

Evidence-based best practices for sustaining open-source communities

Bogdan Vasilescu
@b_vasilescu



<https://octodex.github.com/abtocat/>



MIT Lincoln Laboratory

MIT Lincoln Laboratory is an FFRDC chartered to apply advanced technology to problems of national security.

📍 MIT Lincoln Laboratory ↗ <http://www.ll.mit.edu/>

🏠 Overview

💻 Repositories 63

📦 Packages

👤 People 1

/projects

Popular repositories

LL-Fuzzer

An automated NFC fuzzing framework for Android devices.

● Python ⭐ 116 🏷 23

CEP

Common Evaluation Platform

● C ⭐ 34 🏷 13

SPARTA

Security and Privacy Assurance Research Test and Assessment (SPARTA) tool suite

● Python ⭐ 32 🏷 4

LL-Smartcard

A Python module for interacting with smart cards.

● Python ⭐ 27 🏷 11

LO-PHI

Low-Observable Physical Host Instrumentation: A suite of tools supporting introspection and semantic gap reconstruction for both physical and virtual machines.

● Python ⭐ 24 🏷 7

Seep.jl

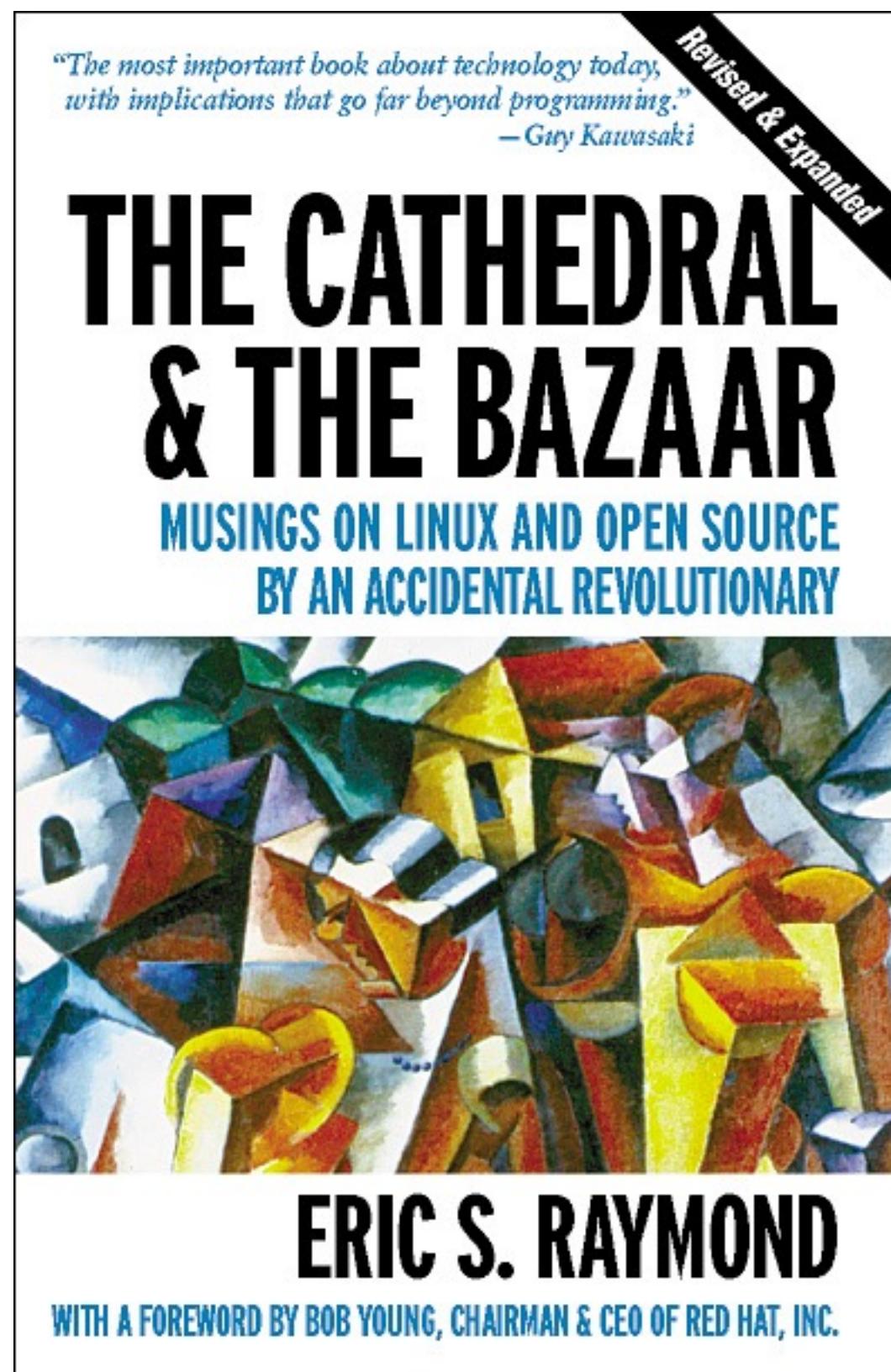
Seep.jl: a tensor flow like library for Julia

● Julia ⭐ 18

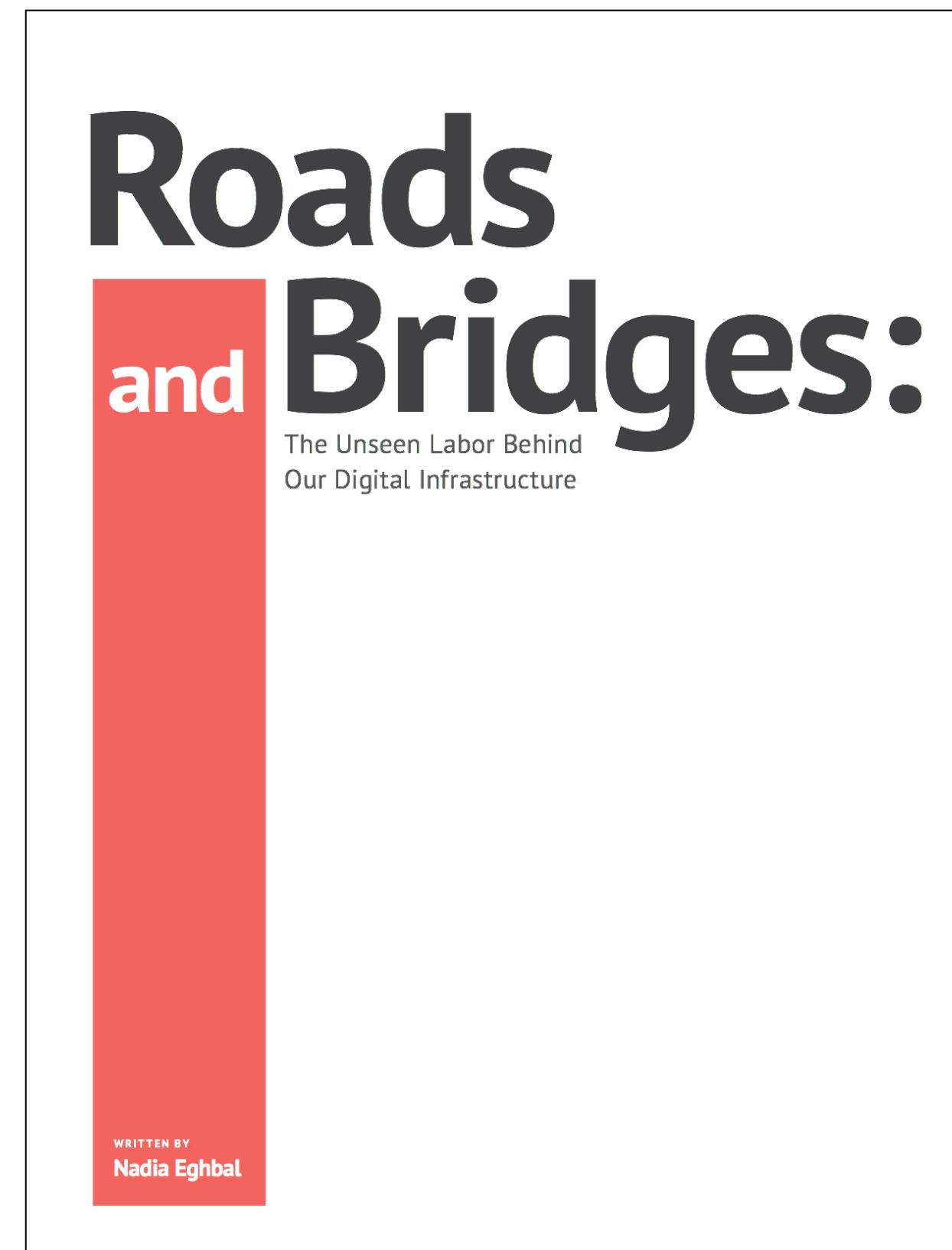
<https://github.com/mit-ll>

Open-Source Software: From curiosity to digital infrastructure

1999



2016



- Open source code as digital roads or bridges:
 - can be used by anyone to build software
- Nearly all software that powers our society relies on open source code
- Everybody uses open source code:
 - Fortune 500 companies
 - government
 - major software companies
 - startups

Open-Source Software: From curiosity to digital infrastructure

 Lincoln Laboratory  @MITLL 

Keylime, a cloud security software from @MITLL, is being adopted into IBM's cloud fleet. It's the latest deployment for this now open-source, @CloudNativeFdn sandbox technology.



Keylime software is deployed to IBM cloud
Developed at Lincoln Laboratory, the technology allows organizations to ensure the security of sensitive data stored in the cloud.
ll.mit.edu

4:55 PM · Jul 27, 2021 

- Open source code as digital roads or bridges:
 - can be used by anyone to build software
- Nearly all software that powers our society relies on open source code
- Everybody uses open source code:
 - Fortune 500 companies
 - government
 - major software companies
 - startups

Sustaining Open Source remains a challenge

Roads and Bridges:

The Unseen Labor Behind
Our Digital Infrastructure

WRITTEN BY
Nadia Eghbal

“Most of us take [Open Source] for granted, the way we take turning on the lights for granted. We don’t think about the human capital necessary to make that happen.”

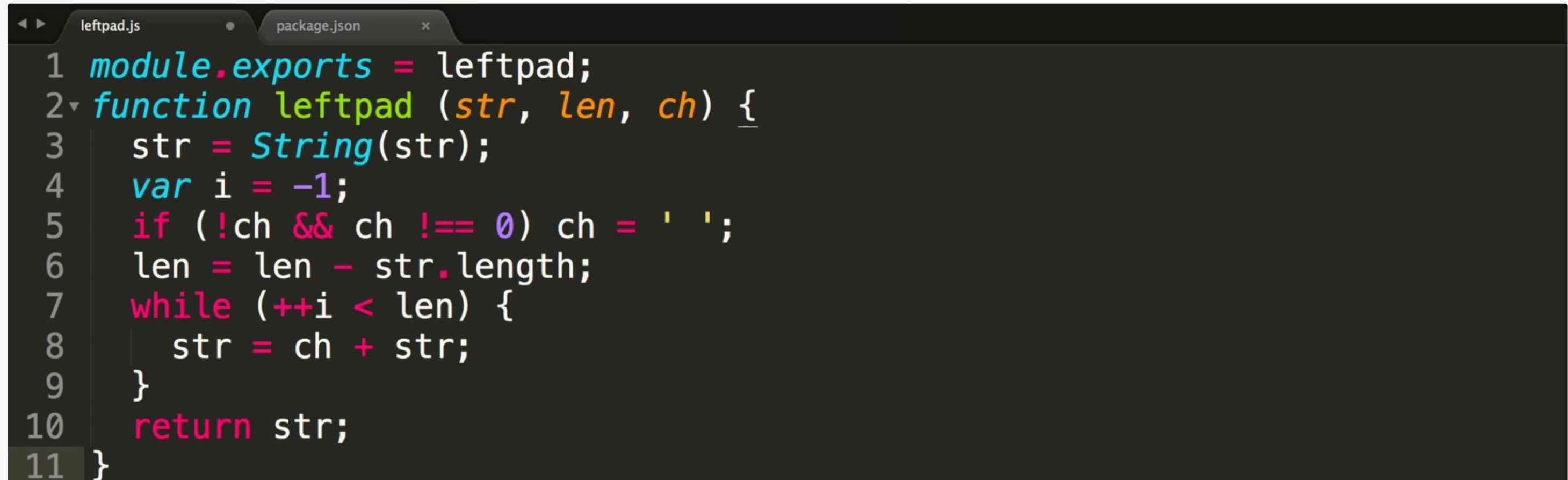
Sustaining Open Source remains a challenge

Invisible infrastructure

NPM ERR!

How one programmer broke the internet by deleting a tiny piece of code

By Keith Collins • March 27, 2016



```
1 module.exports = leftpad;
2 function leftpad (str, len, ch) {
3   str = String(str);
4   var i = -1;
5   if (!ch && ch !== 0) ch = ' ';
6   len = len - str.length;
7   while (++i < len) {
8     str = ch + str;
9   }
10  return str;
11 }
```

<https://qz.com/646467/how-one-programmer-broke-the-internet-by-deleting-a-tiny-piece-of-code/>

Sustaining Open Source remains a challenge

High pressure environment

- Equifax (market cap \$14 billion) built products on top of open-source infrastructure, including Apache Struts
- Equifax did not make any contributions to open source projects
- A flaw in Apache Struts contributed to the breach (CVE-2017-5638)
- Equifax publicly blamed (with national news coverage) Apache Struts for the breach

Equifax confirms Apache Struts security flaw it failed to patch is to blame for hack

The company said the March vulnerability was exploited by hackers.



By Zack Whittaker | September 14, 2017 -- 01:27 GMT (18:27 PDT) | Topic: Security

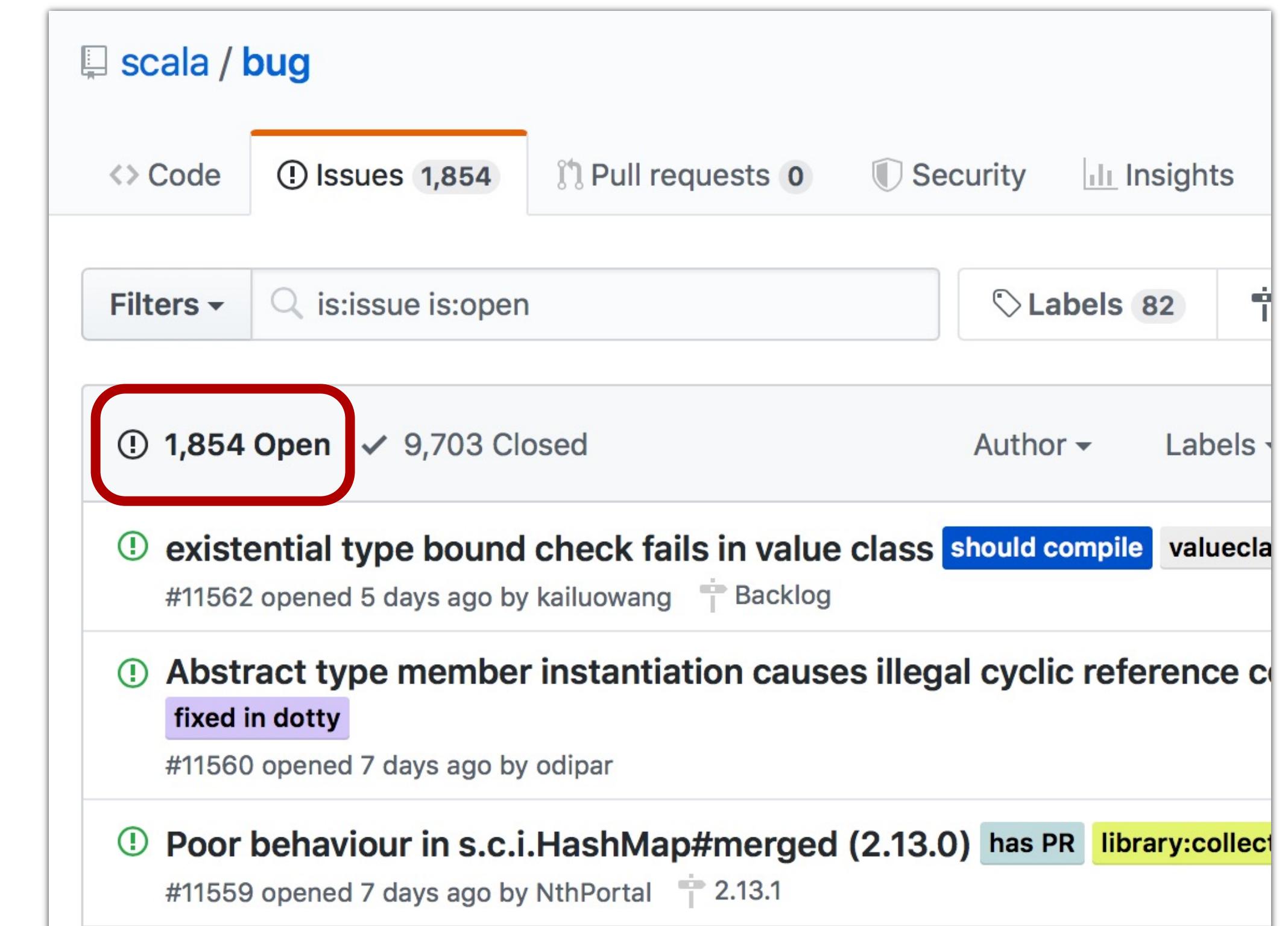


<https://www.zdnet.com/article/equifax-confirms-apache-struts-flaw-it-failed-to-patch-was-to-blame-for-data-breach/>

Sustaining Open Source remains a challenge

High level of demands & stress

- Social pressure to respond quickly
 - Otherwise, off-putting to newcomers
(Steinmacher et al. 2015)
- Entitlement, unreasonable requests from users:
 - *"I have been waiting 2 years for Angular to track the 'progress' event and it still can't get it right?!?!"*
 - *"Thank you for your ever useless explanations."*



Sustaining Open Source remains a challenge

Limited pool of potential contributors

Expectation:



“Code sees *no color or gender*”

- Perceptions of Diversity on GitHub: A User Survey.
Vasilescu, B., Filkov, V., and Serebrenik, A. CHASE 2015

Sustaining Open Source remains a challenge

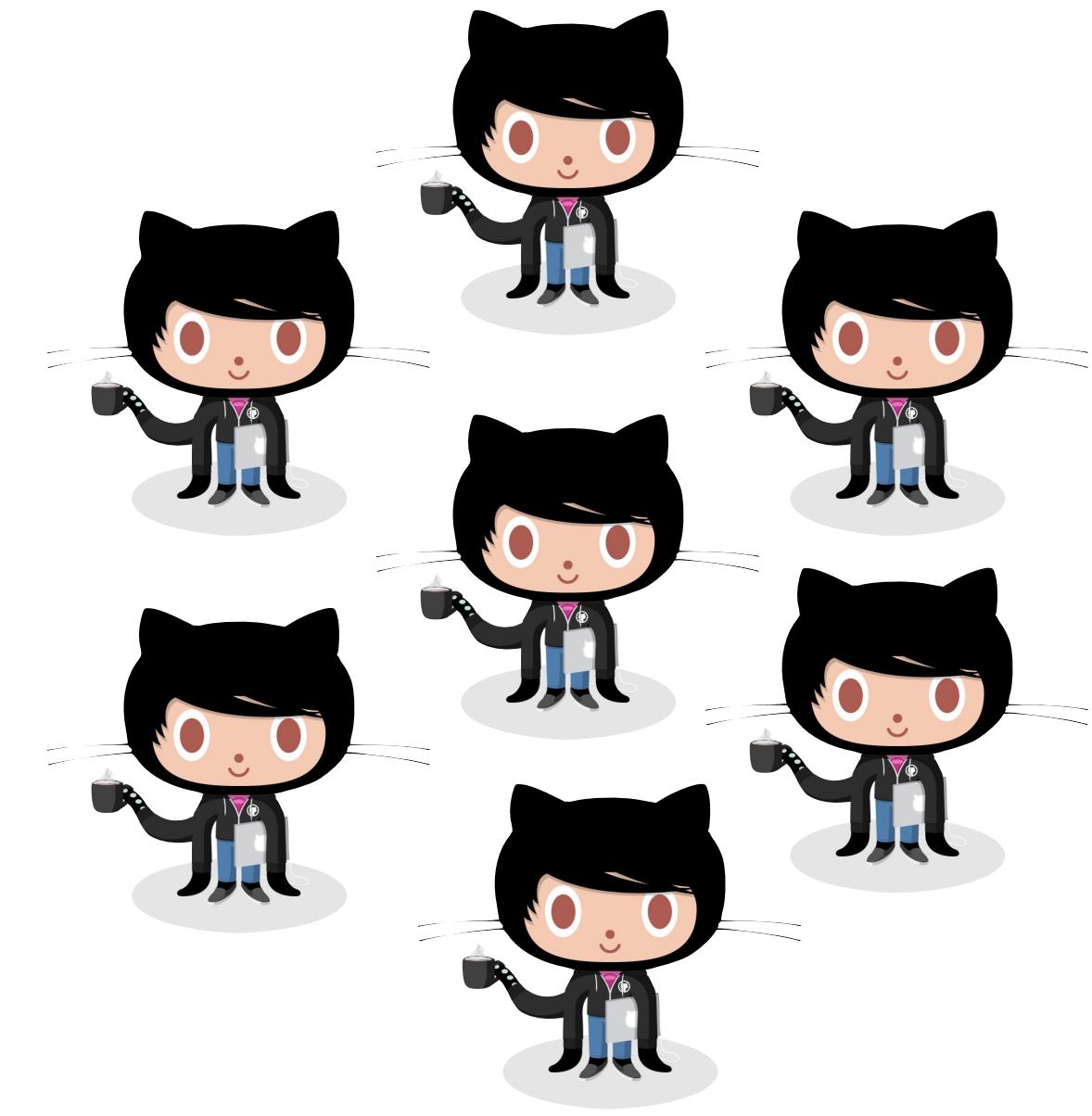
Limited pool of potential contributors

Expectation:

