Language Communities on GitHub

Puja Jena, Hayeon Kim, Frances Tso, Sharon Yu, Jake Zimmerman

March 27, 2017

Software development is social

- Software is built collaboratively
- GitHub is a platform for “social coding”
- GitHub users collaborate on many tasks:
  - reporting bugs
  - discussing future changes
  - reviewing code quality

Programming languages create communities

- Each project has a primary language
- Same language? Same habits & customs
- Language communities accumulate stereotypes
  - C, C++? $\rightarrow$ “old-school”
  - JavaScript, Ruby? $\rightarrow$ “hip”
  - SML, Haskell? $\rightarrow$ “academic”

Key Findings from Existing Research

- Interactions between core and new members differed from interactions among core members.
- Discussions tended to involve core and peripheral developers, not just core developers.
- Prior interaction increases likelihood of pull request acceptance
- Developers with higher stakes in the project are more likely to engage in discussions

Our question:

How does social activity on GitHub vary by programming language community?

For the midterm report, we narrowed in on:

How does emoji use vary by language community?

Emoji on the GitHub platform

GitHub lets users

- write emoji in comment bodies
- add Slack-style “reactions” to other people’s comments

We looked at **reactions**.

Choosing projects to sample

- For 10 hand-selected programming languages,
  - Pick 10 random projects using that language where:
    - the project has between 1,000 and 4,000 stars (moderately popular)
    - the project was updated within the last 3 months (moderately active)

Result: 100 random projects, 10 in each language, where:

- popularity held constant (not just popularity contest)
- project isn't abandoned (still part of the community)

