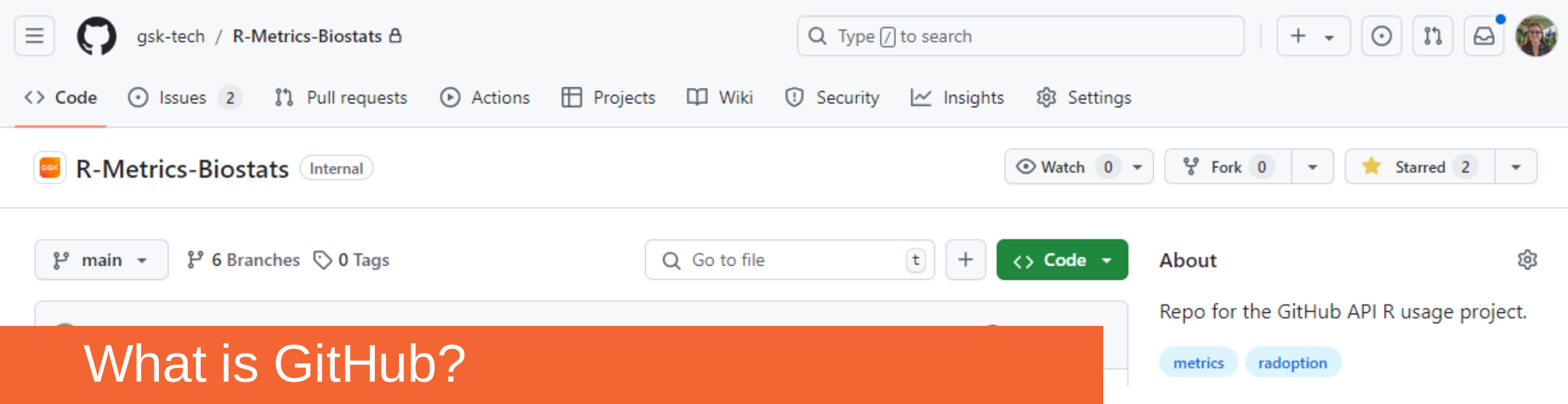
2024

We've been adopting open-source technologies for several years now, but how can we effectively measure our progress?



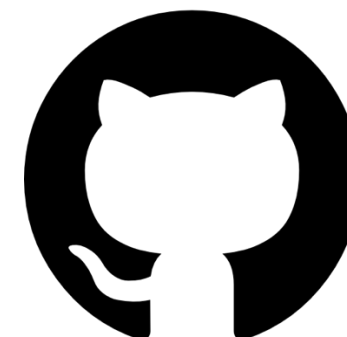


The Solution: GitHub APIs



What is GitHub?

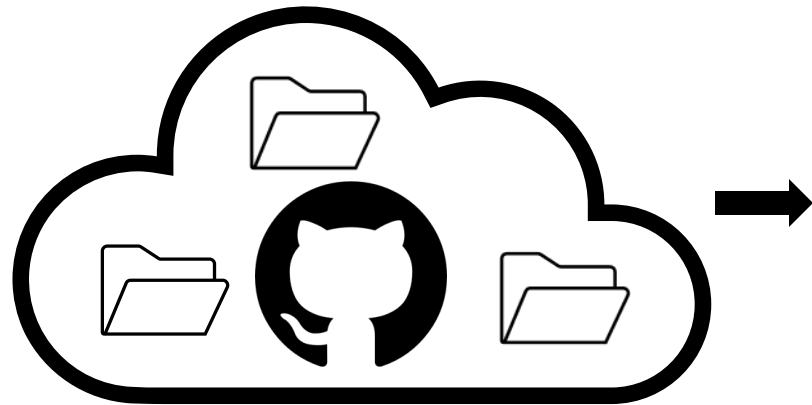
- GitHub is a web-based platform used for version control and collaborative software development.
- GitHub helps developers manage code, track changes, and collaborate on projects efficiently.



What are APIs/GitHub APIs?

What is an API?

APIs enable different software systems to interact seamlessly, exchanging data and commands.



GitHub
API



What is the GitHub API?

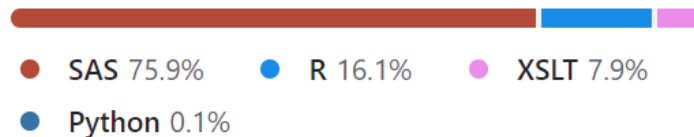
GitHub APIs Allow developers to interact with GitHub programmatically, enabling them to access data and features of GitHub.

What information can be extracted from a repo?