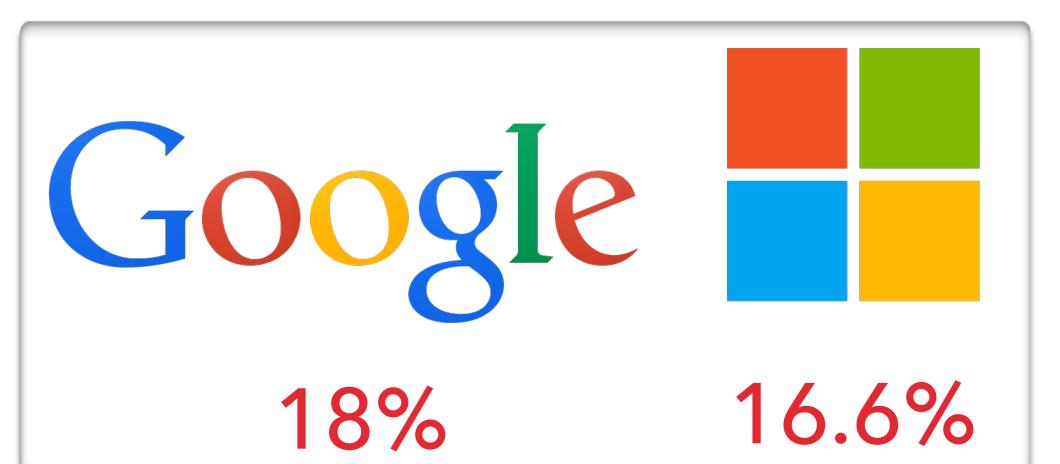
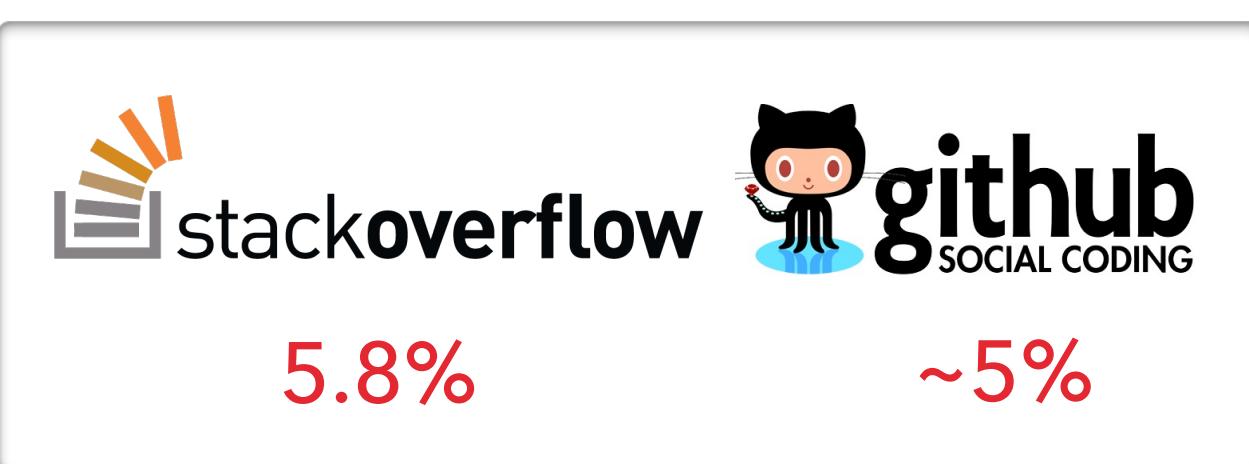


“Code sees *no color or gender*”

Reality:



Women:



- Perceptions of Diversity on GitHub: A User Survey. Vasilescu, B., Filkov, V., and Serebrenik, A. CHASE 2015

- Google Diversity (2015) www.google.com/diversity/index.html#chart
- Inside Microsoft (2015) <https://goo.gl/nT4YiI>

- Stack Overflow 2015 Developer Survey (26,086 people from 157 countries) <http://stackoverflow.com/research/developer-survey-2015#profile-gender>

- Exploring the data on gender and GitHub repo ownership. Alyssa Frazee. <http://alyssafrazee.com/gender-and-github-code.html>

How to create
more sustainable
open-source
communities?

Science is needed for evidence-based recommendations

Anecdotal evidence reliable? One man says “yes”.

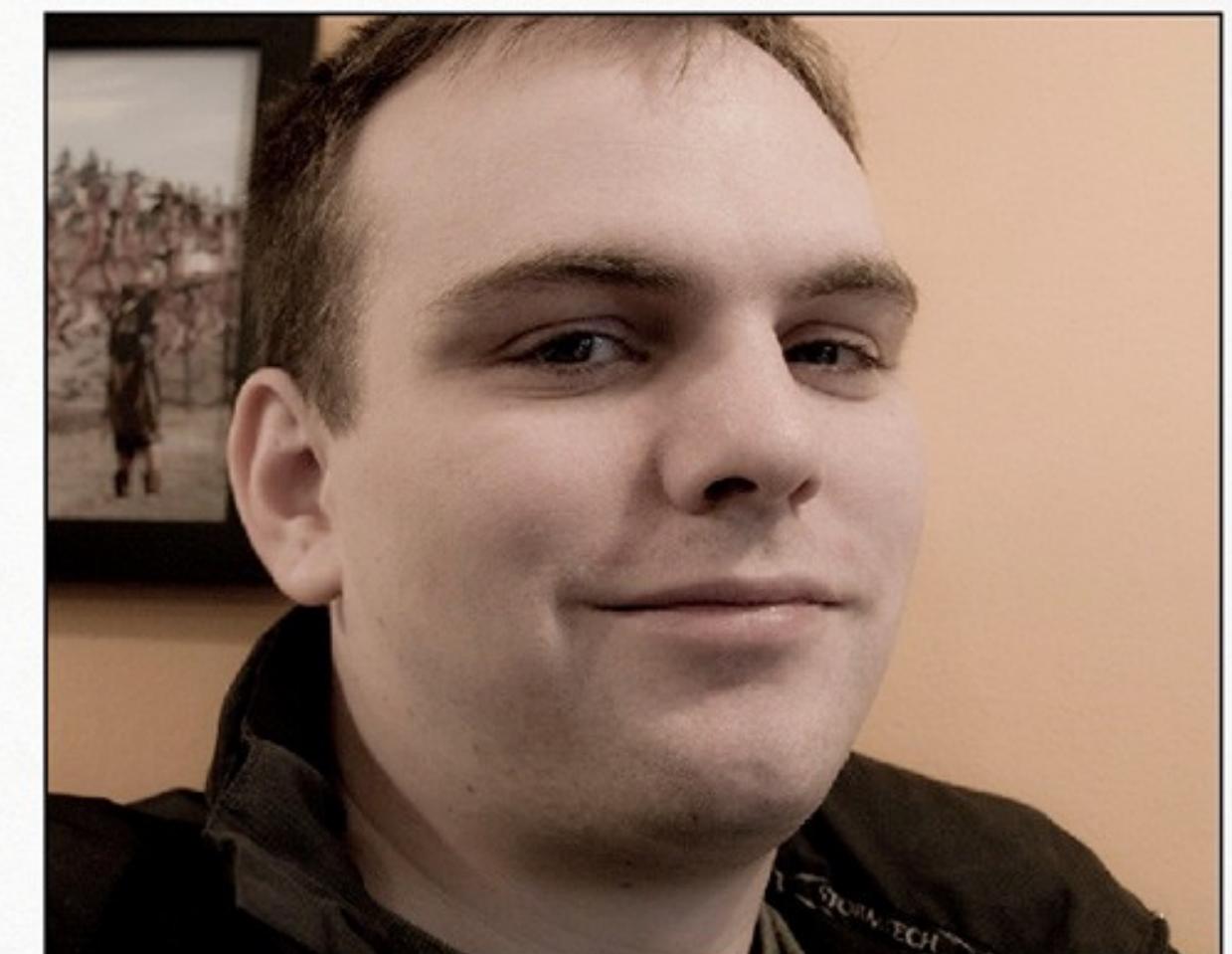
A STUDY CONDUCTED YESTERDAY by a man on himself concluded that self-reported anecdotal evidence is, in fact, both reliable and relevant.

The landmark study, conducted by Mark Mattingly of Virginia Beach in his apartment, concluded with 100% accuracy that data collected from personal experience can disprove other data conducted by reputable scientific institutions, thereby proving once and for all that “statistics can’t be trusted”.

In a press release Mr. Mattingly took aim at his detractors saying that “...this study shows what I’ve been telling people on the internet for years: all your fancy evidence and statistics don’t mean nothing in the real world.”.

A frequenter of internet forums, comment sections, and social media, Mr. Mattingly recounts that he was inspired to undertake the study when someone reportedly kept insisting that he provide evidence for his claims. “I think everyone’s entitled to an opinion, and that my opinion is worth just as much as anyone else’s” Mr. Mattingly said.

Academic types have criticised the study, and papers who are publishing it, saying that it lacks everything and makes no sense. When shown the study, Emeritus Professor James Albrecht of Carnegie Mellon University looked all confused and hopeless before making pining, guttural sounds.



Mr. Mattingly in his apartment looking all smug.

Mr. Mattingly has responded saying that this is just the first of many studies he intends to conduct, and that a meta-analysis of people who have opinions and anecdotal experiences independent of controls, methodological rigor, blinding and peer review are soon to be published, adding further weight to his initial findings.

Published Saturday 22 February 2014 by yourlogicalfallacyis.com/anecdotal

Photo: Weasello

My history

- Assistant Professor @CMU, SCS, ISR
 - STRUDEL: Socio-Technical Research Using Data Excavation Lab
- Researching open source for ~10 years
 - Empirical, socio-technical perspective



STRUDELers
(and strudels)
circa 2018

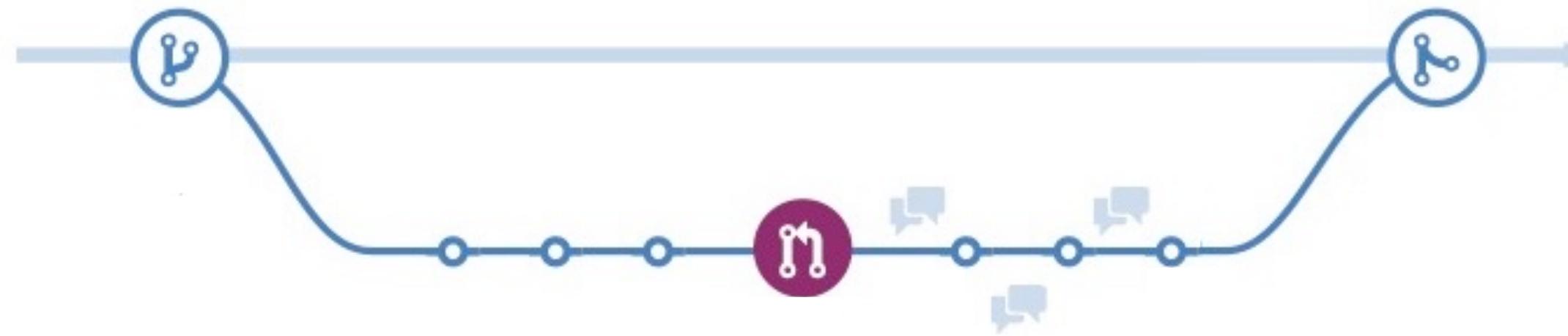
A great opportunity
for research

GitHub standardized the practices

Version control



The Pull Request model



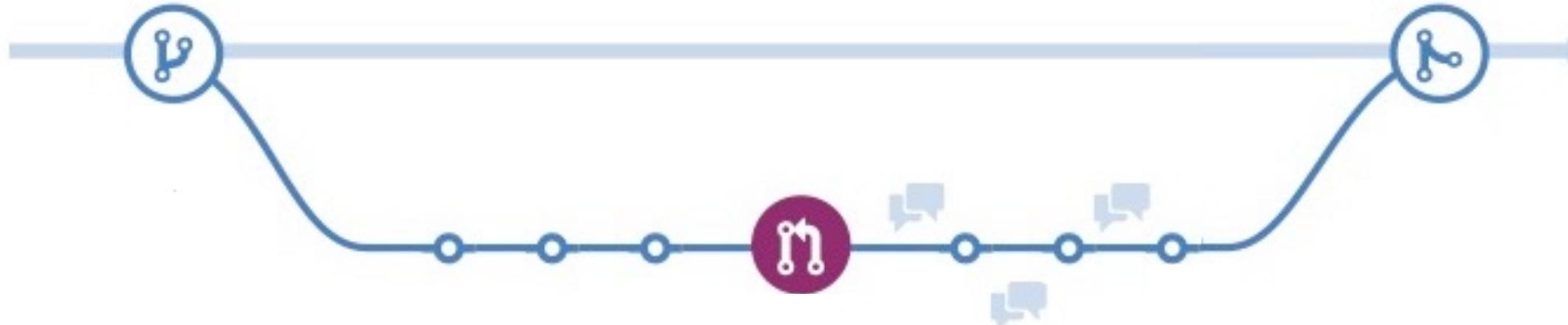
→ Uniform access to contribution data

GitHub standardized the practices

Version control



The Pull Request model



User profile pages

A screenshot of a GitHub user profile page for "Bogdan Vasilescu" (bvasiles). The page includes a profile picture of a man with glasses, a "Follow" button, and a list of popular repositories: "empirical-methods", "jsNaughty", "bvasiles.github.io", and "SuffixTree". The "empirical-methods" repository is described as the "Fall 2018 edition of 17-803 Empirical Methods". The "jsNaughty" repository is described as a "JS reverse minifier based on statistical machine translation". The "bvasiles.github.io" repository is described as "My website". The "SuffixTree" repository is described as "Forked from JDonner/SuffixTree". The page also shows follower and following counts, and links to the user's Carnegie Mellon University profile and GitHub page.

→ Uniform access to contribution and personal data

Heaps of data



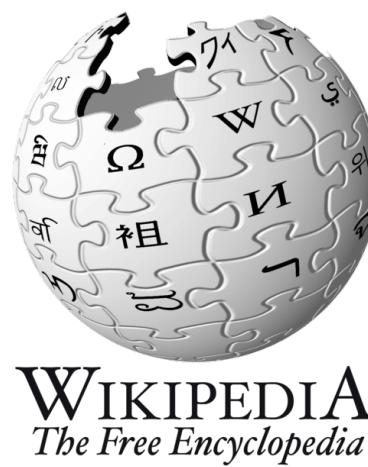
GitHub alone:

More than 50M people and 100M repositories hosted as of August 2019



Beyond GitHub:

“The collection of public Git repositories as a whole [...] exceeds 1.5PB” (Ma et al, 2019)



For reference: English Wikipedia

6M articles and 40M users as of August 2020

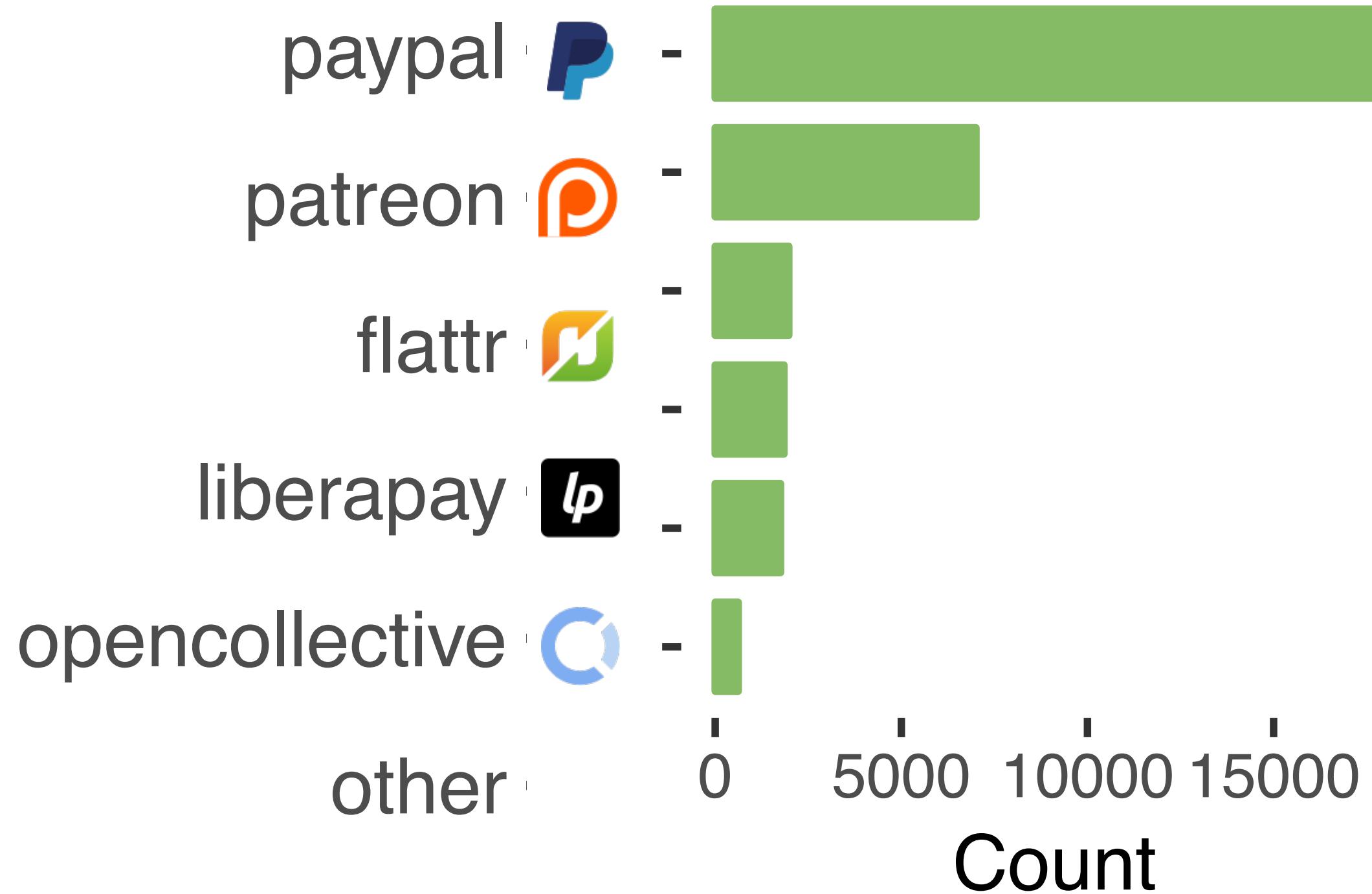
Ma, Y., Bogart, C., Amreen, S., Zaretzki, R., & Mockus, A. (2019, May). World of Code: An infrastructure for mining the universe of open source VCS data. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)* (pp. 143-154). IEEE.