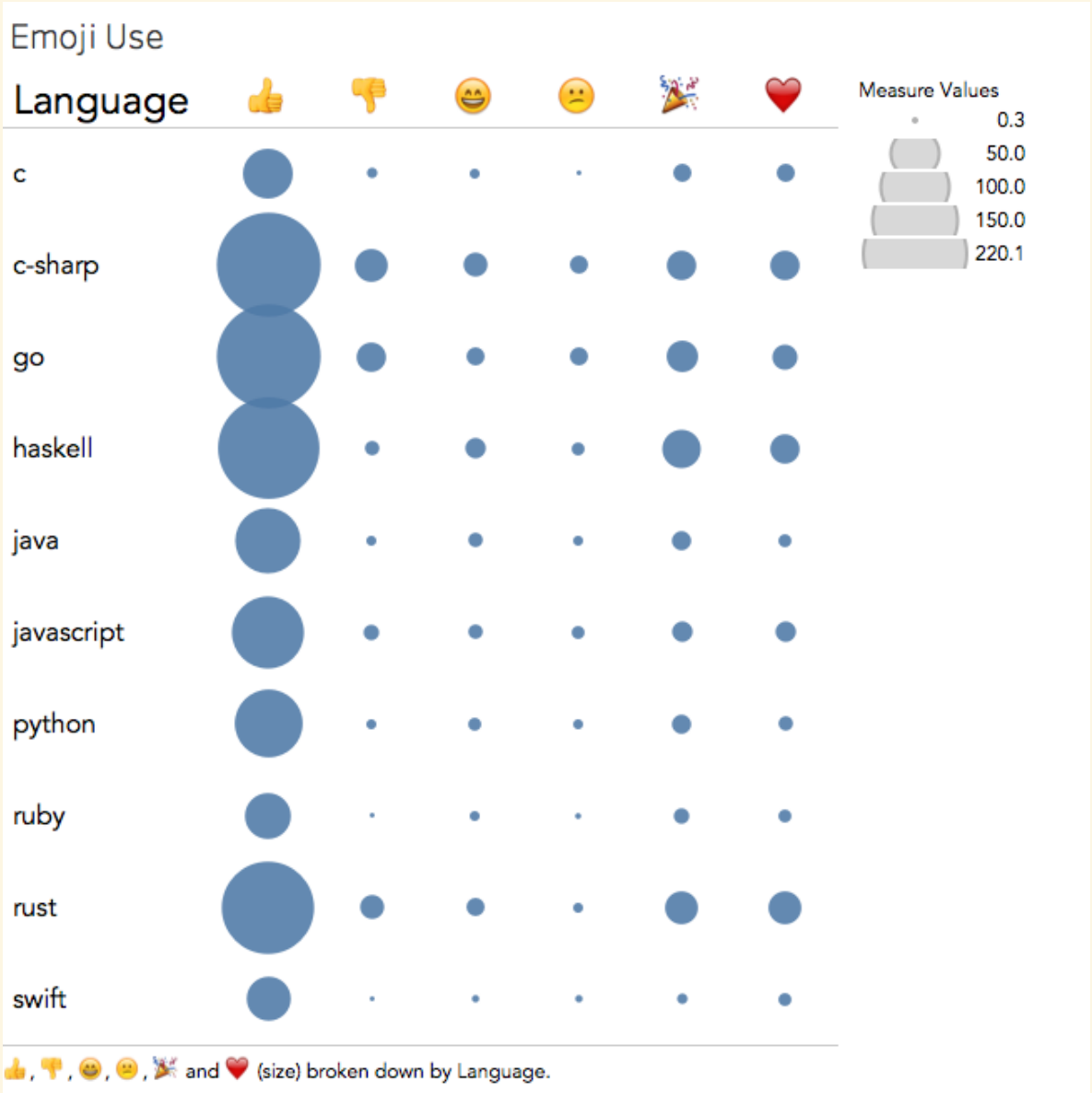
Getting the data

- For each sampled project,
  - Fetch all reactions on all comments
  - Add up how many times each emoji was used
- For each language,
  - Average the number of times each emoji was used

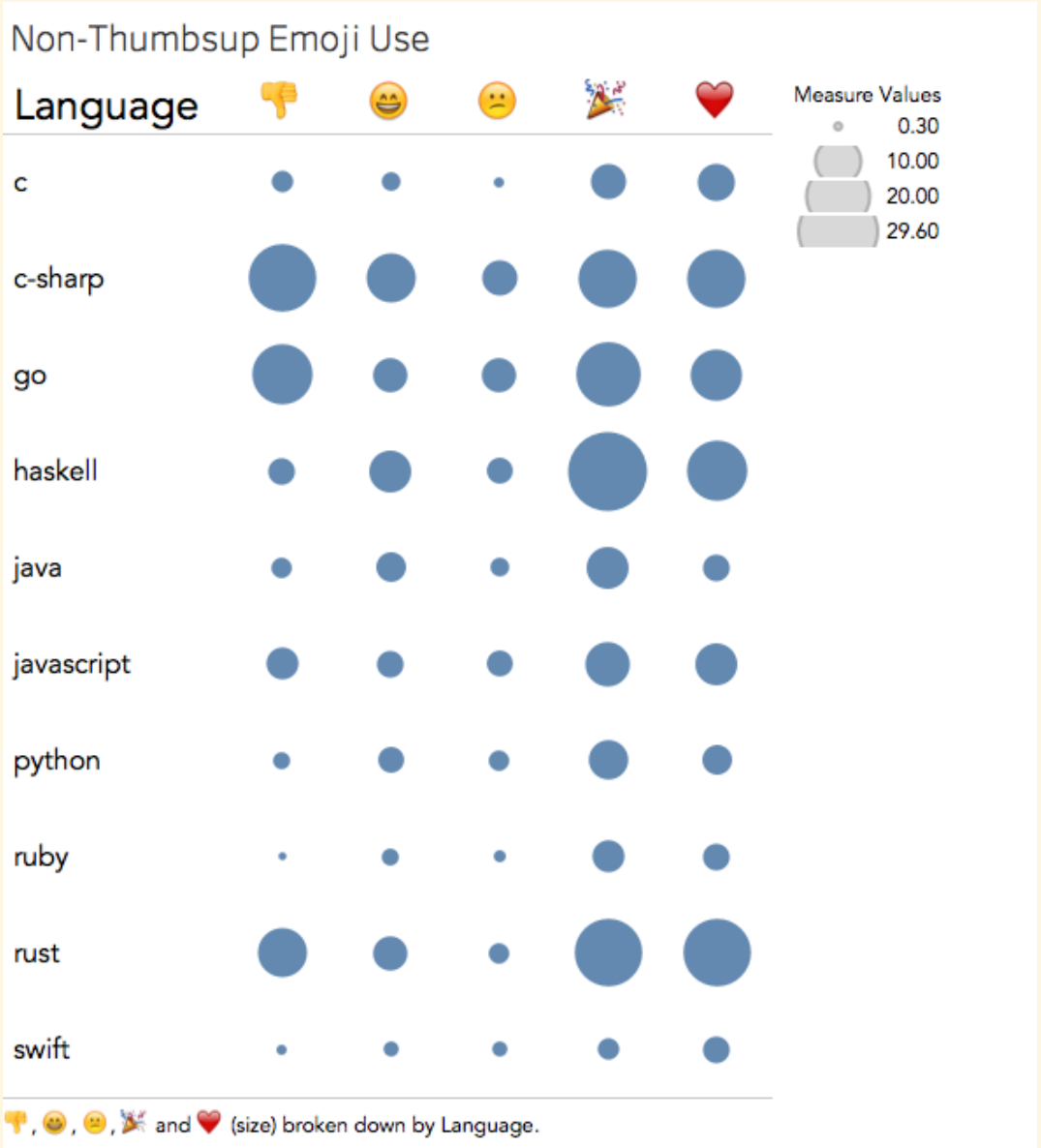
Result: average number of emoji used across 10 languages