- Repository Details
- Files and their contents
- Commits and Pull Requests
- Branches
- Language usage
- Contributors

```
> folder R
> folder dev
> folder image
> folder rsconnect
v folder scripts
  dashboard_data_prep.R
  .gitignore
  R-Metrics-Biostats.Rproj
  README.md
  manifest.json
  r_metrics_biostats_dashboard.Rmd
```

Languages

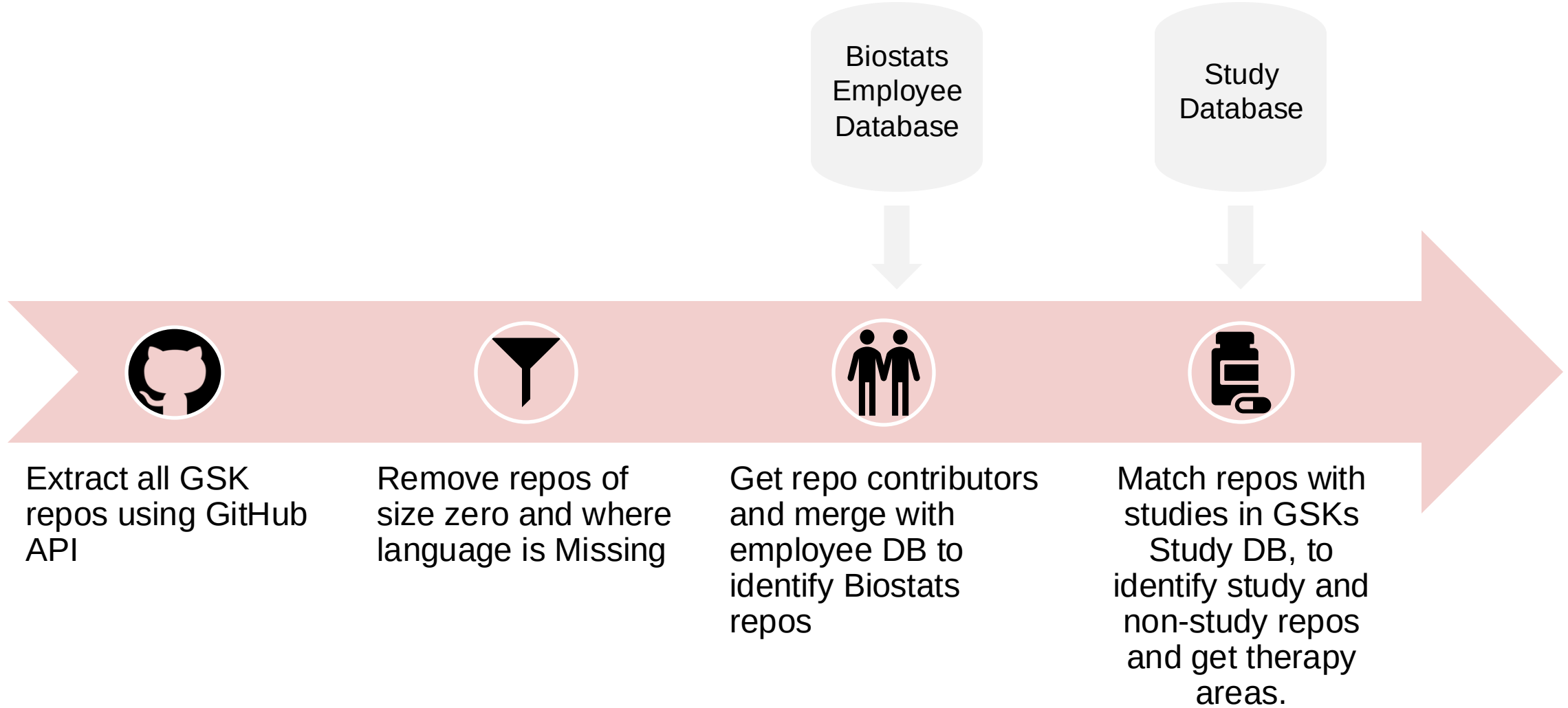


Branch	
dev	
explore_e	
investigate_language	
asset_explore	
revert-192-dev	
quarto_explore	

Contributors 4

- aj641247_gsk Alanah Jonas
- nd258262_gsk Niladri Dasgupta
- bzk67899_gsk Becca Krouse
- bca99787_gsk Benjamin Arancibia

Finding Biostats GitHub Repos: Process Flow





Example

Get a repos language usage

GSK

gh Package

The **gh** package allows you to interact with the GitHub API directly from R.