A great opportunity for research

From anecdotes and small-sample studies to ecosystem-wide censuses and large-scale quantitative models



Overall, 0.04% of repos ask for donations



as of May 23, 2019

The data is naturally longitudinal



All events have timestamps:

- Commits
- Issues
- ...

The compiler for writing next generation JavaScript.

Gitpod ready-to-code

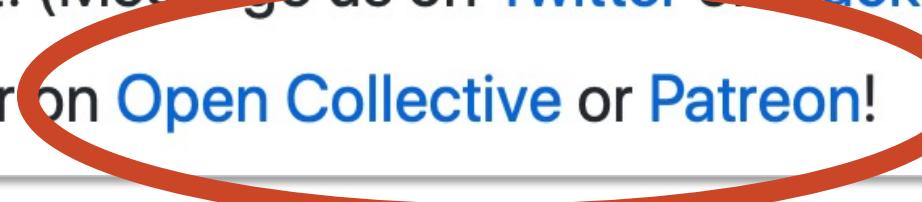
v7 downloads 74M/month v6 downloads 23M/month

travis passing circle passing coverage 91% slack 13112 Follow 47k

Supporting Babel

backers 640 sponsors 270 business model flavortown

ed "babble") is a community-driven project used by many companies and projects, a
unteers. If you'd like to help support the future of the project, please consider:
loper time on the project. (Message us on Twitter or Slack for guidance!)
s by becoming a sponsor on Open Collective or Patreon!



<https://github.com/babel/babel>

Therefore, one can:

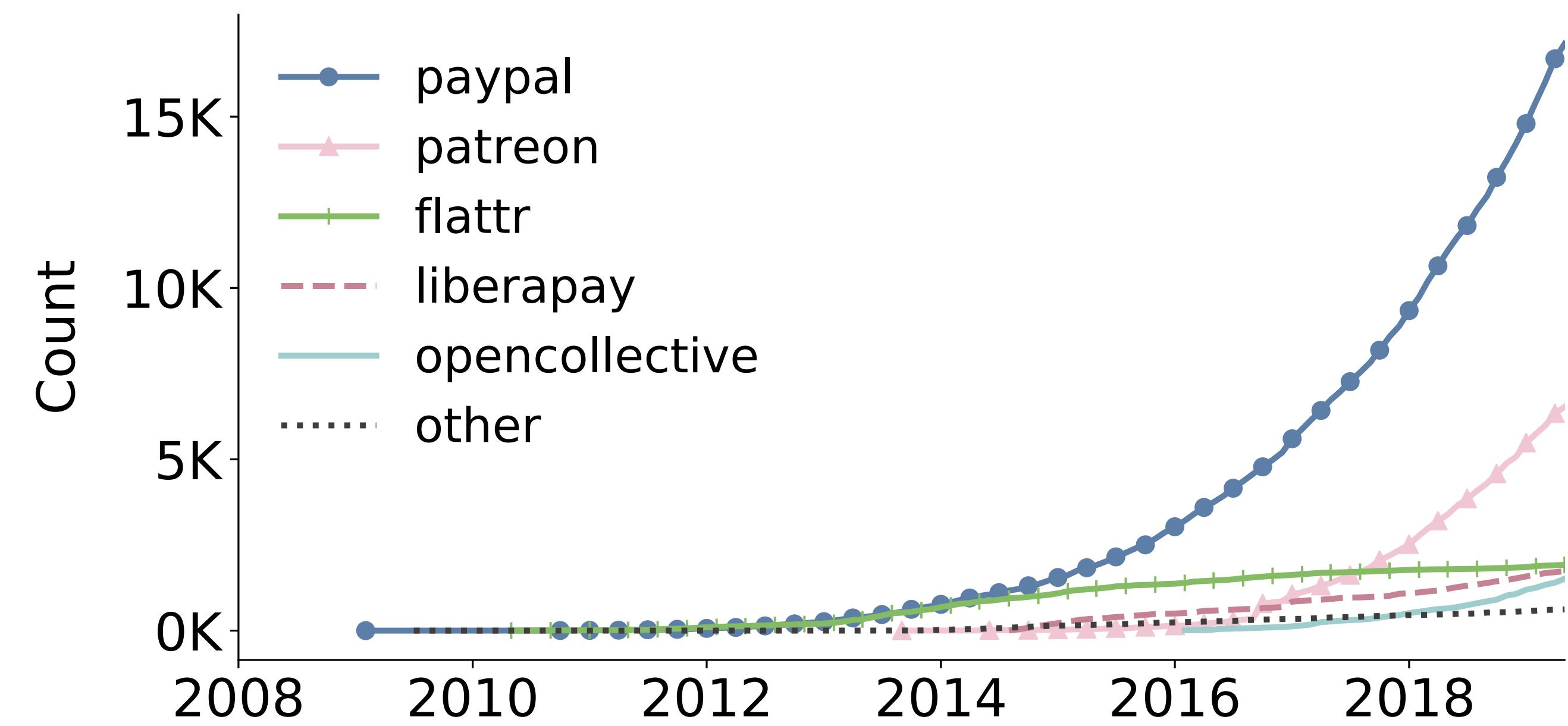
- Track changes to files
- Track people joining and leaving projects
- ...

A great opportunity for research

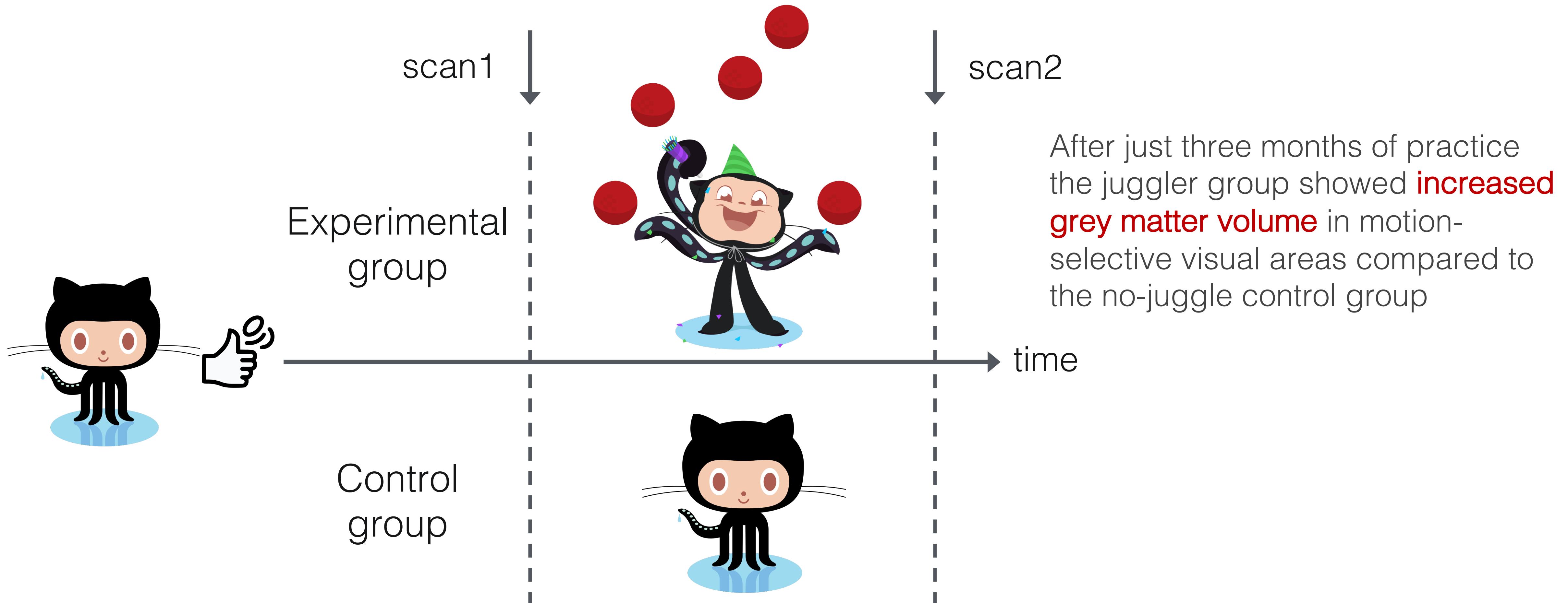
Capture and understand trends over time, analyze time series data



Adoption of donation platforms over time



Juggling as a sustainability intervention?



Bogdan Draganski, Christian Gaser, V. Busch, G. Schuierer, U. Bogdahn, and A. May. "Changes in grey matter induced by training." Nature 427, no. 6972 (2004): 311-312.

Natural experiments: interventions are outside the control of the researchers

Project started receiving donations

Project adopted a certain practice / tool

open collective

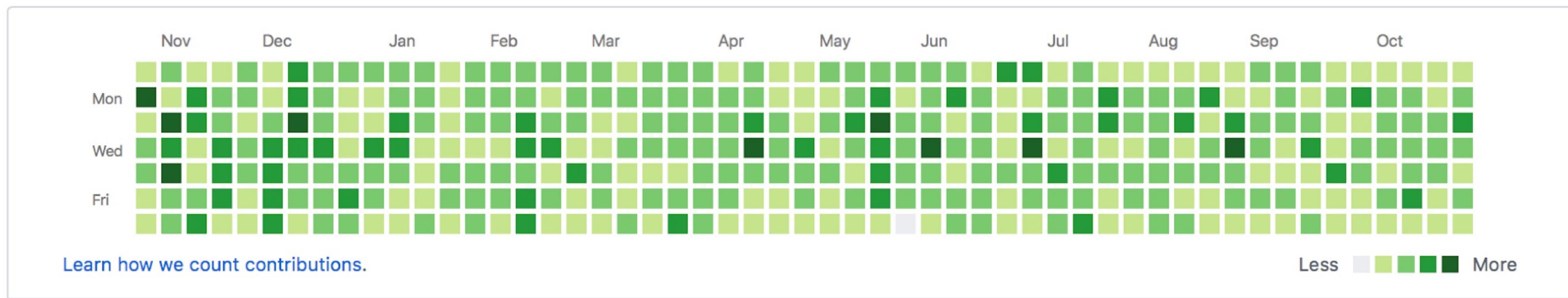
dependencies up to date



But be careful, the data is noisy!

- Hyperactive maintainer? No, bot

5,786 contributions in the last year



fossabot
fossabot

Follow

Your friendly neighborhood badge bot.
Sends PRs to your READMEs when
integrating tools from @fossas to track
scan status. Feedback? Contact
support@fossa.io!

But be careful, the data is noisy!

The screenshot shows two GitHub repository pages side-by-side, both for the `keylime` project.

Left Repository: `keylime / keylime`
Forked from `keylime/keylime`
Issues: 75, Pull requests: 10, Actions, Projects
Branches: master, 31 branches, 36 tags
Last commit by `Marcio Silva` and `mpeters`: Additional fix for the processing of "tpm_policy" (#699)
List of files and commits:

- .ci: make run_local.sh work with newer doc
- .github: Use tox for pylint
- archive: Rename common to config
- auto-ipsec: Remove lock in the command execution
- demo: Set default UUID as lowercase (#699)
- doc: Set default UUID as lowercase (#699)
- docker: Use Fedora 34 as base image for CI container
- ima_stub_service: Port License to Apache 2.0 (#329)
- keylime: Additional fix for the processing of "tpm_policy"
- media/images/logo: Add logos in dedicated media directory (#374)

Right Repository: `amylily1011 / keylime`
forked from `keylime/keylime`
Code, Issues, Pull requests, Actions, Projects, Wiki
Branches: master, This branch
Last commit by `luke`: jetwhiz Update README.md
List of files and commits:

- master: 1 branch, 1 tag

Note at the bottom: Note that this is an MIT Keylime copy released under the original BSD-2 license. The community-maintained version is available at <https://github.com/keylime/keylime>

Let's look at some
concrete examples

STRUDEL sustainability research on ...

Open-source projects

Project practices

- [ICSE 2020](#) (forking)
- [ESEC/FSE 2019](#) (forking)
- [ESEC/FSE 2018](#) (abandonment factors)
- [FSE 2016](#) (breaking changes)

Attracting contributors

- [MSR 2020](#) (Twitter)
- [CSCW 2019](#) (signals)
- [ESEC/FSE 2015](#) (social connections)

Funding models

- [ICSE 2020](#) (donations)

Transparency and signaling

- [ESEC/FSE 2020](#) (diffusion of practices)
- [ICSE 2018](#) (badges)

Open-source people

Stress, burnout, disengagement

- [ICSE NIER 2020](#) (toxic language)
- [ICSE 2019](#) (overwork)
- [OSS 2019](#) (dropout and survival analysis)

Diversity and inclusion

- [ICSE 2019](#) (social capital)
- [CHI 2015](#) (gender & tenure)
- [CHASE 2015](#) (survey)

Today's focus:
Transparency and signaling

Transparency is a defining characteristic of social coding platforms

This screenshot shows a GitHub profile page for a user named 'cv'. The profile features a cartoon cat icon holding a laptop with 'CV' on it. Key sections include:

- Popular repositories:** breakfast-repo (208 stars), x86-kernel (48 stars), jsconf-2015-deck (32 stars), ratpack (32 stars).
- Repositories contributed to:** npm/docs (44 stars), mozilla/publish.webmaker.org (2 stars), npm/marky-markdown (104 stars), artisan-tattoo/assistant-frontend (5 stars), npm/npm-camp (1 star).
- Public contributions:** A heatmap showing activity over time (Feb to Jan) across days of the week (M, W, F). Summary: 1,886 total contributions from January 24, 2015, to January 24, 2016.
- Contributions in the last year:** 1,886 total.
- Longest streak:** 37 days (October 7 – November 12).
- Current streak:** 7 days (January 18 – January 24).
- Organizations:** npm, inc.
- Statistics:** Joined on Oct 31, 2011, 776 Followers, 38 Starred, 15 Following.

This screenshot shows the GitHub repository page for 'caolan/async'. The repository has 23,937 stars and 2,203 forks. Key details include:

- Code:** 1,629 commits, 11 branches, 72 releases, 206 contributors.
- README.md:** Async utilities for node and the browser (<http://caolan.github.io/async/>)
- Tags:** javascript, async, callbacks.
- Build status:** build passing.
- Dependencies:** npm v2.6.0, coverage 99%, gitter, join chat, examples 26348, jsDelivr, 407k hits/month.
- Description:** Async is a utility module which provides straight-forward, powerful functions for working with asynchronous JavaScript. Although originally designed for use with Node.js and installable via `npm install --save async`, it can also be used directly in the browser.

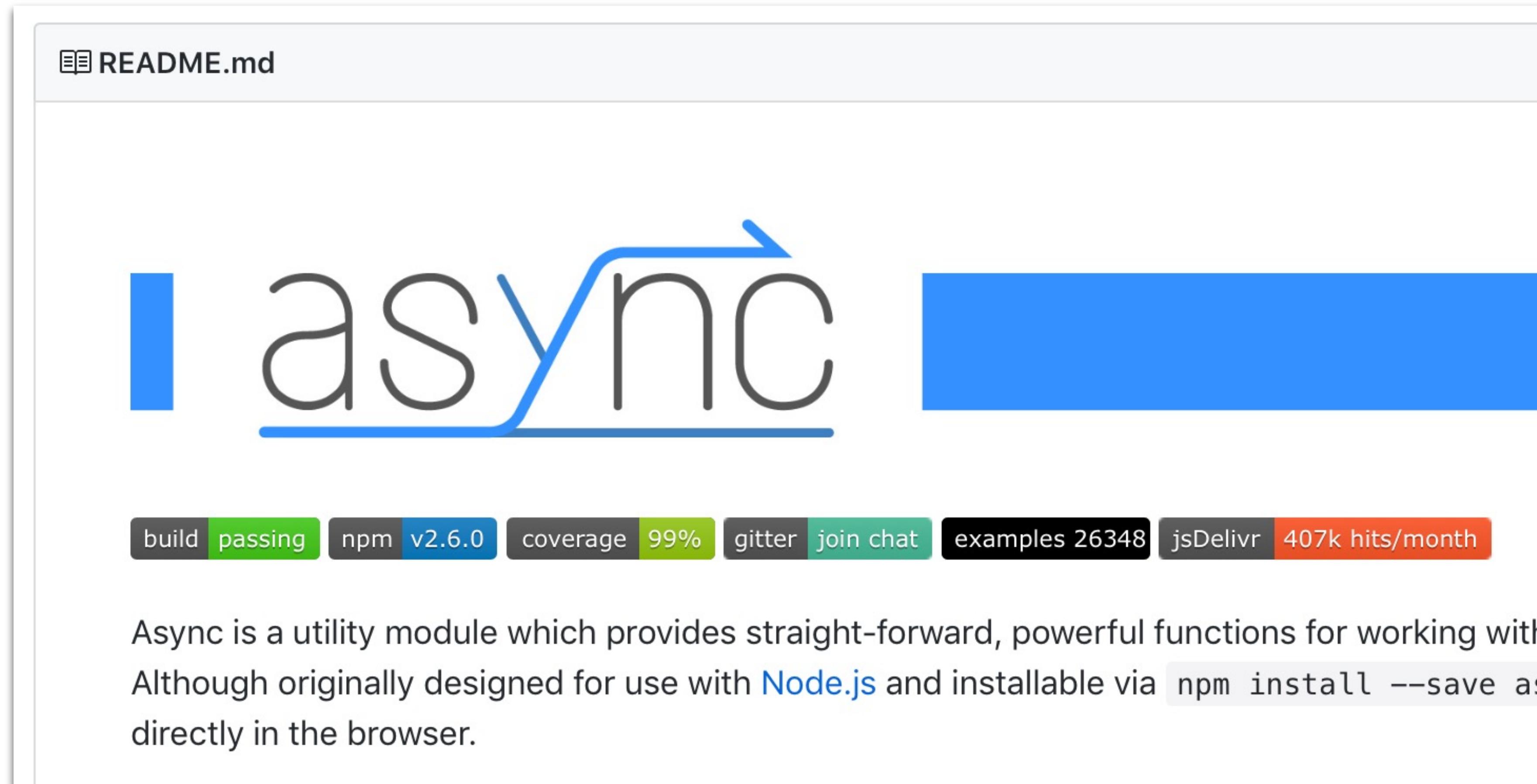
- Rich inferences about people's expertise and level of commitment

- Impacts collaboration, but also recruiting and hiring

(Dabbish et al. 2012), (Marlow et al. 2013), (Marlow and Dabbish 2013)

Some signals are customizable

- E.g., repository badges



- Adding Sparkle to Social Coding: An Empirical Study of Repository Badges in the npm Ecosystem. Trockman, A., Zhou, S., Kästner, C., and Vasilescu, B. *ICSE 2018*

Some signals are customizable

README.md

PANDA

 Parallel Tests failing  Build and Publish Docker Container and Pypanda Docs failing

PANDA is an open-source Platform for Architecture-Neutral Dynamic Analysis. It is an emulator, and so analyses have access to all code executing in the guest and all data and replay executions, enabling iterative, deep, whole system analyses. Further, the results are shareable, allowing for repeatable experiments. A nine billion instruction boot of FreeBSD on a few hundred MB. PANDA leverages QEMU's support of thirteen different CPU architectures and diverse instruction sets possible within the LLVM IR. In this way, PANDA can have a single analysis engine that precisely supports many CPUs. PANDA analyses are written in a simple mechanism to share functionality between plugins, increasing analysis code re-use and development.

It is currently being developed in collaboration with MIT Lincoln Laboratory, NYU, and released under the [GPLv2 license](#).

README.md

Keylime

[build](#) [error](#) [chat](#) [on gitter](#)



Keylime is an open-source scalable trust system harnessing TPM Technology.

What badges signal to viewers

build passing

codacy A

dependencies out of date

vulnerabilities 0

code climate 4.0

coverage 94%

docs

issue resolution 3 h

Survey

32 Maintainers

- What do you intend to signal?
- What effects do you expect?

57 Contributors

- What do badges tell you?