Visualizing the Result

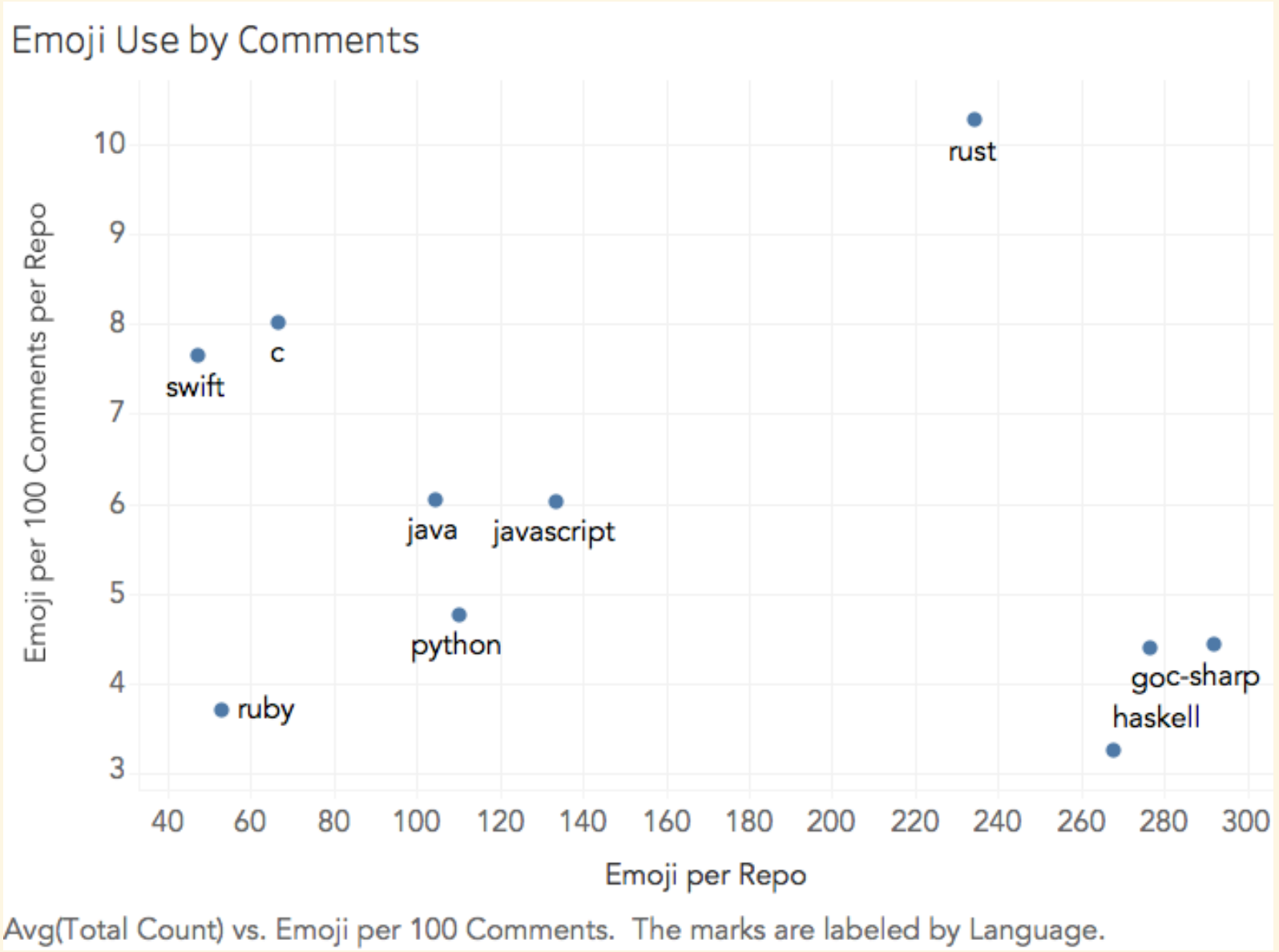
Emoji Use by Language



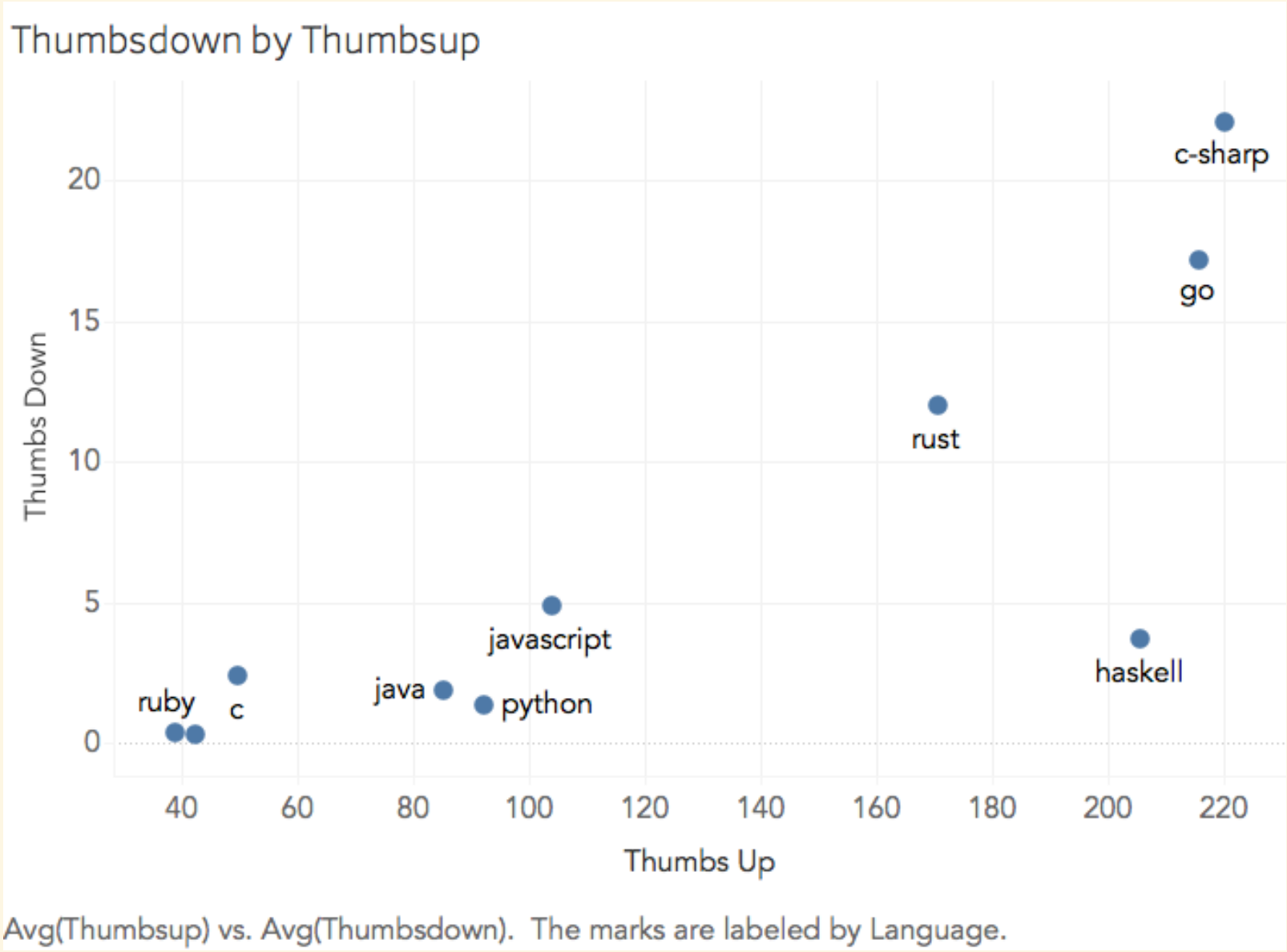
Emoji Use by Language (no Thumbsup)



Emoji Use by Comment Volume



Thumbsdown by Thumbsup



Possible questions for next time:

- Which other language communities does this community talk about?
- What influence does contribution time (day, night, …) have on discussions?
- When do people comment, versus when do they contribute code?
- Are discussions resolved quickly, or do they drag on?