1. Authentication:

Set up an environment variable called `GITHUB_PAT` equal to your GitHub Personal Access Token (PAT).

```
# Set the environment variable  
Sys.setenv(GITHUB_PAT = "your_personal_access_token")
```

2. Requests:

Use the `gh()` function.

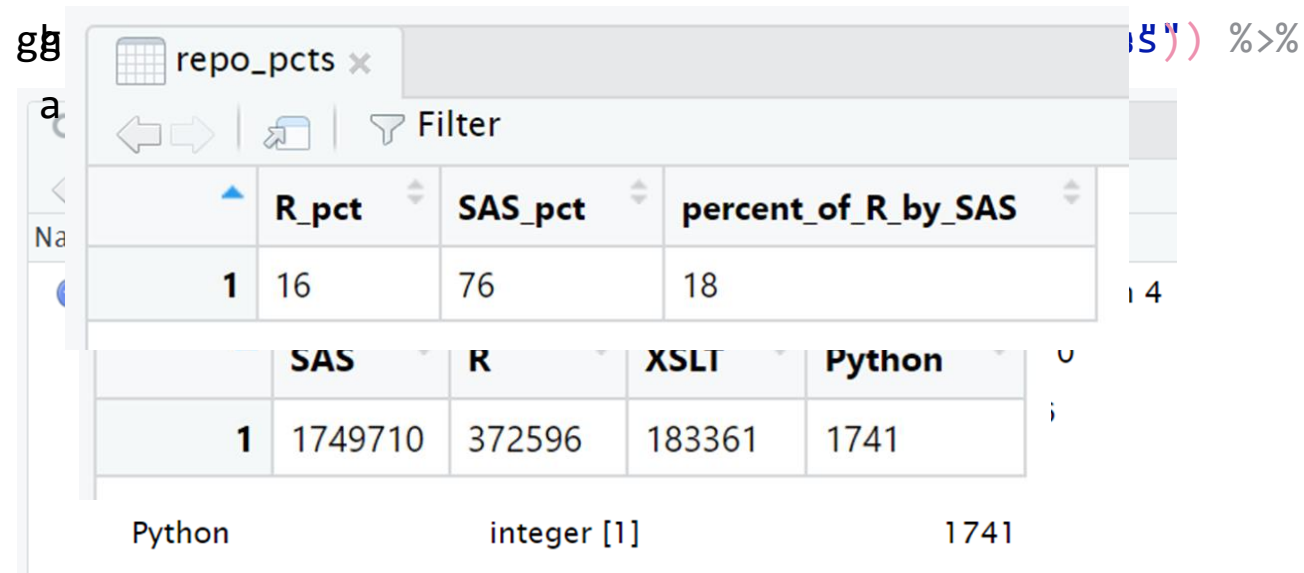
Example: Get a repos languages

To get a list of repository languages we do the following request:

/repos/{owner}/{repo}/languages

How do we go from Bytes to Percentages?

```
repo_languages <-
```



The screenshot shows a data table viewer with a tab labeled 'repo_pcts'. The table has columns for 'R_pct', 'SAS_pct', and 'percent_of_R_by_SAS'. The first row shows values 16, 76, and 18 respectively. Below this, there is a section with columns for 'SAS', 'R', 'XSLT', and 'Python', showing byte counts: 1749710, 372596, 183361, and 1741. The bottom of the image shows the variable 'Python' with its data type 'integer [1]' and the value '1741'.

	R_pct	SAS_pct	percent_of_R_by_SAS
1	16	76	18

	SAS	R	XSLT	Python
1	1749710	372596	183361	1741

Python integer [1] 1741

R Usage Metrics

01

Relative R =

R code size out of
(R and SAS)

02

Any R =

Total Repos with
any usage of R.

03

Substantial R =

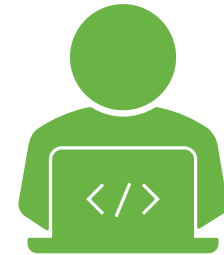
Repos where R
code size (out of R
and SAS) is
 $\geq 10\%$



Metrics have enabled us to clearly see R usage across studies.



Non-studies have higher R usage.



R is used frequently in studies, but in small amounts.

Future Explorations

Enhancements

- Access to private repositories.
- Investigate the number of people on a study writing R code.
- How do we measure 50% R code.
- Measure progress over time.
- Resolve Long processing time and rate limits.

Other API Projects

- Assessing the health of GitHub repositories.
- Creating a Biostatistics Tool Catalogue.

Conclusion

We have shown that we can effectively use GitHub APIs to automate tracking and analysing our open-source efforts, eliminating the need for manual recording.



Any Questions?

GSK

GSK