“welcoming contributions”

“expectations of contribution quality”

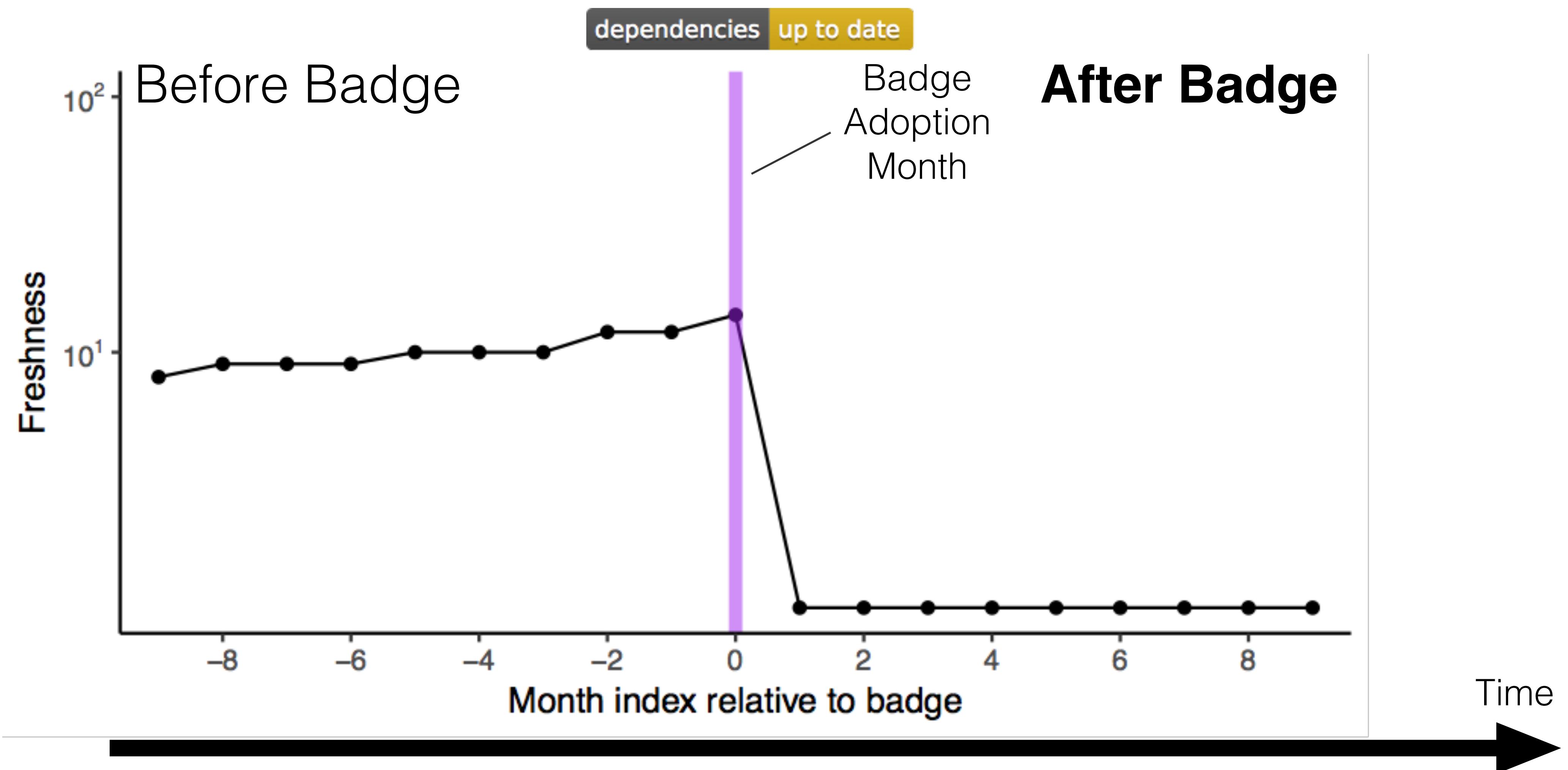
“dedicated to offering support”

“reduced chances of conflicting versions of dependencies”

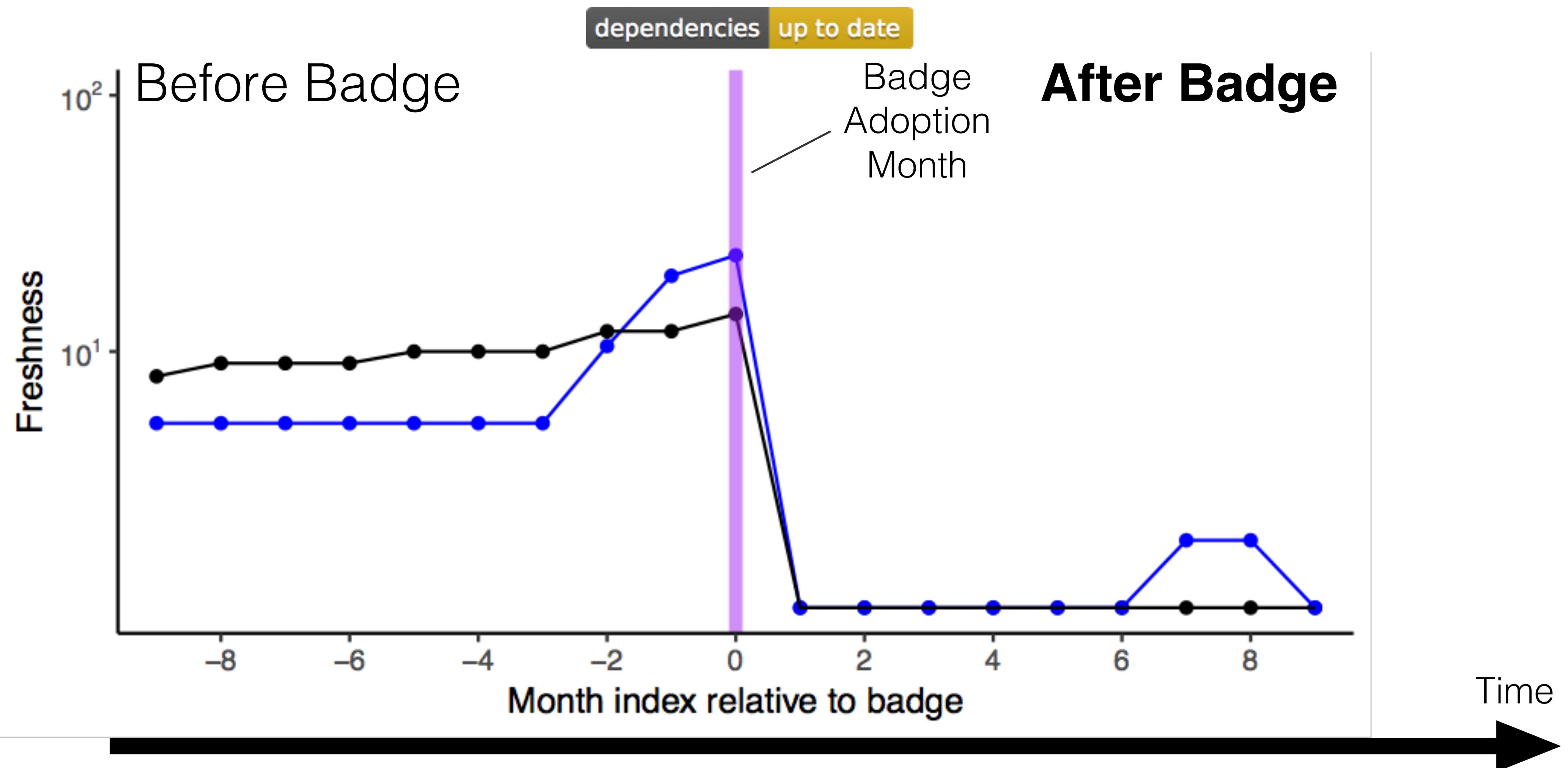
“indicator of product quality”

- Adding Sparkle to Social Coding: An Empirical Study of Repository Badges in the npm Ecosystem. Trockman, A., Zhou, S., Kästner, C., and Vasilescu, B. *ICSE 2018*

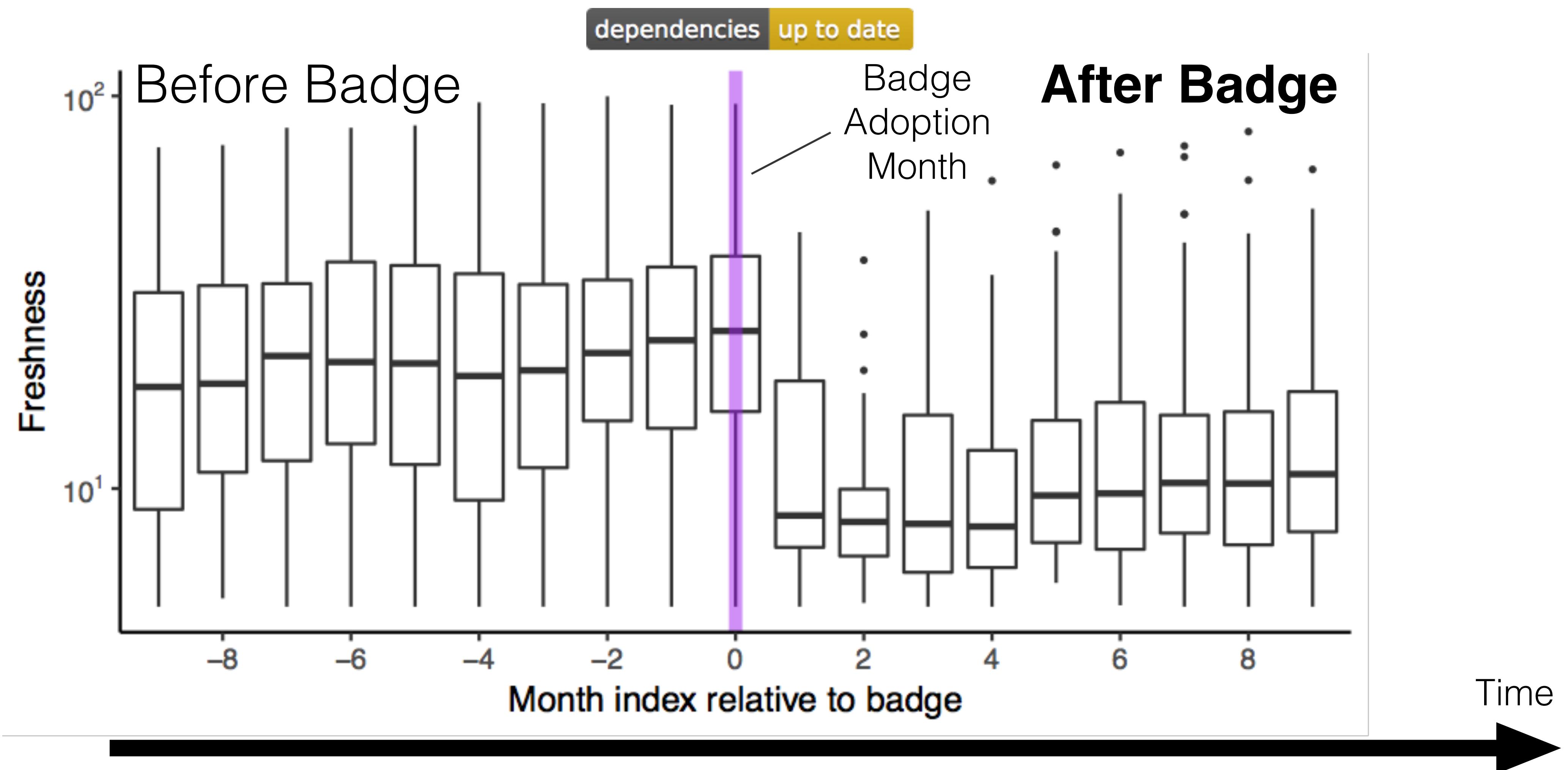
Time Series Analysis



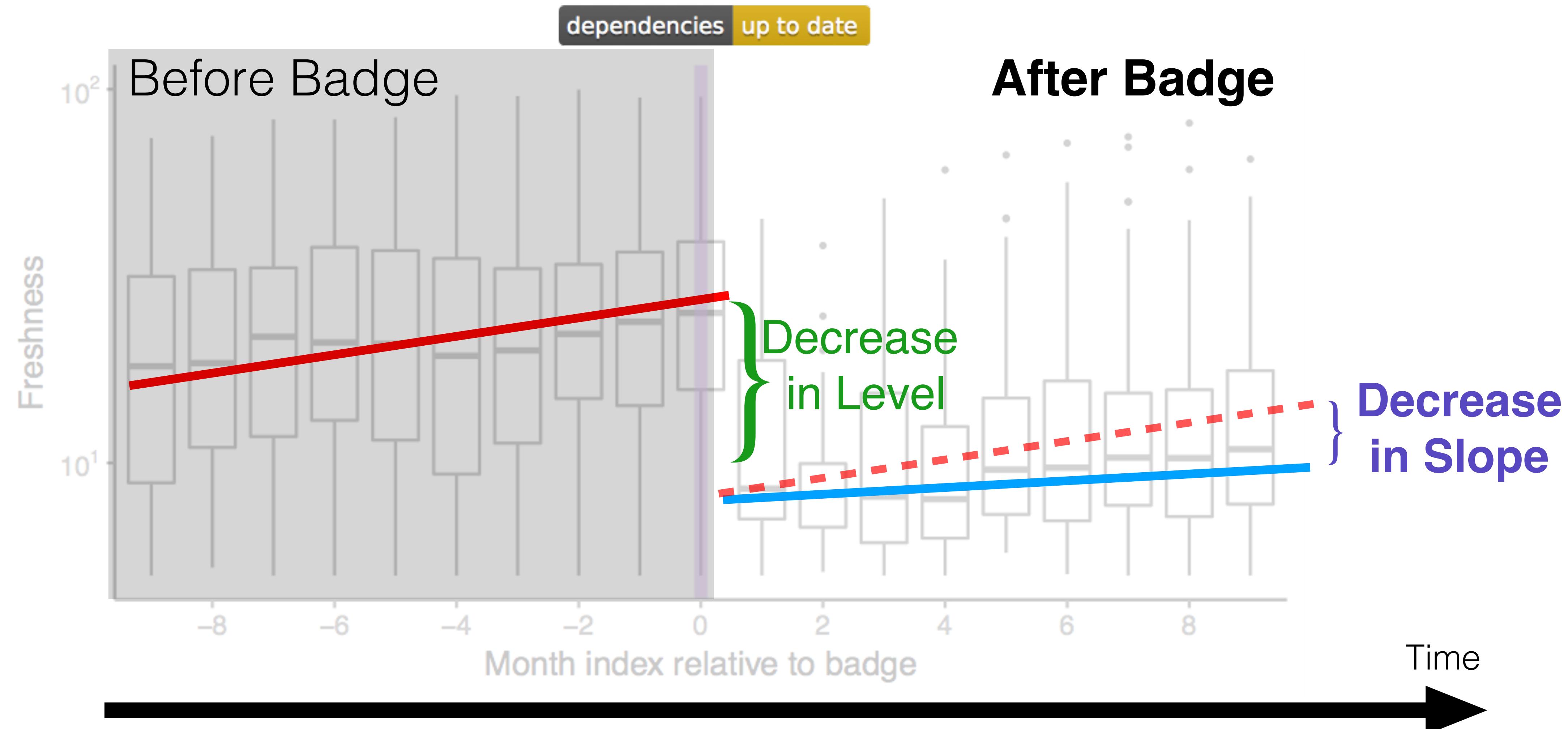
Time Series Analysis



Time Series Analysis



Time Series Analysis



Statistical multi-variate analysis

Basic Model		Full Model		RDD	
response: <i>freshness</i> = 0	17.3% deviance explained	response: <i>freshness</i> = 0	17.4% deviance explained	response: $\log(freshness)$	$R_m^2 = 0.04, R_c^2 = 0.35$
Coeffs (Err.)	LR Chisq	Coeffs (Err.)	LR Chisq	Coeffs (Err.)	Sum sq.
(Inter.) 3.54 (0.03)***		3.50 (0.03)***		1.45 (0.09)***	
Dep. -1.78 (0.01)***	32077.8***	-1.79 (0.01)***	32292.8***	-0.04 (0.02)	3.01
RDep. 0.22 (0.01)***	610.3***	0.21 (0.01)***	560.6***	-0.01 (0.02)	0.11
Stars -0.08 (0.00)***	301.4***	-0.09 (0.00)***	311.2***	0.00 (0.01)	0.00
Contr. -0.24 (0.01)***	500.5***	-0.25 (0.01)***	548.7***	-0.04 (0.02)*	4.39*
lastU -0.65 (0.01)***	12080.9***	-0.64 (0.01)***	11537.9***	0.01 (0.02)	0.37
hasDM		0.24 (0.03)***	116.1***	0.45 (0.08)***	2.43
hasInf		0.11 (0.02)***	48.3***	0.04 (0.05)	0.45
hasDM:hasInf		-0.05 (0.04)	1.9	-0.32 (0.10)**	
hasOther		0.01 (0.01)			
time				0.03 (0.00)***	82.99***
intervention				-0.93 (0.03)***	1373.22***
time_after_intervention				0.11 (0.00)***	455.56***
time_after_intervention:hasDM				-0.10 (0.01)***	230.36***
time_after_intervention:hasInf				-0.00 (0.01)	1.14
time_after_intervention:hasDM:hasInf				0.03 (0.01)**	10.62**

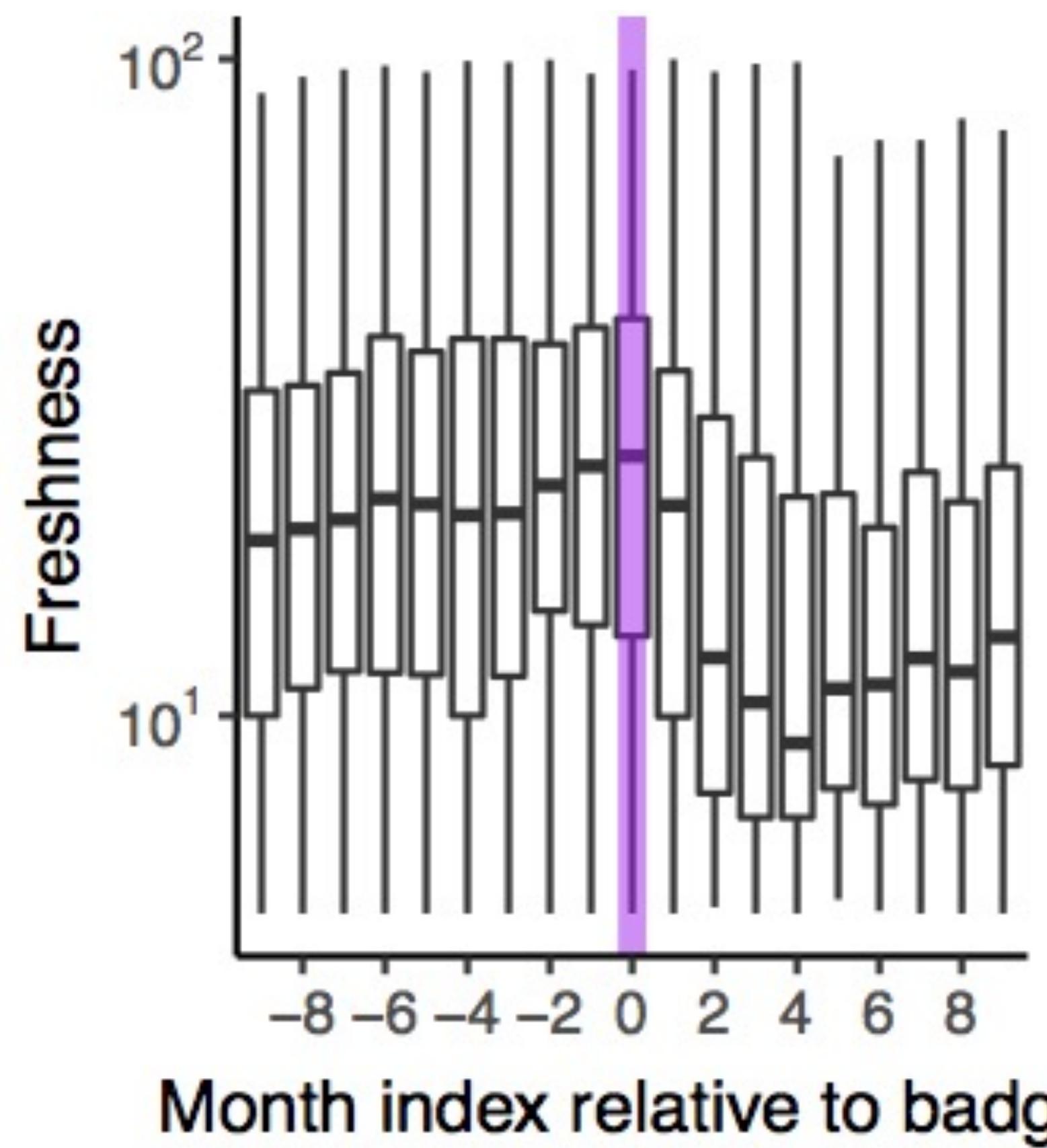
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$;

Dep: dependencies; RDep: dependents; Contr.: contributors; lastU: time since last update;
 hasDM: has dependency-manager badge; hasInf: has information badge; hasOther: adopts
 additional badges within 15 days

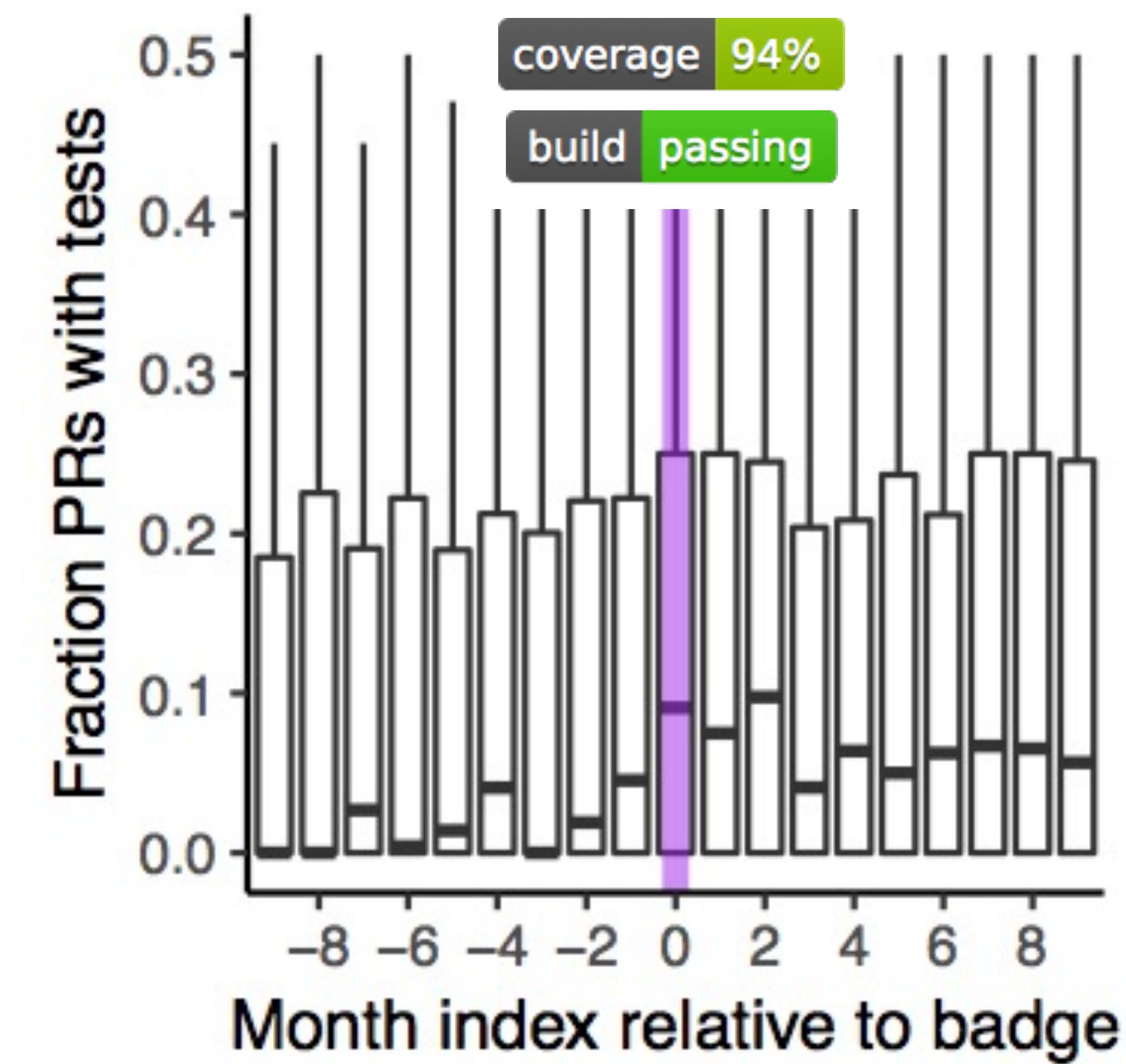
Badges are Reliable Signals

Mostly

dependencies up to date
up-to-date and secure dependencies

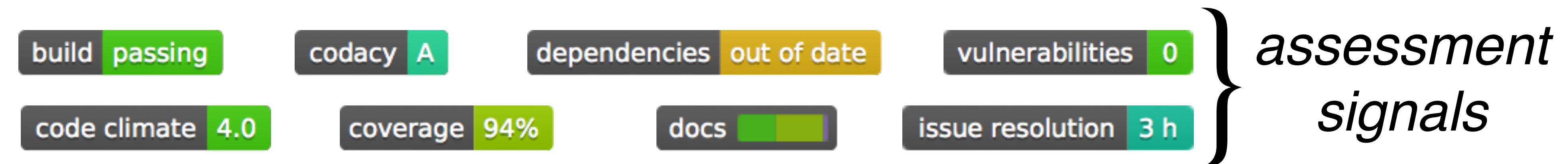


build passing + coverage 94%
tests in PRs

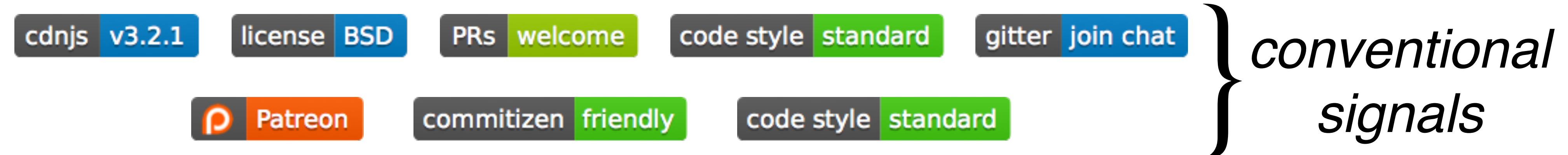


Take-away: Prefer “assessment” badges

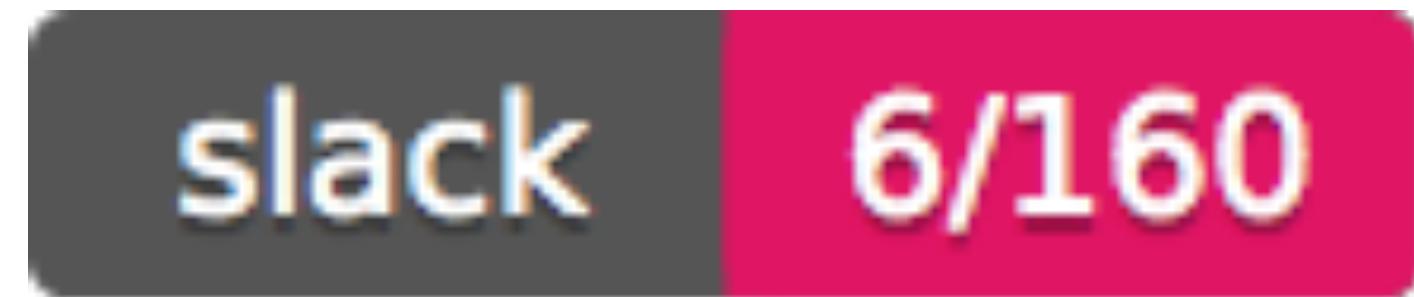
Badges with underlying analyses:



are **stronger predictors** than badges that merely state intentions or provide links:



Take-away: Prefer “assessment” badges



Assessment signal

>



Conventional signal

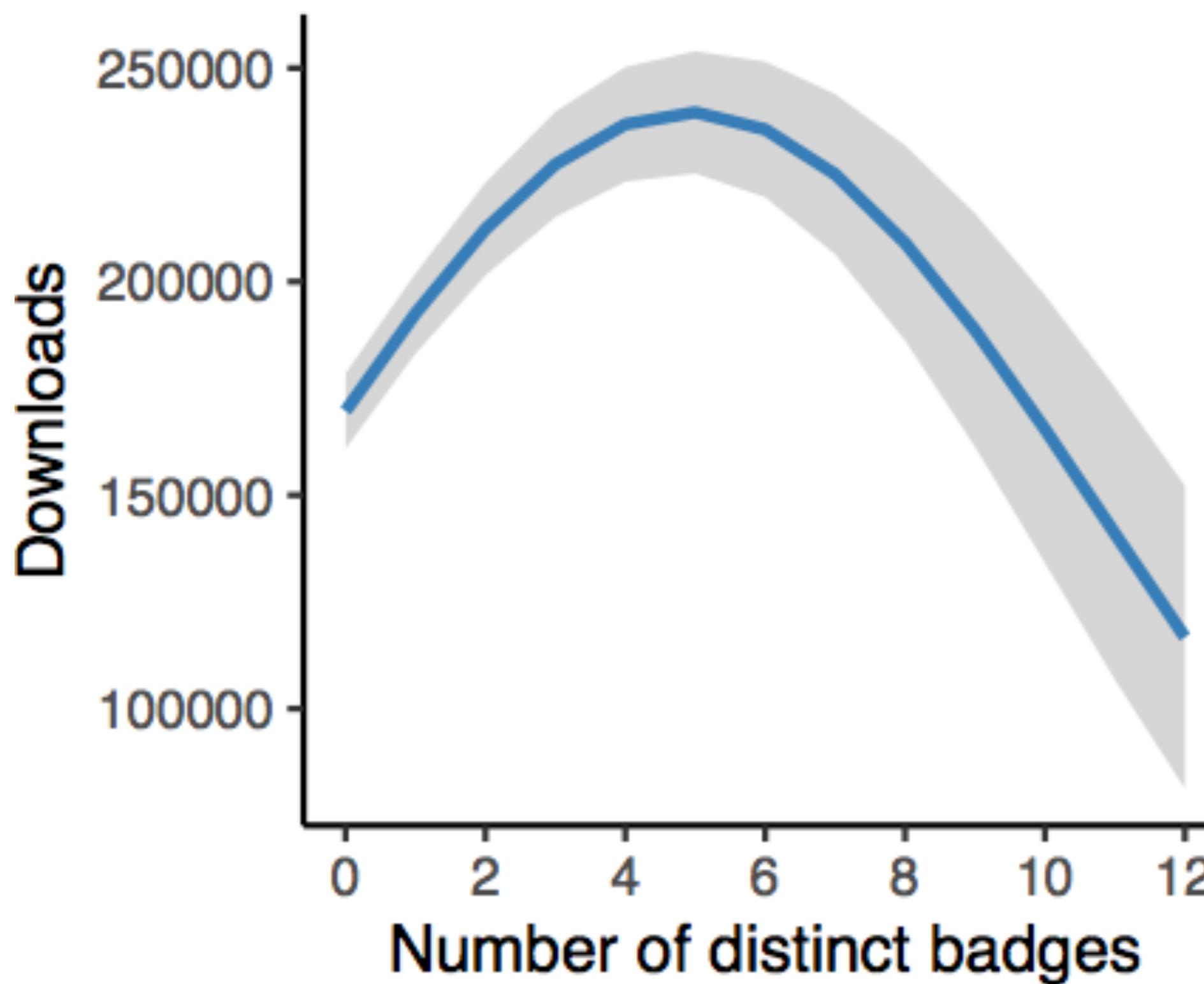
Keylime

build error chat on gitter



Take-away: Don't add too many

Attractiveness wears off beyond 5 badges



“It’s most important that the people seem nice”

How do people choose which project to contribute to?

The **tone of the community** is an important factor in both interviews and model.

maintainers polite ?

🔥 Problem convos

Here are some conversations you should probably check in on

- [outline/outline#2307 \(comment\)](#)
- [outline/outline#2303 \(comment\)](#)
- [outline/outline#2298 \(comment\)](#)

Interviews:

15 GitHub users

Data:

~10K npm packages

Model:

Logistic regression
(has new contributors)

- The Signals that Potential Contributors Look for When Choosing Open-source Projects.
Qiu, S., Li, Yucen., Padala, S., Sarma, A., and Vasilescu, B. *CSCW 2019*

“It’s most important that the people seem nice”

How do people choose which project to contribute to?

The **tone of the community** is an important factor in both interviews and model.

maintainers polite ?

Asking for help explicitly is an important factor in the interviews.

PRs welcome help wanted ?

"help wanted" issues 20 open

Interviews:

15 GitHub users

Data:

~10K npm packages

Model:

Logistic regression
(has new contributors)

- The Signals that Potential Contributors Look for When Choosing Open-source Projects.
Qiu, S., Li, Yucen., Padala, S., Sarma, A., and Vasilescu, B. *CSCW 2019*

“It’s most important that the people seem nice”

How do people choose which project to contribute to?

First Timers Support

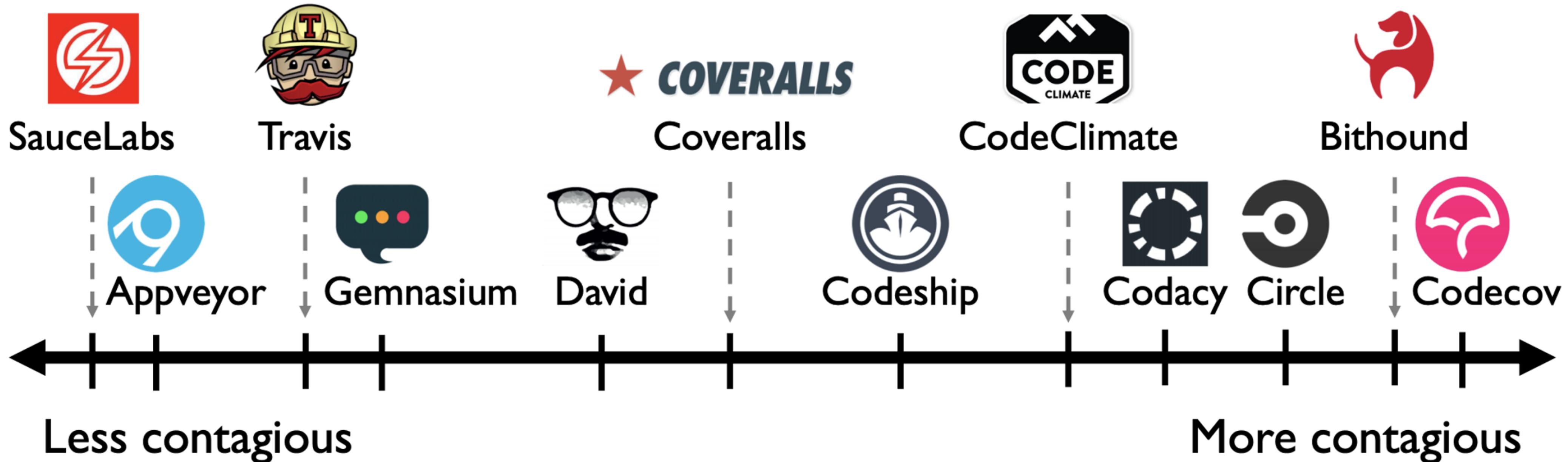
We welcome new contributors to Keylime of any form, including those of you who maybe new to working in an open source project.

So if you are new to open source development, don't worry, there are a myriad of ways you can get involved in our open source project. As a start, try exploring issues with `good first issue` label. We understand that the process of creating a PR can be a barrier for new contributors. These issues are reserved for new contributors like you. If you need any help or advice in making the PR, feel free to jump into our [gitter channel](#) and ask for help there.

Your contribution is our gift to make our project even more robust. Check out [CONTRIBUTORS.md](#) to find out more about how to contribute to our project.

Does transparency also influence
tool adoption?

Empirically, some tools spread faster than others



Contagious*: How quickly the tools spread and get adopted by different projects.

*Computed using the Bass model of innovation diffusion.

Lamba, H., Trockman, A., Armanios, D., Kästner, C., Miller, H., & Vasilescu, B. Heard it through the Gitvine: an empirical study of tool diffusion across the npm ecosystem. ESEC/FSE 2020.

Qualitative studies of tool adoption have hinted at social influence

“Developers are more likely to adopt tools they learn about from their peers than ones they learn about elsewhere”

(Xiao, Witschey, and Murphy-Hill, CSCW 2014)

“Developers who interact with security teams as part of the auditing process are more likely to adopt security tools”

(Witschey et al, ESEC/FSE 2015)

On social coding platforms (e.g., GitHub), there are clear opportunities to observe tool use

The screenshot shows the GitHub repository page for 'caolan/async'. At the top, it displays the repository name 'caolan / async' and various metrics: Watch (715), Star (26.8k), Fork (2.4k). Below the header, there are navigation links for Code, Issues (4), Pull requests (1), Actions, Projects, Wiki, Security, and Insights. The main content area is titled 'README.md' and features a large logo with the word 'async' in a stylized font where the 'y' has a blue arrow pointing right. Below the logo, there's a green character icon holding a magnifying glass over the logo. A horizontal bar at the bottom contains status badges for build (passing), Azure Pipelines (succeeded), npm (v3.2.0), coverage (99%), gitter (join chat), and jsDelivr (401k hits/month). The description text reads: 'Async is a utility module which provides straight-forward, powerful functions for working with asynchronous JavaScript. Although originally designed for use with Node.js and installable via `npm i async`, it can also be used directly in the browser. A ESM/MJS version is included in the main `async` package that should automatically be used with compatible bundlers such as Webpack and Rollup.'

On social coding platforms (e.g., GitHub), there are clear opportunities to observe tool use ...

priorityQueue: Prevent same tick setImmediate #1727 Open with ▾

Merged nearly merged 1 commit into `caolan:master` from `pkarimov:priority_queue` 15 days ago

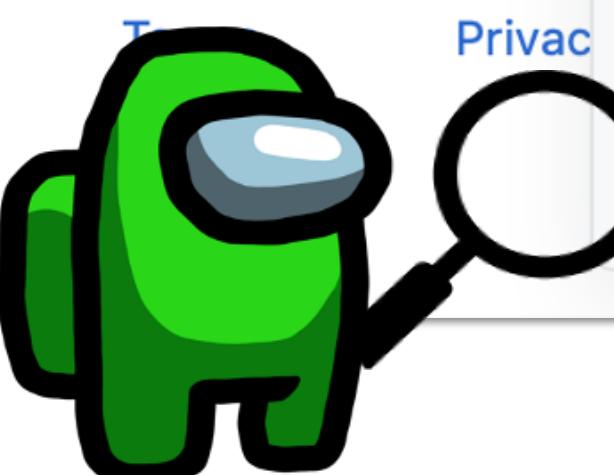
Conversation 1 Commits 1 Checks 1 Files changed 1 +8 -1

Commits on Oct 13, 2020

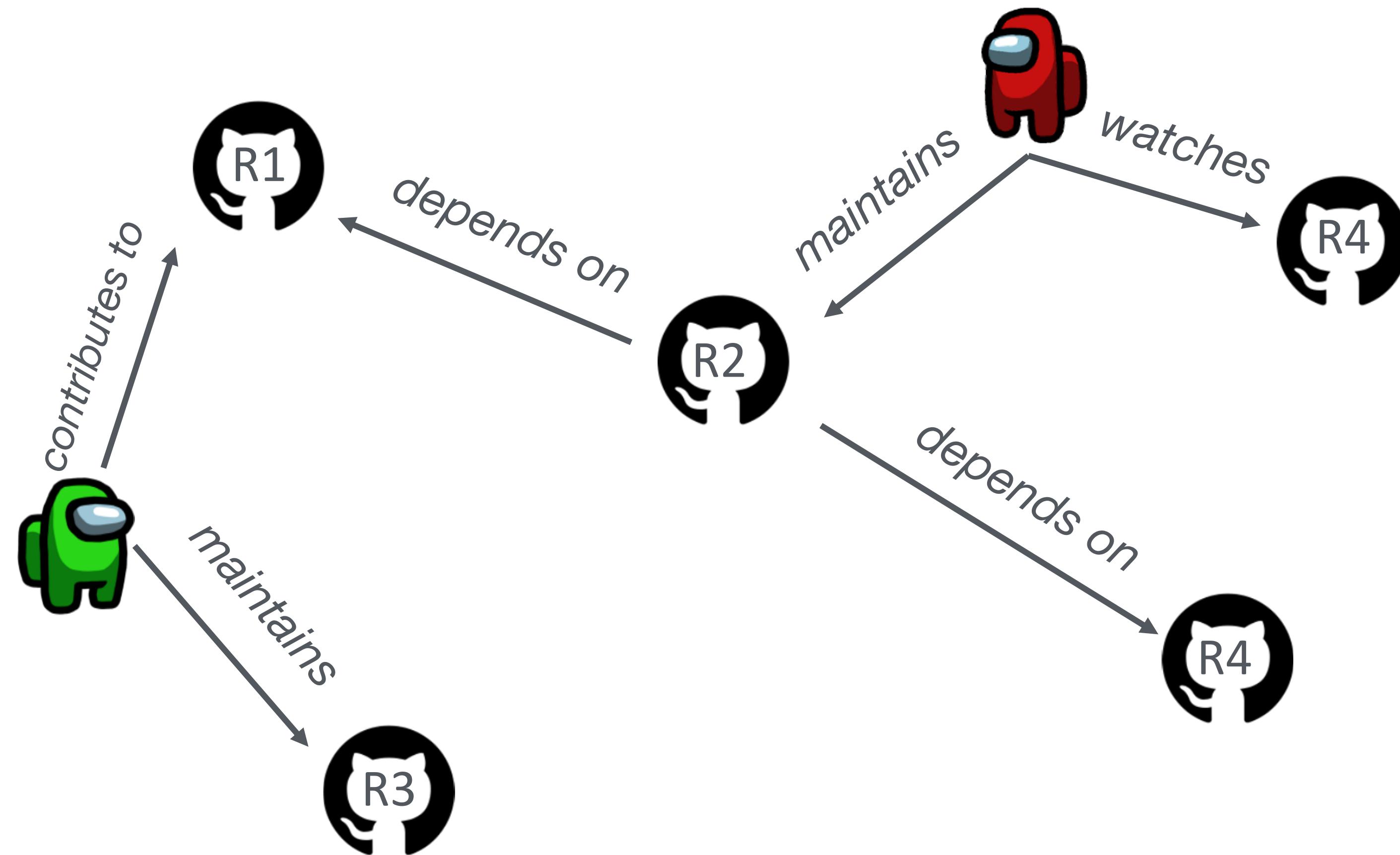
priorityQueue: Prevent same tick setImmediate
paulk-msft committed 20 days ago ×

Some checks were not successful
1 cancelled and 2 successful checks

!  caolan.async — Build #20201013.2 was canceled	Details
✓  continuous-integration/travis-ci/pr — The Travis ...	Details
✓  coverage/coveralls — Coverage increased (+0.003...)	Details

© 2020 GitHub, Inc.   Privacy API Training Blog About

... and to spread those practices across projects



Overview: Multiple case study

12 popular quality assurance tools

<i>Continuous integration</i>	<i>Dependency management</i>	<i>Code coverage reporters</i>	<i>Cross browser testers</i>
build passing	dependencies up to date	coverage 94%	Firefox Chrome 82 7 ✓ 86 7 ✓
Travis Circle Appveyor Codeship	David Bithound Gemnasium	Coveralls Codeclimate Codecov Codacy	Saucelabs

Overview: Multiple case study

12 popular quality assurance tools

Continuous integration

build passing

Dependency management

dependencies up to date

Code coverage reporters

coverage 94%

Cross browser testers

Firefox Chrome

82 7 ✓ 86 7 ✓

Travis
Circle
Appveyor
Codeship

David
Bithound
Gemnasium

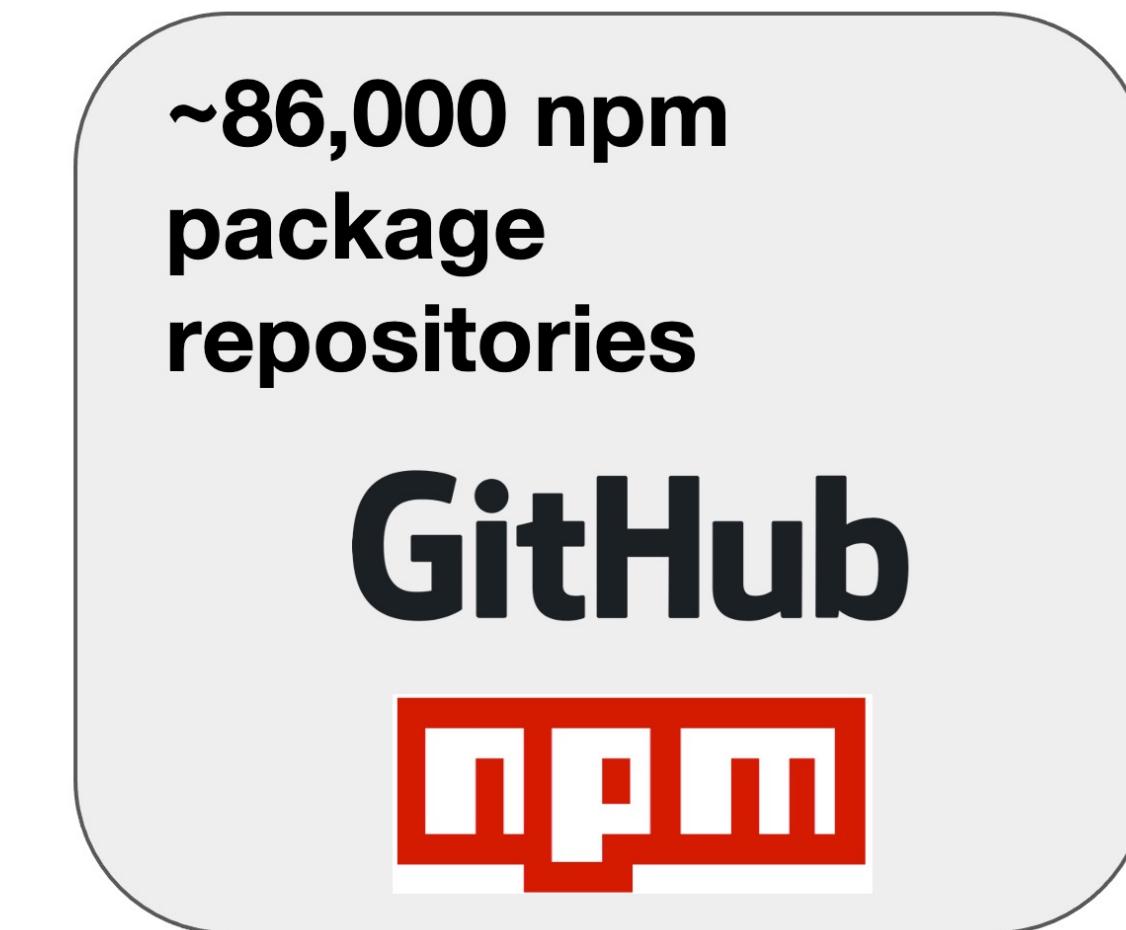
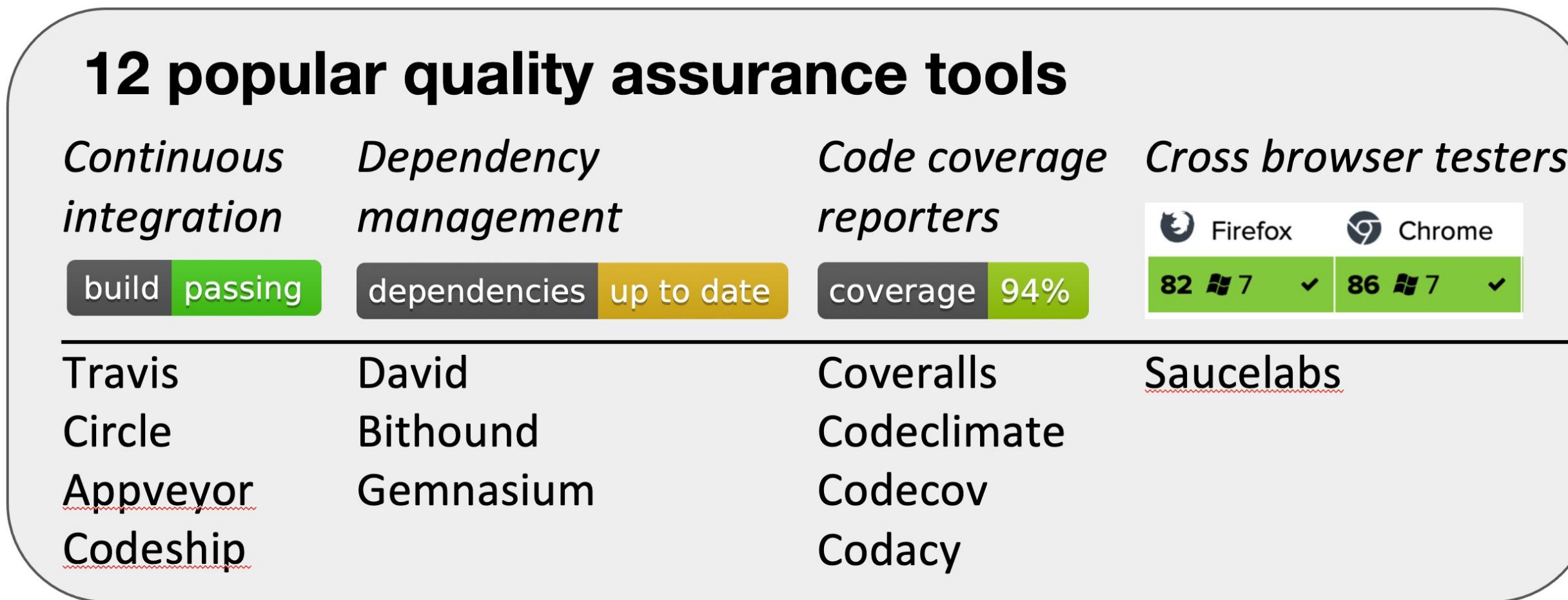
Coveralls
Codeclimate
Codecov
Codacy

Saucelabs

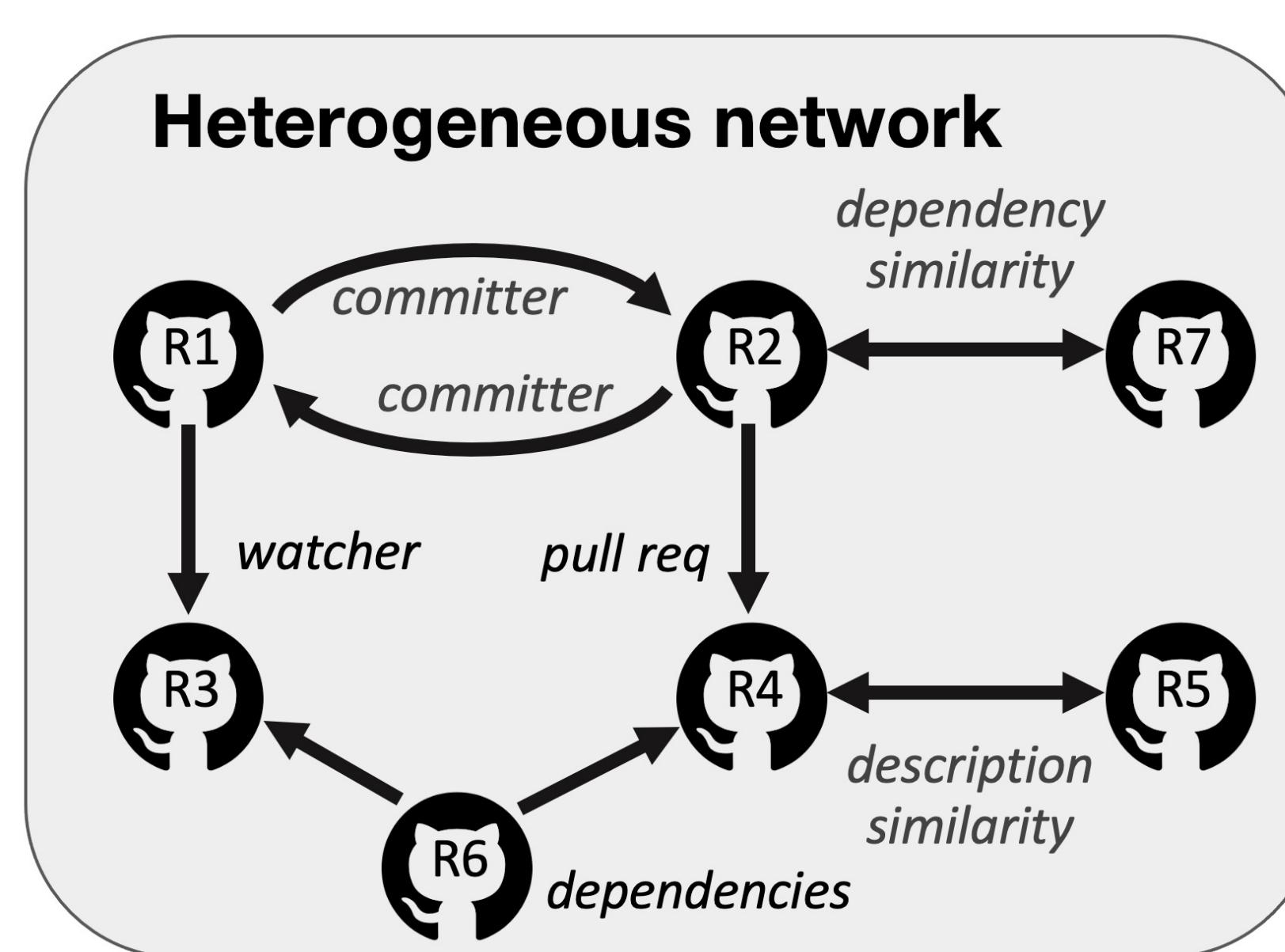
~86,000 npm
package
repositories



Overview: Multiple case study



For each tool:



Overview: Multiple case study

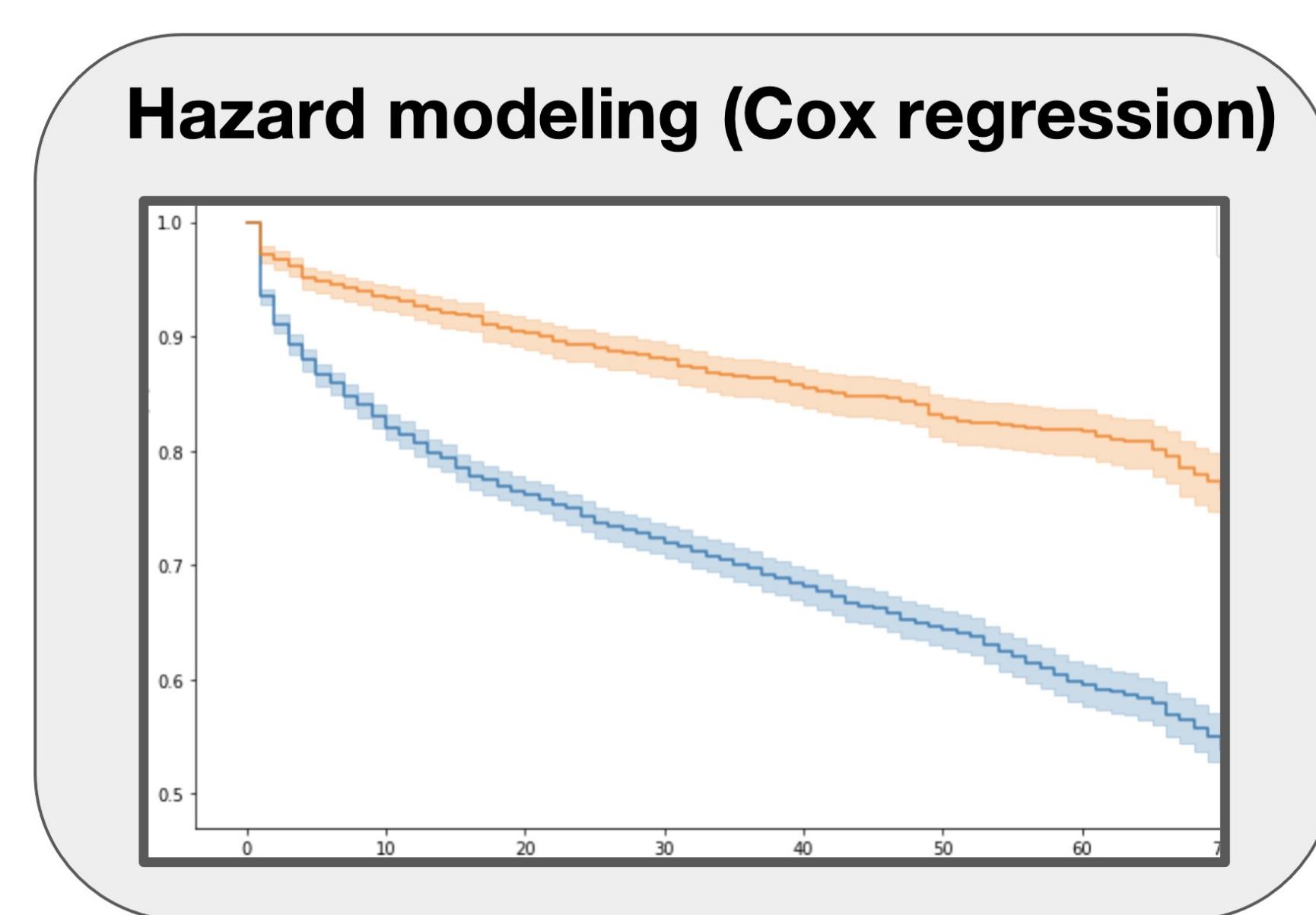
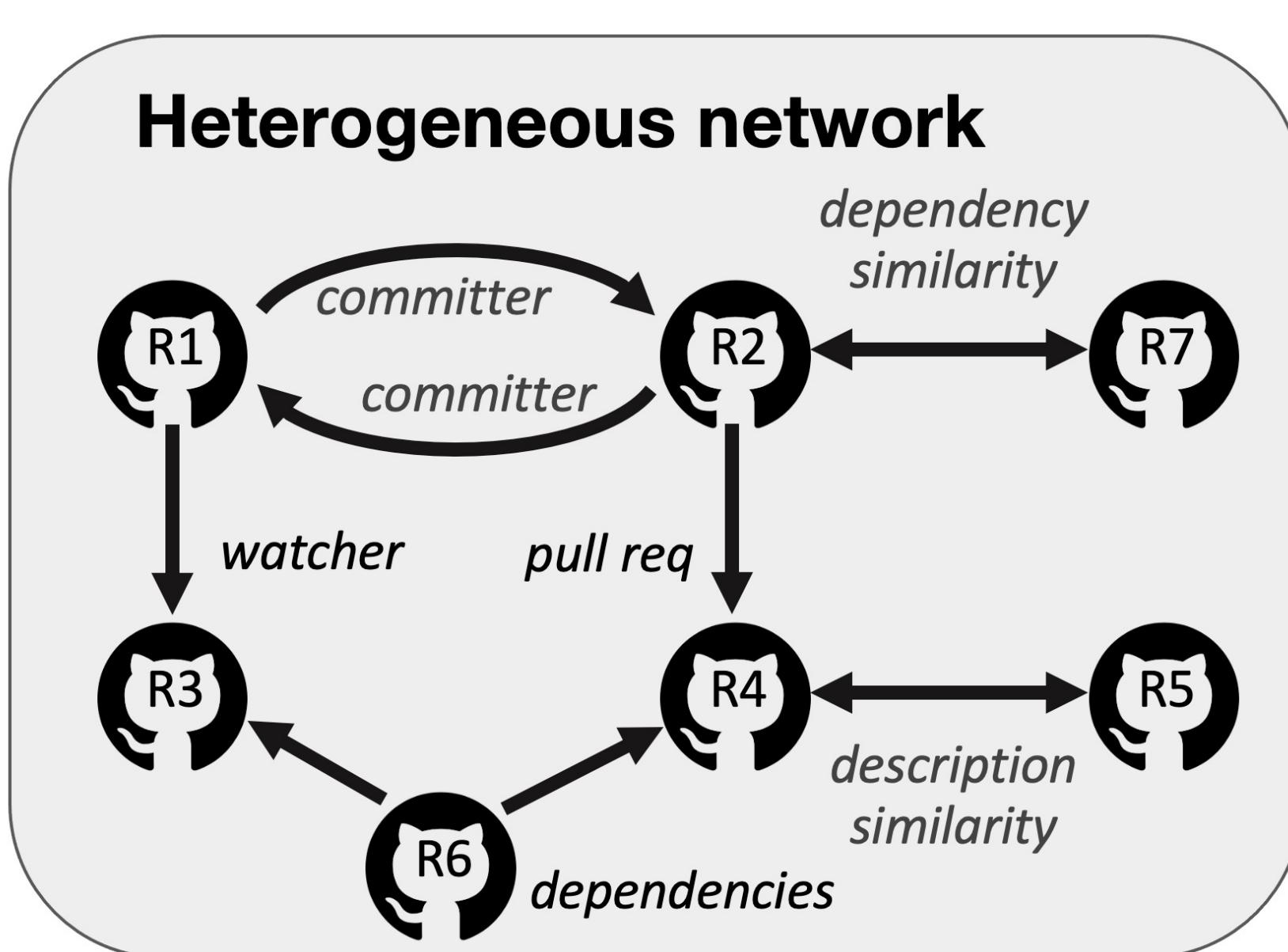
12 popular quality assurance tools

Continuous integration	Dependency management	Code coverage reporters	Cross browser testers
build passing	dependencies up to date	coverage 94%	Firefox Chrome 82 7 ✓ 86 7 ✓
Travis Circle Appveyor Codeship	David Bithound Gemnasium	Coveralls Codeclimate Codecov Codacy	Saucelabs

~86,000 npm package repositories

GitHub
npm

For each tool:



Key takeaway: The stronger the social connections, the more likely to adopt

Overview: Multiple case study

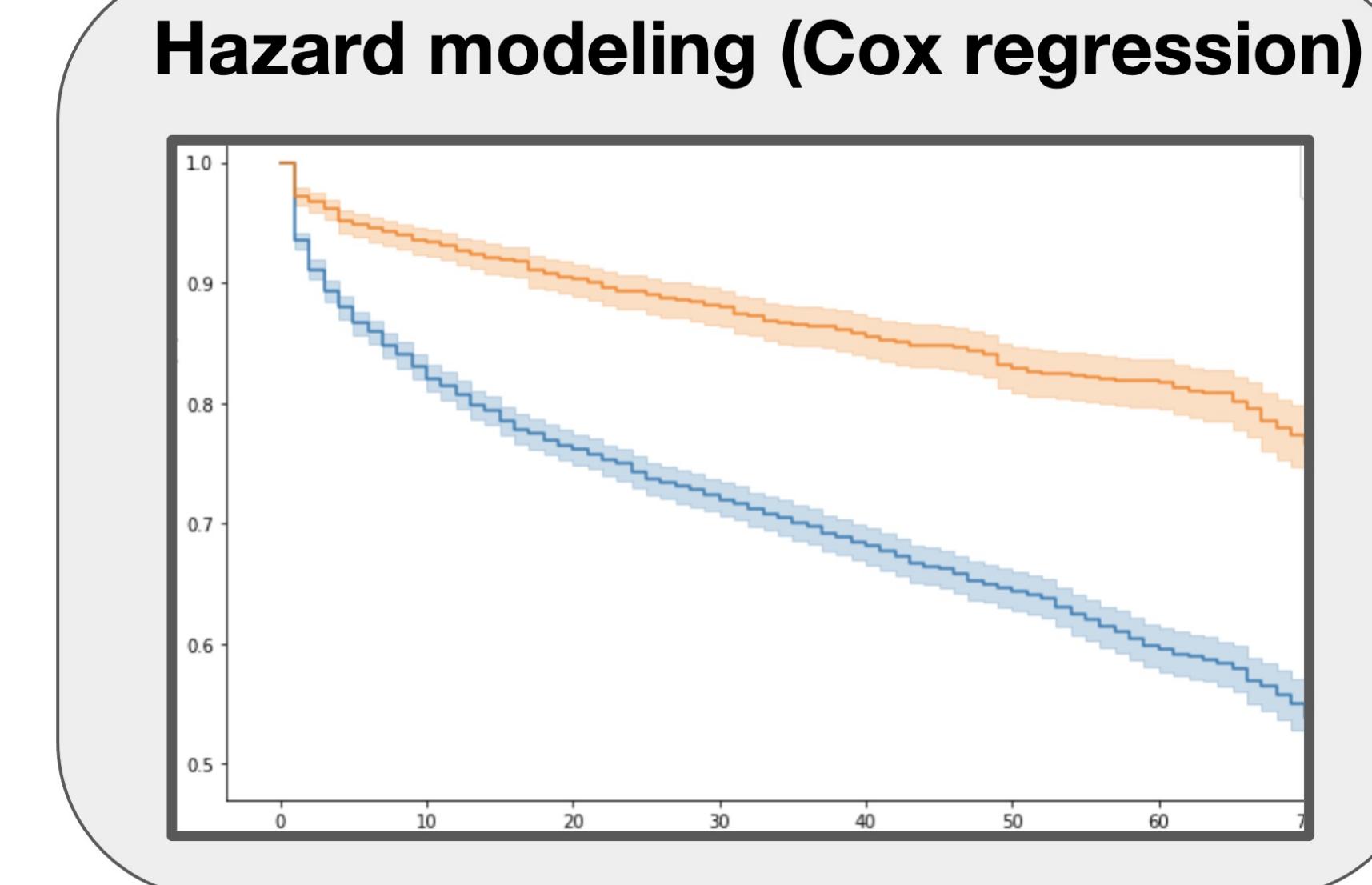
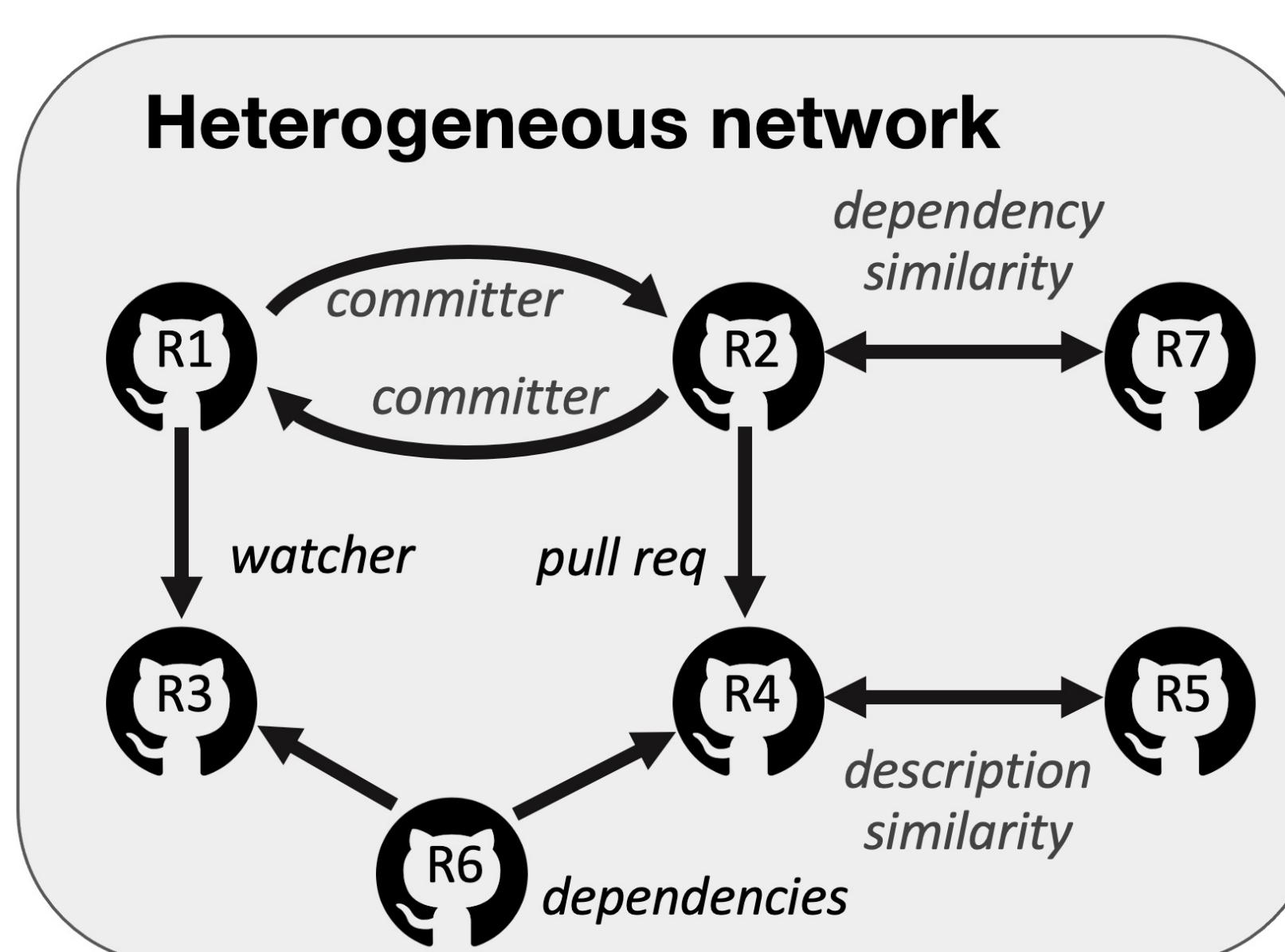
12 popular quality assurance tools

Continuous integration	Dependency management	Code coverage reporters	Cross browser testers
build passing	dependencies up to date	coverage 94%	Firefox Chrome 82 7 86 7
Travis Circle Appveyor Codeship	David Bithound Gemnasium	Coveralls Codeclimate Codecov Codacy	Saucelabs

~86,000 npm package repositories

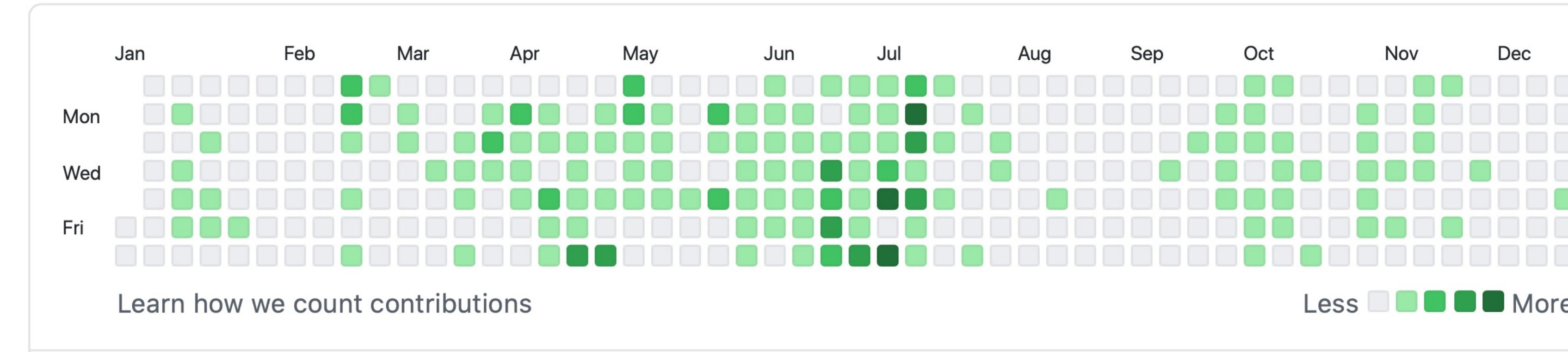
GitHub
npm

For each tool:



The Dark Side of transparency and signaling

Gamification is not always healthy



Contribution graph can be harmful to contributors #627

! Open

mxsasha opened this issue on Apr 1, 2016 · 189 comments



mxsasha commented on Apr 1, 2016

A common well-being issue in open-source communities is the tendency of people to over-commit. Many contributors care deeply, at the risk of saying yes too often harming their well-being. Open-source communities are especially at risk, because many contributors work next to a full-time job.

Any mechanism in our community that motivates people to avoid taking breaks and avoid stepping back, can be harmful to the well-being of contributors and is thereby harmful to open source as a whole. Even though it was probably introduced with the best intentions. If our interests are really in supporting open-source long-term, this graph should be removed or substantially changed so that it no longer punishes healthy behaviour. For example, what if we would give people achievements for taking breaks instead of working non-stop?

I therefore want to ask you to consider removing or substantially changing the contribution graph and

Developers are aware of each other's gender

Survey, 816 responses

Which of the following characteristics of your team members are you aware of?

- 74% • Programming skills
- 48% • **Gender**
- 45% • Real name
- 42% • Social skills
- 40% • Country of residence
- 39% • Personality
- 31% • Reputation as programmer
- 30% • Ethnicity
- 30% • Employment
- 28% • GitHub experience
- 26% • Educational level
- 23% • Age
- 11% • Hobbies
- 4% • Political views

“I have used a fake GitHub handle [...] so that people would assume I was male”



Again, lots of anecdotes

Experiences working in a diverse team

"code sees no color or gender"

Meritocracy; no effects of diversity

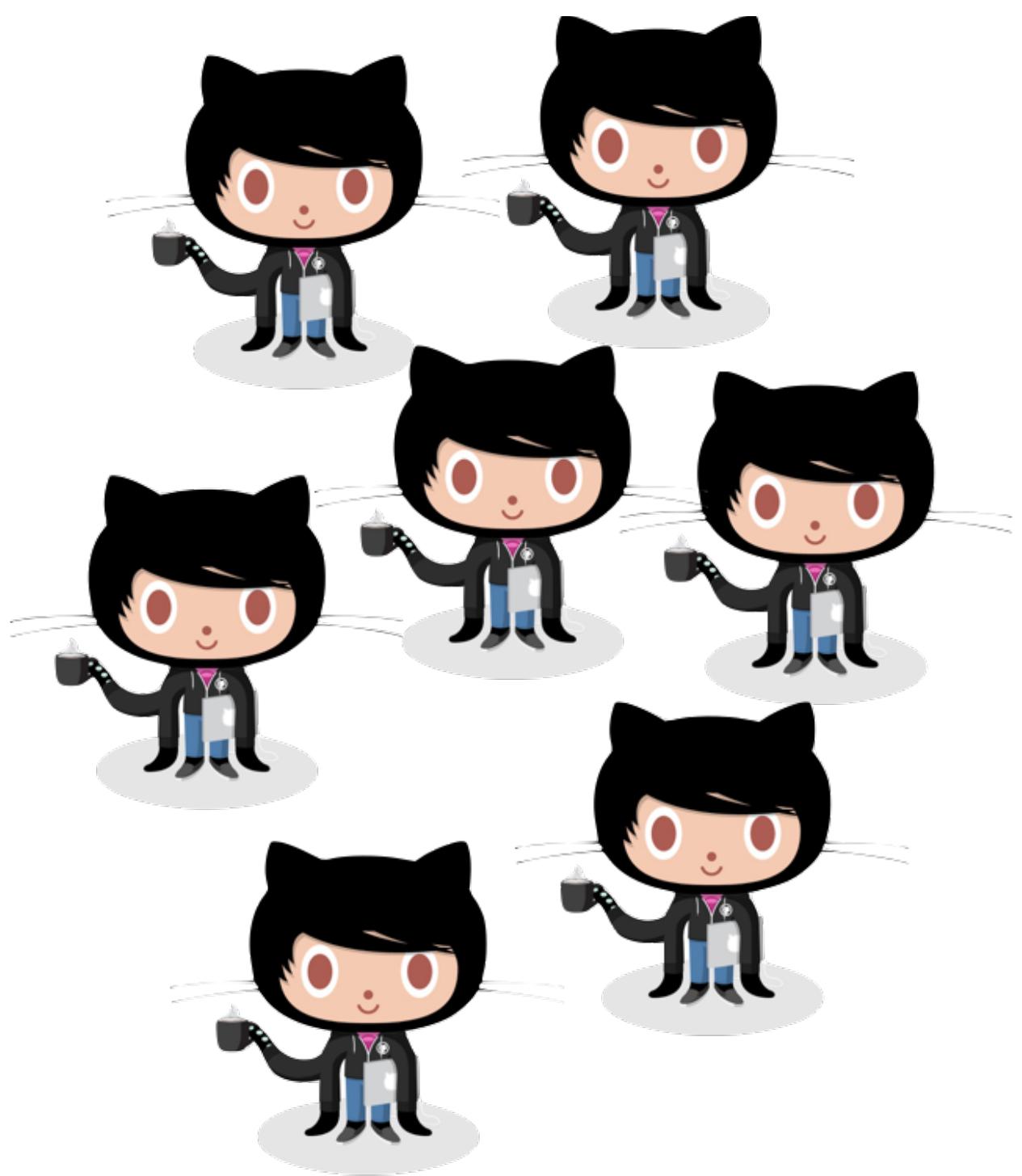
"diverse viewpoints often lead to **lively discussions and new ideas**"

Positive effects of diversity

"I have used a **fake GitHub handle** (my normal GitHub handle is my first name, which is a distinctly female name) **so that people would assume I was male**"

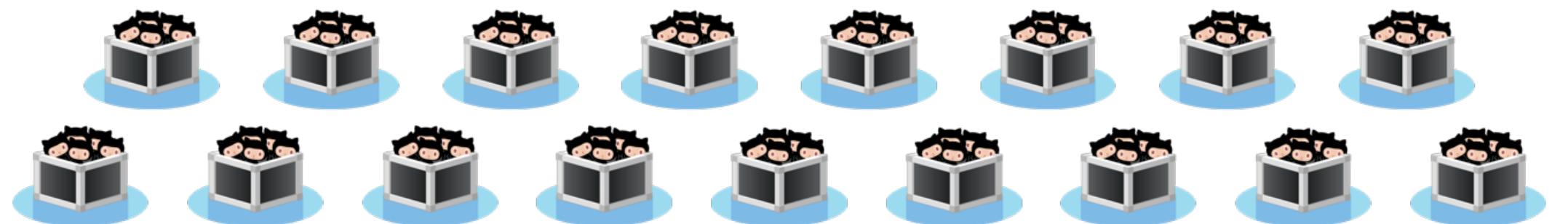
Negative effects of diversity

Which tends to be more effective, on average?



Natural experiment

1. Mine data from many **collaborative projects**



2. Compare **outputs produced per unit time**
in more/less diverse teams



Gender diversity
= mix women/men

Simplifying assumption:
gender is binary



Tenure diversity
= mix junior/senior
GitHub coding experience

Human resources

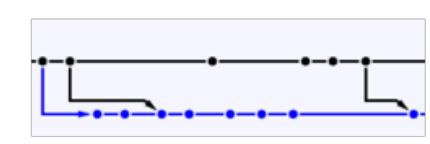


Team size



Experience

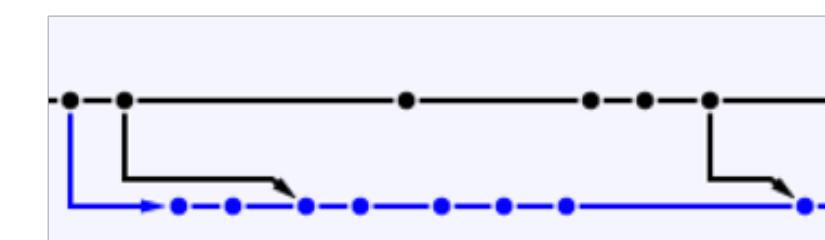
Project size



Total commits

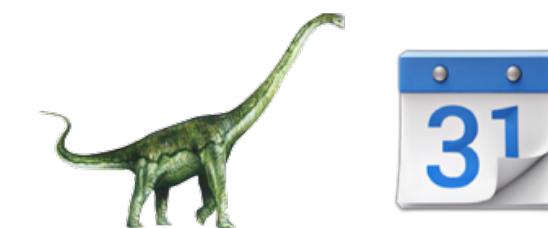
Response

Productivity
(#commits/quarter)



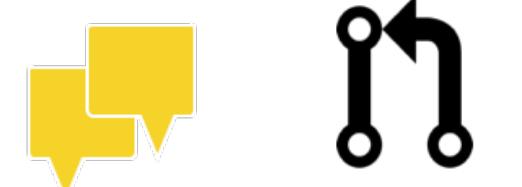
Controls

Evolution of GitHub
& time passing



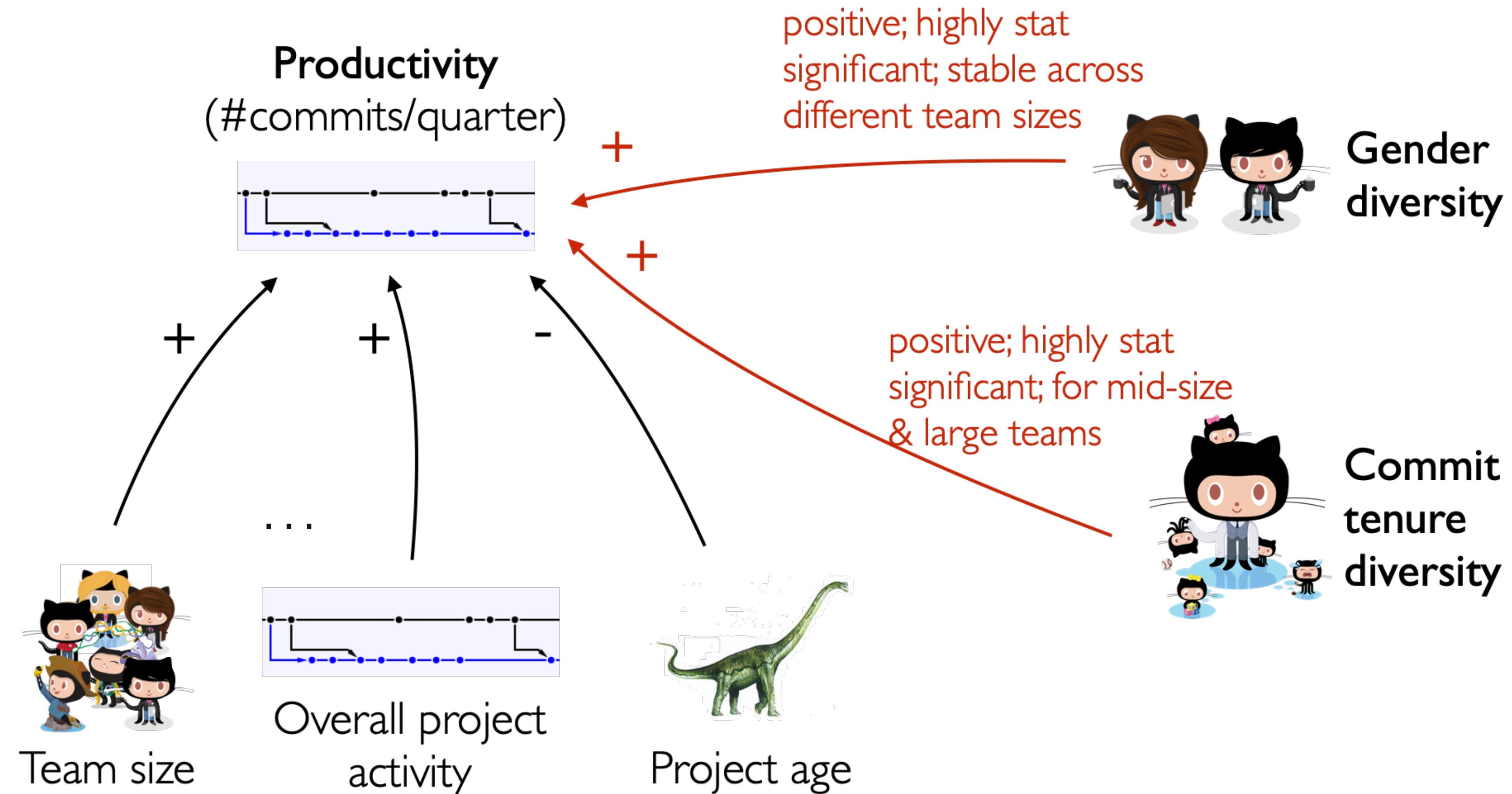
Project age Time

Popularity /
Distributed
development



Comments Forks

Increased diversity correlates with higher productivity



- Gender and tenure diversity in GitHub teams. Vasilescu, B., Posnett, D., Ray, B., Brand, M.G.J. van den, Serebrenik, A., Devanbu, P., and Filkov, V. *CHI 2015*

But small effects!

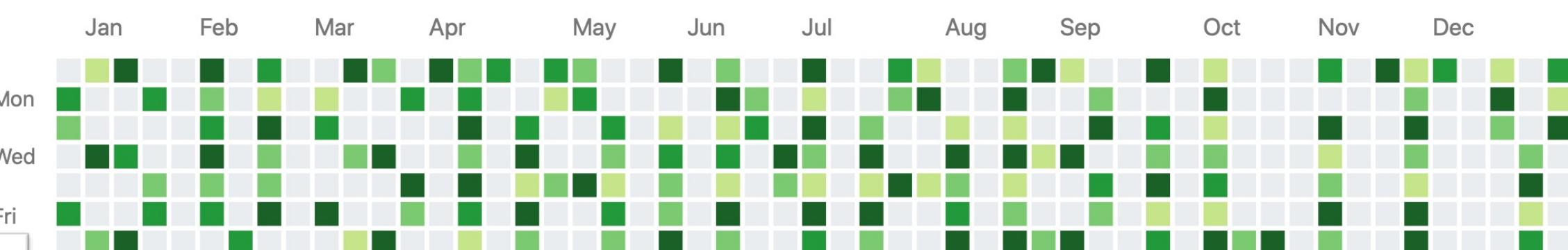
Aside: Inclusivity helps everyone



© Anita Sarma & Margaret Burnett, Oregon State U

Summary

Sustaining Open Source remains a challenge



Heaps of data

Transparency is a defining characteristic of social coding platforms



More than
repositories

STR

Carnegie Mellon University School of Computer Science

- Rich inference and level of automation

Badges are Reliable Signals
Mostly

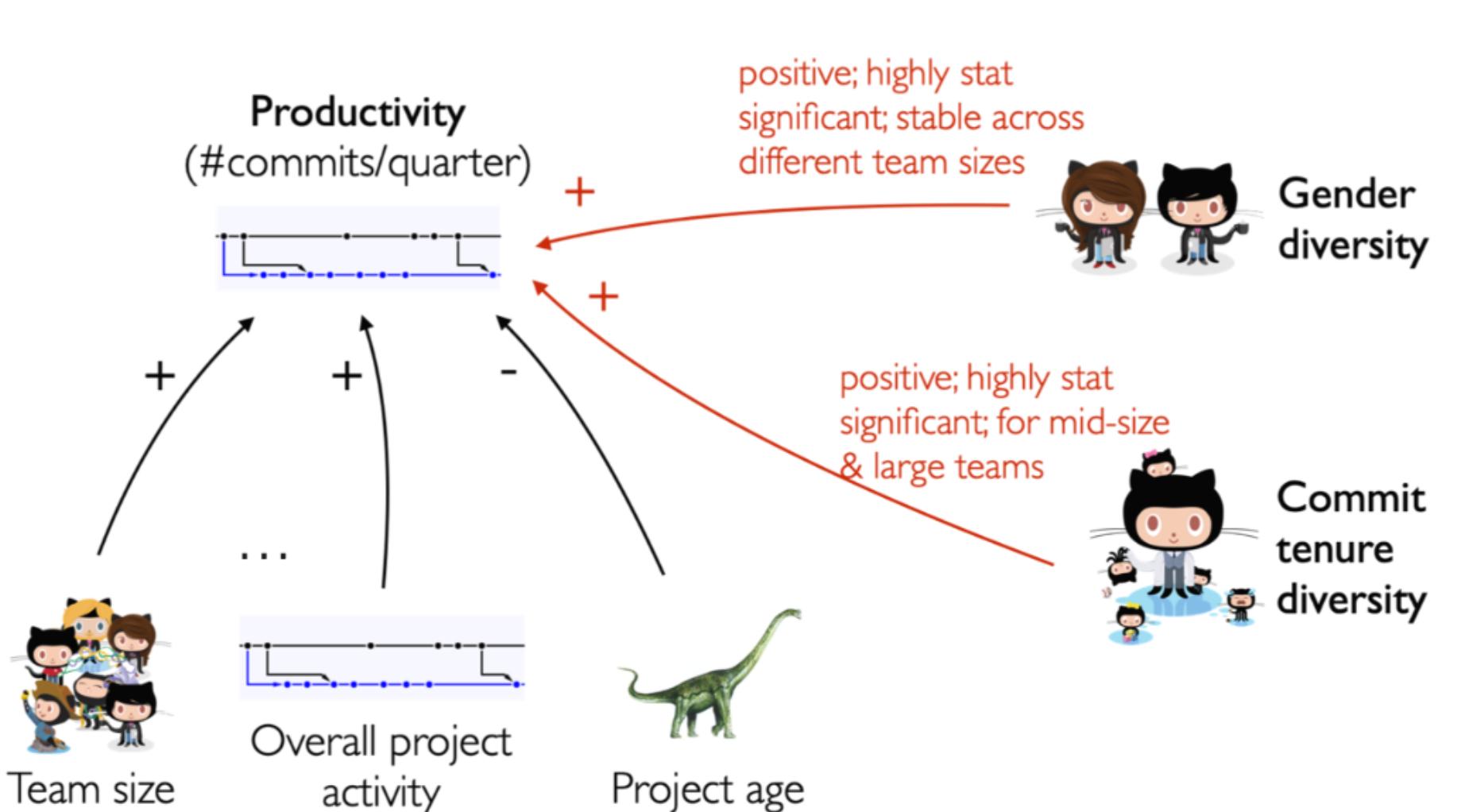
Key takeaway: Stronger social connections explain adoptions

STR

Carnegie Mellon University School of Computer Science

Overview: Multiple case study

Increased diversity correlates with higher productivity



STRUDEL sustainability research on ...

Open-source projects

Project practices

- [ICSE 2020](#) (forking)
- [ESEC/FSE 2019](#) (forking)
- [ESEC/FSE 2018](#) (abandonment factors)
- [FSE 2016](#) (breaking changes)

Attracting contributors

- [MSR 2020](#) (Twitter)
- [CSCW 2019](#) (signals)
- [ESEC/FSE 2015](#) (social connections)

Funding models

- [ICSE 2020](#) (donations)

Transparency and signaling

- [ESEC/FSE 2020](#) (diffusion of practices)
- [ICSE 2018](#) (badges)

Open-source people

Stress, burnout, disengagement

- [ICSE NIER 2020](#) (toxic language)
- [ICSE 2019](#) (overwork)
- [OSS 2019](#) (dropout and survival analysis)

Diversity and inclusion

- [ICSE 2019](#) (social capital)
- [CHI 2015](#) (gender & tenure)
- [CHASE 2015](#) (survey)

Acknowledgements



Courtney Miller



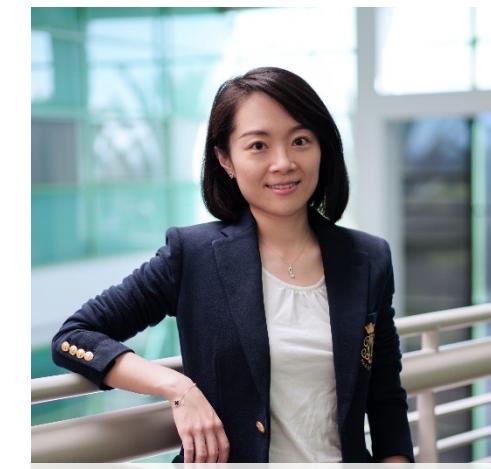
Anita Brown



Asher Trockman



Jim Herbsleb



Shurui Zhou



David Widder



Anita Sarma



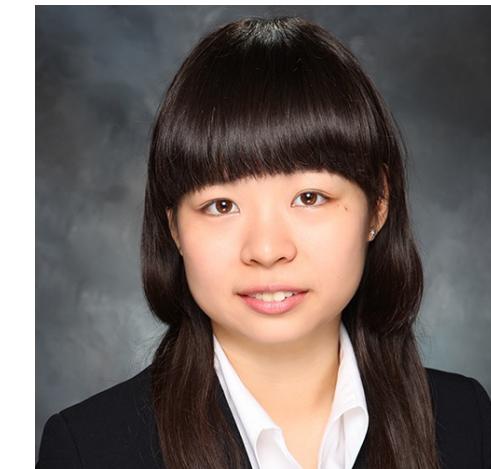
Cassandra Overney



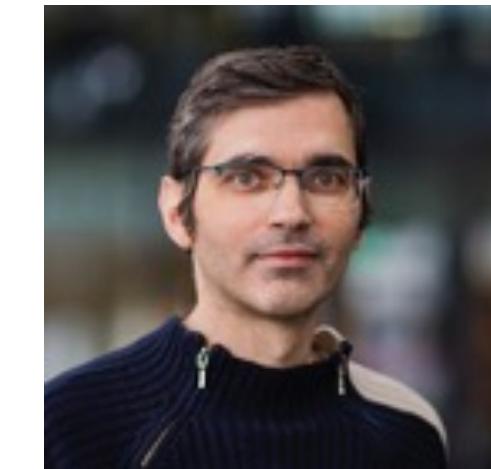
Audris Mockus



Alex Nolte



Sophie Qiu



Alex Serebrenik



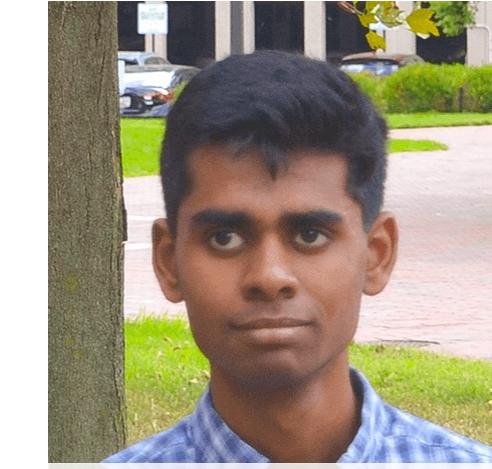
Marat Valiev



Laura Dabbish



Lily Li



Naveen Raman



National Science
Foundation

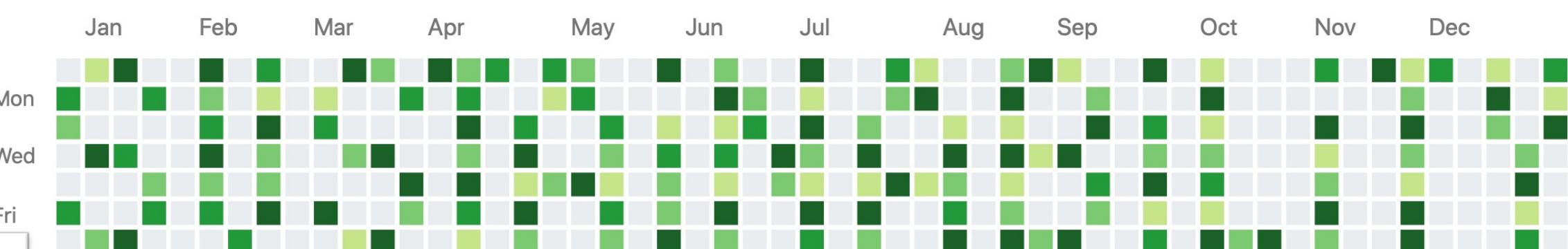


Alfred P. Sloan
FOUNDATION



FORDFOUNDATION

Sustaining Open Source remains a challenge



Heaps of data

Transparency is a defining characteristic of social coding platforms

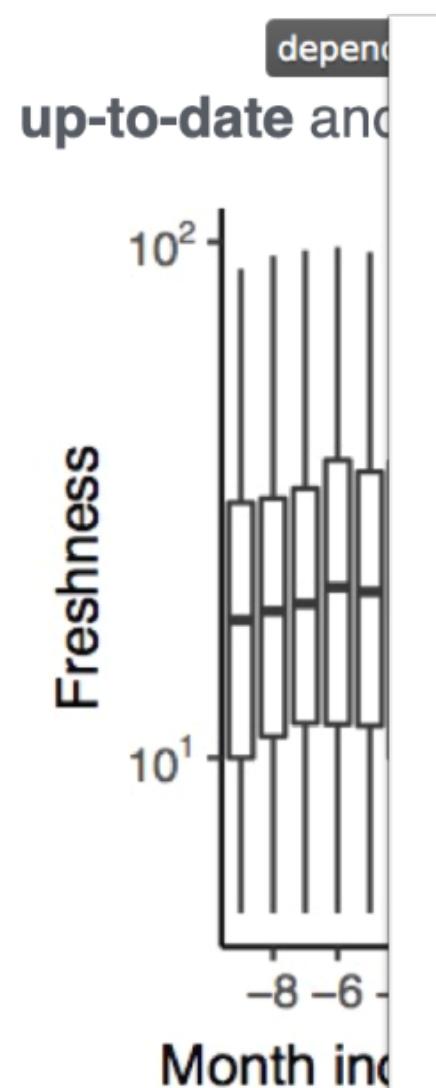


More than
repositories

Carnegie Mellon University School of Computer Science

- Rich inference and level of detail

Carnegie Mellon University School of Computer Science



Carnegie Mellon University School of Computer Science

12 popular CI tools

Continuous integration build passing

Travis Circle Appveyor Codeship

For each tool:

How to use

12 popular CI tools

Continuous integration build passing

Travis Circle Appveyor Codeship

For each tool:

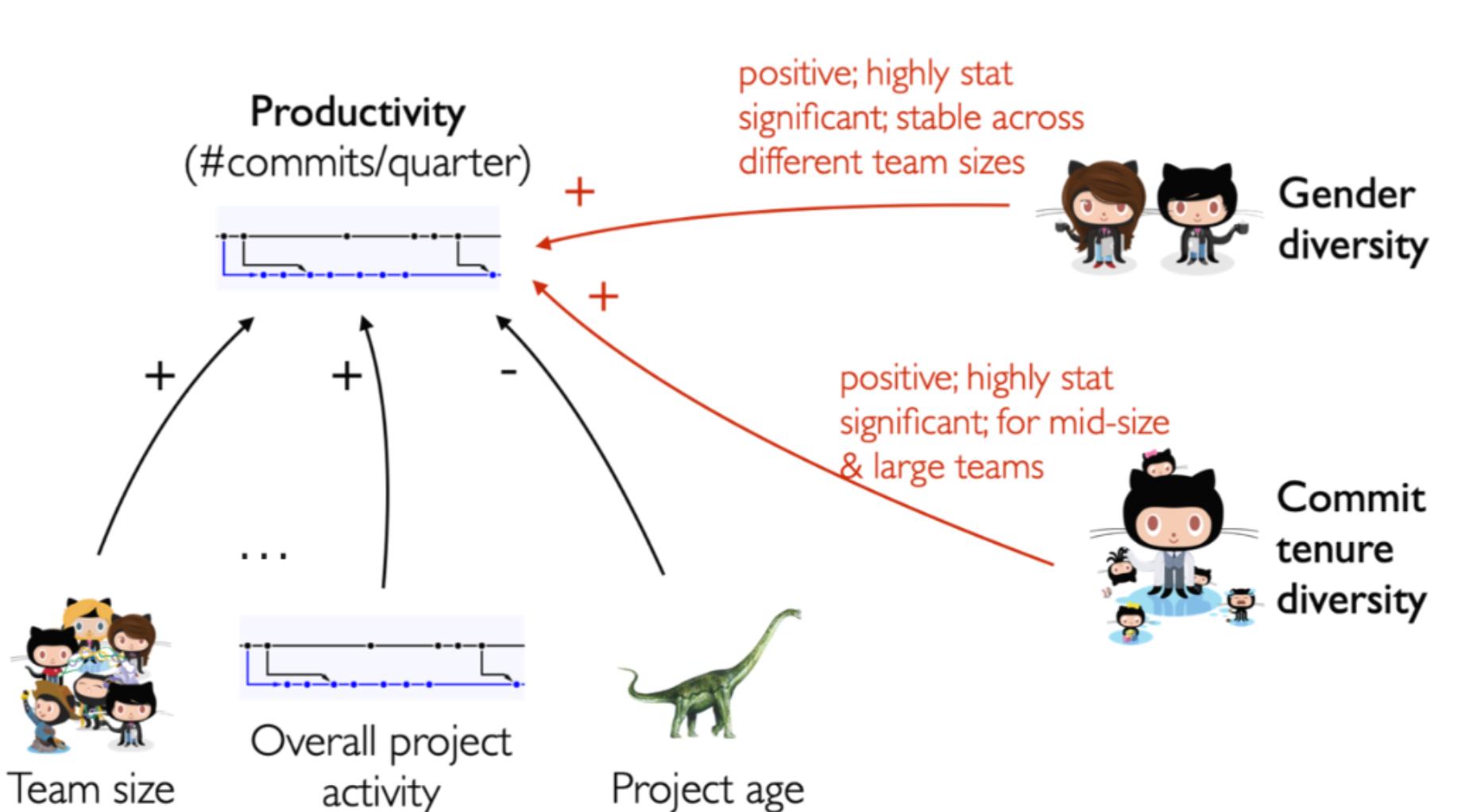
How to use

Badges are Reliable Signals

Mostly

Key takeaway: Stronger social connections explain adoptions

Increased diversity correlates with higher productivity



Bogdan Vasilescu

@b_vasilescu

vasilescu@cmu.edu